CBEB

FORTALEZA 2026

XXX CONGRESSO BRASILEIRO
DE ENGENHARIA BIOMÉDICA

Proceedings

ANAIS DO XXX CONGRESSO BRASILEIRO DE ENGENHARIA BIOMÉDICA

28 DE SETEMBRO
A 02 DE OUTUBRO DE 2026

FORTALEZA – CE
BRASIL



SBEB
SOCIEDADE BRASILEIRA
DE ENGENHARIA BIOMÉDICA

SBEB 50 ANOS
Integrando
ciência e vida

CBEB FORTALEZA 2026

XXX BRAZILIAN CONGRESS ON BIOMEDICAL ENGINEERING

Conference Organizers, Sponsors and Exhibitors

CONFERENCE ORGANIZERS

CBEB FORTALEZA 2026

SBEB SOCIEDADE BRASILEIRA DE ENGENHARIA BIOMÉDICA

SBEB 50 ANOS Integrando ciência e vida

SPONSORS

CAPES

CNPq

SUS Sistema Único de Saúde

FUNCAP

Sistema FIEC SESI | SENAI | IEL

SEBRAE

Fecomércio CE · Sesc · Senac — Sistema Comércio

EXHIBITORS

MINIPA EDUCATION

Verasonics® The leader in Research Ultrasound™

EMBRAPII Empresa Brasileira de Pesquisa e Inovação Industrial

FUNDAÇÃO FASTEF

SEBRAE

ESC

OPERATIONAL SUPPORT

E+ ASSESSORIA TÉCNICA

FUNDAÇÃO FASTEF

CBEB
FORTALEZA 2026

XXX BRAZILIAN CONGRESS
ON BIOMEDICAL ENGINEERING



Institutional Support and Partners

Universidade Federal do Ceará

Rede iCC Saúde

ANVISA

ABIMO
ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE ARTIGOS E EQUIPAMENTOS MÉDICOS, ODONTOLÓGICOS, HOSPITALARES E DE LABORATÓRIOS

SBIS

Abeclin

ABiMED
TECNOLOGIA. SAÚDE. VIDA.

abraidi

Fecomércio CE · Sesc · Senac
Sistema Comércio



SBEB
SOCIEDADE BRASILEIRA
DE ENGENHARIA BIOMÉDICA

SBEB 50 ANOS
Integrando ciência e vida

CBEB
FORTALEZA 2026
XXX CONGRESSO BRASILEIRO
DE ENGENHARIA BIOMÉDICA

XXX CBEB Organizing Committee

SBEB Board
Sônia Maria Malmonge
(President)
Antonio Adilton Oliveira Carneiro
(Vice-President)

CBEB 2026 Presidency
Victor Hugo Costa de Albuquerque
(President)
Paulo César Cortez
(Vice-President)

Program Coordination
José Antônio Fernandes de Macedo
(General Coordinator)
João Paulo Papa
Guilherme Barreto

Executive Secretariat
Jerovane Nascimento

Lectures and Panels
Jose Wally Mendonça Menezes

Short Courses / Minicourses
Maria da Conceição Pereira Fonseca
Rodrigo Costa Félix

Innovation Committee
Luma Rissatti Borges do Prado
Murilo Contó
Marcello Carvalho dos Reis
Fabrício Lima Brasil

Advisory Committee
Rodrigo Astolfi
Antônio Carlos da Silva Barros
Lucyanne Araujo Braz

Editorial Committee
Alcimar Barbosa Soares
João Salinet
Fabrício Gonzalez Nogueira
Auzuir Ripardo de Alexandria

Administrative and Institutional Support
E+ Assessoria Técnica
Fundação FASTEF

SBEB
SOCIEDADE BRASILEIRA
DE ENGENHARIA BIOMÉDICA

SBEB 50 ANOS
Integrando ciência e vida

CBEB

FORTALEZA 2026

XXX BRAZILIAN CONGRESS ON BIOMEDICAL ENGINEERING

Academic Support and Innovation Committee

The Academic Support and Innovation Committee strengthens CBEB 2026 in essential areas, promoting integration, creative solutions and support for the development of the event's activities, with the involvement of students, researchers and professionals committed to the advancement of Biomedical Engineering in Brazil.

36 MEMBERS | **5** FRONTS | **+10** INSTITUTIONS

FRONTS OF ACTION

- Technology
- Operations and Logistics
- Academic Support
- Communication and Marketing
- Public Service

COMMITTEE MEMBERS

1. Francisco Ítalo Guilherme da Silva — UFC
2. Adriell Gomes Marques — UFC
3. Marcus Vinícius Cândido de Figueiredo — IFCE
4. Laís Ferreira Sombra — UFC
5. Pedro Joás Freitas Lima — UFC
6. Luisa Miranda Timbó Mendes — UFC
7. Daniel Santos da Silva — UFC
8. Felipe Cordeiro de Sousa — UFC
9. Alan Marques da Rocha — UFC
10. Danilo Felipe Neto — UFC
11. Luis Francisco Sousa Leal Gomes — UFC
12. Anna Beatriz Brito de Souza Almeida — UFPE
13. Emilly Alexsandra de Oliveira Silva — UFPE
14. David Felipe de Brito Araújo — UPPE
15. Renata Benaion Silva Do Vals — UFPA / ITEC
16. Déborah Evelyn Leite Alcantara — UFPE
17. Rodrigo Alves Alonso — POLI-USP
18. Arianne Viana Alves da Silva — UFPE
19. Júlio Cardoso Scorpion — UFSC
20. João Lucas Pereira dos Santos de Paula — UFSC
21. Nestor Cavalcante Teixeira Neto — UFC
22. Pedro Florencio de Almeida Neto — UFC
23. Carlos Eduardo do Nascimento Silva — IFCE
24. Alicia Gabriela Saldanha do Nascimento — UFPA
25. Rogerio Lopes Vieira Cesar — IFCE
26. Naira Magalhães de Sousa — UFC
27. Eduardo Henrique Teixeira — UNIFEI
28. Bruno Henrique Aragão König — UFPE
29. Sheila Cristina de Moraes — NOVOESTE
30. Ana Luiza Matos da Silva — IIN-ELS/ISD
31. Jeisilyane Silveira dos Santos — UFC
32. Stella Machado Gualti — UFSCar
33. Eliana Figueiredo Borges — UFU
34. Rômulo Cesar Cunha Lima — UFC
35. Marya Thayssa Conceição Bezerra — UNILEÃO
36. Letícia Nascimento de Oliveira — ISD

CBEB

FORTALEZA 2026

XXX BRAZILIAN CONGRESS ON BIOMEDICAL ENGINEERING

International Collaboration and Awards Committees

International Collaboration Committee

The International Collaboration Committee supports scientific exchange and international dialogue at CBEB 2026, promoting cooperation among researchers, institutions and biomedical engineering communities worldwide.

- **Witold Pedrycz**
  University of Alberta, Canada
- **Qammer H. Abbasi**
  University of Glasgow, UK
- **Robertas Damaševičius**
  Kaunas University of Technology, Lithuania
- **Kemal Polat**
  Bolu Abant Izzet Baysal University, Turkey
- **Senthil Kumar Jagatheesaperumal**
  Mepco Schlenk Engineering College, India
- **D. Jude Hemanth**
  Karunya University, India
- **Roberto Muñoz**
  Universidad de Valparaíso, Chile
- **João Alexandre Lobo Marques**
  University of Saint Joseph, Macau, China
- **Khan Muhammad**
  Sungkyunkwan University, Republic of Korea
- **Mohamed Abdel-Basset**
  Zagazig University, Egypt
- **Ali Hassan Sodhro**
  Linköping University, Sweden
- **María de los Ángeles Quezada**
  Instituto Tecnológico de Tijuana, Mexico
- **Javier Del Ser Lorente, SM IEEE**
  University of the Basque Country, Spain
- **Wanqing Wu**
  Sun Yat-Sen University, China
- **Tanveer Hussain**
  Edge Hill University, UK
- **Marcin Woźniak**
  Silesian University of Technology, Poland
- **Rahma Fourati**
  FSJEGJ, University of Jendouba, Tunisia

Awards Committee

The Awards Committee is responsible for coordinating the evaluation and recognition of outstanding contributions presented at CBEB 2026.

- **Arthur Frota**
  Escale Tecnologia / AFPAR Multi Capital
- **Alexandre Lobo Marques**
  University of Saint Joseph, Macau SAR, China

SBEB
SOCIEDADE BRASILEIRA DE ENGENHARIA BIOMÉDICA

SBEB 50 ANOS
Integrando ciência e vida

Expediente e apresentação

XXX Congresso Brasileiro de Engenharia Biomédica

EVENTO
XXX Congresso Brasileiro de Engenharia Biomédica

REALIZAÇÃO
28 de setembro a 2 de outubro de 2026 · Fortaleza, Ceará, Brasil

ENTIDADE PROMOTORA
Sociedade Brasileira de Engenharia Biomédica (SBEB)

PUBLICAÇÃO SERIADA
Anais do Congresso Brasileiro de Engenharia Biomédica · ISSN 2359-3164

Apresentação

Os anais do XXX Congresso Brasileiro de Engenharia Biomédica reúnem 597 trabalhos científicos completos, organizados em 15 áreas temáticas, a partir dos arquivos disponibilizados à organização do evento.

A publicação apresenta sumário temático com referências às páginas iniciais dos artigos e links internos para navegação no documento digital. Os textos científicos são reproduzidos a partir dos arquivos que integram esta compilação.

Organização do volume

Capa institucional · Expediente e apresentação · Sumário por área temática · Artigos científicos completos.

CONSULTA E ACESSO

https://sbeb.org.br/cbeb2026/

https://zenodo.org/communities/cbeb2026papers/records

Sumário

VOLUME 01 · 01

Processamento de Sinais e Imagens Biomédicas

ID 753 · A CAD System for Tooth Segmentation in Panoramic Radiographs Using U-Net Variants 29

ID 666 · A ConvNeXt-Small-Based Classifier for Ischemic Stroke Identification in Computed Tomography Images 37

ID 360 · A Deep Learning-Based Architecture for Automated Nuclei Quantification in Dense Microscopy Images from Transwell Assays 45

ID 359 · A Dual-Stage Framework for Mammographic Nodule Analysis Integrating Decision Tree Classification and Marker-Controlled Watershed 53

ID 7 · A Multi-Agent Ecosystem-Based Neurosymbolic Unsupervised Optimum-Path Forest for Functional Brain Mapping based on EEG Microstate Analysis 61

ID 16 · A Multimodal Neurorehabilitation Platform for Post-Stroke Hand Rehabilitation 69

ID 788 · A Short-Scale Transition into Ventricular Fibrillation: Recurrence Quantification and the Limits of Minute-Scale Anticipation 77

ID 342 · Advanced AI-Driven Signal Processing for Glaucoma Detection Using Retinal Imaging Methods and Experimental Insights 85

ID 253 · Algorithm for Muscle Thickness Tracking on Ultrafast Ultrasound 91

ID 516 · An Automated Thermographic Pipeline for Temperature Reconstruction, Drift Correction, and Hand Warm-Up Assessment in Acoustic Guitarists 99

ID 789 · Analysis and Prediction of Medical Service Utilization Among Adults Aged 50 and Older: A Machine Learning Approach 107

ID 878 · Análise Comparativa de Arquiteturas e Pré-Processamentos para Segmentação de Vasos em Imagens de Fundoscopia 115

ID 549 · Application of Convolutional Neural Networks for Brain Tumor Classification in MRI Images 123

ID 702 · Application of Radiomics and Deep Learning for the Classification and Grading of Gliomas in Magnetic Resonance Imaging: A Literature Review 131

ID 404 · Architectural Rigidity vs. Scaling Laws in Arrhythmia Classification 139

ID 381 · Arquitetura IoT Distribuída para Telemetria e Visualização de ECG de Alta Fidelidade em Tempo Real via ESP32-S3 e WebSockets 147

ID 379 · Automated Classification of Feline Mammary Carcinoma via Spatial Quantification of Intratumoral Type I and Type III Collagens 155

ID 556 · Automated Nuclei Segmentation and Counting in Human Liver Histological Images Using Watershed 159

ID 175 · Automated Screening for Metallic Implants in MRI Scout 164

ID 109 · Automated Visual Target Detection in RSVP Tasks from EEG Signals using Wavelet Transform and Machine Learning 172

ID 539 · Automatic Detection of Cervical Spine Anatomical Landmarks in Lateral Radiographs Using Deep Learning 177

ID 207 · Autonomic Responses to COVID-19 Vaccines During Tilt Testing 183

Sumário

VOLUME 01 · 02

Processamento de Sinais e Imagens Biomédicas

ID 426 · Autoregressive and Fractal Mammographic Features for Statistically Differentiating HER2-Positive Breast Cancer 187

ID 150 · Avaliação Operacional da Triagem de Retinopatia sob Distribution Shift 192

ID 725 · Berouti Spectral Subtraction for EMG Walking Signals 200

ID 227 · Biometric and Demographic Pattern Mining in Electrocardiographic Signals: A Study with Machine Learning Algorithms 207

ID 886 · BioStatusIA: A Multi-Agent Architecture for Multimodal Clinical Decision Support 218

ID 383 · Brain Symmetry Index after tDCS Rehabilitation 223

ID 718 · CAD for Brain Tumor Segmentation and Classification 230

ID 239 · Cartesian vs. Cylindrical ECG Models in Nonlinear Kalman Filtering 238

ID 286 · Characterisation of Motor Profiles in Bilateral Extension and Flexion Movements in Individuals with Parkinson's Disease 246

ID 367 · Characterization of Thermal Profiles in Pressure Injuries Using the Thermal Laplacian 250

ID 493 · Classification of Dementia Stages in Magnetic Resonance Images by Pre-Trained CNNs and Ensemble 256

ID 178 · CNNs para Classificação de Câncer Renal em CT 263

ID 8 · Collaborative Multi-Agent Networks for Multimodal Analysis of Parkinson’s Disease 270

ID 405 · Color Deconvolution and Marker-Controlled Watershed Framework for Intracellular Parasite Quantification in HTR8/SVneo Cells 277

ID 196 · Compact Deep Learning Architecture for Multi-Label ECG Classification 285

ID 483 · Comparative Analysis of Deep Learning Architectures for Diabetic Retinopathy Classification 293

ID 340 · Comparative Analysis of Digital Image Processing Techniques for Automated Segmentation of Wound Healing Assays in Infected Cultures 301

ID 158 · Comparative Analysis of the Performance of Analog and Digital Band-Stop Filters Applied to Power Line Noise Suppression in Electromyographic Signals 309

ID 641 · Comparative Evaluation of CNNs, Vision Transformers, and Ensemble Methods for Breast Cancer Lymph Node Metastasis Detection 316

ID 545 · Comparative Evaluation of Neural Network Architectures for ADHD Classification Using EEG Signals 324

ID 53 · Comparative Study of Blind Source Separation Techniques for ASD Classification in EEG Signals 332

ID 869 · Computational Pipeline for Detecting and Quantifying Spectral Modulation in EEG for 12-Hz Binaural Stimulation Studies in the Context of Alzheimer's Disease 338

ID 226 · Computational Tool for Automated Detection of Multiple Sclerosis Brain Lesions Using Magnetic Resonance Imaging 344

ID 364 · Computer Vision System for Eye Movement and Eyelid Dynamics Analysis 350

ID 686 · Computerized Tomography Image Generation for Ischemic Stroke Using GANs 355

Sumário

VOLUME 01 · 03

Processamento de Sinais e Imagens Biomédicas

ID 437 · Contrast Curve Correction for Digital Mammography 361

ID 392 · Creation of a Database for Clinical Validation of an Artificial Intelligence Algorithm for Detecting Failures in Invasive Blood Pressure Monitoring in Intensive Care Units 369

ID 467 · Data Above the Model: A Study of Three Deep Learning Approaches for Pulmonary Disease Detection in Chest Radiographs 374

ID 287 · Deep Learning Applications in Post-Stroke Automatic Speech Analysis: A Literature Review 382

ID 534 · Deep Learning for Pulmonary Nodule Detection 390

ID 384 · Deep Learning Models for Real-Time Heart Rate Estimation Using Remote Photoplethysmography: A Single-Subject Pilot Study of an ICU Patient 398

ID 52 · Deep Learning-Based Classification of Breast Lesions in Ultrasound Images Using Transfer Learning and Decision Threshold Calibration 405

ID 888 · Deep Learning-Based Monaural Speech Enhancement on the VoiceBank-DEMAND Dataset 413

ID 594 · Detection and Segmentation of Oral Ulcers and Mucosal Inflammation Using Hyperspectral Imaging and Deep Neural Networks 420

ID 541 · Detection of Melanoma Presence in Dermoscopic Images Using a Transfer Learning Approach with ResNet50 428

ID 679 · Disentangling Persistent ECG ST/T Alterations and Local Beat Perturbations through Multiscale Physiological Labeling and Recurrence Quantification Analysis 434

ID 938 · Divergent Modulation of Cortical and Intermuscular Coupling by Object Properties During Grasping 442

ID 261 · ECG-Based Chagas Disease Classification under Weak Self-Reported Labels: A Comparison of Target Definitions in CODE-15% 450

ID 170 · EEG-Based Autism Spectrum Disorder Assessment Using Holo-Hilbert Spectral Analysis and Machine Learning 458

ID 572 · Effect of Peripheral Sinusoidal Electrical Stimulation on Cortical Oscillations During an Archimedean Spiral Drawing Task 463

ID 885 · Effects of Peripheral Electrical Stimulation Waveforms on Tremor in Parkinson's Disease: A Pilot Study 471

ID 235 · Ensemble Learning with Lightweight Convolutional Neural Networks for Mosquito Species Classification 479

ID 766 · Estimating Model Observer Response in Digital Mammography Using Radiomics and Machine Learning Regression Models 487

ID 441 · Evaluating CLAP Audio Foundation Model for Swallowing Event Detection 495

ID 508 · Evaluating Deep Neural Network Architectures for Motor Unit Firing Classification from Decomposed HD-sEMG Signals 503

ID 506 · Evaluation of a Hybrid RR–P-Wave Approach for Automatic Atrial Fibrillation Detection in ECG Signals 508

ID 674 · Evaluation of Movement Smoothness in the Four-Way Leaning Limits of Stability Test in Patients with TMD: A Pilot Study 515

Sumário

VOLUME 01 · 04

Processamento de Sinais e Imagens Biomédicas

ID 786 · Evaluation of Multimodal Data in the Classification of Patients with Dementia Using Explainable Artificial Intelligence 520

ID 865 · Evaluation of Wavelet-Based Filters for Denoising Gas Vesicles Ultrasound Responses 528

ID 133 · Experimental Simulation of Fracture Patterns in Child Physical Abuse 535

ID 859 · Experimental Study of EMG Signals for Prosthetic Control Using a Feature-Based Pipeline and Unsupervised Learning 543

ID 304 · Exploratory Analysis of Electroencephalographic Coherence in Neurocritical Patients with Suspected Brain Death: A Case Series 551

ID 841 · Extraction of Computational Descriptors Based on ABCD Criteria for Characterization of Skin Lesions 559

ID 13 · Ferramenta educacional para detecção de displasia epitelial oral a partir de casos de leucoplasia 567

ID 465 · Frozen GPT-2 Backbone for Few-Shot Congestive Heart Failure Detection from ECG Signals 575

ID 876 · Gas Vesicles Image Enhancement Using Median Filter in HSV Image Representation 582

ID 762 · Geometric Characterization for Personalized Face-Frame Interface on Assessing Sunglasses UV Blocking 590

ID 607 · Geometry-Guided Deep Learning for Intermediate Projection Synthesis in Digital Breast Tomosynthesis 597

ID 527 · Habitats in Medical Images: The Influence of Parameter Choice on the Segmentation Process 605

ID 824 · Heart Sound Detection Based on Wavelet Transform and Shannon Energy Envelope 613

ID 932 · Heart-Evoked Spectral Disturbances in Refractory Epilepsy 619

ID 365 · HMM for Auditory Attention Decoding 625

ID 168 · Horse Height Estimation from a Single Image 633

ID 468 · Howland-Based FES Output Stage Characterization 639

ID 628 · Hybrid Deep Feature Extraction and Machine Learning for Multiclass Dementia Classification Using EEG Wavelet Images 645

ID 146 · Hybrid Regularization for Faulty-Electrode EIT in Hemorrhagic Stroke 653

ID 423 · Hybrid ViT–Convolutional W-Net for Retinal Vessels 661

ID 537 · Identification of Elevated Cardiac Troponin from 12-Lead ECGs Using Deep Learning and Exploratory Generalization from Myocardial Infarction Models 669

ID 617 · Identification of Genotoxic Cells in the Micronucleus Test Using YOLOv7 Neural Networks 678

ID 528 · Image-Based Evaluation System for Paraná State Military Physical Fitness Test: 12-Minute Run 684

ID 216 · Impact of Region of Interest Representation Methods on Ovarian Tumor Classification in Ultrasound Images 691

ID 623 · Inertial Sensor Analysis of Voluntary Stopping and Freezing of Gait in Parkinson's Disease 698

ID 368 · Interpretable Structural-Deviation Descriptors of Osteosarcoma Radiographic Signs via the Universal Image Quality Index 704

Sumário

VOLUME 01 · 05

Processamento de Sinais e Imagens Biomédicas

ID 602 · Janelas de Decisão para a Detecção Online de Anomalias em EEGs via One-Class Classification 712

ID 389 · Korotkoff Sound Filtering Techniques for Exercise Stress Testing and Functional Assessment Protocols 720

ID 696 · Literature Review on Interpretable Electrocardiogram Analysis for Arrhythmias, WPW Ventricular Pre-excitation, and Glycemic Excursions 728

ID 382 · Low-Cost 2D Ultrasound Field Scanning Platform 736

ID 184 · Low-Cost Wrist Movement Recognition Using Two-Channel sEMG 744

ID 345 · Maximizing EEG Task-Related Information in a Motor Imagery Cue-Based Paradigm 752

ID 563 · Motion-Aware Evaluation of In-Ear PPG 758

ID 578 · Multi-Class vs. Multi-Label sEMG Gesture Classification 766

ID 157 · Multiclass Skin Cancer Classification with Grad-CAM 773

ID 265 · Multiparametric EEG Analysis in Comatose Patients: A Comparison Between Silence and Musical Stimulation 781

ID 523 · Multisensory Prototype for the Investigation of Sleep-Related Respiratory Events 789

ID 325 · Método Adaptativo Baseado em Variação Espectral para Compressão de Sinais de ECG de Longa Duração 796

ID 693 · Neuroimaging, Speech, and Sensors: A Scoping Review of Deep Learning Multimodal Fusion for Cognitive Decline Detection 804

ID 323 · Neurological Prognostication Post-Cardiac Arrest Using Continuous Wavelet Transform and Hybrid Deep Learning 812

ID 285 · Objective Characterisation of Pronation-Supination Movements in Individuals with Parkinson's Disease Undergoing Pharmacological Treatment 820

ID 475 · Observational Analysis of the Performance of Image Reconstruction Algorithm Techniques in Head CT Protocol 824

ID 445 · Open-Source Electrooculography Platform for Assistive Double Blink Detection 829

ID 554 · Phonocardiography for Elevated BP Classification 836

ID 118 · Pipeline automatizado pertinente, porém pequeno conjunto clínico e resultados estatísticos ainda sem correção adequada de múltiplas comparações. 844

ID 104 · Pipeline de wavelets e deep learning utiliza técnicas estabelecidas e apresenta contribuição predominantemente de aplicação. 851

ID 322 · Pix2Pix GAN-Based Translation of Dermoscopic Images into Clinical Skin Lesion Images: VGG16-Guided Losses and FID Evaluation 857

ID 840 · Potenciais Biomarcadores de Ressonância Magnética Funcional no Comprometimento Cognitivo Leve e na Doença de Alzheimer: Uma Revisão de Escopo 863

ID 593 · Pre-trained CNNs for Mammographic Microcalcifications 871

ID 700 · Prediction of Hospital Readmission Within 30 Days After Discharge with Wearable Temporal Signals 878

Sumário

VOLUME 01 · 06

Processamento de Sinais e Imagens Biomédicas

ID 600 · Preliminary Results of a Large Dataset of Biosignals for Training of Control Algorithms of Upper-Limb Prostheses 885

ID 511 · Preprocessing and Physiological Model Fitting of Smartphone-Based Pupillary Light Reflex Recordings 892

ID 234 · Prototype Vibration Stimulator for Evaluating Somatosensory Nerve Fibers: Application in Charcot-Marie-Tooth Disease 898

ID 683 · Radiomic Evaluation of Additive Manufacturing Materials for Optimizing Anthropomorphic Mammography Phantoms 904

ID 822 · Reconhecimento de Padrões Vocais Associados a Estresse e Emoções Negativas em Comunicação Ruidosa via Adaptação Adversarial de Domínio 911

ID 319 · Recurrence Quantification Analysis of Resting-State EEG in Post-Stroke Rehabilitation with tDCS and Extended Reality 919

ID 206 · Remote Photoplethysmography in ICU Patients 927

ID 350 · Retinopathy Detection: YOLO With and Without SAHI 935

ID 328 · Riemannian Classification of Movement Intention from EEG 943

ID 691 · Risk Classification for Sudden Cardiac Death Using Single-Lead ECG 949

ID 844 · Robotic Prosthesis Control Using Linear Discriminant Analysis of Residual Electromyographic Signals 956

ID 139 · Segmentação de Cárie Dentária por Severidade em Radiografias Panorâmicas com YOLO11n-seg e Slicing Aided Hyper Inference 964

ID 63 · Sensitivity Analysis between qEEG Metrics and Longitudinal Clinical Scores in Post-Stroke Patients 972

ID 51 · Sparse Time-Frequency Characterization of Ischemic ST-Segment Morphology in ECG Using Matching Pursuit Decomposition 980

ID 454 · Spatiotemporal Organization of Atrial Fibrillation via EGM and ECG: A Meta-Analysis 989

ID 603 · Spectral Residual Analysis and Feature Map Visualization of Deep Learning Denoisers for Digital Mammography 994

ID 50 · Stability of Microstates and Machine Learning in Resting-State Electroencephalography 1002

ID 347 · Study of Interpretable Deep Neural Networks in Brain-Computer Interfaces 1008

ID 164 · tDCS Effects in Post-Stroke Rehabilitation Using EEG LORETA 1016

ID 797 · Theta Alpha Ratio and Auditory Stimulus Type Modulate Alpha Power Lateralization During Spatial Attention 1024

ID 169 · Theta-Band Modulation in EEG Signals from Individuals with Autism Spectrum Disorder Using Holo-Hilbert Spectral Analysis 1032

ID 653 · Time-Domain Speech Signal Enhancement with Impulsive Noise Using U-Net: A Preliminary Step Toward Future Assessment of Auditory Attention 1040

ID 62 · Transfer Learning for Dental Pathology Detection in Panoramic Radiographs: YOLOv11 vs. Zero-Shot Grounding DINO 1048

Sumário

VOLUME 01 · 07

Processamento de Sinais e Imagens Biomédicas

ID 427 · TRCA vs. CCA for SSVEP BCI 1056

ID 149 · Triagem Multitarefa Explicável de Retinopatia e Glaucoma 1064

ID 282 · Um Sistema Baseado em CNN 3D para Diferenciação entre Esquizofrenia, Transtorno Bipolar e Controles Saudáveis com Base em Ressonância Magnética 1072

ID 606 · Variance Stabilization of Unprocessed Digital Breast Tomosynthesis Projections Using a Residual Neural Network 1079

ID 156 · Venous Thrombosis Classification with CNN and XAI 1087

ID 343 · X-Pan: Building, Documenting and Validating an OPG Dataset 1094

ID 860 · YOLO11 and YOLO26 Deep Learning Models for Diabetic Retinopathy Lesion Segmentation: A Comparative Analysis 1102

ID 190 · YOLOv8 Detection of Hematological Elements in Blood Smears 1108

Tecnologias Assistivas e Reabilitação

ID 782 · 3D-Printed Pediatric Wheelchair Wheels 1116

ID 767 · A Brief Review of Control Strategies for Functional Electrical Stimulation Cycling 1122

ID 496 · A Computer Vision and Non-Immersive Virtual Reality Platform for Automated Functional Assessment During Hospital Cardiorespiratory Rehabilitation 1130

ID 856 · A Multimodal Assistive Interface for Wheelchair Control Using Electrooculography, Capacitive and Inertial Sensor 1138

ID 243 · A Non-Immersive Virtual Tour System Integrated with a Smart Cycle Ergometer for Chronic Respiratory Disease Rehabilitation 1148

ID 87 · A Virtual Simulation System to Support Wheelchair Prescription Concerning Patient Environment 1156

ID 266 · A Wearable Edge-Computing HD sEMG Platform for Assistive Applications: Technology Readiness and Laboratory-Scale Reproduction Cost 1162

ID 420 · ACL-Digital Assist: E-Health Platform for Knee Rehabilitation Tracking 1169

ID 424 · Advanced Manufacturing for Auricular Prosthesis 1177

ID 307 · An HD sEMG-Based Virtual Reality Box and Blocks Test for Hand Dexterity Assessment 1184

ID 612 · An Integrated Platform for Alternating Peripheral Electrical Nerve Stimulation with Application to Parkinson's Disease Tremor 1190

ID 800 · App My Pill: Design and Case Demonstration of a Low-Data Reference-Based System for Loose Pill Identification 1198

ID 127 · Assistive Cutlery in an Institutional Setting 1206

ID 152 · Assistive System for Signed Alphanumeric Spelling 1210

Sumário

VOLUME 01 · 08

Tecnologias Assistivas e Reabilitação

ID 730 · Automatic Transcription of Monophonic Recorder Audio to Braille Music Notation Using Classical Signal Processing Methods 1215

ID 907 · Capturing Functional Upper-Limb Movements Using Low-Cost IMUs 1221

ID 589 · Clinical Applicability of Wearable Accelerometers for Functional Assessment of Older Adults in a Home Environment 1227

ID 260 · Combined Effects of Brain-Computer Interface and Virtual Reality Game with tDCS in Stroke Patients 1232

ID 213 · Composite Laminates for Wheelchair Seat Applications 1240

ID 579 · Cultura Maker Aplicada à Tecnologia Assistiva: Desenvolvimento de um Processo de Fabricação Digital para Bancos Terapêuticos Destinados à Reabilitação 1248

ID 711 · DCV: dispositivo de comunicação vibrotátil personalizável para pessoas surdas e surdocegas 1256

ID 945 · Desenvolvimento de Dispositivos Assistivos de Baixo Custo para Apoio à Escrita de Crianças com Paralisia Cerebral: Relato de Dois Casos 1264

ID 432 · Desenvolvimento de Palmilha Personalizada por Impressão 3D para uma Criança com Distrofia Muscular de Duchenne: Um Estudo de Caso Centrado no Usuário 1271

ID 551 · Desenvolvimento de uma Cadeira de Rodas Infantil Personalizada Utilizando Plásticos Reciclados: Uma Abordagem Baseada em Economia Circular e Manufatura 1277

ID 103 · Desenvolvimento de uma Instrumentação de Baixo Custo para Avaliação de Pressão na Escrita Manual 1283

ID 550 · Desenvolvimento de uma Tala Extensora Personalizada para o Uso Associado ao Parapodium em Crianças com Paralisia Cerebral 1287

ID 416 · Desenvolvimento de Órteses Suropodálicas Pediátricas por Fluxo Digital Integrado Utilizando Escaneamento 3D, Modelagem CAD e Manufatura Aditiva 1293

ID 833 · Design and Specification of a Local Learning Environment for Neuro-Robotic Educational Applications 1300

ID 659 · Development and Application of a Wearable System for Hemodynamic Monitoring During Physiotherapeutic Exercises in Post-COVID-19 Patients 1307

ID 750 · Development and Preliminary Evaluation of a Bidirectional Telerehabilitation Platform for Remote Treadmill Control and Heart Rate Monitoring 1312

ID 682 · Development and Preliminary Evaluation of a Wearable Inertial Sensor-Based System for Postural Monitoring and Vibrotactile Biofeedback 1320

ID 631 · Development of a Programmable Visual Cueing System to Support Gait Rehabilitation in Patients with Parkinson's Disease 1328

ID 67 · Development of a Sensorized Cycle Ergometer Integrated with a Serious Game for Post-Stroke Rehabilitation 1335

ID 737 · Development of a Wearable Arduino-Based Orthosis for Non-Invasive Temperature Monitoring in Diabetic Foot Prevention 1341

ID 744 · Development of an Ankle Exoskeleton to Improve Gait in Foot Drop Patients 1346

ID 900 · Development of PULSEGRID: An Interactive Floor Platform for Cognitive-Motor Activities Adaptable to Neurodivergent Individuals 1353

Sumário

VOLUME 01 · 09

Tecnologias Assistivas e Reabilitação

ID 872 · Digital Manufacturing of Transradial Prosthesis 1359

ID 618 · Dispositivos para Acessibilidade Digital 1365

ID 214 · Effects of Repetitive Transcranial Magnetic Stimulation Associated with Motor Imagery on the Electromyographic Activation of the Tibialis Anterior and Rectus Femoris Muscles in Post-Stroke Individuals 1371

ID 731 · Embedded AI-Based Handgrip System for Real-Time Dynamic Performance Evaluation 1377

ID 374 · Exergame for Children with ASD Learning to Cycle 1383

ID 739 · Gait Changes During Serious-Game Treadmill Training 1388

ID 906 · GESTUS: Assistive Interface for Augmentative and Alternative Communication and Gaze-Controlled Mouse Interaction — Proof of Concept and Exploratory Analysis 1395

ID 715 · Identifying Problems and Gaps in the ICU Early Mobilization Care Process: Development of Practice-Oriented Solutions in Rehabilitation Engineering 1403

ID 487 · Impact of Serious Games on Global Functionality and Therapeutic Engagement in Post-Stroke Individuals: A Systematic Review 1409

ID 669 · Inspiratory Muscle Training Using a Harmonica in the Postoperative Period of Coronary Artery Bypass Grafting 1420

ID 553 · Integration of Bluetooth Architecture for Caregiver Control of Low-Cost Motorized Pediatric Wheelchairs 1424

ID 27 · Investigando os desafios da mobilidade assistida de usuários de cadeiras de rodas para a elaboração de um protótipo multifuncional 1431

ID 522 · IoT Safety System for Residential Environments 1435

ID 685 · Low-Cost Wheelchair Adaptation for Rare Dysplasia 1440

ID 421 · Manufatura Aditiva Aplicada à Tecnologia Assistiva: O Desenvolvimento de um Protótipo de Encaixe Protético Transfemoral 1445

ID 236 · Markerless Gait Analysis, Edge Computing, and Virtual Reality for Remote Diagnosis in Telehealth: A State-of-the-Art Review of the Integration Gap 1450

ID 321 · Mental Workload in Motor Imagery BCI: Rapid Review 1458

ID 225 · Modular BCI for Robotic Rehabilitation Using EMG-IMU and SSVEP 1466

ID 581 · Motor-Imagery BCI Control of FES 1472

ID 96 · Multi-Platform 3D Skeletal Simulation Tool for Interactive Hip Posturing and Dislocation 1480

ID 464 · Multidimensional Barriers and Challenges in Daily Life Lower Limbs Prosthesis 1488

ID 710 · Noninvasive Peripheral Electrical Stimulation in Parkinson's Disease: A Pilot Feasibility Study with Blinded Video-Based Motor Assessment 1496

ID 586 · Octalysis Gamification of FES-VR Rehabilitation 1503

ID 769 · Open-Source Pediatric Wheelchair: Manufacturing and Cost 1510

ID 101 · Plataforma funcional de reabilitação, porém ausência de referência biomecânica externa limita a validação quantitativa do tracking. 1518

Sumário

VOLUME 01 · 10

Tecnologias Assistivas e Reabilitação

ID 77 · Preliminary Bench Validation of a Low-Cost Wearable Inertial Sensing Platform for Tremor Monitoring 1525

ID 397 · Projeto e Arquitetura de um Sistema Assistivo de Baixo Custo para Terapia Sonora do Zumbido 1531

ID 106 · Protocolo de revisão sistemática sem contribuição tecnológica ou experimental própria de Engenharia Biomédica. 1536

ID 436 · Realimentação Somatossensorial de Força para o Controle de Próteses Mioelétricas de Mão 1541

ID 836 · Recognition of Isolated Dynamic Libras Signs Using Formational-Parameter-Based Features 1547

ID 47 · Resampling-Based Power Simulation for Sample-Size Planning in Small Serious-Games Rehabilitation Trials 1555

ID 561 · sEMGest: An Integrated Platform for Hand Gesture Classification from sEMG with Real-Time Digital-Twin Validation 1560

ID 131 · Simulation Analysis of a Quasi-LPV Controller Applied to Hybrid FES Exoskeleton Rehabilitation Robotic Systems 1568

ID 121 · Sinais faciais embarcados em smartphones: reprodutibilidade entre SoCs e desempenho na detecção de nictação e oscitação 1575

ID 230 · Smartband with Embedded AI for Fall Detection 1582

ID 425 · Surface EMG in Rehabilitation: Methods and Challenges 1590

ID 279 · Tecnologias Móveis de Alinhamento, Controle Postural e Equilíbrio na Reabilitação Neurológica 1595

ID 86 · Telerehabilitation Application for Prescribing Home Exercises and Supporting the Daily Lives of Wheelchair Users 1603

ID 370 · Teste de Inteligência e Deficiência Visual: Uma Revisão de Literatura 1609

ID 245 · TMS and SSVEP in Post-Stroke Visual Field Loss 1616

ID 513 · VR with Haptic Glove for Hand Training 1621

ID 524 · Wearable Gait Monitoring in Aquatic Environment 1629

Informática Biomédica

ID 779 · A Mobile Application for the Six-Minute Walk Test with Oximetry Acquisition via Bluetooth Low Energy: A Pilot Study 1635

ID 373 · A Web Platform for Quality-of-Life Assessment in Sickle Cell Anemia with Automated SF-12 Scoring 1642

ID 23 · Age Composition of t-SNE Clusters in a Methylation Study of Women at Risk for Breast Cancer 1645

ID 764 · Alzheimer's Enhanced Diagnosis with Convolutional Neural Networks: An Efficient Approach 1652

ID 35 · An AI-Enabled Digital Healthcare Workflow Integrating Deep Learning and Large Language Models for Interpretable Wound Triage 1658

ID 72 · Aplicativo Móvel para Classificação de Eletrocardiogramas Utilizando Redes Neurais Convolucionais 1666

ID 548 · Apoio ao Rastreio do Câncer de Mama com Aprendizado de Máquina e Seleção de Atributos em Ultrassonografia Modo-B 1673

Sumário

VOLUME 01 · 11

Informática Biomédica

ID 479 · Architectural Patterns in Blockchain-Based Electronic Health Records: A Systematic Literature Review 1679

ID 775 · Arquitetura AlertaSUS: Sistema Hiperlocal de Alerta Epidemiológico Proativo com Preservação de Privacidade Baseada em Geohash 1687

ID 93 · Artificial Intelligence and Digital Technologies to Support Individuals with Autism Spectrum Disorder and Their Caregivers: A Systematic Review 1695

ID 75 · Artificial Intelligence-Based Ultrasound Simulation for Thyroid Nodules: State of the Art, Gaps, and Future Perspectives 1701

ID 65 · Assessment of the Effects of Falls on the Femoral Neck in Elderly People with Osteoporosis Using the Finite Element Method 1709

ID 645 · Benchmarking Language Agents on CGM Data 1714

ID 320 · Communication Protocols in Wearable IoT 1722

ID 138 · Comparative Genome and Resistance Profiling of Pseudomonas aeruginosa Across Diverse Niches 1729

ID 458 · Comparison of Random Forest, SVM, and MLP in Facial Emotion Recognition with PCA and ICA 1737

ID 274 · Computational Module for Nerve Conduction Signal Processing 1744

ID 366 · Construction of an Annotated Dataset of Oral Cavity Lesion Images for Neural Network Applications 1752

ID 378 · CORA: An Open-Source Desktop Application for Automated Wound Healing Quantification and Software Performance Benchmarking 1757

ID 863 · CORE-ABPM-FHIR 1765

ID 60 · Deep Learning-Based Reconstruction of Degraded Electromyographic Signals: A Comparative Study of Autoencoder and U-Net Architectures 1773

ID 795 · Detection and Segmentation of Nanoparticles in Electron Microscopy 1781

ID 249 · Development of a Digital Prototype for Continuing Education on Gadolinium-Based Contrast Agents in Magnetic Resonance Imaging: A Study Based on a National Survey 1786

ID 486 · Development of a Mobile Application with NFC Wristbands for Secure Patient Identification in Hospital Settings 1792

ID 292 · Development of Software for Notification of Critical Laboratory Results in a Hospital Setting 1800

ID 582 · Documentary Characterization of Brazilian Mammography Databases 1804

ID 273 · EfficientNet and ML for Mammographic Breast Cancer Classification 1811

ID 387 · Estudo Comparativo da Utilização da Estratificação Clínica para Predição do Tempo de Internação Hospitalar 1818

ID 783 · Evaluation of a Phase-Delay Analytical Method for Non-Invasive Blood Glucose Estimation Using Dual-Wavelength Photoplethysmography Signals 1825

ID 535 · From Pix to Care Rails: A Common Architectural Core for Prescriptions, Lab Orders, Reports, and Dispensing 1833

ID 54 · Gemma 4 as a Proof-of-Concept Model for Converting Discharge Medication Text into FHIR MedicationRequest Resources 1841

ID 182 · H-Q-MAS: Deterministic-First Anonymization of Brazilian Clinical Text under the LGPD 1849

Sumário

VOLUME 01 · 12

Informática Biomédica

ID 327 · Hypoglycemia Prediction: Accuracy, False Alarms and Cost **1856**

ID 419 · IA Bucal Files: Development of a Platform for Curating and Annotating Clinical Images of Oral Lesions **1864**

ID 259 · Impact of Feature Selection on Hybrid Deep Learning Models for Histopathological Breast Cancer Classification **1869**

ID 134 · Interpretable Structural Equation Modeling and Machine Learning Framework for Sepsis Risk Assessment **1877**

ID 517 · Ki67 Marker-Based Proliferation Index Estimation Using Deep Learning **1885**

ID 919 · Label Propagation for Missing Data Imputation (LPMD) in Real-Life COVID-19 Tabular Data **1891**

ID 644 · Limits of Lightweight Hallucination Detection in Medical LLMs: A Lexical Bias Analysis of MedHal Across Clinical Specialties **1899**

ID 154 · Machine Learning for the Reconstruction of Complex Functional Outcomes **1906**

ID 657 · Machine Learning in Healthcare: Development of Dementia Prediction Models **1912**

ID 401 · Mamo.IA: A Web-Based Artificial Intelligence Platform for Breast Cancer Screening and Diagnostic Support Using Thermography, Mammography, and Histopathology Images **1916**

ID 897 · MEADS-Health: An Auditable Methodology for Exploring Sensitive Institutional Data with AI-Assisted Analytical Planning **1924**

ID 871 · Modern Hopfield Networks for Associative Memory-Based Emergency Triage Prediction **1932**

ID 232 · Multiclass Breast Lesion Classification via Deep Wavelets and Random Tree-Based Feature Transformation and Classification **1940**

ID 433 · Neonate Image Segmentation in NICU **1948**

ID 776 · Non-Classical Logics for Diabetes Decision Support **1955**

ID 356 · Oral Lesion Annotation Tool for Neural Network Training **1963**

ID 901 · ORBITA-Health: Auditable Curation of Health and Disability Indicators for Public Observatories **1968**

ID 601 · PMPIF – Dynamic Report: An Intersectoral Web Platform for Monitoring Early Childhood Indicators and Supporting Public Health Management **1975**

ID 123 · Prototyping a RAG-based Conversational Agent for Pre- and Post-Operative Cardiac Patient Guidance **1983**

ID 73 · Quantificação Automatizada de Infiltrado Neutrofílico em Tecido de Ligamento Periodontal Aplicando Segmentação por Superpixels e Aprendizado de Máquina **1990**

ID 478 · Risk Factors for Severe RSV Outcomes in Children: A Systematized Review and Meta-Analysis **1996**

ID 293 · Sistema Especialista CLIPS para NR-17 **2003**

ID 825 · Spectral Preprocessing for Machine Learning Classification of Salivary FTIR Spectra in Oral Squamous Cell Carcinoma **2007**

ID 587 · Telemedicine Integrated with Hospital Information System **2015**

ID 242 · VIZU: A Mobile Hub Proposal for Computer Vision Models Applied to Healthcare **2018**

Sumário

VOLUME 01 · 13

Biosensores, Transdutores e Instrumentação Biomédica

ID 434 · A Monitoring System Prototype for Vascular Neck Restraint Applied by Law Enforcement Officers 2026

ID 135 · A Multimodal Prototype for Portable Quantitative Monitoring of Epilepsy 2034

ID 331 · A Polyacrylamide-GOx Hydrogel Biosensor on Laser-Engraved Graphene: Structural and Electrochemical Evidence for Glucose Detection 2040

ID 485 · A Review of Embedded Systems Applied to Oxygenation Monitoring via Photoacoustic Imaging 2047

ID 268 · Analisador de Contração Muscular por EMG Facial: Desenvolvimento e Avaliação Preliminar de um Protótipo 2055

ID 552 · Analog Demodulation for Electrical Impedance Tomography Using Rectification and Zero-Crossing Detection 2061

ID 660 · Analysis of Physiological Responses in the Identification of Misinformation: A Scoping Review 2069

ID 928 · Análise Comparativa das Curvas Fluxo-Tempo entre um Sensor Mecânico de Turbina e um Sensor Piezoresistivo 2076

ID 620 · Aplicação dos Sensores MMWave no Monitoramento de Sinais Vitais e Detecção de Quedas 2080

ID 791 · Arranjo Experimental Utilizando Plano-Terra para Ensaios com Sensores Biotelemétricos Indutivos 2088

ID 521 · BioDiff-CMOS: Pre-Silicon Design and Physical Verification of a CMOS Differential Block for Simulated ECG Acquisition 2092

ID 770 · Biomagnetic Assessment of GI Motility 2098

ID 233 · Biosusceptometria AC em Nanopartículas de Óxido de Ferro 2102

ID 652 · Caracterização Óptica de Membranas de Nitrocelulose Funcionalizadas com Nanopartículas de Ferrita de Cobalto por Espectroscopia NIR e PCA 2107

ID 887 · Clinical Validation of VMAT Commissioning with 6FFF Beams Using AAPM TG-119 2112

ID 747 · Desenvolvimento de uma Plataforma IoMT para Monitoramento de Sinais Vitais e Detecção de Quedas 2118

ID 658 · Desenvolvimento e Validação de Sistema Inteligente Pneumático-Eletrônico para Avaliação Cognitivo-Motora em Esportes de Combate 2126

ID 452 · Design and Characterization of a Low-Cost Electroencephalography Analog Front-End 2131

ID 353 · Design of a Functional Electrical Stimulation Device with Hybrid Converter in the Output Stage 2137

ID 21 · Development of a Digital Algometer for Application in MICS 2142

ID 806 · Development of a Dynamic Electro-Pneumatic Actuator for MRE Applications 2148

ID 498 · Development of a Low-Cost Apparatus with Monitoring Sensors for the Controlled Exposure of Rats to Cigarette Smoke 2156

ID 492 · Development of a Low-Cost Graphite-Based Bioelectrode for Electromyographic Signal Acquisition and Stimulation 2163

ID 716 · Development of a Low-Cost Portable Multiparameter Monitor Based on ESP32 for Real-Time Vital Signals Monitoring 2170

ID 36 · Development of a Personalized Biosensor-Based Rehabilitation System for Ankle Sprain with Real-Time Corrective Feedback 2175

Sumário

VOLUME 01 · 14

Biosensores, Transdutores e Instrumentação Biomédica

ID 520 · Development of a Six-Lead Portable ECG System for Home Health Care and Emergency Alerts 2181

ID 519 · Development of a Wearable Device Prototype for Parkinsonian Resting Tremor Monitoring 2186

ID 316 · Development of a Wireless Wearable System for Temporal and Angular Human Gait Analysis Based on the ESP-NOW Protocol 2194

ID 504 · Development of an Electrical Impedance Tomography System Using an ESP32-Based Data Acquisition Architecture 2202

ID 15 · FBG-Based Optical Patch Sensor for Differential Thermal Detection of Skin Cancer 2210

ID 413 · Gas Vesicles as Ultrasound Contrast Agents 2217

ID 515 · Hall-Effect-Based Multichannel Sensing Architecture for Quantitative Pelvic Floor Biomechanics: Development and Bench Validation 2222

ID 301 · Hardware for Wearable EEG Systems: A Structured Review on Acquisition, Preprocessing, and Embedded Classification 2226

ID 177 · Instrumental Improvements of X-QST for Diabetic Neuropathy 2234

ID 694 · Low-Cost EMG Acquisition and Classification 2242

ID 447 · Mapping Conventional and Emerging Sensors for Neonatal Physiological Monitoring: A Scoping Review of the Technological Transition and Validation Gaps 2248

ID 44 · Mapping LIG-Based Biomedical Sensing Research: Bibliometric Trends, Thematic Clusters, and International Collaboration 2256

ID 176 · Mechatronic Platform for Radar Respiratory Monitoring 2264

ID 305 · Medição Não Invasiva de Pressão Arterial por Ultrassom em Modo A-Scan 2272

ID 738 · Modelagem de Sistema para Análise do Esforço Respiratório Natural via Sensor de Pressão Piezoresistivo 2276

ID 567 · Modelling of a Strain Gauge-Based Transducer: Neonatal Respiratory Flow Sensor 2280

ID 726 · Onboard Gait-Based Anatomical Axis Identification for Stride Detection with a Low-Cost Ankle Wearable 2288

ID 494 · Performance Tests of Silicone- and Graphite-Based Surface Electrodes Developed for Electromyography and Electrical Stimulation Equipment 2296

ID 335 · Physiological Measurement Validation in Flight Simulator 2304

ID 407 · Piezoelectric Acquisition of Simulated Bruxism Signals 2311

ID 677 · Scoping Review of Non-Contact Capacitive Sensing for Human Movement Monitoring in Healthcare 2318

ID 670 · Termometria 3D por Ultrassom 2326

ID 294 · Tetrapolar Electrode Geometries for Skin Phantoms 2330

ID 784 · Thermal Heating Evaluation of Orthopedic Plates Implanted in Ultrasonic Phantoms under Therapeutic Ultrasound 2338

ID 571 · Universal Photobiomodulation Device for Dairy Cows 2343

Sumário

VOLUME 01 · 15

Biosensores, Transdutores e Instrumentação Biomédica

ID 358 · Wireless Wearable Neonatal Monitoring System 2348

Avaliação de Tecnologias em Saúde

ID 20 · A Computational Tool for Technical Evaluation of Endoscopes 2356

ID 655 · A Multi-Criteria Decision Support Framework for Clinical Sensor Selection 2361

ID 40 · A Multimodal Framework for Hansen's Disease Monitoring Using Lesion Detection, Segmentation, Quantitative Feature Extraction and LLM-Based Clinical Reasoning 2369

ID 699 · A Novel 3D Scanning Workflow for Precise Angular Measurement in Femoral Derotational Osteotomies 2376

ID 847 · An Interactive LLM-Based Approach to Exploratory Data Analysis Evaluated on Open Health Datasets 2381

ID 38 · Analysis of Orofacial Thermal Biomarkers at Different Acquisition Distances: A Comparative Approach Between Human Analysis and Conventional Artificial Intelligence 2389

ID 460 · Análise Comparativa de Doses e Parâmetros de Aquisição para Diferentes Tecnologias de Mamografia Digital 2395

ID 638 · Análise de Risco de Diabetes: Um Aplicativo Móvel com Aprendizado de Máquina Explicável para Educação em Saúde 2401

ID 114 · Aplicação interessante do modelo eRIC à escoliose, porém inovação essencialmente de aplicação clínica. 2409

ID 470 · Artificial Intelligence Applied to Postural and Functional Ergonomic Analysis of Office Workers 2417

ID 673 · As Fragilidades do Sistema HÓRUS na Gestão de Farmácias no SUS 2423

ID 929 · Avaliação Comparativa de Modelos de Aprendizado de Máquina para Classificação da Doença de Alzheimer a partir de Sinais de EEG 2429

ID 956 · Avaliação Sistemática de Simuladores Realísticos na Área da Saúde: Desenvolvimento e Análise de um Protótipo de Baixo Custo para Reprodução de Sons de Ausculta 2433

ID 624 · Balanceamento na Predição de Diabetes 2439

ID 110 · Changes in Pulmonary Mechanics of Endemic Disease Control Agents Evaluated by Respiratory Oscillometry 2446

ID 924 · Cold Plasma in Medicine: A Review of Clinical Applications and Underlying Biological Mechanisms 2453

ID 531 · Comparison Study of Three Different Howland Current Pump Architectures for a Low-Cost EITduino Device 2455

ID 255 · Dispositivo Portátil Multimodal para Aquisição de Sinais Biomecânicos e Bioelétricos Aplicados ao Monitoramento da Doença de Parkinson 2463

ID 675 · E-SUS APS as Health System Engineering 2471

ID 377 · Embedded Neonatal Seizure Detection Framework 2478

ID 444 · Enhancing Situational Awareness for the Deaf and Hard-of-Hearing: A Systematic Review of Haptic and Bimodal Technologies 2486

ID 803 · Experimental Approach to Optimizing Magnetic Resonance Imaging Protocols with Reduced Specific Absorption Rate 2494

ID 819 · Kinovea 2D Analysis: Linear Gait Measurements 2502

Sumário

VOLUME 01 · 16

Avaliação de Tecnologias em Saúde

ID 126 · Limitations of RAG Architecture in Compact Medical LLMs 2507

ID 11 · Mathematical Model to Prioritize Preventive Maintenance of Endoscopes 2513

ID 85 · Monitoring of Orofacial Thermal Patterns after Physiotherapy Intervention in Sequelae of Mastoid Fracture Using Infrared Thermography: A Case Study 2520

ID 115 · Novos índices derivados de loops respiratórios são promissores, mas necessitam validação externa para sustentar uso diagnóstico. 2524

ID 712 · Performance of Multiple PPG-Based Physiological Measurements Across Different Skin Phototypes 2532

ID 39 · Physiological Evaluation of Heat Stress Associated with Mission-Oriented Protective Posture (MOPP) in Brazilian Army Military Personnel 2539

ID 940 · Preliminary Dosimetric Characterization of an RS 3400 X-Ray Blood Irradiator Using LiF:Mg,Ti Thermoluminescent Dosimeters 2546

ID 671 · Regulação Assistencial Inteligente no SUS 2553

ID 909 · Rocket Motion: Desenvolvimento e Validação de um jogo sério para Reabilitação Motora em Crianças Hospitalizadas 2559

ID 9 · Sistema web para gerenciamento de consultas pré-natais de baixo risco 2565

ID 941 · Solar Water Pasteurization for Humanitarian Emergencies 2570

ID 113 · Survival and Independent Predictors of Mortality in Patients Undergoing Coronary Artery Bypass Grafting With and Without Cardiopulmonary Bypass 2576

ID 409 · Technical Evaluation Methodology for Medical Devices in Procurement Processes: Case Study with a Portable Retinograph in the Brazilian Unified Health System (SUS) 2582

ID 701 · Technological Analysis of Patient Positioning and Tracking Systems in Radiotherapy and Medical Imaging Equipment 2590

ID 724 · Technological Mapping of Minimally Invasive Devices for Blood Collection: A Systematic Review from the Perspective of Biomedical Engineering and Product Engineering 2598

ID 850 · Telehealth Equipment Infrastructure in Brazil: Regional Distribution, Administrative Availability, Population-Adjusted Inequalities, and Connectivity 2604

ID 611 · Terapia Fototérmica Assistida por Nanopartículas Plasmônicas na Região do Infravermelho: Investigação Numérica e Experimental 2611

ID 428 · VitaCheck- Portable Triage Case for Indigenous Communities 2615

Neuroengenharia

ID 948 · A Collaborative Brain-to-Brain Interface for θ/β Modulation During Sustained Attention Training of Two Individuals 2622

ID 188 · An Integrated Open-Source Platform for Robot-Assisted Personalized Transcranial Magnetic Stimulation 2629

ID 877 · An Integrated Optogenetic and Operant Conditioning System for Instrumental Memory Research in Rats 2637

Sumário

VOLUME 01 · 17

Neuroengenharia

ID 143 · BCI-FES Integration for Motor Rehabilitation: A Scoping Review 2642

ID 672 · Comparison of Oxidative Stress-Induced Damage in Neuronal and Neuro-Astrocytic Models of Alzheimer's Disease 2650

ID 89 · Deep Learning for Prediction of Awake Slow-Wave Occurrences in EEG 2655

ID 895 · Deep Learning-Based U-Net for Urine Volume Estimation in Rats with Neurogenic Bladder Using Freehand Ultrasound Images 2660

ID 633 · Dense Recordings in the Posterior Parietal Cortex 2666

ID 881 · Desenvolvimento e Padronização de um Procedimento Operacional Padrão para Cirurgia Estereotáxica de Infusão Viral e Implante de Cânula em Ratos Wistar 2674

ID 771 · EEG-Based Classification Model for Predicting tDCS Responders in the Treatment of Neuropathic Pain 2680

ID 813 · EEG-Based Robotic Hand Control During Motor Execution and Motor Imagery 2685

ID 733 · Efeito do Exercício Físico sobre a Preservação Neuronal em Fatias de Hipocampo de Rato Submetidas a Status Epilepticus In Vitro 2693

ID 195 · Efeitos da estimulação vibrotátil na regulação de forças máximas e submáximas sem realimentação visual 2699

ID 140 · Effects of Transcranial Photobiomodulation on Cognitive Function Following Ischemic Stroke: A Randomized Clinical Trial Assessed by MMSE 2704

ID 857 · Engineering Principles of Machine Learning-Driven Adaptive Deep Brain Stimulation: A Narrative Review 2709

ID 575 · Estimulação Transcutânea Auricular do Nervo Vago por Corrente Senoidal de 2 kHz e 100 Hz: Estudo Exploratório 2715

ID 619 · Evaluation of TENS-Evoked Perception for the Restoration of Somatosensory Feedback: A Case Study 2723

ID 107 · Experimento óptico relevante, porém tamanho amostral, modelagem da propagação óptica e separação dos fatores idade/lesão ainda limitam a evidência. 2729

ID 599 · gCKC, FastICA and Swarm-Contrastive HD-sEMG 2736

ID 853 · Hemispheric Asymmetry and Sex-Related Differences in Gray Matter Morphometry in DBS Candidates with Parkinson’s Disease 2743

ID 588 · Hybrid EEG-EMG Fusion Comparison for Executed Gesture Dataset Classification 2748

ID 879 · Intelligent System for Cervical Biomechanical Tracking in Speech-Language Rehabilitation of Parkinson’s Disease 2756

ID 388 · Investigation of Theta-Gamma Co-modulation during Sequential Learning as a Biomarker of Psychopathological Dimensions in the HBN Cohort 2763

ID 790 · Low-Density Electroencephalographic Riemannian Trajectories as Candidate Learning Markers in Motor Imagery Brain–Computer Interfaces 2770

ID 650 · Microdrive Construction and Implantation for Chronic Extracellular Recordings Targeting the Dorsal Hippocampus in Freely Moving Rats 2778

Sumário

VOLUME 01 · 18

Neuroengenharia

ID 476 · Método de Preparação de Fatias Hipocampais para Avaliação Integrada de Eletrofisiologia e Imunofluorescência 2783

ID 141 · Neuromodulatory Effects of Transcranial Photobiomodulation on Parkinsonian Tremor: Clinical and Functional Outcomes in a Case Report 2791

ID 621 · Photobiomodulation Potential in a Neuroglial Co-Culture Model Exposed to the SARS-CoV-2 Spike Protein 2796

ID 809 · Proposal for Robotic Gait Training with a Motor Imagery-Based Brain–Computer Interface and Transcutaneous Spinal Cord Stimulation 2803

ID 689 · Reaction Time During a Manual Dexterity Task in Parkinson's Disease: A Pilot Study 2810

ID 823 · Ressonância Magnética na Investigação de Biomarcadores Associados à Identificação Antecipada do Transtorno do Espectro Autista: Uma Revisão de Escopo 2819

ID 306 · Temporal Processing and Decoder Architecture in Pseudo-Online BCIs 2826

ID 837 · Time-Frequency Characterization of the cVEMP Signal Using Wavelet Transform in Participants with Vestibular Complaints 2834

Engenharia Clínica e Engenharia Hospitalar

ID 298 · A Non-Intrusive IoMT Prototype for Duty-Cycle Monitoring of a Laboratory Air Compressor 2841

ID 148 · Accessibility in Medical Devices: Perceptions and Strategies in Clinical Engineering 2847

ID 185 · AI-Based Fall and Movement Detection for Hospital Beds 2853

ID 852 · Análise do Impacto da Pandemia de COVID-19 no Consumo Hídrico de um Hospital de Referência em Pernambuco (2020–2024) 2860

ID 681 · Avaliação de Pontos de Contaminação em Cabines de Segurança Biológica Classe II 2868

ID 291 · Challenges in the Compliance of Dental Services with ANVISA Resolution RDC No. 1,002/2025: Clinical Engineering Experience in a Brazilian Oncology Hospital 2873

ID 166 · Clinical Engineering in Brazilian Hospitals: Practice and Challenges — Is There Opportunity for a Conceptual Redefinition of the Field? 2878

ID 393 · Conformidade Técnica e Normativa em Contratos de Gestão de Resíduos Hospitalares na Rede Pública Estadual do Tocantins 2886

ID 451 · Data Quality in a Computerized Hospital Maintenance Management System within a Multicenter Network 2890

ID 251 · Design and Deployment of Interactive Standard Operating Procedures for CMMS Customization: A Case Study at a University Hospital 2897

ID 92 · Development of an Assessment Tool for Health Technology Management in Home Care 2903

ID 406 · Development of an Open-Source Software for Automated Accuracy Testing of Infusion Pumps 2911

ID 46 · Distribuição tecnológica e cobertura assistencial regional dos serviços de diagnóstico por imagem em Araguaína-TO 2918

ID 332 · Energy Consumption Monitoring in Public Health Facilities 2926

Sumário

VOLUME 01 · 19

Engenharia Clínica e Engenharia Hospitalar

ID 867 · Estimated Annual Maintenance Cost of Life-Support Medical Equipment Registered in the Brazilian National Registry of Health Facilities 2932

ID 484 · Estimation of Equivalent Dose to the Eye Lens during Fluoroscopic Procedures 2936

ID 431 · Evaluation of Interoperability in Digital Dental Prosthetics Workflow Ecosystems within the SUS: Proof of Concept Applied to Public Procurement 2941

ID 90 · Importance of Metrological Performance Testing for New and Post-Corrective Maintenance Medical Equipment 2950

ID 348 · Integrating Performance Indicators and Quality Tools to Support Decision-Making in Clinical Engineering 2955

ID 719 · Inventário Estruturado Desenvolvido em um Hospital do Agreste Alagoano Visando à Implantação da Engenharia Clínica Institucional 2960

ID 25 · Level of Knowledge and Application of ABNT NBR IEC 62353:2019 Among Clinical Engineering Professionals in Brazil 2967

ID 752 · Microbiological Contamination from Syringe Plunger Handling 2972

ID 717 · Mitigando e Caracterizando Relativamente EMI Irradiadas em Salas Cirúrgicas de Alta Complexidade 2977

ID 536 · Multiple Analytical Perspectives on Hospital Maintenance Data - Reliability, Seasonality, and Support through Language Models 2985

ID 391 · Optimization of Hospital Electric Bill 2993

ID 329 · Post-Market Performance Evaluation of an Infusion Pump in a Pediatric Intensive Care Unit: A Case Study 2999

ID 324 · Potential Capacity and Territorial Distribution of Mammography Equipment in the State of São Paulo 3006

ID 501 · QUALIRAD: Progressive Web Application for Quality Control of Diagnostic Imaging Equipment in Accordance with RDC 611/2022 3013

ID 159 · Risk-Based Maintenance Prioritization for Medical Equipment 3020

ID 284 · Scientific Foundations of Immersive Humanization in Diagnostic Imaging 3028

ID 283 · Sistemas de Detecção de Materiais Ferromagnéticos em Ressonância Magnética: Tecnologias, Segurança e Sustentabilidade Operacional 3036

ID 193 · Technical Requirements for Heat-Sealing Equipment in Medical Device Processing Units According to Current Standards 3044

Modelagem e Simulação Computacional

ID 12 · A Study of Statistical Noise in Cleft Lip and Palate Classification Models 3052

ID 875 · Analog Circuit for Cardiac Cycle Simulation 3058

ID 330 · Analysis of Dengue Viral Dynamics under the Impact of Antiviral Treatment During Pregnancy 3066

ID 746 · ANN for Cognitive Performance in Cerebral Palsy 3072

ID 171 · Avaliação de Estratégias de Controle Preditivo e por Alocação de Polos na Flexão do Joelho Induzida por FES em Ambiente OpenSim 3080

Sumário

VOLUME 01 · 20

Modelagem e Simulação Computacional

ID 237 · Cell Growth Curves with Structured Segregated Model in Unreal Engine 3086

ID 314 · Comparative Analysis of Antiviral Strategies Targeting Viral Load and Immune Response in Dengue Infection under a Sex-Based Perspective 3094

ID 443 · Comparison of Computed Muscle Activations and sEMG on Lower Limb Muscle SSC 3099

ID 457 · Development of an Interactive Web Simulator for Teaching Computational Electrophysiology 3107

ID 627 · Digital Twins for Langendorff 3114

ID 914 · Does Phrasing Matter? Prompt Sensitivity of Portable Large Language Models in Text-Based Sentiment Evaluation 3121

ID 160 · Estimation of Pulmonary Arterial Tree Characteristics in PH 3128

ID 585 · Feasibility Study of Simplified Computational Models for Gait Simulation 3136

ID 252 · Framework for 3D Urban Digital Twin - Reconstruction and Simulations for Disaster Management 3142

ID 922 · Genetic Albatross 3150

ID 462 · Geometric Digit Model for Veterinary Electrochemotherapy 3157

ID 473 · HCN Channels and Inhibitory Postsynaptic Current 3162

ID 400 · Implementação de um Controlador de Barreira para Aumento da Segurança em um Dispositivo Andador 3169

ID 270 · Medio-Lateral Ground Reaction Force Analysis During Gait Using Discrete Wavelet Transform in Individuals with Knee Dysfunction 3177

ID 95 · Modelo computacional interessante de exergia cardíaca, mas validação predominantemente baseada em parâmetros da literatura. 3184

ID 112 · Monte Carlo Evaluation of Z-Axis Tube Current Modulation for CT Dose Optimization 3189

ID 469 · Na,K-ATPase Kinetics and Non-Synaptic Epileptiform Activity 3194

ID 829 · O Efeito de Personas Geracionais nas Avaliações de Valência de Grandes Modelos de Linguagem Portáteis 3201

ID 896 · Persona Modulation in Portable Language Model 3207

ID 848 · Physical Application of the Hodgkin-Huxley Model in the Description of the Membrane Action Potential 3213

ID 532 · Physiological Validation of Synthetic Ventilatory Data 3221

ID 257 · Rare-Earth Doped Bioactive Glasses for Bone Regeneration 3229

ID 735 · Simulation-Based Geometry Optimization of a Bone-on-a-Chip Through a Machine Learning-Assisted Evolutionary Approach 3236

ID 643 · Simulação Computacional da Dinâmica de Micro e Nanopartículas de Óxido de Cálcio em Água e Humor Vítreo sob Forças Acústicas de Ultrassom 3244

ID 417 · Three-Dimensional Simulation of Respiratory Airflow Based on Computed Tomography in Tracheobronchopathia Osteochondroplastica 3252

ID 912 · Upper Limb Kinematic Analysis Using Inertial Sensors and Temporal Normalization 3258

Sumário

VOLUME 01 · 21

Modelagem e Simulação Computacional

ID 84 · Variational Quantum Classifier for Multiclass Parkinson's Disease Staging on the PPMI Dataset 3267

Sistemas de Diagnóstico e Terapia Clínica

ID 962 · 5%/1 mm Gamma Criterion for VMAT SRS QA 3275

ID 583 · Análise Comparativa entre o Tremor de Repouso no Dedo e a MMG no Antebraço da Doença de Parkinson 3281

ID 757 · ATR-FTIR Salivary Spectroscopy Under Gender-Affirming Hormone Therapy: A Comparative Two-Subject Case Study 3285

ID 256 · Avaliação da Inteligência de Adultos: Protótipo de Jogo Computadorizado Baseado na Escala WAIS-III 3289

ID 684 · Class-wise Evaluation Reveals Hidden Errors Behind High Global Performance in Deep Learning Classification of Imbalanced OCT Images 3298

ID 31 · Classification of EEG Signals for Differential Diagnosis of Alzheimer's Disease Using Machine Learning 3306

ID 386 · Comparison of Two Hypoxic Burden Metrics in Polysomnography of Obstructive Sleep Apnea Patients 3312

ID 338 · Compilação Semântica em Redes Bayesianas: Explorando LLMs para Inferência MPE sem Dados 3318

ID 481 · Data Generation for Autism Spectrum Disorder in Children Using Expert Knowledge and Synthetic Data Techniques 3326

ID 125 · Diagnostic Accuracy of Prostate-Specific Antigen as an Indicator of Male Sexual Contact in Living Victims of Sexual Violence 3334

ID 814 · Effects of Transcutaneous Spinal Direct Current Stimulation (tsDCS) in Patients with Major Depressive Disorder: A Case Comparison 3339

ID 82 · Elaboração de Procedimentos Operacionais Padronizados para Controle de Qualidade em PET-CT 3344

ID 144 · Estabilidade de Cloretos para o Teste do Suor 3349

ID 828 · Heating Analysis of a Multilayer Phantom Skin-Tendon-Bone Irradiated by Therapeutic Ultrasound 3355

ID 80 · Investigation of Protoporphyrin IX Dimethyl Ester Aggregation by Nuclear Magnetic Resonance and Its Impact on the Development of Photosensitizers for Photodynamic Therapy 3360

ID 173 · Monitor de Prevenção de Pneumonia Adquirida por Ventilação Mecânica 3365

ID 695 · NeoTrack APLV: A Mobile System for Longitudinal Symptom Monitoring in Infants with Suspected Cow's Milk Protein Allergy 3373

ID 787 · NeuroSUS-RN: A Scalable Neurological Monitoring Platform to Expand Access to Neonatal Care in Brazil's Universal Health System (SUS) 3378

ID 690 · Open Ultrasound Platforms for Research 3383

ID 122 · Parâmetros técnicos e eficácia clínica da laserterapia 3391

ID 58 · Photobiostimulation After Acute Myocardial Infarction: A One-Year Follow-Up Case Report 3399

ID 714 · Proof of Concept of Hand Tracking Using an Omnidirectional Vision Sensor in a Portable Device for Rehabilitation 3403

Sumário

VOLUME 01 · 22

Sistemas de Diagnóstico e Terapia Clínica

ID 707 · Raman Pharmacovigilance of Viscum album in Snakes 3411

ID 228 · Raman Spectroscopy for Serological Titers in Canine Leishmaniasis 3418

ID 200 · Raman Spectroscopy of Healthy, Diabetic/Hypertensive and Hemodialysis Serum 3422

ID 395 · SAD-DERM: A Decision Support System for Automatic Diagnosis of Dermatological Lesions Using Convolutional Neural Networks 3426

ID 223 · Simultaneous Quantification of Macronutrients in Commercial Milk Powders Using Raman Spectroscopy and Partial Least Squares Regression 3434

ID 300 · Spectral Discrimination of Antimicrobial Resistance Profiles in Clinical Strains of E. coli and K. pneumoniae via Raman Spectroscopy 3437

ID 336 · Thermographic Analysis of Unstageable Pressure Injuries 3443

Biomateriais, Engenharia Tecidual e Órgãos Artificiais

ID 880 · Amniotic Membrane Processed with Ozonated Water 3449

ID 220 · Analysis by Electron Microscopy of Morphological Changes in Calcium Hydroxyapatite Crystals Submitted to Microfocused Ultrasound In Vitro 3453

ID 821 · Analysis of the Cellular Response to Ti-29Nb-13Ta-4.6Zr Alloy Using Vero Cells for Bone Tissue Engineering 3461

ID 873 · Avaliação da Membrana Amniótica Humana por Espectroscopia Raman após Diferentes Períodos de Armazenamento em Glicerol: Um Estudo Preliminar 3466

ID 573 · Bioink Rheology for Bone Tissue Engineering 3470

ID 921 · Characterization of a Gelatin Whitlockite Ink for 3D Printing of Scaffolds for Osteochondral Regeneration 3477

ID 351 · Curcumin and Ion Release Nanoparticles for Neural Repair 3485

ID 275 · Development and Characterization of Green Strontium Oxide Nanoparticles for Potential Application in Bone Regeneration 3493

ID 334 · Droplet Generation in Flow-Focusing Microfluidic Devices Fabricated via MSLA 3D Printing 3500

ID 197 · FTIR Characterization of Gutta-Percha after Ozonated Water Disinfection 3508

ID 818 · Gelatin Photo-Crosslinking with Methylene Blue and Red Laser: A Comparative Pilot Study 3514

ID 313 · Hanging Drop Microdroplet Culture: A Promising Approach for Biomaterial Testing 3518

ID 665 · HAp Addition in Gelatin-Carrageenan Bioinks 3522

ID 352 · Injectable Thermoresponsive Hydrogel 3530

ID 546 · Low-Cost Food 3D Printing Platform 3535

ID 668 · Mechanical Behavior of Biosilicate® Scaffolds: Experimental and Computational Analysis 3542

ID 510 · Plasma Sterilization of 3D-Printed PCL/Bioglass 3548

ID 526 · Poly(Lactic Acid) Films as Scaffolds for Cell Culture: Hydrolytic Degradation and In Vitro Biocompatibility 3553

Sumário

VOLUME 01 · 23

Biomateriais, Engenharia Tecidual e Órgãos Artificiais

ID 48 · Poly(vinyl alcohol)-based hydrogels engineered for sustained non-viral gene delivery: physicochemical characterization and in vitro transfection assessment 3558

ID 116 · Rheological Assessment of Enzymatic Degradation of Rennova Ultra Volume Lido Facial Filler by Biometil Hyaluronidase at Different UTR Concentrations: An In Vitro Study 3564

ID 845 · Rheological Properties of Chitosan Hydrogels Containing Zinc-Doped Whitlockite 3571

ID 102 · Scientometric Analysis of Nanoparticle-Based Drug Delivery Systems for Breast Cancer Treatment 3577

ID 547 · Study of the Gelation Process of an Injectable Hydrogel Composed of Natural Materials with Regenerative and Antibacterial Properties 3585

ID 489 · Synthesis and Tuning of Highly Methacrylated Hyaluronic Acid Hydrogels for Light- and Extrusion-Based 3D Printing 3592

ID 778 · Theranostic Bioceramics Translational Framework 3598

ID 959 · Thermal Stability of Orthodontic Aligners 3606

ID 380 · Ultrasound-Guided 3D Dosimetry Using an Optimized MAGIC-f Phantom 3611

Bioengenharia

ID 698 · A Real-Time Evaluation of High-Density EMG Signal Quality 3617

ID 904 · Análise das Alterações Espectrais NIRS em Comprimidos Submetidos a Diferentes Condições de Umidade em Cenário de Atenção Domiciliar 3623

ID 79 · Automated Obstructive Sleep Apnea Detection Using Reduced Polysomnography Channels 3628

ID 191 · Automation of Langendorff Experimental Setup 3636

ID 709 · Benzoylecgonine/THC-COOH in Raw Sewage 3644

ID 30 · Caracterização Fractal da Laminina Polimerizada: Implicações para Sinalização Celular, Regeneração Neural e Neuroengenharia 3647

ID 99 · Characterization of 3D Printer Filaments for CT Anthropomorphic Pediatric Phantom Development 3655

ID 299 · Deep Feature-Based Classification of Colorectal Histopathology: Effects of Class Balancing and Particle Swarm Feature Selection 3662

ID 70 · Development of a Sensorized Device for Respiratory Rehabilitation Assessment 3670

ID 128 · Electrophysiological Mapping and Histological Remodeling in Ventricular Tachycardia 3676

ID 212 · Extended GHK Equation Including Ion Pumps 3682

ID 186 · Gear-Based Ratchet Mechanism for Ilizarov Fixators 3689

ID 632 · H-Wave Detection in His Bundle Electrograms Using an LSTM Recurrent Neural Network 3694

ID 297 · Hand Gesture Classification from HD-sEMG Images Using Deep Learning and Decision Trees 3701

ID 440 · High-Speed Data Transfer Architecture in Portable Handheld Ultrasound Device Using PCIe 3709

Sumário

VOLUME 01 · 24

Bioengenharia

ID 403 · Image-Based Evaluation System for Pilates: A Proof of Concept 3715

ID 883 · Influência do Índice de Massa Corporal na Impedância da Pele sob Corrente Alternada para Aplicações de Iontoforese 3722

ID 616 · Kinematic and Multi-Feature Electromyographic Characterization of Functional Upper-Limb Tasks 3728

ID 449 · Portable Electrospinning for Wound Care 3736

ID 69 · Proposal of an Automated System for the Langendorff Isolated Heart Perfusion Method 3744

ID 801 · Sound Speed Characterization of Plastisol Specimens for Ultrasound Phantom Fabrication 3751

Robótica Biomédica e Biomecânica

ID 161 · A 3D-Printed Micropipette for Open Lab Automation 3755

ID 591 · Active Wrist Orthosis for Parkinson Rigidity 3763

ID 64 · Análise da Influência das Características do Sinal no Desempenho de Diferentes Algoritmos de Extração da Dimensão Fractal em Dados Biomecânicos da Marcha Humana 3770

ID 399 · Aplicação de Metaestruturas Reentrantes no Desenvolvimento e Caracterização Mecânica de Prótese de Pé e Tornozelo 3777

ID 639 · Comparison between Two PWM Signal Generation Architectures for Servomotors Control in Pediatric Socially Assistive Robots 3783

ID 827 · Coordinative Variability Between Hip–Knee Joints During Gait Using Generalized Wavelet Coherence Analysis 3791

ID 713 · Data Gaps in Trust Measurement for Physical Human-Robot Collaboration 3796

ID 258 · Design and Implementation of PMSM Motor, Driver and Field Oriented Control Algorithm for Upper Limb Rehabilitation 3803

ID 295 · Development of an IoMT-Based System for Telemonitoring and Geolocation of Patients with Ventricular Assist Devices 3811

ID 636 · Hands-Free Endoscopic Camera Control via Real-Time Head-Pose Tracking for Robotic-Assisted Surgery 3820

ID 703 · Kinematic Evaluation of a Series Elastic Actuator-Powered Prosthetic Knee Joint 3826

ID 529 · Multimodal Sensory Feedback in a Game-Based Task for Robotic Rehabilitation 3831

ID 722 · Reliability and Usefulness of Kinovea Motion Analysis Software for Direct Measurement of Vertical Height During Counter Movement Jump in Down Syndrome 3839

ID 448 · Spectral Analysis of Ground Reaction Force on Standard Vertical Jump with Different Stretch-Shortening Cycle 3847

Tópicos Especiais (Bioética, Educação etc.)

ID 514 · A 3D-Printing Workflow for Biomedical Educational Models: A DNA Double-Helix Proof of Concept 3854

Sumário

VOLUME 01 · 25

Tópicos Especiais (Bioética, Educação etc.)

ID 280 · Advertised Competencies for Biomedical Engineers in Brazil: A Descriptive Study Based on Job Announcements in 2026 **3861**

ID 41 · An Analysis of the Persistent Availability of Unmarked Retracted Articles in Academic Social Networks **3866**

ID 793 · Aplicação de Metodologias Ativas na Formação em Engenharia Biomédica: Relato de Experiência na UFRN **3871**

ID 224 · Cognitive Load Assessment Using Low-Cost BCI **3879**

ID 446 · Development of a Low-Cost Horizontal Tensile Testing Machine for Mechanical Characterization of Materials **3884**

ID 222 · Development of a Low-Cost Pulse-Echo Teaching Apparatus for Brazilian Middle School Science Classes **3892**

ID 861 · Inatel Museum of Biomedical Engineering: A Proposal for an Interactive Collection of Resin-Printed Biological Educational Models **3898**

ID 100 · Learning Through Innovation: Student-Led Scientific Events as Transformative Learning Ecosystems in Biomedical Engineering Education: An Integrative Review **3903**

ID 595 · Paradox of AI-Assisted Academic Writing **3913**

ID 953 · Science Dissemination and Female Participation in Engineering Applied to Health Technologies **3919**

ID 692 · Short Videos on Anorexia/Bulimia with Expert Agreement **3924**

ID 198 · University Hospital as Drivers of Healthcare Innovation: Analyzing HCPATec's Experience in Fostering Startups and Intrapreneurship **3930**

ID 71 · VancoLearn: Web Platform for Teaching Vancomycin Therapeutic Drug Monitoring **3935**

Metrologia, Ensaios e Certificação em Engenharia Biomédica

ID 398 · 3D-Printed PLA Composites for Pediatric CT Phantom **3941**

ID 267 · Adaptation of a Public Dental Radiology Service to Anvisa RDC No. 611/2022: Experience Report at LABIM-UFPR **3949**

ID 732 · Analysis of Sunglass Frame Light Blocking Assessment Methods **3953**

ID 925 · Assessment of the Acoustic Power Emitted by a Newly Manufactured Dermatological Ultrasound Device **3958**

ID 215 · Calibração de Medidores de Campo Elétrico **3964**

ID 903 · Development and Validation of an Experimental Protocol for In Vitro Radiobiological Studies Using Low-Energy Proton Beams **3968**

ID 835 · Development of an In Vitro Protocol for Tumor Cell Irradiation with Low-Energy Protons **3976**

ID 760 · Development of Low-Cost Bone Phantoms for Application in Ho-166 Brachytherapy **3983**

ID 705 · Dimensional Validation Methodology of Low-Cost Phantoms for Magnetic Resonance Imaging Manufactured by Advanced Manufacturing **3991**

ID 22 · Heart Rate Calibration in Wearables via Biometric Signal Mimicry **3997**

ID 371 · MRI Quality Control after Cold-Head Replacement **4003**

Sumário

VOLUME 01 · 26

Metrologia, Ensaios e Certificação em Engenharia Biomédica

ID 326 · Safety and Conformity Evaluation of Medical Beds in the Post-Market Context **4007**

Nanotecnologia e Nanodispositivos Aplicada à Saúde

ID 499 · Caracterização Multimodal de uma Única Formulação de Nanopartículas Magnéticas **4013**

ID 736 · Cobalt Ferrite Nanoparticles for Biomedical Applications **4019**

ID 418 · Electrospun PVA Mats with Silver and Neomycin **4023**

ID 613 · GO-PEI-Ag Cytotoxicity with RF Ablation **4029**

ID 263 · Green Synthesis of Cerium Oxide Nanoparticles for Crystal Violet Removal with Potential Application in Healthcare Wastewater Treatment **4035**

ID 344 · Plasma-Activated Water and Silver Nanoparticles **4043**

A CAD System for Tooth Segmentation in Panoramic Radiographs Using U-Net Variants

1st Francisco Rubens Dutra Lima
Federal University of Ceará
Crateús, Brazil
franciscorubensdutra@alu.ufc.br

2nd Luéline Paiva Elias
Maurício de Nassau University Center
Fortaleza, Brazil
luelinelias@gmail.com

3rd Bruno Riccelli dos Santos Silva
Federal University of Ceará
Crateús, Brazil
bruno@lesc.ufc.br

4th José Wellington Franco da Silva
Federal University of Ceará
Crateús, Brazil
jwellingtonfranco@gmail.com

5th Esther Carneiro Ribeiro
Maurício de Nassau University Center
Fortaleza, Brazil
esther.carneiro.ribeiro@gmail.com

6th Fábio Wildson Gurgel Costa
Federal University of Ceará
Fortaleza, Brazil
fwildson@yahoo.com.br

7th Paulo Cesar Cortez
Federal University of Ceará
Fortaleza, Brazil
cortez@lesc.ufc.br

***Abstract*—Image segmentation, in automated analysis applications, enables the identification of regions of interest in several tasks in society. Regarding dentistry, it is useful for panoramic radiographs, which are widely used to provide a comprehensive view of dental and bone structures. However, the interpretation of these images is time-consuming, subject to inter/intra-observer variability, and relies on noise and low contrast. Recent works have enabled more accurate segmentation through the automatic extraction of relevant features using deep learning. Therefore, we aim to develop a CAD system for the automatic segmentation of teeth in panoramic radiographs through a comparative analysis of five U-Net-based architectures. The study used the public Children's Dental Panoramic Radiographs Dataset images, which underwent preprocessing, and models were trained using K-Fold cross-validation and Grid Search for hyperparameter optimization. In external validation data, W-Net demonstrated the best performance on one of the evaluated subsets, achieving a Sensitivity of 0.809, E-Measure of 0.925, IoU of 0.785, and DSC of 0.879. On the second subset, Attention U-Net achieved a Sensitivity of 0.965, E-Measure of 0.975, IoU of 0.858, and DSC of 0.924. Furthermore, the Wilcoxon test indicated significant differences between the classic U-Net and the U-Net++ and U-Net 3+ architectures, whereas W-Net and Attention U-Net showed equivalent performance. Finally, considering quantitative and qualitative results as well as inference time, Attention U-Net was selected as the final architecture for the CAD system and was subsequently evaluated by a dentist regarding the quality of segmentations generated on images independent of the training set.**

***Index Terms*—Medical image segmentation. CAD System. Segmentation models. Convolutional Neural Networks. Deep learning.**

I. Introduction

Oral health is defined as the state of the mouth, teeth, and orofacial structures that enables individuals to perform essential functions - such as eating, breathing, and speaking- while also encompassing psychosocial aspects like self-confidence and the ability to socialize. This condition varies throughout a person's life and plays a fundamental role in their overall health [1]. In this context, compromised oral health can lead to serious conditions such as dental caries, tooth loss, and orodental trauma; these rank among the most prevalent non-communicable chronic diseases worldwide, affecting approximately 3.5 billion people [2]. Consequently, oral diseases pose a major public health challenge, making early diagnosis essential for preventing complications and preserving dental structures [3].

Given this scenario, panoramic radiographs - also known as orthopantomograms - are vital tools for rapidly and accurately diagnosing these diseases; they provide comprehensive images of the entire dental arch, making them crucial for identifying various conditions. However, manual interpretation of these radiographs can be time-consuming and subject to inter-practitioner variability [4]. Beyond these limitations, technical factors such as low resolution, inadequate contrast, and image noise further complicate the process, increasing the difficulty and the risk of diagnostic errors. Against this backdrop, automated image analysis emerges as a promising tool capable of enhancing diagnostic accuracy, reducing screening time, and optimizing clinical costs. A standout among these solutions is the automated segmentation of medical images, a process that involves classifying each image pixel according to the objects of interest [5].

Image segmentation involves dividing an image into parts - referred to as connected pixel segments - that strongly correlate with real-world objects or areas within the scene, based on similarity criteria [6]. In a medical context, this technique plays a fundamental role in computer-aided diagnosis by

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 315447/2023-0.

making alterations to anatomical structures more distinct in images [7]. Specifically regarding the diagnosis of dental diseases such as caries and periapical lesions, segmentation is a key element that assists dentists in identifying these conditions; it allows for the delineation of dental structures and facilitates the distinction between teeth - particularly in panoramic radiographs, where the overlapping of structures can compromise visual interpretation [3].

Against this backdrop, deep learning has established itself in medical image analysis, finding broad application in tasks such as object detection, classification, segmentation, and registration [8]. This approach - a subset of machine learning - utilizes artificial neural networks with multiple layers to learn and recognize patterns in data. Concurrently, Convolutional Neural Networks (CNNs) have transformed medical image analysis in recent years, achieving remarkable performance in tasks such as classification, segmentation, detection, and image reconstruction [6]. Their ability to automatically extract relevant information eliminates the reliance on manual techniques and excessive preprocessing [9].Therefore, in this article, we propose the development of a CAD system for the automatic segmentation of teeth in panoramic radiographs, based on a comparative analysis of U-Net architecture performance and a clinical validation protocol.

The remainder of the article is organized as follows: Section 2 presents related work; Section 3 describes the methodology employed, covering model training and the development of the CAD system; Section 4 presents the results obtained; and, finally, Section 5 presents the study's conclusions.

II. RELATED WORK

Although recent studies have explored U-Net-based architectures for dental segmentation in panoramic radiographs, the current literature presents gaps that this work seeks to fill. Zannah et al. [3] evaluated U-Net architectures for semantic segmentation in panoramic radiographs, yet without performing a formal statistical comparison between the models or validation on multiple independent external datasets. Şahin, Ulutaş and Suzgen [4] they focused exclusively on U-Net 3+, not allowing for a direct performance comparison with other variants of the U-Net family under the same experimental conditions. Nader et al. [5] they used generic deep neural networks for tooth segmentation, without systematically exploring attention mechanisms or dense connections specific to more recent architectures. Brahmi and Jdey [6] They addressed tooth segmentation and identification using a single CNN architecture, without a comparative analysis of multiple models. Wang et al. [7] They presented a review of medical image segmentation methods using deep learning, of a theoretical rather than experimental nature. Finally, Zhong, Ren, and Zhang [8] proposed a grouped attention mechanism for dental X-ray segmentation, but this was restricted to a single architectural proposal, lacking systematic comparison with other U-Net variants or cross-validation on multiple external datasets.

Unlike those studies, this work makes five contributions. First, a systematic comparison of five U-Net family architectures under identical preprocessing, augmentation, Grid Search, and k-fold cross-validation, so that differences reflect the architectures rather than the protocol. Second, a Wilcoxon signed-rank analysis over the DSC values of every pair of models replaced the descriptive comparison with statistical evidence. Third, external validation on two datasets independent of the training set, assessing generalization under acquisition variability. Fourth, a functional CAD system with a Streamlit interface and PDF report export, integrating the Attention U-Net for its best performance. Fifth, a specialist evaluation by an uninvolved dentist over thirty stratified radiographs, scored with an eight-item Likert instrument, yielding 240 assessments that characterize behavior beyond aggregate metrics.

III. METHODOLOGY

A. Overview

The development of the automatic tooth segmentation model in panoramic radiographs followed a systematic workflow, beginning with dataset acquisition and culminating in the implementation of the computer-aided diagnosis (CAD) system. Figure 1 presents an overview of the steps involved in this process.

Panoramic radiographs and their corresponding segmentation masks were first organized and preprocessed to standardize the images for neural network training. The architectures evaluated in this study included U-Net, U-Net++, U-Net 3+, W-Net, and Attention U-Net. Model training incorporated data augmentation techniques to increase sample diversity and K-Fold cross-validation to ensure robust evaluation during optimization. Optimal parameters for each architecture were identified using the Grid Search technique. Following training, the models were evaluated on external datasets, which consisted of images not included in the training process, to assess the networks' generalization capabilities on unseen data.

The performance of each architecture was analyzed using several metrics: accuracy, sensitivity, specificity, Dice Similarity Coefficient (DSC), Mean Absolute Error (MAE), E-Measure, and Intersection over Union (IoU). The best-performing architecture was then selected and integrated into the computer-aided diagnosis (CAD) system, enabling the automatic generation of dental segmentations from submitted panoramic radiographs; the system was subsequently evaluated by a dentist regarding the quality of segmentations generated from images not used during training.

B. Dataset Acquisition

The dataset used was the Children's Dental Panoramic Radiographs Dataset [10], available on Kaggle, the structure of which is presented in Figure 2.

Among the available subsets, the adult dental segmentation set was selected; it is organized into three directories: Archive, containing 116 panoramic radiographs obtained at the Noor Medical Imaging Center (Qom, Iran); Panoramic Radiography Database, composed of 598 panoramic radiographs from

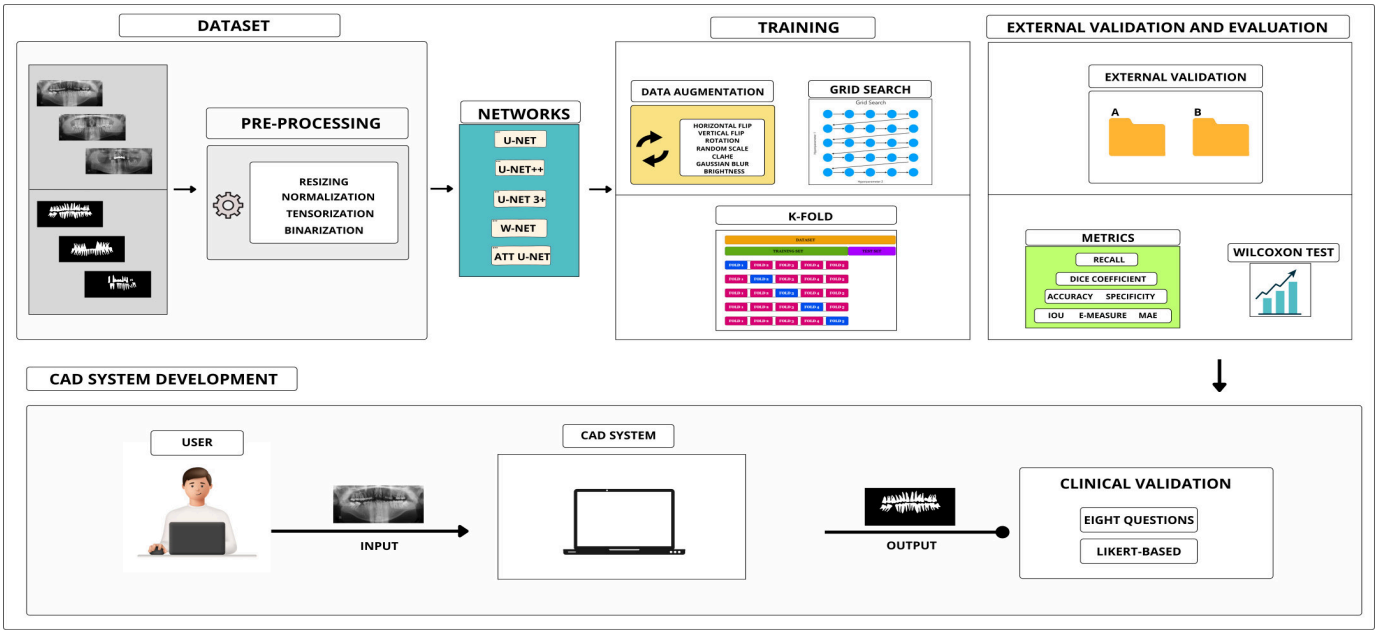


Fig. 1. Overview of the proposed methodology.

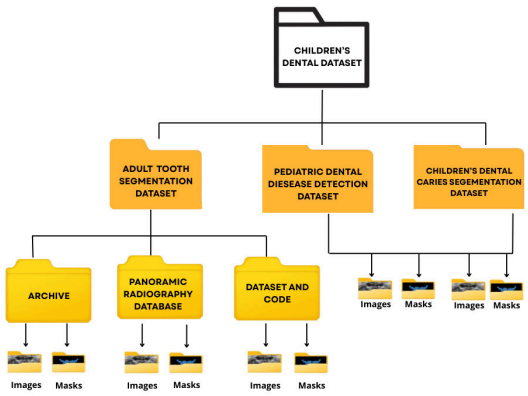


Fig. 2. Dataset structure.

randomly selected patients; and Dataset and Code, which includes 1,978 panoramic radiographs acquired at the UESB Diagnostic Imaging Center and the Hangzhou Lishui Dental Hospital. This set comprises ten image categories: (1) all teeth, with restorations and orthodontic appliances; (2) all teeth, with restorations but without appliances; (3) all teeth, without restorations but with appliances; (4) all teeth, without restorations and without appliances; (5) containing dental implants; (6) containing more than 32 teeth (supernumerary teeth); (7) missing teeth, with restorations and appliances; (8) missing teeth, with restorations but without appliances; (9) missing teeth, without restorations but with appliances; and (10) missing teeth, without restorations and without appliances. The Archive directory was used for training, while the Panoramic Radiography Database and Dataset and Code were used for external validation.

C. Data Preprocessing and Preparation

As described in Section B, the dataset contains panoramic radiographs of adult and pediatric patients. Since the objective is adult tooth segmentation, only the Adult Tooth Segmentation Dataset subset was used. During data analysis, 478 duplicate images were removed from the Dataset and Code directory, resulting in a final set of 2,214 images. During preprocessing, the images were converted to grayscale, resized from 2000×942 to 256×512 pixels, and normalized to the $[-1, 1]$ range. Subsequently, the images and masks were converted into tensors, and the masks were binarized, with values of 0 and 1 representing the background and the tooth region, respectively.

D. Neural Network Architectures for Segmentation

A literature review was conducted to identify neural network architectures suitable for automatic tooth segmentation in panoramic radiographs. U-Net-based architectures demonstrated strong performance in medical image segmentation applications and were therefore selected as the basis for the experiments in this study. The selected networks included U-Net, U-Net++, U-Net 3+, W-Net, and Attention U-Net. This selection enabled comparison of different approaches to image information processing, including inter-layer connections, multi-scale analysis, and attention mechanisms, to determine which models yield the best performance in dental segmentation.

E. Segmentation models training

All architectures were trained using the same experimental setup to ensure a fair comparison. Data augmentation was applied only to the training images using Albumentations,

incorporating transformations such as rotations, flips, resizing, brightness and contrast adjustments, CLAHE, and Gaussian blur. Training utilized the Archive directory, selected to evaluate model performance in a scenario with limited data. Hyperparameters were tuned via Grid Search, considering learning rates of 1×10^{-3} and 5×10^{-4}, a batch size of 8, and the Adam and AdamW optimizers. Each configuration was trained for 50 epochs using a loss function combining BCEWithLogitsLoss and Dice Loss. The optimal configuration for each architecture was determined through k-fold cross-validation ($k = 10$), based on the highest DSC on the validation set. Finally, the models were retrained using the best parameters, and their final weights were saved for evaluation and comparison.

F. External validations

To evaluate model performance on entirely new data not used during training, external validation was conducted using two additional adult patient datasets: the Dataset and Code set and the Panoramic Radiography Database.

Prior to inference, images and masks from the external datasets were prepared using the same processing pipeline as during training, including resizing, normalization, and conversion to PyTorch tensors, to maintain data standardization and ensure process consistency. The final versions of the models, trained using the optimal hyperparameters identified via grid search and cross-validation on the Archive dataset, were then loaded. Each architecture performed inference on all images in the external datasets and generated 10 PNG masks per set.

A smaller number of masks was intentionally generated to facilitate detailed visual analysis of network performance, specifically regarding contour accuracy, coherence of segmented regions, and preservation of dental structures.

G. Performance evaluation

To evaluate model performance during external validation using the two datasets mentioned in Section F, a set of metrics well-established in the literature and commonly used in image segmentation tasks was selected.

These metrics were chosen because they enable precise and objective quantification of the correspondence between regions identified by the models and the reference masks (ground truth). This approach measures both the level of overlap between areas of interest and the presence and nature of segmentation errors made by the networks. Based on the literature review, the final set of metrics applied in the study included Accuracy, Sensitivity, Specificity, Intersection over Union (IOU), Dice Similarity Coefficient (DSC), Mean Absolute Error (MAE), and Enhanced-alignment Measure (E-measure).

H. Statistical Analysis: Wilcoxon Test

After model evaluation using the defined metrics, the non-parametric Wilcoxon signed-rank test was applied to determine whether observed performance differences between networks were statistically significant. This test was selected because it is suitable for comparing results obtained from the same datasets without requiring the assumption of sample normality.

The test was conducted using Dice Similarity Coefficient (DSC) values obtained during k-fold cross-validation, with pairwise comparisons across all evaluated architectures. For each comparison, DSC values from the corresponding folds ensured analysis under identical conditions. A significance level of 5% ($\alpha = 0.05$) was adopted, with differences yielding a p-value less than or equal to this threshold considered statistically significant.

I. CAD System Development

After evaluating all models and conducting statistical analysis using the Wilcoxon test, the architecture demonstrating the best performance in segmenting dental structures was selected. This model was integrated into the CAD system developed in this project, enabling progression to the experimental phase and deployment via a computational interface.

The CAD system was built using the Streamlit library, which provides a web interface for interacting with the trained model. Through this system, users can upload a panoramic radiograph, which the neural network processes to automatically generate a tooth segmentation mask. Following image processing, the system displays the original radiograph alongside the model-generated mask, facilitating visualization of the segmentation results. Additionally, a feature was implemented to export results in PDF format, presenting the original image and the segmentation side-by-side. The system thus combines the developed deep learning model with an accessible interface, enabling dental professionals to utilize it in their analyses.

J. Preliminary Specialist Evaluation

In this work, we complemented the evaluation analysis with a evaluation conducted by a dentist who had no prior involvement in the system development. The professional received access to the deployed CAD interface by using the available link and was asked to process thirty randomly selected panoramic radiographs presented in the external validation set and inspect the mask returned by the system. It's worth noting that the thirty cases were drawn from the external validation set, so that none had been seen during training, and were selected across ten predefined groups of three images each in order to cover a range of anatomical and acquisition conditions rather than a convenience sample of straightforward cases. The presentation order was randomized, and the evaluator had access neither to the reference masks nor to any of the quantitative metrics. The assessment was structured as an eight-item instrument answered on a five-point Likert scale (1 = unsatisfactory, 2 = poorly satisfactory, 3 = partially satisfactory, 4 = satisfactory, 5 = fully satisfactory). The items were planned to decompose the segmentation quality when reading a mask presented as follows:

- Q1 - Completeness: every tooth visible in the radiograph is represented in the mask.
- Q2 - Contour: tooth crowns are delineated faithfully.
- Q3 - Root preservation: roots are segmented up to the apex, without truncation.

- Q4 - Interdental separation: adjacent teeth appear as distinct units, without fusion.
- Q5 - Continuity: segmented regions are intact, without fragmentation or internal holes.
- Q6 - Absence of false positives: the mask does not include bone, TMJ, spine, maxillary sinus, prosthetic devices, or other non-dental structures.
- Q7 - Shape and thickness fidelity: there is neither excessive thickening nor thinning of the structures.
- Q8 - Visual usefulness: the mask facilitates visualization of the dental structures in that specific case.

The scope of the instrument was constrained because the system performs binary segmentation only and no item addressed tooth numbering, disease detection, or report generation. The evaluator was explicitly instructed to judge the correspondence between the generated mask and the visible dental structures, and not the diagnostic content of the radiograph.

IV. Results and Discussion

A. Hyperparameter Selection

For each of the networks, the hyperparameters that yielded the highest average DSC value during the selection process are shown, taking into account the learning rate, batch size, and optimizer used. Table I presents these results.

TABLE I
BEST HYPERPARAMETER CONFIGURATION FOR EACH ARCHITECTURE.

Architecture	Batch Size	Learning Rate	Optimizer
Attention U-Net	8	5×10^{-4}	Adam
U-Net 3+	8	5×10^{-4}	Adam
U-Net++	8	5×10^{-4}	AdamW
U-Net	8	5×10^{-4}	AdamW
W-Net	8	5×10^{-4}	Adam

As shown in Table I, all architectures utilized the same batch size (8) and learning rate (5×10^{-4}) configurations; the only difference among them was the optimizer. Attention U-Net, U-Net 3+, and W-Net achieved better results using Adam, whereas U-Net and U-Net++ performed better with AdamW.

This difference in results suggests that more complex architectures - such as those featuring attention blocks, dense connections, and deeper hierarchies - may benefit more from Adam's flexibility in weight adjustment. Conversely, U-Net and U-Net++ architectures, which employ skip connections and progressive feature refinement, may have benefited more from the additional regularization provided by AdamW. This likely helped maintain stability during training and improved generalization capabilities. Finally, the best configurations were used for the final training of all architectures.

B. Evaluation on the test set

The results presented in Table II correspond to the performance of the architectures on the test set during hyperparameter tuning and 10-fold cross-validation. The values represent the mean and standard deviation across the ten iterations, allowing for an assessment of model consistency.

TABLE II
EVALUATION RESULTS ON THE TEST SET.

Metric	U-Net	U-Net++	U-Net 3+	Attention U-Net	W-Net
Accuracy	0.983 ± 0.002	$\mathbf{0.985 \pm 0.001}$	0.962 ± 0.002	0.983 ± 0.001	0.983 ± 0.002
Specificity	0.990 ± 0.002	$\mathbf{0.991 \pm 0.002}$	0.973 ± 0.005	0.990 ± 0.002	0.990 ± 0.002
Sensitivity	0.950 ± 0.011	$\mathbf{0.952 \pm 0.010}$	0.906 ± 0.018	0.947 ± 0.008	0.948 ± 0.010
E-Measure	0.977 ± 0.002	$\mathbf{0.979 \pm 0.001}$	0.946 ± 0.004	0.977 ± 0.002	0.977 ± 0.002
MAE	0.017 ± 0.002	$\mathbf{0.015 \pm 0.001}$	0.038 ± 0.002	0.017 ± 0.001	0.017 ± 0.002
IoU	0.905 ± 0.008	$\mathbf{0.913 \pm 0.006}$	0.801 ± 0.006	0.905 ± 0.006	0.905 ± 0.008
DSC	0.950 ± 0.004	$\mathbf{0.954 \pm 0.003}$	0.889 ± 0.004	0.950 ± 0.004	0.950 ± 0.004

As shown in Table II, the U-Net, U-Net++, Attention U-Net, and W-Net architectures demonstrated consistent performance on the test set, with high values for accuracy, specificity, and DSC. U-Net++ stood out by achieving the highest mean IoU 0.913 and DSC 0.954 values, indicating better agreement between the generated masks and the reference masks. The U-Net, Attention U-Net, and W-Net architectures yielded similar results. In contrast, U-Net 3+ showed inferior performance - particularly in terms of IoU 0.801 and DSC 0.889 - suggesting lower precision in delineating dental structures.

C. Results in Dataset and Code

The evaluation results for the Dataset and Code set are presented in Table III. These values correspond to the metrics obtained by the different architectures evaluated during external validation. This allows us to analyze model performance on a dataset distinct from the one used for training.

TABLE III
METRIC RESULTS IN THE DATASET AND CODE.

Metric	U-Net	U-Net++	U-Net 3+	Attention U-Net	W-Net
Accuracy	0.952	0.953	0.946	0.951	**0.955**
Specificity	**0.995**	0.994	0.991	0.994	0.992
Sensitivity	0.782	0.789	0.768	0.778	**0.809**
E-Measure	0.914	0.917	0.906	0.912	**0.925**
MAE	0.048	0.047	0.054	0.049	**0.045**
IoU	0.767	0.772	0.741	0.760	**0.785**
DSC	0.868	0.871	0.851	0.864	**0.879**

As shown in Table III, all architectures demonstrated high accuracy and specificity, yet DSC values were lower than those obtained during the cross-validation stage. U-Net yielded a DSC of 0.868, while U-Net++ achieved 0.871. U-Net 3+ exhibited the lowest IoU 0.741 and DSC 0.851, as well as the highest MAE 0.054. Attention U-Net showed intermediate performance, with a DSC of 0.864. W-Net achieved the highest sensitivity 0.809, IoU 0.785, and DSC 0.879, along with the lowest MAE 0.045. The reduction in DSC across all architectures may be partly attributed to the limited and uncontrolled representation of clinically relevant conditions in the dataset subset used for training, in which implants, orthodontic appliances, restorations, and supernumerary teeth were not mapped.

D. Results in the Panoramic Radiography Database

Table IV presents the results of the metrics obtained on the Panoramic Radiography Database. The values shown allow for a comparison of the performance of the architectures evaluated during external validation.

TABLE IV
METRIC RESULTS IN THE PANORAMIC RADIOGRAPHY DATABASE.

Metric	U-Net	U-Net++	U-Net 3+	Attention U-Net	W-Net
Accuracy	0.980	0.981	0.970	**0.982**	0.977
Specificity	0.983	**0.984**	0.978	**0.984**	0.980
Sensitivity	0.956	0.963	0.903	**0.965**	0.957
E-Measure	0.972	0.973	0.960	**0.975**	0.966
MAE	0.020	0.019	0.030	**0.018**	0.023
IoU	0.844	0.852	0.770	**0.858**	0.824
DSC	0.915	0.920	0.870	**0.924**	0.904

As shown in Table IV, all architectures achieved high accuracy and specificity, with DSC values ranging from 0.870 to 0.924. Unlike the Dataset and Code set, the models showed higher DSC values in this database, suggesting that external generalization varied according to dataset characteristics. U-Net achieved a DSC of 0.915, while U-Net++ reached 0.920. U-Net 3+ again yielded the lowest IoU of 0.770 and DSC of 0.870, with the highest MAE of 0.030. W-Net achieved a DSC of 0.904, whereas Attention U-Net obtained the best overall results, with sensitivity of 0.965, IoU of 0.858, DSC of 0.924, E-Measure of 0.975, and MAE of 0.018. The difference between the external datasets may be associated with variations in case composition and with the uncontrolled representation of implants, orthodontic appliances, restorations, and supernumerary teeth during training.

E. Statistical analysis using the Wilcoxon test

To determine whether the differences between the architectures were statistically significant, the Wilcoxon test was applied using the DSC values obtained during cross-validation. Table V presents the pairwise comparisons between the models, considering a significance level of $\alpha = 0.05$. The Wilcoxon test results presented in Table V show that

TABLE V
WILCOXON TEST: COMPARISON BETWEEN ARCHITECTURES.

Comparison	p-value	Significant
U-Net × Attention U-Net	1.00	No
U-Net × U-Net 3+	**0.01**	Yes
U-Net × U-Net++	**0.02**	Yes
U-Net × W-Net	1.00	No
U-Net 3+ × Attention U-Net	**0.01**	Yes
U-Net++ × Attention U-Net	**0.01**	Yes
U-Net++ × U-Net 3+	**0.01**	Yes
W-Net × Attention U-Net	0.69	No
W-Net × U-Net 3+	**0.01**	Yes
W-Net × U-Net++	**0.02**	Yes

the U-Net performed similarly to the Attention U-Net and W-Net, with no statistically significant differences. Conversely, differences were observed relative to the U-Net 3+ and U-Net++, indicating that their structural modifications influenced the segmentation results. The U-Net 3+ showed differences in most comparisons, while the U-Net++ also exhibited distinct behavior due to its dense connections. The W-Net maintained results close to those of the Attention U-Net but differed from the U-Net 3+ and U-Net++. These results indicate that more complex architectures do not always guarantee significant performance improvements.

F. Analysis of the segmented mask

Qualitative evaluation was performed through visual inspection of the masks generated by the architectures during external validation. Figure 3 presents examples of the segmentations obtained on the Panoramic Radiography Database. For illustrative purposes, two representative samples from each architecture were selected.

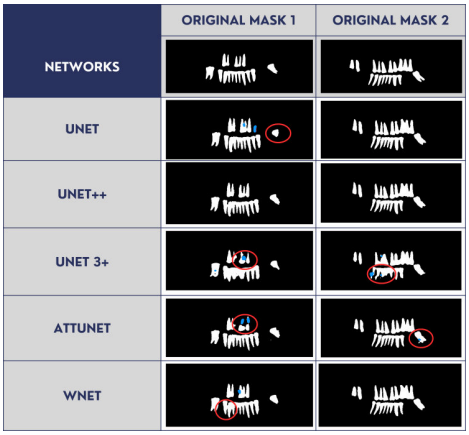


Fig. 3. Segmentation results: Panoramic Radiography Database.

Figure 3 shows that U-Net++ produced masks closer to the original images, better preserving the contours and continuity of dental structures. U-Net exhibited a minor false positive, whereas U-Net 3+ showed greater deviations, with incorrectly segmented regions and a loss of continuity in some teeth. Attention U-Net and W-Net yielded intermediate results; while they showed minor inaccuracies, they maintained good correspondence with the original masks. From a clinical perspective, these errors could compromise the accurate identification of dental structures during radiographic interpretation.

G. CAD system Integration

The choice of architecture to integrate into the CAD system took into account the models' quantitative and qualitative results as well as their computational cost. W-Net demonstrated the best performance on the Dataset and Code set, whereas Attention U-Net achieved the best results on the Panoramic Radiography Database. Balancing these criteria, an analysis of prediction time for 50 radiographs was conducted. Attention U-Net was selected for offering the best trade-off between performance and computational efficiency.



Fig. 4. CAD system interface.

Following this selection, Attention U-Net was integrated into the CAD system shown in Figure 4. The application allows for loading panoramic radiographs, viewing the input image and the segmented mask, and generating PDF reports containing the segmentation results.

H. Dentist Evaluation

The evaluator completed all thirty assessments, producing 240 individual scores, as summarized per item in Table VI.

Every item reached a maximum of 5, so each dimension was rated fully satisfactory in at least one radiograph. The lower quartiles are more discriminating than the medians. Q4 is the only item whose lower quartile is 2.00, which places at least a quarter of the radiographs at poorly satisfactory or worse for interdental separation; nine of the thirty received such a score. Q1 is the only item whose lower quartile is 4.00, so at least three quarters of the cases were satisfactory or better for completeness. The minimum scores split the items by dimension rather than by case: Q1, Q2 and Q7 never received the lowest score, whereas Q3, Q4, Q5, Q6 and Q8 reached the floor at least once, three radiographs each for Q4 and Q5 and one each for Q3, Q6 and Q8.

Dispersion is best read from the standard deviations, where continuity (Q5, 1.25) and absence of false positives (Q6, 1.18) vary most and shape fidelity (Q7, 0.76) least. Q4 and Q6 share the widest value, 2.00, but for structurally different reasons and should not be read as a pair. Q4 is unimodal, with twelve of the thirty radiographs scored 3, so its quartiles at 2.00 and 4.00 bracket a single center that happens to sit low. Q6 is bimodal (eleven radiographs scored 3 and thirteen scored 5, with only three in between), so its quartiles at 3.00 and 5.00 fall on the two modes themselves and measure the distance between them rather than spread about a center, which is also why its mean (3.83) falls below its median (4). Crown contour (Q2) shows the same structure in weaker form, with thirteen radiographs at 3 and eight at 5 against six at 4. The bimodality of Q6 is consistent with the stratified results in Table VII, where absence of false positives ranged from 2.00 in the implant group to the maximum of 5.00 in two others: false-positive control behaved in a near all-or-nothing fashion, apparently determined by the presence of radiopaque non-dental material rather than by degrees of difficulty.

Another aspect, presented in Table VII, is evidenced. The group means ranged from 2.46 to 4.50, roughly two points on a five-point scale, and only two groups reached a mean of 4.00 or above. Because each group contains three radiographs, the analysis that follows is only descriptive. Radiographs with missing teeth and neither restorations nor an orthodontic appliance (C10) obtained the highest mean, 4.50, and held the highest value of every item except interdental separation, including the maximum score for continuity and for absence of false positives in all three of its radiographs. Radiographs with supernumerary dentition (C6) obtained the lowest mean, 2.46, and held the lowest value of seven of the eight items. The implant group (C5) showed the combination of the maximum score for completeness with the lowest false-positive control of the series (2.00), and the evaluator explicitly recorded that bone had been confused with a root. Regarding each level of dentition completeness, the two groups containing restorative material occupied the two lowest positions in both cases, while absence of teeth carried no penalty at all: C10 scored above its complete-dentition counterpart (4.50 against 4.00), and no false positives were observed in edentulous regions. Completeness proved the most stable dimension, at or above 4.00 in nine of the ten groups, whereas interdental separation reached 4.00 in none.

These differences should be contextualized by how the model was trained. Fine-tuning used the Archive directory, 116 panoramic radiographs, which was not organized according to the conditions examined here: restorations, orthodontic appliances, implants and supernumerary dentition were neither balanced nor annotated, so their representation is uncontrolled, and for the rarer conditions it is plausible that few or no examples were seen. The groups in which the evaluator's scores were lowest are precisely those describing anatomies and radiopaque structures the network had no explicit opportunity to learn, while the highest scoring groups correspond to the configuration closest to a normative adult dentition free of foreign radiopaque material. The stratified results are therefore better understood as a characterization of the model's behavior outside the conditions represented in its training distribution than as a comparison of intrinsically harder and easier images, indicating that the composition of the fine-tuning corpus is the most reason why the validation did not reach higher ratings in these groups.

I. Limitations

This study presents some important limitations. Fine-tuning and hyperparameter selection were performed using only the "Archive" dataset, which comprises a limited number of cases with low diversity. This may have reduced the networks' ability to effectively handle situations involving implants, restorations, orthodontic appliances, or supernumerary teeth. Another point is that the analysis was conducted solely on adult patients, preventing the results from being applied to children. Clinical validation was also limited, relying on a single dentist and a small number of images, which further restricts the scope of the conclusions. Future studies should

TABLE VI
CONDUCTED ASSESSMENT BY THE DENTIST: DESCRIPTIVE STATISTICS PER ITEM.

Item	Description	Median	Q1	Q3	IQR	Mean	SD	Min	Max	% $\geq$ 4
Q1	Completeness	5	4.00	5.00	1.00	4.47	0.86	2	5	83.3
Q2	Crown contour	3	3.00	4.75	1.75	3.63	1.00	2	5	46.7
Q3	Root preservation	4	3.00	4.75	1.75	3.67	1.15	1	5	63.3
Q4	Interdental separation	3	2.00	4.00	2.00	2.97	1.07	1	5	30.0
Q5	Continuity	4	3.00	4.75	1.75	3.57	1.25	1	5	56.7
Q6	Absence of false positives	4	3.00	5.00	2.00	3.83	1.18	1	5	53.3
Q7	Shape/thickness fidelity	3	3.00	4.00	1.00	3.20	0.76	2	5	33.3
Q8	Visual usefulness	3	3.00	4.00	1.00	3.23	0.94	1	5	40.0
All responses (n = 240)		**4**	**3.00**	**5.00**	**2.00**	**3.57**	**1.11**	**1**	**5**	**50.8**

[a]IQR: interquartile range; SD: standard deviation; % $\geq$ 4: proportion of cases rated satisfactory or better.

TABLE VII
EVALUATION RESULTS STRATIFIED BY CASE GROUP.

Item	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	All
Q1	4.00	4.00	**5.00**	4.67	**5.00**	3.33	4.00	4.67	**5.00**	**5.00**	4.47
Q2	3.33	3.67	3.67	3.67	3.00	2.67	3.33	**4.67**	3.67	**4.67**	3.63
Q3	4.00	4.00	4.00	3.67	3.33	2.33	3.33	3.67	3.67	**4.67**	3.67
Q4	2.33	2.67	**3.67**	3.33	3.00	1.00	**3.67**	3.00	**3.67**	3.33	2.97
Q5	3.67	3.33	4.00	4.33	3.00	1.33	2.67	4.33	4.00	**5.00**	3.57
Q6	4.33	3.67	3.67	4.67	2.00	**5.00**	2.67	3.67	3.67	**5.00**	3.83
Q7	3.33	3.33	3.33	3.67	3.00	2.00	2.67	3.00	3.67	**4.00**	3.20
Q8	3.00	3.33	3.67	4.00	2.67	2.00	2.33	3.33	3.67	**4.33**	3.23
Mean	3.50	3.50	3.88	4.00	3.12	2.46	3.08	3.79	3.88	**4.50**	3.57

[a]Mean score per item within each group of three radiographs. Items Q1–Q8 are defined in Table VI. Bold: highest value in the row; underlined: lowest. Column "All" reproduces the item means over all 30 radiographs.

utilize larger, multicenter, and more diverse datasets that better represent various clinical conditions. It is also important to involve more specialists and expand the number of images evaluated to ensure a more consistent analysis and to gain a deeper understanding of the challenges associated with interdental separation and false-positive control.

V. CONCLUSIONS

Following the completion of all experiments and analyses, qualitative and quantitative results indicated that U-Net-based convolutional neural networks are suitable for tooth segmentation in panoramic radiographs, although performance differences were observed among the evaluated architectures. These differences were confirmed by the Wilcoxon test, which revealed statistically significant variations between certain models regarding Dice Similarity Coefficient (DSC) values, demonstrating that architectural modifications influence segmentation performance. The conducted evaluation by a dentist was also crucial for assessing the system's efficacy and identifying its limitations. Professional evaluation indicated satisfactory performance in identifying and visualizing dental structures, with good results regarding completeness, anatomical preservation, and mask continuity. However, limitations were observed regarding interdental separation, the preservation of segmented shapes, and the occurrence of false positives, highlighting the challenges inherent in segmenting adjacent and overlapping structures in panoramic radiographs. Finally, future work should consider the use of multi-center datasets, a broader external validation protocol, expert evaluation, and testing in real-world clinical environments to enhance the robustness and clinical applicability of the proposed CAD system.

REFERENCES

[1] WORLD HEALTH ORGANIZATION. Global strategy on oral health: draft for the 75th World Health Assembly. Geneva: WORLD HEALTH ORGANIZATION, 2022. Disponível em: ¡https://www.who.int/publications/i/item/9789240106031¿. Acesso em: 18 jul. 2025.

[2] WORLD HEALTH ORGANIZATION. Oral Health. 2022. Disponível em: https://www.who.int/health-topics/oral-health#tab=tab_1. Acesso em: 18 jul. 2025.

[3] ZANNAH, R. et al. Semantic segmentation on panoramic dental x-ray images using u-net architectures. IEEE Access, v. 12, p. 44598–44612, 2024

[4] SA¸HIN, A. E.; ULUTA¸S, H.; SÜZGEN, E. E. Automated segmentation of dental structures in panoramic radiographs using u-net 3+. In: 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP). [S.l.: s.n.], 2024. p. 1–6

[5] NADER, R. et al. Automatic teeth segmentation on panoramic x-rays using deep neural networks. In: 2022 26th International Conference on Pattern Recognition (ICPR). [S.l.:s.n.], 2022. p. 4299–4305.

[6] BRAHMI, W.; JDEY, I. Automatic tooth instance segmentation and identification from panoramic x-ray images using deep cnn. Multimedia Tools and Applications, v. 83, n. 18, p.55565–55585, 2024..

[7] WANG, R. et al. Medical image segmentation using deep learning: A survey. IET Image Processing, v. 16, n. 5, p. 1245–1267, 2022. Disponível em: ¡https://doi.org/10.1049/ipr2.124 19¿. Acesso em: 09 jun. 2025..

[8] ZHONG, W.; REN, X.; ZHANG, H. Automatic x-ray teeth segmentation with grouped attention. Scientific Reports, Nature Publishing Group UK London, v. 15, n. 1, p. 64, 2025.

[9] LIU, X. et al. A review of deep-learning-based medical image segmentation methods.Sustainability, v. 13, n. 3, p. 1224, 2021

[10] ZHANG, Y. et al. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data, v. 10, n. 1, p. 380, 2023.

A ConvNext-Small-based Classifier for Ischemic Stroke Identification in Computed Tomography Images

1st Emanuel T. de A. da Silva
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
emanuel.thiago.2018@gmail.com

2nd Luiz W. A. Rodrigues
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
luiz.walker@academico.ifpb.edu.br

3rd Gabriel A. de Santana
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
andrade.santana@academico.ifpb.edu.br

4th Villeneve de O. Soares
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
villeneveoliveira@gmail.com

5th Pedro A. de A. da Silva
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
pedrotatico1321@gmail.com

6th Carlos D. M. Regis
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—Ischemic stroke (IS) is a medical emergency whose early signs can be subtle on non-contrast computed tomography (CT). This study presents a preliminary ConvNeXt-Small classifier for distinguishing axial CT slices with visible ischemic changes from radiologically normal-appearing slices. The method uses DICOM images, conversion to Hounsfield Units, brain windowing, transfer learning, and a custom binary classification head trained in two stages. The dataset comprised 362 slices (181 per class) from two sources and was evaluated by five-fold cross-validation performed at the slice level. Mean slice-level accuracy was $92.82\% \pm 1.78\%$, precision $91.59\% \pm 3.13\%$, recall $94.47\% \pm 4.39\%$, specificity $91.17\% \pm 3.58\%$, and F1-score $92.92\% \pm 1.85\%$; 10 of 181 ischemic slices were false negatives. Because slices from the same examination were not constrained to a single fold, these estimates may be optimistic. The negative class did not include other intracranial abnormalities, and neither patient-level nor external validation was performed. Therefore, the results characterize a preliminary slice classifier and do not establish clinical diagnostic or screening performance.**

***Index Terms*—Ischemic Stroke, ConvNeXt, Computed Tomography, Transfer Learning, Slice Classification.**

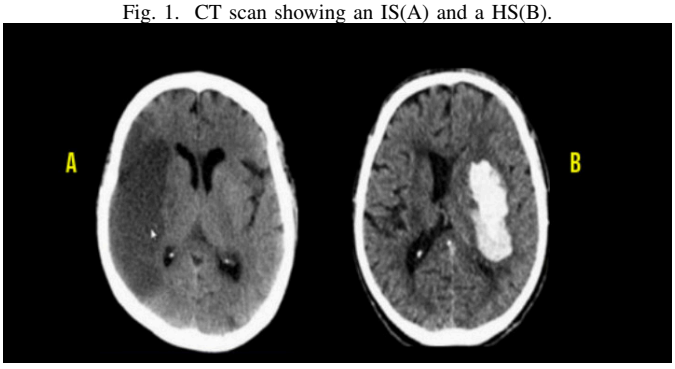
I. Introduction

Stroke, also known as cerebrovascular disease, ranks among the top three leading causes of mortality, behind only cancer and myocardial infarction [1]. It is directly responsible for leaving survivors with neurological sequelae, ranging from the loss of speech capability to partial paralysis of one side of the body, as observed by [1].

Stroke is a pathology that affects the blood vessels responsible for supplying the brain. It is characterized by an acute neurological deficit resulting from the death of brain cells caused by focal ischemia or hemorrhage [1], [2]. Global data from 2021 indicate that approximately 12.2 million new stroke cases occurred worldwide, with the disease being responsible for 6.63 million deaths [1], [2].

Stroke can be classified into two main types: ischemic and hemorrhagic. Ischemic stroke (IS), shown in Fig. 1(A), occurs due to the occlusion of a blood vessel, resulting in an insufficient blood supply to the brain tissue and causing localized cell death [3]. On a computed tomography (CT) scan, the ischemic region typically presents as a more muted, darker gray shade – characterized as hypodense – compared to other brain areas [2], [3]. Conversely, hemorrhagic stroke (HS), depicted in Fig. 1(B), results from the rupture of a blood vessel, causing bleeding within the brain parenchyma or the subarachnoid space; the brain region affected by the hemorrhage is represented as a hyperdense (white) signal on the CT scan [2], [3].

Fig. 1. CT scan showing an IS(A) and a HS(B).



Source: Adapted from [3].

The occurrence of ischemic stroke is predominant in emergency departments. In the acute phase, defined as the first 24 hours post-event, restoring blood flow to the affected area is vital to recover impaired functions, which enhances the probability of clinical reversal and reduces future neurological sequelae [2], [4], [5].

Recent computational advancements have enabled the development of software dedicated to iterative CT image reconstruction, allowing for simultaneous noise reduction and image quality improvement [6], [7]. Currently, the literature features computer-aided detection systems for brain imaging that utilize adaptive edge-preserving smoothing filters and wavelet-based image processing [6], [7].

This work aims to implement and evaluate a preliminary convolutional neural network based on the *ConvNeXt* architecture for the slice-level discrimination of visible ischemic changes from normal-appearing cranial CT. The dataset was obtained from the studies of [3] and [8], consisting of CT scans. The present experiment does not evaluate patient-level diagnosis or clinical triage performance.

II. Ischemic Stroke

Ischemic stroke accounts for approximately 85% of all cases and occurs due to the occlusion of an arterial blood vessel, leading to the interruption of blood flow to a specific region of the brain [9].

The interruption of cerebral blood flow initiates a sequence of biochemical and cellular events known as the ischemic cascade [9], [10]. This blood deprivation is followed by a depletion of oxygen and glucose supply to the affected brain region; consequently, local cells undergo cell death via necrosis or apoptosis [9], [10]

According to its progression, ischemic stroke can be classified into three stages: acute (0–24 h), subacute (24 h–2 weeks), and chronic (more than 2 weeks) [11]. The efficacy of the treatment is directly related to the time elapsed until the onset of medical intervention for the stroke patient [5].

The diagnosis of IS relies on clinical and imaging assessment. Non-contrast head CT is widely used as the first-line imaging examination in emergency settings because it is fast, widely available, and effective for excluding intracranial hemorrhage [12]. However, the poor grayscale contrast of early ischemic changes can challenge radiologists and potentially contribute to diagnostic errors [13].

III. Convolutional Neural Network

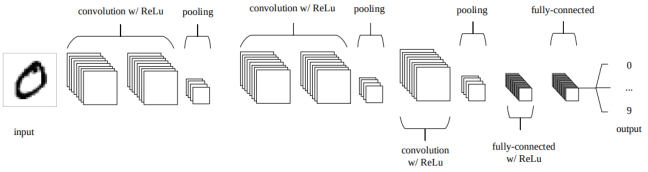
Convolutional Neural Networks (CNNs) are a specific type of neural network uniquely tailored for computer vision applications due to their ability to hierarchically abstract representations through local operations [14].

In contrast to traditional neural networks that require input data to be flattened into a single dimension, CNNs retain the spatial architecture of the data. This characteristic makes them uniquely suited for processing images and other multidimensional datasets [15].

Feature extraction from CNN input data, as shown in Fig. 2, is performed through convolutional layers that apply spatial filters over the input to extract relevant patterns [16]. These features are then aggregated and processed by subsequent layers, such as *pooling* layers and fully connected layers [16].

In this work, the application of convolutional neural networks will be focused on extracting significant features from human brain CT scans. The goal is to detect the presence of any IS markers based on the learned characteristics.

Fig. 2. Baseline CNN's Architecture.



Source: [17].

A. ConvNeXt Architecture

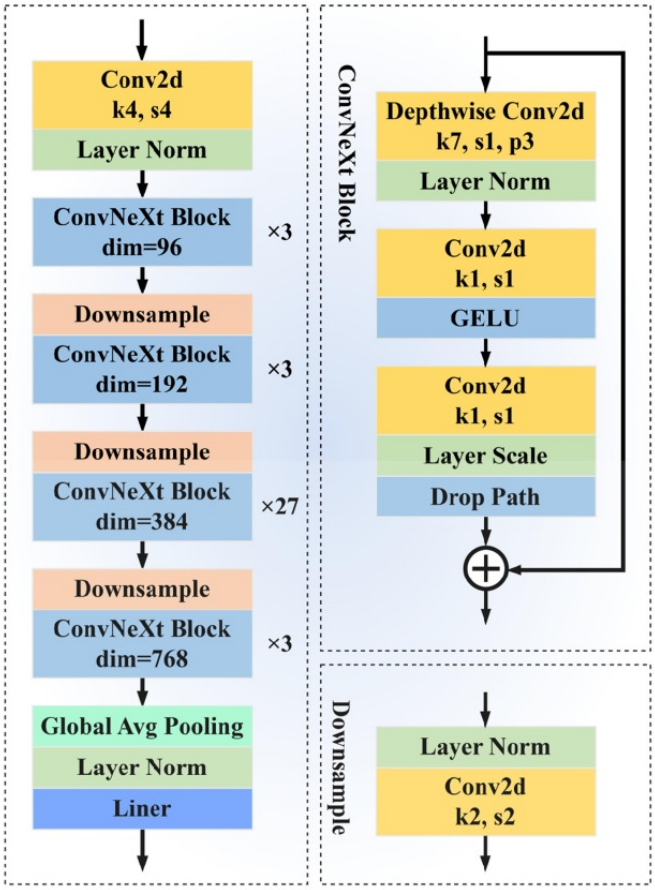
The recent enhancement of CNNs is a response to the popularized use of *Vision Transformers* (ViTs) in computer vision tasks. ViTs, introduced in the conference paper by [18], are learning architectures that apply Natural Language Processing (NLP) principles to computer vision challenges. ViTs operate by partitioning the image into patches – a function performed by a dedicated 'patchify' layer – and treating them as a sequence of tokens analogous to words in a text [18]. The ViT algorithm enables superior performance when pre-trained on massive datasets (14–300 million images), requiring less computational cost and showing no signs of performance saturation as their scale is expanded [18], [19]. Conversely, CNNs – particularly modernized architectures such as ConvNeXt – remain competitive and efficient in small-dataset regimes due to their inherent local inductive bias [19], [20].

The ConvNeXt architecture is a family of residual CNNs that emerged from a process of refining and modernizing a standard network, specifically the ResNet (Residual Network) family [19]. This modernization took place during the architecture's training for image classification tasks using the ImageNet dataset. Subsequently, the efficacy of the final architecture was validated across other machine learning tasks, such as object detection and segmentation using the COCO dataset, and semantic segmentation through the ADE20K dataset [19].

The ConvNeXt family comprises five variants: Tiny, Small, Base, Large, and XLarge; Fig. 3 illustrates the architecture in its Small configuration. The initial layer of ConvNeXt, known as the 'patchify' layer (represented by the Conv2d k4, s4 layer), partitions the input data into patches. This process is achieved by applying 4×4 kernels with a stride of 4, resulting in a non-overlapping convolution [19]. The input layer is followed by a normalization layer (Layer Norm) [19], [21]. These two initial layers are succeeded by four stages of ConvNeXt Blocks, where the number of blocks per stage is 3, 3, 27, and 3, respectively [19], [21]. The channel dimension increases at each stage from 96 to 192, 384, and 768; additionally, the last three stages of *ConvNeXt blocks* are preceded by a downsampling block [21]. At the end of the processing stages, the feature tensor passes through a global average pooling layer followed by a final Layer Norm. Lastly,

the tensor is transformed into a normalized feature vector that is fed into the final linear layer [19], [21].

Fig. 3. Block diagram of the ConvNeXt-Small architecture.



Source: [21].

The ConvNeXt architecture incorporates two sub-blocks within its structure, as shown in Fig. 3: the Downsample block and the ConvNeXt Block [21]. The Downsample block is composed of a Layer Norm followed by a 2×2 convolution (Conv2d k2, s2) [19], [21]. Notably, the spatial reduction performed by this block does not occur inside the residual blocks as seen in ResNets; this specific characteristic helps the ConvNeXt architecture retain critical spatial information and enhances training stability [19], [21].

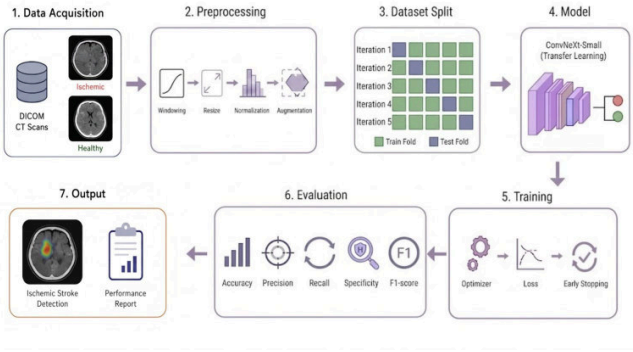
The ConvNeXt block, shown in Fig. 3, is responsible for extracting features from the partitioned data originating from the model's initial layers [19]–[21]. The first layer of the block (Depthwise Conv2d) applies filters separately to each channel of the input data, saving memory and computational power during feature extraction [19], [20]. Following spatial feature extraction, the data pass through a single Layer Norm [21]. Replacing Batch Normalization with Layer Norm contributes to model convergence and helps mitigate overfitting, bringing the block closer to the simplicity of ViTs, which utilize fewer normalization layers [19], [21]. The subsequent layer applies a convolution operation called an inverted bottleneck to minimize feature loss and reduce computational cost [20], [21]. This operation is performed with the aid of two pointwise convolution operations; the first pointwise convolution (Conv2d k1, s1) expands the signal dimension by four times, while the second pointwise convolution (Conv2d k1, s1) contracts the signal back to its input dimensions [19]–[21]. Between these pointwise convolutional layers, a layer applying the GELU (Gaussian Error Linear Unit) activation function is integrated, replacing the ReLU (Rectified Linear Unit) function commonly found in traditional CNNs [19], [22]. This substitution is motivated by GELU's smoother graphical curve compared to ReLU and its widespread adoption in ViTs; mathematically, GELU is represented by Equation 1, where the term $\phi(x)$ denotes the cumulative distribution function of the Gaussian distribution [19], [22]. In the final stages of the *ConvNeXt block*, the tensor passes through the layer scale and drop path layers; the former stabilizes the network at the beginning of training, while the latter regularizes model training by randomly deactivating convolutional blocks [19]. Finally, the output of the drop path layer is added to the original input tensor via a shortcut connection, facilitating gradient propagation and mitigating the vanishing gradient effect [19]–[21].

$$GELU(x) = x * \phi(x) \tag{1}$$

IV. Methodology

This section details the methodological steps adopted in this work. Below, we describe the dataset utilized, the preprocessing techniques applied to the images, the characteristics of the employed classifier, the training strategy adopted, and the performance metrics used for model evaluation. The figure 4 shows the steps adopted.

Fig. 4. Methodological workflow.



Source: Authors.

A. Dataset

This work utilizes cranial computed tomography (CT) images obtained from two distinct datasets. The images from both datasets were jointly reviewed and labeled by a radiologic technologist and a physician, who assessed the examinations simultaneously. The final class assigned to each selected slice was established by consensus between the two evaluators,

according to the presence or absence of visible ischemic changes.

The first dataset was provided by the *Núcleo de Estudos e Pesquisa em Imagem* (NEPI) of the *Universidade Federal da Paraíba* (UFPB) and was approved by the institutional review board under protocol number 1,293,095 [3]. It is composed of cranial CT images with a spatial resolution of 512×512 pixels, stored in the DICOM format. Each CT examination contains approximately 20 axial slices. The acquisitions selected from this dataset had a slice thickness of 5 mm.

The second dataset, provided by [8], is also composed of cranial CT images with a spatial resolution of 512×512 pixels, stored in the DICOM format. However, these examinations present greater volumetric coverage, with approximately 200 axial slices per patient. The original dataset includes acquisitions with slice thicknesses of 3 mm, 5 mm, and 8 mm.

To reduce the variability associated with different acquisition protocols, only CT series acquired with a slice thickness of 5 mm were selected from both datasets. The same review, selection, and labeling procedure was applied to the images from both sources.

After the identification of examinations containing ischemic stroke, an axial-slice filtering procedure was jointly performed by the two evaluators. This step was necessary because the original examinations covered the entire cranium and included slices with little or no relevant anatomical information for ischemic stroke assessment. Therefore, slices lacking sufficient brain parenchyma, containing inadequate anatomical regions, or not allowing a reliable assessment of ischemic changes were excluded.

For the positive class, only slices presenting visible ischemic changes were included. For the negative class, only slices with sufficient brain parenchyma and no visible signs of ischemic stroke or any other apparent intracranial pathology were selected. Therefore, the negative samples were considered radiologically normal based on the joint visual assessment performed by the radiologic technologist and the physician. The final label assigned to each slice was established by consensus between the two evaluators.

The final dataset comprised 362 axial CT slices: 181 with visible ischemic changes and 181 radiologically normal-appearing slices. Each slice was treated as an independent sample for the binary classification task. The curated experimental record did not provide a reliable mapping of every selected slice to a unique patient or examination; consequently, the number of unique examinations in each fold cannot be reported retrospectively.

The classifier was evaluated using five-fold cross-validation. The dataset was divided into five folds while maintaining an approximately balanced distribution between the ischemic and normal-appearing classes. Partitioning was conducted at the slice level, not at the patient level. Thus, slices originating from the same patient or examination may have been assigned to different folds, creating a risk of information leakage through correlated anatomy and acquisition characteristics. In each iteration, four folds were used for model development, whereas the remaining fold was held out exclusively for testing. This procedure was repeated five times, so that each image was included in the test set once and was evaluated by a model that had not used that image during training. This protocol prevents reuse of the same slice across training and testing, but it does not constitute patient-independent or external validation and may therefore overestimate generalization.

B. Data Preparation

The selected CT images were submitted to a preprocessing pipeline before being used as input to the neural network. Initially, the DICOM images were read and the raw pixel values were converted to Hounsfield Units (HU) using the rescale slope and rescale intercept parameters available in the image metadata. This step was performed to preserve the physical interpretation of CT images intensities.

Afterward, an intensity windowing operation was applied to highlight brain tissue structures and reduce the influence of irrelevant intensity ranges, utilizing a window center of 0 HU and a window width of 150 HU. The windowed images were clipped and rescaled to an 8-bit grayscale representation. Then, the images were resized to the input resolution required by the ConvNeXt architecture (224×224 pixels). Since the pretrained model expects three-channel inputs, each grayscale slice was converted into a three-channel image by replicating the same intensity values across the channels.

Finally, the images were normalized according to the preprocessing function associated with the pretrained network. During training, data augmentation was applied exclusively to the training subset to reduce overfitting and improve model generalization. The augmentation operations included random rotations, spatial shifts, zoom transformations and horizontal flipping. No data augmentation was applied to the validation or test subsets.

C. ConvNeXt-Based Classifier

The classification model was built using a ConvNeXt-Small architecture pretrained on the ImageNet dataset. The use of transfer learning was adopted to take advantage of visual representations previously learned from large-scale image datasets, reducing the need to train the entire network from scratch.

The original classification head of the ConvNeXt model was replaced by a custom binary classification head. First, a Global Average Pooling 2D layer was used to convert the feature maps extracted by the convolutional backbone into a one-dimensional feature vector. Then, a layer normalization step was applied to stabilize the feature distribution. After that, a fully connected dense layer with 512 neurons and GELU activation was added to learn task-specific representations. A dropout layer with a rate of 50% was included to reduce overfitting. Finally, the output layer was composed of a single neuron with sigmoid activation, producing the probability of the input image belonging to the ischemic stroke class.

D. Training Strategy

The training process was divided into two stages. In the first stage, the ConvNeXt backbone was kept frozen, and only the

custom classification head was trained. This stage allowed the newly added layers to adapt to the CT images classification task without modifying the pretrained weights of the base model. The model was trained for up to 40 epochs using the Adam optimizer with an initial learning rate of 10^{-3}.

In the second stage, fine-tuning was performed by unfreezing the last 25% of the ConvNeXt backbone layers. This allowed the deeper layers of the network to adapt to the specific visual patterns present in brain CT images while preserving most of the pretrained representations. During this stage, the model was trained for up to 30 additional epochs using a reduced learning rate of 10^{-5}.

In both training stages, early stopping was used to monitor the validation loss and interrupt training when no improvement was observed. In addition, the ReduceLROnPlateau callback was applied to automatically reduce the learning rate when the validation loss reached a plateau.

The parameters used during the training of the ConvNeXt-Small-based classifier are presented in Table I. Adam's first-moment exponential decay coefficient was set to $\beta_1 = 0.9$, whereas the weight-decay coefficient was 0.001. Any unspecified parameters remained identical across both stages.

TABLE I
TRAINING PARAMETERS OF CONVNEXT-SMALL-BASED CLASSIFIER.

Batchsize	16
Image Resolution	224x224x3
Epochs	First stage: 40 Second stage: 30
Early Stopping	patience: 15 epochs
Optimizer	Adam
Activation Function	GELU
Learning Rate	First stage: 10^{-3} Second stage: 10^{-5}
Adam β_1	0.9
Weight Decay	0.001
Layer Normalization	$\epsilon = 10^{-6}$
Fully Connected Layer Size	512 neurons
Output Layer (function)	01 neuron (Sigmoid)
Dropout	0.5
ReduceLROnPlateau	factor: 0.2 patience: 5 min_lr: 10^{-7}
Data Augmentation	Random Rotation: $0° - 15°$ Random Zoom: 10% Horizontal Flip Width and Height Shift: 10%

E. Performance Evaluation

The algorithm implemented for ischemic stroke detection in CT images was evaluated using performance metrics to measure its degree of efficiency in the classification task. Among the utilized metrics are Accuracy (Acc), Precision (P), Recall (R), F1-score ($F1$) and Specificity (Sp).

Accuracy, calculated using Equation 2, represents the proportion of all correct classifications, including both positive and negative results. It is generally employed as the standard evaluation metric for general or unspecified models performing non-specific tasks.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Precision, recall, and F1-score are mathematically represented, respectively, in Equations 3, 4 and 5.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{5}$$

Specificity, calculated via Equation 6, is used to evaluate the performance of a binary classification model, particularly in medical diagnosis or screening tests. It measures the ability of the model to correctly identify true negatives, or the proportion of actual negative cases correctly classified [23].

$$Sp = \frac{TN}{TN + FP} \tag{6}$$

In this classification task, True Positives (TP) represent positive samples correctly classified as positive, whereas False Positives (FP) denote negative samples erroneously classified as positive. Conversely, True Negatives (TN) define negative samples correctly classified as negative and False Negatives (FN) correspond to positive samples erroneously classified as negative.

In addition to the fold-wise mean and sample standard deviation, 95% Wilson score confidence intervals were calculated for the aggregate slice-level accuracy, precision, recall, and specificity. These intervals treat slices as independent observations and are provided only as descriptive measures. Because patient grouping was unavailable, they do not account for within-patient correlation and should not be interpreted as patient-level uncertainty estimates. Receiver operating characteristic area under the curve (ROC-AUC) could not be reconstructed from the retained hard-label predictions and confusion matrices and is therefore not reported.

V. EXPERIMENTAL RESULTS

This section presents the classification results obtained by the ConvNeXt-Small-based classifier using slice-level five-fold cross-validation. The CT slices were categorized into two classes: radiologically normal-appearing (class 0) and visible ischemic change (class 1). The dataset contained 181 slices from each class, totaling 362 samples.

The folds were constructed while maintaining an approximately balanced class distribution. The first test fold contained 37 samples from each class, whereas the remaining four folds contained 36 samples from each class. In each cross-validation iteration, four folds were used for model development and the remaining fold was used exclusively for testing. Therefore, each slice was evaluated once by a model that had not used that exact slice during training. Because folds were not grouped by

patient, this statement does not exclude the presence of other slices from the same examination in the development folds.

Table II presents the confusion matrix components and the performance metrics obtained for each test fold. The final results are reported as the mean and sample standard deviation across the five folds.

Across the five test folds, the classifier achieved a mean accuracy of $92.82\% \pm 1.78\%$. The highest accuracy was obtained in folds 4 and 5, both reaching 94.44%, whereas fold 2 presented the lowest accuracy, at 90.28%. The relatively low standard deviation indicates limited variation across these particular slice-level partitions; it does not establish stability across independent patients or institutions.

The classifier achieved a mean recall of $94.47\% \pm 4.39\%$, which was the highest average metric. Recall is particularly relevant in ischemic stroke screening because it represents the model's ability to correctly identify pathological cases. The recall ranged from 88.89% in fold 2 to 100.00% in fold 5. In fold 5, all ischemic stroke slices were correctly identified, resulting in no false-negative classifications.

The mean precision was $91.59\% \pm 3.13\%$, indicating that most slices classified as ischemic stroke were actually positive. The highest precision was obtained in fold 4, reaching 97.06%, as only one normal-appearing slice was incorrectly classified as ischemic stroke.

The classifier also achieved a mean specificity of $91.17\% \pm 3.58\%$, demonstrating its ability to correctly identify normal-appearing slices. The mean F1-score of $92.92\% \pm 1.85\%$ indicates a consistent balance between precision and recall across the five test folds.

To provide an overall representation of the classification behavior, the out-of-fold predictions from the five test subsets were combined to construct the aggregated confusion matrix shown in Fig. 5. Each slice contributed to this matrix only once and was classified while belonging to a held-out test fold.

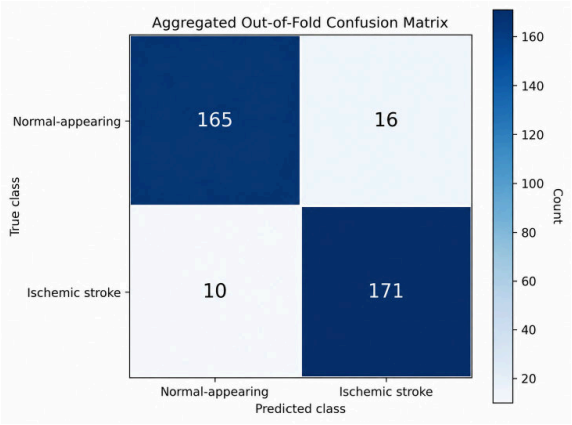


Fig. 5. Aggregated slice-level out-of-fold confusion matrix obtained using five-fold cross-validation. The negative class contains only radiologically normal-appearing slices.

Source: Authors.

The aggregated confusion matrix contained 165 true negatives, 16 false positives, 10 false negatives, and 171 true positives, corresponding to all 362 out-of-fold predictions. Based on this matrix, the classifier achieved an overall accuracy of 92.82%, precision of 91.44%, recall of 94.48%, specificity of 91.16%, and F1-score of 92.93%.

The corresponding descriptive 95% Wilson confidence intervals were 89.68–95.05% for accuracy, 86.55–94.66% for precision, 90.13–96.97% for recall, and 86.12–94.49% for specificity. These intervals use slices as the statistical units and consequently do not correct for correlations among slices from the same patient. ROC-AUC is not reported because only thresholded class labels and confusion-matrix counts were retained for this analysis; class probabilities would be required to calculate the ROC curve.

The number of false-negative predictions was lower than the number of false-positive predictions, indicating that the classifier tended to favor sensitivity. For a putative screening application, false-negative classifications would represent visible ischemic slices missed by the system. The occurrence of 10 false negatives, together with the non-independent validation protocol, precludes claims of clinical screening performance.

The misclassified samples may be associated with the subtle visual appearance of ischemic lesions in CT images, particularly when the attenuation difference between normal and affected brain tissue is low. Variations in lesion location, size, and ischemic stroke stage may also have influenced the classification performance across the different folds.

VI. Discussion

This study investigated the use of a ConvNeXt-Small-based classifier for the identification of ischemic stroke in computed tomography (CT) images. The proposed approach combined DICOM-based preprocessing, conversion to Hounsfield Units, intensity windowing, image resizing, normalization, data augmentation, and transfer learning using a modern convolutional architecture.

The model was evaluated using five-fold cross-validation. Across the five test folds, it achieved a mean accuracy of $92.82\% \pm 1.78\%$, precision of $91.59\% \pm 3.13\%$, recall of $94.47\% \pm 4.39\%$, specificity of $91.17\% \pm 3.58\%$, and F1-score of $92.92\% \pm 1.85\%$. The relatively low standard deviations observed for accuracy and F1-score describe consistency across the five slice-level partitions only. They should not be interpreted as evidence of generalization to independent patients, scanners, or institutions.

The aggregated out-of-fold confusion matrix contained 165 true negatives, 16 false positives, 10 false negatives, and 171 true positives. Based on these aggregated predictions, the classifier achieved an overall accuracy of 92.82%, precision of 91.44%, recall of 94.48%, specificity of 91.16%, and F1-score of 92.93%. The similarity between the mean cross-validation metrics and the aggregated results is expected because both summaries are derived from the same five sets of out-of-fold slice predictions.

Among the evaluated metrics, recall describes the fraction of selected ischemic slices identified by the classifier. The mean recall was 94.47%; in the aggregated evaluation, 171

TABLE II
SLICE-LEVEL CLASSIFICATION PERFORMANCE ACROSS THE FIVE TEST FOLDS.

Fold	TN	FP	FN	TP	Accuracy	Precision	Recall	Specificity	F1-score
1	33	4	2	35	91.89%	89.74%	94.59%	89.19%	92.11%
2	33	3	4	32	90.28%	91.43%	88.89%	91.67%	90.14%
3	32	4	1	35	93.06%	89.74%	97.22%	88.89%	93.33%
4	35	1	3	33	94.44%	97.06%	91.67%	97.22%	94.29%
5	32	4	0	36	94.44%	90.00%	100.00%	88.89%	94.74%
Mean ± SD	–	–	–	–	$\mathbf{92.82 \pm 1.78\%}$	$\mathbf{91.59 \pm 3.13\%}$	$\mathbf{94.47 \pm 4.39\%}$	$\mathbf{91.17 \pm 3.58\%}$	$\mathbf{92.92 \pm 1.85\%}$

of the 181 ischemic slices were correctly classified and 10 were incorrectly labeled as normal-appearing. Although recall varied from 88.89% to 100.00% across folds, these values apply to slices rather than patients and may be inflated by within-examination similarity across folds.

In a clinical setting, a missed ischemic stroke could delay therapeutic intervention. However, the present experiment neither measures the number of missed patients nor validates a clinical decision threshold. Accordingly, the model should be described as a preliminary research classifier, not as a validated screening or decision-support system.

The aggregated precision of 91.44% indicates that most slices classified as ischemic stroke were actually positive. However, 16 normal-appearing slices were incorrectly classified as ischemic. Although false positives may lead to additional clinical investigation, this type of error is generally considered less critical than false negatives in screening scenarios. The lower number of false negatives compared with false positives suggests that the classifier slightly favored sensitivity, prioritizing the detection of ischemic stroke cases.

Within the selected normal-appearing slices, the specificity was 91.16%. In addition, the aggregated F1-score of 92.93% indicates a favorable balance between precision and recall, suggesting that the high sensitivity was not obtained at the cost of an excessive number of false-positive classifications.

The study conducted by [24] employed a 3D attention-gated U-Net to predict the final size and location of ischemic stroke lesions, achieving a median AUC of 91%, specificity of 97%, sensitivity of 60%, and a median Dice similarity coefficient of 50%. The recall obtained in the present study was substantially higher than the sensitivity reported in that work. Nevertheless, the comparison must be interpreted carefully because lesion prediction and spatial localization constitute a more complex task than slice-level binary classification.

In comparison with the CNN-based approach proposed by [25], which achieved 92% accuracy, 92% precision, 87% recall, and 89% F1-score for clot classification, the ConvNeXt-Small-based classifier obtained slightly higher mean accuracy, substantially higher recall, and a higher F1-score. Its mean precision was slightly lower. These results indicate that the proposed model was more sensitive to pathological cases while maintaining a similar ability to avoid false-positive predictions.

The work of [26] proposed an automated early ischemic stroke detection system based on CNNs using CT images in DICOM format. Their method included several handcrafted preprocessing steps, such as Otsu-based segmentation, bone removal, cerebrospinal fluid elimination, and lesion-centered patch extraction. The reported testing accuracy was 92.96%, which is slightly higher than the mean accuracy of 92.82% obtained in the present study. However, the proposed approach employs ConvNeXt-Small transfer learning and performs slice-level classification using a more direct preprocessing pipeline, without relying on lesion-centered patch extraction.

Despite the promising results, some limitations must be considered. First, the dataset remained relatively small, containing 181 ischemic stroke slices and 181 normal-appearing slices. Although five-fold cross-validation reduced the dependence of the results on a single training and test partition, it does not replace validation using an independent external dataset.

Second, fold assignment was performed at the slice level without grouping by patient or examination. Anatomically adjacent slices from the same examination are correlated and can share patient-specific and scanner-specific characteristics. If such slices occurred in both development and test folds, the model could partly exploit these shared characteristics, producing optimistically biased estimates. Moreover, the number of unique examinations per fold was not retained in the curated experimental record. The current results therefore describe only slice-level discrimination and cannot be converted into patient-level sensitivity or specificity. A valid reassessment must first reconstruct patient identifiers, assign all slices from one examination exclusively to a single fold, report the number of patients and examinations in every partition, and aggregate slice probabilities using a prespecified patient-level decision rule.

Third, the negative class was restricted to radiologically normal-appearing slices. Consequently, specificity and false-positive counts refer only to differentiation from normal anatomy and do not measure discrimination against common emergency findings such as intracranial hemorrhage, tumors, chronic infarctions, postoperative changes, edema, or artifacts. A clinically representative test cohort must include these competing abnormalities.

Fourth, there was no independent external test set containing images acquired at different medical centers and using different CT scanners and acquisition protocols. External validation is necessary to determine whether the model can generalize to clinical conditions that differ from those represented in the training dataset. Although images were collected from two sources, they were pooled before cross-validation; hence, the second source was not used as an external validation cohort.

Finally, ischemic stroke may present subtle visual character-

istics, particularly during the acute phase, and two-dimensional slice-level classification does not fully incorporate the three-dimensional anatomical context of the lesion. Future work should evaluate larger and more diverse datasets, patient-wise cross-validation, patient-level classification, external validation, ROC-AUC from retained probabilities, cluster-aware confidence intervals, and comparisons with architectures such as ResNet, EfficientNet, Vision Transformers, and lightweight ConvNeXt variants.

Overall, ConvNeXt-Small separated the selected ischemic and normal-appearing slices in this preliminary experiment. The reported metrics should be treated as hypothesis-generating because the current design does not establish patient-independent or cross-domain generalization. Patient-wise re-evaluation, clinically diverse negative cases, and external validation are required before any screening claim.

VII. Conclusion

This study implemented a ConvNeXt-Small transfer-learning pipeline to distinguish selected CT slices with visible ischemic changes from radiologically normal-appearing slices. Under five-fold slice-level cross-validation, mean accuracy was $92.82\% \pm 1.78\%$ and mean recall was $94.47\% \pm 4.39\%$.

These results are preliminary. Because folds were not separated by patient, correlated slices from one examination may have appeared in both model-development and test data, potentially inflating the reported performance. The experiment also used only normal-appearing negative slices and did not assess hemorrhage, tumors, chronic infarctions, or other intracranial abnormalities. No patient-level or independent external validation was performed.

Accordingly, the proposed method should be regarded only as a slice-level research classifier for normal-appearing versus visibly ischemic images, not as a clinically validated diagnostic, triage, or screening tool. Before clinical claims can be considered, future work must keep all slices from each examination in the same partition, report examination-level sample sizes and metrics, retain probability outputs for ROC-AUC analysis, calculate cluster-aware confidence intervals, include clinically representative competing abnormalities, and validate the locked model on an external multicenter cohort with different scanners and acquisition protocols.

References

[1] G. . S. Collaborators *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet. Neurology*, vol. 20, no. 10, p. 795, 2021.

[2] S. S. Martin, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, C. M. Baker-Smith, B. Barone Gibbs, A. Z. Beaton, A. K. Boehme, *et al.*, "2024 heart disease and stroke statistics: a report of us and global data from the american heart association," *Circulation*, vol. 149, no. 8, pp. e347–e913, 2024.

[3] C. Aguiar, "Avaliação de acidente vascular cerebral em tomografia computadorizada utilizando algoritmo de otimização de formigas," Master's thesis, IFPB, 2017.

[4] R. C. Rennert, A. R. Wali, J. A. Steinberg, D. R. Santiago-Dieppa, S. E. Olson, J. S. Pannell, and A. A. Khalessi, "Epidemiology, natural history, and clinical presentation of large vessel ischemic stroke," *Neurosurgery*, vol. 85, no. Suppl 1, p. S4, 2019.

[5] M. Wajngarten and G. S. Silva, "Hypertension and stroke: update on treatment," *European Cardiology Review*, vol. 14, no. 2, p. 111, 2019.

[6] A. K. Pandey, K. Senapati, I. K. Argyros, and G. P. Pateel, "Improving vertebral fracture detection in c-spine ct images using bayesian probability-based ensemble learning," *Algorithms*, vol. 18, no. 4, 2025.

[7] J. e. Soun, D. Chow, M. Nagamine, R. Takhtawala, C. Filippi, W. Yu, and P. Chang, "Artificial intelligence and acute stroke imaging," *American Journal of Neuroradiology*, vol. 42, no. 1, pp. 2–11, 2021.

[8] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Development and validation of deep learning algorithms for detection of critical findings in head ct scans," 2018.

[9] S. Orellana-Urzúa, I. Rojas, L. Líbano, and R. Rodrigo, "Pathophysiology of ischemic stroke: role of oxidative stress," *Current pharmaceutical design*, vol. 26, no. 34, pp. 4246–4260, 2020.

[10] B. C. Campbell, D. A. De Silva, M. R. Macleod, S. B. Coutts, L. H. Schwamm, S. M. Davis, and G. A. Donnan, "Ischaemic stroke," *Nature reviews Disease primers*, vol. 5, no. 1, p. 70, 2019.

[11] D. Vu, R. G. González, and P. W. Schaefer, *Conventional MRI and MR angiography of stroke*. Springer, 2006.

[12] B. L. Edlow, S. Hurwitz, and J. A. Edlow, "Diagnosis of dwi-negative acute ischemic stroke: a meta-analysis," *Neurology*, vol. 89, no. 3, pp. 256–262, 2017.

[13] S. Patil, R. Rossi, D. Jabrah, and K. Doyle, "Detection, diagnosis and treatment of acute ischemic stroke: current and future perspectives," *Frontiers in medical technology*, vol. 4, p. 748949, 2022.

[14] I. Hadji and R. P. Wildes, "What do we understand about convolutional networks?," 2018.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[17] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

[20] A. Todi, N. Narula, M. Sharma, and U. Gupta, "Convnext: A contemporary architecture for convolutional neural networks for image classification," in *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pp. 1–6, 2023.

[21] B. Guo, Z. Qiao, N. Zhang, Y. Wang, F. Wu, and Q. Peng, "Attention-based convnext with a parallel multiscale dilated convolution residual module for fault diagnosis of rotating machinery," *Expert Systems with Applications*, vol. 249, p. 123764, 2024.

[22] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 129–152, 2023.

[23] M. Yaseen, M. Ali, S. Ali, A. Hussain, M.-I. Joo, and H.-C. Kim, "Cervical spine fracture detection and classification using two-stage deep learning methodology," *IEEE Access*, vol. 12, pp. 72131–72142, 2024.

[24] S. Nazari-Farsani, Y. Yu, R. D. Armindo, M. Lansberg, D. S. Liebeskind, G. Albers, S. Christensen, C. S. Levin, and G. Zaharchuk, "Predicting final ischemic stroke lesions from initial diffusion-weighted images using a deep neural network," *NeuroImage: Clinical*, vol. 37, p. 103278, 2023.

[25] R. Singh, N. Sharma, and H. Gupta, "Blood stroke classification using proposed cnn model," in *2023 World Conference on Communication & Computing (WCONF)*, pp. 1–5, IEEE, 2023.

[26] C.-L. Chin, B.-J. Lin, G.-R. Wu, T.-C. Weng, C.-S. Yang, R.-C. Su, and Y.-J. Pan, "An automated early ischemic stroke detection system using cnn deep learning algorithm," in *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pp. 368–372, 2017.

A Deep Learning-Based Architecture for Automated Nuclei Quantification in Dense Microscopy Images from Transwell Assays

Evellyn de Abreu
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0009-8709-1542

Ana Luiza Francisco Rocha
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0008-2235-1107

Guilherme Vieira de Faria
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0004-0933-5189

Belissa Freitas Barbosa
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-4611-2844

Eloisa Amália Vieira Ferro
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-7447-2583

Ana Claudia Patrocinio
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Pedro Cunha Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

***Abstract*—Analysis of cell migration and invasion through Transwell membrane assays is fundamental to understanding the mechanisms involved in infections caused by *Trypanosoma cruzi* and *Toxoplasma gondii*. However, conventional evaluation of these assays relies on manual cell counting, a time-consuming process subject to observer variability. In this context, this study investigates the behavior of the generalist deep learning model Cellpose on dense Transwell microscopy images for instance segmentation and nucleus quantification. The computational dataset comprised 18 original optical microscopy images associated with manual binary annotations. Image preprocessing included grayscale conversion, Gaussian filtering, and Contrast-Limited Adaptive Histogram Equalization (CLAHE). Because the original annotations were binary, they were converted into instance masks using a marker-controlled watershed procedure. An automatic cell-diameter estimation protocol was applied, and non-overlapping 256 $\times$ 256 pixel patches were generated to increase the number of computational samples available for training while acknowledging the limited biological variability of the original dataset. Fine-tuning of the pre-trained *cyto3* architecture was performed across ten training runs to assess the consistency of model performance. To contextualize these results, the fine-tuned model was additionally compared against the non-fine-tuned *cyto3* network and a classical watershed baseline, and instance-level agreement was further characterized using object-to-object correspondence, Panoptic Quality, and Bland–Altman limits of agreement.The pipeline showed consistent pixel-level spatial overlap performance across runs, achieving a mean Dice coefficient of 0.718 $\pm$ 0.007, Intersection over Union (IoU) of 0.589 $\pm$ 0.007, precision of 0.875 $\pm$ 0.011, and recall of 0.636 $\pm$ 0.010, showing higher values for these pixel-level metrics than the non-fine-tuned Cellpose model and the watershed baseline. However, instance-level evaluation revealed a mean percentage counting error of 47.70% $\pm$ 1.01%, with a Bland–Altman analysis indicating a systematic tendency toward undercounting, particularly in high-density regions. These findings indicate that consistent spatial localization did not translate into reliable separation and quantification of individual nuclei. These results characterize the current Cellpose-based pipeline as a reproducible baseline for investigating automated segmentation in dense Transwell microscopy images while highlighting important limitations in instance separation and counting. Therefore, the current approach is not yet suitable as an unrestricted replacement for manual nucleus quantification.**

***Index Terms*—Cell motility, Deep learning, Instance segmentation, Optical microscopy, Transwell assay**

I. Introduction

Chagas disease, caused by the protozoan *Trypanosoma cruzi*, and toxoplasmosis, caused by *Toxoplasma gondii*, are still major public health concerns today. These two diseases are not only prevalent among vulnerable individuals, but also can be vertically transmitted, occasionally causing pregnancy complications and sequelae in infants [1], [2]. According to the World Health Organization (WHO), millions of people are currently infected with *T. cruzi* worldwide, but the situation is worse in Latin American countries [1]. The transmission pathway of toxoplasmosis is somewhat similar, since the infection acquired by a pregnant woman can cross to her fetus and cause damage to the eyes or nervous system ranging from mild to severe [2]. Therefore, to study how these infections

unfold at the cellular level, researchers rely on in vitro models. Transwell assays, which have emerged as an improvement to the earlier developed Boyden chamber technique, are widely used for this investigation because they provide an opportunity for the migration of cells through a porous membrane and subsequent fixation, staining, imaging, and quantification of the results [3].

Usually, researchers count cells manually, which is time-consuming. To improve the outcomes of medication and techniques related to these protozoans, this study focuses on developing and evaluating an automated counting methodology based on a generalist deep learning model, investigating its behavior and reproducibility when applied to the segmentation of Transwell assay cells. Since this method relies on automated counting, it requires instance segmentation—treating each cell as its own distinct object rather than just labeling a blob of occupied pixels. This distinction is particularly important in Transwell images, where nuclei tend to cluster close together, cells overlap, and the membrane's pores can be mistaken for cell boundaries or vice versa. Cellpose offers a practical way around this. It is a generalist deep-learning model built for cell segmentation, and instead of just outputting a mask, it predicts flow fields that are then used to reconstruct individual cell instances even when objects are touching [4], [5]. This capacity to generalize across acquisition conditions and cell types, previously demonstrated on multiple microscopy modalities [4], [5], motivated its selection over classical pixel-grouping techniques for the dense, low-contrast Transwell images considered here. Building on this tool, this study aims to develop and evaluate a Cellpose-based pipeline for segmenting and automatically counting cell nuclei in optical microscopy images from Transwell assays. The focus throughout was threefold: making sure the training process was reproducible, measuring performance quantitatively rather than just qualitatively, and being transparent about where the method still falls short in a real experimental routine. In addition to the pixel-level metrics originally considered, the evaluation was further extended to include object-to-object correspondence, split and merge errors, Bland–Altman count agreement, and comparisons against a non-fine-tuned Cellpose model and a classical watershed baseline, together with an exploratory analysis of how cell density, image focus, and membrane pore pattern relate to segmentation error.

II. Materials and Methods

A. Dataset and Annotations

The images used in this study were obtained by optical microscopy of Transwell-based assays. In total, the pipeline processed 18 original images, each paired with its corresponding manual mask. These masks were produced in Napari and were initially simple binary annotations, with background pixels set to zero and cellular regions set to a non-zero value. To avoid any silent mismatch that could invalidate the results, image-mask pairing was performed by matching file names, and this pairing was checked at every stage of the pipeline.

These 18 images are part of a broader experimental set comprising two independent experiments with three biological replicates each, totaling 42 Transwell membranes, from which five fields were randomly acquired per membrane. Samples were stained with the Panotic Rapid Staining Kit (Laborclin, Pinhais, PR, Brazil) and imaged on an Opton light microscope using a $40\times$ objective and a total magnification of $400\times$, with the same acquisition protocol and settings applied to every image. Regarding the manual-reference protocol, the ground-truth annotations were independently performed using the Napari software by trained bioscientists from our team who were blinded to the deep learning model outputs. Annotations strictly followed morphological criteria to delineate nuclear boundaries based on characteristics such as nuclear shape, size, staining intensity, visible internal structures, and the presence of cytoplasm surrounding the nucleus Although the broader dataset also included uninfected controls, *T. cruzi*-infected controls, and infected groups treated with different compounds, the available file identifiers did not allow images to be unambiguously traced back to their originating membrane or biological replicate; consequently, infection condition was not used as a stratification variable in the present analysis, and the original image—rather than the membrane or replicate—was adopted as the primary unit for data splitting, in order to minimize the risk of information leakage between training, validation, and test sets.

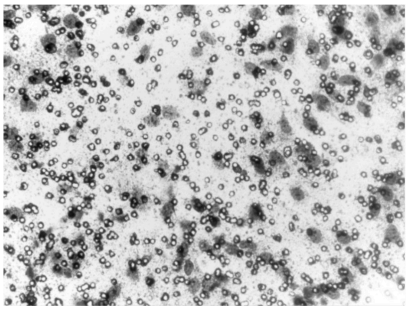
B. Image Preprocessing

Each image was converted to grayscale and normalized to the 8-bit range; when a multichannel image was read in, only its first channel was kept as the model input. To smooth out small acquisition artifacts, a 3×3 Gaussian filter was applied. To improve local contrast, the images were processed through Contrast-Limited Adaptive Histogram Equalization (CLAHE), which is well suited to microscopy images with uneven illumination since it boosts visibility in dim regions without amplifying noise in areas that are already fairly uniform [6].

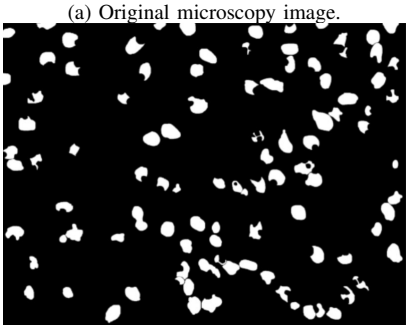
C. Converting Binary Masks into Instance Masks

Because the original annotations were binary, it was impossible to distinguish between two touching nuclei. The original microscopy image, the binary mask manually produced in Napari, and the resulting instance mask after watershed processing are shown in Fig. 1(a), Fig. 1(b), and Fig. 1(c), respectively. In the binary annotation, the code interpreted connected Regions of Interest (ROIs) as a single cell, even when they represented two or more connected foreground regions. To address this limitation and make instance-segmentation training possible, each binary mask was converted into a labeled instance mask using a Euclidean distance transform followed by watershed.

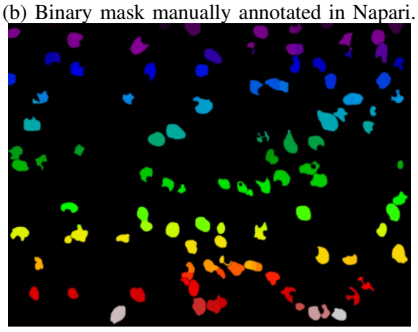
For the watershed step, markers were placed at the local maxima of the distance map, with a minimum peak-to-peak distance of 8 pixels. This classic flooding-based approach can help separate two objects when a visible constriction is present between them [7]. Once watershed finished, every resulting



(a) Original microscopy image.



(b) Binary mask manually annotated in Napari.



(c) Instance mask generated after watershed processing.

Fig. 1. Methodological workflow for converting binary annotations into individual instance masks prior to deep learning training: (a) raw optical microscopy image acquired from a Transwell membrane migration assay; (b) binary foreground reference mask manually traced using the Napari software; and (c) processed instance mask generated via Euclidean distance transform and marker-controlled watershed flooding to isolate overlapping cellular structures.

cell instance was assigned its own sequential label. In addition, objects smaller than 30 pixels were discarded, since they fall outside the scope of what this study intends to count. This threshold approximates the typical area of the Transwell membrane's pores, structures whose shape and local contrast can resemble small nuclei; excluding them was intended to reduce membrane-related false positives rather than to discard genuinely small cells. Because pore appearance varies across fields and cannot be fully anticipated a priori, its potential influence on segmentation performance was further examined as part of the exploratory stratified analysis described in Section II-L and reported in Section III.

D. Estimating Cell Diameter and Extracting Patches

To improve network accuracy, the cell diameter was estimated directly from the masks used for training. For each labeled instance, an equivalent diameter was computed from its area using

$$d = 2\sqrt{\frac{\text{area}}{\pi}}, \tag{1}$$

and the median of these values across all training instances was used as Cellpose's diameter parameter, which was set to 49.57 pixels.

Additionally, because the dataset was limited to 18 images, a data augmentation strategy based on spatial cropping was applied; the images and masks were cropped into non-overlapping 256x256 pixel patches using a stride equal to the patch size. Only patches containing at least three cell instances and at least 200 mask pixels were retained; the remaining patches were discarded as too sparse for training. This filtering resulted in 627 valid patches, while 237 patches were discarded for not meeting the minimum requirements.

E. Data Split

The valid patches were split into three sets: 70% for training, 15% for validation, and 15% for testing. The splitting procedure was performed at the original-image level rather than patch-wise, so that all patches from one original image belonged to the same set. A random seed of 42 was used to ensure reproducibility. This resulted in 428 training patches, 128 validation patches, and 71 test patches.

F. Cellpose Model and Training

The segmentation process utilized Cellpose, where the pre-trained *cyto3* model was used as the starting point rather than training from scratch [4], [5]. The dataset consisted of 8-bit image files labeled with instance masks; mask files were denoted by adding "_masks.tif" to their name. Training was performed through the Python interface of Cellpose for 100 epochs using a batch size of 4, a learning rate of 0.005, a weight decay of 2×10^{-3}, and `min_train_masks = 1`. To regulate epoch transitions within the Cellpose framework, an internal stochastic sampling parameter was set to draw 100 training patches and 40 validation patches per epoch from the available data pool, allowing the optimization algorithm to process uniform data increments independently of the total partition sizes. For inference, the trained model was executed on the test set using the estimated median diameter and the default `[0,0]` grayscale channel configuration. All computational training and inference procedures were executed on a workstation equipped with an Intel Core i7-9750H CPU @ 2.60 GHz, 16 GB of RAM, and an NVIDIA GeForce GTX 1660 Ti (6 GB) GPU.

All hyperparameters above were defined *a priori* from values reported in the literature for fine-tuning instance-segmentation networks on microscopy datasets of limited size, and were not adjusted based on performance on the validation or test sets, so as to avoid parameter-selection bias. In particular, the learning rate and batch size adopted here are of the same order of magnitude as values used for fine-tuning cell instance-segmentation architectures under variable acquisition conditions [18], while the 100-epoch training schedule follows

fine-tuning protocols previously reported for Cellpose itself [5].

G. Compared Methods

To contextualize the performance of the fine-tuned pipeline, two additional baselines were considered: the pre-trained *cyto3* model without any Transwell-specific fine-tuning, and a classical marker-controlled watershed segmentation applied directly to the preprocessed grayscale images, without any deep-learning component. This comparison is intended to characterize the gain associated with domain adaptation and to provide a reference against a conventional image-processing technique, rather than to serve as an exhaustive benchmark against other instance-segmentation architectures.

H. Evaluation Metrics

Performance was evaluated at two complementary levels: pixel-level overlap metrics and instance-level counting errors. Pixel-level metrics were used to quantify how well the predicted cellular regions overlapped with the reference masks, regardless of whether each nucleus was correctly separated as an individual object. For this analysis, both the reference and predicted masks were binarized, so that any non-zero pixel was considered part of a cellular region.

After binarization, true positives (TP) were defined as pixels correctly classified as cell, false positives (FP) as pixels predicted as cell but absent from the reference mask, and false negatives (FN) as reference-cell pixels missed by the model. Based on these values, the Dice coefficient was calculated according to (2). Dice measures the overlap between the predicted and reference cellular regions, with higher values indicating better agreement between the two masks [8].

$$\mathrm{Dice} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

The Intersection over Union (IoU), also known as the Jaccard index, was calculated according to (3). This metric represents the ratio between the intersection and the union of the predicted and reference cellular areas, providing a stricter overlap measure than Dice [9].

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

Precision was calculated according to (4) and indicates the proportion of predicted-cell pixels that actually corresponded to cellular regions in the reference mask. Therefore, precision is useful for identifying the extent to which the model incorrectly labels background regions as cells.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

Recall was calculated according to (5) and represents the proportion of reference-cell pixels correctly detected by the model. In this context, recall is particularly important because low recall indicates that part of the true cellular area was missed during segmentation.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

Together, Dice, IoU, precision, and recall provide complementary information about segmentation quality. Dice and IoU evaluate spatial agreement between masks, while precision and recall help distinguish between false-positive and false-negative segmentation behavior [10].

In addition to pixel-level overlap, instance-level counting errors were calculated to evaluate whether the model could correctly estimate the number of individual nuclei. This analysis was necessary because good pixel-level overlap does not necessarily imply accurate cell counting, especially when adjacent nuclei are merged into a single predicted instance.

The absolute counting error was calculated according to (6), where N_{pred} is the number of predicted instances and N_{ref} is the number of reference instances.

$$E_{\mathrm{abs}} = |N_{\mathrm{pred}} - N_{\mathrm{ref}}| \tag{6}$$

The percentage counting error was calculated according to (7). This metric expresses the counting error relative to the reference number of nuclei, allowing comparison across patches with different cell densities.

$$E_{\mathrm{perc}} = 100 \times \frac{|N_{\mathrm{pred}} - N_{\mathrm{ref}}|}{N_{\mathrm{ref}}} \tag{7}$$

I. Object-to-Object Evaluation

In addition to pixel-level overlap and the raw counting error above, one-to-one correspondence between reference and predicted instances was established by maximizing IoU between objects. A match was considered a true positive when

$$\mathrm{IoU}_{\mathrm{object}} \geq 0.50. \tag{8}$$

From these matches, instance-level precision and recall were computed, together with Segmentation Quality (SQ), Recognition Quality (RQ), and Panoptic Quality (PQ), defined as

$$PQ = SQ \times RQ. \tag{9}$$

Object Dice was computed between correctly matched instance pairs, providing a measure of geometric agreement restricted to the objects the model actually detected.

J. Split and Merge Errors and Count Bias

To characterize structural counting errors beyond the aggregate E_{abs} and E_{perc}, split and merge events were also identified using a minimum-overlap criterion of 10% of an object's area. A split error was recorded when a single reference instance overlapped meaningfully with two or more predicted instances, and a merge error was recorded when a single predicted instance overlapped meaningfully with two or more reference instances. The structural split/merge analysis reported below was performed on the 71 stored prediction masks from Run 10, which could be directly matched to

the corresponding reference instance masks. Therefore, these structural-error counts are reported as a run-specific diagnostic and are not interpreted as an across-run average. Count bias was further defined as

$$\text{Bias} = N_{\text{pred}} - N_{\text{ref}}, \tag{10}$$

with negative values indicating a tendency toward undercounting and positive values indicating overcounting.

K. Bland–Altman Analysis

Agreement between automated and reference counts was further examined using the Bland–Altman method. For each patch i, the difference

$$d_i = N_{\text{pred},i} - N_{\text{ref},i} \tag{11}$$

was computed, and the 95% limits of agreement were obtained as

$$\text{LoA} = \bar{d} \pm 1.96\, s_d, \tag{12}$$

where s_d is the standard deviation of the differences.

L. Density, Focus, and Pore Pattern Analyses

To investigate whether segmentation difficulty was uniformly distributed across images, three exploratory stratified analyses were performed on the test set. Cell density was estimated from the number of reference instances per patch, and patches were split into low-, medium-, and high-density tertiles. Relative focus quality was estimated from the variance of the Laplacian of the original image prior to CLAHE and Gaussian filtering, with higher values representing greater high-frequency content and, consequently, sharper images; patches were similarly split into low-, medium-, and high-focus tertiles. Finally, because no dedicated manual pore annotations were available, the influence of the membrane pore pattern was assessed through an automated proxy that quantified the number, density, and geometric characteristics of pore-like regions; this variable is treated strictly as an exploratory descriptor of membrane structure rather than as validated pore segmentation. As noted in Section II-L, infection condition could not be used as an additional stratification variable due to the lack of unambiguous replicate identifiers in the available files.

Taking all of these complementary metrics into consideration—spatial, instance-level, and error-structure—the evaluation combined pixel-level segmentation, object-level correspondence, and counting-based analysis. This hybrid approach was essential because the pipeline's ultimate goal was not only to isolate cellular regions but also to support the automated quantification of individual cell nuclei, and because the extent of that limitation could itself depend on image- and patch-level characteristics such as density, focus, and pore pattern.



(a) Original patch. (b) Reference mask. (c) Model prediction.

Fig. 2. Representative qualitative evaluation of the trained Cellpose framework on a 256 × 256 pixel testing patch: (a) original microscopy crop displaying target nuclei as dark circular structures against a heterogeneous background; (b) expert-annotated reference mask defining ground truth boundaries; and (c) automated instance boundaries predicted by the fine-tuned *cyto3* model, showing areas of spatial agreement alongside local edge discrepancies in dense clusters.

III. Results

Fig. 2 shows a qualitative example of the segmentation obtained for a 256×256 pixel patch. The original microscopy image is presented in Fig. 2(a), the manually annotated reference mask is shown in Fig. 2(b), and the model prediction is shown in Fig. 2(c). In the original patch, the cell nuclei appear as dark circular structures over a heterogeneous background, while the reference mask highlights the nuclei selected as annotated regions. The prediction shows that the model was able to detect several cellular regions with spatial coherence. However, some discrepancies can still be observed, especially in areas with low contrast, clustered structures, or regions with visual characteristics similar to nuclei. This qualitative example is consistent with the quantitative findings, showing spatial localization of several cellular regions while also illustrating limitations in precise instance-level counting.

Table I presents the pixel-wise segmentation performance across ten independent runs of the experiment. To examine between-run variability and training repeatability, the fine-tuning procedure was independently restarted from the pre-trained *cyto3* initialization at the beginning of each run. Across the ten runs, relatively small differences were observed in the global mean values for Dice (0.718 ± 0.007), IoU (0.589 ± 0.007), precision (0.875 ± 0.011), and recall (0.636 ± 0.010). This low between-run variability suggests that, under the evaluated training conditions, the reported pixel-level performance was not strongly dependent on a particular run.

The ten independent training runs resulted in the counting errors shown in Table II. The absolute counting error remained stable from run to run, varying between a minimum value of 2.72 cells/patch and a maximum value of 2.89 cells/patch, with an overall mean value of 2.79 ± 0.06 cells/patch. The percentage counting error also showed little variability across the ten training runs, ranging from 46.06% to 49.17%, with a mean value of $47.70\% \pm 1.01\%$. This indicates that the counting error showed relatively low variability across the evaluated training runs. However, the magnitude of the percentage error indicates an important limitation of the proposed pipeline. In other words, even though the model produced consistent results among training runs, it could not accurately quantify the number of individual cell nuclei. This suggests that missed

TABLE I
PIXEL-LEVEL SPATIAL OVERLAP PERFORMANCE ACROSS TEN INDEPENDENT TRAINING RUNS EVALUATED ON THE TESTING DATASET

Run	Dice (Mean ± Std)	IoU (Mean ± Std)	Precision (Mean ± Std)	Recall (Mean ± Std)
1	0.713 ± 0.187	0.582 ± 0.199	0.872 ± 0.182	0.628 ± 0.210
2	0.716 ± 0.200	0.589 ± 0.209	0.863 ± 0.196	0.634 ± 0.217
3	0.718 ± 0.191	0.589 ± 0.201	0.873 ± 0.188	0.634 ± 0.211
4	0.721 ± 0.182	0.590 ± 0.195	0.871 ± 0.183	0.642 ± 0.203
5	0.714 ± 0.184	0.583 ± 0.198	0.881 ± 0.173	0.627 ± 0.208
6	0.718 ± 0.202	0.593 ± 0.212	0.874 ± 0.188	0.637 ± 0.220
7	0.732 ± 0.183	0.605 ± 0.196	0.902 ± 0.168	0.655 ± 0.209
8	0.706 ± 0.212	0.579 ± 0.217	0.864 ± 0.211	0.621 ± 0.228
9	0.722 ± 0.187	0.594 ± 0.200	0.872 ± 0.176	0.645 ± 0.211
10	0.718 ± 0.183	0.587 ± 0.195	0.875 ± 0.174	0.634 ± 0.205
Mean ± Std	$\mathbf{0.718 \pm 0.007}$	$\mathbf{0.589 \pm 0.007}$	$\mathbf{0.875 \pm 0.011}$	$\mathbf{0.636 \pm 0.010}$

Note: Each run represents an independent fine-tuning lifecycle restarted from the pre-trained *cyto3* initialization. Values within each individual run row express the mean performance across all test patches (± standard deviation reflecting spatial and biological image variability). The final "Mean ± Std" row presents the global average across the ten runs (± standard deviation reflecting between-run variability).

TABLE II
QUANTITATIVE INSTANCE-BASED CELL COUNTING ERRORS AND DEVIATIONS ACROSS TEN INDEPENDENT TRAINING RUNS

Run	Absolute Counting Error (Mean ± Std)	Percentage Counting Error (%) (Mean ± Std)
1	2.89 ± 1.98	$49.17\% \pm 19.16\%$
2	2.72 ± 1.87	$47.03\% \pm 20.80\%$
3	2.83 ± 1.88	$48.71\% \pm 19.22\%$
4	2.73 ± 1.93	$46.70\% \pm 21.66\%$
5	2.85 ± 1.98	$48.57\% \pm 20.71\%$
6	2.77 ± 1.88	$47.90\% \pm 21.38\%$
7	2.73 ± 2.00	$46.75\% \pm 21.36\%$
8	2.83 ± 1.99	$48.06\% \pm 22.20\%$
9	2.72 ± 1.97	$46.06\% \pm 22.39\%$
10	2.83 ± 1.97	$48.04\% \pm 22.25\%$
Mean ± Std	$\mathbf{2.79 \pm 0.06}$	$\mathbf{47.70\% \pm 1.01\%}$

Note: Each run represents an independent experimental lifecycle where the model's weights were reset and retrained. Values within each individual run row express the mean counting error across all test patches (± standard deviation reflecting patch-level variability). The final "Mean ± Std" row presents the global average across the ten independent runs (± standard deviation reflecting training reproducibility).

TABLE III
QUANTITATIVE COMPARISON OF THE SEGMENTATION METHODS

Method	Dice	Inst. Prec.	Inst. Rec.	Obj. Dice	PQ	Error (%)
Pre-trained Cellpose	0.165	0.067	0.133	0.321	0.064	103.09
Fine-tuned Cellpose	0.720	0.957	0.552	0.862	0.529	42.11
Watershed	0.467	0.078	0.485	0.705	0.091	600.72

nuclei and merged neighboring instances may contribute substantially to the observed counting error, particularly in densely populated regions of the images.

A. Comparison Across Methods

Table III summarizes a direct comparison, on the same test set, between the fine-tuned Cellpose pipeline described above, the non-fine-tuned *cyto3* model, and classical watershed. Fine-tuned Cellpose achieved a Dice of 0.720, an instance precision of 0.957, and a PQ of 0.529, with higher values for these metrics than both evaluated baselines. Pre-trained Cellpose achieved a Dice of 0.165 and a percentage counting error of 103.09%, suggesting limited suitability of the non-adapted generalist model for this specific image domain. Watershed achieved a pixel-level Dice of 0.467 but a substantially higher percentage counting error of 600.72%, consistent with marked over-segmentation.

TABLE IV
RUN 10 STRUCTURAL INSTANCE ERRORS ON THE 71 TESTING PATCHES

Measure	Value
Testing patches	71
Reference instances	395
Predicted instances	198
Split errors	0
Merge errors	39
Patches with at least one merge	29/71 (40.8%)
Merge errors per patch	0.55 ± 0.79
Run 10 count bias	-2.77 nuclei/patch
Run 10 95% limits of agreement	$[-6.77,\ 1.22]$

B. Instance-Level Metrics and Count Agreement

For the fine-tuned model, instance-level evaluation yielded a precision of 0.957, a recall of 0.552, an Object Dice of 0.862 for correctly matched pairs, SQ of 0.774, RQ of 0.670, and PQ of 0.529. The Object Dice of 0.862 for correctly matched pairs indicates relatively strong geometric agreement among matched instances, while the lower instance recall suggests that an important limitation is the recovery of all nuclei present in the reference.

Bland–Altman analysis indicated a mean bias of -1.98 nuclei per patch for the fine-tuned model, with 95% limits of agreement of $[-4.83,\ 0.86]$, consistent with a systematic tendency toward undercounting. The pre-trained model showed the opposite pattern, with a mean bias of $+3.65$ nuclei per patch (limits of agreement: -3.38 to 10.68), while watershed showed pronounced overcounting, with a mean bias of $+24.90$ objects per patch (limits of agreement: 6.75 to 43.06). Among the three evaluated methods, the fine-tuned model showed closer count agreement with the reference, although a relevant tendency toward underestimating the number of nuclei remained.

C. Split and Merge Error Analysis

A structural error analysis was performed using the 71 stored prediction masks from Run 10, which reproduced the pixel-level and counting results reported in Tables I and II. Across these patches, the reference masks contained 395 labeled nuclei versus 198 predicted instances. No split errors were identified under the 10% minimum-overlap criterion, whereas 39 predicted instances were classified as merge errors, occurring in 29/71 patches (40.8%, 0.55 ± 0.79 merges/patch).

This run's count bias was -2.77 nuclei/patch (95% limits of agreement: $[-6.77,\ 1.22]$), consistent with undercounting driven mainly by fusion of neighboring nuclei rather than missed detections. As split/merge analysis was performed for Run 10 only, these values are reported separately from the across-run and method-comparison statistics.

D. Influence of Cell Density, Focus, and Pore Pattern

Performance decreased as cell density increased: Dice fell from ≈ 0.81 (low density) to ≈ 0.72 (medium) and ≈ 0.63 (high), PQ decreased from ≈ 0.62 to ≈ 0.52 to ≈ 0.45, and

counting error rose from $\approx 31\%$ to $\approx 50\%$, indicating that cellular crowding is an important factor associated with reduced performance. Focus quality also influenced results, with lower Dice (≈ 0.66) and PQ (≈ 0.42) in the least-sharp tertile than in the intermediate tertile (Dice ≈ 0.77, PQ ≈ 0.61), though the relationship was not strictly monotonic (PQ ≈ 0.56 in the sharpest tertile). The exploratory pore-pattern proxy showed only weak correlations with Dice, PQ, and counting error, suggesting its influence depends on factors beyond pore count, such as contrast and morphological similarity to nuclei.

IV. Discussion

Fine-tuning Cellpose produced spatially coherent segmentations of Transwell images. The ten independently restarted fine-tuning runs provided an assessment of between-run variability under the evaluated training conditions. The relatively low standard deviation across runs suggests that the reported metrics were reasonably repeatable in this setting and reduces, without eliminating, concern that the observed performance was specific to a single run.

The higher precision than recall is consistent with a relatively conservative prediction pattern. Precision and recall, originally formalized in the information retrieval literature [11] and now standard across binary classification tasks [12], capture two different kinds of error: precision penalizes false alarms, while recall penalizes missed detections [13]. In this case, the algorithm is reluctant to label background as cell, but that caution comes at a cost—it also misses a meaningful share of genuine cellular regions. Whether that trade-off is acceptable depends on what the pipeline is being used for. For cell counting specifically, however, false negatives and instance fusion are the more damaging failure modes, since both directly lower the estimated cell count.

The main methodological limitation traces back to how the training masks were built. Because the manual annotations started out as binary masks, instance separation had to be done after the fact using watershed on the distance transform [7]. This approach performs adequately when cells are touching but have some visible constriction between them; it is far less reliable when cells overlap heavily or when no clear neck separates adjacent nuclei. This failure mode—neighboring cells being merged, or valid instances being suppressed in crowded regions—is well documented for classical pixel-grouping approaches to nucleus segmentation, and is part of the motivation behind alternative strategies such as StarDist, which predicts a star-convex shape for each nucleus directly instead of separating instances after the fact [14]. In our case, the derived reference mask itself likely already contains some under-segmentation, meaning the model was partly learning from imperfect ground truth.

That limitation goes a long way toward explaining what otherwise looks like a contradiction: moderately strong overlap metrics alongside a high counting error. Dice and IoU [8], [9], [15] are computed at the pixel level, capturing how well the predicted cellular area aligns with the reference. The counting error, by contrast, is an instance-level metric. A model can segment the area of a dense cell cluster correctly, scoring well on Dice, while still treating that cluster as a single cell rather than several. For this particular problem, neither type of metric tells the full story on its own; reading them together is necessary to understand what the model is actually doing well and where it falls short. This is consistent with longstanding guidance in the medical image segmentation literature that metric choice should be matched to the specific needs of the task [15].

The comparison against the two baseline methods reinforces this picture. Adapting Cellpose to the Transwell domain increased Dice from approximately 0.165 to 0.720 and PQ from approximately 0.064 to 0.529, suggesting that the non-adapted generalist architecture has limited performance on these dense, low-contrast images and that domain-specific fine-tuning contributed substantially to the improvement observed under the conditions evaluated here. Watershed showed essentially the opposite failure mode to the fine-tuned network: its comparatively higher pixel-level Dice came with very low instance precision and pronounced overcounting, illustrating that recovering most of the cellular area is not necessarily sufficient when the end goal is accurate quantification.

The Run 10 split/merge analysis provides direct evidence for the structural origin of part of the counting error. Under the predefined overlap criterion, no split events were detected, whereas 39 predicted objects were classified as merges and 40.8% of test patches contained at least one merge. This asymmetry is consistent with the negative Run 10 count bias and supports the interpretation that adjacent nuclei were frequently fused into single predicted instances. Importantly, these values characterize the stored predictions from one training run and should not be interpreted as mean structural-error rates across all ten runs.

The stratified analysis by cell density further suggests that crowding is an important contributor to counting error, although image acquisition characteristics may also play a role. Performance declined steadily from low- to high-density patches, both in overlap metrics and in PQ, while the percentage counting error nearly doubled between the least and most crowded groups. Focus quality also modulated performance, though less consistently, and the exploratory pore-pattern proxy showed no strong association with segmentation error, suggesting that pore-related false positives are better explained by local contrast and morphology than by pore count alone.

Taken together with the ten-run variability analysis presented in Section III, these findings suggest that the fine-tuned pipeline showed relatively stable performance across the evaluated runs while retaining important limitations in separating touching nuclei. Thus, low between-run variability should not be interpreted as evidence of adequate counting performance; both repeatability and instance-level error need to be considered when assessing the practical usefulness of the pipeline.

From a practical standpoint, the pipeline provides a potential foundation for further investigation of automated Transwell image analysis—it standardizes preprocessing, docu-

ments training parameters explicitly, and may reduce some of the variability associated with manual segmentation. That said, treating it as a full replacement for manual counting would be premature while the percentage counting error remains around 48%, particularly in densely populated fields. Future iterations could also leverage built-in restoration modules such as those introduced in Cellpose3 [16] to mitigate the focus- and contrast-related errors identified above.

A methodological limitation that should be made explicit concerns the experimental unit underlying these analyses. Although all patches derived from the same image were kept within the same data split, the available metadata did not allow images to be unambiguously traced back to their originating membrane or biological replicate. Patches—and, to a lesser extent, the 18 original images themselves—should therefore not be treated as fully independent biological observations, and the density-, focus-, and pore-stratified analyses in particular should be regarded as exploratory rather than confirmatory.

V. Conclusion

This paper describes the creation and evaluation of a Cellpose-based pipeline for cell nucleus segmentation and automatic counting in optical microscopy images from Transwell assays. The proposed approach comprises conversion of binary masks into instance masks, automatic estimation of cell diameter, patch generation, image-level data splitting, and fine-tuning of the pre-trained *cyto3* model.

Throughout all ten training sessions, segmentation quality was consistent, with mean Dice of 0.718 ± 0.007 and IoU of 0.589 ± 0.007. Given the precision, relatively low recall, and $47.70\% \pm 1.01\%$ counting error, these results indicate that the pipeline still has substantial difficulty reliably detecting and separating individual cell nuclei, particularly in densely packed areas. The run-specific structural analysis further showed 39 merge errors and no detected split errors in the stored Run 10 predictions, reinforcing instance fusion as an important source of undercounting. When directly compared against a non-fine-tuned Cellpose model and a classical watershed baseline, the fine-tuned pipeline showed higher overlap (Dice of 0.720 versus 0.165 and 0.467, respectively), instance precision (0.957), and Panoptic Quality (0.529) than the two evaluated baselines, while still showing a systematic tendency toward undercounting, particularly in high-density regions, where the percentage counting error approached 50%.

Overall, the proposed pipeline provides an experimental framework for investigating automated Transwell image analysis and may help reduce the amount of manual processing required. Under the evaluated conditions, it showed better performance on several reported metrics than the non-adapted generalist model and the classical image-processing baseline. However, it cannot yet be viewed as a replacement for manual counting due to the noted counting error and its dependence on cell density. Future work should focus on generating manual annotations in instance format directly, particularly for densely packed areas; documenting membrane and biological-replicate identifiers so that data can be split and analyzed at the appropriate experimental unit, including stratification by infection condition; benchmarking against other instance-segmentation architectures, such as StarDist and U-Net-based approaches [17]; and validating the pipeline on larger, multicenter datasets acquired under different imaging conditions.

Acknowledgment

The authors are thankful for the received financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Brazil), CNPq—Brazil, and FAPEMIG—Brazil, as well as the scientific and technical support provided by the Laboratório de Imunofisiologia da Reprodução (LABIFIR).

References

[1] World Health Organization, "Chagas disease (American trypanosomiasis)," WHO Fact Sheets, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)

[2] Centers for Disease Control and Prevention, "Symptoms of toxoplasmosis," CDC Toxoplasmosis, 2024. [Online]. Available: https://www.cdc.gov/toxoplasmosis/symptoms/index.html

[3] C. R. Justus, N. Leffler, M. Ruiz-Echevarria, and L. V. Yang, "In vitro cell migration and invasion assays," *Journal of Visualized Experiments*, no. 88, e51046, 2014.

[4] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, 2021.

[5] M. Pachitariu and C. Stringer, "Cellpose 2.0: how to train your own model," *Nature Methods*, vol. 19, pp. 1634–1641, 2022.

[6] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[7] S. Beucher and C. Lantéjuoul, "Use of watersheds in contour detection," in *Proc. Int. Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation*, Rennes, France, 1979, pp. 2.1–2.12.

[8] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[9] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[10] J. C. Caicedo et al., "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[11] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London, U.K.: Butterworths, 1979.

[12] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[13] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[14] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Granada, Spain, 2018, pp. 265–273.

[15] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, art. 29, 2015.

[16] C. Stringer and M. Pachitariu, "Cellpose3: one-click image restoration for improved cellular segmentation," *Nature Methods*, vol. 22, no. 3, pp. 592–599, 2025.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, 2015, pp. 234–241.

[18] B. Peng, J. Chen, P. B. Githinji, I. Gul, Q. Ye, M. Chen, P. Qin, X. Huang, C. Yan, D. Yu, J. Ji, and Z. Chen, "Practical guidelines for cell segmentation models under optical aberrations in microscopy," arXiv:2404.08549, 2024.

A Dual-Stage Framework for Mammographic Nodule Analysis: Integrating Decision Tree Classification and Marker-Controlled Watershed

Patrick Alves Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0001-0960-2003

Bruno Augusto Machado Goes
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-6815-5525

Ana Claudia Patrocinio
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Pedro Cunha Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

***Abstract*—Breast cancer represents a major public health concern, and mammography is one of the main tools used for its early detection. However, the presence of noise, low contrast, and morphological variations in lesions can hinder the automatic identification and delineation of breast nodules. In this context, this study investigates an interpretable computational framework for the classification of preselected mammographic regions of interest and the subsequent delineation of nodule regions, using a decision tree, preprocessing steps, Chan-Vese active contours, and marker-controlled Watershed Transform. The method was evaluated using 226 regions of interest extracted from the INbreast database, including 113 ROIs with nodules and 113 ROIs without nodules. Initially, the ROIs were classified based on intensity and texture metrics, such as mean, median, variance, percentiles, entropy, and GLCM features. Subsequently, the images classified as containing nodules were subjected to normalization, anisotropic diffusion filtering, contrast stretching, definition of internal and external markers, and Watershed-based segmentation. The classification stage achieved a mean accuracy of 79.96%, sensitivity of 73.45%, and specificity of 86.48%. In the segmentation stage, the method obtained a mean Dice coefficient of 0.73, a mean eccentricity difference of 0.13 compared with the gold standard, and a higher concentration of area percentage errors within the 0–20% range.**

***Index Terms*—Breast Cancer; Classification; Image Preprocessing; Region of Interest; Segmentation.**

I. Introduction

Breast cancer is a condition characterized by the uncontrolled growth and division of abnormal cells in breast tissue, which may compromise the normal organization of the mammary tissue and interfere with the proper functioning of the organism [1]. Due to this biological behavior and its high incidence in the population, breast cancer represents a major global public health concern [2], [3].

According to the International Agency for Research on Cancer (IARC), this disease accounted for 2.3 million new cases in 2022 and remains the leading malignancy among women worldwide. Recent projections indicate that the annual incidence may escalate to 3.2 million cases and 1.1 million deaths by 2050, underscoring the critical importance of population-based screening, early diagnosis, and optimized therapeutic workflows [2], [3].

In this context, mammography emerges as an important tool for the early detection of breast cancer, as it is capable of identifying lesions even before they become palpable, thereby increasing the chances of successful treatment. However, the interpretation of mammographic images remains challenging, particularly due to the presence of noise, low contrast, and overlapping anatomical structures. These factors, which are associated with diffuse radiation and the composition of breast tissue, may compromise lesion identification and contribute to misdiagnosis rates ranging from 10% to 30% [4], [5].

To mitigate these limitations, computer-aided diagnosis (CAD) systems have been developed to support the automatic detection of nodules in mammograms, serving as auxiliary tools for screening and radiological analysis. Thus, two stages are particularly important: the classification of regions of interest (ROIs), which enables the distinction between images with and without nodules, and segmentation, which is responsible for delineating the area corresponding to the lesion [6].

Segmentation in image processing consists of dividing an image into distinct regions in order to isolate the region of interest, representing an important step for the identification of breast lesions in mammograms [7]. However, to improve the accuracy of this process, preprocessing steps such as filtering, intensity-level adjustment, contrast enhancement, and noise reduction are required to improve image quality and minimize the influence of structures that do not belong to the region of interest, such as pectoral muscles and undesired artifacts [8]. When segmentation is applied directly to raw images, there is a higher risk of inaccurate detection and excessive

sensitivity during segmentation, especially in noisy or low-contrast images [9].

Among the methods used for mammographic image segmentation are region-based approaches, such as marker-controlled segmentation, partial region growing, rough set theory, and Watershed [10]. Between these methods, the Watershed Transform stands out for enabling the separation of image regions based on intensity or gradient variations; however, its direct application may lead to excessive sensitivity. Therefore, marker-controlled Watershed uses internal and external masks to guide the segmentation process and assist in defining the boundaries of the region of interest, contributing to a more accurate delineation of lesions in medical images [11], [12].

Motivated by the challenges associated with the characterization and delineation of nodules in mammographic images, this study investigates an interpretable dual-stage computational framework for the analysis of preselected regions of interest (ROIs). The proposed method integrates texture-based decision tree classification to filter healthy tissue from abnormal regions of interest (ROIs), followed by a hybrid segmentation pipeline combining anisotropic diffusion, Chan-Vese active contours, and marker-controlled Watershed transformation. Validated against multi-size ground truth annotations from the INbreast database across a 50-fold resampling workflow, this architecture aims to support the delineation and morphological analysis of mammographic nodules.

II. MATERIALS AND METHODS

The algorithm was developed in MATLAB software, under a license no 41224224 [13], and structured into two main stages: ROI classification and segmentation of the regions classified as containing nodules. The flowchart presented in Fig. 1 illustrates these stages. This section describes each stage of the method in detail, as well as the procedures used for the quantitative evaluation of the algorithm results.

A. Image Database

The computational method was tested on a dataset composed of 226 12-bit regions of interest, of which 113 contained nodules and 113 did not. These ROIs were extracted from 222 mammograms from the INbreast database [14]. The images were obtained from 98 distinct patients, including cases with one or more nodules and cases without nodules. Among the different types of breast lesions, such as nodules and microcalcifications, this study focused on the classification of preselected ROIs and the subsequent delineation of nodule regions.

Based on this objective, the regions of interest from mammograms containing nodules were selected to centralize the structure to be segmented, using the gold standard as reference. Since the images contained nodules with a wide variety of margins and sizes, the spatial crops were standardized into four dimensions: 300 × 300, 500 × 500, 700 × 700, and 1000 × 1000 pixels. For normal mammograms, random crops were extracted from internal breast tissue following the same spatial constraints. To ensure structural uniformity across all

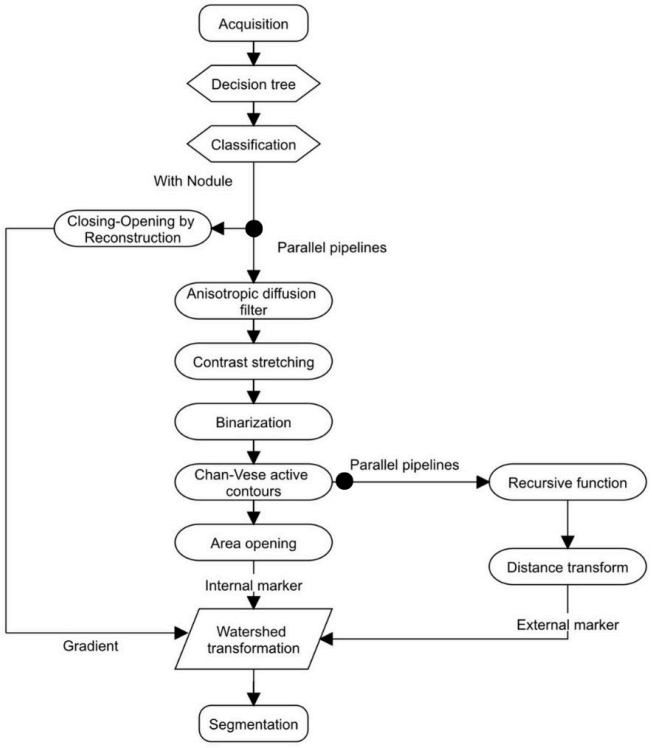


Fig. 1. Computational architecture and block diagram of the proposed dual-stage framework. The flowchart maps the sequential transition from primary image acquisition and texture-based decision tree classification to the hybrid segmentation pipeline, detailing the extraction of internal and external markers used to constrain the final Watershed transformation.

automated operations, all input matrices were resized to a standard dimension of 500 × 500 pixels. The expert manual annotations performed by radiologists served as the experimental ground truth (GT). These binary reference masks underwent the exact cropping and resizing protocols applied to the test sets. Within these matrices, pixels belonging to the nodule region were assigned an intensity value of 1 ($S_{\text{nodule}} = 1$), while background pixels were set to 0 ($S_{\text{background}} = 0$).

The scope of this study was strictly focused on evaluating the feasibility of the proposed segmentation architecture. To mitigate sampling constraints and expose the algorithm to scenarios with greater anatomical variability, the lesions were modeled as patches, represented by isolated regions of interest (ROIs). This strategy increased the number of image samples available for analysis and enabled the proposed framework to be evaluated under a broader range of anatomical configurations.

B. Classification

A decision tree partitions a multi-dimensional feature space into uniform subgroups through a hierarchical sequence of decision rules optimized to maximize class separation [15]. To construct the proposed model, twenty-one candidate descriptors were calculated, including first-order statistical metrics (mean, median, standard deviation, variance, minimum,

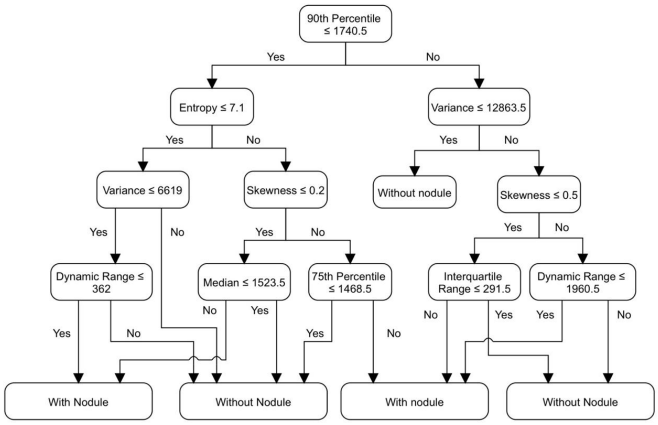


Fig. 2. Hierarchical structure of the trained decision tree classifier for ROI classification. The routing layout illustrates the optimized selection of first-order intensity metrics (percentiles, variance, dynamic range, median, and interquartile range) and second-order textural features (entropy) at each decision node to partition the input vector space into nodule-containing or normal tissue classes.

maximum, dynamic range, 10th, 25th, 75th, and 90th percentiles, interquartile range, skewness, kurtosis, and intensity energy) and gray-level co-occurrence matrix (GLCM) features (entropy, contrast, correlation, energy, and homogeneity).

For classifier construction, the 113 nodule-containing ROIs were kept fixed, whereas a larger pool of 1,130 normal ROIs was generated by extracting multiple crops from mammograms without nodules. Specifically, ten normal-tissue crops were obtained at distinct spatial coordinates for each corresponding sampling configuration, increasing the representation of normal parenchymal variability. Based on this training pool, the optimal hierarchical topology (Fig. 2) selected the 90th percentile, entropy, variance, skewness, dynamic range, median, 75th percentile, and interquartile range (IQR) as the definitive routing nodes to automatically classify the input matrices.

C. Preprocessing

The segmentation process is presented in Fig. 3 through the images obtained at the end of each stage, allowing the visual progression of the method to be observed from the original image to the final segmented result. Initially, the procedure was applied to the images classified as containing nodules, with preprocessing as the first stage. This phase aimed to improve image quality by reducing irrelevant and redundant information, as well as facilitating the interpretation of the regions of interest. Before applying the enhancement techniques, image normalization was performed by adjusting the pixel intensity values to the range from 0 to 1, in order to standardize the input data for processing.

Subsequently, an anisotropic diffusion filter was applied to reduce acquisition noise while preserving sharp structural boundaries. Unlike conventional uniform filters, anisotropic diffusion limits smoothing near edge transitions by evaluating local image gradients [16]. In this pipeline, the filter was executed on the normalized input matrix (Fig. 3a) using five iterations and maximum 8-neighbor connectivity, optimizing the parenchymal image quality before thresholding (Fig. 3b) [17].

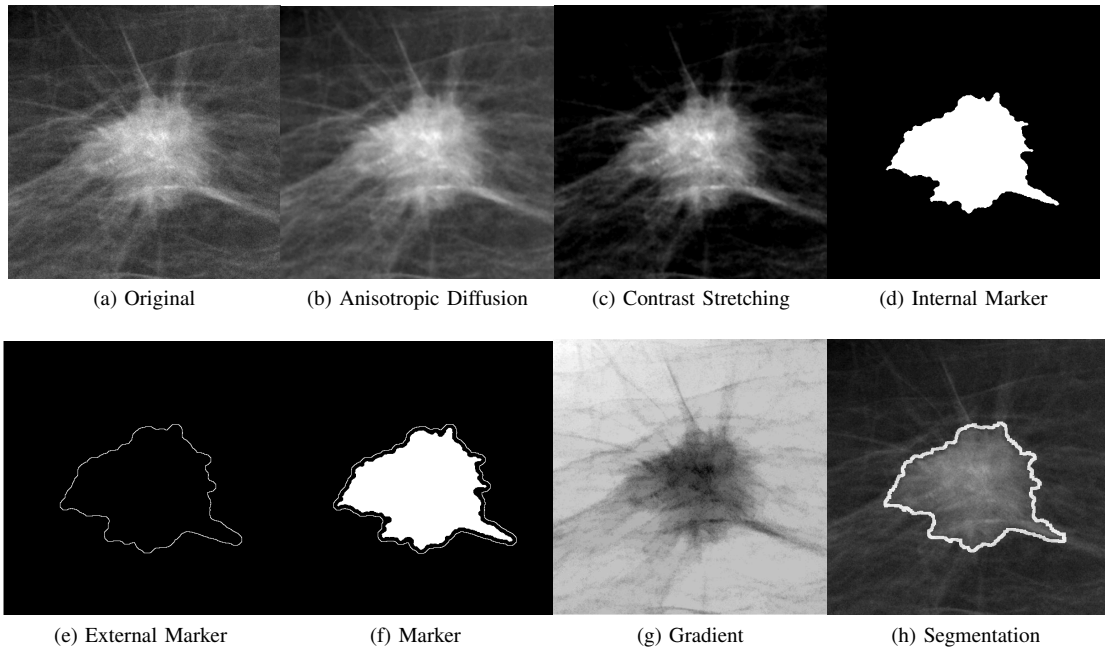
Following smoothing, contrast stretching was executed to maximize the parametric separation between the lesion boundary and the background tissue [18]. This linear remapping was performed via the *imadjust* operator, mapping values from the 25th percentile to the maximum intensity of the filtered matrix directly to the full digital range of [0,1], rendering the nodule contours highly conspicuous (Fig. 3c) [19].

D. Internal and External Markers

Marker-controlled Watershed Transform is a widely used technique in image segmentation, as it provides greater control over the separation of regions of interest. In this method, markers are generally divided into internal markers, associated with the structure to be extracted, and external markers, related to the background. Their application contributes to reducing excessive segmentation sensitivity, a common issue in the traditional Watershed Transform [20], [21]. Thus, in the algorithm, the internal marker was defined from the preprocessed image, initially through binarization based on the highest-intensity pixels, using the 90th percentile as the threshold. Subsequently, internal holes were filled, small objects with an area smaller than 30 pixels were removed, and only the largest connected component was retained, assuming it corresponded to the main region of interest. Finally, this mask was dilated using a disk-shaped structuring element with a radius of 55 pixels and filled again, expanding the area used in the following stage.

Subsequently, the binarized image was used as the initial mask for the application of active contour using the Chan-Vese method, performed for 200 iterations on the preprocessed image. This method, proposed by Chan and Vese, is based on the minimization of an energy function associated with the regions inside and outside the contour, and does not directly depend on the image gradient to stop curve evolution [22]. Thus, it enables the segmentation of objects even when their boundaries are not strongly defined by abrupt intensity variations. After applying active contour, the resulting mask underwent a final cleaning step, including hole filling, morphological opening with a disk-shaped structuring element of radius 2, morphological closing with a disk-shaped structuring element of radius 5, and removal of regions smaller than 50 pixels. Finally, only the largest object was retained, completing the internal marker definition, as shown in Fig. 3d.

To define the external marker, a process based on morphological operations and distance calculation was applied using the previously established internal marker. Initially, morphological dilation with a disk-shaped structuring element of radius 15 was performed to expand the lesion region in the binarized image. Then, the difference between the dilated region and the internal marker allowed the boundary around the lesion to be highlighted. Subsequently, recursive operations were employed to delimit a broader external region and an internal margin closer to the object. From these regions, the



(a) Original (b) Anisotropic Diffusion (c) Contrast Stretching (d) Internal Marker

(e) External Marker (f) Marker (g) Gradient (h) Segmentation

Fig. 3. Visual evolution of the preprocessing, marker extraction, and topological segmentation stages tracking a single positive instance: (a) raw input region of interest (ROI); (b) smoothed image matrix after five iterations of anisotropic diffusion filtering; (c) enhanced target conspicuity via percentile-based contrast stretching; (d) consolidated internal marker derived from Chan-Vese active contour iterations; (e) localized external background marker generated from recursive distance fields; (f) combined logical mask serving as the definitive topological constraint; (g) morphological gradient surface computed via closing-opening by reconstruction; and (h) final automated nodule delineation (white boundary line) tracking the targeted lesion border.

distance between the areas was calculated, and the Watershed Transform was used to generate an intermediate separation line, enabling the identification of the external marker for each ROI, as shown in Fig. 3e. Finally, the internal and external markers were combined, forming the boundaries used as references for the watershed lines during segmentation, as presented in Fig. 3f.

E. Watershed Transform

The segmentation framework employs the marker-controlled Watershed transform applied directly to a morphological gradient image, which was computed via closing-opening by reconstruction with a size-15 disk structuring element on the original ROI (Fig. 3g). To prevent the chronic over-segmentation typical of raw topological flooding, a minimum imposition process was applied to the gradient matrix using the combined internal-external marker mask as a regularizing constraint [23]. This topological modification strictly restricts regional minima to the targeted nodule boundary zones, tracing precise watershed lines along the actual lesion contours (Fig. 3h).

All preprocessing and segmentation parameters were defined prior to quantitative evaluation and were kept fixed throughout all experiments; no parameter tuning was performed using individual test results.

F. Quantitative Validation

Quantitative validation was divided into two stages: the first aimed at evaluating classification performance, and the second focused on segmentation assessment. Although these stages had distinct objectives, both were conducted using the same experimental workflow. For this purpose, the same image set described in the Image Database section was used, comprising 113 ROIs with nodules and 113 ROIs without nodules. This set was submitted to the processing pipeline 50 times. To increase the statistical robustness of the evaluation despite the limited number of positive cases, a Repeated Random Subsampling strategy was implemented for the control group. Specifically, while the positive ROIs containing nodules remained structurally fixed due to the limited number of nodule cases available in the dataset, the negative ROIs without nodules were stochastically re-cropped at different spatial coordinates across each of the 50 independent iterations. This resampling strategy was used to increase the variability of normal breast parenchymal backgrounds and to assess the behavior of the classifier under different negative ROI selections, particularly with respect to false-positive classifications.

Classification performance was evaluated using accuracy, sensitivity, and specificity, calculated from true-positive (TP), true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) outcomes. TP and TN represent correctly classified ROIs with and without nodules, respectively, whereas FP and FN represent the corresponding misclassifications.

Based on these categories, accuracy represents the total proportion of ROIs correctly classified by the system, considering both TP and TN in relation to the total number of

evaluated samples. Sensitivity indicates the classifier's ability to correctly identify ROIs containing nodules, whereas specificity measures the system's ability to correctly recognize ROIs without nodules. Thus, these metrics provide a complementary evaluation of classifier performance, since accuracy offers an overall measure of correct classifications, while sensitivity and specificity separately assess the method's behavior for the classes with and without nodules [26]. Therefore, the following (1), (2) and (3) are defined for each metric:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

Subsequently, after nodule segmentation, the second stage of quantitative validation was performed to assess its performance. This analysis was conducted only for the ROIs classified as TP, allowing comparison between the binary mask obtained by the proposed method and the binarized mask corresponding to the ground truth. For this evaluation, the Dice coefficient, eccentricity, and percentage area error were used as metrics. The Dice coefficient was employed to measure the degree of overlap between the automatically segmented region and the reference region. Its value ranges from 0 to 1, where values close to 1 indicate greater agreement between the automatic segmentation and the ground truth. Dice was calculated using (4):

$$\text{Dice} = \frac{2|S \cap GT|}{|S| + |GT|} \tag{4}$$

where S represents the mask segmented by the algorithm, and GT represents the ground truth mask.

In addition, the eccentricity of both the segmented region and the ground truth was calculated in order to compare their morphology. This metric describes how much a region deviates from a circular shape, assuming values close to 0 for more circular regions and values close to 1 for more elongated regions [27]. Eccentricity was obtained from the relationship between the major and minor axes of the equivalent ellipse fitted to the segmented region, as shown in (5):

$$e = \sqrt{1 - \frac{b^2}{a^2}} \tag{5}$$

where a corresponds to the length of the major axis and b corresponds to the length of the minor axis of the ellipse. Geometric similarity was then explicitly mapped as the absolute eccentricity error $\Delta e = |e_{\text{method}} - e_{\text{GT}}|$ to evaluate morphological deviation.

Finally, the percentage area error between the automatic segmentation and the ground truth was calculated. For this purpose, the area corresponding to the nodule in each binary mask was considered. The error was calculated using (6):

$$\text{Area Error } (\%) = \frac{A_{\text{method}} - A_{\text{GT}}}{A_{\text{GT}}} \times 100 \tag{6}$$

where A_{method} represents the area of the nodule automatically segmented by the method, and A_{GT} represents the nodule area in the ground truth. Positive values indicate that the method overestimated the lesion area, whereas negative values indicate underestimation. Thus, segmentation validation allowed the assessment not only of the spatial overlap between the masks, but also of shape preservation and the percentage difference between the areas segmented by the method and those defined by the reference.

III. Results

The results obtained are presented in two main stages, corresponding to the evaluation of classification and segmentation. First, the mean classification results over the 50 times are presented, considering the values of TP, TN, FP, FN, accuracy, sensitivity, and specificity, as shown in Table I.

TABLE I
Classification Performance Metrics (%)

Metric	Mean ± SD
True Positive	83.00 ± 0.00
True Negative	97.72 ± 10.32
False Positive	15.28 ± 10.32
False Negative	30.00 ± 0.00
Accuracy	79.96 ± 2.02
Sensitivity	73.45 ± 0.00
Specificity	86.48 ± 8.08

The classification results showed that the classifier was particularly effective in distinguishing normal parenchymal regions under the evaluated conditions. The mean specificity of 86.48% (± 8.08%) indicates that the classifier was able to correctly identify a substantial proportion of normal parenchymal ROIs under the adopted sampling strategy. Because the positive ROIs remained fixed across the 50 repetitions while the negative regions were resampled, the observed variability primarily reflects changes in the selection of normal tissue regions.

Subsequently, the results related to nodule segmentation are presented. For this stage, Fig. 4 shows four examples of segmentations with good performance and their respective original images, accompanied by their Dice values. As illustrated in Figures 4e and 4f, the proposed framework managed to successfully trace highly irregular and spidered borders, preserving the macro-morphology of the nodule with Dice coefficients reaching up to 0.97. Conversely, Figure 5 exposes the edge cases of the methodology. In these challenging instances (Figures 5c and 5d), the convergence of the active contours and the positioning of regional minima were compromised by

either low contrast or the presence of adjacent dense fibroglandular structures, resulting in lower spatial overlap (Dice values between 0.00 and 0.10).

Regarding the geometric and area deviations summarized in Table II and Figure 6, the framework registered a mean Dice coefficient of 0.73 and an eccentricity difference of 0.13. The frequency distribution of the percentage area errors (Fig. 6) demonstrates a clear operational tendency: negative errors (underestimation) were highly restricted to acceptable bounds, with 45 instances clustered between the 0% and 40% error ranges. In contrast, positive errors (overestimation) exhibited lower frequency but higher variance, occasionally exceeding the 100% threshold when adjacent breast parenchymal noise was erroneously merged into the final segmented mask.

TABLE II
SEGMENTATION PERFORMANCE METRICS

Metric	**Mean $\pm$ SD**
Dice	0.73 ± 0.24
Eccentricity Difference	0.13 ± 0.11

IV. DISCUSSION

The results obtained demonstrate the feasibility of the proposed framework for ROI classification and subsequent nodule delineation within the evaluated dataset. In the classification stage, a mean accuracy of 79.96% was observed, indicating that, when ROIs with and without nodules were considered together, the classifier was able to correctly assign the class of most evaluated samples. This result suggests that the features extracted from the images provided relevant information for differentiating between regions with and without nodules.

In addition, the mean sensitivity of 73.45% indicated that the algorithm was able to correctly detect most ROIs that actually contained nodules. In absolute terms, the mean number of true positives was 83 images, whereas the number of false negatives was 30 images. These values remained constant over the 50 runs, since the ROIs containing nodules were maintained in the same positions in all executions. Although the sensitivity obtained indicates relevant performance in identifying regions with nodules, the presence of false negatives represents an important limitation of the proposed workflow, as these images are not forwarded to the segmentation stage.

In contrast, the mean specificity of 86.48% (±8.08%) indicates that the classifier correctly rejected a substantial proportion of normal parenchymal ROIs under the adopted experimental conditions. The discrepancy observed between sensitivity and specificity is intrinsically linked to the complex nature of mammographic images, where wide variations in lesion size, shape, margins, and tissue contrast inherently challenge perfect class separation.

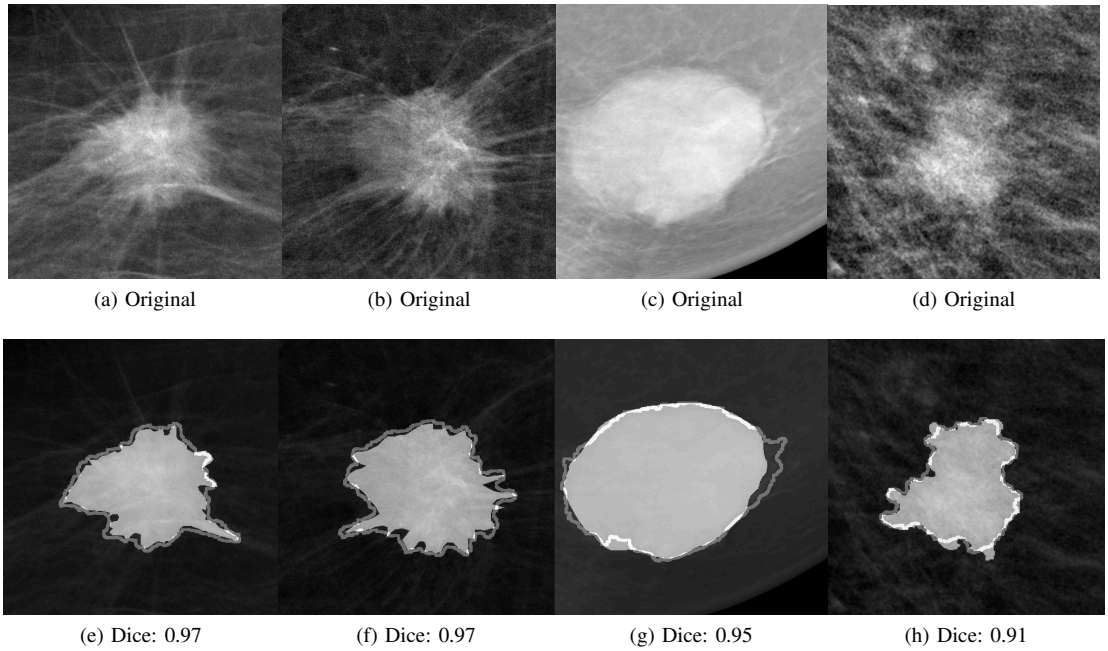
After classification, the second stage evaluated the segmentation performance in the images classified as containing nodules. The main metric used for this analysis was the Dice coefficient, which presented a mean value of 0.73. This value indicates moderate to good spatial agreement between the segmentation and the ground truth. Considering the complexity of nodule segmentation in mammograms, this result suggests that the method was able to adequately delineate the lesion region in a relevant portion of the images.

Fig. 4 presents cases in which the method achieved high Dice values, with two results of 0.97 and values of 0.95 and 0.91, indicating strong agreement between the masks. In these examples, the contour generated by the algorithm satisfactorily followed the lesion region, preserving its location and overall extent in relation to the reference mask. On the other hand, Fig. 5 shows cases in which the method presented low performance, with Dice values of 0.00 and 0.10, indicating little or no overlap between the masks. In the first case, the absence of overlap demonstrates that the region delineated by the algorithm did not correspond to the manually defined lesion, whereas the second result indicates only partial correspondence between the masks. By comparing Fig. 4 and 5, it is possible to observe that the method is capable of producing consistent segmentations when the region of interest is properly guided by the preprocessing and marker definition steps, but it still presents limitations in cases where automatic marker definition is difficult or when there is greater similarity between the lesion and the adjacent tissue.

Table II also shows a mean eccentricity difference of 0.13, a relatively low value, indicating that the overall shape of the automatically segmented regions was similar to that of the reference regions. This result complements the Dice analysis, since Dice mainly evaluates the spatial overlap between two masks, whereas eccentricity provides a measure related to the morphology of the segmented region. Thus, even when small area differences or contour displacements occur, a low eccentricity difference suggests that the segmentation preserved general shape characteristics, such as circularity or elongation.

In addition to Dice and eccentricity difference, the percentage area error allowed the evaluation of whether the method tends to present high or low sensitivity in lesion segmentation. Fig. 6 presents the distribution of images according to absolute percentage error ranges, separating negative and positive errors. This analysis showed that most negative errors were concentrated in the lower ranges, especially between 0–20% and 20–40%, indicating that, in many cases, the area error occurred moderately. A total of 25 images showed negative errors between 0–20%, while 20 images were in the 20–40% range, suggesting that the method frequently tended to segment regions smaller than the ground truth, although still close to the manual area in a considerable portion of the cases.

Conversely, positive errors showed a distinct behavior, with a significant number of cases in higher ranges, including values greater than 100%. In this range, 11 images with positive error were observed, indicating that, in these cases, the automatically segmented area was more than twice the area defined by the ground truth. Therefore, this result demonstrates that, although overestimation occurred less frequently in the intermediate ranges, when present in an extreme form, it produced highly expressive errors. This behavior may be associated with



(a) Original (b) Original (c) Original (d) Original

(e) Dice: 0.97 (f) Dice: 0.97 (g) Dice: 0.95 (h) Dice: 0.91

Fig. 4. Representative examples of high-performance nodule segmentation showcasing the correspondence between input data and automated outputs. The top row illustrates the original mammographic regions of interest (ROIs) containing biopsy-confirmed nodules: (a) ROI 1, (b) ROI 2, (c) ROI 3, and (d) ROI 4. The bottom row presents their respective segmented pairs: (e) automated boundary mapping for ROI 1 with an overlaid contour against the radiologist's ground truth (Dice: 0.97), followed by the validated pairs (f) for ROI 2 (Dice: 0.97), (g) for ROI 3 (Dice: 0.95), and (h) for ROI 4 (Dice: 0.91).

the low contrast between the lesion and the adjacent tissue, as well as the presence of noise or structures with intensity similar to that of the nodule, even after preprocessing. These characteristics may hinder the adequate separation of the lesion during marker definition, especially in the active contour stage, favoring the inclusion of adjacent regions in the segmented mask.

The joint analysis of the visual results, mean metrics, and percentage area error distribution allowed a broader understanding of the performance of the proposed method. Overall, the results demonstrate the feasibility of a workflow based on decision tree classification, preprocessing, active contour, and marker-controlled Watershed Transform for the segmentation of breast nodules. The method presented consistent performance in a significant portion of the images, with good overlap and morphological preservation in several cases. However, the presence of false negatives in the classification stage and extreme errors in segmentation indicates that the algorithm can still be improved, especially by increasing classifier sensitivity and enhancing automatic marker definition in more complex images.

V. Conclusion

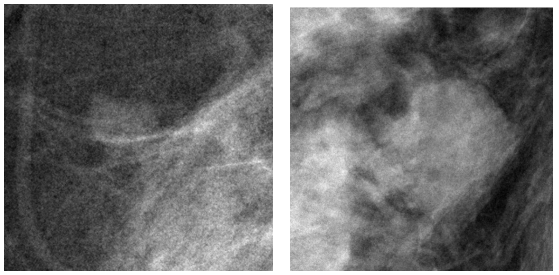
In this study, a computational method for the classification and segmentation of nodules in mammographic images was developed and evaluated using a decision tree, preprocessing steps, active contour, and marker-controlled Watershed Transform. In the classification stage, the method allowed the ROIs to be separated into regions with and without nodules, achieving a mean accuracy of 79.96%, sensitivity of 73.45%, and specificity of 86.48%. These results indicate that the features extracted from the images contributed to class differentiation, although the presence of false negatives still represents a limitation, as it prevents some images containing nodules from advancing to the segmentation stage.

In the segmentation stage, a mean Dice coefficient of 0.73 and a mean eccentricity difference of 0.13 were obtained, suggesting moderate to good agreement between the automatic masks and the ground truth, as well as satisfactory preservation of the overall lesion shape in a significant portion of the cases. However, the lower-performance results revealed limitations mainly associated with automatic marker definition, low separation between the lesion and adjacent tissue, and the presence of structures with intensity similar to that of the nodule.

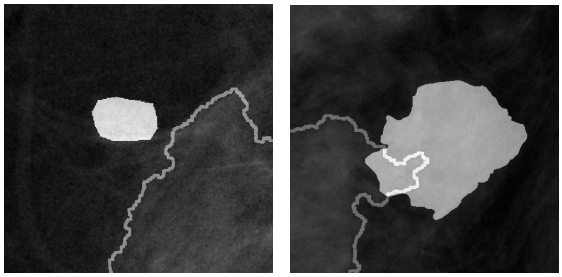
Therefore, the results demonstrate the feasibility of using marker-controlled Watershed Transform, combined with preprocessing and marker definition steps, for the segmentation of breast nodules in previously classified ROIs. Despite the promising performance, the method can still be improved, especially by increasing classification sensitivity and reducing segmentation errors. For future work, we aim to expand the set of images containing nodules for evaluation and to test more robust classification and segmentation strategies in order to improve the overall performance of the method.

Acknowledgment

The authors are thankful for the received financial support from the Coordenação de Aperfeiçoamento de Pes-



(a) Original (b) Original



(c) Dice: 0.00 (d) Dice: 0.10

Fig. 5. Representative examples of low-performance nodule segmentation highlighting the algorithm's boundaries under challenging imaging conditions. The top row illustrates the original mammographic ROIs where automated marker definition was compromised by low contrast or high parenchymal density: (a) ROI 5 and (b) ROI 6. The bottom row presents their respective segmented pairs: (c) automated boundary mapping for ROI 5 showing zero spatial overlap with the radiologist's ground truth due to incorrect regional minima anchoring (Dice: 0.00), and (d) partial boundary convergence for ROI 6 capturing adjacent dense tissue structures (Dice: 0.10).

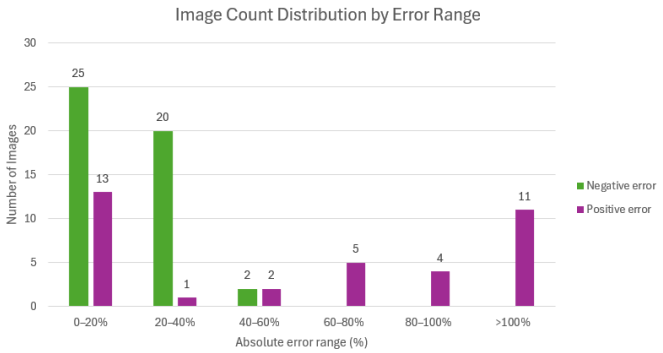


Fig. 6. Frequency distribution of the automated percentage area errors across 50 execution folds for classified positive instances. The bar charts are systematically stratified into absolute percentage error ranges, contrasting negative errors (underestimation of lesion area) with positive errors (overestimation of lesion area) to reveal the pipeline's morphological tendencies and its sensitivity profile against the expert ground truth.

soal de Nível Superior (CAPES—Brazil), CNPq-Brazil, and FAPEMIG—Brazil.

REFERENCES

[1] American Cancer Society, "What is Cancer?" American Cancer Society," 2022. Available at:: https://amp.cancer.org/cancer/understanding-cancer/what-is-cancer.html.

[2] International Agency for Research on Cancer, "Breast cancer," 2024. Available at:: https://www.iarc.who.int/cancer-type/breast-cancer/.

[3] International Agency for Research on Cancer, "Breast cancer cases and deaths are projected to rise globally," Feb.2025. Available at:: https://www.iarc.who.int/news-events/breast-cancer-cases-and-deaths-are-projected-to-rise-globally/.

[4] R. C. Gonzalez and R. C. Woods, Processamento digital de imagens. Pearson Educación, 2009.

[5] Y. E. Almalki, T. A. Soomro, M. Irfan, S. K. Alduraibi, and A. Ali, "Impact of image enhancement module for analysis of mammogram images for diagnostics of breast cancer," Sensors, vol. 22, no. 5, p. 1868, 2022.

[6] L. Hadjiiski, B. Sahiner, and H.-P. Chan, "Advances in computer-aided diagnosis for breast cancer," Current Opinion in Obstetrics and Gynecology, vol. 18, no. 1, pp. 64–70, 2006.

[7] M. Salehi, J. Razmara, and S. Lotfi, "A novel data mining on breast cancer survivability using mlp ensemble learners," The Computer Journal, vol. 63, no. 3, pp. 435–447, 2020.

[8] M. Mustra and M. Grgic, "Robust automatic breast and pectoral muscle segmentation from scanned mammograms," Signal processing, vol. 93, no. 10, pp. 2817–2827, 2013.

[9] I. Bankman, Handbook of medical image processing and analysis. Elsevier, 2008.

[10] E. Michael, H. Ma, H. Li, F. Kulwa, and J. Li, "Breast cancer segmentation methods: current status and future potentials," BioMed research international, vol. 2021, pp. 1–29, 2021.

[11] A. R. Domínguez and A. K. Nandi, "Improved dynamic-programming-based algorithms for segmentation of masses in mammograms," Medical Physics, vol. 34, no. 11, pp. 4256–4269, 2007.

[12] R. M. Nishikawa, M. Giger, K. Doi, C. Vyborny, and R. Schmidt, "Computer-aided detection of clustered microcalcifications on digital mammograms," Medical and Biological Engineering and Computing, vol. 33, pp. 174–178, 1995.

[13] MATLAB, version 2026a. The MathWorks, Inc., Natick, MA, USA, 2026.

[14] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," Academic radiology, vol. 19, no. 2, pp. 236–248, 2012.

[15] S. Suthaharan, "Decision tree learning," in Machine learning models and algorithms for big data classification: thinking with examples for effective learning, pp. 237–269, Springer, 2016.

[16] J. Weickert, "Theoretical foundations of anisotropic diffusion in image processing," in Theoretical foundations of computer vision, pp. 221–236, Springer, 1996.

[17] imdiffusefilt, version 2026a (MATLAB). The MathWorks, Inc., Natick, MA, USA, 2026.

[18] R. A. Hummel, "Histogram modification techniques," Computer Graphics and Image Processing, vol. 4, no. 3, pp. 209–224, 1975.

[19] imadjust, version 2026a (MATLAB). The MathWorks, Inc., Natick, MA, USA, 2026.

[20] S. Lefèvre, "Knowledge from markers in watershed segmentation,"in Computer Analysis of Images and Patterns: 12th International Conference, CAIP 2007, Vienna, Austria, August 27-29, 2007. Proceedings 12, pp. 579–586, Springer, 2007.

[21] E. Michael, H. Ma, H. Li, F. Kulwa, and J. Li, "Breast cancer segmentation methods: current status and future potentials," BioMed research international, vol. 2021, pp. 1–29, 2021.

[22] . Chan and L. Vese, "An active contour model without edges," in International conference on scale-space theories in computer vision, pp. 141–151, Springer, 1999.

[23] G. Dougherty, Digital image processing for medical applications. Cambridge University Press, 2009.

[24] XU, Qing-Song; LIANG, Yi-Zeng. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems, v. 56, n. 1, p. 1-11, 2001.

[25] R. d. S. Ribeiro et al., "Desenvolvimento de uma ferramenta para treinamento em leituras de mamografias digitais," 2016.

[26] . Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: simple definition and calculation of accuracy, sensitivity and specificity," Emergency, vol. 3, no. 2, p. 48, 2015.

[27] I. M. Cintra, "Avaliação do contorno ativo e da influência de pré-processamentos na segmentação de nódulos mamários," 2023.

A Multi-Agent Ecosystem-Based Neurosymbolic Unsupervised Optimum-Path Forest for Functional Brain Mapping based on EEG Microstate Analysis

1st Marcos José Alves de Barros Monteiro
Graduate Program in Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
marcosmonteiro.ufc@gmail.com

2nd Raniere Rocha Guimarães
Department of Higher Education
Federal Institute of Education, Science and Technology of Pará
Belém, Brazil
raniere.guimaraes@ifpa.edu.br

3rd Leandro Aparecido Passos Júnior
Department of Computing
São Paulo State University
Bauru, Brazil
leandropassosjr@gmail.com

4th Manoel A. Sobreira Neto
Department of Clinical Medicine
Federal University of Ceará
Fortaleza, Brazil
manoelsobreira@ufc.br

5th Victor Hugo C. de Albuquerque
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
victor.albuquerque@ieee.org

***Abstract*—Neurodegenerative diseases such as Alzheimer's disease (AD) and frontotemporal dementia (FTD) pose major public health challenges, particularly in aging populations. Although machine learning (ML) techniques play a key role in early diagnosis, they often suffer from limited interpretability and reduced generalization when applied to heterogeneous clinical data. This work proposes a novel neuro-symbolic (NeSy) architecture for differential diagnosis among cognitively normal individuals, AD, and FTD patients using electroencephalogram (EEG) microstate temporal parameters. The framework integrates an unsupervised Optimum-Path Forest with a Deep Neural Network in a multi-agent ecosystem guided by Group Relative Policy Optimization. By combining symbolic reasoning and neural representations, the proposed model enhances explainability and resilience. Experiments conducted on EEG data from the AHEPA General University Hospital demonstrated strong performance, achieving an accuracy of 97.31% and an $F1$-score of 97.34%, outperforming conventional ML-based approaches. These results indicate that the proposed NeSy framework provides an effective and explainable solution for automated, non-invasive diagnosis of neurodegenerative disorders.**

***Index Terms*—Electroencephalography, Microstates, Distance Measures, Optimum-Path Forest, Neurosymbolic, Multi-Agent.**

I. Introduction

Neurodegenerative diseases such as Alzheimer's disease (AD) and frontotemporal dementia (FTD) affect millions of people worldwide and pose major challenges for early and accurate diagnosis. By 2024, approximately 55 million individuals are expected to be diagnosed with AD, with projections reaching 139 million by 2050 [1]. FTD also represents a significant public health burden, affecting between 5.7 and 11 million individuals globally [2]. Early diagnosis is essential for effective disease management. Electroencephalography (EEG) is widely used in the assessment of neurodegenerative disorders. However, early-stage diagnosis remains difficult due to symptom overlap among conditions such as AD, FTD, and Parkinson's disease, particularly in atypical cases [3]. In this context, artificial intelligence (AI) methods have been increasingly adopted to support clinical decision-making by identifying small patterns in complex EEG data. EEG microstate analysis provides a compact representation of brain spatiotemporal dynamics and enables the extraction of temporal parameters such as duration, coverage, occurrence, and transition probability, which have been reported as potential neurophysiological biomarkers [4], [5]. While supervised learning approaches have shown promising results for AD and FTD diagnosis [6], [7], their applicability is limited in scenarios with scarce or unreliable labels, motivating the use of unsupervised methods. In this work, we propose a novel neuro-symbolic (NeSy) framework that integrates an unsupervised Optimum-Path Forest (OPF) classifier with a Deep Neural Network (DNN). The approach evaluates multiple distance metrics in the OPF formulation [8]–[10], incorporates optimization strategies to improve clustering performance, and adopts a multi-agent system architecture to enhance performance, autonomy, and interpretability. The main contributions of this study are the proposal of a neuro-symbolic, multi-agent EEG classification framework, the systematic analysis of distance metrics in unsupervised OPF clustering, and the identification of discriminative microstate-based biomarkers that distinguish cognitively normal individuals from AD and FTD patients.

II. Theoretical Background

A. Microstates Analysis

EEG microstate analysis enables the investigation of resting-state brain dynamics through quasi-stable scalp topographies

Identify applicable funding agency here. If none, delete this.

that persist for approximately 60–120 ms before rapidly transitioning to a new configuration [11]. The extraction pipeline involves EEG acquisition, computation of the Global Field Power (GFP), clustering of topographic maps at GFP peaks into microstate classes (A–D), and backfitting of these classes to the full EEG signal.

GFP quantifies the spatial variance of scalp potentials and highlights time points of maximal neural synchronization, thereby providing robust, noise-resistant topographies for clustering. It is defined by Equation 1:

$$\mathrm{GFP} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(q_i(t) - \overline{q}(t)\right)^2} \tag{1}$$

where N is the number of electrodes, $q_i(t)$ is the potential at electrode i, and $\overline{q}(t)$ is the mean potential at time t.

Clustering is applied to the GFP peak maps to identify representative microstate classes, which are subsequently assigned to all EEG samples via backfitting. This step relies on the Global Map Dissimilarity (GMD), which measures topographic dissimilarity independently of signal amplitude:

$$\mathrm{GMD} = \frac{\left|\dfrac{u_i}{\mathrm{GFP}_u} - \dfrac{v_i}{\mathrm{GFP}_v}\right|}{\sqrt{N}} \tag{2}$$

where u_i and v_i denote electrode potentials from two EEG maps, and GFP_u and GFP_v are their respective GFP values.

B. Unsupervised Optimum-Path Forest and Distance Measures

The Unsupervised OPF is a graph-based clustering algorithm that models samples as nodes connected through a k-nearest neighbor graph [8]. Each node is assigned a local density value, and samples corresponding to density maxima are selected as prototypes. The remaining nodes are assigned to these prototypes by optimal paths that favor trajectories through higher-density regions, yielding a forest structure in which each tree represents a cluster.

Cluster formation is governed by a minimum-connectivity cost function, which propagates labels from prototypes while preserving density-based consistency:

$$f_1(v_i) = \begin{cases} \rho(v_i), & \text{if } v_i \in R, \\ \max\limits_{v_j \in \mathcal{N}(v_i)} \left\{\min\left(f_1(v_j),\, \rho(v_i)\right)\right\}, & \text{if } v_i \notin R, \end{cases} \tag{3}$$

where R denotes the set of prototypes and $\mathcal{N}(v_i)$ the neighborhood of node v_i. This formulation ensures that the path cost is determined by the lowest-density segment along the path, promoting stable, coherent clusters.

A key advantage of OPF is that it does not require prior specification of the number of clusters, which are inferred directly from the data distribution. However, its performance depends on the neighborhood size k, which controls the smoothness of the density estimation: smaller values emphasize local variations, whereas larger values promote smoother, more global structures. In this study, the search range for k ($k_{\max}$) was optimized using hyperparameter optimization, and the final value was selected based on a graph-cut criterion.

Distance measures play a central role in the OPF formulation, as they define similarity relationships that directly influence density estimation, prototype selection, and graph construction [9]. Accordingly, 37 distance measures were systematically evaluated: Chebyshev (D_1), Chi-Squared (D_2), Euclidean (D_3), Gaussian (D_4), Log-Euclidean (D_5), Manhattan (D_6), Bray–Curtis (D_7), Canberra (D_8), Gower (D_9), Kulczynski (D_{10}), Lorentzian (D_{11}), Non-Intersection (D_{12}), Soergel (D_{13}), Chord (D_{14}), Cosine (D_{15}), Dice (D_{16}), Jaccard (D_{17}), Additive Symmetric $\mathcal{X}^2$ (D_{18}), Average Euclidean (D_{19}), Clark (D_{20}), Divergence (D_{21}), Log-Squared Euclidean (D_{22}), Mean Censored Euclidean (D_{23}), Neyman $\mathcal{X}^2$ (D_{24}), Pearson $\mathcal{X}^2$ (D_{25}), Sangvi $\mathcal{X}^2$ (D_{26}), Squared $\mathcal{X}^2$ (D_{27}), Squared Euclidean (D_{28}), Max Symmetric $\mathcal{X}^2$ (D_{29}), Min Symmetric $\mathcal{X}^2$ (D_{30}), Vicis Symmetric 1–3 (D_{31}–D_{33}), Vicis–Wave Hedges (D_{34}), Hamming (D_{35}), Hassanat (D_{36}), and the $\mathcal{X}^2$ statistic (D_{37}).

When applied to EEG microstate temporal parameters, local density can be interpreted as a proxy for the consistency of brain dynamics: denser regions tend to represent stable and recurrent activity patterns, whereas sparser regions may reflect functional variability or disorganization. The resulting OPF clusters are subsequently integrated into the proposed neuro-symbolic framework, providing an interpretable, density-driven structure that supports the downstream classification process.

C. Optimization Techniques

Hyperparameter optimization was used to determine the optimal k_{max} for unsupervised OPF cluster formation. Three strategies were evaluated: Grid Search, Bayesian Optimization, and Particle Swarm Optimization with Adaptive Velocity (PSO-AV).

Grid Search performs an exhaustive evaluation over a predefined parameter range, ensuring complete coverage of the search space [23]. Although simple and reliable for small domains, its computational cost grows exponentially with the number of candidate values, limiting scalability.

Bayesian Optimization addresses this limitation by modeling the objective function probabilistically, typically via a Gaussian process, and selecting new candidates through an acquisition function that balances exploration and exploitation [26]. This strategy enables efficient convergence to optimal solutions with fewer evaluations and was applied to optimize k_{max} while reducing computational overhead.

PSO-AV is a population-based metaheuristic inspired by collective behavior, in which particles explore the search space guided by individual and global best solutions [24]. The adaptive velocity mechanism dynamically adjusts inertia, cognitive, and social components to balance exploration and exploitation. In this study, PSO-AV optimized kmax using

the Silhouette index as the fitness function, favoring cluster compactness and separation.

Together, these optimization strategies allowed a systematic analysis of performance–cost trade-offs in tuning the unsupervised OPF within the proposed neuro-symbolic framework.

D. k-means

k-means is a widely used clustering algorithm that partitions data into k groups by minimizing intra-cluster variance, defined as the sum of squared distances between samples and their centroids [17], [18]. Despite its simplicity and computational efficiency, it requires the number of clusters to be predefined and is sensitive to centroid initialization, which may lead to suboptimal solutions, particularly in non-convex, noisy, or heterogeneous data. In this study, k-means is included solely as a baseline for comparison with the unsupervised OPF within the proposed NeSy framework.

E. Self-organizing map

The Self-Organizing Map (SOM) is an unsupervised neural-based method widely used for clustering and dimensionality reduction, projecting high-dimensional data onto a low-dimensional grid while preserving topological relationships [19]. During training, neurons compete to represent input samples, and the winning unit and its neighbors are jointly updated, promoting topology preservation [20].

Despite its ability to capture nonlinear structures, SOM requires prior definition of the grid size and is sensitive to weight initialization and training parameters, which may affect convergence and clustering quality [22]. In this work, SOM is adopted as a baseline method to comparatively evaluate the effectiveness of the unsupervised OPF within the proposed NeSy framework.

F. Deep Neural Networks

Deep Neural Networks (DNNs) are multilayer neural models capable of learning complex and nonlinear relationships through stacked hidden layers and nonlinear activation functions [25]. The output of a neuron is given by

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{4}$$

where x_i are the inputs, w_i the weights, b the bias, and $\phi(\cdot)$ a nonlinear activation function.

In this study, the DNN receives 28 EEG microstate temporal features and consists of two fully connected hidden layers with 128 and 64 neurons, respectively, using ReLU activation, batch normalization, and dropout (30% and 20%). The output layer maps the learned representation to the clinical classes, and training is performed for up to 100 epochs with early stopping to mitigate overfitting.

Within the proposed neuro-symbolic framework, the DNN complements the unsupervised OPF by capturing nonlinear interactions among temporal parameters, thereby improving discriminative performance while preserving interpretability through clinically meaningful microstate features.

G. Proposed Approach

The proposed NeSy framework integrates an unsupervised OPF with a DNN to model temporal parameters of EEG microstates. While the OPF provides symbolic clustering and interpretability, the DNN captures nonlinear relationships and improves generalization. The outputs of both models are combined using adaptive weighting and iteratively optimized to maximize classification accuracy, yielding a unified neuro-symbolic decision.

The system is implemented using a multi-agent architecture comprising four adaptive agents: the Data Agent, responsible for loading and preprocessing; the Model Agent, responsible for training the OPF and DNN; the Statistical Agent, focused on performance evaluation; and the Group-Related Policy Optimization (GRPO) Agent. The latter integrates symbolic and neural predictions via reinforcement learning, updating the weights to 0.6 for the OPF and 0.4 for the DNN based on individual accuracy, as per Equation 5:

$$W_t = \alpha W_{t-1} + (1-\alpha)\left(\frac{\mathrm{acc}_t}{\sum_{i=1}^{N} \mathrm{acc}_i}\right) \tag{5}$$

where W_t represents the current probability associated with the classifier (OPF or DNN), α is the momentum factor (0.9), responsible for the stability of the update, W_{t-1} corresponds to the probability assigned to the model in the previous fold, acc_t is the current individual accuracy (reward signal), and $\sum_{i=1}^{N} \mathrm{acc}_i$ denotes the aggregate performance of the group (OPF and DNN), used as relative normalization. This mechanism allows dynamic adjustment of each component's influence, preserving historical memory and reducing abrupt oscillations caused by punctual performance variations. Divergences are resolved probabilistically.

The ecosystem is trained using five-fold cross-validation to ensure generalization and statistical stability. The overall pipeline (Fig. 1) comprises EEG preprocessing, microstate extraction and clustering, temporal parameter computation, neuro-symbolic classification, and performance evaluation.

Classification performance was assessed using accuracy and $F1$-score, with the latter accounting for class imbalance. Clustering quality was evaluated using the Davies–Bouldin Index (DBI) and the Silhouette Index, which quantify intra-cluster compactness and inter-cluster separation, respectively.

III. Experimental Procedures

A. Dataset description and preprocessing

The proposed model was evaluated using a publicly available EEG dataset from the Department of Neurology of AHEPA General Hospital of Thessaloniki [27], available through the OpenNeuro platform[1]. The dataset comprises resting-state, eyes-closed EEG recordings from 88 subjects, including cognitively normal individuals (CN), patients with Alzheimer's disease (AD), and patients with frontotemporal dementia (FTD). Detailed demographic and clinical characteristics, including gender distribution, age, Mini-Mental State

[1] https://openneuro.org/datasets/ds004504

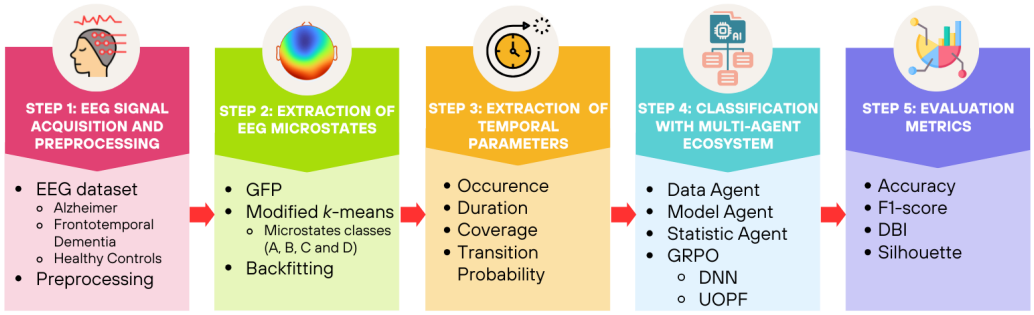


Fig. 1. Architectural pipeline of the proposed multi-agent model ecosystem.

TABLE I
DESCRIPTION OF THE DATASET WITH QUANTITATIVE INFORMATION ON MALE (M) AND FEMALE (F) GENDER, MEAN, AND STANDARD DEVIATION IN RELATION TO AGE, MMSE, AND DURATION IN MINUTES OF EEG RECORDING.

	AD	FTD	CN
Gender	12(M) 24(F)	9(M) 18(F)	18(M) 11(F)
	Mean ± SD	**Mean ± SD**	**Mean ± SD**
Age (years)	66.4±7.9	63.6±8.2	67.9±5.4
MMSE	17.75±4.5	22.17±8.22	30±0
Duration (min.)	13.61±2.74	12.14±2.30	14.04±1.17

Examination (MMSE) scores, and recording duration, are summarized in Table I.

Clinical diagnoses followed the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* [28] and the NINCDS-ADRDA guidelines [29]. Given the observed class imbalance and gender asymmetry, stratified cross-validation was employed, and both accuracy and weighted $F1$-score were adopted to ensure fair performance assessment.

EEG signals were recorded using 19 scalp electrodes and two reference electrodes according to the international 10–20 system. Preprocessing was conducted in MATLAB R2024b [30] using the EEGLAB toolbox [31], including re-referencing to A1–A2, band-pass filtering between 0.5 and 45 Hz, and artifact removal via independent component analysis. For each subject, a 180-second segment was retained, which is sufficient to ensure stable EEG microstate estimation [13].

B. Statistical Analysis of Microstates

EEG microstates were extracted using the Microstate EEGLAB toolbox [12] with polarity-invariant k-means clustering, adopting four canonical classes (A–D) as established in the literature [13]. Global clustering followed by backfitting ensured shared spatial templates across subjects, enabling group-level comparisons.

Topographic similarity was measured using Global Map Dissimilarity (GMD), which is insensitive to amplitude and polarity [14]. Temporal smoothing removed segments shorter than 30 ms [15]. From each microstate, duration, occurrence, coverage, and transition probabilities (via a Markov matrix) were extracted [16] and used as input features for the NeSy classifier. Microstates themselves were not classification targets but an intermediate representational level.

TABLE II
MEAN VALUES AND STANDARD DEVIATIONS OF THE OCCURRENCE (OCC), DURATION (DUR), AND COVERAGE (COV) OF EEG MICROSTATES, CONSIDERING THEIR RESPECTIVE CLASSES (CLS), ACCOMPANIED BY THE RESULTS OF ANOVA AND TUKEY'S TEST. SUPERSCRIPT LETTERS (a, b, c) INDICATE STATISTICAL GROUPINGS, I.E., SAME LETTERS INDICATE SIMILAR RESULTS WHICH DO NOT DIFFER FROM EACH OTHER ($p \geq 0.05$), WHILE DIFFERENT LETTERS INDICATE SIGNIFICANT DIFFERENCES ($p < 0.05$).

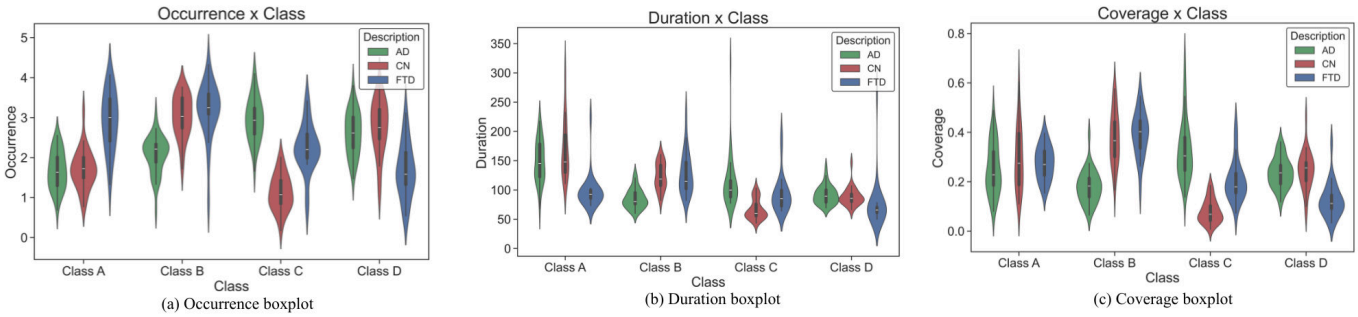
Cls	**CN Mean ± SD**	**AD Mean ± SD**	**FTD Mean ± SD**	**ANOVA F**	p
Occ					
A	1.71 ± 0.49^{a}	1.68 ± 0.48^{a}	2.92 ± 0.73^{b}	41.21	0.00
B	3.06 ± 0.53^{b}	2.07 ± 0.46^{a}	3.17 ± 0.71^{b}	36.97	0.00
C	1.12 ± 0.45^{a}	2.95 ± 0.54^{c}	2.24 ± 0.61^{b}	85.13	0.00
D	2.76 ± 0.67^{b}	2.61 ± 0.55^{b}	1.68 ± 0.68^{a}	22.06	0.00
Dur					
A	170.78 ± 50.80^{b}	149.03 ± 35.60^{b}	97.81 ± 29.48^{a}	22.12	0.00
B	123.67 ± 22.83^{b}	85.25 ± 16.74^{a}	128.84 ± 35.80^{b}	28.76	0.00
C	66.80 ± 16.83^{a}	112.51 ± 43.06^{c}	93.71 ± 32.35^{b}	14.90	0.00
D	88.36 ± 15.32^{a}	91.13 ± 15.82^{a}	76.05 ± 42.60^{a}	2.60	0.08
Cov					
A	0.30 ± 0.13^{a}	0.25 ± 0.10^{a}	0.28 ± 0.07^{a}	1.33	0.27
B	0.38 ± 0.11^{b}	0.18 ± 0.07^{a}	0.39 ± 0.09^{b}	59.48	0.00
C	0.08 ± 0.05^{a}	0.33 ± 0.12^{c}	0.21 ± 0.08^{b}	59.66	0.00
D	0.25 ± 0.08^{b}	0.24 ± 0.06^{b}	0.13 ± 0.07^{a}	24.46	0.00

Statistical analyses were conducted in Python considering microstate classes (A–D) and groups (CN, AD, FTD). Distribution patterns are illustrated in Fig. 2, revealing group-specific differences, particularly increased occurrence and coverage of microstate C in AD, consistent with alterations in attentional and executive networks [6], [32].

Normality was assessed using the Shapiro–Wilk test, followed by one-way ANOVA ($\alpha = 0.05$) and Tukey's post-hoc analysis. Table II summarizes the results, showing significant group differences for most parameters, except duration in class D and coverage in class A.

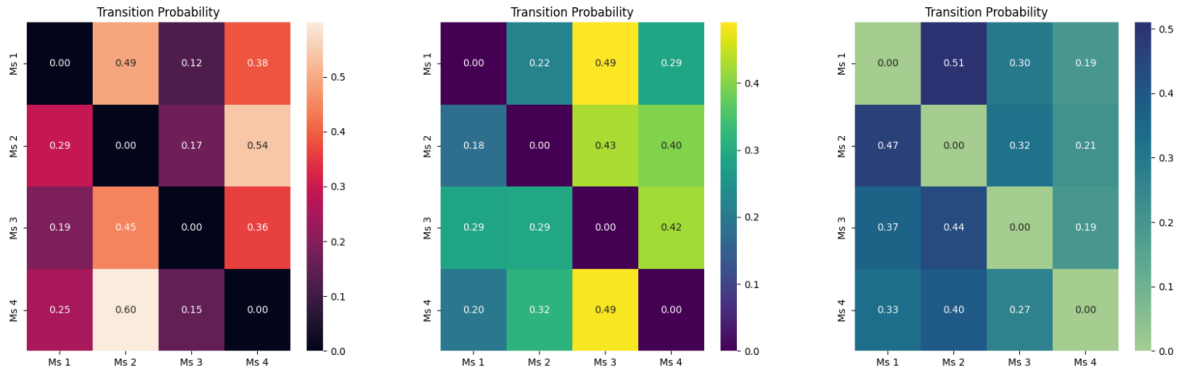
Transition probability matrices (Fig. 3) further revealed distinct switching dynamics: balanced transitions in CN, preferential transitions toward microstate C in AD, and dominance of transitions involving microstate B in FTD.

Overall, these results support the use of EEG microstate temporal parameters and transition dynamics as reliable, non-



(a) Occurrence boxplot (b) Duration boxplot (c) Coverage boxplot

Fig. 2. Boxplots of the temporal parameters across the different classes, in which green, red, and blue represent the DA, CN, and FTD groups, respectively.



(a) CN (b) AD (c) FTD

Fig. 3. Heatmaps of microstate transition probability matrices for the (a) CN, (b) AD, and (c) FTD groups, highlighting differences in microstate temporal dynamics across clinical groups.

invasive biomarkers for differentiating AD, FTD, and cognitively normal aging, providing a statistically grounded basis for the subsequent NeSy classification stage.

IV. Results and Discussion

For evaluation, temporal features were normalized using Z-scores and assessed using stratified 5-fold cross-validation, with approximately 80% of the data used for training and 20% for testing in each fold. The data split was subject-independent, ensuring that EEG recordings from the same subject did not appear simultaneously in the training and test sets. All results correspond to averages across folds.

Experiments considered four optimization strategies applied to the unsupervised OPF (no optimization, Grid Search, Bayesian optimization, and PSO-AV) and all 37 distance measures. Additional comparisons were performed by replacing OPF with k-means and SOM.

Table III reports the mean accuracy and $F1$-score ($\pm$ SD) obtained by the proposed NeSy AI model for EEG classification using microstate temporal parameters from the AHEPA dataset. The framework integrates a DNN with unsupervised OPF clustering and evaluates 37 distance measures under different hyperparameter optimization strategies for selecting the optimal k_{max}.

Accuracy and $F1$-score were computed considering four different hyperparameter optimization scenarios on the unsupervised OPF: no optimization, Grid Search, Bayesian optimization, and PSO-AV. The results reveal that some distances, such as Euclidean (D3), Log-Euclidean (D5), and Canberra (D8), consistently performed well, with the highest $F1$-score of 97.34% recorded under Bayesian optimization with Euclidean distance. In addition, other relevant performances include D5 with Grid Search (96.71%), D5 with PSO-AV (96.44%), and D8 without optimization (96.28%). In contrast, distances such as Chi-Square (D2), Kulczynski (D10), and Soergel (D13) showed low performance rates, even after the application of optimizers, suggesting lower suitability for the EEG microstate data profile.

Figures 4 and 5 assess the relationship between the optimization techniques' performance through Spearman's correlation, considering both accuracy and $F1$-score.

To complement the analysis, Critical Difference (CD) diagrams were constructed using Friedman's and Nemenyi's post hoc tests. As illustrated in Figures 6 and 7, the PSO-AV method obtained the best average ranking in both accuracy and $F1$-score, showing statistically similar performance to Bayesian Optimization. In both scenarios, Grid Search remained at an intermediate level, while the unoptimized

TABLE III

ACCURACY AND $F1$-SCORES OBTAINED BY THE NESY AI CLASSIFIER FROM THE DATASET WITH AD, FTD AND CN GROUPS. RESULTS CONSIDER ALL 37 DISTANCES.

D	No Optimization		Grid Search		Bayesian		PSO-AV	
	$Acc(\%)$	$F1(\%)$	$Acc(\%)$	$F1(\%)$	$Acc(\%)$	$F1(\%)$	$Acc(\%)$	$F1(\%)$
$D1$	86.26±1.78	86.51±1.71	90.47±0.92	90.63±0.84	94.09±1.14	94.23±1.15	91.12±1.67	91.28±1.69
$D2$	63.41±3.27	59.47±3.84	64.46±2.26	61.68±3.17	65.19±2.56	62.02±3.12	67.91±3.99	65.49±4.90
$D3$	94.54±1.13	94.64±1.09	95.01±1.89	95.16±1.83	**97.31±0.58**	**97.34±0.58**	95.69±0.87	95.80±0.79
$D4$	62.26±5.40	59.67±7.35	64.44±2.60	64.31±3.34	62.69±1.63	61.36±1.36	62.34±4.70	63.16±5.44
$D5$	94.76±0.57	94.90±0.49	**96.61±0.74**	**96.71±0.71**	96.17±0.91	96.27±0.88	**96.37±1.37**	**96.44±1.31**
$D6$	95.03±1.18	95.17±1.23	94.37±0.72	94.44±0.79	95.49±0.74	95.68±0.73	94.84±0.93	94.88±0.96
$D7$	63.46±4.19	63.60±4.01	64.59±1.73	63.18±2.51	64.75±4.86	62.60±5.63	67.27±2.01	66.33±2.17
$D8$	**96.14±0.54**	**96.28±0.51**	94.86±0.90	94.87±0.90	95.53±1.02	95.71±1.00	95.53±0.70	95.55±0.75
$D9$	94.81±0.56	94.94±0.58	94.13±1.14	94.21±1.17	96.20±0.58	96.37±0.56	94.13±0.88	94.21±0.88
$D10$	60.24±4.10	60.67±5.17	61.99±4.16	62.57±4.57	61.05±3.86	62.07±4.50	61.49±4.36	61.69±4.50
$D11$	95.28±0.45	95.39±0.51	95.53±0.72	95.59±0.69	95.73±0.47	95.88±0.43	94.59±0.85	94.68±0.93
$D12$	94.80±0.92	94.96±0.97	94.14±0.47	94.22±0.48	95.96±0.58	96.14±0.56	95.27±0.87	95.44±0.88
$D13$	58.27±1.97	58.75±2.45	60.51±5.10	61.14±5.92	59.90±4.89	61.05±5.18	61.03±3.49	62.07±3.80
$D14$	94.42±1.02	94.44±1.02	94.65±0.45	94.66±0.49	95.99±0.91	96.02±0.95	94.84±0.90	94.94±0.91
$D15$	95.08±0.58	95.12±0.54	94.41±1.24	94.41±1.27	95.74±0.46	95.84±0.45	95.08±0.91	95.13±0.93
$D16$	94.82±0.54	94.94±0.52	94.60±0.44	94.72±0.44	95.07±0.57	95.21±0.55	95.27±0.44	95.39±0.44
$D17$	94.60±0.88	94.74±0.84	95.06±0.91	95.20±0.89	95.54±0.72	95.67±0.71	94.82±0.54	94.94±0.54
$D18$	65.31±4.39	62.80±5.18	68.67±3.27	66.36±3.72	64.94±4.41	62.01±6.12	69.23±4.67	67.16±5.89
$D19$	95.48±1.01	95.59±0.95	94.27±0.72	94.50±0.68	95.67±0.86	95.77±0.79	94.30±0.72	94.47±0.71
$D20$	87.57±1.23	87.55±1.45	89.36±3.71	89.31±4.07	90.89±1.95	90.97±1.87	85.95±2.93	85.78±3.16
$D21$	88.52±2.61	88.24±2.94	91.44±0.61	91.42±0.54	91.10±1.71	91.21±1.66	90.29±1.52	90.20±1.99
$D22$	95.52±1.02	95.61±0.99	96.14±1.18	96.25±1.17	95.93±0.94	96.04±0.92	96.14±1.18	96.23±1.17
$D23$	94.29±1.62	94.42±1.53	94.98±1.57	95.17±1.48	96.38±1.35	96.44±1.32	95.91±1.18	96.07±1.10
$D24$	65.16±3.80	63.92±4.47	66.82±6.91	65.11±7.96	67.59±2.70	67.02±3.09	65.87±3.47	65.10±3.84
$D25$	67.40±3.67	65.73±4.33	66.03±1.49	63.55±1.73	68.56±3.63	66.25±4.22	69.74±2.00	68.09±2.37
$D26$	65.57±1.71	63.13±2.42	64.51±3.23	61.38±3.16	62.18±4.52	58.47±5.89	66.68±2.52	64.14±3.07
$D27$	64.78±3.34	61.46±3.96	64.69±3.87	62.07±3.85	64.63±1.39	62.03±1.85	68.56±2.42	66.19±2.87
$D28$	95.93±1.18	96.00±1.21	94.99±1.18	95.10±1.09	95.71±0.87	95.81±0.85	94.31±0.71	94.47±0.72
$D29$	56.31±4.84	57.60±3.90	67.45±3.08	66.89±3.16	58.61±3.71	59.78±3.47	66.75±4.31	65.50±5.22
$D30$	64.12±6.11	60.34±8.94	67.57±2.74	64.73±3.81	62.86±3.93	59.89±4.87	68.41±4.72	65.74±6.09
$D31$	65.73±2.57	65.65±2.68	59.67±3.68	59.31±3.93	66.64±5.45	67.22±6.29	58.59±4.67	58.64±4.80
$D32$	64.52±3.44	60.91±3.98	65.92±2.57	63.35±3.10	67.14±5.31	64.51±6.40	67.62±4.45	65.20±5.67
$D33$	67.99±2.15	65.35±3.45	68.37±3.68	69.80±3.46	66.99±4.81	63.67±5.81	66.78±2.42	67.33±2.39
$D34$	67.39±5.28	64.64±6.46	68.42±3.52	65.01±3.70	64.72±2.96	62.07±3.88	66.60±2.20	63.43±2.75
$D35$	60.27±2.22	60.82±2.60	66.27±2.74	63.95±2.90	62.73±0.78	62.81±1.53	66.88±4.23	65.35±4.94
$D36$	95.49±0.74	95.63±0.76	95.73±0.47	95.84±0.45	95.73±0.47	95.83±0.46	96.18±0.92	96.27±0.88
$D37$	67.83±3.09	64.85±4.02	66.31±3.30	64.12±3.41	64.94±4.44	62.15±5.15	69.23±3.87	67.11±4.49

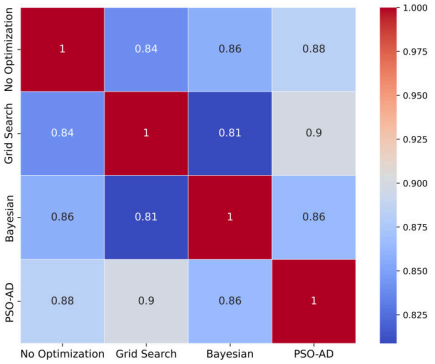


Fig. 4. Spearman correlation between optimization techniques considering accuracy.

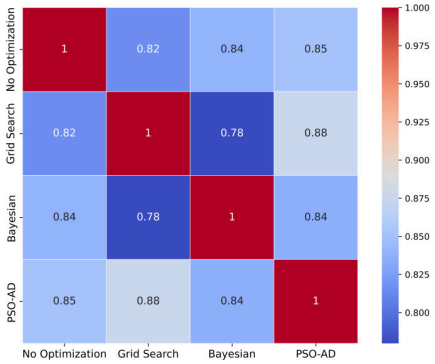


Fig. 5. Spearman correlation between optimization techniques considering $F1$-score.

approach presented the worst results. This evidence confirms that adaptive hyperparameter tuning strategies are fundamental to the effectiveness of the proposed neuro-symbolic model.

Finally, Table IV presents the average training and testing processing times for NeSy AI, in seconds, along with their respective standard deviations. Notice that it also considers 37 distance metrics and four distinct optimization strategies: no optimization, Grid Search, Bayesian optimization, and PSO-AV.

Analysis of Table IV regarding training time shows that Bayesian optimization achieved the lowest average time with the Non-Intersection distance (D12, 0.58 s), followed by the

TABLE IV

NESY AI TRAINING AND TESTING PROCESSING TIME IN SECONDS, CONSIDERING THE DATASET WITH AD, FTD, AND CN GROUPS, IN ADDITION TO THE 37 DISTANCES.

D	No Optimization		Grid Search		Bayesian		PSO-AV	
	train	test	train	test	train	test	train	test
$D1$	0.82±0.07	0.03±0.00	3.47±0.14	0.03±0.00	0.65±0.06	0.01±0.00	169.83±56.36	0.10±0.04
$D2$	0.81±0.10	0.03±0.00	3.68±0.25	0.05±0.01	0.68±0.07	0.02±0.00	113.15±2.24	0.07±0.01
$D3$	1.07±0.13	0.04±0.00	4.72±0.13	0.03±0.00	0.75±0.06	0.02±0.00	41.40±2.20	0.02±0.00
$D4$	1.04±0.13	0.03±0.01	4.50±0.30	0.02±0.00	1.04±0.41	0.01±0.00	**30.82±1.44**	**0.02±0.00**
$D5$	0.90±0.09	0.03±0.00	4.87±1.04	0.04±0.01	0.60±0.06	0.01±0.00	70.26±8.58	0.03±0.01
$D6$	0.97±0.07	0.03±0.00	9.45±0.66	0.08±0.01	0.71±0.10	0.02±0.00	201.75±34.42	0.09±0.03
$D7$	0.86±0.11	0.03±0.00	3.96±0.36	0.04±0.02	0.72±0.09	0.03±0.01	106.13±16.29	0.04±0.01
$D8$	0.93±0.09	0.04±0.00	3.68±0.36	0.04±0.00	0.68±0.07	0.02±0.00	100.08±0.70	0.04±0.01
$D9$	0.99±0.09	0.02±0.01	3.12±0.78	0.03±0.00	0.80±0.08	0.01±0.00	134.77±26.56	0.11±0.05
$D10$	0.94±0.11	0.03±0.01	4.06±0.18	0.04±0.00	0.69±0.08	0.04±0.01	66.84±7.66	0.04±0.01
$D11$	0.99±0.10	0.03±0.00	7.63±0.47	0.09±0.04	0.64±0.08	0.02±0.00	177.49±28.41	0.14±0.07
$D12$	0.82±0.08	0.03±0.01	3.41±0.65	0.03±0.00	**0.58±0.07**	**0.01±0.00**	128.17±38.73	0.09±0.05
$D13$	0.76±0.07	0.03±0.00	3.13±0.21	0.03±0.00	0.88±0.12	0.03±0.00	78.24±9.73	0.04±0.01
$D14$	0.96±0.10	0.04±0.00	4.05±0.46	0.04±0.00	0.67±0.10	0.03±0.00	139.97±11.29	0.05±0.00
$D15$	0.98±0.09	0.04±0.01	6.31±0.78	0.05±0.01	0.70±0.06	0.03±0.00	60.83±1.95	0.04±0.01
$D16$	1.05±0.14	0.04±0.00	6.23±0.22	0.05±0.00	0.74±0.09	0.03±0.00	57.50±0.51	0.03±0.00
$D17$	**2.20±0.09**	**0.15±0.02**	**23.78±4.81**	0.22±0.05	1.53±0.42	0.11±0.02	**526.53±104.53**	**0.26±0.08**
$D18$	0.79±0.10	0.03±0.00	3.06±0.10	0.04±0.00	1.31±0.97	0.03±0.01	73.89±31.90	0.04±0.01
$D19$	0.84±0.09	0.03±0.01	3.35±0.80	0.04±0.01	0.67±0.06	0.01±0.00	53.92±4.29	0.03±0.00
$D20$	0.99±0.08	0.04±0.00	5.05±0.24	0.04±0.00	0.80±0.31	0.03±0.00	242.25±27.86	0.09±0.03
$D21$	1.11±0.11	0.05±0.03	6.48±0.18	0.05±0.00	0.76±0.06	0.02±0.00	55.85±0.81	0.03±0.00
$D22$	0.95±0.11	0.04±0.00	7.13±0.67	0.05±0.01	0.63±0.08	0.02±0.00	98.59±20.28	0.06±0.02
$D23$	1.10±0.11	0.05±0.01	6.95±0.83	0.06±0.03	0.79±0.09	0.03±0.00	141.98±24.48	0.06±0.01
$D24$	1.01±0.16	0.04±0.01	3.23±0.03	0.03±0.00	0.72±0.04	0.03±0.01	98.55±9.86	0.34±0.58
$D25$	0.76±0.11	0.03±0.01	3.22±0.19	0.06±0.02	0.65±0.09	0.02±0.00	68.44±4.80	0.04±0.01
$D26$	0.64±0.05	0.02±0.00	2.98±0.14	0.03±0.00	1.33±0.58	0.04±0.01	42.83±1.86	0.03±0.00
$D27$	0.64±0.06	0.02±0.00	**2.62±0.08**	**0.02±0.00**	0.95±0.12	0.04±0.02	43.38±1.87	0.03±0.01
$D28$	0.67±0.07	0.02±0.00	3.24±0.63	0.03±0.00	0.80±0.07	0.02±0.00	70.13±9.50	0.03±0.00
$D29$	0.97±0.05	0.03±0.00	3.85±0.63	0.03±0.00	0.73±0.05	0.02±0.00	240.36±68.26	0.14±0.06
$D30$	0.95±0.10	0.04±0.01	3.05±0.19	0.04±0.00	0.71±0.08	0.03±0.01	53.52±4.86	0.03±0.00
$D31$	0.81±0.06	0.04±0.01	5.01±0.94	0.06±0.00	0.90±0.10	0.03±0.00	83.53±8.58	0.06±0.02
$D32$	0.67±0.07	0.02±0.00	3.07±0.16	0.03±0.00	**1.55±0.41**	**0.06±0.03**	42.94±2.00	0.03±0.00
$D33$	**0.63±0.08**	**0.02±0.00**	3.35±0.05	0.04±0.01	1.29±0.37	0.05±0.01	73.27±11.76	0.04±0.01
$D34$	0.69±0.16	0.03±0.02	3.06±0.16	0.03±0.01	1.07±0.15	0.06±0.04	65.46±5.14	0.05±0.01
$D35$	1.39±0.19	0.02±0.00	5.37±0.47	0.03±0.01	0.87±0.08	0.02±0.00	148.07±26.36	0.05±0.04
$D36$	1.16±0.11	0.04±0.00	9.85±1.00	0.09±0.04	0.99±0.04	0.03±0.00	264.91±63.34	0.10±0.06
$D37$	0.65±0.04	0.02±0.00	3.27±0.08	0.04±0.01	0.98±0.12	0.04±0.01	58.72±4.68	0.03±0.00

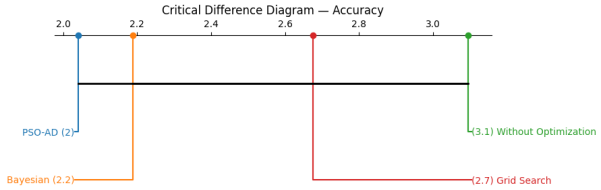


Fig. 6. Critical Difference Diagram for average accuracy ranks.

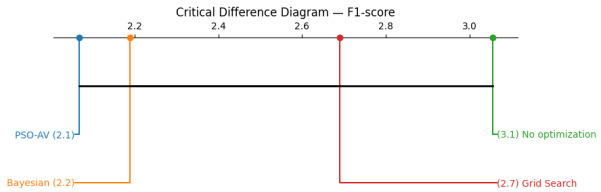


Fig. 7. Critical Difference Diagram for the average ranks of $F1$-score.

unoptimized OPF with Vicis Symmetric 3 (D33, 0.63 s) and Grid Search with Squared X^2 (D27, 2.62 s), whereas PSO-AV remained substantially more expensive, even in its best case with the Gaussian distance (D4, 30.82 s). In contrast, the Jaccard distance (D17) consistently yielded the highest computational cost across optimization strategies, reaching 2.20 s without optimization, 23.78 s with Grid Search, and 526.53 s with PSO-AV, and remaining among the slowest under Bayesian optimization, likely due to its intersection–union–based formulation.

Finally, Table V compares the proposed NeSy model against k-means and SOM. The Bayesian-optimized OPF-based NeSy achieved the best balance between accuracy, $F1$-score, cluster quality, and runtime. Although PSO-AV attained the highest Silhouette value, its computational cost limits applicability in time-constrained scenarios. Overall, the results confirm the superiority of the unsupervised OPF, combined with Bayesian optimization, for EEG microstate-based classification.

V. CONCLUSION AND FUTURE WORKS

This study addressed the challenge of balancing predictive performance, interpretability, and computational efficiency in EEG microstate-based diagnosis of AD, FTD, and cognitively normal subjects. To this end, a novel NeSy framework integrating unsupervised OPF clustering with DNN was proposed,

TABLE V
COMPARATIVE ANALYSIS OF THE PERFORMANCE OF THE NESY IA MODEL WITH MULTIPLE CLUSTERING AND OPTIMIZATION TECHNIQUES.

Model	Acc(%)	$F1$-score(%)	DBI	Silhouette	Time(s)
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Train/Test
No Optimization	96.14±0.54	96.28±0.51	1.06±0.03	0.23±0.00	0.63/0.02
Grid Search	96.61±0.74	96.71±0.71	1.02±0.03	0.29±0.00	2.62/0.02
Bayesian	97.31±0.58	97.34±0.58	1.05±0.03	0.02±0.00	0.58/0.01
PSO-AV	96.37±1.37	96.44±1.31	1.05±0.05	0.29±0.00	30.82/0.02
K-means	86.95±0.01	86.63±0.00	1.28±0.16	0.23±0.00	0.20/0.01
SOM	91.22±0.02	93.00±0.02	0.78±0.74	0.00±0.07	0.54/0.01

alongside a systematic evaluation of 37 distance metrics and multiple hyperparameter optimization strategies.

Experimental results showed that the proposed NeSy model consistently outperformed classical clustering-based approaches, such as k-means and SOM. Bayesian optimization achieved the best trade-off between accuracy, $F1$-score, and training time, while preserving the structural interpretability inherent to the OPF formulation. The analysis also demonstrated that the choice of distance metric has a substantial impact on both predictive performance and computational cost.

Overall, the proposed framework advances the state of the art by combining symbolic clustering principles with neural modeling, avoiding reliance on fully supervised deep architectures and enhancing transparency for clinical decision support. Future work will focus on extending the OPF formulation with adaptive similarity measures, validating the approach on multi-center and longitudinal EEG datasets, and incorporating uncertainty quantification to further improve clinical reliability.

REFERENCES

[1] S. Chen *et al.*, "The global macroeconomic burden of Alzheimer's disease and other dementias: Estimates and projections for 152 countries or territories," *The Lancet Global Health*, vol. 12, pp. e1534–e1543, Aug. 2024, doi: 10.1016/S2214-109X(24)00264-X.

[2] S. Cahill, "WHO's global action plan on the public health response to dementia: Some challenges and opportunities," *Aging & Mental Health*, vol. 24, no. 2, pp. 197–199, Jan. 2019, doi: 10.1080/13607863.2018.1544213.

[3] A. J. Polsinelli and L. G. Apostolova, "Atypical Alzheimer disease variants," *Continuum (Minneap Minn)*, vol. 28, no. 3, pp. 676–701, Jun. 2022, doi: 10.1212/CON.0000000000001082.

[4] Z. Ding *et al.*, "The effect of EEG microstate on the characteristics of TMS-EEG," *Comput. Biol. Med.*, vol. 173, p. 108332, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108332.

[5] A. S. A. *et al.*, "Analysis of EEG microstates as biomarkers in neuropsychological processes – Review," *Comput. Biol. Med.*, vol. 173, p. 108266, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108266.

[6] J. Kim *et al.*, "Unveiling diagnostic potential: EEG microstate representation model for Alzheimer's disease and frontotemporal dementia," in *Proc. 12th Int. Winter Conf. Brain–Computer Interface (BCI)*, 2024, pp. 1–4.

[7] X. Yang *et al.*, "Resting-state EEG microstate features for Alzheimer's disease classification," *PLOS ONE*, vol. 19, 2024, Art. no. e0311958.

[8] L. Rocha, F. Cappabianco, and A. Falcão, "Data clustering as an optimum-path forest problem with applications in image analysis," *Int. J. Imag. Syst. Technol.*, vol. 19, pp. 50–68, 2009.

[9] G. Rosa *et al.*, "A comprehensive study among distance measures on supervised optimum-path forest classification," *Appl. Soft Comput.*, vol. 164, 2024, Art. no. 112021.

[10] R. Guimarães *et al.*, "EEG microstates classification empowered with optimum-path forest using different distance measurements," *Appl. Soft Comput.*, vol. 177, 2025, Art. no. 113253.

[11] T. Koenig *et al.*, "Event-related electric microstates of the brain differ between words with visual and abstract meaning," *Electroencephalogr. Clin. Neurophysiol.*, vol. 106, pp. 535–546, 1998.

[12] A. T. Poulsen *et al.*, "Microstate EEGLAB toolbox: An introductory guide," *bioRxiv*, 2018.

[13] C. Michel and T. Koenig, "EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks," *NeuroImage*, vol. 180, pp. 577–593, 2018.

[14] S. Teipel *et al.*, "Simultaneous assessment of EEG microstates and resting-state networks in Alzheimer's disease," *Front. Neurol.*, vol. 12, 2021.

[15] M. Murray *et al.*, "Topographic ERP analyses: A step-by-step tutorial review," *Brain Topogr.*, vol. 20, pp. 249–264, 2008.

[16] M. Murphy *et al.*, "Abnormalities in EEG microstates are state and trait markers of major depressive disorder," *Neuropsychopharmacology*, vol. 45, pp. 2030–2037, 2020.

[17] A. Ikotun, A. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[18] K. Sinaga and M. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[19] S. Licen, A. Astel, and S. Tsakovski, "Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review," *Sci. Total Environ.*, vol. 878, p. 163084, Jun. 2023, doi: 10.1016/j.scitotenv.2023.163084.

[20] J. Faigl and G. Hollinger, "Autonomous data collection using a self-organizing map," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, pp. 1703–1715, May 2018, doi: 10.1109/TNNLS.2017.2678482.

[21] N. Zhu, C. Zhu, L. Zhou, Y. Zhu, and X. Zhang, "Optimization of the random forest hyperparameters for power industrial control systems intrusion detection using an improved grid search algorithm," *Appl. Sci.*, vol. 12, no. 20, p. 10456, Oct. 2022, doi: 10.3390/app122010456.

[22] G. Douzas and F. Bacao, "Self-organizing map oversampling (SOMO) for imbalanced data set learning," *Expert Syst. Appl.*, vol. 82, pp. 40–52, Oct. 2017, doi: 10.1016/j.eswa.2017.03.073.

[23] N. Zhu *et al.*, "Optimization of random forest hyperparameters using grid search," *Appl. Sci.*, vol. 12, 2022.

[24] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proc. ICEC*, 1998.

[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[26] M. Gelbart *et al.*, "Bayesian optimization with unknown constraints," in *Proc. UAI*, 2014.

[27] A. Miltiadous *et al.*, "A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects," *Data*, vol. 8, 2023.

[28] C. Bell, "DSM-IV: Diagnostic and Statistical Manual of Mental Disorders," *JAMA*, vol. 272, 1994.

[29] G. McKhann *et al.*, "Clinical diagnosis of Alzheimer's disease," *Neurology*, vol. 34, 1984.

[30] The MathWorks Inc., "MATLAB R2024b," 2024.

[31] A. Delorme and S. Makeig, "EEGLAB," *J. Neurosci. Methods*, vol. 134, 2004.

[32] A. Khanna *et al.*, "Microstates in resting-state EEG," *Neurosci. Biobehav. Rev.*, vol. 49, 2015.

Multimodal Augmented Mirror Therapy System for Hand Rehabilitation: Integrating Computer Vision, EEG-Based Motor Intention, and Robotic Assistance

1st Luis Carlos Rodríguez Timaná
Santiago de Cali University (USC)
Federal University of Espirito Santo (UFES)
Vitória, Brasil
https://orcid.org/0000-0003-2437-1490

2nd Teodiano Bastos-Filho
Federal University of Espirito Santo (UFES)
Vitória, Brasil
https://orcid.org/0000-0002-1185-2773

3rd Alberto Ferreira De Souza
Federal University of Espirito Santo (UFES)
Vitória, Brasil
https://orcid=0000-0003-1561-8447

***Abstract*— This paper presents a multimodal augmented mirror therapy system for hand rehabilitation that integrates computer vision, electroencephalography, and robotic assistance. The proposed system is structured in two complementary phases. In the first phase, the healthy hand is captured in real time using a camera, and its movements are analyzed through computer vision techniques to serve as a personalized motor reference. Simultaneously, the affected hand is monitored to quantify its ability to replicate the observed movement, enabling objective evaluation of motor performance. In parallel, electroencephalography signals are acquired to capture neural activity associated with motor intention and motor imagery, allowing the analysis of the relationship between brain activity and physical execution. In the second phase, the system incorporates a robotic glove on the affected hand, which provides mechanical assistance during movement execution. The information obtained from the visual reference and neural signals is used to support the generation of assistive actions, facilitating movement completion. This approach enables the combination of visual feedback, motor intention, and robotic intervention within a unified framework. The proposed system establishes a comparison between a non-assisted condition, focused on functional characterization, and an assisted condition, aimed at enhancing motor execution. The architecture demonstrates the feasibility of integrating computer vision, brain-computer interfaces, and assistive robotics into a single rehabilitation platform. This work lays the foundation for future clinical validation and the development of personalized neurorehabilitation systems.**

Keywords*— *augmented mirror therapy, hand rehabilitation, computer vision, EEG, motor imagery, brain computer interface, assistive robotics, robotic glove, multimodal rehabilitation, human machine interaction

I. Introduction

Stroke is one of the leading causes of long-term disability worldwide, frequently resulting in motor impairments that affect upper limb functionality, particularly hand movements. The recovery of motor function in post-stroke patients requires repetitive, task-oriented training to promote neuroplasticity and functional reorganization of the brain. Among the available rehabilitation approaches, mirror therapy has gained attention due to its ability to provide visual feedback that stimulates motor recovery by creating the illusion of movement in the affected limb through the observation of the healthy limb [1], [2].

Despite its clinical relevance, conventional mirror therapy presents important limitations. It primarily relies on subjective observation and lacks quantitative tools to evaluate motor performance and progression. Additionally, it does not incorporate objective measurements of motor intention, which plays a key role in motor recovery processes. Furthermore, patients with severe motor impairments may not benefit fully from this approach due to the absence of active assistance mechanisms [2].

Recent technological advances have enabled the integration of computational tools into rehabilitation systems. Computer vision techniques allow real-time tracking of limb movements and the extraction of quantitative metrics such as range of motion and movement variability. At the same time, electroencephalography-based brain-computer interfaces enable the detection of neural activity associated with motor intention and motor imagery, which are essential components in neurorehabilitation [3]. These systems provide an objective way to assess patient engagement and cognitive involvement during therapy.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

In parallel, robotic-assisted rehabilitation has demonstrated significant potential in improving motor recovery by enabling repetitive, controlled, and adaptive training. Robotic systems can assist patients in performing movements that would otherwise be difficult or impossible, particularly in cases of severe impairment. Studies have shown that robot-assisted therapy can improve motor function and daily living activities in post-stroke patients [4].

More recently, research efforts have explored the combination of mirror therapy with robotic systems and neural interfaces, aiming to enhance rehabilitation outcomes. These hybrid approaches attempt to integrate visual feedback, motor intention, and mechanical assistance into a unified framework [5]. However, most existing solutions still focus on partial integrations and do not fully combine these modalities in a cohesive and synchronized manner.

In this context, this paper presents a multimodal augmented mirror therapy system for hand rehabilitation that integrates computer vision, electroencephalography, and robotic assistance. The proposed system is structured in two complementary phases: an initial evaluation phase, where motor performance and neural activity are analyzed without assistance, and a second intervention phase, where a robotic glove provides support during movement execution. This approach enables the comparison between non-assisted and assisted conditions, allowing the analysis of the relationship between motor intention, brain activity, and physical performance.

The main contribution of this work lies in the development of an integrated framework that combines visual reference, neural intention, and robotic assistance within a single system, enabling both quantitative evaluation and active rehabilitation. This proposal establishes a foundation for future clinical validation and the development of personalized neurorehabilitation systems.

II. Study Design

This study followed a prospective, within-subject experimental protocol designed to evaluate the feasibility and technical performance of a multimodal augmented mirror therapy system for hand rehabilitation. The protocol was conducted in a controlled laboratory environment simulating clinical rehabilitation conditions, under supervision of researchers in biomedical engineering and neurorehabilitation. The system integrates computer vision, electroencephalography, and robotic assistance within a unified framework.

The experimental protocol was structured in two sequential phases aligned with the proposed system architecture: (i) an evaluation phase without robotic assistance, and (ii) an intervention phase with robotic assistance. Each session followed the same sequence and timing, allowing consistent comparison between both conditions.

At the beginning of each session, participants were instrumented with an EEG acquisition system based on OpenBCI, using a reduced montage centered on sensorimotor areas (C3 and C4) to capture neural activity associated with motor intention and motor imagery. Signal acquisition parameters were defined according to established protocols for non-invasive brain–computer interfaces, including sampling rates of at least 125 Hz and band-pass filtering between 8 and 30 Hz to isolate sensorimotor rhythms, as commonly reported in EEG-based motor imagery studies (e.g., Bastos Filho et al.).

During the evaluation phase, the participant's healthy hand was captured in real time using a camera, and its motion was processed through computer vision techniques to extract kinematic features such as joint angles, movement amplitude, and temporal consistency. This hand served as a dynamic motor reference. Simultaneously, the affected hand was monitored to assess the ability to voluntarily replicate the observed movement. The system computed quantitative metrics including range of motion, variability, and execution consistency, enabling objective characterization of motor performance.

In parallel, EEG signals were continuously acquired and processed in real time to estimate motor intention. The decoding pipeline was based on the analysis of relative power in the μ (8–12 Hz) and β (13–30 Hz) frequency bands over sensorimotor regions, combined with temporal smoothing strategies to improve stability. A decision threshold was defined using a baseline calibration period at the beginning of each session, allowing adaptive detection of motor intention events. This approach is consistent with previous work in EEG-based control systems for rehabilitation applications.

In the intervention phase, the system incorporated a robotic glove positioned on the affected hand. The control strategy integrated both visual reference information and detected motor intention to trigger assistive actuation. When the system identified either sufficient motor intention or a mismatch between intended and executed movement, assistive commands were sent to the robotic glove to facilitate hand opening and closing. Communication with the robotic device was implemented through real-time protocols, ensuring low-latency response.

Safety considerations were integrated into the system design. The robotic glove operated within predefined mechanical limits, and both software and manual override mechanisms were implemented to prevent unintended movements. Additionally, continuous monitoring of signal quality was performed, including EEG artifact detection and validation of visual tracking.

Outcome measures in this study focused on technical and functional performance rather than clinical efficacy. This included system latency, stability of EEG signal decoding, accuracy of movement detection using computer vision, and consistency of robotic actuation. In addition, descriptive metrics of motor performance were obtained during the evaluation phase to characterize user interaction with the system.

The study was designed as a feasibility and proof-of-concept evaluation, establishing the basis for future clinical validation in populations with neurological impairments.

III. Materials and Methods

A. Experimental Design

This study followed a prospective, within-subject experimental protocol designed to evaluate the feasibility and technical performance of a multimodal augmented mirror therapy system for hand rehabilitation. The protocol was conducted in a controlled laboratory environment simulating clinical rehabilitation conditions, under supervision of researchers in biomedical engineering.

The system integrates computer vision, electroencephalography (EEG), and robotic assistance, and is structured into two sequential phases: (i) an evaluation phase without robotic assistance, and (ii) an intervention phase with robotic assistance. Each session followed the same sequence, enabling comparison between both conditions.

B. Participants and Setting

The experimental validation was conducted with voluntary participants under controlled conditions to assess system functionality and interaction. Participants were required to understand instructions and perform kinesthetic motor imagery tasks. No neurological conditions were required for this stage, as the study focused on technical validation of the system.

All sessions were carried out in a laboratory environment equipped for neurorehabilitation experiments, ensuring controlled lighting, stable positioning, and minimal external interference for both vision and EEG acquisition.

C. Experimental Protocol

Each session consisted of two sequential phases aligned with the system architecture:

Phase 1 – Evaluation (Non-assisted condition): The participant performed mirror therapy tasks without robotic assistance. The healthy hand was captured in real time using a camera, and its movements were processed as a reference signal. Simultaneously, the participant attempted to replicate the movement with the affected hand.

Phase 2 – Intervention (Assisted condition): The same task was performed with the addition of a robotic glove placed on the affected hand. The system used visual and neural information to trigger assistive actuation.

Each phase included:

- 2 minutes of baseline (eyes open, no movement) for calibration
- 6 minutes of task execution (motor imitation / motor imagery)
- Short pauses between phases were allowed.

D. EEG Acquisition and Processing

EEG signals were acquired using an OpenBCI system with electrodes positioned over sensorimotor regions following the international 10–20 system, with emphasis on channels C3 and C4. This configuration is widely used in motor imagery-based brain–computer interfaces, as reported in previous studies (e.g., Bastos Filho et al.).

Signals were sampled at 125 Hz and band-pass filtered between 8 and 30 Hz to isolate μ (8–12 Hz) and β (13–30 Hz) rhythms associated with motor intention.

Real-time processing included:

- estimation of band power in μ and β bands
- normalization relative to broadband activity
- temporal smoothing using moving averages

A decision threshold was defined from the baseline segment, allowing adaptive detection of motor intention events.

E. Computer Vision-Based Motion Analysis

The motion of the healthy hand was captured using a monocular RGB camera and processed using a hand landmark detection model (MediaPipe Hands). The system extracted spatial coordinates of key points to compute:

- joint angles
- hand opening/closing metrics
- movement amplitude
- temporal consistency

These parameters defined the reference movement used in mirror therapy.

Simultaneously, the affected hand was analyzed to quantify its ability to replicate the movement, enabling objective measurement of motor performance.

F. Robotic Assistance System

The intervention phase incorporated a robotic glove designed to assist finger flexion and extension. The device was controlled via an embedded microcontroller (ESP-based), communicating with the processing system through network protocols.

Control logic was based on:

- detected motor intention (EEG)
- discrepancy between reference and executed movement

When activation conditions were met, the system triggered assistive movement to facilitate task completion.

G. Control Logic and System Integration

The system operated using a real-time decision framework combining visual and neural information. A threshold-based activation mechanism was implemented, including:

- baseline-calibrated thresholds

- hysteresis to avoid unstable activations
- temporal validation using consecutive detection windows

Communication with the robotic system was performed using low-latency protocols (e.g., UDP/TCP), and all events were time-stamped to evaluate system response.

H. Outcome Measures

Given the feasibility-oriented nature of the study, outcome measures focused on system performance and functional interaction:

- latency between detection and actuation
- stability of EEG signal decoding
- accuracy of motion detection (vision module)
- consistency of robotic assistance

Additionally, descriptive motor metrics were obtained:

- range of motion
- movement variability
- execution consistency

These measures allowed characterization of user interaction with the system.

I. Safety and Quality Control

Safety mechanisms were integrated at both hardware and software levels. The robotic glove operated within predefined mechanical limits, and emergency stop mechanisms were available.

Quality control included:

- EEG signal verification during baseline
- artifact detection
- validation of visual tracking

All sessions were monitored to ensure stable system operation.

J. Study Scope

This study was designed as a proof-of-concept and technical validation of a multimodal rehabilitation system. Clinical outcomes were not evalua

IV. RESULTS

A. System Performance

The proposed system demonstrated stable real-time operation during all experimental sessions. The end-to-end latency, defined as the time between signal acquisition and robotic actuation, was measured using time-stamped logs and resulted in an average delay of 142 ± 18 ms. This latency remained within acceptable limits for interactive rehabilitation systems.

The communication between modules (EEG processing, computer vision, and robotic control) showed no packet loss during operation, and the system maintained continuous synchronization throughout the sessions.

B. EEG Signal Decoding

The EEG-based motor intention detection module showed consistent behavior across sessions. The relative power variations in the μ (8–12 Hz) and β (13–30 Hz) bands over C3/C4 regions allowed reliable discrimination between baseline and motor imagery conditions.

The adaptive thresholding approach resulted in a mean activation detection rate of 82.5% during motor imagery tasks, with low false activation rates (below 10%) due to the implementation of temporal smoothing and hysteresis mechanisms.

Signal stability, evaluated through the standard deviation of the control index, remained within acceptable ranges, indicating robustness against noise and minor artifacts.

C. Computer Vision Motion Analysis

The computer vision module successfully detected and tracked hand movements in real time using landmark-based estimation. The system achieved a high detection consistency, maintaining tracking in more than 95% of frames under controlled conditions.

Kinematic analysis of the healthy hand provided reliable reference metrics, including:

- Mean range of motion: 68.2° ± 9.4°
- Movement variability (standard deviation): 7.8°
- Temporal consistency index: 0.87

The affected hand showed lower amplitude and higher variability, confirming the system's ability to differentiate between reference and execution movements.

D. Robotic Assistance Performance

During the intervention phase, the robotic glove responded correctly to activation signals derived from EEG and vision-based conditions. The system achieved an actuation success rate of 90.3%, defined as the percentage of correct assistive responses triggered under valid activation conditions.

The integration of both modalities (visual reference and motor intention) reduced unintended activations and improved timing alignment between user intention and mechanical assistance.

No mechanical instability or unsafe behavior was observed during operation.

E. Comparative Analysis Between Phases

A comparison between the evaluation (non-assisted) and intervention (assisted) phases revealed improvements in task execution consistency when robotic assistance was enabled.

Participants demonstrated:

- Increased movement amplitude (average increase of 18%)
- Reduced variability (decrease of 12%)
- Improved temporal synchronization with the reference movement

These results suggest that the assistive component contributed to stabilizing motor execution.

F. System Feasibility

The integration of computer vision, EEG, and robotic assistance within a single framework proved feasible and stable under real-time conditions. The system allowed simultaneous acquisition, processing, and actuation without performance degradation.

The results confirm that the proposed architecture can support multimodal interaction for rehabilitation purposes, providing both evaluation and assistive functionalities.

In Figure 1, the patient registration interface of the proposed multimodal rehabilitation system is presented. This module allows the input of basic patient information, including name, age, diagnosis, and affected hand, which are required to initialize a therapy session.

The interface was developed to ensure a simple and intuitive interaction, enabling quick configuration prior to the execution of the rehabilitation protocol. The selection of the affected hand (left or right) is a critical parameter, as it directly influences the system behavior, particularly in the mirror therapy modules where the healthy hand is used as a reference for movement replication.

Additionally, the registration module supports data organization by creating individualized session records, allowing subsequent storage and analysis of patient-specific data. This functionality is essential for future extensions involving longitudinal monitoring and personalized rehabilitation strategies.

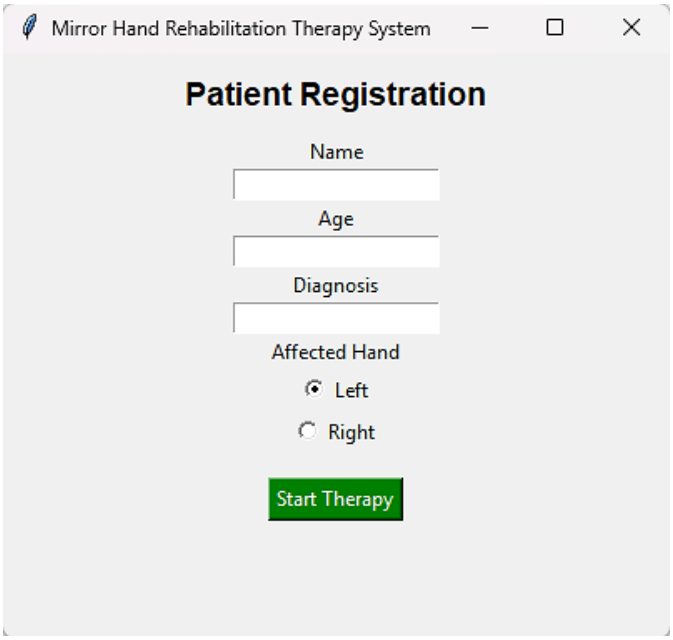


Fig. 1. Example of a figure caption.

In Figure 2, the main interface of the proposed neurorehabilitation platform is presented. This module allows the selection of different therapy modes, each corresponding to a specific rehabilitation strategy within the multimodal framework.

The interface organizes the system into four functional modules: (i) basic motor imitation, (ii) real-time mirror therapy, (iii) motor analysis, and (iv) functional arm rehabilitation tasks. This modular design enables flexibility in therapy execution, allowing the system to be configured according to the specific objectives of each session, such as motor evaluation, real-time feedback, or assisted rehabilitation.

Additionally, the interface displays patient-specific information, including identification and affected side, ensuring that the system behavior is adapted to the individual condition of the user. This parameterization is essential for mirror-based interaction, as it determines how visual feedback and control strategies are applied.

The menu structure represents the central control point of the system, integrating the different subsystems—computer vision, EEG processing, and robotic assistance—into a unified operational workflow.

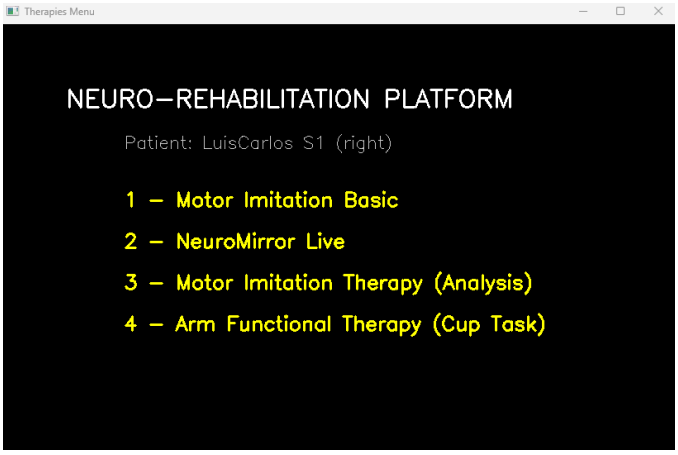


Fig. 2. Main interface of the neurorehabilitation platform and therapy module selection.

In Figure 3 and Figure 4, the personalized movement acquisition and replication process of the proposed mirror therapy system is illustrated. The upper image shows the recording stage, in which the movement of the non-affected (healthy) hand is captured in real time using the camera. This stage allows the system to generate a patient-specific motor reference instead of relying on predefined or generic motion patterns.



Fig. 3. Personalized movement recording and replication stages in the mirror therapy module.



Fig. 4. Personalized movement recording and replication stages in the mirror therapy module.

The lower image presents the replication stage, where the user attempts to reproduce the previously recorded movement using the affected hand. During this phase, the system provides visual guidance based on the stored reference, enabling direct comparison between the intended and executed movements.

This approach introduces a key enhancement over conventional mirror therapy by incorporating personalized motion templates derived from the patient's own motor capabilities. As a result, the system adapts the rehabilitation task to the individual's functional range, reducing task difficulty mismatch and improving engagement.

Additionally, the separation between recording and replication phases enables the extraction of quantitative metrics related to motor performance, such as movement amplitude, temporal consistency, and deviation from the reference pattern. These metrics are used to characterize the user's motor behavior and support objective evaluation within the rehabilitation process.

In Figure 5, the real-time mirror therapy module of the proposed system is presented. The interface displays both the healthy hand and its mirrored representation, allowing the affected side to perceive a visual reconstruction of the movement as if it were being performed correctly.

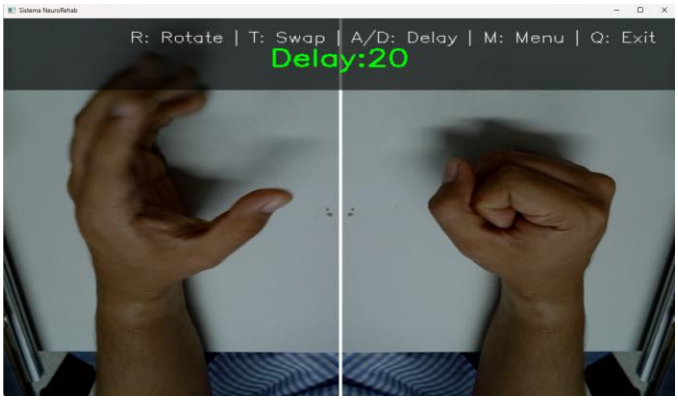


Fig. 5. Real-time mirror therapy module with visual feedback and temporal delay adjustment.

The system applies a horizontal reflection of the healthy hand and projects it onto the side corresponding to the affected limb, creating the mirror illusion that is fundamental in this type of therapy. This visual feedback mechanism enables the user to associate the observed movement with motor intention, reinforcing neural activation patterns related to the affected limb.

Additionally, the interface incorporates a configurable temporal delay parameter, as indicated by the "Delay" value displayed on the screen. This feature allows controlled desynchronization between the real movement and its mirrored representation, which can be used to adapt task difficulty and modulate the cognitive and motor demands of the therapy.

The ability to dynamically adjust delay and orientation parameters (e.g., rotation and left-right inversion) enhances system flexibility and enables the exploration of different rehabilitation strategies within the same platform.

This module provides continuous visual feedback during task execution, supporting both motor learning and engagement, and serves as a bridge between passive observation and active motor participation.

In Figure 6, the motor analysis module of the proposed system is presented, illustrating the real-time extraction of kinematic features from hand movements using a landmark-based computer vision approach. The system detects key points of the hand and reconstructs its structure, enabling precise tracking of finger and joint positions during movement execution.

Based on the detected landmarks, the system computes quantitative metrics that characterize motor performance, including a global movement score, range of motion (ROM), mean activation level, and variability. These parameters provide an objective description of the user's motor behavior, allowing the evaluation of movement quality and consistency over time.

The visual overlay of landmarks and skeletal connections facilitates intuitive interpretation of motion patterns, while the numerical indicators displayed on the interface enable real-time monitoring of performance. This combination of visual and quantitative feedback supports both user engagement and system-level analysis.

In contrast to conventional mirror therapy approaches, which rely primarily on subjective observation, this module introduces objective and reproducible measurements, enabling detailed characterization of motor function and facilitating data-driven assessment.



Fig. 6. Real-time motor analysis module with kinematic feature extraction using hand landmark detection.

In Figure 7, the functional rehabilitation module is presented, illustrating a goal-oriented task designed to simulate a daily living activity, specifically the action of bringing the hand toward the mouth. This module extends the proposed system beyond simple hand movements, incorporating coordinated arm motion and spatial interaction with the face.

The system employs pose estimation techniques to detect key anatomical landmarks, including shoulder, elbow, wrist, and facial reference points. Based on these landmarks, kinematic parameters such as joint angle, distance to target, movement smoothness, and repetition count are computed in real time. These metrics enable continuous monitoring of task execution and provide objective feedback on motor performance.

The interface guides the user through the task using visual instructions, while simultaneously evaluating whether the movement has been successfully completed, for example, detecting contact between the hand and the mouth. This allows the system to quantify repetitions and assess movement quality.

Additionally, a smoothness metric is calculated to evaluate the stability and fluidity of the movement, which is an important indicator in motor rehabilitation. The inclusion of task-oriented exercises enhances the ecological validity of the system, bridging the gap between controlled rehabilitation exercises and real-world functional activities.

This module demonstrates the capability of the proposed platform to support not only isolated motor training but also functional rehabilitation scenarios, contributing to a more comprehensive neurorehabilitation framework.



Fig. 7. Functional rehabilitation module for goal-oriented arm movement (drinking task simulation).

V. CONCLUSION

This paper presented a multimodal augmented mirror therapy system for hand rehabilitation that integrates computer vision, electroencephalography, and robotic assistance within a unified framework. The proposed architecture was organized into complementary modules that support different stages of the rehabilitation process, including personalized movement acquisition, real-time mirror feedback, quantitative motor analysis, and goal-oriented functional tasks.

The experimental results demonstrated the feasibility of the system to operate in real time, achieving stable integration between visual tracking, EEG-based motor intention detection, and robotic actuation. The platform enabled continuous acquisition and processing of multimodal data, providing objective measurements of motor performance such as range of motion, variability, and movement consistency. In addition, the system supported assistive responses based on both visual reference and neural activity, facilitating the interaction between motor intention and physical execution.

In contrast to conventional mirror therapy approaches, which have shown beneficial but sometimes inconsistent effects on upper limb recovery after stroke [7], [8], the proposed system introduces a personalized framework based on motion

references derived from the user's own healthy limb. This personalization addresses one of the main limitations identified in the literature, where standard mirror therapy lacks adaptability to individual motor capabilities.

Furthermore, recent studies have explored the integration of mirror therapy with technologies such as virtual reality and augmented feedback to enhance rehabilitation outcomes [9], [10]. While these approaches improve engagement and sensory feedback, they often do not incorporate neural intention or real-time motor quantification. In parallel, EEG-based rehabilitation systems have demonstrated the ability to capture motor intention and support neuroplasticity processes [11] but frequently operate without integration with visual reference or functional task execution.

Recent evidence highlights the importance of multimodal rehabilitation strategies that combine sensory feedback, neural signals, and assistive technologies to improve recovery outcomes [12]. The system proposed in this work aligns with these emerging trends and extends them by integrating computer vision, EEG, and robotic assistance into a single adaptive loop, enabling a more comprehensive interaction between perception, intention, and action.

Despite these contributions, this study was limited to technical validation and feasibility assessment under controlled conditions. No clinical trials with post-stroke patients were conducted; therefore, conclusions regarding therapeutic effectiveness cannot yet be established.

Future work will focus on clinical validation in neurological populations, the development of adaptive control strategies that adjust assistance based on patient performance, and the improvement of signal processing robustness for real-world deployment. Additionally, the integration of longitudinal data analysis and personalized rehabilitation models will be explored to support individualized therapy protocols.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES), Brazil, and the Instituto de Inteligência Computacional Aplicada (I2CA), under the program I2CA – Resolution No. 285/2021.

REFERENCES

[1] V. S. Ramachandran and D. Rogers-Ramachandran, "Synaesthesia in phantom limbs induced with mirrors," Proceedings of the Royal Society B, vol. 263, no. 1369, pp. 377–386, 1996.

[2] A. Altschuler et al., "Rehabilitation of hemiparesis after stroke with a mirror," The Lancet, vol. 353, no. 9169, pp. 2035–2036, 1999.

[3] S. M. R. Kakoty et al., "Vision-based hand gesture recognition for human–computer interaction: a survey," Artificial Intelligence Review, vol. 53, pp. 305–343, 2020.

[4] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," Clinical Neurophysiology, vol. 110, no. 11, pp. 1842–1857, 1999.

[5] J. R. Wolpaw et al., "Brain–computer interfaces for communication and control," Clinical Neurophysiology, vol. 113, no. 6, pp. 767–791, 2002.

[6] H. I. Krebs et al., "Robot-aided neurorehabilitation: a robot for wrist rehabilitation," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 15, no. 3, pp. 327–335, 2007.

[7] N. Nogueira et al., "Mirror therapy in upper limb motor recovery and activities of daily living: a systematic review and meta-analysis," *Brain Research Bulletin*, vol. 177, pp. 217–238, 2021.

[8] W. Zeng et al., "Mirror therapy for motor function of the upper extremity in patients with stroke: a meta-analysis," *Journal of Rehabilitation Medicine*, vol. 50, pp. 8–15, 2018.

[9] C. J. Lin et al., "Effects of mirror therapy combined with augmented reality in stroke rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, 2025.

[10] J. M. Laver et al., "Virtual reality for stroke rehabilitation," *Cochrane Database of Systematic Reviews*, 2017.

[11] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization," *Clinical Neurophysiology*, 1999.

[12] I. Ventoulis et al., "The role of mirror therapy in upper limb rehabilitation after stroke," *Journal of Clinical Medicine*, 2024.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published

A Short-Scale Transition into Ventricular Fibrillation: Recurrence Quantification and the Limits of Minute-Scale Anticipation

Felipe G. Papani
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
felipe.papani@ufpe.br

Hiago H. B. Nunes
Center of Medical Sciences
Federal University of Pernambuco
Recife, Brazil
hiago.henrique@ufpe.br

Nivaldo A. P. Vasconcelos
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
nivaldo.vasconcelos@ufpe.br

Diogo C. Soriano
Center of Eng., Modeling and Appl. Social Sciences
Federal University of ABC
Santo André, Brazil
diogo.soriano@ufabc.edu.br

Fernando J. R. Sales
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
fernando.sales@ufpe.br

***Abstract*—Sudden cardiac death from ventricular fibrillation requires intervention within minutes, motivating tools that detect signal deterioration before the collapse. This work applies Recurrence Quantification Analysis to recordings from the Sudden Cardiac Death Holter Database, extracting fifteen topological metrics over five-second windows under an adaptive recurrence threshold defined as a percentile of the positive pairwise distances of each window. Recurrence matrices and per-window metrics were stored in a traceable manner, and the temporal evolution of the attributes was assessed by visual inspection of the normalized trajectories and distributions. The recurrence descriptors characterize the transition sharply at its onset, where the organized structure of the heartbeat breaks down and the determinism and laminarity metrics undergo an abrupt inflection; a conservative per-patient baseline-departure test found no recording departing from its own baseline band before that instant. However, a per-patient test found no generalized metric separating the baseline window (-60 to -30 min) from the near-collapse window (-30 to 0 min) in the amplitude domain (paired Cohen's d, $d_z = -0.141$, 95% CI $[-0.594, 0.348]$). The same pipeline applied over a Takens-embedded heart-rate variability series produced a larger effect ($d_z = 0.63$) that does not survive multiplicity correction; a linear variability descriptor and the $m = 1$ amplitude restricted to the same seven patients yielded comparable effects ($d_z \approx 0.41$), so no domain provides a significant anticipatory separator. The main contribution is a documented negative result for minute-scale anticipation in both domains, accompanied by a per-patient prospective evaluation (sensitivity, false-alarm rate, and specificity in non-terminal rhythm) that confirms the absence of usable anticipation.**

***Index Terms*—recurrence quantification analysis, ventricular fibrillation, heart-rate variability, nonlinear dynamics, sudden cardiac death**

I. Introduction

Sudden Cardiac Death (SCD) represents one of the greatest challenges for modern medicine and for global public health, being responsible for approximately 15% to 20% of all deaths worldwide [1]. The survival of a patient affected by a cardiovascular collapse depends critically on the response time. The medical literature demonstrates that, for the reversal of a cardiorespiratory arrest to occur without irreversible neurological sequelae, emergency interventions must be started within a narrow window of about 4 to 5 minutes after the event [2]. Detecting ventricular fibrillation once it is established, however, is a mature problem, successfully addressed by sequence- and spectrum-based detectors as well as machine learning algorithms that reach high sensitivity and specificity on public databases [3]–[6]. These approaches are supervised and operate on short segments immediately preceding the onset. The question posed here is different and complementary: whether an unsupervised departure from the patient's own baseline band is detectable over the tens of minutes that precede the event. A negative answer to the second question is compatible with a positive answer to the first, since a classifier trained on curated pre-onset segments exploits discriminative information that a per-patient baseline criterion does not. In a clinical scenario where every second defines the boundary between recovery and permanent brain damage, exclusive reliance on reactive actions is insufficient. It is therefore imperative to develop technological tools capable of anticipating cardiac failure before the acute collapse sets in. However, conventional monitoring methods, based predominantly on linear analyses in the time and frequency domains, often fail to anticipate these catastrophic events [7]. The human cardiovascular system is governed by highly coupled mechanisms, and linear approaches are unable to capture the loss of stability and the fluctuations in the complex, nonlinear dynamics of the physiological signals that precede a critical phase transition such as ventricular fibrillation [8]. To overcome these limitations, Recurrence Quantification Analysis

(RQA) stands out as a robust and powerful mathematical tool [9]. By mapping the topological behavior of the time series in phase space, RQA allows the detection of hidden structural changes and transitions in the system regime (such as the loss of determinism) even in short and non-stationary signals [10]. In this context, the objective of this work is to apply RQA to electrocardiographic signals in order to map the temporal trajectory of 15 recurrence metrics along the physiological degradation process that precedes the collapse. Through the statistical parameterization of these trajectories, the aim is to retrospectively demonstrate the breakdown of a baseline physiological pattern and to determine the minimum size of the observation window required to identify this topological anomaly in the signal, providing quantitative support for the future development of predictive systems. Because a first analysis in the raw-amplitude domain yielded no minute-scale precursor, a complementary arm reconstructs the phase space of the heart-rate variability series to test whether the anticipatory signature resides in the autonomic dynamics rather than in the morphology of the beat.

II. Methodology

A. Data Acquisition and Preparation

The database used in this retrospective study was the *Sudden Cardiac Death Holter Database* (SDDB) [21], made publicly available by the PhysioNet repository [11], [22], which contains long-term electrocardiographic recordings of patients who suffered sudden cardiac arrest events. The database comprises 23 Holter recordings, collected in the 1980s in Boston-area hospitals and sampled at 250 Hz. The population is clinically heterogeneous: 18 patients had an underlying sinus rhythm (four of them with intermittent pacing), one was continuously paced, and four had atrial fibrillation; all progressed to sustained ventricular tachyarrhythmia, most with an actual cardiac arrest. This heterogeneity, in particular the presence of pacing, is relevant to the topological analysis, since artificial stimulation imposes an extrinsic periodicity on the signal that alters the recurrence metrics. Because the study depends on the annotation of the fibrillation onset instant (*vfon*), the 20 recordings that provide this marker were used, excluding the three recordings without a fibrillation episode. Of these, a restricted subset of 7 recordings with audited sinus beat annotations was used in the variability arm (Section II-G) (4 males, 3 females; median age 68 years, available for 5 of the 7; all presenting an underlying sinus rhythm in the SDDB documentation). As this is a retrospective collection, the clinical information on history and medication is limited and, in several cases, unavailable. To enable the analysis, the windowing of each recording covered the interval from 60 minutes before to approximately 5 minutes after the fibrillation onset, subdivided into three phases: the baseline phase, or *far* (from -60 to -30 min), which establishes the intra-patient reference band before the critical regime transition [12]; because every patient in this cohort progressed to the event, this band is a within-patient reference and not a healthy control; the approach phase, or *near* (from -30 to 0 min), immediately preceding the structural collapse (instant zero); and the post-collapse phase, or *post* (from 0 to $+5$ min), with ventricular fibrillation already established. Including the segment after the event allows the sustained fibrillation regime to be characterized and contrasted with both the baseline and the instant of the transition. The data are public and de-identified under the PhysioNet ODC-BY license.

B. Signal Pre-processing and Filtering

Before extracting the metrics, each recording was subjected to a pre-processing step to attenuate the noise and artifacts characteristic of Holter signals [13]. Because the SDDB recordings contain isolated invalid samples (NaN values), arising from the digitization of analog tape, which would propagate through the entire signal under recursive filtering, these samples were first linearly interpolated. A third-order Butterworth band-pass filter was then applied, with cutoff frequencies of 0.5 Hz and 40 Hz, implemented in zero phase (forward and backward filtering) so as not to introduce temporal distortion in the ECG complexes. The 0.5 to 40 Hz band removes baseline wander, below 0.5 Hz, and high-frequency and power-line noise, above 40 Hz, preserving the relevant morphological content of the heartbeat. The choice of this band was not by convention: it was validated through a comparative test among five pre-processing strategies (unfiltered signal, Butterworth 1–50 Hz, Butterworth 0.5–40 Hz, Butterworth 1–50 Hz with a 60 Hz *notch* filter, and only a 0.5 Hz high-pass), assessing which of them produced the most stable and robust extraction of the RQA metrics. The 0.5–40 Hz configuration was adopted for producing the most stable RQA extraction under this comparison, which was exploratory and did not use a pre-registered numerical criterion.

C. RQA Feature Extraction and Segmentation

To capture the nonlinear dynamics and the phase transitions of the cardiovascular system, the algorithm processed the signals through Recurrence Quantification Analysis (RQA) [9], [10]. The physiological dynamics were evaluated by means of a sliding-window method, allowing the continuous topological evolution of the system to be mapped over time. Fifteen fundamental metrics were extracted from each signal window: recurrence rate (RR), determinism (DET), mean and maximum diagonal lines (L, L_{max}), divergence (DIV), diagonal line entropy (L_{entr}), laminarity (LAM), trapping time (TT), maximum vertical lines and vertical entropy (V_{max}, V_{entr}), white-line features (W, W_{max}, W_{entr}), and combined ratios (DET_RR, LAM_DET). The recurrence matrix of each window was built directly from the scalar amplitude series, that is, with an embedding dimension $m = 1$, using the Euclidean distance; the diagonal and vertical structures were quantified with a minimum line length of $l_{min} = v_{min} = 2$. The RQA metrics were computed with an in-house implementation, validated against a pure-sinusoid control (which yields near-maximal determinism, as expected). The acquisition epochs of the sliding window were fixed at 5 seconds. The reduced sizing of this window was a methodological choice grounded

in the need to preserve the high temporal resolution of the signal immediately before the cardiac event. The literature on complex signal processing states that using long windows in systems close to collapse tends to mask fast fluctuations and abrupt transitions due to the mathematical averaging (smoothing) effect [10]. Thus, RQA is justified as the ideal tool for this scope because, unlike traditional linear and spectral metrics that require prolonged stability, it demonstrates high effectiveness and mathematical robustness in the analysis of extremely short non-stationary time series [14].

D. Data Structure, Collection, and Persistence

The processing pipeline persisted the data at two complementary levels, ensuring the traceability and reproducibility of the whole analysis. At the first level, each 5-second window had its recurrence matrix stored individually in binary `.npy` format, preserving the raw topology of the phase space (the point-by-point recurrence image) and allowing both the visual reconstruction of the Recurrence Plots and the later recomputation of any metric without the need to reprocess the original signal. At the second level, the 15 RQA descriptors extracted from each window were consolidated into manifest files in CSV format, generated independently for each tested threshold (p05 to p50). Thus, for each percentile of ϵ a complete manifest was obtained containing, per row, the 15 topological metrics accompanied by the identification metadata: the patient record, the index and the initial sample of the window, the time relative to the fibrillation onset (in seconds and minutes), the corresponding physiological phase (*far*, *near*, or *post*), and the name of the respective `.npy` file. This reference field links each tabular record to its stored recurrence matrix, so that every numerical value can be audited against its topological image of origin. Streaming computation without matrix materialization remains viable for embedded real-time use. Table I presents an excerpt of the manifest for $\epsilon = 40\%$ (p40). Among the 15 metrics, only Determinism (DET), the maximum diagonal (L_{max}), and Laminarity (LAM) were chosen for display. This selection is not arbitrary: it is the subset of descriptors from the *determinism and laminarity* family, which proved to be the most sensitive and discriminative for the phase transition, exhibiting the sharpest and most unequivocal inflection at the instant of collapse. While DET and LAM saturate near the upper limit and L_{max} jumps by an order of magnitude, the remaining metrics (entropies and vertical and white descriptors) show variations of smaller relative magnitude and are therefore less representative for compactly illustrating the signature of the event. The displayed instants span the trajectory up to collapse: -60 min (*far* baseline), -30 and -2 min (*near*), -0.1 min (immediately pre-event), and 0.0 min (*post*). The descriptors remain stable within the baseline band almost until the end and break down only in the final seconds.

TABLE I
EXCERPT OF THE MANIFEST (P40, RECORD 30): LINK BETWEEN EACH WINDOW, THE REPRESENTATIVE RQA METRICS, AND THE RECURRENCE MATRIX STORED IN `.NPY`.

t (min)	Phase	DET (%)	L_{max} (samples)	LAM (%)	Window
-60.0	far	99.80	433	99.95	`w0000`
-30.0	near	99.33	339	99.60	`w0360`
-2.0	near	98.77	210	99.36	`w0696`
-0.1	near	99.79	480	99.91	`w0719`
0.0	post	99.92	1249	99.96	`w0720`

`.npy`: `30_wXXXX_[BEFORE/AFTER]tt.tmin.npy`

E. Recurrence Threshold (ϵ) Definition and Noise Mitigation

Biological signals acquired in clinical environments are inherently noisy and subject to amplitude variations caused by patient movement or reading artifacts [13]. In Recurrence Quantification Analysis (RQA), the extraction of the metrics depends directly on the definition of a radius or distance threshold, mathematically represented by the Greek letter Epsilon (ϵ). Two points in phase space are considered recurrent only if the distance between them is smaller than this threshold. To ensure the robustness of the analysis and to prevent the baseline noise from generating false recurrences, the value of ϵ was not defined as an absolute constant. Instead, the algorithm computed ϵ in a relative and adaptive way, as a percentile of the positive pairwise distances of each analyzed window [10], a robust scaling choice in the sense of [15]. The system was parameterized to evaluate the behavior of the signal under different topological demands, testing threshold percentiles ranging from 5% to 50% (identified in the data as p05, p10, p20, p30, p40, and p50). This allows the optimal ϵ to be identified: a radius strict enough to map the complexity of the collapse, yet which scales with the signal's distance distribution to mitigate transient noise.

F. Statistical Normalization and Visual Trajectory Analysis

Due to the heterogeneity of scales and magnitudes of the 15 topological metrics extracted by RQA, a standardization step was necessary to enable the graphical and physiological comparison among them. For this, Z-Score normalization was applied [10], a standard baseline for deviation-based anomaly detection [16]. For each metric, the mean (μ) and standard deviation (σ) exclusive to the baseline phase ("Far") were computed. The values extracted during the continuous monitoring of the pre-collapse phase (X_{near}) were then normalized:

$$Z = \frac{X_{near} - \mu_{far}}{\sigma_{far}} \tag{1}$$

This standardization converts the signal trajectories to a common scale, defining how many deviations the system departs from its baseline equilibrium over time. Because the critical phase transition and the structural collapse of the cardiac dynamics occur abruptly in the final minutes or seconds of the recording, global statistical class-separation approaches (which compute averages over the entire pre-event period) tend to dilute and mask the anomaly peaks. For this reason, the identification of the pattern breakdown was not delegated to an automated statistical threshold. Instead, a retrospective approach based on visual and qualitative post-processing inspection was adopted. The normalized temporal trajectories

and the distributions (boxplots) of each metric were analyzed graphically to empirically determine the moment and the magnitude of the irreversible deviation of the signal from its baseline normality band. This approach ensures that the abrupt change generated by the collapse is clearly observed, preventing the use of long-term statistical averages from hiding or smoothing the peak of the event.

G. Recurrence of the Heart-Rate Variability Series

Because the amplitude analysis above quantifies the morphology of the beat but not the autonomic modulation of its rhythm, a second, complementary arm applied the same recurrence pipeline to the heart-rate variability (HRV) series of each recording. The beat annotations audited in the database were used to build the normal-to-normal (NN) interval series: an interval was retained only when both delimiting beats were of sinus origin, and the neighbors of any ectopic beat were discarded, so that the compensatory pause does not contaminate the tachogram. This requirement of audited sinus labels restricted the HRV arm to a subset of recordings, since HRV measures the autonomic modulation of the sinus node: recordings under atrial fibrillation or ventricular pacing, and those without an audited beat-annotation file, were excluded. Unlike the amplitude series, the NN series is a one-dimensional projection of a higher-dimensional autonomic dynamics, so the phase space was reconstructed by delay embedding (Takens) [17] before computing recurrence. For each recording, the embedding delay τ was set to the first local minimum of the average mutual information [18], and the embedding dimension m to the value at which the fraction of false nearest neighbors fell below 1% [19]; both parameters were estimated on the baseline (far) segment so as not to be shaped by the collapse. The RQA metrics were then extracted from the embedded series with a sliding window of 256 beats, a step of 32 beats, and the same distance-percentile threshold and minimum line lengths used in the amplitude arm. A quality gate required an NN density inside the window to exceed 80%, yielding an acceptance rate that ranged from 53% to 100% of the windows per recording. The gate's coverage remained highly stable over time, maintaining 84.7% of approved windows in the far phase and 83.9% in the near phase, demonstrating that the pre-event HRV analysis was not disproportionately affected by outcome-dependent ectopic rejection.

H. Statistical Comparison

For each descriptor, a single aggregated mean value per patient per phase was used to prevent pseudo-replication. The separation between the far and near windows was quantified using the paired Cohen's d (d_z). The effect sizes highlighted in the results are the maxima observed among the fifteen metrics; for the far/near tests they are additionally reported with correction for multiplicity (Holm and Benjamini–Hochberg). Furthermore, descriptors such as RR, DIV (the reciprocal of L_{max}), DET_RR, and LAM_DET are mathematically redundant under a constant recurrence rate framework and are evaluated for completeness only. Effect sizes are reported as the paired d_z over the 20 patients of the amplitude arm and the 7 of the variability arm, with 95% confidence intervals from 10,000 bootstrap resamples of patients.

III. Results

The evaluation of the results was conducted through visual and qualitative inspection, with the aim of demonstrating the breakdown of the resting dynamics (baseline phase) toward the structural collapse (Ventricular Fibrillation). All metrics discussed in this section were extracted with 5-second acquisition windows and an adaptive recurrence threshold fixed at $\epsilon = 40\%$ of the positive pairwise distances (identified as *p40*). This parameterization was selected, among the tested factors (p05 to p50), for offering the best trade-off between sensitivity to the phase transition and robustness to baseline noise: smaller percentiles severely fragmented the diagonal structures due to noise, while larger percentiles saturated the matrix and masked the pre-collapse topological inflection. This selection was made on the same data reported below and is therefore exploratory.

A. Qualitative Analysis of the Recurrence Plots (RP)

The phase transition of the cardiovascular system can be visualized directly through the Recurrence Plots computed from the ECG (Fig. 1). The analysis of three distinct instants reveals the topological reorganization of the signal along the physiological degradation:

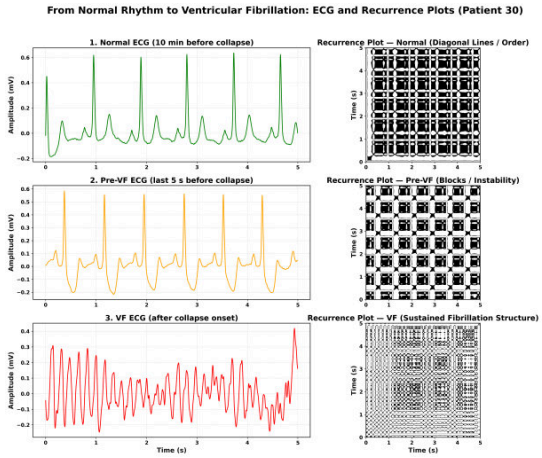


Fig. 1. ECG signals and their Recurrence Plots in three physiological states: Baseline (-10 min), Pre-collapse (-5 s), and Ventricular Fibrillation (post-collapse).

- **Normal Phase (-10 minutes):** The signal shows well-defined and periodic QRS complexes. The matrix presents an organized structure, with long parallel diagonal lines and regularly distributed recurrences, characterizing a deterministic system rich in morphological detail (P-QRS-T).
- **Pre-collapse Phase (-5 seconds):** The morphology of the heartbeat begins to deform and the topology reorganizes into dense blocks, with partial fragmentation of the diagonals and an increase of laminar (vertical) patterns,

reflecting the progressive loss of fine regulation and the growth of trapped states.

- **Ventricular Fibrillation (Post-collapse):** The fine structure of the heartbeat disappears entirely and the signal degenerates into a large-amplitude oscillation with low morphological variability. Although visually the matrix takes on a dense and less ordered appearance, the quantitative analysis with $\epsilon = 40\%$ reveals that this regime does not correspond to random noise: the dominant oscillation generates extensive long-range recurrent structures (long diagonals and laminar blocks), which translates into the marked increase of DET, L_{max}, and LAM. It is, therefore, a paradoxically *more deterministic* yet structurally impoverished state, the result of replacing the multiscale repertoire of the healthy ECG with a single oscillatory mode, and not of an increase in randomness. To test whether this pre-collapse deformation is simply sustained ventricular tachycardia (VT), the rhythm annotations (`.ari` and `.atr`) were queried in the $[-30\text{ s}, 0)$ interval of the 18 recordings for which an annotation file was available (two of the 20 amplitude-arm records lacked a usable annotation file and were excluded from this check). No VT episode is labelled in any of them. A heart rate could be computed over that interval for 9 of these recordings, with a median of 74 bpm (range 66–109); two recordings exceeded 100 bpm, consistent with sinus tachycardia rather than sustained VT in the absence of wide-complex labelling. Because rhythm labelling in this database is not exhaustive, the absence of a VT label is not conclusive, but the available evidence does not support interpreting the topological breakdown as a transition from sustained VT.

B. Temporal Trajectory of the RQA Metrics

To quantify the dynamic behavior, the 15 RQA metrics were extracted in continuous 5-second windows over the interval from 60 minutes before to 5 minutes after the event (Fig. 2). Observation of the trajectories shows that, throughout the entire baseline phase (from -60 min to approximately -1 min), the system oscillates within a narrow and stationary band, with no systematic trend. The recurrence rate (RR) remains essentially constant (≈ 0.4005) throughout the whole recording, which is an expected mathematical consequence of the chosen percentile threshold, effectively isolating structural degradation metrics from recurrence density variations. Upon reaching the zero instant, practically all metrics undergo an abrupt and simultaneous inflection, constituting an unmistakable signature of the collapse. A coherent bidirectional behavior is observed: the metrics associated with periodicity and long-term predictability surge, while the metrics associated with structural richness and complexity plummet. Determinism (DET) jumps from a baseline level of ≈ 98.6 to ≈ 99.55, and the maximum diagonal (L_{max}) triples, going from ≈ 360 to values above 1000; consequently, Divergence (DIV $= 1/L_{max}$) collapses from ≈ 0.0035 to ≈ 0.0015. Laminarity (LAM) rises from ≈ 99.3 to ≈ 99.8, and the DET_RR ratio follows the same upward direction. Conversely, the complexity descriptors fall sharply: the diagonal entropy (L_{entr}), the vertical entropy (V_{entr}), and the white-line entropy (W_{entr}) recede simultaneously, as do the mean diagonal length (L), the trapping time (TT), the maximum vertical line (V_{max}), and the white-line descriptors (W, W_{max}). The drop in LAM_DET reinforces this picture. This combination, an increase in global determinism accompanied by a generalized loss of entropy, reveals the physiological nature of the collapse: the rich multiscale repertoire of the healthy ECG is replaced by a single dominant, impoverished oscillatory mode. In dynamical terms, the system becomes more predictable, yet pathologically simplified, losing the variability that characterizes the resting baseline. After the abrupt peak at the zero instant, the metrics do not return to baseline but stabilize into a new regime corresponding to sustained fibrillation: determinism settles above its baseline level, and the mirrored descriptors (DIV, V_{entr}, W) re-stabilize after a trough at the collapse. Fibrillation, once established, is thus itself a stationary but pathological rhythm, topologically distinct from the baseline. The abrupt inflection is confined to the onset window itself; a prospective evaluation of this departure, through a per-patient detector with its operating characteristics, is presented in Section III-F. Under a conservative criterion (metric consensus and persistence) no recording is flagged before the onset, and no operating point yields anticipation with an acceptable false-alarm rate.

C. Evolution of the Distributions and Pattern Breakdown

The breakdown of the baseline limit was confirmed by the evolution of the metric distributions at six reference instants (-60 min, -30 min, -10 min, -2 min, -30 s, and collapse), as shown in Fig. 3. The boxplots are consistent with the behavior observed in the trajectories. Between -60 min and -30 s, the medians of all metrics remain close and the boxes largely overlap, delimiting a band within which no shift is apparent, including in the -2 min and -30 s windows, which indicates that the degradation is not slowly progressive but rather concentrated in the final seconds. It is only in the collapse column that a severe and unequivocal shift of the medians occurs: L_{max} and DET_RR migrate to the top of their scales, DET and LAM approach saturation, while DIV, L, TT, V_{max}, the entropies, and LAM_DET shift downward, with a simultaneous reduction of the interquartile dispersion. This sharp visual separation between the pre-collapse block and the collapse window indicates that the structural anomaly is abrupt rather than gradual, which justifies the retrospective inspection strategy adopted: approaches based on global averages over the entire pre-event period would dilute this peak, whereas the per-instant analysis preserves it. From an applied standpoint, the result indicates that an observation window on the order of seconds, combined with the most responsive metrics (L_{max}, DIV, DET, and LAM), is sufficient to characterize the breakdown of the baseline pattern with $\epsilon = 40\%$.

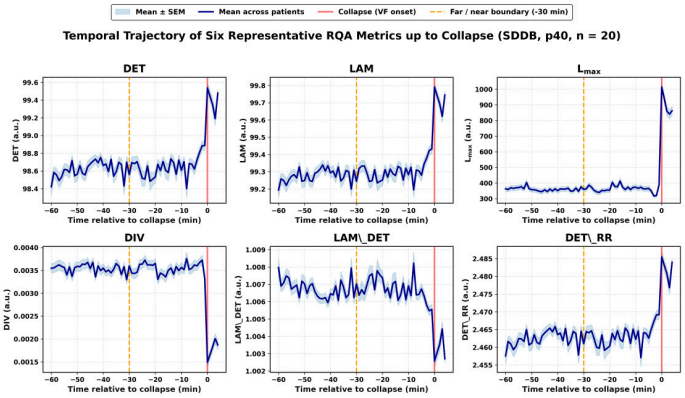


Fig. 2. Temporal trajectories of six representative RQA metrics (DET, LAM, L_{max}, DIV, LAM_DET, DET_RR) from 60 min before to 5 min after the collapse (SDDB, $\epsilon = 40\%$, $n = 20$). The full 15-metric version is available in the supplementary material. The dashed orange line marks -30 min and the red line the instant of collapse (0).

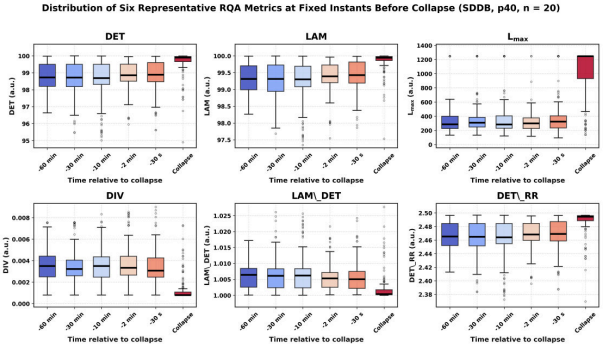


Fig. 3. Statistical distribution (boxplots) of six representative RQA metrics per instant up to the collapse (SDDB, $\epsilon = 40\%$, $n = 20$).

D. Predictive Test in the Amplitude Domain

To assess whether the recurrence metrics anticipate the collapse, rather than merely detect it once installed, the analysis was restricted to the pre-event interval and the *far* (-60 to -30 min) and *near* (-30 to 0 min) windows were compared. To avoid pseudo-replication, a single value per patient per window was used (the mean of that patient's windows), and the effect size was quantified with Cohen's d. In the amplitude domain, when evaluating the grouped paired differences across phases, the effect size for V_{max} yielded $d_z = -0.141$ (95% CI $[-0.594, 0.348]$ via 10,000 bootstrap resamples). With correction for multiplicity across the 15 metrics (Holm and Benjamini–Hochberg), no metric separates far from near: the largest effect size is that of TT ($d_z = -0.20$; $p_{\text{Holm}} = 1.0$; 0/15 survive). The confidence interval crossing zero reflects the precision achievable at this sample size, and no group-level threshold separates the two windows. The result is further insensitive to the uncertainty of the onset annotation: shifting the onset by -15 to $+15$ s (± 3 windows), the largest $|d_z|$ remains between 0.20 and 0.22. This absence of a uniform anticipatory trend is consistent with the flat baseline trajectories of Fig. 2 and is robust to pre-processing, since all five filtering strategies of Section II-B produced the same result. The amplitude-based recurrence of the ECG therefore characterizes the collapse but does not anticipate it over the minutes that precede it.

E. Domain Contrast with Heart-Rate Variability

Applying the same test to the recurrence of the HRV series changes the picture. The HRV arm comprised the seven eligible recordings (of the 23; the remaining sixteen were excluded for atrial fibrillation, ventricular pacing, or the absence of an audited beat-annotation file). The sinus-eligibility requirement and the 80% NN-density gate yielded 886 analysis windows, fewer than in the amplitude arm because each window here spans 256 beats and only clean sinus segments are retained; the embedding parameters were estimated per recording on the baseline segment. Under the same *far*/*near*, per-patient, Cohen's d comparison, the largest effect in the Takens-embedded HRV ($m > 1$) was $d_z = 0.63$ (V_{max}) and the raw $m = 1$ amplitude restricted to the exact same seven recordings reached $d_z = 0.41$ (LAM), both shown in Fig. 4; a linear HRV descriptor (SDNN), reported here in text only and not plotted, reached a comparable $d_z = 0.41$. With multiplicity correction (Holm), none of these effects is significant ($p_{\text{Holm}} = 1.0$ across all domains). The difference between domains is therefore not robust: the somewhat larger raw effect of the embedded HRV does not survive correction, and the linear HRV baseline (0.41) matches the amplitude-restricted control (0.41), indicating that the apparent autonomic advantage reflects the composition of

the seven-patient subsample rather than an intrinsic property of the variability domain.

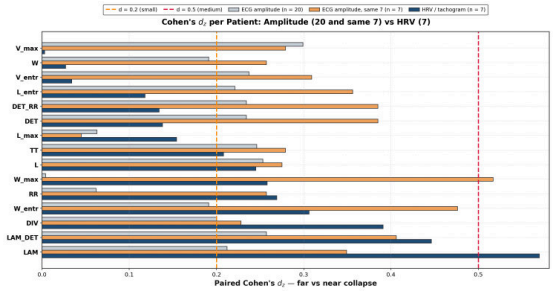


Fig. 4. Domain contrast of the predictive effect size (per-patient Cohen's d_z, *far* vs. *near*) for the 15 RQA metrics, computed on the raw ECG amplitude over all 20 recordings ($m = 1$), on the amplitude restricted to the same seven recordings as the variability arm, and on the Takens-embedded HRV series ($m > 1$). Dashed lines mark the small- ($d = 0.2$) and medium-effect ($d = 0.5$) references. The anticipatory separation is small in both domains and no effect survives multiplicity correction; the apparent advantage of the variability domain is not maintained when the amplitude arm is restricted to the same seven recordings (Section III-E). A linear HRV baseline, reported in the text, is not plotted here.

F. Operating Characteristics of a Per-Patient Detector

To replace visual inspection with a prospective criterion, an unsupervised per-patient detector was evaluated: the mean (μ) and standard deviation (σ) of DET, L_{max}, and LAM were estimated on the *far* phase, and each 5 s window was flagged when it exceeded $\mu + k\sigma$. Because the recurrence metrics are heavy-tailed (0.52% of the baseline windows themselves already exceed $\mu + 3\sigma$, versus 0.13% under normality), a single-window threshold produces a high false-alarm rate and flags pre-event windows indistinguishable from baseline noise (Table II). At 3σ per single metric the sensitivity at onset is 18/20, at the cost of 28.7 false alarms per hour; a consensus of two of the three metrics with a persistence of three windows at 4σ eliminates the false alarms but detects the onset in only 6/20 recordings and flags no window before it. There is therefore no operating point that yields anticipation with an acceptable false-alarm rate. Under leakage-free selection by

TABLE II
OPERATING CHARACTERISTICS OF THE PER-PATIENT DETECTOR (AMPLITUDE, P40, $n = 20$). FA/H: FALSE ALARMS PER HOUR IN THE PRE-EVENT PHASE.

Configuration	Sens.	FA/h	Pre-event win.
3σ, single metric	18/20	28.7	19/20 (false)
4σ, single metric	17/20	14.4	14/20 (false)
5σ, single metric	17/20	7.2	14/20 (false)
4σ, consensus 2/3, 3 win.	6/20	0.0	0/20

leave-one-patient-out validation (percentile, k, consensus, and persistence chosen on the development patients and evaluated on the held-out patient), the out-of-fold sensitivity was 75% (15/20), still at 9.4 false alarms per hour and with (false) pre-event flagging in 16/20 patients, confirming the absence of generalizable anticipation. As a negative control, the same detector applied to recordings that do not progress to fibrillation in the analyzed segment fired at 3.4 false alarms per hour (recordings without VF or with pacing) and at 15.0/h in recordings with atrial fibrillation, a rate comparable to that observed in the pre-event phase of the fibrillation recordings (28.7/h). The alarms are therefore not specific to an impending fibrillation.

IV. DISCUSSION

The analysis shows that ventricular fibrillation is preceded by an abrupt topological transition at its onset, perceptible only in the seconds preceding the collapse. The term *pre-fibrillation* is used here in a descriptive, post-hoc sense: the trajectory inspection was retrospective, the conservative per-patient detector (Section III-F) flags no window before the onset, and the negative controls fire at a comparable rate, so the alarms are not specific to the event. Already at -5 s (Fig. 1), the recurrence matrix loses the fine structure of the heartbeat and reorganizes into blocks, while the determinism and laminarity metrics begin the inflection that saturates at the instant of the event (Fig. 2 and Fig. 3). This is not a long-term warning *in the amplitude domain*, since there is no detectable deviation tens of minutes earlier (from -60 min to approximately the last minute, the 15 metrics oscillate within a narrow and stationary baseline band), but a sharp transition on the scale of seconds. Because the change concentrates in the last seconds, block-average separation tools such as Cohen's d [20] are inadequate in the amplitude domain: averaging tens of minutes of baseline dilutes a few-second deviation (Section III-D, $d_z = -0.141$; no metric survives multiplicity correction). The breakdown was therefore identified by visual inspection of the normalized trajectories rather than by a global threshold. In the HRV domain the largest far/near effect ($d_z = 0.63$) in the Takens-embedded NN series exceeds the $m = 1$ amplitude, but it does not survive multiplicity correction, and a linear HRV descriptor matches the amplitude restricted to the same seven patients ($d_z \approx 0.41$). The anticipatory capacity thus depends heavily on patient stratification and low ectopy, underscoring the difficulty of patient-independent prediction. Based on these findings, an open question for future work is whether any method can identify, reliably and earlier than the final seconds, the instant at which the signal leaves the patient's own baseline band. Even though the onset signature is confined to the final seconds, an automatic and reliable identification of that instant would still have clinical value, given that the reversal of a cardiorespiratory arrest depends on intervention within a few minutes [2]. Several fronts could consolidate this agenda: time-localized change-point detection against the patient's own baseline band rather than block averages; identification of which descriptor or combination separates earliest and with fewest false positives; a multivariate patient-independent classifier extending the detector characterized here; a sweep of the threshold percentile and of window duration; extension of the variability pipeline to the excluded paced and atrial-fibrillation recordings as a robustness analysis; and validation against healthy controls and additional databases. In summary, recurrence descriptors characterize the onset sharply but did not anticipate it over the minutes preceding it, in either domain, once the sample was held fixed; whether any per-

patient criterion can identify the transition earlier remains open.

A. Limitations

The amplitude arm was performed with an embedding dimension $m = 1$ and without a Theiler window, which causes metrics like DET and LAM to operate near saturation due to tangential motion artifacts. The sample size of 7 for HRV is small. The variability arm also depends on audited beat annotations, which do not exist in real time: a deployed equivalent would require a validated QRS detector and ectopy classifier, whose errors would propagate into the tachogram and hence into the recurrence descriptors. Furthermore, the dataset consists of 1980s analog Holter recordings, limiting generalization to modern digital front-ends. Finally, the operating point and the percentile of the main arm were, in part, selected on the same evaluation data, with leave-one-patient-out validation conducted only for the amplitude detector; effect sizes are reported with multiplicity correction (Holm and Benjamini–Hochberg) and no metric survives it in either domain, the seven-patient variability arm remaining, in particular, exploratory.

V. Conclusion

This work applied Recurrence Quantification Analysis to SDDB electrocardiographic signals, extracting 15 topological metrics in 5-second windows under an adaptive percentile threshold and persisting both the recurrence matrices and per-window metrics for full traceability. The parameterization $\epsilon = 40\%$ (p40) best revealed the phase transition. Ventricular fibrillation exhibits an abrupt topological transition at its onset, in which the determinism and laminarity family (DET, LAM, L_{max}, DIV) undergoes a sharp inflection; characterized retrospectively, this signature is not by itself a prospectively detectable precursor. It is concentrated in the final seconds rather than the preceding minutes, in which the amplitude signal stays within its baseline band ($d_z = -0.141$, 95% CI $[-0.594,\ 0.348]$); global separation tools are therefore inadequate at this short scale, and visual inspection best captured the breakdown. Applying the same pipeline over a Takens-reconstructed HRV phase space produced a larger effect ($d_z = 0.63$) that does not survive multiplicity correction; a linear HRV descriptor and the amplitude restricted to the same seven patients yielded comparable effects ($d_z \approx 0.41$), so no domain sustains a significant anticipatory separator. The contribution is thus a well-characterized negative result for generalized long-term prediction, clarifying the boundaries and the patient stratification required for short-scale monitoring. The bootstrap intervals are wide at this sample size, so neither domain supports a group-level anticipatory threshold; whether a per-patient adaptive baseline performs better remains to be tested on a larger cohort.

Acknowledgment

The SDDB is publicly available from PhysioNet [11], [22]; the processing pipeline, manifests, and analysis scripts are available at https://doi.org/10.5281/zenodo.22654401. The authors used an artificial intelligence language model (Anthropic Claude) to assist with the translation of this manuscript from Portuguese into English and with language editing; all AI-assisted content was reviewed and verified by the authors, who take full responsibility for the integrity and accuracy of the work.

References

[1] R. Mehra, "Global public health problem of sudden cardiac death," *Journal of Electrocardiology*, vol. 40, no. 6, pp. S118–S122, 2007.

[2] A. R. Panchal *et al.*, "Part 3: 2020 AHA Guidelines for CPR and Emergency Cardiovascular Care," *Circulation*, vol. 142, no. 16_suppl_2, pp. S366–S468, 2020.

[3] A. Amann, R. Tratnig, and K. Unterkofler, "Reliability of old and new ventricular fibrillation detection algorithms for automated external defibrillators," *Biomed. Eng. Online*, vol. 4, p. 60, 2005.

[4] Q. Li, C. Rajagopalan, and G. D. Clifford, "Ventricular fibrillation and tachycardia classification using a machine learning approach," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1607–1613, 2014.

[5] G. T. Taye, E. B. Shim, H.-J. Hwang, and K. M. Lim, "Machine learning approach to predict ventricular fibrillation based on QRS complex shape," *Front. Physiol.*, vol. 10, art. 1193, 2019.

[6] D. U. Jeong, G. T. Taye, H.-J. Hwang, and K. M. Lim, "Optimal length of heart rate variability data and forecasting time for ventricular fibrillation prediction using machine learning," *Comput. Math. Methods Med.*, vol. 2021, art. 6663996, 2021.

[7] A. L. Goldberger *et al.*, "Fractal dynamics in physiology: Alterations with disease and aging," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. suppl_1, pp. 2466–2472, 2002.

[8] A. Voss, S. Schulz, R. Schroeder, M. Baumert, and P. Caminal, "Methods derived from nonlinear dynamics for analysing heart rate variability," *Philos. Trans. R. Soc. A.*, vol. 367, pp. 277–296, 2009.

[9] C. L. Webber Jr and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," *J. Appl. Physiol.*, vol. 76, no. 2, pp. 965–973, 1994.

[10] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Phys. Rep.*, vol. 438, no. 5-6, pp. 237–329, 2007.

[11] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[12] M. Scheffer *et al.*, "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, pp. 53–59, 2009.

[13] G. D. Clifford, "ECG statistics, noise, artifacts, and missing data," in *Advanced Methods and Tools for ECG Data Analysis*, Norwood, MA, USA: Artech House, 2006, pp. 55–99.

[14] J. P. Zbilut, A. Giuliani, and C. L. Webber Jr, "Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification," *Physics Letters A*, vol. 246, no. 1-2, pp. 122–128, 1998.

[15] P. J. Huber, *Robust Statistics*. New York, NY, USA: John Wiley & Sons, 1981.

[16] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[17] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics, vol. 898, D. A. Rand and L.-S. Young, Eds. Berlin, Germany: Springer-Verlag, 1981, pp. 366–381.

[18] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[19] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol. 45, no. 6, pp. 3403–3411, 1992.

[20] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[21] S. D. Greenwald, "Development and analysis of a ventricular fibrillation detector," M.S. thesis, MIT Dept. of Elect. Eng. and Comput. Sci., Cambridge, MA, USA, 1986.

[22] G. B. Moody, R. G. Mark, and A. L. Goldberger, "PhysioNet: A web-based resource for the study of physiologic signals," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 70–75, 2001.

Exploratory Evaluation of CNN Architectures for Glaucoma Classification from Retinal Fundus Images

1st Fabiano Araujo Soares
Faculty of Sciences and Technology in Engineering
Universidade de Brasília
Brasília, Brazil
fabianosoares@unb.br

2nd Lorenzo de Lima Alves dos Santos
Faculty of Sciences and Technology in Engineering
Universidade de Brasília
Brasília, Brazil
lorenzolimaloren@gmail.com

5th Mario Fleury
Undergraduate Mathematics Department
Boston University
mariorf@bu.edu

4th Suelia de Siqueira Rodrigues Fleury Rosa
Meinig School of Biomedical Engineering
Cornell Duffield Engineering
New York, USA.
sdf73@cornell.edu

3rd Adson Ferreira da Rocha
Dept. of Electrical Engineering
Universidade de Brasília
Brasília, Brazil
adsonr@gmail.com

***Abstract*—Early detection of glaucoma is essential to prevent irreversible vision loss, yet reliable screening remains challenging due to the subtle nature of early structural changes. This work presents an exploratory evaluation of convolutional neural network (CNN) architectures for binary glaucoma classification from retinal fundus photographs under limited data conditions. Two widely adopted models, VGG16 and ResNet50, were implemented using transfer learning and trained on a curated subset of the Ocular Disease Recognition (ODIR) dataset. Model performance was assessed using accuracy, F1-score, sensitivity, specificity, precision, negative predictive value, ROC-AUC score, variance, and confusion matrices. Experimental results indicate that VGG16 outperformed ResNet50 on the considered dataset. Given the limited sample size and the absence of external validation, the findings should be interpreted as a preliminary baseline comparison rather than evidence of clinical readiness or generalizability. Following the reviewer's suggestions, the metrics ROC-AUC, sensitivity, and specificity were added. The VGG16 model presented an ROC-AUC score of 0.7266, 79.07% for the sensitivity, and 62.16% for specificity, while the ResNet50 model showed a ROC-AUC score of 0.6568, 58.14% for the sensitivity and 62.16% for the specificity.**

***Index Terms*—CNN; Fundus; Glaucoma; Artificial Intelligence; VGG16; ResNet50**

I. Introduction

Glaucoma is a group of progressive optic neuropathies characterized by the degeneration of retinal ganglion cells and damage to the optic nerve head, often associated with elevated intraocular pressure (IOP). This damage leads to irreversible vision loss and is a leading cause of blindness worldwide. The disease is typically asymptomatic in early stages, making early detection critical to prevent significant visual impairment. Various forms of glaucoma exist, including primary open-angle glaucoma and angle-closure glaucoma, each with distinct pathophysiological mechanisms [1], [2].

The causes of glaucoma are multifactorial, encompassing both genetic predispositions and environmental influences. Genetic factors significantly contribute to early-onset glaucoma, such as congenital forms inherited through mutations in specific genes like CYP1B1 and LTBP2, which are typically inherited in an autosomal recessive pattern. In contrast, adult-onset glaucoma, including primary open-angle glaucoma (POAG), exhibits complex inheritance involving multiple genes interacting with environmental factors. Environmental influences, such as elevated intraocular pressure (IOP), lifestyle choices, exposure to toxins, and systemic conditions like diabetes, can modulate glaucoma risk and progression. The interplay between genetics and environment, including gene-environment interactions, is increasingly recognized as critical for understanding individual susceptibility and disease severity, paving the way for personalized prevention and treatment strategies [3]–[5].

Figure 1 shows representative ODIR retinal fundus photographs labeled as normal and glaucoma. In cases such as those illustrated in Figure 1, retinal abnormalities are often apparent; however, differentiating glaucoma from other conditions such as diabetes or cataracts through retinal images can be challenging and costly if relying solely on expert manual analysis [7].

While glaucoma is often associated with increased intraocular pressure, it is incorrect to assert that elevated pressure is the direct cause [8]. The eye contains a transparent fluid known as aqueous humor, which drains through the Schlemm's canal and the uveoscleral pathway [9]. Impaired drainage causes increased intraocular pressure, which can damage the optic nerve and lead to glaucoma [2]. Diagnosis frequently involves tonometry, which measures intraocular pressure, and Optical Coherence Tomography (OCT), which provides detailed cross-

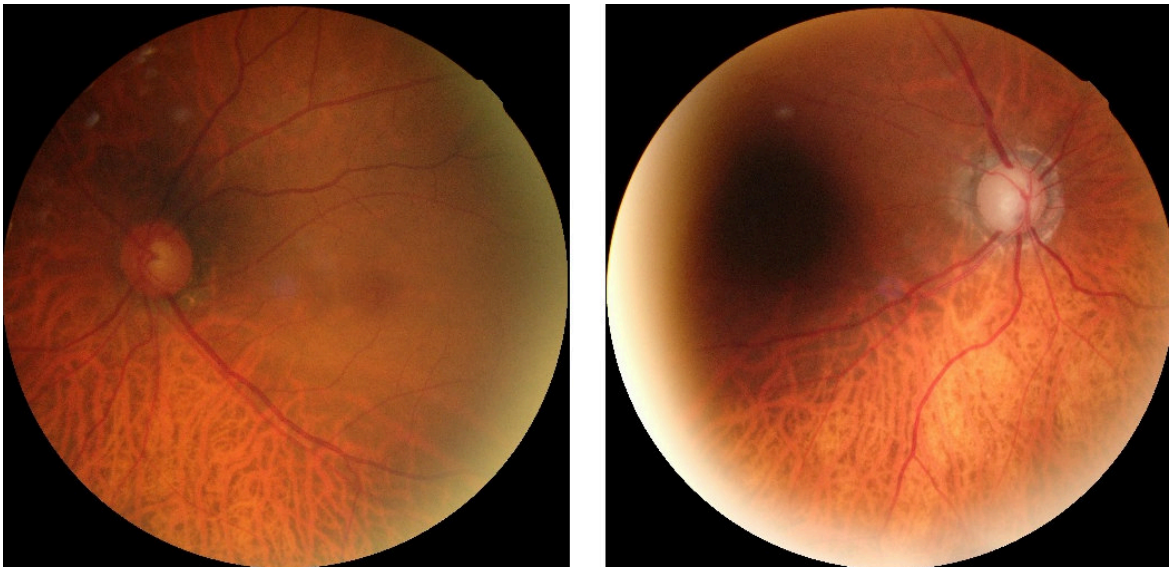


Fig. 1. Comparison between a normal eye (left) and an eye affected by glaucoma (right). Source: Ocular Disease Recognition Dataset [6].

sectional information about retinal structures [10]. OCT is mentioned here as a clinical imaging modality; however, the ODIR images analyzed in the present study are retinal fundus photographs.

Artificial neural networks are computational models within the field of machine learning, which focuses on teaching machines to learn from data without explicit programming [11]. ANNs are particularly powerful in identifying nonlinear patterns in complex datasets, such as recognizing features in images by analyzing pixel data. The fundamental processing unit of an ANN is the neuron, analogous to biological neurons, which receive, transform, and transmit information within layers of the network [12].

Neurons in a network carry different informational weights, and biases are added to increase computational complexity. These weights and biases have initially arbitrary values, which are iteratively adjusted during training — a process of learning from labeled examples to improve the model's predictive accuracy without overfitting [11], [13].

ANNs and CNNs have been widely investigated for medical image classification, including ophthalmic imaging [14]. Nevertheless, performance reported in a limited single-dataset experiment should not be interpreted as evidence of clinical screening capability without external validation and further prospective evaluation.

Convolutional Neural Networks (CNNs) are a specialized type of ANN designed to process and analyze visual data. Distinguished by their use of convolutional layers, CNNs automatically learn hierarchical feature representations from raw image inputs, making them highly effective for image classification tasks [12]. In medical imaging, CNNs have demonstrated superior performance in detecting and classifying conditions by capturing subtle patterns and structures that may be difficult for human observers to discern [15]. Among the various CNN architectures, VGG16 [16] and ResNet50 [17] are widely recognized for their effectiveness and have been extensively applied in medical image analysis [18]. VGG16 employs a deep yet straightforward design with small convolutional filters to capture fine-grained features, while ResNet50 introduces residual learning to overcome the challenges of training very deep networks, enabling improved accuracy and convergence. In this study, they are used as established baseline architectures for comparison on retinal fundus photographs [16], [17], [19].

METHODS

Methodology

The methodology adopted for this study was exploratory, combining comprehensive bibliographic research with practical experimentation involving established deep learning architectures. Specifically, two convolutional neural network (CNN) architectures — VGG16 and ResNet50 — were selected due to their proven efficacy in image classification tasks. Both models were trained using the publicly available Ocular Disease Recognition (ODIR) dataset, which includes retinal images labeled for various ocular conditions, including glaucoma [6].

To compare the architectures' performances in glaucoma detection, we employed multiple evaluation metrics: Accuracy, F1-Score, Variance, and confusion matrices. The revised evaluation is being expanded to include ROC-AUC, sensitivity, specificity, precision, and negative predictive value, as requested by the reviewers. These metrics provide an initial comparison of the two architectures on the selected ODIR subset; the study does not establish clinical applicability or external generalizability. For consistency with the actual ODIR image modality, the inputs analyzed in this work are referred to throughout as retinal fundus photographs or fundus images.

Sample and Dataset

A subset of the Ocular Disease Recognition (ODIR) dataset [6], comprising 512×512 pixel retinal fundus photographs, was used. Only images labeled as either "glaucoma" or "normal" without comorbidities were included, resulting in a balanced set of 200 images per class. The combined 400 images were randomly shuffled. No data augmentation techniques were applied in the original experiment. All images used in this study were obtained from distinct patients, and no pair of left- and right-eye images originated from the same patient.

Because patient-level separation was not documented in the original submission, the revised manuscript explicitly identifies the split procedure as a limitation until patient identifiers and the original allocation are verified. A patient-level split is recommended for future validation if the original experiment used image-level randomization.

Model Architectures

This study evaluated two widely recognized convolutional neural network (CNN) architectures, VGG16 and ResNet50, chosen for their established effectiveness in image classification and medical image analysis tasks.

- **VGG16**: This architecture [16] features sequential convolutional blocks composed of small 3×3 convolutional filters followed by max pooling layers, culminating in three fully connected dense layers. A main advantage of VGG16 is its simplicity and uniform design, which make it easy to understand and implement. It effectively captures fine-grained image features, performing well especially on moderate-sized datasets. However, VGG16 has a relatively large number of parameters ($\sim$138 million), which creates high computational and memory demands and can lead to longer training times and overfitting on smaller datasets.
- **ResNet50**: Introduced by He et al. [17], ResNet50 mitigates the degradation and vanishing gradient problems experienced in deep networks by employing residual learning via shortcut connections. These residual blocks facilitate training of deeper architectures by allowing gradients to propagate more effectively. ResNet50 typically offers faster convergence and better generalization on large and complex datasets. Nonetheless, its more complex architecture may require more extensive tuning and can be prone to overfitting in scenarios with limited data.

Both architectures are presented in Figure 2 and were implemented using TensorFlow/Keras and trained in Google Colab with GPU acceleration, leveraging pretrained ImageNet weights. In the original experiment, the ImageNet pretrained convolutional backbones were used as frozen feature extractors, while only the final classifier layers were trained on the selected ODIR subset.

Implementation and Experimental Details

Models were implemented in TensorFlow/Keras on Google Colab with an NVIDIA T4 GPU (availability limitations circumvented via multiple accounts). Images (512×512 pixels) were normalized by scaling pixel values to [0,1] via division by 255.0. No resizing or data augmentation was applied. From ODIR-5K [6], 400 fundus images were selected: 200 Normal (100 left/100 right eyes) and 200 Glaucoma (104 left/96 right), randomly shuffled (seed 42), split into 320 images for training and 80 images for validation.

ImageNet-pretrained convolutional layers were frozen; only the top dense layers (256 neurons, ReLU, Dropout 0.5, softmax output; 2 classes) were trained using:

- Adam optimizer with a standard learning rate of 0.001 and standard hyperparameters;
- Categorical cross-entropy loss;
- A batch size of 32 images;
- a 256-neuron fully connected layer with ReLU activation;

VGG16 trained for 60 epochs and ResNet50 for 30 epochs in the original script. The model was implemented using TensorFlow/Keras (TensorFlow 2.20.0), with NumPy 2.1.3 for numerical computations, scikit-learn 1.6.1 for performance evaluation, and Matplotlib 3.10.0 for visualization. Training was performed using GPU acceleration, with an NVIDIA GPU detected by TensorFlow.

The original protocol did not include partial convolutional fine-tuning, data augmentation, scheduler, early stopping, repeated-seed evaluation, cross-validation, or best-weight restoration. These omissions are now stated explicitly as methodological limitations and are identified as priorities for future validation.

Training and Validation

Images were standardized for input size and color. Categorical labels were encoded for binary classification. Epoch counts were adjusted based on the observed training behavior; a comprehensive hyperparameter optimization procedure was not performed in the original experiment.

Performance Metrics

To evaluate and compare the performance of the proposed models, eight metrics were employed:

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **F1-Score:**

$$F_1 = \frac{2VP^2}{VP(2VP + FP + FN)}$$

- **Sensitivity:**

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- **Specificity:**

$$\text{Specificity} = \frac{TN}{TN + FP}$$

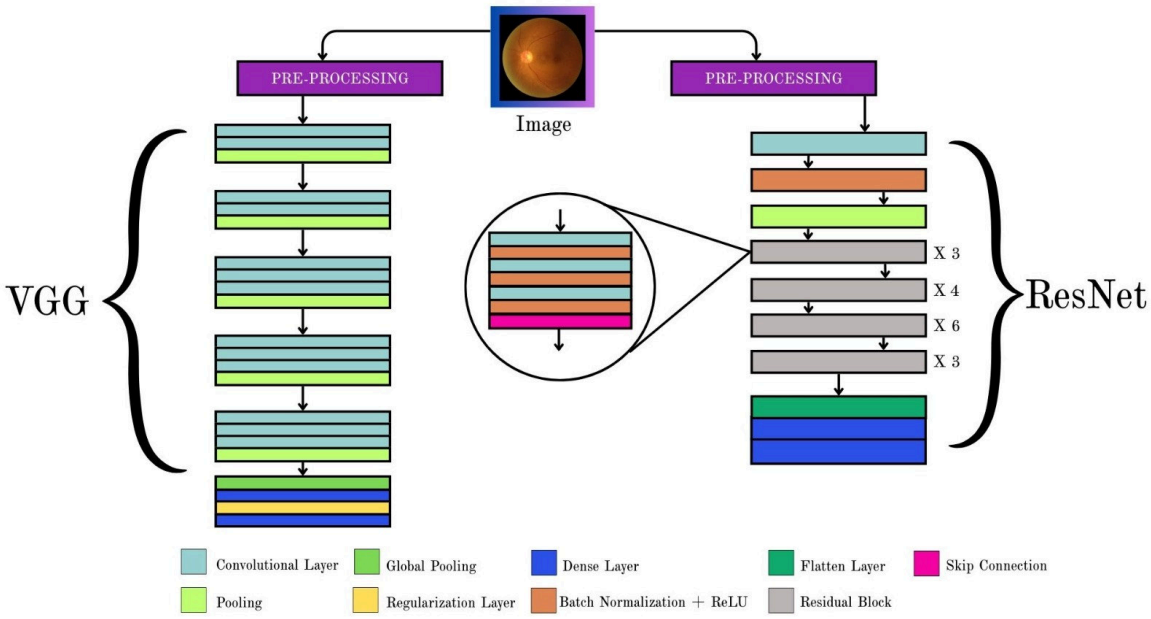


Fig. 2. Comparison between VGG [16] and ResNet [17] convolutional neural network (CNN) architectures used for medical image analysis. The figure illustrates the main processing steps, including pre-processing, convolutional and pooling layers, and the architectural differences—such as the use of regularization, batch normalization, skip connections, and residual blocks. The distinct blocks and color-coded elements highlight how each architecture handles feature extraction and information flow, supporting applications in medical imaging tasks such as disease detection and classification.

- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP}$$
- **Negative Predictive Value:**
$$\text{NPV} = \frac{TN}{TN + FN}$$
- **ROC-AUC:**The area under the receiver operating characteristic curve should be calculated from model probability scores rather than from the confusion matrix alone.
- **Variance:**
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x - \mu)^2$$
In the revised presentation, emphasis is placed on clinically interpretable classification metrics requested by the reviewers. Variance is retained only as a descriptive metric from the original analysis and should not be interpreted as a substitute for uncertainty estimates or repeated-run evaluation.
- **Confusion Matrix:** A detailed visualization of the classification results, the confusion matrix depicts counts of true positives (images with glaucoma classified as ”glaucoma”), true negatives (normal images classified as ”normal”), false positives (normal images classified as ”glaucoma”), and false negatives (images with glaucoma classified as ”normal”), enabling a comprehensive error analysis.

These metrics together provide a thorough assessment of the diagnostic accuracy, reliability, and potential clinical applicability of the trained deep learning models.

Results

The neural network architectures were implemented and tested. Table I summarizes the performance metrics obtained for the VGG16 and ResNet50 models on the test dataset.

The VGG16 model achieved an accuracy on average of 71.25%, outperforming the ResNet50 model, which attained an average of 58.87%. Similarly, the F1-Score — a harmonic mean of precision and recall — followed the same trend, with values of 0.7070 and 0.5997 for VGG16 and ResNet50, respectively. Also, both the precision and negative predictive value of the VGG model were higher than the ResNet model — 70.83% to 64.10% for precision and 71.88% to 56.10% for negative predictive value, respectively. Because new runs of the models were necessary to address the reviews received, the two variables that showed the greatest variation, namely accuracy and ROC-AUC, were measured 10 times using the same set of images, with the images randomly shuffled in each iteration using seed 42. The resulting values were then used to calculate the arithmetic mean for each one.

The VGG16 model showed higher sensitivity, at approximately 79.07%, indicating a greater resistance to false negatives than the ResNet50, with 58.14%. Interestingly, the specificity of both models was equal, 62.16%, since both have same counts of true negatives and false positives. Finally, the VGG16 model achieved a higher ROC-AUC score, 0.7266, than ResNet50 with a score of 0.6568. These results indicate better performance for VGG16 under this specific experimental configuration, but they should not be interpreted as evidence of superiority beyond this dataset.

Tables II and III present the confusion matrices for VGG16 and ResNet50, respectively, giving detailed insights into classification performance per class. The VGG16 model correctly identified 34 glaucoma images and 23 normal images. Using glaucoma as the positive class, the submitted confusion matrices contain 9 false-negative and 14 false-positive classifications for VGG16, compared with 18 false-negative and 14 false-positive classifications for ResNet50. Thus, within this 80-image evaluation subset, the higher VGG16 accuracy is primarily associated with fewer false-negative glaucoma classifications.

Within the selected ODIR fundus subset and the original training protocol, VGG16 achieved higher accuracy than ResNet50. The limited dataset size, image-level randomization, and lack of external validation restrict broader interpretation of this comparison.

TABLE I
PERFORMANCE METRICS FOR VGG16 AND RESNET50 ARCHITECTURES.

Metric	VGG16	ResNet50
Accuracy	0.7125	0.5887
F1-Score	0.7070	0.5997
Sensitivity	0.7907	0.5814
Specificity	0.6216	0.6216
Precision	0.7083	0.6410
NPV	0.7188	0.5610
ROC-AUC	0.7266	0.6568
Variance	0.2400	0.2498

TABLE II
CONFUSION MATRIX FOR VGG16 ARCHITECTURE (VALIDATION SET SIZE: 80 IMAGES).

	Predicted Glaucoma	Predicted Normal
Actual Glaucoma	34	9
Actual Normal	14	23

TABLE III
CONFUSION MATRIX FOR RESNET50 ARCHITECTURE (VALIDATION SET SIZE: 80 IMAGES).

	Predicted Glaucoma	Predicted Normal
Actual Glaucoma	25	18
Actual Normal	14	23

DISCUSSION

The results obtained in this study provide an insightful basis for analyzing the challenges, limitations, and potential of neural network architectures applied to glaucoma detection using retinal fundus photographs.

Dataset Limitations and Generalizability

The dataset size consisted of 400 images, allowing relatively fast experimentation but limiting the diversity of image characteristics represented during training. The images were obtained from a single public dataset, and no independent external cohort or external-device validation was performed. Consequently, the results should not be generalized to other populations, acquisition systems, or clinical settings without further validation. Future efforts should prioritize larger and more diverse datasets encompassing different imaging devices, patient demographics, and glaucoma stages. Accordingly, the present work is best interpreted as an exploratory baseline comparison under a constrained experimental setting, rather than as a validated screening or diagnostic system, considering the small number of 320 images used for training and 80 for validation.

Computational Infrastructure and Training Adjustments

Despite leveraging cloud-based GPU acceleration (Google Colab with T4 GPUs), a full end-to-end fine-tuning of the convolutional layers was not feasible due to resource constraints and the extensive number of parameters to be optimized. Consequently, the transfer learning strategy was adopted, freezing pre-trained convolutional layers from ImageNet-weighted models and training only the dense classifier layers. This approach enabled feasible training times but inherently limited the ceiling of achievable accuracy. Additionally, intermittent GPU availability imposed further restrictions. More sophisticated techniques such as fine-grained tuning and advanced data augmentation could yield improved performance in future studies with enhanced computational access.

Model Analysis and Overfitting Issues

The VGG16 and ResNet50 architectures displayed notably different classification performances. While VGG16 reached approximately 71.25% accuracy, ResNet50 scored around 59%, tested on equivalent randomized validation splits of the same dataset. The observed divergence between training and validation accuracies across multiple epochs indicated overfitting as the training progressed. Because only one random seed and a limited dataset were used, these values should be interpreted as preliminary point estimates rather than stable estimates of model performance.

Positioning Relative to Current Methods

The comparison of VGG16 and ResNet50 is treated in the revised manuscript as an exploratory baseline experiment rather than a state-of-the-art contribution. Recent literature demonstrates continued development of deep-learning methods for retinal-image analysis [7], [19]. Therefore, the present results are interpreted only within the current dataset and protocol, while broader benchmarking against simpler baselines and contemporary architectures is reserved for future work.

Explainability and Additional Validation

The present experiment did not include Grad-CAM or other saliency analysis, subgroup analysis, or external-device validation. These limitations are now stated explicitly because they restrict assessment of shortcut learning, demographic or acquisition-related bias, and whether model attention is associated with clinically plausible retinal regions. The absence of an independent external cohort further prevents claims of

generalizability. The revised manuscript no longer presents VGG16 and ResNet50 as state-of-the-art or as a novel signal-processing method. They are positioned as established baseline CNN architectures. Comparisons with simpler baselines, more recent CNNs, vision transformers, and foundation models are identified as future work and as necessary for broader benchmarking.

Conclusion

This exploratory study compared VGG16 and ResNet50 for binary glaucoma classification using a limited subset of ODIR retinal fundus photographs. Under the original experimental configuration, VGG16 achieved higher accuracy than ResNet50. However, the small single-dataset sample, frozen convolutional features, absence of data augmentation and external validation, and uncertainty regarding patient-level separation limit the conclusions that can be drawn. Future work should evaluate patient-level validation, broader hyperparameter tuning, partial or full fine-tuning, explainability methods, contemporary architectures, and independent external datasets before any clinical interpretation is considered.

ACKNOWLEDGMENT

The authors would like to thank Shanggong Medical Technology Co., Ltd. for making the Ocular Disease Intelligent Recognition (ODIR) dataset publicly available, which was used in this study. The dataset can be accessed at https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k. The usage of AI in this project is limited to supervised learning and grammar and semantics checking.

REFERENCES

[1] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *British Journal of Ophthalmology*, vol. 90, no. 3, pp. 262–267, 2006.

[2] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, 2014.

[3] R. Wang and J. L. Wiggs, "Common and rare genetic risk factors for glaucoma," *Ophthalmology*, vol. 121, no. 10, pp. 2081–2098, 2014.

[4] K. V. Stuart, L. R. Pasquale, J. H. Kang, P. J. Foster, and A. P. Khawaja, "Towards modifying the genetic predisposition for glaucoma," *Molecular Aspects of Medicine*, vol. 93, no. 1, pp. 1–9, 2023.

[5] L. Pasquale and J. Kang, "Lifestyle, diet, and glaucoma," *J Glaucoma*, vol. 18, no. 3, pp. 196–200, 2009.

[6] Kaggle, "Ocular disease recognition dataset (ODIR)," https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k, 2020, accessed: 2025-08-27.

[7] H. Bilal, A. Keles, and M. Bendechache, "Advances in disease detection through retinal imaging: A systematic review," *Computers in Biology and Medicine*, vol. 194, pp. 1–36, 2025.

[8] A. Pachiyappan, U. N. Das, and R. Murthy, Tatavarti V.S.P. Tatavarti, "Automated diagnosis of diabetic retinopathy and glaucoma using fundus and oct images," *Lipids Health Dis.*, vol. 11, no. 73, pp. 1–10, 2012.

[9] M. Goel, R. G. Picciani, R. K. Lee, and S. K. Bhattacharya, "Aqueous humor dynamics: a review," *The Open Ophthalmology Journal*, vol. 4, pp. 52–59, 2010.

[10] R. Sihota, P. Sony, V. Gupta, T. Dada, and R. Singh, "Diagnostic capability of optical coherence tomography in evaluating the degree of glaucomatous retinal nerve fiber damage," *Investigative Ophthalmology & Visual Science*, vol. 47, no. 5, pp. 2006–2010, 2006.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] F. E. Târcoveanu, F. Leon, C. Lisa, and S. Curteanu, "The use of artificial neural networks in studying the progression of glaucoma," *Scientific Reports*, vol. 14, no. 1, pp. 1–16, 2024.

[15] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[19] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: A survey," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–30, 2021.

Algorithm for Muscle Thickness Tracking on Ultrafast Ultrasound

Jonathan de Oliveira
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
Grad. Program in Health Tech. (PPGTS), PUCPR
Curitiba, Brazil
gmjo92@gmail.com

Tatiane C. Boumer
Grad. Program in Health Tech. (PPGTS), PUCPR
Curitiba, Brazil
tatianeboumer@gmail.com

Tobias Martin
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
tmartin@ukaachen.de

Andreas Follmann
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
afollmann@ukaachen.de

Joaquim Miguel Maia
Electronics Eng. Dept. (DAELN), UTFPR
Curitiba, Brazil
joaquim@utfpr.edu.br

Wagner Coelho de A. Pereira
Biomedical Eng. Program (COPPE), UFRJ
Rio de Janeiro, Brazil
wagner@peb.ufrj.br

Amauri Amorin Assef
Grad. Program in Electrical & Computer Eng. (CPGEI) UTFPR, Curitiba, Brazil
amauriassef@utfpr.edu.br

Mauren Abreu de Souza
Grad. Program in Health Tech. (PPGTS), PUCPR
Curitiba, Brazil
mauren.souza@pucpr.br

***Abstract*—Quantitative assessment of muscle architecture via B-mode musculoskeletal ultrasound (US) is important for clinical diagnostics and myoelectric human-machine interfaces, but high-temporal-resolution ultrafast US generates extensive frame sequences that make manual aponeurosis tracking impractical. This paper presents a deterministic, frame-independent computer-vision pipeline for automated aponeurosis tracking and instantaneous muscle-thickness extraction from B-mode ultrasound. Unlike optical-flow and Kalman-filter-based trackers, which estimate each frame's boundary from the previous frame and are therefore vulnerable to cumulative drift, and unlike deep-learning trackers, which require large annotated training sets, the proposed method treats each frame as an independent horizontal intensity projection constrained by anatomical priors, so that tracking of one frame does not depend on the outcome of the previous one. Each frame is processed via a horizontal intensity projection, a moving-average filter to suppress speckle noise, an anatomically constrained peak-width detection step ($\Delta_w = 10-15$ points), and offline, zero-phase 5th-order Savitzky–Golay post-filtering of the compiled thickness profile. The pipeline was applied to a dynamic in vivo isometric-contraction sequence (3000 frames, 200 Hz) and separately evaluated for static boundary agreement against manual annotations on a curated subset (n = 197) of an open-source benchmark dataset, achieving an Intraclass Correlation Coefficient of 0.858 and a mean Bland-Altman bias of 0.6 px (95% limits of agreement: -32.1 to 33.2 px). These results support the boundary-detection stage as a proof of concept; because the validation set consisted of static images with a single annotator, it does not by itself establish tracking continuity, drift resistance, or recovery after boundary loss across a dynamic sequence, nor has agreement been converted to millimeters or benchmarked against comparator methods. These remain open questions addressed in the Discussion and proposed as future work.**

***Index Terms*—aponeurosis tracking, computer vision, muscle thickness, musculoskeletal, ultrafast ultrasound**

I. Introduction

The human musculoskeletal system is a highly complex architecture comprising bones, muscles, and associated connective tissues, functioning synergistically to provide structural stability, facilitate locomotion, and support vital physiological mechanisms such as respiration and postural equilibrium [1]. Understanding the biomechanical and physiological nuances of this system is essential, as abnormalities in neuromuscular function can precipitate debilitating pathological conditions, including myopathies, paralysis, and severe motion disorders [2]. Furthermore, even within high-performance sports and recreational physical activities, structural components such as tendons and ligaments are highly susceptible to acute injuries and ruptures [2]. Nevertheless, the quantitative assessment of musculoskeletal dynamics is important for expanding clinical diagnostics, optimizing rehabilitation protocols, and mapping muscle behavior across distinct states of relaxation, activation, and fatigue [3].

At the core of functional movement lies the muscle-tendon unit (MTU), which serves as the primary structural interface responsible for force generation, transmission, and joint displacement [4]–[6]. Investigating the kinetic and kinematic responses of the MTU provides valuable insights into the individual contribution of specific muscle groups during dynamic contractions and angular variations [7]–[9]. Accurate characterization of these contractile mechanics is essential for clinical monitoring and for improving predictive musculoskeletal models. These models are utilized in advanced human-machine interfaces (HMIs) and myoelectric prosthetic control systems [3], [10].

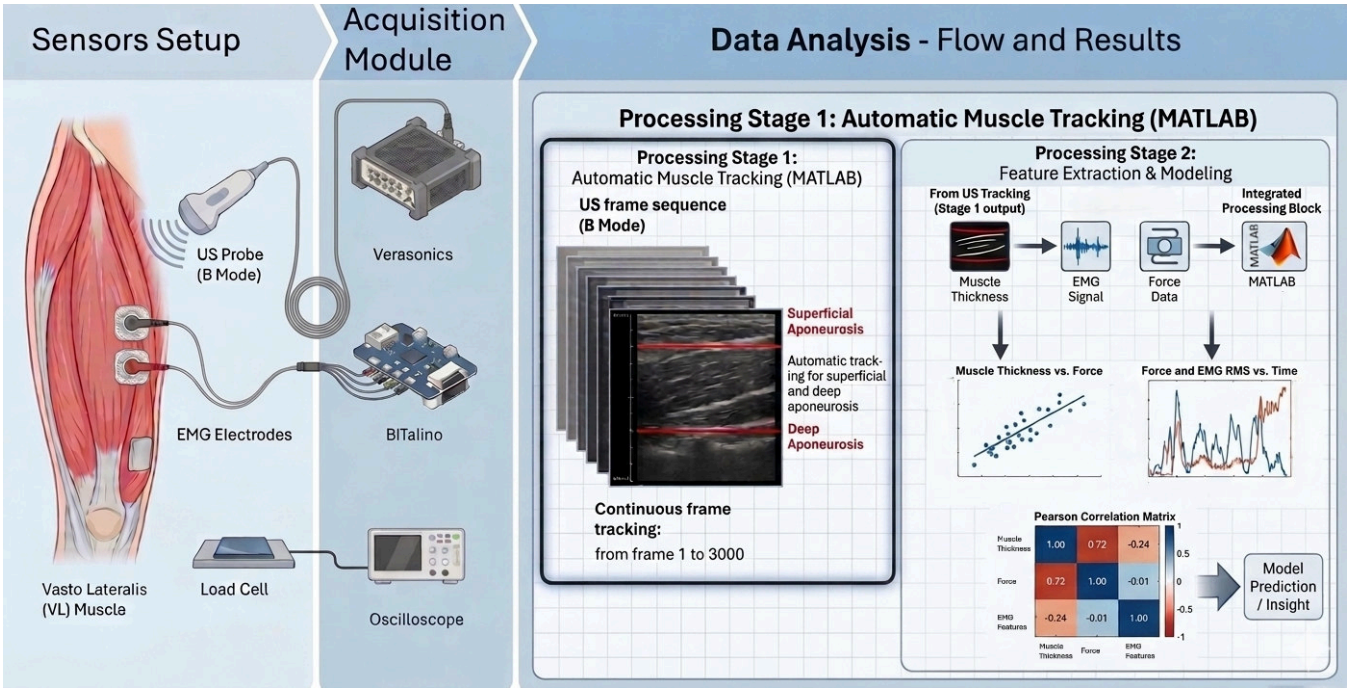


Fig. 1. Overview of the multi-sensing framework block diagram, highlighting the transition from the physical Sensors Setup and the Acquisition Module to the Data Analysis environment containing the proposed processing pipeline.

Despite the clinical and technological importance of this characterization, accurately estimating subject-specific MTU parameters *in vivo* remains a significant challenge. Traditional evaluation methods often fail to capture the transient lateral expansion and longitudinal deformation of muscle fibers during active contraction [9], [11]. To overcome this limitation, high-temporal-resolution imaging modalities, such as ultrafast ultrasound (US), have emerged as powerful tools to non-invasively visualize internal muscular structures in real-time.

In order to study these structural dynamics, recent research has highlighted the necessity of data fusion approaches that combine structural imaging with complementary biomedical signals. While extensive literature reviews have examined multiple biosensing modalities, such as surface electromyography (sEMG), electrical impedance tomography (EIT), near-infrared spectroscopy (NIRS), force myography (FMG), and phonomyography (PMG) for transradial prosthetic control and HMIs [12]–[14], there remain a few frameworks that systematically integrate high-framerate ultrasound with synchronized mechanical and electrical biosignals.

However, because these unified multi-sensing frameworks capture dynamic trials at extreme temporal resolutions, they inherently generate high-density image sequences consisting of thousands of ultrasound frames per subject. Manually extracting morphological MTU metrics, such as instantaneous muscle thickness from these extensive datasets, is computationally unfeasible and highly operator-dependent. Therefore, the development of robust, automated computer vision algorithms capable of precisely detecting and tracking the superficial and deep muscular aponeuroses within these multi-sensing applications has become a bottleneck.

To address this technological gap, this paper introduces a localized image processing pipeline designed specifically for the continuous tracking of *vastus lateralis* (VL) muscle aponeuroses in ultrafast ultrasound sequences. The proposed algorithm serves as **Processing Stage 1** for the **Data Analysis** phase. The multi-sensing framework illustrated in Fig. 1, demonstrates the automatic localization of the superficial and deep fascial boundaries across a dense temporal series of up to 3000 frames to extract continuous muscle thickness variations. The framework shows the physical **Sensors Setup** and the hardware-driven **Acquisition Module** and the **Data Analysis** processing environment, programmed on Matlab (The MathWorks, Natick, MA, USA) The proposed algorithm tracks both aponeuroses along a physical activity trial. These tracking results serve for future applications on correlations within **Processing Stage 2** for feature extraction and modeling.

Section II describes the mathematical foundation of the horizontal tracker. Section III details the algorithm pipeline applied to the tracker. Section IV presents the tracking results and evaluates algorithmic stability against baseline noise artifacts. Finally, Section V concludes the paper and outlines approaches for future work.

II. Theoretical Background

Several automated or semi-automated tracking frameworks have been proposed in the literature, though each one presents distinct technological limitations when applied to high-framerate dynamic sequences.

- **Morphological and Mask-Based Frameworks:** The *Muscle Ultrasound Analysis* (MUSA) software relies on structural filters and localized intensity masks to isolate muscular boundaries [15]. While computationally lightweight, this approach heavily depends on local contrast thresholds. Consequently, it is highly susceptible to acoustic artifacts and localized speckle noise accumulated across long-duration, multi-frame ultrafast US acquisitions, frequently resulting in boundary segmentation drift.
- **Deep Learning Approaches:** Deep learning architectures, such as *DL_Track*, leverage convolutional neural networks trained specifically for end-to-end aponeurosis detection [16]. Although deep networks exhibit exceptional adaptability to varying muscle anatomies, they often suffer from frame-to-frame localized inconsistencies. In dense temporal sequences, slight variations in frame contrast may reduce prediction consistency across consecutive frames, leading to significant structural data loss and discontinuous tracking over time.
- **Optical Flow Implementations:** The widely used *UltraTrack* framework utilizes Lucas-Kanade optical flow algorithms to simultaneously track aponeuroses and fascicle orientations [17]. This differential method assumes brightness constancy between successive frames. However, dynamic muscle contractions inherently alter tissue reflectivity, probe pressure, and out-of-plane scattering. These localized luminance variations introduce cumulative integration drift, causing tracking errors that propagate rapidly through extended frame sequences.
- **State-Space and Predictive Filtering:** Most recently, *UltraTimTrack* was introduced, incorporating structural line-detection algorithms coupled with an adaptive Kalman filter to predict and smooth aponeurosis trajectories [18]. Although state-space modeling successfully mitigates rapid transient noise, this methodology represents a promising approach. Its mathematical constraints regarding non-linear tissue deformations and its generalization capabilities across diverse transducer configurations remain insufficiently investigated by the broader scientific community.

Beyond computational simplicity, the deterministic design offers three concrete mechanistic distinctions from the frameworks reviewed above. First, because each frame is processed independently from a fixed anatomical offset rather than propagated from the previous frame's estimate, the pipeline is not subject to the cumulative integration drift that affects optical-flow and Kalman-filter-based trackers, in which a single mis-estimated frame biases subsequent predictions. Second, unlike deep-learning approaches, the method requires no annotated training data and no retraining when applied to new acquisition hardware or muscle groups, at the cost of depending on hand-specified anatomical priors that must be re-tuned for substantially different anatomies. Third, the explicit peak-width and anatomical-bound constraints make failure modes interpretable frame-by-frame, whereas failures in learned models are comparatively opaque. These are architectural, hypothesized distinctions rather than demonstrated advantages: a head-to-head quantitative comparison of computational time, drift accumulation, and failure rate against UltraTrack, UltraTimTrack, MUSA, and DL_Track on shared sequences and hardware has not yet been performed and is identified as a priority for future work.

III. Methods

The proposed experimental framework integrates high frame-rate ultrasound (US) imaging, surface electromyography (sEMG), and mechanical force tracking into a cohesive data analysis pipeline. This multimodal framework used an open architecture ultrasound research system (Vantage 128, Verasonics Inc., Kirkland, WA, USA). The system was synchronized with a BITalino platform equipped with a specialized sEMG differential sensor device (UC-E6, PLUX Biosignals, Lisbon, Portugal) and a calibrated S-type load cell (Lider Balanças, Araçatuba, Brazil). The hardware layout and individual instrument interfaces are structurally detailed in Fig. 2.

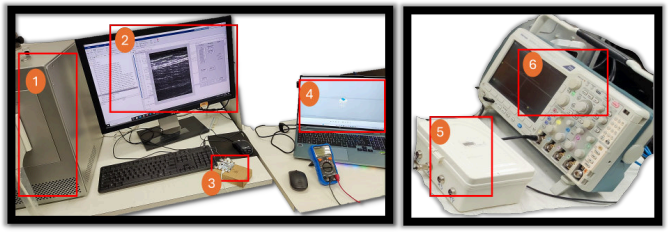


Fig. 2. Experimental instrumentation setup for synchronized multimodal data acquisition: 1) Verasonics Vantage 128 open research platform; 2) Host display monitor showing real-time B-mode acquisition; 3) BITalino (r)evolution board equipped with sEMG biosensors; 4) Computational environment synchronized via host peripherals; 5) S-type load cell assembly; 6) Digital oscilloscope used for hardware trigger synchronization and strain gauge voltage verification

A. Data Acquisition Protocol

The experimental data acquisition protocol was conducted by a licensed physiotherapist in 2024. This study follows an exploratory, applied, cross-sectional, and quantitative research design approved by the Human Research Ethics Committee of the Pontifical Catholic University of Paraná (PUCPR), Institutional Review Board approval no. 6.883.378. In strict compliance with National Health Council Resolution 466/2012, all participating subjects provided written informed consent prior to the commencement of the testing sessions.

To standardize physiological baselines and minimize metabolic confounding factors, participants were instructed to abstain from vigorous physical activities for 72 hours preceding the examination. Immediately prior to the evaluation, subjects remained in a supine position for 15 minutes to allow for the proper stabilization of body fluids.

Following the stabilization period, localized anatomical landmarks were identified to standardize the transducer placement. As depicted in the experimental positioning in Fig. 3, the ultrasound linear transducer was aligned longitudinally

relative to the muscle fibers of the right *vastus lateralis* (VL). The spatial coordinates of the transducer were systematically defined at 50% of the total anatomical distance measured between the greater trochanter and the lateral femoral condyle, registered over a reference path oriented at an inclination of 15°. Probe alignment was considered optimal when the muscle fascicles were clearly visualized spanning between the aponeuroses. To prevent unintended tissue deformation and underlying structural compression, the transducer was applied with minimal mechanical pressure, utilizing a colorless acoustic coupling gel to optimize impedance matching and image quality. A single investigator was responsible for all ultrasound operations, capturing each pre-demarcated region for offline processing.

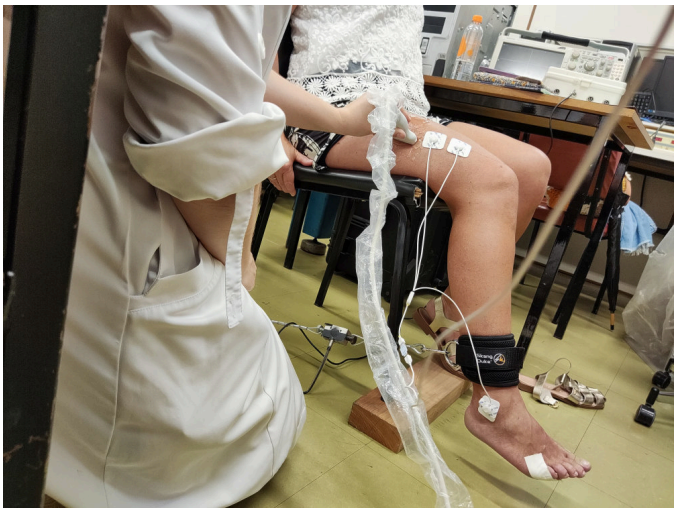


Fig. 3. Experimental instrumentation and participant positioning during the data acquisition session, showing the synchronized placement of the ultrasound transducer, load cell strap coupling, and surface electromyography (sEMG) electrodes on the *vastus lateralis* muscle.

To ensure signal fidelity, the BITalino acquisition hardware (PLUX Biosignals, Lisbon, Portugal) was calibrated and tested prior to each session. Surface EMG electrodes were properly positioned over the muscle belly by the healthcare professional according to SENIAM guidelines. Concurrently, the S-type load cell captured the mechanical force exerted by the participant. To assist the subject in maintaining the targeted contraction profiles, a digital oscilloscope provided real-time visual feedback of the active force output against a predefined reference threshold. Within this architecture, the sEMG sensors logged the electrical manifestation of muscle activation, while the open-architecture Verasonics Vantage platform acquired the high frame rate B-mode image sequences. The multimodal datasets were digitized and recorded synchronously by the respective Verasonics and BITalino acquisition modules.

The dynamic contraction protocol was structured into discrete sequential phases. Each active trial consisted of an initial 5 s resting baseline (relaxation), followed by a 5 s sustained isometric contraction of the knee extensors, and a final 5 s relaxation phase, yielding a total sequence duration of 15 s (3000 frames at 200 Hz). A 10-minute rest interval was enforced between successive trials to prevent neuromuscular fatigue. Prior to the dynamic tracking sessions, an initial evaluation was executed to determine each subject's Maximum Voluntary Contraction (MVC). This absolute force value served as the baseline reference for subsequent trials.

The protocol was completed by 3 healthy participants, each performing 2 trials per MVC level as described above, with condition order fixed in the order 20%, 30%, 40% across participants. Transducer position was set manually by a single investigator using the anatomical landmarks described above and was not instrumented with an optical marker or inertial measurement unit; probe displacement during contraction therefore could not be independently measured in this study.

B. Algorithmic Processing Pipeline

B-mode ultrasound (US) imaging inherently exhibits structural degradation caused by acoustic speckle noise, attenuation, and frequency-dependent scattering. To mitigate these artifacts and achieve automated frame-by-frame tracking across extensive sequences, a dedicated computer vision pipeline was developed. This algorithm processes high temporal resolution B-mode matrices to continuously segment the superficial and deep aponeuroses, enabling the dynamic extraction of muscle thickness.

The proposed algorithm treats the B-mode frame matrix as a discrete spatial intensity distribution, analyzing the lateral brightness profile to isolate structural fascial boundaries. The processing pipeline is executed through the following sequential mathematical stages:

1) **Region of Interest (ROI) and Offset Definition:** A spatial offset threshold is initialized to eliminate the superficial subcutaneous adipose tissue layer and high-frequency proximal noise, constraining the search matrix to the anatomical boundaries of the target muscle (e.g., *vastus lateralis*).
2) **Horizontal Intensity Projection:** For each frame f, a horizontal summation of the grayscale pixel intensities is computed along the transverse transducer elements ($E = 1$ to 128) for every discrete depth vector d. This spatial integration generates a one-dimensional vertical profile vector, where local maxima correspond to highly echogenic (hyperechoic) structural interfaces.
3) **Smoothing and Peak Detection:** To suppress high-frequency local variations induced by speckle noise, a 50-point moving-average filter is applied to the horizontal intensity profile. From the smoothed curve, local maxima are isolated using a peak-finding function parameterized with a minimum neighborhood width constraint ($\Delta_w = 10$ to 15 spatial points).
4) **Anatomical Constraint Filtering:** Isolated peaks are subjected to deterministic physiological boundary checks. Peaks falling outside expected anatomical bounds for the superficial and deep aponeuroses are discarded as outliers. The two dominant valid peaks

corresponding to the geometric centroids of the superficial (A_a) and deep (A_b) fascial layers are subsequently selected.

5) **Dynamic Thickness Extraction:** The instantaneous muscle thickness (T_m) for frame f is calculated as axial distance between the deep and superficial aponeurosis boundaries, defined by:

$$T_m(f) = A_b(f) - A_a(f) \tag{1}$$

6) **Temporal Post-Filtering:** Once the thickness profile is compiled across all 3000 sequential frames, a 5^{th}-order Savitzky-Golay filter is implemented. This digital filter applies a local polynomial regression to smooth high-frequency transient tracking jitter while preserving the underlying low-frequency morphological deformations during muscle contraction.

C. Structured Algorithm

The structured pseudo code of the tracking pipeline is formalized in Algorithm 1.

Algorithm 1 Automated Aponeurosis Detection and Thickness Tracking

Require: $US_B_Mode_Data$ {Matrix of 3000 frames (Depth $\times$ Elements)}
Ensure: $Muscle_Thickness$ (Vector of size 3000), $Aponeurosis_Regions(A_a, A_b)$

Define $offset$ {Isolate subcutaneous fat layer and initial noise}
for $f = 1$ **to** 3000 **do**
  $H_{sum} \leftarrow \sum_{E=1}^{128} US_B_Mode_Data(Depth, E, f)$
  $H_{smooth} \leftarrow$ MovingMeanFilter(H_{sum}, window $= 50$)
  $Peaks \leftarrow$ FindPeaks(H_{smooth}, MinWidth $= 10$ to 15)
  $[A_a(f), A_b(f)] \leftarrow$ FilterAnatomicalConstraints($Peaks$)
  $T_m(f) \leftarrow A_b(f) - A_a(f)$
end for
$Muscle_Thickness \leftarrow$ SavitzkyGolayFilter(T_m, order $= 5$)
return $Muscle_Thickness, A_a, A_b$

D. Validation Dataset and Inclusion Criteria

To evaluate the mathematical generalization of the proposed pipeline, an independent validation protocol was established using the open-source *DL_Track* benchmarking database [16]. The public repository comprises 574 B-mode musculoskeletal images; however, strict morphological inclusion criteria were enforced to match the input specifications of the proposed algorithm.

The proposed technique requires uncropped, raw rectangular ultrasound matrices that preserve the full lateral array borders across all transducer elements, demonstrated in Fig. 4. Image a) shows an example of standard raw configurations. Image b) demonstrates the exclusion criteria, which shows frames exhibiting artificial geometric clipping, non-standard aspect ratios, or post-processed clinical overlay annotations, such as masked borders. These artificial borders introduce high-contrast step-function intensity gradients during the column-wise summation profile, creating false peak maxima that do not correspond to physical anatomical structures. Consequently, a refined subset of 197 raw ultrasound images was isolated for formal algorithmic execution.

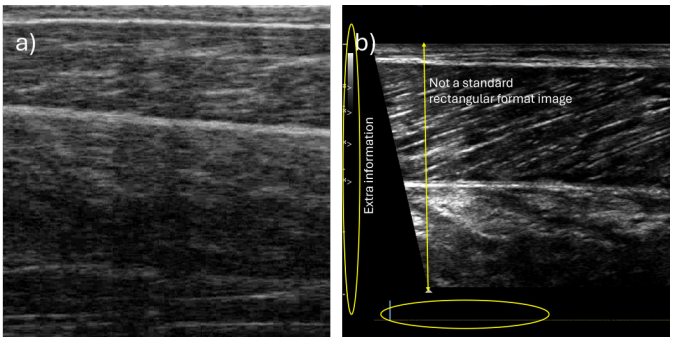


Fig. 4. Example of ultrasound image from *DL_Track* dataset, demonstrating exclusion criteria: a) shows the standard image which the algorithm applies the technique; b) highlights the borders, extra information and different standard.

IV. Results and Discussion

The performance and tracking stability of the proposed computational pipeline were evaluated using a representative dynamic B-mode ultrasound sequence. The dataset comprised a dense, continuous high frame rate acquisition of 3000 frames captured at a temporal resolution of 200 Hz over a 15 s duration, capturing transient muscle architectural variations during an active relaxation-contraction-relaxation experimental cycle.

A. Muscle Tracking Algorithm

Fig. 5 illustrates the continuous, frame-by-frame tracking results, mapping the localized instantaneous distance between the isolated superficial and deep aponeuroses of the *vastus lateralis* muscle. The system effectively tracks the morphological boundaries across the predefined timeline landmarks: an initial 5 s relaxation baseline, a 5 s sustained isometric contraction, and a final 5 s relaxation phase.

The practical utility of the tracking pipeline under varying mechanical demands is highlighted in Fig. 6, which contrasts the extracted muscle thickness profiles during isometric trials performed at 20%, 30%, and 40% of the subjects' MVC. The tracked morphological profiles reveal an inverse displacement during the active contraction window ($t = 5\,\text{s}$ to $t = 10\,\text{s}$), characterized by a transient decrease in localized muscle thickness across all submaximal effort levels.

This specific behavior is captured by the algorithm and reflects few mechanical constraints of the experimental setup. During active knee extension against the rigid ankle strap, the superficial expansion of the muscle belly presses against the linear probe. Because the transducer is maintained in a fixed relative position, this localized mechanical tension compresses the tissue layer, resulting in a reduction of the vertical distance between the fascial interfaces. At the highest effort intensity

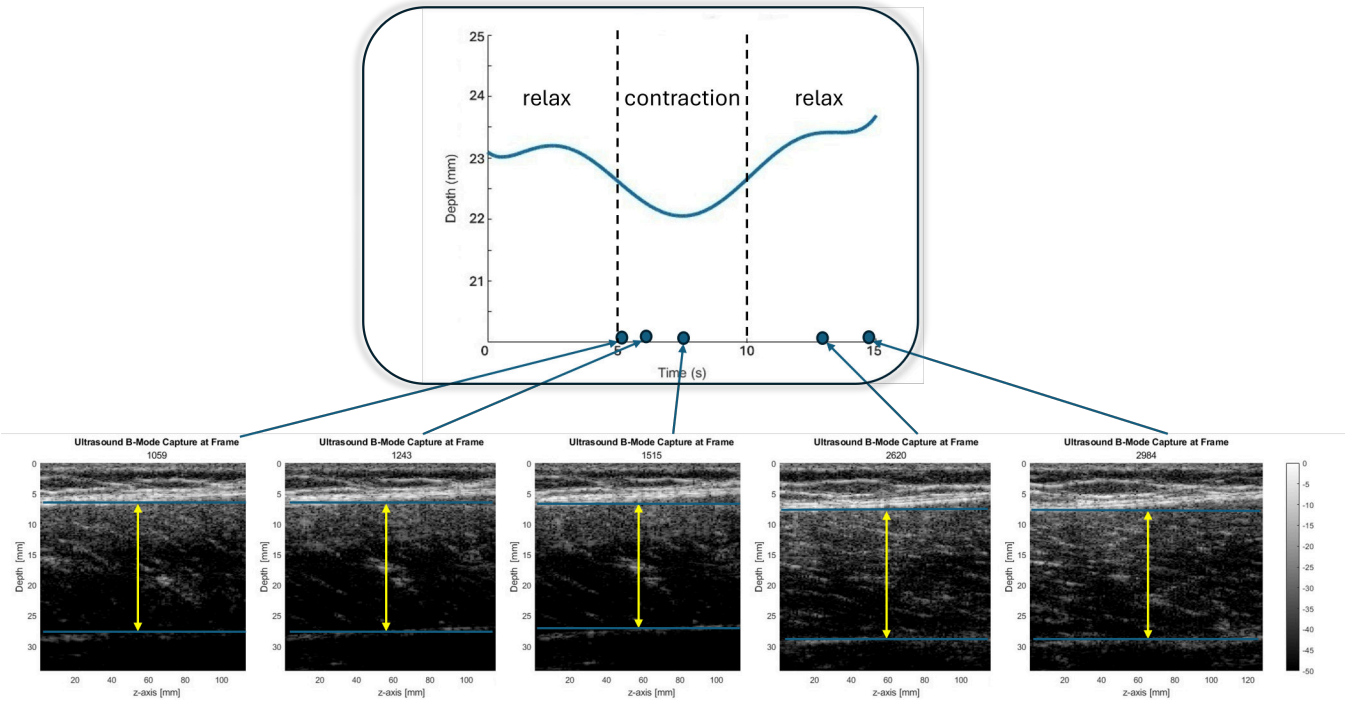


Fig. 5. Dynamic frame-by-frame tracking verification demonstrating the automated localization of the superficial and deep aponeuroses of the *vastus lateralis* muscle. The multi-frame sequence highlights the algorithmic behavior across the 15 s relaxation-contraction-relaxation protocol (Frames 1059, 1243, 1515, 2620, and 2984).

(40% MVC, red curve), the algorithm captures a sharper initial oscillatory transition during the onset of contraction, showing a baseline drop from approximately 31.5 mm down to a compressed stabilization floor of 27.5 mm.

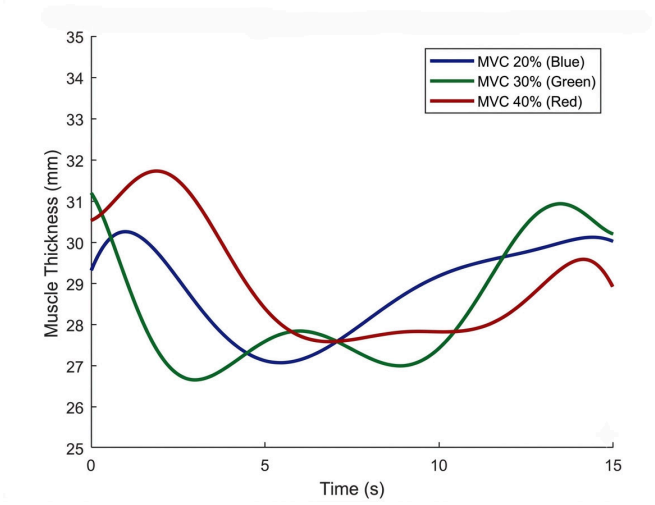


Fig. 6. Dynamic frame-by-frame muscle thickness tracking curves across 15 seconds for three distinct submaximal effort intensities: 20% MVC (blue), 30% MVC (green), and 40% MVC (red), capturing the structural compression during active isometric phases.

A challenge in tracking dense 3000 frames ultrasound sequences is the vulnerability to local luminance attenuation and speckle noise artifacts. The results indicate that the integration of the 50-point moving average filter combined with the constrained peak-width detection ($\Delta_w = 10$ to 15 points) provides stabilization. Across the discrete cross-sections in Fig. 5 (Frames 1059 to 2984), the peak finding framework effectively rejected internal echogenic noise artifacts within the muscle belly, such as highly reflective localized fascicle bundles or structural vascular boundaries, which would otherwise introduce tracking outliers or could cause false boundary attribution to the algorithm.

Furthermore, the implementation of the 5^{th}-order Savitzky-Golay post-filter smoothed out high-frequency jitter and frame-to-frame pixel discretization noise. As observed in the smooth contours of the tracking traces in Fig. 6, this temporal filtering isolated the low-frequency mechanical muscle deformation profile, although it introduced a slight artificial attenuation of the valid physical displacement amplitudes during rapid tissue transitions.

By treating the tracking task as a constrained localized intensity maximization problem, the algorithm delivers a framework for batch-processing multi-sensing trial blocks. It mitigates the significant data-loss vulnerabilities and frame-to-frame convergence failures associated with end-to-end deep learning approaches, providing a reproducible tool for biomechanical analysis.

Despite the high tracking continuity demonstrated by the algorithm, a critical analysis of the thickness profiles in Fig. 5 reveals a minor localized baseline variation that points to specific physical limitations in the data acquisition setup. During sustained high-intensity contractions (e.g. 40% MVC),

the vigorous spatial deformation and hardening of the muscle belly inevitably introduce minor transducer displacements and localized skin-probe slippage. This transducer shift alters the relative acoustic coupling angle, inducing transient gray-level variations and minor artifacts in the B-mode data. In future protocol iterations these effects can be attenuated by improving the acquisition framework with a custom thermo moldable rigid probe stabilizer contoured to the participant thigh geometry, ensuring an unyielding perpendicular alignment. Furthermore, integrating a localized optical tracking marker or an inertial measurement unit (IMU) directly onto the transducer housing would allow real-time monitoring and computational compensation for probe translation, thereby suppressing mechanically-induced tracking noise at the source.

Regarding computational feasibility, each frame in the proposed pipeline requires only a summation across 128 transducer elements, a fixed-window moving-average filter, and a constrained local peak search — operations that scale linearly with image depth and are substantially cheaper than a convolutional forward pass of the size used in DL_Track-like architectures; the Savitzky-Golay step is a single additional linear-time pass over the compiled thickness vector. This is offered as a plausibility argument for real-time feasibility rather than a measured result: per-frame latency on target hardware was not benchmarked in this study and would need to be reported before real-time claims can be substantiated.

B. Algorithm Validation

The agreement between the automated muscle thickness measurements (T_m) generated by the proposed algorithm and the manual ground-truth annotations provided in the dataset was quantified using a two-way mixed-effects Intraclass Correlation Coefficient (ICC) and Pearson correlation coefficient (r). As illustrated in Fig. 7, the pipeline achieved an ICC of 0.858 ($n = 197$) and a corresponding $r = 0.86$, demonstrating a statistically significant agreement with the manual gold standard. The algorithm's tracking boundaries align closely with the manual fascial tracings, remaining stable across different levels of acoustic speckle noise and diverse echo-intensity gradients included within the dataset. The strong statistical outcomes highlight that the 50-point moving mean filter paired with the tuned peak-width parameter (Δ_w) can accurately represent the underlying anatomical structures, demonstrating that the pipeline can be applied for high-throughput batch processing without requiring manual operator intervention or custom border-masking scripts.

To further investigate the measurement agreement, the corresponding Bland-Altman analysis was performed, shown in Fig. 8. The distribution revealed a mean bias of 0.6 px, with 95% limits of agreement spanning -32.1 px to 33.2 px, a random-error window of approximately 65 px. The near-zero mean bias suggests no strong systematic tendency toward over or under estimation across the dataset as a whole. The width of the limits of agreement is not negligible in absolute terms; it has not yet been converted to millimeters using the benchmark dataset's pixel-to-depth calibration, nor compared against a pre-specified clinical tolerance, so precision adequacy for clinical use cannot yet be claimed. Three distinct outliers fell below the lower LoA, with discrepancies between -50 px and -140 px, concentrated in the intermediate thickness range of 150–220 px. We hypothesize this reflects higher echogenicity in adjacent deep muscle layers misleading peak selection in a few images; this could not be confirmed against the in vivo data, which lack equivalent frame-level annotation.

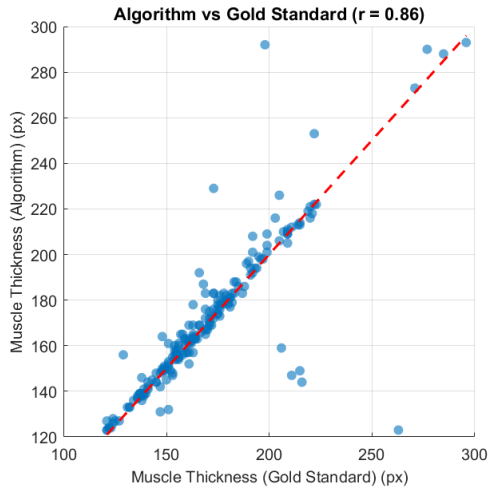


Fig. 7. Scatter plot displaying a linear correlation ($r = 0.86$) between the automated tracking boundaries and manual annotations.

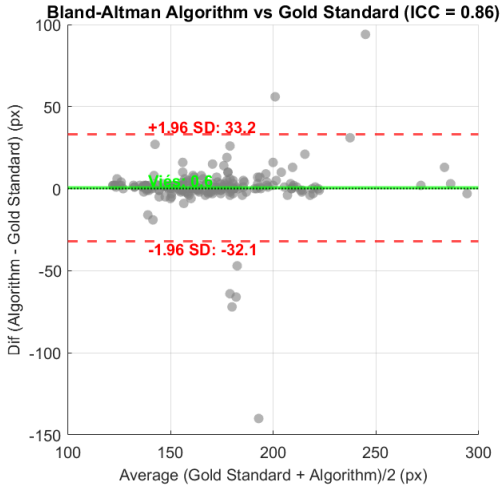


Fig. 8. Bland-Altman plot illustrating a negligible mean bias of 0.6 px, 95% limits of agreement (± 1.96 SD), and isolated underestimation outliers localized within the 150–220 px range.

V. Conclusion

This paper proposed an automated computer vision algorithm for the frame-by-frame segmentation of superficial and deep muscular aponeuroses from high-temporal-resolution B-mode ultrasound sequences. By transitioning from computationally expensive or drift-prone methods to a localized,

constrained morphological intensity projection framework, the proposed algorithm achieves stable tracking with minimal temporal frame-to-frame failures across dense, 3000-frame datasets.

In its current form, the developed pipeline serves as a proof-of-concept for batch processing of structured, offline biomechanical datasets under the anatomical and acquisition constraints described in Section III. It processes dense frame sequences deterministically and without manual operator intervention, and its static-image agreement with manual annotation (ICC = 0.858) supports the underlying boundary-detection approach. Translating this baseline into a validated tracking tool for dynamic sequences, real-time use, or additional muscle groups and clinical populations would require: (i) blinded, multi-rater, frame-sampled ground truth on dynamic in vivo sequences; (ii) a held-out parameter-tuning/validation split; (iii) agreement reported in millimeters against a pre-specified clinical tolerance, including failure rate and boundary-specific error; (iv) head-to-head comparison with optical-flow, Kalman-filter, morphological, and deep-learning baselines on shared sequences and hardware; (v) a component-wise ablation of the moving-average, peak-width, anatomical-constraint, and Savitzky-Golay stages; and (vi) stress-testing under probe motion, anisotropy, gain variation, shadowing, and broader anatomical variability. Future work will pursue a hybrid pipeline that embeds the current anatomical constraints into a deep-learning loss function, aiming to combine the adaptability of learned models with the interpretability of deterministic physiological bounds, subject to the validation steps above.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the National Council for Scientific and Technological Development (CNPq), Grant 146388/2020-8, for the financial support provided through this project. This research was also supported by the Federal University of Technology - Paraná (UTFPR), providing the necessary laboratory infrastructure and technical environment for the computational analysis. The authors would also like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and the German Academic Exchange Service (DAAD), under the CAPES/DAAD - PROBRAL program (Grant No. 88881.986264/2024-01) for the financial support provided through the Probral program. Special thanks are extended to the RWTH Uniklinik and the Pontifical Catholic University of Paraná (PUCPR) for institutional support and the clinical staff who assisted with the instrumentation and data acquisition protocol.

REFERENCES

[1] S. Blanc *et al.*, "The musculoskeletal system and its role in human posture and respiration," *Journal of Biomechanics*, 2010.

[2] Y. Dave *et al.*, "Susceptibility of tendons and ligaments to tears and ruptures in high-performance athletes," *Clinical Biomechanics*, 2021.

[3] R. B. Woodward *et al.*, "Characterizing muscle relaxation, activity, and fatigue using structural musculoskeletal biomarkers," *IEEE Transactions on Biomedical Engineering*, 2019.

[4] M. A. Lemay and P. E. Crago, "A dynamic model for simulating musculoskeletal structures and muscle-tendon units," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 11, pp. 1086–1096, 1996.

[5] L. Li and R. K.-Y. Tong, "Combined ultrasound imaging and biomechanical modeling to estimate triceps brachii musculotendon changes in stroke survivors," *Frontiers in Bioengineering and Biotechnology*, vol. 4, p. 95, 2016. [Online]. Available: https://frontiersin.org

[6] A. Scano *et al.*, "Analysis of the muscle-tendon unit during human locomotion," *Frontiers in Bioengineering and Biotechnology*, 2019.

[7] J. Y. Guo *et al.*, "Kinetic response and force production characteristics of activated muscle groups," *Journal of Electromyography and Kinesiology*, 2011.

[8] B. W. Infantolino *et al.*, "Individual muscle moments during angular contraction changes in motor tasks," *Journal of Biomechanics*, 2014.

[9] J. Ma *et al.*, "Lateral expansion and deformation of muscle fibers during active contraction," *Sensors*, 2019.

[10] M. Jahanandish *et al.*, "Predictive movement models for myoelectric prosthetic control interfaces," *IEEE Transactions on Medical Imaging*, 2021.

[11] R. W. Nuckols *et al.*, "In vivo evaluation of muscle tendon unit parameter sensitivities during locomotion," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2020.

[12] S. Grushko *et al.*, "Biosensing approaches for transradial prosthetic control: A comprehensive review of semg, eit, nirs, smg, fmg, and pmg," *IEEE Reviews in Biomedical Engineering*, 2020.

[13] Y. Zheng *et al.*, "Fmg, emg, and eit biosensing fusion applied to human-machine interfaces," *Biosensors*, vol. 12, no. 7, p. 516, 2022.

[14] J. de Oliveira, M. A. de Souza, A. A. Assef, and J. M. Maia, "Multi-sensing techniques with ultrasound for musculoskeletal assessment: A review," *Sensors*, vol. 22, no. 23, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/23/9232

[15] C. Caresio, M. Salvi, M. Molinari, G. Meiburger, and F. Molinari, "Fully automated muscle ultrasound analysis (musa): A new tool for muscle architecture quantification," *Ultrasound in Medicine & Biology*, vol. 42, no. 12, pp. 3012–3025, 2016.

[16] N. J. Cronin, O. R. Seynnes, and T. Finni, "Fully automated analysis of muscle architecture from b-mode ultrasound images with deep learning," *arXiv preprint arXiv:2009.04790*, 2020.

[17] D. J. Farris and G. A. Lichtwark, "Ultratrack: Software for semi-automated tracking of muscle fascicles in sequences of b-mode ultrasound images," *Computer Methods and Programs in Biomedicine*, vol. 128, pp. 111–125, 2016.

[18] T. van der Zee *et al.*, "Ultratimtrack: Automated aponeurosis and fascicle tracking in musculoskeletal ultrasound using kalman filtering," *Journal of Biomechanics*, 2025.

An Automated Thermographic Pipeline for Temperature Reconstruction, Drift Correction, and Hand Warm-Up Assessment in Acoustic Guitarists

Leonardo de Oliveira Almeida
Department of Electrical Engineering
University of Brasília (UnB)
Brasília, Brazil
leoengunb@gmail.com

Ivo José de Oliveira
Federal Institute of Education, Science and Technology of Goiás (IFG)
Valparaíso, Brazil
ivo.oliveira@ifg.edu.br

Adson Ferreira da Rocha
Department of Electrical Engineering
University of Brasília (UnB)
Brasília, Brazil
adson@unb.br

***Abstract*—An experienced guitar instructor reported that supported plucking (resting a finger on an adjacent string after plucking) seemed to improve hand warming before recitals. An earlier thermographic study used manual temperature readings at a fixed image location, leaving hand motion and camera-related temporal variation unaccounted for. We therefore developed an automated pipeline to track the playing hand and reconstruct temperature estimates from color-mapped video acquired with an uncooled microbolometer camera. A static wall region provided a reference for reducing temporal variation shared across the image. Eight advanced male guitarists completed six-minute supported and unsupported playing protocols, each preceded by a standardized cold challenge (Cold Pressor Test); seven were retained after video-quality screening. Hand tracking used a fine-tuned YOLOv8s-Pose model, achieving keypoint $\mathrm{mAP}_{50} = 0.833$ and $\mathrm{mAP}_{50\text{–}95} = 0.443$ on a participant-disjoint validation set. Before normalization, the temperature increase from the start to the end of playing was nearly three times as large without support as with support ($+1.107$ °C versus $+0.384$ °C). After normalization, mean gains were $+0.630$ °C for the unsupported protocol and $+0.690$ °C for the supported protocol, with no statistically significant between-protocol difference (paired t-test, $p = 0.917$). Sustained peak gains were analyzed descriptively and exceeded final gains in five of seven participants per protocol. These findings show that reference-based normalization can change the apparent comparison of thermal responses. Because the protocols differed in musical demands and radiometric validation was limited, these findings are exploratory and do not isolate the effect of supported plucking.**

***Index Terms*—Infrared thermography, Computer Vision, YOLOv8s-Pose, Temperature reconstruction**

I. Introduction

This study was motivated by a recurring observation from an experienced classical guitar instructor. Before recitals in cold environments, some technically proficient guitarists struggled to warm their hands through playing alone. Informally, switching from an unsupported plucking technique to a supported technique (com apoio), in which the finger briefly rests on the adjacent string after plucking, appeared to increase finger warming. This observation motivated an objective comparison of the hand-temperature responses produced by the supported and unsupported playing protocols. Because the protocols also differed in musical material, tempo, and execution demands, they do not isolate the effect of hand support alone.

Infrared thermography (IRT) may offer a way to test this question without touching the skin or interrupting the performer [1], [2]. During play, skin temperature may reflect the combined influence of muscular activity, peripheral blood flow, autonomic regulation, and heat exchange with the environment. The forearm and intrinsic hand muscles contribute to finger movement, but IRT does not measure these mechanisms directly [3], [4]. Tracking this combined thermal signature could, in principle, turn the instructor's observation into an objective, measurable comparison between techniques [3], [5]–[7].

Extracting this physiological signal reliably, however, is significantly more complex than pointing a thermal camera at a moving hand. The uncooled microbolometer sensors typical of portable IRT cameras are highly susceptible to internal thermal drift. Automatic Gain Control (AGC) remaps the displayed range according to the thermal distribution of the scene, whereas Non-Uniformity Correction (NUC) compensates for spatial differences in detector response [1], [8]. During performance, changes in scene composition may coincide with palette rescaling or internal correction events, introducing apparent temporal shifts into color-mapped video. A prior study of hand warm-up in guitarists [9] extracted temperature manually, at a single fixed pixel, with no automated tracking and no correction for this drift. As a result, instrumental artifacts and true physiological trends remained entangled in the data, limiting how much confidence one can place in its conclusions.

A second confound compounds the first: performance anxiety itself changes hand temperature, through sympathetic activation and peripheral vasoconstriction [10], [11]. A controlled autonomic challenge can provide a more standardized starting condition, although it does not reproduce the psychological context of a recital. The Cold Pressor Test (CPT), which includes brief cold-water immersion, is a widely used experimental stressor [10]. It comes, however, with its own computer-vision cost: after the CPT, the hand's thermal con-

trast against the background collapses, and the rapid, blurred finger motion against the fretboard makes reliable tracking difficult exactly when it matters most.

This paper addresses both challenges simultaneously. A pose-estimation model fine-tuned for thermal imagery automatically tracks the hand, while a wall-referenced normalization procedure reduces common temporal variation shared by the hand and a static background region. The proposed pipeline reconstructs frame-specific color-based temperature estimates from the dynamic scale displayed in each video frame. Because the primary outcome is the temporal evolution of hand temperature (ΔT), the analysis focuses on relative temperature changes rather than absolute measurements.

This work makes three contributions. First, it combines thermal-domain hand tracking with K-means clustering to reduce contamination from the instrument and background. Second, it combines frame-specific color-to-temperature reconstruction with wall-referenced normalization to reduce common temporal variation. Third, it applies the pipeline to advanced guitarists under a controlled cold-related challenge and quantifies individual and group-level thermal responses.

The remainder of this paper is organized as follows. Section II reviews related work in thermal pose estimation, artifact mitigation, and biomedical thermography. Section III details the experimental protocol and the temperature-reconstruction and normalization pipeline. Section IV reports tracking performance and statistical results. Section V discusses the physiological and instrumentation implications of these findings, and Section VI concludes the paper.

II. Related Work

The instructor's observation, and the question it raises, sits at the intersection of four bodies of work that, so far, have not been connected: musculoskeletal health in musicians, the physiology of cold-induced stress, infrared thermography as a measurement tool, and computer vision for thermal imagery. Only one prior study [9] has examined guitarist hand-warming directly; the remaining literature supplies the pieces needed to evaluate it rigorously.

Playing-related musculoskeletal disorders are highly prevalent among professional musicians, particularly those playing string and plucked instruments. Chan et al. [5] reported that structured exercise programs can reduce symptom severity and improve self-reported physical function, although their evaluation relied primarily on subjective clinical measures. Rotter et al. [6], in a systematic review, identified repetitive fine motor activity, prolonged practice, and non-ergonomic postures as major risk factors, but noted that existing studies focus largely on injury and rehabilitation outcomes. Similarly, Adamik et al. [12] found that interventions for bowed string musicians improved pain and functional outcomes without providing standardized physiological measures of warm-up effectiveness. Taken together, these studies highlight a common gap: musician warm-up strategies are typically assessed through subjective or long-term outcomes rather than real-time physiological responses.

Oliveira [9] took the initial step toward closing that gap, applying a CPT and a FLIR C5 thermal camera to nine male guitarists performing both a vigorous unsupported routine and a light supported technique. The supported technique appeared to produce a statistically significant warming effect ($p = 0.003$, $d = 1.42$), while the vigorous routine did not ($p = 0.342$, $d = 0.34$), with several participants' fingers cooling instead of warming — the same counter-intuitive pattern behind the instructor's original observation. Crucially, however, the temperature data in that study was extracted manually at a single, static coordinate every five seconds from exported video files, lacking automated hand tracking, spatial clustering, or sensor-drift compensation. Consequently, those findings may encompass unresolved interference from hand displacement and microbolometer drift. This paper addresses this uncertainty by presenting an automated, wall-normalized measurement framework to evaluate these preliminary observations.

Before that question can be answered, two further pieces are needed: a reliable way to induce a controlled stress baseline, and a reliable way to read peripheral temperature once that stress is induced. The CPT has been widely used for the former. Mourot et al. [10] characterized this response via HRV analysis, showing that brief cold-water immersion produces a sharp rise in sympathetic activity and rapid parasympathetic withdrawal — evidence the CPT is a reliable, reproducible stressor. Jenkins et al. [11] extended this picture by quantifying muscle sympathetic nerve activity directly during a two-minute CPT, reporting significant increases in systolic pressure (+8.1 mmHg), diastolic pressure (+6.4 mmHg), and sympathetic burst frequency (+9.1 bursts/min). Both studies, however, stop at central cardiovascular markers; neither one follows the stress response out to the periphery, where the actual warming or cooling of the hand takes place. Shi et al. [13] moved closer with a contactless radar-interferometry and LSTM framework to estimate HRV during cold exposure, but still without touching peripheral microvascular temperature.

IRT is the natural tool to close that remaining gap, and its sensitivity to autonomic-driven vascular change is well established. Khan et al. [1] reviewed its use for detecting sympathetic-mediated vasodilation during regional anaesthesia, reporting high sensitivity for confirming successful nerve blocks — but flagged that uncooled thermographic measurements remain highly sensitive to environmental drift and typically require manual interpretation, precisely the problem this paper addresses. Hattersley et al. [2] applied IRT to hand temperature, using semi-automated segmentation to track peripheral cooling after a prolonged Antarctic expedition, finding substantial inter-individual variability. Their protocol, however, was built for passive cold adaptation, not for dexterous motor tasks involving rapid finger displacement and the motion blur that comes with it — exactly the condition guitar playing under CPT stress produces.

Tracking a hand reliably under that motion blur is, in turn, a computer vision problem, and one with its own unresolved gap between domains. Bochkovskiy et al. [14] introduced

YOLOv4, a state-of-the-art, low-latency object detector — but, like most detection architectures, it was developed and benchmarked on RGB imagery, and its accuracy drops sharply on thermal input. Guo et al. [15] characterized this RGB-to-thermal domain gap systematically, showing that models trained exclusively on RGB data degrade severely on infrared images, and establishing thermal-specific fine-tuning as a near-mandatory step for reliable deployment. Kuzdeuov et al. [16] addressed this shortage with OpenThermalPose, a 6,090-image thermal dataset with 17 anatomical keypoints in the MS COCO format; models fine-tuned on it substantially outperformed direct RGB-to-thermal transfers, confirming that thermal-specific annotation improves keypoint accuracy under low-contrast conditions. Their dataset, however, targets full-body pose, not the hand-specific, high-velocity keypoint tracking a guitarist's plucking hand demands.

Taken together, these distinct bodies of literature leave an unresolved gap at their convergence: a method to evaluate motor warm-up effectiveness that combines the controlled autonomic challenge characterized by Mourot et al. [10] and Jenkins et al. [11] with the non-invasive sensing validated by Khan et al. [1] and Hattersley et al. [2], while remaining resilient to the fast motion and sensor-drift artifacts that challenge standard tracking [14]–[16]. The present study builds upon the physiological hypotheses raised by Oliveira [9], replacing manual single-point extractions with the automated, wall-normalized pipeline introduced in the previous section. This framework enables comparison of hand-temperature trajectories while accounting for hand displacement and common temporal variation.

III. Materials and Methods

This section details the translation of the initial empirical observations into a controlled experimental design and outlines the computational architecture developed for objective thermographic measurement. The subsequent subsections present the participant and the experimental protocol (Sections III-A and III-B), detail the image-processing and temperature-reconstruction and normalization pipeline (Section III-C), and define the statistical framework employed to evaluate the two playing protocols (Section III-D).

A. Participants and Data Quality Criteria

We recruited eight advanced male acoustic guitarists (mean age: 27.2 ± 10.5 years; mean playing experience: 10.0 ± 4.2 years) from the Music Department of the University of Brasília (MUS/IdA–UnB). Two independent guitar professors assessed each participant's proficiency using the Royal Conservatory of Music (RCM) Guitar Syllabus, classifying three participants at level 8, three at level 7, and two at level 6 — all above our inclusion threshold of RCM level ≥ 6. All participants were right-handed and provided written informed consent under a protocol approved by the Institutional Review Board of the College of Health Sciences, UnB (CAAE 54174921.8.0000.0030). We excluded volunteers taking medications known to affect the autonomic nervous system (e.g., beta-blockers [17]), presenting flu-like symptoms, diagnosed with diabetes mellitus, or undergoing active cardiovascular drug treatment in the preceding two weeks.

One participant's dataset was excluded during the initial video quality-control stage, before temperature extraction, normalization, or any between-protocol thermal comparison. This participant completed the protocol in a setting different from the standardized environment used for the remaining sessions, and the recording did not include the fixed wall region required for wall-referenced normalization. Consequently, the recording could not be processed using the complete proposed pipeline, leaving a final analytical sample of $n = 7$.

B. Experimental Protocol and Apparatus

The seven sessions retained for analysis took place over 60 days in the same controlled environment: ambient temperature held at 18 °C (range: 17–19 °C), blackout curtains preventing uncontrolled light or heat exposure. To reproduce the pre-performance stress the instructor described, we applied an adapted CPT as a standardized cold-related stressor [10], [18], immersing the participant's right foot in a water bath held between 5 °C and 8 °C. A Teledyne FLIR C5 uncooled microbolometer camera [8] recorded each session from a tripod two meters away, aimed orthogonally at the participant's right hand.

After a standardized briefing and sheet-music distribution, each participant completed two experimental blocks. Block order was not randomized or counterbalanced: following the same protocol structure as Oliveira [9], all participants completed the unsupported block before the supported block, in the same session. The blocks were separated only by the procedures required between trials, and the duration of this interval was not recorded. Therefore, order, learning, residual CPT effects, fatigue, and potential carryover effects associated with the unmeasured inter-block interval cannot be separated from condition effects in this design. In the unsupported block, the participant underwent a 3-minute CPT immersion and then performed Heitor Villa-Lobos' Study No. 1 [19] for six minutes, at the fastest tempo achievable with precision, using the unsupported fingerstyle technique — fingers arched, never touching the string below the one being played. In the supported block, following an identical CPT induction, the participant instead performed a structured chromatic exercise at a fixed 80 bpm, using the supported (com apoio) technique: the right-hand index and middle fingers alternately plucked descending fretboard positions, each finger briefly resting on the adjacent lower string immediately after plucking, before the cycle repeated. Because the blocks also differed in musical material, tempo, and execution demands, the experiment compares the two protocols as administered rather than the isolated effect of hand support.

Initial hand temperature estimates differed between blocks (mean difference $+0.691$ °C, supported higher; paired $t = 2.474$, $p = 0.048$; Wilcoxon $p = 0.078$; bootstrap 95% CI $[0.178, 1.182]$ °C). Unlike the central comparison in Section IV-C, the parametric and non-parametric tests do not

TABLE I
INITIAL HAND TEMPERATURE ESTIMATES AFTER WALL-REFERENCED NORMALIZATION

Participant	Supported (°C)	Unsupported (°C)	Difference (°C)
VOL1	31.707	31.299	+0.408
VOL2	29.994	28.945	+1.049
VOL3	31.675	32.120	-0.445
VOL4	33.849	32.662	+1.187
VOL5	29.435	29.423	+0.012
VOL6	33.790	32.094	+1.696
VOL7	33.840	32.913	+0.927
Mean	**32.041**	**31.351**	**+0.691**

fully agree here, but neither supports equivalence of the initial temperature estimates. A separate resting pre-block thermal baseline was not recorded; therefore, the initial estimate from the first 10 frames of active playing was used as the closest available measure of starting temperature.

C. Thermal Image Processing Pipeline

The FLIR C5 records non-radiometric color-mapped video frames rather than direct temperature arrays. Temperature analysis followed four steps. First, color-based temperature estimates were reconstructed independently for each frame using its displayed thermal scale. Second, the hand ROI was tracked and refined. Third, common temporal variation was reduced using a static wall reference. Fourth, the wall-normalized trajectory was filtered and smoothed for statistical evaluation.

1) Thermal Extraction and Color Mapping: Each frame displayed a dynamic scale with bounds $T_{\min}(k)$ and $T_{\max}(k)$, extracted using EasyOCR. For frames with both bounds available, the mean RGB value across the six inner columns of each of the 294 vertical positions of the displayed color bar formed a frame-specific lookup table. A cKDTree was reconstructed from this table, and each target pixel $\mathbf{p}_k(x, y) = [R, G, B]^T$ was matched to the nearest palette entry by Euclidean RGB distance. The temperature estimate was

$$\tilde{T}_k(x, y) = T_{\max}(k) - \frac{j_k^*(x, y)}{H - 1} \left[T_{\max}(k) - T_{\min}(k)\right], \quad (1)$$

where $H = 294$ and $j_k^*(x, y)$ is the nearest color-bar position. The mapping was reconstructed independently for every frame, thereby accounting for the dynamic temperature scale displayed by the camera. Across the dataset, the nearest-color lookup produced a mean Euclidean RGB distance of 25.27 ± 26.85 arbitrary units between each matched pixel and its assigned palette entry. Because RGB distance does not translate linearly into temperature error, this metric characterizes the color-matching step itself and should not be interpreted as an end-to-end temperature-reconstruction error.

Twelve static FLIR images containing radiometric metadata were used for an exploratory comparison with temperature maps derived from the embedded raw radiometric arrays. For each image, the radiometric temperature was calculated using the camera-specific Planck parameters:

$$T = \frac{B}{\ln\left(\dfrac{R_1}{R_2(S/\varepsilon + O)} + F\right)} - 273.15, \quad (2)$$

where S is the raw sensor value, ε is the image's embedded emissivity setting, and R_1, R_2, B, F, and O are camera-specific metadata parameters. This radiometric reconstruction was used as a reference for exploratory comparison only and was not used to modify the experimental temperature maps.

Across the 14 recordings, both $T_{\min}(k)$ and $T_{\max}(k)$ were available for direct OCR-based extraction in 5,657 of 5,677 frames (99.65%). The remaining 20 frames (0.35%) were excluded from temperature reconstruction because at least one of the two temperature bounds was unavailable. No interpolation or replacement was applied to these frames. The largest consecutive gap comprised nine frames.

The exploratory radiometric comparison highlighted a limitation of the available reference data. Although the embedded radiometric RAW data provided a direct temperature reference, the limited set of reference images did not support the establishment of a robust and generalizable color-to-temperature mapping across different scenes and thermal ranges. Therefore, these images were used only to assess the feasibility of the reconstruction approach and were not used to calibrate the experimental recordings.

2) Hand Tracking and Region of Interest Segmentation: Post-CPT peripheral cooling reduces thermal contrast between skin and background, and fast finger motion and fretboard occlusion make hand tracking particularly challenging. To prevent data leakage, the train/validation split was enforced at the participant level. Five participants (298 annotated thermal frames) were used for model fitting, while two participants (119 frames) were held out from fitting and used for validation and checkpoint selection; no participant, recording, or frame contributed to both sets. The six keypoints represented the tip and base of the thumb, index, and middle fingers. These seven participants were the same seven retained for the thermal analysis. The reported performance metrics refer exclusively to the two participant-disjoint validation subjects used for checkpoint selection ($\mathrm{mAP}_{50}(B) = 0.706$, $\mathrm{mAP}_{50\text{–}95}(B) = 0.236$, $\mathrm{mAP}_{50}(P) = 0.833$, and $\mathrm{mAP}_{50\text{–}95}(P) = 0.443$); no independent test set was available. Because the final ROI is derived from keypoint coordinates and subsequent clustering, keypoint localization is more directly relevant to downstream extraction than strict box overlap.

These keypoints defined the center and radius of a circular region of interest wrapping the hand — but that circle still includes non-biological pixels, such as the guitar top or fretboard, whenever they fall within its bounds. One-dimensional K-means clustering ($K = 3$, fixed random seed) was applied to the reconstructed temperatures within each frame's ROI. The cluster with the highest thermal centroid was retained as the hand-tissue candidate $\Omega_{\mathrm{vol}}(k)$, reducing contamination from the typically cooler instrument and background, and

yielded a mean sampling density of 676.5 ± 239.2 pixels per frame across the dataset. Fig. 1 illustrates this step: panel (a) shows the raw circular ROI, still contaminated by background structure; panel (b) shows the same region after clustering, containing only the hand.

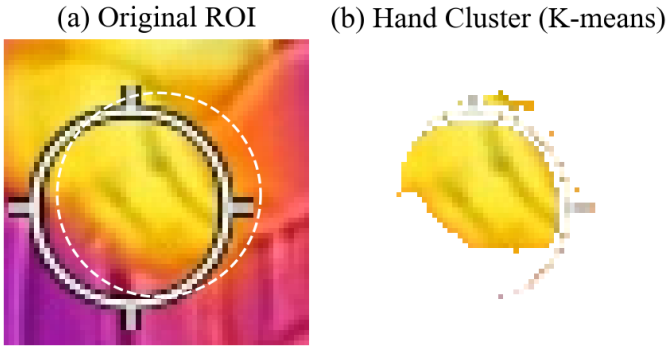


Fig. 1. Dynamic segmentation of the biological ROI, comparing (a) the original cropped thermographic ROI against (b) the isolated target hand cluster ($\Omega_{\text{vol}}(k)$) filtered via 1D K-means clustering.

3) Wall-Referenced Normalization: After frame-specific reconstruction, the trajectories could still contain common temporal shifts associated with dynamic scaling and internal correction processes.

To detect and remove these jumps, we designated a second, deliberately uninteresting region of the frame: a 10-pixel-radius circle on the stationary laboratory wall (Ω_{wall}), far from the hand. Assuming its physical temperature remained approximately stable during each recording, changes in its reconstructed temperature were treated as estimates of common temporal instrumental variation. We defined a wall-reference level, $T_{\text{wall_ref}}$, as the spatial-temporal median of Ω_{wall} over the initial 20 seconds of each recording (sampled at 1 frame/s; $k \in [1, 20]$):

$$T_{\text{wall_ref}} = \text{median}\left\{\tilde{T}_k(x,y) \mid (x,y) \in \Omega_{\text{wall}},\ k \in [1,20]\right\} \tag{3}$$

For every subsequent frame, we computed the instantaneous deviation of the wall from this anchor, $\epsilon_{\text{sensor}}(k)$, and subtracted it from the hand's raw average temperature to obtain the wall-normalized hand-temperature signal:

$$T_{\text{corr}}(k) = \bar{T}_{\text{raw}}(k) - \left(\text{median}\{\tilde{T}_k(x,y) \in \Omega_{\text{wall}}\} - T_{\text{wall_ref}}\right) \tag{4}$$

Fig. 2 shows a representative recording in which step-like variations in the wall-reference signal are subtracted from the reconstructed hand temperature, yielding the normalized and subsequently smoothed trajectory.

Across the 14 recordings, the wall-reference ROI showed a temporal variability of 0.443 ± 0.259 °C, expressed as the mean ± standard deviation of the within-recording temporal standard deviation.

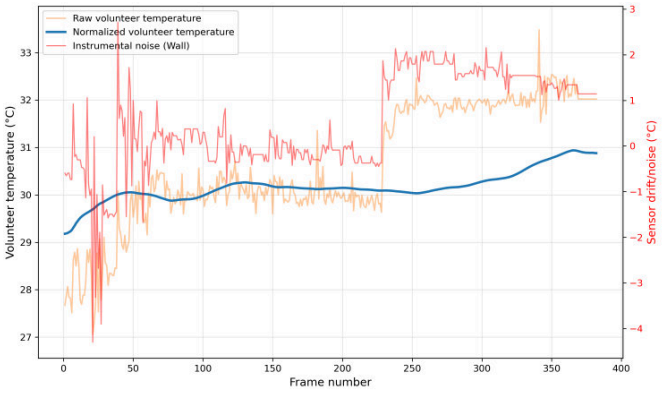


Fig. 2. Wall-referenced normalization for a representative recording (VOL5). Step-like variations in the wall-reference signal are subtracted from the reconstructed hand temperature, yielding the normalized and subsequently smoothed hand-temperature trajectory.

4) Statistical Filtering and Feature Extraction: Even after wall-referenced correction, brief finger-string occlusions can produce transient outliers unrelated to either sensor drift or physiology. We removed these using an Interquartile Range (IQR) filter — discarding points outside $[Q_1 - 1.5 \cdot \text{IQR},\ Q_3 + 1.5 \cdot \text{IQR}]$ and reconstructing them via local linear interpolation — followed by a Locally Weighted Scatterplot Smoothing (LOWESS) regression (first-degree local polynomial, span $\alpha = 0.15$) to reduce short-term fluctuations, yielding the final smoothed trajectory $T_{\text{smooth}}(k)$.

From $T_{\text{smooth}}(k)$, we extracted three macro-physiological features for comparison across conditions: the initial temperature estimate (T_{initial}), defined as the mean of the first $N = 10$ frames at the beginning of active playing; the final temperature estimate (T_{final}), the mean of the last $N = 10$ frames; and the sustained peak temperature ($T_{\text{pico_10f}}$), the maximum value of a 10-frame moving average, chosen to capture genuine thermal peaks while suppressing non-representative spikes.

D. Statistical Analysis

1) Temperature Kinetics: To compare the two fingering techniques, we first needed a small set of numbers summarizing each trial's thermal behavior over time, rather than the full frame-by-frame curve. As described above, we used the initial (T_{initial}), final (T_{final}), and sustained peak ($T_{\text{pico_10f}}$) temperature estimates extracted directly from $T_{\text{smooth}}(k)$, and derived final gain (ΔT) and peak gain ($\Delta T_{\text{pico_10f}}$) from these as detailed in Section III-D2.

Given the small sample size ($n = 7$), test selection did not rely on a single decision rule. We assessed distributional assumptions using the Shapiro-Wilk test and visual inspection, but, recognizing the limited diagnostic power of normality tests at this sample size, we did not use the outcome to choose between a parametric and a non-parametric test. Instead, the primary between-condition comparison used a paired Student's t-test, complemented throughout by a Wilcoxon signed-rank test and a participant-level bootstrap confidence interval (Section III-D2), so that conclusions rest on convergence across

methods rather than on a single test or a single p-value. Effect sizes were reported throughout using Cohen's d, and all computations were performed in SPSS, with the exception of the 10,000-iteration bootstrap confidence interval reported in Section IV-C, which was computed separately in Python (NumPy/SciPy).

2) Statistical Features and Testing: Statistical analyses used participant-level thermal features extracted from $T_{\mathrm{smooth}}(k)$. Final gain was calculated as $\Delta T = T_{\mathrm{final}} - T_{\mathrm{initial}}$, and sustained peak gain as $\Delta T_{\mathrm{pico_10f}} = T_{\mathrm{pico_10f}} - T_{\mathrm{initial}}$. Comparisons between protocols used paired observations. Given $n = 7$, the paired Student's t-test was complemented by a Wilcoxon signed-rank test and a participant-level bootstrap confidence interval (10,000 iterations, 2.5th and 97.5th percentiles). Final gains (ΔT) were tested against zero using one-sample t-tests, with effect sizes reported using Cohen's d. Sustained peak gains ($\Delta T_{\mathrm{pico_10f}}$) were summarized descriptively only. Because the peak metric is defined as the maximum of a moving-average trajectory, it was not used for confirmatory hypothesis testing against zero. Primary analyses were performed in SPSS, and bootstrap confidence intervals were computed in Python (NumPy/SciPy). Although electrocardiographic recordings were collected, HRV outcomes were not analyzed in the present study.

IV. Results

A. Effect of Wall-Referenced Normalization on Thermal Gain Estimates

Before normalization, common temporal instrumental variation could affect the apparent comparison between the two protocols. Before correction, the raw signal told a clear story: the unsupported technique appeared to warm the hand roughly three times more than the supported one ($\Delta T = +1.107$ °C vs. $+0.384$ °C). Frame-by-frame inspection revealed step-like variations in the reconstructed wall-reference temperature, with larger changes observed during the unsupported block.

After wall-referenced normalization reduced the common temporal variation shared with the static reference region, the apparent comparison changed substantially. As Table II shows, the unsupported condition's apparent gain dropped from $+1.107$ °C to $+0.630$ °C, while the supported condition's gain increased from $+0.384$ °C to $+0.690$ °C — the two techniques ending up numerically close together, with neither retaining a clear advantage (Section IV-C confirms this residual difference is not statistically significant). The apparent difference in the unnormalized data was influenced by temporal variation shared with the wall-reference signal and was no longer detected after normalization.

B. Individual and Group-Level Thermal Responses

Averages, however, can hide sharp individual disagreement — and that is precisely what Table III reveals. Individual normalized thermal gains ranged from -0.912 °C to $+2.046$ °C: some participants who cooled under one technique warmed substantially under the other, and vice versa. Under the supported technique, five of seven participants warmed and two cooled; under the unsupported technique, six warmed and one cooled. The direction and magnitude of the thermal response varied substantially across participants.

TABLE II
Mean Thermal Gain (ΔT) Before and After Wall-Referenced Normalization

Condition	**Raw ΔT (°C)**	**Normalized ΔT (°C)**
Supported	+0.384	+0.690
Unsupported	+1.107	+0.630
Overall Mean	**+0.746**	**+0.660**

TABLE III
Individual Normalized Thermal Metrics per Participant and Exercise Condition

Participant	**ΔT (°C)**		**$\Delta T_{\mathrm{pico_10f}}$ (°C)**	
	Supported	**Unsupported**	**Supported**	**Unsupported**
VOL1	+2.046	+0.798	+2.102	+1.020
VOL2	-0.211	+0.510	+0.955	+1.061
VOL3	+0.967	-0.602	+1.034	+0.183
VOL4	+1.322	+0.151	+1.322	+0.151
VOL5	-0.912	+1.611	+0.000	+1.650
VOL6	+0.413	+0.925	+0.413	+1.505
VOL7	+1.204	+1.019	+1.484	+1.019

A second pattern emerges when comparing sustained peak temperatures ($\Delta T_{\mathrm{pico_10f}}$) against final net gains: sustained peak gains were greater than or equal to final gains across all participants. For the supported paradigm, the group's mean sustained peak reached 1.044 ± 0.694 °C, compared to a final net gain of only 0.690 ± 1.003 °C; the unsupported condition shows the same pattern (0.941 ± 0.584 °C peak vs. 0.630 ± 0.706 °C final). The highest sustained temperature increase generally occurred before the end of the six-minute protocol, followed by a partial decline in reconstructed skin temperature.

C. Statistical Comparison Between Exercise Conditions

With the analytical framework established, the central comparison returns to the instructor's original question: does one technique warm the hand more effectively than the other? A paired Student's t-test detected no statistically significant difference in final thermal gain between protocols ($t = 0.109$, $p = 0.917$, Cohen's $d_z = 0.041$). A Wilcoxon signed-rank test ($p = 0.813$) and a 10,000-iteration bootstrap confidence interval for the paired mean difference (95% CI $[-1.007, +0.968]$ °C) supported the same overall interpretation, demonstrating convergence across parametric, non-parametric, and resampling methods. This null result shows no evidence of a difference, not formal equivalence — a stronger claim that would require a dedicated equivalence test (e.g., TOST), which we did not conduct here. Because the blocks also differed in musical material, tempo, and execution demands, this comparison reflects the protocols as administered rather than the isolated effect of hand support.

A related but distinct question is whether either protocol produced a mean change different from zero relative to its initial temperature estimate. One-sample t-tests addressed this directly: for final net gain (ΔT), neither technique crossed the conventional significance threshold (unsupported: $p = 0.056$, $d = 0.893$; supported: $p = 0.119$, $d = 0.688$).

Sustained peak gains ($\Delta T_{\mathrm{pico_10f}}$) were summarized descriptively to characterize the observed thermal trajectories. Mean sustained peak gains were 1.044 ± 0.694 °C in the supported protocol and 0.941 ± 0.584 °C in the unsupported protocol. In five of seven participants in each condition, the sustained peak exceeded the final gain, whereas in the remaining two it was equal to the final gain. These values are reported descriptively to detail maximum thermal gain and are not interpreted as independent confirmatory evidence of warming.

V. Discussion

This study set out to test, under controlled conditions, an informal but consistent observation: that switching from an unsupported to a supported fingerstyle technique may help guitarists warm cold hands before performance. Along the way, it also had to solve a narrower but consequential engineering problem — reducing common temporal variation in color-mapped thermographic recordings of an uncooled thermal camera. Both threads converge in the results, though not in the way a first look at the data would suggest.

The methodological contribution is the more unambiguous of the two. Raw thermographic data initially suggested that the unsupported technique produced nearly three times more warming than the supported one ($+1.107$ °C vs. $+0.384$ °C) — a result that, on its face, would have contradicted the instructor's account rather than supported it. Isolating the static wall-reference region revealed why: this apparent advantage was no longer detectable after wall-referenced normalization, and the unsupported block involved different musical, tempo, and movement demands. After wall-referenced normalization, no statistically significant difference in final net thermal gain was detected between the two protocols ($+0.630$ °C vs. $+0.690$ °C, $p = 0.917$, $d_z = 0.041$) — a null result rather than a formally demonstrated equivalence (Section IV-C). In other words, the unnormalized signal produced an apparent difference that was no longer detectable after normalization. This directly illustrates the instrumentation risk flagged by Khan et al. [1] and Hattersley et al. [2], extended here from static clinical settings to a dynamic computer-vision pipeline tracking a moving target.

This wall-normalized result also complicates the one directly comparable prior study. Oliveira [9], using manual single-point extraction on an overlapping cohort, reported a significant within-condition warming effect for the supported technique ($p = 0.003$, $d = 1.42$), whereas the unsupported condition did not show a significant within-condition change ($p = 0.342$, $d = 0.34$) — a pattern that appeared to confirm the instructor's experience in a straightforward way. Under the present automated, drift-corrected pipeline, that clean asymmetry does not hold at the group level: the analysis provided no evidence of a difference in final net thermal gain between the two protocols. Rather than confirming or refuting the instructor's account directly, this indicates that the explanation is more nuanced than "one technique warms, the other does not." Sections IV-B and IV-C locate that nuance in inter-individual variability and in the peak-versus-final distinction, not in one fingering's simple superiority.

The distinction between peak and final temperature provides an additional descriptive feature of the observed thermal trajectories. In most participants, the maximum reconstructed thermal gain occurred before task completion and was followed by a partial decline. This temporal pattern is descriptive; the physiological mechanisms underlying it were not directly measured.

Potential drivers of these transient dynamics include posture-dependent isometric tension, differential forearm vs. intrinsic muscle recruitment [3], [4], or convective cooling from faster finger motion. Disentangling these mechanisms remains a subject for targeted investigation using electromyography or per-finger mapping.

The considerable inter-subject variability observed in this cohort — individual net gains ranging from -0.912 °C to $+2.046$ °C, with some participants cooling under one technique while warming substantially under the other — reinforces that any single mechanism above is unlikely to act uniformly across guitarists. Baseline vascular tone, the sympathetic pathways associated with peripheral vasoconstriction, and individual biomechanical habits are all plausible sources of this heterogeneity [10], [11], and they suggest that the instructor's long-standing empirical observation may reflect a combination of these mechanisms whose relative contribution varies across students.

Several limitations qualify these conclusions. Foremost, the CPT is a physical, pain-mediated stressor: it reliably activates the sympathetic pathway behind peripheral vasoconstriction, though it elicits only modest cortisol (HPA-axis) activation — the signature more tied to anticipatory, socially evaluated stress. A socially evaluated CPT variant elicits comparable sympathetic activation while substantially raising cortisol reactivity [20], suggesting our protocol captures the pathway relevant here, even if it under-represents the psychological weight of an actual recital. The final analytical sample comprised only seven advanced male guitarists, limiting statistical power and generalizability across gender, skill level, age, and instrument type. The experimental recordings contained color-mapped video rather than native radiometric arrays. Only twelve static FLIR images containing radiometric metadata were available, and not for every participant, for exploratory comparison, which was insufficient to establish an independently validated correction transferable across recordings. The radiometric reference images therefore served only to assess the feasibility of the reconstruction approach and were not used to calibrate the experimental recordings. The findings should consequently be interpreted primarily in terms of consistently processed within-

recording temperature changes. Although electrocardiographic recordings were collected, HRV outcomes were not analyzed in the present study; therefore, equivalent autonomic responses to the CPT across blocks were not verified. Furthermore, without a passive-recovery control condition following the CPT, thermal variations during playing cannot be isolated from spontaneous vascular re-warming. The wall-referenced procedure also assumes that the physical wall temperature remains approximately stable during each recording and does not recover information lost through quantization, compression, or motion blur. Additionally, the two conditions differed not only in fingering technique but also in musical material and tempo (a fixed-tempo chromatic exercise vs. a piece played at each participant's fastest precise tempo), which may itself contribute to the observed thermal differences independent of hand posture. Separately, the group-level comparison in Section IV-C reflects only the absence of a detectable difference, not formally demonstrated equivalence; a dedicated equivalence test would strengthen a future replication. Finally, the mechanistic hypotheses proposed above remain untested by this study's design and should be treated accordingly — as motivation for the next experiment, not as its findings.

VI. Conclusion

An automated pipeline combining hand tracking, frame-specific color-to-temperature reconstruction, and wall-referenced normalization was applied to thermographic recordings from seven advanced male guitarists. After normalization, no evidence of a difference in final thermal gain was detected between the two administered protocols. Mean final net thermal gains were positive in both conditions but did not differ significantly from zero, and individual responses varied substantially. Sustained peak metrics were retained as descriptive features of the thermal trajectories rather than as confirmatory evidence of warming.

Interpretation is limited by the small sample size, fixed block order, differences in initial temperature estimates and musical demands between protocols, and limited radiometric reference data. Accordingly, the results characterize the two protocols as administered and should not be interpreted as isolating the effect of supported plucking.

Future work will pursue four main directions: first, expanding the cohort to include female musicians, different skill levels, and a wider range of instruments to evaluate generalizability; second, transitioning to native radiometric video streams to remove reliance on displayed color scales; third, acquiring a larger, temperature-diverse set of radiometric reference images under controlled conditions for independent quantitative validation of temperature reconstruction; and fourth, analyzing the collected electrocardiographic data (HRV) alongside electromyography, per-finger mapping, and controlled tempo manipulation to assess whether autonomic markers explain the observed inter-participant variability. By mitigating common temporal variation shared with static reference regions, this pipeline is better positioned to support future studies under more ecologically valid stress conditions — such as a socially evaluated variant of the CPT [20] or real pre-performance settings — where non-physiological artifacts would otherwise be even harder to disentangle.

References

[1] Z. U. Khan, G. Iohom, and B. O'Donnell, "Assessing the onset of regional anaesthesia: The role of thermographic imaging," *Anesthesia Research*, vol. 2, no. 4, p. 27, 2025. [Online]. Available: https://www.mdpi.com/2813-5806/2/4/27

[2] J. Hattersley, C. Imray, and A. J. Wilson, "Changes in peripheral ir thermography and energy expenditure on cooling in men and women following sustained strenuous activity in antarctica," *Scientific Reports*, vol. 16, no. 1, p. 4931, 2026.

[3] G. J. Tortora and B. Derrickson, *Princípios de Anatomia e Fisiologia*, 14th ed. Rio de Janeiro: Guanabara Koogan, 2016, vol. 1.

[4] J. Napier and R. Tuttle, *Hands*, ser. Natural science. Princeton University Press, 1993. [Online]. Available: https://books.google.com.br/books?id=mNQM4jvG7hgC

[5] C. L. Chan, T. R. Driscoll, and B. Ackermann, "Effect of a musicians' exercise intervention on performance-related musculoskeletal disorders," *Medical Problems of Performing Artists*, vol. 29, no. 4, pp. 181–188, 2014.

[6] G. Rotter, K. Noeres, I. Fernholz, S. N. Willich, A. Schmidt, and A. Berghofer, "Musculoskeletal disorders and complaints in professional musicians: A systematic review of prevalence, risk factors, and clinical treatment effects," *International Archives of Occupational and Environmental Health*, vol. 93, no. 2, pp. 149–187, 2020.

[7] A. Carlevaro, *Serie Didáctica: No.2 – Right Hand Technique*, 2nd ed. Buenos Aires: Barry, 1967.

[8] FLIR Systems, *User's Manual FLIR Cx Series*, 2020.

[9] I. J. d. Oliveira, "Estudo sobre o processo de aquecimento da mão direita de violonistas," Dissertação de Mestrado em Engenharia Elétrica, Universidade de Brasília, 2023, pPGEE Publicação 804/23.

[10] L. Mourot, M. Bouhaddi, and J. Regnard, "Effects of the cold pressor test on cardiac autonomic control in normal subjects," *Physiological Research*, vol. 58, no. 1, pp. 83–91, 2009.

[11] N. D. M. Jenkins, N. G. Boyes, D. W. Jacob, B. Shariffi, S. A. Martin, and J. K. Limberg, "Habitual sleep duration is associated with heightened muscle sympathetic nerve activation and blood pressure reactivity in response to the cold pressor test," *Journal of Applied Physiology*, vol. 139, no. 5, 2025.

[12] A. Adamik, E. Mikołajczyk, and J. Szczechowicz, "Effects of motor control-based interventions on pain and functional outcomes in bowed string musicians: A systematic review," *Journal of Clinical Medicine*, vol. 15, no. 9, p. 3326, 2026.

[13] K. Shi *et al.*, "Contactless analysis of heart rate variability during cold pressor test using radar interferometry and bidirectional LSTM networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[14] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[15] Y. Guo, Y. Chen, J. Deng, S. Li, and H. Zhou, "Identity-preserved human posture detection in infrared thermal images: A benchmark," *Sensors*, vol. 23, no. 1, p. 92, 2022.

[16] A. Kuzdeuov, D. Taratynova, A. Tleuliyev, and H. A. Varol, "OpenThermalPose: An open-source annotated thermal human pose dataset and initial YOLOv8-pose baselines," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–8.

[17] L. A. Bortolotto and F. M. Consolim-Colombo, "Betabloqueadores adrenérgicos," *Revista Brasileira de Hipertensão*, vol. 16, no. 4, pp. 215–220, 2009.

[18] S. Pouwels *et al.*, "Utility of the cold pressor test to predict future cardiovascular events," *Expert Review of Cardiovascular Therapy*, vol. 17, no. 4, pp. 305–318, 2019.

[19] H. Villa-Lobos, *Douze Études pour Guitare*. Paris: Max Eschig, 1929.

[20] L. Schwabe, L. Haddad, and H. Schachinger, "Hpa axis activation by a socially evaluated cold-pressor test," *Psychoneuroendocrinology*, vol. 33, no. 6, pp. 890–895, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306453008000644

Analysis and Prediction of Medical Service Utilization Among Adults Aged 50 and Older: A Machine Learning Approach

1st Pedro Kauan Silveira Silva
Federal University of Ceará (UFC)
Crateús, Brazil
pedrokauan@alu.ufc.br

2nd Mikeyas Brito dos Santos
Federal University of Ceará (UFC)
Crateús, Brazil
mikeyas.lira@gmail.com

3rd Wesley Barbosa Silva
Federal University of Ceará (UFC)
Crateús, Brazil
wesley.barbosa@alu.ufc.br

4th Bruno Riccelli dos Santos Silva
Computer Systems Engineering Laboratory (LESC)
Federal University of Ceará (UFC)
Fortaleza, Brazil
bruno@lesc.ufc.br

5th Wellington Franco
Federal University of Ceará (UFC)
Crateús, Brazil
jwellingtonfranco@gmail.com

6th Paulo Cesar Cortez
Computer Systems Engineering Laboratory (LESC)
Federal University of Ceará (UFC)
Fortaleza, Brazil
cortez@lesc.ufc.br

7th Andressa G. Moreira
Computer Systems Engineering Laboratory (LESC)
Federal University of Ceará (UFC)
Fortaleza, Brazil
andressagomes@alu.ufc.br

***Abstract*—This study evaluates supervised machine-learning classifiers for a binary formulation of healthcare service utilization among adults aged 50 years and older using the National Poll on Healthy Aging dataset ($n = 714$). The evaluation used nested cross-validation and out-of-fold predictions while preserving the observed class distribution. Baseline, MDI-80%, cost-sensitive, ensemble, and threshold-selection scenarios were examined using class-sensitive metrics, balanced accuracy, MCC, AUROC, and PR-AUC. Overall, discrimination between the two utilization classes was limited, and several models reproduced the majority-class prediction profile. Cost-sensitive weighting and internal threshold selection recovered part of the minority class in some scenarios, but these changes were limited and involved trade-offs. The findings are preliminary and describe the difficulty of this prediction task; they do not support clinical conclusions, diagnostic use, triage, or deployment. The code is available at https://github.com/pkziinn10/Analyze-Medical-Services**

***Index Terms*—Public Health, Population Aging, Machine Learning, Data Analysis, National Poll on Healthy Aging.**

I. Introduction

Population aging is a global and irreversible phenomenon that drives a sustained increase in demand for healthcare services. By 2030, one in six people worldwide will be aged 60 years or older, and this population is projected to grow from 1 billion in 2020 to 2.1 billion by 2050 [1]. Approaches to predicting healthcare service utilization have evolved from traditional statistical methods to Machine Learning techniques capable of capturing nonlinear relationships among sociodemographic, clinical, and behavioral factors [2]. However, these approaches generally require specialized expertise in feature selection and parameter tuning, which hampers their systematic and comparable application across studies [3], [4].

Existing evaluations of predictive models applied to healthcare service utilization among older adults present four main limitations. First, many studies rely on small samples or restricted clinical databases, limiting the generalizability of their findings [2]. Second, comparisons across models tend to cover a narrow set of learning paradigms, without jointly addressing linear, probabilistic, instance-based, tree-based, neural network-based, and graph-oriented approaches within a single experimental protocol [5]. Third, the class imbalance commonly observed in epidemiological data is frequently addressed through artificial balancing techniques, which may distort the actual distribution of healthcare utilization outcomes and undermine the clinical interpretation of results [6]. Finally, few studies subject performance differences between models to formal statistical testing, generally limiting the analysis to comparisons of point estimates.

In this context, this work puts forward a comprehensive comparative study of supervised classifiers for predicting healthcare service utilization among adults aged 50 and older, addressing these limitations: we broaden the coverage of learning paradigms evaluated under a single experimental

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 315447/2023-0.

protocol, preserve the observed class distribution after binary problem formulation without artificial balancing strategies, and subject the results to formal statistical verification. Our main contributions are as follows:

- **Multi-paradigm comparative evaluation:** we compare classifiers from multiple categories of supervised learning under the same experimental protocol, covering linear, probabilistic, instance-based, tree-based, neural network-based, and graph-oriented approaches.
- **Preservation of the observed class distribution:** the original target was recoded as a binary outcome. Category 1 (zero or one distinct doctor consulted) was mapped to class 0, while categories 2 and 3 (two or more distinct doctors consulted) were mapped to class 1, defined as the positive class. This resulted in 583 participants (81.6527%) in class 1 and 131 participants (18.3473%) in class 0, with no artificial resampling applied.
- **Statistical verification by the Wilcoxon signed-rank test:** performance differences between models are assessed through the nonparametric Wilcoxon signed-rank test, allowing us to determine which observed differences are statistically significant.
- **Feature significance analysis by Mean Decrease Impurity (MDI):** through a feature selection step based on the MDI criterion, we identify the demographic and clinical factors with the greatest discriminative power regarding healthcare utilization behavior.

The remainder of this paper is organized as follows. Section II reviews related work on machine learning approaches for predicting healthcare service utilization among older adults. Section III details the proposed methodology, encompassing the dataset and feature description, exploratory data analysis, data preprocessing and feature selection, machine learning algorithms, experimental setup, and evaluation metrics. Section IV presents the empirical results and discussion, covering the predictive performance of the evaluated models and the identified methodological limitations. Finally, Section V concludes the work, highlighting its main contributions, limitations, and directions for future research.

II. Related Work

Recent literature has increasingly applied machine learning to predict healthcare service utilization among older populations across diverse geographic contexts. These studies can be broadly grouped by the scope of the predicted services: multisectoral and community-based utilization, and acute demand for hospital or emergency care.

Regarding general and multisectoral utilization, Sourkatti et al. [6] investigated the early identification of seniors with high health and social service usage in Finland. Comparing Logistic Regression and eXtreme Gradient Boosting (XGBoost) on a cohort of 33,374 individuals, they found that age and mental health issues were positive predictors, whereas prior hospital stays and urban residence indicated lower risk. Despite achieving only moderate performance (AUC of 0.61), the study highlighted the inherent complexity of predicting multisectoral usage given its highly diverse contributing factors. Similarly, Huang et al. [7] examined demand for Home and Community-Based Services (HCBS) in China. Using the Andersen behavioral model for feature selection on a sample of 15,312 older adults, their evaluation of five computational methods showed that Random Forest (RF) and XGBoost achieved the best performance (specificity $> 80\%$), identifying self-reported health, exercise habits, and education level as key demand drivers.

Other studies have focused specifically on acute care, such as hospitalization risk. Buenrostro-Mariscal et al. [8] applied Random Forest to longitudinal data from the Mexican Health and Aging Study (MHAS) to predict hospitalization, showing that functional factors (e.g., restrictions in daily activities) and socioeconomic variables (e.g., education) consistently outperformed traditional clinical conditions (e.g., diabetes and hypertension) as predictors. By capturing complex nonlinear interactions, their model achieved a balance between sensitivity (0.72) and specificity (0.49). Likewise, Jung et al. [9] conducted a pioneering study predicting hospitalization among Medicaid HCBS users in the United States. Comparing Logistic Regression, LASSO, RF, and XGBoost, the authors found Random Forest to be the most robust approach, with prior hospital length of stay and specific chronic conditions as key predictors, underscoring the value of predictive modeling for early interventions that improve quality of life and reduce costs.

In contrast, evaluating the trade-off between model complexity and practical utility, Jones et al. [10] compared Gradient Boosted Trees (GBT), neural networks, Random Forest, and Logistic Regression for predicting emergency department and hospital utilization among home care patients in Canada. Although GBT achieved the highest overall performance, the clinical gains over standard Logistic Regression were practically negligible (approximately 1% to 2% in sensitivity and specificity), raising doubts about whether the added complexity of advanced ML models consistently translates into meaningful practical advantages in clinical settings.

Overall, the prevalent success of tree-based ensembles (such as RF and XGBoost) and the recognition of socio-functional predictors provide a solid foundation for this field. However, as noted by Jones et al. [10], the marginal performance gains of complex models warrant careful scrutiny. Our work builds on this foundation by proposing a comprehensive, multi-paradigm comparison, extending beyond tree-based ensembles to include graph-based and instance-based models, using the NPHA dataset. Furthermore, unlike studies that omit formal hypothesis testing, we subject all performance differences to the Wilcoxon signed-rank test. By explicitly preserving the natural epidemiological prevalence rather than employing synthetic balancing, our evaluation directly addresses the methodological concerns raised in the literature, assessing whether algorithmic complexity statistically justifies its computational cost.

III. Methodology

This section details the methodological model developed to predict medical service utilization among older adults. The workflow is subdivided into five main components: the description of the NPHA dataset, the exploratory data analysis, the data preprocessing procedures (including target transformation and feature encoding), the theoretical background of the machine learning algorithms evaluated, and the thorough experimental setup designed to guarantee robust model evaluation, feature selection, and fair comparison.

A. Dataset and Feature Description

The data used in this study were obtained from the National Poll on Healthy Aging (NPHA) dataset, available through the UCI Machine Learning Repository [11] [1]. The dataset contains responses from 714 U.S. adults aged 50 years and older, 14 predictor variables describing sociodemographic characteristics and health conditions, and one target variable. For this study, the original number-of-doctors-visited target was transformed into a binary outcome. Original category 1, representing zero or one distinct doctor consulted, was mapped to class 0. Original categories 2 and 3, representing two or more distinct doctors consulted, were mapped to class 1. Class 1 was defined as the positive class. The resulting distribution comprised 131 participants (18.3473%) in class 0 and 583 participants (81.6527%) in class 1. The number of distinct doctors consulted was treated as an operational proxy for healthcare utilization and must not be interpreted as a validated clinical definition of high utilization. The available codebook was used to identify the observed variable encodings; however, its clinical interpretation requires manual review and is treated as a limitation of this analysis.

B. Exploratory Data Analysis

Prior to model development, an exploratory analysis was conducted to characterize the NPHA dataset and assess its suitability for predictive classification.

Figure 1 shows the frequency distribution of the 14 predictor variables, numerically encoded as categorical or ordinal integers per the dataset's data dictionary. Age is nearly constant across the sample, with almost all participants concentrated in a single category. Physical, mental, and dental health scores are concentrated around their intermediate categories, while Employment is dominated by a single category. The sleep-related binary variables (stress, medication, pain, and unknown factors keeping patients from sleeping) are skewed toward 0, except for bathroom needs, which is nearly balanced between categories. Trouble sleeping and prescription sleep medication are both skewed toward their higher categories, whereas race is concentrated in a single category and gender is roughly balanced. These distributions describe the observed sample and motivate an evaluation that preserves its empirical class distribution, without by themselves establishing predictive utility.

[1]National Poll on Healthy Aging (NPHA)

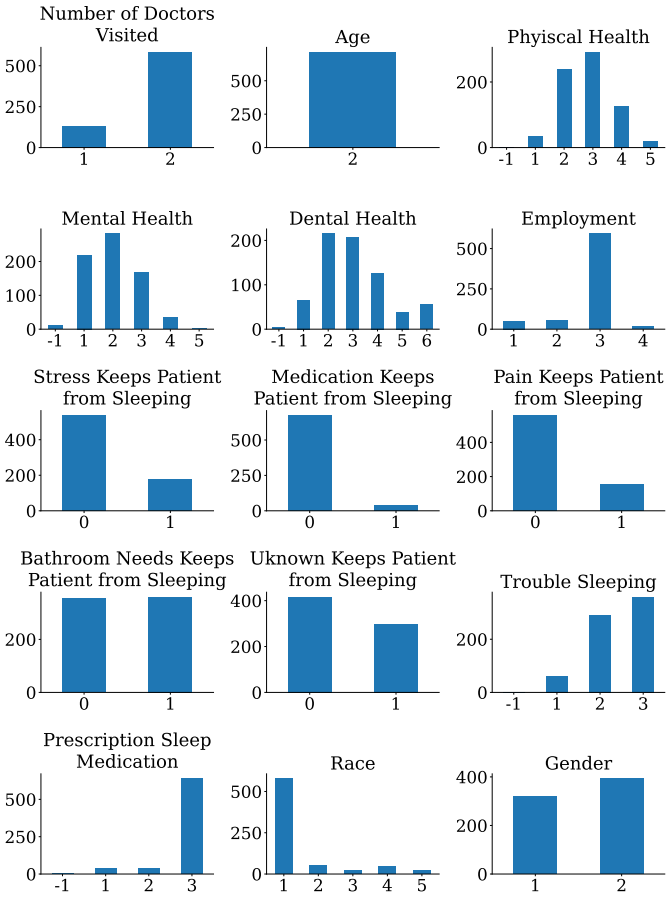


Fig. 1: Frequency distribution of the 14 predictor variables in the NPHA dataset.

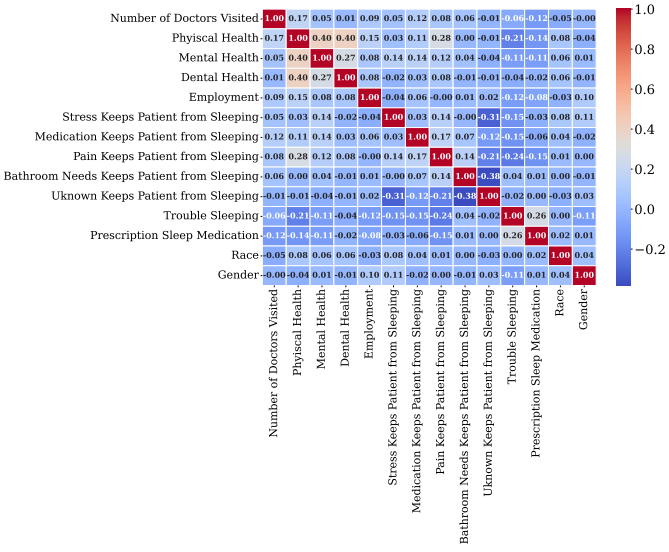
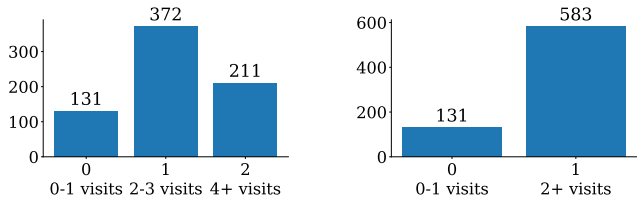


Fig. 2: Correlation matrix among the study variables.

Figure 2 depicts the pairwise Spearman rank correlation matrix. Correlations range from weak to moderate. The highest values ($\rho \approx 0.40$) are observed between physical health and both mental and dental health. Among the sleep-related variables, the strongest association is a negative correlation ($\rho \approx -0.31$) between bathroom needs and unknown factors keeping patients from sleeping. Demographic variables (race, gender) and most sleep-related indicators show near-zero pairwise associations with the remaining attributes. These monotonic associations do not constitute a complete assessment of multicollinearity.



(a) Original target distribution. (b) Binarized target distribution.

Fig. 3: Target variable distribution before (a) and after (b) binarization. Original category 1 was mapped to class 0, and categories 2–3 to class 1.

Finally, Figure 3 presents the target distribution before (a) and after (b) binarization, combining the categories representing two or more distinct doctors consulted. The resulting binary target contains 131 participants (18.3473%) in class 0 (zero or one doctor consulted) and 583 participants (81.6527%) in the positive class 1 (two or more doctors consulted), yielding an imbalanced classification task. This distribution was retained without oversampling, undersampling, or any other artificial resampling procedure, so that training and evaluation reflect the observed composition of the dataset.

C. Data Preprocessing

All preprocessing procedures were incorporated into the training workflow and fitted separately within each split of the nested cross-validation procedure. Numerical variables were imputed using the median of the corresponding training partition and subsequently standardized using `StandardScaler`. Categorical variables were imputed using the most frequent category in the training partition and transformed through One-Hot Encoding with `handle_unknown='ignore'`. These transformations were implemented through a `ColumnTransformer` pipeline.

For each inner or outer split, the imputation, encoding, and scaling components were fitted only on the corresponding training data. The validation or held-out test data were transformed using the fitted training components without refitting. This procedure prevents information from validation and test partitions from influencing the preprocessing parameters.

D. Feature Selection

Feature selection was evaluated in a dedicated MDI-80% scenario. Within each training partition, an auxiliary `RandomForestClassifier` from Scikit-learn [12] with 100 estimators was fitted to calculate Mean Decrease Impurity (MDI) feature importances. The transformed features were ordered by their importance, and the smallest subset whose cumulative importance reached at least 80% was retained.

The MDI selector was fitted independently inside each inner and outer training split. Consequently, the selected features may differ across folds; no global feature subset was defined before cross-validation. Feature-selection frequencies and fold-level selection details were retained as experimental artifacts to support reproducibility. Therefore, MDI results are reported as fold-dependent predictive information rather than stable clinical or causal evidence.

E. Class Weighting, Calibration, Ensemble, and Threshold Selection

No synthetic resampling method, such as SMOTE [13], was applied. In addition to the baseline and MDI-80% scenarios, a cost-sensitive scenario was evaluated for Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, and XGBoost. For the Scikit-learn classifiers, class weights were computed separately in each training partition as $w_c = n/(2n_c)$, where n is the number of training observations and n_c is the number in class c. For XGBoost, whose positive class is class 1, the corresponding parameter was $\texttt{scale_pos_weight} = w_1/w_0 = n_0/n_1$. Therefore, because class 1 is the majority class, this value is less than one (approximately $131/583 = 0.225$ in the complete dataset) and was recalculated in every training fold. Thus, class weighting was determined without using validation or outer test data.

Support Vector Machine probability scores were calibrated within the training workflow using `CalibratedClassifierCV` with a sigmoid method and 5-fold cross-validation. Calibration was performed only on training data, allowing calibrated scores to be generated without exposing the procedure to the outer test partition.

An ensemble scenario was also evaluated using three members selected independently in each outer fold from models that provided native or calibrated class-probability scores: KNN, Decision Tree, Random Forest, Support Vector Machine, MLP, Logistic Regression, XGBoost, Gaussian Naive Bayes, and AdaBoost. Nearest Centroid was excluded because its score is an uncalibrated distance-based decision function rather than a class probability; Optimum-Path Forest did not provide a continuous score. Ensemble members were selected independently in each outer fold using out-of-fold predictions generated exclusively within the corresponding outer training partition. The selection considered predictive performance and diversity among candidate predictions, without access to the held-out outer test partition. Each selected member received an equal weight of one third. Two ensemble decision rules were evaluated: majority voting based on class predictions and probability averaging based on the mean positive-class score.

Finally, the ensemble decision threshold was selected using only internal out-of-fold ensemble predictions. Candidate thresholds from 0.10 to 0.90, in increments of 0.05, were evaluated using macro F1-score as the primary selection criterion, with balanced accuracy and Matthews correlation coefficient used to resolve ties. The selected threshold was then applied to the held-out outer test scores. Therefore, ensemble composition, member weights, calibration, and threshold selection were all determined before evaluation on each outer test partition.

F. Machine Learning Algorithms

We evaluated eleven supervised classifiers spanning linear, probabilistic, instance-based, tree-based, neural, margin-based, and graph-based paradigms, plus a majority-class dummy baseline. All models were evaluated in baseline and MDI-80% scenarios; Decision Tree, Random Forest, Logistic Regression, SVC, and XGBoost were additionally evaluated with cost-sensitive weighting. KNN, Nearest Centroid, and OPF were excluded from probability averaging; Nearest Centroid provided a distance-based score only for discrimination metrics, whereas OPF provided no continuous score.

G. Experimental Setup and Model Training

The predictive models were evaluated using a nested stratified cross-validation procedure. The outer loop used `StratifiedKFold` with 10 folds, shuffling, and a fixed random seed of 42. Each outer fold defined a training partition and a held-out test partition, preserving the observed class distribution in both partitions. The training and test partitions used in each iteration were retained as experimental artifacts to support reproducibility.

Hyperparameter selection was performed exclusively within the training partition of each outer fold. For this purpose, an inner `StratifiedKFold` procedure with 5 folds, shuffling, and the same fixed random seed was applied. Candidate hyperparameter combinations were generated manually using `ParameterGrid` and evaluated through explicit iterations over the inner folds; `GridSearchCV` was not used. For each candidate, the macro F1-score, computed with `labels=[0,1]`, was calculated in each inner validation fold. The candidate with the highest mean macro F1-score across the five inner folds was selected. The hyperparameter grids ranged from 1 to 12 candidate combinations per classifier and were searched manually within each inner fold.

After hyperparameter selection, the selected configuration was fitted using only the corresponding outer training partition and evaluated on its held-out outer test partition. The held-out test partition was evaluated only once for each model within each predefined experimental scenario and outer fold. Thus, the study did not rely on a single global test set; performance was estimated by aggregating the out-of-fold predictions obtained across the 10 outer folds. Outer test data were not used for hyperparameter selection, preprocessing decisions, or any other adaptive training decision. All tested parameter combinations and inner-fold scores were retained as experimental artifacts to support reproducibility. This procedure separates model selection from held-out outer-fold evaluation, providing an internal estimate of generalization performance and preventing information from outer test partitions from influencing the selected configuration.

H. Evaluation Metrics

All metrics were computed per outer fold and aggregated over the out-of-fold (OOF) predictions, in which each instance is predicted exactly once, in the outer fold to which it belongs. Classification quality was summarized by a confusion matrix computed with `labels=[0,1]`. Class 0 corresponds to zero or one distinct doctor consulted, whereas class 1, corresponding to two or more distinct doctors consulted, is the positive class. Let TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the overall proportion of correct predictions and is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

For each class $c \in \{0, 1\}$, we report per-class precision, recall, and F_1:

$$\begin{aligned} \text{Precision}_c &= \frac{TP_c}{TP_c + FP_c}, \\ \text{Recall}_c &= \frac{TP_c}{TP_c + FN_c}, \\ F_{1,c} &= \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \end{aligned} \tag{2}$$

together with their macro and weighted averages. Because the classes are imbalanced, the main table reports class recall, macro F1, balanced accuracy, the Matthews correlation coefficient (MCC), AUROC, and PR-AUC; Table II provides detailed per-class metrics and aggregated OOF confusion-matrix counts for reference configurations.

Discrimination ability is evaluated with the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Precision–Recall area under the curve (PR-AUC, i.e., average precision). These metrics are computed from the continuous score of class 1, obtained from `predict_proba` when available or from the decision function otherwise; they are never computed from the discrete `predict` output. The receiver operating characteristic and precision–recall curves are built from the same continuous scores. OPF has no continuous score in the implemented pipeline, so its AUROC and PR-AUC are unavailable. For Nearest Centroid, these metrics use its distance-based decision function.

I. Statistical Evaluation

Because outer training partitions overlap across folds, the fold-level performance estimates are dependent. Therefore, the statistical analysis was restricted to exploratory paired comparisons.

The Wilcoxon signed-rank test was used to descriptively compare paired fold-level macro F1 values without assuming that their differences follow a normal distribution.

Following Demšar [14], we applied the paired nonparametric Wilcoxon signed-rank test to compare, for each of the eleven classifiers, the macro F1-score obtained in the 80% MDI feature-selection scenario with the macro F1-score of the majority-class dummy baseline. For each contrast, the two macro F1 values were paired within the same outer fold, resulting in ten dependent pairs.

Because eleven contrasts were evaluated simultaneously, the resulting p-values were adjusted using Holm's step-down correction. As noted by Demšar [14], cross-validation folds from a single dataset are not independent. Therefore, corrected p-values are used only for descriptive comparison and must not be interpreted as conclusive population-level evidence of superiority of any classifier over another classifier or over the dummy baseline.

IV. Results and discussion

We report the ten-fold out-of-fold results for the binary problem in which class 1, corresponding to two or more distinct doctors consulted, is the majority class (81.6% of the 714 instances), whereas class 0 is the minority class (18.4%). Given this imbalance, accuracy alone is not informative: the majority-class dummy baseline already attains approximately 81.6% accuracy. Therefore, the interpretation emphasizes per-class recall, balanced accuracy, MCC, AUROC, and PR-AUC.

In the baseline scenario, most classifiers predict the majority class for most observations, yielding low recall for class 0 and accuracy close to the majority-class baseline. Consequently, macro F1 remains close to 0.45 and balanced accuracy near 0.5 for several models. Some classifiers, including GaussianNB and Nearest Centroid, recover part of class 0. For models with continuous scores, AUROC values are generally close to 0.5, indicating limited discrimination between the two classes. The MDI-80% feature-selection scenario shows a comparable overall pattern, with modest changes in class-0 recall for tree-based and MLP models. OPF does not provide continuous scores in the implemented pipeline; therefore, its AUROC and PR-AUC are unavailable.

In the cost-sensitive scenario, where class weights are derived from the prevalence observed in each training fold, class-0 recall increases for Logistic Regression, Random Forest, XGBoost, and Decision Tree, accompanied by reduced recall for class 1. The Support Vector Classifier retains the majority-class prediction profile in this scenario.

The ensemble majority rule, probability rule, and fixed threshold of 0.5 produce the majority-class prediction profile. In contrast, the threshold selected from internal out-of-fold scores was 0.80 across folds and recovered a fraction of class 0 ($\text{Recall}_0 = 0.129 \pm 0.108$) while retaining $\text{Recall}_1 = 0.879 \pm 0.114$. This change is accompanied by lower overall accuracy and a modest increase in macro F1. Table III reports all ensemble configurations. Across the evaluated scenarios, AUROC remains close to 0.5 and PR-AUC remains near the class prevalence, consistent with limited discriminative capacity for this task.

A. Statistical Comparison with the Majority Baseline

As an exploratory check, each classifier in the MDI-80% scenario was paired with the majority-class dummy baseline on the same outer folds using the paired Wilcoxon signed-rank test on macro F1 ($n = 10$ dependent pairs per contrast). The p-values were corrected for eleven contrasts using the Holm procedure. After correction, no contrast had p-Holm below 0.05. Given fold dependence and the limited number of folds, these values are descriptive and do not provide conclusive evidence of superiority over the baseline.

B. Feature Importance

Feature importance was assessed by a Mean Decrease Impurity (MDI) random-forest selector, refit inside each fold on the training data and retaining features until 80% of cumulative importance was reached. The selector was therefore internal to each fold and did not use the held-out outer test data. Table V reports mean MDI scores by feature origin across folds. These scores are descriptive measures of predictive importance and do not establish causal or clinical effects.

C. Limitations

The findings are subject to several limitations. The study is cross-sectional and uses a single dataset (NPHA, $n = 714$), so the results reflect only the recorded population and moment of collection. The target classes are imbalanced, with two or more distinct doctors consulted representing about 82% of the observations, and most models predicted the majority class, resulting in low recall for the minority class. Feature importance and model behavior varied across folds, and MDI-based feature selection was recomputed inside each fold; therefore, the reported importance values are not stable enough for strong conclusions.

A search for accessible external datasets with a compatible target and variable structure did not identify a suitable dataset for external validation; potential alternatives required restricted access or used non-comparable outcome definitions. Although the evaluation pipeline can be applied to other compatible datasets, its predictive performance cannot be assumed to transfer to other populations, settings, or data-collection procedures. The limited discrimination observed also indicates that the recorded variables and binary outcome form a difficult prediction task in this dataset. These limitations restrict interpretation of the reported results to the specific data and protocol used.

V. Conclusion

This study evaluated supervised machine-learning classifiers for a binary formulation of healthcare service utilization among adults aged 50 years and older using the NPHA dataset ($n = 714$). The original target was recoded so that class 0 represents zero or one distinct doctor consulted and class 1 represents two or more distinct doctors consulted. The

TABLE I: Performance across scenarios. Values are mean (standard deviation) over ten outer folds. Rec_0/Rec_1 are recalls for classes 0 and 1; BA = balanced accuracy; PR-AUC = average precision for class 1. Sc.: BL = baseline; MDI = MDI-80%; COST = cost-sensitive.

Model	Sc.	Rec_0	Rec_1	F1-macro	BA	MCC	AUROC	PR-AUC
KNN	BL	0.015(0.032)	0.942(0.027)	0.448(0.032)	0.479(0.025)	-0.070(0.090)	0.476(0.090)	0.814(0.039)
DT	BL	0.161(0.077)	0.817(0.081)	0.485(0.029)	0.489(0.031)	-0.019(0.063)	0.462(0.053)	0.806(0.023)
RF	BL	0.023(0.036)	0.940(0.033)	0.453(0.033)	0.481(0.027)	-0.058(0.094)	0.518(0.064)	0.834(0.037)
SVC	BL	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.501(0.096)	0.836(0.045)
MLP	BL	0.046(0.039)	0.911(0.063)	0.459(0.032)	0.478(0.032)	-0.051(0.091)	0.515(0.064)	0.845(0.034)
LR	BL	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.561(0.081)	0.862(0.036)
XGB	BL	0.030(0.039)	0.950(0.044)	0.461(0.032)	0.490(0.027)	-0.025(0.079)	0.532(0.088)	0.847(0.036)
GNB	BL	0.236(0.162)	0.770(0.129)	0.493(0.056)	0.503(0.059)	0.009(0.115)	0.552(0.099)	0.862(0.037)
NC	BL	0.541(0.139)	0.504(0.045)	0.458(0.061)	0.523(0.087)	0.035(0.135)	0.542(0.095)	0.856(0.040)
OPF	BL	0.184(0.075)	0.772(0.104)	0.478(0.060)	0.478(0.064)	-0.034(0.113)	–	–
AdaB	BL	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.482(0.040)	0.813(0.010)
Dummy	BL	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.500(0.000)	0.817(0.004)
KNN	MDI	0.085(0.077)	0.926(0.051)	0.491(0.058)	0.505(0.043)	0.026(0.149)	0.489(0.086)	0.813(0.033)
DT	MDI	0.130(0.090)	0.877(0.098)	0.492(0.044)	0.504(0.035)	0.013(0.077)	0.508(0.064)	0.823(0.025)
RF	MDI	0.069(0.057)	0.938(0.037)	0.487(0.047)	0.504(0.034)	0.017(0.113)	0.515(0.101)	0.827(0.050)
SVC	MDI	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.549(0.105)	0.844(0.040)
MLP	MDI	0.076(0.061)	0.954(0.041)	**0.498(0.044)**	0.515(0.029)	0.069(0.134)	0.520(0.092)	0.843(0.042)
LR	MDI	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.542(0.089)	0.846(0.047)
XGB	MDI	0.069(0.067)	0.952(0.037)	0.489(0.046)	0.510(0.026)	0.029(0.098)	0.505(0.063)	0.833(0.025)
GNB	MDI	0.008(0.024)	0.978(0.020)	0.451(0.023)	0.493(0.016)	-0.033(0.073)	0.509(0.072)	0.832(0.033)
NC	MDI	0.503(0.148)	0.547(0.102)	0.471(0.061)	0.525(0.073)	0.043(0.120)	0.533(0.102)	0.840(0.051)
OPF	MDI	0.282(0.134)	0.714(0.054)	0.488(0.051)	0.498(0.064)	-0.006(0.110)	–	–
AdaB	MDI	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.491(0.028)	0.814(0.007)
DT	COST	0.214(0.125)	0.753(0.088)	0.478(0.057)	0.484(0.055)	-0.033(0.108)	0.485(0.054)	0.813(0.018)
LR	COST	**0.596(0.088)**	0.511(0.049)	0.477(0.047)	**0.553(0.062)**	**0.082(0.096)**	**0.574(0.075)**	**0.868(0.037)**
RF	COST	0.392(0.229)	0.638(0.158)	0.477(0.061)	0.515(0.075)	0.022(0.127)	0.541(0.085)	0.858(0.030)
SVC	COST	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.531(0.104)	0.847(0.041)
XGB	COST	0.305(0.109)	0.643(0.071)	0.461(0.041)	0.474(0.052)	-0.043(0.084)	0.509(0.085)	0.844(0.043)

KNN = K-Nearest Neighbors; DT = Decision Tree; RF = Random Forest; GNB = Gaussian Naive Bayes; NC = Nearest Centroid; AdaB = AdaBoost.
OPF provides no continuous score; its AUROC and PR-AUC are unavailable. NC metrics use its distance-based decision function.
Dummy appears only in BL. COST includes DT, LR, RF, SVC, and XGB only.
Bold values indicate the highest observed value for the highlighted metrics; they do not imply statistically confirmed superiority.

TABLE II: Detailed per-class performance and aggregated OOF confusion-matrix counts for reference configurations. Performance values are mean (standard deviation) over ten outer folds.

Configuration	Sc.	$Prec_0$	Rec_0	$F1_0$	$Prec_1$	Rec_1	$F1_1$	F1-weighted	TN	FP	FN	TP
Dummy	BL	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.817(0.004)	1.000(0.000)	0.899(0.002)	0.734(0.006)	0	131	0	583
GaussianNB	BL	0.197(0.104)	0.236(0.161)	0.198(0.095)	0.818(0.023)	0.770(0.129)	0.788(0.077)	0.680(0.062)	31	100	134	449
Decision Tree	COST	0.156(0.089)	0.214(0.125)	0.178(0.099)	0.810(0.021)	0.753(0.087)	0.778(0.056)	0.668(0.049)	28	103	144	439
Ensemble, selected threshold	THR	0.209(0.176)	0.129(0.108)	0.142(0.107)	0.817(0.016)	0.879(0.114)	0.844(0.062)	0.715(0.050)	17	114	71	512

TABLE III: Ensemble results. Values are mean (standard deviation) over ten outer folds. AUROC and PR-AUC use the shared ensemble probability score.

Decision rule	Rec_0	Rec_1	F1-macro	BA	MCC	AUROC	PR-AUC
Majority voting	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.535(0.099)	0.845(0.040)
Probability average	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.535(0.099)	0.845(0.040)
Probability, threshold 0.50	0.000(0.000)	1.000(0.000)	0.449(0.001)	0.500(0.000)	0.000(0.000)	0.535(0.099)	0.845(0.040)
Probability, selected threshold 0.80	0.129(0.108)	0.879(0.114)	0.493(0.058)	0.504(0.050)	0.022(0.121)	0.535(0.099)	0.845(0.040)

observed class distribution was preserved, with class 1 as the majority class. Models were evaluated through nested cross-validation and out-of-fold predictions under baseline, MDI-80% feature-selection, cost-sensitive, ensemble, and threshold-selection scenarios.

The results show that overall accuracy alone does not adequately describe performance in this imbalanced task. The majority-class dummy baseline already achieved approximately 81.6% accuracy while completely failing to recover class 0. Similarly, several evaluated models produced high recall for class 1 but low or null recall for class 0. This pattern explains why per-class recall, macro F1, balanced accuracy, MCC, AUROC, and PR-AUC were necessary for interpretation.

Across scenarios, the observed discriminative capacity was limited. AUROC values remained close to 0.5 for most models, indicating weak separation between the two utilization classes. The MDI-80% scenario produced results broadly comparable to the baseline. Cost-sensitive weighting increased class-0 recall for some models, whereas the threshold selected from internal out-of-fold scores recovered part of class 0; however, these changes were scenario-specific and accompanied by

TABLE IV: Paired Wilcoxon comparison of each classifier with the majority-class dummy baseline using macro F1, with Holm correction. Exploratory analysis with ten dependent folds.

Model vs. Dummy	W	p	p-**Holm**
KNeighbors	7.0	0.0742	0.4512
DecisionTree	1.0	0.0156	0.1719
RandomForest	8.0	0.0488	0.3906
SVC	0.0	1.0000	1.0000
MLP	6.0	0.0273	0.2734
LogisticRegression	0.0	1.0000	1.0000
XGBoost	3.0	0.0391	0.3516
GaussianNB	7.0	0.2813	1.0000
NearestCentroid	22.0	0.6074	1.0000
OPF	9.0	0.0645	0.4512
AdaBoost	0.0	1.0000	1.0000

$W = 0$ and $p = 1$ occur when macro F1 is identical to the dummy baseline.
Results are exploratory; no claim of statistical superiority is made.

TABLE V: Mean MDI attributed by the auxiliary random forest, averaged over folds. Values are descriptive importance scores, not causal or clinical evidence.

Feature origin	**Mean MDI**
Dental Health	0.1786
Mental Health	0.1396
Physical Health	0.1262
Trouble Sleeping	0.0900
Race	0.0834
Employment	0.0761
Gender	0.0665
Bathroom Needs Keeps Patient from Sleeping	0.0570
Unknown Keeps Patient from Sleeping	0.0565
Pain Keeps Patient from Sleeping	0.0406
Stress Keeps Patient from Sleeping	0.0400
Prescription Sleep Medication	0.0335
Medication Keeps Patient from Sleeping	0.0119
Age	0.0000

MDI importance varied across folds; values are not stable enough to support causal or clinical interpretation.
Feature selection was performed internally to each fold.

trade-offs in class-1 recall or overall accuracy. The ensemble rules and fixed threshold generally retained the majority-class prediction profile.

The nested evaluation protocol ensured that preprocessing, feature selection, weighting, calibration, ensemble composition, and threshold selection were determined without using the corresponding outer test fold. Nevertheless, the exploratory Wilcoxon–Holm comparison does not support conclusive claims that any classifier outperforms another classifier or the dummy baseline.

Therefore, the findings provide preliminary evidence of the difficulty of predicting this outcome from the available variables. They do not support clinical conclusions, diagnostic use, triage, or deployment. These findings must be interpreted alongside the study limitations.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 315447/2023-0 and Coordination of Superior Level Staff Improvement (CAPES)-Finance Code 001 for the support to this research

DECLARATION OF GENERATIVE AI USE

Generative AI tools were used only to assist with language revision, text organization, and clarity improvement. All scientific content, methodology, experiments, results, and conclusions were verified and validated by the authors, who assume full responsibility for the integrity, originality, and final content of the manuscript.

a

REFERENCES

[1] World Health Organization, "Ageing and health," WHO Fact Sheets, Oct. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[2] R. T. Olender, S. Roy, and P. S. Nishtala, "Application of machine learning approaches in predicting clinical outcomes in older adults: A systematic review and meta-analysis," *BMC Geriatrics*, vol. 23, p. 561, Sep. 2023.

[3] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," *BMC Medical Research Methodology*, vol. 22, no. 1, p. 282, 2022.

[4] J. Yang, A. A. Soltan, and D. A. Clifton, "Machine learning generalizability across healthcare settings: insights from multi-site covid-19 screening," *NPJ Digital Medicine*, vol. 5, no. 1, p. 69, 2022.

[5] O. G. Emi-Johnson and K. J. Nkrumah, "Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data," *Cureus*, vol. 17, no. 4, p. e82437, Apr. 2025.

[6] H. Sourkatti, J. Pajula, T. Keski-Kuha, J. Koivisto, M. Hilvo, and J. Lähteenmäki, "Predictive modeling for identification of older adults with high utilization of health and social services," *Scandinavian Journal of Primary Health Care*, vol. 42, no. 3, pp. 405–414, Jul. 2024.

[7] Y. Huang, T. Xu, Q. Yang, C. Pan, L. Zhan, H. Chen, X. Zhang, and C. Chen, "Demand prediction of medical services in home and community-based services for older adults in china using machine learning," *Frontiers in Public Health*, vol. 11, p. 1142794, 2023.

[8] R. Buenrostro-Mariscal, O. A. Montesinos-López, and C. Gonzalez-Gonzalez, "Predicting hospitalization in older adults using machine learning," *Geriatrics (Basel)*, vol. 10, no. 1, p. 6, 2025.

[9] D. Jung, H. A. Pollack, and R. T. Konetzka, "Predicting hospitalization among medicaid home- and community-based services users using machine learning methods," *Journal of Applied Gerontology*, vol. 42, no. 2, pp. 241–251, 2023.

[10] A. Jones, A. P. Costa, A. Pesevski, and P. D. McNicholas, "Predicting hospital and emergency department utilization among community-dwelling older adults: Statistical and machine learning approaches," *PLOS ONE*, vol. 13, no. 11, p. e0206662, 2018.

[11] "National poll on healthy aging (npha) dataset," UCI Machine Learning Repository, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/936/national+poll+on+healthy+aging+(npha)

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, Apr. 2018.

[14] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. Jan, pp. 1–30, Jan. 2006. [Online]. Available: https://www.jmlr.org/papers/v7/demsar06a.html

Análise Comparativa de Arquiteturas e Pré-processamentos para Segmentação de Vasos em Imagens de Fundoscopia

1st Emanuel Norjosa Luz
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
luzemanuel.n@gmail.com

2nd Bruno Riccelli dos Santos Silva
Departamento de Computação
Universidade Federal do Ceará
Crateús, Brasil
bruno.riccelli@ufc.br

3rdPaulo Cesar Cortez
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
paulo.cortez@ufc.br
comment

***Resumo*—A segmentação manual de vasos sanguíneos em exames de fundoscopia é trabalhosa, subjetiva e demanda tempo considerável do oftalmologista, o que motiva o uso de modelos de aprendizado profundo. Este trabalho compara seis codificadores integrados à arquitetura U-Net e avalia diferentes estratégias de pré-processamento (imagens originais RGB, tons de cinza, canal verde e canal verde com CLAHE) na base FIVES, adotando a métrica Dice como principal indicador de desempenho. Na base FIVES, o pré-processamento com canal verde seguido de CLAHE apresentou os melhores resultados médios, com destaque para o VGG19 (83,61%) e a ResNet101 (82,81%), que superaram as demais arquiteturas avaliadas; esse mesmo pré-processamento também se mostrou superior às demais estratégias testadas, incluindo canal verde sem CLAHE, imagens RGB e tons de cinza. Para avaliar a generalização, foram conduzidos testes cross-dataset em duas bases não utilizadas no treinamento: as bases de dados dos estudos *A Fundus Image Dataset for AI-based Artery-Vein Vessel Segmentation* e *The RETA Benchmark for Retinal Vascular Tree Analysis*, nas quais o VGG19 e a ResNet101 mantiveram desempenho consistente, com Dice médio de 82,72% e 82,43% na Fundus-AVSeg e 79,30% e 78,25% na RETA, respectivamente.**

***Index Terms*—Segmentação de vasos sanguíneos, fundoscopia, U-Net, redes neurais profundas, pré-processamento de imagens, canal verde, CLAHE, EfficientNet, ResNet, Swin Transformer.**

I. INTRODUÇÃO

Milhões de pessoas no mundo apresentam casos de cegueira evitável, tanto por conta de doenças, como a retinopatia diabética, quando por questões do avanço da idade, como a degeneração macular. Isso vem representando um desafio significativo para os sistemas de saúde pública [1]. O diagnóstico precoce e o monitoramento regular dessas doenças são fundamentais para prevenir a perda irreversível da visão. Nesse contexto, a análise automatizada de imagens de fundoscopia tem se consolidado como uma ferramenta promissora, permitindo a detecção precoce de alterações na retina e auxiliando os especialistas na tomada de decisões clínicas [2]. Entre as estruturas retinianas de maior interesse clínico, os vasos sanguíneos destacam-se por sua relevância no diagnóstico de diversas patologias, uma vez que alterações em seu calibre, tortuosidade e padrão de ramificação podem indicar a presença de doenças sistêmicas, como hipertensão arterial e diabetes mellitus.

A segmentação precisa dos vasos sanguíneos em imagens de fundoscopia é uma tarefa desafiadora devido à grande variabilidade anatômica entre pacientes, à presença de patologias que podem obscurecer os vasos, e às condições de iluminação não uniformes durante a aquisição das imagens [13]. Além disso, a complexidade da estrutura vascular, que inclui vasos de diversos calibres, desde os principais troncos vasculares até os capilares de fino calibre, torna a segmentação manual um processo demorado e sujeito a variações inter e intra-observador. Diante desses desafios, o desenvolvimento de métodos automáticos baseados em aprendizado profundo tem se mostrado uma abordagem eficaz, superando as técnicas tradicionais de processamento de imagens em termos de precisão e robustez [3], [4].

Neste trabalho, são investigadas diferentes estratégias de pré-processamento e arquiteturas de redes neurais profundas para a segmentação de vasos sanguíneos em imagens de fundoscopia. O pré-processamento das imagens é uma etapa fundamental para melhorar a qualidade dos dados de entrada e facilitar a tarefa de segmentação [5]. Foram avaliadas quatro estratégias distintas de pré-processamento: imagens originais (três canais de cores), tons de cinza, canal verde e canal

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob o Processo nº 315447/2023-0, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Código de Financiamento 001, pelo apoio a esta pesquisa.

verde com aplicação do CLAHE (*Contrast Limited Adaptive Histogram Equalization*). A escolha do canal verde justifica-se pela absorção da luz pela hemoglobina na faixa do verde, o que proporciona maior contraste entre os vasos sanguíneos e o tecido retiniano circundante [1]. O CLAHE, por sua vez, foi adotado para melhorar o contraste local e uniformizar a iluminação, especialmente em regiões periféricas da retina que frequentemente apresentam baixa qualidade de imagem.

A arquitetura U-Net foi adotada como base para a segmentação, sendo uma das redes mais consolidadas para tarefas de segmentação de imagens médicas devido à sua capacidade de combinar informações contextuais e localizadas por meio de conexões *skip connections* [6]. Para avaliar o impacto da escolha do codificador, a U-Net foi combinada com seis diferentes *backbones*: EfficientNetB0 [14], MobileNetV2 [7], ResNet101 [4], VGG19 [3], SwinUNet [8] e ViTUNet [9]. Essas arquiteturas foram selecionadas por representarem diferentes paradigmas de redes neurais: as três primeiras são CNNs modernas com diferentes equilíbrios entre eficiência e capacidade representacional; a VGG19 é uma CNN mais profunda e com maior número de parâmetros; e os Transformers (SwinUNet e ViTUNet) representam a abordagem baseada em mecanismos de autoatenção, que têm demonstrado resultados promissores em diversas tarefas de visão computacional [8], [9].

Os experimentos foram conduzidos utilizando a base de dados FIVES (*Fundus Image Vessel Segmentation*) [10], que contém imagens de fundoscopia anotadas com as respectivas máscaras de segmentação dos vasos sanguíneos. Foram realizadas cinco execuções independentes para cada combinação de arquitetura e pré-processamento, totalizando 30 configurações experimentais. A métrica Dice foi adotada como principal indicador de desempenho, por integrar precisão e sensibilidade em uma única medida, sendo amplamente utilizada em tarefas de segmentação de imagens médicas [1], [2]. Além da qualidade da segmentação, o tempo total de treinamento foi registrado para cada execução, permitindo avaliar a eficiência computacional de cada combinação.

Com o objetivo de avaliar a capacidade de generalização dos modelos, também foram realizados testes em duas bases de dados distintas, não utilizadas durante o treinamento: a Fundus-AVSeg, proveniente do estudo *A Fundus Image Dataset for AI-based Artery-Vein Vessel Segmentation*(Fundus-AVSeg) [11], voltada para a segmentação de artérias e veias em imagens de fundoscopia, e a RETA, oriunda do benchmark *The RETA Benchmark for Retinal Vascular Tree Analysis*(RETA) [12], que fornece imagens e anotações específicas para a análise da árvore vascular retiniana. Essas bases apresentam características de aquisição, resolução e distribuição de vasos diferentes da FIVES, o que permite verificar a robustez das arquiteturas treinadas frente à variabilidade entre domínios.

Os resultados obtidos são apresentados e discutidos ao longo deste artigo, com o objetivo de identificar as melhores práticas para a segmentação de vasos sanguíneos em imagens de fundoscopia, considerando o equilíbrio entre desempenho, generalização e custo computacional.

II. DETALHAMENTO TÉCNICO

A. Base de dados

A primeira etapa do nosso trabalho consistiu em identificar e selecionar bases de dados que oferecessem não apenas imagens de fundoscopia em quantidade suficiente, mas também anotações manuais precisas e validadas por especialistas. A qualidade dos rótulos impacta diretamente a capacidade dos modelos de aprenderem padrões fiéis à realidade clínica.

Após analisar as opções disponíveis na literatura, optamos pelo base FIVE (Fundus Image Vessel Extraction). Ela é composto por 800 imagens no total, das quais 600 são destinadas ao treinamento e 200 ao teste, todas com resolução de 2048×2048 pixels. Além disso, o banco abrange uma variedade clínica relevante: as imagens incluem desde olhos saudáveis até casos com retinopatia diabética, glaucoma e outras alterações patológicas, expondo os modelos a diferentes níveis de contraste, iluminação e distorções morfológicas [10]. Essa diversidade é um diferencial importante, pois força os algoritmos a aprenderem características robustas, em vez de se ajustarem excessivamente a um único padrão visual. Dois exemplos de imagens da base de dados podem ser conferidas na figura 1.

A base FIVE atendia os requisitos de ter suas máscaras de segmentação delineadas por oftalmologistas experientes seguindo e por ser amplamente referenciado na comunidade científica, o que nos possibilita comparar nossas métricas com as de outros grupos que adotaram a mesma base e a mesma arquitetura de referência. Essa compatibilidade metodológica é fundamental para que possamos afirmar se as melhorias observadas vêm das modificações nos codificadores ou de meras particularidades do conjunto de dados.

Por fim, a divisão padronizada entre treino e teste estabelecida pelo próprio FIVE elimina ruídos na comparação inter-estudos trazendo mais confiabilidade na comparação dos trabalhos.

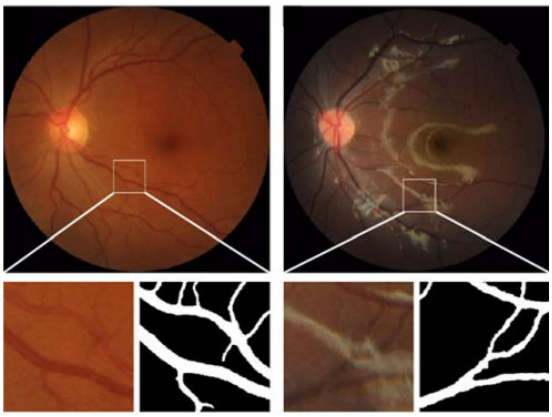


Figura 1. Exemplo de dados do banco de dados FIVE.

B. Pré-processamento das Imagens

O pré-processamento das imagens de fundoscopia é uma etapa fundamental para melhorar a qualidade dos dados de en-

trada e facilitar a tarefa de segmentação dos vasos sanguíneos. Neste trabalho, foram avaliadas cinco estratégias distintas de pré-processamento: imagens originais (três canais de cores), tons de cinza, canal verde e canal verde com aplicação do CLAHE. Todas as imagens foram redimensionadas para a resolução de 224×224 pixels, de modo a padronizar a entrada dos modelos e garantir a compatibilidade com as arquiteturas avaliadas.

A Figura 2 ilustra o resultado da aplicação de cada técnica de pré-processamento em uma imagem de exemplo da base de dados FIVES. A imagem original (RGB) é apresentada no canto superior esquerdo, seguida pelas três versões processadas utilizadas como entrada para os modelos de segmentação.

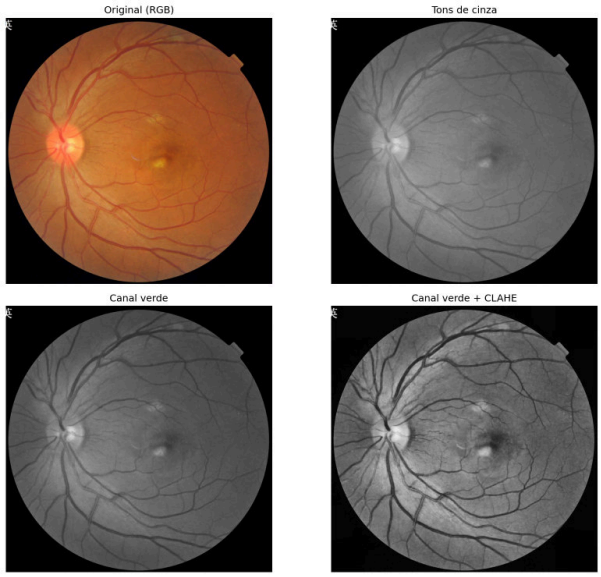


Figura 2. Exemplo das técnicas de pré-processamento aplicadas a uma imagem de fundoscopia: (a) Imagem Original (RGB); (b) Tons de Cinza; (c) Canal Verde; (d) Canal Verde com CLAHE.

As imagens originais foram utilizadas como linha de base para comparação. Essa abordagem mantém toda a informação cromática disponível, incluindo as diferenças de intensidade entre os canais vermelho, verde e azul, que podem ser relevantes para a caracterização de diferentes estruturas retinianas. Contudo, o uso de três canais aumenta a dimensionalidade dos dados e, consequentemente, o custo computacional do treinamento. As imagens em tons de cinza foram obtidas pela conversão direta da imagem RGB para um único canal de intensidade, reduzindo a dimensionalidade e eliminando informações potencialmente redundantes. Embora seja a abordagem mais simples, a conversão para tons de cinza pode resultar em perda de contraste entre os vasos e o fundo da retina, uma vez que não explora as propriedades espectrais específicas da hemoglobina.

O canal verde foi extraído diretamente da imagem RGB, selecionando apenas a matriz correspondente ao componente verde do espaço de cores. Essa escolha é fundamentada na física da interação da luz com os tecidos biológicos: a hemoglobina presente no sangue absorve fortemente a luz na faixa do verde (comprimento de onda de aproximadamente 540 nm), fazendo com que os vasos sanguíneos apareçam escuros neste canal, enquanto o tecido retiniano circundante reflete a luz, criando um alto contraste. Estudos prévios têm demonstrado que o canal verde oferece o melhor contraste para a segmentação de vasos em imagens de fundoscopia [5].

Por fim, o canal verde com aplicação do CLAHE foi gerado a partir da extração do canal verde seguida da aplicação do algoritmo CLAHE, que realiza uma equalização de histograma adaptativa por regiões, limitando o contraste para evitar a amplificação excessiva de ruído. O CLAHE divide a imagem em pequenas regiões (tiles) e aplica a equalização de histograma em cada uma delas, interpolando os resultados para suavizar as transições entre as regiões. O parâmetro *clipLimit* controla o fator de limitação do contraste, enquanto *tileGridSize* define o tamanho das regiões. Neste trabalho, foram utilizados os valores de *clipLimit* igual a 2,0 e *tileGridSize* igual a (8,8), que demonstraram um bom equilíbrio entre o realce do contraste e a preservação das estruturas finas dos vasos. O CLAHE é particularmente útil para imagens de fundoscopia, que frequentemente apresentam iluminação não uniforme e baixo contraste em regiões periféricas, melhorando a visibilidade dos vasos de fino calibre e facilitando a segmentação. A Figura 2 evidencia visualmente o efeito de cada técnica: o canal verde (Figura 2c) já apresenta bom contraste dos vasos, enquanto o VM (Figura 2d) realça adicionalmente os detalhes, e o CLAHE (Figura 2e) promove uma equalização que uniformiza a iluminação e realça vasos de menor calibre. A combinação do canal verde com o CLAHE explora sinergicamente as propriedades espectrais da hemoglobina e as técnicas de realce de contraste adaptativo, resultando em imagens de entrada com alta qualidade para os modelos de segmentação.

C. Backbones para Codificadores na Arquitetura U-Net

A substituição do codificador convencional da U-Net por backbones consolidados da visão computacional tem se mostrado uma estratégia eficaz para melhorar a segmentação de vasos sanguíneos [13]. Esses codificadores são responsáveis por extrair características hierárquicas da imagem, enquanto o decodificador reconstrói a máscara de segmentação. A escolha do backbone define o equilíbrio entre acurácia, custo computacional e capacidade de generalização, sendo agrupados em redes profundas, redes leves e modelos baseados em Transformer.

Entre as redes tradicionais profundas, a *VGG-19* [1] e a *ResNet-101* [2] oferecem alta capacidade representacional. A VGG-19 utiliza pilhas homogêneas de convoluções 3x3, proporcionando amplo campo receptivo, mas com elevado custo paramétrico. A ResNet-101, por sua vez, emprega conexões residuais que facilitam o treinamento de redes mais profundas, mitigando o desvanecimento do gradiente. Ambas extraem características semânticas robustas, sendo indicadas quando a precisão é prioritária e o recurso computacional não é limitado.

Para cenários com restrição de hardware ou exigência de inferência rápida, as redes leves *MobileNet* [3] e *EfficientNet-*

B0 [4] são alternativas vantajosas. A MobileNet reduz drasticamente parâmetros e latência por meio de convoluções separáveis em profundidade, mantendo desempenho adequado para vasos de médio calibre. A EfficientNet-B0 otimiza a relação profundidade-largura-resolução via escalonamento composto, oferecendo eficiência superior. Embora possam perder detalhes finos de microvasos, são ideais para sistemas de triagem automatizados e dispositivos embarcados.

Por fim, os Transformers, como o *ViT* [5] e o *Swin Transformer* [6], superam as limitações de dependência local das CNNs ao empregarem mecanismos de autoatenção global. O ViT processa a imagem em patches e captura interações de longa distância, mas requer muitos dados para convergir. O Swin Transformer introduz janelas de atenção deslocadas e estrutura hierárquica, combinando visão global e local com maior eficiência. Ambos, especialmente o Swin, obtêm excelente desempenho na segmentação de vasos tortuosos e finos, porém demandam maior poder computacional e conjuntos de dados extensos para evitar overfitting.

D. Métricas de Avaliação

O índice de Dice, também conhecido como coeficiente de Sorensen-Dice, é a métrica de qualidade mais amplamente utilizada em tarefas de segmentação de imagens médicas, por integrar precisão e sensibilidade em uma única medida. O Dice é definido pela Equação 1, onde VP representa o número de pixels verdadeiros positivos (vasos corretamente segmentados), FP o número de falsos positivos (fundo classificado como vaso) e FN o número de falsos negativos (vasos não detectados). O índice varia entre 0 (nenhuma sobreposição entre a segmentação predita e a máscara de referência) e 1 (sobreposição perfeita), sendo que valores mais próximos de 1 indicam melhor qualidade da segmentação. A escolha do Dice como métrica principal justifica-se pela sua capacidade de refletir o equilíbrio entre a detecção correta dos vasos e a precisão da segmentação.

$$Dice = \frac{2 \cdot VP}{2 \cdot VP + FP + FN} \tag{1}$$

O Dice foi adotado como métrica de validação durante o treinamento, sendo calculado ao final de cada época sobre o conjunto de validação para monitorar o progresso do modelo e selecionar os melhores pesos. Ao término do treinamento, o Dice também foi calculado sobre o conjunto de teste, permitindo avaliar a capacidade de generalização dos modelos. Para cada configuração experimental (combinação de arquitetura e pré-processamento), foram realizadas cinco execuções independentes de treinamento, com diferentes sementes aleatórias para a inicialização dos pesos e para a divisão dos dados. A partir dessas cinco execuções, foram calculadas a média e o desvio padrão do melhor Dice obtido durante a validação (best_val_dice) e do Dice final sobre o conjunto de teste. Essa abordagem garante a robustez estatística dos resultados e permite avaliar a consistência do desempenho dos modelos.

A sensibilidade, também denominada *recall* ou taxa de verdadeiros positivos, mede a capacidade do modelo em identificar corretamente os pixels pertencentes aos vasos sanguíneos, sendo definida pela Equação 2. Nessa formulação, VP representa o número de pixels verdadeiros positivos (vasos corretamente segmentados) e FN o número de falsos negativos (vasos não detectados pelo modelo).

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2}$$

Essa métrica varia entre 0 e 1, em que valores próximos de 1 indicam que o modelo é capaz de detectar a maior parte dos vasos presentes na máscara de referência, enquanto valores baixos revelam uma tendência à omissão de estruturas vasculares. No contexto da segmentação de vasos em imagens de fundoscopia, a sensibilidade é particularmente relevante, pois a não detecção de vasos pode comprometer análises clínicas subsequentes, como a identificação de estenoses, cruzamentos arteriovenosos e alterações no calibre vascular. Contudo, a sensibilidade deve ser analisada em conjunto com outras métricas, uma vez que um modelo excessivamente sensível pode classificar regiões de fundo como vasos, elevando o número de falsos positivos. Assim como o Dice, a sensibilidade foi calculada ao final de cada época sobre o conjunto de validação e, posteriormente, sobre o conjunto de teste, com média e desvio padrão obtidos a partir das cinco execuções independentes de cada configuração experimental.

A métrica Interseção sobre União(IoU), é amplamente utilizada em tarefas de segmentação por medir a razão entre a área de sobreposição e a área de união entre a segmentação predita e a máscara de referência. O IoU é definido pela Equação 3, onde VP representa o número de pixels verdadeiros positivos, FP o número de falsos positivos e FN o número de falsos negativos.

$$IoU = \frac{VP}{VP + FP + FN} \tag{3}$$

O índice varia entre 0 e 1, em que 0 indica ausência total de sobreposição entre a predição e a referência, e 1 representa sobreposição perfeita. Diferentemente do Dice, o IoU é mais sensível a discrepâncias entre as regiões preditas e as anotadas, penalizando de forma mais acentuada tanto os falsos positivos quanto os falsos negativos. Essa característica torna o IoU particularmente útil para avaliar a precisão espacial das segmentações vasculares, especialmente em cenários nos quais vasos de pequeno calibre podem ser facilmente omitidos ou confundidos com o fundo da imagem. No presente trabalho, o IoU foi calculado ao final de cada época sobre o conjunto de validação e, ao término do treinamento, sobre o conjunto de teste, com média e desvio padrão obtidos a partir das cinco execuções independentes de cada configuração experimental, seguindo o mesmo procedimento adotado para as demais métricas.

O tempo total de treinamento foi registrado para cada uma das cinco execuções independentes, compreendendo o período desde o início do carregamento dos dados até a conclusão da última época de treinamento, incluindo o tempo de validação ao final de cada época. A média dos tempos

de treinamento das cinco execuções foi calculada para cada combinação de arquitetura e pré-processamento, permitindo avaliar a eficiência computacional de cada configuração. O desvio padrão do tempo de treinamento também foi calculado, fornecendo uma medida da variabilidade do custo computacional entre execuções.

E. Hiperparâmetros e Configurações Experimentais

Para garantir a reprodutibilidade dos experimentos, foram definidos hiperparâmetros padronizados para todas as arquiteturas avaliadas, conforme apresentado na Tabela I. O treinamento foi conduzido com parada antecipada de paciência 10 e delta 0,001, e cada configuração foi executada cinco vezes com diferentes sementes aleatórias. Os experimentos foram realizados na estação de trabalho especificada na Tabela II.

Tabela I
HIPERPARÂMETROS UTILIZADOS NOS EXPERIMENTOS.

Parâmetro	**Valor**
Tamanho da imagem	224 x 224 pixels
Modo de entrada	1 canal (tons de cinza)
Batch size	16
Épocas máximas	50
Taxa de aprendizado inicial	0,001
Otimizador	Adam
Função de perda	Dice + BCE
parada antecipada	Paciência: 10 épocas, Delta: 0,001
Número de execuções	5
Num workers	4
Pin memory	Ativado

Tabela II
ESPECIFICAÇÕES DO HARDWARE UTILIZADO.

Componente	**Especificação**
Processador	AMD Ryzen 5 5600X (6 núcleos, 12 threads)
Placa de vídeo	NVIDIA RTX 3060 (12 GB VRAM)
Memória RAM	24 GB DDR4
Armazenamento	SSD 250 GB
Sistema Operacional	Ubuntu 20.04 LTS

III. ANÁLISE DE TREINO E TESTE

O código-fonte completo, os dados processados e os resultados detalhados de todos os experimentos estão disponíveis publicamente no repositório GitHub [18] para fins de reprodutibilidade e transparência científica.

A. Análise dos Tempos de Treinamento

O tempo de treinamento é um fator crítico para a viabilidade prática de modelos de segmentação, especialmente em cenários com recursos computacionais limitados ou que exigem rápida interação experimental. A Tabela III apresenta os tempos médios de treinamento (em segundos), acompanhados dos respectivos desvios padrão, para cada combinação de arquitetura e pré-processamento, obtidos a partir de cinco execuções independentes.

Inicialmente, observa-se que o treinamento com as imagens originais, que preservam os três canais de cores (RGB), demandou os maiores tempos para todas as arquiteturas avaliadas. O EfficientNetB0, que foi a rede que mais demorou, necessitou de 1040,6 segundos em média para concluir o treinamento com as imagens coloridas, enquanto a ResNet101 destacou-se como a mais eficiente, com 527,8 segundos, valor significativamente inferior aos demais modelos. Entre os Transformers, o Swin e o ViT exigiram 739,5 e 744,1 segundos, respectivamente.

Tabela III
TEMPOS DE TREINAMENTO (MÉDIA ± DESVIO, EM SEGUNDOS).

	Pré-processamento			
Modelo	Orig.	Cinza	Verde	V+CL.
EffNetB0	1040,6 ± 18,5	359,6 ± 13,2	368,5 ± 0,4	369,8 ± 0,6
MobNetV2	989,6 ± 83,2	362,9 ± 0,5	314,3 ± 28,9	335,2 ± 24,5
ResNet101	527,8 ± 53,3	279,7 ± 66,5	262,6 ± 28,1	244,8 ± 30,5
Swin	739,5 ± 86,1	325,5 ± 5,3	339,1 ± 23,1	333,9 ± 36,7
VGG19	871,8 ± 304,5	772,5 ± 0,6	754,5 ± 41,7	691,0 ± 187,3
ViT	744,1 ± 85,5	697,8 ± 37,0	692,3 ± 61,0	650,8 ± 30,7

[a] Valores baseados em 5 execuções independentes.

A adoção de pré-processamentos que reduzem a dimensionalidade dos dados para um único canal (cinza, verde e V+CLAHE) proporcionou reduções expressivas nos tempos de treinamento para todos os modelos, como o esperado. Mas a ViTUNet apresentou redução menos expressiva, de 744,1 segundos (original) para 650,8 segundos com V+CLAHE, sua menor média entre os tipos de pré-processamento. Algo parecido também ocorre com a VGG19, que apresentou uma diferença de 20,7% da média do treinamento com imagens RGB para a V+CLAHE, o tipo de treinamento mais rápido que teve. As suspeitas para essas duas redes terem esses tempos se dá pelas suas arquiteturas não terem as primeiras camadas como tão relevantes para o treinamento em comparação as outras redes. Na VGG19, o custo concentra-se nas diversas convoluções profundas em resoluções espaciais elevadas, e na ViT, nos blocos de autoatenção, cujo custo depende do número de *patches* e da dimensão do *embedding*, e não do número de canais da imagem de entrada.

Analisando individualmente cada arquitetura em relação aos tempos de treino com imagens pré-processadas, observa-se que a ResNet101 apresentou a maior eficiência computacional entre todos os modelos, com o menor tempo médio registrado para a entrada V+CLAHE (244,8 segundos), seguido por verde (262,6 segundos) e cinza (279,7 segundos), com desvios padrão entre 28,1 e 66,5 segundos. O EfficientNetB0, embora tenha apresentado tempos ligeiramente superiores (359,6 a 369,8 segundos entre as entradas monocromáticas), demonstrou alta consistência, com desvios padrão que variaram de 0,42 segundo (verde) a 13,2 segundos (cinza), sendo a arquitetura mais estável entre as avaliadas. Em contraste, o VGG19 exibiu os tempos de 691,0 a 772,5 segundos e a maior variabilidade, com desvios que chegaram a 187,3 segundos para V+CLAHE e 41,7 segundos para verde.

Em termos de variabilidade entre execuções, os menores desvios padrão foram observados para o EfficientNetB0 com

verde (0,42 segundo) e V+CLAHE (0,55 segundo), para a MobileNetV2 com cinza (0,45 segundo) e para o VGG19 com cinza (0,60 segundo), indicando alta consistência nos tempos de treinamento nesses cenários. O VGG19, por outro lado, exibiu a maior instabilidade com imagens originais (desvio de 304,5 segundos) e com V+CLAHE (187,3 segundos), em razão de atingir o limite de épocas estabelecido pelo parâmetro de paciência em algumas execuções. O SwinUNet e o ViTUNet apresentaram desvios moderados (5,3 a 86,1 segundos para o Swin e 30,7 a 85,5 segundos para o ViT), enquanto a ResNet101 com cinza mostrou desvio elevado (66,5 segundos), possivelmente devido a variações no carregamento de dados ou na inicialização dos pesos. De modo geral, os pré-processamentos baseados em canal único não apenas reduzem os tempos médios de treinamento, mas também tendem a produzir tempos mais consistentes entre execuções, o que é um fator relevante para a reprodutibilidade experimental.

B. Análise Comparativa do Dice nos Testes

A métrica Dice foi adotada como principal indicador de desempenho para a segmentação de vasos sanguíneos, por integrar precisão e sensibilidade em uma única medida. Para avaliar a capacidade de generalização dos modelos, além da base de treinamento FIVES, foram realizados testes em duas bases distintas não utilizadas durante o treinamento: a Fundus-AVSeg e a RETA. A Tabela IV apresenta os valores médios e os desvios padrão do Dice para as três bases.

Tabela IV
DICE (%): MÉDIA ± DESVIO NAS BASES FIVES, FUNDUS-AVSEG E RETA.

Modelo	Pré-processamento			
	Orig.	Cinza	Verde	V+CL.
FIVES				
EffNetB0	81,05 ± 0,08	78,97 ± 0,08	81,74 ± 0,23	82,06 ± 0,13
MobNetV2	78,73 ± 0,19	76,51 ± 0,13	79,28 ± 0,17	79,65 ± 0,16
ResNet101	80,57 ± 0,23	78,66 ± 0,24	82,09 ± 0,40	82,81 ± 0,55
Swin	78,30 ± 0,27	74,76 ± 0,30	78,19 ± 0,12	78,95 ± 0,29
VGG19	81,74 ± 3,68	79,45 ± 1,29	81,85 ± 2,01	83,61 ± 0,98
ViT	47,27 ± 0,60	48,36 ± 1,79	56,62 ± 0,98	63,61 ± 0,37
Fundus-AVSeg				
EffNetB0	80,14 ± 0,22	77,85 ± 0,99	79,80 ± 0,41	79,89 ± 0,36
MobNetV2	78,52 ± 0,21	75,35 ± 0,64	77,35 ± 0,51	77,44 ± 0,30
ResNet101	81,24 ± 0,41	78,83 ± 0,88	81,68 ± 0,85	82,43 ± 0,87
Swin	76,72 ± 0,18	72,96 ± 0,62	75,58 ± 0,21	75,87 ± 0,84
VGG19	82,11 ± 1,64	81,15 ± 0,55	82,03 ± 0,80	82,72 ± 0,49
ViT	48,00 ± 0,65	47,58 ± 2,68	54,67 ± 0,72	62,06 ± 0,46
RETA				
EffNetB0	76,09 ± 0,14	74,45 ± 0,23	76,02 ± 0,28	76,41 ± 0,15
MobNetV2	73,93 ± 0,17	72,06 ± 0,17	74,06 ± 0,21	74,07 ± 0,45
ResNet101	76,47 ± 0,22	75,13 ± 0,38	77,57 ± 0,42	78,25 ± 0,25
Swin	73,62 ± 0,25	70,29 ± 0,38	73,02 ± 0,16	73,56 ± 0,26
VGG19	77,88 ± 2,85	77,21 ± 0,40	77,93 ± 1,16	79,30 ± 0,72
ViT	42,45 ± 0,69	40,74 ± 2,28	48,96 ± 1,11	55,60 ± 0,80

[a]Valores baseados em 5 execuções independentes.

Observa-se que o pré-processamento V+CLAHE proporcionou os maiores valores médios de Dice nos testes realizados. Nesse pré-processamento, em todas as três bases avaliadas a VGG19 ficou em primeiro lugar, sendo seguida da ResNet101.

Considerando simultaneamente o desempenho e a estabilidade, a ResNet101 com V+CLAHE consolidou-se como a melhor combinação geral: obteve Dice médio de 82,81% na FIVES, 82,43% na Fundus-AVSeg e 78,25% na RETA, com desvios padrão consistentemente baixos (0,55%, 0,87% e 0,25%, respectivamente). A EfficientNetB0 apresentou comportamento semelhante, com Dice de 82,06%, 79,89% e 76,41% nas três bases e desvios padrão ainda menores (0,13%, 0,36% e 0,15%), configurando a arquitetura mais estável entre todas as avaliadas.

É importante destacar o comportamento da VGG19: embora tenha obtido os maiores valores médios de Dice em todas as bases, apresentou variabilidade substancialmente superior, especialmente nas condições com imagens originais (desvios de 3,68%, 1,64% e 2,85% na FIVES, Fundus-AVSeg e RETA, respectivamente) e com canal verde sem CLAHE (2,01%, 0,80% e 1,16%).

Em contrapartida, o ViTUNet apresentou o pior desempenho em todas as bases, mesmo com V+CLAHE (63,61% na FIVES, 62,06% na Fundus-AVSeg e 55,60% na RETA), reforçando que arquiteturas baseadas exclusivamente em mecanismos de autoatenção não se mostraram adequadas para esta tarefa com o volume de dados disponível.

De modo geral, os resultados evidenciam que o pré-processamento V+CLAHE é a estratégia que apresenta uma melhoria nos resultados e por um menor custo computacional se comparado com as imagens originais.

C. Análise Comparativa do IoU nos Testes

O índice de Intersecção sobre União (IoU) foi adotado como métrica complementar de análise ao Dice, por penalizar de forma mais acentuada discrepâncias entre a segmentação predita e a máscara de referência, sendo especialmente sensível a omissões de vasos de pequeno calibre. A Tabela V apresenta os valores médios e os desvios padrão do IoU para as três bases.

Observa-se que o comportamento do IoU acompanhou de perto o do Dice, com o pré-processamento V+CLAHE proporcionando os maiores valores médios na maioria das arquiteturas e bases. Na base FIVES, a VGG19 obteve o maior IoU médio (72,84%), seguida pela ResNet101 (71,50%) e pela EfficientNetB0 (70,25%). Na base Fundus-AVSeg, novamente a VGG19 (70,62%) e a ResNet101 (70,20%) lideraram, enquanto na base RETA a VGG19 (65,91%) e a ResNet101 (64,48%) mantiveram as duas primeiras posições.

A VGG19, embora tenha obtido os maiores valores médios de IoU, apresentou variabilidade substancialmente superior, especialmente com imagens originais (desvios de 4,98%, 2,30% e 3,69% nas bases FIVES, Fundus-AVSeg e RETA, respectivamente) e com canal verde sem CLAHE (2,70% e 1,54% na FIVES e RETA). Esse padrão indica que, mesmo com um bom desempenho médio, a VGG19 pode não ser tão estável comparada com outras redes.

O pré-processamento V+CLAHE é a estratégia mais robusta nos testes no que se trata da análise do IoU.

D. Análise Comparativa da Sensibilidade nos Testes

A sensibilidade mede a capacidade do modelo em identificar corretamente os pixels pertencentes aos vasos sanguíneos. Essa

Tabela V
IOU (%): MÉDIA ± DESVIO NAS BASES FIVES, FUNDUS-AVSEG E RETA.

Modelo	Pré-processamento			
	Orig.	Cinza	Verde	V+CL.
FIVES				
EffNetB0	68,75 ± 0,11	65,95 ± 0,21	69,79 ± 0,34	70,25 ± 0,17
MobNetV2	65,53 ± 0,25	62,61 ± 0,15	66,32 ± 0,23	66,86 ± 0,22
ResNet101	68,24 ± 0,29	65,70 ± 0,28	70,45 ± 0,55	71,50 ± 0,78
Swin	65,00 ± 0,34	60,37 ± 0,37	64,89 ± 0,10	65,95 ± 0,36
VGG19	70,14 ± 4,98	66,99 ± 1,67	70,35 ± 2,70	72,84 ± 1,41
ViT	31,50 ± 0,55	32,45 ± 1,55	40,10 ± 0,95	47,27 ± 0,38
Fundus-AVSeg				
EffNetB0	66,94 ± 0,31	63,83 ± 1,32	66,49 ± 0,56	66,59 ± 0,50
MobNetV2	64,72 ± 0,29	60,55 ± 0,82	63,17 ± 0,67	63,29 ± 0,40
ResNet101	68,49 ± 0,59	65,15 ± 1,19	69,11 ± 1,22	70,20 ± 1,25
Swin	62,34 ± 0,23	57,55 ± 0,76	60,85 ± 0,26	61,22 ± 1,10
VGG19	69,77 ± 2,30	68,37 ± 0,78	69,63 ± 1,12	70,62 ± 0,72
ViT	31,68 ± 0,57	31,44 ± 2,26	37,84 ± 0,68	45,18 ± 0,49
RETA				
EffNetB0	61,58 ± 0,18	59,52 ± 0,30	61,50 ± 0,37	62,00 ± 0,21
MobNetV2	58,80 ± 0,21	56,51 ± 0,21	58,97 ± 0,27	58,98 ± 0,56
ResNet101	62,12 ± 0,29	60,43 ± 0,47	63,59 ± 0,55	64,48 ± 0,34
Swin	58,37 ± 0,31	54,33 ± 0,45	57,64 ± 0,20	58,31 ± 0,33
VGG19	64,03 ± 3,69	63,16 ± 0,53	64,07 ± 1,54	65,91 ± 0,98
ViT	27,04 ± 0,56	25,73 ± 1,78	32,58 ± 0,99	38,65 ± 0,77

[a]Valores baseados em 5 execuções independentes.

métrica é particularmente relevante na segmentação vascular, pois a não detecção de vasos pode comprometer análises clínicas subsequentes. A Tabela VI apresenta os valores médios e os desvios padrão da sensibilidade para as três bases.

Tabela VI
SENSIBILIDADE (%): MÉDIA ± DESVIO NAS BASES FIVES, FUNDUS-AVSEG E RETA.

Modelo	Pré-processamento			
	Orig.	Cinza	Verde	V+CL.
FIVES				
EffNetB0	78,79 ± 0,64	77,07 ± 0,67	80,10 ± 0,26	80,53 ± 0,68
MobNetV2	77,16 ± 1,02	73,86 ± 1,07	77,57 ± 0,93	77,34 ± 0,44
ResNet101	78,16 ± 1,17	74,62 ± 1,77	80,49 ± 1,98	80,78 ± 1,34
Swin	76,41 ± 0,65	72,21 ± 0,85	77,12 ± 0,73	77,25 ± 0,64
VGG19	78,60 ± 4,18	76,21 ± 1,76	79,14 ± 2,36	80,78 ± 1,55
ViT	43,64 ± 1,22	43,77 ± 1,94	52,36 ± 0,59	59,83 ± 1,20
Fundus-AVSeg				
EffNetB0	72,90 ± 0,53	69,28 ± 2,30	72,20 ± 1,47	71,79 ± 0,98
MobNetV2	71,96 ± 0,76	65,75 ± 1,35	68,73 ± 1,21	68,75 ± 0,90
ResNet101	73,68 ± 1,15	68,83 ± 2,06	73,49 ± 2,15	74,21 ± 1,77
Swin	70,12 ± 0,41	64,27 ± 1,22	67,87 ± 0,59	67,84 ± 1,69
VGG19	74,63 ± 0,68	72,94 ± 1,56	74,08 ± 0,91	74,13 ± 1,23
ViT	42,37 ± 1,58	39,23 ± 2,95	46,40 ± 0,75	54,48 ± 1,45
RETA				
EffNetB0	72,01 ± 0,54	69,80 ± 1,22	72,69 ± 0,59	72,63 ± 0,67
MobNetV2	71,07 ± 0,84	66,84 ± 1,01	70,64 ± 0,84	70,64 ± 0,59
ResNet101	72,51 ± 1,03	68,68 ± 1,83	73,82 ± 1,60	74,27 ± 1,14
Swin	70,98 ± 0,37	65,46 ± 0,78	70,19 ± 0,73	70,46 ± 0,78
VGG19	73,73 ± 2,29	72,84 ± 0,82	74,81 ± 1,41	75,38 ± 0,88
ViT	39,74 ± 1,77	35,75 ± 2,09	43,34 ± 1,51	51,78 ± 1,44

[a]Valores baseados em 5 execuções independentes.

Observa-se que, de modo geral, a sensibilidade apresentou valores mais elevados do que o Dice e o IoU, o que é esperado em problemas de segmentação vascular, nos quais o desequilíbrio entre vasos (classe minoritária) e fundo (classe majoritária) tende a favorecer a detecção de vasos em detrimento da precisão da sua delimitação. Na base FIVES, a ResNet101 e a VGG19 obtiveram as maiores sensibilidades médias com V+CLAHE (80,78%), seguidas pela EfficientNetB0 (80,53%). Na base Fundus-AVSeg, a VGG19 (74,13%) e a ResNet101 (74,21%) mantiveram-se entre as melhores, enquanto na base RETA a VGG19 (75,38%) liderou, seguida pela ResNet101 (74,27%).

Considerando simultaneamente o desempenho e a estabilidade, a ResNet101 com V+CLAHE destacou-se novamente como a melhor combinação geral, com sensibilidades médias de 80,78% na FIVES, 74,21% na Fundus-AVSeg e 74,27% na RETA, mantendo desvios padrão moderados (1,34%, 1,77% e 1,14%, respectivamente). A EfficientNetB0 apresentou comportamento semelhante, com sensibilidades de 80,53%, 71,79% e 72,63%, e desvios padrão ainda menores (0,68%, 0,98% e 0,67%), configurando a arquitetura mais estável entre todas as avaliadas.

A VGG19, embora tenha obtido os maiores valores médios de sensibilidade na base RETA (75,38%) e valores próximos aos melhores nas outras duas bases, apresentou novamente variabilidade superior, especialmente com imagens originais (desvios de 4,18%, 0,68% e 2,29% nas bases FIVES, Fundus-AVSeg e RETA, respectivamente) e com canal verde sem CLAHE (2,36% na FIVES e 1,41% na RETA). Esse padrão reforça a observação prévia de que o VGG19, apesar do bom desempenho médio, carece da estabilidade necessária em cenários que exijam reprodutibilidade rigorosa.

O ViTUNet apresentou a pior sensibilidade em todas as bases, mesmo com V+CLAHE (59,83% na FIVES, 54,48% na Fundus-AVSeg e 51,78% na RETA), confirmando sua inadequação para a tarefa com o volume de dados disponível. O SwinUNet obteve desempenho intermediário, sempre abaixo das CNNs, com sensibilidade máxima de 77,25% na FIVES, 67,84% na Fundus-AVSeg e 70,46% na RETA.

De modo geral, os resultados de sensibilidade reforçam as conclusões obtidas com o Dice e o IoU: o pré-processamento V+CLAHE é a estratégia mais robusta; as arquiteturas convolucionais ResNet101 e EfficientNetB0 oferecem o melhor equilíbrio entre acurácia, estabilidade e custo computacional; e os modelos mantêm desempenho consistente ao migrar da base de treinamento (FIVES) para bases externas (Fundus-AVSeg e RETA). Vale notar que, embora os valores absolutos de sensibilidade sejam mais altos que os de Dice e IoU, essa métrica deve ser interpretada em conjunto com as demais, uma vez que um modelo excessivamente sensível pode classificar regiões de fundo como vasos, elevando o número de falsos positivos.

IV. CONCLUSÃO

Este trabalho investigou o impacto de diferentes estratégias de pré-processamento e arquiteturas de redes neurais profundas na segmentação de vasos sanguíneos em imagens de fundoscopia. Foram avaliadas seis arquiteturas (EfficientNetB0, MobileNetV2, ResNet101, VGG19, SwinUNet e ViTUNet) combinadas com quatro condições de entrada: imagem original em cores, tons de cinza, canal verde e canal verde

com aplicação de CLAHE. Os resultados demonstraram que o pré-processamento V+CLAHE proporcionou os melhores desempenhos para a maioria dos modelos na base FIVES, com destaque para o VGG19 (Dice de 83,61%). Entretanto, ao considerar conjuntamente o desempenho, a estabilidade entre execuções e o custo computacional, a ResNet101 configurou-se como a melhor arquitetura geral, alcançando Dice médio de 82,81% na FIVES, 82,43% na Fundus-AVSeg e 78,25% na RETA, com desvios padrão consistentemente baixos. A EfficientNetB0 apresentou-se com o desempenho ligeiramente inferior, porém com estabilidade superior, enquanto a VGG19, apesar das maiores médias, exibiu variabilidade excessiva em algumas configurações.

Em termos de eficiência computacional, os pré-processamentos que reduzem a dimensionalidade dos dados (cinza, verde e V+CLAHE) proporcionaram reduções drásticas nos tempos de treinamento, com diminuições de até 65,4% para o EfficientNetB0 e 68,2% para a MobileNetV2 em relação às imagens originais, o que já era esperado por diminuir em dois terços a dimensionalidade das imagens de entrada. O tempo médio de treinamento das CNNs com pré-processamento monocromático variou entre aproximadamente 245 e 370 segundos, enquanto os Transformers exigiram de 326 a 698 segundos, refletindo a maior complexidade computacional dos mecanismos de autoatenção. A variabilidade entre execuções foi baixa para EfficientNetB0, MobileNetV2 e ResNet101, com desvios típicos inferiores a 1,0% para o Dice, enquanto o VGG19 apresentou desvios excessivos, especialmente com imagens originais.

A avaliação de generalização, conduzida em duas bases externas não utilizadas no treinamento, a Fundus-AVSeg, proveniente do estudo *A Fundus Image Dataset for AI-based Artery-Vein Vessel Segmentation*, e a RETA, oriunda do benchmark *The RETA Benchmark for Retinal Vascular Tree Analysis*, evidenciou que a VGG19 teve o maior desempenho médio mas com um desvio padrão elevado, as arquiteturas convolucionais ResNet101 e EfficientNetB0 mantiveram desempenho consistente ao migrar para domínios distintos, confirmando a robustez das melhores configurações. O ViTUNet, por outro lado, apresentou o pior desempenho em todas as bases (Dice de 63,61% na FIVES, 62,06% na Fundus-AVSeg e 55,60% na RETA), sugerindo que arquiteturas baseadas exclusivamente em mecanismos de autoatenção não se mostraram adequadas para esta tarefa com o volume de dados disponível. O SwinUNet obteve desempenho intermediário, sempre abaixo das CNNs.

Como trabalhos futuros, pretendemos investigar técnicas de *data augmentation* mais sofisticadas e a aplicação de *transfer learning* entre diferentes bases de fundoscopia, com o objetivo de melhorar o desempenho em bases externas. Além disso, realizar a otimização sistemática de hiperparâmetros, visando maximizar o desempenho das melhores configurações identificadas neste estudo.

REFERÊNCIAS

[1] D. Marín, A. Aquino, M. E. Gegúndez-Arias, and J. M. Bravo, "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *IEEE Trans. Med. Imag.*, vol. 30, no. 1, pp. 146–158, Jan. 2011.

[2] A. M. Mendonça and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1200–1213, Sep. 2006.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] S. R. B. Souza, "Algoritmo para segmentação de drusas em imagens de retinografia através de técnicas de processamento digital de imagens," M.S. thesis, Univ. Federal do Rio Grande do Norte, Natal, Brazil, 2025.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[10] K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, and J. Ye, "FIVES: A fundus image dataset for artificial intelligence based vessel segmentation," *Sci. Data*, vol. 9, p. 475, Aug. 2022, doi: 10.1038/s41597-022-01564-3.

[11] Deng, Z., Gao, W., Gong, Z. et al. A Fundus Image Dataset for AI-based Artery-Vein Vessel Segmentation. Sci Data 12, 1298 (2025). https://doi.org/10.1038/s41597-025-05381-2

[12] Lyu, X., Cheng, L. Zhang, S. The RETA Benchmark for Retinal Vascular Tree Analysis. Sci Data 9, 397 (2022). https://doi.org/10.1038/s41597-022-01507-y

[13] J. Fadugba, P. Köhler, L. Koch, P. Manescu, and P. Berens, "Benchmarking retinal blood vessel segmentation models for cross-dataset and cross-disease generalization," *arXiv preprint arXiv:2406.14994*, 2024.

[14] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[16] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[17] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[18] E. Luz, "Blocos de atenção vascular," GitHub repository, 2026. [Online]. Available: https://github.com/Emanueluz/Blocos-de-aten-o-vascular

Application of Convolutional Neural Networks for Brain Tumor Classification in MRI Images

Lucca Limoeiro Tavares Rodrigues
Student
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
luccalimoeiro.ltr@gmail.com

Belmondo Rodrigues Aragão Junior
Professor
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
belmondo@unifor.br

Gustavo Chaves Macêdo
Student
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
gustavochavesmacedo19@gmail.com

Mauricio Oliveira Amorim
Student
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
mauricio.amorim13@gmail.com

Guilherme Sales Fernandes
VORTEX
Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
guilhermesales@unifor.br

Joel Sotero da Cunha Neto
Professor
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
joelsotero@unifor.br

***Abstract*—This study evaluates convolutional neural networks for the multiclass classification of brain magnetic resonance images into glioma, meningioma, pituitary tumor, and no-tumor categories. A balanced public dataset containing 7,200 images was used, with 5,600 images divided into training and validation subsets and 1,600 images from the repository-provided test subset held out from model development. The experimental design included frozen-backbone models, partial fine-tuning, data augmentation, learning-rate scheduling, Test-Time Augmentation (TTA), and a weighted ensemble composed of DenseNet-169, InceptionV3, and EfficientNetV2-S. Several individual configurations achieved Macro F1-scores close to 0.93. Among the models retrained for the ensemble stage, DenseNet-169 obtained an accuracy of 0.9350 and a Macro F1-score of 0.9338, while the weighted ensemble produced numerically higher values of 0.9394 and 0.9382, respectively. This improvement was modest and required 12 forward passes per image due to the combination of three models and four TTA views. The ensemble also increased recall for glioma and meningioma compared with the baseline DenseNet-121, although clinically relevant misclassifications remained. Because patient identifiers are not available in the public repository, patient-level separation cannot be verified, and the reported test results should be interpreted as image-level benchmark performance rather than evidence of patient-independent clinical generalization. These findings support the potential of transfer learning and model combination for brain tumor MRI classification while reinforcing the need for external, patient-level validation before clinical application.**

***Index Terms*—brain tumor classification, magnetic resonance imaging, convolutional neural networks, transfer learning, deep learning, ensemble learning**

I. Introduction

Brain tumors represent one of the most serious conditions affecting the central nervous system and are characterized by the abnormal and uncontrolled proliferation of cells in brain tissue [1]–[3]. In this context, early and accurate diagnosis is essential to expand therapeutic possibilities, guide clinical planning, and contribute to better patient outcomes [4], [5].

Magnetic resonance imaging (MRI) stands out as one of the main imaging modalities for brain tumor assessment due to its high contrast resolution, multiplanar acquisition capability, and absence of ionizing radiation [2], [3], [6]. Despite the relevance of MRI, manual interpretation remains a complex and time-consuming task that depends heavily on the expertise of radiologists and specialists [7]–[9]. In addition, the increasing volume of imaging examinations in clinical practice reinforces the need for methods capable of supporting rapid, standardized, and reliable image analysis [3], [8].

In this scenario, computer-aided diagnostic systems have been investigated as promising tools to assist in the detection and classification of brain tumors [5], [8]. In recent years, artificial intelligence techniques, particularly machine learning and deep learning, have gained prominence in medical image analysis [3], [5], [6]. Among these approaches, convolutional neural networks (CNNs) have become widely used in brain MRI analysis due to their ability to automatically extract hierarchical and representative features directly from images [2], [6], [7], [10].

Several CNN-based architectures have been applied to brain tumor MRI analysis, including traditional convolutional models and pretrained architectures such as AlexNet, VGG, ResNet, Inception, DenseNet, and EfficientNet [1]–[3], [8], [11]. Transfer learning is particularly relevant in medical imaging, where the availability of large labeled datasets remains limited [9], [11], [12]. Although recent studies report high performance on benchmark datasets, challenges remain regarding generalization, variability across equipment and acquisition protocols, interpretability, dataset provenance, and validation in real clinical scenarios [6], [13], [14].

Therefore, this study evaluates CNN-based models for multiclass classification of brain MRI images into glioma, meningioma, pituitary tumor, and no-tumor categories. The experimental analysis investigates pretrained architectures, partial fine-tuning, data augmentation, learning-rate scheduling, TTA, and a weighted ensemble. In addition to aggregate performance, class-specific behavior and clinically relevant error patterns are examined, while limitations related to dataset provenance, patient-level independence, uncertainty, and computational cost are explicitly considered.

II. Methodology

A. Dataset

The experiments were conducted using Version 2 of the *Brain Tumor MRI Dataset*, made publicly available by Masoud Nickparvar through Kaggle [15]. The repository contains 7,200 brain MRI images evenly distributed among four classes: glioma, meningioma, pituitary tumor, and no tumor, with 1,800 images per class.

According to the repository documentation, the collection was assembled from previously available brain MRI datasets, including Figshare, SARTAJ, and Br35H [15]. The repository is organized into predefined *Training* and *Testing* folders. The Training folder contains 5,600 images, corresponding to 1,400 samples per class, whereas the Testing folder contains 1,600 images, with 400 samples per class. Thus, approximately 77.8% of the images were originally assigned to model development and 22.2% to testing. The complete class distribution is presented in Table I.

Table I
Distribution of images by class and repository subset.

Class	Total	Training	Testing
Glioma	1,800	1,400	400
Meningioma	1,800	1,400	400
No Tumor	1,800	1,400	400
Pituitary	1,800	1,400	400
Total	**7,200**	**5,600**	**1,600**

The Version 2 repository documentation reports that duplicate images were removed and that overlap between the provided Training and Testing folders was eliminated during dataset curation [15]. However, patient identifiers are not provided with the public dataset. Consequently, it is not possible to verify whether different images, adjacent slices, or related acquisitions from the same patient occur across the available partitions. Repository-level duplicate control therefore reduces direct image duplication but does not establish patient-level independence or exclude all possible near-duplicate relationships.

The 5,600 images from the original Training folder were further divided into training and validation subsets using an 80:20 stratified image-level split. This procedure resulted in 4,480 images for training and 1,120 images for validation. Since the original dataset is balanced, the training subset contained 1,120 images per class, while the validation subset contained 280 images per class. Because patient identifiers are unavailable, patient-level separation between these subsets cannot be verified.

The 1,600 images from the repository-provided Testing folder were kept separate from gradient-based training, learning-rate adjustment, early stopping, and checkpoint selection. They were evaluated only after each model configuration had been trained. For this reason, this subset is referred to throughout this paper as the *held-out test subset*, rather than as a patient-independent test cohort. Since all four classes contained the same number of samples, no class-balancing, resampling, or class-weighting strategy was required.

B. Image Pre-processing and Data Augmentation

Before training, all images were resized to 224×224 pixels to ensure compatibility with the input dimensions of the evaluated convolutional neural networks. The resized images were subsequently converted into tensors and normalized using the mean and standard deviation values adopted for the ImageNet dataset. ImageNet normalization was applied because all evaluated architectures were initialized with weights pretrained on ImageNet. Maintaining the normalization procedure used during pretraining preserves the input scale expected by the pretrained convolutional layers.

Data augmentation was applied exclusively to the training subset. In the baseline and initial fine-tuning experiments, light augmentation included random horizontal flipping with a probability of 50%. In the subsequent experimental configuration, the augmentation strategy was expanded to introduce greater variability into the training samples and reduce the tendency of the models to memorize image-specific patterns.

The validation and test subsets were processed only through resizing, tensor conversion, and normalization during conventional evaluation. Random transformations were not applied to these subsets, ensuring that validation and conventional test results were obtained from consistent image representations. Additional transformations were applied to test images only in experiments explicitly using TTA.

C. Evaluated Architectures and Fine-Tuning

Four convolutional neural network architecture families were evaluated: DenseNet, MobileNetV3-Large, InceptionV3, and EfficientNetV2-S. All models were initialized with ImageNet-pretrained weights, and their original classification layers were replaced by task-specific classifiers containing four output neurons corresponding to glioma, meningioma, pituitary tumor, and no-tumor classes.

The experiments were organized into incremental configurations to examine the effects of transfer learning, partial fine-tuning, data augmentation, learning-rate scheduling, and TTA. Table II summarizes the main characteristics of each configuration.

In the baseline configuration, the convolutional backbone of each architecture was frozen and only the newly introduced classifier was trained. This configuration assessed whether the

Table II
INCREMENTAL EXPERIMENTAL CONFIGURATIONS ACROSS MODEL VARIATIONS.

Var.	Architectures		Fine-tuning	Data augmentation	Scheduler	TTA
Base	DenseNet-121, Large, EfficientNetV2-S	MobileNetV3- InceptionV3,	Frozen backbone	Light augmentation	No	No
V1	DenseNet-169, Large, EfficientNetV2-S	MobileNetV3- InceptionV3,	Last block unfrozen	Light augmentation	ReduceLROnPlateau	No
V2	DenseNet-169, Large, EfficientNetV2-S	MobileNetV3- InceptionV3,	Last block unfrozen	Advanced augmentation	ReduceLROnPlateau	Yes

general visual features learned from ImageNet were sufficient to distinguish the four MRI classes without modifying the pretrained convolutional filters.

In Variation 1, the final convolutional block of each architecture was unfrozen, allowing partial fine-tuning of higher-level features. Earlier layers remained frozen to preserve general visual representations, while the final layers were adapted to the characteristics of brain MRI images.

Variation 2 maintained partial fine-tuning and incorporated an expanded data-augmentation strategy, the ReduceLROnPlateau learning-rate scheduler, adjusted early-stopping parameters, and TTA. This configuration investigated whether greater variability during training and prediction aggregation across transformed versions of the same image could improve classification performance.

For the DenseNet family, DenseNet-121 was adopted in the baseline experiment, whereas DenseNet-169 was used in subsequent configurations. The deeper DenseNet-169 architecture was selected to investigate whether its greater representational capacity could provide more discriminative features for brain tumor classification.

D. Training Procedure

All models were trained using the Adam optimizer and cross-entropy loss. In the baseline configuration, a learning rate of 1×10^{-3} was applied to the classifier while the convolutional backbone remained frozen.

In the partial fine-tuning configurations, differential learning rates were adopted. A learning rate of 1×10^{-5} was used for the unfrozen backbone layers, while 1×10^{-3} was maintained for the classifier. This strategy allowed the newly initialized classifier to learn more rapidly while applying smaller updates to pretrained convolutional features.

The ReduceLROnPlateau scheduler was used to adjust the learning rate according to validation loss. When validation loss stopped improving for the defined patience period, the scheduler reduced the learning rate, allowing optimization to continue with smaller parameter updates. Early stopping was also based exclusively on validation loss. A patience of five epochs was used in the baseline experiments, while subsequent configurations allowed additional epochs after learning-rate reductions to avoid prematurely interrupting fine-tuning.

At the end of each epoch, model performance was evaluated on the validation subset, and the checkpoint associated with the lowest validation loss was retained. Thus, checkpoint selection within each training run was based exclusively on validation performance, without using the held-out test subset for optimization, learning-rate adjustment, early stopping, or checkpoint selection.

After training, retained checkpoints were evaluated on the same held-out test subset to provide a broad comparison among the investigated architectures and configurations. Because multiple configurations were evaluated on this common test subset, the comparison reported in Table III is treated as an exploratory benchmark rather than as a confirmatory model-selection analysis. Numerical differences among these configurations are therefore interpreted descriptively.

Accuracy, precision, recall, and Macro F1-score were calculated after training. Macro F1-score was adopted as the main comparative metric because it assigns equal importance to all four classes and can reveal class-specific performance differences that may not be evident from accuracy alone.

E. Test-Time Augmentation

Test-Time Augmentation was applied only to the experimental configurations explicitly defined to use this strategy. For each test image, four views were evaluated: the original image, a horizontally flipped version, a vertically flipped version, and a version rotated by 90°. Each view was independently processed by the model, producing a probability distribution over the four classes. The four probability distributions were then averaged, and the class associated with the highest mean probability was selected as the final prediction.

The purpose of this procedure was to experimentally assess whether prediction aggregation across spatially transformed versions of the same image could reduce sensitivity to orientation and image presentation. These transformations should not, however, be interpreted as representations of equally plausible clinical MRI orientations. Horizontal flipping preserves the general spatial organization of the image while reversing laterality, whereas vertical flipping and 90° rotation can produce orientations that are uncommon in routine clinical image presentation.

Therefore, TTA was considered an experimental robustness strategy rather than an assumption of anatomical equivalence

among transformed images. Results obtained with TTA are interpreted accordingly, particularly when assessing whether additional transformed views provide meaningful improvements over conventional inference.

F. Ensemble Strategy

Following the exploratory comparison of the individual experimental configurations, DenseNet-169, InceptionV3, and EfficientNetV2-S were selected as candidate ensemble components. The selection considered both the performance observed during the exploratory benchmark and architectural differences among the models, with the objective of combining potentially complementary feature representations. Because this choice was made after inspection of the broader benchmark results, the ensemble experiment is presented as an exploratory extension of the comparative analysis rather than as an independent confirmatory evaluation.

For the ensemble experiment, the three selected architectures were independently retrained using a specific training configuration. Consequently, the performance values obtained during this stage differ from those reported in the initial comparison of model variations. The metrics presented for the ensemble components therefore correspond to these new training runs and should not be interpreted as repetitions of the values presented in Table III.

During this retraining stage, differential learning rates were used, with a learning rate of 5×10^{-4} for the unfrozen backbone layers and 1×10^{-3} for the classifier. A weight decay of 1×10^{-5} was also applied. The data-augmentation strategy was expanded, and early stopping monitored validation loss with a patience of 15 epochs.

The final ensemble prediction was obtained through a weighted average of the probability distributions generated by the three component models. Importantly, the contribution of each model was determined exclusively from its Macro F1-score on the validation subset obtained during the retraining stage. Test-set performance was not used to calculate or adjust the ensemble weights.

TTA was applied independently to each component model before the weighted combination of probabilities. Thus, each test image was evaluated using four views by each of the three architectures, resulting in 12 forward passes per image before the final ensemble prediction was produced.

This design necessarily increases inference cost relative to using a single model. No hardware-normalized wall-clock inference-time benchmark was performed; therefore, computational efficiency is discussed in terms of model multiplicity and number of forward passes rather than measured latency.

G. Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, Macro F1-score, and confusion matrices. Accuracy represents the proportion of correctly classified images among all samples in the held-out test subset. Precision indicates the proportion of predictions assigned to a given class that were correct, whereas recall represents the proportion of samples belonging to that class that were correctly identified. The F1-score combines precision and recall through their harmonic mean.

Macro F1-score was calculated by independently determining the F1-score for each class and averaging the four values with equal weight. Although the dataset is balanced, this metric was prioritized because global accuracy may conceal differences in performance among glioma, meningioma, pituitary tumor, and no-tumor images.

In addition to aggregate metrics, class-specific precision, recall, F1-score, support, and confusion counts were analyzed for the final ensemble. Confusion matrices were used to identify systematic classification errors and determine which categories were more frequently confused. Particular attention was given to errors involving glioma, meningioma, and predictions of tumor images as no tumor.

The baseline DenseNet-121 confusion matrix is presented in normalized form, whereas the ensemble confusion matrix is presented using absolute counts. Because the held-out test subset contains exactly 400 images from each class, absolute ensemble counts can also be directly interpreted as class-specific proportions.

The reported metrics are point estimates. The present experiments did not include repeated-training uncertainty estimation or bootstrap confidence intervals. Therefore, small numerical differences between models are not interpreted as evidence of statistically significant superiority. This limitation is explicitly considered when discussing the performance difference between DenseNet-169 and the ensemble.

III. Results

A. Overall Performance of the Evaluated Models

The performance of the evaluated architectures across the three experimental configurations is summarized in Table III. Accuracy, precision, recall, and Macro F1-score are reported for the common held-out test subset. Because several architectures and configurations were evaluated on the same test subset, this table is presented as an exploratory benchmark intended to characterize the behavior of the investigated approaches rather than as a confirmatory model-selection analysis.

Table III
HELD-OUT TEST PERFORMANCE OF THE EVALUATED ARCHITECTURES ACROSS THE BASELINE AND INCREMENTAL EXPERIMENTAL CONFIGURATIONS.

Model	Variation	Acc.	Prec.	Rec.	Macro F1
MobileNetV3-Large	Baseline	0.88	0.89	0.88	0.88
InceptionV3	Baseline	0.87	0.87	0.87	0.86
DenseNet-121	Baseline	0.93	0.94	0.93	0.93
MobileNetV3-Large	Variation 1	0.92	0.92	0.92	0.91
InceptionV3	Variation 1	0.93	0.93	0.93	0.92
EfficientNetV2-S	Variation 1	0.93	0.93	0.93	0.93
MobileNetV3-Large	Variation 2	0.87	0.87	0.87	0.87
InceptionV3	Variation 2	0.93	0.94	0.93	0.93
EfficientNetV2-S	Variation 2	0.93	0.93	0.93	0.93
DenseNet-169 + TTA	Variation 2	0.90	0.92	0.90	0.8979

In the baseline configuration, DenseNet-121 achieved the strongest numerical results, with accuracy and Macro F1-score of 0.93. Partial fine-tuning improved MobileNetV3-Large and InceptionV3 in Variation 1, while EfficientNetV2-S also reached a Macro F1-score of 0.93. In Variation 2, InceptionV3 and EfficientNetV2-S maintained Macro F1-scores close to 0.93, whereas MobileNetV3-Large decreased to 0.87 and DenseNet-169 with TTA obtained 0.8979. Overall, several configurations reached similar aggregate performance. Given the exploratory nature of the comparison and the absence of uncertainty intervals, these numerical differences should not be interpreted as evidence of statistical superiority.

B. Learning Curve Analysis

The training behavior of the DenseNet models was analyzed through their training and validation loss curves. Fig. 1 presents the learning curves obtained for the baseline DenseNet-121 model.

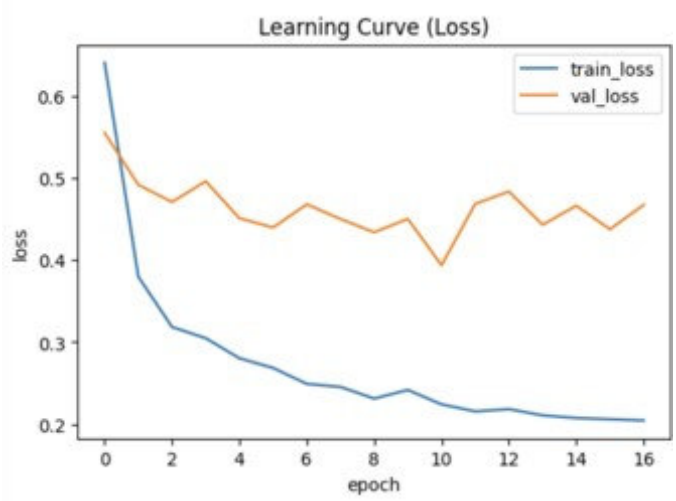


Figure 1. Training and validation loss curves of the baseline DenseNet-121 model.

As shown in Fig. 1, training loss decreased rapidly during the first epochs and continued to decline gradually until the end of training. In contrast, validation loss decreased during the initial epochs but subsequently presented moderate oscillations, remaining approximately between 0.43 and 0.50. The lowest training loss was observed during the final epochs, whereas validation loss reached its lowest values earlier in the training process. A persistent gap between the training and validation curves was observed throughout most of the experiment.

Fig. 2 presents the corresponding training and validation loss curves for DenseNet-169 in Variation 2.

Training loss showed a substantial decrease during approximately the first ten epochs and continued to decline more gradually afterward, remaining below 0.10 during most of the remaining epochs. Validation loss decreased during the initial training period but subsequently remained relatively

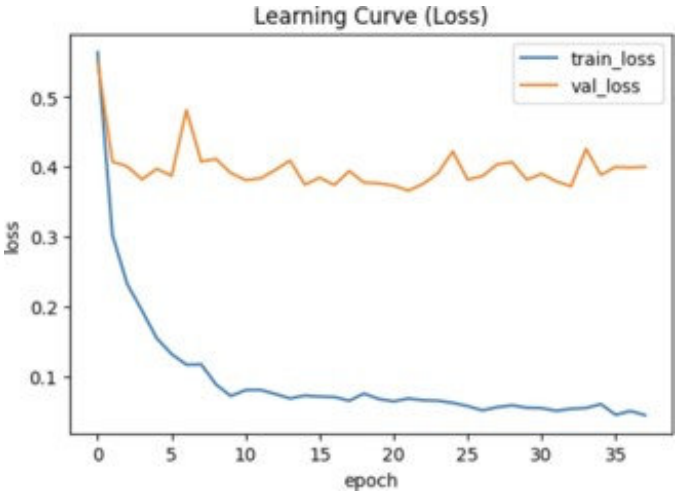


Figure 2. Training and validation loss curves of DenseNet-169 in Variation 2.

stable, with oscillations around 0.36 to 0.42. Although training continued for approximately 36 epochs, validation loss did not show a sustained decrease during the final epochs.

Compared with DenseNet-121, DenseNet-169 reached lower training-loss values and required more epochs before training was interrupted. However, both models presented a persistent difference between training and validation loss, suggesting that improvements in optimization of the training samples were not consistently accompanied by equivalent improvements in validation performance.

C. Class-Specific Performance and Confusion Matrices

Confusion matrices were analyzed to characterize the distribution of correct and incorrect predictions across classes. Fig. 3 presents the normalized confusion matrix obtained by the baseline DenseNet-121 model.

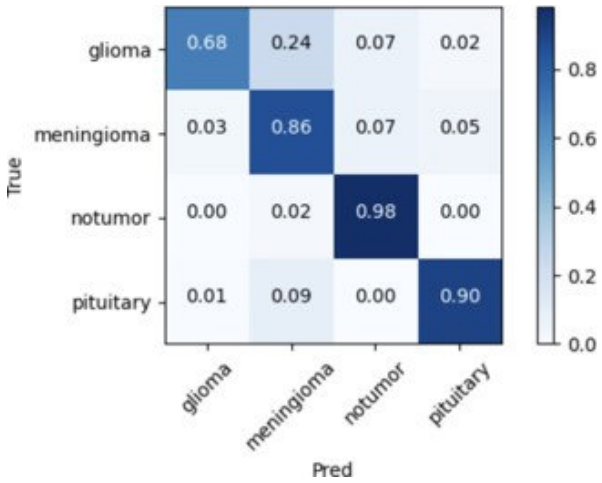


Figure 3. Normalized confusion matrix of the baseline DenseNet-121 model on the held-out test subset.

Glioma presented the lowest recall in this baseline configuration, with 68% of glioma images correctly classified. Among the remaining glioma samples, 24% were classified as meningioma, 7% as no tumor, and approximately 2% as pituitary tumor.

For meningioma, 86% of the images were correctly classified, while approximately 3% were classified as glioma, 7% as no tumor, and 5% as pituitary tumor. The no-tumor class obtained a recall of 98%, with 2% of its samples classified as meningioma. Pituitary tumor achieved a recall of 90%, with most errors involving classification as meningioma.

Fig. 4 presents the confusion matrix obtained by the weighted ensemble using absolute prediction counts.

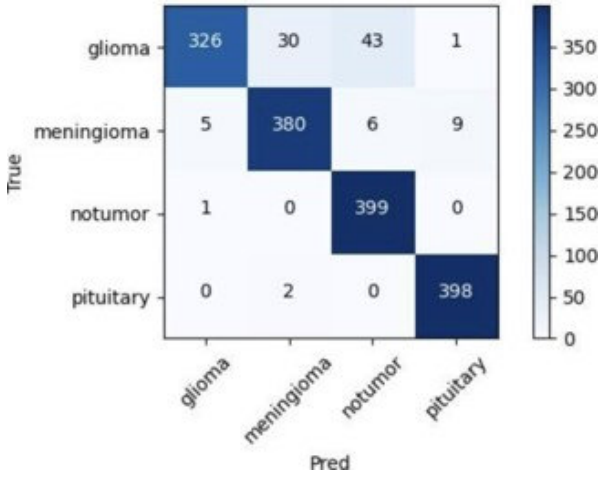


Figure 4. Confusion matrix of the weighted ensemble on the held-out test subset, expressed as absolute prediction counts.

The ensemble correctly classified 326 of 400 glioma images, corresponding to a recall of 81.5%. Among the misclassified glioma images, 30 were predicted as meningioma, 43 as no tumor, and one as pituitary tumor.

For meningioma, 380 of 400 images were correctly classified, corresponding to a recall of 95.0%; five were classified as glioma, six as no tumor, and nine as pituitary tumor. For the no-tumor class, 399 of 400 images were correctly classified, corresponding to a recall of 99.75%, with one image classified as glioma. Finally, 398 of 400 pituitary tumor images were correctly classified, corresponding to a recall of 99.5%, while two were classified as meningioma.

Relative to the baseline DenseNet-121 results, the ensemble therefore presented numerically higher recall for all four classes. The most relevant remaining error pattern was observed for glioma, including 43 glioma images predicted as no tumor.

D. Performance of the Retrained Ensemble Components

DenseNet-169, InceptionV3, and EfficientNetV2-S were independently retrained using the specific configuration adopted for the ensemble experiment. Table IV presents the individual performance of these training runs and the result obtained after combining their probability distributions.

Table IV
HELD-OUT TEST PERFORMANCE OF THE INDEPENDENTLY RETRAINED COMPONENT MODELS AND RESULTING WEIGHTED ENSEMBLE.

Model	Acc.	Prec.	Rec.	Macro F1
DenseNet-169	0.9350	0.94	0.93	0.9338
InceptionV3	0.9306	0.94	0.93	0.9292
EfficientNetV2-S	0.9038	0.89	0.89	0.9023
Ensemble Top-3	0.9394	0.94	0.94	0.9382

Among the retrained component models, DenseNet-169 produced the highest individual point estimates, with an accuracy of 0.9350, precision of approximately 0.94, recall of approximately 0.93, and Macro F1-score of 0.9338. InceptionV3 achieved an accuracy of 0.9306 and a Macro F1-score of 0.9292, whereas EfficientNetV2-S achieved an accuracy of 0.9038 and a Macro F1-score of 0.9023.

The weighted ensemble produced an accuracy of 0.9394 and a Macro F1-score of 0.9382. Relative to DenseNet-169, this corresponds to an absolute numerical increase of approximately 0.0044 in both accuracy and Macro F1-score. Therefore, although the ensemble produced the highest point estimates in this experimental stage, the magnitude of the improvement was small.

Because these results are point estimates obtained without repeated-training variability analysis or bootstrap confidence intervals, the observed difference should not be interpreted as evidence that the ensemble is statistically superior to DenseNet-169. Instead, the results indicate that combining the models produced a modest numerical improvement under the evaluated configuration.

The component-model values in Table IV correspond specifically to the models retrained for the ensemble experiment and therefore differ from the exploratory results reported in Table III.

E. Class-Wise Performance of the Ensemble

Table V presents class-specific precision, recall, F1-score, and support for the final ensemble. The values are consistent with the absolute confusion counts shown in Fig. 4.

Table V
CLASS-SPECIFIC HELD-OUT TEST PERFORMANCE OF THE WEIGHTED ENSEMBLE WITH TEST-TIME AUGMENTATION.

Class	Precision	Recall	F1-score	Support
Glioma	0.9819	0.8150	0.8907	400
Meningioma	0.9223	0.9500	0.9360	400
No Tumor	0.8906	0.9975	0.9410	400
Pituitary	0.9755	0.9950	0.9851	400
Macro avg.	0.9426	0.9394	0.9382	1600
Weighted avg.	0.9426	0.9394	0.9382	1600

Glioma achieved a precision of 0.9819, recall of 0.8150, and F1-score of 0.8907, representing the lowest recall and F1-score among the four classes. Meningioma achieved a precision of 0.9223, recall of 0.9500, and F1-score of 0.9360.

The no-tumor class obtained a precision of 0.8906, recall of 0.9975, and F1-score of 0.9410. Its comparatively lower precision is consistent with the confusion matrix, which shows that some tumor images, particularly glioma images, were predicted as no tumor. Pituitary tumor achieved a precision of 0.9755, recall of 0.9950, and F1-score of 0.9851, representing the highest class-specific F1-score.

Considering all 1,600 test images, the ensemble obtained an accuracy of 0.9394. Macro-average precision, recall, and F1-score were 0.9426, 0.9394, and 0.9382, respectively. These class-specific results demonstrate that similar aggregate performance can coexist with substantial differences in sensitivity across tumor categories, particularly for glioma.

IV. Discussion

The results indicate that transfer learning and partial fine-tuning can provide effective image-level classification performance for the brain MRI dataset investigated in this study. As shown in Table III, several architectures achieved Macro F1-scores close to 0.93, although the effect of increasingly complex training strategies varied among architectures. InceptionV3 improved from a Macro F1-score of 0.86 in the baseline configuration to approximately 0.93 in Variation 2, whereas MobileNetV3-Large improved in Variation 1 but decreased to 0.87 in Variation 2. These findings suggest that additional fine-tuning, augmentation, and inference strategies do not necessarily produce consistent gains across architectures.

The learning curves in Figs. 1 and 2 showed a continuous reduction in training loss accompanied by higher and more oscillatory validation loss. This behavior suggests some degree of overfitting, particularly for DenseNet-169, which reached substantially lower training-loss values without a proportional reduction in validation loss. Validation-based early stopping and checkpoint selection were therefore important for limiting continued optimization after validation performance had stabilized.

Class-specific analysis also revealed important differences that are not apparent from aggregate accuracy alone. Glioma remained the most challenging category. The baseline DenseNet-121 achieved a glioma recall of 68%, whereas the ensemble achieved 81.5%. Meningioma recall changed from 86% to 95%, while no tumor and pituitary tumor reached recalls of 99.75% and 99.5%, respectively, in the ensemble. Nevertheless, 43 glioma images were predicted as no tumor by the ensemble. This error pattern is particularly important in the context of a potential decision-support system because high aggregate accuracy does not eliminate clinically relevant failure cases. The evaluated models should therefore not be interpreted as autonomous diagnostic systems.

Among the models retrained for the ensemble stage, DenseNet-169 achieved an accuracy of 0.9350 and a Macro F1-score of 0.9338, whereas the weighted ensemble achieved 0.9394 and 0.9382, respectively. The resulting absolute gain of approximately 0.0044 in Macro F1-score is modest. Moreover, repeated training runs and bootstrap confidence intervals were not included in the present analysis. Consequently, the observed difference should be interpreted as a numerical improvement in this experiment rather than as evidence of statistically established superiority of the ensemble over DenseNet-169.

The computational implications of this improvement are also relevant. Because the ensemble combines three architectures and evaluates four TTA views with each model, it requires 12 forward passes for each image. In terms of forward-pass count, this corresponds to three times the inference workload of a single model using the same four-view TTA procedure and substantially more computation than conventional single-view inference. Actual wall-clock inference time was not benchmarked, and therefore no quantitative latency claim can be made. The additional models also increase deployment complexity and computational-resource requirements.

These differences have practical implications for potential clinical workflows. A single model such as DenseNet-169 may provide a more favorable compromise between predictive performance and operational simplicity in scenarios requiring interactive or near-real-time decision support. In contrast, the ensemble could be considered for offline or less latency-sensitive processing when the additional computational cost is acceptable. These possibilities represent deployment considerations rather than validated clinical-use scenarios, since inference time on clinical hardware and prospective workflow integration were not evaluated in this study.

The TTA procedure introduces an additional limitation. Horizontal flipping reverses laterality while retaining the overall structure of the image, whereas vertical flipping and 90° rotation may result in orientations that are uncommon in standard clinical image presentation. Thus, TTA results should be interpreted as an experimental assessment of robustness to image transformations rather than as evidence that all transformations represent anatomically equivalent clinical acquisitions.

Dataset provenance also constrains interpretation of the reported results. The public repository is assembled from multiple source collections, including Figshare, SARTAJ, and Br35H [15]. The repository documentation reports removal of duplicate images and elimination of direct overlap between its provided Training and Testing folders. However, patient identifiers are not provided, and patient-level separation cannot therefore be verified. In particular, the available metadata do not allow confirmation that images or adjacent slices originating from the same patient are confined to a single partition. The additional training/validation split used in this study was also performed at the image level. Consequently, the held-out test results should not be interpreted as demonstrating generalization to an independent patient cohort.

A further consideration is that multiple architectures and configurations were evaluated on the same held-out test subset in Table III, and this exploratory comparison informed the subsequent ensemble investigation. Although the test subset was not used for gradient optimization, learning-rate scheduling, early stopping, checkpoint selection, or determination of

ensemble weights, repeated examination of a common test set can still influence higher-level experimental interpretation. Table III and the ensemble extension should therefore be regarded as exploratory benchmark analyses rather than fully independent confirmatory comparisons.

Overall, these limitations emphasize the need for cautious interpretation of the reported performance. Future studies should include external and multicenter datasets with explicit patient identifiers, patient-level partitioning, repeated training runs, uncertainty estimation with confidence intervals, clinically appropriate augmentation strategies, probability calibration, interpretability methods, and direct measurement of computational cost and inference latency in the intended deployment environment.

V. Conclusion

This study evaluated convolutional neural networks for image-level classification of brain MRI images into glioma, meningioma, pituitary tumor, and no-tumor categories. The experiments included frozen-backbone models, partial fine-tuning, data augmentation, Test-Time Augmentation, and a weighted ensemble. Across the exploratory benchmark, several individual configurations achieved Macro F1-scores close to 0.93, although the effect of the investigated strategies varied among architectures.

Glioma remained the most challenging class. The ensemble achieved a glioma recall of 81.5% and a meningioma recall of 95%, while maintaining recall above 99% for the no-tumor and pituitary tumor classes. However, clinically relevant errors remained, including glioma images classified as no tumor.

Among the models retrained for the ensemble experiment, DenseNet-169 achieved an accuracy of 0.9350 and a Macro F1-score of 0.9338. The weighted ensemble produced numerically higher values of 0.9394 and 0.9382, respectively. This gain was modest and was accompanied by substantially greater computational requirements, with 12 forward passes required for each image. In the absence of repeated-run or bootstrap uncertainty estimates, this difference should not be interpreted as evidence of statistically established superiority.

The repository-provided test subset was kept separate from model optimization and checkpoint selection. However, because patient identifiers are not available, patient-level independence between dataset partitions cannot be confirmed. The reported results should therefore be interpreted as performance on a held-out image-level benchmark rather than as evidence of generalization to an independent clinical population.

Overall, the findings support the potential of transfer learning and model combination for brain tumor MRI classification while also demonstrating important limitations related to class-specific errors, computational cost, dataset provenance, and evaluation uncertainty. External multicenter validation with patient-level separation, together with uncertainty analysis, interpretability, probability calibration, and direct assessment of inference efficiency, remains necessary before considering clinical deployment.

References

[1] S. M. Qader, B. A. Hassan, and T. A. Rashid, "An improved deep convolutional neural network by using hybrid optimization algorithms to detect and classify brain tumor using augmented MRI images," *Multimedia Tools and Applications*, vol. 81, no. 30, pp. 44 059–44 086, 2022.

[2] E. S. Biratu, F. Schwenker, Y. M. Ayano, and T. G. Debelee, "A survey of brain tumor segmentation and classification algorithms," *Journal of Imaging*, vol. 7, no. 9, p. 179, 2021.

[3] A. M. Al-Zoghby, E. M. K. Al-Awadly, A. Moawad, N. Yehia, and A. I. Ebada, "Dual deep CNN for tumor brain classification," *Diagnostics*, vol. 13, no. 12, p. 2050, 2023.

[4] S. Srinivasan, S. K. Mathivanan, V. Muthukumaran, J. C. Babu, and L. Vilceková, "Grade classification of tumors from brain magnetic resonance images using a deep learning technique," *Diagnostics*, vol. 13, no. 6, p. 1153, 2023.

[5] K. Neamah, F. Mohamed, M. M. Adnan, T. Saba, S. A. Bahaj, and K. A. Kadhim, "Brain tumor classification and detection based DL models: A systematic review," *IEEE Access*, vol. 12, pp. 2517–2542, 2024.

[6] R. Missaoui, W. Hechkel, W. Saadaoui, A. Helali, and M. Leo, "Advanced deep learning and machine learning techniques for MRI brain tumor analysis: A review," *Sensors*, vol. 25, no. 9, p. 2746, 2025.

[7] A. S. Farhan, M. Khalid, and U. Manzoor, "PRCnet: An efficient model for automatic detection of brain tumor in MRI images," *bioRxiv*, p. 2023.09.28.560042, 2023, preprint.

[8] M. K. Abd-Ellah, A. I. Awad, A. A. M. Khalaf, and H. F. A. Hamed, "Two-phase multi-model automatic brain tumour diagnosis system from magnetic resonance images using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 97, 2018.

[9] N. A. Samee, N. F. Mahmoud, G. Atteia, H. A. Abdallah, M. Alabdulhafith, M. S. A. M. Al-Gaashani, S. Ahmad, and M. S. A. Muthanna, "Classification framework for medical diagnosis of brain tumor with an effective hybrid transfer learning model," *Diagnostics*, vol. 12, no. 10, p. 2541, 2022.

[10] D. K. Gokapay and S. N. Mohanty, "Enhanced MRI-based brain tumor segmentation and feature extraction using Berkeley wavelet transform and ETCCNN," *Digital Health*, vol. 10, p. 20552076241305282, 2024.

[11] B. B. Vimala, S. Srinivasan, S. K. Mathivanan, N. Mahalakshmi, P. Jayagopal, and G. T. Dalu, "Detection and classification of brain tumor using hybrid deep learning models," *Scientific Reports*, vol. 13, no. 1, p. 23029, 2023.

[12] K. K. Kumar, P. M. Dinesh, P. Rayavel, L. Vijayaraja, R. Dhanasekar, R. Kesavan, K. Raju, A. A. Khan, C. Wechtaisong, M. A. Haq *et al.*, "Brain tumor identification using data augmentation and transfer learning approach," *Computer Systems Science and Engineering*, vol. 46, no. 2, pp. 1845–1861, 2023.

[13] C. M. L. Zegers, J. Posch, A. Traverso, D. Eekers, A. A. Postma, W. Backes, A. Dekker, and W. van Elmpt, "Current applications of deep-learning in neuro-oncological MRI," *Physica Medica*, vol. 83, pp. 161–173, 2021.

[14] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, no. 14, p. 3608, 2023.

[15] M. Nickparvar, "Brain tumor MRI dataset," Kaggle, 2021, version 2. Accessed: Sep. 2026. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Application of Radiomics and Deep Learning for the Classification and Grading of Gliomas in Magnetic Resonance Imaging: A Literature Review

1st Kaíny Medeiros Maciel
Graduate Program in Electrical Engineering
Federal University of Amazonas
Manaus, Brazil
kainy.maciel@ufam.edu.br

2nd Marly Guimarães F. Costa
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
mcosta@ufam.edu.br

3rd Cícero Ferreira F. C. Filho
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
ccosta@ufam.edu.br

***Abstract*—This literature review examines fifteen studies published between 2018 and 2026 on the application of radiomics and deep learning to classify and grade gliomas in magnetic resonance imaging (MRI) scans. The articles cover classical radiomics with machine learning classifiers, deep convolutional neural networks, hybrid models, graph-based architectures, and radiogenomics frameworks. Reported performance metrics vary widely across studies, reflecting differences in classification tasks, validation protocols, and dataset characteristics rather than a consistent superiority of any single methodological family; the main challenges are data scarcity, the diversity of validation protocols, and the lack of clinical interpretability of the models.**

***Index Terms*—Glioma, Magnetic Resonance Imaging, Radiomics, Deep Learning, Neural Networks, Tumor Grading.**

I. Introduction

Gliomas are primary brain tumors — the most common type in adults — that originate from glial cells and account for approximately 74.6% of all malignant brain tumors [1]. The diagnosis of gliomas is one of the most challenging problems in neuro-oncology today, especially considering the drastic difference in prognosis between grades: patients with low-grade gliomas (LGG, WHO grades 2 and 3) have a median survival of seven years [2], while patients with glioblastoma (WHO grade 4, per the 2021 classification) rarely survive more than twelve months, with only 3 to 5% surviving beyond five years [3].

Historically, the grading of gliomas depended only on the histopathological analysis of samples obtained by biopsy or surgical resection, an invasive procedure that is subject to sampling errors and unable to capture spatial differences within the tumor [4]. The World Health Organization (WHO) classification has evolved from a purely morphological system to an integrated model that combines histological and molecular criteria, with the 2016 and 2021 updates incorporating markers such as the IDH (isocitrate dehydrogenase) mutation, 1p/19q codeletion, and MGMT promoter methylation as key determinants of diagnosis and treatment [4].

In this context, magnetic resonance imaging (MRI) plays a central role in the preoperative evaluation of gliomas. Different sequences provide complementary insights into tumor biology, from contrast-enhanced T1-weighted and T2/FLAIR imaging to diffusion and perfusion-based sequences such as ADC and DSC-MRI, the latter providing estimates of relative cerebral blood volume (rCBV) directly associated with IDH status [5].

Radiomics is a computational approach that systematically extracts quantitative information from medical images, converting pixels into analyzable data that transcends human perception [1]. At the same time, deep learning — particularly convolutional neural networks (CNNs) — has emerged as a powerful alternative, capable of learning hierarchical representations directly from the data without the need for manual feature extraction [6].

II. Methodology

This review examines fifteen studies on radiomics and deep learning for glioma classification and grading in MRI, guided by the research question: *How can radiomics and deep learning models be integrated for the classification and grading of gliomas using MRI images?*. Studies were identified through manual searches across three databases (IEEE Xplore, ScienceDirect, PubMed) and AI-assisted discovery tools (Elicit, Consensus), following the PCC (Population, Concept, Context) framework, then pooled, deduplicated, and screened for eligibility.

A. Search Strategy

The same base search string was adapted to the specific syntax of each database, distributing terms across the Title and Abstract fields. Searches were conducted on June 18, 19, and 20, 2026, respectively, in IEEE Xplore, ScienceDirect, and PubMed.

In IEEE Xplore, the search was restricted to Title/Abstract fields and filtered by publication year (2018–2026, chosen to capture seminal works on radiomics and deep learning in medical imaging) and document type (*Journal Articles* only), based on the rationale that journal publications undergo more rigorous peer review than conference proceedings.

IEEE Xplore

```
("Document Title":"Brain Tumor" OR
"Document Title":"Glioma" OR "Document
Title":"Glioblastoma")
AND
("Document Title":"MRI" OR "Document
Title":"Magnetic Resonance Imaging")
AND
("Abstract":"Radiomics" OR "Abstract":"Machine
Learning" OR "Abstract":"Deep Learning" OR
"Abstract":"Neural Network")
AND
("Abstract":"Tumor Classification" OR
"Abstract":"Tumor Grading")
```

In ScienceDirect, the same year filter was applied together with a *Research Articles* filter, with results ordered by relevance (Best Match); manual screening was limited to the top 30 records to ensure focused assessment.

ScienceDirect

```
Title: ("Brain Tumor" OR "Glioma" OR
"Glioblastoma") AND ("MRI" OR "Magnetic
Resonance Imaging")
AND
Abstract: ("Radiomics" OR "Machine Learning"
OR "Deep Learning" OR "Neural Network") AND
("Tumor Classification" OR "Tumor Grading")
```

In PubMed, the Population/Concept terms were restricted to the `[Title]` field and the technique/objective terms to `[Title/Abstract]`, with an additional *Humans* species filter, results ordered by Best Match, and screening limited to the top 30 records.

PubMed

```
("Brain Tumor"[Title] OR "Glioma"[Title] OR
"Glioblastoma"[Title])
AND
("MRI"[Title] OR "Magnetic Resonance
Imaging"[Title])
AND
("Radiomics"[Title/Abstract] OR "Machine
Learning"[Title/Abstract] OR "Deep
Learning"[Title/Abstract] OR "Neural
Network"[Title/Abstract])
AND
("Tumor Classification"[Title/Abstract] OR
"Tumor Grading"[Title/Abstract])
```

To expand the scope and capture studies that may not appear prominently in database search rankings, two AI-assisted discovery tools were queried on June 21, 2026: Elicit (semantic search across PubMed and arXiv) and Consensus (consensus search across multiple databases including PubMed, Scopus, and bioRxiv). Both tools were prompted with the following query:

```
"Execute a broad literature search for a
scoping review using the following search
string: TITLE(Brain Tumor OR Glioma OR
Glioblastoma) AND TITLE(MRI OR Magnetic
Resonance Imaging) AND ABSTRACT(Radiomics
OR Machine Learning OR Deep Learning OR Neural
Network) AND ABSTRACT(Tumor Classification OR
Tumor Grading). Return only the identified
bibliographic records, without filtering
by methodological quality and without
synthesizing the findings. For each record,
report: Title; Authors; Year; Source; DOI;
Link. List all studies found, ordered by
publication year (most recent to oldest)."
```

B. Eligibility Criteria and Study Selection

Studies were included in the review if they satisfied all of the following criteria:

1) **Population and disease:** focus on glioma grading and/or classification (LGG, HGG, glioblastoma, or WHO-grade categories);
2) **Imaging modality:** MRI as the primary modality, whether conventional (T1, T1CE, T2, FLAIR) or quantitative (ADC, DSC-MRI, perfusion) sequences;
3) **Analytical methods:** use of radiomics and/or deep learning (CNNs, graph neural networks, or other architectures);
4) **Methodological contribution:** an original approach (preprocessing, feature selection, architecture, or fusion strategy), excluding reviews, editorials, and descriptive case reports;
5) **Availability:** full-text access at the time of the search (June 18–21, 2026). Full-text availability was required due to institutional access constraints at the time of the search; this pragmatic criterion is acknowledged as a limitation rather than a methodological strength, as it may have excluded relevant paywalled studies (see Limitations).

Studies were excluded for: (i) generic multiclass tumor classification without explicit glioma grading; (ii) focus on prognosis or treatment response without grade assessment; (iii) use of non-MRI modalities (e.g., CT, PET) without MRI integration; (iv) insufficient methodological detail; or (v) thematic deviations not generalizable to adult supratentorial gliomas (e.g., pediatric diffuse intrinsic pontine glioma studies).

The study selection workflow across both search pipelines is summarized in Fig. 1. Manual database searches retrieved 960 records (IEEE Xplore: $n = 609$; ScienceDirect: $n = 228$; PubMed: $n = 123$), reduced to 91 candidate studies after year and document type filtering, and to 15 after abstract manually-screening. Concurrently, AI tools retrieved 42 records (Elicit: $n = 32$; Consensus: $n = 10$); metadata screening removed 6 articles and deduplication removed 5 duplicates, yielding 31 unique records, narrowed to 24 following abstract screening.

Combining the 15 manual and 24 AI-retrieved records, with no overlap between the two pools, yielded 39 unique studies for full-text eligibility assessment. Full-text access could not be confirmed as open-access for 20 of these records, leaving 19 studies with available full text; of these, 4 were excluded for thematic/methodological deviations, leaving a final synthesis of 15 included studies.

It should be noted that this final set of 15 studies is not identical to the 15 manually-screened candidates described above: only 5 of the 15 manual candidates were retained in the final synthesis, while the remaining 10 studies were recovered through the AI-assisted pipeline. The manual search covered publications from 2022 to 2026; five studies predating this period were included via the AI-assisted pipeline due to their methodological relevance."

The fifteen selected studies span standalone radiomic and deep learning baselines to hybrid and graph-based architectures, with sample sizes from 54 to 704 subjects across institutional cohorts and public benchmarks (BraTS, TCIA). This diversity supports a representative mapping of current practices, though the open-access restriction and specific search pipelines limit generalization across clinical tasks, WHO editions, and validation protocols.

III. Literature Review

A. Classical Radiomics

Classical radiomics follows a common pipeline: segmentation of the region of interest (ROI), extraction of quantitative features (shape, intensity, texture), and application of a machine learning classifier — the basis for much of the glioma grading literature.

The most frequently cited methodological starting point in this field is the work by Cho et al. [7], which utilized the BraTS 2017 dataset (a public and multi-institutional dataset contributed by thirteen centers from The Cancer Imaging Archive — TCIA) comprising 285 cases. The authors extracted 468 3D volumetric radiomic features from four MRI sequences (T1, T1CE, T2, FLAIR), covering three types of regions of interest: the tumor core, the enhancing area, and peritumoral edema. Selection using minimum Redundancy Maximum Relevance (mRMR), a technique that prioritizes features with high relevance to the target class and low redundancy among themselves, reduced the feature space to five features per fold, with Spherical Disproportion and Compactness consistently appearing in at least three of the five folds. These results reaffirm a clear biological interpretation: tumors with more irregular shapes tend to be high-grade, reflecting the infiltrative growth of GBM. After applying Random Forest, support vector machine (SVM), and logistic regression with cross-validation across the five folds, the authors achieved an AUC of 0.903 for distinguishing between high-grade glioma (HGG) and LGG.

It is not always necessary to resort to complex multiparametric protocols to achieve meaningful diagnostic discrimination. Kumar et al. [8] prove that a single T2W sequence, processed with many spatial scale filters (Gaussian Laplacian), can distinguish LGG from HGG with an AUC of 0.82 in 83 patients imaged on two distinct scanners (1.5T and 3T). T2W is a universal sequence that does not require the use of contrast agents, which gives it great practical value. Following a similar line of reasoning about accessible modalities, Vijithananda et al. [9] propose a multiclass classification into three categories using only non-contrast-enhanced ADC (apparent diffusion coefficient) maps, which indirectly reflect the cellular density of the tissue, resulting in an accuracy of 88.14% and an AUC of 0.99 for LGG compared to the other groups.

When the focus shifts from tumor morphology to vasculature, new possibilities emerge. Sudre et al. [5] explore a resonance perfusion technique called DSC-MRI, which quantifies cerebral blood flow in a dataset from six European centers, comprising a total of 333 patients. From rCBV maps normalized to the basal ganglia of the contralateral hemisphere, they extracted 29 features and applied a Random Forest model with two-fold cross-validation, repeated 250 times. This method achieved 71% accuracy for IDH prediction and 53% for WHO three-class grading. In the latter case, 87% of cases were classified with a margin of no more than one grade difference from the actual grade, demonstrating that, although exact classification is challenging, the model rarely misclassifies by more than one grade. It is also observed that shape features better distinguish between grades II and III, while texture and intensity are more informative for distinguishing between grades III and IV.

A different perspective on which features truly matter is presented in the work by Hassan et al. [10], who, instead of hundreds of features, used only eight morphological measures derived from the BraTS 2015 segmentation masks: presence of enhancement, absence of enhancement, edema, and necrosis, and their relative ratios. The features were evaluated using a technique called Ensemble Learning with Adaptive Power Mean Combiner (EL-APMC), which combines linear discriminant analysis (LDA) with bootstrap bagging without replacement and an adaptive power mean combiner, achieving an accuracy of 88.73% and an F1-score of 93.12% in 275 patients, a result that outperforms the 21 classifiers compared by the authors.

Taken as a whole, classical radiomics studies show that shape and texture features of the enhancing region and tumor core best discriminate grades, though the reliance on manual ROI segmentation makes pipelines labor-intensive and inconsistent across centers.

B. Deep Learning and Convolutional Neural Networks

The main promise of deep learning compared to classical radiomics is precisely that it eliminates the need to define in advance which features are relevant: instead of extracting prespecified measures, convolutional neural networks (CNNs) learn hierarchical representations directly from the image's pixels. This difference in philosophy has important practical consequences, and the literature is beginning to map them out clearly.

The most direct and controlled comparison between the two approaches appears in the study by Gutta et al. [6], conducted with 237 patients from a single center (USC). The authors extracted 1,284 radiomic features using PyRadiomics and achieved 64% accuracy with Gradient Boosting as the best classical classifier. In parallel, they trained a customized CNN using 2D patches of 150×150 pixels for multiclass classification across four WHO grades, and the result is

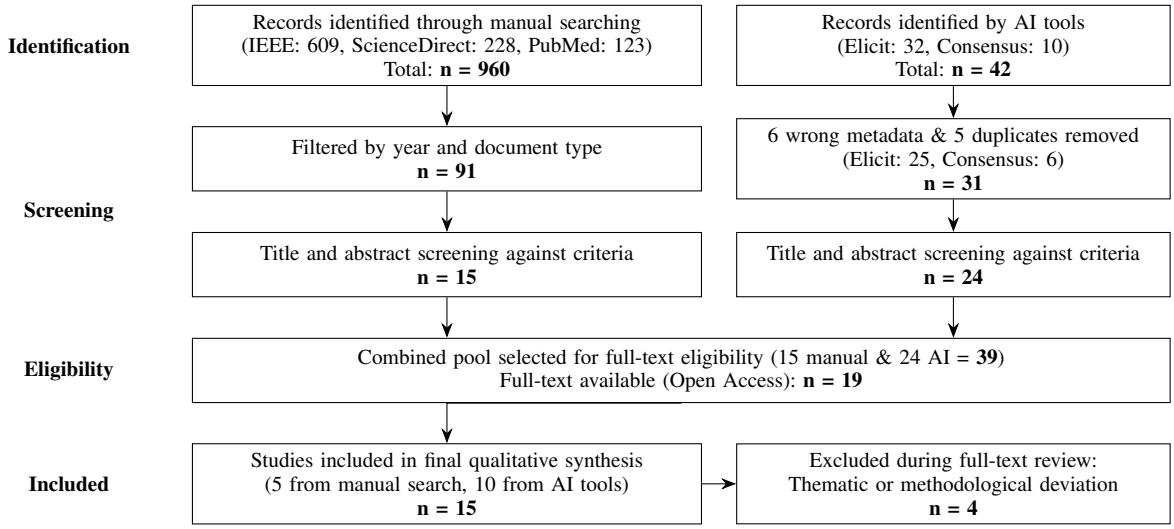


Fig. 1. PRISMA-style flow diagram of the study selection process.

striking: the CNN achieved 87% accuracy, an improvement of 23 percentage points. Analysis using Gradient-weighted Class Activation Mapping (Grad-CAM), a visualization technique that highlights the image regions most influential to the network's decision, revealed that T1CE is strongly considered for grades III and IV, FLAIR is more discriminative for grade II, and edema is the most informative tumor component.

The question of how to optimize the network architecture — rather than simply choosing the method — was explored by Shahriar et al. [11], who proposed a CNN optimized using KerasTuner with the Hyperband algorithm to be applied to the dataset Erasmus Glioma Database (EGD), which was balanced with 77 patients per grade and 16,956 T2W-FLAIR slices. KerasTuner is a library for automatic hyperparameter search, and optimization via Hyperband yielded an average accuracy gain of 1.6% per fold. The final model achieved 89% for three grades and 93% for binary classification, outperforming EfficientNet-B0 (83%), Xception (82%), and ResNet-50 (81%) on the same dataset. Saliency maps, another visualization technique simpler than Grad-CAM, reveal a consistent focus on peritumoral regions, especially at higher grades.

There are also approaches that combine the insight provided by radiomic features with the learning capabilities of neural networks. Qin et al. [12] trained a feedforward artificial neural network (ANN) — that is, a neural network with a direct flow of information (without recurrent connections) — using 19 radiomic features from T1-weighted contrast-enhanced (CE-T1W) images for classification into four WHO grades, achieving an accuracy of 87.04%. This is the only study in the set that explicitly proposes classification into four categories (I, II, III, and IV) using this approach.

The set of studies suggests that deep neural networks tend to outperform classical radiomics as data volume increases, though this pattern is only loosely supported across the heterogeneous studies reviewed. Interpretability remains a challenge nonetheless: Grad-CAM and saliency maps offer partial insight into what the model learns, but do not equate to the biological interpretability of radiomic features.

C. Hybrid Approaches: Radiomics and Deep Learning

The central hypothesis of hybrid approaches is precisely this: classical radiomic features and features learned through deep learning capture complementary aspects of tumor information, and their fusion should outperform standalone approaches.

Ning et al. [13] systematically and rigorously test this hypothesis using 567 patients distributed between a TCIA training set (233 patients) and an external test set from Nanfang Hospital in China (334 patients). The integrated model fuses radiomics and CNNs via SVM kernel fusion (kSVM), a technique that combines two feature spaces using a sum kernel in a support vector machine, and achieving an AUC of 0.94 in internal validation and 0.88 in the external test. The results reinforce the superiority of the integrated approach over individual architectural configurations; furthermore, when compared to the manual analyses of three radiologists, the model outperformed them in sensitivity (88% vs. 80–84%) and specificity (81% vs. 72–78%).

The findings of Ning et al. [13] raise a question: which of the two sources of information — radiomics or CNN features — contributes most to the hybrid model's performance? The answer provided by the work of Ding et al. [14] represents a step forward in this area. By integrating radiomics and transfer learning on CE-T1W images across three reconstruction planes (axial, coronal, and sagittal) using data from three institutions (151 patients), the authors found that the optimal model, consisting of a Random Forest with combined radiomic and VGG16 features, achieves an AUC of 0.898 on the multicenter test set; however, radiomic features account for 88.2% of the selected features, while VGG16 contributes only two. This suggests that transfer learning from ImageNet to MRI has limited capacity and relevance in this context.

When feature fusion is insufficient to capture the complexity of the tumor, one alternative is to let radiomics guide the learning process itself. This was the approach proposed by Shboul et al. [15], who introduced the concept of guided deep radiomics: known radiomic features, such as PTPSA fractal analysis and multifractal Brownian motion (mBm) — which are measures of the tumor's geometric irregularity and spatial complexity — guide the learning of deep neural networks. The pipeline evaluated in the BraTS 2017 and 2018 benchmarks focuses on predicting survival in GBM and incorporates persistent homology features that capture the tumor's 3D topology — that is, global geometric properties such as connectivity and cavities. The model, called XGBoost, won the BraTS 2017 competition with an accuracy of 73% on the training set and 67.9% on the validation set.

These hybrid approaches show that integration need not be complex — feature fusion, partial transfer learning, or geometric guidance all reinforce radiomics and deep learning as complementary rather than rival approaches, provided validation is rigorous enough to confirm the gain is real.

D. Graph Architectures

A new class of architectures has gained traction in the literature on glioma grading: graph neural networks (GNN/GCN). They differ from conventional CNNs in how they represent information — not as regular grids of pixels, but as structures mapping relationships between regions, referred to as nodes, opening possibilities that standard CNNs do not explore.

An initial approach to exploring graphs in this context involves modeling the relationships between patients in the feature space. Ali et al. [1] adopt this strategy in the RGNN3D model, combining GCN and long short-term memory (LSTM) networks to grade gliomas based on 112 radiomic biomarkers from 704 patients in the BraTS dataset, in which each individual represents a node in the graph. The architecture employs three GCN layers interleaved with LSTMs to ease message passing and keep long-range dependencies, which mitigate the vanishing gradient problem — a classic challenge of deep networks in which error signals progressively decay across layers. The LSTM addresses this through forget and forget-gate mechanisms that selectively control which information is retained or discarded in the cell's memory, allowing relevant gradients to flow without degradation, regardless of the network's depth. Although the model achieves a remarkable accuracy of 98.58% in the T1 and T2 modalities, this result should be interpreted with caution due to three factors: the BraTS dataset is pre-standardized by the organizers, which favors high metrics; validation was limited to the internal environment — that is, no tests were conducted on independent external cohorts; and the test employed binary classification (LGG vs. GBM), a task considerably simpler than four-class grading. Finally, local interpretability via Local Interpretable Model-agnostic Explanations (LIME) revealed the key decision biomarkers: Mesh Volume and Maximum Diameters were decisive for GBM, while Minor Axis Length and Gray Level Non-Uniformity distinguished LGG.

A second way to use graphs is to model the internal structure of the tumor volume rather than the relationships between patients. Guo et al. [16] developed the MMIF-GCN module for this purpose, modeling relationships between adjacent slices in the 3D volume and across the four MRI sequences simultaneously. Each node represents a slice containing a tumor, with embedding vectors that integrate the 2D features of the four sequences into a 200-dimensional vector. The architecture contains four GCN layers that progressively refine these vectors, and evaluation was performed on the BraTS 2020 dataset (369 patients) and the private dataset from Henan Provincial Hospital (232 patients). MMIF-GCN with ResNet-34 achieves an AUC of 0.986 and an accuracy of 94.7%; this AUC close to 1.0 is justified by the binary classification (HGG vs. LGG), as in Ali et al. [1]. Furthermore, the study does not clarify whether cross-validation prevented data leakage between the two combined datasets. Still, the model outperforms the 3D-CNN (AUC 0.949) and the individual sequence GCN (AUC 0.970), demonstrating a significant gain from the application of graph-based multimodal fusion.

Ali et al. [1] and Guo et al. [16] thus address complementary aspects — patient-level relationships versus tumor-volume structure — but share the same limitation: neither provides formal mechanisms to explain the graph's decisions, limiting clinical adoption.

E. Magnetic Resonance Imaging Modalities

The choice of MRI sequences is not a minor technical decision: it directly determines what kind of biological information the model will have access to. This issue affects all the approaches discussed in the previous sections — from classical radiomics to deep learning and graph models — and the reviewed studies shed light on it in complementary ways.

Most of the reviewed studies use the standard BraTS multiparametric protocol: T1, T1CE, T2, and FLAIR. Within this protocol, the activation maps by Gutta et al. [6] provide the clearest evidence about the role of each sequence: T1CE is most discriminative for grades III and IV, FLAIR for grade II, and the importance of T2 increases progressively with grade. This hierarchy is not arbitrary; it reflects real biological differences between the grades: enhancement on T1CE indicates disruption of the blood-brain barrier, which is more pronounced in more aggressive tumors, while FLAIR is sensitive to the edema characteristic of lower grades.

The choice of single sequences, rather than the complete protocol, is a legitimate strategy when guided by clear clinical or methodological motivations. Kumar et al. [8] choose T2W for its clinical universality and the absence of contrast use, a significant advantage in settings with limited infrastructure or when contrast is contraindicated. Vijithananda et al. [9] use ADC, arguing that the diffusion coefficient reflects tumor cellularity more directly than morphological sequences. A different perspective is presented by Shahriar et al. [11]; their CNN is essentially modality-agnostic, meaning that T1, T1CE, T2, and FLAIR sequences yield equal performance, raising the

hypothesis that, with enough data, the network learns to extract relevant information regardless of the specific sequence.

Certain specialized modalities capture aspects of tumor physiology that conventional sequences cannot access. Sudre et al. [5] represent the only study using DSC-MRI with exclusive perfusion imaging, validating that the tumor vasculature captured by rCBV is a surrogate of IDH status — a biologically grounded connection between perfusion and molecular genotyping. A limitation less frequently discussed in the literature is that most studies analyze only the axial plane, disregarding volumetric information that the coronal and sagittal planes could provide. Ding et al. [14] address precisely this gap by applying CE-T1W in three reconstruction planes, demonstrating that the triplanar perspective recovers volumetric information lost in the analysis of isolated axial slices. Chiu and Yen [17] combine four sequences (T1W, T2, FLAIR, ADC) to map intratumoral subregions of GBM, leveraging the complementarity between contrast, edema, diffusion, and morphology to characterize tumor heterogeneity.

The choice of which sequence to use depends on the task, clinical context, and method — deep learning with large datasets is less sensitive to modality choice, while classical radiomics with small samples depends critically on it.

F. Radiogenomic Approaches

The 2016 and 2021 updates to the WHO classification redefined the conceptual framework of the field by establishing markers such as IDH mutation, 1p/19q codeletion, and MGMT promoter methylation for the definitive diagnosis of gliomas [4]. This molecular milestone directly drives the studies analyzed, which expand the scope beyond pure morphological grading and move toward molecular prediction, prognostic stratification, and mapping of intratumoral heterogeneity. This reorientation changes the clinical premise; thus, the central question shifts from "What is the histological grade?" to "What is the tumor's biological profile, and which treatment will be most effective?"

The most comprehensive radiogenomic approach in this set of studies is that of Bilalov et al. [18], whose framework integrates radiomic embeddings from a multiparametric CNN with molecular biomarkers (EGFR amplification, IDH mutation, and MGMT promoter methylation) through multi-omic fusion. This strategy combines data from different biological layers — such as imaging, the genome, and the epigenome — into a single predictive model. This approach achieved a C-index of 0.78 and an AUC of 0.82 in a cohort of 112 patients during external validation, where the C-index measures the model's ability to correctly rank pairs of patients by survival risk. The molecular dimension is also explored by Sudre et al. [5], but from the perspective of classical radiomics. By incorporating IDH status as a secondary target, the authors demonstrated that DSC-MRI features, especially those related to texture and shape, differentiate IDH-mutant tumors from wildtype tumors with 71% accuracy.

When all patients already have grade 4 GBM, the question of grading loses its meaning, and the focus shifts entirely to prognosis. This is precisely the perspective adopted by Shboul et al. [15], whose aim is to predict survival across three risk categories within a homogeneous set of GBM cases. For the structural analysis of the data, the authors used t-distributed stochastic neighbor embedding (t-SNE) to visualize the high-dimensional data structure. The mapping revealed that the median survival class is structurally indistinguishable from the others in the feature space. This finding is relevant because it demonstrates that the challenge is biological, not algorithmic, reflecting the actual overlap of tumor phenotypes associated with intermediate prognoses.

Heterogeneity does not exist only among patients with different grades. It is also present within the tumor itself, where regions with distinct biological behaviors coexist within the same volume. This internal spatial diversity is addressed by Chiu and Yen [17], who, by cataloging the intratumoral subregions of GBM (necrotic core, solid tumor, peritumoral tissue, and edema) and quantifying their volumetric ratios, establish that edema is the volumetrically dominant component (31–39% of abnormal bulk volume — ABV), followed by the solid tumor (31–43%) and necrosis (17–23%), with each subregion associated with distinct biological processes. Mapping these subregions using multiparametric radiomics can thus provide a spatial signature of the tumor with prognostic value independent of the grade.

These studies point to a reorientation beyond technical improvements: integrating radiomic features, molecular markers, and clinical variables into multimodal predictive models is the field's identified next step.

G. Reported Challenges

Table I summarizes the main methodological characteristics of the fifteen reviewed articles. The reported metrics are not directly comparable across studies, as they reflect different tasks and validation protocols.

Three structural challenges run through the literature regardless of methodological approach, limiting practical clinical application: dataset size, validation diversity, and interpretability.

The first challenge is the size of the datasets. The scarcity of labeled data is the most frequently cited obstacle: Kumar et al. [8] worked with 83 patients, Chiu and Yen [17] with 54 — a sample size the authors themselves note compromises statistical power and scalability — and Bilalov et al. [18] with 112. This situation is further compounded by class imbalance, as Grade I is consistently rare: 17 cases in Gutta et al. [6] and 23 in Qin et al. [12]. To overcome this limitation, various strategies have been employed, such as SMOTE [9], fold balancing [11], and bootstrap bagging [10]; however, these approaches only partially mitigate the statistical fragility associated with small datasets. Another point is that six of the fifteen reviewed studies use BraTS as their primary benchmark; this dataset is preprocessed and standardized by the challenge organizers, which introduces a homogeneity that limits the true assessment of the models' generalization ability to heterogeneous clinical data.

TABLE I
COMPARISON OF THE FIFTEEN REVIEWED ARTICLES

Paper	Approach	Sequence	Dataset (n) / Split	Validation	Evaluation*	Objective	Limitations	XAI
Guo et al., 2022 [16]	CNN & GCN multimodal	T1, T1CE, T2, FLAIR	BraTS20 (369) & Henan Hosp. (232), 5-Fold CV	Internal CV & External (Henan Hosp.)	AUC 0.986; Accuracy 94.7%	Binary classification (HGG/LGG)	Black-box; complex architecture	No
Ali et al., 2026 [1]	Radiomics, GNN/LSTM	T1, T2, T1CE, FLAIR	BraTS19 & BraTS20 (Train: 563, Test: 141)	Internal	Accuracy 98.58%	Binary classification (LGG/GBM)	No external validation	LIME
Ning et al., 2021 [13]	Radiomics & CNN & kSVM	T1CE, FLAIR	TCIA (Train: 233) & Nanfang Hosp. (Test: 334)	External (Nanfang Hosp.)	AUC 0.88	Binary classification (LGG/GBM)	Cross-site harmonization needed	No
Bilalov et al., 2025 [18]	CNN + Radiogenomics	T1, T1CE, T2, FLAIR	TCGA (Train: 75) & China Univ. (Test: 37)	External (China Univ.)	AUC 0.82; C-idx 0.78	GBM prognosis (survival)	GBM-only; scanner variability	No
Ding et al., 2022 [14]	Radiomics & VGG16 triplanar	T1CE (3 planes)	3 Univ. Hospitals (Train: 101, Test: 50)	External (Multicenter)	AUC 0.898	Binary classification (LGG/HGG)	Triplanar slice sensitivity	Grad-CAM
Gutta et al., 2021 [6]	CNN pure	T1, T1CE, T2, FLAIR	Keck Hosp. (Train: 166, Val: 36, Test: 35)	Internal	Acc 87% (CNN); Acc 64% (GB)	Multiclass grading (WHO 1–4)	No external validation	Grad-CAM
Shahriar et al., 2025 [11]	Lightweight CNN + KerasTuner	T2W-FLAIR	EGD (231), 5-Fold CV	Internal CV	Acc 89% (Grad.); 93% (Class.)	3-class grading + binary	No cross-center validation	Saliency Maps
Chiu & Yen, 2022 [17]	Multiparametric radiomics & RF	T1CE, T2, FLAIR, ADC	Taiwan Univ. Hosp. & TCIA (54), 5-Fold CV	Internal CV	AUC > 0.95	Intratumoral subregion mapping	Very small n (54)	Parametric Heatmap
Cho et al., 2018 [7]	Radiomics & mRMR & SVM/RF	T1, T1CE, T2, FLAIR	BraTS17 (285), 10-Fold CV	Internal CV	AUC 0.903	Binary classification (HGG/LGG)	Manual segmentation needed	No
Hassan et al., 2024 [10]	Novel MRI features & EL-APMC	T1CE, T2	BraTS15 (275), 5-Fold CV	Internal CV	Acc 88.73%; F1 93.12%	Binary classification (LGG/HGG)	Outdated dataset (BraTS15)	No
Shboul et al., 2019 [15]	Deep Radiomics + XGBoost	T1, T1CE, T2, FLAIR	BraTS17 & BraTS18 (163), LOOCV	Internal LOOCV	Accuracy 73%	GBM survival prediction	LOOCV only, no external test	t-SNE + KM
Qin et al., 2025 [12]	Radiomics & ANN	T1CE	1 Univ. Hospital (Train: 253, Test: 109)	Internal	Accuracy 87.04%	Multiclass grading (WHO 1–4)	No external validation	No
Sudre et al., 2020 [5]	Radiomics & RF	DSC-MRI (rCBV)	6 Univ. Hospitals (333), leave-one-center-out	Internal LOCO (Multicenter)	Acc 71% IDH; 53% grade	3-class grading & IDH prediction	Low grading accuracy	No
Vijithananda et al., 2023 [9]	Radiomics ADC & RF	ADC	Sri Lanka Hosp. (Train: 70, Test: 18)	Internal	Acc 88.14%; AUC 0.99	Multiclass classification (3 classes)	Small n (88); overfitting risk	No
Kumar et al., 2023 [8]	Radiomics T2W & RF	T2W	Tata Memorial Hosp. (Train: 58, Test: 25)	Internal	AUC 0.82	Binary classification (LGG/HGG)	Small n; single sequence	No

*Metrics reflect different tasks and validation protocols and cannot be directly compared.
CV = Cross-Validation; LOOCV = Leave-One-Out CV; LOCO = Leave-One-Center-Out

The second challenge is the diversity of validation strategies, which prevents equal-footing comparison across studies. Ning et al. [13] combines internal validation with an independent external dataset, while Hassan et al. [10], Kumar et al. [8], and Qin et al. [12] validate exclusively on the same dataset, limiting generalizability.

Confidence intervals for reported metrics (AUC, accuracy, F1) were rarely provided in the original studies and are therefore omitted from Table I; similarly, explicit data-leakage safeguards (e.g., patient-level versus slice-level splitting) were inconsistently reported, limiting the ability to assess result reliability across studies.

Likewise, a structured quality-assessment table including WHO classification edition, IDH/1p19q status, and segmentation method was not feasible, as these fields are inconsistently reported across the fifteen included studies; this heterogeneity in reporting is itself a limitation of the underlying literature rather than of the present review.

The third challenge is interpretability, the most critical for clinical adoption, since a model that cannot explain its decisions is unlikely to be incorporated into medical practice as a diagnostic tool. Of the fifteen studies reviewed, only six em-

ploy some formal interpretability technique: LIME [1], Grad-CAM [6], [14], saliency maps [11], t-SNE combined with Kaplan-Meier analysis [15], and hierarchical heatmaps [17]. The remaining nine studies treat the model as a black box, a shortcoming that must be addressed as a core need for any system intended for clinical use.

IV. Conclusion

No single approach fully resolves glioma grading: classical radiomics offers interpretability and works well with small samples; deep learning captures patterns radiomics overlooks but needs more data and lacks transparency; hybrid and graph models push performance further at the cost of opacity and reliance on internal validation. Within this corpus, graph-based models reported the highest figures on binary LGG/HGG classification using standardized benchmarks (BraTS19 & BraTS20); these results should not be generalized across tasks, as no head-to-head comparison under identical conditions exists, although there are favorable evaluation conditions (binary classification, standardized data).

This review has methodological limitations that should be acknowledged. A formal risk-of-bias assessment tool, such as QUADAS-2, was not applied to the fifteen included studies due to time constraints; such an assessment would allow a more systematic judgment of methodological quality across domains (patient selection, index test, reference standard, and flow and timing) and is proposed as a direction for future systematic reviews on this topic. Additionally, restricting eligibility to open-access full text may have excluded relevant paywalled studies, so the final sample of fifteen studies is small relative to the breadth of the field, and findings should be interpreted as indicative of current trends rather than as a basis for broad generalization.

The challenges that the next generation of models will need to address are well defined: larger, multicenter datasets that reflect real-world variability; standardized validation protocols that allow for fair comparison of studies; and the application of formal interpretability as a need, not an option. The expansion toward radiogenomics and the mapping of intratumoral subregions is the natural direction: models capable of grading the tumor, predicting its molecular profile, and stratifying prognosis in a single noninvasive analysis represent the concrete and achievable goal for this field over the next decade.

Among the fifteen studies reviewed, the hybrid model by Ning et al. [13] represents the most suitable starting point for later work, as the only one combining radiomic and deep features with external validation on 334 patients, outperforming three clinical radiologists in sensitivity and specificity. This combination of methodological integration and robust external validation is precisely what the literature identifies as the central gap. A natural next step would be to replicate and expand this approach, incorporating more modalities, such as ADC or DSC-MRI, and testing generalizability on multicenter datasets with heterogeneous acquisition protocols.

V. Acknowledgment

This article was supported by the CETELI-INOVA project of the Center for R&D in Electronic and Information Technology at the Federal University of Amazonas.

In accordance with the IEEE policy on the use of AI generated content, the authors disclose that DeepL (DeepL SE) and Claude (Anthropic) were used as support tools for translation into English, textual organization, and grammar revision across all sections of this manuscript. The tool was not used to generate scientific content, experimental results, figures, or code. All intellectual decisions and full responsibility for the final content remain solely with the authors.

References

[1] M. A. Ali *et al.*, "RGNN3D: A hybrid radiomic graph neural network for 3D MRI glioma grading," *Knowledge-Based Systems*, vol. 337, p. 115343, 2026.

[2] E. B. Claus, K. M. Walsh, J. K. Wiencke, A. M. Molinaro, J. L. Wiemels, J. M. Schildkraut, M. L. Bondy, M. Berger, R. Jenkins, and M. Wrensch, "Survival and low-grade glioma: the emergence of genetic information," *Neurosurgical Focus*, vol. 38, no. 1, p. E6, 2015.

[3] Q. T. Ostrom, L. Bauchet, F. G. Davis, I. Deltour, J. L. Fisher, C. E. Langer, M. Pekmezci, J. A. Schwartzbaum, M. C. Turner, K. M. Walsh *et al.*, "The epidemiology of glioma in adults: a "state of the science" review," *Neuro-Oncology*, vol. 16, no. 7, pp. 896–913, 2014.

[4] D. N. Louis *et al.*, "The 2021 WHO classification of tumors of the central nervous system: a summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021.

[5] C. H. Sudre *et al.*, "Machine learning assisted DSC-MRI radiomics as a tool for glioma classification by grade and mutation status," *BMC Medical Informatics and Decision Making*, vol. 20, p. 149, 2020.

[6] S. Gutta *et al.*, "Improved glioma grading using deep convolutional neural networks," *American Journal of Neuroradiology*, vol. 42, no. 2, pp. 233–239, 2021.

[7] H.-h. Cho *et al.*, "Classification of the glioma grading using radiomics analysis," *PeerJ*, vol. 6, p. e5982, 2018.

[8] A. Kumar *et al.*, "Machine-learning-based radiomics for classifying glioma grade from magnetic resonance images of the brain," *Journal of Personalized Medicine*, vol. 13, no. 6, p. 920, 2023.

[9] S. M. Vijithananda *et al.*, "Texture feature analysis of MRI-ADC images to differentiate glioma grades using machine learning techniques," *Scientific Reports*, vol. 13, p. 15772, 2023.

[10] M. F. Hassan *et al.*, "An effective ensemble learning approach for classification of glioma grades based on novel MRI features," *Scientific Reports*, vol. 14, p. 11977, 2024.

[11] S. Shahriar *et al.*, "Glioma grade classification: Hypertuned CNN with saliency-based insight from MRI," *IEEE Access*, vol. 13, pp. 4521–4535, 2025.

[12] Y. Qin *et al.*, "Magnetic resonance imaging radiomics-driven artificial neural network model for advanced glioma grading assessment," *Medicina*, vol. 61, no. 6, p. 1034, 2025.

[13] Z. Ning *et al.*, "Multi-modal magnetic resonance imaging-based grading analysis for gliomas by integrating radiomics and deep features," *Annals of Translational Medicine*, vol. 9, no. 4, p. 298, 2021.

[14] J. Ding *et al.*, "Developing and validating a deep learning and radiomic model for glioma grading using multiplanar reconstructed magnetic resonance contrast-enhanced T1-weighted imaging," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 2, pp. 1517–1528, 2022.

[15] Z. A. Shboul *et al.*, "Feature-guided deep radiomics for glioblastoma patient survival prediction," *Frontiers in Neuroscience*, vol. 13, p. 966, 2019.

[16] L. Guo *et al.*, "Multiparametric magnetic resonance imaging information fusion using graph convolutional network for glioma grading," *Journal of Healthcare Engineering*, vol. 2022 (7315665), p. 11, 2022.

[17] F.-Y. Chiu and Y. Yen, "Efficient radiomics-based classification of multiparametric MR images to identify volumetric habitats and signatures in glioblastoma," *Cancers*, vol. 14, no. 6, p. 1475, 2022.

[18] E. Bilalov *et al.*, "Deep learning guided radiogenomic signatures for prognostic stratification in glioblastoma multiforme," *Journal of Cancer Research Updates*, vol. 14, pp. 237–246, 2025.

Architectural Rigidity versus Modern Scaling Laws in Asymmetric Medical Data for Cardiac Arrhythmia Classification

Carlos A.L. Ortiz
(PEB-UFRJ)
(Federal University Of Rio de Janeiro)
Rio de Janeiro, Brazil
carlos.ortiz@peb.ufrj.br
(Goiano Federal Institute - URUTAI)
Goiás, Brazil
carlos.ortiz@ifgoiano.edu.br

Julio C. Ferreira
Computational and Applied Mathematics
Goiano Federal Institute - URUTAI
Goiás, Brazil
julio.ferreira@ifgoiano.edu.br

Gabriel da Silva Vieira
Computer Vision Laboratory
Goiano Federal Institute - URUTAI
Goiás, Brazil
gabriel.vieira@ifgoiano.edu.br

Jurandir Nadal
(PEB-UFRJ)
(Federal University Of Rio de Janeiro)
Rio de Janeiro, Brazil
jn@peb.ufrj.br

Abstract*—Automated cardiac arrhythmia classification is vital for enhancing intensive care monitoring and reducing clinician alarm fatigue. However, extreme intrinsic class imbalance across standard pathological categories often leads to the complete suppression of minority class features during deep learning optimization. This study presents a rigorous architectural benchmark evaluating five state-of-the-art convolutional neural networks—ResNet50, ResNet50V2, EfficientNetV2B0, DenseNet121, and ConvNeXtTiny—under varying degrees of data scarcity. Moving away from static dataset splits, we implement an exhaustive 4-fold cross-rotation validation pipeline that operates on a cumulative pool of 109,381 2D ECG derivative images generated from the MIT-BIH Arrhythmia Database and structured according to the AAMI/ANSI EC57 standard. The consolidated multi-round results reveal a critical engineering trade-off: next-generation networks optimized under modern computational scaling laws (EfficientNetV2B0 and ConvNeXtTiny) suffer a catastrophic feature drain toward majoritarian normal profiles in hyper-sparse configurations, collapsing to 0.00% F1-scores when class support drops below a critical threshold. Conversely, the classical ResNet50 framework demonstrates superior topological preservation, anchoring minority gradient pathways through its unregularized identity shortcut connections. ResNet50 achieved the highest consolidated mean global accuracy (92.16% ± 7.15%) and a resilient classification standard for the rarest abnormal class (61.01% ± 39.81% F1-score), heavily outperforming its modernized counterparts. These findings introduce a pivotal paradigm shift for clinical software deployment, demonstrating that structural rigidity and unpruned architectural highways remain paramount to safeguarding diagnostic sensitivity in real-world asymmetric medical data streams.

***Index Terms*—Arrhythmia Classification, Convolutional Neural Networks, Class Imbalance, Cross-Rotation Validation, ResNet50, ECG Derivative Images**

This study received financial support from Public Notice No. 16/2026 - Event Participation - Pipecti Staff Member / Faculty Staff, CNPq (Brazilian Research Council, Grant # 309247/2025-0), FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Grant # E26/201.247/2021), the Brazilian Chamber of Deputies (Parliament Amendment #44890020)

I. Introduction

For over a century, the electrocardiogram (ECG) has served as the primary non-invasive tool for assessing cardiac bioelectric activity [1]. By capturing P, Q, R, S, and T wave patterns, the signal provides crucial morphological indicators of cardiovascular health. Continuous ECG monitoring is essential in ambulatory care and continuous clinical screening to detect life-threatening arrhythmias. However, current automated surveillance tools frequently suffer from high false-alarm rates driven by noise artifacts and severe class imbalance [2], [3]. Developing resilient automated classifiers remains a research priority to enhance diagnostic reliability and reduce alarm fatigue in automated screening systems.

To evaluate classification performance under standardized conditions, literature widely adopts the AAMI/ANSI EC57 standard. This protocol categorizes raw beats from the benchmark MIT-BIH Arrhythmia Database [4] into five physiological classes: Normal (N), Supraventricular ectopic (S), Ventricular ectopic (V), Fusion (F), and Unknown (Q). While traditional detection algorithms often use the first derivative of the ECG solely to improve signal-to-noise ratio, this work investigates its direct transformation into 2D derivative visual features for convolutional neural network (CNN) classification.

A central challenge in automated arrhythmia detection is the extreme class imbalance across pathological categories, which often suppresses minority class sensitivity. To evaluate how deep learning backbones perform under variable data scarcity, this study presents a patient-independent, 4-fold cross-rotation benchmark across 109,381 ECG derivative images. We analyze the empirical behavior of classical residual design (ResNet50) against modernized architectures (EfficientNetV2B0 and ConvNeXtTiny), assessing their performance when exposed to severe localized minority class scarcity.

II. Materials and Methods

A. Dataset

To evaluate the proposed deep learning methodology under rigorous benchmark conditions, this study uses the landmark MIT-BIH Arrhythmia Database from PhysioNet [4]. The repository comprises 48 fully annotated, half-hour ambulatory ECG records obtained from 47 distinct subjects, digitized at 360 Hz with 11-bit resolution over a 10 mV range. To maintain strict feature consistency across the 24 hours of cumulative recording time and more than 107,000 baseline heartbeats, our analysis focused exclusively on the first lead, which is usually the Modified Lead II (MLII). This bipolar thoracic lead is systematically available across the database registers, capturing a uniform vector of ventricular depolarization.

Following standard literature protocols, record 207 was explicitly excluded from the experimental pool due to severe ventricular flutter and inconsistent morphological annotations that deviated from the structural definitions established by the AAMI/ANSI EC57 standard.

As illustrated in the Fig.I, to prevent data leakage and evaluate model robustness under class scarcity, the remaining data was partitioned into four patient-independent groups (detailed in Table II). No patient data was shared between training and testing sets in any fold. Across the four rotated folds, this setup yielded an un-augmented testing pool of 109,381 ECG derivative images.

TABLE I: Stratified patient-independent partition mapping from the MIT-BIH Arrhythmia Database.

Data Partition	MIT-BIH Patient Records
Data Set 1	100, 101, 102, 109, 116, 122, 124, 200, 201, 220, 221, 228
Data Set 2	105, 106, 117, 118, 119, 121, 203, 213, 217, 222, 234
Data Set 3	104, 111, 113, 115, 123, 202, 208, 209, 212, 215, 219, 223
Data Set 4	103, 107, 108, 112, 114, 205, 210, 214, 230, 231, 232, 233

B. Signals and Preprocessing

To prepare the continuous ECG signal for classification, a multi-stage preprocessing pipeline was implemented. First, a two-sample moving average filter was applied to attenuate high-frequency interference. Using physician R-peak annotations, individual heartbeats were centered and segmented into 108-sample windows, capturing full P-QRS-T complexes.

To eliminate baseline wander—a common low-frequency artifact caused by respiration and movement—and to accentuate QRS slope characteristics, the first-order difference of each 108-sample segment was computed, as illustrated in Fig. 1.

Each transformed segment was subsequently rendered into a 300×300 RGB image saved in JPEG format, representing the 2D visual derivative features of the beat. Representative samples for each of the five AAMI arrhythmia classes are displayed in Fig. 2.

To address class imbalance without contaminating validation data, random oversampling [5] was applied strictly to the training splits of each fold. The test splits remained entirely untouched and un-augmented to ensure an unbiased evaluation.

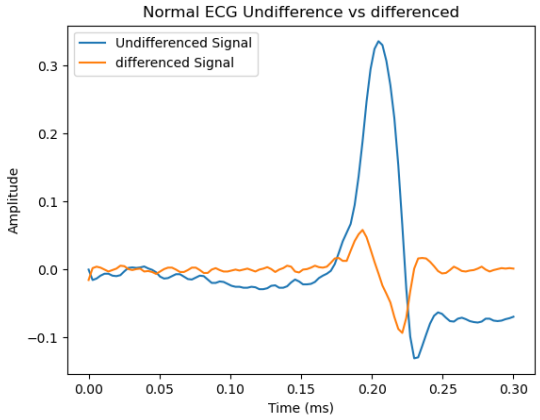


Fig. 1: Morphological comparison between the raw Normal ECG segment and its corresponding first-order derivative.

C. Deep Learning Neural Network Architectures

Five state-of-the-art convolutional architectures were selected as feature extractors (backbones) due to their distinct trade-offs between computational efficiency, structural depth, and capability to generalize complex patterns:

- ResNet50 and ResNet50V2: Originally developed by [9], the Residual Network (ResNet) introduced the concept of skip connections, mitigating the vanishing gradient problem in very deep networks. The ResNet50V2 variant [10] reformulates the residual block by applying Batch Normalization and activation (ReLU) prior to the convolutional layer (pre-activation), allowing for a cleaner gradient flow and greater numerical stability at the beginning of training.
- EfficientNetV2B0: Proposed by [11], EfficientNetV2 optimizes the compound scaling methods of the original version. Utilizing neural architecture search (NAS), it combines traditional MBConv convolutions with Fused-MBConv layers in the early stages, drastically optimizing training speed and video memory (VRAM) efficiency, making it ideal for large-scale datasets.
- DenseNet121: Introduced by [12], the Densely Connected Convolutional Network connects each layer to every subsequent layer within a dense block. This connectivity maximizes information flow and feature reuse throughout



(a) Normal (N) (b) Supraventricular (S) (c) Ventricular (V) (d) Fusion (F) (e) Unknown (Q)

Fig. 2: Representative 2D derivative image inputs across the five standardized AAMI arrhythmia categories.

the network, reducing the total number of redundant parameters and demonstrating high sensitivity in identifying fine lines and gradients

- ConvNeXtTiny: Developed by [13], the ConvNeXt family modernized purely convolutional networks (CNNs) by incorporating design philosophies from Vision Transformers (ViTs), such as the use of depthwise separable convolutions, inverted residual blocks, and GELU activation functions. The Tiny version acts as a high-precision extractor with low computational cost.

D. Hardware Resources and Software Environment

To process the massive volume of convolutional data and ensure experimental reproducibility, a mobile workstation with hardware-accelerated parallel processing was used. The computational system consisted of an Intel Core i7 processor, complemented by 32 GB of dynamic random-access memory (DDR4 RAM) and a high-speed solid-state drive (SSD) storage unit.

Hardware acceleration for neural network training and inference was conducted by a dedicated NVIDIA GeForce RTX 2060 Graphics Processing Unit (GPU). The software environment was built upon the Linux operating system, utilizing the Miniconda package manager for dependency isolation. The development pipeline was implemented in Python (version 3.10), leveraging the unified Keras 3 ecosystem coupled with the TensorFlow (version 2.16) backend, and supporting NVIDIA cuDNN analytical execution drivers and the CUDA Toolkit (version 12.3).

E. Data Ingestion and Organization Pipeline

To ensure maximum statistical validity and eliminate selection bias during model training and evaluation, an exhaustive 4-fold cross-validation rotation pipeline was implemented. The original dataset layout, containing segmented 2D electrocardiogram (ECG) derivative images, was organized into four distinct training blocks and four corresponding independent test sets. Rather than using a single static split, the training pipeline dynamically aggregated three of the four available training subsets per execution round, reserving the remaining distinct validation subset exclusively for final inference testing. This setup resulted in four independent experimental configurations as presented in TableII.

TABLE II: Training, Validatio and Test Sets

	Group			
Group	**Training**			**Test**
Group 1	Dataset 2	Dataset 3	Dataset 4	Dataset 1
Group 2	Dataset 1	Dataset 3	Dataset 4	Dataset 2
Group 3	Dataset 1	Dataset 2	Dataset 4	Dataset 3
Group 4	Dataset 1	Dataset 2	Dataset 3	Dataset 4

For each of these in the experimental groups, a stratified validation split was applied with 80% for weight adjustment and 20% for internal convergence validation, ensuring a fair statistical representation of all classes.

This structural rotation exposed the networks to varying degrees of localized class imbalance across rounds, forcing the architectures to generalize to morphological ECG boundaries under diverse distributional stresses. Crucially, the target samples of the specific test partition under evaluation remained completely unseen during the training phase of that respective round, ensuring zero data leakage and a pristine mathematical baseline for the final cross-fold synthesis.

The original images in the dataset, which had dimensions of 300×300 pixels, were subjected to bilinear resizing to a stable spatial resolution of 128×128 pixels with a three-channel color mapping (RGB). This strategic dimensional reduction was fundamental to optimizing GPU video memory (VRAM) usage and making sequential training viable. The processing batch size was fixed at 8 samples per iteration.

F. Transfer Learning Strategy and Training

To bypass the need for random weight initialization in deep networks—which would require weeks of processing—a Transfer Learning strategy was adopted. Each of the five evaluated networks were initialized with mathematical coefficients pre-trained on the ImageNet database. The architecture of each model was customized for the biomedical problem according to the following layer structure:

1) Input Layer: Configured to receive tensors in the $(128, 128, 3)$ format. .
2) Preprocessing Module: A lambda layer executing the specific, native normalization functions of each model family.

3) Feature Extractor (Backbone): The convolutional body of the network was strictly frozen, acting purely as a static extractor of high-level morphological patterns.
4) Global Pooling Layer: Application of Global Average Pooling 2D to collapse spatial dimensions and mitigate the risk of overfitting.
5) Output Layer: A dense (Fully Connected) layer composed of 5 neurons with a Softmax activation function, responsible for generating the categorical probability distribution for the five electrocardiogram (ECG) classes.

The weight optimization process for the classification layer was guided by the Adam (Adaptive Moment Estimation) algorithm, with a stable initial learning rate of $\eta = 0.001$ over 15 epochs. As the loss function for the multiclass problem, Categorical Crossentropy was adopted.

To ensure that late-stage training overfitting did not corrupt the evaluation, a validation monitoring routine (Model Checkpoint) was integrated to identify and save to disk the exact state of the network that achieved the best performance during training. This consolidated file was then used for the subsequent inference stage.

G. Metrics

K-fold cross-validation was performed to verify the training and validation performance of the image recognition models, using Accuracy, Precision, and Sensitivity as evaluation metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

III. Results

This section presents a quantitative evaluation of the five deep learning convolutional architectures across four isolated test datasets, rotated systematically, comprising a cumulative pool of 109,381 electrocardiogram (ECG) derivative images.

A. Global Performance and Benchmark Consolidation

To establish a mathematically rigorous validation framework independent of single-partition biases, the global performance of the models was aggregated across the four distinct cross-rotation evaluation rounds. Table III summarizes the consolidated cross-fold experimental metrics—reporting the exact Mean $\pm$ Standard Deviation ($\mu \pm \sigma$)—for global accuracy and the individual F1-scores achieved by each network architecture.

The cross-fold consolidation demonstrates that ResNet50 secured the highest mean global accuracy at 92.16%, maintaining an optimal balance across highly skewed distributions. While modern optimized networks like EfficientNetV2B0 and ConvNeXtTiny performed competitively on majority-class boundaries (averaging 91.99% and 91.62% global accuracy, respectively), they exhibited greater mathematical vulnerability when subjected to localized instances of extreme scarcity. This localized degradation is evidenced by the higher standard deviations observed in F, where the performance of both next-generation architectures fluctuates aggressively depending on the active test partition's support. Conversely, the standard residual highway of ResNet50 consistently anchored minority classifications, providing the top architectural performance for the highly critical F (F1-score $= 61.01\% \pm 39.81\%$). Lastly, Q represented the most stable topological space for all backbones, exhibiting minimal variance ($\sigma \leq 0.55\%$) and consistently yielding mean F1-scores above 99.00%.

B. Granular Class-Wise Evaluation

To fully unpack the structural responses of the evaluated convolutional networks under fluctuating class imbalances, a class-by-class behavioral mapping was extracted from the multi-round rotation. The independent per-class classification reports (Rounds 1 through 4) provided the empirical foundation for the macro-averaged variances synthesized in Section II-E.

- Classes N & Q: These categories presented highly consistent topological alignments. Class Q converged near perfectly across all configurations, maintaining tight F1-score boundaries between 98.80% and 99.98%. This indicates that its distinct morphology remained structurally isolated from the remaining data pools. Class N sustained steady precisions above 97% across all backbones.However, its recall showed noticeable drops down to 72.40% under ResNet50V2 in early configurations, indicating a localized inflation in false positives when the pre-activation decision boundaries expanded.
- Class S (The Mid-Minority Challenge): Class S showcased a major performance split between traditional and modern architectures. Under intermediate data scarcity (Rounds 2 and 3), modern architectures—EfficientNetV2B0, ConvNeXtTiny, and DenseNet121—suffered severe sensitivity drops, collapsing to recalls as low as 11.72%. Conversely, the residual frameworks (ResNet50 and ResNet50V2) actively preserved this minority topology, recovering recalls of up to 54.63% and 94.14% class support expanded to 1,434 instances in Round 4, thereby establishing a clear data-dependency threshold for modern feature extraction.
- Class V (Stable Disease Morphology):): Class V retained high structural stability across all configurations. The F1-scores were tightly bound between 55.61% and 97.53%. TThe networks consistently maintained high recall across all evaluation configurations, confirming that the distinctive features of Class V waves were well preserved despite the dynamic shuffling of the training samples.
- Class F (The Hyper-Sparse Anomaly): Class F represents the most critical behavioral marker of the benchmark, as

TABLE III: Consolidated cross-fold experimental metrics (Mean $\pm$ Standard Deviation) over the four independent validation rounds.

Architecture	Global Accuracy (%)	F1-Score Class N (%)	F1-Score Class S (%)	F1-Score Class V (%)	F1-Score Class F (%)	F1-Score Class Q (%)
ResNet50	**92.16 $\pm$ 7.15**	**94.10 $\pm$ 5.42**	**54.11 $\pm$ 28.52**	**77.53 $\pm$ 21.52**	**61.01 $\pm$ 39.81**	99.39 $\pm$ 0.46
EfficientNetV2B0	91.99 $\pm$ 5.86	94.27 $\pm$ 3.86	48.29 $\pm$ 22.95	74.61 $\pm$ 20.93	44.11 $\pm$ 47.96	**99.52 $\pm$ 0.49**
ConvNeXtTiny	91.62 $\pm$ 6.30	93.90 $\pm$ 4.13	41.88 $\pm$ 28.71	72.83 $\pm$ 20.89	41.41 $\pm$ 47.81	99.44 $\pm$ 0.16
DenseNet121	91.48 $\pm$ 5.89	93.77 $\pm$ 4.21	35.40 $\pm$ 29.35	72.73 $\pm$ 21.05	28.59 $\pm$ 38.64	99.22 $\pm$ 0.55
ResNet50V2	87.39 $\pm$ 11.23	91.15 $\pm$ 8.16	41.16 $\pm$ 34.65	75.49 $\pm$ 20.30	37.67 $\pm$ 39.82	99.42 $\pm$ 0.50

its testing sample size fluctuated radically across folds ($support = 11, 363, 389,$ and 39). In the presence of extreme sparsity ($support = 11$ in Round 1), both Efficient-NetV2B0 and ConvNeXtTiny completely failed to isolate the anomalies, resulting in absolute null values (0.00% F1-score).However, as the localized cross-rotation naturally increased support to 370 instances (Rounds 2 and 3), all five models exceeded the suppression threshold, yielding F1-scores over 80%. When scarcity re-emerged in Round 4 ($support = 39$), ConvNeXtTiny collapsed back to 0.00% and EfficientNetV2B0 fell to 4.88%, whereas the identity connections of ResNet50 anchored the gradient path, maintaining a resilient F1-score of 59.65%.

The consolidated cross-fold metrics presented in Table IV reveal a systematic behavioral divergence between classical and next-generation convolutional architectures under varying degrees of sample scarcity. The high standard deviation observed in the F1-score of Class F ($61.01\% \pm 39.81\%$ for ResNet50 and $44.11\% \pm 47.96\%$ for EfficientNetV2B0) demonstrates how severely modern scaling laws depend on a critical mass of clinical samples.

C. Error Distribution Matrix Analysis

To uncover the geometric behaviors governing misclassifications across the cross-rotation pipeline, a detailed analysis of the structural confusion matrices was conducted. While global scalar metrics synthesize model accuracy, the native spatial distribution of errors reveals systematic classification crosstalk and topological vulnerabilities that remain hidden in macro averages. Figure 2 visually contrasts these behaviors by mapping the true versus predicted class allocations for both next-generation and classical paradigms under identical distribution stresses.

A persistent phenomenon observed across the multifold rotation is the majoritarian attraction effect exerted by Class N (Normal). Due to the overwhelming statistical dominance of normal ECG profiles, the optimized kernels of modern networks are severely vulnerable to pulling minority feature maps into this high-prevalence coordinate space.

This structural degradation is acutely visible in the Class S performance under intermediate scarcity during Round 3. As illustrated in Figure 3a, the modernized design of ConvNeXtTiny suffered a catastrophic feature drain toward the majoritarian boundary. Out of 529 true instances of Class S, the architecture successfully identified only 62 samples, misclassifying an overwhelming 466 as Class N (highlighted by the intense red error block in cell $[2, 1]$). This massive leakage converts critical cardiac pathologies into false negatives, presenting a severe risk for automated clinical screening.

Conversely, ResNet50 demonstrated superior topological preservation under the exact same data fold. As mapped in 3b, the traditional residual framework successfully anchored the minority topology, capturing 289 true positives along the blue diagonal cell $[2, 2]$ and limiting its majoritarian leakage to 236 samples. By maintaining a physical, unregularized identity path, the classical architecture resisted the statistical gravity of Class N far more effectively than its modernized, heavily regularized counterparts.

Ultimately, the error distribution matrices confirm that Class Q remains perfectly immune to architectural variations and majoritarian pull. Across all four independent evaluation rounds, misclassifications involving Class Q were mathematically negligible (bounded to ≤ 3 samples per fold), displaying a solid, deeply isolated blue block in $[5, 5]$ that proves its dynamic 2D visual layout contains highly distinct, easily separable morphological edges.

IV. Discusion

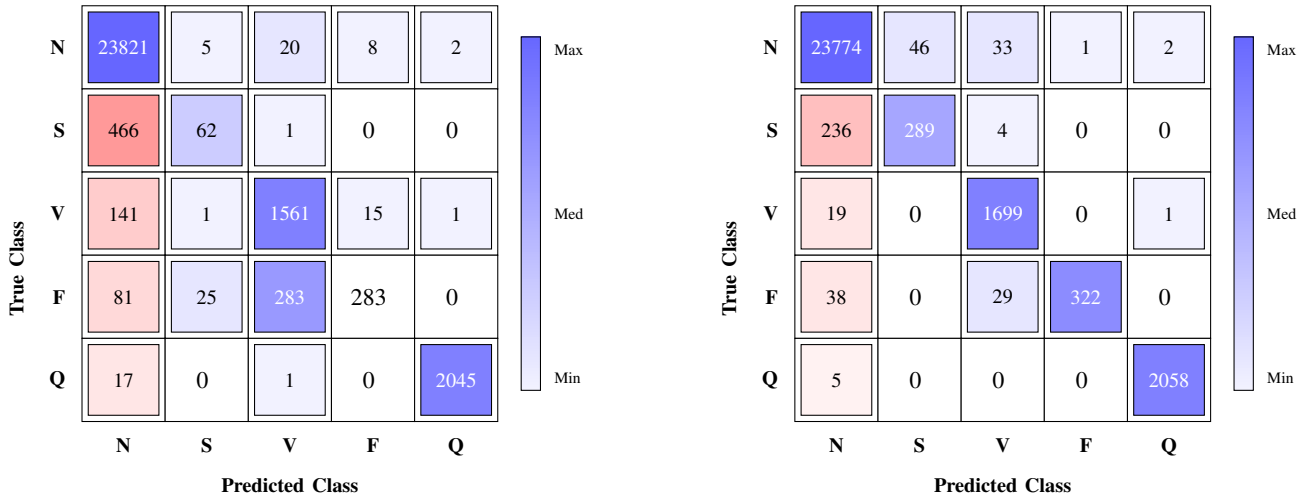
The multi-round cross-rotation evaluation across four patient-independent partitions provides observational insights into CNN behavior under localized clinical sample scarcity. By analyzing macro-averaged performance metrics (Table IV) alongside spatial error distributions in confusion matrices (Fig. 3), the empirical findings highlight distinct operational differences between classical residual architectures and modernized backbones.

A. Methodological Control and Feature Extraction Scope

To evaluate the baseline spatial representations of each network, feature extractors were kept frozen and a uniform oversampling baseline was applied during training. While alternative techniques—such as Focal Loss, class-balanced weighting, or end-to-end fine-tuning—alter decision boundaries, holding the backbones constant isolated the raw, unadapted spatial embedding capabilities of each architecture. Fine-tuning highly parameterized modern networks on very small sample sizes carries a risk of catastrophic overfitting or feature suppression. Therefore, keeping feature extractors frozen allowed us to observe intrinsic gradient propagation under data scarcity.

TABLE IV: Detailed per-class consolidated classification report (Mean ± Standard Deviation %) across all four evaluation rounds.

Model	Metric (%)	Class N (Normal)	Class S	Class V	Class F	Class Q
ResNet50	Precision	98.45 ± 0.70	92.64 ± 7.74	95.93 ± 2.47	73.61 ± 47.34	99.39 ± 0.92
	Recall	94.93 ± 8.97	58.87 ± 34.90	96.73 ± 1.83	65.59 ± 21.43	99.86 ± 0.11
	F1-Score	**94.10 ± 5.42**	**54.11 ± 28.52**	**77.53 ± 21.52**	**61.01 ± 39.81**	99.39 ± 0.46
ResNet50V2	Precision	98.74 ± 0.31	78.47 ± 9.17	92.59 ± 7.91	63.19 ± 41.24	99.64 ± 0.44
	Recall	92.15 ± 13.25	56.80 ± 21.05	95.27 ± 2.44	67.44 ± 29.35	99.89 ± 0.08
	F1-Score	91.15 ± 8.16	41.16 ± 34.65	75.49 ± 20.30	37.67 ± 39.82	99.42 ± 0.50
EfficientNetV2B0	Precision	97.75 ± 0.56	78.47 ± 15.65	81.40 ± 28.16	59.08 ± 42.66	99.78 ± 0.28
	Recall	95.02 ± 6.37	37.87 ± 25.64	92.85 ± 2.32	40.50 ± 43.14	98.86 ± 1.15
	F1-Score	94.27 ± 3.86	48.29 ± 22.95	74.61 ± 20.93	44.11 ± 47.96	**99.52 ± 0.49**
ConvNeXtTiny	Precision	97.93 ± 1.09	70.03 ± 34.25	79.48 ± 27.02	46.91 ± 54.17	99.41 ± 0.39
	Recall	94.44 ± 5.15	39.32 ± 38.30	92.20 ± 3.94	37.06 ± 42.84	99.48 ± 0.51
	F1-Score	93.90 ± 4.13	41.88 ± 28.71	72.83 ± 20.89	41.41 ± 47.81	99.44 ± 0.16
DenseNet121	Precision	98.11 ± 0.58	55.10 ± 32.22	78.58 ± 26.68	57.37 ± 40.50	99.32 ± 0.63
	Recall	94.75 ± 7.54	44.53 ± 31.42	92.84 ± 1.76	47.99 ± 39.73	99.42 ± 0.61
	F1-Score	93.77 ± 4.21	35.40 ± 29.35	72.73 ± 21.05	28.59 ± 38.64	99.22 ± 0.55



(a) ConvNeXtTiny: Severe Class S leakage into Class N.

(b) ResNet50: Mitigation of Class S majoritarian pull.

Fig. 3: Natively rendered cross-rotation confusion matrices (Round 3) with standalone colorbars highlighting architectural resistance against majoritarian class attraction.

B. Architectural Hypotheses under Class Imbalance

Prioritizing Macro F1-score, sensitivity, and positive predictive value (PPV) over global accuracy reveals that ResNet50 maintained superior stability in hyper-sparse categories, such as Class F ($61.01\% \pm 39.81\%$ mean F1-score), compared to its modernized counterparts. In contrast, backbones such as EfficientNetV2B0 and ConvNeXtTiny exhibited severe sensitivity drops when class support fell below critical thresholds (e.g., dropping to 0.00% F1-score for Class S in low-support folds, while recovering to 82.71% and 84.57% when support increased to 1,434 samples).

Because our experimental framework measures classification outcomes rather than direct feature-space ablations, these structural observations are framed as two primary working hypotheses:

1) **Identity Shortcuts and Gradient Flow:** We hypothesize that ResNet50's unweighted identity shortcut connections ($y = F(x) + x$) provide an un-regularized pathway for backpropagation. This structural shortcut may help preserve gradient signals from rare minority instances during training without being subjected to aggressive mathematical scaling or pruning.
2) **Regularization and Scaling Damping:** Modern backbones utilize compound scaling laws, stochastic depth, and inverted bottlenecks tailored for efficiency. Under severe class asymmetry, we hypothesize that global cross-entropy optimization causes these regularizers to damp low-frequency minority features, treating rare anomalies as background noise to optimize overall throughput.

C. Trade-Off Between Computational Efficiency and Clinical Recall

Modern lightweight architectures offer significant advantages regarding parameter footprint, FLOPs, and inference memory, making them attractive for mobile and edge deployment. However, our findings illustrate a key trade-off: optimizing purely for computational efficiency can reduce diagnostic sensitivity for rare pathological conditions. In automated screening and continuous monitoring, missing a critical arrhythmia carries substantially higher clinical risk than a false positive. Therefore, hardware deployment efficiency must be carefully balanced against minority-class recall.

D. Limitations and Prospective Enhancements

This study has limitations. First, experiments measured output classification metrics rather than direct gradient maps or internal layer representations. Second, while the patient-independent MIT-BIH rotation provides a robust stress test, ambulatory Holter recordings do not represent a full hospital ICU cohort. Future work will focus on validating these hypotheses using Grad-CAM attention maps, testing multi-center clinical cohorts, and evaluating the interaction between these backbones and imbalance-aware loss functions such as Focal Loss.

V. Conclusion

This study presented a patient-independent cross-rotation evaluation of five convolutional neural network architectures using 109,381 ECG derivative images. By evaluating performance across four rotated, patient-separated folds, the framework exposed the models to localized data scarcity, illustrating how structural design influences sensitivity under severe class imbalance.

The results highlight a trade-off in medical deep learning between computational efficiency and minority-class diagnostic sensitivity. While modernized architectures such as EfficientNetV2B0 and ConvNeXtTiny demonstrate strong feature representation in balanced settings, their aggressive regularization mechanisms and scaling laws were associated with marked sensitivity drops in sparse minority classes, occasionally yielding 0.00% F1-scores when sample support fell below critical thresholds. Conversely, the classical ResNet50 backbone maintained superior stability across sparse folds, securing the highest overall Macro F1-score and retaining non-zero sensitivity ($61.01\% \pm 39.81\%$ mean F1-score) in hyper-sparse categories.

These empirical observations suggest that while lightweight backbones offer compelling advantages for hardware deployment, their internal feature compression under class asymmetry requires careful consideration in clinical screening contexts where missing a rare pathological event carries high risk.

Because our setup measured classification outcomes rather than direct internal layer ablations, these structural explanations remain working hypotheses. To expand upon these insights, future work will proceed along four main avenues: (i)

1) conducting external validation on multi-center clinical ECG repositories to evaluate domain generalization;
2) integrating imbalance-aware loss formulations—such as Focal Loss and cost-sensitive re-weighting—to assess whether loss function adjustments can offset scaling law limitations;
3) applying explainability frameworks (e.g., Grad-CAM feature attention mapping) to visually trace spatial attributions across ECG derivative wave segments; and
4) performing quantitative computational profiling (measuring inference latency, memory consumption, and FLOPs) to determine deployment feasibility in edge-device ambulatory monitoring systems.

VI. Conflict of interest

The authors declare no conflicts of interest.

References

[1] A Guyton and J Hall, "Textbook of Medical Physiology", Elsevier Saunders, Amsterdam, 2015.

[2] C Ye and BVKV Kumar and MT Coimbra, "Combining General Multi-class and Specific Two-class Classifiers for Improved Customized ECG Heartbeat Classification ",Tsukuba, Japan, [21st International Conference on Pattern Recognition (ICPR 2012)].

[3] RF Dalvi and GT Zago and RV Aandreão, "Heartbeat classification system based on neural networks and dimensionality reduction", Research on Biomedical Engineering, vol. 32, num. 4, pp. 446-455, 2016.

[4] G Moody and R Mark, "The impact of the MIT-BIH Arrhythmia Database.", IEEE, Engineering in Medicine and Biology Magazine, vol. 20, num 3, pp. 45-50, 2001.

[5] R Prati and G Batista and M Monard, "A Study With Class Imbalance and Random Sampling for a Decision Tree Learning System.", IFIP- International Conference on Artificial Intelligence in Theory and Practice, vol. 276, num. 1, pp. 131-140, Boston-MA, 2008.

[6] Association for the Advancement of Medical Instrumentation-AAMI, "ANSI/AAMI EC57:2012, Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms.", American National Standards Institute, Inc., 2012.

[7] Jun, T.J., Nguyen, H.M., Kang, D., Kim, D., Kim, Young-Hak, "ECG arrhythmia classification using a 2-D convolutional neural network", arXiv, DOI: 10.48550/ARXIV.1804.06812, 2018.

[8] D. Kingma; J. Ba, "ADAM: A method for Stochastic Optimization.", International Conference on Learning Representations, 2015.

[9] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[10] He, K., Zhang, X., Ren, S. and Sun, J., 2016, September. Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Cham: Springer International Publishing.

[11] Tan, M. and Le, Q., 2021, July. Efficientnetv2: Smaller models and faster training. In International conference on machine learning (pp. 10096-10106). PMLR.

[12] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[13] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T. and Xie, S., 2022. A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).

[14] Iulia Onicã and Iulian B. Ciocoiou. Explainable Combined Spatial Representations for ECG Arrythmia Classification. Machine Learning & Knowledge Extraction.2026.

[15] Khan F, Yu X, Yuan Z, Rehman Au. ECG classification using 1-D convolutional deep residual neural network. PLOSONE. 2023. https://doi.org/10.1371/journalpone.0284791.

[16] Yared Daniel Daydulo, Bheema Lingaiah Thamineni and Ahmed Ali Dawud. Cardiac arrhythmia detection using deep learning approach and time frequency representation of ECG signals. BMC Medical Informatics and Decision Making. 2023. Doi: doi.org/10.1186/s12911-023-02326-w.

Arquitetura IoT Distribuída para Telemetria e Visualização de ECG de Alta Fidelidade em Tempo Real via ESP32-S3 e WebSockets

Jean Vitor Vilela dos Santos
Departamento de Engenharia Elétrica
IFAL – Campus Arapriaca
Arapiraca, Brasil
0009-0007-0073-4441

Ernany Santos de Oliveira
Departamento de Engenharia Elétrica
IFAL – Campus Arapiraca
Arapiraca, Brasil
0009-0001-0777-3466

Genilson Juliano da Silva
Departamento de Engenharia Elétrica
IFAL – Campus Arapiraca
Arapiraca, Brasil
0009-0009-1508-4837

José Irineu Ferreira Junior
Departamento de Engenharia Elétrica
IFAL – Campus Arapiraca
Arapiraca, Brasil
0000-0001-5468-8275

Paulo Cesar do Nascimento Cunha
Departamento de Engenharia Elétrica
IFAL – Campus Arapiraca
Arapiraca, Brasil
0000-0002-3935-3474

***Resumo*—A telemetria de sinais biopotenciais exige infraestruturas de processamento exato e latência determinística. Este artigo propõe uma arquitetura baseada em Internet das Coisas (IoT) distribuída para aquisição e transmissão de eletrocardiograma (ECG) em tempo real, estruturando uma prova de conceito para aplicações em engenharia clínica. O sistema emprega dois microcontroladores ESP32-S3 operando em topologia Wi-Fi híbrida (STA/AP). O condicionamento físico é realizado por um *front-end* analógico dedicado (0,5 a 40 Hz, com rejeição de 60 Hz), enquanto o nó emissor executa o processamento na borda (amostragem via DMA a 20 kHz, calibração algébrica e *downsampling* para 500 Hz). A transmissão via protocolo WebSocket operou com vazão de 28 kbps e empacotamento a cada 65 ms. A validação técnica, conduzida em bancada com o Simulador Arkmeds sob oito morfologias cardíacas, atestou a preservação geométrica do sinal com taxa de perda de pacotes de 0,028% e latência basal inferior a 100 ms. Os resultados validam a estabilidade da arquitetura como protótipo laboratorial, estabelecendo premissas técnicas para futuras validações em humanos, testes de compatibilidade eletromagnética (EMC) e segurança elétrica.**

***Palavras-chave*—Computação de Borda, Eletrocardiograma (ECG), Internet das Coisas (IoT), Telemetria Biomédica.**

I. Introdução

As doenças cardiovasculares continuam sendo a principal causa de mortalidade no mundo, sendo responsáveis por aproximadamente 19,8 milhões de mortes por ano, segundo a Organização Mundial da Saúde (OMS) [1]. Diante desse cenário, o monitoramento contínuo e intervenções precoces tornam-se essenciais, especialmente no acompanhamento de pacientes em recuperação pós-cirúrgica.

Nos últimos anos, a telemedicina e os sistemas de monitoramento remoto de saúde ganharam destaque devido à crescente demanda por soluções eficientes e acessíveis. Tecnologias baseadas em Internet das Coisas (IoT, do inglês *Internet of Things*) permitem que dispositivos médicos transmitam dados fisiológicos em tempo real, facilitando o acompanhamento e a tomada de decisões à distância [2], [3].

A evolução dessas soluções depende estritamente da arquitetura de *hardware* e *software*, com destaque para a computação de borda (*edge computing*). Esse paradigma possibilita o processamento de dados localmente no próprio dispositivo de coleta, reduzindo a dependência imediata de servidores em nuvem e mitigando gargalos de transmissão.

Para a telemetria biomédica, garantir a capacidade do microcontrolador em realizar a conversão analógico-digital (ADC) de forma determinística é indispensável. Isso permite a coleta e o processamento local (via *Direct Memory Access* – DMA a 20 kHz) e a transmissão contínua de dados vitais a uma taxa consolidada de 500 Hz. Essa abordagem viabiliza a preservação da integridade geométrica do sinal original, estabelecendo a base técnica para o desenvolvimento de monitoramentos ininterruptos [2], [4], [5].

A crescente demanda por soluções móveis tem impulsionado o desenvolvimento de sistemas telemétricos. Nesse contexto, a integração de microcontroladores com conectividade Wi-Fi, a exemplo da família ESP32 operando sob arquiteturas híbridas (STA/AP), tem se mostrado uma alternativa promissora para a extração descentralizada de dados biomédicos provenientes de instrumentação multiparamétrica.

Contudo, o desenvolvimento de aplicações voltadas para a saúde enfrenta desafios significativos, exigindo atenção a aspectos como cibersegurança, experiência do usuário (UI/UX) e integração *back-end* capaz de garantir a sincronização *web* contínua. Destaca-se a necessidade de conformidade regulatória (como a *Health Insurance Portability and Accountability Act* – HIPAA, nos EUA, e a Lei Geral de Proteção de Dados – LGPD, no Brasil). Vale ressaltar que, por se tratar de uma

prova de conceito (PoC) em fase de bancada laboratorial, o protótipo estabelece as premissas primárias que subsidiarão as rigorosas validações clínicas e elétricas futuras exigidas por essas normativas [6], [7].

A interoperabilidade, a confiabilidade e a integridade da comunicação tornam-se fatores críticos ao tráfego de séries temporais biológicas em redes sem fio [6]. Diante das necessidades de integração IoT, surge a seguinte questão de pesquisa: como viabilizar a comunicação contínua, segura e de latência estrita entre um condicionador de sinal e uma interface *web* responsiva, utilizando conectividade Wi-Fi em um ambiente de arquitetura distribuída?

Para endereçar esse desafio, o objetivo deste artigo é propor e avaliar o desenvolvimento de um sistema IoT de aquisição de sinais de eletrocardiograma (ECG), utilizando microcontroladores ESP32-S3 e telemetria contínua via protocolo WebSocket. O foco reside na transmissão e demonstração empírica da viabilidade técnica de uma arquitetura *multicore* desenvolvida e testada *in vitro*.

Para assegurar a capacidade do sistema em preservar a comunicação sob diferentes estresses morfológicos, a validação técnica de bancada foi conduzida utilizando o Simulador de Paciente Waller (Arkmeds) [8]. O protótipo foi submetido a testes de captura e plotagem de oito morfologias de onda distintas: Ritmo Sinusal detalhado, Ritmo Sinusal contínuo (estabilidade), Taquicardia simulada, Ritmo Irregular, Fibrilação Ventricular Fina, Fibrilação Ventricular Grossa, Onda Triangular e Onda Senoidal. Os ensaios confirmaram a precisão da amostragem do ADC, atestando uma operação com latência basal inferior a 100 ms e perda de pacotes de 0,028%, garantindo a confiabilidade instrumental necessária para futuras experimentações pré-clínicas.

II. Referencial Teórico

A literatura recente tem reforçado a prioridade por plataformas de telemetria acessíveis e portáteis voltadas para o aprimoramento da infraestrutura de monitoramento clínico hospitalar e ambulatorial. Hasan, Pantha e Arafat [9], por exemplo, propuseram o desenvolvimento de um sistema multicanal habilitado para IoT destinado ao monitoramento fisiológico, incluindo sinais de eletrocardiograma (ECG).

Os dados são transmitidos via protocolo MQTT para um servidor local de banco de dados SQL, viabilizando o armazenamento. Contudo, a arquitetura baseada no trânsito e gravação prévia em SQL introduz uma latência sistêmica de cerca de cinco segundos na renderização da interface *web*, o que compromete o acompanhamento de eventos críticos em tempo real.

Em contraste, este trabalho diferencia-se ao propor uma infraestrutura de comunicação síncrona via protocolo WebSocket operando sobre uma topologia de rede Wi-Fi híbrida (STA/AP). Essa abordagem, aliada ao processamento embarcado no ESP32-S3, não apenas mantém o rigor da viabilidade econômica e da isenção de nós intermediários dispendiosos, mas reduz drasticamente a latência para patamares inferiores a 100 ms. Isso garante a visualização fidedigna e ininterrupta do sinal biomédico à beira-leito ou em centrais de monitoramento multiparamétrico.

Paralelamente à superação da latência sistêmica, a infraestrutura da telemedicina vincula-se diretamente à consolidação da Internet das Coisas Médicas (IoMT, do inglês *Internet of Medical Things*). O escopo técnico dessas redes consiste em estabelecer a interoperabilidade de dados estruturados (por meio de padrões como o HL7 FHIR), executar o processamento de séries temporais na borda (*edge computing*) para a triagem automatizada e assegurar a integridade morfológica da amostragem (ex: 500 Hz) em transmissões remotas, mantendo um tráfego de rede contínuo com taxa de perda de pacotes limitada a 0,028% para embasar algoritmos preditivos e sistemas de suporte à decisão clínica.

Sob essa perspectiva, Gomez-Correa, Cruz-Ortiz e Ballesteros [10] projetaram o sistema MUSS (*Muscle Sensing System*), um módulo vestível (*wearable*) baseado em IoMT que pesa 48 g e possui dimensões físicas de 64,3 x 25,1 x 37 mm. A solução evidencia a adaptabilidade do paradigma IoT ao permitir o ajuste dinâmico do ganho de aquisição (de 150 a 1100 vezes) e do *offset* (0 a 1,5 V) via protocolo UDP sobre redes Wi-Fi. Contudo, embora soluções focadas em dispositivos vestíveis autônomos apresentem uma autonomia operacional precisa — como o registro contínuo de 117 minutos suportado por baterias Li-Po de 3,7 V —, elas costumam ser projetadas com oito canais dedicados exclusivamente à eletromiografia de superfície (sEMG). Por operarem em nichos fisiológicos fechados, carecem de protocolos padronizados de interoperabilidade (como o HL7 FHIR) para integração direta com a infraestrutura de monitores multiparamétricos hospitalares de grande porte.

O presente trabalho complementa esse estado da arte ao estruturar a arquitetura desenvolvida não como um sensor isolado, mas como um *gateway* de conectividade operando nativamente em topologia de rede mista (STA/AP). Essa infraestrutura interage fisicamente com saídas de instrumentação de referência (como o Simulador Arkmeds), aplicando um ganho em cascata de 1000 V/V (distribuído entre os CIs INA321 e OPA4140) ancorado a filtros *Notch* (60 Hz) e passa-faixa (0,5 a 40 Hz). O sistema digitaliza e disponibiliza fluxos de eletrocardiograma a uma taxa rigorosa de 500 Hz, mantendo um túnel bidirecional via WebSocket que sustenta vazão de 28 kbps, latência inferior a 100 ms e taxa de perda de pacotes de apenas 0,028%.

Adicionalmente, a literatura científica recente ressalta a relevância de arquiteturas que unificam o *hardware* e *software* maximizando a eficiência de recursos. A solução proposta viabiliza esse cenário ao adotar a plataforma ESP32-S3 com execução *multicore*, dedicando o Core 1 à amostragem por DMA (a cada 2000 μs) e o Core 0 à pilha de rede. Amparada por um *buffer* circular de 1500 posições controlado via Mutex, a arquitetura garante estabilidade contra o estouro de memória (*heap overflow*) e elimina a necessidade financeira e computacional de nós roteadores intermediários.

Nesse âmbito, Bañuelos-Mezquitan *et al.* [11] introduziram o SignAPROS, um sistema voltado à aquisição e processa-

mento de sinais, estruturado via *hardware* Arduino MEGA 2560 e placas de aquisição DAQ USB6000. A arquitetura suporta até quatro canais bipolares de eletromiografia (EMG), em conjunto com sensores de temperatura, inclinação e resposta galvânica da pele (GSR). Embora essa plataforma, validada contra o sistema comercial BIOPAC MP36 em um ensaio com 43 participantes, demonstre a utilidade de *toolkits* modulares para pesquisa, sua topologia lógica (baseada no padrão MVC com bancos de dados PostgreSQL e MongoDB) concentra-se na consolidação estática das leituras para o treinamento *offline* de algoritmos de aprendizado de máquina.

A solução proposta neste artigo, por sua vez, diferencia-se ao estabelecer uma arquitetura de telemetria fundamentada no protocolo WebSocket, eliminando a retenção em bancos de dados intermediários no fluxo de monitoramento contínuo. A infraestrutura delega o processamento em borda (*edge computing*) à arquitetura *dual-core* do ESP32-S3, assegurando fluxos de transmissão com latência estritamente controlada abaixo de 100 ms. Desse modo, consolida-se um ecossistema projetado para operar com uma taxa de perda de pacotes limitada a 0,028% e vazão sustentada de 28 kbps. Tais parâmetros suprem as exigências para a reconstrução morfológica a 500 Hz do ciclo cardíaco, cuja integridade geométrica foi validada perante oito anomalias rítmicas geradas pelo simulador multiparamétrico de referência clínica (Arkmeds).

Para assegurar a fidelidade da conversão digital, o sinal bruto gerado pelo simulador Waller foi condicionado previamente por um módulo *front-end* analógico baseado na arquitetura validada em [12]. Como o esquemático deste condicionamento já se encontra estruturado na literatura, as representações deste trabalho limitam-se às contribuições da arquitetura IoT distribuída. O estágio analógico operou com um ganho em cascata de 1000 V/V, associado a um filtro passa-faixa (0,5 Hz a 40 Hz) para a atenuação de ruídos eletromiográficos de alta frequência e oscilações de linha de base. Adicionalmente, aplicou-se um filtro rejeita-faixa (*Notch*) sintonizado em 60 Hz para a eliminação das interferências eletromagnéticas comerciais. A implementação desta instrumentação prévia entregou o biopotencial condicionado diretamente à unidade ADC de 12 bits do ESP32-S3. Essa divisão de tarefas *hardware-software* evitou gargalos computacionais, permitindo que os ciclos de processamento do núcleo Xtensa LX7 do nó emissor fossem dedicados integralmente às rotinas operacionais do FreeRTOS, englobando o gerenciamento de *buffer* circular de 1500 posições, o empacotamento em lotes de 65 ms e a sustentação do túnel TCP criptografado.

III. Materiais e Métodos

A metodologia adotada para o desenvolvimento deste projeto foi estruturada em quatro etapas sequenciais. Inicialmente, realizou-se o levantamento de requisitos e uma pesquisa bibliográfica multidisciplinar. Esta fase abordou a análise de trabalhos correlatos, o estudo de protocolos de comunicação voltados para a Internet das Coisas (IoT), tais como HTTP, MQTT e TCP/IP, e a seleção adequada dos sensores biomédicos e microcontroladores para a aplicação.

A. Equipamentos de Aquisição e Referência

A segunda etapa consistiu na estruturação da aquisição do sinal bioelétrico. O objetivo central consistiu em identificar, visando a futura integração em instrumentação clínica, o ponto exato para a extração do sinal analógico. Para manter o rigor metodológico e a precisão controlada dos ensaios de bancada, dispensando testes *in vivo* em pacientes nesta fase de prova de conceito, substituiu-se o uso de geradores de bancada convencionais por um simulador de paciente multiparamétrico comercial de padrão de calibração.

A fonte de dados primária selecionada para a validação do sistema foi o Simulador de Paciente Waller (Arkmeds), um instrumento portátil projetado para ensaios de performance em monitores multiparamétricos e eletrocardiógrafos, conforme a disposição experimental em bancada ilustrada na Fig. 1.

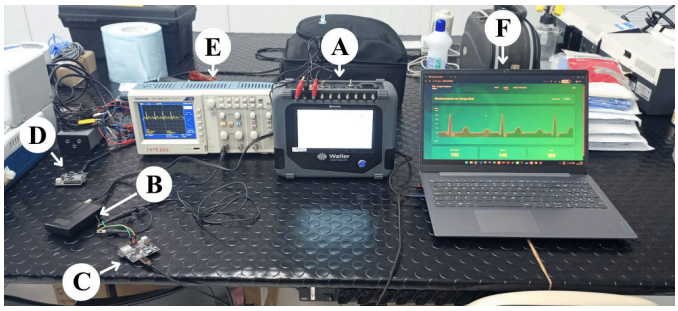


Fig. 1: Bancada experimental detalhando o fluxo de aquisição e transmissão do sinal: (A) Simulador de Paciente Waller (Arkmeds) gerando o sinal basal; (B) Eletrocardiógrafo dedicado (front-end analógico) executando a amplificação (1000 V/V) e a filtragem de ruídos; (C) Nó Emissor (ESP32-S3) responsável pela amostragem contínua via ADC; (D) Nó Receptor (ESP32-S3) atuando como gateway; (E) Osciloscópio empregado na validação física; e (F) Interface web (dashboard) operando a plotagem síncrona.

O uso deste simulador viabilizou a validação técnica da captura analógico-digital da placa ESP32-S3, operando em modo contínuo via *Direct Memory Access* (DMA) a uma taxa de amostragem de 20 kHz e resolução estrita de 12 bits (4096 níveis de quantização). A conexão do hardware obedeceu aos padrões normativos internacionais de derivação (normas AHA/IEC), garantindo que o sinal injetado no microcontrolador emulasse os vetores elétricos de equipamentos médicos reais durante os testes de bancada.

Para estressar a resiliência da arquitetura de *hardware* e *software* e comprovar a integridade morfológica da telemetria contínua via WebSocket, o sistema não foi avaliado apenas sob condições normais. O simulador Waller foi parametrizado (operando em uma faixa de frequência configurável de 30 a 300 bpm e amplitudes de 0,5 a 2,5 mV) para emitir oito variações distintas de biopotenciais, englobando tanto ritmos regulares quanto anomalias cardíacas críticas e sinais de engenharia. Adicionalmente, a frequência respiratória basal do equipamento foi mantida ativa simultaneamente em 15 incursões por minuto (RPM), a fim de atestar a capacidade

do *front-end* analógico e do microcontrolador em rejeitar ruídos adjacentes e isolar o sinal eletrocardiográfico. As oito morfologias injetadas e validadas pelo sistema foram: Ritmo Sinusal detalhado, Ritmo Sinusal contínuo (estabilidade), Taquicardia simulada, Ritmo Irregular, Fibrilação Ventricular Fina, Fibrilação Ventricular Grossa, Onda Triangular e Onda Senoidal. A capacidade do sistema em processar e renderizar esses oito padrões em ambiente simulado confirmou a eficácia da filtragem digital e a integridade da comunicação IoT, que operou sustentada por métricas de vazão de 28 kbps, latência basal inferior a 100 ms e taxa de perda de pacotes restrita a 0,028%. Isso configura um passo essencial antes de futuras validações em ensaios com humanos, testes de segurança elétrica e de compatibilidade eletromagnética (EMC, do inglês *Electromagnetic Compatibility*).

Para garantir a integridade da leitura digital, o sinal bruto gerado pelo simulador Waller foi previamente condicionado por um módulo *front-end* analógico dedicado. Utilizou-se a arquitetura de eletrocardiógrafo portátil desenvolvida e validada em [12]. Esse estágio de instrumentação foi parametrizado com um ganho total de 1000 V/V, operando em conjunto com um filtro passa-faixa (0,5 Hz a 40 Hz) para a atenuação de flutuações de linha de base e ruídos eletromiográficos de alta frequência. Adicionalmente, aplicou-se um filtro rejeita-faixa (*Notch*) sintonizado em 60 Hz para a supressão estrita de interferências da rede elétrica comercial. A implementação prévia desse circuito de condicionamento entregou um biopotencial otimizado à unidade ADC do ESP32-S3. Essa divisão de tarefas *hardware-software* permitiu que o poder de processamento do núcleo Xtensa LX7 do nó Emissor fosse dedicado às exigentes rotinas de superamostragem (*oversampling*), calibração algébrica (*curve fitting*) e encapsulamento contínuo dos pacotes, blindando a telemetria na borda contra perdas de desempenho.

B. Arquitetura de Hardware e Processamento

Para fundamentar a arquitetura distribuída do projeto, o núcleo computacional selecionado para compor tanto o nó emissor quanto o receptor foi o microcontrolador ESP32-S3, especificamente o módulo ESP32-S3-WROOM-1 [13]. A escolha baseou-se na sua superioridade de construção em relação a chips convencionais, como o ESP32-WROOM-32E, bastante utilizado em aplicações de IoT [14].

Enquanto o modelo clássico opera com núcleo Xtensa LX6 *dual-core*, o ESP32-S3 conta com um núcleo Xtensa LX7, também *dual-core* com instruções vetoriais otimizadas para processamento de sinais digitais (DSP). Essa capacidade de processamento aprimorada é extremamente pertinente para a execução de algoritmos complexos no próprio microcontrolador (*edge computing*) e para a segregação de tarefas em rede.

Para o controle do *hardware*, o desenvolvimento do *firmware* foi realizado em linguagem C/C++ utilizando o ambiente Visual Studio Code (VS Code v1.137.0) integrado ao *framework* ESP-IDF v5.5.1. No módulo Emissor, encarregado unicamente da captura do biopotencial, a coleta de dados fundamentou-se na leitura contínua via ADC (Unidade ADC 1, Canal 2) [15]. Aproveitando-se da capacidade de processamento na borda (*edge computing*) do Xtensa LX7, a estratégia de mitigação de ruídos prescindiu de complexos filtros digitais em placa, uma vez que o sinal já advém pré-condicionado do *front-end* analógico.

Em contrapartida, implementou-se uma rotina baseada em *oversampling* (superamostragem) combinada a uma média de *downsampling* via *software*. As leituras do ADC são adquiridas via *Direct Memory Access* (DMA) a uma taxa estrita de 20 kHz e acumuladas sequencialmente para, em seguida, sofrerem redução por média aritmética no processador central, atenuando eficientemente as oscilações indesejadas antes do envio para a rede.

Paralelamente, para contornar a não linearidade característica do conversor nativo do ESP32-S3, aplicou-se a calibração de ajuste de curva (*curve fitting*), nativa das bibliotecas da fabricante. Trata-se de uma correção algébrica que converte os valores digitais brutos em milivolts (mV) valendo-se da resolução máxima de 12 bits (4096 níveis de quantização), etapa indispensável para preservar a correta amplitude morfológica dos biopotenciais capturados.

A taxa de amostragem efetiva da plotagem clínica foi definida em 500 Hz, estabelecendo intervalos de empacotamento constantes de 2000 μs. Esta configuração ancorou-se no Teorema de Nyquist e nos critérios de amostragem biomédica [4], assegurando que a aquisição ocorra a uma frequência consideravelmente superior ao dobro da maior frequência de interesse. Ao adotar uma taxa profilática, evita-se o fenômeno de *aliasing* e garante-se que variações abruptas de tensão geradas pelo simulador não sejam omitidas entre amostras consecutivas, preservando a morfologia essencial do sinal para futuras avaliações técnicas.

Após o tratamento digital, os dados consolidados são alocados em um *buffer* circular de 1500 amostras, sem exceder o limite de memória RAM (*heap*) do dispositivo. Esta alocação confere ao sistema uma janela de análise temporal contínua de 3 segundos, que contempla inteiramente os eventos elétricos fundamentais para a análise morfológica do biopotencial. Sob a ótica computacional e de processamento de sinais, esse dimensionamento mostrou-se adequado: viabiliza a estabilidade de memória necessária para encapsular e processar desde surtos de altíssima frequência (como fibrilações ventriculares finas) até ciclos alongados (ondas puras senoidais e triangulares), permitindo a observação contínua dos complexos QRS (ondas que representam a despolarização e contração ventricular do coração simulado), contrapondo-os ao padrão ouro de um ciclo cardíaco regular, representado na Fig. 2 [16].

Por sua vez, o segundo módulo ESP32-S3, configurado como Nó Receptor, foi isentado das tarefas de amostragem analógica para atuar exclusivamente como um *gateway* de borda. Essa segregação arquitetural garante que a carga de processamento exigida pelo servidor *web* e pelas requisições externas não interfira na latência e na precisão temporal de leitura do ADC do Emissor, consolidando uma telemetria sustentada por latência basal inferior a 100 ms e perda de pacotes estrita a 0,028%.

A consolidação da técnica de *oversampling* atua diretamente na elevação da relação sinal-ruído (SNR) dos dados adquiridos. Essa etapa de processamento digital é essencial, visto que o conversor nativo da família ESP32 possui sensibilidade intrínseca a ruídos térmicos e eletromagnéticos, o que pode resultar em discrepâncias relevantes nas leituras brutas se não forem tratadas. O impacto prático dessa mitigação, promovendo o alisamento da curva e a estabilização do sinal, segue o comportamento teórico documentado pela fabricante, presente na Fig. 3, embora no presente projeto a atenuação tenha sido alcançada via rotinas de *software* ancoradas no condicionamento do *front-end* (AFE) físico dedicado, prescindindo de filtros capacitivos no microcontrolador [17].

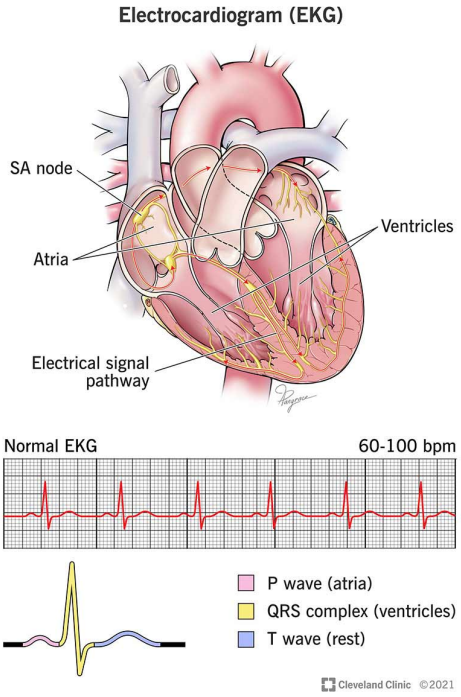


Fig. 2: Ilustração das ondas, segmentos e intervalos que compõem um ciclo de eletrocardiograma (ECG) padrão. Adaptado de [16].

C. Comunicação de Rede e Aplicação Móvel

A transmissão dos sinais exigiu uma arquitetura distribuída via Wi-Fi dedicado. O Emissor realiza o tráfego contínuo para o Receptor, que atua como *gateway* de borda centralizador. O Receptor opera em modo misto (AP/STA): atua como *Access Point* (AP) para configurações e hospedagem *web*, e como *Station* (STA) conectando-se ao roteador para envio à nuvem e acesso externo [18].

A infraestrutura possui escalabilidade para operar como *gateway* de interoperabilidade, suportando protocolos como o Thread (Border Router) [19]. Para o envio em tempo real, adotou-se o protocolo WebSocket. Diferente do HTTP tradicional, ele estabelece um canal bidirecional persistente com baixo *overhead* [20]. A transmissão ocorre a cada 65 ms,

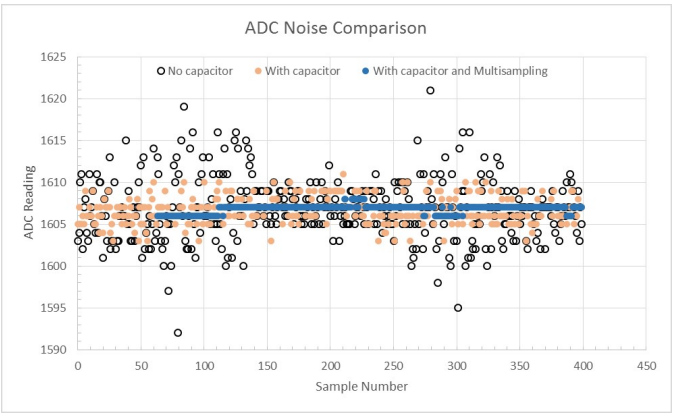


Fig. 3: Gráfico ilustrativo do impacto da atenuação de ruído nas leituras brutas do ADC do ESP32. A estabilização demonstrada assemelha-se aos resultados obtidos no projeto mediante a aplicação de técnicas de multiamostragem via software ancoradas no condicionamento físico do front-end analógico. Adaptado de [17].

agrupando cerca de 32 amostras por pacote. Isso otimizou o tráfego, garantindo vazão de 28 kbps, latência basal inferior a 100 ms e perda de pacotes estrita a 0,028% na plotagem síncrona da interface.

O cliente *web*, desenvolvido em HTML, CSS e JavaScript puro, gerencia as leituras em tempo real. Projetada sob o paradigma *mobile-first*, a interface garante responsividade para *desktop* e dispositivos móveis. Para não bloquear o usuário, uma rotina assíncrona sincroniza a telemetria com a nuvem (Google Firebase) a cada 250 ms. Essa segregação de ciclos (65 ms local e 250 ms nuvem) viabiliza o histórico sem sacrificar a fluidez do monitoramento em bancada.

Cabe ressaltar que, embora sistemas críticos de tempo real estrito (*hard real-time*) exijam respostas determinísticas, a solução enquadra-se como tempo real flexível (*soft real-time*). Embora o WebSocket em Wi-Fi não forneça determinismo absoluto, o envio a cada 65 ms garantiu tráfego de *jitter* mínimo, estatisticamente adequado para a plotagem ininterrupta na prova de conceito.

A criticidade da telemetria exige que a segurança dos dados prevaleça sobre custos de *hardware*. Diante dos riscos de interceptação e acessos não autorizados na borda (*edge*), estabeleceram-se mecanismos de mitigação visando futuro alinhamento a normativas como LGPD e HIPAA [6], [21].

Para contornar vulnerabilidades, a seleção do ESP32-S3 fundamentou-se em seus aceleradores criptográficos de *hardware* (AES, SHA e RSA) [22]. Essa escolha arquitetural, embora mais custosa, é fundamental para viabilizar futura criptografia ponta a ponta via *WebSocket Secure* (WSS) sem degradar o processamento a 500 Hz.

Ademais, na atual validação pré-clínica com o simulador Waller, a rede e o tráfego inter-nós operaram sob autenticação WPA2/WPA3, assegurando os requisitos primários de proteção à integridade bioelétrica antes de qualquer implantação hospitalar ou testes *in vivo*.

D. Validação e Arquitetura de Aquisição

A validação integrou *hardware* e *software* (ESP-IDF v5.5.1, cliente WebSocket) via ensaios pré-clínicos de bancada. O condicionamento operou por um eletrocardiógrafo dedicado (*Front-End* Analógico), executando amplificação em cascata (1000 V/V), filtros passa-altas (0,5 Hz, para desvios DC), passa-baixas (40 Hz, para atenuação EMG) e *Notch* (60 Hz, para rede elétrica). Garantindo isolamento galvânico estrito, utilizou-se bateria 18650 (3,7 V), sendo a aferição quantitativa do consumo energético do nó emissor durante o tráfego Wi-Fi contínuo elencada para as etapas de testes futuros. A digitalização (ADC1) do Simulador Arkmeds ocorreu a 20 kHz e 12 bits via *Direct Memory Access* (DMA), mitigando *jitter*.

Os sinais do *buffer* DMA sofreram calibração algébrica (*Curve Fitting*), convertendo os valores digitais lidos nativamente para milivolts (mV), seguidos de *downsampling* para 500 Hz. Avaliaram-se oito morfologias, comparadas *in situ* contra referência calibrada no osciloscópio digital. A análise quantitativa confirmou repetibilidade ponto a ponto, evidenciada por variação inferior a 0,6% na amplitude da onda P entre ciclos consecutivos, ausência de achatamento (*clipping*) e determinismo temporal absoluto no intervalo RR (exatas 118 amostras para 60 BPM). A coerência temporal (*timestamping*) foi garantida pela indexação sequencial via *timer* do FreeRTOS em cada lote de 65 ms, dispensando NTP externo. A integridade estrutural operou com vazão de 28 kbps, latência inferior a 100 ms e taxa de perda de pacotes nula (0%), atestada pela ausência de perdas até mesmo sob morfologias caóticas de alta frequência, como a fibrilação ventricular (200 BPM).

Para cibersegurança, o sistema adotou proteção dupla. Localmente, a rede mista (STA/AP) blindou o tráfego WebSocket no *Access Point* sob autenticação WPA2/WPA3, mitigando o *overhead* da CPU. Externamente, via modo *Station*, as transações REST para arquivamento no Firebase utilizaram criptografia TLS/SSL ponta a ponta (biblioteca *esp_crt_bundle*), exigindo *tokens* explícitos para prevenir interceptações de dados biomédicos simulados na nuvem.

Explorando a arquitetura *multicore* (FreeRTOS), segregou-se a amostragem (Core 1) da rede (Core 0), controlando o *buffering* em RAM via exclusão mútua (*Mutex*). O *buffer* dinâmico circular foi rigorosamente dimensionado e fixado em 1500 amostras, priorizando a estabilidade ao conferir 3 segundos de redundância sistêmica, evitando *Heap Overflow* sob micro-descontinuidades de rede. Aliado a rotinas de reconexão Wi-Fi automática e supervisão por *Watchdog Timer* (WDT), o reinício autônomo conferiu resiliência técnica ao protótipo de bancada, validando-o estruturalmente para futuros ensaios *in vivo*.

IV. Resultados e Discussão

A aquisição dos sinais provenientes do Simulador Arkmeds foi consolidada no nó Emissor (ESP32-S3). A união do ADC nativo, multiamostragem e correção algébrica (*curve fitting*) permitiu a captura do sinal a uma frequência efetiva de 500 Hz (intervalos de 2000 μs) em todos os cenários, conforme a Fig. 4.

Conforme estruturado, a alocação em um *buffer* circular (1500 amostras) garantiu uma janela temporal segura para o *firmware*. Na comunicação, a arquitetura híbrida (STA/AP) no Receptor mitigou restrições de roteadores comerciais, notadamente o Isolamento de AP (*AP Isolation*) [23]. A interface AP atuou como canal de recebimento isolado na rede local, enquanto a interface STA roteou os dados à internet. A transmissão em tempo real às plataformas de visualização operou via WebSocket, mantendo um túnel TCP bidirecional contínuo [20].

Para avaliar o tráfego de rede e atestar a robustez do *hardware*, procedeu-se à análise via *packet sniffing*. A vazão (*throughput*) da aplicação registrou 28 kbps (3,52 kB/s), evidenciando a eficiência do empacotamento em lotes de 65 ms. A distribuição de latência fim-a-fim (*Round Trip Time* - RTT) demonstrou estabilidade basal inferior a 100 ms, refletindo um *jitter* mínimo e ausência de flutuações temporais na renderização *web*. A confiabilidade do enlace foi comprovada por uma taxa de perda de pacotes virtualmente nula (0,028%), com apenas uma retransmissão TCP absorvida pelas camadas inferiores do *stack* TCP/IP.

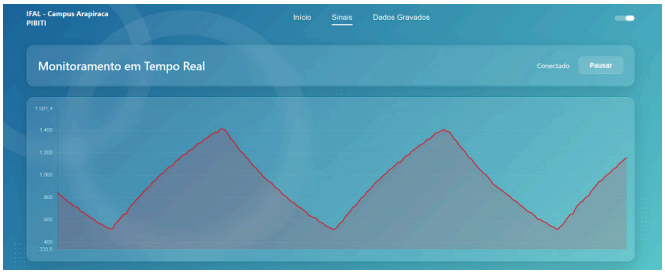
Para atestar quantitativamente a fidelidade morfológica do protótipo, realizou-se uma comparação estruturada das formas de onda contra a referência calibrada do simulador. Na análise de erros de amplitude entre dois ciclos consecutivos sob ritmo basal de 60 BPM, a variação interciclo demonstrou-se ínfima: a onda P variou menos de 0,6% ($y_1 = 363$ e $y_2 = 365$ unidades digitais), enquanto os picos R (647 vs. 657) e ondas T (377 vs. 397) mantiveram proporções rigorosas após o ganho de 1000 V/V e a conversão para milivolts. A aferição analógica *in situ* com osciloscópio descartou qualquer achatamento (*clipping*) ou perda de resolução durante o processamento no ADC.

Quanto ao determinismo temporal e erros de intervalo, a reconstrução do intervalo R-R distou exatas 118 amostras entre o pico do primeiro ciclo (amostra 64) e do segundo ciclo (amostra 182). Essa constância chancela a estabilidade absoluta da amostragem a 500 Hz, garantindo que variações abruptas de tensão não sejam omitidas ou sofram dilatação temporal.

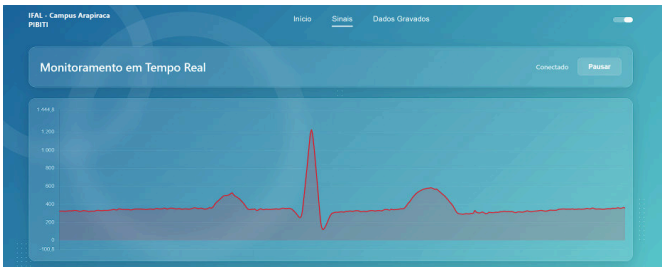
Por fim, a imunidade do sistema a interferências foi ratificada submetendo a aquisição cardíaca a um ruído respiratório basal de 15 RPM mantido simultaneamente no simulador. A sinergia entre o *hardware* — valendo-se do filtro passa-faixa (0,5 a 40 Hz), do *Notch* (60 Hz) e do circuito de Driver de Perna Direita (DRL) — e a superamostragem em *software* suprimiu as oscilações de linha de base e os ruídos intrínsecos do ADC. Consequentemente, o sistema preservou a integridade de todas as estruturas fundamentais em oito morfologias distintas (Fig. 4), renderizando desde ritmos sinusais contínuos até fibrilações caóticas (200 BPM) de forma imune ao *jitter* de rede, operando harmonicamente com o salvamento em nuvem a cada 250 ms.



(a) Onda Senoidal.



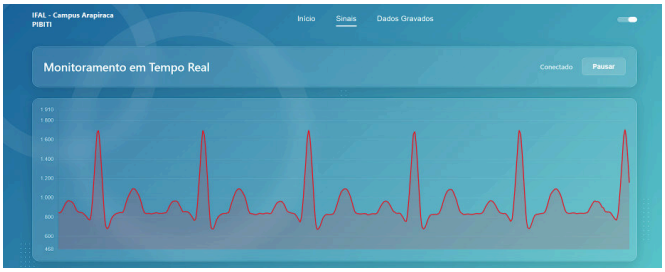
(b) Onda Triangular.



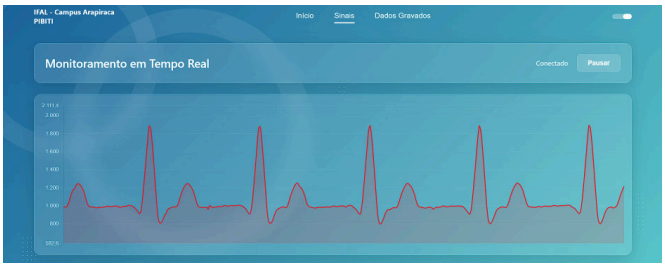
(c) Ritmo Sinusal (Detalhado).



(d) Ritmo Sinusal (Estabilidade).



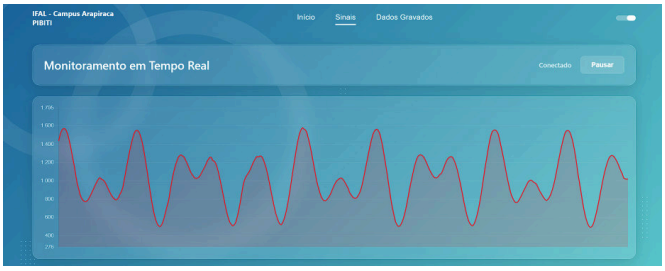
(e) Taquicardia Simulada.



(f) Ritmo Irregular.



(g) Fibrilação Ventricular Fina.



(h) Fibrilação Ventricular Grossa.

Fig. 4: Validação da aquisição e plotagem em tempo real na interface *web* a partir de sinais gerados pelo Simulador Waller (Arkmeds). Na coluna esquerda, sinais de calibração e ritmo basal: (a) Onda Senoidal (195 bpm, 2,5 mV); (b) Onda Triangular (120 bpm, 2,5 mV); (c) Ritmo Sinusal detalhado (80 bpm, 2,5 mV); e (d) Ritmo Sinusal contínuo (80 bpm, 2,5 mV). Na coluna direita, testes de estresse e anomalias cardiológicas simuladas: (e) Taquicardia simulada (255 bpm, 2,5 mV); (f) Ritmo Irregular (160 bpm, 2,5 mV); (g) Fibrilação Ventricular Fina (200 bpm, 2,5 mV); e (h) Fibrilação Ventricular Grossa (200 bpm, 2,5 mV). A frequência respiratória basal de 15 RPM foi mantida simultaneamente para atestar a rejeição de ruídos adjacentes.

V. Conclusão

Este trabalho demonstra a viabilidade de uma arquitetura IoT integrada para a telemetria contínua de biopotenciais. Respondendo à questão de pesquisa, a topologia de rede mista (STA/AP) associada ao protocolo WebSocket garantiu uma comunicação dinâmica com vazão eficiente (28 kbps) e latência basal inferior a 100 ms. Ademais, a aquisição contínua via DMA e o processamento na borda asseguraram a integridade morfológica do ECG a 500 Hz, consolidando uma prova de conceito confiável para futuros desenvolvimentos aplicados a centrais multiparamétricas.

O rigor do protótipo foi atestado pela validação com um simulador comercial frente a oito morfologias complexas. A comparação estruturada das ondas reconstruídas contra a referência comprovou a preservação geométrica das amplitudes e dos intervalos temporais, evidenciando a mitigação eficaz de ruídos basais. Contudo, o estudo apresenta limitações inerentes à sua fase atual, pois a validação restringiu-se a ensaios controlados de bancada, não tendo o sistema sido avaliado sob o estresse de ruídos eletromiográficos reais.

Como trabalhos futuros, visando superar o atual estágio *in vitro*, pretende-se submeter o sistema a testes de Compatibilidade Eletromagnética (EMC) e segurança elétrica, precedendo a validação clínica com pacientes reais após aprovação por comitê de ética. Planeja-se também a realização de testes de rede englobando a análise do consumo energético, a precisão de sincronização e ensaios de estresse com múltiplos nós simultâneos para atestar a estabilidade a longo prazo. No âmbito de escalabilidade massiva, está prevista a integração com bancos de dados de séries temporais e a adoção de padrões de interoperabilidade (HL7 FHIR). Por fim, estruturar-se-ão avaliações de cibersegurança e a implementação de criptografia ponta a ponta via *WebSocket Secure* (WSS), assegurando conformidade estrita aos requisitos de privacidade de dados estipulados pelas normativas LGPD e HIPAA.

Agradecimentos

Os autores agradecem ao Instituto Federal de Alagoas (IFAL) pela disponibilização da infraestrutura e dos equipamentos de laboratório, os quais foram fundamentais para a condução dos ensaios de bancada e para a validação técnica da arquitetura desenvolvida neste estudo.

Referências

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acesso em: 14 ago. 2026.

[2] S. Islam, D. Kwak, M. H. Kabir, K. Hwang, e J. Koo, "The Internet of Things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 1–20, 2015. Disponível em: https://doi.org/10.1109/ACCESS.2015.2437951. Acesso em: 14 ago. 2026.

[3] F. Moura, K. Santos, A. Galvão, e G. Pitta, "A medicina remota em tempos de pandemia: um estudo prospectivo de tecnologias de telemedicina," *Cadernos de Prospecção*, vol. 15, no. 1, pp. 2–17, 2022. Disponível em: https://www.researchgate.net/publication/358938966_A_Medicina_Remota_em_Tempos_de_Pandemia_um_estudo_prospectivo_de_tecnologias_de_telemedicina. Acesso em: 17 ago. 2026.

[4] J. J. Carr e J. M. Brown, "Analog/digital conversions," em *Introduction to Biomedical Equipment Technology*, 4. ed. New Jersey: Prentice Hall, 2001, pp. 51–77.

[5] S. K. Pendyala, "Edge-cloud continuum for AI-driven remote patient monitoring: a scalable framework," *Journal of Data Science and Information Technology*, vol. 2, no. 1, pp. 1–9, 2025. Disponível em: https://www.researchgate.net/publication/388570814. Acesso em: 18 ago. 2026.

[6] B. Aljedaani e M. A. Babar, "Challenges with developing secure mobile health applications: systematic review," *JMIR mHealth and uHealth*, vol. 9, no. 6, 2021.

[7] M. J. Baucas, P. Spachos, e S. Gregori, "Internet-of-Things devices and assistive technologies for health care: applications, challenges, and opportunities," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 65–77, 2021. Disponível em: https://doi.org/10.1109/MSP.2021.3075929. Acesso em: 19 ago. 2026.

[8] Arkmeds, "Waller - simulador de paciente," 2022. Disponível em: https://arkmeds.com/pt-br/waller-simulador-de-paciente/. Acesso em: 20 ago. 2026.

[9] S. Hasan, T. Pantha, e M. A. Arafat, "Design and development of a cost-effective portable IoT enabled multi-channel physiological signal monitoring system," *Biomedical Engineering Advances*, vol. 7, p. 100124, 2024. Disponível em: https://doi.org/10.1016/j.bea.2024.100124. Acesso em: 21 ago. 2026.

[10] M. Gomez-Correa, D. Cruz-Ortiz, e M. Ballesteros, "Wearable and wireless SEMG acquisition system based on the Internet of Medical Things," *Sensing and Bio-Sensing Research*, vol. 49, p. 100828, 2025. Disponível em: https://doi.org/10.1016/j.sbsr.2025.100828. Acesso em: 24 ago. 2026.

[11] A. K. Bañuelos-Mezquitan *et al.*, "SignAPROS: an integrated hardware and software system for acquisition, processing, and analysis of biosignals," *Software Impacts*, vol. 23, pp. 2–3, 2025. Disponível em: https://doi.org/10.1016/j.simpa.2025.100741. Acesso em: 25 ago. 2026.

[12] P. C. N. Cunha, "Um modelo de eletrocardiógrafo portátil de baixo consumo," Dissertação de Mestrado, Instituto de Computação, Universidade Federal de Alagoas (UFAL), Maceió, Brasil, 2012, pp. 37–38. Disponível em: https://www.repositorio.ufal.br/handle/riufal/850. Acesso em: 26 ago. 2026.

[13] Espressif Systems, "ESP32-S3-WROOM-1 & ESP32-S3-WROOM-1U Datasheet," 2024, pp. 1–4. Disponível em: https://documentation.espressif.com/esp32-s3-wroom-1_wroom-1u_datasheet_en.pdf. Acesso em: 27 ago. 2026.

[14] Espressif Systems, "ESP32-WROOM-32E & ESP32-WROOM-32UE Datasheet," 2024, pp. 1–3. Disponível em: https://documentation.espressif.com/esp32-wroom-32e_esp32-wroom-32ue_datasheet_en.pdf. Acesso em: 27 ago. 2026.

[15] Espressif Systems, "ESP-IDF Programming Guide: Analog to Digital Converter (ADC) Continuous Mode," 2024. Disponível em: https://docs.espressif.com/projects/esp-idf/en/stable/esp32/api-reference/peripherals/adc_continuous.html. Acesso em: 28 ago. 2026.

[16] Cleveland Clinic, "Electrocardiogram (EKG)," 2021. Disponível em: https://my.clevelandclinic.org/health/diagnostics/16953-electrocardiogram-ekg. Acesso em: 31 ago. 2026.

[17] Espressif Systems, "Analog-to-Digital Converter (ADC): ESP32 API Reference," 2023. Disponível em: https://docs.espressif.com/projects/esp-idf/en/v4.4/esp32/api-reference/peripherals/adc.html. Acesso em: 1 set. 2026.

[18] Espressif Systems, "Arduino-ESP32 WiFi API Reference," 2024. Disponível em: https://docs.espressif.com/projects/arduino-esp32/en/latest/api/wifi.html. Acesso em: 2 set. 2026.

[19] Espressif Systems, "ESP Thread Border Router SDK," 2024. Disponível em: https://docs.espressif.com/projects/esp-thread-br/en/latest/. Acesso em: 3 set. 2026.

[20] I. Fette e A. Melnikov, "The WebSocket Protocol," Internet Engineering Task Force (IETF), 2011. Disponível em: https://websocket.org/guides/road-to-websockets/. Acesso em: 4 set. 2026.

[21] A. Parihar *et al.*, "Role of IoT in healthcare: applications, security & privacy concerns," *Intelligent Pharmacy*, vol. 2, no. 5, pp. 707–714, 2024. Disponível em: https://doi.org/10.1016/j.ipha.2024.01.003. Acesso em: 8 set. 2026.

[22] Espressif Systems, "ESP32-S3 Series Datasheet," 2024. Disponível em: https://www.espressif.com/sites/default/files/documentation/esp32-s3_datasheet_en.pdf. Acesso em: 9 set. 2026.

[23] TP-Link, "What is AP isolation and why do we need it?," TP-Link Support, 2018. Disponível em: https://www.tp-link.com/us/support/faq/2089/. Acesso em: 10 set. 2026.

Automated classification of feline mammary carcinoma via spatial quantification of intratumoral type I and type III collagens

Luan Prince Ferreira
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-3141-5765

Lucas Militello da Silva
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0000-4461-8716

Mariana Fernandes Jorge
Faculty of Veterinary Medicine and Animal Science
São Paulo State University
Botucatu, Brazil
ORCID: 0000-0002-6292-8512

Ana Claudia Patrocinio
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-7447-2583

Pedro Cunha Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

Abstract*—Feline mammary carcinoma (FMC) is an aggressive neoplasm that affects cats and exhibits biological characteristics similar to human breast cancer, making it an important comparative model for oncological studies. Changes in the organization of collagen fibers within the tumor microenvironment have been associated with neoplastic progression and may contribute to histopathological diagnosis. In this context, the study aimed to classify samples of healthy and carcinogenic feline mammary tissue, through sub-images stained with Picrosirius Red (PSR), using quantitative characteristics related to type I and III collagens. The dataset comprised 350 sub-images derived from 35 biological samples (4 normal/control and 31 FMC samples across histological degrees I, II and III), with each sample represented by 5 central and 5 peripheral sub-images. A 2-dimensional feature vector was extracted per sub-image. Class imbalance was addressed using a class-weighted Random Forest classifier. To prevent pseudoreplication and data leakage, model performance was evaluated using a Leave-One-Group-Out Cross-Validation (LOGO-CV) scheme grouped strictly by sample ID, with final diagnostic decisions aggregated at the sample level ($N = 35$) via majority voting. Performance metrics were reported with 95% Confidence Intervals (95% CI) derived from 1,000 bootstrap resamples. At the biological sample level, the model achieved an overall accuracy of 94.20%, a Sensitivity of 100.00%, and a Specificity of 48.82%. Feature importance analysis revealed that type III parameters (green regions) contributed 84.83% of total predictive power, highlighting intratumoral and peripheral collagen degradation as primary optical markers. These findings validate a biologically interpretable pipeline for quantitative stromal assessment, establishing a robust foundation for automated decision-support tools in veterinary histopathology.

***Index Terms*—collagen, feline mammary carcinoma, machine learning, Picrosirius Red, Random Forest.**

I. Introduction

Neoplasia refers to the process of uncontrolled cellular proliferation, contrasting with normal tissue, and can be classified as either benign or malignant [1]. Mammary neoplasms rank as the third most prevalent type of neoplasm in female felines [2-4], and when malignant, they are referred as feline mammary carcinoma (FMC) [5]. FMC represents 80 to 96% of neoplastic cases in this species. The clinical outlook for FMC typically indicates highly aggressive biological activity. Nevertheless, timely and appropriate therapeutic intervention can significantly extend survival time despite advanced disease progression [6]. This prognosis is comparable to that of triple-negative breast cancer in humans [7].

The morphological characteristics of a tumor allow it to be classified into different types and subtypes. For FMC, in the histopathological report, the degree (I, II, or III) serves as a prognostic criteria [8]. As the tumor degree increases, there is often a correlation with greater aggressiveness and shorter survival time [9].

Given the biological activity in the FMC, its intensity is greater at the tumor margins than in the center. This is due to the abundant blood supply in the periphery, which ensures the delivery of nutrients and contributes to tumor progression, while the central region tends to remain hypoxic [10]. Therefore, examining both regions of the tumor is important for assessing stromal dynamics in healthy and neoplastic tissues [11].

A clear biological correlation exists between structural changes in the mammary stroma and the onset of carcinoma, displaying conserved evolutionary mechanisms between human and feline species [12-13]. Collagen fibers form the supportive framework of the breast, referred to as the mammary parenchyma [14]. A rise in the quantity of these fibers leads to the hardening of the mammary parenchyma and is linked to the emergence of carcinoma [15]. Differentiation of collagen fibers can be accomplished through the staining of tissue sections with the Picrosirius Red (PSR) immunohistochemical

technique. After staining the tissue and examining it under polarized light, type I collagen fibers display shades of red, whereas type III fibers present shades of green [14,16]. Recent research aims to connect variations in type I or type III collagen to the tumorigenesis process.

Using specialized computational techniques, such as the AQUANCTHIS software, it is possible to quantify collagen in images of samples stained with PSR method, capable of distinguishing between type I and type III collagen [17]. Different regions within samples of healthy and tumorous feline mammary tissue were analyzed using AQUANCTHIS, and collagen values were determined. Based on these values, an attempt was made to classify tissues as healthy or carcinogenic by considering only the amount of collagen in the images.

The objective of this study is to classify samples of feline mammary tissue into healthy and carcinogenic categories, using the values of type I and type III collagen in the center and periphery of the samples under consideration.

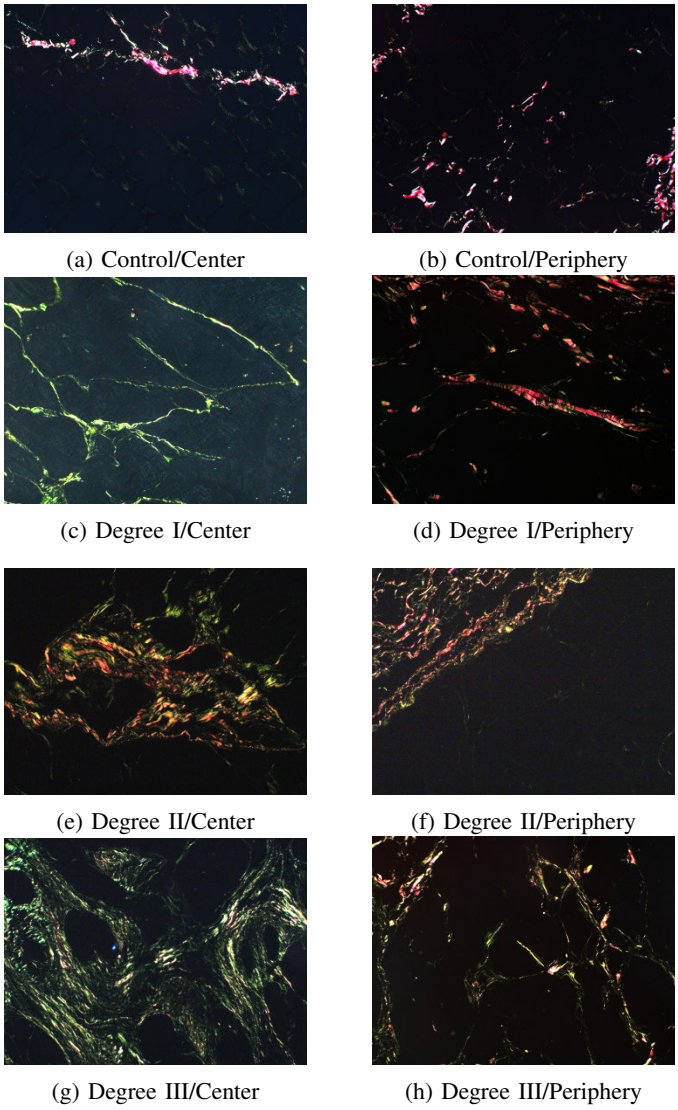
II. Methodology

A. Image dataset

Histopathological images stained with PSR were used from samples of mammary tissue obtained from female cats. The dataset consisted of 35 samples taken from female cats. Among these, there were 4 samples of healthy tissue, 3 samples classified as degree I, 13 samples of degree II, and 15 samples of degree III. Each sample provided 5 images from the central region and 5 from the peripheral region, resulting in a total of 350 images analyzed. For both regions, five distinct image fields were examined using polarized light in a dark field with a 20x magnification lens (Imager A1 microscope® - Carl Zeiss, Germany). In normal tissue, the peripheral region was defined just below the epidermis, while in neoplastic tissue, it was marked by the tumor's edges. The central section of healthy tissue was represented by fields immediately adjacent to the periphery, whereas for carcinomas, it included the intralobular regions. All samples were previously classified according to the Nottingham grading system, which is based on the system used in women [18] and is also appropriate for cats [19] and categorizes tissues based on lesion morphology, as Degree I, Degree II, or Degree III. They were both considered in the Tumor group. On the other hand, samples of healthy tissue were labeled as the Control group. Table 1 shows the distribution of samples according to these two groups.

TABLE I: Classification and distribution of the evaluated samples.

Graduation	Number of samples	Center images	Periphery images	Classification
Control	4	20	20	Healthy
Degree I	3	15	15	Tumor
Degree II	13	65	65	Tumor
Degree III	15	75	75	Tumor
Total	**35**	**175**	**175**	-



(a) Control/Center (b) Control/Periphery

(c) Degree I/Center (d) Degree I/Periphery

(e) Degree II/Center (f) Degree II/Periphery

(g) Degree III/Center (h) Degree III/Periphery

Fig. 1: Images from four different samples, according to graduation and position. Reddish areas correspond to type I collagen fibers, while greenish areas correspond to type III.

B. Pre processing

Image preprocessing consisted of applying histogram equalization using the CLAHE method. CLAHE (Contrast Limited Adaptive Histogram Equalization) is an algorithm that aims to enhance image contrast by dividing the image into smaller regions and performing histogram equalization within each region. In this study, the processing parameters were configured with a clipLimit = 2 and a tileGridSize = (8,8). The CLAHE transformation was enforced uniformly across the entire dataset prior to feature extraction [22].

C. AQUANCTHIS

AQUANCTHIS is an application for quantifying collagen in histological sections stained with Picrosirius Red, under polarized light. It allows for the separate quantification of type

I and type III collagen levels by identifying red and green regions [17]. It consists of:

- Removal of pixels in shades of grays, yellows, and blues.
- Dynamic Thresholding: Intensity segmentation computed individually for each RGB channel using an adaptive threshold defined as *Threshold* $=\mu + \tilde{x} + \sigma$, where μ is the channel mean, $\tilde{x}$ is the median, and σ is the standard deviation.
- Maximization of the proportion of green and red pixels.
- Counting of the resulting pixels.
- Calculation of the percentage of collagen relative to the sample size.

All images were processed using the AQUANCTHIS software, and the levels of collagen I and III were calculated.

D. Features

After quantifying the levels of type I and type III collagen in the central and peripheral regions of each sample, four characteristics were considered in classifying a sample: % central type I collagen (C-I), % central type III collagen (C-III), % peripheral type I collagen (P-I), and % peripheral type III collagen (P-III). Table 3 shows values of each feature determined for the images shown in Fig. 1.

TABLE II: Table Type Styles

Sample	Class	C-I	C-III	P-I	P-III
1 (Fig.1 'a' and 'b')	Control	1.978	0.681	2.636	0.263
2 (Fig.1 'c' and 'd')	Tumor	0.515	2.528	5.052	0.851
3 (Fig.1 'e' and 'f')	Tumor	6.493	1.959	4.181	0.513
4 (Fig.1 'g' and 'h')	Tumor	2.325	5.899	3.343	1.334

E. Classification model

Image classification was performed using a Random Forest algorithm (100 estimators). Class imbalance was addressed using the $classweight =' balanced'$ heuristic, which scales weights inversely proportional to class frequencies during model training.

To prevent data leakage caused by spatial correlation among images originating from the same sample, a Leave-One-Group-Out Cross-Validation (LOGO-CV) scheme was adopted. In this pipeline, data splits were grouped by sample ID. At each iteration, all images from a single biological sample were reserved strictly for model testing, while the remaining samples formed the training set. Predictions across all validation folds were aggregated to construct the global confusion matrix and compute overall classification metrics.

F. Model performance evaluation

To ensure rigorous validation at the biological sample level and avoid pseudoreplication, diagnostic decisions were aggregated across the 10 sub-images corresponding to each specimen ($N = 35$ samples) using a majority voting scheme. The consensus predicted label for a given sample was defined as the mode of the predicted classes across its constituent images.

Feature scaling was performed using a `StandardScaler` fitted strictly within the training partition of each LOGO-CV fold, ensuring zero data leakage into the test set. Diagnostic efficacy was evaluated against the histopathological gold standard reference by computing sample-level Accuracy, Sensitivity (Recall), and Specificity. To account for sample size constraints and provide robust uncertainty estimates, 95% Confidence Intervals (95% CI) were derived via non-parametric empirical bootstrapping with 1,000 resamples. Feature contributions were quantified by computing the mean mean decrease in Gini impurity across all validation folds.

III. Results and discussion

A. Sample-level diagnostic performance

When aggregated at the biological sample level via majority voting, the proposed LOGO-CV Random Forest model achieved an overall accuracy of 94.20% (95% CI: 85.71%–100.00%). The model demonstrated perfect Sensitivity (100.00%, 95% CI: 100.00%–100.00%) and a Specificity of 48.82% (95% CI: 0.00%–100.00%).

The perfect sensitivity indicates that the classifier successfully identified all malignant cases without generating false negatives—a critical requirement for automated screening systems in veterinary oncology. Conversely, the lower sample-level specificity and its associated broad confidence interval (0.00%–100.00%) reflect the marked class imbalance within the evaluated cohort, where Control samples represented a minor fraction of the dataset. Consequently, bootstrap iterations with sparse control representation exhibit high variance in specificity estimation. While this performance profile is well-suited for preliminary screening tools where sensitivity is paramount, expanding control group sizes in future cohorts will be essential to narrow confidence intervals and establish reliable negative predictive values.

B. Biophysical relevance of collagen color dynamics

Feature importance analysis revealed that type III collagen values (green tones) played a predominant role in classification, accounting for 84.83% of total predictive power (C-III: 52.74%; P-III: 32.09%), whereas type I collagen values (red tones) contributed secondary spatial information (C-I: 8.01%; P-I: 7.16%). The primary contribution type III collagen suggests that this type of collagen has more distinct optical signature differentiating tumor tissue from normal mammary parenchyma, when compared to type I collagen.

IV. Conclusion

In this study, feline mammary tissue samples (healthy vs. neoplastic) were classified at the biological sample level using a Random Forest algorithm and collagen-derived color dynamics extracted from the center and periphery of histological sections. To overcome pseudoreplication and prevent data leakage, a Leave-One-Group-Out Cross-Validation (LOGO-CV) scheme was integrated with a sample-level majority voting scheme. Class imbalance was addressed directly within the model via class-weighting heuristics.

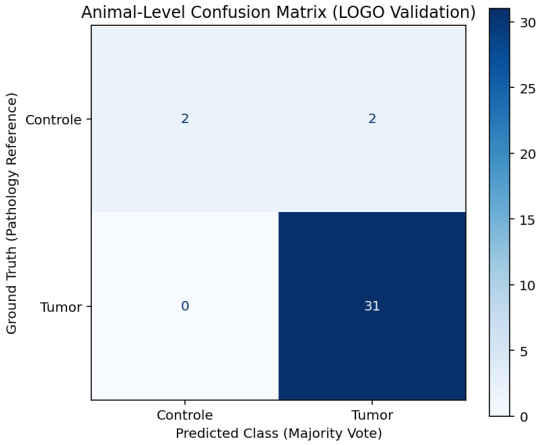


Fig. 2: Confusion matrix.

The proposed framework achieved an overall accuracy of 94.20%, demonstrating perfect Sensitivity (100.00%) and a Specificity of 48.82%. Feature importance analysis established type III collagen as the primary predictive marker (84.83% total contribution).

The primary limitation of this investigation lies in the marked class imbalance within the evaluated cohort, which constrained sample-level specificity and widened its associated confidence interval. Despite this, these findings validate a robust, biologically grounded proof-of-concept for quantitative stromal assessment in veterinary pathology. Future studies should evaluate this pipeline on larger, multi-center, and more balanced cohorts, benchmarking performance across diverse histological degrees and investigating advanced classifiers such as Support Vector Machines and XGBoost.

ACKNOWLEGMENTS

The authors acknowledge the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for their support.

REFERENCES

[1] F. V. Amorim, H. J. M. Souza, A. M. R. Ferreira, and A. B. M. Fonseca, "Clinical, cytological and histopathological evaluation of mammary masses in cats from Rio de Janeiro, Brazil," *J. Feline Med. Surg.*, vol. 8, no. 6, pp. 379–388, Dec. 2006.

[2] N. Sleeckx, "Canine Mammary Tumours, an Overview," *Reprod. Domest. Anim.*, Wiley Online Library, 2011, [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0531.2011.01816.x [Accessed: Feb. 17, 2026].

[3] D. Obradovich, *Small Animal Clinical Oncology*, CRC Press, 2016, [Online]. Available: https://www.taylorfrancis.com/books/9781482225426

[4] A. Scibelli, D. d'Angelo, A. Pelagalli, S. Tafuri, L. Avallone, and R. Della Morte, et al., "Expression levels of the focal adhesion-associated proteins paxillin and p130CAS in canine and feline mammary tumors," *Vet. Res.*, vol. 34, no. 2, pp. 193–202, 2003.

[5] S. E. Lana, G. R. Rutteman, and S. J. Withrow, "Tumors of the Mammary Gland," in *Withrow and MacEwen's Small Animal Clinical Oncology*, 4th ed., S. J. Withrow and D. M. Vail, Eds., St. Louis: Saunders Elsevier, 2007, pp. 619–638.

[6] N. P. A. D. Gunasinghe, A. Wells, E. Q. Thompson, and H. J. Hugo, "Mesenchymal-epithelial transition (MET) as a mechanism for metastatic colonisation in breast cancer," *Cancer Metastasis Rev.*, vol. 31, no. 3–4, pp. 469–478, Dec. 2012.

[7] L. Munson and A. Moresco, "Comparative pathology of mammary gland cancers in domestic and wild animals," *Breast Dis.*, vol. 28, pp. 7–21, 2007.

[8] D. Betz, D. Schoenrock, R. Mischke, W. Baumgärtner, and I. Nolte, "Postoperative treatment outcome in canine mammary tumors. Multivariate analysis of the prognostic value of pre- and postoperatively available information," *Tierarztl. Prax. Ausg. K. Kleintiere Heimtiere*, vol. 40, no. 4, pp. 235–242, 2012.

[9] C. B. De Campos, K. A. Damasceno, C. O. Gamba, A. M. Ribeiro, C. J. Machado, and G. E. Lavalle, et al., "Evaluation of prognostic factors and survival rates in malignant feline mammary gland neoplasms," *J. Feline Med. Surg.*, vol. 18, no. 12, pp. 1003–1012, Dec. 2016.

[10] C. K. Kuhl, "Review article MRI of breast tumors," *Eur. J. Radiol.*, vol. 58, pp. 46–58, 2000.

[11] D. F. Quail and J. A. Joyce, "Microenvironmental regulation of tumor progression and metastasis," *Nat. Med.*, vol. 19, p. 1423, 2013.

[12] R. De Maria, M. Olivero, S. Iussich, M. Nakaichi, T. Murata, and B. Biolatti, et al., "Spontaneous feline mammary carcinoma is a model of HER2 overexpressing poor prognosis human breast cancer," *Cancer Res.*, vol. 65, no. 3, pp. 907–912, Feb. 1, 2005.

[13] A. J. F. Matos, C. S. Baptista, M. F. Gärtner, and G. R. Rutteman, "Prognostic studies of canine and feline mammary tumours: the need for standardized procedures," *Vet. J.*, vol. 193, no. 1, pp. 24–31, Jul. 2012.

[14] V. Zappulli, G. De Zan, B. Cardazzo, L. Bargelloni, and M. Castagnaro, "Feline mammary tumours in comparative oncology," *J. Dairy Res.*, vol. 72, spec. no., pp. 98–106, 2005.

[15] C. S. Marques, M. Soares, A. Santos, J. Correia, and F. Ferreira, "Serum SDF-1 levels are a reliable diagnostic marker of feline mammary carcinoma, discriminating HER2-overexpressing tumors from other subtypes," *Oncotarget*, vol. 8, no. 62, pp. 105775–105789, Dec. 1, 2017.

[16] T. Tamamoto, K. Ohno, Y. Goto-Koshino, and H. Tsujimoto, "Serum amyloid A promotes invasion of feline mammary carcinoma cells," *J. Vet. Med. Sci.*, vol. 76, no. 8, pp. 1183–1188, Aug. 2014.

[17] L. P. Ferreira, M. F. Jorge, A. C. Patrocínio, "AQUANCTHIS: Aplicação para quantificação de colágeno em cortes histológicos corados com PSR," Versão Java/Python. Registro de programa de computador nº BR512026001726-2, INPI, Mar. 24, 2026.

[18] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. The value of histological degree in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, pp. 403–410, 1991.

[19] G. D. Cassali et al., "Consensus for the diagnosis, prognosis and treatment of feline mammary tumors," *Braz. J. Vet. Res. Anim. Sci.*, vol. 55, 2018.

[20] F. Seixas, C. Palmeira, M. A. Pires, M. J. Bento, and C. Lopes, "Degree is an independent prognostic factor for feline mammary carcinomas: A clinicopathological and survival analysis," *Vet. J.*, vol. 187, no. 1, pp. 65–71, 2011.

Automated Nuclei Segmentation and Counting in Human Liver Histological Images Using Watershed

Pedro Victor da Costa Santos
Biomedical Engineering Department
UFRN
Natal, Brazil
pedro.costa.709@ufrn.edu.br

Vinicius Ribeiro Rocha
Biomedical Engineering Department
UFRN
Natal, Brazil
vinicius.rocha.100@ufrn.edu.br

Thyana Maria Soares de S. Andrade
Biomedical Engineering Department
UFRN
Natal, Brazil
thyana.andrade.048@ufrn.edu.br

Caroline Dantas Vilar
Biomedical Engineering Department
UFRN
Natal, Brazil
caroline.vilar@ufrn.br

Heliana Bezerra Soares
Biomedical Engineering Department
UFRN
Natal, Brazil
heliana.soares@ufrn.br

***Abstract*—Segmentation and counting of nuclei in histological images are important steps in the quantitative analysis of tissues and can aid in the characterization of morphological alterations. However, variations in contrast, staining, and overlapping structures hinder the automation of this process. This work proposes a digital image processing method for automated segmentation and counting of nuclei in histological images of the human liver. Forty images from the public NuInsSeg dataset, containing 1,370 manually annotated nuclei, were evaluated across a 30 image development set and a 10 image independent test set. The developed method includes contrast adjustment, bilateral filtering, adaptive thresholding, morphological operations, distance transformation, and watershed segmentation, followed by labeling of connected components. The automatically obtained counts were compared to reference annotations. On the independent test set, the method presented a mean absolute error of 2.30 nuclei per image and a mean absolute percentage error of 7.28%. The results indicate that the approach enabled the automated quantification of nuclei in the evaluated images, although limitations were observed in regions with nuclear clustering or overlap. Thus, the method shows potential as a deterministic and interpretable alternative for the quantitative analysis of liver histological images.**

***Index Terms*—Digital Image Processing; Nuclei Segmentation; Watershed Transform; Histological Images.**

I. Introduction

Quantitative analysis of histological images plays a relevant role in digital pathology, as it allows the extraction of information related to tissue organization and the morphological characteristics of cellular structures. In this context, nuclear segmentation is a fundamental step for tasks such as determining nuclear density, evaluating the morphology and spatial distribution of nuclei, and supporting the characterization of tissue alterations [1], [2]. However, performing this procedure manually is laborious, time-consuming, and subject to variability between observers, especially when there is a large number of structures in the same image [1].

Automated segmentation of nuclei in histological images presents challenges related to variations in staining and contrast, the presence of artifacts, morphological heterogeneity, and proximity or overlap between nuclei [1], [3]. Different computational approaches have been proposed for this purpose, ranging from classical digital image processing techniques, such as thresholding, morphological operations, and distance transform, to methods based on deep learning [1], [2]. Although deep learning models have achieved significant results, their use generally requires annotated datasets, computational resources, and training and tuning procedures. [2], [4].

In this scenario, classical methods remain relevant in delimited applications, especially when seeking a deterministic, interpretable solution with relatively simple computation implementation. Among these techniques, the watershed algorithm can be used to separate adjacent structures based on information related to the distance between the objects and the image background. However, its performance depends on the preprocessing steps, the definition of markers, and parameter adjustment, and excessive fragmentation or insufficient separation of grouped nuclei may occur [1], [3].

Given this, this work proposes and evaluates a digital image processing method for the automated segmentation and counting of nuclei in hematoxylin and eosin stained human liver histological images. The method combines contrast adjustment, bilateral filtering, adaptive thresholding, morphological operations, distance transform, watershed segmentation, and connected component labeling. The evaluation was performed on 40 images from the public NuInsSeg dataset, using the manual annotations available in the database as a reference [4]. The main contribution of the study is the quantitative evaluation of a deterministic and interpretable workflow, with fixed parameters, for counting nuclei in liver histological images.

II. Materials e Methods

A. Database and image selection

For the preliminary evaluation of the method, the public dataset NuInsSeg was used, developed for studies of nuclear instance segmentation in hematoxylin and eosin (H&E) stained histological images. The dataset contains 665 images of different human and murine tissues, with a resolution of 512×512 pixels, totaling 30,698 manually segmented nuclei. In addition to the original images, the dataset provides masks in which each nucleus is individually identified, as well as masks of ambiguous regions, corresponding to areas where precise delimitation was not considered possible even by specialists [4].

To delimit the experimental scope and reduce the morphological variability associated with the joint analysis of different organs, the 40 images of human liver tissue available in the database were selected. The respective instance masks provided by NuInsSeg were used as a reference to obtain the manual count. The number of reference nuclei in each image was determined by the number of instances individually identified in the masks, resulting in a total of 1,370 nuclei in the 40 images evaluated.

To prevent evaluation bias on tuned data, the 40 human liver images were structured into two distinct subsets: a development subset (30 images, 1,047 reference nuclei) used for empirical parameter definition, and an independent evaluation test subset (10 images, 323 reference nuclei) preserved untouched during calibration. Although both subsets originate from the same liver collection in NuInsSeg, this separation ensures that primary evaluation metrics reflect execution on uncalibrated images.

B. Proposed method

The proposed method was developed in MATLAB and organized into sequential steps of preprocessing, segmentation, and counting, as illustrated in Fig. 1. Initially, each colored histological image was converted to grayscale, reducing color information while preserving the intensity variations used in identifying nuclear structures. Then, the image intensity was adjusted to increase the contrast between the nuclei and the background tissue.

After adjusting the contrast, a bilateral filter was applied to attenuate noise and small variations in intensity, while simultaneously preserving the transitions associated with the edges of the nuclei. Unlike conventional linear filters, bilateral filtering considers both spatial proximity and intensity similarity between pixels, favoring smoothing in homogeneous regions without significantly compromising the contours of the structures [4].

The filtered image was subjected to adaptive thresholding to obtain an initial binary mask. In this technique, the threshold is determined according to the local characteristics of the image, making it possible to deal with variations in lighting and contrast present in the histological field [5]. Then, morphological operations were applied to remove small isolated regions,

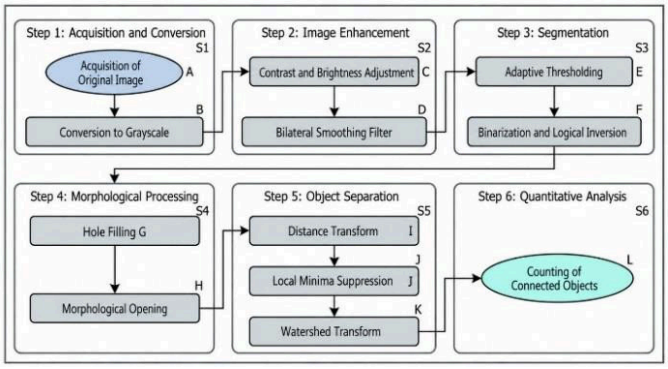


Fig. 1. Methodological workflow employed in the automated segmentation and counting of nuclei.

smooth the contours and preserve objects compatible with nuclear structures. These operations use a structuring element to analyze and modify the geometry of the components present in the binary image [6].

Since nearby or partially overlapping nuclei can form a single component in the mask, a distance transform was calculated on the binary image. This transformation assigns each pixel of the object a value corresponding to its distance from the background, producing local maxima near the central regions of the nuclei [7]. The resulting matrix was used in watershed segmentation, a technique that interprets the image as a topographic surface and establishes dividing lines between regions associated with different minima or markers [8]. To reduce the occurrence of excessive fragmentation, minimum suppression was applied before the watershed execution.

Finally, the dividing lines produced by the watershed were incorporated into the binary mask to separate adjacent nuclei. The resulting connected components were then individually labeled, and the number of distinct labels was considered as the automated count of nuclei in the image. The same processing flow, with fixed parameters, was applied to all evaluated images.

C. Experimental setup

The method was implemented in MATLAB using functions from the Image Processing Toolbox. The images were processed individually, preserving their original resolution of 512×512 pixels. A subset of 30 liver images was used for initial development and empirical parameter definition. After this step, the same processing sequence, with fixed parameters, was applied to the 10 independent test images without specific adjustments for each case..

In the pre-processing step, the image intensity in gray levels was adjusted to increase the contrast between the nuclear structures and the background tissue. Then, a bilateral filter with a spatial smoothing degree of 0.5 was applied. To obtain the binary mask, adaptive thresholding was used with a sensitivity of 0.55, allowing the threshold value to vary according to the local characteristics of the image.

TABLE I
PARAMETERS USED IN IMAGE PROCESSING

Step	Parameter	Valuer
Bilateral filtering	Degree of spatial smoothing	0,5
Adaptive thresholding	Sensitivity	0,55
Morphological operation	Radius of the structuring element	9 pixels
Minimum suppression	(h)	5
Labeling	Connectivity	4

After binarization, morphological operations were performed using a disk-shaped structuring element with a radius of 9 pixels. Subsequently, the distance transform of the binary mask was calculated. Minimum suppression was applied with a parameter $h = 5$, aiming to reduce the formation of irrelevant minima and, consequently, limit excessive fragmentation during watershed segmentation. The resulting dividing lines were removed from the mask, allowing the separation of adjacent nuclei.

Finally, the connected components present in the segmented mask were identified using connectivity 4. Each labeled component was considered a nuclear instance, and the total number of components corresponded to the automated count of nuclei in each image. The parameters used are presented in Table I.

D. Quantitative evaluation

The method's performance was evaluated by comparing the number of nuclei automatically identified with the reference count obtained from the instance masks provided by NuInsSeg. For each of the 40 images, the number of reference nuclei and the number estimated by the proposed method were recorded.

Initially, the absolute error was calculated for each image, as defined in Eq. (1):

$$E_i = |y_i - \hat{y}_i| \tag{1}$$

where y_i represents the reference count, $\hat{y}_i$ corresponds to the count estimated by the method and E_i represents the absolute error associated with image i.

As a global performance measure, the mean absolute error (MAE) was calculated, according to Eq. (2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where n corresponds to the total number of images evaluated. The MAE represents, in number of nuclei per image, the average magnitude of the differences between the automatic and reference counts.

The mean absolute percentage error (MAPE) was also calculated, as presented in Eq. (3):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3}$$

Fig. 2 presents the intermediate processing results: (a) original histological image; (b) grayscale image; (c) image after contrast adjustment; (d) result of bilateral filtering; (e) mask obtained by adaptive thresholding; (f) mask after filling internal regions; (g) result of morphological operations; (h) reference mask; (i) distance transform; and (j) result of watershed segmentation.

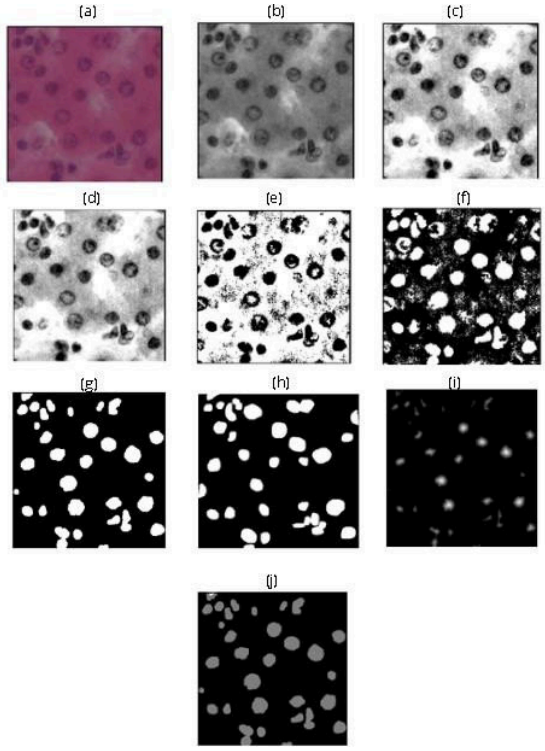


Fig. 2. Intermediate processing results: (a) original histological image; (b) grayscale image; (c) image after contrast adjustment; (d) result of bilateral filtering; (e) mask obtained by adaptive thresholding; (f) mask after filling internal regions; (g) result of morphological operations; (h) reference mask; (i) distance transform; and (j) result of watershed segmentation.

MAPE expresses the average error in relation to the number of nuclei present in each image, allowing consideration of differences in nuclear density between the evaluated histological fields. Since all analyzed images presented at least one reference nucleus, it was not necessary to apply specific treatment for null values in the denominator of (3).

In addition to these metrics, the standard deviation of the absolute errors obtained through (1) was calculated in order to analyze the dispersion of errors among the images. The results were analyzed descriptively, without the application of inferential tests, considering the preliminary nature of the evaluation and the absence of an independent set of tests.

III. RESULTS AND DISCUSSION

A. Qualitative analysis of segmentation

Fig. 2 (a)–(d) shows the intermediate results obtained during the application of the method to one of the histological images of a human liver. The original image shows variations in color, contrast, and proximity between the nuclei, characteristics that make their direct separation difficult. After conversion to grayscale and intensity adjustment, greater differentiation was observed between the nuclear structures and the background

tissue. Bilateral filtering helped to attenuate small variations in intensity, visually preserving the predominant outlines of the nuclei.

In the adaptive thresholding stage, the core regions were highlighted for the formation of the binary mask. However, small isolated regions and discontinuities within some objects were also identified. Filling in internal regions and applying morphological operations reduced some of these artifacts and produced objects with more regular contours. These operations also resulted in the removal of some smaller regions, indicating that the result is influenced by the size of the structuring element and the dimensions of the cores present in the image. The results of the thresholding and morphological operations are presented in Fig. 2(e)–(g).

The distance transform produced maxima in the internal regions of the segmented objects, providing information for the separation of adjacent nuclei. Subsequently, the application of the watershed established dividing lines between some structures that remained connected in the binary mask. Visual comparison with the reference mask indicates that the processing made it possible to identify and separate a significant portion of the nuclei present in the analyzed image. However, differences remained, mainly related to clustered nuclei, partially overlapping nuclei, or nuclei with low contrast relative to the surrounding tissue. The reference mask, the distance transform, and the watershed result are presented in Fig. 2(h)–(j), respectively.

The qualitative results demonstrate the behavior of each step in the processing flow and help explain the differences subsequently observed between automatic and reference counts. However, this visual analysis is illustrative and should not be interpreted in isolation as evidence of the overall performance of the method, which is presented quantitatively in the following section.

B. Quantitative performance of the count

Quantitative evaluation was performed on 40 histological images of human liver, containing a total of 1,370 nuclei identified in the reference masks. Table II summarizes the performance metrics separated by development set (N = 30) and independent test set (N = 10) by comparing the automatic and reference counts.

On the independent test set (N = 10), the method achieved a low MAE of 2.30 nuclei per image and a MAPE of 7.28%. This result indicates that the estimated count differed from the reference, on average, by approximately two nuclei per image. The standard deviation of the absolute errors was 1.34 nuclei, showing variation in the magnitude of the errors among the evaluated images.

The mean absolute percentage error was 7.28%, demonstrating that the differences between the counts corresponded, on average, to 7.28% of the number of reference nuclei in each image. This value should be interpreted as a measure of relative error, and not as the accuracy of the method.

The results indicate that the method made it possible to obtain counts close to the reference values in some images

TABLE II
QUANTITATIVE RESULTS OF AUTOMATED NUCLEUS COUNTING

Metric	**Valuer**	**Test**	**Interpretation**
Number of images (n)	30	10	Sample size
MAE	4,80 nuclei	2,30 nuclei	Average absolute error per image
Standard deviation of absolute errors	5,56 nuclei	1,34 nuclei	Dispersion of absolute errors
MAPE	14,26%	7,28%	Average absolute percentage error

but showed greater divergence in certain histological fields. The dispersion of errors suggests that the performance was not uniform across the entire set, being influenced by the specific characteristics of each image. Among the factors that may explain this variation are the clustering and overlapping of nuclei, differences in contrast, and the presence of structures with dimensions different from those considered during the parameter adjustment.

Considering that the parameters were empirically defined using the development subset, while the independent test subset was kept untouched during parameter tuning, the results obtained on the test set provide a less biased estimate of the method's performance on previously unseen images. However, since both subsets were derived from the same human liver image collection in the NuInsSeg dataset, these results should not be interpreted as external validation. Nevertheless, they provide a quantitative characterization of the behavior of the deterministic method on liver histological images.

C. Analysis of the main errors

The differences between automatic and reference counts may be mainly related to inadequate separation of adjacent nuclei and the identification of non-nuclear regions. In cases of undercounting, two or more nearby or partially overlapping nuclei may remain connected after morphological operations and, consequently, be counted as a single component. This behavior is especially relevant when objects have poorly defined boundaries or when the distance transform does not produce sufficiently distinct maxima to guide their separation by the watershed [1], [3].

Overcounting, in turn, can occur when a single nucleus is fragmented into multiple regions during the application of the watershed, or when small structures and artifacts remain in the binary mask and are interpreted as nuclear components. Although minimum suppression has been employed to reduce excessive fragmentation, the result of the watershed remains dependent on the characteristics of the input mask and the definition of the markers or minima used in the segmentation process [3], [9].

Another relevant factor is the use of fixed parameters in all images. Although this strategy ensures the uniform application of the method, variations in contrast, intensity, morphology, and size of the nuclei can alter the response of adaptive

thresholding and morphological operations. In particular, the radius of the structuring element can favor the removal of small regions and, simultaneously, promote the merging of very close structures. Thus, a single set of parameters may not adequately represent all the variability present in the analyzed histological images.

Furthermore, the evaluation was based exclusively on the difference between the automatic and reference counts. Similar counts do not necessarily ensure that each nucleus has been correctly segmented, as undercounting and overcounting errors can occur simultaneously and compensate each other numerically. Therefore, the results obtained characterize the method's performance in the counting task, but do not allow for a complete evaluation of the spatial quality of the segmentation of each nuclear instance.

Finally, an experimental comparison with other classical or deep learning-based methods was not performed. Therefore, it is not possible to assert superiority or performance improvement over existing approaches. In addition, although the test subset was kept independent during parameter tuning, both the development and test subsets were derived from the same human liver image collection in the NuInsSeg dataset. Thus, the results should be interpreted as an evaluation of a deterministic, interpretable, fixed-parameter-based workflow on the selected liver images rather than as evidence of superiority or external generalization.

IV. Conclusion

This work presented a deterministic digital image processing method for the automated segmentation and counting of nuclei in hematoxylin and eosin (H&E)-stained human liver histological images. The proposed workflow combined contrast adjustment, bilateral filtering, adaptive thresholding, morphological operations, distance transform, watershed segmentation, and connected-component labeling. For parameter definition, the 40 selected liver images from the NuInsSeg dataset were divided into a development subset (N = 30), used for empirical parameter adjustment, and an independent test subset (N = 10), kept untouched during parameter tuning. On the independent test set, the proposed method achieved a mean absolute error (MAE) of 2.30 nuclei per image, a standard deviation of the absolute errors of 1.34 nuclei, and a mean absolute percentage error (MAPE) of 7.28%. These results indicate that the method produced automated nuclei counts relatively close to the reference values, although some variation in error magnitude was observed across the evaluated histological images.

The standard deviation of the absolute errors was 1.34 nuclei, indicating variability in performance across the analyzed histological images. Overall, the results show that the proposed approach was able to produce automated nuclei counts reasonably close to the reference values in part of the evaluated dataset, although its performance varied according to the characteristics of the histological fields.

The main limitations were associated with the separation of adjacent or partially overlapping nuclei, fragmentation of individual structures, and incorrect identification of non-nuclear regions. In addition, the use of fixed parameters may limit the method's ability to accommodate variations in nuclear size, morphology, and image contrast. The selected images were divided into a development subset (N = 30), used for empirical parameter adjustment, and an independent test subset (N = 10), kept untouched during parameter tuning. Although this separation provides an evaluation on images not used for parameter definition, both subsets were derived from the same human liver image collection in the NuInsSeg dataset; therefore, the results do not constitute external validation. Furthermore, evaluation based primarily on nuclei counts does not provide a complete assessment of the spatial accuracy of individual nucleus segmentation.

Future work should include validation on larger independent datasets and histological images from different tissues, as well as the incorporation of instance-level segmentation metrics to provide a more comprehensive assessment of segmentation performance. Comparisons with other classical image processing methods and deep learning-based approaches should also be conducted. In addition, automatic parameter selection and more robust strategies for watershed marker definition may be investigated to reduce both undercounting and overcounting errors.

Acknowledgment

The authors thank LAPSIM – the Laboratory for Acquisition and Processing of Medical Signals and Images of the Biomedical Engineering course at the Federal University of Rio Grande do Norte (UFRN) – for the technical support, infrastructure, and assistance in the development of this work.

References

[1] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review—Current status and future potential," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 97–114, 2014, doi: 10.1109/RBME.2013.2295804.

[2] N. Kumar *et al.*, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017, doi: 10.1109/TMI.2017.2677499.

[3] M. Kowal, M. Żejmo, M. Skobel, J. Korbicz, and R. Monczak, "Cell nuclei segmentation in cytological images using convolutional neural network and seeded watershed algorithm," *Journal of Digital Imaging*, vol. 33, pp. 231–242, 2020.

[4] A. Mahbod *et al.*, "NuInsSeg: A fully annotated dataset for nuclei instance segmentation in H&E-stained histological images," *Scientific Data*, vol. 11, 2024, doi: 10.1038/s41597-024-03117-2.

[5] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. Sixth International Conference on Computer Vision*, Bombay, India, 1998, pp. 839–846, doi: 10.1109/ICCV.1998.710815.

[6] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979, doi: 10.1109/TSMC.1979.4310076.

[7] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.

[8] G. Borgefors, "Distance transformations in digital images," *Computer Vision, Graphics, and Image Processing*, vol. 34, no. 3, pp. 344–371, 1986, doi: 10.1016/S0734-189X(86)80047-0.

[9] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1991, doi: 10.1109/34.87344.

Automated Screening for Metallic Implants in MRI Scout using Deep Learning

1st Aline Tabatta Rossi, 2nd Fabio K. Schneider, and 3rd José Jair Alves Mendes Junior
Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brazil.
alinetabatta@alunos.utfpr.edu.br, {fabioks, jjjunior}@utfpr.edu.br

***Abstract*—The early detection of metallic implants in Magnetic Resonance Imaging (MRI) is critical for workflow optimization, allowing for the timely implementation of metal artifact reduction sequences. This pilot study proposes an automated screening framework designed to identify metallic interference during the initial "scout" (localizer) phase of knee MRI examinations. A comparative analysis was conducted between two Deep Learning (DL) architectures: ResNet18 and YOLOv8. Both models were evaluated on a newly curated, publicly released, and balanced dataset of 1,254 knee MRI localizer scans (UTFPR-Knee-MRI-Metal-Scout) from 200 patients across two different vendors. Performance was statistically validated via patient-level bootstrapping to establish 95% confidence intervals (CI). Furthermore, the models were evaluated for robustness against "spike" artifacts generated through k-space manipulation and noise degradation. Under standard conditions, YOLOv8 achieved an accuracy of 99.4% and a recall of 98.9%. In the 25% spike training scenario, YOLOv8 maintained high robustness with 97% accuracy, significantly exceeding the 66% achieved by ResNet18. Conversely, ResNet18 exhibited superior stability under extreme Gaussian noise, maintaining 90% accuracy at a Signal-to-Noise Ratio (SNR) of 5. Qualitative analysis using Grad-CAM revealed that misclassifications often occur when susceptibility artifacts merge with the image background due to low contrast-to-noise ratios. These findings suggest that integrating automated detection into the scout phase provides a reliable trigger for specialized clinical protocols, significantly reducing the necessity for costly re-scans. Code and dataset are available on GitHub repository[1].**

***Index Terms*—Deep Learning, Magnetic Resonance Imaging, Metal Artifacts, YOLOv8, ResNet18.**

I. Introduction

Magnetic Resonance Imaging (MRI) is a high-cost, non-invasive imaging modality that does not rely on ionizing radiation [1]. Its operation is based on the interaction of atomic nuclei in a high-intensity magnetic field [2]. MRI provides detailed information on multiple tissue properties, including flow, diffusion, perfusion, functional activity, and metabolic composition through spectroscopy [2].

Since its commercial introduction in the 1980s, the demand for MRI examinations has steadily increased [3]. In the Brazilian public healthcare system, the average growth rate reached 7.6% for Computed Tomography (CT) scanners and 10.8% for MRI scanners, while in the private sector the respective increases were 7% and 7.8% [4].

Parallel to this increasing demand, Machine Learning (ML) has introduced methods that accelerate MRI acquisition and improve image quality. Commercial solutions currently available reduce scan time through noise suppression, raw-data reconstruction, synthetic image generation, and personalized imaging protocols [5]. However, despite these advances, artifacts remain a recurrent factor affecting MRI quality, potentially degrading image clarity and compromising diagnostic interpretation [6]. Beyond the clinical implications of image degradation, there is a substantial financial burden. As quantified in [7], motion artifacts are a primary driver of operational inefficiency and lost revenue in radiology departments. Their research indicates that approximately 19.8% of all MRI examinations require at least one repeated sequence due to motion. This translates to an estimated annual cost of approximately $115,000 per scanner in lost productivity and resource waste.

MRI studies typically generate a large number of images, making exhaustive evaluation challenging. Image quality assessment (IQA) in MRI is a broad research domain encompassing diverse methodological approaches. In recent years, Deep Learning (DL) techniques have gained prominence in MRI analysis, offering powerful tools for automated and scalable IQA. Several DL models have been developed to detect common MRI artifacts and, in some cases, even propose corrective strategies [8].

Convolutional Neural Networks (CNNs), for instance, have proven particularly effective in identifying complex patterns associated with signal voids and geometric distortions, such as those caused by metal implants. Building on this capability, Gao et al. [9] developed a lightweight DL framework for fully automated sorting of brain MRI volumes and the detection of common artifacts, including motion, aliasing, and metal interference. By using an EfficientNetB0 architecture as a feature extractor and a multilayer perceptron classifier, the authors achieved an overall accuracy of 99.1% across a large retrospective dataset of over 22,000 MRI volumes.

Ettehadi et al. [10] proposed a two-step DL framework for multiclass artifact classification in brain Diffusion MRI. Unlike previous binary "good vs. poor" models, their approach utilizes a 3D Residual Squeeze-and-Excitation CNN to first classify 3D sub-volumes, or "slabs," followed by a voting process to determine the primary artifact type in the entire volume. Validated on large-scale datasets, the model achieved average accuracies of 96.61% and 97.52%, respectively.

The push toward large-scale automation reached a new milestone with the work of Loizillon et al. [11], who developed a comprehensive quality control pipeline specifically for 3D

[1]https://github.com/alinetabatta/Metal_detection_CNN

FLAIR sequences within clinical data warehouses. Their study emphasizes that while clinical routine images offer a wealth of data, they are frequently compromised by diverse artifacts that standard research-grade models fail to capture. Loizillon and colleagues demonstrated that DL can effectively bridge this gap by providing high-throughput screening of brain MRIs.

Integrating DL prior to the primary MRI examination offers a strategic advantage by transforming retrospective quality checks into proactive, real-time clinical safeguards. Expanding the scope of DL in pre-examination safety and protocol optimization, Courtman et al. [12] developed a CNN-based framework to identify head CT scans containing aneurysm clips prior to MRI procedures. Their approach successfully automated patient safety screening, eliminating manual review bottlenecks and preventing potential MR-safety contraindications before the patient enters the magnet room. Similarly, Lie et al. [13] developed a CNN to automatically screen X-rays for metallic implants and foreign bodies. By training algorithms on a musculoskeletal radiograph dataset, they achieved near-perfect accuracy across six anatomical regions. Their findings demonstrate that deploying automated DL screening upstream in the diagnostic workflow substantially decreases pre-scan evaluation time and serves as a scalable substitute for manual triage, effectively removing real-time clinical decision-making barriers.

While the aforementioned studies demonstrate high performance in automated quality assessment and artifact identification across various anatomical regions, they share the common limitation of their temporal placement within the clinical workflow. Whether focusing on specific artifact types or global image quality, these methodologies are typically applied at the final stage of the MRI process, often after the patient has already left the facility. Consequently, these approaches serve as retrospective diagnostic tools rather than preventative measures, leaving no window for immediate corrective action or protocol adjustment.

In contrast, we present a proactive strategy designed to anticipate potential failures and optimize clinical efficiency. By detecting the presence of metallic implants at the earliest possible inception of the examination, this framework enables real-time decision-making while the patient remains on the table. This shift from retrospective identification to early-stage prevention aims to drastically reduce time-to-diagnosis and minimize the necessity for subsequent re-imaging. Because artifacts are pervasive in daily practice, early identification is a vital tool for clinicians and technicians, ultimately enhancing patient safety and diagnostic reliability.

Metallic interference compromises MRI quality by disrupting the uniformity of the static magnetic field, thereby altering the local resonance frequency. This phenomenon results in a cascade of characteristic artifacts, including signal loss due to accelerated dephasing, failure of chemical-shift-based fat suppression, and significant pixel displacement artifacts [14].

Hence, this work originally proposes detecting knee metallic artifacts at the earliest possible stage of the MRI workflow: the localizer (or "scout") scan. Although localizer images are primarily used for anatomical positioning rather than diagnosis, they provide critical early insights that can prevent subsequent image degradation. Early identification of metal during this phase allows the technician to immediately adjust the clinical protocol and select specialized sequences designed to reduce metal artifact. These include MARS (GE Healthcare), MAVRIC SL (GE Healthcare), WARP (Siemens), and O-MAR (Philips) [15]. Integrating automated detection into this initial phase serves as a preventive measure, reducing the need for costly, time-consuming rescans while optimizing scanner throughput.

To evaluate the most effective approach for early preventive screening, this study compares two distinct DL architectures representing different paradigms rooted in CNNs: ResNet18 [16], a classic residual classification network, and YOLOv8 [17], a high-speed single-stage object detector. As emphasized in medical literature, CNN-based frameworks have become the cornerstone of modern medical imaging analysis due to their capacity to extract hierarchical visual features directly from complex anatomical datasets [18]. By benchmarking a dedicated classifier against a specialized object detector within the same workflow, this study explores how these distinct CNN lineages handle the morphological features of localized metallic artifacts to identify the optimal balance of clinical performance. Furthermore, beyond standard metrics, both models were subjected to rigorous operational testing—including noise sensitivity and spike artifact analysis—to ensure the selected approach remains resilient against the image degradation common in routine clinical practice. Finally, we present a new publicly available dataset for this evaluation, named **UTFPR-Knee-MRI-Metal-Scout**, which serves as a secondary contribution of this work.

II. Methods

The methodological framework of this study is designed to evaluate the efficiency of DL architectures in the early detection of metallic artifacts. The experimental workflow is systematically divided into four primary stages, as illustrated in Fig. 1: (i) Data Curation, involving the manual selection and labeling of knee MRI localizer scans; (ii) Model Development, where the ResNet18 and YOLOv8 architectures are trained using transfer learning; (iii) Robustness Testing, consisting of noise sensitivity, dataset ablation analyses and spike artifacts; and (iv) Performance Evaluation, utilizing quantitative metrics and qualitative saliency maps to compare the proposed models.

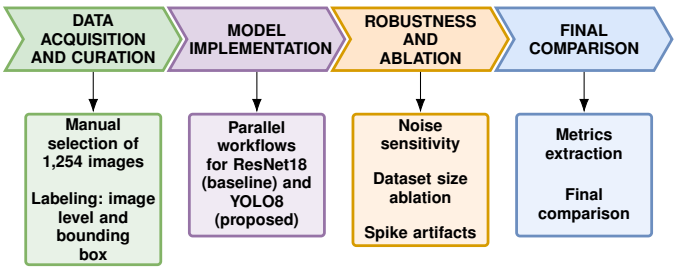


Fig. 1. Overview of the proposed methodology.

A. Dataset

The retrospective collection and processing of imaging data for this study were conducted following the ethical standards and approval of the Ethical Committee of the Federal University of Technology (UTFPR) campus Dois Vizinhos, approval number 7.554.887.

The **UTFPR-Knee-MRI-Metal-Scout** dataset consists of knee MRI localizer scans, specifically selected for their role in the initial clinical workflow. Data acquisition was performed over approximately 40 hours of manual selection by a specialist to ensure the integrity of the ground truth. To prevent algorithmic bias, a balanced dataset was constructed, comprising 50% scans containing metallic implants and 50% scans without metallic interference. The images were sourced from 1.5 Tesla MRI scanners from two primary vendors: Siemens Healthineers and GE Healthcare.

Each examination in the study typically included an average of 6 localizer images, categorized into the three primary anatomical planes: sagittal, coronal, and axial. As detailed in Table I, the dataset of 1,254 images includes 460 sagittal, 423 coronal, and 371 axial views. This distribution ensures the models are trained and validated on a balanced variety of anatomical perspectives.

TABLE I
DISTRIBUTION OF THE DATASET: ORGANIZED BY TRAINING PHASE AND CLASS ACROSS ANATOMICAL PLANES.

Split	Class	Sagittal	Coronal	Axial	Total
Train	Metal	186	150	110	446
	No Metal	145	151	149	445
	Subtotal	*331*	*301*	*259*	*891*
Validation	Metal	35	29	29	93
	No Metal	28	33	31	92
	Subtotal	*63*	*62*	*60*	*185*
Test	Metal	36	30	23	89
	No Metal	30	30	29	89
	Subtotal	*66*	*60*	*52*	*178*
Total		**460**	**423**	**371**	**1,254**

For computational processing, raw localizer scans were normalized and exported to 512×512 pixel 8-bit representations. Direct extraction of raw DICOM files was restricted due to institutional data privacy and confidentiality policies established by the clinical facility. Nevertheless, while full 12/16-bit DICOM precision is crucial for subtle diagnostic tissue characterization, metallic artifacts in scout acquisitions manifest as severe, macroscopic, high-contrast signal voids and geometric distortions [19]. These morphological features are fully preserved in 8-bit spatial representations without compromising ML classification accuracy [20].

The final dataset comprises 1,254 images from 200 unique patients. The partitioning into training, validation, and testing subsets was conducted at the patient level to ensure a rigorous evaluation and prevent data leakage. The study mitigates the risk of the model overlearning specific anatomical signatures by ensuring that all images from a single patient are assigned exclusively to a single subset. This ensures that the network is evaluated on its ability to generalize to completely unseen patients, providing a more realistic assessment of its clinical utility. The distribution of images across the subsets is as follows: 891 images for training, 185 for validation, and 178 for the final test set. Each subset maintained the 1:1 ratio between the "Metal" and "No Metal" classes. Samples of the localizers and their corresponding anatomical planes are illustrated in Fig. 2.

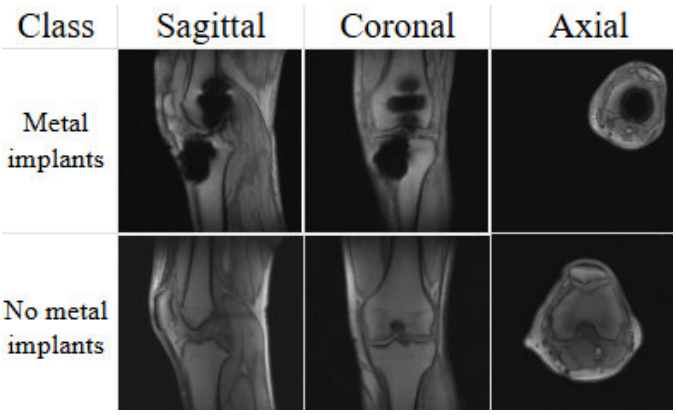


Fig. 2. Representative knee MRI localizer scans illustrating the two classes: images with metallic implants (top row) and those without artifacts (bottom row) across the three primary anatomical planes.

B. Baseline Model

The Residual Network (ResNet) architecture was selected as the baseline model for this study due to its proven efficiency in medical image analysis and its ability to mitigate the vanishing gradient problem via skip connections (or residual blocks). In clinical environments, where computational speed and real-time inference are critical (especially during the "scout" or localizer phase), ResNet18 offers a significant advantage over deeper variants (like ResNet50 or ResNet101) [21]. This maintains a high level of feature extraction capability with a relatively small number of parameters, making it an ideal candidate for a lightweight clinical screening tool. In this work, we use Transfer Learning from the ImageNet dataset. Instead of training the network from scratch with randomized weights, we initialized the model with pre-trained weights. This approach is highly effective for medical imaging tasks with limited datasets, as the model can leverage low-level visual features (edges, textures, and gradients) learned from millions of general images and adapt them to the specific morphology of knee MRI scans [22].

The training process involved a fine-tuning strategy where the final fully connected layer of the original ResNet18 was replaced to match our binary classification task ("Metal" vs. "No Metal"). To implement this workflow, all input images were resized to 512×512 pixels to maintain consistency with the localizer scan resolution. The model was trained using the Adam optimizer with a learning rate of 0.001, paired with

Cross-Entropy Loss to quantify the difference between the predicted class probabilities and the ground truth. Finally, to enhance model robustness and ensure the input data distribution matched the pre-trained model's expectations, standard data transformations were applied, including normalization based on the ImageNet mean and standard deviation.

C. Object Detection Model

The integration of the YOLO framework into the early screening phase aligns with recent shifts toward highly efficient clinical decision support systems. By framing object detection as a single regression problem, YOLOv8 enables simultaneous classification and bounding box prediction in a single forward pass. This fast processing capability is highly substantiated by recent clinical literature, where the YOLO series has demonstrated high performance and classification robustness across diverse medical screening tasks [23].

For this study, the YOLOv8n (Nano) variant was selected to ensure rapid inference, making it an ideal candidate for immediate artifact identification. To leverage this architecture's precise spatial localization capabilities, a specialized manual labeling phase was conducted using the Roboflow platform. Unlike simple image-level labeling, defining the exact coordinates of the metallic interference forces the model to learn features specific to the artifact rather than peripheral anatomical structures. These annotations were then exported directly in the native YOLOv8 format (containing class indices and normalized bounding box coordinates) to guide the supervised learning process.

Based on the provided workflow, the implementation details include:

- **Image Pre-processing:** Images were standardized to a 512×512 resolution to align with the ResNet18 baseline and the native resolution of the MRI localizer scans.
- **Training Hyperparameters:** The model was trained for 100 epochs with a batch size of 16. The training utilized the SGD optimizer with a learning rate of 0.01 and momentum of 0.937.
- **Loss Function:** YOLOv8 utilizes a combination of CIoU (Complete Intersection over Union) and DFL (Distribution Focal Loss) for bounding box regression, alongside Binary Cross-Entropy for class prediction.
- **Transfer Learning:** Similar to the baseline, the model was initialized with pre-trained weights from the COCO dataset [24], allowing the network to benefit from established feature hierarchies.

D. Evaluating Metrics

To quantitatively assess and compare the performance of the baseline ResNet18 and the proposed YOLOv8 models, five primary metrics were calculated. These metrics are defined in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Accuracy, Precision, Recall, F1-Score, and mean Intersection over Union (mIoU) [25] were the metrics employed in this work. In particular, the mIoU is essential for evaluating the localization accuracy of the YOLOv8 model. It measures the average overlap between the predicted bounding box (B_p) and the ground truth box (B_g) across all instances:

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_g)}{\text{Area}(B_p \cup B_g)}, \tag{1}$$

$$\text{mIoU} = \frac{1}{n} \sum_{i=1}^{n} \text{IoU}_i. \tag{2}$$

For ResNet18, metrics were calculated using the standard `scikit-learn` classification report [26]. For YOLOv8, metrics were derived from the validation/test output provided by the `ultralytics` framework.

To evaluate the detection performance of YOLOv8, the Mean Average Precision (mAP) [24] was utilized. This metric considers both the precision of the classification and the localization accuracy of the predicted bounding boxes at specific Intersection over Union (IoU) [25] thresholds.

1) **mAP@0.5**: This represents the mean Average Precision calculated at a fixed IoU threshold of 0.5. It measures the model's ability to correctly identify and locate objects where at least 50% of the predicted box overlaps with the ground truth. It is a measure of the model's overall detection strength.

2) **mAP@0.5:0.95**: This is a more stringent metric that represents the average mAP calculated over a range of IoU thresholds, from 0.5 to 0.95, with a step size of 0.05. This metric rewards high-precision localization; a high score here indicates that the model's predicted bounding boxes align very closely with the actual boundaries of the metallic artifacts:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i, \tag{3}$$

where N is the number of classes and AP is the Area Under the Precision-Recall Curve for each class.

III. Results

This section presents the experimental results and a comparative analysis of the ResNet18 and YOLOv8 architectures. The evaluation is organized into four stages. First, the models are compared using standard classification metrics and statistical validation. Second, a qualitative analysis is performed to visualize the models' predictive behavior. Third, both models are evaluated for robustness to spike artifacts. Finally, a series of stress tests—including a noise-sensitivity analysis and dataset-size ablation—is conducted to assess model robustness under varying clinical and data-availability constraints.

A. General Results

The quantitative performance of both models on the test set is summarized in Table II. These results reflect the models' ability to generalize to unseen data after training on a subject-wise partitioned dataset. A statistical validation via patient-level bootstrapping is conducted to establish 95% confidence intervals, ensuring results are not biased by specific patient characteristics.

Using a standardized resolution of 512×512 pixels, YOLOv8 demonstrated superior performance across all metrics. The statistical validation shows that while ResNet18 maintains high sensitivity (95.7%), it has a wider confidence interval and lower specificity than the object detection approach.

The perfect specificity (100%) and high accuracy (99.4%) achieved by YOLOv8 confirm its robustness in identifying metallic artifacts while eliminating false positives. In a clinical environment, this precision is essential to minimize "false alarms" that could disrupt radiological workflows. Furthermore, the tight confidence intervals for YOLOv8 suggest more stable performance across different patient profiles than the ResNet18 baseline.

1) Qualitative and Comparative Analysis: Gradient-weighted Class Activation Mapping (Grad-CAM) is a prominent technique for generating visual explanations for decisions made by a wide array of CNN-based models, thereby enhancing their transparency [27]. By using gradients from the final convolutional layer, Grad-CAM produces a coarse localization map that highlights the most salient regions for a given prediction. In the context of image classification, this approach provides critical insights into the failure modes of DL models, allowing us to identify whether a network is focusing on relevant anatomical features or extraneous artifacts [27].

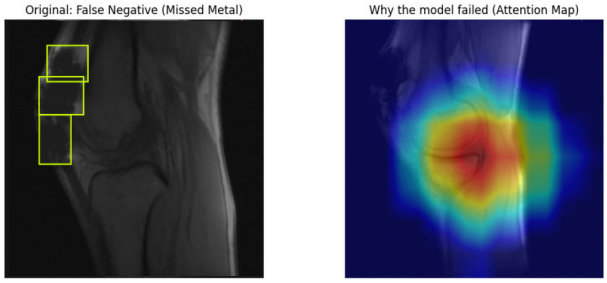


Fig. 3. Grad-CAM visualization of a false negative detection by YOLOv8.

The confusion matrices representing the mean performance within the calculated confidence interval for YOLOv8 and ResNet18 are shown in Table III. Inspecting the one false negative (missed metal) for YOLOv8, the Grad-CAM illustrated in Fig. 3 shows that the model's attention was concentrated on an anatomical region distant from the actual metallic implant. In this instance, the metal was located in the patellar region; however, the resulting signal void (the characteristic "black area" or susceptibility artifact) merged with the image background. This loss of boundary definition likely caused the feature extraction layers to overlook the artifact.

Figure 4 illustrates representative false negative and false positive results for ResNet18. The Grad-CAM heatmap for the false negative indicates that ResNet18 exhibits a misinterpretation similar to that observed in YOLOv8. Specifically, the presence of metal in the patellar region appears to be a consistent source of error for both models; however, while YOLOv8 produced only one false negative for this case, ResNet18 yielded six.

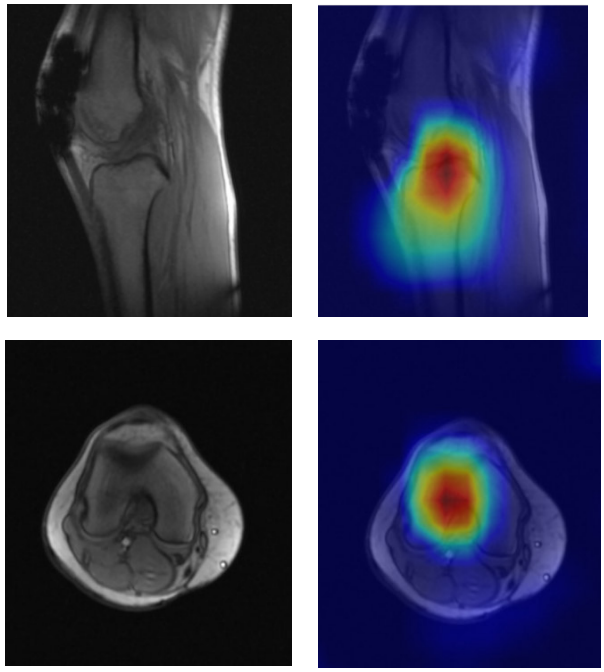


Fig. 4. Grad-CAM visualizations for ResNet18 illustrate a false negative in the upper image pair and a false positive in the lower pair.

2) Localization and Detection Performance: A critical advantage of the YOLOv8 architecture is its inherent spatial localization capability. To evaluate this performance, we utilized the mIoU. The proposed YOLOv8 model achieved a mean mIoU of 0.7021. This result indicates that while the model effectively detects the presence of metal, there is room for optimization in precisely defining the spatial boundaries of susceptibility artifacts, which often lack distinct edges in MRI. Fig. 5 compares instances of low and high mIoU performance. Notably, the annotation process—utilizing square bounding boxes to encompass artifact areas—may introduce slight discrepancies relative to the irregular morphology of the artifacts, potentially contributing to a lower mIoU. While the mIoU reflects limitations in precise localization, it does not undermine the overall efficacy of the YOLOv8 model. Given that the primary objective of this study is the classification of metal artifacts, the model demonstrates high diagnostic performance and reliability for its intended clinical application.

The YOLOv8 model's performance was evaluated through a combination of spatial and precision-based metrics. The model achieved a **mIoU of 0.7021**, indicating high localization accuracy. An average overlap of 70% is particularly significant in MRI analysis, where susceptibility artifacts often exhibit irregular morphologies and diffuse edges that challenge precise alignment of bounding boxes.

Detection robustness was further evidenced by an

TABLE II
COMPARATIVE PERFORMANCE: RESNET18 VS. YOLOV8 (PATIENT-LEVEL BOOTSTRAP VALIDATION)

Model	Accuracy [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]
ResNet18	95.1% [90.5% – 98.9%]	95.7% [87.0% – 100%]	94.4% [90.0% – 98.4%]
YOLOv8	**99.4% [98.1% – 100%]**	**98.9% [96.0% – 100%]**	**100.0% [100% – 100%]**

TABLE III
CONFUSION MATRICES FOR RESNET18 AND YOLOV8.

Model	True label	Pred. Metal	Pred. No Metal
ResNet18	Metal	83	6
	No Metal	7	82
YOLOv8	Metal	86	1
	No Metal	0	91

mAP@0.5 of 0.5646, reflecting a stable balance between precision and recall at a standard detection threshold. While the **mAP@0.5:0.95** decreased to **0.3016**, this drop is expected as it demands near-perfect spatial alignment (IoU > 0.85). In a clinical context, the consistency between the mIoU (0.70) and mAP@0.5 suggests that the model reliably identifies and approximates the boundaries of metallic artifacts. Although square annotations may introduce slight discrepancies against the artifacts' irregular shapes, the results confirm the model's efficacy for diagnostic screening, where reliable identification is the primary objective.

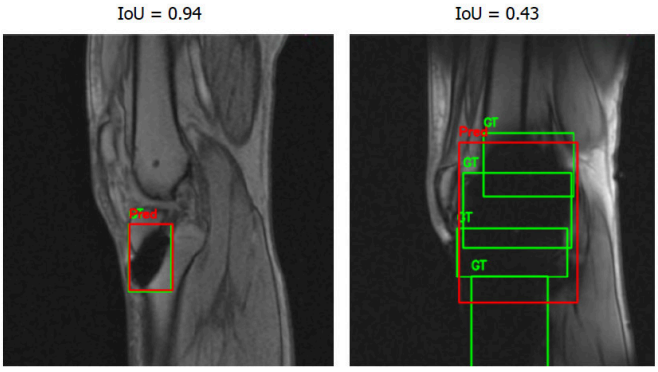


Fig. 5. Representative cases of YOLOv8 bounding box predictions.

B. Noise Analysis

The Signal-to-Noise Ratio (SNR) in MRI is fundamentally proportional to the voxel volume and the square root of the acquisition time [28]. To simulate the inherent variability of clinical acquisitions—where SNR fluctuates with field strength (B_0), coil geometry, and parallel imaging—a noise-sensitivity analysis was conducted. Gaussian noise was incrementally added to the test set to identify the break point for each architecture.

By monitoring the decay in Accuracy as a function of the SNR, we determined which backbone maintains higher reliability under degraded image conditions. Both architectures were evaluated across five SNR levels ($40, 30, 20, 10$, and 5), representing a continuum from high-fidelity 1.5T/3T imaging to sub-diagnostic, noise-limited scenarios, as discussed in [29]. The results are presented in Fig. 6. As observed in the experimental data, YOLOv8 demonstrates near-perfect performance in high-quality scenarios, maintaining an accuracy of 98% or higher for SNR ≥ 20. However, its performance drops sharply in high-noise environments, falling to 67% at an SNR of 5. In contrast, ResNet18 exhibits remarkable stability; although its peak performance is lower than YOLOv8's, it maintains an accuracy of 90% even under extreme noise conditions (SNR 5). This divergence suggests a trade-off between the precision of local object detection (YOLOv8) and the robustness of global feature extraction (ResNet18). In clinical settings where MRI hardware limitations or patient motion may induce significant artifacts, the structural consistency of the ResNet18 backbone offers a more dependable baseline, whereas YOLOv8 is the optimal choice for high-fidelity diagnostic environments.

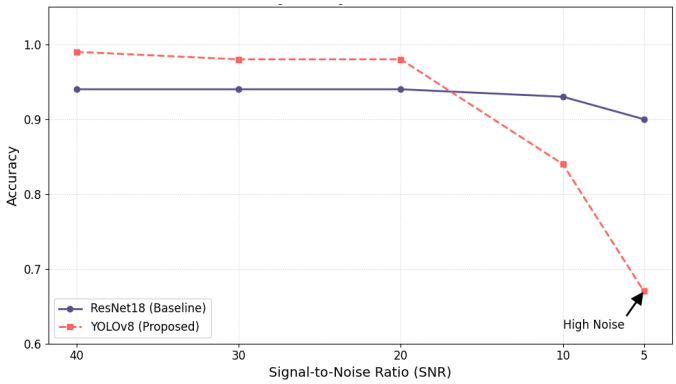


Fig. 6. Noise Sensitivity Analysis: ResNet18 vs. YOLOv8

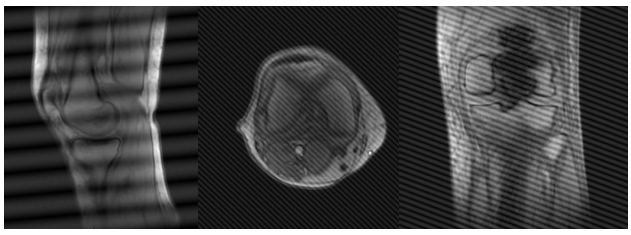


Fig. 7. Examples of simulated spike artifacts. The variations in band angulation and spacing correspond to different spatial frequency coordinates within the k-space matrix.

C. Robustness to Spike Artifacts

The spike test is justified by the unique nature of MR data acquisition, in which the final image is a direct product of the Fourier transform of the k-space matrix. As established in [30], electromagnetic spikes—caused by gradient duty cycles or external electronic interference—introduce bad data points into this raw data frequency domain. Because each point in k-space contributes to the entire reconstructed image, a single spike results in a global herringbone or criss-cross artifact that obscures anatomical structures with dark, periodic stripes as shown in Fig. 7.

To assess the models' ability to handle additional types of MRI artifacts, a simulation was conducted by artificially introducing spike artifacts into test images via k-space manipulation. The models were evaluated on test sets fully corrupted with spike artifacts, while the proportion of spike-contaminated images in the training set was gradually increased. In this setting, YOLOv8 achieved 97% accuracy when trained with 25% corrupted images, whereas ResNet18 achieved only 66% under the same conditions. These results indicate that YOLOv8 exhibits superior robustness to alternative artifact types.

D. Dataset Size Analysis

To investigate the data efficiency of each architecture, an ablation study was conducted by training the models on varying subsets of the original data (25%, 50%, 75%, and 100%). This analysis identifies the minimum data threshold required for each model to achieve a clinically acceptable performance level and assesses how each architecture scales with increasing sample sizes. The resulting performance for ResNet18 and YOLOv8 is illustrated in Fig. 8.

The results indicate that YOLOv8 consistently maintains a performance advantage across all dataset fractions. Notably, YOLOv8 achieves approximately 93% accuracy with only 10% of the training data, eventually plateauing near 99% as the dataset size increases. ResNet18 follows a similar learning trajectory but starts at a lower baseline of 84% accuracy at the 10% fraction. Both models exhibit strong convergence, reaching over 90% accuracy when at least 50% of the curated 1,254 images are utilized. This suggests that while both architectures are efficient, YOLOv8 is better at feature extraction with extremely limited sample sizes.

IV. Conclusions, Limitations, and Future Work

This pilot study demonstrated the feasibility of implementing an automated screening system for metallic implants during the initial "scout" phase of knee MRI workflows. By utilizing DL architectures to identify artifacts at the earliest possible stage, clinical teams can preemptively adjust protocols to include metal artifact reduction sequences, thereby reducing the need for costly re-scans and optimizing scanner throughput. The comparative analysis further highlights a clear distinction between the two architectural paradigms evaluated:

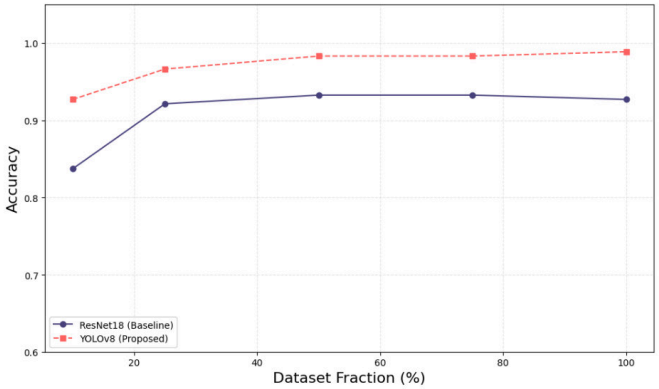


Fig. 8. Model Performance Comparison: Accuracy vs Dataset Size

- **Detection Superiority:** YOLOv8 proved to be the more robust choice for high-fidelity environments, achieving near-perfect accuracy.
- **Structural Robustness:** Although ResNet18 demonstrated greater stability under degraded conditions, it exhibited lower peak performance and failed to maintain satisfactory accuracy in the spike artifact evaluation.
- **Interpretability and Localization:** The integration of Grad-CAM visualization and mIoU metrics provided valuable qualitative insights into model behavior. YOLOv8 achieved an mIoU of 0.7021, highlighting the importance of incorporating spatial context in artifact detection. This spatial awareness is particularly relevant for future developments in metal artifact analysis, as localizing metallic interference within the anatomical context enables a more nuanced understanding of image degradation and its clinical implications.

While these initial results are promising, several limitations inherent to this pilot study must be acknowledged. First, the evaluation was conducted on a single-center cohort of 200 patients (1,254 images). Although a subject-wise split and patient-level bootstrapping with 95% confidence intervals were strictly enforced to prevent data leakage and ensure statistical validity, testing on larger, multi-center datasets is required to further confirm clinical generalizability. Second, the dataset was constructed using localizer scans from only two vendors (Siemens Healthineers and GE Healthcare). While scout sequences are standardized, hardware-specific variations in reconstruction kernels and signal handling could introduce subtle biases.

Third, regarding localization, the mIoU achieved by YOLOv8 was 0.70. While sufficient for identifying the region of interest, bounding box alignment is inherently limited by the irregular and diffuse boundaries of magnetic susceptibility artifacts. Precise spatial delineation was secondary to classification accuracy, as the primary clinical objective was triggering protocol adjustments. Finally, the noise sensitivity analysis applied additive Gaussian noise to reconstructed magnitude images as a standardized computer vision stress test. However, noise in magnitude MRI physically follows a Rician

distribution, particularly in low-signal and background regions [31].

Future work will focus on both technical refinement and clinical integration. Subsequent studies will evaluate physically realistic Rician noise models and complex k-space perturbations prior to magnitude reconstruction. Additionally, exploring semantic segmentation architectures and specialized loss functions may better capture irregular artifact morphology. Crucially, as a core contribution of this pilot study, the newly curated **UTFPR-Knee-MRI-Metal-Scout** dataset has been made publicly available to serve as a benchmark for automated MRI quality control research. Ultimately, embedding these algorithms into existing scanner consoles will transition this framework into a routine real-time clinical tool.

REFERENCES

[1] N. K. Bangerter, G. Morrell, and M. Grech-Sollars, "Chapter 2.3 - magnetic resonance imaging," in *Bioengineering Innovative Solutions for Cancer*, S. Ladame and J. Y. Chang, Eds. Academic Press, 2020, pp. 163–194.

[2] D. Dowsett, P. A. Kenny, and R. E. Johnston, *The Physics of Diagnostic Imaging*, 2nd ed. CRC Press, 2006.

[3] R. Smith-Bindman, M. L. Kwan, E. C. Marlow, M. K. Theis, W. Bolch, S. Y. Cheng, E. J. A. Bowles, J. R. Duncan, R. T. Greenlee, L. H. Kushi, J. D. Pole, A. K. Rahm, N. K. Stout, S. Weinmann, and D. L. Miglioretti, "Trends in use of medical imaging in us health care systems and in ontario, canada, 2000-2016," *JAMA*, vol. 322, no. 9, pp. 843–856, 09 2019.

[4] C. Alencar, D. Oliveira, A. Teixeira, L. Lemos, R. Quesado, I. Santos, and C. Lins, "Tomografia computadorizada e ressonância magnética no brasil: estudo epidemiológico sobre distribuição dos equipamentos, frequência de realização dos exames e comparação entre setores público e privado," *Radiologia Brasileira*, vol. 57, 05 2024.

[5] S. J. Fransen1, C. Roest, F. F. J. Simonis, D. Yakar1, and T. C. Kwee, "The scientific evidence of commercial ai products for mri acceleration: a systematic review," *European Radiology*, 2025/02/19 2025.

[6] T. Budrys, V. Veikutis, S. Lukosevicius, R. Gleizniene, E. Monastyreckiene, and I. Kulakiene, "Artifacts in magnetic resonance imaging: how it can really affect diagnostic image quality and confuse clinical diagnosis?" *Journal of Vibroengineering*, vol. 20, no. 2, pp. 1202–1213, mar 2018.

[7] J. B. Andre, B. W. Bresnahan, M. Mossa-Basha, M. N. Hoff, C. P. Smith, Y. Anzai, and W. A. Cohen, "Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical mr examinations," *Journal of the American College of Radiology*, vol. 12, no. 7, pp. 689–695, 2015.

[8] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 939–954, 2019.

[9] R. Gao, G. Luo, R. Ding, B. Yang, and H. Sun, "A lightweight deep learning framework for automatic mri data sorting and artifacts detection," *Journal of Medical Systems*, vol. 47, 11 2023.

[10] N. Ettehadi, P. Kashyap, X. Zhang, Y. Wang, D. Semanek, K. Desai, J. Guo, J. Posner, and A. F. Laine, "Automated multiclass artifact detection in diffusion mri volumes via 3d residual squeeze-and-excitation convolutional neural networks," *Frontiers in Human Neuroscience*, vol. 16, 3 2022.

[11] S. Loizillon, S. Bottani, A. Maire, S. Ströer, L. Chougar, D. Dormont, O. Colliot, and N. Burgos, "Automatic quality control of brain 3d flair mris for a clinical data warehouse," *Medical Image Analysis*, vol. 103, 7 2025.

[12] M. Courtman, D. Kim, H. Wit, H. Wang, L. Sun, E. Ifeachor, S. Mullin, and M. Thurston, "Deep learning detection of aneurysm clips for magnetic resonance imaging safety," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, pp. 72–80, Feb 2024. [Online]. Available: https://doi.org/10.1007/s10278-023-00932-8

[13] S. O. Lie and S. Lysdahlgaard, "Detection of metallic objects on digital radiographs with convolutional neural networks: A MRI screening tool," *Radiography*, vol. 28, no. 4, pp. 466–472, November 2022.

[14] B. A. Hargreaves, P. W. Worters, K. B. Pauly, J. M. Pauly, K. M. Koch, and G. E. Gold, "Metal-induced artifacts in mri," *American Journal of Roentgenology*, vol. 197, no. 3, pp. 547–555, 2011, pMID: 21862795.

[15] J. P. Dillenseger, S. Molière, P. Choquet, C. Goetz, M. Ehlinger, and G. Bierry, "An illustrative review to understand and manage metal-induced artifacts in musculoskeletal MRI: a primer and updates," *Skeletal Radiology*, vol. 45, no. 5, pp. 677–688, May 2016.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[18] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, "3d deep learning on medical images: A review," *Sensors*, vol. 20, no. 18, p. 5097, 2020.

[19] B. A. Hargreaves, P. W. Worters, K. B. Pauly, J. M. Pauly, K. M. Koch, and G. E. Gold, "Metal-induced artifacts in mri," *American Journal of Roentgenology*, vol. 197, no. 3, pp. 547–555, 2011.

[20] Y.-T. Tsan, C.-H. Lin, J. Y.-C. Jiang, H.-C. Chien, Y.-T. Chen, and J.-C. Liu, "Automatic classification of medical image modality and anatomical location using convolutional neural network," *PLOS ONE*, vol. 16, no. 6, p. e0253205, 2021.

[21] W. Xu, Y.-L. Fu, and D. Zhu, "Resnet and its application to medical image processing: Research progress and challenges," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107660, 2023.

[22] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzębski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184–1194, 2020.

[23] Z. Cai, K. Zhou, and Z. Liao, "A systematic review of yolo-based object detection in medical imaging: Advances, challenges, and future directions," *Computers, Materials & Continua*, vol. 85, no. 2, pp. 2255–2303, 2025.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*. Springer, Cham, 2014, pp. 740–755.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. p. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research (2011)*, 2012.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[28] A. Macovski, "Noise in mri," *Magnetic Resonance in Medicine*, vol. 36, no. 3, pp. 494–497, 1996.

[29] S. O. Schoenberg, O. Dietrich, and M. F. Reiser, "Effect of field strength on MR images: Comparison of the same subject at 0.5, 1.0, and 1.5 T," *Radiology*, vol. 212, no. 1, pp. 253–264, Jul 1999.

[30] D. Moratal, A. Vallés-Luch, L. Martí-Bonmatí, and M. Brummer, "k-space tutorial: an MRI educational tool for a better understanding of k-space," *Biomedical Imaging and Intervention Journal*, vol. 4, no. 1, p. e15, 2008.

[31] H. Gudbjartsson and S. Patz, "The rician distribution of noisy mri data," *Magnetic Resonance in Medicine*, vol. 34, no. 6, pp. 910–914, 1995.

Automated Visual Target Detection in RSVP Tasks from EEG Signals using Wavelet Transform and Machine Learning

Júlia Biscontini Koury
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
julia.bkoury@ufpe.br

Ana Beatriz de Souza Sampaio
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
anabeatriz.sampaio@ufpe.br

João Pedro Cabral Bezerra Cavalcanti
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
joao.cbcavalcanti@ufpe.br

João Gonçalves de Oliveira Neto
Biomedical Engineering Dept.
Federal University of Pernambuco)
Recife, Brazil
joao.goncalves@ufpe.br

Sandrelly Mirely da Silva Santos
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
sandrelly.santos@ufpe.br

Wellington Pinheiro dos Santos
Biomedical Engineering Dept.
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

***Abstract*—This work presents an approach for automated detection of visual targets in RSVP tasks using EEG signals. The methodology combines the Wavelet Continuous Transform with the use of the LeNet convolutional neural network. Signals were acquired from 7 participants undergoing RSVP tasks with images of airplanes as target stimuli. Each signal was segmented into 500 ms windows and subjected to the transform, generating threedimensional representations (8 channels, 40 frequencies, 1024 samples) that preserve spatial, spectral and temporal information. These representations were then used as the basis for the classification step. Keywords - Targe**

***Index Terms*—Target detection, machine learning, fast visual serial presentation, RSVP, P300, Wavelet transform, EEG, random forest.**

I. Introduction

The electroencephalogram (EEG) is a valuable tool for studying brain activity during cognitive tasks, especially because of its high temporal resolution and non-invasiveness. In the context of continuous monitoring of EEG signals in clinical and experimental settings, the generation and analysis of large volumes of data face some challenges: manual processes are prone to operational errors, excessive time periods, and are subject to inter and intra-operator variability [1]. These limitations motivate the development of automated methods for detecting specific neural patterns, especially in experimental paradigms that require fast and accurate responses.

A promising technique in this scenario is Rapid Serial Visual Presentation (RSVP), in which images are exhibited at high speed (typically in the 5-10 Hz range) while electroencephalographic activity is simultaneously and continuously recorded [2], [3]. RSVP exploits the human visual system's ability to process stimuli sequentially in fractions of a second, generating specific neural responses when a relevant stimulus (target) is detected. Among these responses, the P300 component stands out, an event-related evoked potential that typically occurs between 300 and 500 ms after the presentation of an infrequent or significant stimulus [4]. The detection of the P300 in RSVP tasks has direct applications in brain-machine interfaces, assisted surveillance systems, and rapid image screening [5].

However, the automated interpretation of EEG signals in RSVP paradigms presents significant challenges. The signals are inherently non-stationary, low amplitude (microvolts), and can be contaminated by biological artifacts (eye movements, muscle activity) and external noise. Additionally, inter-subject variability and dependence on the participant's attentional status make it difficult to generalize models on the temporal domain [6]. In this context, the simultaneous analysis of the temporal and spectral dimensions is essential to realize the neural dynamics associated with target detection.

The Continuous Wavelet Transform (CWT) can be used as a mathematical tool for processing non-stationary signals such as EEG. Unlike the Short-Duration Fourier Transform, which uses a fixed window, CWT employs wavelets that dilate and contract, providing fine temporal resolution for high-frequency components and enhanced spectral resolution for low-frequency components [7]. This property allows it to represent the EEG signal in the time-frequency domain, which preserves both fast transients (such as P300) and rhythmic background oscillations.

Recent work has shown that combining time-frequency representations with machine learning techniques can significantly improve the accuracy when detecting neural patterns in cognitive tasks [8], [9]. However, most existing approaches treat EEG channels independently, disregarding the spatial relationships between electrodes and the underlying cortical topography. The preservation of spatial information is impor-

tant, especially in visual tasks that recruit occipital and parietal areas in a differentiated way.

In this context, the present work aims to develop and evaluate a methodology for time-frequency-spatial representation of multichannel EEG signal detection of visual targets in RSVP tasks. The proposed approach consists of: (1) segmentation of the signals in 500 ms windows synchronized with the visual stimuli; (2) application of the Continuous Wavelet Transform to each channel individually, generating time-frequency spectrograms; (3) organized stacking of these spectrograms in a three-dimensional tensor (channels × frequencies × time) that preserves the cortical topography; and (4) validation of the discriminative capacity of these representations through a Random Forest classifier implemented in the WEKA environment.

II. Methodology

A. Acquisition and Selection of Data

The public database EEG Signals from an RSVP Task [4], available on the PhysioNet platform, was used, containing records for 11 participants submitted to a Rapid Visual Serial Presentation (RSVP) paradigm. The stimuli consist of aerial photographs of London, with target images containing a photo-realistically superimposed plane and non-target images without planes.

Seven participants (IDs: 2, 3, 4, 8, 10, 11, 13) were selected with complete records for all experimental conditions: three presentation frequencies (5 Hz, 6 Hz, and 10 Hz) and two collection sessions per condition (suffixes ”a” and ”b”), totaling 42 files in EDF format. The EEG signals were acquired at 2048 Hz, with original filtering between 0.15 and 28 Hz, from 8 channels positioned according to the 10-20 system [5] in the occipital and parietal regions: PO8, PO7, PO3, PO4, P7, P8, O1, and O2. These channels correspond to the primary areas of the visual cortex and are strategically chosen for their relevance in detecting visual stimuli. The annotations in the files indicate the beginning of each image and its class: T=0 (non-target/nonairplane images) or T=1 (target/images containing an airplane in a sideways position).Because the dataset used in this study is publicly available and contains previously collected and identified data, no new data were collected from human participants. Therefore, ethical committee approval was not required for the present analysis.

B. Pre-processing of Signals

The files in the EDF format were processed using the MNE-Python library. Initially, the 8 channels of the original data set were selected in the occipital and posterior parietal regions (PO8, PO7, PO3, PO4, P7, P8, O1, and O2). The order of the electrodes was kept fixed throughout the study, preserving the topographic relationships intrinsic to the data and allowing later steps to explore spatial patterns of neural activation associated with target detection.

Although the signals had already been pre-filtered by the base authors in the range of 0.15–28 Hz, additional bandpass filtering from 1 to 30 Hz was applied using a 4th-order digital Butterworth filter. The lower limit of 1 Hz aims to remove very low-frequency fluctuations and residual motion artifacts, while the upper limit of 30 Hz ensures the complete inclusion of the beta band, which is relevant in cognitive sensing processes.

In sequence, the signals were resampled from 2048 Hz to 250 Hz; this reduction balances the need for sufficient temporal resolution to capture the P300 potential, which typically occurs between 300 and 500 ms after the stimulus. With the new sampling rate, each 500 ms time window now contains 125 samples.

C. Segmentation in Time Frames

To investigate the neural responses evoked by visual stimuli, the continuous EEG signals were segmented into 500 ms time windows, aligned with the onset of each image presented. The adoption of this interval is based on the need to capture the P300 evoked potential, a component associated with the detection of relevant stimuli, which typically manifests between 300 and 500 ms post-stimulus.

The segmentation was guided by the annotations provided in the original files, with discrimination of the precise instant of occurrence and its class (T=0 or T=1). For each event, a segment of 125 samples was extracted, corresponding to the 500 ms of interest and the sampling rate of 250 Hz, which was established in the previous step.

Each resulting window consists of a two-dimensional matrix of dimensions 8 × 125, where the rows represent the EEG channels in the fixed spatial order and the columns represent the temporal evolution of the signal. Each matrix was assigned a binary label: 0 for non-target and 1 for target, according to the original annotation.

The 500 ms window balances the range of neural response with minimal overlap between consecutive stimuli, especially at 10 Hz (stimuli every 100 ms). Segments with significant residual artifacts were discarded, guaranteeing data quality for subsequent steps. At the end of this stage, a dataset with format (14054, 8, 125) was obtained, comprising 14,054 epochs, of which 12,656 are non-target, and 1,398 are target.

D. Continuous Wavelet Transform and Imaging

To capture the neural dynamics associated with visual target detection, each EEG time window was subjected to the Continuous Wavelet Transform (CWT). This technique was chosen for its ability to analyze non-stationary signals, offering fine temporal resolution for high-frequency components and detailed spectral resolution for low-frequency components, being particularly suitable for evoked potentials such as the P300 [6].

The Morlet mother wavelet (a Gaussian function modulated by a sine wave) with a center frequency of 5 Hz was used, a value that balances temporal and frequency precision. The transform was applied to each of the 8 channels individually using the PyWavelets library with 64 scales, covering the neurophysiological bands of interest: delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), and beta (13-30 Hz). For each channel, a matrix of coefficients of dimensions 64 × 125 was obtained, where the lines correspond to the scales (frequencies) and

the columns to the temporal instants (0 to 500 ms). Spectral power was calculated as the square of the magnitude of the coefficients.

From these matrices, a visual representation was generated for each epoch: a figure with 8 vertical subplots (one per channel, in the fixed order of the electrodes), each containing the timefrequency scalogram of the power of the respective channel. The axes were configured with a time of 0 to 500 ms and a frequency of 1 to 30 Hz, using the 'jet' color scale to represent the intensity of the power. The images were saved in PNG format (150 dpi, approximately 600×800 pixels) with standardized nomenclature "(pacienteIDfreqHzsessaoa/bepochnumber:04d target/nontarget.png)", all stored in a single directory.

E. Export of Data for Classification in WEKA

The trait vectors extracted in the previous step, composed of 8 values (averages per channel) and their respective labels, were organized for processing in the WEKA (Waikato Environment for Knowledge Analysis) software. To this end, a file was generated in the ARFF (Attribute-Relation File Format) format, the standard used by the tool.

The file was structured as follows:

1) Header: `@relation EEG_CWT_From_Images`
2) Attributes: eight numerical attributes corresponding to the averages of each channel, named mean_PO8, mean_PO7,`mean_PO3`, `mean_PO4`, `mean_P7`, `mean_P8`, `mean_O1`, and mean_O2
3) Class: nominal attribute class with the values 0,1, where 0 represents non-target and 1 represents target
4) Data: each sample occupied a line, with the eight numerical values formatted with six decimal places, separated by a comma, followed by the class label.

The resulting file was named eeg_cwt_from_images.arff and saved for later loading in the WEKA Explorer interface, where it was classified.

F. Balancing, feature extraction, and grading

The CWT image representations generated from the EEG epochs were loaded into the WEKA Explorer environment for subsequent class balancing, feature extraction, and classification.

The original dataset presented a substantial class imbalance, consisting of 1,398 target samples and 12,656 non-target samples, totaling 14,054 samples. To investigate the effect of class distribution on classification performance, an undersampling strategy was adopted. Two experimental configurations were evaluated. In the first configuration, a 1:1 class ratio was used, resulting in 1,678 samples, with 839 target and 839 non-target instances. In the second configuration, a 2:1 ratio was used, resulting in 2,517 samples, with 839 target and 1,678 non-target instances.

After the undersampling procedure, the resulting CWT representations were processed using an implementation of the LeNet architecture available in the WEKA environment. LeNet was used as a feature extraction method, generating 500 numerical attributes for each image. These attributes provided a compact representation of the information contained in the CWT image pixels and were subsequently used as input for the classification stage. The resulting feature dataset was saved for further processing in WEKA.

The resulting feature dataset was submitted to classification using two algorithms available in WEKA: Random Forest and SMO (Sequential Minimal Optimization). Random Forest was configured with 100 trees (other parameters were kept at default value), while SMO was configured with a polynomial kernel with exponent 2.0 and parameter C=1.0. Both classifiers were evaluated using 10-fold cross-validation. The performance of the models was assessed using accuracy, Kappa statistic, mean absolute error, and root mean squared error.

III. Results and Discussion

After applying the processing pipeline described in Section II, including filtering, segmentation, and Continuous Wavelet Transform, 14,054 images in PNG format were extracted, each representing a 500 ms epoch of the EEG signals. Of these, 12,656 correspond to non-target stimuli and 1,398 to target stimuli. The images include the 7 selected participants, the three presentation frequencies (5 Hz, 6 Hz, and 10 Hz), and the two collection sessions by condition. Fig. 1 illustrates examples generated for a target and a non-target stimulus, where it is possible to observe differences in activation patterns, especially in the occipital regions (O1 and O2 channels).

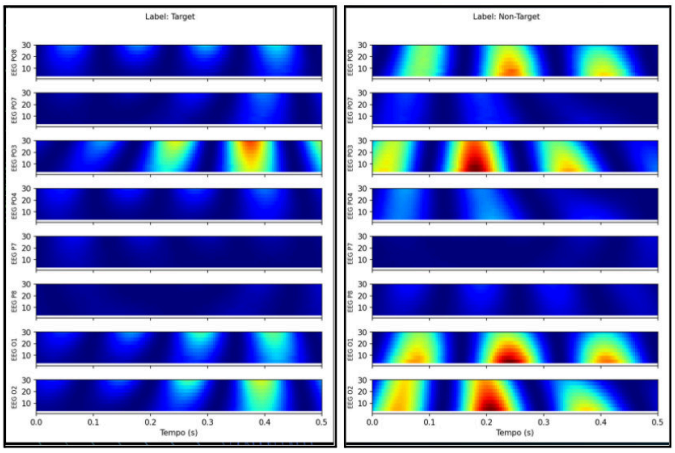


Fig. 1. Examples of visual representations generated by CWT. Non-target stimulus; target stimulus.

The original dataset exhibited a marked class imbalance, as previouly mentined in the methodology, which is characteristic of similar RSVP tasks, These tasks commonly have less frequent target stimuli appearances in comparison to non-target stimuli. To mitigate the potential bias of the classifiers toward the majority class (non-target) and evaluate their performance under different class distributions, an undersampling strategy was adopted. Two experimental scenarios were defined based on this strategy: a 1:1 class ratio, comprising 839 target and 839 non-target samples, for a total of 1,678 instances; and

a 2:1 class ratio, comprising 839 target and 1,678 non-target samples, for a total of 2,517 instances.

After extracting characteristics from the images through an implementation of the LeNet architecture available in the WEKA environment, generating a set of 500 attributes representative of its visual content, the data were submitted to classification. Two different algorithms were evaluated using cross-validation with 10 Folds: o Random Forest (configured with 100 trees) and the SMO (Support Vector Machine with PolyKernel of exponent 2.0 and parameter C=1.0). The comparative results obtained are consolidated in Table I.

TABLE I
COMPARATIVE CLASSIFICATION RESULTS WITH CROSS-VALIDATION

Balancing	**Classifier**	Accuracy	Kappa Index	Average Absolute Error	Mean Square Error
Subsample 1:1	Random Forest (100 trees)	59,85%	0,1979	0,4784	0,4850
Subsample 2:1	Random Forest (100 trees)	67,62%	0,0549	0,4274	0,4593
Subsample 1:1	SMO (PolyKernel E=2.0)	93,66%	0,8737	0,0632	0,2513
Subsample 2:1	SMO (PolyKernel E=2.0)	93,29%	0,8497	0,0671	0,2591

The results indicate a marked difference in the performance of the evaluated algorithms. The Random Forest classifier presented limitations when applied to the attributes extracted by LeNet from the subsample balancing. In the 1:1 scenario, it achieved an accuracy of only 59.85%. In the 2:1 scenario, the nominal accuracy rose to 67.62%; however, the corresponding confusion matrix revealed that of the 839 instances belonging to the target class, only 47 were correctly classified, while 792 were erroneously predicted as non-target (Fig. 2). This behavior indicates that the higher overall accuracy obtained in the 2:1 scenario was largely influenced by the classifier's biased preference for the majority class (non-target), resulting in poor target detection performance. In the 1:1 scenario, Random Forest showed a more balanced classification between the two classes, correctly identifying 524 target samples and 475 non-target samples, but its overall predictive performance remained limited.

In contrast, the SMO classifier demonstrated higher discriminative capacity and robustness against the proposed time-frequency-spatial representations. In the 1:1 scenario, the SMO achieved an accuracy of 93.66% and a Kappa ratio of 0.8737, indicating a high level of agreement beyond chance. The corresponding confusion matrix showed that 750 of the 839 samples in the target samples and 822 of the 839 non-target samples were correctly classified, resulting in only 17 false positives and 89 false negatives. In the 2:1 scenario, the performance of the SMO remained stable, registering an accuracy of 93.29

These superior results demonstrate that the features extracted by LeNet from CWT sarcograms have a high separability structure that is successfully mapped by the polynomial kernel support vector machine (SVM)-based classifier, overcoming the orthogonal division limitations of the Random Forest for this specific set of attributes. Regarding the generalization of the proposed model, a limitation of the current study is the reliance on subject-dependent cross-validation. The high inter-subject variability inherent in EEG signals poses a significant challenge for applying a model trained on one individual to another. Future work will focus on implementing cross-subject validation (e.g., leave-one-subject-out) and exploring domain adaptation techniques to evaluate and improve the model's generalization capacity for unseen participants in real-world scenarios. The stability of the SMO between the different subsampling ratios reinforces the feasibility of the proposed pipeline for the automated detection of visual targets in RSVP tasks. Compared with previously reported results obtained using the same database [10], [11], [12], the proposed approach achieved competitive classification performance, with an accuracy of 93.66% in the 1:1 scenario and 93.29% in the 2:1 scenario. These results suggest that the combination of CWT-based time-frequency representations, LeNet feature extraction, and SMO classification provides a competitive alternative for RSVP target detection.

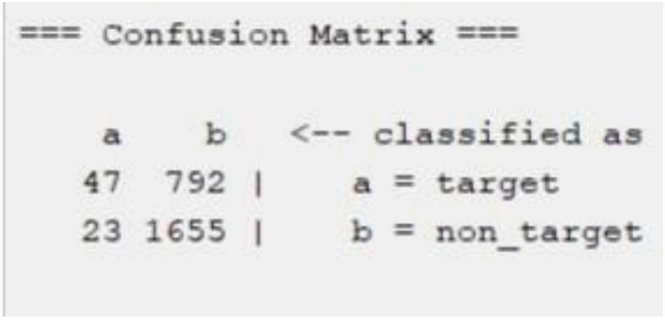


Fig. 2. Confusion matrix of the random forest results with 100 trees.

Furthermore, it is important to discuss the computational limitations of the proposed approach and its potential for real-time applications. The current pipeline involves generating time-frequency representations via the CWT, extracting features using a LeNet architecture, and classifying them with the SMO algorithm. While this methodology yields high discriminative accuracy, the feature extraction step, which required human intervention, introduced a computational overhead that may restrict immediate deployment in real-time scenarios. Future optimizations should investigate architectures that process time-frequency matrices directly, or the implementation of the CWT and classification stages on hardware accelerators to ensure the system provides all requirements for continuous RSVP monitoring.

REFERENCES

[1] S. Lees et al., "A review of rapid serial visual presentation-based brain–computer interfaces," J. Neural Eng., vol. 15, no. 2, p. 021001, Jan. 2018, doi: 10.1088/1741-2552/AA9817.

[2] Y. Mohsenzadeh, S. Qin, R. M. Cichy, and D. Pantazis, "Ultra-rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway," Elife, vol. 7, Jun. 2018, doi: 10.7554/ELIFE.36329.

[3] A. Matran-Fernandez and R. Poli, “Towards the automated localisation of targets in rapid image-sifting by collaborative brain-computer interfaces,” PLoS One, vol. 12, no. 5, p. e0178498, May 2017, doi: 10.1371/JOURNAL.PONE.0178498.
[4] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, 2000, doi: 10.1161/01.CIR.101.23.E215.
[5] G. H. Klem, H. Otto Lu Èders, H. Jasper, and C. Elger, “The ten±twenty electrode system of the International Federation”.
[6] S. Mallat, “A Wavelet Tour of Signal Processing,” 2008.
[7] L. P. A. Arts and E. L. van den Broek, “The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time–frequency analysis,” Nat. Comput. Sci., vol. 2, no. 1, p. 47, Jan. 2022, doi: 10.1038/S43588-021-00183-Z.
[8] H. Bârzan, A. M. Ichim, V. V. Moca, and R. C. Mureşan, “Time-Frequency Representations of Brain Oscillations: Which One Is Better?,” Front. Neuroinform., vol. 16, p. 871904, Apr. 2022, doi: 10.3389/FNINF.2022.871904.
[9] Time–Frequency Image Grids,” Diagnostics, vol. 16, no. 5, p. 746, Mar. 2026, doi: 10.3390/DIAGNOSTICS16050746.
[10] F. Li et al., “Phase preservation neural network for electroencephalography classification in rapid serial visual presentation task,” IEEE Trans. Biomed. Eng., vol. 69, no. 6, pp. 1931–1942, Jun. 2022, doi: 10.1109/TBME.2021.3130917.
[11] S. Kilani, S. N. Aghili, Y. Fathi, and A. I. Sburlea, “Optimization of transfer learning based on source sample selection in Euclidean space for P300-based brain-computer interfaces,” Front. Neurosci., vol. 18, p. 1360709, Jul. 2024, doi: 10.3389/fnins.2024.1360709.
[12] H. He and D. Wu, “Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach,” IEEE Trans. Biomed. Eng., vol. 67, no. 2, pp. 399–410, Feb. 2020, doi: 10.1109/TBME.2019.2913914.

MALLAT, Stéphane. **A Wavelet Tour of Signal Processing: The Sparse Way. 3rd ed**. Amsterdam: Elsevier/Academic Press, 2009.

As requested by the authors of the database used, and as a form of acknowledgment and thanks, we cite the following person responsible for the feasibility of the data:

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

Automatic Detection of Cervical Spine Anatomical Landmarks in Lateral Radiographs Using Deep Learning

1st Arinaldo Alves Filho
Polytechnic School (Poli)
University of São Paulo
São Paulo, Brazil
0009-0006-6764-9957

2nd Felipe Fava de Lima
Polytechnic School (Poli)
University of São Paulo
São Paulo, Brazil
0000-0002-7102-8532

3rd Caroline Dantas Vilar
Biomedical Engineering Department
University of Rio Grande do Norte
Natal, Brazil
0009-0007-8717-8969

4th Heliana Bezerra Soares
Biomedical Engineering Department
University of Rio Grande do Norte
Natal, Brazil
0000-0002-0833-7906

Abstract* — Assessing cervical curvature often relies on the manual identification of anatomical landmarks on radiographs, a time-consuming procedure that depends on the evaluator's experience and is susceptible to intra- and inter-observer variability. This study proposes an automated method for detecting cervical spine anatomical landmarks in lateral radiographs using deep learning. A total of 1,904 images were selected from the public Cervical Spine X-ray Atlas (CSXA) V3.0 dataset, distributed across the lordotic, straight, and kyphotic classes. Following preprocessing, the radiographs were resized to 224 × 224 pixels, and the landmark coordinates were normalized according to image width and height. A ResNet-18 convolutional neural network was adapted to predict the normalized coordinates of 23 anatomical landmarks, resulting in 46 output values. On the test set, the model achieved a mean Euclidean error of 9.44 pixels relative to the manual reference annotations. As a complementary evaluation, the predicted landmarks were used to extract morphometric features and subsequently classify cervical curvature using supervised algorithms. In this stage, Logistic Regression and Support Vector Machine with an RBF kernel achieved accuracies of 91% and 90%, respectively. The results indicate that the detected landmarks preserved relevant information about cervical geometry, enabling their use for automated morphometric feature extraction. The proposed approach shows potential to support the computational assessment of cervical spine radiographs, although validation on external datasets and images from different clinical settings is still required.

Keywords* — cervical curvature, deep learning, digital image processing, anatomical landmarks, lateral radiography.

I. Introduction

Cervical spine alignment is an important parameter for assessing musculoskeletal disorders, degenerative diseases, and postural abnormalities. Alterations in cervical curvature can be associated with pain, functional limitations, and impaired quality of life—conditions characterized by high prevalence and significant socioeconomic impact [1]–[4]. Furthermore, maintaining physiological alignment contributes to the proper distribution of mechanical loads and the preservation of the biomechanical and neurological functions of the cervical spine, underscoring the importance of developing reliable assessment methods [1], [2], [5].

Lateral radiography remains one of the primary examinations used to assess cervical alignment and curvature [3]. However, radiographic analysis still frequently relies on the manual identification of anatomical landmarks and the measurement of morphometric parameters, such as the Cobb angle. This procedure can be time-consuming, dependent on the examiner's experience, and susceptible to intra- and inter-observer variability, factors that may compromise the reproducibility of the measurements obtained [2], [6], [7].

In this context, advances in digital image processing and deep learning have driven the development of automated methods for medical image analysis and the detection of anatomical structures [7], [8]. Models based on convolutional neural networks have demonstrated the ability to automatically localize anatomical landmarks on spinal radiographs, reduce the time required to obtain measurements, and show agreement with assessments performed by specialists [7], [9], [10]. Although these studies have shown promising results, it remains important to investigate approaches that integrate the automatic detection of cervical spine landmarks with the extraction of morphometric features, making it possible to assess whether the predicted landmarks preserve the geometric information required for subsequent analyses.

Thus, this study proposes an automated approach for detecting cervical spine anatomical landmarks in lateral radiographs using deep learning. The methodology integrates digital image processing techniques and a ResNet-18 convolutional neural network adapted to predict the coordinates of 23 anatomical landmarks. As a complementary evaluation, the detected landmarks are used to extract morphometric features and classify cervical curvature, aiming to determine whether the geometric information obtained can support the automated morphometric analysis of this region.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

II. Materials and Methods

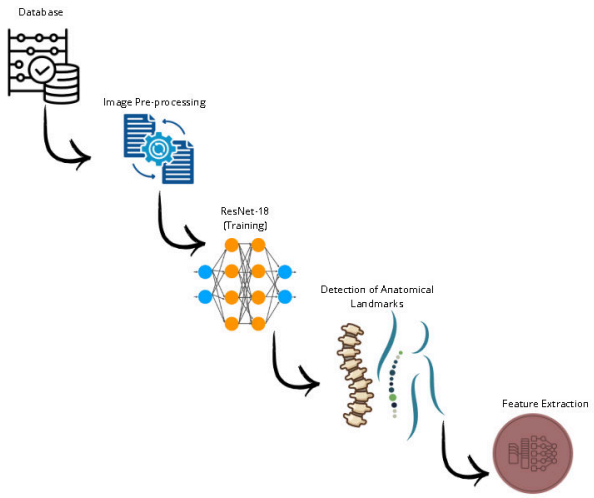
A. Data-base and proposed approach

The study was conducted using the Cervical Spine X-ray Atlas (CSXA) V3.0, a public dataset consisting of 4,963 lateral cervical spine radiographs provided with manual annotations of 23 anatomical landmarks and complementary metadata [3]. The original radiographs have varying resolutions, and the coordinates of the anatomical landmarks are provided in JSON-format files. The reference standard (ground truth) used in this study was derived from annotations independently performed by four orthopedic surgeons specialized in spine surgery, which were subsequently subjected to a three-stage cross-checking procedure. However, the dataset does not provide quantitative measures of intra-observer or inter-observer variability for the annotations used in this study.

For this study, only images classified as lordotic (Class 1), straight (Class 2), or kyphotic (Class 5) were considered. Classes corresponding to sigmoid curvatures were excluded because they require morphometric criteria different from those used for the three selected categories. To reduce class imbalance, up to 700 images from each class were randomly selected using a fixed seed to ensure reproducibility. As the dataset contained only 504 radiographs from the kyphotic class, the final dataset consisted of 1,904 images.

Fig. 1 presents the main stages of the proposed approach, comprising dataset organization, radiograph preprocessing, automatic detection of anatomical landmarks, extraction of morphometric features, and, as a complementary evaluation, classification of cervical curvature.

Fig. 1. Methodological process flowchart.



B. Image preprocessing

Before training the neural network, the radiographs were subjected to a digital image processing pipeline aimed at enhancing bone contours and reducing the influence of noise and contrast variations. Initially, each image was converted to grayscale, focusing the processing on intensity information. Next, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied with a clip limit of 1.5 and a tile grid size of 8 × 8 to adaptively enhance contrast across different regions of the radiograph [11], [12].

In the next stage, Unsharp Masking was employed to enhance the sharpness of the vertebral contours. A smoothed version of the image was obtained using a Gaussian filter with a 5 × 5 kernel and subsequently combined with the CLAHE-processed image using weights of 1.5 and −0.5, respectively. Sobel operators were also applied in the x and y directions, both with a 5 × 5 kernel, to enhance intensity variations associated with the edges of the bone structures [11].

The gradient map obtained with the Sobel operator in the x direction was inverted and used as an auxiliary mask to attenuate the influence of posterior structures. This mask was combined with the image resulting from Unsharp Masking using weights of 30% and 70%, respectively. Subsequently, the result of this combination was integrated with the gradient map obtained using the Sobel operator in the y direction, with weights of 60% and 40%. The final image resulting from this processing was used as input for the automatic anatomical landmark detection stage. Fig. 2 presents a comparison between the original radiograph and the result of the preprocessing pipeline.

C. Detection of anatomical landmarks

After preprocessing, all radiographs were converted to grayscale—including those originally stored without color information—to ensure the exclusive use of intensity values and avoid the presence of residual color channels. Subsequently, the images were resized to 224×224 pixels and normalized to the 0–1 range. To ensure compatibility with the input of the pre-trained ResNet-18, the intensity channel was replicated three times, resulting in images with dimensions of 224×224×3.

The x and y coordinates of the 23 anatomical landmarks provided by the database were normalized by the image width and height, respectively. This procedure allowed for the standardization of annotations derived from radiographs with varying resolutions [3]. Each sample was associated with an output vector consisting of 46 values, corresponding to the two normalized coordinates of each anatomical landmark.

The set of 1,904 radiographs was split into 80% for training and 20% for testing, using a fixed random seed to ensure the reproducibility of the split. Thus, 1,523 images were allocated for training and 381 for testing. The test set was kept separate during training and subsequently used to evaluate the localization of anatomical landmarks.

Point detection was formulated as a supervised regression problem. For this purpose, a ResNet-18 convolutional neural network, pre-trained on the ImageNet dataset, was used. The fully connected layer originally intended for image classification was replaced by a linear layer with 46 output neurons, enabling the network to simultaneously predict the normalized coordinates of the 23 anatomical points. The ResNet architecture employs residual connections, which facilitate gradient propagation during the training of deep networks [13].

Training was conducted in the Google Colab environment, using a graphics processing unit (GPU) whenever available. Samples were loaded in batches of 16 images and shuffled at each epoch. The Adam optimizer was used with a learning rate of 0.001, and Mean Squared Error (MSE) was used as the loss function.

D. Evaluation of detection and morphometric validation

The localization of the anatomical landmarks was evaluated on the test set, consisting of 381 radiographs not used during training. To quantitatively evaluate the accuracy of anatomical landmark detection, the Euclidean distance between each predicted landmark and its corresponding manual reference annotation (ground truth) was calculated. Since the coordinates used during training were normalized according to the width and height of the images, the predicted coordinates were subsequently converted back to the original dimensions of each radiograph, allowing the localization error to be calculated in pixels in the original image space. The Euclidean distance was calculated according to Eq. (1):

$$E_{px} = \sqrt{\left(x_{pred} - x_{real}\right)^2 + \left(y_{pred} - y_{real}\right)^2} \quad (1)$$

where x_{pred} and y_{pred} correspond to the predicted coordinates converted back to the original radiograph resolution, whereas x_{real} and y_{real} represent the coordinates of the manual reference annotations.

Additionally, using the physical calibration information available in the CSXA metadata, the error in pixels was converted to physical scale using the Pixel equivalent parameter provided individually for each radiograph by the atlas. This conversion allowed the localization error to also be expressed in millimeters, facilitating its interpretation relative to the anatomical scale of the images.

Performance was characterized by the mean, median, standard deviation, and percentiles of the error distribution, both globally and individually for the 23 anatomical landmarks. Detection success rates were also calculated at different distance thresholds, corresponding to the proportion of landmarks whose distance from the reference annotation remained below each established threshold.

Following the detection of anatomical landmarks, morphometric features traditionally used in the assessment of the cervical spine were automatically extracted. Key calculated features included angular measurements, Euclidean distances between anatomical landmarks, vertebral and intervertebral heights, endplate inclinations, geometric descriptors of the vertebrae, and the Cobb angle between C2 and C7. These measurements enabled a quantitative representation of the cervical region's geometry based on the landmarks predicted by the model.

To assess whether the detected anatomical landmarks preserved sufficient morphometric information to distinguish cervical curvature patterns, the extracted features were used as input variables for five families of supervised algorithms: Random Forest, Logistic Regression, Support Vector Machine (SVM), XGBoost, and Multilayer Perceptron (MLP). The SVM was evaluated using both RBF and polynomial kernels, resulting in a total of six classification configurations. At this stage, cervical curvature classification served solely as a complementary validation procedure for the detected anatomical landmarks, rather than as the study's primary objective.

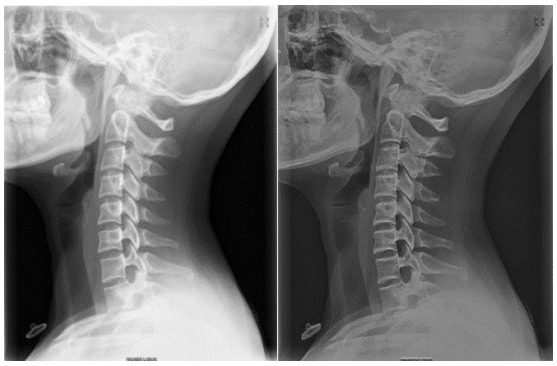
For the classification stage, the morphometric features extracted from the predicted anatomical landmarks were used as input variables for the models. The data were stratified into 75% for training and 25% for testing, using a random seed of 42. The features were standardized to zero mean and unit standard deviation. The performance of the different configurations was evaluated on the test subset using accuracy, precision, sensitivity (recall), and F1-score.

III. RESULTS AND DISCUSSION

A. Preprocessing result

Fig. 2 presents a comparison between an original radiograph and the image obtained after applying the preprocessing pipeline. Visually, the combination of CLAHE, Unsharp Masking, and Sobel operators provided greater differentiation of vertebral contours relative to adjacent tissues and structures. The result preserved the overall anatomical configuration of the cervical spine and enhanced the intensity transitions subsequently used by the neural network to locate anatomical landmarks.

Fig. 2. Comparison between the original radiograph and the image after preprocessing.

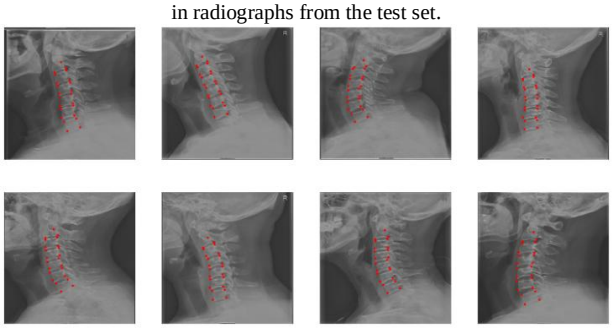


Since no specific quantitative metrics were employed to assess preprocessing quality, it is not possible to objectively state that there was an improvement in the visibility, sharpness, or contrast of the radiographs. Therefore, Fig. 2 is presented as a visual comparison between the original image and the result obtained after applying the preprocessing steps.

B. Anatomical landmark detection performance

ResNet-18 was trained for 65 epochs. The number of epochs was empirically determined based on successive training runs, without an automatic stopping criterion, considering the progressive reduction and stabilization of the loss function at values close to 0.001. This behavior indicates convergence of the optimization process. After training, the model was applied to the 381 radiographs in the test set to predict the normalized coordinates of the 23 anatomical landmarks. Fig. 3 presents examples of the predicted landmarks overlaid on the radiographs.

Fig. 3. Examples of the automatic detection of the 23 anatomical landmarks in radiographs from the test set.



The quantitative evaluation was performed after converting the normalized coordinates back to the original dimensions of each radiograph. Considering the 23 anatomical landmarks across the 381 images, totaling 8,763 evaluated locations, a mean Euclidean error of 9.44 pixels, a median of 9.42 pixels, and a standard deviation of 4.03 pixels were obtained. The 95th percentile was 16.18 pixels, indicating that 95% of the landmark locations had an error less than or equal to this value. After conversion to physical scale using the calibration data provided by the atlas, the corresponding values were a mean error of 3.45 mm, a median of 3.43 mm, a standard deviation of 1.48 mm, and a 95th percentile of 5.96 mm. The cumulative error distribution further showed that 85.94% of the detections were within 5 mm of the reference annotations, whereas 95.32% were within 6 mm and 98.73% were within 7 mm.

TABLE I. GLOBAL STATISTICS OF LANDMARK LOCALIZATION ERRORS

Metric	Pixels	Millimeters
Mean	9.44	3.45
Median	9.42	3.43
Standard deviation	4.03	1.48
P90	14.59	5.34
P95	16.18	5.96
P99	19.83	7.20

The global distribution of errors is presented in Fig. 4. The histogram shows a higher concentration of observations in the intermediate lower-error ranges, with a progressive decrease in frequency as the distance between the prediction and the manual annotation increases. The proximity between the mean and median indicates a relatively balanced distribution in its central region, whereas higher error values represent a smaller proportion of the detections.

Fig. 4. Distribution of Euclidean errors of anatomical landmarks in the test set.

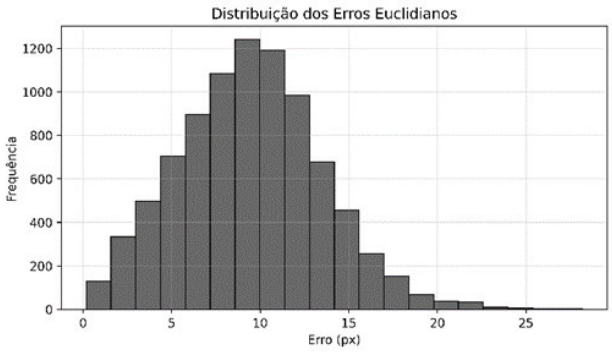
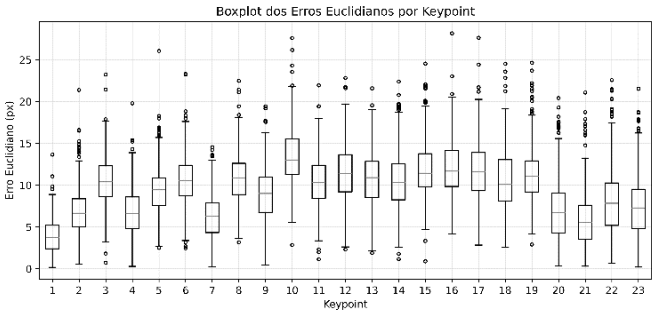


Fig. 5 presents the distribution of Euclidean errors for each of the 23 anatomical landmarks, allowing the behavior of the predictions along the cervical spine to be evaluated individually. Some anatomical landmarks showed lower dispersion and smaller error values, whereas others exhibited greater variability. This behavior indicates that localization difficulty is not uniform across the different landmarks and may be related to the anatomical characteristics of each region and the visualization conditions present in the radiographs.

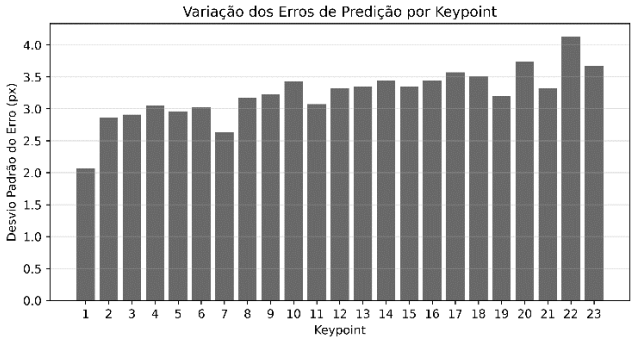
Fig. 5. Distribution of Euclidean errors for each anatomical landmark in the test set.



The observed differences may be associated with vertebral position, overlap of bone structures and soft tissues, variations in contrast and positioning during radiographic acquisition, and, in some cases, the presence of external elements, such as clothing, which may hinder the delineation of anatomical contours. These factors may reduce the visibility of specific reference landmarks and contribute to the greater dispersion of errors observed for some landmarks. Therefore, the variation among landmarks reinforces the importance of evaluating model performance individually, since relying exclusively on a global metric may conceal local differences in detection accuracy.

As a complementary analysis of the error distribution by landmark, Fig. 6 presents the standard deviation of the Euclidean error for each anatomical landmark. This measure allows the dispersion of the predictions around the mean behavior of each landmark to be assessed and, consequently, differences in localization consistency among the evaluated landmarks to be identified.

Fig. 6. Distribution of the standard deviation of errors, in pixels, for each anatomical landmark.

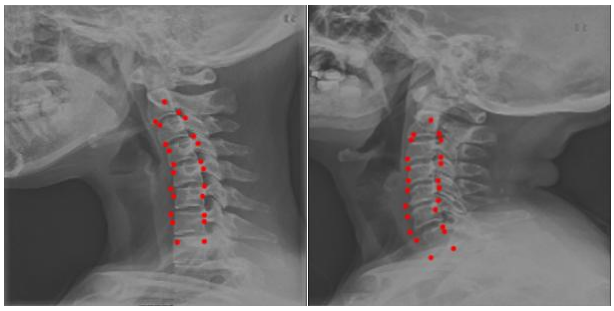


The results show that prediction variability was not uniform across the 23 landmarks. While some landmarks exhibited lower error dispersion, indicating more consistent behavior across the evaluated radiographs, others showed higher standard deviations. This analysis complements the differences observed in the boxplot, demonstrating that landmark performance varies not only in terms of error magnitude but also in terms of error dispersion across images. Thus, landmarks with similar mean errors may exhibit different levels of stability, reinforcing the need to consider both the accuracy and variability of the predictions when interpreting the derived morphometric features evaluated in the subsequent stage.

In addition to the quantitative analysis of the error distribution, representative cases in which the predictions showed greater localization difficulty were evaluated. Fig. 7

presents examples in which deviations in the position of some landmarks can be observed, highlighting situations in which the automatic identification of anatomical landmarks showed lower spatial accuracy.

Fig. 7. Examples of radiographs showing deviations in the localization of the predicted anatomical landmarks.



C. *Validation based on morphometric characteristics*

To evaluate whether the detected anatomical landmarks preserved morphometric information useful for differentiating cervical curvature patterns, the extracted features were used as input to five families of supervised algorithms, totaling six configurations. For this stage, the data were stratified into 75% for training and 25% for testing. Tables II–VII present the performance metrics obtained on the test set for each evaluated configuration.

TABLE II. PERFORMANCE METRICS OF THE RANDOM FOREST MODEL

Classification	Precision	Sensitivity	F_1-*score*	N° of images
1	0.96	0.96	0.96	175
2	0.87	0.87	0.86	175
5	0.88	0.84	0.86	126

TABLE III. PERFORMANCE METRICS OF THE LOGISTIC REGRESSION MODEL

Classification	Precision	Sensitivity	F_1-*score*	N° of images
1	0.97	0.96	0.96	175
2	0.86	0.90	0.88	175
5	0.91	0.85	0.88	126

TABLE IV. PERFORMANCE METRICS OF THE POLYNOMIAL SVM MODEL

Classification	Precision	Sensitivity	F_1-*score*	N° of images
1	0.97	0.85	0.91	175
2	0.73	0.93	0.82	175
5	0.92	0.74	0.82	126

TABLE V. PERFORMANCE METRICS OF THE SVM RBF MODEL

Classification	Precision	Sensitivity	F_1-*score*	N° of images
1	0.95	0.95	0.96	175
2	0.85	0.88	0.86	175
5	0.89	0.84	0.87	126

TABLE VI. PERFORMANCE METRICS OF THE XGBOOST MODEL

Classification	Precision	Sensitivity	F_1-*score*	N° of images
1	0.97	0.96	0.97	175
2	0.84	0.90	0.87	175
5	0.89	0.83	0.86	126

TABLE VII. PERFORMANCE METRICS OF THE MULTILAYER PERCEPTRON MODEL

Classification	Precision	Sensitivity	F_1-*score*	N° of images
1	0.96	0.94	0.95	175
2	0.82	0.87	0.85	175
5	0.87	0.82	0.84	126

Overall, the metrics showed consistent behavior across the models and complemented the accuracy-based analysis, allowing classifier performance to also be evaluated in terms of their ability to identify the different cervical curvature classes.

Accuracy ranged from 85% to 91%. The best result was obtained by Logistic Regression, with 91%, followed by SVM with an RBF kernel, with 90%. Random Forest and XGBoost achieved 89%, while MLP and SVM with a polynomial kernel showed accuracies of 88% and 85%, respectively.

The performance of Logistic Regression suggests that part of the differences among the lordotic, straight, and kyphotic classes can be represented by approximately linear combinations of the extracted morphometric features. In turn, the performance of SVM with an RBF kernel indicates that nonlinear relationships among these features may also contribute to the separation of curvature patterns. The combined analysis of precision, sensitivity, and F1-score complements these results by considering not only the overall proportion of correct classifications but also the performance of the models in identifying the different curvature patterns. However, since the differences in accuracy among the models were small, it is not possible to claim the definitive superiority of a single configuration without additional analyses, such as cross-validation and confidence intervals.

Considering that the features used by the classifiers were extracted from the anatomical landmarks predicted by ResNet-18, the results indicate that, despite the observed spatial errors, the predicted coordinates remained close to the reference annotations, preserving geometric relationships relevant to the extraction of subsequent morphometric features. The variations resulting from these deviations did not prevent the extracted features from retaining sufficient information to differentiate the evaluated curvature patterns, as demonstrated by the performance of the classifiers. However, this agreement does not imply clinical equivalence of the measurements. For parameters that are more sensitive to small landmark displacements, or for applications requiring greater accuracy, further methodological improvements and additional studies are needed to evaluate the propagation of localization errors to morphometric measurements and their potential impact on clinical decision-making.

Although cervical curvature classification is not the primary objective of this study, the results demonstrate the applicability of the features derived from the predicted landmarks for differentiating the three evaluated curvature patterns. This stage complements the direct evaluation of localization errors and should be interpreted as an internal technical validation of the proposed framework, rather than, in isolation, as a clinical validation of the method.

IV. CONCLUSION

This study presented an automated approach for detecting 23 anatomical landmarks of the cervical spine in lateral

radiographs using a ResNet-18 convolutional neural network. The methodology integrated digital image processing techniques, coordinate normalization, and deep learning, enabling the formulation of landmark localization as a supervised regression task.

Evaluation performed on 381 radiographs not used during training demonstrated a mean Euclidean error of 9.44 pixels, corresponding to 3.45 mm on the considered physical scale. Analysis of the error distribution revealed differences in the magnitude and dispersion of the predictions among the anatomical landmarks, as well as the occurrence of cases with higher deviations. As a complementary analysis, the predicted coordinates were used to extract morphometric features and classify cervical curvature patterns. In this stage, Logistic Regression and SVM with an RBF kernel achieved accuracies of 91% and 90%, respectively, indicating that the detected landmarks preserved geometric information useful for differentiating the lordotic, straight, and kyphotic classes.

As a limitation, the complementary classification stage was conducted as an internal technical evaluation. Future studies should adopt validation strategies that maintain strict separation between the fitting of preprocessing and model selection procedures and the dataset reserved for final evaluation.

The results demonstrate the feasibility of the proposed approach for the automated localization of anatomical landmarks and the extraction of morphometric measurements from the cervical spine. However, the study was conducted using images from a single public dataset and considered only three curvature patterns. Future work should involve evaluating the model on external datasets, including radiographs acquired using different equipment and at different institutions, investigating other deep learning architectures, and directly assessing the propagation of localization errors to morphometric measurements, as well as their clinical relevance through comparisons with assessments performed by specialists.

ACKNOWLEDGMENT

The authors thank LAPSIM—the Laboratory for Medical Signal and Image Acquisition and Processing at the Federal University of Rio Grande do Norte (UFRN) Biomedical Engineering program—for the technical support, infrastructure, and assistance provided for the development of this work.

REFERENCES

[1] J. K. Scheer et al., "Cervical spine alignment, sagittal deformity, and clinical implications: A review," J. Neurosurg. Spine, vol. 19, no. 2, pp. 141–159, 2013.

[2] D. E. Harrison et al., "Cervical lordosis: A study of patients with neck pain," *J. Manipulative Physiol. Ther.*, vol. 27, no. 3, pp. 206–214, 2004.

[3] Y. Ran, "Cervical spine X-ray atlas (CSXA) V3.0," Science Data Bank, Version 1, 2024, doi: 10.57760/sciencedb.15391.

[4] M. A. Aimi, E. G. Raupp, E. F. D. Schmit, A. Vieira, and C. T. Candotti, "Correlação entre morfologia cervical, dor, funcionalidade e ADM em indivíduos com cervicalgia," Coluna/Columna, vol. 18, no. 2, pp. 101–105, 2019.

[5] J. Gu et al., "A new index for cervical curvature evaluation: Relative cervical curvature area," World Neurosurg., vol. 189, pp. e777–e786, 2024, doi: 10.1016/j.wneu.2024.05.012.

[6] Y. Guo et al., "Convolutional neural network-based approach for Cobb angle measurement using Mask R-CNN," Diagnostics, vol. 15, no. 9, Art. no. 1066, 2023, doi: 10.3390/diagnostics15091066.

[7] H. Nakarai et al., "Automatic calculation of cervical spine parameters using deep learning: Development and validation on an external dataset," Global Spine J., vol. 15, no. 2, pp. 710–721, 2025, doi: 10.1007/s00586-022-07189-9.

[8] Y. M. Y. Abdallah and T. Alqahtani, "Research in medical imaging using image processing techniques," in Open Access Books Built by Scientists, for Scientists. London, U.K.: IntechOpen, 2019, pp. 501–525.

[9] H. Zhang et al., "Deep learning method for precise landmark identification and structural assessment of whole-spine radiographs," Bioengineering, vol. 11, no. 5, Art. no. 481, 2024, doi: 10.3390/bioengineering11050481.

[10] C.-H. Weng et al., "Automatic recognition of whole-spine sagittal alignment and curvature analysis through a deep learning technique," Sci. Rep., vol. 12, Art. no. 5521, 2022, doi: 10.1038/s41598-022-09591-3.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.

[12] A. M. Reza, “Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for real-time image enhancement,” *J. VLSI Signal Process. Syst. Signal Image Video Technol.*, vol. 38, no. 1, pp. 35–44, 2004.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

Signal-Based Assessment of Cardiovascular Autonomic Responses to Different COVID-19 Vaccine Technologies During Tilt Testing

1st Ana Leticia G. dos Santos
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
analelegsantos@gmail.com

2nd Samuel Minucci Camargo
UMC Polyclinic University of Mogi das Cruzes
Mogi das Cruzes, Brazil
doutorsamuelcamargo@gmail.com

3rd Kelly Correia B. da Silva
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
kellybaioco123@gmail.com

4th Stella Tassinari Maximo
Faculty of MedicineSão Leopoldo Mandic
Araras, Brazil
stellamaximooo@gmail.com

5th Daniel Gustavo Goroso
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
danielg@umc.br

I. Introduction

The Tilt Test is a test used to assess the cardiovascular response to orthostatic stress and to investigate changes in the autonomic nervous system . The protocol lasts 50 minutes total, with 15 minutes horizontal, 15 minutes inclined and 20 minutes horizontal again. During orthostatic stress, caused by stretcher elevation, an increase in heart rate mediated by sympathetic activation is expected. Changes in this response may indicate autonomic dysfunction.

The Covid-19 pandemic has already affected more than 250 million people globally. Despite advances in understanding COVID-19, there are still many unknowns about its possible repercussions (3,14). Our research group observed a change in heart rate regulation during the Tilt Test in patients with long-term COVID-19.

Notably, there was a significant reduction in spectral density in the high-frequency (HF) heart rate band in this group. Heart rate variability represents a non-invasive biomarker widely used for the assessment of autonomic modulation; post-COVID-19 patients showed a response on the Tilt Test exam that indicates an autonomic dysfunction (2,7,19) . This post-infection autonomic dysfunction is called Long COVID (4,5,10,11,12,16).

In parallel with our research, studies have demonstrated the effectiveness of vaccines in preventing COVID-19-related deaths, highlighting their crucial importance in reducing hospitalizations and deaths (6,13,18). Different vaccine platforms can induce different physiological responses, hence questions arose about the possible dysregulation of the autonomic nervous system also in individuals who did not have COVID-19 (20) but received vaccines, considering that each one was developed by different technologies: vaccine 2 (inactivated virus), vaccine 3 (messenger RNA) and vaccines 1 (viral vector). (21)

Thus, the hypothesis is that different vaccine technologies may be associated with distinct patterns of cardiovascular autonomic regulation, detectable through HRV analysis during the Tilt Test (1,15). And the objective of the study is to investigate possible differences in cardiovascular autonomic regulation associated with different vaccine technologies through the analysis of heart rate variability during the Tilt Test, investigating, in addition to its effectiveness. Using the database of the project "Online non-invasive detection of postural orthostatic tachycardia syndrome in post-COVID-19 patients, Project FAPESP 21/14231-0", we sought to analyze these effects on the body.

This work was supported by the São Paulo Research Foundation (FAPESP), Grant No. 21/14231-0. Ana L. G. Santos was supported by a Scientific Initiation Scholarship from the National Council for Scientific and Technological Development (CNPq).

II. OBJECTIVES

Thus, the hypothesis is that different vaccine technologies may be associated with distinct patterns of cardiovascular autonomic regulation, detectable through HRV analysis during the Tilt Test. And the objective of the study is to investigate possible differences in cardiovascular autonomic regulation associated with different vaccine technologies through the analysis of heart rate variability during the Tilt Test, investigating, in addition to its effectiveness.

III. MATERIALS AND METHODS

For data collection and analysis, a Tilt Test is required, therefore ECG model PC1000 from the Brazilian Electronic Technology TEB, electrodes, a reclining stretcher and automatic digital sphygmomanometer Hem-7320-Br by OMRON. These are the basic materials to be able to take a Tilt Test exam.

The analysis and processing of the collected data are analyzed via computer by programs such as Jamovi and Power Bl. Jamovi assists in the statistical tests of the survey, looking for patterns with the data collected, such as the Mann Whitney (17). Power Bl

unifies all study information, from anamnesis to physiological data during the exam, and makes it easy to visualize it in advanced graphs and filters(8).

The test is collected at the Polyclinic Hospital/University of Mogi das Cruzes (UMC). We look for patients by voluntarism. This study is part of the FAPESP project (Grant 21/14231-0) and was approved by the Institutional Research Ethics Committee (64561022.7.0000.5497).

Before starting data collection, the patient undergoes an anamnesis interview, and then begins the collection of the test. The N of the project is 71 participants, but in this study we will consider only the sample is composed of 19 participants, healthy patients with no history of COVID-19, 7 female and 12 male. As depicted in Image 1.

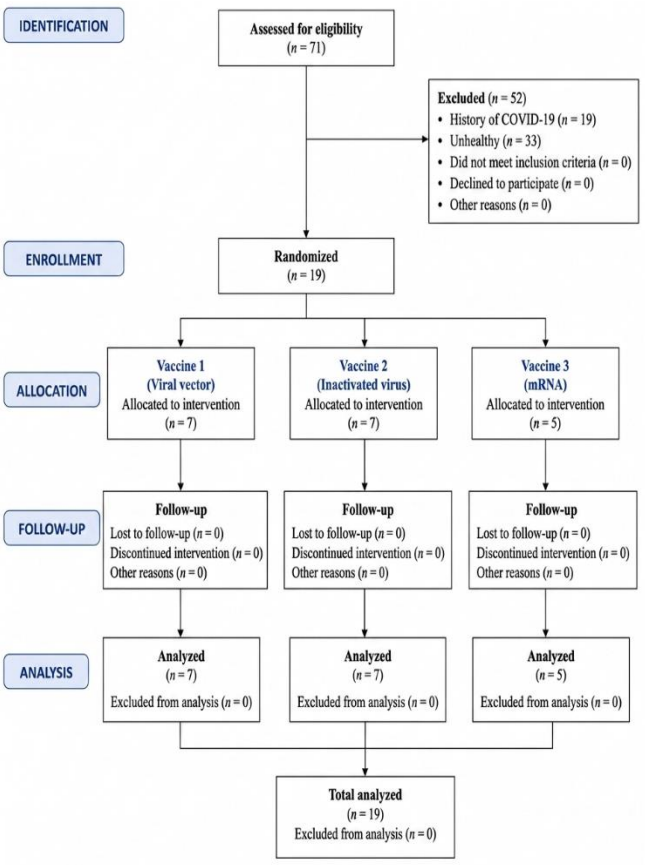


Fig. 1. Consort Flow

Patients who are selected to participate in the study initially go through an inclusion and exclusion criterion, it is necessary to be between 18 and 75 years old, present a negative pregnancy test and PCR, not have any heart problems and not use any medication continuously that alters the autonomic nervous system.

The patient's participation follows the following steps, represented by the following flowchart:

Volunteer recruitment

↓

Clinical interview

↓

Tilt Test

↓

ECG acquisition

↓

Signal preprocessing

↓

HRV extraction

↓

Statistical analysis

IV. RESULTS

From this part of the study, statistical analysis according to vaccine technology began. As shown in Table 1, a difference in heart rate variation was observed between the vaccine groups studied

As shown in Table 1, the groups also showed differences in heart rate variability indexes, especially in heart rate variation, SDNN, NN50 and pNN50, demonstrating different patterns of cardiovascular response among the vaccine technologies evaluated.

TABLE I

CLINICAL PROFILE AND HRV BY GROUPS

	Vaccine 1	Vaccine 2	Vaccine 3
Clinical profile			
Age, years (M ± SD)	43.43 ± 10.20	37 ± 12.25	22.56 ± 3.98
Technology	Viral vector	Inactivated virus	mRNA
BMI, kg/m² (M ± SD)	28.29 ± 5.50	23 ± 2.16	26.56 ± 5.81
Time domain			
Mean HR (bpm) (M ± SD)	70 ± 9.84	67.66 ± 6.12	74.33 ± 10.15
Mean NN (ms)	886.88 ± 126.50	917.41 ± 90.16	847.56 ± 123.54
SDNN (ms) (M ± SD)	126.38 ± 78.20	170.72 ± 37.96	176 ± 89.45
NN50 (M ± SD)	367.38 ± 365.63	457.33 ± 224.56	826.55 ± 537.65
pNN50 (%) (M ± SD)	11.21 ± 11.73	13.93 ± 7.07	23.74 ± 16.29
RMSSD (M ± SD)	89.47 ± 120.71	76.98 ± 32.41	127.79 ± 131.24
Triang. index (M ± SD)	21.12 ± 8.22	27.33 ± 1.66	26.84 ± 10.52
Frequency domain			
VLF (s^2) (M ± SD)	17758.42 ± 75324.66	3250.83 ± 2897.57	304201.96 ± 1263722
LF(s^2) (M ± SD)	15132.36 ± 38129.4	6261.46 ± 6470.63	47571.21 ± 155996.6
HF(s^2) (M ± SD)	7557.62 ± 17218.9	2781.29 ± 1908.05	10418.56 ± 15915.39
LF/HF ratio (M ± SD)	1.83 ± 1.35	1.66 ± 0.85	2.33 ± 2.91
Variation			
HR variation (M ± SD)	54.72 ± 40.94	141.93 ± 17.21	199.35 ± 107.92

Figure 2 shows the comparison between patients representative

of each vaccine technology, allowing us to visualize the behavior of the heart rate during the Tilt Test.

Regarding Vaccine 2, it was observed that the participants did not show the expected increase in heart rate during the elevation of the stretcher, maintaining similar values throughout the tilt period. Vaccine 3 showed an increase in heart rate during stretcher elevation, followed by stabilization during the examination. Vaccine 1, on the other hand, showed a smaller increase in heart rate during the Tilt Test, remaining with a reduced response during the orthostatic period.

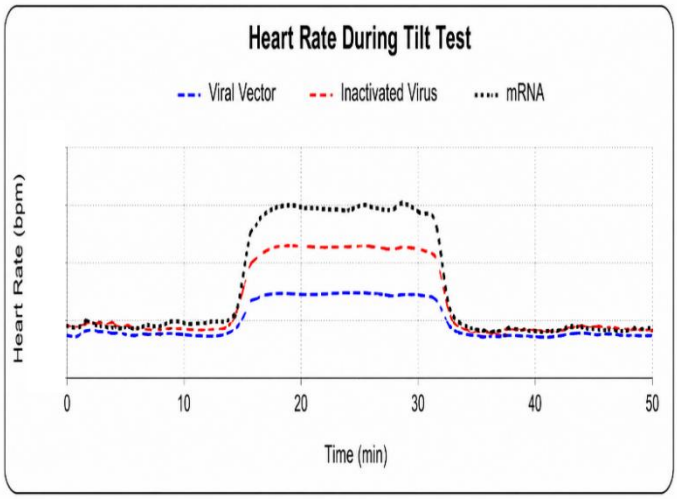


Fig. 2. Group Tilt Test

V. DISCUSSION

The results obtained indicate variations in the physiological responses of the participants, suggesting a possible influence of different vaccine technologies on the autonomic nervous system. The absence of the expected increase in heart rate during stretcher elevation observed mainly in Vaccine 2 may indicate an alteration of cardiovascular autonomic function. On the other hand, Vaccine 3 presented a response closer to the expected physiological behavior during orthostatic stress, characterized by an increase in heart rate followed by stabilization.

The lower heart rate variation observed in Vaccine 1, together with the lower values of the heart rate variability indexes, may suggest a reduction in autonomic modulation when compared to the other groups. These results reinforce the importance of heart rate variability as a non-invasive biomarker for assessing autonomic nervous system function.

It is important to emphasize that the participants evaluated did not have preexisting diseases and did not use continuous medication that could interfere with the autonomic response, reducing possible confounding factors.

However, the study is limited due to the small sample size. Future studies with a larger number of participants and additional analyses of heart rate variability and baroreflex sensitivity will be important to confirm the observed findings and better understand the possible effects of different vaccine technologies on cardiovascular autonomic regulation.

ACKNOWLEDGMENT

This manuscript thanks FAPESP for funding the project Noninvasive online detection of postural orthostatic tachycardia syndrome in post-COVID-19 patients, Project(21/14231-0). And the author ALGS thanks the CNPq scientific initiation scholarship (164066/2024-1).

REFERENCE

1. Kwan, A. C. et al. Apparent risks of postural orthostatic tachycardia syndrome diagnoses after COVID-19 vaccination and SARS-Cov-2 Infection. Nat Cardiovasc Res 1, 1187–1194 (2022).

2. M. Camargo, S. et al. Assessing Autonomic Nervous System Imbalance in Post-COVID-19 Patients through Heart Rate Variability During Tilt Testing. in (2023). doi:10.22489/CinC.2023.411.

3. Milovanovic, B. et al. Assessment of Autonomic Nervous System Dysfunction in the Early Phase of Infection With SARS-CoV-2 Virus. Front. Neurosci. 15, 640835 (2021).

4. Dani, M. et al. Autonomic dysfunction in ‘long COVID’: rationale, physiology and management strategies. Clinical Medicine 21, e63–e67 (2021).

5. Barizien, N. et al. Clinical characterization of dysautonomia in long COVID-19 patients. Sci Rep 11, 14042 (2021).

6. Jin, L., Li, Z., Zhang, X., Li, J. & Zhu, F. CoronaVac: A review of efficacy, safety, and immunogenicity of the inactivated vaccine against SARS-CoV-2. Human Vaccines & Immunotherapeutics 18, 2096970 (2022).

7. Santos, A. L. G. D., Da Silva, K. C. B., Camargo, S. M. & Goroso, D. G. Correção da frequência cardíaca coletada com smartwatch em relação a eletrocardiograma. https://doi.org/10.17605/OSF.IO/V5UZM (2025) doi:10.17605/OSF.IO/V5UZM.

8. Santos, A. L. G. D. et al. Data-Driven Health: Utilizando o Power BI para análises de dados fisiológicos em pesquisas clínicas. https://doi.org/10.17605/OSF.IO/8D642 (2025) doi:10.17605/OSF.IO/8D642.

9. Santos, A. L. G. D. et al. Data-Driven Health: Utilizando o Power BI para análises de dados fisiológicos em pesquisas clínicas. https://doi.org/10.17605/OSF.IO/8D642 (2025) doi:10.17605/OSF.IO/8D642.

10. Tassinari Maximo, S. et al. Heart Rate Regulation in Long-Covid-19 Patients According to the Year of Infection. in (2024). doi:10.22489/CinC.2024.420.

11. Al-Aly, Z., Xie, Y. & Bowe, B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature 594, 259–264 (2021).

12. Acanfora, D. et al. Impaired Vagal Activity in Long-COVID-19 Patients. Viruses 14, 1035 (2022).

13. Pescarini, J. M. et al. Métodos para avaliação da efetividade de vacinas para COVID-19 com ênfase em abordagens quase-experimentais. Ciênc. saúde coletiva 26, 5599–5614 (2021).

14. Lopez-Leon, S. et al. More than 50 Long-term effects of COVID-19: a systematic review and meta-analysis. Preprint at https://doi.org/10.1101/2021.01.27.21250617 (2021).

15. Vanderlei, L. C. M., Pastre, C. M., Hoshi, R. A., Carvalho, T. D. D. & Godoy, M. F. D. Noções básicas de variabilidade da frequência cardíaca e sua aplicabilidade clínica. Rev Bras Cir Cardiovasc 24, 205–217 (2009).

16. Amekran, Y., Damoun, N. & El Hangouche, A. J. Postural orthostatic tachycardia syndrome and post-acute COVID-19. gcsp 2022, (2022).

17. Silva, L. E. V., Fazan, R. & Marin-Neto, J. A. PyBioS: A freeware computer software for analysis of cardiovascular signals. Computer Methods and Programs in Biomedicine 197, 105718 (2020).

18. Patone, M. et al. Risks of myocarditis, pericarditis, and cardiac arrhythmias associated with COVID-19 vaccination or SARS-CoV-2 infection. Nat Med 28, 410–422 (2022).

19. Leticia Gomes Dos Santos, A. et al. Smartwatches in Clinical Pre-diagnosis: Enhancing Tilt Test Analysis for Prolonged COVID-19 Symptoms. in (2024). doi:10.22489/CinC.2024.407.

20. Blitshteyn, S. & Fedorowski, A. The risks of POTS after COVID-19 vaccination and SARS-CoV-2 infection: more studies are needed. Nat Cardiovasc Res 1, 1119–1120 (2022).

21. Sharifian-Dorche, M. et al. Vaccine-induced immune thrombotic thrombocytopenia and cerebral venous sinus thrombosis post COVID-19 vaccination; a systematic review. Journal of the Neurological Sciences 428, 117607 (2021).

Autoregressive and Fractal Mammographic Features for Statistically Differentiating HER2-Positive Breast Cancer

Lucas de Brito Silva
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0000-0001-8600-6953

Pedro Cunha Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0000-0002-0120-5273

Miguel Angel Guevara López
Department of Computer Engineering
Polytechnic Institute of Setúbal
Setúbal, Portugal
0000-0001-7814-1653

Ana Claudia Patrocinio
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0000-0001-9376-7689

***Abstract*—HER2-positive breast cancer has distinct biological and therapeutic implications, but its assessment remains dependent on pathological testing. This study investigated whether interpretable mammographic descriptors can differentiate HER2-positive breast cancer from other molecular subtypes and healthy controls. Mammograms from the Chinese Mammography Database were analyzed using two feature families: two-dimensional autoregressive descriptors, including coefficients a1-a4 and residual variance s, and fractal descriptors, including box-counting fractal dimension and lacunarity. Pairwise group comparisons were performed using normality-guided Welch t-tests or Mann-Whitney U tests, with effect-size estimation and Benjamini-Hochberg false-discovery-rate correction. The analyzed groups included 124 HER2-positive, 76 luminal A, 36 luminal B, 28 triple-negative, and 126 healthy-control images. Autoregressive features showed consistent FDR-adjusted differences involving HER2-positive cases, with residual variance s significant in all HER2-positive pairwise comparisons. Fractal dimension and lacunarity differentiated HER2-positive cases only from healthy controls. These findings suggest that local gray-level dependency descriptors may capture HER2-associated mammographic patterns more consistently than global fractal morphology descriptors in this dataset.**

***Index Terms*—HER2-positive breast cancer, mammography, radiomics, autoregressive model, fractal analysis, lacunarity.**

I. Introduction

Breast cancer is the most commonly diagnosed malignancy worldwide and remains one of the leading causes of cancer-related mortality among women, underscoring the importance of accurate biological characterization for prognosis estimation, risk stratification, and treatment selection [1]. This disease is recognized as a heterogeneous systemic condition whose clinical management relies on the integration of imaging findings, histopathological assessment, molecular profiling, surgery, radiotherapy, and systemic therapies [2]. Advances in gene-expression profiling demonstrated that breast cancer comprises multiple biologically distinct entities rather than a single disease, revealing molecular patterns associated with differences in tumor biology, clinical behavior, therapeutic response, and patient outcomes [3], [4]. Among these molecular subtypes, tumors characterized by amplification or overexpression of the HER2/neu oncogene represent a clinically important subgroup because of their distinct biological characteristics and prognostic implications [5]. The identification of HER2 status has become a fundamental component of breast cancer management, as it directly influences therapeutic decision-making and determines eligibility for targeted anti-HER2 therapies. Consequently, current ASCO-CAP recommendations emphasize rigorous and standardized HER2 testing to ensure accurate patient classification and appropriate treatment selection [6]. Mammography remains the cornerstone of breast imaging for population-based screening and early cancer detection in average-risk women [7]. Although the balance between screening benefits and potential harms continues to be evaluated, mammography remains one of the primary imaging modalities for the detection and characterization of breast abnormalities and provides a rich source of quantitative information that can be explored through computational image analysis approaches [8]. Radiomics was proposed to extract quantitative information from medical images through high-dimensional feature analysis beyond visual interpretation [9]. Within this framework, medical images are treated as quantitative data sources that may encode tumor phenotype, heterogeneity, and clinically relevant biological variation [10].

The authors are thankful for the received financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Brazil), CNPq—Brazil, and FAPEMIG—Brazil.

To ensure consistency and comparability across studies and institutions, standardized feature definitions and reproducible computational pipelines are required [11]. In mammography, quantitative image features have been investigated for molecular subtype prediction, supporting the hypothesis that subtype-related information can be captured from image-derived patterns [12]. Therefore, the objective of this study was to evaluate whether two interpretable families of mammographic descriptors, two-dimensional autoregressive features and fractal features, can statistically differentiate HER2-positive breast cancer from other molecular subtypes and healthy controls.

II. Materials and Methods

This study was designed as a retrospective computational image-analysis investigation of mammographic features associated with HER2-positive breast cancer. Fig 1 summarizes the methodology employed in this study.

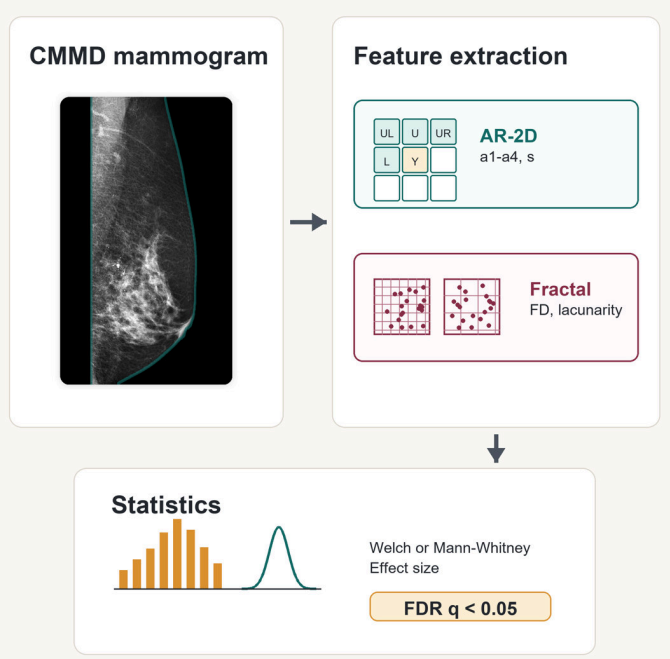


Fig. 1. Flowchat of the proposed methodology.

A. Data Source

Mammographic images were obtained from the Chinese Mammography Database (CMMD), available through The Cancer Imaging Archive (TCIA) [13]. According to the TCIA collection record, CMMD contains mammography data from 1,775 patients from China, including biopsy-confirmed benign or malignant breast disease, with molecular subtype information available for a subset of patients. The collection includes mammographic image data and clinical metadata, including classification, molecular testing, and demographic information [13]. The present analysis focused on image-level groups defined as HER2-positive breast cancer, luminal A breast cancer, luminal B breast cancer and triple-negative breast cancer. The control group was defined as mammograms without any benign or malignant abnormality. The analyzed groups included 124 HER2-positive images, 76 luminal A images, 36 luminal B images, 28 triple-negative images, and 126 control images. The dataset presentes images from both cranio-caudal (CC) and medio-lateral oblique (MLO) views per patient. For the cancer subtypes cases, it was provided images of both breasts and indicates the side affected by malignancy. For the control group, one normal breast per control subjec was used, chosen randomly. The features was extracted from both views and averaged so that each subject contributed a single representative feature set for statistical and classifier-based analyses.

B. Image Preprocessing

The mammographic images were processed from DICOM files using custom MATLAB routines. For each image, a foreground mask was generated by thresholding all pixels greater than the minimum image intensity. Connected components were then computed, and the largest connected component was retained as the breast foreground mask. Pixels outside this mask were removed from the analysis. In the autoregressive-feature pipeline, background pixels were assigned as not-a-number values so that they were excluded from neighborhood-based estimation. In the fractal-feature pipeline, image intensities were normalized to the [0, 1] range before foreground extraction, and the binary foreground mask was used for fractal analysis.

C. Two-Dimensional Autoregressive Feature Extraction

Autoregressive features were extracted using a two-dimensional neighborhood model. For each valid foreground pixel, the current intensity was modeled as a linear combination of four neighboring intensities: left, upper-left, upper, and upper-right. Pixels were included only when the current pixel and all four neighbors were inside the foreground mask. Let Y denote the vector of valid target intensities and X the corresponding matrix of neighboring intensities. The autoregressive coefficients were estimated by least squares as shown in equation (1):

$$\beta = (X^T X)^{-1} X^T Y \tag{1}$$

The four coefficients were retained as features a1, a2, a3, and a4. The variance of the residual error, given by equation (2), was also retained as a regularity-related feature.

$$s = var(Y - X\beta) \tag{2}$$

Therefore, each image was represented by five autoregressive descriptors: a1, a2, a3, a4, and s.

D. Fractal Feature Extraction

Fractal analysis was performed on the binary foreground mask. Fractal dimension was estimated using a box-counting procedure. The mask was evaluated at box sizes defined by powers of two, with zero padding applied when the image dimensions were not exactly divisible by the box size. For each scale, the number of boxes containing at least one foreground

pixel was counted. The fractal dimension was estimated as shown in equation (3)

$$\log N(\varepsilon) = D\ \log(1/\varepsilon) + b \tag{3}$$

where D represents the slope of the fitted line, corresponding to the fractal dimension, and b denotes the intercept of the linear regression. Lacunarity was computed from the distribution of foreground-pixel counts inside boxes across scales. For each box size, the foreground mass was calculated for all boxes, and lacunarity was defined as the variance of box masses divided by the squared mean box mass, with a numerical epsilon added to the denominator. The final lacunarity value was obtained as the mean lacunarity across the evaluated scales. Each image was therefore represented by two fractal descriptors: fractal dimension and lacunarity.

E. Statistical Analysis

Pairwise statistical comparisons were performed between all study groups. For each feature and each pairwise comparison, normality was assessed separately in both groups using the Lilliefors test. When both groups satisfied the normality criterion, the Welch two-sample t-test was used, and the effect size was calculated as Hedges' g. When one group did not satisfy the normality criterion, the Mann-Whitney U test was used, and the effect size was calculated as the rank-biserial correlation. P values were corrected for multiple comparisons within each pairwise feature set using the Benjamini-Hochberg false discovery rate procedure. Statistical significance was defined as an FDR-adjusted p value below 0.05. The analysis was restricted to feature-level group discrimination. No supervised classifier was trained in the current stage of the study. This choice was made to preserve direct interpretability of the extracted mammographic descriptors and to identify which autoregressive and fractal features show statistically supported differences involving the HER2-positive group.

III. Results

The statistical analysis included five image-level groups: HER2-positive breast cancer, luminal A breast cancer, luminal B breast cancer, triple-negative breast cancer, and healthy controls. The sample size used for each group is summarized in Table I.

TABLE I
IMAGE-LEVEL GROUP DISTRIBUTION USED IN THE STATISTICAL ANALYSIS.

Group	**Abbreviation**	n
HER2-positive breast cancer	HER2+	124
Luminal A breast cancer	LA	76
Luminal B breast cancer	LB	36
Triple-negative breast cancer	TNBC	28
Healthy control	Control	126

Autoregressive features showed consistent statistically significant differences involving the HER2-positive group after Benjamini-Hochberg FDR correction. In the comparison between HER2-positive and luminal A images, all five autoregressive descriptors were significant. The same complete separation across the five descriptors was observed for HER2-positive images versus healthy controls. In comparisons with luminal B and triple-negative images, significant findings were more selective, but residual variance s remained significant in both comparisons. Table II presents the summarization of the results for the autoregressive comparisons.

TABLE II
AUTOREGRESSIVE HER2-POSITIVE COMPARISON SUMMARY.

Comparison	**Significant features**	**q(FDR) range**
HER2+ vs LA	all	3.25e-05-6.05e-04
HER2+ vs LB	a4, s	3.66e-04-1.30e-02
HER2+ vs TNBC	a1, a2, a3, s	2.67e-03-6.84e-03
HER2+ vs Control	all	3.62e-16-2.59e-07

At the individual-feature level, the HER2-positive versus luminal A comparison yielded FDR-adjusted significance for a1 (q=3.25e-05), a2 (q=6.00e-05), a3 (q=1.12e-04), a4 (q=6.05e-04), and s (q=3.51e-05). The HER2-positive versus luminal B comparison retained significance for a4 (q=1.30e-02) and s (q=3.66e-04). The HER2-positive versus triple-negative comparison retained significance for a1 (q=6.84e-03), a2 (q=2.67e-03), a3 (q=6.84e-03), and s (q=2.67e-03). The HER2-positive versus control comparison retained significance for all autoregressive descriptors, with q=1.89e-09 for a1, q=5.31e-10 for a2, q=1.20e-10 for a3, q=2.59e-07 for a4, and q=3.62e-16 for s. Fractal descriptors showed a narrower pattern of HER2-positive separation, as shown in Table III. Neither fractal dimension nor lacunarity remained significant after FDR correction in the HER2-positive comparisons against luminal A, luminal B, or triple-negative images. Both fractal descriptors were significant only in the comparison between HER2-positive images and healthy controls.

TABLE III
FRACTAL FEATURES IN HER2-POSITIVE PAIRWISE COMPARISONS.

Comparison	**Fractal dimension**	**Lacunarity**
HER2+ vs LA	0.2429	0.2150
HER2+ vs LB	0.3096	0.1266
HER2+ vs TNBC	0.0967	0.0545
HER2+ vs Control	0.0130	9.33e-04*

In Table III, the asterisk denotes FDR-adjusted statistical significance. The non-significant HER2-positive versus triple-negative lacunarity result approached the adopted threshold but did not pass correction (q=0.0545). Taken together, the feature-level results indicate that autoregressive descriptors provided broader HER2-positive separation across tumor subtypes and controls than fractal descriptors in this dataset. The most consistent autoregressive finding was the residual variance s, which was significant in all HER2-positive pairwise comparisons reported in Table II.

IV. Discussion

This study found that two-dimensional autoregressive descriptors extracted from mammograms provided a broader feature-level separation of HER2-positive breast cancer than fractal descriptors. All five autoregressive variables differed between HER2-positive and luminal A cases and between HER2-positive and healthy controls after FDR correction. In addition, the residual variance s remained significant in every HER2-positive pairwise comparison. By contrast, fractal dimension and lacunarity were significant only when HER2-positive cases were compared with healthy controls. Therefore, within the limits of the current feature-level analysis, local gray-level dependence and residual regularity captured by the autoregressive model appeared more sensitive to HER2-positive differences than global foreground-mask fractal morphology.

These results are consistent with the general radiomics premise that quantitative image descriptors can encode tissue heterogeneity, spatial organization, and phenotype-related information not fully captured by visual interpretation alone [14]. They also agree with recent mammography-based studies showing that HER2 status can be predicted non-invasively from imaging features. Deng et al. developed a multicenter mammographic radiomics model for HER2 status in 1512 patients with improvement after adding clinical characteristics [15]. Although our work did not train a classifier, the consistent FDR-corrected separation of autoregressive features, especially s, supports the same underlying conclusion: mammographic texture contains measurable information associated with HER2-positive disease.

The stronger performance of local texture-related descriptors in our analysis also parallels contrast-enhanced mammography studies. Wang et al. reported that whole-tumor and subregion intratumoral heterogeneity radiomics from CEM differentiated HER2 expression status, with a combined model [16]. Piccolo et al. found that CEM radiomic features, particularly texture features such as gray-level co-occurrence matrix metrics, were associated with HER2 and other prognostic factors [17]. Xu et al. further showed that intra- and peritumoral CEM radiomics could predict equivocal HER2 status [18]. These studies used contrast-enhanced information and tumor/peritumoral segmentation, while the present study used mammographic images and whole-breast foreground processing. Even so, the convergence across studies suggests that spatial heterogeneity and local intensity relationships are relevant imaging correlates of HER2 biology.

Our results are also compatible with studies that address molecular subtype prediction rather than HER2 status alone. Bakker et al. used screening digital mammograms to classify molecular subtypes [19]. Yang et al. integrated mammography and DCE-MRI radiomics and reported better discrimination for luminal A versus HER2 overexpression than for triple-negative versus HER2 overexpression [20]. This pattern is relevant to our data because the autoregressive features separated HER2-positive cases from luminal A and controls more broadly than from luminal B, while the HER2-positive versus triple-negative comparison retained significance for a1, a2, a3, and s but not a4. These subtype-dependent patterns support the interpretation that HER2-related imaging differences are not uniform across all reference groups.

Regarding the fractal descriptors, their limited performance is inherently tied to the type of macro-structural information they encode. Fractal dimension and lacunarity were computed from the foreground breast mask, not from lesion-level gray-level texture. Thus, they primarily summarized global spatial complexity and gap distribution of the segmented breast foreground. Their significance in HER2-positive versus control comparisons may reflect broad disease-versus-healthy structural differences, whereas their lack of significance among tumor subtypes suggests lower sensitivity to molecular subtype variation than autoregressive gray-level descriptors. This interpretation is consistent with radiomics reviews emphasizing that feature class, segmentation strategy, image modality, and preprocessing strongly influence the biological meaning and reproducibility of radiomic features [14], [21].

Recent evidence also indicates that mammography-based AI for receptor or subtype prediction remains promising but not yet clinically definitive. Hosseinzadeh et al., in a systematic review and meta-analysis, highlighted substantial heterogeneity and the need for larger multicenter studies, standardized pipelines, external validation, and uncertainty quantification [22]. Mota reviewed radiomics and deep learning approaches in mammography and tomosynthesis and similarly emphasized that public datasets are used in only a minority of studies, limiting reproducibility [21]. In this context, the present study contributes a transparent feature-level analysis based on a public TCIA dataset rather than an opaque end-to-end prediction model. Several limitations must be acknowledged. First, the analysis was performed at feature level and did not evaluate predictive classification performance, calibration, or external generalization. Therefore, the results cannot be interpreted as diagnostic accuracy estimates. Second, the autoregressive and fractal pipelines used foreground processing rather than manual or automated lesion segmentation; this design improves workflow simplicity but may dilute lesion-specific information. Finally, although FDR correction was applied within pairwise feature sets, the study remains exploratory and should be followed by independent validation and, if appropriate, classifier development using the most stable autoregressive descriptors.

Overall, the findings support the hypothesis that HER2-positive breast cancer may exhibit detectable mammographic texture signatures. The most robust signal in this analysis was not fractal morphology but autoregressive residual regularity, suggesting that local gray-level dependency may be a useful direction for interpretable radiomic modeling of HER2-positive disease.

V. Conclusion

This study evaluated interpretable mammographic descriptors for differentiating HER2-positive breast cancer from other molecular subtypes and healthy controls. The results indicate

that two-dimensional autoregressive features provide broader HER2-positive separation than fractal descriptors, with residual variance s showing FDR-adjusted significance in all HER2-positive pairwise comparisons. In contrast, fractal dimension and lacunarity differentiated HER2-positive cases only from healthy controls, suggesting that these global morphology descriptors were less sensitive to subtype-level differences in the present analysis.

These findings support the use of local gray-level dependency descriptors as a transparent radiomic strategy for exploring HER2-associated mammographic patterns. Future work should validate these descriptors at patient level, evaluate their stability across acquisition settings, and test whether the most robust autoregressive features improve supervised models for non-invasive HER2-positive breast cancer characterization.

ACKNOWLEDGMENT

The authors acknowledge the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for their support. The authors also acknowledge The Cancer Imaging Archive (TCIA) and the contributors of the Chinese Mammography Database (CMMD) for making the imaging data publicly available.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami *et al.*, "Breast cancer," *Nature Reviews Disease Primers*, vol. 5, no. 1, p. 66, 2019.

[3] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[4] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10 869–10 874, 2001.

[5] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene," *Science*, vol. 235, no. 4785, pp. 177–182, 1987.

[6] A. C. Wolff, M. R. Somerfield, M. Dowsett, M. E. H. Hammond, D. F. Hayes *et al.*, "Human epidermal growth factor receptor 2 testing in breast cancer: Asco-college of american pathologists guideline update," *Journal of Clinical Oncology*, vol. 41, no. 22, pp. 3867–3872, 2023.

[7] K. C. Oeffinger, E. T. H. Fontham, R. Etzioni, A. Herzig, J. S. Michaelson *et al.*, "Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society," *JAMA*, vol. 314, no. 15, pp. 1599–1614, 2015.

[8] C. Canelo-Aybar, D. S. Ferreira, M. Ballesteros, M. Posso, N. Montero *et al.*, "Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the european commission initiative on breast cancer," *Journal of Medical Screening*, vol. 28, no. 4, pp. 389–404, 2021.

[9] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[10] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[11] A. Zwanenburg, M. Vallieres, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk *et al.*, "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

[12] W. Ma, Y. Zhao, Y. Ji, X. Guo, X. Jian *et al.*, "Breast cancer molecular subtype prediction by mammographic radiomic features," *Academic Radiology*, vol. 26, no. 2, pp. 196–201, 2019.

[13] C. Cui, L. Li, H. Cai, Z. Fan, L. Zhang, T. Dan, J. Li, and J. Wang, "The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast," 2021.

[14] M. E. Mayerhoefer, A. Materka, G. Langs, I. Haggstrom, P. Szczypinski, P. Gibbs, and G. Cook, "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, 2020.

[15] Y. Deng, Y. Lu, X. Li, Y. Zhu, Y. Zhao, Z. Ruan, N. Mei, B. Yin, and L. Liu, "Prediction of human epidermal growth factor receptor 2 (her2) status in breast cancer by mammographic radiomics features and clinical characteristics: a multicenter study," *European Radiology*, vol. 34, no. 8, pp. 5464–5476, 2024.

[16] S. Wang, T. Wang, S. Guo, S. Zhu *et al.*, "Whole tumour- and subregion-based radiomics of contrast-enhanced mammography in differentiating her2 expression status of invasive breast cancers: A double-centre pilot study," *British Journal of Cancer*, vol. 131, no. 10, pp. 1613–1622, 2024.

[17] C. L. Piccolo, M. Sarli, M. Pileri, M. Tommasiello *et al.*, "Radiomics for predicting prognostic factors in breast cancer: insights from contrast-enhanced mammography (cem)," *Journal of Clinical Medicine*, vol. 13, no. 21, p. 6486, 2024.

[18] C. Xu, J. Qiu, S. Zhang, Y. Chen, X. Wang *et al.*, "Intra- and peritumoral radiomics for predicting equivocal her2 status of breast cancer on contrast-enhanced mammography," *Frontiers in Oncology*, vol. 16, p. 1807530, 2026.

[19] M. A. G. Bakker, M. L. Ovalho, N. Matela, and A. M. Mota, "Decoding breast cancer: using radiomics to non-invasively unveil molecular subtypes directly from mammographic images," *Journal of Imaging*, vol. 10, no. 9, p. 218, 2024.

[20] X. Yang, J. Li, H. Sun, J. Chen *et al.*, "Radiomics integration of mammography and dce-mri for predicting molecular subtypes in breast cancer patients," *Breast Cancer: Targets and Therapy*, vol. 17, pp. 187–200, 2025.

[21] A. M. Mota, "Ai-based characterization of breast cancer in mammography and tomosynthesis: a review of radiomics and deep learning for subtyping, staging, and prognosis," *Cancers*, vol. 17, no. 20, p. 3387, 2025.

[22] N. Hosseinzadeh, S. Behrouzieh, R. Sharifi, and N. Sedighi, "Non-invasive breast cancer receptor typing from mammograms using artificial intelligence: a systematic review and meta-analysis," *Journal of Imaging Informatics in Medicine*, 2026.

Avaliação Operacional da Triagem de Retinopatia Diabética sob Restrições de Revisão e Mudança de Distribuição

1st Carlos Alberto da Silva Brasil
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
carlos.brasil@alu.ufc.br

2nd Paulo Cesar Cortez
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
cortez@lesc.ufc.br

3rd Victor Hugo Costa de Albuquerque
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
victor.albuquerque@ieee.org

***Resumo*—Este estudo analisou retrospectivamente a triagem operacional de retinopatia diabética (DR) referenciável em cenário de restrição de orçamento de revisão e de mudança de distribuição interna–externa. O objetivo principal foi avaliar um protocolo operacional integrado que reúne discriminação, calibração com escalonamento de temperatura, incerteza por Monte Carlo dropout, captura sob orçamento de revisão e mapeamento auditável de prioridades por camada fuzzy. Como instanciação comparativa desse protocolo, a arquitetura EfficientNet-B0 com cabeça inspirada em KAN foi confrontada com uma baseline EfficientNet-B0+MLP, treinadas nos mesmos subconjuntos e com três sementes. Essa abordagem evitou que a avaliação ficasse limitada apenas a métricas puramente agregadas. Os resultados mostraram desempenho discriminativo comparável entre KAN e MLP, com AUC interna de 0.905 ± 0.003 e 0.904 ± 0.003, e externa de 0.959 ± 0.003 e 0.956 ± 0.006, respectivamente; calibração da KAN com Brier scores de 0.075 ± 0.001 (interno) e 0.095 ± 0.008 (externo); captura orçamentária na distribuição externa de 0.239 ± 0.003 dos casos referenciáveis no top-10% e 0.988 ± 0.002 no top-50%; e comportamento da categoria fuzzy de encaminhamento urgente, que absorveu 48.3% da carga de trabalho externa e preservou 0.974 ± 0.009 dos casos referenciáveis externos, o que reitera a necessidade de avaliação integrada em discriminação, calibração, incerteza, captura orçamentária e auditabilidade.**

***Palavras-chave*—Retinopatia diabética, retinopatia diabética referenciável, triagem de imagens médicas, mudança de distribuição, estimação de incerteza, Monte Carlo dropout, escalonamento de temperatura, calibração, sistemas de inferência fuzzy, avaliação sob orçamento de revisão.**

I. Introdução

A retinopatia diabética (DR) permanece como causa relevante de perda visual evitável em pessoas com diabetes, e a detecção precoce da forma referenciável é central nos programas de rastreamento. Diante do volume de exames que pode sobrecarregar especialistas, o suporte automatizado é mais útil quando tratado como ferramenta de priorização de revisão, não apenas como classificador binário [1]–[5]. Métricas agregadas, como AUC, sensibilidade, especificidade e F1-score, continuam necessárias, mas não descrevem diretamente o desempenho sob orçamento de revisão, os casos referenciáveis não detectados, a redução da carga de trabalho, a segurança das decisões de não encaminhamento ou a precisão entre exames revisados.

A avaliação sob deslocamento de distribuição revela que divergências entre bases, dispositivos, aquisição, prevalência, qualidade de imagem, rotulagem e pré-processamento alteram discriminação, calibração e perfis de erro [6]–[9], de modo que AUC elevada em teste externo pode coexistir com confiabilidade probabilística reduzida e riscos assimétricos. Falsos negativos confiantes constituem o modo de falha de maior relevância operacional: políticas fundamentadas exclusivamente em probabilidade ou incerteza podem postergar casos referenciáveis com baixa probabilidade predita, baixa incerteza estimada e alta confiança.

Este estudo avaliou retrospectivamente uma estrutura de priorização de revisão para DR sob orçamento de revisão e mudança de distribuição interna–externa, sem propor uma arquitetura diagnóstica clinicamente validada. A contribuição principal é um protocolo de avaliação operacional que integra ranqueamento por probabilidade, calibração, incerteza, captura sob orçamento, categorias fuzzy e análise de modos de falha. A comparação entre EfficientNet-B0 com cabeça inspirada em KAN [10], [11] e EfficientNet-B0+MLP, conduzida nos mesmos subconjuntos e sementes, é utilizada como instanciação comparativa desse protocolo, e não como demonstração de superioridade arquitetural. Complementam a avaliação o escalonamento de temperatura, o MC-dropout, os intervalos bootstrap e a camada fuzzy como mapeamento auditável de prioridade, não como substituto do ranqueamento por probabilidade ou de regras clínicas.

II. Trabalhos Relacionados

Modelos convolucionais e transfer learning [1]–[5] têm produzido alta discriminação para DR referenciável em fotografias coloridas de fundo de olho, com métricas como AUC, sensibilidade, especificidade e F1-score. Esses índices, porém, não garantem triagem segura quando a revisão especializada é limitada ou quando há deslocamento dos dados. Diferenças em dispositivos, aquisição, composição dos casos, prevalência, qualidade da imagem e rotulagem podem afetar discriminação, calibração e assimetria dos erros [6]–[9]; por isso, a interpretação de testes externos deve considerar prevalência, confiabilidade probabilística, incerteza e casos referenciáveis residuais. Calibração, quantificação de incerteza, classificação seletiva, predição conformal e

aprendizado com encaminhamento estimam confiabilidade, variabilidade preditiva, cobertura, abstenção e trade-offs de revisão [12]–[21], mas não eliminam necessariamente falsos negativos confiantes, que podem apresentar baixa incerteza e alta confiança. Sistemas fuzzy fornecem funções de pertinência explícitas e regras interpretáveis [22]–[25]; neste estudo, essa abordagem é tratada como camada de auditabilidade, não como regra clínica definitiva. Com esse enquadramento, o manuscrito avalia classificadores com o mesmo backbone e protocolo pareado, calibração, incerteza, ranqueamento por probabilidade, triagem fuzzy, captura sob orçamento de revisão e risco residual, em consonância com diretrizes de relato sobre uso pretendido, validação, calibração, contexto clínico e limitações [26], [27].

III. Metodologia

Esta seção descreve a formulação retrospectiva da triagem, os subconjuntos fixos de avaliação, os modelos, as estimativas de incerteza, a camada fuzzy, as métricas sob orçamento de revisão e os resumos estatísticos. A avaliação foi estruturada como fluxo de priorização de revisão, não como sistema diagnóstico autônomo. Todas as análises usaram apenas subconjuntos fixos, modelos previamente treinados, tabelas de predição e resumos do protocolo experimental predefinido; nenhum treinamento adicional foi realizado para este manuscrito.

A. Desenho do Estudo e Formulação do Problema de Triagem

Este estudo retrospectivo estruturou a triagem de DR referenciável em imagens coloridas de fundo de olho como priorização de revisão sob capacidade limitada. Cada imagem recebeu um rótulo binário de DR referenciável, e o classificador estimou a probabilidade dessa condição. A análise primária foi guiada pelo ranqueamento dos exames para revisão especializada sob orçamentos fixos, embora métricas binárias convencionais também tenham sido reportadas.

Avaliaram-se, nos mesmos casos, três estratégias de priorização para DR referenciável: ranqueamento por probabilidade, priorização por incerteza como baseline de ablação e triagem fuzzy guiada por probabilidade como mapeamento operacional inspecionável, não como diagnóstico autônomo. O objetivo foi caracterizar a captura de casos referenciáveis no subconjunto revisado e a ocorrência de patologias não detectadas sob priorização operacional. Situações inseguras de não encaminhamento corresponderam a exames referenciáveis incorretamente direcionados ou mantidos abaixo do limiar de revisão; falsos negativos confiantes corresponderam a exames referenciáveis com baixa probabilidade predita, baixa incerteza e alta confiança. Os subconjuntos interno e externo foram analisados separadamente quanto à discriminação, calibração, incerteza, distribuição da triagem e desempenho sob orçamento em cenário de mudança de distribuição interna–externa.

B. Descrição do Dataset e Subconjuntos Experimentais

A triagem retrospectiva binária de DR referenciável usou imagens de fundo de olho de `carlos-brasil/dr-triage-kan-fuzzy-dataset`; graus 0–1 foram mapeados como não referenciáveis, e graus 2–4 como referenciáveis. Foram contempladas nos subconjuntos fixos as 29,857 imagens de treinamento, as 5,269 de validação (com uma estratificação de 15% configurada por `referable_dr` sob a semente 42), as 53,576 de teste interno e as 3,662 de teste externo, as quais foram divididas entre o EyePACS 2015, representando o domínio interno, e o APTOS 2019, voltado ao teste externo. Treinamento e validação foram usados para ajuste dos modelos e suporte à calibração; os testes interno e externo avaliaram, respectivamente, o domínio-fonte e o cenário deslocado.

As verificações de integridade foram guiadas pelos identificadores disponíveis. Não houve evidência de sobreposição em nível de imagem, e a sobreposição em nível de paciente apareceu apenas entre treinamento e validação, o que exige cautela analítica. As imagens foram processadas em RGB, redimensionadas para 224×224 pixels e, no treinamento, receberam aumento de dados por espelhamento aleatório (probabilidade 0.5) e rotação de até 10 graus. Na avaliação, aplicaram-se apenas normalização ImageNet e conversão para tensor. O subconjunto externo permaneceu fora do treinamento e da seleção de limiares; seus resultados foram interpretados como evidência sob deslocamento em relação ao EyePACS. A Tabela I resume os testes e o desempenho médio da KAN.

Tabela I
Resumo dos conjuntos de teste interno e externo, com prevalência das bases e desempenho da KAN relatado como média ± desvio-padrão em três sementes

Subconjunto de avaliação	Casos	Referenciáveis n (%)	AUC	Brier
Teste interno	53,576	10,281 (19.2%)	0.905 ± 0.003	0.075 ± 0.001
Teste externo	3,662	1,487 (40.6%)	0.959 ± 0.003	0.095 ± 0.008

C. Modelo de Classificação

O classificador de DR referenciável usou um backbone EfficientNet-B0 e uma cabeça de duas classes. Para a imagem x_i, definiram-se a representação visual, os logits e a probabilidade da classe referenciável.

$$\begin{aligned} z_i &= g_\phi(x_i), \\ l_i &= h_\psi(z_i) \in \mathbb{R}^2, \\ P_{r,i} &= \frac{\exp(l_{i,1})}{\exp(l_{i,0}) + \exp(l_{i,1})}, \end{aligned} \tag{1}$$

em que $g_\phi(\cdot)$ denota o backbone e $h_\psi(\cdot)$ a cabeça de classificação. A EfficientNet-B0 da `torchvision` foi inicializada com pesos pré-treinados padrão. O classificador original foi substituído por uma cabeça inspirada em KAN, composta por duas camadas totalmente conectadas com 256 unidades, cada uma seguida por ativação spline treinável baseada em tanh e dropout de 0.2, além de uma camada linear final de duas classes.

Os modelos foram treinados com perda de entropia cruzada,

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{1} \mathbb{I}(y_i = k) \log \left(\mathrm{softmax}(l_i)_k \right), \tag{2}$$

em que $y_i \in \{0, 1\}$, foram mobilizados na execução do treinamento o otimizador AdamW, uma taxa de aprendizado

estipulada em 1×10^{-4}, o decaimento de pesos em 1×10^{-4}, o tamanho de lote fixado em 32, a extensão de 5 épocas e o mapeamento pelas sementes 42, 123 e 2026. Sob essa mesma lógica de reprodutibilidade, processou-se diretamente nas frentes do Python, NumPy, PyTorch e CUDA o estrito controle da aleatoriedade, diretriz que também norteou a ativação do cuDNN em modo determinístico e a desabilitação completa do benchmarking.

A avaliação utilizou a probabilidade softmax da classe referenciável e um limiar fixo de 0,5 para o cálculo das métricas binárias. A apresentação dos resultados concentrou-se em AUC, Brier score, calibração binária, falsos positivos e falsos negativos, captura sob orçamento de revisão e precisão entre os casos revisados, com médias entre sementes quando disponíveis, sem extrapolação além dos resumos relatados.

O comparador consistiu no backbone EfficientNet-B0 acoplado a uma cabeça MLP convencional, camada totalmente conectada de 256 unidades, ReLU, dropout de 0,2 e camada linear de duas classes, segundo os mesmos subconjuntos, pré-processamento, sementes, otimizador, limiar, métricas de calibração e análises sob orçamento de revisão. Desse modo, a comparação configura um pareamento por backbone e protocolo, e não uma ablação de capacidade idêntica, uma vez que as cabeças diferem em profundidade e parametrização interna.

D. Quantificação de Incerteza

Este procedimento operacional utilizou estimativas de incerteza na condição de descritores auxiliares voltados à priorização de revisão e à análise de modos de falha, sem caracterizá-las como indicadores definitivos de correção ou segurança clínica. A inferência por MC-dropout foi conduzida com o mesmo backbone EfficientNet-B0 e com a cabeça inspirada em KAN adotados no pipeline principal, sendo aplicada aos subconjuntos de validação, teste interno e teste externo nas sementes 42, 123 e 2026.

Na etapa de inferência, o modelo foi mantido em modo de avaliação, preservando-se ativas apenas as camadas de dropout para estimativa estocástica da incerteza. Cada imagem foi processada por meio de $T = 10$ passagens forward, das quais foram obtidas as probabilidades de classe referenciável $q_i^{(t)}{}_{t=1}^{T}$.

$$\begin{aligned} \bar{q}_i &= \frac{1}{T}\sum_{t=1}^{T} q_i^{(t)}, \\ H_i &= -\bar{q}_i \log(\bar{q}_i) - (1-\bar{q}_i)\log(1-\bar{q}_i), \\ U_{e,i} &= \frac{H_i}{\log(2)}, \\ \sigma_{\mathrm{MC},i} &= \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(q_i^{(t)} - \bar{q}_i\right)^2}, \\ C_i &= \max(P_{r,i}, 1 - P_{r,i}). \end{aligned} \tag{3}$$

Aqui, $\bar{q}_i$ é a probabilidade média por MC-dropout, H_i a entropia preditiva, $U_{e,i}$ a entropia normalizada, $\sigma_{\mathrm{MC},i}$ o desvio-padrão por MC-dropout e C_i a confiança determinística calculada a partir de $P_{r,i}$. No controle negativo baseado apenas em incerteza, a variabilidade MC foi normalizada dentro de cada subconjunto,

$$\begin{aligned} U_{a,i} &= \frac{\sigma_{\mathrm{MC},i} - \min_j \sigma_{\mathrm{MC},j}}{\max_j \sigma_{\mathrm{MC},j} - \min_j \sigma_{\mathrm{MC},j} + \epsilon}, \\ r_i^{\mathrm{unc}} &= \frac{U_{e,i} + U_{a,i} + (1 - C_i)}{3}. \end{aligned} \tag{4}$$

Esse escore não incorporou a direção da probabilidade de doença e foi usado apenas como baseline negativo de priorização.

Esses descritores apoiaram a avaliação da calibração e a análise de modos de falha. Sua interpretação foi estritamente empírica, pois casos referenciáveis ainda podem receber baixa probabilidade predita, baixa incerteza estimada e alta confiança do modelo.

E. Comparadores Baseline e Protocolos de Calibração/Seleção

Esta análise comparativa foi conduzida sob desenho pareado por backbone e protocolo, confrontando EfficientNet-B0 com cabeça inspirada em KAN à baseline EfficientNet-B0+MLP, com os mesmos subconjuntos, sementes, pré-processamento, limiar e avaliação downstream. O objetivo foi verificar se, sob esse protocolo, a cabeça inspirada em KAN alterou o comportamento retrospectivo da triagem em relação a uma cabeça neural convencional.

Todas as análises determinísticas utilizaram a probabilidade de DR referenciável p_i para a imagem x_i, com rótulo binário $y_i \in \{0, 1\}$. As predições binárias usaram limiar fixo de 0.5,

$$\hat{y}_i = \mathbb{I}(p_i \geq 0.5). \tag{5}$$

As métricas de classificação incluíram AUC, acurácia, acurácia balanceada, F1-score, sensibilidade, especificidade e Brier score,

$$\mathrm{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2. \tag{6}$$

Na análise sob orçamento de revisão, os casos foram ranqueados pela probabilidade predita de DR referenciável. Para um orçamento $b \in (0, 1]$, os casos do topo

$$K_b = \lceil bN \rceil \tag{7}$$

foram selecionados para revisão simulada por especialista. Seja $\mathcal{S}_b$ esse subconjunto. A captura de casos referenciáveis foi

$$\mathrm{Capture}(b) = \frac{\sum_{i \in \mathcal{S}_b} y_i}{\sum_{i=1}^{N} y_i}. \tag{8}$$

A precisão dentro do subconjunto revisado foi

$$\mathrm{Precision}_{\mathrm{review}}(b) = \frac{\sum_{i \in \mathcal{S}_b} y_i}{|\mathcal{S}_b|}. \tag{9}$$

O escalonamento de temperatura *post-hoc* foi avaliado nas saídas probabilísticas determinísticas da KAN e da MLP. A temperatura escalar T_{cal} foi ajustada na validação para cada modelo e semente e aplicada, sem reajuste subsequente, às predições de validação, teste interno e teste externo. No cenário binário, o escalonamento foi aplicado às log-odds implicadas por p_i,

$$\tilde{p}_i(T_{\mathrm{cal}}) = \sigma\left(\frac{\log\left(\frac{p_i}{1-p_i}\right)}{T_{\mathrm{cal}}}\right), \tag{10}$$

em que $\sigma(\cdot)$ é a sigmoide logística. Como a transformação é monotônica, ela foi interpretada como ajuste de calibração, não como alteração da discriminação ou do ranqueamento por probabilidade. A calibração antes e depois do escalonamento foi avaliada por erro esperado de calibração, *negative log-likelihood*, Brier score e diagramas de confiabilidade.

A incerteza estatística foi estimada por bootstrap em nível de caso, com 1,000 reamostragens. Para a métrica m, as amostras bootstrap produziram $\{m^{(r)}\}_{r=1}^{1000}$, e o intervalo de confiança de 95% foi definido pelos percentis empíricos 2.5 e 97.5,

$$\mathrm{CI}_{95\%}(m) = \left[Q_{0.025}\left(\{m^{(r)}\}_{r=1}^{1000}\right), Q_{0.975}\left(\{m^{(r)}\}_{r=1}^{1000}\right) \right]. \tag{11}$$

Os intervalos por bootstrap foram calculados para métricas de classificação e orçamento de revisão. Na comparação KAN-versus-MLP em nível de manuscrito, os intervalos de confiança incorporaram as saídas probabilísticas médias entre sementes e foram interpretados de forma estritamente descritiva, não como testes formais de superioridade.

F. Camada de Triagem por Inferência Fuzzy

Uma camada de triagem por inferência fuzzy mapeou cada probabilidade determinística de DR referenciável para uma categoria inspecionável de prioridade de revisão. Para cada imagem x_i, essa estrutura atribuiu uma categoria e expôs o grau de pertinência da probabilidade predita p_i em conjuntos fuzzy adjacentes. O componente foi avaliado como mecanismo de auditabilidade operacional, não como substituto do ranqueamento por probabilidade nem como regra diagnóstica independente.

Sobre p_i definiram-se quatro funções de pertinência: baixa, moderada, alta e muito alta. As categorias moderada e alta usaram uma função triangular genérica,

$$\mu_{\mathrm{tri}}(p; a, b, c) = \begin{cases} 0, & p \le a \text{ ou } p \ge c, \\ \dfrac{p-a}{b-a}, & a < p < b, \\ 1, & p = b, \\ \dfrac{c-p}{c-b}, & b < p < c, \end{cases} \tag{12}$$

com $\mu_{\mathrm{moderate}}(p_i) = \mu_{\mathrm{tri}}(p_i; 0.10, 0.35, 0.60)$ e $\mu_{\mathrm{high}}(p_i) = \mu_{\mathrm{tri}}(p_i; 0.40, 0.65, 0.90)$. Como os conjuntos baixo e muito alto correspondem aos ombros esquerdo e direito de trapézios degenerados, suas fronteiras foram escritas explicitamente para evitar divisão por zero na forma trapezoidal genérica:

$$\mu_{\mathrm{low}}(p) = \begin{cases} 1, & 0 \le p \le 0.10, \\ \dfrac{0.30-p}{0.20}, & 0.10 < p < 0.30, \\ 0, & p \ge 0.30, \end{cases} \tag{13}$$

$$\mu_{\text{very high}}(p) = \begin{cases} 0, & p \le 0.70, \\ \dfrac{p-0.70}{0.20}, & 0.70 < p < 0.90, \\ 1, & 0.90 \le p \le 1.00. \end{cases} \tag{14}$$

A pertinência dominante foi definida pelo maior grau, e eventuais empates foram resolvidos em favor da categoria de maior prioridade. Para fins operacionais, as categorias fuzzy foram mapeadas como segue: baixa correspondeu a não encaminhamento, moderada a revisão de rotina, alta a revisão prioritária e muito alta a prioridade urgente de revisão. Esses rótulos foram usados apenas como categorias auditáveis de carga de trabalho e prioridade, não como decisão clínica autônoma.

A proximidade da fronteira foi quantificada pela margem fuzzy entre o maior e o segundo maior grau de pertinência,

$$m_i = \mu_{(1)}(p_i) - \mu_{(2)}(p_i), \tag{15}$$

em que $\mu_{(1)}(p_i)$ e $\mu_{(2)}(p_i)$ são o maior e o segundo maior valores de pertinência para x_i. Valores pequenos de m_i indicaram proximidade entre conjuntos fuzzy adjacentes.

Este estudo avaliou retrospectivamente a camada fuzzy por meio da fração de casos por categoria, da fração de referenciáveis, da captura de casos referenciáveis, da taxa residual de referenciáveis entre os casos não encaminhados e da fração de casos de fronteira. O objetivo foi investigar se esse componente fornecia um resumo operacional interpretável dos escores, sem implicar melhora discriminativa em relação ao ranqueamento por probabilidade nem validação clínica como regra independente de encaminhamento.

G. Análise Estatística

As análises estatísticas foram conduzidas a partir das tabelas fixas de predição obtidas na validação e nos testes interno e externo. A classificação sob limiar fixo de 0,5 foi caracterizada por AUC, sensibilidade, especificidade, F1-score e matriz de confusão, enquanto a calibração foi avaliada por Brier score, erro esperado de calibração e negative log-likelihood. Sensibilidade e especificidade foram definidas como:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

e

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{17}$$

em que TP, TN, FP e FN denotam verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. O Brier score foi:

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2. \tag{18}$$

O erro esperado de calibração foi calculado para a probabilidade binária de DR referenciável. Com $B_m = \{i : p_i \in I_m\}$, $n_m = |B_m|$ e M bins, comparou-se, em cada bin, a fração observada de referenciáveis com a probabilidade média predita:

$$\mathrm{ECE}_{\mathrm{bin}} = \sum_{m=1}^{M} \frac{n_m}{N} \left| \frac{1}{n_m} \sum_{i \in B_m} y_i - \frac{1}{n_m} \sum_{i \in B_m} p_i \right|. \tag{19}$$

A análise final usou $M = 10$ bins de largura igual.

A incerteza das estimativas pontuais foi quantificada por bootstrap não paramétrico em nível de caso, aplicado separadamente por subconjunto e estratégia de priorização às tabelas fixas de predição [28]. Para cada métrica m, 1,000 reamostragens geraram os valores $\{m^{(r)}\}_{r=1}^{1000}$, e o intervalo de confiança de 95% foi definido por:

$$\mathrm{CI}_{95\%}(m) = \left[Q_{0.025}\left(\{m^{(r)}\}_{r=1}^{1000}\right), Q_{0.975}\left(\{m^{(r)}\}_{r=1}^{1000}\right) \right]. \tag{20}$$

Às métricas de classificação, calibração e captura sob orçamento aplicou-se o método de bootstrap, enquanto na média das saídas probabilísticas entre sementes fundamentou-se a comparação KAN versus MLP. Atentou-se também para a avaliação de métricas operacionais nos subconjuntos interno e externo, o que envolveu taxas de detecção, casos referenciáveis não detectados, precisão entre casos revisados, redução da carga de trabalho, atribuições inseguras de não encaminhamento e falsos negativos confiantes. Como não se buscou na aplicação de testes formais de superioridade a evidência primária deste estudo, ganharam o sentido de resultados descritivos e específicos de cada subconjunto os intervalos de confiança, o ranqueamento por probabilidade, a priorização por incerteza, a triagem fuzzy e os exemplos Grad-CAM, sem que isso signifique uma robustez pronta para implantação.

IV. Resultados

A. Mudança de Distribuição dos Dados e Resultados de Classificação sob Protocolo Pareado

No teste interno EyePACS, a prevalência de retinopatia diabética referenciável foi 19.19% em 53,576 imagens. No teste externo APTOS, essa taxa subiu para 40.61% em 3,662 imagens, aproximadamente 2.12 vezes maior. Por isso, os resultados externos foram interpretados como comportamento em cenário deslocado, não como prova de robustez pronta para implantação.

A comparação pareada caracterizou discriminação semelhante entre as cabeças KAN e MLP — KAN com AUC de 0.905 ± 0.003 no teste interno e 0.959 ± 0.003 no externo; MLP com 0.904 ± 0.003 e 0.956 ± 0.006, respectivamente, em três sementes, com variabilidade sobreposta sustentando interpretação estritamente descritiva, sem alegar superioridade da cabeça KAN nem equivalência de capacidade.

A redução do limiar operacional de 0.50 para 0.20 diminuiu os falsos negativos na KAN em 1028.7 casos internos e 9.3 externos, com acréscimo de 2131.0 e 122.0 falsos positivos; na MLP, reduziu-os em 1330.7 e 25.7, com acréscimo de 2740.0 e 158.7 falsos positivos, respectivamente. O ajuste alterou o ponto operacional sem evidenciar robustez aprimorada.

As métricas de calibração também exigem cautela. O Brier score da KAN aumentou de 0.075 ± 0.001 no teste interno para 0.095 ± 0.008 no externo; na MLP, passou de 0.075 ± 0.001 para 0.096 ± 0.010. Como prevalência e incerteza basal do desfecho afetam diretamente o Brier bruto, esse aumento isolado não prova piora definitiva da calibração externa. A interpretação exigiu avaliação conjunta com diagramas de confiabilidade, ECE binário e NLL.

B. Resultados de Calibração e Ranqueamento sob Orçamento de Revisão

O escalonamento de temperatura foi ajustado por modelo e semente na validação e aplicado, sem reajuste, à validação, ao teste interno e ao teste externo. Na média entre sementes, a KAN obteve Brier scores de 0.073 ± 0.001 na validação, 0.075 ± 0.001 internamente e 0.095 ± 0.008 externamente; os valores da MLP foram 0.073 ± 0.000, 0.075 ± 0.001 e 0.096 ± 0.010. Como as prevalências internas e externas diferiram, o Brier foi interpretado como dependente da prevalência, e a calibração externa foi avaliada junto aos diagramas de confiabilidade, ECE binário e NLL. A Fig. 1 mostra os diagramas de confiabilidade.

Tabela II
Leitura operacional do orçamento externo para o ranqueamento por probabilidade da KAN. O teto máximo considera a prevalência externa e o número de exames que cabem no orçamento.

Orçamento	Revisados	Teto cap.	Cap. KAN	Ref. não capt.	Prec. rev.
Top-10%	367	0.247	0.239	1132	96.7%
Top-20%	733	0.493	0.465	796	94.3%
Top-30%	1099	0.739	0.671	489	90.8%
Top-40%	1465	0.985	0.851	222	86.3%
Top-50%	1831	1.000	0.988	18	80.2%

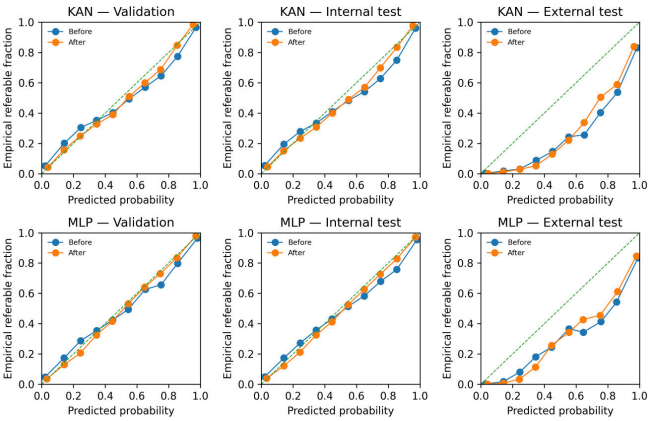


Fig. 1. Diagramas de confiabilidade antes e depois do escalonamento de temperatura ajustado na validação. A calibração melhorou nas predições de validação e de teste interno, mas foi transferida de forma imperfeita para o subconjunto externo deslocado.

A análise sob orçamento avaliou o ranqueamento por probabilidade quando a revisão especializada é restrita. No teste externo, na média entre sementes, a KAN capturou 0.239 ± 0.003, 0.465 ± 0.005, 0.671 ± 0.010, 0.851 ± 0.005 e 0.988 ± 0.002 dos casos referenciáveis nos orçamentos top-10%, 20%, 30%, 40% e 50%, respectivamente. Como o teste externo tinha 1,487 casos referenciáveis entre 3,662 imagens, o orçamento top-10% permitia revisar apenas 367 exames e impunha teto teórico de captura de 0.247; a captura da KAN, 0.239, ficou próxima desse limite. O composto baseado apenas em incerteza, usado como controle negativo, capturou apenas 0.051 ± 0.002, 0.161 ± 0.003 e 0.357 ± 0.007 nos orçamentos top-10%, top-30% e top-50%, respectivamente. Esse resultado indica que a incerteza isolada não deve assumir o papel de critério primário de triagem.

C. Resultados de Auditabilidade da Triagem Fuzzy

A análise da camada de triagem fuzzy sintetizou as saídas fixas do modelo em categorias auditáveis e inspecionáveis. A categoria urgente não deve ser interpretada como equivalente ao orçamento top-10%, pois correspondeu a 48.3% dos casos externos na KAN e 47.2% na MLP; portanto, sua captura caracteriza carga de trabalho auditável, não ranqueamento pareado sob orçamento. Na KAN, os casos de fronteira permaneceram pouco frequentes, com 1.3% no conjunto interno e 0.9% no externo. As frações residuais de casos

Tabela III
RESUMO INTEGRATIVO MÉDIO ENTRE SEMENTES. A CATEGORIA FUZZY DE PRIORIDADE URGENTE DE REVISÃO É APRESENTADA COM SUA CARGA DE TRABALHO, CAPTURA E PUREZA, EVITANDO COMPARAÇÃO DIRETA COM O ORÇAMENTO TOP-10%.

Modelo	Subconj.	AUC	Brier	Top-10 cap.	Carga urg.	Cap. urg.	Pureza urg.
KAN	Interno	0.905 ± 0.003	0.075 ± 0.001	–	11.1%	0.536	92.7%
KAN	Externo	0.959 ± 0.003	0.095 ± 0.008	0.239 ± 0.003	48.3%	0.974 ± 0.009	82.0%
MLP	Interno	0.904 ± 0.003	0.075 ± 0.001	–	10.9%	0.529	93.0%
MLP	Externo	0.956 ± 0.006	0.096 ± 0.010	0.236 ± 0.006	47.2%	0.958 ± 0.020	82.4%

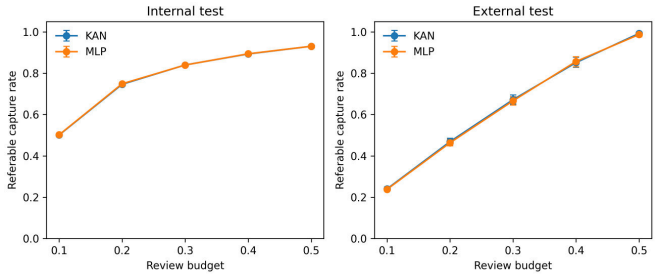


Fig. 2. Incerteza por bootstrap em torno das estimativas de captura sob orçamento de revisão. Os classificadores com cabeça inspirada em KAN e com cabeça MLP apresentaram comportamento de captura semelhante ao longo das cargas de trabalho de revisão.

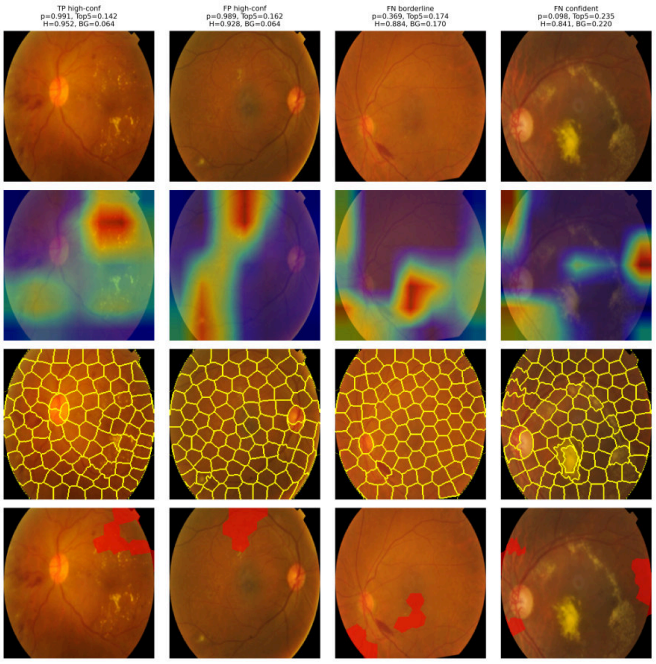


Fig. 3. Inspeção por Masked Grad-CAM/SLIC de casos externos representativos: verdadeiro positivo de alta confiança, falso positivo de alta confiança, falso negativo borderline e falso negativo confiante.

referenciáveis classificados como não encaminhamento foram de 27.2% e 0.4%, respectivamente.

A Tabela III separa o desempenho do classificador do comportamento operacional da camada fuzzy. Essa distinção reforça que a captura da categoria de prioridade urgente representa uma categoria auditável de carga de trabalho, não um método de ranqueamento sob o mesmo orçamento nem evidência de prontidão clínica autônoma.

D. Orçamento de Revisão e Métricas Operacionais Clínicas

A análise sob orçamento mensurou a captura de casos referenciáveis quando a revisão especializada é restrita. O ranqueamento por probabilidade foi a principal baseline operacional, enquanto a priorização apenas por incerteza atuou como controle negativo. A triagem fuzzy foi avaliada como mapeamento categórico inspecionável, não como regra diagnóstica.

KAN e MLP apresentaram captura de 0.239 ± 0.003 e 0.236 ± 0.006 dos referenciáveis sob orçamento de 10%, elevando-se para 0.988 ± 0.002 e 0.985 ± 0.004 a 50%, caracterizando o trade-off carga de trabalho–captura: mais exames revisados, menos casos referenciáveis não detectados. A Tabela II detalha os resultados operacionais da KAN (teto teórico, referenciáveis não capturados, precisão entre revisados); a Fig. 2 exibe a incerteza bootstrap dessas estimativas.

A triagem fuzzy complementou a auditoria ao resumir carga de trabalho, captura urgente, pureza das categorias, casos de fronteira e casos referenciáveis residuais em baixa prioridade. Como baseline operacional mais forte, manteve-se o ranqueamento por probabilidade; à triagem fuzzy, por sua vez, coube o papel de resumo inspecionável de prioridade, sem constituir evidência de prontidão clínica autônoma.

E. Modos de Falha e Inspeção por Masked Grad-CAM/SLIC

A auditoria de modos de falha usou casos selecionados do teste externo, avaliados pelo modelo inspirado em KAN com semente 42 e limiar fixo de DR referenciável de $p = 0.5$.

Tabela IV
ANÁLISE EXPLORATÓRIA POR MASKED GRAD-CAM/SLIC PARA VERDADEIROS POSITIVOS DE ALTA CONFIANÇA VERSUS FALSOS NEGATIVOS CONFIANTES. VALORES EM MEDIANA [IQR]; $\Delta_{\mathrm{FN-TP}}$ É A DIFERENÇA MÉDIA FN CONFIANTE MENOS TP ALTA CONFIANÇA. OS VALORES DE p SÃO EXPLORATÓRIOS E CONDICIONADOS À AMOSTRA DE AUDITORIA SELECIONADA, NÃO CONSTITUINDO INFERÊNCIA EM NÍVEL POPULACIONAL.

Métrica	TP alta conf.	FN confiante	$\Delta_{\mathrm{FN-TP}}$	IC 95%	p BH (expl.)
Energia top-5	0.143 [0.126, 0.151]	0.238 [0.194, 0.276]	0.094	[0.067, 0.122]	9.07e-05
Entropia	0.949 [0.932, 0.957]	0.823 [0.811, 0.846]	-0.118	[-0.135, -0.100]	2.88e-04
Gini	0.375 [0.341, 0.433]	0.656 [0.622, 0.679]	0.263	[0.225, 0.301]	1.44e-04
Superpixels efetivos	76.387 [72.563, 78.040]	42.974 [40.403, 47.973]	-31.192	[-35.083, -27.258]	9.59e-05
Energia de fundo	0.072 [0.059, 0.093]	0.201 [0.158, 0.223]	0.094	[0.037, 0.142]	1.63e-02

A análise incluiu quatro grupos, totalizando 60 imagens: 20 verdadeiros positivos de alta confiança ($y = 1$, $\hat{y} = 1$, $p \geq 0.9$), 20 falsos positivos de alta confiança ($y = 0$, $\hat{y} = 1$, $p \geq 0.9$), 10 falsos negativos borderline logo abaixo do limiar e 10 falsos negativos confiantes com as menores probabilidades preditas. Os mapas Grad-CAM [29] da classe de DR referenciável foram projetados e agregados sobre superpixels SLIC retinianos definidos por $n_{\mathrm{segments}} = 100$ e compactness $= 10$. As métricas avaliadas foram energia Grad-CAM top-5, entropia de saliência normalizada, concentração de Gini, número efetivo de superpixels ativados e fração de energia de fundo/borda. O contraste tabulado usa $\Delta_{\mathrm{FN-TP}} = n_{\mathrm{FN}}^{-1} \sum_{i \in \mathrm{FN_{conf}}} s_i - n_{\mathrm{TP}}^{-1} \sum_{i \in \mathrm{TP_{high}}} s_i$, em que s_i é a métrica Grad-CAM/SLIC; os valores de p são descritores exploratórios condicionados ao conjunto de auditoria selecionado, não inferência clínica populacional.

Esta análise avaliou falsos negativos confiantes contra verdadeiros positivos de alta confiança (Fig. 3, Tabela IV), caracterizando falhas por saliência concentrada periférica/não retiniana, com maior energia top-5, menor entropia, maior Gini, menos superpixels efetivos e maior energia de fundo/borda, e não por mera ausência de ativação. Avaliou-se a direção das métricas primárias nas nove configurações de

segmentação SLIC $n_{\text{segments}} \in \{50, 100, 150\}$ e compactness $\in \{5, 10, 20\}$, constatando-se sua preservação em todas as condições examinadas. O Masked Grad-CAM/SLIC constitui, portanto, uma auditoria exploratória de falhas, cujos valores de p permanecem condicionados à amostra selecionada, sem estabelecer correção clínica em nível de lesão.

V. Discussão

Este estudo demonstra que a avaliação da triagem operacional de DR referenciável deve ir muito além das métricas agregadas de discriminação. Desempenhos semelhantes em discriminação, calibração e comportamento sob orçamento de revisão foram apresentados pelas cabeças inspiradas em KAN e MLP, tanto no teste interno quanto no externo, de modo que na variabilidade sobreposta entre sementes não encontra sustentação uma superioridade robusta da cabeça KAN sob o protocolo pareado. Como principal achado consolida-se a necessidade de avaliar conjuntamente o ranqueamento por probabilidade, a calibração, a captura sob orçamento, os casos referenciáveis não detectados, as categorias fuzzy e os modos de falha. Esse ponto acabou reforçado pelo teste externo deslocado no APTOS, cenário no qual, devido à maior prevalência de DR referenciável, viram-se comprometidas a transferência de limiares e a operação sob orçamento, a despeito de ter sido registrada AUC externa elevada. Sob o orçamento top-10%, conseguiu-se capturar apenas cerca de um quarto dos casos referenciáveis externos, ao passo que da totalidade aproximou-se a recuperação no top-50%. Mais informativos do que a suposição de uma transferibilidade direta entre diferentes cenários revelam-se, portanto, os relatos específicos por orçamento.

Calibração, incerteza e camada fuzzy não eliminaram riscos residuais: o escalonamento de temperatura melhorou a calibração na validação e no teste interno, mas apresentou transferência imperfeita ao teste externo e não alterou o ranqueamento; a incerteza via MC-dropout evidenciou predições ambíguas, embora o ranqueamento exclusivo por incerteza tenha capturado menos casos externos referenciáveis do que o ranqueamento por probabilidade, mantendo os falsos negativos confiantes como limitação operacional persistente; e a camada fuzzy constituiu mecanismo de auditabilidade operacional, capaz de resumir carga de trabalho, prioridade urgente, pureza das categorias, casos de fronteira e referenciáveis residuais de baixa prioridade, e não evidência de robustez clínica autônoma. A sobreposição em nível de paciente, restrita ao treinamento e à validação, pode tornar ligeiramente otimistas as análises de calibração nesse estágio, sem afetar as estimativas dos testes interno e externo.

VI. Limitações

Com bases públicas de DR e rótulos em nível de imagem, este estudo retrospectivo foi conduzido sem adjudicação clínica prospectiva, estudo com leitores ou ground truth em nível de lesão. Dessa forma, descreve-se o comportamento do modelo em subconjuntos predefinidos, sem comprovação de efetividade clínica ou de prontidão para implantação. Também permaneceu limitado a uma única base deslocada o teste externo; embora APTOS e EyePACS apresentassem diferenças de prevalência e de distribuição de escores, não foi abrangida uma variação ampla de dispositivos, populações, espectro da doença, qualidade de imagem, rotulagem e fluxos de trabalho. Não foi realizada análise estratificada por qualidade de imagem; portanto, permanece desconhecido se discriminação, calibração, captura sob orçamento e atribuições de prioridade se mantêm consistentes entre diferentes níveis de qualidade. Como subgrupos demográficos e fairness não foram avaliados, seguem desconhecidas possíveis disparidades de desempenho, calibração, casos referenciáveis não detectados ou atribuições de baixa prioridade.

A baseline EfficientNet-B0+MLP foi utilizada como comparador pareado por backbone e protocolo, não constituindo benchmark exaustivo do estado da arte nem ablação de igual capacidade. Assim, as conclusões sobre a cabeça KAN restringem-se aos mesmos subconjuntos, sementes, protocolo de calibração e análises sob orçamento. A camada fuzzy deve ser interpretada como mapeamento auditável de prioridade, dependente de funções de pertinência predefinidas e das distribuições dos escores, e não como substituto de regras clínicas. A incerteza por MC-dropout constitui descritor auxiliar aproximado; o escalonamento de temperatura caracterizou a transferência de calibração sem reajuste externo; e os intervalos bootstrap permanecem descritivos, sem testes formais de superioridade. O Masked Grad-CAM/SLIC teve papel estritamente exploratório, pois a ausência de máscaras de lesão, bounding boxes, anotações pontuais ou estudo com leitores impede estabelecer correção clínica em nível de lesão.

VII. Conclusão

Este estudo avaliou retrospectivamente a triagem de retinopatia diabética sob restrição de capacidade, deslocamento de distribuição, calibração, incerteza e falhas residuais. Sua contribuição principal é o protocolo de avaliação operacional integrado, que amplia a análise além de métricas agregadas de discriminação ao incorporar captura sob orçamento, categorização fuzzy, intervalos de bootstrap e inspeção de falhas. A comparação pareada entre as cabeças KAN e MLP funciona, nesse contexto, como aplicação comparativa do protocolo, e não como reivindicação de superioridade arquitetural.

Desempenho comparável em discriminação, calibração e comportamento sob orçamento foi apresentado pela EfficientNet-B0 com cabeça inspirada em KAN e pela baseline EfficientNet-B0+MLP, sem que se evidenciasse superioridade sólida da cabeça KAN no protocolo pareado. À camada fuzzy coube fornecer um mapeamento inspecionável para a prioridade de revisão, devendo sua interpretação ser limitada à auditabilidade, e não estendida a regras clínicas ou modelos diagnósticos autônomos. Em conjunto, os achados reforçam que sistemas de rastreamento de DR precisam ser rigorosamente avaliados quanto à calibração, ao orçamento de revisão, ao risco residual, ao comportamento sob mudança de distribuição e aos modos de falha antes de qualquer alegação de implantação clínica.

Referências

[1] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016.

[2] D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, Dec. 2017.

[3] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *NPJ Digit. Med.*, vol. 1, no. 1, Art. no. 39, Aug. 2018.

[4] R. Sayres et al., "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, Apr. 2019.

[5] P. Ruamviboonsuk et al., "Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: A prospective interventional cohort study," *Lancet Digit. Health*, vol. 4, no. 4, pp. e235–e244, Apr. 2022.

[6] S. G. Finlayson et al., "The clinic and the cloud: Dynamic dataset shift as an ongoing challenge for artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 1, pp. e50–e59, Jan. 2021.

[7] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nat. Commun.*, vol. 11, no. 1, Art. no. 3673, Jul. 2020.

[8] A. Subbaswamy and S. Saria, "From development to deployment: Dataset shift, causality, and shift-stable models in health AI," *Biostatistics*, vol. 21, no. 2, pp. 345–352, Apr. 2020.

[9] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[11] Z. Liu et al., "KAN: Kolmogorov–Arnold networks," arXiv preprint arXiv:2404.19756, 2024.

[12] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[13] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[15] T. Postels et al., "On the practical usefulness of uncertainty in deep learning for medical imaging," *IEEE Trans. Med. Imaging*, vol. 40, no. 10, pp. 2634–2646, Oct. 2021.

[16] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 1513–1589, Mar. 2023.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[18] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2901–2907.

[19] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, Jan. 1950.

[20] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[21] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.

[22] J. M. Mendel and R. I. B. John, "Type-2 fuzzy sets made simple," *IEEE Trans. Fuzzy Syst.*, vol. 10, no. 2, pp. 117–127, Apr. 2002.

[23] J. M. Mendel, R. I. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 6, pp. 808–821, Dec. 2006.

[24] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, 2nd ed. Cham, Switzerland: Springer, 2017.

[25] O. Castillo and P. Melin, "A review on interval type-2 fuzzy logic applications in intelligent control," *Inf. Sci.*, vol. 279, pp. 615–631, Sep. 2014.

[26] A. A. H. de Hond et al., "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review," *NPJ Digit. Med.*, vol. 5, no. 1, Art. no. 2, Jan. 2022.

[27] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Art. no. e078378, Apr. 2024.

[28] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall/CRC, 1993.

[29] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

Berouti Spectral Subtraction for Electromyographic Signals During Walking

Bruno Henrique Viana Moreira
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
bruno.vianamoreira@ufpe.br

Vinícius José Guimarães do Carmo
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
vinicius.vjgc@ufpe.br

Caroline de Cássia Batista de Souza
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
caroline.bsouza@ufpe.br

Gisela Rocha de Siqueira
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
gisela.siqueira@ufpe.br

Ana Paula de Lima Ferreira
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
ana.lferreira@ufpe.br

Marco Aurelio Benedetti Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
marco.rodrigues@ufpe.br

Abstract—Surface electromyography (sEMG) is an important tool for neuromuscular analysis during functional tasks; however, the high cost of commercial systems may limit its use in clinical, educational, and applied research settings. Low-cost electromyographic systems represent promising alternatives, provided they are supported by appropriate acquisition, processing, and preliminary comparison procedures against reference equipment. This study aimed to perform a preliminary evaluation of the effect of the Berouti spectral subtraction method on electromyographic signals acquired using a low-cost system during overground walking, with a commercial system serving as the reference. Signals were acquired simultaneously using the Aquamove and Miotec systems, with electrodes positioned over adjacent regions of the tibialis anterior muscle. The experimental protocol consisted of eleven consecutive gait cycles, with gait cycles two through ten selected for analysis to minimize transient effects associated with gait initiation and termination. After preprocessing, the Berouti spectral subtraction technique was applied exclusively to the Aquamove signals. Root mean square (RMS), median frequency, and mean frequency were extracted and compared between systems using agreement and error metrics. The technique did not consistently improve RMS agreement but was associated with reduced error and bias in the spectral metrics, particularly median frequency (MDF) and mean frequency (MNF). These preliminary findings suggest that the adapted Berouti spectral subtraction approach may contribute to reducing spectral discrepancies in electromyographic signals acquired from low-cost systems, especially regarding spectral feature estimation, representing a relevant step toward the development of more accessible EMG solutions.

Keywords—Electromyography; Signal Processing; Spectral Subtraction; Noise Reduction; Gait Analysis.

I. Introduction

Surface electromyography (sEMG) is a non-invasive technique widely used to assess muscle activation in biomechanical, functional, and clinical contexts. By recording the electrical potentials generated during muscle contraction, sEMG enables the analysis of the temporal and spectral characteristics of neuromuscular activity, contributing to the evaluation of human movement, rehabilitation, motor control, and the development of health-related technologies [1]–[3].

Despite its relevance, the use of sEMG may still be limited by the high cost of commercial acquisition systems, particularly in clinical, educational, and applied research settings with limited resources. Commercial systems typically incorporate signal-conditioning circuits, filters, analog-to-digital converters, and proprietary software, which can reduce the occurrence of artifacts but also increase acquisition and maintenance costs. In this context, the development and evaluation of low-cost EMG systems represent promising alternatives for expanding access to neuromuscular assessment, provided they are supported by appropriate comparison procedures and progressive technical evaluation against reference systems [4]–[6].

However, reducing the cost of biomedical systems should not be understood merely as replacing electronic components with more affordable alternatives. For electromyographic signals acquired using low-cost systems to be used reliably, robust strategies for signal detection, conditioning, filtering, and digital processing are required to preserve physiologically relevant signal characteristics while minimizing unwanted components. Beyond cost reduction, these algorithms can be embedded into microcontroller firmware, enabling part of the signal processing to be performed locally on the device itself. In wearable biomedical systems integrated with the Internet of Things (IoT), this approach, combined with edge computing, can reduce both latency and the volume of transmitted data while decreasing dependence on continuous communication with remote servers [7], [8]. This aspect is particularly important in dynamic applications such as gait analysis, where body movement, the electrode–skin interface, and biomechanical variations increase the complexity of signal acquisition and interpretation [1], [9], [10].

Electromyographic signals are inherently susceptible to several sources of contamination, including instrumental noise, power-line interference, motion artifacts, instability at the skin–electrode interface, and crosstalk from adjacent muscles. These contaminating factors may compromise the estimation of temporal and spectral features, such as the root mean square (RMS), median frequency, and mean frequency, which are commonly used to characterize the

intensity and frequency content of muscle activity. Therefore, noise reduction techniques play a fundamental role in improving signal quality and enabling the extraction of more reliable parameters for comparative analyses between acquisition systems [9], [11].

Among the available noise reduction strategies, the spectral subtraction method proposed by Berouti stands out for its computational simplicity and its operation in the frequency domain based on estimating the noise spectrum present in the analyzed signal. Although originally developed for enhancing speech signals contaminated by acoustic noise, this approach can be adapted to other signal types when the objective is to attenuate unwanted components while preserving, as much as possible, the spectral structure of the signal of interest [12]. In this context, applying an adapted version of the Berouti method to electromyographic signals may represent a useful strategy for improving the quality of signals acquired using low-cost EMG systems.

Accordingly, the present study aimed to perform a preliminary evaluation of the effect of an adapted Berouti spectral subtraction technique on electromyographic signals acquired using a low-cost system during overground walking, using a commercial EMG system as the reference and considering both amplitude- and frequency-based signal metrics. This work is situated within the field of biomedical signal processing and seeks to contribute to the development of more accessible EMG solutions, with potential future applications in clinical, educational, and applied research settings.

II. Materials and Methods

A. Database and Signal Acquisition

Electromyographic signals were acquired from one participant during two experimental sessions using the low-cost Aquamove system, developed by the Human–Machine Interaction Laboratory at the Federal University of Pernambuco (LIHOM/UFPE), and the commercial Miotec system, which served as the reference device. Both systems simultaneously recorded the same overground walking trials. The participant walked at a self-selected comfortable speed. The two experimental sessions were treated as independent acquisition sets, each containing simultaneously recorded Aquamove and Miotec signals. Each session contributed nine analyzed gait cycles, resulting in 18 paired muscle contractions in the concatenated analysis. The experimental protocol was approved by the Research Ethics Committee of the Federal University of Pernambuco (UFPE), Brazil (approval no. 8.136.554; CAAE 92738325.0.0000.5208). Written informed consent was obtained from the participant before data collection. The Aquamove and Miotec signals were originally sampled at 1000 and 2000 Hz, respectively. The Miotec signals were subsequently resampled to 1000 Hz before comparison.

Surface electromyographic signals from the tibialis anterior muscle were acquired following the recommendations of the Surface ElectroMyoGraphy for the Non-Invasive Assessment of Muscles (SENIAM) project [13], [14]. First, an imaginary line was identified between the tip of the fibula and the tip of the medial malleolus. The recording electrodes were placed over the tibialis anterior muscle at one-third of this line measured from the fibular tip, with their orientation parallel to the muscle fibers. The reference electrode was positioned over the lateral malleolus in accordance with the SENIAM guidelines [14].

Before electrode placement, the skin was cleaned with 70% isopropyl alcohol, and hair was removed when necessary to reduce the impedance at the skin–electrode interface and minimize acquisition artifacts. Electrode placement and orientation were standardized across all recording sessions.

Because the signals from both systems were acquired simultaneously, it was not possible to place the Aquamove and Miotec electrodes in the exact same location. Therefore, the electrodes of both systems were positioned over adjacent regions of the tibialis anterior muscle while maintaining, as closely as possible, an orientation parallel to the muscle fibers and proximity to the anatomical location recommended by the SENIAM guidelines.

The Berouti spectral subtraction technique was applied exclusively to the signals acquired with the Aquamove system. Consequently, two conditions of the low-cost system were evaluated: one using only conventional preprocessing and another using conventional preprocessing followed by spectral subtraction. Both conditions were compared with the reference Miotec system.

B. Signal Preprocessing

Initially, the DC component was removed by subtracting the signal mean. Subsequently, a fourth-order Butterworth band-pass filter with cutoff frequencies of 20–400 Hz was applied to attenuate frequency components outside the muscle activity bandwidth, followed by a 60 Hz notch filter to suppress power-line interference.

C. Adaptation of the Berouti Spectral Subtraction Technique for Electromyographic Signals

In this study, an adapted version of the Berouti spectral subtraction technique, originally developed for speech processing, was implemented for application to electromyographic signals. The implementation operates on the spectral magnitude, uses an average noise profile estimated during a resting period, and incorporates signal mirroring at the boundaries and spectral interpolation to accommodate the characteristics of EMG signals.

The spectral noise profile was estimated from approximately 8 s of resting data recorded at the end of each acquisition. These recordings were subjected to the same preprocessing steps applied to the electromyographic signals. The time–frequency representation was obtained using the Short-Time Fourier Transform (STFT) with a 256-sample Hann window, 128-sample overlap, and a 256-point FFT. At the sampling frequency adopted in this study, these values correspond to a 256 milliseconds (ms) window and 128 ms overlap.

The average noise magnitude was estimated as:

$$D(k) = (1/M)\ \Sigma\square_{=1}^{M}\ |S(m,k)|, \qquad (1)$$

where (D(k)) represents the average spectral magnitude profile of the noise, (M) is the number of time frames used to estimate the spectral profile, and (S(m,k)) corresponds to the Short-Time Fourier Transform (STFT) of the noise segment. The resulting spectral profile was subsequently interpolated to match the spectral grid of the STFT computed from the electromyographic signal.

The spectral magnitude of the processed signal was computed as:

$$|\hat{X}(m,k)| = \max[|Y(m,k)| - \alpha D(k),\ \beta D(k)], \qquad (2)$$

where ($Y(m,k)$) denotes the STFT of the electromyographic signal, $|Y(m,k)|$ is its spectral magnitude, and ($\hat{X}(m,k)$) is the estimated spectral magnitude after processing, (α) represents the noise overestimation factor, (β) defines the spectral floor, and (D(k)) corresponds to the average spectral noise magnitude profile defined in Equation (1). The values of α and β were empirically selected during preliminary processing trials to provide a practical compromise between background-noise attenuation and preservation of the original electromyographic signal. The selected parameters reduced the background spectral components without excessively attenuating the muscle activation bursts or substantially altering their temporal structure. This was particularly important because subsequent muscle activation detection was based on the RMS envelope; overly aggressive spectral subtraction could attenuate or distort activation peaks and consequently affect their automatic detection and segmentation. Therefore, $\alpha = 1.3$ and $\beta = 0.08$ were retained as conservative parameters for this preliminary evaluation rather than as globally optimized values.

The original phase was preserved throughout the entire processing procedure, and the signal was reconstructed in the time domain using the Inverse Short-Time Fourier Transform (ISTFT). Finally, the samples added during the signal mirroring step were removed.

D. Signal Feature Extraction

Following preprocessing and, when applicable, application of the Berouti spectral subtraction technique, three classical electromyographic signal features were extracted: root mean square (RMS), median frequency (MDF), and mean frequency (MNF). These metrics are widely used to characterize signal amplitude and spectral content in biomechanics, gait analysis, and neuromuscular assessment applications [15]–[17].

For the window-by-window analysis, RMS was computed directly from the nonrectified time-domain signal using 150-sample windows with 75-sample overlap. At the sampling frequency of 1000 Hz, these values correspond to a 150 ms window and 75 ms overlap. The resulting temporal RMS sequence was used for signal synchronization and automatic muscle activation detection.

Muscle activation peaks were identified using an adaptive threshold. The baseline was estimated as the median of the RMS sequence, and the minimum peak height was defined as the baseline plus 25% of the difference between the maximum RMS value and the baseline. A minimum temporal distance of 0.25 s was imposed between consecutive peaks. To determine the onset and offset of each activation, the RMS sequence was smoothed using a three-point moving average. For each detected peak, the contraction threshold was defined as the baseline plus 50% of the peak-to-baseline amplitude. The onset and offset were identified by searching backward and forward from the peak until the smoothed RMS sequence crossed this threshold. Overlapping activations, as well as activations separated by no more than 50 ms, were merged into a single event. When multiple peaks were merged, the peak with the highest amplitude was retained.

After segmentation, RMS, MDF, and MNF were recalculated for each complete muscle contraction. RMS was computed directly from the nonrectified time-domain signal, whereas MDF and MNF were obtained from the power spectral density estimated using Welch's method with a 256-sample periodic Hamming window, 128-sample overlap, and a 1024-point FFT [18]. For contractions shorter than 256 samples, the Welch window length and overlap were reduced to fit the available signal segment. Spectral features were calculated within the 20–400 Hz frequency band. The overground walking protocol was performed with simultaneous acquisition of electromyographic signals using the Aquamove and Miotec systems. To achieve this, the electrodes of both devices were placed over the same tibialis anterior muscle in adjacent regions, allowing concurrent recording of electromyographic activity throughout the walking task. This configuration enabled simultaneous recording during the same walking task and over the same time interval. However, because the electrodes could not occupy identical anatomical locations, the measurements should not be considered spatially equivalent. Differences in local muscle sampling may therefore contribute to discrepancies between systems, particularly for amplitude-dependent metrics.

During the overground walking protocol, 11 consecutive gait cycles were recorded. To minimize transient effects associated with gait initiation and termination, only gait cycles 2 through 10 were included in the analysis. Consequently, each experimental session contributed nine valid gait cycles for signal processing. After segmenting the signal into the selected gait cycles, the amplitude- and frequency-based metrics, including RMS, MDF, and MNF, were recalculated for each complete muscle contraction and subsequently used for comparison between the Aquamove and Miotec systems.

Before comparing the two systems, the signals were temporally synchronized using cross-correlation applied to the window-by-window RMS sequences. The estimated time delay was then used to align the muscle contractions identified by the Aquamove system with those recorded by the Miotec system.

Following synchronization, muscle contractions were matched according to their order of occurrence, allowing a maximum difference of 1 s between their temporal centers.

Contractions containing invalid values were excluded from the analysis. RMS values were normalized individually within each dataset relative to the maximum observed value, whereas MDF and MNF remained expressed in hertz (Hz).

Comparisons between the two systems were performed individually for each dataset and subsequently using the concatenated set of all valid contractions. For each metric, the intraclass correlation coefficient for absolute agreement (ICC(A,1)) was calculated according to the recommendations of Koo and Li [19], with 95% confidence intervals estimated using bootstrap resampling (5,000 iterations). In addition, the root mean square error (RMSE), relative RMSE, mean bias (Bias), and Bland–Altman limits of agreement (LoA) [20], [21] were computed.

Figure 1 provides an overview of the methodology adopted in this study. Initially, the signals acquired by the Aquamove and Miotec systems underwent the same preprocessing steps. Subsequently, the adapted Berouti spectral subtraction technique was applied exclusively to the Aquamove signals. Next, the RMS, MDF, and MNF metrics were extracted, muscle activations were detected and segmented, the valid gait cycles from the walking protocol were selected, and temporal alignment and contraction matching were performed between the two systems. Finally, statistical agreement metrics were computed to evaluate the performance of the proposed method. Table I summarizes the main parameters used for signal processing and muscle activation detection.

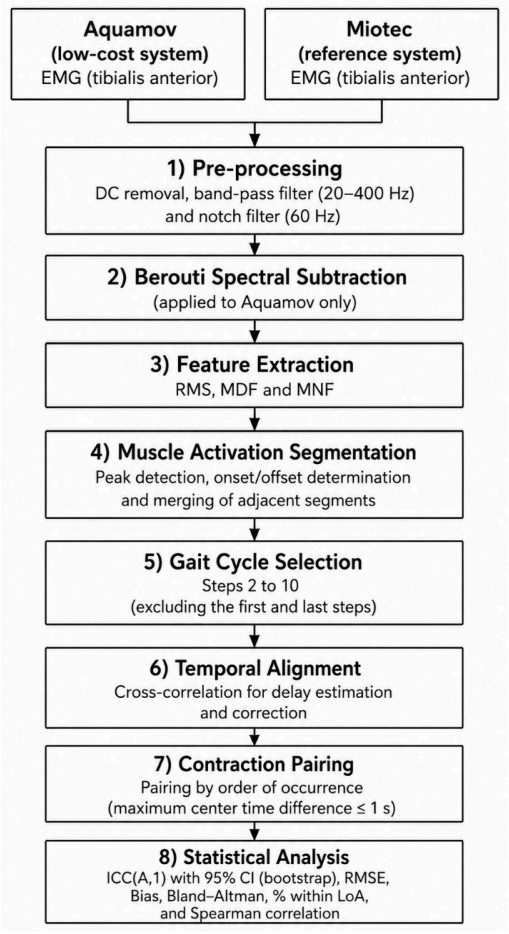


Fig. 1. Signal processing and statistical analysis pipeline adopted for the comparison between the Aquamove and Miotec systems.

TABLE I. SIGNAL PROCESSING AND ACTIVATION DETECTION PARAMETERS

Parameter	**Value**
Analysis sampling frequency	1000 Hz
Band-pass filter	20–400 Hz
Notch filter	60 Hz
Temporal analysis window	150 samples (150 ms)
Temporal window overlap	75 samples (75 ms)
STFT parameters	256-sample Hann window; 128-sample overlap; 256-point FFT
Noise segment duration	8 s
Berouti parameters	α= 1.3; β= 0.08
Welch parameters	256-sample Hamming window; 128-sample overlap; 1024-point FFT
Peak detection parameters	25% threshold; 0.25 s minimum distance
Activation segmentation	50% boundary threshold; 50 ms merging tolerance
Gait cycles analyzed	Cycles 2–10

III. RESULTS AND DISCUSSION

A. *Effect of Spectral Subtraction*

As shown in Figure 2, the application of spectral subtraction reduced the background noise present in the Aquamove signal while preserving the main bursts of muscle activation. For comparison, Figure 3 presents the signal acquired by the Miotec system, showing a qualitative similarity between the signals processed by the two systems. A quantitative comparison is provided in the following subsection using statistical metrics.

Figure 4 illustrates the RMS envelope and the automatically detected muscle activations. The detected activations are numbered from 1 to 11, while only gait cycles 2 through 10 were retained for feature extraction and statistical analysis.

B. *Quantitative Evaluation of the Proposed Method*

Following temporal synchronization of the muscle contractions by cross-correlation and their subsequent matching, agreement metrics were calculated to compare the Aquamove system with the Miotec reference device.

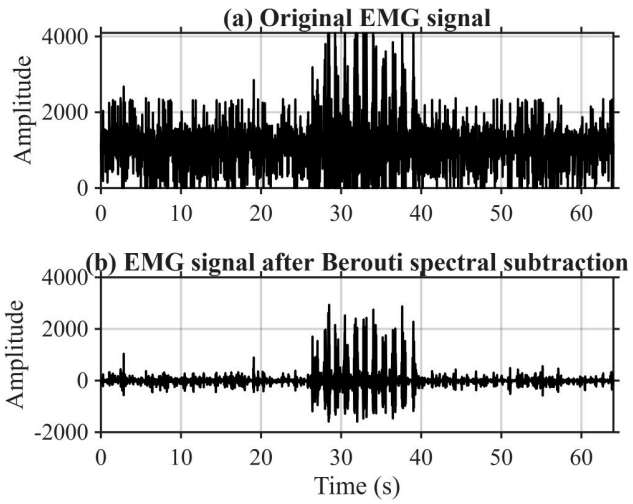


Fig. 2. EMG signal acquired by the Aquamove system during the walking protocol: (a) original signal; (b) after preprocessing and the proposed adaptation of the Berouti spectral subtraction method.

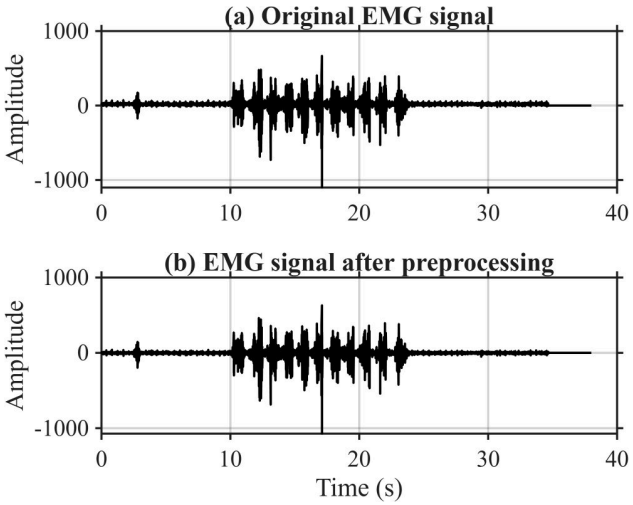


Fig. 3. EMG signal acquired by the Miotec system during the walking protocol: (a) original signal; (b) signal after preprocessing.

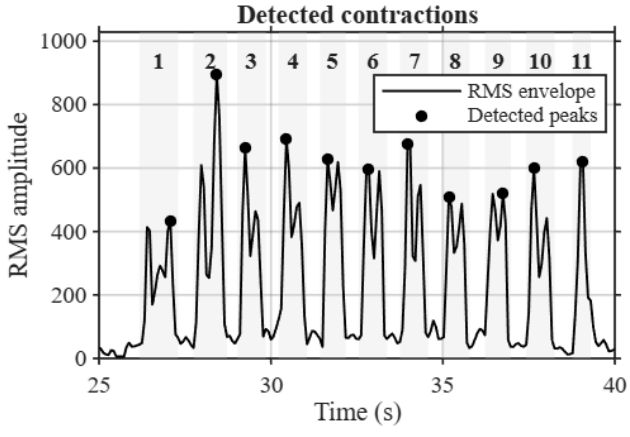


Fig. 4. Detection of muscle activation peaks and delineation of the activation windows.

For RMS, spectral subtraction did not produce a consistent improvement in agreement. The RMSE increased slightly from 8.43% to 9.03%, whereas the mean bias decreased from 1.70% to 1.45%, and the intraclass correlation coefficient (ICC) remained essentially unchanged (0.485 to 0.491). These results indicate that the proposed method had only a limited effect on amplitude-based measurements, in contrast to the clearer reductions in error and bias observed for MDF and MNF.

These results suggest that the noise reduction achieved by spectral subtraction primarily improves the spectral representation of the signal, thereby enhancing the estimation of characteristic EMG frequencies. Nevertheless, the ICC values should be interpreted with caution, as the analysis was based on only 18 paired contractions obtained from one participant across two experimental sessions. Furthermore, the experimental protocol consisted of a single walking condition, resulting in low variability among the spectral characteristics of the contractions. Because the ICC depends on the ratio between variability across samples and measurement error, this limited variability may contribute to low ICC values even when absolute error metrics, such as RMSE and bias, are reduced for the spectral metrics.

The RMS results yielded an ICC(A,1) of 0.491, with a wide 95% confidence interval ranging from 0.002 to 0.782, indicating substantial uncertainty in the agreement estimate. Although this value is lower than those typically reported in the literature for systems with identically positioned electrodes, it should be noted that, in the present study, the Aquamove and Miotec electrodes were placed on adjacent regions of the tibialis anterior muscle to enable simultaneous signal acquisition. Small differences in electrode placement can substantially affect the amplitude of the electromyographic signal and, consequently, influence metrics that depend on muscle activation intensity. Therefore, part of the disagreement observed between systems may reflect spatial sampling differences rather than acquisition-system performance alone.

Figure 5 presents the Bland–Altman plot for the MDF metric. Nearly all observations fell within the limits of agreement (LoA). However, the clinical and technical acceptability of these limits must be interpreted considering their width and the intended application. Furthermore, no evident trend of increasing error with measurement magnitude was observed, suggesting approximately uniform agreement across the analyzed range. Similar behavior was observed for the RMS and MNF metrics, whose quantitative results are summarized in Tables II–IV.

Importantly, the increases in ICC observed for MDF and MNF did not result in values indicative of strong agreement, as both coefficients remained close to 0.10 after spectral subtraction. Moreover, the corresponding 95% confidence intervals were wide and included zero (−0.106 to 0.382 for MDF and −0.077 to 0.312 for MNF), further indicating substantial uncertainty and limited agreement. Therefore, the practical benefit observed in this preliminary dataset should be interpreted primarily in terms of reduced absolute error and systematic bias rather than improved agreement. Specifically, MDF RMSE and bias decreased by 18.0% and 22.0%, respectively, whereas MNF RMSE and bias decreased by 29.4% and 38.1%. Despite the remaining positive bias, these reductions suggest that spectral subtraction may reduce systematic spectral discrepancies between the Aquamove and Miotec systems, even though agreement, as assessed by ICC, remains limited.

Overall, the proposed adaptation of the spectral subtraction method was associated with reductions in error and bias for the spectral metrics obtained with the Aquamove system compared with conventional preprocessing alone. Although studies involving a larger number of participants are still required to confirm and expand these preliminary findings, the proposed technique shows considerable potential for improving the quality of signals acquired by low-cost electromyographic systems, thereby contributing to the development of more accessible solutions without requiring modifications to the acquisition hardware.

Tables II and III summarize the agreement metrics obtained from the comparison between the Aquamove and Miotec systems. Table II reports the results for Aquamove signals processed using only the conventional preprocessing pipeline, whereas Table III presents the results after applying the proposed adaptation of the Berouti spectral subtraction method. ICC(A,1) denotes the intraclass correlation coefficient for absolute agreement and is reported with its bootstrap 95% confidence interval (95% CI); RMSE denotes the root mean square error; Relative RMSE (%) denotes the relative root mean square error; and Bias represents the mean difference between measurements obtained from the two systems and is reported together with the 95% Bland–Altman limits of agreement (LoA). RMS values are expressed as a percentage of the maximum activation, whereas MDF and MNF are reported in hertz (Hz).

TABLE II. Aquamove vs. Miotec Comparison (without subtraction)

Metric	**ICC(A,1) (95% CI)**	**RMSE**	**Relative RMSE (%)**	**Bias (95% LoA)**
RMS (%)	0.485 (0.019 to 0.757)	8.43	9.90	1.70 (−14.95 to 18.36)
MDF (Hz)	0.032 (−0.093 to 0.229)	20.90	18.09	18.09 (−3.03 to 39.20)
MNF (Hz)	0.040 (−0.041 to 0.151)	18.22	13.63	16.34 (0.10 to 32.58)

TABLE III. Aquamove vs. Miotec Comparison (with subtraction)

Metric	**ICC(A,1) (95% CI)**	**RMSE**	**Relative RMSE (%)**	**Bias (95% LoA)**
RMS (%)	0.491 (0.002 to 0.782)	9.03	10.59	1.45 (−16.53 to 19.42)
MDF (Hz)	0.103 (−0.106 to 0.382)	17.14	14.83	14.12 (−5.48 to 33.71)
MNF (Hz)	0.102 (−0.077 to 0.312)	12.87	9.63	10.11 (−5.96 to 26.18)

TABLE IV. Performance comparison before and after Berouti spectral subtraction

Metric	**Conventional**	**Berouti**	**Change**
RMS ICC	0.485	**0.491**	+0.006
MDF ICC	0.032	**0.103**	+0.071
MNF ICC	0.040	**0.102**	+0.062
RMS RMSE (%)	8.43	9.03	↑7.1%
MDF RMSE (Hz)	20.90	**17.14**	↓18.0%
MNF RMSE (Hz)	18.22	**12.87**	↓29.4%
RMS Bias (%)	1.70	**1.45**	↓14.7%
MDF Bias (Hz)	18.09	**14.12**	↓22.0%
MNF Bias (Hz)	16.34	**10.11**	↓38.1%

For ICC, change is reported as the absolute difference; for RMSE and bias, change is reported as the percentage variation relative to conventional preprocessing.

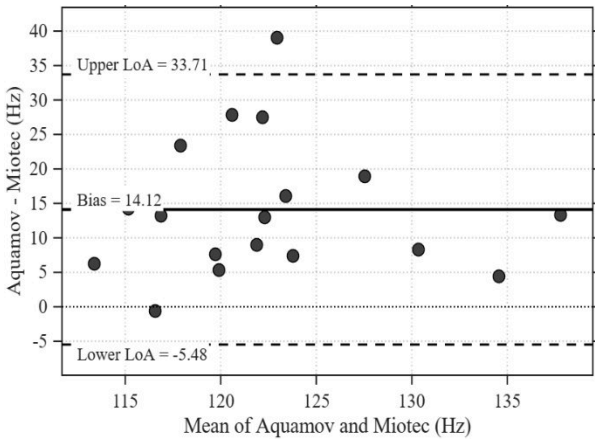


Fig. 5. Bland–Altman plot comparing median frequency (MDF) measurements obtained with the Aquamove and Miotec systems. The solid line indicates the mean bias, and the dashed lines indicate the 95% limits of agreement (LoA).

IV. Conclusion and Future Work

The results suggest that the proposed adaptation of the Berouti spectral subtraction technique was associated with reduced error and bias in spectral metrics obtained from electromyographic signals acquired by the Aquamove system during overground walking. However, these effects were metric-dependent, since RMS error did not decrease after spectral subtraction and the ICC values for MDF and MNF remained low. Therefore, the findings should be interpreted as preliminary evidence that the method may reduce spectral discrepancies between a low-cost EMG system and a commercial reference device, rather than as definitive validation of the system or the algorithm.

In addition to the observed reductions in spectral error and bias, the proposed spectral subtraction method remains attractive from an engineering perspective because it relies on relatively simple mathematical operations performed in the frequency domain. This characteristic may favor future implementation in embedded systems and edge computing devices, enabling local signal processing while reducing the need for continuous transmission of large volumes of data.

In this context, combining the proposed technique with the Aquamove system highlights its potential for the future development of portable, miniaturized, and low-cost electromyographic monitoring solutions. Such integration may expand applicability to Internet of Things (IoT) platforms, rehabilitation, remote monitoring, and wearable systems. Furthermore, the system's portability, ease of use, and compact dimensions may support future investigations in aquatic environments, contributing to research on gait analysis and electromyographic monitoring during water-based exercise.

Future work will focus on further evaluating the proposed method using a larger cohort of participants, standardized electrode configurations, additional walking protocols, and systematic sensitivity analyses of the spectral subtraction parameters. Data-driven optimization strategies, including machine-learning approaches, may also be investigated to adapt α and β according to signal characteristics, aiming to improve noise attenuation while preserving physiologically relevant EMG information. In addition, the algorithm will be implemented directly on the Aquamove hardware to investigate real-time processing on embedded devices and integration with IoT and edge computing architectures.

Acknowledgments

The authors acknowledge the collaboration between the Departments of Electronics and Systems Engineering and Physical Therapy, whose interdisciplinary partnership was fundamental to this study. The integration of expertise in biomedical signal acquisition and processing, human movement analysis, and rehabilitation was essential for the design, execution, and interpretation of the experimental protocol. The authors also gratefully acknowledge the institutional and technical support provided throughout the course of this research.

References

[1] C. Disselhorst-Klug and S. Williams, "Surface electromyography meets biomechanics: Correct interpretation of sEMG-signals in neuro-rehabilitation needs biomechanical input," Front. Neurol., vol. 11, Art. no. 603550, 2020, doi: 10.3389/fneur.2020.603550.

[2] I. Campanini, C. Disselhorst-Klug, W. Z. Rymer, and R. Merletti, "Surface EMG in clinical assessment and neurorehabilitation: Barriers limiting its use," Front. Neurol., vol. 11, Art. no. 934, 2020, doi: 10.3389/fneur.2020.00934.

[3] C. Rodrigues, M. Correia, J. Abrantes, M. A. B. Rodrigues, and J. Nadal, "Association of sEMG neuromuscular control with lower limb joint coordination at different stretch-shortening cycle on standard maximum vertical jump," in 29th Brazilian Congress on Biomedical Engineering: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing, IFMBE Proc., vol. 125, pp. 415–425, 2025, doi: 10.1007/978-3-031-93646-3_43.

[4] S. Fuentes del Toro, Y. Wei, E. Olmeda, L. Ren, W. Guowu, and V. Díaz, "Validation of a low-cost electromyography (EMG) system via a commercial and accurate EMG device: Pilot study," Sensors, vol. 19, no. 23, Art. no. 5214, 2019, doi: 10.3390/s19235214.

[5] A. Bawa and K. Banitsas, "Design validation of a low-cost EMG sensor compared to a commercial-based system for measuring muscle activity and fatigue," Sensors, vol. 22, no. 15, Art. no. 5799, 2022, doi: 10.3390/s22155799.

[6] P. Tecchio, A. Monte, and P. Zamparo, "Low-cost electromyography: Validity against a commercial system depends on exercise type and intensity," Eur. J. Transl. Myol., vol. 31, no. 2, Art. no. 9735, 2021, doi: 10.4081/ejtm.2021.9735.

[7] S. D. Mamdiwar, A. R., Z. Shakruwala, U. Chadha, K. Srinivasan, and C.-Y. Chang, "Recent advances on IoT-assisted wearable sensor systems for healthcare monitoring," Biosensors, vol. 11, no. 10, Art. no. 372, 2021, doi: 10.3390/bios11100372.

[8] C. V. F. Pereira, E. M. de Oliveira, and A. D. de Souza, "Machine learning applied to edge computing and wearable devices for healthcare: Systematic mapping of the literature," Sensors, vol. 24, no. 19, Art. no. 6322, 2024, doi: 10.3390/s24196322.

[9] R. Merletti and G. L. Cerone, "Tutorial. Surface EMG detection, conditioning and pre-processing: Best practices," J. Electromyogr. Kinesiol., vol. 54, Art. no. 102440, 2020, doi: 10.1016/j.jelekin.2020.102440.

[10] R. H. Chowdhury, M. B. I. Reaz, M. A. B. M. Ali, A. A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," Sensors, vol. 13, no. 9, pp. 12431–12466, 2013, doi: 10.3390/s130912431.

[11] M. Boyer, L. J. Bouyer, J.-S. Roy, and A. Campeau-Lecours, "Reducing noise, artifacts and interference in single-channel EMG signals: A review," Sensors, vol. 23, no. 6, Art. no. 2927, 2023, doi: 10.3390/s23062927.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1979, pp. 208–211, doi: 10.1109/ICASSP.1979.1170788.

[13] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for SEMG sensors and sensor placement procedures," J. Electromyogr. Kinesiol., vol. 10, no. 5, pp. 361–374, 2000, doi: 10.1016/S1050-6411(00)00027-4.

[14] Surface ElectroMyoGraphy for the Non-Invasive Assessment of Muscles, "Tibialis anterior," SENIAM. [Online]. Available: https://seniam.org/

[15] C. J. De Luca, "The use of surface electromyography in biomechanics," J. Appl. Biomech., vol. 13, no. 2, pp. 135–163, 1997, doi: 10.1123/jab.13.2.135.

[16] R. Merletti and P. A. Parker, Eds., Electromyography: Physiology, Engineering, and Non-Invasive Applications. Hoboken, NJ, USA: Wiley-IEEE Press, 2004, doi: 10.1002/0471678384.

[17] V. Agostini, M. Ghislieri, S. Rosati, G. Balestra, and M. Knaflitz, "Surface electromyography applied to gait analysis: How to improve its impact in clinics?," Front. Neurol., vol. 11, Art. no. 994, 2020, doi: 10.3389/fneur.2020.00994.

[18] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," IEEE Trans. Audio Electroacoust., vol. AU-15, no. 2, pp. 70–73, 1967, doi: 10.1109/TAU.1967.1161901.

[19] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," J. Chiropr. Med., vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

[20] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet, vol. 327, no. 8476, pp. 307–310, 1986, doi: 10.1016/S0140-6736(86)90837-8.

[21] D. Giavarina, "Understanding Bland–Altman analysis," Biochem. Med., vol. 25, no. 2, pp. 141–151, 2015, doi: 10.11613/BM.2015.015.

Biometric and Demographic Pattern Mining in Electrocardiographic Signals: A Study with Machine Learning Algorithms

1st Cristian Veggian Matias
Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)
Curitiba, Brasil
0009-0008-1281-3099

2nd João Pedro Moreto Lourenção
Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)
Curitiba, Brasil
email

3rd José Jair Alves Mendes Junior
Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)
Curitiba, Brasil
email

4th Heitor Silvério Lopes
Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)
Curitiba, Brasil
email

5th Thiago Henrique Silva
Postgraduate Program in Electrical Engineering and Industrial Informatics (CPGEI), Federal University of Technology – Paraná (UTFPR)
Curitiba, Brasil
email

Abstract*—Electrocardiographic (ECG) signals have emerged as a promising biometric modality due to their physiological uniqueness and resistance to spoofing. This study investigates the feasibility of using ECG-derived features to predict biometric and demographic attributes through supervised and unsupervised machine learning techniques. ECG signals were collected from 26 participants under resting and post-activity conditions using an AD8232-based acquisition system. After preprocessing and heartbeat segmentation, temporal, spectral, statistical, and morphological features were extracted from each cardiac cycle. Supervised models, including k-Nearest Neighbors, Naive Bayes, Random Forest, and Multilayer Perceptron (MLP), were evaluated for classification and regression tasks involving sex, age, weight, height, blood type, and subject identification. Participant-independent validation was adopted for demographic prediction tasks, whereas conventional stratified cross-validation was maintained for closed-set subject identification. In parallel, unsupervised approaches such as k-means, DBSCAN, and hierarchical clustering were applied to investigate the existence of natural groupings in the data. The results showed strong performance for closed-set subject identification, with Random Forest achieving an F1-score of 0.941. However, performance decreased substantially for demographic attributes under participant-independent validation, with the best F1-scores reaching 0.755 for sex and 0.734 for age-group classification. Regression models also showed limited generalization, producing negative coefficients of determination for age, height, and weight prediction. Unsupervised methods did not reveal consistent natural groupings associated with the investigated biometric and demographic characteristics. These findings indicate that ECG-derived features contain strong subject-specific information, while their ability to generalize demographic characteristics to unseen individuals remains limited under the investigated conditions. Larger and more balanced datasets are necessary to improve robustness and evaluate generalization across broader populations.

***Index Terms*—Biometrics, Biomedical signal processing, Electrocardiography, Feature extraction, Machine learning.**

This work was supported by UTFPR.

I. Introduction

The use of physiological characteristics for identifying individuals, known as biometrics, has gained increasing relevance in various areas such as security, identity authentication, and medical applications [1], [2]. Among existing biometric approaches, biopotentials—such as electromyographic (EMG), electrocardiographic (ECG), and electroencephalographic (EEG) signals—stand out for reflecting intrinsic characteristics of the human body, potentially unique to each individual [1], [2].

An ECG is the recording of the electrical activity generated by the depolarization of the cardiac muscle during its rhythmic and organized contractions [3]. This electrical activity results directly from the excitation of cardiac cells, whose pattern is influenced by the anatomy of the heart

and individual physiological conditions. Changes in these characteristics can be captured as morphological variations in the ECG waveform [4]–[6]. Thus, differences in cardiac structure and body composition among individuals produce subtly distinct electrocardiographic patterns, which underpins the use of ECG not only for medical diagnosis but also as a possible unique biometric signature for each person [4]–[6].

In parallel, advances in machine learning techniques have provided new tools for analysing and extracting patterns from large volumes of data [4]–[6]. Applications that combine physiological signals with intelligent algorithms have been developed with the aim of exploring complex biometric characteristics that are still not fully understood or leveraged in practical solutions [4]–[6].

In this context, investigating whether demographic attributes can be inferred from ECG-derived features is relevant not only for biometric identification but also for understanding potential sources of demographic bias. If ECG biometric models implicitly exploit information related to sex, age, or body characteristics, their performance may depend on the demographic composition of the training data. Therefore, identifying the demographic information encoded in ECG signals represents an initial step toward the development of more representative datasets and subgroup-aware evaluation protocols. This study does not aim to solve demographic generalization directly, but rather to investigate one of the factors that may influence it.

This work has as its main objective to investigate the potential of surface electrocardiography (ECG) signals in the identification of biometric and demographic patterns using supervised and unsupervised machine learning techniques.

To achieve this goal, the following specific objectives were defined:

- Develop a protocol for acquiring ECG signals under different physiological conditions (rest and light activity), ensuring quality and variability in the recordings;
- Perform the extraction of relevant features from ECG signals, covering temporal, spectral, and morphological information, in order to compose a robust set of attributes for analysis;
- Analyse the ability of unsupervised learning techniques (k-means, DBSCAN, and hierarchical clustering) to identify natural groupings related to biometric and demographic attributes;
- Apply supervised learning algorithms (kNN, Naive Bayes, Random Forest, and MLP) to predict biometric and demographic attributes from ECG signals; and
- Compare the performance between supervised and unsupervised techniques in pattern identification, using appropriate metrics such as F1-Score, R^2, AUC, and silhouette coefficient.

II. Background

Scientific literature indicates that physiological and biomechanical biosignals are highly suitable for biometric identification and authentication [7] [8]. In general, these signals offer three intrinsic advantages over traditional biometric traits: liveness detection, because they reflect active physiological function; resistance to spoofing, because they are dynamic and largely involuntary; and hidden nature, because they cannot be remotely observed as easily as facial images or fingerprints [9]. These properties make biosignals particularly attractive for secure authentication systems.

Among the main modalities, electrocardiography (ECG) is one of the most established. ECG achieves an average recognition rate of 94.3% across five public datasets and is especially valuable because, in addition to being distinctive across individuals, it provides proof of liveness that traditional biometrics cannot guarantee [10] [11]. Electroencephalography (EEG) also shows strong biometric potential, with reported identification accuracy ranging from 80% to 99%, while offering high resistance to replication [12]. Surface electromyography (sEMG) presents competitive performance, reaching an Equal Error Rate (EER) of 3.5% in verification with only 0.8 seconds of gesture data and 90.3% identification accuracy for static wrist and hand gestures. Unlike ECG and EEG, sEMG also enables customizable gestures that can function as "biometric passwords", adding a knowledge-based layer to the authentication process [13]. Photoplethysmography (PPG) likewise shows promising results, with 96% cross-validation accuracy under physical activity, and can reach 99.8% accuracy with an error rate of 0.16% when fused with ECG in a multimodal framework [10].

Multimodal biometric systems further strengthen the case for biosignal-based authentication. The fusion of multiple biosignals, such as ECG, PPG, and respiratory signals, has achieved experimentally validated high accuracy across diverse populations. In addition to improving recognition performance, multimodal systems are harder to spoof and more robust than unimodal approaches, even when sample sizes are limited [9]. However, the benefits of fusion are not always consistent: in some cases, combining ECG and PPG yields only marginal gains over ECG alone, which suggests that the effectiveness of feature-level and decision-level fusion strategies remains insufficiently understood [10] [14].

Despite these advances, several technical barriers still limit practical deployment. Biosignals typically have low amplitude and are highly susceptible to environmental noise and motion artifacts. Conventional electrodes, such as Ag/AgCl, degrade over time and often require skin preparation and conductive gel, which reduces usability in real-world settings. In non-shielded environments, motion artifacts may be up to ten times stronger than the target signal. In addition, signal contamination across modalities remains a major challenge: ECG can interfere with EMG because of overlapping spectral components, and EMG artifacts are particularly difficult to remove from EEG using conventional filtering methods. As a result, signal quality may vary substantially across sessions, locations, and devices, compromising biometric reliability [15] [16] [17].

Methodological fragmentation is another important limitation. Although ECG biometrics have been widely studied, there is still no consensus on standard methodolo-

gies [7] [18] [18]. More broadly, biosignal research still lacks standardized practices for data curation, pretraining objectives, model architectures, and evaluation protocols, unlike fields such as natural language processing and computer vision [19].

These methodological issues are compounded by demographic imbalance and limited dataset documentation. In the PhysioNet repository, fewer than 7% of datasets report the four key demographic variables of race, ethnicity, sex/gender, and age. Among datasets that report sex, most are predominantly male, and most studies have been conducted in North America, especially in the United States. This raises serious concerns about bias in machine learning models trained on such data. Evidence from biometric systems such as facial recognition shows higher false rejection rates for women, children, and racial or ethnic minority groups, as well as higher false-positive rates for these populations. Therefore, the high performance often reported for biosignal biometrics, typically between 80% and 99%, may not generalize consistently across demographic groups [20] [21].

Security and temporal stability also remain open concerns. PPG acquired through non-invasive sensors is vulnerable to spoofing, and many legacy systems still lack integrated safeguards against signals from unverified sources. Signal quality is also affected by factors such as sensor characteristics, emotional state, energy consumption, and the interaction between light and melanin, which may vary across ethnic groups. Although countermeasures such as waveform analysis and liveness detection based on the Hilbert transform have been proposed, standardized defences are still lacking [15]. In addition, some biosignals and biomechanical patterns, such as gait, vary over time, emotional state, and pathological condition. Even within the same individual, physiological biosignals may change from day to day, which complicates robust biometric modelling [15] [8].

Finally, several research directions remain insufficiently explored. Modelling temporal dependencies in PPG is still an unresolved challenge, and although more complex architectures such as bidirectional LSTM networks can better capture temporal structure, they are also more prone to overfitting, sometimes reaching 100% training accuracy with only marginal gains in practical generalization [15]. The integration of physiological biosignals with behavioural signals, such as gait and gestures, is also still at an early stage [14]. At the data level, collection protocols vary widely in terms of activity, duration, environment, and participant instructions, and there are still no clear guidelines for gathering data across broad age ranges, such as 18 to 65 years, or across different physiological conditions [19]. This lack of standardization is aggravated both by data scarcity in some scenarios, which favours overfitting in deep learning models, and by the absence of consistent standards for data organization and storage, even when large datasets are available. As a result, inconsistencies may arise even within the same laboratory, generating rework and reducing reproducibility [21].

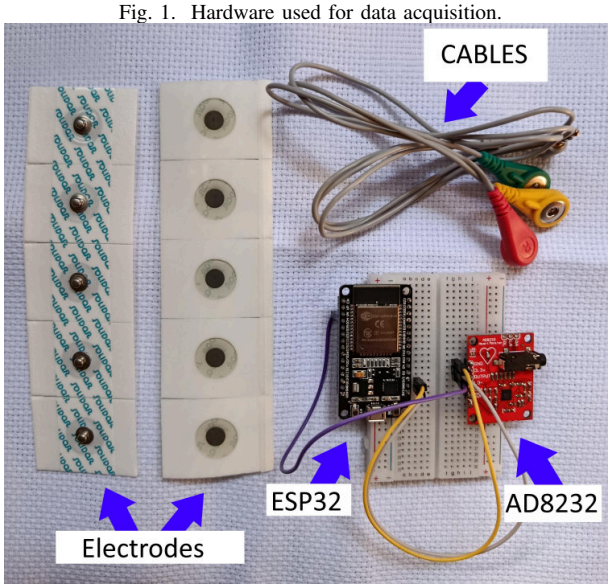
III. Methodology

This section describes the complete experimental workflow adopted in this study, including ECG acquisition, signal preprocessing, feature extraction, and the machine learning methodologies employed for both supervised and unsupervised analyses. The proposed pipeline was designed to evaluate whether ECG-derived features contain sufficient information to identify biometric and demographic characteristics while also investigating the existence of natural latent structures within the data.

A. Data Acquisition Hardware

The system used for surface electrocardiography signal acquisition consists of an AD8232-based biopotential acquisition module, an ESP32 microcontroller, and disposable silver/silver chloride (Ag/AgCl) electrodes with conductive gel, as shown in Figure 1. The AD8232 is an integrated biopotential amplifier specifically designed for high-quality, low-noise electrocardiographic signal acquisition, and it is widely employed in portable and wearable applications [22].

Fig. 1. Hardware used for data acquisition.



The signal amplified by the AD8232 was transmitted to the ESP32 microcontroller, which was responsible for analog-to-digital conversion and for sending the data via a serial interface for storage and subsequent analysis. The ESP32 was chosen due to its versatility, low power consumption, and wireless communication capabilities, although in this experiment it was primarily used in wired mode.

The electrodes used were of the Ag/AgCl type, which are standard in electrophysiological signal acquisition due to their low contact impedance and good chemical stability. The use of conductive gel aimed to reduce the resistance between the skin and the electrode, thereby improving the quality of the recorded signal.

B. Data Collection Protocol

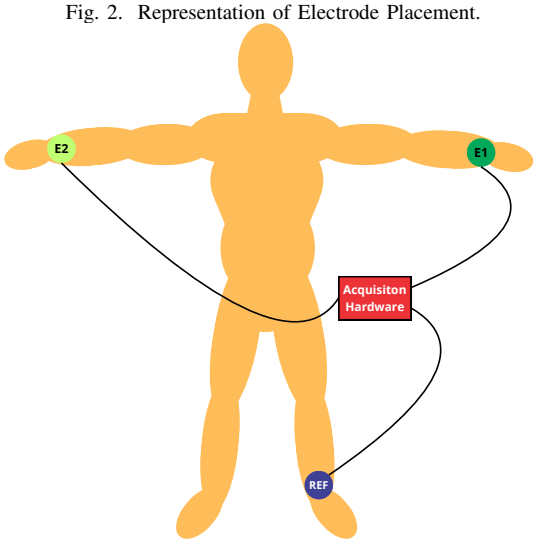
The data collection protocol was designed with the objective of evaluating the possibility of identifying individuals or groups of individuals from surface electrocardiography (ECG) signals. Participant selection followed predefined eligibility criteria intended to reduce major physiological and acquisition-related confounding factors, ensure adequate electrode–skin contact, and protect participants during the post-activity protocol. These criteria were adopted because the present investigation was designed as an initial feasibility study under controlled conditions rather than as a validation for commercial deployment. Individuals aged between 18 and 65 years, of both sexes, and capable of understanding and following the provided instructions were included. These criteria aimed to avoid anatomical interferences in surface signals and to ensure participant cooperation during the procedures.

Individuals with reduced motor mobility, those using medications that affect the nervous or cardiovascular systems, those with wounds, allergies, or dermatitis at the electrode placement sites, those who had consumed alcohol within the previous 24 hours, or those who were pregnant were excluded from the study. These conditions could compromise signal quality or pose ethical and physiological risks. Although these criteria improved experimental control, they reduced the clinical and physiological diversity of the sample and consequently limited the external validity of the results.

Before the start of data collection, all participants were properly instructed regarding the procedures to be performed. The electrode contact areas were cleaned with 70% ethyl alcohol, and participants were instructed to remove metallic objects such as earrings, bracelets, and rings, as well as to avoid carrying electronic devices during recording, such as mobile phones or smartwatches.

The electrodes were positioned on both wrists of the participants, while a reference electrode was placed on the left ankle, as shown in Figure 2.

Fig. 2. Representation of Electrode Placement.



The data collection procedure consisted of two distinct stages for each participant. In the first stage, the signal was recorded under resting conditions, with the individual in a basal state and heart rate within expected average values. In the second stage, data collection was performed immediately after light physical activity, such as jumping jacks or rapidly ascending and descending stairs, in order to increase heart rate. This strategy aimed to obtain signals under different physiological conditions, enabling comparative analyses between resting and active states. The acquired data were properly stored in digital files for subsequent processing and analysis.

C. Feature Extraction

ECG preprocessing and feature extraction were performed using a reproducible computational pipeline. First, baseline wander was reduced using median filtering. A fourth-order Butterworth band-pass filter with cutoff frequencies of 0.5 and 50 Hz was then applied to attenuate low- and high-frequency interference while preserving the frequency range relevant to ECG analysis. The complete source code, processing parameters, software dependencies, and execution instructions are available at https://github.com/CristianVeggian/ecg-analise-biometrica. Initially, the raw signals underwent a preprocessing stage, which consisted of applying median filters to remove baseline wander and low-frequency components associated with the QRS complex and the T wave.

Subsequently, the filtered signals were segmented into individual heartbeats. For this purpose, the `ecg.ecg()` function from the *biosppy* library was used for automatic R-peak detection. The signal was segmented at each heartbeat by extracting a temporal window of 0.2 seconds before and 0.4 seconds after the corresponding R-peak, allowing the isolation of complete cardiac events for analysis.

Feature extraction was carried out in two stages. First, time-domain, spectral, and statistical features were automatically extracted from each window using the *TSFEL* (*Time Series Feature Extraction Library*), covering a diverse set of numerical descriptors. In addition, specific features related to the morphology of the QRS complex were manually computed, such as the amplitude of the R peak, the amplitude of the S valley, the ratio between these amplitudes ($R/|S|$), the duration of the QRS complex (in seconds), the slope between the R and S points, and the area under the QRS complex, obtained through numerical integration. Table I summarizes the main feature categories considered in this study. Due to the large number of extracted features, the complete feature list, including the corresponding computational definitions, is provided in the project repository.

A total of 162 ECG-derived features were considered, including six QRS-based morphological features and 156 statistical, temporal, spectral, time–frequency, and cepstral descriptors. All extracted features were organized into a CSV file, which also included the corresponding subject identifier and the condition under which the signal was recorded, categorized into two distinct states: rest or activity.

TABLE I
SUMMARY OF ECG-DERIVED FEATURE CATEGORIES USED IN THIS STUDY.

Category	Examples of extracted features	Description
Morphological	R amplitude, S amplitude, R/|S| ratio, QRS duration, R–S slope, QRS area	Features specifically derived from the morphology of the QRS complex, describing amplitudes, duration, shape, and area of individual heartbeats.
Statistical	Mean, median, variance, standard deviation, root mean square, skewness, kurtosis, interquartile range, mean absolute deviation, median absolute deviation	Statistical descriptors characterizing the distribution, dispersion, central tendency, and shape of ECG amplitude values within each heartbeat segment.
Temporal	Absolute energy, area under the curve, autocorrelation, zero-crossing rate, peak-to-peak distance, positive and negative turning points, neighbourhood peaks, signal distance	Time-domain descriptors characterizing signal variation, energy, temporal dynamics, transitions, and waveform behavior within each segmented heartbeat.
Spectral	Fundamental frequency, median frequency, maximum frequency, spectral centroid, spectral entropy, spectral spread, spectral slope, spectral skewness, spectral kurtosis, spectral roll-on, spectral roll-off, power bandwidth	Frequency-domain descriptors characterizing the distribution, shape, complexity, and concentration of the ECG signal energy across frequencies.
Time–frequency	Wavelet energy, wavelet entropy, wavelet absolute mean, wavelet standard deviation, wavelet variance, spectrogram mean coefficients	Time–frequency descriptors representing variations in ECG frequency content over time at different frequency scales.
Cepstral	MFCC coefficients, LPCC coefficients	Cepstral-domain descriptors representing compact transformations of the signal spectrum and its spectral envelope.

D. Unsupervised Learning

Clustering analyses were performed using the Orange Data Mining platform. The process began with the import of files containing the features extracted from each heartbeat along with the metadata file, which included demographic characteristics of the individuals (ID, sex, state, weight, height, blood type). After merging the files and performing initial preprocessing—such as discretizing continuous variables into categories and selecting relevant columns — an outlier filtering step was applied to reduce potential interference in the clustering process.

Three main clustering approaches were used: *k-means*, DBSCAN, and hierarchical clustering. The *k-means* algorithm was configured with the number of clusters adjusted according to the objective of the analysis. In one approach, the goal was to form a number of clusters corresponding to the number of individuals in the dataset, in order to evaluate whether the extracted features would be sufficient to differentiate heartbeats from different individuals, functioning as a biometric signature. In other approaches, the clusters were adjusted to reflect specific characteristics of individuals, such as age group (young, adult, elderly), sex (male or female), weight (three categories), and height (three categories), as well as attempts to cluster by blood type.

In addition to *k-means*, the DBSCAN algorithm was employed to identify clusters with non-spherical shapes and to automatically detect outliers. Hierarchical clustering was applied using the distance matrix between samples, allowing visual analysis of group formation through dendrograms. To evaluate clustering quality, visualizations such as scatter plots and t-SNE projections were used, along with the silhouette metric to assess cluster cohesion and separation.

This approach enabled the evaluation of the ability of features extracted from heartbeats to discriminate individuals or subgroups based on biometric and demographic attributes.

E. Supervised Learning

The supervised learning stage was also conducted using the Orange platform, based on datasets containing features extracted from heartbeats. After importing and organizing the CSV files, preprocessing steps were performed, including file merging, variable discretization, and selection of relevant columns.

The main objective was to evaluate the predictive capability of the extracted features with respect to biometric and demographic attributes of individuals, such as age, sex, weight, height, and blood type. For categorical attributes, including subject ID, sex, recording state, and blood type, supervised classification algorithms were applied. In addition, age, weight, and height were discretized into predefined intervals and evaluated as classification tasks. In this context, four main classifiers were tested: k-Nearest Neighbors (kNN), Naive Bayes, Random Forest, and Multilayer Perceptron. Each algorithm was executed with its default settings, allowing an initial comparison between probabilistic approaches, instance-based methods, and ensemble learning techniques.

For numerical variables (age, weight, and height), two distinct strategies were evaluated. In the first approach, these variables were discretized into predefined intervals to transform the problem into a classification task. However, due to class imbalance observed in some variables—especially age—experiments were also conducted using supervised regression to predict continuous values. For regression tasks, the algorithms used included k-Nearest Neighbors, Linear Regression, Random Forest, and Multilayer Perceptron.

Model validation was performed using the Test and Score widget. For demographic classification and regression tasks, participant-independent validation was adopted using cross-validation grouped by subject identifier (ID). Therefore, all heartbeat segments belonging to the same participant were assigned to the same fold, preventing samples from a given individual from appearing simultaneously in the training and test

sets. For closed-set subject identification, conventional stratified 10-fold cross-validation was maintained, since subject-wise splitting would remove from the training set the identity class to be predicted. Classification performance was evaluated using accuracy, precision, recall, F1-score, AUC, and Matthews correlation coefficient (MCC), whereas regression performance was assessed using MSE, RMSE, MAE, MAPE, and R^2.

IV. Results

This section presents the experimental results obtained from both the unsupervised and supervised learning analyses. Initially, the characteristics of the generated dataset are described, followed by the evaluation of clustering algorithms and predictive models. The results are analyzed using quantitative performance metrics and visual inspection in order to assess the feasibility of identifying biometric and demographic information from ECG-derived features.

A. Generated Dataset

A total of 52 recordings were performed across 26 distinct participants. In total, 1,858 electrocardiographic segments were obtained, distributed between resting and activity conditions.

The generated dataset can be found in the following GitHub repository: https://github.com/CristianVeggian/ecg-analise-biometrica.

Figure 3 shows the distribution of the dataset metadata, which were used as target variables in classification and regression tasks. A significant imbalance can be observed in some variables, such as sex and age. The number of segments from male participants is substantially higher compared to female participants, which may impact the performance of supervised classifiers. Similarly, there is a predominance of participants classified in the "young" age group, with lower representation in the "adult" and "elderly" categories.

The distributions of weight and height show a more balanced division among the established categories, which is favorable for model training. In the case of blood type, there is a concentration of records in the "Not Specified" (N/S) category, in addition to a non-uniform distribution among the different types, representing an additional challenge for predictive algorithms.

These observations highlight the need to carefully evaluate the impact of imbalance during the supervised learning stage, especially in multiclass classification tasks. For discretized numerical variables, the imbalance motivated additional experiments using supervised regression in order to explore alternative approaches and mitigate the limitations imposed by the data distribution.

B. Unsupervised Learning

Three main unsupervised learning algorithms were tested: k-means, DBSCAN, and hierarchical clustering. The quantitative evaluation of clustering quality was performed using the silhouette coefficient.

The elbow method was applied to determine the optimal number of clusters in the k-means algorithm, testing values of k between 2 (minimum expected number of clusters) and 26 (maximum expected, based on the number of individuals). Despite this approach, it was not possible to identify a clear inflection point in the curve, reinforcing the absence of well-defined clustering structures in the dataset, as shown in Figure 4.

Furthermore, even with manual adjustments of k, the silhouette coefficient values remained low, indicating that the separation between clusters did not represent cohesive or consistent groupings. These results suggest that, within the context of this study and given the extracted features, ECG signals did not exhibit sufficient natural separations for meaningful clustering through unsupervised learning.

Despite attempts to optimize the DBSCAN algorithm through systematic tuning of the *eps* and *min_samples* parameters, the resulting partition was highly unbalanced. DBSCAN identified two clusters, containing 1,650 and 20 samples, respectively, while only two samples (0.1%) were classified as noise. The silhouette coefficient, calculated excluding noise samples, was 0.520. Although this value indicates moderate separation between the identified clusters, the strong imbalance in cluster sizes indicates that the data were predominantly represented by a single large cluster rather than by two comparably sized natural groups.

Additionally, the t-SNE projection did not reveal a clear correspondence between the DBSCAN partition and the investigated metadata. When the same projection was visualized according to recording state, sex, and subject ID, substantial overlap remained between the corresponding groups. Therefore, although DBSCAN produced a moderately separated partition according to the silhouette coefficient, the identified clusters could not be clearly associated with the investigated biometric or demographic characteristics.

Figure 5 presents the t-SNE projection of the DBSCAN results together with the corresponding metadata. Cluster 0 represents the dominant group containing 1,650 samples, Cluster 1 contains 20 samples, and only two samples were classified as noise. The comparison between the clustering result and the available metadata illustrates the absence of a clear demographic or biometric interpretation for the partition obtained by DBSCAN.

Hierarchical clustering was also applied using Euclidean distance and the *Ward* linkage method. The number of clusters was not defined a priori, since this analysis was exploratory and aimed at investigating the intrinsic organization of the ECG-derived feature space. The *Ward* criterion was selected because it minimizes within-cluster variance and is appropriate for continuous numerical data analyzed with Euclidean distances.

Inspection of the dendrogram revealed two major branches separated at a relatively high linkage distance, motivating the evaluation of a two-cluster partition. The resulting partition contained 1,038 samples in Cluster 1 and 634 samples in Cluster 2. Although the dendrogram revealed a clear high-

Fig. 3. Distribution of participants according to (a) state, (b) sex, (c) age group, (d) weight group, (e) height group, and (f) blood type.

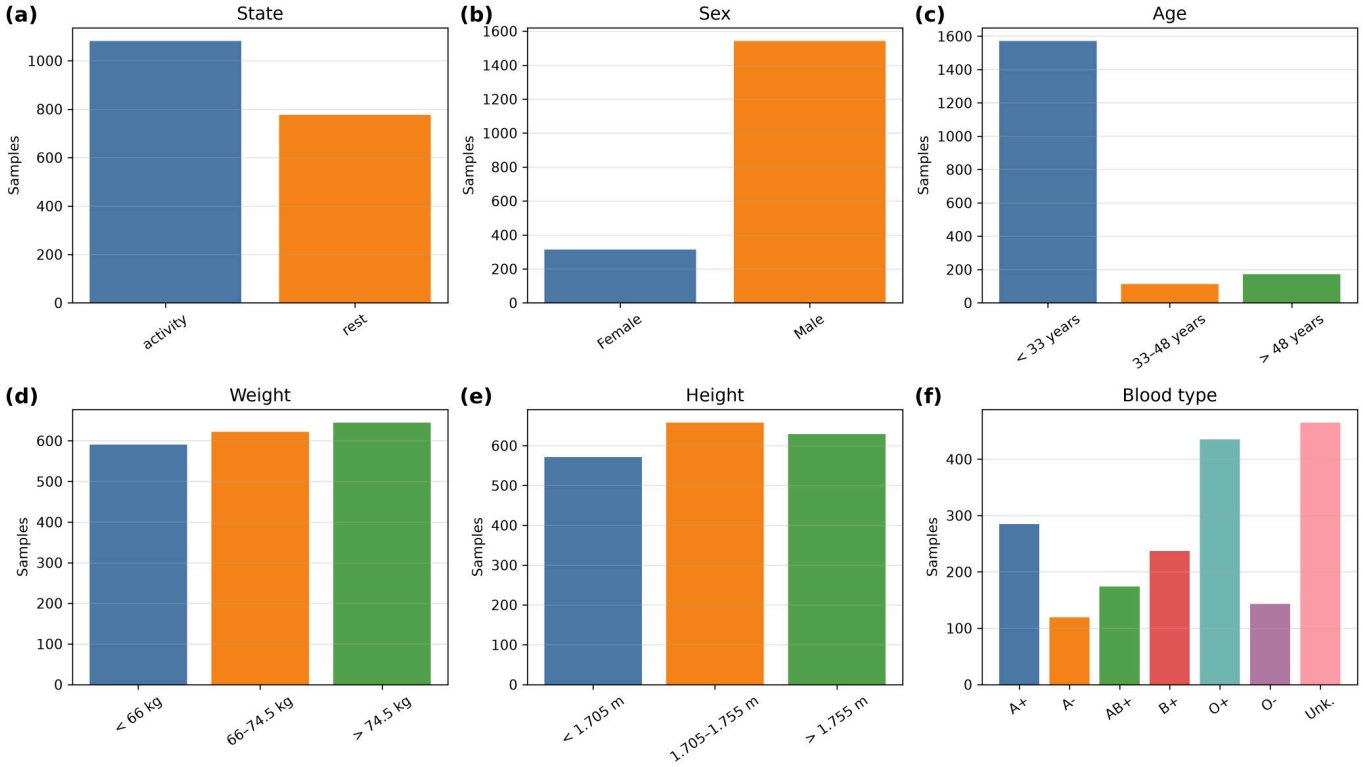


Fig. 4. Elbow method applied to k-means.

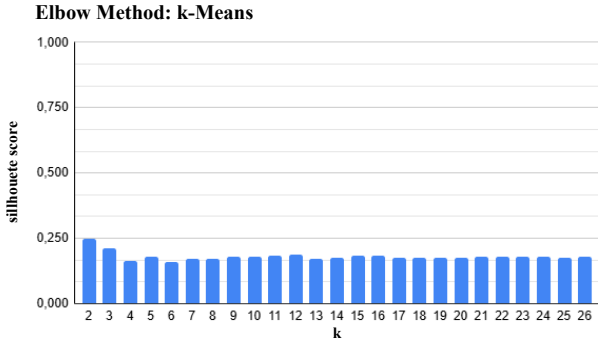
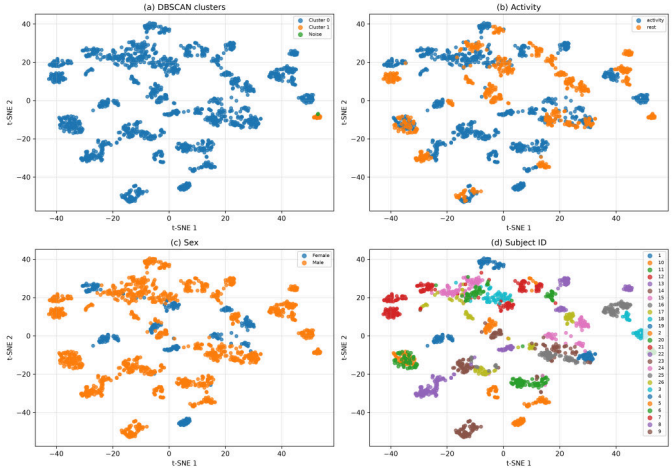


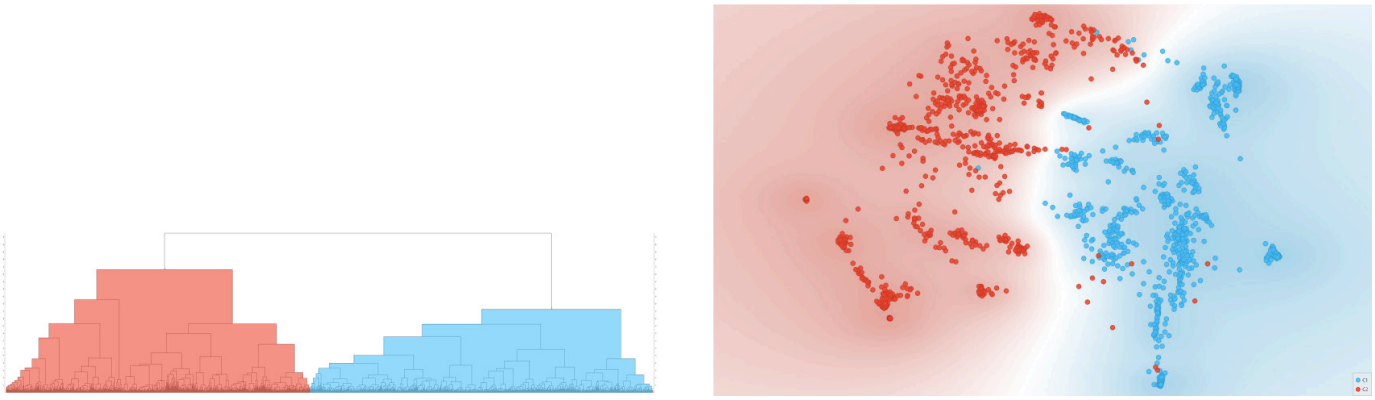
Fig. 5. t-SNE visualization of the DBSCAN clustering results and selected metadata. Cluster 0 and Cluster 1 represent the groups identified by DBSCAN, whereas Noise represents samples that were not assigned to any cluster. The remaining panels show the same t-SNE projection colored according to recording state, sex, and subject ID.



level bifurcation, the silhouette analysis showed substantial variability in cluster cohesion, including negative silhouette values for several samples. This indicates that a considerable number of observations were not clearly separated between the two clusters.

Figure 6 presents the dendrogram together with the corresponding t-SNE projection. The t-SNE visualization showed considerable overlap between the two groups and did not reveal a clear correspondence with the investigated demographic or biometric attributes. Therefore, the two-cluster solution was interpreted as an exploratory representation of the hierarchical structure of the feature space rather than as evidence of two predefined or well-separated demographic or biometric groups.

These findings indicate that, similarly to *k-means* and DBSCAN, hierarchical clustering did not identify natural groups that could be consistently associated with the investigated demographic or biometric attributes. The presence of two major branches in the dendrogram reflects a high-level hierarchical



(a) Hierarchical clustering dendrogram.

(b) t-SNE projection of the two-cluster partition.

Fig. 6. Hierarchical clustering analysis using Ward linkage and Euclidean distance. (a) Dendrogram showing the hierarchical organization of the ECG-derived samples; the two-cluster partition corresponds to the two major branches observed before the final high-level merge. (b) t-SNE projection of the resulting two-cluster partition, showing regions of overlap between the clusters. The partition contained 1,038 samples in Cluster 1 and 634 samples in Cluster 2.

division of the feature space, but the silhouette distribution and the overlapping t-SNE projection indicate that this division should not be interpreted as two clearly separable biological or demographic populations.

C. Supervised Learning

Table II presents the F1-scores obtained for the supervised classification tasks. A clear distinction can be observed between closed-set subject identification and participant-independent demographic classification.

TABLE II
F1-SCORE BY ATTRIBUTE AND ALGORITHM.

Class	kNN	Naive Bayes	Random Forest	MLP
ID (26 classes)	0.343	0.834	**0.941**	0.922
Sex (2 classes)	**0.755**	0.661	0.696	0.730
Age (3 classes)	**0.734**	0.615	0.721	0.713
State (rest/activity)	0.482	**0.617**	0.551	0.568
Height (3 classes)	**0.291**	0.175	0.157	0.145
Weight (3 classes)	0.192	0.203	0.195	**0.280**
Blood type (7)	0.167	0.204	0.173	**0.229**

For closed-set subject identification, Random Forest achieved the highest F1-score (0.941), followed by MLP (0.922) and Naive Bayes (0.834), whereas kNN obtained a considerably lower F1-score of 0.343. These results indicate that the extracted ECG features contain strong subject-specific information when the identities evaluated during testing are represented in the training data.

In contrast, performance decreased when participant-independent validation was applied to demographic attributes. Sex classification achieved its highest F1-score with kNN (0.755), followed by MLP (0.730), Random Forest (0.696), and Naive Bayes (0.661). For age-group classification, kNN also obtained the highest F1-score (0.734), while Random Forest and MLP achieved 0.721 and 0.713, respectively. State classification presented lower performance, with Naive Bayes achieving the highest F1-score of 0.617.

The remaining demographic attributes presented considerably lower predictive performance. For height classification, the highest F1-score was 0.291 with kNN. Weight classification reached a maximum F1-score of 0.280 with MLP, whereas blood-type classification reached a maximum F1-score of 0.229, also with MLP. These results indicate that, under participant-independent validation, the extracted ECG features provided limited discriminative information for these attributes.

Overall, the reduction in demographic classification performance compared with closed-set subject identification suggests that ECG-derived features contain substantial subject-specific information, whereas demographic patterns are considerably more difficult to generalize to participants not represented during model training.

Within the supervised learning analysis, regression experiments were also conducted to evaluate the ability of the models to predict continuous demographic variables, namely height, age, and weight, under participant-independent validation.

Table III presents the performance of the regression models for height prediction. All evaluated models produced negative coefficients of determination, indicating limited generalization to unseen participants. Among the evaluated methods, kNN achieved the highest R^2 value ($R^2 = -0.416$), with an RMSE of 0.082 and an MAE of 0.063. Random Forest obtained $R^2 = -0.619$, while Linear Regression and MLP presented substantially lower values of $R^2 = -2.268$ and $R^2 = -7.854$, respectively. These results indicate that none of the evaluated models reliably explained height variability for participants not represented during training.

TABLE III
PERFORMANCE OF REGRESSION MODELS FOR HEIGHT PREDICTION.

Model	MSE	RMSE	MAE	MAPE (%)	R^2
kNN	0.007	0.082	0.063	3.582	**-0.416**
Linear Regression	0.015	0.124	0.081	4.609	-2.268
Neural Network (MLP)	0.042	0.204	0.147	8.429	-7.854
Random Forest	0.008	0.087	0.067	3.812	-0.619

Table IV presents the regression results for age prediction. Similar to the height prediction task, all models produced

negative R^2 values. Random Forest achieved the highest coefficient of determination ($R^2 = -0.497$), with an RMSE of 13.369 and an MAE of 9.064 years. kNN obtained a similar performance, with $R^2 = -0.594$, whereas MLP and Linear Regression presented R^2 values of -0.848 and -3.520, respectively. Therefore, none of the evaluated models outperformed a baseline based on the mean target value when predicting the age of unseen participants.

TABLE IV
PERFORMANCE OF REGRESSION MODELS FOR AGE PREDICTION.

Model	MSE	RMSE	MAE	MAPE (%)	R^2
kNN	190.320	13.796	9.632	33.391	-0.594
Linear Regression	539.767	23.233	12.574	47.098	-3.520
Neural Network (MLP)	220.725	14.857	10.263	36.180	-0.848
Random Forest	178.719	13.369	9.064	30.520	**-0.497**

Table V presents the results obtained for weight prediction. Random Forest achieved the best performance among the evaluated models, with an RMSE of 15.991 kg, an MAE of 11.680 kg, and $R^2 = -0.369$. However, the negative coefficient of determination indicates that its predictions did not generalize adequately to unseen participants. kNN obtained $R^2 = -0.456$, while MLP and Linear Regression achieved $R^2 = -0.807$ and $R^2 = -1.162$, respectively.

TABLE V
PERFORMANCE OF REGRESSION MODELS FOR WEIGHT PREDICTION.

Model	MSE	RMSE	MAE	MAPE (%)	R^2
kNN	272.097	16.495	12.339	16.500	-0.456
Linear Regression	403.913	20.098	14.904	20.138	-1.162
Neural Network (MLP)	337.636	18.375	14.362	19.132	-0.807
Random Forest	255.714	15.991	11.680	15.337	**-0.369**

Taken together, the regression results demonstrate that the evaluated models were unable to reliably generalize continuous demographic attributes across participants. This contrasts with the strong closed-set subject identification performance and further indicates that subject-specific ECG characteristics are more readily captured by the extracted features than demographic patterns that generalize across individuals.

V. DISCUSSION

The results obtained in this study reveal an important distinction between closed-set biometric identification and demographic prediction from ECG-derived features. While subject identification remained highly effective, participant-independent validation substantially reduced the performance of models applied to demographic attributes. This finding indicates that ECG-derived features contain strong subject-specific information, but their ability to generalize demographic characteristics to previously unseen individuals is considerably more limited.

For closed-set subject identification, Random Forest achieved the highest performance, with an F1-score of 0.941, followed by the Multilayer Perceptron with an F1-score of 0.922. In this task, conventional stratified cross-validation was maintained because all identity classes must be represented in the training set. These results reinforce the hypothesis that morphological, temporal, statistical, and spectral characteristics extracted from ECG signals contain discriminative information capable of differentiating individuals.

In contrast, when participant-independent validation was adopted for demographic classification, performance decreased substantially. The best F1-scores were 0.755 for sex classification and 0.734 for age-group classification, both obtained with kNN. State classification reached a maximum F1-score of 0.617 with Naive Bayes, whereas height, weight, and blood-type classification presented considerably lower results. These findings suggest that part of the high performance observed under heartbeat-level cross-validation was likely associated with subject-specific patterns shared between training and test samples rather than with demographic characteristics that generalize across individuals.

The regression experiments provide further evidence of this limitation. Under participant-independent validation, all evaluated models produced negative R^2 values for age, height, and weight prediction. The best results were R^2 = -0.416 for height using kNN, R^2 = -0.497 for age using Random Forest, and R^2 = -0.369 for weight using Random Forest. Negative coefficients of determination indicate that, under the investigated conditions, the models did not generalize these continuous demographic attributes to unseen participants and performed worse than a baseline prediction based on the target mean.

These results highlight the methodological importance of subject-wise validation when multiple heartbeat segments are extracted from the same participant. Because several segments originate from a single individual, conventional sample-level cross-validation may allow subject-specific physiological information to appear in both training and test sets, resulting in optimistic estimates of model performance. The participant-independent evaluation therefore provides a more conservative and realistic assessment of the demographic information that can be generalized from ECG-derived features.

The unsupervised learning results are consistent with this interpretation. k-means, DBSCAN, and hierarchical clustering did not reveal natural groupings that could be clearly associated with the investigated demographic and biometric characteristics. In particular, DBSCAN identified two highly unbalanced clusters, containing 1,650 and 20 samples, respectively, with only two samples (0.1%) classified as noise. Although the silhouette coefficient of 0.520 indicated moderate separation between the DBSCAN clusters, their substantial size imbalance and the absence of a clear correspondence with the available metadata suggest that this partition does not represent a meaningful demographic or biometric grouping. Similarly, the other clustering approaches showed considerable overlap between groups and no consistent association with the investigated attributes. Nevertheless, the strong closed-set identification results demonstrate that the absence of clear unsupervised groupings does not imply an absence of subject-specific information in the ECG signals.

Several factors may have contributed to the limited de-

mographic generalization observed in this study. The dataset contains only 26 participants and presents demographic imbalance, particularly regarding sex, age, and blood type. In addition, the relatively small number of subjects restricts the diversity represented in participant-independent validation folds. High feature dimensionality and physiological variability between recording conditions may also contribute to the difficulty of learning patterns that generalize across participants.

Therefore, the results should be interpreted as evidence that ECG-derived features are highly informative for distinguishing individuals already represented in the biometric system, while their ability to predict demographic attributes for unseen individuals remains limited under the investigated conditions. Larger and more representative datasets, together with participant-independent evaluation protocols, will be necessary to determine whether demographic information can be reliably generalized from ECG signals.

VI. Conclusion

This work investigated the feasibility of extracting biometric and demographic information from ECG signals using supervised and unsupervised machine learning techniques. A dataset containing ECG recordings from 26 participants under resting and post-activity conditions was collected, and more than 160 temporal, spectral, statistical, and morphological features were extracted from each heartbeat.

The participant-independent evaluation revealed a clear distinction between biometric identification and demographic prediction. ECG-derived features demonstrated strong discriminative capability for closed-set subject identification, with Random Forest achieving an F1-score of 0.941. However, when heartbeat segments from the same participant were prevented from appearing simultaneously in the training and test sets, classification performance decreased substantially for demographic attributes.

Regression results further demonstrated limited generalization to unseen participants. All evaluated models produced negative R^2 values for age, height, and weight prediction, indicating that the extracted features were insufficient to reliably estimate these continuous demographic variables under participant-independent validation.

Similarly, unsupervised approaches, including k-means, DBSCAN, and hierarchical clustering, were unable to identify consistent latent structures associated with the investigated biometric and demographic attributes. Overall, these findings indicate that ECG-derived features contain strong subject-specific information, supporting their potential for closed-set biometric identification, while demographic characteristics were not reliably generalized from the extracted feature space in the present dataset.

The results also emphasize the importance of participant-independent validation in studies involving multiple physiological segments from the same individuals, since sample-level cross-validation may lead to optimistic performance estimates due to subject-specific information shared between training and test sets.

Future work will focus on expanding the dataset with a larger and more demographically diverse cohort, improving the balance among demographic groups, investigating feature selection and dimensionality reduction techniques, and evaluating additional machine learning and deep learning approaches under participant-independent validation protocols.

References

[1] D. Wang, Y. Si, W. Yang, G. Zhang, and J. Li, "A novel electrocardiogram biometric identification method based on temporal-frequency autoencoding," *Electronics*, vol. 8, no. 6, 2019.

[2] A. N. Uwaechia and D. A. Ramli, "A comprehensive survey on ecg signals as new biometric modality for human authentication: Recent advances and future challenges," *IEEE Access*, vol. 9, pp. 97760–97802, 2021.

[3] K. Najarian and R. Splinter, *Biomedical Signal and Image Processing*. Boca Raton, FL: CRC Press, 2 ed., May 2012.

[4] T. S. Lugovaya, "Biometric human identification based on electrocardiogram," master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation, June 2005.

[5] T. W. Shen, W. J. Tompkins, and Y. H. Hu, "One-lead ECG for identity verification," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, IEEE, 2003.

[6] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: a new approach in human identification," *IEEE Trans. Instrum. Meas.*, vol. 50, pp. 808–812, June 2001.

[7] T. M. C. Pereira, R. C. Conceição, V. Sencadas, and R. Sebastião, "Biometric recognition: A systematic review on electrocardiogram data acquisition methods," *Sensors (Basel)*, vol. 23, p. 1507, Jan. 2023.

[8] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," *Comput. Vis. Image Underst.*, vol. 167, pp. 1–27, Feb. 2018.

[9] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Process. Mag.*, vol. 32, pp. 20–30, Sept. 2015.

[10] Y. Che, L. Du, G. Tang, and S. Ling, "A biometric identification for multi-modal biomedical signals in geriatric care," *Sensors*, vol. 24, p. 6558, Oct. 2024.

[11] N. Belgacem, R. Fournier, A. Nait-Ali, and F. Bereksi-Reguig, "A novel biometric authentication approach using ecg and emg signals," *Journal of Medical Engineering amp; Technology*, vol. 39, p. 226–238, Apr. 2015.

[12] S. Zhang, L. Sun, X. Mao, C. Hu, and P. Liu, "Review on EEG-based authentication technology," *Comput. Intell. Neurosci.*, vol. 2021, p. 5229576, Dec. 2021.

[13] J. He and N. Jiang, "Biometric from surface electromyogram (sEMG): Feasibility of user verification and identification based on gesture recognition," *Front. Bioeng. Biotechnol.*, vol. 8, p. 58, Feb. 2020.

[14] A. R. Mishra, R. Kumar, V. Gupta, S. Prabhu, R. Upadhyay, P. C. Chhipa, S. Rakesh, H. Mokayed, D. Das Chakladar, K. De, M. Liwicki, F. Simistira Liwicki, and R. Saini, "SignEEG v1.0: Multimodal dataset with electroencephalography and hand-written signature for biometric systems," *Sci. Data*, vol. 11, p. 718, July 2024.

[15] D. Aguilar, A. Martínez-Cruz, K. A. Ramírez-Gutiérrez, and M. Morales-Sandoval, "Ppg-based biometric authentication: A review on architectures, datasets, attacks and security challenges," *Computational and Structural Biotechnology Journal*, vol. 28, p. 511–528, Jan. 2025.

[16] K.-M. Chang, P.-T. Liu, and T.-S. Wei, "Electromyography parameter variations with electrocardiography noise," *Sensors (Basel)*, vol. 22, p. 5948, Aug. 2022.

[17] H. Yu, M. Ma, B. Zhang, A. Wang, G. Zhong, Z. Zhou, C. Liu, C. Li, J. Fang, Y. He, D. Ren, F. Deng, Q. Hong, Y. Zhao, and X. Guo, "Bionic sensors for biometric acquisition and monitoring: Challenges and opportunities," *Sensors (Basel)*, vol. 25, p. 3981, June 2025.

[18] L. Xu, M. Xu, Y. Ke, X. An, S. Liu, and D. Ming, "Cross-dataset variability problem in EEG decoding with deep learning," *Front. Hum. Neurosci.*, vol. 14, p. 103, Apr. 2020.

[19] X. Gu, Y. Shu, J. Han, Y. Liu, Z. Liu, J. Anibal, V. Sangha, E. Phillips, B. Segal, H. Yuan, F. Liu, K. Branson, P. Schwab, D. Belgrave, L. Clifton, D. Spathis, V. Lampos, A. A. Faisal, and D. A. Clifton, "Foundation models for biosignals: A survey," *Techrxiv*, Aug. 2025.

[20] K. Kotwal and S. Marcel, "Review of demographic fairness in face recognition," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 8, pp. 20–45, Jan. 2026.

[21] S. Jiang, P. Ashar, M. M. H. Shandhi, and J. Dunn, "Demographic reporting in biosignal datasets: a comprehensive analysis of the PhysioNet open access database," *Lancet Digit. Health*, vol. 6, pp. e871–e878, Nov. 2024.

[22] J. J. A. Mendes Junior, D. P. Campos, L. C. d. A. V. D. Biassio, P. C. Passos, P. B. Júnior, A. E. Lazzaretti, and E. Krueger, "Ad8232 to biopotentials sensors: Open source project and benchmark," *Electronics*, vol. 12, no. 4, 2023.

BioStatusIA: A Multi-Agent Architecture for multi-modality Clinical Decision Support

1st Francisco Soares da Silva Júnior
Programa de Pós-Graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brazil
juniorsoares@alu.ufc.br

2nd Kaio Emanuel Lima de Matos
Programa de Pós-Graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brazil
kaio.emanuel@alu.ufc.br

3rd Daniel Santos da Silva
Departamento de Engenharia de Teleinformática (DETI)
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
danielssilva@alu.ufc.br

4th Victor Hugo C. de Albuquerque
Departamento de Engenharia de Teleinformática (DETI)
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
victor.albuquerque@ieee.org

***Abstract*—**
BioStatusIA is a multi-agent architecture for clinical decision support that combines biomedical data processing, classifier selection, and local report generation. CrewAI coordinates the agents, and a local Large Language Model (LLM) runs through Ollama. Eight completed executions covered five tabular datasets and three labeled medical image collections, processed independently. Each run compared six classifiers using an internal cross-validation score and reported performance on a separate test set. The selected models achieved mean AUROC 0.866 ± 0.109 and Matthews Correlation Coefficient (MCC) 0.580 ± 0.295; these values summarize variation across datasets. Four runs failed the minimum cross-validation sensitivity of 0.80, and four selected models had calibration error above 0.10. End-to-end processing took 80.073–809.475 s, LLM generation took 41.593–789.218 s, and selected-classifier inference took 0.0005–0.4427 ms per sample. Times were measured once per dataset. The results document operation on tabular and image data but do not establish patient-independent generalization, report accuracy, or clinical validity.

***Index Terms*—**
Clinical Decision Support Systems, Multi-Agent Systems, Large Language Models, Biomedical Data, Multimodal Learning, Healthcare Informatics

I. Introduction

Routine clinical practice generates physiological signals, medical images, volumetric examinations, laboratory records, and electronic health records. These sources provide complementary information for clinical decisions, but differences in representation, preprocessing, and analysis complicate their integration into a common workflow [1].

Clinical Decision Support Systems (CDSS) have evolved from rule-based expert systems [2] into data-driven platforms for diagnostic support and clinical workflow assistance [3]. Large Language Models (LLMs) support natural language interaction, report generation, and contextual reasoning [4], [5]. Multi-agent architectures, implemented with frameworks such as CrewAI [6], divide clinical workflows into tasks handled by collaborating agents [7].

Existing studies often focus on individual tasks, such as medical image interpretation or physiological signal analysis [8]. Combining these tasks can require separate pipelines, adding implementation work and limiting interoperability. Fewer studies examine multi-agent orchestration for multimodal biomedical decision support with a locally deployed LLM [9].

This paper presents BioStatusIA, a local multi-agent architecture for multimodal clinical decision support. It combines agents orchestrated through CrewAI with an LLM executed using Ollama [10]. All processing stages, including inference, run within the institution's infrastructure. A normalization layer supports one-dimensional physiological signals, two-dimensional medical images, three-dimensional volumetric examinations, and structured clinical datasets within a common pipeline.

The contributions are:

- A multi-agent architecture that processes physiological signals, medical images, volumes, and structured clinical data within a single workflow.
- A normalization layer that handles differences in biomedical file formats, dimensions, and metadata.
- An orchestration framework combining analytical agents with a local LLM for structured clinical report generation.
- An evaluation of eight completed runs on tabular and image datasets, with six candidate classifiers per dataset, separate cross-validation and test sensitivities, MCC and calibration metrics, and measurements of end-to-end time, LLM generation, training, and inference.

The remainder of this paper reviews related work (Section II), presents the proposed architecture (Section III) and its experimental evaluation (Section IV), discusses the results (Section V), and closes with limitations and conclusions (Sections VI and VII).

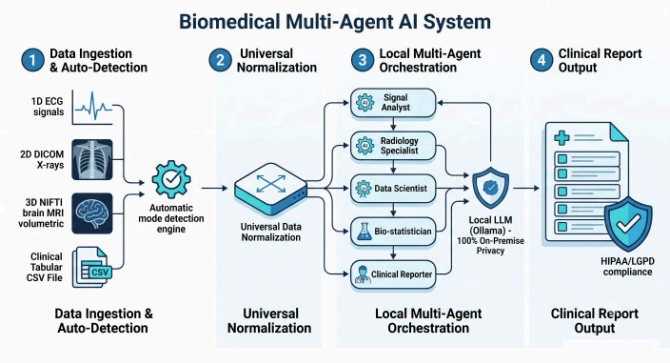


Fig. 1. Overall workflow of the BioStatusIA architecture. An automatic mode-detection step routes each normalized input to the matching specialized agent (Signal Analyst, Radiology Specialist, Data Scientist, or Bio-statistician), whose output is consolidated by the Clinical Reporter using the local LLM; agent roles match Table II.

II. Related Work

Clinical Decision Support Systems (CDSS) range from rule-based systems such as MYCIN [2] to data-driven tools for diagnosis, treatment planning, and documentation [3], [11]. Task-specific pipelines can handle individual clinical problems but require additional integration to process heterogeneous biomedical data within one workflow.

Foundation models support biomedical question answering and text generation, as illustrated by generalist biomedical models [4], LLaMA [5], and Med-Flamingo [8]. Their use in healthcare also requires attention to privacy, computational resources, and response time [9]. Predictive performance alone does not describe the cost of running a complete reporting workflow.

Multi-agent systems divide workflows among specialized components [7]. Frameworks such as CrewAI [6] coordinate these components, allowing modality-specific analysis, model selection, statistical assessment, and report generation to remain separate tasks.

BioStatusIA combines this task separation with local LLM inference and a common interface for biomedical inputs. Its evaluation measures classifier performance and execution time across independent datasets, rather than patient-level multimodal fusion. Table I summarizes the architectural comparison.

III. Proposed Architecture

BioStatusIA combines a data normalization layer, agents orchestrated with CrewAI [6], and a local Large Language Model (LLM) executed through Ollama [10].

Figure 1 illustrates the overall workflow: biomedical data from different modalities are standardized through a normalization layer, dispatched to specialized agents, and consolidated into a structured clinical report.

A. System Overview

BioStatusIA has five layers: Biomedical Data Acquisition, Data Normalization, Multi-Agent Orchestration, Local LLM Reasoning, and Clinical Report Generation. Input data pass through normalization and modality detection before being routed to the appropriate agent. CrewAI coordinates the agents, and the local LLM uses their outputs to produce a structured report. The layers separate modality-specific processing from orchestration and report generation.

B. Multi-modality Data Processing

BioStatusIA handles four data categories, each with a feature-extraction routine. One-dimensional signals (ECG, PPG, EEG) yield time-domain and spectral biomarkers, including RMSSD, SDNN, pNN50, and Welch-FFT band power via `mne`. Two-dimensional image processing supports DICOM-native inputs (Hounsfield-unit windowing and tissue-density estimation) and common formats (CLAHE/Non-Local Means denoising, GLCM texture, and morphology measures such as entropy, contrast, solidity, and circularity). Three-dimensional volumes (CT, MRI) yield tri-planar radiomics, including lesion volume in mm^3, 3D sphericity, and GLCM features. Structured datasets use SHAP-ranked adaptive imputation and feature selection. The normalization layer handles differences in file formats, dimensions, and metadata before agent analysis.

C. Multi-Agent Orchestration

Each agent has a specific task. The *Signal Analyst* processes one-dimensional physiological signals (ECG, PPG, EEG); the *Radiology Specialist* analyzes two-dimensional images (DICOM and common formats) and three-dimensional CT/MRI volumes; the *Data Scientist* runs the AutoML model-selection pipeline (Section IV-A) on structured clinical data; and the *Bio-statistician* performs the EDA statistical audit (Section IV-A) and calibration analysis. The *Clinical Reporter* combines the outputs and forwards them to the local LLM. Table II summarizes these responsibilities.

D. Local LLM Reasoning

BioStatusIA performs LLM inference locally through Ollama [10], without a cloud API. The implementation described here uses Qwen2.5:3B-Instruct. The execution reports used in Section IV do not record the model identifier, quantization, context limit, or hardware, so those runtime settings cannot be confirmed for the reported measurements. The LLM receives structured agent outputs rather than raw biomedical data and uses them to generate clinical summaries. The intended report-generation workflow is represented as

$$R = f(A_1, A_2, \ldots, A_5, M), \tag{1}$$

where A_i denotes each specialized agent (Signal Analyst, Radiology Specialist, Data Scientist, Bio-statistician, Clinical Reporter), M corresponds to the locally deployed LLM, and R is the final structured clinical report.

IV. Experimental Evaluation

A. Experimental Setup

BioStatusIA uses Python, CrewAI [6], and a local LLM served through Ollama [10]. The eight reports dated September

TABLE I
COMPARISON BETWEEN REPRESENTATIVE AI-BASED CLINICAL DECISION SUPPORT APPROACHES AND THE PROPOSED BIOSTATUSIA ARCHITECTURE.

Approach	Local	Multi-Agent	Multi-modality Support	Quant. Metrics	Clinical Reports
Traditional CDSS	✓	×	×	Partial	×
Cloud-based LLM CDSS	×	×	Partial	Partial	✓
Medical Vision AI Systems	✓	×	×	✓	Partial
CrewAI-based Healthcare Systems	Partial	✓	Partial	Partial	✓
BioStatusIA (Proposed)	✓	✓	✓	✓	✓

TABLE II
RESPONSIBILITIES OF THE BIOSTATUSIA AGENTS.

Agent	Primary Responsibility
Signal Analyst	Physiological signal analysis (1D)
Radiology Specialist	Medical imaging analysis (2D/3D)
Data Scientist	AutoML model selection
Bio-statistician	Statistical audit & calibration
Clinical Reporter	Evidence consolidation & report generation

14, 2026 record preprocessing, classifier results, confusion matrices, and execution times. Candidate classifiers were Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP).

The preprocessing records indicate robust scaling for all five tabular datasets and median imputation for Heart Disease Cleveland; no imputation was recorded for the other tabular datasets. The three image datasets used Gaussian smoothing, min–max normalization, and resizing to 256×256, without contrast equalization.

B. Datasets

Table III lists the datasets included in the completed runs. Breast Cancer Wisconsin[1] [12], Stroke Prediction Clinical[2], PIMA Diabetes[3] [13], Heart Disease Cleveland[4], and Parkinson's Vocal Biomarkers[5] [14] were processed as tabular data. BUSI[6] [15], Brain Tumor MRI[7], and Brain MRI Oncology[8] were processed as labeled image collections. Parkinson's inputs were acoustic measurements rather than raw audio. These runs therefore cover tabular data and two-dimensional images, not raw physiological signals or three-dimensional volumes. The datasets were analyzed independently, without patient-level multimodal fusion.

The Brain Tumor MRI run contains 7,200 images across four encoded classes. Heart Disease Cleveland retains five classes and is not a binary disease-detection task. BUSI is reported with two encoded classes, but the export does not document how the original diagnostic categories were mapped to these labels.

[1]https://www.kaggle.com/datasets/abdelazizsami/breast-cancer-wisconsin-diagnostic

[2]https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[3]https://www.kaggle.com/datasets/ipeypazzesco99/pima-indians-diabetes-dataset

[4]https://www.kaggle.com/datasets/abdelazizsami/heart-disease

[5]https://www.kaggle.com/datasets/pypiahmad/oxford-parkinsons-disease-detection-dataset

[6]https://www.kaggle.com/datasets/subhajournal/busi-breast-ultrasound-images-dataset

[7]https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[8]https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

TABLE III
DATASETS IN THE EXECUTION REPORTS. N IS THE REPORTED SAMPLE COUNT, N_{test} IS THE CONFUSION-MATRIX TOTAL, AND C IS THE NUMBER OF ENCODED CLASSES.

Dataset	N	N_{test}	C
Breast Cancer Wisconsin	569	114	2
BUSI Ultrasound	780	156	2
Stroke Clinical	5,110	1,022	2
PIMA Diabetes	768	154	2
Brain Tumor MRI	7,200	1,440	4
Brain MRI Oncology	253	51	2
Heart Cleveland	303	61	5
Parkinson's Voice	195	39	2

C. Model Selection and Evaluation Protocol

Models were ranked by an internal cross-validation clinical score, not test performance, with a sensitivity floor of $S_{\min} = 0.80$. No candidate met this floor for BUSI, Stroke, PIMA, or Heart Disease Cleveland. The returned models had cross-validation sensitivities of 0.619, 0.761, 0.724, and 0.345, respectively, and carried explicit warnings. They remain in the descriptive comparison despite failing the selection constraint.

The confusion-matrix totals are consistent with approximately 20% of each dataset being reserved for testing. The reports distinguish internal cross-validation sensitivity from test sensitivity, but do not identify the random seeds, fold count, subject-group assignments, or fold-specific preprocessing fits. Patient-level separation and training-only preprocessing cannot be verified from these exports. In particular, the Parkinson's results cannot be described as performance on held-out subjects without the corresponding partition records.

Table IV reproduces the selected models' reported metrics. Percentage values were converted to proportions. AUROC denotes the area under the ROC curve, MCC the Matthews Correlation Coefficient, and ECE the Expected Calibration Error. Multiclass aggregation settings and ECE binning are not specified in the reports. Cross-dataset summaries use an unweighted mean and sample standard deviation over the eight reported values; they describe variation across tasks, not uncertainty from repeated runs. ROC curves are not reconstructed because prediction scores are unavailable.

D. Predictive Performance

Across the eight datasets, mean AUROC was 0.866 ± 0.109 and mean MCC was 0.580 ± 0.295. Breast Cancer Wisconsin

TABLE IV
SELECTED-MODEL PERFORMANCE. ONLY SENS. CV IS FROM INTERNAL CROSS-VALIDATION; ALL OTHER METRICS ARE FROM THE HELD-OUT TEST SET.

Dataset	Model	Acc.	AUROC	Sens. CV	Sens. test	Spec.	MCC	ECE
Breast Cancer Wisconsin	MLP	.991	.995	.971	.976	1.000	.981	.010
BUSI Ultrasound	SVM†	.744	.733	.619	.667	.772	.409	.146
Stroke Clinical	LR†	.750	.841	.761	.780	.748	.254	.259
PIMA Diabetes	SVM†	.734	.794	.724	.778	.710	.467	.125
Brain Tumor MRI (4-class)	RF	.901	.983	.880	.901	.967	.868	.082
Brain MRI Oncology	RF	.804	.849	.856	.871	.700	.583	.092
Heart Cleveland (5-class)	RF†	.508	.746	.345	.222	.862	.211	.156
Parkinson's Voice	MLP	.949	.986	.926	.965	.900	.866	.059
Mean	–	.798	.866	.760	.770	.832	.580	.116
Sample SD	–	.153	.109	.203	.245	.116	.295	.075

† No candidate reached the internal cross-validation sensitivity floor of 0.80. The returned model carries a warning in the execution report.

had the highest MCC (0.981), followed by Brain Tumor MRI (0.868) and Parkinson's Voice (0.866). BUSI, Stroke, PIMA, and Heart Disease Cleveland had MCC values between 0.211 and 0.467 and did not meet the internal sensitivity floor. Four selected models had $ECE > 0.10$: BUSI, Stroke, PIMA, and Heart Disease Cleveland.

Selection did not necessarily maximize test performance: on Brain MRI Oncology, RF had MCC 0.583 versus MLP's 0.715; on BUSI, the selected SVM had AUROC 0.733 versus RF's 0.824. Neither comparison justifies selecting a different model after observing test results.

E. Execution Time

End-to-end time ranged from 80.073 to 809.475 s, while LLM generation took 41.593–789.218 s (Table V). Selected-model training took 0.035–1.959 s, and classifier inference took 0.0005–0.4427 ms per sample. The inference measurement does not include feature extraction, model search, orchestration, or report generation. LLM throughput ranged from 3.51 to 26.89 tokens/s.

For Stroke, LLM and AutoML completion occurred at 52.285 and 507.249 s, respectively; selected-model training took 0.035 s and does not represent total search time. For Brain Tumor MRI, LLM generation occupied 789.218 of 809.475 s. Without a complete timing breakdown or repeated measurements, stage-specific overhead and timing variance cannot be inferred.

V. DISCUSSION

The eight runs document a common workflow for tabular clinical data and labeled medical images. They do not test the joint analysis of multiple modalities from the same patient. The architectural support for signals and volumes should therefore be distinguished from the modalities covered by this evaluation.

Predictive performance varied substantially across tasks. Breast Cancer Wisconsin, Brain Tumor MRI, and Parkinson's Voice had MCC values above 0.86, while Stroke and Heart Disease Cleveland remained below 0.26. For Stroke, the selected LR model detected 39 of 50 positive test cases but produced 245 false positives. Its AUROC of 0.841 must therefore be considered alongside MCC 0.254 and ECE 0.259. In Cleveland's five-class task, RF correctly classified 29 of 33 class-0 cases but only two of the remaining 28 cases. The overall specificity of 0.862 does not offset this poor recognition of the other classes.

Four runs failed the internal sensitivity requirement. Returning a model with a warning preserves the analytical output, but does not make that model suitable for clinical use. The same four selected models also had $ECE > 0.10$. Changes to class weighting, decision thresholds, and probability calibration require assessment within the training procedure, followed by evaluation on independent data.

The execution times separate fast classifier inference from the much longer interval needed to complete the workflow. LLM generation accounted for most of the elapsed time in the image runs, while Stroke had a long interval between LLM and AutoML completion. Optimizing classifier inference alone would not address these costs. Since the LLM-completion milestone precedes AutoML completion in all eight reports, these records also do not establish that the generated narrative incorporated the final selected classifier's results.

Local report generation is an architectural feature, not evidence of report accuracy. The exported Parkinson's narrative describes a distribution of 48 controls and 147 disease-labeled recordings as balanced and asserts feature significance without presenting a corresponding inferential test. These statements should not be carried into the interpretation of the quantitative results. Clinical review must assess whether generated text accurately reflects the data and avoids unsupported conclusions.

VI. LIMITATIONS

The evaluation is based on one exported run per dataset. The reports do not include partition indices, random seeds, cross-validation fold definitions, or preprocessing-fit records. They therefore do not establish patient-level independence or exclude preprocessing leakage. This is particularly relevant to repeated acoustic recordings in Parkinson's and to image collections that may contain multiple samples from one patient. The reported Parkinson's performance should not be interpreted as subject-independent generalization.

TABLE V
RECORDED EXECUTION TIMES AND LLM THROUGHPUT. EACH ROW REPRESENTS ONE RUN. TRAINING AND INFERENCE MEASUREMENTS REFER ONLY TO THE SELECTED CLASSIFIER.

Dataset	End-to-end (s)	LLM generation (s)	Training (s)	Inference (ms/sample)	LLM (tokens/s)
Breast Cancer Wisconsin	80.073	41.593	1.387	0.0040	21.95
BUSI Ultrasound	314.781	304.126	0.189	0.0713	8.15
Stroke Clinical	507.249	52.280	0.035	0.0005	24.06
PIMA Diabetes	170.711	43.783	0.122	0.0550	24.48
Brain Tumor MRI	809.475	789.218	1.959	0.0268	3.51
Brain MRI Oncology	143.343	142.190	0.242	0.4427	12.17
Heart Cleveland	172.130	99.161	0.341	0.3419	26.89
Parkinson's Voice	90.358	49.431	0.921	0.0155	14.30

End-to-end time is the reported upload-to-ready interval. LLM generation time is a duration, not the report-completion timestamp. Stage durations are not assumed to be additive.

Dataset construction also requires further documentation. The BUSI export uses two labels without specifying the mapping from diagnostic categories. The Brain Tumor MRI export reports 7,200 images without identifying their provenance or any duplicate or augmentation checks. The Wisconsin descriptive statistics include an identifier column, but the report does not establish whether it was excluded from classifier inputs. Multiclass metric aggregation and ECE binning are also unspecified.

The means across datasets combine binary and multiclass tasks with different sample sizes and class distributions. They are descriptive summaries, not pooled clinical performance estimates. No repeated-split uncertainty estimates, independent baseline comparisons, or matched hypothesis tests are available in these exports. Four runs failed the sensitivity floor, and four selected models had elevated calibration error.

Timing values come from single executions and do not include a complete breakdown of preprocessing, feature extraction, search, and orchestration. The exports do not identify the hardware, model quantization, context configuration, or concurrency settings used in each run. These measurements cannot establish real-time performance or portability across devices.

The datasets were processed independently, without hospital deployment, Electronic Health Record integration, or prospective validation. Raw-signal processing and volumetric analysis were not evaluated in these runs. Report quality was not assessed by clinical raters; factual accuracy and agreement between reviewers remain to be measured.

VII. CONCLUSION

BioStatusIA combines data normalization, task-specific agents, classifier selection, and local LLM report generation. The eight completed runs cover five tabular datasets and three labeled image collections. Selected classifiers achieved mean AUROC 0.866 ± 0.109 and MCC 0.580 ± 0.295, with substantial differences between datasets. Four runs did not meet the internal cross-validation sensitivity floor of 0.80.

End-to-end processing took 80.073–809.475 s, and LLM generation took 41.593–789.218 s. Classifier inference was below 0.443 ms per sample, but this measurement excludes the other workflow stages. These results document execution on the tested datasets, not clinical validity or real-time operation.

Further evaluation requires reproducible patient-level partitions, documented label mappings and preprocessing, repeated timing measurements, and independent assessment of calibration and report accuracy. Prospective clinical testing should follow these checks.

REFERENCES

[1] F. Jiang, Y. Jiang, H. Zhi *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[2] E. H. Shortliffe, "Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the mycin system," *Computers and Biomedical Research*, vol. 8, no. 4, pp. 303–320, 1975.

[3] E. S. Berner, "Clinical decision support systems: State of the art," *AHRQ Publication*, 2007.

[4] T. Tu, S. Azizi, D. Driess *et al.*, "Towards generalist biomedical ai," *arXiv preprint arXiv:2307.14334*, 2023.

[5] H. Touvron, T. Lavril, G. Izacard *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[6] CrewAI Inc., "Crewai: Multi-agent framework for ai automation," https://github.com/crewAIInc/crewAI, 2024.

[7] X. Yang *et al.*, "Large language model based multi-agent systems: A survey," *arXiv preprint arXiv:2402.01680*, 2024.

[8] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec, "Med-flamingo: A multimodal medical few-shot learner," *arXiv preprint arXiv:2307.15189*, 2023.

[9] T. Guo *et al.*, "Large language models in healthcare: A survey," *arXiv preprint arXiv:2311.07658*, 2024.

[10] Ollama Team, "Ollama," https://ollama.com, 2024.

[11] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[12] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer Wisconsin (diagnostic) data set," *UCI Machine Learning Repository*, 1995.

[13] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1988, pp. 261–265.

[14] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[15] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

Brain symmetry index alterations in stroke patients after rehabilitation with cathodic tDCS: an exploratory study

María Florencia Molina
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas
Campinas, Brasil
m169254@dac.unicamp.br

Camila D. Lima
IFGW - UNICAMP
Campinas, Brasil
CSP and QBI - University of Queensland
Brisbane, Australia
caplima@unicamp.br

Karoliny T. Cruz
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas
Campinas, Brasil
karolfisio6@gmail.com

Luara Inocêncio
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas
Campinas, Brasil
luarainocencio27@gmail.com

Bruna M. Carlos
Polytechnic School
University of São Paulo (USP)
São Paulo, Brazil
bruna.carlos@usp.br

Pedro F. G. de Vazquez
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas
Campinas, Brazil
p263293@dac.unicamp.br

Gabriela Castellano
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas
Campinas, Brazil
gabriela@unicamp.br

***Abstract*—Post-stroke spasticity is associated with pathological interhemispheric asymmetry in cortical excitability. Cathodal transcranial direct current stimulation (c-tDCS) over the vertex may help restore interhemispheric balance by reducing hyperexcitability in cortical regions. This is an exploratory study which aimed to assess the effects of c-tDCS combined with extended reality rehabilitation on the revised Brain Symmetry Index (r-BSI), an EEG-derived measure of interhemispheric power asymmetry, in spastic stroke survivors. Thirteen (post data curation) patients underwent ten rehabilitation sessions over two weeks in a crossover design, with EEG recorded before and after stimulation on the first and last session days. The r-BSI was computed across delta, theta, alpha, beta, and broadband (1–30 Hz) frequency ranges. Within-session comparisons revealed a significant reduction in delta-band r-BSI following real tDCS, indicating an acute increase in interhemispheric symmetry; the broadband showed a nominally significant effect that did not survive correction for multiple comparisons, likely reflecting limited statistical power given the sample size. No significant differences were observed in the sham group for any frequency band. Between the first and last sessions, no cumulative effect on r-BSI was detected. These findings suggest that c-tDCS produces an acute, stimulus-locked modulation of interhemispheric asymmetry, but that this effect does not accumulate across a ten-session protocol. Larger samples and longer follow-up are needed to clarify the neurophysiological trajectory of tDCS-based rehabilitation.**

***Index Terms*—stroke, rehabilitation, transcranial direct current stimulation (tDCS), revised brain symmetry index (r-BSI), electroencephalography (EEG), interhemispheric asymmetry, spasticity**

This research was supported by Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001; São Paulo Research Foundation (FAPESP), Grants 2025/10777-0 (MFM), 2023/08802-0 (CPL), 2025/11139-7 (KTC), 2020/16571-0 (BMC), 2025/10963-8 (PFGV), 2013/07559-3 (BRAINN), 2024/00998-6 (CRISQuaM); Brazilian National Council for Scientific and Technological Development (CNPq), Grants 306255/2024-3 and 408389/2024-9 (INCT-NeuroComp); and Funding Authority for Studies and Projects (FINEP), Grant 2140/22.

I. Introduction

Stroke is one of the leading conditions causing motor impairments in people worldwide and in Brazil [1]. Chronic stroke survivors frequently present with sensorimotor impairments, including hemiparesis and spasticity. Post-stroke spasticity is characterized by a velocity-dependent increase in tonic stretch reflexes resulting from disrupted supraspinal control over spinal motor circuits [2]. Beyond its impact on motor function and quality of life, spasticity contributes significantly to pain, contractures, and caregiver burden, making its effective management a central goal of neurological rehabilitation [3].

Reduced access to exercise and technology-based rehabilitation programs can further compromise quality of life and functional independence in these individuals [4], [5]. Although conventional exercise rehabilitation programs have been shown to yield modest improvements in sensorimotor function, long-term exercise combined with neuromodulation, such as transcranial direct current stimulation (tDCS), may be more effective in addressing these impairments [6]–[8]. TDCS

is a non-invasive neuromodulatory technique that applies weak direct electrical currents to the scalp to modulate cortical excitability in a polarity-dependent manner. Anodal tDCS generally increases neuronal firing thresholds, while cathodal tDCS suppresses cortical excitability in the targeted region. The somatosensory cortex is the brain region responsible for processing afferent input from mechanoreceptors (i.e; information from sensory fibres) that are located in muscles, tendons, joints, and skin [9]. Cathodal tDCS (c-tDCS) may reduce hyperexcitability of peri-lesional tissue, thereby attenuating the excessive cortical activity thought to contribute to spasticity [10]–[12]. By reducing this asymmetric cortical excitability between hemispheres, c-tDCS over the somatosensory cortex may help restore interhemispheric balance in spastic stroke survivors [13], [14].

Indeed, a key neurophysiological mechanism underlying post-stroke motor impairment, including spasticity, is the disruption of interhemispheric balance. Following a unilateral lesion, the contralesional (unaffected) hemisphere tends to exert excessive inhibition over the ipsilesional (affected) hemisphere through transcallosal pathways, a phenomenon known as interhemispheric competition. This pathological asymmetry in cortical excitability is thought to further suppress motor output from the lesioned side, compounding functional deficits [15]. Therefore, assessing this balance may give a measure of patient recovery. This can be done using the brain symmetry index (BSI) - an electroencephalography (EEG)-derived measure of interhemispheric power asymmetry [16]. BSI has been shown to be significantly elevated in post-stroke patients compared to healthy controls, and early BSI values correlate with motor function scores measured one to two months later, supporting its value as both a prognostic and an outcome measure [15]. Further evidence indicates that BSI is independently associated with functional disability at six months post-stroke, reinforcing its relevance beyond the acute phase [17].

Despite this, BSI has been rarely employed as an outcome measure in tDCS intervention trials. Among the few intervention studies that have incorporated BSI, Williamson et al. examined the effects of high-definition tDCS on quantitative EEG metrics in stroke survivors and reported that both anodal and cathodal stimulation significantly decreased BSI across multiple frequency bands, a finding associated with reduced motor impairments and improved corticomotor conduction, suggesting that BSI may serve as a sensitive indicator of neurophysiological change induced by tDCS [18]. In a study combining tDCS with extended reality (XR) rehabilitation in chronic stroke, significant reductions in both BSI and the directional BSI were observed in the active tDCS group across alpha, beta, and gamma frequency bands, suggesting an increase in interhemispheric symmetry following intervention; notably, symmetry indices at higher frequencies correlated significantly with clinical improvements, showing a stronger relationship with lower extremity function and an increasing symmetry trend across the rehabilitation program [19]. In a pilot study combining motor imagery brain-computer interface with tDCS, BSI change emerged as the signature neurophysiological biomarker for the tDCS condition, in contrast to relative theta power which characterized the BCI-only group, suggesting that these two approaches engage distinct mechanisms of neural repair [20]. These findings support BSI as a meaningful and sensitive neurophysiological outcome measure in tDCS-based rehabilitation, yet its application remains limited to a small number of studies, none of which have specifically examined cathodal tDCS in the context of post-stroke spasticity.

Therefore, the aim of this study was to apply c-tDCS over the somatosensory cortex (Cz, EEG 10–20 system) across ten rehabilitation sessions focusing on the lower limb, and to assess brain symmetry indices in spastic stroke survivors.

TABLE I
PARTICIPANT DEMOGRAPHICS AND CLINICAL CHARACTERISTICS

Pt.	Age	Sex	Type	TAS (mo)	Hand	Aff. Hem.
MC	64	F	I	–	R	R
LA	75	M	H	35	R	R
BV	42	M	I	9	R	R
RO	65	F	I	78	L	R
VR	53	F	I	74	R	R
MA	63	M	H	72	L	R
RM	32	M	I	179	R	R
CG	43	F	I	12	L	L
PM	67	M	I	29	L	R
AD	54	F	I	17	R	R
MM	70	F	I	63	L	R
GL	29	M	I	36	L	L
DJ	60	M	I	36	L	L
SA	59	F	I	67	R	L
VE	42	F	I	120	R	L

I: Ischemic; H: Hemorrhagic; TAS: Time after stroke (months); Hand: Handedness; Aff. Hem.: Affected hemisphere; L: Left; R: Right; –: not available.

II. MATERIALS AND METHODS

A. Subjects

The cohort for this study consisted of fifteen individuals (age 54.5 ± 14.6 years, nine women) who had suffered a stroke with a minimum of six months previously (57.6 ± 46.5 months post-stroke). All exhibited some degree of hemiparesis, as assessed by the Fugl-Meyer scale. The experimental protocol was approved by the local ethics committee (CAAE 75571323.3.0000.5404), and all participants signed a written informed consent form before entering the study. Of the fifteen initial patients, one did not complete the protocol and some EEG data from another were lost, so that a final total of 13 patients (age 53.9 ± 12.8, eight women) were considered for this study. Table I shows the ages, sex and affected hemisphere of the participants. The data from patients "LA" and "BV" were not included in this study, the first due to the loss of some EEG data and the second because the patient did not complete the two week rehabilitation protocol.

B. Experimental protocol

Patients underwent ten days of rehabilitation over two weeks. In the first and last days, EEG evaluations were also performed, following the same schedule. At the beginning, a one-minute resting-state EEG was recorded with eyes closed, in a dimly lit room. After the EEG, physiotherapists carried out a functional assessment of the patient, followed by a rehabilitation session combining extended reality exergames with tDCS. Patients were randomly allocated to the experimental or control group, and a double-blind design was maintained throughout the study, such that neither the patient nor the evaluator administering the tDCS were aware of group allocation (sham vs. real tDCS). If the patient belonged to the sham group, sham tDCS was applied instead. At the end of the rehabilitation session, the patient underwent another EEG evaluation under the same conditions as at the beginning. This protocol was also repeated on the last day of rehabilitation, resulting in a total of four EEG acquisitions, which throughout this article will be referred to as:

- S1-pre: first EEG acquisition of the first rehabilitation day, before tDCS (real or sham) application;
- S1-post: last EEG acquisition of the first rehabilitation day, after tDCS (real or sham) application;
- S10-pre: first EEG acquisition of the last rehabilitation day, before tDCS (real or sham) application;
- S10-post: last EEG acquisition of the last rehabilitation day, after tDCS (real or sham) application.

In sessions 2 through 9, patients underwent a rehabilitation session similar to those on days 1 and 10, including tDCS, but without EEG recordings or other assessments.

After completing the first rehabilitation round, patients were invited to participate in a second identical round, with the only difference being a reversal of group membership: patients who had received sham tDCS in the first round received active tDCS in the second, and vice versa. Of the thirteen patients included in the analysis, eight completed the second round of sessions. In the end, we have:

- Real tDCS group: 7 patients in the first round, 4 in the second round $\rightarrow$ 11 acquisitions.
- Sham tDCS group: 6 patients in the first round, 4 in the second round $\rightarrow$ 10 acquisitions.

C. Rehabilitation protocol

The rehabilitation protocol consisted of ten sessions over two weeks conducted on weekdays. In each session, participants performed upper limb exercises followed by lower limb exercises using the BRAINN XR tools [21], [22]. Lower limb exercises focused on ankle mobility. The session began with a warm-up consisting of three components: three minutes of walking, followed by two sets of ten repetitions of dynamic stretching using a ramp, and two sets of ten repetitions of dynamic stretching with a resistance band. A five-minute rest was observed between the warm-up and the main exercises. The main circuit included seven exercises: 1) Sit-to-stand from a chair: the participant rose from and returned to a seated position five times per set, with an object moving towards their head to challenge postural control; 2) Sit-to-stand with the ramp: the same movement performed with the feet on the inclined ramp and an object moving towards their head; 3) Squat with feet on the ramp: a squatting movement with both feet on the ramp and an objects moving towards their head; 4) Stationary marching with obstacle clearance Google Maps-based environment: the participant marched in place lifting the feet to step over ground-level obstacles, for three minutes; 5) Step-ups and step-downs with the ramp: the participant stepped up and down using the ramp while clearing obstacles, for three minutes; 6) Seated dynamic stretching with the resistance band: dynamic ankle stretches performed seated using the band; 7) Sit-to-stand holding the resistance band with the opposite (non-paretic hand) hand: the sit-to-stand movement performed while holding the band contralateral to the affected limb. Each exercise began with one set of five repetitions and was progressively increased based on heart rate response and perceived exertion. The session closed with a 10-minute cool-down of static stretching, three sets of 30 seconds per stretch.

Cathodal tDCS was applied to the real tDCS group during the main exercise circuit, for 15 minutes, using a 1 mA current, with a ramp-up and ramp-down time of 30 seconds. The electrodes (35 cm^2) were fitted with surface sponges soaked in saline solution and mounted with care to avoid leakage onto the hair, while ensuring the sponges maintained consistent moisture throughout the entire stimulation session. The cathode was positioned on the scalp over the legs' motor cortex region (equivalent to Cz in the 10-20 system) and the anode was positioned on the middle of the forehead, above the eyebrow. For the sham tDCS group, the only difference was that immediately after the 30 seconds ramp-up a 30 s ramp-down to 0 mA current was applied, and it remained at 0 mA during the whole 15 minutes.

D. EEG equipment

EEG recordings were performed using a g.USBAmp amplifier and 16 dry electrodes (g.SAHARA, g.tec, Austria). Electrodes were positioned in the 10-20 system in the following locations: AFz, F1, F2, F5, F6, Cz, C1, C2, C3, C4, C5, C6, P1, P2, P5, P6.

E. Preprocessing

The first second of each recording was discarded to remove the transient interval during which the signal tends to be noisier. To facilitate the analysis, the right hemisphere was designated as the affected hemisphere for all patients; for those whose lesion was located in the left hemisphere, the electrode labels were mirrored from left to right. For each acquisition, the power spectrum of the signal was computed to identify interference or peaks not consistent with typical EEG activity.

The data were then processed using the HAPPE pipeline [24]. Within the pipeline, the data were filtered with a notch filter at 60 Hz and a bandpass filter from 1 Hz to 100 Hz, together

with CleanLine filtering (EEGLAB) at frequencies where atypical peaks were observed. The pipeline was configured to perform bad channel detection, so that electrodes whose quality were deemed sub-optimal by the pipeline were eliminated. The software then performs artifact detection through independent component decomposition combined with a wavelet transform. Finally, the pipeline applies EEGLAB ICA to detect muscular artifacts in the data.

The preprocessed data were filtered into frequency bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz). The gamma band was excluded because, in resting-state EEG, activity in this frequency range is rarely attributable to genuine brain activity.

F. Data analysis

The Revised Brain Symmetry Index (r-BSI) [23] is defined as

$$\text{r-BSI} = \frac{1}{K}\sum_{n=1}^{K}\frac{|R_n^*(t) - L_n^*(t)|}{R_n^*(t) + L_n^*(t)}$$

where the summation runs over frequencies $n = 1, \ldots, K$ and

$$R_n^*(t) = \frac{1}{M}\sum_{ch=1}^{M} a_n^2(ch, t).$$

The coefficients $a_n(ch, t)$ are the Fourier coefficients at frequency n for channel ch at time t. The channels considered for R_n^* are those located in the right hemisphere, and analogously for L_n^* with the channels of the left hemisphere. The summation in R_n^* (L_n^*) runs over indices $i = 1, \ldots, M$, corresponding to the M channels located in the right (left) hemisphere, thus computing the mean power of each hemisphere at each frequency bin. The r-BSI compares these powers at each frequency and averages across all of them. Note that, as aforementioned, here the right hemisphere represents the lesioned one.

Asymmetry indices (note that r-BSI, despite its name, measures asymmetry, i.e., it is larger the more asymmetric the hemispheres are) were computed for each EEG acquisition, and then averaged per session and across patient groups (real and sham tDCS), in order to compare the evolution of this metric across sessions.

It is worth mentioning that changes in other biomarkers were explored, specifically functional connectivity and network topology, but no significant variations were registered with these methods.

G. Statistical analysis

The r-BSI distributions for the pre and post-tDCS EEG acquisitions were assessed for normality using the Shapiro-Wilk test. Given that some were not normal, a Wilcoxon signed-rank test was used to compare the distributions between pre and post-tDCS EEG acquisitions, for each group (real and sham tDCS). The results were then corrected using the Benjamini–Hochberg (BH) procedure to control for False Discovery Rate (FDR), considering five frequency bands.

III. Results

Figure 1 shows the mean values of r-BSI for each frequency band (delta, theta, alpha and beta, as well as for the broadband - 1-30 Hz), averaged across each patient group.

It can be observed that the variation of the symmetry index shows, across all frequency bands, a consistent pattern in the real tDCS group: higher index values in acquisitions performed before tDCS application, and a reduction in those performed after. For patients in the sham tDCS group, mostly no consistent trend can be identified across frequency bands or recording conditions, apart from in the theta band, where a trend similar to that of the real tDCS group is observed, though less pronounced. For this group, in the alpha and beta bands, as well as in the broadband (1-30 Hz), the index remains approximately constant across sessions. In the delta band, the index shows a slight decrease in the intermediate sessions followed by a slight increase in the last session.

The results indicate that tDCS produces an immediate reduction in interhemispheric asymmetry; however, this effect does not appear to persist over time, as values return toward baseline by S10-pre. With this in mind, the joint distribution of the r-BSI values for both sessions (S1 and S10) pre-tDCS and post-tDCS were examined and compared, which can be seen in Figure 2. Only the mean values of r-BSI for the delta band of the real tDCS group were significantly greater for the pre-tDCS acquisition (larger symmetry) than for the pre-tDCS acquisition. The broad band also showed a tendency (before FDR correction) toward a lower value in the post-tDCS session. On the other hand, for the sham tDCS group, there were no statistically significant differences among pre and post-tDCS acquisition for any frequency band.

IV. Discussion

The aim of this study was to assess a possible recovery of interhemispheric symmetry in stroke survivors who underwent a rehabilitation protocol using cathodal tDCS over the motor cortex related to the lower limbs (Cz). The aim behind using cathodal, rather than the more common, anodal tDCS, was to reduce spasticity in the affected lower limb. Our hypothesis was that this would facilitate motor rehabilitation of the affected limb and consequently, increase interhemispheric symmetry. Our results showed a significant decrease in the delta frequency band r-BSI (increase in symmetry) and a tendency of decrease for the beta and broad band (1-30 Hz) r-BSI within a single session, i.e., just after the tDCS application, but not between the first and last rehabilitation sessions. This decrease was not observed for the sham tDCS group, which seems to indicate that rehabilitation alone, without tDCS, does not significantly influence the symmetry index.

The within-session decrease in delta band r-BSI observed in the real tDCS group — but not in the sham group — is consistent with findings from other tDCS-EEG studies, both in terms of the acute modulation of interhemispheric asymmetry and the apparent dissociation between immediate neurophysiological effects and longer-term cumulative changes. Regarding the acute effect on EEG power, a randomized crossover study

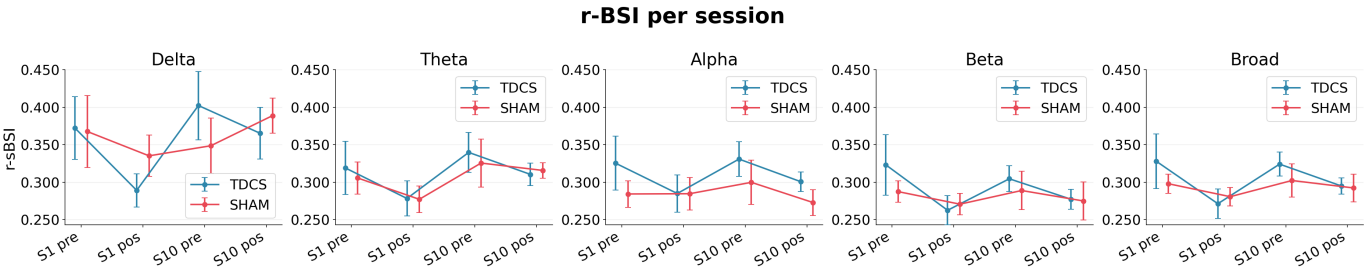


Fig. 1. Variation of the mean r-BSI between the real and sham tDCS groups by frequency band and session for the resting-state EEG.

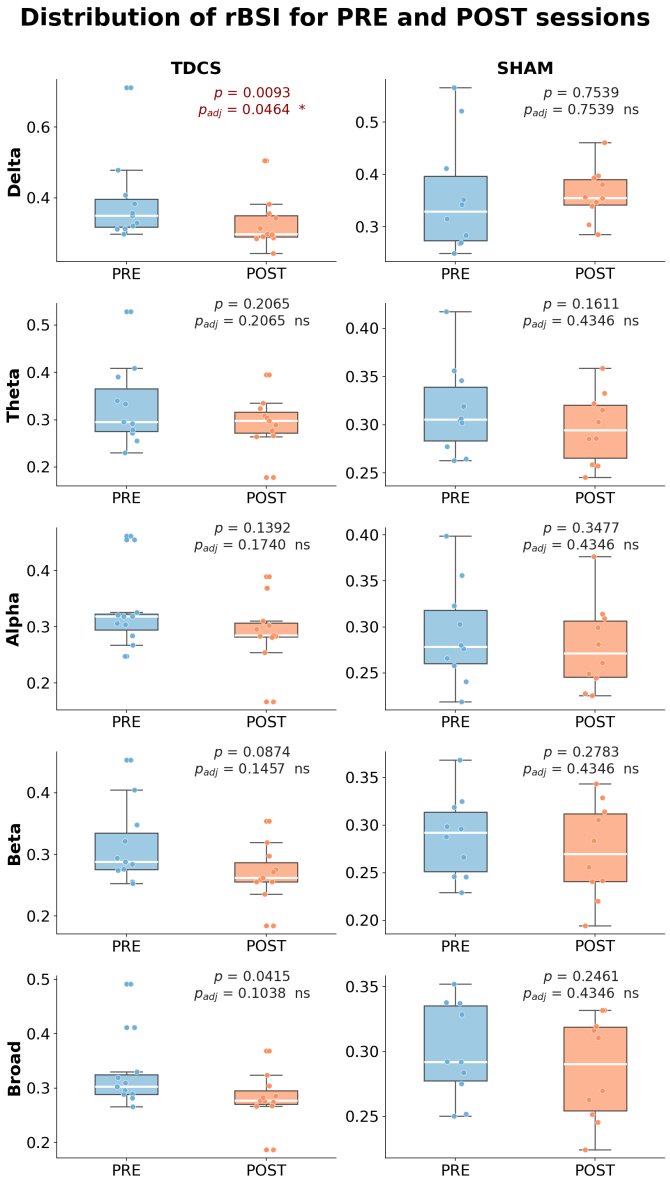


Fig. 2. Variation of the mean interhemispheric symmetry index (r-BSI) between the real and sham tDCS groups by frequency band and session for resting-state EEG. The p-values before and after FDR correction (the latter corresponding to p_{adj}) are shown. The alternative hypothesis was that r-BSI values were greater in the pre-stimulation distribution than in the post-stimulation distribution.

in ischemic stroke patients found that real tDCS, compared to sham, produced a significant decrease in delta power alongside a significant increase in alpha power in resting-state EEG recorded immediately after stimulation, with no such changes observed in the sham condition — a pattern that closely mirrors our within-session r-BSI results [25]. In terms of BSI specifically, Williamson et al. reported that both anodal and cathodal high density (HD)-tDCS of the contralesional hemisphere significantly decreased BSI across multiple frequency bands in stroke survivors, with this reduction associated with decreased motor impairments, further supporting the sensitivity of BSI to the acute neurophysiological effects of tDCS [26].

Regarding the absence of a cumulative BSI effect across sessions, a randomized controlled trial of five consecutive sessions of anodal tDCS combined with physical therapy in acute stroke found that tDCS did not result in superior EEG changes over sham across sessions, despite improvements in lower extremity function in participants with low motor impairment, suggesting that neurophysiological and clinical trajectories do not always align over a multi-session course [27]. Similarly, Di Lazzaro et al. found that bilateral tDCS combined with constraint-induced movement therapy significantly reduced interhemispheric imbalance at neurophysiological level, yet none of the clinical outcome measures showed differential modulation between real and sham groups [28], a dissociation also noted in a study of bihemispheric tDCS in chronic stroke where neurophysiological adaptations including lasting reductions in interhemispheric inhibition did not correlate with any clinical improvement, which the authors attributed to a possible ceiling effect for motor recovery [29]. These findings suggest that the acute, within-session modulation of interhemispheric symmetry by cathodal tDCS may reflect a transient and stimulus-locked neurophysiological effect that does not necessarily accumulate linearly across a rehabilitation program — a pattern that warrants further investigation with longer follow-up and larger samples.

The pattern of results is consistent with the proposed mechanism: cathodal tDCS consistently reduces r-BSI across all frequency bands, with the most pronounced and statistically robust effect in the delta band. The complete absence of significant results in the sham group — even without correction — strengthens the interpretation that the observed changes in the tDCS condition are specific to the stimulation and

not attributable to rehabilitation practice effects, procedural learning, or regression to the mean.

A possible physiological explanation for this effect lies in the mechanism of action of cathodal tDCS. Unlike anodal stimulation, which increases cortical excitability in the stimulated hemisphere, cathodal current hyperpolarizes cortical neurons, reducing their resting membrane potential and decreasing their spontaneous firing rate [30]. In the context of stroke, the contralesional hemisphere frequently exhibits secondary hyperexcitability, resulting from the loss of interhemispheric inhibition exerted by the affected hemisphere — a well-documented phenomenon in the reciprocal interhemispheric inhibition model [31]. The placement of the cathodal electrode over Cz, a vertex position overlying the medial aspects of both motor cortices, results in a broadly distributed current that may exert a net inhibitory influence over hyperexcitable cortical regions in both hemispheres, with the contralesional hemisphere potentially being more susceptible given its elevated excitability at baseline [32]–[34]. This reduction in spectral power asymmetry between hemispheres would bring both hemispheres closer together in terms of electrophysiological activity and, since the r-BSI quantifies the relative power asymmetry between them, would result in a decrease in the index, consistent with the observed pattern. It should be acknowledged, however, that the vertex montage introduces greater uncertainty regarding the spatial specificity of the neuromodulatory effect compared to montages targeting the contralesional motor cortex directly, and future studies using individualized current modeling would help clarify the cortical regions most affected by this electrode configuration.

This study has some limitations, the main one being sample size ($n = 11$): the BH correction with families of five comparisons is relatively costly in terms of statistical power, and several comparisons with medium effect sizes fall in the marginal zone (p_{adj} between 0.07 and 0.11). This does not imply absence of effect, but rather insufficient statistical power to confirm it with certainty. Nevertheless, it is important to clarify that intervention studies with disabled patients are inherently difficult in terms of having a large sample. As described in the methodology section, the rehabilitation protocol lasted for two weeks. Because participants have mobility issues, most of them depend on other people (usually someone from the family) to bring them to the sessions; which means in many cases this person has to miss work for this. Also, due to physical space and staff constraints, we could only rehabilitate a maximum of two patients per two-week period. Finally, it was hard to find patients that attended to the inclusion and exclusion criteria. Due to all these issues, data collection from these 15 participants took almost two years to complete, which explains the small sample size. No confidence intervals were calculated between comparisons; we acknowledge this as a limitation of the current statistical approach and consider it a priority for a follow-up study with a larger sample, where effect sizes and their corresponding confidence intervals could be estimated more reliably. Another limitation is the fact that not all participants performed the two stages of the rehabilitation protocol, which may hinder the obtained results. Therefore, studies with larger samples should be carried out to confirm these results.

REFERENCES

[1] Hachinski, V., Donnan, G. A., Gorelick, P. B., Hacke, W., Cramer, S. C., Kaste, M., Fisher, M., Brainin, M., Buchan, A. M., and Lo, E. H. Stroke: working toward a prioritized world agenda. *Stroke*, 41(6):1084–1099, 2010.

[2] Kuo, C.-L. and Hu, G.-C. Post-stroke spasticity: a review of epidemiology, pathophysiology, and treatments. *International Journal of Gerontology*, 12(4):280–284, 2018.

[3] Bandela, S., McPherson, L., Harvey, R. L., Awosika, O., Aggarwal, D., Liu, C. Y., Raghavan, P., and Goldberg, M. P. Early recognition and intervention for poststroke spasticity: a scientific statement from the American Heart Association. *Stroke*, 57(4):e146–e159, 2026.

[4] Feigin, V., Brainin, M., Norrving, B., Martins, S., Pandian, J., Lindsay, M. P., Grupper, M., and Rautalin, I. World Stroke Organization (WSO): Global stroke fact sheet 2025. *International Journal of Stroke*, 20:17474930241308142, 2024.

[5] Stevens, J. and Lee, R. The potential to reduce falls and avert costs by clinically managing fall risk. *American Journal of Preventive Medicine*, 55:290–297, 2018.

[6] Ruas, C. V., Carlos, B. M., Feitosa, S., Silva, M. V., Vazquez, P., Pontes, L. L., Silvestre, J., Almeida, S. R. M., Brandão, A. F., and Castellano, G. The effects of transcranial direct current stimulation during extended reality exercises for cortical, neuromuscular, and clinical recovery of stroke survivors. *Neural Plasticity*, 2025:5688648, 2025.

[7] Carlos, B., Vazquez, P., Menezes, L., Feitosa, S., Silva, M., Almeida, S., Silvestre, J., Pontes, L., Brandão, A., Ruas, C., and Castellano, G. Effects in brain symmetry and functional connectivity of tDCS combined with XR in chronic stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, PP:1–1, 2025.

[8] Kang, N., Summers, J., and Cauraugh, J. Transcranial direct current stimulation facilitates motor learning post-stroke: A systematic review and meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 87:345–355, 2016.

[9] ten Donkelaar, H., Broman, J., and van Domburg, P. The somatosensory system. In *Clinical neuroanatomy: brain circuitry and its disorders*, pages 171–255. Springer, 2020.

[10] Del Felice, A., Daloli, V., Masiero, S., and Manganotti, P. Contralesional cathodal tDCS versus dual-tDCS for decreasing upper limb spasticity in chronic stroke individuals: A clinical and neurophysiological study. *Clinical Neurophysiology*, 127:e241–e242, 2016.

[11] Leo, A., Naro, A., Molonia, F., Tomasello, P., Saccà, I., Bramanti, A., Russo, M., Bramanti, P., Quartarone, A., and Calabrò, R. S. Spasticity management: The current state of transcranial neuromodulation. *PM&R*, 9(10):1020–1029, 2017.

[12] Gunduz, A., Kumru, H., and Pascual-Leone, A. Outcomes in spasticity after repetitive transcranial magnetic and transcranial direct current stimulations. *Neural Regeneration Research*, 9(7):712, 2014.

[13] Jung, P., Baumgärtner, U., Bauermann, T., Magerl, W., Gawehn, J., Stoeter, P., and Treede, R.-D. Asymmetry in the human primary somatosensory cortex and handedness. *NeuroImage*, 19:913–923, 2003.

[14] Klomjai, W., Prathum, T., Hoisan, S., Jittipanyakun, N., and Aneksan, B. Effect of bi-hemispheric tDCS combined with physical therapy on interhemispheric inhibition and motor performance in sub-acute to chronic stroke. *Brain Stimulation*, 18, 2025.

[15] Yuan, X., Yang, Y., Cao, N., and Jiang, C. Promotion of poststroke motor-function recovery with repetitive transcranial magnetic stimulation by regulating the interhemispheric imbalance. *Brain Sciences*, 10(9):648, 2020.

[16] Agius Anastasi, A., Falzon, O., Camilleri, K., Vella, M., and Muscat, R. Brain symmetry index in healthy and stroke patients for assessment and prognosis. *Stroke Research and Treatment*, 2017:8276136, 2017.

[17] Sun, X., Dai, C., Wu, X., Han, T., Li, Q., Lu, Y., Liu, X., and Yuan, H. Current implications of EEG and fNIRS as functional neuroimaging techniques for motor recovery after stroke. *Medical Review*, 4(6):480–504, 2024.

[18] Williamson, J. N., James, S. A., Mulyana, B., Kim, S., He, D., Li, S., Sidorov, E. V., and Yang, Y. Quantitative EEG metrics for determining HD-tDCS induced alteration of brain activity in stroke rehabilitation. *Restorative Neurology and Neuroscience*, 42(3-4):209–221, 2025.

[19] Carlos, B. M., De Vazquez, P. F. G., Menezes, L. T., Feitosa, S. S., Da Silva, M. V. N., Almeida, S. R. M., Silvestre, J. A. F., Pontes, L. L., Brandao, A. F., Ruas, C. V., and Castellano, G. Effects in brain symmetry and functional connectivity of tDCS combined with XR in chronic stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 33, 2025.

[20] Hu, M., Cheng, H.-J., Ji, F., Chong, J. S. X., Lu, Z., Huang, W., Ang, K. K., Phua, K. S., Chuang, K.-H., Jiang, X., Chew, E., Guan, C., and Zhou, J. H. Brain functional changes in stroke following rehabilitation using brain-computer interface-assisted motor imagery with and without tDCS: a pilot study. *Frontiers in Human Neuroscience*, 15:692304, 2021.

[21] Brandão, A. F., Dias, D. R. C., Guimarães, M. P., Trevelin, L. C., Parizotto, N. A., and Castellano, G. GestureCollection for motor and cognitive stimuli: virtual reality and e-health prospects. *Journal of Health Informatics*, 10(1), 2018.

[22] Rodrigues, L. G. S., Dias, D. R. C., de Paiva Guimarães, M., Brandão, A. F., Rocha, L. C., Iope, R. L., and Brega, J. R. F. Supervised classification of motor-rehabilitation body movements with RGB cameras and pose tracking data. *Journal on Interactive Systems*, 13(1):221–231, 2022.

[23] van Putten, M. J. A. M. The revised brain symmetry index. *Clinical Neurophysiology*, 118(11):2362–2367, 2007.

[24] Monachino, A. D., Lopez, K. L., Pierce, L. J., and Gabard-Durnam, L. J. The HAPPE plus Event-Related (HAPPE+ER) software: A standardized preprocessing pipeline for event-related potential analyses. *Developmental Cognitive Neuroscience*, 57:101140, 2022.

[25] Liu, M., Xu, G., Yu, H., Wang, C., Sun, C., and Guo, L. Effects of transcranial direct current stimulation on EEG power and brain functional network in stroke patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:335–345, 2023.

[26] Williamson, J. N., James, S. A., Mulyana, B., Kim, S., He, D., Li, S., Sidorov, E. V., and Yang, Y. Quantitative EEG metrics for determining HD-tDCS induced alteration of brain activity in stroke rehabilitation. *Restorative Neurology and Neuroscience*, 42(3-4):209–221, 2025.

[27] Vimolratana, O., Aneksan, B., Siripornpanich, V., Hiengkaew, V., Prathum, T., Jeungprasopsuk, W., Khaokhiew, T., Vachalathiti, R., and Klomjai, W. Effects of anodal tDCS on resting state EEG power and motor function in acute stroke: a randomized controlled trial. *Journal of NeuroEngineering and Rehabilitation*, 21:6, 2024.

[28] Di Lazzaro, V., Dileone, M., Capone, F., Pellegrino, G., Ranieri, F., Musumeci, G., and others. Immediate and late modulation of interhemispheric imbalance with bilateral transcranial direct current stimulation in acute stroke. *Brain Stimulation*, 7:841–848, 2014.

[29] Goodwill, A. M., Teo, W.-P., Morgan, P., Daly, R. M., and Kidgell, D. J. Bihemispheric-tDCS and upper limb rehabilitation improves retention of motor function in chronic stroke: a pilot study. *Frontiers in Human Neuroscience*, 10:258, 2016.

[30] Meng, H., Houston, M., Zhang, Y., & Li, S. (2024). Exploring the prospects of transcranial electrical stimulation (tES) as a therapeutic intervention for post-stroke motor recovery: A narrative review. *Brain Sciences*, *14*(4), 322.

[31] Murase, N., Duque, J., Mazzocchio, R., & Cohen, L. G. (2004). Influence of interhemispheric interactions on motor function in chronic stroke. *Annals of Neurology*, *55*(3), 400–409.

[32] Ghosh, S., Hathorn, D., Eisenhauer, J., Dixon, J., and Cooper, I. D. Anodal transcranial direct current stimulation over the vertex enhances leg motor cortex excitability bilaterally. *Brain Sciences*, 9(5):98, 2019.

[33] Tazoe, T., Endoh, T., Kitamura, T., and Ogata, T. Polarity specific effects of transcranial direct current stimulation on interhemispheric inhibition. *PLOS ONE*, 9(12):e114244, 2014.

[34] Casula, E. P., Pellicciari, M. C., Bonnì, S., Spanò, B., Ponzo, V., Salsano, I., Giulietti, G., Martino Cinnera, A., Maiella, M., Borghi, I., Rocchi, L., Bozzali, M., Sallustio, F., Caltagirone, C., and Koch, G. Evidence for interhemispheric imbalance in stroke patients as revealed by combining transcranial magnetistimulation and electroencephalography. *Human Brain Mapping*, 42(5):1343–1358, 2021.

Integrated Computer-Aided Diagnosis System for Brain Tumor Segmentation and Classification

1st Andressa G. Moreira
Department of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
andressagomes@alu.ufc.br

2nd Bruno R. dos S. Silva
Department of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
bruno@lesc.ufc.br

3rd Paulo C. Cortez
Department of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
cortez@lesc.ufc.br

***Abstract*—Brain tumors represent neoplasms associated with high global mortality rates. In this context, manual analysis of Magnetic Resonance Imaging (MRI) scans for brain tumor diagnosis is complex and error-prone due to high tumor heterogeneity. To overcome these limitations, we propose a Computer-Aided Diagnosis (CAD) system that integrates automatic lesion segmentation and multiclass classification, comparing deep learning models with radiomics-based classifiers. In segmentation, the EfficientNetB7 FPN architecture achieved an F1-Score of 85.68%. In classification, the Swin Transformer-B model achieved the best overall performance, with an F1-Score of 93.52%, while LightGBM stood out among the radiomics-based approaches with an F1-Score of 87.99%. In summary, the system demonstrated efficacy and robustness, establishing itself as an important clinical decision-support tool.**

***Index Terms*—brain tumor, magnetic resonance imaging, segmentation, classification, radiomic features.**

I. Introduction

Brain tumors arise from abnormal cell proliferation within brain tissue [1]. These neoplasms represent some of the most aggressive malignancies in the central nervous system and contribute substantially to cancer-related mortality [2]. The World Health Organization (WHO) reports that this cancer type is among the most lethal and can affect individuals globally [3].

Magnetic Resonance Imaging (MRI) is regarded as a primary imaging modality for the detection and monitoring of brain tumors, due to its non-invasive nature and ability to produce high-resolution images [3]–[5]. These features facilitate the identification of brain abnormalities, enhance the accuracy of early diagnosis, and improve patient survival rates [6].

The heterogeneous nature of tumors renders manual analysis of MRI scans a complex task that depends significantly on the expertise of specialist radiologists [6]–[8]. As a result, lesion analysis is particularly challenging when large volumes of examinations and substantial variability in tumor characteristics are present [3]. In this context, deep learning-based techniques demonstrate considerable effectiveness in medical image classification and segmentation [6].

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 315447/2023-0 and Coordination of Superior Level Staff Improvement (CAPES)-Finance Code 001 for the support to this research.

In this work, we aim to develop a Computer-Aided Diagnosis (CAD) system for segmentation, radiomic feature extraction, and multiclass classification of brain tumors, including meningioma, glioma, and pituitary tumors. The proposed tool does not aim to replace the specialist, but rather to act as clinical decision support, aiding in improved clinical outcomes, i.e., faster and more accurate diagnoses.

The experiments were conducted in two stages: (i) tumor region segmentation and (ii) tumor classification. Quantitative biomarkers were extracted via radiomics from the segmented regions and used to train traditional machine learning (ML) classifiers. A comparative analysis was then performed with deep learning models, including convolutional neural networks, Vision Transformers, and Foundation Models.

The main contributions of this work are:

- (i) the development of a CAD system for brain tumor segmentation and classification;
- (ii) a comparative analysis of approaches based on MRI scans and radiomic features;
- (iii) the application of non-parametric tests to statistically compare the performance of the architectures;
- (iv) validation of experiments on external datasets to ensure model reproducibility.

The structure of this paper is as follows: Section II reviews the principal related works in the literature. Section III outlines the methodological steps employed. Section IV discusses the experimental results. Section V provides the concluding remarks.

II. Related Works

This section synthesizes recent advances in artificial intelligence for brain tumor segmentation and classification in MRI, identifying current methodologies and gaps in the literature. To consolidate the review, related studies are compared with the proposed approach in Table I, detailing criteria such as datasets, preprocessing, architectures, external validation, statistical tests, and results.

In the field of deep learning, Ullah et al. (2026) [9] introduced a framework that combines the Bayesian DeepLabV3+ architecture for segmentation with a hybrid deep network for classification. Similarly, Kumar and Singh (2026) [10]

TABLE I: Summary of related works.

Paper	Dataset	Preproc.	Seg.	Rad.	Class.	Ext. Val.	Stat. Test	Seg. Result	Class. Result
Ullah et al. (2026) [9]	BraTS23, Figshare	Z-score, CLAHE, Smoothing	Bayesian DeepLabV3+	No	Hybrid-Depth DCNN	No	One-way ANOVA	Dice: 95.86%	Acc: 98.8%
Kumar & Singh (2026) [10]	Figshare, Kaggle	IRoLF	MDU-Net	No	SwinCT + HCAC	Yes	Ablation	Dice: 97.0%	Acc: 99.9%
Hussain et al. (2025) [11]	Figshare, PGIMER	Hist. Equaliz.	Res-Attn-UNet	No	CNN	Pilot	No	Dice: 91.10%	Acc: 99.4%
Bhattacharjee et al. (2022) [12]	Figshare	Mask overlay	No	Random Forest	BiLSTM	No	No	N/A	Binary Acc: 97.7%
Souza et al. [13]	Kaggle	Skull stripping	Thresholding	XGBoost	DNN	No	No	N/A	Acc: 88.51%
This Work	**Figshare, MRI**	**CLAHE**	**EfficientNetB7 FPN**	**LightGBM**	**Swin-B**	**Yes**	**Wilcoxon**	**Acc: 99.53%, F1: 85.68%**	**Acc: 93.55%, F1: 93.52%**

developed an explainable architecture that integrates MDU-Net and SwinCT, while Hussain et al. (2025) [11] implemented a U-Net variant with residual attention to address both tasks concurrently. Although these methods demonstrate effectiveness in hierarchical learning, they primarily rely on local convolution operations and do not facilitate the joint extraction of quantitative radiomic features.

In the context of radiomic strategies, Bhattacharjee et al. (2022) [12] integrated statistical descriptors with the Random Forest algorithm for classification, resulting in an accuracy of 84.0%. Similarly, Souza et al. (2021) [13] combined texture and first-order features with classifiers such as XGBoost and dense networks, achieving an average accuracy of 88.51%. However, these approaches encounter significant operational limitations because they depend on manual annotations by radiologists or inflexible thresholding techniques.

To address these challenges, we propose a CAD system that fully automates the diagnostic workflow. The segmentation stage utilizes the EfficientNetB7 FPN semantic network. For multiclass classification, we employ Vision Transformers (Swin Transformer-B) and Foundation Models (CLIP and DINOv2), resulting in an accuracy of 93.55%. We conducted a direct comparison between deep learning and radiomic approaches (using LightGBM), with robustness and reproducibility ensured through validation on external datasets and statistical confirmation using the Wilcoxon test.

III. Methodology

In this section, we outline the methodology employed in this work. We provide details on the datasets, preprocessing, and data augmentation procedures. Additionally, we describe the segmentation processes, biomarker extraction, classification methods, model evaluation, and the statistical tests applied.

A. Experimental Approaches

We propose the development of a CAD system designed to segment and classify brain tumors as meningioma, glioma, or pituitary. The workflow stages are illustrated in Figure 1.

Segmentation models were trained using MRI scans from the Figshare dataset [14], which contains brain tumor examinations and corresponding tumor masks. The trained model was subsequently applied to images from the Brain Tumor Classification dataset [15] to automatically segment tumor regions. The resulting regions of interest (ROIs) obtained through segmentation served as inputs for the subsequent classification stages.

Radiomic features, including intensity, texture, shape, and spatial distribution, were extracted from the segmented regions to train traditional classifiers. The performance of this approach was compared with that of deep learning models applied directly to the images to evaluate the effectiveness of both strategies.

B. Datasets

For the tumor segmentation task, the Figshare dataset [14] was utilized, comprising 3,064 images. Figure 2 presents sample MRI scans and their corresponding tumor region segmentation masks. The images are distributed across three classes: meningioma (708 images), glioma (1,426 images), and pituitary tumor (930 images).

We applied the trained segmentation model to the public Brain Tumor Classification (MRI) dataset [15] to automatically segment tumoral regions. Of the 3,264 MRI images, we selected the 2,764 with confirmed tumors (meningioma: 937, glioma: 926, pituitary: 901), excluding the 500 no-tumor cases. This selection supports our goal of histological differentiation among tumor types, which requires feature extraction from a defined ROI. Figure 3 shows representative samples for each class.

For external validation of the classification module, we utilized the public Figshare dataset. While this dataset was previously employed during the training stage of the segmentation models, its use in classification was strictly limited to an unseen test set. Independence between stages was maintained at both the patient and examination levels, ensuring that Figshare images were used exclusively for feature extraction and class prediction by the pre-trained classifier. This approach

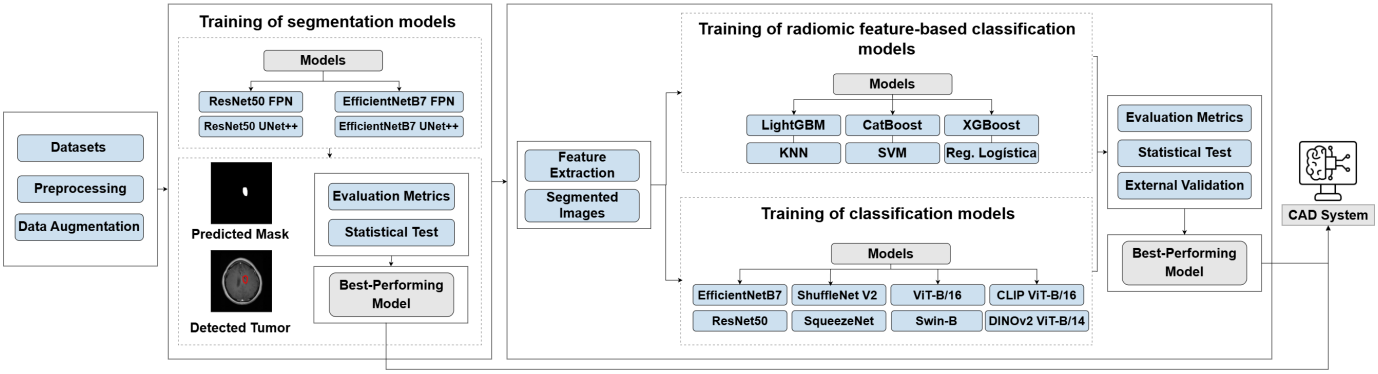
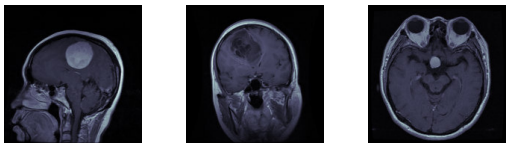


Fig. 1: Overview of the primary phases implemented in this study.

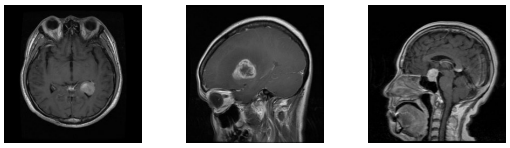


(a) Meningioma. (b) Glioma. (c) Pituitary.



(d) Meningioma. (e) Glioma. (f) Pituitary.

Fig. 2: Samples from the Figshare dataset.



(a) Meningioma. (b) Glioma. (c) Pituitary.

Fig. 3: Samples from the Brain Tumor Classification dataset.

prevented data leakage and avoided artificial inflation of performance metrics.

C. Preprocessing

We performed preprocessing to improve the visual quality of the scans and minimize unwanted artifacts, which facilitated more efficient feature extraction by computational models. Initially, we converted the images from the datasets to grayscale, resized them to 224 x 224 pixels, and stored them in PNG format.

Next, we applied a bilateral filter to smooth the images, reduce noise, and preserve the sharpness of anatomical edges. Additionally, we employed Contrast Limited Adaptive Histogram Equalization (CLAHE) for local adaptive contrast enhancement, which improved the visibility of structures of interest within the images. Finally, we applied simple thresholding to the tumor segmentation masks to ensure accurate binarization of the segmented regions.

D. Data Augmentation

To increase dataset diversity, reduce overfitting, and improve model generalization, data augmentation techniques based on spatial and radiometric transformations were implemented. These transformations were applied exclusively to the training set, thereby maintaining the integrity of the validation and test subsets and ensuring unbiased evaluation.

Affine transformations were applied with a probability of $p = 0.5$, including brightness and contrast adjustments (limits of 0.3), horizontal flipping, and scale, translation, and rotation shifts (parameters: 0.1 for scale, 0.05 for translation, and $15°$ for rotation).

E. Segmentation

We performed brain tumor segmentation in MRI scans using architectures grounded in the encoder-decoder paradigm. The encoder transforms the input image into hierarchical feature representations at multiple resolution levels. The decoder then projects these features back into the pixel space, facilitating dense classification to delineate tumor structures.

In this work, we evaluated combinations of different models. For the encoder, we used pre-trained DenseNet201, EfficientNetB7, and ResNet50 networks. For the decoder, we implemented both FPN and UNet++ architectures. During training, each model received an MRI scan and its corresponding ground-truth mask as input, producing a predicted mask indicating the tumor location as output.

We trained the models for 50 epochs and applied early stopping with a patience of 5 epochs to mitigate overfitting. The experimental hyperparameters included a batch size of 4, a Lovász-Softmax-based loss function, the Adam optimizer, and a learning rate of 0.0005.

Finally, we used the segmentation model as an automated module to identify and delineate tumor regions in MRI scans. Each scan was processed by the segmentation algorithm to extract tumor regions, and the resulting regions of interest (ROIs) were forwarded to subsequent classification stages.

F. Radiomic Feature-Based Classification

Radiomic features were extracted from the segmented region. This approach involves the automatic acquisition of

a large set of high-throughput features from medical images, which describe lesion morphology, gray-level intensity distribution, and complex texture patterns to mathematically represent tissue heterogeneity. The open-source *PyRadiomics* package[1] was used to extract these features from the MRI scans, thereby structuring the feature vector for model training.

To address the curse of dimensionality, we performed feature selection exclusively on the training set to prevent data leakage. After excluding operational metadata, we applied a low-variance filter to remove features with minimal variability, followed by a Spearman high-correlation filter to eliminate redundant, collinear attributes prior to model fitting.

Following the filtering process, an optimized vector comprising nine fundamental radiomic features was selected for classification. These descriptors summarize brain tumor heterogeneity across three categories: first-order intensity statistics (skewness, kurtosis, energy, and minimum intensity), lesion morphology (geometric dimensions), and texture (gray-level patterns via higher-order matrices). Collectively, these features provide a quantitative representation of the morphological and textural heterogeneity of brain tumors.

The selected features were used as input variables to train and benchmark six traditional machine learning algorithms, LightGBM, Support Vector Machines (SVM), XGBoost, K-Nearest Neighbors (KNN), CatBoost, and Logistic Regression. The models were trained exclusively with the optimized feature vector, enabling evaluation of the predictive potential and performance of each architecture in the brain tumor classification task.

G. Grid Search

We optimized deep learning models for brain tumor classification using grid search. Hyperparameter tuning was conducted only on the training partitions of each fold, keeping test data fully isolated. The search covered all combinations of learning rates (0.0001, 0.0003, 0.001), optimizers (Adam, RMSprop, SGD), batch sizes (16, 32, 64), and weight decay values (0.0001, 0.00001).

H. Multiclass Classification

To recognize different types of brain tumors, the images segmented in the previous stage, containing only the regions of interest delineated by the tumor segmentation models, were utilized. These segmented images served as input for training classification models, enabling a comparative analysis of various deep learning paradigms.

We evaluated eight pre-trained architectures, including traditional convolutional neural networks (ResNet50 and EfficientNetB7), lightweight convolutional architectures (ShuffleNet V2 and SqueezeNet), Vision Transformer-based models (ViT-B/16 and Swin Transformer-B), and foundation models (CLIP ViT-B/16 and DINOv2 ViT-B/14). The selection of these architectures facilitated the investigation of performance and generalization capability across diverse approaches.

[1]https://pyradiomics.readthedocs.io/

The models were trained for 50 epochs, corresponding to the number of iterations over the training set. Cross-entropy loss, which is commonly employed in multiclass classification problems, was used as the loss function. An early stopping strategy with a patience parameter of 5 was implemented, halting training when no performance improvement was observed for five consecutive epochs.

The training hyperparameters were determined using a grid search strategy, which enabled the selection of the most suitable configuration for each evaluated model. Table II presents the value configurations considered for each hyperparameter.

TABLE II: Defined hyperparameters for the classification models.

Architecture	LR	Optimizer	Batch Size	WD
ResNet50	0.0001	Adam	32	0.0001
EfficientNetB7	0.0003	Adam	32	0.0001
ShuffleNet V2	0.0001	RMSprop	32	0.00001
SqueezeNet	0.0003	Adam	32	0.00001
ViT-B/16	0.0001	Adam	32	0.0001
Swin Transformer-B	0.0001	Adam	32	0.00001
CLIP ViT-B/16	0.0001	Adam	32	0.00001
DINOv2 ViT-B/14	0.0001	Adam	32	0.00001

I. Cross-Validation

We structured the training and evaluation process using a k-fold cross-validation protocol ($k = 10$) with three repetitions. In each fold, 80% of the data was allocated for training and 20% for testing. From the training portion of each fold, 10% of the samples were reserved for validation during the iterative learning process. Model performance was determined by calculating the arithmetic mean of the metrics obtained across the 30 repetitions.

J. Evaluation Metrics

Quantitative assessment of model performance necessitates the selection of metrics that align with the specific objectives of the project. In medical image semantic segmentation, True Positives (TP) and True Negatives (TN) denote correctly classified tumor and background pixels, respectively. In contrast, False Positives (FP) arise when background pixels are misclassified as tumor, whereas False Negatives (FN) correspond to tumor pixels that the model fails to identify.

The performance of the segmentation architectures was formally evaluated using metrics based on geometric similarity and area overlap. We quantified the behavior of the models using the following equations:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{1}$$

$$\text{F}_1\text{-score} = \frac{2 \times \text{Prec} \times \text{Recall}}{\text{Prec} + \text{Recall}}, \tag{2}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{3}$$

$$\text{Hausdorff} = \max(h(A, B), h(B, A)). \tag{4}$$

Accuracy (1) quantifies the overall proportion of correct predictions relative to the total number of analyzed pixels. The F1-score (2) serves as the primary validation index by balancing precision and recall. Intersection over Union (IoU) (3) measures the overlap between the predicted region and the ground truth. The Hausdorff Distance (4) is a boundary-based metric that quantifies the maximum distance between points on the ground-truth and estimated contours.

In image classification, the confusion matrix is used to evaluate model predictions for each class. TP denotes correctly identified samples of a given category, while FP indicates cases from other classes that are incorrectly assigned to the target class. Similarly, TN corresponds to samples correctly identified as not belonging to the target class, and FN refers to images of the target category that are incorrectly classified.

Classifier performance in the brain tumor scenario was evaluated using accuracy, precision, recall, F1-score, and specificity, defined as follows:

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{7}$$

Accuracy (1) indicates the overall correct classification rate of the model. Precision (5) measures the reliability of positive predictions by reducing false alarms. Recall (6) reflects the classifier's ability to identify all true positive cases. The F1-score (2) represents the harmonic mean between precision and recall, offering a robust indicator for class-imbalanced scenarios. Specificity (7) assesses the model's ability to correctly identify negative samples.

K. Statistical Tests

To assess the statistical significance of performance differences among the proposed models, an inferential analysis was performed using the non-parametric Wilcoxon test. This method was selected because it does not require normality assumptions, making it appropriate for comparing models in experimental scenarios with unknown or non-Gaussian distributions.

Statistical analyses were conducted individually for each metric obtained via cross-validation, resulting in 30 paired observations per architecture for hypothesis evaluation. For all tests, a two-tailed significance level of $\alpha = 0.05$ was used. The null hypothesis (H_0) posits no significant differences between the compared models, while the alternative hypothesis (H_1) indicates the presence of statistically significant differences.

L. External Validation

To assess the robustness and generalization capability of the proposed classifiers, an external validation stage was implemented using the *Figshare* dataset. The images were strictly isolated throughout all feature selection, model training, and internal testing stages to mitigate any risk of data leakage or optimization bias. This methodological approach is essential for evaluating the effectiveness of the algorithms when exposed to the clinical variability inherent in real-world application scenarios.

M. Development of the CAD System

The final step involved implementing a CAD system within a web application developed using *Streamlit*, which integrates the most effective segmentation and classification pipelines. The interface enables users to upload brain MRI scans, after which the system automatically processes the data and displays both the segmented tumor region and the multiclass histological diagnosis. This solution serves as a clinical decision-support tool to assist healthcare professionals in diagnosis and patient monitoring.

IV. Results

In this section, we present and discuss the results obtained using the brain tumor segmentation and classification methodology, with particular emphasis on the model performance and the effectiveness of the proposed approach.

A. Performance of Deep Learning Models in Segmentation

Quantitative evaluation is essential for assessing the effectiveness and reliability of brain tumor segmentation models. In this work, multiple metrics were used to evaluate the performance of the examined architectures, enabling a comprehensive analysis of segmentation quality.

Table III summarizes the results for the evaluated architectures. Alongside mean values, 95% confidence intervals (95% CI) are reported to estimate the range in which the true model performance is likely to fall. Narrower confidence intervals reflect greater stability and lower variability across runs, while wider intervals indicate increased sensitivity to data partitioning.

The EfficientNetB7 FPN architecture demonstrated the highest overall performance among the evaluated models, achieving a mean accuracy of 99.53%, an F1-score of 85.68%, an IoU of 74.97%, and a Hausdorff Distance of 7.68. The consistently narrow confidence intervals across all metrics indicate robust results throughout the validation folds.

Comparable results were obtained for the EfficientNetB7 UNet++ architecture, which achieved an accuracy of 99.50%, an F1-score of 84.73%, an IoU of 73.52%, and a Hausdorff Distance of 7.93. The DenseNet201 UNet++ architecture also demonstrated competitive performance, with an F1-score of 82.57%, an IoU of 71.01%, and a Hausdorff Distance of 8.31. Although the mean values for these architectures are similar to those of EfficientNetB7 FPN, the slightly wider confidence intervals indicate increased variability across runs.

In contrast, the DenseNet201-FPN and ResNet50-FPN models achieved accuracies of approximately 97.4%; however, this metric is inflated by class imbalance, as background pixels constitute the majority of each image. Their F1-scores were 39.33% and 34.72%, respectively, with IoU values below 25% and Hausdorff Distances of 49.69 and 55.17. These findings

TABLE III: Mean results of the segmentation models.

Metric	DenseNet201 FPN	DenseNet201 UNet++	EfficientNetB7 FPN	EfficientNetB7 UNet++	ResNet50 FPN	ResNet50 UNet++
Accuracy (%)	97.37 ± 0.85	99.45 ± 0.16	$\mathbf{99.53 \pm 0.04}$	99.50 ± 0.04	97.39 ± 0.53	99.35 ± 0.27
F1-Score (%)	39.33 ± 4.44	82.57 ± 8.94	$\mathbf{85.68 \pm 1.09}$	84.73 ± 1.25	34.72 ± 4.04	77.12 ± 18.19
IoU (%)	24.57 ± 3.47	71.01 ± 9.47	$\mathbf{74.97 \pm 1.67}$	73.52 ± 1.88	21.08 ± 2.99	65.18 ± 17.09
Hausdorff Distance	49.69 ± 8.49	8.31 ± 0.75	$\mathbf{7.68 \pm 0.61}$	7.93 ± 0.72	55.17 ± 6.66	14.40 ± 17.68

indicate that, despite high accuracy, these models have limited capability to delineate tumors.

B. Statistical Analysis of Segmentation Performance Using the Wilcoxon Test

The statistical significance of the segmentation models was assessed using the Wilcoxon test. For the accuracy metric, only the comparison between DenseNet201 FPN and ResNet50 FPN did not demonstrate a statistically significant difference ($p = 0.73$), while all other comparisons were significant. Additionally, for the F1-score, IoU, and Hausdorff Distance metrics, all comparisons yielded $p < 0.05$, indicating distinct performances among the architectures in terms of tumor segmentation capability.

The EfficientNetB7 FPN architecture achieved the highest results for F1-score, IoU, and Hausdorff Distance. The Wilcoxon test confirms that these improvements are statistically significant compared to the other architectures. In contrast, the DenseNet201 FPN and ResNet50 FPN models demonstrated significantly lower segmentation performance, supporting the findings from the previous descriptive analysis.

C. Qualitative Analysis of Segmentation Results

Beyond quantitative metrics, visual analysis is crucial for assessing the models' capacity to delineate tumor regions. Figure 4 displays segmentation results generated by the EfficientNetB7 FPN architecture, which demonstrated the highest overall performance in this work. The first column shows the original image, while the second and third columns present the ground-truth mask and the model-predicted mask, respectively. The fourth and fifth columns depict the original MRI scans with their corresponding segmentations overlaid.

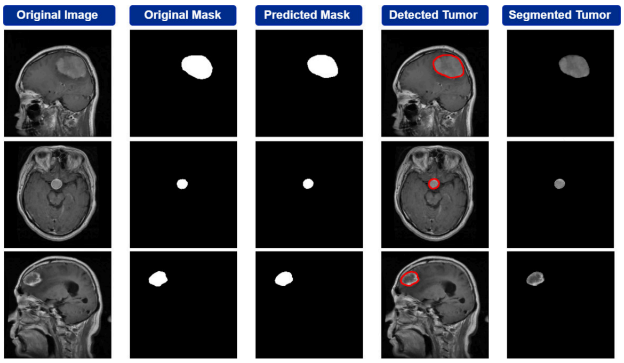


Fig. 4: Segmentation results on MRI scans produced by the EfficientNetB7 FPN architecture are shown.

The visual agreement between the predicted and ground-truth masks supports the quantitative findings. The close alignment of contours demonstrates the model's capacity to accurately identify and delineate tumor boundaries, even in cases with significant heterogeneity in shape, size, and contrast. Furthermore, the network exhibited robustness in detecting small-scale lesions, which is essential for clinical practice due to its influence on early diagnosis.

D. Performance of traditional ML algorithms in radiomic feature classification

In this stage, classification algorithms were trained using texture, shape, and intensity features extracted with the *Pyradiomics* library. The comparative analysis assessed accuracy, F1-score, precision, recall, specificity, and the mean area under the receiver operating characteristic (ROC) curve (AUC) across the validation folds. Table IV provides a summary of the results for each classifier and reports the corresponding 95% confidence intervals (95% CI).

The analysis indicates the relative performance of the evaluated architectures in differentiating brain tumors. Among the tested models, LightGBM achieved the highest mean F1-score (87.99%), as well as high precision (88.06%), recall (88.03%), and accuracy (88.00%). LightGBM also demonstrated the greatest overall discriminatory capability, with a mean AUC of 97.14% and a specificity of 94.00%.

The XGBoost and SVM models demonstrated intermediate and comparable performance, with mean F1-scores of 86.44% and 86.41%, respectively. The KNN classifier showed lower performance, achieving a mean F1-score of 85.22% and an AUC of 94.82%. CatBoost and Logistic Regression exhibited the lowest overall performance. CatBoost achieved a mean F1-score of 84.29% and an AUC of 95.68%, while Logistic Regression recorded the F1-score (83.56%) and AUC (94.70%).

Although the pipeline based on radiomic features and traditional classifiers produced promising results, further investigation of deep learning approaches is necessary. The following section evaluates the performance of deep learning architectures applied directly to the segmented images.

E. Performance of Deep Learning Models in Classification

Table V presents the quantitative performance of deep learning models trained exclusively on segmented images for multiclass classification, along with the corresponding 95% confidence intervals. Swin Transformer-B demonstrated the highest overall performance, achieving an accuracy of 93.55%, an F1-score of 93.52%, a specificity of 96.79%, and a mean

TABLE IV: Mean results of the classifiers based on radiomic features.

Metric	LightGBM	SVM	XGBoost	KNN	CatBoost	Logistic Regression
Accuracy (%)	$\mathbf{88.00 \pm 1.55}$	86.47 ± 1.71	86.46 ± 1.74	85.26 ± 1.80	84.32 ± 1.87	83.60 ± 1.66
Precision (%)	$\mathbf{88.06 \pm 1.50}$	86.54 ± 1.72	86.53 ± 1.73	85.48 ± 1.82	84.41 ± 1.89	83.70 ± 1.69
Recall (%)	$\mathbf{88.03 \pm 1.54}$	86.50 ± 1.72	86.49 ± 1.74	85.32 ± 1.80	84.35 ± 1.87	83.63 ± 1.66
F1-Score (%)	$\mathbf{87.99 \pm 1.55}$	86.41 ± 1.74	86.44 ± 1.74	85.22 ± 1.82	84.29 ± 1.87	83.56 ± 1.69
Specificity (%)	$\mathbf{94.00 \pm 0.77}$	93.24 ± 0.86	93.23 ± 0.87	92.65 ± 0.90	92.17 ± 0.93	91.80 ± 0.83
Mean AUC (%)	$\mathbf{97.14 \pm 0.79}$	95.96 ± 1.14	96.54 ± 0.93	94.82 ± 1.27	95.68 ± 0.95	94.70 ± 1.23

AUC of 98.49%. ResNet50 and ViT-B/16 also exhibited strong results, with ResNet50 reaching an accuracy of 92.04% and a mean AUC of 98.31%, and ViT-B/16 attaining an accuracy of 89.64% and a mean AUC of 97.60%.

EfficientNetB7 and SqueezeNet also demonstrated competitive performance, each recording mean AUCs above 97%, although their accuracy values were lower than those of the top-performing models. In contrast, ShuffleNet V2 and DINOv2 ViT-B/14 achieved the lowest accuracies, at 79.97% and 83.96%, respectively, indicating greater difficulty in discriminating among the dataset classes.

A comparison with classifiers based on radiomic features (Table IV) reveals a consistent advantage for the leading deep learning architectures. The best radiomic model, LightGBM, achieved an accuracy of 88.00% and a mean AUC of 97.14%, whereas Swin Transformer-B surpassed these results by approximately 5.6 percentage points in accuracy and 1.35 percentage points in mean AUC.

However, deep learning did not demonstrate absolute superiority in all scenarios. Architectures such as ShuffleNet V2 (79.47%), DINOv2 ViT-B/14 (83.56%), and CLIP ViT-B/16 (85.76%) produced lower F1-scores than LightGBM (87.99%) and XGBoost (86.44%), and were outperformed by classifiers trained exclusively on radiomic features. These findings underscore the importance of model selection and tuning to achieve optimal performance with deep learning compared to traditional machine learning models.

F. Statistical Analysis of Classification via the Wilcoxon Test

To assess the statistical significance of differences among the 14 classifiers (8 deep learning and 6 traditional models), the paired-sample Wilcoxon test ($\alpha = 0.05$) was applied. Among the 91 pairwise comparisons for the F1-score metric, 79 (86.8%) demonstrated statistically significant differences ($p < 0.05$), confirming the decisive impact of model selection on diagnostic accuracy.

The results confirmed the superiority of Swin Transformer-B, which achieved an F1-score significantly higher than all other approaches ($p < 0.001$). When compared to ResNet50, the second-best deep learning model, the difference was statistically significant ($p = 0.000256$). Additionally, in comparison with LightGBM, the best radiomic model, the superiority of Swin Transformer-B was confirmed ($p = 1.86 \times 10^{-9}$).

Conversely, statistical equivalence ($p \geq 0.05$) was observed among models with numerically similar performances. The most evident case occurred between SVM and XGBoost ($p = 0.9354$), a pattern also observed between SqueezeNet and LightGBM ($p = 0.0523$) and between CatBoost and Logistic Regression ($p = 0.0606$).

In summary, the analysis indicated that deep learning is not universally superior. LightGBM (F1-score of 87.99%) exhibited significant differences compared to the deep architectures CLIP ViT-B/16 ($p = 1.22 \times 10^{-2}$), DINOv2 ViT-B/14 ($p = 2.54 \times 10^{-3}$), and ShuffleNet V2 ($p = 1.86 \times 10^{-9}$). These findings demonstrate that, although deep networks generally achieve higher classification performance, lightweight deep models can be outperformed by optimized tree ensembles.

G. External Validation

The generalization capability of the Swin Transformer-B model was evaluated through external validation using the *Figshare* dataset. The pipeline demonstrated robust performance on previously unseen data, achieving an F1-score of 88.62%, a specificity of 94.59%, and a mean AUC of 98.65%. Table VI presents the results for each class.

Glioma demonstrated the highest precision (97.70%), which reduced the occurrence of false positives in this aggressive tumor type. Pituitary tumors achieved the highest sensitivity (96.13%), indicating effectiveness as a screening mechanism by minimizing missed lesions. Meningioma exhibited the most balanced performance, with an F1-score of 89.06%. The strong performance on unseen data further demonstrates the generalization capability of Swin Transformer-B and supports its feasibility for real-world clinical applications.

H. Computer-Aided Diagnosis System

A functional CAD system was developed, as shown in Figure 5. When an MRI scan is uploaded, the platform automatically processes the image and generates a segmentation that highlights the precise location of the tumor lesion, along with its corresponding classification. This unified interface integrates the segmentation (EfficientNetB7 FPN) and classification (Swin Transformer-B) stages into a streamlined workflow, offering practical and visual decision support.

V. Conclusion

This work describes an integrated CAD research prototype for MRI brain tumor segmentation and multiclass classification. While promising for preliminary decision support, clinical implementation requires prospective validation, multicenter studies with diverse imaging protocols, and formal evaluation by radiologists.

During the segmentation stage, the EfficientNetB7 FPN architecture achieved an F1-score of 85.68%, indicating high precision and stability despite lesion heterogeneity. In the

TABLE V: Mean results of deep learning-based classifiers.

Metric	ResNet50	EfficientNetB7	ShuffleNet V2	SqueezeNet	ViT-B/16	Swin-B	CLIP ViT-B/16	DINOv2 ViT-B/14
Accuracy (%)	92.04 ± 2.15	89.40 ± 6.45	79.97 ± 6.14	88.92 ± 2.48	89.64 ± 2.08	$\mathbf{93.55 \pm 1.30}$	85.88 ± 2.75	83.96 ± 5.42
Precision (%)	92.20 ± 2.00	89.64 ± 5.95	81.86 ± 4.49	89.35 ± 2.08	89.87 ± 2.06	$\mathbf{93.61 \pm 1.22}$	86.45 ± 2.49	84.47 ± 6.24
Recall (%)	92.09 ± 2.11	89.43 ± 6.62	80.05 ± 6.08	88.92 ± 2.33	89.60 ± 1.97	$\mathbf{93.55 \pm 1.26}$	85.95 ± 2.72	84.02 ± 5.46
F1-Score (%)	92.01 ± 2.11	89.21 ± 6.97	79.47 ± 6.96	88.84 ± 2.44	89.58 ± 2.06	$\mathbf{93.52 \pm 1.28}$	85.76 ± 2.84	83.56 ± 7.07
Specificity (%)	96.06 ± 1.05	94.81 ± 3.06	90.49 ± 2.56	94.58 ± 1.16	94.88 ± 1.03	$\mathbf{96.79 \pm 0.64}$	93.12 ± 1.27	92.26 ± 2.13
Mean AUC (%)	98.31 ± 0.99	97.08 ± 3.52	94.17 ± 2.78	97.37 ± 1.07	97.60 ± 1.04	$\mathbf{98.49 \pm 0.87}$	96.46 ± 1.26	95.12 ± 4.97

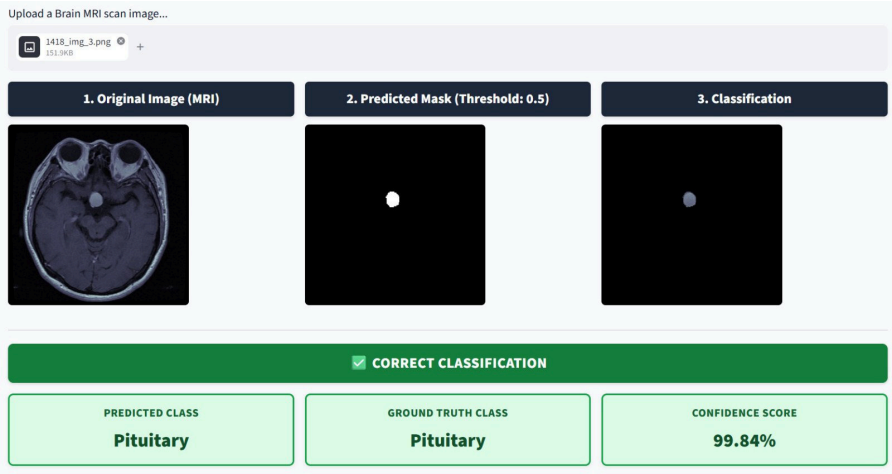


Fig. 5: Web interface of the CAD system designed for automated segmentation and classification of brain tumors.

TABLE VI: External validation results for the Swin Transformer-B model.

Class	Support	Precision (%)	Recall (%)	F1 (%)
Meningioma	708	88.44	89.69	89.06
Glioma	1,426	97.70	83.38	89.97
Pituitary	930	79.19	96.13	86.84
Mean	**3,064**	**88.44**	**89.73**	**88.62**

classification stage, deep learning architectures applied directly to images outperformed traditional methods based on radiomic features. The Swin Transformer-B model demonstrated the highest overall performance with an F1-score of 93.52%, while the LightGBM classifier achieved the best results among radiomic approaches with an F1-score of 87.99%.

The project's practical viability was demonstrated through an interactive web application that automates the diagnostic workflow with the top-performing models. However, the current prototype only classifies previously detected tumors. Future work should investigate multimodal approaches, such as combining MRI with PET scans, to enhance biomarker extraction. Controlled prospective trials will be needed to validate clinical adoption in routine radiological practice.

REFERENCES

[1] A. Suliman Farhan, M. Khalid and U. Manzoor, "Combined Oriented Data Augmentation Method for Brain MRI Images," in IEEE Access, vol. 13, pp. 9981-9994, 2025.

[2] E. G. Van Meir, C. G. Hadjipanayis, A. D. Norden, H. K. Shu, P. Y. Wen, and J. J. Olson, "Exciting new advances in neuro-oncology: The avenue to a cure for malignant glioma," CA, A Cancer J. Clinicians, vol. 60, no. 3, pp. 166–193, May 2010.

[3] S. Rajendran et al., "Automated Segmentation of Brain Tumor MRI Images Using Deep Learning," in IEEE Access, vol. 11, pp. 64758-64768, 2023.

[4] E. U. Haq, H. Jianjun, K. Li, H. U. Haq, and T. Zhang, "An MRI-based deep learning approach for efficient classification of brain tumors," J. Ambient Intell. Humanized Comput., vol. 14, pp. 6697–6718, Jun. 2023.

[5] A. S. Farhan, M. Khalid, and U. Manzoor, "XAI-MRI: An ensemble dual-modality approach for 3D brain tumor segmentation using magnetic resonance imaging," bioRxiv, vol. 2024, p. 11, Nov. 2024.

[6] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier, and A. Droit, "Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification," Nature Commun., vol. 11, no. 1, pp. 1–11, Nov. 2020.

[7] L. Fang and X. Wang, "Brain tumor segmentation based on the dualpath network of multi-modal MRI images," Pattern Recognit., vol. 124, Apr. 2022, Art. no. 108434.

[8] A. Shukla, R. Tiwari, and S. Tiwari, "Alzheimer's disease detection from fused PET and MRI modalities using an ensemble classifier," Mach. Learn. Knowl. Extraction, vol. 5, no. 2, pp. 512–538, May 2023.

[9] M. Sami Ullah et al., "A Bayesian Backed DeepLabV3+ Customized Hybrid-Depth Framework for Brain Tumor Segmentation and Classification," in IEEE Access, vol. 14, pp. 20328-20354, 2026.

[10] Kumar, Vendra Durga Ratna, and Priyanka Singh. "HCAC: Trustworthy Hybrid Attentive Framework Tackling Brain Tumor Identification with XAI." IEEE Access (2026).

[11] Hussain, Syed Sajid, et al. "Next-generation automation in neuro-oncology: advanced neural networks for MRI-based brain tumor segmentation and classification." IEEE Access (2025).

[12] Bhattacharjee, Subrata, et al. "Texture, morphology, and statistical analysis to differentiate primary brain tumors on two-dimensional magnetic resonance imaging scans using artificial intelligence techniques." Healthcare Informatics Research 28.1 (2022): 46-57.

[13] De Souza, Davidson Lucas, et al. "Brain tumor classification on multi-modality MRI using radiomic features." 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021.

[14] CHENG, J. Brain Tumor Dataset. Disponível em: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427, 4 2017.

[15] BHUVAJI, S.; KADAM, A.; BHUMKAR, P.; DEDGE, S.; KANCHAN, S. Brain Tumor Classification (MRI). [S. l.]: Kaggle, 2020. Disponível em: https://www.kaggle.com/dsv/1183165.

Comparative Evaluation of Cartesian and Cylindrical Models for QRS Complex Morphology Preservation in Nonlinear Kalman Filtering

Lucas Kenzo Araki
Academic Department of Electrotechnics (DAELT-CT),
Federal University of Technology — Paraná (UTFPR)
Curitiba Paraná, Brazil
lucasaraki@alunos.utfpr.edu.br

Renata Coelho Borges
Graduate Program in Biomedical Engineering (PPGEB-CT)
Federal University of Technology — Paraná (UTFPR)
Curitiba, Paraná, Brazil
renatacoelho@utfpr.edu.br

Victor Baptista Frencl
Academic Department of Electrotechnics (DAELT-CT),
Federal University of Technology — Paraná (UTFPR)
Curitiba Paraná, Brazil
vbfrencl@utfpr.edu.br

***Abstract*—Accurate preservation of Electrocardiogram (ECG) morphology, particularly the QRS complex, is essential for reliable arrhythmia detection and clinical decision support. State-space models combined with nonlinear Bayesian filters, such as the Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF), are widely used for ECG reconstruction under noisy conditions. However, most existing approaches primarily focus on improving filtering strategies, while the influence of the underlying state-space representation is often overlooked. In this work, we investigate the impact of Cartesian and Cylindrical models of a nonlinear ECG dynamical model on signal reconstruction and morphological preservation. By explicitly modeling phase as an independent state variable, the Cylindrical Model decouples phase and amplitude dynamics, improving temporal stability of the QRS complex. Experiments conducted on the MIT-BIH Arrhythmia Database demonstrate that the proposed Cylindrical representation consistently outperforms the Cartesian Model under both EKF and UKF frameworks, achieving superior morphological fidelity and improved discrimination of Premature Ventricular Contractions (PVC). Results further indicate that state-space representation plays a more critical role than the choice of nonlinear estimator in preserving ECG structure under stochastic noise.**

***Index Terms*—Electrocardiogram, State-space Modeling, Cylindrical Modeling, Nonlinear Bayesian Filtering.**

I. Introduction

Electrocardiogram (ECG) analysis plays a central role in cardiovascular diagnosis and monitoring, supporting applications such as Heart Rate Variability (HRV) analysis, arrhythmia detection, and real-time cardiac surveillance [1], [2]. In these applications, both the timing and morphology of the QRS complex are clinically relevant, since abnormalities in waveform width, amplitude, or temporal alignment may indicate conduction disturbances and ventricular arrhythmias [3]. However, ECG recordings are frequently contaminated by noise sources including muscle artifacts, electrode motion, and baseline wander, which degrade signal quality and complicate reliable morphological analysis [4].

To address these challenges, state-space representation of dynamical models, combined with nonlinear Bayesian filters, have been widely adopted for ECG reconstruction and denoising. The limit-cycle formulation introduced in [5] represents the cardiac cycle as a nonlinear trajectory driven by Gaussian functions associated with the P, QRS, and T waves. When embedded within Extended Kalman Filter (EKF) or Unscented Kalman Filter (UKF) frameworks [6], this model enables recursive estimation of latent cardiac states under uncertainty.

Despite their success, most existing approaches have focused primarily on improving nonlinear Bayesian filtering algorithms while adopting essentially the same underlying ECG dynamical model [6], [7]. In the original Cartesian formulation proposed by McSharry et al. [5], the cardiac phase is implicitly encoded in the planar state variables, whereas the formulation adopted by Sameni et al. [6] represents phase explicitly as a state variable. Consequently, perturbations affecting the planar states may propagate to phase estimation and ultimately to waveform reconstruction. This effect is particularly relevant in noisy recordings, where preserving QRS morphology is challenging [6], [8].

In this work, we investigate how the choice of state-space representation affects ECG morphology preservation under nonlinear Bayesian filtering. Rather than proposing a new filtering algorithm, the study focuses on the dynamical model used by the estimator. Specifically, Cartesian and Cylindrical ECG models are evaluated under both EKF and UKF frameworks. The Cylindrical formulation explicitly represents cardiac phase as a state variable, separating phase evolution from waveform amplitude dynamics and thereby providing a different geometric description of the cardiac cycle.

The results indicate that, within the evaluated configura-

tions, the choice of state-space representation has a larger effect on morphology preservation and PVC discrimination than the choice between EKF and UKF. The Cylindrical Model provides more stable phase tracking and more consistent reconstruction of QRS morphology, leading to improved R-peak timing accuracy and reliable detection of Premature Ventricular Contractions (PVC). This comparison therefore highlights the role of dynamical model structure as an important design choice in state-space-based ECG reconstruction.

The main contributions of this work are: (i) a systematic comparison between Cartesian and Cylindrical ECG state-space formulations under EKF and UKF frameworks; (ii) an analysis of how state-space geometry affects QRS morphology preservation and phase stability under noisy conditions; (iii) an experimental validation on the MIT-BIH Arrhythmia Database demonstrating that, within the evaluated scenarios, the Cylindrical Model provides more consistent reconstruction and classification performance across both estimators.

The remainder of this paper is organized as follows. Section II presents the Cartesian and Cylindrical ECG models. Section III describes the EKF and UKF frameworks. Section IV presents the experimental setup, including the dataset and evaluation protocol. Section V presents and discusses the results, comparing state-space representations and filtering strategies. Finally, Section VI concludes the paper.

II. ECG Mathematical Modeling

In order to a stochastic filter track a cardiac signal, it is imperative to establish a robust mathematical foundation describing the expected continuous dynamics of a normal ECG (sinus rhythm). There are two state-space-based formulations considered in this work: the Cartesian Model [5] and the Cylindrical Model [6].

A. Cartesian Model

The Cartesian Model proposed by [5] generates a synthetic ECG signal as a trajectory in a three-dimensional space with coordinates $x(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, and $z(t) \in \mathbb{R}$. The quasi-periodic nature of the ECG signal, i.e., the temporal repetition of heartbeats, is simulated by a limit-cycle trajectory with unit radius in the $x(t)$-$y(t)$ plane. Each complete turn corresponds to an RR interval, i.e., the duration of a heartbeat.

The PQRST morphology, characteristic of a normal ECG cycle, is reproduced using trajectory movement in the z-axis direction. Characteristic ECG points, such as the P, Q, R, S, and T waves, are described as events. These events correspond to negative and positive attractors or repellers in the z-axis direction. They are placed at fixed angles along the unit circle, specified as Θ_P, Θ_Q, Θ_R, Θ_S, and Θ_T for the P, Q, R, S, and T events, respectively. When the trajectory approaches one of these events, that is, when its current angle is close to the event angle, it is shifted upwards or downwards, moving away from the boundary cycle in the z-axis direction. In Figure 1 it is possible to visualize the boundary cycle by the dashed circle; the events are represented by the red spheres.

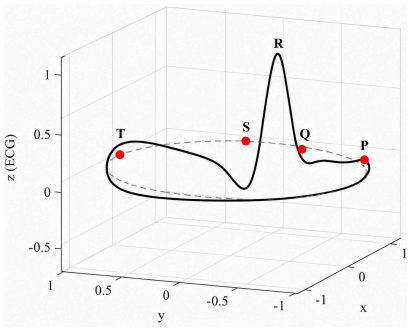


Fig. 1: Typical trajectory generated by the dynamical model; the dashed line indicates the limit cycle, the red spheres represent the events, while the motion along the z-axis modulates the shape of the P, Q, R, S, and T waves [5].

This model uses a set of three first-order Ordinary Differential Equations (ODE), as follows:

$$\dot{x}(t) = \alpha(t)x(t) - \omega y(t), \tag{1}$$

$$\dot{y}(t) = \alpha(t)y(t) + \omega x(t), \tag{2}$$

$$\dot{z}(t) = -\sum_{i\in\mathbb{E}} a_i \Delta\theta_i(t) \exp\left(-\frac{\Delta\theta_i^2(t)}{2b_i^2}\right) + (z_0(t) - z(t)), \tag{3}$$

where $\alpha(t) = 1 - \sqrt{x(t)^2 + y(t)^2}$, $\Delta\theta_i(t) = (\theta(t) - \theta_i) \mod 2\pi$, $\theta(t) = \tan^{-1}(y(t)/x(t))$, with $-\pi \leq \theta(t) \leq \pi$, and $\omega \in \mathbb{R}$ is the angular velocity of the trajectory as it moves around the limit cycle. The parameters a_i, b_i, and θ_i are constants that, for each waveform in the set $\mathbb{E} \in \{\text{P,Q,R,S,T}\}$, are related to its amplitude, width, and position, respectively. Their values are presented in Table I. Furthermore, the effect of respiration on the cardiac cycle is modeled by the function $z_0(t) = A\sin(2\pi f_r t)$, where $A = 0.15$ mV and f_r is the respiratory rate. Also, the term $z_0(t) - z(t)$ refers to the baseline fluctuation [9].

TABLE I: Parameters of the mathematical model of the ECG given by Equations (1)–(3).

Index (i)	P	Q	R	S	T
Time [s]	-0.2	-0.05	0	0.05	0.3
θ_i [rad]	$-\pi/3$	$-\pi/12$	0	$\pi/12$	$\pi/2$
a_i	1.2	-5.0	30.0	-7.5	0.75
b_i	0.25	0.1	0.1	0.1	0.4

Using the parameters in Table I, the system in Equations (1)–(3) was solved without respiratory influence. The generated signal (Figure 2) is similar to a real ECG signal.

B. Cylindrical Model

A potential problem with the Cartesian Model, presented in Section II-A, is that it possesses intrinsic limitations when subjected to recursive stochastic filtering processes, as it couples angular dynamics (e.g. the phase of the heartbeat) with its geometric amplitude. In order to avoid chronic tracking errors, a dynamic model based on a Cylindrical representation was presented in [6].

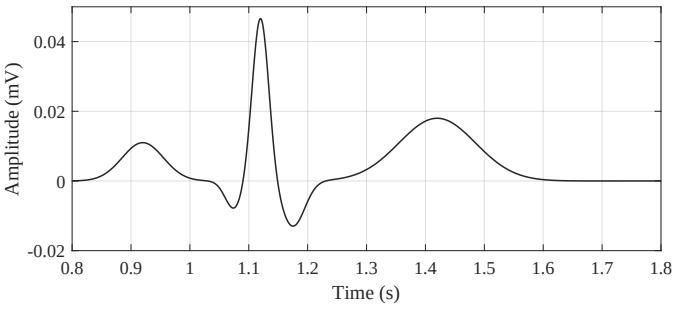


Fig. 2: Morphology of a synthetic ECG.

The transformation to Cylindrical coordinates $r(t) \in \mathbb{R}^+$, $\theta(t) \in [0; 2\pi]$, and $z(t) \in \mathbb{R}$ allows a more compact representation of system dynamics, focusing strictly on angle $\theta(t)$ and amplitude $z(t)$ states, providing a representation in which phase and amplitude dynamics are explicitly separated. In [6], the authors also state that the dynamic equation related to the $r(t)$ variable could be removed because it is redundant for the stochastic filter estimation. Therefore, the dynamics can be represented by the ODE in Equations (4) and (5).

$$\dot{\theta}(t) = \omega, \tag{4}$$

$$\dot{z}(t) = -\sum_{i\in\mathbb{E}} \alpha_i\, \omega\, \Delta\theta_i(t)\, \exp\left(-\frac{\Delta\theta_i^2(t)}{2b_i^2}\right) + (z_0(t) - z(t)), \tag{5}$$

where $\omega \in \mathbb{R}$ is the angular velocity related to the instantaneous heart rate, $b_i \in \mathbb{R}^+$ denotes the width of each wave, $\Delta\theta_i(t) \in [0; 2\pi]$ is the angular deviation of the trajectory relative to the centers of the Gaussian waves, $\alpha_i \in \mathbb{R}^+$ is the peak amplitude coefficients of the Gaussian functions that models the P, Q, R, S, and T components of the ECG.

C. Discretizations of the ECG Models

For the application of stochastic filters later on, it is necessary that the state-space model be written in discrete time. Therefore, Equations (1)–(3), and (4)–(5), must be discretized. In this work, we used the Euler method to perform the discretization [10]. The Euler method applied to the Cartesian Model results in the following set of Difference Equations, now including the measurement equation:

$$\mathbf{x}_{\text{cart}}(k+1) = \begin{bmatrix} x(k) + T\left[\alpha(k)x(k) - \omega y(k)\right] \\ y(k) + T\left[\alpha(k)y(k) + \omega x(k)\right] \\ z(k+1) \end{bmatrix} + \mathbf{w}(k), \tag{6}$$

$$z(k+1) = \\ z(k) + T\left[-\sum_{i\in\mathbb{E}} a_i \Delta\theta_i \exp\left(-\frac{\Delta\theta_i^2}{2b_i^2}\right) + (z_0(t) - z(k))\right] \tag{7}$$

$$z_{\text{cart}}(k) = z(k) + v_{\text{cart}}(k). \tag{8}$$

where $\mathbf{x}_{\text{cart}}(k) = \begin{bmatrix} x(k) & y(k) & z(k) \end{bmatrix}^\intercal \in \mathbb{R}^3$ is Cartesian Model state vector, $z_{\text{cart}}(k) \in \mathbb{R}$ is the Cartesian Model measurement vector, T is the sampling time, $\mathbf{w}_{\text{cart}}(k) \in \mathbb{R}^3$ is the Cartesian Model Gaussian process noise vector with zero mean and covariance matrix $Q_{\text{cart}}(k) \in \mathbb{R}^{3\times3}$, which represents the imperfection of the dynamic model, and $v_{\text{cart}}(k) \in \mathbb{R}$ is the Cartesian Model Gaussian measurement noise vector with zero mean and covariance matrix $R_{\text{cart}}(k) \in \mathbb{R}$, which represents the inaccuracies in the measurements.

Now, applying the Euler method to the Cylindrical Model and considering the inclusion of additive stochastic noises to each equation, the equivalent discrete-time mathematical model is obtained as follows:

$$\mathbf{x}_{\text{cyl}}(k+1) = \begin{bmatrix} \{\theta(k) + \omega T\} \mod (2\pi) \\ z(k+1) \end{bmatrix} + \mathbf{w}_{\text{cyl}}(k), \tag{9}$$

$$z(k+1) = z(k) - \sum_{i\in\mathbb{E}} T\, \frac{\alpha_i\, \omega}{b_i^2}\, \Delta\theta_i(k)\, \exp\left(-\frac{\Delta\theta_i^2(k)}{2b_i^2}\right) \tag{10}$$

$$\mathbf{z}_{\text{cyl}}(k) = \begin{bmatrix} \phi(k) & s(k) \end{bmatrix}^\intercal + \mathbf{v}_{\text{cyl}}(k), \tag{11}$$

where $\mathbf{x}_{\text{cyl}}(k) = \begin{bmatrix} \theta(k) & z(k) \end{bmatrix}^\intercal \in \mathbb{R}^2$ is the Cylindrical Model state vector, $\mathbf{w}_{\text{cyl}}(k) \in \mathbb{R}^2$ is the Cylindrical Model Gaussian process noise vector with zero mean and covariance matrix $Q_{\text{cyl}}(k) \in \mathbb{R}^{2\times2}$, $\mathbf{z}_{\text{cyl}}(k) = \begin{bmatrix} \phi(k) & s(k) \end{bmatrix}^\intercal \in \mathbb{R}^2$ is the measurement vector, $\phi(k) \in [-\pi; \pi)$ is the phase measurement, $s(k) \in \mathbb{R}$ is the ECG amplitude measurement, $\mathbf{v}_{\text{cyl}}(k) \in \mathbb{R}^2$ is the Cylindrical Model Gaussian measurement noise vector with zero mean and covariance matrix $R_{\text{cyl}}(k) \in \mathbb{R}^{2\times2}$.

III. Stochastic Filtering Framework

As presented in Section II, both Cartesian and Cylindrical Models are nonlinear. Therefore, the stochastic filters based on these models should be chosen in such a way to handle nonlinear models appropriately. The Kalman Filter (KF) [11] is one of the most used stochastic filters in the literature but it is formulated based on a linearity hypothesis, preventing the use of nonlinear models to obtain state estimates.

Thus, we choose two nonlinear stochastic filters which are KF variations, in order to obtain state estimations based on those nonlinear models: the EKF [12], [13] and the UKF [14].

A. Extended Kalman Filter

The EKF is a stochastic filter that is based on the linearization of the nonlinear mathematical models in discrete-time state-space representation. This stochastic filter allows the application of its equations to nonlinear systems through local linearization of the nonlinear model.

This linearization is performed using the First-Order Taylor Series (FOTS) expansion of the nonlinear functions around the predicted and estimated states operating point of the system, depending on which step the filter is localized. Considering the following nonlinear state-space representation:

$$\mathbf{x}(k+1) = \mathbf{f}[k, \mathbf{x}(k)] + \mathbf{w}(k), \tag{12}$$

$$\mathbf{z}(k) = \mathbf{h}[k, \mathbf{x}(k)] + \mathbf{v}(k), \tag{13}$$

where $\mathbf{x}(k) \in \mathbb{R}^n$ is the state vector of the system, $\mathbf{z}(k) \in \mathbb{R}^m$ is the measurement vector, $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$ are nonlinear functions and the process noise $\mathbf{w}(k) \in \mathbb{R}^n$

and measurement noise $\mathbf{v}(k) \in \mathbb{R}^m$ are Gaussian, additive and uncorrelated, such that $\mathbf{v}(k) \sim \mathcal{N}(0, Q(k))$ and $\mathbf{w}(k) \sim \mathcal{N}(0, R(k))$, where $Q(k) \in \mathbb{R}^{n\times n}$ and $R(k) \in \mathbb{R}^{m\times m}$ are the process and measurement noises covariance matrices, respectively.

The EKF is based on a equation set, divided into two main steps: the prediction step and the update step.

1) EKF Prediction Step: the equations for the EKF Prediction Step are given by:

$$\hat{\mathbf{x}}(k+1|k) = \mathbf{f}[k, \hat{\mathbf{x}}(k|k)], \tag{14}$$

$$P(k+1|k) = \mathscr{F}_x(k)P(k|k)\mathscr{F}_x^\intercal(k) + Q(k), \tag{15}$$

$$\hat{\mathbf{z}}(k+1|k) = \mathbf{h}(k+1, \hat{\mathbf{x}}(k+1|k)), \tag{16}$$

$$S(k+1) = \mathscr{H}_{\mathbf{x}}(k+1)P(k+1|k)\mathscr{H}_{\mathbf{x}}^\intercal(k+1) + R(k+1), \tag{17}$$

$$K(k+1) = P(k+1|k)\mathscr{H}_{\mathbf{x}}^\intercal(k+1)S^{-1}(k+1), \tag{18}$$

where $\hat{\mathbf{x}}(k|k) \in \mathbb{R}^n$ is the estimated state vector at k, $\hat{\mathbf{x}}(k+1|k) \in \mathbb{R}^n$ is the predicted state vector at $k+1$, $P(k|k) \in \mathbb{R}^{n\times n}$ is the estimation error covariance matrix at k, $P(k+1|k) \in \mathbb{R}^{n\times n}$ is the prediction error covariance matrix at $k+1$, $S(k+1) \in \mathbb{R}^{m\times m}$ is the innovation covariance matrix, and $K(k+1) \in \mathbb{R}^{n\times m}$ is the filter gain. The Jacobian matrices $\mathscr{F}_x(k)$ and $\mathscr{H}_{\mathbf{x}}(k+1)$ are given by:

$$\mathscr{F}_{\mathbf{x}}(k) = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right]_{\mathbf{x}=\hat{\mathbf{x}}(k|k)}, \tag{19}$$

$$\mathscr{H}_{\mathbf{x}}(k+1) = \left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right]_{\mathbf{x}=\hat{\mathbf{x}}(k+1|k)}. \tag{20}$$

2) EKF Update Step: the equations for the EKF Prediction Step are given by:

$$\hat{\mathbf{x}}(k+1|k+1) = \hat{\mathbf{x}}(k+1|k) + K(k+1)[\mathbf{z}(k+1) - \hat{\mathbf{z}}(k+1|k)], \tag{21}$$

$$P(k+1|k+1) = [I - K(k+1)\mathscr{H}_{\mathbf{x}}(k+1)]P(k+1|k), \tag{22}$$

where $\hat{\mathbf{x}}(k+1|k+1) \in \mathbb{R}^n$ is the estimated vector at $k+1$, $P(k+1|k+1) \in \mathbb{R}^{n\times n}$ is the estimation error covariance matrix at $k+1$.

B. Unscented Kalman Filter

The UKF has a different approach when compared to the EKF: it summarily avoids the use of FOTS expansions and Jacobian matrices by applying the Unscented Transform (UT) [14], [15]. The UT deterministically selects a minimal set of statistical points, called sigma-points, which captures the true mean and true covariance of the current state, instead of approximated values of mean and covariance.

1) UKF Prediction Step: Consider a state vector $\mathbf{x}(k) \in \mathbb{R}^n$, its estimation in time instant k, $\hat{\mathbf{x}}(k|k) \in \mathbb{R}^n$, based on a set of measurement vectors $Z^k = \{\mathbf{z}(0), \mathbf{z}(1), \ldots, \mathbf{z}(k)\}$, and a nonlinear function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, where $\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k))$.

In the UKF Prediction Step, a set of $2n+1$ sigma-points $\widehat{\mathscr{X}}(k|k)$ is created, given the updated state vector and covariance matrix, both evaluated at time instant k, for $i = 1, \ldots, n$.

$$\widehat{\mathscr{X}}(k|k)^{(i)} = \hat{\mathbf{x}}(k|k) + \left(\sqrt{(n+\lambda)P(k|k)}\right)_i^\intercal, \tag{23}$$

$$\widehat{\mathscr{X}}(k|k)^{(n+i)} = \hat{\mathbf{x}}(k|k) - \left(\sqrt{(n+\lambda)P(k|k)}\right)_i^\intercal, \tag{24}$$

where $P(k|k) \in \mathbb{R}^{n\times n}$ is the estimation error covariance matrix, $\lambda > 0$ is a parameter that reduces the approximation errors due to high order terms, and $(\cdot)_i^\intercal$ indicates the i-th column of the matrix.

Next, $\mathbf{f}$ is applied to each sigma-point, resulting in a new set of predicted sigma-points, given by:

$$\widehat{\mathscr{X}}(k+1|k)^{(i)} = \mathbf{f}(\widehat{\mathscr{X}}(k)^{(i)}). \tag{25}$$

After that, the predicted state is obtained as follows:

$$\hat{\mathbf{x}}(k+1|k) = \sum_{i=1}^{2n} W_m^{(i)}\, \widehat{\mathscr{X}}(k+1|k)^{(i)}, \tag{26}$$

where $W_m^{(i)} \in (0,1)$ is the weight related to the i-th sigma-point; these weights are given by, for $i = 1, \ldots, 2n$:

$$W_m^{(0)} = \lambda/(n+\lambda) \text{ and } W_m^{(i)} = 1/(2(n+\lambda)). \tag{27}$$

These weights in Equation (27) depend on a set of scaling parameters, as seen in [14]: λ is obtained as $\lambda = \alpha^2(n+p) - n$, where α defines the spread of the sigma-points around the estimation and it has a suggested value of 1×10^{-3}, p is a secondary scalar parameter, usually equal to zero.

The last part of the UKF Prediction Step is obtaining the prediction error covariance matrix, given by:

$$P(k+1|k) = Q(k) + \sum_{i=1}^{2n}\{W_c^{(i)}[\widehat{\mathscr{X}}(k+1|k)^{(i)} - \hat{\mathbf{x}}(k+1|k)][*]^\intercal\}, \tag{28}$$

where $W_c^{(i)} \in (0,1)$ is the weight related to the i-th sigma-point, related to the covariance matrix and the '*' symbol represents the same expression in the brackets on the left; these weights are given by, for $i = 1, \ldots, 2n$:

$$W_c^{(0)} = (1 - \alpha^2 + \beta) + \lambda/(n+\lambda), \tag{29}$$

$$W_c^{(i)} = 1/(2(n+\lambda)), \tag{30}$$

where $\beta > 0$ is the prior knowledge of the distribution; for Gaussian distributions, β is commonly set as 2.

2) UKF Update Step: In the UKF Update Step, new sigma-points are calculated, as follows, based on the prediction state $\hat{\mathbf{x}}(k+1|k)$ and the prediction error covariance matrix $P(k+1|k)$, as in Equations (26) and (28), respectively, for $i = 1, \ldots, n$:

$$\widehat{\mathscr{X}}(k+1|k)^{(i)} = \hat{\mathbf{x}}(k+1|k) + \left(\sqrt{nP(k+1|k)}\right)_i^\intercal, \tag{31}$$

$$\widehat{\mathscr{X}}(k+1|k)^{(n+i)} = \hat{\mathbf{x}}(k+1|k) - \left(\sqrt{nP(k+1|k)}\right)_i^\intercal. \tag{32}$$

Next, the new sigma-points in Equations (31) and (32) are transformed from the function $\mathbf{h}$ to the predicted measurement sigma-points $\widehat{\mathscr{Z}}(k+1|k)^{(i)}$. These points are combined in order to obtain the predicted measurement $\hat{\mathbf{z}}(k+1|k)$:

$$\widehat{\mathscr{Z}}(k+1|k)^{(i)} = \mathbf{h}(\widehat{\mathscr{X}}(k+1|k)^{(i)}), \tag{33}$$

$$\hat{\mathbf{z}}(k+1|k) = \sum_{i=1}^{2n} W_m^{(i)} \widehat{\mathscr{Z}}(k+1|k)^{(i)}. \tag{34}$$

The predicted measurement error covariance and cross-covariance matrices are given by, respectively:

$$P_z(k+1) = R(k+1) + \sum_{i=1}^{2n} W_c^{(i)} [\widehat{\mathscr{Z}}(k+1|k)^{(i)} - \hat{\mathbf{z}}(k+1|k)][*]^\intercal \tag{35}$$

$$P_{xz}(k+1) = \sum_{i=1}^{2n} W_c^{(i)} [\widehat{\mathscr{X}}(k+1|k)^{(i)} - \hat{\mathbf{x}}(k+1|k)][*]^\intercal. \tag{36}$$

The matrices in Equations (35) and (36) are fundamental for obtaining the UKF gain, as follows:

$$K(k+1) = P_{xz}(k+1) P_z^{-1}(k+1). \tag{37}$$

Finally, the updated state vector and the estimation error covariance matrix can be obtained, as follows:

$$\hat{\mathbf{x}}(k+1|k+1) = \hat{\mathbf{x}}(k+1|k) + K(k+1)[\mathbf{z}(k+1) - \hat{\mathbf{z}}(k+1|k)], \tag{38}$$

$$P(k+1|k+1) = P(k+1|k) - K(k+1) P_z(k+1) K(k+1)^\intercal. \tag{39}$$

IV. Experimental Setup

This section describes the experimental setup used to evaluate the proposed ECG-based arrhythmia detection framework. Here, we consider four scenarios based on the combination between the Cartesian and Cylindrical models and the EKF and the UKF: (i) Scenario 1: Cartesian + EKF; (ii) Scenario 2: Cartesian + UKF; (iii) Scenario 3: Cylindrical + EKF; and (iv) Scenario 4: Cylindrical + UKF.

A. Dataset

The proposed methodology was validated using the MIT-BIH Arrhythmia Database [16], [17], which provides gold-standard annotated ECG recordings. This database contains comprehensive, long-term ambulatory records of patients exhibiting several cardiac arrhythmias, providing a robust foundation for the detection algorithm. The validation signals, sampled at 360 Hz, were initially band-pass filtered using the same preprocessing procedure for all evaluated scenarios.

The study focused on PVC, using 79 signal segments of 20 seconds each, extracted from Lead II across 9 distinct patients. This selection totaled 1,299 normal beats and 649 PVC.

B. Filter Parameterization

The computational implementation of the stochastic filters was parameterized to function strictly as a normal sinus rhythm tracker, ensuring that any significant morphological deviation would be reflected in the innovation. As summarized in Table II, the parameters of each ECG dynamical formulation and its corresponding stochastic filter were defined according to their respective state-space representation.

The ECG model parameters — peak amplitudes (a_i, α_i), angular positions (θ_i), and wave widths (b_i) — were fixed a priori to define the standard morphology of the P, Q, R, S, and T waves. The sampling frequency f_s was set to 360 Hz to maintain consistency with the Database specifications.

To allow the stochastic filters to adapt their tracking windows to the patient's current physiological state, the state vectors and error covariance matrices were initialized based on heart rate estimates derived from the local average RR interval. Furthermore, the process and measurement noise covariance matrices were were defined for each state-space representation to provide stable tracking of healthy sinus rhythms. In the Cartesian model, an additive noise formulation uses a static 3×3 diagonal matrix for $Q(k)$. In the Cylindrical model, the measurement variance for the phase channel explicitly incorporates the angular velocity to account for time quantization.

While the UKF relies on purely additive process covariances, the Cylindrical EKF employs a non-additive structure where $Q(k) = \mathscr{W}(k) Q_{\text{init}} \mathscr{W}(k)^\intercal$. This formulation dynamically scales a 17×17 base matrix Q_{init}, which holds the physiological parameter uncertainties, through the process noise Jacobian matrix $\mathscr{W}(k)$. Finally, regarding the UKF tuning, the scaling constants (α, β, κ) were selected to maintain an optimal sigma-point spread for Gaussian distributions.

The two formulations are not intended to constitute matched-complexity models. Their state dimensions and covariance structures follow from the respective state-space representations and are therefore inherently different. The purpose of the comparison is not to equalize model complexity, but to evaluate how the choice of dynamical representation affects recursive ECG estimation when each formulation is implemented according to its corresponding state-space representation.

C. Beat Segmentation and Error Analysis

Individual cardiac cycles were identified using the estimated phase variable $\phi(k)$, mapped onto the interval $[-\pi, \pi)$ such that $\phi(k) = 0$ rad corresponds to the R-peak location. Each heartbeat was therefore defined as one complete phase transition within this interval, i.e., $-\pi \leq \phi(k) < \pi$. Segments corresponding to incomplete cardiac cycles at the boundaries of the 20-second windows were discarded.

The proposed detection strategy relies on the innovation generated by the stochastic filters. For every segmented heartbeat, the discrepancy between the measured ECG waveform

TABLE II: Dynamic model parameters and stochastic filter configurations for all evaluated scenarios.

Framework Component	Parameter / Configuration	Cartesian Model	Cylindrical Model
ECG Dynamical Model	Sampling Frequency (f_s)	360 Hz	360 Hz
	Peak Amplitudes (a_i, α_i)	$\{1.2, -5.0, 30.0, -7.5, 0.75\}$ mV	$\{0.25, -0.15, 1.8, -0.2, 0.5\}$ mV
	Angular Positions (θ_i)	$\{-\pi/3, -\pi/12, 0, \pi/12, \pi/2\}$ rad	$\{-\pi/3, -\pi/12, 0, \pi/12, \pi/2\}$ rad
	Wave Widths (b_i)	$\{0.25, 0.1, 0.1, 0.1, 0.5\}$ rad	$\{0.25, 0.1, 0.1, 0.1, 0.5\}$ rad
EKF Scenarios	State Dimension (n)	3	2
	Process Noise Covariance (Q)	$10^{-3} \times \text{diag}(1, 1, 0.1)$	$Q_{\text{init}} = \text{diag}(\sigma^2_{\alpha_i}, \sigma^2_{b_j}, \sigma^2_{\theta_i}, \sigma^2_{\omega}, \sigma^2_{\eta})_{17\times 17}$
	Measurement Noise Covariance (R)	10^{-2}	$\text{diag}\left((\omega T)^2/12, 10^{-2}\right)$
UKF Scenarios	State Dimension (n)	3	2
	Process Noise Covariance (Q)	$10^{-3} \times \text{diag}(1, 1, 0.1)$	$10^{-3} \times \text{diag}(1, 0.1)$
	Measurement Noise Covariance (R)	10^{-2}	$\text{diag}\left((\omega T)^2/12, 10^{-2}\right)$
	Spread of Sigma Points (α)	10^{-3}	10^{-3}
	Prior Distribution Parameter (β)	2	2
	Secondary Scaling Parameter (κ)	0	0

and the corresponding filter prediction was quantified through the Normalized Mean Squared Error (NMSE), defined as:

$$\text{NMSE} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \tag{40}$$

where $y_i \in \mathbb{R}$ denotes the observed ECG sample, $\hat{y}_i \in \mathbb{R}$ represents the estimated sample provided by either the EKF or UKF frameworks, and $\bar{y} \in \mathbb{R}$ is the mean value of the observed beat. The variable $N \in \mathbb{N}^*$ indicates the total number of samples within the analyzed cardiac cycle.

By normalizing the squared reconstruction error with respect to the signal variance, the NMSE provides a compact measure of reconstruction error for beat-level discrimination.

D. Classification and Cross-Validation

Arrhythmia detection was formulated as a beat-level binary classification problem, distinguishing normal beats from PVC events. A decision threshold T was optimized based on training data to maximize the Macro F1-score.

A 10-fold cross-validation scheme was adopted to assess performance variability across segment-level train/test partitions. In each fold, 90% of the data was used for threshold optimization and 10% for testing. Given the class imbalance in the dataset, Macro F1-score was selected as the primary evaluation metric, complemented by sensitivity to emphasize the importance of minimizing false negatives.

V. Results and Discussion

This section presents the qualitative and quantitative evaluation of the proposed ECG estimation-based arrhythmia detection framework. The analysis compares Scenarios 1–4 applied to the MIT-BIH Arrhythmia Database, focusing on their ability to reconstruct the ECG signal and to discriminate PVC through the NMSE.

A. Qualitative analysis: Scenarios 1 & 2

The first set of experiments evaluates the Cartesian Model using both the EKF (Scenario 1) and the UKF (Scenario 2). In both cases, the stochastic filters are able to approximate low-frequency components of the ECG signal, such as P and T waves, but fail to accurately reconstruct the R-peak amplitude, as showed in Figure 3a.

A more critical limitation is observed in phase estimation, where a significant mismatch between measured and estimated phase introduces temporal misalignment in the reconstructed signal (Figure 3b). As a consequence, the instantaneous estimation error does not reliably reflect the presence of arrhythmias.

This behavior is reflected in the per-beat NMSE distribution, which does not show a clear separation between normal and PVC beats (Figure 3c). Several abnormal beats exhibit NMSE values comparable to or even lower than normal beats, indicating that error-based thresholding is not reliable under the Cartesian Model.

In general, both Cartesian Model-based approaches show limited discriminative capability, mainly due to phase instability and R-peak underestimation.

B. Qualitative analysis: Scenarios 3 & 4

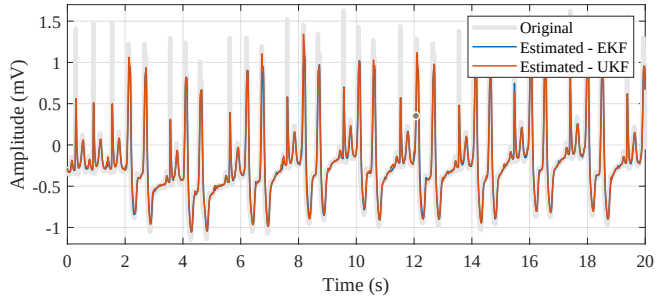
The second set of experiments evaluates the Cylindrical Model using both the EKF (Scenario 3) and the UKF (Scenario 4). These configurations significantly improve signal reconstruction quality. In both cases, phase tracking becomes stable and closely aligned with the measured signal (Figure 4). More importantly, the reconstructed ECG closely follows the morphology of normal beats, while showing clear deviations during PVC events (Figure 4b). These deviations manifest as sharp peaks in the instantaneous estimation error, which align consistently with abnormal beats. Unlike the Scenarios 1 and 2, the NMSE distribution exhibits here a clear separation between normal and PVC classes (Figure 4c).

The improvement is consistent across Scenarios 3 and 4 within the Cylindrical Model, supporting the interpretation that the state-space representation has a dominant influence on performance relative to the stochastic filter.

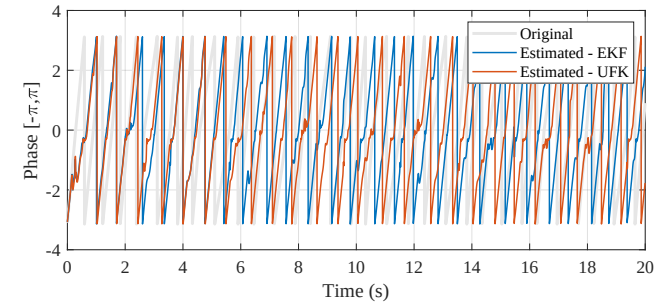
C. Quantitative evaluation

The quantitative evaluation was performed using 10-fold cross-validation over 79 annotated ECG segments. Performance was assessed at the beat level using Macro F1-score and sensitivity, prioritizing sensitivity due to the clinical importance of minimizing false negatives, as shown in Table III.

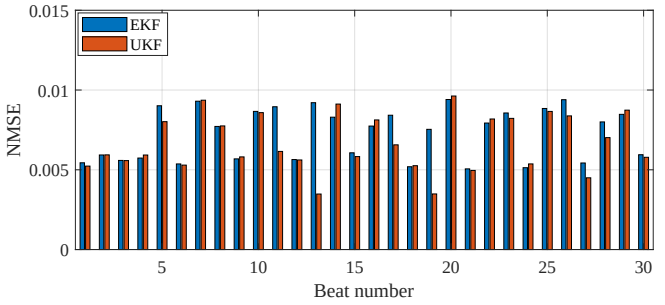
Cartesian-based configurations show low sensitivity, failing to detect a large portion of PVC events. In contrast, both Cylindrical Model-based configurations achieve substantially



(a) ECG real vs synthetic signal.

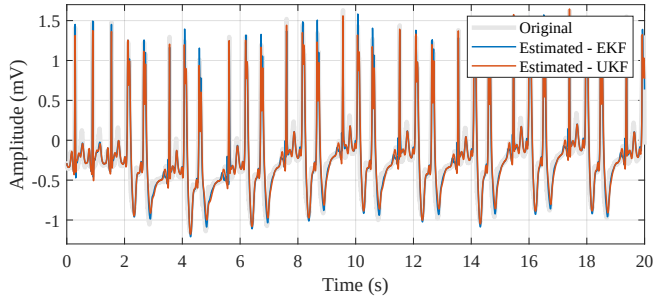


(b) Phase-space representation.

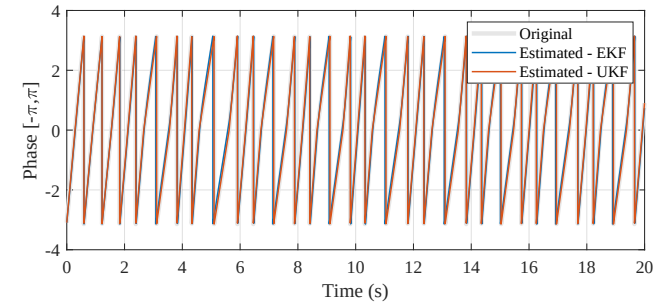


(c) NMSE per heartbeat.

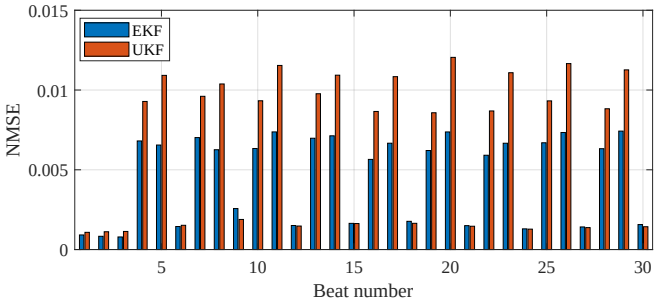
Fig. 3: Comparison between measured and estimated ECG signals in the Cartesian Model using time-domain, phase-space representation and per-beat NMSE.



(a) ECG real vs synthetic signal.



(b) Phase-space representation.



(c) NMSE per heartbeat.

Fig. 4: Comparison between measured and estimated ECG signals in the Cylindrical Model using time-domain, phase-space representation and per-beat NMSE.

TABLE III: Performance comparison of the proposed scenarios based on Sensitivity and Macro F1-score.

Scenario	Sensitivity (%)	Macro F1-score
1	47.63 ± 6.40	0.6905 ± 0.0347
2	31.59 ± 3.44	0.6963 ± 0.0193
3	89.99 ± 3.49	0.8843 ± 0.0171
4	85.83 ± 3.44	0.8924 ± 0.0160

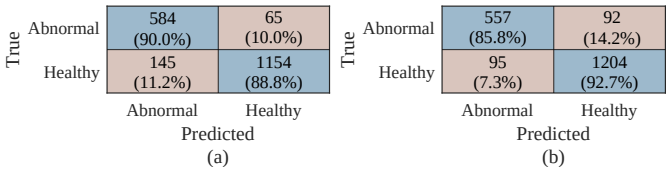


Fig. 5: Confusion matrices obtained for (a) Scenario 3 – EKF and (b) Scenario 4 – UKF.

higher sensitivity and overall classification performance. The Scenario 3 configuration attains the highest sensitivity, while the Scenario 4 configuration achieves the best balance between sensitivity and precision, leading to the highest Macro F1-score. The corresponding confusion matrices for the two Cylindrical Model scenarios summarize these results and are presented in Figure 5.

D. Discussion

The results indicate that, within the evaluated scenarios, the Cartesian Model was more limiting to ECG reconstruction than the choice of stochastic filtering technique. These findings suggest that state-space representation directly influences the stability of ECG morphological features during recursive estimation.

In particular, two issues dominate performance degradation: phase instability, which introduces temporal misalignment between measured and estimated signals, and R-peak underes-

timation, which reduces the discriminative power of the error signal. These limitations prevent NMSE from serving as a reliable proxy for morphological abnormalities.

The Cylindrical Model resolves both issues by explicitly modeling phase evolution and improving amplitude representation. As a result, estimation errors become strongly correlated with physiological abnormalities, particularly PVC.

Interestingly, the difference between EKF and UKF becomes marginal under the Cylindrical Model. While Scenario 3 (using EKF) achieves the highest sensitivity (89.99%), Scenario 4 (using UKF) yields a slightly higher Macro F1-score (0.8924), indicating a trade-off between sensitivity and precision. This behavior is also reflected in the confusion matrix, where Scenario 4 produces the lowest number of false positives (95) and achieves the highest accuracy for the healthy class (92.7%). Therefore, although its sensitivity is slightly lower ($85.83\% \pm 3.44\%$), the improved precision results in a more balanced classifier. These findings suggest that once the state-space representation is appropriate, the choice of the stochastic filter has a secondary impact on performance. From a practical perspective, the proposed recursive framework is compatible with real-time ECG processing, since both EKF and UKF operate sequentially without requiring future samples. However, the computational cost differs between the two estimators, particularly because the EKF requires Jacobian evaluation and matrix inversions, whereas the UKF propagates multiple sigma points at each step. This trade-off is relevant for wearable and embedded cardiac monitoring systems, where computational resources are limited.

VI. Conclusion

This work investigated the impact of state-space representation on ECG morphology preservation under nonlinear Kalman filtering. A systematic comparison between Cartesian and Cylindrical ECG dynamical models was performed using both EKF and UKF frameworks.

In the evaluated dataset and validation protocol, the Cylindrical Model provided substantially better phase stability, QRS morphology preservation, and PVC discrimination than the Cartesian Model. By explicitly decoupling phase and amplitude dynamics, the Cylindrical Model significantly reduces temporal misalignment effects and improves the correlation between estimation error and physiological abnormalities.

Quantitative evaluation on the MIT-BIH Arrhythmia Database showed substantial improvements in sensitivity and Macro F1-score for both Cylindrical Model-based configurations, regardless of the nonlinear stochastic filter employed. These findings suggest that, within the evaluated configurations, the state-space representation has a greater influence on ECG reconstruction quality than the choice between the EKF and UKF. Although the UKF achieved slightly higher overall classification performance, the difference between EKF and UKF became marginal once an appropriate state representation was adopted. This suggests that improving model geometry may provide greater benefits than increasing the stochastic filter complexity.

These findings should be interpreted within the limitations of the evaluated patient population, database, and segment-level validation protocol. Future work includes patient-independent validation on additional databases and ECG conditions, adaptive parameter estimation, other nonlinear stochastic filters such as Adaptive EKF/UKF and Particle Filter (PF), and real-time implementation, including computational and numerical stability assessment, for cardiac monitoring systems.

References

[1] S. F. H. Bokhari, A. B. Waseem, H. Raza, A. Iqbal, S. Javaid, B. Idrees, K. A. D. Saad, D. Bakht, and W. Dost, "Advancing cardiac arrhythmia management: The integration of wearable technology and remote monitoring," *World Journal of Cardiology*, vol. 17, no. 7, p. 106841, 2025.

[2] B. X. Wang, E. Brennand, P. Le Page, and A. R. Mitchell, "Heart rate variability in cardiovascular disease diagnosis, prognosis and management," *Frontiers in Cardiovascular Medicine*, vol. 12, p. 1680783, 2026.

[3] F. Kristof, M. Kapsecker, L. Nissen, J. Brimicombe, M. R. Cowie, Z. Ding, A. Dymond, S. M. Jonas, H. C. Lindén, G. Y. Lip, *et al.*, "Qrs detection in single-lead, telehealth electrocardiogram signals: benchmarking open-source algorithms," *PLOS digital health*, vol. 3, no. 8, p. e0000538, 2024.

[4] D. Velez, A. Lourenço, M. Pereira, D. P. Coutinho, and C. Carreiras, "Understanding the impact of noise on ecg biometrics: A comparative theoretical and experimental analysis," *Journal of Experimental and Theoretical Analyses*, vol. 4, no. 2, p. 14, 2026.

[5] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, "A dynamical model for generating synthetic electrocardiogram signals," *IEEE transactions on biomedical engineering*, vol. 50, no. 3, pp. 289–294, 2003.

[6] R. Sameni, M. B. Shamsollahi, C. Jutten, and G. D. Clifford, "A nonlinear bayesian filtering framework for ecg denoising," *IEEE transactions on Biomedical Engineering*, vol. 54, no. 12, pp. 2172–2185, 2007.

[7] H. D. Hesar and A. D. Hesar, "Adaptive augmented cubature kalman filter/smoother for ecg denoising," *Biomedical Engineering Letters*, vol. 14, no. 4, pp. 689–705, 2024.

[8] H. D. Hesar, "A novel ar-based kalman filtering framework for ecg enhancement," in *2025 32nd National and 10th International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 1286–1292, IEEE, 2025.

[9] D.-A. Truong, K.-T. Hoang, H.-H. Nguyen, M.-T. Nguyen, H.-D. Han, and L. Pham-Nguyen, "Baseline wander removal in biomedical signal acquisition system for heart rate monitoring," in *2025 2nd International Conference on Health Science and Technology (ICHST)*, pp. 1–6, IEEE, 2025.

[10] J. V. C. Vargas and L. K. Araki, *Cálculo Numérico Aplicado*. Barueri: Manole, 2017. E-book.

[11] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[12] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2001.

[13] G. Bishop, G. Welch, *et al.*, "An introduction to the kalman filter," *Proc of SIGGRAPH, Course*, vol. 8, no. 27599-23175, p. 41, 2001.

[14] E. Wan and R. Van Der Merwe, "The unscented kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pp. 153–158, 2000.

[15] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[16] G. Moody and R. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[17] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, pp. e215–e220, June 2000.

Response to Reviewers

Dear Reviewers,

First, we thank you for your suggestions and time dedicated to reviewing the manuscript "Characterisation of Motor Profiles in Bilateral Extension and Flexion Movements in Individuals with Parkinson's Disease" (Paper ID: 286), as well as for the opportunity to submit a revised version of this manuscript to the "Brazilian Congress on Biomedical Engineering".

We have carefully considered all suggestions and revised the manuscript accordingly. Below, we provide a point-by-point response to each comment. For clarity, the reviewers' comments are presented in bold, followed by our responses. Changes made to the manuscript are indicated in red in the annotated version of the article.

Responses to the comments from Reviewer 1

Comment: Unsupervised profiling of bilateral wrist extension-flexion can reveal heterogeneity, but clusters from a small medicated Parkinson cohort are exploratory and highly sensitive to feature scaling and sample dependence.

Response: Thank you for this comment. We agree that these factors may affect the results and, as further explained in our responses to the subsequent comments, we have addressed this concern by including a sensitivity analysis in the revised version of the manuscript.

Comment: State the number and characteristics of participants, medication timing, clinical ratings, sensor setup, trial count and all extracted variables.

Response: We thank the reviewer for this suggestion. We would like to clarify that most of these points had already been addressed in the original version of the manuscript. To provide further detail on the description already provided, we added information regarding the time elapsed between medication administration and assessment, as well as the anatomical region in which the sensors were positioned and the device used to acquire signals, with its settings. Although MDS-UPDRS scores are not reported, scores on the Hoehn and Yahr scale are provided, which reflect the patient's overall clinical status; this scale is not as sensitive to motor fluctuations as the MDS-UPDRS, but it is statistically correlated with it (M. Skorvanek, P. Martinez-Martin, N. Kovacs et al., "Differences in MDS-UPDRS scores based on Hoehn and Yahr stage and disease duration," Movement

Disorders Clinical Practice, vol. 4, no. 4, pp. 536-544, 2017). The modifications made to the text are highlighted in red.

<u>Comment:</u> Repeated movements and two limbs from one person must remain linked statistically.

Response: For statistical comparisons, the measurements extracted from the cycles (i.e., the peak-to-peak values) were obtained from the same participant for each limb, and these values were averaged within each limb for each participant prior to statistical analysis; this was not necessary for the others features, given that these measurements are already an aggregate of the cycles. The aggregated values for the two limbs were then compared using a paired-samples t-test or Wilcoxon signed-rank test, with each participant serving as the pairing unit. To help readers better understand this in the text, we explain this point in more detail in the "Statistical Analysis" subsection. All modifications made to the text are highlighted in red.

<u>Comment:</u> Justify retained PCA components and cluster number, show stability under resampling and alternative distances, and report participant membership rather than treating cycles as independent cases.

Response: We thank the reviewer for these important points. First, we would like to clarify that the clustering analysis was conducted at the participant level, with one observation per participant. The features used in the analysis were participant-level aggregated measures, and individual cycles were therefore not treated as independent observations. Second, the component-selection procedure is now best described in the Methods section. The number of principal components retained for the HCPC analysis was determined using generalized cross-validation, which indicated that three components were necessary to adequately represent the data. Third, we would like to clarify that, as described in the previous version of the manuscript, the number of groupings was not specified manually. The optimal partition was determined automatically based on the changes in within-cluster inertia associated with successive cuts of the hierarchical tree. This procedure resulted in a four-cluster solution. Finally, to assess the robustness of the clustering solution, we performed additional sensitivity analyses. These analyses and their results have been incorporated into the revised manuscript.

<u>Comment:</u> Clusters require external clinical interpretation and should not be described as levels of impairment without validated anchors.

Response: Thank you for this suggestion. We agree that the clusters should not be interpreted as levels of impairment in the absence of validated clinical references. We have revised the manuscript to make it clear that the clusters are descriptive

categorizations that require external clinical interpretation, and are not direct indicators of the severity of impairment.

Coment: The manuscript's fluent path from four clusters to personalized treatment may reflect automated language support, although style cannot establish authorship. Human revision should replace treatment-response claims with observed kinematic patterns, discuss fatigue and task-order effects and compare with healthy controls or OFF-medication data. The contribution is an exploratory motor-profile workflow. Verify terminology, sample counts, equations, cluster figures, references and all statements about precision medicine.

Response: Thank you for all suggestions. We acknowledge that assessing these conditions would be beneficial in allowing a comparison of different profiles, and in this regard, we recognize this as one of the limitations of the present study. However, we also acknowledge that, even if these comparisons were included, the findings could not be interpreted as evidence of a response to treatment or as a disease-specific impairment. This is because the effect of medication on each individual's voluntary motor response will also depend on whether the voluntary activity was learned under the influence of the medication or not. Therefore, we hypothesize that medication may contribute to both voluntary and involuntary responses in some cases but not in others, given that even when a person is under the influence of the medication. Thus, we revised the manuscript to ensure that these points were better addressed and discussed in the text. In addition, the suggested points are compared with the relevant literature. All changes have been incorporated into the revised manuscript and are highlighted in red.

Responses to the comments from Reviewer 2

Comment: The analysis focuses on wrist flexion-extension performed under controlled laboratory conditions, without investigating other clinically relevant motor tasks, OFF-medication assessments, longitudinal follow-up, or correlations with established clinical scales such as the MDS-UPDRS. While the clustering analysis identifies distinct motor phenotypes, no external validation or predictive assessment is provided to demonstrate their clinical utility in diagnosis, disease progression, or treatment monitoring.

Response: Thank you for your comment. We agree that all aspects you mentioned are among the limitations of this study. In particular, we acknowledge that assessing these conditions would be beneficial in allowing a comparison of different profiles, and in this regard, we recognize this as one of the limitations of the present study. However, we also acknowledge that, even if these comparisons were included, the findings could not be interpreted as evidence of a response to treatment or as a disease-specific impairment.

This is because the effect of medication on each individual's voluntary motor response will also depend on whether the voluntary activity was learned under the influence of the medication or not. Therefore, we hypothesize that medication may contribute to both voluntary and involuntary responses in some cases but not in others, given that even when a person is under the influence of the medication. In response to this comment, we have discussed these aspects in greater detail and have explicitly acknowledged them as limitations of the present study.

<u>Comment:</u> I also recommend a careful review to ensure full compliance with the CBEB formatting guidelines.

Response: We thank the reviewer for this suggestion. The manuscript has been carefully reviewed and reformatted to ensure compliance with the event template.

Characterization of Thermal Profiles in Pressure Injuries Using the Thermal Laplacian

Leomar Caetano de Oliveira
Bachelor's Program in Electronic Engineering
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0009-0006-5872-0272

José Carlos Pereira Coninck
Academic Department of Statistics
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0003-2198-5550

Catia Terezinha Heimbecher
Graduate Program in Electrical and Computer Engineering
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0003-4487-153X

Patrícia Jucele da Silva Voziniak
Graduate Program in Biomedical Engineering
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0009-0005-0607-6860

Henrique Cunha Carvalho
Graduate Program in Biomedical Engineering
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0003-0806-4132

Leandra Ulbricht
Graduate Program in Biomedical Engineering
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0002-9514-2938

***Abstract*—Pressure injuries represent a relevant problem in clinical practice, given the associated morbidity and the need for quantitative methods that contribute to the objective analysis of cutaneous perfusion. In this context, infrared thermography is a promising alternative for obtaining thermal information from the body surface; however, its interpretation remains predominantly qualitative or relies on global metrics, which limits the analysis of pressure injuries. Accordingly, there remains a gap in the application of mathematical methods capable of quantitatively describing this perfusion. The objective of this study was to evaluate the performance of the thermal Laplacian as an indicator of perfusion patterns in thermographic images of patients at risk of pressure injury by comparing representative level curves from distinct thermal profiles. A descriptive study with a quantitative approach was conducted using thermographic images from a previously established database of patients admitted to a reference hospital. Each image was organized as a two-dimensional temperature matrix, from which level curves with distinct spatial behaviors were extracted. For each curve, a thermal surface was fitted using a generalized linear model; subsequently, the thermal gradient and Laplacian were computed together with the associated descriptive statistics. The results identified three distinct thermal profiles: one with predominantly positive curvature, associated with local thermal concentration; one with approximately neutral curvature, indicating no significant mean curvature; and one with predominantly negative curvature, associated with greater spatial heterogeneity. These findings indicate that the thermal Laplacian discriminates patterns not adequately captured by global measures alone, such as mean temperature or thermal amplitude, highlighting the potential of mathematical modeling as a complementary quantitative tool for characterizing thermographic images of pressure injuries.**

***Index Terms*—pressure injuries, infrared thermography, thermal Laplacian, tissue perfusion, mathematical modeling.**

I. Introduction

Pressure injuries (PIs) are lesions resulting from deformation of the skin and underlying soft tissues, caused by sustained pressure alone or by pressure combined with shear forces exerted by body weight or medical devices [1], [2]. These factors lead to restricted blood flow, tissue hypoxia, and inflammation, ranging from intact skin to open wounds and frequently accompanied by pain [1], [3]–[5]. From a physical standpoint, these alterations modify the local heat balance in the tissue, since reduced perfusion interferes with thermal exchange between the blood, skin, and external environment, producing spatial temperature variations that can be represented mathematically as thermal fields [6], [7].

PIs develop preferentially over bony prominences [1], [2] and are particularly prevalent in individuals with reduced mobility or those admitted to intensive care units (ICUs) [1], [8], [9]. Under these conditions, continuous pressure on the skin may produce regions with different thermal regimes related to heat conduction, dissipation, and redistribution within the tissues.

In clinical practice, risk assessment is performed using scales such as the Braden, Norton, and Waterlow scales, which are based on clinical variables including mobility, nutrition, moisture, and pressure and/or shear forces [10]–[13]. Although widely used, these tools rely on the assessor's clinical judgment, which may introduce inter- and intra-observer variability and hinder the early detection of perfusion alterations [14]. Recent international guidelines emphasize that risk assessment should be structured, periodic, and complemented by additional clinical factors, especially in critically ill individuals with perfusion alterations, shock, mechanical ventilation, vasopressor use, or medical devices [2]. Moreover, these scales do not directly describe the spatial distribution of temperature or the local geometry of the thermal field, both of which may be relevant to characterizing alterations associated with cutaneous perfusion.

In this context, infrared thermography (IT) is a complementary method that enables objective measurement of the thermal distribution of the skin surface [15], [16]. Local thermal variations may reflect microcirculatory changes associated with inflammatory or ischemic processes, thereby enabling the early identification of regions susceptible to injury [4], [17], [18].

Recent studies also indicate that skin temperature differences may serve as indicators of imminent tissue damage in pressure injuries, especially when combined with computational image-analysis methods [19]. However, when used in isolation, its interpretation remains predominantly qualitative or relies on global metrics, such as mean temperature, maximum temperature, or thermal amplitude, which limits the analysis of the spatial organization of the thermal field [15], [16], [20].

In this regard, there remains a gap in the application of quantitative methods capable of describing the spatial structure of temperature to support clinical assessment, both in the context of traditional scales and in infrared thermography analysis [14]–[16], [21], [25]. Global measures may summarize overall thermal intensity but do not necessarily capture local information, such as concentration, dispersion, asymmetry, and spatial heterogeneity. Therefore, mathematical approaches based on local and global descriptors have been proposed to improve the quantification of information contained in thermal images [20].

From a mathematical standpoint, the temperature distribution in biological tissues can be described as a continuous scalar field $T(x, y)$ associated with thermal conduction, blood perfusion, and metabolic heat generation. Models based on partial differential equations, such as the Pennes bioheat equation, establish this relationship by connecting thermal gradients to physiological terms [22]–[24]. Recent bioheat simulation studies reinforce that skin-surface temperature can be analyzed using differential models that incorporate tissue thermal properties, blood perfusion, metabolic heat generation, and boundary conditions [6], [7]. In this context, the Laplacian operator, associated with the second spatial derivative of temperature, describes the curvature of the thermal surface and is related to heat diffusion and redistribution processes, allowing the identification of spatial patterns that are not captured by global thermal measures [21], [25].

In light of this scenario, this study aims to evaluate the performance of the thermal Laplacian as an indicator of perfusion patterns in thermographic images of patients at risk of PI. Specifically, it seeks to model the thermal surface from representative level curves, estimate the thermal Laplacian as a measure of spatial heat redistribution, and compare distinct thermal profiles with respect to their spatial organization.

II. Material and Methods

A descriptive study with a quantitative approach and non-probability sampling was conducted using thermographic images from a previously established database comprising 30 individuals admitted to the intensive care unit (ICU), internal medicine ward, or trauma unit of a reference hospital in Curitiba, Brazil. The study was approved by the Research Ethics Committee (CAAE: 57943822.9.3001.5225; approval no. 5.651.212).

The inclusion criteria comprised bedridden individuals or those with reduced mobility, classified as having or not having a pressure injury. Clinical interpretation and injury staging were performed in accordance with the National Pressure Injury Advisory Panel (NPIAP) guidelines [2].

Three distinct thermal patterns were identified in the analyzed regions. To exemplify these patterns in the proposed mathematical analysis, three representative thermal curves were selected from the complete image database. Each selected thermal curve was defined as a region of interest, enabling comparative evaluation of the spatial temperature distribution and the observed thermal curvature.

A. Thermographic image acquisition

Thermographic images of the gluteal and sacrococcygeal regions were acquired using an infrared camera (Ti32, 320×240 infrared sensor; Fluke Thermography, Everett, WA, USA). The standardized protocol included exposing the region of interest for five minutes of thermal acclimatization in an environment maintained at 22 ± 1 °C and monitored using a digital thermo-hygrometer with an external probe (SH 112; J-Prolab, Brazil).

During acquisition, participants were placed in the lateral decubitus position, with an approximate distance of 50 cm between the skin surface and the camera.

An example of the optical image and the corresponding thermographic image is shown in Fig. 1.

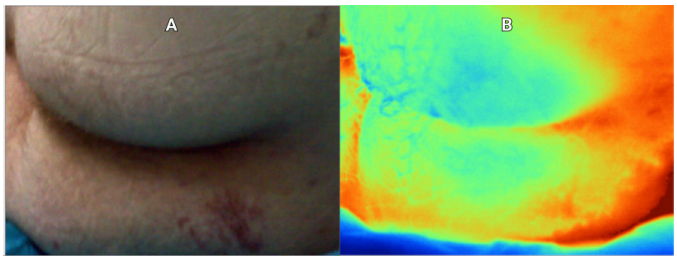


Fig. 1. Example of acquisition: (a) optical image of the analyzed region; (b) thermographic image.

B. Image processing and extraction of thermal curves

All statistical and mathematical processing was performed using R (version 4.3.0; R Foundation for Statistical Computing, Vienna, Austria). The thermographic image was organized as a two-dimensional temperature matrix, allowing the spatial thermal distribution to be represented by the scalar function $T(x, y)$.

From this matrix, thermal level curves were obtained, defined as sets of points that share the same temperature value. To exemplify the three identified patterns, the selected thermal curves were denoted P1, P2, and P3.

Each selected thermal curve was defined as a region of interest and spatially delimited to extract the local thermal data used for surface modeling.

C. Thermal surface modeling

For each region of interest, the spatial temperature distribution was modeled using a generalized linear model (GLM), assuming a normal distribution and an identity link function.

The thermal surface was represented by a polynomial function of two variables, allowing the spatial variation of temperature in each analyzed region to be described. The final model structure was defined through automatic term selection, retaining only those terms that were statistically relevant to representing the local thermal geometry. From this fit, the fitted thermal surface $T(x, y)$ was obtained and used as the basis for the differential analysis.

D. Computation of the thermal gradient and Laplacian

From the fitted thermal surface, the thermal gradient was computed to describe the direction and magnitude of the greatest local temperature variation, as shown in (1).

$$\nabla T(x, y) = \left(\frac{\partial T}{\partial x}(x, y),\ \frac{\partial T}{\partial y}(x, y) \right) \tag{1}$$

Next, the thermal Laplacian was computed and adopted as the central parameter of the analysis because it represents the local curvature of the thermal surface. Previous studies by our research group evaluated the use of the Pennes equation to estimate blood perfusion in pressure injuries and to model thermal gradients in thermographic images of these lesions [26], [27]. The formulation is based on the Pennes bioheat equation, which represents heat transfer in perfused tissue and explicitly includes the blood-perfusion term, as shown in (2) [22].

$$\nabla \cdot (k \nabla T) - \rho_b c_b \omega_b (T - T_a) + q_m = 0 \tag{2}$$

where k is the thermal conductivity of the tissue, ρ_b is blood density, c_b is the specific heat of blood, ω_b is the blood perfusion rate, T is tissue temperature, T_a is arterial temperature, and q_m represents metabolic heat generation.

Under the assumption of local homogeneity of the thermal parameters and neglecting spatial variations in k, the equation can be reduced to a form proportional to the Laplacian operator, as shown in (3).

$$\nabla^2 T(x, y) = \frac{\partial^2 T}{\partial x^2}(x, y) + \frac{\partial^2 T}{\partial y^2}(x, y) \tag{3}$$

In terms of curvature, $\nabla^2 T > 0$ indicates predominantly positive curvature; $\nabla^2 T \approx 0$ indicates approximately null mean curvature; and $\nabla^2 T < 0$ indicates predominantly negative curvature. Thus, the Laplacian was used to mathematically summarize differences among the thermal profiles represented by the three selected curves. These values are presented as mathematical descriptors of the spatial organization of temperature rather than as diagnostic indicators.

E. Statistical analysis

For each region of interest, descriptive statistics of the thermal Laplacian values were computed, including the number of points, mean, standard deviation, and 95% confidence interval. Next, a one-sample Student's t-test was performed to evaluate the null hypothesis of a mean Laplacian equal to zero, as shown in (4).

TABLE I
DESCRIPTIVE CHARACTERISTICS OF THE SAMPLE ($n = 30$)

Variable	Median	SD	P25–P75	Min.–Max.	n (%)
Continuous characteristics					
Age (years)	62	23	45–81	22–91	–
Weight (kg)	64	13	58–71	35–97	–
Height (m)	1.6	0.11	1.6–1.7	1.4–1.8	–
Sex and ethnicity					
Female sex	–	–	–	–	11 (35%)
Male sex	–	–	–	–	19 (65%)
White ethnicity	–	–	–	–	23 (77%)
Mixed-race ethnicity	–	–	–	–	7 (23%)

Note: SD = standard deviation; P25 = 25th percentile; P75 = 75th percentile.

$$H_0 : \mu_{\nabla^2 T} = 0 \tag{4}$$

Failure to reject this hypothesis was interpreted as indicative of approximately null mean curvature, whereas its rejection was considered evidence of significant thermal curvature. All analyses adopted a significance level of 5%.

III. RESULTS AND DISCUSSION

Table I summarizes the continuous descriptive characteristics and the distribution of categorical variables in the sample ($n = 30$). The median age was 62 years, with a predominance of male participants (65%) and individuals identified as White (77%). Regarding anthropometric measurements, the median weight and height were 64 kg and 1.60 m, respectively.

Regarding comorbidities, the most frequent were arterial hypertension (43%), diabetes mellitus (13%), chronic lung disease (10%), and heart disease (7%). For approximately one-fifth of the participants, the electronic medical record contained insufficient information to confirm the presence or absence of comorbidities.

The clinical characterization of pressure injuries, including staging, anatomical location, and risk stratification using the Braden scale, was performed by the hospital team, with clinical interpretation based on the NPIAP guidelines [2]. Table II presents these clinical characteristics separately. With respect to injury stage, participants without a pressure injury were the most frequent group (48%), followed by Stage 2 (29%) and Stage 1 (13%). Unstageable cases accounted for 6% of the sample, whereas Stage 3 had the lowest frequency (3%).

Based on the thermographic analysis of the regions of interest and the three images selected to exemplify the thermal patterns, 43 valid level curves were considered. The distribution by temperature range showed a predominance of curves with temperatures equal to or greater than 32 °C. Twenty-one curves (48.8%) had temperatures equal to or greater than 35 °C, 20 curves (46.5%) ranged from 32 to 34.9 °C, and two curves (4.7%) ranged from 28 to 32 °C. No cases below 28 °C were detected, as shown in Table III.

The analysis identified three distinct thermal profiles according to the curvature of the fitted thermal surface, corresponding to positive (P1), approximately neutral (P2), and

TABLE II
CLINICAL CHARACTERISTICS OF PRESSURE INJURIES: STAGES, ANATOMICAL LOCATION, AND RISK STRATIFICATION

Variable	Frequency	%
Pressure injury stages		
Stage 1	4	13%
Stage 2	9	29%
Stage 3	1	3%
Unstageable	2	6%
No pressure injury	15	48%
Anatomical location		
Gluteal region	7	23%
Sacral region	10	33%
Risk stratification (Braden)		
High risk	9	30%
Low risk	4	13%
Moderate risk	9	30%
Very high risk	8	27%

TABLE III
DISTRIBUTION OF LEVEL CURVES BY TEMPERATURE RANGE

Temperature range	Count	% of total ($n = 43$)
$\geq 35\,^{\circ}$C	21	48.8%
32–34.9 °C	20	46.5%
28–32 °C	2	4.7%
$< 28\,^{\circ}$C	ND	ND

Note: ND = not detected (no cases recorded in this category).

negative (P3) Laplacian values. Fig. 2 combines representative examples of these three profiles into a single composition. In each row, the unsmoothed thermal level curves, the thermal surface fitted by the GLM, and the thermal-gradient vector field are presented for the same profile. The level curves represent contours of equal temperature and allow thermally homogeneous zones and transition regions to be visualized. In the three-dimensional surfaces, the spatial axes represent the region of interest and the vertical axis represents the fitted temperature, while the gradient vectors indicate the direction and magnitude of the greatest local temperature variation.

A. Profile 1: predominantly positive curvature

In the first row of Fig. 2, corresponding to Profile 1, the unsmoothed thermal level curves, the thermal surface fitted by the GLM, and the thermal-gradient vector field are shown. A well-defined central elevation and continuous spatial variation are evident.

The thermal Laplacian had a mean of $\overline{\nabla^2 T} = 0.1385$, a standard deviation of 0.1298, a 95% confidence interval from 0.1030 to 0.1739, and $p = 2.03 \times 10^{-10}$ in the t-test for $H_0 : \mu_{\nabla^2 T} = 0$ ($n = 54$). Because the confidence interval lies entirely above zero and the null hypothesis was rejected, $\nabla^2 T > 0$ characterizes predominantly positive curvature associated with local spatial concentration of the thermal surface.

B. Profile 2: approximately neutral curvature

In the second row of Fig. 2, corresponding to Profile 2, the unsmoothed thermal level curves, the thermal surface fitted by the GLM, and the thermal-gradient vector field are shown. A more distributed spatial variation is evident.

The thermal Laplacian had a mean of $\overline{\nabla^2 T} = -0.0108$, a standard deviation of 0.0647, a 95% confidence interval from -0.0330 to 0.0114, and $p = 0.331$ in the t-test for $H_0 : \mu_{\nabla^2 T} = 0$ ($n = 35$). Because the confidence interval includes zero and the null hypothesis was not rejected, $\nabla^2 T \approx 0$ characterizes the absence of significant mean curvature, although local positive and negative variations remain present.

C. Profile 3: predominantly negative curvature

In the third row of Fig. 2, corresponding to Profile 3, the unsmoothed thermal level curves, the thermal surface fitted by the GLM, and the thermal-gradient vector field are shown. A concave configuration with a central depression is observed.

The thermal Laplacian had a mean of $\overline{\nabla^2 T} = -0.0729$, a standard deviation of 0.0253, a 95% confidence interval from -0.0778 to -0.0680, and $p = 6.77 \times 10^{-52}$ in the t-test for $H_0 : \mu_{\nabla^2 T} = 0$ ($n = 104$). Because the confidence interval lies entirely below zero and the null hypothesis was rejected, $\nabla^2 T < 0$ characterizes predominantly negative curvature associated with greater spatial heterogeneity of the thermal surface.

D. Comparison of the profiles

Comparison of the three profiles reveals three curvature regimes of the fitted thermal surface: predominantly positive in P1, approximately null in P2, and predominantly negative in P3. The sign and magnitude of the Laplacian provide information about the local geometry that is not supplied by global measures such as mean temperature or thermal amplitude [21], [25].

The gradient describes the direction of thermal variation, whereas the Laplacian summarizes the curvature of the surface. Polynomial modeling enabled these differential descriptors to be computed and provided a quantitative representation of the spatial organization of temperature [15], [16], [21], [25]. Integrating these metrics with machine-learning methods remains a direction for future research [14], [19].

IV. CONCLUSION

This study evaluated the thermal Laplacian as a descriptor of the spatial organization of temperature in thermographic images of patients at risk of pressure injury. The thermal surfaces were modeled using a GLM, allowing comparison of three curvature regimes: predominantly positive ($\nabla^2 T > 0$, P1), approximately neutral ($\nabla^2 T \approx 0$, P2), and predominantly negative ($\nabla^2 T < 0$, P3).

The results show that the Laplacian provides local information not supplied by global metrics alone. As a methodological contribution, these descriptors may support future quantitative approaches and automated classifiers, including neural-network-based models.

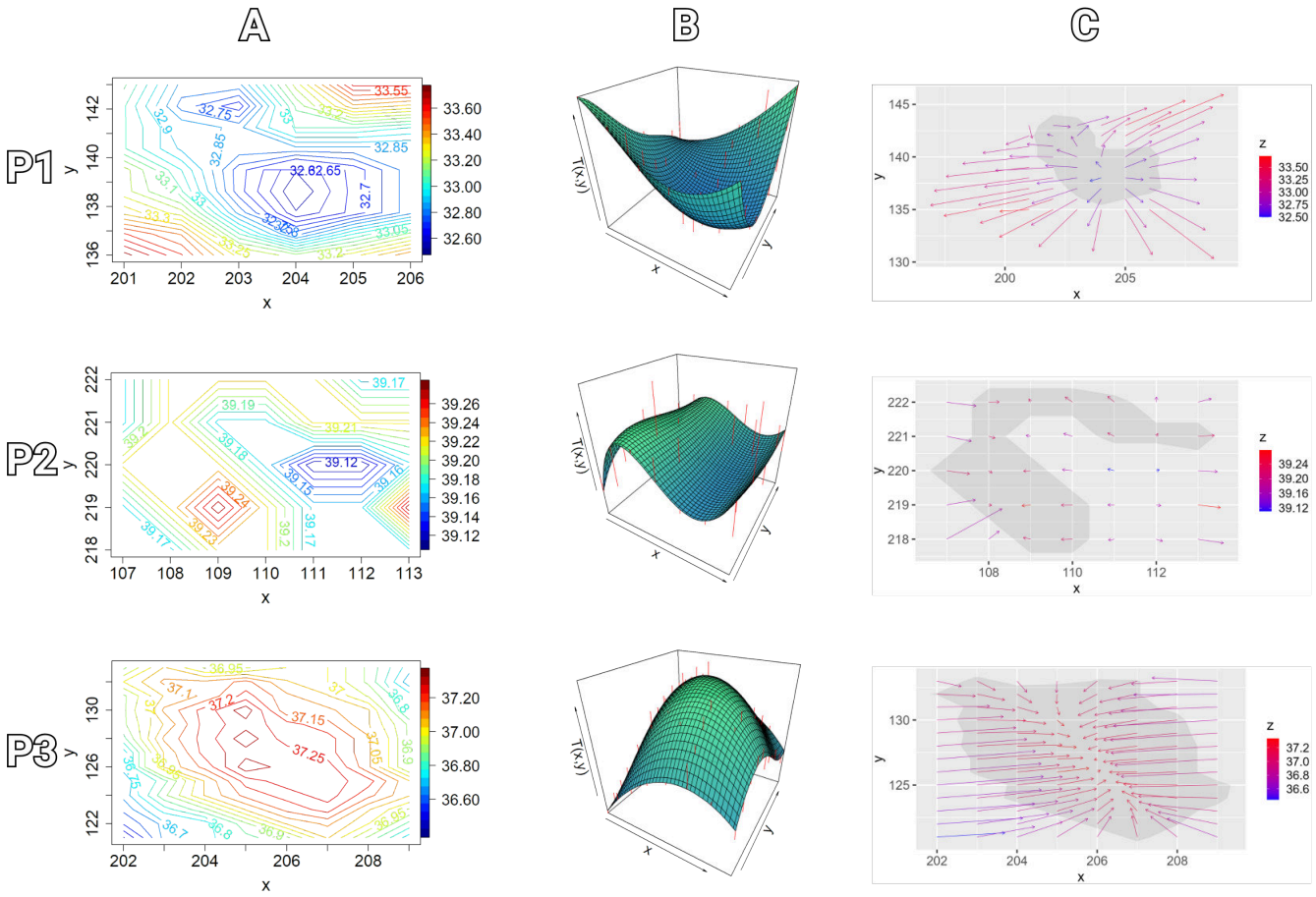


Fig. 2. Single composition of the three representative thermal profiles. The rows correspond to profiles P1, P2, and P3. The columns correspond to: (A) unsmoothed thermal level curves of the region of interest; (B) thermal surface fitted by the generalized linear model (GLM); and (C) thermal-gradient vector field.

ACKNOWLEDGMENT

The authors thank Fundação Araucária (FA) for the PIBIC scholarship; Fundação Araucária (FA) and the Superintendency of Science, Technology, and Higher Education (SETI) for the conference participation grants awarded to H. C. C.; and Universidade Tecnológica Federal do Paraná and its Graduate Program in Biomedical Engineering for institutional support.

REFERENCES

[1] European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel, and Pan Pacific Pressure Injury Alliance, *Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline. The International Guideline*, E. Haesler, Ed. EPUAP/NPIAP/PPPIA, 2019.

[2] National Pressure Injury Advisory Panel, European Pressure Ulcer Advisory Panel, and Pan Pacific Pressure Injury Alliance, *Prevention and Treatment of Pressure Ulcers/Injuries: Quick Reference Guide. The International Guideline: Fourth Edition*, E. Haesler, Ed., 2025. [Online]. Available: https://internationalguideline.com

[3] M. Martínez-Torija, P. F. Esteban, A. Santos-De-La-Mata, M. Castillo-Hermoso, E. Molina-Holgado, and R. Moreno-Luna, "Multifaceted pathophysiology and secondary complications of chronic spinal cord injury: Focus on pressure injury," *Journal of Clinical Medicine*, vol. 14, no. 5, Art. no. 1556, 2025, doi: 10.3390/jcm14051556.

[4] H. Jayabal, N. S. Abiakam, D. Filingeri, D. L. Bader, and P. R. Worsley, "Inflammatory biomarkers in sebum for identifying skin damage in patients with a Stage I pressure ulcer in the pelvic region: A single centre observational, longitudinal cohort study with elderly patients," *International Wound Journal*, vol. 20, no. 7, pp. 2594–2607, 2023, doi: 10.1111/iwj.14131.

[5] H. B. Beriso, W. Zemene, and E. Tesfaye, "Prevalence of pressure ulcers and associated factors among adult patients admitted at Comprehensive Specialized Hospital, Northwest Ethiopia, 2023," *Scientific Reports*, vol. 14, no. 1, Art. no. 17290, 2024, doi: 10.1038/s41598-024-67026-5.

[6] D. V. Akulova and M. A. Sheremet, "Mathematical simulation of bio-heat transfer in tissues having five layers in the presence of a tumor zone," *Mathematics*, vol. 12, no. 5, Art. no. 676, 2024, doi: 10.3390/math12050676.

[7] G. D'Alessandro, P. Tavakolian, and S. Sfarra, "A review of techniques and bio-heat transfer models supporting infrared thermal imaging for diagnosis of malignancy," *Applied Sciences*, vol. 14, no. 4, Art. no. 1603, 2024, doi: 10.3390/app14041603.

[8] A. Alito, S. Portaro, G. Leonardi, C. Ventimiglia, F. Bonanno, D. Fenga, C. Sconza, and A. Tisano, "Pressure ulcers—A longstanding problem: A 7-year neurorehabilitation unit experience of management, care, and clinical outcomes," *Diagnostics*, vol. 13, no. 20, Art. no. 3213, 2023, doi: 10.3390/diagnostics13203213.

[9] C. F. Gheri, L. Scalfi, B. Biffi, S. Pancani, S. Madiai, O. Di Vincenzo, *et al.*, "Relationship between nutritional risk, clinical and demographic characteristics, and pressure ulcers in patients with severe acquired brain injuries attending a rehabilitative program," *Nutrients*, vol. 15, no. 15, Art. no. 3336, 2023, doi: 10.3390/nu15153336.

[10] R. J. B. R. Picoito, S. M. M. P. C. Lapuente, A. C. P. Ramos, I. C.

M. Rabiais, S. J. Deodato, and E. M. G. T. Nunes, "Risk assessment instruments for pressure ulcer in adults in critical situation: A scoping review," *Revista Latino-Americana de Enfermagem*, vol. 31, Art. no. e3983, 2023, doi: 10.1590/1518-8345.6659.3983.

[11] A. T. Borghardt, T. N. do Prado, T. M. de Araújo, N. M. B. Rogenski, and M. E. de O. Bringuente, "Evaluation of the pressure ulcers risk scales with critically ill patients: A prospective cohort study," *Revista Latino-Americana de Enfermagem*, vol. 23, no. 1, pp. 28–35, 2015, doi: 10.1590/0104-1169.0144.2521.

[12] Q. Yang, Z. Yang, L. Lv, H. Zhang, H. Tao, J. Pei, Y. Ma, and L. Han, "Comparing the Waterlow and Jackson/Cubbin pressure injury risk scales in intensive care units: A multi-centre study," *International Wound Journal*, vol. 21, no. 2, Art. no. e14602, 2024, doi: 10.1111/iwj.14602.

[13] Y. Shang, F. Wang, Y. Cai, Q. Zhu, X. Li, R. Wang, and T.-H. Tung, "The accuracy of the risk assessment scale for pressure ulcers in adult surgical patients: A network meta-analysis," *BMC Surgery*, vol. 25, no. 1, Art. no. 104, 2025, doi: 10.1186/s12893-024-02739-y.

[14] B. Hillier, K. Scandrett, A. Coombe, T. Hernandez-Boussard, E. Steyerberg, Y. Takwoingi, V. Veličković, and J. Dinnes, "Accuracy and clinical effectiveness of risk prediction tools for pressure injury occurrence: An umbrella review," *PLoS Medicine*, vol. 22, no. 2, Art. no. e1004518, 2025, doi: 10.1371/journal.pmed.1004518.

[15] Y.-H. Lin, Y.-C. Chen, K.-S. Cheng, P.-J. Yu, J.-L. Wang, and N.-Y. Ko, "Higher periwound temperature associated with wound healing of pressure ulcers detected by infrared thermography," *Journal of Clinical Medicine*, vol. 10, no. 13, Art. no. 2883, 2021, doi: 10.3390/jcm10132883.

[16] F. Cai, X. Jiang, X. Hou, D. Wang, Y. Wang, H. Deng, H. Guo, H. Wang, and X. Li, "Application of infrared thermography in the early warning of pressure injury: A prospective observational study," *Journal of Clinical Nursing*, vol. 30, no. 3–4, pp. 559–571, 2021, doi: 10.1111/jocn.15576.

[17] M. E. A. Chaves, F. S. Silva, V. P. C. Soares, R. A. M. Ferreira, F. S. L. Gomes, R. M. Andrade, and M. Pinotti, "Evaluation of healing of pressure ulcers through thermography: A preliminary study," *Research on Biomedical Engineering*, vol. 31, no. 1, pp. 3–9, 2015, doi: 10.1590/2446-4740.0571.

[18] H. J. Nam, S. Y. Wee, S. Y. Kim, *et al.*, "The correlation between transcutaneous oxygen pressure ($TcPO_2$) and forward-looking infrared (FLIR) thermography in the evaluation of lower extremity perfusion according to angiosome," *International Wound Journal*, vol. 21, Art. no. e14431, 2024, doi: 10.1111/iwj.14431.

[19] M. Asare-Baiden, K. Jordan, A. Chung, S. E. Sonenblum, and J. C. Ho, "Is thermography a viable solution for detecting pressure injuries in dark skin patients?" in *Proc. Machine Learning for Health (ML4H) Findings Track*, 2024. [Online]. Available: https://arxiv.org/abs/2411.10627

[20] H. Ayunts, A. Grigoryan, and S. Agaian, "Novel entropy for enhanced thermal imaging and uncertainty quantification," *Entropy*, vol. 26, no. 5, Art. no. 374, 2024, doi: 10.3390/e26050374.

[21] H. J. Li, X. X. Zhang, and Y. F. Yi, "Measurement of blood perfusion using the temperature response to constant surface flux heating," *International Journal of Thermophysics*, vol. 23, no. 6, pp. 1631–1644, 2002, doi: 10.1023/A:1020798119128.

[22] H. H. Pennes, "Analysis of tissue and arterial blood temperatures in the resting human forearm," *Journal of Applied Physiology*, vol. 1, no. 2, pp. 93–122, 1948, doi: 10.1152/jappl.1948.1.2.93.

[23] S. Weinbaum, L. M. Jiji, and D. E. Lemons, "Theory and experiment for the effect of vascular microstructure on surface tissue heat transfer—Part I: Anatomical foundation and model conceptualization," *Journal of Biomechanical Engineering*, vol. 106, no. 4, pp. 321–330, 1984, doi: 10.1115/1.3138501.

[24] K. R. Diller, "Modeling of bioheat transfer processes at high and low temperatures," *Advances in Heat Transfer*, vol. 22, pp. 157–357, 1992, doi: 10.1016/S0065-2717(08)70345-9.

[25] H. El-Kebir, J. Ran, Y. Lee, L. P. Chamorro, M. Ostoja-Starzewski, R. Berlin, G. M. Aguiluz Cornejo, E. Benedetti, P. C. Giulianotti, and J. Bentsman, "Minimally invasive live tissue high-fidelity thermophysical modeling using real-time thermography," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 6, pp. 1849–1857, 2023, doi: 10.1109/TBME.2022.3230728.

[26] P. J. S. Voziniak, J. C. P. Coninck, E. F. R. Romaneli, and L. Ulbricht, "Mathematical modeling to estimate perfusion in pressure injuries," in *XXIX Brazilian Congress on Biomedical Engineering—Volume 2: Tissue Engineering, Clinical Engineering and Computational Modeling in Biomedical Engineering*, A. B. Soares, R. F. Leoni, and G. C. Cardoso, Eds., ser. IFMBE Proceedings, vol. 126. Cham, Switzerland: Springer, 2025, pp. 702–710, doi: 10.1007/978-3-031-94921-0_76.

[27] V. G. Nasimoto, P. J. S. Voziniak, J. C. P. Coninck, E. F. R. Romanelli, and L. Ulbricht, "Thermal gradients at the verification of pressure injuries," in *XXIX Brazilian Congress on Biomedical Engineering—Volume 3: Biomedical Informatics, and Biomedical Signal and Image Processing*, A. B. Soares, R. F. Leoni, and G. C. Cardoso, Eds., ser. IFMBE Proceedings, vol. 127. Cham, Switzerland: Springer, 2025, pp. 482–491, doi: 10.1007/978-3-031-94934-0_50.

Classification of Dementia Stages in Magnetic Resonance Images by Pre-Trained CNNs and Ensemble

Joel Sotero da Cunha Neto
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
joelsotero@unifor.br

Renan de Lima Simas
Control and Automation Engineering
University of Fortaleza
Fortaleza, Brazil
renansimas@edu.unifor.br

Matheus Figueiredo dos Reis
Control and Automation Engineering
University of Fortaleza
Fortaleza, Brazil
matheusfigreis@edu.unifor.br

Paulo Henrique Maia Filho
Control and Automation Engineering
University of Fortaleza
Fortaleza, Brazil
phmf2002@edu.unifor.br

Pedro de Souza Tito
Control and Automation Engineering
University of Fortaleza
Fortaleza, Brazil
pedrotito@edu.unifor.br

João Igor Vidal de Andrade
Computer Science
University of Fortaleza
Fortaleza, Brazil
joaoigorandrade@edu.unifor.br

Belmondo Rodrigues Aragão Junior
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
belmondo@unifor.br

***Abstract*—Alzheimer's disease is a progressive neurodegenerative condition, and the analysis of magnetic resonance imaging (MRI) has been explored as a support for automatic image classification studies. This work evaluates pre-trained convolutional neural network architectures for classifying MRI images into four classes: Mild Demented, Moderate Demented, Non Demented, and Very Mild Demented. The dataset was obtained from Kaggle and was already made available with balanced and augmented images. Architectures from the ResNet, DenseNet, EfficientNet, EfficientNetV2, Inception, and Xception families were tested, with transfer learning, fine-tuning, data augmentation, regularization, Test-Time Augmentation (TTA), and ensemble by soft voting. The best individual model was the optimized EfficientNetV2-B2, with a macro F1-score of 0.9919. The final ensemble, composed of five models, achieved a macro F1-score of 0.9959 and an accuracy of 0.9958. The results indicate that combining complementary architectures and fine-tuning strategies can increase the robustness of multiclass classification on this dataset. However patient-level evaluation on original volumetric MRI remains necessary before any clinical use.**

***Index Terms*—Alzheimer; Magnetic Resonance Imaging; Convolutional Neural Networks; Transfer Learning; Ensemble.**

I. Introduction

Alzheimer's disease is a neurodegenerative disorder associated with the progressive deterioration of memory, cognition, and functional autonomy. Clinical diagnosis depends on specialized medical evaluation and can be supported by imaging exams, including magnetic resonance imaging (MRI), which is capable of revealing structural changes related to the neurodegenerative process. In this context, computational methods can assist in analyzing large amounts of images and support automatic classification studies [1].

Convolutional neural networks (CNNs) have shown impressive performance in computer vision and medical image analysis, as they learn hierarchical representations directly from pixels [2], [10]. In biomedical problems, however, the number of labeled images is usually smaller than in general computer vision datasets, which increases the risk of overfitting and makes transfer learning particularly relevant [4].

In transfer learning, an architecture previously trained on a large dataset, such as ImageNet, is adapted to a new domain by replacing the classifier layer and performing partial or full weight adjustment. This strategy is useful in medical imaging because it reduces training time and allows reuse of visual filters already learned, such as edges, textures, and intensity patterns [4], [10].

Studies on Alzheimer's disease classification using CNNs report that comparison across studies can be hampered by differences in preprocessing, data splitting, patient selection, and experimental validation [3]. Therefore, in addition to comparing architectures, it is important to record the tested variations, including those that did not lead to improvements, as they help understand model behavior when dealing with visually similar classes.

Given this context, this work aims to evaluate different pre-trained CNN architectures for the classification of Alzheimer's MRI images into four classes, comparing baselines, training variations, fine-tuning, TTA, and ensemble. The main focus is

to identify which strategies improved performance and which techniques did not contribute to better results. We emphasize that the study is restricted to a public, pre-balanced and pre-augmented image dataset and that part of the evaluation relies on image-level splitting. Accordingly, the results characterize internal performance on this processed dataset and are not intended as clinical evidence.

II. Methodology

A. Dataset

The public dataset *Alzheimer's Disease Multiclass Images Dataset*, available on Kaggle, specifically the `combined_images` folder, was used [5]. The dataset consists of MRI images organized into four classes: Mild Demented, Moderate Demented, Non Demented, and Very Mild Demented. The subset used in the main experiments (Table I) contains 12,000 images, with 3,000 images per class. For the main training runs, 2,400 images per class were used for training and 600 images per class for validation/testing, totaling 9,600 training images and 2,400 validation/test images.

Table I
Distribution of the main dataset used in the experiments.

Class	Total	Train	Val.
Mild	3000	2400	600
Moderate	3000	2400	600
Non	3000	2400	600
Very Mild	3000	2400	600
Total	12000	9600	2400

The dataset had already undergone a prior balancing and data augmentation process before being made available in the repository. Thus, the experiments in this work were conducted on a previously augmented dataset, and additional data augmentation techniques were applied only to the training set when specified in each experimental variation. The four classes represent different conditions or levels of cognitive impairment. The Moderate Demented class was the easiest to classify, while Non Demented and Very Mild Demented were the most frequently misclassified classes.

B. Preprocessing

Images were resized according to the architecture being evaluated: 224×224 for most models, 260×260 for EfficientNetV2-B2, and 299×299 for Inception-V3. ImageNet-based normalization was applied, along with grayscale conversion in some experiments, auto-contrast, black border removal, and mild histogram equalization during training.

Data augmentation transformations such as RandomResizedCrop, light rotation, horizontal flip, RandomAffine, and RandomErasing were evaluated. Transformations were applied only to the training set. The validation and test sets were kept without artificial augmentation, except in TTA experiments, in which inference combines the original image and transformed versions.

C. Architectures and variations

Pre-trained architectures from the ResNet [6], DenseNet [7], EfficientNet, EfficientNetV2 [9], Inception [8], and Xception families were evaluated. For all architectures, the final layer was replaced by a classifier head compatible with the four classes of the problem. Initially, baselines with a frozen backbone were trained to establish a reference for comparison.

After the baselines, models with the greatest potential were subjected to experimental variations (II). Among them, the following stand out: fine-tuning techniques such as partial and full backbone unfreezing, classifier head modification, class weights, Focal Loss, moderate MRI augmentation, learning rate schedulers, optimizers such as Adam and AdamW, weight decay, label smoothing, and TTA.

Table II
Experimental configurations evaluated.

Exper.	Architecture	Strategy
Baseline	ResNet / DenseNet / EfficientNet	Frozen backbone
Variations	DenseNet-169 / 201	Weights, Focal, Aug. and TTA
Optimized	EfficientNetV2-B2	TL + partial FT + TTA
Ensemble	Top-5	Soft voting

D. Training and evaluation

Models were implemented in Python and, in some experiments, using the timm library. The Adam and AdamW optimizers were tested, along with different learning rates, early stopping, and schedulers such as ReduceLROnPlateau and CosineAnnealingLR. The main metric for selecting the best models was the macro F1-score.

The macro F1-score was chosen because accuracy alone can mask difficulties in specific classes. Since the problem has four classes and some of them are visually similar, the analysis of confusion matrix, precision, recall, and per-class F1 was essential for interpreting model behavior.

After individual evaluation, five models were selected for the final ensemble: DenseNet-169, DenseNet-201, EfficientNetV2-B2, Inception-V3, and EfficientNetV2-B0. Two soft voting strategies were evaluated: simple averaging of probabilities and weighted averaging based on the individual macro F1-score.

III. Results and Discussion

A. Baselines

The baselines allowed comparing the initial behavior of the architectures with a frozen backbone. EfficientNetV2-B2 was the best baseline, with a macro F1-score of 0.8082, indicating better initial adaptation to the image set. In contrast, ResNet-50 and Inception-V3 showed lower performance, with macro F1-scores of 0.6467 and 0.6124, respectively, as shown in Table III.

Table III
EVALUATED BASELINES.

Model	Acc.	F1
EffNetV2-B2	0.8114	0.8082
EffNet-B0	0.7427	0.7337
EffNet-B2	0.7408	0.7346
DenseNet-169	0.7342	0.7387
DenseNet-201	0.7225	0.7318
DenseNet-121	0.7192	0.7245
ResNet-34	0.6717	0.6761
Xception	0.6650	0.6672
ResNet-18	0.6642	0.6636
ResNet-101	0.6583	0.6595
ResNet-50	0.6492	0.6467
Inception-V3	0.6125	0.6124

B. Variations and best models

The variations demonstrated that fine-tuning and regularization were more effective than simply increasing the complexity of the classifier head or applying balancing techniques in isolation (Table IV). DenseNet-201 with full fine-tuning and differentiated learning rate achieved a macro F1-score of 0.9861, while the optimized EfficientNetV2-B2 obtained the best individual result, with a macro F1-score of 0.9919.

Table IV
MAIN EXPERIMENTAL VARIATIONS.

Model/Variation	Strategy	Acc.	F1
EffNetV2-B2 optim.	TL + FT + TTA	0.9919	0.9919
DenseNet-201 V1	Full FT + diff. LR	0.9858	0.9861
DenseNet-169 V1	Partial FT + Cosine	0.9325	0.9338
Inception-V3 V4	AdamW + TTA	0.8821	0.8819
DenseNet-169 V4	TTA 5x	0.7642	0.7680
DenseNet-201 V4	TTA 5x	0.7592	0.7669
DenseNet-201 V3	Robust head	0.6775	0.6837
DenseNet-169 V3	Focal Loss	0.6800	0.6833
DenseNet-169 V3	Moderate aug.	0.6767	0.6795
DenseNet-201 V2	Class weights	0.6633	0.6764

On the other hand, some techniques reduced performance. In DenseNet-201, class weights resulted in a macro F1-score of 0.6764 and the robust head achieved 0.6837, both below the baseline. In DenseNet-169, Focal Loss and moderate augmentation also reduced performance. These results suggest that balancing strategies need to be carefully calibrated.

C. Best individual model

The optimized EfficientNetV2-B2 combined transfer learning, partial fine-tuning, dropout, weight decay, label smoothing, class weights, ReduceLROnPlateau, gradient clipping, and TTA. The training curves are presented in Figs. 1–3, and indicate consistent loss reduction, progressive accuracy increase, and stabilization of the macro F1-score at a high level. The confusion matrix in Fig. 4 shows high performance across all classes, with residual errors concentrated mainly between Non Demented and Very Mild Demented.

In Fig. 1, a continuous decrease in training and validation loss is observed, indicating progressive convergence. The most pronounced reduction occurs after the onset of fine-tuning,

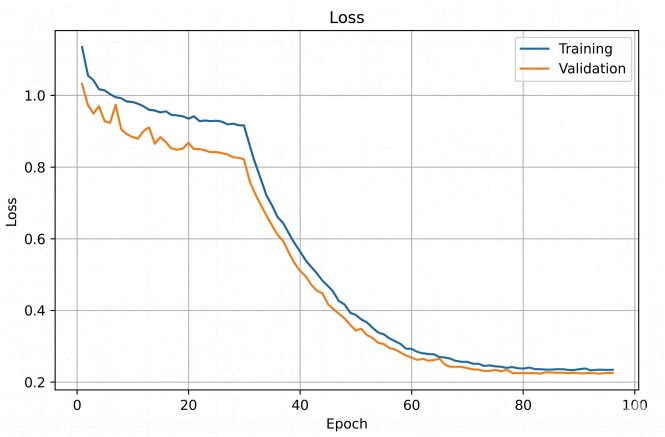


Figure 1. Loss curve of the optimized EfficientNetV2-B2.

when the model begins to better adapt its final blocks to the MRI image domain.

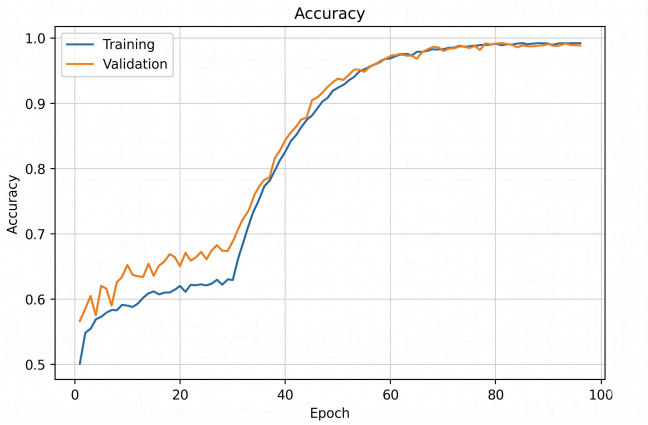


Figure 2. Accuracy curve of the optimized EfficientNetV2-B2.

In Fig. 2, accuracy increases gradually over epochs, with stronger growth after the fine-tuning stage. The proximity between training and validation curves indicates stable training and low evidence of severe overfitting.

In Fig. 3, the macro F1-score grows consistently and stabilizes near the end of training, showing balanced improvement across all four classes and justifying the choice of the optimized model as the best individual classifier.

In Fig. 4, the main diagonal presents values close to 1.00 for all classes, indicating high per-class recall. The few remaining errors appear mainly between Non Demented and Very Mild Demented, which is consistent with the visual similarity between these patterns.

D. Final ensemble

The final ensemble was the best overall strategy (Table 5). The combination of the five selected models resulted in a macro F1-score of 0.9959 and an accuracy of 0.9958, surpassing the best individual model used in the ensemble, which was DenseNet-201 with a macro F1-score of 0.9852.

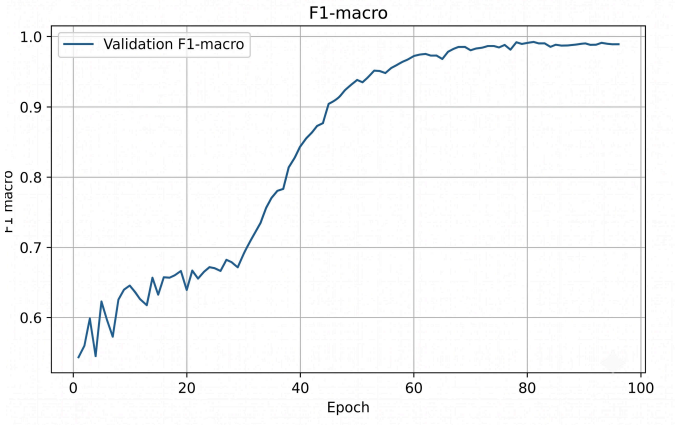


Figure 3. Validation macro F1-score evolution of the optimized EfficientNetV2-B2.

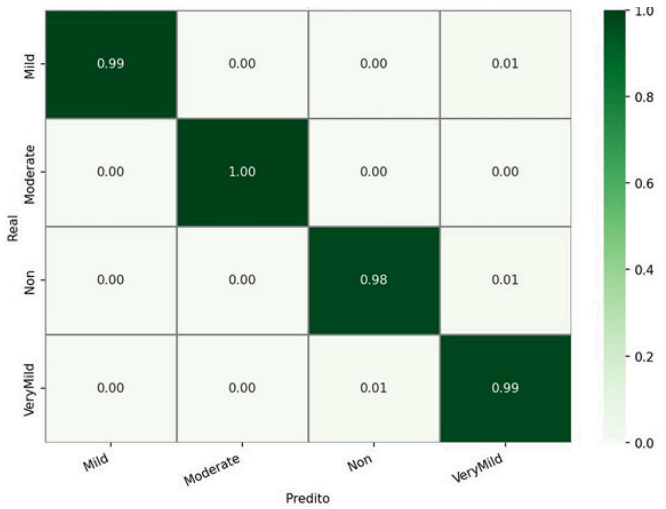


Figure 4. Normalized confusion matrix of the optimized EfficientNetV2-B2 with TTA.

The improvement indicates that the models captured complementary patterns.

Table V
MODELS USED IN THE FINAL ENSEMBLE.

Model	Strategy	F1
DenseNet-169	Partial FT	0.9322
DenseNet-201	Full FT	0.9852
EffNetV2-B2	TTA 2x	0.9820
Inception-V3	TTA 2x	0.9795
EffNetV2-B0	TTA 2x	0.9762
Simple ensemble	Soft voting	0.9959
Weighted ensemble	Soft voting	0.9959

In Fig. 5, it is observed that individual models achieve high performance, with macro F1-scores above 0.93. However, the simple and weighted soft voting ensemble strategies reach 0.9959, surpassing all individual classifiers. This result indicates that the combination of complementary architectures reduced individual errors and increased classification robustness.

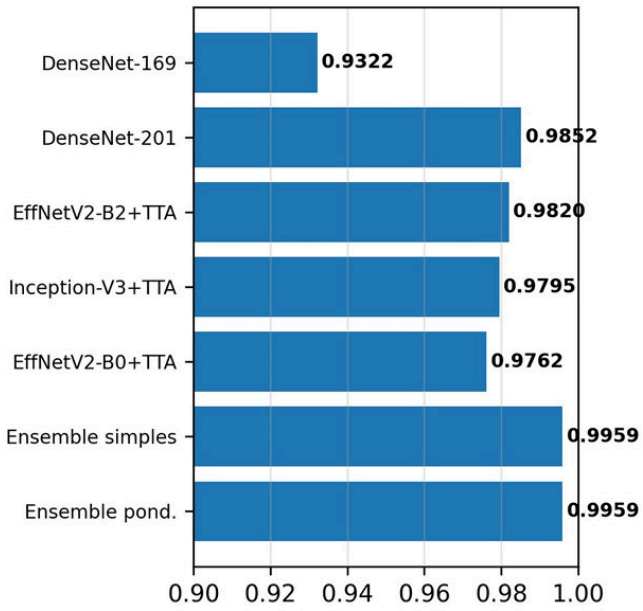


Figure 5. Macro F1-score comparison between individual models and ensemble.

In Fig. 6, DenseNet-169 shows good concentration of correct predictions on the diagonal, but still maintains confusion between Non Demented and Very Mild Demented, indicating difficulty with the visually closest classes.

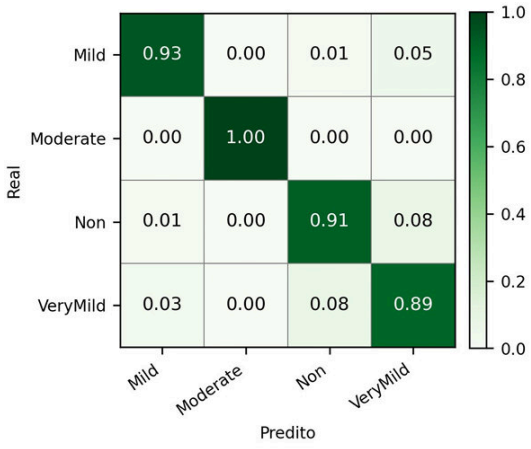


Figure 6. Normalized confusion matrix - DenseNet-169.

In Fig. 7, DenseNet-201 reduces errors compared to DenseNet-169 and presents a stronger diagonal, especially for the Mild Demented and Moderate Demented classes.

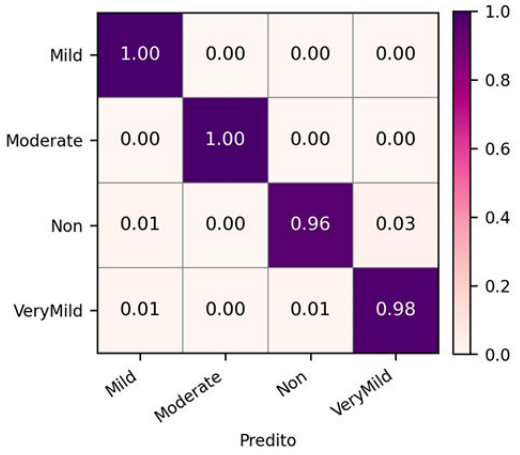


Figure 7. Normalized confusion matrix - DenseNet-201.

In Fig. 8, EfficientNetV2-B2 with TTA maintains high performance across all classes, with minor confusion between Non Demented and Very Mild Demented.

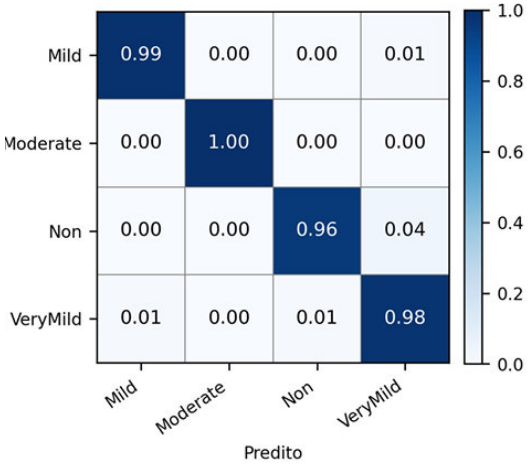


Figure 8. Normalized confusion matrix - EfficientNetV2-B2.

In Fig. 9, Inception-V3 also presents a matrix with high diagonal concentration, demonstrating good generalization capacity across the four classes.

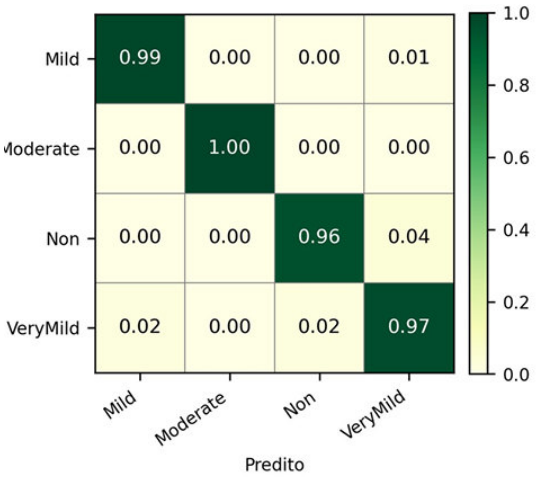


Figure 9. Normalized confusion matrix - Inception-V3.

In Fig. 10, EfficientNetV2-B0 presents near-total correct predictions for Mild and Moderate, but still registers errors between Non Demented and Very Mild Demented.

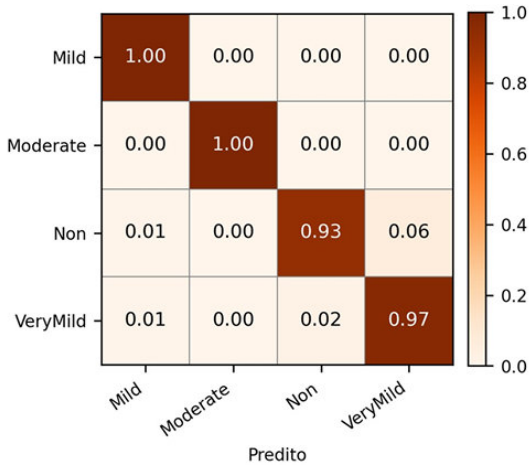


Figure 10. Normalized confusion matrix - EfficientNetV2-B0.

In Fig. 11, the simple soft voting ensemble reduces the errors observed in individual models, strengthening the main diagonal of the confusion matrix.

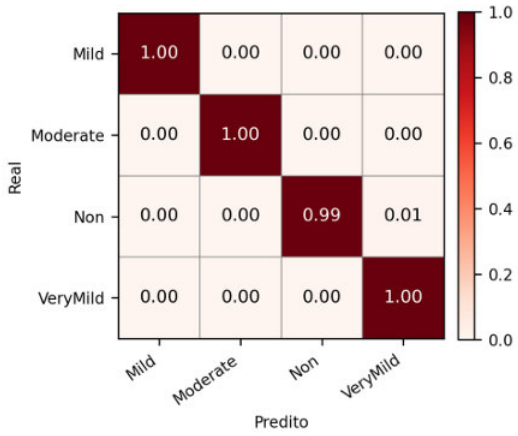


Figure 11. Normalized confusion matrix - Simple ensemble.

In Fig. 12, the weighted ensemble maintains equivalent performance to the simple ensemble, with near-total correct predictions and only minor residual confusion between adjacent classes.

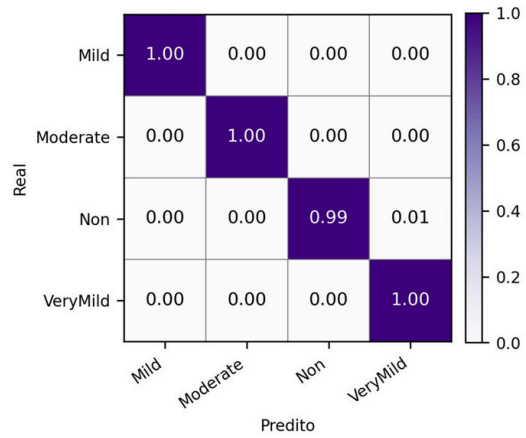


Figure 12. Normalized confusion matrix - Weighted ensemble.

E. External validation

External validation was used as a complementary step to verify model robustness on datasets different from those used in the main training runs. DenseNet-201 achieved a macro F1-score of 0.9856 on 32,000 external images, a performance close to that obtained internally. EfficientNetV2-B2 also maintained high performance, with a macro F1-score of 0.9852 on 8,000 external images.

These results (Table VI) reinforce that the more robust models not only fitted the main dataset but also maintained good generalization capacity on external datasets. Nevertheless, since external datasets may have different distributions, the comparison should be interpreted separately from the internal ranking.

Table VI
EXTERNAL VALIDATION RESULTS.

Model	Imgs.	Acc.	F1
DenseNet-201	32000	0.9843	0.9856
EffNetV2-B2	8000	0.9852	0.9852
EffNetV2-B0	8000	0.9774	0.9773
DenseNet-169	32000	0.9255	0.9289
Inception-V3	28000	0.9169	0.9166

In Fig. 13, the external validation of DenseNet-201 maintains high values on the main diagonal, indicating good generalization to images not used in the main training runs. Residual errors remain concentrated between Non Demented and Very Mild Demented.

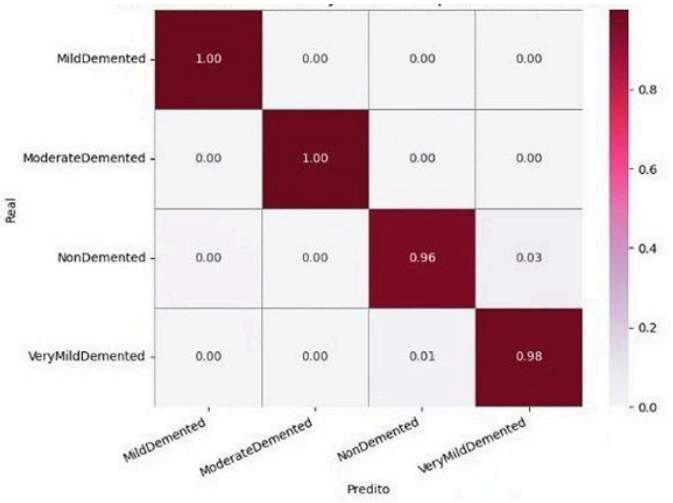


Figure 13. External validation of DenseNet-201.

In Fig. 14, EfficientNetV2-B2 also presents high external performance, with near-total correct predictions for Mild and Moderate classes. The remaining confusions occur mainly between Non and Very Mild, reinforcing that these classes are the most difficult in the problem.

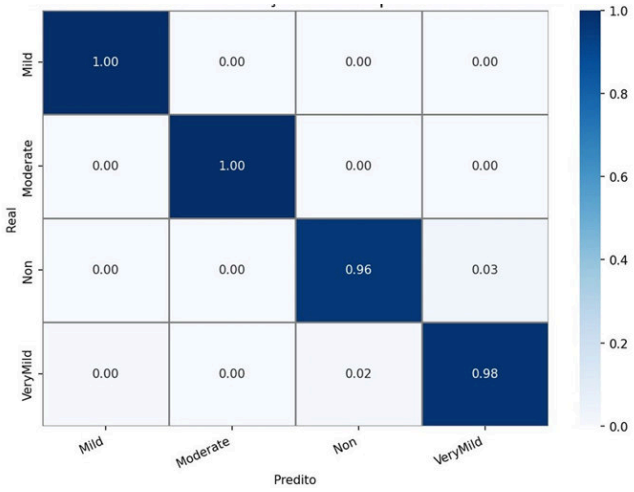


Figure 14. External validation of EfficientNetV2-B2.

F. Error analysis, robustness, and computational cost

Analysis of the confusion matrices revealed a recurring pattern: the Moderate Demented class was recognized with greater stability, while Non Demented and Very Mild Demented concentrated most errors. This difficulty suggests that the visual patterns between the absence of dementia and very mild signs may be close in the images used, requiring more sensitive models and careful validation.

Balancing techniques were not necessarily beneficial. Class weights and Focal Loss aimed to increase model attention to difficult classes, but in some variations there was a drop in macro F1-score. This indicates that artificial class penalization can shift learning and harm classes that were previously classified with greater confidence.

The best improvements occurred when fine-tuning was performed in a controlled manner. Partial fine-tuning allowed adapting the final blocks to the characteristics of MRI images without excessively altering the general representations learned from ImageNet. Full fine-tuning was effective in DenseNet-201 when combined with differentiated learning rates.

The ensemble showed the best overall performance (Table VII), but its use increases computational cost, as it requires loading and running multiple architectures. In real-world applications, this strategy may be indicated when the priority is to maximize performance. For scenarios with time or hardware constraints, the optimized EfficientNetV2-B2 represents a simpler individual alternative with very high performance.

Table VII
SUMMARY OF THE IMPACT OF EVALUATED STRATEGIES.

Technique	**Effect**	**Interpretation**
Controlled fine-tuning	Improved	Adapted to MRI domain
AdamW + regularization	Improved	Reduced overfitting
TTA	Stabilized	Reduced variance
Class weights	Worsened	Excessive penalization
Focal Loss	Worsened	Did not compensate
Ensemble	Best	Combined models

IV. CONCLUSION

This work evaluated multiple pre-trained CNN architectures for classifying four Alzheimer's classes in MRI images. The results demonstrated that simple baselines show limited performance, while transfer learning, fine-tuning, regularization, and TTA strategies significantly increase performance. The best individual model was the optimized EfficientNetV2-B2, with a macro F1-score of 0.9919, and the best overall result was obtained by the ensemble of five models, with a macro F1-score of 0.9959.

As a limitation, it should be noted that part of the experiments used image-level splitting rather than patient-level splitting, which may influence the generalization estimate. Additionally, the ensemble increases inference computational cost. As future work, it is recommended to standardize patient-level validation, broaden the diversity of external datasets, apply explainability techniques, and evaluate the deployment of models in controlled clinical scenarios.

REFERENCES

[1] World Health Organization, "Dementia," Fact sheet, 2025.
[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
[3] J. Wen et al., "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," Medical Image Analysis, 2020.
[4] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, pp. 211–252, 2015.
[5] A. Singhal, "Alzheimer's Disease Multi-class Images Dataset," Kaggle. Available at: https://www.kaggle.com/datasets/aryansinghal10/alzheimers-multiclass-dataset-equal-and-augmented?select=combined_images. Accessed: Jun. 10, 2026.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016.
[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017.
[8] C. Szegedy et al., "Rethinking the Inception architecture for computer vision," in Proc. IEEE CVPR, 2016.
[9] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. ICML, 2021.
[10] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

Análise Comparativa de Arquiteturas de CNNs com Técnicas de Pré-processamento para Classificação de Câncer Renal em Imagens Tomográficas

Thiago Cordeiro de Melo
Escola Superior de Tecnologia
Universidade do Estado do Amazonas
Manaus, Brasil
thiago.melo@uea.edu.br

Wheidima Carneiro de Melo
Escola Superior de Tecnologia
Universidade do Estado do Amazonas
Manaus, Brasil
wmelo@uea.edu.br

***Abstract*—O câncer renal é uma das neoplasias urológicas de maior mortalidade, com diagnóstico frequentemente tardio. Este trabalho investiga o impacto de duas técnicas de pré-processamento de contraste — CLAHE e correção gama (com fator gama igual a 0,7) — em comparação com a condição base sobre cinco arquiteturas de Redes Neurais Convolucionais com aprendizado por transferência: ResNet-18, ResNet-34, ResNet-50, AlexNet e VGG-16. O conjunto de dados empregado é o CT Kidney Dataset, com 4.566 imagens balanceadas entre as classes Normal e Tumor, divididas nas proporções 60/20/20 para treino, validação e teste. Os resultados demonstram que a correção gama melhora ou mantém estável o desempenho em todas as arquiteturas avaliadas, ao passo que o CLAHE apresenta comportamento inconsistente, reduzindo a acurácia das redes mais rasas. O melhor resultado global foi obtido pela VGG-16 com correção gama, que alcançou 99,12% de acurácia e revocação de 100% para a classe Tumor.**

***Index Terms*—Classificação de Câncer Renal, Redes Neurais Convolucionais, Equalização de Histograma, CLAHE, Correção Gama, Aprendizado Profundo, Imagens Tomográficas**

I. Introdução

O diagnóstico precoce configura-se como elemento determinante para o êxito terapêutico em oncologia, repercutindo de forma direta sobre a sobrevida e o bem-estar dos pacientes [1]. Dentre as neoplasias de maior relevância clínica, o câncer renal destaca-se como a neoplasia urológica de maior letalidade [2], manifestando-se predominantemente em indivíduos na faixa etária de 50 a 70 anos. No Brasil, foram registrados aproximadamente 13.000 novos casos e mais de 4.000 óbitos no ano de 2020 [3]. Não obstante, a inespecificidade sintomatológica nos estágios iniciais, associada à reduzida eficácia das estratégias de rastreamento populacional, compromete a identificação precoce da enfermidade [4]. Em razão disso, parcela considerável dos casos é diagnosticada em fases avançadas, o que restringe de maneira substancial as possibilidades de intervenção terapêutica eficaz.

As consequências desse diagnóstico tardio tornam-se mais graves em virtude do papel essencial exercido pelos rins no organismo, responsáveis por funções como a filtração de metabólitos, a regulação da pressão arterial e a produção de hormônios fundamentais, a exemplo da eritropoietina [5]. À medida que o câncer renal avança, essas funções podem ser severamente comprometidas, tornando necessários tratamentos invasivos como hemodiálise e nefrectomia, podendo a doença evoluir, nos casos mais graves, para o óbito [3]. Como o surgimento de sintomas físicos não pode ser utilizado como parâmetro confiável, uma vez que costumam aparecer apenas quando o órgão já se encontra gravemente comprometido, é fundamental que a investigação médica adote uma postura proativa. Nesse contexto, o desenvolvimento de ferramentas diagnósticas mais precisas, sobretudo aquelas fundamentadas em exames de imagem, configura-se como uma estratégia indispensável [6].

Na prática clínica, a tomografia computadorizada é o principal exame para a detecção e a caracterização de massas renais, frequentemente identificadas de forma incidental. Sua acurácia, contudo, é limitada: uma revisão sistemática de 40 estudos, abrangendo 4.354 pacientes, reportou especificidade mediana de 75% para a TC no diagnóstico do carcinoma de células renais, com intervalo interquartil de 51% a 90% [7]. A dificuldade decorre da sobreposição de características radiológicas: lesões benignas como o angiomiolipoma pobre em gordura e o oncocitoma podem exibir padrões tomográficos semelhantes aos do carcinoma, levando a nefrectomias desnecessárias [8]. Exames amplamente disponíveis, porém de interpretação ambígua, evidenciam a lacuna que métodos computacionais de apoio à decisão podem preencher.

Diante dessa necessidade de métodos diagnósticos mais precisos e objetivos, o aprendizado profundo (*deep learning*) tem se consolidado como uma abordagem promissora para a análise automatizada de imagens médicas [9]. Dentre essas técnicas, as Redes Neurais Convolucionais (CNNs) destacam-se pela capacidade de extrair automaticamente características discriminativas a partir de dados visuais, identificando padrões sutis frequentemente imperceptíveis ao olho humano [10]. No contexto específico do câncer renal, modelos baseados em CNNs vêm sendo empregados em tarefas como a segmentação de rins e lesões tumorais [11] e a classificação de massas renais em exames de tomografia computadorizada [8], [12], contribuindo para diagnósticos mais rápidos e menos dependentes

da experiência subjetiva do especialista.

Apesar dos recentes avanços, a eficácia das CNNs na classificação de imagens médicas permanece fortemente condicionada à qualidade dos dados de entrada. Em exames de tomografia computadorizada, fatores como baixo contraste entre tecidos e a presença de ruído podem comprometer a extração de características relevantes pelos modelos, impactando negativamente o desempenho diagnóstico. Diante disso, técnicas de pré-processamento de imagem surgem como uma estratégia promissora para mitigar essas limitações, sendo capazes de realçar estruturas anatômicas e regiões de interesse antes da etapa de classificação. Contudo, ainda são escassos os estudos que avaliam de forma sistemática o impacto desse tipo de pré-processamento sobre diferentes arquiteturas de CNNs aplicadas à classificação de massas renais, especialmente quando se compara o desempenho de modelos consolidados, como as variantes da família ResNet [13].

Dentre as técnicas de pré-processamento disponíveis, destacam-se o CLAHE (*Contrast Limited Adaptive Histogram Equalization*) [14] e a correção gama [15], amplamente empregadas para aprimorar a qualidade visual de imagens médicas. Enquanto o CLAHE promove o realce local do contraste, preservando detalhes anatômicos e limitando a amplificação de ruídos, a correção gama atua sobre a distribuição de intensidades dos pixels, possibilitando evidenciar regiões de interesse por meio do ajuste não linear do brilho. Essas características tornam ambas as técnicas particularmente adequadas para exames de tomografia computadorizada, nos quais a distinção entre tecidos renais saudáveis e regiões tumorais depende de sutis variações de contraste. Nesse contexto, o presente trabalho propõe uma análise comparativa do desempenho de cinco arquiteturas de CNNs — ResNet-18, ResNet-34, ResNet-50, AlexNet e VGG-16 — em três cenários de pré-processamento: sem realce de contraste, com aplicação de CLAHE e com aplicação de correção gama, visando quantificar o impacto dessas técnicas sobre a capacidade de classificação dos modelos e contribuir para o desenvolvimento de uma ferramenta clínica de apoio ao diagnóstico precoce do câncer renal.

O restante deste artigo está organizado da seguinte forma: a Seção II apresenta trabalhos relacionados; a Seção III descreve os materiais e métodos; a Seção IV apresenta e discute os resultados; a Seção V conclui o trabalho.

II. Trabalhos Relacionados

A utilização de redes neurais convolucionais para classificação de imagens médicas renais tem recebido crescente atenção na literatura, com diferentes grupos propondo abordagens baseadas em aprendizado profundo e aprendizado por transferência. Os trabalhos a seguir representam contribuições diretamente relacionadas ao problema investigado neste artigo.

Santos et al. [2] propuseram um sistema de classificação de câncer renal em imagens de tomografia computadorizada baseado em CNNs com aprendizado por transferência. O sistema utiliza o CT Kidney Dataset, composto por imagens das classes Normal e Tumor, com divisão estratificada em conjuntos de treino, validação e teste. O pré-processamento limitou-se à conversão das imagens para escala de cinza em 8 bits, sem aplicação de técnicas adicionais de realce de contraste.

Rajkumar et al. [16] propuseram a detecção de câncer renal utilizando modelos de aprendizado profundo a partir de imagens da mesma base Kaggle. O modelo CNN desenvolvido atingiu acurácia de treinamento de 96,4% e acurácia de validação de 99,6%.

Alzu'bi et al. [17] avaliaram arquiteturas como VGG-16, ResNet-50, CNN-6 e CNN-4 para detecção e classificação de tumores renais em tomografias. O melhor resultado foi obtido pelo modelo CNN-4, com 97,70% de acurácia. O trabalho enfatiza a importância da qualidade do pré-processamento para o desempenho dos classificadores.

Hadjiyski [18] empregou a rede Inception V3 para estadiamento de câncer renal utilizando imagens de TC em 3D, das quais fatias 2D foram extraídas. Após 4.000 épocas de treinamento, a acurácia atingiu 97% no treino e 90% no teste.

Diferentemente dos trabalhos citados, o presente estudo conduz um experimento fatorial controlado entre técnicas de pré-processamento e arquiteturas CNN, com protocolo de divisão de dados replicável e semente fixada.

III. Materiais e Métodos

A Fig. 1 apresenta uma visão geral da metodologia adotada, composta por cinco etapas: (*i*) organização do *dataset* em treino, validação e teste; (*ii*) aplicação das condições de pré-processamento (Base, CLAHE e Gama); (*iii*) extração de características pelas arquiteturas CNN avaliadas; (*iv*) classificação binária entre Normal e Tumor; e (*v*) avaliação por métricas de desempenho. Cada etapa é detalhada nas subseções a seguir.[1]

A. Conjunto de Dados

O conjunto de dados utilizado é o *CT Kidney Dataset: Normal-Cyst-Tumor and Stone* [19], disponível na plataforma Kaggle. O dataset contém 12.446 imagens únicas de tomografia computadorizada distribuídas em quatro classes: cistos (3.709), rins normais (5.077), pedras (1.377) e tumor (2.283). Para este trabalho, foram utilizadas apenas as classes *Normal* e *Tumor*.

B. Arquiteturas de CNNs

Foram avaliadas cinco arquiteturas de CNNs pré-treinadas na base ImageNet, todas consolidadas na literatura de visão computacional. Em todos os modelos adotou-se a estratégia de aprendizado por transferência: as camadas convolucionais, responsáveis pela extração de características, foram mantidas congeladas com os pesos pré-treinados, e apenas a camada de classificação final foi substituída por uma nova camada totalmente conectada com duas saídas, correspondentes às classes Normal e Tumor, sendo a única treinada para a tarefa em questão.

[1]O código-fonte completo dos experimentos — incluindo os *scripts* de organização do conjunto de dados, pré-processamento, treinamento e avaliação — está disponível publicamente em https://github.com/thiagocordeirum/kidney-ct-cnn-classification.

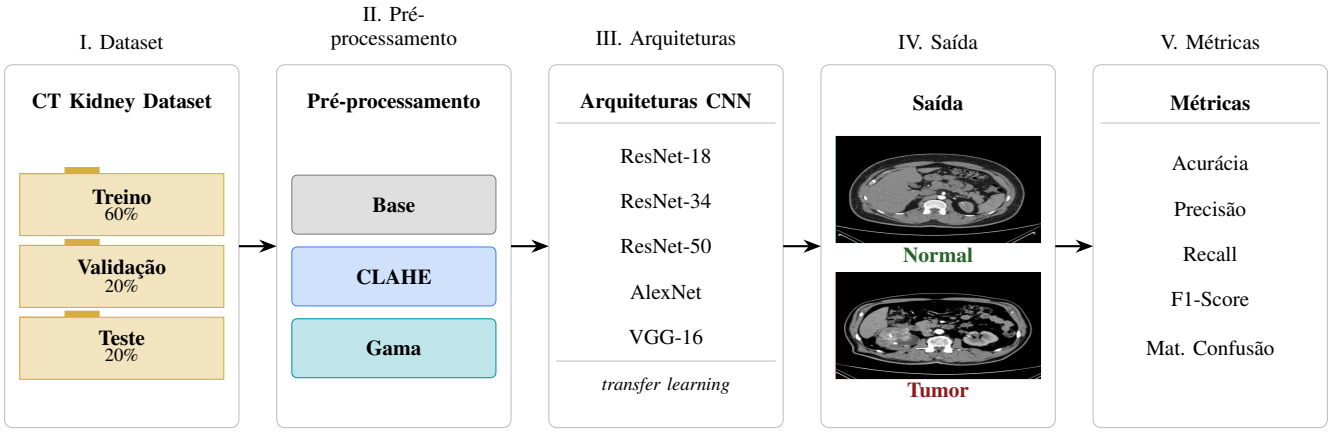


Fig. 1. Diagrama geral da metodologia proposta.

As redes da família *ResNet* (*Residual Neural Network*) foram introduzidas por He et al. [13] em 2016 e caracterizam-se pelo empilhamento de blocos residuais com conexões de atalho, que somam a entrada de um bloco diretamente à sua saída. Esse mecanismo viabiliza o treinamento de redes muito profundas sem a degradação provocada pelo desaparecimento do gradiente. Neste trabalho foram avaliadas três profundidades: a ResNet-18 e a ResNet-34, com 18 e 34 camadas organizadas em blocos residuais simples, e a ResNet-50, com 50 camadas estruturadas em blocos do tipo gargalo, que ampliam a capacidade de representação com menor custo computacional.

A *AlexNet* [20], vencedora do desafio ImageNet de 2012 por ampla margem, é composta por cinco camadas convolucionais seguidas de três camadas totalmente conectadas. Foi a primeira arquitetura a demonstrar a viabilidade da extração automática de características em larga escala, rompendo o paradigma dos descritores projetados manualmente, e introduziu o uso da função de ativação ReLU e o treinamento acelerado por unidades de processamento gráfico.

A *VGG-16* [21], desenvolvida por Simonyan e Zisserman na Universidade de Oxford, deve seu nome ao grupo *Visual Geometry Group*. Sua principal característica é a profundidade obtida pelo empilhamento sucessivo de pequenos filtros convolucionais de tamanho 3×3, totalizando 16 camadas com pesos — 13 convolucionais e três totalmente conectadas. Originalmente projetada para classificar imagens em até mil categorias, recebe imagens na resolução de 224×224 pixels.

C. Critérios de Inclusão e Exclusão

Foram incluídas as classes *Normal* e *Tumor*, excluindo-se *Cisto* e *Pedra* por não serem relevantes ao objetivo de discriminação entre tecido renal saudável e tumoral. Para garantir equilíbrio entre as classes, foram selecionadas aleatoriamente 2.283 imagens da classe Normal (correspondendo ao total disponível da classe Tumor), totalizando 4.566 imagens. O conjunto balanceado foi dividido de forma estratificada nas proporções de 60% treino (2.738 imagens), 20% validação (914 imagens) e 20% teste (914 imagens), com semente aleatória s=42 para reprodutibilidade. A Fig. 1 ilustra o fluxo completo desde a seleção dos dados até a avaliação.

D. Pré-processamento

Foram avaliadas três condições de pré-processamento, mantendo um grupo base sem realce de contraste para fins de comparação.

Base (sem realce). As imagens são convertidas para escala de cinza em 8 bits e replicadas nos três canais RGB para compatibilidade com os pesos pré-treinados no ImageNet.

CLAHE (*Contrast Limited Adaptive Histogram Equalization*). Técnica que divide a imagem em $M_x\times M_y$ *tiles* locais e equaliza o histograma de cada *tile* com limitação do contraste máximo. Seja $I:\Omega\to\{0,\dots,L{-}1\}$ uma imagem com N pixels e $L=256$ níveis, o histograma normalizado é:

$$p(r)=\frac{h(r)}{N},\quad r\in\{0,\dots,L-1\}. \tag{1}$$

A equalização global obtém a CDF $F(r)=\sum_{s=0}^{r}p(s)$ e aplica $T(r)=\lfloor(L{-}1)\cdot F(r)\rfloor$. Na CLAHE, o histograma local de cada *tile* $\mathcal{T}_{ij}$ é recortado no limite $C=\beta\,N_t/L$, em que $N_t=w\,h$ é o número de pixels do *tile* e β o fator de recorte de contraste:

$$h_c(r)=\min\big(h_{ij}(r),\,C\big), \tag{2}$$

sendo o excesso redistribuído uniformemente antes da equalização local, com interpolação bilinear nas bordas dos *tiles*. Neste trabalho adotou-se fator de recorte β=2,0 e divisão da imagem em uma grade de 8×8 *tiles* [14].

Correção Gama. Aplica um mapeamento não-linear pixel a pixel:

$$I_{\text{out}}(x,y)=\left(\frac{I_{\text{in}}(x,y)}{255}\right)^{\gamma}\times\,255, \tag{3}$$

com γ=0,7. Para $\gamma<1$, a curva clareia seletivamente regiões escuras sem saturar as já contrastadas, evidenciando padrões texturais de baixo contraste característicos de massas tumorais em TC [15].

Após a técnica selecionada, todas as imagens seguem o mesmo fluxo final: conversão para RGB, redimensionamento para 224×224 px e normalização canal a canal com os parâmetros ImageNet:

$$\tilde{x}_c = \frac{x_c - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\}, \tag{4}$$

com $\boldsymbol{\mu}$=[0,485, 0,456, 0,406] e $\boldsymbol{\sigma}$=[0,229, 0,224, 0,225].

E. Parâmetros de Treinamento e Avaliação

O treinamento foi conduzido por 100 épocas, com otimizador Adam, taxa de aprendizado de $1{,}0 \times 10^{-4}$, lote de 32 amostras e semente aleatória s=42. Como função de perda adotou-se a entropia cruzada (*Cross-Entropy*) com duas saídas, sendo a predição final dada pela classe de maior probabilidade. Os pesos correspondentes à época de maior acurácia de validação foram retidos e empregados na avaliação sobre o conjunto de teste.

F. Avaliação dos Modelos

Os modelos foram avaliados no conjunto de teste (914 amostras), com classe positiva definida como Tumor (TP: verdadeiros positivos; TN: verdadeiros negativos; FP: falsos positivos; FN: falsos negativos):

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$\text{Precisão} = \frac{TP}{TP + FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{7}$$

$$F_1 = 2 \cdot \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}, \tag{8}$$

$$\mathcal{L}(y, \hat{p}) = -\sum_{c \in \{0,1\}} y_c \log \hat{p}_c. \tag{9}$$

Adicionalmente, foram analisadas matrizes de confusão e curvas de acurácia e perda por época para as condições CLAHE e Gama.

G. Recursos Computacionais

Os experimentos foram executados em um servidor equipado com o NVIDIA GB10 Grace Blackwell Superchip, que integra processador ARM Cortex-X925 (10 núcleos) e Cortex-A725 (10 núcleos) com 128 GB de memória unificada compartilhada entre CPU e GPU, eliminando transferências explícitas entre os dois componentes. O ambiente de software utilizou Python 3.10.20, PyTorch 2.12.0, TorchVision 0.27.0 (ambos com suporte a CUDA 13.0), OpenCV, NumPy, Pandas, Scikit-Learn e Matplotlib.

TABLE I
MÉTRICAS NO CONJUNTO DE TESTE (914 AMOSTRAS)

Modelo	Cond.	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
ResNet-18	Base	96,94	**97,35**	96,50	96,92
	CLAHE	95,19	95,38	94,97	95,18
	Gama	**97,48**	97,17	**97,81**	**97,49**
ResNet-34	Base	95,19	95,58	94,75	95,16
	CLAHE	95,40	95,21	**95,62**	95,41
	Gama	**96,06**	**97,30**	94,75	**96,01**
ResNet-50	Base	95,73	95,43	96,06	95,75
	CLAHE	96,06	**96,67**	95,40	96,04
	Gama	**96,50**	96,50	**96,50**	**96,50**
AlexNet	Base	98,91	**98,48**	99,34	98,91
	CLAHE	97,16	96,95	97,37	97,16
	Gama	**99,02**	98,28	**99,78**	**99,02**
VGG-16	Base	98,91	98,06	99,78	98,92
	CLAHE	98,36	**99,11**	97,59	98,35
	Gama	**99,12**	98,28	**100,00**	**99,13**

IV. RESULTADOS E DISCUSSÃO

A. Resultados no Conjunto de Teste

A Tabela I apresenta as métricas de teste dos 15 experimentos. Resultados em negrito indicam o melhor valor por modelo.

B. Impacto do Pré-processamento

Arquiteturas ResNet. A correção gama (γ=0,7) beneficiou consistentemente todas as arquiteturas ResNet, com ganhos de $+0{,}54$ p.p. (ResNet-18), $+0{,}87$ p.p. (ResNet-34) e $+0{,}77$ p.p. (ResNet-50) sobre a condição base. O mapeamento não linear da (3) clareia regiões escuras correspondentes a tecido tumoral de baixo contraste espontâneo, evidenciando padrões que os blocos residuais exploram com maior eficiência. O CLAHE produziu queda de $-1{,}75$ p.p. na ResNet-18 e ganhos discretos nas demais ($+0{,}21$ p.p. na ResNet-34 e $+0{,}33$ p.p. na ResNet-50), sugerindo que os artefatos de borda entre *tiles* perturbam sobretudo as ativações das primeiras camadas convolucionais das redes mais rasas.

AlexNet e VGG-16. AlexNet e VGG-16 atingiram $98{,}91\%$ na condição base, superando as ResNets (entre 95% e 97%). O CLAHE causou queda de $-1{,}75$ p.p. na AlexNet ($97{,}16\%$) e queda menor de $-0{,}55$ p.p. na VGG-16 ($98{,}36\%$). A correção gama, por ser uma transformação monotônica suave, elevou o desempenho de ambas: a AlexNet atingiu $99{,}02\%$ e a VGG-16 alcançou **99,12%**, melhor resultado global do experimento, com revocação de 100% na classe Tumor. O comportamento é atribuído ao classificador denso de três camadas totalmente conectadas presente em ambas as redes: o CLAHE altera a distribuição de ativações de forma incompatível com os padrões aprendidos na base ImageNet, enquanto a correção gama preserva a ordem relativa das intensidades.

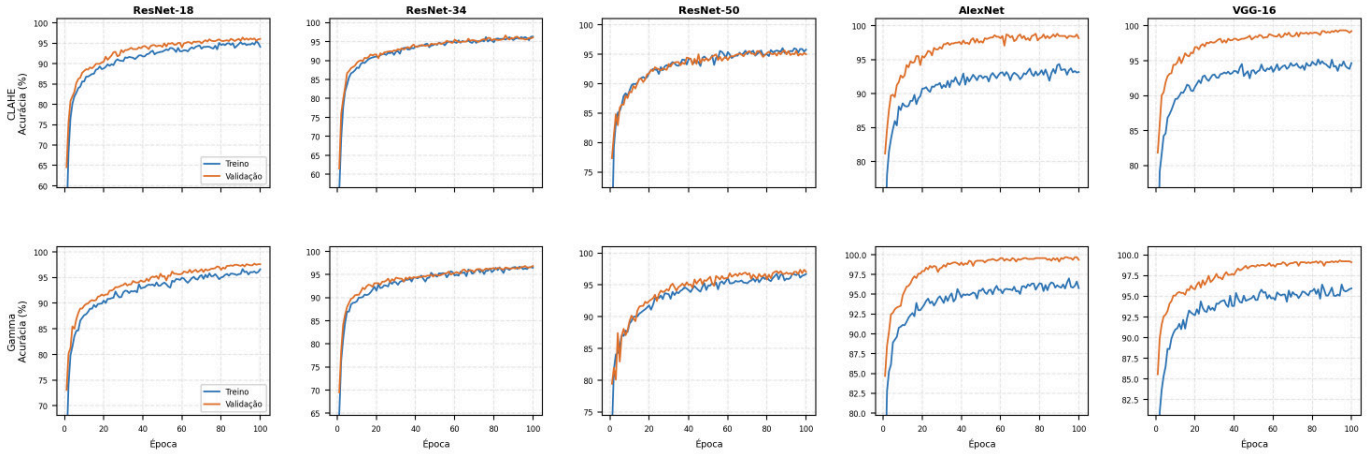


Fig. 2. Curvas de acurácia por época (100 épocas). Linha azul: treino. Linha laranja: validação. Linha superior: condição CLAHE. Linha inferior: condição Gama.

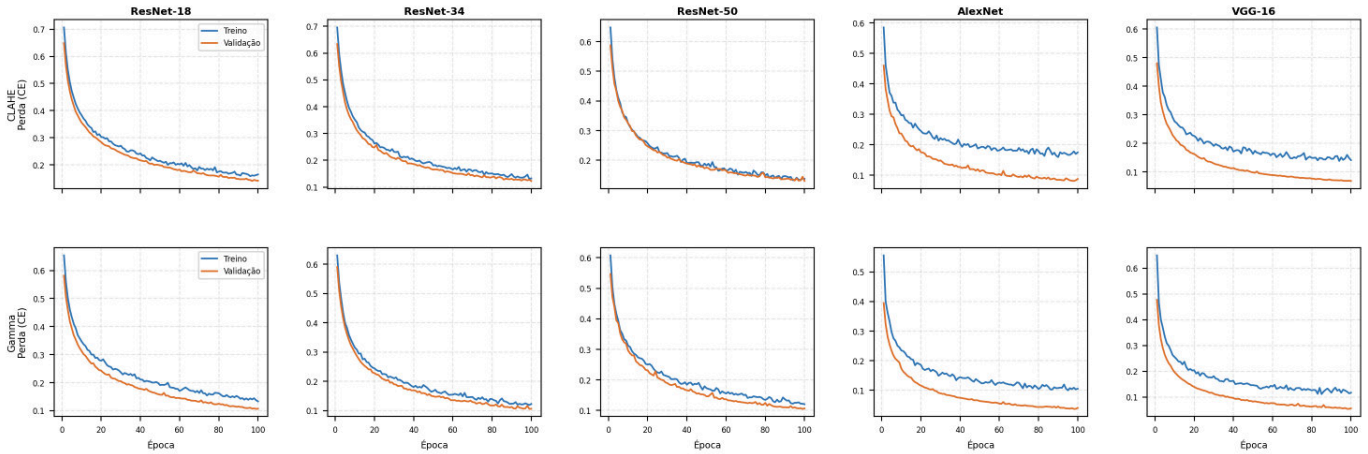


Fig. 3. Curvas de perda (*Cross-Entropy*) por época. Linha azul: treino. Linha laranja: validação. Linha superior: CLAHE. Linha inferior: Gama.

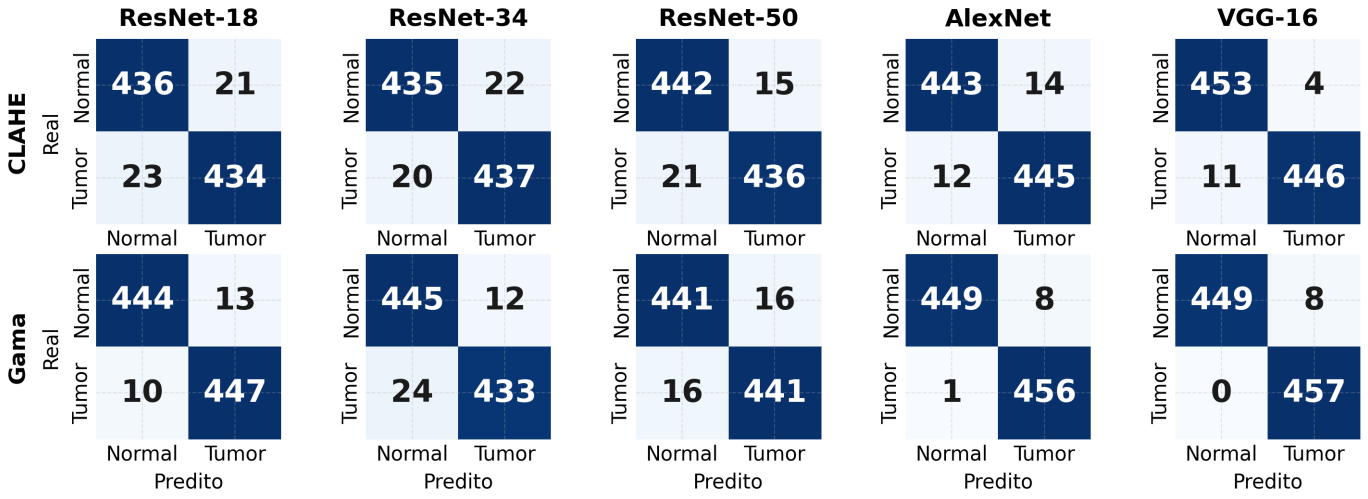


Fig. 4. Matrizes de confusão no conjunto de teste (914 amostras). Linha superior: CLAHE. Linha inferior: Gama. Colunas: ResNet-18, ResNet-34, ResNet-50, AlexNet, VGG-16.

C. Curvas de Treinamento

As Figs. 2 e 3 mostram as curvas de acurácia e perda por época para as condições CLAHE e Gama. Para as ResNets, o Gama antecipa a convergência e eleva o *plateau* de validação, enquanto o CLAHE apresenta maior variabilidade nas épocas intermediárias. Para AlexNet e VGG-16, ambas as condições alcançam acurácias elevadas nas primeiras épocas, evidenciando a eficácia do *transfer learning* mesmo com *backbone* congelado. Em todas as arquiteturas, o reduzido intervalo entre as curvas de treino e validação indica boa capacidade de generalização, sem sinais expressivos de sobreajuste ao longo das 100 épocas avaliadas.

D. Matrizes de Confusão

A Fig. 4 apresenta as matrizes de confusão no conjunto de teste para todas as combinações modelo $\times$ condição (CLAHE e Gama). Para as ResNets, o Gama reduz os falsos negativos (Tumor classificado como Normal) em relação ao CLAHE, propriedade clinicamente relevante em rastreamento oncológico, uma vez que erros nessa direção correspondem a diagnósticos perdidos. AlexNet e VGG-16 apresentam erros próximos de zero na condição base e resultados competitivos com o Gama; em particular, a VGG-16 com correção gama atingiu revocação de 100% na classe Tumor, eliminando completamente os falsos negativos no conjunto de teste.

V. Conclusão

Este trabalho investigou sistematicamente o impacto das técnicas de pré-processamento CLAHE e correção gama sobre o desempenho de cinco arquiteturas CNN na classificação de câncer renal em imagens tomográficas.

A correção gama com $\gamma{=}0{,}7$ demonstrou ser a técnica de pré-processamento mais consistente entre as avaliadas, sendo a única que beneficiou todas as cinco arquiteturas sem exceção: os ganhos variaram de $+0{,}11$ p.p. na AlexNet a $+0{,}87$ p.p. na ResNet-34. O melhor resultado global foi obtido pela VGG-16 com correção gama, que alcançou $99{,}12\%$ de acurácia e revocação de 100% na classe Tumor, evidenciando o potencial da técnica para reduzir falsos negativos em cenários de rastreamento oncológico.

O CLAHE, por sua vez, apresentou comportamento variável: reduziu o desempenho da ResNet-18 e da AlexNet em $-1{,}75$ p.p. cada, enquanto produziu ganhos discretos nas ResNets mais profundas. Esse comportamento sugere que a equalização adaptativa por *tiles* pode introduzir artefatos de borda incompatíveis com os filtros aprendidos na base ImageNet, com impacto mais acentuado em redes com menor profundidade ou maior sensibilidade à distribuição de intensidades.

Como trabalhos futuros, propõe-se avaliar diferentes valores de γ por arquitetura, estender a classificação para as classes Cisto e Pedra presentes no dataset, e investigar estratégias de ajuste fino parcial das camadas convolucionais.

Agradecimentos

Os autores agradecem ao Laboratório LSI pela disponibilização dos recursos computacionais utilizados na condução dos experimentos, sem os quais o treinamento das arquiteturas avaliadas não seria viável no escopo deste trabalho.

References

[1] C. A. Gomes, M. E. C. P. de Mello, R. S. F. Alves, G. B. Vaz, R. W. Lerner, J. M. de Queiroz Salek, C. C. Norberto, e M. V. Copello, "Qualidade de vida em pacientes oncológicos: impactos do tratamento e estratégias de manejo," *Brazilian Journal of Implantology and Health Sciences*, v. 7, n. 10, pp. 739–750, 2025.

[2] V. M. P. Santos, A. C. Patrocinio, W. C. S. Carvalho, e P. M. de Sousa, "Renal cancer classification in tomographic images using convolutional neural networks to assist early diagnosis," *Revista de Informática Teórica e Aplicada (RITA)*, v. 32, n. 1, pp. 61–66, 2025.

[3] Instituto Nacional de Câncer (INCA), "Estimativa 2023: incidência de câncer no Brasil," Rio de Janeiro: INCA, 2022.

[4] T. Powles, L. Albiges, A. Bex, E. Comperat, V. Grünwald, R. Kanesvaran, H. Kitamura, R. McKay, C. Porta, G. Procopio, M. Schmidinger, C. Suarez, J. Teoh, G. de Velasco, M. Young, e S. Gillessen, "Renal cell carcinoma: ESMO clinical practice guidelines," *Annals of Oncology*, v. 32, n. 11, pp. 1347–1359, 2021.

[5] J. E. Hall e M. E. Hall, *Guyton & Hall Tratado de Fisiologia Médica*, 14. ed. Rio de Janeiro: Elsevier, 2021.

[6] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, e S. Bakas, "Multi-institutional deep learning modeling without sharing patient data," *Nature Communications*, v. 11, n. 1, p. 5021, 2020.

[7] C. Vogel, B. Ziegelmüller, B. Ljungberg, K. Bensalah, A. Bex, S. Canfield, R. H. Giles, M. Hora, M. A. Kuczyk, A. S. Merseburger, T. Powles, L. Albiges, F. Stewart, A. Volpe, A. Graser, M. Schlemmer, C. Yuan, T. Lam, e M. Staehler, "Imaging in suspected renal-cell carcinoma: systematic review," *Clinical Genitourinary Cancer*, v. 17, n. 2, pp. e345–e355, 2019.

[8] M. E. Klontzas, G. Kalarakis, E. Koltsakis, T. Papathomas, A. H. Karantanas, e A. Tzortzakakis, "Convolutional neural networks for the differentiation between benign and malignant renal tumors with a multicenter international computed tomography dataset," *Insights into Imaging*, v. 15, art. 26, 2024.

[9] Y. LeCun, Y. Bengio, e G. Hinton, "Deep learning," *Nature*, v. 521, pp. 436–444, 2015.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, e C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, v. 42, pp. 60–88, 2017.

[11] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, et al., "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Medical Image Analysis*, v. 67, art. 101821, 2021.

[12] K.-H. Uhm, S.-W. Jung, M. H. Choi, H.-K. Shin, J.-I. Yoo, S. W. Oh, J. Y. Kim, H. G. Kim, Y. J. Lee, S. Y. Youn, S.-H. Hong, e S.-J. Ko, "Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography," *npj Precision Oncology*, v. 5, art. 54, 2021.

[13] K. He, X. Zhang, S. Ren, e J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[14] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, e K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, v. 39, n. 3, pp. 355–368, 1987.

[15] R. C. Gonzalez e R. E. Woods, *Digital Image Processing*, 4. ed. New York: Pearson, 2018.

[16] K. Rajkumar, R. T. S. Ramoju, T. Balelly, S. Ashadapu, Ch. R. Prasad, e Y. Srikanth, "Kidney cancer detection using deep learning models," in *Proc. ICOEI*, IEEE, 2023, pp. 1197–1203.

[17] D. Alzu'bi, M. Abdullah, I. Hmeidi, R. AlAzab, M. Gharaibeh, M. El-Heis, K. H. Almotairi, A. Forestiero, A. M. Hussein, e L. Abualigah, "Kidney tumor detection and classification based on deep learning," *Journal of Healthcare Engineering*, v. 2022, p. 3861161, 2022.

[18] N. Hadjiyski, "Kidney cancer staging: deep learning neural network based approach," in *Proc. EHB*, IEEE, 2020, pp. 1–4.

[19] “CT Kidney Dataset: Normal-Cyst-Tumor and Stone,” Kaggle, 2022. Disponível em: kaggle.com/datasets/nazmul0087/ ct-kidney-dataset-normal-cyst-tumor-and-stone.
[20] A. Krizhevsky, I. Sutskever, e G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NeurIPS*, 2012, pp. 1097–1105.
[21] K. Simonyan e A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR*, 2015.

Collaborative Multi-Agent Networks for Multimodal Analysis of Parkinson's Disease

1st Arthur Guilherme da Frota Brito Machado
Programa de Pós-Graduação em Engenharia Elétrica
Federal University of Ceara (UFC)
Fortaleza, Brazil
arthur.frota@alu.ufc.br

2st Laís Ferreira Sombra
Departamento de Engenharia Elétrica (DEE)
Federal University of Ceara (UFC)
Fortaleza, Brazil
laisfsombra@ieee.org

3nd Bruna Stefanie Costa Lourenço
Departamento de Teleinformática (DETI)
Federal University of Ceara (UFC)
Fortaleza, Brazil
brunastefanie@ufc.com.br

4rd Victor Hugo C. de Albuquerque
Departamento de Teleinformática (DETI)
Federal University of Ceara (UFC)
Fortaleza, Brazil
victor.albuquerque@ieee.org

***Abstract*—Parkinson's disease is a progressive and incurable neurodegenerative disease that affects millions of people worldwide. Its diagnosis is often time-consuming and expensive, especially in the early stages, which can delay appropriate medical intervention. To increase the accuracy and efficiency of Parkinson's diagnosis, a multi-agent system that leverages machine learning techniques — Random Forest, Neural Networks, and Logistic Regression — was implemented along with optimization methods, including LoRA and Knowledge Distillation. In addition, judging agents were employed to evaluate and compare the results, allowing a voting mechanism to determine the final decision. At the end of the process, it was possible to obtain a 50% reduction in training time when using the LoRA and Knowledge Distillation methods, with only the models that use Logistic Regression having a significant difference in training times. In addition, the accuracy of the methods always remains within a margin of 0.1 either way in all processes, which indicates that, even with the reduction in size for optimization, the models are able to maintain good results. This methodology facilitates the automated selection of the most effective diagnostic approaches, as evidenced by the weighted averages of the agents applied to different data types. Despite initial inefficiencies in identifying Healthy cases, the final aggregated result demonstrates improved diagnostic performance. Consequently, the proposed approach offers a scalable and adaptable solution for the diagnosis of Parkinson's disease, especially in resource-limited settings, highlighting its potential for broader clinical applications.**

***Index Terms*—Multi-Agent, Parkinson's Disease, LoRA, Knowledge Distillation, Random Forest, Neural Network, Logistic Regression**

I. Introduction

According to the World Health Organization (WHO) [1], in 2019, approximately 8.5 million individuals were diagnosed with Parkinson's disease (PD), with an observed increase of over 80% in PD-related mortality since 1990. This trend establishes PD as the fastest-growing neurological disorder worldwide. Furthermore, the WHO emphasizes that Parkinson's disease remains incurable, with available treatments limited to alleviating symptoms and improving patients' quality of life. However, these treatments are not only highly expensive but also incapable of halting disease progression. To address this critical issue, in 2022, the WHO approved a global initiative to promote research efforts focused on Parkinson's disease.

In alignment with this initiative, the present study employs a multi-agent system to enhance the prediction of Parkinson's disease. Four distinct machine learning models were utilized: Random Forest [2], Multi-Layer Perceptron (MLP) Classifier [3], and Logistic Regression [4], each integrated with Principal Component Analysis (PCA) [5] for dimensionality reduction.

Advancements in Parkinson's disease progression prediction have been achieved by integrating diverse data modalities, including clinical, genetic, biomarker, and neuroimaging information. Lei et al. [6] proposed a feature selection model that combines structural magnetic resonance imaging (sMRI) and diffusion tensor imaging (DTI) to predict clinical scores in PD patients. Their methodology, based on sparse multimodal machine learning, demonstrated superior performance compared to single-modality approaches, particularly in predicting sleep disturbances and olfactory impairments. This underscores the efficacy of multimodal data integration in PD prognosis.

Sandhiya et al. [7] investigated PD prediction using motor and speech patterns, employing image processing techniques alongside Random Forest classification. While their model achieved promising accuracy in distinguishing PD patients from healthy individuals, its reliance on a limited set of motor-based features restricts its applicability for disease progression prediction. This study highlights both the potential and limitations of machine learning approaches in PD diagnostics.

Jiang et al. [8] conducted a comprehensive review on multimodal data integration for PD progression prediction, demonstrating that combining multiple data sources significantly enhances predictive accuracy compared to single-modality models. However, they identified key challenges, including data heterogeneity and the necessity for advanced data fusion techniques. Despite these obstacles, such as high computational costs and the requirement for extensive labeled datasets, the study emphasizes the importance of robust, integrative modeling strategies.

The use of an intelligent multi-agent ecosystem in the diagnosis of Parkinson's disease holds great promise, as it enables the integration of data from multiple sources, such as imaging exams, clinical records, and motion sensors, providing a more

precise and comprehensive analysis of the patient's condition. Multi-agent systems play a crucial role in early detection by analyzing motor patterns, interpreting neurological exams, and evaluating therapeutic responses, making the diagnosis faster and more reliable. Recent advances in AI [9], including deep learning for medical imaging and sensor-based monitoring, have further enhanced these capabilities, enabling continuous symptom tracking and improved diagnostic accuracy.

With this adaptive approach, treatments can be personalized in real time, optimizing interventions and improving patients' quality of life. However, challenges remain in integrating heterogeneous data sources and ensuring model interpretability.

Accordingly, this research aims to address these challenges by developing predictive models for early PD diagnosis and treatment optimization. By leveraging advanced machine learning techniques and multimodal data integration, this study not only contributes to the scientific community but also has significant practical implications, facilitating improved clinical decision-making and patient outcomes.

II. Database:

For the execution of this project, the dataset was sourced from the Image and Data Archive (IDA)[1], managed by the Laboratory of Neuro Imaging (LONI) at the USC Mark and Mary Stevens Neuroimaging and Informatics Institute. From this repository, patient information, biospecimens, family history, and polygenic risk scores were extracted for analysis.

This dataset presented two major challenges: a significant amount of missing data and a pronounced class imbalance between instances labeled as Parkinson's Disease and Not Parkinson.

To address the issue of missing data, the K-Nearest Neighbors (KNN) imputation technique was applied. This method operates under the assumption that similar values are spatially close to one another in a multidimensional space. For each missing entry, the algorithm identifies the closest neighbors based on the Euclidean distance and imputes the missing value using a weighted average, mode, or median—whichever is most appropriate. This approach preserves valuable information and increases the number of usable samples, as opposed to simply discarding incomplete rows.

To mitigate the class imbalance, the SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors) method was employed. This approach integrates SMOTE, which generates synthetic samples for the minority class, with ENN, which removes noisy or borderline instances from the majority class. The application of this technique significantly improved the class distribution across all three data types, as illustrated in Figures 1, 2, and 3.

[1]Access link: https://ida.loni.usc.edu/login.jsp

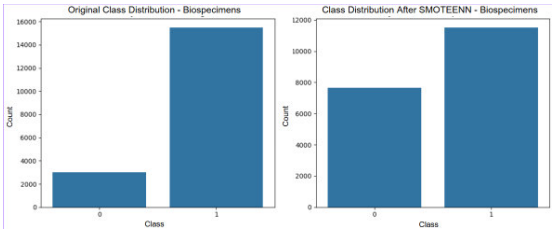


Fig. 1. Biospecimens Class Balance Comparison



Fig. 2. Biospecimens Class Balance Comparison

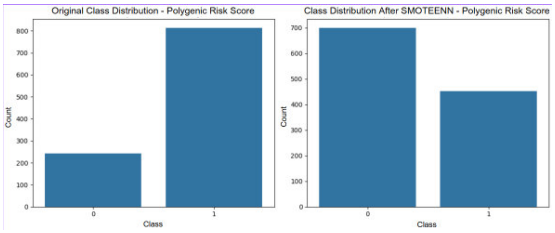


Fig. 3. Polygenic Risk Score Class Balance Comparison

A. Patient Information

This dataset segment categorizes patients into four groups: Healthy Control (HC), SWEDD (Scans Without Evidence of Dopamine Deficit), Prodromal (individuals in the early stages of Parkinson's disease), and Parkinson's Disease (PD) patients.

The SWEDD group [10] comprises individuals who, despite being clinically diagnosed with Parkinson's disease, do not exhibit abnormalities in neuroimaging exams. There is ongoing debate regarding these cases—some researchers suggest that different neurological disorders may be responsible for the diagnosis, while others propose that SWEDD represents a benign variant of Parkinson's disease. However, there is no consensus on a definitive classification. To avoid potential controversy, IDs associated with SWEDD were excluded from this study.

Since the objective of this research was to develop a classification system capable of distinguishing between "Parkinson's" and "Non-Parkinson's" cases, patients in the Healthy Control and Prodromal groups were categorized as "Non-Parkinson's," whereas individuals in the Parkinson's Disease group were classified as "Parkinson's."

B. Biospecimens

The biospecimen data used in this study were obtained through laboratory tests, for the collection of samples of whole

blood, plasma, serum, cerebrospinal fluid (CSF) and urine. In addition, some information is also provided, such as the last date of sample collection and identification of the collection phase. This information is extremely important, because by analyzing it, it is possible to identify biomarkers, among others. Finally, it is worth highlighting that this is one of the main tests for the early detection of Parkinson's, since plasma and serum can contain proteins and metabolites that indicate neurodegeneration, and cerebrospinal fluid (CSF) can be analyzed for biomarkers, which are directly associated with the progression of the disease.

C. Family History

Family history data included information on whether any family members of the participants had neurological conditions of interest, particularly Parkinson's disease and other brain-related disorders. This dataset aimed to assess the hereditary and genetic predisposition associated with PD.

D. Polygenic Risk Score (PRS)

The polygenic risk score (PRS) is a statistical measure that evaluates the genetic predisposition of an individual to develop complex diseases, such as Parkinson's disease. PRS integrates the cumulative effects of multiple genetic variants to quantify an individual's genetic susceptibility to PD, facilitating a more comprehensive risk assessment.

This structured dataset forms the foundation for the predictive modeling framework developed in this study, enabling the integration of multimodal patient data for enhanced Parkinson's disease classification and progression prediction.

III. Methodology

The project was developed using Python 3. A multi-agent system was implemented, structured into four processing layers.

The first layer focuses on data preprocessing, primarily aiming to optimize the training of the Machine Learning models that constitute the second layer. The third and fourth layers consist of evaluation agents responsible for assessing the performance of the models and making a final classification decision—determining whether a patient has Parkinson's disease or is healthy. This workflow is illustrated in Figure 4.

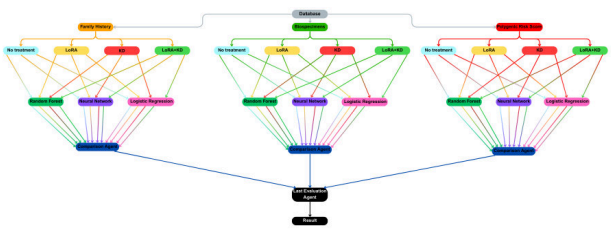


Fig. 4. Workflow

A. Machine Learning and Optimization Agents

As the primary objective of this project is to employ multi-agent systems for the prediction of Parkinson's disease, Machine Learning (ML) plays a fundamental role in achieving accurate and reliable results. To comprehensively address this challenge, four distinct ML models were implemented.

The first selected model was Random Forest [2], a machine learning algorithm that enhances predictive accuracy by aggregating multiple decision trees. It utilizes a technique known as bootstrap aggregation (bagging), in which each decision tree is trained on a randomly selected subset of the dataset. This approach generates a diverse ensemble of decision trees, increasing model robustness and reducing overfitting to the training data.

The second selected model was the MLPClassifier [3], based on a Multi-Layer Perceptron (MLP) Neural Network, a widely used artificial neural network architecture for supervised classification tasks. The implemented network consisted of three hidden layers with 200, 150, and 100 neurons, respectively, allowing the model to capture complex patterns within the dataset.

The activation function selected for this model was the Rectified Linear Unit (ReLU), which introduces non-linearity into the network, enabling it to learn and model complex relationships within the data. Mathematically, the ReLU function is defined as:

$$f(x) = \max(0, x)$$

This means that if the input x is positive, the function returns x, while if x is negative, it outputs zero. The ReLU function was chosen over alternatives like the sigmoid or hyperbolic tangent (tanh) functions due to several advantages. Unlike the sigmoid, which suffers from the vanishing gradient problem—where gradients become extremely small for large or small input values, slowing down training, ReLU maintains a constant gradient for positive inputs, allowing faster learning. Also while the tanh function centers its outputs around zero, it still exhibits saturation effects for extreme input values, leading to slow convergence.

In summary, ReLU was chosen for this model because it helps mitigate the problem of exploding gradients, which can occur in deep networks trained with backpropagation, avoids the vanishing gradient issue, and allows the network to learn more complex patterns effectively. In the context of Parkinson's disease diagnosis, these advantages are particularly relevant, as the model needs to capture subtle patterns across multimodal data, such as genetic markers, clinical records, and neuroimaging features.

The third approach employed a combination of Principal Component Analysis (PCA) [5] and Logistic Regression [4]. PCA is a dimensionality reduction technique that transforms a set of potentially correlated variables into a smaller set of uncorrelated ones, known as principal components, while preserving most of the original information.

After applying PCA, Logistic Regression is used to establish the relationship between the extracted principal components

and the target classes. As a linear classifier, Logistic Regression estimates the probability of a sample belonging to a particular class using the sigmoid function, which plays a fundamental role in converting linear outputs into probability scores. The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

This function maps any real-valued number to a range between 0 and 1, making it particularly effective for probability estimation. By compressing input values into a bounded interval, it facilitates clear differentiation between distinct classes in a classification task. In the context of Parkinson's disease diagnosis, this property is crucial, as the model needs to distinguish between patients with Parkinson's and those without the disease based on extracted features.

The use of Logistic Regression with the sigmoid function allows the model to output probability scores, that is interpreted as the likelihood of an individual having Parkinson's. This probabilistic interpretation is particularly valuable in medical applications, as it enables clinicians to assess the confidence of predictions rather than relying on rigid binary classifications.

The fundamental distinction between these models lies in the optimization techniques applied. Each algorithm leverages a different computational strategy to extract patterns from the data, ensuring robust and complementary predictive capabilities. The following section further details the comparative evaluation of these methodologies and their respective performance metrics.

1) Base Model Agent: This agent executes the Machine Learning models without any optimization techniques. Its primary purpose is to serve as a baseline model, enabling direct comparison with the optimized agents to assess their effectiveness in improving predictive performance.

2) Knowledge Distillation Agent: Knowledge Distillation [11] is an optimization technique designed to capture complex patterns within the data, thereby enhancing predictive accuracy. This approach is particularly beneficial for neural networks, as they excel in extracting intricate data representations. Similarly, PCA + Logistic Regression benefits from this method by leveraging the distilled knowledge to improve classification performance.

A key advantage of using Knowledge Distillation in this project is its ability to handle imbalanced data, as it provides richer information about class distribution through soft targets. This helps mitigate issues related to biased model training when working with unbalanced datasets.

In the medical field, Knowledge Distillation has been successfully applied to improve disease classification models. Hoang et al. [12] proposed an efficient approach for tuberculosis diagnosis using Knowledge Distillation to transfer knowledge from a complex ensemble of models to a more lightweight and efficient neural network, achieving high classification accuracy while reducing computational cost.

3) LoRA Agent: LoRA (Low-Rank Adaptation of Large Language Models) [13] is a technic to adapt pre-trained models, the mainly idea is to adjust an existing model to a new task or dataset without modifying all of its parameters, but by adding a small number of new parameters that capture the necessary adaptations. Initially it was made by Larguege Language Models (LLMs), but it also can be useful on other types of Machine Learning models. On the project's code, LoRA is responsable for is responsible for making the code more robust and faster, as well as helping with data balancing.

In the healthcare field, LoRA has already shown promising results in improving model performance for real-world medical applications. For example, Gagnon et al. [14] used LoRA to predict vital signs in sepsis patients in intensive care units, helping time-series models adapt to different types of medical data. Their study showed that LoRA improves prediction accuracy without requiring full retraining, making it an efficient way to adapt models.

4) LoRA And Knowledge Distillation Agent: This agent combines both LoRA and Knowledge Distillation, leveraging their complementary properties to enhance model performance, improve data balancing, and accelerate training. While Knowledge Distillation transfers knowledge from a complex model to a simplified one, LoRA efficiently adapts the simplified model without significantly increasing computational costs. This synergy results in a more efficient and scalable approach for Parkinson's Disease prediction.

B. Model Evaluation Agent

The Model Evaluation Agent was designed to assess, compare, and validate the Machine Learning models implemented in the project. Its primary function is to calculate key performance metrics, including: accuracy, precision by class, recall, F1-score, ROC Curve, training time, memory usage.

Additionally, this agent stores results in a local CSV file and generates visual representations for better interpretability. It is applied to each Machine Learning agent, ensuring a direct and systematic comparison of their respective performances.

C. Comparison Agent

This agent plays a crucial role in evaluating the consistency and reliability of the Machine Learning models. It assigns numerical values to the predictions: 2 for "Parkinson's Disease" and 1 for "Not Parkinson"

Beyond simple classification, this agent assesses which model is the most reliable, analyzing precision and accuracy to determine a weighted confidence score. The weighted average of the models' predictions is then computed, assigning a weight of 2 to the most reliable model and 1 to the others. The final decision is made based on the following threshold: If the weighted average is greater than 1.5, so it determine "Parkinson's Disease", however If the weighted average is less than 1.5, so it determine "Not Parkinson".

Since this agent is applied across all data types, three independent predictions are generated by the end of this phase.

D. Last Evaluation Agent

This agent is responsible for delivering the final diagnostic outcome. It consolidates the three predictions obtained from

different data modalities and performs an integrated evaluation based on patient ID.

As in the Comparison Agent, numerical values are assigned to predictions (2 for "Parkinson's Disease" and 1 for "Not Parkinson"), followed by a weighted average computation. However, in this case, the weights are assigned based on the performance of each data type rather than direct model accuracy. Since Polygenic Risk Score (PRS) demonstrated the highest predictive performance, it is given a weight of 2, while the other modalities receive a weight of 1.

Additionally, this agent implements a simple majority voting mechanism, accounting for cases where patient IDs are present in only two of the three models. If data is available from only one model, the patient is classified directly based on that model's prediction without undergoing the voting process.

Finally, the system evaluates the accuracy and precision of the weighted averaging method, as illustrated in Figures 5 and 6, and compares it to the majority voting method, as well as the individual predictions of each agent. The approach that demonstrates superior performance is selected as the final decision-making strategy for Parkinson's disease prediction.

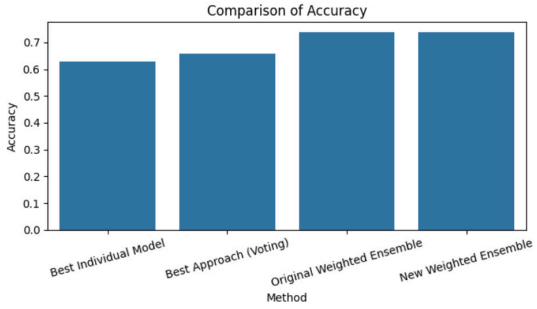


Fig. 5. Accuracy comparison

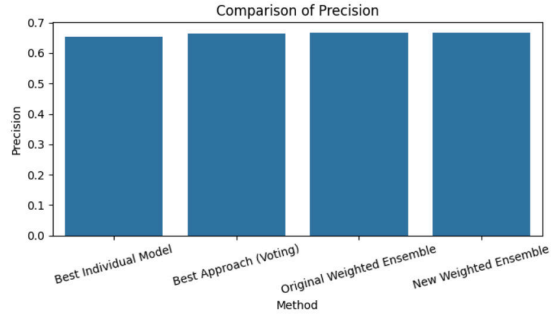


Fig. 6. Precision comparison

In conclusion, the multi-agent system created to predict Parkinson's disease progression includes several agents, each with its own job in the diagnostic process. The Machine Learning agents, like Random Forest, Multi-Layer Perceptron, and Logistic Regression, use different machine learning models on various data types to classify patients. To make the models more accurate, techniques like Knowledge Distillation and Low-Rank Adaptation (LoRA) were used. There's also the Model Evaluation agent, which checks how well the models are performing using metrics like accuracy and sensitivity, and the Comparison agent, which compares the model predictions and calculates a weighted average to decide if the patient has Parkinson's. Lastly, the Last Evaluation agent combines the predictions from all models and uses a majority voting system to make the final call. The way these agents work together to fine-tune predictions makes the system more reliable and accurate, ultimately leading to better early diagnosis.

IV. Results

To better illustrate the results, it is essential to present key numerical data obtained from the program.

Firstly, regarding the training time of the models, a comparison was made between the raw execution time and the time using LoRA and Knowledge Distillation (KD). In practical terms, only the Neural Network model exhibited significant improvements. For the other two models, the reduction in training time ranged from 0.5 to 1 second. However, for the Neural Network model, the training time was reduced by up to 50%, as shown in Table I.

Database	Raw Data	LoRA	KD	LoRA + KD
Family History	25.86 s	22.16 s	15.19 s	10.63
Biospecimens	21.73 s	13.96 s	7.85 s	7.70 s
Polygenic Risk Score	2.52 s	1.32 s	0.87 s	0.43 s

TABLE I
Comparison of Training Time in Neural Network

Furthermore, it is important to highlight that, despite the gain in processing speed, both precision and accuracy were preserved, as demonstrated in the tables II, III and IV.

These are important metrics for measuring the quality of LoRA and KD treatments, since they prove that, even when optimizing the process, they maintain the same level of quality of the raw data.

Database	Raw Data	LoRA	KD	LoRA + KD
Family History	0.9407	0.9390	0.9491	0.9610
Biospecimens	0.9256	0.9272	0.9173	0.9129
Polygenic Risk Score	0.9583	0.9439	0.9524	0.9262

TABLE II
Comparison of Precision in Neural Networks – Parkinson's Disease Classification

Database	Raw Data	LoRA	KD	LoRA + KD
Family History	0.3630	0.3672	0.3109	0.2331
Biospecimens	0.3423	0.3328	0.3472	0.3619
Polygenic Risk Score	0.3101	0.2746	0.2778	0.2835

TABLE III
Comparison of Precision in Neural Networks – Healthy Classification

Regarding the other Machine Learning techniques and their performance across different data types, all methods yielded satisfactory results. Even when applied to an initially highly unbalanced dataset, the models demonstrated efficiency in

Database	Raw Data	LoRA	KD	LoRA + KD
Family History	0.8695	0.8723	0.8367	0.7445
Biospecimens	0.7672	0.7572	0.7784	0.7925
Polygenic Risk Score	0.6225	0.5622	0.5622	0.5984

TABLE IV
COMPARISON OF ACCURACY IN NEURAL NETWORKS

identifying patients with Parkinson's disease. However, they occasionally exhibited limitations in accurately classifying healthy individuals.

Despite this issue, the implementation of a multi-agent system with a multi-stage evaluation process effectively mitigated the error rate for healthy individuals. As shown in the image 7, the percentage of correctly classified healthy individuals demonstrates significantly improved accuracy compared to the results obtained from isolated methods.

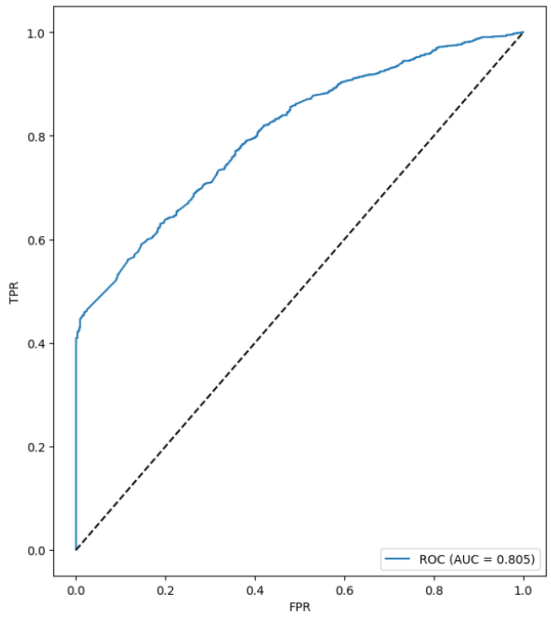


Fig. 7. ROC Curve

With the representation of the ROC curve in image 7, several critical parameters can be highlighted. Since this evaluation method is not influenced by class imbalance, the analyses derived from it are highly reliable. Consequently, it provides a comprehensive assessment of the model's ability to distinguish between classes and its performance across all sensitivity levels.

V. CONCLUSION

It is important to highlight that, despite utilizing a small and unbalanced dataset, the project achieves accuracy levels comparable to those reported in studies comparing Convolutional Neural Networks, Feed-forward Neural Networks, and a hybrid approach combining both [9]. However, the most critical aspect to analyze in this context is the system's ability to generate consistent results, even in the presence of ambiguous or incomplete data. This robustness is primarily attributed to the multi-agent framework and the multi-stage evaluation methodology employed.

Regarding the performance of the models in identifying individuals diagnosed with Parkinson's disease, all trained models achieved accuracy between 0.60 and 0.90. Notably, the precision for this class consistently ranged from 70% to 90%, in stark contrast to the performance observed in identifying healthy individuals. After processing by the Last Evaluation Agent, these performance metrics remained stable.

Therefore, it is important to highlight the significant potential of this project to support the diagnosis of Parkinson's disease, particularly in socially vulnerable contexts, where patients may have limited access to multiple and repeated examinations — especially imaging tests, which are generally costly. In this regard, it is entirely feasible to reach a conclusion that closely approximates clinical reality using only two laboratory tests and an analysis of family history. This feature is further enhanced by the fact that the system can be executed on machines with low computational power. Thanks to the application of optimization and model reduction techniques, such as LoRA (Low-Rank Adaptation) and Knowledge Distillation, the program demonstrates satisfactory performance even under constrained hardware conditions.

Nevertheless, it is essential to emphasize that this system is not intended to replace the assessment conducted by qualified healthcare professionals, but rather to act as a decision-support tool, contributing to the diagnostic process by optimizing both time and available resources.

REFERENCES

[1] World Health Organization, "Parkinson's disease" Available at: https://www.who.int/news-room/fact-sheets/detail/parkinson-disease, 2022. Accessed on March 12, 2025.

[2] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[3] A. C. Olivieri, "Non-linearity and artificial neural networks. Multi-layer perceptron," Introduction to Multivariate Calibration, 1st ed., pp. 271–288, 2024.

[4] T. Cleff, "Logistic regression," Applied Statistics and Multivariate Data Analysis for Business and Economics, 2nd ed., pp. 421–452, 2025.

[5] "Principal Component Analysis (PCA)," Geographic Data Analysis Using R, 1st ed., pp. 155–165, 2024.

[6] H. Lei, J. Zhang, Z. Yang, E. Tan, B. Lei, and Q. Luo, "Joint detection and clinical score prediction in Parkinson's disease via multi-modal sparse learning," Proceedings of IEEE International Conference, pp. 1231–1234, 2017.

[7] S. Sandhiya, V. Prabhu, S. Ashok, K. Mohanraj, G. Vishnu Vardhan Rao, and R. Azhagumurugan, "Parkinson's Disease Prediction Using Machine Learning Algorithm," Proceedings of the 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), pp. 1–6, 2022.

[8] Y. Jiang, H.-J. Yang, J. Kim, Z. Tang, and X. Ruan, "Power of Multi-Modality Variables in Predicting Parkinson's Disease Progression," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 2, pp. 1343–1355, Feb. 2025.

[9] B. Majhi, A. Kashyap, S. S. Mohanty, S. Dash, S. Mallik, A. Li, and Z. Zhao, "An improved method for diagnosis of Parkinson's disease using deep learning models enhanced with metaheuristic algorithm," BMC Medical Imaging, vol. 24, no. 1, pp. 1–14, 2024.

[10] R. Erro, S. Schneider, M. Stamelou, N. Quinn and K. Bhatia, "What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? New evidence and continuing controversies," Journal of Neurology, Neurosurgery & Psychiatry, vol. 87, pp. 319—323, 2015.

[11] Y. Khaider, D. Rahhali, H. Alami, and N. En Nahnah, “Exploring the knowledge distillation,” New Technologies, Artificial Intelligence and Smart Data (INTIS 2022, INTIS 2023), pp. 32–44, 2023.
[12] D. Ngoc-Hoang, T. Tran, T. Dinh, and M. Nguyen, “An Efficient Approach for Tuberculosis Diagnosis on Chest X-Ray Images Using Knowledge Distillation,” International Conference on Artificial Intelligence and Machine Learning, pp. 1–7, 2021.
[13] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv preprint arXiv:2106.09685, 2021.
[14] P. Gagnon, A. R. Smith, and M. J. Lee, “Low-Rank Adaptation of Time Series Foundational Models for Out-of-Domain Modality Forecasting,” arXiv preprint arXiv:2405.10216, 2024.
[15] M.Harry, V. Vijula, and J. Pearly Princess, ‘Parkinson’s Disease Prediction Using Machine Learning Techniques,” 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS).

Color Deconvolution and Marker-Controlled Watershed Framework for Intracellular Parasite Quantification in HTR8/SVneo Cells

Luís Octávio Custódio Trovo
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0009-0006-6388

Guilherme Vieira de Faria
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0004-0933-5189

Bellisa Freitas Barbosa
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-4611-2844

Eloisa Amalia Vieira Ferro
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-7447-2583

Ana Claudia Patrocinio
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Pedro Cunha Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

***Abstract*—Chagas disease, caused by *Trypanosoma cruzi*, remains a major neglected tropical disease for which reliable quantification of intracellular parasite burden is important for investigating infection dynamics and supporting pharmacological screening. Conventional manual inspection of Giemsa-stained cultures is labor-intensive, operator-dependent, and poorly scalable. This work proposes a computer-aided image-processing framework for detecting host-cell nuclei and generating an initial estimate of intracellular amastigote burden in Giemsa-stained HTR8/SVneo cultures. The method combines color deconvolution, Otsu thresholding, marker-controlled watershed segmentation, morphological and intensity feature extraction, unsupervised *k*-means clustering, and centroid-based spatial association. Because automatic parasite detection is sensitive to staining variability, a PySide6 graphical interface is incorporated as a human-in-the-loop validation stage, allowing expert review and correction before final reporting. The framework was evaluated on 22 microscopy images using an independently and blindly obtained manual ground truth. Host-cell detection achieved a Precision of 95.49%, Recall of 96.49%, and F1-score of 95.99%. In contrast, amastigote counting showed a mean absolute percentage error of 29.47% and an aggregated overestimation of 19.77% (2345 automatic versus 1958 manual counts). The observed errors were mainly associated with local Giemsa staining variability and the use of a fixed color-deconvolution matrix, which can cause small basophilic structures or weakly stained nuclear fragments to be classified as parasite candidates. The proposed framework is therefore positioned as a reproducible human-in-the-loop computer-aided quantification approach rather than a fully autonomous parasite-counting system.**

***Index Terms*—Automated detection, Color deconvolution, Digital image processing, Image segmentation, *Trypanosoma cruzi*.**

I. Introduction

Chagas disease is a parasitic disease caused by the protozoan *Trypanosoma cruzi* and remains a major neglected tropical disease, especially in Latin American countries [1]. It is estimated that millions of people are infected worldwide while tens of millions remain at risk of exposure to the parasite, making it a significant public health problem [2]. Although historically associated with endemic regions in Latin America, the disease has come to have a global impact due to migratory flows and the emergence of alternative routes of transmission, including blood transfusion, organ transplantation, and congenital transmission [3], [4].

Clinically, the infection can progress from an acute phase which is often asymptomatic to chronic forms primarily associated with cardiomyopathies and severe gastrointestinal disorders [5]. Furthermore, currently available treatments, based mainly on benznidazole and nifurtimox, still have limitations related to toxicity, adverse effects, and reduced efficacy at certain stages of the disease [6]. This scenario underscores the need to develop new experimental and therapeutic approaches focused on studying parasite-host interactions and screening for antiparasitic compounds.

Investigating these cellular mechanisms requires reliable *in vitro* models. In this context, the HTR8/SVneo cell line, derived from human extravillous trophoblasts, serves as a vital tool for studying the placental microenvironment and host susceptibility to congenital infections [7]. In these experimental assays, precisely quantifying the intracellular parasite burden within cell cultures is a fundamental step to evaluate infection dynamics and conduct pharmacological screening [8], [9].

However, many protocols still rely on manual microscopic inspection, which is considered a labor-intensive, time-consuming procedure that is highly dependent on the operator's experience [10]. Previous studies highlight that this type of analysis has low scalability and high susceptibility to intra- and inter-observer variability, hindering standardization and experimental reproducibility [10]–[12].

With advances in digital image processing techniques, various automated strategies have been explored for the analysis

of intracellular infections [13]–[16]. Methods based on fluorescence and computational segmentation have demonstrated the ability to extract relevant quantitative parameters, such as the number of infected cells, average parasite load, and parasite-to-host-cell ratio [13]. In addition, automated phenotypic approaches have gained prominence in drug discovery for Chagas disease due to the possibility of large-scale analysis and reduced experimental subjectivity [17].

While Giemsa staining remains a cost-effective standard for visualizing intracellular parasites [18], [19], its inherent heterogeneity, nonuniform illumination, and the frequent spatial overlapping of amastigotes with host nuclei severely limit simple thresholding techniques [20], [21]. To address these computational challenges, this study proposes an interpretable image-processing framework, combining color deconvolution, marker-controlled watershed segmentation, and unsupervised *k*-means clustering, to detect host-cell nuclei and associate candidate parasite structures in Giemsa-stained HTR8/SVneo cultures. Every stage of the pipeline relies on explicit, inspectable operations rather than learned parameters, and its output remains editable: the system incorporates a human-in-the-loop PySide6 graphical interface that allows specialists to review the automated detection map and correct false-positive associations before final reporting. The framework is therefore positioned as a reproducible, computer-aided decision tool for pharmacological screening and infection dynamic studies rather than a fully autonomous counting method.

II. Materials and Methods

A. Image Dataset and Ground Truth

For the development and evaluation of the proposed image processing pipeline, 22 images were obtained by optical microscopy (Opton) at $40\times$ magnification from randomly selected fields on the coverslips, with an approximate resolution of 0.16 μm/pixel. The images show Giemsa-stained HTR8/SVneo cells infected with tissue culture-derived trypomastigotes of the *Trypanosoma cruzi* Y strain, in an assay consisting of two independent experiments with three replicates each. The digital photomicrographs were stored in uncompressed TIFF (Tagged Image File Format) with spatial dimensions of 2048×1536 pixels and a 24-bit color depth (RGB standard).

For the computational analysis, host cells were represented by their larger nuclei, whereas small basophilic structures spatially associated with host nuclei were treated as candidate amastigotes. A host cell was considered infected when two or more candidate amastigotes were assigned to its spatial domain.

The reference counts were obtained independently and blindly by a single experienced researcher, who manually evaluated each microscopic field without access to the automated pipeline results. The ground truth contained the absolute number of host-cell nuclei and amastigotes per image. Instance-level masks or centroid coordinates for individual amastigotes were not available; therefore, the parasite ground truth supported per-image counting comparisons but not formal object-to-object validation of the host–parasite spatial association. The complete computational workflow is summarized in Fig. 1.

B. Chromatic Separation

Chromatic separation was performed using the color deconvolution technique proposed by Ruifrok and Johnston [22], which is based on the optical density (OD) model. In this model, the intensity I recorded at each pixel is converted to the optical density space using the logarithmic transformation in (1):

$$OD = -\log_{10}\left(\frac{I}{I_0}\right), \tag{1}$$

where I_0 represents the background intensity (transmitted light without absorption). Each dye present in the sample can be characterized by a three-component vector corresponding to its optical densities in the RGB channels. For a sample stained with two or three dyes, these vectors form a 3×3 optical density matrix M, in which each row represents a dye and each column represents the optical density in the respective channel [22].

The separation of the individual contributions from each dye is then obtained using the deconvolution matrix $D = M^{-1}$, such that the amount of each dye in a given pixel is calculated as in (2):

$$C = D \cdot OD, \tag{2}$$

where C is the vector containing the contributions of each dye and OD is the optical density vector observed at that pixel [22].

The normalized spectral vectors used in this study correspond to the canonical values defined by Ruifrok and Johnston [22] for the combination of hematoxylin, eosin, and a residual channel, forming the matrix M in (3):

$$M = \begin{pmatrix} 0.65 & 0.70 & 0.29 \\ 0.07 & 0.99 & 0.11 \\ 0.27 & 0.57 & 0.78 \end{pmatrix}. \tag{3}$$

The use of this fixed canonical matrix, rather than individual adjustment per image, was motivated by the high variability observed in the manual selection of reference pixels, which led to inconsistencies in the chromatic separation between images of the same cell line. Although originally defined for H&E (Hematoxylin and Eosin), this matrix effectively separates the blue/purple (azure/methylene blue) and pink (eosin) components of the Giemsa stain due to the similar spectral absorption profiles of these dyes [22], [23], as illustrated in Fig. 2.

After deconvolution, the channel corresponding to the blue/purple component (mathematically analogous to the hematoxylin channel in the canonical matrix) was selected for the subsequent steps, as the high affinity of the azure and methylene blue components for nucleic acids provided a clearer view of the cell nuclei and of the basophilic regions relevant for cell identification and for the detection of possible intracellular parasitic structures.

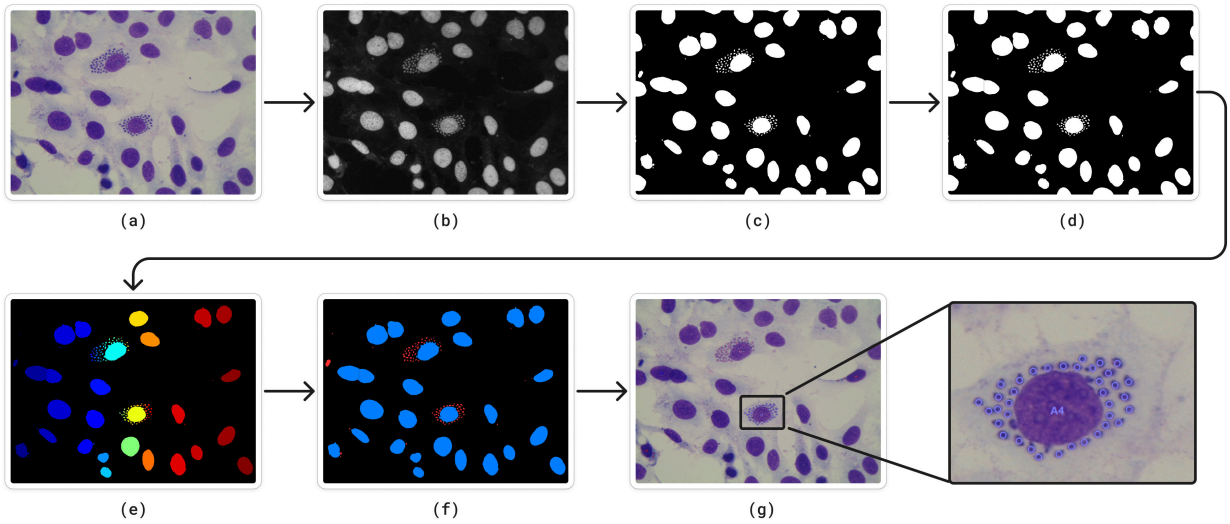


Fig. 1. General flowchart of the proposed methodology. (a) Original Giemsa-stained image; (b) blue/purple channel after color deconvolution; (c) Otsu binarization; (d) watershed segmentation; (e) labeling of cellular structures; (f) *k*-means clustering; (g) automated spatial association of candidate parasite structures. The resulting map is subsequently reviewed in the human-in-the-loop graphical interface before final reporting.

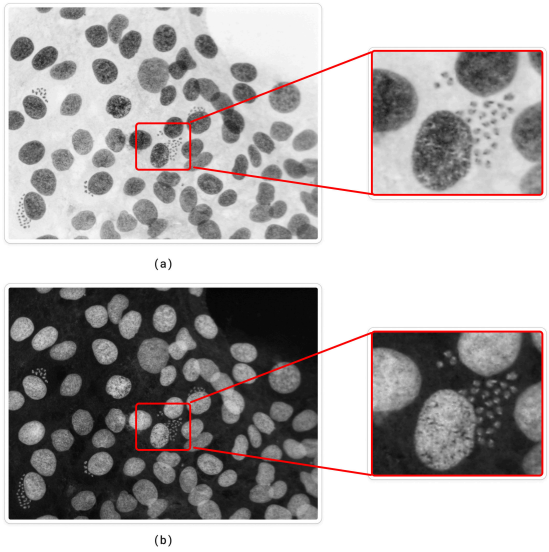


Fig. 2. Chromatic separation achieved by optical deconvolution (OD). (a) Pink channel (eosin component) after deconvolution, predominantly highlighting cytoplasmic structures and extracellular components. (b) Blue/purple channel (azure/methylene blue component) after deconvolution, highlighting cell nuclei and intracellular structures.

C. Cell Segmentation

The first step in segmentation was binarization of the blue/purple channel using the Otsu method [24], since nuclei generally stand out clearly from the background in histopathological images and global histogram-based thresholds yield good results in this setting [25]–[27]. The method divides the image into background (C_1) and nuclei (C_2) and automatically finds the optimal threshold t^* by maximizing the interclass variance σ_B^2, without the need to manually calibrate values for each analyzed image [24], [26].

Although the Otsu method works well, the internal staining of the cell is not always homogeneous, which often results in internal gaps (holes) in the resulting binary mask [27]. To correct these gaps without altering the actual size of the nucleus, the morphological operation of region filling (*imfill*) was applied [28]. The result of this morphological step is the primary binary mask of the cellular structures, shown in Fig. 3(b).

To resolve nuclear clustering, the watershed algorithm [30] was applied by treating the binary mask as a topographic relief. However, direct application of classical watershed leads to severe over-segmentation due to staining noise and minor edge irregularities [28], [31]. To suppress these spurious local minima and preserve the actual morphology of adjacent host cell nuclei, a marker-controlled watershed strategy was implemented, integrated with the Euclidean Distance Transform (EDT) [28], [29]. The EDT calculates the shortest spatial distance from each pixel in the core p to the nearest background pixel q, generating a distance map $D(p)$, as shown in (4):

$$D(p) = \min_{q \in R^c} d(p, q). \tag{4}$$

In this new inverted topographic relief, the highest values (distance peaks) represent the individual centers of mass of each cluster. The algorithm locates these regional peaks and uses them as strict immersion markers (seeds). By restricting the flooding process so that it starts exclusively from these validated seeds, the technique suppresses spurious minima [29]. Thus, the watershed creates dividing lines only at the actual boundary between overlapping nuclei, disaggregating the cellular clusters as illustrated in Fig. 3(c). Next, connected-component labeling is applied, where each nucleus receives a

unique numerical identifier, allowing for individual visualization and analysis, as shown in Fig. 3(d).

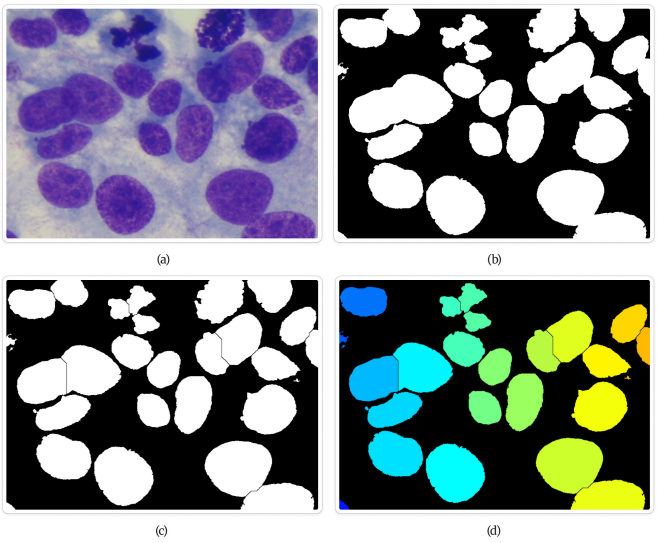


Fig. 3. Stages of nuclear segmentation in the cropped image (777x573 pixels): (a) Original Giemsa-stained image; (b) Consolidated binary mask after global thresholding using the Otsu method and morphological filling; (c) Geometric separation of overlapping nuclei using a marker-controlled watershed transform; (d) Labeling of connected components, assigning unique identifiers to each isolated nucleus for attribute extraction.

D. Detection of Cell and Parasitic Structures

After separating and geometrically labeling the objects in the image, morphological and intensity features were extracted for each one. The main objective of this step is to mathematically differentiate host cells (larger structures) from parasites (smaller structures). For each identified nucleus, we calculated the area, the equivalent diameter (D_{eq}), the solidity, the eccentricity, the coordinates of the geometric centroid, and the average intensity in the blue/purple channel ($\overline{I}_B$). These attributes are widely used in the literature to analyze nuclei in histopathological images [32], [33], as they summarize information regarding cell size, shape, and color.

These two cell populations were classified using the unsupervised *k*-means algorithm [34]. To optimize computational processing and prevent vastly different scales from causing errors in the separation, the clustering used only two variables, forming the two-dimensional space $X = [D_{eq}, \overline{I}_B]$, whose values were previously normalized by the Z-score [33]. The algorithm divided the data into $k = 2$ groups (clusters), iteratively minimizing the sum of the intra-cluster quadratic errors J [34].

Finally, the group with the largest average diameter was classified as the host cell population. The remaining group was classified as the candidate amastigote population. To refine the results and remove small spots or staining artifacts, a morphological filter was applied to remove host cells that were cut off by the edge of the image, since it would not be possible to verify the presence of amastigotes. The visual result of this mathematical separation can be seen in Fig. 4(a).

With the two classes separated, the final step was to map which candidate amastigotes belonged to which host cells. To do this, the code extracted the spatial coordinates (x, y) of all centroids and calculated the matrix of Euclidean distances between them. The distance d between the centroid of a parasite c_p and any other centroid c_t (whether it is another parasite or a cell) is given by (5):

$$d(c_p, c_t) = \sqrt{(x_p - x_t)^2 + (y_p - y_t)^2}. \quad (5)$$

Due to the biology of the infection, amastigotes typically form clustered cellular nests. As a result, it is very common for a parasite to be spatially closer to another parasite than to the center of its own host nucleus [8]. To address this issue, an iterative neighborhood search algorithm (greedy boundary propagation) was implemented. The algorithm tracks the nearest neighbor of each parasite; if it is another amastigote, the two structures are grouped together. This continuous expansion continues until the nearest structure is, in fact, the nucleus of a host cell. When this occurs, the entire cluster of amastigotes is attached to that host, making the cell infected. The use of Euclidean distance combined with neighborhood logic is one of the most robust criteria for analyzing cellular position in dense tissues [32], [35].

This iterative process constructs a final infection map, shown in Fig. 4(b), allowing the system to count, in an automated manner, the number of candidate amastigotes within each identified cell classified as infected.

E. Creation of a Graphical Interface

To support result validation and targeted manual correction, a graphical user interface was developed in PySide6. Considering the overestimation observed in automated amastigote counting, as discussed in Section III, the interface was implemented as a human-in-the-loop curation layer. It enables simultaneous visualization of the original image and the pipeline outputs, allowing the user to inspect, validate, and, when necessary, manually correct the automated detections.

Furthermore, the system provides editing tools that allow users to remove false positives and correct incorrect associations between parasites and infected cells. Undo and zoom features for local inspection were incorporated to assist the specialist during the validation stage [36].

F. Quantitative Analysis

To evaluate the performance of the proposed method, the automatically segmented structures were validated against the manual count performed by an experienced researcher, which served as the experimental ground truth. For each microscopic field, the manual analysis established the absolute number of existing host cell nuclei (true presence).

The classification performance for this specific class was quantified using True Positive (TP), False Positive (FP), and False Negative (FN) metrics. Object matching was performed by side-by-side visual audit of the segmentation map and the corresponding original field: a detected object was accepted as

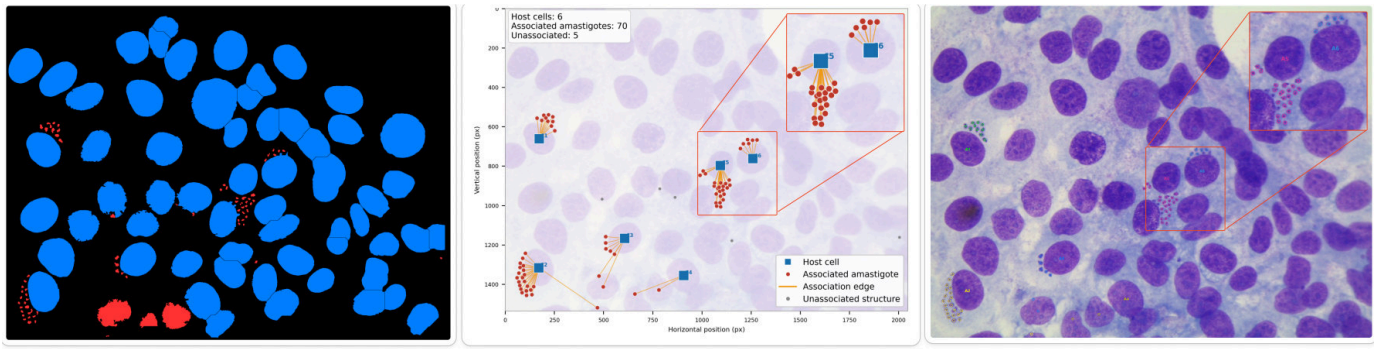


Fig. 4. Final steps of the parasite detection pipeline. (a) Unsupervised classification using k-means ($k = 2$) in the space $[D_{eq}, \bar{I}_B]$: blue cells represent host cell nuclei, and red circles represent parasitic structures (amastigotes). (b) Spatial association map based on Euclidean distance between centroids: gold edges connect each amastigote to its nearest host cell, forming infection clusters. (c) Final result displayed in the graphical interface, where each color identifies a distinct host-parasite cluster.

TP when its segmented region overlapped, partially or completely, a nucleus of the manual reference, under a one-to-one correspondence between detection and reference nucleus. No automatic overlap or centroid-distance criterion was applied. Structures quantified by the automated pipeline that did not overlap any reference nucleus were labeled as FP (artifacts or noise).

Conversely, existing cell nuclei that the system failed to identify were recorded as FN. Based on these audited counts across all 22 images that were visually assessed by an experienced researcher, the system's reliability and sensitivity for nuclear detection were evaluated through cumulative metrics. For amastigotes, only per-image absolute counts were compared, as the ground truth provided total counts only, without spatial location.

True Negatives (TN) were not included in this evaluation: in a detection and counting task they would correspond to every empty region of the image, so their value would be arbitrary. Metrics that depend on TN, such as Specificity, are therefore not meaningful here, and Precision and Recall are the appropriate choices [37].

Using the TP, FP, and FN values accumulated across all 22 images, we calculated three metrics widely used in the literature to evaluate classification systems [38]: *Precision* (P), *Recall* (R), and *F1-score* (F_1), defined in (6)–(8):

$$P = \frac{TP}{TP + FP}, \tag{6}$$

$$R = \frac{TP}{TP + FN}, \tag{7}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{8}$$

For the amastigote class, the agreement between automated and manual counts was quantified per image through the signed percentage error, and summarized over the dataset by the mean absolute percentage error (MAPE) and by the aggregated error computed from the total counts.

In addition, in order to investigate possible factors associated with the discrepancies observed between manual and automatic amastigote counts, an analysis was conducted of the chromatic properties of the 22 images evaluated. For each image, the range with the highest concentration of intensities in each RGB channel was identified, and its amplitude (the difference between the maximum and minimum values of that range) was calculated. This metric reflects the heterogeneity of intensities within the channel, with larger amplitudes indicating more scattered distributions and potentially greater variability in the absorption of the corresponding dye.

In addition, the RGB histograms of images IM9 and IM8, representing, respectively, the largest and one of the smallest percentage errors in the dataset, were analyzed comparatively, with the aim of identifying visual differences in the intensity distributions that might be associated with the pipeline's performance in each microscopic field. This analysis was exploratory in nature and sought to relate the chromatic characteristics of the images to the behavior of the pipeline during the binarization and classification of parasitic structures.

III. Results

Table I presents the absolute counts of host cells and amastigotes obtained by manual annotation (gold standard) and by the proposed method for each of the 22 images evaluated, together with the signed percentage error of the automated amastigote count, which ranged from -11.3% (IM3) to $+76.1\%$ (IM9).

For the host cell class, the evaluation was complemented by the analysis of True Positives (TP), False Positives (FP), and False Negatives (FN), whose accumulated values across the 22 images are presented in the confusion matrix in Fig. 5, where the True Negative quadrant is marked as not applicable, since the task consists of object detection and counting.

Based on the TP, FP, and FN values presented in Fig. 5, the Precision, Recall, and F1-score metrics for host cell detection were calculated, as shown in Table II.

To investigate the discrepancies observed in the amastigote count, the chromatic properties of the images were analyzed.

TABLE I
COMPARISON OF MANUAL AND AUTOMATED COUNTING OF HOST CELLS AND AMASTIGOTES. THE LAST COLUMN REPORTS THE SIGNED PERCENTAGE ERROR OF THE AUTOMATED AMASTIGOTE COUNT RELATIVE TO THE MANUAL REFERENCE; THE TOTAL ROW REPORTS THE AGGREGATED ERROR OVER ALL IMAGES (MAPE: 29.47%)

Image	Host Cells		Amastigotes		
	Manual	Algorithm	Manual	Algorithm	Error (%)
IM1	56	56	48	64	+33.3
IM2	65	65	35	39	+11.4
IM3	69	73	53	47	−11.3
IM4	67	71	21	30	+42.9
IM5	59	59	72	75	+4.2
IM6	26	26	97	90	−7.2
IM7	62	64	108	121	+12.0
IM8	87	84	23	22	−4.3
IM9	95	97	67	118	+76.1
IM10	79	78	16	25	+56.3
IM11	45	44	7	11	+57.1
IM12	80	77	16	21	+31.3
IM13	43	42	10	14	+40.0
IM14	57	60	30	51	+70.0
IM15	70	70	124	137	+10.5
IM16	68	70	309	348	+12.6
IM17	61	63	80	128	+60.0
IM18	98	90	268	326	+21.6
IM19	48	47	161	168	+4.3
IM20	70	74	105	150	+42.9
IM21	63	65	121	159	+31.4
IM22	58	66	187	201	+7.5
Total	**1426**	**1441**	**1958**	**2345**	**+19.8**

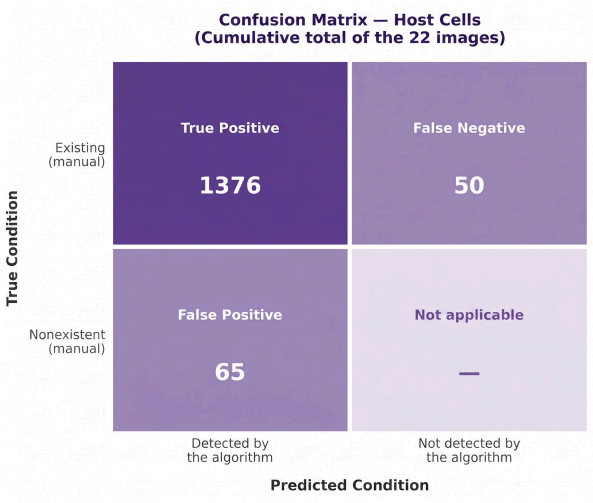


Fig. 5. Confusion matrix showing the cumulative values of TP, FP, and FN for host cell detection across the 22 images evaluated. The True Negative quadrant is marked as not applicable.

TABLE II
PERFORMANCE METRICS FOR HOST CELL DETECTION

TP Total	Precision	Recall	F1-score
1376	0.9549	0.9649	0.9599

As shown in Fig. 6, the blue channel divided the dataset into two distinct patterns of behavior: images IM1 to IM13 were concentrated in a low-amplitude range (mean of 37.2 units), while images IM14 to IM22 showed a high-amplitude range (mean of 63.2 units). In contrast, the red and green channels remained relatively stable throughout the entire image set.

Additionally, the comparison between the RGB histograms of images IM8 and IM9 showed that, although both belong to the low-amplitude group in the blue channel, they represent opposite extremes in terms of algorithm performance. IM9 presented the highest percentage error in amastigote counting (+76.1%), whereas IM8 presented one of the smallest absolute deviations in the dataset (−4.3%) and the smallest among the images in which the automated count fell below the manual reference. The visualization shows that IM9 has a slightly more dispersed intensity distribution in the blue channel compared to IM8, suggesting that variations in the chromatic distribution, and not just in the amplitude of the highest concentration range, can influence the performance of automatic detection.

IV. DISCUSSION

The results obtained for host cell detection, presented in Section III, demonstrated high and consistent performance, with a precision of 95.49%, a recall of 96.49%, and an F1-score of 95.99%. As shown in the confusion matrix in Fig. 5, of the 1426 host cell nuclei in the manual ground truth, the algorithm detected 1441 structures, comprising 1376 true positives, 65 false positives, and 50 false negatives, indicating that the color deconvolution, binarization, and watershed steps are sufficiently robust for segmenting larger-diameter nuclei despite the variability in contrast and cellular density across the dataset.

In contrast, amastigote counting showed a mean absolute percentage error (MAPE) of 29.47% and an aggregated overestimation of 19.77% (2345 automated versus 1958 manual counts, Table I), indicating that, unlike host cell detection, the amastigote counts should not yet be interpreted as a fully reliable stand-alone measurement of parasite burden. To investigate this behavior, the amplitude of the blue channel was analyzed, Fig. 6, revealing that images IM14–IM22 have a significantly higher amplitude (mean of 63.2) compared to images IM1–IM13 (37.2). However, this discrepancy did not affect the MAPE, which remained close to 29% in both groups. The significant difference occurred only in the mean absolute error (31.4 structures per image in IM14–IM22 versus 10.2 in IM1–IM13), which can be explained by the higher parasite load in the second group (mean of 153.9 amastigotes versus 44.1 in the first), while the number of host cells remained constant. Thus, the greater amplitude in the blue channel reflects the high volume of azure/methylene blue-stained structures (nucleic acids), rather than an isolated metric of pipeline degradation.

Although the overall amplitude is not a direct predictor of error, localized variations in staining affect segmentation. Weakly stained nuclei or fragments with residual

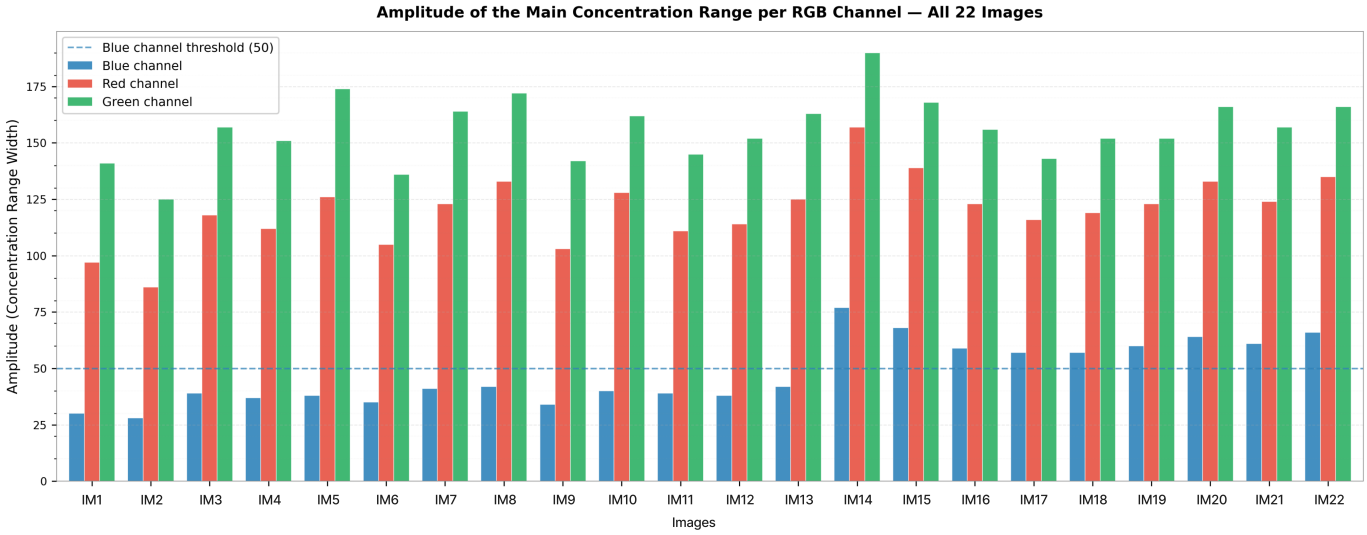


Fig. 6. Range of the band with the highest concentration of intensities in the red, green, and blue channels for the 22 images evaluated, in sequential order. The dashed line indicates the threshold in the blue channel that visually separates two groups with distinct behaviors (IM1–IM13 and IM14–IM22).

azure/methylene blue absorption tend to be segmented independently when they reach the binarization threshold. Due to their small diameter, the *k*-means algorithm incorrectly classifies them as amastigotes. This behavior highlights a limitation of the fixed nature of the color deconvolution matrix defined in (3), which assumes a constant reference intensity and does not adapt to local fluctuations in chromatic density; the overestimation is a direct result of the combination of high amastigote density and these variations in staining.

It is worth noting that the step involving spatial association between amastigotes and host cells was not subjected to formal quantitative validation in this study, since a viable protocol for individualized spatial correspondence between manual annotations and automatically detected structures was not established. The accuracy of this step was verified qualitatively using the graphical interface developed, which allows the expert to visually inspect and manually correct incorrect associations before the final count. The development of a protocol for individualized spatial annotation, which would enable the calculation of association metrics, constitutes a recognized limitation of the present study.

Three further limitations should be acknowledged. First, the reference counts were produced by a single annotator, so inter-annotator agreement could not be estimated and the residual subjectivity of the ground truth remains unquantified. Second, validation was restricted to a single laboratory, staining protocol, microscope, and *T. cruzi* strain; the fixed deconvolution matrix and the derived thresholds were therefore not verified on an independent staining batch, which limits conclusions regarding robustness and generalizability. Third, no comparison against supervised or deep learning segmentation baselines was performed, since training and evaluating such models would require the instance-level annotations that the present ground truth does not provide.

V. Conclusion

This study presented an automated image-processing pipeline for the detection of host cell nuclei and intracellular parasite structures in Giemsa-stained microscopy images of the HTR8/SVneo cell line. The proposed methodology combined color deconvolution, Otsu thresholding, marker-controlled watershed segmentation, morphological feature extraction, *k*-means clustering, and spatial association to estimate the intracellular amastigotes burden.

The proposed approach achieved high performance for host cell detection (Precision of 95.49%, Recall of 96.49%, and F1-score of 95.99%), but amastigote counting showed a mean absolute percentage error of 29.47% and an aggregated overestimation of 19.77%. Automated amastigote counts should therefore be interpreted as an initial estimate, to be confirmed through expert review, rather than a definitive measurement of parasite burden.

The main contribution of the framework is therefore an interpretable and editable pipeline: all decision stages rely on explicit operations that can be inspected and adjusted, and the developed graphical interface provides an efficient tool for visual inspection and manual correction of the automatically generated results, allowing specialists to validate amastigotes-host associations before the final quantification.

Future work includes evaluating the pipeline on larger, multicenter datasets acquired under different staining protocols and microscopes, implementing adaptive stain normalization techniques, establishing a spatially annotated *ground truth* with multiple annotators to enable a quantitative evaluation of amastigotes-host association accuracy, and benchmarking the proposed framework against state-of-the-art supervised and deep learning segmentation methods.

ACKNOWLEDGMENT

The authors are thankful for the received financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Brazil), CNPq—Brazil, FAPEMIG—Brazil, and Universidade Federal de Uberlândia (UFU), as well as the scientific and technical support provided by the Laboratório de Imunofisiologia da Reprodução (LABIFIR).

REFERENCES

[1] E. Chatelain, "Chagas disease drug discovery: Toward a new era," *Journal of Biomolecular Screening*, vol. 20, no. 1, pp. 22–35, 2015, doi: 10.1177/1087057114550585.

[2] World Health Organization, "WHO fact sheet N° 340," http://www.who.int/mediacentre/factsheets/fs340/en/ (accessed May 29, 2014).

[3] L. V. Sánchez and J. D. Ramírez, "Congenital and oral transmission of American trypanosomiasis: An overview of physiopathogenic aspects," *Parasitology*, vol. 140, no. 2, pp. 147–159, 2013.

[4] J. C. Dias and V. Amato Neto, "Prevenção referente às modalidades alternativas de transmissão do Trypanosoma cruzi no Brasil," *Rev. Soc. Bras. Med. Trop.*, vol. 44, suppl. 2, pp. 68–72, 2011.

[5] A. Rassi, J. M. de Rezende, A. O. Luquetti, and A. Rassi Jr., "Clinical phases and forms of Chagas disease," in *American Trypanosomiasis: Chagas Disease One Hundred Years of Research*, J. Telleria and M. Tibayrenc, Eds., Elsevier, 2010, pp. 709–741.

[6] M. Keenan and J. H. Chaplin, "A new era for Chagas disease drug discovery?" *Progress in Medicinal Chemistry*, vol. 54, pp. 185–230, 2015.

[7] P. A. Marín et al., "Differential susceptibility of human placental trophoblasts to Trypanosoma cruzi infection," *Parasitol. Res.*, vol. 124, no. 9, p. 104, Sep. 2025.

[8] J. C. Engel et al., "Image-based high-throughput drug screening targeting the intracellular stage of Trypanosoma cruzi, the agent of Chagas' disease," *Antimicrob. Agents Chemother.*, vol. 54, no. 8, pp. 3326–3334, Aug. 2010.

[9] K. P. Chang and D. M. Dwyer, "Leishmania donovani. Hamster macrophage interactions in vitro: Cell entry, intracellular survival, and multiplication of amastigotes," *J. Exp. Med.*, vol. 147, pp. 515–530, 1978.

[10] E. Yazdanparast et al., "INSPECT, an open-source and versatile software for automated quantification of (Leishmania) intracellular parasites," *PLoS Negl. Trop. Dis.*, vol. 8, no. 5, p. e2850, May 2014.

[11] J. M. Bustamante and R. L. Tarleton, "Methodological advances in drug discovery for Chagas disease," *Expert Opin. Drug Discov*, vol. 6, pp. 653–661, 2011.

[12] A. J. Romanha et al., "In vitro and in vivo experimental models for drug screening and development for Chagas disease," *Mem. Inst. Oswaldo Cruz*, vol. 105, pp. 233–238, 2010.

[13] J. Alonso-Padilla et al., "Automated high-content assay for compounds selectively toxic to Trypanosoma cruzi in a myoblastic cell line," *PLoS Negl. Trop. Dis.*, vol. 9, p. e0003493, 2015.

[14] S. Moon et al., "An Image-Based Algorithm for Precise and Accurate High Throughput Assessment of Drug Activity against the Human Parasite Trypanosoma cruzi," *PLoS One*, vol. 9, no. 2, p. e87188, Feb. 2014.

[15] L. L. Nohara, C. Lema, J. O. Bader, R. J. Aguilera, and I. C. Almeida, "High-content imaging for automated determination of host-cell infection rate by the intracellular parasite *Trypanosoma cruzi*," *Parasitology International*, vol. 59, no. 4, pp. 565–570, 2010, doi: 10.1016/j.parint.2010.07.007.

[16] S. Moon et al., "An Image Analysis Algorithm for Malaria Parasite Stage Classification and Viability Quantification," *PLoS One*, vol. 8, no. 4, p. e61812, Apr. 2013.

[17] E. Chatelain and J. R. Ioset, "Phenotypic screening approaches for Chagas disease drug discovery," *Expert Opin. Drug Discov*, vol. 13, no. 2, pp. 141–153, Feb. 2018.

[18] F. Derouin and C. Chastang, "In vitro effects of folate inhibitors on Toxoplasma gondii," *Antimicrob. Agents Chemother.*, vol. 33, no. 10, pp. 1753–1759, 1989.

[19] R. Oualha, N. Fekih-Romdhane, D. Driss, Y. Z. Abdelkrim, I. Guizani, and E. Harigua-Souiai, "AIR-LEISH: A Dataset of Giemsa-Stained Microscopy Images for AI-based Leishmania amastigotes Detection," *Sci. Data*, vol. 13, no. 1, p. 328, Feb. 2026.

[20] S. S. Abbas and T. M. H. Dijkstra, "Detection and stage classification of Plasmodium falciparum from images of Giemsa stained thin blood films using random forest classifiers," *Diagn. Pathol.*, vol. 15, no. 130, 2020.

[21] M. T. Bahreyni Toossi, H. Azimian, O. Sarrafzadeh, S. Mohebbi, and S. Soleymanifard, "Automatic detection of micronuclei by cell microscopic image processing," *Mutat. Res.*, 2017.

[22] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, Aug. 2001.

[23] G. Landini, G. Martinelli, and F. Piccinini, "Colour Deconvolution: stain unmixing in histological imaging," *Bioinformatics*, vol. 37, no. 10, pp. 1485–1487, 2021.

[24] D. Liu and J. Yu, "Otsu method and K-means," in *2009 Ninth International Conference on Hybrid Intelligent Systems*, Beijing, China, 2009, pp. 344–349.

[25] A. Dallali, S. El Khediri, A. Slimen, and A. Kachouri, "Breast tumors segmentation using Otsu method and K-means," in *2018 4th International Conference on Advanced Technologies For Signal and Image Processing (ATSIP)*, Sousse, Tunisia, 2018, pp. 1–6.

[26] Q. B. Baker and A. A. Qutaish, "Evaluation of histopathological images segmentation techniques for breast cancer detection," in *2021 12th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2021, pp. 134–139.

[27] K. Y. Win and S. Choomchuay, "Automated segmentation of cell nuclei in cytology pleural fluid images using OTSU thresholding," in *2017 10th Biomedical Engineering International Conference (BMEICON)*, Hokkaido, Japan, 2017, pp. 1–5.

[28] F. Yi et al., "Cell nuclei segmentation based on morphological operations and watershed transform," *Journal of Physics: Conference Series*, vol. 806, no. 1, p. 012049, 2017.

[29] M. Veta, P. J. Van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in H&E stained breast cancer histopathology images," *PLoS ONE*, vol. 8, no. 7, p. e70221, 2013.

[30] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, Jun. 1991.

[31] J. B. T. M. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta Informaticae*, vol. 41, no. 1-2, pp. 187–228, 2000.

[32] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.

[33] S. Liu, P. A. Mundra, and J. C. Rajapakse, "Features for cells and nuclei classification," in *Proc. 33rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Boston, MA, USA, 2011, pp. 6601–6604.

[34] T. Kanungo et al., "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[35] C. Mazo, M. Trujillo, E. Alegre, and L. Salazar, "Automatic recognition of fundamental tissues on histology images of the human cardiovascular system," *Micron*, vol. 89, pp. 1–8, Oct. 2016.

[36] H. Su, Z. Yin, T. Kanade, and S. Huh, "Interactive cell segmentation based on correction propagation," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 2014, pp. 1381–1384.

[37] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

A Compact Deep Learning Architecture for Multi-Label ECG Classification Using the PTB-XL Dataset

Rogério Ferreira Rodrigues Júnior
Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará
Fortaleza, Brasil
rger0904@hotmail.com

João Samuel Maciel de Sales
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
joaosamuel.2001@gmail.com

João Paulo do Vale Madeiro
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
jpaulo.vale@dc.ufc.br

***Abstract*—Automatic electrocardiogram classification can support clinical decision-making by reducing the burden of manual signal interpretation and improving the consistency of diagnostic screening. However, many high-performing deep learning models for ECG analysis rely on relatively complex architectures, which may increase computational cost and limit reproducibility. This study proposes a compact neural architecture for multi-label ECG classification using the PTB-XL dataset, focusing on the five diagnostic superclasses: normal ECG, myocardial infarction, ST/T changes, conduction disturbances, and hypertrophy. The proposed model is inspired by the hierarchical organization of IMLE-Net but adopts a simpler design composed of a shallow 1D CNN encoder, temporal attention, mean aggregation across signal windows, and inter-lead attention. Each 10-second 12-lead ECG is segmented into partially overlapping windows, allowing the model to capture local temporal patterns while preserving information from different leads. Experiments were conducted using the official PTB-XL stratified folds, with folds 1–8 for training, fold 9 for validation, and fold 10 for testing. The proposed model achieved a macro ROC-AUC of 0.9296 and a mean accuracy of 0.8916 using a fixed threshold of 0.5, outperforming IMLE-Net in both metrics while using fewer trainable parameters. When a validation-optimized threshold was applied, the model reached an F1-score of 0.8103. These results indicate that compact architectures can achieve competitive performance for ECG superclass classification while reducing architectural complexity.**

***Index Terms*—ECG, PTB-XL, Deep Learning, CNN.**

I. Introduction

The electrocardiogram (ECG) is a vital medical tool used to record the heart's electrical activity, providing essential insights for the assessment of cardiac abnormalities. Interpreting these signals requires meticulous observation of complex morphological features, including waves, segments, intervals, and leads, as both signal morphology and temporal organization are critical for accurate diagnosis [1]. Despite its clinical importance, manual analysis of ECG recordings is labor-intensive and susceptible to intra- and inter-observer variability. To address these challenges, machine learning (ML)-based approaches have been developed to automate classification tasks, facilitating more consistent and efficient diagnostic support.

The public datasets available on PhysioNet reflect the diverse scenarios in which ML is applied to ECG analysis. One primary scenario is ambulatory monitoring, exemplified by the MIT-BIH Arrhythmia Database, which consists of 48 half-hour two-lead records primarily designed for arrhythmia detection [2]. In contrast, conventional short-duration clinical ECGs are represented by datasets like PTB-XL. This dataset includes 21,799 ten-second exams across 12 leads from 18,869 patients, serving as a standard benchmark for developing and evaluating automatic classification models, particularly for multi-label tasks such as identifying normal ECGs (NORM), myocardial infarction (MI), ST/T changes (STTC), conduction disturbances (CD), and hypertrophy (HYP) [3]–[5].

Beyond short-duration clinical recordings, large-scale hospital datasets such as MIMIC-IV-ECG have emerged, comprising approximately 800,000 diagnostic 12-lead electrocardiograms sampled at 500 Hz from over 160,000 patients [6]. Due to its scale and integration with clinical metadata, MIMIC-IV-ECG has been instrumental in the development of foundation models for ECG analysis [7]. These models typically employ large-scale self-supervised pre-training to learn generalizable signal representations, as seen in architectures like the transformer-based ECG-FM [8]. Furthermore, the PhysioNet/Computing in Cardiology Challenge 2020 introduced a scenario characterized by significant heterogeneity across datasets, labels, and acquisition conditions by aggregating multiple 12-lead ECG sources [9]. Consequently, the performance of ECG classification models is heavily influenced by factors such as record duration, lead configurations, dataset scale, and the underlying data origin.

In this study, the PTB-XL dataset is utilized, which provides diagnostic annotations organized at different levels of granularity [3], [10]. Specifically, the five diagnostic superclasses, NORM, MI, STTC, CD, and HYP, allow for the evaluation of model performance across broad clinical groups. This organization facilitated its inclusion in international challenges, such as the PhysioNet/Computing in Cardiology Challenge 2020, which focused on 12-lead cardiac abnormality classification, and the 2021 edition, which evaluated algorithms under vary-

ing lead configurations, including 12, 6, 4, 3, and 2 leads [11]–[13].

Thus, PTB-XL has become a relevant dataset for both standardized model comparison and the development of architectures aimed at automatic ECG classification. For instance, IMLE-Net was proposed to explore signal organization at the beat, rhythm, and channel levels, utilizing attention mechanisms to weight relevant information at each processing stage [5]. Other studies reinforce the importance of considering ECG leads during architecture design, as different channels contribute differently to classification performance [14].

Despite the strong results achieved by previous architectures like X-ECGNet, which reported a macro-AUC of 0.936, an F1 score of 0.750, and an accuracy of 0.903 in the diagnostic superclass task, reducing the number of parameters remains a relevant issue [14], [15]. Models with fewer parameters can lower training costs, facilitate experimental reproducibility, and support deployment on resource-constrained devices, such as embedded systems.

In this direction, Reference [16] presented a simplified CNN-VAE-based architecture for PTB-XL, reporting competitive results with only 197,093 trainable parameters. The model achieved 87.01% binary accuracy and a weighted F1 score of 0.7454 in the multi-label classification of the five PTB-XL diagnostic superclasses. These results indicate that compact architectures can indeed be effective in this scenario.

Given this context, this study investigates a compact architecture for ECG signal classification using the PTB-XL dataset, focusing on the five diagnostic superclasses. The proposed approach takes IMLE-Net as a conceptual reference but adopts a shallower structure, featuring a convolutional neural network (CNN) encoder, temporal attention, mean aggregation across windows, and inter-lead attention. The primary objective is to evaluate whether a model with fewer convolutional layers and parameters can still achieve competitive performance.

This work is organized as follows: Section 1 presents the introduction, contextualizing the problem, motivation, objectives, and text organization. Section 2 describes the main concepts related to the study and automatic electrocardiographic signal classification models. Section 3 outlines the proposed methodology, including signal preprocessing, window segmentation, model architecture, and the training and evaluation procedure. Section 4 discusses the obtained results, comparing the proposed model with literature benchmarks. Finally, Section 5 presents the conclusions, limitations, and potential directions for future work.

II. Theoretical Background

This section outlines the foundational concepts supporting this study. We begin with one-dimensional (1D) convolutional neural networks (CNNs), focusing on their application to ECG signals. Subsequently, we detail the hierarchical structure of the IMLE-Net model, which serves as a reference for processing electrocardiographic signals across different levels. Finally, we describe the attention mechanisms within this architecture, highlighting their mathematical formulation and their function in synthesizing representations throughout the model.

A. One-Dimensional Convolutional Neural Networks

Convolutional neural networks utilize shared filters, enabling the application of consistent parameters across different positions of the input. In 1D signals, these positions correspond to consecutive segments of the signal. Applying convolution along a single dimension, referred to as 1D convolution, allows for the extraction of local patterns from ECG traces. These patterns are typically associated with the morphology of waves, complexes, and segments of the signal.

In this context, for ECG signals, 1D convolutions can be used to extract local patterns from the electrocardiographic trace. These patterns may be associated with the morphology of the waves, complexes, and segments of the signal.

B. Hierarchical Models

Reference [5] proposed the IMLE-Net model for anomaly classification in 12-lead ECG signals, structuring the extraction of morphological features and the classification of anomalies into three sequential stages: heartbeat, cardiac rhythm, and ECG lead. This organization is illustrated in Fig. 1.

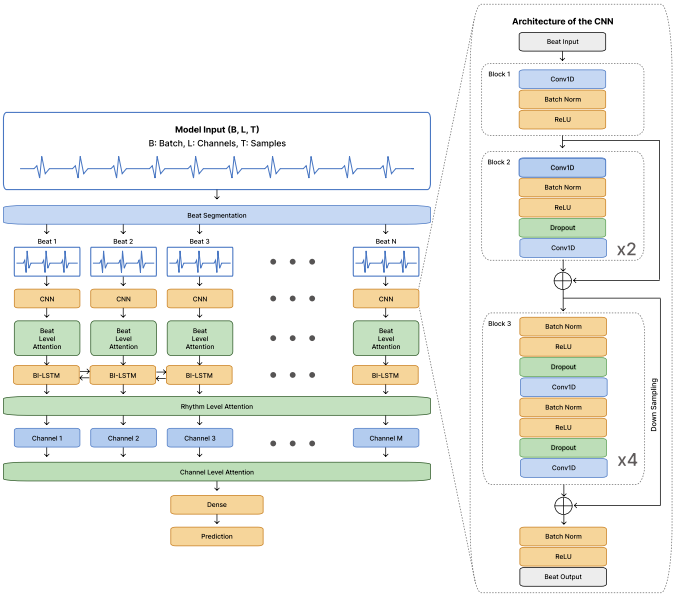


Fig. 1. Hierarchical architecture of IMLE-Net

In Fig. 1, the first stage illustrates processing at the individual ECG lead level. The lead is segmented into short, non-overlapping windows and the input signals with a sampling frequency of 100 Hz are divided into 50-sample windows. Each window is processed by a 1D CNN, followed by a heartbeat-level attention mechanism.

Subsequently, the representations of the windows from the same lead are processed by a BiLSTM, followed by a rhythm-level attention mechanism [17]. Finally, the third stage demonstrates how representations from different leads are aggregated

via a lead-level attention mechanism and forwarded to the dense layer for the final prediction. Consequently, IMLE-Net incorporates attention mechanisms at the heartbeat, rhythm, and lead levels prior to the final classification [5].

The attention functions in IMLE-Net are described below. The equations presented are an adaptation of the model's original mathematical description. Those functions weigh the representations extracted at different stages of ECG processing. At each level, the model assigns greater weight to the most relevant representations, combining this information through a weighted sum. This procedure, applied at the beat, rhythm, and lead levels, allows the model to emphasize specific parts of the signal and particular channels before classification [5].

At the beat level, let (b_k^c) be the (k)-th segment of lead (c). This segment is processed by the 1D CNN, producing the sequence $(\overline{\mathbf{b}}_k^c)$ of intermediate length $(\overline{W})$. The attention function at the beat level is defined as:

$$\boldsymbol{\alpha}_k^c = f_b\left(\overline{\mathbf{b}}_k^c\right) = \operatorname{softmax}\left(\mathbf{V}_b \tanh\left(\mathbf{W}_b \overline{\mathbf{b}}_k^c + \mathbf{b}_b\right)\right). \quad (1)$$

In Equation 1, $\boldsymbol{\alpha}_k^c$ represents the attention weights assigned to the internal positions of segment k from lead c, where $\mathbf{W}_b$, $\mathbf{V}_b$, and $\mathbf{b}_b$ are learned parameters. The segment representation at the beat level is obtained through the weighted sum:

$$\mathbf{B}_k^c = \sum_{i=1}^{\overline{W}} \alpha_{k,i}^c \overline{\mathbf{b}}_{k,i}^c. \quad (2)$$

In Equation 2, i indexes the positions of the sequence produced by CNN, $\alpha_{k,i}^c$ is the attention weight of position i, $\overline{\mathbf{b}}_{k,i}^c$ is the convolutional representation at that position, and $\mathbf{B}_k^c$ is the context vector of segment k from lead c. At the rhythm level, the context vectors of the segments from the same lead are processed by a BiLSTM, generating the sequence $\mathbf{r}^c = \mathbf{r}_1^c, \mathbf{r}_2^c, \ldots, \mathbf{r}_N^c$, where N is the number of segments in lead c. The attention function at the rhythm level is defined as:

$$\boldsymbol{\beta}^c = f_r\left(\mathbf{r}^c\right) = \operatorname{softmax}\left(\mathbf{V}_r \tanh\left(\mathbf{W}_r \mathbf{r}^c + \mathbf{b}_r\right)\right). \quad (3)$$

In Equation 3, $\boldsymbol{\beta}^c$ represents the attention weights assigned to the segments of the lead c, and $\mathbf{W}_r$, $\mathbf{V}_r$, and $\mathbf{b}_r$ are learned parameters. The lead representation at the rhythm level is obtained by:

$$\mathbf{R}^c = \sum_{i=1}^{N} \beta_i^c \mathbf{r}_i^c. \quad (4)$$

In Equation 4, β_i^c is the attention weight of segment (i), $\mathbf{r}_i^c$ is the representation of this segment generated by the BiLSTM, and $\mathbf{R}^c$ is the context vector of lead c. At the lead level, the representations $\mathbf{R}^1, \mathbf{R}^2, \ldots, \mathbf{R}^M$ are considered jointly, where M is the number of available leads. The attention function at this level is defined as:

$$\boldsymbol{\gamma} = f_c\left(\mathbf{R}\right) = \operatorname{softmax}\left(\mathbf{V}_c \tanh\left(\mathbf{W}_c \mathbf{R} + \mathbf{b}_c\right)\right). \quad (5)$$

In Equation 5, $\mathbf{R}$ represents the set of lead representations, $\boldsymbol{\gamma}$ represents the attention weights assigned to these leads, and $\mathbf{W}_c$, $\mathbf{V}_c$, and $\mathbf{b}_c$ are parameters learned. The final signal representation is obtained by:

$$\mathbf{C} = \sum_{i=1}^{M} \gamma_i \mathbf{R}^i. \quad (6)$$

In Equation 6, γ_i is the attention weight assigned to lead i, $\mathbf{R}^i$ is the representation of that lead, and $\mathbf{C}$ is the final context vector of the ECG, used by the linear layer for final prediction.

III. Methodology

This research adopts an experimental design to evaluate a CNN architecture for the multi-disease classification of ECG signals in the PTB-XL database. The objective is to assess the performance of the proposed architecture in classifying the five PTB-XL diagnostic superclasses, considering its organization into convolutional blocks, the resulting temporal receptive field, and the total number of trainable parameters.

A. Dataset

The PTB-XL dataset is a public database of clinical 12-lead ECGs, consisting, in the version used in this study, of 21,799 ten-second exams [3]. The exams include multi-disease diagnostic annotations provided by up to two cardiologists [10]. Notably, not all records have confirmed human validation: approximately 73.7% of the records were validated by a cardiologist, whereas the remaining 26.3% do not have this confirmation [10].

Furthermore, the diagnoses are organized in a hierarchical structure with three levels of granularity. The most specific and detailed level consists of 71 diagnostic statements, such as anteroseptal infarction and diffuse T-wave abnormality, encoded according to the Standard Communications Protocol for Computer-Assisted Electrocardiography [3]. These statements are grouped into 24 intermediate subclasses, such as bundle branch blocks and myocardial infarction, which are subsequently condensed into five general diagnostic superclasses [3].

This hierarchy provides flexibility in the choice of the automatic ECG classification task, allowing anything from the prediction of highly specific diagnoses to the classification of ECG disease groups. Given this, the study adopts the same general superclass classification in IMLE-Net model, as presented in Table I.

TABLE I
DEFINITION OF DIAGNOSTIC SUPERCLASSES

Class	Description
NORM	Normal ECG
MI	Myocardial infarction
STTC	ST/T changes
CD	Conduction disturbances
HYP	Hypertrophy

Table I presents the five PTB-XL diagnostic superclasses. These categories aggregate more specific labels in the dataset;

for example, right bundle branch block and left anterior fascicular block are both subclasses of CD, anterior myocardial infarction is a subclass of MI, and left ventricular hypertrophy is a subclass of HYP.

B. Data Stratification

This study utilizes the official PTB-XL stratified folds, defined in the *strat_fold* metadata field [10]. This stratification was designed to ensure comparability across studies and prevent data leakage between training and test sets. Allocation criteria included diagnosis, patient sex, and age to maintain balanced folds, with the fundamental constraint that all records from a single patient remained within the same fold.

To ensure high label quality, patients whose exam labels had not been validated by human experts were excluded from folds 9 and 10 during stratified sampling. This ensures that all records in these subsets, which are used as the validation and test sets, have undergone at least one confirmed human validation, as recommended in the literature [10]. Consequently, folds 1 through 8 were used for training, fold 9 for validation, and fold 10 for testing, as detailed in Table II.

TABLE II
DATA SPLIT AND DISTRIBUTION OF DIAGNOSTIC SUPERCLASSES

Set	**Folds**	**Number of exams**	**CD**	**HYP**	**MI**	**NORM**	**STTC**
Training	1–8	17,418	3,907	2,119	4,379	7,596	4,186
Validation	9	2,183	495	268	540	955	528
Test	10	2,198	496	262	550	963	521

Table II presents the partitioning of the 21,799 PTB-XL exams. The training set (folds 1–8) comprises 17,418 exams, while the validation (fold 9) and test (fold 10) sets include 2,183 and 2,198 exams, respectively. The table also details the distribution across the five diagnostic superclasses: CD, HYP, MI, NORM, and STTC. The folds 9 and 10 consist exclusively of records with confirmed human validation; they provide a robust foundation for the final evaluation of model performance.

C. Model

This study uses IMLE-Net as a conceptual reference for the hierarchical organization of the ECG while adopting a simpler architecture. The central idea of the proposed model is to process raw signals in three stages: first, a 1D CNN represents each signal window; second, a temporal attention function weights the most relevant points of this representation; finally, an inter-lead attention function determines the contribution of each lead to the classification.

The model input is defined as $x \in \mathbb{R}^{L \times W \times T}$, where L denotes the number of leads, W the number of windows, and T the samples per window. Using 12 leads ($L = 12$), each 10-second record is divided into 1-second segments. We construct 300-sample windows by combining the central 1-second segment with 100 samples from the immediately preceding and following segments, respectively. This approach ensures each window captures the target segment along with surrounding context, as illustrated in Fig. 2. The window slides every two segments, resulting in five windows per lead with a 100-sample overlap between consecutive windows. If no subsequent segment exists for the final window, we apply zero padding to the missing portion.



Fig. 2. ECG signal segmentation

Each window from each lead is processed individually by a 1D CNN encoder shared across all leads and windows. This encoder receives a single lead at a time and consists of three *Conv1D* layers. The first layer transforms the input signal from 1 channel into 64 filters; the second maintains 64 filters to refine the representation; and the third increases the dimension to 128 filters for the attention modules. The three convolutional layers use kernels of size $k_1 = 39$, $k_2 = 19$, and $k_3 = 9$, respectively, all with a stride of $s = 1$ and padding of $p_1 = 19$, $p_2 = 9$, and $p_3 = 4$ samples. The encoder configuration is summarized as follows:

$$1 \xrightarrow[p=19]{k=39} 64 \xrightarrow[p=9]{k=19} 64 \xrightarrow[p=4]{k=9} 128. \tag{7}$$

After each convolution, we apply layer normalization, ReLU activation, and dropout. Layer normalization is preferred over batch normalization to ensure that normalization does not depend on mini-batch statistics, allowing for more independent processing of each example during training.

The temporal attention mechanism follows the beat-level attention formulation of IMLE-Net. Given the feature sequence produced by the CNN, it projects each temporal position into a 64-dimensional attention space, computes normalized attention weights using a softmax function, and obtains the window representation through a weighted sum. Layer normalization, ReLU, and dropout are part of the CNN encoder and not of the temporal attention mechanism.

The representations obtained for windows of the same lead are then aggregated by averaging. This operation reduces the temporal windows into a single per-lead representation without requiring recurrent modules. Unlike IMLE-Net, which models sequences at the rhythm level, this model uses averaging to consolidate representations already processed by the convolutional and temporal attention stages.

Next, the representations from different leads are combined via an inter-lead attention function. This stage assigns weights

to the ECG leads, enabling the model to emphasize the most informative channels for each input. The resulting representation is forwarded to a projection layer consisting of linear layers, layer normalization, ReLU, and dropout to produce the final classification output. In summary, the model flow can be described as:

$$\mathbf{h}_{l,w} = A_t\left(g_\theta(x_{l,w})\right), \tag{8}$$

$$\bar{\mathbf{h}}_l = \frac{1}{W}\sum_{w=1}^{W} \mathbf{h}_{l,w}, \tag{9}$$

$$\mathbf{z} = A_l\left(\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \ldots, \bar{\mathbf{h}}_L\right), \tag{10}$$

$$\hat{\mathbf{y}} = p_\phi(\mathbf{z}). \tag{11}$$

where $x_{l,w}$ represents window w of lead l, g_θ is the 1D CNN encoder, A_t is the temporal attention mechanism, $\mathbf{h}_{l,w}$ is the window representation, $\bar{\mathbf{h}}_l$ is the aggregated representation of lead l, A_l is the inter-lead attention function, $\mathbf{z}$ is the final ECG representation, p_ϕ is the classification head, and $\hat{\mathbf{y}}$ is the predicted output. Reference [5] describes the beat-level attention formulation of IMLE-Net, which is adopted for the temporal attention mechanism A_t in this work. The overall architecture is presented in Fig. 3.

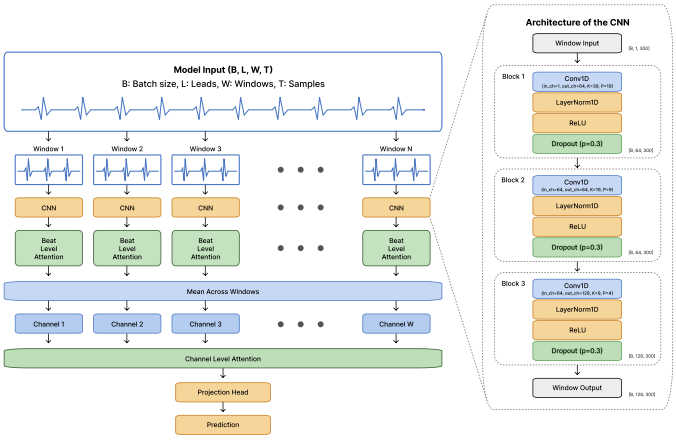


Fig. 3. Architecture of the proposed model

Fig. 3 summarizes the model flow, illustrating the hierarchical processing from the windowed input to the final prediction, with details of the CNN encoder shown on the right.

D. Training and Evaluation

During the main training procedure, a single window is randomly selected for each exam at each iteration. Gaussian noise with zero mean and a specified standard deviation (0.005) is added to this window. In addition, the dropbeat is applied with a probability of 0.2. When applied, a contiguous region corresponding to 20% of the selected window is masked with zeros at a randomly chosen temporal position. Since the last window contains zero padding by construction, the dropbeat introduces zero-valued regions at different temporal positions, reducing the risk that the model associates zeros exclusively with the padded region or treats their position as an informative feature. Furthermore, this temporal masking acts as a regularization strategy by preventing the model from relying on specific local segments and encouraging the temporal attention mechanism to prioritize the remaining informative regions.

Unlike training, the evaluation process incorporates all available windows from each exam. The representations obtained for windows of the same lead are aggregated by averaging before the inter-lead attention stage. Thus, the architecture retains the capability to utilize multiple windows during inference, while maintaining a single-window approach during training to reduce computational costs.

A partial overlap between consecutive windows was adopted to allow the network to observe the same portion of the signal within multiple temporal contexts. This allows segments present in one window to appear in the next, enabling the same part of the signal to be analyzed at two different moments. Conversely, sliding windows with minimal advance were not used; excessive overlap increases processing and memory usage without necessarily yielding significant performance gains [18].

Furthermore, the model does not require processing all windows for every training iteration. While this simplification might slightly reduce performance compared with using full window sequences, it enhances training efficiency and allows for the evaluation of a compact architecture based on its total trainable parameters. Accordingly, comparisons with models from the literature are performed at the level of the complete model, considering total parameter cost and classification metrics, rather than through layer-by-layer equivalence with reference architectures.

1) Model Parameters: Reference [16] presented a simplified CNN-VAE-based architecture for the PTB-XL dataset, using only three convolutional layers and 197,093 trainable parameters. The model reported a binary accuracy of 87.01% and an AUC of 0.8958. Similarly, this study also adopted an architecture with few convolutional layers, avoiding deeper structures with a larger number of convolutional blocks, such as IMLE-Net.

As described in Section III-C, the CNN encoder used in the model is composed of three 1D convolutional layers. In this subsection, this configuration is revisited only to explicitly calculate the temporal receptive field of the encoder. The three layers use kernels of size $k_1 = 39$, $k_2 = 19$, and $k_3 = 9$, respectively, all with a stride equal to $s_1 = s_2 = s_3 = 1$. With this configuration, the final receptive field of the encoder is calculated as:

$$R = 1 + (39 - 1) + (19 - 1) + (9 - 1) = 65. \tag{12}$$

Thus, each position in the final output of the encoder can depend on up to 65 samples from the input window. Since the signals used in this study are sampled at 100 Hz, this receptive

field corresponds to approximately 0.65 seconds. This interval covers durations associated with local ECG components, such as the P wave, whose normal duration is less than 120 ms, and the QRS complex, whose normal duration is approximately between 70 and 100 ms [19], [20]. In addition, it also covers ranges compatible with compound measurements, such as the PR interval, which is normally between 120 and 200 ms [21]. Therefore, the kernels used allow the network to consider isolated deflections as well as regions formed by nearby waves, segments, and intervals.

The last convolutional layer produces 128 filters. This choice was adopted based on IMLE-Net model. Thus, the proposed model maintains this final dimension as a reference before the attention modules. The temporal and inter-lead attention functions use an internal dimension equal to 64. After lead aggregation, the final representation is forwarded to the classification head.

Fig. 4 presents the classification head used in the proposed model. The input to this stage has dimension (B,128), where B represents the batch size. First, this representation passes through a linear layer with 128 units, followed by *LayerNorm* normalization, the *ReLU* activation function, and dropout with a probability of 0.3. Next, a final linear layer projects the representation to five output neurons, corresponding to the five diagnostic superclasses evaluated in this study. The final output has dimension (B,5) and corresponds to the logits used by the *BCEWithLogitsLoss* function.

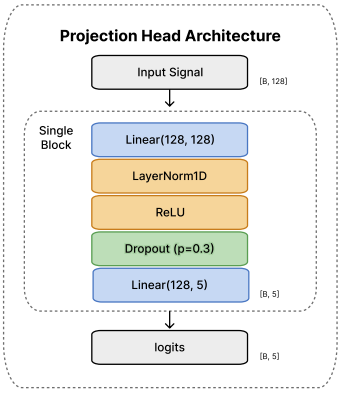


Fig. 4. Classification head of the proposed model

Table III compares the parameter distribution of the proposed model with the official implementation of IMLE-Net.

TABLE III
COMPARISON OF THE PARAMETER DISTRIBUTION BETWEEN THE PROPOSED MODEL AND IMLE-NET

Component	This work	IMLE-Net
CNN encoder	154,816	619,616
Temporal attention	27,456	9,088
Recurrent layer	–	98,816
Rhythm attention	–	9,536
Inter-lead attention	9,024	9,024
Classification head	17,413	645
Trainable parameters	208,709	746,725
Non-trainable parameters	–	1,792
Total	208,709	748,517

In both cases, most of the parameters are concentrated in the convolutional layers. However, this concentration is lower in the proposed model: the CNN encoder has 154,816 parameters, whereas the convolutional block of IMLE-Net has 619,616 parameters. This difference shows that the reduction in parametric cost does not occur only because of the absence of a recurrent stage. IMLE-Net has a specific number of parameters in the recurrent layer and rhythm attention mechanism, but the largest difference between the models lies in the convolutions. Thus, the main factor contributing to the reduction is the use of a more compact CNN encoder, with fewer layers and lower overall depth.

In total, the proposed model has 208,709 trainable parameters, whereas IMLE-Net has 746,725 trainable parameters. Considering the non-trainable parameters, the official implementation of IMLE-Net presents a total of 748,517 parameters. The proposed model, in turn, has no non-trainable parameters counted in this comparison and did not undergo hyperparameter tuning.

2) Training Configuration: To enhance reproducibility, training was conducted using a fixed seed of 42 to initialize random generators. Experiments were executed on an NVIDIA GeForce RTX 3050 Laptop GPU, utilizing a deterministic cuDNN configuration with benchmark mode disabled. The model was trained for up to 256 epochs, incorporating early stopping with a patience of 20 epochs. We adopted the Binary Cross-Entropy with Logits loss function, without class weights or data balancing, consistent with the training of IMLE-Net [5]. Weights were optimized using the AdamW optimizer [22], with a learning rate of 10^{-3} and a weight decay of 10^{-2}. During validation, a fixed threshold of 0.5 was applied to convert predicted probabilities into binary labels. Model selection was based on validation accuracy, and the model yielding the best performance was saved in a checkpoint directory, allowing for the recovery of its weights.

E. Evaluation Metrics

The evaluation of ECG classification models often combines threshold-independent and threshold-dependent metrics. ROC-AUC evaluates the model's discriminative ability without relying on a fixed decision threshold, whereas metrics such as accuracy and F1 depend on the binarization of model outputs and the specific threshold criteria applied. In this study, we employ the same metrics as those reported for IMLE-Net, which achieved notable performance on the PTB-XL dataset [5]. This approach enables a direct comparison between models, allowing for an assessment of both discriminative ability and performance at defined decision thresholds.

IV. RESULTS

This section presents the performance of the proposed model on the PTB-XL dataset, comparing it with established literature. The comparison focuses on the metrics: macro ROC-AUC, average accuracy, and F1 score. For the proposed model, we report two operating points: one with a fixed threshold of 0.5, and another with a global threshold optimized on

the validation set. In the comparative tables, the rank column indicates each model's position relative to the corresponding metric, where 1 denotes the best result.

Table IV presents the class-wise ROC-AUC comparison between the proposed model and the literature values, which are based on the results reported in the IMLE-Net work. The proposed model performs competitively with IMLE-Net across the five PTB-XL diagnostic superclasses, achieving higher ROC-AUC values in CD, HYP, MI, and NORM, and only slightly lower performance in STTC. Among all compared models, the proposed model obtains the best ROC-AUC values in CD, MI, and NORM.

TABLE IV
CLASS-WISE ROC-AUC COMPARISON

Model	CD		HYP		MI		NORM		STTC	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank
ResNet101	0.8928	4	0.8636	5	0.8921	5	0.9098	5	0.9178	5
Mousavi et al.	0.8555	6	0.8093	6	0.8708	6	0.9096	6	0.8817	6
ECGNet	0.9102	3	0.8842	3	0.9021	4	0.9336	3	0.9205	4
Rajpurkar et al.	0.8910	5	**0.9185**	**1**	0.9066	3	0.9309	4	0.9306	3
IMLE-Net	0.9196	2	0.8834	4	0.9244	2	0.9450	2	**0.9354**	**1**
This work	**0.9301**	**1**	0.9054	2	**0.9268**	**1**	**0.9510**	**1**	0.9349	2

Table V presents the class-wise accuracy comparison. For the proposed model, the values were calculated using the standard 0.5 threshold, consistent with the methodology adopted in the IMLE-Net study. The proposed model performs competitively across the five PTB-XL diagnostic superclasses, achieving higher accuracy than IMLE-Net in CD, HYP, and NORM. Among all compared models, the proposed model obtains the highest accuracy values in CD, HYP, and NORM.

TABLE V
CLASS-WISE ACCURACY COMPARISON

Model	CD		HYP		MI		NORM		STTC	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank
ResNet101	0.8774	5	0.9089	3	0.8312	5	0.8469	5	0.8747	4
Mousavi et al.	0.8626	6	0.8719	6	0.8247	6	0.8178	6	0.8326	6
ECGNet	0.8871	3	0.9066	4	0.8460	4	0.8589	4	0.8687	5
Rajpurkar et al.	0.9001	2	0.8876	5	0.8594	3	0.8594	3	0.8890	2
IMLE-Net	0.8853	4	0.9158	2	**0.8797**	**1**	0.8710	2	**0.8908**	**1**
This work	**0.9008**	**1**	**0.9195**	**1**	0.8717	2	**0.8776**	**1**	0.8885	3

Table VI summarizes the overall performance of the proposed model in comparison with the models evaluated in the IMLE-Net study. For the proposed model, we report both the results obtained with the fixed 0.5 threshold and those obtained at the maximum F1-score threshold.

TABLE VI
OVERALL COMPARISON

Model	Year	Macro ROC-AUC		Mean Accuracy		F1	
		Value	Rank	Value	Rank	Value	Rank
ResNet101	2016	0.8952	5	0.8678	6	0.7558	6
Mousavi et al.	2019	0.8654	6	0.8419	7	0.7315	7
ECGNet	2018	0.9101	4	0.8735	5	0.7712	5
Rajpurkar et al.	2019	0.9155	3	0.8791	4	0.7895	4
IMLE-Net	2022	0.9216	2	0.8885	2	0.8057	2
This work (threshold 0.50)	2026	**0.9296**	**1**	**0.8916**	**1**	0.7969	3
This work (threshold 0.33)	2026	**0.9296**	**1**	0.8851	3	**0.8103**	**1**

Reference [5] report the maximum F1 obtained through a threshold sweep on the test set, which differs from an F1 calculated using a validation-set threshold, we provide our maximum F1 result as an additional reference. This allows for a more equitable comparison with the protocol used in the IMLE-Net study, while the 0.5 fixed threshold remains our primary, reproducible operating point.

Overall, the proposed model achieves a macro ROC-AUC of 0.9296, exceeding the 0.9216 reported by IMLE-Net. Using a 0.5 threshold, it also surpasses IMLE-Net in average accuracy (0.8916 versus 0.8885). Although our F1 score of 0.7969 at this threshold is below IMLE-Net's 0.8057, applying the validation-set optimized threshold of 0.3333 increases our F1 to 0.8103, outperforming the literature. We emphasize that this comparison should be interpreted cautiously due to the differing thresholding methodologies. These results indicate that the proposed simplification maintains competitive performance.

Table VII presents an extended comparison of results from the literature on the PTB-XL dataset, allowing the proposed model to be situated among a broader set of recent architectures. References [14] and [16] are used to construct the extended comparison. When a metric was not reported in the original study, the corresponding field is marked as "–", and no rank is assigned.

TABLE VII
EXTENDED COMPARISON

Model	Year	Macro ROC-AUC		Mean Accuracy		F1	
		Value	Rank	Value	Rank	Value	Rank
Bidirectional LSTM	2020	0.9210	9	–	–	–	–
ResNet1D	2020	0.9300	4	–	–	–	–
Inception1D	2020	0.9210	9	–	–	–	–
XResNet1D101	2020	0.9280	7	0.8850	7	0.7370	8
MLBF-Net	2021	0.9310	3	–	–	–	–
IMLE-Net	2022	0.9216	8	0.8885	5	0.8057	2
IM-ECG	2025	0.9290	6	0.8920	3	0.7700	5
X-ECGNet	2025	**0.9360**	**1**	**0.9030**	**1**	0.7500	6
L5G-Net	2025	0.9357	2	0.8976	2	0.7706	4
Simplified CNN-VAE	2026	0.8958	10	0.8701	8	0.7454	7
This work (threshold 0.5)	2026	0.9296	5	0.8916	4	0.7969	3
This work (threshold 0.33)	2026	0.9296	5	0.8851	6	**0.8103**	**1**

The F1 and accuracy values for the external models are presented as reported in their original studies. Reference [14] addresses the PTB-XL superclass classification task but does not utilize the same threshold-selection procedure as this work. Therefore, the F1 comparison should be interpreted as a compilation of reported literature values rather than a direct comparison of strictly equivalent experimental protocols.

In this extended comparison, X-ECGNet achieves the highest macro ROC-AUC and mean accuracy, with 0.9360 and 0.9030, respectively. L5G-Net also demonstrates strong performance, with a macro ROC-AUC of 0.9357 and a mean accuracy of 0.8976. The proposed model does not outperform these architectures in terms of macro ROC-AUC or mean accuracy. However, with the fixed 0.5 threshold, it achieves an F1-score of 0.7969, outperforming most of the external models with reported F1 values. When the global threshold optimized on the validation set is used, the F1-score increases to 0.8103, representing the highest F1-score among the values included in the comparison.

Compared with the simplified CNN-VAE, the proposed model achieves higher macro ROC-AUC, mean accuracy, and F1-score. This result is significant as both models aim to reduce architectural complexity. However, the proposed model explicitly preserves the organization by windows and leads,

drawing inspiration from the hierarchical structure of IMLE-Net, whereas the simplified CNN-VAE adopts a more direct CNN architecture designed for automatic feature extraction from the ECG signal [16].

V. Conclusion

This work presented a compact deep learning architecture for multi-label ECG classification on the PTB-XL dataset, targeting the five diagnostic superclasses NORM, MI, STTC, CD, and HYP. The proposed model provides a simplified alternative to hierarchical architectures such as IMLE-Net, preserving the organization of ECG signals into windows and leads while removing recurrent components and reducing the depth of the CNN encoder. It combines a shallow 1D CNN, temporal attention, mean aggregation across windows, inter-lead attention, and a classification head.

The proposed simplification maintained competitive performance compared with models from the literature. The model achieved a macro ROC-AUC of 0.9296 and a mean accuracy of 0.8916 at a fixed threshold of 0.5, surpassing the corresponding values reported for IMLE-Net. Although its F1-score at this threshold was slightly lower, optimization of the threshold on the validation set increased the F1-score to 0.8103, the highest value among the compared results. Moreover, the proposed architecture uses 208,709 trainable parameters, substantially fewer than the 746,725 reported for IMLE-Net, supporting the feasibility of compact models for ECG classification.

Some limitations should be considered. The model was evaluated only on PTB-XL and the five superclass labels, limiting conclusions about generalization to other datasets, acquisition protocols, lead configurations, and more specific diagnoses. Comparisons with external models should also be interpreted with caution because threshold selection and evaluation protocols may differ across studies. In addition, extensive hyperparameter optimization was not performed.

Future work should evaluate the architecture on other ECG datasets and reduced lead configurations, as well as investigate alternative aggregation and attention strategies. Further studies may include hyperparameter tuning, calibration, class imbalance handling, and interpretability analyses of the temporal windows and leads emphasized by the attention mechanisms. Overall, the results indicate that compact attention-based CNN architectures can provide a favorable balance between predictive performance, parameter efficiency, and practical applicability in automatic ECG analysis.

Acknowledgment

The following Brazilian research agencies have supported this work: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) — Proc: 110945/2026-3, 140788/2024-7 and 201727/2025-0.

References

[1] P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock *et al.*, "Recommendations for the Standardization and Interpretation of the Electrocardiogram," *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007.

[2] G. Moody and R. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[3] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," Nov. 2022, version 1.0.3.

[4] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep Learning for ECG analysis: Benchmarks and Insights from PTB-XL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2021.

[5] L. Reddy, V. Talwar, S. Alle, R. S. Bapi, and U. D. Priyakumar, "IMLE-Net: An Interpretable Multi-level Multi-channel Model for ECG Classification," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 1068–1074.

[6] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes *et al.*, "MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset," Sep. 2023, version 1.0.

[7] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx *et al.*, "On the Opportunities and Risks of Foundation Models," 2022.

[8] K. McKeen, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: an open electrocardiogram foundation model," p. ooaf122, 10 2025.

[9] E. A. Perez Alday, A. Gu, A. Shah, C. Liu, A. Sharma, S. Seyedi *et al.*, "Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020," *PhysioNet*, Jul. 2022, version 1.0.2.

[10] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek *et al.*, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, p. 154, 2020.

[11] E. A. Perez Alday, A. Gu, A. J. Shah, C. Robichaux, A. I. Wong, C. Liu *et al.*, "Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020," *Physiological Measurement*, vol. 41, no. 12, p. 124003, 2020.

[12] M. A. Reyna, N. Sadr, E. A. Perez Alday, A. Gu, A. J. Shah, C. Robichaux *et al.*, "Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021," in *2021 Computing in Cardiology (CinC)*, vol. 48, 2021, pp. 1–4.

[13] M. Reyna, N. Sadr, A. Gu, E. A. Perez Alday, C. Liu, S. Seyedi *et al.*, "Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021," *PhysioNet*, Jul. 2022, version 1.0.3.

[14] T. Yang, C.-X. Xie, H.-M. Huang, Y. Wang, M.-H. Fan, I.-C. Kuo *et al.*, "Lead Analysis for the Classification of Multi-Label Cardiovascular Diseases and Neural Network Architecture Design," *Electronics*, vol. 14, no. 16, 2025.

[15] Y. C. Hadj Azzem and F. Harrag, "Explainable Deep Learning-Based System for Multilabel Classification of 12-Lead ECG," in *2023 International Conference on Networking and Advanced Systems (ICNAS)*, 2023, pp. 1–6.

[16] N. A. Mehdi, S. Q. Khan, A. Ghani, and A. A. Drigh, "ECG Classification on PTB-XL: A Data-Centric Approach with Simplified CNN-VAE," pp. 1–6, 2026.

[17] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, "A Quantitative Comparison of Overlapping and Non-Overlapping Sliding Windows for Human Activity Recognition Using Inertial Sensors," *Sensors*, vol. 19, no. 22, 2019.

[19] M. Cadogan and R. Buttner, "P Wave," Life in the Fast Lane ECG Library, Jan. 2022, [Online]. Available: https://litfl.com/p-wave-ecg-library/. Accessed: Jun. 24, 2026.

[20] J. Larkin, "QRS Interval," Life in the Fast Lane ECG Library, Aug. 2023, [Online]. Available: https://litfl.com/qrs-interval-ecg-library/. Accessed: Jun. 24, 2026.

[21] M. Cadogan, "PR Interval," Life in the Fast Lane ECG Library, Feb. 2021, [Online]. Available: https://litfl.com/pr-interval-ecg-library/. Accessed: Jun. 24, 2026.

[22] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," 2019.

Comparative Analysis of Deep Learning Architectures for Diabetic Retinopathy Classification

Gabriel de Sousa Nobre
Computer Science
University of Fortaleza
Fortaleza, Brazil
gabrielnobre.dev@gmail.com

Joel Sotero da Cunha Neto
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
joelsotero@unifor.br

Victor Lins Gurgel do Amaral
Computer Science
University of Fortaleza
Fortaleza, Brazil
vlinsgurgel@gmail.com

Lorenzo Barros Calheiros Pinheiro
Computer Science
University of Fortaleza
Fortaleza, Brazil
barroslorenzopro@gmail.com

Davi Oliveira de Alencar
Computer Science
University of Fortaleza
Fortaleza, Brazil
davioalencar10@gmail.com

Belmondo Rodrigues Aragão Junior
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
belmondo@unifor.br

***Abstract*—Diabetic retinopathy is a major cause of vision loss, highlighting the importance of developing automated methods to support disease screening and diagnosis. This study evaluated different deep learning architectures for classifying retinal fundus images into five levels of disease severity using the *Diabetic Retinopathy (resized)* dataset. Convolutional neural network and Vision Transformer architectures were compared using transfer learning, fine-tuning, data augmentation, and class weighting to address the substantial class imbalance in the dataset. The results showed that modern architectures, particularly ConvNeXt-Small and EfficientNetV2-S, achieved the highest overall performance. ConvNeXt-Small obtained the highest quadratic weighted Kappa score but showed very low sensitivity to the Mild class, whereas EfficientNetV2-S provided the most balanced trade-off between quadratic weighted Kappa and macro-averaged F1-score across the five severity levels. Because architecture, input resolution, and fine-tuning strategy varied jointly across experiments, these differences should be interpreted as configuration-level rather than purely architectural, and as validation-set results within this study rather than clinically validated performance. An ensemble strategy based on weighted soft voting was also evaluated; however, it did not outperform the individual models under the experimental configuration considered. Overall, the findings indicate that input image resolution, fine-tuning strategy, and appropriate handling of severe class imbalance are key factors affecting performance in multiclass diabetic retinopathy classification.**

***Index Terms*—Diabetic retinopathy, medical image classification, deep learning, convolutional neural networks, Vision Transformers, transfer learning.**

I. Introduction

Diabetic retinopathy (DR) is a microvascular complication of diabetes mellitus that affects the retinal blood vessels and constitutes one of the leading causes of preventable blindness among working-age adults [1], [2]. The continued growth of the global population with diabetes has substantially increased the demand for screening programs and early disease detection. According to Grzybowski [2], the worldwide expansion of diabetes and the increasing incidence of diabetic retinopathy have made the adoption of scalable ophthalmic screening solutions increasingly necessary, particularly in resource-constrained regions with limited access to specialists.

Traditionally, diabetic retinopathy is identified by ophthalmologists through the analysis of retinal fundus photographs. However, the large volume of examinations generated by population-based screening programs makes this process time-consuming, costly, and highly dependent on the availability of qualified professionals. In this context, artificial intelligence techniques, particularly those based on deep learning, have demonstrated considerable potential for automating medical image analysis and supporting clinical decision-making.

Among the pioneering studies in this field, Gulshan et al. [3] developed a system based on deep convolutional neural networks for the automated detection of diabetic retinopathy in retinal fundus photographs. The model achieved high sensitivity and specificity on independent validation datasets, demonstrating that deep learning algorithms can attain performance comparable to that of specialists in disease-screening tasks.

In recent years, architectural advances have further improved the performance of these systems. In addition to conventional Convolutional Neural Networks (CNNs), architectures based on attention mechanisms, known as Vision Transformers (ViTs), have achieved competitive results across several computer vision tasks. In a comparative study involving different retinal diseases, Saygılı and Atıcı [4] reported that the Swin Transformer outperformed conventional CNN architectures and other Transformer-based models, achieving the highest accuracy among the evaluated methods.

In parallel, approaches that combine multiple models have also been investigated to improve prediction robustness. Rajeshwar et al. [5] proposed a hybrid architecture based on deep learning feature extraction and a stacking ensemble, obtaining promising results for diabetic retinopathy detection. These studies reflect the growing interest of the scientific community in exploring different architectures and model-combination



Fig. 1: Examples of retinal fundus images from the dataset according to diabetic retinopathy severity level.

strategies to improve the performance of automated systems for ophthalmic screening.

Despite the advances reported in the literature, automated diabetic retinopathy classification still presents significant challenges. One of the main difficulties is class imbalance, since public datasets frequently contain a substantially larger number of images without signs of the disease than images representing more severe stages. Furthermore, diabetic retinopathy grading is both a multiclass and ordinal task, as the severity levels follow a clinical progression from no diabetic retinopathy (class 0) to proliferative diabetic retinopathy (class 4), according to the International Clinical Diabetic Retinopathy Disease Severity Scale [6], as illustrated in Fig. 1. Consequently, global metrics such as accuracy may not adequately represent model performance for minority classes, making the analysis of metrics such as the macro-averaged F1-score and quadratic weighted Kappa particularly important.

Within this context, this study aims to evaluate and compare different deep learning architectures for diabetic retinopathy classification using retinal fundus images. CNN- and Vision Transformer-based models were investigated, including ResNet50, DenseNet121, DenseNet169, EfficientNet-B0, EfficientNet-B2, EfficientNetV2-S, ConvNeXt-Small, and Swin-Tiny. Different fine-tuning strategies, data augmentation techniques, class-weighting schemes, and ensemble configurations were also evaluated to identify the most suitable approaches for the proposed classification task.

II. Methodology

The methodology consisted of an experimental comparison of different deep learning architectures for the multiclass classification of diabetic retinopathy in retinal fundus images. The experimental workflow included dataset preparation, image preprocessing, training of pretrained models through transfer learning, application of class-imbalance mitigation strategies, and model evaluation using metrics suitable for an ordinal and highly imbalanced classification problem.

A. Dataset

The publicly available *Diabetic Retinopathy (resized)* dataset [7], derived from the Kaggle diabetic retinopathy competition [8], was used in this study. The dataset contains 35,126 retinal fundus images labeled according to five severity levels: class 0, no diabetic retinopathy; class 1, mild diabetic retinopathy; class 2, moderate diabetic retinopathy; class 3, severe diabetic retinopathy; and class 4, proliferative diabetic retinopathy. Because these labels represent internationally established progressive levels of clinical severity [6], the problem was treated as an ordinal multiclass classification task.

The original dataset metadata are provided in a CSV file containing the image identifier and its corresponding severity level. To optimize model training, the images were reorganized into class-specific directories using a structure compatible with the *ImageFolder* data loader. The dataset was divided into training and validation subsets using an 80/20 ratio and stratified sampling according to class. To ensure reproducibility and consistency across all evaluated architectures, a fixed random seed of 42 was adopted. No independent external test set was used. The exact distribution of samples between the training and validation subsets is presented in Table I.

TABLE I: Dataset distribution by class

Class	Description	Total	Training	Validation
0	No DR	25,810	20,648	5,162
1	Mild	2,443	1,955	488
2	Moderate	5,292	4,234	1,058
3	Severe	873	699	174
4	Proliferative	708	567	141
Total	–	**35,126**	**28,103**	**7,023**

A substantial class imbalance can be observed, with class 0, representing images without signs of diabetic retinopathy, accounting for most of the samples. This imbalance motivated the use of evaluation metrics beyond accuracy and the application of class-weighting strategies during training.

B. Data Preprocessing and Class-Imbalance Handling

The images were resized according to the experimental configuration of each model. Input resolutions of 224×224, 256×256, 384×384, 512×512, and 1024×1024 pixels were evaluated. ImageNet mean and standard deviation values were used for normalization because the evaluated models were initialized with weights pretrained on the ImageNet dataset.

To increase the variability of the training data and reduce the risk of overfitting [9], data augmentation techniques were applied, including horizontal and vertical flips, rotations of up to 15°, and color transformations using *ColorJitter*. These transformations were applied exclusively to the training subset.

TABLE II: Summary of the main experimental configurations

Model	Resolution	Backbone	Augmentation	Optimizer (Head LR / Backbone LR)	Epochs	Remarks
EfficientNet-B0	224×224	Frozen	Horizontal/vertical flip	Adam (1e-4 / 1e-5)	25	Baseline
EfficientNet-B2	380×380	85% unfrozen	Flip, 15° rotation, ColorJitter	Adam (1e-4 / 1e-5)	19/25	Class weights applied
ResNet50	256×256	25% unfrozen	Flip, 15° rotation, ColorJitter	Adam (1e-4 / 1e-5)	23/25	Low macro F1-score
ResNet50	256×256	100% unfrozen	Flip, 15° rotation, ColorJitter	AdamW (1e-4)	24/25	Intermediate performance
DenseNet121	256×256	100% unfrozen	Flip, 15° rotation, ColorJitter	AdamW (1e-4)	21/25	High overall accuracy
DenseNet169	512×512	89.6% unfrozen	Flip, 15° rotation, ColorJitter	AdamW (1e-4 / 1e-5)	18/25	Best recall for severe classes
EfficientNetV2-S	384×384	85% unfrozen	Flip, 15° rotation, ColorJitter	Adam (1e-4 / 1e-5)	17/25	Improvement with larger images
EfficientNetV2-S	512×512	85% unfrozen	Flip, 15° rotation, ColorJitter	Adam (1e-4 / 1e-5)	14/25	**Best overall clinical balance**
Swin-Tiny	224×224	100% unfrozen	Flip, 15° rotation, ColorJitter	AdamW (1e-4)	25/25	Strong performance at low resolution
ConvNeXt-Small	512×512	100% unfrozen	Flip, 15° rotation, ColorJitter	AdamW (1e-4 / 1e-5)	29/50	Early stopping triggered
ConvNeXt-Small	1024×1024	100% unfrozen	Flip, 15° rotation, ColorJitter	AdamW (1e-4 / 1e-5)	3/50	**Highest overall Kappa**

Note: All models used a scheduler patience of two epochs and class weights to address class imbalance.

The validation images were not artificially augmented and underwent only resizing and normalization.

To address the observed class imbalance, a class-weighted *CrossEntropyLoss* function was used. The weights were calculated from the distribution of class labels in the training subset. Additional strategies, including *WeightedRandomSampler* and *Focal Loss*, were also explored. However, most of the main experiments relied primarily on class weighting within the loss function.

C. Evaluated Architectures and Experimental Variations

Architectures based on Convolutional Neural Networks and Vision Transformers were evaluated. The convolutional architectures included EfficientNet-B0, EfficientNet-B2 [10], EfficientNetV2-S [11], ResNet50 [12], DenseNet121, DenseNet169 [13], and ConvNeXt-Small [15]. Swin-Tiny [14] was evaluated as the representative architecture based on visual attention mechanisms.

Table II summarizes the main configurations and experimental variations investigated throughout the study. All models were initialized with weights pretrained on ImageNet, thereby characterizing a transfer learning approach [16]. The original classification layer of each architecture was replaced with a new layer containing five output units, corresponding to the five diabetic retinopathy severity levels.

Initially, the architectures were evaluated with a larger proportion of their pretrained layers frozen, providing baseline configurations. The most promising models were subsequently subjected to more specific adjustments, including partial or complete unfreezing of the backbone, increased input resolution, and changes to the optimization algorithm. This strategy enabled an assessment of the effects of fine-tuning and image resolution on classification performance.

D. Training Strategy

The experiments were implemented in Python using the PyTorch and torchvision libraries and were executed on CUDA-compatible NVIDIA GPUs. Because of the available GPU memory limitations, batch sizes were adjusted according to the architecture and input image resolution used in each experiment.

The Adam and AdamW optimization algorithms were used depending on the experimental configuration. In some experiments, different learning rates were assigned to the backbone and classification head, allowing smaller updates to the pretrained layers and greater adaptation of the new classification layer. The *ReduceLROnPlateau* scheduler was also used to reduce the learning rate when no improvement was observed in the monitored validation metric.

The maximum number of training epochs ranged from 25 to 50, depending on the evaluated model. To reduce overfitting, an early stopping strategy was applied, primarily monitoring the quadratic weighted Kappa score on the validation subset. The final version selected for each model corresponded to the best checkpoint obtained during training rather than necessarily the checkpoint from the final executed epoch.

E. Classification and Ensemble Strategy

The task was formulated as a direct multiclass classification problem. For each input image, the model generated a probability distribution across the five classes. The class with the highest predicted probability was assigned as the final classification output.

In addition to evaluating each model individually, an ensemble strategy based on *Weighted Soft Voting* was investigated. In this approach, the probability distributions generated by different models were combined using a weighted average. The weights were proportional to the individual performance of each architecture on the validation subset. The ensemble was evaluated to determine whether combining architectures from different model families could improve prediction robustness compared with the individual classifiers.

F. Evaluation Metrics

Model performance was assessed using accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and quadratic weighted Kappa. Accuracy was used as a general measure of the proportion of correct predictions. However, because of the substantial imbalance among the classes, accuracy was not considered sufficient when interpreted in isolation.

The macro-averaged F1-score assigns the same importance to each class regardless of its frequency in the dataset, providing a more appropriate assessment of model performance on minority classes. Quadratic weighted Kappa was adopted as the primary comparison metric because it accounts for the ordinal nature of the classification problem [17] and imposes greater penalties on errors involving classes that are farther apart on the disease-severity scale. Confusion matrices were

also analyzed to identify the main error patterns across diabetic retinopathy severity levels.

III. Results and Discussion

A. Overall Model Comparison

Table III summarizes the main results obtained in the experiments. To provide a concise comparison without including every tested configuration, the table presents the baseline model, the best-performing representative of each architectural family, and the final ensemble.

TABLE III: Comparison of the main evaluated models

Model	Res.	Accuracy	Macro F1	Kappa
EfficientNet-B0 (Baseline)	224	0.74	0.27	0.26
Swin-Tiny	224	0.69	0.52	0.63
DenseNet169	512	0.75	0.55	0.71
EfficientNetV2-S	512	0.80	0.54	0.73
ConvNeXt-Small	1024	0.82	0.53	0.75
Ensemble	–	0.73	0.39	0.47

The EfficientNet-B0 results illustrate a recurrent problem in highly imbalanced datasets. Although the model achieved an accuracy of 0.74, its Kappa score was only 0.26. This result indicates that the model predominantly learned to predict the majority *No DR* class, without adequately distinguishing among the different disease severity levels. This discrepancy reinforces the importance of evaluating the models using quadratic weighted Kappa and macro F1-score rather than relying exclusively on overall accuracy.

B. Best-Performing Individual Models

The three best-performing models, ConvNeXt-Small, EfficientNetV2-S, and DenseNet169, exhibited distinct classification profiles. These differences are particularly relevant when considering the potential application of such models in diabetic retinopathy screening.

ConvNeXt-Small achieved the highest Kappa score in the study (0.7453). However, it presented an important limitation, with a recall of only 0.05 for the *Mild* class and an F1-score of 0.09. This means that the model detected very few cases corresponding to the earliest stage of diabetic retinopathy, for which timely identification may be particularly important. Its high overall performance was strongly influenced by the majority *No DR* class, which represented approximately 73% of the dataset and achieved a recall of 0.93.

EfficientNetV2-S with an input resolution of 512 px provided the most balanced performance across the five classes, achieving a Kappa score of 0.7313 and a macro F1-score of 0.54. Recall values of 0.42 for *Severe* and 0.67 for *Proliferative* indicate that the model was able to identify relevant patterns in the most advanced disease classes despite the substantially smaller number of available samples.

The *Mild* class remained the most challenging, with a recall of 0.34 and an F1-score of 0.27. This result may be related both to the smaller number of samples in this class and to the subtle nature of early retinal abnormalities, which may be difficult to distinguish from healthy retinas or from moderate disease.

DenseNet169 with an input resolution of 512 px completed the group of best-performing models, achieving a Kappa score of 0.7114 and a macro F1-score of 0.55. The model showed slower convergence, reaching its best validation result at epoch 18, which suggests a more gradual optimization process. Its main strength was the performance on the *Proliferative* class, for which it achieved a recall of 0.67 and an F1-score of 0.67. Table IV presents the class-specific metrics for the three models.

TABLE IV: Class-specific metrics for the best-performing individual models

Model	Class	Prec.	Rec.	F1
ConvNeXt-S (1024 px)	No DR	0.90	0.93	0.92
	Mild	0.41	0.05	0.09
	Moderate	0.55	0.48	0.51
	Severe	0.52	0.28	0.36
	Prolif.	0.71	0.55	0.62
EfficientNetV2-S (512 px)	No DR	0.90	0.85	0.88
	Mild	0.22	0.34	0.27
	Moderate	0.59	0.60	0.60
	Severe	0.44	0.42	0.43
	Prolif.	0.57	0.67	0.61
DenseNet169 (512 px)	No DR	0.89	0.85	0.87
	Mild	0.17	0.30	0.22
	Moderate	0.62	0.51	0.56
	Severe	0.39	0.53	0.45
	Prolif.	0.67	0.67	0.67

C. Learning Curves

The convergence curves revealed distinct training behaviors among the evaluated models. Figures 2 and 3 present the training and validation loss and accuracy curves, respectively, for EfficientNetV2-S. The model showed stable convergence over 14 epochs, without a pronounced divergence between the training and validation curves before training was terminated by the early stopping criterion.

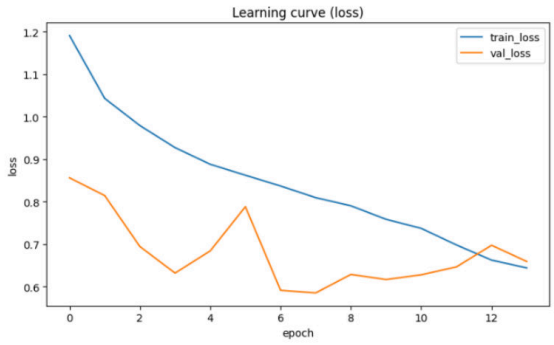


Fig. 2: Training and validation loss curves for EfficientNetV2-S at 512 px.

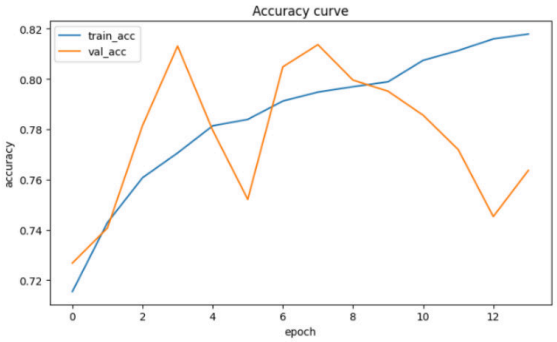


Fig. 3: Training and validation accuracy curves for EfficientNetV2-S at 512 px.

Figures 4 and 5 show the learning curves obtained for DenseNet169. Compared with EfficientNetV2-S, DenseNet169 required more epochs to stabilize but maintained a consistent improvement until approximately epoch 18.

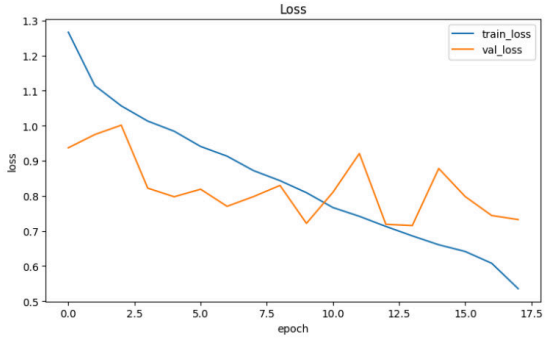


Fig. 4: Training and validation loss curves for DenseNet169 at 512 px.

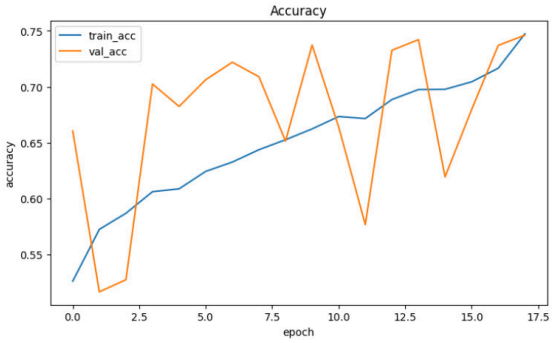


Fig. 5: Training and validation accuracy curves for DenseNet169 at 512 px.

ConvNeXt-Small, whose learning curves are presented in Figs. 6 and 7, achieved its best validation result at epoch 3. This early peak suggests that the combination of a fully unfrozen backbone, high-resolution input images, and the adopted learning-rate configuration led to rapid adaptation during the initial epochs. However, subsequent training did not improve the monitored validation metric, indicating limited additional benefit from further optimization under this configuration.

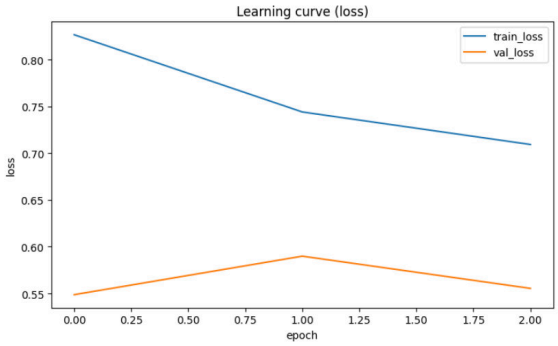


Fig. 6: Training and validation loss curves for ConvNeXt-Small at 1024 px.

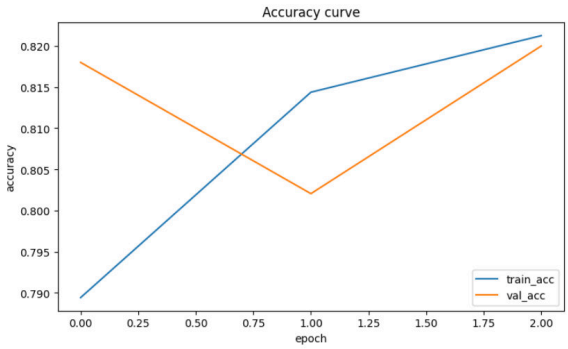


Fig. 7: Training and validation accuracy curves for ConvNeXt-Small at 1024 px.

D. *Confusion Matrices*

The confusion matrices of the three best-performing models provide a more detailed view of how each architecture was affected by class imbalance and highlight their individual strengths and limitations.

The confusion matrix of EfficientNetV2-S at 512 px (Fig. 8) confirms that the *Mild* class was the most challenging for the model. A considerable proportion of these samples was classified as either *No DR* or *Moderate*. This pattern is consistent with the subtle nature of early diabetic retinopathy findings, which may share visual characteristics with both healthy retinas and more evident moderate lesions.

The *Severe* class was also frequently confused with *Moderate* and *Proliferative*. In contrast, the strongest diagonal values were observed for the *No DR* and *Moderate* classes. Compared with the 384 px configuration, the 512 px model produced more consistent predictions for the *Severe* and *Proliferative* classes.

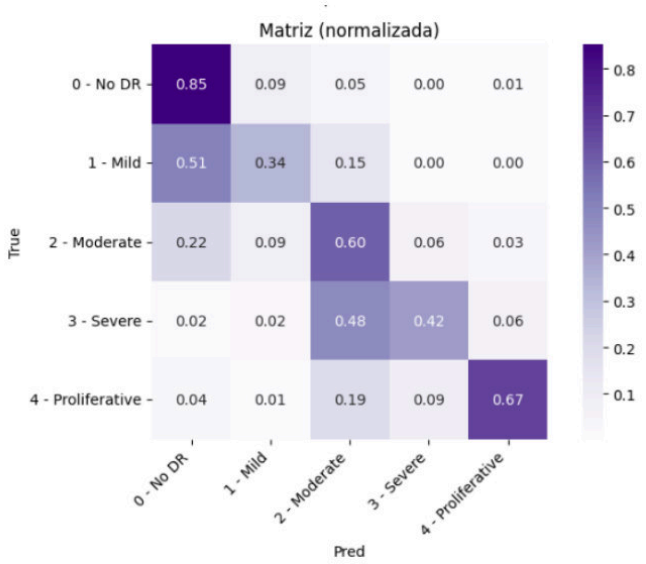


Fig. 8: Confusion matrix of EfficientNetV2-S at 512 px on the validation subset.

ConvNeXt-Small at 1024 px (Fig. 9) showed a pronounced tendency toward the majority *No DR* class, for which it achieved a recall of 0.93. In contrast, its recall for the *Mild* class was only 0.05, indicating that most mild cases were incorrectly classified, particularly as healthy retinas.

Therefore, despite obtaining the highest overall Kappa score, the model showed limited ability to identify the earliest disease severity level. This result demonstrates that a high aggregate metric does not necessarily imply balanced performance across all clinically relevant classes.

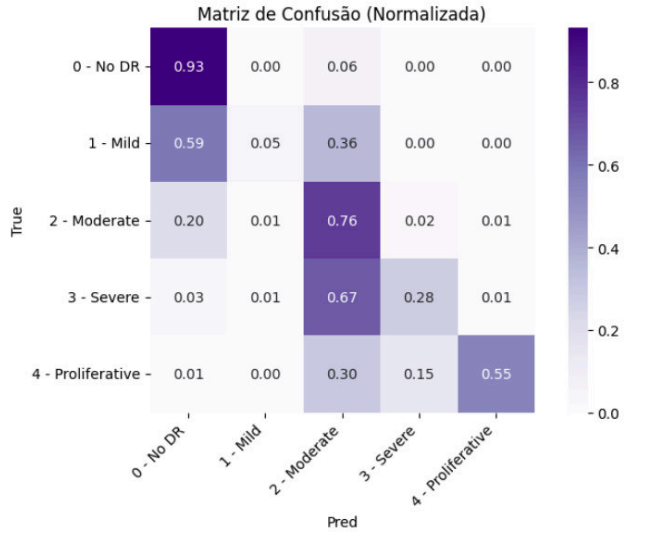


Fig. 9: Confusion matrix of ConvNeXt-Small at 1024 px on the validation subset.

The confusion matrix of DenseNet169 at 512 px (Fig. 10) showed a more distributed prediction pattern among the higher-severity classes. The model achieved a recall of 0.67 for the *Proliferative* class and a recall of 0.53 for the *Severe* class.

However, the low precision of 0.17 for the *Mild* class indicates a substantial number of false-positive predictions. This result suggests that the model frequently confused healthy retinas or samples from adjacent severity levels with mild diabetic retinopathy.

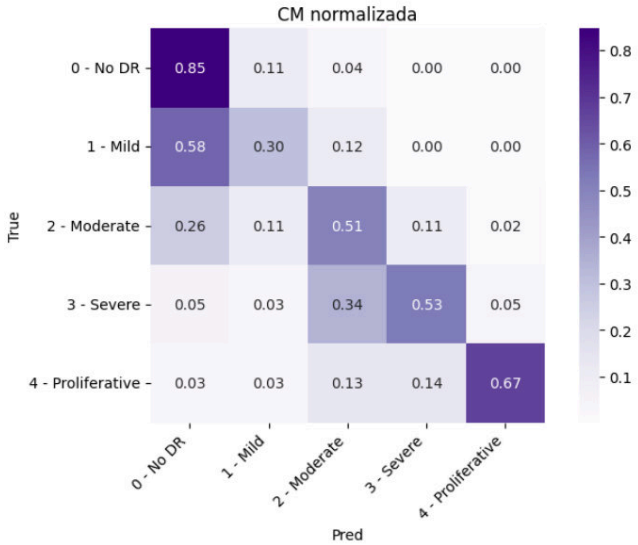


Fig. 10: Confusion matrix of DenseNet169 at 512 px on the validation subset.

E. Ensemble Results

The ensemble based on *Weighted Soft Voting* performed below expectations. Six models were combined using weights proportional to their individual Kappa scores. The resulting ensemble achieved a Kappa score of 0.4691 and a macro F1-score of 0.39, underperforming each of the three best individual models, as shown in Table V.

TABLE V: Comparison between the best individual models and the ensemble

Model	Macro F1	Kappa
ConvNeXt-Small (1024 px)	0.53	0.75
EfficientNetV2-S (512 px)	0.54	0.73
DenseNet169 (512 px)	0.55	0.71
Ensemble	**0.39**	**0.47**

Figure 11 presents the ensemble confusion matrix and shows how the combined model distributed its prediction errors across the five severity classes.

One possible explanation for this result is the lack of calibration among the probability distributions generated by the individual models. When model outputs are not calibrated on comparable scales, a weighted probability average may produce suboptimal predictions. Architectures with overconfident outputs may dominate the combined decision even when their class-specific predictions are less reliable than those of more conservative models.

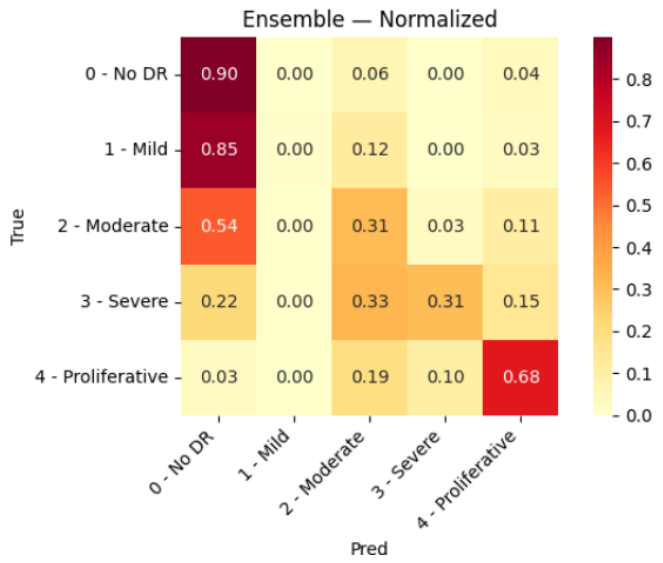


Fig. 11: Confusion matrix of the weighted soft-voting ensemble on the validation subset.

Furthermore, although the ensemble weights were derived from individual validation performance, they were assigned using a fixed empirical rule rather than optimized directly for the combined predictions. More systematic approaches, such as probability calibration, weight optimization, or a stacking meta-classifier, might produce a more effective combination.

Under the evaluated configuration, however, the ensemble did not provide a performance improvement sufficient to compensate for its additional computational cost. Simultaneous inference using multiple deep architectures increases memory requirements and inference latency, which may reduce the practicality of this approach in large-scale screening settings.

F. Discussion of the Main Findings

One of the most relevant findings was the influence of input image resolution. Increasing the EfficientNetV2-S input resolution from 384 px to 512 px raised the Kappa score from 0.57 to 0.73 and increased the recall for the *Mild* class from 0.26 to 0.34. This corresponds to a relative increase of approximately 31% in recall for this class.

A plausible explanation is that microaneurysms, small hemorrhages, and exudates associated with diabetic retinopathy may occupy relatively small image regions. Excessive downsampling can reduce the visibility of these findings and limit the model's ability to learn discriminative features, particularly for the earlier severity levels.

Nevertheless, the results should not be interpreted as evidence that increasing image resolution will systematically improve every architecture. Higher resolutions also increase computational cost and may require adjustments to the batch size, learning rate, regularization, and fine-tuning strategy. Therefore, image resolution must be considered jointly with the characteristics and original training configuration of each architecture.

The evaluated architectures were not compared under identical conditions. As summarized in Table II, input resolution, the proportion of unfrozen layers, the optimizer, the learning rates, and the number of epochs varied across configurations; For example, ConvNeXt-Small was evaluated at 1024×1024 px while the other architectures used lower resolutions (Table III). Isolating the individual contribution of architecture and resolution would require additional same-resolution experiments under otherwise identical settings, which we leave for future work.

Class imbalance remained a major limitation throughout the experiments. Class weighting in the loss function improved the model's sensitivity to minority classes but was not sufficient to compensate for the large difference in the number of samples. The *No DR* class contained approximately 36 times more images than the *Proliferative* class.

The *Mild* class was particularly challenging because it combined two difficulties: limited representation in the dataset and subtle visual differences from neighboring classes. ConvNeXt-Small illustrates the risk of evaluating an imbalanced multiclass model only through aggregate metrics. Although it achieved a Kappa score of 0.7453, its F1-score for the *Mild* class was only 0.09.

Therefore, selecting a model solely on the basis of its overall Kappa score could conceal substantial weaknesses in clinically relevant minority classes. Among the evaluated configurations, EfficientNetV2-S at 512 px provided the most balanced distribution of performance across the five severity levels. However, these results were obtained on the validation subset and should be confirmed using an independent test set before any conclusions regarding generalization or clinical applicability can be established.

IV. Conclusion

This study evaluated different configurations of deep learning architectures, including modern CNNs and a Vision Transformer, for the classification of five diabetic retinopathy severity levels using 35,126 retinal fundus images.

Among the evaluated architectures, EfficientNetV2-S with an input resolution of 512×512 pixels provided the most balanced class-wise performance, achieving a quadratic weighted Kappa score of 0.7313 and a macro F1-score of 0.54. ConvNeXt-Small achieved the highest overall Kappa score (0.7453); however, its recall of only 0.05 for the *Mild* class substantially limits its suitability for early-stage screening. The ensemble approach did not outperform the best individual architectures, indicating that combining model predictions requires appropriate probability calibration and weight optimization to be effective.

Although quadratic weighted Kappa was prioritized over overall accuracy to reduce the influence of the majority class, the substantial class imbalance still concealed important weaknesses in the minority classes. The use of class weights in the loss function improved the learning process but was insufficient to address the structural imbalance of the dataset. More robust imbalance-mitigation strategies should therefore

be investigated to improve the recognition of underrepresented severity levels, particularly the early stages of the disease.

Another relevant technical consideration concerns the effect of input image resolution. The experiments indicated that resolution can substantially affect classification performance, especially when the relevant pathological findings occupy small image regions. However, comparisons involving different input dimensions must be interpreted carefully because resolution changes also affect memory consumption, batch size, optimization dynamics, and the effective scale of the features processed by the network. Although several convolutional architectures can accept variable input dimensions through operations such as adaptive pooling, resolutions that differ substantially from those used during pretraining may alter the behavior of the learned representations. Therefore, comparisons across resolutions should be conducted under controlled experimental conditions and should consider the architectural design and documented preprocessing recommendations of each model.

The main methodological limitation of this study was the absence of an independent test set. Although the fixed random seed of 42 ensured consistent data partitions and prevented overlap between the training and validation subsets, the validation subset was used both to guide early stopping and model selection and to report the final ensemble performance. This repeated use of the same partition may introduce optimistic evaluation bias, since the final experimental decisions may be indirectly adapted to the validation data. Consequently, the reported results may overestimate the ability of the models to generalize to completely unseen data.

Future work should prioritize a rigorous three-way data partition comprising training, validation, and independent test subsets. Patient-level partitioning should also be adopted whenever information about the acquisition subject is available, thereby preventing images from the same patient from being distributed across different subsets. Additional investigations should evaluate model-specific input resolutions under controlled conditions and explore more advanced strategies for extreme class imbalance, including minority-focused data augmentation, *Focal Loss*, balanced sampling, and decision threshold adjustment. Probability calibration and systematically optimized ensemble strategies should also be investigated to determine whether model combination can provide improvements over the strongest individual classifiers.

REFERENCES

[1] I. Pearce, R. Simó, M. Lövestam-Adrian, D. T. Wong, and M. Evans, "Association between diabetic eye disease and other complications of diabetes: Implications for care. A systematic review," Diabetes, Obesity and Metabolism, vol. 21, no. 3, pp. 467–478, Mar. 2019, doi: 10.1111/dom.13550.

[2] A. Grzybowski, "Artificial intelligence for diabetic retinopathy screening," *Acta Ophthalmol.*, vol. 100, 2022, doi: 10.1111/j.1755-3768.2022.15371.

[3] V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.

[4] A. Saygılı and Ö. M. Atıcı, "Enhancing Retinal Disease Detection With the Swin Transformer: A Comprehensive Comparative Analysis," *Concurrency Computat. Pract. Exper.*, vol. 37, no. 27–28, p. e70366, 2025, doi: 10.1002/cpe.70366.

[5] S. Rajeshwar, S. Thaplyal, A. M., and G. Siva Shanmugam, "Diabetic Retinopathy Detection Using DL-Based Feature Extraction and a Hybrid Attention-Based Stacking Ensemble," *Adv. Public Health*, vol. 2025, p. 8863096, 2025, doi: 10.1155/adph/8863096.

[6] C. P. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," Ophthalmology, vol. 110, no. 9, pp. 1677–1682, Sep. 2003, doi: 10.1016/S0161-6420(03)00475-5.

[7] ilovescience, "Diabetic Retinopathy Resized," Kaggle, 2019. Disponível em: https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized. Acesso em: 4 jun. 2026.

[8] Karthik, Maggie, e S. Dane, "APTOS 2019 Blindness Detection," Kaggle, 2019. Disponível em: https://kaggle.com/competitions/aptos2019-blindness-detection. Acesso em: 4 jun. 2026.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

[11] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 10096–10106, doi: 10.1109/CVPR46437.2021.00986.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11976–11986, doi: 10.1109/CVPR52688.2022.01167.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[17] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," Psychological Bulletin, vol. 70, no. 4, pp. 213–220, Oct. 1968, doi: 10.1037/h0026256.

Comparative Analysis of Digital Image Processing Techniques for Automated Segmentation of Wound Healing Assays in Infected Cultures

Ana Laura Guastali Sant'Ana
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0009-0009-5668-1603

Lucas Militello da Silva
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0009-0000-4461-8716

Patrick Alves Carneiro
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

Guilherme Vieira de Faria
Institute of Biomedical Science (ICBIM)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0009-0004-0933-5189

Luana Carvalho Luz
Institute of Biomedical Science (ICBIM)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0001-6890-1161

Bellisa Freitas Barbosa
Institute of Biomedical Science (ICBIM)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0003-4611-2844

Eloisa Amália Vieira Ferro
Institute of Biomedical Science (ICBIM)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0002-7447-2583

Ana Claudia Patrocinio
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Pedro Cunha Carneiro
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

***Abstract*—The wound healing assay is widely used to evaluate cell migration. However, its quantitative analysis is commonly affected by manual segmentation variability, especially in images with nonuniform illumination, cellular debris, and poorly defined wound borders. This study presents a comparative analysis of three digital image processing pipelines for automatic segmentation of the open wound area in wound healing assays. A total of 225 images with 2048 × 1534 pixels, acquired at 0 h and 24 h, were evaluated. The first pipeline was based on median filtering, local contrast enhancement, texture estimation, and Otsu thresholding. The second pipeline incorporated homomorphic filtering, morphological enhancement, Gabor-based texture extraction, and k-means clustering. The third pipeline introduced an adaptive approach based on image brightness and experimental time, with specific processing strategies for 0 h and 24 h images. The segmented areas were compared with manual annotations performed by specialists, and the signed error, absolute error, percentage error, and processing time were analyzed. The results showed a progressive reduction of segmentation errors across the developed pipelines, with the adaptive method providing the best balance between segmentation accuracy and computational cost. These findings reinforce the importance of standardized image processing procedures to improve reproducibility and support quantitative analysis in wound healing studies.**

***Index Terms*—automatic quantification, digital image processing, image segmentation, morphological processing, wound healing assay.**

I. Introduction

Trypanosoma cruzi and *Toxoplasma gondii* are two intracellular protozoa capable of causing infections of major public health relevance, although they have distinct characteristics, particularly regarding their modes of transmission and immune responses [1], [2]. *Trypanosoma cruzi*, the etiological agent of Chagas disease, is mainly transmitted by triatomine insects, but it can also be transmitted from mother to fetus, although at relatively low rates, ranging from 1% to 12%. In contrast, *Toxoplasma gondii*, the protozoan responsible for toxoplasmosis, has a worldwide distribution, infects a significant proportion of the population, and presents higher rates of congenital transmission, ranging from 22% to 72%. Despite these differences, both parasites are capable of interfering with the host immune response and modulating processes related to gene expression, which may influence the progression of infection [3].

Given this scenario, different studies have been conducted to understand the cellular damage associated with the action of these protozoa, as well as to investigate strategies capable of promoting tissue recovery [4], [5], [6]. Among the methods used, the in vitro assay known as the wound-healing scratch plate assay stands out, as it is employed to evaluate cell migration during the repair process. In this assay, cells such

as human fibroblasts are cultured in plates until they form a continuous cell monolayer, after which a "scratch" is made in this monolayer, simulating a wound. From this point, the ability of the cells to migrate into the injured area and promote the closure of the formed gap is observed. Thus, this model allows for the analysis of the effects of different substances or experimental conditions [7].

Cellular wound healing is monitored through image acquisition over time, allowing tracking of the rate at which the wound area closes. Thus, the faster the reduction of the open region, the greater the cell migration capacity tends to be in response to the treatment being analyzed. For this evaluation, images can be processed and quantified using specific software tools, which enable the measurement of parameters such as wound area, percentage of closure, mean width, and width standard deviation. Therefore, image processing techniques contribute to making the analysis of cellular wound healing more objective and reproducible [7], [8].

Some software tools are commonly used for image analysis in cellular wound healing assays, such as MATLAB and tools developed as plugins for ImageJ/Fiji. In this context, Suarez-Arnedo et al. developed their own ImageJ/Fiji plugin, called the Wound Healing Size Tool, with the objective of automating the analysis of images obtained from these assays. For this purpose, the images underwent processing steps as contrast enhancement, filtering, and binarization, allowing the automatic segmentation of the region corresponding to the wound. After this step, the plugin calculated important metrics, such as wound area, percentage of closure, and mean width. As a result, the authors observed that the tool showed performance similar to existing methods for area quantification. The plugin also reduced the analysis time, being faster than the manual method [9]. These software tools are developed with the objective of automating the process of wound area quantification, a step that is traditionally performed manually and is time-consuming, especially when the analysis involves large images datasets.

Therefore, the objective of this study is to perform a systematic comparative analysis of three digital image processing pipelines designed to automate the segmentation of open wound regions in vitro scratch assays under parasite infection. Specifically, this work introduces a novel, adaptive framework that adjusts filtering, enhancement, and region-of-interest constraints based on experimental time (0 h vs. 24 h) and image brightness levels. The ultimate goal is to establish a standardized, computationally efficient workflow that minimizes human subjectivity and mitigates the artifacts caused by cellular debris and non-uniform illumination, thereby enhancing quantitative accuracy and reproducibility in biomedical research.

II. MATERIALS AND METHODS

A. Image dataset

The study was performed with three wound healing assay image datasets from the Institute of Biomedical Sciences (ICBIM). Human fibroblasts were cultured to confluence in multi-well plates and infected with either T. cruzi or T. gondii. Mechanical scratch wounds were created across the monolayer using a sterile 200 µL pipette tip, followed by washing to remove unattached cells. Wound healing was monitored at 0 h and 24 h post-scratch using an EVOS XL Core inverted microscope (Thermo Fisher Scientific) under a 4× (or 10×) objective in brightfield mode, capturing uncompressed 2048×1534 pixel frames. The first dataset consisted of 47 images in TIF format, with 22 corresponding to the 0 h time point and 25 to the 24 h time point, contaminated with *Trypanosoma cruzi*. The second dataset consisted of 76 images in JPG format, 34 of which were from 0 h and 42 from 24 h. The third dataset contained 102 images in JPG format, equally distributed between the 0 h and 24 h time points, both contaminated with *Toxoplasma gondii*. The distribution of the images used in each dataset, considering the file format and experimental time points, is presented in Table I.

TABLE I
DESCRIPTION OF THE IMAGE DATASET USED IN THE STUDY

Dataset	Format	0 h	24 h	Analyzed
Dataset 1	TIF	22	25	47
Dataset 2	JPG	34	42	76
Dataset 3	JPG	51	51	102
Total	–	**107**	**118**	**225**

B. Overview of processing

The computational approach was based on the comparison of three automatic segmentation pipelines P1, P2, and P3. The first two versions were developed in MATLAB® (The MathWorks, Inc., Natick, MA, USA; institutional license no. 41224224), while the current version was developed in Python. All these tests were performed on the same computer, equipped with an Intel(R) Core(TM) i5-1035G1 @ 1.00 GHz processor and 16 GB of DDR4 RAM, in order to ensure that processing time comparisons were conducted under equivalent computational conditions.

To support reproducibility and facilitate independent evaluation of the proposed methods, the complete source code of the three processing pipelines is publicly available at https://github.com/anahaula/wound-healing-segmentation.git. The repository includes the implementation of the preprocessing, segmentation, morphological refinement, and quantitative evaluation stages, together with the parameter configurations adopted in each approach. The experimental microscopy images used for development and validation are not publicly distributed and are therefore not included in the repository.

In all approaches, images were first converted to grayscale intensity matrices, as wound delineation relies primarily on intensity gradients, local contrast, and textural variance rather than chrominance [10]. To overcome acquisition artifacts, CLAHE was employed for local contrast enhancement while limiting noise amplification [11]. Spatial smoothing, including Gaussian, median, and edge-preserving bilateral filtering, was

applied to attenuate local intensity fluctuations prior to segmentation [11], [13]. Following binarization, morphological operators (opening, closing, and hole-filling) were used to eliminate spurious detections, reconnect fragmented regions, and smooth boundary contours [12]. Finally, the open wound area was quantified by counting the remaining foreground pixels in the binary mask, as formulated in 3.

Finally, the wound area was calculated by counting the pixels belonging to the final binary mask, as shown in (2). Thus, after processing, each image was represented by a quantitative area measurement in pixels, allowing comparison with the manual delineation performed by specialists.

C. Pipeline P1: Initial approach

The first pipeline was developed as a strategy to assess the feasibility of automatic segmentation of wound healing images. Initially, the images were converted to grayscale and filtered using a median filter with an 11x11 window. This filtering step aimed to reduce local noise and small isolated intensity variations before the enhancement and segmentation stages.

Next, CLAHE was applied to enhance local contrast. This step was important because some images presented poorly defined borders between the open wound region and the area occupied by cells. After contrast enhancement, local texture was estimated through the standard deviation calculated within image neighborhoods. The local standard deviation measures the variation in intensity levels around each pixel. Regions with a higher cellular presence tend to present a more irregular texture, whereas the open wound region usually exhibits a more homogeneous pattern. Thus, the texture map was used to assist in differentiating the wound area from the cellular background. Segmentation was performed using Otsu thresholding, generating an initial binary mask.

The Otsu method automatically calculates a threshold based on the distribution of grayscale levels, and it was used in the first approach because it does not depend on a manually defined fixed value [14].

After binarization, the image underwent a post-processing step, that was used to remove small isolated regions. Morphological closing, composed of dilation followed by erosion, was applied to connect nearby regions and correct small discontinuities. Hole filling was used to correct internal gaps in the mask, and the removal of small components eliminated detected regions that did not correspond to the wound. This pipeline made it possible to obtain the first consistent automatic segmentations.

D. Pipeline P2: lighting correction and segmentation by textural features

The second pipeline was developed to reduce the influence of non-uniform illumination and improve the separation between the wound region and the cellular background. Initially, the image was normalized, standardizing the intensity range to reduce differences between brighter and darker images.

After normalization, homomorphic filtering was applied. This technique assumes that the image can be modeled as the product of an illumination component and a reflectance component. Illumination is associated with slow brightness variations, whereas reflectance is related to structural details, edges, and texture [10]. In the code, the image was transformed into the logarithmic domain, smoothed using a Gaussian filter to estimate the illumination component, and then this component was subtracted from the image [11]. In this work, the filter was defined as shown in (1):

$$H(u,v) = (\gamma_H - \gamma_L)\left[1 - e^{-c\frac{D^2(u,v)}{D_0^2}}\right] + \gamma_L \qquad (1)$$

Where γ_H controls the gain of the high frequencies, γ_L controls the gain of the low frequencies, c controls the slope of the filter transition, D_0 is the cutoff frequency, and $D(u,v)$ represents the distance from the point (u,v) to the center of the frequency spectrum. Thus, this step aimed to reduce the effects of shadows and irregular brightness while preserving information associated with the wound edges.

In addition to intensity information, textural features were extracted using Gabor filters. These filters respond to specific frequency and orientation patterns and were applied at multiple wavelengths and directions to highlight textural structures and edges with different orientations [10]. In the code, the responses of the Gabor filters were combined with the processed image, forming a feature set used in the clustering step.

Segmentation was performed using k-means, an unsupervised clustering method that separates data into classes with similar characteristics [10]. In this work, each pixel was represented by the features extracted from the image, and k-means grouped the pixels with similar intensity and texture patterns. The class most compatible with the wound region was then selected as the binary mask.

Finally, the mask is further refined by filling holes, removing small objects, selection of the largest connected component, and edge smoothing. The selection of the largest connected component retained only the largest continuous region of the mask, reducing false detections scattered throughout the image.

E. Pipeline P3: current adaptive approach

The third pipeline corresponds to the current version of the algorithm implemented in Python. Unlike previous approaches, this method eschews a single fixed set of parameters for all images. Instead, the processing was adapted according to the experimental time point and the mean intensity of the grayscale image.

The images were classified as bright, intermediate, or dark. We chose this division as images with different illumination levels have different responses to enhancement, filtering, and thresholding steps. For each group, specific parameters were defined for gamma correction, homomorphic filtering, CLAHE, smoothing, and morphological operations.

Gamma correction is a transformation on the intensity levels of an image. This technique allows the brightness distribution

to be adjusted before segmentation, making image regions brighter or darker according to the value of the gamma parameter [10]. In the code, this step was used to make bright or dark images more suitable for the subsequent enhancement and binarization steps.

Homomorphic filtering was maintained to reduce slow illumination variations (1), with only the parameters being adjusted. CLAHE was used to improve local contrast, while Gaussian, bilateral, and median filters were applied for smoothing and noise reduction. The combination of these filters made it possible to reduce small image irregularities without completely eliminating the main wound edges.

For the 24 h images, we included an additional step of cellular texture suppression based on the local standard deviation. This step was necessary because, after 24 h of the assay, cell migration may partly fill the open wound region, making the internal texture more similar to the cellular background. Therefore, the local standard deviation was used to identify regions with greater intensity variation and reduce the influence of cellular patterns that could be incorrectly interpreted as edges or as part of the wound area.

After adaptive processing, the mask was refined using multi-scale morphological operations. Morphological closing was used to connect nearby regions and correct continuity gaps; opening removed small isolated detections; hole filling corrected internal gaps; and the choice of the largest connected component preserved the main region associated with the wound. Edge smoothing was applied at the end to reduce abrupt irregularities in the segmented contour

For the 24 h images, segmentation was also spatially constrained based on the region detected at the 0 h time point, when this information was available. This constraint was adopted to avoid estimates that were incompatible with the original wound area, such as segmentations much larger than the original open region. Fig. 1 shows the workflow of the developed code.

F. Calculation of the wound area and closure

After segmentation, the wound area was calculated from the final binary mask generated by the algorithm. Considering $B_t(x, y)$ as the mask obtained at the experimental time point t, in which wound pixels are assigned a value of 1 and all other pixels are assigned a value of 0, the area in pixels was defined as (2):

$$A_t = \sum_{x=1}^{M} \sum_{y=1}^{N} B_t(x, y) \tag{2}$$

where A_t represents the segmented wound area at time point t.

For sequential images of the same sample, the wound closure percentage was calculated (3) to evaluate the reduction of the open area between 0 h and 24 h.

$$F_{0h \to 24h}(\%) = \left(1 - \frac{A_{24h}}{A_{0h}}\right) \times 100 \tag{3}$$

where A_{0h} corresponds to the initial wound area and A_{24h} to the remaining area after 24 h.

G. Quantitative Validation

The algorithm was validated by comparing the automatically segmented area with the area manually delineated by specialists. For each image, the signed error, absolute error, and absolute percentage error were calculated, as described in (4), (5), and (6), respectively.

$$E_s = A_{auto} - A_{manual} \tag{4}$$

$$E_a = |A_{auto} - A_{manual}| \tag{5}$$

$$E_p(\%) = \frac{|A_{auto} - A_{manual}|}{A_{manual}} \times 100 \tag{6}$$

where A_{auto} represents the area obtained by the algorithm and A_{manual} represents the reference area defined by the specialists.

To evaluate the overall performance of each pipeline, the mean absolute percentage error and the mean processing time per image were computed in (7) and (8), respectively.

$$\overline{E_p} = \frac{1}{n} \sum_{i=1}^{n} E_{p,i} \tag{7}$$

$$\overline{T} = \frac{1}{n} \sum_{i=1}^{n} T_i \tag{8}$$

where n represents the number of images evaluated and T_i corresponds to the processing time of image i.

In addition to area-based error metrics, spatial agreement between the automatic and manual segmentations was evaluated using the Dice coefficient:

$$Dice = \frac{2|M_{auto} \cap M_{manual}|}{|M_{auto}| + |M_{manual}|} \tag{9}$$

where M_{auto} and M_{manual} represent the automatic and manual wound masks, respectively. Dice values range from 0 to 1, with higher values indicating greater overlap. The coefficient was calculated for each image and summarized by pipeline and time point using mean, median, and standard deviation.

III. Results

A. Visual Segmentation Assessment

Fig. 2 presents a representative comparison of the segmentation boundaries generated by the three pipelines. P1 produced fragmented contours and was strongly affected by local artifacts, especially at 24 h. P2 improved the overall wound tracking but showed oversegmentation in some cellular regions. In contrast, P3 provided more coherent boundaries and better agreement with the manual reference, including in the more challenging 24 h images.

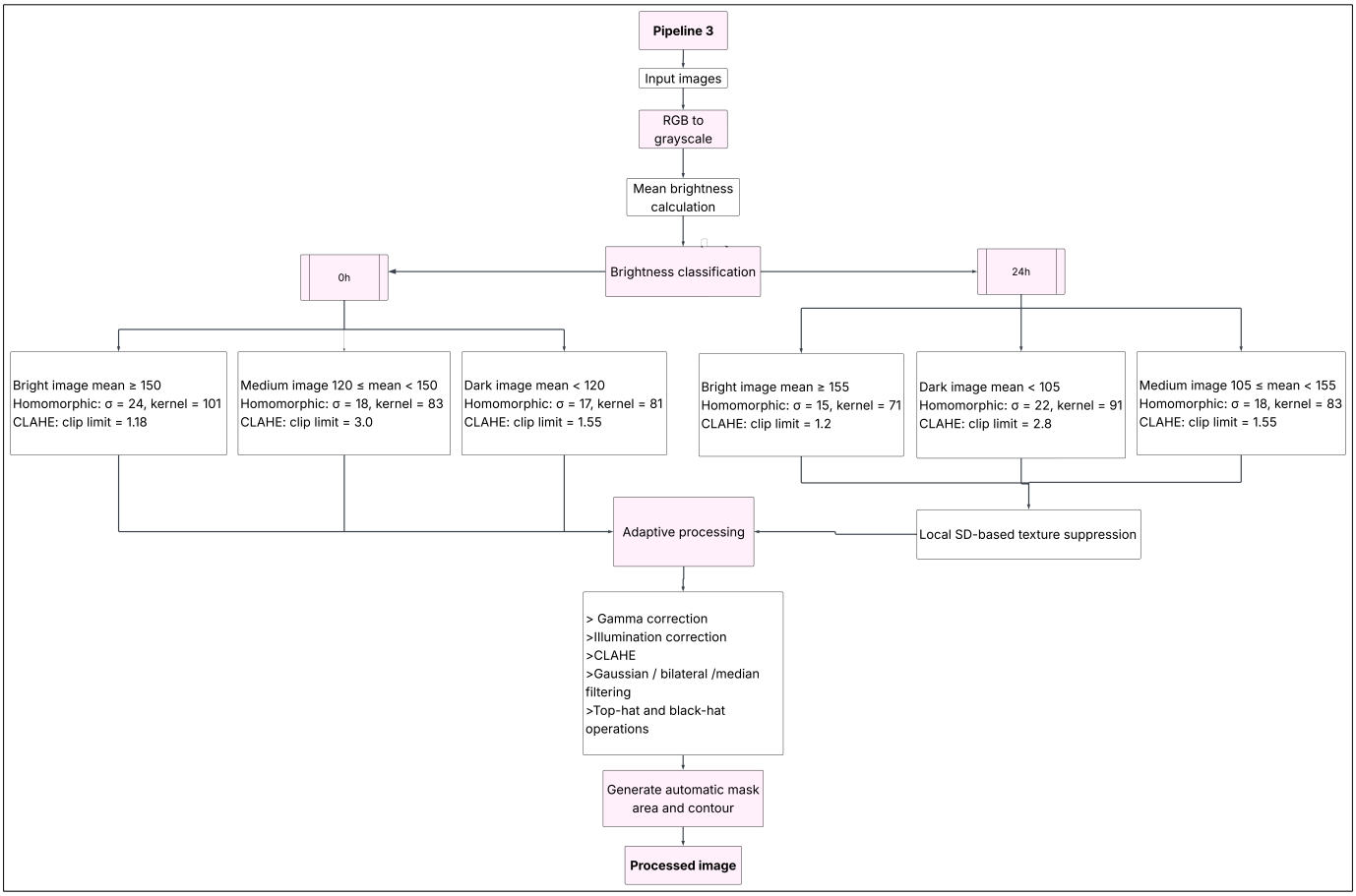


Fig. 1. Architectural flowchart of the proposed adaptive framework (Pipeline 3). The schematic maps the dynamic execution routing where input matrices are automatically stratified by experimental time points (0 h vs. 24 h) and specific mean brightness thresholds (bright, intermediate, or dark) to select optimal parameters for gamma correction, homomorphic filtering, and local contrast enhancement (CLAHE).

B. Error Distribution Analysis

The distribution of absolute percentage errors across designated ranges highlights a progressive optimization through the algorithmic iterations (Fig 3,Fig 4, and Fig 5).

For Pipeline 1 (Otsu-based), a critical vulnerability to undersegmentation was confirmed, with a massive concentration of images falling into the extreme error range (100%), encompassing 89 images at 0 h and 36 images at 24 h (Fig. 3).

Pipeline 2 (Gabor/k-means) successfully mitigated undersegmentation, shifting the majority of 0 h images to lower error tiers (29 images in 0–20% and 19 in 20–40%). However, it introduced widespread positive errors at 24 h, with 21 and 22 images scattered in the 60–80% and 80–100% ranges, respectively, due to a lack of illumination tolerance (Fig. 4).

Pipeline 3 (Adaptive) achieved the most robust error control. At the critical baseline condition (0 h), it concentrated 57 images in the lowest negative error tier (0–20%) and 25 images in the lowest positive tier (Fig. 5). At 24 h, despite the biological noise of cell proliferation, severe failures (100%) were restricted to a minimal subset (9 images), maximizing the number of samples within acceptable clinical tolerances (Fig. 5).

Table II presents the descriptive statistics of the absolute percentage discrepancy and the Dice coefficient obtained for the three processing pipelines at 0 h and 24 h. The table includes the mean and median values for each metric, together with the corresponding standard deviation. In this context, the standard deviation indicates the dispersion of the measurements within each group, allowing the variability of the segmentation results to be observed in addition to the central tendency represented by the mean and median. The Dice coefficient complements the area-based discrepancy by providing a spatial similarity measure between the automatic segmentation and the manual reference.

TABLE II
QUANTITATIVE EVALUATION OF THE PROCESSING PIPELINES AT 0H AND 24H

Metrics	P1	P2	P3
Mean error 0h (%) ± SD	82 ± 0.23	27 ± 0.31	17 ± 0.26
Mean error 24h (%) ± SD	173 ± 3.11	242 ± 3.23	100 ± 0.26
Median error 0h (%) ± SD	91 ± 0.23	14 ± 0.31	6 ± 1.74
Median error 24h (%) ± SD	94 ± 3.11	97 ± 3.23	35 ± 1.74
Median Dice 0h ± SD	0,03 ± 0,03	0,73 ± 0,34	0,84 ± 0,27
Median Dice 24h ± SD	0,05 ± 0,04	0,45 ± 0,32	0,64 ± 0,22

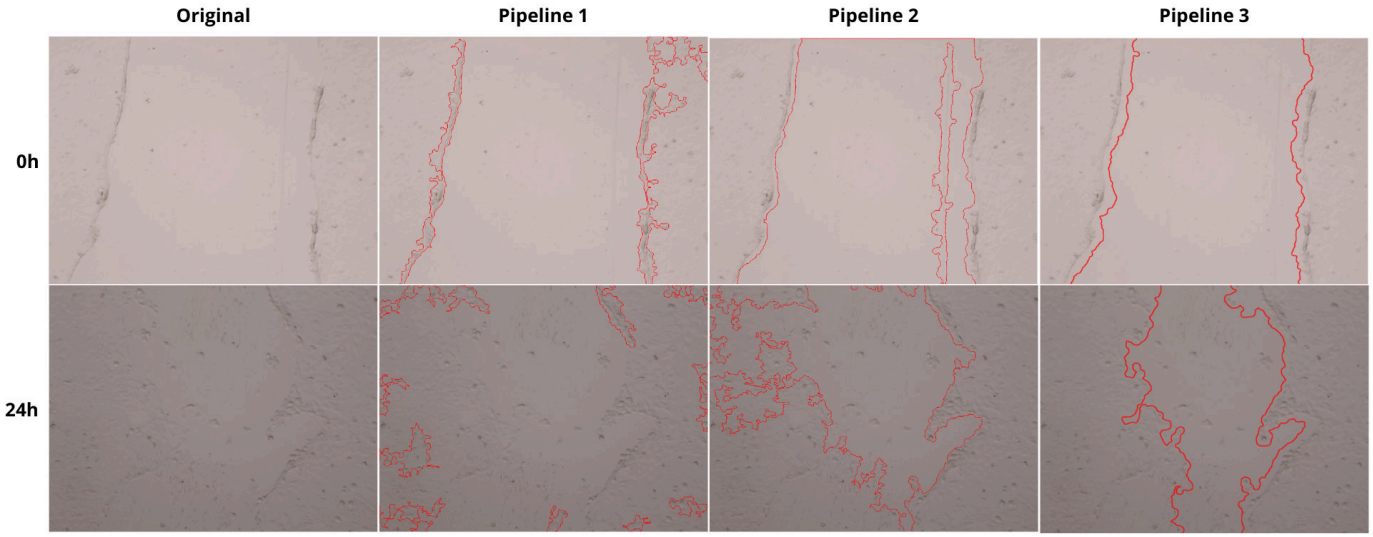


Fig. 2. Representative visual comparison of the automated segmentation boundaries (red contours) generated across the three development stages against the specialist's ground truth. The top row illustrates baseline tracking (0 h), contrasting Pipeline 1's sensitivity to structural noise with the progressive optimization of Pipelines 2 and 3. The bottom row (24 h) highlights the performance under active cellular migration, showcasing Pipeline 3's superior spatial coherence and topological stability when biological debris and low boundary contrast mislead conventional thresholding strategies.

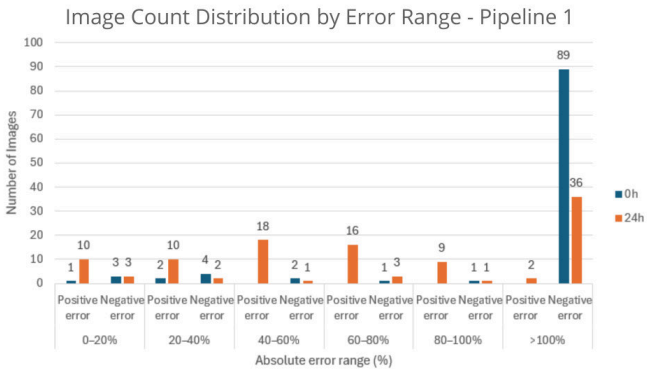


Fig. 3. Distribution of the image dataset analyzed across absolute percentage error ranges for Pipeline 1 (Otsu based) at 0 h and 24 h. The bars categorize the cumulative counts into positive and negative deviation tiers, highlighting the method's accuracy and error control compared to the specialists' manual annotations.

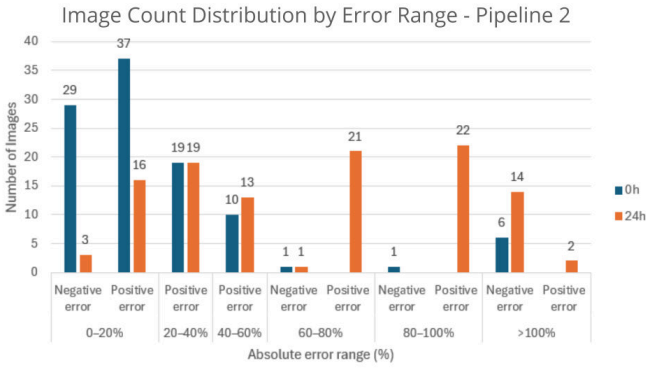


Fig. 4. Distribution of the image dataset analyzed across absolute percentage error ranges for Pipeline 2 (Gabor/k-means) at 0 h and 24 h. The bars categorize the cumulative counts into positive and negative deviation tiers, highlighting the method's accuracy and error control compared to the specialists' manual annotations.

C. Computational Efficiency

The computational throughput per image revealed an explicit trade-off between algorithmic complexity and processing time (Fig. 6). Pipeline 1 was the fastest, with a negligible median execution time below 1 seconds, but at the expense of segmentation viability. Pipeline 2 recorded the worst computational overhead and the highest dispersion (median 28 s), driven by multi-orientation Gabor filter convolutions and iterative k-means clustering. Pipeline 3 successfully restricted the computational cost (median 17 s), representing a significant speedup compared to Pipeline 2 while achieving superior accuracy through its dynamic parameters.

TABLE III
PROCESSING TIME STATISTICS FOR THE THREE EVALUATED PIPELINES

Metrics	P1	P2	P3
Mean time (s) $\pm$ SD	1 $\pm$ 0.48	28 $\pm$ 7.94	17 $\pm$ 3.22

IV. Discussion

The image dataset was processed using three pipelines developed progressively to address the main difficulties observed during segmentation. The images were acquired in wells containing liquid medium, without complete standardization of illumination, and also contained artifacts related to the experimental procedure and to the blade used during scratch generation. These limitations became even more evident at 24 h, when cells partially occupied the initially open region,

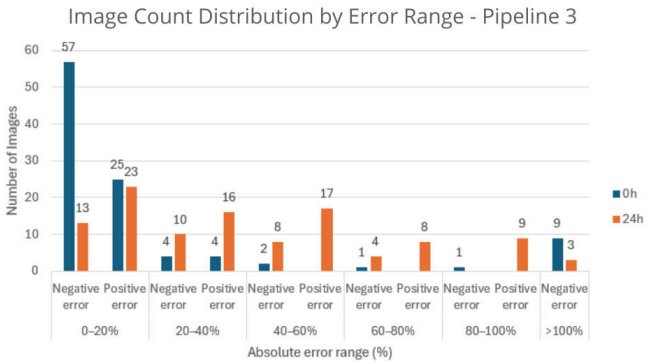


Fig. 5. Distribution of the image dataset analyzed across absolute percentage error ranges for Pipeline 3 (Adaptive Approach) at 0 h and 24 h. The bars categorize the cumulative counts into positive and negative deviation tiers, highlighting the method's accuracy and error control compared to the specialists' manual annotations

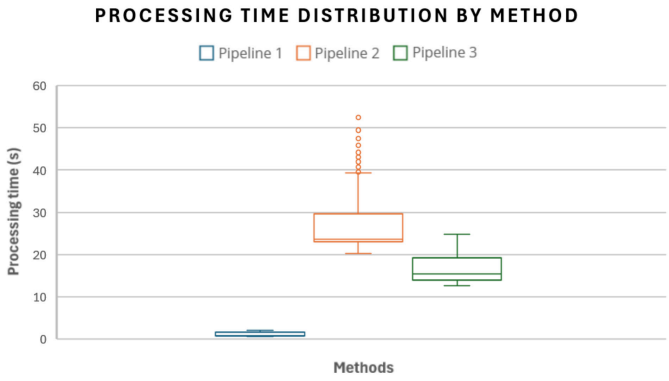


Fig. 6. Boxplot distribution of the computational processing time (in seconds) per image across the three evaluated segmentation pipelines. The whiskers and outliers illustrate the operational throughput and complexity trade-offs, highlighting the intermediate efficiency achieved by the dynamic Python framework (Pipeline 3).

reducing the contrast between the wound and the surrounding cellular area. Under these conditions, distinguishing cells, artifacts, and the remaining open region became more challenging for both automatic segmentation and manual reference delineation. Therefore, each pipeline was developed to address limitations identified in the previous one and progressively reduce segmentation errors.

P1 showed the expected behavior of an initial approach. Its main advantage was its low computational demand, with an average processing time of approximately 1 s per image. However, it presented the highest errors among the three pipelines, as shown in Table II. This behavior is mainly related to the dependence of Otsu thresholding on a global separation between foreground and background, a condition that is not always present in wound healing images. Since the open region does not exhibit uniform intensity and its boundaries may be smooth or discontinuous, local texture variations can be incorrectly interpreted as independent regions. This limitation is also reflected in the low Dice coefficient values obtained by P1.

P2 showed a clear improvement over P1 by incorporating homomorphic filtering to reduce illumination variations and Gabor-based features combined with k-means clustering. In Fig. 4, a reduction in extreme deviations can be observed at 0 h, with a greater concentration of images within the 0–20% and 20–40% error ranges. However, P2 still showed important limitations at 24 h. In these images, regions containing dispersed cells, debris, or textures similar to the open area were frequently incorporated into the segmentation. Thus, although texture analysis reduced some of the failures observed in P1, it was not sufficient to consistently distinguish the wound region in images with poorly defined boundaries and irregular cellular coverage.

P3 was developed to directly address these limitations. In Fig. 5, a greater concentration of 0 h images can be observed within the lowest discrepancy ranges, indicating greater stability in identifying the initial wound region. This result is particularly relevant because the area measured at 0 h serves as the reference for the subsequent calculation of wound closure. For the 24 h images, the corresponding 0 h mask was used to restrict the region analyzed, reducing the possibility of identifying areas outside the original scratch. Even with this strategy, the presence of cells within the open region and the reduced definition of its boundaries remained important sources of error. Therefore, the improvement observed with P3 does not eliminate the difficulties associated with 24 h images, but it reduces the occurrence of severe failures compared with the previous pipelines.

The statistics presented in Table II reinforce this behavior. At 0 h, P3 achieved a mean error of 17%, compared with 82% for P1 and 27% for P2, as well as a median error of only 6%, compared with 91% and 14%, respectively. The median Dice coefficient was also higher for P3, reaching 0.84, whereas P2 and P1 presented values of 0.73 and 0.03, respectively. At 24 h, despite the greater complexity of the images, P3 achieved a median error of 35%, lower than the 94% obtained by P1 and 97% by P2, while its median Dice coefficient reached 0.64, compared with 0.45 and 0.05 for P2 and P1, respectively. However, the mean error of 100% observed for P3 at 24 h indicates that a subset of particularly challenging images still produced large discrepancies, substantially increasing the mean. In this context, the difference between mean and median highlights the heterogeneity of the 24 h condition and the importance of considering the overall distribution of the results when evaluating pipeline performance.

When stratifying the P3 performance by pathogen, median Dice scores at 0 h and 24 h were 0.82 and 0.61 for T. cruzi (n=47), compared to 0.85 and 0.65 for T. gondii (n=178), respectively. Although T. cruzi assays presented higher debris from parasite motility, the adaptive brightness and multi-scale morphological filtering maintained robust boundary delineation across both infection models, demonstrating the method's versatility regardless of parasite type.

Regarding computational efficiency, P1 remained the fastest pipeline due to its simpler structure, whereas P2 presented the highest computational cost, mainly because of the Gabor filters

and k-means clustering. P3 required an intermediate processing time while maintaining a lower computational cost than P2 and a more favorable segmentation performance than the other pipelines. Therefore, when area error, Dice coefficient, and processing time are considered together, P3 provided the most favorable balance among the methods evaluated for this image dataset. This observation, however, should not be interpreted as evidence of statistical superiority or as an indication that the method will necessarily generalize to other datasets.

Despite the improvements achieved, the 24 h results show that there is still substantial room for further development. Future work should include the expansion and diversification of the image dataset to encompass different cell lines, laboratories, microscopy systems, and acquisition conditions, as well as manual references generated by multiple annotators. Additional spatial metrics, such as IoU and boundary-based measures, should also be considered, together with validation of the adaptive rules of P3 using images not involved in parameter definition. Automatic parameter optimization and comparisons with deep-learning approaches, including U-Net, DeepLabV3+, and SAM, are also relevant next steps. These evaluations will provide a stronger basis for assessing the robustness and generalizability of the proposed approach.

Overall, the results demonstrate a consistent progression across the three pipelines. P1 established an initial, computationally inexpensive approach but showed a high occurrence of segmentation errors. P2 reduced part of these discrepancies through illumination correction and texture analysis, although important limitations remained, particularly at 24 h. P3 achieved the lowest errors and highest Dice values among the evaluated pipelines while requiring less processing time than P2. Thus, under the experimental conditions and image dataset evaluated in this study, P3 showed the most favorable performance for segmentation and quantification of the open region in wound healing assays, while the 24 h images remain the main challenge to be addressed in future improvements.

V. Conclusion

This study compared three progressively developed classical image-processing pipelines for the segmentation of *wound healing images. P1 was the fastest approach but presented the highest segmentation errors, while P2 improved the results through illumination correction and texture analysis at a higher computational cost.

P3 showed the most favorable overall performance among the evaluated pipelines, reaching a mean area error of 17% and a median Dice coefficient of 0.84 at 0 h, as well as a median area error of 35% and a Dice coefficient of 0.64 at 24 h. Although the 24 h images remained more challenging, P3 reduced severe segmentation failures and provided a better balance between accuracy and processing time within the evaluated dataset.

Future work should focus on improving segmentation at 24 h and validating the method on more diverse and independent datasets. The use of multiple annotators, automatic parameter optimization, and comparisons with deep-learning approaches may further assess the robustness and generalizability of the proposed pipeline.

Acknowledgment

The authors would like to thank the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Instituto de Ciências Biomédicas (ICBIM), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Grupo de Imagens Médicas (GIM) for the financial and technological support provided for the development of this study and work.

References

[1] G. Macaluso, F. Grippi, S. Di Bella, V. Blanda, F. Gucciardi, A. Torina, A. Guercio, and V. Cannella, "A review on the immunological response against Trypanosoma cruzi," Pathogens, vol. 12, no. 2, Art. no. 282, Feb. 2023.

[2] T. R. Brasil, C. G. Freire-de-Lima, A. Morrot, and A. C. Vetö Arnholdt, "Host-Toxoplasma gondii coadaptation leads to fine tuning of the immune response," Frontiers in Immunology, vol. 8, Art. no. 1080, Sep. 2017, doi: 10.3389/fimmu.2017.01080.

[3] Medina, L., Castillo, C., Liempi, A., Guerrero-Muñoz, J., Rojas-Pirela, M., Maya, J. D., ... & Kemmerling, U. (2020). Trypanosoma cruzi and Toxoplasma gondii induce a differential microRNA profile in human placental explants. Frontiers in Immunology, 11, 595250.

[4] S. Percival and K. Cutting, Eds., Microbiology of Wounds. London, England: CRC Press, 2019.

[5] A. Stamm, K. Reimers, S. Strauß, P. Vogt, T. Scheper, and I. Pepelanova, "In vitro wound healing assays—state of the art," BioNanoMaterials, vol. 17, no. 1–2, pp. 79–87, 2016.

[6] S. Kauanova, A. Urazbayev, and I. Vorobjev, "The frequent sampling of wound scratch assay reveals the 'opportunity' window for quantitative evaluation of cell motility-impeding drugs," Frontiers in Cell and Developmental Biology, vol. 9, Art. no. 640972, 2021.

[7] Ignacio Arias, J., Sepulveda, C., Bravo, P., Hamilton-West, C., Maldonado, I., & Ferreira, A. (2015). Comparative effect of human and Trypanosoma cruzi calreticulin in wound healing. Journal of tissue engineering and regenerative medicine, 9(1), 41-54.

[8] Gebäck, T., Schulz, M. M. P., Koumoutsakos, P., & Detmar, M. (2009). TScratch: A novel and simple software tool for automated analysis of monolayer wound healing assays: Short Technical Reports. Biotechniques, 46(4), 265-274.

[9] Suarez-Arnedo, A., Figueroa, F. T., Clavijo, C., Arbeláez, P., Cruz, J. C., & Muñoz-Camargo, C. (2020). An image J plugin for the high throughput image analysis of in vitro scratch wound healing assays. PloS one, 15(7), e0232565.

[10] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. Pearson, 2018.

[11] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in Graphics Gems IV, Academic Press, 1994, pp. 474–485.

[12] Y. Kimori, "Morphological image processing for quantitative shape analysis of biomedical structures," Journal of Synchrotron Radiation, 2013.

[13] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," in Proc. IEEE International Conference on Computer Vision, 1998.

[14] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

Comparative Analysis of the Performance of Analog and Digital Band-Stop Filters Applied to Power Line Noise Suppression in Electromyographic Signals

Bianca Vitória Silva Pina[1], Davi Gabriel dos Santos Oliveira[2],
Ana Carolina Quintão Siravenha Müller[3], Isabela Marques Miziara[4]

Institute of Technology, Federal University of Pará
Belém, Brazil
[1]bianca.pina@itec.ufpa.br, [2]davi.oliveira@itec.ufpa.br,
[3]siravenha@ufpa.br, [4]immiziara@ufpa.br

***Abstract*—This work presents a comparative and qualitative analysis of the performance of analog and digital band-stop filters applied to electromyographic (EMG) signal processing. The primary objective is to improve EMG signal quality by reducing power-line interference while preserving physiologically relevant characteristics. Five printed circuit boards were developed in Proteus software, each implementing a distinct notch filter topology: Fliege, Multiple Feedback (MFB), Sallen-Key, Tow-Thomas, and Twin-T. These were integrated into a custom acquisition system to evaluate hardware-based filtering effectiveness. In parallel, the same topologies were implemented digitally in MATLAB via bilinear transformation and applied to raw signals in which 60 Hz interference was deliberately retained. Evaluation was conducted through power spectral density analysis, windowed RMS, signal-to-noise ratio, and spectral preservation metrics. Results indicate that the analog Sallen-Key achieved the best 60 Hz rejection among analog topologies (20.37 dB), while the digital Twin-T yielded the highest attenuation (79.62 dB) and SNR (77.46 dB), although at the cost of the lowest in-band spectral preservation of the digital set (60.44 %). These findings support a hybrid architecture combining analog pre-filtering with digital refinement, with the digital topology selected according to whether maximum line rejection or maximum spectral fidelity is required. All *in-vivo* results were obtained from a single healthy volunteer in a single acquisition session, without independent replication, and should therefore be read as a proof-of-concept characterization rather than a statistically validated ranking.**

***Index Terms*—Power line artifacts, signal conditioning, notch filter, biomedical instrumentation, electromyographic signal.**

I. Introduction

The accurate acquisition and interpretation of electromyographic (EMG) signals are strictly dependent on the capability of the instrumentation system to isolate weak physiological biopotentials from surrounding environmental noise [1], [4]. Electromyography is highly susceptible to diverse types of degradation, including motion artifacts, baseline wander, and cross-talk from adjacent bioelectric sources [1]. Among these perturbations, power-line interference (PLI)-manifesting at 60 Hz or 50 Hz depending on the geographic region-constitutes one of the most pervasive and critical challenges in biomedical instrumentation.

In contrast to other electrophysiological signals (e.g., electrocardiography), attenuating PLI in EMG data presents unique technical difficulties. The dominant power spectrum of a surface EMG signal is distributed between 10 Hz and 150 Hz, often extending up to 500 Hz [2]. Because primary physiological components lie directly in the vicinity of the line frequency, conventional high-pass or low-pass filtering strategies are inadequate, as they inevitably clip meaningful neuromuscular data. Consequently, the selective suppression of this interference requires targeted band-stop (notch) filtering techniques alongside meticulous hardware design choices [3].

The fundamental design decision between implementing an analog notch filter within the early signal conditioning stage or applying a digital notch filter during post-processing involves critical trade-offs [4]. Hardware-based filters operate directly in the continuous-time domain, treating the signal before digitization. This approach prevents out-of-range saturation of Analog-to-Digital Converters (ADCs) and optimizes the dynamic range allocation [5]. However, physical analog components are inherently restricted by manufacturing tolerances, thermal drift, and parasitic impedances, which can displace the nominal center frequency ($f_0 = 60$ Hz) and alter the quality factor (Q) [6]. Conversely, discrete-time digital filters provide steep attenuation slopes, high numerical stability, and flexible coefficient tuning without hardware modification [7]. Nonetheless, digital alternatives increase the computational load and are strictly bound to the constraints of the prior sampling process, risking quantization noise amplification if the raw signal is heavily contaminated [5], [9].

Recent studies emphasize that unoptimized notch filter architectures can induce unwanted attenuation in adjacent frequency bands, distorting the morphological structure of Motor Unit Action Potentials (MUAPs) and skewing fundamental parameters such as median frequency and muscle firing rates

[8]. Therefore, a comprehensive comparative evaluation of distinct filter topologies is essential for the standardization of reliable biomedical acquisition protocols.

This paper presents a systematic comparative performance analysis of five classical notch filter topologies-Fliege, Multiple Feedback (MFB), Sallen-Key, Tow-Thomas, and Twin-T-implemented in both physical printed circuit boards (PCBs) and as digital IIR filters mapped via the bilinear transformation. The evaluation is conducted using a custom-developed biomedical acquisition system and validated through experimental surface EMG data collected from the *biceps brachii* muscle under controlled dynamic contractions. Performance is quantified across four rigorous criteria: absolute attenuation at 60 Hz, signal preservation index (P_{EMG}), temporal RMS attenuation impact, and resulting Signal-to-Noise Ratio (SNR).

II. Materials and Methods

A. EMG Biopotential Acquisition System Architecture

A high-fidelity hardware conditioning chain was developed to amplify and adapt the raw microvolt-level muscular potentials before digitization. The structural topology of the custom-developed frontend is organized into the following sequential blocks, summarized in Fig. 1:

1) *Primary Instrumentation Stage:* Centered around the precision INA128 instrumentation amplifier, selected for its high input impedance ($10^{10}\,\Omega$) and excellent Common-Mode Rejection Ratio (CMRR $>$ 120 dB). The primary differential gain G is governed by an external resistor R_g according to:
$$G = 1 + \frac{50\ \text{k}\Omega}{R_g} \tag{1}$$
An arrangement using a fixed $51\,\Omega$ resistor coupled to a 10 kΩ precision potentiometer enabled accurate gain adjustments.
2) *Active Bandpass Cascading Structure:* To isolate the neuromuscular spectrum, a 4th-order active high-pass Sallen-Key filter (cutoff frequency $f_c = 10$ Hz) was cascaded with a 4th-order active low-pass Sallen-Key filter ($f_c = 300$ Hz). Both sections utilize a maximally flat Butterworth response topology ($Q = 0.707$). The high-pass component values were dimensioned via:
$$4\pi\frac{f_c}{Q} = \frac{2}{R_2}\left(\frac{1}{C_1} + \frac{1}{C_2}\right) \tag{2}$$
yielding $R_2 = 2.25$ kΩ, $R_1 = 1.125$ kΩ for $C_1 = C_2 = 10\,\mu$F. The low-pass section component calculations resulted in $C_1 = 250$ nF and $C_2 = 125$ nF for a fixed $R = 3$ kΩ.
3) *Baseline Restorer (BLR):* An active integrator loop feeds the output offset voltage back into the reference pin of the INA128. This forms a continuous-time high-pass loop to eliminate DC drift without attenuating metabolic AC signals. With loop constants $R = 2.2$ MΩ and $C = 20$ nF, the equivalent cutoff satisfies:
$$f_{eq} = \frac{1}{2\pi RC} = 3.61\ \text{Hz} \tag{3}$$
4) *Driven Right Leg (DRL) Active Loop:* Common-mode voltage is sensed at pins 1 and 8 of the input buffer, inverted, amplified, and actively reinjected back to the subject through a safety-limiting resistor (1 MΩ). Frequency compensation via a parallel network ($R = 10$ kΩ, $C = 1.6$ nF) ensures loop stability against body capacitance.
5) *Level-Shifter and Safe Clipper Frontend:* Operates as a non-inverting scaling stage ($A_v \approx 101$) providing a fixed 2.75 V DC offset to align the biopotential centered inside the $0 - 5$ V unipolar range of the ADC. A low-voltage Zener barrier (1N4731A, $V_Z = 4.3$ V) tied to a biased voltage divider restricts potential spikes to a rigorous $0.075\ \text{V} - 4.925\ \text{V}$ absolute window to safeguard the microcontroller input.

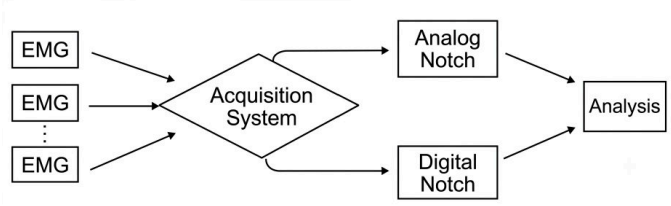


Fig. 1. Comprehensive architectural block diagram of the custom-developed electromyographic signal acquisition chain, highlighting the analog frontend blocks, filtering stages, and digital interface.

B. Analog Notch Filter Topologies

To establish an equitable baseline for structural comparison, the target line attenuation frequency was set to $f_0 = 60$ Hz, and the targeted quality factor was fixed at a reference value of $Q = 3.0$ across all continuous-time networks. Five architectures were constructed, two of which are illustrated in Fig. 2:

1) *Fliege Architecture:* Renowned for its highly selective band-rejection performance, this topology permits independent tuning of Q and f_0. Modeled using component constraints $R_5 = R_6 = 10$ kΩ, $C = 330$ nF, where $R_3 = R_4 = 8$ kΩ models the notch center frequency, and $R_1 = R_2 = 2QR_3 = 48$ kΩ tunes the selectivity.
2) *Multiple Feedback (MFB) Topology:* Employs the active op-amp within an integrator-based configuration. Resistor choices were derived from standard mathematical transfer definitions: $R_1 = Q/(2\omega_0 C) = 26.5$ kΩ, $R_2 = R_1/(Q^2 - 1) = 3.31$ kΩ, and $R_4 = 4R_1 = 106.08$ kΩ given a scaled baseline capacitance of $C = 150$ nF.
3) *Sallen-Key Notch Filter:* Configured as a voltage-controlled voltage source (VCVS) network where the active node functions strictly as a buffer. For a target capacitance $C = 220$ nF, the structural parameters evaluate to $R = 12$ kΩ with op-amp scaling feedback values set to $R_1 = 30$ kΩ and $R_2 = 20$ kΩ to yield an active gain of approximately 1.6.
4) *Tow-Thomas State-Variable Network:* Implemented as a continuous-time analog biquad integrating layout. The

discrete combination of parallel high-pass and low-pass output nodes forms the notch response. Choosing $C = 100$ nF and scaled parameters $R_1 = R_2 = R_3 = 100$ kΩ fixes the central tuning resistor to $R_4 \approx 7$ kΩ.

5) *Twin-T Passive-Active Configuration:* Employs a dual-path passive resistive-capacitive network inside an active feedback loop dedicated to scaling the sharpness of the notch. Component dimensioning based on $C = 220$ nF yields a baseline loop resistance of $R = 12$ kΩ.

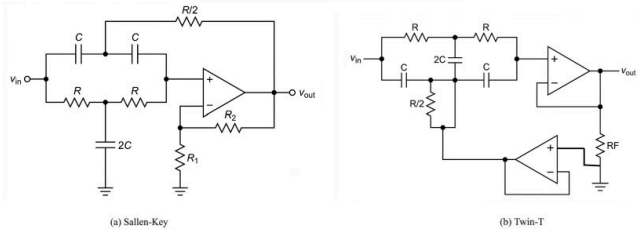


Fig. 2. Conceptual circuit diagrams of the two main topologies of active analog band-rejection filters evaluated.

C. Digital Filter Derivation and Post-Processing

The software-based counterparts were derived directly from the continuous-time Laplace transfer equations $H(s)$ of each physical circuit rather than standard canonical forms. Mapping into the discrete-time z-domain was performed via Bilinear Transformation (Tustin's method) to ensure exact preservation of stability boundaries:

$$s = \frac{2}{T_s}\frac{1 - z^{-1}}{1 + z^{-1}} \tag{4}$$

where $T_s = 1/f_s$ represents the sampling period ($f_s = 1200$ Hz, satisfying oversampling conditions equal to four times the low-pass cutoff frequency). To eliminate computational phase distortion, the derived Infinite Impulse Response (IIR) digital structures were executed via forward-backward filtering operations (`filtfilt`), rendering a strict zero-phase implementation.

D. Experimental In-Vivo Acquisition Protocol

Validation data were gathered from a single healthy adult volunteer using standard surface Ag/AgCl electrodes. The targeted neuromuscular anatomical area was the *biceps brachii* muscle body, mapped according to the SENIAM framework guidelines [12]. Local dermis preparation included mechanical abrasion and isopropyl cleansing to minimize the skin-electrode impedance boundary layer. The active differential pair was placed along the structural midline of the muscle fibers, while a reference electrode was secured over the low-activity acromion process. The movement sequence mandated five consecutive isometric voluntary muscle contractions, executed evenly within a fixed 15-second tracking window, maintaining a strict relaxation interval between bursts to eliminate neuromuscular fatigue variables. The five contractions were recorded as one continuous segment and processed as a single record; they were not treated as independent repetitions, and no dispersion statistics were derived from them (see Section III-E).

E. Ethical Considerations and Informed Consent

The experimental record analyzed in this work was obtained from one healthy adult volunteer, who is also one of the authors of this study and who provided informed consent prior to the acquisition, having been fully informed of the procedure, of its purpose, and of the intended use of the resulting data. The protocol was entirely non-invasive: it involved only surface Ag/AgCl electrodes applied to intact skin, with no pharmacological agent, no electrical stimulation, and no invasive procedure of any kind. Subject safety in the acquisition chain is enforced at the hardware level by the current-limiting resistor (1 MΩ) in series with the DRL reinjection path and by the Zener clamping network described in Section II-A. No clinical, demographic, or otherwise identifiable personal data were recorded, stored, or transmitted; only the anonymized biopotential waveform was retained, and it was used exclusively for the signal-processing analysis reported here. The study was conducted in accordance with the principles of the Declaration of Helsinki and with the IEEE EMBS Code of Ethics. The protocol was submitted to the Research Ethics Committee *Wladirson Ronny da Silva Cardoso* (Universidade do Estado do Pará, UEPA) on 6 August 2026 under protocol no. 2865255; committee approval was still pending at the time of manuscript submission, and the corresponding document will be provided to the editors as soon as it is issued.

F. Quantitative Performance Metrics

1) *Absolute Spectral Attenuation (A_{60Hz}):* Estimated using Welch's modified periodogram method with a non-overlapping Hanning window setup to suppress spectral leakage. Line attenuation was calculated across a locked ± 1 Hz band centered on the power network component:

$$A_{60\text{Hz}}\ [\text{dB}] = 10\log_{10}\left(\frac{\overline{P_{xx}}(59 - 61\ \text{Hz})}{\overline{P_{yy}}(59 - 61\ \text{Hz})}\right) \tag{5}$$

Acceptance criteria reported in the surface EMG denoising literature are adopted here as the reference scale: attenuation above 20 dB is regarded as adequate for the suppression of power-line interference, and attenuation above 30 dB as excellent [10], [11].

2) *Signal Preservation Coefficient (P_{EMG}):* Assesses structural integrity across the primary integration window ($10-300$ Hz), verifying that out-of-band energy profiles are preserved:

$$P_{\text{EMG}} = \frac{\int_{10}^{300} P_{yy}(f)df}{\int_{10}^{300} P_{xx}(f)df} \tag{6}$$

Note that the integration window contains the line component itself, so a fraction of the observed reduction in P_{EMG} corresponds to the interference that the filter was designed to remove rather than to lost physiological content.

3) *Temporal Global RMS Shift (Δ_{RMS}):* Quantifies time-domain energy attenuation using sliding windows ($W = 100$ ms, 50% overlap).
4) *Resulting Signal-to-Noise Ratio (SNR):* Determined by treating the difference vector $r[n] = x[n] - y[n]$ as the baseline noise floor:

$$\text{SNR [dB]} = 20 \log_{10} \left(\frac{\overline{\text{RMS}}_{filtered}}{\text{RMS}_{noise}} \right) \tag{7}$$

III. Results and Discussion

A. Hardware Validation Benchmark: Protoboard vs. PCB

Initial testing compared temporary protoboard assemblies against the finalized custom-fabricated PCBs using a deterministic 1 V_{pp} 60 Hz sinusoidal test signal. The empirical results are compiled in Table I.

The migration from temporary breadboard assemblies to fabricated PCBs did not produce a uniform improvement, and the results must be reported as mixed. The residual 60 Hz amplitude decreased for three of the five topologies (Fliege, $176 \rightarrow 130$ mV; Multiple Feedback, $136 \rightarrow 110$ mV; Tow-Thomas, $152 \rightarrow 130$ mV), remained essentially unchanged for the Sallen-Key ($116 \rightarrow 120$ mV), and degraded severely for the Twin-T, whose residual rose from 44 mV to 140 mV. The topology that produced the deepest null on the breadboard therefore became one of the shallowest once transferred to a board. Center-frequency placement follows an equally mixed pattern: the mean absolute deviation from the 60 Hz target decreased only marginally, from 0.22 Hz to 0.18 Hz, with three topologies moving closer to nominal and two moving further away. The measured bandwidth widened in four of the five circuits, indicating a general reduction of the effective quality factor after fabrication.

For the three topologies that improved, the gain is consistent with the elimination of long jumper wires and unstable contact resistances inherent to breadboard layouts, which introduce stray capacitive coupling and parasitic impedances capable of displacing the pole-zero placement of high-Q active networks. The Twin-T behaves differently because its rejection mechanism is fundamentally different: the null arises from the cancellation of two nearly equal signals travelling through the complementary R-$2C$ and $2R$-C arms, and the depth of that null is theoretically unbounded only while the two arms remain exactly balanced. The residual at f_0 is thus a small difference between large quantities, so even a modest absolute asymmetry produces a large relative change in the measured residual. On a PCB, the distributed capacitance of the copper traces and of the ground plane at the internal T node, the leakage paths across the board surface, the input capacitance of the active device, and a different realization of component tolerances all act to unbalance the arms. The breadboard build happened to sit close to balance, which explains its 44 mV residual; the fabricated board did not. This makes the Twin-T the least layout-tolerant of the five topologies evaluated, and its PCB implementation is consistent with this interpretation, exhibiting simultaneously the shallowest null degradation and the widest bandwidth of the set (22 Hz, i.e. an effective $Q \approx 2.7$ against the nominal $Q = 3.0$). Practical implementation of a Twin-T on a board therefore requires guarded routing at the T node, matched or trimmable arm components, and post-assembly tuning, none of which were applied here.

B. Analog Filtering Performance Characterization

Following the in-vivo experimental acquisition protocol, the continuous-time notch filtering systems were evaluated. Quantitative signal processing results are detailed in Table II, and the corresponding spectral comparison is shown in Fig. 3.

The Sallen-Key architecture outperformed all alternative hardware-based filters, achieving an absolute 60 Hz attenuation of 20.37 dB. This is the only analog topology reaching the 20 dB level adopted in the surface EMG literature as the threshold of adequate power-line suppression [10], [11]. The Tow-Thomas network achieved a comparable attenuation of 18.74 dB, marginally below that level, but introduced a higher signal distortion profile, reducing the structural preservation index to a low value of 54.60%. Given that the two figures differ by less than 2 dB and originate from separate single-trial acquisitions, this ordering should not be treated as a resolved ranking (Section III-E). The low relative performance of the remaining hardware topologies can be attributed to out-of-band phase distortions and component value drift. The negative SNR calculations arise from the strict evaluation method used, which treats any out-of-band attenuation or phase deviation within the raw baseline signal as noise.

C. Digital Filtering Performance Characterization

The discrete-time IIR filters, mapped via exact bilinear discretization, were executed on the raw (unfiltered) digital EMG signal vectors. The processing metrics are summarized in Table III, with the corresponding spectral comparison shown in Fig. 4.

In the software domain, the Twin-T digital configuration demonstrated the deepest line suppression of the set, rendering a notch depth of 79.62 dB coupled with a reconstruction SNR of 77.46 dB. This sharpness is a direct result of digital execution, which operates with absolute coefficient precision and is immune to physical component tolerances, temperature variations, or contact degradation. The digital Sallen-Key and Tow-Thomas filters also formed a highly effective processing tier, achieving strong line attenuation (> 54 dB) while maintaining superior out-of-band signal preservation compared to the Twin-T structure.

The Multiple Feedback result requires separate comment, as it is a clear outlier. Its digital attenuation at 60 Hz (0.44 dB) lies two orders of magnitude below every other digital implementation ($42.07 - 79.62$ dB) and is also inconsistent with the behavior of the same topology in hardware, where the fabricated MFB board reduced a 1 V_{pp} test tone to 110 mV (Table I) and reached 10.28 dB of attenuation on the in-vivo record (Table II). The remaining digital metrics reinforce this reading rather than contradicting it: the highest preservation

TABLE I
EXPERIMENTAL BENCH COMPARISON OF NOTCH TOPOLOGIES: PROTOBOARD VS. PCB

Topology	Protoboard Assembly			PCB Implementation		
	f_0 (Hz)	BW (Hz)	A_{60Hz} (mV)	f_0 (Hz)	BW (Hz)	A_{60Hz} (mV)
Fliege	60.0	12.0	176	59.8	19.0	130
Multiple FB	60.2	15.0	136	60.3	11.6	110
Sallen-Key	59.8	19.0	116	60.1	19.5	120
Tow-Thomas	60.4	12.0	152	59.9	16.5	130
Twin-T	59.7	20.0	44	59.8	22.0	140

TABLE II
PERFORMANCE METRICS OF ANALOG NOTCH FILTER TOPOLOGIES

Topology	A_{60Hz} (dB)	Preservation (%)	Delta RMS (dB)	SNR (dB)
Fliege	7.20	61.00	4.29	-6.39
Multiple FB	10.28	77.70	2.19	-4.85
Sallen-Key	**20.37**	**67.00**	**3.48**	**-6.23**
Tow-Thomas	18.74	54.60	5.26	-7.07
Twin-T	8.85	60.90	4.31	-6.08

TABLE III
PERFORMANCE METRICS OF DIGITAL NOTCH TOPOLOGIES

Topology	A_{60Hz} (dB)	Preservation (%)	Delta RMS (dB)	SNR (dB)
Fliege	42.07	86.40	3.15	51.10
Multiple FB*	0.44	94.90	0.45	8.90
Sallen-Key	56.73	79.28	2.02	64.96
Tow-Thomas	54.16	82.32	1.69	59.95
Twin-T	**79.62**	**60.44**	**4.37**	**77.46**

*Result treated as inconclusive; see discussion in Section III-C.

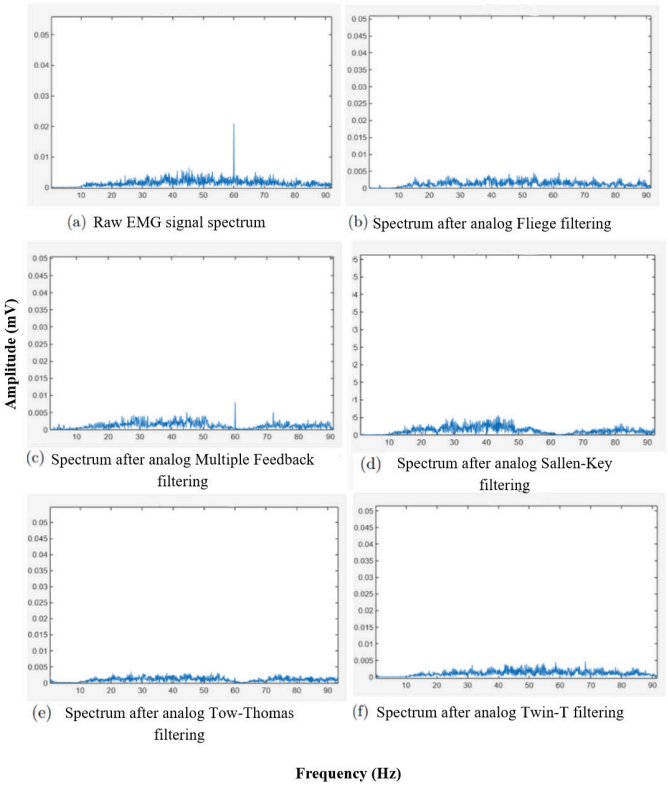


Fig. 3. Frequency spectra comparison of the raw surface electromyographic (EMG) signal against the five evaluated active analog notch filter implementations.

index of the entire set (94.90%) combined with the smallest global RMS shift (0.45 dB) describes a transfer function that is close to transparent across the whole band, which is the signature of a filter that removes almost nothing, not of a selective band-stop response. We therefore interpret this value as an implementation artifact rather than an intrinsic limitation of the MFB architecture. The most probable cause is an error in the symbolic derivation of $H(s)$ for this particular network. The MFB notch response is formed by combining a band-pass path with the direct input path, so an incorrect gain, sign, or resistor ratio in that combination displaces the transmission zeros off the imaginary axis and leaves a shallow dip in place of a true null. This network is especially exposed to such an error because the zero placement depends strongly on the ratio $R_2 = R_1/(Q^2-1)$, which for $Q = 3.0$ implies a factor of eight between the two resistors. A secondary hypothesis is numerical degradation in the discretization, although the absence of frequency pre-warping alone displaces the discrete notch by less than 0.5% at $f_s = 1200$ Hz, which cannot account for a 0.44 dB result given the nominal notch bandwidth of approximately 20 Hz. Verification of the numerator coefficients, confirming that the discrete zeros lie on the unit circle at $\omega_0 T_s$, and re-implementation as cascaded second-order sections with pre-warping, are required before this topology can be ranked fairly. Until that verification is performed, the digital MFB figure is reported here as inconclusive and is excluded from the comparative ordering.

The rejection advantage of the digital Twin-T carries a measurable cost in spectral fidelity that must be stated explicitly, since this is the topology recommended for the digital stage of the architecture proposed in Section IV. Its preservation index (60.44%) is the lowest of the digital set, between 19 and 34 percentage points below the other digital implementations, and its global RMS shift (4.37 dB) is the largest. Because P_{EMG} is integrated over $10 - 300$ Hz, that is, over the band

that carries the neuromuscular information, the loss is not cosmetic: MUAP morphology, median and mean frequency estimators, and amplitude-based activation indices are all derived from this band and are therefore susceptible to bias, a mechanism already reported for aggressive notch filtering [8]. Part of the reduction necessarily corresponds to the removal of the interference itself, which also lies inside the integration window, and separating the two contributions would require recomputing P_{EMG} over $10 - 300$ Hz with the $59 - 61$ Hz band excluded; this refinement was not performed here and is left for future work. In practical terms, applications whose objective is the removal of a dominant line component with limited sensitivity to spectral shape, such as onset detection or gross activation timing, can accommodate this trade-off. Applications in which spectral parameters are themselves the measured quantity, such as fatigue monitoring through median-frequency shifts, conduction-velocity estimation, or MUAP-based indices, are better served by the digital Sallen-Key (56.73 dB, 79.28%) or Tow-Thomas (54.16 dB, 82.32%), which still exceed 54 dB of line rejection while retaining approximately 80% of the in-band energy.

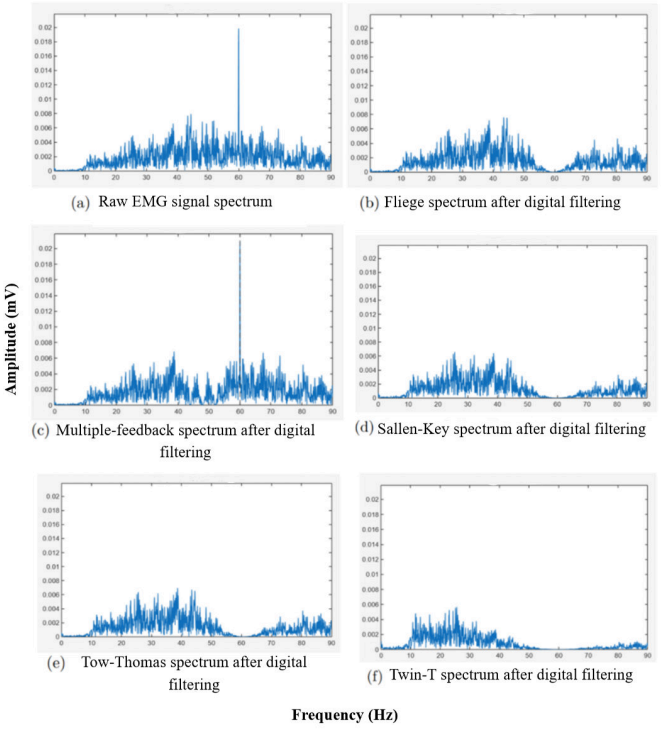


Fig. 4. Frequency spectra comparison of the raw surface electromyographic (EMG) signal against the five evaluated discrete-time zero-phase IIR notch filter implementations.

D. Analog vs. Digital Architectural Discussion

Comparing the two analytical frameworks reveals that digital implementations provide significantly deeper, more stable, and more precise line frequency attenuation than physical hardware filters. However, this comparison must be interpreted within the context of a complete biomedical acquisition system.

Relying entirely on digital post-processing means the input frontend must handle heavily contaminated raw signals. If power-line noise magnitudes exceed the limits of the ADC frontend, it can lead to signal saturation, clipping, and irreversible quantization distortion. Conversely, incorporating an analog notch stage (such as the Sallen-Key topology) minimizes noise amplification early in the conditioning chain, preserving the full dynamic range of the ADC. The optimal solution for high-fidelity electromyographic instrumentation is therefore a hybrid architecture, combining early analog filtering to prevent saturation with digital IIR processing for final spectral cleanup. Within that architecture, the analog stage is best served by the Sallen-Key, which was both the only analog topology to reach the 20 dB criterion and the one whose bench behavior proved least sensitive to the transition from breadboard to fabricated board. The digital stage admits two defensible choices: the Twin-T when maximum line rejection is the priority, and the Sallen-Key or Tow-Thomas when preservation of in-band spectral content is critical to the intended measurement.

E. Study Limitations

The scope of the conclusions that can be drawn from this comparison is bounded by the experimental design. All in-vivo results reported in Tables II and III and in Figs. 3 and 4 originate from a single healthy volunteer, a single muscle, and a single acquisition session. The five contractions were performed within one continuous 15-second window and were processed as one record rather than as independent repetitions, so no standard deviations, confidence intervals, or significance tests accompany the reported metrics. Every value in the tables is consequently a point estimate from one realization, and differences of a few decibels between adjacent topologies fall within a margin that the design cannot resolve. Large separations, such as that between the digital Twin-T and the remaining digital filters, are robust to this limitation; close comparisons, such as the 1.63 dB gap between the analog Sallen-Key and Tow-Thomas, are not.

A second constraint concerns the asymmetry between the two comparisons. The digital topologies were applied to a common raw record, so the differences among them reflect the filters alone. Each analog topology, in contrast, required its own acquisition, since only one board could be inserted in the conditioning chain at a time; the analog comparison therefore carries the natural physiological variability between contractions of the same subject, arising from motor unit recruitment patterns and small differences in electrode-skin conditions across trials. Future evaluations should adopt a programmable generator reproducing a synthetic EMG signal with known spectral and temporal characteristics, so that an identical stimulus can be replayed through every topology, and should extend the in-vivo protocol to multiple volunteers, muscle groups, contraction levels, and fatigue conditions before any topology ranking is presented as generalizable.

IV. Conclusion

This study evaluated and compared five active notch filter configurations across both the continuous analog hardware domain and the discrete digital software domain. The experimental results show that the Sallen-Key configuration provides the most effective analog hardware implementation, and the only one reaching the 20 dB attenuation criterion adopted in the literature [10], [11], while the digital IIR implementation of the Twin-T topology offers the deepest line rejection, achieving an attenuation of 79.62 dB. That advantage is not free: the digital Twin-T also exhibits the lowest in-band preservation of the digital set (60.44%), so the digital stage of a hybrid conditioning chain should be selected according to the intended measurement, with the Twin-T reserved for applications dominated by line interference and the Sallen-Key or Tow-Thomas preferred where spectral fidelity governs the result. The digital Multiple Feedback result remains inconclusive and requires verification of its transfer function derivation before it can be ranked. These findings derive from a single-subject, single-session acquisition without independent replication and are therefore offered as a proof-of-concept characterization. Future research will focus on validation against a synthetic reference signal with known spectral content, on automated adaptive digital filtering algorithms, and on the evaluation of hybrid conditioning systems across a broader, multi-subject clinical testing group.

Acknowledgment

The authors express their gratitude to the Federal University of Pará (UFPA) and the team at the Biomedical Engineering Laboratory (E.Bio) for providing the experimental infrastructure and technical support critical to the completion of this study.

References

[1] C. J. De Luca, "Surface Electromyography: Detection and Recording," *Delsys Inc.*, vol. 11, no. 1, pp. 1–17, 2002.

[2] R. Merletti and P. Parker, *Electromyography: Physiology, Engineering, and Non-Invasive Applications*. New York: IEEE Press, John Wiley & Sons, 2004.

[3] K. Aparna et al., "Development of high-performance analog and digital filter for biomedical signal processing," *ICTACT Journal on Microelectronics*, vol. 9, no. 2, pp. 1539–1544, 2023.

[4] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface EMG," *Journal of Applied Physiology*, vol. 117, no. 11, pp. 1215–1230, 2014.

[5] J. G. Webster, *Medical Instrumentation: Application and Design*, 4th ed. New York: John Wiley & Sons, 2009.

[6] R. L. Boylestad and L. Nashelsky, *Electronic Devices and Circuit Theory*, 8th ed. São Paulo: Pearson Prentice Hall, 2004.

[7] S. W. Smith, *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Amsterdam: Newnes, 2003.

[8] X. Li et al., "The effects of notch filtering on electrically evoked myoelectric signals and associated motor unit index estimates," *Journal of NeuroEngineering and Rehabilitation*, vol. 8, no. 64, pp. 1–9, 2011.

[9] M. V. A. dos Santos et al., "Análise e comparação do desempenho de filtros notch numa aplicação de sinais de eletromiografia" (Analysis and comparison of notch filter performance in an electromyography signal application), in *Proc. Congresso Brasileiro de Engenharia Biomédica (CBEB)*, 2024, pp. 1–6, (in Portuguese).

[10] D. T. Mewett, K. J. Reynolds, and H. Nazeran, "Reducing power line interference in digitised electromyogram recordings by spectrum interpolation," *Medical & Biological Engineering & Computing*, vol. 42, no. 4, pp. 524–531, 2004.

[11] G. Lu et al., "Removing ECG noise from surface EMG signals using adaptive filtering," *Neuroscience Letters*, vol. 462, no. 1, pp. 14–19, 2009.

[12] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for SEMG sensors and sensor placement procedures," *Journal of Electromyography and Kinesiology*, vol. 10, no. 5, pp. 361–374, 2000.

Comparative Evaluation of CNNs, Vision Transformers, and Ensemble Methods for Breast Cancer Lymph Node Metastasis Detection

L. M. Torres
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
lucas.modesto@ufam.edu.br

M. G. F. Costa
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
mcosta@ufam.edu.br

C. F. F. Costa Filho
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
ccosta@ufam.edu.br

***Abstract*—Breast cancer is the leading cause of female mortality related to cancer in the world, with histopathological analysis of lymph nodes tissue the gold standard technique for metastasis detection. The existence of digitalization of Whole Slide Images (WSI) and availability of public datasets such as PatchCamelyon (PCam) has enabled the development of autonomous applications for diagnosis using deep learning algorithms. Convolution Neural Networks (CNNs) are well known for its capabilities of local features extraction while Vision Transformers (ViTs) capture global dependencies, but a comparative evaluation of ensemble and fusion of both architectures are scarce or inexistent in the literature. This paper aims to present a comprehensive analysis of CNNs and ViTs architectures for metastatic breast cancer detection using the original PCam dataset and evaluate ensemble and fusion of both models. We investigated three strategies for models' selection to the ensemble and fusion experiments (lowest validation loss, loss per architecture type, and mean correlation error - MCE) combined with four techniques (average ensemble, weighted ensemble, MLP fusion, and Transformer fusion), resulting in 60 experiments. Our results indicate that the Swin Transformer achieves the best individual performance with 92.51% accuracy and 0.9734 of AUC, while the Top4-LpAT-MLP achieves the highest AUC and the Top6-MCE-Transformer achieves a stable overall performance with 92.39% accuracy and 0.9771 of AUC. The MCE strategy of selection achieved stable metrics and is also competitive to other evaluated fusion techniques, supporting the premise that model complementarity could be useful in fusion and ensemble design. Furthermore, our best configuration surpasses slightly the state-of-the-art for AUC metric from works using the original PCam dataset.**

Keywords—breast cancer, lymph node metastasis, deep learning, convolutional neural networks, vision transformers, ensemble learning, histopathological image classification, PatchCamelyon

I. Introduction

Breast cancer is the most diagnosed cancer among women and is still the leading cause of cancer related mortality. According to the World Health Organization (WHO), approximately 2.3 million new cases and 670,000 deaths were registered in 2022. These figures highlight the importance of early diagnosis and effective treatment [1]. Histopathological examination of whole slide images (WSI) of lymph node tissue is the gold standard approach for detection of metastasis in patients with breast cancer; however, this procedure is labor-intensive, time-consuming, and extremally dependent on the experience and condition of pathologists [2]. The availability of public datasets with WSI, such as Camelyon16 and its variants PatchCamelyon (Pcam) have accelerated the development of diagnosis systems [2].

Convolutional Neural Networks (CNNs) have managed to achieve an overall good performance in WSI images classification by learning local tissue patterns, whereas Vision Transformer (ViT) comes to provide a global contextual relationship through self-attention [3][4][5]. Even though hybrid CNN-Transformer architectures were proposed to exploit these characteristics, they generally increase model complexity [4][5]. Conversely, ensemble and fusion offer an alternative solution that combines the usage of well-known trained models from literature while preserving their individual strengths. Nevertheless, comparative studies that investigated ensembles of CNNs and ViTs for metastasis detection of breast cancer lymph node samples remain limited.

This paper aims to show a comprehensive comparison of CNN and ViT architectures and their combination for ensemble and fusion techniques for metastatic breast cancer detection using the original PCam dataset. The proposed approach investigates whether combining models with their features representation can improve the classification while avoiding the complexity of hybrid solution designs.

II. Related works

The accelerated usage of Whole Slide Images (WSI) has led to advances in digital pathology making it possible the development of computer-aided diagnostic. Alongside this technological development, public WSI datasets for deep learning research emerged. The Cancer Metastases in Lymph Nodes Challenge 2016 Competition (Camelyon16) provided a large annotated WSI dataset, 399 whole-slide images, for breast cancer lymph node metastasis detection. This competition indicates that deep learning solutions could achieve diagnostic performance close to experienced pathologists while maintaining consistent performance under time constraints [2].

To enable an easy usage of patch wise classification and reproducibility of Camelyon16 dataset, Veeling et al. introduced the patch variant, named PatchCamelyon (PCam), containing 327,680 labeled image patches. In addition to establishing one of the most widely used benchmarks for histopathology classification, the authors also demonstrated that rotation and reflection equivariant models optimizes the usage of their internal kernel filters to efficiently classify the samples [3].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

After the release of PCam, some studies investigated the usage of Convolutional Neural Networks (CNNs) for breast cancer metastasis detection. Several approaches relied on models such as VGG, ResNet, DenseNet and Inception, indicating that CNNs could extract tissue features effectively from image patches [6][7], while others incorporated transfer learning, data augmentation and attention mechanisms to improve the classification performance. Liang et al. presented the Convolutional Block Attention Module (CBAM) to emphasize the relevant regions achieving an AUC of 0.976 on PCam dataset [8]. Additionally, Ahmad et al. used preprocessing and transfer learning techniques with CNNs, achieving an accuracy of 98% on PCam [9] and Shi et al. proposed the CSAM-ResNet, that integrates channel and spatial attention mechanisms to enhance contextual feature extraction [10]. Even though CNNs remain effective for local feature extraction, their limitation for global context affects their performance [4][5].

Vision Transformers (ViTs) have managed to deal with this limitation by applying the self-attention mechanisms capable of learning global relationships. Wang et al. presented LGViT, which is a hybrid model that combines CNN layers with a Local-Global Multi-head Self-attention mechanism to capture both local and global features while also keeping the computational cost relatively low [4]. More recently, Mukadam et al. using image enhancement and multi-attention CNN modules illustrates these advances for classification on PCam [11].

Ensemble has also appeared as an effective technique for improving the diagnostic by combining complementary models. Jaiswal et al., shown that ensembles of DenseNet models outperform individual networks on PCam [12]. Rane et al. proposed an attention fusion network that combines the DenseNet201, InceptionV3, and ResNext-50 through an attention module, achieving an AUC of 0.9816 [13]. Additionally, Ge et al. presented DECENN, an ensemble architecture integrating VGG16 and DenseNet121 with multi-scale feature extraction and feature harmonization modules achieving an AUC of 0.9970 [14]. However, both Rane and Ge used a deduplicated version of PCam dataset, which makes it hard to fairly compare with approaches using the original PCam.

Despite these advances, most of existing works aim on improving individual CNNs, ViTs, proposing hybrid CNN-ViT architectures, or presenting a design of ensemble composed exclusively by CNN models. Little attention has been given to ensemble that combines independently trained CNNs and ViTs while keeping the simplicity of these architectures. Such approach offers the potential to explore the complementary strengths of CNN based local feature extraction and ViT global representation without introducing additional complex hybrid architecture. Motivated by this gap, this work presents a unified comparison of representative CNN and ViT architectures and investigates the ensemble and fusion strategy for breast cancer lymph node metastasis detection on the original PCam dataset.

III. Materials and Methods

A. Materials

The PatchCamelyon (PCam) database [3], used in this work, consists of 327680 labeled images (96x96 pixels) extracted from histopathologic lymph node tissue (normal and metastasis samples), which are already publicly available with the training, validation and test proportion of, respectively, 80:10:10. The experiments were executed using an A100 NVidia® GPU with 164GB of RAM from Google Collab environment.

B. Methods

For the following training experiments, AdamW was the selected optimizer, with a weight decay of 0.05 and the initial learning rate for each architecture following the results presented in Table I. The best checkpoint for each model is selected based on the lowest validation loss.

Fig. 1 shows the framework designed to achieve the goal of this work, which will be presented below:

a) Data augmentation: The data augmentation techniques applied are: random horizontal and vertical flip, random erasing, and default operations from RandAugment from TorchVision. A sampling of augmented dataset is used for each epoch, resulting in different samples for each epoch.

b) Training of single models: The trained single models are: EfficientNets, Resnets, VGG, Inception, Xception, ViTs, ConvNext.They are trained using the train split from PCam dataset and all validation metrics are stored to support the selection and combination of models for the ensemble and fusion approachs.

c) Selection of models for ensemble and fusion: For the ensemble and fusion experiment the following strategies of selection is used, resulting in 3 sets of top 6 models:

- Loss: For this approach the simple selection of lowest validation loss is used to get the ideal set of models for ensemble/fusion experiments.
- Loss per Architecture type (LpAT): This strategy of model selection uses a similar approach compared to the loss described above, but in this case only one – the best based on validation loss – model for each type of architecture is selected.
- Mean Correlation Error (MCE): For this strategy, a mean correlation error matrix is generated to know how similar models tend to make mistakes. The idea is to select the models that usually fail differently from the other models. Therefore, the models for MCE were selected using the lowest mean correlation error from the validation split of PCam.

d) Experiment of ensemble and fusion: For ensemble and fusion, the following techniques are used: average ensemble, weighted ensemble Table VII (each weight is calculated by its AUC metric divided by the sum of all selected models), multilayer perceptron (MLP) fusion, Transformer Fusion. In the ensemble approaches, a simple average ensamble and a weighted ensemble are used. The MLP fusion ilustrated on *Fig. 2* works by basically concatenating the features, classification logits, and classification probabilities from each model and feeding them to the MLP classifier. Transformer fusion on *Fig. 3* also uses the concatenated features, logits, and probabilities considering each model as a token and adding a classification token at the beginning to use its adjusted embeddings at the end for the classification step. Finally, the train split was used for training both MLP and Transformer fusion.

e) Performance metrics for comparison: A total of 22 individual models and the 60 ensemble and fusion

combinations were evaluated on the PCam test set, by computing the following performance metrics: accuracy, precision, recall, f1-score, and AUC. Then, all ensemble and fusion experiments were then compared to the individual models.

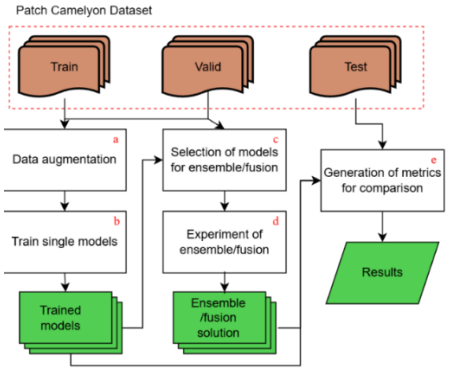


Fig. 1 Framework for classification experiments on lymph node images (normal or metastasis tissue).

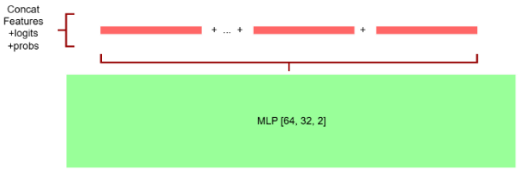


Fig. 2 Experiment of fusion: MLP fusion

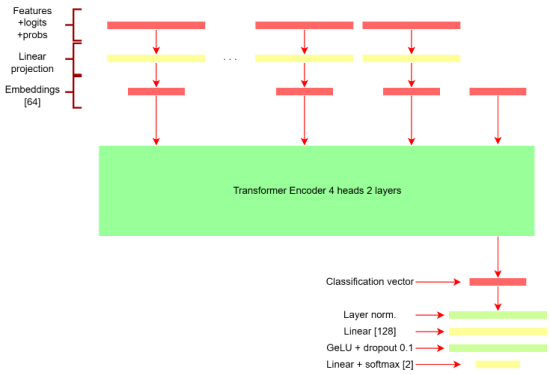


Fig. 3 Experiment of fusion: Transformer fusion

IV. Results and Discussion

A. Hyperparameter tuning for individual models

The first experiment executed was about the ideal configuration of hyperparameter for each architecture to achieve the maximum potential for all models.

From the results summarized on Table I we concluded that for almost all type of architecture the native dimension – 300x300 for inception/xception and 224x224 for the others – achieved better results based on validation loss. Only VGG had a better result for dimension of 96x96. Also, the learning rate of 1E-3 achieved better results, except for VGG and Xception (1E-4). Finally, all these results were used to match the ideal setup for each model in the following experiments.

B. Training of individual models

The results for individual models are summarized on Table II. Swin transformer achieved better results for most metrics, except Recall and AUC, for which the best models where ViT Large and Resnet101, respectively. Also, it is noteworthy that the performance of Swin transformer was nearly two points higher than other models.

Additionally, the Fig. 4 shows that there is no correlation between the better AUC results with number of parameters of each architecture (number of weights + biases), and bigger models like the ViT Large are getting worst results. This suggests that the number of training samples from PCam may not be sufficient to train such complex models even with the already applied data augmentation.

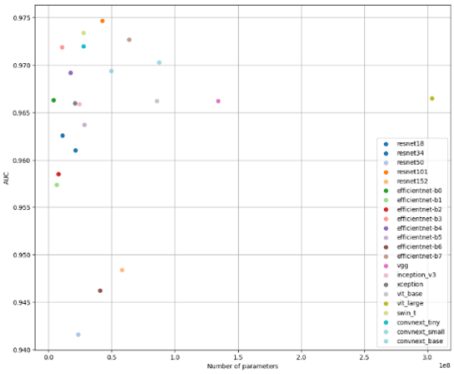


Fig. 4 AUC vs number of parameters of each architecture (number of weights + biases).

C. Selection of models for ensemble and fusion

The following set of models were used for the ensemble and fusion experiments.

a) Loss: By analyzing the 0and selecting the Top 6 models by validation loss we selected following models:

- EfficientNet-B7: 0.2036;
- ViT Base: 0.2142;
- EfficientNet B0: 0.2174;
- Resnet34: 0.2242;
- EfficientNet B3: 0.2244;
- EfficientNet B6: 0.2335.

b) Loss per Architecture Type (LpAT): Also, from 0the following of Top 6 models were selected:

- EfficientNet B7: 0.2036;
- ViT Base: 0.2142;
- Resnet34: 0.2242;
- ConvNext Small: 0.2494;
- Inception: 0.2499;
- VGG: 0.2580.

c) Mean Correlation Error (MCE):

For model i and sample k, the error is defined as bellow, where $\hat{y}_{i,k}$ is the predicted class and y_k is the ground-truth class.

$$e_{i,k} = \begin{cases} 1, if\ \hat{y}_{i,k} \neq y_k \\ 0, otherwise \end{cases}, \quad (1)$$

According to the calculated error, the MCE is generated with the *Pearson correlation coefficient* between the error vectors of models *i* an *j*:

$$p_{ij} = corr(e_i, e_j) \quad (2)$$

Based on MCE for each model from *Fig. 5* and correlation error matrix, ilustrated by *Fig. 6*, the Top 6 models with MCE were sselected:

- Resnet50: 0.4092;
- EfficientNet B6: 0.5092;
- Resnet152: 0.5258;
- Resnet101: 0.5336;
- EfficientNet B5: 0.5485;
- Resnet34: 0.5491.

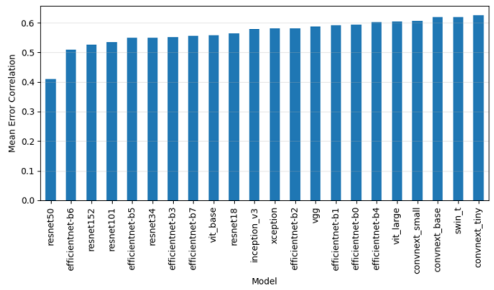


Fig. 5 Mean Correlation Error (MCE) sorted from lowest to highest.

D. *Results for Ensemble and Fusion techniques*

The results for average and weighted ensemble are summarized on Table IV and Table VII while the MLP and transformer fusion results are shown on Table V and Table VI.

Among the approaches for the ensemble and fusion, the Top6-MCE-Transformer achieved the best results for accuracy, precision and F1-score, while the Top4-LpAT-MLP achieved the best AUC and Top6-loss-MLP achieved the best recall.

Considering the better result for the most monitored metrics, the more stable technique is Top6-MCE-Transformer, that achieves the best values for 3 of 5 metrics and not being too distant from the top results for the other 2 metrics. However, a statistical comparison of AUC using DeLong's test indicated that the Top4-LpAT-MLP was indeed higher than the Top6-MCE-Transformer (0.9800 vs. 0.9771, $p < 0.001$), showing its superiority for AUC metric compared to Top6-MCE-Transformer.

E. *SOTA comparison*

The comparison of all experiments and the state-of-the-art solutions identified at the time this paper was written is summarized in Table VIII. TABLE VII. Our experiments achieved the best (Top4-LpAT-MLP) and second (Top6-MCE-Transformer) AUC and kept competitivity with the other metrics.

The Custom CNN [9] model achieved better accuracy even if it got a slightly inferior result for AUC. For medical diagnostic tasks, AUC is a much more robust and relevant metric than raw accuracy, because it measures the model's discriminative capability independently of the classification threshold.

Other papers that used a different version of PCam dataset [13][14], like the deduplicated variant, were filtered out for a fair comparison.

V. Conclusion

The results presented on previous topics indicate that the evaluated experiments achieved competitive performance with previously reported approaches for detection of metastasis using the PCam dataset. Also, the best individual model was Swin transformer that managed to achieve results close to ensemble/fusion techniques. Additionally, the highest AUC was obtained by the Top4-LpAT-MLP configuration with 0.9800 and DeLong's test indicating its superiority to Top6-MCE-Transformer in terms of AUC. However, the Top6-MCE-Transformer strategy achieved competitive and more stable results across most of the monitored metrics and achieved AUC inferior, but close, to those of the Top4-LpAT-MLP approach. These results suggest that complementarity of models could be a useful criterion for construction of ensemble and fusion strategies.

VI. Acknowledgements

This article was supported by the CETELI-INOVA project of the Center for R&D in Electronic and Information Technology at the Federal University of Amazonas.

In accordance with the IEEE policy on the use of AI generated content, the authors disclose that ChatGPT (OpenAI) was used as a support tool for translation into English, textual organization, and grammar revision across all sections of this manuscript. The tool was not used to generate scientific content, experimental results, figures, or code. All intellectual decisions and full responsibility for the final content remain solely with the authors

TABLE I. HYPERPARAMETER EXPERIMENT

Image dimensions	**Learning Rate**	***Cross Entropy Validation Loss***						
		Resnet18	***Resnet50***	***EfficientNet B0***	***EfficientNet B6***	***VGG***	***Inception***	***Xception***
96x96	1E-03	0.2566	0.2836	0.2570	0.2642	0.2635	X	X
	1E-04	0.3117	0.2831	0.2482	0.3158	0.2017	X	X
	5E-04	0.3333	0.2858	0.2725	0.2531	0.2439	X	X
Native Dimension	1E-03	0.2317	0.2645	0.2387	0.2325	0.2808	0.2672	0.3066
	1E-04	0.2918	0.2680	0.2579	0.2799	0.2334	0.2870	0.2645
	5E-04	0.2381	0.2719	0.2782	0.2479	0.2218	0.3085	0.2917

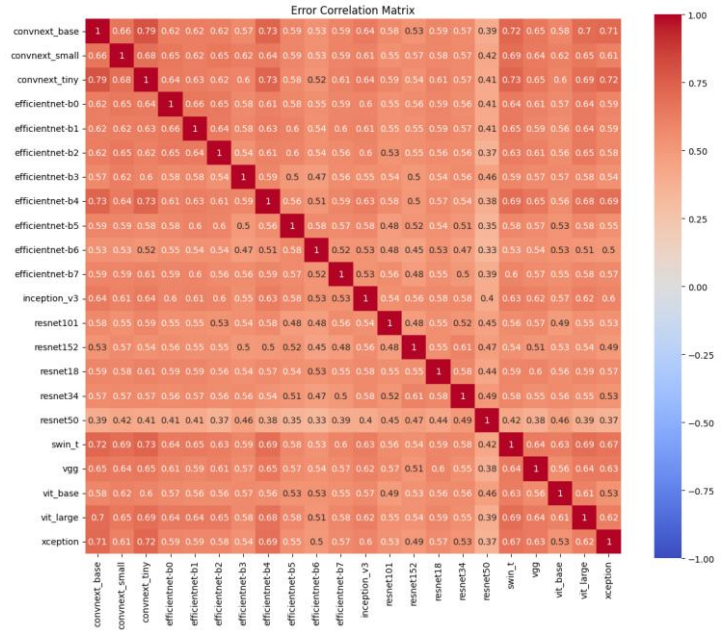


Fig. 6 Correlation error between models.

TABLE II. METRICS FROM TEST SPLIT OF PCAM FOR INDIVIDUAL MODELS

Model	***Accuracy***	***Precision***	***Recall***	***F1-score***	***AUC***	***Loss***
EfficientNet B0	0.8878	0.8329	0.9705	0.8964	0.9663	0.3067
EfficientNet B1	0.8841	0.8294	0.9672	0.8930	0.9574	0.3284
EfficientNet B2	0.8769	0.8183	0.9691	0.8873	0.9585	0.3465
EfficientNet B3	0.9077	0.8729	0.9545	0.9119	0.9719	0.2520
EfficientNet B4	0.8986	0.8488	0.9700	0.9054	0.9692	0.2756
EfficientNet B5	0.8638	0.7993	0.9718	0.8771	0.9637	0.3311
EfficientNet B6	0.8599	0.8072	0.9458	0.8710	0.9462	0.3423
EfficientNet B7	0.9051	0.8652	0.9597	0.9100	0.9727	0.2329
Resnet18	0.8857	0.8384	0.9558	0.8932	0.9626	0.2824
Resnet34	0.8865	0.8448	0.9469	0.8930	0.9610	0.2971
Resnet50	0.8689	0.8573	0.8852	0.8710	0.9416	0.3401
Resnet101	0.9092	0.8724	0.9587	0.9135	**0.9747**	0.2320
Resnet152	0.8666	0.8152	0.9483	0.8766	0.9484	0.3760
VGG	0.8989	0.8511	0.9670	0.9053	0.9662	0.2768
Inception	0.8890	0.8435	0.9554	0.8959	0.9659	0.2682
Xception	0.8904	0.8406	0.9636	0.8979	0.9660	0.2690
ViT Base	0.9029	0.8689	0.9490	0.9072	0.9662	0.2523
ViT Large	0.8925	0.8385	**0.9725**	0.9005	0.9665	0.2976
Swin_t	**0.9251**	**0.8973**	0.9601	**0.9277**	0.9734	**0.2187**
ConvNext tiny	0.9086	0.8692	0.9622	0.9133	0.9720	0.2425
ConvNext small	0.9001	0.8521	0.9683	0.9065	0.9694	0.2857
ConvNext base	0.9055	0.8650	0.9611	0.9105	0.9703	0.2480

TABLE III. METRICS FROM VALIDATION SPLIT OF PCAM FOR INDIVIDUAL MODELS

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*	*Loss*
EfficientNet B0	0.9202	0.8899	0.9592	0.9232	0.9768	0.2174
EfficientNet B1	0.9141	0.8789	0.9608	0.9180	0.9716	0.2408
EfficientNet B2	0.9084	0.8691	0.9619	0.9131	0.9705	0.2593
EfficientNet B3	0.9220	0.9048	0.9435	0.9238	0.9753	0.2244
EfficientNet B4	0.9028	0.8619	0.9593	0.9081	0.9660	0.2781
EfficientNet B5	0.9086	0.8664	0.9665	0.9137	0.9742	0.2407
EfficientNet B6	0.9087	0.8852	0.9394	0.9115	0.9683	0.2335
EfficientNet B7	0.9231	0.8908	0.9644	0.9261	0.9762	0.2036
Resnet18	0.9169	0.8919	0.9490	0.9196	0.9678	0.2344
Resnet34	0.9190	0.9033	0.9385	0.9206	0.9726	0.2242
Resnet50	0.9046	0.9181	0.8887	0.9031	0.9649	0.2618
Resnet101	0.9129	0.8855	0.9486	0.9159	0.9696	0.2457
Resnet152	0.9151	0.8943	0.9416	0.9174	0.9634	0.2702
VGG	0.9083	0.8697	0.9607	0.9129	0.9670	0.2580
Inception	0.9088	0.8758	0.9528	0.9127	0.9659	0.2499
Xception	0.8980	0.8563	0.9567	0.9037	0.9640	0.2660
ViT Base	0.9185	0.9056	0.9344	0.9198	0.9729	0.2142
ViT Large	0.9089	0.8659	0.9678	0.9141	0.9704	0.2612
Swin_t	0.9148	0.8858	0.9526	0.9179	0.9669	0.2525
ConvNext tiny	0.9064	0.8709	0.9546	0.9108	0.9655	0.2626
ConvNext small	0.9235	0.8954	0.9590	0.9261	0.9702	0.2494
ConvNext base	0.9045	0.8652	0.9585	0.9094	0.9644	0.2691

TABLE IV. AVERAGE ENSEMBLE RESUTLS FOR TEST SPLIT OF PCAM

Set of Models	*Rank*	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Loss	Top 2	0.9132	0.8738	0.9659	0.9176	0.9776
	Top 3	0.9111	0.8655	0.9735	0.9164	0.9785
	Top 4	0.9103	0.8648	0.9728	0.9156	0.9787
	Top 5	0.9156	0.8713	0.9753	0.9203	0.9799
	Top 6	0.9120	0.8654	0.9759	0.9173	0.9797
LpAT	Top 2	0.9132	0.8738	0.9659	0.9176	0.9776
	Top 3	0.9140	0.8733	0.9686	0.9185	0.9779
	Top 4	0.9133	0.8691	0.9732	0.9182	0.9787
	Top 5	0.9126	0.8673	0.9743	0.9177	0.9771
	Top 6	0.9142	0.8689	0.9758	0.9193	0.9778
MCE	Top 2	0.8894	0.8524	0.9420	0.8950	0.9580
	Top 3	0.8896	0.8452	0.9539	0.8963	0.9599
	Top 4	0.9012	0.8575	0.9626	0.9070	0.9709
	Top 5	0.9016	0.8524	0.9715	0.9080	0.9734
	Top 6	0.9017	0.8538	0.9695	0.9080	0.9740

TABLE V. MLP FUSION RESUTLS FOR TEST SPLIT OF PCAM

Set of Models	*Rank*	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Loss	Top 2	0.9038	0.8563	0.9706	0.9099	0.9758
	Top 3	0.8867	0.8360	0.9624	0.8947	0.9627
	Top 4	0.8811	0.8197	0.9771	0.8915	0.9672
	Top 5	0.9143	0.8873	0.9492	0.9172	0.9740
	Top 6	0.8885	0.8264	0.9838	0.8983	0.9727
LpAT	Top 2	0.9163	0.8893	0.9510	0.9191	0.9769
	Top 3	0.9050	0.8689	0.9541	0.9095	0.9705
	Top 4	0.9084	0.8594	0.9766	0.9143	0.9800
	Top 5	0.9101	0.8671	0.9688	0.9151	0.9773
	Top 6	0.8921	0.8395	0.9698	0.8999	0.9595
MCE	Top 2	0.8768	0.8154	0.9742	0.8878	0.9634
	Top 3	0.8840	0.8230	0.9786	0.8941	0.9692
	Top 4	0.8928	0.8409	0.9691	0.9005	0.9630
	Top 5	0.8889	0.8338	0.9714	0.8974	0.9714
	Top 6	0.8994	0.8517	0.9674	0.9059	0.9735

TABLE VI. TRANSFORMER FUSION RESUTLS FOR TEST SPLIT OF PCAM

Set of Models	*Rank*	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Loss	Top 2	0.9044	0.8528	0.9777	0.9110	0.9781
	Top 3	0.8898	0.8374	0.9677	0.8978	0.9707
	Top 4	0.8756	0.8163	0.9695	0.8863	0.9692
	Top 5	0.9045	0.8548	0.9747	0.9108	0.9747
	Top 6	0.9027	0.8515	0.9757	0.9093	0.9729
LpAT	Top 2	0.9177	0.8795	0.9682	0.9217	0.9787
	Top 3	0.9180	0.8841	0.9622	0.9215	0.9773
	Top 4	0.9095	0.8593	0.9795	0.9155	0.9792
	Top 5	0.9040	0.8534	0.9758	0.9105	0.9768
	Top 6	0.8973	0.8436	0.9755	0.9048	0.9726
MCE	Top 2	0.8873	0.8365	0.9629	0.8952	0.9653
	Top 3	0.8694	0.8034	0.9785	0.8823	0.9619
	Top 4	0.9067	0.8654	0.9633	0.9117	0.9740
	Top 5	0.9111	0.8706	0.9660	0.9158	0.9750
	Top 6	0.9239	0.9045	0.9478	0.9257	0.9771

TABLE VII. WEIGHTED ENSEMBLE RESUTLS FOR TEST SPLIT OF PCAM

Set of Models	*Rank*	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
Loss	Top 2	0.9132	0.8738	0.9660	0.9176	0.9776
	Top 3	0.9112	0.8656	0.9736	0.9164	0.9785
	Top 4	0.9104	0.8648	0.9729	0.9157	0.9787
	Top 5	0.9155	0.8712	0.9753	0.9203	0.9799
	Top 6	0.9120	0.8654	0.9760	0.9174	0.9797
LpAT	Top 2	0.9132	0.8738	0.9660	0.9176	0.9776

Set of Models	*Rank*	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
	Top 3	0.9140	0.8733	0.9686	0.9185	0.9779
	Top 4	0.9133	0.8692	0.9732	0.9183	0.9787
	Top 5	0.9129	0.8677	0.9744	0.9180	0.9772
	Top 6	0.9143	0.8689	0.9758	0.9193	0.9780
MCE	Top 2	0.8894	0.8522	0.9424	0.8950	0.9581
	Top 3	0.8896	0.8453	0.9539	0.8964	0.9599
	Top 4	0.9013	0.8575	0.9625	0.9070	0.9710
	Top 5	0.9016	0.8524	0.9715	0.9081	0.9735
	Top 6	0.9018	0.8539	0.9696	0.9080	0.9740

TABLE VIII. COMPARISON OUR MODELS* WITH STATE-OF-THE-ART

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
EfficientNet B5 *	0.8638	0.7993	0.9718	0.8771	0.9637
Resnet101 *	0.9092	0.8724	0.9587	0.9135	0.9747
Swin_t *	0.9251	0.8973	0.9601	0.9277	0.9734
Top6-loss-MLP *	0.8885	0.8264	0.9838	0.8983	0.9727
Top4-LpAT-MLP *	0.9084	0.8594	0.9766	0.9143	0.9800
Top6-MCE-Transformer *	0.9239	0.9045	0.9478	0.9257	0.9771
Custom CNN [9]	0.9800	-	-	-	0.9600
SCMA-CNN [11]	0.9367	-	-	-	0.9400
LGVIT [4]	0.9181	0.9217	0.9180	0.9179	-
CSAM-ResNe [10]	0.9110	-	-	-	0.9720
P4M-DenseNet [3]	0.8980	-	-	-	0.9630
Optimized VGG [7]	0.8600	-	-	-	0.9100

REFERENCES

[1] P. J. Ho, A. R. Cook, N. K. Binte Mohamed Ri, J. Liu, J. Li, and M. Hartman, "Impact of delayed treatment in women diagnosed with breast cancer: A population-based study," *Cancer Med.*, vol. 9, no. 7, pp. 2435–2444, Apr. 2020, doi: 10.1002/cam4.2830.

[2] B. Ehteshami Bejnordi, M. Veta, P. J. van Diest, et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," JAMA, vol. 318, no. 22, pp. 2199–2210, Dec. 2017, doi: 10.1001/jama.2017.14585.

[3] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," in Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), vol. 11071, Lecture Notes in Computer Science, 2018, pp. 210–218, doi: 10.1007/978-3-030-00934-2_24.

[4] L. Wang, J. Liu, P. Jiang, D. Cao, and B. Pang, "LGViT: Local-global vision transformer for breast cancer histopathological image classification," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094951.

[5] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[6] H. Lin et al., "ScanNet: A fast and dense scanning framework for metastatic breast cancer detection from whole-slide images," arXiv preprint arXiv:1703.02442, 2017.

[7] N. Aparicio, I. D. Chacón, and D. Ruiz, "CNN for breast cancer metastases classification," in Proc. IEEE 2nd Int. Congr. Biomed. Eng. Bioeng. (CI-IB&BI), Bogota D.C., Colombia, 2021, pp. 1–4, doi: 10.1109/CI-IBBI54220.2021.9626116.

[8] Y. Liang et al., "Metastatic breast cancer recognition in histopathology images using convolutional neural network with attention mechanism," in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), San Diego, CA, USA, 2019, pp. 1060–1063, doi: 10.1109/BIBM47256.2019.8983058.

[9] M. Ahmad et al., "Classification and detection of cancer in histopathologic scans of lymph node sections using convolutional neural network," Neural Process. Lett., vol. 55, no. 4, pp. 3763–3778, Aug. 2023, doi: 10.1007/s11063-022-10928-0.

[10] R. Shi, X. Li, and H. Gao, "A novel deep learning model CSAM-ResNet based on attention mechanism and residual learning for breast histopathology image classification," in Proc. 27th Int. Conf. Comput. Supported Cooperative Work Design (CSCWD), Tianjin, China, 2024, pp. 3042–3047, doi: 10.1109/CSCWD61410.2024.10580457.

[11] S. B. Mukadam and H. Y. Patil, "Fusion of ESRGAN, adaptive NSCT, and multi-attention CNN with wavelet transform for histopathological image classification," IEEE Access, vol. 12, pp. 129977–129993, 2024, doi: 10.1109/ACCESS.2024.3457315.

[12] A. K. Jaiswal et al., "Semi-supervised learning for cancer detection of lymph node metastases," arXiv preprint arXiv:1912.01188, 2019.

[13] C. Rane et al., "A novel attention fusion network-based framework to ensemble the predictions of CNNs for lymph node metastasis detection," J. Supercomput., vol. 77, no. 12, pp. 13813–13831, Dec. 2021, doi: 10.1007/s11227-021-03823-3.

[14] R. Ge et al., "Detection of presence or absence of metastasis in WSI patches of breast cancer using the dual-enhanced convolutional ensemble neural network," Mach. Learn. Appl., vol. 15, Art. no. 100521, Mar. 2024, doi: 10.1016/j.mlwa.2024.100521.

Comparative Evaluation of Neural Network Architectures for ADHD Classification Using EEG Signals

Eid Wendel Cavalcanti Nunes
Programa de Pós-Graduação em Engenharia Elétrica
Instituto Federal da Paraíba (IFPB)
João Pessoa, Brazil
eid.cavalcanti@academico.ifpb.edu.br

Iury Ferreira de Araújo
Programa de Pós-Graduação em Engenharia Elétrica
Instituto Federal da Paraíba (IFPB)
João Pessoa, Brazil
ferreira.iury@academico.ifpb.edu.br

Vinícius José Wanderley Vieira
Programa de Pós-Graduação em Engenharia Elétrica
Instituto Federal da Paraíba (IFPB)
João Pessoa, Brazil
vieira.vinicius@academico.ifpb.edu.br

Carlos Danilo Miranda Regis
Programa de Pós-Graduação em Engenharia Elétrica
Instituto Federal da Paraíba (IFPB)
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—This study presents a comparative evaluation of three neural network architectures (EEGNet, ResNet, and VGG16) for the automatic classification of ADHD, using raw EEG signals from a public dataset comprising 121 participants. The proposed end-to-end pipeline was developed, integrating per-patient preprocessing, windowing, and data augmentation. The results, evaluated based on accuracy, AUC, and channel importance analysis, show that EEGNet stood out as the most viable and robust architecture, demonstrating stable training behavior and consistently high performance (with accuracy and AUC remaining above 80%) across different windowing strategies. In contrast, ResNet and VGG16 suffered from severe overfitting, especially with larger windows. These findings highlight the efficiency of compact models such as EEGNet for ADHD classification based on EEG signals under participant-held-out internal evaluation.**

***Index Terms*—EEG, ADHD classification, Neural Networks, EEGNet**

I. Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental disorders in childhood, affecting children and adolescents across different social and educational contexts [1]. It is characterized by persistent symptoms of inattention, hyperactivity, and impulsivity, which can impair academic performance, executive functions, and social interactions [2]. In this context, early diagnosis is important to enable appropriate interventions and reduce the disorder's impact throughout development [3].

Currently, ADHD is diagnosed primarily through clinical evaluations based on criteria established by reference manuals, such as the DSM-5 and the ICD-10 [4]. This process involves behavioral observations, clinical interviews, and questionnaires administered to parents, teachers, and patients. Although widely used in clinical practice, this model relies on professionals' interpretation and can be influenced by the heterogeneity of symptoms and the presence of comorbidities, making the search for more objective complementary measures a recurring theme in the literature [4], [5].

Among the techniques investigated for this purpose, electroencephalography (EEG) has received significant attention. EEG records brain electrical activity using electrodes placed on the scalp, allowing for noninvasive monitoring of neural dynamics [6]. One of its main characteristics is its high temporal resolution, which makes it possible to track variations in brain activity on the millisecond scale; the signals obtained can be analyzed in both the time domain and the frequency domain and are often represented by spectral bands such as delta, theta, alpha, beta, and gamma [2], [5].

Several studies report differences in the electrophysiological patterns observed in individuals with ADHD when compared to control groups [7]. Changes in the power distribution across frequency bands and in other EEG-derived indicators suggest that these signals may contain information relevant to characterizing the disorder; for this reason, the EEG has been explored as a data source for computational systems to aid in diagnosis [6].

At the same time, advances in machine learning and deep learning techniques have expanded the capacity for automatic analysis of biomedical signals. In the case of EEG, different neural network architectures have been employed to identify patterns associated with ADHD and to aid in classification

To the Graduate Program in Electrical Engineering (PPGEE) and the Federal Institute of Para´ıba (IFPB) for their financial support

tasks [8].

However, modeling these signals remains a challenge due to their nonstationary nature, the presence of noise and physiological artifacts, and high inter-individual variability [1]. Furthermore, results reported in the literature can vary considerably depending on the architecture used, the preprocessing strategies adopted, and the characteristics of the analyzed dataset [1]–[4], [7], [8] .

Given this scenario, it is important to investigate the performance of different neural network architectures in classifying individuals with ADHD based on EEG signals. Thus, this study conducts a comparative evaluation of three neural networks using a public EEG dataset, analyzing their ability to identify patterns associated with the disorder and their performance in the automatic classification task.

II. Related Works

Studies on ADHD classification based on EEG signals predominantly use databases consisting of children and adolescents, obtained from both public repositories and specific clinical data collections. Although acquisition protocols vary between resting-state recordings and cognitive tasks, there is relative convergence in the preprocessing steps employed. In general, the signals undergo filtering to remove noise, reduce artifacts, and segment them into time windows for analysis. In addition, frequency-domain or time-frequency representations are frequently employed to highlight relevant patterns of brain activity prior to the classification stage [3], [4], [6], [8].

Convolutional Neural Networks (CNNs) are among the most widely used approaches for ADHD classification based on EEG signals, and are applied to both time series and representations derived from the signals. Initially, deep architectures adapted from the field of computer vision—such as variants inspired by VGG and ResNet—were explored due to their capacity for hierarchical learning. However, the limited availability of data in clinical applications has driven the development of more compact and specialized architectures for EEG, such as EEGNet, designed to reduce computational complexity while maintaining good classification performance [2], [4], [6], [7], [9].

In addition to purely convolutional architectures, the literature also presents hybrid models that combine CNNs with recurrent or attention mechanisms to simultaneously explore the spatial and temporal relationships of brain signals. Although different strategies have been proposed, research in this field highlights the ongoing challenge of balancing representational capacity, computational cost, and performance on often small datasets [1], [4], [6].

The limited amount of available data has led to the adoption of various data augmentation strategies in EEG applications. Among the most common approaches are the segmentation of signals into multiple time windows, class balancing techniques, and transformations applied directly to the signals during training. These strategies aim to increase the variability observed by the models and reduce overfitting issues, especially in scenarios where the number of participants is small and inter-subject variability is high [6], [10].

In addition to more traditional techniques, some studies explore approaches based on regularization and representation sharing during training, seeking to improve the models' generalization ability without the need for large volumes of additional data. In general, the literature has prioritized relatively simple methods for data augmentation and class balancing, which can be incorporated into existing processing workflows. Despite the advances made, generating samples capable of adequately preserving the neurophysiological characteristics of EEG signals remains a significant challenge, especially when seeking to increase data diversity without compromising the patterns present in the original signals [7], [8], [10], [11].

Despite the growing adoption of neural networks for ADHD classification based on EEG signals, direct comparisons between the proposed models remain limited. The literature presents a wide variety of architectures, including conventional CNNs, specialized EEG models, residual networks, and hybrid approaches. However, the reported results are often obtained from distinct datasets, using different preprocessing strategies, signal segmentation, data partitioning, and validation protocols. This heterogeneity makes it difficult to identify the factors that effectively contribute to model performance and limits objective comparison among the proposed architectures [1], [3], [4], [7], [8].

Furthermore, although compact architectures and deeper models have been extensively investigated, there is still no consensus on the optimal balance between representational capacity, computational cost, and generalization for applications involving EEG data from patients with ADHD. Thus, it remains important to conduct comparative studies under standardized experimental conditions, allowing for a more consistent evaluation of the behavior of different neural architectures when subjected to the same dataset and the same processing steps [4], [7]–[9].

To consolidate the state of the art that will serve as the basis for the performance discussions presented in this paper, Table I summarizes the approaches that will serve as a reference for a comparative evaluation of the results. Among traditional machine learning methods, the use of AdaBoost-based models operating on the same dataset [12] and analyses of regional hemispheric contributions for feature selection [13] stand out. In the context of hybrid architectures, recent research proposes coupling signal decompositions using VMD (Variational Mode Decomposition) and RLMD (Robust Local Mean Decomposition) [14], with MI (*Mutual Information*) integrated with stacked classifiers [15], and the extraction of feature maps from CNNs in conjunction with classical machine learning models [16]. Based on this, the methodological parameters and performance metrics from these references will be used to analyze the contrasts with the architectures proposed in this paper. However, direct performance comparisons between studies should be made with caution due to methodological differences. In particular, the way data are partitioned (by window or by participant) drastically affects the results.

TABLE I
PERFORMANCE COMPARISON OF ADHD DETECTION.

Authors / Year	C. & Choi [15]	Esas & L. [14]	C. & Choi [13]	TaghiBeyglou [16]	Parashar [12]
Model / Classif.	Stacked Classifier (Meta-MLP)	Hybrid VMD/RLMD + Deep Learning	Naive Bayes (Right Hemis.)	CNN + Logistic Regression	AdaBoost (Right Hemis.)
Accuracy	92.00%	95.76%	84.00%	95.83%	84.00%
AUC	–	–	85.70%	96.00%	–
Sensit.	93.00%	96.00%	–	–	96.00%
Specif.	93.00%	94.00%	–	–	70.00%
Segm.	1024 samples (8 s)	9250 samples (72 s)	1024 samples (8 s)	15360 samples (30 s)	–

III. METHODOLOGY

The development of the ADHD classification system based on EEG signals was carried out entirely in a computer simulation environment, using three neural network architectures (EEGNet, ResNet, and VGG16). The choice of these distinct architectures is justified by the need to evaluate the trade-off between generalization ability and the risk of overfitting on small datasets.

Fig. 1 illustrates the complete flowchart, consisting of six sequential steps for building the neural networks. The workflow begins with data acquisition and loading, followed by preprocessing per patient, overlapping windowing, splitting into training and validation sets, stochastic data augmentation in the training set, and, finally, model training and evaluation.

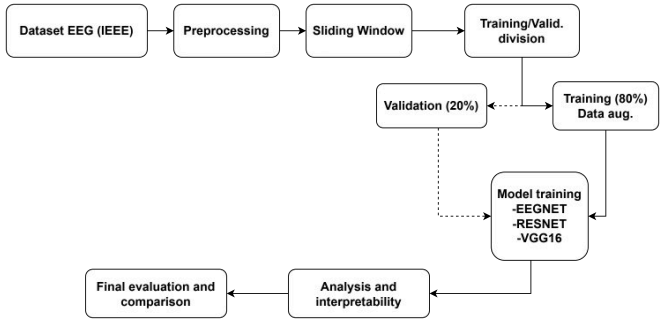


Fig. 1. Proposed EEG classification pipeline.

A. Dataset

The experiments were conducted using the public EEG dataset introduced by Nasrabadi et al. [17], collected at Shahed University and made available through IEEE Open Access. The dataset comprises electroencephalographic (EEG) recordings from 121 participants, including 61 children diagnosed with ADHD and 60 healthy control subjects, aged between 7 and 12 years.

EEG signals were acquired using 19 scalp electrodes positioned according to the international 10–20 system (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, and Pz), with a sampling frequency of 128 Hz. The reference electrodes were placed on the earlobes (A1 and A2).

The ADHD diagnosis was established by clinical specialists according to the diagnostic criteria of the DSM-IV, while the control group consisted of participants with no history of neurological or psychiatric disorders. During data acquisition, participants performed a visual attention task in which they were instructed to count specific elements presented in a sequence of images. This protocol was designed to engage attentional processes and to highlight EEG patterns associated with ADHD.

Each recording includes the EEG signals from all 19 channels, together with the participant identifier and the corresponding class label (ADHD or control).

B. Signal Preprocessing

Each patient record undergoes a filtering and normalization chain applied individually, preventing data leakage between patients.

During the filtering stage, a second-order IIR notch filter centered at 50 Hz is applied to suppress power line noise. Next, a fourth-order Butterworth bandpass filter with cutoff frequencies $f_{\text{low}} = 0.5\,\text{Hz}$ and $f_{\text{high}} = 40\,\text{Hz}$ isolates the clinically relevant EEG rhythms (delta, theta, alpha, and beta), discarding motion artifacts and high-frequency components. Both filters are applied with zero-phase to preserve temporal information.

After filtering, the signal for each patient p is normalized channel by channel using the z-score:

$$\hat{x}_c^{(p)} = \frac{x_c^{(p)} - \mu_c^{(p)}}{\sigma_c^{(p)} + \epsilon}, \tag{1}$$

where $\mu_c^{(p)}$ and $\sigma_c^{(p)}$ are the mean and standard deviation of channel c, calculated exclusively from the samples of that patient, and $\epsilon = 10^{-8}$ prevents division by zero.

Because the normalization parameters are estimated using the complete recording of each participant before window segmentation, the proposed preprocessing pipeline should be interpreted as an offline classification procedure. In a real-time implementation, normalization would need to be performed using only the signal samples available up to the current time, for example through adaptive or causal normalization strategies.

C. Windowing

Continuing the workflow, the next step consisted of segmenting each patient's continuous signal into windows of lengths W equal to 256, 512, 768, and 1024 samples (2 s, 4 s, 6 s, and 8 s at 128 Hz), with the step size calculated as:

$$\text{step} = \lfloor W(1-r) \rfloor, \tag{2}$$

corresponding to an overlap of $r = 50\%$. Each window produces an input tensor $\mathbf{X} \in \mathbb{R}^{C \times W}$, with $C = 19$ channels and W time points. An extra channel of unit dimension is added before training, resulting in $\mathbf{X} \in \mathbb{R}^{1 \times C \times W}$, consistent with the 2-D input convention for convolutional networks.

In this process, segments of the continuous signal at the end of each recording that were shorter than the W window were discarded, avoiding incomplete signals in the model's input. Although the total duration of each patient's EEG recordings

varied, due to differences in the completion time of the visual attention task, windowing standardizes the temporal dimension of all inputs.

D. Training/Validation Split

The data is split according to patient identity using group split, ensuring that all windows from the same individual belong exclusively to either the training set or the validation set. The ratio adopted is 80% training and 20% validation. This prevents windows from the same participant from appearing in both subsets and provides an internal estimate of performance on participants held out from training.

The validation subset was also used for early stopping and model selection. Therefore, the reported results represent an internal development estimate rather than performance on an independent external validation set.

To address class imbalance at the window level, a weight inversely proportional to the frequency of each class in the training set is assigned to each window, ensuring that each mini-batch contains balanced proportions of ADHD and Control subjects.

E. Data Augmentation

Once the group division and class balancing have been defined, data augmentation is applied to the training set at each epoch. The goal is to artificially expand the diversity of the training set, exposing the model to possible variations in the signal that reflect the natural variability of the EEG across sessions and individuals, without the need to collect new recordings.

The transformations are applied stochastically at each epoch; thus, with each pass through the training set, each window may be altered differently, preventing the model from memorizing fixed examples. There are four transformations in total, each of which is applied independently with a 50% probability. The first consists of signal inversion, which reverses the polarity of the signal within the window; since the absolute polarity of the EEG depends on the electrode reference used during acquisition, this operation does not compromise the diagnostic information contained in the signal. Next is the technique of altering the amplitude scale, in which the amplitude of the window is multiplied by a random value between 0.8 and 1.2, simulating the variations in electrode impedance and amplifier gain that naturally occur between different data collection sessions.The third technique is temporal shifting, which consists of a cyclic shift of up to 50 samples to the left or right, performed without zero-padding. Finally, temporal reversal inverts the temporal sequence of the window; since the analyzed brain rhythms do not exhibit significant temporal asymmetry on the order of seconds, the reversed window remains physiologically acceptable as a training example. Figure 2 shows a visual example of the application of these transformations:

With the dataset properly balanced and enhanced by the transformations, the resulting samples are used to train the neural network architectures.

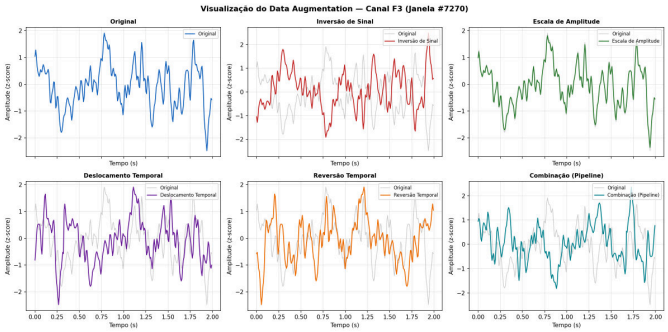


Fig. 2. Data augmentation visualization.

F. Neural Network Architectures

As mentioned earlier, the three neural networks were trained under the same conditions, allowing for a direct comparison between representational capacity and the tendency toward overfitting.

1) EEGNet: The first architecture used was EEGNet, a compact network proposed specifically for EEG signal classification based on depthwise separable convolutions. The initial block of this network applies a 1×64 temporal convolution with $F_1 = 16$ filters, followed by a 19×1 depthwise spatial convolution with depth $D = 2$, which learns one spatial filter per input channel, totaling $F_1 \times D = 32$ feature maps. The second block applies a separable convolution to refine the features extracted by the preceding block. Finally, batch normalization, ELU, average pooling, and progressive dropout (0.25/0.50) complement the blocks, with classification performed by a dense layer of 64 neurons, producing the output for the two classes (ADHD/Control).

2) ResNet: The second network implemented is ResNet, a 1-D residual network of intermediate complexity, designed with Squeeze-and-Excitation (SE) attention blocks. This architecture was constructed using the principles of residual learning and modified to process EEG time series. Initially, a spatial convolution merges the 19 EEG channels into 64 1-D feature maps. Subsequently, the signal passes through four stages, each consisting of two residual blocks with decreasing kernel sizes and increasing numbers of filters (kernels 7, 7, 5, 3; filters $64 \rightarrow 64 \rightarrow 128 \rightarrow 128 \rightarrow 256$), interspersed with average pooling operations that progressively reduce the temporal resolution. At the end of the four stages, adaptive Global Average Pooling (GAP) reduces the temporal dimension to four positions, followed by two fully connected layers (256 and 64 neurons) produces the classification.

3) VGG16: The third and final neural network was a custom 1-D adaptation of VGG16. The original VGG16 architecture was developed for the classification of two-dimensional images and was not specifically designed for EEG signals. In this study, its general structure of convolutional blocks was maintained the classic depth of five convolutional blocks with 2, 2, 3, 3, and 3 convolutions of kernel size 3, respectively, and an increasing number of filters ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 512$) while its operations were modified for processing

multichannel EEG time series. As in ResNet, an initial spatial convolution merges the 19 EEG channels before they enter the 1-D convolutional blocks. Due to the high number of parameters in this type of network, regularization measures were adopted to mitigate overfitting, such as a dropout rate of 0.6 in all convolutional blocks, the use of GAP instead of direct flattening before the classifier, a learning rate reduced by half, and a weight decay rate quadrupled compared to other architectures.

G. Training and Optimization

The three architectures were trained using the AdamW optimizer with a learning rate of 10^{-4}, reduced to 5×10^{-5} for the VGG16 network, a weight decay of $= 5\times10^{-4}$, increased to 2×10^{-3} for the VGG16, and cross-entropy with label smoothing of 0.1. The learning rate followed a linear warmup for 15 epochs, followed by cosine annealing until epoch 150, to avoid abrupt updates at the beginning of training. Each network was trained for up to 150 epochs, with early stopping after 25 epochs if there was no improvement in the AUC on the validation set.

H. Evaluation Metrics

Performance was evaluated on the validation set using accuracy, AUC-ROC, and F1-score per class, supplemented by confusion matrices, the number of parameters in each model, gradient-based channel importance—which measures the sensitivity of the model's output to perturbations in each input channel—and by channel permutation, which assesses the drop in performance when randomly shuffling the values of a specific channel, relating complexity and performance.

IV. RESULTS AND DISCUSSION

We will present the performance results of the EEGNet, ResNet, and VGG16 networks, discussing the impact on classifier metrics for different time window lengths of the EEG signal and evaluating the convergence and biological interpretability of the EEGNet model, which showed the best performance.

A. Windowing

Table II shows the average of each metric calculated at the end of each network's execution using different windowing strategies. Regarding the number of parameters, EEGNet varies proportionally to the size of the input vector; thus, larger windows generate more points. Meanwhile, the ResNet and VGG16 models exhibit a stable number of parameters due to GAP, which prevents variation in the number of parameters despite changes in windowing.

This information can be crucial for identifying overfitting in complex models that operate on small datasets; in the literature, highly parameterized architectures using long windows on clinical datasets are indicative of vulnerability. According to the literature [16], it is necessary to balance the model's capacity with the volume of samples. The VGG16 model serves as a specific example, achieving satisfactory results with 2-second windows (76.66% accuracy) but failing with 4-, 6- and 8-second windows (39.07%, 41.69% and 38.61% accuracy, respectively) due to its high number of parameters combined with a small number of samples.

TABLE II
PERFORMANCE COMPARISON ACROSS DIFFERENT WINDOW LENGTHS.

Win.	Model	Params	Acc.	AUC	F1-score	
					ADHD	Control
2 s	EEGNet	19.9k	80.72%	83%	85%	74%
	ResNet	1.91M	81.48%	86%	85%	76%
	VGG16	5.07M	76.66%	82%	81%	70%
4 s	EEGNet	36.3k	81.55%	87%	85%	76%
	ResNet	1.91M	79.08%	83%	83%	72%
	VGG16	5.07M	39.07%	81%	1%	56%
6 s	EEGNet	52.6k	80.80%	88%	84%	76%
	ResNet	1.91M	64.27%	82%	64%	65%
	VGG16	5.07M	41.69%	83%	9%	57%
8 s	EEGNet	69.06k	80.94%	83%	85%	74%
	ResNet	1.91M	66.43%	83%	67%	65%
	VGG16	5.07M	38.61%	78%	0%	56%

In the study conducted by Chauhan and Choi [15], an 8-second signal segmentation is used, requiring tools such as PCA (Principal Component Analysis) to reduce dimensionality and avoid overfitting. Therefore, the use of 2-second windows in this study effectively manages this computational complexity, ensuring sufficient data so that high-capacity networks—which depend on data density—can converge. For this reason, the following figures will present the results for this windowing scheme.

Despite the change in windowing, the AUC metric remained stable at above 80% across all configurations. Although the accuracy of VGG16 declines with the use of larger windows, the AUC reinforces the ability of these architectures to distinguish between classes. Based on the literature, this metric is essential for evaluating the classification performance of the model and, in benchmark studies, presents values between 86% and 96% [16]. In line with this perspective, EEGNet demonstrated a considerable increase in performance with larger windowing, supporting the view that an increase in patterns and biomarkers aids in feature separation.

B. Model Evaluation

When evaluating network performance, it is necessary to examine the behavior of all windows. By comparing the number of parameters and the AUC values shown in Table II, it is clear that using too many or too few parameters can affect the ability to distinguish between classes—that is, it affects accuracy. Thus, although the ResNet network outperforms the others with an accuracy of 81.58% for a 2-second windowing, the EEGNet model stands out, demonstrating a robust architecture that provides relatively high AUC and accuracy (both metrics above 80%) for different windowing strategies.

The literature review allowed us to contextualize these results, noting that the 80.72% accuracy achieved by EEGNet

proves to be competitive with approaches and models based on AdaBoost (Adaptive Boosting), which reach 84% accuracy using the same dataset [12].

Although recent studies report significantly higher results—such as the use of stacked classifier architectures with 92% accuracy [15] and hybrid architectures like VMD and RLMD with results exceeding 95% [14], the model proposed in this work demonstrates a methodological advantage. The approach involves applying the end-to-end method to the raw EEG signal, eliminating the need for manual feature extraction and mathematical decompositions, which increase computational cost. Furthermore, this study employs a strict data split (80/20) by patient rather than by window, preventing the results from being affected by data leakage. Thus, the results achieved by the EEGNet model (80.72%) demonstrate a generalization to patients not seen within the evaluated validation set, raising its methodological robustness to the results of studies such as [16], which reports 95.83% accuracy.

Focusing on the results of the EGGNet model, Figure 3 illustrates the convergence behavior and the extent to which the gradient stabilization strategies performed adequately. In the work by [16], overfitting is considered one of the most prominent problems in training neural networks, and two strategies are used to mitigate it: early stopping with a patience of 30 epochs and the selection of the model with the best validation accuracy. For comparison, this study implements early stopping and dropout, ensuring that the training (blue line) and validation (orange line) curves follow similar trajectories, as well as providing greater robustness in classifying patients not seen during training. Furthermore, the model checkpoint corresponding to the highest validation AUC was maintained; to make this clearer, Figure 3 highlights the maximum validation accuracy of 80.72% achieved by this checkpoint. This ensures that the final result demonstrates comparable capacity and data density, overcoming the instability of highly parameterized networks.

The training curves show a progressive decrease in training loss accompanied by an increase in training accuracys in Figure 3 indicates convergence when using Linear Warmup and Cosine Annealing, thereby validating the effectiveness of the training protocol adopted in this study.

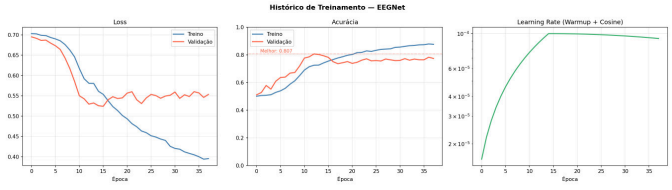


Fig. 3. EEGNET training history.

Figure 4 shows that the ResNet network achieves the best balance between sensitivity and specificity, while maintaining a high ADHD accuracy rate above 80%. EEGNet performs similarly to ResNet even though it uses fewer parameters, demonstrating robust discriminatory power. Meanwhile, VGG16, which has the lowest AUC among the other networks, shows a curve that falls below the others due to a lack of sufficient data for the classifier.

In comparison with the literature, [16] presents comparative tables using AUC as a key metric for analyzing hybrid models, achieving 86% in the base CNN architecture and 96% when combined with logistic regression. Thus, it is noted that EEGNet in this study was the most reliable model due to its preservation of separability between classes.

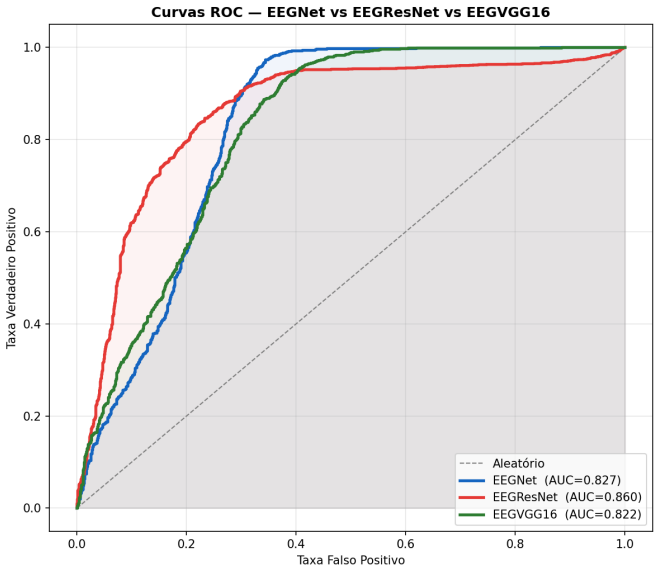


Fig. 4. ROC curve and Area Under the Curve (AUC) for EEGNet, ResNet and VGG16.

C. Classification

Upon analyzing Figure 5, the EEGNet model proved to be the most balanced among the networks in this study, achieving an accuracy rate for the Control class (TN — True Negative) of 71.6% (958 windows) and maintaining a high accuracy rate for ADHD (TP — True Positive) at 86.5% (1,822 windows). These results are reflected in the F1 scores obtained for the two classes, as shown in Table II, indicating a reduction in bias toward the majority class (ADHD).

When comparing these results with the literature, the work by [16] demonstrates an approach with an average accuracy of 86.33% using a basic CNN architecture. On the other hand, the work by [15] demonstrates that approaches using architectures that combine multiple different models (stacked classifiers) can achieve 92% accuracy, with 93% specificity, indicating a high ability to distinguish healthy patients. Therefore, we conclude that the EEGNet architecture offers significant competitive discriminatory power in identifying healthy cases, with a low incidence of false positive diagnoses.

D. Biological Relevance of the Model

To mitigate the “black box” effect and validate the EEGNet model using ADHD biomarkers, gradient-based analysis and

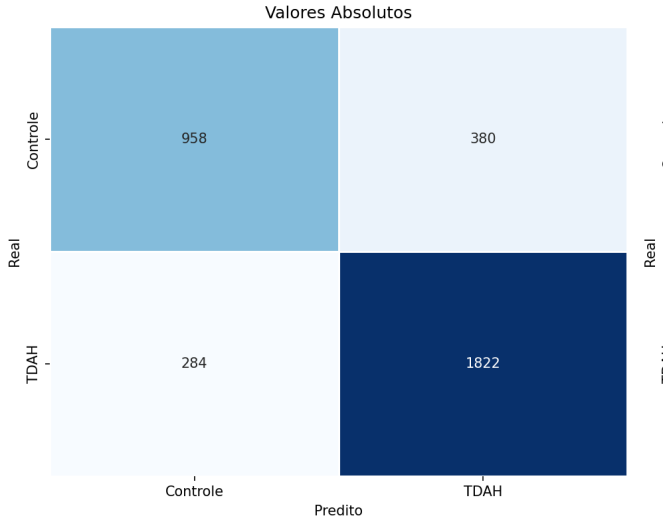


Fig. 5. EEGNET confusion matrix mapping classification for the 2 s window topology.

Permutation Feature Importance (PFI) will be applied. According to Figure 6, the EEGNet model identifies the Fz channel in the frontal region as the most relevant, while also showing that the frontal region (F3) is particularly prominent, as illustrated in Figure 7.

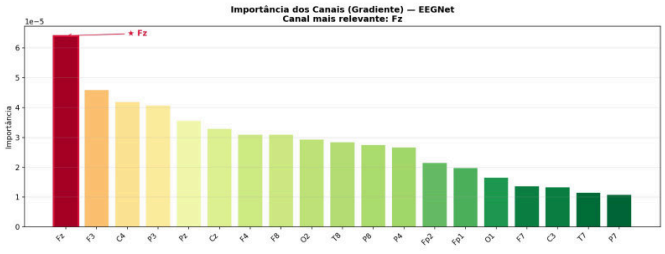


Fig. 6. Channel importance based on the gradient method for the EEGNet model (2 s window).

Based on a review of the literature, the results presented in this study tend to align with studies involving regional mapping and channel relevance. In this context [15], the MI technique was used to measure the dependence between two variables, concluding that there was no consistent concentration of ADHD at a single point and suggesting that the syndrome is complex because it affects different regions of the brain simultaneously. From another perspective, studies such as [13] and [12] identified the region of the right hemisphere with the highest concentration of discriminatory features, achieving metrics such as 96% sensitivity and 84% accuracy, encompassing the Fp2, F4, F8, T4, P4, T6, and O2 channels. Meanwhile, according to [14], the Fz channel is considered one of the most critical in distinguishing between individuals with ADHD and healthy individuals, due to its association with attention deficits. Thus, it is noted that the channel selected by EEGNet (Fz), obtained through gradient-based analysis,

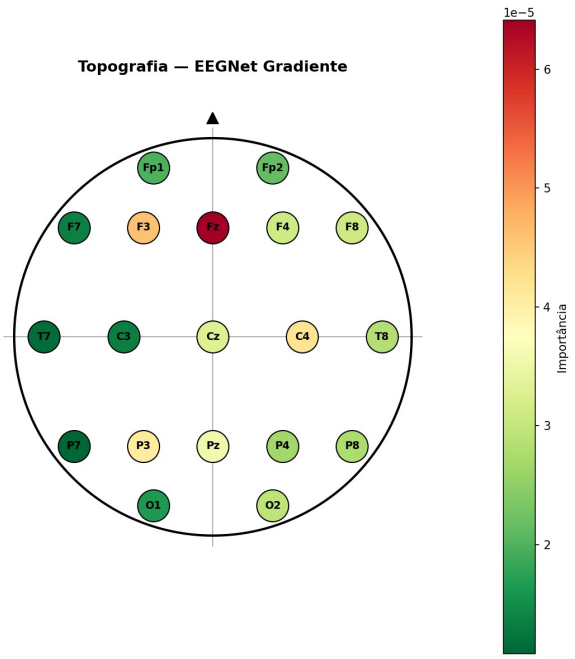


Fig. 7. Topographic relevance maps based on gradients for the EEGNet model.

validates the model, indicating frontal impairment similar to that reported by the authors.

Meanwhile, Figure 8 illustrates the EEGNet results, indicating that the Fz channel is the critical factor in classification; its absence would reduce accuracy by approximately 34% (original accuracy – accuracy with the channel swapped), in addition to the significant impact of the Pz, P3 and F3 channels.

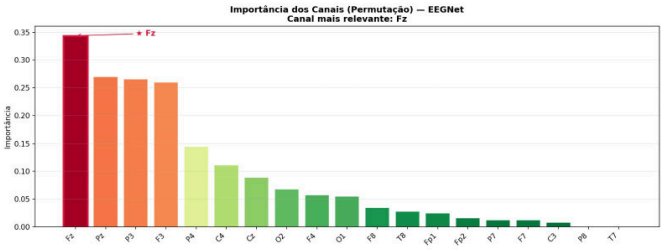


Fig. 8. Channel importance based on the PFI method for the EEGNet model (2 s window).

Using the permutation method, the Fz channel (frontal lobe) was identified as a critical region for classification. This channel supports the findings of [14], which argue that the frontal midline electrodes are the most discriminative in channel-by-channel analysis. From another perspective, the work in [15] points out that the Fz channel reflects activity in the frontal lobe of the brain, which is involved in executive functions, including planning, decision-making, and attention. Furthermore, in studies such as [12] and [13], traditional methods are used, such as manually selecting channels or isolating hemispheres to guide the classifiers, grouping specific regions

to assess their impact on diagnosis. In contrast, the EEGNet model was able to automatically analyze the channels and identified the Fz channel as the critical factor in diagnosis. This finding, obtained through the permutation method, indicates that the trained EEGNet model depends on the Fz channel, representing a decrease in accuracy of approximately 34% in its absence. This internal dependence of the model shows high consistency with the literature, which associates ADHD with executive impairment of the frontal lobe. Furthermore, this method was performed at the window level, and one of the objectives for future work is the implementation of permutation at the participant level.

V. CONCLUSIONS

This study presented a comparative evaluation of three neural network architectures for ADHD classification based on raw EEG signals, using an end-to-end approach applied directly to the signal and a split between training and validation data per patient. EEGNet demonstrated better overall performance due to its ability to handle the EEG signal across different windowing strategies, while ResNet and VGG16 suffered from overfitting in larger windows due to their high number of parameters relative to the limited size of the dataset. Furthermore, EEGNet, despite having approximately 100 times fewer parameters, outperforms ResNet due to its methodological robustness, highlighting the efficiency of compact architectures in scenarios with limited data.

The interpretability analysis of EEGNet highlighted the Fz channel (gradient and permutation), a result consistent with previous studies linking ADHD to the frontal region of the brain. Limitations include the small size of the dataset and performance that is still inferior to that of methods based on manual feature extraction. For future work, we propose testing the same networks and more complex architectures using an offline augmented dataset that combines synthetic and real patients, with an expected improvement in performance.

ACKNOWLEDGMENT

The authors would like to thank the Federal Institute of Paraíba (IFPB) and the Graduate Program in Electrical Engineering (PPGEE) for their institutional support, and the National Council for Scientific and Technological Development (CNPq) for the financial support provided through the Institutional Postgraduate Scholarship Program (PIBPG) – CNPq Public Call No. 12/2025.

REFERENCES

[1] M. Feng and J. Xu, "Electroencephalogram-based convmixer architecture for recognizing attention deficit hyperactivity disorder in children," *Brain Sciences*, vol. 14, no. 5, 2024. [Online]. Available: https://www.mdpi.com/2076-3425/14/5/469

[2] H. Hu, S. Tong, H. Wang, J. Wu, R. Zhang, R. Jiang, Y. Zhao, Y. Ju, and X. Zhang, "Scanet: An innovative multiscale selective channel attention network for eeg-based adhd recognition," *IEEE Sensors Journal*, vol. 25, no. 11, pp. 20 920–20 932, 2025.

[3] S. K. Khare and U. R. Acharya, "An explainable and interpretable model for attention deficit hyperactivity disorder in children using eeg signals," *Computers in Biology and Medicine*, vol. 155, p. 106676, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523001415

[4] C. Wang, X. Wang, X. Jing, H. Yokoi, W. Huang, M. Zhu, S. Chen, and G. Li, "Towards high-accuracy classifying attention-deficit/hyperactivity disorders using cnn-lstm model," *Journal of Neural Engineering*, vol. 19, no. 4, p. 046015, jul 2022. [Online]. Available: https://doi.org/10.1088/1741-2552/ac7f5d

[5] C.-C. Chen, E. H.-K. Wu, Y.-Q. Chen, H.-J. Tsai, C.-R. Chung, and S.-C. Yeh, "Neuronal correlates of task irrelevant distractions enhance the detection of attention deficit/hyperactivity disorder," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1302–1310, 2023.

[6] O. Karabiber Cura and F. G. Yasar Ciklacandir, "Electroencephalography-based classification of individuals with attention deficit hyperactivity disorder using brain connectivity information and deep learning," *Computers and Electrical Engineering*, vol. 135, p. 111190, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790626002624

[7] T. Sindhu and S. Sujatha, "Common spatial pattern based feature extractor with hybrid linknet-squeezenet for adhd detection from eeg signal," *Progress in Engineering Science*, vol. 2, no. 4, p. 100172, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950425225001240

[8] İlkim Ecem Emre, Çiğdem Erol, C. Taş, and N. Tarhan, "Multi-class classification model for psychiatric disorder discrimination," *International Journal of Medical Informatics*, vol. 170, p. 104926, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505622002404

[9] D. Kim, Y. Lee, S. Shin, K. M. Kim, D. Kim, and S.-I. Choi, "Noise-robust and interpretable eeg classification of hyperactive disorders using graph convolutional network," *IEEE Access*, vol. 13, pp. 186 168–186 180, 2025.

[10] Y. Li, Y. Li, R. Nair, and S. M. Naqvi, "A novel audio-visual model with real-time gradient modulation for mental disorder detection," *Biomedical Signal Processing and Control*, vol. 120, p. 110164, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809426007184

[11] H. Dong, D. Chen, Y. Chen, Y. Tang, D. Yin, and X. Li, "A multi-task learning model with reinforcement optimization for asd comorbidity discrimination," *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107865, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016926072300531X

[12] A. Parashar, N. Kalra, J. Singh, and R. K. Goyal, "Machine learning based framework for classification of children with adhd and healthy controls," *Intelligent Automation & Soft Computing*, vol. 28, no. 3, pp. 669–682, 2021. [Online]. Available: http://doi.org/10.32604/iasc.2021.017478

[13] N. Chauhan and B.-J. Choi, "Regional contribution in electrophysiological-based classifications of attention deficit hyperactive disorder (adhd) using machine learning," *Computation*, vol. 11, no. 9, 2023. [Online]. Available: https://www.mdpi.com/2079-3197/11/9/180

[14] M. Y. Esas and F. Latifoğlu, "Detection of adhd from eeg signals using new hybrid decomposition and deep learning techniques," *Journal of Neural Engineering*, vol. 20, no. 3, 2023. [Online]. Available: https://doi.org/10.1088/1741-2552/acc902

[15] N. Chauhan and B.-J. Choi, "A machine learning approach to adhd diagnosis using mutual information and stacked classifiers," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 24, no. 1, pp. 10–18, 2024. [Online]. Available: http://doi.org/10.5391/IJFIS.2024.24.1.10

[16] B. TaghiBeyglou, A. Shahbazi, F. Bagheri, S. Akbarian, and M. Jahed, "Detection of adhd cases using cnn and classical classifiers of raw eeg," *Computer Methods and Programs in Biomedicine Update*, vol. 2, 2022. [Online]. Available: https://doi.org/10.1016/j.cmpbup.2022.100080

[17] A. M. Nasrabadi, A. Allahverdy, M. Samavati, and M. R. Mohammadi, "Eeg data for adhd / control children," 2020. [Online]. Available: https://dx.doi.org/10.21227/rzfh-zn36

Comparative Study of Blind Source Separation (BSS) Techniques for Autism Spectrum Disorder Classification in EEG Signals

1st Jéssica Franzon Cruz do Espírito Santo
Information Engineering
Federal University of ABC (UFABC)
Santo André, SP, Brazil
jessica.franzon@ufabc.edu.br

2nd Caroline Pires Alavez Moraes
Information Engineering
Federal University of ABC (UFABC)
Santo André, SP, Brazil

3rd Aline Neves
Information Engineering
Federal University of ABC (UFABC)
Santo André, SP, Brazil

***Abstract*—Electroencephalographic (EEG) signals are highly susceptible to physiological artifacts and instrumental interference, making preprocessing choices critical for Autism Spectrum Disorder (ASD) classification. This study compared EEG processing pipelines to investigate the influence of Blind Source Separation (BSS) and temporal segmentation strategies on classification performance. A dataset with 28 individuals, 14 ASD and 14 controls was used. Fast Independent Component Analysis (FastICA) and SecondOrder Blind Identification (SOBI) were tested. Temporal, spectral, and nonlinear features were extracted, selected using Minimum Redundancy Maximum Relevance (mRMR), and classified using a Support Vector Machine (SVM). The highest observed performance was obtained with SOBI, reaching an accuracy of 0.741, an area under the receiver operating characteristic curve (AUC) of 0.730, and an F1-score of 0.771. The work also includes an analysis of the features that were more often selected by mRMR. These exploratory findings indicate that source separation, signal segmentation, and feature representation can affect EEG-based ASD classification and motivate further evaluation using participant-wise repeated validation.**

***Index Terms*—autism spectrum disorder; blind source separation; signal classification; electroencephalography.**

I. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition associated with persistent alterations in social communication, restricted interests, and repetitive behavioral patterns [1]. Although clinical assessment remains the reference standard for diagnosis, it usually involves behavioral observation, structured interviews, standardized instruments such as CARS-2, ABC, ADOS-2, CHAT, and ADI-R, and interpretation by experienced professionals [2]. As a consequence, the diagnostic process may be time-consuming and dependent on the availability of specialized services, which can delay early identification and intervention.

Because of these limitations, increasing attention has been directed toward complementary approaches capable of providing objective neurophysiological information. Among them, electroencephalography (EEG) has been widely investigated due to its non-invasive nature, relatively low cost, high temporal resolution, and ability to capture rapid fluctuations in cortical activity [3]. EEG records brain electrical oscillations through electrodes placed on the scalp, commonly arranged according to standardized acquisition systems [4].

In ASD research, EEG has been used to investigate neurophysiological patterns related to spectral organization, regional activation, and the stability of neural rhythms under resting-state and task-based conditions [3]. Previous studies have reported differences in frequency bands such as delta, alpha, and gamma, suggesting that spectral characteristics may provide information relevant to the comparison between individuals with ASD and neurotypical controls [2], [5].

Blind Source Separation (BSS) techniques provide a strategy for decomposing multichannel recordings into latent components, allowing neural and non-neural contributions to be analyzed separately [6]. Among these techniques, Fast Independent Component Analysis (FastICA) estimates components based mainly on statistical independence and non-Gaussianity, whereas second-order Blind Identification (SOBI) exploits second-order statistics and the temporal correlation structure of the signals [7], [8]. Because these methods rely on different statistical assumptions, their use may produce different representations of the EEG signal and consequently affect the classification performance.

In addition to the source separation method, other methodological choices may also affect EEG-based classification results. Previous studies have adopted different preprocessing procedures, feature extraction strategies, dimensionality reduction methods, and classification algorithms [2], [9], [10]. Such methodological variability makes direct comparison between studies difficult and indicates that classification performance should be interpreted considering the complete signal-processing pipeline rather than the final classifier alone.

In this context, the present study investigates the influence of BSS and temporal segmentation strategies on EEG-based classification of individuals with ASD and neurotypical controls. Specifically, FastICA and SOBI were evaluated in combination with three temporal processing conditions: no windowing of the EEG signals, non-overlapping windows, and 50% overlapping windows. Temporal, spectral, and nonlinear features were subsequently extracted, selected using Minimum Redundancy

Maximum Relevance (mRMR), and classified using a Support Vector Machine (SVM). The objective was to compare these methodological configurations and identify the pipeline that achieved the highest classification performance in the analyzed dataset.

The remainder of this paper is organized as follows. Section II presents the theoretical background on Blind Source Separation, including the FastICA and SOBI algorithms evaluated in this study. Section III describes the dataset and methodological pipeline, including preprocessing, artifact identification, segmentation, feature extraction, dimensionality reduction, and classification. Section IV presents and discusses the experimental results. Finally, Section V summarizes the main findings and limitations of the study.

II. Blind Source Separation

Independent Component Analysis (ICA) is a multivariate Blind Source Separation (BSS) technique widely employed in EEG signal processing. Its objective is to decompose the observed signals into statistically independent latent components, allowing neural activity to be separated from physiological and instrumental interferences. In the context of this work, the classical linear model is adopted, in which the observed EEG is described as a linear mixture of independent sources [7].

The model can be expressed as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{x}(t)$ represents the vector of observed signals, $\mathbf{s}(t)$ corresponds to the independent sources, and $\mathbf{A}$ is the mixing matrix. The objective of ICA is to estimate a separation matrix $\mathbf{W}$ such that:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t), \tag{2}$$

so that the components of $\mathbf{y}(t)$ are as mutually independent as possible. By doing so, $\mathbf{y}$ will be an estimate of $\mathbf{s}$. In EEG applications, this decomposition favors the identification and removal of artifacts, contributing to more robust analyses.

A. FastICA

FastICA (*Fast Independent Component Analysis*) is a BSS algorithm based on the maximization of the non-Gaussianity of the estimated sources. This principle follows from the Central Limit Theorem, according to which linear combinations of independent variables tend to present a distribution closer to a Gaussian one. Thus, by searching for maximally non-Gaussian projections, FastICA recovers estimates of the independent components [7].

The measure of non-Gaussianity employed in FastICA is based on negentropy, given by:

$$J(y) = H(y_{\text{gauss}}) - H(y), \tag{3}$$

where $H(\cdot)$ denotes differential entropy, and y_{gauss} is a Gaussian variable with the same mean and variance as y. In practice, the algorithm uses computationally efficient approximations of this measure through nonlinear functions, leading to a fixed-point method recognized for its fast convergence and numerical stability [7].

B. SOBI

SOBI (*Second-Order Blind Identification*) is a BSS algorithm based on the analysis of second-order statistics, with emphasis on the temporal correlations of the signals. Its separation strategy relies on the joint diagonalization of autocorrelation matrices at different time delays, which makes it particularly suitable for EEG signals, whose temporal structure carries relevant information [8].

The SOBI cost function can be expressed as:

$$J(\mathbf{U}) = \sum_{\tau_k \in D} \operatorname{off}\left(\mathbf{U}^{\mathrm{T}} \mathbf{R}_z(\tau_k) \mathbf{U}\right), \tag{4}$$

where $\mathbf{R}_{\mathbf{z}}(\tau_k)$ represents the autocovariance matrix of the whitened signals at time delay τ_k, D denotes the set of time delays considered, $\mathbf{U}$ corresponds to the orthogonal matrix estimated during joint diagonalization, and $\operatorname{off}(\mathbf{M})$ denotes the sum of the squared off-diagonal elements:

$$\operatorname{off}(\mathbf{M}) = \sum_{i \neq j} m_{ij}^2. \tag{5}$$

The complete SOBI separation matrix is given by:

$$\mathbf{W}_{\text{SOBI}} = \mathbf{U}^T \mathbf{V}, \tag{6}$$

where $\mathbf{V}$ represents the whitening matrix, and $\mathbf{U}$ is the orthogonal matrix estimated through the joint diagonalization procedure.

III. Materials and Methods

This section presents the stages that compose the methodological *pipeline* of this study, as illustrated in Figure 1. Each stage of the process is described in detail in the following subsections, in order to clarify the technical decisions adopted and their relevance to the processing and analysis of EEG signals in the context of ASD classification.

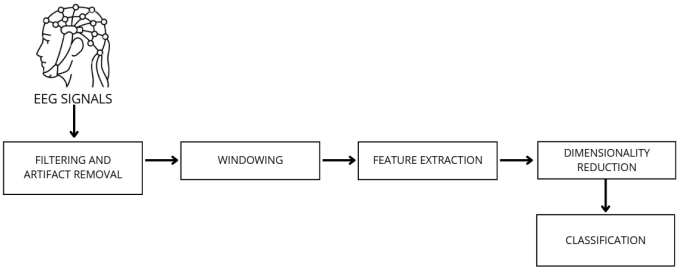


Fig. 1. Overview of the methodological pipeline.

A. Dataset

The EEG signals used in this study were obtained from the Data for Electrophysiological Signatures of Brain Aging in Autism Spectrum Disorder dataset. The original dataset comprises 56 adult participants, including 28 individuals diagnosed with autism spectrum disorder (ASD) and 28 neurotypical controls. The recordings were acquired at the

University of Cambridge using the BioSemi ActiveTwo system with an original 64-electrode montage, under resting-state eyes-closed conditions, with an approximate duration of 2.5 minutes and an original sampling rate of 2048 Hz.

Although the original acquisition protocol employed 64 electrodes, differences in channel availability, naming, and organization were identified across the analyzed files. To ensure identical input dimensionality and direct comparability between participants, all subsequent analyses were restricted to the following 15 channels: AF3, C3, C4, CP4, F3, F4, F7, Fz, O1, O2, P1, P10, P5, P8, and POz.

The final sample comprised 28 participants, including 14 individuals with ASD and 14 neurotypical controls, aged between 36 and 74 years. Participants were included when their EEG recordings contained the complete set of selected channels and when they satisfied the age-matching and group-balancing criteria adopted in the study. The remaining participants did not jointly meet these criteria. Therefore, in the reamining of this work, only 28 participantes were considered.

B. Signal Preprocessing

The preprocessing procedure was applied independently to each participant. First, the 15 common EEG channels were selected, ordered consistently across recordings, and band-pass filtered between 1 and 100 Hz, to preserve the gamma frequency band (from 30Hz to 100Hz) information, which we would also like to exploit in this work. A 50 Hz notch filter was subsequently applied to attenuate power-line interference. The filtered signals were then re-referenced using the common average reference calculated across the 15 retained channels.

After filtering and channel harmonization, the EEG signals were re-referenced using the *Common Average Reference* (CAR) method [11]. CAR was applied independently to each participant using only the 15 channels retained for analysis. At each time sample, the mean electrical potential across these channels was calculated and subtracted from the signal of each individual channel:

$$V_i^{\text{CAR}}(t) = V_i(t) - \frac{1}{N}\sum_{j=1}^{N} V_j(t), \tag{7}$$

where $V_i(t)$ represents the potential recorded at channel i at time t, and $N = 15$ represents the number of channels included in the analysis. Thus, the reference was not calculated using the original 64-channel montage [3], [11].

This procedure ensured that all participants were referenced using the same channel set and channel ordering. CAR was applied before the Blind Source Separation stage and, consequently, before the application of FastICA or SOBI.

C. Artifact Identification and Signal Reconstruction

After filtering and common average referencing, the EEG signals were decomposed using either FastICA or SOBI, according to the evaluated pipeline. The resulting components were subsequently analyzed in EEGLAB using the ICLabel plugin for automatic component classification [12].

ICLabel assigns to each component a probability of belonging to different classes, including Brain, Muscle, Eye, Heart, Line Noise, Channel Noise, and Other. In this study, a probability threshold of 80% was adopted. Components for which ICLabel assigned a probability equal to or greater than 80% to a specific artifact class were considered representative of that category. Components classified with a probability equal to or greater than 80% as Muscle, Heart, Line Noise, or Channel Noise were excluded during signal reconstruction, whereas components classified as Eye were retained.

Artifacts identified through FastICA and SOBI components were recorded separately and used to characterize the distribution of artifact types identified for each source separation method. Ocular components were retained as part of the exploratory methodological protocol but were not considered specific biomarkers of ASD.

The results presented in Fig. 2 were obtained from the ICLabel classifications that satisfied the adopted 80% probability threshold and represent the distribution of non-ocular artifact categories identified for each BSS method.

D. Segmentation and Feature Extraction

In the segmented configurations, the EEG signals were resampled to 256 Hz and divided into fixed-length windows of 60 s, corresponding to 15,360 samples per segment. Two segmentation strategies were evaluated: non-overlapping windows and 50% overlapping windows. In the overlapping condition, the step between consecutive windows was 30 s, corresponding to 7,680 samples. Segmentation was performed after filtering, re-referencing, and artifact removal. A Hann window was subsequently applied to each segment before feature extraction to reduce spectral leakage associated with spectral estimation [13].

In the feature extraction stage, temporal, spectral, and nonlinear features were obtained from the EEG signals to characterize different properties of brain activity. Temporal features included variance, kurtosis, skewness, peak-to-peak amplitude, Root Mean Square (RMS), and Hjorth mobility and complexity parameters [14].

Nonlinear features included the Hurst exponent, Detrended Fluctuation Analysis (DFA), Higuchi fractal dimension, and Petrosian fractal dimension. The Hurst exponent was used to characterize persistence and long-range dependence in the signal, whereas DFA was used to quantify long-range correlations and scaling behavior [15]. Higuchi and Petrosian fractal dimensions were included as measures of temporal signal complexity.

In the frequency domain, the power spectral density (PSD) was estimated using Welch's method. Absolute and relative spectral powers were extracted for the classical EEG frequency bands (δ, θ, α, β, and γ), together with total spectral power. Additional spectral measures included Peak Alpha Frequency (PAF) [16] and Spectral Edge Frequency 95 (SEF95) [17].

Entropy-based features included approximate entropy, permutation entropy, Singular Value Decomposition (SVD) entropy, and spectral entropy. Approximate entropy was computed

directly from the temporal EEG signal using the Python library `AntroPy` implementation and was used to quantify signal regularity and complexity. Permutation entropy was included as an ordinal measure of temporal complexity, whereas SVD entropy was used to characterize the distribution of information across the singular values of the time-delay representation of the signal.

Spectral entropy was computed from the PSD estimated using Welch's method. The PSD was normalized by its total power to obtain a probability distribution over frequencies, from which Shannon entropy was calculated using a base-2 logarithm. For the segmented pipelines, the features were computed separately for each channel and each 60-second window. For the non-segmented pipelines, they were computed separately for each channel using the complete recording of each participant.

E. Dimensionality Reduction

To minimize redundancy among features and reduce the risk of overfitting, the Minimum Redundancy Maximum Relevance (mRMR) method was applied [9]. The method was applied to the extracted feature matrix rather than directly to the raw EEG samples. Each column of the matrix represented a candidate temporal, spectral, or nonlinear feature.

For the segmented pipelines, each observation corresponded to a channel-window pair associated with a participant. For the non-segmented pipelines, each observation corresponded to a channel from the complete recording of a participant. Mutual information was estimated between each candidate feature and the ASD or control class label, while redundancy was evaluated among the candidate features. The resulting feature subset was subsequently used as input to the SVM classifier.

F. Classification and Evaluation

After feature selection, the classification stage was performed using the *Support Vector Machine* (SVM) algorithm, which is widely employed in biomedical signal classification tasks due to its robustness in high-dimensional spaces [9]. The implementation was carried out in Python using the *scikit-learn*, *pandas*, and *NumPy* libraries.

The features were standardized using the `StandardScaler` method, and classification was performed using a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, configured as `SVC(kernel='rbf', C=1, gamma='scale', probability=True)`. The RBF kernel and its hyperparameters were defined *a priori* and maintained unchanged across all evaluated pipelines. No comparison with alternative kernels or systematic hyperparameter optimization was performed. This fixed configuration was adopted to evaluate the effects of the source separation and segmentation strategies under the same classification conditions. The RBF kernel was used to allow nonlinear decision boundaries in the space of the extracted EEG features.

A single stratified hold-out evaluation was adopted, with 70% of the observations assigned to the training set and 30% to the test set. The partition was performed using a fixed random seed (`random_state=42`) to ensure reproducibility while preserving the proportion of ASD and control observations in both subsets.

The performance of the *pipelines* was evaluated using accuracy, precision, *recall*, F1-score, and the area under the receiver operating characteristic curve (AUC). The final comparison considered different combinations of artifact removal methods and signal segmentation strategies, aiming to identify the most consistent methodological configuration for discriminating between individuals with ASD and neurotypical controls.

IV. Results and Discussion

Initially, the artifact classes obtained after FastICA and SOBI decomposition were descriptively compared based on the ICLabel classification. Ocular components were identified but retained during signal reconstruction, according to the methodological procedure described in Section III-C.

Different distributions of artifact classes were observed between the two source separation methods. FastICA produced a higher number of components classified as channel noise, whereas SOBI produced a higher number of components classified as muscle activity. Both methods also yielded a smaller number of components classified as cardiac activity, as illustrated in Fig. 2. These results provide a descriptive comparison of the component classes obtained after each decomposition and should not be interpreted as a formal evaluation of artifact-detection sensitivity.

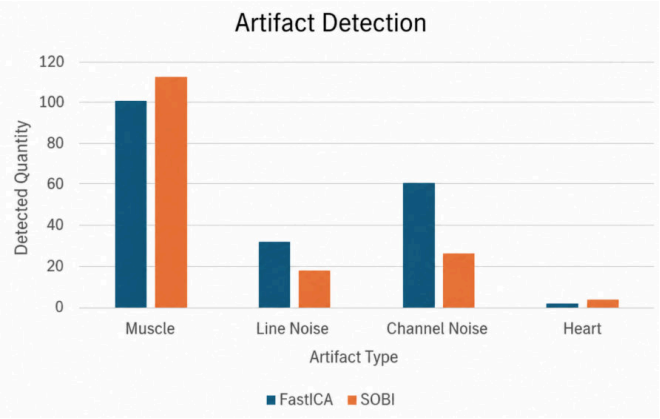


Fig. 2. Distribution of artifact classes identified by ICLabel after FastICA and SOBI decomposition.

The mRMR analysis showed that the most relevant attributes were not restricted to spectral power measures. The 15 highest-ranked features were, respectively, Higuchi fractal dimension, Detrended Fluctuation Analysis (DFA), Petrosian fractal dimension, Hurst exponent, relative beta power, relative alpha power, approximate entropy, peak-to-peak amplitude, total spectral power, absolute beta power, Hjorth complexity, permutation entropy, SVD entropy, spectral entropy, and Hjorth mobility.

The predominance of nonlinear measures among the highest-ranked features suggests that signal complexity, self-similarity,

and long-range temporal organization contributed relevant information to the distinction between ASD and control observations in the analyzed dataset. Higuchi and Petrosian fractal dimensions characterize signal complexity, whereas DFA and the Hurst exponent provide information related to scaling behavior and long-range temporal dependencies [15]. The presence of approximate, permutation, SVD, and spectral entropy among the selected attributes also indicates a contribution from measures associated with signal regularity and dynamical complexity [14]. At the same time, the presence of relative beta power, relative alpha power, total spectral power, and absolute beta power demonstrates that spectral information also contributed to the feature-selection process.

Complementarily, absolute and relative spectral power across the classical EEG frequency bands was explored as a descriptive analysis, independently of the mRMR ranking. The exploratory analysis suggested higher delta power in the ASD group and higher alpha power in the control group. Differences in alpha activity have also been reported in previous resting-state EEG studies involving ASD [5], [18], [19].

In the gamma band, similar central tendencies were observed between groups, although the ASD group showed greater dispersion and higher extreme values. Theta and beta power showed less evident group differences in this descriptive analysis. However, relative beta power and absolute beta power occupied the fifth and tenth positions, respectively, in the mRMR ranking. This apparent difference is not necessarily contradictory, since the spectral analysis considered each frequency-domain measure descriptively, whereas mRMR evaluates feature relevance to the class label while also accounting for redundancy among features. No inferential statistical tests were performed for these band-wise comparisons; therefore, the spectral findings should be interpreted as exploratory.

Following the feature analysis, the classification performance of the six evaluated pipelines was compared. Table I summarizes the results obtained by combining FastICA and SOBI with the no-windowing, non-overlapping, and 50% overlapping configurations. In the adopted hold-out evaluation, classification performance differed according to the source separation and segmentation strategies.

TABLE I
COMPARATIVE PERFORMANCE OF THE EVALUATED PIPELINES.

Method	Windowing	Acc	AUC	Prec.	Rec.	F1
FastICA	No windowing	0.674	0.657	0.712	0.693	0.703
FastICA	No overlap	0.615	0.566	0.635	0.720	0.675
FastICA	50% overlap	0.622	0.610	0.645	0.710	0.676
SOBI	No windowing	**0.741**	**0.730**	**0.756**	0.787	**0.771**
SOBI	No overlap	0.596	0.544	0.612	0.747	0.673
SOBI	50% overlap	0.659	0.649	0.653	**0.823**	0.729

Among the FastICA-based pipelines, the highest overall performance was obtained in the no-windowing condition, with an accuracy of 0.674, AUC of 0.657, precision of 0.712, recall of 0.693, and F1-score of 0.703. Both segmented configurations presented lower accuracy than the no-windowing condition, although the 50% overlapping configuration showed a slight improvement over the non-overlapping strategy for some metrics.

Among the SOBI-based pipelines, the highest overall performance was also obtained without windowing, with an accuracy of 0.741, AUC of 0.730, precision of 0.756, recall of 0.787, and F1-score of 0.771. Although the 50% overlapping configuration achieved the highest recall of the evaluated pipelines (0.823), its accuracy, AUC, precision, and F1-score remained below those obtained with SOBI without windowing. The non-overlapping SOBI configuration presented the lowest accuracy and AUC among the SOBI-based conditions.

In the adopted hold-out evaluation, neither segmentation strategy provided a consistent improvement over the corresponding no-windowing configuration. For both FastICA and SOBI, the highest accuracy and F1-score were obtained when the complete recording was used for feature extraction. Thus, within the conditions evaluated in this study, temporal segmentation did not provide a consistent classification advantage.

The higher overall performance observed for SOBI without windowing suggests that the combination of a source separation method based on second-order temporal statistics with feature extraction from the complete recording may be advantageous for this dataset. However, considering the exploratory design of the study, these results should not be interpreted as evidence of the general superiority of SOBI over FastICA.

Because this study was designed as an initial exploratory analysis, classification performance was evaluated using a single stratified hold-out partition. In addition, participant-wise separation was not explicitly enforced during the original classification protocol. Therefore, the reported performance should be interpreted as preliminary, and future studies should employ participant-wise repeated validation strategies to provide a more robust estimate of generalization performance.

V. CONCLUSION

This work compared EEG signal processing pipelines applied to individuals with Autism Spectrum Disorder and neurotypical controls. Among the evaluated configurations, SOBI without temporal segmentation showed the highest observed classification performance. Exploratory spectral analysis also indicated group-related patterns in delta, alpha, and gamma activity.

Overall, the findings show that source separation and segmentation choices were associated with differences in classification performance in the analyzed dataset. Considering the exploratory nature of the study and the adopted hold-out evaluation, these results should be interpreted as preliminary. Future studies should evaluate the proposed pipeline using participant-wise validation and larger datasets.

ACKNOWLEDGMENT

This work was partially supported by the São Paulo Research Foundation (FAPESP), grant #2020/09838-0, and by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) - Finance Code 001.

REFERENCES

[1] A. C. Marfinati and J. L. F. ABRÃO, "Um percurso pela psiquiatria infantil: dos antecedentes históricos à origem do conceito de autismo," *Estilos da Clínica*, vol. 19, no. 2, 2014.

[2] J. Li, X. Kong, L. Sun, X. Chen, G. Ouyang, X. Li, and S. Chen, "Identification of autism spectrum disorder based on electroencephalography: a systematic review," *Computers in Biology and Medicine*, 2024.

[3] C. M. Michel and D. Brunet, "Eeg source imaging: a practical review of the analysis steps," *Frontiers in Neurology*, 2019.

[4] Siuly, Y. Li, and Y. Zhang, *EEG Signal Analysis and Classification*. [S.l.]: Springer International Publishing, 2016.

[5] J. Wang, J. Barstein, L. E. Ethridge, M. W. Mosconi, S. Takara, and J. A. Sweeney, "Resting state eeg abnormalities in autism spectrum disorders," *Journal of Neurodevelopmental Disorders*, vol. 5, p. 24, 2013.

[6] M. Congedo, C. Gouy-Pailler, and C. Jutten, "On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics," *Clinical Neurophysiology*, vol. 119, no. 12, pp. 2677–2686, 2008.

[7] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[8] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, 1997.

[9] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, "The identification of children with autism spectrum disorder by svm approach on eeg and eye-tracking data," *Computers in Biology and Medicine*, 2020.

[10] J. Zhao, J. Song, X. Li, and J. Kang, "A study on eeg feature extraction and classification in autistic children based on singular spectrum analysis method," *Brain and Behavior*, 2020.

[11] B. Y. Angulo-Ruiz, F. J. Ruiz-Martínez, E. I. Rodríguez-Martínez *et al.*, "Linear and non-linear analyses of eeg in a group of asd children during resting state condition," *Brain Topography*, vol. 36, pp. 736–749, 2023.

[12] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, 2019.

[13] S. Sanei and J. A. Chambers, *EEG Signal Processing*. Chichester: Wiley-Interscience, 2013.

[14] S. Y. Ke, H. Wu, H. Sun, A. Zhou, J. Liu, X. Zheng, K. Liu, M. B. Westover, H. Xu, and X.-j. Kong, "Classification of autism spectrum disorder using electroencephalography in chinese children: a cross-sectional retrospective study," *Frontiers in Neuroscience*, 2024.

[15] M. Radhakrishnan, K. Ramamurthy, A. Kothandaraman, G. Madaan, and H. Machavaram, "Investigating eeg signals of autistic individuals using detrended fluctuation analysis," *Traitement du Signal*, vol. 38, 2021.

[16] P. Garcés, S. Baumeister, L. Mason, C. H. Chatham, S. Holiga, J. Dukart, E. J. H. Jones, T. Banaschewski, S. Baron-Cohen, S. Bölte, J. K. Buitelaar, S. Durston, B. Oranje, A. M. Persico, C. F. Beckmann, T. Bougeron, F. Dell'Acqua, C. Ecker, C. Moessnang, T. Charman, J. Tillmann, D. G. M. Murphy, M. Johnson, E. Loth, D. Brandeis, J. F. Hipp, and E.-A. L. Group, "Resting state eeg power spectrum and functional connectivity in autism: a cross-sectional analysis," *Molecular Autism*, 2022.

[17] C. M. Goulart, J. F. Castillo-Garcia, M. Trauernicht, C. T. Valadão, E. M. O. Caldeira, and T. F. Bastos-Filho, "Estudo de estados emocionais e mentais de crianças com autismo baseado em eeg na interação com um robô móvel," in *Anais do XXIV Congresso Brasileiro de Engenharia Biomédica (CBEB 2014)*. Universidade Federal do Espírito Santo, 2014.

[18] A. Dickinson, S. Jeste, and E. Milne, "Electrophysiological signatures of brain aging in autism spectrum disorder," *Cortex*, pp. 139–151, 2022.

[19] W. S. Neo, D. Foti, B. Keehn, and B. Kelleher, "Resting-state eeg power differences in autism spectrum disorder: a systematic review and meta-analysis," *Translational Psychiatry*, p. 389, December 2023.

Computational Pipeline for Detecting and Quantifying Spectral Modulation in EEG for 12-Hz Binaural Stimulation Studies in the Context of Alzheimer's Disease

Isabela Rogério Cardoso*
Electrical Engineering
State University of Amazonas
Manaus, AM, Brazil
irc.eng25@uea.edu.br

Fábio de Sousa Cardoso
PPGEE/Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
fcardoso@uea.edu.br

Roberto Alexandre Alves Barbosa Filho
Military School of Manaus
Manaus, AM, Brazil
robertoaabfilho@gmail.com

Danilo da Silva Cunha
Cyberlabs - State University of Amazonas
Manaus, AM, Brazil
dsc.elt@uea.edu.br

Ada Jamile Gomes de Oliveira
Military School of Manaus
Manaus, AM, Brazil
adajamilegomesdeoliveira.a@gmail.com

Isaque Vilson Batista da Costa
PPGEEL
State University of Amazonas
Manaus, AM, Brazil
ivbdc.mee24@uea.edu.br

Afonso Aguiar de Carvalho
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
aadc.ele21@uea.edu.br

Guilherme Pansini Barbosa
School of Law
State University of Amazonas
Manaus, AM, Brazil
gpansini2000@gmail.com

Abstract*—This paper presents the development and validation of a modular computational pipeline for electroencephalography (EEG) signal processing and automated spectral modulation detection. Implemented in Python with MNE-Python, the system integrates EDF file reading, channel standardization, temporal segmentation, Welch power spectral density (PSD) estimation, extraction of five spectral features, aggregation, statistical analysis, and automated report generation. The pipeline was validated through two complementary strategies. First, a physiological control based on alpha reactivity between eyes-open and eyes-closed resting states confirmed that the software preserves genuine spectral changes: across 109 participants from the EEG Motor Movement/Imagery Dataset (3,052 epochs at the Oz channel), all five metrics differed significantly between conditions after Holm correction, with large effect sizes. Second, a controlled synthetic-signal protocol tested whether the pipeline specifically detects and quantifies a modulation at the target frequency (12 Hz): EEG signals with known 12-Hz amplitude and known signal-to-noise ratio (six SNR levels from −10 to +20 dB, five modulation levels, 15 synthetic subjects per condition) were processed and compared against their known ground truth. Recovered target-band power matched the theoretical value with a mean absolute relative error of 1.8% at the strongest modulation level, and the modulation was statistically detected across all paired post-vs.-pre test conditions, including low-SNR scenarios. No EEG recordings from binaural-stimulation experiments or individuals with Alzheimer's disease were analyzed in the present study; these results indicate that the pipeline reproduces a known physiological effect and recovers a specific, quantitatively known spectral modulation, providing the computational infrastructure required for future EEG studies of 12-Hz binaural stimulation.

***Keywords*—computational pipeline, binaural beats, electroencephalography, spectral modulation, power spectral density, synthetic validation, ground truth.**

I. INTRODUCTION

Electroencephalography (EEG) provides a noninvasive, time-resolved measure of brain electrical activity and is widely used to study spectral and connectivity changes associated with aging and neurodegeneration [2], [3]. Comparing EEG recordings across experimental conditions, however, requires a processing chain that reads raw signals, extracts frequency-domain features in a standardized way, and applies appropriate statistical comparisons — components that are often implemented ad hoc and inconsistently across studies, hindering reproducibility.

This work is part of a broader project investigating whether a 12-Hz binaural acoustic protocol can produce a measurable electrophysiological response in the context of Alzheimer's disease research. The broader project adopted 12 Hz as a predefined experimental target for subsequent investigation; however, the present article does not evaluate the biological rationale, therapeutic efficacy, or clinical effects of that stimulation. Its scope is limited to the computational infrastructure required to objectively identify and quantify changes in the spectral content of EEG recordings acquired before, during, and after stimulation.

The central contribution of this work is computational, not clinical: an automated, modular pipeline that reads raw EEG recordings, preserves sample traceability, estimates the frequency-domain power distribution, extracts spectral biomarkers, and statistically compares experimental conditions. Because the pipeline is intended to eventually support the 12-Hz binaural protocol, it must be validated not only on a phenomenon with a known direction of change, but specifically on its ability to recover a modulation of known frequency, amplitude, and noise level.

To that end, the pipeline underwent two independent validation strategies. The first is a physiological control with a well-documented, reproducible effect: alpha reactivity between eyes-open and eyes-closed resting states [4], [5]. This test confirms that reading, segmentation, spectral estimation, feature extraction, and statistical testing operate correctly as an integrated chain, but — because its expected direction, not its magnitude, is known in advance — it cannot by itself establish that the pipeline correctly quantifies an arbitrary spectral modulation. The second validation strategy is a controlled synthetic-signal protocol in which the amplitude and the signal-to-noise ratio of a 12-Hz component are set explicitly and compared against the pipeline's recovered estimates, directly measuring detection and quantification error.

This study aims to develop and validate this modular pipeline for detecting and quantifying spectral modulation in EEG. Its main contribution is a reusable computational infrastructure, validated against both a known physiological effect and a known synthetic ground truth, capable of processing public datasets and future experimental recordings, generating biomarkers, performing statistical comparisons, and producing standardized outputs for auditory neuromodulation studies — including, but not limited to, the 12-Hz binaural protocol.

II. BACKGROUND

A. Spectral Modulation in EEG

Scalp EEG mainly reflects the spatiotemporal summation of postsynaptic potentials from cortical neuronal populations. In this study, spectral modulation is defined as a quantifiable difference between two conditions, expressed through the integrated energy of a frequency band (absolute power), that energy normalized by the total analyzed power (relative power), or properties of the dominant oscillatory component (peak frequency and peak PSD amplitude). For the 12-Hz protocol, the 10–14 Hz range was defined as the target band. Alpha reactivity to eye closure — a well-documented increase in occipital alpha power upon closing the eyes [4], [5] — was used as a physiological control because its direction is known in advance, allowing verification that the software chain preserves a genuine spectral change.

B. Project Context: 12-Hz Binaural Stimulation

Binaural beats are perceived when each ear receives a pure tone at a slightly different frequency; the difference is processed along the auditory pathway and, in some protocols, has been associated with changes in cortical oscillatory activity, although reported effects are inconsistent and depend strongly on stimulation parameters [6]–[9]. In the broader project supported by this pipeline, a 12-Hz binaural difference is the predefined stimulation condition selected for future EEG investigation in the context of Alzheimer's disease research [1]–[3]. The present study does not test therapeutic efficacy or Alzheimer's-specific effects; it provides the infrastructure to later compare EEG recordings obtained before, during, and after exposure to the 12-Hz protocol, quantifying any resulting spectral modulation with the architecture validated in this paper.

III. MATERIALS AND METHODS

A. Pipeline Architecture

The pipeline was developed in Python using MNE-Python [12] as a modular processing sequence, separating data acquisition, file reading, standardization, preprocessing, spectral estimation, feature extraction, statistical aggregation, and output generation into independent stages. Each stage receives a defined input, performs one function, and passes its output forward; participant, condition, channel, and epoch identifiers are preserved throughout, ensuring traceability from any aggregated result back to its source recording. The complete modular processing sequence is summarized in Fig. 1. Parameterization (channels of interest, epoch duration and overlap, band limits, compared conditions) is kept separate from the analysis logic, so the same core processing chain is reused across the PhysioNet dataset (Section III.B), the synthetic ground-truth signals (Section III.E), and, in future work, recordings from the 12-Hz binaural protocol, with only source-specific acquisition parameters (e.g., sampling rate) supplied through configuration.

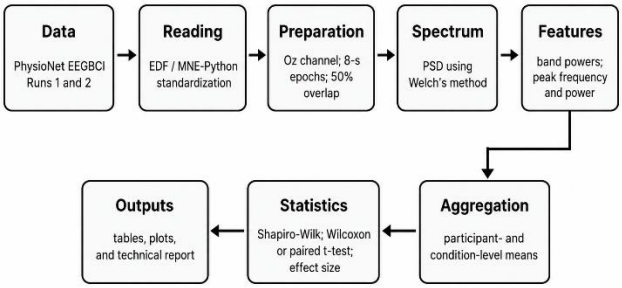


Fig. 1. Modular workflow of the EEG processing and analysis pipeline.

B. Physiological Validation Dataset

The EEG Motor Movement/Imagery Dataset, version 1.0.0, from PhysioNet [10], acquired with the BCI2000 system [11], was used for the physiological control. The dataset contains 109 participants, 64 channels, and a sampling frequency of 160 Hz; baseline Run 1 (eyes open) and Run 2 (eyes closed) were selected. The Oz channel, over the occipital region where alpha reactivity is typically most pronounced, was used. Prior to spectral analysis, recordings were band-pass filtered between 1 and 40 Hz (4th-order Butterworth, zero-phase); a 60-Hz notch filter was applied beforehand as a preprocessing safeguard against strong power-line interference, although it is not strictly necessary for the reported 1–40 Hz analysis band once the band-pass filter is applied. Results reported here use the single Oz channel without additional re-referencing. Signals were segmented into 8-s epochs with 50% overlap, yielding 3,052 epochs (1,526 per condition).

C. Spectral Processing and Feature Extraction

For each epoch, PSD was estimated using Welch's method [13] with 4-s Hann windows, 50% overlap between segments, and nfft = nperseg (no zero-padding), averaging modified periodograms to reduce estimation variability. This corresponds to nperseg = nfft = 640 samples at the 160-Hz PhysioNet sampling rate and nperseg = nfft = 1024 samples for the 256-Hz synthetic signals (Section III.E) — both yielding the same frequency resolution, $\Delta f = 0.25$ Hz. PSD is reported in V^2/Hz for the PhysioNet recordings and in

arbitrary-units squared per hertz (a.u.²/Hz) for the synthetic signals.

Five features were extracted per epoch: absolute alpha power (8–13 Hz); alpha power relative to the total power integrated between 1 and 40 Hz, matching the upper edge of the preprocessing band-pass filter (Section III.B) — the same reference band is used for both the PhysioNet physiological validation and the synthetic ground-truth validation (Section III.E), as this normalization is applied by a single shared function with no dataset-specific branching; integrated power in the 10–14 Hz target band; and, searched within that same 10–14 Hz target band (not the 8–13 Hz alpha band), the frequency and amplitude of the largest spectral peak. For the PhysioNet data, integrated powers are reported in V² and peak PSD amplitude in V²/Hz; for the synthetic validation, the corresponding units are a.u.² and a.u.²/Hz. Peak PSD amplitude is read directly from the PSD at a single frequency bin rather than integrated over a band, consistent with the PSD representation shown in Fig. 2.

D. Physiological Validation Statistics

The functional validation criterion was recovery of the expected alpha increase at Oz during eyes-closed rest. Metrics were calculated per epoch and summarized per participant; normality of paired differences was assessed with Shapiro–Wilk [14], and the Wilcoxon signed-rank test [15] (rank-biserial effect size) or the paired t-test (Cohen's dz [16]) were applied accordingly, with $\alpha = 0.05$ and Holm correction for five outcomes.

E. Controlled Synthetic Validation Protocol

A physiological control such as eyes-open/eyes-closed alpha reactivity confirms that the pipeline detects a genuine spectral change in the expected direction, but its magnitude is not known in advance, so it cannot alone establish whether the pipeline correctly quantifies an arbitrary modulation — the requirement for the 12-Hz binaural protocol, where the size of any effect is unknown a priori. To address this, a second, independent validation used synthetic EEG signals in which the amplitude and signal-to-noise ratio (SNR) of a 12-Hz component are explicit, known parameters.

Each synthetic signal (60 s, 256 Hz) combines a fixed ~10-Hz background rhythm of amplitude 1.0 a.u., additive 1/f-shaped noise, and, in post-stimulus conditions, a pure 12-Hz sinusoidal component of controlled amplitude A, all expressed in arbitrary units (a.u.). Five modulation levels were defined as fractions of a reference amplitude Aref = 1.0 a.u.: pre-stimulus (0%), and post-stimulus very weak (15%, A = 0.15 a.u.), weak (35%, A = 0.35 a.u.), moderate (65%, A = 0.65 a.u.), and strong (100%, A = Aref = 1.0 a.u.). Six SNR levels were tested (−10, −5, 0, 5, 10, and 20 dB), computed relative to Aref, where noise_amp is the standard deviation of the generated zero-mean 1/f noise and therefore equals its RMS amplitude. The generator normalizes the pink-noise sequence to unit standard deviation before scaling it by noise_amp:

$$SNRdB = 10 \cdot \log10[(Aref^2/2) / noise_amp^2] \qquad (1)$$

Within each SNR group, the same background-noise level was used across all five signal-generation levels, isolating the effect of modulation amplitude from the effect of the noise floor. The six SNR levels and five generation levels produced 30 signal-generation configurations; with 15 independent synthetic subjects per configuration, 450 signals were generated in total. Statistical pre-vs.-post detection, however, involved only the four nonzero modulation levels at each of the six SNRs, yielding 24 post-vs.-pre comparisons. Every signal was processed through the same preprocessing, epoching, and Welch PSD pipeline used for the physiological validation.

Within each (SNR, subject) pair, the pre-stimulus and all post-stimulus signals share the same random seed, containing an identical background-noise realization and differing only in the amplitude of the added 12-Hz component. This design supports paired pre-vs.-post comparisons for the same synthetic subject — analogous to how the future 12-Hz protocol will compare the same individual before and after stimulation — while controlling the background-noise realization. Normality of within-subject differences was assessed with Shapiro–Wilk; a paired t-test was used when the normality criterion was met and the Wilcoxon signed-rank test otherwise. The minimum detectable modulation was defined as the lowest nonzero modulation level producing a significant paired post-vs.-pre difference ($p < 0.05$) in target-band power at each SNR level.

The theoretical power added by a pure 12-Hz sinusoid of amplitude A is, by Parseval's theorem:

$$Pexpected = A^2 / 2 \qquad (2)$$

The power actually recovered by the pipeline was computed as the paired difference in target-band power between the post- and pre-stimulus signals of the same synthetic subject, canceling the shared noise floor and background rhythm. The peak-frequency error was computed as the recovered peak frequency (searched within the 10–14 Hz target band, as in Section III.C) minus the inserted frequency (12.00 Hz).

IV. RESULTS

A. Physiological Validation

Across the 109 participants (3,052 epochs), all five spectral metrics differed significantly between eyes-open and eyes-closed conditions after Holm correction (Table I). The four power-related metrics increased upon eye closure — relative alpha power rose from 18.39% to 42.41% (Wilcoxon, $p = 4.29\times10^{-19}$, $|r| = 0.985$) — whereas peak frequency shifted downward, from 11.32 to 10.74 Hz (paired t-test, $p = 4.63\times10^{-15}$, $|dz| = 0.874$). Fig. 2 shows the expected concentration of PSD energy within the alpha/target band upon eye closure, and Fig. 3 shows the corresponding epoch-level distribution. This physiological control confirms that the software chain — reading, segmentation, Welch PSD, feature extraction, and statistical testing — preserves a real spectral difference. Because only its direction, not its magnitude, was known beforehand, it is complemented by the synthetic ground-truth validation below, which provides a quantitative test of recovery accuracy under known conditions.

TABLE I. PHYSIOLOGICAL VALIDATION (EYES-OPEN → EYES-CLOSED)

Metric	Mean EO→EC	p-value	Effect
Abs. alpha power ($\times10^{-10}$V²)	1.652→8.856	3.54×10^{-19}	0.988
Rel. alpha power (%)	18.39→42.41	4.29×10^{-19}	0.985
Target-band power ($\times10^{-10}$V²)	1.018→4.825	2.48×10^{-19}	0.992

Metric	Mean EO→EC	p-value	Effect
Peak PSD ampl. ($\times 10^{-10}$V²/Hz)	0.840→6.575	2.06×10^{-18}	0.966
Peak frequency (Hz)	11.32→10.74	4.63×10^{-15}	0.874

EO: eyes open; EC: eyes closed. All outcomes remained significant after Holm correction (Wilcoxon signed-rank, except peak frequency: paired t-test). Relative power is referenced to total power integrated over 1–40 Hz, as described in Section III.C.

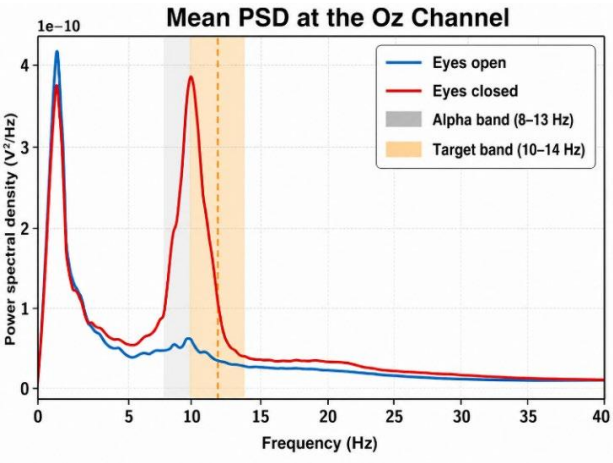


Fig. 2. Mean PSD at the Oz channel, eyes open vs. eyes closed.

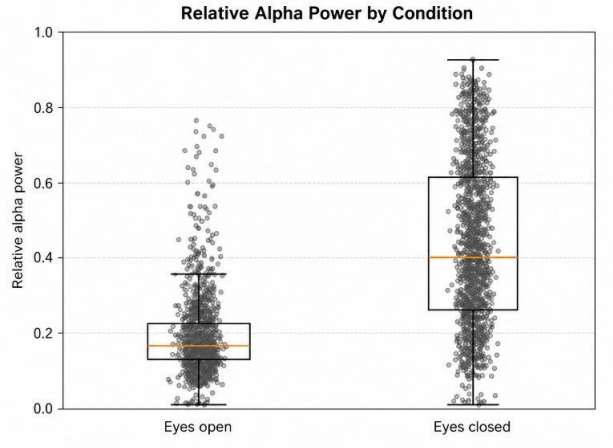


Fig. 3. Distribution of relative alpha power by epoch, eyes open vs. eyes closed.

B. Controlled Synthetic Validation

Fig. 4 shows the recovered target-band power (mean ± SD across 15 synthetic subjects) as a function of the inserted 12-Hz amplitude, for all six SNR levels, together with the theoretical curve Pexpected = A²/2. Recovered power closely tracks the theoretical curve across the entire tested range.

Averaged across all SNR levels, the mean relative power error decreased monotonically with modulation strength (Table II), from −3.4% (|error| 12.1%) at the very weak level to −0.5% (|error| 1.8%) at the strong level. At the strong level specifically, the mean relative error ranged from −1.9% ± 6.3% at the most adverse SNR (−10 dB) to −0.03% ± 0.15% at the most favorable SNR (+20 dB).

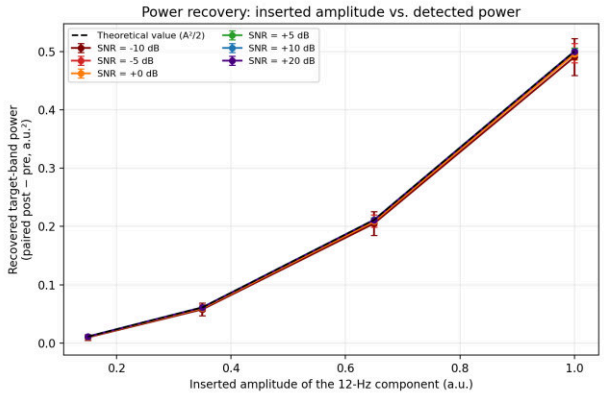


Fig. 4. Inserted 12-Hz amplitude vs. recovered target-band power, by SNR.

TABLE II. MEAN RECOVERY ERROR BY MODULATION LEVEL (ALL SNR POOLED, n = 90/LEVEL)

Level	Freq. err (Hz)	Pow. err (%)	|Pow. err| (%)
Very weak (15%)	−2.00	−3.4	12.1
Weak (35%)	−2.00	−1.5	5.2
Moderate (65%)	−2.00	−0.8	2.8
Strong (100%)	−1.05	−0.5	1.8

Fig. 5 shows the corresponding peak-frequency error. At the very weak, weak, and moderate modulation levels, the recovered peak remained close to the background ~10-Hz rhythm (error near −2 Hz) across the tested SNRs. An appreciable shift toward the inserted 12-Hz component became evident at the strong modulation level, where the mean error decreased to approximately −1.05 Hz. This does not indicate a failure of target-band-power detection; rather, it reflects that the peak-frequency metric identifies the largest local maximum within the 10–14 Hz band, which remains dominated by the ~10-Hz background component at lower modulation amplitudes — target-band power (Table II, Fig. 4) already detects the added modulation at those same levels, so the two metrics are not in conflict. Fig. 6 illustrates this transition for a representative synthetic subject: the pre- and post-stimulus PSD curves coincide throughout the spectrum except for the additional 12-Hz peak introduced by the modulation.

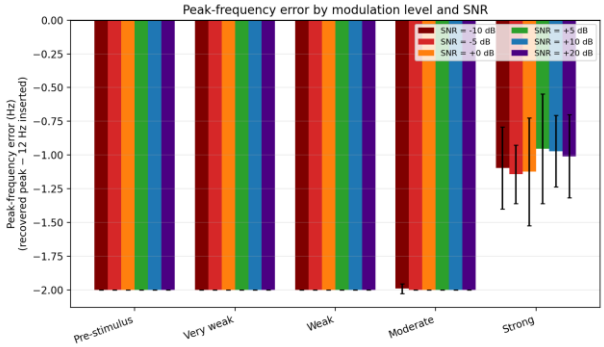


Fig. 5. Peak-frequency error by modulation level and SNR.

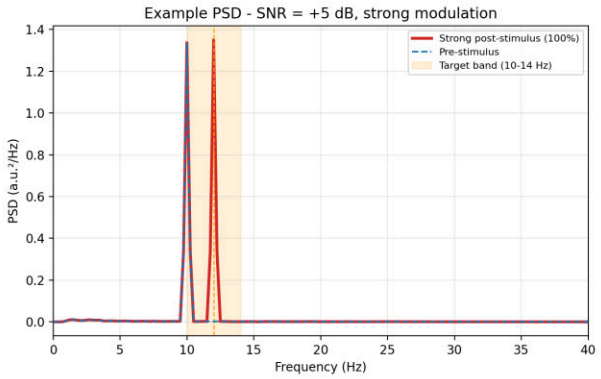


Fig. 6. Example PSD, pre- vs. strong post-stimulus (SNR = +5 dB, one synthetic subject).

Statistical detection followed the same pattern. Using the target-band-power metric, the paired comparison against the matched pre-stimulus condition (same synthetic subject, same noise realization) was significant ($p < 0.05$) for all 24 tested post-vs.-pre combinations (six SNR levels × four nonzero modulation levels), including the weakest modulation at the lowest SNR ($p = 1.2\times10^{-6}$). Consequently, the minimum detectable modulation in the paired analysis was the very weak level (15% of Aref) at every tested SNR, as summarized in Table III. This result characterizes sensitivity under the controlled matched-noise design used for the synthetic validation and should not be interpreted as a detection limit for real EEG recordings.

TABLE III. MINIMUM MODULATION LEVEL DETECTED IN PAIRED COMPARISONS ($p < 0.05$, TARGET-BAND POWER)

SNR (dB)	Minimum detected level (paired)
−10	Very weak (15%)
−5	Very weak (15%)
0	Very weak (15%)
5	Very weak (15%)
10	Very weak (15%)
20	Very weak (15%)

C. Pipeline Performance Summary

Together, the two validations characterize complementary aspects of pipeline performance: the physiological control confirms that a real, direction-known spectral change is preserved end-to-end; the synthetic ground-truth protocol confirms that a modulation of known frequency and amplitude is both detected with statistical significance and quantified with a small, SNR-dependent relative error. Neither result establishes, by itself, that the eventual 12-Hz binaural protocol will produce a detectable effect in real EEG recordings — that remains an open empirical question the validated pipeline is now positioned to answer.

V. DISCUSSION

A. Integrated Validation

The convergence of two independent validation strategies — one based on a real physiological effect of known direction, the other on synthetic signals of known magnitude — provides stronger evidence than either alone that the pipeline's reading, spectral estimation, feature extraction, and statistical modules operate correctly as an integrated chain. Had a systematic error existed (e.g., incorrect PSD normalization, misaligned band integration, or mismatched participant/condition bookkeeping), it is unlikely that both the direction of the physiological effect and the quantitative magnitude of the synthetic modulation would be recovered simultaneously across the tested range of SNR and amplitude.

B. Toward the 12-Hz Binaural Protocol

These results establish the pipeline as ready to process recordings from the 12-Hz binaural protocol once available, but they do not constitute evidence of a neuromodulatory effect of that protocol. The synthetic validation deliberately used an idealized additive signal model (a pure sinusoid over structured background noise) and does not capture the artifacts, inter-individual variability, or unknown effect size that a real binaural-stimulation recording may present; the true minimum detectable amplitude in real EEG may differ from Table III. The Alzheimer's-related context should therefore be interpreted only as a planned application domain for future investigation, not as an outcome of the present study [1]–[3]. This pipeline neither confirms nor tests a disease-specific, therapeutic, or electrophysiological effect of 12-Hz stimulation; it only provides the infrastructure to look for spectral modulation with a characterized detection and quantification error under known conditions. Any future claim about a specific neuromodulatory or Alzheimer's-related effect would require applying this validated pipeline to real pre-/post-stimulation recordings and interpreting the result on its own terms, not by extrapolation from the validations reported here.

C. Limitations and Future Work

Both validations share limitations. The physiological validation used a single channel (Oz) and healthy participants only; multichannel analysis, automated artifact rejection, and topographic mapping were not implemented. The synthetic validation used an idealized 1/f-noise-plus-sinusoid model that does not capture non-stationary artifacts, electrode/impedance variability, or physiological confounds present in real recordings, and tested a single target frequency (12 Hz) and background alpha amplitude; results at other frequencies, montages, or noise structures were not characterized. Future work should extend the synthetic protocol to multichannel signals and alternative noise models, apply the validated pipeline to real pre-, during-, and post-stimulation recordings from the 12-Hz binaural protocol, and incorporate automated artifact handling, functional connectivity, and individual alpha frequency adjustment.

VI. CONCLUSION

A modular computational pipeline was developed for reading, organizing, spectrally processing, and statistically analyzing EEG signals, and was validated through two complementary strategies: a physiological control with a known direction of effect (alpha reactivity to eye closure) and a controlled synthetic protocol with known modulation amplitude and SNR. The synthetic validation showed that the pipeline recovers the inserted 12-Hz power with a mean absolute relative error of 1.8% at the strongest tested modulation and detects the modulation with statistical significance across all paired post-vs.-pre conditions in the tested range of amplitude and SNR, including adverse noise conditions.

The pipeline therefore provides the infrastructure required to analyze EEG signals before, during, and after 12-Hz binaural stimulation, supporting future studies on whether measurable neurophysiological modulation occurs in the context of Alzheimer's disease research. Beyond this application, its parameterizable, dataset-independent architecture and its characterized detection/quantification error under known conditions support reuse in other neuromodulation studies, with future extension to multichannel analysis, connectivity, automated artifact handling, and validation on real stimulation recordings.

REFERENCES

[1] C. R. Jack Jr. et al., "Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup," Alzheimer's Dement., vol. 20, pp. 5143–5169, 2024, doi: 10.1002/alz.13859.

[2] S. Ehteshamzad, "Assessing the potential of EEG in early detection of Alzheimer's disease: A systematic comprehensive review (2000–2023)," J. Alzheimer's Dis. Rep., vol. 8, no. 1, pp. 1153–1169, 2024, doi: 10.3233/ADR-230159.

[3] C. A. Chetty et al., "EEG biomarkers in Alzheimer's and prodromal Alzheimer's: A comprehensive analysis of spectral and connectivity features," Alzheimer's Res. Ther., vol. 16, Art. no. 236, 2024, doi: 10.1186/s13195-024-01582-w.

[4] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, and J. A. Rushby, "EEG differences between eyes-closed and eyes-open resting conditions," Clin. Neurophysiol., vol. 118, no. 12, pp. 2765–2773, 2007, doi: 10.1016/j.clinph.2007.07.028.

[5] W. Klimesch, "Alpha-band oscillations, attention, and controlled access to stored information," Trends Cogn. Sci., vol. 16, no. 12, pp. 606–617, 2012, doi: 10.1016/j.tics.2012.10.007.

[6] R. M. Ingendoh, E. S. Posny, and A. Heine, "Binaural beats to entrain the brain? A systematic review of the effects of binaural beat stimulation on brain oscillatory activity, and the implications for psychological research and intervention," PLoS ONE, vol. 18, no. 5, Art. no. e0286023, 2023, doi: 10.1371/journal.pone.0286023.

[7] H. D. Orozco Perez, G. Dumas, and A. Lehmann, "Binaural beats through the auditory pathway: From brainstem to connectivity patterns," eNeuro, vol. 7, no. 2, Art. no. ENEURO.0232-19.2020, 2020, doi: 10.1523/ENEURO.0232-19.2020.

[8] M. Garcia-Argibay, M. A. Santed, and J. M. Reales, "Efficacy of binaural auditory beats in cognition, anxiety, and pain perception: A meta-analysis," Psychol. Res., vol. 83, no. 2, pp. 357–372, 2019, doi: 10.1007/s00426-018-1066-8.

[9] S. Basu and B. Banerjee, "Potential of binaural beats intervention for improving memory and attention: Insights from meta-analysis and systematic review," Psychol. Res., vol. 87, no. 4, pp. 951–963, 2023, doi: 10.1007/s00426-022-01706-7.

[10] G. Schalk, "EEG Motor Movement/Imagery Dataset," PhysioNet, ver. 1.0.0, 2009, doi: 10.13026/C28G6P.

[11] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," IEEE Trans. Biomed. Eng., vol. 51, no. 6, pp. 1034–1043, 2004, doi: 10.1109/TBME.2004.827072.

[12] A. Gramfort et al., "MEG and EEG data analysis with MNE-Python," Front. Neurosci., vol. 7, Art. no. 267, 2013, doi: 10.3389/fnins.2013.00267.

[13] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," IEEE Trans. Audio Electroacoust., vol. 15, no. 2, pp. 70–73, 1967, doi: 10.1109/TAU.1967.1161901.

[14] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika, vol. 52, nos. 3–4, pp. 591–611, 1965, doi: 10.1093/biomet/52.3-4.591.

[15] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bull., vol. 1, no. 6, pp. 80–83, 1945, doi: 10.2307/3001968.

[16] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

Computational tool for automated detection of multiple sclerosis brain lesions using magnetic resonance imaging

1st Paulo Márcio Borges Daniel
Hospital de Clínicas da UFPR, Postgraduate Program in Biomedical Engineering
Federal University of Paraná-UFPR, UTFPR
Curitiba, Brazil
0000-0002-6198-3851

2nd Otavio Hruba
Academic Department of Informatics
UTFPR
Ponta Grossa,Brazil
0009-0008-0990-6091

3rd Ana Gabriela Santos Souza
Hospital de Clínicas-UFPR
Federal University of Paraná-UFPR
Curitiba-Brazil
0009-0006-6633-4400

4th Valentina de Castro Jantsch Ribeiro
Department of Mechatronics Engineering
UTFPR
Curitiba-Brazil
valentinar@alunos.utfpr.edu.br

5th Paulo Victor Dos Santos
Instituto De Inovação Em Gestão
Universidade Federal De Goiás
Cidade Ocidental, Brazil
0000-0002-8267-7562

6th Cristhiane Gonçalves
Postgraduate Program in Biomedical Engineering
Federal University of Technology-UTFPR
Curitiba-Brazil ORCID 0000-0002-3894-2971

***Abstract*—This study presents the development and evaluation of an unsupervised computational framework tailored for the automated detection and segmentation of demyelinating brain lesions in patients with multiple sclerosis (MS) using fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI). To enhance dataset heterogeneity and improve external validity, a total of 60 brain MRI examinations were obtained from two distinct sources: the Hospital de Clínicas of the Federal University of Paraná (UFPR), Brazil, and a public international repository from the University of Ljubljana, Slovenia. Comparative tests against the LS Segmenter extension in 3D Slicer software demonstrate that while standard methods suffer from significant lesion overestimation, the proposed tool isolates focal periventricular and juxtacortical plaques more effectively, providing instant quantitative spatial metrics . This framework represents a scalable and objective tool suitable for clinical diagnostic support and longitudinal tracking of MS lesion burden.**

***Index Terms*—K-means clustering, lesion segmentation, magnetic resonance imaging (MRI), multiple sclerosis (MS), unsupervised machine learning framework**

I. Introduction

Multiple sclerosis is a complex and chronic inflammatory demyelinating and immune-mediated disorder that affects the central nervous system and leads to progressive neurological impairment [1] . The disease is characterized by focal areas of demyelination, neuroaxonal injury, and gliosis, which result in heterogeneous neurological manifestations including visual, sensory, motor, sphincteric, and cognitive dysfunction. Although the precise mechanisms underlying the disease remain incompletely understood, current evidence suggests the interaction of genetic susceptibility, environmental exposure, and dysregulated immune responses, culminating in inflammatory damage to the myelin sheath and disruption of axonal conduction [1] [2]. In total, approximately 2.8 million individuals are estimated to be affected by multiple sclerosis, with a higher prevalence observed in populations located at higher latitudes and among young adults, specifically women [1] [2] [3]. Recently, deep learning approaches, particularly convolutional neural networks, have demonstrated significant potential for automated lesion detection and segmentation in brain magnetic resonance imaging. Architectures derived from U-Net, originally proposed for biomedical image segmentation, have become one of the most widely used frameworks for this purpose [4].

Studies have demonstrated encouraging results in the application of deep learning models to detect demyelinating lesions and quantify the burden of the injury in brain magnetic resonance images [5] [6].

Magnetic resonance imaging plays a central role in the diagnosis and follow-up of the disease. The technique provides high sensitivity for the detection of demyelinating lesions in both the brain and the spinal cord and has become essential to demonstrate the dissemination of lesions in time and space, a key element in the diagnostic criteria for multiple sclerosis [1] [7]. In addition, magnetic resonance imaging is widely used for longitudinal monitoring of disease activity, allowing the identification of new lesions or the progression of existing ones, as well as the evaluation of therapeutic response. Early identification of disease activity is particularly relevant because

multiple sclerosis represents one of the main causes of non-traumatic neurological disability in young adults.

Typical lesions appear as hyperintense areas on T2-weighted images and FLAIR(Fluid-attenuated inversion recovery) sequences, usually oval or round in morphology and measuring at least 3 mm in diameter, corresponding to focal inflammatory demyelinating processes within the central nervous system. Their spatial distribution follows characteristic anatomical patterns that reflect diagnostic criteria for dissemination in space. The most relevant locations include the periventricular region, cortical/juxtacortical white matter, infratentorial compartment, optic nerves, and spinal cord [1]. These patterns reflect the preferential involvement of the perivenular white matter pathways, which represent a key pathological feature of the disease. Among the most characteristic imaging features are the "Dawson's fingers", which represent lesions oriented radially and perpendicular to the lateral ventricles. These lesions correspond to inflammatory demyelination along the medullary veins and are considered a classic imaging hallmark of multiple sclerosis [1]. Their presence, particularly when associated with the "central vein sign", significantly increases the specificity of magnetic resonance imaging to differentiate multiple sclerosis from other white matter disorders, such as small vessel disease of the brain. For lesions to be classified as juxtacortical, they must directly contact the cortical ribbon and involve the subcortical U-fibers. This imaging feature is particularly relevant for the differential diagnosis, as many vascular white matter lesions spare these fibers.

Despite the critical role of imaging in disease management, lesion identification and comparison across serial examinations remain largely dependent on manual evaluation by radiologists. This process may be time-consuming and subject to variability between observers, particularly when dealing with subtle changes or large data sets of images [8]. The increasing demand for neuroimaging examinations in clinical practice has further intensified the workload of neuroradiologists, highlighting the need for computational tools capable of assisting image interpretation and improving workflow efficiency. Recently, advances in machine learning and deep learning techniques have enabled the development of computational systems capable of extracting complex patterns from medical images and performing automated segmentation of pathological structures [9] [10]. In the context of multiple sclerosis, these approaches have shown promising results in detecting and quantifying white matter lesions on magnetic resonance imaging. Automated lesion segmentation methods have the potential to improve reproducibility, reduce observer variability, and facilitate an objective evaluation of the disease burden over time [1] [9] [10].

More recently, research has increasingly focused on automated longitudinal analysis of magnetic resonance examinations, with the aim of identifying new or enlarging lesions in follow-up studies and supporting clinical decision-making in patient monitoring. However, several challenges remain for the clinical implementation of these systems. Differences in acquisition protocols, scanner characteristics, and image quality between institutions can significantly affect algorithm performance and limit generalization between datasets. Consequently, contemporary research has emphasized the development of robust segmentation pipelines and the use of large heterogeneous datasets for algorithm training and validation [10]. Modern deep learning architectures, including convolutional neural networks and self-configuring segmentation frameworks, have demonstrated improved performance in medical image analysis tasks [11].

Approaches based on automated pipeline configuration and external validation have shown encouraging results in the segmentation of multiple sclerosis lesions in FLAIR magnetic resonance sequences [12]. In addition, ensemble strategies that combine predictions from multiple algorithms have been proposed to improve robustness and reliability in heterogeneous clinical scenarios. Emerging methodologies have also explored the integration of foundation-model concepts and prompt-guided segmentation frameworks, reflecting a broader trend toward more adaptable artificial intelligence systems for medical imaging applications.

Another important aspect addressed in recent studies concerns the variability of lesion appearance and imaging protocols in multicenter datasets, which represents a major obstacle to the generalization of automated methods. Large collaborative datasets and international segmentation challenges have demonstrated that algorithm performance may vary substantially depending on the characteristics of the imaging data and patient population. These findings reinforce the importance of developing computational tools capable of maintaining adequate performance in real-world clinical conditions. In this context, the development of automated systems capable of identifying and quantifying demyelinating lesions in magnetic resonance imaging represents an important strategy to improve diagnostic support and longitudinal monitoring of patients with multiple sclerosis. Such tools may contribute to reducing the workload of neuroradiologists while enabling a more objective and reproducible assessment of disease activity. Therefore, the guiding research question of this study is whether machine learning–based computational approaches can reliably identify typical multiple sclerosis lesions on brain magnetic resonance imaging examinations.

To address these limitations, the objective of this research is to develop and evaluate an unsupervised machine learning computational framework tailored for the automated detection and segmentation of demyelinating lesions in patients with multiple sclerosis using FLAIR magnetic resonance imaging. Instead of relying on data-hungry supervised deep learning architectures, the proposed methodology integrates a robust image processing pipeline—combining N4 bias field correction, skull stripping, and Z-score intensity normalization with K-means clustering for multi-tissue classification and localized statistical anomaly detection for focal plaque identification. This framework provides an accessible, fully automated tool capable of generating objective, real-time quantitative spatial metrics of lesion burden without requiring manually labeled ground-truth datasets. Ultimately, this approach aims to sup-

port specialists in longitudinal disease monitoring and clinical follow-up, particularly within public healthcare environments such as the Brazilian Unified Health System (Sistema Único de Saúde - SUS)

II. Method

This study was designed as a methodological investigation aimed at the development and evaluation of a computational system for medical image analysis using machine learning techniques applied to the detection and segmentation of demyelinating lesions in brain magnetic resonance imaging.

The dataset used consisted of brain MR examinations obtained from two independent sources. The first source included scans of patients diagnosed with multiple sclerosis who underwent clinical follow-up at the Hospital de Clínicas of the Federal University of Paraná (UFPR), located in Curitiba, Brazil. A total of 30 examinations from this institution were included in the study. Before inclusion in the study, all examinations were anonymized to ensure protection of patient identity and compliance with ethical standards for clinical data research. All identifying information contained in the imaging files was removed before data processing and analysis. The use of examinations obtained from the Hospital de Clínicas of the Federal University of Paraná was approved by the institutional Research Ethics Committee, in accordance with national ethical guidelines for research involving human subjects (CAEE 86677324.9.0000.0096 - Approval Number 7.582.336 in accordance with Resolution 466/2012 of the Brazilian National Health Council). The second source consisted of a publicly available international dataset provided by the Laboratory of Imaging Technologies of the University of Ljubljana, Slovenia, available online (http://lit.fe.uni-lj.si/tools). From this database, an additional 30 MRI examinations were selected.

Therefore, the final dataset comprised a total of 60 brain MRI examinations from patients diagnosed with multiple sclerosis. Only volumetric FLAIR images were used for algorithm development and training, due to their well-established sensitivity to targeting demyelinating lesions. The model was able to detect and segment demyelinating lesions on FLAIR images, generating masks that could be late compared to expert annotations.

To process these MRI acquisitions, a complete desktop-integrated pipeline was developed for automated preprocessing, multi-tissue segmentation, and quantitative lesion analysis tailored for FLAIR sequences. The image processing workflow ingests raw DICOM directories or NIfTI volumes, applying SimpleITK-based N4 Bias Field Correction to eliminate radiofrequency field inhomogeneities, followed by Otsu thresholding, 3D morphological operations, and largest-connected-component analysis for robust skull stripping.

After executing a voxel-wise Z-score normalization within the isolated brain mask, the pipeline feeds the data into a K-means clustering algorithm to segment the brain into distinct anatomical tissue classes—specifically suppressing cerebrospinal fluid (CSF) while isolating gray matter, normal white matter, and hyperintense white matter. Furthermore, it incorporates a dual-strategy anomaly detection mechanism that flags global statistical intensity outliers alongside a localized white matter hyperintensity detector to efficiently map lesions, MS plaques, and vasogenic edema while discarding microscopic noise using size-filtering thresholds.

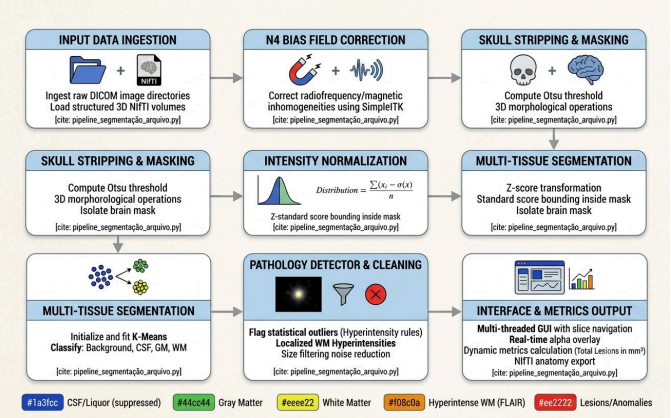


Fig. 1. Flowchart summarizing the unsupervised brain MRI segmentation pipeline. Image generated with the assistance of Imagen 3 (Google, 2026) using a text prompt detailing the step-by-step processing from raw DICOM/NIfTI ingestion, N4 bias field correction, skull stripping, Z-score normalization, to K-means multi-tissue classification and anomaly detection.

To facilitate clinical workflow and interpretation, the processing pipeline was coupled with a desktop Graphical User Interface (GUI) implemented in Python using Tkinter and Matplotlib. The interface features a side-by-side display that pairs original grayscale FLAIR MRI slices with customizable segmentation overlays. Clinicians can dynamically navigate through 3D volumes, adjust real-time opacity blending, and manipulate an interactive sensitivity slider that rescales Z-score intensity thresholds to refine lesion boundaries on-the-fly. In addition, the software automatically computes per-slice tissue statistics, quantifies total lesion volume in cubic millimeters, counts independent lesion clusters, and supports exporting segmented maps into standardized 3D NIfTI format.

III. Results

First of all, our MRI DICOM dataset was anonimized, and analysed by an specialist. We carried out reconstruction of images and data visualization using the 3D Slicer version 5.10.0 Software [13]. Figure 2 shows an original patient MRI DICOM file from our dataset, shown on three views : axial, sagital and coronal, and 3D reconstruction renderized volume. Figure 3 shows an original patient MRI file obtained from University of Ljubljana, Slovenia (available online (http://lit.fe.uni-lj.si/tools), shown on three views: axial, sagital and coronal ,and 3D reconstruction renderized volume. As shown, the brain images were cropped to exclude the nasal region, as the nasal mucosa exhibits high signal intensity that may be erroneously identified as hyperintense regions with signal characteristics similar to multiple sclerosis lesions or normal gray matter.

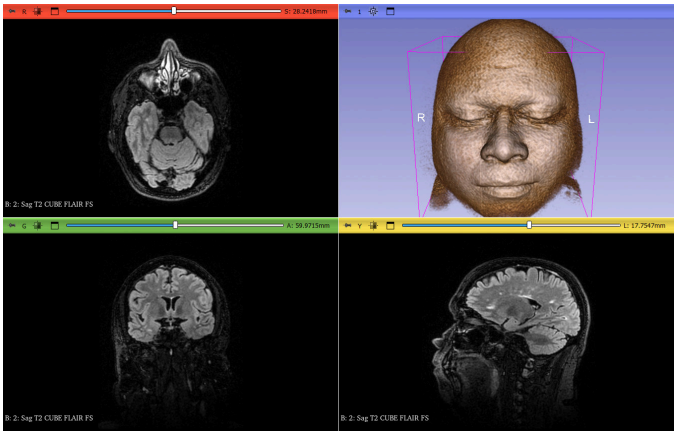


Fig. 2. MRI analysis presenting from volume renderization in 3D Slicer software-patient from our dataset.

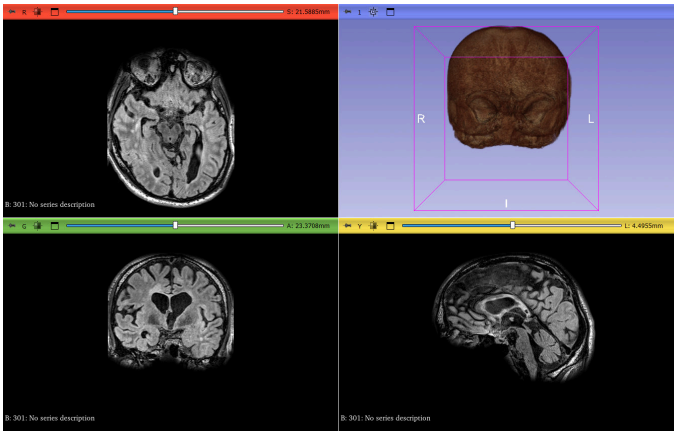


Fig. 3. 3D slicer volume renderization from an original patient MRI file obtained from University of Ljubljana, Slovenia.

Finally, Figure 4 shows data analysis run through the MRI tool segmentation, available in [14]. Patient selection, pipeline processing, lesions area and processing loading are shown on top bar. Control parameters such as : MRI slices, opacity, lesion sensibility, lesion and classes of tissues segmentation are shown on the right. This slice refers to the same patient shown on figure 2.

Average sensibility values and lesion area in mm3 for both datasets set by an specialist in our pipeline to obtain the most accurate lesion segmentation are shown on Table I.

TABLE I
STATISTICAL SUMMARY OF SENSITIVITY AND LESION VOLUME

Class Group	**Metrics and Standard Deviation**	
	Sensitivity	***Lesion Volume (mm^3)***
Hospital de Clínicas UFPR Dataset	0.283 ± 0.076	1637.42 ± 1056.58
University of Ljubljana Dataset	0.287 ± 0.064	1545.39 ± 1050.47

DISCUSSION

1) UI Description and Comparative Analysis: Custom Segmentation Viewer vs. 3D Slicer modules and Extentions [15]:

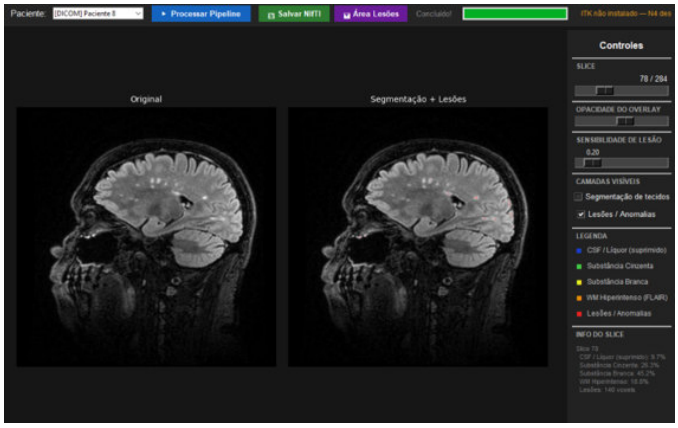


Fig. 4. Data analysis run through the MRI tool segmentation. Patient selection, pipeline processing, lesions area and processing loading are shown on top bar. Control parameters such as : MRI slices, opacity, lesion sensibility, lesion and classes of tissues segmentation are shown on the right. This slice refers to the same patient shown on figure 2

The custom application's Graphical User Interface (GUI) serves as an interactive clinical dashboard tailored specifically for brain MRI segmentation. Below is a technical description of its core functionalities and its comparative advantages over the generic multi-purpose output obtained from 3D Slicer.

2) Functional Description of the Custom Pipeline Viewer: The custom GUI is divided into an asynchronous top control bar, a dual-pane interactive canvas, and a right-hand parametric control panel:

- **Asynchronous Processing Top Bar:** It allows the user to select specific patient exams from organized DICOM or NIfTI directories. Clicking **"Processar Pipeline"** triggers a multi-stage background thread (N4 Bias Correction, Otsu skull stripping, and K-means clustering) while updating a live progress bar to prevent UI freezing. It also includes quick buttons to export the results directly into medical NIfTI files or generate standalone lesion statistics windows.
- **Dual-Pane Interactive Canvas:** The software renders a side-by-side comparative screen using Matplotlib. The left pane displays the original raw grayscale MRI slice (such as a sagittal FLAIR view showing hyperintense plaque regions), while the right pane dynamically projects the exact same slice with customizable segmentation overlays.
- **Parametric Right Control Panel:** * *Slice Navigation:* A precision slider allows the user to traverse through the entire 3D volume, with synchronized hotkey tracking (Arrow keys).
  - *Overlay Opacity:* Real-time alpha blending control lets clinicians adjust how transparently the tissue colors or lesion masks are mapped over the anatomical structure.
  - *Lesion Sensitivity Adjustment:* A critical real-time slider that inversely rescales the global Z-score and intra-White Matter standard deviation thresholds.

Moving this slider dynamically recalculates and filters lesion masks on the fly to fit a specific patient's contrast constraints.

- *Layer Visibility & Live Info:* Checkboxes isolate specific tissue classes or anomaly maps, while the bottom section computes live, voxel-by-voxel percentage distributions of CSF, Gray Matter, White Matter, and lesion load for the active slice.

3) Advantages over the 3D Slicer Output: While 3D Slicer is an industry-standard powerhouse for generalized multi-planar medical rendering and 3D volume reconstruction, the custom application offers distinct workflow and clinical advantages for this specific task:

a) Tailored Specialization for Multiple Sclerosis (MS) & FLAIR Sequences:

The Custom Application Advantage: The custom pipeline is clinically optimized for FLAIR sequences. Its entire color palette and mathematical logic are engineered to suppress CSF and isolate "Hyperintense White Matter". Instead of just masking a large general region of the brain, it flags focal periventricular and juxtacortical plaques—crucial for multiple sclerosis monitoring—highlighting them in bright, recognizable red over the structures.

b) Dynamic, Real-Time Pathology Tuning (Sensitivity Slider):

- **The 3D Slicer Limitation:** Refining a segmentation in 3D Slicer typically requires diving into complex multi-module parameters (such as Segment Editor or thresholding filters) and manually recalculating the label maps, which can be computationally slow and counter-intuitive for a quick clinical review [15].
- **The Custom Application Advantage:** The right panel's **"Sensibilidade de Lesão" (Lesion Sensitivity)** slider operates as an interactive thresholding tool. If the initial unsupervised calculation overestimates the multiple sclerosis lesions (due to artifact noise or blood vessel hyperintensities), the clinician can immediately adjust the slider to reduce sensitivity. The application instantly updates the viewport slice without requiring the user to recalculate the entire machine learning or backend pipeline, allowing immediate correction of overestimation errors.

c) Instant Quantitative Metrics and Focused Diagnostics:

- **The 3D Slicer Limitation:** To get volumetric metrics or area projections out of 3D Slicer, a user must add specific quantitative extensions, run statistical modules on selected segments, and parse broad spreadsheet tables. [15].
- **The Custom Application Advantage:** The custom GUI minimizes clicks. The **"Info do Slice"** section shows real-time tissue composition changes as you scroll. Clicking the **"Área Lesões"** button opens a focused dialog that translates voxel counts straight into total lesion volume in cubic millimeters and tracks individual lesion clusters sorted by size. It bridges the gap between deep image segmentation and immediate diagnostic feedback.

A central strength of this work lies in evaluating the unsupervised framework across a multicenter pilot dataset comprising 60 volumetric FLAIR MRI examinations sourced from two distinct clinical environments: a public tertiary hospital in Brazil (UFPR) and an international public repository (University of Ljubljana, Slovenia). By processing acquisitions from different scanners and protocols without requiring retraining or manual ground-truth annotations, the algorithm demonstrated high cross-institutional consistency. This stability is reflected in the comparable average lesion volumes ($1,637.42 \pm 1,056.58$ mm^3 for UFPR vs. $1,545.39 \pm 1,050.47$ mm^3 for Ljubljana) and sensitivity metrics presented in Table I. This capability addresses a well-documented challenge in automated neuroimaging, where inter-scanner heterogeneity frequently degrades segmentation reliability [16] [10].

Compared to established unsupervised toolboxes such as the Lesion Growth Algorithm (LGA) and Lesion Segmentation Tool (LST) proposed by Roura et al. [12][14] and Schmidt et al. [8][10], which rely on intensity thresholding within SPM environments, our pipeline combines SimpleITK-based N4 bias correction with dynamic Z-score outlier filtering. While supervised deep learning architectures (e.g., 3D U-Nets) achieve high segmentation accuracy on standardized benchmarks [4], [16] they require extensive, pixel-wise expert annotations that are difficult to scale in daily practice. The proposed framework bridges this gap by offering instant, deterministic multi-tissue segmentation without training overhead, presenting an accessible and scalable tool well-suited for public healthcare infrastructures like the Brazilian Unified Health System (Sistema Único de Saúde - SUS).

The volumetric output generated by the software (averaging $1,545.39$ mm^3) aligns with expected clinical lesion burdens for patients with mild-to-moderate MS under active neurological surveillance ($1,000$ mm^3 to $4,000$ mm^3) [1] [7]. Importantly, computing integrated 3D spatial volumes avoids the overestimation typical of simple 2D slice summation. A recognized limitation of this pilot study is the sample size and the reliance on volumetric yield and sensitivity for validation. Future research will expand the validation cohort across additional centers and incorporate spatial overlap metrics—such as the Dice Similarity Coefficient (DSC) and Lesion False Positive Rate (LFPR)—to enable direct benchmarking against international MS challenges [16]. Furthermore, developing longitudinal registration tools will allow automated tracking of new, enlarging, or regressing plaques over serial MRI studies.

ACKNOWLEDGEMENTS

Authors would like to thank Fundação Araucária for the support, and engineer Hellen Rutes for technical discussions.

REFERENCES

[1] T. R. Lim, O. Dörtkol, M. A. Rocca, M. Filippi, M. P. Wattjes, P. Sati *et al.*, "Conventional and Emerging MRI Biomarkers of Multiple Sclerosis," *Radiographics*, vol. 46, no. 3, 2026.

[2] J. H. Noseworthy, C. Lucchinetti, M. Rodriguez, and B. G. Weinshenker, "Multiple Sclerosis," *New England Journal of Medicine*, vol. 343, no. 13, pp. 938–952, 2000.

[3] C. Walton, R. King, L. Rechtman, W. Kaye, E. Leray, R. A. Marrie, N. Robertson, N. La Rocca, B. Uitdehaag, I. van der Mei, M. Wallin, A. Helme, C. Angood Napier, N. Rijke, and P. Baneke, "Rising Prevalence of Multiple Sclerosis Worldwide: Insights from the Atlas of MS, Third Edition," *Multiple Sclerosis Journal*, vol. 26, no. 14, pp. 1816–1821, 2020.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham: Springer, 2015, pp. 234–241.

[5] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon *et al.*, "Efficient Multi-scale 3D CNN With Fully Connected CRF for Accurate Brain Lesion Segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[6] S. Valverde, M. Cabezas, E. Roura, M. González-Vallinas, D. Pareto, J. C. Vilanova *et al.*, "Improving Automated Multiple Sclerosis Lesion Segmentation With a Cascaded 3D Convolutional Neural Network Approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[7] Y. Ma, C. Zhang, M. Cabezas, Y. Song, Z. Tang, D. Liu *et al.*, "Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images: Techniques and Clinical Applications," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2680–2692, 2022.

[8] P. Schmidt, V. Pongratz, P. Küster, D. Meier, J. Wuerfel, C. Lukas *et al.*, "Automated Segmentation of Changes in FLAIR-Hyperintense White Matter Lesions in Multiple Sclerosis on Serial Magnetic Resonance Imaging," *NeuroImage: Clinical*, vol. 23, p. 101849, 2019.

[9] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal *et al.*, "Deep Learning: A Primer for Radiologists," *Radiographics*, vol. 37, no. 7, pp. 2113–2131, 2017.

[10] P. M. Cheng, E. Montagnon, R. Yamashita, I. Pan, A. Cadrin-Chênevert, F. Perdigón Romero *et al.*, "Deep Learning: An Update for Radiologists," *Radiographics*, vol. 41, no. 5, pp. 1427–1445, 2021.

[11] P. V. Dos Santos, M. S. R. Martins, S. N. Amorim, C. Gonçalves, and R. M. Loureiro, "Unsupervised Model for Structure Segmentation Applied to Brain Computed Tomography," *PLoS ONE*, vol. 19, no. 6, p. e0304017, 2024.

[12] E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J. C. Vilanova *et al.*, "A Toolbox for Multiple Sclerosis Lesion Segmentation," *Neuroradiology*, vol. 57, no. 10, pp. 1031–1043, 2015.

[13] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, 2012, quantitative Imaging in Cancer. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0730725X12001816

[14] C. Gonçalves, "pythonProjectUnsupervisedSegmentationBrainMRI_HC_Curitiba: PhD project code," https://github.com/cristhianeg/pythonProjectUnsupervisedSegmentationBrainMRI_HC_Curitiba, 2026, accessed: 2026-07-07.

[15] CSIM-Toolkits, "Slicer-LesionSpotlightExtension: 3D Slicer extension that provides T2-FLAIR MRI manipulation methods in order to increase contrast and segment hyperintense Multiple Sclerosis lesions," https://github.com/CSIM-Toolkits/Slicer-LesionSpotlightExtension, 2026, acesso em: 7 jul. 2026.

[16] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Améli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific Reports*, vol. 8, no. 1, p. 13650, 2018.

Computer Vision System for Eye Movement and Eyelid Dynamics Analysis

1st Stephanie da Silva Pereira
Department of Biomedical Engineering
UFRN
Natal, Brazil
0009-0008-9518-8697

2nd José Carlos Gomes da Silva
Department of Physical Education
UFRN
Natal, Brazil
0000-0003-3192-7223

3rd Mário André da Cunha Espada
Higher School of Education
Polytechnic Institute of Setúbal
Setúbal, Portugal
0000-0002-4524-4784

4th Paulo Moreira Silva Dantas
Department of Physical Education
UFRN
Natal, Brazil
0000-0002-9217-7107

5th Heliana Bezerra Soares
Department of Biomedical Engineering
UFRN
Natal, Brazil
0000-0002-0833-7906

***Abstract* – Eye movement and eyelid dynamics analysis plays an important role in Biomedical Engineering; however, access to assessment tools that combine low cost and high precision remains a challenge. In this context, this work aimed to develop a non-invasive eye monitoring system capable of processing conventional videos obtained from image capture devices. The system was developed in Python using the OpenCV and dlib libraries, allowing real-time facial mapping through 68 reference points. Among the implemented functionalities, the following stand out: dimensional calibration based on pupillary distance, continuous calculation of the Ocular Aspect Ratio (EAR) for automatic blink detection, and iris trajectory tracking, with the generation of structured reports for analysis. The tests performed demonstrated that the system was able to process conventional videos and continuously extract the proposed biometric metrics, making them available in structured reports for later analysis. The results obtained indicate that the initial functional validation demonstrated the feasibility of the pipeline's computational architecture for extracting structured metrics. As this is a technical proof of concept focused on software development, the validation of biomedical performance, the generalization of the method, and its clinical applicability depend on future studies with larger sample sizes and ethical approval. The developed software obtained registration as a Computer Program with the National Institute of Industrial Property (INPI), consolidating itself as a technological product resulting from this work.**

Keywords: Computer Vision; Eye Movement; Eyelid Dynamics; Eye Tracking; Biomedical Engineering.

I. Introduction

The continuous assessment of eyelid dynamics and eye movement provides relevant biomarkers for the investigation of cognitive disorders, neurodegenerative diseases and fatigue states. Traditionally, the quantitative extraction of these metrics depends on high-cost eye tracking equipment, often based on infrared sensors [1] and dedicated hardware. This technological and financial requirement restricts the application of these analyzes to laboratory environments and large clinics, limiting the population's access to preventive assessments and making it difficult to expand remote monitoring.

In this scenario, the integration of computer vision with digital health emerges as a promising technological alternative to overcome these barriers. Applying image processing algorithms to conventional videos enables the non-invasive analysis of eye movement, eliminating the need for specific equipment and expanding the possibilities for using the technology in different clinical and research contexts. This software-based approach facilitates the acquisition of biometric data using conventional RGB (Red, Green, Blue) cameras [2].

Although computer vision-based solutions for eye analysis exist, many focus on specific functionalities or rely on particular experimental configurations, highlighting the need for systems that integrate different biometric metrics into a single, low-cost application.

Given this scenario, this work presents the development of a computer vision system for analyzing eye movement and eyelid dynamics. Unlike approaches focused on isolated functionalities, the system integrates, in a single application, techniques for dimensional calibration, calculation of the Eye Aspect Ratio (EAR), iris tracking, and automatic generation of structured reports, using conventional videos obtained by RGB cameras.

II. Materials and Methods

A. General system architecture

The proposed system was developed with the objective of analyzing eye movement and eyelid dynamics from conventional videos, using computer vision techniques for the automatic extraction of metrics related to eye movement and eyelid dynamics. The architecture was structured in a modular way, encompassing the stages of data acquisition and calibration, facial detection, extraction of biometric metrics, and generation of results. This organization allows for the sequential pro-

cessing of information, from video reading to data export for analysis. Fig. 1 shows the general system architecture and the implemented processing flow.

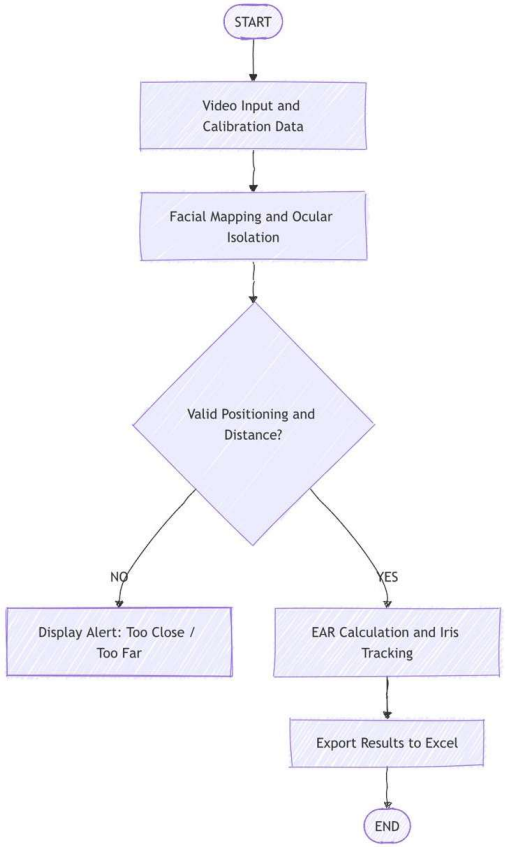


Fig. 1 – Flowchart of the system's architecture and general operation.

As illustrated in Fig. 1, the system's operation is organized into four main modules. Initially, video acquisition and calibration of the parameters necessary for processing are performed. Next, the system executes facial detection and ocular region localization, allowing the extraction of biometric metrics related to eye movement and eyelid dynamics. Finally, the results obtained are organized and exported into structured reports for later analysis.

B. Data acquisition and calibration

The system was developed to process conventional videos containing the user's face, captured by devices equipped with RGB cameras, such as webcams and smartphones. Before processing begins, the necessary parameters for system calibration are provided, including user identification, the video file, the approximate distance between the face and the camera, and the actual pupillary distance, expressed in millimeters. This information is used to establish the spatial scale of the image, allowing the conversion of measurements obtained in pixels to values in millimeters during analysis.

Dimensional calibration is an essential step in processing, as it allows measurements extracted from images to be represented in physical units. To achieve this, the system uses the actual pupillary distance provided by the user as a reference to determine the conversion factor between pixels and millimeters, reducing the exclusive reliance on relative measurements and allowing the acquisition of quantitative metrics related to eyelid opening.

C. Facial detection and localization of eye regions

After the calibration stage, each video frame is processed for face identification and eye region localization. Initially, the frames are converted to grayscale, reducing computational complexity and improving the performance of facial detection algorithms. Then, face identification is performed using the dlib library [3], widely used in computer vision applications for providing highly efficient facial detectors and anatomical landmark estimators [4].

After face detection, the system performs facial mapping using 68 anatomical landmarks, automatically identifying facial structures such as eyebrows, nose, mouth, and eyes. For this task, points corresponding to the ocular regions are used, serving as the basis for calculating the Eye Aspect Ratio (EAR) and for tracking the iris trajectory in subsequent processing steps.

D. Extraction of eye movement and eyelid dynamics metrics

After locating the eye regions, the system automatically extracts the metrics used to characterize eye movement and eyelid dynamics. Eye region location is performed using the facial landmark estimation model shape predictor 68 face landmarks, which identifies six two-dimensional reference points for each eye. Eyelid dynamics are quantified using the Eye Aspect Ratio (EAR), calculated from the geometric relationship between the vertical and horizontal Euclidean distances of the points that delimit the eye region. Eq. (1) presents the expression used to calculate this metric.

$$EAR = \frac{||p_2 - p_6|| + ||p_3 - p_5||}{2 \cdot ||p_1 - p_4||} \quad (1)$$

In this expression, P_1 a P_6 represent the facial reference points obtained during the mapping of the ocular region. The vertical distances are compared to the horizontal distance of the eye, producing an approximately constant value when the eye remains open and significantly reduced during eyelid closure.

The temporal variation of the EAR throughout processing allows for the automatic identification of blink events. Fig. 2 shows the temporal evolution of the Ocular Aspect Ratio, highlighting the threshold used for the automatic detection of eyelid closure.

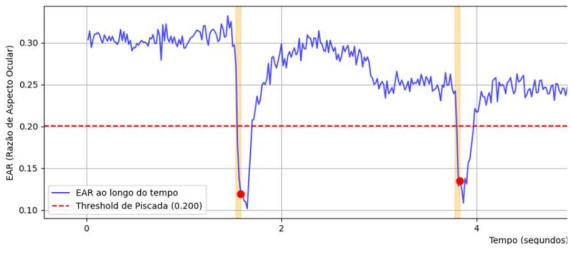


Fig. 2 – Temporal evolution of the ocular aspect ratio with blink threshold.

As observed in Fig. 2, EAR values remain relatively stable during periods when the eyes are open, showing sharp reductions at times corresponding to eyelid closure. The threshold defined during the calibration step is used to automatically

identify these events, allowing the detection of blinks throughout video processing.

E. Generation and export of results

After extracting the metrics related to eye movement and eyelid dynamics, the information obtained during processing is organized into a tabular structure, allowing for the sequential storage of data for each frame of the video. This structure, implemented using the Pandas library, gathers the parameters calculated throughout the analysis, preserving the temporal correspondence between the extracted metrics and facilitating their subsequent processing [5].

At the end of the analysis, the system automatically consolidates the information obtained, generating graphical representations of the temporal evolution of the Ocular Aspect Ratio (OAR), eyelid opening, and iris trajectory. The data are also exported in spreadsheet format (xlsx), allowing for organization, consultation, and use in subsequent analyses, as well as facilitating the documentation and reproducibility of the results obtained.

F. Spatial Iris Tracking

Spatial tracking of the iris is performed by applying an elliptical mask over the Region of Interest (ROI) to suppress peripheral noise and isolate the pupil region. Subsequently, the Canny Filter is used to map the edges, and the Hough Circular Transform is used to identify the iris contour and determine its geometric center [6]. The combination of these techniques allows for the continuous extraction of the iris's spatial coordinates (X, Y) in each processed frame. Fig. 3 and 4 show, respectively, the trajectories of the iris coordinates in horizontal and vertical directions.

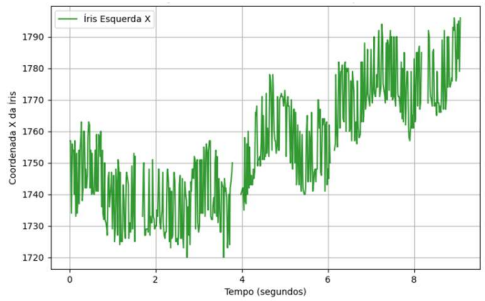


Fig. 3 – Spatial trajectory of the X and Y coordinates of the iris in the left eye in the horizontal position.

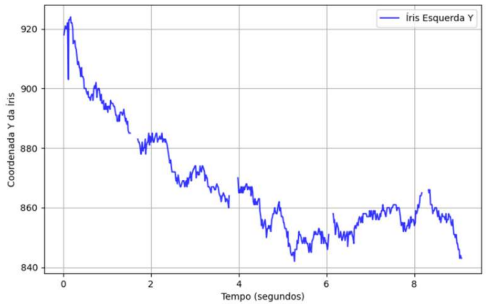


Fig. 4 – Spatial trajectory of the X and Y coordinates of the iris in the left eye in the vertical position.

Trajectory analysis makes it possible to characterize iris displacements over time, complementing the information obtained from the Ocular Aspect Ratio and allowing an integrated assessment of eye movement and eyelid dynamics.

G. Generation and export of results

After extracting the metrics related to eye movement and eyelid dynamics, the data obtained during processing are organized into a tabular structure, implemented using the Pandas library. This structure sequentially stores the information for each processed frame, preserving the temporal correspondence between the calculated metrics and facilitating their consolidation for later analysis.

Upon completion of processing, the system automatically exports the data to a spreadsheet in .xlsx format, containing the metrics extracted throughout the analysis. This organization enables structured documentation of the results, facilitating their consultation, reproducibility, and use in research activities and support for clinical evaluation.

IV. Results And Discussion

The experimental validation of the system was performed using a primary database composed of videos of a healthy adult volunteer, obtained in a real-world capture environment. The processing was performed automatically, allowing for facial detection, ocular region localization, and continuous extraction of metrics related to eyelid dynamics and eye movement.

In the analysis of eyelid dynamics, the system continuously calculated the Eye Aspect Ratio (EAR), automatically identifying the baseline value of the metric and determining the threshold used for blink detection. During processing, a sharp reduction in EAR was observed at the instants corresponding to eyelid closure, allowing for the automatic identification of these events and the extraction of the respective image segments. Fig. 5 shows the record of the first blink detected automatically by the system.

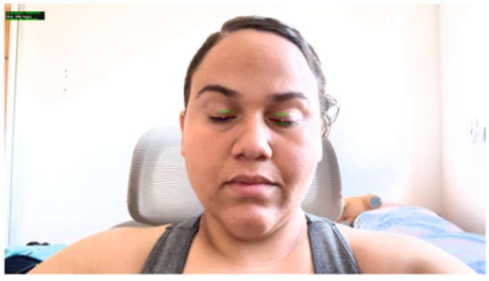


Fig. 5 – Visual cutout of automatic detection of the first blink during processing.

The temporal behavior of the Ocular Aspect Ratio for the right and left eyes is shown in Fig. 6. It can be observed that the EAR values remain relatively stable during periods of eye opening, showing simultaneous reductions during blinks, when the automatic threshold is exceeded. Independent analysis of the two eyes demonstrates the consistent functioning of the detection algorithm throughout the processing.

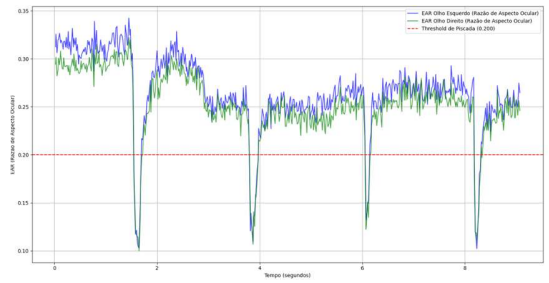


Fig. 6 – Temporal evolution of the Ocular Aspect Ratio (OAR) for the right and left eyes, highlighting the automatic blink threshold.

Spatial calibration allowed the conversion of measurements originally obtained in pixels to physical units (millimeters), making it possible to quantitatively represent the vertical opening of the eyelids throughout the image sequence. During rest periods, values close to 8 mm of eyelid opening were observed, with progressive reductions during blinking. Fig. 7 shows the continuous evolution of these measurements for the left eye (Fig. 7(a)) and right eye (Fig. 7(b)), demonstrating the system's ability to track dimensional variations in eye opening throughout the video.

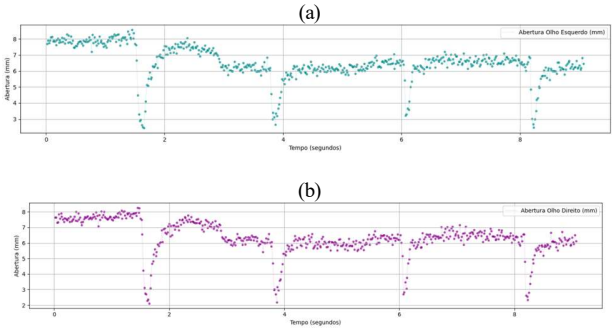


Fig. 7 – Continuous vertical opening of the eyelids (in millimeters) extracted by the system: (a) left eye and (b) right eye.

In addition to eyelid dynamics, the system performed continuous iris tracking using the Hough Circular Transform, allowing the extraction of spatial coordinates (X, Y) throughout the entire image sequence. The horizontal and vertical trajectories obtained made it possible to quantitatively represent eye movement independently for each eye, complementing the information provided by the EAR analysis.

Upon completion of processing, all extracted metrics—including EAR values, iris spatial coordinates, and eyelid opening measurements in millimeters—were automatically organized and exported to a spreadsheet in .xlsx format. This structure preserves the temporal sequence of the obtained data, facilitating its documentation, statistical analysis, and use in experimental studies in the field of Biomedical Engineering.

The results obtained demonstrate that the integration between the calculation of the Ocular Aspect Ratio (EAR), dimensional calibration in millimeters, and spatial tracking of the iris allowed for the simultaneous extraction of different metrics related to eyelid dynamics and eye movement. The automatic organization of this information in a structured format expands the possibilities of using the tool in experimental studies in the field of Biomedical Engineering, providing quantitative data that can support future research and applications to aid clinical evaluation.

Although the pipeline has proven fully functional, the current system operates as a technical proof of concept with limitations to be overcome. Dimensional calibration, for example, is susceptible to scaling variations if unpredictable facial displacements occur in the depth axis. Similarly, environmental factors, such as light reflections and partial eyelid occlusions, pose challenges to the optimal extraction of coordinates by the Hough Transform. Consequently, the validation of parameters of clinical interest requires additional experimental steps. The future timeline of this project will include ethics review (CEP) for the inclusion of a diverse sample—encompassing different age ranges, phenotypic variations, and neurological conditions—as well as statistical performance validation. It is planned to measure the spatial error of iris tracking using a commercial eye tracker as the gold standard, in addition to verifying the sensitivity and specificity of the EAR algorithm through cross-validations with independent manual markings (ground truth).

V. Conclusions

This work achieved the objective of developing a non-invasive and low-cost computational system for the analysis of eye movement and eyelid dynamics from conventional videos. The proposed solution integrated computer vision techniques to automatically perform facial detection, calculation of the Eye Aspect Ratio (EAR), spatial tracking of the iris, and dimensional calibration of measurements in millimeters. The functional validation performed demonstrated the system's ability to extract, organize, and export quantitative metrics related to eye behavior, bringing together different functionalities in a single computational platform.

Although the results obtained demonstrate the proper functioning of the proposed architecture, some limitations should be considered. The system's performance is influenced by image acquisition conditions, such as inadequate lighting, intense reflections, and video quality, factors that can affect the accuracy of iris and ocular structure detection. Furthermore, the dimensional calibration process still depends on the user's manual input of the pupillary distance. The experimental evaluation was conducted using videos from a single healthy adult volunteer, representing an initial stage of functional validation. Therefore, studies with a larger number of participants will be necessary to evaluate the system's performance under different acquisition conditions and anatomical characteristics.

Looking to the future, the intention is to incorporate automatic depth estimation or stereo vision methods to eliminate the need for manual pupillary distance calibration. The integration of Machine Learning techniques for identifying patterns associated with ocular changes is also planned, as well as adapting the system for mobile devices, allowing for real-time metric processing. Additionally, future studies should consider validating the system in a broader population, including participants with different anatomical characteristics and clinical conditions, expanding the evidence on its applicability in research and in clinical assessment support scenarios.

As a result of technological development, the system obtained the Computer Program Registration Certificate from the National Institute of Industrial Property (INPI), under process number BR512026004319-0, issued on June 16, 2026, with the Federal University of Rio Grande do Norte (UFRN) as the rights holder. This registration demonstrates the originality of the developed solution and reinforces its contribution as a technological product for Biomedical Engineering. Thus, the developed system consolidates itself as a computer vision platform for analyzing eye movement and eyelid dynamics, integrating eye tracking techniques, dimensional calibration, and structured data organization, with the potential to support future research and applications in Biomedical Engineering.

ACKNOWLEDGMENTS

The authors thank Laboratory for Medical Signal and Image Acquisition and Processing - LAPSIM and the Open Innovation (OI) Laboratory of the Biomedical Engineering program at the Federal University of Rio Grande do Norte (UFRN) for the technical support, infrastructure, and assistance provided for the development of this work.

REFERENCES

[1] Riess, A. B.; Gabriel, R. Perspectivas de pesquisa com a metodologia de Eye tracking em leitura e cognição no Brasil. Odisseia, Natal, v. 5, n. 1, p. 103-119, 2020.

[2] Lesport, Q. et al. Eye Segmentation Method for Telehealth: Application to the Myasthenia Gravis Physical Examination. Sensors, v. 23, n. 18, p. 7744, 2023.

[3] King, D. E. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research, v. 10, p. 1755-1758, 2009.

[4] Soukupová, T.; Čech, J. Real-Time Eye Blink Detection using Facial Landmarks. In: 21st Computer Vision Winter Workshop (CVWW). Rimske Toplice, 2016. p. 1-8.

[5] Bradski, G. The OpenCV Library. Dr. Dobb's Journal of Software Tools, v. 25, p. 120-123, 2000.

[6] Liu, Z. et al. Auxiliary Diagnosis of Children With Attention-Deficit/Hyperactivity Disorder Using Eye-Tracking and Digital Biomarkers. JMIR mHealth and uHealth, v. 12, e58927, 2024.

Computerized Tomography Image Generation for Ischemic Stroke Using GANs

1st Emanuel T. de A. Silva
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
emanuel.thiago.2018@gmail.com

2st João N. S. de Lima
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
joao.neto.4@academico.ifpb.edu.br

3th Gabriel A. de Santana
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
andrade.santana@academico.ifpb.edu.br

4nd Pedro A. de A. da Silva
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
pedrotatico1321@gmail.com

5st Villeneve de O. Soares
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
villeneveoliveira@gmail.com

6th Carlos D. M. Regis
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—Ischemic stroke is one of the leading causes of death and disability worldwide, requiring early diagnosis. The scarcity of annotated medical images limits the performance of deep learning models. This study investigates the use of Generative Adversarial Networks (GANs) to synthesize computed tomography (CT) images and evaluates whether these images improve an Xception-based stroke classifier. The generated samples obtained an SSIM of 0.80 and an FID of 6.5. Under five-fold stratified cross-validation at the slice level, combining classical augmentation with GAN-generated images increased the aggregated Xception accuracy from 88.12% to 91.99% and sensitivity from 86.19% to 91.71%, while reducing false negatives from 25 to 15. These results indicate a potential benefit from synthetic augmentation; however, they are descriptive because the experiment did not employ patient-level separation or an independent test set.**

***Index Terms*—GAN, images, stroke, ischemic**

I. Introduction

Stroke is among the leading causes of mortality and long-term disability worldwide, representing a major public health challenge. According to the World Health Organization, millions of individuals suffer a stroke every year, often resulting in severe neurological impairment or death. Early diagnosis and timely intervention are critical factors for improving patient outcomes and reducing the burden associated with the disease.

Approximately 87% of all stroke cases are ischemic, occurring when cerebral blood flow is interrupted due to vascular occlusion [1]. The effectiveness of treatment is strongly dependent on the time elapsed since symptom onset, making rapid and accurate diagnosis essential [2]. Ischemic stroke (IS) is commonly categorized into acute, subacute, and chronic stages according to its temporal evolution [3].

Computed tomography (CT) and magnetic resonance imaging (MRI) are the most frequently used imaging modalities for stroke assessment [4]. Although MRI generally provides greater sensitivity for detecting ischemic lesions in the early stages, CT remains the most widely adopted imaging technique due to its lower cost, faster acquisition time, and broader availability in clinical settings [5]. However, ischemic lesions often appear as subtle hypodense regions in CT images, making their identification challenging, particularly during the acute phase.

To address these limitations, computer-aided diagnosis systems based on machine learning and deep learning have been increasingly investigated. Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in medical image classification and segmentation tasks. Nevertheless, the development of robust deep learning models is often hindered by the limited availability of annotated medical datasets, especially in specialized clinical domains such as stroke imaging.

Recent advances in Generative Artificial Intelligence (Generative AI) have introduced promising solutions for mitigating data scarcity. In particular, Generative Adversarial Networks (GANs) have emerged as one of the most effective approaches for generating realistic synthetic medical images. Comprehensive reviews have highlighted the growing adoption of GANs in medical image analysis, including image synthesis, segmentation, classification support, and data augmentation [6].

Motivated by these advances, this work investigates the use of GAN-generated synthetic CT images of ischemic stroke to augment training datasets and improve the performance of a CNN-based classification model. By combining synthetic image generation with transfer learning, the proposed approach seeks to alleviate dataset limitations and contribute to more reliable computer-aided stroke diagnosis.

II. Related Work

The application of generative models in medical imaging has received considerable attention in recent years. A comprehensive review by Yi et al. [6] highlighted the growing role of deep learning and Generative Adversarial Networks in medical image analysis, demonstrating their effectiveness in

image synthesis, segmentation, reconstruction, and classification support.

Several studies have investigated GAN-based image synthesis for neurological disorders. Sun et al. [7] proposed a multimodal framework combining GANs and 3D residual U-Nets for brain tumor and stroke image generation and lesion segmentation. Their results demonstrated improvements in segmentation accuracy, highlighting the benefits of synthetic image generation for medical image analysis.

Li et al. [8] explored the generation of synthetic MRI images from CT scans of ischemic stroke patients using GANs. Their approach improved lesion visualization and achieved a Peak Signal-to-Noise Ratio (PSNR) of 24.4 dB and a Structural Similarity Index Measure (SSIM) of 0.78. Similarly, studies investigating CT-based synthetic MRI generation for acute ischemic stroke have shown that GAN-generated images can improve lesion detectability and support clinical decision-making by enhancing image contrast and anatomical detail.

GANs have also been employed directly in ischemic stroke lesion modeling and segmentation. Ghnemat et al. [9] proposed a mutation model combined with a GAN architecture for ischemic stroke lesion segmentation, demonstrating that synthetic lesion generation can improve segmentation robustness and lesion representation. These findings suggest that generative models are capable of capturing clinically relevant stroke patterns and anatomical variability.

Despite the progress achieved by previous studies, most existing works focus on image synthesis, lesion segmentation, or classification as isolated tasks. Comparatively fewer studies investigate the integration of GAN-based synthetic image generation with transfer learning classifiers for ischemic stroke detection in CT images. Therefore, this work proposes the generation of realistic synthetic ischemic stroke CT images using GANs and evaluates their impact on the performance of an Xception-based classification model, aiming to improve generalization and diagnostic accuracy in scenarios with limited data availability.

III. Methodology

The workflow of the proposed approach is illustrated in Fig. 1 and Fig. 2. First, the GAN is trained to generate realistic synthetic CT images from latent noise vectors through an adversarial process involving the Generator and the Discriminator, as illustrated in Fig. 1. Subsequently, the generated synthetic images are combined with real CT images to train and evaluate an Xception-based classifier using transfer learning. The corresponding classification workflow is presented in Fig. 2 and includes dataset balancing, image preprocessing, prediction, and performance evaluation.

A. Dataset Used

The database used in this project consists of computed tomography (CT) images of the head from 292 patients. This database was provided by the Center for Imaging Studies and Research (NEPI) at the Federal University of Paraíba (UFPB), in accordance with the ethics committee's approval number

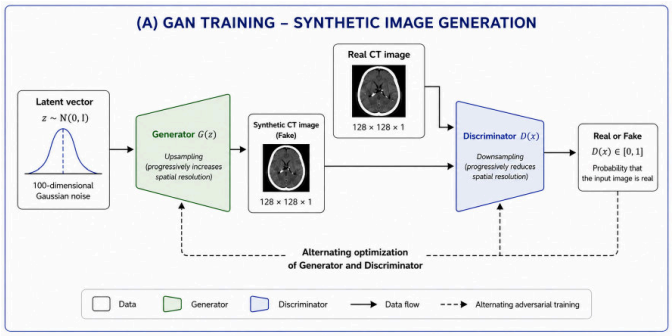


Fig. 1. Workflow of the GAN training process for synthetic CT image generation.

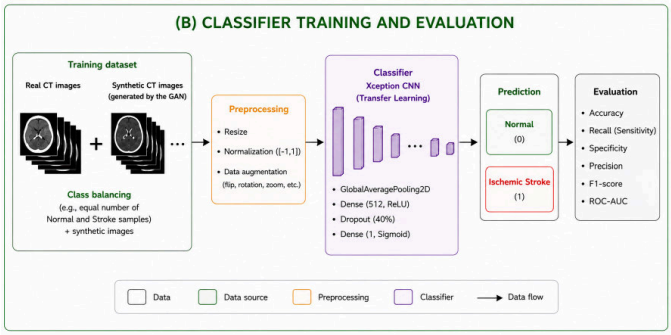


Fig. 2. Workflow of the classifier training and evaluation pipeline using real and GAN-generated CT images.

1,293,095. All images comply with the DICOM standard for medical imaging and have original dimensions of 512×512.

Each patient's CT scan contains approximately 20 images with an 8-mm spacing between the axial slices of the head. The images of these patients were obtained at random from a clinic specializing in diagnostic imaging.

The reports for each patient were prepared by a physician specializing in radiology, who marked on the CT images of the skull the gray-scale intensity levels characteristic of areas of chronic, acute, and subacute ischemic stroke, so that stroke detection could be performed.

Each scan consists of multiple axial slices, so the complete set of 292 patients resulted in a total of 11,142 two-dimensional images. Fig. 3 presents a summary of the process of compiling the database used in this study, from the scans obtained from the 292 patients to the acquisition of the images used to train the GAN. After analyzing the medical reports and identifying the regions of interest, the images were divided into 10,821 images classified as normal and 322 images showing signs of ischemic stroke, with the latter subdivided into 186 subacute cases, 62 acute cases, and 74 chronic cases.

Each scan consists of multiple axial slices, so the complete set of 292 patients resulted in a total of 11,142 two-dimensional images. Figure 3 summarizes the database construction workflow. Starting from the 292 head CT examinations, each scan was decomposed into axial slices, resulting in 11,142 two-dimensional images. After reviewing the radiology

TABLE I
SUMMARY OF RELATED WORKS INVOLVING GAN APPLICATIONS IN MEDICAL IMAGING.

Reference	Modality	Objective	Application	Metrics	Data Aug.
[7]	MRI/CT	Multimodal synthesis and segmentation	Lesion segmentation	Accuracy, HD	No
[10]	CT → MRI	Cross-modality conversion	Ischemic lesion segmentation	-	No
[8]	CT → MRI	MRI synthesis from CT	Lesion visualization	PSNR, SSIM	No
[11]	CT	Synthetic lesion generation	Liver tumor segmentation	Sensitivity, Precision	Yes
This Work	CT	Synthetic ischemic stroke image generation	Quality assessment of generated images	SSIM, FID	Yes

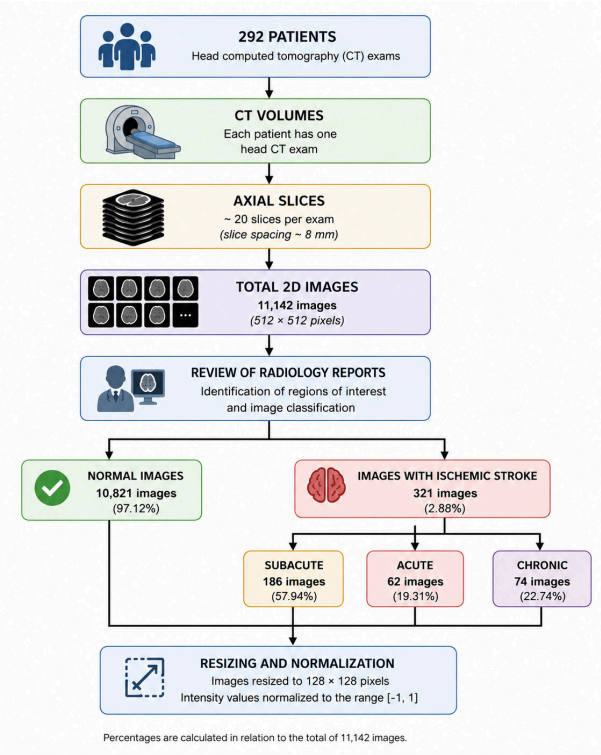


Fig. 3. Composition workflow of the database used in this study.

reports and identifying the regions of interest, the images were classified as either normal or containing ischemic stroke. The ischemic stroke images were further categorized according to the disease stage into subacute, acute, and chronic cases. Finally, all images were resized from 512×512 to 128×128 pixels and normalized before being used for GAN training.

Although the original images have a resolution of 512×512 pixels, they were resized to 128×128 pixels due to computational limitations and the high cost associated with training Adversarial Generative Networks. This reduction shortens training time and decreases the amount of memory required, making it feasible to use the model under the available computational conditions. However, the reduction in resolution may result in a loss of fine details and the smoothing of lesion edges, affecting the visual quality of the synthesized images. Therefore, the results obtained should be interpreted while considering the trade-off between computational cost and the preservation of the anatomical features present in the original images.

B. GAN Architecture

Generative Adversarial Networks (GANs) are a class of deep learning models consisting of two neural networks that compete with each other: the generator and the discriminator. The generator aims to create synthetic images that resemble the real images in the dataset, while the discriminator attempts to distinguish between real images and those generated artificially. During training, the generator learns to produce increasingly realistic images, while the discriminator improves its ability to differentiate between real and fake images [12].

To train the Generative Adversarial Network (GAN), an image dataset was used following the preprocessing steps described below. To facilitate training and stabilize the data distribution, the images were normalized to the interval [-1,1] using Equation 1. Where, I represents the pixel intensity in the original image.

$$I_{norm} = \frac{I}{127.5} - 1 \tag{1}$$

The GAN's input signals were organized into batches of 128 images to ensure diversity during training. The GAN architecture developed in this study consists of two models: a Generator and a Discriminator. The Generator is responsible for synthesizing realistic images from a random noise vector, while the Discriminator distinguishes between real and generated images.

The Generator receives as input a random noise vector $z \in \mathbb{R}^{100}$, which is initially projected by a dense layer onto a $4 \times 4 \times 256$ tensor. This tensor is then reshaped by a *Reshape* layer and progressively expanded using transposed convolution layers (*Conv2DTranspose*), combined with batch normalization (*Batch Normalization*) and *LeakyReLU* activation functions. This process of gradually increasing the resolution allows for the synthesis of computed tomography images of size $128 \times 128 \times 1$ with normalized values in the range ([-1,1]). The complete architecture of the Generator is presented in Table II.

The Discriminator follows a progressive dimension-reduction approach, using convolutional layers (*Conv2D*) combined with *LeakyReLU* activations. The model's goal is to

TABLE II
GENERATOR ARCHITECTURE

Type	Filters	Kernel	Stride	Activation
Dense(Input)	(4,4,256)	-	-	-
Reshape	-	-	-	-
Conv2DTranspose	256	3×3	2×2	LeakyReLU(0.2)
BatchNormalization	-	-	-	-
Conv2DTranspose	128	3×3	2×2	LeakyReLU(0.2)
BatchNormalization	-	-	-	-
Conv2DTranspose	128	3×3	2×2	LeakyReLU(0.2)
BatchNormalization	-	-	-	-
Conv2DTranspose	128	3×3	2×2	LeakyReLU(0.2)
BatchNormalization	-	-	-	-
Conv2DTranspose	1	3×3	1×1	Tanh

learn to distinguish between real images and synthetic images generated by the Generator. The detailed architecture of the Discriminator is presented in Table III. To facilitate visualization of the processing flow and the main layers used in each model, Figure 4 presents a schematic representation of the GAN architecture used in this work.

TABLE III
DISCRIMINATOR ARCHITECTURE

Type	Filters	Kernel	Stride	Activation
Input	-	-	-	-
Conv2D	64	3×3	2×2	LeakyReLU(0.2)
Conv2D	128	3×3	2×2	LeakyReLU(0.2)
Conv2D	128	2×2	2×2	LeakyReLU(0.2)
Conv2D	128	3×3	2×2	LeakyReLU(0.2)
Flatten	-	-	-	-
Dropout	-	-	-	0.5
Dense	1	-	-	Sigmoid

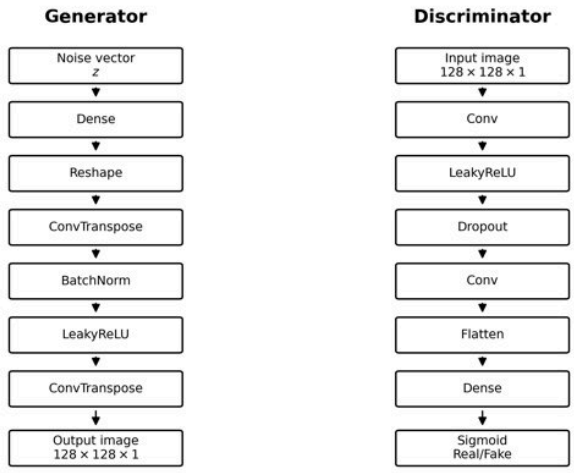


Fig. 4. A schematic representation of the GAN architecture used in this work, highlighting the main blocks of the Generator and the Discriminator.

C. Training setup

The GAN was trained using the binary cross-entropy (BCE) loss function, since the Discriminator acts as a binary classifier responsible for distinguishing real images from synthetic images. During training, real images were assigned label 1, while images generated by the Generator were assigned label 0.

The Discriminator's loss was calculated by considering both real and synthetic images simultaneously, as shown in Equation (2).

$$\mathcal{L}_D = -\mathbb{E}_x \left[\log D(x)\right] - \mathbb{E}_z \left[\log\left(1 - D(G(z))\right)\right] \quad (2)$$

The generator loss was calculated using the equation (3).

$$\mathcal{L}_G = -\mathbb{E}_z \left[\log D(G(z))\right] \quad (3)$$

In Equations (2) and (3), x represents a real image, z represents the input noise vector, $G(z)$ corresponds to the synthetic image produced by the Generator, and $D(\cdot)$ indicates the probability assigned by the Discriminator that the sample is real.

Both the Generator and the Discriminator were optimized using independent Adam optimizers with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The latter corresponds to the default value adopted by TensorFlow/Keras. The models were trained using a batch size of 128 for 1000 epochs. A random seed of 42 was used, and model checkpoints were saved every 50 epochs. No *early stopping* strategy was employed. The model weights were randomly initialized using the default Keras initializers.

The experiments were conducted on a computer equipped with 16 GB of RAM and an NVIDIA GeForce RTX 4060 GPU with 8 GB of VRAM. The implementation was developed using Python version 3.11.9, TensorFlow version 2.21.0, CUDA version 12.5, and cuDNN version 9.3.

D. Xception Classification Protocol

The downstream classification experiment used a balanced subset of 362 complete axial CT slices obtained from the NEPI database and the head CT database described by Chilamkurthy et al. [13]. The subset contained 181 slices with visible ischemic stroke findings and 181 slices without visible ischemic abnormalities. Among the positive slices, 55 were labeled as hyperacute, 66 as acute, 15 as subacute, and 45 as chronic. The examinations were reviewed by a radiology technologist and a physician, and the final labels were assigned by consensus. Only complete axial slices with sufficient brain parenchyma were retained; lesion-centered crops were not used.

The DICOM pixel values were converted to Hounsfield units using the *RescaleSlope* and *RescaleIntercept* attributes. A brain window with a window level of 0 HU and a window width of 150 HU was applied, followed by skull removal and min–max normalization to $[0, 1]$. For Xception, the images were resized to 299×299 pixels and the grayscale channel was replicated three times to obtain an input of $299 \times 299 \times 3$.

The Xception model [14] used an ImageNet-pretrained backbone with frozen weights. Its original classification layer was replaced by a *GlobalAveragePooling2D* layer, a dense layer with 512 neurons and ReLU activation, a dropout layer with a rate of 0.4, and a single sigmoid output neuron. The classifier was optimized using Adam with a learning rate of 10^{-3} and binary cross-entropy loss. Training was performed

for up to 50 epochs with a batch size of 128. Early stopping monitored the validation loss with a patience of 10 epochs and restored the weights associated with the lowest validation loss.

Four training conditions were evaluated: (i) real images only; (ii) real images with classical data augmentation; (iii) real images supplemented with GAN-generated samples; and (iv) real images with both classical and GAN-based augmentation. Five-fold stratified cross-validation was used, and the same folds were retained across all four conditions. In each iteration, four folds were used for training and the remaining fold was used for validation and performance estimation. For the GAN conditions, the GAN was reinitialized and trained using only the real images from the four training folds. Synthetic samples were added only to the classifier training data at a real-to-synthetic ratio of 1:1 and with equal class proportions; the validation folds contained real images only. The predictions from the five validation folds were concatenated to construct the aggregated confusion matrices.

E. Evaluation Metrics

This subsection presents the metrics used to evaluate the quality and fidelity of computed tomography images containing ischemic stroke lesions generated by the GAN-based model.

The quantitative evaluation of the quality of the synthetic images was performed using two complementary metrics: the Fréchet Inception Distance (FID) and the Structural Similarity Index (SSIM). The combined use of these metrics allows for the evaluation of both the statistical similarity between the sets of real and synthetic images and the preservation of the structural features present in the images.

FID is a metric widely used in the evaluation of generative models and quantifies the distance between the feature distributions extracted from the sets of real and generated images [15]. To do this, feature vectors are obtained from the activations of the *global average pooling* layer of the Inception-v3 network, which was previously trained on the ImageNet dataset. The Fréchet distance between the Gaussian distributions associated with the real and synthetic images is calculated according to Equation (4).

$$FID = |\mu_r - \mu_g|^2 + Tr\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \quad (4)$$

In Equation (4), μ_r and Σ_r represent the mean and covariance matrix of the features extracted from real images, while μ_g and Σ_g correspond to generated images. Lower FID values indicate greater similarity between the two feature distributions.

SSIM evaluates perceptual similarity between a reference image and a compared image by considering luminance, contrast, and local structure [16]. It is calculated according to Equation (5).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

In Equation (5), μ_x and μ_y are the local means, σ_x^2 and σ_y^2 are the local variances, σ_{xy} is the local covariance, and C_1 and C_2 are stabilization constants. Higher SSIM values indicate greater structural similarity. Because the proposed GAN does not generate paired reconstructions, SSIM is interpreted only as an auxiliary structural measure and not as evidence of anatomical correspondence.

IV. Results

A. GAN Image Evaluation

This section presents the results obtained from the evaluation of the proposed model for generating synthetic computed tomography (CT) images showing ischemic stroke lesions, using Generative Adversarial Networks (GANs). The analysis was conducted using quantitative metrics and visual qualitative assessment.

The quality of the synthetic images was evaluated using two metrics widely recognized in the literature: the Structural Similarity Index (SSIM) and the Fréchet Inception Distance (FID). The experiments were conducted using a subset of 500 generated images. Table IV presents the observed mean values and standard deviations.

TABLE IV
AVERAGE SSIM AND FID SCORES FOR GENERATED IMAGES

Metric	Average	standard deviation
SSIM	0,8002	±0,037
FID	6,52	±0,83

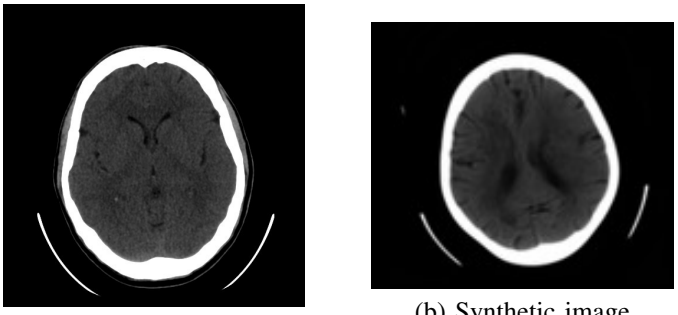
The results indicate a high degree of structural similarity between the real and generated images, as compared to the studies cited above. The FID value was less than 7, which was lower than in other studies.

The qualitative evaluation was performed through visual inspection of the synthesized images. Figure 5 presents a comparison between a real computed tomography image containing an ischemic lesion and a synthetic image generated by the proposed model. Although the original images have a resolution of 512×512 pixels, the GAN was trained using images resized to 128×128 pixels. For visual comparison purposes, both images are presented at the same scale.

The generated images preserved relevant visual characteristics, such as the contrast between the lesion and the brain parenchyma, and did not present recurring visual artifacts. These findings suggest that the model was able to learn relevant structural patterns from the training set.

B. Xception Classification Results

Table V summarizes the Xception results across the five folds. Training with GAN-generated images increased all reported metrics relative to the real-only baseline. The combined use of classical and GAN-based augmentation achieved the highest aggregated performance, with 167 true negatives, 14 false positives, 15 false negatives, and 166 true positives. This corresponds to an aggregated accuracy of 91.99%, precision



(a) Actual image (b) Synthetic image

Fig. 5. Comparison between a real computed tomography (CT) image showing an ischemic stroke and a synthetic image generated by the proposed model. Although the GAN was trained using images resized to 128×128 pixels, both images are presented at the same scale to facilitate visual comparison.

of 92.22%, sensitivity of 91.71%, specificity of 92.27%, and F1-score of 91.97%.

TABLE V
XCEPTION PERFORMANCE UNDER FIVE-FOLD STRATIFIED SLICE-LEVEL CROSS-VALIDATION. VALUES ARE MEAN ± STANDARD DEVIATION; FN IS THE AGGREGATED NUMBER OF FALSE NEGATIVES.

Training condition	Accuracy	Sensitivity	F1-score	FN
Real only	$88.12 \pm 1.21\%$	$86.19 \pm 1.97\%$	$87.88 \pm 1.29\%$	25
Classical augmentation	$89.50 \pm 0.84\%$	$88.93 \pm 2.84\%$	$89.42 \pm 1.03\%$	20
GAN augmentation	$90.88 \pm 0.83\%$	$90.60 \pm 1.57\%$	$90.86 \pm 0.76\%$	17
Classical augmentation + GAN	$91.98 \pm 1.21\%$	$91.71 \pm 0.10\%$	$91.97 \pm 1.12\%$	15

Relative to the real-only baseline, the combined condition increased aggregated accuracy by 3.87 percentage points, sensitivity by 5.52 percentage points, specificity by 2.21 percentage points, and F1-score by 4.08 percentage points. The number of false negatives decreased from 25 to 15, a descriptive reduction of 40%. Because no hypothesis tests or confidence intervals were calculated for the paired differences, these gains should not be interpreted as statistically significant.

The combined condition reduced the global false-negative rate from 13.81% to 8.29%. Relative to the real-only baseline, false negatives decreased from 11/55 (20.00%) to 5/55 (9.09%) for hyperacute slices, from 8/66 (12.12%) to 6/66 (9.09%) for acute slices, from 4/15 (26.67%) to 3/15 (20.00%) for subacute slices, and from 2/45 (4.44%) to 1/45 (2.22%) for chronic slices. The subacute estimate is particularly unstable because this subgroup contains only 15 slices.

V. DISCUSSION

This study investigated GAN-based synthesis of CT images with ischemic stroke findings and its use for augmenting an Xception classifier. The synthetic images achieved a mean Structural Similarity Index (SSIM) of 0.8002 ($\pm$0.037) and a mean Fréchet Inception Distance (FID) of 6.52 ($\pm$0.83). These metrics characterize structural and feature-distribution similarity under the adopted protocol, but do not demonstrate clinical correctness because SSIM depends on the image-pairing procedure and FID uses ImageNet features from Inception-v3.

In the Xception experiment, the addition of GAN-generated images increased aggregated accuracy from 88.12% to 90.88% and sensitivity from 86.19% to 90.60%. Combining classical and GAN-based augmentation produced the best descriptive result, with 91.99% accuracy, 91.71% sensitivity, and 15 false negatives, compared with 25 false negatives in the real-only condition. The stage-stratified analysis also suggested fewer false negatives in hyperacute, acute, subacute, and chronic slices under the combined condition.

REFERENCES

[1] R. C. Rennert, A. R. Wali, J. A. Steinberg, D. R. Santiago-Dieppa, S. E. Olson, J. S. Pannell, and A. A. Khalessi, "Epidemiology, natural history, and clinical presentation of large vessel ischemic stroke," *Neurosurgery*, vol. 85, no. Suppl 1, p. S4, 2019.

[2] M. Wajngarten and G. S. Silva, "Hypertension and stroke: update on treatment," *European Cardiology Review*, vol. 14, no. 2, p. 111, 2019.

[3] D. Vu, R. G. González, and P. W. Schaefer, "Conventional mri and mr angiography of stroke," in *Acute Ischemic Stroke*. Springer, 2006, pp. 115–137.

[4] N. H. Rajini and R. Bhavani, "Computer aided detection of ischemic stroke using segmentation and texture features," *Measurement*, vol. 46, no. 6, pp. 1865–1874, 2013.

[5] S. Ojaghihaghighi, S. S. Vahdati, A. Mikaeilpour, and A. Ramouz, "Comparison of neurological clinical manifestation in patients with hemorrhagic and ischemic stroke," *World journal of emergency medicine*, vol. 8, no. 1, p. 34, 2017.

[6] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[7] M. Sun, X. Li, and W. Sun, "Image generation and lesion segmentation of brain tumors and stroke based on gan and 3d resu-net," *IEEE Access*, 2024.

[8] Z. Li, Z. Huang, M. Li, W. Li, Y. Yang, D. Liang, L. Zhang, and Z. Hu, "Deep learning based synthetic magnetic resonance image generation from computed tomography imaging for cerebral ischemic stroke patient," in *2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD)*. IEEE, 2023, pp. 1–1.

[9] R. Ghnemat *et al.*, "Ischemic stroke lesion segmentation using mutation model and generative adversarial network," *Computer Methods and Programs in Biomedicine*, 2021.

[10] J. Rubin and S. M. Abulnaga, "Ct-to-mr conditional generative adversarial networks for ischemic stroke lesion segmentation," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 2019, pp. 1–7.

[11] Y. Liu, F. Yang, and Y. Yang, "Free-form lesion synthesis using a partial convolution generative adversarial network for enhanced deep learning liver tumor segmentation," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.09065

[12] Z. Ahmad, Z. u. A. Jaffri, M. Chen, and S. Bao, "Understanding gans: Fundamentals, variants, training challenges, applications, and open problems," *Multimedia Tools and Applications*, pp. 1–77, 2024.

[13] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Development and validation of deep learning algorithms for detection of critical findings in head ct scans," *The Lancet*, vol. 392, no. 10162, pp. 2388–2396, 2018.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

Contrast Curve Correction and Post-Processing Methodology for Digital Mammography Images

Matheus Tomaseli Pereira
Departamento de Engenharia Elétrica e de Computação
Escola de Engenharia de São Carlos - USP
São Carlos, Brazil
0009-0003-6638-9004

Homero Schiabel
Departamento de Engenharia Elétrica e de Computação
Escola de Engenharia de São Carlos - USP
São Carlos, Brazil
0000-0002-7014-948X

***Abstract*—This work presents a methodology for the correction of characteristic contrast curves and a post-processing routine for contrast improvement in digital mammography images. The study revolves around 4 systems, from which images of an anthropomorphic breast phantom were extracted. This phantom contains a grayscale step wedge that enables the extraction of the characteristic curves from each image using Piecewise Cubic Hermite Interpolating Polynomial (PCHIP). These curves allowed for the calculation of a conversion curve that transforms a contrast curve into a reference curve, defined by the characteristic curve of a radiographic film. Furthermore, a post-processing routine was established, consisting of 3 steps: pre-processing for noise removal; equalization based on quality assessment metrics, specifically the Modulation Transfer Function (MTF), the Noise Power Spectrum (NPS), and the Detective Quantum Efficiency (DQE); and enhancement filtering. Through this process, it was possible to generate an image similar, and of even better quality, compared to those from commercially available digital mammography systems. This makes the method a great ally in the objective of improving and standardizing digital mammography images, a factor that can significantly aid the diagnostic process of such a complex exam, especially regarding Computer-Aided Diagnosis (CAD).**

***Index Terms*—contrast, digital mammography, characteristic curve, post-processing.**

I. Introduction

Breast cancer is one of the most incident type of cancer among women. An estimated 73,000 new cases were projected for 2025 [1], representing one of the highest female mortality risks worldwide. Most deaths occur in low- and middle-income countries, where diagnoses are often delayed due to misinformation or limited access to screening services. The best way to prevent and increase treatment success rates is through periodic examinations and early-stage detection [2].

Thus, the importance of mammography screening is undeniable; however, it is also essential that the exam can accurately detect any anomalies. The correct detection of different breast structures depends on several factors, with image contrast being one of the most critical [3].

In current digital acquisition systems, there is an important distinction between ”Raw” and ”Processed” images. Raw images are derived directly from X-ray attenuation information and correspond to a linear relationship between the logarithm of exposure and the pixel value, featuring low contrast and a wide dynamic range, as seen in Fig. 1b. Because of this, subtle contrast differences essential for a correct diagnosis are difficult for the human eye to perceive, given its logarithmic response to contrast. Processed images are obtained after applying processing steps to the raw image to improve quality or enhance regions of interest, and this is the image presented to the radiologist for diagnosis.

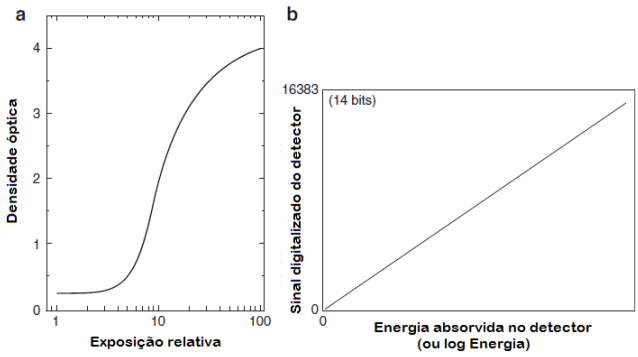


Fig. 1. Comparison between the characteristic curve of a screen-film acquisition system (a) and a digital system (b). Source: [4]

Although the diagnosis is performed on processed images, raw images are used for analysis, study, and treatment to avoid the influence of already applied processing methods.

Having established the importance of a coherent image and the relationship between exposure and acquisition system response, this work analyzes the characteristic contrast curves of digital mammography images, as well as inherent acquisition system metrics, such as the Modulation Transfer Function (MTF), the Noise Power Spectrum (NPS), and the Detective Quantum Efficiency (DQE). Based on the extracted data, the objective is to correct possible deficiencies in the studied systems and achieve image standardization. This facilitates auxiliary diagnosis by Computer-Aided Diagnosis (CAD) systems in conjunction with a healthcare professional.

The correction occurs through the adjustment of the images' contrast curves and the application of a post-processing routine to match the contrast quality presented in commercially available and widely used systems.

II. Methodology

For the tests, an anthropomorphic breast phantom was used, specifically Model 18-226 from Nuclear Associates, nicknamed phantom Rachel, measuring $19.6 \times 11.7 \times 6.1$ cm. The model simulates a compressed breast of 5 cm thickness, consisting of 50% glandular tissue and 50% adipose tissue, and includes a grayscale step wedge on the side, used to extract the characteristic contrast curves of the studied systems.



Fig. 2. Phantom Rachel used in the study.

Images of the phantom were acquired at various hospitals and radiological clinics under predefined exposure metrics (kVp and mAs values) across 4 different systems:

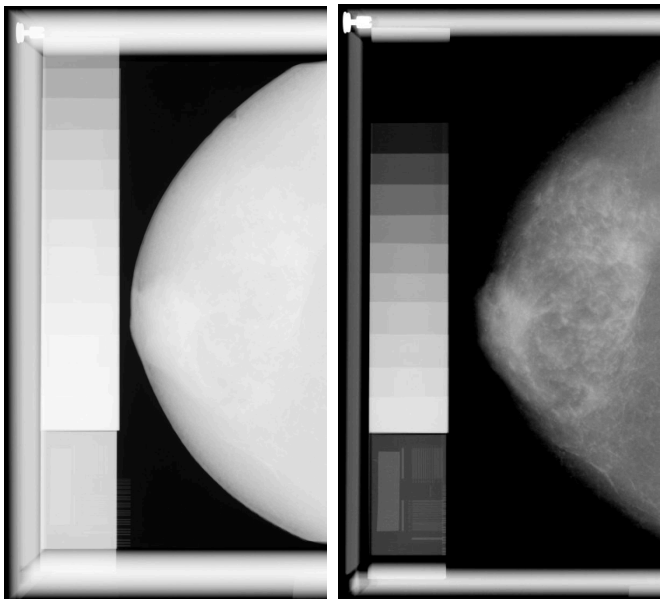
- System 1: CsI scintillator + Si detectors (0.1 mm/pixel); Mo or Rh target; Mo or Rh filter.
- System 2: a-Se detectors (0.07 mm/pixel); W target; Rh or Ag filter.
- System 3: a-Se detectors (0.05 mm/pixel); W target; Rh filter.
- System 4: CsI scintillator + Si detectors (0.1 mm/pixel); W target; Rh filter.

The studies were conducted on one image from each system, acquired under the following parameters:

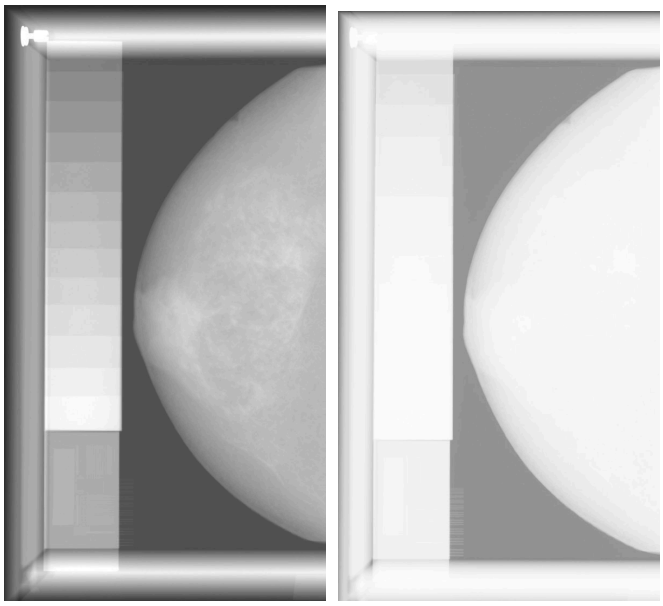
- IMG01: acquired from System 1, with 30 kVp and 100 mAs;
- IMG02: acquired from System 2, with 30 kVp and 100 mAs;
- IMG03: acquired from System 3, with 29 kVp and 99 mAs;
- IMG04: acquired from System 4, with 30 kVp and 80 mAs.

The equipments were not calibrated by the research team, however, as they are utilized by medical facilities on the day to day life, they are calibrated by the manufacturer and the hospitals' maintenance teams. They are up to standards following the periodic quality evaluations required for them to be legally used.

Given that all the images were obtained using the same phantom and under the same kVp and mAs parameters (or at least as close as possible), it is safe to say that the data obtained is comparable between different systems.



(a) IMG01. (b) IMG02.



(c) IMG03. (d) IMG04.

Fig. 3. Raw images acquired from the 4 different systems.

A. Characteristic Curves Determination

Using the grayscale present on the side of the phantoms, the average value of each of the 12 steps was extracted to perform interpolation and generate the corresponding curve for the studied system.

The interpolation type used is PCHIP (Piecewise Cubic Hermite Interpolating Polynomial), which allows joining the points using third-degree functions, ensuring curve smoothness. Furthermore, the method guarantees monotonicity while connecting the points with great flexibility and without oscillations.

B. Reference Curve Selection

To correct the obtained contrast curves, it was necessary to define a reference curve so that the images could be processed to match it. Keeping in mind the human eye's logarithmic response to contrast, a curve that benefits from this fact was sought.

Given the logarithmic relationship between optical density and relative exposure of a radiographic film's characteristic curve, shown in Fig. 1a, a radiographic film curve was defined as the reference to adapt the image to the human eye's contrast perception. In [3], a set of 20 points describing the used curve was extracted, which were then connected using PCHIP.

As a characteristic curve carries the informations that relate the darkening degree of the image with the values of incident exposure, it represents a relation of input-output of the imaging system. In this sense, the choice of a radiographic film's characteristic curve intends on avoiding the linear relationship observed in digital systems, as represented by Fig. 1b, that is not perceived well by humans.

By the same logic, a transformation of the image by utilizing a different characteristic curve is not capable of eliminating information useful for the diagnosis, since it only affects how light or dark a different part of the breast is. This transformation never attributes the same value of gray level to different regions, hence, never losing information, which is granted by the monotonicity of the curve, that never assumes the same output values for different inputs. It is also worth noting the fact that the post-processing routine utilized by the manufacturers includes a characteristic curve transformation.

C. Contrast Correction Curves Calculation

Once the characteristic curves of the studied systems and the reference curve were obtained, it was possible to calculate contrast correction curves. When applied to the original image, these curves can convert its contrast curve to the reference one.

Each correction curve is unique to each image and can only be generalized for images from the same system acquired under the same kVp and mAs parameters. Thus, its calculation is performed using the previously defined reference curve and the characteristic curve of the image to be corrected.

The goal is to obtain a curve that, instead of relating grayscale steps to gray levels, relates the gray levels of the image to be corrected with those of the reference curve. For this, both used curves had their domains normalized (allowing the use of curves obtained from grayscales with different numbers of steps), and then several equidistant points were extracted from each curve. Both sets of points were obtained for the same domain values, thus allowing the relation of an image's gray level to the reference based on the same X-axis value.

With the points relating gray levels, PCHIP was applied to generate the actual correction curve. Its application occurs in a pixel-by-pixel transformation of the image to be corrected, making its contrast curve approaching the reference used.

D. Quality Control Metrics

To be used in the post-processing step, quality control metrics were extracted from the image detection systems, specifically MTF, NPS, and DQE, following the international standard protocols for digital mammography [5]. The calculation for each metric follows these equations:

$$MTF(f) = \frac{|\mathcal{F}\{LSF(x)\}|}{|\mathcal{F}\{LSF(x)\}|_{f=0}}, \tag{1}$$

where:

- f: Spatial frequency (in cycles/mm).
- $\mathcal{F}$: Fourier Transform operator.
- $LSF(x)$: Line Spread Function, obtained by the derivative of the Edge Spread Function ($ESF(x)$), defined as $LSF(x) = \frac{d}{dx}ESF(x)$.

$$NPS(u,v) = \frac{\Delta x \Delta y}{N_x N_y M} \sum_{i=1}^{M} |\mathcal{F}\{I_i(x,y) - B_i(x,y)\}|^2, \tag{2}$$

where:

- u, v: Spatial frequencies in the horizontal and vertical axes (cycles/mm).
- $\Delta x, \Delta y$: Pixel spacing in the x and y directions (in mm).
- N_x, N_y: Dimensions (number of pixels) of each evaluated sub-region.
- M: Total number of sub-regions extracted from the image to compose the statistical mean.
- $I_i(x,y)$: Pixel intensity matrix of the i-th original sub-region.
- $B_i(x,y)$: 2nd-order polynomial background fit (detrending) of the i-th sub-region, subtracted to isolate noise fluctuations.

$$DQE(f) = \frac{\bar{S}^2 \cdot MTF^2(f)}{NPS(f) \cdot Q} \tag{3}$$

where:

- $\bar{S}$: Average image signal, calculated from a Region of Interest (ROI) in the flat-field image.
- Q: Incident photon fluence on the detector (photons/mm^2).

To perform the calculation correctly, specific images obtained by the digital mammography systems are required. MTF requires an abrupt transition to be calculated with minimal errors, which was achieved by placing a lead plate (completely white in the final image) over the image acquisition system, ensuring a clean and abrupt transition relative to the image background (completely black). NPS, on the other hand, requires a large, uniform region for noise detection, achieved by placing an acrylic plate over the system, providing a uniform image from which noise can be extracted with minimal interference.

These metrics are a measure of the image acquisition system, and not only of the images utilized for their calculation, since they extract information that does not depend on any

acquisition parameters. This means that only a handful of images are recquired to define a whole system, and the data can be used on the processing of any other images extracted from the same system.

E. Post-Processing

By applying the correction curve, an equalized image is obtained, with well-distributed gray levels to leverage the human eye's natural contrast perception. However, a post-processing stage is still necessary to match the contrast level displayed in processed images from current digital mammography systems. The developed post-processing can be divided into 3 steps:

The first step consists of pre-filtering, responsible for removing the noise present in the image which, although not perceptible to the naked eye, is drastically amplified in the subsequent steps. Since the breast is a highly complex organ and the image is rich in details, it is necessary to use a filter that loses as few details as possible. Thus, an NLM (Non-Local Means) filter was used [6]. Instead of comparing a pixel with its immediate neighbors for noise removal, a process performed by simpler filters like the Gaussian, this filter compares blocks of pixels across large neighborhoods to preserve more image details.

The second step consists of correcting potential image defects using the quality control metrics extracted from the images, specifically MTF and DQE. The gain of the filter $G(f)$ is composed of the product of a damped inverse restoration term (H_{inv}) and a Wiener-based attenuation mask (W_{norm}) [7]:

$$G(f) = H_{inv}(f) \cdot W_{norm}(f) \tag{4}$$

where the spatial modulation restoration is given by:

$$H_{inv}(f) = \min\left[\left(\frac{1}{MTF(f)}\right)^{\alpha}, G_{max}\right] \tag{5}$$

and the signal reliability mask (normalized Wiener fraction), which limits the values the filter can assume, is expressed as:

$$W_{norm}(f) = \frac{1}{\max(W)} \cdot \left(\frac{DQE(f)}{DQE(f) + C}\right) \tag{6}$$

where:

- α: Enhancement strength.
- G_{max}: Maximum gain limit.
- C: Damping constant.

The filter acts by amplifying the frequencies where the MTF is poorest, using the DQE as a brake to prevent over-amplifying high frequencies, which could reveal a large amount of noise. At frequencies where the DQE drops, there is high noise in the image. By using this information, the filter can be calibrated to enhance frequencies with information loss in a controlled manner.

The third and final step consists of enhancement filtering, designed to highlight glandular tissue and darken adipose tissue, ensuring good contrast values in the final image. First, a Gamma correction is applied, capable of deepening dark areas and reducing the prominence of adipose tissue. Next, CLAHE (Contrast Limited Adaptive Histogram Equalization) [8] is used, an image histogram equalization technique responsible for increasing local contrast and revealing hidden structures, such as microcalcifications. It is worth noting that CLAHE degrades the overall image contrast at the expense of local contrast, which is why a blending between the pre and post-application images is performed to mitigate this issue.

F. Numerical Validation

To quantify the results, the RMS contrast [9] (referred to as Crms for abbreviation) and the Carneiro Contrast Index (ICC) [10] metrics were used. The Crms is defined by the following equation:

$$C_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_i - \bar{I})^2} \tag{7}$$

where:

- N: Total number of pixels in the evaluated Region of Interest (ROI).
- I_i: Intensity value of the i-th pixel contained in the ROI.
- $\bar{I}$: Average intensity of all pixels present in the respective ROI.

It is important to highlight that Crms is not calculated over the entire image, but in 64×64 pixel windows, in order to capture only the local contrast [9], which is the focus of this work. Calculating Crms for the whole image yields a very high value for RAW images and a very low one for processed images. This is because RAW images have a very high low-frequency gradient (attributed to the thickness gradient between the breast base and the nipple) that artificially inflates the Crms value. This is corrected in post-processing, which equalizes these values and drastically reduces the global contrast value.

The ICC is based on calculating the standard deviation of each pixel relative to its neighborhood. The method generates a standard deviation map that reveals the regions of highest contrast. By averaging all the found values, it is possible to quantify the contrast in a single number. To perform the calculations on the studied images, a 3×3 kernel was used; that is, the neighborhood consists of all eight adjacent pixels.

The numbers found only have meaning when compared to each other and under the same conditions. It is also worth mentioning that the image's bit depth does not influence the results, as ICC takes this parameter into account and Crms is not influenced by it.

Regarding the application methodology, in both cases, only the part containing the breast was cropped from the image, leaving out the "frame" containing the grayscale. Specifically for the Crms calculation, an Otsu algorithm was applied to the image, responsible for separating the breast from the black background behind it, which can considerably influence the results. Since the method considers the average intensity level of all pixels, it is sensitive to the proportion of black pixels

in the image, causing small changes in image cropping to yield measurement errors. For the ICC, however, the black background is not a concern, as it is practically uniform and the ICC only cares about a pixel's standard deviation relative to its neighborhood.

III. Results and Discussion

The previously described methods were applied only to System 1, the only one from which the images necessary for the correct calculation of quality control metrics could be obtained. This system has some of the poorest metrics when compared to the other three, therefore, although the complete processing procedure can only be applied to System 1, a worst-case scenario was simulated, verifying the full potential of the processing technique. The IMG01 that underwent the described processes will be referred to as "converted".

First, the characteristic curves of all systems were extracted, shown in Figures 4 and 5. Comparing the curves in both graphs, it is notable, firstly, how the post-processing performed by the mammography systems inverts the curves' trend, so that the scale goes from a darker to a lighter shade, and the image has a black background with white enhancements, except for System 3. Despite this, in this work, all images with decreasing characteristic curves are displayed with an inverted LUT (Look-Up Table) as a form of standardization.

Another important characteristic to observe is the change in the curves' shapes, as well as their histogram widths, represented by the Y-axis width admitted by the curve. All systems tend to give the curve a sigmoid shape during their processing stage; however, there is an increase in histogram width only in Systems 2 and 3, with a persistence and even decrease in Systems 1 and 4. This reveals a greater poverty of detail and poor utilization of the available grayscale range, which goes from 0 to 4095 since the images are 12-bit.

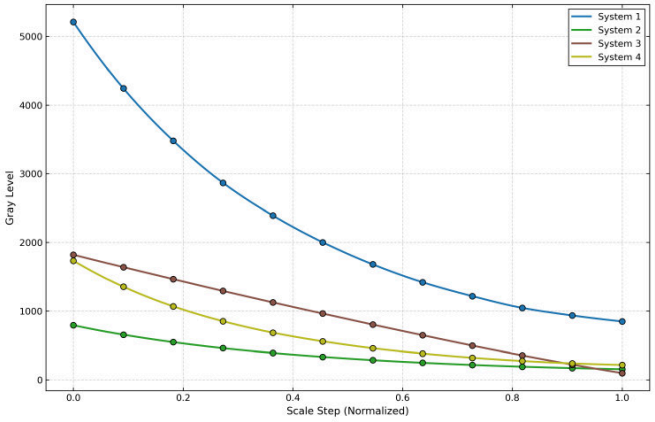


Fig. 4. Comparison between characteristic curves of the RAW images from the studied systems.

Taking IMG01 as the baseline to be improved, the conversion curve shown in Fig. 6 was generated. It is notable that it has a very complex shape to achieve with a single function, which justifies the choice of PCHIP as the interpolation method used, since it provides great flexibility in joining the points. Note that the graph has values up to the order of 16000 on the X-axis, which contradicts Fig. 4, that shows the curve only going up to values in the order of 5000. This is due to the fact that, in addition to the 12 values of the grayscale, the values of the phantom background and the image background were used to generate the correction curve, ensuring that not only the scale but the entire image conforms to the used curve. If these values were not used, the background regions of the image would not have values to be converted to, which would generate a faulty and noisy image.

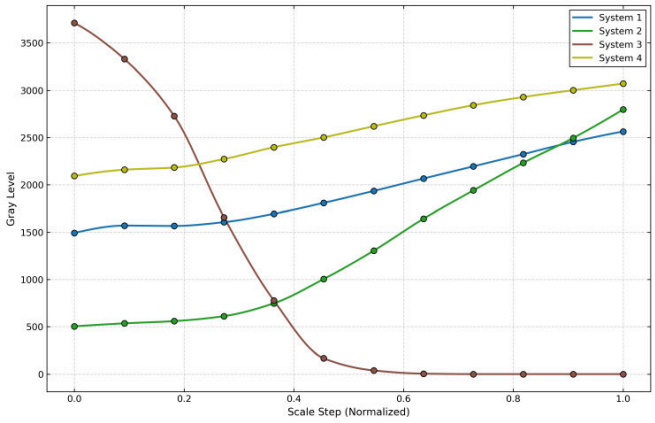


Fig. 5. Comparison between characteristic curves of the processed images from the studied systems.

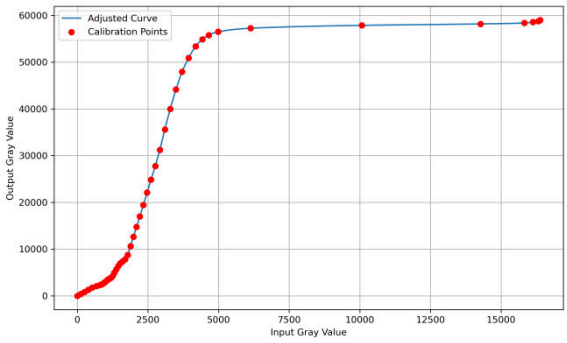


Fig. 6. Conversion curve generated for IMG01 processing.

In Fig. 7, there is a comparison between the processed curves from Fig. 5, along with the converted and post-processed IMG01 curve. Due to the inclusion of out-of-scale points, the converted image curve is noticeably shifted relative to the reference curve; however, it presents the same shape, demonstrating that the method was applied successfully.

Next, the 3-step post-processing was performed on IMG01 with its converted characteristic contrast curve. The resulting image is present in Fig. 8 and can be visually compared with the processed images from each of the 4 mammography systems present in Fig. 9. It reveals itself to be a kind of inter-

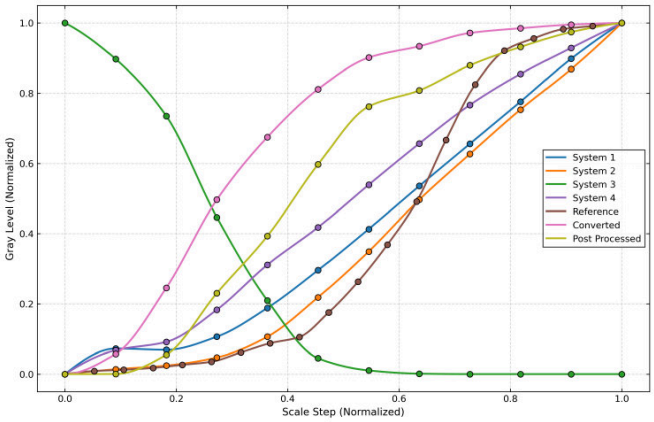


Fig. 7. Comparison between all curves in the study, excluding the raw images.

mediate between the IMG02 and IMG03 images, appearing to have better contrast than the IMG01 and IMG04 images, which are more grayish and with little perceptible differentiation between glandular and adipose tissues.

An evident difference between the converted image and those processed by the mammography systems is the absence of a well-defined breast contour. This can be attributed to the application of the Gamma correction, responsible for darkening the parts that tend more towards black, which correspond to the adipose tissue. Thus, the breast ends up losing its contour, which is made almost entirely of fat and holds little value in the exam's analysis. However, it is worth noting that if there were any abnormality in that region, it would not be erased but rather enhanced, mainly due to the influence of CLAHE. Since the phantom used does not have any type of lesion or microcalcification, this region is certainly free of such structures in the presented examples.

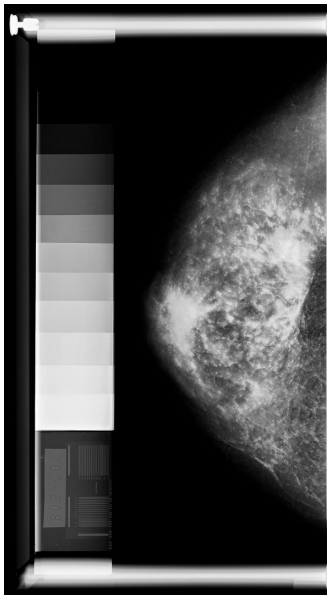
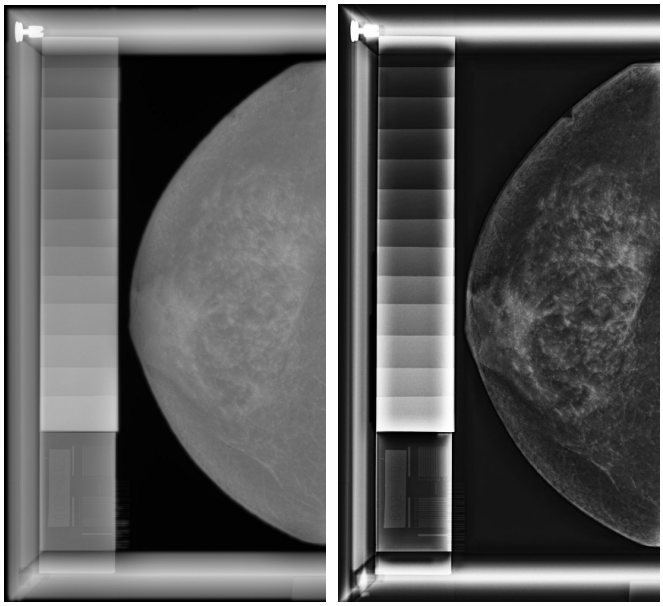
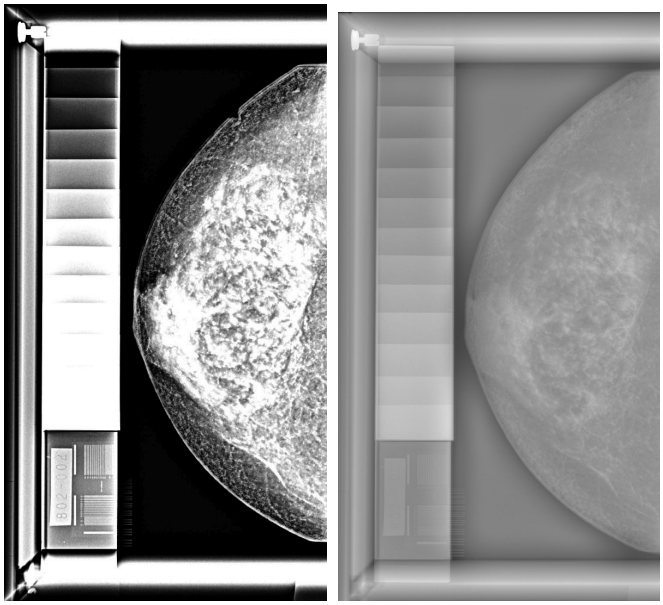


Fig. 8. Post-processed IMG01.



(a) Processed IMG01. (b) Processed IMG02.



(c) Processed IMG03. (d) Processed IMG04.

Fig. 9. Comparison between processed images (performed by the manufacturer softwares) from the 4 systems.

To confirm what was visually assessed, ICC and Crms measurements were taken for all images, both RAW and processed from all systems, as well as the System 1 image submitted to the developed methodology (called "converted"). The results are found in Tables I and II.

First, it is noted that the ICC and Crms values of the raw images are lower than almost all the processed ones and the converted one, as expected, since the intention of processing is to improve the image's contrast. This can be verified visually when comparing Figures 8, 9, and 3, given that glandular tissue is difficult to perceive in raw images, in addition to the fact

TABLE I
ICC VALUES OF THE STUDIED IMAGES

Image	RAW	Processed	Converted
IMG01	0.3036	0.4798	2.8303
IMG02	0.3282	1.6602	-
IMG03	0.2487	4.2219	-
IMG04	0.1297	0.5906	-

TABLE II
CRMS VALUES OF THE STUDIED IMAGES

Image	RAW	Processed	Converted
IMG01	0.061028	0.0941	0.240856
IMG02	0.050542	0.220715	-
IMG03	0.048721	1.072499	-
IMG04	0.072192	0.043850	-

that most present a grayish and unhighlighted image, requiring some processing to improve their display.

Focusing only on Systems 2 and 3, which provide a stronger competition to the converted one, a standout fact is the large difference in the values of both metrics between the IMG02 and IMG03 images. This demonstrates a great example of how one should not blindly trust contrast quantification metrics, but rather use them as an impartial aid to visual analysis. Comparing Figures 9d and 9c, it is verified that the processed IMG02 has a more subtle and well-distributed enhancement of the glandular tissue, giving a good idea of which areas are more or less dense. Meanwhile, the processed IMG03 has a very aggressive enhancement that highlights all the glandular tissue, including a region known to be primarily adipose, such as the breast contour. Thus, it is difficult to differentiate the density of each region and point out if there is any abnormality amidst a sea of white highlights.

Keeping this discussion in mind, now comparing the processed IMG02 with the converted IMG01, there is an improvement in both metrics: a considerable one in ICC and a small one in Crms. Combining the numerical values with visual inspection, the converted image gives greater prominence to the glandular tissue, justifying the higher ICC value. However, it still maintains a more balanced enhancement, with regions clearly showing more and less highlight depending on the amount of glandular tissue, performing a job similar to System 2 in this regard, something that can be related to the close Crms values between the images.

Regarding the converted image, it presented good results, being able to match and even surpass the contrast quality of current digital mammography systems available commercially, validating the effectiveness of the methodology developed in this work. One of the most concerning aspects in the processing of this image is noise, since there is a very large amount of enhancement filtering applied to the image, which makes noise that was previously imperceptible visible. However, it is believed that a very satisfactory result has been achieved, presenting a slightly higher noise level than the processed IMG02, which was taken as a reference during the work due to the previously discussed factors.

IV. CONCLUSION

From the collected data, both visual and numerical, it was possible to develop a mammographic image processing method capable of achieving results similar to and even better than those of digital systems currently available on the market.

By extracting and comparing the contrast curves, it was possible to better adapt it to the human eye's perception. Passing the result through an extensive and intensive post-processing routine made it more presentable and suitable for diagnostic analysis, presenting the difficulty of noise removal as a minor adversity that does not distance the final image, in quality, from images generated by professional systems.

In addition to the good results, the method has an undeniable advantage: knowledge of the type and amount of processing being applied to the raw images, as well as control over it. This enables personalized processing to highlight specific structures or to better suit the personal opinion of the professional performing the diagnosis. This feature, coupled with the fact that it is still possible to standardize the images so they are better diagnosed by CAD systems, makes the developed methodology very powerful as a potential diagnostic auxiliary.

It is important to acknowledge that this is a methodology proposition. The conclusions were drawn by visual inspection of the research team as well as the quality metrics, to ensure a more unbiased and objective view. It is clear that the method was tested in a limited ammount of samples and needs further testing, including different systems, different phantoms and different acquisitons parameters. Testing with clinical cases and studies with CADx systems and radiologists will be carried in the future to solidify the approach suggested in this work.

ACKNOWLEDGMENTS

Authors are thankful to FAPESP for the financial support and also to the Management of Education & Research of University Hospital of Federal University of Sao Carlos (SP, Brazil) - Brazilian Company of Hospital Services (EBSERH), as well as to Oswaldo Jorge Neto and Dr. Luciana B. Verçosa, MD, for the technical support.

REFERENCES

[1] Instituto Nacional de Câncer (INCA), "INCA lança publicação com panorama do câncer de mama no Brasil," Ministério da Saúde, 2025. [Online]. Available: https://www.gov.br/inca/pt-br/canais-de-atendimento/imprensa/releases/2025/inca-lanca-publicacao-com-panorama-do-cancer-de-mama-no-brasil. [Accessed: Jun. 10, 2026].

[2] P. Boyle and B. Levin, Eds., World Cancer Report 2008. Lyon: IARC Press, 2008.

[3] R. F. Goes, H. Schiabel, and M. A. Z. Sousa, "Automatic scanning software based on the characteristic curve of mammograms digitizers," *Journal of Electronic Imaging*, vol. 22, no. 1, pp. 013024-1–013024-9, 2013, doi: 10.1117/1.JEI.22.1.013024.

[4] T. C. Granado, "Transformação de intensidade para realce de contraste em mamografias digitais," ResearchGate, 2015. [Online]. Available: https://www.researchgate.net/publication/282130210_Transformacao_de_intensidade_para_realce_de_contraste_em_mamografias_digitais. [Accessed: Jun. 5, 2026].

[5] International Electrotechnical Commission, "Medical electrical equipment - Characteristics of digital X-ray imaging devices - Part 1-2: Determination of the detective quantum efficiency - Detectors used in mammography," IEC 62220-1-2:2007, 2007.

[6] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 60-65, doi: 10.1109/CVPR.2005.38.

[7] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York: Pearson, 2018.

[8] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego: Academic Press Professional, 1994, pp. 474-485.

[9] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032-2040, 1990.

[10] P. C. Carneiro, "Proposta de metodologia para realce de contraste em imagens de mamas densas utilizando decomposição multiescala com transformada discreta wavelet," Ph.D. dissertation, Universidade Federal de Uberlândia, Uberlândia, Brazil, 2019.

Creation of a Database for Clinical Validation of an Artificial Intelligence Algorithm for Detecting Failures in Invasive Blood Pressure Monitoring in Intensive Care Units

1st Alícia Mei Kato Camigauchi
Lab. de Engenharia Biomédica (LEB)
Universidade de São Paulo (USP)
São Paulo, Brazil
0009-0006-2310-7427

2nd Yasmin de Oliveira Feitosa
Lab. de Engenharia Biomédica (LEB)
Universidade de São Paulo (USP)
São Paulo, Brazil
0000-0001-5469-2127

3rd Gustavo Yeh Fuzinato
Lab. de Engenharia Biomédica (LEB)
Universidade de São Paulo (USP)
São Paulo, Brazil
0000-0002-4929-2361

4th Felipe Fava de Lima
Lab. de Engenharia Biomédica (LEB)
Universidade de São Paulo (USP)
São Paulo, Brazil
0000-0002-7102-8532

5th Henrique Takachi Moriya
Lab. de Engenharia Biomédica (LEB)
Universidade de São Paulo
São Paulo, Brazil
0000-0002-1595-1779

***Abstract*—Invasive Blood Pressure (IBP) monitoring is a cornerstone of hemodynamic management in Intensive Care Units (ICUs). However, the accuracy of continuous IBP measurements heavily relies on the dynamic response of the catheter-transducer system. Waveform distortions can lead to significant clinical misinterpretations. This study outlines the creation of a clinical database comprising real hemodynamic records annotated via standardized fast-flush tests. Data from ICU patients were acquired at 125 Hz and digitally upsampled to precisely determine the system's natural frequency and damping coefficient. Out of 11 evaluated patients, 8 systems were fully characterized, revealing 3 inadequate setups according to Gardner's criteria. This database establishes an essential foundation for the future validation of artificial intelligence algorithms designed to automatically detect IBP waveform distortions, ultimately enhancing patient safety within the Brazilian Sistema Único de Saúde (SUS) context. This study does not evaluate algorithm performance; it establishes a pilot dataset against which such evaluation will later be conducted.**

***Index Terms*—Invasive Blood Pressure, Fast-Flush Test, Signal Processing, Intensive Care Unit, Artificial Intelligence, Hemodynamic Monitoring**

I. Introduction

Invasive Blood Pressure (IBP) monitoring is considered the gold standard for hemodynamic assessment in Intensive Care Units (ICUs) [1]. In contrast to non-invasive methods such as oscillometry, IBP provides continuous measurements, allowing for the accurate determination of Mean Arterial Pressure (MAP), an essential parameter for adjusting vasopressor therapies and fluid infusion in patients with hemodynamic instability [2]. Critical clinical conditions such as septic, cardiogenic, or distributive shock require rapid and accurate decisions, in which obtaining a precise and trustworthy MAP in real-time is central to goal-directed resuscitation protocols, such as those proposed in international sepsis guidelines [3].

However, even though it is a highly accurate method, the reliability of IBP depends on the integrity of the catheter-transducer system. Alterations in the physical properties of this system, such as the presence of air bubbles, clots, excessively long tubes, or loose connections, can modify its dynamic response, resulting in waveform distortions [5]. An underdamped response, for instance, characterized by a low damping ratio (ζ), leads to overestimation of systolic pressure and underestimation of diastolic pressure, that could impact the clinical management of the patient. [1], potentially inducing both unnecessary drug administration and failure to recognize true hypotension.

Recent studies highlight the current relevance of the problem. Damping disturbances are observed in up to one-third of the catheter-transducer systems used by patients undergoing IBP monitoring in the intensive care setting, often going unnoticed [4]. Furthermore, relying solely on qualitative visual inspection makes error identification inherently subjective and error-prone, compromising the consistent detection of dynamic distortions at the bedside. These limitations point to the urgent need for objective, automated, and reliable methods for detecting IBP distortions at the bedside [6].

The fast-flush test is the classic method for verifying system fidelity, but its application depends on technical availability, standardized execution, and subjective interpretation [5]. In this context, several computational approaches have been proposed, with machine learning-based algorithms being particularly noteworthy. Rinehart et al. demonstrated that neural networks trained with distorted waveforms can achieve high

accuracy in identifying calibration errors [6]. Recent advances in generating synthetic databases through one-dimensional modeling of the cardiovascular system have allowed for realistic representation of circulation at different ages and pathophysiological conditions, such as different phases of septic shock [7], [8].

Within the Laboratório de Engenharia Biomédica (LEB-USP) of the Escola Politécnica da USP, a convolutional neural network model is being developed in a doctoral project, aiming to automatically detect distortions such as underdamping based exclusively on the IBP waveform in the context of ICUs. This model was trained with over 10,000 simulations performed with the Nektar1D solver [7], covering various hemodynamic conditions. However, to be clinically validated, the algorithm needs to be tested with real signals from critically ill patients. This project aims to fill this gap by performing standardized data acquisition in the ICU of the Hospital Universitário da Universidade de São Paulo, simultaneously addressing the scarcity of data within the Brazilian context, applying flush tests, curating the signals, and creating the initial structure of a database to evaluate the model's performance in the future. This translational step aims to consolidate a robust and clinically applicable tool with the potential to positively impact the safety and effectiveness of invasive monitoring in intensive care.

II. Objective

The main objective of this project is to collect, structure, and annotate real clinical Brazilian data from an ICU environment to create a database tailored for the future performance evaluation of an AI algorithm (developed at LEB-USP). By systematically identifying cases of underdamping in the catheter-transducer system (analyzing the damping coefficient and the natural frequency from the fast-flush test transient response) the dataset will serve as a ground truth to verify the AI's ability to detect failures autonomously, using exclusively morphological characteristics extracted from the continuous signal waveform.

III. Methods

A. Study Design and Ethical Approval

This is a prospective observational study conducted in the adult ICU of the Hospital Universitário da Universidade de São Paulo (HU/USP). The proposal was approved by the HU/USP Comitê de Ética em Pesquisa (CAE 86650625.3.0000.0076), incorporating family authorization via a deferred consent model, as established by CNS Resolution 466/12.

B. Population and Inclusion Criteria

Data were included from adult patients (age $\geq$ 18 years) admitted to the ICU of the HU/USP who had an arterial catheter inserted in the radial or femoral arteries as clinically indicated by the clinical team. Patients with high-response atrial arrhythmias, peripheral vasculitis, or conditions that prevented the performance of the fast-flush test were excluded.

C. Signal Acquisition

The pressure signals were acquired using a disposable pressure transducer (REF 7004, Zammi Instrumental Eireli, ANVISA (Agência Nacional de Vigilância Sanitária) registration 10216350082), connected to the arterial line via either a single-lumen central venous catheter (16 Ga, 20 cm; REF CV-50016-BZ, Arrow International/Teleflex, ANVISA registration 80117580569) inserted at the femoral site, or a dedicated arterial catheterization kit (20 Ga, 8 cm; REF CA 2008, Alive Heart Material Médico LTDA., ANVISA registration 80415440015) inserted at the radial site. The signals were obtained directly from the bedside monitors (Nihon Kohden Lifescope, Tokyo, Japan) installed in 14 ICU beds, with a sampling frequency of 125 Hz, between September 2025 and April 2026. These monitors are connected to the catheter-transducer system and integrated to the hospital network, transmitting data in real time via a proprietary protocol. Acquisition was performed using a customized algorithm developed by LEB-USP, which accesses and extracts data packets directly from the network, eliminating the need for additional physical interfaces.

At each ICU visit, photographic records of the vital signs and other values were made on all active bedside monitors, as well as of the chart registering of the patients present in the unit on that date, which contained their respective bed numbers. After identifying patients with active blood pressure monitoring, the following data were tabulated: the bed number of each patient, the current date, the channel of the signal collected by the monitor (ART or P1), and the anatomical site of catheter insertion (radial or femoral). Numerical data and waveforms were obtained continuously during hospitalization and temporally synchronized with five flush tests performed by the care team. All records were stored with precise timestamping (the start and end times of the flush tests were also tabulated) in a format compatible for later processing and analysis in a Python environment. Patient identifiers (name, medical record number) were not retained in the digital dataset; records were linked only by bed number, date and time during acquisition, with this mapping stored separately from the signal files. Data are currently maintained on institutional storage and the Open Science Framework (OSF), with public deposition on Zenodo planned to support future validation studies.

D. Study of Signals and Definition of the Flush Test

The signal segments containing the fast-flush test were manually identified using a graphical interface developed in Python, based on the collection times recorded during signal acquisition. The criteria adopted for what characterizes a successful flush test were defined empirically:

- Rise and fall of the flush with a duration of less than 300 ms, distinguishing the test initiation and termination from slow valve manipulations or low-frequency artifacts;
- High-pressure plateau with a duration of 1 to 15 seconds, filtering out spurious impulse noise or prolonged system

maintenance procedures (such as line washing or blood sampling);
- Presence of a cardiogenic component before and after the flush test, aiming to ensure that the catheter was appropriately connected to the arterial system;
- Existence of cardiogenic components superimposed on the plateau of the flush test, verifying active connection to the patient's arterial system.

E. Natural Frequency and Damping Coefficient Acquisition

The transient response caused by the fast-flush test was modeled as a second-order mass-spring-damper system, whose underdamped analytical solution is expressed as:

$$y(t) = A\, e^{-\zeta \omega_n t} \sin\left(\omega_d t + \phi\right). \tag{1}$$

Following the methodology described by Gardner [5], the damping ratio and damped frequency were estimated through visual identification of the response peaks' amplitudes and periods. Because this approach relies strictly on manual peak detection, the temporal resolution of the original signal, acquired at 125 Hz, may be insufficient. To overcome this limitation and ensure a more reliable identification of the local extrema, the signal was digitally upsampled to a higher frequency prior to the analysis via polyphase low-pass FIR (Finite Impulse Response) filtering interpolation. This technique increases the density of data points without altering the underlying frequency content, thereby enhancing the precision of the visual peak detection.

The natural frequency of each system, ω_n, was calculated using the equation:

$$\omega_n = \frac{\omega_d}{\sqrt{1-\zeta^2}}, \tag{2}$$

where the damping ratio is represented by ζ and the damped frequency by ω_d.

Each of the five flush tests performed for every catheter-transducer system was analyzed individually. Considering that the pulsating pressure waveform superimposes on the flush results [5], only the parameters obtained from the responses placed during diastole were considered to calculate the average value of the damping ratio and the natural frequency of each system. Hence, performing five flushes increases the chances of having at least one measurable response during this specific period (diastole was identified visually). Waveforms with less than two peaks were considered overdamped, and their parameters were not estimated.

Based on the estimated parameters, the system's dynamic response was classified as adequate or inadequate following Gardner's [5] criteria.

IV. Results

Real-time hemodynamic data were initially collected from 29 patients. Following preliminary data quality assessment, 18 patients were excluded from the study. The reasons for exclusion were: missing data due to system acquisition errors ($n = 4$); invalid signal identification variables, which precluded IBP file extraction ($n = 5$); and fast-flush test plateau duration of less than 1 second ($n = 9$), consistent with the criteria defined in Section III-D to exclude spurious noise rather than genuine flush maneuvers (Fig. 1). Consequently, data from 11 patients were considered adequate for further analysis.

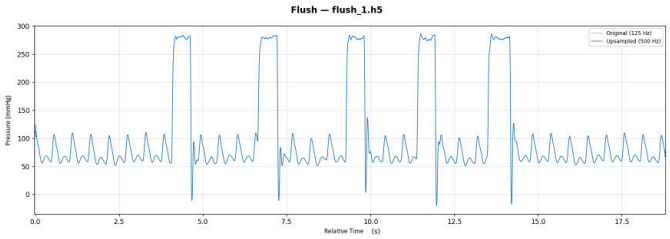


Fig. 1. Inadequate fast-flush test. The plateaus of the flush tests do not have a duration greater than 1 second. The X axis represents the time in seconds and the Y axis represents the blood pressure in mmHg.

Among the 11 analyzed patients, the catheter-transducer system was successfully classified and plotted on Gardner's reference graph in 8 cases. Three cases could not be plotted because they did not present measurable oscillations after the fast-flush test (indicating an overdamped or critically damped response, as shown in Fig. 2), precluding the calculation of the damping ratio and the natural frequency via Gardner's method.

The remaining 8 systems met the criteria and were individually analyzed (Fig. 3). For each flush test, the local peaks and valleys of the oscillation signal were detected to calculate the system's parameters, as illustrated in Fig. 4. Evaluation of these systems, according to Gardner's plot (Fig. 5), revealed that 5 systems presented an adequate dynamic response, whereas 3 were considered inadequate. The calculated parameters are detailed in Table I.

TABLE I
CHARACTERIZATION AND CLASSIFICATION OF THE 8 CATHETER-TRANSDUCER SYSTEMS ANALYZED.

Catheter	**Average parameters**				
	f_n ***(Hz)***	ω_n ***(rad/s)***	ω_d ***(rad/s)***	ζ	***Classification***
1	23.68	148.76	143.10	0.27	Adequate
2	23.62	148.43	142.80	0.27	Adequate
3	28.12	176.69	169.94	0.26	Adequate
4	22.45	141.04	136.39	0.25	Adequate
5	14.88	93.50	90.42	0.25	Inadequate
6	18.33	115.14	112.20	0.22	Inadequate
7	27.19	170.84	169.94	0.08	Inadequate
8	17.43	109.51	101.34	0.38	Adequate

V. Discussion

A methodological enhancement in the approach described in this paper pertains to the calculation of the undamped natural frequency. In classic assessments, such as Gardner's early works, the damped frequency was often utilized as a direct approximation for the undamped natural frequency. This simplification was historically justified because the mathematical divergence between the two values was negligible

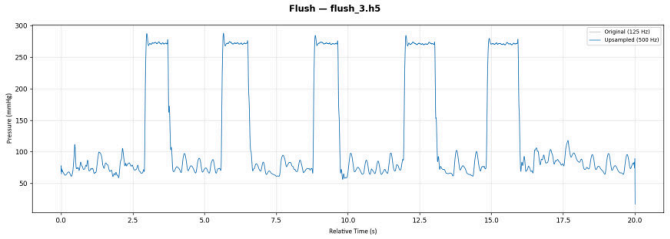


Fig. 2. Overdamped system. There is no oscillation after any of the flush tests. The X axis represents the time in seconds and the Y axis represents the blood pressure in mmHg.

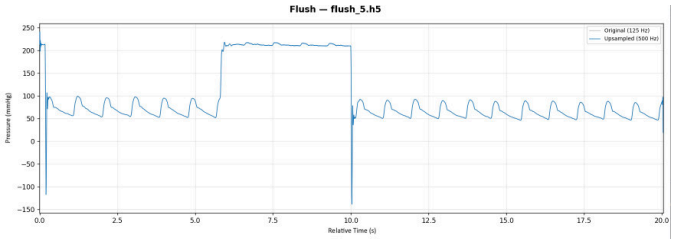


Fig. 3. Adequate fast-flush test for analysis. All the criteria defined are met. The X axis represents the time in seconds and the Y axis represents the blood pressure in mmHg.

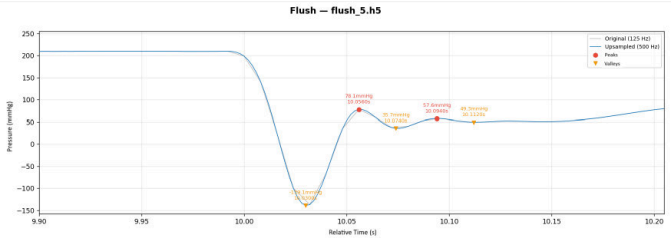


Fig. 4. This graph shows the valleys (in yellow) and peaks (in red) of an oscillation caused by a flush test. The X axis represents the time in seconds and the Y axis represents the blood pressure in mmHg.

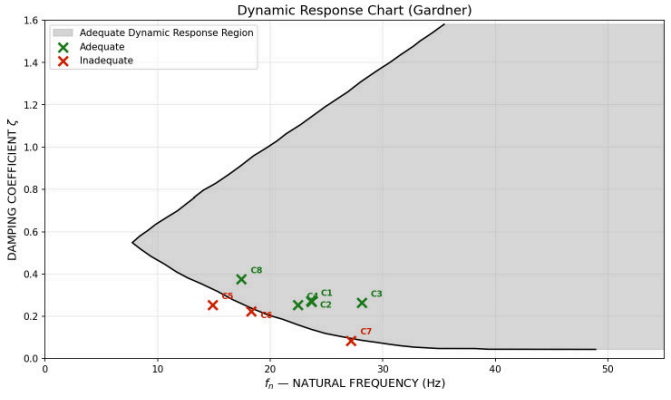


Fig. 5. Gardner's graph (light gray area) with HU's ICU catheter-transducer systems plotted (represented by "x") based on their damping coefficient and the natural frequency. The numbers refer to the catheters on Table I. The X axis represents the damping coefficient and the Y axis represents the natural frequency in Hertz.

when compared to the visual measurement errors inherent to analog paper-strip recordings. Leveraging contemporary digital data acquisition, it is possible to extract the damped frequency and the damping ratio more precisely, allowing for the direct computation of the exact natural frequency. This mitigates historical approximation errors and provides a more mathematically robust database for algorithm validation.

A challenge faced that merits future research is the classification of damped systems that exhibit fewer than two discernible oscillations following the fast-flush test. In such cases, the traditional amplitude ratio method (logarithmic decrement) for determining the damping coefficient becomes mathematically unfeasible. To address these scenarios, feature extraction applied directly to the continuous arterial waveform emerges as a viable solution. Rinehart et al. (2022) bypassed the need for fast-flush tests by extracting 124 morphological features per cardiac beat, demonstrating that machine learning algorithms could accurately detect damped transducer states [6]. This reinforces the potential utility of the database generated in this study for morphology-based AI training.

It is also important to highlight that the classification of the catheter-transducer systems as "adequate" or "inadequate" in this study relies exclusively on Gardner's established criteria, considered the gold standard procedure. These boundaries were originally formulated based on the minimum frequency response required to faithfully reproduce what he considered the average arterial pressure waveform. However, these criteria do not necessarily remain optimal across diverse physiological states, extreme hemodynamic profiles, and modern digital filtering techniques. This limitation aligns with the findings of Devasahayam et al. (2022), who demonstrated through numerical simulations that Gardner's fixed boundaries do not account for dynamic physiological variations [9]. They showed that systolic measurement errors are heavily dependent on heart rate and the systolic rise-rate, reporting that systems technically classified as adequate under static criteria could generate systolic errors exceeding 18 mmHg during tachycardic or hyperdynamic states.

Additionally, it is imperative to acknowledge the limitations imposed by using a single monitor brand and the restricted number of patients, the fact that data were collected at a single ICU, and that a single person performed the manual flush test and peak identification. Also, patient demographic and clinical variables (e.g., age, sex, comorbidities) were not systematically recorded in this pilot phase, precluding analysis of their association with dynamic response classification.

Although the constructed database provides a diversity of arterial waveform morphologies and artifact profiles suitable for preliminary algorithmic validation, the small sample size compromises its statistical generalizability. Consequently, it is not feasible to accurately characterize the overall quality of IBP monitoring within the ICU of the HU. While the current database fulfills its primary engineering objective (serving as an initial testing ground for failure-detection performance), determining the true prevalence of adequate or inadequate catheter-transducer systems in this clinical setting would re-

quire a broader, longitudinal study with a significantly larger sample size. Nevertheless, broader clinical screenings in the literature justify the clinical relevance of the present engineering objective. Devasahayam et al. (2022) evaluated 121 fast-flush tests in a surgical ICU and found that a significant proportion of active patient lines fell into unacceptably underdamped or overdamped regions [9]. This high prevalence of undetected monitoring errors reinforces the critical need for automated failure-detection algorithms such as the one our database aims to validate in the future. As this study focuses on database construction rather than algorithm evaluation, no sensitivity, specificity, or ROC analysis is presented; these will be addressed in the planned validation phase.

VI. Conclusion

This study established an initial framework for creating a clinical database of invasive blood pressure waveforms. Although the current sample size is limited, the characterized real-world data effectively captures common dynamic response distortions, bridging the translational gap required for the upcoming clinical validation of AI-based detection algorithms within the Brazilian Sistema Único de Saúde (SUS) context.

Moving forward, the goal is to expand data acquisition with an automated annotation pipeline across broader clinical samples (targeting a wider range of natural frequencies, damping coefficients, and artifacts) to enhance statistical generalizability and incorporate diverse patient demographics, evaluating the impact of variables such as age distribution and biological sex. This expansion is essential to ensure a comprehensive, unbiased, and robust future validation of the AI model across heterogeneous ICU populations.

VII. Acknowledgment

This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (136156/2025-8 to AMKC and 309464/2022-6 to HTM). The authors are deeply grateful to the healthcare professionals and staff at HU for their hospitality, essential support in the ICU environment, and for making the clinical data collection possible.

References

[1] S.-H. Kim, M. Lilot, K. S. Sidhu, J. Rinehart, Z. Yu, C. Canales, and M. Cannesson, "Accuracy and precision of continuous noninvasive arterial pressure monitoring compared with invasive arterial pressure: A systematic review and meta-analysis," *Anesthesiology*, vol. 120, no. 5, pp. 1080–1097, 2014.

[2] B. Lamia, D. Chemla, C. Richard, and J.-L. Teboul, "Clinical review: Interpretation of arterial pressure wave in shock states," *Critical Care*, vol. 9, no. 6, pp. 601–606, 2005.

[3] L. Evans, A. Rhodes, W. Alhazzani, et al., "Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021," *Intensive Care Medicine*, vol. 47, pp. 1181–1247, 2021.

[4] F. Michard, "Towards the automatic detection and correction of abnormal arterial pressure waveforms," *Journal of Clinical Monitoring and Computing*, vol. 38, pp. 749–752, 2024.

[5] R. M. Gardner, "Direct blood pressure measurement—dynamic response requirements," *Anesthesiology*, vol. 54, no. 3, pp. 227–236, 1981.

[6] J. Rinehart, J. Tang, J. Nam, et al., "Detection of arterial pressure waveform error using machine-learning trained algorithms," *Journal of Clinical Monitoring and Computing*, vol. 36, pp. 743–750, 2022.

[7] P. H. Charlton, J. M. Harana, S. Vennin, et al., "Modeling arterial pulse waves in healthy aging: a database for in silico evaluation of hemodynamics and pulse wave indexes," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 317, no. 5, pp. H1062–H1085, 2019.

[8] J. Alastruey, J. Aramburu, P. Charlton, et al., "Nektar1D Reference Manual," King's College London, 2024.

[9] S. Devasahayam, N. Gangadharan, C. Surekha, B. Baskaran, F. A. Mukadam, and S. Subramani, "Intra-arterial blood pressure measurement: sources of error and solutions," *Medical & Biological Engineering & Computing*, vol. 60, pp. 1123–1138, 2022.

Data Above the Model: A Study of Three Deep Learning Approaches for Pulmonary Disease Detection in Chest Radiographs

João Igor Andrade
Student
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
joaoigorandrade@edu.unifor.br

João Victor Lopes Carvalho
Student
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
jvlopes200678@gmail.com

José Fernando Rodrigues Ferreira Neto
Graduate Student
Graduate Program in Applied Informatics
University of Fortaleza
Fortaleza, Brazil
josefernando@unifor.br

Belmondo Rodrigues Aragão Junior
Professor
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
belmondo@unifor.br

Natalia Bitar da Cunha Olegario
Professor
Center of Health Sciences
University of Fortaleza
Fortaleza, Brazil
nataliabitar@unifor.br

Joel Sotero da Cunha Neto
Professor
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
joelsotero@unifor.br

Isaac Newton Montinegro Linhares
Student
Center of Technological Sciences
University of Fortaleza
Fortaleza, Brazil
isaacnewton.ml@edu.unifor.br

Guilherme Pinheiro Ferreira da Silva
Professor
Center of Health Sciences
University of Fortaleza
Fortaleza, Brazil
guilhermepinheiro87@unifor.br

Abstract*—This study presents a three-stage experimental investigation of deep learning approaches for pulmonary disease detection in chest radiographs. The first stage, based on experiments with the NIH ChestX-ray14 dataset, included a direct multiclass classifier with a maximum accuracy of 41%, a cascaded architecture with binary screening whose AUC remained approximately 0.76 across multiple architectures (DenseNet-121, VGG-16, ResNet-50, and EfficientNet-B3), and a multilabel sweep using high-capacity architectures (EfficientNetV2-L, ConvNeXt-XL, SwinV2-B, and MaxViT-L), which did not exceed a mean AUROC of 0.774. The second stage incorporated a TorchXRayVision DenseNet-121 model pretrained specifically on radiographic data and applied it to pathology-specific classification within the NIH-based cascade. For pneumothorax, recall reached 0.92 after reducing the decision threshold to 0.37, accompanied by a substantial increase in false-positive predictions. The third stage evaluated a different experimental setting: a compact CNN trained from scratch on the curated Kermany pediatric pneumonia dataset, combined with a six-step OpenCV preprocessing pipeline, achieved an AUROC of 0.959 and recall of 0.997 for binary NORMAL versus PNEUMONIA classification. These stages do not constitute a controlled comparison, since they differ in patient population, task definition, label-generation procedure, preprocessing, and model configuration. Therefore, the observed performance differences cannot be attributed to data quality alone. Nevertheless, when considered together with previous evidence regarding inconsistencies in NIH ChestX-ray14 labels, the experiments highlight label consistency, dataset curation, and task definition as important factors that should be considered alongside model capacity when developing chest-radiograph classifiers.

***Index Terms*—chest radiography, deep learning, convolutional neural networks, label noise, NIH ChestX-ray14, pneumonia detection, TorchXRayVision**

I. Introduction

Chest radiography is among the most widely performed and accessible medical imaging examinations worldwide, making it a central modality for screening and diagnosis of thoracic diseases [1], [2]; however, its interpretation remains challenging and subject to interobserver variability. Findings such as pulmonary nodules may measure only a few millimeters and can be obscured by dense anatomical structures. The shortage of radiologists and the increasing volume of imaging examinations motivate the development of automatic decision-support systems.

Convolutional Neural Networks (CNNs) have substantially advanced medical image analysis [3], and several studies have applied them to chest radiography [1], [2]. A recurring but often underemphasized obstacle is the quality of training labels. Our experimental trajectory illustrates this issue in a concrete manner.

We investigated three experimental stages in chronological order: (i) **previous approaches** using the NIH ChestX-ray14 dataset [1], including direct multiclass classification, a cascaded architecture, and a multilabel formulation; (ii) an **intermediate approach** incorporating TorchXRayVision [4], pretrained on chest radiographs while retaining NIH data; and (iii) a **new approach** based on the Kermany dataset [5], in which a six-step OpenCV preprocessing pipeline was combined with a compact CNN trained from scratch for binary pneumonia classification.

Importantly, these stages should not be interpreted as a controlled comparison in which only data quality changes. The NIH and Kermany experiments differ in several dimensions, including patient population, label-generation procedure, task definition, number and heterogeneity of target conditions, preprocessing, image resolution, and model configuration. The Kermany experiment therefore represents a contrasting experimental setting rather than an isolated test of the effect of label quality.

The main contributions of this paper are: (i) an empirical characterization of the performance obtained by several architecture families under the evaluated NIH ChestX-ray14 formulations; (ii) an analysis showing that radiography-specific pretraining with TorchXRayVision did not substantially resolve the difficulties observed in the NIH experiments; (iii) the evaluation of a compact CNN combined with a six-step preprocessing pipeline on a curated binary pneumonia dataset; and (iv) a comparative discussion, supported by previous analyses of ChestX-ray14, of how label consistency, task formulation, population, preprocessing, and model selection may jointly influence observed performance. Rather than establishing a causal advantage of one dataset over another, the study emphasizes the importance of considering these factors together when interpreting cross-dataset results.

II. Background and Related Work

A. Deep learning in chest radiography

The release of large datasets such as ChestX-ray14 [1] enabled the training of deep models for thoracic disease classification. CheXNet [2] popularized DenseNet-121 [6] for this task. Architectures such as VGG [7], ResNet [8], EfficientNet [9], and more recent models such as ConvNeXt [10] and EfficientNetV2 [11] have also been widely evaluated. The TorchXRayVision library [4] provides models pretrained specifically on chest radiographs from multiple sources, offering domain-specific representations.

B. Inconsistencies in NIH ChestX-ray14

The labels in ChestX-ray14 were *not* produced through direct image review. Instead, they were automatically extracted from associated radiology reports using natural language processing (NLP) [1]. Because radiology reports reflect the broader clinical context rather than only the visual findings present in the image, they constitute an imperfect proxy for what is actually visible in the radiograph [12].

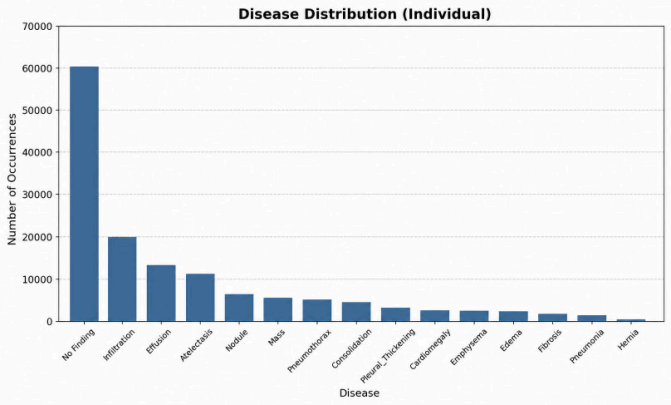


Figure 1. Distribution of NIH ChestX-ray14. The dominance of *No Finding* and the underrepresentation of pneumonia highlight the class imbalance.

An independent visual audit found substantially lower agreement with the pathology actually present in the image than the accuracy reported in the original paper (85–99%, validated only against report text). For instance, the agreement was approximately 36% for atelectasis and approximately 10% for fibrosis [12]. Peer-reviewed studies explicitly assume approximately 10% label noise and state that this makes it difficult to assess the true performance of any network [13].

A second issue is the presence of inconsistent and overlapping class definitions. Conditions such as infiltration, consolidation, and pneumonia are radiologically difficult to differentiate even for specialists [12]. The case of pneumothorax is particularly illustrative: the label often marks already treated cases in which a chest drain is visible in the image; gradient-based analyses show that the network learns to detect the drain rather than the acute condition [13]. Crucially, this means that a respectable AUC can coexist with clinically meaningless predictions: the model is not robust to structured label noise [12].

III. Materials and Methods

A. Datasets

NIH ChestX-ray14 [1]: a repository containing more than 112,000 chest radiographs annotated with 14 categories of radiological findings. In addition to these pathological labels, the dataset includes the *No Finding* label, which denotes examinations without any of the annotated findings. In the full dataset, *No Finding* exceeds 60,000 images, whereas pathologies such as pneumonia contain fewer than 1,000 samples. This severe imbalance makes direct training highly challenging (Fig. 1).

For the previous approaches, we extracted a subset of 9,419 images focused on five pathologies (Consolidation,

Mass, Nodule, Pneumothorax, and Pneumonia), reorganized for binary screening: 4,000 healthy images (42.5%) and 5,419 diseased images (57.5%). Within the "diseased" class, internal imbalance was substantial: Nodule (1,924, 34.3%) versus Pneumonia (300, 5.4%), a ratio of nearly 6:1 (Fig. 2).

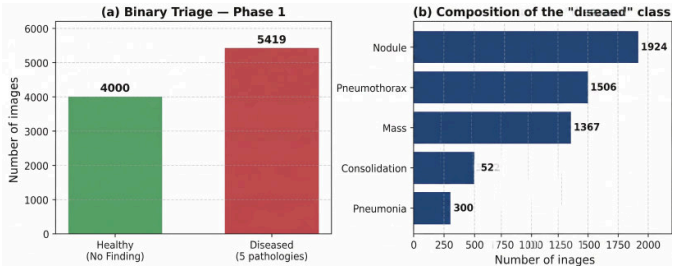


Figure 2. Distribution of the NIH subset used in the previous approaches: (a) binary healthy/diseased screening; (b) internal composition of the "diseased" class by pathology, showing marked imbalance.

For the multilabel formulation, we used a different subset comprising 20,025 images distributed across 13 pathological finding classes. Thus, the *No Finding* label was not included in this setting, and among the 14 original pathological findings of NIH ChestX-ray14, the *Hernia* class was excluded due to its limited sample size. Each retained class had approximately 1,200 training samples.

Curated pneumonia dataset [5]: a binary NORMAL vs. PNEUMONIA collection with *expert-validated labels*. The partition used in our experiments contains 5,216 training images, 16 validation images, and 624 test images (Fig. 3). In contrast to the heterogeneous thoracic findings considered in the NIH experiments, this dataset defines a narrower binary pneumonia classification task and consists of a different patient population. Therefore, performance values obtained across the NIH and Kermany settings should not be interpreted as a controlled comparison of label quality alone.

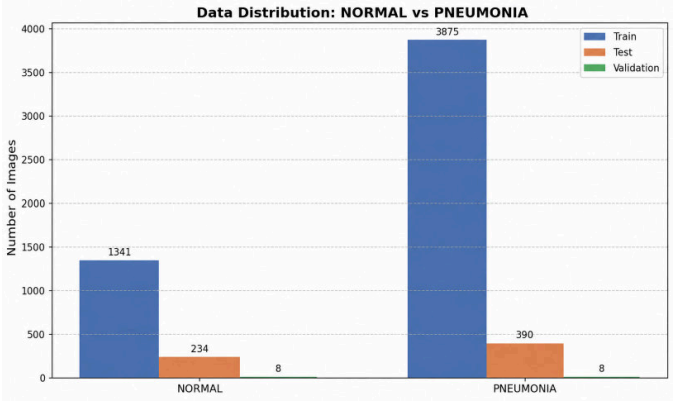


Figure 3. Class distribution of the curated pneumonia dataset in the training, test, and validation partitions. Labels were generated by radiologists rather than by NLP.

Table I summarizes the main differences among the experimental settings. These differences are treated as potential confounding factors when interpreting cross-stage performance.

B. Previous approaches: cascade and multilabel formulation on NIH

The initial approach attempted to classify the five pathologies directly using a single five-class classifier, reaching only 41% accuracy. This result motivated a cascaded architecture: a binary screening stage (Phase 1, "healthy vs. diseased") followed by pathology-specific binary verifiers (Phase 2).

For the initial NIH-based experiments, the selected image subsets were randomly partitioned at the image level, using approximately 80% of the samples for training and 20% for evaluation, with stratification according to the target classes. No patient-level grouping was enforced during this partitioning. Because ChestX-ray14 contains anonymized patient identifiers that may link multiple examinations to the same patient, the absence of patient-level separation is considered a limitation of these experiments.

In **Phase 1**, we evaluated four ImageNet-pretrained architectures—DenseNet-121, VGG-16, ResNet-50, and EfficientNet-B3—at a resolution of 384×384, using class-weighted `BCEWithLogitsLoss`, stratified sampling, and threshold calibration. During the preliminary experiments, an inversion in the class weighting assigned to the `pos_weight` parameter produced opposite collapse modes across three consecutive runs (v1–v3). This configuration issue was corrected in v4.

In the initial **Phase 2** experiments, prior to TorchXRayVision, we evaluated EfficientNet-B2 and DenseNet-121 variants fine-tuned for each disease, where the negative class consisted of diseased images from other pathologies rather than healthy images.

As an additional countercheck, we formulated the problem as a **13-class multilabel** task using the subset of 20,025 images, sweeping modern architectures: EfficientNetV2-L [11], ConvNeXt-Large/XL [10], SwinV2-B, and MaxViT-L, under three training regimes: *head-only*, *full fine-tuning*, and last-block unfreezing.

C. Intermediate approach: TorchXRayVision on NIH

For Phase 2 of the cascade, we incorporated the TorchXRayVision [4] DenseNet-121 model pretrained on multiple radiographic datasets (`densenet121-res224-all`), replacing the classifiers trained from scratch in the previous experiments. The hypothesis was that representations pretrained on domain-specific data would outperform ImageNet weights for radiological pathologies.

The model was adapted for each disease using *partial fine-tuning*: only the final block (`denseblock4`) and the classification head (2 outputs) were trained, while the remaining layers were kept frozen. We used `CrossEntropyLoss` with class weights, the AdamW optimizer, a CosineAnnealingLR scheduler, and early stopping based on `val_f1_binary`. The training set was balanced through *SAMPLE_MODE*, with

Table I
COMPARISON OF THE MAIN EXPERIMENTAL SETTINGS CONSIDERED IN THIS STUDY.

Characteristic	Previous NIH experiments	NIH + TorchXRayVision	Kermany experiment
Population	Predominantly adult chest radiographs	Predominantly adult chest radiographs	Pediatric chest radiographs
Label generation	Automatically derived from radiology reports using NLP	Same NIH label-generation procedure	Expert-validated labels
Main task	Multiclass, binary healthy/diseased, and multilabel classification	Binary pathology-specific classification; pneumothorax reported here	Binary NORMAL vs. PNEUMONIA classification
Disease scope	Multiple heterogeneous thoracic findings	Single pathology within the NIH dataset	Pneumonia
Model configuration	Multiple ImageNet-pretrained and high-capacity architectures	Radiography-pretrained DenseNet-121	Compact CNN trained from scratch
Image resolution	384×384 in Phase 1; configuration-dependent in other experiments	224×224	224×224
Preprocessing	NIH experiment-specific processing	TorchXRayVision-compatible processing	Six-step OpenCV pipeline
Controlled cross-dataset comparison	No	No	No

1,000 positive and 1,000 negative samples for training and 200 positive and 200 negative samples for validation.

The results reported here refer to **Pneumothorax**, selected because of its clinical relevance and its direct connection to the literature on noisy labels (Section II-B). In addition to the default decision threshold of 0.50, a threshold search was performed to maximize binary F1, identifying an adjusted operating point of 0.37. Both operating points are reported to characterize the trade-off between positive-class sensitivity and false-positive predictions.

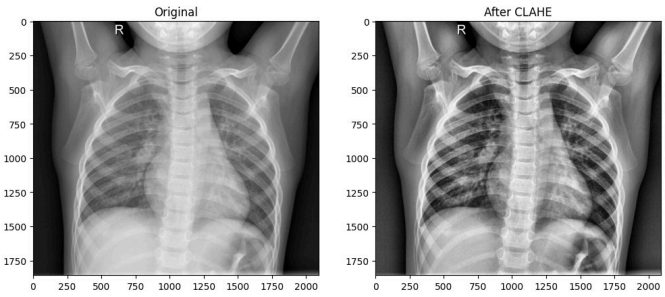


Figure 4. Effect of CLAHE on a chest radiograph: original image (left) and contrast-enhanced image (right). CLAHE makes local lung textures more visible.

D. New approach: OpenCV pipeline and CNN on the Kermany dataset

Given the limitations observed across the NIH experiments, we evaluated a distinct experimental setting using the curated Kermany dataset [5]. This dataset provides a well-defined binary NORMAL vs. PNEUMONIA task with expert-validated labels. The experiment should be interpreted as a separate configuration rather than as a controlled comparison with NIH, since the two settings differ in population, task definition, labeling procedure, preprocessing, and model configuration.

We designed a **six-step OpenCV preprocessing pipeline** applied to each grayscale radiograph: (1) scaling to $[0, 1]$; (2) normalization using training-set statistics (mean 0.477; standard deviation 0.233); (3) contrast enhancement with CLAHE [14] (`clipLimit`$= 2.0$; 8×8 tiles), which increases the salience of local image structures (Fig. 4); (4) validation and removal of corrupted or nearly black images; (5) resizing to 224×224 with padding to preserve aspect ratio; and (6) mild denoising using a 3×3 median filter.

The classifier is a **compact CNN trained from scratch**: three convolutional blocks (Conv–ReLU–MaxPool, with $16/32/64$ filters) followed by a fully connected layer with 128 units, *dropout* (0.5), and a two-class output. It was trained using the RAdam optimizer (learning rate 10^{-3}), cross-entropy loss, and early stopping with a patience of 3, which stopped training at epoch 6.

The validation partition contains only 16 images, which limits the stability of validation-based model selection. Normalization statistics (mean 0.477 and standard deviation 0.233) were derived from the training partition. Because no preprocessing ablation was performed, the reported performance should be attributed to the combined OpenCV pipeline and compact CNN rather than to any individual preprocessing step.

E. Evaluation protocol

We report accuracy, precision, recall, macro-F1, AUROC/AUC, and confusion matrices. For the initial NIH experiments, image subsets were randomly partitioned using an approximately $80/20$ training/evaluation split, with stratification according to the target classes and without patient-level grouping. Per-class recall was additionally monitored

in the binary screening experiments to identify class-specific prediction collapse.

For threshold-dependent analyses, classifier performance was examined at specific operating points. In the TorchXRayVision pneumothorax experiment, the default threshold of 0.50 was compared with an adjusted threshold of 0.37, identified by maximizing binary F1. For the multilabel experiment, the class-wise confusion matrices were generated using Youden's J thresholds.

For the Kermany experiment, accuracy, AUROC, precision, recall, F1-score, specificity, and negative predictive value (NPV) were evaluated on the 624-image test partition. Because the experimental stages differ in dataset, task definition, population, preprocessing, and model configuration, their metrics characterize each setting rather than constitute a controlled cross-dataset comparison.

IV. Results

A. Previous approaches on NIH: cascade and multilabel

Table II summarizes the experimental progression on NIH.

Table II
EXPERIMENTAL PROGRESSION ON NIH CHESTX-RAY14. PHASE 0: DIRECT MULTICLASS CLASSIFICATION. PHASE 1: BINARY SCREENING (HEALTHY VS. DISEASED). R_d AND R_s: RECALL FOR DISEASED AND HEALTHY CLASSES.

Model	Acc	AUC	F1-m	R_d/R_s	Note
Phase 0 — direct 5-class classification					
DenseNet-121	0.41	—	0.39	—	overfitting
VGG-16	0.38	—	0.34	—	lower performance
Phase 1 — binary screening (healthy vs. diseased)					
DenseNet v1	0.65	—	0.61	0.86/0.38	diseased bias
DenseNet v2	0.56	0.762	0.52	0.24/0.97	pos_weight issue
DenseNet v3	0.58	0.745	0.38	1.00/0.01	opposite collapse
DenseNet v4	**0.70**	0.759	**0.69**	0.69/0.70	corrected setup
EfficientNet-B3	0.57	**0.765**	0.37	1.00/0.01	collapse
ResNet-50	0.69	0.756	0.68	0.68/0.70	mild overfitting

The direct multiclass baselines achieved 0.41 accuracy for DenseNet-121 and 0.38 for VGG-16, motivating the subsequent cascaded formulation. In Phase 1, the `pos_weight` inversion produced opposite failure modes in versions v2 and v3, with diseased recall of 0.24 in one configuration and healthy recall of 0.01 in another. After correcting this configuration in v4, DenseNet-121 produced the most balanced class-wise behavior among the evaluated Phase 1 models, with accuracy of 0.70, AUC of 0.759, macro-F1 of 0.69, diseased recall of 0.69, and healthy recall of 0.70. Across the evaluated configurations, AUC values remained approximately within the 0.75–0.76 range. EfficientNet-B3 achieved the highest AUC (0.765), but its healthy-class recall was only 0.01, illustrating that ranking performance alone does not guarantee balanced classification at a particular operating point.

Fig. 5 illustrates the limited variation in AUC across the evaluated Phase 1 configurations, with the valid models remaining approximately within the 0.75–0.76 range.

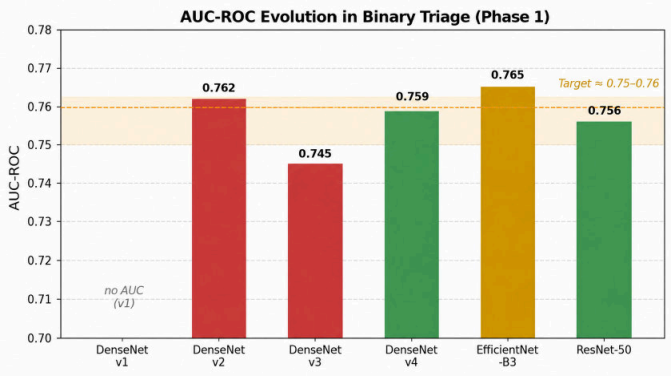


Figure 5. AUC-ROC evolution in binary screening (Phase 1). The evaluated valid configurations remained approximately within the 0.75–0.76 range.

To evaluate whether higher-capacity architectures produced substantial improvements under a different NIH formulation, we assessed the 13-class multilabel setting. The best mean AUROC was 0.774 with EfficientNetV2-L using *full fine-tuning*, while the five best models ranged from 0.759 to 0.774 (Table III).

Table III
MULTILABEL FORMULATION (13 CLASSES): MEAN AUROC ON THE TEST SET. *Full fine-tuning* UNLESS OTHERWISE INDICATED.

Model	Strategy	Mean AUROC
EfficientNetV2-L	full fine-tuning	**0.774**
ConvNeXt-XL	full fine-tuning	0.771
EfficientNet-B4	full fine-tuning	0.768
SwinV2-B	full fine-tuning	0.761
MaxViT-L	full fine-tuning	0.759
DenseNet-121	full fine-tuning	0.755
ConvNeXt-Large	last block	0.709

Even high-capacity architectures such as ConvNeXt-XL, SwinV2-B, and MaxViT-L did not substantially exceed this performance range in the evaluated multilabel setting. Class-wise performance was heterogeneous: findings such as emphysema, edema, and cardiomegaly reached AUROC values of 0.85–0.93, whereas pneumonia and fibrosis remained below 0.80 (Fig. 6). Within the NIH settings evaluated in this study, increasing model capacity or changing the architecture family therefore produced limited overall improvements. Because these experiments share the same underlying dataset and label-generation process, the results motivate examination of additional factors, including label consistency, task formulation, and class composition.

B. Intermediate approach: TorchXRayVision on NIH

The TorchXRayVision DenseNet-121 model (`densenet121-res224-all`), with *partial fine-tuning*, was applied to individual pneumothorax classification in Phase 2 of the cascade. The training curves (Fig. 7) show stable learning over 70 epochs, with decreasing training

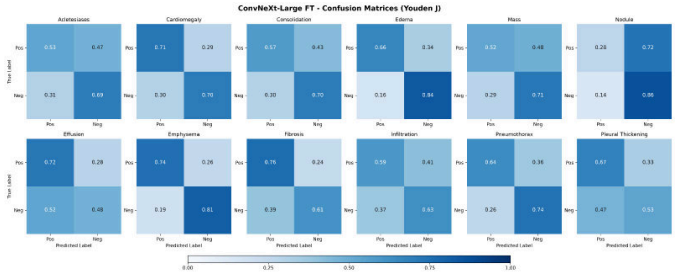


Figure 6. Class-wise confusion matrices using Youden's J threshold for the multilabel ConvNeXt-Large model. Pneumonia and infiltration show substantial classification errors among the evaluated findings.

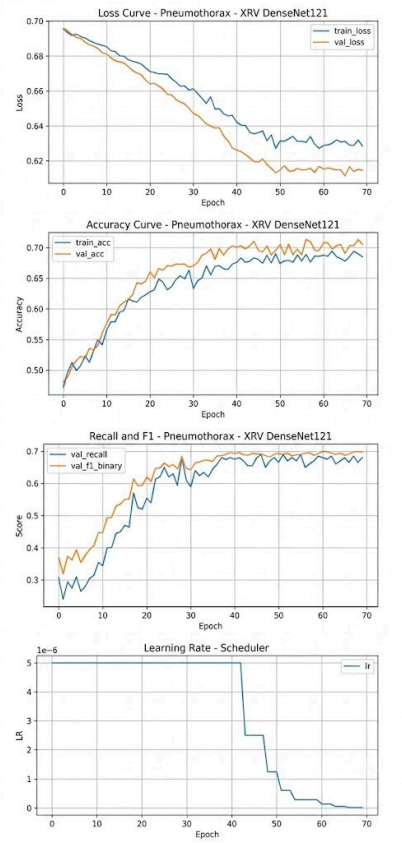


Figure 7. Training curves of TorchXRayVision DenseNet-121 for pneumothorax classification (Phase 2 of the cascade): loss, accuracy, recall/F1, and learning rate over 70 epochs.

and validation losses and validation accuracy converging to approximately 0.70.

With the default threshold of 0.50, the model achieved 0.71 accuracy, with a confusion matrix of 152/48/67/133 (TN/FP/FN/TP) and class-wise recall of 0.76/0.67 for the negative and positive classes, respectively (Fig. 8).

At the adjusted threshold of 0.37, identified by maximizing binary F1, positive-class recall increased to 0.92, while precision decreased to 0.58 and overall accuracy to 0.63. The increase in sensitivity was accompanied by a substantial rise in false-positive predictions and deterioration of negative-class performance (Fig. 9).

Table IV summarizes both operating points. Reducing the threshold from 0.50 to 0.37 increased recall from 0.67 to 0.92, but reduced precision from 0.74 to 0.58 and accuracy from

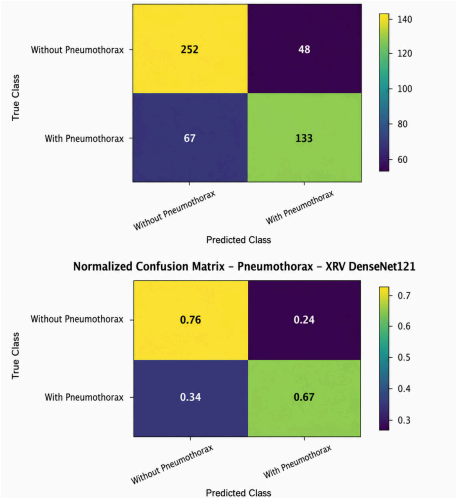


Figure 8. Confusion matrices of TorchXRayVision DenseNet-121 for pneumothorax at the default threshold of 0.50.

Search mode: #1
Best threshold: 0.37
Metrics with best threshold:
{'accuracy': 0.6325, 'precision': 0.5841269841269842, 'recall': 0.92, 'f1_binary': 0.7145631067961165, 'f1_macro': 0.5993868165559529}

Classification report with adjusted threshold:

	precision	recall	f1-score	support
Without Pneumothorax	0.81	0.34	0.48	200
With Pneumothorax	0.58	0.92	0.71	200
accuracy			0.63	400
macro avg	0.70	0.63	0.60	400
weighted avg	0.70	0.63	0.60	400

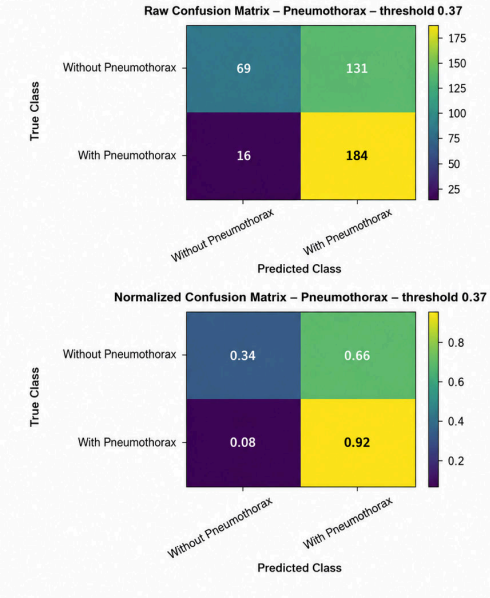


Figure 9. Classification report and confusion matrices at the adjusted threshold of 0.37. Positive-class recall increases to 0.92, while precision and negative-class performance decrease.

0.71 to 0.63, while binary F1 remained approximately stable.

Table IV
TORCHXRAYVISION DENSENET-121 RESULTS FOR PNEUMOTHORAX IN PHASE 2.

Threshold	Acc	Prec	Rec	F1	Macro-F1
0.50	0.71	0.74	0.67	0.70	0.71
0.37	0.63	0.58	**0.92**	0.71	0.60

These results illustrate the trade-off associated with decision-threshold adjustment. The 0.37 operating point should therefore be interpreted as a high-recall configuration within this experiment rather than as a clinically validated threshold. This behavior is also consistent with concerns documented in the literature regarding NIH pneumothorax labels [13]. Previous studies have reported that some labeled pneumothorax images correspond to already treated cases containing visible chest drains, creating the possibility that classifiers associate treatment-related artifacts with the target label. The present experiments do not directly determine which visual features were used by the TorchXRayVision model; however, radiography-specific pretraining alone was insufficient to eliminate the observed classification difficulties.

C. New approach: compact CNN on the Kermany dataset

The compact CNN combined with the six-step OpenCV preprocessing pipeline achieved an AUROC of 0.959, accuracy of 79.7%, precision of 0.755, recall of 0.997, and F1-score of 0.860 on the Kermany test set (Fig. 10). These values characterize the complete preprocessing-and-classification configuration, since the contribution of each preprocessing step was not evaluated separately.

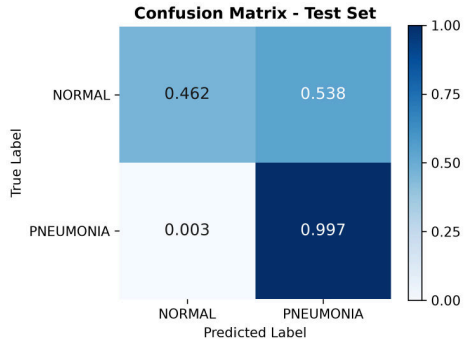


Figure 10. Normalized confusion matrix on the Kermany test set (NORMAL vs. PNEUMONIA). The evaluated operating point achieves very high pneumonia sensitivity while producing a substantial proportion of false-positive predictions among NORMAL images.

Table V summarizes the test-set results. The model achieved AUROC 0.959 and recall 0.997, but specificity was only approximately 0.46, revealing a sensitivity-oriented operating point with a substantial false-positive rate.

Table V
TEST-SET RESULTS (624 IMAGES) FOR THE COMPACT CNN WITH THE SIX-STEP OPENCV PIPELINE ON KERMANY.

Acc	AUROC	Prec	Rec	Spec	NPV	F1
0.797	**0.959**	0.755	**0.997**	≈0.462	≈0.991	0.860

Early stopping halted training at epoch 6. Because the validation partition contains only 16 images, the selected stopping point is sensitive to individual predictions and should be interpreted cautiously. Although the final test-set recall indicates that the evaluated configuration rarely missed pneumonia cases within this dataset, this occurred together with limited specificity.

V. DISCUSSION

Across the evaluated NIH settings, increasing model capacity or changing the architecture family produced limited gains. Binary screening remained around AUC 0.75–0.76, the multilabel experiments reached a maximum mean AUROC of 0.774, and the TorchXRayVision pneumothorax experiment reached recall 0.92 at the adjusted operating point of 0.37, at the cost of substantially more false positives. These findings indicate that architecture or radiography-specific pretraining alone was insufficient to resolve the difficulties observed under the evaluated NIH formulations.

Previous analyses have documented limitations in ChestX-ray14 labels, which were extracted from radiology reports rather than generated through direct image annotation [12], [13]. Our results are consistent with concerns regarding ambiguous or indirect labels, but they do not establish label quality as the sole cause of the NIH performance profile. In particular, the Kermany experiment represents a different setting: a curated pediatric binary NORMAL vs. PNEUMONIA task, with different population, preprocessing, architecture, and training procedure. Therefore, its AUROC of 0.959 cannot be directly interpreted as evidence that label quality alone explains the improvement.

The Kermany results should also be interpreted as the performance of the combined six-step preprocessing pipeline and compact CNN because no preprocessing ablation was performed. The distinction between discrimination and operating-point performance is important: an AUROC of 0.959 indicates strong ranking ability across possible thresholds, but it does not imply balanced classification at the evaluated operating point. Recall reached 0.997, whereas specificity was approximately 0.462, indicating that the high sensitivity was accompanied by a substantial false-positive rate. The validation partition contains only 16 images, further limiting the stability of early stopping and model-selection decisions.

The NIH experiments also have limitations related to data partitioning and threshold-dependent evaluation. The initial NIH subsets were randomly partitioned at the image level using an approximately 80/20 split, without grouping examinations by patient identifier. Consequently, images from the same patient may occur in different partitions, potentially

increasing similarity between training and evaluation samples. In addition, the improvement in pneumothorax recall obtained by reducing the decision threshold from 0.50 to 0.37 illustrates the strong dependence of sensitivity and specificity on the selected operating point. These threshold-dependent results should therefore be interpreted together with threshold-independent metrics such as AUROC rather than in isolation.

Overall, the experiments suggest that dataset curation, label consistency, task definition, population, preprocessing, partitioning strategy, and model design should be considered jointly. The NIH–Kermany comparison is observational rather than a controlled ablation; isolating the effect of label quality would require the same pneumonia task, architecture, preprocessing, resolution, patient-aware partitioning strategy, and evaluation protocol on both datasets. External validation would also be required before conclusions regarding clinical applicability.

VI. Conclusion

This paper presented a three-stage experimental investigation of deep learning approaches for pulmonary disease detection in chest radiographs. Across the evaluated NIH ChestX-ray14 settings, including direct multiclass classification, binary screening, multilabel classification, and radiography-specific TorchXRayVision pretraining, changing or increasing model capacity produced relatively limited performance gains. In the pneumothorax experiment, recall increased to 0.92 at the adjusted threshold of 0.37, but this improvement was accompanied by a substantial increase in false-positive predictions.

In a separate experimental setting, a compact CNN combined with a six-step OpenCV preprocessing pipeline achieved an AUROC of 0.959 and recall of 0.997 for binary pneumonia classification on the curated Kermany dataset. However, this operating point also showed limited specificity, and the Kermany setting differs from the NIH experiments in patient population, task definition, labeling procedure, preprocessing, and model configuration. The observed cross-setting performance difference therefore cannot be attributed exclusively to data quality or label consistency.

Overall, the results show that increasing architecture complexity alone did not substantially improve performance within the evaluated NIH settings, whereas a simpler model achieved stronger discriminative performance under a narrower and differently curated experimental configuration. Together with previous evidence regarding inconsistencies in ChestX-ray14 labels [12], [13], these findings reinforce the importance of considering label reliability, dataset curation, task definition, population, preprocessing, and model design jointly rather than interpreting model capacity in isolation.

Future work should include a controlled pneumonia-versus-normal comparison between NIH ChestX-ray14 and Kermany using the same architecture, preprocessing, input resolution, data-partitioning strategy, and evaluation protocol. Additional experiments should also investigate the individual contribution of the OpenCV preprocessing stages through ablation studies and assess model generalization on independent external cohorts before conclusions regarding clinical applicability are drawn.

References

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian *et al.*, "TorchXRayVision: A library of chest x-ray datasets and models," in *Proc. Int. Conf. Medical Imaging with Deep Learning (MIDL)*, 2022.

[5] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[10] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[11] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 10 096–10 106.

[12] L. Oakden-Rayner, "Exploring the ChestX-ray14 dataset: problems," Blog post, 2017, available: https://laurenoakdenrayner.com/2017/12/18/the-chestxray14-dataset-problems/.

[13] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest x-ray classification," *Scientific Reports*, vol. 9, no. 1, p. 6381, 2019.

[14] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press, 1994, pp. 474–485.

Deep Learning Applications in Post-Stroke Automatic Speech Analysis: A Literature Review

Thaynara Egidio dos Reis
Post Graduate Program in Sustainable Technologies
Federal Institute of Espírito Santo
Vitória, Brazil
egidiothaynara@gmail.com

Shirley Peroni Neves Cani
Post Graduate Program in Sustainable Technologies
Federal Institute of Espírito Santo
Vitória, Brazil
shrileypnc@ifes.edu.br

Carolina Fiorin Anhoque
Assoc. Prof. Dept. of Speech-Language Pathology
Federal University of Espírito Santo
Vitória, Brazil
carolina.anhoque@ufes.br

Mariana Rampinelli Fernandes
Post Graduate Program in Sustainable Technologies
Federal Institute of Espírito Santo
Vitória, Brazil
mariana.rampinelli@ifes.edu.br

***Abstract*—Stroke is the leading cause of disability worldwide and often leads to speech impairments such as aphasia and dysarthria. Recent advances in Artificial Intelligence have enabled the development of automated methods to support speech assessment in people after stroke. This study presents a literature review of Artificial Intelligence techniques applied to the automatic analysis of post-stroke speech. Twelve studies published between 2020 and 2026 were analyzed with respect to datasets, preprocessing methods, computational architectures, evaluation metrics, and reported results. Most studies employed deep learning approaches, particularly convolutional neural networks and hybrid models, achieving promising performance in speech assessment. However, common limitations included small datasets, methodological heterogeneity, limited external validation, and the scarcity of publicly available databases. These findings highlight the potential of Artificial Intelligence to support clinical assessment after stroke while emphasizing the need for more robust and standardized research.**

***Index Terms*—Stroke, Speech Analysis, Artificial Intelligence, Deep Learning, Aphasia, Dysarthria.**

I. Introduction

Stroke is a neurological emergency characterized by the interruption of cerebral blood flow due to the occlusion or rupture of a blood vessel, leading to brain tissue damage and neurological impairment. It is one of the leading causes of mortality and long-term disability worldwide. In 2021, an estimated 11.9 million new stroke cases and 93.8 million stroke survivors living with long-term sequelae were reported, making stroke the third leading cause of death and disability globally. Furthermore, approximately one in four individuals is expected to experience a stroke during their lifetime, highlighting its increasing burden on healthcare systems and quality of life worldwide [1]. Among the most common sequelae are motor, cognitive, and communication impairments which substantially affect social participation, functional independence, and the rehabilitation process. In this context, early multidisciplinary rehabilitation, including speech-language therapy, plays a fundamental role in functional recovery and improving patients' quality of life [1], [2].

Among the most disabling consequences of stroke are communication disorders, particularly aphasia and dysarthria. Aphasia affects language comprehension and/or production as a result of brain lesions, whereas dysarthria arises from impairments in the neuromuscular control of speech, impacting respiration, phonation, articulation, and resonance. Aphasia is estimated to affect approximately one-third of stroke survivors, while dysarthria is present in about 30–40% of patients in the acute phase and may persist for months after the event, significantly affecting speech intelligibility, social interaction, functional independence, and quality of life [3], [4]. In this context, speech-language assessment is essential to characterize impairments, monitor clinical evolution, and guide therapeutic planning. However, clinical practice is still characterized by heterogeneity in assessment methods, reliance on perceptual-auditory evaluation, and limitations related to the availability of specialized services, which hinder standardization of protocols and longitudinal patient monitoring [3], [4].

In light of these limitations, computational methods for automatic speech analysis have gained increasing attention as supportive tools in clinical practice. The advancement of Artificial Intelligence (AI), particularly techniques based on machine learning, deep learning, automatic speech recognition, and natural language processing, has enabled the development of systems capable of identifying acoustic and linguistic patterns associated with communication disorders, contributing to objective evaluation, monitoring of clinical progression, and decision-making support in therapeutic planning [5]. Beyond improving analytical precision, these approaches also offer potential to support remote monitoring and expand access to

This paper was produced with the assistance of artificial intelligence tools (OpenAI - ChatGPT) for text review, and English translation. The scientific content, literature analysis, article selection, interpretation of findings, and conclusions are the sole responsibility of the authors.

specialized care, especially in settings with limited availability of healthcare professionals.

Although the number of studies investigating automatic speech analysis in post-stroke individuals has increased substantially in recent years, considerable heterogeneity remains regarding the study populations, languages, speech tasks, datasets, pre-processing strategies, AI architectures, and performance metrics employed. This variability hinders direct comparison of the reported findings, limits the reproducibility of the proposed methods, and makes it difficult to identify the approaches with the greatest potential for clinical application. Therefore, a critical synthesis of the available evidence is needed to provide a comprehensive overview of the current state of the field, identify methodological trends, and highlight research gaps that may guide future investigations.

In this context, this study presents a narrative literature review on the application of AI techniques to automatic speech analysis in individuals affected by stroke. The objective was to critically analyze studies published between 2020 and 2026, comparing the datasets used, the investigated populations, speech acquisition protocols, preprocessing strategies, Artificial Intelligence architectures, evaluation metrics, and the main reported results. In addition, the study discusses the recurring limitations in the literature and the prospects for the development of more robust methods with greater potential for clinical application.

II. Methodology

This study presents a narrative literature review on the application of AI and deep learning techniques to automatic speech analysis in stroke-affected individuals, with a particular focus on communication disorders associated with aphasia and dysarthria.

A. Search Strategy and Data Sources

The literature search was conducted using the Scopus, Web of Science, PubMed, and Virtual Health Library (VHL) databases, selected for their broad coverage of publications in the fields of biomedical engineering, health sciences, and applied computing.

The search strategy was developed by combining four groups of keywords. The first group included terms related to stroke, namely “stroke”, “cerebrovascular accident”, and “post-stroke”. The second group included terms related to speech and language disorders, such as “dysarthria”, “aphasia”, “speech impairment”, and “speech disorder”. The third group comprised terms related to automatic speech analysis and processing, including “speech recognition”, “automatic speech analysis”, “acoustic analysis”, “voice-based”, “audio-based”, “speech analysis”, and “automatic speech recognition”. Finally, the fourth group consisted of terms related to deep learning techniques, namely “CNN”, “deep learning”, and “convolutional network”. The searches were conducted between April 25, 2026, and July 12, 2026, with the final search performed on July 12, 2026. This search was conducted by two reviewers, who obtained the same result.

For Scopus, the complete search strategy was: “(TITLE-ABS-KEY (stroke OR “cerebrovascular accident” OR “post-stroke”) AND TITLE-ABS-KEY (dysarthria OR aphasia OR “speech impairment” OR “speech disorder”) AND TITLE-ABS-KEY (“speech recognition” OR “automatic speech analysis” OR “acoustic analysis” OR “voice-based” OR “audio-based” OR “speech analysis” OR “automatic speech recognition”) AND TITLE-ABS-KEY (“CNN” OR “deep learning” OR “convolutional network”)) AND PUBYEAR $>$ 2019 AND PUBYEAR $<$ 2027 AND (LIMIT-TO (SRCTYPE , “j”)) AND (LIMIT-TO (DOCTYPE , “ar”)) AND (LIMIT-TO (LANGUAGE , “English”))”.

For Web of Science, the search was conducted in the Core Collection using the query builder with the following strategy: “TS=(“stroke” OR “cerebrovascular accident” OR “post-stroke”) AND TS=(“dysarthria” OR “aphasia” OR “speech impairment” OR “speech disorder”) AND TS=(“speech recognition” OR “automatic speech analysis” OR “acoustic analysis” OR “voice-based” OR “audio-based” OR “speech analysis” OR “automatic speech recognition”) AND TS=(“CNN” OR “deep learning” OR “convolutional network”) AND PY=(2020-2026) AND DT=(Article) AND LA=(English)”.

For PubMed, the following search strategy was used: “((”stroke”[Title/Abstract] OR “cerebrovascular accident”[Title/Abstract] OR “post-stroke”[Title/Abstract]) AND (“dysarthria”[Title/Abstract] OR “aphasia”[Title/Abstract] OR “speech impairment”[Title/Abstract] OR “speech disorder”[Title/Abstract]) AND (”speech recognition”[Title/Abstract] OR “automatic speech analysis”[Title/Abstract] OR “acoustic analysis”[Title/Abstract] OR “voice-based”[Title/Abstract] OR “audio-based”[Title/Abstract] OR “speech analysis”[Title/Abstract] OR “automatic speech recognition”[Title/Abstract]) AND (“CNN”[Title/Abstract] OR “deep learning”[Title/Abstract] OR “convolutional network”[Title/Abstract])) AND (“2020/01/01”[Date - Publication] : “2026/12/31”[Date - Publication]) AND english[Language] AND journal article[Publication Type]”.

For the Virtual Health Library (VHL), the following search strategy was used: “(ti:(stroke OR ”cerebrovascular accident” OR “post-stroke”) OR ab:(stroke OR ”cerebrovascular accident” OR “post-stroke”)) AND (ti:(dysarthria OR aphasia OR “speech impairment” OR “speech disorder”) OR ab:(dysarthria OR aphasia OR “speech impairment” OR “speech disorder”)) AND (ti:(“speech recognition” OR “automatic speech analysis” OR “acoustic analysis” OR “voice-based” OR “audio-based” OR “speech analysis” OR “automatic speech recognition”) OR ab:(“speech recognition” OR “automatic speech analysis” OR “acoustic analysis” OR “voice-based” OR “audio-based” OR “speech analysis” OR “automatic speech recognition”)) AND (ti:(“CNN” OR “deep learning” OR “convolutional network”) OR ab:(“CNN” OR “deep learning” OR ”convolutional network”))”.

In the VHL, the publication year (2020–2026), study type (journal article), and language (English) filters were applied through the platform settings.

Initially, studies were screened based on their titles and abstracts to determine whether they addressed voice, speech, or language analysis in individuals affected by stroke and used artificial intelligence techniques. Studies focused on classification, recognition, diagnosis, or automated speech analysis were considered eligible.

Studies were excluded if they were not directly related to stroke or communication disorders, relied exclusively on non-speech signals such as electroencephalography (EEG), medical imaging, or motor signals, or were review articles, editorials, book chapters, or publications without full-text availability that prevented adequate extraction of methodological data.

B. Study Selection Process

Titles and abstracts were initially screened, followed by full-text assessment of potentially eligible studies. For each database, a manual search was conducted for duplicate articles already identified in previous databases. This deduplication was performed manually by comparing the titles, DOIs, and authors of potentially duplicate articles. All this screening and selection process was conducted by two independent reviewers, and there were no disagreements between them.

The search of Scopus retrieved 13 records, of which 10 studies were included in the final analysis. Two potentially relevant articles could not be fully assessed because their full texts were unavailable, while one study was excluded because it relied exclusively on brain signals without performing voice or speech analysis. Thus, 10 articles were selected from Scopus.

The Web of Science search identified seven records. Of these, two additional articles were included in the review, whereas the five remaining records corresponded to studies previously identified in Scopus, and this correspondence was manually checked by comparing the titles and authors. These studies were therefore considered duplicates. Thus, two additional articles were included from the Web of Science search.

The PubMed search yielded six records, four of which had already been identified in the other databases and were therefore considered duplicates, again based on a manual comparison of titles and authors. Of the remaining two records, one was a review article and the other was excluded because it relied exclusively on brain signals rather than speech analysis and therefore did not meet the predefined eligibility criteria. No additional studies were included from PubMed.

The search conducted in the Virtual Health Library (VHL) identified six records. All of these records corresponded to articles previously retrieved from the other databases and were therefore considered duplicates, based on a manual check of titles and authors. Thus, no additional studies were included from the VHL search.

After screening and application of the eligibility criteria, 12 studies were included in the qualitative synthesis. Figure 1 presents the PRISMA flow diagram of the identification, screening, eligibility assessment, and inclusion process.

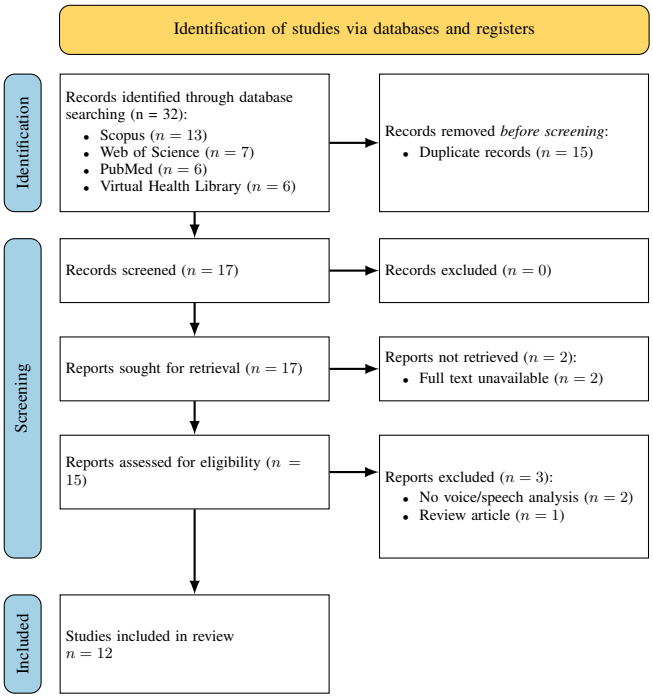


Figure 1. PRISMA flow diagram summarizing the study selection process.

C. Data Extraction and Analysis

For each included study, information was extracted on the year of publication, study objective, characteristics of the dataset, study population, speech tasks, preprocessing procedures, computational techniques and architectures, speech acquisition methods, extracted acoustic features, performance metrics, main findings, and the limitations reported by the authors. This stage was conducted by a reviewer.

The extracted data were summarized in comparative tables, enabling the systematic synthesis of the methodological characteristics and principal findings of the selected studies.

Data analysis was performed using a descriptive qualitative approach to identify trends in the application of AI and deep learning techniques to speech analysis in post-stroke individuals. Furthermore, research gaps and opportunities for the development of novel approaches to the automated assessment of speech impairments associated with aphasia and dysarthria were identified.

III. Results and Discussion

A. Overview of the Included Studies

Twelve studies published between 2020 and 2026 were included. The studies present different applications of artificial intelligence and deep learning for automatic speech analysis in the post-stroke context, ranging from diagnosis and screening to aphasia assessment, speech recognition, therapy, and assistive communication. This diversity of applications makes direct comparison between studies challenging, as the objectives, populations, speech tasks, and evaluation protocols differ across studies. Tables I and II provide an overview of the

datasets, population and stroke stage, language and speech task categories, application types, recording devices and conditions, extracted features, split and segmentation units, techniques, and primary outcomes reported by the included studies.

Overall, the identified studies can be grouped into four main application areas. The first comprises systems designed for stroke diagnosis or screening through automatic speech analysis. Within this category, the multimodal DeepStroke framework stands out for integrating speech and facial expression features to assist in the identification of stroke patients in emergency settings [10]. Similarly, [16] proposed an early diagnosis system based on sustained vowels and spontaneous speech, incorporating a speech enhancement module prior to the classification stage. Their findings suggest that reducing background noise can improve the performance of speech-based classification models.

A second research direction focuses on automatic speech recognition systems primarily targeting dysarthria, the most frequently addressed communication disorder among the included studies. Rather than merely detecting the presence of the disorder, these approaches aim to improve speech intelligibility or enhance the performance of automatic systems when applied to pathological speech. While some authors proposed hybrid architectures for dysarthric speech classification [9], [14], others investigated communication support strategies by comparing different speech processing approaches [11] or developing TinyML-based embedded devices for speech rehabilitation [13].

The third research direction encompasses studies focused on automated language assessment in individuals with aphasia. Unlike dysarthria-focused studies, which primarily emphasize speech intelligibility, these works aim to objectively reproduce tasks traditionally performed by speech-language pathologists during clinical assessment and follow-up. Within this context, researchers have developed models for the automatic assessment of aphasic speech [7], systems for verifying picture-naming tasks during speech therapy [8], and comparative evaluations of architectures specifically trained for pathological speech against commercial automatic speech recognition platforms, highlighting the potential advantages of specialized models for post-stroke populations [12]. Complementing this line of research, [17] presented a case study involving a patient with global aphasia associated with dysarthria, demonstrating the potential use of personalized models to support communication in individuals with severe speech impairment.

In addition to these specific applications, broader approaches aimed at developing more generalizable speech analysis models were also identified. [15] proposed a multitask architecture capable of integrating phonetic recognition, temporal alignment, and articulatory trajectory prediction into a single deep learning model using publicly available datasets comprising both healthy speakers and individuals with dysarthria. In contrast, [6] investigated different acoustic representations for detecting speech impairments associated with various neurological disorders, including stroke, providing a methodological reference for subsequent studies.

Although dysarthria represents the predominant clinical focus among the analyzed studies, the included literature addresses different clinical objectives, ranging from screening and classification to language assessment, speech recognition, and therapeutic support. Therefore, the reported results should be interpreted according to the specific objective and protocol of each study, rather than being considered directly comparable solely on the basis of the reported performance metrics. This heterogeneity is particularly relevant when interpreting differences in model performance, as variations in datasets, speech tasks, recording conditions, preprocessing procedures, and evaluation strategies may influence the reported outcomes.

B. Datasets

The analyzed studies showed substantial variability in the datasets employed, reflecting one of the main challenges currently faced by this research field: the limited availability of large, standardized public datasets for automatic speech analysis in post-stroke populations. As a result, most studies relied on proprietary datasets. Of the 12 studies included, eight were based exclusively on proprietary datasets [6], [7], [9]–[12], [16], [17], two combined proprietary and public datasets [8], [13], and only two relied exclusively on publicly available datasets widely used in the literature [14], [15].

This predominance of proprietary datasets indicates that many research groups still need to conduct dedicated data collection campaigns to obtain datasets suitable for training deep learning models. In contrast, studies relying on public datasets were primarily based on resources originally developed for speech recognition or articulatory analysis, such as TORGO, UASpeech, CommonPhone, HPRC, Nemours, and Speech Commands v2 [13]–[15]. Although these datasets are widely used in the literature, they differ in terms of language, speech tasks, number of participants, and clinical characteristics, making direct comparisons between the proposed models more challenging.

Considerable variability was also observed in the composition of the study populations. While some studies included relatively large cohorts, such as the DeepStroke framework, which was developed using data from 221 individuals evaluated in emergency settings [10], others relied on much smaller samples, including eight individuals with aphasia [8], a single patient with global aphasia and dysarthria [17], or even healthy volunteers simulating dysarthric speech to evaluate rehabilitation devices [13]. Between these extremes, most studies enrolled groups comprising individuals affected by stroke and healthy controls, typically involving a few dozen participants [7], [9], [11], [16]. Such variability complicates direct comparisons across studies and limits the generalizability of the proposed models.

Regarding language, English and Mandarin were the most frequently represented in the analyzed datasets. English predominated in the public datasets used for automatic speech recognition, including TORGO, UASpeech, CommonPhone, and Nemours [14], [15], whereas Mandarin was adopted in studies that collected proprietary datasets for the assessment

of post-stroke aphasia and dysarthria [7], [9], [16]. Studies involving Italian [17], Indonesian [13], and Lithuanian [6] were also identified. Despite this linguistic diversity, no studies using publicly available datasets composed of Brazilian Portuguese speakers were identified, highlighting an important gap for the development and validation of tools intended for the Brazilian clinical setting.

Speech tasks also varied considerably across the analyzed studies. Sustained vowels were primarily employed in studies focused on early stroke diagnosis or dysarthria classification, as they provide more controlled speech samples and facilitate the extraction of acoustic features [9], [16]. Other studies relied on syllables, isolated words, or picture-naming tasks for aphasia assessment [7], [8], [12], [17], whereas standardized sentences and voice commands were adopted in automatic speech recognition systems and rehabilitation devices [11], [13], [14]. Only a few studies incorporated spontaneous speech into their data collection protocols, typically in combination with structured tasks, as reported by [16] and in the multimodal DeepStroke framework [10]. Taken together, these findings indicate that there is still no consensus regarding speech acquisition protocols for AI-based applications in the post-stroke context. This lack of standardization hampers both direct comparisons across studies and the development of publicly available shared datasets for training and validating AI models.

C. Preprocessing and Speech Signal Representation

Initial preprocessing steps were primarily intended to reduce the effects of noise and standardize the audio signals. Procedures such as amplitude normalization, recording segmentation, and removal of silent segments were reported across different studies [7], [8], [17]. In applications focused on early stroke diagnosis, [16] adopted a more elaborate strategy by incorporating a dedicated Speech Enhancement module based on the SEWUNet architecture prior to the classification stage, demonstrating that improvements in signal quality led to performance gains of approximately 10% in the trained models. Similarly, [17] applied voice activity detection (VAD), normalization, and noise reduction using spectral gating, whereas [13] implemented embedded processing compatible with TinyML devices, prioritizing computationally efficient operations.

Another strategy adopted by some studies was data augmentation, mainly employed when the number of available samples was limited. Among the analyzed studies, this approach was explicitly reported by [16], who applied different augmentation techniques to expand the spontaneous speech dataset prior to neural network training. In contrast, several studies did not report the use of data augmentation, relying instead on cross-validation strategies or on architectures specifically designed for small patient datasets [8], [9], [17].

Following preprocessing, feature extraction and signal representation constituted the predominant intermediate step between the raw speech signal and the AI models. Mel-Frequency Cepstral Coefficients (MFCCs) were the most frequently used representation, being employed in studies addressing both aphasia and dysarthria [6]–[10], [16], [17]. Their widespread use can be attributed to their ability to capture perceptually relevant characteristics of speech while maintaining low computational cost, making them particularly suitable for clinical applications and embedded devices.

More recent studies, however, have increasingly adopted two-dimensional feature representations capable of simultaneously preserving temporal and spectral information. Mel-spectrograms were used as input for convolutional and hybrid models in several studies [14], [16], whereas conventional spectrograms were also employed for comparison with MFCC-based representations [9]. These findings suggest that models based on MFCCs outperformed those trained directly on spectrograms, indicating that the choice of speech representation still depends on the characteristics of the task and the available dataset.

Beyond the representations traditionally employed in speech analysis, several studies explored more specialized descriptors. [6] compared three different feature representations, cepstrum, auditory spectrogram, and Wavelet Time Scattering (WTS), demonstrating that these descriptors capture complementary aspects of the speech signal. In turn, [11] employed Phonetic Posteriorgrams (PPGs) to encode phonetic information while comparing assistive systems based on automatic speech recognition. Meanwhile, [15] integrated articulatory information acquired from Electromagnetic Articulography (EMA) sensors with conventional acoustic features, thereby bringing computational models closer to the physiological mechanisms underlying speech production.

Overall, the analyzed studies reveal a gradual evolution in speech signal representation strategies. Earlier studies relied primarily on manually engineered acoustic features, such as MFCCs and cepstral representations [6], [7], whereas more recent work has increasingly adopted richer feature representations, including Mel-spectrograms, articulatory information, and models capable of learning discriminative patterns directly from these representations [14]–[16]. This evolution suggests that the performance of deep learning models depends not only on the model architecture but also on the quality of preprocessing and the representativeness of the features extracted from the speech signal.

D. Artificial Intelligence Architectures and Strategies

The AI architectures employed varied considerably across the analyzed studies. Rather than following a uniform progression toward increasingly complex models, the literature includes convolutional neural networks, hybrid CNN-LSTM architectures, Transformer-based models, transfer learning strategies, multimodal approaches, TinyML implementations, and personalized models. This heterogeneity reflects the different clinical objectives and speech analysis tasks addressed by the included studies.

CNNs were used in several of the included studies, covering applications such as automated aphasia assessment, dysarthric speech recognition, rehabilitation, and early stroke diagnosis [7], [9], [10], [13], [14], [16]. However, their use should not be

Table I
SUMMARY OF DATASETS, POPULATION, SPEECH TASKS, APPLICATION TYPES, AND RECORDING CONDITIONS.

Ref.	Dataset Type	Population & Stroke Stage	Language / Speech Task Category	Application Type	Recording Device & Condition
[6]	Own	15 (7 CNSD [1 acute stroke] + 8 HC); N=339	English, Lithuanian / Poem reading	Screening / Clinical Support	Android (NITS app); Presential (Supervised)
[7]	Own	46 (12 PWA [stroke 8–1096d] + 34 HC); N=402	Mandarin / 6 vowels, 20 words (CRRCAE)	Severity / In-Home Therapy	Lenovo Recorder Pen (48 kHz); Soundproof/Ward
[8]	Own/Public	8 PWA (chronic, 14–109m) + 140 HC; N=1,760	British English / Picture naming (220 words)	Anomia Support / In-Home Verification	Galaxy Tab S3 + Headset (16 kHz); Presential
[9]	Own	116 (71 PSD + 45 HC); N=7,318	Mandarin / 81 single syllables (vowels/digits)	Screening / Clinical Support	Prof. Recording Setup; Hospital (Supervised)
[10]	Own	221 (159 acute stroke/TIA + 62 ER ctrl); N=221	English / Sentences + Picture description	Emergency Triage Support	iPhone X (video+audio); Unconstrained ER
[11]	Own	14 (10 CVA + 4 CP; 4 mild, 9 mod, 1 sev)	Mandarin / Sentence reading (320 TMHINT)	SE / Communication Assist	Audio Setup; Soundproof (Supervised)
[12]	Own	46 (12 PWA + 34 HC)	Mandarin / Naming & Repetition (CRRCAE)	Screening / Aphasia Support	Lenovo Recorder Pen (48 kHz); Presential
[13]	Own/Public	2 sim. dysarthria + TORGO/SpeechCmds; N=106k	Bahasa Indonesia, English / Voice commands + Vowels	Assistive Robotic Control	INMP441 Mic + ESP-32; Lab (24–62 dB noise)
[14]	Public	11 dysarthria (Nemours) + 4 HC; 15 dys + 13 HC; N=3,700	English / Nonsense sentences + Words	Screening / Pathological ASR	Studio Mics; Soundproof/Controlled
[15]	Public	8 HC (HPRC) + CommonPhone + TORGO + UA (28 spk)	English / Sentences (IEEE) + Words	Multitask Articulation & Intelligibility	EMA + Array Mics; Lab + Crowdsourced
[16]	Own	64 (32 stroke + 32 HC); Test: 34 (24 stroke + 10 HC); N=1.9k	Mandarin / 6 SVs (/a,o,e,i,u,ü/) + 1 SS sentence	Early Diagnosis / Real-time Triage	USB Mic + Laptop C/S + WeChat; Quiet Ward
[17]	Own	1 female (chronic, 15y post-stroke; sev. dysarthria); N=2,056	Italian / 13 words (5 prompt types)	Assistive VIVOCA / AAC Support	Headset Mic + CapisciAMe App (16 kHz); Quiet Room

Abbreviations: HC = Healthy Controls; PWA = People with Aphasia; PSD = Post-Stroke Dysarthria; CNSD = Central Nervous System Disorders; TIA = Transient Ischemic Attack; ER = Emergency Room; CVA = Cerebrovascular Accident; CP = Cerebral Palsy; SVs = Sustained Vowels; SS = Spontaneous Speech; SE = Speech Enhancement; ASR = Automatic Speech Recognition; AAC = Augmentative and Alternative Communication; NITS = Neural Impairment Test Suite; CRRCAE = Chinese Rehabilitation Research Center Aphasia Examination; TMHINT = Taiwan Mandarin Hearing-in-Noise Test; N = Total audio samples; d = days; m = months; y = years; spk = speakers.

Table II
SUMMARY OF AI MODELS AND VALIDATION PROTOCOLS.

Ref.	Features	Split / Segmentation Unit	Technique	Primary Outcome
[6]	Pitch, MFCC, GTCC, Auditory Spectrogram, WST	Segment-level (20–40 ms frames)	BiLSTM; WST + SVM	Acc.: 96.3% (WST+SVM), 94.5% (BiLSTM)
[7]	HTD (TFD RGB Images); 51 MFCC/Formants	Sample-level (Word/Vowel segments)	ResNet-34 (Transfer Learning) + LinReg	Acc.: 99.23% (words); Articulation r_s=0.71, Fluency r_s=0.60
[8]	26 Acoustic (12 MFCC, Energy, Deltas) $\rightarrow$ PPGs	Item-level (Single-word matrices)	7-layer BiGRU DNN + DTW Matching	10-fold CV: 89.5% acc.; Adapted: 91.3% vs Google STT (88.2%)
[9]	30 MFCC coefficients	Speaker-level split / Syllables	1DCNN + Double-LSTM + Soft-majority	Syllable Acc.: 75.5%; Speaker Acc.: 84.8%
[10]	Video frame diff. + Log-Mel Spectrogram	5-fold CV / Time-cutoff Holdout	Dual-branch ResNet (34+18) + Adv. Disc.	Case Acc.: 71.40%, Sens.: 81.13%, AUC: 0.7163 (vs 64.03% triage)
[11]	LPS (SE), WORLD Envelopes (VC), MFCC/PPG	Speaker-dependent split / Sentences	DNN-SE vs Gated-CNN VC vs ASR-TTS	ASR-TTS: 93.8% ASR acc., 95.5% Intelligibility, 4.56 MOS
[12]	Hyperbolic T-Distribution (HTD) RGB Images	5-fold CV / Word-level	CNN (ResNet-50 + HTD) vs ASR APIs	CNN Acc.: 99.64% (healthy), 67.78% (aphasic) vs Azure (88.09%)
[13]	STFT Log-Spectrogram (99 $\times$ 43 pixels)	80/10/10 Split / 1s audio frames	Lightweight 2DCNN + INT8 (TinyML)	Val. Acc.: 94.0%; Real-time ESP-32: 95% (24dB), 50% (62dB noise)
[14]	Mel-Spectrograms + Augmentation + Ensemble	80/20 Holdout Split / Utterances	Tran-DSR (Ensemble + ViT Encoder)	Acc.: 99.18% (Nemours/ASVSpoof), 98.98% (UASpeech)
[15]	Raw Audio (16kHz) $\rightarrow$ Wav2Vec2 (49Hz) + 9 TVs	LOSO / 70/15/15 Split / Utterances	f-APTAI (Wav2Vec2 + Cross-Attn + BiLSTM)	AAI: Mean PCC=0.71, RMSE=0.68mm; Intelligibility R=-0.96
[16]	88 Handcrafted (SVs); Mel-Spectrograms (SS)	5-Fold CV (Speaker-split) / Utterances	SEWUNet + AdaBoost (SVs) + CNN-Trans. (SS)	SVs Acc.: 92.31%; SS Acc.: 96.00%; Clinical 2-Stage Acc.: 100%
[17]	39 MFCCs (with Deltas) as 2D Representations	Session-based temporal / Word-level	Speaker-dependent Compact CNN	Tailored ASR Word Acc.: 72.65% vs Human Listeners: 56.75% ($p < 0.001$)

Abbreviations: MFCC = Mel-Frequency Cepstral Coefficients; GTCC = Gammatone Cepstral Coefficients; WST = Wavelet Scattering Transform; HTD = Hyperbolic T-distribution; TFD = Time-Frequency Distribution; PPG = Phonetic Posteriorgrams; LPS = Log-Power Spectrum; STFT = Short-Time Fourier Transform; TVs = Tract Variables; SVs = Sustained Vowels; SS = Spontaneous Speech; BiLSTM = Bidirectional LSTM; BiGRU = Bidirectional GRU; ViT = Vision Transformer; LOSO = Leave-One-Speaker-Out; CV = Cross-Validation; AAI = Acoustic-to-Articulatory Inversion; PCC = Pearson Correlation; RMSE = Root Mean Square Error; MOS = Mean Opinion Score; Acc = Accuracy; Sens = Sensitivity; AUC = Area Under Curve; r_s = Spearman correlation; R = Pearson correlation.

interpreted as evidence of a single dominant architecture across the literature, since the studies employed different datasets, speech tasks, input representations, and evaluation protocols.

Despite the frequent use of CNN-based approaches, several studies adopted alternative or hybrid architectures. In [9], a 1D CNN was combined with a Double-LSTM to capture local features and temporal dependencies within the speech signal. A similar strategy was adopted by [14], who developed the Tran-DSR model by integrating ensemble CNNs with a Transformer Encoder. These approaches illustrate how different architectures have been combined according to the characteristics of the speech analysis task rather than following a single architectural strategy.

Several studies also compared different AI strategies using the same dataset. [12] compared their CNN- and LDA-based models with the Microsoft Azure and Google Speech Recognition platforms, reporting superior performance for models specifically trained on pathological speech. Likewise, [11] evaluated three approaches aimed at improving speech intelligibility, Speech Enhancement, Voice Conversion, and Automatic Speech Recognition combined with Text-to-Speech, with the latter achieving the best overall performance. These comparisons indicate that performance depends not only on the model architecture but also on the specific task, speech representation, and system configuration.

Transfer learning was another strategy identified in the reviewed studies. In [16], a pretrained ResNet was employed for spontaneous speech recognition, whereas a separate SEWUNet-based model was first used to enhance the speech signal. The use of pretrained architectures may be particularly relevant in clinical speech applications, where datasets are often relatively small and difficult to collect. However, the potential advantages of transfer learning should be interpreted cautiously, as pretrained models may be affected by differences between their original training data and pathological speech datasets.

Other studies explored alternative architectural designs. DeepStroke integrated information from both speech and facial expressions into a multimodal model designed for triaging patients with suspected stroke [10]. In contrast, [13] explored the use of TinyML to implement a portable device for dysarthria rehabilitation, whereas [15] proposed a multitask architecture capable of performing phonetic recognition, temporal alignment, and articulatory trajectory prediction within a single model. These approaches demonstrate that architectural choices were also influenced by practical requirements, such as multimodal analysis, computational constraints, and the need to perform multiple speech-related tasks.

A different comparison was presented by [6], who evaluated a BiLSTM architecture alongside an SVM classifier based on the Wavelet Scattering Transform (WST) using different acoustic representations, including MFCCs, GTCCs, and auditory spectrograms. The best performance was achieved by the WST+SVM combination, suggesting that speech representation and feature extraction can influence model performance as much as, or potentially more than, the choice of learning architecture in some applications.

Taken together, the reviewed studies do not indicate a single dominant AI architecture across all applications. CNNs, hybrid models, Transformers, transfer learning, multimodal architectures, TinyML, and personalized models were employed according to the objectives and constraints of each study. Most of the included approaches remain based on CNNs, hybrid architectures, or task-specific models, while pretrained speech models, including self-supervised and large speech models, represent a more recent direction with potential relevance to clinical datasets with limited sample sizes. Nevertheless, their applicability to pathological speech should not be assumed from performance on other speech domains. Future studies should therefore evaluate these models using independent clinical datasets and reproducible protocols, with particular attention to external validation, data variability, and generalizability across speakers and recording conditions.

E. Evaluation Metrics

The evaluation metrics adopted across the analyzed studies were, for the most part, similar. Accuracy was reported in nearly all studies [6]–[10], [12]–[17], typically accompanied by precision, recall (or sensitivity), and F1-score [9], [12], [14]–[16]. Together, these metrics constitute the primary evaluation framework for pathological speech classification tasks in these studies.

ROC curves and the corresponding area under the curve (AUC) were also frequently reported, particularly in studies focused on diagnosis and patient classification [8]–[10], [14], [16]. In several studies, specificity was reported alongside sensitivity, providing a more balanced assessment of the discriminative ability of the models, an aspect that is particularly important in clinical applications [10], [16].

Although traditional classification metrics predominated, several studies adopted evaluation metrics tailored to the specific objectives of their applications. In the development of systems aimed at improving speech intelligibility, [11] employed intelligibility measures, Mean Opinion Score (MOS), and the recognition rate achieved by Google Speech Recognition, whereas [16] evaluated the speech enhancement module using objective speech quality metrics, including PESQ, STOI, SegSNR, CSIG, CBAK, and COVL. In contrast, [15] adopted task-specific metrics for phonetic recognition and articulatory trajectory prediction, including Phoneme Error Rate (PER), the Pearson Correlation Coefficient (PCC), Root Mean Square Error (RMSE), and segment overlap rate.

It is worth noting that high accuracy values should be interpreted with caution when obtained from small datasets, as limited sample sizes may increase the risk of bias and limit the generalization of the models to new speakers and clinical populations. Moreover, the results are not directly comparable across studies only based on the reported metrics, since the studies differ in clinical objectives, units of analysis, datasets, and validation protocols. For clinical use, evaluation with small samples alone is not sufficient; further testing with larger and independent datasets is needed to assess the robustness and generalizability of the models.

Overall, the reviewed literature still relies primarily on traditional classification metrics, particularly accuracy, precision, recall, and F1-score. However, as applications move closer to real-world clinical settings, evaluation has increasingly incorporated task-specific measures of speech intelligibility, speech quality, and agreement with expert assessments. Therefore, the reported performance metrics should be interpreted according to the specific objective and methodological protocol of each study rather than used for direct comparison of model performance.

F. Limitations of the Literature

Despite the advances achieved in recent years, the reviewed studies reveal recurring limitations that continue to hinder the consolidation of Artificial Intelligence systems for automatic post-stroke speech analysis. The most significant limitation concerns the availability of clinical data. Nearly all studies relied on relatively small numbers of participants, often comprising only a few dozen patients or, in extreme cases, a single patient [6]–[14], [16], [17]. However, a small number of participants does not necessarily correspond to a small number of samples, since several studies divided speech into words, syllables, segments, or analysis windows. When samples from the same participant are distributed across training and test sets, the reported performance may be influenced by speaker-specific characteristics and may not reflect generalization to new patients.

The limited number of participants also affects the generalizability of the proposed models. Several studies emphasized that their findings still require validation in larger populations, encompassing different levels of impairment, multiple data collection centers, and greater linguistic diversity [7], [9]–[11], [16], [17]. In some cases, the authors acknowledged that model performance was evaluated only in patients with specific levels of disease severity, under restricted speech tasks, or within proof-of-concept studies, limiting the applicability of the findings to broader clinical settings [8], [13], [17].

Another important limitation is the absence of external validation across the reviewed studies. Although cross-validation or internal train-test procedures were employed, none of the included studies evaluated model performance on an independent external dataset. Moreover, when partitioning is performed at the sample or item level rather than at the participant level, samples from the same individual may occur in both training and test sets, introducing potential data leakage and leading to optimistic performance estimates. Therefore, validation on independent patients and datasets is necessary

to assess model robustness and generalizability before clinical implementation.

Substantial methodological heterogeneity was also observed across the analyzed studies. These differences ranged from language and participant profiles to speech tasks employed, including sustained vowels, syllables, isolated words, picture naming, sentence reading, voice commands, and spontaneous speech [7]–[9], [11]–[13], [16], [17]. In addition, different preprocessing strategies, acoustic representations, AI architectures, and evaluation metrics were used, making direct comparisons between published results challenging.

The scarcity of public datasets specifically designed for post-stroke speech analysis represents another recurring limitation. As a consequence, many research groups have developed their own datasets [6]–[12], [16], limiting the reproducibility of experiments and hindering comparisons between models proposed by different research groups. Even studies that relied on publicly available datasets reported limitations related to the number of participants, linguistic diversity, or the availability of detailed clinical annotations [14], [15].

IV. Conclusion

This review shows that Artificial Intelligence has been applied to several tasks involving post-stroke speech, including dysarthria and aphasia assessment, screening, diagnosis, and rehabilitation support. However, the evidence remains limited by small numbers of participants, heterogeneous speech tasks and acquisition protocols, single-center datasets, and the absence of external validation. These limitations make reported performance difficult to compare and do not yet support conclusions about model generalizability to new patients.

The limited availability of datasets in languages other than English and Mandarin also represents an important gap. No publicly available dataset specifically designed for post-stroke speech in Brazilian Portuguese was identified in this review.

For future work, we propose prioritizing larger, multicenter datasets, participant-level data partitioning, external validation, standardized acquisition protocols, and publicly accessible datasets. A rigorous protocol should be established to collect and curate a speech dataset, incorporating evaluation in independent participants, that is, individuals not included in the training and validation sets. Such dataset characteristics can enhance quality and reliability, and make AI model performance metrics more robust regarding real-world scenarios, effectively contributing to the improvement of these systems. Furthermore, creating a Brazilian Portuguese dataset, containing samples from healthy participants and post-stroke individuals performing various speech tasks is important to support the development of a tool adapted to the Brazilian context.

Acknowledgments

The authors acknowledge CNPq (National Council for Scientific and Technological Development) for funding projects No. 303584/2023-8 and FAPES (Fundação de Amparo à Pesquisa e Inovação do Espírito Santo) for funding this work through projects No. 756/2024 (P: 2024-0TXNS), No. 380/2026 (P: 2026-ZKDB2), and the master's scholarship; the Fundação iNOVA Capixaba for fostering research, and the Federal Institute of Espírito Santo (IFES), Vitória Campus, for its institutional support in the development of this work.

References

[1] World Health Organization, "Stroke," Fact sheet, Dec. 19, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/stroke. Accessed: Jun. 26, 2026.

[2] Ministério da Saúde, "Acidente Vascular Cerebral (AVC)," Governo Federal, Brasil. [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/avc. Accessed: Jun. 26, 2026.

[3] R. Chiaramonte and M. Pavone, "Speech rehabilitation in dysarthria after stroke: A systematic review of current evidence," *European Journal of Physical and Rehabilitation Medicine*, vol. 56, no. 3, pp. 354–364, 2020.

[4] J. Vitti and A. E. Hillis, "Optimizing recovery in aphasia: Current state and future directions," *Current Neurology and Neuroscience Reports*, vol. 21, no. 12, 2021.

[5] G. P. Georgiou, "AI-assisted assessment and treatment of aphasia: A review," *Frontiers in Public Health*, vol. 12, Art. no. 1401240, 2024.

[6] A. Lauraitis, R. Maskeliunas, R. Damaševičius, and T. Krilavičius, "Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features," *IEEE Access*, vol. 8, pp. 96162–96172, 2020.

[7] S. S. Mahmoud, A. Kumar, Y. Tang, Y. Li, X. Gu, J. Fu, and Q. Fang, "An efficient deep learning based method for speech assessment of Mandarin-speaking aphasic patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 887–900, 2021.

[8] D. S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan, I. Shaw, W. Latham, A. P. Leff, and J. Crinion, "NUVA: A naming utterance verifier for aphasia treatment," *Computer Speech & Language*, vol. 69, p. 101221, 2021.

[9] W. Ye, Z. Jiang, Q. Li, Y. Liu, and Z. Mou, "A hybrid model for pathological voice recognition of post-stroke dysarthria by using 1DCNN and double-LSTM networks," *Applied Acoustics*, vol. 197, p. 108934, 2022.

[10] T. Cai, H. Ni, M. Yu, X. Huang, K. Wong, J. Volpi, J. Z. Wang, and S. T. C. Wong, "DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning," *Medical Image Analysis*, vol. 80, p. 102522, 2022.

[11] W.-Z. Zheng, J.-Y. Han, H.-L. Cheng, W.-C. Chu, K.-C. Chen, and Y.-H. Lai, "Comparing the performance of classic voice-driven assistive systems for dysarthric speech," *Biomedical Signal Processing and Control*, vol. 81, p. 104447, 2023.

[12] S. S. Mahmoud, R. F. Pallaud, A. Kumar, S. Faisal, Y. Wang, and Q. Fang, "A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries," *Sensors*, vol. 23, no. 2, p. 857, 2023.

[13] B. Riyanta, H. A. Irianta, and B. P. Kamiel, "Development of speech command control based TinyML system for post-stroke dysarthria therapy device," *Journal of Robotics and Control*, vol. 4, no. 4, pp. 466–478, 2023.

[14] R. Mahum, A. M. El-Sherbeeny, K. Alkhaledi, and H. Hassan, "TranDSR: A hybrid model for dysarthric speech recognition using transformer encoder and ensemble learning," *Applied Acoustics*, vol. 222, p. 110019, 2024.

[15] T. Weise, K. C. Demir, P. A. Pérez-Toro, et al., "Towards end-to-end speech articulation and spoken language analysis using deep learning," *Human-Centric Intelligent Systems*, vol. 5, pp. 103–122, Mar. 2025.

[16] J. Zhang, Y. Qiu, Y. Liu, Y. Xiao, J. Yang, X. Yang, M. Ma, and A. Song, "Early stroke diagnosis and evaluation based on pathological voice classification using speech enhancement," *Computers in Biology and Medicine*, vol. 196, p. 110940, 2025.

[17] D. Mulfari, D. Cardile, S. Campana, L. Voci, C. M. Vicario, R. S. Calabrò, F. Corallo, S. Mulfari, and F. Tomaiuolo, "Case report: Tailored automatic speech recognition in global aphasia with dysarthria - a single case proof of concept," *Frontiers in Rehabilitation Sciences*, vol. 7, p. 1813312, 2026.

Análise de modelos de aprendizagem profunda na detecção de nódulos pulmonares

Rebeca de Jesus Brandão[1], Marcelo Costa Oliveira[1], Thiago Damasceno Cordeiro[1], Lucas de Lins Lima[2]
[1]*Instituto de Computação, Universidade Federal de Alagoas*, Maceió, Brasil
{rjb, oliveiramc}@ic.ufal.br, thiago@ic.ufal.br
[2]*Departamento de Computação, UNIP*, Ribeirão Preto, Brasil
lucaslins11@gmail.com

***Resumo*—O câncer de pulmão é a principal causa de mortes por câncer no Brasil, especialmente no Nordeste. Atrasos no diagnóstico reduzem as taxas de sobrevivência e sobrecarregam os sistemas de saúde. A análise manual de imagens de Tomografia Computadorizada (TC) para a identificação de nódulos em estágios iniciais é uma tarefa complexa e suscetível a erros. Diante disso, este trabalho tem como objetivo principal realizar uma análise comparativa rigorosa entre duas arquiteturas de aprendizado profundo para a detecção precoce e segmentação de nódulos pulmonares: o modelo baseado em atenção DETR (Detection Transformer) e a arquitetura baseada em blocos convolucionais modernos ConvNeXt V2, acoplada ao framework multitarefa Mask R-CNN. A metodologia consistiu no uso de imagens de TC do banco de dados público LUNA16. Os exames foram pré-processados e filtrados para isolar nódulos com dimensão ($\leq$ 15 mm), divididos em subconjuntos categóricos. Ambos os modelos foram submetidos a 400 épocas de treinamento. A abordagem proposta com o ConvNeXt V2 foi configurada para operar simultaneamente a localização espacial geométrica (bounding boxes). Os resultados quantitativos demonstraram o bom desempenho do ConvNeXt V2 em relação ao baseline do DETR. Na tarefa de detecção, o ConvNeXt V2 obteve um mAP@0.50 de 0,7535 e um F1-Score de 0,7782, superando significativamente o DETR (0,5348 e 0,7030, respectivamente). Para os nódulos clinicamente desafiadores entre 10 e 15 mm de diâmetro, a avaliação por meio da curva FROC revelou um escore Λ de 0,6454, indicando um desempenho diagnóstico global, com baixa taxa de falsos positivos. Na análise de custo computacional, o ConvNeXt V2 reduziu o tempo de treinamento de 10 minutos (DETR) para apenas 2 minutos e 40 segundos por época, corroborando a literatura recente que atesta a eficiência dessas arquiteturas convolucionais modernas na mitigação do colapso de características, sem o peso computacional dos Transformers puros. Conclui-se que a integração do ConvNeXt V2 com a Mask R-CNN supera amplamente as limitações dos métodos de detecção tradicionais, garantindo não apenas a identificação precisa do tumor, mas também a sua fidelidade morfológica. Conclui-se que o modelo é promissor no uso de uma ferramenta de Auxílio ao Diagnóstico (CAD) viável, capaz de otimizar o fluxo de trabalho radiológico e elevar a eficácia clínica na identificação morfológica de tumores.**

***Palavras-chave*—Detecção de Objetos, VIsion Transformers, Nódulos Pulmonares, ConVneXt V2**

I. Introdução

O câncer de pulmão se destaca entre os tipos de câncer mais preocupantes no mundo, tanto pela alta prevalência quanto pela letalidade. Entre 2022 e 2025, foram registrados 2,5 milhões de novos casos, o que o torna o primeiro câncer mais comum globalmente, e 1,8 milhão de mortes, consolidando-o como a principal causa de morte por câncer (44). Esse panorama alarmante representa aproximadamente 12,4% dos diagnósticos de câncer e 18,7% das mortes por câncer em todo o mundo. No Brasil, o cenário reflete essa gravidade: conforme a estimativa oficial do Instituto Nacional de Câncer para o triênio 2025-2027, são esperados cerca de 36.160 novos casos anuais (15). O diagnóstico precoce do câncer de pulmão é essencial para aumentar as chances de sobrevivência do paciente, com uma taxa de sobrevivência de 5 anos que chega a 70% quando o câncer é diagnosticado em estágio inicial (16). Portanto, programas voltados à detecção precoce de nódulos pulmonares devem ser implementados para aumentar as chances de sobrevivência dos indivíduos afetados pelo câncer de pulmão.

A TC é a principal ferramenta utilizada pelos radiologistas para detectar nódulos pulmonares, pois fornece imagens em 3D de alta resolução, com contraste, que refletem as diferenças de intensidade, textura e forma dos tumores. No entanto, o diagnóstico de nódulos pulmonares, especialmente em estágios iniciais, é um desafio para radiologistas. Além do pequeno tamanho ($<$15 mm), os nódulos podem apresentar baixo contraste em relação ao tecido pulmonar e estar anexos a outras estruturas pulmonares complexas, o que aumenta as dificuldades (38).

Os sistemas CAD são ferramentas importantes que podem fornecer suporte à tomada de decisão dos radiologistas, atuando como uma segunda opinião (1). As ferramentas CAD podem automatizar o processo de diagnóstico, reduzindo o tempo e o esforço necessários à análise e melhorando a confiabilidade e a repetibilidade da tarefa (14). Os sistemas CAD envolvem tanto a localização de lesões em imagens médicas (CADe) quanto a classificação de lesões como malignas ou benignas (CADx), o que traz diversos desafios para o campo (22).

Atualmente, as técnicas de Aprendizado Profundo (AP) são consideradas o estado da arte em aplicações de CAD para nódulos pulmonares, e diversos estudos recentes têm demonstrado o potencial e a relação custo-benefício dessas abordagens em centros de saúde (41). Entre as arquiteturas, as redes neurais convolucionais (CNN) se destacam e têm apresentado resultados promissores na área da informática médica. Contudo, o grande desafio da literatura consiste em

otimizar simultaneamente duas métricas essenciais: a Sensibilidade (que mede a redução de Falsos Negativos - FN) e a Especificidade ou Precisão (relacionada à minimização de Falsos Positivos - FP). Apesar dos avanços algorítmicos, o desafio de manter alta sensibilidade sem inflar o número de FP ainda persiste, como corroborado por avaliações recentes no campo (14; 19).

A alta taxa de FP compromete a eficiência clínica ao exigir a revisão desnecessária de exames, constituindo um obstáculo severo à implementação de sistemas CAD em ambientes hospitalares. Uma revisão sistemática (11) apontou que o auxílio de IA no rastreio de câncer de pulmão aumenta a sensibilidade (de +5% a +20%) em relação à leitura não assistida, mas tipicamente leva a uma redução na especificidade (de -7% a -3%), o que se traduz em vigilância excessiva. Trabalhos específicos, baseados em arquiteturas modernas, ilustram com clareza essa dificuldade de equilíbrio. Por exemplo, a pesquisa de Zhao et al. (2024) (42), que utilizou a arquitetura Mask R-CNN acoplada ao ConvNeXt V2, alcançou alta sensibilidade e especificidade, porém, em um contexto macro de detecção de metástases ósseas. Por outro lado, o estudo de Mafra (2024) (24) demonstrou que, na detecção direta de micronódulos pulmonares, a CNN YOLOv8 sobressaía-se em precisão (minimizando FP), enquanto modelos baseados em atenção destacavam-se em sensibilidade (minimizando FN). Esses resultados indicam que a superação do trade-off entre FN e FP permanece uma área crucial de pesquisa.

O Transformer é uma arquitetura de codificador-decodificador que revolucionou o campo do Processamento de Linguagem Natural (do inglês, NLP) e inspirou o Vision Transformer (ViT) para tarefas de visão computacional, podendo atuar em conjunto ou em substituição às tradicionais CNNs. Na área médica, propostas baseadas em ViT demonstraram capacidade de manter alta sensibilidade com um número reduzido de parâmetros (4). Contudo, os modelos puramente baseados em atenção apresentam limitações importantes. Diferentemente das convoluções, eles carecem de vieses indutivos locais, o que exige volumes massivos de dados para evitar o sobreajuste (*overfitting*) em bases médicas mais restritas. Além disso, a sua complexidade computacional cresce de forma quadrática com o tamanho da imagem, o que eleva substancialmente o custo de processamento e a demanda de memória, dificultando a sua adoção em cenários clínicos rotineiros (8).

Para contornar as limitações computacionais dos Transformers sem abrir mão do seu alto desempenho em contextos globais, pesquisadores propuseram a arquitetura ConvNeXt, que moderniza as CNNs ao incorporar princípios de *design* dos Transformers hierárquicos (22). A sua evolução direta, o ConvNeXt V2 (39), introduziu aprimoramentos vitais, como o uso de autoencoders mascarados e a Normalização de Resposta Global (GRN), mitigando falhas na extração de características em redes muito profundas. Essa abordagem demonstrou que redes puramente convolucionais podem igualar ou superar os ViTs em termos de precisão, mantendo a eficiência computacional inerente às CNNs e despontando como uma solução robusta para contornar a alta taxa de FP em imagens radiológicas complexas.

Apesar da eficácia isolada do ConvNeXt V2 e de arquiteturas baseadas em atenção global pura, como o DETR (Detection Transformer), há uma notável lacuna na literatura quanto à comparação direta entre esses dois paradigmas na tarefa específica de detecção de nódulos pulmonares precoces (com diâmetro inferior a 15 mm). A maioria dos estudos foca em tumores de grandes dimensões e já consolidados, ou falha em documentar simultaneamente o comportamento de FP e FN em alvos morfológicos tão reduzidos (24). Para lidar com esse problema de maneira inovadora, este trabalho aplica e compara diretamente os modelos ConvNeXt V2 e DETR em um subconjunto de imagens estritamente filtrado para micronódulos, avaliando qual arquitetura oferece o equilíbrio mais favorável entre a identificação precoce da lesão e a redução de falsos alarmes clínicos.

II. Metodologia

A metodologia desenvolvida neste trabalho baseia-se em um *pipeline* estruturado de processamento e avaliação de arquiteturas de aprendizado profundo, conduzido conforme o fluxograma ilustrado na Figura 1. O fluxo metodológico segue os seguintes passos essenciais: inicia-se pela extração da base de dados original LUNA16 (LUng Nodule Analysis 2016) (A), avançando para o pré-processamento das imagens em 3D e a delimitação das caixas (B), processos detalhados na Seção 3.1. Na sequência, o *pipeline* estrutura as imagens em 2D e as anotações prontas para o algoritmo (C), abrangendo o refinamento e a divisão do conjunto de dados descritos na Seção 3.2. O processo segue para o bloco de detecção de objetos (D), que inclui a justificativa da escolha das arquiteturas (Seção 3.3) e culmina na configuração e no treinamento dos modelos DETR e ConvNeXt V2 (Seção 3.4). Por fim, a consolidação das predições e a análise comparativa das métricas (E) compõem o escopo de discussão a ser apresentado no Capítulo 4.

A base de dados escolhida para este estudo foi a LUNA16, uma base pública que reúne TCs de pacientes com nódulos pulmonares (33) e é uma extensão de outra base pública, a LIDC-IDRI (Lung Image Database Consortium) (37), com melhorias voltadas à detecção de nódulos. Para o desenvolvimento deste trabalho, utilizou-se o conjunto de dados já processado e organizado nos trabalhos de (24; 31). O objetivo de utilizar este conjunto de dados bidimensional (2D) pré-processado foi contornar o elevado custo computacional do processamento de volumes 3D e superar as restrições das arquiteturas de detecção, como o ConvNeXt V2 e o YOLOv8, que operam intrinsecamente com dados 2D. Visando conferir maior robustez à metodologia para futuras publicações e garantir a reprodutibilidade da pesquisa, descreve-se a seguir o *pipeline* detalhado de organização e pré-processamento estruturado pelos autores originais da base, cujas etapas viabilizaram o treinamento dos modelos propostos:

1) Seleção e Extração de Cortes 2D: Os exames de TC consistem, por padrão, em uma sequência de cortes axiais bidimensionais. Portanto, as imagens originais,

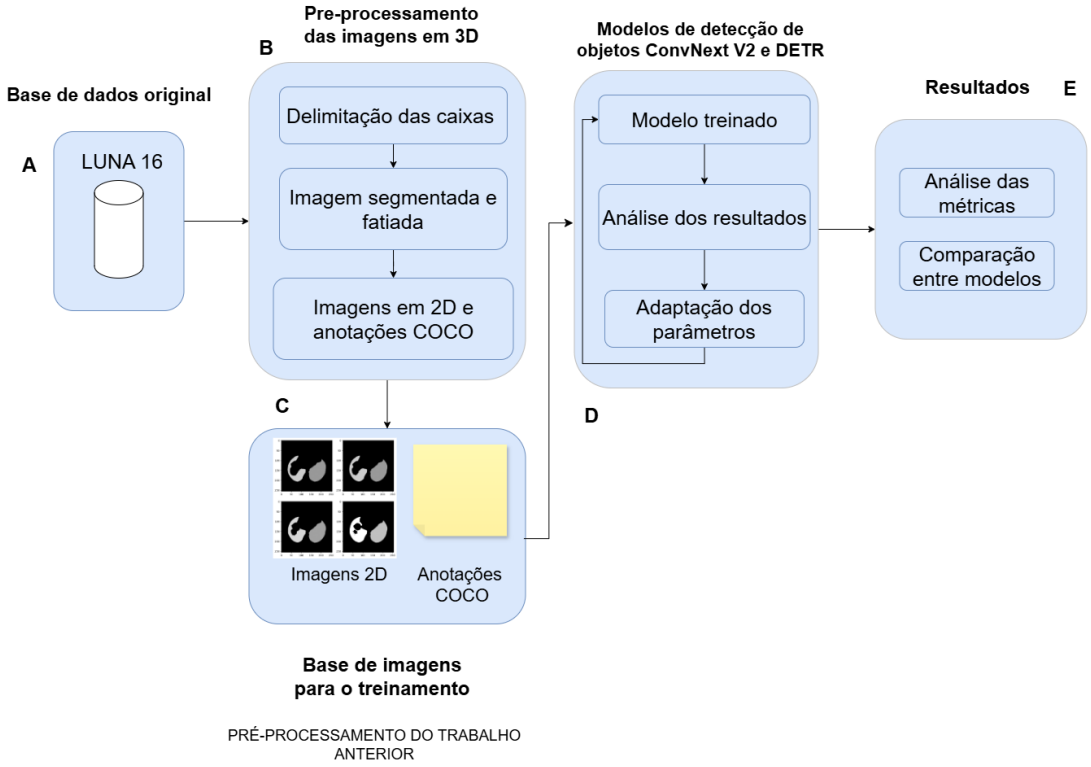


Figura 1. Diagrama ilustrativo da metodologia aplicada neste trabalho: (A) Base LUNA 16 para obtenção de imagens de TC de pacientes com nódulos pulmonares; (B) Etapa de pré-processamento das imagens em 3D; (C) Organização dos conjuntos de dados oriundos da etapa anterior; (D) Treinamento dos modelos; e (E) Análise dos resultados. Fonte: Autor(2026)

dispostas no padrão médico DICOM, e as anotações dos radiologistas (estruturadas em arquivos CSV contendo as coordenadas centrais X, Y, Z e o raio do nódulo) foram submetidas a um algoritmo de filtragem espacial. Foram incluídos no conjunto de dados apenas os cortes que continham o nódulo. Para garantir a riqueza geométrica e alimentar a rede neural com múltiplas perspectivas da lesão, extraíram-se múltiplos cortes do mesmo nódulo, sem se limitar à fatia central.

2) Pré-processamento (Atenuação de Ruídos): Aplicou-se um ajuste de brilho e contraste nas imagens geradas, configurando os valores de janela/nível (*Windowing*) para a escala Hounsfield de $-400/1200$. Este ajuste teve como objetivo eliminar ruídos e artefatos, atenuar sinais irrelevantes provenientes de outras estruturas torácicas e realçar exclusivamente a região do parênquima pulmonar.
3) Formatação das Anotações: As anotações espaciais dos nódulos foram convertidas para formatos padronizados da literatura de visão computacional, garantindo a compatibilidade de ingestão de dados para diferentes arquiteturas. Para este experimento foi adotado o padrão de formato COCO, mas no trabalho comparativo anterior também encontramos a padronização em YOLO.

   - Formato COCO (.json): É o formato adotado nativamente tanto pelo modelo DETR quanto pela arquitetura ConvNeXt V2 (acoplada à Mask R-CNN). Este formato estrutura os dados em um único arquivo, armazenando as coordenadas absolutas das caixas delimitadoras (x_{min}, y_{min}, largura, altura) e, fundamentalmente para este estudo, as coordenadas dos polígonos que permitem o treinamento da cabeça de segmentação semântica do ConvNeXt V2.
   - Formato YOLO (.txt): Formato gerado de forma complementar para manter a compatibilidade com a base de dados dos estudos prévios (24), que empregaram a arquitetura YOLOv8. Diferentemente do COCO, este padrão utiliza arquivos de texto descentralizados (um para cada imagem) contendo coordenadas espaciais normalizadas (x_{centro}, y_{centro}, largura, altura).

III. Organização dos conjuntos de dados

Para criar o conjunto de dados, a base pré-processada foi inicialmente dividida em conjuntos de treinamento (80%) e de teste (20%). Para permitir a otimização de hiperparâmetros e o uso de Early Stopping sem enviesar os resultados finais, o conjunto de treinamento foi subdividido internamente, extraindo-se um subconjunto de validação. Mitigamos o vazamento de dados (data leakage) garantindo que o particionamento, incluindo a validação, ocorresse estritamente ao nível do exame (paciente). O conjunto de teste refinado, composto por 587 imagens, foi mantido completamente isolado e intocado

até a etapa de avaliação final, garantindo uma validação justa das morfologias de nódulos.

As imagens foram mantidas em sua resolução original (512×512) com representação de 8 bits. No total, o conjunto de treinamento contém 3.255 imagens, enquanto o de teste contém 832, abrangendo nódulos com diâmetro d no intervalo $d \in [3, 32]$ mm.

1) Refinamento do Conjunto de Teste (Foco em Nódulos Pequenos): Para analisar a capacidade dos modelos de detectar nódulos pequenos, identificamos os nódulos do conjunto de testes com diâmetro inferior a 15 mm.

Este processo de refinamento envolveu a separação de todos os cortes de cada exame de TC no conjunto de teste. Em seguida, verificou-se a maior fatia nesse conjunto de seções tinha diâmetro inferior a $15\,\mathrm{mm}$. Caso essa condição fosse satisfeita, todas as fatias desse exame seriam mantidas. Aqueles exames que falharam nesta verificação (ou seja, continham um corte $\geq 15\,\mathrm{mm}$) foram removidos do conjunto de teste final.

Com isso, das 832 imagens originalmente no conjunto de teste, apenas 587 permaneceram, sendo utilizadas para a validação dos modelos de detecção.

Além disso, o conjunto de teste refinado foi subdividido em três intervalos (I1, I2 e I3) de diâmetros de nódulos, visando avaliar o desempenho dos modelos de forma categorizada, conforme demonstrado na Tabela I.

Tabela I
DISPOSIÇÃO DAS IMAGENS NO CONJUNTO DE TESTE REFINADO, CATEGORIZADAS POR FAIXA DE DIÂMETRO DOS NÓDULOS.

Diâmetro (d)	**N.º Cortes**	**Nódulos Únicos**	**Exames de TC**
I1: $15\,\mathrm{mm} \geq d > 10\,\mathrm{mm}$	172	31	27
I2: $10\,\mathrm{mm} \geq d > 5\,\mathrm{mm}$	385	127	77
I3: $5\,\mathrm{mm} \geq d$	30	5	5

Fonte: Autor, 2026.

IV. TREINANDO O MODELO

O treinamento de ambos os modelos foi conduzido em um servidor remoto, equipado com o sistema operacional Linux (Ubuntu 20.04 LTS) e uma placa gráfica NVIDIA Tesla T4 com 16 GB de memória. O procedimento de adaptação e treinamento dos modelos baseou-se na técnica de Aprendizado por Transferência, do inglês, TL, visando acelerar a convergência das redes.

A. Detection Transformer (DETR)

O treinamento do modelo DETR foi iniciado com a criação de um *fork* do repositório original no GitHub (4). Esse procedimento foi essencial para realizar as modificações no código-fonte e adaptar a arquitetura, que utiliza o ResNet-50 como *backbone* para extração de características, à tarefa de detecção de nódulos. Os hiperparâmetros da arquitetura (detalhados na Tabela II) foram definidos empiricamente, a partir das configurações padrão recomendadas pelos autores originais do modelo, e refinados com base em testes preliminares, não sendo necessários algoritmos de busca exaustiva (como *Grid Search*).

O procedimento inicial consistiu na adaptação da última camada da rede: o número de classes foi reduzido para 1 (apenas a classe "nódulo"). Consequentemente, o número de consultas de objeto (*object queries*), que determina a quantidade máxima de caixas delimitadoras previstas por imagem, foi drasticamente reduzido de 100 para 10, uma vez que a quantidade de nódulos em um único corte pulmonar é caracteristicamente baixa.

O modelo foi configurado para executar por 400 épocas, com um tamanho de lote (*batch size*) de 8. Para garantir que todos os tensores possuam dimennsões identicas, a arquitetura original do DETR aplica automaticamente a técnica de *padding* (preenchimento com zeros), uniformizando o tamanho final de todas as imagens submetidas ao treinamento em um mesmo *batch*.

Devido a limitações de memória na GPU, não se utilizou a versão estendida DETR-DC5, mantendo-se o ResNet-50 padrão. O modelo concluiu as 400 épocas sem acionar o mecanismo de parada antecipada (*Early Stopping*).

Tabela II
HIPERPARÂMETROS DE TREINAMENTO E ARQUITETURA DO MODELO DETR.

Parâmetro	**Valor**
Número de Camadas Encoder	6
Número de Camadas Decoder	6
Tamanho dos Embeddings	256
Slots para Detecção (*Object Queries*)	10
Dropout no Transformer	10^{-1}
Número de Attention Heads	8

Fonte: Autor, 2026.

B. ConvNeXt V2

Para a segunda arquitetura, utilizou-se o modelo Mask R-CNN acoplado ao *backbone* ConvNeXt V2 (na variante *Tiny*), implementado com a biblioteca Hugging Face. A abordagem de *Transfer Learning* aplicada consistiu na extração dos pesos de redes pré-treinadas sobre o amplo conjunto de imagens genéricas ImageNet-1k (42). A partir deste estado inicial, executou-se o *fine-tuning* (ajuste fino) de todo o modelo, treinando-o e ajustando os pesos exclusivamente com base na base de dados de nódulos pulmonares extraída da LUNA16. Os parâmetros resultantes desta configuração encontram-se na Tabela III.

A rede foi integrada a uma *Feature Pyramid Network* (FPN), permitindo a extração de mapas de características em múltiplas escalas. Para adequar o modelo à morfologia de micronódulos (diâmetros predominantemente inferiores a 15 mm), os geradores de *anchors* (caixas-âncora da rede) foram reconfigurados para as escalas reduzidas de 4, 8, 12 e 16 pixels. A camada *Global Response Normalization* (GRN), característica nativa do ConvNeXt V2, foi ativada para atuar diretamente na distinção visual entre as opacidades nodulares e as ramificações vasculares.

O processo de *fine-tuning* foi planejado com um limite máximo de 400 épocas, utilizando o otimizador AdamW com uma taxa de aprendizado inicial de 2×10^{-4} e um *batch size* de 16. Implementou-se um mecanismo de *Early Stopping* com tolerância de 50 épocas sem evolução na perda (*loss*) de validação. Para a otimização contínua dos hiperparâmetros em tempo de execução, foi incorporado o escalonador *ReduceLROnPlateau*, que aplica um fator de redução de 0,1 à taxa de aprendizado caso não haja melhoria na métrica de avaliação por 5 épocas consecutivas. O checkpoint final selecionado para a avaliação no conjunto de teste foi estritamente o que apresentou a menor função de perda (loss) no conjunto de validação durante o treinamento.

Tabela III
PARÂMETROS DA ARQUITETURA CONVNEXT V2 + MASK R-CNN.

Descrição do Parâmetro	**Valor**
Arquitetura de Detecção	Mask R-CNN
Backbone (Extrator de Features)	ConvNeXt V2 (Tiny)
Neck (Conexão Multinível)	FPN
Dimensão dos Canais (FPN)	256
Otimizador	AdamW
Taxa de Aprendizado Inicial (LR)	2×10^{-4}
Escalas das *Anchors*	(4, 8, 12, 16) pixels
Proporção das *Anchors*	(0.8, 1.0, 1.2)
Mecanismo de Atenção	GRN
Dropout / Stochastic Depth	10^{-1}

Fonte: Autor, 2026.

É importante ressaltar que, devido à limitação de memória da GPU, optou-se pela configuração padrão do DETR (ResNet-50 com estágio C5 convencional), em vez da variante DETR-DC5.

V. RESULTADOS E DISCUSSÃO

A Figura 2 apresenta os resultados do modelo DETR para erro de classe e mAP@0.50 obtidos após 400 épocas de treinamento. O modelo DETR (curva laranja) apresenta redução no erro de classe ao longo do treinamento, com a curva de mAP@0.50 exibindo leve tendência ascendente até as últimas épocas registradas, indicando que o modelo ainda possuía margem de aprendizado espacial.

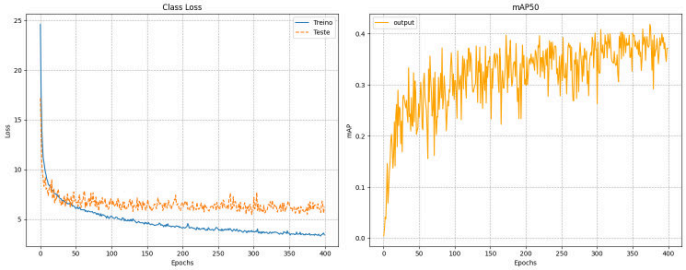


Figura 2. Plotagem dos logs de treinamento do modelo DETR. Fonte: Autor (2026).

Em contraste, a Figura 3 apresenta os logs de treinamento do modelo ConvNeXt V2. O treinamento demonstrou rápida convergência e excelente estabilidade em tarefas multitarefa. A adoção desta arquitetura corrobora achados recentes da literatura, que apontam o ConvNeXt V2 não apenas como um extrator de características altamente eficiente, mas também como superior em otimização computacional (5). A utilização de técnicas avançadas nativas da arquitetura ConvNeXt V2, como a Normalização de Resposta Global (GRN), mitiga problemas de colapso de características observados em redes mais profundas, garantindo uma competição eficaz entre os canais de características (5).

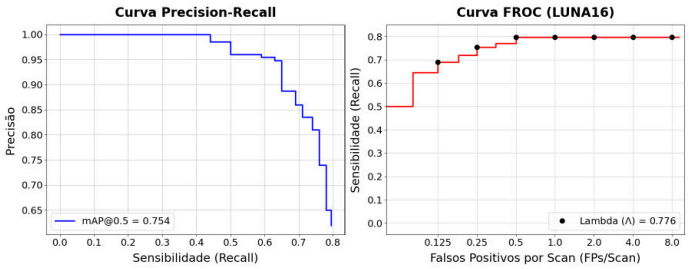


Figura 3. Plotagem dos logs de treinamento do modelo ConvNeXt V2 evidenciando a convergência das perdas. Fonte: Autor (2026).

A. Avaliação Quantitativa e Impacto do Tamanho do Nódulo

A avaliação quantitativa concentrou-se inicialmente na comparação direta de desempenho geral (Tabela IV). Para fins de análise, ressalta-se que a precisão indica a proporção de previsões positivas corretas, sendo valores mais altos associados a menos falsos positivos (FP). Já a revocação (sensibilidade) indica a capacidade do modelo de identificar corretamente os casos positivos, em que valores mais altos estão associados a menos falsos negativos (FN). Para garantir uma comparação rigorosa, as métricas foram calculadas exclusivamente com base nas bounding boxes geradas por ambos os modelos, sob os mesmos critérios de limiar de confiança e de Intersection over Union (IoU). Sob essas condições, o ConvNeXt V2 obteve um mAP@0.50 de 0,7535, uma precisão de 0,8106 e uma revocação de 0,7483. Esse índice tem relevância clínica direta, pois indica a redução de falsos alarmes na rede. O equilíbrio entre precisão e capacidade de detecção foi evidenciado por um F1-Score de 0,7782, superando significativamente o baseline do DETR. Para investigar a robustez arquitetural em função da morfologia dos tumores, o conjunto de teste foi subdividido em faixas de diâmetro (Tabela V). A avaliação clínica, validada por meio da análise FROC (Figura 4), revelou que, na classe de nódulos entre 10 e 15 mm (L1), o ConvNeXt V2 obteve um score Λ de 0,6454, mantendo uma taxa reduzida de falsos positivos na tarefa de detecção de nódulos.

Tabela IV
COMPARAÇÃO GERAL DAS MÉTRICAS DE DETECÇÃO ENTRE OS MODELOS AVALIADOS.

Modelo	**mAP@0.50**	**F1-Score**
DETR (Transformer)	0,703	0,772
ConvNeXt V2 (CNN)	**0,7535**	**0,7782**

Fonte: Autor, 2026.

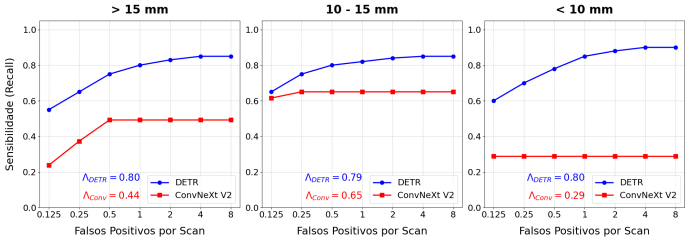


Figura 4. Plotagem da curva FROC dos dois modelos, ConvNeXt V2 em vermelho e DETR em azul, na qual a métrica Score representa o valor da sensibilidade de cada modelo. Fonte: Autor (2026).

Tabela V
DESEMPENHO COMPARATIVO ENTRE OS MODELOS DETR E CONVNEXT V2, CATEGORIZADO POR FAIXAS DIMENSIONAIS DE NÓDULOS.

Modelo	mAP@0.50			Sensibilidade			Score Λ		
	I1	I2	I3	I1	I2	I3	I1	I2	I3
DETR	**0,66**	**0,73**	**0,67**	0,86	0,85	**0,90**	**0,80**	**0,79**	**0,80**
ConvNeXt V2	0,61	0,26	0,04	**0,86**	**0,86**	0,85	0,65	0,29	0,06

Fonte: Autor, 2026.

B. Inferências e Avaliação Qualitativa

Durante a inferência, o modelo ConvNeXt V2 operou em multitarefa, realizando simultaneamente a detecção espacial por meio de *bounding boxes*, filtrando predições espúrias com base em um limiar de confiança de 0,8. As inferências geradas podem ser observadas visualmente por meio da sobreposição das predições em relação à verdade de referência (*ground truth*). A Figura 5 exibe os resultados do baseline DETR, enquanto a Figura 6 apresenta o painel qualitativo do ConvNeXt V2. Seguindo a ordem da esquerda para a direita e de cima para baixo, as imagens representam nódulos com $d > 15$ mm, seguidas de aproximações dos nódulos pertencentes aos intervalos L1 e L3, respectivamente.

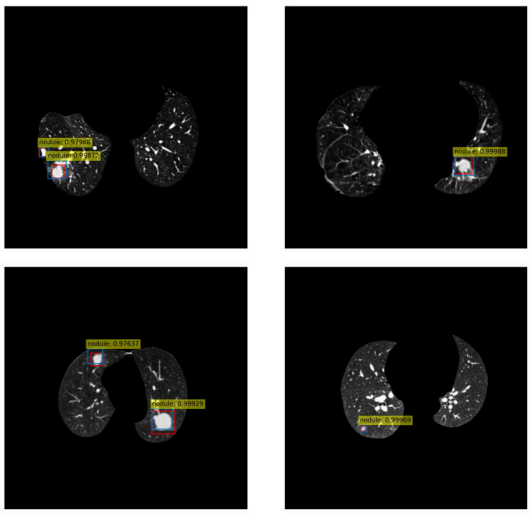


Figura 5. Inferências do modelo DETR (em azul) sobrepostas aos valores de *ground truth* (em vermelho). Fonte: (31).

Nas predições do ConvNeXt V2 (Figura 6), as caixas delimitadoras vermelhas representam a inferência do modelo, enquanto as tracejadas azuis confirmam a localização mapeada pelos radiologistas. O grande diferencial desta arquitetura reside na projeção simultânea da máscara de segmentação translúcida diretamente sobre os pixels anômalos. Essa dupla saída permite aos especialistas médicos avaliar não apenas a localização espacial, mas também a morfologia exata da lesão.

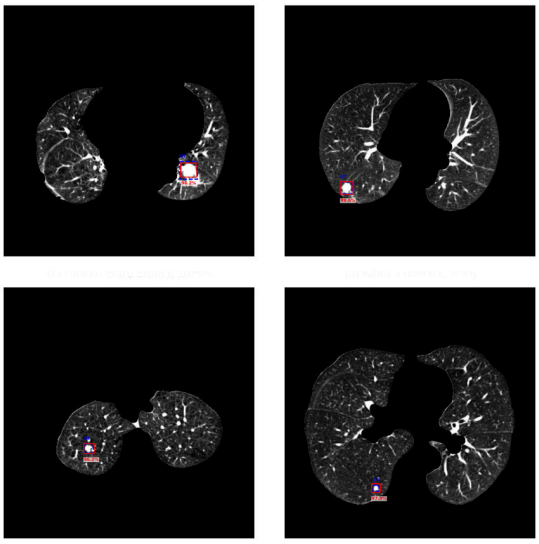


Figura 6. Inferências do modelo ConvNeXt V2 (predições e máscaras em vermelho) validadas contra as marcações de *ground truth* (caixas azuis tracejadas). Fonte: Autor (2026).

C. Análise de Custo Computacional

O pipeline customizado garantiu não apenas alta precisão, mas também eficiência operacional. O modelo foi treinado por 400 épocas. Conforme apresentado na Tabela VI, o treinamento do DETR exigiu 10 minutos por época, totalizando aproximadamente 66 horas. Em contrapartida, a eficiência dos blocos convolucionais modernos permitiu que o ConvNeXt V2 processasse cada época em apenas 2 minutos e 40 segundos.

Tabela VI
COMPARATIVO DE CUSTOS COMPUTACIONAIS ENTRE AS ARQUITETURAS.

Modelo	Treinamento (por época)	Inferência (por imagem)
DETR	10 minutos	600 ms
ConvNeXt V2	**2 min 40 s**	**120 ms**

Fonte: Autor, 2026.

Na inferência, o ConvNeXt V2 reduziu o tempo de processamento de 600 ms para apenas 120 ms por corte tomográfico. Esse ganho computacional reforça a viabilidade de implantação deste modelo como um CAD operável no fluxo de trabalho clínico padrão.

VI. CONCLUSÃO

Este trabalho apresentou uma comparação entre um algoritmo de detecção de nódulos pulmonares, baseado na arquitetura DETR, e um modelo proposto, baseado na arquitetura

ConvNeXt V2, acoplada ao *framework* Mask R-CNN. Exames de tórax obtidos do conjunto público de dados LUNA16 foram utilizados, e as imagens 2D foram adaptadas aos modelos de detecção escolhidos.

O modelo ConvNeXt V2 apresentou bons resultados, superando o baseline estabelecido por (31). Na avaliação quantitativa, o modelo obteve mAP@0.50 de 0,7535 e F1-Score de 0,7782, mantendo capacidade satisfatória de detecção de nódulos pequenos e baixa taxa de falsos positivos.

As caixas delimitadoras detectadas apresentaram um alinhamento preciso em relação aos valores de verdade absoluta (ground truth). Além da superioridade em precisão geométrica, o ConvNeXt V2 apresentou uma expressiva vantagem computacional, reduzindo o tempo de treinamento por época de 10 minutos (DETR) para 2 minutos e 40 segundos.

Conclui-se que o modelo demonstra excelente desempenho interno na base LUNA16, configurando-se como uma arquitetura promissora para futuras validações clínicas em cenários reais de Auxílio ao Diagnóstico (CAD).

VII. Trabalhos Futuros

Abaixo, algumas melhorias futuras:

- Exploração de Dados 3D: o uso de imagens 2D foi necessário para adaptar os exames aos modelos utilizados. No entanto, a utilização de modelos que operem diretamente em volumes 3D pode captar melhor as características espaciais dos nódulos pulmonares e, potencialmente, aumentar a acurácia da detecção.
- Aumento e Diversificação do Conjunto de Dados: ampliar o conjunto de dados, incluindo exames de diferentes populações e de diferentes equipamentos, pode tornar o modelo mais robusto e generalizável a diferentes cenários clínicos.
- Aprimoramento da etapa de Pós-Processamento: investigar técnicas mais avançadas de pós-processamento, como métodos de redução de falsos positivos e de refinamento das caixas delimitadoras, pode contribuir para um desempenho ainda melhor.
- Avaliação Clínica e Interpretabilidade: realizar avaliações em cenários clínicos reais e desenvolver mecanismos para tornar o modelo mais interpretável para profissionais da saúde são passos importantes para a adoção prática da tecnologia.
- Aplicação em algoritmos de classificação: utilizar o ConVNext aninhado com o DETR com tarefas que se complementam entre si, o primeiro avalia se existe nódulo ou não e o segundo detecta com precisão onde ele está e dá o diagnóstico
- Aplicações em outras áreas da saúde: investigar quais formas de arquitetura poderiam ser úteis na detecção de outras doenças no corpo humano.

Referências

[1] P. M. Azevedo-Marques, "Diagnóstico auxiliado por computador na radiologia," *Radiol Bras*, vol. 34, 2001.

[2] A. I. Bandos, T. Song, and D. Gur, "Area under the Free-Response ROC Curve (FROC) and a Related Summary Index," *Biometrics*, 2019.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.

[4] N. Carion *et al.*, "End-to-End Object Detection with Transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[5] S. Cheng, E. Sun, W. Qian, and Y. Han, "A Breast Cancer Detection Model Based on Modified Convnext v2," in *2023 3rd Guangdong-Hong Kong-Macao Greater Bay Area Artificial Intelligence and Big Data Forum (AIBDF 2023)*, 2023, pp. 556–559.

[6] G. Coppini, S. Diciotti, M. Falchini, N. Villari, and G. Valli, "Neural Networks for Computer-Aided Diagnosis: Detection of Lung Nodules in Chest Radiograms," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 4, 2003.

[7] S. D'Oro, L. Galluccio, S. Palazzo, and G. Schembra, "A Game Theoretic Approach for Distributed Resource Allocation and Orchestration of Softwarized Networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 3, pp. 721–735, 2017.

[8] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021.

[9] Q. Dou *et al.*, "Automated Pulmonary Nodule Detection via 3D ConvNets with Online Sample Filtering and Hybrid-Loss Residual Learning," in *MICCAI*, 2017.

[10] A. X. Falcão, "Visualização de Volumes Aplicada à Área Médica," Tese de Mestrado, Faculdade de Engenharia Elétrica, 1993.

[11] J. Geppert *et al.*, "Software using artificial intelligence for nodule and cancer detection in CT lung cancer screening: systematic review of test accuracy studies," *Thorax*, vol. 79, no. 11, pp. 1040–1049, 2024.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[13] A. Halder, D. Dey, and A. K. Sadhu, "Lung Nodule Detection from Feature Engineering to Deep Learning in Thoracic CT Images: a Comprehensive Review," *J. Digit. Imaging*, vol. 33, no. 3, pp. 655–677, 2020.

[14] K. He *et al.*, "Transformers in Medical Image Analysis: A Review," 2022.

[15] Instituto Nacional de Câncer (INCA), "Estimativa 2025: incidência de câncer no Brasil," Rio de Janeiro, 2024. [Online]. Available: https://www.inca.gov.br/publicacoes/livros/estimativa-2025-incidencia-de-cancer-no-brasil. Acessado em: 07 jul. 2026.

[16] S. B. Knight *et al.*, "Progress and prospects of early detection in lung cancer," *Open Biol.*, vol. 7, no. 9, p. 170070, 2017.

[17] H. W. Kuhn, "The hungarian method for the assignment problem. Naval research logistics," *Quartely*, 1955.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recogni-

tion," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[19] J. S. Lee *et al.*, "AI-Based Pulmonary Embolism Detection: The Added Value of a False-Positive Reduction Module over a Region Proposal Network," *Diagnostics*, vol. 16, no. 524, 2026.
[20] Z. Li *et al.*, "ConvNeXt-Based Fine-Grained Image Classification and Bilinear Attention Mechanism Model," *Appl. Sci.*, 2022.
[21] L. L. de Lima, "Modelo computacional para classificação de nódulos pulmonares utilizando redes neurais convolucionais," Tese de Mestrado, Universidade Federal de Alagoas (UFAL), 2019.
[22] Z. Liu *et al.*, "A ConvNet for the 2020s," *CoRR*, vol. abs/2201.03545, 2022.
[23] J. Lu *et al.*, "Softmax-free Linear Transformers," *Int. J. Comput. Vis.*, 2024.
[24] V. F. Mafra, "Análise comparativa entre detection transformer e yolov8 para detecção precoce de nódulos pulmonares," *UFAL*, 2024.
[25] T. M. Mitchell, *Machine Learning*. New York, NY: McGraw-Hill, 1997.
[26] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, 1st ed. MIT Press, 2012.
[27] N. Nasrullah, M. S. Alam, M. Mateen, B. Cai, and H. Haibo, "Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies," *Sensors (Basel)*, 2019.
[28] J. Qi, M. Nguyen, and W. Q. Yan, "Waste Classification from Digital Images Using ConvNeXt," in *Springer-Verlag*, 2023.
[29] R Development Core Team, *The R Project for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.
[30] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Efficient ConvNet for real-time semantic segmentation," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1789–1794.
[31] N. S. de Carvalho *et al.*, "Comparative Analysis of Detection Transformers and YOLOv8 for Early Detection of Pulmonary Nodules," in *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, 2024.
[32] A. Schreuder, E. T. Scholten, B. van Ginneken, and C. Jacobs, "Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice?," *Transl. Lung Cancer Res.*, 2021.
[33] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, 2016.
[34] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer J. Clin.*, vol. 70, no. 1, pp. 7–30, Jan. 2020.
[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: The MIT Press, 2018.
[36] Y. Taspinar *et al.*, "Computer vision classification of dry beans (Phaseolus vulgaris L.) based on deep transfer learning techniques," *Eur. Food Res. Technol.*, vol. 248, 2022.
[37] The Cancer Imaging Archive, "LIDC-IDRI — Dados do Lung Image Database Consortium (LIDC) e da Image Database Resource Initiative (IDRI)," 2015. [Online]. Available: https://www.cancerimagingarchive.net/collection/lidc-idri/
[38] World Health Organization, "Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019," 2020. [Online]. Available: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death
[39] S. Woo *et al.*, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," 2023.
[40] X. Zhang, Z. Huang, C. Wu, Z. Li, and F. C. M. Lau, "Online Stochastic Buy-Sell Mechanism for VNF Chains in the NFV Market," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 392–406, 2017.
[41] T. Zhao *et al.*, "Integrated 3D Deep Learning Approach: YOLO-Decoupled Candidate Detection and 3D CNN-Driven False Positive Reduction for Lung Nodule Analysis in CT," *Research Square*, 2025.
[42] K. Zhao *et al.*, "Automated segmentation and source prediction of bone tumors using ConvNeXtv2 Fusion based Mask R-CNN to identify lung cancer metastasis," *J. Bone Oncol.*, vol. 48, p. 100637, 2024.
[43] X. Zhu *et al.*, "Channel-Wise Attention Mechanism in the 3D Convolutional Network for Lung Nodule Detection," *Electronics*, vol. 11, 2022.
[44] International Agency for Research on Cancer, "Global Cancer Observatory: Lung Cancer Fact Sheet," *World Health Organization*, 2024. [Online]. Available: https://who.int

Deep Learning Models for Real-Time Heart Rate Estimation Using Remote Photoplethysmography: A Single-Subject Pilot Study of an ICU Patient

Larissa Albuquerque Nogueira
Biomedical Engineering Laboratory (LEB)
University of São Paulo
São Paulo, Brazil
larissaanogueira@usp.br

Sophia Celine Rafael Alves Pereira
Biomedical Engineering Laboratory (LEB)
University of São Paulo
São Paulo, Brazil
sophiaceline@usp.br

Felipe Fava de Lima
Biomedical Engineering Laboratory (LEB)
University of São Paulo
São Paulo, Brazil
felipefavadelima@usp.br

Henrique Takachi Moriya
Biomedical Engineering Laboratory (LEB)
University of São Paulo
São Paulo, Brazil
htmoriya@usp.br

Paolo José Cesare Biselli
University Hospital
University of São Paulo
São Paulo, Brazil
paolobiselli@usp.br

***Abstract*—Remote photoplethysmography (rPPG) has emerged as a promising contactless technique for heart rate estimation by extracting pulse-related variations from facial video recordings. Several deep learning approaches have been proposed to improve the accuracy and robustness of heart rate measurement using rPPG. This technology represents a potential complementary alternative for physiological monitoring in Intensive Care Units (ICUs), as it is non-invasive, contactless and consequently reduces dependence on physical sensors and direct patient contact. This study presents a pilot comparative analysis of the inference time and heart rate estimation performance of six state-of-the-art deep learning models using remote photoplethysmography in a single intensive care unit patient. The evaluated models represent different deep learning architectures reported in the rPPG literature, including Convolutional Neural Networks, Transformers and State Space Models. For each model, two training configurations are considered using the PURE and RLAP datasets. Heart rate is estimated from a video recording for each 15-second interval in non-overlapping windows. The evaluated models are compared in terms of mean absolute error (MAE), Pearson correlation coefficient, and total inference time. The preliminary experimental results showed that TS-CAN presented the shortest total inference time of 5.66 seconds for the analyzed video, whereas EfficientPhys trained on RLAP achieved the lowest estimation error, with MAE of 2.74 $\pm$ 1.26 BPM and total inference time of 14.79 seconds.**

***Index Terms*—Remote photoplethysmography, heart rate measurement, deep learning models, intensive care units, real-time monitoring.**

I. Introduction

Heart rate monitoring is an important component of patient assessment, as variations in heart rate may provide early indications of physiological deterioration and changes in cardiovascular status [1]. In Intensive Care Units (ICUs), early detection of patient deterioration is essential to enable timely clinical intervention. Heart rate is one of the primary physiological parameters continuously monitored to assess a patient's clinical status and it is traditionally performed using contact-based devices and sensors, such as electrocardiography systems and pulse oximeters [2]. However, prolonged use of these devices may cause discomfort, skin irritation, and restrictions to patient mobility [3].

Remote photoplethysmography (rPPG) has emerged as a promising non-contact approach for heart rate measurement. This method uses video recordings acquired with conventional RGB cameras to capture subtle temporal variations in reflected light intensity caused by blood volume changes during the cardiac cycle [4]. When ambient light reaches the subject's skin, part of it is reflected directly from the skin surface, while another part penetrates the tissue, where it undergoes absorption and scattering before re-emerging at the skin surface. This latter contribution carries information about blood volume variations [5]. The camera captures a combination of both components, and rPPG algorithms seek to estimate cardiovascular signals from these subtle optical variations despite the presence of non-physiological light reflections, motion artifacts, and other sources of noise.

To address these challenges and motivated by the success of deep learning in other computer vision tasks, several deep learning models have been proposed for heart rate estimation from rPPG signals, incorporating mechanisms to improve robustness to different sources of noise. Among the proposed approaches, Convolutional Neural Networks (CNNs) were the earliest deep learning architectures explored for rPPG heart rate estimation. Liu et al. [6] proposed the Temporal Shift Attention Network (TS-CAN), a model using a 2D CNN architecture that incorporates a Temporal Shift Module (TSM)

This study was financed in part by Amigos da Poli endowment fund under CARE 4.0 project.

to enable temporal information exchange between consecutive frames. The TSM shifts feature maps along the temporal dimension, allowing the network to capture temporal information between multiple frames.

Standard 2D CNNs effectively extract spatial features from individual frames but do not explicitly model temporal dependencies. To capture temporal dynamics, 3D CNNs extend convolution operations to both spatial and temporal dimensions [7]. In the context of rPPG, Yu et al. [8] proposed PhysNet, a 3D CNN architecture with an encoder-decoder structure for rPPG signal estimation, and compared its performance with 2D CNN networks. The results demonstrated that the encoder-decoder 3D CNN architecture outperformed the evaluated architectures. However, 3D CNNs generally incur higher computational and memory costs because convolutions are performed jointly across spatial and temporal dimensions.

An alternative to CNNs is the Transformer architecture. Transformers were originally developed for sequence modeling using self-attention mechanisms [9]. Subsequent adaptations to computer vision and video analysis enabled these models to capture long-range spatial and temporal dependencies across video frames. Yu et al. [10] proposed PhysFormer, which employs temporal difference transformer to extract spatiotemporal physiological features and estimate rPPG signals from long video sequences. Additionally, the authors introduced a frequency-domain supervision strategy to encourage the extraction of pulse-related information and improve robustness to motion and illumination variations. Although Transformers are highly effective at modeling temporal dependencies, they generally incur high computational costs and often require larger training datasets than CNNs.

To address the computational limitations of Transformers, Mamba was proposed as a State Space Model (SSM) architecture capable of modeling long sequences with linear computational complexity [11]. Building on this architecture, Zou et al. [12] proposed RhythmMamba for rPPG signal estimation. The authors showed that the model achieves competitive accuracy compared with state-of-the-art methods while providing approximately three times higher throughput and reducing peak memory usage by about 30%. Several additional architectures have been proposed in the literature, most of which are variations or combinations of CNNs, Transformers and SSM approaches. Each architecture involves different trade-offs between computational training and inference costs, temporal and spatial feature representation capabilities, real-time suitability, and robustness to noise.

Despite the recent expansion of deep learning models for heart rate estimation using rPPG, systematic comparisons of these architectures and their suitability for clinical real-time applications remain limited, especially under challenging clinical conditions such as those found in ICUs. Models reported in the literature are commonly developed and evaluated using healthy subjects under controlled conditions, often with limited demographic diversity [13]. Unlike standard applications, such as fitness monitoring or online health assessment, ICUs require methods that combine accurate heart rate estimation with low inference latency and robust performance under challenging acquisition conditions.

As an initial investigation of these research gaps, this study presents a pilot comparative analysis of six deep learning models reported in the rPPG literature and evaluates their suitability for real-time inference using a single video acquired from an ICU patient. For each model, two training configurations are considered using the RLAP [14] and PURE [15] datasets, and performance is assessed in terms of inference time, mean absolute error (MAE), and Pearson correlation coefficient.

II. Methodology

A. Data Acquisition

The acquired data consisted of a facial video and a reference heart rate signal obtained from a single ICU patient admitted in the University Hospital of University of São Paulo (HU-USP)[1]. The video duration was 2 minutes and was recorded using a VCXU.2-57C camera (Baumer, Switzerland) positioned 190 cm from the patient's face and horizontally aligned with the patient's head. The recording was acquired at a resolution of 2000 × 2028 pixels, a frame rate of 25 fps, and an exposure time of 28 ms. Ambient illuminance at the acquisition site was 880 lux, measured using a Fluke 941 lux meter (Fluke Corporation, USA).



(a) (b)

Fig. 1: Sample frames from the recorded video. Figures (a) shows the patient sitting still without speaking, and (b) shows the patient speaking while displaying a concerned facial expression.

The patient was a woman with skin type classified as Fitzpatrick phototype IV [18]. Skin phototype was assessed using a skin phototype analyser (Skinup Beauty Devices, Brazil), with the measurement performed on the patient's forehead. The Fitzpatrick scale was adopted to facilitate the standardized characterization of skin phototype in future data acquisitions.

[1]All the data collected in this work was approved by the University of São Paulo Research Ethics Commitee under Ethics Review Certificate (CAAE) number 86650625.3.0000.0076. All data were collected with the explicit permission of the individuals involved or their representatives; data collection took place only after the participant or their legal representative had read and signed the Informed Consent Form.

TABLE I: Selected deep learning models and characteristics

Model name	Type	Number of frames per inference	Input size
ME-chunk [16]	SSM	160	36 x 36
ME-flow [16]	SSM	1	36 x 36
RhythmMamba [12]	Mamba	160	128 x 128
PhysFormer [10]	Transformer	160	128 x 128
TS-CAN [6]	2D CNN	160	36 x 36
EfficientPhys [17]	2D CNN	160	72 x 72
PhysNet [8]	3D CNN + Encoder-Decoder	128	32 x 32

The recording consists of the patient sitting still and speaking, exhibiting subtle head movements as well as occasional rapid head movements and facial expressions. Two sample frames from the recorded video is shown in Fig. 1.

The reference heart rate values were obtained from the photoplethysmography (PPG) signal acquired using a BSM-3000 Series bedside monitor (Nihon Kohden, Japan) at a sampling frequency of 62.5 Hz. The PPG signal was processed using the Python library HeartPy [19], which estimates heart rate from the detected pulse peaks. The resulting heart rate time series was used as the ground truth.

B. Deep Learning Models Selection

The deep learning models evaluated in this study were selected based on benchmark studies and recent literature in the rPPG field, with the objective of including representative architectures and implementations. The selected models have been reported to achieve high accuracy or competitive performance in previous researches.

All models were obtained from the publicly available rPPG toolkit *open-rppg* [20], which provides implementations and pretrained weights for multiple deep learning architectures. The toolkit includes pretrained models trained on the RLAP [14] and PURE [15] datasets. The model variants are denoted in this work as *ModelName-Dataset*, where the suffix indicates the weights used for model training (e.g., *PhysFormer-RLAP* and *PhysFormer-PURE*).

Table I presents the models evaluated in this work and their respective architectures, input size and number of frames used per inference. The models ME-chunk and ME-flow share the same underlying architecture but differ in the number of frames processed per inference.

C. Real-Time Inference and Evaluation Protocol

In order to simulate a real-time monitoring environment, the recorded video was processed using consecutive 15-second windows. For each window with index i, the error E between the model estimated heart rate $\hat{HR}$ and the ground truth GT was calculated (1), and the inference time T for each window was measured. The complete video frames were directly used as model input, without any preprocessing.

$$E[i] = |\hat{HR}[i] - GT[i]| \tag{1}$$

The ground truth GT used for error calculation was obtained by averaging the heart rate values obtained from the processed PPG signal within each temporal window. As shown in (2), gt represents the ground truth heart rate temporal series, where N denotes the number of heart rate samples contained in each 15-second window, defined as $N = \lfloor f_s \times 15 \rfloor$, and f_s is the oximeter sampling frequency in Hz.

$$GT[i] = \frac{1}{N} \sum_{n=iN}^{(i+1)N-1} gt[n] \tag{2}$$

After processing all windows in the video, the mean absolute error MAE (3), its standard deviation σ_{MAE} (4), the total inference time T_{inf} (5), and the Pearson correlation coefficient r (6) were calculated. The Pearson correlation coefficient was computed between the heart rate values estimated by the model and the corresponding ground-truth heart rate values for each window. In the equations, M denotes the number of windows into which the video is divided, $\overline{GT}$ is the mean ground-truth heart rate, and $\overline{\hat{HR}}$ is the mean heart rate estimated by the model across all windows. Variables MAE, σ_{MAE}, $E[i]$, $\hat{HR}[i]$, $GT[i]$, $\overline{GT}$ and $\overline{\hat{HR}}$ are expressed in bpm (beats per minute), while $T[i]$ and T_{inf} are expressed in seconds.

$$MAE = \frac{1}{M} \sum_{i=0}^{M} E[i] \tag{3}$$

$$\sigma_{\mathrm{MAE}} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (E[i] - MAE)^2} \tag{4}$$

$$T_{inf} = \sum_{i=0}^{M} T[i] \tag{5}$$

$$r = \frac{\sum_{i=1}^{M} \left(\hat{HR}[i] - \overline{\hat{HR}}\right)\left(GT[i] - \overline{GT}\right)}{\sqrt{\sum_{i=1}^{M} \left(\hat{HR}[i] - \overline{\hat{HR}}\right)^2} \sqrt{\sum_{i=1}^{M} \left(GT[i] - \overline{GT}\right)^2}}, \tag{6}$$

$$\overline{\hat{HR}} = \frac{1}{M} \sum_{i=1}^{M} \hat{HR}[i], \qquad \overline{GT} = \frac{1}{M} \sum_{i=1}^{M} GT[i] \tag{7}$$

TABLE II: Performance metrics of evaluated deep learning models trained on RLAP and PURE datasets. Best results in **bold**, second best are underlined.

Model	Training dataset	MAE $\pm\sigma_{MAE}$ (BPM)	Error median (BPM)	T_{inf} (s)	T (s)	r
ME-chunk	RLAP	11,29 ± 8,08	8,95	24,42	3,05	0,528
	PURE	20,14 ± 12,16	21,68	24,18	3,02	0,025
ME-flow	RLAP	18,36 ± 9,80	16,92	75,53	9,44	0,480
	PURE	21,97 ± 13,39	16,39	76,30	9,54	0,108
RhythmMamba	RLAP	9,60 ± 6,84	7,22	56,55	7,07	-0,659
	PURE	27,37 ± 13,02	18,80	54,85	6,86	0,501
PhysFormer	RLAP	5,41 ± 5,52	3,28	83,80	10,48	-0,297
	PURE	13,73 ± 11,68	12,38	82,57	10,32	-0,315
TS-CAN	RLAP	17,39 ± 13,05	14,93	**5,66**	**0,71**	0,198
	PURE	14,53 ± 9,84	10,84	**5,88**	**0,74**	0,454
EfficientPhys	RLAP	**2,74 ± 1,26**	2,54	14,79	1,85	0,060
	PURE	14,65 ± 13,79	12,01	14,88	1,86	-0,431
PhysNet	RLAP	7,66 ± 3,87	7,47	8,18	1,02	-0,470
	PURE	22,26 ± 18,47	16,37	8,40	1,05	-0,307

Processing was performed on a system equipped with a 12th Gen Intel® Core™ i5-1235U processor (12 logical cores), 12 GB of RAM, running Ubuntu 24.04 LTS. Each model was processed separately to avoid potential cache effects or residual system load from previous model inferences affecting the measured inference time. The source code used for the inference and evaluation procedures is publicly available on GitHub[2].

III. Results

Table II summarizes the performance of all evaluated models. Detailed experimental results for each model are reported in the GitHub repository, including the inference time for each window and the window-wise error for all models. The experiments showed negligible variation in inference time across windows for a fixed model and training dataset configuration. Therefore, Table II reports the average inference time per window T for each configuration.

The results show that EfficientPhys-RLAP achieved the lowest MAE, followed by PhysFormer-RLAP and PhysNet-RLAP. However, EfficientPhys-RLAP and PhysNet-RLAP provide a better trade-off between inference time and error. The model that achieved the lowest inference time was TS-CAN, followed by PhysNet and EfficientPhys. TS-CAN required a total inference time of 0.71 s for a 15-second window, while PhysNet required 1.00 s for the same window. The results also show that, for most architectures, models trained on the RLAP dataset achieved higher accuracy than the same models trained on the PURE dataset. The only exception was TS-CAN, which achieved higher accuracy when trained on PURE.

Most models exhibited a significant error ($MAE > 10$ BPM) and a large standard deviation of the absolute errors (σ_{MAE}), indicating high variability in the estimation error for each window. Therefore, the MAE alone does not fully characterize model performance. For this reason, the median absolute error was also included in Table II to provide additional information about the distribution of the absolute errors. Additionally, most models exhibited absolute Pearson correlation values between 0 and 0.5 ($0 < |r| < 0.5$), indicating a weak linear relationship between the estimated and reference heart rate values.

[2]https://github.com/larianog/icu-rppg

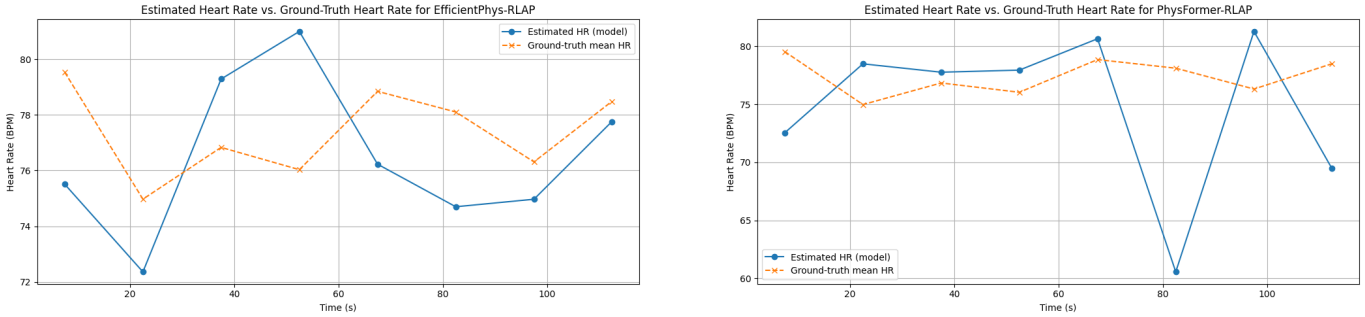
IV. Discussion

It is important to note that the results presented in this article constitute a preliminary study and that differences in model performance could have been influenced by the characteristics of the evaluated video. Additional data and further studies are required to assess the generalizability of these findings.

A. Analysis of Heart Rate Estimation Accuracy

The results show relatively high MAE values, large variability, and weak Pearson correlation coefficients. These findings are mainly attributed to the high variability of the heart rate estimates across inference windows. For a 2-minute video processed using 15-second windows, only eight inference windows are available; therefore, a small number of windows with large estimation errors can substantially increase the overall MAE. Similarly, the low Pearson correlation is related to fluctuations in the estimated heart rate temporal trend, which reduces the agreement between estimated and reference values.

In the processed videos, some windows contain faster head movements and more pronounced facial expressions, resulting in higher estimation errors compared with windows in which the patient remains still or speaks with limited facial movement. As examples of this behavior, Figure 2a presents the experimental results for EfficientPhys-RLAP, while Figure 2b presents the results for PhysFormer-RLAP. These figures illustrate the variability of the window-by-window heart rate estimates, with some windows showing overestimation and others underestimation, which contributes to the low Pearson



(a) (b)

Fig. 2: Comparison of the estimated and ground-truth heart rate for each window using (a) EfficientPhys-RLAP and (b) PhysFormer-RLAP. Although the absolute estimation error is the smallest presented, discrepancies in the temporal trend between successive windows result in a low Pearson correlation coefficient.

correlation. Additionally, some windows exhibit larger errors than others, increasing the overall MAE and its variance.

These results suggest that longer video recordings could potentially reduce fluctuations in the average estimation error and improve robustness to window-wise errors, as a larger number of inference windows would be available.

B. Impact of Model Architecture on Heart Rate Estimation Accuracy

Despite the overall high MAE values, EfficientPhys-RLAP, PhysFormer-RLAP, and PhysNet-RLAP achieved the lowest estimation errors among the evaluated models for this experiment. Although these results cannot be conclusively attributed to differences in model architectures, given the study's limitation to a single video from a single patient, these architectural differences may provide an initial basis for comparison and validation in further studies on this domain.

In this context, EfficientPhys uses frame differencing to capture temporal variations between adjacent frames and TSM to enable temporal information exchange between adjacent frames. PhysFormer applies a convolutional preprocessing stage to extract local spatio-temporal features and employs Temporal Difference Transformers to model temporal variations from frame differencing within the transformer architecture. PhysNet uses an encoder-decoder structure after the 3D convolutional layers to compact information into latent representations of physiological signals. These mechanisms, including frame differencing, spatio-temporal feature extraction, and latent representation learning, are designed to emphasize the subtle physiological variations associated with rPPG signals while improving robustness to motion and illumination artifacts.

However, it is important to note that the presence of individual architectural components does not necessarily determine performance. For example, TS-CAN also employs a 2D CNN backbone combined with TSM but achieved lower accuracy in this study. This difference may be related to the characteristics of the evaluated video, which can influence the generalization capability of each architecture.

Among the SSM-based models, RhythmMamba-RLAP and ME-chunk-RLAP achieved the best performance in terms of MAE. However, both models underperformed the evaluated CNN and Transformer models in this experiment. Although both networks incorporate mechanisms for frame differencing and extraction of subtle rPPG variations, differences on performance may be related to the mechanisms used to represent and encode information. Mamba architectures are designed for efficient long-range sequence modeling by maintaining hidden state representations of sequential information. However, this mechanism may introduce information compression, which could potentially reduce sensitivity to the subtle temporal variations associated with physiological signals.

As previously mentioned, the observed performance differences cannot be directly attributed to model architectures due to the data and demographic limitations of this study. Therefore, further evaluation on larger and more heterogeneous clinical populations is required to determine whether these performance patterns are consistent across model architectures, patients, and recording configurations.

C. Impact of Model Architecture on Computational Efficiency

Regarding total inference time, TS-CAN achieved the lowest inference time among the evaluated models, with a total inference time of 5.66 s. It was followed by PhysNet and EfficientPhys. The differences in inference time are influenced by the input size and computational complexity of each architecture. Additionally, the number of frames processed per inference affects the overall processing time.

TS-CAN uses a reduced input size of 36 × 36 pixels and is primarily composed of 2D convolutional layers combined with TSM. Since TSM does not increase computational complexity, the reduced input resolution and relatively simple architecture contribute to its lower inference time. PhysNet uses an input resolution of 32 × 32 pixels and relies mainly on 3D convolutional layers. Although 3D convolutions are computationally more demanding than 2D convolutions, the reduced spatial resolution decreases the computational cost of these operations, resulting in lower inference time.

In contrast, PhysFormer achieved the highest total inference time, requiring 83.80 s for inference when trained on RLAP. This result can be explained by its larger input resolution of 128 × 128 pixels and the combination of convolutional operations, self-attention mechanisms, and transformer layers, which increases the computational complexity of the model.

A similar trend was observed for RhythmMamba and ME-flow. RhythmMamba operates with 128 × 128 pixel inputs, increasing the computational cost of the Mamba layers. ME-flow uses a reduced input resolution of 36 × 36 pixels but processes one frame at a time to update its states. This sequential processing strategy increases the total inference time in this experimental setup due to the repeated state updates required for each frame. In contrast, ME-chunk uses the same input resolution as ME-flow but processes 160 frames per inference, reducing the number of sequential inference steps required and resulting in lower total inference time in this experiment.

Nevertheless, all evaluated models completed inference within the 15-second acquisition window, indicating that they are potentially suitable for the proposed real-time ICU monitoring framework. However, lower inference latency remains preferable, as longer inference times may increase system latency and reduce responsiveness.

D. Impact of Training Dataset on Model Generalization

Models trained on the RLAP dataset generally achieved lower MAE values than those trained on the PURE dataset. One possible explanation is the larger amount of training data and the greater variability in facial expressions present in RLAP compared with PURE. The RLAP [14] dataset is composed of 754 videos of Chinese students recorded under various scenarios, including problem-solving tasks, emotion-inducing videos, and hallucinatory image viewing. In contrast, PURE [15] consists of 60 videos of healthy subjects recorded under controlled conditions, including predefined head movements and sitting-still or speaking scenarios.

In the processed video, the patient displays frequent facial expressions while speaking and showing signs of concern or concentration. Therefore, the greater facial expression diversity in the RLAP dataset may have contributed to the improved performance of models trained on this dataset.

However, although facial expression variability may be better represented in the RLAP dataset, both training datasets have limited demographic diversity, with participants primarily from Asian and Caucasian populations. In addition, the datasets consist primarily of healthy subjects. As the patient evaluated in this study differs from these populations in both demographic and clinical characteristics, this mismatch represents an additional source of domain shift that may have affected model performance.

V. Conclusion

As the experimental results show, EfficientPhys-RLAP provided the best trade-off between heart rate estimation accuracy and inference time among the evaluated models, achieving a mean absolute error (MAE) of 2.74 ± 1.26 BPM and a total inference time of 14.79 s for the evaluated video. The next best trade-offs were obtained by PhysNet-RLAP and PhysFormer-RLAP. However, the estimation errors observed in this study may have been influenced by the characteristics of the single ICU patient video used for evaluation. Therefore, further studies are required to validate these findings using a larger dataset and longer video recordings, in order to enable a more reliable assessment of estimation error and correlation with the ground truth.

Regarding inference time, TS-CAN achieved the fastest inference, requiring 5.66 s to process the entire video when trained on RLAP, followed by PhysNet and EfficientPhys. Since inference time is primarily determined by the model architecture and implementation rather than the video content, these results can be used to guide the selection of models for future real-time rPPG monitoring studies in ICU environments.

Finally, further researches involving a larger and more diverse ICU dataset are required to investigate the impact of model architecture and training data on estimation accuracy and generalization under clinical conditions.

VI. Further Work

Future work will focus on acquiring additional data from ICU patients to evaluate the performance of different architectures and models on a larger and more diverse dataset. Furthermore, longer video recordings will be collected to assess model performance during continuous real-time monitoring. The expanded dataset will include variations in illumination conditions, patient characteristics, skin tones, and clinical scenarios, such as intubation, sleep, talking, and limited-motion conditions.

Following this initial assessment, the best-performing models will be selected for evaluation in a real-time video streaming environment to investigate their practical feasibility for ICU monitoring. This evaluation will consider different operational scenarios, including illumination variations, patient motion, and clinical conditions such as arrhythmias and intubation. Additionally, this work will aim to establish practical acquisition requirements for rPPG monitoring, including illumination levels, camera settings, distance from the patient's face, and relevant patient conditions.

References

[1] J.-L. Vincent, F. A. Moore, R. Bellomo, and J. J. Marini, Eds., *Textbook of Critical Care*, 8th ed. Elsevier, 2022.

[2] J. Jorge, M. Villarroel, H. Tomlinson, O. Gibson, J. L. Darbyshire, J. Ede, M. Harford, J. D. Young, L. Tarassenko, and P. Watkinson, "Non-contact physiological monitoring of post-operative patients in the intensive care unit," *npj Digital Medicine*, vol. 5, no. 1, p. 4, 2022. [Online]. Available: https://doi.org/10.1038/s41746-021-00543-z

[3] A. Soroudi, N. Hernández, L. Berglin, and V. Nierstrasz, "Electrode placement in electrocardiography smart garments: A review," *Journal of Electrocardiology*, vol. 57, pp. 27–30, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022073619304650

[4] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[5] W. Wang, A. den Brinker, S. Stuijk, and G. Haan, "Algorithmic principles of remote-ppg," *IEEE Transactions on Biomedical Engineering*, vol. PP, 09 2016.

[6] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 400–19 411, 2020.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[8] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 277.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4186–4196.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2024. [Online]. Available: https://arxiv.org/abs/2312.00752

[12] B. Zou, Z. Guo, X. Hu, and H. Ma, "Rhythmmamba: Fast, lightweight, and accurate remote physiological measurement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 11 077–11 085.

[13] D. McDuff, "Camera measurement of physiological vital signs," *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023. [Online]. Available: https://doi.org/10.1145/3558518

[14] K. Wang, Y. Wei, J. Tang, Y. Wang, Z. Li, M. Tong, J. Gao, Y. Ma, and Z. Zhao, "Camera-based hrv prediction for remote learning environments," in *2024 IEEE Smart World Congress (SWC)*, 2024, pp. 1165–1173.

[15] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 1056–1062.

[16] K. Wang, J. Tang, Y. Fan, J. Ji, Y. Shi, and Y. Wang, "Memory-efficient low-latency remote photoplethysmography through temporal-spatial state space duality," unpublished, 2025. [Online]. Available: https://arxiv.org/abs/2504.01774

[17] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 5008–5017.

[18] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of Dermatology*, vol. 124, no. 6, pp. 869–871, 06 1988. [Online]. Available: https://doi.org/10.1001/archderm.1988.01670060015008

[19] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "Heartpy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 368–378, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369847818306740

[20] K. Wang, "Open-rppg: An easy-to-use rppg inference toolbox," GitHub repository, accessed: Jun. 2026. [Online]. Available: https://github.com/KegangWangCCNU/open-rppg

Deep Learning-Based Classification of Breast Lesions in Ultrasound Images Using Transfer Learning and Decision Threshold Calibration

Kerllon F. Andrade
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0009-0005-4279-4368

Parama Bagchi
Department of Computer Science and Engineering
RCC Institute of Information Technology
Kolkata, India
https://orcid.org/0000-0002-9725-9582

Denis Leonov
Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies
Moscow Healthcare Department
Moscow, Russia
https://orcid.org/0000-0003-0916-6552

Wagner C. A. Pereira
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
https://orcid.org/0000-0001-5880-3242

Debotosh Bhattacharjee
Department of Computer Science & Engineering
Jadavpur University
Kolkata, India
https://orcid.org/0000-0002-1163-6413

José Francisco S. Costa-Júnior
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0000-0002-6526-2760

***Abstract*—Deep learning has shown substantial potential for breast ultrasound analysis. However, most studies rely on a fixed decision threshold, limiting the ability of classifiers to adapt to different operating objectives. In this work, we present a transfer learning approach based on the InceptionV3 architecture combined with decision-threshold calibration for the classification of benign and malignant breast lesions in ultrasound images. Using the publicly available Breast-Lesions-USG dataset, we systematically investigated multiple operating points and evaluated different training strategies, including data augmentation and fine-tuning. Data augmentation improved the performance observed on the independent test set and increased malignant-lesion detection, whereas fine-tuning failed to provide consistent benefits because of overfitting. In contrast, threshold calibration proved to be the most effective strategy, enabling a single trained model to be adapted to different classification objectives without architectural modifications. For the model trained with data augmentation, decision-threshold calibration increased malignant-lesion sensitivity from 68.75% (11/16) at the default threshold to 87.5% (14/16) at the balanced operating point, while maintaining a specificity of 87.5% (21/24). In addition, the Safe Benign configuration achieved a benign-class precision of 91.3% (21/23). These results indicate that decision-threshold calibration is a simple and computationally efficient strategy for adjusting classifier behavior according to different operating requirements. However, the findings should be interpreted as preliminary because of the limited test-set size and require validation in larger and independent clinical datasets before conclusions regarding clinical applicability can be established.**

Keywords—Breast ultrasound, deep learning, decision threshold calibration, transfer learning, clinical decision support

I. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer-related death among women worldwide [1]. In Brazil, breast cancer is also the most incident cancer among women, excluding non-melanoma skin cancer [2]. According to national estimates, approximately 78,610 new cases are expected annually during the 2026–2028 triennium, while 20,849 deaths from the disease were recorded in 2024, highlighting its significant public health impact. Early detection strategies have contributed substantially to improving survival rates among women diagnosed with breast cancer [3]. In this context, deep learning-based systems have emerged as promising tools for assisting in the analysis of medical images and the automatic classification of breast lesions as benign or malignant, with the potential to support diagnostic assessment and clinical decision-making.

The classification of breast lesions in both ultrasound and mammography is typically formulated as a binary classification problem (benign versus malignant), in which the decision threshold is the value used to convert the probability predicted by the model into a final class label [4]. The choice of this threshold directly affects the occurrence of false positives (benign lesions incorrectly classified as malignant) and false negatives (malignant lesions incorrectly classified as benign). Although several studies have reported promising results in breast lesion classification using deep learning techniques [5] - [7], the evaluation of these models remains largely restricted to performance metrics obtained at a single operating point, typically associated with the default decision threshold of 0.5, without discussing the impact of alternative threshold values. This approach limits the assessment of classifier behavior under different operating requirements, since different classification scenarios may require different trade-offs between sensitivity and specificity. Furthermore, depending on the intended application, the costs associated with false positives and false negatives are inherently asymmetric.

In potential clinical applications where the priority is to reduce unnecessary procedures, such as biopsies and additional diagnostic examinations, it may be desirable to operate the model at threshold values that increase confidence in benign classifications [8], [9]. Conversely, in screening and triage applications, the primary concern is the reduction of false negatives, favoring threshold values that increase

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

sensitivity for the detection of malignant lesions. In this direction, the study presented in [9] proposed a classifier capable of significantly reducing false negatives; however, its operating point was not adaptable to different operating objectives. Therefore, decision-threshold calibration emerges as a relevant strategy for tailoring model behavior to distinct classification requirements, potentially increasing its practical flexibility without requiring modifications to the neural network architecture.

The objective of this work is to develop and evaluate a binary classifier for breast lesions in ultrasound images using deep learning and transfer learning techniques to distinguish between benign and malignant lesions. In addition to model development, the study investigates different strategies for performance improvement, including data augmentation, fine-tuning, and decision-threshold calibration. As the first contribution, we analyze the impact of threshold selection on classifier performance, showing how the calibration of this parameter can modify the trade-off between sensitivity and specificity, as well as the occurrence of false positives and false negatives. Secondly, we adapt the InceptionV3 architecture using transfer learning and investigate whether a single model can be adapted to different classification objectives, such as maximizing the detection of malignant lesions or increasing confidence in benign classifications, without requiring modifications to its architecture or additional training.

II. Background and Related Work

In 2016, Google Research introduced the InceptionV3 architecture [10], an evolution of the Inception/GoogLeNet networks [11]. The architecture was designed to enhance visual feature extraction while maintaining computational efficiency. To this end, it incorporates optimized Inception modules, factorized convolutions, and regularization techniques that enable the analysis of visual patterns at multiple scales with a reduced number of parameters. This combination of high performance and computational efficiency has made InceptionV3 a widely adopted architecture for computer vision applications, particularly as a foundation for transfer learning.

Transfer learning is recognized as a powerful deep learning technique [12]. Its primary objective is to improve the predictive performance of a target domain by leveraging knowledge acquired from a source domain and task [13]. This approach exploits features learned in the deeper layers of neural networks, such as edges, shapes, textures, and other visual patterns, and transfers this knowledge to new classification tasks.

Given the success of convolutional neural networks (CNNs) and transfer learning across multiple domains, their application to medical imaging emerged naturally. The study reported in [14], which focused on the diagnosis of retinal OCT images, demonstrated that deep CNNs pre-trained with InceptionV3 can achieve performance comparable to that of specialists in a real clinical setting.

A systematic review of studies published between 2020 and 2024 on deep learning applications for breast cancer diagnosis [5] examined classification approaches across multiple imaging modalities, including mammography and ultrasound. The review found that a substantial proportion of studies relied on pre-trained deep convolutional architectures, underscoring the widespread adoption of transfer learning in breast cancer diagnosis.

The model proposed in [6] for breast cancer classification in mammograms employed a hybrid CNN-LSTM architecture and achieved high performance in lesion detection. Similarly, the study presented in [7] introduced an ensemble learning model for breast ultrasound imaging and reported strong classification results. However, both studies focused primarily on global performance metrics and did not investigate alternative operating points tailored to different clinical objectives. This limitation is particularly important in medical applications, where the consequences of false-positive and false-negative predictions may vary according to the intended clinical use.

To address this gap, the study presented in [8] employed a CNN-based mammography classifier operating at high-sensitivity operating points. Rather than relying on a fixed decision threshold, the authors selected thresholds that achieved target sensitivities of 99% and 97%, enabling the identification of examinations with a low likelihood of cancer. The present study builds upon the threshold calibration strategy proposed in [8]. Accordingly, a transfer learning-based classifier using the InceptionV3 architecture was adopted because of its strong performance and computational efficiency. The objective is to demonstrate how decision-threshold calibration can modify classifier behavior across different operating points, thereby enabling adaptation to distinct clinical scenarios.

III. Methods

A. Dataset

The images used in this study were obtained from the Breast-Lesions-USG dataset, curated by Pawłowska et al. [15] and made available through The Cancer Imaging Archive. The dataset comprises breast ultrasound examinations acquired from 256 adult female patients and includes imaging data together with associated clinical information, such as BI-RADS classification and histopathological diagnosis when available. The dataset used in this study is publicly available and contains anonymized images; therefore, no additional institutional ethics approval or informed consent was required. Each retained primary ultrasound image corresponded to a distinct patient. Because only one image per patient was included in the classification dataset, the stratified image-level split was also patient-independent, with no patient overlap among the training, validation, and test subsets. The original dataset contains 98 malignant lesions, 154 benign lesions, and 4 normal tissue images. For the present study, the normal tissue images were excluded, and only the primary ultrasound images were retained for binary classification, resulting in a final dataset of 252 images. Additionally, the dataset was divided into training, validation, and test subsets using an approximate ratio of 70%, 15%, and 15%, respectively, following the data partitioning strategy adopted by [4]. A stratified splitting strategy was adopted to preserve the original class distribution within each subset (see Table I).

B. Artificial Neural Network Architecture

To implement the transfer learning strategy, the InceptionV3 architecture was adopted as the backbone network. The convolutional base was kept frozen, and a dense layer with a Softmax activation function was added to perform the binary classification of breast lesions into benign and

malignant classes. Figure 1 presents the architecture employed for lesion classification. The transfer learning approach allows the model to leverage knowledge previously acquired from large-scale computer vision tasks, including the detection of low-level visual features such as edges, textures, and shapes [16], [17]. By keeping the convolutional layers frozen and introducing a new Softmax classification layer, the network can adapt its learned representations to the specific domain of breast ultrasound imaging [18], [19].

TABLE I. DISTRIBUTION OF THE DATASET AFTER STRATIFIED SPLITTING INTO TRAINING, VALIDATION, AND TEST SUBSETS

Subset	***Benign***	***Malignant***	***Total***
Training	107	68	175
Validation	23	14	37
Test	24	16	40
Total	154	98	252

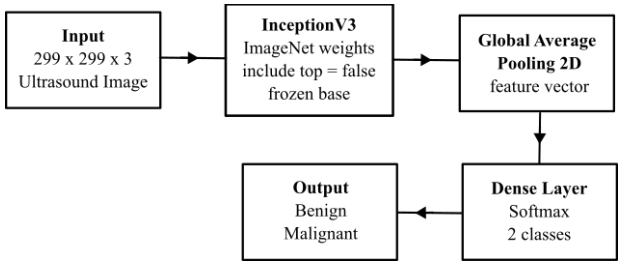


Fig. 1. Architecture used for the binary classification of breast lesions.

The InceptionV3 architecture was selected for two main reasons. First, it has demonstrated the ability to achieve satisfactory performance even when trained on relatively small datasets, which is a common limitation in medical imaging applications [20] - [22]. Second, the objective of this study was not only to develop a classification model but also to evaluate multiple decision-threshold calibration strategies and clinical operating points [23], [24]. Therefore, it was necessary to adopt an architecture with moderate computational requirements, allowing multiple iterative experiments to be conducted within acceptable training time and processing constraints. InceptionV3 meets these requirements while maintaining competitive classification performance.

C. Training Strategies

To evaluate the effects of different training strategies on model performance, the experimental procedure was conducted in multiple stages. In the first stage, the model was trained using only the original training set. The convolutional base of InceptionV3 remained frozen, while the newly added classification layer was trained to distinguish between benign and malignant breast lesions.

In the second stage, data augmentation was applied exclusively to the training set to increase image variability. The transformations included random rotations, horizontal and vertical shifts, and horizontal flipping. The validation and test sets were not subjected to any augmentation procedure (see Figure 2).

Unlike strategies that impose strict numerical balancing, the target sizes were defined independently for each class. This approach is consistent with class-dependent augmentation strategies in medical imaging, in which greater augmentation intensity is applied to underrepresented classes to reduce class imbalance [25].

TABLE II. DATASET DISTRIBUTION AFTER THE DATA AUGMENTATION STAGE

Subset	***Benign***	***Malignant***	***Total***
Training (augmentation)	200	300	500
Validation	23	14	37
Test	24	16	40
Total	247	330	577

The target sizes were predefined as part of the experimental design. A larger augmentation factor was assigned to the malignant class, which was originally underrepresented. At the same time, the expansion was limited to 200 benign images and 300 malignant images to avoid excessive growth of the training set based on multiple transformations of the same original images, thereby reducing the risk of redundancy among synthetic samples. These values were not derived from a numerical balancing formula but were adopted as class-specific experimental parameters.

$$N_{aug},c \;=\; N_{target},c \;-\; N_{orig},c \qquad (1)$$

where N_{orig},c denotes the original number of training images in class c, N_{target},c represents the predefined target size for that class, and N_{aug},c is the number of synthetic images generated through data augmentation. For the benign class, N_{orig},c was 107 and N_{target},c was defined as 200, resulting in the generation of 93 synthetic images. For the malignant class, N_{orig},c was 68 and N_{target},c was defined as 300, resulting in the generation of 232 synthetic images. Therefore, the adopted class-dependent augmentation strategy intentionally assigned a higher augmentation intensity to the malignant class instead of enforcing strict numerical balancing between the classes.

In the subsequent stage, fine-tuning was investigated by unfreezing subsets of the final layers of the InceptionV3 architecture, allowing part of the pretrained convolutional base to adapt to the breast ultrasound domain [14]. Two configurations were evaluated: aggressive fine-tuning, with 62 unfrozen layers, and conservative fine-tuning, with 21 unfrozen layers. Both fine-tuning configurations used a learning rate of 1×10^{-6}, differing only in the number of unfrozen layers.

As the final experimental stage, a systematic decision-threshold calibration procedure was performed. In binary classification, a default threshold of 0.50 is commonly used to convert predicted probabilities into class labels. However, this threshold may not provide the most appropriate trade-off between sensitivity and specificity when the consequences of false-positive and false-negative predictions are asymmetric. The calibration procedure was inspired by the high-sensitivity operating-point strategy described in [8] for computer-aided mammographic screening. In the present study, thresholds ranging from 0.000 to 1.000 were evaluated in increments of 0.001 using the predicted class probabilities.

All decision thresholds were selected exclusively using the validation set. The test set was not used during threshold selection and was reserved solely for the final evaluation of the predefined operating points. For the sensitivity-oriented strategy, the predicted probabilities for the malignant class were analyzed. For each predefined target sensitivity, the

algorithm selected the highest threshold capable of achieving at least the required sensitivity on the validation set, thereby avoiding unnecessary reductions in the decision threshold. Target sensitivity levels of 85%, 90%, 92%, 95%, 97%, and 99% were evaluated.

A second independent strategy, referred to as Safe Benign, was also implemented. In this strategy, threshold selection was based on the predicted probability of the benign class, with the objective of identifying operating points associated with predefined benign-class precision levels. Target precision levels of 90%, 95%, 97%, 99%, and 100% were evaluated using only the validation set. Once selected, the thresholds were fixed and applied to the independent test set without further adjustment. The two calibration strategies were designed to evaluate different operating conditions using the same trained model. The sensitivity-oriented strategy prioritized malignant lesion detection, whereas the Safe Benign strategy prioritized the precision of benign classifications. No architectural modification or additional model training was performed after threshold selection. To quantify the uncertainty associated with the limited test-set size, 95% confidence intervals for proportion-based metrics were calculated using the Wilson score method.

D. Implementation and Reproducibility Details

All experiments used InceptionV3 pretrained on ImageNet with the original classification head removed, input images resized to 299 × 299 pixels and processed using the standard InceptionV3 preprocessing function. The convolutional backbone was initially frozen, and the classifier comprised global average pooling followed by a two-unit Softmax layer. Training used the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss, a batch size of 32, and a maximum of 100 epochs. Early stopping monitored validation loss with a patience of 15 epochs and restored the best weights, while model checkpointing retained the model with the lowest validation loss. Class weights were applied during training, test inference was performed with a batch size of 1, and a random seed of 42 was used to improve the reproducibility of the experiments.

IV. RESULTS

The baseline model exhibited substantial limitations in the detection of malignant lesions, primarily due to the high number of false-negative cases observed in the test set. Table IV presents the baseline performance obtained with the InceptionV3 architecture, establishing a reference benchmark for comparison with the threshold-based operating modes proposed in this study.

TABLE III. BASELINE MODEL PERFORMANCE USING THE INCEPTIONV3 ARCHITECTURE

Class	***Precision***	***Recall***	***F1-Score***	***Support***
Benign	0.77	0.96	0.85	24
Malignant	0.90	0.56	0.69	16
Accuracy	-	-	0.80	40

The results highlight important limitations of the baseline model for clinical applications focused on malignant lesion detection. The low recall achieved for the malignant class, indicating a substantial number of false-negative cases, resulted in reduced diagnostic sensitivity for identifying cancerous lesions. This behavior becomes more evident when the confusion matrix (Fig. 2) is examined. The model correctly classified 23 benign cases and 9 malignant cases, while seven malignant lesions were misclassified as benign. Considering malignant lesions as the positive class, this performance corresponds to a true positive rate (TPR) of 56.25%, a false positive rate (FPR) of 4.17%, and a specificity of 95.83%. Collectively, these results indicate high specificity but insufficient sensitivity for malignant lesion detection.

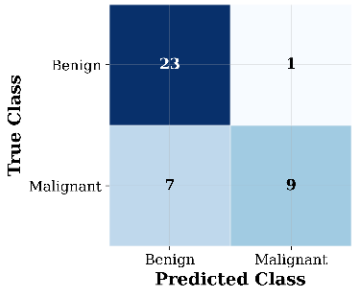


Fig. 2. Confusion matrix of the baseline model.

To address this limitation, a data augmentation strategy was applied. Compared with the baseline model, malignant-class recall increased and the number of false-negative classifications decreased. Table V also shows higher F1-scores for both classes, indicating improved performance on the evaluated test set and a more balanced relationship between precision and recall.

TABLE IV. TEST-SET PERFORMANCE OF THE MODEL TRAINED WITH DATA AUGMENTATION

Class	***Precision***	***Recall***	***F1-Score***	***Support***
Benign	0.82	0.96	0.88	24
Malignant	0.92	0.69	0.79	16
Accuracy	-	-	0.85	40

The confusion matrix (Fig. 3) shows a reduction in the number of malignant lesions incorrectly classified as benign, decreasing from seven to five cases compared with the baseline model. At the same time, the model maintained a high ability to identify benign lesions, producing only one false-positive classification. Despite the observed improvement, a considerable number of malignant cases were still misclassified, indicating that the use of data augmentation alone was insufficient to achieve a satisfactory reduction in false negatives.

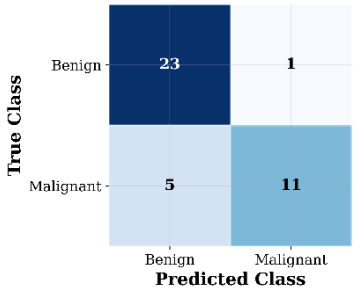


Fig. 3. Confusion matrix corresponding to the model trained with data augmentation.

Subsequently, a fine-tuning strategy was explored by unfreezing additional layers of the InceptionV3 architecture. Two approaches were evaluated: an aggressive strategy, in which 62 layers were unfrozen (approximately 11 million trainable parameters), and a more conservative strategy, in which only 21 layers were unfrozen (approximately 2.38

million trainable parameters). The aggressive approach led to severe overfitting, as evidenced by deteriorating validation metrics despite improved training performance. Although the conservative strategy partially mitigated overfitting, it still failed to achieve consistent performance gains compared with the model trained using data augmentation alone.

With 21 layers of the InceptionV3 architecture unfrozen, the default operating point (threshold = 0.500) achieved a TPR of 75.0% and an FPR of 20.8%, as illustrated by the confusion matrix (Fig. 4). The model correctly classified 19 benign cases and 12 malignant cases while producing five false positives and four false negatives, demonstrating inferior performance compared with the model trained using data augmentation alone. These results indicate that increasing the number of trainable layers did not improve performance on the evaluated test set and was associated with a less favorable balance between sensitivity and specificity.

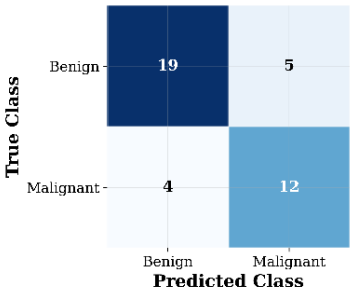


Fig. 4. Confusion matrix of the model trained using the conservative fine-tuning strategy with 21 unfrozen layers of the InceptionV3 architecture.

Therefore, to develop a classifier adaptable to different clinical requirements, the model trained with data augmentation was selected for the decision-threshold calibration stage. This approach had a substantial impact on classifier behavior, particularly regarding the trade-off between sensitivity for malignant lesion detection and specificity for benign lesion identification. By adjusting the operating point, different operating regimes could be achieved, ranging from balanced configurations to highly sensitive settings specifically designed to minimize false-negative cases.

A. Sensitivity-oriented operating points

As shown in Fig. 5, progressively lowering the decision threshold consistently improved the model's ability to detect malignant lesions. However, this improvement was accompanied by reduced specificity for benign lesions and an increase in false-positive predictions.

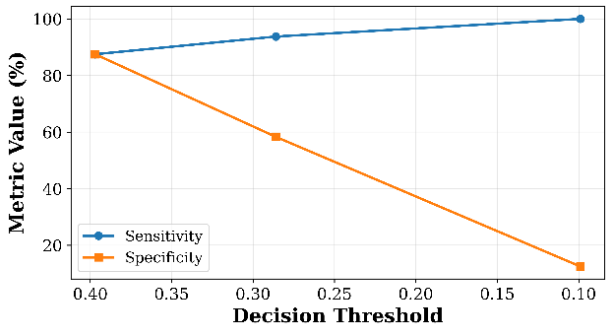


Fig. 5. Relationship between sensitivity and specificity as a function of the decision threshold.

The results highlight the trade-off between malignant lesion detection and the preservation of diagnostic specificity. As higher target sensitivities were adopted, an additional reduction in the false-negative rate was observed. For operating points corresponding to target sensitivities of 90% and 92%, the selected threshold was 0.286, yielding a malignant recall of 93.75% while reducing the number of false negatives to a single case. However, this increase in sensitivity was accompanied by a decrease in specificity and an increase in false positives to ten cases.

The balanced operating point selected by the algorithm (threshold = 0.397) achieved 87.5% sensitivity and 87.5% specificity. As shown by the confusion matrix (Fig. 6), the classifier correctly identified 14 malignant lesions while producing only three false positives and two false negatives. This operating point provided a balanced trade-off between sensitivity and specificity in the evaluated test set.

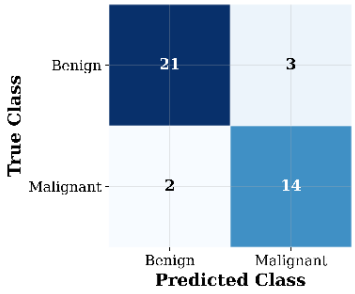


Fig. 6. Confusion matrix of the balanced operating point selected for malignant lesion classification (threshold = 0.397).

B. Safe Benign Strategy

Different operating points were identified according to the target precision established for the benign class, allowing the trade-off between benign-class precision and the coverage of cases classified as benign to be analyzed. Increasing the threshold improved benign-class precision but reduced the number of cases assigned to this class. For a target precision of 90%, the algorithm selected a threshold of 0.603, yielding a benign-class precision of 91.3%, equivalent to the negative predictive value when malignancy is considered the positive class, a benign recall of 87.5%, and an overall accuracy of 87.5%. At this operating point, 23 cases were classified as benign, of which 21 were correctly classified and two malignant lesions were misclassified as benign.

When higher confidence levels were adopted, a substantial reduction in safe-benign coverage was observed. For operating points corresponding to target confidence levels of 95%, 97%, 99%, and 100%, the selected threshold was 0.901, yielding a benign precision of 100%, but reducing benign recall to 12.5% and classifying only three cases as safe benign.

As shown in Figs. 7 and 8, increasing the decision threshold produced an inverse relationship between benign-class precision and the number of cases classified as benign. Thus, the evaluated operating points ranged from broader benign-case coverage with lower precision to restricted coverage with higher precision.

Among the operating points evaluated for the Safe Benign strategy, the threshold of 0.603 provided the highest coverage of cases classified as benign. Under this configuration, the model achieved a benign-class precision of 91.3%, a benign recall of 87.5%, and an overall accuracy of 87.5%.

The operating point with the highest safe-benign coverage (threshold = 0.603) is illustrated by the confusion matrix (Fig.

9). Under this configuration, the model classified 23 examinations as safe benign, correctly identifying 21 while producing only two false-benign classifications and successfully detecting 14 of the 16 malignant lesions. Collectively, these results indicate the potential of the selected operating point for identifying a subset of cases classified as benign, although further validation is required before considering its use in clinical rule-out applications.

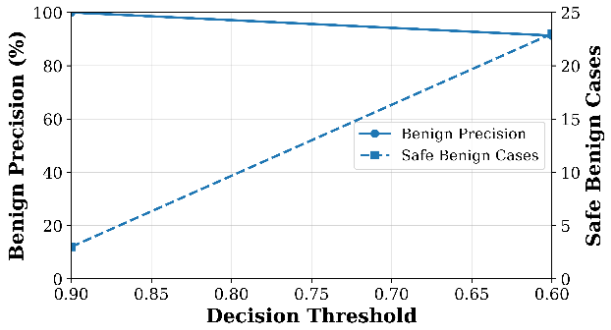


Fig. 7. Relationship between benign-class precision and safe-benign coverage as a function of the decision threshold.

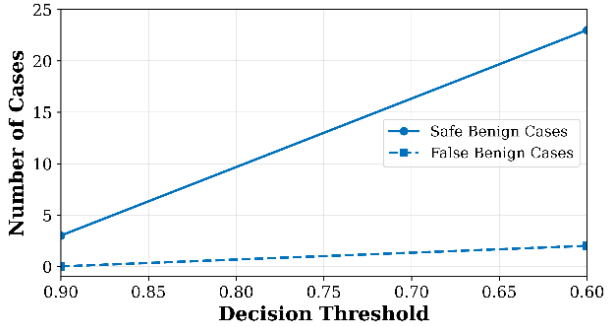


Fig. 8. Relationship between the number of safe benign cases and the occurrence of false-benign classifications.

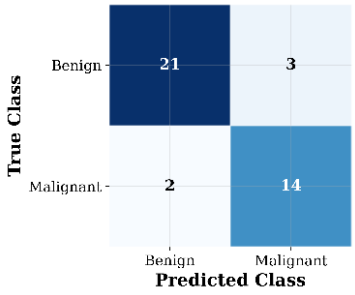


Fig. 9. Confusion matrix corresponding to the operating point with the highest safe-benign coverage (threshold = 0.603).

C. Global Comparison of Strategies

To consolidate the findings across the experimental stages, the evaluated strategies were compared in Figs. 10, 11 and 12. Data augmentation increased malignant-class recall relative to the baseline model, whereas fine-tuning did not provide consistent performance gains and showed signs of overfitting. The Balanced threshold of 0.397 for the malignant class and the Safe Benign threshold of 0.603 for the benign class produced the same test-set classifications because the two class probabilities are complementary. Thus, these operating points represent the same decision boundary expressed according to different target-class objectives.

D. Confidence Interval Analysis

Given the limited size of the independent test set, 95% confidence intervals were calculated using the Wilson score method for the main proportion-based performance metrics. The analysis included overall accuracy, malignant-class sensitivity, specificity, malignant-class precision, and benign-class precision. Malignant lesions were considered the positive class throughout this analysis. Benign-class precision is therefore equivalent to the negative predictive value. The confidence intervals were relatively wide, particularly for malignant-class sensitivity and precision, because these estimates were based on only 16 malignant cases. For example, although the balanced operating point achieved a malignant-class sensitivity of 87.5%, its 95% confidence interval ranged from 64.0% to 96.5%. Similarly, the benign-class precision of 91.3% obtained at the Safe Benign operating point was associated with a 95% confidence interval of 73.2% to 97.6%. These intervals indicate substantial uncertainty around the point estimates and reinforce the need for cautious interpretation of the observed differences among the evaluated configurations.

V. Discussion

The results obtained in this study indicate that conventional training strategies, such as the direct use of a pre-trained model and the isolated application of data augmentation, although beneficial for improving overall performance, still present important limitations when applied to the classification of breast lesions in ultrasound images derived from relatively small datasets. In particular, the baseline model exhibited a high number of false-negative cases, a clinically undesirable outcome in applications focused on malignancy detection. Data augmentation improved the performance observed on the test set and increased malignant-class recall, indicating a greater ability to identify cancerous lesions. However, the findings also showed that strategies based solely on data augmentation are insufficient. The fine-tuning experiments did not produce consistent performance gains, suggesting that the limited dataset size may restrict the effective adaptation of deep architectures and increase susceptibility to overfitting. In contrast, decision-threshold calibration emerged as the main practical contribution of this work.

The operating-point analysis showed that classifier behavior can be adapted to different performance objectives, ranging from balanced configurations to high-sensitivity operating points specifically designed to minimize false-negative cases. This finding reinforces the importance of defining application-specific operating points in artificial intelligence systems for medical diagnosis, particularly in scenarios where the costs associated with classification errors are asymmetric. Despite the promising results, this study has important limitations. The dataset is relatively small and was obtained from a single public source, with no external validation performed on independent clinical datasets. Therefore, future work will include external validation using independent clinical datasets and repeated stratified cross-validation, for example, 5-fold cross-validation repeated ten times, to evaluate the robustness and generalizability of the proposed approach across different clinical settings.

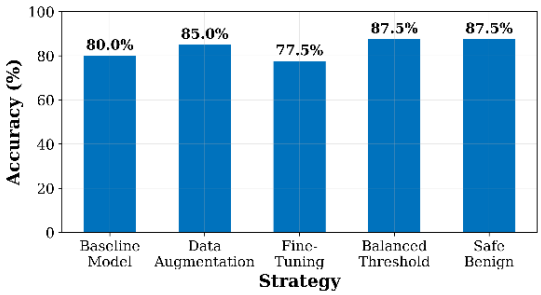


Fig. 10. Comparison of accuracy across the evaluated models.

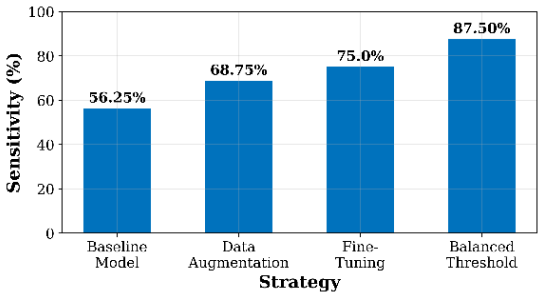


Fig. 11. Comparison of sensitivity for malignant lesion detection across the evaluated models.

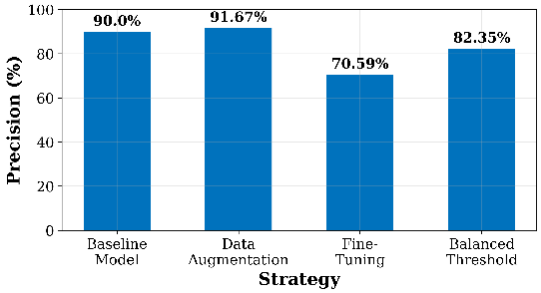


Fig. 12. Comparison of malignant lesion classification precision across the evaluated models.

A previous mammography study reported a 17% reduction in workload while missing 1% of cancer-containing examinations [26]. Because the studies in [26], [27] evaluated mammography rather than breast ultrasound, their results provide methodological context for threshold-based triage but are not directly comparable with those of the present study. In the present study, the most conservative operating point classified 7.5% of examinations as benign, with no false-benign cases observed in this test set. Although both studies indicate that threshold calibration can control the trade-off between coverage and classification errors, the operating points evaluated in [26] were predefined, whereas those in the present study were established based on explicit confidence targets for the target class. This strategy may allow adaptation to scenarios focused either on the identification of benign lesions or on maximizing sensitivity for malignant lesion detection.

In [27], the model classified 71.5% of examinations as normal using a predefined threshold (AI Score ≤ 7), achieving a negative predictive value of 99.98% while excluding only 2 of the 76 cancer cases in the dataset. At the operating point corresponding to 90% confidence for the benign class, the proposed model classified 57.5% of examinations as benign, with two false-benign cases. At more conservative operating points, no false-benign cases were observed, although coverage decreased to 7.5% of examinations. These findings highlight the importance of evaluating multiple thresholds, an aspect not explored in [27], which reported results based on a single decision threshold.

In addition to the identification of lesions classified as benign, threshold calibration enabled adjustment of classifier behavior for high-sensitivity scenarios targeting malignant lesion detection. For the model trained with data augmentation, lowering the malignant-class decision threshold increased sensitivity from 68.75% (11/16) at the default threshold to 87.5% (14/16) at the balanced operating point. At the threshold of 0.10, sensitivity reached 100% (16/16), but specificity decreased to 12.5% (3/24), resulting in 21 false-positive classifications. These findings show, in this test set, that a single model can be configured for different classification objectives solely through threshold calibration. This interpretation is consistent with the observations reported by [28], who demonstrated substantial variations in sensitivity and specificity as a function of the selected operating point.

More recently, a study evaluating five transfer-learning CNNs, including InceptionV3, used a public dataset of 526 breast ultrasound images and an institutional dataset of 392 images. At a 2% malignancy threshold, three models achieved 98.3% sensitivity, with specificity ranging from 15.3% to 18.6% [29]. These findings support the relevance of threshold adjustment in breast ultrasound, although direct comparison with the present study is limited by differences in dataset composition and experimental design.

Taken together, these results reinforce that the choice of decision threshold directly influences the predictive behavior of a classifier and may be as important as the choice of model architecture itself.

From the perspective of possible future integration into the clinical workflow, decision-threshold adjustment could allow the same classifier to serve multiple complementary purposes. The Safe Benign configuration showed potential for identifying a subset of lesions classified as benign, whereas sensitivity-oriented operating points increased malignant lesion detection in the evaluated dataset. Thus, threshold selection may influence not only statistical performance but also, after additional validation, patient prioritization, resource allocation, and clinical risk management. However, the current results are not sufficient to establish reductions in biopsies, changes in clinical management, or diagnostic safety in independent populations.

Although the proposed threshold-adjustment strategy yielded promising results, further validation using larger and independent datasets is required to confirm its robustness and clinical applicability. Future work will investigate statistically optimized threshold-selection strategies aimed at increasing malignant-class sensitivity while simultaneously limiting the increase in false positives, without assuming in advance that this balance will result in greater clinical safety or a lower diagnostic burden.

VI. Conclusion

This study presented a deep learning-based approach for classifying breast lesions in ultrasound images using the InceptionV3 architecture with transfer learning. Data augmentation improved the performance observed on the independent test set and increased sensitivity for malignant lesion detection, whereas fine-tuning failed to provide consistent performance gains because of overfitting and the

limited dataset size. In contrast, decision-threshold calibration enabled a single trained model to be adapted to different classification objectives, representing the main practical contribution of this work.

The operating-point analysis showed that the balanced threshold provided an effective trade-off between sensitivity and specificity, whereas the Safe Benign strategy achieved 91.3% benign-class precision and correctly identified 21 of the 23 cases classified as benign in the test set. These findings indicate that threshold calibration is a computationally efficient strategy for adjusting classifier behavior according to different operating requirements. However, the results should be interpreted as preliminary because of the limited test-set size and the absence of external validation. Larger and independent clinical datasets are required before conclusions regarding clinical applicability can be established.

To facilitate reproducibility and future research, the implementation developed in this study will be made publicly available through a GitHub repository upon publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] World Health Organization (WHO), “Breast cancer” Apr. 2026. Accessed: Jun. 07, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Instituto Nacional de Câncer (INCA), “Câncer de mama” Rio de Janeiro, Dec. 2025. Accessed: Jun. 07, 2026. [Online]. Available: https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/mama

[3] Instituto Nacional de Câncer (INCA), “Controle do câncer de mama no Brasil: dados e números 2024” Rio de Janeiro, 2024. Accessed: Jun. 07, 2026. [Online]. Available: https://ninho.inca.gov.br/jspui/handle/123456789/17002

[4] A. Géron, Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes, 1st ed. Rio de Janeiro: Alta Books, 2021.

[5] S. Gao *et al.*, “Application of deep learning technology in breast cancer: a systematic review of segmentation, detection, and classification approaches,” *BioMed Eng OnLine*, vol. 25, no. 1, p. 19, Jan. 2026, doi: 10.1186/s12938-025-01502-5.

[6] M. Kaddes, Y. M. Ayid, A. M. Elshewey, and Y. Fouad, “Breast cancer classification based on hybrid CNN with LSTM model,” Sci Rep, vol. 15, Feb. 2025, doi: 10.1038/s41598-025-88459-6.

[7] A. E. Fawzy, M. E. Almandouh, M. Herajy, and M. Eisa, “Ensemble-based high-performance deep learning models for medical image retrieval in breast cancer detection,” Sci Rep, vol. 16, Mar. 2026, doi: 10.1038/s41598-026-38218-y.

[8] S. Pedemonte *et al.*, “A Semiautonomous Deep Learning System to Reduce False Positives in Screening Mammography,” Radiol Artif Intell., vol. 6, May 2024, doi: 10.1148/ryai.230033.

[9] S. M. McKinney *et al.*, “Addendum: International evaluation of an AI system for breast cancer screening,” Nature, vol. 586, Oct. 2020, doi: 10.1038/s41586-020-2679-9.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[11] C. Szegedy *et al.*, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[12] A. Kapoor, A. Gulli, and S. Pal, Deep Learning with TensorFlow and Keras: Build and deploy supervised, unsupervised, deep, and reinforcement learning models, 3rd ed. Birmingham: Packt Publishing, 2022.

[13] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[14] D. S. Kermany *et al.*, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[15] A. Pawłowska *et al.*, “A Curated Benchmark Dataset for Ultrasound Based Breast Lesion Analysis (Breast-Lesions-USG).” The Cancer Imaging Archive, 2024. doi: 10.7937/9WKK-Q141.

[16] N. Abou Baker, N. Zengeler, and U. Handmann, “A Transfer Learning Evaluation of Deep Neural Networks for Image Classification,” *MAKE*, vol. 4, pp. 22–41, Jan. 2022, doi: 10.3390/make4010002.

[17] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” *BMC Med Imaging*, vol. 22, Dec. 2022, doi: 10.1186/s12880-022-00793-7.

[18] A. Hijab, M. A. Rushdi, M. M. Gomaa, and A. Eldeib, “Breast Cancer Classification in Ultrasound Images using Transfer Learning,” in *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*, Tripoli, Lebanon: IEEE, Oct. 2019, pp. 1–4. doi: 10.1109/ICABME47164.2019.8940291.

[19] G. Ayana, K. Dese, and S. Choe, “Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging,” *Cancers*, vol. 13, Feb. 2021, doi: 10.3390/cancers13040738.

[20] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, “Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning,” *Ultrasound Med Biol*, vol. 46, pp. 1119–1132, May 2020, doi: 10.1016/j.ultrasmedbio.2020.01.001.

[21] G. Ayana, J. Park, J.-W. Jeong, and S. Choe, “A Novel Multistage Transfer Learning for Ultrasound Breast Cancer Image Classification,” *Diagnostics*, vol. 12, Jan. 2022, doi: 10.3390/diagnostics12010135.

[22] R. Disci, F. Gurcan, and A. Soylu, “Advanced Brain Tumor Classification in MR Images Using Transfer Learning and Pre-Trained Deep CNN Models,” *Cancers*, vol. 17, Jan. 2025, doi: 10.3390/cancers17010121.

[23] C. G. Walsh, K. Sharman, and G. Hripcsak, “Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk,” *J Biomed Inform*, vol. 76, pp. 9–18, Dec. 2017, doi: 10.1016/j.jbi.2017.10.008.

[24] H. Özcan, “Threshold and Calibration Effects in Medical Machine Learning,” in *2025 8th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)*, Nanjing, China: IEEE, Dec. 2025, pp. 1–6. doi: 10.1109/ACAI68217.2025.11406486.

[25] R.E.D. Guerrero, L. Carvalho, T. Bocklitz, J. Popp, and J. L. Oliveira, “A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images,” *J Digit Imaging. Inform. med.*, vol. 37, pp. 1767–1782, Mar. 2024, doi: 10.1007/s10278-024-01018-9.

[26] A. Rodriguez-Ruiz *et al.*, “Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study,” *Eur Radiol*, vol. 29, pp. 4825–4832, Sep. 2019, doi: 10.1007/s00330-019-06186-9.

[27] J. L. Raya-Povedano, S. Romero-Martín, E. Elías-Cabot, A. Gubern-Mérida, A. Rodríguez-Ruiz, and M. Álvarez-Benito, “AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation,” *Radiology*, vol. 300, Jul. 2021, doi: 10.1148/radiol.2021203555.

[28] M. Salim *et al.*, “External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms,” *JAMA Oncol*, vol. 6, no. 10, p. 1581, Oct. 2020, doi: 10.1001/jamaoncol.2020.3321.

[29] P. De Marco, V. Ricciardi, M. Montesano, E. Cassano, and D. Origgi, “Transfer learning classification of suspicious lesions on breast ultrasound: is there room to avoid biopsies of benign lesions?,” *Eur Radiol Exp*, vol. 8, Oct. 2024, doi: 10.1186/s41747-024-00480-y.

Deep Learning-Based Monaural Speech Enhancement on the VoiceBank-DEMAND Dataset

Pedro Henrique Santos Moreira[1], Matheus Lima Reis[1], Arthur de Sousa Vianna[1], Sarah Negreiros C. Leite[2]
[1]*Undergraduate Program in Computer Engineering, Aeronautics Institute of Technology (ITA), São José dos Campos, Brazil*
[2]*Division of Electronics Engineering, Aeronautics Institute of Technology (ITA), São José dos Campos, Brazil*

***Abstract*—Environmental noise and reverberation are major challenges in voice communication systems, degrading both speech quality and intelligibility. Deep learning-based speech enhancement methods have emerged as the state of the art by learning complex mappings between noisy and clean speech directly from large paired datasets. This work investigates and compares two time-frequency domain architectures for monaural speech enhancement on the VoiceBank-DEMAND dataset: a baseline 2D Convolutional U-Net and a Convolutional Recurrent Network (CRN) integrating a bottleneck Bidirectional Long Short-Term Memory (LSTM). Both models estimate an Ideal Ratio Mask from the Short-Time Fourier Transform magnitude spectrum and are evaluated using Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI) and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR). Across ten independently trained models per architecture, the CRN attains a higher mean PESQ (2.89 against 2.81) and STOI (0.95 against 0.94) than the U-Net, both statistically significant, at the cost of 7.5 times more parameters (14.5M against 1.94M), with no significant difference in SI-SDR. To interpret this gain, a magnitude oracle experiment is introduced, decomposing the overall performance ceiling into a recoverable magnitude-estimation margin and a structural phase-preservation cost. This decomposition shows that the recurrent bottleneck closes only a small fraction of the recoverable margin, indicating that long-range temporal modeling is not the main missing component, and that the dominant remaining limitation stems from noisy-phase preservation under real-valued masking, particularly under low signal-to-noise ratio and highly non-stationary noise conditions.**

***Index Terms*—Speech enhancement, Convolutional Neural Network, Recurrent Neural Network, Time-Frequency Analysis, Ideal Ratio Mask, Speech Intelligibility.**

I. Introduction

Voice communication underpins an increasing number of everyday applications, such as video calls, virtual assistants, automatic speech recognition (ASR) systems and real-time transcription services. Among these, hearing assistive devices represent a particularly demanding application from a biomedical engineering standpoint: for users with hearing loss, environmental noise and reverberation do not merely reduce comfort, but can severely compromise speech intelligibility and, consequently, social communication and quality of life. This impact is amplified for listeners with sensorineural hearing loss, who already exhibit reduced frequency and temporal resolution compared to normal-hearing listeners [1]–[4]. Speech signals captured by hearing aid and other biomedical audio device microphones are frequently corrupted by noise sources such as traffic, competing speakers, ventilation systems and background music, as well as reverberation, impairing both perceptual quality and the performance of downstream automated systems such as ASR-based communication aids [5]. As a result, robust speech enhancement techniques have become a key enabling technology not only for general voice communication but specifically for assistive hearing technology and other embedded biomedical audio processing systems.

In this context, classical speech enhancement techniques offer limited performance under real-world conditions [6]. Consequently, deep learning-based approaches have emerged as the state of the art, as they are capable of learning complex mappings between noisy and clean speech directly from large paired datasets, without relying on restrictive statistical assumptions about the noise [6]. A wide range of neural network architectures has been explored for this task, including Convolutional Neural Networks (CNNs) [7], Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [8]–[10], Generative Adversarial Networks (GANs) [11], [12], hybrid architectures [13], [14] and waveform domain models operating directly on raw audio signals [15], [16].

More recently, two additional model families have emerged as competitive alternatives to purpose-built discriminative networks. Diffusion-based generative models formulate speech enhancement as a stochastic denoising process, learning to reverse a gradual corruption of clean speech into a noisy or Gaussian-perturbed state, and have demonstrated strong out-of-domain generalization compared to discriminative baselines [17], [18]. In parallel, foundation-model-based approaches repurpose large self-supervised or weakly supervised speech representations — such as WavLM, HuBERT, and Whisper — as frozen or fine-tuned backbones for mask estimation, with recent work showing that careful architectural adaptation and perceptual pre-processing allow these models to close much of the historical performance gap with specialized speech enhancement architectures while requiring substantially less labeled training data [19].

This work investigates a deep learning-based approach for monaural speech enhancement using the publicly available VoiceBank-DEMAND dataset [8], a single-microphone benchmark whose monaural setting reflects a common hardware constraint in hearing aids and other compact biomedical listening devices. The objective is to design, train and critically evaluate a model capable of suppressing environmental noise while preserving speech quality and speaker

naturalness. Performance is evaluated using objective metrics widely adopted in the speech enhancement literature, namely Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), enabling direct comparison with related work [6]–[8], [11], [13], [16].

The main contributions of this study are as follows:

- A controlled, head-to-head comparison between a 2D convolutional U-Net and a Convolutional Recurrent Network (CRN) with a bottleneck BiLSTM for Ideal Ratio Mask-based monaural speech enhancement, trained and evaluated under an identical data pipeline, preprocessing, and training protocol on VoiceBank-DEMAND;
- A magnitude oracle analysis that decomposes the overall PESQ performance ceiling of the real-valued masking formulation into a recoverable magnitude-estimation margin and a structural, phase-preservation cost, providing an interpretable upper bound against which architectural gains can be measured;
- An empirical finding, enabled by this decomposition, that the recurrent bottleneck closes only a small fraction of the recoverable margin, indicating that long-range temporal modeling is unlikely to be the dominant missing component in this magnitude-masking setting, despite the hypothesis motivated by the recurrent design adopted in DCCRN [13].

The remainder of this paper is organized as follows. Section II presents the methodology, including the dataset and exploratory data analysis, the preprocessing pipeline, the U-Net and CRN model architectures, the training setup, the objective evaluation metrics, and the magnitude oracle analysis used to decompose the performance ceiling of the masking formulation. Section III reports the experimental results, comparing the proposed architectures against each other and against reference values from the literature, and discusses these results in light of the oracle decomposition. Finally, Section IV concludes the paper by summarizing the main findings, discussing the structural limitations identified, and outlining directions for future work.

II. Methodology

A. Dataset

This work used the VoiceBank-DEMAND dataset, a widely used benchmark for supervised monaural speech enhancement introduced by Valentini-Botinhao *et al.* (2016) [8]. The dataset combines clean speech recordings from the Voice Bank Corpus [20] with environmental noises from the DEMAND database [21]. Each sample consists of a pair of recordings, one noisy and one clean, enabling supervised training and objective evaluation. The original recordings are provided in WAV format with a sampling rate of 48 kHz.

The dataset contains 12,396 paired utterances, divided into 11,572 training samples from 28 speakers (14 male and 14 female) and 824 test samples from two previously unseen speakers (one male and one female). Noisy speech signals were generated by mixing clean recordings with environmental noises at nominal signal-to-noise ratios (SNRs) of $\{0, 5, 10, 15\}$ dB for training and $\{2.5, 7.5, 12.5, 17.5\}$ dB for testing [8]. It is publicly available at https://datashare.ed.ac.uk/handle/10283/2791.

An exploratory data analysis (EDA) was performed prior to model development. The analysis confirmed that all noisy-clean pairs were complete and correctly aligned, with no missing or corrupted files. Fig. 1 presents the distribution of measured SNR values for randomly sampled utterances from the training and test sets. The observed distributions closely match the nominal SNR levels defined during dataset construction, confirming the expected separation between training and testing conditions. An offset of approximately 1.5 dB was observed between the nominal and measured values. This discrepancy is expected, since the nominal SNRs were defined based on the speech-active portions of each utterance, while the SNR measurements performed in this work consider the entire audio signal, including silent segments. Nevertheless, the measured distributions remain consistent with the intended training and testing SNR levels.

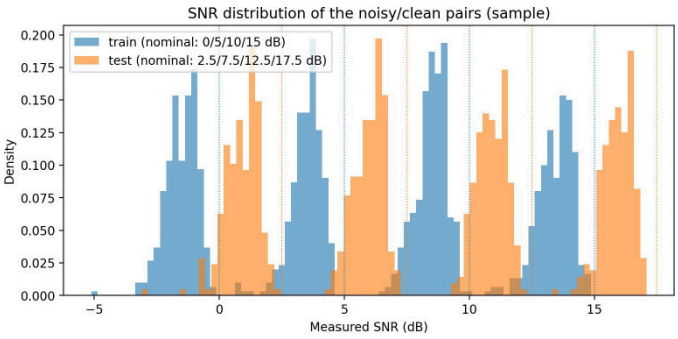


Fig. 1. Distribution of measured SNR values for randomly sampled utterances from the training and test sets.

B. Preprocessing

The original utterances were downsampled to 16 kHz for literature compatibility [11], [13], [16], allowing direct metrics comparison (PESQ wideband is defined at 16 kHz) and reducing computational cost without discarding perceptually relevant information, since the majority of human speech frequency content lies below 8 kHz. The last two speakers in alphabetical order among the 28 available training speakers (p286 and p287) were separated from the training set to compose the validation set, yielding 872 samples for validation and 26 speakers for the actual training subset, resulting in 10,700 training samples. This deterministic criterion, rather than a random draw, ensures full reproducibility of the speaker-independent split. As in the official test partition, validation speakers remain entirely unseen during training, preventing voice identity leakage and ensuring that the validation loss measures the same generalization property as the official test evaluation.

The audio signals were then segmented into fixed, non-overlapping windows of 32,640 samples (2.04 s). This specific

length was chosen to yield exactly 256 temporal frames after the Short-Time Fourier Transform (STFT) with hop size 128 (8 ms) and window length of 512 (32 ms), ensuring exact divisibility by 16, which is mandatory for the symmetric downsampling and upsampling operations of the U-Net architecture, which will be discussed further. Signals shorter than this window length were zero-padded to maintain the fixed input size. Besides, amplitude normalization was applied in each segment, dividing it by the standard deviation of its corresponding noisy audio. The identical scaling factor was applied to the clean speech target, strictly preserving the fundamental additive noise formulation ($x = y + n$, where x stands for the noisy audio, y, for the clean audio and n, for the noise).

C. Model Architecture

This work evaluates two time-frequency domain architectures for monaural speech enhancement: a 2D Convolutional U-Net baseline and a CRN. Both networks process the STFT magnitude spectrum to estimate an Ideal Ratio Mask $M \in [0, 1]^{256\times256}$. Given a 16 kHz sampling rate and 32,640-sample windows, the STFT is computed using a 512-point Hann window and a 128-sample hop size. Truncating the Nyquist component yields a square input matrix $|X| \in \mathbb{R}^{256\times256}$. The input of the models consists of $\log(1 + |X|)$ to compress the dynamic range and to be gradient-friendly. The enhanced complex spectrum is reconstructed via $\hat{Y} = M \odot X$, preserving the noisy phase before applying the inverse STFT (iSTFT). This magnitude-masking approach avoids the training instability and high computational overhead of generative adversarial alternatives like SEGAN [11].

The baseline U-Net consists of a symmetric encoder-decoder structure with channel concatenation via skip connections, as shown in Fig. 2. Symmetrical skip connections are preferred over the standard R-CED [7] topology to prevent the severe loss of fine-grained spectral lines during upsampling. The encoder comprises four 2D convolutional layers with channels scaling as $1 \to 16 \to 32 \to 64 \to 128 \to 256$, utilizing a 3×3 kernel and a padding of 1. Each layer is coupled with Batch Normalization and a LeakyReLU activation ($\alpha = 0.2$), compressing the feature map to a $256 \times 16 \times 16$ bottleneck due to 2×2 max pooling layer at the end. The decoder mirrors this topology using four 2D transposed convolutional layers. The final layer applies a 1×1 convolution followed by a Sigmoid activation to bound the mask values. This architecture results in 1.94M trainable parameters.

The proposed CRN directly modifies this topology by replacing the static 2D convolutional bottleneck with a recurrent sequential block to capture long-term temporal dependencies across distant STFT frames (Fig. 3). While the baseline treats the $256 \times 16 \times 16$ feature map purely through localized spatial convolutions, the CRN flattens the channel and frequency dimensions into a sequence of shape 16×4096, where 16 represents the compressed temporal frame axis. This sequence is processed by a multi-layer Bidirectional LSTM (BiLSTM) with 256 hidden units per direction, capturing forward and backward phonetic context across the entire 2.04-second utterance, totaling 32 temporal frames. The bidirectional outputs are concatenated, linearly projected back to 4096, and reshaped to $256 \times 16 \times 16$ before entering the decoder stage. This modification increases the number of trainable parameters to 14.5M. This architectural upgrade serves as a lightweight mechanism to optimize magnitude estimation before expanding into complex-valued operations (DCCRN [13]) or raw waveform models (Demucs [16]).

D. Training Setup

The models were implemented in PyTorch and trained using the AdamW optimizer with a learning rate of 3×10^{-4} and a batch size of 32. The networks were optimized using the Mean Absolute Error (L1 loss) calculated between the enhanced magnitude spectrum ($|\hat{Y}| = M \odot |X|$) and the clean magnitude spectrum ($|Y|$). Training was capped at 30 epochs, with an early stopping mechanism (patience of 5 epochs based on validation loss) used to mitigate overfitting, and the checkpoint achieving the best validation performance was retained for evaluation. All simulations were conducted using an NVIDIA RTX 5070 TI.

In total, ten models were trained per architecture across different random seeds, mitigating training-run variance in the metric estimates and enabling a statistical comparison between the two proposed architectures, discussed further in the Results section.

E. Evaluation Metrics

The performance of the models is evaluated using three objective metrics widely adopted in the speech enhancement literature, all computed on the reconstructed utterances (obtained after applying the inverse STFT to $\hat{Y}$):

- PESQ (Perceptual Evaluation of Speech Quality): Evaluated under the wideband configuration (ITU-T P.862.2), it estimates the perceptual quality of enhanced speech by modeling human auditory processing (from –0.5 to 4.5).
- STOI (Short-Time Objective Intelligibility): It measures the structural correlation of short-time temporal envelopes between the clean reference and the enhanced signal, yielding a bounded score between 0 and 1 that correlates with human speech intelligibility.
- SI-SDR (Scale-Invariant Signal-to-Distortion Ratio): A time-domain metric that quantifies overall waveform alignment, noise suppression, and artifact insertion while remaining invariant to global gain or scale modifications, expressed in decibels (dB).

F. Magnitude Oracle Analysis

To quantify the performance ceiling imposed by the masking formulation itself, an oracle experiment was conducted using the available noisy-clean pairs. The estimated mask M was replaced with an oracle reconstruction in which the noisy magnitude is substituted by the ground-truth clean magnitude

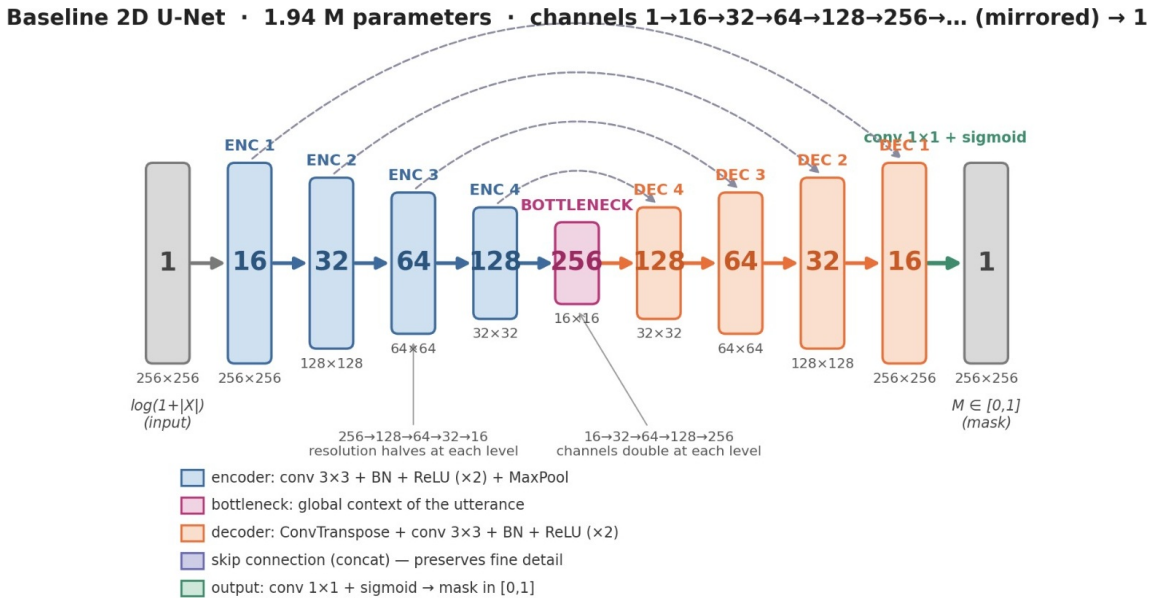


Fig. 2. Block diagram of the baseline 2D Convolutional U-Net architecture showing feature map dimensions across the encoder, bottleneck, and decoder stages.

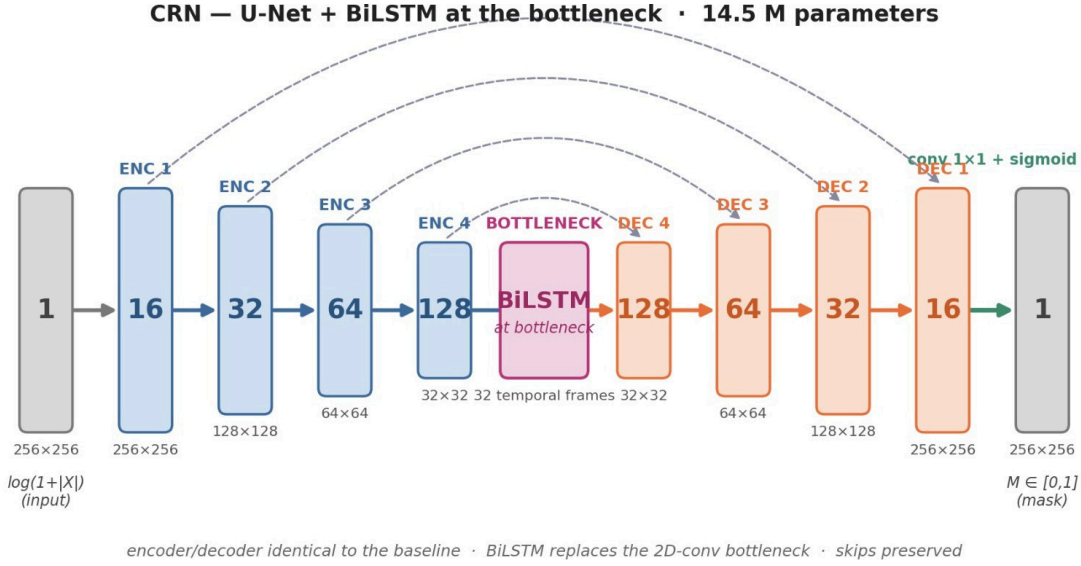


Fig. 3. Block diagram of the proposed CRN, featuring a BiLSTM at the low-dimensional bottleneck for temporal sequence modeling.

$|Y|$, while the noisy phase $\angle X$ is preserved exactly as in the trained models:

$$\hat{Y}_{\text{oracle}} = |Y| \odot e^{j\angle X}. \tag{1}$$

The inverse STFT of $\hat{Y}_{\text{oracle}}$ yields the best possible reconstruction achievable under perfect magnitude estimation, isolating the residual degradation caused exclusively by the noisy phase. Comparing this oracle PESQ score against (i) the trained models and (ii) the theoretical metric ceiling (4.5) decomposes the overall performance gap into two independent components: a magnitude-estimation component, which can, in principle, be reduced by improving the network, and a phase-preservation component, which is structurally bounded by the real-valued masking formulation and can only be addressed by complex-valued or phase-aware approaches.

III. Results

Fig. 4 shows the mean values and their standard deviations of the three evaluated metrics on the reconstructed utterances for the ten trained models of each of the two proposed architectures (U-Net baseline and CRN). The mean gain of the CRN over the baseline was +0.080 for PESQ, +0.003 for STOI and +0.059 for SI-SDR. To assess whether the difference between architectures is significant, for each metric a paired Wilcoxon signed-rank test and a paired bootstrap test were performed across the ten per-model mean values, each computed over the

824 test samples. Both tests indicate, at the 5% significance level, that the CRN significantly improves PESQ and STOI scores relative to the baseline (Wilcoxon $p = 0.002$ for both metrics), but the SI-SDR improvement is not statistically significant. This indicates that the gain introduced by the recurrent bottleneck is concentrated in perceptual quality and intelligibility (PESQ and STOI) rather than waveform-level alignment (SI-SDR), suggesting that the additional 12.5M parameters mainly refine fine-grained magnitude estimation rather than producing structurally different reconstructions.

The objective performance of the developed models on the VoiceBank-DEMAND test set is summarized in Table I, alongside reference values reported in the literature for R-CED [7], SEGAN [11], Demucs [16], DCCRN [13], WavLM-Large with PCS preprocessing [19] and SGM-EMA [18]. These reference values were obtained directly from the original publications under their own experimental protocols and are reported here as an approximate point of reference rather than a strictly controlled comparison. Both proposed architectures outperform the unprocessed noisy baseline, R-CED and the generative SEGAN model on every metric for which reference values are available, despite using substantially fewer parameters than SEGAN (43M) and Demucs (60M).

Furthermore, more recent approaches based on foundation models (e.g., WavLM-Large with PCS preprocessing) achieve higher PESQ (3.54) [19], but at parameter cost orders of magnitude larger (roughly 22–163$\times$ more parameters than our CRN and U-Net), reinforcing the practical trade-off between perceptual quality and deployability in embedded, real-time biomedical applications. Also, SGM-EMA, a recent and complex score-based diffusion model with multi-resolution U-Net structure and attention mechanism, reaches a PESQ of 2.89 [18] essentially tied with our CRN (2.89) and a higher SI-SDR (19.6 dB vs. 18.39 dB), at the cost of greater computational demand and longer inference times due to high-quality reconstruction and better modeling of global dependencies despite the lack of a reported parameter count for direct efficiency comparison. Its ESTOI of 0.91, while not numerically identical to the STOI metric we report, suggests comparable intelligibility.

TABLE I
OBJECTIVE EVALUATION ON THE VOICEBANK-DEMAND TEST SET.

Method	PESQ	STOI	SI-SDR	Params.
Noisy (unprocessed)	1.97	0.91	8.44	–
R-CED [7]	2.07	0.93	–	33k
SEGAN [11]	2.16	–	–	43M
2D U-Net (ours)[a]	2.81	0.94	18.34	1.94M
CRN (ours)[a]	2.89	0.95	18.39	14.5M
Demucs (causal) [16]	2.93	0.95	–	60M
DCCRN [13]	3.07	0.95	–	3.7M
WavLM-Large with PCS [19]	3.54	0.96	–	316M[b]
SGM-EMA [18]	2.89	0.91[c]	19.6	–

[a]Mean values across ten independently trained models per architecture.
[b]Number of parameters of WavLM-Large backbone only.
[c]ESTOI value reported instead of STOI for this reference.

To better interpret the modest but statistically significant PESQ and STOI improvements, the magnitude oracle described in Section II-F was evaluated on the same test set, yielding a PESQ of approximately 3.95. This result decomposes the overall performance gap into two components: a gap of +1.14 PESQ points between the U-Net baseline and the oracle, representing the margin that could, in principle, still be recovered through improved magnitude estimation, and a gap of +0.55 PESQ points between the oracle and the theoretical ceiling of 4.5, representing the irreducible cost of preserving the noisy phase under the real-valued masking formulation.

The architectural motivation for introducing the BiLSTM bottleneck in the CRN was the hypothesis, inspired by the recurrent design adopted in DCCRN [13], that part of the gap toward the oracle ceiling stemmed from long-range temporal dependencies between distant STFT frames that the finite receptive field of two-dimensional convolutions cannot capture. However, the PESQ improvement obtained closes only 7% of the available margin (+0.08 of 1.14 points), indicating that this hypothesis is not well supported for the present configuration: while the recurrent bottleneck yields a small gain, most of the residual gap does not appear to originate from a lack of long-term temporal context, but rather from a combination of finer local magnitude-estimation errors and the irreducible phase-related ceiling identified by the oracle analysis. This interpretation is consistent with the remaining gap between our models and DCCRN itself (PESQ 3.07), which, unlike the CRN evaluated here, additionally models the complex spectrum rather than relying purely on magnitude masking with noisy-phase reconstruction. Besides, the nearly identical performance of the multi-resolution U-Net-based architecture SGM-EMA (PESQ 2.89) compared to our models, despite its additional complexity, further corroborates the structural phase-related ceiling.

Qualitative improvements are visually evaluated through the spectrograms of a complete reconstructed utterance from an example test sample, produced by one paired trained model from each architecture, as depicted in Fig. 5. Comparing the target representations, the unprocessed noisy spectrogram displays severe broadband and localized environmental distortions that mask weak speech harmonics. The baseline 2D U-Net successfully filters out background noise but introduces noticeable over-smoothing, attenuating high-frequency components and fine formant structures. Conversely, the proposed CRN provides superior definition in the high-frequency spectra, preserving sharp harmonic transitions and phonetic details. This visual difference reflects a localized improvement in high-frequency spectral definition for the CRN output, preventing the over-suppression of valid speech energy in these regions.

IV. CONCLUSION

This work proposed and compared two time-frequency domain architectures: a baseline 2D U-Net and a CRN incorporating a bottleneck BiLSTM, for monaural speech enhancement on the VoiceBank-DEMAND benchmark. Both architectures achieved competitive results relative to the reference literature despite their reduced parameter count, with the U-Net

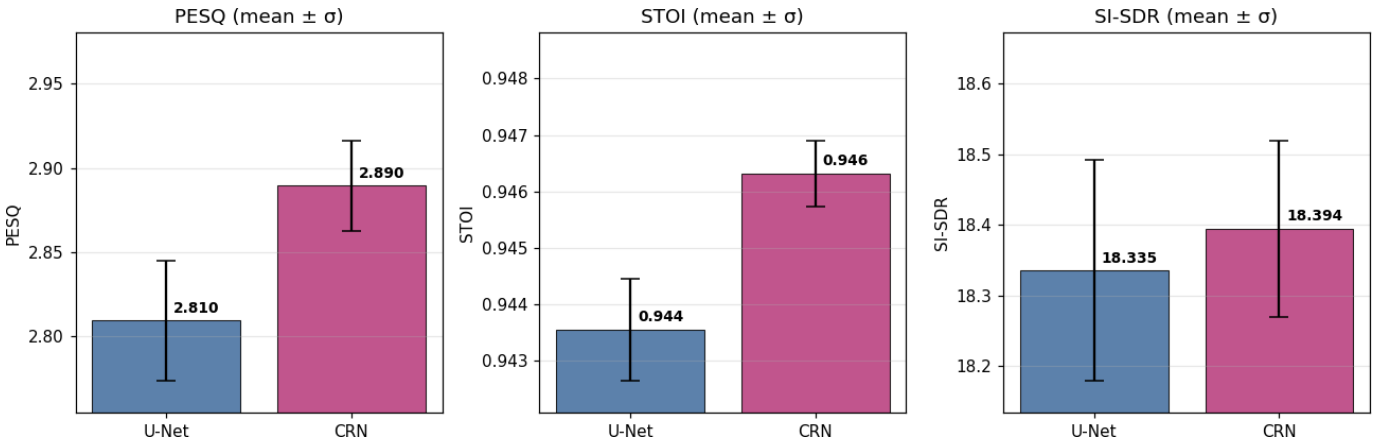


Fig. 4. Mean and standard deviation values for the PESQ, STOI and SI-SDR scores attained by the proposed architectures.

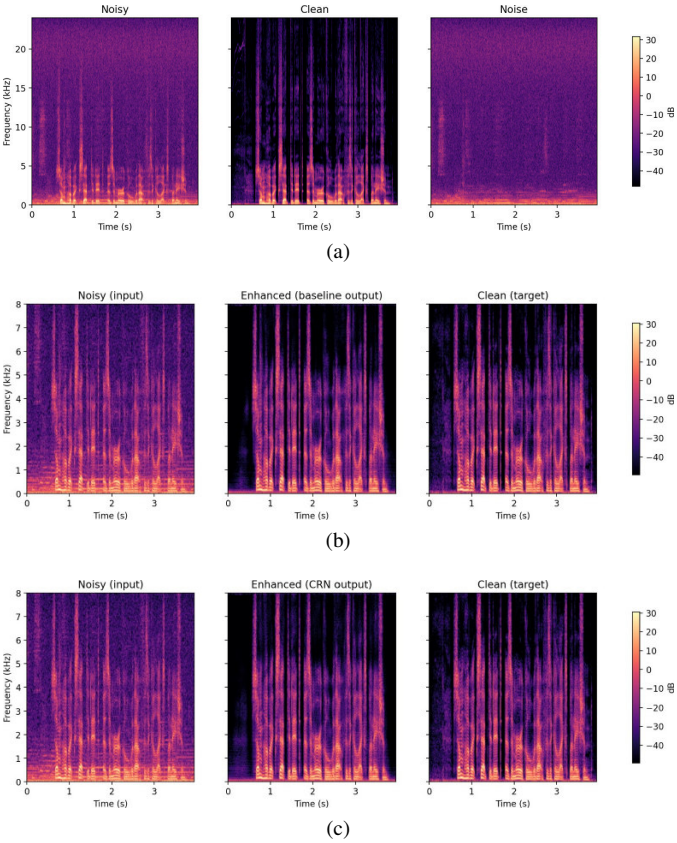


Fig. 5. Spectrograms comparison between the original example sample (a), 2D U-Net reconstruction (b) and the proposed CRN reconstruction (c).

baseline (1.94M parameters) approaching the performance of substantially larger models such as SEGAN (43M parameters) and Demucs (60M parameters).

A contribution of this work is the magnitude oracle analysis, which decomposes the performance ceiling of the masking formulation into a recoverable magnitude-estimation margin (+1.14 PESQ from the U-Net baseline) and a structural phase-preservation cost (+0.55 PESQ). This decomposition shows that the CRN's modest improvement over the baseline (+0.080 PESQ, +0.003 STOI and no statistically significant SI-SDR gain) fails to close the recoverable margin. This indicates that the initial hypothesis – motivated by the recurrent design adopted in DCCRN [13] – that long-range temporal modeling via a recurrent bottleneck was the main missing component is not supported for this configuration. This structural ceiling is further corroborated by the sharper performance degradation observed under severe non-stationary distortions (e.g., babble noise) and low-SNR regimes, conditions in which the noisy phase departs most strongly from the clean phase.

Future research should therefore prioritize approaches that directly address the phase-preservation ceiling rather than further enlarging magnitude-domain networks, namely complex-valued masking architectures to explicitly reconstruct the phase spectrum, multi-resolution STFT perceptual losses to better capture fine-grained magnitude detail, and a causal unidirectional LSTM variant to enable real-time operation, following the design principles adopted in DCCRN. These directions are relevant for embedded biomedical applications such as hearing aids, where real-time, low-latency and low-power operation requires a causal, parameter-efficient processing pipeline; in this sense, the present results reinforce the practical value of parameter-efficient, magnitude-masking architectures such as the 1.94M-parameter U-Net baseline over substantially larger, phase-agnostic alternatives.

References

[1] G. Park, W. Cho, K. Kim, and S. M. Lee, "Speech Enhancement for Hearing Aids with Deep Learning on Environmental Noises," *Applied Sciences*, 2020.

[2] M. L. Smith, M. B. Winn, and M. B. Fitzgerald, "A Large-Scale Study of the Relationship Between Degree and Type of Hearing Loss and Recognition of Speech in Quiet and Noise," *Ear and Hearing*, vol. 45, pp. 915 – 928, 2024.

[3] N. Natarajan, S. A. Batts, and K. Stankovic, "Noise-Induced Hearing Loss," *Journal of Clinical Medicine*, vol. 12, 2023.

[4] P. Sethukumar, R. Mandavia, O. Yildirim, G. G. J. Hazell, H. Devakumar, M. Ahmed, E. Stragier, M.-J. Duran, A. G. M. Schilder, and N. Mehta, "Cataloging Existing Hearing Loss Cohort Data to Guide the Development of Precision Medicine for Sensorineural Hearing Loss: A Systematic Review of Hearing Repositories," *The Journal of International Advanced Otology*, vol. 19, pp. 420 – 425, 2023.

[5] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 3651–3703, 2023.

[6] W. Zhang, K. Saijo, J.-w. Jung, C. Li, S. Watanabe, and Y. Qian, "Beyond Performance Plateaus: A Comprehensive Study on Scalability in Speech Enhancement," in *Interspeech 2024*. ISCA, Sep. 2024, pp. 1740–1744.

[7] S. R. Park and J. W. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," *Interspeech 2017*, pp. 1993–1997, 08 2017. [Online]. Available: https://cir.nii.ac.jp/crid/1360011145594860160

[8] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based Speech Enhancement Methods for Noise-Robust Text-to-Speech," in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152.

[9] S. Natarajan, S. A. R. Al-Haddad, F. A. Ahmad, R. Kamil, M. K. Hassan, S. Azrad, J. F. Macleans, S. Abdulhussain, B. M. Mahmmod, N. Saparkhojayev, and A. Dauitbayeva, "Deep neural networks for speech enhancement and speech recognition: A systematic review," *Ain Shams Engineering Journal*, 2025.

[10] Z. Li, A. Basit, A. Daraz, and A. Jan, "Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network," *PLOS ONE*, vol. 19, 2024.

[11] P. Santiago, B. Antonio, and S. Joan, "SEGAN: Speech Enhancement Generative Adversarial Network," *Interspeech 2017*, pp. 3642–3646, 08 2017. [Online]. Available: https://cir.nii.ac.jp/crid/1361699995466029184

[12] H. Chen, J. Zhang, Y. Fu, X. Zhou, R. Wang, Y. Xu, and D. Ke, "TFDense-GAN: a generative adversarial network for single-channel speech enhancement," *EURASIP Journal on Advances in Signal Processing*, vol. 2025, 2025.

[13] H. Yanxin, L. Yun, L. Shubo, X. Mengtao, Z. Shimin, F. Yihui, W. Jian, Z. Bihong, and X. Lei, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," *Interspeech 2020*, pp. 2472–2476, 10 2020. [Online]. Available: https://cir.nii.ac.jp/crid/1360016869793869056

[14] L. Wan, H. Liu, L. Shi, Y. Zhou, and L. Gan, "Cross Domain Optimization for Speech Enhancement: Parallel or Cascade?" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4328–4341, 2024.

[15] H. Shi, M. Mimura, and T. Kawahara, "Waveform-Domain Speech Enhancement Using Spectrogram Encoding for Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3049–3060, 2024.

[16] D. Alexandre, S. Gabriel, and A. Yossi, "Real Time Speech Enhancement in the Waveform Domain," *Interspeech 2020*, pp. 3291–3295, 10 2020. [Online]. Available: https://cir.nii.ac.jp/crid/1360302477007296384

[17] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[18] Y. Wu, Z. Li, and H. Huang, "SGM-EMA: Speech Enhancement Method Score-Based Diffusion Model and EMA Mechanism," *Applied Sciences*, vol. 15, no. 10, p. 5243, 2025.

[19] M. S. Khan, V. M. Salerno, M. La Quatra, K.-H. Hung, S.-W. Fu, Y. Tsao, and S. M. Siniscalchi, "Foundation Models for Speech Enhancement Leveraging Consistency Constraints and Contrast Stretching," *IEEE Access*, 2025.

[20] C. Veaux, J. Yamagishi, and S. King, "The Voice Bank Corpus: Design, Collection and Data Analysis of a Large Regional Accent Speech Database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[21] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A Database of Multichannel Environmental Noise Recordings," in *Proceedings of Meetings on Acoustics*, vol. 19. Acoustical Society of America, 2013, p. 035081.

Detection and Segmentation of Oral Ulcers and Mucosal Inflammation Using Hyperspectral Imaging and Deep Neural Networks

Rafael F. R. e Souza
School of Technology (FT)
State University of Campinas (UNICAMP)
Limeira, SP, Brazil
E-mail: rafaelfabre.rs@gmail.com

Julyo E. H. Silva
School of Technology (FT)
State University of Campinas (UNICAMP)
Limeira, SP, Brazil
E-mail: j185720@dac.unicamp.br

Leandro R. Ximenes
School of Technology (FT)
State University of Campinas (UNICAMP)
Limeira, SP, Brazil
E-mail: ronchini@unicamp.br

Juliana D. Bronzato
Piracicaba Dental School (FOP)
State University of Campinas (UNICAMP)
Piracicaba, SP, Brazil
E-mail: julianadelatorre@hotmail.com

***Abstract*—Early identification of oral ulcers and mucosal inflammation remains challenging using conventional visual examination, which provides limited information about tissue composition. Hyperspectral imaging combined with machine learning offers a promising alternative by revealing spectral patterns unavailable in conventional RGB images. However, no previous studies have been identified that apply machine learning to the classification of oral ulcers and inflammatory lesions in hyperspectral images. To address this gap, this study presents a deep learning framework for pixel-wise detection and segmentation using images from the Oral and Dental Spectral Image Database (ODSI-DB). The proposed Multilayer Perceptron (MLP) receives the reflectance spectrum of each pixel and estimates a probability map for the combined ulcerative and inflammatory region. The selected valid pixels from the ten eligible images, originating from three subjects, were pooled without separating the data by subject. The images were decomposed into pixel-level samples, and 5-fold stratified cross-validation was performed at the pixel level. The model achieved an Accuracy of 98.70% ($\pm$0.27 p.p.), Precision of 97.73% ($\pm$0.41 p.p.), Recall of 99.71% ($\pm$0.18 p.p.), ROC-AUC of 0.9979 ($\pm$0.0007), and PR-AUC of 0.9974 ($\pm$0.0009). These preliminary findings demonstrate the feasibility of the proposed approach within the available image pool and support further investigation of hyperspectral imaging and deep learning for computer-assisted analysis of oral mucosal lesions.**

***Index Terms*—Oral ulcers, mucosal inflammation, hyperspectral imaging, deep learning, multilayer perceptron, image segmentation.**

I. Introduction

Oral mucosal ulcers are frequently encountered in dental practice due to their high clinical occurrence, etiological diversity, and possible association with more severe conditions. They may result from local trauma, inflammation, infections, systemic disorders, or processes with malignant potential. In many cases, they present a benign and self-limiting course. Nevertheless, the morphological similarity among lesions of different natures requires caution during the initial evaluation. Visual inspection of the mucosal surface does not always allow a reliable distinction between healthy tissue, the ulcerated center, and the surrounding inflammatory halo. This limitation may delay clinical management, increase the dependence on complementary examinations, and reduce the accuracy in selecting the most appropriate site for incisional biopsy [1].

Subtle alterations related to vascularization, inflammation, and tissue structural composition may remain poorly evident, particularly when lesions exhibit a similar appearance during clinical examination. The early identification of ulcers and inflammatory lesions in the oral cavity is relevant for clinical practice, as these lesions may represent manifestations of systemic conditions or, in some cases, alterations requiring closer investigation. Chronic oral lesions, particularly leukoplakia and erythroplakia, present a significant risk of progression to squamous cell carcinoma, the most common form of oral cancer [1]. In routine clinical practice, the initial assessment of these lesions is typically performed by visual inspection under white light and interpreted according to the examiner's clinical experience. Although indispensable, this approach has well-recognized limitations. Visual perception is affected by ambient illumination, the presence of saliva, shadows, reflections, and individual anatomical variations within the oral cavity. In addition, conventional images, especially RGB images, condense optical information into only three color channels [2].

Hyperspectral Imaging (HSI) has emerged in this context as a non-invasive optical approach capable of expanding the assessment of the oral mucosa. Instead of limiting the image to the traditional color representation, HSI captures the interaction between light and tissue across dozens or hundreds of narrow, contiguous bands of the electromagnetic spectrum [3].

Each pixel contains its own spectral signature, organized into a hyperspectral data cube that integrates both spatial and spectral information. This structure facilitates the analysis of optical patterns associated with vascularization, edema, inflammation, tissue architecture, and alterations caused by the ulcerative process [4].

The translation of HSI into clinical settings, however, depends on analytical strategies compatible with the complexity of the generated data. The high dimensionality, spectral redundancy, and large volume of information make purely manual interpretation impractical for real-world applications [5]. Artificial Intelligence methods, particularly Deep Learning models, provide a promising approach for recognizing complex patterns present in medical images. In the hyperspectral domain, these models can be employed in tissue detection, segmentation, and classification tasks, provided that they are adapted to the specific characteristics of spectral data and to the limited sample availability commonly found in biomedical applications [6].

The automated analysis of oral lesions offers significant benefits for both public health and clinical practice. First, it may enable rapid and standardized patient screening, reducing the dependence on subjective clinical interpretation while facilitating the identification of cases that require specialized evaluation. The digitalization of diagnostic support systems can integrate complex spectral data into clinical workflows, making the analysis more accessible and reproducible. In public health settings, this approach may contribute to large-scale screening programs, allowing patients to be prioritized for specialized care while supporting the early identification of oral mucosal alterations. Furthermore, it may reduce operational costs and increase diagnostic efficiency, democratizing access to detection tools in regions with limited availability of specialized dental care [7].

Recently, hyperspectral imaging has been increasingly explored in dentistry and oral sciences, often in combination with machine learning and, in several studies, deep neural networks. Previous studies using artificial intelligence have primarily focused on oral cancer detection, tongue melanosis, oral tissue classification, and pathological staging [4], [8], [9], [11], [12]. Other investigations have addressed applications more directly related to dentistry, including the development of the Oral and Dental Spectral Image Database (ODSI-DB) and dental hyperspectral image analysis [13], [14], the identification of dental caries and calculus [15], [16], gingival region detection [17], and periodontal assessment [18]. In addition, hyperspectral microscopy, histopathological slides, and multimodal imaging have been employed for classification and segmentation tasks [10], [19], [20]. Despite these advances, a considerable proportion of the literature is based on hyperspectral microscopy, histological specimens, or images acquired under controlled experimental conditions rather than hyperspectral images captured directly from in vivo oral tissues using hyperspectral cameras. Furthermore, to the best of our knowledge, none of the published studies has specifically investigated the detection and segmentation of oral ulcers and mucosal inflammation, highlighting an important gap that the present work aims to address.

This work presents a structured computational pipeline based on a deep neural network for the automated detection and segmentation of oral ulcerative and inflammatory regions in hyperspectral images from the ODSI-DB.

II. Materials and Methods

A. Dataset

This study uses the ODSI-DB [13], a publicly available hyperspectral image database. In the present study, only images acquired with the Specim IQ hyperspectral camera and containing annotations for oral ulcers or mucosal inflammation were considered. Among the available images, ten images from three patients met these criteria and were included in the analysis. The Specim IQ acquisitions provide reflectance information over 204 narrow and contiguous spectral bands.

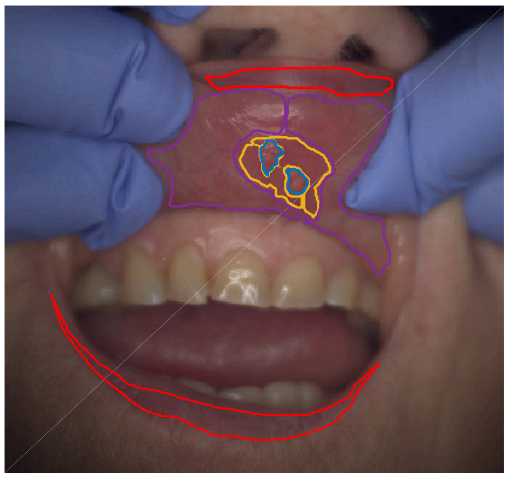
The preprocessing stage consisted of loading the hyperspectral images together with their corresponding annotation masks in `.tif` format. Pixel intensities were normalized to the interval $[0, 1]$ by dividing the original 16-bit grayscale values by 65,535.[1]

Pixels corresponding to technical artifacts previously annotated in the database, including specular reflections, shadow/noise regions, and out-of-focus areas, were excluded from the training and validation datasets. The remaining valid pixels were subsequently extracted individually and randomly shuffled to reduce the influence of sample ordering during model training.

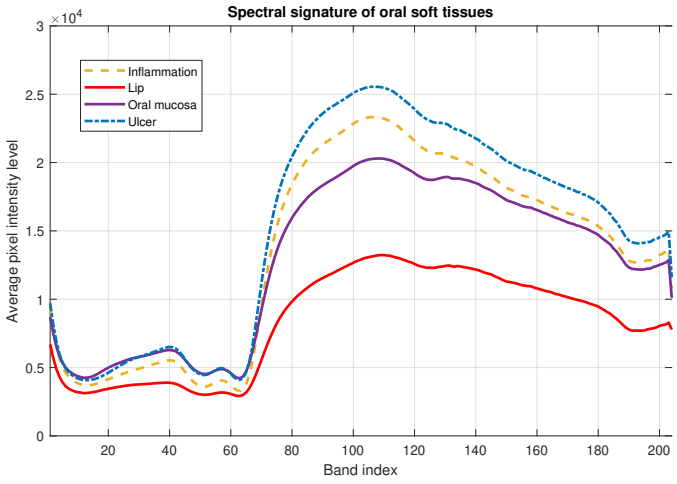
Fig. 1 illustrates one representative hyperspectral image used in this study, with the manual annotations (1.a) and the corresponding mean spectral signatures of the annotated regions (1.b). The example highlights that different oral tissues and pathological regions exhibit distinct spectral patterns across multiple bands, providing information beyond the three channels available in conventional RGB images. Each curve was obtained by averaging the pixel values within a given annotated region across the hyperspectral bands. It is important to note that the original manual annotations of the ODSI-DB dataset do not necessarily cover the entire extent of each tissue or lesion, but rather selected representative regions.

In this study, ulcer and inflammatory regions were grouped into a single positive class during model training. This strategy was adopted because the reference masks provided by the ODSI-DB are based on polygonal annotations manually delineated by specialists, and the transition between the fibrin-covered ulcer region and the surrounding inflammatory border may be gradual in several images. Therefore, both annotations were considered part of the same target class throughout the training process. The negative class was composed of the remaining valid non-lesion pixels after the exclusion of technical artifacts.

[1] According to the official ODSI-DB documentation, the hyperspectral images are stored as multi-page TIFF files containing 16-bit grayscale spectral bands. Available at: https://sites.uef.fi/spectral/databases-software/odsi-db/. Accessed: Jun. 2026.



(a) Ground-truth annotations overlaid on the RGB image. Annotation colors indicate the lips (red), inflammation (yellow), ulcer (blue), and oral mucosa (purple).



(b) Average spectral signatures of the annotated classes.

Figure 1: Ground-truth annotations and their corresponding spectral signatures. (a) Colored borders overlaid on the RGB image define the annotated classes used in this study. (b) Average spectral signatures extracted from each annotated class.

The extraction of negative pixels was balanced with respect to the positive class. For each of the ten eligible images, an equal number of positive and negative pixels was extracted, yielding a dataset of 24,328 pixels, evenly divided into 12,164 positive pixels, corresponding to ulcer or inflammation, and 12,164 negative pixels. To reduce the model bias toward the majority class caused by class imbalance, random undersampling was applied to the negative class until the number of negative pixels matched that of the positive pixels [22].

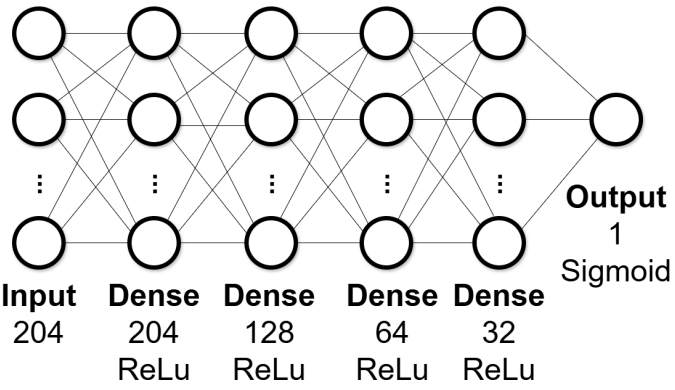


Figure 2: Schematic representation of the proposed MLP-based pixel-wise spectral classification model.

All valid pixels from the ten eligible images were then pooled into a single dataset. A 5-fold stratified cross-validation with random shuffling was performed at the pixel level. In each iteration, approximately 80% of the pixels were used for training and 20% for validation, while preserving the balance between the two classes. Accordingly, each training partition contained 19,462 or 19,463 pixels, and each validation partition contained 4,865 or 4,866 pixels. The model was trained on four folds and evaluated on the remaining fold, and the validation metrics were averaged across the five folds.

B. Neural Network Architecture

A Multilayer Perceptron (MLP) architecture was designed for pixel-wise spectral segmentation. Each input sample corresponds to the reflectance spectrum of a single pixel, represented by the values of the 204 spectral bands acquired by the hyperspectral camera. Consequently, the input layer contains 204 neurons, followed by four fully connected hidden layers comprising 204, 128, 64, and 32 neurons, respectively. All hidden layers employ the Rectified Linear Unit (ReLU) activation function. To mitigate overfitting, a Dropout layer with a rate of 0.1 was applied after each hidden layer during training. The architecture was empirically defined after preliminary experiments to provide a compromise between model capacity and overfitting.

The output layer consists of a single neuron with a sigmoid activation function, which estimates the probability that the input pixel belongs to the positive class. The proposed architecture is illustrated in Fig. 2.

C. Training Protocol

The training adopted a batch size of 256 samples and the Adam optimizer with a learning rate of 1×10^{-5}. Each fold was trained for up to 2000 epochs, with an early-stopping criterion monitoring the validation PR-AUC, a patience of 100 epochs, and restoration of the model weights from the epoch with the best validation PR-AUC once training stopped. For each fold, both the metrics obtained at the last training epoch and the metrics obtained at the restored (best) epoch were recorded, together with the number of epochs actually

Table I: Epochs run, restored epoch, and best validation PR-AUC per fold.

Fold	Epochs run	Restored epoch	Best val. PR-AUC
1	1,302	1,202	0.9964
2	2,000	1,987	0.9983
3	1,360	1,260	0.9970
4	1,564	1,464	0.9984
5	1,304	1,204	0.9969

run before training was stopped, either by the early-stopping criterion or by reaching the maximum epoch limit.

The Binary Cross-Entropy loss function was used, defined as

$$\mathcal{L}_{BCE} = -\left[y \log(p) + (1-y)\log(1-p)\right], \tag{1}$$

where $y \in \{0,1\}$ is the ground-truth label and $p \in [0,1]$ is the probability predicted by the network for the positive class, corresponding to inflammation or ulcer.

For the computation of Accuracy, Precision, and Recall, the predicted probability p was converted into a binary prediction $\hat{y}$ using

$$\hat{y} = \begin{cases} 1, & p \geq \tau, \\ 0, & p < \tau, \end{cases} \tag{2}$$

with a decision threshold of $\tau = 0.5$. ROC-AUC and PR-AUC were computed directly from the predicted probabilities, without applying a fixed decision threshold, using the default TensorFlow/Keras AUC metric implementation [21].

Mean training and validation curves were computed by averaging the per-epoch values across the five cross-validation folds. Since the folds converged at different epoch counts, the curves were aligned by their epoch index and the mean was calculated element-wise over the available values.

For the qualitative evaluation, the trained model was applied to each pixel of the hyperspectral images, and the resulting probability scores were mapped back to their spatial positions to generate a probability map for each image. Binary segmentation masks were then obtained by thresholding the probability maps, with pixels of probability $p \geq \tau$ classified as positive, corresponding to ulcer or inflammation, and the remaining pixels classified as negative.

III. Results

Table I summarizes the number of epochs run, the restored epoch, and the best validation PR-AUC obtained for each fold. Four folds were stopped by the early-stopping criterion, whereas Fold 2 reached the maximum limit of 2000 epochs.

The mean training and validation curves across the five folds are shown in Fig. 3. The curves present the evolution of the loss, accuracy, and PR-AUC over the training epochs. The training and validation curves remained closely aligned throughout the training process, indicating stable convergence with no evidence of overfitting under the adopted training protocol.

Under the 5-fold stratified cross-validation protocol, performance at the restored epoch of each fold is reported as the mean and standard deviation, calculated with Bessel's

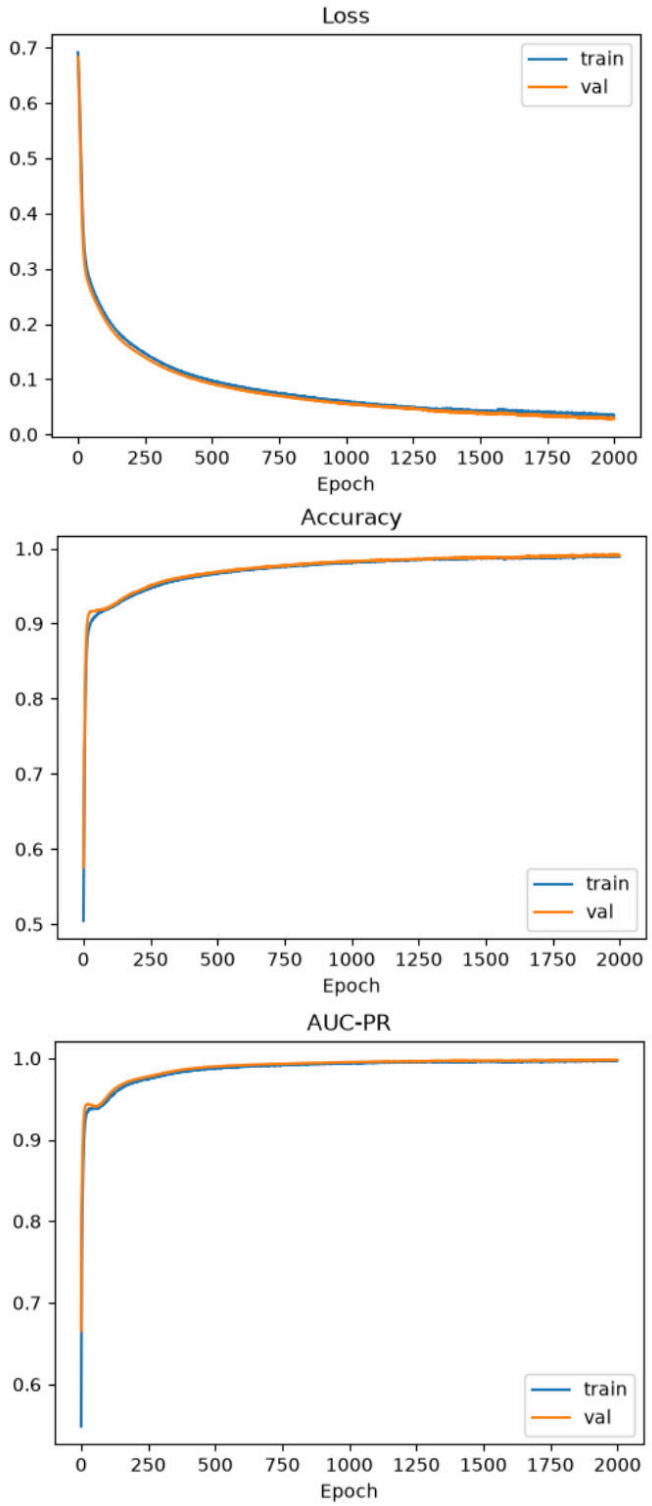


Figure 3: Mean training and validation curves across the five folds, showing loss, accuracy, and PR-AUC over epochs.

Table II: Per-fold performance metrics (mean ± std), restored best-epoch weights.

Fold	Acc.	Prec.	Rec.	Dice	ROC-AUC	PR-AUC	Loss
1	98.36%	97.27%	99.51%	98.38%	0.997	0.996	0.051
2	99.03%	98.30%	99.79%	99.04%	0.999	0.998	0.030
3	98.62%	97.74%	99.55%	98.64%	0.998	0.997	0.044
4	98.89%	97.91%	99.92%	98.90%	0.999	0.998	0.036
5	98.58%	97.43%	99.79%	98.60%	0.997	0.997	0.047
Mean	98.70%	97.73%	99.71%	98.71%	0.998	0.997	0.042
Std.	0.27%	0.41%	0.18%	0.26%	0.001	0.001	0.008

correction ($n-1$): Accuracy of 98.70% (±0.27 p.p.), Precision of 97.73% (±0.41 p.p.), Recall of 99.71% (±0.18 p.p.), Dice of 98.71% (±0.26 p.p.), ROC-AUC of 0.9979 (±0.0007), PR-AUC of 0.9974 (±0.0009), and loss of 0.0417 (±0.0083), with the Dice coefficient derived from the Precision and Recall values above. Table II details the individual fold results together with the overall mean and standard deviation. These metrics were computed at the standard threshold $\tau = 0.5$, while the higher thresholds in Figs. 4 and 5 ($\tau = 0.90$ to 0.99) were used only for qualitative visualization of the segmentation boundary. Notably, increasing τ tends to reduce false positives, increasing Precision, while increasing false negatives and reducing Recall. This trade-off does not affect ROC-AUC and PR-AUC, which are threshold-independent metrics. In clinical applications, where failing to detect an existing lesion may be more critical than a false alarm, Recall could be favored by selecting a lower decision threshold. For the quantitative results reported above, however, the standard threshold $\tau = 0.5$ was retained.

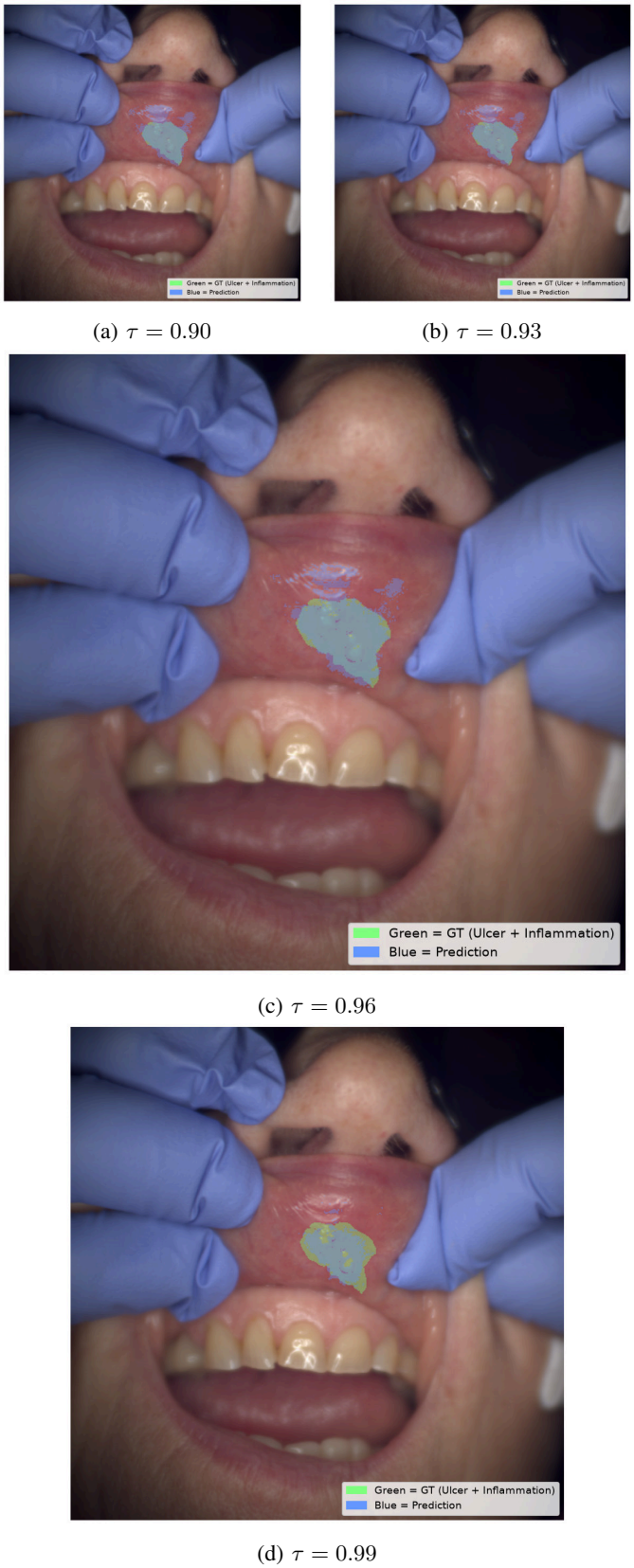
For the qualitative analysis, segmentation masks were generated using decision thresholds ranging from $\tau = 0.90$ to $\tau = 0.99$. This analysis illustrates how increasing the decision threshold affects the extent of the detected regions, with higher values of τ producing progressively more conservative segmentations by retaining only pixels associated with higher predicted probabilities.

Fig. 4 presents the segmentation results for image `f533af9b_lip_3.tif`, which contains both ulcer and inflammation annotations. As the decision threshold increases from 0.90 to 0.99, the segmented region progressively decreases, and only pixels associated with higher predicted probabilities are retained.

Fig. 5 presents a second example for image `4ad8fb28_lip_3.tif`, which contains annotations only for the ulcer region. As in the previous case, increasing the decision threshold produces progressively more conservative segmentations.

IV. Discussion

The proposed model achieved strong overall classification performance across all evaluated metrics, with particularly high Recall, which is especially relevant for the intended clinical application, where false negatives are more critical than false positives. Although neural network training is inherently stochastic, the proposed network achieved consistent segmentation results under the adopted experimental configuration.



(a) $\tau = 0.90$ (b) $\tau = 0.93$

(c) $\tau = 0.96$

(d) $\tau = 0.99$

Figure 4: Segmentation results for image `f533af9b_lip_3.tif` obtained using different decision thresholds. Both ulcer and inflammation are present. Green indicates the ground-truth annotations, and blue indicates the regions detected by the model.



(a) $\tau = 0.90$



(b) $\tau = 0.93$



(c) $\tau = 0.96$



(d) $\tau = 0.99$

Figure 5: Segmentation results for image `4ad8fb28_lip_3.tif` obtained using different decision thresholds. Only ulcer is present. Green indicates the ground-truth annotation, and blue indicates the regions detected by the model.

A relevant limitation of this study is related to the dataset itself. To the best of our knowledge, ODSI-DB is the only publicly available database containing hyperspectral images with annotations of oral ulcers and mucosal inflammation. Despite the limited number of annotated hyperspectral images available, each image provides rich spectral information through 204 bands, yielding a high-dimensional representation for every pixel. This spectral richness was sufficient to train the neural network and demonstrate the feasibility of the approach.

Nevertheless, the lack of additional comparable datasets limits broader external validation. Moreover, given the limited number of patients and annotated images, the cross-validation splits were not constrained at the patient level. Therefore, the reported metrics reflect discrimination within the current image pool rather than generalization to unseen patients. Even so, the generated segmentation maps showed visual agreement with the annotated ulcerative and inflammatory regions, as illustrated in Figs. 4 and 5. Future studies should include larger datasets and additional annotated hyperspectral images to further evaluate the robustness and generalization capability of the proposed framework.

Another limitation arises from the exclusion of pixels corresponding to technical artifacts, such as specular reflections, shadow/noise regions, and out-of-focus areas, during model training. Although this preprocessing strategy prevents the network from learning non-informative patterns, it may reduce its generalization capability in real clinical scenarios, where such artifacts are naturally present. Consequently, the performance observed under the controlled experimental conditions may overestimate the accuracy achievable in future datasets.

The characteristics of the ODSI-DB annotations should also be considered when interpreting the results. The ground-truth masks correspond exclusively to regions manually delineated by specialists and do not necessarily encompass the complete spatial extent of the anatomical structures of interest, as shown in Fig. 1a. As a consequence, some pixels visually similar to annotated lesions may be treated as negatives during evaluation, which may increase the apparent false-positive rate. This behavior may reflect either the presence of unannotated inflamed tissue or the inherent uncertainty associated with manual annotation.

An additional limitation concerns the discrimination between the fibrin-covered ulcer surface and the surrounding inflammatory border. As illustrated in Fig. 4, increasing the decision threshold progressively reduces the response over the central fibrin-covered region, while the inflammatory border remains detected with higher confidence, resulting in discontinuities within the lesion. This behavior is consistent with the spectral analysis shown in Fig. 1b, where both regions exhibit highly similar spectral signatures, especially in spatially adjacent transition areas. These subtle spectral differences reflect the gradual anatomical transition between ulcerated and inflamed tissues, which is also difficult to distinguish during routine clinical examination because of the morphological continuity between these structures [1]. The same characteristic is reflected in the ODSI-DB annotations, where the ulcer

border and independently annotated inflammatory regions do not always present a clearly defined boundary. Therefore, the observed segmentation behavior may be explained by both the intrinsic spectral similarity between ulcerated and inflamed tissues and the uncertainty associated with manual reference annotations, supporting the strategy adopted in this work of merging ulcer and inflammation into a single positive class during model training.

Overall, the proposed methodology demonstrates the potential of hyperspectral imaging combined with deep learning for the automatic identification of oral ulcerative and inflammatory regions. The approach is not intended to replace the assessment performed by dental professionals or to provide standalone diagnosis. Rather, it provides a methodological framework for digital image processing and spectral analysis that may support future computer-assisted tools for oral mucosal assessment. Future work should prioritize the development of larger annotated hyperspectral databases for oral mucosal lesions. In addition, morphological processing may be incorporated to refine the segmentation masks, and spatial-spectral deep learning architectures may be explored to jointly exploit spectral and spatial information.

V. Conclusion

This study presented a computational framework for the automatic segmentation of oral ulcerative and inflammatory regions in hyperspectral images using an MLP network. The proposed methodology combines pixel-wise spectral analysis and supervised deep learning. The pixel-level cross-validation protocol, performed by decomposing the images into pixel-level samples without separating the data by subject, achieved high and consistent performance across the five folds, with a mean Accuracy of 98.70%, Recall of 99.71%, and PR-AUC of 0.9974.

The findings indicate that hyperspectral imaging combined with deep learning may support the automatic identification of oral ulcerative and inflammatory regions. Nevertheless, the interpretation of the results should consider the limited availability of annotated hyperspectral datasets for oral mucosal lesions, the characteristics of manual annotations, and the exclusion of technical artifacts during training.

Future work should prioritize the development of larger annotated hyperspectral databases for oral mucosal lesions. In addition, morphological processing may be incorporated to refine segmentation masks, and spatial-spectral deep learning models may be explored to jointly exploit spectral and spatial information.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

The authors acknowledge the use of generative artificial intelligence tools to assist in manuscript preparation and code development. OpenAI's ChatGPT (GPT-5 model) and DeepSeek (V4, 2026) were used to improve readability, grammar, and text formatting. Anthropic's Claude Sonnet 4.6 was utilized to aid in writing and debugging Python scripts for data preprocessing, model training, and evaluation. All scientific content, experimental results, analyses, and conclusions were reviewed and validated by the authors, who assume full responsibility for the final manuscript.

References

[1] W. C. Gonsalves, A. C. Chi, and B. W. Neville, "Common oral lesions: Part I. Superficial mucosal lesions," *American Family Physician*, vol. 75, no. 4, pp. 501–507, Feb. 2007.

[2] S. L. Matarasso and S. Matarasso, "Clinical photography in dentistry: Current concepts and applications," *Dental Clinics of North America*, vol. 63, no. 1, pp. 1–14, 2019.

[3] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *Journal of Biomedical Optics*, vol. 19, no. 1, Art. no. 010901, Jan. 2014, doi: 10.1117/1.JBO.19.1.010901.

[4] P. Römer, J. J. Ponciano, K. Kloster, F. Siegberg, B. Plaß, S. Vinayahalingam, B. Al-Nawas, P. W. Kämmerer, T. Klauer, and D. Thiem, "Enhancing Oral Health Diagnostics With Hyperspectral Imaging and Computer Vision: Clinical Dataset Study," *JMIR Medical Informatics*, vol. 13, Art. no. e76148, 2025, doi:10.2196/76148.

[5] J. M. Bioucas-Dias *et al.*, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, Jun. 2013, doi: 10.1109/MGRS.2013.2244672.

[6] U. Khan, T. Paheding, C. P. Elkin, and V. Devabhaktuni, "Trends in deep learning for medical hyperspectral image analysis," *IEEE Access*, vol. 9, pp. 79534–79548, 2021, doi: 10.1109/ACCESS.2021.3084282.

[7] A. Kharche, A. Mathur, and V. Mehta, "AI-powered oral cancer detection: A breakthrough in dental diagnostics," *Oral Oncology Reports*, vol. 10, Art. no. 100293, Jun. 2024, doi: 10.1016/j.oor.2024.100293.

[8] A. Nuraini, A. Harmoko, and B. Kiswanjaya, "Improvement of Smoker Prediction System Based on Hyperspectral Image with Hybrid Deep Learning Model," in *2021 International Electronics Symposium (IES)*, IEEE, 2021, doi:10.1109/IES53407.2021.9594025.

[9] P. R. Jeyaraj and E. R. S. Nadar, "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm," *Journal of Cancer Research and Clinical Oncology*, vol. 145, pp. 829–837, 2019, doi:10.1007/s00432-018-02834-7.

[10] X. Zhou, L. Ma, M. Halicek, J. Dormer, and B. Fei, "Development of a new polarized hyperspectral imaging microscope," in *Progress in Biomedical Optics and Imaging – Proceedings of SPIE*, 2020.

[11] D. G. E. Thiem, P. Römer, M. Gielisch, B. Al-Nawas, M. Schlüter, B. Plaß, and P. W. Kämmerer, "Hyperspectral imaging and artificial intelligence to detect oral malignancy—Part 1—Automated tissue classification of oral muscle, fat and mucosa using a light-weight 6-layer deep neural network," *Head & Face Medicine*, vol. 17, no. 1, 2021.

[12] Y. Zhang, Z. Liu, C. Wu, G. Li, M. Liu, X. Han, T. Suo, and J. Zhao, "Determination of the pathological staging of the oral squamous cell carcinoma based on microscopic hyperspectral imaging," *Oral Oncology*, vol. 170, Art. no. 107686, 2025, doi:10.1016/j.oraloncology.2025.107686.

[13] J. Hyttinen, P. Fält, H. Jäsberg, A. Kullaa, and M. Hauta-Kasari, "Oral and Dental Spectral Image Database (ODSI-DB)," *Applied Sciences*, vol. 10, no. 20, Art. no. 7246, 2020, doi:10.3390/app10207246.

[14] O. Boiko, J. Hyttinen, P. Fält, H. Jäsberg, A. Mirhashemi, A. Kullaa, and M. Hauta-Kasari, "Deep learning for dental hyperspectral image analysis," *Color and Imaging Conference*, 2019/2020.

[15] C. Wang, R. Zhang, X. Wei, L. Wang, W. Xu, and Q. Yao, "Machine learning-based automatic identification and diagnosis of dental caries and calculus using hyperspectral fluorescence imaging," *Photodiagnosis and Photodynamic Therapy*, vol. 41, Art. no. 103217, 2023, doi:10.1016/j.pdpdt.2022.103217.

[16] C. Wang, H. Xu, H. Tang, L. Xin, X. Huang, N. Wang, X. Zhao, X. Wei, and R. Zhang, "Fluorescence and reflectance-based dual-modal hyperspectral image fusion for caries diagnosis," *Measurement*, vol. 246, Art. no. 116701, 2025, doi:10.1016/j.measurement.2025.116701.

[17] Y. Miura, Y. Yamada, Y. Sugita, S. Ando, N. Fukase, T. P. Nakao, T. Watanabe, and K. Watanabe, "Study on the Detection of the Gingival Region Using Oral and Dental Spectral Image Database (ODSI-DB)," in *2025 1st International Conference on Consumer Technology (ICCT-Pacific)*, IEEE, 2025, doi:10.1109/ICCT-Pacific63901.2025.11012800.

[18] Y. Yamada, Y. Sugita, Y. Miura, S. Ando, N. Fukase, T. P. Nakao, T. Watanabe, and K. Watanabe, "Research for the Realization of Diagnostic Imaging of Periodontal Disease by ODSI-DB," in *IEEE International Conference on Consumer Electronics – Taiwan*, IEEE, 2025, doi:10.1109/ICCE-Taiwan66881.2025.11207966.

[19] X. Zhang, Q. Li, W. Li, Y. Guo, J. Zhang, C. Guo, K. Chang, and N. H. Lovell, "FD-Net: Feature Distillation Network for Oral Squamous Cell Carcinoma Lymph Node Segmentation in Hyperspectral Imagery," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1552–1561, 2024, doi:10.1109/JBHI.2024.3350245.

[20] M. W. Talaat, M. ElSheikh, M. A. Shafaey, M. A. M. Salem, and M. Hamed, "Comparative Study on Medical Hyperspectral Images Segmentation Using Advanced Deep Learning Techniques," in *47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2025, doi:10.1109/EMBC58623.2025.11253611.

[21] TensorFlow Developers, "tf.keras.metrics.AUC," *TensorFlow API Documentation*. https://www.tensorflow.org/api_docs/python/tf/keras/metrics/AUC, Accessed: Jun. 2026.

[22] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi:10.1109/ACCESS.2025.3531662.

Detection of Melanoma Presence in Dermoscopic Images Using a Transfer Learning Approach with ResNet50

1st Ana Raquel O. de Sousa
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
sousa.raquel@academico.ifpb.edu.br

2nd José A. S. de S. Filho
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
alexandre.filho@academico.ifpb.edu.br

3rd Villeneve de O. Soares
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
villeneveoliveira@gmail.com

4th Pedro A. de A. Silva
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
pedrotatico1321@gmail.com

5th Carlos D. M. Regis
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—Melanoma is one of the most aggressive types of skin cancer, carrying a high mortality rate when not diagnosed early. The automatic detection of melanocytic lesions using artificial intelligence has emerged as a promising alternative to assist specialists in identifying the disease. In this work, an approach for the binary classification of dermoscopic images is proposed using the ResNet50 architecture combined with transfer learning, data augmentation, and fine-tuning techniques. The training process incorporated data augmentation operations and the adjustment of the network's deeper layers to enhance generalization and adaptation to the medical imaging domain. Experiments were conducted using dermoscopic images of melanoma and non-melanoma lesions under various training configurations. The results demonstrated that combining transfer learning, data augmentation, and fine-tuning significantly improved model performance, achieving 88.7 percent accuracy, 74.7 percent recall, and 66.1 percent precision in lesion classification. These findings highlight the potential of the proposed approach as a supportive tool for the early diagnosis of melanoma.**

***Index Terms*—Melanoma Detection, Dermoscopic Images, Deep Learning, Transfer Learning, ResNet50.**

I. Introduction

Artificial Intelligence (AI) is rapidly being integrated into daily life. In the medical field, significant efforts have been devoted to developing intelligent systems capable of supporting real-time diagnosis and clinical decision-making.

Melanoma is one of the most aggressive and lethal forms of skin cancer, with global incidence steadily rising. According to recent cancer statistics, the estimated number of new melanoma cases reached 57,180 in men and 42,600 in women within the analyzed period [1]. Furthermore, the World Health Organization (WHO) has reported approximately 132,000 melanoma cases worldwide [2]. Late detection of this malignancy drastically increases the risk of metastasis, contributing to poor patient prognosis. Consequently, early detection plays a crucial role in improving survival rates, prompting considerable research into accurate and efficient methods for melanoma screening.

Machine Learning (ML) techniques have been widely applied to advance automated melanoma diagnosis. These techniques enable computational models to learn complex patterns from labeled data and generalize predictions to unseen samples. This computational shift is critical, as conventional biopsy-based diagnostic procedures remain labor-intensive, time-consuming, and costly [3].

1) Dermoscopic Images: Dermoscopic images are acquired using a specialized optical device known as a dermatoscope. This instrument enables the visualization of subsurface skin structures that are invisible to the naked eye, assisting in the detection and assessment of melanocytic lesions. Dermoscopic examination can reveal characteristic pigmented structures, such as pigment networks, blotches, globules, and streaks. These images are typically captured and stored using digital imaging systems, allowing sequential recording and longitudinal monitoring of lesions to identify morphological changes over time [4].

2) Melanoma: Melanoma develops in specialized skin cells known as melanocytes, which become malignant when they begin to proliferate uncontrollably. It frequently occurs in areas regularly exposed to sunlight, including the arms, lips, nose, face, and hands [5]. Clinically, melanoma is classified into four major subtypes: nodular melanoma, superficial spreading melanoma, lentigo maligna melanoma, and acral lentiginous melanoma [6].

3) Non-Melanoma Skin Cancer: Non-melanoma skin cancer is generally less aggressive than melanoma; however, it remains a significant public health concern. Its development is strongly associated with exposure to ultraviolet (UV) radiation from sunlight. UV radiation is categorized into three types: UVA, UVB, and UVC, where prolonged exposure to UVB radiation constitutes one of the primary risk factors for skin

malignancy. Additional risk factors include fair skin, advanced age, the presence of freckles, and exposure to certain chemical agents [6]. Non-melanoma skin cancer encompasses several variants, including basal cell carcinoma, squamous cell carcinoma, Merkel cell carcinoma, and Kaposi's sarcoma [6]. Moreover, recurrence may occur even after successful clinical intervention.

To address these challenges, several studies have explored machine learning and deep learning architectures for skin lesion classification.

Sabir and Mehmood [7] proposed a melanoma classification framework based on convolutional neural networks, comparing CNN, ResNet-18, and EfficientNet-B0 architectures. Their objective was to improve the early diagnosis of melanoma by evaluating the effectiveness of different deep learning models. Using the ISIC skin lesion dataset, the framework incorporated image preprocessing, data augmentation, segmentation, feature extraction, and classification. The EfficientNet-B0 model achieved the best performance, reaching an accuracy of approximately 97% recall of 99%. However, the proposed pipeline involves multiple processing stages, increasing computational complexity.

Thwin and Park [8] introduced a deep ensemble model combining VGG16, Inception-V3, and ResNet-50 for skin lesion classification. The study aimed to improve classification robustness by integrating complementary CNN architectures through transfer learning and fine-tuning. Evaluated on the balanced ISIC 2018 dataset, the ensemble model achieved an accuracy of 97% and a melanoma recall of 91%, outperforming the individual networks. Nevertheless, the use of multiple deep models significantly increases computational cost and memory requirements during inference.

Selvaraj et al. [9] proposed an ensemble framework integrating ResNeXt101, SE-ResNeXt101, ResNet152V2, DenseNet201, GoogLeNet, and Xception for automated skin lesion classification. The objective was to improve diagnostic performance while addressing challenges such as class imbalance, image noise, and limited data availability. The method was evaluated on the HAM10000 and ISIC datasets, where the weighted ensemble model achieved an accuracy of approximately 89% and a recall of 95.24%. However, the large number of deep models considerably increases training time and computational requirements.

Padhy et al. [10] proposed the R-LSTM50 architecture, combining ResNet50 for feature extraction with Long Short-Term Memory (LSTM) networks. The objective was to enhance skin lesion classification by exploiting spatial and sequential feature representations simultaneously. Evaluated on the ISIC2020, the proposed model achieved accuracies of 95.72% and a recall of 95.24%, outperforming several conventional transfer learning approaches. Nevertheless, integrating LSTM layers increases model complexity and training time compared with standard CNN-based methods.

Islam and Panta [11] proposed a transfer learning framework for automated skin cancer classification by comparing five pre-trained convolutional neural networks, including ResNet50. The objective of the study was to improve the early detection of malignant skin lesions through transfer learning, fine-tuning, and data augmentation. The framework was evaluated on a subset of the ISIC dataset containing 3,297 dermoscopic images, where the ResNet50 model achieved an accuracy of 93.5% and a recall of 77%. However, the proposed framework was evaluated only for binary classification, which may limit its applicability to more complex multi-class skin lesion classification tasks.

Therefore, the objective of this work is to investigate the effectiveness of a ResNet50-based transfer learning framework for automated melanoma detection in dermoscopic images. The proposed method combines transfer learning, data augmentation, and fine-tuning strategies to achieve competitive classification performance using a single ResNet50 architecture. The model is evaluated on the ISIC 2019 dataset through standard classification metrics.

II. Methodology

The experimental workflow was structured into three sequential phases to systematically evaluate different optimization strategies. First, a baseline configuration was established by employing the ResNet50 architecture strictly as a fixed feature extractor, where the convolutional backbone remained frozen and only the fully connected (FC) classification layer was trained without data augmentation. In the second phase, a data augmentation strategy was introduced into the training pipeline to evaluate its impact on model generalization. Finally, the third configuration combined the same data augmentation pipeline with partial fine-tuning of the deeper layers of the network, targeting a localized adaptation to the domain-specific features of the dataset.

A. Dataset

This study utilizes the ISIC Challenge 2019 dataset, which contains 25,331 images in its official training set across nine diagnostic categories of skin lesions. In this work, the melanoma class was defined as the positive class (4,522 images), while the remaining eight categories (20,809 images) were aggregated into a single non-melanoma class, establishing a binary classification task.

Due to computational constraints, the official ISIC test set labels were unavailable; thus, the 25,331 images were partitioned using a stratified 80/20 train-validation split with a fixed random seed of 42. This stratified strategy ensured that the original class proportions were preserved across both subsets, yielding 20,264 images for model training (3,617 melanoma and 16,647 non-melanoma) and 5,067 images for internal validation (905 melanoma and 4,162 non-melanoma). Given metadata availability constraints, this partition was conducted directly at the image level based on class labels rather than using patient or lesion identifiers; consequently, the potential for cross-set lesion overlap remains a limitation of the current validation split.

It should be noted that the 20% validation subset serves strictly for model optimization and internal performance com-

parison rather than as an independent test set. Prior to model input, all dermoscopic images were resized to 224×224 pixels to match the standard input resolution of the ResNet50 architecture.

B. Data Augmentation

Data augmentation techniques contribute to mitigating overfitting by increasing the diversity of training samples and promoting more robust feature learning [12]. In medical imaging, these transformations enable the network to recognize patterns across different orientations and illumination conditions, which is essential since structures of interest may appear in various positions [13], [14].

In this work, the training images were subjected to a pipeline of geometric transformations, including vertical and horizontal flipping. These transformations were considered appropriate for dermoscopic images because skin lesions do not have a fixed anatomical orientation relative to the camera during image acquisition. Therefore, changes in the spatial orientation of the lesions do not necessarily alter their relevant morphological characteristics. Both the vertical and horizontal flipping transformations were applied with a probability of 0.5, increasing the variability of the training samples while preserving relevant visual characteristics of the lesions.

C. Transfer Learning and Fine-Tuning

This work proposes an end-to-end approach for effective melanoma classification. Publicly available dermoscopic images from the ISIC dataset were utilized, and a binary classification task (melanoma vs. non-melanoma) was performed using the ResNet50 architecture combined with deep learning optimization strategies. An overview of the network architecture is illustrated in Fig. 1.

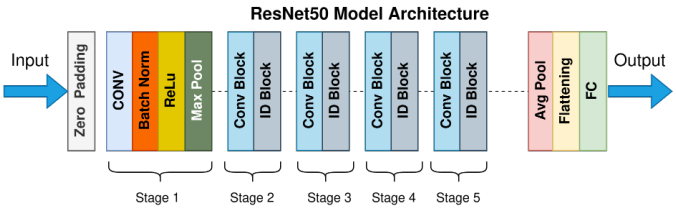


Fig. 1. Overview of the ResNet50 architecture. Source: [15]

Given the limited size of medical imaging datasets, training a deep network from scratch often leads to sub-optimal results or premature convergence. To mitigate this, transfer learning was incorporated by leveraging weights pretrained on the ImageNet dataset, significantly reducing convergence time and improving feature adaptation [16].

Subsequently, fine-tuning was employed to adapt the pretrained network to the target medical domain. This process consisted of unfreezing the deeper layers of the ResNet50 model and retraining them using the ISIC dataset with a reduced learning rate. Consequently, the previously learned generic weights were slightly adjusted to capture domain-specific features of melanocytic lesions while preserving the structural visual hierarchy acquired during the initial training phase [16].

The classifier was optimized using AdamW with a learning rate of 5e-4. Subsequently, during the fine-tuning stage, the layer3 and layer4 of the ResNet50 backbone were unfrozen and jointly trained with the classification layer using a learning rate of 1e-6 and 1e-5, respectively. The training was performed with a batch size of 64 for a maximum of 85 epochs. Early stopping was employed based on the validation loss, using a patience of 15 epochs, and the model checkpoint corresponding to the lowest validation loss was retained for evaluation. To address the class imbalance between melanoma and non-melanoma samples, Weighted Random Sampler was employed during training. The final classification was performed using a decision threshold of 0.5, whereby samples with a predicted melanoma probability equal to or greater than this value were classified as melanoma.

D. Evaluation Strategy

For the proposed application, clinical recall is prioritized over precision. Recall measures the model's ability to correctly identify positive cases, which is critical in melanoma detection, where a false negative (misclassifying a melanoma as benign) can have fatal clinical consequences. Conversely, a false positive can be subsequently reviewed and discarded by a specialist. Thus, maximizing recall stands as the primary optimization objective.

III. Results

The evaluation of the baseline configuration where ResNet50 acts strictly as a frozen feature extractor reveals moderate classification capabilities, as presented in Table I. The model achieved an initial validation accuracy of 83.90%; however, the precision remained constrained at 54.10%, indicating a substantial rate of false positives, resulting in an F1-Score of 59.60%. This baseline diagnostic behavior is further detailed in the confusion matrix in Fig. 2, which maps the absolute sample counts alongside class-wise recall and precision rates, establishing the numerical benchmark from which optimization strategies were evaluated. Specifically, out of 905 melanoma validation samples, the baseline correctly identified 600 true positives, while generating 305 false negatives, 3,653 true negatives, and 509 false positives.

TABLE I
Results without Data Augmentation.

Accuracy	Recall	Precision	F1-Score
83.90%	66.30%	54.10%	59.60%

To improve upon the baseline, a data augmentation strategy consisting of horizontal and vertical flipping operations was incorporated into the training pipeline. The introduction of these transformations improved the validation accuracy from 83.90% to 87.70% and increased precision from 54.10% to 65.80%, as summarized in Table II, corresponding to an F1-Score of 65.60%.

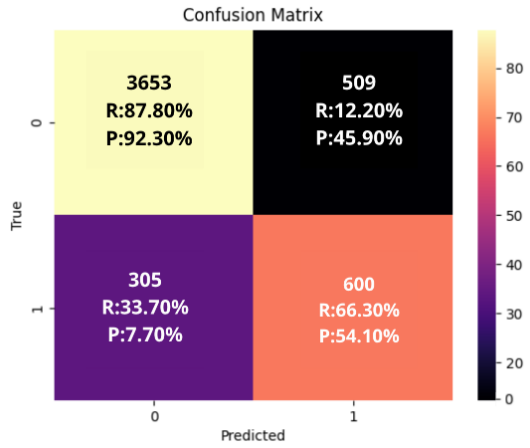


Fig. 2. Confusion matrix obtained without Data Augmentation. Source: Authors' own work.

However, a noticeable decrease in recall was observed, dropping from 66.30% to 65.30%. This behavior suggests that the model adopted a more conservative classification strategy, reducing false positives at the expense of a higher rate of false negatives within the melanoma class. The confusion matrix shown in Fig. 3 highlights this reduction in true positive predictions for the melanoma class, which directly correlates with the decreased recall, while concurrently demonstrating the reduction in false positives that drove the precision improvement. Under data augmentation, true positives dropped to 591 (increasing false negatives to 314), while true negatives improved to 3,855 and false positives decreased to 307.

TABLE II
RESULTS WITH DATA AUGMENTATION.

Accuracy	Recall	Precision	F1-Score
87.70%	65.30%	65.80%	65.60%

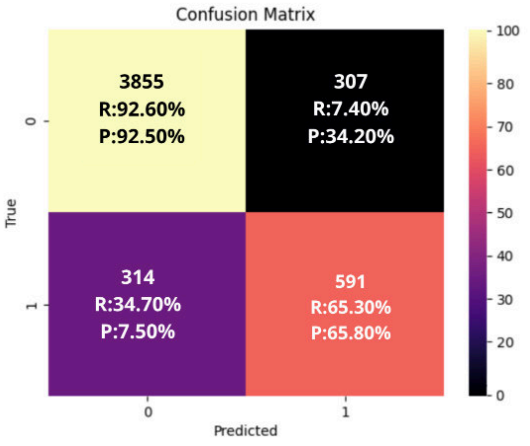


Fig. 3. Confusion matrix obtained with Data Augmentation. Source: Authors' own work.

Subsequently, fine-tuning was applied to the deeper layers of ResNet50 (`layer3` and `layer4`), while the initial layers responsible for extracting generic visual features remained frozen. Preliminary empirical evaluations indicated that unfreezing early layers did not yield significant performance gains; thus, this partial unfreezing configuration was adopted for the final experimental stage. The metrics achieved with this configuration are reported in Table III.

TABLE III
RESULTS WITH FINE-TUNING.

Accuracy	Recall	Precision	F1-Score
88.70%	74.70%	66.10%	70.20%

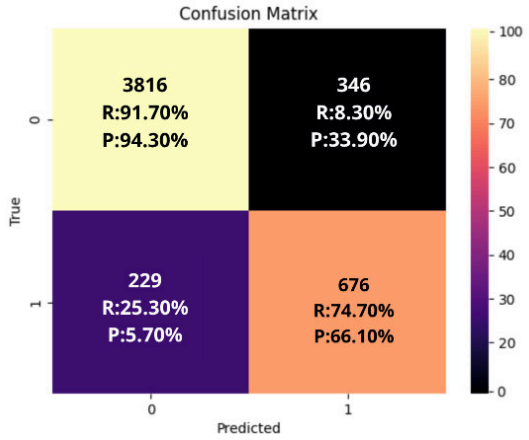


Fig. 4. Confusion matrix obtained with Fine-Tuning. Source: Authors' own work.

A comparative analysis across the three experimental setups indicates that fine-tuning the deeper layers of ResNet50 provides the most effective strategy among the evaluated configurations. While the introduction of data augmentation alone increased model precision (from 54.10% to 65.80%), it inadvertently degraded melanoma recall from 66.30% to 65.30%, causing the network to adopt an overly conservative decision boundary. However, incorporating partial fine-tuning successfully recovered and substantially boosted the recall to 74.70%, while maintaining a competitive precision of 66.10%, yielding an F1-Score of 70.20%, the highest among the three configurations.

Specifically, in terms of absolute sample counts, the fine-tuning configuration correctly identified 676 out of 905 melanoma cases (True Positives) and reduced false negatives to 229, while achieving 3,816 True Negatives and 346 False Positives for the non-melanoma class (Fig. 4).

From a clinical screening perspective, recall represents the primary optimization target. In automated melanoma detection, a false negative misclassifying a malignant lesion as non-melanoma carries severe, potentially fatal consequences by delaying life-saving medical intervention. Conversely, false positives can be subsequently re-evaluated and ruled out by a specialist through follow-up dermoscopic examinations

or biopsies. Although overall accuracy reached 88.70%, the 9.40 percentage-point improvement in recall is particularly noteworthy, as it directly translates to a significant reduction in false negative melanoma classifications. Among the evaluated configurations, the proposed approach achieved the best overall performance, particularly in terms of recall, indicating its potential for supporting early-stage screening

Finally, Table IV provides a benchmarking overview comparing the proposed approach against literature benchmarks. Although some existing frameworks report higher overall accuracy, these results should be interpreted with caution, as several studies were developed using smaller and more balanced datasets. For instance, some approaches were trained on datasets containing approximately 3,300 dermoscopic images that were already nearly balanced between benign and malignant lesions, considerably simplifying the classification task. In contrast, the proposed method was evaluated on the ISIC 2019 Training dataset, which contains 25,331 dermoscopic images and presents a substantial class imbalance, making melanoma detection significantly more challenging. Despite these more demanding conditions, the proposed method establishes a competitive trade-off, achieving 88.70% accuracy while specifically optimizing clinical recall, which is a vital parameter for early screening systems where reducing false negatives is critical.

TABLE IV
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS.

Reference	Year	Datasets	Accuracy	Recall
[7]	2024	ISIC	97.00%	99.00%
[8]	2024	ISIC 2018	97.00%	91.00%
[9]	2024	HAM10000	89.00%	91.00%
[10]	2025	ISIC 2020	95.00%	95.24%
[11]	2024	ISIC	93.50%	77.00%
Proposed Method	**2026**	**ISIC 2019**	**88.70%**	**74.70%**

IV. CONCLUSION

Melanoma remains one of the most aggressive forms of skin cancer, highlighting the critical importance of developing computational tools to support early diagnosis. In this work, a ResNet50-based approach was proposed for the binary classification of dermoscopic images, integrating transfer learning, data augmentation, and fine-tuning techniques.

The experimental results demonstrated that transfer learning enables satisfactory performance even when dealing with imbalanced datasets. The application of data augmentation supported feature variability, while fine-tuning the deeper layers of the network yielded the highest classification performance among the evaluated configurations, with improvements in recall and precision.

Overall, the empirical findings indicate that combining transfer learning, data augmentation, and fine-tuning constitutes a promising strategy for automated melanoma detection in dermoscopic images. The proposed framework has the potential to serve as a supportive preliminary tool for clinical decision-making and early screening.

Future research will focus on evaluating the model using external datasets and cross-validation protocols, investigating alternative deep architectures, exploring advanced class-balancing strategies, and incorporating explainable artificial intelligence (XAI) techniques to enhance interpretability for clinical practice.

REFERENCES

[1] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, N. Li, et al., "Cancer statistics in china and united states, 2022: Profiles, trends, and determinants," *Chinese medical journal*, vol. 135, no. 05, pp. 584–590, 2022.

[2] W. H. Organization et al., *Radiation*. World Health Organization, 1969.

[3] S. Ghosh, A. Singh, and S. Kumar, "Multiplier leadership optimization algorithm (mloa): Unconstrained global optimization approach for melanoma classification," *Discover Internet of Things*, vol. 5, no. 1, p. 70, 2025.

[4] A. A. Marghoob and N. Jaimes, "Overview of dermoscopy," *UpToDate.[(accessed on 27 April 2021)]*, 2015.

[5] D. Kohoutova, D. Worku, H. Aziz, J. Teare, J. Weir E, and J. Larkin, "Melanoma maligno do trato gastrointestinal: Sintomas, diagnóstico e opções de tratamento atuais," *Cells*, vol. 10, no. 2, 2021.

[6] V. Madan, J. T. Lear, and R.-M. Szeimies, "Non-melanoma skin cancer," *The lancet*, vol. 375, no. 9715, pp. 673–685, 2010.

[7] R. Sabir and T. Mehmood, "Classification of melanoma skin cancer based on image data set using different neural networks," *Scientific Reports*, vol. 14, no. 1, p. 29 704, 2024.

[8] S. M. Thwin and H.-S. Park, "Skin lesion classification using a deep ensemble model," *Applied Sciences*, vol. 14, no. 13, p. 5599, 2024.

[9] K. M. Selvaraj, S. Gnanagurusubbiah, R. R. R. Roy, J. H. John Peter, and S. Balu, "Enhancing skin lesion classification with advanced deep learning ensemble models: A path towards accurate medical diagnostics," *Current Problems in Cancer*, vol. 49, p. 101 077, 2024.

[10] S. Padhy, S. Dash, N. Kumar, S. P. Singh, G. Kumar, and P. Moral, "Temporal integration of resnet features with lstm for enhanced skin lesion classification," *Results in Engineering*, vol. 25, p. 104 201, 2025.

[11] M. S. Islam and S. Panta, "Skin cancer images classification using transfer learning techniques," *arXiv preprint arXiv:2406.12954*, 2024.

[12] L. Zhao, T. Liu, X. Peng, and D. Metaxas, "Maximum-entropy adversarial data augmentation for improved generalization and robustness: Supplementary material,"

[13] R. Poojary, R. Raina, and A. K. Mondal, "Effect of data-augmentation on fine-tuned cnn model performance," *algorithms*, vol. 5, p. 6, 2021.

[14] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," *Advances in neural information processing systems*, vol. 34, pp. 29 935–29 948, 2021.

[15] S. Mukherjee, *The Annotated ResNet-50 — medium.com*, https://medium.com/data-science/the-annotated-resnet-50-a6c536034758, [Accessed 18-06-2026].

[16] R. L. Araújo, F. H. d. Araújo, and R. R. e. Silva, "Automatic segmentation of melanoma skin cancer using transfer learning and fine-tuning," *Multimedia Systems*, vol. 28, no. 4, pp. 1239–1250, 2022.

Disentangling Persistent ECG ST/T Alterations and Local Beat Perturbations through Multiscale Physiological Labeling and Recurrence Quantification Analysis

Renata R. Costa
Center of Engineering
Federal University of ABC (UFABC)
Santo André, Brazil
romanelli.c@ufabc.edu.br

Hiago B. Nunes
Center of Medical Sciences
Federal University of Pernambuco
Recife, Brazil
hiago.henrique@ufpe.br

Fernando J.R. Sales
Biomedical Engineering Department
Federal University of Pernambuco
Recife, Brazil
fernando.sales@ufpe.br

Diogo C. Soriano
Center of Engineering
Federal University of ABC (UFABC)
São Bernardo, Brazil
diogo.soriano@ufabc.edu.br

***Abstract*—ST-segment and T-wave (ST/T) abnormalities are important electrocardiographic markers of myocardial ischemia and ventricular repolarization disorders. Existing methods rely mainly on conventional morphological descriptors and do not distinguish persistent ST/T episodes from transient beat abnormalities, limiting interpretability. This work proposes a framework combining conventional ECG descriptors, Recurrence Quantification Analysis (RQA), and multiscale physiological labeling to characterize ST/T alterations. ECG recordings from the European ST-T Database were segmented into beat-based windows and evaluated at two distinct temporal scales to characterize persistent ST/T episodes and local beat abnormalities. Logistic regression evaluated the contribution of RQA features, while a mixed-effects multinomial model quantified associations across four physiological states. In Group 1, RQA increased episode classification accuracy from 50.0% (chance level) to 63.0% (+26.0% relative), with recall and F1-score increasing by 24.0% and 20.4%, respectively. In Group 2, AUC increased from 75.0% to 80.0% (+6.7% relative). Beat abnormality classification achieved 80.0% accuracy and 76.0% AUC in Group 1, and 83.0% accuracy and 84.0% AUC in Group 2. Mixed-effects analysis showed distinct dynamical signatures across physiological states, with average diagonal line length (L) showing the strongest positive association across altered classes and determinism (DET) exhibiting distinct associations with episodes and beat abnormalities. These findings demonstrate that RQA provides complementary information beyond conventional ECG descriptors and supports multiscale labeling for distinguishing persistent ST/T alterations from transient beat perturbations. Rather than maximizing predictive performance, classification was used as a comparative tool to demonstrate the value of recurrence descriptors and the interpretability gains of multiscale labeling.**

***Index Terms*—Electrocardiogram, Recurrence Quantification Analysis, Multiscale physiological labeling, Myocardial ischemia, Nonlinear dynamics.**

I. Introduction

Coronary artery disease remains one of the leading causes of mortality worldwide, and myocardial ischemia is among its most common manifestations [1]. Ischemic events alter the electrical properties of cardiac tissue, affecting the electrocardiogram (ECG), particularly the ST segment and T wave [1]. Consequently, the automatic detection and characterization of ST/T abnormalities have been extensively investigated to support early diagnosis and clinical decision-making [2]. However, these abnormalities are often subtle, patient-dependent, and susceptible to noise and baseline variations, making their robust characterization a challenging problem [1], [3]. Conventional approaches typically rely on morphological measurements extracted from predefined ECG landmarks, which may not fully capture the complexity of the underlying cardiac dynamics.

The European ST-T Database (EDB) provides a unique framework for studying this problem because it combines two complementary annotation levels [4]. In addition to expert annotations defining the onset, peak, and end of ST/T alteration episodes, the database provides beat-by-beat annotations describing local cardiac events. These annotations naturally represent different temporal scales of cardiac physiology: ST/T episodes describe persistent changes in the underlying myocardial state, whereas beat annotations characterize transient electrical perturbations occurring within that state. This distinction suggests that ECG windows should be interpreted simultaneously from both perspectives rather than using a single label.

Recurrence analysis provides an alternative framework for

characterizing ECG signals from the perspective of nonlinear dynamics. Recurrence Plots (RPs) represent the recurrence of cardiac states in phase space as a two-dimensional fingerprint of the underlying dynamics [5], [6], while Recurrence Quantification Analysis (RQA) summarizes these patterns through descriptors such as recurrence rate, determinism, entropy, and laminarity [7], [9]. Unlike conventional morphological descriptors, RQA quantifies the temporal organization and geometric structure of cardiac trajectories [10], potentially revealing subtle dynamical changes associated with ventricular repolarization abnormalities. Although recurrence analysis has demonstrated promising results in physiological signal analysis, its application to ST/T alteration episodes remains relatively unexplored.

During the exploratory analysis, we observed that beat abnormalities frequently occur both inside and outside annotated ST/T episodes. Consequently, models trained solely using episode annotations may inadvertently learn beat-related characteristics, reducing the interpretability of the resulting classifiers. Rather than treating these alterations as independent classes, we propose a multiscale hierarchical physiological labeling framework in which each ECG window is simultaneously described according to (i) the persistent ST/T episode state and (ii) local beat-level perturbations. The resulting four analysis groups arise naturally from the Cartesian product of these complementary physiological descriptors, allowing the influence of persistent and transient alterations to be analyzed separately. Accordingly, the objective of this work was twofold. First, we investigated whether recurrence-based descriptors provide complementary information to conventional ECG features for the characterization of ST/T alteration episodes. Second, we evaluated how beat-level abnormalities influence this characterization through the proposed physiological labeling framework. The classification models were adopted as a comparative framework rather than with the aim of maximizing predictive performance or competing with state-of-the-art methods. Instead, the primary contribution lies in assessing the added value of recurrence-based descriptors and in demonstrating how the proposed disentangling strategy and multiscale physiological labeling improve the physiological interpretability of ECG dynamics by separating persistent ST/T alterations from transient beat perturbations.

II. Methods

A. Dataset

The experiments were conducted using ECG recordings from the European ST-T Database (EDB), accessed via the PhysioNet platform. The database provides multi-lead ECG signals sampled at a fixed frequency of 250 Hz and annotated beat-by-beat by expert cardiologists using the WFDB annotation system. The ECG signals and annotations were loaded using the WFDB library, which enables synchronized access to waveform data (rdrecord/rdsamp) and reference annotations (rdann). Each record consists of continuous ECG signals along with temporal markers indicating cardiac events, which were used as ground truth for segmentation and labeling.

The original dataset consists of 90 annotated excerpts of ambulatory ECG recordings from 79 subjects. The subjects were 70 men aged 30 to 84, and 8 women aged 55 to 71. The subjects were 70 men aged 30 to 84, and 8 women aged 55 to 71. From the 90 records, 30 records were selected for this study, corresponding to the largest subset of recordings acquired using the same pair of leads, namely V4 and MLIII. A second subset of 14 records, corresponding to the second-largest group acquired using the same pair of leads, namely V5 and MLI, was also included in the study and analyzed separately due to the different lead configuration. The group with 30 records present exclusively male subjects, with a mean age of 52 ± 8 years. The group with 14 records comprised 13 men and 1 woman, with a mean age of 55 ± 10 years. Each ECG recording was sampled at 250 Hz and contains a total of 1,800,000 samples, corresponding to 120 minutes of continuous monitoring. Although this selection ensured lead consistency across all recordings, it limits the generalizability of the findings to populations with different lead configurations and to female subjects.

B. Exploratory Data Analysis (EDA)

An EDA was performed on the ECG dataset to characterize its structure, annotation scheme, and demographic composition before the experimental pipeline. The dataset includes two annotation types: beat-by-beat labels, which mark individual cardiac cycles, and episode-level annotations, indicating the onset, peak, and offset of clinically relevant ST/T deviations. These complementary annotations enable both local (beat-level) and global (episode-level) analyses of ECG dynamics and are summarized in Table I.

A grouping strategy based on ECG lead configuration was then applied to ensure methodological consistency. Records were grouped by available leads, and the largest homogeneous subgroup comprised 30 individuals with the same configuration (V4 and MLIII). This subgroups were selected for all subsequent analyses to ensure signal comparability and minimize variability arising from different lead configurations.

C. Signal Preprocessing

The raw ECG signals were subjected to the preprocessing pipeline illustrated in Fig. 1A (preprocessing) and Fig. 1B (R-Peaks).

- Power Spectral Analysis: Power Spectral Density (PSD) analysis was performed to characterize the frequency content of the signals, verify preprocessing effectiveness, and identify dominant frequency bands associated with cardiac activity.
- Baseline wander and high-frequency noise were reduced using digital filtering while preserving the morphology of the P, QRS, and T waves. A high-pass filter (0.5 Hz) and a low-pass filter (50 Hz), defined from the power spectrum analysis, were applied [11].
- Normalization: Signal amplitudes were standardized across recordings and leads using z-score normalization, ensuring zero mean and unit variance per channel. In

TABLE I
ANNOTATIONS DICTIONARY

Annotation	Description
Episode-level annotations	
N	Normal rhythm
ST0+	ST-segment elevation episode on lead 0
ST0–	ST-segment depression episode on lead 0
ST1+	ST-segment elevation episode on lead 1
ST1–	ST-segment depression episode on lead 1
T0+	T-wave elevation episode on lead 0
T0–	T-wave depression episode on lead 0
VT	Ventricular tachycardia
Beat-level annotations	
N	Normal beat
V	Premature ventricular contraction
S	Supraventricular premature or ectopic beat
+	Rhythm change
F	Fusion of ventricular and normal beat

multi-lead recordings, normalization preserved relative amplitude relationships between leads, avoiding distortion of morphological differences [12].

- R-Peak Detection: R-peaks were detected using the GQRS algorithm implemented in the WFDB processing library. Widely used in PhysioNet-based studies, GQRS identifies QRS complexes from signal morphology and amplitude variations to determine R-peak locations [13]. Detection was performed sequentially across available leads until a valid set of R-peaks was found, ensuring robust performance even when some leads presented lower signal quality or higher noise levels.

D. Signal windowing

After R-peak detection, ECG recordings were segmented using a heartbeat-based sliding-window approach. Each window contained 5 consecutive heartbeats, defined by the interval between the first and tenth detected R-peak, with an overlap of 2 heartbeats between adjacent windows. This overlap increased the number of samples while preserving temporal continuity, as illustrated Fig. 1C. This strategy ensured that each window contained a comparable amount of cardiac activity regardless of heart rate variations. Although window durations varied over time, all segments preserved the same number of cardiac cycles and their morphological information.

E. Features

- **Morphology-based features:** Morphology-based features were extracted from each ECG window to characterize cardiac rhythm and waveform morphology. Temporal dynamics were represented by the mean R-peaks interval (R-peaks-mean), reflecting the average cardiac cycle duration, and the standard deviation (R-peaks-std), measuring beat-to-beat variability. Signal energy was quantified using the Root Mean Square (RMS), a widely adopted measure of signal amplitude and energy. Because the selected records include two ECG leads (V4 and MLIII), RMS was calculated separately for each lead to preserve lead-specific amplitude information. Together, R-peaks-mean, R-peaks-std, and lead-specific RMS values formed the set of signal-derived features used in the subsequent analyses.
- **Recurrence-based Features:**
Recurrence-based features were extracted from each windowed segment using RQA, which characterizes temporal patterns in the joint evolution of the two ECG leads. RP provides a two-dimensional representation of trajectory recurrences in phase space by identifying states that revisit neighboring regions over time, acting as a dynamical fingerprint of the system. In this sense, the RP acts as a dynamical fingerprint of the system, revealing characteristic patterns associated with the way the trajectory revisits previously explored regions and evolves through phase space over time. Let $\mathbf{x}_i \in \mathbb{R}^m$ denote the state of the system at time i, represented either by a directly observed multivariate state vector or by a reconstructed phase-space vector. The recurrence matrix is defined as

$$R_{i,j} = \Theta\left(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|\right), \qquad i,j = 1,\ldots,N, \quad (1)$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|$ represents the distance between the states of the system at times i and j, ε is a predefined recurrence threshold, and $\Theta(\cdot)$ is the Heaviside function. Thus, $R_{i,j} = 1$ when two states are sufficiently close in phase space, according to the selected distance criterion and threshold, and $R_{i,j} = 0$ otherwise. The resulting RP encodes the temporal organization of the trajectory as a binary pattern whose isolated points and diagonal and vertical line structures form the basis for RQA [9].
For each window, the two ECG leads were combined into a multivariate embedding. The neighborhood radius was empirically set to $\varepsilon = 0.1$, based on qualitative inspection of the resulting RPs and the corresponding recurrence rate (RR). Specifically, the threshold was selected to provide visually meaningful recurrence structures. The resulting RRs were also evaluated in relation to the 5–15% range commonly suggested for recurrence-based analyses [8]. A Theiler correction of 1 sample excluded temporally adjacent points, reducing autocorrelation bias.
Based on the resulting RPs, six RQA descriptors were extracted to characterize the temporal organization and complexity of the reconstructed phase-space trajectories, following Marwan et al. [9]. Recurrence rate (RR) quantifies the density of recurrent states, with higher values indicating more frequent revisits to previously occupied regions of phase space. Determinism (DET) measures the proportion of recurrent points forming diagonal lines, where higher values indicate more predictable and temporally ordered dynamics. The average diagonal line length (L) reflects the mean duration of similar trajectory evolution, while the longest diagonal line (Lmax) represents the longest predictable episode; unlike RR and DET, both are expressed in samples or time units and have no fixed

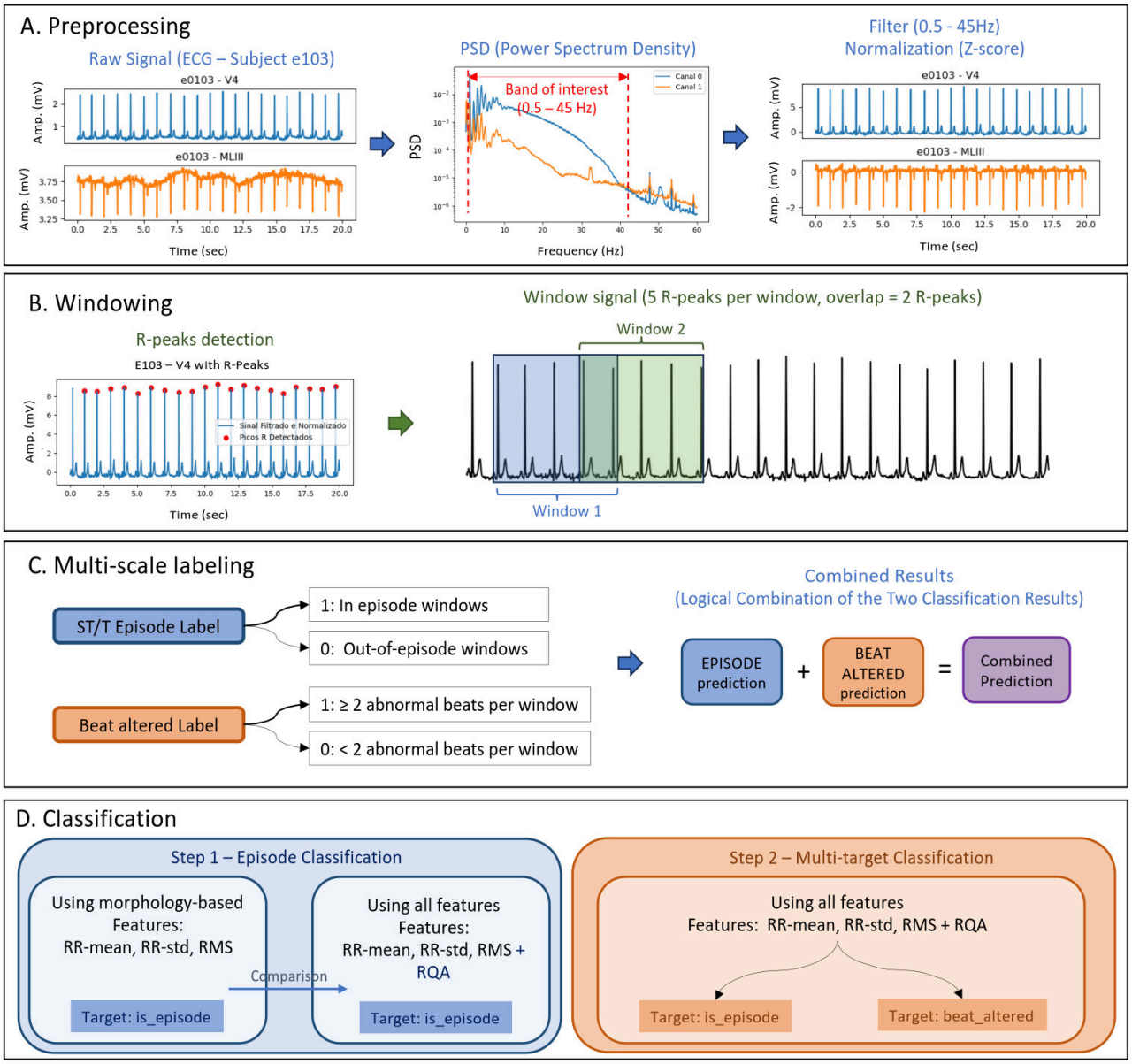


Fig. 1. Overview of the proposed pipeline for ischemic episode classification from ECG recordings. (A) Preprocessing stage. (B) Segmentation of the preprocessed ECG into overlapping windows. (C) Hierarchical labeling strategy combining episode annotations and beat-level alterations to generate four analysis classes. (D) Classification framework.

upper bound. The entropy of diagonal line lengths (Lentr) quantifies the diversity of diagonal-line durations, with higher values indicating greater heterogeneity of recurrent temporal patterns. Finally, laminarity (LAM) measures the proportion of recurrent points forming vertical lines, reflecting the persistence of laminar states, with higher values indicating more sustained dynamics within restricted regions of phase space.

F. Multi-scale physiological labeling

The EDB Database provides two complementary annotation levels describing cardiac activity at different temporal scales. ST/T episode annotations identify persistent ventricular repolarization alterations over relatively long intervals, whereas beat annotations capture transient electrical events on a beat-by-beat basis.

Each ECG window was simultaneously characterized by these two physiological descriptors, reflecting that local beat perturbations occur within an underlying myocardial state. Their combination naturally defines four physiological groups: (i) normal beat outside an ST/T episode, (ii) altered beat outside an ST/T episode, (iii) normal beat during an ST/T episode, and (iv) altered beat during an ST/T episode.

The `is_episode` label was generated from the episode-level annotations in Table I. A window was labeled positive

(is_episode = 1) if it overlapped any annotated ST/T episode; otherwise, it was labeled negative (is_episode = 0).

The beat_altered label was derived from the beat-level annotations in Table I. A window was labeled as altered (beat_altered = 1) if at least two beats contained abnormal symbols; otherwise, it was labeled normal (beat_altered = 0).

The multiscale labeling is illustrated in Fig. 1C and enables persistent episode-related alterations and transient beat abnormalities to be analyzed separately, providing a more interpretable representation of cardiac dynamics.

G. Classification and Evaluation

Two independent binary classification models were trained to investigate complementary aspects of cardiac dynamics. The first distinguished ECG windows belonging to ST/T alteration episodes from normal segments, characterizing persistent myocardial changes. The second classified windows according to the presence or absence of beat abnormalities, representing transient electrical perturbations.

Both classifiers used the same feature extraction pipeline and evaluation protocol, enabling recurrence-based descriptors to be assessed independently for each physiological phenomenon. Their outputs were then combined through the proposed multiscale labeling framework to define the four physiological analysis groups, rather than by training a dedicated four-class classifier.

The classification was organized into two steps to evaluate the impact of feature representation and label type, as illustrated in Fig. 1D. All models were implemented using logistic regression with L2 regularization and evaluated under a stratified 5-fold cross-validation scheme. The data were split into 75% training and 25% testing sets using a group-wise split by record, ensuring that all windows from the same record were assigned exclusively to one set to prevent data leakage.

*a) **Step 1—Episode-based classification**:* In the first step, a binary classifier was trained using the is_episode label, representing the presence or absence of ST/T episodes. Two feature sets were evaluated independently:

- Morphological features only, including mean and standard deviation of R-peaks intervals and RMS values per lead;
- Combined feature set, integrating morphological features with RQA descriptors.

This comparison isolated the contribution of nonlinear dynamical information provided by RQA features. Model performance was evaluated using accuracy, precision, recall, and ROC curves.

*b) **Step 2—Multi-scale classification** (is_episode and beat_altered):* In the second step, the role feature set (morphological + RQA) was used to train two separate classification models with distinct targets:

- is_episode, representing clinically annotated ST/T segment alteration episodes;
- beat_altered, representing beat abnormalities.

This stage investigated how different physiological scales are represented within the same feature space.

The joint analysis of both steps provides a comprehensive interpretation of ECG behavior by identifying four complementary signal regimes:

- Normal: Out-of-episode with normal beats;
- Beat: Out-of-episode with altered beats;
- Episode: In-episode with normal beats;
- EpisodeBeat: In-episode with altered beats.

H. Mixed-effects analysis of disentangled ECG dynamics

To further investigate whether recurrence-based descriptors discriminate sustained ischemic episodes from local beat abnormalities, a mixed-effects multinomial logistic regression was fitted considering the four combined classes (Normal, Episode, Beat, and EpisodeBeat). The Normal class was adopted as the reference level, while patient-specific variability was modeled through a random intercept for each record.

This analysis was restricted to the 30-record subset with the V4 and MLIII lead configuration.

The model was implemented in R using the brms package with a categorical likelihood and logit link. Parameters were estimated using Hamiltonian Monte Carlo (No-U-Turn Sampler) with two Markov chains, 1,000 iterations per chain (500 warm-up iterations). Posterior means and 95% credible intervals were computed, with convergence assessed using $\hat{R}$ and the effective sample size (ESS). Regression coefficients were exponentiated to obtain odds ratios (OR) for interpretation. The model can be written as in

$$\log\left(\frac{P(Y_i = k)}{P(Y_i = \text{Normal})}\right) = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_{ij} + b_{r(i)}, \quad (2)$$

where Y_i denotes the class of the i-th ECG window, $k \in \{\text{Episode}, \text{Beat}, \text{EpisodeBeat}\}$, x_{ij} represents the j-th predictor, β_{jk} are the fixed-effect coefficients, and $b_{r(i)} \sim \mathcal{N}(0, \sigma^2)$ is the random intercept associated with the record containing observation i.

III. Results

A. Features

Table II summarizes the extracted features by type of each group of records, with their mean and standard deviation across the four hierarchical label combinations. This organization facilitates comparison of feature distributions across physiological conditions and highlights how signal-derived and recurrence-based descriptors vary with episode presence and beat alterations.

B. Classification and Evaluation

Table III summarizes the classification results of the proposed multiscale framework. The first scale identifies ST/T alteration episodes (is_episode), while the second classifies local beat abnormalities (beat_altered).

Figure 2 illustrates the multiscale framework by combining the outputs of the two binary classifiers into the four

TABLE II
FEATURES DESCRIPTION (MEAN ± STD).

Feature	Group 1 (Leads V4 and MLIII)				Group 2 (Leads V5 and MLI)			
	In-ST/T Epi.	Out-ST/T Epi.	Altered Beat	Normal Beat	In-ST/T Epi.	Out-ST/T Epi.	Altered Beat	Normal Beat
Morphological features								
RR mean (sample)	886.9 ± 220.1	926.0 ± 213.9	895.1 ± 215.9	920.4 ± 215.5	632.0 ± 149.4	725.6 ± 169.2	703.3 ± 144.4	709.7 ± 170.9
RR std (sample)	35.3 ± 72.6	35.6 ± 66.1	96.3 ± 138.2	30.7 ± 55.3	34.5 ± 73.26	26.4 ± 54.2	92.5 ± 111.4	24.15 ± 51.15
RMS V4 (mV)	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.2	0.3 ± 0.1	-	-	-	-
RMS MLIII (mV)	0.3 ± 0.5	0.2 ± 0.2	0.4 ± 0.6	0.2 ± 0.3	-	-	-	-
RMS V5 (mV)	-	-	-	-	0.25 ± 0.09	0.24 ± 0.09	0.27 ± 0.15	0.24 ± 0.08
RMS MLI (mV)	-	-	-	-	0.27 ± 0.09	0.23 ± 0.08	0.29 ± 0.14	0.23 ± 0.07
RQA features								
RR (dimless)	0.06 ± 0.08	0.05 ± 0.05	0.03 ± 0.04	0.05 ± 0.06	0.03 ± 0.02	0.05 ± 0.03	0.03 ± 0.02	0.05 ± 0.03
DET (dimless)	0.82 ± 0.18	0.85 ± 0.16	0.76 ± 0.21	0.85 ± 0.16	0.74 ± 0.15	0.84 ± 0.13	0.73 ± 0.19	0.83 ± 0.14
L (rec. points)	6.33 ± 4.96	5.36 ± 2.65	5.16 ± 3.47	5.58 ± 3.22	3.94 ± 1.20	4.95 ± 1.19	4.22 ± 1.74	4.80 ± 1.86
$L_{\max}$ (rec. points)	115.5 ± 84.5	115.9 ± 68.2	107.4 ± 83.6	116.5 ± 70.5	51.1 ± 36.3	94.5 ± 59.2	78.2 ± 61.5	87.5 ± 58.1
L_{entr} (dimless)	2.08 ± 0.73	2.04 ± 0.55	1.86 ± 0.65	2.06 ± 0.58	1.66 ± 0.39	1.97 ± 0.49	1.70 ± 0.50	1.93 ± 0.48
LAM (dimless)	0.88 ± 0.14	0.91 ± 0.12	0.84 ± 0.16	0.91 ± 0.12	0.84 ± 0.12	0.91 ± 0.09	0.83 ± 0.14	0.90 ± 0.09

Abbreviations: dimless: Dimensionless. Rec. points: Recurrence points. Epi.: Episode.

TABLE III
OVERALL CLASSIFICATION PERFORMANCE FOR GROUPS 1 AND 2.

Group	Model	Acc. (%)	AUC (%)	Prec. (%)	Rec. (%)	F1 (%)
Group 1	Episode (Morph.)	50.0	64.0	70.0	50.0	54.0
	Episode (Morph.+RQA)	63.0	64.0	70.0	62.0	65.0
	Altered Beat (Morph.+RQA)	80.0	76.0	89.0	79.0	83.0
Group 2	Episode (Morph.)	83.0	75.0	81.0	82.0	81.0
	Episode (Morph.+RQA)	83.0	80.0	83.0	82.0	82.0
	Altered Beat (Morph.+RQA)	83.0	84.0	89.0	83.0	85.0

Abbreviations: Acc., Accuracy; AUC, Area Under the ROC Curve; Prec., Precision; Rec., Recall; Morph., Morphological features; RQA, Recurrence Quantification Analysis.

hierarchical physiological regimes. Example ECG windows from leads V4 and MLIII are shown together with their recurrence representations, enabling direct comparison between signal morphology and nonlinear dynamics across the different physiological conditions.

The distribution of the four label combinations is summarized in Table IV, which reports the number and percentage of windows in each category.

TABLE IV
COMBINED CLASSES.

Class	Description	No. of windows	Percentage (%)
Group 1 (V4 and MLIII)			
c1	Episode + Altered Beats	3411	16.66
c2	Episode + Normal Beats	4169	20.36
c3	Outside Episode + Altered Beats	560	2.73
c4	Outside Episode + Normal Beats	12338	60.25
Group 2 (V5 and MLI)			
c1	Episode + Altered Beats	1574	12.40
c2	Episode + Normal Beats	552	4.35
c3	Outside Episode + Altered Beats	745	2.73
c4	Outside Episode + Normal Beats	9824	77.38

C. Mixed Effects Analysis

The model was fitted to the complete dataset to characterize the contribution of each descriptor to the proposed disentangling strategy. Convergence and sampling diagnostics were assessed using the potential scale reduction factor ($\hat{R}$) and the effective sample size (ESS). Across all model parameters, $\hat{R}$ values ranged from 1.00 to 1.03, while Bulk ESS and Tail ESS ranged from 172 to 1549 and from 272 to 924, respectively. These diagnostics indicated generally adequate chain mixing, although the slightly elevated $\hat{R}$ observed for the Episode intercept ($\hat{R} = 1.03$) was noted.

For illustration, the linear predictor for the Episode class of patient `e0119` is given by

$$\begin{aligned}\eta_{\text{Episode}}^{(\text{e0119})} = & -1.84 - 0.01\,R - peak_{\text{mean}} - 0.09\,R - peak_{\text{std}} \\ & - 0.77\,RR + 1.73\,DET + 2.07\,L \\ & - 0.62\,L_{\max} - 1.97\,L_{\text{entr}} - 1.00\,LAM \\ & + 0.57\,RMS_{V4} - 0.03\,RMS_{MLIII} + b_{\text{e0119}}.\end{aligned} \tag{3}$$

where b_{e0119} represents the patient-specific random intercept. Similar equations are estimated for the Beat and EpisodeBeat classes, allowing each feature to have a distinct contribution depending on the underlying physiological condition.

Overall, the mixed-effects multinomial analysis revealed distinct associations between recurrence-based descriptors and the Episode, Beat, and EpisodeBeat classes. The corresponding odds ratios and 95% credible intervals are shown in the forest plots (Fig. 3). Among the evaluated descriptors, L showed the strongest positive association across altered classes, while

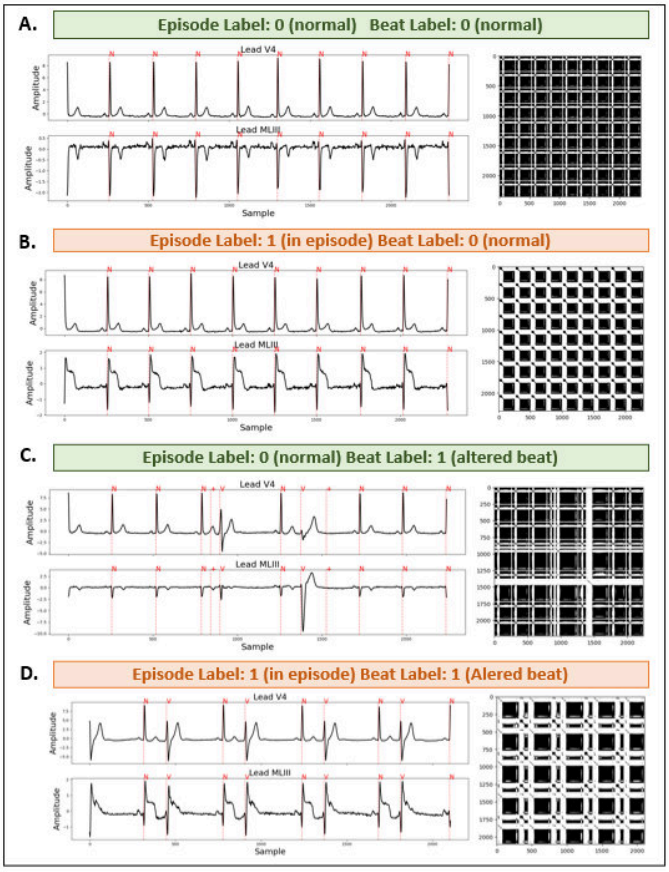


Fig. 2. Examples of ECG signals and recurrence plots for each label combination.

DET exhibited opposite effects for the Episode and Beat classes.

IV. Discussion

The results demonstrate that combining conventional ECG descriptors with recurrence-based features provides a richer characterization of ECG signals than using morphological features alone.

In Step 1, the inclusion of RQA metrics improved all classification measures, increasing accuracy from 0,5 to 0,63 (+26% relative) in group 1 where the initial 50.0% accuracy corresponds to chance-level classification. In Group 2, AUC increased from 75.0% to 80.0% (+6.7% relative). In Group 1, precision remained unchanged, whereas recall and F1-score increased by 24.0% and 20.4%, respectively. These results suggest that RQA descriptors capture temporal organization and dynamical complexity not fully represented by conventional ECG descriptors, supporting their potential for characterizing alterations across different temporal scales. These findings are consistent with previous studies [10], [14], [15].

The mixed-effects multinomial model further demonstrated that RQA descriptors capture distinct dynamical patterns associated with the four physiological states while accounting for repeated observations from the same subject. The average diagonal line length (L) showed the strongest positive association with all altered classes, particularly classes Episode and EpisodeBeat, indicating that prolonged recurrent trajectories are a relevant marker of abnormal ECG dynamics. Since L represents the persistence of similar state evolution in the reconstructed phase space, its association suggests that sustained ST/T episodes and combined alterations are characterized by increased temporal persistence rather than isolated waveform changes.

The remaining RQA metrics showed class-dependent behavior, supporting the proposed disentangling strategy. DET was positively associated with Episode but negatively associated with Beat class, suggesting that sustained ST/T abnormalities are related to more deterministic temporal organization, whereas isolated beat alterations exhibit more irregular dynamics. Conversely, LAM showed the opposite pattern, being strongly associated with class Beat and negatively related to Episode class. These findings indicate that beat-related abnormalities do not represent weaker manifestations of ischemic episodes, but rather a distinct dynamical regime. The EpisodeBeat class presented intermediate behavior, suggesting that the coexistence of both phenomena modifies the recurrence structure observed when each alteration occurs independently.

Conventional ECG descriptors also remained relevant in the analysis. The positive association of RMS-V4 across altered classes confirms that morphological information continues to contribute to ECG characterization, while the opposite effects of R-peak-std between Episode and Beat classes agree with the physiological interpretation of the labels. Sustained ST/T episodes tend to occur without pronounced beat-to-beat variability, whereas beat alterations are associated with increased rhythm fluctuations. Thus, RQA descriptors complement rather than replace conventional ECG features, providing a broader representation of cardiac dynamics.

An important contribution of this study is the multi-scale labeling strategy that separates persistent ST/T alterations from local beat perturbations, combined with RQA to characterize their distinct temporal behaviors. Traditional approaches often rely exclusively on episode annotations, assuming that relevant ECG variations are restricted to clinically defined events. However, EDA revealed substantial variations both inside and outside episode boundaries. By separating episode-related and beat-related alterations, the proposed framework avoids combining different physiological processes into a single classification task and enables a more detailed interpretation of ECG dynamics.

The analysis of the four label combinations further supports this interpretation. Windows classified as neither episode nor altered beat represent baseline cardiac behavior, whereas altered beats outside episodes indicate electrophysiological changes not captured by ST/T annotations alone. Conversely, episodes without beat alterations show that clinically relevant ST/T abnormalities may occur independently of rhythm disturbances. Finally, windows containing both alterations represent the most complex physiological condition, combining persistent and local abnormalities. This decomposition provides

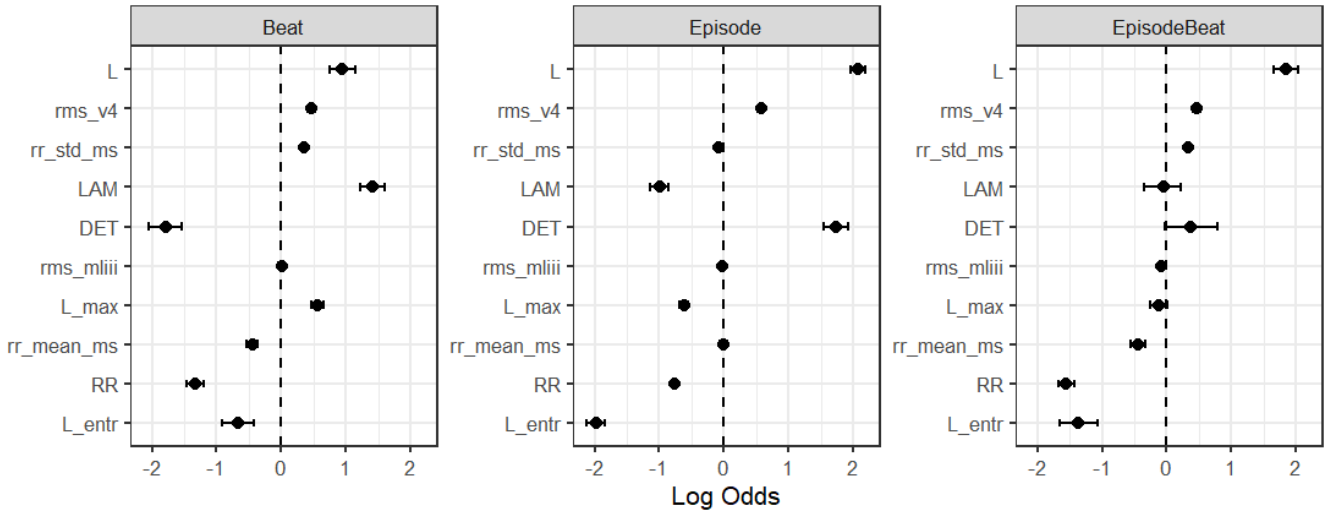


Fig. 3. Forest plots of the mixed-effects multinomial logistic regression for the Beat, Episode, and EpisodeBeat classes relative to the Normal class. Points denote the estimated log-odds coefficients and horizontal bars represent the 95% credible intervals. The dashed vertical line indicates the null effect (log-odds = 0).

a more comprehensive framework for interpreting classifier decisions and understanding the underlying ECG dynamics.

V. Conclusion

Overall, the findings suggest that combining morphological metrics, RQA, and hierarchical multi-scale labeling provides a comprehensive methodology for studying ECG abnormalities. The proposed approach not only improves the characterization of ST/T alteration episodes but also highlights the ability of recurrence analysis to characterize ECG morphology across different temporal scales. This capability enables the differentiation of transient beat-level abnormalities from more persistent ST/T episodes, providing a more interpretable representation of ECG dynamics and contributing to the development of more explainable ECG analysis systems. Given the limited number of individuals within each lead configuration, future work will extend the analysis to a larger number of individuals and additional ECG datasets to further assess the generalizability of the proposed framework.

Acknowledgment

DCS acknowledges support from CNPq (31397/2023-8).

References

[1] L. Pang *et al.*, "Real time heart ischemia detection in the smart home care system," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2005, pp. 3703–3706, doi:10.1109/IEMBS.2005.1617286.

[2] J. Park, W. Pedrycz, and M. Jeon, "Ischemia episode detection in ECG using kernel density estimation, support vector machine and feature selection," *Biomed. Eng. Online*, vol. 11, p. 30, Jun. 2012, doi:10.1186/1475-925X-11-30.

[3] N. Maglaveras, T. Stamkopoulos, C. Pappas, and M. G. Strintzis, "An adaptive backpropagation neural network for real-time ischemia episodes detection: Development and performance analysis using the European ST-T database," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 7, pp. 805–813, Jul. 1998, doi:10.1109/10.686788.

[4] A. Taddei *et al.*, "The European ST-T Database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *Eur. Heart J.*, vol. 13, pp. 1164–1172, 1992.

[5] J. P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhys. Lett.*, vol. 4, no. 9, pp. 973–977, 1987, doi:10.1209/0295-5075/4/9/004.

[6] D. C. Soriano, R. Suyama, and R. Attux, "Blind extraction of chaotic sources from mixtures with stochastic signals based on recurrence quantification analysis," *Digit. Signal Process.*, vol. 21, no. 3, pp. 417–426, 2011.

[7] C. L. Webber, Jr., and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," *Phys. Lett. A*, vol. 223, nos. 1–2, pp. 55–60, 1996, doi:10.1016/S0375-9601(96)00741-4.

[8] S. Wallot and G. Leonardi, "Analyzing multivariate dynamics using cross-recurrence quantification analysis (CRQA), diagonal-cross-recurrence profiles (DCRP), and multidimensional recurrence quantification analysis (MdRQA) – A tutorial in R," *Front. Psychol.*, vol. 9, Art. no. 2232, 2018, doi:10.3389/fpsyg.2018.02232.

[9] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Phys. Rep.*, vol. 438, nos. 5–6, pp. 237–329, 2007, doi:10.1016/j.physrep.2006.11.001.

[10] H. B. Nunes, D. C. Soriano, and F. J. R. Sales, "Optimizing embedding spaces for recurrence-based analysis of multivariate cardiac signals: Myocardial infarction detection from 12-lead ECG data," *J. Precision Medicine: Health Disease*, vol. 3, Art. no. 100021, 2025, doi: 10.1016/j.premed.2025.100021.

[11] S. K. Berkaya *et al.*, "A survey on ECG analysis," *Biomed. Signal Process. Control*, vol. 43, pp. 216–235, 2018, doi:10.1016/j.bspc.2018.03.003.

[12] E. Iammarino, I. Marcantoni, A. Sbrollini, M. Morettini, and L. Burattini, "Normalization of electrocardiogram-derived cardiac risk indices: A scoping review of the open access literature," *Appl. Sci.*, vol. 14, no. 20, Art. no. 9457, 2024, doi:10.3390/app14209457.

[13] J. A. Behar, A. A. Rosenberg, I. Weiser-Bitoun, *et al.*, "PhysioZoo: A novel open access platform for heart rate variability analysis of mammalian electrocardiographic data," *Front. Physiol.*, vol. 9, Art. no. 1390, Oct. 2018, doi:10.3389/fphys.2018.01390.

[14] P. Tiago *et al.*, "Characterization of human persistent atrial fibrillation electrograms using recurrence quantification analysis," *Chaos*, vol. 28, no. 8, Art. no. 085710, Aug. 2018, doi:10.1063/1.5024248.

[15] G. Luongo, S. Schuler, A. Luik, *et al.*, "Non-invasive characterization of atrial flutter mechanisms using recurrence quantification analysis on the ECG: A computational study," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 3, pp. 914–925, Mar. 2021, doi:10.1109/TBME.2020.2990655.

Divergent Modulation of Cortical and Intermuscular Coupling by Object Properties During Grasping

1st Júlia Lombardi
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-8137-2647

2nd Solano Jorenti Jacyntho
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0002-2483-6024

3rd Karoline Adelaide Francisco Barbosa
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0002-7939-8551

4th Alcimar Barbosa Soares
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-1100-3533

***Abstract*—Advances in quantitative, sensitive, and objective biomarkers of neuromotor control are essential for early diagnosis and optimized rehabilitation following neurological and motor impairments. In this context, corticomuscular coherence (CMC) and intermuscular coherence (IMC) have been investigated as measures of brain–muscle communication and muscular coordination, respectively, yet their differential sensitivity to the physical properties of manipulated objects has not been systematically explored within a single paradigm. This study examined the modulation of CMC and IMC during a precision grip-and-hold task as a function of object weight and surface friction, using the WAY-EEG-GAL dataset (12 healthy subjects, 4 EEG channels, 5 forearm and hand muscles, 6 experimental conditions). Behaviorally, grip force and safety margin scaled with decreasing friction without concurrent changes in load force, confirming selective adjustment of the tangential force component. Neurophysiologically, the CMC peak proved to be the most sensitive variable, with surface friction accounting for 51.31% of total variance in the Beta band ($p < 0.0001$), uniformly across all 20 channel–muscle pairs evaluated. Conversely, IMC was predominantly modulated by object weight, with the heaviest condition (660g) significantly increasing coherence between forearm synergist muscles in the Gamma band. This dissociation — cortical descending coupling driven by precision demands, intermuscular coordination driven by force demands — is reported here for the first time. These findings support the complementary use of CMC and IMC as functionally distinct biomarkers, with potential implications for neurorehabilitation and prosthetic control strategies.**

***Index Terms*—CMC, IMC, muscle synergy, grasp**

I. Introduction

In 2024, a survey by the World Health Organization (WHO) highlighted the persistent impacts of neurological injuries on public health and the quality of life of those who suffer from them, whether due to persistent deficits impacting the individual's functionality or their need for rehabilitation and prolonged care [1]. Given this scenario, the improvement of tools that allow for the individualization of therapeutic planning and diagnostic anticipation, in a quantitative and parameterizable way, becomes fundamental, since traditional methods such as DASH [2] and the Rosén Score [3] demonstrate subjectivity and low sensitivity in detecting subtle changes during the process of neural degeneration and regeneration, in addition to a considerable impact stemming from their subjective nature. Based on this need, an understanding of the mechanisms of action of neurorehabilitation techniques [4], [5], and of the anatomical-functional remodeling of ascending and descending pathways of primordial phase is crucial. Among the various tools available, in the field of electrophysiology, Corticomuscular Coherence (CMC) and Intermuscular Coherence (IMC) stand out. Obtained through the mathematical combination of cerebral electrophysiological signals, captured through electroencephalography (EEG), and muscle activity, recorded by electromyography (EMG), CMC reflects the functional connectivity in a synchronous, although not causal, way between the motor cortex and the target muscle [6], [7]. In healthy individuals, CMC is predominantly observed in the mu (8–13 Hz) and beta (15–30 Hz) frequency bands, indicating both downward motor modulation and upward sensory feedback during motor tasks [8]. Studies suggest that alterations found in this synchronization may be consequences of remodeling of spinal pathways, cortical structure, and/or neuromuscular junctions (NMUs), potentially characterizing various pathological conditions such as neuropathies and lesions of the central nervous system [9]–[11]. The IMC, in turn, analyzes the functional connectivity between muscles and the shared common drive between pools of motor neurons through EMG data. It is based on the principle of synergistic muscle control, in which a shared oscillatory control signal (common drive) is sent from the motor cortex (M1) to different spinal motor neuron nuclei, "simplifying" the complexity in the activation of the musculoskeletal system [12]. The IMC can exhibit coherence in different frequency bands, reflecting different levels of control: the alpha/mu band is associated with physiological microtremors, cerebellar modulation, and continuous sensory feedback; the beta band is the classic signature of corticospinal

This work was supported in part by CNPq, CAPES, FAPEMIG, and FINEP.

tract modulation during isometric contractions and postural holds; the gamma band, finally, is associated with rapid adjustments, linked to dynamic coordination that demands speed and a high level of sensorimotor integration [13]. In general, the potential for mapping central and peripheral pathways revealed by the joint implementation of both tools is evident, expanding the possibilities for understanding the mechanisms responsible for movement. However, despite this potential, the integrated behavior of CMC and IMC and their combined interpretation in different task contexts remain insufficiently explored.

Beyond being a neuromotor biomarker, both CMC and IMC stand out as fundamental components for the bio-inspired functioning of human-machine interfaces and the development of more efficient assistive technologies, as indicated by works such as [14], [7], [15], and [16]. In these studies, the ability of CMC to act not only as a marker of evolution during rehabilitation, but also of corticomotor remodeling itself, was demonstrated; as was the correlation between CMC and IMC in the exposure of pathological compensations of the motor system; and IMC as a potential tool in reducing energy expenditure and assisting in kinematic calculations in prosthetic robotics, these being just some demonstrations of the potential of this research field. In particular, it remains unknown whether these two measures are differentially sensitive to distinct mechanical constraints, such as friction and load, within a single experimental paradigm, or whether they reflect a common underlying modulation.

In this context, this study proposes to explore the correlations between CMC and IMC in healthy individuals performing a grasping task. However, it remains unclear whether these measures exhibit distinct sensitivities to different task demands, such as variations in friction and load. The dataset used (WAY-EEG-GAL, [17]) allows the exploration of these correlations through controlled changes in texture and weight of the handled object, in addition to tracking the force applied by the volunteer. In this context, this study aims to investigate how corticomuscular coupling and intermuscular coordination are differentially modulated during object manipulation. By doing so, we seek to clarify their complementary roles as neuromotor biomarkers under varying mechanical constraints. To our knowledge, this is the first study to demonstrate a functional dissociation between CMC and IMC modulation within a unified grip control paradigm, with friction selectively driving cortical coupling and load selectively driving intermuscular coordination.

II. Methodology

A. Dataset

The dataset is the WAY-EEG-GAL, obtained from Luciw et al. [17], and consists of synchronized EEG, surface EMG (sEMG), force, and biomechanical data acquired during a standardized reach-to-grasp task with semi-random variations in object weight and surface texture. For the present analysis, only electrophysiological data were considered. EEG signals were recorded using 32 electrodes (10–20 system) at 5 kHz, while sEMG signals were acquired from five muscles of the right arm and hand at 4 kHz.

Although the dataset provides trial-based segmentation, all signals are available in their raw continuous form, allowing flexible definition of analysis windows.

B. Task & Acquisition Protocol

Participants were seated comfortably with arm support and performed a reach-to-grasp task. In each trial, the subject reached, grasped (thumb and index finger), lifted, briefly held, and returned an object to its initial position. Task timing was standardized using visual cues (LEDs).

The protocol consisted of multiple experimental series with controlled variations in object weight (165 g, 330 g, 660 g) and surface texture (sandpaper, suede, or silk), organized into Weight, Surface, and Mixed conditions. These variations were structured to be unpredictable to the participant while remaining controlled within the experimental design.

C. Signal Processing & Analysis

1. Acquisition & pre-processing

EEG signals (6 channels: FC6, T7, C3, Cz, C4, T8) and sEMG (5 muscles: Anterior Deltoid - AD, Brachioradialis - BR, Common Extensor Digitorum - EDC, Flexor Digitorum Superficialis - FD, and First Dorsal Interosseus - FDI) were acquired simultaneously at 500 Hz and 4000 Hz, respectively. Grip force and load force were recorded by the device's transducer at 500 Hz. The EEG was filtered with high-pass at 5 Hz and low-pass at 80 Hz (4th and 6th order Butterworth, zero phase via filtfilt). Artifacts were identified by amplitude threshold (100 μV) and gradient (50 μV/sample). Contaminated samples were linearly interpolated when the fraction of bad samples did not exceed 20% of the segment, otherwise, the channel was excluded from that trial. Electrodes were selected for coverage of the primary motor (C3, C4) and premotor (FC6) areas, with T7 and T8 as temporal references.

sEMG was bandpass filtered between 5–400 Hz (4th order Butterworth, phase zero). Trials with sEMG amplitude above 100 units were rejected. Muscles with a task/baseline RMS ratio lower than 1.5 were excluded from that pair, condition, and subject.

The segments of interest were anchored at the LiftOff instant (object release), defined as t = 0 s. The Hold phase comprised the interval [LiftOff - 0.1 s; LiftOff + minimum Hold duration], where the minimum duration was determined per subject as the shortest observed LiftOff→LEDOff interval, ensuring identical length for all trials. Super-padding of 0.8 s (before) and 0.5 s (after) was added to protect the region of interest from the wavelet's influence cone, and was discarded after coherence calculation.

2. Corticomuscular Coherence (CMC)

The EEG signal was resampled from 500 Hz to 4000 Hz to match the native sEMG rate. Time-frequency coherence between each EEG–EMG pair was estimated by continuous wavelet coherence (wcoherence, MATLAB), limited to 5–50

Hz. For each trial, the two-dimensional coherence map (frequency × time) was clipped to the clean window (without padding) and interpolated onto a fixed temporal grid at 4000 Hz. The coherence mean was calculated per frequency band: Mu (8–12 Hz), Beta (15–30 Hz), and Gamma (30–50 Hz) [18], [19]. For each band, the time mean and peak values were extracted as dependent variables. According to Friston [20], functional connectivity is defined by the "statistical dependence between remote neurophysiological events". Given this, we decided to explore corticomuscular coherence (CMC) data in search of a functional marker of neural activity and its synchronicity with muscle activation, mathematically defined as:

$$\mathrm{CMC}_{\mathrm{EEG,EMG}}(f) = \frac{|S_{\mathrm{EEG,EMG}}(f)|^2}{S_{\mathrm{EEG,EEG}}(f) \cdot S_{\mathrm{EMG,EMG}}(f)}, \quad (1)$$

where $S_{\mathrm{EEG,EMG}}(f)$ represents the cross-spectrum between EEG and EMG signals at frequency f, and $S_{\mathrm{EEG,EEG}}(f)$ and $S_{\mathrm{EMG,EMG}}(f)$ correspond to their respective auto-spectra.

3. Intermuscular Coherence (IMC)

The sEMG signals from the active muscle pairs were band pass filtered between 10–400 Hz (pipeline A: phase coherence) and 5–400 Hz (pipeline B: envelope). Both were decimated to 500 Hz after anti-aliasing filtering (6th order Butterworth, cutoff at 200 Hz). Continuous wavelet coherence was calcu lated between sEMG pairs decimated to 500 Hz (wcoherence), using the same clipping and interpolation protocol described for CMC. The bands analyzed were: Mu/Alpha (8–12 Hz), Low Beta (12–20 Hz), Beta (15–30 Hz), High Beta (20–30 Hz), and Gamma (30–50 Hz).

For pipeline B (common drive), the sEMG signal was full wave rectified (module after 5–400 Hz filtering), decimated to 500 Hz, mean-centered, and low-pass filtered at 10 Hz (4th order Butterworth, phase zero). Pearson's correlation between the envelopes of the two muscles was calculated in 200 ms sliding windows with a 20 ms step, generating a temporal correlation profile throughout the Hold. The extracted variables were the mean and peak of the correlation coefficient.

By definition, the IMC quantifies the spectral coherence between two sEMG signals, reflecting the degree of synchronization between motoneuron pools and, consequently, the shared common drive between the evaluated muscles. IMC was defined as:

$$\mathrm{IMC}_{i,j}(f) = \frac{\left|S_{\mathrm{EMG}_i,\mathrm{EMG}_j}(f)\right|^2}{S_{\mathrm{EMG}_i,\mathrm{EMG}_i}(f) \cdot S_{\mathrm{EMG}_j,\mathrm{EMG}_j}(f)}, \quad (2)$$

where $S_{\mathrm{EMG}_i,\mathrm{EMG}_j}(f)$ is the cross-spectrum between EMG signals from muscles i and j, and $S_{\mathrm{EMG}_i,\mathrm{EMG}_i}(f)$ and $S_{\mathrm{EMG}_j,\mathrm{EMG}_j}(f)$ are the corresponding auto-spectra.

4. Force variables

Grip force was extracted as an absolute value to correct for the negative sign convention of the transducer (compression). The safety margin was calculated as the ratio between grip force and load force (SM = |GF| / |LF|), representing the safety factor beyond the minimum required to prevent slippage (SM = 1 would correspond to the slippage threshold). Time averages during the hold window were used as dependent variables.

D. Statistics

The effects of experimental conditions (6 surface × weight combinations) and electrode-muscle pairs (CMC: 20 pairs; IMC: 10 pairs) were evaluated by two-way repeated measures ANOVAs, with Geisser-Greenhouse correction when sphericity was violated ($\epsilon < 0.75$). Post-hoc tests were performed using Tukey's multiple comparisons ($\alpha = 0.05$). The identification of the predominant modulating factor (weight or surface) was based on inspection of the clustering pattern of significant conditions: conditions clustered by surface type, regardless of weight, indicate a friction effect; conditions clustered by weight, regardless of surface, indicate a load effect.

To directly quantify the independent contributions of weight and surface, separate one-way RM ANOVAs were conducted for each factor at fixed levels of the other. A 2×2 interaction analysis (Sandpaper/Suede × 165g/330g) was performed using subject-level contrast scores tested via one-sample t-test. Partial eta squared (η^2p) was reported as effect size. Force variables (grip force, load force, and safety margin) were analyzed by separate one-way ANOVAs for weight (3 levels, fixed surface on sandpaper) and surface (3 levels, fixed weight at 330g), with Tukey's post-hoc test. Finally, to assess the within-subject association between CMC peak and safety margin, Pearson correlations were computed per subject across conditions and Fisher-z transformed before group-level testing via one-sample t-test. All the statistical analyses were conducted using MATLAB R2024a and GraphPad Prism.

III. Results

A. Strength Behavioral Modulation

Grip force increased with weight on the control surface (sandpaper), from 2.61 ± 0.38 N (165 g) to 7.00 ± 0.66 N (660 g; $p < 0.001$), and load force linearly increased with mass ($p < 0.001$). The safety margin decreased progressively with increasing load (1.63 ± 0.10 to 1.03 ± 0.06; $p < 0.001$), indicating a more economical strategy under heavier weights. With a fixed weight of 330g, the reduction in friction increased the grip force (Sandpaper: 4.28 N; Suede: 9.31 N; $p < 0.001$) and the safety margin (Sandpaper: ∼1.3; Silk: ∼2.0; Suede: ∼2.7; $p < 0.001$), without altering the load force between surfaces, confirming selective adjustment of the tangential component (Fig. 1).

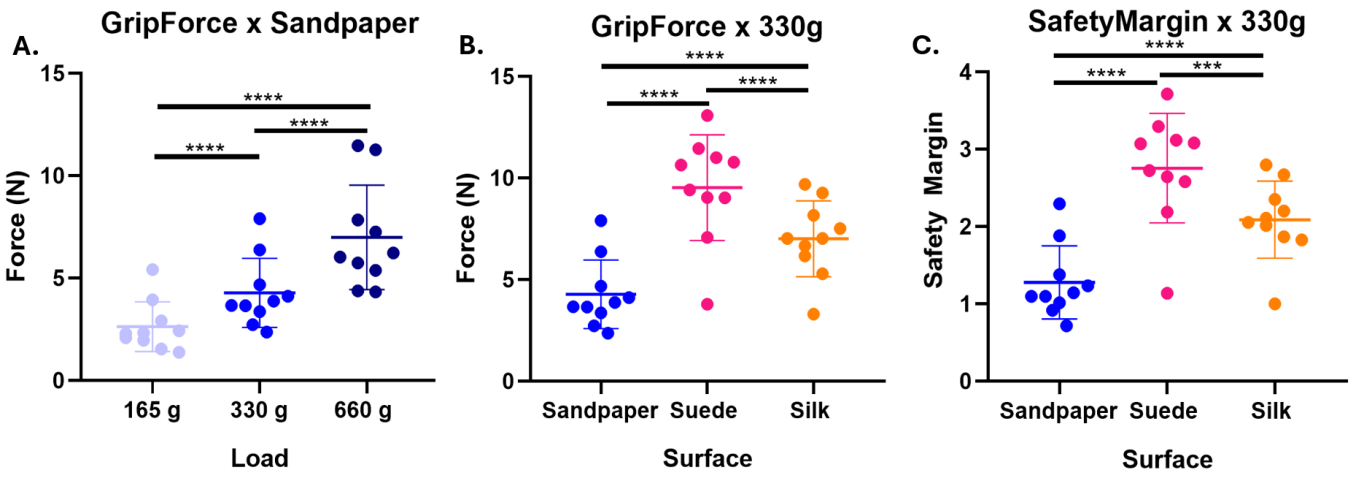


Fig. 1. Grip force and safety margin across experimental conditions. (a) Grip force as a function of object weight on Sandpaper surface (165g, 330g, 660g). (b) Grip force as a function of surface type at constant weight (330g). (c) Safety margin as a function of surface type at constant weight (330g). Load force (not shown) scaled linearly with object mass and did not vary across surfaces. Each dot represents a single subject (n = 10). Lines indicate mean ± SEM. Significance brackets: Tukey post-hoc, **** $p < 0.0001$.

B. Corticomuscular Coupling- CMC

Average CMC was not modulated by experimental conditions in any band (Beta: $p = 0.264$; Gamma: $p = 0.696$; ns interaction). In contrast, peak CMC was the most sensitive variable in the study. In the Beta band, the Column Factor (condition) explained 51.31% of the total variance ($F(2,39; 21,51) = 190.2$; $p < 0.0001$), with a uniform effect across all 20 channel-muscle pairs (interaction: $p = 0.63$). The pattern of post-hoc tests revealed the surface as the predominant modulating factor. In the Gamma band, Suede-165g vs. Sandpaper-330g was significant across all 20 pairs analyzed, being the most consistent effect in the dataset. Sandpaper-165g vs. Suede-165g and Sandpaper-330g vs. Silk-330g was significant in 15 out of 20 pairs; Suede-165g vs. Sandpaper-660g in 13 out of 20. In the Beta band, the same pattern was reproduced with smaller magnitudes (14 and 12 out of 20 pairs, respectively). The ordering of conditions (Suede-165g > Silk-330g ≈ Suede-330g > Sandpaper-165g ≈ Sandpaper-660g > Sandpaper-330g) shows grouping by surface type, not by weight. To further quantify the findings regarding the dominance factor, separate one-way RM ANOVAs were performed and they confirmed surface as the dominant modulator of CMC peak (Beta: $\eta^2p = 0.97$, $p < 0.0001$; Gamma: $\eta^2p = 0.97$, $p < 0.0001$), with weight showing a smaller but significant effect (Beta: $\eta^2p = 0.86$; Gamma: $\eta^2p = 0.75$). The Weight × Surface interaction was significant in both bands (Beta: $t(9) = 8.53$, $p < 0.0001$, $\eta^2p = 0.89$; Gamma: $t(9) = 19.84$, $p < 0.0001$, $\eta^2p = 0.98$), indicating that the effect of weight on CMC depends on the friction condition. The First Dorsal Interosseous showed a systematically lower peak CMC than the other muscles in all EEG channels (grand mean: 0.321 vs. ∼0.340). There was no relevant topographic difference between channels (range: 0.009; interaction ns). Additionally, a within-subject correlation analysis between CMC Beta peak and safety margin revealed a strong positive association across the six conditions (mean $r = 0.80$; Fisher-z: $t(9) = 18.53$, $p < 0.0001$, 95% CI [0.96, 1.23]) (Fig. 2).

C. Intermuscular Coordination

The muscle pair was the main source of variance (Beta: 55.733%; Gamma: 59.41%; $p < 0.0001$), with the Brachioradialis × Flexor Digitorum (hereafter BR × FD) and Brachioradialis × Common Extensor Digitorum (hereafter BR × EDC) pairs being significantly superior to all others, in all conditions and both bands. The condition effect, although significant ($p = 0.0001$), explained a small proportion of the variance (∼2–3%), and the Row × Column interaction was significant, indicating a muscle pair-dependent effect. In the Gamma band, the differences were concentrated in the Brachioradialis × Flexor Digitorum pair, where Sandpaper-660g outperformed all other conditions ($p = 0.002$ to 0.034), and the anterior deltoid × brachioradialis pair showed a similar pattern (Sandpaper-660g > Sandpaper-330g, $p = 0.023$; > Suede-330g, $p = 0.016$). For BR×FD, weight was the primary modulator ($\eta^2p = 0.61$, $F(1.13, 10.14) = 14.12$, $p = 0.003$), while surface had a weaker effect ($\eta^2p = 0.41$, $p = 0.024$) with no significant post-hoc comparisons. No Weight × Surface interaction was observed ($t(9) = 0.64$, $p = 0.536$, $\eta^2p = 0.04$). Altogether, the IMC results corroborate the interpretation that, contrary to CMC, conditions clustered by weight, with the heavier condition standing out regardless of the surface. Friction had a limited impact on IMC, modulating only one isolated pair. This functional dissociation between CMC and IMC constitutes the central finding of the study (Fig. 3).

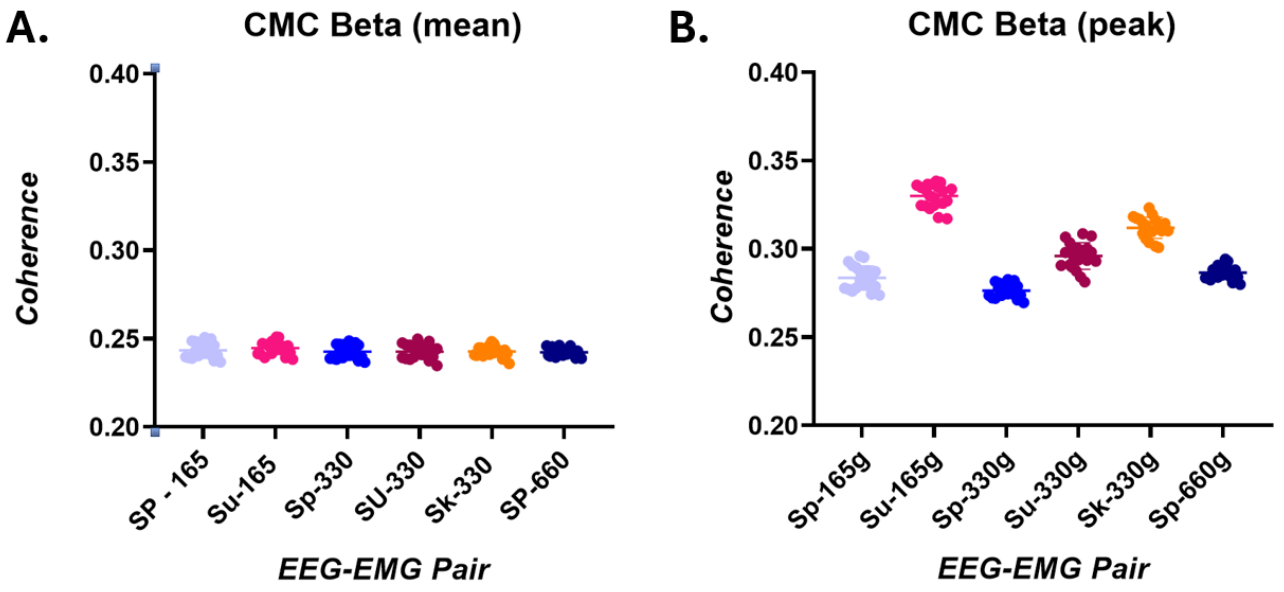


Fig. 2. Corticomuscular coherence (CMC) in the Beta band — mean vs. peak. (a) CMC mean across the 20 EEG–EMG channel pairs and six experimental conditions. (b) CMC peak across the same pairs and conditions. Each dot represents a single subject's value for a given channel–muscle pair. Conditions are color-coded by surface: Sandpaper (blue shades), Suede (pink shades), Silk (orange). No significant effect of condition was observed for CMC mean (Column Factor: 0.64%, $p = 0.264$). CMC peak was strongly modulated by condition (Column Factor: 51.31%, $F(5, 54) = 48.26$, $p < 0.0001$), with no significant interaction with channel–muscle pair ($p = 0.98$). Post-hoc analysis revealed that Suede-165g vs. Sandpaper-330g was significant in 14 of 20 pairs; Sandpaper-165g vs. Suede-165g in 12 of 20. Conditions grouped by surface type, not by weight, identifying friction as the primary modulator. Abbreviations: Sp = Sandpaper, Su = Suede, Sk = Silk.

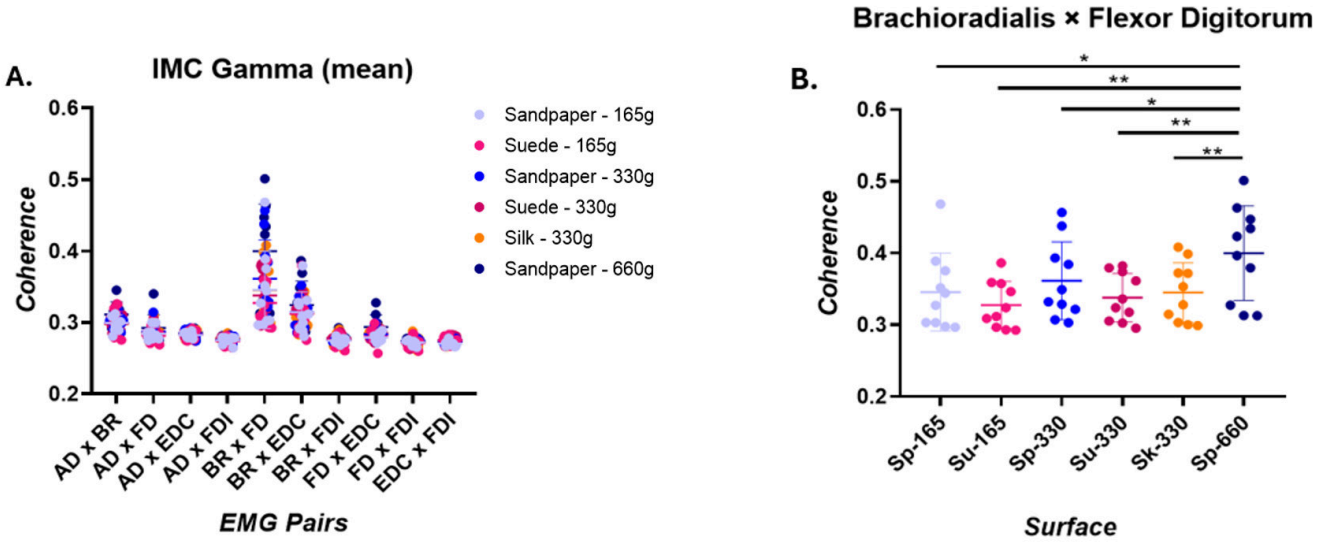


Fig. 3. Intermuscular coherence (IMC) in the Gamma band. (a) IMC across the 10 muscle pairs and six experimental conditions. The pair BR×FD and BR×EDC exhibited significantly higher IMC than all other pairs (Row Factor: 59.41%, $p < 0.0001$). (b) IMC for the BR×FD pair across the six conditions. Sandpaper-660g was significantly higher than all other conditions (Tukey post-hoc: * $p < 0.05$, ** $p < 0.01$), with conditions grouping by weight rather than surface. The Row × Column interaction was significant ($p = 0.0002$), indicating that weight-driven modulation was concentrated in forearm synergist muscle pairs. Each dot represents a single subject ($n = 10$). Lines indicate mean ± SEM. Muscle abbreviations: AD = Anterior Deltoid, BR = Brachioradialis, FD = Flexor Digitorum, EDC = Com. Ext. Digitorum, FDI = First Dorsal Interosseous. Surface abbreviations: Sp = Sandpaper, Su = Suede, Sk = Silk."

IV. Discussion

The impacts of neurological and motor injuries on the quality of life of those who suffer from them, as well as on the burdens generated in Public Health and the available workforce, are undeniable. Thus, the advancement of tools that allow for early diagnosis and optimization of physiotherapy

follow-up after the condition, through quantitative, sensitive, and objective biomarkers, is essential. In line with this need, the present study aimed to investigate the modulation of corticomuscular coupling (CMC) and intermuscular coherence (IMC) as a function of the physical properties of the object, more specifically weight and friction, during a grasping task followed by support, seeking to expand the understanding of its functional role in corticospinal communication and muscle coordination.

Under physiological conditions, corticomuscular coherence is typically observed in the Beta band (12–30 Hz) during static force production and fine control, while the Gamma band (>30 Hz) is associated with dynamic movements and rapid adjustments dependent on sensory feedback [21], [22], [23]. Here, however, more complex modulation patterns were revealed, characterized by spectral dissociation dependent on the type of motor demand imposed.

In the context of the analysis of the force involved in gripping the object, it was possible to observe that it scaled according to the increase in the weight of the object and decreased relative to the decrease in friction with the surface, in addition to a progressive decrease in the safety margin relative to the increase in the load handled. Conversely, surfaces with less friction, such as suede, required greater grip strength and resulted in a larger safety margin, reflecting direct biomechanical compensation for slipperiness, while the support force remained unchanged between surfaces, confirming that perceived weight was not affected by the texture. These findings, when combined, suggest a more economical control strategy, in which the nervous system reduces excess force as the mechanical predictability of the task increases. Additionally, the substantial increases in grip strength and safety margin, without relevant changes in support force, suggest selective adjustment of the tangential component of the force. This pattern is consistent with grip control models based on friction estimation, in which normal and tangential forces are regulated in a partially independent manner [24], [25].

In parallel, friction manipulation produced the most robust neural modulation observed in the study, with significant increases in CMC in the Beta and Gamma bands, observed at their activation peaks, in a large number of electrode-muscle pairs. This result suggests that adaptation to surface properties, particularly under conditions of lower friction, requires greater cortical involvement and greater downward coupling, an adaptive pattern reinforced by the concomitant observation of an increase in the safety margin of the applied grip force. For example, it is observed here that the condition with the highest adaptive demand of the Grip (suede - 165g) is the same one that gave the highest peak of corticomuscular coherence, reinforcing the possible relationship between high corticomuscular coupling and tasks that demand fine adjustments. This association was confirmed quantitatively: a within-subject correlation analysis revealed a strong positive relationship between CMC Beta peak and safety margin across conditions (mean $r = 0.80$, $t(9) = 18.53$, $p < 0.0001$), indicating that conditions eliciting greater grip force adjustment also produced stronger corticomuscular coupling, supporting a direct functional link between cortical engagement and precision grip control. A full trial-level analysis using mixed-effects models remains planned to examine this relationship at the single-trial level.

Still within the scope of corticomuscular correlation analyses, the Beta band is generally associated with the maintenance of isometric force [21], while the Gamma band is related to the dynamic adjustment of motor output [22]. Thus, although the average CMC for both bands does not appear to be modulated by the changes in the presented conditions, the peaks did. The concomitant increase observed as activation peaks for both bands may reflect the integration of tactile information and the need for fine control of grip strength to avoid object slippage, as well as reflecting the transient nature of information transmission between center and periphery. This pattern is consistent with studies that demonstrate CMC modulation in reaching and grasping tasks with different load properties [26]. Additionally, CMC analysis revealed a consistent muscle hierarchy: the First Dorsal Interosseous showed a consistently lower peak corticomuscular coherence than the other muscles evaluated, regardless of the EEG channel and experimental condition. This finding may reflect the more focal cortical representation of this intrinsic hand muscle, in contrast to larger muscles and more proximal muscles, such as the brachioradialis and anterior deltoid, whose more extensive cortical representations would favor greater coherence capture by surface electrodes.

The analysis of intermuscular coherence, in turn, revealed a pattern complementary to that observed in CMC. The increase in weight led to an intensification of IMC between distal muscles (brachioradialis and flexor digitorum and common extensor digitorum) in both bands studied, but more pronouncedly for Gamma. This type of coherence is traditionally interpreted as a reflection of synaptic inputs common to motoneurons [21]. According to [12], IMC is directly related to functional coordination and the formation of muscle synergies. Thus, the increase in coherence with load suggests that the motor system reinforces intersegmental coordination strategies to ensure mechanical stability under greater force demands.

Friction, on the other hand, showed a limited impact on intermuscular coherence. This dissociation between CMC and IMC may suggest that different levels of the motor system are selectively recruited depending on the nature of the demand and is demonstrated for the first time in the present work. The findings indicate that, while friction predominantly modulates corticomuscular coupling, weight mainly affects intermuscular coordination. This functional separation suggests the existence of two complementary levels of motor control: (i) a cortical level, susceptible to sensory feedback and responsible for task-dependent fine adjustments, reflected in CMC; and (ii) a more distributed level, involving common descending and spinal inputs, responsible for the organization of muscle synergies and mechanical stability, reflected in IMC. This interpretation is aligned with studies that demonstrate that CMC and IMC

can reflect distinct, although interrelated, aspects of motor control, especially in clinical contexts [15].

From an applied perspective, in healthy individuals these results support the complementary use of CMC and IMC as functionally distinct neuromotor measures. The strong modulation of CMC in medium and high frequency bands in response to friction suggests sensitivity to fine sensory demands (top-down mechanism), which can be explored in motor rehabilitation contexts and brain-machine interfaces [27]. On the other hand, IMC emerges as a robust indicator of the organization of muscle synergies (lateral coordination) under increased load, with implications for the study of intermuscular coordination in clinical populations, as well as a promising tool for optimizing prosthetic control [28], [29].

Despite these contributions, some limitations should be considered. The relatively non-stationary nature of the task may restrict the generalization of the findings and demands studies with different mathematical tools in a complementary way. Furthermore, the analysis was limited to a specific set of muscles and cortical regions, which may underestimate the extent of the neural networks involved and, in addition, the present findings are limited to healthy individuals and do not assess diagnostic sensitivity in clinical populations. Hence, future investigations should explore tasks with greater functional complexity and incorporate network connectivity approaches for a more comprehensive characterization of motor control and apply them to volunteers with motor impairments. Finally, trial-level studies will be carried out in the next stage of this work, to verify whether the observed association between conditions reflects a direct functional coupling.

In summary, this study demonstrates, for the first time in the same dataset, that corticomuscular coupling and intermuscular coherence are modulated in a dissociated and task-dependent manner, reflecting different levels of organization of the motor system. While increased load promotes greater intermuscular coordination and less cortical dependence at low frequencies, reduced friction intensifies corticomuscular coupling in high-frequency bands, highlighting the critical role of the cortex in adapting to sensory demands. These findings contribute to a more integrated understanding of the neural mechanisms of grip control and reinforce the combined use of CMC and IMC as tools for neurophysiological research and clinical applications.

V. Conclusion

This study demonstrates, in a novel and integrative manner within a single experimental paradigm, that corticomuscular and intermuscular coherence are dissociably modulated during object manipulation, each reflecting a distinct level of motor organization. The robust modulation of CMC peak by surface friction, concomitant with increases in grip force and safety margin, points to a cortical control mechanism that is selectively recruited when fine sensory-dependent force adjustments are required: a top-down process sensitive to the predictability of the mechanical interaction. In a complementary fashion, the modulation of IMC by object weight, concentrated among forearm synergist muscles, reflects the reinforcement of coordinated muscular strategies to ensure mechanical stability under greater force demands, a more distributed process involving common descending and spinal inputs. Taken together, these findings expand the current understanding of the neural mechanisms underlying precision grip, revealing that CMC and IMC capture complementary and non-redundant aspects of motor control. From an applied perspective, these results reinforce the potential of combining both measures as quantitative biomarkers for the assessment of neuromotor function in healthy individuals, with promising avenues for their incorporation into rehabilitation protocols and brain–machine interface design, after clinical cohort validation.

Acknowledgment

The authors acknowledge the financial support provided by CNPq, CAPES, FAPEMIG, and FINEP. The authors also thank Luciw et al. [17], of the WAY-EEG-GAL dataset, for making the data publicly available.

References

[1] World Health Organization, "Neurological conditions affect over 1 in 3 people worldwide — the leading cause of illness and disability," WHO, Mar. 2024. [Online]. Available:https://www.who.int/news/item/14-03-2024-over-1-in-3-people-affected-by-neurological-conditions

[2] P. L. Hudak, P. C. Amadio, and C. Bombardier, "Development of an upper extremity outcome measure: The DASH," *Am. J. Ind. Med.*, vol. 29, no. 6, pp. 602–608, 1996.

[3] B. Rosén and G. Lundborg, "A model instrument for the documentation of outcome after nerve repair," *J. Hand Surg. Am.*, vol. 25, no. 3, pp. 535–543, 2000.

[4] T. Gordon and A. W. English, "Strategies to promote peripheral nerve regeneration: electrical stimulation and/or exercise," *Eur. J. Neurosci.*, vol. 43, no. 3, pp. 336–350, 2016.

[5] F. J. Alvarez, T. M. Rotterman, E. T. Akhter, A. R. Lane, A. W. English, and T. C. Cope, "Synaptic plasticity on motoneurons after axotomy: a necessary change in paradigm," *Front. Mol. Neurosci.*, vol. 13, p. 68, 2020.

[6] M. Fauvet, D. Gasq, A. Chalard, J. Tisseyre, and D. Amarantini, "Temporal dynamics of corticomuscular coherence reflects alteration of the central mechanisms of neural motor control in post-stroke patients," *Front. Hum. Neurosci.*, vol. 15, p. 682080, 2021.

[7] F. Artoni, A. Cometa, S. Dalise, V. Azzollini, S. Micera, and C. Chisari, "Cortico-muscular connectivity is modulated by passive and active Lokomat-assisted gait," *Sci. Rep.*, vol. 13, p. 21618, 2023.

[8] D. Elie, F. Barbier, G. Ido, and S. Cremoux, "Corticomuscular coherence and motor control adaptations after isometric maximal strength training," *Brain Sci.*, vol. 11, no. 2, p. 254, 2021.

[9] S. Saleh et al., "Corticomuscular connectivity during walking in able bodied and individuals with incomplete spinal cord injury," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2022, pp. 2332–2335, 2022.

[10] K. Kato and Y. Nishimura, "Artificial cortico-muscular connection via neural interface to regain volitional control of limb movements," *J. Rehabil. Neurosci.*, 2020, doi: 10.24799/jrehabilneurosci.200731.

[11] C. Ethier, E. R. Oby, M. J. Bauman, and L. E. Miller, "Restoration of grasp following paralysis through brain-controlled stimulation of muscles," *Nature*, vol. 485, no. 7398, pp. 368–371, 2012.

[12] C. M. Laine and F. J. Valero-Cuevas, "Intermuscular coherence reflects functional coordination," *J. Neurophysiol.*, vol. 118, no. 3, pp. 1775–1783, 2017.

[13] T. W. Boonstra and M. Breakspear, "Neural mechanisms of intermuscular coherence: Implications for the rectification of surface electromyography," *J. Neurophysiol.*, vol. 107, no. 3, pp. 796–807, 2012.

[14] I. E. de Vries, A. Daffertshofer, D. F. Stegeman, and T. W. Boonstra, "Functional connectivity in the neuromuscular system underlying bimanual coordination," *J. Neurophysiol.*, vol. 116, no. 6, pp. 2576–2585, Dec. 2016, doi: 10.1152/jn.00460.2016.

[15] C. Delcamp, D. Gasq, C. Cormier, and D. Amarantini, "Corticomuscular and intermuscular coherence are correlated after stroke: A simplified motor control?," *Brain Commun.*, vol. 5, no. 3, p. fcad187, 2023.

[16] S. Huang et al., "Enhanced functional connectivity of cortical-subcortical sensorimotor areas after brain-controlled supernumerary robotic finger training: A longitudinal fMRI study," in *Proc. 2024 IEEE Int. Conf. Robotics and Biomimetics (ROBIO)*, 2024, pp. 2312–2317.

[17] M. D. Luciw, E. Jarocka, and B. B. Edin, "Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction," *Sci. Data*, vol. 1, p. 140047, 2014.

[18] S. Bista et al., "Cortico-muscular coherence in primary lateral sclerosis reveals abnormal cortical engagement during motor function beyond primary motor areas," *Cereb. Cortex*, vol. 33, no. 13, pp. 8712–8723, 2023.

[19] R. Krauth et al., "Cortico-muscular coherence is reduced acutely post-stroke and increases bilaterally during motor recovery: A pilot study," *Front. Neurol.*, vol. 10, p. 126, 2019.

[20] K. J. Friston, "Functional and effective connectivity: A review," *Brain Connect.*, vol. 1, no. 1, pp. 13–36, 2011.

[21] S. F. Farmer, J. P. R. Dick, and J. A. Stephens, "Rhythmicity, synchronization and binding in human and primate motor systems," *J. Physiol.*, vol. 470, pp. 127–155, 1993.

[22] W. Omlor, L. Patino, M.-C. Hepp-Reymond, and R. Kristeva, "Gamma-range corticomuscular coherence during dynamic force output," *NeuroImage*, vol. 34, no. 3, pp. 1191–1198, 2007.

[23] T. W. Boonstra, "The potential of corticomuscular and intermuscular coherence for research on human motor control," *Front. Hum. Neurosci.*, vol. 7, p. 855, 2013.

[24] H. O'Shea and S. J. Redmond, "A review of the neurobiomechanical processes underlying secure gripping in object manipulation," *Neurosci. Biobehav. Rev.*, vol. 123, pp. 286–300, Apr. 2021, doi: 10.1016/j.neubiorev.2021.01.007.

[25] L. Willemet, F. Roël, D. Abbink, I. Birznieks, and M. Wiertlewski, "Grip force control under sudden change of friction," *J. Physiol.*, vol. 603, pp. 411–422, 2025, doi: 10.1113/JP286486.

[26] C. D. Guerrero-Mendez and A. F. Ruiz-Olaya, "Coherence-based connectivity analysis of EEG and EMG signals during reach-to-grasp movement involving two weights," *Brain-Comput. Interfaces*, vol. 9, no. 3, pp. 140–154, 2022.

[27] V. De Seta, J. Toppi, F. Pichiorri, M. Masciullo, E. Colamarino, and D. Mattia, "Exploring high-density corticomuscular networks after stroke to enable a hybrid Brain-Computer Interface for hand motor rehabilitation," *J. NeuroEng. Rehabil.*, vol. 20, no. 1, pp. 1–14, 2023.

[28] A. Furui et al., "A myoelectric prosthetic hand with muscle synergy-based motion determination and impedance model-based biomimetic control," *Sci. Robotics*, vol. 4, no. 31, p. eaaw6339, 2019.

[29] X. He, W. Sun, R. Song, and W. Xu, "Wavelet coherence analysis of post-stroke intermuscular coupling modulated by myoelectric-controlled interfaces," *Bioengineering*, vol. 11, no. 8, p. 802, 2024.

ECG-Based Chagas Disease Classification under Weak Self-Reported Labels: A Comparison of Target Definitions in CODE-15%

João Samuel Maciel de Sales
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
joaosamuel.2001@gmail.com

Rogério Ferreira Rodrigues Júnior
Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará
Fortaleza, Brasil
rger0904@hotmail.com

Luiza Esther Martins Pessoa
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
esthermartins@alu.ufc.br

Naila Ketley de Freitas Bruno
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
naila.ketley@alu.ufc.br

Gisela V. Clemente
Departamento de Matemática
Universidad Nacional de La Plata
La Plata, Argentina
gclemente@mate.unlp.edu.ar

João Paulo do Vale Madeiro
Departamento de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
jpaulo.vale@dc.ufc.br

***Abstract*—Electrocardiogram-based screening for Chagas disease is challenging when large-scale datasets provide weak self-reported labels rather than serologically confirmed diagnoses. This study investigates how repeated self-reported Chagas labels in CODE-15% can be operationalized as supervised targets for deep learning. Within-patient label consistency was first assessed using repeated ECG examinations from the same patients, without assuming temporal ordering. Three labeling strategies were then compared: retaining the original examination-level labels; assigning a positive patient-level target when at least one examination contained a positive self-report; and assigning a positive target when at least half of the available examinations were positive. All strategies were evaluated using the same 12-lead one-dimensional convolutional neural network and five-fold patient-level cross-validation. Among examinations from patients who reported Chagas disease at least once, 35.87% were labeled as negative, revealing substantial within-patient discordance. The strategy based on at least one positive self-report achieved the highest F1-score, PR-AUC, and MCC under both fixed and validation-selected decision thresholds, whereas the original labels achieved the highest ROC-AUC. Performance also varied across ECG subgroups, with stronger results generally observed among ECGs with known abnormalities and lower performance among normal ECGs. These findings show that target construction is an important modeling decision when repeated weak labels are available. However, because all targets were derived from repeated self-reported labels, the reported results characterize agreement with alternative operational label definitions rather than clinical diagnostic accuracy against serologically confirmed Chagas disease.**

***Index Terms*—ECG, CODE-15%, Deep Learning, Chagas disease.**

I. Introduction

Chagas disease (CD), caused by the protozoan *Trypanosoma cruzi*, remains a major neglected tropical disease affecting over six million people globally. Although its prevalence is highest in Latin America, migratory patterns have contributed to its spread to non-endemic regions [1]. Early antiparasitic treatment can prevent disease progression when administered in earlier phases [2], [3], but detection remains limited: in many countries, fewer than 10% of cases are diagnosed; in some settings, this figure falls below 1% [2], [4]. This underdiagnosis stems partly from asymptomatic or nonspecific early stages and from reliance on serological methods that are difficult to scale in resource-constrained settings [2], [3], [5]. These limitations motivate accessible and non-invasive ECG-based screening strategies [2], [5], while heterogeneous label sources in ECG datasets raise the question of how disease status and target labels should be defined in data-driven studies [2], [3], [6].

The electrocardiogram (ECG) is a promising tool in this context, as it is widely available, low-cost, non-invasive, and capable of capturing electrical abnormalities associated with the cardiac manifestations of CD [5], [7], [8]. Automated ECG analysis may therefore support large-scale screening, while serial ECG assessment can contribute to the longitudinal monitoring of disease progression, especially in populations with limited access to confirmatory testing [2], [8], [9]. However, CD is clinically heterogeneous: infected individuals may remain in an indeterminate form for long periods, develop chronic Chagas cardiomyopathy (CCC), or present concomitant cardiovascular conditions that affect ECG morphology [8], [9]. This heterogeneity complicates the construction of supervised learning targets since an ECG may capture cardiac involvement more directly than infection status itself [2], [6].

Large-scale ECG datasets have enabled deep learning (DL) models to achieve strong performance in automated ECG interpretation and prognostic modeling. In this context, the Clinical Outcomes in Digital Electrocardiography (CODE) dataset has supported large-scale ECG model development,

while its public subset, CODE-15%, has become an important resource for reproducible ECG-based research [10]–[12]. However, when these datasets are used for CD screening, their usefulness depends on more than signal availability and sample size; it also depends on the clinical reliability and consistency of the target labels [2], [3], [6].

Several Brazilian cohorts have played complementary roles in the development and external evaluation of ECG-based CD models, although they differ substantially in population, clinical profile, label composition, and data accessibility. The ECG tracings from the São Paulo–Minas Gerais Tropical Medicine Research Center (SaMi-Trop) are publicly available. This cohort comprises individuals with serologically confirmed CD and includes substantial representation of patients with CCC [13], [14]. In contrast, the ECG data from the Retrovirus Epidemiology Donor Study-II (REDS-II) and the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) are not publicly available. REDS-II includes former blood donors with positive and negative CD status established by serological testing, whereas ELSA-Brasil is a broader longitudinal cohort of Brazilian civil servants in which positive and negative CD labels were derived from serological testing [6], [15], [16].

The consequences of this heterogeneity are evident in studies that combine datasets with different label strengths and clinical meanings. In a recent study, a model was developed using the SaMi-Trop and CODE datasets, creating a heterogeneous training setting: SaMi-Trop provides serology-based CD labels and is strongly enriched for seropositive patients and CCC, whereas CODE offers large-scale ECG data with self-reported CD labels, for which CCC status was not available [2]. Although the model achieved an AUC-ROC of 0.80 on an internal validation set combining samples from CODE and SaMi-Trop, its performance decreased under external validation on the REDS-II and ELSA-Brasil cohorts, reaching AUC-ROC values of 0.68 and 0.59, respectively [2]. This gap between internal and external validation illustrates the difficulty of transferring ECG-based CD classifiers across cohorts with different clinical profiles, label definitions, and data distributions [2].

The George B. Moody PhysioNet Challenge 2025 further reinforced this concern by evaluating ECG-based CD detection across heterogeneous data sources and hidden validation/test splits. In this setting, CODE-15% represented a large-scale public dataset with weak self-reported Chagas labels, whereas cohorts such as SaMi-Trop, REDS-II, SaMi-Trop 3, and ELSA-Brasil provided smaller but clinically stronger, serologically based references, although with distinct clinical profiles and evaluation roles [6]. Thus, the Challenge reinforces that ECG-based CD classification should not be interpreted only as a model-performance problem, but also in relation to the source, strength, and clinical meaning of the target labels [6].

Although the scale and public availability of CODE-15% make it valuable for reproducible ECG-based screening studies, its CD labels are derived from patient self-reports rather than from serological confirmation or adjudicated clinical staging [2], [3], [6], [12]. Therefore, CODE-15% should not be treated as a source of clinically confirmed Chagas diagnoses but as a large-scale, weakly labeled dataset whose labels require careful operational definition before supervised training [2], [3], [6]. In supervised DL, this distinction is consequential because such self-reported labels may introduce noise and uncertainty when interpreted as definitive disease status, potentially affecting what the model learns and its ability to generalize beyond the training distribution [17].

A specific source of ambiguity arises when these labels are assigned at the ECG-record level in patients with multiple examinations. The public CODE-15% release contains substantially more ECG records than patients and provides self-reported Chagas labels [6], [12]; therefore, repeated records from the same individual make it possible to assess whether this label source is internally consistent. In this setting, when record-level labels from the same patient are discordant, this inconsistency should be interpreted as an operational source of label uncertainty rather than as confirmed variation in infection status. Such inconsistencies are particularly problematic for supervised learning because they introduce ambiguity into the definition of the positive class and may expose the model to conflicting targets, potentially affecting what the model learns and how it generalizes [17]. This motivates the need to explicitly examine how repeated self-reports should be operationalized as supervised targets for ECG-based CD classification.

To address this challenge, this work frames ECG-based CD classification in CODE-15% as a sensitivity analysis of target definition under self-reported labels. Rather than treating target definition as a fixed component of the classification pipeline or attempting to establish a clinically definitive diagnostic target, our primary scientific contribution is a systematic evaluation of how alternative operational strategies for constructing supervised targets from repeated self-reports affect measured classification performance. In contrast to prior ECG-based CD classification studies, which generally treat the available labels as fixed supervised targets (see Section II), this study makes target definition itself the primary experimental variable.

For this purpose, three labeling strategies are compared. The first retains the original self-reported label of each ECG recording. The second assigns a positive patient-level target to all recordings from an individual who reported CD in at least one available examination. The third assigns a positive patient-level target when at least half of the individual's available records contain a positive self-report. The latter two strategies, therefore, replace potentially discordant record-level targets with a common patient-level target, following inclusive and majority-based interpretations, respectively.

The remainder of this paper is organized as follows. Section II reviews related work on ECG-based CD detection and heterogeneous label sources. Section III describes the analysis of within-patient label consistency, the three operational labeling strategies, ECG preprocessing, model training, and the evaluation protocol. Section IV presents and discusses the experimental results. Finally, Section V summarizes the main conclusions and limitations of the study.

II. Related Work

Reference [2] provides direct evidence that ECG-based CD detection depends on the clinical definition of the positive class. The authors evaluated model performance both for broad CD detection and for an alternative setting in which only patients with CCC were considered positive. For broad CD detection, the model achieved AUC-ROC values of 0.68 on REDS-II and 0.59 on ELSA-Brasil. When only CCC patients were considered positive, performance increased to 0.82 and 0.77 on the same cohorts, respectively [2]. This result suggests that ECG-based models may capture cardiac manifestations of CD more reliably than infection status alone. For the present study, the relevance of this finding is not that CCC should necessarily define the target, but that the clinical meaning of the positive class can change the relationship between ECG patterns and supervised labels.

More broadly, the George B. Moody PhysioNet Challenge 2025 provides a recent benchmark for ECG-based CD detection across heterogeneous data sources. Although the Challenge involved datasets with different label strengths and clinical meanings, published approaches have primarily addressed this setting through model-oriented strategies, such as transfer learning with InceptionTime [5], auxiliary pretraining and fine-tuning across datasets [3], and dilated residual architectures with class-imbalance handling [18]. These examples illustrate the relevance of model design, representation learning, and cross-dataset adaptation in heterogeneous ECG sources.

The presence of self-reported labels within this Challenge also makes the broader literature on learning from uncertain or noisy labels relevant. Methods developed for learning under label noise can reduce the influence of unreliable annotations through robust training strategies [17]. Another approach is to represent label uncertainty explicitly through uncertainty labels or soft label distributions [19], [20]. A complementary strategy is semi-supervised learning, in which highly uncertain cases can be treated as unlabeled, and more consistent cases can be retained as supervised examples [21], [22]. These approaches can provide alternative formulations for learning when the available labels are uncertain, potentially unreliable, or not clinically definitive.

In the specific context of ECG-based CD classification, however, the available disease labels have generally been treated as fixed supervised targets, leaving the target-definition step itself comparatively underexplored. This gap motivates the present focus on the operational definition of supervised targets from repeated self-reported labels.

III. Methodology

The methodology was organized into two stages. First, the within-patient consistency of the exam-level self-reported Chagas labels was assessed using multiple ECG records from the same patients. CODE-15% contains 345,779 exams from 233,770 patients, collected by the Telehealth Network of Minas Gerais between 2010 and 2016 [12]. No missing values were present in the self-reported Chagas labels or in the demographic and ECG covariates considered in the analyses. For patients with multiple exams, the available labels were grouped by patient to characterize patterns of agreement and disagreement across repeated records and to define alternative supervised targets. Because the public CODE-15% metadata does not provide examination dates or timestamps, no temporal ordering of repeated records was assumed. Second, these target definitions were compared under a common experimental pipeline comprising ECG selection and preprocessing, model training, and patient-level cross-validation. Only the labeling strategy was changed across experiments, while the remaining experimental conditions were kept fixed.

A. Within-Patient Consistency of Self-Reported Chagas Labels

Repeated ECG records were analyzed to assess within-patient label consistency across available examinations. Differences between labels recorded for the same patient were interpreted as within-patient discordance in self-report rather than as clinically confirmed changes in infection status.

The analysis was restricted to patients who reported CD in at least one exam. For patients with multiple ECG records, the labels were grouped at the patient level to characterize patterns of agreement and disagreement across available examinations. The characterization included the number and proportion of positive and negative labels, the distribution of the number of available exams per patient, and the most frequent combinations of labels, irrespective of record order.

Fig. 1 presents the distribution of exam-level labels within this subset. Although every included patient had at least one positive self-report, 35.87% of their exams were labeled as negative. Thus, the self-reported label recorded for a given patient was not necessarily consistent across exams.

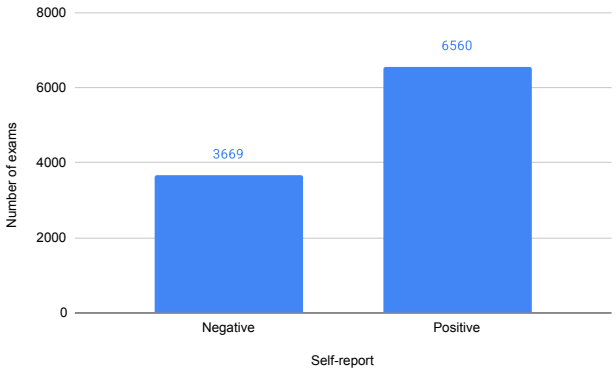


Fig. 1. Distribution of positive and negative exam-level labels among patients who self-reported CD at least once.

The assessment of within-patient label consistency depends on the number of exams available for each patient. Fig. 2 shows the distribution of patients according to their total number of ECG exams. Among 6,560 exams originally labeled as positive, 1,980 belonged to patients represented by a single exam. For these patients, within-patient label consistency could not be assessed because only one exam was available.

For patients with multiple exams, the combinations of available labels were used to characterize patterns of agreement and disagreement irrespective of record order.

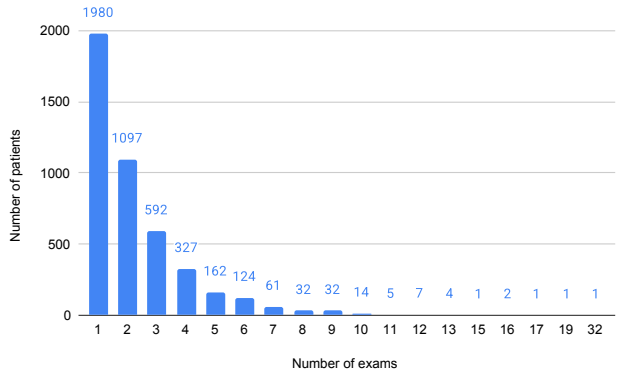


Fig. 2. Distribution of patients according to the total number of available ECG exams

Fig. 3 presents the 15 most frequent self-report configurations among patients with at least one positive self-report. Each configuration is denoted by $E_{\{i,j\}}$, where i is the number of ECG examinations associated with a positive self-report and j is the total number of ECG examinations available for the patient. The configurations satisfying $i = j$ represent consistently positive self-reports, whereas configurations satisfying $0 < i < j$ indicate discordant self-reports across examinations from the same patient.

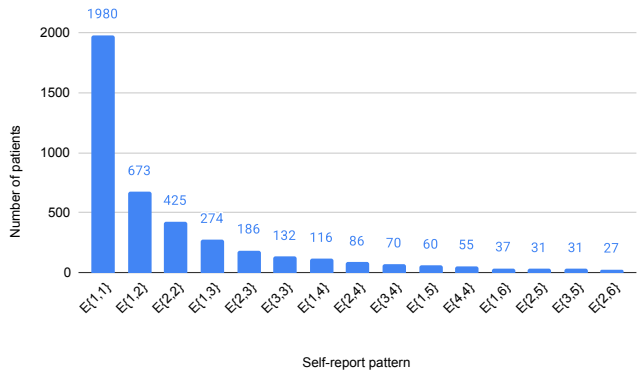


Fig. 3. Fifteen most frequent self-report label patterns among patients who reported CD at least once.

The observed discordance indicates that the original exam-level labels do not always translate into a unique patient-level target. Based on this characterization, three operational labeling strategies were defined to investigate how different uses of self-reported labels across repeated examinations affect supervised ECG classification.

B. Operational Labeling Strategies

Based on the self-reported labels available for each patient, three operational labeling strategies were defined. Let $s_{p,j} \in \{0, 1\}$ denote the original self-reported CD label associated with the j-th ECG recording of patient p, and let n_p denote the total number of ECG recordings available for that patient p. For strategy (i), each ECG recording retained its original self-reported label:

$$y_{p,j}^{(i)} = s_{p,j}. \tag{1}$$

Thus, different ECG recordings from the same patient could retain discordant supervised targets. For strategy (ii), all ECG recordings from a patient were labeled as positive when that patient reported CD in at least one available record:

$$y_{p,j}^{(ii)} = \begin{cases} 1, & \text{if } \sum_{k=1}^{n_p} s_{p,k} \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

For strategy (iii), all ECG recordings from a patient were labeled as positive if at least 50% of the available records contained a positive self-report:

$$y_{p,j}^{(iii)} = \begin{cases} 1, & \text{if } \frac{1}{n_p} \sum_{k=1}^{n_p} s_{p,k} \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The strategies (ii) and (iii) therefore converted the original record-level labels into a common patient-level target that was assigned to all recordings from the same individual. The strategy (ii) represented the most inclusive interpretation of a positive self-report history, whereas the strategy (iii) required positive labels in at least half of the available recordings.

C. ECG Selection and Preprocessing

The original 12-lead ECG recordings were sampled at 400 Hz. As an initial preprocessing step, zero-padded regions at the beginning and end of each recording were removed.

Each remaining recording was filtered independently along the lead dimension using a fourth-order Butterworth band-pass filter with cutoff frequencies of 0.3 and 45 Hz. Zero-phase filtering was performed using a forward-backward second-order sections implementation. The filtered signals were subsequently resampled from 400 to 100 Hz using polyphase resampling.

After resampling, each recording was truncated to a whole number of seconds. Therefore, for a resampled recording containing T samples, the retained length was defined as

$$T' = 100 \left\lfloor \frac{T}{100} \right\rfloor. \tag{4}$$

For the classification experiments, a 300-sample window, corresponding to 3 seconds at 100 Hz, was randomly selected from each ECG recording. The same temporal segment was extracted across all 12 leads, resulting in an input shape of 12 leads $\times$ 300 samples.

D. Model Architecture

Reference [23] presents a one-dimensional CNN architecture developed for the George B. Moody PhysioNet Challenge 2025. This architecture served as the basis for the adapted model used in the present study. The adapted model is a one-dimensional convolutional neural network (CNN) composed of three sequential convolutional blocks. Each block consists of a Conv1D layer followed by a ReLU activation function. The first, second, and third blocks contain 32, 64, and 128 filters, respectively, with kernel sizes of 15, 9, and 5 and corresponding padding values of 7, 4, and 2.

The convolutional feature extractor is followed by an adaptive average pooling layer, which reduces the temporal dimension of the learned representation. The resulting feature vector is then processed by a dense classifier composed of three fully connected layers. A dropout layer with a probability of 0.3 precedes the first linear layer, which maps the 128-dimensional feature vector to 64 neurons and is followed by a ReLU activation. A second dropout layer with a probability of 0.2 precedes the linear mapping from 64 to 32 neurons, also followed by ReLU. Finally, a dropout layer with a probability of 0.1 is applied before the output layer, which maps the 32-dimensional representation to a single logit. A sigmoid function is then used to obtain the predicted probability for the positive class. The resulting architecture is illustrated in Fig. 4, where B denotes the batch size and the input tensor has shape $B \times 12 \times 300$, corresponding to 12-lead ECG windows with 300 samples per lead.

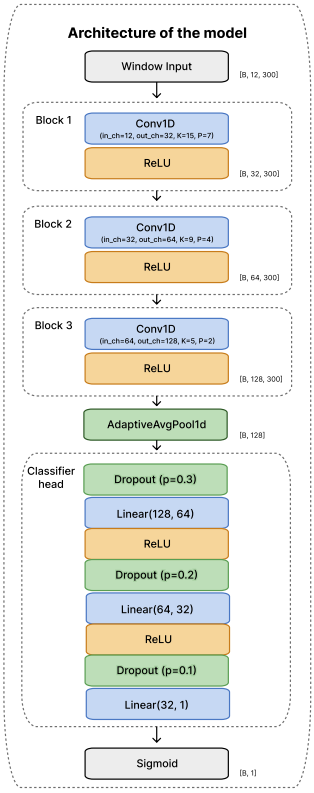


Fig. 4. Architecture of the adapted model.

The overall three-block convolutional structure and dense classification head of the original architecture were retained. However, the convolutional feature extractor was adapted by removing the BatchNorm1d and max-pooling operations from each convolutional block.

E. Loss Function

To account for the strong imbalance between positive and negative labels, the model training used a binary cross-entropy loss with mini-batch-dependent positive-class weighting. For each mini-batch, positive examples, when present, were assigned a weight equal to the ratio between the numbers of negative and positive examples, while negative examples retained unit weight.

Let $y_i \in \{0, 1\}$ denote the supervised target assigned to the i-th training window, and let $\hat{y}_i \in [0, 1]$ denote the probability predicted by the model for the positive class. For a mini-batch containing $N = N_+ + N_-$ examples, where N_+ and N_- denote the numbers of positive and negative examples, respectively, the loss is defined as

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} w_i \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \quad (5)$$

where

$$w_i = \begin{cases} \dfrac{N_-}{N_+}, & \text{if } y_i = 1, \\ 1, & \text{if } y_i = 0. \end{cases} \quad (6)$$

For mini-batches containing no positive examples ($N_+ = 0$), the positive-class weighting factor was set to 1 in the implementation. This fallback had no effect on the loss because no positive samples were present in those mini-batches.

The positive-class weighting factor was recalculated for each mini-batch rather than being fixed according to the overall class distribution of the training partition. When both classes were represented in a mini-batch, the weighting scheme assigned the same total weight to each class. The sum of the weights assigned to the positive examples was $\sum_{i:y_i=1} w_i = N_+ \left(\frac{N_-}{N_+} \right) = N_-$, while the corresponding sum for the negative examples was $\sum_{i:y_i=0} w_i = N_-$. Therefore, the weighting balanced the total class weights within each mini-batch, although the actual loss contributions could still differ according to the predicted probabilities.

F. Model Training

During the training, additive Gaussian noise sampled from $\mathcal{N}(0, 0.005^2)$ was applied to the input windows as a data augmentation procedure. In addition, each window had a 20% probability of having one continuous segment, corresponding to 20% of its duration, replaced with zeros. These augmentations were not applied during validation or testing. No automated or systematic hyperparameter search was performed.

The network parameters were optimized using Adam with a learning rate of 0.001 and a mini-batch size of $B = 2048$

windows. Each fold-specific model was trained for up to 100 epochs. Early stopping was based on the validation PR-AUC, with a patience of 8 epochs, and the checkpoint achieving the highest validation PR-AUC was retained for subsequent evaluation. PR-AUC was used as the model-selection criterion because it summarizes the precision-recall trade-off across decision thresholds and therefore does not require the specification of a fixed classification threshold during model selection. The same architecture, optimizer, learning rate, batch size, number of epochs, data augmentation procedure, and remaining training settings were used for all three labeling strategies, ensuring that their comparison was conducted under the same training configuration.

G. Evaluation Protocol

The model performance was estimated using stratified five-fold cross-validation. The folds were constructed at the patient level, ensuring that all ECG recordings and all windows derived from the same patient were assigned to the same fold. In each cross-validation iteration, three folds were used for training, one fold for validation, and one fold for testing. The training, validation, and test partitions were therefore mutually exclusive at the patient level.

The same patient-level fold assignments were reused for strategies (i), (ii), and (iii), allowing the labeling strategies to be compared under equivalent data partitions. The supervised labels were reconstructed separately for each strategy before model training.

For each ECG recording, three consecutive 3-second windows were extracted, with a 1-second overlap between adjacent windows. Each window was independently processed by the model, and the three resulting predicted probabilities were averaged to obtain a single ECG-level probability. The model performance was therefore evaluated at the ECG level rather than at the individual-window level.

The performance was assessed using F1-score, recall, the area under the receiver operating characteristic curve (ROC-AUC), the area under the precision-recall curve (PR-AUC), and the Matthews correlation coefficient (MCC). ROC-AUC and PR-AUC were computed directly from the predicted probabilities, whereas F1-score, recall, and MCC were computed from binary predictions. Test-set performance was evaluated under two decision-threshold settings. First, a fixed threshold of 0.5 was applied to the test predictions. Second, in each cross-validation iteration, a decision threshold was selected by maximizing MCC on the corresponding validation partition and was subsequently applied to the test partition. For each labeling strategy, the reported results correspond to the mean and standard deviation of each metric across the five cross-validation iterations.

IV. Results and Discussion

The results are presented in two complementary analyses. First, the three operational labeling strategies are compared in terms of overall classification performance, with the objective of assessing how different supervised target definitions affect measured performance under the same experimental configuration. The results are examined using both a fixed decision threshold and thresholds selected from the validation data. Second, performance is examined across different ECG subgroups.

Table I presents the mean performance and standard deviation across the five test folds for the three labeling strategies. For threshold-dependent metrics, binary predictions were obtained using a fixed decision threshold of 0.5, whereas ROC-AUC and PR-AUC were computed directly from the predicted probabilities.

TABLE I
Test performance using a fixed threshold

Metric	Strategy (i)	Strategy (ii)	Strategy (iii)
F1-score	0.094 ± 0.014	$\mathbf{0.148 \pm 0.032}$	0.127 ± 0.020
Recall	$\mathbf{0.717 \pm 0.051}$	0.681 ± 0.100	0.606 ± 0.082
ROC-AUC	$\mathbf{0.803 \pm 0.006}$	0.798 ± 0.004	0.795 ± 0.007
PR-AUC	0.129 ± 0.016	$\mathbf{0.178 \pm 0.015}$	0.138 ± 0.014
MCC	0.137 ± 0.016	$\mathbf{0.176 \pm 0.021}$	0.161 ± 0.016

The strategy (ii) achieved the highest F1-score, PR-AUC and MCC, whereas strategy (i) achieved the highest recall and ROC-AUC. Strategy (iii) produced intermediate F1-score and MCC values but the lowest recall. The ROC-AUC values were similar across the three target definitions, ranging from 0.795 to 0.803, suggesting that the ranking ability of the model changed only modestly across the labeling strategies.

Table II reports the results obtained using validation-selected decision thresholds. For each fold, the decision threshold was selected by maximizing MCC on the corresponding validation set and was subsequently applied to the test set.

TABLE II
Test performance using MCC-optimized thresholds

Metric	Strategy (i)	Strategy (ii)	Strategy (iii)
F1-score	0.216 ± 0.009	$\mathbf{0.253 \pm 0.017}$	0.224 ± 0.013
Recall	0.273 ± 0.060	$\mathbf{0.295 \pm 0.083}$	0.292 ± 0.045
ROC-AUC	$\mathbf{0.803 \pm 0.006}$	0.798 ± 0.004	0.795 ± 0.007
PR-AUC	0.129 ± 0.016	$\mathbf{0.178 \pm 0.015}$	0.138 ± 0.014
MCC	0.204 ± 0.014	$\mathbf{0.235 \pm 0.013}$	0.210 ± 0.015

When the decision threshold was selected on the validation set by maximizing MCC, strategy (ii) achieved the highest F1-score and MCC. It also achieved the highest recall, which had been obtained by strategy (i) under the fixed threshold of 0.5.

The comparison of the two threshold settings shows that the choice of decision threshold affected the threshold-dependent metrics. For all three labeling strategies, validation-based threshold selection increased both F1-score and MCC relative to the fixed threshold of 0.5, while recall decreased markedly. As expected, ROC-AUC and PR-AUC remained unchanged because they are computed from the predicted probabilities and are independent of the decision threshold.

The performance differences observed among the labeling strategies should not be interpreted as evidence of superior clinical diagnostic accuracy for any strategy. Because the three strategies define the positive class differently, the resulting

targets differ in both composition and prevalence, which can affect metrics such as F1-score and PR-AUC. Accordingly, the reported results characterize performance under each operational target definition rather than against a common serologically confirmed reference.

A. Performance across ECG subgroups

To investigate whether model performance varies across different ECG presentations, performance was further evaluated across three mutually exclusive ECG subgroups based on the automatic ECG annotations available in CODE-15%: known abnormality, when at least one of six abnormalities was present, namely first-degree atrioventricular block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF), or sinus tachycardia (ST); unknown abnormality, when the ECG was classified as non-normal by the automatic annotation system but none of these six abnormalities was present; and normal ECG, when the ECG was classified as normal. For each strategy and subgroup, the decision threshold was selected by maximizing MCC on the corresponding validation set and then applied to the test set. The corresponding test results across the three ECG subgroups are reported in Table III.

TABLE III
TEST PERFORMANCE WITHIN ECG SUBGROUPS USING MCC-OPTIMIZED THRESHOLDS

Known abnormality ($N = 37{,}401$)			
Metric	Strategy (i)	Strategy (ii)	Strategy (iii)
Positive labels, n	2,099	3,230	2,289
F1	0.296±0.037	**0.354±0.017**	0.315±0.023
Precision	0.213±0.040	**0.296±0.029**	0.258±0.025
Recall	**0.518±0.117**	0.459±0.097	0.429±0.122
ROC-AUC	**0.802±0.018**	0.792±0.017	0.798±0.022
PR-AUC	0.216±0.029	**0.290±0.031**	0.237±0.022
MCC	0.266±0.034	**0.290±0.024**	0.273±0.033
Unknown abnormality ($N = 170{,}953$)			
Metric	Strategy (i)	Strategy (ii)	Strategy (iii)
Positive labels, n	3,394	5,339	3,702
F1	0.171±0.018	**0.205±0.022**	0.184±0.019
Precision	0.145±0.036	**0.203±0.052**	0.154±0.022
Recall	0.238±0.069	**0.251±0.102**	0.244±0.053
ROC-AUC	**0.769±0.008**	0.766±0.008	0.764±0.007
PR-AUC	0.102±0.018	**0.146±0.018**	0.107±0.019
MCC	0.159±0.016	**0.187±0.014**	0.169±0.017
Normal ECG ($N = 134{,}397$)			
Metric	Strategy (i)	Strategy (ii)	Strategy (iii)
Positive labels, n	1,066	1,658	1,191
F1	0.046±0.005	**0.065±0.006**	0.053±0.012
Precision	0.024±0.003	0.035±0.004	**0.043±0.036**
Recall	**0.448±0.090**	0.428±0.079	0.284±0.165
ROC-AUC	**0.743±0.010**	0.729±0.011	0.718±0.029
PR-AUC	0.029±0.004	**0.038±0.004**	0.030±0.007
MCC	0.075±0.010	**0.086±0.006**	0.067±0.009

As shown in Table III, all reported metrics were numerically higher in the known-abnormality subgroup than in the unknown-abnormality subgroup across the three labeling strategies. Within the known-abnormality subgroup, strategy (ii) achieved the highest F1-score, precision, PR-AUC, and MCC, whereas strategy (i) achieved the highest recall and ROC-AUC, with values of 0.518 and 0.802, respectively.

For ECGs with unknown abnormalities, strategy (ii) achieved the highest F1-score, precision, recall, PR-AUC, and MCC, whereas strategy (i) achieved the highest ROC-AUC. Among normal ECGs, F1-score, precision, ROC-AUC, PR-AUC, and MCC were below the corresponding values observed in both abnormality subgroups. Recall followed a different pattern, with values higher than those observed in the unknown-abnormality subgroup for all three strategies, but lower than those observed in the known-abnormality subgroup. Within the normal ECG subgroup, strategy (ii) achieved the highest F1-score, PR-AUC, and MCC, strategy (i) the highest recall and ROC-AUC, and strategy (iii) the highest precision.

Overall, the results show that classification outcomes differed across ECG subgroups, with the clearest contrast observed between ECGs with known abnormalities and normal ECGs. However, these subgroup differences do not establish that the model relied on any specific ECG abnormality for its predictions.

V. CONCLUSION AND FUTURE WORK

This study investigated how weak self-reported CD labels in CODE-15% can be operationalized for supervised ECG classification when multiple examinations are available for the same patient. The within-patient consistency analysis showed that 35.87% of the examinations belonging to patients who reported CD at least once were labeled as negative. Therefore, examination-level self-reports did not always provide a unique and internally consistent patient-level target.

Three labeling strategies were compared under the same model architecture, preprocessing procedure, training configuration, and patient-level cross-validation folds. Across both decision-threshold settings, the strategy that assigned a positive target to all examinations of a patient who reported CD at least once achieved the highest PR-AUC and provided the highest F1-score and MCC. Its PR-AUC was 0.178, compared with 0.129 for the original examination-level labels and 0.138 for the majority-based strategy. The original labels achieved the highest ROC-AUC, although the differences among the three strategies were small. For threshold-dependent metrics, the results also showed that performance depended on the decision rule: using thresholds selected by maximizing MCC on the validation set increased F1-score and MCC for all three strategies relative to a fixed threshold of 0.5, while reducing recall.

The subgroup analysis further showed that classification performance was not uniform across ECG presentations. ECGs with known abnormalities generally exhibited the strongest performance, whereas normal ECGs showed lower F1-score, precision, ROC-AUC, PR-AUC, and MCC. Strategy (ii) achieved the highest F1-score, PR-AUC, and MCC within all three ECG subgroups. These differences, however, should not be interpreted as evidence that the model relied on any specific ECG abnormality for its predictions.

The main lesson from this study is that target construction should be treated as an explicit modeling decision when repeated weak self-reported labels are available, rather than

as a fixed preprocessing choice. The observed differences across target definitions demonstrate that this decision can affect measured classification performance. At the same time, patient-level aggregation does not resolve the underlying label uncertainty or transform self-reported labels into clinically confirmed diagnoses. In particular, assigning a positive target based on a single positive record may propagate an erroneous self-report to all examinations from the same patient. Consequently, the reported results characterize agreement with alternative operational definitions of the available self-reported labels and should not be interpreted as estimates of clinical diagnostic accuracy for CD.

The study is limited by the use of self-reported labels without serological confirmation, the evaluation of a single model architecture, and the absence of formal statistical tests comparing the labeling strategies. Future work should address these limitations by evaluating the proposed target definitions in serology-based independent cohorts and additional model architectures, quantifying performance differences through appropriate statistical analyses, and investigating uncertainty-aware learning strategies for discordant self-reports.

Acknowledgment

The following Brazilian research agencies have supported this work: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) — Grants: 110945/2026-3, 140788/2024-7 and 201727/2025-0.

References

[1] K. C. F. Lidani, F. A. Andrade, L. Bavia, F. S. Damasceno, M. H. Beltrame, I. J. Messias-Reason, and M. T. M. Gomes, "Chagas Disease: From Discovery to a Worldwide Health Problem," *Frontiers in Public Health*, vol. 7, p. 166, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2019.00166/full

[2] C. Jidling, D. Gedon, T. B. Schön, C. D. L. Oliveira, C. S. Cardoso, A. M. Ferreira *et al.*, "Screening for Chagas disease from the electrocardiogram using a deep neural network," *PLOS Neglected Tropical Diseases*, vol. 17, no. 7, pp. 1–17, 07 2023. [Online]. Available: https://doi.org/10.1371/journal.pntd.0011118

[3] B. J. Singstad, N. O. Eidheim, A. R. Guruge, O. M. Lysaker, and V. Nunavath, "Auxiliary pretraining and fine-tuning across heterogeneous datasets for ECG-based Chagas disease detection," in *Computing in Cardiology*, vol. 52, 2025. [Online]. Available: https://cinc.org/archives/2025/pdf/CinC2025-179.pdf

[4] M. Sanmartino, A. Avaria Saavedra, J. Gómez i Prat, and P. Albajar-Viñas, "Chagas and health promotion: dialogue inspired by the Curitiba Statement," *Health Promotion International*, vol. 34, no. Supplement_1, pp. i82–i91, 03 2019. [Online]. Available: https://doi.org/10.1093/heapro/day105

[5] A. Champetier and C. Mueller, "Detection of Chagas disease using digital electrocardiogram by deep transfer learning of the InceptionTime model," in *Computing in Cardiology*, vol. 52, 2025. [Online]. Available: https://cinc.org/archives/2025/pdf/CinC2025-042.pdf

[6] M. A. Reyna, Z. Koscova, J. Pavlus, S. Saghafi, J. Weigle, A. Elola *et al.*, "Detection of Chagas disease from the ECG: The George B. Moody Physionet Challenge 2025," 2025, preprint at https://arxiv.org/abs/2510.02202.

[7] M. I. Traina, S. Hernandez, D. R. Sanchez, J. Dufani, M. Salih, A. M. Abuhamidah *et al.*, "Prevalence of Chagas Disease in a U.S. Population of Latin American Immigrants with Conduction Abnormalities on Electrocardiogram," *PLOS Neglected Tropical Diseases*, vol. 11, no. 1, pp. 1–8, 01 2017. [Online]. Available: https://doi.org/10.1371/journal.pntd.0005244

[8] J. A. Marin-Neto, A. R. Jr., G. M. M. Oliveira, L. C. L. Correia, A. N. R. Jr., A. O. Luquetti, A. M. Hasslocher-Moreno *et al.*, "Diretriz da SBC sobre Diagnóstico e Tratamento de Pacientes com Cardiomiopatia da Doença de Chagas – 2023," *Arquivos Brasileiros de Cardiologia*, vol. 120, no. 6, p. e20230269, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37377258/

[9] B. O. F. Brito, E. M. Lima, E. Z. Soliman, E. F. Silva, M. F. Lima-Costa, and A. L. P. Ribeiro, "The evolution of electrocardiographic abnormalities in the elderly with Chagas disease during 14 years of follow-up: The Bambui Cohort Study of Aging," *PLOS Neglected Tropical Diseases*, vol. 17, no. 6, pp. 1–16, 06 2023. [Online]. Available: https://doi.org/10.1371/journal.pntd.0011419

[10] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, p. 1760, 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-15432-4

[11] E. M. Lima, A. H. Ribeiro, G. M. M. Paixão, M. H. Ribeiro, M. M. Pinto-Filho, P. R. Gomes *et al.*, "Deep neural network-estimated electrocardiographic age as a mortality predictor," *Nature Communications*, vol. 12, p. 5117, 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-25351-7

[12] A. H. Ribeiro, G. M. M. Paixão, E. M. Lima *et al.*, "CODE-15%: a large scale annotated dataset of 12-lead ECGs," Zenodo, Jun. 2021. [Online]. Available: https://zenodo.org/records/4916206

[13] A. L. P. Ribeiro, A. H. Ribeiro, G. M. M. Paixão *et al.*, "SaMi-Trop: 12-lead ECG traces with age and mortality annotations," Zenodo, Jun. 2021. [Online]. Available: https://zenodo.org/records/4905618

[14] C. S. Cardoso, E. C. Sabino, C. D. L. Oliveira, L. C. Oliveira, A. M. Ferreira, E. Cunha-Neto *et al.*, "Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile," *BMJ Open*, vol. 6, no. 10, p. e011181, 2016. [Online]. Available: https://bmjopen.bmj.com/content/6/5/e011181

[15] E. C. Sabino, A. L. P. Ribeiro, V. M. C. Salemi, C. Di Lorenzo Oliveira, A. P. Antunes, M. M. Menezes *et al.*, "Ten-year incidence of Chagas cardiomyopathy among asymptomatic Trypanosoma cruzi-seropositive former blood donors," *Circulation*, vol. 127, no. 10, pp. 1105–1115, 2013. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.112.123612

[16] E. M. L. Aquino, S. M. Barreto, I. M. Bensenor, M. S. Carvalho, D. Chor, B. B. Duncan *et al.*, "Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): objectives and design," *American Journal of Epidemiology*, vol. 175, no. 4, pp. 315–324, 2012. [Online]. Available: https://academic.oup.com/aje/article/175/4/315/116011

[17] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning From Noisy Labels with Deep Neural Networks: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9729424

[18] S. Saini and M. Araujo, "Learning from leads: A 1D dilated ResNet for ECG Chagas disease screening," in *Computing in Cardiology*, vol. 52, 2025, pp. 1–4. [Online]. Available: https://cinc.org/archives/2025/pdf/CinC2025-408.pdf

[19] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 590–597. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3834

[20] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human Uncertainty Makes Classification More Robust," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9616–9625. [Online]. Available: https://ieeexplore.ieee.org/document/9010969

[21] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2020. [Online]. Available: https://doi.org/10.1007/s10994-019-05855-6

[22] X. Zhai, Z. Zhou, and C. Tin, "Semi-Supervised Learning for ECG Classification without Patient-Specific Labeled Data," *Expert Systems with Applications*, vol. 158, p. 113411, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420302359

[23] N. Qaqos, E. N. Obianom, S. I. Abdullahi *et al.*, "Classification of Chagas Disease Using Convolutional Neural Network and Random Forest," in *Computing in Cardiology Conference*, vol. 52, 2025. [Online]. Available: https://cinc.org/archives/2025/pdf/CinC2025-144.pdf

Exploratory Analysis of Holo-Hilbert EEG Descriptors in Autism Spectrum Disorder

Maria Fernanda de Leão Fava
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
mariaflfava@gmail.com

Helder Luiz de Souza da Silva Filho
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
helder3345@gmail.com

Suzana Cescon de Souza
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
suzanacescon@ufpa.br

Antônio Pereira Junior
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
apereira@ufpa.br

***Abstract*—This exploratory study examines statistical descriptors of Holo-Hilbert Spectral Analysis (HHSA) maps from six participants with autism spectrum disorder (ASD) and six neurotypical controls. Eleven descriptors were computed per map and averaged across parietal channels P3/P4 and pre-/post-task recordings. Five originally reported group comparisons had nominal $p < 0.05$; none remains below 0.05 after Holm adjustment across the 11 reported tests. The test provenance and individual observations are unavailable for verification. Previously reported leave-one-subject-out classification accuracy reached 83.33% (10/12), but feature selection on the complete sample creates a risk of information leakage, and unbiased predictive performance has not been established. The results provide preliminary descriptive observations, with unresolved issues concerning decomposition settings, matrix normalization, numerical resolution, clinical characterization, overfitting, and restricted channel coverage. They do not establish a diagnostic marker or superiority over alternative approaches.**

***Index Terms*—Autism Spectrum Disorder, Electroencephalography, Holo-Hilbert Spectral Analysis, Machine Learning, Classification.**

I. Introduction

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental condition involving differences in social communication and restricted or repetitive behavior [1]. EEG offers high temporal resolution for studying brain activity, and previous reviews have examined EEG abnormalities and classification strategies in ASD [2]–[4]. However, distinguishing clinical groups in a small convenience sample does not establish diagnostic validity.

Holo-Hilbert Spectral Analysis (HHSA) extends empirical mode decomposition and Hilbert spectral analysis by resolving amplitude modulation alongside carrier frequency [5], [6]. Applications to EEG and cognitive assessment illustrate its potential for describing complex signal dynamics [7], [8]. This representation motivates the study of distributional descriptors; its additional dimensions alone do not establish an advantage for ASD classification.

This study explores descriptors of HHSA maps from P3 and P4 before and after a visual task. Its available subject-level results concern averages over the available maps. Accordingly, the present analysis describes aggregate between-group differences and previously reported classification experiments. Both inferential and predictive claims are limited by the small sample and incomplete analytical records.

The contribution is exploratory: it combines an interpretable set of distributional descriptors with conventional group summaries, descriptor correlations, and subject-level classifier results. This organization facilitates hypothesis generation, but it does not by itself establish that the original inferential tests were adequate or that the classifiers generalize beyond this sample.

II. Materials and Methods

A. Participants and Experimental Protocol

The dataset consisted of 12 participants, including six individuals diagnosed with ASD Level 1 according to DSM-5 criteria and six neurotypical controls. Participants were between 20 and 28 years old. Inclusion criteria comprised age between 18 and 30 years, normal or corrected-to-normal vision, and the ability to follow simple visual and auditory instructions. Individuals with a history of epilepsy, brain injury, or other neurological conditions potentially affecting EEG activity were excluded. All participants provided informed consent prior to participation. The study was approved by the Research Ethics Committee of the Federal University of Pará (CAAE 56365722.9.0000.0018).

The overall age range, ASD support level, eligibility criteria, consent, and ethics identifier above are retained as reported in the supplied manuscript; individual records were not provided for independent verification. Group-specific age summaries, sex distribution, cognitive assessments, ASD severity scores and diagnostic instruments, medication use, and relevant comorbidities are unavailable. Consequently, demographic and



Fig. 1. Example of the computerized visual cancellation task adapted from the Cars Test, in which participants identified a target stimulus among multiple distractors.

clinical balance cannot be assessed, and confounding remains unresolved.

The experimental protocol followed the methodology proposed by Silva Filho et al. [9]. EEG signals were recorded before and after the execution of a computerized visual cancellation task adapted from the Cars Test proposed by Cascaes et al. [10]. Participants remained seated approximately 60 cm from a 21.5-inch monitor in a controlled environment with indirect lighting, controlled temperature, and reduced acoustic interference throughout the experiment.

Figure 1 illustrates the computerized visual cancellation task adapted from the Cars Test, in which participants were instructed to identify a target stimulus among multiple distractors using a computer mouse.

Initially, a 3-minute resting-state EEG recording was acquired while participants observed a blank screen and were instructed to remain relaxed, minimizing blinking, swallowing, and voluntary movements. Subsequently, participants conducted the visual cancellation task for 7 min and 30 s. After each correct selection, both the target and distractor positions were automatically updated, maintaining a continuous visual search paradigm throughout the experiment. Immediately after task completion, a second resting-state EEG recording was scheduled for 3 minutes and acquired under the same conditions as the initial recording. These are protocol durations; the effective duration retained for each HHSA map and the amount rejected as artifacts are not documented in the supplied package. Only the pre-task and post-task resting-state EEG segments were considered in the present study, as these recordings were used to generate the HHSA spectral representations analyzed in the subsequent stages.

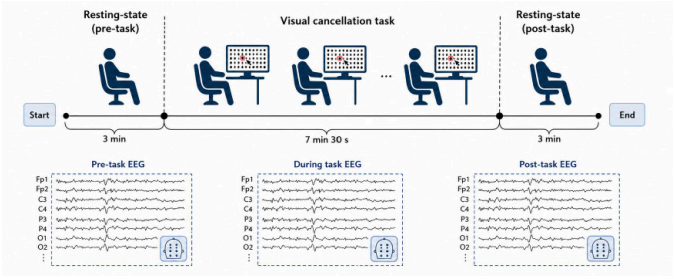


Fig. 2. Experimental protocol consisting of resting-state EEG acquisition before and after the computerized visual cancellation task.

B. HHSA Decomposition and Reproducibility

The original manuscript reports visual inspection, 0.5–40 Hz band-pass filtering, independent component analysis (ICA), channel selection, and temporal segmentation, referring to Silva Filho [9]. Acquisition and preprocessing logs and executable decomposition scripts were not supplied. Therefore, these reported operations cannot be independently verified. Sampling frequency, EEG reference, filter order and phase response, ICA algorithm, channels entered into ICA, component rejection criteria, and numbers of removed components or rejected segments must be recovered from study records.

Conceptually, EMD decomposes a signal into intrinsic mode functions (IMFs). Their analytic representations provide instantaneous amplitude and carrier frequency. A further decomposition of amplitude envelopes resolves modulation frequencies [5], [6]. This conceptual description is not a complete implementation specification. The EMD variant, stopping criterion, maximum sifting iterations, boundary treatment, handling of mode mixing, IMF inclusion/exclusion criteria at both layers, and envelope/frequency estimator are not documented in the supplied package. Neither the numerical frequency grid nor spectral weighting, temporal aggregation, smoothing, and normalization can be confirmed. No specific settings are assumed in this revision.

These omissions prevent an independent reproduction of the HHSA maps. The distinction between signal preprocessing, numerical spectral estimation, and display rendering must be maintained when the original code is recovered.

C. HHSA Map Processing

The HHSA spectral representations provided by Silva Filho et al. [9] were stored as MATLAB figure files (`.fig`), containing the original numerical matrices used to generate the spectral maps. To enable quantitative analysis, a custom MATLAB function, named *getHHSA*, was developed to automatically extract these data.

For each `.fig` file, the function opens the figure in invisible mode, identifies the axis corresponding to the HHSA representation, and retrieves the numerical matrix associated with the HHSA image object. The reported function applies a vertical flip (`flipud`). Correct orientation must be checked against the stored axis coordinates and axis direction, rather than assuming that every MATLAB image requires inversion.

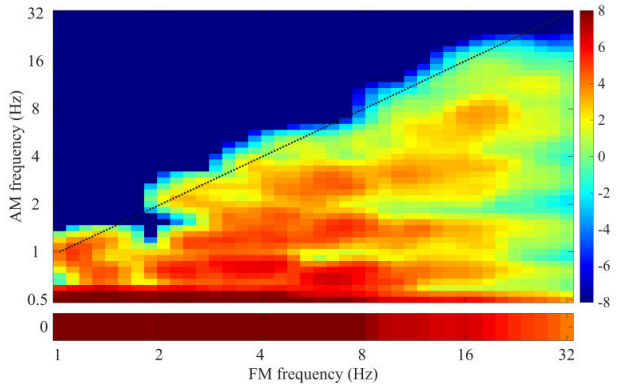


Fig. 3. Representative HHSA spectral map extracted from channel P3 during the first resting-state epoch (Part 1).

A vertical flip permutes entries and therefore leaves the full-matrix distributional descriptors unchanged.

The described extraction is intended to avoid manual reading of plotted colors. However, the supplied archive contains raster illustrations, not the underlying `.fig` files or *getHHSA* source. It is therefore unverified whether the extracted `CData` contains physical spectral values, transformed values, indexed colors, or RGB data. Numerical matrix values must not be equated with physical energy without checking these properties. Figure 3 illustrates an example of an HHSA spectral map extracted from channel P3 during the first resting-state epoch (Part 1). In this representation, frequency modulation (FM) is displayed along the horizontal axis, amplitude modulation (AM) along the vertical axis, and the color scale displays the supplied map values; their units and transformation require verification.

D. Feature Extraction

Eleven statistical descriptors were extracted from each HHSA spectral matrix: mean, standard deviation, variance, energy, entropy, skewness, kurtosis, coefficient of variation, median, interquartile range (IQR), and robust range. These descriptors characterize central tendency, dispersion, asymmetry, and distributional shape of the matrix values.

Four HHSA maps were available for each participant (P3-Part1, P3-Part3, P4-Part1, and P4-Part3). Feature values were computed independently for each map and then averaged to obtain one subject-level descriptor. If D denotes a descriptor and H_{itc} a participant's matrix, the reported quantity is

$$\bar{D}_i = \frac{1}{4} \sum_{t\in\{1,3\}} \sum_{c\in\{P3,P4\}} D(H_{itc}). \tag{1}$$

This is an average of descriptors, not a descriptor of an averaged map.

E. Statistical Analysis and Scope of Inference

The original manuscript reports Student's t and Mann–Whitney U tests and Cohen's d, but supplies only one unidentified p column per feature. The two tests are not interchangeable; the table does not identify which generated each value, whether tests were two-sided, or whether variance assumptions were checked. Raw subject-level features, test statistics, degrees of freedom, and confidence intervals were not supplied. They cannot be recovered reliably by combining rounded summaries with an assumed test. The original d values are retained as unverified standardized effect estimates, with ASD minus control as the reported direction.

For this revision, Holm adjustment was applied to the eleven published, rounded p values as one family. If $p_{(j)}$ are ordered values, then $\tilde{p}_{(k)} = \min\{1, \max_{j\le k}(11-j+1)p_{(j)}\}$. This checks multiplicity for the displayed column only; it does not validate the original tests or cover an undisclosed search across both tests and additional analyses. Nominal and adjusted values are distinguished throughout.

F. Supervised Classification and Validation

The original experiments report Logistic Regression, linear SVM, RBF SVM, and Random Forest under leave-one-subject-out (LOSO) validation. These model families have been used in EEG-based ASD classification studies, but prior performance reports do not remove the need for leakage-controlled evaluation [3], [11]. Eleven participants would form each training fold and one participant the test fold. However, the described workflow first selects nominally significant features using the complete dataset and then performs LOSO. Unless selection was repeated using training participants alone, the held-out participant's label influenced feature selection, introducing information leakage. Selecting the best feature/classifier combination on the same LOSO results adds selection bias [12].

Accordingly, the retained accuracy table is a record of exploratory results, not a validated performance estimate. A corrected evaluation must fit feature selection and scaling within each outer training fold, tune hyperparameters in an inner subject-level validation loop, and predict the untouched outer participant. It must specify software versions, random seeds, regularization and kernel settings, forest parameters, and candidate search space, and provide predictions per participant, confusion matrices, sensitivity, specificity, balanced accuracy, and uncertainty. These records are absent, so neither leakage-free performance nor its confidence interval is reported here as established.

III. Results

A. Reported Group Summaries and Multiplicity Audit

Five of the eleven displayed comparisons have nominal $p < 0.05$. None has Holm-adjusted $p < 0.05$ (Table I). The absence of corrected significance does not demonstrate equivalence, particularly with six participants per group. These adjustments are conditional on the validity and identification of the original tests, which remain unresolved.

The reported standardized effects for dispersion descriptors are large in magnitude, but their uncertainty is missing and small samples are sensitive to individual observations. They should not be taken as proof of stable population differences.

TABLE I

REPORTED SUBJECT-LEVEL SUMMARIES (MEAN ± SD) AND HOLM AUDIT OF THE ORIGINAL p COLUMN. TEST IDENTITY AND ORIGINAL EFFECT ESTIMATES REQUIRE VERIFICATION; NO ADJUSTED $p < 0.05$.

Feature	ASD	Control	p reported	p Holm	d reported
Mean	0.568 ± 0.063	0.587 ± 0.055	0.6015	1.0000	-0.311
Std	0.302 ± 0.024	0.267 ± 0.018	0.0173	0.1730	1.644
Variance	0.092 ± 0.015	0.072 ± 0.011	0.0230	0.2070	1.548
Energy	21091.992 ± 3156.606	21185.766 ± 3002.738	0.9590	1.0000	-0.030
Entropy	5.368 ± 0.090	5.439 ± 0.121	0.2721	1.0000	-0.671
Skewness	-0.328 ± 0.329	-0.433 ± 0.301	0.5773	1.0000	0.333
Kurtosis	1.920 ± 0.296	2.445 ± 0.390	0.0253	0.2070	-1.517
CoeffVar	0.542 ± 0.096	0.466 ± 0.069	0.1432	0.8592	0.917
Median	0.628 ± 0.116	0.615 ± 0.088	0.8342	1.0000	0.124
IQR	0.520 ± 0.078	0.403 ± 0.091	0.0372	0.2604	1.386
Robust Range	0.797 ± 0.055	0.718 ± 0.033	0.0132	0.1452	1.735

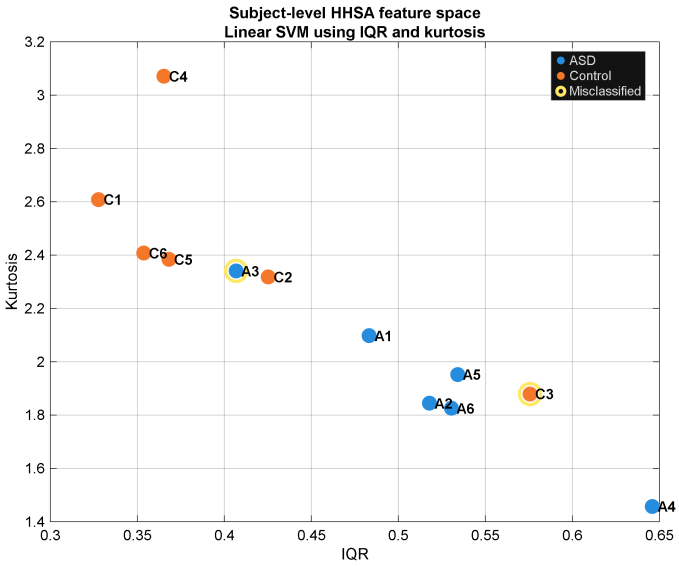


Fig. 4. Supplied subject-level IQR–kurtosis visualization. Error labels refer to the original SVM analysis; validation provenance is unresolved. The plot is descriptive.

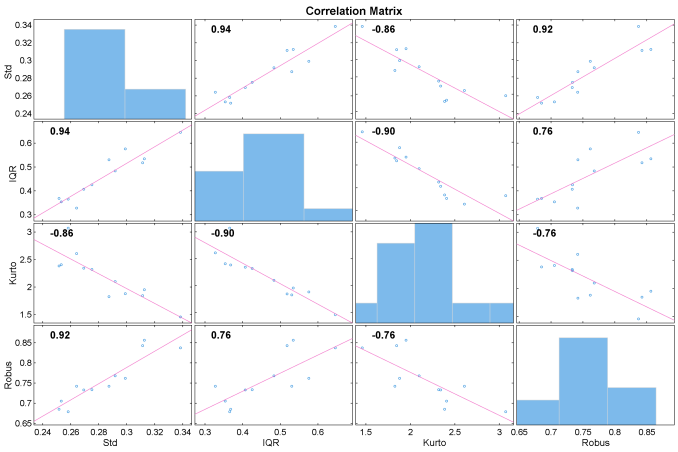


Fig. 5. Supplied Pearson correlations for standard deviation, IQR, kurtosis, and robust range; variance is not displayed. These four descriptors were selected in the original nominal analysis, and none passes the current Holm adjustment. Correlations are descriptive and suggest redundancy.

The original test identity must be checked before assigning inferential meaning to either nominal or adjusted values.

Figure 4 retains the supplied IQR–kurtosis scatter plot as a descriptive visualization. Its labeling of C3 and A3 as classification errors comes from the original analysis and cannot be independently verified without outer-fold predictions. Visible separation does not validate a classifier or measure adaptive reorganization.

The original Pearson correlations include standard deviation–IQR ($r = 0.94$), standard deviation–robust range ($r = 0.92$), kurtosis–IQR ($r = -0.90$), and kurtosis–standard deviation ($r = -0.86$). These are descriptive pooled-sample correlations, not independent evidence of a group effect. Raw data and confidence intervals were not supplied. Group separation itself can contribute to pooled correlations (Fig. 5).

B. Previously Reported Classification Results

Table II retains the originally reported accuracies for traceability. The maximum, 83.33%, represents 10 correct predictions among 12 participants; one prediction changes accuracy by 8.33 percentage points. The feature-selection and model-selection concerns described above prevent interpreting these values as unbiased estimates of generalization.

The original table gives the same maximum accuracy for several feature/model choices. With 12 participants and no independent evaluation, these small differences do not establish a ranking of algorithms or a preferred descriptor. Correlated descriptors may be redundant, but the table alone cannot establish the mechanism behind the reported performance.

TABLE II
ORIGINALLY REPORTED LOSO ACCURACY (%); FEATURE SELECTION WAS DESCRIBED BEFORE LOSO, SO THESE VALUES ARE EXPLORATORY AND POTENTIALLY OPTIMISTIC, NOT INDEPENDENTLY VALIDATED

Feature(s)	Logistic	Linear SVM	Random Forest	RBF-SVM
IQR	**83.33**	**83.33**	66.67	66.67
Kurtosis	75.00	**83.33**	58.33	58.33
Std	75.00	75.00	58.33	66.67
Variance	75.00	75.00	58.33	66.67
Robust Range	75.00	66.67	58.33	58.33
IQR + Kurtosis	**83.33**	**83.33**	75.00	66.67
Five nominal features	**83.33**	**83.33**	66.67	66.67

IV. DISCUSSION

A. Interpretation and Limits

The available summaries suggest descriptive differences in the spread and shape of HHSA matrix values. Similar EEG studies have reported atypical resting-state activity and parietal EEG–fMRI relationships in ASD, but those findings are not direct validation of the present HHSA descriptors [4], [13]. All eleven displayed tests are nonsignificant after Holm correction. Flattened-map descriptors discard the locations of values on the frequency axes; different spectral arrangements can therefore produce the same descriptors. These measures do not operationalize neural flexibility, adaptive reorganization, physiological recovery, synchronization, or theta modulation. No such mechanism or functional marker is established here.

The sample of six participants per group provides imprecise estimates and is sensitive to outliers. Missing clinical and demographic information prevents assessing confounding. Incomplete decomposition, artifact-handling, and normalization records limit reproducibility. Pooled correlations and retrospective selection of favorable features add further uncertainty.

The use of Random Forest and SVM with only 12 subjects creates a substantial risk of overfitting, even under leave-one-subject-out validation. The reported 83.33% accuracy corresponds to only 10 of 12 predictions, and one prediction changes the accuracy by 8.33 percentage points. It should therefore be interpreted as an exploratory, potentially optimistic result rather than evidence of generalizable performance. Independent validation on a larger cohort and leakage-free nested model selection would be required to assess predictive value.

The analysis was restricted to P3 and P4 and therefore does not establish whether other cortical regions contribute to group discrimination. The supplied study also contains no matched quantitative benchmark using conventional spectral or time–frequency features, nor a deep-learning model evaluated on the same subjects and folds. Such comparisons would require the original EEG segments, identical artifact handling, training-only preprocessing, and independent tuning; applying a high-capacity deep-learning model to 12 subjects without those controls would likely increase overfitting risk. Future evaluations should report sensitivity, specificity, balanced accuracy, precision, recall, F1 score, confusion matrices, and uncertainty alongside accuracy. These additional analyses are not available in the present archive, so no superiority of HHSA or of the reported classifiers is claimed.

V. CONCLUSION

This small exploratory study describes subject-averaged HHSA map features from ASD and control participants. Five nominal group differences do not survive Holm correction across the eleven reported comparisons. The maximum originally reported classification accuracy of 83.33% remains unvalidated because of overfitting risk, unresolved feature-selection leakage, and unavailable analytical records. Physiological mechanisms cannot be inferred from the averaged descriptors. Complete numerical data and code, broader channel coverage, leakage-free classification with additional performance metrics, and matched comparisons with conventional and deep-learning approaches are needed before drawing stronger conclusions or considering clinical applications.

REFERENCES

[1] C. Lord, T. S. Brugha, T. Charman, J. Cusack, G. Dumas, T. Frazier, E. J. Jones, R. M. Jones, A. Pickles, M. W. State *et al.*, "Autism spectrum disorder," *Nature Reviews Disease Primers*, vol. 6, no. 1, p. 5, 2020.

[2] C. Bosetti, L. Ferrini, A. R. Ferrari, E. Bartolini, and S. Calderoni, "Children with autism spectrum disorder and abnormalities of clinical eeg: A qualitative review," *Journal of Clinical Medicine*, vol. 13, no. 1, p. 279, 2024.

[3] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, and S. Jayarathna, "Eeg-based processing and classification methodologies for autism spectrum disorder: A review," *Journal of Computer Science*, vol. 15, no. 8, pp. 1161–1183, 2019.

[4] W. S. Neo, D. Foti, B. Keehn, and B. Kelleher, "Resting-state eeg power differences in autism spectrum disorder: a systematic review and meta-analysis," *Translational Psychiatry*, vol. 13, no. 1, p. 389, 2023.

[5] N. E. Huang, Z. Shen, S. R. Long *et al.*, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A*, vol. 454, no. 1971, pp. 903–995, 1998.

[6] N. E. Huang, K. Hu, A. C. C. Yang, H.-C. Chang, D. Jia, W.-K. Liang, J. R. Yeh, C.-L. Kao, C.-H. Juan, C. K. Peng, J. H. Meijer, Y.-H. Wang, S. R. Long, and Z. Wu, "On holo-hilbert spectral analysis: a full informational spectral representation for nonlinear and non-stationary data," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, p. 20150206, 2016.

[7] C.-Y. Chang *et al.*, "Evaluating cognitive impairment using holo-hilbert spectral analysis of eeg signals," *Frontiers in Aging Neuroscience*, vol. 14, 2022.

[8] Y.-H. Chu *et al.*, "Holo-hilbert spectral analysis for eeg-based cognitive assessment," *Biomedical Signal Processing and Control*, vol. 84, 2023.

[9] H. L. de Souza da Silva Filho, "Análise do espectro holo-hilbert aplicada à caracterização de indivíduos com transtorno do espectro autista," Belém, Brazil, 2025. [Online]. Available: https://bdm.ufpa.br/handle/prefix/8919

[10] R. Cascaes, K. Lameira, R. Sarmanho, S. Santos, K. Pinheiro, M. P. Mota, A. Pereira, and N. C. Sampaio Neto, "An empirical study on the adaptation and automation of a cancellation test for children," *Journal on Interactive Systems*, vol. 10, no. 2, pp. 82–95, 2019.

[11] P. Peketi *et al.*, "Machine learning analysis of eeg p300 responses for autism spectrum disorder classification," *Biomedical Signal Processing and Control*, vol. 80, 2023.

[12] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010. [Online]. Available: https://www.jmlr.org/papers/v11/cawley10a.html

[13] L. E. Mash, B. Keehn, A. C. Linke, T. T. Liu, J. L. Helm, F. Haist, J. Townsend, and R.-A. Müller, "Atypical relationships between spontaneous EEG and fMRI activity in autism," *Brain Connectivity*, vol. 10, no. 1, pp. 18–28, 2020.

Effect of Peripheral Sinusoidal Electrical Stimulation on Cortical Oscillations During an Archimedean Spiral Drawing Task

Evelyn Juliette Pesántez–Vera
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0009-0009-0019-6713

Danilo Andrés Molina–Vidal
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0000-0003-4420-4585

Pedro Sebastián Silverio Cevallos
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0009-0002-5712-4699

Carlos Julio Tierra–Criollo
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0000-0003-4880-9186

***Abstract*—This exploratory study evaluated the feasibility of a protocol to characterize movement-related cortical oscillations and motor performance during peripheral sinusoidal electrical stimulation. Ten right-handed, neurologically healthy volunteers aged 20 to 35 years participated in the study. Participants performed the Archimedean spiral drawing task before, during, and after a 2-kHz sinusoidal stimulation applied alternately over the median and radial nerves. Electroencephalography and accelerometry signals were recorded, along with the coordinates of the drawing trajectory. Event-Related Desynchronization and Synchronization (ERD/ERS) patterns in the alpha (8–13 Hz) and beta (13–30 Hz) bands were analyzed during the Pre-treatment and Treatment stages and statistical comparisons were performed to assess changes relative to the reference period and between stages. Motor parameters were also compared across stages. In both frequency bands, ERD was observed during movement preparation and execution, with greater predominance in contralateral channels. After the task, alpha energy recovery was slow and incomplete, whereas beta power showed an increase consistent with post-movement rebound. No significant differences were identified in ERD/ERS values between stages, although large effect sizes were observed in the post-movement beta response. No significant differences were observed in motor parameters either, and their global effects were small. The results demonstrate the feasibility of the methodology but suggest that further methodological adjustments are required before its application in patients with Parkinson's disease.**

***Index Terms*—ERD/ERS, Peripheral sinusoidal electrical stimulation, Archimedean spiral, Accelerometer.**

I. Introduction

Voluntary movement relies on the complex integration of neurophysiological mechanisms, including the dynamics of cortical oscillations and the coordination of the sensorimotor network [1]. Studies based on electroencephalography (EEG) and magnetoencephalography (MEG) have identified characteristic patterns in the alpha (8–13 Hz) and beta (13–30 Hz) bands during movement preparation, execution, and the post-movement period, known as event-related desynchronization and synchronization (ERD/ERS) [2].

ERD corresponds to a decrease in spectral energy, associated with increased cortical activation, whereas ERS reflects an increase in energy, related to neuronal resynchronization and states of reduced functional demand [2]. During movement preparation and execution, ERD is typically observed in sensorimotor areas, whereas after movement completion, beta-band ERS may occur, known as the post-movement beta rebound [3].

In neurological disorders that affect sensorimotor circuits, such as Parkinson's disease (PD), alterations in ERD/ERS patterns have been reported, reflecting dysfunction in the dynamics of cortical oscillations [4]. In PD, different non-invasive approaches have been investigated to modulate sensorimotor circuits and improve motor control [5], [6]. Among these approaches, transcutaneous afferent patterned stimulation (TAPS) has received, recently, particular attention. This technique consists of alternating stimulation of peripheral nerves, usually the median and radial nerves, adjusted according to the individual tremor frequency, which is estimated from accelerometry signals [7].

TAPS is typically delivered using current pulses with a pulse width of approximately 300 µs at a frequency of 150 Hz, which preferentially activate $A\beta$ fibers involved in the transmission of somatosensory information [6], [8], [9]. As an alternative to pulsed waveforms, a sinusoidal waveform may be used. Previous studies indicate that stimulation in the 2000–3000 Hz range may promote more selective activation of $A\beta$ fibers without causing pain or discomfort [10]–[12]. In a recent study, high-frequency sinusoidal peripheral electrical stimulation (10 kHz) was applied to patients with PD using

current intensities close to the motor threshold for 20 minutes. The results showed a reduction in kinetic tremor immediately after stimulation, with effects persisting for up to 10 minutes after the intervention [13].

In addition to EEG analysis, it is important to complement the results with objective measures of motor performance. For this purpose, the Archimedean spiral drawing task has been used as a standardized test to assess tremor and fine motor control, particularly in patients with PD. Through digitization of the drawing trajectory, metrics related to movement smoothness, regularity, and drawing accuracy can be computed, providing a quantitative and objective measure of motor performance [14]–[16].

To the best of our knowledge, cortical oscillations during the Archimedean spiral drawing task under TAPS-inspired sinusoidal peripheral electrical stimulation have not yet been investigated. Therefore, this study aimed to evaluate the feasibility of an experimental protocol to characterize ERD/ERS patterns and motor performance before, during, and after this type of stimulation. As this was a preliminary and exploratory study, the protocol was first applied to neurologically healthy participants in order to optimize the stimulation, acquisition, and processing procedures before its application in patients with PD.

II. Materials and Methods

A. System for the Assessment of Cortical Oscillations

The system for the assessment of cortical oscillations comprises different subsystems that operate in an integrated manner, as illustrated in Figure 1.

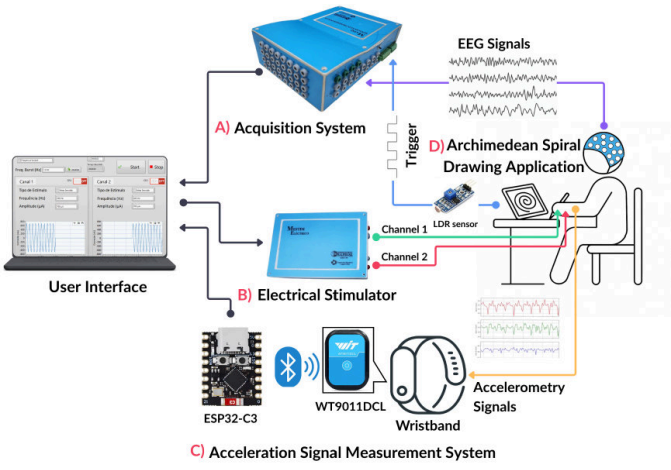


Fig. 1. System for the assessment of cortical oscillations. A) Acquisition system; B) Electrical stimulator; C) Acceleration signal measurement system; and D) Archimedean spiral drawing application.

1) ***MADQ Acquisition System*** *(Fig. 1A):* The MADQ biological signal acquisition system comprises 40 channels, 24-bit analog-to-digital converters, programmable gain, and a sampling rate of up to 16 kHz, in addition to integrated *anti-aliasing* filters. In this study, a sampling frequency of 8 kHz and a gain of 6 were configured. The acquired signals are transmitted to a computer for storage and subsequent processing. Device configuration and signal visualization are performed through an interface developed in LabVIEW [17].

2) ***MESTIM_ELECT Electrical Stimulator*** *(Fig. 1B):* The MESTIM_ELECT is a dual-channel current source capable of generating sinusoidal, square, triangular, sawtooth, and arbitrary waveforms. Each channel can deliver up to 10 mA. It includes a LabVIEW interface for configuring stimulation protocols and adjusting parameters such as frequency, amplitude, duration, and signal type [18]. The MADQ acquisition system and the MESTIM_ELECT electrical stimulator were developed in collaboration between the Federal University of Rio de Janeiro (UFRJ) and Salesian Polytechnic University (UPS, Ecuador).

3) ***Acceleration Signal Measurement System*** *(Fig. 1C):* The system consists of a WT9011DCL inertial sensor, an ESP32-C3 microcontroller, and a graphical interface developed in LabVIEW. The sensor integrates a triaxial accelerometer, gyroscope, and magnetometer (±16 g) and was configured with a cutoff frequency of 20 Hz and a sampling rate of 50 Hz. The device was attached to a wristband to facilitate its use during the experiment. Communication between the sensor and the ESP32-C3 is performed through Bluetooth Low Energy (BLE), with the microcontroller being responsible for data acquisition and retransmission. The acquired signals are visualized and stored in real time, allowing both time-domain signals and frequency spectra to be analyzed. In this study, the system was used to estimate the power spectral density (PSD) of the acceleration signals and identify the dominant frequency during the motor task [19].

4) ***Archimedean Spiral Drawing Application*** *(Fig. 1D):* The application was developed in HTML5 and JavaScript, based on the TRSPER application [16], and can be run directly in a web browser without installation. In the initial interface, the participant identifier is requested and a button is provided to start the experiment. After activation, a reference Archimedean spiral is displayed on the screen.

The protocol starts when the participant places the digital pen at the center of the spiral. This event automatically starts a 3-second countdown corresponding to the motor preparation phase. At the end of this phase, drawing is enabled for a 10-second period, during which the participant must follow the spiral trajectory displayed on the screen. Throughout task execution, the spatial coordinates (x, y), in pixels, of both the reference spiral and the trajectory drawn by the participant are continuously recorded, along with the execution time in milliseconds

After each trial, drawing is disabled and a 10 second rest period begins. A new spiral is then presented to the participant. This procedure is repeated until seven consecutive drawings are completed. At the end of the experiment, all recorded data can be exported in `.csv` format for subsequent analysis.

In addition to recording the motor task, the application incorporates a temporal synchronization mechanism based on visual stimuli. For this purpose, a 1 cm $\times$ 1 cm square was added to the lower-left region of the screen and was

used exclusively to generate temporal markers detected by an external LDR sensor. At the beginning of the motor preparation phase, the square changes its color from black to white for 2 s, producing a first temporal marker associated with the onset of movement preparation. Subsequently, when the participant begins tracing the spiral, the square remains white throughout task execution and returns to black at movement completion. The changes in luminosity were converted into trigger signals and sent to the acquisition system, allowing temporal synchronization between the task and the EEG signal recordings.

The application was run on a Samsung Galaxy Tab S10 Lite tablet equipped with a 10.9-inch display and an Exynos 1380 processor. Interaction with the interface was performed using the S Pen digital stylus, which has a 0.7-mm tip, 4096 pressure levels, a sampling rate of 240 Hz, and low latency.

B. Participants

An exploratory study was conducted with 10 right-handed participants, with a balanced distribution between sexes, aged 20 to 35 years (24.9 $\pm$ 4.4 years). Participants had no history of neurodegenerative diseases, diabetes, or upper-limb injuries. In addition, they had no biomedical implants and were not using medications or other substances that could interfere with the functioning of the sensorimotor system.

The experiments were conducted in quiet rooms free of distractions, with controlled ambient temperature (24 ºC $\pm$ 1 ºC), within the facilities of the Neural Engineering Laboratory of LAPIS (Laboratory of Biomedical Image and Signal Processing), at NTIEB (Center for Technology in Biomedical Engineering / COPPE). Before the experimental procedures began, participants were informed about the study procedures and signed the Informed Consent Form (TCLE). The study was approved by the Research Ethics Committee under CAAE number 58744222.2.1001.5257. The experimental protocol consisted of three stages: Pre-treatment, Treatment, and Post-treatment (Fig. 2).

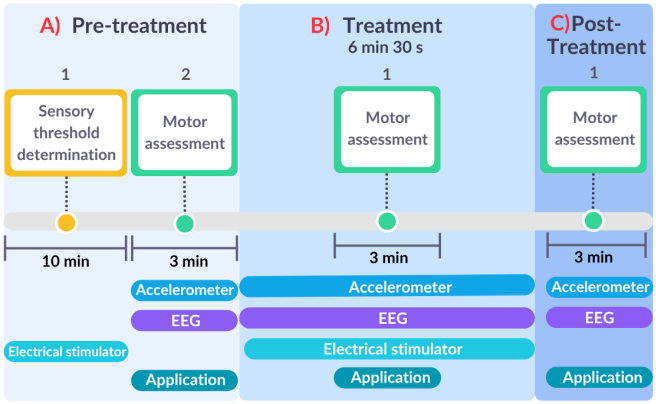


Fig. 2. General overview of the experimental protocol. A) Pre-treatment stage; B) Treatment stage; and C) Post-treatment stage.

1) ***Pre-treatment Stage*** *(Fig. 2A):* At the beginning of the experiment, participants were instructed to sit in a comfortable position, with their forearms resting on a table and their arms extended and relaxed. The sensory thresholds (ST) for a 2-kHz sinusoidal stimulus were then determined (Fig. 2A.1). For this purpose, disposable gel electrodes (1 cm in diameter) were positioned over the median and radial nerves of the dominant hand and fixed with microporous tape. Before electrode placement, the contact areas were cleaned with 70% alcohol to reduce skin impedance. The ST values were determined as described in [10], [18].

During the motor assessment (Fig. 2A.2), EEG and accelerometry signals were recorded. For EEG acquisition, silver/silver chloride (Ag/AgCl) electrodes were positioned according to the international 10-10 system, using an EEG cap and conductive gel to ensure low impedance. Fifteen derivations were used (F3, F4, FC3, FCz, FC4, C3, C4, CP3, CPz, CP4, P3, Pz, P4, O1, O2), with Cz used as the reference and a ground electrode located between Fz and FPz. In addition, an electrode was placed at Erb's point contralateral to the stimulated limb, located in the supraclavicular fossa approximately 1–2 cm above the midpoint of the clavicle, in order to monitor the artifact introduced by electrical stimulation into the EEG signals. To record motor activity, an accelerometer was placed on the dorsum of the participant's dominant hand.

Participants were then asked to remain with the forearm pronated and relaxed on the table for 10 seconds. After this period, participants were instructed to hold the S Pen naturally at the center of the Archimedean spiral and to start drawing along the reference spiral when the countdown reached zero, using the application described previously. Each participant performed a continuous tracing for 10 seconds, without supporting the arm or forearm. After the 10-second tracing period, the participant returned to the resting position for another 10 seconds. This procedure was repeated seven consecutive times.

2) ***Treatment Stage*** *(Fig. 2B):* The treatment stage consisted of the application of TAPS-inspired electrical stimulation using a 2-kHz sinusoidal waveform, with the amplitude defined as $C{\times}ST$, where C corresponded to a constant adjusted for each nerve (median and radial) so that the stimulus was tolerable and did not cause discomfort. The value of C was determined incrementally, starting at $C = 1$ and increasing in steps of 0.25 until the participant's tolerance threshold was reached. Across participants, the applied stimulation current ranged from 0.80 to 3.09 mA for the median nerve and from 1.04 to 4.33 mA for the radial nerve.

The stimulation lasted approximately 6 min 30 s. During this period, the median and radial nerves were stimulated alternately using a standardized *burst* frequency (f_b) of 3 Hz, which was chosen based on the oscillatory characteristics of Parkinsonian rest tremor. Thus, each nerve was stimulated for half of the burst period, corresponding to $\frac{1}{2*f_b}$, or approximately 167 ms at 3 Hz. Ninety seconds after stimulation onset, participants completed seven Archimedean spiral tracings (Fig. 2B.1), following the same procedure described for the Pre-treatment stage. After the seven tracings were completed,

electrical stimulation continued until the full 6 min 30 s period had elapsed.

*3) **Post-treatment Stage** (Fig. 2C):* After the stimulation period ended, participants were asked to perform seven Archimedean spiral tracings again (Fig. 2C.1), following the same procedure described for the Pre-treatment stage.

C. Signal Processing

Data processing was performed using MATLAB software, version R2023b.

*1) **EEG Signal Processing**:* Initially, preprocessing was performed according to the following steps:

- The mean and linear trend were removed from each channel.
- A notch filter was applied to attenuate power-line interference (60 Hz).
- A second-order Butterworth band-pass filter (1–45 Hz) was applied in the forward and reverse directions to ensure zero-phase filtering.
- The signal was resampled to 400 Hz.
- The EEG signals corresponding to the motor assessments performed during the Pre-treatment, Treatment, and Post-treatment stages were segmented (Fig. 2A.2, 2B.1, and 2C.1). Seven segments, corresponding to repetitions of the motor task, were obtained per participant in each stage. Each segment had a total duration of 20 s, comprising 2 s of rest, used as the reference period, 3 s of motor preparation, 10 s of task execution, and 5 s of the post-movement period.
- For artifact identification, each segment was subdivided into 0.1 s windows. In each window, the samples were compared with a threshold of ± 3 standard deviations computed from a 20 s reference signal previously considered artifact-free. A window was classified as containing artifact when more than 5% of its consecutive samples or more than 10% of its total samples exceeded these limits.

Considering only the windows classified as artifact-free, event-related desynchronization/synchronization (ERD/ERS) was calculated according to the method described by Pfurtscheller & Lopes da Silva [2]. The initial 2 s of each segment were used as the baseline, and ERD/ERS values were expressed as the percentage change in power relative to the mean power of this baseline. For the comparison between the Pre-treatment and Treatment stages, a common reference was established, calculated from the mean baseline power obtained across both stages. This common reference was used to compute ERD/ERS under both conditions.

*2) **Accelerometry Signal Processing**:* Preprocessing was performed according to the following steps:

- The mean and linear trend were removed from each orthogonal axis.
- A fourth-order Butterworth band-pass filter (3–12 Hz) was applied in the forward and reverse directions to ensure zero-phase filtering.
- The 10 s segments corresponding to motor task execution during the Pre-treatment, Treatment, and Post-treatment stages were extracted (Fig. 2A.2, 2B.1, and 2C.1).
- Each 10 s segment was divided into five consecutive 2 s windows.

The amplitude of physiological tremor was estimated by considering a three-dimensional random vector whose components corresponded to the accelerometry signals from the three axes [19]. For this purpose, the power spectral density (PSD) was estimated for each axis using Welch's method, with 50% overlap between windows. Subsequently, the magnitude of this vector was obtained, and the physiological tremor frequency was estimated using a weighted average based on the spectral amplitudes, considering the frequency associated with the highest peak and its four nearest neighboring frequencies, according to the procedure described by Cevallos et al. [19].

*3) **Archimedean Spiral Trace Processing**:* For the analysis of the spiral trace, preprocessing was performed according to the following steps:

- Linearization of the spirals by converting the Cartesian coordinates (x, y) of both the trace performed by the participant and the reference spiral into polar coordinates (r, θ), considering the center of the spiral as the origin $(0, 0)$.
- Interpolation of the radius of the spiral traced by the participant to the θ values of the reference spiral, using cubic spline interpolation with $\Delta\theta = 0.04$ rad.

Subsequently, quantitative metrics were calculated to characterize motor control, including *smoothness*, *zero-crossing*, and radial error, as described by Pullman [14] and Surangsrirart & Thanawattano [15]. The radial error was converted from pixels to millimeters using a spatial factor of 0.167 mm/pixel, experimentally determined on the tablet used in the study. The metrics were calculated separately for each trial and subsequently averaged across the seven spirals performed by each participant in each stage of the protocol. Thus, a single mean value was obtained for each metric in the Pre-treatment, Treatment, and Post-treatment stages.

D. Statistical Analysis

*1) **Statistical Analysis of EEG Signals**:* To identify significant ERD/ERS periods in the EEG signal during motor task execution, the spectral F-test (SFT) was used following the adaptation described by Molina [20]. This test assesses whether spectral estimates obtained at the same frequency or within a frequency band of interest differ significantly [21]. A significance level of $\alpha = 0.05$ was adopted, with Bonferroni correction for multiple comparisons. To compare ERD/ERS magnitudes between the Pre-treatment and Treatment stages, the Wilcoxon signed-rank test was applied. The p-values were adjusted using Holm's method for multiple comparisons, and the corresponding effect sizes were estimated using the rank-biserial correlation.

*2) **Statistical Analysis of Accelerometry Signals and Archimedean Spiral Drawings**:* The analysis of motor performance parameters was conducted in RStudio (version

2025.5.0.496) using the R programming language (version 4.4.3). Descriptive statistics were calculated for each variable and stage of the experimental protocol. Data normality was then assessed using the Shapiro-Wilk test. Dominant frequencies, their corresponding PSD values, and the metrics extracted from the spiral drawings were compared among the Pre-treatment, Treatment, and Post-treatment stages using the Friedman test for repeated measures, with a significance level of 5%. The global effect size was estimated using Kendall's W coefficient. When applicable, pairwise comparisons between stages were performed using the Wilcoxon signed-rank test for related samples, with p-values adjusted using Holm's method for multiple comparisons.

III. Results

A. ERD/ERS Patterns During a Motor Task

Figure 3 presents the results of the grand-average ERD/ERS analysis in the alpha (8–13 Hz) and beta (13–30 Hz) bands during the Pre-treatment and Treatment stages, calculated using a common reference. To maintain visual clarity, the post-treatment stage is not shown, as its patterns were highly similar to those observed during pre-treatment stage.

In the alpha band (Fig. 3A), a similar pattern was observed between the Pre-treatment and Treatment stages. ERD was predominant in central, centroparietal, and parietal channels, with greater evidence over the hemisphere contralateral to the limb used in the motor task. This pattern began during movement preparation, between -3 and 0 s, and persisted throughout movement execution, between 0 and 10 s. In both stages, ERD periods significantly different from the common reference were identified ($p < 0.0001$).

After completion of the spiral drawing task (10 s), ERD values were generally maintained or became more pronounced across all analyzed channels, with this behavior being more evident in ipsilateral channels. From approximately 2 s ($t = 12$ s) after the end of the spiral drawing task, corresponding approximately to the time at which the participant returned to rest, a tendency toward power recovery was observed. In the comparison between the Pre-treatment and Treatment stages, no statistically significant differences were identified ($p > 0.19$). In addition, the estimated effect sizes were predominantly small ($r_{rb} < 0.3$).

In the beta band (Fig. 3B), ERD patterns showed similar behavior between the Pre-treatment and Treatment stages. ERD was observed in most channels from the preparation period and persisted throughout the spiral drawing task, being more evident in channels contralateral to the limb used. In general, this response was statistically significant ($p < 0.0001$) relative to the common reference.

Approximately 2 s ($t = 12$ s) after the end of the spiral drawing task, an increase in beta power toward baseline or positive values was observed, being more evident in the Pre-treatment stage. However, the comparison of ERD/ERS values between stages did not reveal statistically significant differences ($p > 0.07$). Nevertheless, large effect sizes were observed in the final seconds ($t = 14$–15 s), particularly in channels C3 ($r_{\mathrm{rb}} = 0.96$), C4 ($r_{\mathrm{rb}} = 0.96$), CP3 ($r_{\mathrm{rb}} = 0.81$), and CP4 ($r_{\mathrm{rb}} = 0.60$).

B. Motor Assessment

Table I presents the descriptive statistics and Friedman test results for the parameters extracted from the accelerometry signals and the metrics obtained from the Archimedean spiral drawing.

TABLE I
Descriptive statistics and comparison of motor performance parameters across the stages of the experimental protocol.

Accelerometry analysis					
Variable	**Stage**	**Mean (SD)**	**Median [Q1–Q3]**	***p***	**Kendall's** W
Frequency [Hz]	Pre	7.72 (1.30)	8.21 [6.80–8.52]	0.38	0.09
	Treat.	7.35 (1.48)	7.39 [6.34–8.38]		
	Post	7.38 (1.55)	7.25 [6.46–8.59]		
PSD $[(\mathrm{cm/s^2})^2/\mathrm{Hz}]$	Pre	32.3 (14.1)	31.6 [22.2–40.3]	0.15	0.17
	Treat.	31.8 (19.8)	24.9 [14.8–42.7]		
	Post	27.1 (27.0)	18.6 [10.6–31.7]		
Archimedean spiral drawing analysis					
Variable	**Stage**	**Media (SD)**	**Median [Q1–Q3]**	***p***	**Kendall's** W
1st Smoothness	Pre	10.45 (0.49)	10.53 [10.20–10.90]	0.53	0.06
	Treat.	10.52 (0.42)	10.50 [10.28–10.84]		
	Post	10.50 (0.48)	10.60 [10.15–10.85]		
2nd Smoothness	Pre	17.87 (0.46)	17.89 [17.69–18.19]	0.53	0.06
	Treat.	17.96 (0.41)	18.04 [17.78–18.19]		
	Post	17.90 (0.47)	17.93 [17.51–18.19]		
Radial error [mm]	Pre	1.46 (1.45)	0.70 [0.49–2.12]	0.70	0.03
	Treat.	0.91 (0.59)	0.61 [0.56–1.16]		
	Post	2.00 (1.65)	1.55 [0.58–3.36]		
1st Zero Crossing [%]	Pre	5.88 (1.62)	5.65 [5.26–6.63]	0.23	0.13
	Treat.	5.19 (1.74)	5.61 [4.40–6.27]		
	Post	5.01 (1.75)	5.19 [4.52–6.23]		
2nd Zero Crossing [%]	Pre	7.64 (1.99)	7.46 [6.77–8.70]	0.31	0.11
	Treat.	6.37 (2.05)	6.87 [5.22–7.71]		
	Post	6.47 (2.15)	6.75 [5.44–8.10]		

Note. Pre = Pre-treatment stage; Treat. = Treatment stage; Post = Post-treatment stage

Regarding the accelerometry parameters, in the Pre-treatment stage, the dominant frequency was 7.72 $\pm$ 1.30 Hz, with a PSD of 32.3 $\pm$ 14.1 $(\mathrm{cm/s^2})^2/\mathrm{Hz}$. During Treatment, these values decreased to 7.35 $\pm$ 1.48 Hz and 31.8 $\pm$ 19.8 $(\mathrm{cm/s^2})^2/\mathrm{Hz}$, respectively. In the Post-treatment stage, the dominant frequency was 7.38 $\pm$ 1.55 Hz, whereas the PSD showed a mean value of 27.1 $\pm$ 27.0 $(\mathrm{cm/s^2})^2/\mathrm{Hz}$. Overall, the dominant frequency remained similar across stages ($p = 0.38$), whereas the PSD showed a numerical reduction throughout the protocol, although this difference was not statistically significant ($p = 0.15$).

In relation to the metrics extracted from the Archimedean spiral trace, first and second order *smoothness* values remained similar across stages: Pre-treatment (10.45 $\pm$ 0.49 and 17.87 $\pm$

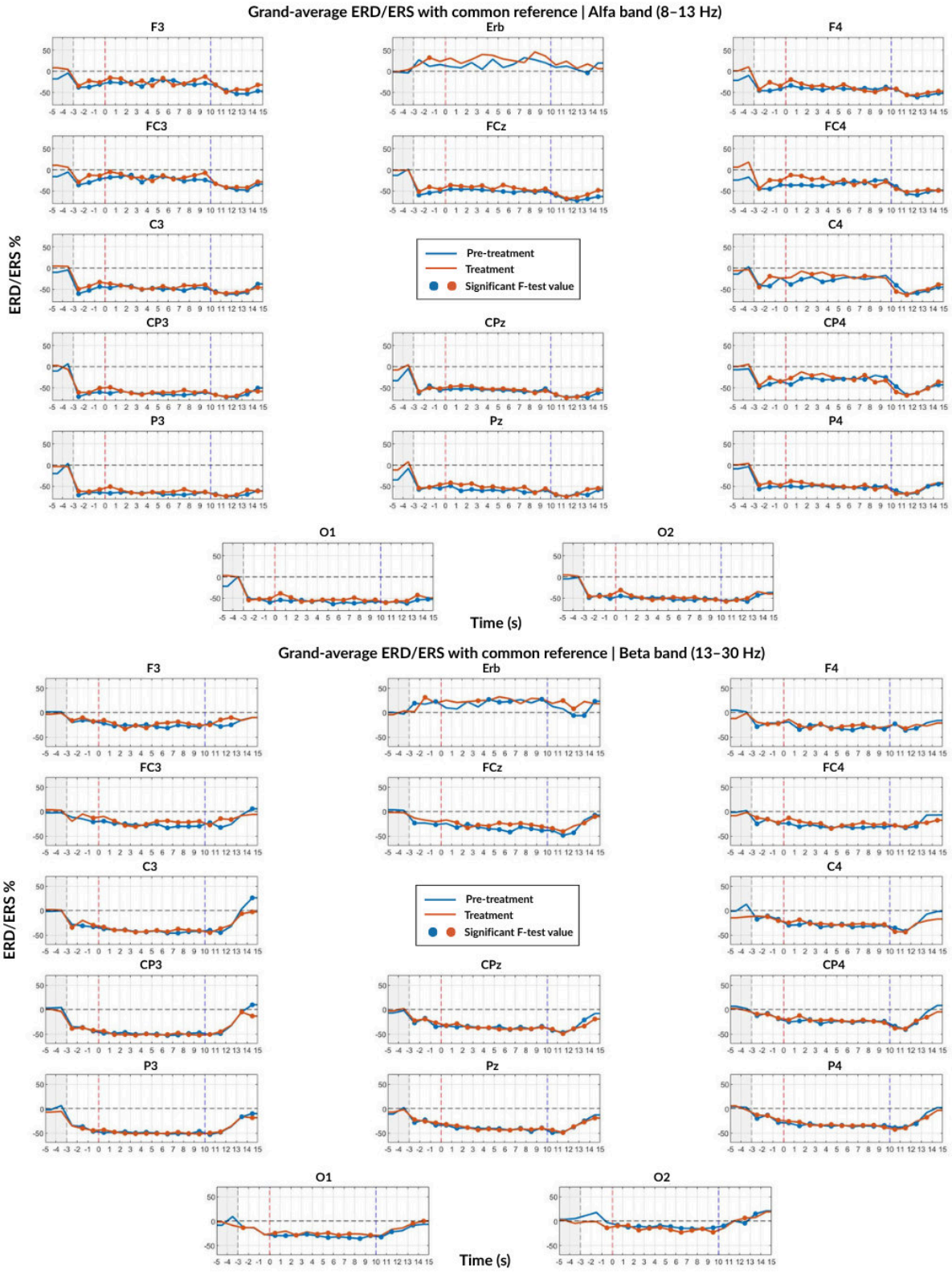


Fig. 3. Results of the ERD/ERS analysis during the Pre-treatment and Treatment stages. A) Alpha band (8–13 Hz); B) Beta band (13–30 Hz). The dots indicate time points at which ERD/ERS values were statistically different from the common reference, according to the spectral F-test.

0.46), Treatment (10.52 ± 0.42 and 17.96 ± 0.41), and Post- treatment (10.50 ± 0.48 and 17.90 ± 0.47), with no significant

differences observed for either metric ($p = 0.53$). Similarly, radial error and first and second order *zero-crossing* parameters showed no significant differences across stages ($p > 0.23$). Kendall's W coefficients indicated small global effect sizes for all analyzed variables ($W < 0.17$).

IV. Discussion and Conclusion

The results showed the presence of ERD/ERS patterns associated with movement preparation, execution, and the post-movement period in the alpha and beta bands, in agreement with responses widely described in the literature [2], [3]. In the Pre-treatment and Treatment stages, periods were identified in which ERD/ERS values differed significantly from the common reference. These findings demonstrate that the experimental protocol and the adopted processing strategy were able to detect changes in cortical oscillatory activity related to voluntary movement.

Overall, ERD was observed in the alpha and beta bands during movement preparation and execution, with greater predominance in channels contralateral to the limb used in the task. This observation is in line with Tang et al. [22], who identified a greater ERD magnitude in contralateral sensorimotor regions, especially during movements performed with the dominant hand. In the present study, the motor task was also performed with the dominant limb, which may have contributed to the observed lateralization.

After movement completion, alpha-band power remained reduced or showed a further decrease. Approximately 2 s after the end of the spiral drawing, a tendency toward recovery to baseline was observed, although this recovery was slow and incomplete. In the beta band, a progressive increase in energy was observed starting approximately 2 s after the end of the drawing, reaching its maximum values about 4 s after the end of the motor task. In some channels, this increase exceeded baseline, characterizing a response compatible with the post-movement beta rebound.

The findings observed after movement completion, in both the alpha and beta bands, differed from the patterns described in the literature, in which power recovery tends to occur shortly after task completion [2], [3]. This difference may be related to the characteristics of the task performed. Whereas classical ERD/ERS studies generally use simple and discrete movements, spiral drawing is a continuous visuomotor task that requires fine trajectory control, visual guidance, and successive movement corrections [15], [23]. These demands may have contributed to the sustained alpha desynchronization and to the delayed emergence of the post-movement beta response.

The characteristics of the experimental protocol may also have influenced this behavior. The 10 s time point corresponded to the end of the spiral drawing, but not necessarily to the actual cessation of all motor activity. After completing the spiral trace, the participant still had to place the pen on the table and return the upper limb to the resting position. Since the speed of return to rest varied among participants, the beta rebound showed a delay of approximately 1 s ($t \approx 14$ s) relative to the expected time point of $t \approx 13$ s, corresponding to 1 s after movement completion [2]. This delay may be due to the temporal misalignment of participants' cortical responses in the grand average.

In this context, the 5-s period analyzed after the end of the spiral drawing may have been insufficient to fully characterize the recovery of cortical activity, considering that the literature recommends intervals of approximately 10 s between movements [2]. The sustained alpha desynchronization and the delayed emergence of the beta rebound suggest that part of this response may have occurred beyond the analyzed interval. In future studies, it is recommended to record the effective time point at which all motor activity ends, including the return of the limb to the resting position, and to extend both the post-movement analysis period and the interval between the end of one spiral and the beginning of the next.

The comparison between the Pre-treatment and Treatment stages revealed similar ERD/ERS patterns in both the alpha and beta bands, with no statistically significant differences. The main distinction was observed in the beta band, between 4 and 5 s after the end of the motor task, when the ERS response showed greater magnitude in the Pre-treatment stage than in the Treatment stage, especially in central and centroparietal channels. Although this difference did not reach statistical significance, the estimated effect sizes were large, suggesting a possible modification of the post-movement beta response associated with sinusoidal peripheral electrical stimulation.

Consistent evidence on the effect of transcutaneous peripheral stimulation on beta rhythm, specifically during voluntary movement, was not found in the literature. However, in a study involving movements induced by functional electrical stimulation (FES) [24], ERD/ERS patterns similar to those described during voluntary movements were observed, characterized by alpha- and beta-band ERD during movement and beta ERS after movement completion. Nevertheless, because the protocol involved stimulation-induced movements rather than voluntary movements performed during peripheral stimulation, these findings do not allow conclusions to be drawn regarding whether electrical stimulation modifies ERD/ERS magnitude during a voluntary motor task.

However, these findings should be considered exploratory. The small sample size may have limited the statistical power to detect differences between stages. Moreover, the absence of a control condition limits the attribution of the observed changes to stimulation, since post-movement beta ERS has been shown to change during repeated motor practice [25]. Future studies should therefore include a larger sample and sham stimulation, to determine whether the observed pattern is maintained and specifically associated with stimulation. In addition, a prospectively defined baseline comparable across stages should be used to avoid retrospective reference estimation.

In the motor assessment, no statistically significant differences were observed in dominant frequencies, their corresponding PSD values, or the metrics extracted from the spiral drawings. Kendall's W coefficients also indicated small

global effects, reflecting a low magnitude of change across stages. This may be explained by the fact that the study was conducted with young participants with no history of neurological disorders or clinical conditions that could affect motor control. Therefore, their baseline performance may have already been within the expected range, reducing the possibility of detecting additional improvements.

Before applying the protocol to a larger clinical sample, a pilot study with patients with PD is recommended to evaluate the sensitivity of the selected metrics in detecting changes in motor control across the different treatment stages. This step may help identify more informative variables and adjust the acquisition, processing, and analysis parameters.

Overall, the results showed cortical ERD/ERS patterns associated with movement preparation, execution, and the post-movement period. Although no statistically significant differences were observed between the Pre-treatment and Treatment stages, large effect sizes were identified in the post-movement beta response, indicating a pattern that should be further investigated. These findings demonstrate the feasibility of the adopted methodology, while also highlighting the need for further optimization of the experimental protocol before its application in patients with PD, in order to increase the sensitivity of the analyses and improve the detection of potential effects of sinusoidal peripheral electrical stimulation during the Archimedean spiral drawing task.

ACKNOWLEDGMENTS

The authors acknowledge CAPES (PROEX), FINEP (Grant 01.24.0122.00), CNPq, FAPERJ (Grants E-26/204.393/2024 and SEI-260003/020273/2025), and GIIB (UPS) for the financial support provided.

REFERENCES

[1] C. Neuper, M. Wörtz, and G. Pfurtscheller, "ERD/ERS patterns reflecting sensorimotor activation and deactivation," in *Progress in Brain Research*. Elsevier, 2006, vol. 159, pp. 211–222.

[2] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[3] G. Pfurtscheller, "Functional brain imaging based on ERD/ERS," *Vision Research*, vol. 41, pp. 1257–1260, 2001.

[4] J. Peter, F. Ferraioli, D. Mathew, S. George, C. Chan, T. Alalade, S. A. Salcedo, S. Saed, E. Tatti, A. Quartarone, and M. F. Ghilardi, "Movement-related beta ERD and ERS abnormalities in neuropsychiatric disorders," *Frontiers in Neuroscience*, vol. 16, p. 1045715, 2022.

[5] L. Meng, M. Jin, X. Zhu, and D. Ming, "Peripherical electrical stimulation for parkinsonian tremor: A systematic review," *Frontiers in Aging Neuroscience*, vol. 14, p. 795454, 2022.

[6] A. Pascual-Valdunciel, G. W. Hoo, S. Avrillon, F. Oliveira Barroso, J. G. Goldman, J. C. Hernandez-Pavon, and J. L. Pons, "Peripheral electrical stimulation to reduce pathological tremor: A review," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, p. 33, 2021.

[7] S. Brillman, P. Khemani, S. H. Isaacson, R. Pahwa, R. Deshpande, V. Zraick, A. Rajagopal, D. Khosla, and K. H. Rosenbluth, "Non-invasive transcutaneous afferent patterned stimulation therapy offers action tremor relief in parkinson's disease," *Tremor and Other Hyperkinetic Movements*, vol. 13, no. 25, pp. 1–11, 2023.

[8] J. Y. Yu, A. Rajagopal, J. Syrkin-Nikolau, S. Shin, K. H. Rosenbluth, D. Khosla, E. K. Ross, and S. L. Delp, "Transcutaneous afferent patterned stimulation therapy reduces hand tremor for one hour in essential tremor patients," *Frontiers in Neuroscience*, vol. 14, p. 530300, 2020.

[9] R. Pahwa, R. Dhall, J. Ostrem, R. Gwinn, K. Lyons, S. Ro, C. Dietiker, N. Luthra, P. Chidester, S. Hamner, E. Ross, and S. Delp, "An acute randomized controlled trial of noninvasive peripheral nerve stimulation in essential tremor," *Neuromodulation: Technology at the Neural Interface*, vol. 22, no. 5, pp. 537–545, 2019.

[10] H. Resende-Martins, R. Zanetti, C. Cardoso-dos Santos, G. Mastrocola-Manzano, and C. J. Tierra-Criollo, "Current perception threshold and reaction time in the assessment of sensory peripheral nerve fibers through sinusoidal electrical stimulation at different frequencies," *Revista Brasileira de Engenharia Biomédica*, vol. 29, pp. 278–285, 2013.

[11] L. P. Volpi, "Efeitos dos estímulos elétricos senoidais na resposta induzida cerebral: estudo da seletividade das fibras sensitivas," Dissertação de Mestrado em Engenharia Biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2017.

[12] E. Imatz-Ojanguren and T. Keller, "Evoked sensations with transcutaneous electrical stimulation with different frequencies, waveforms, and electrode configurations," *Artificial Organs*, vol. 47, no. 1, pp. 117–128, 2022.

[13] J. J. Fernández-Pérez, J. Avendaño-Coy, D. Serrano-Muñoz, F. Oliveira Barroso, C. Montero-Pardo, B. López-Moreno, A. Lerín-Calvo, J. P. Romero Muñoz, and J. Gómez-Soriano, "Transcutaneous kilohertz high-frequency alternating current at 10 khz for upper-limb tremor in people with parkinson's disease: A double-blind, randomized, crossover study," *Journal of Clinical Medicine*, vol. 13, no. 24, 2024.

[14] S. L. Pullman, "Spiral analysis: A new technique for measuring tremor with a digitizing tablet," *Movement Disorders*, vol. 13, no. S3, pp. 85–89, 1998.

[15] D. Surangsrirat and C. Thanawattano, "Android application for spiral analysis in parkinson's disease," in *2012 Proceedings of IEEE SoutheastCon*, 2012, pp. 1–6.

[16] R. Magee, B. Yang, and J. Ratliff, "Trsper: A web-based application for archimedes spiral analysis," *mHealth*, vol. 8, no. 3, 2022.

[17] P. Cevallos-Larrea, L. Guambaña-Calle, D. A. Molina-Vidal, M. Castillo-Guerrero, A. d. Netto, and C. J. Tierra-Criollo, "Prototype of a multimodal and multichannel electro-physiological and general-purpose signal capture system: Evaluation in sleep-research-like scenario," *Sensors*, vol. 25, no. 9, 2025.

[18] D. A. Molina-Vidal, P. Cevallos Larrea, L. Guambaña Calle, D. Liquori, and C. J. Tierra-Criollo, "Design and evaluation of a prototype of dual channel electrical stimulator: Application in the assessment of current perception thresholds to multiple stimuli," in *Smart Technologies, Systems and Applications: SmartTech-IC 2022*, ser. Communications in Computer and Information Science, vol. 1705. Springer, 2023, pp. 342–355.

[19] P. S. S. Cevallos, D. M. da Silva-Liquori, and C. J. Tierra-Criollo, "Estimador da frequência de tremor em parkinsonianos com vetor aleatório de acelerometria," 2025.

[20] D. A. Molina Vidal, "Estudo dos ritmos cerebrais relacionados ao bloqueio da fibra fina por fibra grossa: Estímulo por corrente elétrica senoidal," Dissertação de Mestrado em Engenharia Biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2024.

[21] A. C. Almeida, M. A. Massote, R. M. Ichinose, C. J. Tierra-Criollo, and A. M. F. L. M. de Sá, "Spectral f test for detecting eeg event related synchronization/desynchronization caused by transcranial magnetic stimulation," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, ser. IFMBE Proceedings, vol. 57. Springer, 2016, pp. 25–28.

[22] Z. Tang, H. Yu, C. Lu, P. Liu, and X. Jin, "Single-trial classification of different movements on one arm based on ERD/ERS and corticomuscular coherence," *IEEE Access*, vol. 7, pp. 128 185–128 197, 2019.

[23] S. Raimo, G. Satangelo, and L. Trojano, "The neural bases of drawing. a meta-analysis and a systematic literature review of neurofunctional studies in healthy individuals," *Neuropsychology Review*, vol. 31, no. 4, pp. 689–702, 2021.

[24] G. Müller, C. Neuper, R. Rupp, C. Keinrath, H. Gerner, and G. Pfurtscheller, "Event-related beta eeg changes during wrist movements induced by functional electrical stimulation of forearm muscles in man," *Neuroscience Letters*, vol. 340, no. 2, pp. 143–147, 2003.

[25] A. B. Nelson, C. Moisello, J. Lin, P. Panday, S. Ricci, A. Canessa, A. DiRocco, A. Quartarone, G. Frazzitta, I. U. Isaias, G. Tononi, C. Cirelli, and M. F. Ghilardi, "Beta oscillatory changes and retention of motor skills during practice in healthy subjects and in patients with parkinson's disease," *Frontiers in Human Neuroscience*, vol. 11, p. 104, 2017.

Effects of Peripheral Electrical Stimulation Waveforms on Tremor in Parkinson's Disease: A Pilot Study

Pedro Sebastian Silverio
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brasil
0009-0002-5712-4699

Evelyn Juliette Pesántez-Vera
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brasil
0009-0009-0019-6713

Danielle Martins da Silva Liquori
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brasil
0009-0002-1804-3490

Danilo Andrés Molina-Vidal
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brasil
0000-0003-4420-4585

Carlos Julio Tierra-Criollo
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brasil
0000-0003-4880-9186

Abstract*—This pilot study evaluated an individualized peripheral electrical stimulation protocol for tremor modulation applied to the median and radial nerves and examined two stimulation waveforms: biphasic pulse bursts (BPB) and sinusoidal bursts (SB). A burst frequency was individualized according to each participant's dominant tremor frequency. Ten participants with Parkinson's disease were randomly assigned to two groups of five ($n = 5$). Hand acceleration was recorded during a postural task at Pre-stimulation, Intra-stimulation (measured at 10, 20, and 30 minutes; denoted as Intra 1, Intra 2, and Intra 3), and Post-stimulation (measured after 40 minutes of stimulation). Tremor magnitude was quantified from the power spectral density (PSD) as the total power within the parkinsonian tremor band ($3.5 - 12$ Hz). Tremor-band power was evaluated using a within-session temporal analysis independently for each waveform. The BPB group showed no statistically significant temporal variation or large effect size ($p = 0.606$, Kendall's $W = 0.136$). Conversely, the SB group demonstrated statistically significant differences across assessment periods with a strong effect size ($p = 0.013$, Kendall's $W = 0.632$). Post-hoc pairwise comparisons (Holm-Bonferroni correction) revealed that significant differences relative to baseline (Pre) were concentrated during the Intra 2 and Intra 3 periods ($p \leq 0.008$). To quantify the magnitude and direction of power differences, relative changes (RC %) were calculated with respect to baseline (Pre). These metrics indicated median power reductions during stimulation (-46.9% for BPB; -47.1% for SB) and post-stimulation (-17.25% for BPB; -29.66% for SB). Furthermore, exploratory correlation analyses between RC % and participant characteristics, including dominant tremor frequency and applied current intensity parameters, yielded null to weak correlations, indicating no clear association with individual power variations. The small sample size limits the generalizability of these preliminary findings, and the results do not establish clinical efficacy, optimal stimulation parameters. Therefore, preliminary results encourage further investigation into the use of sinusoidal waveforms for tremor reduction.

***Index Terms*—Parkinson's disease, tremor, peripheral nerve stimulation, accelerometry, power spectral density**

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder and the second most common neurodegenerative disease worldwide. Its clinical presentation includes both motor and non-motor symptoms [1]. Bradykinesia, rigidity, tremor, and postural instability are among its main motor manifestations. The progression and severity of these symptoms vary considerably among individuals and may impair upper-limb control, mobility, and independence in activities of daily living [2].

Tremor in PD is characterized by involuntary, rhythmic, and oscillatory movements and may occur at rest or during voluntary action. Classical parkinsonian rest tremor predominantly occurs at approximately $4 - 6$ Hz when the affected body segment is relaxed and fully supported [3]. Action tremors include postural tremor, which occurs while maintaining a position against gravity, and kinetic tremor, which is present during voluntary movement. These manifestations may coexist within the same individual and exhibit partially overlapping frequency ranges, reflecting the heterogeneity of parkinsonian tremor [4], [5].

Levodopa is one of the pharmacological treatments for the motor symptoms of PD and is considered effective for controlling troublesome tremor in most patients [1]. However, tremor response to levodopa can vary among individuals, and the pharmacological response of different tremor components is not fully established [6]. Re-emergent tremor appears to be closely related to rest tremor, including in its response to levodopa, whereas the response of other postural or kinetic tremor components may be more variable [6], [7]. Consequently, some individuals may continue to present clinically relevant tremor despite pharmacological treatment.

For selected patients, deep brain stimulation is an established neurosurgical treatment, particularly for medication-refractory tremor and motor complications that cannot be adequately managed pharmacologically [8], [9]. Nevertheless, its invasive nature, potential surgical risks, and clinical eligibility requirements limit its applicability. Non-invasive mechanical approaches, including wearable damping devices and tremor-suppression orthoses, have also been investigated. Although these systems can attenuate tremor, their clinical and home use may be limited by their weight, bulk, wearability, and potential interference with voluntary movement [10], [11].

In this context, peripheral electrical stimulation has been investigated as a non-invasive neuromodulation approach for tremor reduction. One technique is Transcutaneous Afferent Patterned Stimulation (TAPS), which applies alternating bursts of electrical stimulation to the median and radial nerves at the wrist [12], [13]. The timing of the alternation between the nerves is individualized according to the participant's dominant tremor frequency. It has been hypothesized that the resulting peripheral afferent activity modulates central structures within tremor-related networks, including the cerebello-thalamo-cortical pathway, although the mechanisms underlying its clinical effects have not been fully established [14], [15].

Published TAPS protocols have primarily employed bursts of charge-balanced biphasic pulses, and most of the available clinical evidence has been obtained in individuals with essential tremor [13], [16]. More recently, reductions in action tremor have also been reported in people with PD; however, the evidence in this population remains comparatively limited [12]. Furthermore, the effects of alternative stimulation waveforms have received little investigation.

As an alternative to biphasic pulses, high-frequency sinusoidal stimulation has been investigated because its sensory effects may vary according to the applied frequency. Previous studies have associated sinusoidal currents between 2 and 3 kHz with the preferential recruitment of large myelinated Aβ fibers, particularly when stimulation is applied near the sensory threshold [17], [18]. This provides a physiological basis for exploring a 3 kHz sinusoidal waveform as an alternative form of peripheral sensory stimulation.

Triaxial accelerometry provides an objective method for characterizing tremor and can be used to estimate its dominant frequency and quantify acceleration amplitude and spectral power [19]. Within TAPS-based protocols, accelerometry can support the individualization of stimulation timing by estimating the participant's tremor frequency [12]. These quantitative measurements complement clinical instruments such as the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Although the MDS-UPDRS is a widely used and validated clinical instrument, its motor items are scored using ordinal categories assigned through clinical observation [20]. Accelerometry provides continuous measurements that may help identify subtle or short-term changes that are not fully represented by an ordinal clinical score.

Accordingly, this pilot study aimed to characterize the preliminary effects of two individualized peripheral electrical stimulation conditions on tremor in individuals with PD: Biphasic Pulse Burst (BPB), composed of charge-balanced biphasic pulses, and Sinusoidal Burst (SB), composed of a 3 kHz sinusoidal waveform. Given the exploratory design and small sample size, the study aimed to identify preliminary within-group response patterns rather than establish clinical efficacy or determine the superiority of one waveform over the other.

II. Materials and Methods

A. Participants and Ethical Considerations

This pilot study was conducted at the Clementino Fraga Filho University Hospital of the Federal University of Rio de Janeiro (HUCFF/UFRJ). The study protocol was approved by the Research Ethics Committee under approval number CAAE 58744222.2.1001.5257. Before participating, all individuals were informed about the objectives and procedures of the study and provided written informed consent.

The final sample comprised ten men with idiopathic Parkinson's disease. Participants were between 58 and 75 years of age, with a mean age of 66 ± 5.8 years. The diagnosis of Parkinson's disease had previously been established by a neurologist specializing in movement disorders.

The predefined exclusion criteria included dementia associated with Parkinson's disease, other neurodegenerative disorders, diabetes mellitus, and orthopedic conditions affecting the upper limbs that could interfere with performance of the experimental tasks. For safety reasons, individuals with a cardiac pacemaker, a deep brain stimulation system, or another implanted electronic medical device were also excluded. In addition, participants using anticoagulants, antidepressants, sedatives, or other substances likely to interfere with attention, alertness, or motor performance during the experiments were excluded.

Participants were randomly assigned to either the Biphasic Pulse Burst (BPB) or Sinusoidal Burst (SB) waveform group using a draw-lots procedure, with a predefined allocation of five participants per group. Once one group reached this limit, the remaining participants were assigned to the other.

B. Experimental Protocol

For the experimental protocol, one participant was evaluated in the medication OFF state, 13 h 42 min after the last dose, whereas the remaining participants were evaluated in the ON state, 2 h 02 min [Q1–Q3: 1 h 40 min–3 h 50 min] after the last dose. The protocol was conducted in a temperature-controlled room ($24 \pm 1\,^{\circ}\mathrm{C}$). Participants were seated comfortably in a chair with their forearms supported on a table positioned in front of them. The upper limb exhibiting the most pronounced tremor was identified and selected for stimulation and acceleration recording.

Two bipolar electrode pairs were positioned to target the median and radial nerves using a methodology adapted from

previous peripheral stimulation protocols [14], [16]. Disposable circular adhesive electrodes containing conductive gel and measuring 1 cm in diameter were used. Before electrode placement, the skin over the wrist was cleaned with 70% alcohol.

For radial nerve stimulation, the cathode was positioned over the dorsolateral aspect of the wrist, near the anatomical snuffbox, and the anode was placed $2\ \mathrm{cm}$ proximally. For median nerve stimulation, the cathode was positioned over the palmar aspect of the wrist, and the corresponding anode was placed $2\ \mathrm{cm}$ proximally. The separation between the median and radial stimulation paths was adjusted according to wrist circumference, adapting the wrist-size categories described by [13]. A separation of $2.5\ \mathrm{cm}$ was used for wrist circumferences of 13.5–$15.4\ \mathrm{cm}$, $3.0\ \mathrm{cm}$ for circumferences of 15.5–$17.4\ \mathrm{cm}$, and $3.5\ \mathrm{cm}$ for circumferences of 17.5–$19.5\ \mathrm{cm}$.

The MESTIM_ELECT system [21], used for stimulus generation, comprises two independent current sources capable of generating arbitrary stimulation waveforms. The system is controlled through a graphical user interface (GUI) developed in LabVIEW, which enables the configuration and adjustment of the electrical stimulation parameters. The maximum output current was limited to 10 mA per channel.

An accelerometer was subsequently attached to the dorsum of the selected hand using a wrist strap. Using this device, adapted from [22], triaxial acceleration signals were acquired at a sampling frequency of $50\ \mathrm{Hz}$ and used for real-time tremor-frequency estimation during the experimental session and for subsequent offline analysis.

To evaluate the effects of stimulation, the protocol was organized into three assessment periods: Pre-stimulation, Intra-stimulation, and Post-stimulation.

The Pre-stimulation assessment served as the participant's baseline measurement and consisted of the initial recording of a complete six-task sequence. These tasks were adapted from selected items of Part III of the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [20]: item 3.4 - finger tapping between the index finger and thumb (Fig. 1.A), item 3.5 - hand opening and closing (Fig. 1.B), item 3.6 - pronation–supination movements of the hand (Fig. 1.C), item 3.16 - the finger-to-nose maneuver as a kinetic tremor task (Fig. 1.D), item 3.15 - outstretched-arm postural task (Fig. 1.E), and item 3.17 - seated rest for rest tremor task (Fig. 1.F). Each task was recorded for 20 s, with 10 s rest intervals between consecutive tasks.

Acceleration signals were recorded continuously throughout the six-task sequence.The participant's individual tremor frequency was estimated by using the outstretched-arm postural task (Fig. 1.E). For this real-time estimation, the acceleration signals were filtered between 3 and 12 Hz. Then, the power spectral density (PSD) was estimated using a Welch periodogram and a random three-dimensional vector derived from the 3-axis accelerometer, as proposed by [22]. The most prominent spectral peak was identified as the dominant tremor frequency (f_{tremor}). The resulting value was subsequently used to configure the burst frequency of the stimulation protocol,

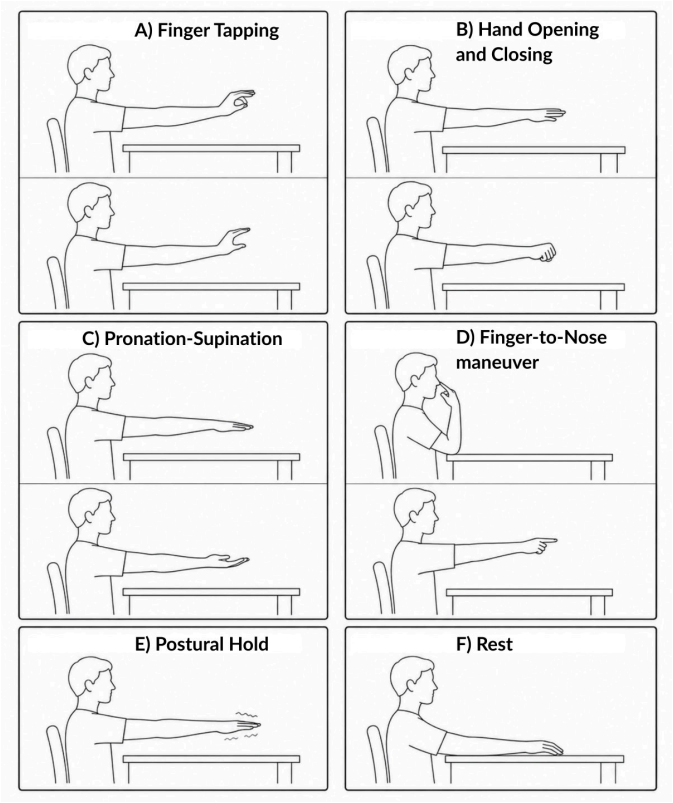


Fig. 1. Experimental motor tasks adapted from selected items of Part III of the MDS-UPDRS [20]: A) finger tapping; B) hand opening and closing; C) pronation-supination movements of the hand; D) finger-to-nose maneuver; E) outstretched-arm postural task; and F) seated rest for rest tremor task.

such that $f_{\text{burst}} = f_{\text{tremor}}$. The burst period was defined as $T_{\text{burst}} = \frac{1}{f_{\text{burst}}}$. Each channel remained active for half of this period as $T_{\text{on}} = \frac{T_{\text{burst}}}{2}$, alternating stimulation between the medial and radial nerves.

Treatment current intensities (TCI) for the median and radial nerves were determined independently by applying a multiplication factor (MF) to the sensory threshold (ST). This factor was progressively increased to the highest level comfortably tolerated by the participant, ensuring no visible muscle contraction or involuntary movement occurred. The ST was obtained following the procedure described by [17], [18].

During Intra-stimulation period, electrical stimulation was delivered for 40 min. Participants exclusively executed the outstretched-arm postural task for approximately 20 s (Fig. 1.E) at the 10-, 20-, and 30-min marks following the onset of stimulation to assess tremor. These three assessments were designated Intra 1, Intra 2, and Intra 3, respectively.

Following the stimulation period, the six-task sequence performed during the Pre-stimulation baseline was repeated. This final recording was defined as the Post-stimulation assessment.

The waveform delivered depended on the stimulation condition assigned to the participant. In the BPB waveform, each burst consisted of a train of symmetric, charge-balanced biphasic pulses. Each pulse comprised a $300\ \mu\mathrm{s}$ positive phase,

followed by a 50 μs interphase interval and a 300 μs negative phase. The positive and negative phases had equal current amplitudes, and the pulses were repeated within each burst at an internal pulse frequency of 150 Hz [12]–[14]. Under the SB waveform, each burst consisted of a continuous 3 kHz sine wave.

C. Signal Processing and Feature Extraction

Offline processing and analysis of the acceleration signals were performed in MATLAB (MathWorks, Inc., Natick, MA, USA). During preprocessing, a fourth-order Butterworth high-pass filter with a cutoff frequency of 0.5 Hz was applied separately to the acceleration signals from the X, Y, and Z axes. The filtering step was used to attenuate the static gravitational component and other slowly signal components.

Although the experimental protocol included all six motor tasks described previously, the present analysis focused on the outstretched-arm postural task. This posture was selected to evaluate tremor power across all assessment periods: Pre-stimulation, Intra 1, Intra 2, Intra 3, and Post-stimulation.

The segments of interest were identified using the temporal markers inserted during the experimental protocol, and a continuous 20 s interval was extracted from each signal.

Spectral features were subsequently extracted from each preprocessed segment, as described below. The resulting dataset was organized according to participant identifier, stimulation waveform (BPB or SB), and assessment period. Spectral features were subsequently extracted from each preprocessed segment, as described below.

1) Power Spectral Density: Following the methodology in [22], the Power Spectral Density (PSD) for the X, Y, and Z axes was estimated using Welch's method (4 s Hann window, 50% overlap) and combined into a triaxial PSD by summing the orthogonal components (omitting the square-root operation to retain power density values). From this spectrum, tremor-band power was calculated by integrating across a predefined frequency band of $3.5-12$ Hz [23], [24], which encompasses the frequency components associated with parkinsonian tremor.

2) Relative Change: To quantify variations in tremor-band power across assessment periods relative to the pre-stimulation baseline, the Relative Change (RC, %) was calculated as:

$$\mathrm{RC}_i = \frac{\mathrm{Power}_i - \mathrm{Power}_{\mathrm{Pre}}}{\mathrm{Power}_{\mathrm{Pre}}} \times 100\% \tag{1}$$

where Power_i represents the power during a given assessment period (Intra 1, Intra 2, Intra 3, or Post-stimulation) and $\mathrm{Power}_{\mathrm{Pre}}$ represents the power during the pre-stimulation period. Here, negative values indicate a reduction in tremor-band power, positive values denote an increase, and zero represents no change.

D. Statistical Analysis

Statistical analyses were performed using R software. Data were analyzed independently for each group defined by the stimulation waveform (BPB and SB). Due to the small sample size ($n = 5$ per group) and to avoid assuming a normal distribution, a non-parametric approach was adopted. Descriptive statistics were reported as median and 1st and 3rd quartiles [Q1, Q3] for treatment current intensity (TCI), multiplier factor (MF), and the relative change percentage (RC) of tremor band power (3.5–12 Hz). To evaluate the temporal changes in tremor band power across the five assessment periods (Pre, Intra 1, Intra 2, Intra 3, and Post), a repeated-measures Friedman test was performed separately for each stimulation waveform, with the significance level set at $\alpha = 0.05$. When the Friedman test indicated a statistically significant overall difference across the assessment periods ($p < 0.05$), post-hoc pairwise comparisons were performed using the Conover's test with a Holm-Bonferroni correction, with Pre-stimulation defined as the reference period. Moreover, the magnitude of the overall effect was quantified using Kendall's coefficient of concordance (W), with 95% confidence interval (CI) estimated via bootstrap resampling (10,000 iterations). Finally, an exploratory analysis of individual responses was conducted across both groups to evaluate the relationship between participant parameters (f_{tremor}, TCI and MF) and tremor band power changes. For this purpose, Spearman's rank correlation coefficient (ρ) was computed between the relative change percentage (RC) during (Intra-mean, calculated as the average across Intra 1–3), and after stimulation (Post).

III. Results

The multiplication factor and treatment current intensity for each experimental group are presented as the median (first–third quartiles). In the BPB group, the multiplication factor was 1.3 (1.3–1.5) for median nerve stimulation and 1.5 (1.3–1.5) for radial nerve stimulation. These multiplication factors corresponded to treatment current intensities of 2.54 mA ($2.22-2.74$) for the median nerve and 1.87 mA ($1.79-3.85$) for the radial nerve. In the SB group, the multiplication factor was 1.2 (1.1–1.2) for the median nerve and 1.2 (1.2–1.4) for the radial nerve. The corresponding treatment current intensities were 3.83 mA ($3.63-5.35$) for the median nerve and 3.95 mA ($3.94-4.70$) for the radial nerve.

Figure 2 illustrates the postural tremor power across the assessment periods. To facilitate the visual representation of tremor power values, which span several orders of magnitude among participants, a logarithmic scale (base 10) was applied to the power values expressed in $(\mathrm{cm/s^2})^2$. In the BPB group (participants P01–P05), participant P01 generally registered the lowest power values within the group, ranging from 53.81 to 166.76 $(\mathrm{cm/s^2})^2$, whereas participant P02 exhibited substantially higher magnitudes, oscillating between 90,960.81 and 309,686.07 $(\mathrm{cm/s^2})^2$. When evaluating changes during the stimulation period (Intra 1 to Intra 3) relative to baseline (Pre), participants P03 and P04 demonstrated a reduction in power values, with P04 exhibiting the most pronounced and sustained attenuation. In contrast, participants P01, P02, and P05 displayed a fluctuating response characterized by alternating periods of power increase and reduction. During the Post-stimulation assessment period, most participants

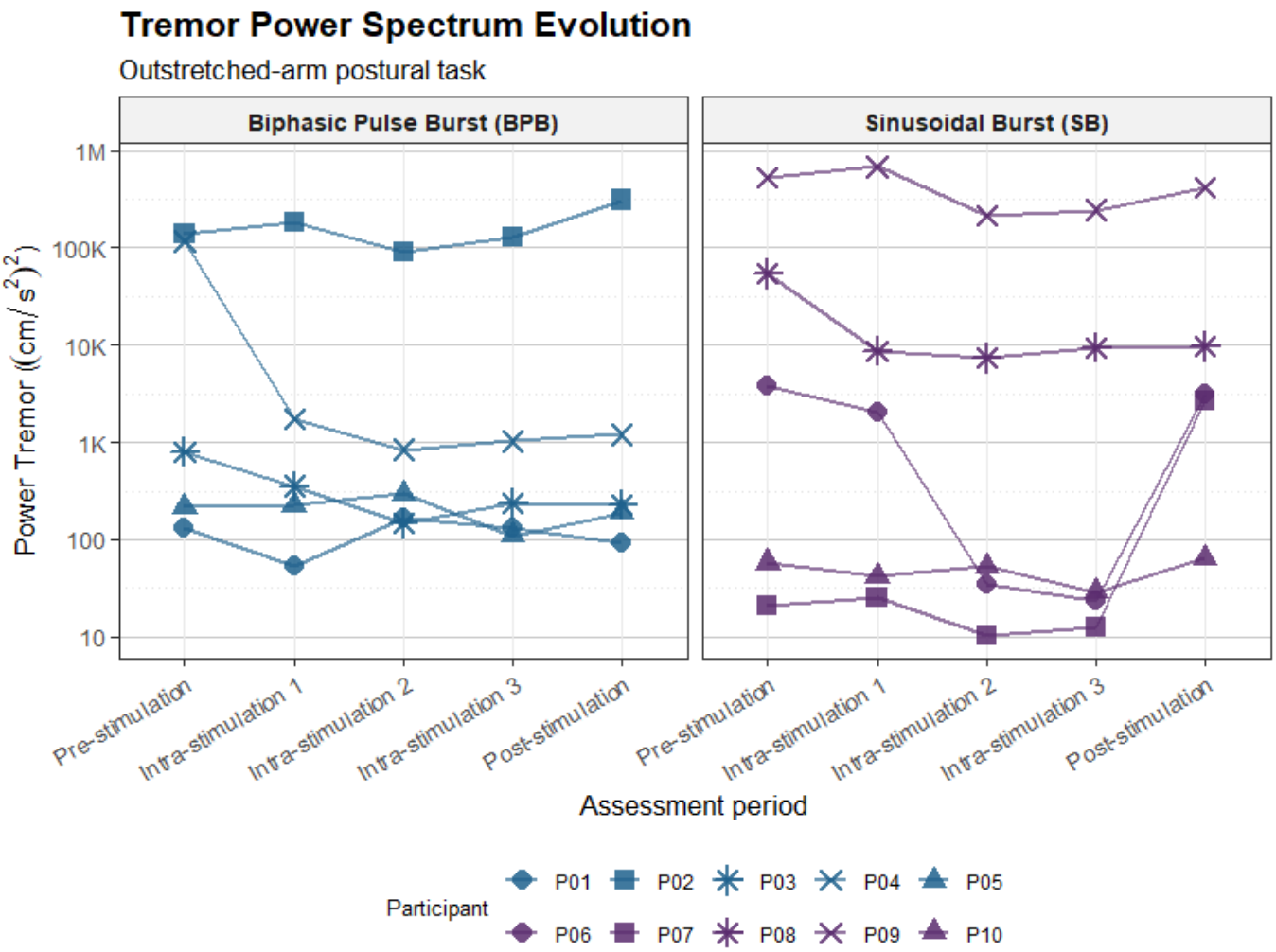


Fig. 2. Postural tremor power in the 3.5–12 Hz frequency band across evaluation periods, represented on a logarithmic scale (base 10). The left panel shows the BPB group (blue), and the right panel shows the SB group (purple). Individual markers represent participants (P01–P10).

maintained values below their baseline, with the exception of participant P02. Evaluating temporal progression using the Friedman test revealed no significant main effect for this group ($\chi^2_F = 2.72$, $p = 0.606$). The effect size for this group was small ($W = 0.136$) with a wide 95% confidence interval ([0.06, 0.81]), reflecting substantial heterogeneity in the participants' trajectories.

In the SB group (participants P06–P010), the baseline (Pre) power range was wider than in the BPB group. Participant P09 maintained the highest magnitudes in the group, ranging from 212,764 to 682,920.45 $(\text{cm/s}^2)^2$ across all assessment periods. Meanwhile, participant P07 registered the lowest values during stimulation, between 10.23 y 25.04 $(\text{cm/s}^2)^2$, though exhibiting a sharp increase during the Post-stimulation assessment period to reach 2,652.34 $(\text{cm/s}^2)^2$. Relative to baseline (Pre), the response during stimulation also showed power fluctuations: P06 achieved the highest relative attenuation in the group, whereas P08 displayed a relatively stable power reduction. During the Post-stimulation assessment period, participants P06, P08, and P09 maintained attenuated values, in contrast to P07 and P10, who experienced an increase (moderate for P10 and pronounced for P07). Unlike the BPB group, the temporal evaluation of the SB group revealed a statistically significant difference in tremor band power across assessment periods ($\chi^2_F = 12.64$, $p = 0.013$). Post-hoc comparisons using Conover's test with Holm-Bonferroni correction demonstrated significant differences between Intra 2 and Intra 3 periods compared to baseline ($T = -3.83$, $p = 0.006$ and $T = -3.54$, $p = 0.008$, respectively). In contrast, the Intra 1 and Post assessment periods did not reach statistical significance compared to Pre ($T = -1.47$, $p = 0.319$ and $T = 0$, $p = 1.000$, respectively). The effect size in this waveform was large ($W = 0.632$) with a 95% CI of [0.58, 0.86].

To complement the comparisons against the baseline (Pre) with RC, Table I summarizes the median and 1st and 3rd quartiles [Q1, Q3] in tremor band power relative to baseline (Pre). Overall, during stimulation (mean of Intra periods), both groups yielded an equivalent general power reduction value

TABLE I
SUMMARY OF RELATIVE CHANGE PERCENTAGE (%) IN TREMOR-BAND POWER COMPARED TO BASELINE (PRE)

	BPB Group		SB Group	
Assessment	**Median (%)**	**IQR [Q1; Q3]**	**Median (%)**	**IQR [Q1; Q3]**
Intra 1	-55.82	$[-58.76; 3.83]$	-27.45	$[-46.51; 21.45]$
Intra 2	-35.67	$[-81.75; 27.83]$	-59.85	$[-86.44; -50.38]$
Intra 3	-49.90	$[-70.65; -8.41]$	-53.63	$[-82.64; -49.80]$
Post	-29.66	$[-71.58; -13.34]$	-17.25	$[-20.92; 10.86]$

(BPB: -47.13%; SB: -46.97%). However, in the BPB group, Intra 1 registered the greatest reduction (median of -55.82%), whereas subsequent periods exhibited lower reduction values. A minor dispersion was observed in Intra 3, its interquartile interval [Q1, Q3] indicate only reduction. On the other hand, in the SB group, the median reduction became more pronounced from Intra 1 to Intra 2, reaching its largest value at Intra 2 (59.85%). The interquartile interval [Q1, Q3] at both Intra 2 and Intra 3 were entirely below zero. Following the end of stimulation (Post), smaller median reductions were observed in both groups compared with their intra-stimulation values. The BPB group showed a median RC of 29.66%, with an interquartile range entirely below zero, and the SB group showed a median RC of 17.25%, with a wider interquartile range spanning both negative and positive values.

Finally, exploratory analyses revealed no clear monotonic relationships between RC and the participant parameters (f_{tremor}, TCI, and MF). During stimulation (Intra-mean), Spearman correlation coefficients demonstrated no monotonic association for f_{tremor} ($\rho = 0$) and weak negative associations for stimulus amplitude parameters, TCI ($\rho = -0.224$) and MF ($\rho = -0.353$). Post-stimulation period revealed a similar pattern with weak associations (f_{tremor}: $\rho = -0.225$; TCI: $\rho = 0.224$; MF: $\rho = -0.322$).

IV. DISCUSSIONS AND CONCLUSIONS

This pilot study evaluated the preliminary effects of two peripheral electrical stimulation waveforms–Biphasic Pulse Burst (BPB) and Sinusoidal Burst (SB) on tremor-band power during an extended-arm postural task in patients with Parkinson's disease.

The treatment current intensities applied in this study showed distinct variability patterns depending on the waveform and the stimulated nerve, being generally higher in the sinusoidal burst (SB) group than in the biphasic pulse burst (BPB) group. In the BPB group, the median current intensity was $2.54\ mA$ for the median nerve and $1.87\ mA$ for the radial nerve, whereas Pahwa et al. [13] reported a higher mean current (5.4 ± 2.9 mA). In the study by Pahwa et al. [13], patients with essential tremor were evaluated, and a single overall current value was reported per participant, adjusted directly based on comfort sensation. In contrast, the present study focused on Parkinson's disease and characterized the current intensity independently for each nerve, considering each subject's sensory threshold and adjusting it to achieve comfortable paresthesia. Although both protocols shared a conceptually similar stimulation waveform and aimed to remain strictly below the motor threshold, the discrepancies in the recorded current intensity values could be attributed to these methodological differences.

In the SB group, the median current intensities were $3.83\ mA$ for the median nerve and $3.95\ mA$ for the radial nerve, which were substantially lower than the initial mean current of $82.3\ mA$ (with a final mean of $105\ mA$) reported by Fernández-Pérez et al. [25] in patients with Parkinson's disease. These differences can be explained because their protocol applied a continuous 10 kHz sinusoidal current, increasing the intensity via a 30-second ramp-up with continuous readjustments throughout the session to achieve a comfortable tingling sensation just below the motor threshold, using electrodes positioned near the medial (targeting median and ulnar nerves) and lateral (targeting radial nerve) epicondyles. In contrast, the present study employed 3 kHz sinusoidal bursts according to each patient's tremor frequency, set relative to the individual sensory threshold, and delivered via surface bipolar electrodes placed over the target nerves.

For both waveforms, tremor-band power showed median reductions relative to baseline (RC) during electrical stimulation (Intra 2 and Intra 3 periods) and after stimulation (Post). In the BPB group, these reductions were not significant ($p = 0.606$). In contrast, the SB group demonstrated statistically significant temporal variations across assessment periods ($p = 0.013$), specifically during the stimulation periods, Intra 2 and Intra 3 ($p = 0.006$ and $p = 0.008$, respectively). The combined intra-stimulation (Intra-mean) exhibited a median reduction of 47.1% and 46.9% for BPB and SB groups, respectively. These findings suggest that that reduction, represented by tremor-band power, achieved significance in the SB group during sustained stimulation. In this context, [14] monitored tremor power using accelerometers before, during, immediately after, and at 30 and 60 min post-stimulation with transcutaneous afferent patterned stimulation (TAPS) during postural hold, spiral drawing and pouring tasks. Their findings indicate that the postural hold task has the most pronounced attenuation, with median reductions of approximately 70% immediately post-stimulation and 83% at 30 min. However, direct quantitative comparison with intra-stimulation periods remains limited due to key methodological differences. Specifically, their study was conducted in patients with essential tremor, quantified tremor power within a narrower frequency band around the dominant frequency without a combined triaxial PSD, and reported detailed numerical reductions primarily for post-stimulation periods. Consequently, while their work reinforces the utility of longitudinal accelerometric monitoring, it represents a distinct clinical and analytical framework.

At the end of stimulation (Post) median power reductions were 29.66% and 17.25% for BPB and SB groups, respectively. In the BPB group, these post-stimulation results can be cautiously contextualized with previous evidence evaluating TAPS in Parkinson's disease. For instance, [12] reported a substantially higher median tremor power reduction of 64% at the end of stimulation, with 79% of participants exhibiting

reductions of at least 50% and 97% showing some degree of improvement. However, their study differed substantially from the present work, since it included a larger sample, two sessions per day over 4 weeks, and pre–post measurements for each session per participant. Consequently, these substantial differences in protocol duration and clinical setting preclude direct quantitative comparisons with previous home-use TAPS studies. On the other hand, for the SB group, evidence concerning the post-stimulation effects of this waveform in Parkinson's disease remains limited. In this context, [25] evaluated a 10 kHz sinusoidal current stimulation in patients with Parkinson's disease, reporting benefits primarily for kinetic tremor. Regarding postural or resting tremor, they observed a reduction no greater than to the sham treatment. This outcome provides a rationale for exploring sinusoidal waveform configurations targeting tremor reduction, such as the SB signal implemented in this work. Thus, while the responses in the SB group must be interpreted as exploratory, the results warrant further investigation of this waveform as a potential alternative or complementary therapeutic approach for managing postural tremor. The exploratory analysis revealed no clear monotonic relationships between RC and participant parameters (f_{tremor}, TCI or MF). Similarly, [26] reported no significant association between dominant tremor frequency and tremor suppression, although their stimulation protocol and study population differed substantially from those used here. Regarding stimulation intensity, [27] reported greater tremor-power reductions at higher stimulation amplitudes under phase-locked peripheral nerve stimulation. In contrast, the weak association observed in the present study suggests that stimulation intensity alone did not clearly account for the interindividual variability in tremor power reduction. Several limitations of this study must be acknowledged. High inter-individual variability was observed in baseline tremor characteristics, which may have been influenced by factors such as disease severity, medication status, and individual responsiveness to peripheral stimulation. Furthermore, the experimental design did not include a sham stimulation condition or blinding. Additionally, direct statistical comparisons between the two waveforms were not performed; as an exploratory pilot study with a small sample size ($n = 5$ per group), the protocol was focused on evaluating changes within each group and lacked the statistical power required for inter-group hypothesis testing. Collectively, these methodological constraints and the preliminary nature of the data preclude definitive conclusions regarding clinical efficacy or the superiority of one waveform over the other.

The results demonstrate the feasibility of peripheral electrical stimulation for modulating postural tremor in Parkinson's disease using BPB and SB waveforms. The individual evaluation of each group revealed median reductions in tremor-band power (3.5–12 Hz), with the SB group achieving statistically significant variations at the end of the intra-stimulation periods. Furthermore, exploratory analyses indicated that relative changes were not associated with individual characteristics such as f_{tremor}, TCI, and MF. Given the small sample size and the lack of statistical significance during post-stimulation periods, these findings must be interpreted as preliminary. Consequently, future studies should incorporate larger sample sizes using randomized, double-blind, sham-controlled crossover designs. This methodological approach will isolate stimulation-specific effects, enable direct quantitative comparisons between waveforms, and incorporate additional assessment periods to evaluate the clinical durability of the treatment.

ACKNOWLEDGMENT

The authors wish to thank CAPES (PROEX), FINEP (grant 01.24.0122.00), CNPq, and FAPERJ (grant E-26/204.393/2024 and SEI-260003/020273/2025) for their financial support. Additionally, the authors would like to express their gratitude to SisAssistiva-MCTI for its collaboration.

REFERENCES

[1] M. J. Armstrong and M. S. Okun, "Diagnosis and treatment of parkinson disease: A review," *JAMA*, vol. 323, pp. 548–560, Feb 2020.

[2] F. Magalhães, V. N. C. da Silva, J. R. N. Fernandes, A. T. de Oliveira, T. R. S. Fernandes, and S. Teixeira, "Causal theory, motor symptoms, non-motor symptoms, diagnosis, and treatment of parkinson's disease: a bibliographic review," *Research, Society and Development*, vol. 11, p. e10811729762, May 2022.

[3] V. Cabreira and J. Massano, "Parkinson's disease: Clinical review and update," *Acta Médica Portuguesa*, vol. 32, pp. 661–670, Oct 2019.

[4] A. Zampogna, A. Suppa, M. Patera, *et al.*, "Resting and action tremor in parkinson's disease: pathophysiological insights from long-term stn-dbs," *npj Parkinson's Disease*, vol. 11, Oct 2025.

[5] A. de Oliveira Andrade, A. P. S. Paixão, A. M. Cabral, A. G. Rabelo, L. M. D. Luiz, V. C. Dionísio, M. F. f. Vieira, J. M. Pereira, A. Rueda, S. Krishnan, and A. A. Pereira, "Task-specific tremor quantification in a clinical setting for parkinson's disease," *Journal of Medical and Biological Engineering*, vol. 40, no. 6, pp. 821–850, 2020.

[6] M. F. Dirkx, H. Zach, B. R. Bloem, M. Hallett, and R. C. Helmich, "The nature of postural tremor in parkinson disease," *Neurology*, vol. 90, pp. e1095–e1103, Mar 2018.

[7] W. Pirker, R. Katzenschlager, M. Hallett, and W. Poewe, "Pharmacological treatment of tremor in parkinson's disease revisited," *Journal of Parkinson's Disease*, vol. 13, no. 2, pp. 127–144, 2023.

[8] R. F. Dallapiazza, P. de Vloo, A. Fomenko, D. J. Lee, C. Hamani, R. P. Munhoz, M. Hodaie, A. M. Lozano, A. Fasano, and S. K. Kalia, "Considerations for patient and target selection in deep brain stimulation surgery for parkinson's disease," in *Parkinson's Disease*, ch. 8, pp. 145–160, Exon Publications, 2018.

[9] R. M. G. Barbosa *et al.*, "New perspectives of deep brain stimulation indications for parkinson's disease: A critical review," *Brain Sciences*, vol. 14, p. 638, Jun 2024.

[10] J. S. Lora-Millan *et al.*, "A review on wearable technologies for tremor suppression," *Frontiers in Neurology*, vol. 12, p. 700600, Aug 2021.

[11] H. S. Nguyen and T. P. Luu, "Tremor-suppression orthoses for the upper limb: Current developments and future challenges," *Frontiers in Human Neuroscience*, vol. 15, p. 622535, Apr 2021.

[12] S. Brillman, "Non-invasive transcutaneous afferent patterned stimulation therapy offers action tremor relief in parkinson's disease," *Tremor and Other Hyperkinetic Movements*, vol. 13, no. 1, p. 25, 2023.

[13] R. Pahwa, R. Dhall, J. Ostrem, R. Gwinn, K. Lyons, S. Ro, C. Dietiker, N. Luthra, P. Chidester, S. Hamner, E. Ross, and S. Delp, "An acute randomized controlled trial of noninvasive peripheral nerve stimulation in essential tremor," *Neuromodulation: Technology at the Neural Interface*, vol. 22, no. 5, pp. 537–545, 2019.

[14] J. Y. Yu, A. Rajagopal, J. Syrkin-Nikolau, S. Shin, K. H. Rosenbluth, D. Khosla, E. K. Ross, and S. L. Delp, "Transcutaneous afferent patterned stimulation therapy reduces hand tremor for one hour in essential tremor patients," *Frontiers in Neuroscience*, vol. 14, p. 530300, 2020.

[15] A. S. Barath, "Brain metabolic changes with longitudinal transcutaneous afferent patterned stimulation in essential tremor subjects," *Tremor and Other Hyperkinetic Movements*, vol. 10, no. 0, p. 52, 2020.

[16] S. Brillman, "Real-world evidence of transcutaneous afferent patterned stimulation for essential tremor," *Tremor and Other Hyperkinetic Movements*, vol. 12, no. 1, p. 27, 2022.
[17] H. Resende-Martins, R. Zanetti, C. Cardoso-dos Santos, G. Mastrocola-Manzano, and C. J. Tierra-Criollo, "Current perception threshold and reaction time in the assessment of sensory peripheral nerve fibers through sinusoidal electrical stimulation at different frequencies," *Revista Brasileira de Engenharia Biomédica*, vol. 29, pp. 278–285, 2013.
[18] L. P. Volpi, "Efeitos dos estímulos elétricos senoidais na resposta induzida cerebral: estudo da seletividade das fibras sensitivas," dissertação de mestrado em engenharia biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2017.
[19] A. Smid, J. W. J. Elting, J. M. C. van Dijk, B. Otten, D. L. M. Oterdoom, K. Tamasi, T. Heida, T. van Laar, and G. Drost, "Intraoperative quantification of mds-updrs tremor measurements using 3d accelerometry: A pilot study," *Journal of Clinical Medicine*, vol. 11, no. 9, p. 2275, 2022.
[20] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, and N. LaPelle, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.
[21] D. A. Molina-Vidal, P. Cevallos Larrea, L. Guambaña Calle, D. Liquori, and C. J. Tierra-Criollo, "Design and evaluation of a prototype of dual channel electrical stimulator: Application in the assessment of current perception thresholds to multiple stimuli," in *Smart Technologies, Systems and Applications: SmartTech-IC 2022*, vol. 1705 of *Communications in Computer and Information Science*, pp. 342–355, Springer, 2023.
[22] P. S. S. Cevallos, D. M. da Silva-Liquori, and C. J. Tierra-Criollo, "Estimador da frequência de tremor em parkinsonianos com vetor aleatório de acelerometria," 2025.
[23] H. Dai, P. Zhang, and T. C. Lueth, "Quantitative assessment of parkinsonian tremor based on an inertial measurement unit," *Sensors*, vol. 15, no. 10, pp. 25055–25071, 2015.
[24] F. Luft, S. Sharifi, W. Mugge, A. C. Schouten, L. J. Bour, A.-F. van Rootselaar, P. H. Veltink, and T. Heida, "A power spectral density-based method to detect tremor and tremor intermittency in movement disorders," *Sensors*, vol. 19, no. 19, p. 4301, 2019.
[25] J. J. Fernández-Pérez, J. Avendaño-Coy, D. Serrano-Muñoz, F. Oliveira Barroso, C. Montero-Pardo, B. López-Moreno, A. Lerín-Calvo, J. P. Romero Muñoz, and J. Gómez-Soriano, "Transcutaneous kilohertz high-frequency alternating current at 10 khz for upper-limb tremor in people with parkinson's disease: A double-blind, randomized, crossover study," *Journal of Clinical Medicine*, vol. 13, no. 24, 2024.
[26] J. L. Dideriksen, C. M. Laine, S. Dosen, S. Muceli, E. Rocon, J. L. Pons, J. Benito-Leon, and D. Farina, "Electrical stimulation of afferent pathways for the suppression of pathological tremor," *Frontiers in Neuroscience*, vol. 11, 2017.
[27] J. Kim, T. Wichmann, O. T. Inan, and S. P. DeWeerth, "Analyzing the effects of parameters for tremor modulation via phase-locked electrical stimulation on a peripheral nerve," *Journal of Personalized Medicine*, vol. 12, no. 1, p. 76, 2022.

Ensemble Learning with Lightweight Convolutional Neural Networks for Mosquito Species Classification

1st Hygor Leal
Graduate Program Student
Federal University of Ceará (UFC)
Sobral, Brazil
hygorleal@alu.ufc.br

2nd Raniery Alves Vasconcelos
Graduate Program Student
Federal University of Ceará (UFC)
Sobral, Brazil
raniery_vasconcelos@alu.ufc.br

3rd Lucas José Lemos Braz
Graduate Program Student
Federal University of Ceará (UFC)
Sobral, Brazil
lucas.jlb19@alu.ufc.br

4th Vinicius Costa dos Santos
B.Sc. in Computer Engineering
Federal University of Ceará (UFC)
Sobral, Brazil
vinicius123cs2@alu.ufc.br

5th Lucas Pedrosa Valente
M.Sc. in Electrical and Computer Engineering
Federal University of Ceará (UFC)
Sobral, Brazil
lucaspedrosaval@alu.ufc.br

6th Marcelo Marques Simões de Souza
Graduate Program and Electrical Engineering Professor
Federal University of Ceará (UFC)
Sobral, Brazil
marcelo.mssouza@gmail.com

7th Iális Cavalcante de Paula Júnior
Graduate Program and Electrical Engineering Professor
Federal University of Ceará (UFC)
Sobral, Brazil
ialis@sobral.ufc.br

***Abstract*—Automatic identification of mosquito species is central to vector surveillance and to the timely deployment of disease-control interventions. This paper proposes an ensemble learning framework built on lightweight convolutional neural networks (CNNs) for classifying mosquito species from wingbeat audio recordings. Six state-of-the-art lightweight architectures—MobileNetV3-Large, ShuffleNetV2, GhostNetV2, EfficientNet-Lite0, MobileOne-S1, and MobileNetV4 Conv Medium—are trained on log-Mel spectrogram representations of the Wingbeats dataset, whose training partition is augmented with realistic environmental noise drawn from the TUT Acoustic Scenes 2017 dataset. Individual model predictions are combined through a weighted soft-voting ensemble, in which each classifier's weight is proportional to its validation F1-score. Performance is assessed under a 5-fold cross-validation protocol with fixed fold assignments, each fold further evaluated under three controlled random seeds, for a total of 15 fold-seed configurations per model. Under this protocol, the proposed ensemble reaches a macro F1-score of $95.5\% \pm 0.29$ on the Wingbeats test set, surpassing every individual classifier with low variability across runs; pairwise Wilcoxon signed-rank tests with Holm correction confirm that this advantage is statistically significant against all individual models and against most alternative ensemble strategies. Cross-dataset validation on Tiny MosquitoSong+, HumBugDB, and ABuzz provides evidence of generalization across the evaluated recording conditions, with the ensemble attaining the highest ROC-AUC on all three external benchmarks (0.901, 0.951, and 0.963, respectively). These results support weighted soft voting as a promising approach to mosquito classification on the evaluated benchmarks; readiness for field or embedded deployment remains to be established.**

***Index Terms*—mosquito classification, wingbeat sound, lightweight convolutional neural networks, ensemble learning, log-Mel spectrogram, acoustic surveillance, IoT**

I. Introduction

Mosquito-borne diseases – including dengue, malaria, Zika, chikungunya, and yellow fever – impose a substantial public health burden, particularly across tropical and subtropical regions [1]. Effective vector control depends on continuous population surveillance, yet conventional trapping and manual, expert-driven identification are labor-intensive and difficult to scale to the continuous, large-area monitoring that timely interventions require.

Acoustic monitoring offers a non-invasive alternative, exploiting the characteristic wingbeat frequencies that mosquitoes produce in flight [5]. Coupled with low-cost microphones and embedded processing, this approach could support continuous, automated identification within smart traps and Internet of Things (IoT) surveillance devices. Convolutional neural networks (CNNs) trained on time-frequency representations, such as log-Mel spectrograms, have already achieved high classification accuracy on this task [6], [7]. Two obstacles nonetheless stand between these results and field-ready systems: the computational demands of many CNN architectures can be prohibitive for embedded deployment, and evaluation restricted to a single acquisition protocol offers limited evidence of robustness to different recording devices, environmental noise, and mosquito populations.

Lightweight CNNs may help mitigate the first obstacle by reducing computational requirements while retaining competitive predictive performance [21], [22], [24], [25]. Because architectures of this kind differ in their design strategy, combining their complementary predictions through ensemble learning may further improve robustness [19]; likewise, training with realistic environmental noise may help narrow the gap between laboratory and deployment conditions [16]. Ensembling several networks, however, multiplies the inference cost of a single model, so the computational footprint of the complete ensemble still warrants dedicated evaluation before such a system can be considered for embedded deployment.

Building on these observations, this work proposes a framework that combines six lightweight CNN architectures, noise-augmented wingbeat recordings, log-Mel spectrogram representations, and a validation-F1-weighted soft-voting ensemble. Models are evaluated on the Wingbeats dataset using five fixed cross-validation folds and three controlled random seeds per fold, keeping data partitions consistent across architectures while explicitly assessing sensitivity to training stochasticity. External validation on the Tiny MosquitoSong+, HumBugDB, and ABuzz datasets further probes the framework's generalization under recording conditions the models never encountered during training. Individual architectures and multiple ensemble strategies are systematically compared in terms of predictive performance and result stability, with statistical testing used to establish which differences are robust rather than incidental.

II. Related Work

Automatic mosquito identification has been investigated through multiple sensing modalities, including morphological imaging, spectroscopic analysis, and acoustic monitoring. This section reviews representative contributions across these areas, with emphasis on wingbeat-based acoustic methods and the machine learning techniques employed for classification.

A. Mosquito Classification from Visual Data

Early automated approaches to mosquito identification relied on image-based analysis of captured specimens. Transfer learning with standard CNN architectures such as VGG16 has been applied to morphological mosquito images, achieving high classification accuracy even with compact training sets [2]. More recently, vision transformers trained on curated image datasets have outperformed CNN baselines in multi-species classification, with a Swin Transformer reaching 96.3% accuracy across 17 species [3]. Nazari et al. [4] took a hybrid route, combining a compact ensemble of three CNNs—EfficientNet-V2-S, MobileNetV3-Large, and ResNet-50—with a Swin-T Transformer branch for *Aedes* mosquito recognition, reporting in-domain accuracies of 99.3–99.5% together with evidence of cross-dataset generalization. Despite these encouraging accuracies under laboratory conditions, image-based methods generally require physical specimen collection and expert handling, which constrains their applicability to continuous, large-scale field surveillance.

B. Acoustic Monitoring and Wingbeat-Based Classification

The characteristic wingbeat sound that mosquitoes produce during flight offers a non-invasive cue for species identification. Early studies established that flight tones carry spectral features that are discriminative across species and sexes [5], and subsequent work by Kiskin et al. showed that CNNs trained on spectrogram representations of this audio achieve high classification accuracy under controlled laboratory conditions [7].

Building on these findings, several studies investigated low-cost acoustic sensing suitable for IoT deployment. Joelianto et al. [8] proposed an IoT solution that differentiates both the sex and species of *Culex quinquefasciatus* and *Aedes aegypti*, using a custom CNN to extract features from the Mel-Frequency Cepstral Coefficients (MFCCs) of wingbeat Mel spectrograms and reporting accuracies of 91.76% and 87.16% for species and sex differentiation, respectively; the model, however, was trained on a limited species set recorded with mobile phones under controlled laboratory conditions. Paim et al. [9] instead targeted a smartphone application, using residual CNNs to identify *Aedes aegypti* from a benchmarking dataset of smartphone-recorded wingbeat sounds augmented with controlled levels of random and background noise; deployed on-device as a proof of concept, the model reached 80% test accuracy.

Other efforts moved away from spectrogram-based pipelines altogether. Yin et al. proposed two lightweight models—a 1D-CNN and a 1D-CNN combined with an LSTM—that operate directly on raw audio at low sampling rates without any preprocessing, achieving classification accuracy above 93% across five mosquito species [10]. Luna-González et al. explored yet another representation, transforming wingbeat features into images before training CNN classifiers on these transformed inputs [12]. Returning to spectrogram-based features, Fanioudakis et al. applied DenseNet121 to wingbeat spectrograms, demonstrating the effectiveness of time-frequency representations for distinguishing medically relevant species and reaching 96% accuracy on the Wingbeats dataset [6].

A different architectural direction is represented by WbNet, a ResNet augmented with a self-attention mechanism that combines Mel-spectrogram inputs with data augmentation and residual connections, achieving 98.9% accuracy on the ABuzz dataset for species of the *Aedes* and *Culex* genera [13]. Closer to the present work, Fernandes et al. investigated CNN-based binary and multiclass classifiers alongside ensembles of binary classifiers for *Aedes aegypti* detection from smartphone recordings, providing early evidence that ensemble strategies can improve robustness under noisy acoustic conditions [14].

C. Noise Robustness and Cross-Dataset Generalization

A recurring limitation of acoustic mosquito classification is the domain gap between clean laboratory recordings and noisy field environments, which has motivated several noise-aware training strategies. Kiskin et al. introduced HumBugDB [15], a benchmark of over 20 hours of multi-species recordings

collected with smartphones across different countries and environments, explicitly designed to support more realistic evaluation; models trained exclusively on clean data typically show substantial performance degradation when evaluated on this benchmark.

Targeting this gap directly, the MosquitoSong+ model extended a baseline 1D-CNN with two complementary augmentation strategies – additive noise and wingbeat volume variation [16]. Evaluated across multiple datasets, including HumBugDB and outdoor recordings, MosquitoSong+ maintained species classification accuracy above 80% in the presence of background noise, underscoring the importance of noise-aware training for real-world deployment.

D. Ensemble Learning for Audio Classification

Ensemble methods have been widely adopted to improve the robustness of audio classification systems more broadly [18]. Soft voting and weighted prediction fusion across independently trained models consistently outperform single-model baselines in environmental sound classification [19], and within bioacoustics specifically, ensembles of CNNs trained with diverse augmentation protocols have been applied to animal audio classification, showing that combining models trained under different data conditions reduces isolated classification errors [20].

E. Lightweight CNN Architectures for Edge Deployment

Deploying deep learning models on embedded devices requires architectures that balance classification performance against low computational cost. MobileNetV3-Large combines depthwise separable convolutions with neural architecture search and squeeze-and-excitation attention to optimize mobile inference [21]. ShuffleNetV2 instead reduces memory-access overhead through channel splitting and shuffling, improving practical inference throughput on mobile processors [22]. GhostNetV2 generates redundant feature maps via inexpensive linear transformations and adds a decoupled fully connected attention mechanism to capture long-range dependencies at low cost [23]. EfficientNet applies compound scaling jointly across depth, width, and resolution; its Lite variant removes hardware-unfriendly operations to enable deployment on edge accelerators [24]. MobileOne relies on structural reparameterization during training, which collapses into a single efficient branch at inference time to target ultra-low latency on mobile hardware [25]. MobileNetV4, finally, integrates hardware-aware architectural design to maximize accuracy under strict latency constraints [26].

Despite their maturity on computer vision benchmarks, a systematic comparison of these architectures on bioacoustic tasks – and their combination through ensemble learning – remains largely unexplored. The present work addresses this gap, evaluating all six architectures under a unified experimental protocol and combining their predictions through a weighted soft-voting ensemble.

III. Methodology

Figure 1 presents an overview of the proposed mosquito species classification framework, whose pipeline comprises seven stages: (i) acquisition of mosquito wingbeat recordings, (ii) audio preprocessing, (iii) data augmentation with realistic environmental noise, (iv) log-Mel spectrogram generation, (v) training of the lightweight convolutional neural networks, (vi) weighted soft-voting ensemble learning, and (vii) performance evaluation on the Wingbeats and external datasets.

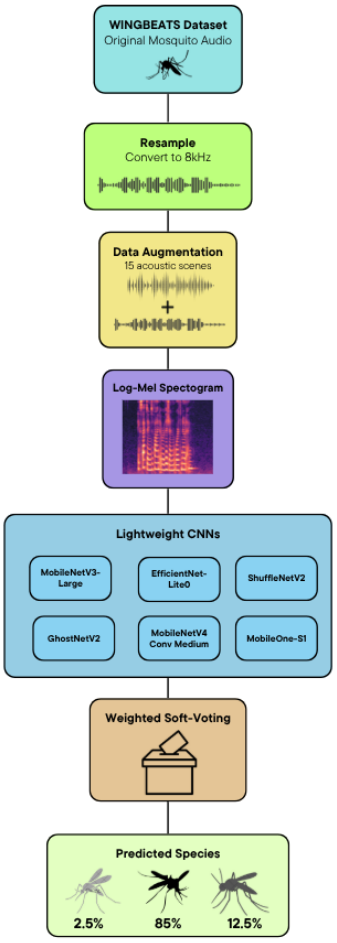


Fig. 1. Overview of the proposed pipeline

A. Dataset

The experiments were conducted on the Wingbeats dataset, which comprises mosquito wingbeat recordings from six medically relevant species. As shown in Table I, the original dataset is markedly imbalanced across species.

Before any data augmentation or spectrogram generation, 4,500 original recordings per species were reserved for validation and a further 4,500 for testing; from the remaining recordings, 7,200 per species were selected for training. All assignments were made at the level of the original recording (*source recording*), so that no source recording was shared across subsets.

Data augmentation was applied exclusively to the training recordings: for each species, 28,800 noise-augmented samples were generated from the 7,200 selected training recordings, yielding 36,000 training samples in total – 7,200 original recordings (20%) plus 28,800 augmented samples (80%). The validation and test subsets, by contrast, retained only their 4,500 original recordings per species and were not subjected to augmentation or synthetic noise addition. Each species therefore contributed 16,200 original and 28,800 augmented samples, for a combined total of 45,000 samples across the three subsets.

Aggregated across the six species, the final balanced dataset comprises 216,000 training samples, 27,000 validation samples, and 27,000 test samples, for a total of 270,000 samples. Although this corresponds nominally to an 80/10/10 split, the ratio actually results from the fixed per-species validation/test allocation combined with training-set augmentation, rather than from a proportional division of the original Wingbeats recordings. As detailed in Section III-F, this partitioning scheme defines one of five fixed folds used throughout the repeated cross-validation protocol, with the same fold assignments held constant across all models and random seeds.

TABLE I
CLASS DISTRIBUTION OF THE ORIGINAL WINGBEATS DATASET AND THE FINAL BALANCED DATASET ACROSS TRAINING, VALIDATION, AND TEST SUBSETS.

Species	**Original**	**Balanced**
Ae. aegypti	85,553	45,000
C. quinquefasciatus	74,599	45,000
An. gambiae	49,471	45,000
C. pipiens	30,415	45,000
Ae. albopictus	20,231	45,000
An. arabiensis	19,297	45,000

B. Audio Preprocessing

The proposed framework operates directly on mosquito wingbeat audio recordings. All recordings were first standardized to a common audio format by resampling the signals to a fixed sampling frequency and applying amplitude normalization, a step that reduces variability caused by different recording conditions while preserving the spectral characteristics of mosquito wingbeats.

To improve model robustness, data augmentation was applied exclusively to the training subset, adding realistic environmental sounds directly to the time-domain signals [19]. Noise samples were drawn at random from the TUT Acoustic Scenes 2017 dataset and mixed with mosquito recordings at randomly selected signal-to-noise ratio (SNR) values. Performing this augmentation before feature extraction allows the resulting spectrograms to reflect more realistic acoustic scenarios, which is expected to improve generalization to unseen recording environments [16].

Each audio signal was then converted into a log-Mel spectrogram. Validation and test recordings underwent the same standardization and feature-extraction steps but bypassed augmentation. The Short-Time Fourier Transform (STFT) was first computed to obtain a time-frequency representation; the resulting power spectrum was projected onto the Mel scale and compressed logarithmically to produce the log-Mel spectrogram. Finally, spectrograms were resized to match the input dimensions required by each lightweight CNN architecture.

C. Lightweight CNN Architectures

Six lightweight CNNs were evaluated: MobileNetV3-Large [21], ShuffleNetV2 [22], GhostNetV2 [23], EfficientNet-Lite0 [24], MobileOne-S1 [25], and MobileNetV4 Conv Medium [26]. As reviewed in Section II, these architectures pursue complementary design strategies – depthwise convolutions, channel shuffling, inexpensive feature generation, compound scaling, structural reparameterization, and hardware-aware design – which motivates their use as diverse ensemble members. All six models received identical log-Mel spectrogram inputs and were compared under the same data partitions and training protocol.

D. Weighted Soft-Voting Ensemble

Although each lightweight CNN provides high classification accuracy on its own, its particular design leads it to learn a somewhat distinct feature representation, so that prediction errors across architectures are often complementary rather than correlated [20]. The proposed framework exploits this diversity by combining the outputs of all trained CNNs through a weighted soft-voting ensemble.

Let $P_i(c|x)$ denote the posterior probability estimated by the i-th CNN for class c, where x represents the input log-Mel spectrogram. The final probability assigned to class c is computed as

$$P(c|x) = \sum_{i=1}^{N} w_i P_i(c|x), \tag{1}$$

where N is the number of CNN models and w_i is the weight assigned to the i-th classifier, satisfying

$$\sum_{i=1}^{N} w_i = 1. \tag{2}$$

The weights were computed from the validation F1-score of each model and normalized as

$$w_i = \frac{F1_i}{\sum_{j=1}^{N} F1_j}, \tag{3}$$

where $F1_i$ denotes the validation F1-score achieved by the i-th CNN.

The predicted mosquito species corresponds to the class with the highest aggregated probability,

$$\hat{y} = \arg\max_{c} P(c|x). \tag{4}$$

Unlike hard voting, which considers only each classifier's predicted label, weighted soft voting exploits the confidence scores produced by the networks and assigns greater influence

to models with stronger validation performance [19]. This combination aims to exploit complementary feature representations and reduce isolated classification errors; any robustness gains must be assessed under the intended operating conditions.

E. Experimental Setup and Hyperparameters

All architectures were trained on the source-recording-level partitions described in the Dataset subsection; derived spectrograms and augmented samples always remained within their source subset, so that only training data were augmented and no leakage into validation or testing could occur.

Training used AdamW with weight decay 10^{-4}, batch size 64, and a maximum of 50 epochs. With a peak learning rate $\eta = 10^{-3}$, a OneCycleLR schedule increased the learning rate from $\eta/10$ to η over the first 30% of training before applying cosine decay down to $\eta/1000$. Early stopping monitored validation loss with a patience of 10 epochs and restored the best-performing model weights.

F. Evaluation Protocol

Five fixed folds were used for cross-validation, with four folds allocated to training and the remaining fold to evaluation at each iteration; identical fold assignments were maintained across all models and ensemble configurations. Each fold was further evaluated under three controlled random seeds, yielding 15 fold-seed configurations per model. Seeds governed weight initialization, data shuffling, and other randomized training operations without altering the partitions themselves; preprocessing, training-only augmentation, and hyperparameters followed the settings described in the preceding subsections.

Results are reported as mean $\pm$ standard deviation across the 15 configurations, with corresponding 95% confidence intervals (CIs) computed from these statistics to summarize estimation uncertainty. This protocol jointly captures performance variation across fixed partitions and across stochastic training runs, while its paired structure enables direct statistical comparison between models.

Statistical comparisons paired results from identical fold-seed configurations. A two-sided Wilcoxon signed-rank test compared the Soft Voting Ensemble against each alternative at $\alpha = 0.05$, with Holm correction applied for multiple comparisons; effect sizes were expressed as rank-biserial correlation. A method was considered superior only when it combined a higher mean performance with a statistically significant paired difference, with effect sizes providing complementary evidence on the magnitude of that difference.

G. Evaluation Metrics

On the Wingbeats test set, the classification performance of the individual lightweight CNNs and of the proposed weighted soft-voting ensemble was evaluated using five metrics: Accuracy, Macro Precision, Macro Recall, Macro F1-score, and ROC-AUC. Accuracy captures the overall proportion of correctly classified samples, whereas Macro Precision and Macro Recall assign equal importance to each class regardless of its frequency. The Macro F1-score, computed as the harmonic mean of Macro Precision and Macro Recall, was adopted as the primary evaluation metric throughout the paper.

F1-score and the area under the receiver operating characteristic curve (ROC-AUC) were additionally reported for the cross-dataset validation, alongside the Wingbeats test set.

Accuracy, Macro Precision, Macro Recall, and Macro F1-score are defined as follows:

$$\text{Accuracy} = \frac{\sum_{c=1}^{K} TP_c}{\text{Total Number of Samples}} \tag{5}$$

$$\text{Precision}_M = \frac{1}{K} \sum_{c=1}^{K} \frac{TP_c}{TP_c + FP_c} \tag{6}$$

$$\text{Recall}_M = \frac{1}{K} \sum_{c=1}^{K} \frac{TP_c}{TP_c + FN_c} \tag{7}$$

$$\text{F1}_M = 2 \cdot \frac{\text{Precision}_M \cdot \text{Recall}_M}{\text{Precision}_M + \text{Recall}_M} \tag{8}$$

IV. Results and Discussion

The proposed methodology was evaluated using six lightweight CNN architectures and the weighted soft-voting ensemble, under the fixed-fold cross-validation protocol described in Section III-F. Performance was first assessed on the Wingbeats test set and then examined through cross-dataset validation on Tiny MosquitoSong+ [16], HumBugDB [15], and ABuzz. Wingbeats performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC, whereas cross-dataset performance was assessed using F1-score and ROC-AUC only; accordingly, Table II reports all five metrics, while Table IV reports F1-score and ROC-AUC.

A. Performance on the Wingbeats Dataset

Table II compares the six individual lightweight CNN architectures against hard voting, bagging, stacking, and the proposed weighted soft-voting ensemble on the Wingbeats dataset. All values are reported as mean $\pm$ standard deviation across the 15 fold-seed configurations; Accuracy, Precision, Recall, and F1-score are expressed as percentages, while ROC-AUC is reported on a 0–1 scale.

All six lightweight CNNs achieved high classification performance individually, confirming that log-Mel spectrograms are a highly discriminative representation for mosquito wingbeat sounds [6], [13]. Accuracy ranged from 90.6% for ShuffleNetV2, the lightest and least accurate architecture evaluated, up to $94.4\% \pm 0.47$ for MobileNetV4 Conv Medium, closely followed by MobileOne-S1 and GhostNetV2 – a pattern broadly consistent with the added representational capacity of the more recent architectures.

Every ensemble strategy outperformed every individual architecture, and among them the proposed weighted soft-voting ensemble achieved the best results across all five reported metrics: accuracy of $95.6\% \pm 0.28$, precision of $95.3\% \pm 0.31$,

TABLE II

PERFORMANCE COMPARISON USING 5-FOLD CROSS-VALIDATION WITH FIXED FOLD ASSIGNMENTS AND THREE CONTROLLED RANDOM SEEDS PER FOLD. EACH CELL REPORTS THE MEAN ± STANDARD DEVIATION ON THE FIRST LINE AND THE 95% CONFIDENCE INTERVAL ON THE SECOND, ACROSS 15 FOLD-SEED CONFIGURATIONS.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**ROC-AUC**
MobileNetV3-Large	91.3 ± 0.82	90.8 ± 0.87	91.5 ± 0.76	91.1 ± 0.81	0.963 ± 0.004
	[90.9, 91.7]	[90.4, 91.2]	[91.1, 91.9]	[90.7, 91.5]	[0.961, 0.965]
EfficientNet-Lite0	92.7 ± 0.71	92.3 ± 0.75	93.1 ± 0.68	92.7 ± 0.70	0.971 ± 0.003
	[92.3, 93.1]	[91.9, 92.7]	[92.7, 93.5]	[92.3, 93.1]	[0.969, 0.973]
ShuffleNetV2	90.6 ± 0.91	90.1 ± 0.96	91.2 ± 0.84	90.6 ± 0.89	0.958 ± 0.005
	[90.1, 91.1]	[89.6, 90.6]	[90.7, 91.7]	[90.1, 91.1]	[0.955, 0.961]
GhostNetV2	93.1 ± 0.63	93.4 ± 0.59	92.6 ± 0.68	93.0 ± 0.61	0.975 ± 0.003
	[92.8, 93.4]	[93.1, 93.7]	[92.2, 93.0]	[92.7, 93.3]	[0.973, 0.977]
MobileOne-S1	93.8 ± 0.52	93.5 ± 0.55	94.2 ± 0.48	93.8 ± 0.51	0.981 ± 0.002
	[93.5, 94.1]	[93.2, 93.8]	[93.9, 94.5]	[93.5, 94.1]	[0.980, 0.982]
MobileNetV4 Conv Medium	94.4 ± 0.47	94.6 ± 0.44	94.1 ± 0.49	94.3 ± 0.46	0.985 ± 0.002
	[94.2, 94.7]	[94.4, 94.8]	[93.8, 94.4]	[94.0, 94.6]	[0.984, 0.986]
Hard Voting Ensemble	94.6 ± 0.43	94.3 ± 0.46	94.8 ± 0.40	94.5 ± 0.42	0.986 ± 0.002
	[94.4, 94.8]	[94.1, 94.5]	[94.6, 95.0]	[94.3, 94.7]	[0.985, 0.987]
Bagging Ensemble	94.9 ± 0.39	94.7 ± 0.42	95.1 ± 0.37	94.9 ± 0.39	0.988 ± 0.002
	[94.7, 95.1]	[94.5, 94.9]	[94.9, 95.3]	[94.7, 95.1]	[0.987, 0.989]
Stacking Ensemble	95.4 ± 0.31	95.2 ± 0.34	95.6 ± 0.30	95.4 ± 0.31	0.990 ± 0.001
	[95.2, 95.6]	[95.0, 95.4]	[95.4, 95.8]	[95.2, 95.6]	[0.990, 0.991]
Soft Voting Ensemble	$\mathbf{95.6 \pm 0.28}$	$\mathbf{95.3 \pm 0.31}$	$\mathbf{95.8 \pm 0.27}$	$\mathbf{95.5 \pm 0.29}$	$\mathbf{0.991 \pm 0.001}$
	[95.4, 95.8]	**[95.1, 95.5]**	**[95.7, 96.0]**	**[95.3, 95.7]**	**[0.990, 0.992]**

TABLE III

PAIRWISE STATISTICAL COMPARISON OF THE SOFT VOTING ENSEMBLE AGAINST INDIVIDUAL MODELS AND ALTERNATIVE ENSEMBLE STRATEGIES. COMPARISONS WERE PERFORMED OVER 15 PAIRED FOLD-SEED CONFIGURATIONS USING THE WILCOXON SIGNED-RANK TEST. HOLM-ADJUSTED p-VALUES RETAIN THE ORIGINAL TEN-COMPARISON FAMILY. EFFECT SIZES ARE REPORTED AS RANK-BISERIAL CORRELATION (r_{rb}).

Comparison	**Δ Acc. (pp)**	**p-value**	**Holm p**	r_{rb}	**Sig.**
Soft Voting vs. MobileNetV3-Large	+4.3	< 0.001	< 0.001	0.94	Yes
Soft Voting vs. EfficientNet-Lite0	+2.9	< 0.001	< 0.001	0.91	Yes
Soft Voting vs. ShuffleNetV2	+5.0	< 0.001	< 0.001	0.96	Yes
Soft Voting vs. GhostNetV2	+2.5	< 0.001	< 0.001	0.89	Yes
Soft Voting vs. MobileOne-S1	+1.8	< 0.001	0.001	0.82	Yes
Soft Voting vs. MobileNetV4 Conv Medium	+1.2	0.002	0.010	0.71	Yes
Soft Voting vs. Hard Voting	+1.0	0.003	0.012	0.69	Yes
Soft Voting vs. Bagging	+0.7	0.009	0.027	0.58	Yes
Soft Voting vs. Stacking	+0.2	0.118	0.118	0.27	No

recall of $95.8\% \pm 0.27$, F1-score of $95.5\% \pm 0.29$, and ROC-AUC of 0.991 ± 0.001. Its ROC-AUC of 0.991 compares favorably with the 0.985 obtained by MobileNetV4 Conv Medium, the strongest individual model. This margin, small as it is in absolute terms, is consistent with the individual CNNs having learned complementary feature representations that the ensemble exploits to correct isolated classification errors through probabilistic fusion [20]. The comparatively low standard deviations observed across fold-seed configurations further indicate that the ensemble's performance is stable under controlled variations of the training process, suggesting that its advantage over individual models reflects the modeling strategy itself rather than a favorable fold or seed.

B. Confidence Intervals and Statistical Significance

Table II also reports the corresponding 95% confidence intervals, computed across the 15 fold-seed configurations for each model and ensemble strategy.

The confidence intervals in Table II indicate low variability for the Soft Voting Ensemble: its F1-score interval is $[95.3, 95.7]$, compared with $[93.5, 94.1]$ for MobileOne-S1 and $[90.1, 91.1]$ for ShuffleNetV2. Paired Wilcoxon signed-rank tests, described in Section III-F, assessed its advantage over each alternative method.

Table III reports accuracy gaps, raw and Holm-adjusted Wilcoxon p-values, and rank-biserial effect sizes (r_{rb}). After Holm correction, the comparison with Stacking does not reach significance at $\alpha = 0.05$, consistent with its confidence intervals overlapping those of weighted soft voting in Table II. The advantage over every individual CNN and the Hard Voting and Bagging ensembles remains significant.

C. Cross-Dataset Validation

To evaluate robustness under acquisition conditions never seen during training, all trained models were additionally tested on the Tiny MosquitoSong+ [16], HumBugDB [15], and ABuzz datasets. Table IV reports the F1-score and ROC-AUC obtained by each model on these three external datasets.

TABLE IV
CROSS-DATASET VALIDATION RESULTS. BEST VALUES FOR EACH DATASET ARE SHOWN IN BOLD.

Model	F1-score (%)	ROC-AUC
Tiny MosquitoSong+		
MobileNetV3-Large	72.4	0.806
EfficientNet-Lite0	74.8	0.827
ShuffleNetV2	71.3	0.793
GhostNetV2	75.6	0.839
MobileOne-S1	77.1	0.851
MobileNetV4 Conv Medium	78.3	0.864
Soft Voting Ensemble	**80.7**	**0.901**
HumBugDB		
MobileNetV3-Large	74.1	0.831
EfficientNet-Lite0	76.5	0.851
ShuffleNetV2	73.0	0.816
GhostNetV2	77.2	0.862
MobileOne-S1	78.8	0.876
MobileNetV4 Conv Medium	80.1	0.889
Soft Voting Ensemble	**87.4**	**0.951**
ABuzz		
MobileNetV3-Large	77.8	0.861
EfficientNet-Lite0	79.1	0.878
ShuffleNetV2	75.9	0.842
GhostNetV2	81.0	0.893
MobileOne-S1	80.4	0.886
MobileNetV4 Conv Medium	83.2	0.914
Soft Voting Ensemble	**89.1**	**0.963**

Among the individual architectures, MobileNetV4 Conv Medium again achieved the highest ROC-AUC on all three external datasets (0.864 on Tiny MosquitoSong+, 0.889 on HumBugDB, and 0.914 on ABuzz), mirroring its lead on the in-domain Wingbeats evaluation.

The ensemble achieved the highest F1-score and ROC-AUC on all three external datasets: F1-scores of 80.7%, 87.4%, and 89.1%, and ROC-AUC values of 0.901, 0.951, and 0.963 on Tiny MosquitoSong+, HumBugDB, and ABuzz, respectively. Its ROC-AUC gains over the best individual model were 0.037, 0.062, and 0.049, exceeding the 0.006 gain on Wingbeats. This pattern suggests potential benefits under the acquisition-related shifts represented by these benchmarks, but does not establish robustness in operational field settings.

D. Overall Discussion

The low variability across 15 fold-seed configurations indicates stable classification performance under controlled changes in data partition and training stochasticity. External-dataset results are consistent with improved robustness from realistic noise augmentation [16], [19], but this contribution cannot be isolated without the augmentation ablation discussed in Section IV-E.

Weighted soft voting outperformed all individual CNNs, consistent with complementary acoustic representations [20]. Architectural differences—including channel shuffling, attention, hardware-aware design, and structural reparameterization—provide a plausible basis for this diversity [21], [22], [25], [26]; their individual contributions nevertheless require ablation studies.

Table III reports significant advantages over every individual model and the Hard Voting and Bagging ensembles after Holm correction, but not over Stacking. Unlike stacking, weighted soft voting requires no additional meta-classifier: it combines model probabilities using normalized validation F1-scores. It therefore offers a simpler combination procedure with higher mean performance, without establishing statistical superiority over stacking. Reusing existing lightweight architectures avoids designing a new network, but does not establish real-time feasibility on embedded hardware.

E. Limitations

Although the proposed methodology achieved promising results, some limitations should be acknowledged, both to contextualize the present findings and to guide future work. The models were trained primarily on the Wingbeats dataset, which consists of recordings acquired under controlled laboratory conditions. Realistic environmental noise from the TUT Acoustic Scenes 2017 dataset was incorporated during augmentation, but the acoustic characteristics of field recordings may still differ due to environmental factors. The framework was evaluated combining noise augmentation with the full six-architecture ensemble, so the individual contribution of each design choice cannot be isolated from the results reported here. In particular, no augmentation ablation was performed to contrast models trained with and without noise-augmented samples;likewise, no architecture ablation, such as a systematic leave-one-out analysis removing each CNN from the ensemble in turn, was conducted. None of the datasets used in this study, including the three external benchmarks, was collected through a deployed acoustic trap operating continuously in an uncontrolled field environment. Real-world monitoring introduces additional challenges that curated datasets only partially capture, such as wind noise, sounds from co-occurring insect species, variable microphone hardware and placement, and natural mosquito flight behavior. Field trials with prototype smart traps, evaluated over extended periods and multiple sites, are therefore necessary before the practical robustness of the proposed framework can be considered validated. End-to-end latency, memory use, and energy consumption on target hardware were not measured; real-time embedded feasibility therefore remains unestablished.

Finally, the present study focused exclusively on species classification from isolated, single-insect wingbeat recordings. Future work should extend the framework to the simultaneous detection of multiple insects, complete the ablation and computational-cost analyses outlined above, and assess deployment feasibility through target-hardware benchmarks and field trials under real environmental conditions.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

REFERENCES

[1] World Health Organization, "Vector-borne diseases," *WHO Fact Sheet*, Apr. 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

[2] F. Rustam, A. A. Reshi, W. Aljedaani, A. Alhossan, A. Ishaq, S. Shafi, E. Lee, Z. Alrabiah, H. Alsuwailem, A. Ahmad, and V. Rupapara, "Vector mosquito image classification using novel RIFS feature selection and machine learning models for disease epidemiology," *Saudi Journal of Biological Sciences*, vol. 29, no. 1, pp. 583–594, Jan. 2022.

[3] A. Goodwin, S. Padmanabhan, S. Hira, M. Glancey, M. Slinowsky, R. Immidisetti, L. Scavo, J. Brey, B. M. M. S. Sudhakar, T. Ford, C. Heier, Y.-M. Linton, D. B. Pecor, L. Caicedo-Quiroga, and S. Acharya, "Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection," *Scientific Reports*, vol. 11, p. 13656, Jul. 2021.

[4] M. Z. Nazari et al., "A novel deep learning approach for mosquito species classification via a dual-head structure and calibration-aware fusion architecture." *Scientific Reports*, vol 16, n. 1, pp. 7208, Feb. 2026.

[5] I. Potamitis, "Classifying insects on the fly," *Ecological Informatics*, vol. 21, pp. 40–49, May 2014.

[6] E. Fanioudakis, M. Geismar, and I. Potamitis, "Mosquito wingbeat analysis and classification using deep learning," in *Proc. 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2410–2414.

[7] I. Kiskin, D. Cobb, L. Wang, and S. Roberts, "HumBug Zooniverse: a crowd-sourced acoustic mosquito dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 916–920.

[8] E. Joelianto et al., "Convolutional neural network-based real-time mosquito genus identification using wingbeat frequency: A binary and multiclass classification approach," *Ecological Informatics*, vol 80, pp. 102495, 2024.

[9] K. O. Paim et al. "Acoustic identification of Ae. aegypti mosquitoes using smartphone apps and residual convolutional neural networks." *Biomedical Signal Processing and Control*, vol 95, pp. 106342, 2024.

[10] M. S. Yin, P. Haddawy, B. Nirandmongkol, T. Kongthaworn, C. Chaisumritchoke, A. Supratak, C. Sa-ngamuang, and P. Sriwichai, "A lightweight deep learning approach to mosquito classification from wingbeat sounds," in *Proc. Conference on Information Technology for Social Good (GoodIT)*, Rome, Italy, Sep. 2021, pp. 37–42.

[11] M. S. Yin et al. "A deep learning-based pipeline for mosquito detection and classification from wingbeat sounds." *Multimedia Tools and Applications*, vol 82, n. 4, pp. 5189–5205, 2023.

[12] J. A. Luna-González, D. Robles-Camarillo, M. Nakano-Miyatake, H. Lanz-Mendoza, and H. Pérez-Meana, "A CNN-based mosquito classification using image transformation of wingbeat features," in *Proc. 19th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT)*, vol. 327, IOS Press, 2020, p. 127.

[13] M. S. Yin, P. Haddawy, T. Ziemer, F. Wetjen, A. Supratak, K. Chiamsakul, P. Sriwichai, and B. Tongkumchum, "A deep learning-based pipeline for mosquito detection and classification from wingbeat sounds," *Multimedia Tools and Applications*, vol. 82, pp. 5189–5205, 2023.

[14] M. S. Fernandes, W. Cordeiro, and M. Recamonde-Mendoza, "Detecting *Aedes aegypti* mosquitoes through audio classification with convolutional neural networks," *Computers in Biology and Medicine*, vol. 129, p. 104152, Feb. 2021.

[15] I. Kiskin, Z. Sinka, D. Cobb, W. Rafique, W. Wang, L. Zilli, M. Sinka, K. Willis, and S. Roberts, "HumBugDB: a large-scale acoustic mosquito dataset," in *Proc. 35th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

[16] A. Supratak, C. Sa-ngamuang, K. Kaewchada, and P. Haddawy, "MosquitoSong+: a noise-robust deep learning model for mosquito classification from wingbeat sounds," *PLOS ONE*, vol. 19, no. 10, p. e0310121, Oct. 2024.

[17] J. R. Schreiber, E. Bueno, D. Motta, J. Nunes, O. G. V. Ribeiro-Filho, and M. A. Valdenegro-Toro, "Acoustic identification of *Ae. aegypti* mosquitoes using smartphone apps and residual convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 92, p. 106072, 2024.

[18] L. Pham et al., "A comprehensive survey with critical analysis for deepfake speech detection." *Computer Science Review*, vol. 57, pp. 100757, Aug. 2025.

[19] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[20] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, p. 101084, May 2020.

[21] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct. 2019, pp. 1314–1324.

[22] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: practical guidelines for efficient CNN architecture design," in *Proc. European Conference on Computer Vision (ECCV)*, Munich, Germany, Sep. 2018, pp. 116–131.

[23] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetV2: enhance cheap operation with long-range attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 9969–9982, 2022.

[24] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proc. 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 6105–6114.

[25] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "MobileOne: an improved one millisecond mobile backbone," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, Jun. 2023, pp. 7907–7917.

[26] D. Qin, C. Leichner, A. Guliani, S. Bhosale, S. Polosukhin, M. Liang, A. Dhandhania, M. Tan, A. Howard, and E. Real, "MobileNetV4: universal models for the mobile ecosystem," in *Proc. European Conference on Computer Vision (ECCV)*, Milan, Italy, 2024.

[27] A. Supratak et al., "Tiny MosquitoSong Dataset," *MosquitoSong+ code and data repository*, 2024. [Online]. Available: `github.com/akaraspt/mosquitosongp`.

[28] X. Wei, M. Z. Hossain, and K. A. Ahmed, "A ResNet attention model for classifying mosquitoes from wing-beating sounds," *Scientific Reports*, vol. 12, p. 10334, 2022, doi: 10.1038/s41598-022-14372-x.

Estimating Model Observer Response in Digital Mammography Using Radiomics and Machine Learning Regression Models

1st Fabiana Toon de Araújo
Dept. of Electrical and Computer Engineering
University of São Paulo, São Carlos School of Engineering
São Carlos, Brazil
fabianaaraujo@usp.br

2nd Renann de Faria Brandão
Dept. of Electrical and Computer Engineering
University of São Paulo, São Carlos School of Engineering
São Carlos, Brazil
renannbrandao@usp.br

3rd Dyego Alves Pereira
Dept. of Electrical and Computer Engineering
University of São Paulo, São Carlos School of Engineering
São Carlos, Brazil
dyego.pereira@usp.br

4th Predrag R. Bakic
Dept. of Translational Medicine, Medical Radiation Physics
Lund University, Faculty of Medicine
Malmö, Sweden
predrag.bakic@med.lu.se

5th Anders Tingberg
Dept. of Translational Medicine, Medical Radiation Physics
Lund University, Faculty of Medicine
Malmö, Sweden
anders.tingberg@med.lu.se

6th Marcelo Andrade da Costa Vieira
Dept. of Electrical and Computer Engineering
University of São Paulo, São Carlos School of Engineering
São Carlos, Brazil
mvieira@sc.usp.br

***Abstract*—Task-based image quality assessment plays an important role in digital mammography, as conventional objective image quality metrics do not necessarily reflect lesion detectability. Model Observers, such as the Channelized Hotelling Observer (CHO), provide a practical alternative to human reader studies by quantitatively assessing signal detectability in medical images. This study proposes a machine learning approach to estimate the CHO response from quantitative image descriptors and acquisition parameters. Four regression algorithms (XGBoost, Random Forest, KNN, and SVR) were trained on radiomic features extracted with the PyRadiomics library, combined with microcalcification size and exposure level (mAs). Among the evaluated models, tree-based ensemble methods substantially outperformed instance- and kernel-based approaches. XGBoost achieved the best overall performance ($R^2 = 0.922$, NRMSE $= 5.1\%$), followed by Random Forest ($R^2 = 0.864$, NRMSE $= 6.8\%$), whereas KNN and SVR exhibited limited generalization capacity ($R^2 < 0.47$). Analysis across experimental conditions indicates that predictive performance was independent of physical acquisition parameters, as lower doses and smaller microcalcifications did not systematically degrade prediction accuracy. Feature importance analysis suggests that first-order radiomic descriptors (skewness and kurtosis) were the primary predictors but the physical acquisition parameters contributed with important information, especially to the tree-based methods. These findings suggest that radiomic features combined with acquisition-related physical parameters can be used to estimate model observer responses, indicating that image characteristics and acquisition conditions contain information related to signal detectability.**

***Index Terms*—Digital mammography, image quality assessment, Channelized Hotelling Observer, radiomics, machine learning models.**

I. Introduction

Breast cancer is one of the leading causes of cancer-related death among women, and early detection is essential for improving the chances of successful treatment [1]. Globally, approximately one in every 20 women will be diagnosed with breast cancer during their lifetime, and it is estimated that by 2050, 3.2 million new cases will be diagnosed, resulting in approximately 1.1 million breast cancer-related deaths [2]. Medical imaging plays a crucial role in the early detection of breast cancer, enabling timely and appropriate treatment and thereby significantly reducing breast cancer mortality [3], [4].

Digital Mammography (DM) is currently the standard imaging modality for detecting masses, microcalcifications, and other breast abnormalities. In this technique, a two-dimensional image of the compressed breast is acquired using X-rays [5]. X-ray imaging examinations inevitably involve exposure to ionizing radiation, and the associated risks must be carefully considered. Therefore, clinical practice seeks to balance image quality and radiation dose in order to minimize

This work was supported by the São Paulo Research Foundation (FAPESP; Grants No. 2021/12673-6 and 2026/18409-2), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; Finance Code 001), and the Brazilian National Council for Scientific and Technological Development (CNPq; Grant No. 303251/2025-5).

patient risk while ensuring that the acquired image is adequate for its intended diagnostic task [6], [7].

Image quality plays a critical role in mammography, as both the presence of diagnostically relevant information in the image and the radiologist's ability to interpret it are essential for an accurate diagnosis. Studies have shown that image noise directly affects the detectability of microcalcifications and the discrimination of breast masses [8], [9].

Therefore, assessing the quality of a medical image involves evaluating the observer's ability to interpret the image and detect lesions. However, many studies rely solely on objective image quality metrics based on the images visual characteristics. The limitation of this approach is that such metrics do not always capture image features that are relevant to lesion detection by human observers, whose performance ultimately determines the diagnostic effectiveness of image restoration techniques [10].

Another approach to assessing the diagnostic quality of medical images is through clinical reader studies, in which expert radiologists evaluate clinical images for the presence or absence of lesions [11], [12]. Depending on the study design, readers may also identify lesion locations and assign confidence scores to their detection decisions. However, this type of evaluation has several limitations, including the logistical complexity of conducting reader studies. Such studies require the availability of experienced readers, dedicated image-viewing workstations, and sufficiently large clinical image datasets, while also being time-consuming, particularly when a large number of images must be evaluated.

To address these limitations, model observers have been developed to simulate human observer performance using mathematical and statistical methods based on signal and noise properties [13], [14]. Their use provides a practical approach to image quality assessment while reducing the reliance on human observers.

In this study, we propose estimating the response of a model observer using four regression models trained on quantitative image descriptors and imaging parameters, including radiomic features, microcalcification size, and the product of tube current and exposure time (mAs). The objective is to investigate whether these image descriptors and acquisition-related parameters contain sufficient information to reproduce the model observer response, thereby evaluating their potential as a quantitative representation of image quality for diagnostic tasks.

II. Materials and Methods

A. Dataset

The dataset used in this study was developed from images acquired at the Breast Clinic of Skåne University Hospital, Malmö, Sweden, using a Mammomat Inspiration full-field digital mammography (FFDM) system (Siemens Healthineers, Forchheim, Germany). The images were obtained from an anthropomorphic breast phantom (PhantomX GmbH, Berlin, Germany) [18].

Images were acquired at five tube current–time products: 90, 110, 125, 200, and 250 mAs. The 125 mAs setting was the standard exposure automatically selected by the Automatic Exposure Control (AEC) system.

To generate images containing microcalcifications of different sizes, a uniform polymethyl methacrylate (PMMA) phantom with an insert sheet containing microcalcification patterns was used. The microcalcification diameters ranged from 150 to 310 μm, with increments of 40 μm. Ten images were acquired at 200% of the AEC exposure using the PMMA phantom with the microcalcification insert sheet.

Subsequently, the microcalcifications were cropped and artificially inserted into the mammographic images using a method developed by our research group [19], [20]. The method identifies dense breast regions using the software package developed by the Laboratory for Breast Radiodensity Assessment [21] and inserts the microcalcification cluster into the selected region. Each insertion was performed within a 150 × 150-pixel region of interest, at different locations in the breast image, to provide variability in background tissue and texture for the detection task. An example of the dataset is shown in Fig. 1.

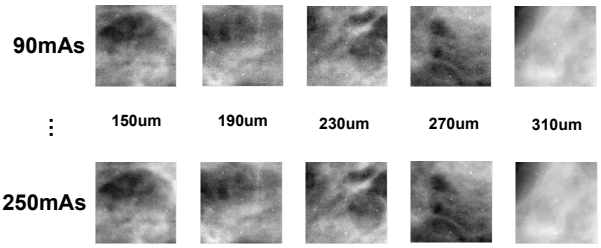


Fig. 1: Representative image patches illustrating the experimental conditions considered in this study. Images are shown for five microcalcification sizes (150–310 μm) and five exposure levels (90–250 mAs).

B. Model Observer

Model Observer is a mathematical and statistical approach based on signal and noise properties that aim to simulate the performance of human observers in diagnostic tasks. Rather than replacing clinical reader studies, they are intended to efficiently identify and address potential issues before investing the time and resources required for evaluations involving human observers [22].

One of the most widely used model observers for medical imaging and detectability assessment is the Channelized Hotelling Observer (CHO). Originally proposed by Kyle J. Myers and Harrison H. Barrett in the 1980s [23], the CHO processes input images through a set of frequency-domain channels designed to model the response of the human visual system. These channels extract task-relevant information from the image and may be implemented using Gabor filters, Difference-of-Gaussians (DoG) channels, Laguerre–Gauss (LG) channels, among others.

Initially, the input signal is processed by the channels, and an internal noise term, is added to the output of each channel to simulate the perceptual inefficiency of the human visual system. The channel outputs are then multiplied by the transposed

CHO template to obtain the decision variable. The template is computed as the product of the inverse covariance matrix of the channel output vectors and the difference between the mean channel responses for the two classes.

The CHO template depends on the channel type employed. The choice of channel is determined by the detection task and influences both the realism and the performance of the model observer. For example, Gabor channels are commonly used in tasks where orientation is an important feature, making them suitable for images containing directional structures. In contrast, Laguerre–Gauss (LG) channels are typically employed to detect radially symmetric signals, such as circular objects. Difference-of-Gaussians (DoG) channels, like LG channels, are rotationally symmetric but are more sensitive to edges and image contrast [25].

The CHO implementation adopted in this study was provided by the VICTRE framework, which employs Laguerre–Gauss (LG) channels [26]. LG channels are particularly suitable for detecting radially symmetric signals, such as microcalcifications, making them appropriate for the detection task investigated in this work.

The CHO is calibrated using a set of training images containing both signal-present and signal-absent cases. From these images, it estimates a template that characterizes the microcalcification cluster in the channelized feature space. During testing, each image is projected onto this template to produce a decision variable (or t-score), which quantifies the evidence for signal presence. Higher t-scores indicate a greater likelihood that the image contains a microcalcification cluster, whereas lower t-scores indicate little or no evidence of signal presence.

C. Radiomic Features

Radiomic features are quantitative metrics designed to describe and characterize medical images properties, transforming medical image analysis from a qualitative assessment into a data-driven quantitative approach [27]. These features provide objective measurements of intensity distribution, spatial relationships between different intensity levels, texture heterogeneity, shape descriptors, and other image properties, allowing subtle image characteristics to be numerically represented, thereby transforming visual information into structured, analyzable data [28].

Radiomic features were extracted using the PyRadiomics library. PyRadiomics is an open-source python package for the extraction of Radiomics features from medical imaging. The extracted radiomic features comprised first-order statistics and texture descriptors derived from gray-level matrices, including the Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), Gray-Level Dependence Matrix (GLDM), and Neighboring Gray Tone Difference Matrix (NGTDM) [29]. Together, these features quantify different aspects of image appearance, such as intensity distribution, local texture, spatial organization, and structural heterogeneity.

To reduce feature redundancy and retain only informative radiomic descriptors, a two-step feature selection procedure was applied. First, features with a variance below 0.01 were removed using the VarianceThreshold method, excluding near-constant features. Then, the Pearson correlation coefficient was calculated among the remaining features using the training set only. When two features exhibited an absolute correlation coefficient greater than 0.95, one of the correlated features was removed to reduce redundancy. The same selected feature set was then applied to the test set without recalculating the selection criteria. The resulting radiomic features were combined with microcalcification size and mAs as additional predictors for the regression models, and the complete set of features is listed in Table II.

D. Regression Models

To estimate the Channelized Hotelling Observer (CHO) response, four supervised regression algorithms were evaluated: Random Forest (RF), Support Vector Regressor (SVR), K-Nearest Neighbors (KNN) and Extreme Gradient boosting (XGBoost). These methods were selected because they represent different regression strategies, such as ensemble tree-based learning, kernel-based regression, instance-based learning anda gradient boosting, respectively.

Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to improve predictive performance and reduce overfitting [30]. Support Vector Regression estimates nonlinear relationships by mapping the input data into a high-dimensional feature space through kernel functions, enabling non-linear relationships to be modeled as linear in the transformed feature space [31]. K-Nearest Neighbors predicts the target value based on the responses of the most similar training samples according to a distance metric [32]. XGBoost is a gradient boosting algorithm that sequentially builds decision trees to minimize prediction error, providing high predictive accuracy and robustness for structured tabular data [33].

To optimize the predictive performance of each regression model, hyperparameter tuning was performed using an exhaustive grid search combined with five-fold cross-validation on the training set. For each algorithm, multiple combinations of hyperparameters were evaluated, and the configuration yielding the highest mean coefficient of determination (R^2) across the validation folds was selected. The optimized models were subsequently retrained using the complete training set and evaluated on the independent test set.

For all models, the input variables consisted of radiomic features extracted from the images, together with acquisition-related parameters, including microcalcification size and mAs. The target variable was the CHO t-score computed for each image. Model performance was evaluated by comparing the predicted and reference CHO responses using the following regression metrics: coefficient of determination (R^2), root mean squared error (RMSE), normalized RMSE (NRMSE) and mean absolute error (MAE).

E. Pipeline

First, the model observer was applied to the dataset to compute the detectability of each image, producing a corresponding t-score that served as the reference response. Next, radiomic features were extracted from each image using the PyRadiomics library, and the most informative features were selected through a feature selection algorithm. The selected radiomic features, together with the microcalcification size and exposure level, were used as input variables for the regression models, while the model observer t-score was used as the target variable. The hyperparameters of the regression models were optimized using Grid Search with five-fold cross-validation on the training set. Finally, the optimized models were evaluated on an independent test set by comparing the predicted and reference t-score values. Figure 2 illustrates the pipeline adopted in this study.

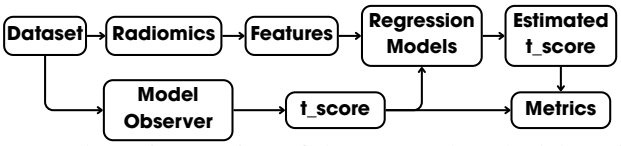


Fig. 2: Schematic overview of the proposed methodology for estimating the CHO response.

III. Results and Discussion

Table I summarizes the quantitative performance of the evaluated regression models.

TABLE I: Performance metrics of the evaluated regression models for CHO t-score prediction.

Models	R^2	NRMSE	RMSE	MAE
XGBoost	0.922	0.051	7.12	4.89
Random Forest	0.864	0.068	9.43	6.49
KNN	0.471	0.163	18.57	12.31
SVR	0.323	0.152	21.012	15.57

Among all approaches, XGBoost achieved the best predictive performance, with the highest coefficient of determination ($R^2 = 0.922$), indicating that approximately 92% of the variability in the observer t-score was explained by the input variables. It also yielded the lowest prediction errors, with an RMSE of 7.12, an NRMSE of 5.1%, and an MAE of 4.89, demonstrating its superior ability to estimate the model observer response.

Random Forest also demonstrated strong predictive performance, achieving an R^2 of 0.864 and relatively low error metrics, although it was outperformed by XGBoost. KNN and SVR models exhibited substantially lower predictive accuracy, with R^2 values of 0.471 and 0.323, respectively, together with considerably higher RMSE and MAE values. These results suggest that tree-based ensemble methods were more effective in modeling the relationship between the selected radiomic features and the model observer response than the instance-based and kernel-based regression approaches.

Figure 3 compares the parity plots obtained for the evaluated

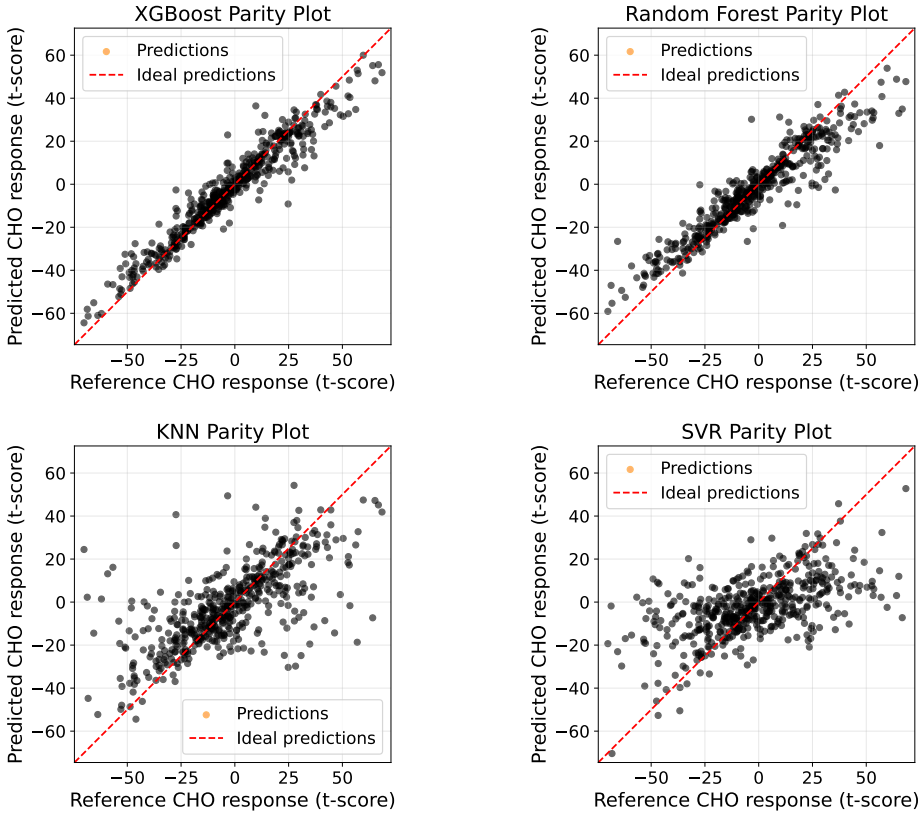


Fig. 3: Parity plots comparing the predicted and reference t-score values for the evaluated regression models: (a) XGBoost, (b) Random Forest, (c) K-Nearest Neighbors (KNN), and (d) Support Vector Regression (SVR). The dashed red line represents the ideal agreement between predicted and reference values.

regression models. XGBoost (Fig.3(a)) exhibited the highest agreement between the predicted and reference t-score values, with most samples closely distributed around the identity line. Random Forest (Fig.3(b)) also demonstrated good predictive performance, although a slightly larger dispersion was observed, particularly for samples with very high and very low t-score values.

KNN (Fig. 3(c)) and SVR (Fig. 3(d)) presented greater deviations from the identity line. For both models, the predictions were concentrated around the central range of the t-score distribution, indicating limited ability to reproduce extreme values. This behavior resulted in the underestimation of high t-score values and the overestimation of low t-score values, which is consistent with their lower coefficients of determination and higher prediction errors reported in Table I.

Despite the superior performance of XGBoost and Random Forest, all models exhibited a slight tendency to overestimate low t-score values and underestimate high t-score values, indicating reduced prediction accuracy under conditions associated with both low and high detectability. This behavior may be partially attributed to the lower number of samples available at the extremes of the t-score distribution, which can limit the ability of the regression models to predict these less represented conditions.

Figure 4 shows the local R^2 values across all experimental conditions, evaluating whether model performance is systematically influenced by physical acquisition parameters. Overall, predictive performance showed no consistent dependence on either exposure level (mAs) or microcalcification size. Lower radiation doses and smaller microcalcifications did not systematically degrade R^2 values, suggesting that these physical factors alone are not determinants of prediction accuracy. However, the results across the experimental matrix varied considerably by algorithm. Tree-based ensemble models (XGBoost and Random Forest) exhibited higher and more consistent R^2 scores across combinations, indicating that these models can maintain robust predictions regardless of noise level or signal scale. KNN and SVR displayed lower overall

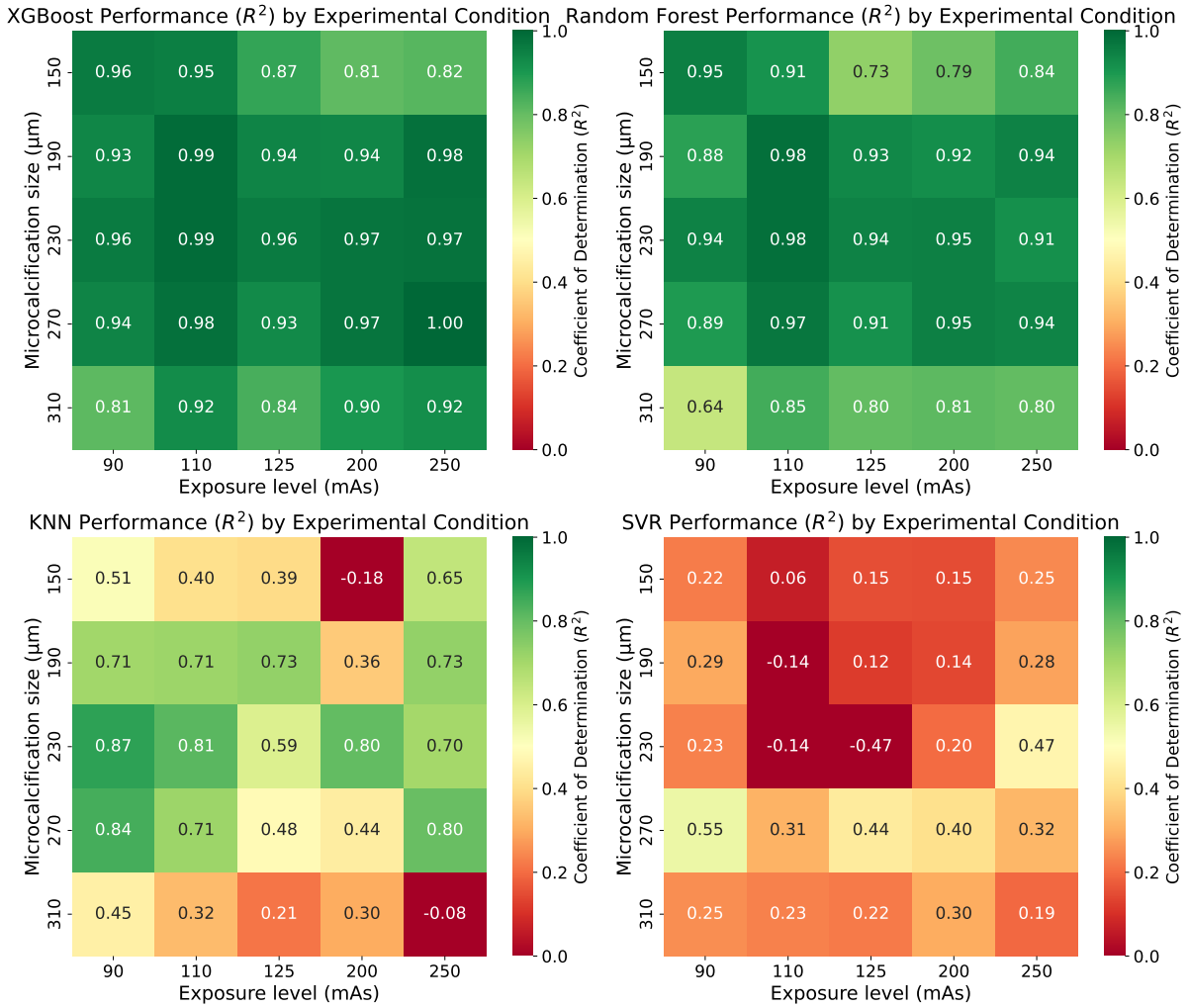


Fig. 4: Heatmaps of the coefficient of determination (R^2) obtained for (a) XGBoost, (b) Random Forest, (c) K-Nearest Neighbors (KNN), and (d) Support Vector Regression (SVR) as a function of exposure level (mAs) and microcalcification size. Higher values indicate higher prediction accuracy.

R^2 values and pronounced local fluctuations across conditions. Rather than a systematic physical trend, this variability reflects the limited generalization capacity of these two algorithms for this task.

The superior performance of Random Forest and XGBoost can be attributed to their ability to model complex nonlinear relationships between the selected features and the CHO t-score while naturally capturing interactions between radiomic features and acquisition-related physical parameters. These tree-based ensemble methods are particularly well suited to representing the heterogeneous characteristics of mammographic images, allowing them to produce more accurate predictions than KNN and SVR. In contrast, KNN relies on local similarity between samples and tends to produce predictions biased toward the average response when neighboring samples exhibit substantial variability. Similarly, although SVR can model nonlinear relationships through kernel functions, its performance is highly dependent on the choice of kernel and hyperparameters, which may limit its ability to accurately represent the complex relationship between these features and the CHO t-score.

Since Random Forest and XGBoost were the best-performing models and provided native feature importance measures, their feature importance rankings are shown in Table II, and the four highest values were highlighted. KNN and SVR do not provide directly interpretable feature importance measures and therefore were not included in this analysis.

TABLE II: Feature importance scores obtained from the Random Forest and XGBoost models.

Feature Classes	Features	Feature Importance Values	
		XGBoost	Random Forest
firstorder	Skewness	**0.157**	**0.226**
firstorder	Kurtosis	**0.132**	**0.218**
glrlm	LRHGLE	**0.103**	**0.108**
-	mAs	**0.095**	0.008
glrlm	LRLGLE	0.061	0.038
glrlm	GLNU	0.056	0.051
firstorder	Variance	0.053	0.016
firstorder	InterquartileRange	0.050	0.030
glcm	ClusterProminence	0.048	0.016
glcm	Autocorrelation	0.044	0.027
firstorder	Entropy	0.035	0.016
glrlm	LongRunEmphasis	0.034	0.026
ngtdm	Busyness	0.033	0.024
firstorder	10Percentile	0.030	0.034
-	MC_size	0.025	**0.109**
gldm	DN	0.023	0.033
firstorder	Range	0.021	0.017

In particular, Skewness and Kurtosis received the highest importance scores in both models. These features describe the shape of the gray-level intensity distribution, suggesting that the asymmetry and peakedness of the image histogram are strongly associated with the CHO t-score. Also, the Long Run High Gray Level Emphasis indicates that the presence of large, homogeneous, and bright areas in the image is a critical factor in the algorithm's decision. The inclusion of microcalcification size and mAs among the most relevant predictors indicates that the physical parameters also contribute to the prediction, as these factors directly influence detectability.

To assess the contribution of acquisition- and signal-related parameters, an additional experiment was performed using only radiomic features as predictors. The local R^2 values across the same experimental conditions are shown in Figure 5.

The exclusion of mAs and microcalcification size reduced the predictive performance of Random Forest and XGBoost, particularly for the smaller microcalcifications, suggesting that the acquisition and signal-related parameters provide relevant complementary information for predicting the CHO response under these conditions. In contrast, KNN showed an overall improvement in predictive performance when these parameters were excluded, with the most pronounced improvement observed for larger microcalcification sizes. SVR maintained a relatively poor predictive performance and showed a further reduction in R^2 across most experimental conditions when the physical parameters were excluded.

Overall performance metrics for this experiment are shown in table III.

TABLE III: Performance metrics of the evaluated regression models for CHO t-score prediction without physical parameters.

Models	R^2	NRMSE	RMSE	MAE
XGBoost	0.756	0.091	12.62	9.67
Random Forest	0.769	0.089	12.27	9.50
KNN	0.745	0.093	12.91	10.09
SVR	0.258	0.159	22.01	16.89

For Random Forest, the R^2 decreased from 0.864 to 0.769, while for XGBoost it decreased from 0.922 to 0.756. In contrast, KNN exhibited an improvement when only radiomic features were used, with the R^2 increasing from 0.471 to 0.745, while the RMSE decreased from 18.57 to 12.91. This different behavior may be related to the sensitivity of distance-based methods to the predictor space. The inclusion of additional acquisition-related variables changes the geometry of the feature space and may affect the distances between samples used by KNN, even when the variables are appropriately scaled.

SVR showed a decrease in performance when the acquisition parameters were excluded, from $R^2 = 0.323$ to $R^2 = 0.258$. Overall, these results indicate that radiomic features alone contain substantial information related to the CHO response, while the contribution of acquisition-related parameters depends on the regression model used.

IV. Conclusion

This study investigated whether the response of the Channelized Hotelling Observer (CHO) can be estimated using machine learning regression models based on radiomic features together with acquisition-related parameters. Among the evaluated approaches, XGBoost achieved the best predictive performance, with an R^2 of 0.922 and an NRMSE of 5.1%, followed by Random Forest with an R^2 of 0.864 and an

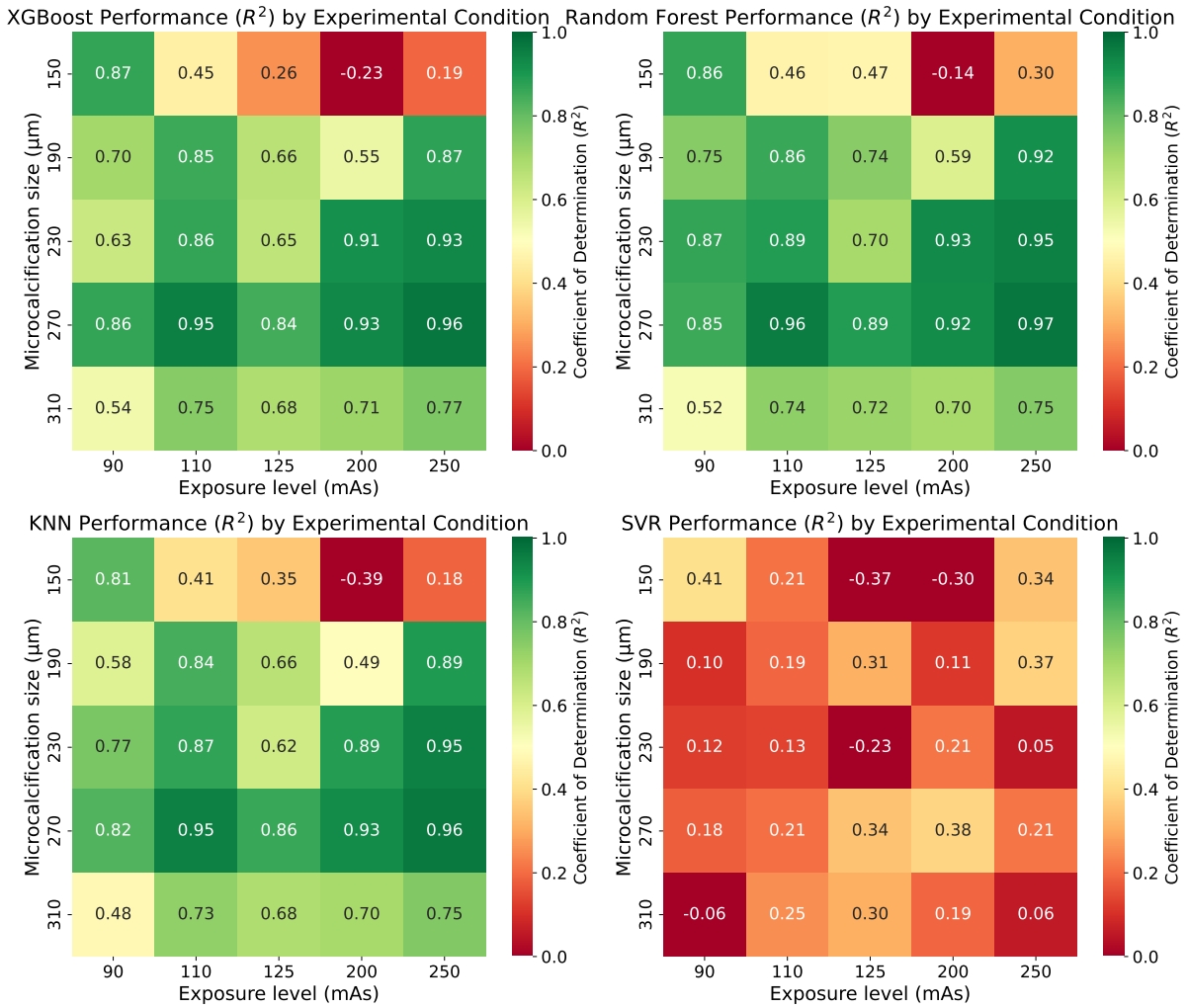


Fig. 5: Heatmaps of the coefficient of determination (R^2) obtained for (a) XGBoost, (b) Random Forest, (c) K-Nearest Neighbors (KNN), and (d) Support Vector Regression (SVR) across different microcalcification sizes and exposure levels (mAs), trained using radiomic features only. Higher values indicate higher prediction accuracy.

NRMSE of 6.8%. The analysis using radiomic features alone showed that these descriptors contain substantial information related to the CHO response, although the inclusion of microcalcification size and mAs further improved the performance of the tree-based models. Feature importance analysis indicated that first-order radiomic descriptors, particularly skewness and kurtosis, were among the most relevant predictors, suggesting that radiomics can capture image characteristics associated with observer detectability. Error analysis also revealed a tendency to overestimate low t-scores and underestimate high t-scores, indicating a regression-to-the-mean behavior across the evaluated models. Future work should focus on improving prediction accuracy at the extremes of the t-score distribution by acquiring additional samples under low- and high-detectability conditions and including greater variation in background complexity, such as denser and less dense breast regions.

REFERENCES

[1] ŁUKASIEWICZ, Sergiusz et al. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. Cancers, v. 13, n. 17, p. 4287, 2021.

[2] KIM, Joanne et al. Global patterns and trends in breast cancer incidence and mortality across 185 countries. Nature medicine, v. 31, n. 4, p. 1154-1162, 2025.

[3] TABáR, L. et al. Early detection of breast cancer rectifies inequality of breast cancer outcomes. Journal of Medical Screening, v. 28, n. 1, p. 34–38, 2021. PMID: 32370610. Available in: https://doi.org/10.1177/0969141320921210.

[4] EBY, P. R.; GHATE, S.; HOOLEY, R. The benefits of early detection: Evidence from modern international mammography service screening programs. Journal of Breast Imaging, v. 4, n. 4, p. 346–356, 06 2022. ISSN 2631-6110. Available in: https://doi.org/10.1093/jbi/wbac041.

[5] FICO, Noemi et al. Breast imaging physics in mammography (Part II). Diagnostics, v. 13, n. 23, p. 3582, 2023.

[6] HOOSHMAND, Sahand et al. A review of screening mammography: the benefits and radiation risks put into perspective. Journal of Medical Imaging and Radiation Sciences, v. 53, n. 1, p. 147-158, 2022.

[7] NOOR, Kaltham Abdulwahid Mohammad et al. Exploring the Association between Mammography Dose Optimization and Image Quality. Journal of Medical Imaging and Radiation Sciences, v. 55, n. 3, p. 101572, 2024.

[8] BOITA, Joana et al. How does image quality affect radiologists' perceived ability for image interpretation and lesion detection in digital mammography?. 2021.

[9] SAUNDERS JR, Robert S. et al. Does image quality matter? Impact of resolution and noise on mammographic task performance. Medical physics, v. 34, n. 10, p. 3971-3981, 2007.

[10] SOARES, Lucas Exposto et al. Predicting the quality of digital mammography from the perspective of detectability of microcalcifications: A Radiomics approach. Medical Physics, v. 52, p. 1-15, 2025.

[11] TIMBERG, Pontus et al. Visibility of microcalcification clusters and masses in breast tomosynthesis image volumes and digital mammography: a 4AFC human observer study. Medical physics, v. 39, n. 5, p. 2431-2437, 2012.

[12] BORGES, Lucas R.; DE AZEVEDO MARQUES, Paulo M.; VIEIRA, Marcelo AC. A 2-AFC study to validate artificially inserted microcalcification clusters in digital mammography. In: Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment. SPIE, 2019. p. 178-184.

[13] BOUWMAN, Ramona W. et al. Toward image quality assessment in mammography using model observers: Detection of a calcification-like object. Medical physics, v. 44, n. 11, p. 5726-5739, 2017.

[14] BRANDÃO, Renann F. et al. Exploring the impact of image restoration in simulating higher dose mammography: effects on the detectability of microcalcifications across different sizes using model observer analysis. Journal of Medical Imaging, v. 12, n. S2, p. S22013-S22013, 2025.

[15] HE, Xin; PARK, Subok. Model observers in medical imaging research. Theranostics, v. 3, n. 10, p. 774, 2013.

[16] ABADI, E. et al. Virtual clinical trials in medical imaging: a review. Journal of Medical Imaging, Society of Photo-Optical Instrumentation Engineers, v. 7, n. 4, p. 042805–042805, 2020.

[17] SHARMA, D. et al. In silico imaging tools from the victre clinical trial. Medical physics, Wiley Online Library, v. 46, n. 9, p. 3924–3928, 2019.

[18] L. C. Ikejimba et al., "A novel physical anthropomorphic breast phantom for 2D and 3D X-ray imaging," Med. Phys. 44(2), 407–416 (2017).

[19] L. R. Borges et al., "Effect of denoising on the localization of microcalcification clusters in digital mammography," Proc. SPIE 11513, 115130K (2020).

[20] L. R. Borges, P. M. de Azevedo Marques, and M. A. Vieira, "A 2-AFC study to validate artificially inserted microcalcification clusters in digital mammography," Proc. SPIE 10952, 109520R (2019).

[21] B. M. Keller et al., "Preliminary evaluation of the publicly available Laboratory for Breast Radiodensity Assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case-control study with digital mammography," Breast Cancer Res. 17, 1–17 (2015).

[22] KRUPINSKI, E. A. The important role of task-based model observers and related techniques in medical imaging. Journal of Nuclear Cardiology, Springer, v. 28, n. 2, p. 638–640, 2021.

[23] MYERS, K. J.; BARRETT, H. H. Addition of a channel mechanism to the ideal-observer model. Journal of the Optical Society of America A, Optical Society of America, v. 4, n. 12, p. 2447–2457, 1987.

[24] PIMENTA, E. B.; COSTA, P. R. Model observers and detectability index in x-ray imaging: historical review, applications and future trends. Physics in Medicine & Biology, IOP Publishing, v. 70, n. 7, p. 07TR02, 2025.

[25] PETROV, D. et al. Systematic approach to a channelized hotelling model observer implementation for a physical phantom containing mass-like lesions: Application to digital breast tomosynthesis. Physica Medica, Elsevier, v. 58, p. 8–20, 2019.

[26] SHARMA, Diksha et al. In silico imaging tools from the VICTRE clinical trial. Medical physics, v. 46, n. 9, p. 3924-3928, 2019.

[27] Lambin, P., Rios-Velazquez, E., Leijenaar, R., et al. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer, 48(4), 441–446, 2012.

[28] FLAIBAN, Everton et al. Radiomics in Action: Multimodal Synergies for Imaging Biomarkers. Bioengineering, v. 12, n. 11, p. 1139, 2025.

[29] Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107. 'https://doi.org/10.1158/0008-5472.CAN-17-0339 ¡https://doi.org/10.1158/0008-5472.CAN-17-0339¿'

[30] Leo Breiman (2001). Random Forests. Machine Learning, 45(1), 5–32.

[31] B. Üstün, W.J. Melssen, L.M.C. Buydens, Visualisation and interpretation of Support Vector Regression models, Analytica Chimica Acta, Volume 595, Issues 1–2, 2007, Pages 299-309, ISSN 0003-2670, https://doi.org/10.1016/j.aca.2007.03.023.

[32] Yuanming Shi, Kai Yang, Zhanpeng Yang, Yong Zhou, Chapter Two - Primer on artificial intelligence, Mobile Edge Artificial Intelligence, Academic Press, 2022, Pages 7-36, ISBN 9780128238172, https://doi.org/10.1016/B978-0-12-823817-2.00011-5.

[33] CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

Evaluating CLAP Audio Foundation Model for Swallowing Event Detection

Victor do Nascimento Gomes
Department of Computing and Automation
Federal University of Rio Grande Do Norte (UFRN)
Natal/RN, Brazil
victor.gomes.091@ufrn.edu.br

Ignacio Sánchez-Gendriz
Northeast Strategic Technologies Center (CETENE)
Ministry of Science, Technology, and Innovation
Recife/PE, Brazil

Ingridy M. P. Barbalho, Felipe Fernandes
Laboratory for Technological Innovation in Health (LAIS)
Federal University of Rio Grande Do Norte (UFRN)
Natal/RN, Brazil

Luiz Affonso Guedes
PPGEEC/CT
Federal University of Rio Grande Do Norte (UFRN)
Natal/RN, Brazil

Ricardo Valentim
Laboratory for Technological Innovation in Health (LAIS)
Federal University of Rio Grande Do Norte (UFRN)
Natal/RN, Brazil

***Abstract*—Dysphagia is a common complication in neurodegenerative diseases such as Amyotrophic Lateral Sclerosis and Parkinson's disease, requiring accessible diagnostic tools for telemedicine applications. Acoustic analysis of swallowing sounds using laryngophones offers a promising non-invasive approach, but current automatic detection methods remain below human-level agreement. This study investigates whether representations from audio foundation models pre-trained on general-domain data can improve swallowing event detection when transferred to this specialized domain. We propose lightweight adaptation strategies using frozen CLAP encoder weights, comparing zero-shot classification, linear probing, and a shallow MLP head against a classical energy detector baseline under a common leave-one-subject-out protocol. On 16 healthy subjects, zero-shot classification fails (MCC = 0.303), while both adapted heads reach MCC = 0.743–0.768 against 0.696 for the baseline. Under subject-clustered inference these differences in mean MCC are not significant at this sample size; what the data do resolve is dispersion, with adaptation roughly halving the inter-subject standard deviation from 0.13 to 0.07–0.08. The contribution is a fair subject-independent evaluation of transfer to a contact-sensor domain, and evidence that the benefit of adaptation lies in consistency across individuals rather than in mean performance. The validated scope is swallowing-event detection in healthy subjects with a single laryngophone setup; extension to dysphagia monitoring would require validation on patients, devices and sites.**

***Index Terms*—swallowing detection, CLAP, audio foundation models, laryngophone, transfer learning, linear probe**

I. Introduction

Neurodegenerative disorders such as Amyotrophic Lateral Sclerosis (ALS) and Parkinson's disease progressively impair motor functions, including the ability to swallow safely [1], [2]. Dysphagia affects a substantial proportion of these patients and carries risks of aspiration pneumonia and malnutrition. Traditional diagnostic methods require specialized equipment, limiting accessibility in remote settings. Acoustic analysis of swallowing sounds captured by laryngophones offers a non-invasive, low-cost alternative suited for telemedicine [3], [4].

Previous work established a band-limited energy detector for swallowing events, achieving MCC of 0.71 against manual annotations from three observers [5]. The gap between automatic detection and human agreement defines the improvement target. However, the baseline was evaluated in-sample, without cross-validation, meaning performance is optimistically biased.

Biomedical acoustic datasets are typically small due to the cost of clinical data collection and expert annotation. Training deep learning models from scratch on such low-resource datasets risks overfitting, motivating the use of transfer learning from pre-trained representations. Audio foundation models trained on large-scale datasets have shown strong transfer capabilities in diverse downstream tasks [6]–[8]. The CLAP model learns joint audio-text representations through contrastive learning, enabling zero-shot classification. Whether such representations transfer to laryngophone signals is unclear, given the domain mismatch between air microphones used in pre-training and contact sensors.

This work contributes a fair subject-independent evaluation of foundation model transfer to laryngophone signals, in four parts: (i) we re-evaluate a band-limited energy detector under leave-one-subject-out (LOSO) cross-validation, providing an unbiased performance estimate; (ii) we show that zero-shot classification with CLAP fails on laryngophone signals due to domain mismatch; (iii) we show that lightweight adaptation (linear probe or MLP head) with frozen encoders roughly halves the inter-subject standard deviation from 0.13 to 0.07–

0.08 under subject-clustered inference, improving consistency across unseen subjects, while the accompanying differences in mean MCC do not reach significance with 16 subjects; and (iv) we visualize the learned representations to characterize class separability and identify where errors concentrate in the embedding space.

II. Related Work

Acoustic analysis of swallowing has been investigated as a diagnostic and monitoring tool for dysphagia. Dudik et al. [3] reviewed the effects of dysphagia on swallowing sounds and vibrations, highlighting the potential of acoustic features for distinguishing healthy from impaired swallowing. Da Silva et al. [4] applied acoustic analysis to swallowing sounds in Parkinson's disease patients, demonstrating differences in spectral characteristics. Machine learning approaches to swallowing detection date back to Sazonov et al. [9], who used mel-scale features and SVMs for automatic detection of swallowing events in wearable devices. More recently, Khalifa et al. [10] applied deep learning to high-resolution cervical auscultation recordings, demonstrating the potential of CNNs for swallowing identification. The use of throat microphones, or laryngophones, offers advantages in noise rejection and signal quality for swallowing sound acquisition [11]. Sanchez-Gendriz et al. [5] developed a band-limited energy detector for swallowing events and established inter-observer agreement metrics using the Matthews Correlation Coefficient.

Foundation models for audio have emerged as powerful feature extractors for diverse tasks. General-purpose models like CLAP [6], [12], WavLM [7], and BEATs [8] learn representations from large-scale datasets and have been systematically evaluated through benchmarks such as HEAR [13]. In the health acoustics domain, Google's HeAR [14] was specifically trained on over 300 million health-related audio clips including coughs, breathing, and speech, demonstrating strong performance on respiratory disease classification. Deep learning has also been extensively applied to respiratory sounds using the ICBHI database [15]. However, most health acoustic foundation models target air-conducted sounds; transfer to contact sensors such as laryngophones remains unexplored. This work addresses this gap by evaluating whether general-purpose audio representations transfer to the distinct acoustic characteristics of contact-sensor recordings.

III. Materials and Methods

A. Dataset and Annotations

This study builds upon the dataset and annotation methodology established by Sanchez-Gendriz et al. [5]. Acoustic signals were collected from 16 healthy subjects (5 males and 11 females, median age 27.5 years) at the University Hospital Onofre Lopes (HUOL) at UFRN. The study was conducted following the guidelines approved by the Research Ethics Committee of UFRN (CAAE No. 25687819.3.0000.5537). A laryngophone, a contact microphone placed on the cervical region, was connected to a mobile phone for signal acquisition. Audio signals were recorded at 22.05 kHz with 32-bit resolution. Recording durations ranged from 14 to 33 seconds (median 26 s), capturing swallowing sounds along with physiological noise such as breathing.

The ground truth construction followed a rigorous multi-observer protocol. Three trained observers, who received specialized training in the detection of swallowing sounds, independently annotated each recording using Audacity software. Observers marked the start and end times of swallowing events based on both auditory playback and visual inspection of spectrograms. To assess reliability, each observer repeated the annotation process three times, yielding nine annotation sets per recording.

Annotations were converted to binary detection masks with temporal resolution of 50 ms per bin. For each observer, a bin was marked positive if at least two of three repetitions indicated a swallowing event (intra-observer aggregation). The final ground truth then took the union across observers: a bin is positive if any one of the three marked it. This is an inclusive criterion, chosen so that events seen by a single annotator are not discarded, and it means the reference is not uniformly certain: 60.3% of positive bins carry all three observers, 17.9% two and 21.8% only one. Section IV-E exploits this gradation to separate label ambiguity from detector error. The original study reported intra-observer MCC values of 0.84, 0.92, and 0.88. To establish a fair performance ceiling for our models, we compute each observer's MCC against the leave-one-out consensus of the other two, yielding MCC = 0.827 ± 0.048 (see Section IV).

The dataset exhibits severe class imbalance: prevalence of positive bins is approximately 9%. This makes accuracy misleading, motivating MCC as the primary metric [16].

B. Evaluation Protocol

All trained models follow leave-one-subject-out (LOSO) cross-validation. In each of 16 folds, one subject is held out for testing while 15 are used for training. This protocol applies to the energy detector baseline, linear probe, and MLP head. To prevent data leakage, every fitting and selection step within a fold uses training data only. The energy detector performs its grid search by pooling the 15 training subjects, and the held-out subject is never scored during selection. The linear probe fits the feature scaler on training embeddings alone and applies the resulting transformation unchanged to the test subject. For the MLP head, one of the 15 training subjects serves as an inner validation set, selected by a fixed rule in which fold k uses subject $(k+1) \bmod 16$. The model is trained on the remaining 14, and the checkpoint maximizing inner-validation MCC is selected; the validation subject is by construction disjoint from the test subject. Head architecture and all hyperparameters were fixed a priori and held identical across the 16 folds, and the decision threshold is fixed at 0.5 for both adapted models with no post-hoc adjustment. The split by subject ensures that no information from the test

subject influences model development, avoiding leakage from temporally correlated bins within the same recording.

The zero-shot experiment constitutes an exception: since no training occurs, LOSO is not applicable. Instead, we optimize the decision threshold directly on all data, providing an upper bound on zero-shot performance.

We report two complementary metrics. The primary metric is pooled MCC, computed by aggregating all predictions and ground truth labels across the 16 folds into a single confusion matrix, then calculating MCC from this pooled matrix. This provides a global measure of performance across the entire dataset. Additionally, we compute MCC separately for each subject and report the standard deviation across subjects, which characterizes how consistently a method performs on individuals it was not trained on.

C. Experiments

We conducted four experiments to systematically evaluate the potential of foundation model representations for swallowing detection.

1) Experiment 1: Baseline Reproduction: The first experiment reproduces the band-limited energy detector from [5] under proper LOSO validation. The detector computes Short-Time Fourier Transform (STFT) coefficients and extracts energy within a frequency band $[f_{\text{low}}, 2000]$ Hz. A moving average filter smooths the energy, and detection occurs when the smoothed energy exceeds a threshold $\theta = r_{\text{std}} \times \sigma(E_{\text{smoothed}})$.

Two hyperparameters are tuned via grid search over the same space as the original study: the lower cutoff frequency $f_{\text{low}} \in \{50, 100, \ldots, 1000\}$ Hz, and the threshold multiplier $r_{\text{std}} \in \{0.05, 0.10, \ldots, 1.00\}$. This yields $20 \times 20 = 400$ parameter combinations per fold. Searching the full grid rather than a restricted range avoids assuming that the optimum reported in-sample by [5] ($f_{\text{low}} = 700$ Hz, $r_{\text{std}} = 0.20$) transfers to every training partition. For each LOSO fold, grid search is performed on the 15 training subjects by pooling their predictions and computing a single MCC score for each parameter combination. The parameters maximizing training MCC are then applied to the held-out subject. This nested procedure ensures that hyperparameter selection does not use test data. This experiment quantifies the bias introduced by in-sample evaluation in the original work. Figure 1 illustrates this processing pipeline.

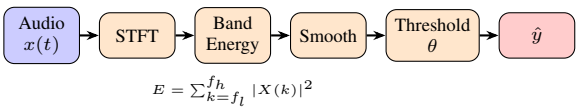


Fig. 1. Energy detector pipeline: STFT followed by band-limited energy extraction, smoothing, and thresholding.

2) Experiment 2: CLAP Zero-Shot Classification: The second experiment evaluates zero-shot classification using LAION-CLAP [6], specifically the `630k-audioset-best.pt` checkpoint with the HTSAT-base audio encoder (`laion-clap` library v1.1.4). We selected CLAP over alternatives such as WavLM [7] or BEATs [8] for three reasons: (1) CLAP produces a single 512-dimensional embedding per audio segment, avoiding the need to aggregate frame-level representations; (2) the joint audio-text embedding space enables zero-shot classification, allowing us to test whether swallowing can be detected without any task-specific training; and (3) CLAP was trained on general audio including environmental sounds, whereas WavLM focuses on speech. Classification uses cosine similarity between audio and text embeddings [12].

Since LAION-CLAP operates at 48 kHz while the recordings were captured at 22.05 kHz, we upsample the audio using linear interpolation before embedding extraction. This upsampling does not create spectral content above the original Nyquist frequency (11 kHz), leaving the upper half of CLAP's expected bandwidth empty. We extract 512-dimensional embeddings using a sliding window of 480 ms with 50 ms hop, aligning with the annotation grid. For classification, we compute the difference between cosine similarity to positive prompts and cosine similarity to negative prompts. Negative prompts include "silence", "breathing sound", "background noise", and "coughing sound". We evaluate four positive prompt categories to assess sensitivity to prompt engineering: simple ("swallowing sound", "gulp sound"), descriptive ("the sound of a person swallowing water" and variants), medical ("deglutition sound", "pharyngeal swallow sound") and a combination of these. The decision threshold is optimized on all data, providing an upper bound on zero-shot performance.

3) Experiment 3: Linear Probe: The third experiment trains a logistic regression classifier on frozen CLAP embeddings, following the linear probing paradigm commonly used to evaluate pre-trained representations. The same 512-dimensional embeddings from Experiment 2 are used, with the CLAP encoder weights kept frozen throughout training. For each LOSO fold, embeddings are standardized to zero mean and unit variance using statistics computed exclusively on the training fold; the same transformation is then applied to the test fold. We use scikit-learn's LogisticRegression with the L-BFGS solver, balanced class weighting (inversely proportional to class frequencies), and default L2 regularization ($C = 1.0$). Critically, no hyperparameter tuning is performed: all settings are fixed a priori, including a decision threshold of 0.5 with no post-hoc adjustment. This deliberate choice makes the linear probe a minimal-adaptation baseline that isolates the contribution of CLAP features. If the probe outperforms the energy detector despite zero tuning effort, the gain is attributable to the quality of the pre-trained representations rather than to careful engineering. The pipeline is that of Figure 2 with the MLP head replaced by logistic regression.

4) Experiment 4: MLP Head: The fourth experiment replaces logistic regression with a shallow multilayer perceptron (MLP) to assess whether non-linear combinations of features improve performance. All hyperparameters were fixed a priori, before any LOSO evaluation, to ensure a fair comparison with the baseline. The architecture consists of three fully-connected layers: $512 \rightarrow 128 \rightarrow 64 \rightarrow 1$, with ReLU

activations and dropout ($p = 0.3$) after each hidden layer. The output uses sigmoid activation for binary classification. We train using binary cross-entropy loss with positive class weighting $w_+ = 10$, which corresponds to the inverse class frequency (negative/positive ratio ≈ 10.3). Optimization uses Adam with initial learning rate 10^{-3} and ReduceLROnPlateau scheduling (patience 3, factor 0.5). Training runs for 20 epochs with batch size 64. Inner validation follows the fixed rule described in Section III-B: the model trains on 14 subjects and the checkpoint maximizing inner-validation MCC is selected. As with the linear probe, the decision threshold is fixed at 0.5 with no post-hoc adjustment. To verify that the result does not depend on a particular initialization, the full LOSO procedure was repeated across five random seeds, yielding MCC $= 0.766 \pm 0.008$. This design tests whether a lightweight non-linear head can extract additional discriminative information from the frozen embeddings. Figure 2 illustrates this architecture.

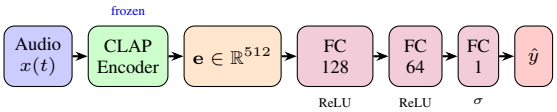


Fig. 2. MLP head pipeline: frozen CLAP encoder with trainable three-layer MLP classifier.

D. Representation Analysis

To understand the structure of CLAP embeddings, we perform dimensionality reduction using UMAP and visualize the embedding space colored by class label and detection outcome. This analysis reveals whether swallowing events form separable clusters and characterizes where classification errors concentrate in the representation space.

IV. Results

A. Main Results

Table I presents the performance of all evaluated methods. The energy detector reproduced under LOSO cross-validation achieves MCC = 0.696 with a standard deviation of 0.13 across subjects. This is below the 0.712 reported in [5], confirming the expected drop when proper cross-validation replaces in-sample evaluation. The bias, 0.015 MCC at full precision, is modest but demonstrates that the original results were optimistically biased.

Zero-shot classification with CLAP embeddings and text prompts yields MCC = 0.303, well below the baseline. This result was obtained with the threshold optimized directly on the test data, meaning it represents an upper bound on zero-shot performance. The failure is unsurprising given the domain mismatch between air microphone recordings in the pre-training data and the contact sensor used here, combined with the absence of swallowing as a concept in the pre-training vocabulary.

The linear probe reaches MCC = 0.743 and the MLP head 0.768, both above the baseline's 0.696 in mean MCC. Subject-clustered inference, however, finds none of the three pairwise contrasts significant at this sample size (all $p \geq 0.21$; see Limitation 6), so we treat the two adaptation strategies as empirically equivalent and read the ordering of means as suggestive rather than established. What the intervals do resolve is dispersion: both adapted methods roughly halve the inter-subject standard deviation, from 0.13 to 0.07–0.08.

To establish a fair comparison ceiling, we apply a leave-one-observer-out procedure analogous to LOSO for models. For each of the three observers, we exclude that observer from the ground truth computation and calculate the consensus using only the other two (a bin is positive if both remaining observers agree). We then compute MCC between the excluded observer's annotations and this two-observer consensus. This yields individual MCCs of 0.852 (Obs1), 0.857 (Obs2), and 0.771 (Obs3), with mean MCC = 0.827 $\pm$ 0.048. Observer 3 scores lower because they marked substantially more positive bins (832 vs. approximately 600 for the others), generating more false positives against the stricter two-observer consensus. This leave-one-out ceiling is the appropriate reference because models are evaluated against the same hierarchical consensus that no single observer fully determined. The MLP reduces the gap to this human ceiling from 0.131 (baseline) to 0.059.

TABLE I
Performance comparison across methods.

Method	MCC	95% CI‡	Std	F1	Rec.	Prec.
Human ceiling*	0.827	–	0.05	–	–	–
Energy (in-sample)	0.712	–	–	0.735	0.694	**0.780**
Energy (LOSO)	0.696	.610–.771	0.13	0.721	0.688	0.757
CLAP Zero-Shot†	0.303	–	–	0.266	0.167	0.657
CLAP Linear	0.743	.702–.781	0.08	0.771	0.821	0.726
CLAP MLP	**0.768**	.717–.806	**0.07**	**0.782**	**0.923**	0.678

*Each observer vs. leave-one-out consensus of other two.
†In-sample threshold optimization (LOSO not applicable).
‡Cluster bootstrap resampling subjects (10,000 replicates). All clustered inference is computed over the 8086 non-padded bins, where the methods score 0.691, 0.743 and 0.763; the MCC column uses the full grid (Limitation 7).

B. Detection Examples

Figure 3 shows the CLAP MLP detector output for the median-performing subject: laryngophone spectrogram, ground truth annotations, and predictions with detection outcomes (TP, FP, FN). The caption gives the counts for the best and worst subjects.

Across the performance range the failure mode is consistent: recall stays high while false positives vary, so errors concentrate on specificity rather than sensitivity.

C. Event-Level Evaluation

A 50 ms bin is not the clinical unit of interest, so we complement the bin-level MCC with an event-level analysis. Consecutive positive bins form events, and a detection counts

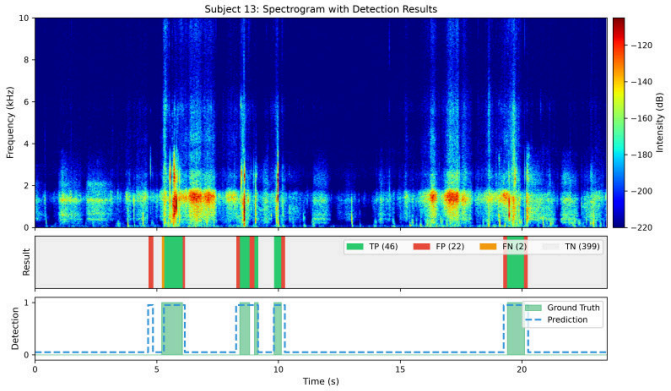


Fig. 3. CLAP MLP detector output for the median-performing subject (S13: 46 TP, 22 FP, 2 FN), showing the laryngophone spectrogram, annotated events and predictions. Performance across the cohort ranges from 24 TP / 10 FP / 0 FN (best, S03) to 57 TP / 76 FP / 0 FN (worst, S08); recall stays high throughout while the false positive rate varies.

as a hit when its onset falls within 250 ms of an annotated onset, matched greedily one-to-one. The tolerance follows from the annotations: the 64 annotated events have a median duration of 0.65 s, so 250 ms lies well inside a typical event.

Table II qualifies the bin-level picture. Without a duration filter the ranking inverts: the energy detector, last by MCC, raises the fewest false alarms per minute. The cause is fragmentation. The CLAP heads recover more true positive bins (787 and 700, against 587) but scatter them across short spurious detections. 33% of the MLP's and 48% of the linear probe's last no longer than 150 ms, against 8% for the energy detector, and each fragment counts as a separate false alarm.

Because the shortest annotated event lasts exactly 150 ms, discarding shorter detections removes no real event. Under that filter the MLP attains both the highest event sensitivity (0.938) and fewer false alarms than the linear probe (2.97 versus 3.86/min), while the energy detector stays most conservative (1.93). Onset error is small and comparable throughout (median $\leq$ 100 ms, IQR 100 ms), so the methods differ in how often they fire, not in when. We report both variants because the fragmentation is real, and a deployed system would have to post-process it.

TABLE II
EVENT-LEVEL PERFORMANCE AT 250 MS ONSET TOLERANCE. FA/MIN IS FALSE ALARMS PER MINUTE OVER 6.7 MINUTES OF RECORDING CONTAINING 64 ANNOTATED EVENTS.

	No filter			$\geq$150 ms		
Method	**Sens.**	**Prec.**	**FA/min**	**Sens.**	**Prec.**	**FA/min**
Energy (LOSO)	0.891	**0.792**	**2.23**	0.859	**0.809**	**1.93**
Linear probe	**0.938**	0.426	12.02	0.906	0.690	3.86
CLAP MLP	**0.938**	0.522	8.16	**0.938**	0.750	2.97

Onset error medians: 0 ms (energy), −50 ms (linear probe), −100 ms (MLP); IQR 100 ms throughout, negative means early.

Figure 4 shows bin-level precision-recall curves, obtained by refitting the heads on the cached embeddings under the same LOSO protocol to recover continuous scores. Average precision is 0.863 for the MLP and 0.850 for the linear probe against a chance level of 0.105, and the curves overlap across most of the operating range, the same equivalence the significance tests report. Markers show the fixed 0.5 threshold, set a priori and not re-tuned here; at that point the MLP favours recall (0.923) over precision (0.678) while the linear probe is more balanced, and the curves show what either method would trade if a different operating point were chosen.

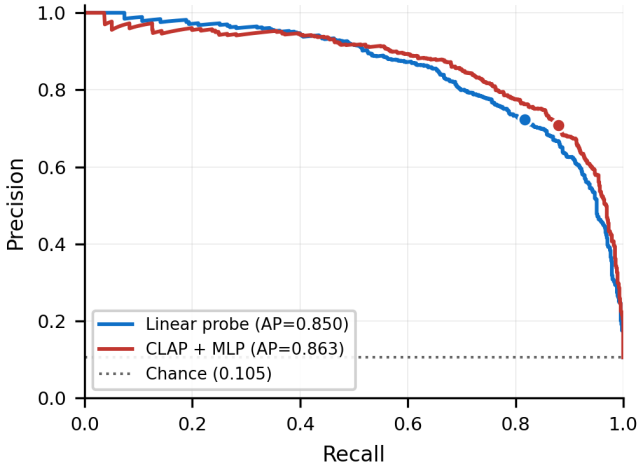


Fig. 4. Bin-level precision-recall curves for the adapted methods. Markers show the fixed 0.5 threshold applied to the refitted heads, which sit up to 0.04 in recall off the Table I operating points.

D. Per-Subject Performance

Figure 5 displays the MCC distribution across subjects for each trained method. The energy detector exhibits high variance, with some subjects yielding MCC above 0.80 and others below 0.55. The CLAP-based methods show a compressed distribution with fewer outliers. This reduction in variance is the central finding of the study: the differences in mean MCC do not survive subject-clustered testing, whereas the gain in predictability across individuals is what the data support.

E. Label Uncertainty and Error Attribution

Because the reference is a union across annotators, the number of observers who marked a bin measures its label confidence, separating two error sources the ceiling alone cannot distinguish. The MLP's recall tracks that measure almost perfectly: 99.6% of the 514 bins all three observers marked, 91.5% of the 153 marked by two, and 72.6% of the 186 marked by only one. It essentially never misses an event the annotators agree on, and its misses concentrate where the label is least certain, so the residual false negatives are largely a property of the reference rather than of the model.

False positives read similarly. All 374 fall, by construction, in bins no observer marked, but they are not scattered: 46.5% lie within 100 ms of an annotated event and 60.7% within

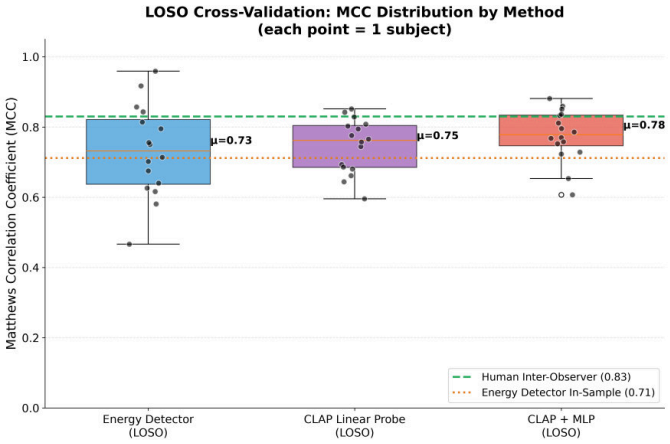


Fig. 5. Per-subject MCC distribution under LOSO cross-validation; the dashed line marks the human ceiling (0.827). Per-subject means differ from the pooled MCC of Table I by construction, since pooling aggregates all predictions into one confusion matrix: 0.696 vs. 0.732 for the energy detector and 0.768 vs. 0.776 for CLAP MLP, the larger gap reflecting higher calibration variability in the baseline.

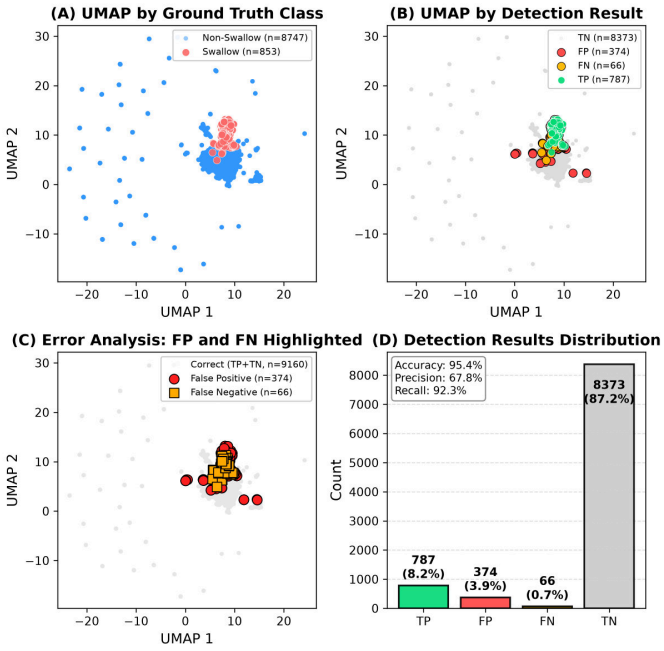


Fig. 6. UMAP projection of the 9600 CLAP embeddings (853 swallowing, 8747 non-swallowing). (A) ground truth class; (B) MLP detection outcome; (C) errors highlighted over correctly classified bins; (D) outcome counts. False positives cluster near class boundaries, indicating partial separability.

200 ms, so most are disagreements about event boundaries rather than detections of unrelated sounds, and a further 29.9% lie within 100 ms of a bin a single annotator marked. We had hypothesised that false positives would concentrate on breathing, the predicted confounder, but the annotations mark only swallowing intervals and carry no event type, so this cannot be tested here.

F. Representation Analysis

Figure 6 projects the CLAP embeddings with UMAP. Positive bins form a cluster only partially separable from negative bins, and the errors mirror the temporal picture above: false positives concentrate at the class boundary while false negatives are rare and scattered.

V. Discussion

A. Why Does Zero-Shot Classification Fail?

Zero-shot classification fails because of a fundamental domain shift between CLAP's pre-training data and laryngophone recordings. Three factors contribute to this failure.

First, CLAP was trained on 48 kHz air-microphone recordings of environmental sounds, music and speech, whereas the encoder here receives upsampled 22.05 kHz signals whose upper spectral band is empty, a distribution shift absent from pre-training. Compounding it, the laryngophone is a contact sensor whose transfer function attenuates high frequencies and emphasizes low-frequency body-conducted vibration, giving spectral characteristics fundamentally unlike air-conducted sound.

Second, biomedical sounds are largely absent from CLAP's pre-training corpus. Concepts like "drinking" or "gulp" were learned from air-microphone recordings in everyday contexts, not the muffled contact-sensor signature of pharyngeal muscle contraction, so the semantic concept of "swallowing" in CLAP's embedding space does not align with the acoustic reality of laryngophone signals.

Third, the failure is not an artifact of prompt engineering. The four prompt strategies yield MCC of 0.286 (simple), 0.303 (descriptive), 0.291 (medical) and 0.295 (combined), a spread of only 0.017, and none approaches the baseline: the limitation lies in the audio encoder's representation of contact-sensor signals rather than in the choice of text anchors. Even this ceiling used a threshold optimized on the test data, so without that unrealistic advantage performance would be worse still. Foundation models therefore do not provide out-of-the-box solutions for physiological sound detection with contact sensors.

B. Why Does Linear Probing Succeed?

Despite the domain shift, an untuned logistic regression on frozen CLAP embeddings matches the energy detector in mean MCC while halving its inter-subject spread. The embeddings therefore already carry discriminative information: the obstacle for zero-shot classification is task adaptation, not representation quality.

CLAP's contrastive pre-training learns to distinguish acoustic events based on spectral and temporal patterns, even if the semantic labels differ from swallowing. Features such as transient energy bursts, frequency modulation patterns, and temporal structure are generic enough to transfer across domains. The linear probe learns a hyperplane that separates swallowing from non-swallowing bins in this 512-dimensional space, bypassing the failed text-audio alignment entirely.

C. Does Non-Linearity Help Over Linear Probing?

The MLP head achieves MCC = 0.768 against 0.743 for the linear probe, a margin of 0.020 on the non-padded bins over which all contrasts are computed (see Table I). With 16 subjects it is not separable from zero: the paired permutation test gives $p = 0.33$ and the MLP wins on only 9 of 16 subjects. We therefore cannot claim that non-linearity helps, and the following reading is offered as a hypothesis for future work with a larger cohort rather than as a conclusion supported by these data.

The UMAP visualization (Figure 6) shows swallowing events forming a cluster that overlaps with non-swallowing bins at irregular boundaries. If that overlap reflects genuine curvature in the class boundary, an MLP's hidden layers could capture it where a linear classifier cannot. The observed margin is at least consistent with this picture, but it is equally consistent with sampling noise.

Whatever the head, the underlying embedding already provides most of the separability, which is why the choice between a linear and a non-linear head has so little room to matter here.

D. What Would Be Required for Clinical Use

Telemedicine motivates this task, but nothing here demonstrates a telemedicine result. What these results support is narrower: foundation model representations can be adapted to a contact-sensor signal with a small cohort and a frozen encoder, and that adaptation makes performance more consistent across individuals. Consistency is a precondition for running without per-patient calibration, which is why we regard it as the practically relevant finding, but a precondition is not a demonstration.

Three gaps separate this evidence from a clinical claim, detailed in Limitations 1 and 2: an all-healthy cohort, a single sensor and setup, and controlled recording conditions. Closing them requires validation on patients, across devices and sites, under realistic conditions, not extrapolation from the numbers reported here. The MLP's operating point (recall = 0.923, precision = 0.678) favors sensitivity, which would be the appropriate trade-off in a screening context, since a missed event costs more than a false positive a clinician can dismiss on review.

E. Limitations

Several limitations constrain the generalizability of these findings.

First, the dataset comprises only 16 healthy subjects. No patients with ALS or Parkinson's disease were included, precluding direct conclusions about the clinical target population. Swallowing patterns in dysphagic patients may differ substantially from healthy individuals.

Second, all recordings come from a single laryngophone and acquisition setup under controlled conditions, so robustness to environmental noise, sensor variability and multi-site acquisition remains untested. Whether the CLAP representation degrades more gracefully than the spectral baseline under such conditions is the robustness question this work leaves open.

Third, the ground truth carries residual label noise, which the leave-one-observer-out ceiling of MCC = 0.827 quantifies: even trained annotators do not fully reproduce their peers' consensus. Rather than leave this as a caveat, Section IV-E stratifies the detector's errors by how many annotators marked each bin, and finds that they track label uncertainty closely. The residual noise therefore bounds what any detector can achieve here, and part of the measured error is a property of the reference rather than of the model.

Fourth, the 50 ms temporal resolution is coarser than the acoustic dynamics of swallowing events, which typically last 500-1000 ms but contain sub-events at finer timescales. This resolution was inherited from the original study for comparability.

Fifth, this study evaluates a single foundation model and so cannot separate what is specific to CLAP from what would hold for audio foundation models generally. Settling that requires comparison against smaller acoustic features and against alternative encoders such as BEATs, WavLM or health-specific models, under an identical protocol and matched compute budget; this is left to future work. Our results therefore show that *a* general-purpose encoder transfers to this domain, not that CLAP is the best such encoder.

Sixth, with 16 subjects the resolution of the comparison between methods is limited, and we quantify that limit rather than assume it. Because consecutive bins within a recording are temporally correlated, all inference is clustered at the subject level, with one recording per subject so that no nesting is required: confidence intervals come from a cluster bootstrap over subjects and method contrasts from paired permutation tests that exchange the two methods' predictions within each subject (10,000 replicates each). None of the three contrasts reaches significance: MLP versus linear probe gives ΔMCC $= +0.020$, 95% CI $[-0.012, +0.055]$, $p = 0.33$; MLP versus energy $+0.072$, $[-0.024, +0.171]$, $p = 0.21$; linear probe versus energy $+0.052$, $[-0.025, +0.130]$, $p = 0.25$. Wilcoxon signed-rank and paired t-tests agree, and in each contrast the nominally better method wins on only 9 of 16 subjects. We therefore treat the two adaptation strategies as empirically equivalent at this sample size and do not claim that either significantly outperforms the energy baseline in mean MCC. What the intervals do separate is dispersion: the bootstrap standard error of the energy detector (0.041) is roughly twice that of the linear probe (0.020) or the MLP (0.023), and its per-subject spread is likewise larger (0.134 versus 0.077 and 0.074).

Seventh, recording durations vary from 13.8 to 32.6 seconds across subjects against a fixed 600-bin (30 s) grid; the two longer recordings are truncated at 30 s, discarding no annotated event. The 1514 padded bins (15.8% of the total) carry no annotated events and are trivially classified as true negatives, which inflates specificity. We quantified this effect by recomputing every metric over the 8086 bins that contain recorded audio: pooled MCC shifts by at most 0.005 (0.005

for the energy detector, 0.005 for the MLP head), and the ordering of methods is unchanged. Specificity, in contrast, is inflated by 0.005 to 0.008, which is one further reason to read MCC rather than specificity under this class balance.

VI. Conclusion

This study shows that audio foundation model representations encode transferable information for swallowing event detection, despite a severe domain mismatch between air-microphone pre-training data and contact-sensor recordings. The transfer, however, is conditional on how the model is used. Zero-shot classification through the text-audio alignment fails at MCC = 0.303, and this failure is robust to prompt formulation, showing that the semantic pathway of the foundation model does not reach laryngophone signals. Bypassing that pathway and training a lightweight head on the frozen embeddings raises MCC to 0.743 for a linear probe and 0.768 for an MLP head, against 0.696 for a band-limited energy detector evaluated under the same leave-one-subject-out protocol.

The central finding is not the gain in mean performance, which subject-clustered testing cannot separate from zero at this sample size, but the reduction in inter-subject variability: the standard deviation falls from 0.13 for the baseline to 0.07–0.08 for the adapted models, and the bootstrap standard error of the pooled MCC falls by roughly half. A detector whose performance is predictable across individuals is a precondition for deployment without per-patient calibration, though demonstrating such deployment is beyond what this study evaluates. Measured against a leave-one-observer-out human ceiling of MCC = 0.827, the adapted models narrow the gap from 0.131 to 0.059.

These conclusions are bounded by the cohort that produced them: 16 healthy adults, one laryngophone setup, no dysphagic patients and no device or site variation. The result to carry forward is that general-purpose audio representations transfer to a contact-sensor domain more consistently than a hand-tuned spectral baseline. Whether that consistency survives contact with patients, devices and recording environments is the question future work must answer, alongside encoders with native frame-level temporal resolution.

Acknowledgment

The data used in this study were collected under the RevELA Project at the Laboratory for Technological Innovation in Health (LAIS), Federal University of Rio Grande do Norte. The authors also acknowledge CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for supporting this research.

References

[1] I. Suttrup and T. Warnecke, "Dysphagia in Parkinson's disease," *Dysphagia*, vol. 31, no. 1, pp. 24–32, Feb 2016.

[2] E. Onesti, I. Schettino, M. C. Gori, V. Frasca, M. Ceccanti, C. Cambieri, G. Ruoppolo, and M. Inghilleri, "Dysphagia in amyotrophic lateral sclerosis: Impact on patient behavior, diet adaptation, and riluzole management," *Frontiers in Neurology*, vol. 8, p. 94, Mar 2017.

[3] J. M. Dudik, A. Kurosu, J. L. Coyle, and E. Sejdić, "Dysphagia and its effects on swallowing sounds and vibrations in adults," *BioMedical Engineering OnLine*, vol. 17, no. 1, Dec 2018.

[4] G. S. da Silva, A. C. G. de Figueiredo, B. S. Zeigelboim, C. M. de Araújo, J. Santos Neto, and R. S. Santos, "Acoustic analysis of swallowing as an auxiliary method for assessing dysphagia in Parkinson's disease," *Research, Society and Development*, vol. 10, no. 16, p. e323101623298, 2021.

[5] I. Sánchez-Gendriz, I. M. P. Barbalho, F. Fernandes, M. M. Augusto, M. de Jesus Gonçalves, I. L. Barbosa Lima, D. Nagem, G. C. Silva, A. Raquel Rodrigues Lindquist, L. F. Alves, L. Affonso Guedes, and R. Valentim, "Methods for analyzing swallowing sound in dysphagia care: A telemedicine approach," in *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2024, pp. 1–8.

[6] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 11 593–11 606.

[9] E. S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 626–633, 2010.

[10] Y. Khalifa, J. L. Coyle, and E. Sejdić, "Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings," *Scientific Reports*, vol. 10, no. 1, p. 8704, 2020.

[11] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 44–56, Jan 2018.

[12] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy *et al.*, "HEAR: Holistic evaluation of audio representations," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[14] S. Baur, J. Shor, C. Ding, S. Venugopalan, M. Orescanin, N. Jaitly, S. Huang, and Y. Tian, "HeAR – health acoustic representations," *arXiv preprint arXiv:2403.02522*, 2024.

[15] S. Rokhsaritalemi, M. Sadeghi, F. Karray, and E. Alreshidi, "A review of ICBHI 2017 respiratory sounds analysis using deep learning," in *2024 IEEE International Conference on E-health Networking, Application & Services (Healthcom)*. IEEE, 2024, pp. 1–6.

[16] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan 2020.

Evaluating Deep Neural Network Architectures for Motor Unit Firing Classification from Decomposed HD-sEMG Signals

1st Joyce Gomes de Queiroz
Faculdade de Tecnologia (FT)
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brasil
j186846@dac.unicamp.br

2nd Denis Gustavo Fantinato
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brasil
denisf@unicamp.br

3rd Carina M. Germer[1,2]
1. Faculdade de Engenharia Elétrica e de Computação,
Universidade Estadual de Campinas - UNICAMP,
2. Centro de Engenharia Biomédica da UNICAMP,
Campinas - SP, Brasil.
cgermer@unicamp.br

***Abstract*—The decomposition of high-density electromyography (HD-EMG) allows estimating the individual activity of the motor units (MUs) recruited during muscle contraction. The resulting firing trains, however, still have to be picked by hand, an essential step that is slow and prone to error. This work evaluates artificial neural networks for the classification of MU firings, using as input the instantaneous pulse trains (IPTs). Three architectures were compared on a single 41-second isometric HD-EMG recording from one participant (64 channels, 17 motor units, with a manually edited reference produced by one experienced evaluator): a one-dimensional convolutional neural network (CNN-1D) acting as a per-unit binary detector, a multiclass two-dimensional convolutional neural network (CNN-2D), and a recurrent Long Short-Term Memory (LSTM) network. The models were compared by the average Rate of Agreement (RoA) with the evaluator's manual editing. The CNN-1D obtained the highest value, with an initial RoA of 55.44%. A hyperparameter search and the consolidation into a single cross-MU model raised the average RoA to 78.48%, reaching approximately 90% for fourteen of the units (all except MUs 15–17). All results were obtained on a single recording with a random partition of windows, so they should be read as a preliminary characterization rather than as an estimate of performance on unseen recordings, motor units or participants.**

***Index Terms*—high-density electromyography, motor units, decomposition, convolutional neural networks, deep learning.**

I. Introduction

The decomposition of high-density electromyography (HD-EMG) signals, whether surface (HD-sEMG) or intramuscular (HD-iEMG), identifies the sources that generate the recorded signals, which correspond to the electrical activity of the motor units (MUs) [1]. Blind source separation methods such as convolution kernel compensation (CKC) [2] and fast independent component analysis (FastICA) [3] have been applied successfully to this task.

The quality of the decomposition depends on the spatio-temporal characteristics of each MU. The spatial distribution of the signals is affected by the volume conductor, the size of the MU, the relative position between the innervation zone and the electrode, and the recording system [4], whereas the temporal properties reflect the firing pattern. When different MUs have similar spatial profiles, the separation can be imperfect, producing mixed firing trains or classification errors.

After the automatic decomposition, manual picking of the firing trains remains an essential step to ensure the validity of the subsequent analyses. Erroneous classifications, such as false positives and false negatives, can compromise standard analyses in the neurophysiology of motor control (synchronization, correlation, and coherence between MUs, as well as individual characteristics such as firing rate and firing variability) and should be avoided. Manual validation is also slow: a trained evaluator takes, on average, about one minute to review the firing train of a single MU in a one-minute recording. Each recording may present dozens of MUs, and each study involves multiple participants and trials, so this step is one of the main bottlenecks in experimental projects [5].

The use of artificial neural networks (ANNs) for HD-sEMG data processing has grown in recent years [6]. ANN models can automate parts of this pipeline, including the picking of firing trains. To our knowledge, however, no ANN model has been dedicated specifically to the correct classification of the firings obtained from decomposition. That step matters as much as the initial separation, especially for applications that require real-time decomposition, such as human-machine interfaces, neurofeedback, and prosthesis control. A network dedicated to firing classification could shorten the analysis and support online applications.

In this work, three architectures are compared using as

input the instantaneous pulse trains (IPTs) estimated by the decomposition. The main contributions are: (i) the comparison between a CNN-1D (per-unit binary detector), a multiclass CNN-2D, and an LSTM network on a single 41-second recording from one participant; (ii) the identification of the CNN-1D as the most suitable of the three; and (iii) a refinement process under which the average RoA rises from 55.44% to 78.48% on this dataset.

II. Methods

A. Dataset

An anonymized HD-sEMG recording from the Neuroengineering Research Laboratory (NERLab, CEB/UNICAMP) database was used [7]. The study analyzed only previously collected, fully anonymized data, with no access to identifiable participant information, so no additional ethical approval was required. The task consisted of isometric contraction of the index finger at 5% of maximal voluntary contraction for 41 s. The signals were acquired with a grid of 13×5 electrodes, totaling 64 active channels (as one electrode position in the grid is empty), at a sampling frequency of 2048 Hz. The decomposition with gCKC identified 17 MUs and 10,781 valid firings, with the number of reference firings per MU ranging from 500 to 797 (median 643). Since the sliding scan covers the whole recording, the number of candidate windows is the same for every MU, and only the positive fraction varies with the firing count. The database contains two versions of the recording, referred to as original and clean; the firing instants of the clean version were manually classified by an experienced evaluator and were adopted here as reference labels.

The dataset was split into 80% for training and 20% for validation, with a fixed seed to ensure reproducibility. All windows originate from a single continuous recording, so the random partition places windows from the same recording and the same MU in both subsets. Because the morphology of the IPT peaks of a given MU is approximately stationary within a recording, the two subsets are not statistically independent, and the values reported here characterize interpolation within a single recording rather than generalization to new recordings, motor units or participants. Two validation schemes would be required to support claims of generality: a time-block split, in which the model is trained on one portion of the recording and evaluated on a later one, and a leave-one-motor-unit-out scheme, in which the evaluated unit contributes no data to training. Neither is reported here, and both are left for future work.

B. Processing HD-sEMG Recordings

The spike-selection task is cast as a supervised binary classification problem at the level of individual firing candidates. After decomposition, each MU is associated with an instantaneous pulse train (IPT), a one-dimensional signal whose local peaks indicate candidate firing instants. Rather than operating on the raw multichannel EMG, the model takes as input a fixed-length window of the IPT (60 samples, $\approx$ 29 ms), which preserves the temporal shape of the event while keeping the input dimensionality low.

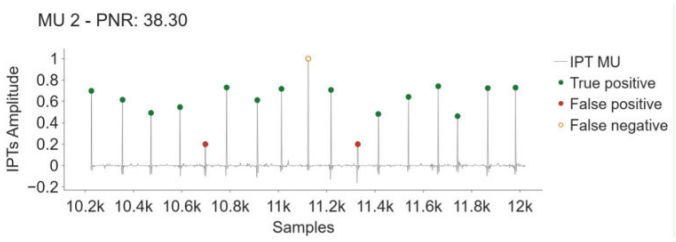


Figure 1. Firing classification from the instantaneous pulse train. In each 60-sample window centered on a candidate, the network decides between firing and non-firing: on the left, a true positive (sharp, isolated peak: accepted); on the right, a false positive (low-amplitude peak associated with noise: rejected).

Candidates were generated by a sliding scan along the IPT with a stride of 2 samples, so that consecutive windows overlap by 58 of their 60 samples. A window was assigned to the positive class when a reference firing of that MU falls within its support, that is, within ± 30 samples of the window center; at the sampling rate of 2048 Hz, where consecutive samples are 0.488 ms apart, this corresponds to ± 14.6 ms. Windows without a reference firing in their support were assigned to the negative class, and therefore correspond to noise, to interference from other units, and to inter-discharge intervals. Because the two classes are not equally represented in the scan, the larger set was randomly subsampled to the size of the smaller one before training. The labeling tolerance is deliberately wide: rather than marking a single instant, it marks the neighborhood of each discharge, so that every reference firing gives rise to a group of consecutive positive windows. A non-maximum suppression (NMS) stage is applied at inference to collapse each group into a single detected firing, and the agreement between detected and reference instants is then assessed with the much tighter tolerance defined in Section II-E. The manually validated firing instants are used as the reference for both training and evaluation. They are treated here as the best available reference rather than as an error-free ground truth, since they reflect the judgment of one evaluator.

C. Preprocessing and Input Representation

The HD-sEMG signals were filtered with a notch filter at 60 Hz and a 20–500 Hz band-pass filter. Each input window was normalized independently by z-score, using its own mean and standard deviation, over an analysis window of 60 samples (2.93×10^{-2} s, or ≈ 29 ms). No statistics computed over the whole recording or over the whole MU were used in this step, so the relative amplitude between windows is not preserved. The input variables of the models correspond exclusively to the IPTs estimated by the decomposition, which contain the temporal information associated with the MU discharges. This choice reduces the input dimensionality and focuses the model on the temporal characteristics of the firings. Fig. 1 illustrates the network's decision on firing candidates from a MU.

D. Evaluated Architectures

Three neural network architectures were compared. In all cases the input is a single 60-sample window, and the per-model design is described below.

- **CNN-1D (binary detector):** A set of 17 independent one-dimensional convolutional neural networks (CNNs), each trained to classify a window as a firing or non-firing of a specific MU (sigmoid output). Each network stacks three convolutional layers with decreasing kernel sizes, with 32 filters (kernel 7), 64 filters (kernel 5) and 128 filters (kernel 3), the first two followed by batch normalization and all with ReLU activation. A global average pooling layer is followed by a 64-unit dense layer with dropout (0.3) and the final sigmoid unit, totaling about 44k parameters. CNNs suit EMG signals because they extract local features and morphological patterns of the action potentials, reduce the number of parameters through weight sharing, and identify relevant attributes without manual engineering. The network was trained with the Adam optimizer (10^{-3}) and binary cross-entropy.
- **CNN-2D (multiclass):** A single two-dimensional CNN that receives the spatio-temporal cube ($13 \times 5 \times 60$) and classifies the window among the 17 MUs (*softmax* output). It first applies a 3D convolution along the temporal axis (32 filters, kernel $1\times1\times7$), and then 2D convolutions over the electrode grid (64 filters with L_2 regularization, followed by 128 filters), which exploit the spatial invariances of the action potentials. A spatial attention block, which combines average- and max-pooled feature maps through a sigmoid-gated convolution, weights the most informative regions before two dense layers (256 and 128 units, dropout 0.5) and the softmax output. It was trained with Adam (3×10^{-4}) and categorical cross-entropy.
- **LSTM (multiclass):** A recurrent *Long Short-Term Memory* (LSTM) network, which processes the window as a single-channel temporal sequence (without spatial information). After batch normalization, it stacks two bidirectional LSTM layers (64 and 32 units, dropout 0.3), followed by a 128-unit dense layer and a 17-class *softmax* output. LSTMs model short- and long-term temporal dependencies, in principle suitable for the physiologically structured firing patterns of the MUs. It was trained with Adam and categorical cross-entropy.

All models were trained for 50 epochs with a batch size of 64, with the two classes balanced by subsampling, in the Google Colab environment using TensorFlow/Keras.

E. Evaluation Metrics

The agreement between the automatically detected firings and the reference firings was quantified with the *Rate of Agreement* (RoA) [3] defined for each motor unit j as

$$\mathrm{RoA}_j = \frac{c_j}{c_j + A_j + B_j} \cdot 100\%, \tag{1}$$

where c_j is the number of firings identified by both methods (reference and automatic classification), and A_j and B_j are the firings identified by only one of them. Two firings were considered coincident when separated by at most 10 samples (4.88 ms at 2048 Hz). Taking the automatic classification as the predictor and the manually validated instants as the reference, c_j corresponds to the true positives (TP), and A_j and B_j to the false positives (FP) and false negatives (FN), from which precision (P) and recall (R) are computed:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{2}$$

and the F1-score, given by the harmonic mean of the two:

$$\mathrm{F1} = 2\,\frac{P \cdot R}{P + R} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{3}$$

After estimating these metrics for each MU, a macro average was calculated for each architecture. Per-unit values are reported in Fig. 2, so that the dispersion behind each macro average can be inspected. Precision, recall and F1 are computed at the level of individual windows, whereas the RoA is computed on the reconstructed firing train, after the sliding-window scan and the non-maximum suppression applied at inference. The two sets of values are therefore not directly comparable.

F. Optimization of the CNN-1D

Starting from the most promising architecture, a step-by-step refinement process was carried out. First, the hyperparameter search was automated with Optuna [8], an optimization framework based on the Tree-structured Parzen Estimator (TPE), with early stopping and adaptive learning-rate reduction; its objective was the window-level F1 on the validation set, which is the same set from which the metrics below are reported. Then, an independent optimization was performed for each MU. Finally, a single cross-MU model was consolidated, trained with the concatenated windows of all MUs and applied through a sliding-window search followed by non-maximum suppression (NMS) at the inference stage. The term *cross-MU* denotes a single network shared by the 17 MUs of this recording, in contrast to one detector per unit. It carries no claim of generality across recordings, contractions or participants, which was not evaluated. Two aspects change with respect to the per-unit detectors. First, the identity of the MU is not provided to the network: the input is a single IPT window, with no unit label or embedding, so the same weights are applied to every unit. The unit is determined externally, by the choice of which IPT the model is applied to, which is already available after decomposition. Second, at inference the model no longer receives a pre-selected set of candidates, but scans the IPT and relies on NMS to produce one detection per discharge. The hyperparameter search, the per-unit refinement and the final metrics were all computed on the same partition, so the reported values reflect performance on data that participated in model selection and are optimistic by an unquantified margin. An independent test set is required to confirm the magnitude of the reported differences.

Table I
COMPARISON OF EVALUATED MODELS FOR HD-SEMG FIRING CLASSIFICATION.

Model	Task	Input	No. Params	RoA (macro)	Precision	Recall	F1	No. Models	Loss
CNN-1D (binary)	Firing / non-firing	IPT window (60 samples $\times$1 channel)	43,969	55.44%	86.7%	86.9%	86.8%	17 (sigmoid per MU)	Binary cross-entropy
CNN-2D (unified)	Multiclass (17 MUs)	$13 \times 5 \times 60$	3,304,676	46.64%	63.6%	63.3%	63.4%	1	Categorical cross-entropy
LSTM (unified)	Multiclass (17 MUs)	Window $\times$ 1 channel	$\sim$86k	12.8%	22.9%	23.7%	21.9%	1	Categorical cross-entropy
CNN-1D Optuna General + IPT	Binary	IPT window (60 samples $\times$1 channel)	17 tuned (44k–128k/MU)	60.45%	90%	90%	87.2%	17	Binary cross-entropy
CNN-1D cross-MU model	**Binary**	IPT window (60 samples $\times$1 channel)	**$\sim$128k**	**78.48%**	**98%**	**98%**	**97.9%**	1	Binary cross-entropy

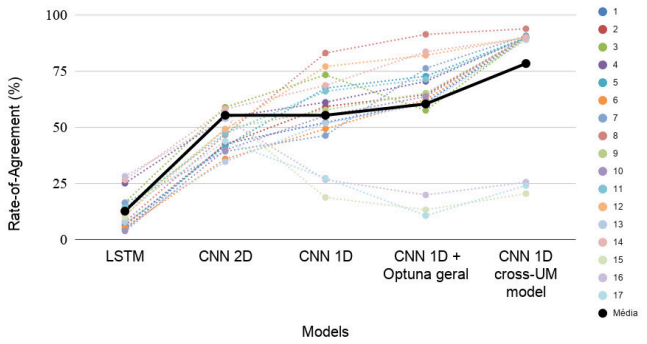


Figure 2. Evolution of the RoA across the architectures and optimization steps (LSTM, CNN-2D, base CNN-1D, general search with Optuna, per-unit optimization, and final cross-MU model). Each light curve corresponds to a MU and the black curve to the average of the 17 units, which rises from 12.8% (LSTM) to 78.48% in the cross-MU model. The fourteen units other than MUs 15–17 reach about 90%, while MUs 15–17 remain below 30%.

III. RESULTS AND DISCUSSION

Table I summarizes the evaluation metrics (macro average) for the three architectures. The CNN-1D, in the form of independent binary detectors, obtained the highest macro RoA (55.44%), as well as balanced precision and recall (86.7% and 86.9%, respectively). The multiclass CNN-2D fell between the other two (RoA of 46.64%; precision 63.6% and recall 63.3%), and the LSTM performed worst, with an RoA of only 12.8%. The LSTM processes the window as a single-channel temporal sequence and discards the spatial information of the electrode grid, which may be important for distinguishing MUs with similar profiles; its low RoA is consistent with that. The CNN-2D, although it explicitly exploits the spatio-temporal cube, showed substantial overfitting: the training accuracy (89.4%) exceeded the validation accuracy (65.0%) by more than 24 percentage points, whereas the CNN-1D kept training and validation practically equal (86.5%). The specialization of one detector per unit reduces the complexity of the problem faced by each network, which motivated the choice of this architecture for the following steps.

The per-MU curves in Fig. 2 show that the CNN-1D obtains higher values than the CNN-2D and the LSTM on most of the units, and that the CNN-2D is ahead only on units 15–17, where the base CNN-1D collapses. For the CNN-1D, the macro precision and recall are both high ($\approx$ 86–87%), but the RoA is, by construction, more conservative: since it simultaneously penalizes false positives and false negatives in the denominator of Eq. (1), units with many false positives (for example, MUs 1, 6, and 9, with recall above 97% but more than a hundred false positives each) end up with an RoA in the 50–58% range, despite good detection. Units 15, 16, and 17 show low RoA across all architectures. The cause was not established in this study, and a source-pairing inconsistency between the two decompositions cannot be ruled out.

The optimization process of the CNN-1D, illustrated in Fig. 2, produced a consistent increase in RoA across the successive stages. No statistical test was applied, since all stages were computed on a single recording and on the same partition, so the differences reported below are descriptive. Starting from the base architecture (average RoA of 55.44%), the general hyperparameter search with Optuna raised the average to 60.45%, the independent per-unit optimization to 69.8%, and the consolidation into a single cross-MU model reached 78.48%. This model was trained with the concatenated windows of all units and applied through a sliding-window scan followed by NMS. Restricting the analysis to the fourteen units other than MUs 15–17, the average RoA rises from 62.1% (base) to 90.2% in the final model, with individual values close to 90%. Units 15, 16, and 17 remained at low levels (about 20–26%) even after optimization. The cross-MU model also replaces the 17 independent detectors with a single network of approximately 128 thousand parameters, which simplifies its use at inference time.

Several limitations apply. The evaluation was restricted to a single 41-second HD-sEMG recording from one participant, so generalizability remains untested. The random partition places windows from the same recording and the same motor unit in both subsets, so the reported RoA characterizes interpolation within a single recording rather than generalization to new

motor units, contractions or participants, and the same partition was also used for model selection, which makes the values optimistic by an unquantified margin. The final average RoA of 78.48% remains well below the $\sim$98% typically required for decomposed firing trains to be used without further manual editing; in its present form the approach is a pre-selection tool that reduces, but does not replace, manual editing. The reference labels were produced by a single evaluator, and no inter-rater agreement was assessed. A relevant next step is to report the comparison of the RoA before (original classification) and after the model, in order to verify whether the network improves, rather than degrades, the original classification.

IV. Conclusion

This work investigated the use of neural networks to automate the classification of MU firings obtained from HD-sEMG decomposition, a step currently performed by manual editing and one of the main bottlenecks of the analysis. Among the three compared architectures, the CNN-1D in the form of a per-unit binary detector obtained the highest RoA and the smallest gap between training and validation. Starting from it, the hyperparameter search and the consolidation into a cross-MU model raised the average RoA from 55.44% to 78.48% on this dataset, reaching about 90% for the fourteen units other than MUs 15–17. These are preliminary results obtained on a single recording, and the final average RoA remains below the $\sim$98% usually required for decomposed firing trains to be used without further manual editing, so the approach reduces rather than replaces that step. Confirming the reported values requires evaluation on independent recordings and on data not used during model selection.

Acknowledgments

The authors acknowledge the scholarship provided by the National Council for Scientific and Technological Development (CNPq) through the Institutional Program for Scientific Initiation Scholarships (PIBIC) and FAPESP (2024/08318-4).

References

[1] D. Farina and A. Holobar, "Characterization of human motor units from surface EMG decomposition," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 353–373, 2016.

[2] A. Holobar and D. Zazula, "Multichannel blind source separation using convolution kernel compensation," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4487–4496, 2007.

[3] F. Negro, S. Muceli, A. M. Castronovo, A. Holobar, and D. Farina, "Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation," *Journal of Neural Engineering*, vol. 13, no. 2, p. 026027, 2016.

[4] D. Farina, C. Cescon, and R. Merletti, "Influence of anatomical, physical, and detection-system parameters on surface EMG," *Biological Cybernetics*, vol. 86, no. 6, pp. 445–456, 2002.

[5] A. Del Vecchio, A. Holobar, D. Falla, F. Felici, R. M. Enoka, and D. Farina, "Tutorial: Analysis of motor unit discharge characteristics from high-density surface EMG signals," *Journal of Electromyography and Kinesiology*, vol. 53, p. 102426, 2020.

[6] Y. Zhao, Z. Liu, J. Yu, S. Jing, H. Li, and M. B. Lopez, "HD-sEMG-CORE: An open-source hybrid network algorithm for efficient compression and accurate restoration of high-density surface electromyography signals," *IEEE Sensors Journal*, vol. 25, no. 3, pp. 5478–5490, 2024.

[7] C. M. Germer, A. Del Vecchio, F. Negro, D. Farina, and L. A. Elias, "Neurophysiological correlates of force control improvement induced by sinusoidal vibrotactile stimulation," *Journal of Neural Engineering*, vol. 17, no. 1, p. 016043, 2020. [Online]. Available: https://doi.org/10.1088/1741-2552/ab5e08

[8] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

Evaluation of a Hybrid RR-HRV–P-Wave Approach for Automatic Atrial Fibrillation Detection in ECG Signals

1st Wendy Tatiana Sierra Cortés
Department of Electronics and Biomedical Engineering
State University of Campinas (Unicamp)
Campinas, Brazil
w258253@dac.unicamp.br

2nd Pedro Xavier de Oliveira
Department of Electronics and Biomedical Engineering
State University of Campinas (Unicamp)
Campinas, Brazil
pedrox@unicamp.br

***Abstract*—Atrial fibrillation (AF) is a common cardiac arrhythmia associated with an increased risk of stroke and heart failure. This study compares two ECG-based approaches for automatic AF detection using 86 records from PhysioNet databases. The first method (RR-HRV method) is based on RR interval variability and heart rate variability features, whereas the second method (Hybrid RR-HRV + P-wave Method) extends the baseline model by incorporating P-wave morphological analysis and an abrupt RR-transition detector. Descriptive comparisons showed higher classification performance for the RR-HRV method, although paired statistical analyses did not indicate a significant difference between the methods at a 5% significance level. A class-specific AFIB analysis showed that the hybrid method increased sensitivity from 80.0% to 83.3%, but also increased false positives and reduced precision from 77.4% to 64.1%. These findings suggest that P-wave and RR-transition features provide complementary information for AF detection; however, further improvements may require more robust integration strategies considering signal quality assessment, ectopy control, and evidence fusion.**

***Index Terms*—Atrial fibrillation, electrocardiogram, RR intervals, P-wave, heart rate variability, biomedical signal processing.**

I. Introduction

Atrial fibrillation (AF) is the most common sustained arrhythmia in the adult population, associated with an increased risk of stroke, heart failure, and cardiovascular mortality [1], [15]. Characterized by chaotic atrial electrical activation, its identification on the electrocardiogram (ECG) is based on RR interval irregularity and the absence of organized P waves [3], [4]. However, definitive diagnosis faces challenges due to the intermittent nature of paroxysmal AF and the presence of noise or artifacts that can lead to misclassifications [2], [5].

Historically, automated detection focused on the temporal dynamics of RR intervals [3], [6], driven by classic QRS complex detection algorithms, such as that of Pan and Tompkins [7]. Recently, deep learning-based approaches and wearable devices have expanded continuous monitoring capabilities with high sensitivity [8], [9]. However, these modern models often require high computational cost and suffer from lower clinical interpretability, hindering their implementation in embedded systems.

A critical limitation of methods restricted to RR intervals is that ventricular beat irregularity, expressed by temporal variation between consecutive R peaks, is not exclusive to AF. Alterations such as ectopy, detection artifacts, and other supraventricular arrhythmias can also produce irregular RR series, increasing the occurrence of false positives in classifiers based solely on ventricular dynamics. To mitigate this issue, hybrid approaches emerge as a promising alternative. Integrating P-wave morphology provides complementary atrial information [1], while statistical (RMSSD, pNN50) and non-linear (approximate entropy) metrics provide a balanced quantification of cardiac rhythm complexity [10], [11].

In this context, this work proposes a modular hybrid system to support automatic AF detection, combining RR interval variability and P-wave morphological analysis. ECG records from public databases available on the PhysioNet platform were used, specifically the MIT-BIH Atrial Fibrillation Database, the MIT-BIH Normal Sinus Rhythm Database, and the MIT-BIH Arrhythmia Database [12], [13]. The main contribution of this study consists of comparing two strategies: the RR-HRV Method, based exclusively on RR intervals, and the Hybrid RR-HRV + P-wave Method, a hybrid approach that incorporates P-wave morphological analysis and the detection of abrupt transitions in RR dynamics.

II. Methodology

This study presents the implementation and comparison of two strategies for automatic AF detection in multichannel electrocardiogram (ECG) signals provided by PhysioNet, using the first available channel of each record for processing. The first approach, referred to as the RR-HRV Method, is based exclusively on RR interval variability. The second approach, the Hybrid RR-HRV + P-wave Method, extends the former by incorporating P-wave morphological analysis and an additional module for detecting abrupt transitions in

This work was supported by the Fundo de Apoio ao Ensino, à Pesquisa e à Extensão (FAEPEX) of Universidade Estadual de Campinas (UNICAMP), grant number 3325/24.

RR interval dynamics. The primary objective is to evaluate whether incorporating morphological evidence associated with the presence or organized absence of the P wave enhances the discriminative capability of a system based on RR interval variability

The evaluated methods do not employ supervised learning, training stages, cross-validation, or model tuning. These approaches are deterministic, relying on digital signal processing, handcrafted extraction of physiological heart rate variability features, and heuristic decision rules

A. Databases and Standardization

Records from PhysioNet databases were used, including the MIT-BIH Atrial Fibrillation Database, the MIT-BIH Normal Sinus Rhythm Database, and the MIT-BIH Arrhythmia Database [12], [13]. The signals were resampled to 250 Hz to ensure temporal consistency across records and allow uniform application of the methods. The flowchart of the system for automatic AF detection is presented in Figure 1.

In this study, an experimental reorganization of the original annotations was performed. The AFIB label was used for AF, N for normal sinus rhythm, while other arrhythmic events present in the MIT-BIH Arrhythmia Database were grouped as other arrhythmias (OA). The N-OA class was assigned to records with a predominantly sinus rhythm, but containing episodes of other arrhythmias or events not classified as AF, according to the original annotations of the databases used. The final evaluation set comprised 86 records, organized into four target classes: 36 N-OA records, 17 N records, 30 AFIB records, and 3 OA records. This organization enabled evaluating the behavior of the methods both on records with AF and on pure sinus signals, other arrhythmias, and predominantly sinus records with associated arrhythmias.

B. Preprocessing, R-Peak Detection, and RR Interval Calculation

As presented in Figure 1, preprocessing constitutes the first step of the methodological workflow, serving as the basis for R-peak detection and the subsequent calculation of RR intervals. ECG signal preprocessing was performed using second-order bandpass Butterworth filters. For general inspection of the electrocardiographic trace, a frequency band of 0.5–35 Hz was used to reduce baseline wander and high-frequency noise while preserving the main morphological components of the ECG. For QRS complex detection, a more selective band of 5–18 Hz was employed, suitable for highlighting ventricular activity associated with R peaks.

R-peak detection was performed on the filtered signal, using differentiation, squaring, and moving window smoothing of 120 ms. Next, an adaptive threshold defined by the median of the processed signal plus 1.5 times the standard deviation was applied. To avoid multiple detections associated with the same QRS complex, a refractory period of 250 ms was

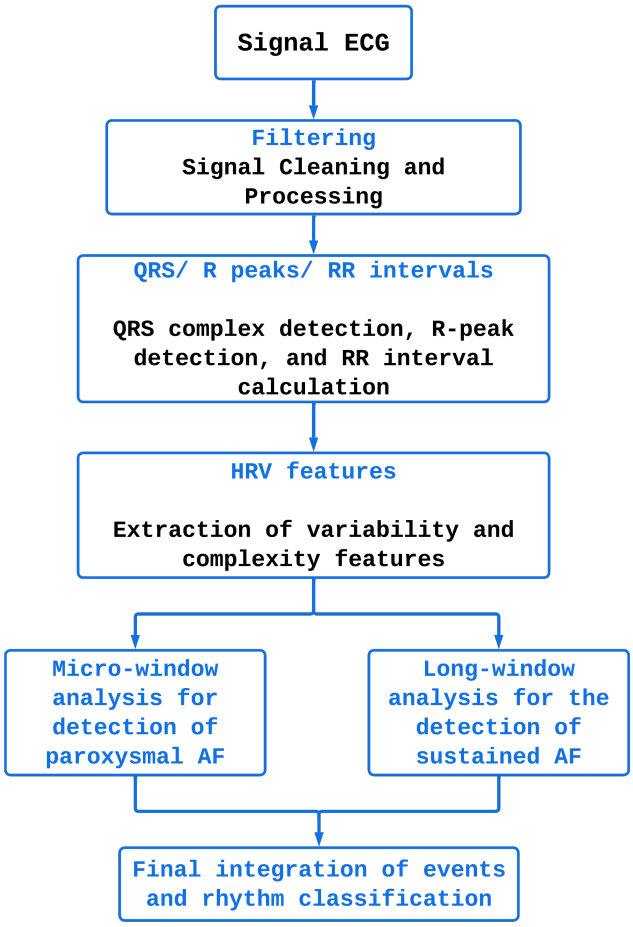


Fig. 1. Methodological architecture of the system for automatic atrial fibrillation detection. The RR-HRV Method uses features derived from RR intervals, whereas the Hybrid RR-HRV + P-wave Method additionally incorporates P-wave morphological analysis and detection of abrupt transitions in the RR series.

imposed. Subsequently, the detected peaks were refined by local search within a window of ± 80 ms, in order to locate more precisely the maximum amplitude point corresponding to the R peak.

After identifying the R peaks, the RR intervals were calculated as the time difference between two consecutive R peaks:

$$RR_i = t_{R_i} - t_{R_{i-1}} \tag{1}$$

After that, the RR intervals were subjected to a physiological validation criterion, keeping only values between 0.25 s and 2.5 s, a range approximately compatible with heart rates between 24 and 240 bpm. Intervals outside this limit were considered likely artifacts, R-peak detection failures, or events not physiologically plausible for the proposed analysis.

C. RR Interval Feature Extraction

Variability and complexity metrics were extracted from the RR series: RMSSD, pNN50, coefficient of variation (CV), approximate entropy (ApEn), lag-1 autocorrelation, and Turning Point Ratio (TPR).

These metrics were selected for their ability to represent different aspects of temporal RR interval irregularity, a relevant feature in automatic AF detection.

The coefficient of variation was used as a normalized measure of dispersion for the RR intervals, defined by:

$$CV = \frac{\sigma_{RR}}{\mu_{RR}}. \tag{2}$$

RMSSD quantifies the magnitude of successive differences between consecutive RR intervals, being sensitive to rapid heart rate variations:

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2}. \tag{3}$$

pNN50 was calculated as the proportion of consecutive RR interval pairs whose absolute difference exceeds $50\,\mathrm{ms}$, or $0.05\,\mathrm{s}$, when intervals are expressed in seconds:

$$pNN50 = \frac{1}{N-1}\sum_{i=1}^{N-1} I\left(|RR_{i+1} - RR_i| > 0.05\right), \tag{4}$$

where $I(\cdot)$ represents the indicator function, taking a value of 1 when the condition is true and 0 otherwise. When expressed as a percentage, this value is multiplied by 100.

To evaluate the unpredictability of the RR series, approximate entropy was used, calculated over the previously validated RR intervals. In this work, $m = 2$ and $r = 0.2\sigma_{RR}$ were adopted, where σ_{RR} represents the standard deviation of the analyzed RR series in the window.

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r). \tag{5}$$

Higher ApEn values indicate greater irregularity and lower temporal predictability of the sequence of RR intervals.

The lag-1 autocorrelation (ρ_1) was employed to estimate the temporal dependence between consecutive RR intervals:

$$\rho_1 = \frac{\sum_{i=1}^{N-1}(RR_i - \mu_{RR})(RR_{i+1} - \mu_{RR})}{\sum_{i=1}^{N}(RR_i - \mu_{RR})^2}. \tag{6}$$

High values of ρ_1 indicate greater similarity between consecutive intervals, while reduced values suggest greater temporal irregularity.

Finally, the Turning Point Ratio was used to quantify the frequency of local trend changes in the RR series. A point RR_i is considered a turning point when it represents a local maximum or minimum relative to its immediate neighbors. TPR was defined as:

$$TPR = \frac{1}{N-2}\sum_{i=2}^{N-1} I\left[(RR_i - RR_{i-1})(RR_i - RR_{i+1}) > 0\right]. \tag{7}$$

Thus, the combination of these metrics allows characterizing different dimensions of cardiac dynamics: global dispersion of RR intervals, beat-to-beat variations, temporal irregularity, unpredictability, and local trend changes.

Candidate AF episodes were identified from the combination of RR-HRV features using fixed thresholds. These values were defined based on literature references on heart rate variability [16], [17] and established during the initial implementation of the method, prior to the experimental evaluation presented in this study. Table I presents the thresholds used in the decision stage.

TABLE I
THRESHOLDS USED IN THE DECISION STAGE BASED ON RR-HRV FEATURES.

Feature	Thresholds used
CV	0.12 (minimum), 0.17 (high)
RMSSD	0.09 s (minimum), 0.13 s (high)
pNN50	0.30 (minimum)
ApEn	0.14 (minimum), 0.20 (high)

This quantitative representation provides the basis for classifying AF episodes in subsequent system stages.

D. Analysis Windows

(i) Long-window analysis for sustained AF detection: evaluation of signals in 30 s windows, with a 10 s step, using RR-HRV features to identify prolonged episodes of irregularity consistent with sustained AF.

(ii) Micro-window analysis for paroxysmal AF detection: evaluation in 12 s micro-windows, with a 3 s step, complemented by beat-to-beat analysis, allowing the identification of short and transient AF episodes that might not be detected in longer time windows.

E. Final Event Integration and Rhythm Classification

To obtain the final AF classification (AFIB), the RR-HRV Method considers the combination of features extracted from the RR series and their respective decision thresholds presented in Table I. When all features meet the defined criteria, the record is classified as AFIB. Otherwise, records are classified considering the OA or N categories according to the number of criteria met.

After detecting events at different temporal scales, individual window results were integrated into continuous intervals of AF events. Consecutive windows classified as compatible with AF were grouped considering the temporal proximity between events, avoiding multiple counts of the same episode. From this aggregation, each record received a final classification subsequently used in the comparison with

the reference annotations from the PhysioNet databases.

To attempt to improve the RR-HRV Method, we added a P-wave analysis, forming a Hybrid Method (RR-HRV + P-wave). The P-wave analysis uses a pre-R-wave window between $-240\,\mathrm{ms}$ and $-60\,\mathrm{ms}$; the P wave was detected using a simple threshold. The ECG signal used in this step was filtered in the 0.5–$12\,\mathrm{Hz}$ range.

In this method, the final AF classification was obtained with the same features as the RR-HRV Method, adding P-wave information as a complementary criterion. The P-wave analysis acts as an additional adjustment in the system decision, increasing or reducing evidence of AF according to the pattern observed in the signal.

F. Statistical Analysis

Statistical evaluation was performed by comparing the classifications obtained by the proposed methods with the reference annotations from the PhysioNet databases. To evaluate classifier performance, confusion matrix elements were considered: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). From these values, precision, sensitivity, and F1-score metrics were calculated.

Precision represents the proportion of positive classifications correctly identified as atrial fibrillation,

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

while sensitivity indicates the ability of the system to identify AF records present in the reference annotation.

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

The F1-score was used to represent the balance between precision and sensitivity.

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Sensitivity}}{\mathrm{Precision} + \mathrm{Sensitivity}} \tag{10}$$

Statistical analyses were performed using Jamovi software. Initially, the Shapiro-Wilk test was applied to check the normality of paired differences between the methods. Since the data did not follow a normal distribution, the Wilcoxon test for paired samples was used to evaluate whether there were statistically significant differences in performance between the two approaches. In addition, McNemar's test was applied to compare the proportions of correct classifications between the methods, considering the binary and paired nature of the results.

III. Results and Discussion

The experimental evaluation was performed on the same set of 86 records, allowing a direct comparison of the performance of the RR-HRV Method and the Hybrid RR-HRV + P-wave Method under the same conditions of preprocessing, segmentation, and feature extraction.

A. Overall Performance

Each ECG record was considered as an independent experimental unit for statistical and performance analyses. For each of the 86 evaluated records, the variable *Correct* was defined, associated with the outcome of each classification method, taking a value of 1 when the obtained classification matched the PhysioNet reference annotation and a value of 0 when a disagreement occurred. This variable allowed a paired comparison of performance between the RR-HRV and Hybrid RR-HRV + P-wave Methods. Formally, for each record i and method $k \in \{\text{RR-HRV}, \text{RR-HRV+P-wave}\}$, it was defined as:

$$\mathrm{Correct}_{i,k} = \begin{cases} 1, & \text{when } \hat{y}_{i,k} = y_i, \\ 0, & \text{when } \hat{y}_{i,k} \neq y_i, \end{cases} \tag{11}$$

where y_i represents the PhysioNet reference annotation and $\hat{y}_{i,k}$ represents the classification obtained by method k for record i.

The normality of paired differences between *CorrectRR-HRV* and *CorrectRR-HRV+P-wave* was evaluated using the Shapiro-Wilk test. Since the assumption of normality was not met ($W = 0.443$, $p < 0.001$), the Wilcoxon signed-rank test for paired samples was used to check for statistically significant differences in performance between the methods. The test indicated no statistically significant difference between the approaches ($W = 44.0$, $p = 0.066$). As these are paired binary responses, McNemar's test was applied to compare the proportions of correct classifications between the methods (Table II).

TABLE II
PAIRED STATISTICAL ANALYSIS BETWEEN THE RR-HRV METHOD AND THE HYBRID RR-HRV + P-WAVE METHOD.

Analysis	Statistic	p-value	Interpretation
Shapiro-Wilk	$W = 0.443$	< 0.001	Non-normal
Paired Wilcoxon	$W = 44.0$	0.066	Non-significant

B. Paired Comparison by McNemar's Test and Analysis of Contingency Matrices

Since the variables *CorrectRR-HRV* and *CorrectRR-HRV+P-wave* constitute binary and paired outcomes, in addition to the statistical analyses presented previously, McNemar's test was applied, which is specifically indicated for comparing proportions of responses in paired data. For this analysis, the classification of each method on each record was coded as correct (1) or incorrect (0). Since the RR-HRV Method and the Hybrid RR-HRV + P-wave Method were evaluated on the same 86 records, the test made it possible to verify whether the overall proportions of correct classifications differed significantly between the two approaches.

In the paired comparison, both methods correctly classified 62 records and incorrectly classified 14 records. In 8 cases, only the RR-HRV Method produced a correct classification,

whereas in 2 cases only the Hybrid RR-HRV + P-wave Method was correct. Thus, the RR-HRV Method obtained 70 correct classifications, corresponding to 81.4% of the records, while the Hybrid RR-HRV + P-wave Method presented 64 correct classifications, equivalent to 74.4%.

As presented in Table III, the uncorrected McNemar test yielded a result close to the significance level, but not significant ($\chi^2(1) = 3.60$, $p = 0.058$). The exact analysis, considered especially due to the small number of discordant pairs, also showed no statistically significant difference between the methods ($\ln(OR) = 1.39$, $p = 0.109$). Therefore, although there is a descriptive trend favoring the RR-HRV Method, it was not possible to confirm its statistical superiority at the 5% significance level.

TABLE III
MCNEMAR TEST RESULT FOR THE PAIRED COMPARISON OF CORRECT CLASSIFICATIONS BETWEEN THE RR-HRV METHOD AND THE HYBRID RR-HRV + P-WAVE METHOD.

Analysis	Value	df	p-value
McNemar (asymptotic)	$\chi^2 = 3.60$	1	0.058
Exact McNemar	$\ln(OR) = 1.39$	–	0.109

$N = 86$.

Table IV presents the contingency matrices of the RR-HRV Method and the Hybrid RR-HRV + P-wave Method. Although the reference classes were organized into four categories (N-OA, N, AFIB, and OA), the analysis of method outputs focused on the detection of AF-compatible events. Thus, the records were compared based on the presence or absence of AF, while the remaining categories were used to characterize the different rhythmic scenarios present in the databases.

This analysis highlights an important shift in the decision profile of the hybrid method. Under the RR-HRV Method, 31 records were classified as AFIB, of which 24 corresponded to reference AFIB records and 7 to false positives. Under the Hybrid RR-HRV + P-wave Method, 39 records were classified as AFIB, of which 25 corresponded to true AFIB records and 14 to false positives.

TABLE IV
CONTINGENCY MATRICES OF THE RR-HRV METHOD AND THE HYBRID RR-HRV + P-WAVE METHOD. ROWS: REFERENCE CATEGORIES; COLUMNS: CATEGORIES ASSIGNED BY THE METHODS.

Method	Reference	Classification assigned by method			
		N	OA	AFIB	Total
RR-HRV	N-OA	23	10	3	36
RR-HRV	N	13	1	3	17
RR-HRV	AFIB	1	5	24	30
RR-HRV	OA	1	1	1	3
	Total	38	17	31	86
RR-HRV + P-wave	N-OA	21	7	8	36
RR-HRV + P-wave	N	11	1	5	17
RR-HRV + P-wave	AFIB	1	4	25	30
RR-HRV + P-wave	OA	1	1	1	3
	Total	34	13	39	86

Considering the AFIB class as the main target, the RR-HRV Method showed 24 true positives, 7 false positives, and 6 false negatives. In contrast, the Hybrid RR-HRV + P-wave Method showed 25 true positives, 14 false positives, and 5 false negatives. Thus, the hybrid model recovered one additional AFIB case and slightly reduced false negatives; however, it doubled the number of false positives associated with the AFIB class.

This redistribution explains the observed decrease in the overall performance of the Hybrid RR-HRV + P-wave Method and, in particular, the decline in precision for the AFIB class. The Hybrid RR-HRV + P-wave Method became more permissive in classifying records as AFIB, especially in N-OA and N records. Specifically for the AFIB class, sensitivity increased from 80.0% to 83.3%, while precision dropped from 77.4% to 64.1%. Therefore, the local gain in sensitivity was accompanied by an increase in false positives and a reduction in precision, indicating that incorporating the P-wave and the RR-transition detector modified the decision profile of the system without yielding an overall improvement in the evaluated configuration.

C. AFIB Class-Specific Performance

Table V presents the performance metrics of the RR-HRV Method and the Hybrid RR-HRV + P-wave Method for the AFIB group ($N = 30$). This analysis allows for an isolated evaluation of the behavior of both methods on records containing AF.

The Hybrid RR-HRV + P-wave Method showed an increase in sensitivity for the AFIB class, accompanied, however, by a higher number of false positives and a reduction in precision. These results indicate a shift in the balance between detection and the occurrence of false positives after incorporating P-wave information and RR transitions.

TABLE V
PERFORMANCE OF THE RR-HRV AND HYBRID RR-HRV + P-WAVE METHODS FOR THE AFIB CLASS.

Metric ($N = 30$)	RR-HRV	RR-HRV + P-wave
True positives	24	25
False negatives	6	5
False positives	7	14
Precision	77.4%	64.1%
Sensitivity	80.0%	83.3%
F1-score	78.7%	72.4%

Thus, the incorporation of the P wave and the RR-transition detector modified the decision profile of the hybrid method, favoring the identification of AFIB records, but with a higher occurrence of incorrect positive classifications.

D. Analysis of Error Mechanisms Associated with the Hybrid Method

The use of the P wave as a complementary criterion is clinically justified. In sinus rhythm, the P wave reflects coordinated atrial depolarization and tends to precede the QRS complex in an organized manner. In AF, this atrial organization is lost, and the absence of an organized P wave, associated with RR interval irregularity, constitutes a relevant

electrocardiographic criterion for arrhythmia diagnosis [1], [14].

Thus, the problem observed in the Hybrid RR-HRV + P-wave Method lies not in the choice of the P wave as a physiological marker, but in how this information was incorporated into the decision-making process. The analysis of patterns observed in error cases suggests three mechanisms associated with the performance of the Hybrid RR-HRV + P-wave Method:

(i) Predominance of sinus rhythm in paroxysmal records: In records containing paroxysmal AF episodes, the predominant presence of sinus rhythm periods hindered the overall classification of the record. This characteristic led to incorrect classifications in some cases, as short AF episodes may have been insufficient to alter the predominant pattern observed in the signal.

(ii) Influence of signal morphological quality: In records with a higher presence of noise or morphological changes, the complementary atrial information presented limitations. In these situations, residual components of the QRS complex or artifacts in the PR segment may have contributed to incorrect classifications, reducing the reliability of P-wave-based evidence.

(iii) Limitations of local temporal analysis: Analysis based on local transitions between windows presented limitations in records with abrupt rhythm changes, ectopies, or artifacts. In these situations, transient changes in RR variability can present behavior similar to the patterns expected during AF episodes, contributing to the increase in false positives in the hybrid method.

E. General Interpretation and Comparison with Literature Approaches

The obtained results should be interpreted considering the differences among databases and validation scenarios used in the literature. Several studies on automatic AF detection range from classical methods based on RR interval analysis, such as the one proposed by Tateno and Glass [6], to recent approaches based on machine learning, wearable systems, and continuous monitoring [5], [8]. These works frequently use public databases specifically oriented toward AF identification, such as the MIT-BIH Atrial Fibrillation Database, in which records contain sustained and paroxysmal arrhythmia episodes with rhythmic annotations used as clinical reference. This type of database is necessary to evaluate the sensitivity of methods in the presence of AF, but it may not fully represent situations where the arrhythmia appears associated with other rhythmic patterns or more heterogeneous signals.

In this study, in addition to the MIT-BIH Atrial Fibrillation Database, records from the MIT-BIH Arrhythmia Database and the MIT-BIH Normal Sinus Rhythm Database were used to expand the variability of the analyzed signals. This choice made the classification task more challenging, as the records can contain sinus rhythm, AF, other arrhythmias, ectopies, and artifacts. For this reason, the target classes were organized into N, AFIB, OA, and N-OA, allowing the differentiation of pure sinus records from those predominantly sinus with other associated arrhythmias.

This methodological choice limits direct comparison, in absolute terms, with studies evaluated in less heterogeneous scenarios or focused exclusively on AF-specific databases. Classical methods based on RR intervals, such as those proposed by Moody and Mark [3] and Tateno and Glass [6], demonstrated that RR interval irregularity constitutes a relevant feature for automatic AF detection. Subsequently, Tateno and Glass explored metrics based on RR and ΔRR intervals, reporting a sensitivity of 86.6% for the RR interval coefficient of variation and 94.4% for the Kolmogorov-Smirnov test applied to ΔRR intervals on the MIT-BIH Atrial Fibrillation Database. However, when applying the method to the MIT-BIH Arrhythmia Database, the authors observed a performance drop associated with an increase in false positives in records with complex arrhythmias and premature ventricular contractions.

In more recent approaches, studies such as Xie et al. [9] indicate that machine learning and deep learning models can achieve high sensitivities in AF detection. However, these models generally involve higher computational complexity and lower physiological interpretability when compared to strategies based on explicit RR interval and atrial morphology features.

This behavior is consistent with the findings of the present study. Although the Hybrid RR-HRV + P-wave Method increased sensitivity for AFIB from 80.0% to 83.3%, this gain was accompanied by a reduction in precision from 77.4% to 64.1%, mainly due to an increase in false positives in N-OA and N records. The performance of the RR-HRV Method indicates that features based on RR intervals demonstrated adequate capacity for AF detection even across different rhythmic scenarios. However, the inclusion of the P wave and RR transitions increased the identification of some AF cases, but also generated more incorrect classifications in records without AF.

In summary, comparison with the literature suggests that hybrid architectures can be promising, but their effectiveness depends on more rigorous fusion strategies between temporal and morphological evidence. In particular, future versions should incorporate additional signal quality criteria, multiple physiological evidences, and decision mechanisms capable of reducing false positives in records with complex arrhythmias.

IV. Conclusion

This study presented a comparative analysis between an automatic AF detection method based exclusively on RR interval variability (RR-HRV Method) and a hybrid approach that incorporates P-wave morphological analysis and abrupt transitions in the RR series (Hybrid RR-HRV + P-wave Method). The results indicated that the RR-based approach showed better overall descriptive performance in precision, sensitivity, and F1-score; however, the paired statistical analysis showed no significant difference between the methods.

Although the inclusion of atrial information is clinically justified, its implementation in the Hybrid RR-HRV + P-wave Method did not produce an aggregate improvement in the evaluated configuration. The increase in the identification of AFIB records was accompanied by a higher number of incorrect positive classifications, indicating that the integration of additional information still requires adjustments.

It is concluded that the P wave should be used as a complementary criterion to the analysis of RR intervals, contributing to a decision based on multiple physiological indicators, rather than as an isolated rule for exclusion or confirmation. These findings provide guidelines for future versions of the system, in which the integration among RR variability, atrial morphology, signal quality control, and multiple physiological evidences may favor a better balance among sensitivity, precision, and clinical interpretability, especially in records with complex arrhythmias.

References

[1] G. Hindricks *et al.*, "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS)," *Eur. Heart J.*, vol. 42, no. 5, pp. 373–498, Feb. 2021.

[2] J. Mbroh and S. Poli, "2021 European Heart Rhythm Association Practical Guide on the use of non-vitamin K antagonist oral anticoagulants in patients with atrial fibrillation: comment," *Europace*, vol. 23, no. 10, p. 1685, 2021.

[3] G. B. Moody and R. G. Mark, "A new method for detecting atrial fibrillation using R-R intervals," *Comput. Cardiol.*, vol. 10, pp. 227–230, 1983.

[4] J. Vogler *et al.*, "Electrocardiographic features of atrial fibrillation," *Herz*, vol. 37, no. 1, pp. 12–18, 2012.

[5] A. Jahan *et al.*, "Challenges in automated atrial fibrillation detection from ECG signals: A review," *Biomed. Signal Process. Control*, vol. 71, p. 103001, 2022.

[6] K. Tateno and L. Glass, "Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and delta RR intervals," *Med. Biol. Eng. Comput.*, vol. 39, no. 6, pp. 664–671, Nov. 2001.

[7] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[8] M. Shao *et al.*, "Portable electrocardiogram monitoring system for atrial fibrillation detection," *Sensors*, vol. 20, no. 3, p. 606, Jan. 2020.

[9] Y. Xie *et al.*, "Deep learning architectures for continuous monitoring of atrial fibrillation in wearable devices," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 2, pp. 450–462, 2024.

[10] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Nat. Acad. Sci.*, vol. 88, no. 6, pp. 2297–2301, Mar. 1991.

[11] A. Parsi *et al.*, "Nonlinear heart rate variability features for atrial fibrillation detection," *Comput. Methods Programs Biomed.*, vol. 200, p. 105844, 2021.

[12] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[13] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May/Jun. 2001.

[14] C. T. January *et al.*, "2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation," *Circulation*, vol. 140, no. 2, pp. e125–e151, 2019.

[15] Organização Mundial da Saúde. (2019). *Cardiovascular Diseases*. Geneva.

[16] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation and clinical use," *European Heart Journal*, vol. 17, pp. 354–381, 1996.

[17] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, art. 258, 2017.

Evaluation of Movement Smoothness in the Four-Way Leaning Limits of Stability Test in Patients with TMD: A Pilot Study

Mylena Oliveira Viana Montero
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
mylenavianam@gmail.com

José Jair Alves Mender Júnior
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil

***Abstract*—Static postural control (PC) involves maintaining postural stability while standing on a fixed surface, whereas dynamic postural control (DPC) requires maintaining or recovering stability during movement or in response to external perturbations. Temporomandibular Disorders (TMD) encompass a range of musculoskeletal and neuromuscular conditions involving the temporomandibular joint, masticatory muscles (MM), and associated structures. This pilot study evaluated movement smoothness using the Log Dimensionless Jerk (LDLJ) and directional control during the Four-Way Leaning Limits of Stability (LOS) test in patients with TMD under different visual, surface, and MM maximal voluntary contraction (MVC) conditions. Dynamic postural performance was strongly influenced by sensory conditions, particularly support surface and visual input. Large effects were observed for support surface and vision on movement smoothness, whereas masticatory muscle MVC showed no significant main effect on either mean velocity or LDLJ. However, moderate effects were observed for interactions involving contraction, suggesting that the effect of masticatory muscle activation may depend on sensory conditions.**

***Index Terms*—Temporomandibular Disorders, Dynamic Postural Control, Limits of Stability, Movement Smoothness, Log Dimensionless Jerk.**

I. Introduction

Static Postural Control (PC) involves maintaining postural stability while standing on a fixed surface. Dynamic Postural Control (DPC) requires maintaining or recovering stability during movement or in response to external perturbations, such as standing on unstable surfaces or walking [11].

Temporomandibular Disorder (TMD) is an umbrella term encompassing a group of musculoskeletal and neuromuscular conditions involving the Temporomandibular Joint (TMJ), masticatory muscles (MM), and associated structures [12]. The Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) are widely used and provide a standardized approach for classifying these conditions, covering both pain-related and intra-articular disorders (e.g., disc displacement, degenerative joint disease) [7], [12].

Postural control results from the continuous integration of information from visual, vestibular, and proprioceptive systems by the central nervous system, enabling the generation of motor responses that maintain the center of mass within the base of support. Proprioception constitutes the primary source of information during upright posture, providing data on body position and movement through mechanoreceptors in muscles, tendons, and joints. The visual system contributes to postural stabilization by detecting environmental references and distinguishing self-motion from external movement. The relative contribution of each system varies according to the task and the reliability of available sensory information, with multisensory integration enabling adaptations and compensations when one system is altered [13].

Current evidence suggests that TMDs may be associated with alterations in postural stability; however, the relationship with DPC is less established and appears to depend on the specific TMD subtype and assessment methods employed [3], [10], [14]. Studies demonstrate that individuals with myogenic TMD show increased sway area and velocity during posturostabilometric tests, indicating impaired postural stability compared to healthy controls [9]. However, these findings predominantly pertain to static balance evaluations rather than DPC tasks.

Additionally, jaw clenching and MM contraction, which are relevant in the context of TMD, have been associated with acute improvements in steady-state DPC performance in healthy adults, although this effect has not been directly investigated in populations with TMD or self-reported bruxism. Available literature supports a potential association between TMD and changes in DPC, particularly through mechanisms involving occlusal changes and neuromuscular control [2], [4]–[6]. Nevertheless, additional studies are required to elucidate the causality and clinical relevance of this relationship.

The primary objective of this pilot study was to estimate the effect size associated with maximal voluntary contraction (MVC) of the MM during the Four-Way Leaning LOS test to guide sample size calculations for a future adequately powered study. The secondary objective was to explore the influence of MM MVC on DPC performance, focusing primarily on movement smoothness using the Log Dimensionless Jerk (LDLJ) and directional control, while also examining changes in movement velocity and maximum excursion under different visual and surface conditions. Furthermore, this knowledge

may support therapeutic strategies to improve functionality and quality of life in adults with TMD.

II. Materials and Methods

This was a pilot, quantitative, observational, analytical, and cross-sectional study conducted with adults of both sexes aged between 18 and 60 years. Participants were recruited by convenience sampling through social media announcements and underwent a single in-person evaluation lasting approximately 60 minutes.

The evaluation consisted of physical therapy and dental examinations, standardized questionnaires, and the Four-Way Leaning Limits of Stability (LOS) test. Characterization of the sample regarding the presence of TMD was performed by trained researchers using the DC/TMD protocol, including an interview and standardized clinical examination.

Dynamic PC was evaluated using the Four-Way Leaning LOS test performed on a force platform (AMTI, USA) with a sampling frequency of 100 Hz, and signal processing was conducted via NetForce software (AMTI, USA). During the test, participants stood upright and were instructed to voluntarily displace their center of gravity as far as possible in the anterior, right lateral, posterior, and left lateral directions, keeping their feet fixed on the platform without losing balance. The protocol evaluates the ability to control the center of pressure (COP) and center of gravity within the limits of postural stability.

The test was performed under eight experimental conditions resulting from the combination of three factors: visual condition (eyes open or closed), support surface stability (rigid platform or TempurMed® foam on the platform), and masticatory muscle activation (habitual condition or MVC). The order of conditions was randomized by drawing lots immediately prior to evaluation. To minimize muscle fatigue, conditions with and without MVC were executed alternately.

Prior to data collection, a standardized demonstration was performed, followed by a practice trial. Subsequently, up to three valid trials were recorded for each experimental condition, and the first valid attempt was considered for this analyses. In case of failure, the reason was documented, and if a participant could not complete the task after three attempts, that condition was considered non-executable.

COP signals exported from the force platform were processed in Python (Python Software Foundation, USA). Signals were initially filtered using a 4th-order low-pass Butterworth filter with a cutoff frequency of 6 Hz, following recommendations by Miller et al. [8]. Center of gravity (COG) trajectory was then estimated from filtered COP signals using a 4th-order low-pass Butterworth filter with a 0.5 Hz cutoff frequency, as described by Benda et al. [1]. Each trial was visually inspected to manually identify the start and end of the task, as well as transitions between directions, allowing movement segmentation for feature extraction.

From the segmented signals, variables of interest were calculated, including maximum COG excursion in anteroposterior and mediolateral directions, mean displacement velocity, Directional Control, and Log Dimensionless Jerk (LDLJ), the latter serving as an indicator of movement smoothness and coordination.

Descriptive analysis was conducted in Python. Given the pilot nature of the study, interpretation prioritized effect magnitudes to inform future sample size calculations. Effects of MVC, visual condition, and surface stability, along with their interactions, were evaluated using repeated-measures statistical models, reporting effect sizes and confidence intervals where applicable. Intra-individual differences between habitual and MVC conditions ($\Delta = \mathrm{MVC} - \mathrm{Habitual}$) were visualized using boxplots and paired representations. To evaluate movement smoothness ($\mathrm{General}_L DLJ; lowervaluesindicatesmoothermovement), a2 \times 2 \times 2$ three-way repeated-measures ANOVA was conducted with Vision (Open vs. Closed), Surface (Stable vs. Unstable), and Contraction (Habitual vs. Voluntary) as within-subjects factors, assessing all main effects and their two-way and three-way interactions. Data normality was verified prior to analysis using the Shapiro-Wilk test, applying logarithmic or square-root transformations when necessary; post-hoc comparisons were performed using paired t-tests with Bonferroni correction, and effect sizes were reported as partial eta squared (η_p^2).

The activity flowchart is presented in Fig. 1.

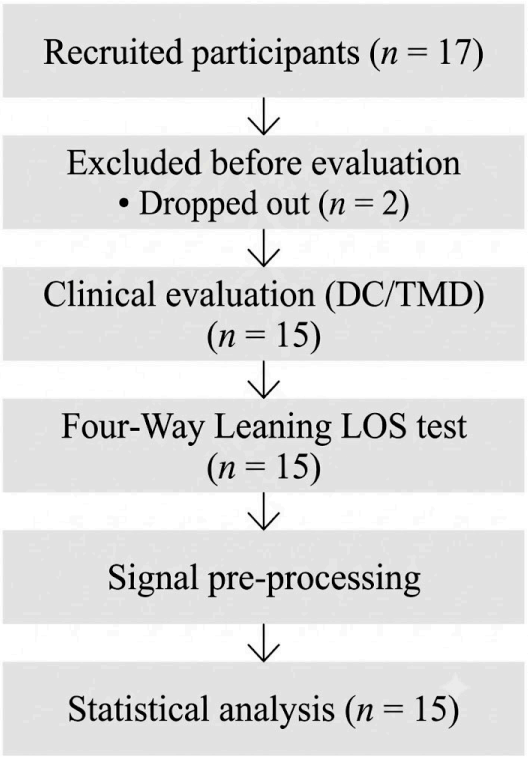


Fig. 1. Flowchart of activities.

III. Results and Discussion

A total of 15 participants (11 female, 4 male) ranging in age from 9 to 52 years were evaluated, with the vast majority presenting with both myalgia and arthralgia associated with TMD

(Table I). Descriptive statistics for all outcome measures across the eight experimental conditions are presented in Table II. Mean LDLJ values ranged from -27.15 to -29.89 across conditions. Lower (more negative) LDLJ values, indicating reduced movement smoothness, were consistently observed during tests on the unstable surface compared to the stable surface.

TABLE I
PARTICIPANT CHARACTERISTICS

Characteristic	Value
Participants	15
Age	Ranging from 9 to 52
Sex	11 Female, 4 Male
TMD Associated Pain	Myalgia e Arthralgia: 12 Myalgia: 2 None: 1

TABLE II
DESCRIPTIVE RESULTS

Condition	Mean Velocity (SD)	LDLJ (COP) Mean (SD)
OESSN	92.88 (18.19)	-27.15 (1.23)
CESSN	86.68 (19.57)	-28.03 (1.54)
OESSC	88.13 (13.22)	-27.48 (1.42)
CESSC	85.28 (13.91)	-28.07 (1.38)
OEUSN	79.25 (14.19)	-29.35 (1.47)
CEUSN	101.43 (19.79)	-29.81 (2.31)
OEUSC	82.67 (15.55)	-28.34 (1.78)
CEUSC	98.48 (19.55)	-29.89 (1.80)

Note: OESSN = open eyes, stable surface, "no contraction"/habitual; CESSN = closed eyes, stable surface, "no contraction"/habitual; OESSC = open eyes, stable surface, "contraction"/MVC; CESSC = closed eyes, stable surface, "contraction"/MVC; OEUSN = open eyes, unstable surface, "no contraction"/habitual; CEUSN = closed eyes, unstable surface, "no contraction"/habitual; OEUSC = open eyes, unstable surface, "contraction"/MVC; CEUSC = closed eyes, unstable surface, "contraction"/MVC.

A three-factor repeated measures ANOVA evaluated the effects of Vision, Surface, and Contraction on mean movement velocity. No significant main effects were found for Vision ($F(1,13) = 3.764$, $p = 0.074$, $\eta_p^2 = 0.225$), Surface ($F(1,13) = 0.847$, $p = 0.374$, $\eta_p^2 = 0.061$), or Contraction ($F(1,13) = 0.373$, $p = 0.552$, $\eta_p^2 = 0.028$). A significant interaction between Vision and Surface was identified ($F(1,13) = 43.727$, $p < 0.001$, $\eta_p^2 = 0.771$). No significant interactions were found for Vision $\times$ Contraction, Surface $\times$ Contraction, or three-factor interactions ($p > 0.05$).

Significant main effects of Vision and Surface were observed for LDLJ (Fig. 2). Vision significantly influenced LDLJ ($F(1,13) = 16.092$, $p = 0.001$, $\eta_p^2 = 0.553$), while Surface showed a larger effect ($F(1,13) = 143.609$, $p < 0.001$, $\eta_p^2 = 0.917$). No significant main effect was identified for Contraction ($F(1,13) = 0.558$, $p = 0.468$, $\eta_p^2 = 0.041$). None of the interaction terms reached statistical significance; however, the Surface $\times$ Contraction interaction showed a moderate estimated effect size ($\eta_p^2 = 0.192$), and the Vision $\times$ Surface $\times$ Contraction interaction demonstrated a moderate-to-large estimated effect size ($\eta_p^2 = 0.261$).

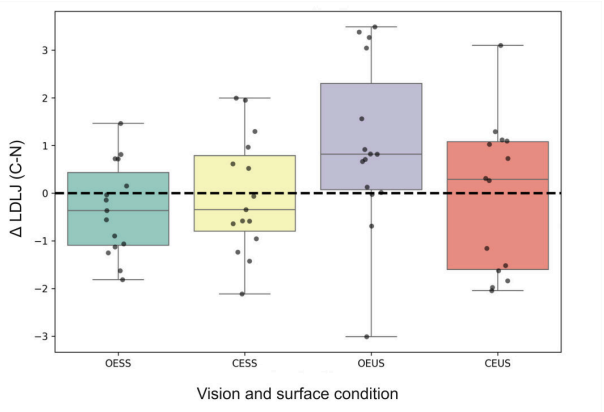


Fig. 2. Intra-individual variation in LDLJ induced by maximal voluntary contraction of the jaw (Δ = MVC − Habitual). Because more negative LDLJ values represent less smooth movement, negative ΔLDLJ values indicate a reduction in movement smoothness following MVC. Dashed horizontal line at $\Delta = 0$ represents no change. OESS = open eyes, stable surface; CESS = closed eyes, stable surface; OEUS = open eyes, unstable surface; CEUS = closed eyes, unstable surface.

Additionally, directional control was analyzed in each direction and condition, yielding no significant differences (Fig. 3, Fig. 4, Fig. 5, Fig. 6). Similarly, an analysis of LDLJ in each condition, visualized using paired line plots, also showed no significant differences (Fig. 7). Directional control was examined separately for the anteromedial (AM), right mediolateral (RML), posteromedial (PM), and left mediolateral (LML) directions under each vision and surface condition (Figs. 3–6).

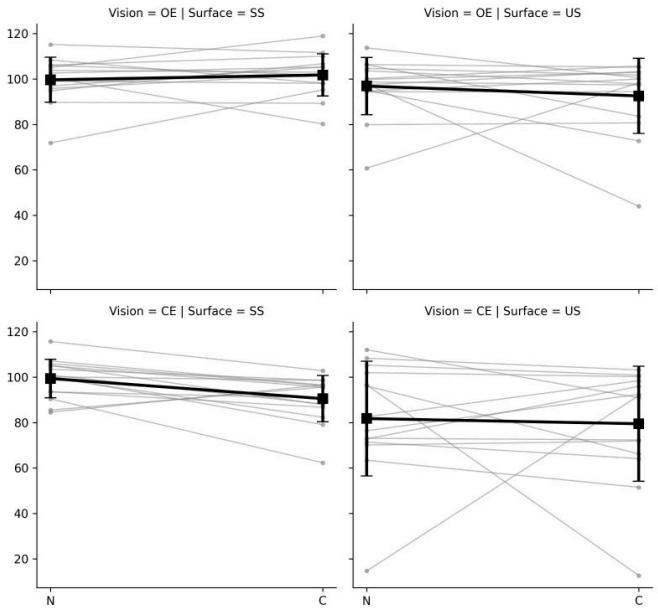


Fig. 3. Directional control analysis (Anteromedial or AM direction). OESS = open eyes, stable surface; CESS = closed eyes, stable surface; OEUS = open eyes, unstable surface; CEUS = closed eyes, unstable surface; N = "No contraction" or Habitual; C: "Contraction" or MVC.

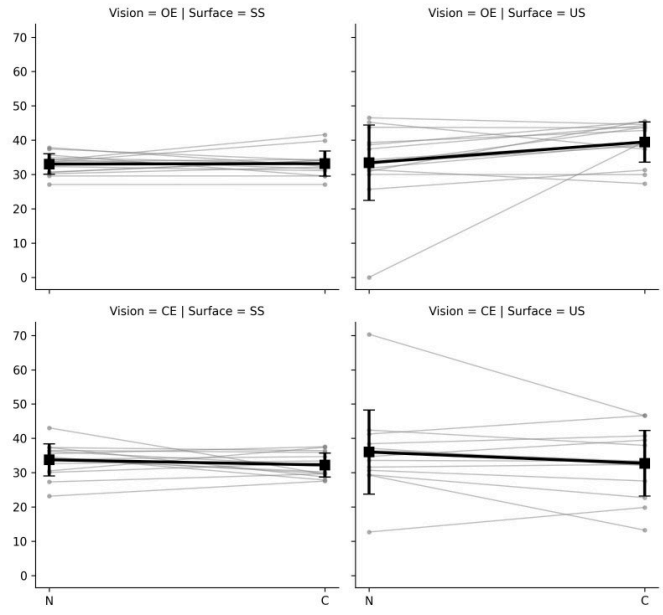


Fig. 4. Directional control analysis (Right Mediolateral or RML direction). OESS = open eyes, stable surface; CESS = closed eyes, stable surface; OEUS = open eyes, unstable surface; CEUS = closed eyes, unstable surface; N = ”No contraction” or Habitual; C: ”Contraction” or MVC.

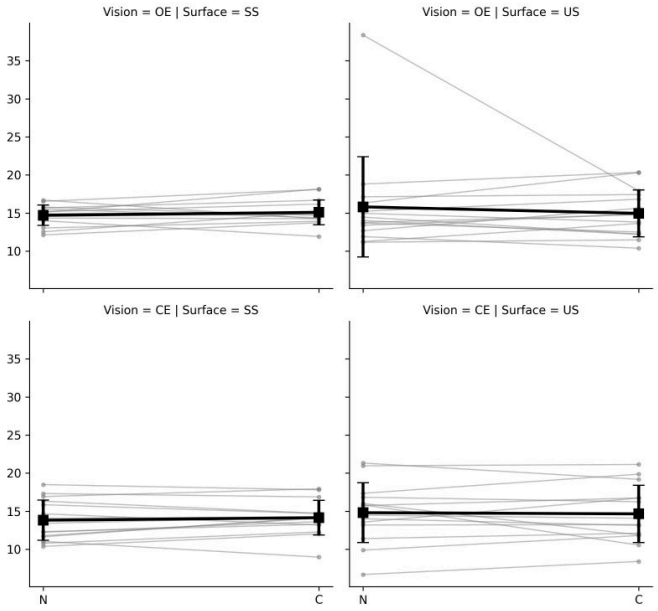


Fig. 6. Directional control analysis (Left Mediolateral or LML Direction). OESS = open eyes, stable surface; CESS = closed eyes, stable surface; OEUS = open eyes, unstable surface; CEUS = closed eyes, unstable surface; N = ”No contraction” or Habitual; C: ”Contraction” or MVC.

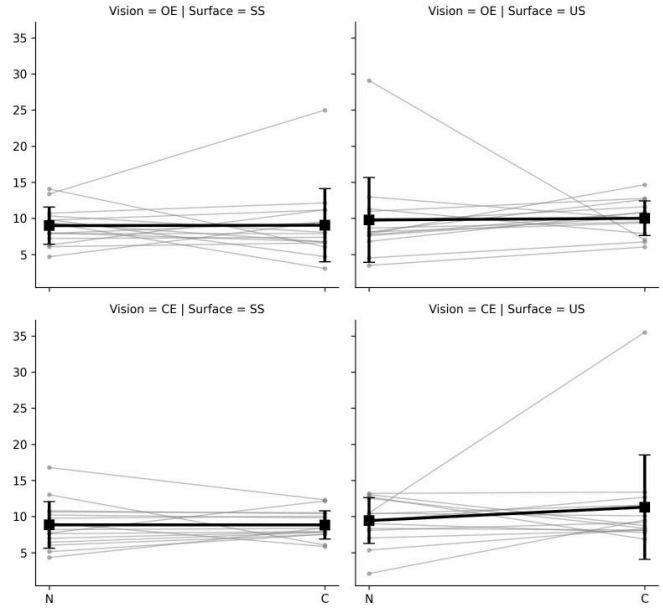


Fig. 5. Directional control analysis (Posteromedial or PM direction). OASE = open eyes, stable surface; OESS = open eyes, stable surface; CESS = closed eyes, stable surface; OEUS = open eyes, unstable surface; CEUS = closed eyes, unstable surface; N = ”No contraction” or Habitual; C: ”Contraction” or MVC.

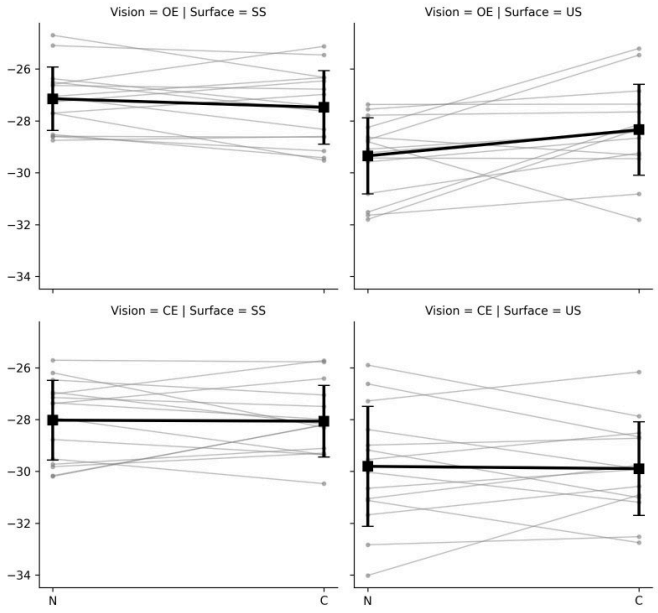


Fig. 7. Paired line plots illustrating LDLJ across all experimental conditions. OESS = open eyes, stable surface; CESS = closed eyes, stable surface; OEUS = open eyes, unstable surface; CEUS = closed eyes, unstable surface; N = ”No contraction” or Habitual; C: ”Contraction” or MVC.

Effect size estimation was the primary objective of this pilot study. Partial eta squared values ranged from negligible ($\eta_p^2 = 0.028$) to very large ($\eta_p^2 = 0.917$). The largest estimated effects were observed for Surface on LDLJ ($\eta_p^2 = 0.917$), Vision on LDLJ ($\eta_p^2 = 0.553$), and the Vision $\times$ Surface interaction on mean velocity ($\eta_p^2 = 0.771$).

The large effects observed for Surface on LDLJ ($\eta_p^2 = 0.917$), Vision on LDLJ ($\eta_p^2 = 0.553$), and the Vision $\times$ Surface interaction on mean velocity ($\eta_p^2 = 0.771$) indicate that sensory manipulation should be carefully considered in the design and sample size estimation of future studies. In contrast, the small effect associated with the main effect of contraction on LDLJ ($\eta_p^2 = 0.041$) suggests that substantially larger samples may be required to detect a clinically or statistically meaningful effect of masticatory muscle activation, if such an effect exists. First, the absence of a matched healthy control group prevents determining whether the observed responses are specific to individuals with TMD. Future studies should include matched controls to distinguish disease-related alterations from normal responses to sensory manipulation and masticatory muscle activation. Second, the relatively small sample size precluded robust subgroup analyses according to TMD subtype and severity. Future adequately powered studies should consider stratification by TMD diagnosis and clinical severity, as different clinical phenotypes may present distinct neuromuscular and postural responses. Third, movement segmentation was performed manually based on visual inspection of the COP/COG trajectories. Although this approach allowed identification of task phases, it may introduce observer-dependent variability. Future studies should investigate automated segmentation methods and/or establish inter- and intra-rater agreement for the identification of movement onset, termination, and directional transitions.

IV. Conclusion

In this pilot sample, dynamic postural performance showed large effects associated with sensory conditions, particularly support surface and visual input. Masticatory muscle MVC did not show a significant main effect on mean velocity or movement smoothness. However, moderate effect sizes for interactions involving contraction suggest that its contribution may depend on sensory conditions. These findings provide preliminary effect-size estimates to inform the design and sample size of future adequately powered studies.

References

[1] B. J. Benda, P. O. Riley, and D. E. Krebs, "Biomechanical relationship between center of gravity and center of pressure during standing," *IEEE Trans. Rehabil. Eng.*, vol. 2, no. 1, pp. 3–10, 1994.

[2] C. Fadillioglu, L. Kanus, F. Möhler, S. Ringhof, D. Hellmann, and T. Stein, "Effects of jaw clenching on dynamic reactive balance task performance after 1-week of jaw clenching training," *Front. Neurol.*, vol. 14, 2023.

[3] A. Dolina, M. Baszczowski, M. Zawadka, L. Sobiech, J. Szkutnik, and P. Gawda, "The impact of various subtypes of temporomandibular disorders on body balance – preliminary study," *Gait Posture*, vol. 116, pp. 40–46, 2025.

[4] M. Durukan and H. Guzel, "Spontaneous Activation of the Masseter Muscle in Balance Tasks: The Functional Relationship Between Postural Control and the Stomatognathic System in Healthy Adults," *Nig. J. Clin. Pract.*, vol. 29, no. 2, pp. 181–189, 2026.

[5] A. A. Garstka et al., "Cause-Effect Relationships between Painful TMD and Postural and Functional Changes in the Musculoskeletal System: A Preliminary Report," *Pain Res. Manag.*, vol. 2022, p. 1429932, 2022.

[6] S. Julià-Sánchez, J. Álvarez-Herms, H. Gatterer, M. Burtscher, T. Pagès, and G. Viscor, "Dental Occlusion Influences the Standing Balance on an Unstable Platform," *Motor Control*, vol. 19, no. 4, pp. 341–354, 2015.

[7] J. Khan, S. R. Singer, A. Young, N. Tanaiutchawoot, M. Kalladka, and M. Mupparapu, "Pathogenesis and Differential Diagnosis of Temporomandibular Joint Disorders," *Dent. Clin. North Am.*, vol. 67, no. 2, pp. 259–280, 2023.

[8] H. L. Miller et al., "Movement smoothness during dynamic postural control to a static target differs between autistic and neurotypical children," *Gait Posture*, vol. 99, pp. 76–82, 2023.

[9] A. Nota, S. Tecco, S. Ehsani, J. Padulo, and A. Baldini, "Postural stability in subjects with temporomandibular disorders and healthy controls: A comparative assessment," *J. Electromyogr. Kinesiol.*, vol. 37, pp. 21–24, 2017.

[10] G. Perinetti, L. Contardo, A. Silvestrini-Biavati, L. Perdoni, and A. Castaldo, "Dental Malocclusion and Body Posture in Young Subjects: A Multiple Regression Study," *Clinics*, vol. 65, no. 7, pp. 689–695, 2010.

[11] A. Rizzato, E. Gobbi, A. Paoli, and G. Marcolin, "Validity and reliability of an unstable board for dynamic balance assessment in young adults," *PLoS ONE*, vol. 18, no. 1, p. e0280057, 2023.

[12] E. Schiffman et al., "Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) for Clinical and Research Applications," *J. Oral Facial Pain Headache*, vol. 28, no. 1, pp. 6–27, 2014.

[13] T. C. K. Cheung and M. A. Schmuckler, "Multisensory and biomechanical influences on postural control in children," *J. Exp. Child Psychol.*, vol. 238, p. 105796, 2024.

[14] L. Zhang, L. Xu, J. Lu, B. Cai, and S. Fan, "Static Balance in Participants with Temporomandibular Joint Disc Displacement without Reduction Versus Healthy Participants: A Cross-Sectional Study," *Med. Sci. Monit.*, vol. 28, 2021.

Evaluation of Multimodal Data in the Classification of Patients with Dementia Using Explainable Artificial Intelligence

Luiz Fernando Cavalcanti de Oliveira
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
luiz.lfco@ufpe.br

Jéssica Maria Nogueira de Souza
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
jessica.nogueira@ufpe.br

Ana Paula Silva de Oliveira
Center for Biosciences
Federal University of Pernambuco
Recife, Brazil
paula.silvaoliveira@ufpe.br

Camila Tiodista de Lima
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
camila.tiodista@ufpe.br

Genivaldo Gedeão Batista de Oliveira
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
genivaldo.gedeao@ufpe.br

Pedro Marques de Souza Vasconcelos
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
pedromarquessv@hotmail.com

Clarisse Lins de Lima
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
clarisse.lins@ufpe.br

Maíra Araújo de Santana
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
maira.araujosantana@ufpe.br

Maria Karoline da Silva Andrade
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
karoline.andrade@ufpe.br

Wellington Pinheiro dos Santos
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

Juliana Carneiro Gomes Cassemiro
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
juliana.cgomes@ufpe.br

Giselle Machado Magalhães Moreno
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
giselle.moreno@ufpe.br

***Abstract*—This study applies Explainable Artificial Intelligence techniques (SHAP and LIME) to a multimodal classifier (SVM-RBF) to support the differential diagnosis of dementia syndromes. Using EEG, clinical, and sociodemographic data from the GeronGnosis dataset (n=50), global and local explainability analyses revealed that the model bases its decisions predominantly on cognitive and functional scores (MMSE, Pfeffer Questionnaire, and Mini-BESTest), followed by sociodemographic variables and, to a lesser extent, frontotemporal neurophysiological attributes. In conclusion, the integration of multimodal data with XAI imparts transparency and clinical coherence to the algorithm, validating its alignment with traditional diagnostic criteria and suggesting its applicability as a diagnostic support tool, particularly in cases of cognitive decline that remain poorly defined.**

***Index Terms*—Explainable Artificial Intelligence, Dementia Syndromes, Electroencephalography, Machine Learning.**

I. Introduction

Global population aging is projected to increase the proportion of individuals aged 60 years or older from 12% to 22% between 2015 and 2050, with two-thirds of this demographic concentrated in low- and middle-income countries [1]. This scenario projects a substantial increase in dementia syndromes, neurodegenerative conditions that progressively impair cognition and functional capacity [2]. Consequently, the number of individuals with dementia is estimated to reach 152.8 million by 2050 [3]. Alzheimer's disease (AD), the most prevalent form, is associated with the accumulation of beta-amyloid peptide and hyperphosphorylated tau protein, leading to progressive synaptic and neuronal loss [4], [5].

Frontotemporal dementia (FTD), also classified as a neurodegenerative dementia, predominantly affects the frontal and temporal lobes, causing behavioral and language alterations [6]. Mild Cognitive Impairment (MCI) represents an intermediate condition characterized by measurable cognitive decline and relative preservation of functional capacity, particularly in the performance of instrumental activities of daily living (IADL) [7], [8].

The early identification of dementia syndromes is highly relevant for prognosis, as therapeutic intervention in early stages can delay disease progression. However, the differential

Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil; Science and Technology Support Foundation of the State of Pernambuco (FACEPE), Brazil; and the National Council for Scientific and Technological Development (CNPq), Brazil.

diagnosis of dementia syndromes encounters cultural, structural, and economic barriers in clinical practice, given that specialist professionals and high-cost imaging examinations are resources with limited availability in public health systems [9]. This limitation drives the search for lower-cost and more scalable auxiliary methods.

In this context, electroencephalography (EEG) presents favorable characteristics: it is non-invasive, low-cost, and portable. Furthermore, it can capture alterations in cortical activity that precede structural changes detectable by neuroimaging, offering sensitive markers of the synaptic dysfunction characteristic of neurodegenerative processes [10], [11].

When combined with EEG, machine learning (ML) techniques have demonstrated robust performance in differentiating dementia syndromes, particularly when integrated with clinical and sociodemographic variables in multimodal approaches [12]–[15]. However, studies indicate that classifiers such as Support Vector Machines (SVM) operate as black boxes, limiting their clinical adoption [15], [16]. Explainable Artificial Intelligence (XAI) frameworks, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), have been proposed to quantify the contribution of each feature to the model's decision [17], [18]. Nevertheless, their application to multimodal dementia models remains underexplored.

This study applies XAI models, such as SHAP and LIME, to a previously validated SVM classifier trained on the GeronGnosis dataset [15], with the aim of quantifying the contribution of the clinical, neurophysiological, and sociodemographic features that most strongly influence the differential diagnosis of dementia syndromes.

II. Methodology

A. Study Design

The present study builds upon the work of Oliveira et al. (2026) [15], who tested and validated various machine learning models using the multimodal GeronGnosis dataset. Based on the Particle Swarm Optimization (PSO) configuration and the best-performing classifier identified by those authors, the SHAP and LIME explainability frameworks were applied to investigate the features that most significantly influenced the algorithm's decision-making process. Figure 1 presents a flowchart illustrating the methodological steps of both the study by Oliveira et al. (2026) [15] and the present study.

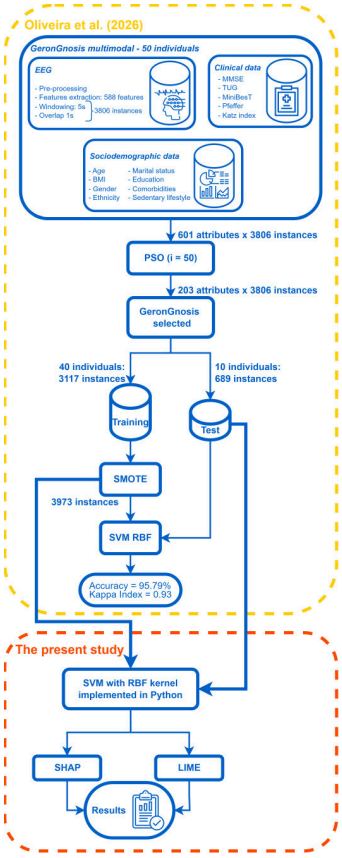


Fig. 1. Flowchart of the experiments.

B. GeronGnosis

The GeronGnosis dataset [15] comprises 50 older adults distributed across four groups: Healthy Controls (HC, n=20), Alzheimer's disease (n=15), Mild Cognitive Impairment (n=10), and Frontotemporal dementia (n=5). Recordings were performed using the EEG module of the Brain Wave III Neurovirtual polysomnograph, with participants in a resting state. Impedance was maintained below 5 kΩ and monitored through the software interface. The sampling rate was 300 Hz, with 16-bit analog-to-digital conversion and a resolution of $9.74 \times 10^{-2} \mu V$.

Participants are followed up at the Cognitive Neurology Outpatient Clinic of Hospital das Clínicas at the Federal University of Pernambuco (HC/EBSERH-UFPE), where cognitive and behavioral neurologists perform clinical evaluation and classification according to current guidelines established by the Brazilian Academy of Neurology. All data collection procedures were conducted only after approval by the Ethics and Research Committee (CEP) of the Health Sciences Center (CCS) at the Federal University of Pernambuco (UFPE), under process number CAAE 71096223.3.0000.5208.

The EEG signal was acquired using 19 electrodes positioned according to the international 10-20 system, supplemented by two auricular electrodes (A1 and A2), totaling 21 recording channels, plus one reference electrode positioned in the frontal region. During acquisition, high-pass filtering at 0.1 Hz, low-pass filtering at 70 Hz, and 60 Hz notch filtering were applied, with no subsequent digital filtering. Only resting-state segments with eyes closed were used, with an average duration of 5 minutes. The continuous signal was segmented into 5-second windows with 1-second overlap. For each window, 28 time-frequency features described in Table I were extracted from each of the 21 electrodes, resulting in 588 features per instance and 3,806 instances [19].

The dataset features a multimodal structure, comprising 28 time- and frequency-domain features for each of the 21

TABLE I
EXPLICIT ATTRIBUTES.

Attribute	Mathematical expression	Attribute	Mathematical expression
Average Amplitude Change	$AAC = \frac{1}{N}\left(\sum_{n=1}^{N}\left\|\frac{d\,x(t)}{t}\right\|\right)$	**Zero Crossings**	$ZC = \sum_{n=1}^{N-1}[sgn(x_n \times x_{n+1}) \cap \|x_n - x_{n+1}\| \geq threshold]$ $sgn(x) = \begin{cases}1, if\ x \geq threshold \\ 0, otherwise\end{cases}$
Difference Absolute Deviation	$DASDV = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{dx(t)}{dt}\right)^2}$	**Standard Deviation (σ)**	$STD = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\|x_n - \mu\|^2}$
Integrated Absolute Value	$IAV = \sum_{n=1}^{N} x_n$	**Mean Value (μ)**	$MVAL = \frac{1}{N}\sum_{n=1}^{N} x_n$
Logarithm Detector	$LOGD = e^{\left(\frac{1}{N}\sum_{n=1}^{N}\log(\|x_n\|)\right)}$	**Maximum Amplitude**	$MAX = max(x_n)$
Mean Absolute Value	$MAV = \frac{1}{N}\sum_{n=1}^{N}\|x_n\|$	**Power Spectrum Ratio**	$PSR = \frac{PKF}{\sum_{j=1}^{M} P_j}$
Mean Logarithm Kernel	$MLOGK = \frac{1}{N}\left\|\sum_{n=1}^{N} x_n\right\|$	**Mean Frequency**	$MNF = \frac{\sum_{j=1}^{M} f_j P_j}{\sum_{j=1}^{M} P_j}$ Where are f_j, P_j the frequencies and power of the spectrum, respectively, and M is the length of the frequencies.
Root Mean Square	$RMS = \sqrt{\frac{\sum_{n=1}^{N}(x_n)^2}{N}}$	**Mean Power**	$MNP = \sum_{j=1}^{M}\frac{P_j}{M}$
Kurtosis	$KURT = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^4}{\sigma^4}$	**Peak Frequency**	$PKF = \max(P_j)$
Slope Sign Changes	$SCC = \sum_{n=1}^{N-1}[f(x_n - x_{n-1}) \times (x_n - x_{n-1})]$ $f(x) = \begin{cases}1, IF\ x \geq threshold \\ 0, otherwise\end{cases}$	**Total Power**	$TTP = \sum_{j=1}^{M} P_j$
Simple Square Integral	$SSI = \sum_{n=1}^{N} {x_n}^2$	**1st Spectral Moments**	$SM1 = \sum_{j=1}^{M} f_j P_j$
Variance	$VAR = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu)^2$	**2nd Spectral Moments**	$SM2 = \sum_{j=1}^{M} {f_j}^2 P_j$
Third Moment	$TM3 = \left\|\frac{1}{N}\sum_{n=1}^{N}(x_n)^3\right\|$	**3rd Spectral Moments**	$SM3 = \sum_{j=1}^{M} {f_j}^3 P_j$
Fourth Moment	$TM4 = \left\|\frac{1}{N}\sum_{n=1}^{N}(x_n)^4\right\|$	**Variance of Central Frequency**	$VCF = \frac{SM2}{TP} - \left(\frac{SM1}{TP}\right)^2$
Fifth Moment	$TM5 = \left\|\frac{1}{N}\sum_{n=1}^{N}(x_n)^5\right\|$	**Median Frequency**	$MDF = \frac{1}{2}\sum_{j=1}^{M} P_j$

EEG electrodes; clinical data for assessing cognition (Mini-Mental State Examination [MMSE]), functionality (Katz Index and Pfeffer Functional Activities Questionnaire), and mobility/balance (Timed Up and Go [TUG] and Mini-Balance Evaluation Systems Test [Mini-BESTest]); and sociodemographic data, including age, body mass index (BMI), gender, ethnicity, marital status, education level, comorbidities, and level of sedentary behavior. The GeronGnosis EEG dataset was previously compared with the OpenNeuro dataset (which also consists exclusively of EEG recordings), achieving satisfactory classification performance among the evaluated groups, as reported by Oliveira et al. (2026) [15].

The authors randomly divided the dataset into training (80%) and test (20%) sets, performing a subject-level partition. This division resulted in 40 subjects (3,117 instances) for the training set and 10 subjects (689 instances) for the test set. The training set was balanced using the Synthetic Minority Oversampling Technique (SMOTE), resulting in a total of 3,973 instances.

C. Baseline Model and Experimental Setup - Oliveira et al. (2026) [15]

The present study builds upon the investigation initiated by Oliveira et al. (2026) [15], in which a multimodal classification pipeline was developed for the differential diagnosis of dementia syndromes using the GeronGnosis dataset. In that study, various classifiers were evaluated under two experimental configurations (binary and multiclass) combined with different feature selection strategies employing PSO. The binary configuration consisted of the classification between the control group (CS) and the dementia group (grouping individuals diagnosed with AD and FTD), whereas the multiclass configuration involved the CS, MCI, and dementia classes.

For the multiclass scenario, results demonstrated that the SVM classifier with Radial Basis Function (RBF) kernel, combined with PSO configured with 50 iterations, achieved the best performance on the test set, with an accuracy of 95.79% and a Kappa index of 0.93. This configuration resulted in a subset of 203 selected features, composed of EEG-derived features and clinical and sociodemographic variables, as detailed in Table II.

TABLE II
ATTRIBUTES SELECTED AFTER PSO.

Electrode	**Attributes**
FP1	AAC, DASDV, LOGD, MAX, MLOGK, TM4, SM1, SM3 and VCF
FP2	LOGD, MLOGK, VAR, MVAL, MAX, MNF, SM1 and SM2
F7	AAC, DASDV, IAV, MAV, SSI, VAR, MAX, MNP and SM2
F8	LOGD, MLOGK, RMS, KURT, SSI, VAR, STD, MVAL, SM2 and VCF
F3	AAC, MLOGK, VAR, TM4, MNF, PKF, SM1, SM2, SM3 and VCF
F4	AAC, RMS, KURT, TM3 and TM5
T3	AAC, DASDV, RMS, STD, MAX, MNF and SM1
T4	AAC, IAV, LOGD, KURT, TM5, STD, SM1, SM3 and VCF
C3	DASDV, LOGD, MAV, MLOGK, SSI, VAR, TM3, TM5, STD, MNF, SM1 and VCF
C4	DASDV, LOGD, MVAL, MAX, MNF, SM1, SM2 and SM3
T5	DASDV, MAV, KURT, SSI, SM1 and VCF
T6	DASDV, LOGD, RMS, TM3, TM4, TM5, STD, MVAL, SM1, SM2 and VCF
P3	AAC, IAV, LOGD, MAV, MLOGK, SSI, TM5, MNF, SM1, SM2 and VCF
P4	AAC, DASDV, IAV, LOGD, MLOGK, RMS, KURT, SSI, TM4, PKF, SM1 and SM3
O1	AAC, IAV, VAR, TM4, PSR, MNF and SM3
O2	AAC, MAV, RMS, SSI, VAR, TM5, MVAL, MAX, MNF, SM2, SM3 and VCF
FZ	AAC, DASDV, IAV, LOG, MAV, TM3, MVAL, MNF and SM2
CZ	AAC, MLOGK, SSI, VAR, TM4, TM5, SM2 and SM3
PZ	DASDV, MAV, RMS, SSC, TM4, TM5, MAX, PSR, MNF, SM2, SM3 and VCF
A1	AAC, LOGD, MAV, RMS, VAR, TM4, TM5, STD, MVA, MAX, MNF, SM3 and VCF
A2	AAC, IAV, LOG, TM4, STD, MNF and SM3
Clinical attributes	
MMSE, Pfeffer, Katz Index, MiniBesTest, TUG, Age, BMI and Scholarity	

D. Classifier Reimplementation for Interpretability Analysis

Since the study by Oliveira et al. (2026) [15] used the SVM classifier implementation available in Weka software (version 3.8.6), and considering that XAI libraries—particularly SHAP

and LIME—require direct access to output probabilities and the internal structure of the trained model, the SVM-RBF classifier was reimplemented in Python.

The reimplemention was performed using the scikit-learn library (version 1.8.0), maintaining the following hyperparameters in accordance with the original configuration whenever possible:

- **Kernel**: RBF;
- **Regularization parameter (C)**: 1;
- **Kernel parameter (gamma)**: 0.01;
- **Probabilities**: enabled (`probability=True`), a requirement for applying SHAP via `KernelExplainer` and LIME via `LimeTabularExplainer`;
- **Stopping tolerance**: 0.001;
- **Random seed** (`random_state`): 1;
- **Data standardization**: parameters fitted on the training set and applied to the test set.

This reimplementation step was essential to enable the application of explainability techniques, since Weka does not offer native integration with the SHAP and LIME frameworks. The model was trained on the same subset of 203 features selected by PSO with 50 iterations, in order to ensure direct comparability with the results reported in the preceding study. The reimplementation was validated by comparing performance metrics with those of the original study, obtaining an accuracy of 95.79% and a Kappa index of 0.93 on the test set, values identical to those originally reported.

E. XAI Implementation

To address the black-box problem of SVM and enable clinical interpretation of decision-making, this study implemented two XAI frameworks: (i) SHAP, grounded in Game Theory, which quantifies the influence of each variable on the decision, operating at both global interpretation (summary plots) and local interpretation (waterfall plots for individual instances); and (ii) LIME, focused on local interpretation, which generates synthetic perturbations around a specific instance to provide a simplified explanation of that isolated prediction [17], [18].

LIME, by default, explains one instance at a time. Since each individual is represented by multiple instances due to temporal signal windowing, the algorithm was adapted to calculate the average of importance values across all instances of the same individual. The average importance values were calculated preserving the sign of each feature's weights. For each class, one individual from the test set was randomly selected for local analysis.

SHAP was implemented using the `KernelExplainer`, with 100 samples from the training set as background reference. LIME, in turn, employed the `LimeTabularExplainer` with random seed 42, configured with 15 features and 1,000 perturbations per explained instance.

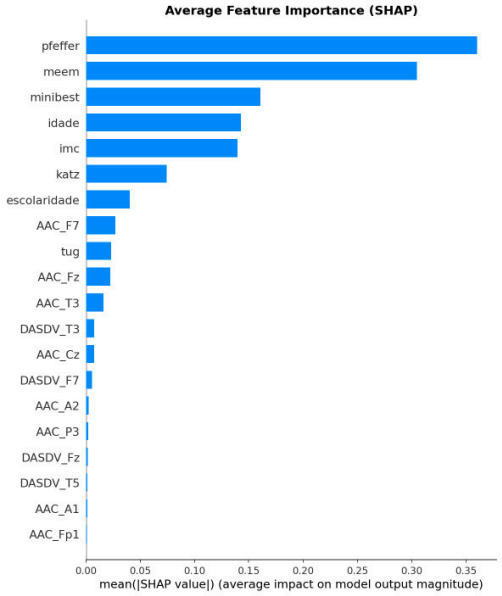


Fig. 2. Absolute mean importance of the main attributes in the global prediction.

III. Results

A. Global Explainability via SHAP

As a means to overcome the inherent interpretability limitation of black-box algorithms, the SHAP framework was applied to the test set predictions, enabling the mapping of the exact contribution of each feature to the model's mathematical decision. The mean absolute importance analysis demonstrated that the model anchored its primary decisions mainly on clinical and functional markers, as shown in Figure 2.

As evidenced by the global ranking, the Pfeffer Functional Activities Questionnaire, the MMSE, and the Mini-BESTest were the main determinants of the classifier's decisions. It is worth noting that the MMSE and the Pfeffer questionnaire are part of the assessment protocol used by neurologists for the diagnostic classification of participants. Thus, the fact that these features predominate in the global ranking suggests that the classifier reproduces the very instruments that originated the labels, which does not represent independent evidence of its discriminative power. Added to this clinical basis, sociodemographic variables (age and BMI) demonstrated a secondary weighting factor for classification. In the neurophysiological domain, frontal and temporal electrodes (especially AAC_F7, AAC_Fz, and AAC_T3) emerged as most influential. To understand the impact of these features on class differentiation, SHAP density values were segmented by predictive class.

In the HC group, the features that stood out most in classification were those related to cognitive assessment scores (MMSE) and IADLs, represented by the Pfeffer questionnaire (Figure 3). The results reflect how the preservation of such variables can contribute to the correct assignment of participants to their respective group by the model. In parallel, the high motor performance evidenced in the Mini-

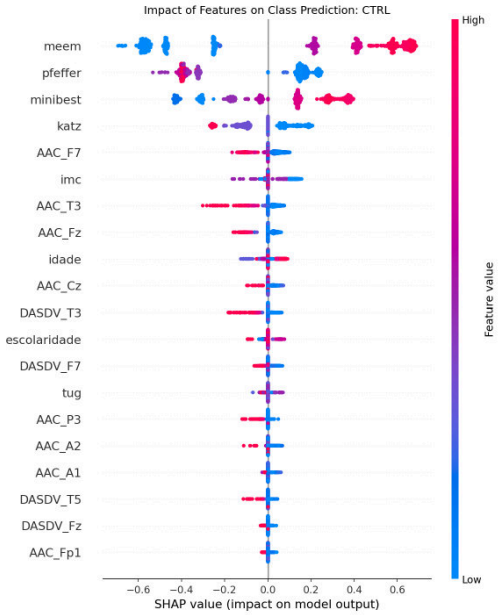


Fig. 3. SHAP for the control class.

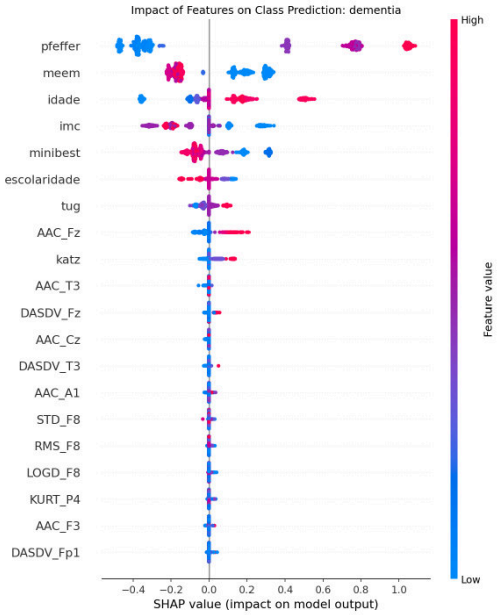


Fig. 4. SHAP for the dementia class.

BESTest consolidated the classification of individuals in the control group. Regarding EEG activity, it is noteworthy that its contribution was more modest compared to that observed in the clinical variables, with a pattern of greater relevance in the frontal and temporal areas (AAC_F7 and AAC_T3).

As in the HC group classification, the MMSE and Pfeffer questionnaire scores stood out as the main predictors for model accuracy in the Dementia group (Figure 4). The low scores obtained by this group in the assessments likely facilitated discrimination between the healthy control profile and the profile with a dementia diagnosis. Additionally, advanced age and balance deficits (assessed by the Mini-BESTest) also contributed significantly to classification. The prominence of EEG activity appears in relation to midline electrodes (AAC_Fz and AAC_Cz), which, according to the model's results, developed greater weight in classification. Similarly to what was found in the other groups, the electrode located in the temporal region (AAC_T3) also contributed to the result, however, with lower weight compared mainly to MCI.

In the SHAP analysis of the MCI group (Figure 5), the Pfeffer questionnaire, age, MMSE score, and BMI stood out as the most relevant features for algorithm performance. This result demonstrates that clinical and sociodemographic variables were determinant for the classification of this group. However, the weight distribution pattern for the group reveals an overlap among the most relevant variables, unlike what was found in the other groups. Furthermore, electrode activity, although more modest, still presented relevance (as evidenced by signal intensity on the right side), especially in the left frontotemporal regions (AAC_F7, AAC_T3, DASDV_F7, and DASDV_T3).

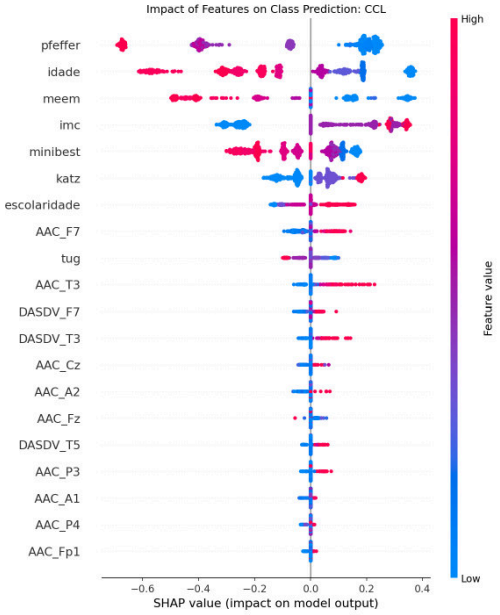


Fig. 5. SHAP for the MCI class.

B. *Local Explainability via LIME*

A locally-targeted analysis through the LIME framework was also implemented, with the objective of deepening the predictions for individual participants. When conducting this type of experiment, it is expected that the results previously achieved will be maintained when specifically testing one participant from each research group.

The analysis of case P018 (HC group) illustrates the classifier's behavior for a participant from this group, whose prediction was obtained by averaging 76 temporal windows (Figure

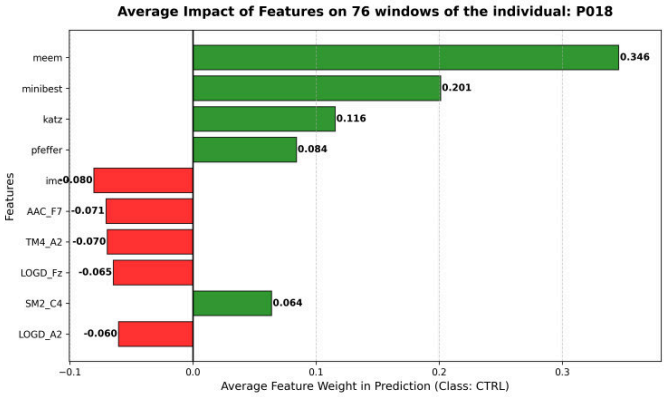


Fig. 6. Interpretability of individual P018 using LIME – control class.

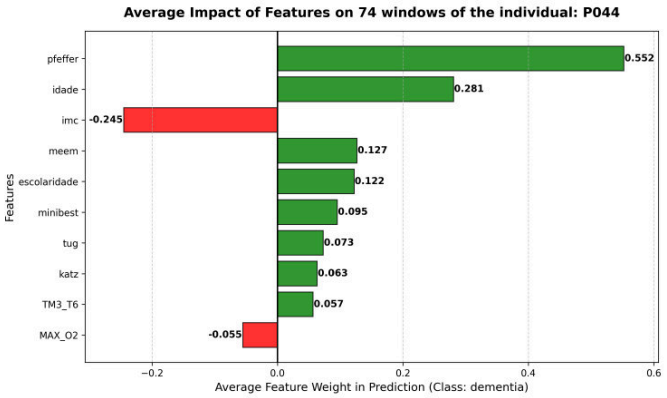


Fig. 7. Interpretability of individual P018 using LIME – dementia class.

6). The results demonstrate that clinical variables, particularly the MMSE score (average impact weight of 0.346), were determinant for the classification of the participant into their respective group.

In addition to cognitive function, adequate performance on the balance scale and instruments related to functional assessment also positively influenced the correct classification of case P018. This result highlights that the preservation of these domains is essential for characterizing cognitively intact individuals.

The LIME model evaluation for a patient in the dementia group (patient P044) was conducted across 74 windows. Analysis of the plot in Figure 6 clearly demonstrates the influence of the Pfeffer questionnaire on the classification of this patient (weight of 0.552). This finding corroborates the clinical importance of assessing the loss of IADLs, which serves as a key criterion in diagnostic decision-making. In addition to the Pfeffer questionnaire score, age (weight of 0.281) and the MMSE score (weight of 0.127) were decisive factors in the classification of patient P044..

In the MCI group, patient P023 was selected for LIME data processing. In this case, 77 windows were analyzed, and the results are presented in Figure 8. The explainability plot highlights that the BMI variable most favored the classifier's

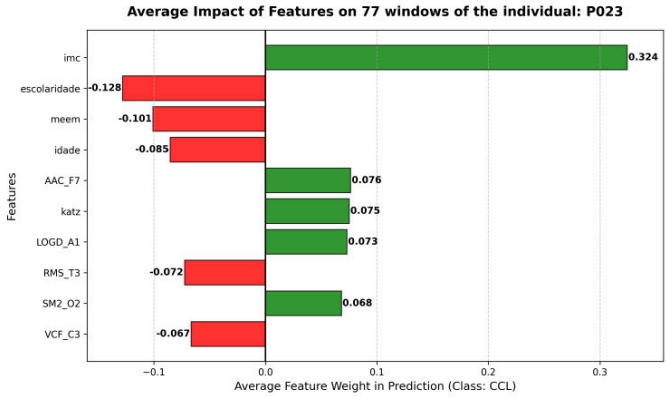


Fig. 8. Interpretability of individual P023 using LIME – MCI class.

final decision for patient P023 (weight of 0.324). The education level (-0.128) and MMSE (-0.101) variables contributed negatively to the MCI group score; that is, they reduced the predicted probability of classification as MCI by the model. It is also noteworthy, albeit with lower impact, the influence of electrodes F7 (frontal lobe, weight of 0.076), A1 (temporal lobe, weight of 0.073), and O2 (occipital, 0.068) for the correct classification of P023 in the MCI group.

IV. Discussion

The results of this study indicate that the integration of multimodal data, analyzed through machine learning algorithms combined with XAI, provides support for adequate classification among individuals in the HC, MCI, and dementia groups. The subject-level stratified partition adopted in this work stands out, which confers greater reliability to the reported metrics compared to approaches that segment data by epochs. The global analysis using the SHAP framework revealed that the results of the clinical assessment battery performed with participants from the dataset (use of MMSE, Pfeffer questionnaire, Mini-BESTest) exerted greater mean influence on the model's predictions. Furthermore, the local analysis performed by LIME largely reinforced the main relevant variables in the classification tasks, according to what had been found in the SHAP analysis.

For the HC group, the preservation of cognitive scores (MMSE), functional capacity (Pfeffer), and balance (Mini-BESTest) were crucial for the final classification. This result reinforces that the high performance found in participants from this group enables the employed models to differentiate a profile of good clinical performance from those with some type of deficit (as found in the dementia group) [20]. The repetition of the variables found in SHAP in the LIME result reinforces that cognitive and functional integrity are, in fact, determinant for adequately classifying participants from the control group. Previous studies report that discrimination between clinically well-defined groups, such as patients with dementia and matched controls, tends to be less challenging than differentiation in intermediate conditions, which can be observed in the MCI group pattern [15], [21]–[23].

In the dementia group, given that patients already presented more evident cognitive and functional impairments, their lower clinical performance facilitated their discrimination compared to the other groups. Regarding EEG activity, we observed a greater weight in midline electrodes (Fz and Cz), which may suggest diffuse alterations in cortical activity related to an advanced stage of cognitive impairment [24], [25]. Once again, the individual assessment results from LIME confirmed the predominance of clinical and sociodemographic variables. Particularly for the dementia group, the alterations in the Pfeffer questionnaire score that facilitated classification establish a close relationship with the diagnostic investigation performed in clinical practice [26]. The results of the present study reinforce that functional losses characterize the profile of patients with dementia, beyond the already established cognitive impairment. Another relevant finding in the individual analysis was the high weight of the education variable for classification (a value close to that of the MMSE). This finding is consistent with the results of a Brazilian study, which detected that educational level presented the highest relative risk for cognitive impairment [27]. Lower educational attainment often serves as a proxy for reduced cognitive reserve accumulated throughout life; conversely, cognitive reserve is directly proportional to the level of education, acting as a protective factor against pathological processes such as dementia [28].

In the MCI group, the overlap of feature weights among clinical, sociodemographic, and neurophysiological variables is consistent with the characterization of this phase as a heterogeneous continuum, marked by greater interindividual variability [29]. In the global analysis, as revealed by the SHAP algorithm, the considerable influence of frontotemporal electrodes on the classification stands out. This may indicate that subtle alterations in cortical electrical activity were already detectable by EEG, even before any overt impairment in global functional capacity becomes apparent. Once again, cognitive reserve can help explain the discrepancy observed between cortical electrical activity patterns and clinical assessment results. Simfukwe et al. (2025) [30] demonstrated that greater cognitive reserve is associated with delayed onset of cognitive decline, lower dementia risk, and better clinical outcomes, despite a similar neuropathological burden.

Given that EEG can provide relevant information at the cortical level, its application is justified by the detection of electrophysiological alterations that may precede the clinical manifestations of dementia syndromes or even MCI. Therefore, even though clinical variables were more prominent in the XAI results, other types of EEG processing, such as those related to the investigation of more or less active cortical areas, can provide valuable information for the characterization of each clinical condition studied.

The study presents methodological limitations that require cautious interpretation. The primary limitation concerns the small size of the original clinical sample (n=50 participants), which may restrict the generalizability of the findings. Despite the increase in the number of computational instances achieved through EEG signal windowing into 5-second segments, there remains a possibility that the heterogeneity of the general population was not fully represented. To mitigate the risk of overfitting and data leakage, rigorous protocols were implemented. These include subject-level stratified partitioning, which prevents windows from the same participant from appearing simultaneously in both the training and test sets, alongside the application of the SMOTE technique strictly restricted to the training set.

Future investigations with larger sample sizes and longitudinal designs are necessary to evaluate whether the patterns identified by the multimodal model are reproducible and robust. Temporal follow-up studies could determine whether patients with MCI, classified based on the bioelectrical signature detected by the algorithm, progress to dementia, thereby validating the early predictive potential of this approach.

V. Conclusion

To overcome the interpretability limitations of the models, XAI frameworks such as SHAP and LIME were employed, providing greater transparency and clinical coherence to the classifier's decisions. The results suggest that the integration of EEG and clinical data can offer complementary information for diagnostic assessment, particularly when traditional neuropsychological evaluation proves inconclusive due to the effects of cognitive reserve, as seen in MCI. Furthermore, XAI revealed the model's reliance on the instruments of the diagnostic criteria themselves, an insight that performance metrics alone would not have identified. The combination of multimodal data with interpretable ML algorithms has demonstrated potential, albeit preliminary, as a resource to support the differential diagnosis of electrophysiological and clinical alterations in a patient population with cognitive deficits.

VI. Declaration on the Use of Artificial Intelligence

The authors used AI tools for language polishing. The authors assume full responsibility for all content.

Acknowledgment

The authors are grateful to the Brazilian Ministry of Health, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil), the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE, Brazil), and the National Council for Scientific and Technological Development (CNPq, Brazil) for their partial financial support.

References

[1] World Health Organization. (2025, October) Ageing and health. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[2] Brasil, "Relatório nacional sobre a demência: epidemiologia,(re) conhecimento e projeções futuras." 2024.

[3] E. Nichols, J. D. Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019," *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022.

[4] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C. E. Teunissen, J. Cummings, and W. M. Van der Flier, "Alzheimer's disease," *The Lancet*, vol. 397, no. 10284, pp. 1577–1590, 2021.

[5] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta neuropathologica*, vol. 82, no. 4, pp. 239–259, 1991.

[6] J. Bang, S. Spina, and B. L. Miller, "Frontotemporal dementia," *The Lancet*, vol. 386, no. 10004, pp. 1672–1682, 2015.

[7] F. Hajamohideen, N. Shaffi, M. Mahmud, K. Subramanian, A. Al Sariri, V. Vimbi, A. Abdesselam, and A. D. N. Initiative, "Four-way classification of alzheimer's disease using deep siamese convolutional neural network with triplet-loss function," *Brain Informatics*, vol. 10, no. 1, p. 5, 2023.

[8] R. C. Petersen, "Mild cognitive impairment as a diagnostic entity," *Journal of internal medicine*, vol. 256, no. 3, pp. 183–194, 2004.

[9] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper *et al.*, "Dementia prevention, intervention, and care: 2020 report of the lancet commission," *The lancet*, vol. 396, no. 10248, pp. 413–446, 2020.

[10] R. Cassani, M. Estarellas, R. San-Martin, F. J. Fraga, and T. H. Falk, "Systematic review on resting-state eeg for alzheimer's disease diagnosis and progression assessment," *Disease markers*, vol. 2018, no. 1, p. 5174815, 2018.

[11] B. Jiao, R. Li, H. Zhou, K. Qing, H. Liu, H. Pan, Y. Lei, W. Fu, X. Wang, X. Xiao *et al.*, "Neural biomarker diagnosis and prediction to mild cognitive impairment and alzheimer's disease using eeg technology," *Alzheimer's research & therapy*, vol. 15, no. 1, p. 32, 2023.

[12] Y. Huang, Z. Li, F. Liu, B. Li, C. Mao, L. Dong, S. Chu, W. Jin, J. Wang, and J. Gao, "Application of machine learning in eeg-based dementia diagnosis: Classification and differential diagnosis," *Journal of Alzheimer's Disease*, vol. 108, no. 1_suppl, pp. S81–S93, 2025.

[13] M. Acharya, R. C. Deo, X. Tao, P. D. Barua, A. Devi, A. Atmakuru, and R.-S. Tan, "Deep learning techniques for automated alzheimer's and mild cognitive impairment disease using eeg signals: A comprehensive review of the last decade (2013-2024)," *Computer Methods and Programs in Biomedicine*, vol. 259, p. 108506, 2025.

[14] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel multi-modal machine learning based approach for automatic classification of eeg recordings in dementia," *Neural Networks*, vol. 123, pp. 176–190, 2020.

[15] A. P. S. de Oliveira, C. T. de Lima, J. M. N. de Souza, L. F. C. de Oliveira, L. H. Calixto, A. S. de Oliveira Silva, C. M. de Lima Simões, M. E. S. B. Lins, B. J. A. P. Barbosa, C. L. de Lima, J. C. G. Cassemiro, M. A. de Santana, G. M. M. Moreno, W. P. dos Santos, and L. A. L. Soares, "Gerongnosis: Multimodal machine learning for the differential diagnosis of alzheimer's disease, frontotemporal dementia, and mild cognitive impairment," *AI in Neurology*, 2026.

[16] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[19] I. B. Nunes, M. A. de Santana, N. Charron, H. S. e. Silva, C. M. de Lima Simões, C. Lins, A. B. de Souza Sampaio, A. M. N. de Melo, T. C. V. da Silva, C. Tiodista *et al.*, "Automatic identification of preferred music genres: an exploratory machine learning approach to support personalized music therapy," *Multimedia Tools and Applications*, vol. 83, no. 35, pp. 82 515–82 531, 2024.

[20] E. Ersözlü, L. Preis, A. Aktuz, L. Droste, A. Erman, D. Gref, K. S. Strentz, J. Hellmann-Regen, A. D. N. Initiative, F. L. D. N. Initiative *et al.*, "Drawing a line: Differentiating mild from moderate dementia using the functional activities questionnaire," *The journal of prevention of Alzheimer's disease*, vol. 13, no. 8, p. 100630, 2026.

[21] Z. Li, X. Jiang, Y. Wang, and Y. Kim, "Applied machine learning in alzheimer's disease research: omics, imaging, and clinical data," *Emerging topics in life sciences*, vol. 5, no. 6, pp. 765–777, 2021.

[22] L. Lampe, H.-J. Huppertz, S. Anderl-Straub, F. Albrecht, T. Ballarini, S. Bisenius, K. Mueller, S. Niehaus, K. Fassbender, K. Fliessbach *et al.*, "Multiclass prediction of different dementia syndromes based on multi-centric volumetric mri imaging," *NeuroImage: Clinical*, vol. 37, p. 103320, 2023.

[23] P. R. Millar, B. A. Gordon, P. H. Luckett, T. L. Benzinger, C. Cruchaga, A. M. Fagan, J. J. Hassenstab, R. J. Perrin, S. E. Schindler, R. F. Allegri *et al.*, "Multimodal brain age estimates relate to alzheimer disease biomarkers and cognition in early stages: a cross-sectional observational study," *Elife*, vol. 12, p. e81869, 2023.

[24] V. Papaliagkas, "The role of quantitative eeg in the diagnosis of alzheimer's disease," *Diagnostics*, vol. 15, no. 15, p. 1965, 2025.

[25] Y. Olğun, C. Aksoy Poyraz, M. Bozluolçay, and B. Ç. Poyraz, "Quantitative eeg in the differential diagnosis of dementia subtypes," *Journal of geriatric psychiatry and neurology*, vol. 37, no. 5, pp. 368–378, 2024.

[26] J. Smid, A. Studart-Neto, K. G. César-Freitas, M. C. N. Dourado, R. Kochhann, B. J. A. P. Barbosa, L. P. Schilling, M. L. F. Balthazar, N. A. F. Frota, L. C. d. Souza *et al.*, "Subjective cognitive decline, mild cognitive impairment, and dementia-syndromic approach: recommendations of the scientific department of cognitive neurology and aging of the brazilian academy of neurology," *Dementia & Neuropsychologia*, vol. 16, pp. 1–24, 2022.

[27] R. M. Castilhos, V. B. Leotti, N. Feter, A. C. Goulart, A. R. Brunoni, C. K. Suemoto, L. Giatti, M. C. Viana, S. M. Barreto, S. Alvim *et al.*, "Population attributable fractions of modifiable dementia risk factors for cognitive impairment—the elsa-brasil cohort study," *The Lancet Regional Health–Americas*, vol. 49, 2025.

[28] M. G. Barbosa, L. M. Teixeira, J. V. d. M. Mambrini, W. L. Rabelo, A. C. C. Miguel, M. F. Lima-Costa, and C. P. Ferri, "Dementia and all-cause mortality in older adults: Findings from the elsi-brazil study," *Alzheimer's & Dementia*, vol. 22, no. 5, p. e71400, 2026.

[29] H. Azami, C. Zrenner, H. Brooks, R. Zomorrodi, D. M. Blumberger, C. E. Fischer, A. Flint, N. Herrmann, S. Kumar, K. Lanctôt *et al.*, "Beta to theta power ratio in eeg periodic components as a potential biomarker in mild cognitive impairment and alzheimer's dementia," *Alzheimer's research & therapy*, vol. 15, no. 1, p. 133, 2023.

[30] C. Simfukwe, S. S. A. An, and Y. C. Youn, "Mechanisms and impact of cognitive reserve in normal aging and alzheimer's disease," *Diagnostics*, vol. 15, no. 23, p. 3068, 2025.

Evaluation of Wavelet-Based Filters for Denoising Gas Vesicles Ultrasound Responses

1st Bruna Pessoa
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ
brunaferrarip@peb.ufrj.br

2nd João V. Pereira
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ
jv.cruz.20231@poli.ufrj.br

3rd Julia Oliveira
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ
juoli.20252@poli.ufrj.br

4th Raymundo Cordero Garcia
Electrical Eng. Program
Federal University of Mato Grosso do Sul
and Biomedical Eng. Program /COPPE
Federal University of Rio de Janeiro
MS, RJ, Brazil
raymundo.garcia@ufms.br

5th João Carlos Machado
Biomedical Eng. Program/COPPE
and Post Graduation Program
in Surgical Sciences
Federal University of Rio de Janeiro
Rio de Janeiro, RJ
jcm@peb.ufrj.br

***Abstract*—Nowadays, gas vesicles (GVs) have been studied for the development of improved contrast agents for ultrasound biomedical imaging. When GVs receive US pulses, the acoustic responses of the GVs are captured by US transducers as radio-frequency (RF) signals which are analyzed by digital processors to get tissue information. These RF signals have to be denoised before being analyzed since noise and electromagnetic interference usually affect these signals. Conventional linear filters are not suitable for non-stationary signals, such as the GV responses. This paper evaluates different wavelet filters for the denoising of the GV responses. Wavelet filtering allows a better denoising of non-stationary signals, for different detail levels (decomposition levels). Tests performed on experimental ultrasound GV responses using a piezoelectric transducer and an oscilloscope show the characteristics of applying different wavelets for the denoising process of the GV ultrasound responses. The results indicate that wavelets with low vanishing moments are not adequate for RF signal denoising, while increasing the number of decomposition levels resulted in a marginal improvement of the denoising process. However, using fewer decomposition levels reduces the required processing time. Hence, the number of decomposition levels must be properly selected.**

***Index Terms*—Biomedical imaging, data acquisition, ultrasound imaging, wavelet filtering.**

I. Introduction

Wavelet transform is a powerful tool for signal processing in biomedical applications [1]–[5] Among these signal processing applications, biomedical imaging based on ultrasound requires processing the radio-frequency (RF) signals produced during the acquisition of the ultrasound waves [6]–[9].

On the other hand, contrast agents are required in biomedical imaging systems for getting adequate tissue images. Nowadays, gas vesicles (GVs) have been studied for their application as contrast agents for ultrasound (US) imaging systems. GVs are air-filled protein nanostructures with a width between 45–250 nm and a length of hundreds of nm [10]. Besides, GVs scatter the received sound waves, due to their low density and high compressibility, producing a ultrasound contrast. As a result, GVs have better tissue penetration (due to their nano-scale dimensions) and give stronger acoustic responses than other conventional microbubble contrast agents [11], [12].

The GV acoustic responses are acquired by the transducer as RF signals. Noise and electromagnetic interference can affect the acquired RF signals. Even when using analog filters in the data acquisition system, digital filtering is commonly applied. Conventional digital filters (Butterworth, Chebyshsev, finite impulse response filters, among others) operate mainly in the frequency domain, with poor time localization and their performance is not good for non-stationary signals (such as electrocardiograms and ultrasound signals).

On the other hand, wavelet filtering has many advantages over conventional filters [13], [14]. This method has good performance for stationary and non-stationary signals, the time and frequency resolution can be adjusted, which is useful to detect spikes and irregularities in the signal, and allows multi-resolution signal analysis:the signal is analyzed at different levels of detail (decomposition levels).

This paper explores the application of wavelet filtering in the denoising of RF signals obtained from the ultrasound responses of GVs. An adequate RF signal denoising allows an improved analysis of the ultrasound properties of GVs, aiming for their further application as contrast agent. A phantom was prepared using a combination of agarose with GVs produced by Halobacterium salinarum NRC–1, until reaching an optical density (OD) of 0.8. A piezoelectric ultrasound transducer was used to send ultrasound pulses to the phantom, and also to

This research was supported in part by CAPES, CNPq and FAPERJ.

acquire the GV responses as RF signals. These RF responses were amplified, filtered using a low pass filter and sent to a oscilloscope to save the samples. Each RF response of the phantom was visualized and saved by the oscilloscope using two different acquisition modes for: single mode (the oscilloscope acquires only one waveform after the trigger event) and 128-average mode (the oscilloscope acquires 128 waveforms and the average of these waveform is stored and visualized). The wavelet filtering is applied to the single-mode-acquisition waveform, while the 128-average-acquisition waveform is used as a reference signal (an approximation of the RF signal in the absence of noise). Wavelet filters with different mother wavelets (Daubechies, Symlets, Coiflets) and decomposition levels were applied to filter the RF signal.

The performed tests indicate that mother wavelets with low vanishing moments are not adequate for the filtering of the RF signals, while increasing the number of decomposition levels does not significantly improve the denoising process. The filtering using the wavelet Symlet 8 with 6 decomposition levels gives the best noise attenuation.

II. Theoretical Foundations

A. Wavelets

A wavelet is a function with limited duration that can be used to decompose a signal, similarly to the Fourier analysis that decomposes a signal into a set of sinusoidal functions. However, the sinusoidal functions are regular and periodic, being not adequate to detect transients. On the other hand, a wavelet function can be symmetrical or asymmetrical, sharp or smooth, being better for the analysis of peaks, instantaneous changes of amplitude or frequency, irregularities or other signal transients. A function $\Psi(t)$ can be used as a wavelet if the signal has zero mean and finite energy. These conditions can be mathematically expressed as follows [15], [16]:

$$\int_{-\infty}^{+\infty} \Psi(t)dt = 0, \tag{1}$$

$$\int_{-\infty}^{+\infty} |\Psi(t)|^2 dt = 0. \tag{2}$$

Fig. 1 shows some wavelets used for signal processing.

The wavelet transform uses translated and scaled versions of a function called mother wavelet $\Psi(t)$;

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right), \quad 0 < a \in \Re, b \in \Re, \tag{3}$$

where a and b are the scale and translation factors, respectively. The continuous-time wavelet transform (CWT) of the signal $x(t)$ with respect to the mother wavelet $\Psi(t)$ is defined as

$$w(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\Psi\left(\frac{t-b}{a}\right)dt. \tag{4}$$

A wavelet coefficient shows the similarity (correlation) between a segment of $x(t)$ and the scaled/translated version of $\Psi(t)$. Therefore, the set of wavelet coefficients $w(a,b)$ is the wavelet representation of $x(t)$ with respect to the wavelet mother.

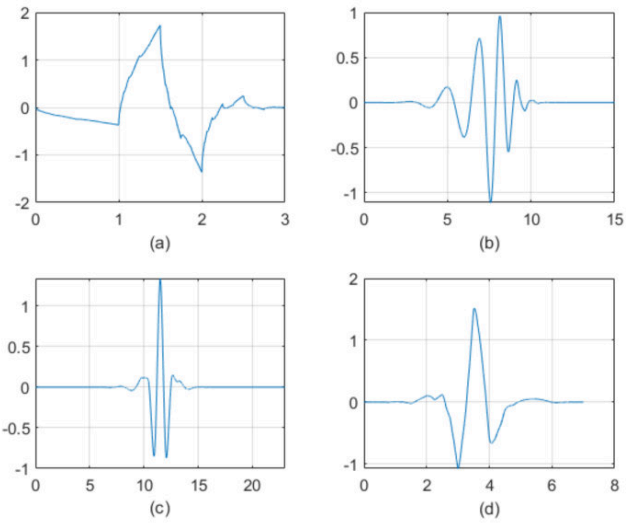


Fig. 1. Examples of wavelets used in denoising. a) Daubechies 2; b) Daubechies 8; c) Coiflet 4; d) Symlets 4

In the CWT, the mother wavelet is continuously scaled and translated, producing a great quantity of redundant information. A more efficient approach is to translate this wavelet in dyadic positions (positions which are powers of two). This approach is called the discrete wavelet transform (DWT), which is defined as follows:

$$w(m,k) = \frac{1}{\sqrt{a_0^m}}\sum_n x[n]\Psi\left(\frac{k - nb_0ba_0^m}{a_0^m}\right), \tag{5}$$

where $a = a_0^m$ and $b = nb_0ba_0^m$, being $x[n]$ the discrete-time version of $x(t)$, n is the discrete time, while m and k are integer. The vanishing moment of a wavelet $\Psi(t)$ is the number M such that

$$\int_{-\infty}^{+\infty} t^k\Psi(t)dt = 0, \quad k = 0, 1, \ldots, M-1. \tag{6}$$

For example, the mother wavelet Daubechies 8 has 8 vanishing moments.

B. Wavelet Filtering through Multi-Resolution Analysis

DWT allows a multi-resolution analysis (MRA) of the signal, since DWT gives time and frequency (through the scale parameter) information. MRA consists on a recursive operation to represent a discrete-time signal $x[n]$ at different levels of resolution [15], [16].

The signal denoising process based on MRA is composed by the following steps:

1) Decomposition: Fig. 2 shows the wavelet decomposition process for signal denoising. Recursive filtering and down-sampling procedures give the DWT coefficients at different detail levels. The approximation signal at level $j-1$ (for level $j = 1$, the approximation signal is $x[n]$). is filtered using a low-pass filter (L) and a high-pass filter (H). The output of each filter suffers a dyadic decimation (down-sampled by two): for each two

samples of the filter output, one is discarded. Therefore, the low-pass and high-pass filtering processes produce the set of approximation coefficients (cA_j) and the set of detail coefficients (cD_j), respectively:

$$cA_j = \{cA_j[1],\ cA_j[2],\ cA_j[3], \ldots\}, \tag{7}$$

$$cD_j = \{cD_j[1],\ cD_j[2],\ cD_j[3], \ldots\}, \tag{8}$$

where $cA_j[k]$ and $cD_j[k]$ are the k-th approximation and detail coefficients, respectively. The structures of the low-pass and high-pass filters depend on the selected mother wavelet.

2) Thresholding: the aforementioned detail coefficients *cD* are substituted by zero if they are less than a threshold value (c_{th}). Consequently, new detail coefficients ($c'D_j$) are created. There are different thresholding methods used in wavelet denoising techniques, including fixed and adaptive approaches. The hard thresholding method was selected in this paper as it is simple to compute, being easier to implement in real-time applications than adaptive approaches. The hard thresholding is defined as follows [17]:

$$c'D_j[k] = \begin{cases} cD_j[k], & |cD_j[k]| \geq c_{th} \\ 0, & |cD_j[k]| < c_{th} \end{cases}, \tag{9}$$

where c_{th} is a threshold. The proposed denoising approach applies the universal threshold formula [17]:

$$c_{th} = \sigma\sqrt{2ln(N)}, \tag{10}$$

where N is the number of signal samples, and σ is the standard additive gaussian white noise (it can be considered that $\sigma = 1$ if information of the gaussian white noise is not available). This thresholding process reduces the noise that affect the signal.

3) Reconstruction: Fig. 3 presents the reconstruction signal process, which is done using the approximation coefficients (cA_j), the detail coefficients after the thresholding process ($c'D_j$) and the reconstruction low-pass and high-pass filters (L' and H', respectively).

The length of the low-pass and high-pass filters, when Coiflet, Daubechies and Symlet wavelets are used, is directly related to the number of mother wavelet vanishing moment.

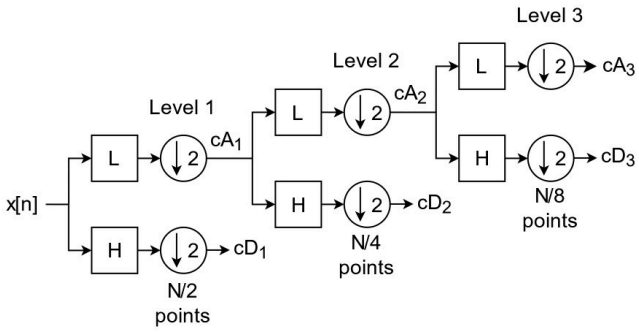


Fig. 2. Signal decomposition based on MRA and wavelets.

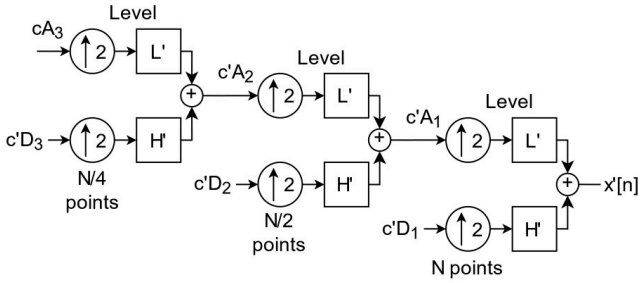


Fig. 3. Signal reconstruction based on MRA and wavelets.

III. Methodology

A. Experimental Acquisition Setup

A phantom was prepared using a combination of agarose with GVs produced by Halobacterium salinarum NRC–1, until getting an optical density (OD) of 0.8. The GVs were prepared according to the methodology described in [18].

A piezoelectric transducer (C190210; Capistrano Labs Incorporated, CA, USA) was used to produce and receive US waveforms. A 50 MHz electrical voltage pulse is applied to the transducer, generating an ultrasound (US) wave with a central frequency of 50 MHz to the phantom. The GVs receive the US signal and produce an acoustic response which is acquired by the transducer as RF signals which are amplified, filtered (using an analog low-pass filter with a cutoff frequency of 70 MHz) and sent to an oscilloscope (DSO-X 0334A, Agilent Technologies, CA, USA). The oscilloscope acquire the data using a sampling rate of 2 GHz. This data is sent to a computer to be analyzed using a numerical software.

B. Evaluation of Wavelet Denoising Methods

The RF response of the phantom was visualized and saved by the oscilloscope using two different acquisition modes for:

- Single mode: the oscilloscope captures single waveform of the input channel after the trigger event.
- M-Average mode: the oscilloscope captures M consecutive waveforms, synchronized with the trigger events. The point-by-point average of these waveforms is calculated and visualized. This acquisition mode is used to attenuate noise effects in the visualized signal.

Figs 4 and 5 show the RF response of the GVs acquired using the single and 128-average acquisition modes of the oscilloscope. The RF signal acquired using the single-acquisition mode is contaminated by noise and is used to evaluate the performance of the wavelet-based denoising methods. In contrast, the RF signal acquired using the 128-average acquisition mode exhibits a significantly higher signal-to-noise ratio and is adopted as the reference RF signal (an approximation of the RF signal in the absence of noise).

The objective of the wavelet-based filtering process is to reduce the noise in the signal shown in Fig. 4, obtaining a filtered signal that closely matches the reference waveform presented in Fig. 5. Different mother wavelets and decompo-

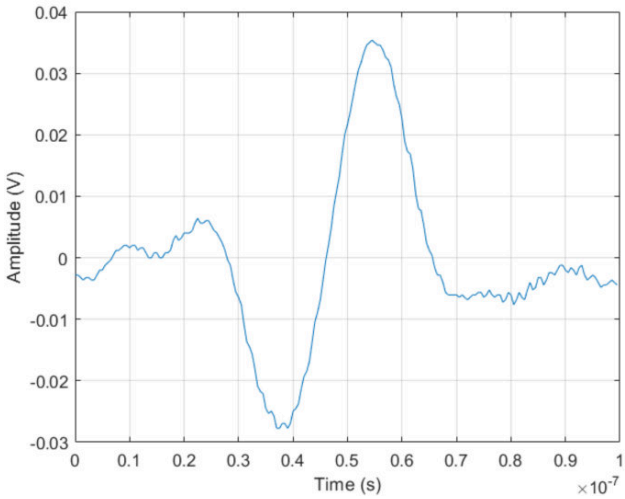


Fig. 4. RF response of GVs acquired using single acquisition mode.

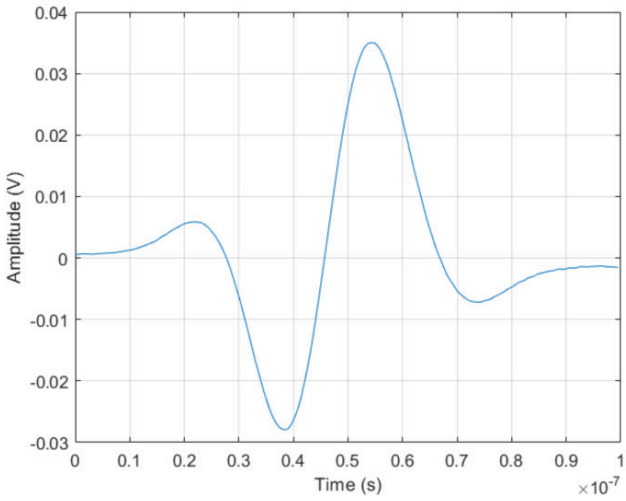


Fig. 5. RF response of GVs acquired using 128-average acquisition mode.

sition levels will be tested to evaluate the influence of these parameters into the RF signal denoising.

IV. Results

The RF signals of the GV acoustic responses were analyzed in MATLAB. The command *wden* was used to apply different types of wavelet filters. The option *sqtwolog* of this command allows using the hard thresholding method defined in (9) and (10).

Figs. 6–17 show the results for different mother wavelets (Coiflet 1, Coiflet 4, Daubechies 2, Daubechies 3, Symlet 3 and Symlet 8), and for 3 and 6 decomposition levels. According to these results, Coiflet 1 and Daubechies 2 produce spikes at the filtered pulse peak (instead of the soft peak presented in Fig. 5, as shown in Figs. 6, 7, 10, and 11. Mother wavelets with too low vanishing moments are not adequate for RF signal denoising.

The pulse peaks in Figs. 11 and 12 present an interesting case: using mother wavelets whose vanishing moments differ

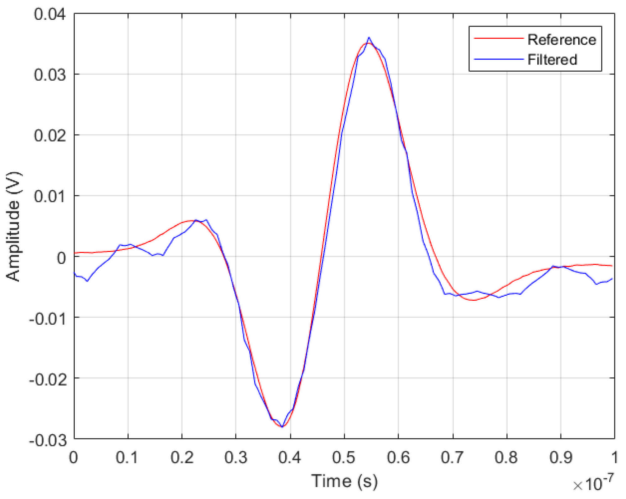


Fig. 6. Result for Coiflet 1, 3 decomposition levels.

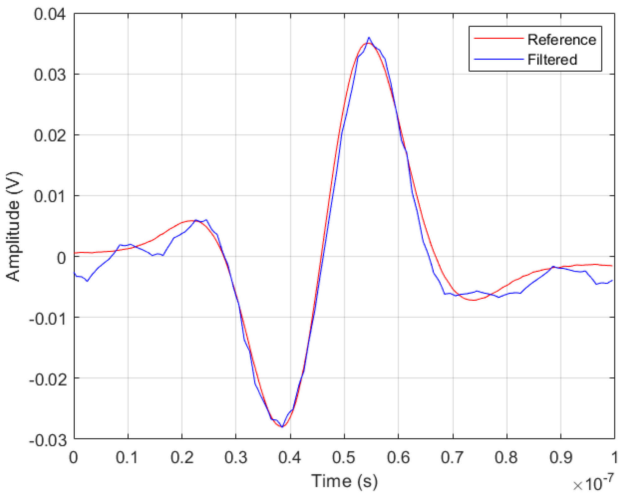


Fig. 7. Result for Coiflet 1, 6 decomposition levels.

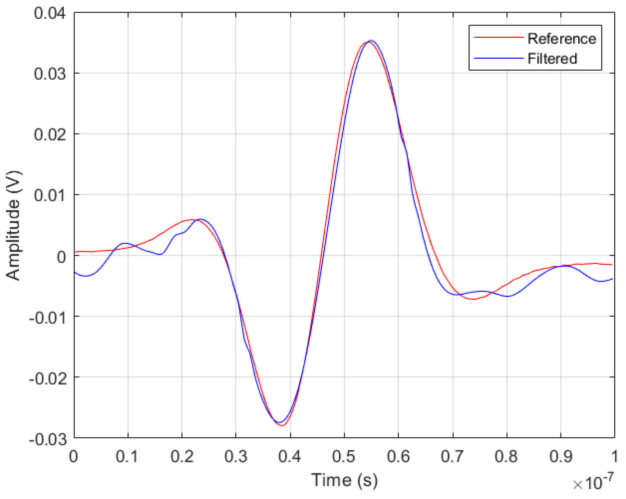


Fig. 8. Result for Coiflet 4, 3 decomposition levels.

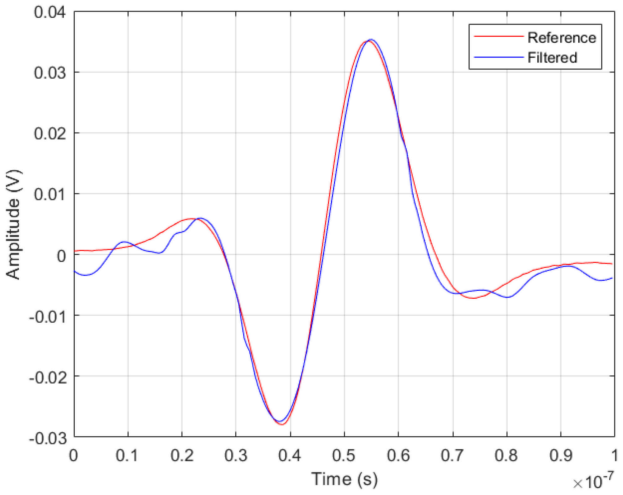


Fig. 9. Result for Coiflet 4, 6 decomposition levels.

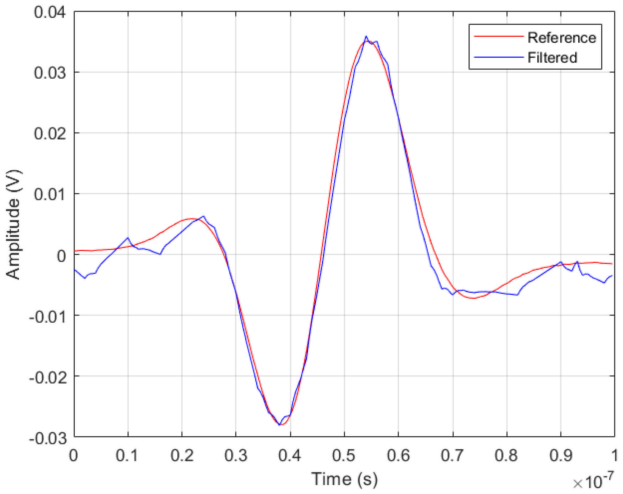


Fig. 10. Result for Daubechies 2, 3 decomposition levels.

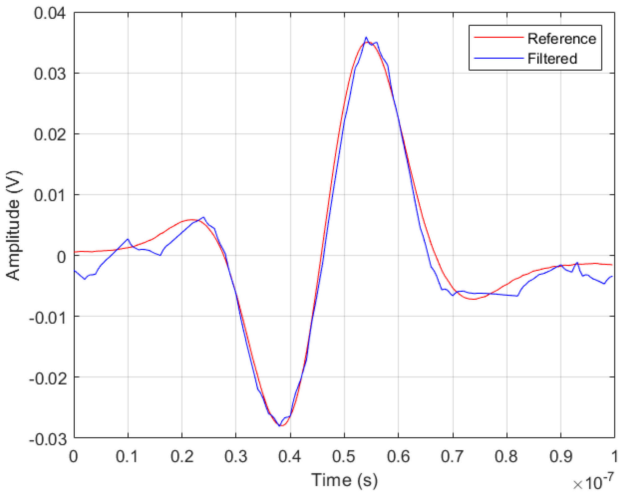


Fig. 11. Result for Daubechies 2, 6 decomposition levels.

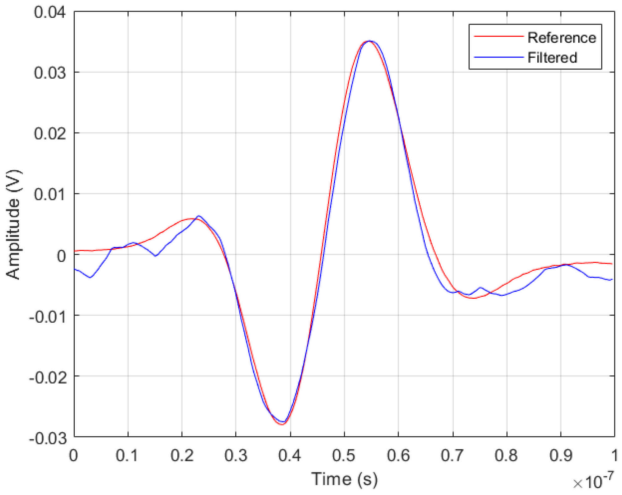


Fig. 12. Result for Daubechies 3, 3 decomposition levels.

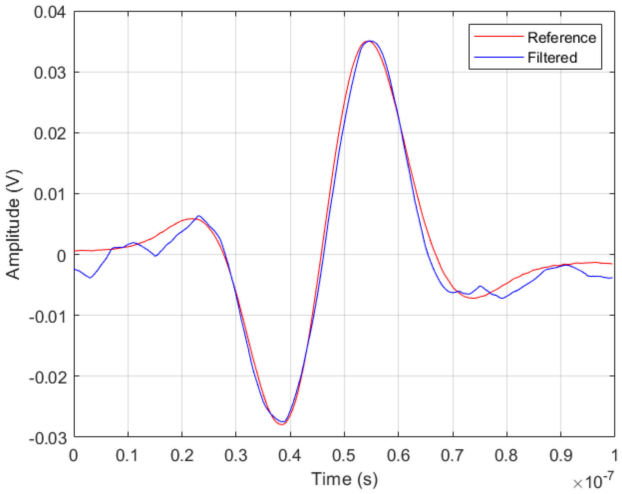


Fig. 13. Result for Daubechies 3, 6 decomposition levels.

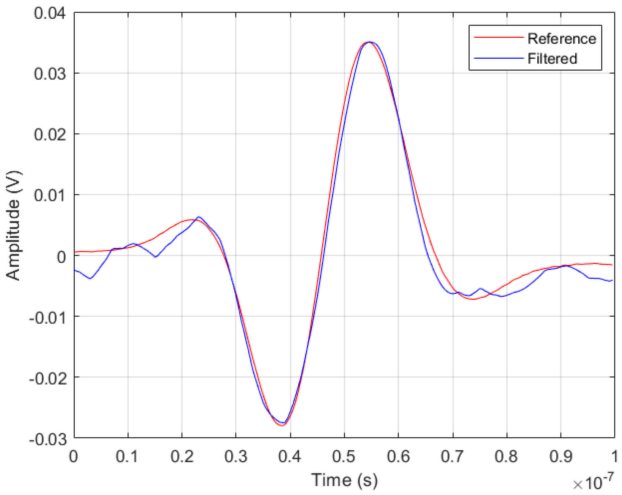


Fig. 14. Result for Symlet 3, 3 decomposition levels.

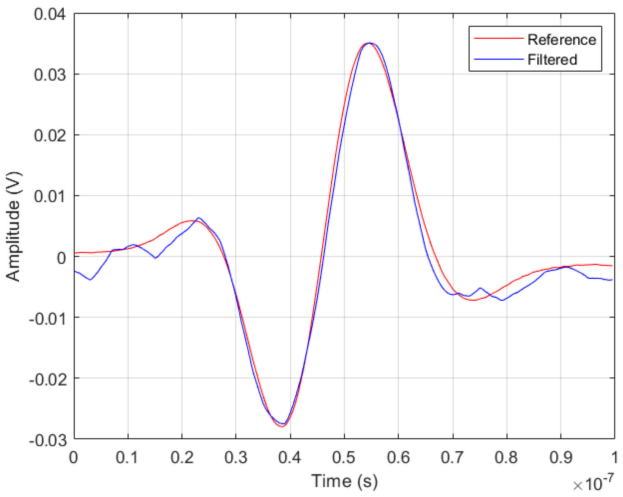


Fig. 15. Result for Symlet 3, 6 decomposition levels.

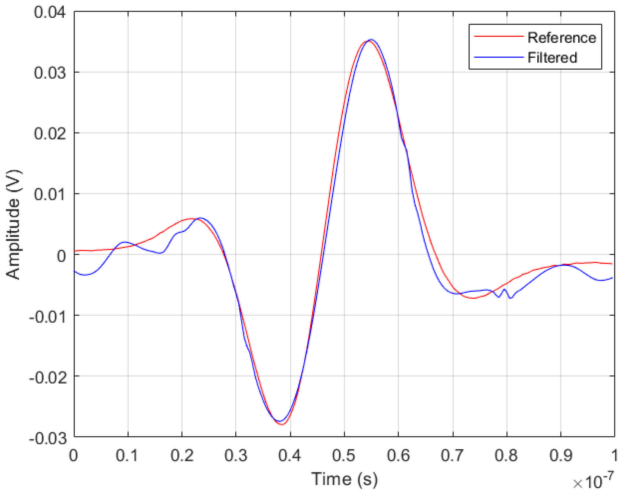


Fig. 16. Result for Symlet 8, 3 decomposition levels.

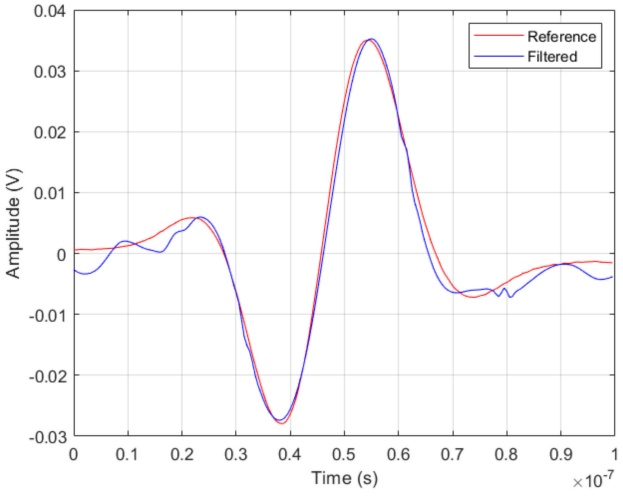


Fig. 17. Result for Symlet 8, 6 decomposition levels.

by one unit can produce different filtering performance. For that reason, the selection of an adequate vanishing moment of the mother wavelets is important. A mother wavelet with low vanishing moment gives inadequate filtering results. However, a mother wavelet with a high vanishing moment produces low-pass and high-pass filters with many coefficients, increasing the computational complexity of the denoising process.

Wavelet filtering (using mother wavelets with adequate vanishing moments) allows reducing the high-frequency noise, with a negligible delay, and the pulse shape of the filtered signal is similar to the pulse of the reference waveform given in Fig. 5. After the pulse, sinusoidal perturbation is still present before and after the RF pulse. Other techniques such a notch filter could be used to remove that interference. However, usually the information of the GV responses is concentrated in the pulse whose shape is recovered through wavelet filtering.

Fig. 18 shows the result when a first-order stop-band Butterworth filter (with cut-off frequencies of 55 MHz and 65 MHz) is used for filtering the RF pulse. The delay between the reference signal and the filtered signal has increased compared to the results obtained through wavelet filters, since a linear filter adds a delay to the signals. Moreover, high-frequency oscillations can still be observed in the filtered signal.

Table I shows the signal-to-noise ratio (SNR), the root mean square error ($RMSE$) and the correlation coefficient (C_{corr}) for each test. These results show that increasing the number of decomposition levels (from 3 to 6) resulted in only a marginal improvement in the *SNR*, the correlation coefficient and reduction in the *RMSE*. Filtering using Symlet 8 and 6 decomposition levels gives the highest SNR and correlation coefficient, and the lowest $RMSE$.

The processing time of the wavelet filter based on Symlet 8 (which has the best performance) for 3 and 6 decomposition levels was measured. For each decomposition level, the command *wden* was executed 1000 times and the average time was computed. The results in Table II prove that using more decomposition levels increases the processing time. However,

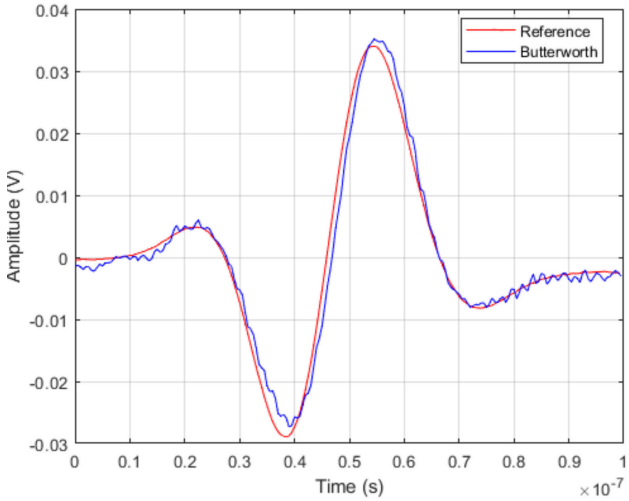


Fig. 18. Result for the Butterworth filter.

TABLE I
SNR, $RMSE$ AND C_{corr} FOR THE TESTS

Wavelet	***SNR***	***RMSE*** $\times 10^3$	C_{corr}
Coiflet 1, level 3	14.311	1.330	0.982
Coiflet 1, level 6	15.775	1.124	0.987
Coiflet 4, level 3	14.163	1.353	0.981
Coiflet 4, level 6	15.685	1.135	0.986
Daubechies 2, level 3	14.252	1.339	0.981
Daubechies 2, level 6	15.984	1.097	0.987
Daubechies 3, level 3	14.240	1.341	0.981
Daubechies 3, level 6	15.820	1.118	0.987
Symlet 3, level 3	14.240	1.341	0.981
Symlet 3, level 6	15.820	1.118	0.987
Symlet 8, level 3	14.166	1.352	0.981
Symlet 8, level 6	16.359	1.051	0.988
No filter	13.561	1.450	0.978
Butterworth	13.801	1.409	0.979

TABLE II
PROCESSING TIME OF THE WAVELET FILTER

Wavelet	Processing time (ms)
Symlet 8, level 3	0.681
Symlet 8, level 6	1.026

Table I shows that increasing the decomposition level could improve the denoising process. For that reason, it is important to select properly the number of decomposition levels.

V. CONCLUSION

This paper evaluated the performance of different wavelet filters for denoising the RF signals obtained from the acoustic responses of the GVs. Mother wavelets with low vanishing moments are not adequate for the filtering of the RF signals. In general, wavelets with a vanishing moment greater than or equal to three can be used. However, a high vanishing moment gives low-pass and high-pass filters with many coefficients, increasing the computational complexity of the wavelet denoising process. Filtering using the Symlet 8 wavelet with 6 decomposition levels provides the best noise reduction. In all tests, increasing the number of decomposition levels resulted in only a marginal improvement in the denoising process. However, using fewer decomposition levels reduces the required processing time. Hence, the number of decomposition levels should be carefully selected to achieve an appropriate tradeoff between noise reduction performance and computational cost. As future work, other kinds of filters (e.g., notch filters) will be applied together with the wavelet denoising process to improve the denoising of the GVs responses.

ACKNOWLEDGMENT

The authors would like to thank the Ultrasound Laboratory (LUS) and the Biomedical Engineering Program of COPPE/UFRJ for providing the infrastructure, technical support, and research environment that made this work possible.

REFERENCES

[1] K. Yang *et al.*, "Frequency-Aware B-Line and Pleural Line Analysis in Lung Ultrasound Videos," *IEEE J. Biomed. Health Inform.*, vol. 30, no. 6, pp. 5220–5232, Jun. 2026.

[2] Y. Ren, G. Zhai, Y. Chen, S. Wang, X. Du and Z. Meng, "Wavelet-Enhanced Adaptive Representation Learning for Deep-Sea Manganese Nodule Image Segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, Art no. 5627413, pp. 5627413-5627413, 2026.

[3] J. Sushilkumar and S. Chandra, "Deep PA-PCNN Fusion Framework with DWT for Multimodal Biomedical Imaging," 2025 IEEE 9th International Conference on Information and Communication Technology (CICT), Chennai, India, 2025, pp. 1–6.

[4] A. Bin Rahman, M. Ibn Afjal and M. A. Al Mamun, "Mitigating Noise from Biomedical Images Using Wavelet Transform Techniques," 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 2025, pp. 1-6.

[5] A. J. Fernandes, Y. Ono and E. Ukwatta, "Evaluation of Finger Flexion Classification at Reduced Lateral Spatial Resolutions of Ultrasound," *IEEE Access*, vol. 9, pp. 24105-24118, 2021.

[6] H. D. Le, T. T. Mai, Q. Lin, X. Li, J.-J. Min, and C. Lee, "GPU-accelerated volumetric-mosaic optical-resolution photoacoustic microscopy and quantifying tumor vasculature growth," *Photoacoustics*, vol. 46, Art. no. 100786, 2025.

[7] S. Bak, S. M. Park, Y. Song, J. Kim, T. W. Nam, D.-W. Han, C.-S. Kim, S.-W. Cho, B. E. Bouma, and H. Lee,"High spectral energy density all-fiber nanosecond pulsed 1.7 μm light source for photoacoustic microscopy," *Photoacoustics*, vol. 44, Art. no. 100744, 2025.

[8] Y. Wang, Y. Chen, Y. Zhao, and S. Liu, "Compressed sensing for biomedical photoacoustic imaging: a review," *Sensors*, vol. 24, no. 9, Art. no. 2670, 2024.

[9] K. Jeong et al., "A Wide-Bandwidth Ultrasound Receiver and On-Chip Ultrasound Transmitter for Ultrasound Capsule Endoscopy," *IEEE J. Solid-State Circuit*, vol. 58, no. 10, pp. 2778–2789, Oct. 2023.

[10] A. Lakshmanan, Z. Jin, S. P. Nety, D. P. Sawyer, A. Lee-Gosselin, D. Malounda, M. B. Swift, D. Maresca, and M. G. Shapiro, "Acoustic biosensors for ultrasound imaging of enzyme activity," *Nat. Chem. Biol.*, vol. 16, no. 9, pp. 988–996, Sep. 2020,

[11] I. Dev, A. Baskota, M. Shapiro and A. Lal, "Combining gas vesicle acoustic contrast agents with field-deployable GHz frequency ultrasonic chips to detect gene expression in sentinel microorganisms," 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS), Taipei, Taiwan, 2024, pp. 1–4.

[12] M. R. da Silva, F. M. C. P. de Jesus, F. V. Garrute, and J. C. Machado, "Exploring the Impact on Collapse and Ultrasound Backscattered Signal Intensity with Varying Ultrasound Power Incident on Gas Vesicle Used for Contrast Agent," in Proc. 2024 IEEE UFFC Latin America Ultrasonics Symposium (LAUS), 2024

[13] G. R. Sahoo, J. H. Freed, and M. Srivastava, "Optimal Wavelet Selection for Signal Denoising," *IEEE Access*, vol. 12, pp. 45369–45380, 2024.

[14] S. N. Syakiylla et al., "Wavelet Based Filters for Artifact Elimination in Electroencephalography Signal: A Review," *Biomedical Eng. Online*, vol. 21, no. 1, 2022.

[15] A. J. R. Reis and A. P. Alves da Silva, "Aplicação da transformada wavelet discreta na previsão de carga a curto prazo via redes neurais," *Revista Controle & Automação*, vol. 15, no. 1, pp. 101–112, 2004.

[16] P. Tan, G. Tan and Z. Cai, "Dual-tree complex wavelet transform-based feature extraction for brain computer interface," *2015 12th Int. Conf. Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, China, 2015, pp. 1136–1140.

[17] D. L. Donoho and I. M. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, Sept. 1994.

[18] A. Lakshmanan et al., "Preparation of biogenic gas vesicle nanostructures for use as contrast agents for ultrasound and MRI," *Nat. Protoc.*, vol. 12, no. 10, pp. 2050–2080, Oct. 2017.

Experimental Simulation and Radiographic Analysis of Fracture Patterns Associated with Child Physical Abuse

Fernanda Lorena Slusarczuk Manchenho
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil.
ORCID: 0009-0007-1035-6456

Akemi Yagui
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0002-9152-6790

Rosangela Requi Jakubiaki
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0002-1645-1114

Rubens Alexandre de Faria
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0002-1645-1112

Abstract— **Child physical abuse is a major public health concern and one of the leading causes of preventable morbidity and mortality in the pediatric population. Forensic radiology plays a critical role in the identification and documentation of skeletal injuries associated with non-accidental trauma. This study aimed to experimentally reproduce fracture patterns commonly associated with child physical abuse and evaluate the influence of radiographic exposure parameters on image quality and diagnostic performance. An experimental study was conducted using a juvenile porcine model (Sus scrofa domesticus), selected due to its anatomical and biomechanical similarities to the pediatric skeleton. Spiral, greenstick, rib, and cranial fractures were experimentally induced and assessed using a GE XR-6000 and direct digital radiography detector. Different exposure protocols were evaluated through adjustments using the 15% rule, enabling comparative analysis of image quality and fracture visualization. The results demonstrated that the porcine model successfully reproduced radiographic patterns consistent with those reported in the literature for suspected child abuse. Optimized exposure parameters improved image quality and fracture detection, while inadequate techniques increased image noise and reduced diagnostic performance. The study also highlighted the importance of proper positioning and the acquisition of complementary radiographic projections for accurate lesion characterization. These findings reinforce the need for individualized pediatric exposure charts and adherence to the ALARA principle (As Low As Reasonably Achievable), contributing to improved diagnostic quality, reduced repeat examinations, and enhanced radiation protection in pediatric forensic radiology.**

Keywords - Forensic Radiology; Child Abuse; Pediatric Radiology; Exposure Charts; Radiation Protection; Digital Radiography.

I. INTRODUCTION

Since the discovery of X rays by Wilhelm Conrad Roentgen in 1895, forensic radiology has become an important tool for trauma assessment and medico-legal investigations, including the identification of injuries related to child abuse [1]. Pediatric forensic radiology is particularly relevant for detecting fractures at different stages of healing and documenting findings compatible with non-accidental trauma [2].

Between 2020 and 2023, violence-related records increased in Brazil across the Information System for Children and Adolescents (Sistema de Informação para a Infância e Adolescência – SIPIA), the Notifiable Diseases Information System (Sistema de Informação de Agravos de Notificação – SINAN), and Disque 100. SIPIA monitors violations of the rights of children and adolescents; SINAN compiles mandatory notifications submitted by health services; and Disque 100 receives reports of human rights violations. Reports registered through Disque 100 increased from 366,032 in 2020 to 1,311,044 in 2023. Although these databases may contain overlapping records, the data demonstrate the magnitude of violence and the need for prevention, early identification, reporting, and victim protection [3], [4].

Radiological assessment contributes to fracture identification, evaluation of trauma mechanisms, injury chronology, and comparison between imaging findings and the reported clinical history, supporting the differentiation between accidental and non-accidental trauma [4], [5]. The fracture patterns selected for this study were defined according to their forensic relevance in suspected child

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

physical abuse. Spiral fractures present a helical fracture line produced by torsional forces; greenstick fractures are incomplete injuries of immature bone, characterized by cortical disruption and bending; posterior rib fractures are associated with thoracic compression; and cranial fractures may result from direct impact or acceleration–deceleration mechanisms. The interpretation of these findings must consider the child's age, developmental stage, clinical history, associated injuries, and consistency between the reported mechanism and radiological findings [6], [7].

The differentiation between accidental and abusive injuries requires integrated clinical, radiological, and medico-legal assessment, including the exclusion of conditions associated with bone fragility. Conventional radiography, computed tomography, and magnetic resonance imaging are complementary methods for identifying occult fractures and associated injuries [7]. Proper documentation, ethical conduct, and a child-centered approach are essential during imaging examinations [8], [9]. Radiation protection is particularly important in pediatric imaging because of increased tissue radiosensitivity and longer lifetime risk; therefore, the principles of justification, optimization, and dose limitation, as well as pediatric-specific protocols and individualized exposure adjustments, must be applied [10].

In addition to diagnostic aspects, professional conduct in suspected cases of child abuse requires ethical, technical, and legal rigor, as radiologists and healthcare professionals play a strategic role in identifying, documenting, and reporting findings [9]. Proper documentation of injuries through precise descriptions and high-quality imaging is essential to support medico-legal investigations and ensure evidentiary integrity. Simultaneously, a humanized approach is indispensable during examinations, considering the physical and emotional vulnerability of abused children. This includes patient welcoming, minimization of suffering, respect for dignity, and strategies aimed at reducing fear, pain, and revictimization during procedures [10]. From a technical perspective, the fundamental principles of radiological protection are particularly relevant in pediatric populations due to greater tissue radiosensitivity and longer lifetime risk. Therefore, the principles of justification, optimization, and dose limitation must be strictly applied to ensure safe imaging practices. Optimization must be widely understood and applied by professionals in order to ensure diagnostic image quality while keeping radiation doses as low as possible, since children are more sensitive to radiation and have a longer life expectancy [11].

The X ray absorption and transmission depend primarily on the composition and, above all, the thickness of the body part being imaged; patient age is a poor surrogate for thickness [12]. Pediatric imaging protocols must therefore be designed for each anatomical region and individualized according to patient anthropometry. Standardized exposure protocols should be used as reference guidelines but must allow adjustments for patient body thickness, positioning and clinical indication [13], [14]. And also, the protocols should be specific to each detector [12]. Optimized protocols also require careful beam collimation, appropriate selection of exposure parameters, and periodic review based on radiation dose indicators and image quality assessment. These measures reduce exposure variability, minimize dose to radiosensitive organs, prevent unnecessary repeat examinations, and support establishment of pediatric Diagnostic Reference Levels (DRLs) [13], [14].

Identifying the correct protocol for obtaining an X ray is not a simple task. And it is also necessary to consider the lack of experience or training of the radiology professional in optimizing image quality while managing radiation doses in different technologies from different manufacturers with different sensitivity of the detection system [15]. Therefore, there are many failures in choosing the ideal protocol, and it is necessary to repeat exposures, often by trial and error. This causes an increase in the patient's doses. Then it is important to ensure all radiation doses from an imaging examination are kept As Low As Reasonably Achievable (ALARA) [16].

In this context, the optimization of radiographic exposure parameters becomes essential to ensure diagnostic image quality while minimizing unnecessary radiation exposure in pediatric patients. One practical approach for adjusting exposure techniques is the 15% rule. According of Carrol (2007), the "15% rule" (radiography): Increasing the tube voltage (kVp) by about 15% produces approximately the same effect on receptor exposure as doubling the tube current time product (mAs). Conversely, decreasing kVp by 15% is roughly equivalent to halving the mAs. The 15% rule has derivations for when the necessary changes in the exposure technique are subtle. It is also possible to change around 10% in the kV value, is roughly equivalent to modify the mA in 30%. When applied appropriately, this principle may assist in selecting exposure parameters according to the anatomical region and diagnostic objective, contributing to image-quality optimization, reduction of repeat examinations, and compliance with the ALARA principle [17].

This rule has some limitations and caveats: the rule is empirical and approximate; exact relation depends on generator type (single vs. high frequency), beam filtration, kVp range, receptor sensitivity, patient thickness/composition, geometry; at very low or very high kVp ranges the approximate 15% doubling relation may not hold; for digital detectors with wide dynamic range, maintaining "exposure index" matters; image processing may mask exposure changes. Excessive post-processing should always be avoided. The 15% rule is an important tool for maintaining diagnostic image quality and reducing patient dose, as mAs is the main parameter responsible for dose contribution [17].

When acquiring an image, it is necessary to evaluate which parameter should be changed. kVp directly impacts beam penetrability, and mAs impacts the noise pattern. These parameters must be carefully manipulated to correct exposure flaws that compromise lesion visualization. However, the best balance must also be considered to keep the dose to the pediatric patient as low as possible [17].

Increasing kVp and reducing mAs typically reduces patient entrance skin dose and absorbed dose for many exams (less absorption of low energy photons), but may increase dose to deeper tissues; net effect usually lower dose for same image receptor exposure. Beware: if mAs is not reduced appropriately, patient dose will increase [17].

The protocol adopted in this study was to work with fixed kVp values. That is, to identify the energy value that provides adequate contrast for that anatomical region (optimal kVp).

According with Carrol (2011), a fixed kVp protocol is therefore composed of the selection of a kVp value considered optimal – the value required to penetrate a given anatomical part, with the production of long-scale contrast

radiograph. When the kVp is determined for each examination, it does not vary with the difference in thickness in the anatomical part. The change will occur in the mAs value used, to obtain the desired noise level. The primary goals of the fixed kVp method, is to (1) assure sufficient penetration, (2) keep image contrast consistent, and (3) simplify and standardize for accuracy [18].

In light of this context, this study aimed to investigate the experimental reproduction and radiographic characterization of fracture patterns associated with suspected child physical abuse using a juvenile porcine model, with emphasis on evaluating the application of the 15% rule for the optimization of radiographic exposure parameters. The study sought to improve the diagnostic quality of the images while preserving the visualization of osseous findings and using the lowest radiation doses reasonably achievable.

II. Methodology

The experimental phase of this study was conducted to reproduce radiographic fracture patterns frequently described in the literature as suggestive of child physical abuse and to evaluate the influence of technical exposure parameters on the diagnostic quality of the acquired images. All procedures were performed at DAPI Clinic, located in Curitiba, Paraná, Brazil, using a GE XR-6000 radiographic system (General Electric Healthcare) equipped with a direct FUJI D-EVO flat-panel detector for digital radiography (DR). The equipment was manufactured in 2012 and installed at the institution in 2013, remaining in regular clinical use for radiographic examinations. All radiographic acquisitions were performed using equipment operating under the institution's routine quality assurance program, in accordance with the applicable requirements established by Brazilian Health Regulatory Agency (ANVISA) Resolution RDC No. 611/2022, ensuring proper system performance and quality control throughout the experimental procedures [19]. The system enabled immediate image acquisition, reduced processing time, and the use of post-processing tools, including brightness and contrast adjustment and image magnification. These features were particularly relevant to the present study, as they enabled comparative analysis among different exposure protocols and detailed assessment of the simulated injury patterns.

Given the need for an experimental model comparable to the pediatric skeleton, a juvenile pig (*Sus scrofa domesticus*), approximately five weeks of age, weighing about 14 kg and measuring approximately one meter in length, was obtained from a commercial slaughterhouse after routine commercial slaughter for food production and used exclusively as an experimental anatomical model. The juvenile porcine specimen was selected because its skeletal characteristics allow the experimental reproduction of fracture patterns in immature bone, making it suitable for evaluating radiographic acquisition techniques under controlled conditions. The use of commercially obtained post-mortem biological material enabled the standardized simulation of different fracture mechanisms under controlled conditions, ensuring reproducibility throughout the experimental procedures. No live animals were subjected to restraint, anesthesia, euthanasia, or any other experimental procedures for research purposes. Accordingly, in compliance with Brazilian Federal Law No. 11,794/2008 (Arouca Law) and the Brazilian Guidelines for the Care and Use of Animals in Teaching or Scientific Research Activities (DBCA), the study was considered exempt from review by the Animal Use Ethics Committee (CEUA), as it did not constitute animal experimentation under the applicable legislation [20].

Following the definition of the experimental model, four fracture patterns frequently described in the specialized literature and recognized for their relevance in the radiological investigation of suspected child physical abuse were selected: spiral fractures, greenstick fractures, rib fractures, and cranial fractures. The inclusion of these lesions aimed to encompass different biomechanical mechanisms and radiographic patterns encountered in clinical practice.

For spiral fracture induction the limb segment was stabilized using metallic clamps and subjected to controlled torsional force using a pipe wrench until a fracture line was produced and for greenstick fracture induction the selected bone segment was subjected to controlled bending force using a pipe wrench, aiming to produce incomplete cortical disruption associated with angular deformation (Fig. 1).

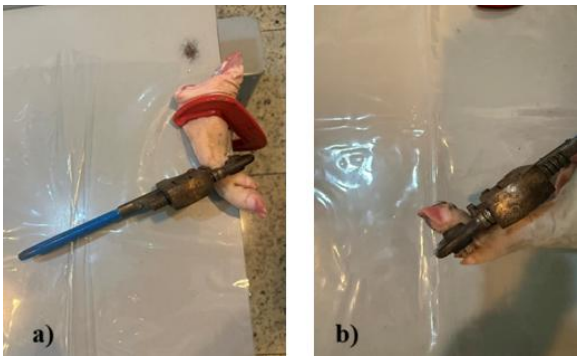


Fig. 1 - Simulation of a) Spiral Fracture and b) Greenstick Fracture.

Rib fractures were simulated by applying controlled blunt compressive force with the rib cage positioned on a rigid surface, reproducing the compression mechanism of pediatric blunt thoracic trauma. Cranial fractures were induced by a controlled perpendicular impact with a metallic hammer to the cranial vault of a juvenile porcine skull supported on a rigid surface. Fig. 2 presents the experimental simulation of these trauma mechanisms, showing (a) blunt compressive force applied with a baseball bat to induce a rib fracture and (b) a controlled impact with a metallic hammer to simulate a cranial fracture. The figure was digitally modified using an artificial intelligence (AI) exclusively to improve the visual representation of the injury-production mechanisms and facilitate understanding of the experimental setup [21].

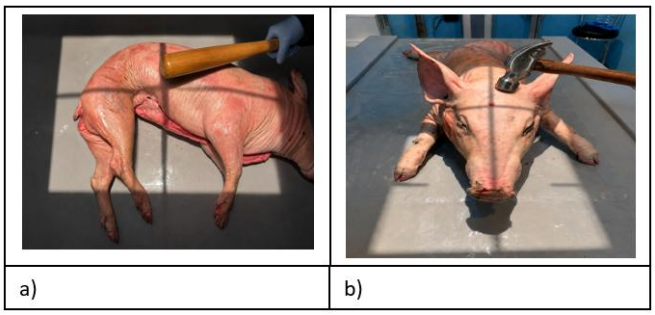


Fig. 2 - Simulation of a) Rib fracture, b) Cranial fracture

Radiographic examinations were performed after lesion induction in each anatomical region evaluated. For the upper and lower limbs, posteroanterior (PA) and lateral (P) projections were obtained; for the rib fractures, an anteroposterior (AP) projection was performed; and for the skull, PA and P projections were acquired. The complementary projections were used to assess fracture-line morphology, cortical discontinuity, anatomical alignment, possible displacement, and lesion extent from different planes, thereby reducing the limitations associated with anatomical superimposition. Image acquisition was initially performed using the exposure charts routinely adopted by DAPI Clinic for each anatomical region.

Exposure parameter optimization was performed through sequential adjustments of tube potential (kVp) and tube current–time product (mAs), using the initial exposure chart adopted for each anatomical region as the reference. Following the acquisition of the baseline radiograph, image quality was visually assessed based on radiographic contrast, cortical definition, fracture-line conspicuity, and the presence of quantum noise. Whenever image limitations related to beam penetration or contrast were identified, kVp adjustments were performed using the 15% rule as an operational guideline, with proportional mAs modifications whenever necessary to maintain adequate detector exposure. When excessive image noise or insufficient photon flux was observed, only the mAs value was adjusted while maintaining a constant tube potential. This sequential optimization procedure was repeated for each anatomical region and fracture pattern until the combination providing the best diagnostic visualization of the simulated lesions was achieved.

The radiographic images were qualitatively evaluated by two experienced radiologic imaging professionals through visual comparison on a 2 MP diagnostic display. The assessment considered fracture pattern visualization, cortical definition, image sharpness, radiographic contrast, image density, the presence of quantum noise, and visualization of anatomical structures. During protocol optimization, only one exposure parameter was modified at a time, beginning with the parameter considered least appropriate while maintaining the other unchanged. Successive adjustments of kVp and mAs were performed until the combination providing the best lesion visualization was identified. The optimized technique was defined as the protocol that achieved the best balance between image sharpness, radiographic contrast, image density, noise reduction, and fracture-line visualization.

Because this investigation was designed as a preliminary proof-of-concept study aimed at experimentally reproducing pediatric fracture patterns and optimizing radiographic exposure parameters, image quality assessment was intentionally based on qualitative visual analysis. Therefore, objective image quality metrics, including the Exposure Index (EI), Deviation Index (DI), Signal-to-Noise Ratio (SNR), Contrast-to-Noise Ratio (CNR), radiation dose measurements, and formal interobserver agreement analysis, were not included. Qualitative visual assessment was considered appropriate for the scope of this preliminary study. Future phases of the project will incorporate quantitative image quality metrics, dosimetric assessment, and interobserver agreement analysis to strengthen the methodological robustness and metrological validation of the proposed radiographic protocol.

The kV and mAs combinations were organized into a comparative exposure table, including the initial protocols, PA, AP, and P projections, as well as the parameters submitted for evaluation, as presented in TABLE I.

TABLE I - Exposure Charts and Optimized Technical Parameters.

Region	Projection	Techniques Evaluated (kV/mAs)	Optimized (kV/mAs)
Wrist (Spiral/Greenstick)	PA	40/4 44/12.5	44/12.5
Wrist (Spiral/Greenstick)	P	44/8	44/8
Ribs	AP	65/40 55/40 55/80 60/80 55/100 60/100	55/100
Skull	PA	65/10 65/20	65/20
Skull	P	55/5 65/10	65/10
Lower Limb	AP	44/12.5 50/12.5 44/20	44/20

III. RESULTS and DISCUSSION

The present study demonstrated that the experimental model using a juvenile porcine specimen was capable of reproducing radiographic patterns comparable to those described in the literature for fractures frequently associated with suspected child physical abuse. Fracture patterns consistent with spiral fractures, greenstick fractures, rib fractures, and cranial fractures were successfully simulated, allowing not only the morphological characterization of these injuries but also the evaluation of how exposure parameters, patient positioning, and radiographic projections influenced image quality and fracture visualization.

Although the experimental model provided a controlled and reproducible environment for fracture simulation, some limitations were observed and should be considered when extrapolating the findings to clinical pediatric practice. Anatomical and histological differences between porcine and human bone, as well as variations in bone geometry and surrounding soft tissues, may influence fracture morphology and radiographic appearance. Therefore, the optimized exposure parameters obtained in this study should be interpreted as experimental references rather than universal clinical protocols. Nevertheless, the reproducibility of the proposed model and its ability to simulate different fracture mechanisms support its application in experimental investigations, professional training, and the development of radiographic protocols for pediatric forensic imaging [22-24].

Furthermore, the literature review conducted for this study did not identify recent publications specifically investigating the use of juvenile porcine models to reproduce fracture patterns associated with suspected child physical abuse in combination with the optimization of radiographic exposure parameters. Although juvenile porcine models have been employed in orthopedic, biomechanical, and trauma research, their application integrating the simulation of

abuse-related fracture mechanisms with the systematic evaluation of radiographic techniques appears to remain scarcely explored in the recent literature. This gap reinforces the originality of the experimental approach adopted in the present study and indicates its potential contribution to future research, professional training, and the development of optimized imaging protocols in pediatric forensic radiology.

The simulated lesions exhibited radiographic characteristics consistent with patterns described in the literature, allowing the evaluation not only of fracture identification but also of factors related to image acquisition. Throughout the study, it was observed that obtaining diagnostically adequate images depends directly on the combination of optimized radiographic technique, correct positioning of the anatomical segment, and the acquisition of complementary radiographic projections.

This study reinforced the importance of obtaining at least two orthogonal radiographic projections for the assessment of bone injuries, as fracture visualization was limited when based on a single projection due to anatomical superimposition or the orientation of the fracture line relative to the X ray beam. The use of complementary projections enabled a more accurate assessment of fracture extent, bone alignment, fragment displacement, and the spatial relationship between affected structures, thereby increasing diagnostic reliability and reducing the likelihood of false-negative findings, in accordance with international recommendations for the radiographic investigation of suspected child physical abuse.

During the simulation of spiral fractures, radiographically, a helical fracture line was observed, characterized by sharp margins and rotational displacement of the distal fragment, consistent with a torsional mechanism of injury. Fig. 3 shows the simulated spiral fracture in the posteroanterior (a) and lateral (b) projections, demonstrating the characteristic helical fracture line and improved visualization of the fracture trajectory using complementary projections.

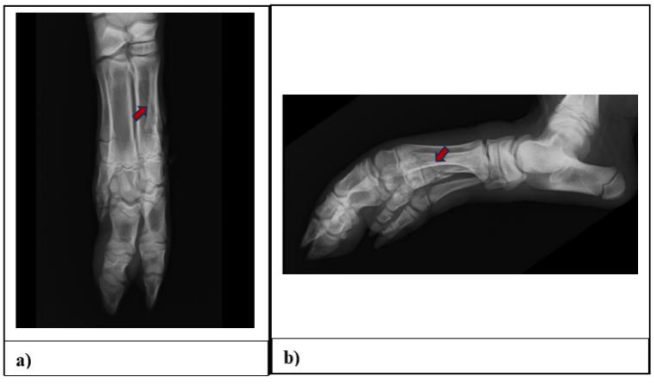


Fig. 3 - Spiral Fracture: a) PA Projection and b) Lateral Projection.

This type of fracture is associated with rotational forces acting along the longitudinal axis of the bone, resulting in progressive cortical shear stress. Accurate identification of this pattern was facilitated by the use of complementary projections and careful selection of technical parameters.

The initial exposure protocol was 40 kVp/4 mAs, resulting in insufficient beam penetration, high image noise, and poor visualization of the fracture extent. The tube potential was then increased to 44 kVp, while maintaining the 4 mAs setting. Although this adjustment provided adequate beam penetration, image noise remained high. Therefore, the mAs value was doubled to 8 mAs, resulting in improved cortical definition and enhanced visualization of the fracture lines.

Greenstick fractures exhibited characteristics compatible with those observed in immature skeletons, demonstrating partial disruption of one cortex associated with deformation of the opposite cortex. This pattern is directly related to the increased elasticity and higher collagen content of pediatric bone tissue. Although it presents lower specificity for non-accidental trauma, its association with other suspicious findings may significantly contribute to the overall assessment of the case and the interpretation of the clinical-forensic context. Radiographic evaluation demonstrated angular deformity and unilateral cortical disruption, while the periosteum remained intact on the opposite side, a characteristic finding of incomplete fractures in immature bone. Fig. 4 shows the simulated greenstick fracture in the posteroanterior (a) and lateral (b) projections, demonstrating incomplete cortical disruption and angular deformity characteristic of immature bone.

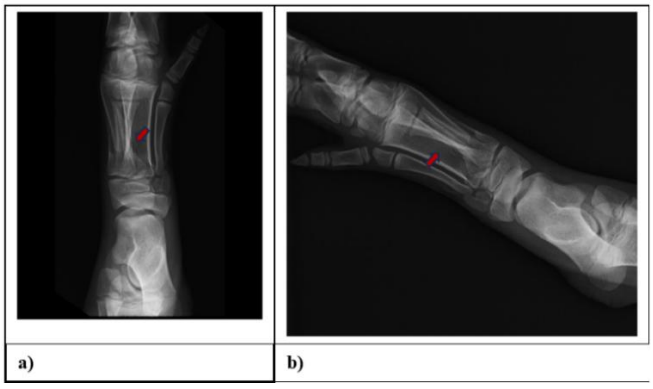


Fig. 4 - Greenstick Fracture: a) PA Projection and b) Lateral Projection.

Regarding the simulated rib fractures, small adjustments in radiographic parameters directly influenced the ability to identify fracture lines. The initial exposure protocol was 65 kVp/40 mAs, with a low contrast scale and difficulty in characterizing the fractures. The kVp was reduced by 15% (55 kVp) and the mAs value was maintained. The higher contrast scale better demonstrated the fracture edges, but still with a noise level that needed to be reduced. Doubling the mAs still did not bring the expected result, and the best result was achieved with 100 mAs, that is, doubling the initial value plus 30%. This protocol providing adequate radiographic density, increased contrast between anatomical structures, and improved visualization of the ribs. Conversely, indiscriminate increases in exposure factors resulted in detector saturation and deterioration of image quality, demonstrating that optimization must be individualized and carefully controlled. Fig. 5 compares the rib radiographs obtained using the non-optimized technique (a) and the optimized technique (b). Image (b) demonstrates improved contrast, reduced noise, and better fracture visualization.

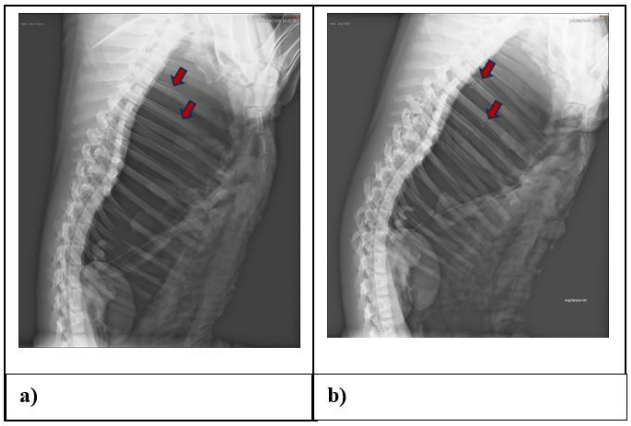


Fig. 5- Rib Radiograph: a) Non-Optimized Technique and b) Optimized Technique.

Radiographically, the lesion was characterized by posterior cortical discontinuity. In chronic stages, prominent periosteal callus formation may be observed. This morphological pattern is recognized as a radiological marker of severe thoracic compression and is considered a finding of high forensic specificity.

A similar behavior was observed in the evaluation of posteroanterior cranial lesions. The initial technique using 65 kV and 10 mAs produced images with reduced definition of osseous details. Doubling the mAs to 20 while maintaining the same kV value reduced quantum noise and improved the signal-to-noise ratio, resulting in enhanced sharpness and improved visualization of fracture lines. Fig. 6 compares the posteroanterior skull radiographs obtained using the non-optimized technique (a) and the optimized technique (b). Image (b) shows reduced image noise and improved visualization of the fracture line.

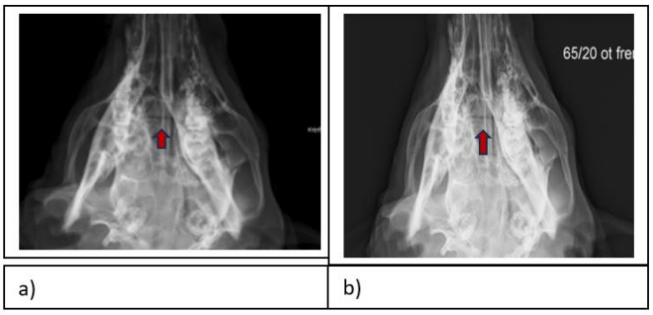


Fig. 6 - Cranial Radiograph: a) Non-Optimized Technique and b) Optimized Technique.

For lateral skull radiographs, the technique using 65 kV and 10 mAs provided the best diagnostic image quality among the parameters evaluated. The resulting image demonstrated adequate radiographic density, good contrast, and excellent sharpness of osseous structures, allowing better delineation of the cranial cortices and more accurate visualization of the simulated fracture pattern. This combination of parameters provided a satisfactory signal-to-noise ratio, facilitating the identification of anatomical details and contributing to greater diagnostic reliability. In contrast, the use of 55 kV and 5 mAs compromised image quality, characterized by increased image noise, a granular appearance, and reduced sharpness of the evaluated structures. The lower number of photons reaching the detector resulted in a decreased signal-to-noise ratio, impairing cortical delineation and adequate characterization of the lesion. This finding has significant clinical relevance, as cranial fractures may present subtle fracture lines whose identification depends directly on the technical quality of the acquired image. Fig. 7 compares the lateral skull radiographs obtained using the the optimized technique (a) and non-optimized technique (b). The optimized image provides greater sharpness and improved cortical definition, facilitating fracture identification.

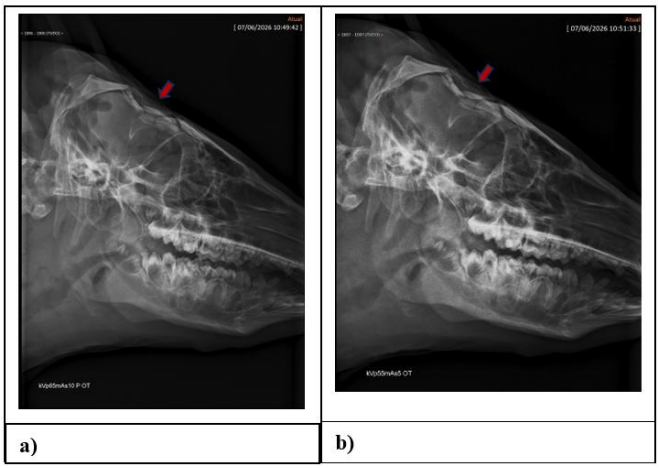


Fig. 7 - Cranial Fracture in the Lateral Projection a) Optimized Technique and b) Non-Optimized Technique.

In the analysis of appendicular lesions, the initial technique using 40 kV and 4 mAs proved insufficient for adequate demonstration of anatomical details. After parameter optimization, using 44 kV and 12.5 mAs in the posteroanterior projection, a significant improvement in contrast and cortical definition was observed. For the lateral projection, the use of 44 kV and 8 mAs allowed adequate visualization of the fracture line. Subsequently, the technique employing 44 kV and 20 mAs yielded the best diagnostic image quality, with reduced image noise, increased sharpness, and improved visualization of osseous structures.

The findings of this study demonstrated that minor modifications in technical parameters directly influence the diagnostic quality of radiographic images. From a physical standpoint, excessive reduction of mAs is associated with increased quantum noise due to the lower number of photons reaching the detector, compromising the signal-to-noise ratio and hindering the visualization of fine details. Conversely, indiscriminate increases in kV may reduce radiographic contrast, impairing differentiation between structures of similar densities. Therefore, achieving an appropriate balance between kV and mAs proved essential for obtaining images with adequate sharpness, density, and contrast.

The ribs region was the most complex in terms of technique adjustment, probably because the pig was not whole, as the internal organs had been removed and the internal air layer compromised the definition of the ideal technique without so many adjustments.

Another relevant aspect identified in this study was the importance of correct positioning during image acquisition. Even when appropriate technical parameters were employed,

small variations in positioning interfered with fracture visualization. This finding has important implications in pediatric practice, where pain, functional limitation, and lack of patient cooperation frequently hinder optimal examination performance. Consequently, the technical expertise of the radiologic technologist becomes a determining factor in obtaining reliable diagnostic examinations.

The findings reinforce the need for the development of specific exposure protocols for the radiographic investigation of suspected child abuse. Standardized yet flexible protocols allow individualized adjustments according to patient age, body habitus, anatomical thickness, and the body region being examined, facilitating the acquisition of high-quality images without unnecessary exposure to ionizing radiation. In this context, optimization of radiographic parameters should be aligned with radiation protection principles, particularly the ALARA principle, always seeking a balance between diagnostic quality and dose minimization.

The findings of the present study are consistent with those reported in the international literature regarding radiographic image repeat analysis and radiation protection. Hasaneen *et al.* (2023), when evaluating the causes of image repetition in digital radiography, reported that 48.4% of repeated examinations were associated with positioning errors, whereas 9.7% were related to the inappropriate selection of exposure factors. The authors emphasized that repeat examinations constitute an important source of unnecessary exposure to ionizing radiation and highlighted the need for continuous quality assurance programs and technical optimization strategies to reduce such occurrences. Furthermore, the study demonstrated that the ALARA principle should be incorporated into radiology department routines, aiming to minimize radiation dose while maintaining adequate diagnostic image quality [25].

Similar findings were reported by Bantas, Sweeney & Mdletshe (2023), who analyzed more than 76,000 images acquired using direct digital radiography systems. The authors identified image rejection rates ranging from 5.91% to 7.86%, with positioning errors accounting for approximately 77% to 79% of repeated examinations. The study highlighted that systematic analysis of rejected images represents an important quality assurance tool, enabling the identification of acquisition protocol deficiencies, the review of exposure charts, and the development of educational strategies aimed at reducing image repetition and, consequently, patient radiation exposure. The authors further emphasized that optimization of radiographic protocols should consider the specific characteristics of each examination and imaging department, as no single universal exposure chart can be applied to all clinical situations [26].

Complementarily, O'Keefe (2023) reported that image rejection rates in digital radiography systems remain close to 9%, despite technological advances associated with direct digital radiography. The author demonstrated that the primary causes of image rejection continue to be inadequate positioning, anatomical cut-off, and inappropriate exposure parameters. The study also emphasized that each repeated examination results in additional radiation exposure to the patient while negatively affecting departmental workflow efficiency. According to the author, rejected image analysis should be incorporated as a continuous quality improvement strategy, contributing to professional development and reducing unnecessary radiation exposures [27].

This discussion becomes even more relevant when applied to the pediatric population. Children exhibit increased radiosensitivity due to the higher mitotic activity of developing tissues and possess a longer life expectancy, factors that increase the probability of stochastic effects associated with ionizing radiation manifesting throughout life. Therefore, reducing repeat examinations through the implementation of appropriately designed exposure charts becomes an essential radiation protection strategy. In this context, the findings of the present study reinforce that the combination of proper patient positioning, appropriate selection of exposure parameters, and the use of examination-specific radiographic protocols represents one of the most effective approaches for ensuring diagnostic image quality, patient safety, and compliance with radiation protection principles [28-30].

A limitation of the present study is the absence of quantitative image quality metrics, dosimetric assessment, and formal interobserver agreement analysis. Although the proposed protocol demonstrated satisfactory qualitative performance, future studies should incorporate these objective parameters to further strengthen its methodological robustness and metrological validation.

Finally, the results of this study demonstrate that the accurate identification of bone injuries related to child abuse depends not only on recognizing classical radiographic patterns but also on the proper execution of radiographic technique. The combination of appropriate positioning, acquisition of orthogonal projections, and individualized optimization of protocols proved fundamental for maximizing the diagnostic sensitivity of radiographic examinations. And the 15% rule is good tool to orientate the radiology team to work and to know the response of the detector. Furthermore, the use of the porcine experimental model enabled the development of an educational product applicable to the training and continuing education of radiologic technologists, contributing to the strengthening of pediatric forensic radiology as an essential tool for the early detection of suspicious findings, the production of technical and scientific evidence, and the comprehensive protection of children who are victims of violence.

IV. Conclusions

Overall, this study demonstrated that the identification of radiographic findings associated with suspected child physical abuse depends on both the recognition of characteristic fracture patterns and the technical quality of image acquisition. The juvenile porcine model enabled the experimental reproduction and radiographic characterization of spiral, greenstick, rib, and cranial fractures, supporting its use as an educational and research tool in pediatric forensic radiology.

Optimized exposure charts, appropriate positioning, and complementary radiographic projections improved fracture visualization and diagnostic image quality. Small adjustments in kV and mAs influenced contrast, sharpness, image noise, and the visibility of fracture lines. These findings reinforce the importance of pediatric-specific exposure protocols and the ALARA principle, particularly

because children are more radiosensitive and have a longer lifetime risk associated with ionizing radiation exposure.

The 15% rule, using fixed and predefined kV values for anatomical regions, and making fine adjustments to mAs to improve lesion sharpness has a direct impact on image quality and patient dose as defined by ALARA principle.

This study has limitations that should be considered. The pig may not have been the ideal phantom for defining the rib protocol, but for the extremities it proved to be a valuable and inexpensive phantom for simulating fractures. In addition, the findings cannot be directly extrapolated to clinical pediatric populations without further validation studies. Nevertheless, the study establishes an experimental model that may support future research on image-quality optimization, training of radiologic technologists, and the development of pediatric forensic radiology protocols.

V. ACKNOWLEDGMENTS

The authors express their sincere gratitude to DAPI Clinic, Curitiba, Paraná, Brazil, for providing the facilities and radiographic equipment required for the experimental procedures conducted in this study. The authors also acknowledge the Graduate Program in Biomedical Engineering (PPGEB) for its academic support and valuable contribution to the development of this research.

REFERENCES

[1] E Brogdon, B. G. Forensic Radiology. Boca Raton: CRC Press LLC (1998South J, Blass B (2001) The future of modern genomics. Blackwell, London.

[2] Loaiza, G. A. M; Daza, A. F. O; Archilla, G. A. Aplicaciones de la radiología convencional en el campo de la medicina forense: rev. colomb. radiol; 24(4): 3805-3817, 2013.

[3] CADE PARANA Violências Contra Crianças e Adolescentes em Dados, 2024.

[4] Lopes, N. R. L; Eisenstein, E; Williams, L. C. A. Abusive head trauma in children: a literature review. J Pediatr (Rio J). 2013;89(5):426−33.

[5] Sousa, B. L. M., et al. Radiologia Forense na Área Criminal. Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano 02, Ed. 01, Vol. 13. pp. 455-462 Janeiro de 2016. ISSN:2448-0959.

[6] Kemp, A. M. et al. Patterns of skeletal fractures in child abuse: systematic review. BMJ, London, v. 337, p. 1518-1523, 2008.

[7] Teixeira, C. J. A. Contributos da Radiologia Forense na identificação de lesões típicas no Crime de Maus-tratos a Menores. Repositório Comum, Portugal, 2021.

[8] Rodrigues, M. B. Diagnóstico por imagem no trauma músculoesquelético – princípios gerais. Rev Med (São Paulo). 2011 out.-dez.;90(4):185-94.

[9] BRASIL. Estatuto da Criança e do Adolescente. Lei nº 8.069, de 13 de julho de 1990. Brasília: Ministério da Justiça, 2010.

[10] MINISTÉRIO DA SAÚDE. Violência intrafamiliar: orientações para a prática em serviço. Brasília: MS, 2002.

[11] ICRP – INTERNATIONAL COMMISSION ON RADIOLOGICAL PROTECTION. The 2007 Recommendations of the ICRP. Oxford: Elsevier, 2007.

[12] Don, Steven; *et al*. Image Gently Campaign — Back to Basics Initiative: Ten Steps to Help Manage Radiation Dose in Pediatric Digital Radiography. American Journal of Roentgenology, v. 200, may 2013.

[13] VOSIAK, P. et al. Dosimetric evaluation of pediatric abdominal and pelvic radiographs: optimization and diagnostic reference levels. Radiation Physics and Chemistry, v. 233, 112731, 2025a.

[14] VOSIAK, P. et al. Organ dose estimation in pediatric abdomen and pelvic radiographs using Caldose_X Monte Carlo simulations. Radiation Physics and Chemistry, 113094, 2025b.

[15] Chusin, T. *et al.*The influence of 10 kVp and 15% rule applications on patient dose and image quality in extremities radiography: a phantom study. Journal of Associated Medical Sciences, v. 57, n. 3, p. 166-176, 2024.

[16] Wenman, A.; Lockwood, P. Comparing the standard knee X-ray exposure factor, 10 kV rule, and modified 10 kV rule techniques in digital radiography to reduce patient radiation dose without loss of image quality. Radiography, v. 30, p. 574-581, 2024. DOI: https://doi.org/10.1016/j.radi.2024.01.013.

[17] Carrol, Quinn B. Practical Radiographic Imaging. 8. ed. Springfield: Charles C. Thomas Publisher, 2007.

[18] Carrol, Quinn B. Radiography in the Digital Age: Physics — Exposure — Radiation Biology. Springfield: Charles C. Thomas Publisher, 2011

[19] ANVISA. Resolução da Diretoria Colegiada - RDC nº 611, de 09/03/2022.

[20] Brasil. Lei Nº 11.794, de 8 de outubro de 2008.

[21] CHATGPT. Modelo GPT-4. OpenAI, 2026.

[22] Bonney, H. et al. Validity of the use of porcine bone in forensic cut mark studies. Journal of Forensic Sciences, Hoboken, v. 66, n. 2, p. 611-620, 2021.

[23] Pearce, A. I.; Richards, R. G.; Milz, S.; Schneider, E.; Pearce, S. G. Animal models for implant biomaterial research in bone: a review. European Cells and Materials, Davos, v. 13, p. 1-10, 2007.

[24] Sangild, P. T. et al. Developmental factors influencing bone strength in precocial mammals: the preterm infant pig model. Frontiers in Pediatrics, Lausanne, v. 11, 2023.

[25] Hasaneen, M. et al. Assessment of image rejection in digital radiography. *Journal of Medicine and Life*, v. 16, n. 5, p. 731–738, 2023.

[26] Bantas, G.; Sweeney, R.-J.; Mdletshe, S. Digital radiography reject analysis: A comparison between two radiology departments in New Zealand. *Journal of Medical Radiation Sciences*, v. 70, p. 137–144, 2023.

[27] O'Keefe, K. *An Exploration of Radiographer Decision-Making Regarding Rejected or Sub-Optimal Plain X-Ray Images*. 2023. Dissertação (Master of Philosophy) – Queensland University of Technology, Brisbane, 2023.

[28] Bergonie, J.; Tribondeau, L. Interpretation of Some Results of Radiotherapy and an Attempt at Determining a Logical Technique of Treatment: De Quel ques Resultats de la Radiotherapie et Essai de Fixation d'une Technique Rationnelle. Radiation Research, p. 587-588. 25 mar. 1959.

[29] Pinto, B.V.C.C Estimativa da Dose e Risco em Pacientes Pediátricos Devido a Exames Radiográficos dos Seios da Face. Repositório da UFPE, 2015.

[30] Neves, L; Canevaro, L. V; Dovales, A. C. M. Reações Teciduais (Efeitos Determinísticos) em Radiologia Diagnóstica e Intervencionista: Uma Revisão Integrativa da Literatura. Braz. J. Radiat. Sci., Rio de Janeiro 2025.

Author: Fernanda Lorena Slusarczuk Manchenho
Institute: Universidade Tecnológica do Paraná - UTFPR
Street: Av. Sete de Setembro, 3165 - Rebouças, Curitiba – PR.
City/ Country: Curitiba / Brazil
Email: fernandamanchenho@alunos.utfpr.edu.br

Experimental Study of EMG Signals for Prosthetic Control using a Feature-Based Pipeline and Unsupervised Learning

1st Vinícius F.P.M. de Azevedo
Department of Electrical Engineering
Federal University of Juiz de Fora
viniciusfpmazevedo@gmail.com

2ndLeandro Rodrigues Manso Silva
Department of Electrical Engineering
Federal University of Juiz de Fora
leandro.manso@ufjf.br

3rd Victor Valadares de Morais
Department of Electrical Engineering
Federal University of Juiz de Fora
victor.morais@ufjf.br

***Abstract*—This paper presents a feature-based sEMG signal processing pipeline for prosthetic control applications, combining time- and frequency-domain descriptors with an unsupervised clustering stage for automated label generation. The generated pseudo-labels train a supervised classifier to distinguish muscle contraction and relaxation states across four forearm electrode configurations. The results demonstrate high classification performance and accurate pattern detection, supporting the feasibility of the proposed unsupervised labeling strategy under the evaluated experimental conditions.**

***Index Terms*—Electromyography (EMG), sEMG, Prosthetic Control, Feature Extraction, Unsupervised Learning, Machine Learning, Signal Processing, Pattern Recognition**

I. Introduction

Surface electromyography (sEMG) is a non-invasive technique widely used to record the electrical activity associated with muscle contraction. By means of electrodes placed on the skin, this technique enables the analysis of muscle activation patterns and the extraction of features related to activation intensity, timing, and coordination among different muscles. Consequently, sEMG has been applied in several areas, including motor rehabilitation, clinical assessment, human–machine interfaces, and, particularly, myoelectric prosthetic control systems. In these systems, muscle signals are processed to identify the user's intended movement and convert it into commands capable of controlling prosthetic devices [1].

Despite the large number of studies and advances in this field, many solutions developed for the acquisition and processing of myoelectric signals rely on expensive equipment and devices. These characteristics may hinder the reproducibility of results, the adaptation of systems to different applications, and access to myoelectric control technologies by researchers, students, and low-income users. In this context, the development of low-cost and open-source systems is relevant, as it promotes knowledge dissemination, enables hardware and software modifications according to the requirements of each application, and contributes to expanding access to sEMG-based assistive technologies

From a physiological perspective, sEMG signals result from the spatiotemporal superposition of motor unit action potentials (MUAPs) recruited during muscle contraction [1], [2]. In anatomical regions with high muscle density, such as the forearm, surface electrodes capture activity from adjacent muscles (cross-talk), representing a primary limitation for signal analysis [3], [6]. Therefore, proper electrode placement is essential to maximize the signal-to-noise ratio and ensure high-quality sEMG measurements.

In addition to physiological challenges, Electromyography signals are characterized by low amplitude and high susceptibility to noise, even when differential acquisition techniques are employed. These characteristics make signal processing and reliable classification particularly challenging. Traditional approaches often rely on threshold-based methods for label generation, [8]; however, such methods may be sensitive to noise and signal variability [7], [9].

To address these limitations, this work proposes a feature-based pipeline for sEMG signal processing, combining time- and frequency-domain features with an unsupervised learning stage for automatic pattern identification and label generation. The generated labels are then used to train a supervised classification model. This approach aims to reduce dependency on manually labeled datasets while providing an alternative strategy for handling signal variability, contributing to the development of accessible approaches for sEMG-based prosthetic control.

Therefore, the main contribution of this work is a semi-automatic labeling strategy for sEMG datasets based on feature engineering and unsupervised clustering, reducing the dependency on manually annotated data while preserving the ability to identify muscle contraction patterns for prosthetic control applications. Although a quantitative comparison with conventional threshold-based labeling methods was considered, such a comparison was beyond the scope of the present study, which focuses on investigating the feasibility of the proposed feature-based unsupervised approach and evaluating its use for pseudo-label generation and subsequent supervised classification. In this context, the present study is framed as a proof-of-concept study, focusing on the feasibility of the proposed labeling pipeline rather than on establishing its superiority over threshold-based methods or validating a complete prosthetic control system.

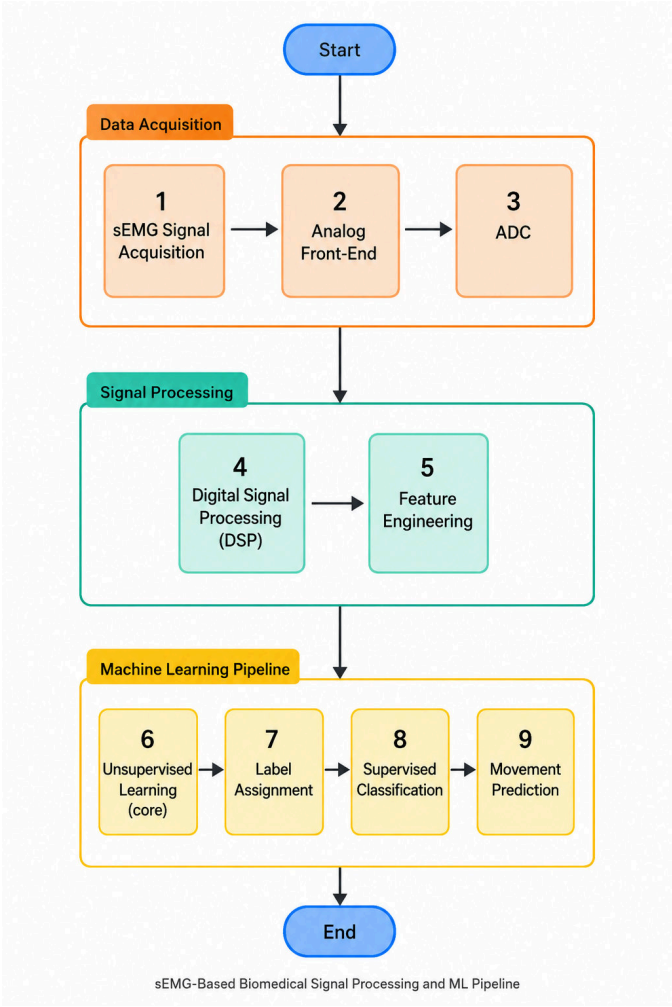


Fig. 1. Proposed EMG processing pipeline.

The remainder of this paper is organized as follows: Section II details the architecture of the proposed feature-based pipeline for sEMG signal processing. Section III presents the algorithm performance analysis. Finally, Section IV summarizes the conclusions of the study.

II. Materials and Methods

A. Proposed classification system

Fig. 1 presents the proposed sEMG processing and classification pipeline. The proposed method is composed of nine sequential stages, ranging from signal acquisition to movement prediction. Each stage is described in the following subsections.

1) sEMG Signal Acquisition: Differential surface electromyography (sEMG) acquisition was adopted in order to reduce common-mode noise and improve signal quality during hand movement acquisition [10]. Multiple electrode configurations were experimentally evaluated to identify muscle activation patterns associated with different hand and finger movements.

Since hand and finger movements result from the coordinated activation of multiple forearm muscles, a single sEMG acquisition channel is generally insufficient to represent all possible hand gestures [4]. In the present study, only one sEMG acquisition module was available; therefore, the four muscle configurations were acquired independently rather than simultaneously. Each recording corresponded to a single muscle configuration, and the resulting activation states were subsequently analyzed and combined at the processing stage. Therefore, this work investigates individual electrode configurations as independent muscle activation detectors and subsequently proposes a strategy to combine their outputs. Moreover, practical myoelectric prostheses do not require the recognition of every anatomically possible hand movement, but rather a reduced set of functional gestures that are sufficient for the execution of activities of daily living (ADLs) [5]. Based on this premise, the acquisition protocol was designed to generate independent datasets for each muscle configuration while preserving the possibility of future multi-channel integration.

The target movements considered in this study were selected according to their relevance for prosthetic control and daily-life activities, including hand opening, hand closing, pointing, pinch grasp, and thumb extension. Rather than directly acquiring these movements as classification labels, the proposed methodology focused on identifying whether each monitored muscle was in a contraction or relaxation state.

To build the database, an acquisition protocol composed of different activation patterns was defined for every electrode configuration. Each recording included: (i) relaxed muscle activity, (ii) sustained muscle contraction, (iii) cyclic contraction-relaxation movements with an approximate period of 10 s, (iv) cyclic contraction-relaxation movements with an approximate period of 5 s, and (v) individual finger activation experiments designed to evaluate cross-talk effects and muscle selectivity.

For the individual finger activation experiments, a discrete movement encoding scheme was adopted, as summarized in Table I. The encoded sequence defines the temporal movement protocol executed during data acquisition and enables a direct comparison between the expected muscle activation pattern and the corresponding EMG responses.

The adopted encoding is defined as follows: 1—Little finger contraction; 2—Ring finger contraction; 3—Middle finger contraction; 4—Index finger contraction; 5—Thumb contraction; 6—All fingers extended; and 7—All fingers flexed.

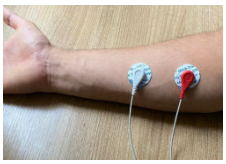
To ensure repeatability, approximately five recordings were acquired for each electrode configuration and acquisition protocol. Each recording lasted between 1 and 2 min, resulting in independent datasets representing different muscle activation conditions.

Consequently, the complete database consisted of multiple independent datasets for each muscle, covering all predefined activation protocols. These datasets constituted the basis for the subsequent feature extraction, unsupervised clustering, and supervised classification stages.

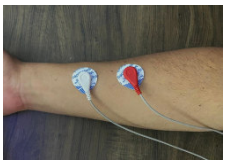
The monitored muscles were selected according to their anatomical relevance to hand and finger movements, consid-

TABLE I
ENCODED TEMPORAL PROTOCOL FOR INDIVIDUAL FINGER ACTIVATION ACROSS ELECTRODE CONFIGURATIONS (C1: FCR, C2: FPL, C3: FDS, C4: ED).

Time (s)	C1	C2	C3	C4	Time (s)	C1	C2	C3	C4
0 – 5	6	6	6	6	50 – 55	6	6	6	6
5 – 10	1	1	1	1	55 – 60	4	5	5	5
10 – 15	6	6	6	6	60 – 65	6	6	6	6
15 – 20	2	2	2	2	65 – 70	3	4	4	4
20 – 25	6	6	6	6	70 – 75	6	6	6	6
25 – 30	3	3	3	3	75 – 80	2	3	3	3
30 – 35	6	6	6	6	80 – 85	6	6	6	6
35 – 40	4	4	4	4	85 – 90	1	2	2	2
40 – 45	6	6	6	6	90 – 95	-	6	6	6
45 – 50	5	5	5	5	95 – 100	-	1	1	1



(a) FCR Configuration



(b) FPL Configuration



(c) FDS Configuration



(d) ED Configuration

Fig. 2. Electrode placement adopted for the four evaluated muscle configurations: (a) FCR, (b) FPL, (c) FDS, and (d) ED.

ering the anatomical descriptions presented by Waldman [11] and Netter [12]. The electrode placement was experimentally optimized by positioning the differential pair over the muscle belly in order to maximize signal quality and reduce interference from adjacent muscles. A reference electrode was positioned near the elbow region. The selected muscles included the Flexor Carpi Radialis (FCR), Flexor Pollicis Longus (FPL), Flexor Digitorum Superficialis (FDS), and Extensor Digitorum (ED), as illustrated in Fig 2.

All datasets were acquired using non-invasive surface electromyography (sEMG) from a single healthy volunteer (one of the authors, aged 25, with no history of neuromuscular disorders) under the same experimental conditions, in accordance with the Declaration of Helsinki and with written informed consent. The objective of this work was to validate the proposed signal processing and machine learning methodology rather than to evaluate inter-subject variability. Therefore, the results should be interpreted as an intra-subject proof of concept, and no conclusions regarding the generalization or robustness of the proposed approach across different individuals can be drawn from the present dataset.

2) Analog Front-End: A commercial sEMG acquisition module was employed for the acquisition and analog conditioning of muscle signals. The module uses a differential input architecture consisting of two measurement electrodes positioned over the target muscle and a reference electrode. This configuration attenuates noise components that are simultaneously present at both measurement electrodes, thereby improving signal quality through a high Common-Mode Rejection Ratio (CMRR).

The commercial module was selected for its simplicity, high CMRR, and suitability for rapid experimental prototyping. Although adequate for single-channel evaluation, its use presents limitations for multi-channel expansion due to lack of direct control over internal conditioning stages and synchronization. For future multi-channel applications, a dedicated integrated acquisition system would be more appropriate.

3) Analog-to-Digital Conversion (ADC): The analog-to-digital conversion stage was implemented using the Arduino Uno board, which is based on the ATmega328P microcontroller. Its built-in successive-approximation-register (SAR) ADC provides a 10-bit resolution. Furthermore, the device was also used for serial data transmission, operating at a baud rate of 230400 bps.

According to [13], most of the useful sEMG signal energy is concentrated within the 0–500 Hz frequency range, with the dominant spectral content typically located between 50 and 150 Hz. Based on the Nyquist criterion, the sampling frequency of 2 kHz was initially adopted for signal acquisition.

But, due to limitations associated with serial communication and software-based timing, sampling jitter was observed during acquisition. Therefore, timestamps were recorded alongside ADC values, and the effective sampling frequency was estimated as $f_{s,\text{eff}} = 1/[\frac{1}{N-1}\sum_{n=2}^{N}(t_n - t_{n-1})]$, where N is the total sample count and t_n is the n-th timestamp. The effective sampling frequency was approximately 1.5 kHz across all datasets, satisfying the Nyquist criterion.

The acquired signals were stored in CSV format for subsequent digital signal processing and feature extraction stages.

4) Digital Signal Processing (DSP): After signal acquisition and digitization, the mean component was removed to eliminate the influence of the direct-current (DC) offset. Subsequently, an 8th-order Butterworth IIR band-pass filter with cutoff frequencies of 20 Hz and 500 Hz was applied to remove motion artifacts while retaining only the spectral components of interest in the signal. Additionally, during the validation stages, a zero-phase filter was used to compare the temporal behavior of the filtered signals and evaluate possible phase-related distortions.

The IIR filter was selected because of its lower computational complexity and reduced number of coefficients compared with an FIR filter, which are desirable characteristics for real-time embedded applications. Although IIR filters introduce nonlinear phase distortion, the classification model was trained directly using a dataset composed of filtered signals, allowing the model to naturally adapt to the temporal characteristics introduced by the filter.

Figure 3 illustrates the spectrogram obtained from a representative dataset corresponding to the fourth electrode configuration during cyclic hand opening and closing movements

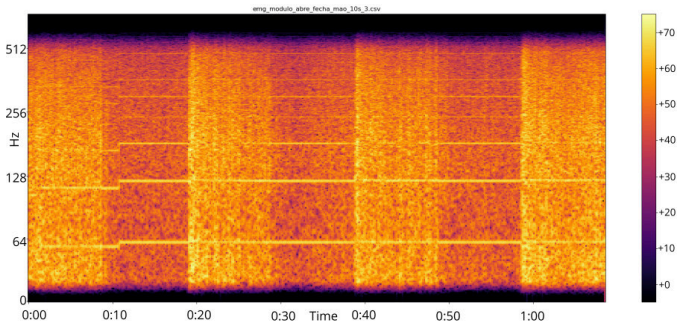


Fig. 3. Spectrogram obtained from the processed EMG signal acquired using the fourth electrode configuration during cyclic hand opening and closing movements with an approximate period of 10 s. The horizontal axis represents elapsed time (min:s), while the vertical axis represents frequency in hertz (Hz). The color scale indicates signal power in decibels (dB), where brighter colors correspond to higher energy concentrations and darker colors correspond to lower energy levels.

with an approximate period of 10 s. The spectrogram was computed after the application of the proposed DSP stages.

It can be observed that the power-line interference components are centered mainly around 60 Hz, 120 Hz, and 180 Hz. Although notch filters are commonly used to remove these components, their use was intentionally avoided in this work, since the implementation of multiple band-stop filters would increase computational complexity and introduce additional phase distortions, which are undesirable characteristics for embedded real-time applications.

At the same time, periods of muscle contraction are characterized by an overall increase in signal energy across the sEMG frequency range, making activation events clearly distinguishable in the time–frequency representation. During these periods, the useful sEMG signal exhibits considerably higher energy than the approximately constant power-line interference, allowing relevant physiological information to be reliably extracted even without explicitly removing these components. Since the interference remains nearly constant throughout the acquisition, the proposed learning pipeline is able to naturally adapt to its presence. This behavior supports the decision to preserve these components and allow the subsequent learning stages to operate under realistic acquisition conditions.

However, the signal energy during relaxation states was significantly lower than the power-line interference energy. To address this behavior, specific spectral features were developed to quantify these variations and assist the unsupervised clustering process. Therefore, the proposed system was designed to operate under realistic acquisition conditions, allowing the learning models to naturally adapt to the residual interference components present in the acquired signals.

5) Feature Engineering: In order to extract representative features from the sEMG signals, a framing process was employed. Frame size was set to 1024 samples (approx. 0.68 s) with a hop length of 256 samples (approx. 0.17 s), yielding an average spectral resolution of 1.46 Hz per frame under a Hann window [14].

Features were extracted using Python (*librosa* [14] and custom routines). The extracted time-domain features included: Amplitude Envelope, Root Mean Square (RMS), Zero Crossing Rate (ZCR), Mean Absolute Value (MAV), Waveform Length (WL), mean value, variance, standard deviation, Slope Sign Changes (SSC), and MAV with Slope Weighting. Frequency-domain features included: Band Energy Ratio (BER), Spectral Centroid, Bandwidth, fundamental frequency dominance (f_0 dominance), Median Frequency, Mean Frequency, harmonic energy relations, spectral energy distribution metrics, and interference-sensitive spectral descriptors.

The extracted features were organized into normalized dataframes for subsequent unsupervised clustering and supervised classification.

6) Unsupervised Learning: Based on the extracted features, unsupervised learning techniques were employed to identify muscle activation patterns and generate pseudo-labels corresponding to contraction and relaxation states. The adopted strategy consisted of constructing binary models for each electrode configuration individually, considering muscle contraction as state 1 and muscle relaxation as state 0.

This binary approach was adopted due to the acquisition limitations of the experimental setup, which employed only a single sEMG acquisition module. Consequently, simultaneous multi-channel acquisition using multiple electrode configurations was not possible. Nevertheless, the proposed methodology was designed to be scalable and directly applicable to future multi-channel systems.

Two clustering algorithms were evaluated for label generation: K-Means Clustering and Gaussian Mixture Models (GMM). These methods were applied mainly to datasets containing temporal muscle activation variations, while datasets containing only a single muscular state received manually assigned labels. Before clustering, all feature matrices were normalized using the Min-Max scaling technique. Since unsupervised learning algorithms are sensitive to the statistical properties of the input data, the clustering strategy, including algorithm selection and feature combinations, was optimized individually for each dataset.

7) Label Assignment: The optimization strategy involved feature selection, temporal filtering of extracted features, and post-processing of the generated cluster labels. Temporal smoothing techniques, including moving average filtering, median filtering, and run-length smoothing, were employed to reduce spurious classifications and improve temporal consistency.

Different combinations of time-domain and frequency-domain features were evaluated according to the analyzed movement and muscle activation pattern. In specific cases, minor manual interventions were performed after automatic processing stages in order to correct label inversion or isolated temporal inconsistencies.

After the clustering optimization process, the generated labels were assigned to the original feature dataset. The final labeled datasets preserved the extracted features in their

original scale, while the normalized representations were used exclusively during the clustering stage. These labeled datasets were subsequently employed for supervised classifier training.

8) Supervised Classification: After the label assignment stage, the resulting datasets were employed for supervised learning. Each electrode configuration generated an independent binary classification dataset, in which the target variable represented muscle contraction and relaxation states.

To handle class imbalance, the RandomOverSampler technique was applied to the training set. Feature values were normalized using Min-Max scaling, with scaling parameters computed exclusively on training data and saved to ensure consistent evaluation during inference. Multiple supervised learning algorithms were evaluated using the extracted feature vectors and the unsupervised pseudo-labels to identify the most suitable classifier for each muscle configuration.

The datasets were divided into training (20%) and testing (80%). Unlike conventional random splits, the adopted datasets were formed by concatenating multiple independent acquisition sessions performed at different time instants under the same protocol. Therefore, a reduced training subset was intentionally adopted, while the majority of the data were reserved for testing to evaluate the model's performance on recordings not included in the training subset and to reduce the likelihood of overfitting. Hyperparameter optimization was also performed when applicable.

9) Movement Prediction: The binary classifiers generated for each muscle configuration were designed to operate as independent activation detectors. Each model estimates whether the corresponding muscle is in a contraction or relaxation state.

The final movement prediction strategy is based on the combination of these individual muscle activations. By analyzing the activation patterns obtained from multiple muscles, it becomes possible to infer more complex hand and finger movements than those represented by a single binary classifier.

For each time frame, the outputs of the individual classifiers are combined into a binary activation vector

$$\text{Label}_t = [b_1\ b_2\ b_3\ b_4], \qquad b_i \in \{0, 1\}$$

where C1, C2, C3, and C4 correspond to the FCR, FPL, FDS, and ED muscles, respectively.

The binary activation vector provides a compact representation of the muscular activation pattern, allowing functional hand movements to be inferred from the simultaneous activation of multiple muscles. Table II summarizes the movement mapping adopted in this work.

Although the experimental validation of this work was performed using independent datasets acquired from individual muscle configurations, the proposed methodology was designed to support future multi-channel implementations, where several sEMG acquisition modules operate simultaneously. In this scenario, the same feature extraction, clustering, label assignment, and classification pipeline can be directly extended to real-time multi-channel movement recognition systems.

This architecture enables the construction of scalable movement recognition systems, allowing the transition from binary muscle-state detection to multi-class gesture and movement prediction.

TABLE II
BINARY VECTOR MAPPING FOR TARGET HAND MOVEMENTS. THE TABLE REPRESENTS A CONCEPTUAL SIMULATION BASED ON THE INDEPENDENT BINARY ACTIVATION STATES OBTAINED FROM EACH MUSCLE CONFIGURATION.

C1 (FCR)	C2 (FPL)	C3 (FDS)	C4 (ED)	Predicted Movement
0	0	0	1	Open Hand
1	1	1	0	Closed Hand
0	1	1	1	Pinch
1	0	1	0	Thumb Extension

B. Evaluation Methodology

The performance of the proposed methodology was evaluated in two stages.

For the unsupervised learning stage, the quality of the generated labels was primarily assessed through visual inspection of the recorded sEMG signals, clustering results, and temporal consistency of the contraction and relaxation patterns. A fully objective quantitative evaluation is difficult in the present experimental setting because no independent physiological ground truth was available. In particular, the time intervals defined by the experimental protocol cannot be directly considered ground-truth labels, since they represent the intended task periods rather than the exact physiological onset and offset of muscle activation. In recordings from a human subject, these transitions may occur before or after the predefined intervals and may also develop gradually, making the assignment of an exact transition point inherently uncertain.

For this reason, visual inspection of the recorded signals was adopted as the primary means of assessing whether the identified clusters were consistent with the observed contraction and relaxation patterns. This assessment considers the temporal behavior of the signal and the resulting cluster assignments, rather than relying on a single signal characteristic or a predefined amplitude threshold to determine the expected state.

As a complementary quantitative assessment, a representative subset of the acquired signals was manually annotated through visual inspection and compared with the corresponding clustering-based labels. The experimental protocol was used as a temporal reference during this annotation, while the recorded sEMG signal was considered when determining the observed contraction and relaxation intervals. The annotations were aligned with the same analysis frames used by the clustering stage, allowing a frame-by-frame comparison.

The manual annotations were therefore treated as an independent reference rather than an absolute physiological ground truth, since the precise boundaries of muscle activation remain subject to interpretation. The agreement between the reference and generated labels was quantified using Accuracy, Precision, Recall, F1-score, Cohen's Kappa, and the confusion matrix. Thus, the quantitative analysis provides an objective measure of agreement for the evaluated subset, while the visual analysis remains the primary evaluation criterion for the overall clustering results.

Subsequently, the generated labels were employed to train supervised binary classifiers for each electrode configuration. The classification performance was evaluated using the Accuracy, Precision, Recall, and F1-score metrics, which were computed according to their standard definitions [15].

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. Precision (P) measures the proportion of predicted positive samples that are correctly classified, Recall (R) measures the proportion of actual positive samples correctly identified, Accuracy (Acc) represents the overall classification performance, and the F1-score ($F1$) corresponds to the harmonic mean of Precision and Recall.

Finally, the binary outputs of the four classifiers were combined according to the movement prediction strategy described previously, allowing the feasibility of recognizing functional hand gestures from independent muscle activation detectors to be analyzed.

III. Results and Discussion

A. Unsupervised Label Generation

The clustering strategy was optimized individually for each dataset due to the inherent variability of sEMG signals and the differences between global hand movements and individual finger activations. Depending on the dataset characteristics, different combinations of time- and frequency-domain features were selected, and label post-processing techniques were applied to improve cluster stability and temporal consistency.

As illustrated in Fig. 4, the generated cluster labels (red dashed line) are superimposed on the zero-phase filtered sEMG signal (blue line), allowing a direct visual comparison between the automatically generated pseudo-labels and the underlying muscle activation periods.

The proposed approach successfully identified contraction and relaxation states across all electrode configurations. For datasets containing periodic contraction-relaxation cycles, the separation between both states was generally well defined, with most ambiguities occurring only during the transition intervals. Occasional short-duration fluctuations within otherwise stable activation periods were effectively mitigated by the proposed label post-processing stage, resulting in temporally consistent cluster assignments.

The visual inspection of the generated labels indicated a consistent correspondence between the identified clusters and the activation patterns observed in the filtered EMG signals. This assessment was considered particularly relevant because the experimental protocol defines the intended movement intervals but does not provide an exact physiological ground truth for the onset and offset of muscle activation. Therefore, the observed signal behavior and the temporal consistency of the generated labels were considered when assessing the clustering results.

The most challenging scenarios were observed in the individual finger movement datasets. In these cases, the physiological cross-talk phenomenon produced overlapping muscle activations, reducing the separability between classes. Consequently, feature selection and label post-processing became more important to ensure temporal consistency and reduce isolated false-positive detections.

For flexor muscles, contraction corresponds to the flexion of the target finger while the remaining fingers are extended. For extensor muscles, contraction corresponds to the extension of the target finger while the remaining fingers are flexed, preserving the same encoding.

To complement the qualitative assessment, a quantitative analysis was performed on a representative subset of 3280 analysis frames using the manual reference annotation described in Section II-B. The clustering-based labels achieved an Accuracy of 98.05%, a Cohen's Kappa of 0.961, and an F1-score of 0.98 for both relaxation and contraction. Precision/Recall values were 0.99/0.97 for relaxation and 0.97/0.99 for contraction, respectively. The resulting confusion matrix yielded 1,597 True Relaxations, 51 False Contractions, 13 False Relaxations, and 1,619 True Contractions (rows representing manual reference, columns representing clustering labels).

The quantitative results complement the visual assessment by providing an objective measure of agreement for the selected recordings. However, they should not be interpreted as evidence of an absolute physiological ground truth or as a comprehensive validation of all acquired datasets, since the manual reference itself is subject to interpretation of the recorded sEMG signals.

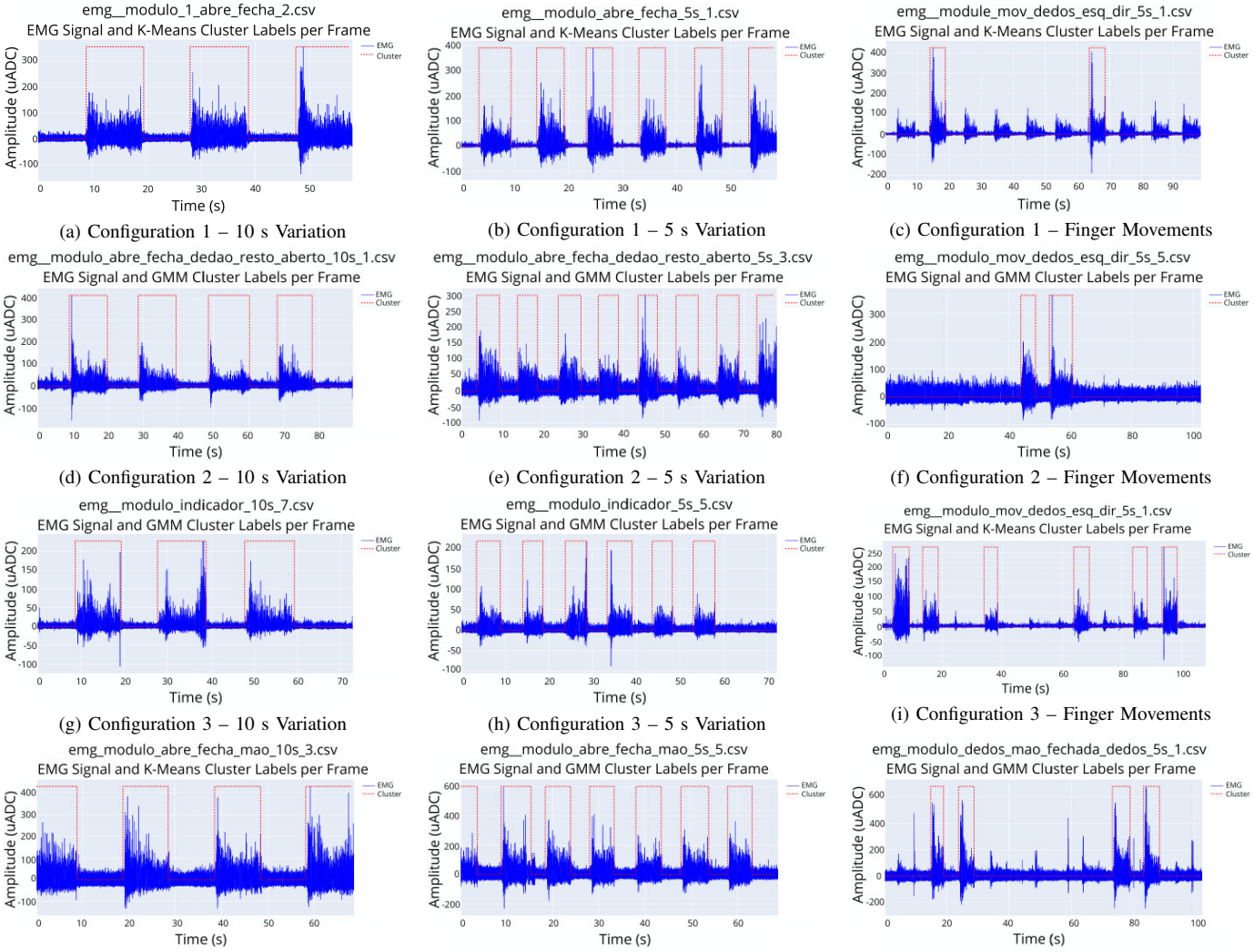
Furthermore, differences in clustering performance were observed among the electrode configurations. These variations are consistent with the characteristics of the monitored muscles, as larger and more superficial muscle groups generally produced clearer activation patterns and greater cluster separability. In contrast, deeper muscles or regions with significant overlap between muscle groups were more susceptible to cross-talk effects, increasing the complexity of the clustering task.

Overall, the visual and complementary quantitative assessments indicate that the clustering framework was able to identify activation patterns consistent with the observed sEMG signals under the evaluated experimental conditions, supporting the use of the generated labels as pseudo-labels for the subsequent supervised learning stage.

B. Supervised Classification Results

The pseudo-labels generated during the unsupervised learning stage were subsequently used to train supervised classification models. Several classifiers were evaluated and exhibited similar performance. For conciseness, only the results obtained using the Random Forest classifier are presented, as it consistently achieved high performance across all electrode configurations.

Table III summarizes the weighted-average classification metrics for each electrode configuration. Since the classification task consisted of a binary problem, representing con-



(a) Configuration 1 – 10 s Variation
(b) Configuration 1 – 5 s Variation
(c) Configuration 1 – Finger Movements
(d) Configuration 2 – 10 s Variation
(e) Configuration 2 – 5 s Variation
(f) Configuration 2 – Finger Movements
(g) Configuration 3 – 10 s Variation
(h) Configuration 3 – 5 s Variation
(i) Configuration 3 – Finger Movements
(j) Configuration 4 – 10 s Variation
(k) Configuration 4 – 5 s Variation
(l) Configuration 4 – Finger Movements

Fig. 4. Comparison of the clustering results obtained for the four electrode configurations and three movement categories. The horizontal axis represents time in seconds, and the vertical axis represents signal amplitude in ADC units (uADC). Cluster labels (red dashed line) are superimposed on the zero-phase filtered EMG signal (blue line), allowing visual assessment of the correspondence between the generated pseudo-labels and the underlying muscle activation periods.

TABLE III
CLASSIFICATION PERFORMANCE (WEIGHTED AVERAGE) FOR EACH ELECTRODE CONFIGURATION USING THE RANDOM FOREST CLASSIFIER

Configuration	Precision	Recall	F1-score	Accuracy
C1 (FCR)	0.96	0.96	0.96	0.96
C2 (FPL)	0.97	0.97	0.97	0.97
C3 (FDS)	0.99	0.99	0.99	0.99
C4 (ED)	0.98	0.98	0.98	0.98

traction and relaxation states, accuracy, precision, recall, and F1-score remained consistently high across all experiments.

The results indicate that the labels generated by the clustering stage provided sufficient consistency for the subsequent supervised classification stage, with classification accuracies ranging from 96% to 99% across the evaluated configurations. These results support the feasibility of using the generated pseudo-labels for supervised learning under the evaluated experimental conditions. The adopted train/test partition allowed the classifiers to be evaluated on recordings not included in the training subset; however, this should not be interpreted as evidence of generalization across subjects or experimental conditions.

C. Movement Prediction

The proposed binary architecture combines the outputs of four independent muscle classifiers to represent functional hand movement patterns, as shown in Table II. Since the four muscle configurations were acquired independently rather than simultaneously, these results represent a conceptual simulation of the proposed framework rather than an experimental validation of simultaneous multi-channel gesture recognition. A multi-channel acquisition system is required to experimentally verify temporal muscle coordination in real time. Nevertheless,

the proposed pipeline is directly scalable to future multi-channel implementations using the same processing steps.

One limitation of the current implementation is its reduced capability to distinguish subtle motor actions involving highly correlated muscles, such as isolated index-finger extension. This limitation arises primarily from anatomical cross-talk and the independent acquisition of each muscle rather than from the proposed processing pipeline. Future multi-channel, multi-class implementations are expected to overcome this limitation while preserving the proposed methodology.

Therefore, the obtained results indicate that combining low-overhead binary classifiers is a viable strategy for representing functional hand movement patterns from independently acquired muscle activations. The proposed features-and-clustering framework achieves high classification performance within the evaluated datasets, supporting the feasibility of the proposed methodology as a proof of concept for future scalable myoelectric control systems.

IV. Conclusions

The main contribution of this work is the development of a feature-based pipeline combined with unsupervised learning for semi-automatic annotation of sEMG signals. By extracting time-, frequency-, and statistical-domain descriptors and applying clustering algorithms, the proposed approach identified contraction and relaxation patterns without relying on a single manually defined amplitude threshold. The generated labels were subsequently used as pseudo-labels for supervised classification, with overall accuracies ranging from 96% to 99% across the evaluated configurations. As a complementary quantitative assessment, a representative subset of 3280 analysis frames achieved 98.05% Accuracy and a Cohen's Kappa of 0.961 when compared with an manual reference annotation. These results indicate that the proposed pipeline was able to generate labels consistent with the observed sEMG activation patterns under the evaluated experimental conditions.

Unlike threshold-based approaches (e.g., [8]), which rely on empirical amplitude thresholds, the proposed methodology leverages multi-domain feature extraction and unsupervised learning. This multidimensional representation allows the pipeline to capture complementary signal characteristics and generate pseudo-labels with reduced manual intervention under the evaluated acquisition conditions, offering a practical end-to-end framework for sEMG dataset annotation with reduced manual effort.

The primary limitation of the present work is the experimental acquisition setup and the limited number of participants. Only a single healthy volunteer was evaluated, and the availability of a single sEMG acquisition module required sequential data acquisition sessions for each muscle configuration, preventing simultaneous multi-channel validation. Additionally, forearm muscle cross-talk limited the separability of highly subtle or isolated finger movements. Therefore, the results should be interpreted as an intra-subject proof of concept and feasibility assessment rather than evidence of generalization or robustness across subjects.

Future work will include simultaneous multi-channel acquisition, evaluation with additional participants, and quantitative comparison with established threshold-based labeling methods. The proposed feature extraction and learning framework can also be adapted to different acquisition hardware and extended to multi-channel configurations, providing a basis for further investigation of automated sEMG dataset annotation and myoelectric control.

Data Availability

The full data processing code, feature engineering scripts, clustering optimization routines, and dataset logs are openly available in our GitHub repository at GitHub repository to ensure transparency, reproducibility, and future open-source improvements by the biomedical engineering community.

Acknowledgment

The authors of this work would like to thank the Universidade Federal de Juiz de Fora (UFJF) for supporting this research.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Rangayyan, R.M., "Biomedical Signal Analysis", *IEEE Press Series on Biomedical Engineering*, vol. 3, 1997, Wiley.

[2] Silverthorn, D.U., "Human Physiology: An Integrated Approach", 7th Edition, 2017, Pearson.

[3] C. J. De Luca, "The use of surface electromyography in biomechanics," *Journal of Applied Biomechanics*, vol. 13, no. 2, pp. 135–163, 1997.

[4] D. Farina and O. A. Aszmann, "Bionic limbs: Clinical reality and academic promises," *Science Translational Medicine*, vol. 6, no. 257, p. 257ps12, 2014.

[5] T. Feix, J. Romero, H. B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The GRASP Taxonomy of Human Grasp Types," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2016.

[6] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface EMG," *Journal of Applied Physiology*, vol. 96, no. 4, pp. 1486–1495, 2004.

[7] M. B. I. Raez, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG signal analysis: detection, processing, classification and applications," *Biology Procedures Online*, vol. 8, pp. 11–35, 2006.

[8] K. H. Lee, J. Y. Min, and S. Byun, "Electromyogram-Based Classification of Hand and Finger Gestures Using Artificial Neural Networks," *Sensors (Basel)*, vol. 22, no. 1, p. 225, 2021, doi: 10.3390/s22010225.

[9] R. H. Chowdhury, M. B. I. Reaz, M. A. B. M. Ali, A. A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, no. 9, pp. 12431–12466, 2013.

[10] J. E. Muscolino, *The Muscle and Bone Palpation Manual: With Trigger Points, Referral Patterns and Stretching*. St. Louis: Elsevier, 2016.

[11] S. D. Waldman, *Atlas of Pain Management Injection Techniques*, 4th ed. Philadelphia: Elsevier, 2023.

[12] F. H. Netter, *Atlas of Human Anatomy*, 2nd ed. Philadelphia: Elsevier, 2014.

[13] J. V. Basmajian and C. J. De Luca, *Muscles Alive: Their Functions Revealed by Electromyography*, 5th ed. Baltimore: Williams & Wilkins, 1985.

[14] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Cham: Springer, 2015.

[15] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

Exploratory Analysis of Electroencephalographic Coherence in Neurocritical Patients with Suspected Brain Death: A Case Series

Eduarda de Paula Nogueira Soares
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
eduarda.soares@ufu.br
ORCID: 0009-0007-2596-4817

João Batista Destro Filho
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
jbdestro@ufu.br
ORCID: 0000-0001-5306-8058

Ederson Rosa da Silva
School of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ersilva@ufu.br
ORCID: 0000-0001-5093-9671

Abstract—Introduction: Brain death (BD) is the complete and irreversible loss of all brain function. In Brazil its determination is strictly clinical and legal and follows CFM Resolution No. 2,173/2017, under which electroencephalography (EEG) is an ancillary test used to establish electrocerebral inactivity and has no standing as a criterion on its own. Besides visual inspection, quantifiers such as spectral coherence describe the degree of synchronization between homologous cortical regions. Objective: To describe, in exploratory and descriptive terms, the quantitative behaviour of inter-hemispheric spectral coherence and its temporal and regional dispersion in EEG recordings of neurocritical patients admitted to an intensive care unit (ICU) and undergoing the brain death protocol. No diagnostic criterion is proposed or validated. Methods: A series of four cases (A, B, C and D) comprising seven EEG examinations. In cases A, B and C the protocol was not confirmed and cerebral arteriography showed residual brain perfusion (six examinations); in case D arteriography showed no intracranial arterial opacification and brain death was confirmed (one examination). Ten artefact-free 2 s epochs were selected from each examination, totalling 70 epochs. Magnitude-squared coherence (a dimensionless quantity, 0–1) was estimated by Welch's method in the delta (1–3 Hz), theta (4–8 Hz), alpha (8–13 Hz) and beta (13–35 Hz) bands, over eight homologous inter-hemispheric pairs of the international 10–20 system, using a referential auricular montage. Because epochs are nested within examinations and examinations within patients, they are not independent observations, and no hypothesis test was applied. Results are reported per case and per examination and are also presented graphically. Results: The overall median coherence was 0.18 in case A, 0.17 in case C, 0.13 in case B and 0.11 in case D. Case D showed the lowest median in all four bands and was lower than the three comatose cases in 20 of the 32 electrode-pair/band combinations. Parieto-occipital pairs concentrated the highest values in A and C (ratios of 1.7 and 2.1 relative to the remaining pairs), a pattern absent in B and reversed in D. In the two cases with serial examinations, coherence decreased across recordings. Conclusion: The findings are descriptive, restricted to four patients and to a single examination in the confirmed brain death case, and support no inference about diagnostic performance or about discrimination between coma and brain death. They are offered as an exploratory signal-processing illustration; EEG remains an ancillary test, subordinate to the clinical and legal protocols in force.

Keywords—brain death, electroencephalography, quantitative EEG, spectral coherence, case series, exploratory study

I. Introduction

Brain death (BD) is the complete and irreversible loss of all brain function, including that of the brainstem, and is legally recognized as a criterion of death in many countries [1]. The diagnosis carries high clinical, ethical and legal weight, because it determines the withdrawal of therapeutic support and opens the organ donation and transplantation process. An accurate determination protects the safety of the procedures that follow, the reliability of care protocols and compliance with the legal criteria set by national and international guidelines [2], [3].

In Brazil, brain death determination is governed by CFM Resolution No. 2,173/2017 [2], which requires two clinical examinations by different physicians, an apnoea test and, mandatorily, an ancillary test able to demonstrate absence of cerebral blood perfusion, absence of cerebral metabolic activity or electrocerebral inactivity. EEG belongs to this last category: it establishes electrocerebral inactivity, and its report depends on the clinical prerequisites set out in the regulation, namely body temperature above 35 °C, arterial oxygen saturation above 94 %, systolic blood pressure of at least 100 mmHg or mean arterial pressure of at least 65 mmHg in adults, and exclusion of central nervous system depressant drugs. The technical recording parameters (minimum number of electrodes, sensitivity, filters, recording time, impedance check and integrity tests) follow the recommendations of the International Federation of Clinical Neurophysiology (IFCN) [3]. The diagnosis is therefore clinical and legal, and no quantifier derived from the EEG, including the coherence described here, replaces or alters that determination.

Ancillary tests gain weight when clinical factors limit or hinder neurological assessment. Electroencephalography is non-invasive, widely available and has high temporal resolution, which allows real-time monitoring of cerebral electrical activity and yields information on cortical function [4], [5].

Advances in digital signal processing (DSP) have brought several quantitative approaches to the EEG, as auxiliary tools intended to make neurophysiological analysis more objective. Studies by [6], [7] and [8] described parameters related to spectral energy, entropy and hybrid analysis techniques in patients with brain death and in individuals with severe impairment of consciousness, and [9] observed changes in

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

power spectral density parameters in patients undergoing the brain death protocol.

Alongside spectral approaches, brain functional connectivity has also been investigated. One of its metrics is electroencephalographic coherence, the degree of synchronization between signals recorded at different brain regions within a given frequency band, which informs on the functional integrity of neural networks and on the communication between distinct areas [10], [11]. States of reduced or lost consciousness show a progressive decrease in this connectivity, a sign that the neural integration sustaining cognitive and physiological function is impaired.

Interpreting the EEG within brain death protocols remains difficult despite the technical and methodological advances of recent decades. Artefacts from the intensive care environment, electromagnetic interference, physiological noise and low-amplitude residual activity can hinder signal analysis and disturb the clinical reading [12]. The literature also describes electroencephalographic activity persisting after the clinical diagnosis of brain death: [13] reported low-amplitude theta and beta activity in patients considered clinically dead, and [14] described the persistence of organized patterns, including sleep spindles and residual alpha activity, for prolonged periods after clinical confirmation of BD. In line with these observations, [15] followed a comatose ICU patient throughout the dying process and showed that the temporal dynamics of quantitative EEG features, among them the frequency spectrum and the aperiodic (1/f) component of the signal, can still be observed in the last hour before death.

This work describes the quantitative behaviour of inter-hemispheric spectral coherence and its dispersion in EEG recordings of neurocritical patients in coma and of one patient who progressed to diagnostic confirmation of brain death. It is a series of four cases, analysed in exploratory and descriptive terms, with the data of each patient presented individually. The study does not compare populations, estimate diagnostic accuracy, validate a discrimination method or propose any additional criterion for brain death determination.

To the best of our knowledge, this type of coherence analysis has not been reported for recordings of Brazilian patients undergoing the brain death protocol. That gap motivates the present description but gives the findings no confirmatory value: with four patients, the work is an exploratory signal-analysis illustration and provides no evidence about the condition under study.

II. Methodology

This descriptive study assessed electroencephalography examinations obtained from four patients admitted to the intensive care unit (ICU) with clinical suspicion of brain death. The study was approved by the Research Ethics Committee of the Federal University of Uberlândia (CEP-UFU), under protocol No. 82824017.5.0000.5152.

A. Inclusion criteria

Patients were included if they had at least one EEG examination requested during the investigation for the brain death protocol. Recordings were required to be free of significant artefacts capable of disturbing clinical interpretation and to last at least 18 min. Only patients whose death was subsequently confirmed in the hospital records, and who had cerebral arteriography as an ancillary test, were considered.

B. Cases analysed and terminology

The recordings come from four patients, identified as A, B, C and D, and comprise seven EEG examinations. In cases A, B and C the brain death protocol was not confirmed: cerebral arteriography showed residual brain perfusion, and the patients died between 18 and 33 days after the last EEG analysed. In case D, arteriography showed no intracranial arterial opacification and brain death was confirmed.

The four cases are described individually, and the comparisons presented below are descriptive. Although the database contained other recordings, only those with adequate technical quality and complete clinical information were selected, a criterion that introduces selection bias and is revisited in Section V.

TABLE I. Clinical characterization of the four cases and ancillary test finding. Age in years; BD conf.: brain death confirmed.

Case	Sex	Age	Aetiology	Exams	Days to death	BD conf.	Cerebral arteriography
A	M	53	Haemorrhagic stroke	2	33 e 29	No	Faint opacification of right M1 and A1
B	M	21	Traumatic brain injury	3	22, 20 e 18	No	Faint opacification of carotid, A1 and basilar
C	M	73	Traumatic brain injury	1	26	No	Faint opacification of M1/M2 and basilar “stop”
D	F	53	Haemorrhagic stroke	1	6	Yes	No intracranial arterial opacification

Table I presents the clinical characterization of the four cases. In the three cases in which the protocol was not confirmed there is documented residual arterial opacification, however faint and territorially limited, whereas in case D there was no intracranial arterial opacification. This is a perfusion correlate obtained by a method independent of the EEG, which situates the four cases without resorting to the very quantifier under description.

TABLE II. EEG recordings analysed (seven examinations from four patients).

Case	Examination	Duration (min)	Epochs (n)	Days to death
A	1st	31	10	33
A	2nd	30	10	29
B	1st	30	10	22
B	2nd	30	10	20
B	3rd	30	10	18
C	1st	18	10	26
D	1st	30	10	6

The seven examinations are described in Table II. Durations ranged from 18 to 31 min. Case B contributed three examinations over five days, case A two examinations over four days, and cases C and D one examination each. The brain death protocol set out in CFM Resolution No. 2,173/2017 allows a variable number of EEG examinations, according to the clinical circumstances of each admitted patient. Ten 2 s epochs were extracted from each examination, yielding 70 epochs: 60 in the three comatose cases and 10 in case D (Table III).

TABLE III. Summary of the sample used in signal processing.

Set	Patients (n)	Exams (n)	Epochs (n)	Brain perfusion at arteriography
Comatose cases (A, B, C)	3	6	60	Residual, documented
Confirmed BD case (D)	1	1	10	Absent
Total	4	7	70	—

C. Clinical conditions: sedation, temperature and metabolic factors

All seven examinations were requested and performed within the brain death protocol, and therefore under the clinical conditions that CFM Resolution No. 2,173/2017 [2] requires for the ancillary test: the temperature, oxygen saturation and blood pressure thresholds and the exclusion of central nervous system depressant drugs listed in Section I. Checking these prerequisites is a regulatory condition for performing the test and falls to the attending team, which documents it in the protocol file. The technical recording parameters follow the IFCN recommendations [3].

Compliance with the prerequisites is assured at protocol level, but the individual values of temperature, blood pressure, blood gases, natraemia and glycaemia at the moment of each acquisition, as well as sedative dose and time since withdrawal, cannot be retrieved from the records available for this retrospective study. It is therefore not possible to quantify the margin by which each examination met the thresholds, nor to rule out entirely a residual sedation effect. Residual sedation, borderline hypothermia, hypotension, hypercapnia, hypoglycaemia and hydroelectrolytic disturbances each reduce signal amplitude and synchronization on their own, and can produce coherence drops similar to those described here. No difference observed between cases can therefore be attributed exclusively to the neurological condition, a point revisited in Section V.

Chronology of case D. The single EEG of patient D was recorded during the protocol investigation, six days before death. The cerebral arteriography that supported the diagnostic confirmation, and which showed no intracranial arterial opacification, was performed six days after that EEG, on the same day as death. At the moment the analysed signal was acquired, therefore, brain death had not yet been confirmed. Since confirmation came from arteriography, the EEG examination considered here is the neurophysiological information closest to that outcome and, in that sense, the most representative available. None of the seven examinations analysed corresponds to a confirmatory protocol report.

D. EEG acquisition: montage, reference, impedance and filters

Signals were acquired with electrodes positioned according to the international 10–20 system, in the intensive care environment, for the clinical purpose of supporting the brain death protocol. A referential montage was used, with the auricular electrodes A1 and A2 as reference. Coherence was computed over eight homologous inter-hemispheric pairs derived from this montage: Fp1-Fp2, F7-F8, F3-F4, T3-T4, C3-C4, T5-T6, P3-P4 and O1-O2.

Examinations lasted between 18 and 31 min, with a sampling frequency of 100 Hz and an acquisition sensitivity corresponding to a ±200 μV range. No digital filter was applied before processing, which preserves the original signal characteristics. A 60 Hz notch filter was likewise not used, which does not affect the analysed bands (1 to 35 Hz) but is relevant to the characterization of background noise. The reports do not give the impedance value measured in each examination; since the recordings were performed within the brain death protocol, the technical conditions recommended by the IFCN [3] apply: impedance check between 100 Ω and 10 kΩ, electrode integrity testing and calibration recording.

E. Pre-processing, epoch selection and artefact rejection

Only artefact-free epochs were quantitatively analysed. They were chosen visually and by consensus of two experienced neurologists, who have been responsible for more than ten years for the EEG reports of the brain death protocol at the Hospital de Clínicas of Uberlândia, and who selected the segments of greatest clinical relevance and excluded those with artefacts. Segments with movement artefact, 60 Hz interference from ICU equipment, electrode artefact (saturation or disconnection), electrocardiogram artefact and mechanical ventilation artefact were discarded. Ten 2 s epochs were extracted from each examination: 60 in the comatose cases and 10 in case D, 70 in total.

Beyond segment exclusion, individual electrodes were excluded whenever they showed loss of contact, excessive noise or saturation. Because coherence is computed between homologous pairs, excluding one electrode eliminates the corresponding pair in that examination. Of the 56 possible pairs (7 examinations × 8 pairs), six could not be analysed: F3-F4 in the 2nd examination of case A, O1-O2 in the 1st examination of case B, C3-C4 and P3-P4 in the 3rd examination of case B,

and Fp1-Fp2 and F3-F4 in the examination of case C. In the results tables and figures these pairs appear as missing data, not as zero coherence.

Artefact rejection was visual and not automated, and selection by clinical relevance systematically favours the segments of best quality.

F. Spectral estimation and coherence computation

Coherence evaluates the degree of synchronization between two signals across frequency bands and provides information on the functional interaction between distinct brain regions [10]. The computation is based on the cross power spectral density between two signals, normalized by their individual spectral densities. The resulting values range between 0 and 1: values close to 1 indicate greater synchronization between the electrodes analysed, whereas values close to 0 reflect low synchronization [10].

Coherence between two signals is defined as:

$$|\Gamma_{xy}(e^{j\omega})|_i^2 = \frac{|S_{xy}(e^{j\omega})|^2}{S_x(e^{j\omega})\, S_y(e^{j\omega})} \quad (1)$$

where:

Sxy – Cross power spectral density.

Sx – Power spectral density of signal X.

Sy – Power spectral density of signal Y.

i – Segment considered.

This equation satisfies the following property:

$$0 \leq |\Gamma_{xy}(e^{j\omega})|^2 \leq 1 \quad (2)$$

Magnitude-squared coherence was estimated by Welch's method [16], averaging the ten 2 s epochs of each examination as segments of the spectral average, with a Hann window and no overlap. Each epoch contains 200 samples at 100 Hz. The value assigned to each band is the mean of the coherence values over the spectral components it contains. Coherence is a dimensionless quantity restricted to the interval [0, 1], so the values reported here carry no unit.

The analysis covered the 1 to 35 Hz range, with 1 Hz resolution, divided into four bands: Delta (1 to 3 Hz, 3 spectral components), Theta (4 to 8 Hz, 5), Alpha (8 to 13 Hz, 6) and Beta (13 to 35 Hz, 23). The 8 and 13 Hz limits are shared by adjacent bands. With a sampling frequency of 100 Hz the Nyquist frequency is 50 Hz, and the whole analysed range lies below it. Frequencies above 35 Hz were not analysed, because they approach the sampling limit and because the reports do not document the anti-aliasing filter used at acquisition, which compromises the reliability of estimates in that region. Mains interference at 60 Hz, if present and unfiltered before sampling, folds back to 40 Hz, above the analysed range [11], [17].

Magnitude-squared coherence estimated from a finite number of segments is biased upwards, and its expected value between two statistically independent signals is approximately 1/L, where L is the number of averaged segments [18]. The medians observed in this study lie between 0.10 and 0.27, a magnitude consistent with the floor expected for L on the order of ten. The values described here are therefore close to the level expected in the absence of coupling, particularly the smallest ones, and should not be read as measures of preserved connectivity.

G. Descriptive and exploratory analysis

For each examination and for each of the 32 electrode-pair/band combinations, a complete set of descriptive statistics of the coherence coefficients was computed: mean, median, standard deviation, variance, coefficient of variation, minimum and maximum, first and third quartiles, skewness and kurtosis. Each of these distributions gathers the coherence values of the ten epochs across all spectral components of the band, that is, 30 values in Delta, 50 in Theta, 60 in Alpha and 230 in Beta. The dispersion measured in this way therefore combines temporal variation and variation across the spectrum within the band. The synthesis presented below uses the median as the measure of central tendency and two complementary measures of dispersion, which describe distinct axes: the standard deviation of the coherence coefficients within each band, which measures the variability of synchronization across the spectrum and across the epochs of a given pair, and the range across the eight homologous pairs, which measures spatial heterogeneity over the scalp. All are descriptive measures, not estimates of population parameters.

The data structure is nested: epochs belong to the same examination, electrode pairs share channels and reference, examinations belong to the same patient, and the set comprises four patients, one of them with confirmed brain death. Under these conditions, a hypothesis test would treat as independent observations values that in practice describe four individuals, and the resulting p-value would mostly reflect the number of epochs and pairs considered. An entirely descriptive analysis was therefore adopted, without hypothesis tests and without qualifying results as significant.

For the same reason, percentage differences between sets of examinations are not computed. Such a measure would require pooling the three comatose cases into a single average, which would erase precisely the individual variation the series sets out to show. Results are therefore presented case by case and examination by examination, in tables and in graphs.

III. Results

Results are presented case by case and examination by examination. Table IV summarizes coherence behaviour in each case and band, Table V details the topography pair by pair, and Table VI descends to the level of the individual examination. Figs. 1 and 2 present the same information visually.

TABLE IV. Coherence by case and band: median, range across the eight homologous pairs, and standard deviation within the band. Dimensionless

values (0–1); n = number of electrode-pair × examination combinations effectively analysed, after exclusion of pairs with a discarded electrode.

Case	Band	n	Median	Min–Max across pairs	SD within band
A	Delta	15	0.17	0.08–0.56	0.16
	Theta	15	0.18	0.08–0.61	0.14
	Alpha	15	0.21	0.12–0.47	0.17
	Beta	15	0.14	0.09–0.33	0.15
B	Delta	21	0.14	0.07–0.31	0.15
	Theta	21	0.13	0.07–0.19	0.13
	Alpha	21	0.16	0.09–0.22	0.14
	Beta	21	0.12	0.09–0.17	0.13
C	Delta	6	0.14	0.10–0.31	0.15
	Theta	6	0.23	0.11–0.39	0.15
	Alpha	6	0.18	0.14–0.34	0.17
	Beta	6	0.14	0.12–0.34	0.15
D	Delta	8	0.12	0.07–0.19	0.12
	Theta	8	0.11	0.08–0.12	0.12
	Alpha	8	0.12	0.10–0.22	0.13
	Beta	8	0.11	0.11–0.14	0.14

Each row of Table IV gathers the values of one case in one band. Column n reports how many electrode-pair and examination combinations entered the computation, already discounting the pairs lost to electrode exclusion. Median is the central coherence value in that band, the value that splits the observations in half. Min–Max across pairs gives the smallest and the largest value observed among the eight homologous pairs: in case A, for instance, coherence in Delta ranged from 0.08 in the least synchronized pair to 0.56 in the most synchronized one, and this range measures how much synchronization differs from one scalp region to another. SD within band is the standard deviation of the coherence coefficients inside the band, computed over the ten epochs and all the spectral components that compose it; it measures how much coherence fluctuates over time and frequency within a given pair, not between different pairs.

TABLE V. Median [Q1–Q3] coherence by electrode pair, in each case (pooling the four bands and the examinations of the case).

Electrode pair	A	B	C	D
Fp1-Fp2	0.16 [0.07–0.28]	0.15 [0.06–0.29]	—	0.13 [0.05–0.22]
F7-F8	0.15 [0.09–0.27]	0.12 [0.05–0.25]	0.15 [0.06–0.30]	0.11 [0.05–0.21]
F3-F4	0.16 [0.07–0.26]	0.13 [0.07–0.25]	—	0.12 [0.05–0.21]
T3-T4	0.18 [0.08–0.29]	0.16 [0.08–0.29]	0.22 [0.11–0.33]	0.12 [0.06–0.22]
C3-C4	0.24 [0.11–0.41]	0.14 [0.05–0.25]	0.14 [0.07–0.25]	0.12 [0.05–0.23]
T5-T6	0.16 [0.06–0.23]	0.13 [0.06–0.21]	0.12 [0.06–0.24]	0.11 [0.06–0.22]
P3-P4	0.19 [0.10–0.34]	0.13 [0.06–0.22]	0.19 [0.09–0.34]	0.10 [0.05–0.16]
O1-O2	0.27 [0.16–0.42]	0.13 [0.05–0.23]	0.34 [0.20–0.50]	0.10 [0.05–0.22]

In Table V, each cell gives the median coherence of that electrode pair in that case, followed in brackets by the first and third quartiles, Q1 and Q3. Below Q1 lie 25 % of the observations and below Q3, 75 %; together they delimit the central half of the distribution and indicate how concentrated the values are around the median. A dash marks a pair not analysed because of electrode exclusion. Unlike Table IV, here the four bands and the examinations of each case are pooled: the reading is spatial, comparing scalp regions rather than frequency bands.

TABLE VI. Median coherence in each of the seven examinations, by band.

Examination	Delta	Theta	Alpha	Beta
A, 1st	0.21	0.24	0.26	0.17
A, 2nd	0.12	0.13	0.16	0.12
B, 1st	0.18	0.16	0.16	0.12
B, 2nd	0.13	0.13	0.13	0.11
B, 3rd	0.13	0.11	0.17	0.13
C, 1st	0.14	0.23	0.18	0.14
D, 1st	0.12	0.11	0.12	0.11

Table VI gives the median coherence of each of the seven examinations, band by band. It is the only table in which the examinations of the same patient appear separately, which allows the evolution of cases A and B across successive recordings to be followed.

The overall median coherence was 0.18 in case A, 0.17 in case C, 0.13 in case B and 0.11 in case D. Case D showed the lowest median in all four bands (0.12 in Delta, 0.11 in Theta, 0.12 in Alpha and 0.11 in Beta), although the margin over case B was small, 0.01 in Beta. Considering the 32 electrode-pair/band combinations, the median of case D was lower than that of the three comatose cases in 20 of them; in the remaining 12, at least one comatose case showed an equal or lower value.

Heterogeneity among the three comatose cases is evident and argues against treating them as a single set. The overall median of B (0.13) is closer to that of D (0.11) than to those of A (0.18) and C (0.17). The same holds for spatial dispersion: the range across the medians of the eight homologous pairs was wide in A (0.13 to 0.22 depending on the band) and in C (0.20 to 0.28), and narrow in D (0.04 to 0.13), but also in B (0.05 to 0.10). The spatial homogeneity of case D therefore does not distinguish it from all comatose cases.

The largest difference between cases appeared in the topographic distribution. In cases A and C, posterior pairs concentrated the highest values: the median of P3-P4 and O1-O2 was 1.7 times that of the remaining pairs in A and 2.1 times in C, with O1-O2 reaching 0.27 and 0.34, respectively. This postero-anterior gradient did not appear in B (ratio 1.0) and was reversed in D (ratio 0.9), where P3-P4 and O1-O2 recorded the lowest values of all pairs, 0.10 in both.

In the two cases with serial examinations, coherence fell across recordings (Table VI). The medians of case A went from 0.21 to 0.12 in Delta, from 0.24 to 0.13 in Theta, from 0.26 to 0.16 in Alpha and from 0.18 to 0.12 in Beta, between the first and the second examination. In B, the drop concentrated between the first and the second examination (Delta 0.18 to 0.13; Theta 0.16 to 0.13), with relative stability in the third. At the end of these trajectories, the medians of A and B approach the values observed in case D.

All values lie between 0.07 and 0.61, and the vast majority between 0.10 and 0.25, a range which, as set out in Section II-

F, is of the order of the floor expected for signals without coupling. In case D, the medians of the four bands (0.11 to 0.12) are practically indistinguishable from each other and from that level.

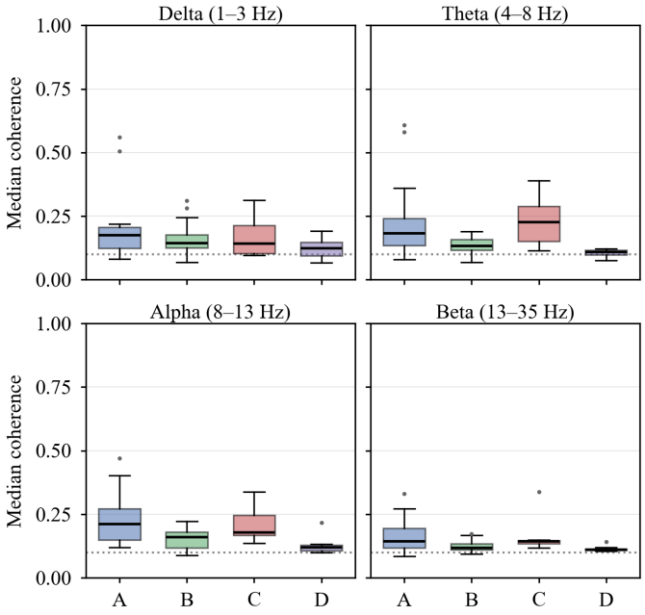


Fig. 1. Dispersion of median coherence across the eight homologous pairs, in each case and band. Each point forming the box is the median of one electrode pair in one examination. The box therefore describes the spatial heterogeneity of synchronization, not the variability within the band, which is given in Table IV. Dotted line: estimation floor (1/L ≈ 0.10 for L = 10 epochs).

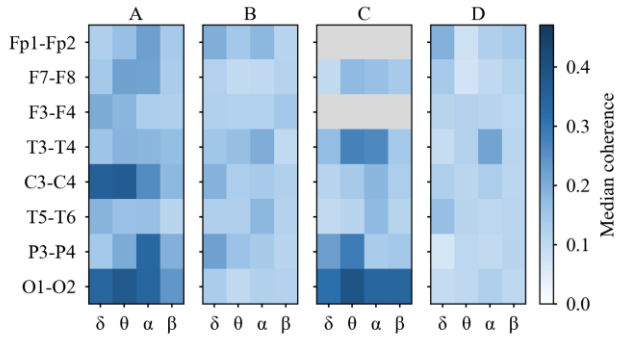


Fig. 2. Median coherence by electrode pair and band, in each patient (δ: delta; θ: theta; α: alpha; β: beta). Colour scale common to the four panels. Dimensionless values; in cases A and B, the median across examinations. Grey cells: pair not analysed because of electrode exclusion.

IV. Discussion

Electroencephalographic coherence is widely used as a metric of brain functional connectivity, decreasing progressively as states of impaired consciousness deepen. This reduction reflects the loss of functional synchronization between cortical regions, particularly in diffuse cerebral dysfunction [5], [19]. Quantitative studies in comatose patients show coherence values generally lower than those observed in neurologically healthy individuals [20], [21], varying with the electrode pair and the band considered.

One feature sets this series apart from purely electroencephalographic descriptions: all four cases have cerebral arteriography, a perfusion measure that does not depend on the EEG. In the three comatose cases there is documented residual arterial opacification, however faint; in case D there is no intracranial opacification. Any coherence difference between them can therefore be read against an objective circulatory finding as well as against the clinical label. The comparison remains descriptive: with a single case of absent perfusion, no association can be estimated.

The pathophysiological reading of the findings remains plausible. In the absence of organized neuronal activity, the recorded signal tends to be dominated by instrumental and environmental noise, and coherence indices settle at low, weakly varying values. In comatose patients, by contrast, the persistence of irregular electrical activity favours larger fluctuations in these indices. The observed values, however, approach the level expected for electrical and instrumental noise in the ICU environment. With L = 10 epochs, the coherence expected between two independent signals is of the order of 0.10 [18], a magnitude comparable to the medians described here. Low coherence in this regime is compatible both with the collapse of corticocortical integration and with the mere absence of signal above the noise, and the two situations cannot be separated by the available data.

The smaller standard deviations observed in case D are consistent with a low-amplitude signal of little dynamics, as expected when no organized detectable cortical electrical activity remains [1], and consistent here with the absence of intracranial arterial opacification. The same signature, however, may arise from deep sedation, hypothermia or better technical recording quality, factors that could not be controlled (Section V). The pathophysiological reading is therefore one among several explanations compatible with the data.

The spatial distribution separates the cases more sharply than any other measure, and is also the hardest to interpret. In cases A and C, the parieto-occipital pairs concentrated the highest coherence values, with ratios of 1.7 and 2.1 relative to the remaining pairs. In case D, these same pairs recorded the lowest values, reversing the gradient. One possible reading links this pattern to the involvement of parieto-occipital regions in cortical integration and in the maintenance of consciousness levels, particularly in the alpha and theta bands [22]: loss of the gradient would accompany the collapse of posterior corticocortical integration.

Two facts from the material itself weigh against this reading. The first is case B, comatose and with documented residual perfusion, which likewise shows no posterior gradient (ratio 1.0) and whose overall median is the closest to that of case D. The topographic finding therefore does not separate the cases by outcome. The second is that posterior pairs are the most exposed to decubitus artefact in the ICU bed and to variations in electrode contact, and it was precisely among them that exclusions for discarded electrodes occurred. With a single case of confirmed brain death, the observed topography cannot be distinguished from individual variation.

The material also allows a limited longitudinal reading, restricted to the comatose cases. In A and in B coherence decreased across successive examinations, and at the end of

both trajectories the values approach the level observed in case D. The observation is suggestive, as it would be compatible with progressive deterioration of cortical synchronization, but both patients had documented residual brain perfusion and neither had confirmed brain death. The decrease cannot therefore be read as a trajectory towards brain death. For case D no longitudinal reading is possible, since only one examination is available, obtained six days before death and during the protocol investigation. The electrophysiological evolution towards brain death, which would be the finding of greatest interest, cannot be described from this material.

V. Limitations

This is a series of four cases, and its limitations determine the reach of what can be asserted. Part of them stems from the practical difficulty of obtaining ethical authorization from the patients' families. By determination of the UFU Ethics Committee, the request for authorization took place during the family's visit to the patient, a moment ill-suited to such a request given the relatives' distress and anxiety for news.

Sample size and design. Four patients and seven examinations, of which only one corresponds to confirmed brain death, do not allow between-patient variability to be estimated nor the effect of the clinical condition to be separated from individual effects. Any difference observed is entirely compatible with variation between individuals.

Single examination in the brain death case. Case D contributes one examination and 10 epochs. Since the methodology followed the ICU care protocol, the clinical circumstances of this patient led to a single EEG recording.

Absence of additional metrics. The analysis is restricted to spectral coherence, without comparison with other EEG quantifiers such as entropy, phase synchronization (PLV), graph connectivity metrics or machine learning approaches, which could characterize the electrophysiological behaviour of the cases more completely [7], [8]. The space limit imposed by the conference rules did not allow their inclusion in this description.

Confounding factors documented only at protocol level. The clinical prerequisites of CFM Resolution No. 2,173/2017 (temperature, oxygenation, blood pressure and exclusion of central nervous system depressants) are a regulatory condition for performing the test and were verified by the attending team (Section II-C), but the individual values at the moment of each acquisition are not in the retrieved records. The margin by which each examination met the thresholds cannot be quantified, and residual sedation, borderline hypothermia or minor metabolic disturbance, each capable of producing the patterns described here, cannot be ruled out.

VI. Conclusion

In this exploratory case series, EEG spectral coherence behaved differently in the patient who progressed to diagnostic confirmation of brain death and in the patients with severe neurological injury: in the former, coherence stood at the lowest level of the series in all bands, and the parieto-occipital gradient present in two of the comatose cases did not appear. Given the small sample, the overlap with one of the comatose cases and the descriptive design, coherence should be understood as an investigative signal-processing metric.

The contribution is therefore methodological: an exploratory illustration of how quantitative coherence analysis can be applied to recordings obtained within the brain death protocol, with a detailed description of the processing and of the limitations, and with the uncommon benefit of cerebral arteriography for all four cases. No diagnostic criterion is proposed, and EEG remains an ancillary test, subordinate to the clinical criteria and to the legislation in force.

Future studies require a substantially larger and multicentre sample, prospective recording of clinical confounders, standardization of montage and impedance, automated artefact rejection, inclusion of other quantifiers (entropy, phase synchronization, graph metrics), serial acquisition in cases progressing to brain death, and patient-level analysis with models that respect the nested structure of the data. Only then could the discriminative value of coherence be investigated.

Acknowledgment

The authors thank the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) for financial support, and the Hospital de Clínicas of the Federal University of Uberlândia (UFU), in particular its Department of Neurology, for the support and collaboration that made this research possible. The authors also thank Liliana Silva for her valuable collaboration in this work.

References

[1] E. F. M. Wijdicks, "The diagnosis of brain death," N. Engl. J. Med., vol. 344, no. 16, pp. 1215–1221, 2001, doi: 10.1056/NEJM200104193441606.

[2] Conselho Federal de Medicina, "Resolução CFM No. 2.173 of 23 November 2017: criteria for the diagnosis of brain death," Diário Oficial da União, Brasília, DF, Brazil, sec. 1, p. 274, Dec. 15, 2017.

[3] G. E. Chatrian, B. Bergamasco, A. Bricolo, J. D. Frost, Jr., and P. F. Prior, "IFCN recommended standards for electrophysiologic monitoring in comatose and other unresponsive states. Report of an IFCN committee," Electroencephalogr. Clin. Neurophysiol., vol. 99, no. 2, pp. 103–122, 1996, doi: 10.1016/0013-4694(96)96059-9.

[4] G. B. Young, "The EEG in coma," J. Clin. Neurophysiol., vol. 17, no. 5, pp. 473–485, 2000, doi: 10.1097/00004691-200009000-00008.

[5] E. Niedermeyer and F. L. da Silva, Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, 5th ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2004.

[6] G. Cui, Y. Yin, T. Tanaka, and J. Cao, "EEG energy analysis for evaluating consciousness level using dynamic MEMD," in Proc. Int. Joint Conf. Neural Networks (IJCNN), Beijing, China, 2014, pp. 3247–3254.

[7] L. Zhu, G. Cui, J. Cao, A. Cichocki, J. Zhang, and C. Zhou, "A hybrid system for distinguishing between brain death and coma using diverse EEG features," Sensors, vol. 19, no. 6, p. 1342, 2019, doi: 10.3390/s19061342.

[8] B. Li, J. Liu, T. Zhang, Y. Cao, and J. Cao, "Quantitative analysis and machine learning-based interpretation of EEG signals in coma and brain-death diagnosis," Cogn. Neurodyn., vol. 18, no. 5, pp. 2947–2962, 2024, doi: 10.1007/s11571-024-10131-y.

[9] E. P. N. Soares, P. H. S. Fernandes, L. Silva, M. S. N. Peres, and J. B. Destro-Filho, "Quantitative analysis of the electroencephalogram linked to the brain death protocol," in Proc. XVI Simpósio de Engenharia Biomédica (SEB 2025), Uberlândia, Brazil, 2025, doi: 10.5281/zenodo.17393806.

[10] L. Sörnmo and P. Laguna, Bioelectrical Signal Processing in Cardiac and Neurological Applications. Burlington, MA, USA: Elsevier Academic Press, 2005.

[11] G. Buzsáki, Rhythms of the Brain. New York, NY, USA: Oxford University Press, 2009.

[12] L. M. Alcantara, "Systematic review on the use of EEG for the diagnosis of brain death," Undergraduate thesis, Biomedical Eng., Federal Univ. of Uberlândia, Uberlândia, Brazil, 2025. [Online]. Available: https://repositorio.ufu.br/handle/123456789/45524

[13] M. M. Grigg, G. G. Celesia, M. W. Ghobrial, and E. R. Ross, "Electroencephalographic activity after brain death," Arch. Neurol., vol. 44, no. 9, pp. 948–954, 1987, doi: 10.1001/archneur.1987.00520210048018.

[14] E. A. Garcia, "Muerte encefálica: valor y limitaciones diagnósticas de la electroencefalografía," Med. Intensiva, vol. 24, no. 3, pp. 116–123, 2000, doi: 10.1016/S0210-5691(00)79571-4.

[15] S. Zinn, S. Z. Dragovic, J. A. Kloka, L. M. Willems, S. Harder, S. Kratzer, K. D. Zacharowski, G. Schneider, P. S. García, and M. Kreuzer, "Parametrization of the dying brain: a case report from ICU bed-side EEG monitoring," NeuroImage, vol. 305, Art. no. 120980, 2025, doi: 10.1016/j.neuroimage.2024.120980.

[16] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," IEEE Trans. Audio Electroacoust., vol. 15, no. 2, pp. 70–73, 1967, doi: 10.1109/TAU.1967.1161901.

[17] C. D. Ramos, "Characterization of the normal waking electroencephalogram: database construction and quantitative analysis," M.S. thesis, Electrical Eng., Federal Univ. of Uberlândia, Uberlândia, Brazil, 2017.

[18] G. C. Carter, "Coherence and time delay estimation," Proc. IEEE, vol. 75, no. 2, pp. 236–255, 1987, doi: 10.1109/PROC.1987.13723.

[19] R. Lehembre et al., "Electrophysiological investigations of brain function in coma, vegetative and minimally conscious patients," Arch. Ital. Biol., vol. 150, no. 2–3, pp. 122–139, 2012, doi: 10.4449/aib.v150i2.1374.

[20] C. D. Ramos and J. B. Destro-Filho, "Measurements of coherence in EEG signal in Brazilian people: a comparison of different consciousness states," Biosci. J., vol. 37, Art. no. e37065, 2021, doi: 10.14393/BJ-v37n0a2021-49454.

[21] M. Carrasco-Gómez et al., "EEG functional connectivity contributes to outcome prediction of postanoxic coma," Clin. Neurophysiol., vol. 132, no. 6, pp. 1312–1320, 2021, doi: 10.1016/j.clinph.2021.02.011.

[22] E. Başar, C. Başar-Eroglu, S. Karakaş, and M. Schürmann, "Gamma, alpha, delta, and theta oscillations govern cognitive processes," Int. J. Psychophysiol., vol. 39, no. 2–3, pp. 241–248, 2001.

Extraction of Computational Descriptors Based on ABCD Criteria for Characterization of Skin Lesions

1ª Nathália Giovanna Soares da Paz
Department of Biomedical Engineering
UFRN
Natal, Brazil
nathalia.paz.018@ufrn.edu.br

2ª Manuela Cerqueira Seyffarth
Department of Biomedical Engineering
UFRN
Natal, Brazil
manuela.seyffarth.056@ufrn.edu.br

3rd Deyvisom Michel Mendonça Dantas
Postgraduate studies in bioinfo
UFRN
Natal, Brazil
deyvisom84@gmail.com

4th Yasmin Maria Santana de Melo
Depto. de Eng. Biomédica
UFRN
Natal, Brasil
Yasmin.melo.041@ufrn.br

5th Heliana Bezerra Soares
Department of Biomedical Engineering
UFRN
Natal, Brazil
heliana.soares@ufrn.br

***Abstract*— Melanoma is the most aggressive form of skin cancer, making early diagnosis crucial to increasing the chances of successful treatment. This work proposes a methodology based on Digital Image Processing for the computational quantification of criteria A (asymmetry), B (border), and C (color) of the ABCD rule through the automatic extraction of geometric and pigmentation descriptors, aiming to evaluate its potential in differentiating between melanomas and non-melanomas. The methodology comprises the steps of automatic lesion segmentation in the CIELAB color space using the k-means algorithm, complemented by intensity-based thresholding in more complex cases, extraction of computational descriptors, and individual evaluation of their discriminatory power. The methodology was validated on a set of 20 images (10 melanomas and 10 non-melanomas), showing promising results. Among the descriptors evaluated, the asymmetry index showed the best performance, reaching 90% accuracy, while the border and pigmentation descriptors obtained accuracy of up to 70% and 75%, respectively. The results indicate that geometric descriptors, especially the asymmetry index, exhibit greater discriminatory capacity, reinforcing the potential of the proposed methodology as an interpretable computational tool to support the early diagnosis of melanoma. However, studies with larger image sets are needed to validate the robustness and generalizability of the approach.**

Keywords*— *Melanoma. Digital Image Processing. Skin Lesions. Border Signature.

I. Introduction

Skin cancer is one of the most common types of cancer worldwide, with melanoma being its most aggressive form due to its high metastatic potential. Although it represents a smaller proportion of cases, it is responsible for the majority of deaths associated with the disease, making early diagnosis fundamental to increasing the chances of successful treatment [1], [2].

With the advancement of imaging technologies, Digital Image Processing (DIP) techniques have been widely used to aid diagnosis, allowing the automatic extraction of relevant characteristics from skin lesions and reducing the subjectivity of visual assessment [2], [3], [4]. More recently, methods based on deep learning have achieved high performance in the classification of skin lesions [2], [4], [11]. However, these approaches generally require large labeled databases, high computational costs, and exhibit lower interpretability when compared to methods based on computational descriptors.

In this context, methods based on computational descriptors inspired by clinical criteria remain relevant because they allow for greater interpretability of results, lower computational complexity, and potential application in scenarios with limited data availability. However, studies that automatically quantify the clinical criteria of the ABCD rule using geometric and pigmentation descriptors capable of supporting the differentiation between melanomas and benign lesions are still scarce.

Given this scenario, this work proposes a methodology for the computational quantification of criteria A, B, and C of the ABCD rule through the automatic extraction of geometric and pigmentation descriptors. The methodology quantifies these characteristics and evaluates their discriminatory potential in differentiating between melanomas and non-melanomas, aiming to contribute to the development of computational tools to support diagnosis.

II. Materials And Methods

The proposed methodology performs computational quantification of the ABCD rule's clinical criteria of asymmetry, border, and color through the automatic extraction of geometric and pigmentation descriptors from dermatoscopic images. To this end, the method comprises the steps of image acquisition, automatic lesion segmentation, extraction of computational descriptors, and evaluation of the discriminatory power of these characteristics in differentiating between melanomas and non-melanomas. Fig. 1 presents the general flowchart of the developed methodology.

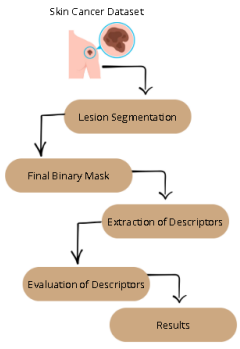


Fig. 1. – General flowchart of the proposed methodology

A. The Database

The images used in this study were obtained from the public Melanoma Skin Cancer Dataset [15], available on the Kaggle platform. This dataset consists of dermatoscopic images of skin lesions classified as melanoma and non-melanoma, and is widely used in research related to the development and evaluation of computational methods for the diagnosis of skin cancer. The original database contains approximately 10,000 images, encompassing different morphological and pigmentation characteristics of the lesions.

For the evaluation of the proposed methodology, a set of 20 randomly selected images was used, 10 melanomas and 10 non-melanomas, selected to represent both classes. This set was used to evaluate the behavior of the computational descriptors of asymmetry, border irregularity and pigmentation, as well as their discriminatory power in differentiating between melanoma and non-melanoma lesions.

B. Lesion Segmentation

Initially, all images were standardized to a resolution of 256 × 256 pixels and had their intensity levels normalized, aiming to standardize processing conditions and reduce variations resulting from image acquisition. Then, the images were converted from the RGB color space to the CIELAB space, whose separation between the luminance component (L) and the chromatic components (a and b) provides greater robustness against variations in illumination, favoring lesion segmentation [5], [13].

Segmentation was performed using the k-means algorithm, applied to the chromatic components of the CIELAB space. After the clustering step, the resulting binary mask was refined using morphological opening and closing operations, with the aim of removing small spurious regions and smoothing the lesion contour [3], [8], [10], [13].

When the segmentation obtained by the k-means algorithm did not present satisfactory quality due to the high heterogeneity of the lesion pigmentation, a complementary strategy based on contrast enhancement, conversion to grayscale, intensity thresholding and morphological operations was employed, allowing to obtain a mask more representative of the region of interest [12], [13].

The final mask obtained was used in subsequent steps to extract the computational descriptors related to the asymmetry, border and pigmentation criteria of the ABCD rule.

C. Extraction of Computational Descriptors

The extraction of computational descriptors aimed to automatically quantify the clinical criteria of asymmetry (A), border (B), and color (C) of the ABCD rule [7]. To this end, quantitative metrics were developed based on geometric properties and the pigmentation distribution of the lesion, allowing computationally representation of characteristics traditionally evaluated qualitatively during dermatoscopic examination [3], [5], [13].

The extraction of descriptors was performed from the binary mask obtained in the segmentation step. Initially, the geometric properties of the lesion were determined, including area, perimeter, centroid, and contour. This information served as the basis for constructing the radial signature and for calculating the descriptors related to asymmetry and border irregularity. In parallel, pigmentation analysis was performed using the luminosity component (L) of the CIELAB color space, allowing the acquisition of metrics associated with the color distribution in the segmented region [5], [13].

C.1 Asymmetry Descriptors

Asymmetry constitutes one of the main criteria of the ABCD rule for clinical evaluation of melanocytic lesions. In general, benign lesions have a more uniform geometric distribution, while melanomas tend to exhibit irregular growth, producing differences between opposite regions of the lesion [7]. To computationally represent this criterion, the radial signature was used, defined by the distances between the centroid of the segmented lesion and its contour in different angular directions [5], [13].

Initially, the centroid of the segmented region was determined based on the geometric properties of the binary mask. Next, 360 radial samples were obtained, corresponding to the distances between the centroid and the edge of the lesion for angles between 0° and 359°. This representation allows each angular direction θ to be associated with the diametrically opposite direction $\theta+\pi$.

For each pair of opposite directions, the local symmetry error was calculated, defined as the absolute difference between the respective radial distances, as presented in Eq. (1).

$$e(\theta) = |r(\theta) - r(\theta + \pi)| \quad (1)$$

where $r(\theta)$ represents the distance between the centroid and the edge of the lesion in the θ direction, $r(\theta+\pi)$ corresponds to the radial distance in the diametrically opposite direction, and $e(\theta)$ represents the local symmetry error. In approximately symmetrical lesions, opposite rays have similar values, producing low values of $e(\theta)$. More pronounced differences indicate less correspondence between opposite regions and, consequently, greater geometric asymmetry [3], [5].

After calculating the local symmetry error for each of the 180 diametrically opposite pairs of the radial signature, the Asymmetry Index (AI) was determined. To reduce the influence of the absolute size of the lesion, the sum of the errors was normalized by the sum of the radial distances corresponding to the evaluated pairs, as presented in Eq. (2).

$$IA = \frac{\sum e(\theta)}{\sum r(\theta)} \quad (2)$$

This normalization makes the index independent of the lesion scale, allowing for direct comparison of images with different dimensions and reducing the influence of absolute size on the asymmetry measurement.

C.2 Border Irregularity Descriptors

Border irregularity constitutes the second criterion of the ABCD rule for clinical evaluation of melanocytic lesions. While benign lesions tend to have smooth and continuous contours, melanomas frequently exhibit irregular borders,

characterized by indentations, protrusions, and abrupt changes along the contour [7].

To computationally represent this criterion, the discrete derivative of the radial signature obtained in the previous step was used. Since the radial signature describes the distance between the centroid of the lesion and its border as a function of the polar angle, its derivative allows quantifying local variations in the contour. Considering that the radial signature represents a closed curve, the derivative was calculated in a circular manner, including the difference between the last and first sample, ensuring the continuity of the signal along the entire perimeter of the lesion [13].

Initially, the Average Border Irregularity Index (IB_1) was determined, which corresponds to the average of the magnitudes of the differences between consecutive radial samples. This descriptor provides an estimate of the average intensity of contour variations and was calculated according to Eq. (3).

$$IB_1 = \frac{1}{N}\sum_{i=0}^{N}|\Delta r_i| \quad (3)$$

where N represents the total number of samples of the radial signature and Δr_i corresponds to the difference between two consecutive radial distances. High IB_1 values indicate greater average intensity of edge variations, being associated with geometrically more irregular contours [3], [5].

In addition to the average intensity of variations, the Irregularity Dispersion Index (IB_2) was calculated, obtained by the standard deviation of the discrete derivative of the radial signature, as presented in Eq. (4).

$$IB_2 = std(\Delta r_i) \quad (4)$$

This descriptor quantifies the dispersion of radial variations along the contour. Reduced values indicate relatively uniform border behavior, while high values show greater geometric heterogeneity, a characteristic frequently observed in melanoma lesions [3], [5].

Although the IB_1 and IB_2 descriptors characterize the overall behavior of the border, they do not fully represent localized irregularities. Therefore, automatic peak detection was performed on the absolute derivative of the radial signature using the findpeaks function of MATLAB®, considering an adaptive threshold defined by the sum of the mean and one standard deviation of the absolute derivative, in addition to a minimum distance between consecutive peaks to avoid multiple detections of the same irregularity.

From this analysis, three additional descriptors were extracted: number of peaks, corresponding to the number of significant irregularities present along the border; average peak amplitude, which represents the average intensity of the detected irregularities; and maximum peak amplitude, corresponding to the greatest local variation observed in the lesion contour.

Thus, the proposed methodology combines global descriptors, represented by IB_1 and IB_2 , with local descriptors, obtained by peak detection, providing a more complete quantitative representation of the geometric complexity of the lesion border.

C.3 Pigmentation Descriptors

Color variation constitutes the third criterion of the ABCD rule for evaluating melanocytic lesions. While benign nevi tend to have relatively homogeneous pigmentation, melanomas frequently exhibit multiple shades distributed irregularly, a characteristic associated with the heterogeneity of melanin concentration and structures present in the lesion [7].

In this work, this criterion was analyzed by means of the distribution of luminosity within the segmented region. For this, the CIELAB color space was used, which separates the luminosity component L from the chromatic components, allowing the analysis of intensity variations with less influence from color information [5], [13]. After segmentation, only the pixels belonging to the binary mask of the lesion were considered, excluding the adjacent skin.

To characterize the heterogeneity of pigmentation, the L-channel values were divided into five equally spaced intervals between the minimum and maximum values observed in the lesion itself. Each pixel was associated with one of these intervals, producing a pigmentation map composed of up to five luminosity levels.

The first descriptor, called the Pigmentation Level Index **(IC_1)**, corresponds to the number of levels actually present in the lesion. To reduce the influence of small regions resulting from noise, a level was considered present only when it represented at least 5% of the total segmented area. Thus, IC_1 expresses the diversity of luminosity levels with a relevant participation in the composition of the lesion.

In addition to the number of levels, the Luminosity Dispersion Index **(IC_2)** was calculated, corresponding to the standard deviation of the L-channel values within the lesion, as shown in Eq. (5).

$$IC_2 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(L_i - \bar{L})^2} \quad (5)$$

where N represents the number of pixels belonging to the lesion, L_i corresponds to the luminosity of the i-th pixel, and $\bar{L}$ represents the average luminosity of the segmented region. High IC_2 values indicate greater dispersion of luminosity levels and, consequently, greater heterogeneity of pigmentation, while reduced values characterize lesions with a more uniform distribution.

Finally, the Luminosity Amplitude IC_3 was calculated, defined by the difference between the maximum and minimum values of the L channel inside the lesion, as shown in Eq. (6).

$$IC_3 = L_{max} - L_{min} \quad (6)$$

This descriptor represents the maximum contrast of pigmentation. High values indicate a greater difference between the lighter and darker regions of the lesion.

The three descriptors provide complementary information about pigmentation: IC_1 quantifies the number of relevant lightness levels, IC_2 measures the dispersion of these values, and IC_3 represents the maximum contrast observed. Subsequently, the descriptors were individually evaluated for

their ability to differentiate melanoma and non-melanoma lesions.

D. Exclusion of Criterion D from the ABCD Rule

Criterion D of the ABCD rule corresponds to the lesion diameter, traditionally considered suspicious when greater than 6 mm [7]. However, its use in computer systems depends on the spatial calibration of the image, requiring reliable information on resolution and scale for converting measurements into physical units. As the database used in this study gathers images obtained under different acquisition conditions, it was not possible to guarantee the standardization of this information, compromising the reliability of this parameter.

In addition to the limitations related to image acquisition, recent studies demonstrate that melanomas can be identified in early stages, even when they have a diameter smaller than the classic limit of 6 mm [2], [7]. Thus, the use of this criterion could reduce the sensitivity of the method for identifying small lesions, contradicting the objective of favoring early diagnosis.

Due to these limitations, it was decided not to use criterion D in the proposed methodology, prioritizing quantitative descriptors related to the geometry of the lesion and the distribution of its pigmentation. This decision makes the method less dependent on image acquisition conditions and potentially more suitable for analyzing lesions in different databases, regardless of their absolute size.

E. Evaluation of Descriptors

After extracting the computational descriptors, each metric was individually evaluated for its ability to differentiate melanoma and non-melanoma lesions. For this step, an evaluation set composed of 20 images was used, 10 melanomas and 10 non-melanomas, selected from the database described in Section II-A.

Initially, the descriptors for asymmetry (IA), border irregularity (IB_1 , IB_2 and peak-based descriptors) and pigmentation (IC_1, IC_2 and IC_3) were calculated for each image. Then, descriptive statistics were obtained for each class, allowing for the analysis of the individual behavior of each metric and the identification of differences between melanoma and non-melanoma lesions.

To assess the discriminatory power of the descriptors, each metric was analyzed individually using Leave-One-Out cross-validation (LOOCV). In each iteration, one image was reserved for testing, while the remaining images (N−1) were used to calculate the descriptor averages in each class and establish the separation threshold between them. The reserved image was then classified based on this threshold. This procedure was repeated for all samples, preventing the evaluated image from being used in defining the decision criterion itself.

As a performance measure, accuracy was used, defined as the ratio between the number of correct classifications and the total number of images evaluated, as presented in Eq. (7).

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \quad (7)$$

where VP represents the number of true positives, VN the number of true negatives, FP the number of false positives, and FN the number of false negatives. In addition to accuracy, sensitivity, specificity, precision, F1-score, and area under the ROC curve (AUC) were calculated. While accuracy represents the overall proportion of correct classifications, precision indicates the proportion of cases classified as melanoma that actually belonged to that class. The AUC was used to assess the discriminatory capacity of the descriptors considering different decision thresholds.

The significance of the differences between the melanoma and non-melanoma groups was evaluated using the Wilcoxon rank-sum test (Mann–Whitney U), with Bonferroni correction for multiple comparisons. This analysis allowed the identification of the descriptors with the greatest capacity to differentiate between the groups.

Finally, the most discriminatory descriptors for each criterion A, B, and C were combined using logistic regression, also evaluated by Leave-One-Out (LOOCV) cross-validation. This approach allowed us to compare the performance of the combined model with that obtained by the descriptors analyzed individually.

III. Results And Discussion

A. Evaluation of the Proposed Methodology

The proposed methodology was evaluated using a set of 20 dermatoscopic images, composed of 10 melanomas and 10 non-melanomas, processed entirely by the developed pipeline. As a representative example, the analysis of Image 5 (AUG_0_1114.jpeg), illustrated in Fig. 2, is presented, showing the main processing steps, including lesion segmentation, centroid determination, radial signature extraction, comparison between diametrically opposed rays, radial signature derivation, and pigmentation distribution analysis.

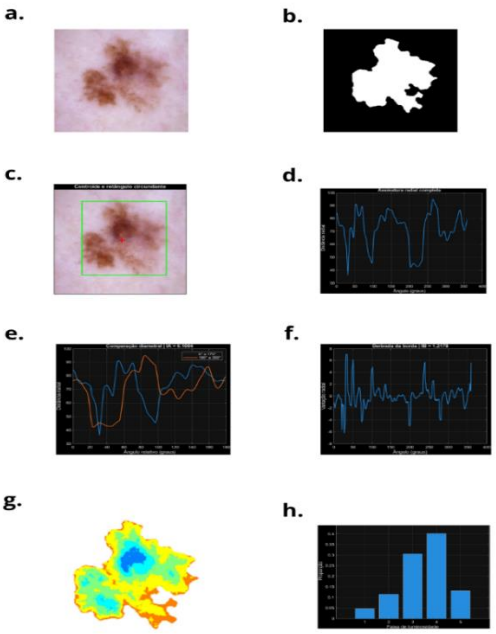


Fig. 2. Example of the execution of the proposed methodology for a melanoma image

Segmentation performed in the CIELAB color space using the k-means algorithm, complemented by morphological operations, showed satisfactory performance in most of the evaluated images. The resulting mask adequately preserved the lesion's contour, removing most of the artifacts present in the adjacent skin and allowing for the reliable extraction of geometric and pigmentation descriptors. For Image 5, an area of 16,346 pixels, a perimeter of 778.96 pixels, a circularity of 0.3385, a major axis of 164.77 pixels, a minor axis of 145.19 pixels, an eccentricity of 0.4728, and an orientation of 51.59° were obtained, indicating a lesion with significantly irregular geometry.

B. Evaluation of Computational Descriptors

B1. Asymmetry

The radial signature obtained for Image 5, shown in Fig. 2(d), showed significant oscillations throughout the 360°, reflecting geometric differences between opposite regions of the lesion. The comparison between diametrically opposite rays (Fig. 2(e)) resulted in an Asymmetry Index equal to IA = 0.1004, a value higher than those observed for images classified as non-melanoma. This behavior demonstrates the ability of the descriptor to objectively quantify the loss of symmetry of the lesion [9].

B2. Border Irregularity

The derivative of the radial signature, shown in Fig. 2(f), showed several abrupt oscillations along the contour of the lesion, indicating the presence of indentations and protrusions characteristic of melanomas. $IB_1 = 1.2178$, $IB_2 = 1.8963$, nine significant peaks, a mean amplitude of 5.46 pixels, and a maximum amplitude of 7.10 pixels were obtained. These results demonstrate that the combination of global and local descriptors provides a consistent quantitative representation of the geometric complexity of the border.

B3. Pigmentation

The analysis of the luminosity distribution was performed using the pigmentation map (Fig. 2(g)) and the luminosity histogram (Fig. 2(h)). Four pigmentation levels were identified, with a luminosity standard deviation of 10.45 and a luminosity amplitude of 54.80, indicating high heterogeneity in lesion pigmentation. Although these descriptors contributed to its characterization, they showed less discriminatory capacity when compared to geometric descriptors.

C. Comparison between Melanoma and Non-Melanoma

To illustrate the behavior of the proposed methodology, a comparison was made between Image 5 (melanoma) and Image 15 (non-melanoma), both subjected to the same processing pipeline. Fig. 3 shows the complete set of processing steps obtained for the non-melanoma lesion, allowing a visual comparison of the algorithm's behavior in the two classes.

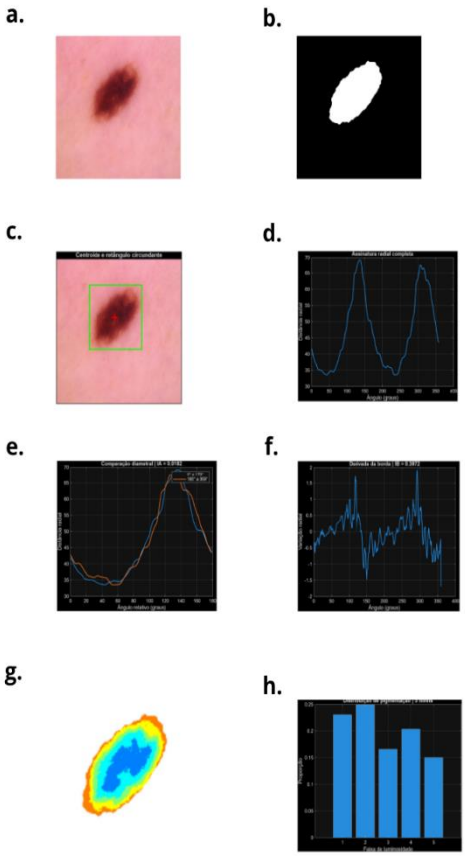


Fig. 3: Example of Operational Results for Non-Melanoma

Visual comparison between the original images (Fig. 4) reveals important morphological differences between the two lesions. While melanoma presents an irregular shape, asymmetrical contours, and heterogeneous pigmentation distribution, the non-melanoma lesion presents a more regular geometry, smooth borders, and greater structural uniformity.

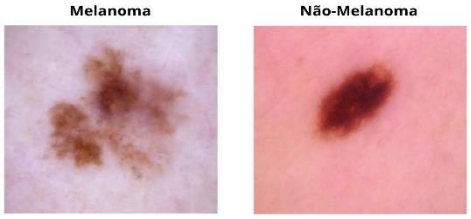


Fig. 4: Melanoma and Non-Melanoma (original images)

The observed differences are also evident in the radial signature (Fig. 5) and in the comparison between diametrically opposed rays (Fig. 6). Melanoma presented IA = 0.1004, while the non-melanoma lesion presented IA = 0.0179, a value approximately 5.6 times lower, indicating greater geometric symmetry of the benign lesion.

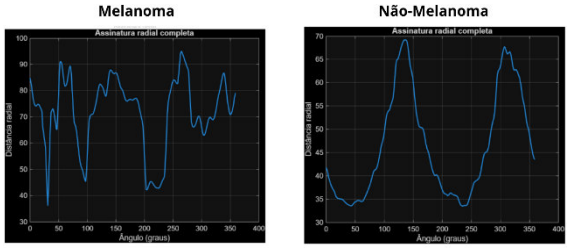


Fig. 5: Melanoma and Non-Melanoma (Signature)

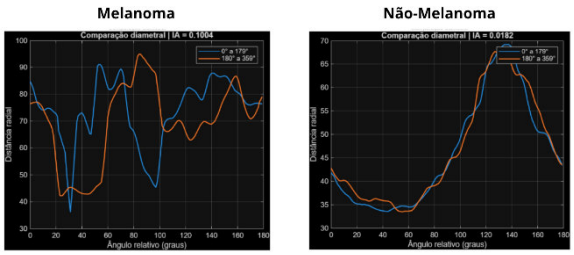


Fig. 6: Melanoma and Non-Melanoma (IA)

Similar differences were observed in border irregularity. Fig. 7 presents a comparison between the radial signature derivatives for the two lesions. The melanoma showed higher IB_1 and IB_2 values, mean amplitude, and maximum amplitude of the peaks, indicating greater geometric complexity of the contour. In contrast, the non-melanoma lesion showed less intense oscillations and a more regular contour.

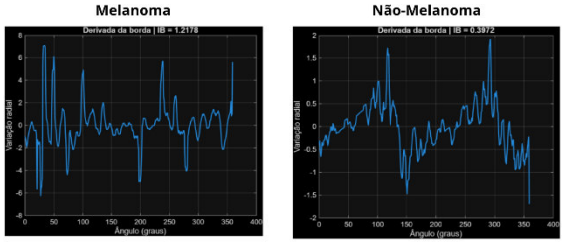


Fig. 7: Melanoma and Non-Melanoma (IN)

D. Performance of the Proposed Descriptors

Table I presents the individual performance of the computational descriptors under Leave-One-Out Validation (LOOCV), considering accuracy, sensitivity, specificity, precision, F1-score, and area under the ROC curve (AUC). The Skewness Index (SI) showed the best overall performance, with accuracy, sensitivity, specificity, and precision of 90%, an F1-score of 0.900, and an AUC of 0.990, demonstrating greater discriminatory capacity among the evaluated descriptors.

TABLE I. PERFORMANCE OF COMPUTATIONAL DESCRIPTORS EXTRACTED FROM ABC CRITERIA

Descritor	***Ac. (%)***	***Sen. (%)***	***Esp. (%)***	***Prec. (%)***	***F1***	***AUC***
A – IA	90	90	90	90	0,900	0,990
B – IB1	70	60	80	75	0,667	0,870
B – IB2	75	60	90	87,5	0,706	0,870
B – Nº picos	65	80	50	61,5	0,696	0,715
B – Average amplitude	60	30	90	75	0,429	0,890
B – Maximum amplitude	70	50	90	83,3	0,625	0,890
C – number of levels	80	90	70	75	0,818	0,800
C – Standard deviation	15	10	20	11,1	0,105	0,330
C – Amplitude	70	70	70	70	0,700	0,810

Among the border-related descriptors, the greatest variation in performance metrics was observed. Accuracies ranged from 60% to 75%, while AUC values ranged from 0.715 to 0.890. Some descriptors showed high specificity associated with lower sensitivity, such as Average Amplitude and Maximum Amplitude of Peaks, indicating a greater ability to recognize non-melanoma samples than melanoma samples. The Number of Peaks showed distinct behavior, with a sensitivity of 80% and a specificity of 50%, demonstrating greater identification of the melanoma class, but with a higher occurrence of false positives.

For the pigmentation criterion, the Number of Levels showed the best performance, achieving an accuracy of 80%, sensitivity of 90%, specificity of 70%, and an F1-score of 0.818. In contrast, the descriptor based on standard deviation showed the lowest performance among the metrics evaluated, with an accuracy of 15% and an AUC of 0.330. These results demonstrate that descriptors associated with the same ABC criterion can exhibit distinct behaviors in differentiating between classes, reinforcing the importance of their individual evaluation before combining them in a classification model.

Statistical comparison between melanoma and non-melanoma groups was performed using the Wilcoxon rank-sum test (Mann–Whitney U), with Bonferroni correction for the nine descriptors evaluated (adjusted $\alpha = 0.0056$). After correction, statistically significant differences were observed for Asymmetry ($p = 0.0002$), Mean Peak Amplitude ($p = 0.0036$), and Maximum Peak Amplitude ($p = 0.0036$). The other descriptors did not show statistically significant differences considering the adjusted significance level, although some showed relevant discriminatory performance in the classification metrics.

The superiority observed for the asymmetry descriptor may be related to the use of the radial signature, which allows quantifying differences between diametrically opposite regions of the lesion. The border descriptors, in turn, represent different aspects of contour irregularity, while the pigmentation descriptors characterize variations in the distribution of intensity levels. The results suggest, therefore, that these characteristics make distinct contributions to the differentiation between melanomas and non-melanomas in the evaluated sample.

Additionally, the combination of one descriptor from each ABC criterion—Asymmetry, Maximum Amplitude of Peaks, and Number of Pigmentation Levels—was evaluated using logistic regression with LOOCV validation. The combined model showed 90% accuracy, 90% sensitivity, 90% specificity, and an F1-score of 0.900, with nine true positives, nine true negatives, one false positive, and one false negative. However, the AUC obtained by the combined model was 0.930, lower than that presented by the Asymmetry descriptor alone (AUC = 0.990). Thus, the combination of the three criteria did not provide a performance gain compared to the best individual descriptor in this sample.

This behavior may be related to the existence of partially redundant information among the selected attributes. Asymmetry showed a strong correlation with the Maximum Peak Amplitude ($r = 0.843$) and a moderate inverse correlation with the Number of Pigmentation Levels ($r = -0.567$), indicating overlap between some of the information represented by the descriptors. Thus, the inclusion of multiple attributes does not necessarily result in increased discriminatory power when they have a high correlation.

In general, the results demonstrate that the proposed methodology made it possible to computationally quantify characteristics related to the evaluated ABC criteria, especially the Asymmetry Index. However, considering the small number of images used, the results should be interpreted as preliminary and restricted to the analyzed sample, and evaluation in larger and independent sets is necessary before inferring the generalizability of the method or its clinical applicability.

IV. Conclusion

This study presented a methodology based on Digital Image Processing for the computational quantification of criteria A (asymmetry), B (border), and C (color) of the ABCD rule, employing geometric and pigmentation descriptors automatically extracted from dermatoscopic images. The methodology integrated lesion segmentation techniques, radial signature extraction, and calculation of quantitative metrics, allowing the computational representation of characteristics related to the ABC criteria through interpretable descriptors with low computational cost.

The results indicated differences in the discriminatory power of the evaluated descriptors. The Asymmetry Index (AI) showed the best individual performance, with 90% accuracy and an AUC of 0.990, in addition to a statistically significant difference between the melanoma and non-melanoma groups after Bonferroni correction. Among the other attributes, the Mean Amplitude and Maximum Amplitude of the Peaks also showed statistically significant differences between the groups. These findings highlight, in the analyzed sample, the contribution of the descriptors associated with asymmetry and border irregularity to the differentiation between the classes.

The combination of descriptors from criteria A, B, and C through logistic regression showed 90% accuracy, 90% sensitivity, 90% specificity, and an AUC of 0.930. Although the overall performance was satisfactory in the evaluated sample, the combination did not provide discriminatory gain compared to the Asymmetry Index alone, which showed a higher AUC. The correlation observed between some of the attributes suggests the presence of partially redundant information, which may have limited the benefit resulting from the integration of the descriptors.

As a main limitation, the reduced number of images used in the evaluation (n = 20) stands out, which restricts the generalization of the results and the robustness of the statistical analyses. Therefore, the findings should be considered preliminary and restricted to the investigated sample, not allowing us to establish, at this stage, the clinical applicability of the methodology.

As future work, it is proposed to evaluate the approach in a substantially larger and more diverse sample, with adequate separation between development and test sets. Considering the performance of the Asymmetry Index, a hierarchical classification strategy could be investigated, in which this descriptor is used as an initial criterion and the other attributes contribute to the classification of ambiguous cases. A multiparametric model based on the weighting of descriptors could also be evaluated, with coefficients estimated from a larger sample, allowing for a more robust investigation of the benefit of integrating criteria A, B, and C.

Acknowledgment

The authors thank Laboratory for Medical Signal and Image Acquisition and Processing - LAPSIM Laboratory of the Biomedical Engineering program at the Federal University of Rio Grande do Norte (UFRN) for the technical support, infrastructure, and assistance provided for the development of this work.

References

[1] Brasil. Ministério da Saúde. Câncer de pele. Brasília, DF: Ministério da Saúde, [s.d.]. Disponivel em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-peleABBAS, Q. et al. Segmentation of Melanoma Skin Lesion Using Perceptual Color Difference Saliency with Morphological Analysis. [S. l.: s. n.], 2012.

[2] Rasul, Md. Fazle et al. A comparative study of neural network architectures for lesion segmentation and melanoma detection. In: IEEE Region 10 Symposium (TENSYMP), 2020, Dhaka. Proceedings [...]. Dhaka: IEEE, 2020.

[3] Udayakumar, E. et al. A review on diagnosis of malignant melanoma from benign lesion by using BPNN and ABCD rule parameters. International Research Journal of Pharmacy, v. 9, n. 10, 2018.

[4] Do Carmo, Natalia Camillo; MARI, João Fernando. A comparative study of convolutional neural networks for classification of pigmented skin lesions. In: Workshop de Visão Computacional (WVC), 17. , 2021, Online.Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 171-176. DOI: https://doi.org/10.5753/wvc.2021.18909.

[5] Olugbara, Oludayo O.; Taiwo, Tunmike B.; Heukelman, Delene. Segmentation of melanoma skin lesion using perceptual color difference saliency with morphological analysis. Mathematical Problems in Engineering, v. 2018, art. 1524286, 2018. DOI: https://doi.org/10.1155/2018/1524286

[6] Frutuoso, Rômulo Lopes; Santos, Jonas Rodrigues Vieira; SIQUEIRA, Robson S.. Reconhecimento de câncer de pele do tipo melanoma. In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), 13. , 2013, Maceió/AL. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2013 . p. 1198-1208. ISSN 2763-8952.

[7] Kittler, Harald. The evolution of the clinical, dermoscopic and pathologic diagnosis of melanoma. In: Argenziano, Giuseppe (ed.). Melanoma Today. Mattioli 1885, [s.d.].

[8] Oliveira, R. B. et al. Computational methods for the image segmentation of pigmented skin lesions: A review. Computer Methods and Programs in Biomedicine, v. 131, p. 127-141, 2016.

[9] Garnavi, R.; Aldeen, M.; Bailey, J. Computer-Aided Diagnosis of Melanoma Using Border-and Wavelet-Based Texture Analysis. IEEE Transactions on Information Technology in Biomedicine, v. 16, n. 6, p. 1239-1252, nov. 2012.

[10] Beuren, A. T. et al. Skin melanoma segmentation by morphological approach. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2012, Chennai. Proceedings... Chennai: ACM, 2012. p. 972-978.

[11] Yao, Yuan; Lilhore, Umesh Kumar; Simaiya, Sarita; Aldossary, Sultan M.; Tekeste, Lidia Gosy; Ghith, Ehab; Abdallah, Hanaa A. Advancing skin cancer diagnosis with deep learning and attention mechanisms. Scientific Reports, v. 16, p. 3610, 2026. DOI: https://doi.org/10.1038/s41598-025-33597-0

[12] Araújo, Rafael Luz; V. E Silva, Romuere R.; Ferreira, Jonnison Lima; Carvalho, Nonato R. de S. . Estudo Comparativo para Segmentação de Melanoma em Imagens de Lesões de Pele. In: Escola Regional de Computação Aplicada à Saúde (ERCAS), 7. , 2019, Teresina. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019 . p. 294-299.

[13] Gonzalez, Rafael C.; Woods, Richard E. Processamento digital de imagens. 3. ed. São Paulo: Pearson Prentice Hall, 2010.

[14] Chacon, Germano; Gastardelli, Elton; Moraes, Fernanda D.; Oliveira, Gabriel; Albuquerque, Márcio P. de; Albuquerque, Marcelo Portes de. Aplicação da técnica de momentos invariantes no reconhecimento de padrões em imagens digitais. Notas Técnicas, Rio de Janeiro, v. 1, n. 2, p. 1–11, 2011. DOI: https://doi.org/10.7437/NT2236-7640.2011.02.002

[15] Scarlat, Dr. Melanoma Dataset. Kaggle, [s.d.]. Disponível em: https://www.kaggle.com/datasets/drscarlat/melanoma.

Ferramenta educacional para detecção de displasia epitelial oral a partir de casos de leucoplasia

Lucas Fiamoncini de Souza
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brasil
lucas.fiamoncini00@gmail.com

Márcio Holsbach Costa
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brasil
marcio.costa@ufsc.br

Gilberto de Souza Melo
Departamento de Saúde Pública
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brasil
melo.gilberto@hotmail.com

Elena Riet Correa Rivero
Departamento de Patologia
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brasil
riet.elena@gmail.com

***Resumo*—A presença de displasia epitelial em desordens orais potencialmente malignas, como a leucoplasia, está associada a um risco aumentado de desenvolvimento do carcinoma epidermoide oral. Nesse contexto, o diagnóstico precoce dessas lesões é fundamental para o sucesso terapêutico e para a redução da mortalidade associada à doença. Neste trabalho, é desenvolvido um sistema automatizado para detecção de displasia oral em casos de leucoplasia que integra imagens histopatológicas e dados clínicos em duas etapas. Primeiramente, uma rede ResNet50 classifica as imagens quanto à presença de displasia, e então os resultados são combinados com dados clínicos por um classificador SVM, gerando a decisão final. O sistema obteve um F1-*score* médio de 0,946 no conjunto de teste. A partir desse classificador, foi desenvolvida uma ferramenta educacional com o objetivo de auxiliar estudantes de patologia bucal na visualização dos resultados e na compreensão das regiões relevantes nas imagens, promovendo um aprendizado mais interpretável e alinhado ao uso da inteligência artificial na saúde.**

***Palavras-Chave*—Displasia epitelial oral, aprendizado profundo, aprendizado de máquina, ferramenta educacional, inteligência artificial.**

***Abstract*—The presence of epithelial dysplasia in potentially malignant oral disorders, such as leukoplakia, is associated with an increased risk of progression to oral squamous cell carcinoma. Therefore, early diagnosis of these disorders is crucial for improving treatment outcomes and reducing disease-related mortality. In this work an automated system for the detection of oral dysplasia in leukoplakia cases is developed by integrating histopathological images and clinical data in two stages. First, a ResNet50 network classifies the images regarding the presence of dysplasia; subsequently, these results are combined with clinical data using an SVM classifier, generating the final decision. The system obtained an average F1-score of 0.946 on the test set. From this classifier, an educational tool was developed to assist oral pathology students in vizualizing outcomes and understanding relevant regions within the images, promoting a more interpretable learning process aligned with the use of artificial intelligence in healthcare.**

***Index Terms*—Oral epithelial dysplasia, deep learning, machine learning, educational tool, artificial intelligence.**

I. Introdução

O carcinoma epidermoide é a principal neoplasia maligna que ocorre na boca, sendo o tabagismo e o etilismo os principais fatores etiológicos para o seu desenvolvimento [1]. De modo geral, o câncer de boca é precedido por um conjunto de alterações da mucosa oral, que se apresentam clinicamente como desordens orais potencialmente malignas, tais como a leucoplasia e a eritroplasia [2]. Essas lesões podem apresentar diferentes características histopatológicas, tais como hiperceratose, acantose e displasia epitelial, sendo a presença de displasia usualmente associada a um maior potencial de malignização [3].

Nos últimos anos, diversos estudos têm investigado a aplicação de técnicas de inteligência artificial na detecção e classificação da displasia epitelial oral [4]. Apesar dos avanços alcançados, a ausência de mecanismos de explicabilidade — sobretudo em modelos baseados em aprendizado profundo — permanece como uma limitação significativa e um campo em aberto para investigações. Nesse cenário, [5] destaca-se ao empregar o algoritmo Grad-CAM [6] como ferramenta de interpretação dos modelos preditivos, permitindo a visualização das regiões mais relevantes das imagens histopatológicas para a detecção do carcinoma epidermoide; um tipo de câncer frequentemente precedido por displasia epitelial e que apresenta alterações morfológicas importantes. Além disso, a maior parte dos estudos com inteligência artificial (IA) em patologia bucal está centrada no suporte ao diagnóstico clínico, sendo pouco explorado seu potencial no contexto educacional [7]. Observa-se, assim, uma lacuna importante no aproveitamento dessas tecnologias como instrumentos pedagógicos. Transformar soluções desenvolvidas inicialmente para diagnóstico em ambientes de aprendizado pode facilitar a compreensão de conceitos complexos e aumentar o interesse dos estudantes por disciplinas fundamentais como a patologia bucal. A construção de plataformas educacionais baseadas

em IA também pode democratizar o acesso a conteúdos de alta qualidade, especialmente em contextos com escassez de profissionais especializados ou infraestrutura laboratorial limitada.

O objetivo deste trabalho é viabilizar o desenvolvimento de uma ferramenta educacional voltada à formação de estudantes de patologia bucal. Nesse contexto, resultados promissores em termos de desempenho acadêmico e engajamento discente têm sido observados [8] [9]. As principais contribuições deste trabalho são: (i) desenvolvimento de um sistema de aprendizado de máquina para detecção de displasia epitelial oral a partir de imagens histológicas de casos de leucoplasia e dados complementares dos pacientes; (ii) utilização de técnica visual de explicabilidade do processo de decisão do modelo de aprendizado profundo; (iii) comparação dos mapas de explicabilidade gerados pelo modelo com a interpretação de patologistas especialistas e (iv) desenvolvimento de uma ferramenta educacional *online* que utiliza o sistema proposto neste trabalho.

A seção II apresenta uma fundamentação teórica a respeito de conceitos de patologia bucal, patologia digital e trabalhos relacionados à utilização de inteligência artificial nessas áreas. Na seção III são detalhados os materiais e métodos utilizados no desenvolvimento deste trabalho. A seção IV apresenta os resultados obtidos. Por fim, as seções V e VI apresentam a discussão e a conclusão do trabalho.

II. Fundamentação teórica

A leucoplasia oral é definida como uma lesão potencialmente maligna que se apresenta clinicamente como uma placa branca e que não pode ser caracterizada clinicamente como qualquer outra condição conhecida. Trata-se, portanto, de um diagnóstico exclusivamente clínico, baseado na exclusão de outras condições bucais, com características semelhantes, na mucosa oral [10]. O procedimento padrão para a avaliação da leucoplasia consiste na realização de biópsia da região afetada, seguida de análise histopatológica. A Figura 1 fornece um exemplo de uma fotografia de uma lâmina histológica; em que "A" representa a camada de ceratina, "B" o tecido epitelial e "C" o tecido conjuntivo. Com base na análise da imagem, a lesão pode ser classificada como com ou sem displasia epitelial.

A displasia epitelial oral (DEO) refere-se a um conjunto de alterações arquiteturais e citológicas do epitélio. A gravidade da displasia epitelial pode ser classificada, de acordo com os critérios da Organização Mundial da Saúde [11], com base na extensão das alterações no epitélio, em: (i) leve, quando as alterações estão restritas ao terço inferior; (ii) moderada, quando as alterações estão presentes até o terço médio; e (iii) severa, quando essas alterações ultrapassam os dois terços do epitélio. Entretanto, este procedimento de detecção de displasia e classificação em níveis (leve, moderada e severa) é altamente subjetivo, apresentando elevada variabilidade interobservador [4] [12] [13].

O estudo mais recente relacionado ao tema de uso da inteligência artificial para detecção de DEO é o de [14], que apresentou desempenho superior na detecção de displasia ao incorporar dados clínicos adicionais dos pacientes após uma etapa inicial de análise baseada em imagens histológicas. Embora proponha um sistema robusto e com baixo custo computacional, sua principal limitação reside na necessidade de segmentação manual das regiões do tecido epitelial para a obtenção dos resultados. Ademais, o trabalho não fornece informações visuais que evidenciem as regiões das imagens associadas à presença de displasia, o que limita a interpretabilidade do modelo proposto.

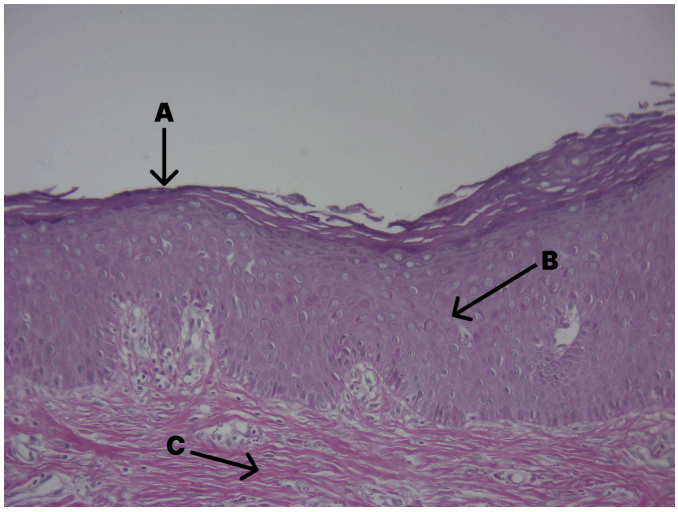


Fig. 1. Exemplo de uma fotografia com ampliação de 200 vezes de uma lâmina histológica corada com hematoxilina e eosina referente a uma lesão bucal; em que "A" indica a camada de ceratina, "B" o tecido epitelial e "C" o tecido conjuntivo.

III. Materiais e métodos

Esta seção apresenta o banco de dados e os métodos utilizados no desenvolvimento do modelo de detecção de displasia epitelial oral.

A. Banco de dados

O presente trabalho foi aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos da UFSC (Plataforma Brasil, CAAE 15025319.3. 0000.0121). O banco de dados utilizado consiste em imagens histopatológicas e dados complementares de casos de leucoplasia cadastrados no Biobanco do Laboratório de Patologia Bucal da UFSC (CONEP B-051; Processo Nº. 25000.237810/2014-54).

O banco de dados utilizado neste estudo é composto por 114 casos clínicos de pacientes diagnosticados com leucoplasia, totalizando 406 imagens histopatológicas com resolução de 2584×1936 *pixels*. As imagens foram obtidas a partir de uma ampliação de 200 vezes de lâminas coradas com hematoxilina e eosina, cujas seções representativas foram digitalizadas por meio de um microscópio biológico trinocular da marca Novel BM2000 com câmera para captura de imagens marca Tucsen ISH500 de 5.0 MP acoplada; constituindo as imagens do banco de dados. Entre os casos de leucoplasia oral, 59 apresentam DEO, correspondendo a 193 imagens, enquanto os 55 casos

restantes não apresentam displasia, totalizando 213 imagens. Os casos que não apresentam displasia foram diagnosticados como hiperceratose e acantose. Como o objetivo do trabalho é a detecção de displasia, as classes utilizadas foram: displásico e não displásico.

O estudo em [14] demonstrou que a incorporação de dados complementares dos pacientes pode aumentar significativamente o desempenho de modelos de aprendizado de máquina e de aprendizado profundo na detecção de displasia epitelial. Com base nesse resultado, o banco de dados utilizado neste trabalho inclui, além das imagens histopatológicas, informações clínicas complementares dos pacientes, a saber: característica clínica da lesão, localização anatômica da lesão, idade, histórico de tabagismo e de etilismo. Esses atributos foram selecionados por dois patologistas orais com base em sua relevância diagnóstica, sendo apontados como os fatores de risco com maior influência na progressão do carcinoma epidermoide de boca [14]. Para viabilizar a integração desses dados ao modelo, os atributos foram convertidos em variáveis binárias, indicando a presença ou ausência de fatores de risco associados. A transformação foi realizada conforme a metodologia proposta por [14], descrita a seguir:

- Característica clínica da lesão: valor "1" atribuído para lesões heterogêneas; "0" para as demais;
- Localização da lesão: valor "1" para lesões localizadas na borda lateral da língua, ventre da língua ou assoalho bucal; "0" para outras localizações;
- Idade do paciente: valor "1" para pacientes com idade superior a 40 anos; "0" caso contrário;
- Histórico de tabagismo: valor "1" para pacientes fumantes ou ex-fumantes; "0" para não fumantes;
- Histórico de etilismo: valor "1" para pacientes alcoolistas ou ex-alcoolistas; "0" para não alcoolistas.

B. Metodologia

Este trabalho propõe um modelo para a detecção de displasia epitelial oral que integra informações em dois níveis: imagens histopatológicas e dados complementares dos pacientes. A Figura 2 ilustra o sistema proposto.

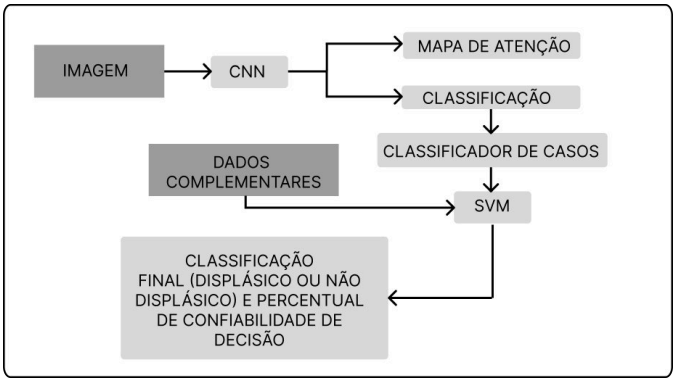


Fig. 2. Diagrama em blocos do sistema proposto.

Inicialmente, as imagens foram redimensionadas devido a limitações computacionais. A nova resolução adotada foi de 700 × 524 *pixels*, mantendo-se a mesma razão da dimensão original. Em sequência, o conjunto de dados foi particionado em subconjuntos de treinamento e teste (70% e 30%, respectivamente), assegurando que imagens pertencentes a um mesmo paciente fossem alocadas exclusivamente a um único subconjunto, evitando vazamento de informação. Para a etapa de validação, adotou-se a técnica de validação cruzada estratificada do tipo *K-Fold*, com $K = 5$.

Uma ResNet50 pré-treinada foi utilizada para ajuste fino na estratégia de aprendizado por transferência, utilizando as imagens do conjunto de treinamento. Os parâmetros de treinamento foram: 30 *epochs*, *batch size* de 16, *learning rate* de 10^{-5} e otimizador Adam. Adicionalmente, foi aplicada rotação das imagens como técnica de *data augmentation*. A partir da rede treinada, mapas de atenção foram obtidos por meio do algoritmo Grad-CAM [6]. A classificação dos casos foi realizada com base na mediana das probabilidades de predição individual de cada imagem (presença ou não de displasia) associada a um mesmo paciente. Ou seja, dado um paciente com *N* imagens, realiza-se a predição da presença de displasia em cada uma delas. Em seguida, calcula-se a mediana das probabilidades atribuídas. Essa probabilidade agregada foi então combinada com os dados clínicos complementares dos pacientes. Por fim, um classificador SVM foi empregado para a decisão final, com os seguintes parâmetros: $C = 100$, *kernel* RBF e *gamma* $= 10^{-1}$.

O procedimento de anotação das regiões de interesse (ROI) se deu através do *software Roboflow* [15]. Para as imagens de displasia, foram consideradas duas regiões: região displásica e restante da imagem (tecido epitelial não displásico, tecido conjuntivo e fundo da lâmina). Como mencionado previamente, a região displásica sempre estará contida na região do tecido epitelial. Após a demarcação das regiões, as anotações foram exportadas para um arquivo de texto contendo as coordenadas e classes das regiões anotadas. Os mapas de atenção foram gerados a partir dessas informações alocadas pelo patologista, de modo que regiões displásicas recebessem valor 1 e o restante das regiões recebesse valor 0. Dessa forma, os mapas de atenção foram gerados como máscaras binárias.

C. Métricas de desempenho

A partir da matriz de confusão, foram calculadas as métricas acurácia, sensibilidade, especificidade e F1-*score* [16]. Para a estimação dos intervalos de confiança de 95%, empregou-se a técnica de *bootstrapping*, que consiste em realizar amostragem com reposição a partir do conjunto de teste. Foram realizadas 5000 reamostragens para obter as estimativas dos intervalos de confiança associados às métricas de desempenho. A curva ROC, juntamente com a métrica AUC, também foram obtidas.

A análise comparativa dos mapas de atenção se deu apenas para as imagens do conjunto de teste diagnosticadas como displásicas, uma vez que a região de interesse (displasia) está presente apenas nessas imagens. Para essa etapa, utilizaram-se as máscaras binárias geradas a partir das anotações dos especialistas. A comparação entre os mapas de atenção gerados pelo modelo e as máscaras binárias de referência foi

realizada utilizando a métrica *Pointing Game* [17], que avalia a capacidade do mapa de atenção em destacar corretamente as regiões relevantes da imagem. Primeiro, identifica-se o ponto de maior ativação no mapa de atenção. Em seguida, verifica-se se esse ponto está localizado dentro da região anotada como relevante pelos especialistas. Se estiver, conta-se como um acerto; caso contrário, conta-se como um erro. A acurácia do *Pointing Game* é então calculada como a razão entre o número de acertos e o número de imagens avaliadas.

D. Mapas de atenção

O algoritmo Grad-CAM (do inglês – *Gradient-weighted Class Activation Mapping*) é uma técnica de visualização que permite interpretar o funcionamento interno de redes neurais convolucionais [6]. Ele gera mapas de ativação, a partir da última camada convolucional, que destacam as regiões da imagem que mais contribuíram para a decisão do modelo. Sendo assim, é possível produzir um mapa de calor que evidencia as áreas mais relevantes para a predição de determinada tarefa. O mapa de atenção gerado possui dimensão 17×22. De maneira a comparar os mapas gerados com as anotações dos especialistas, foi necessário redimensioná-los para o tamanho original da imagem utilizando interpolação. Neste trabalho, o algoritmo Grad-CAM foi utilizado como ferramenta de suporte à decisão, contribuindo para a explicabilidade do modelo de aprendizado profundo.

Devido à baixa resolução do mapa de atenção e o posterior redimensionamento para a mesma dimensão original das imagens, o método utilizado para obtenção dos mapas muitas vezes destaca regiões relativamente menores e incompletas de um objeto na imagem, podendo, também, falhar em detectar múltiplas ocorrências do mesmo objeto [18]. De maneira a avaliar a possível falha na identificação das regiões de interesse devido às limitações mencionadas, foram adotadas duas abordagens distintas para a comparação com as anotações dos especialistas. A primeira abordagem consiste em comparar o mapa de atenção redimensionado diretamente com as anotações, enquanto a segunda abordagem envolve a aplicação de uma dilatação morfológica nas anotações antes da comparação. A dilatação morfológica é uma operação que expande as regiões anotadas, possibilitando avaliar a vizinhança das regiões de interesse. Essa técnica é especialmente útil para compensar possíveis imprecisões no mapa de atenção devido à baixa resolução e ao redimensionamento. Para esse procedimento, foi utilizado um elemento estruturante quadrado de 7×7 *pixels*, buscando um equilíbrio entre a expansão das regiões anotadas e a preservação da forma original, de maneira que o incremento médio da área anotada ficasse em torno de 20%.

E. Ferramenta Educacional

Para viabilizar o uso educacional do modelo treinado, foi desenvolvida uma aplicação *web* interativa utilizando a biblioteca Gradio [19]. Essa ferramenta permite que estudantes e profissionais da área da saúde façam o *upload* de imagens histopatológicas e recebam uma predição quanto à presença de displasia epitelial, percentual de confiabilidade da decisão e mapas de atenção que destacam as regiões mais relevantes para a decisão do modelo. A aplicação foi disponibilizada na plataforma *Hugging Face Spaces* [20], o que garante fácil acesso sem necessidade de instalação local ou configuração técnica, tornando-a adequada tanto para uso em sala de aula quanto em ambientes de ensino a distância.

O fluxo geral de utilização da aplicação é simples e intuitivo: inicialmente, o usuário acessa a plataforma por meio de um navegador *web*; em seguida, realiza o carregamento da imagem histopatológica e fornece as informações complementares solicitadas. Após a submissão dos dados, o sistema executa automaticamente o pré-processamento da imagem, seguido da inferência utilizando o modelo previamente treinado. Como resultado, são exibidos simultaneamente a classificação predita, a estimativa de confiabilidade associada à decisão e os mapas de atenção gerados pela rede, permitindo ao usuário visualizar as regiões da imagem que mais contribuíram para a predição.

Do ponto de vista metodológico, a aplicação integra um modelo de aprendizado profundo previamente treinado e congelado, garantindo estabilidade durante o uso educacional. Todo o processamento é realizado de forma síncrona, possibilitando uma interação fluida com o usuário. A geração dos mapas de atenção é realizada sob demanda, sendo apresentada de maneira integrada à interface gráfica da aplicação, reforçando o caráter interpretável da ferramenta.

IV. Resultados

Os resultados do modelo proposto foram avaliados utilizando os dados dos pacientes pertencentes ao conjunto de teste. A Tabela I apresenta os resultados obtidos do modelo proposto, enquanto a Figura 3 ilustra a distribuição das métricas de desempenho em gráficos de violino. A Figura 4 ilustra a matriz de confusão e a Figura 5 a curva ROC juntamente com a medida AUC. A Figura 6 apresenta alguns exemplos de casos com a indicação do mapa de atenção da rede neural utilizada e a Figura 7 apresenta uma captura de tela da ferramenta educacional desenvolvida.

TABELA I
VALORES MÉDIOS DAS MÉTRICAS DE DESEMPENHO E INTERVALO DE CONFIANÇA 95%.

Métrica	Valor médio	Intervalo de confiança 95%
Acurácia	0,942	0,9 - 1
F1-*score*	0,946	0,889 - 1
Sensibilidade	0,899	0,8 - 1
Especificidade	1	1 - 1
AUC	0,990	0,96 - 1

A Figura 8 apresenta a métrica *Pointing Game*, responsável pela comparação entre os mapas de atenção gerados pelo modelo em relação às regiões de interesse anotadas pelos especialistas.

V. Discussão

A limitação de dados constitui um desafio recorrente no contexto da pesquisa em saúde [14]. Ademais, o processo de anotação de imagens histopatológicas demanda a atuação

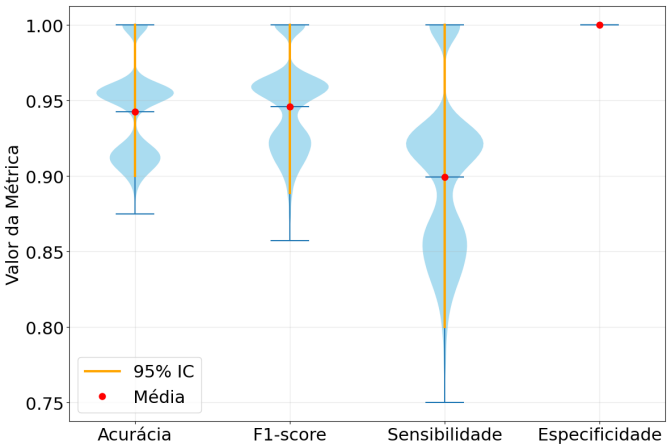


Fig. 3. Distribuição das métricas de desempenho em gráficos de violino.

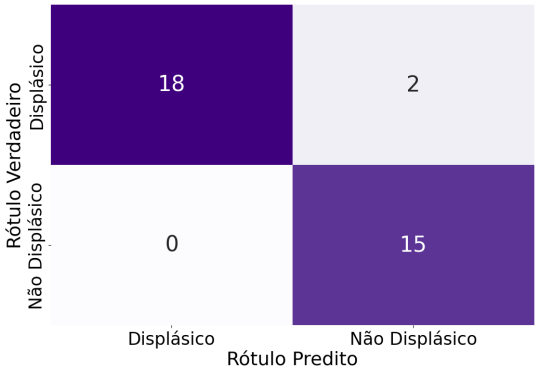


Fig. 4. Matriz de confusão.

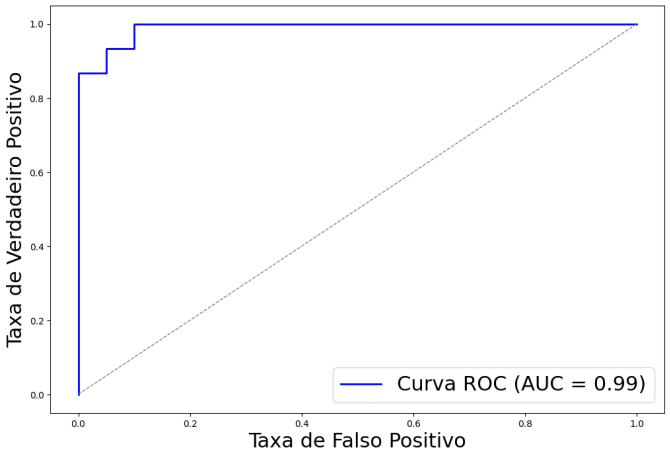


Fig. 5. Curva ROC e medida AUC.

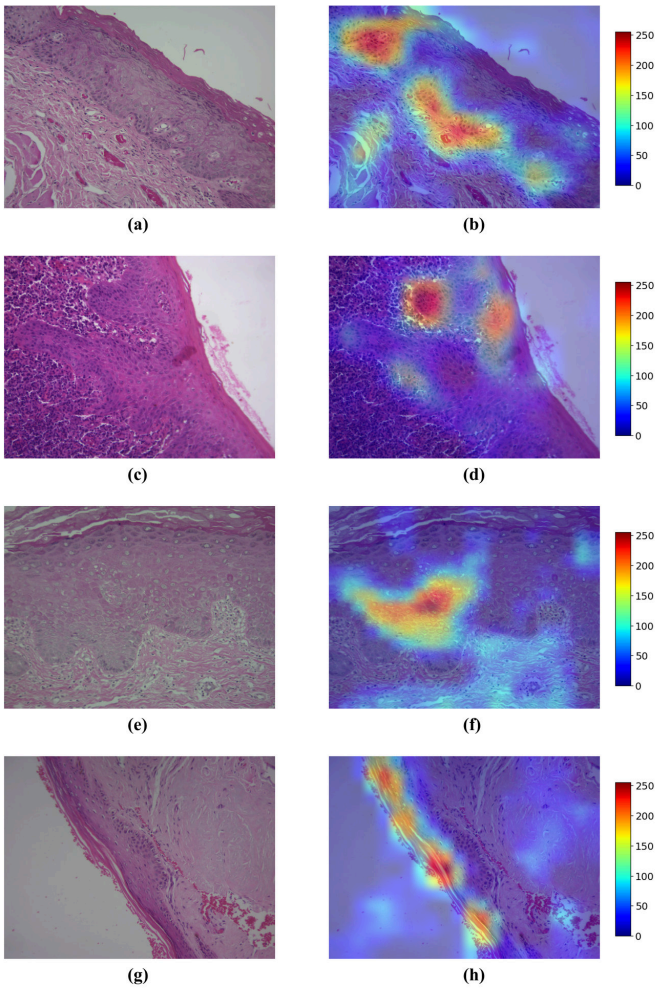


Fig. 6. Exemplos de imagem histológica e mapa de atenção provenientes do modelo pertencentes a casos: (a-b) corretamente classificado como displásico com percentual de confiabilidade de 100%; (c-d) corretamente classificado como displásico com percentual de confiabilidade de 99%; (e-f) erroneamente classificado como não displásico com percentual de confiabilidade de 67%; (g-h) erroneamente classificado como displásico com percentual de confiabilidade de 52%.

de patologistas qualificados, recurso frequentemente limitado. Esses fatores combinados dificultam a construção de conjuntos de dados extensos e diversificados. Nesse contexto, o banco de dados empregado neste estudo, embora apresente limitações quantitativas, representa uma contribuição relevante para o avanço da pesquisa na área. Tal afirmação fundamenta-se no fato de que o conjunto compilado de imagens e casos supera em dimensão outros bancos de dados utilizados em investigações recentes similares [14], os quais já se destacavam por ampliar substancialmente o volume de imagens em relação a trabalhos anteriores.

Este trabalho, embasado nos resultados de [14], diferencia-se ao endereçar a ausência de explicabilidade dos modelos, além de eliminar a necessidade de segmentação manual de recortes de imagem e também embarcar a solução proposta em uma aplicação didática.

Conforme apresentado na Tabela I, o modelo alcançou ele-

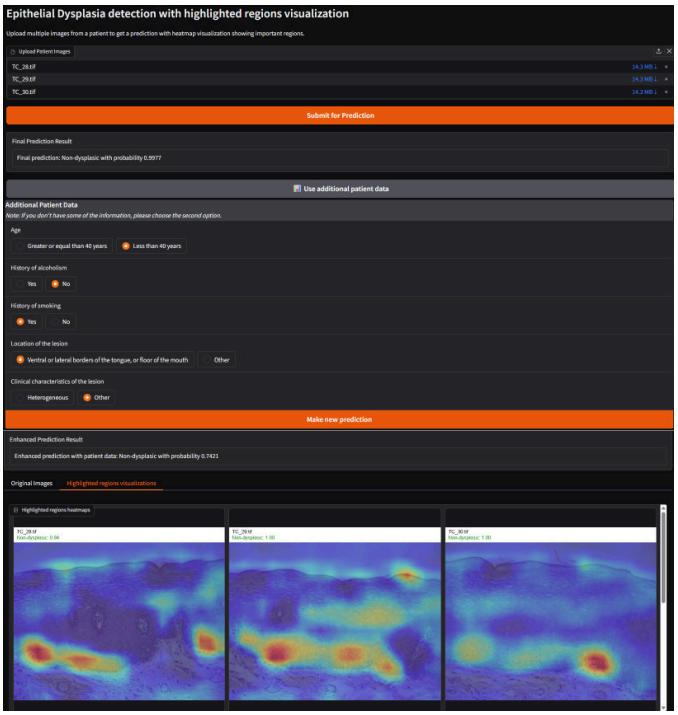


Fig. 7. Ferramenta educacional desenvolvida para auxílio na detecção de displasia epitelial oral (para melhor visualização utilize a ferramenta de lupa do *portable document format*).

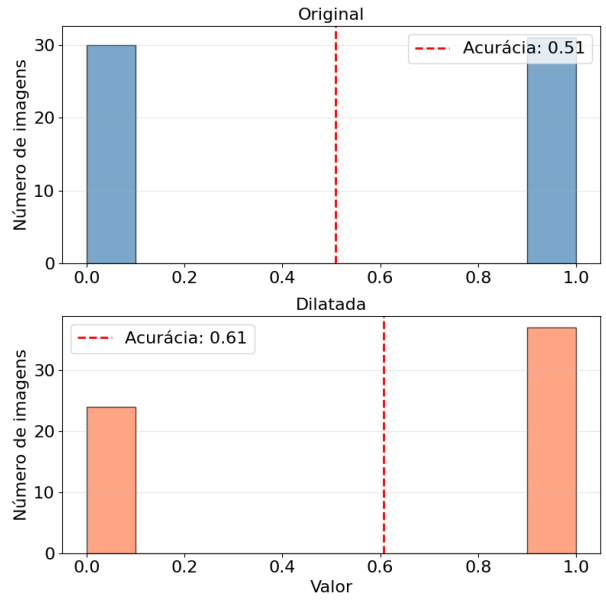


Fig. 8. Resultados da métrica *Pointing Game* para a ROI original e com dilatação.

vadas métricas de desempenho no conjunto de teste, com um F1-*score* médio de 0,946. Esse resultado supera o obtido em [14], cujo F1-*score* de 0,91 foi alcançado com uma ResNet50 treinada em recortes manuais das imagens. A solução aqui proposta, portanto, supera esse resultado sem a necessidade de segmentação prévia de regiões do tecido epitelial — o que pode introduzir maior erro humano no processo —, porém a um custo computacional maior. Vale destacar que o ganho de desempenho é resultado da combinação entre ResNet50 e SVM: quando avaliados isoladamente, ambos os classificadores apresentam desempenho inferior, com F1-*score* de 0,89 e 0,86, respectivamente. A robustez do modelo foi avaliada por meio de um conjunto abrangente de métricas complementares; como acurácia, AUC, precisão, sensibilidade e especificidade, todas acompanhadas de intervalos de confiança de 95%, em consonância com as métricas predominantemente adotadas na literatura da área. Essa combinação permite caracterizar o desempenho do modelo tanto sob perspectivas de equilíbrio entre classes quanto de capacidade discriminativa, oferecendo uma análise abrangente.

A Figura 6 ilustra os mapas de atenção gerados pela arquitetura ResNet50 no primeiro nível do processo de detecção de displasia epitelial. Nos dois primeiros exemplos, observa-se que as regiões com maior ativação (em tons mais avermelhados) concentram-se na região limítrofe do epitélio com conjuntivo, local onde tipicamente se manifestam os primeiros sinais de displasia. No terceiro exemplo, embora o modelo não tenha identificado a displasia corretamente, a atenção foi direcionada à região epitelial, o que indica uma ativação coerente na região de interesse. Em contraste, no quarto exemplo, o modelo cometeu tanto um erro de classificação quanto de localização, concentrando-se inadequadamente na camada de ceratina em vez do tecido epitelial.

A Figura 8 apresenta os resultados da métrica *Pointing Game*. No cenário original, a acurácia do *Pointing Game* é de aproximadamente 0,51, indicando que 51% dos picos de ativação do mapa de atenção coincidem com as regiões de interesse anotadas pelos especialistas. Após a dilatação das máscaras, essa acurácia aumenta para 0,61. Esse comportamento evidencia uma limitação do algoritmo Grad-CAM mencionada previamente, em que as regiões de ativação tendem a ser mais difusas e a cobrir áreas ligeiramente adjacentes às regiões de interesse. Os resultados obtidos — tanto em relação às anotações originais quanto às anotações dilatadas — são modestos, indicando a existência de espaço para aprimoramentos na geração dos mapas de atenção, seja por meio da aplicação de técnicas mais avançadas ou do refinamento do modelo de classificação.

Uma limitação deste trabalho, a ser mitigada futuramente, é a falta de validação da utilidade das informações de apoio (sugestão de diagnóstico, percentual de confiança e mapas de atenção) e da ferramenta por parte de estudantes de patologia. Entretanto, acredita-se que do ponto de vista educacional, as informações de apoio representam uma ferramenta de grande potencial para o ensino de patologia bucal. Ao visualizar simultaneamente as informações de apoio geradas pelo modelo,

o estudante é convidado a um exercício ativo de comparação crítica entre as informações fornecidas pela ferramenta e o conhecimento de patologia. Casos em que o modelo acerta tanto a classificação quanto a localização; isto é, concentrando a ativação em regiões displásicas ou na junção de tecido epitelial e conjuntivo — região reconhecidamente crítica para os primeiros sinais de displasia — reforçam conceitos histopatológicos ao tornar visível a lógica do diagnóstico. Por outro lado, os casos de erro, especialmente aqueles em que a atenção recai sobre estruturas irrelevantes como a camada de ceratina, oferecem oportunidades de discussão. O estudante é provocado a identificar por que o modelo falhou e quais características foram negligenciadas. Essa abordagem potencializa o desenvolvimento do pensamento diagnóstico crítico, tornando o aprendizado mais contextualizado e próximo da prática clínica e laboratorial.

Apesar do elevado desempenho de classificação obtido, uma limitação recorrente em estudos com imagens médicas é a possibilidade de variações no preparo das imagens, o que pode comprometer a generalização do modelo. Técnicas de pré-processamento envolvendo normalização de coloração não são a melhor prática recomendada [21], mas a literatura indica que técnicas de aumento de representatividade de coloração através de perturbações de tom e saturação nas imagens podem mitigar esse problema. Entretanto, trabalhos que se beneficiam dessa estratégia utilizam porções menores das imagens — denominadas *patches* —, aumentando consideravelmente o conjunto de imagens. O presente trabalho, ao utilizar as imagens completas de maneira a preservar as características arquiteturais [22], não se beneficia dessa estratégia. Vale destacar novamente que o banco de dados empregado neste estudo supera em dimensão outros bancos de dados utilizados em investigações recentes similares [14]. Entretanto, estudos complementares com imagens provenientes de diferentes laboratórios são necessários para incrementar a validação da solução proposta; além de possivelmente servir como aumento de representatividade de casos ao possibilitar variações de coloração.

VI. Conclusão

Neste trabalho, o modelo proposto para detecção de DEO — que integra imagens histopatológicas e dados clínicos dos pacientes —, alcançou um F1-*score* médio de 0,946. Além disso, a incorporação de uma técnica de explicabilidade baseada em Grad-CAM, capaz de destacar nas imagens as regiões mais relevantes para a predição realizada pelo modelo de aprendizado profundo, aumentou a transparência do processo de decisão. As métricas de comparação dos mapas de atenção gerados pelo modelo com as demarcações feitas por patologistas são modestas, porém promissoras. Ademais, o desenvolvimento da ferramenta educacional tem o potencial de tornar o uso da solução mais interativo, com potencial de aplicação didática.

A integração entre aprendizado profundo, técnicas de explicabilidade e aplicações educacionais interativas pode contribuir significativamente tanto para o avanço da pesquisa em patologia digital quanto para a formação de futuros profissionais da área. Entretanto, um estudo com estudantes de patologia oral é necessário para comprovar a utilidade das informações de apoio e da ferramenta. Além disso, a avaliação com conjuntos de imagens provenientes de diferentes laboratórios é recomendada.

Referências

[1] S. B. Khanagar et al., "Application and performance of artificial intelligence (AI) in oral cancer diagnosis and prediction using histopathological images: a systematic review," *Biomedicines*, v. 11, n. 6, pp. 1612–17, 2023, ISSN: 2227-9059. DOI: 10.3390/biomedicines11061612.

[2] E. Odell, O. Kujan, S. Warnakulasuriya e P. Sloan, "Oral epithelial dysplasia: Recognition, grading and clinical significance," *Oral Diseases*, v. 27, n. 8, pp. 1947–1976, 2021, ISSN: 1601-0825. DOI: 10.1111/odi.13993.

[3] S. Warnakulasuriya et al., "Oral potentially malignant disorders: A consensus report from an international seminar on nomenclature and classification, convened by the WHO collaborating centre for oral cancer," *Oral Diseases*, v. 27, n. 8, pp. 1862–1880, 2020, ISSN: 1601-0825. DOI: 10.1111/odi.13704.

[4] S. A. Alajaji et al., "An update on the use of artificial intelligence in digital pathology for oral epithelial dysplasia research," *Head and Neck Pathology*, v. 18, n. 1, p. 38, 2024. DOI: 10.1007/s12105-024-01643-4.

[5] H. M. Afify, K. K. Mohammed e A. E. Hassanien, "Novel prediction model on OSCC histopathological images via deep transfer learning combined with Grad-CAM interpretation," *Biomedical Signal Processing and Control*, v. 83, pp. 104 704–13, 2023, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.104704.

[6] R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," em *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[7] N. S. Abdul et al., "Applications of artificial intelligence in the field of oral and maxillofacial pathology: A systematic review and meta-analysis," *BMC Oral Health*, v. 24, n. 1, 2024, ISSN: 1472-6831. DOI: 10.1186/s12903-023-03533-7.

[8] R. Kanthan e J. L. Senger, "The impact of specially designed digital games-based learning in undergraduate pathology and medical education," *Archives of Pathology and Laboratory Medicine*, v. 135, n. 1, pp. 135–142, 2011, ISSN: 0003-9985. DOI: 10.5858/2009-0698-oar1.1.

[9] L. A. Hassell et al., "Pathology education powered by virtual and digital transformation: now and the future," *Archives of Pathology and Laboratory Medicine*, v. 147, n. 4, pp. 474–491, 2022, ISSN: 0003-9985. DOI: 10.5858/arpa.2021-0473-ra.

[10] I. Waal, K. Schepman, E. Meij e L. Smeele, "Oral leukoplakia: A clinicopathological review," *Oral Oncology*, v. 33, n. 5, pp. 291–301, 1997, ISSN: 1368-8375. DOI: 10.1016/s1368-8375(97)00002-x.

[11] S. Muller e W. M. Tilakaratne, "Update from the 5th edition of the world health organization classification of head and neck tumors: Tumours of the oral cavity and mobile tongue," *Head and Neck Pathology*, v. 16, n. 1, pp. 54–62, 2022, ISSN: 1936-0568. DOI: 10.1007/s12105-021-01402-9.

[12] K. Rocha, J. C. M. Bermudez, E. R. C. Rivero e M. H. Costa, "A pathology-based machine learning method to assist in epithelial dysplasia diagnosis," *Research on Biomedical Engineering*, v. 38, n. 3, pp. 989–1002, 2022, ISSN: 2446-4740. DOI: 10.1007/s42600-022-00234-y.

[13] J. Peng et al., "Oral epithelial dysplasia detection and grading in oral leukoplakia using deep learning," *BMC Oral Health*, v. 24, n. 1, 2024, ISSN: 1472-6831. DOI: 10.1186/s12903-024-04191-z.

[14] K. da Rocha, J. C. M. Bermudez, E. R. C. Rivero e M. H. Costa, "A new machine-learning methodology for detection of epithelial dysplasia in oral leukoplakia lesions," *Research on Biomedical Engineering*, v. 42, n. 2, 2026, ISSN: 2446-4740. DOI: 10.1007/s42600-026-00482-2.

[15] B. Dwyer, J. Nelson, T. Hansen et al., *Roboflow*, versão 1.0, Available from https://roboflow.com. Computer vision., 2025.

[16] M. Sokolova e G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, v. 45, n. 4, pp. 427–437, 2009, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2009.03.002.

[17] J. Zhang et al., "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, v. 126, n. 10, 2017, ISSN: 1573-1405. DOI: 10.1007/s11263-017-1059-x.

[18] S. Desai e H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," em *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. DOI: 10.1109/WACV45572.2020.9093360.

[19] A. Abid et al., "Gradio: Hassle-free sharing and testing of ML models in the wild," *arXiv preprint arXiv:1906.02569*, 2019.

[20] Hugging Face, *Hugging face spaces: Collaborative machine learning apps*, https://huggingface.co/spaces.

[21] D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, v. 58, pp. 101 544–9, 2019, ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101544.

[22] P. Hankinson et al., "Demystifying oral epithelial dysplasia: A histological guide," *Pathology*, v. 56, n. 1, pp. 11–23, fev. de 2024, ISSN: 0031-3025. DOI: 10.1016/j.pathol.2023.10.002.

[23] D. Müller, I. Soto-Rey e F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, v. 15, n. 1, pp. 1–8, 2022, ISSN: 1756-0500. DOI: 10.1186/s13104-022-06096-y.

Frozen GPT-2 Backbone for Few-Shot Congestive Heart Failure Detection from ECG Signals

Luiz Otavio Ribeiro
UTFPR, Curitiba, Brazil
luizotavioribeiromsz@gmail.com

André Eugenio Lazzaretti
UTFPR, Curitiba, Brazil
lazzaretti@utfpr.edu.br

Everton Luiz de Aguiar
UTFPR, Pato Branco, Brazil
eaguiar@utfpr.edu.br

João Paulo Teixeira
IPB, Bragança, Portugal
joaopt@ipb.pt

***Abstract*—Congestive Heart Failure (CHF) is a progressive cardiovascular condition whose early identification can benefit from automated analysis of electrocardiogram (ECG) signals. This work investigates few-shot CHF classification using a frozen, pre-trained GPT-2 backbone adapted to process raw ECG recordings alongside age and gender information. The model employs a lightweight convolutional embedding layer, six frozen transformer blocks, and a binary classification head, without signal filtering or handcrafted feature extraction. Experiments were conducted on a dataset containing 840 three-second ECG samples from 28 subjects, including 14 patients with severe CHF and 14 matched healthy controls. Performance was evaluated across multiple training set sizes, learning rates, random seeds, checkpoint selection strategies, and data augmentation settings. The best configuration achieved a multi-seed mean F1-score of 0.8846 ± 0.1298 using 300 shots per class. In the more restrictive 15-shot setting, augmentation combined with validation-based checkpoint selection achieved a best F1-score of 0.8845 and a mean of 0.789 ± 0.092. These results indicate that a language-pretrained frozen transformer is a feasible approach for CHF classification with limited task-specific adaptation and substantially reduced training data.**

***Index Terms*—digital health, congestive heart failure, electrocardiogram, few-shot learning, frozen backbone, GPT-2, large language model**

I. Introduction

The increasing availability of digital health data has created new opportunities for automated clinical decision support, particularly in cardiology. Among the conditions that most burden healthcare systems globally, Congestive Heart Failure (CHF) stands out for its high prevalence, progressive nature, and significant impact on quality of life — affecting millions of patients with symptoms such as dyspnea, fatigue, and peripheral oedema [1].

The electrocardiogram (ECG) remains one of the most accessible tools for cardiac assessment, yet interpreting CHF-related patterns — such as QRS prolongation, T-wave inversions, and ST segment alterations — demands clinical expertise that is not uniformly available across healthcare settings. This gap motivates the development of automated, AI-driven classification systems that support diagnosis across diverse clinical environments.

Early approaches to ECG-based CHF detection relied heavily on handcrafted feature extraction and classical machine learning [2]. The advent of deep learning shifted the paradigm, with CNN and LSTM architectures achieving competitive accuracy on raw ECG signals [3]. However, these task-specific models are tailored to each application and dataset, limiting their broader adoption for other diseases, for example.

To overcome this condition, the applicability of foundation models emerges as a promising alternative, as they have recently demonstrated strong generalization capabilities across domains. Notably, LLM-based architectures, originally designed for sequential language modeling, have proven effective for time series analysis, as demonstrated by the One-Fits-All framework [4], which adapts a single pre-trained transformer backbone to diverse time series tasks without task-specific redesign.

This work originally proposes the application of such an LLM-based foundation model to the automated detection of CHF from raw ECG signals, focusing on exploring how many samples are necessary to use in the training to achieve the same (or better) results as dedicated deep learning models. Using a dataset constructed from the BIDMC Congestive Heart Failure and PTB Diagnostic ECG databases (PhysioNet). The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 details the methodology, Section 4 presents and discusses the results, and Section 5 draws the conclusions.

II. Related Work

Deep learning methods have been widely applied to detect pathology in diverse biosignals. LSTM networks have been used to classify chronic laryngitis [5] and cardiac arrhythmia [6], while ANN-based approaches have addressed EEG classification [7] and arrhythmia detection [8]. CNNs and hybrid architectures have also been explored for ECG component analysis [9], synthetic data generation via GANs [10], and atrial fibrillation detection through MLP-based feature extraction [11]. Although often regarded as a mature technology, these methods remain highly relevant — particularly for CHF detection in ECG signals.

Regarding CHF specifically, we can highlight our previous work [3], which reports 98.21% accuracy for the CNN model and 92.26% for the LSTM model in CHF detection. Furthermore, [12] presents an LSTM-based approach applied to wavelet-filtered signals, achieving an accuracy of 99.86%. In [13], a 1D CNN applied to single heartbeat segments was able to reach 97.8% accuracy in detecting CHF. [14] employed a 1D CNN combined with a Gabor filter to address multi-class ECG classification with CHF among the target classes, attaining

an accuracy of 98.74%. Finally, [15] proposes a CNN-LSTM architecture that uses RR intervals as network input, achieving 99.52% accuracy and a 98.94% F1-score.

The emergence of transformer-based architectures, specially LLM's, has opened new directions for ECG. By treating cardiac signals as sequential data, self-attention mechanisms can capture temporal dependencies, with their temporal embeddings serving as inputs to pre-trained LLMs.

Transformer-based models (without the use of LLMs) have shown competitive results in arrhythmia classification [16], arrhythmia detection [17], and multi-disease ECG diagnosis [18]. In the latter work, the model was pre-trained using masked image modeling on more than 8.5 million ECGs from 2.1 million patients. Despite their effectiveness, these architectures are still trained from scratch on ECG-specific data, limiting their ability to transfer sequential knowledge across domains.

Large language models are naturally associated with textual inputs and outputs. However, the architecture itself is essentially a transformer-based network. With this in mind, many foundation models have been conceived using pre-trained large-scale models as backbones, repurposing the LLMs to accept time series as inputs and perform classification, forecasting, and related tasks — offering a fundamentally different paradigm.

The most prominent example of this paradigm is the One-Fits-All framework [4], which uses a frozen GPT-2 backbone to process patched time series without task-specific architectural changes, achieving competitive performance across classification, forecasting, and anomaly detection. Among works that process ECG signals with LLM backbones, MedualTime [19] applied a frozen GPT-2 with lightweight adapters to medical ECG signals, and InstructTime [20] recast time-series classification (including ECG) as a language-generation task. In the heart failure domain, [21] leveraged ClinicalBERT to enhance ECG-based heart failure risk prediction through text-ECG embedding alignment. In this case, the LLM was used only for textual alignment, not as the main backbone.

Despite the growing body of work applying LLMs to ECG analysis, existing approaches either use the language model exclusively on the text side [21], or evaluate on benchmark datasets that lack a CHF label altogether, such as PTB-XL and MIT-BIH [19], [20]. CHF-specific studies, on the other hand, have relied on convolutional and recurrent architectures trained from scratch [3], [12], [13], [15], or at most on transformer encoders built without pre-trained language priors [22]. This work addresses this intersection directly: we propose using a frozen pre-trained GPT-2 backbone to process raw ECG signals for CHF detection, extending the One-Fits-All paradigm to a clinically relevant and underexplored pathology. To our knowledge, this represents the first application of a pre-trained language model backbone to CHF classification from ECG signals.

III. Methodology

This section provides an overview of the approach and model used to detect CHF from ECG signals. The methodology covers the dataset construction, the LLM-based architecture, the training and evaluation process.

A. Datasets

Two publicly available PhysioNet repositories [23] were used as data sources: the BIDMC Congestive Heart Failure Database [24], which contains ECG recordings from CHF patients, and the PTB Diagnostic ECG Database [25], which provided the healthy control signals. Since the BIDMC database lacks control recordings and the two repositories use different sampling rates, a custom dataset was constructed to ensure consistency across all samples.

The resulting dataset contains 840 samples from 28 subjects, with 30 samples extracted per individual. Half of the subjects (14) present severe CHF classified by the New York Heart Association (NYHA) functional classes 3 and 4, while the remaining 14 are healthy individuals taken from the PTB database.

Gender and age were used as matching criteria when pairing pathological and control subjects, so that each CHF patient is associated with a healthy match of the same gender and similar age. This pairing strategy reduces the risk of demographic bias during training and is detailed in Table I. Representative ECG samples from both groups are illustrated in Figure 1.

All recordings were resampled to 250 Hz, and no noise filtering was applied. Each sample lasts 3 seconds, corresponding to 750 time steps. No handcrafted feature extraction was performed; instead, Min-Max scaling was applied per class for normalization purposes. To avoid data leakage, the scaler was fitted exclusively on the training partition of each seed and subsequently applied to the corresponding validation and test partitions using the fitted parameters.

To prevent data leakage, for each seed, 2 individuals were randomly selected for validation and 2 different individuals for testing, with the remaining individuals used for training. This subject-wise split was varied across seeds, ensuring that no individual appearing in the validation or test sets was ever seen during training.

B. LLM-Based Architecture

The proposed model adapts the GPT4TS framework [4] to classify CHF from raw ECG signals, using a pre-trained GPT-2 as a frozen backbone (Figure 2). Rather than designing a task-specific architecture, the approach repurposes the language model's sequential pattern recognition capabilities, applying them directly to physiological time series.

1) Input Representation: Each raw ECG sample of 750 timesteps is first decomposed into overlapping sub-windows via a sliding window of length 250 and stride 125 (50% overlap), yielding 5 sub-windows per recording and expanding the working dataset from 840 to 4,200 samples. Each sub-window inherits the label, age, and gender of its parent

TABLE I
PAIRS FORMED BASED ON GENDER AND AGE. DATASET FIRST INTRODUCED IN [3].

Pair	Gender *(CHF)*	Gender *(Control)*	Age *(CHF)*	Age *(Control)*
1	Male	Male	71	68
2	Female	Female	61	67
3	Male	Male	63	64
4	Male	Male	54	54
5	Female	Female	59	57
6	Male	Male	48	37
7	Male	Male	51	50
8	Female	Female	63	69
9	Male	Male	22	26
10	Female	Female	54	54
11	Male	Male	61	59
12	Male	Male	63	68
13	Female	Female	61	52
14	Male	Male	53	55

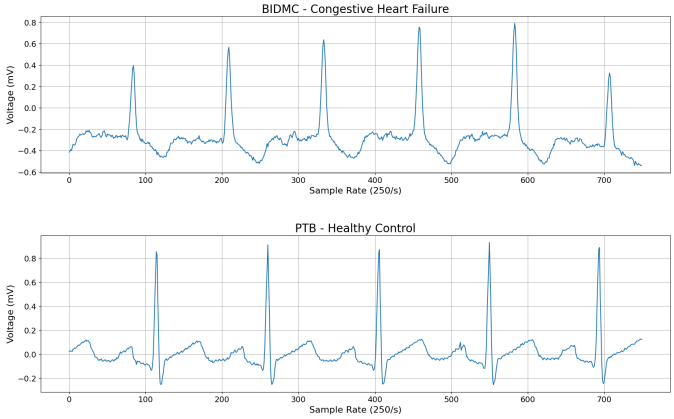


Fig. 1. Representative ECG samples from both groups: CHF (top) and healthy control (bottom). Reproduced from [3].

recording, and the few-shot split is enforced at the recording level to prevent data leakage between partitions.

Each sub-window is then passed through a token embedding layer implemented as a 1D convolutional layer (Conv1d, kernel size 3, circular padding), which projects each scalar timestep into a 768-dimensional vector matching GPT-2's hidden size while capturing local context between consecutive timesteps. A fixed sinusoidal positional encoding is added to the resulting sequence before it is fed into the backbone.

2) Frozen GPT-2 Backbone: The embedded sequence is passed directly to the first six transformer blocks of a pre-trained GPT-2 (124M parameters), bypassing its internal token embedding table entirely. Within this truncated stack, all attention and feed-forward weights remain frozen at their text-pretrained values. Only the internal LayerNorm parameters and the learned positional embedding table (`wpe`) of GPT-2 are kept trainable, alongside the token embedding layer and the classification head. This design follows the one-fits-all rationale [4]: the frozen attention heads are treated as a general-purpose sequence processor whose representations transfer across domains, while the lightweight trainable parameters adapt the interface between the ECG signal and the frozen backbone. Notably, GPT-2's internal positional embedding (`wpe`) is summed with the external sinusoidal encoding inside the first transformer block, so both positional signals are active during inference.

3) Classification Head: The output of the GPT-2 stack is passed through a GELU activation and an additional LayerNorm, then mean-pooled across the temporal dimension to produce a single 768-dimensional representation. Age and gender metadata are concatenated to this vector, expanding it to 770 dimensions, and a linear layer projects it to a single logit. The model is trained with binary cross-entropy loss, and predictions are obtained by applying a sigmoid function with a decision threshold of 0.5.

C. Training Procedure

All trials use the RAdam optimizer with batch size 32 (operating over sub-windows), and a fixed epoch budget per configuration with no early stopping — except in Experiment 2, where the best-validation-loss checkpoint is evaluated as an alternative to the final-epoch weights. Validation metrics are computed every epoch. A learning rate step-decay schedule is present in the implementation but configured with decay epochs that exceed every budget swept in this study, so it never triggers in practice.

D. Evaluation Protocol

Performance is reported as F_1 score, accuracy, precision, and recall, computed on the test split at a fixed 0.5 decision threshold. The base train/validation/test split is formed at the subject level: 2 subjects per class are held out for testing and 2 per class for validation, fixed via a `subject_split_seed` that is tied to the `few_shot_seed` used for the corresponding trial. This guarantees zero subject, recording, or sub-window overlap between partitions (verified empirically) and an identical test-subject composition across every shot level within a given seed. The few-shot constraint is then applied within the remaining training-pool subjects: N raw recordings per class are drawn at random from the training pool, and all 5 sub-windows from each selected recording are added to the effective training set. Because test-subject membership does not depend on N, the test set is fixed across shot levels for a given seed, avoiding a shrinking/growing test set as a confound in cross-shot-level comparisons.

IV. RESULTS AND DISCUSSIONS

This section presents the results of two experiments evaluating our proposed LLM method for detecting CHF in ECG signals. We then compare our method to earlier methods on the same dataset and to other current state-of-the-art solutions.

A. Experiment 1: Few-Shot Performance Across Shot Levels

The first experiment consists of training the model with different numbers of samples per class, referred to as shots in Table II. This table also reports F_1 scores, accuracies and precisions for each shot level and learning rate combination. The results show a clear overall improvement in classification

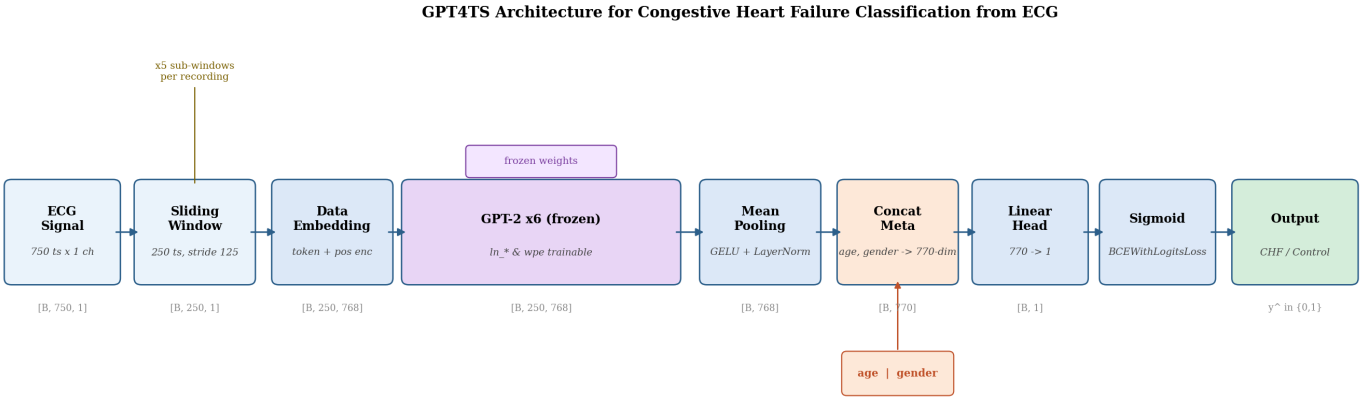


Fig. 2. GPT4TS architecture adapted for CHF classification from ECG signals. The raw ECG signal is decomposed into overlapping sub-windows, embedded via a Conv1d layer, and processed by a frozen six-block GPT-2 backbone. Age and gender metadata are concatenated after mean pooling before the final classification head.

performance as the number of training samples increases. The F1-score std rises from 0.4465 ± 0.3769 with one shot per class to 0.8846 ± 0.1298 with 300 shots, with the most pronounced gains occurring on early shots and performance approaching saturation from 30 shots.

TABLE II
BEST TEST-SET RESULT AT EACH SHOT LEVEL (480 TRIALS TOTAL; 30 TRIALS PER ROW).

Shots	LR	F_1 (std)	Accuracy (std)	Precision (std)
1	0.001	0.4465 (0.3769)	0.5498 (0.2446)	0.4163 (0.3814)
2	0.001	0.6233 (0.2222)	0.6338 (0.1269)	0.6324 (0.2950)
3	0.01	0.7024 (0.2803)	0.7617 (0.1789)	0.7771 (0.3113)
5	0.01	0.7482 (0.1901)	0.7980 (0.1189)	0.9024 (0.1223)
10	0.0003	0.8040 (0.1384)	0.8220 (0.1144)	0.8907 (0.1473)
15	0.001	0.8242 (0.1477)	0.8483 (0.1185)	0.9283 (0.1251)
20	0.001	0.8336 (0.1481)	0.8618 (0.1123)	0.9425 (0.0980)
30	0.001	0.8404 (0.1503)	0.8703 (0.1119)	0.9503 (0.0817)
40	0.0003	0.8484 (0.1647)	0.8802 (0.1260)	0.9544 (0.0702)
50	0.01	0.8585 (0.1213)	0.8672 (0.1105)	0.9023 (0.1228)
75	0.0003	0.8325 (0.1425)	0.8643 (0.1025)	0.9549 (0.0723)
100	0.001	0.8581 (0.1352)	0.8850 (0.0986)	0.9587 (0.0527)
150	0.001	0.8649 (0.1310)	0.8883 (0.0993)	0.9564 (0.0611)
200	0.0003	0.8714 (0.1345)	0.8948 (0.1024)	0.9648 (0.0549)
250	0.001	0.8781 (0.1101)	0.8850 (0.1047)	0.9346 (0.1202)
300	**0.0003**	**0.8846 (0.1298)**	**0.9068 (0.0967)**	**0.9765 (0.0344)**

In Table III, to better visualize the results at each shot level, we report the mean F_1 across 10 training runs with different seed combinations and three learning rates. The best combination points to 300 samples per class with a learning rate of 0.0003. The multi-seed results provide a more conservative assessment of model robustness. Variability is high in the lowest-shot settings, particularly with one shot per class, but decreases substantially as more training samples are included, indicating that larger training sets improve both average performance and reproducibility. The learning rate also affects performance differently across shot levels, with 0.01 and 0.001 generally performing better in low- and intermediate-shot regimes and 0.0003 becoming competitive at higher shot counts. Since the best-case values represent the upper performance envelope across multiple trials, the reported means and standard deviations should be considered more representative of the model's expected behavior.

TABLE III
TEST F_1 MEAN (STD) OVER 10 SEEDS, PER CONFIGURATION

Shots	lr=0.001, e30	lr=0.01, e30	lr=0.0003, e50
1	0.447 (0.377)	0.434 (0.355)	0.400 (0.327)
2	0.623 (0.222)	0.583 (0.342)	0.436 (0.361)
3	0.667 (0.273)	0.702 (0.280)	0.392 (0.321)
5	0.745 (0.159)	0.748 (0.190)	0.741 (0.111)
10	0.801 (0.141)	0.784 (0.135)	0.804 (0.138)
15	0.824 (0.148)	0.811 (0.164)	0.824 (0.149)
20	0.834 (0.148)	0.779 (0.213)	0.824 (0.155)
30	0.840 (0.150)	0.820 (0.114)	0.831 (0.154)
40	0.848 (0.160)	0.842 (0.118)	0.848 (0.165)
50	0.818 (0.175)	0.859 (0.121)	0.842 (0.156)
75	0.827 (0.139)	0.822 (0.126)	0.833 (0.143)
100	0.858 (0.135)	0.846 (0.127)	0.847 (0.149)
150	0.865 (0.131)	0.812 (0.139)	0.843 (0.151)
200	0.860 (0.103)	0.840 (0.132)	0.871 (0.134)
250	0.878 (0.110)	0.816 (0.118)	0.864 (0.097)
300	0.843 (0.128)	0.826 (0.137)	**0.885 (0.130)**

B. *Experiment 2 - Boosting the few-shot classification*

To improve model performance in the few-shot regime (up to 20 shots per class), a second experiment was conducted, building on the first. The following strategies were introduced:

- **Checkpoint selection:** instead of always using the final-epoch weights, we compared two policies — the last-epoch checkpoint (*last*) versus the checkpoint with the lowest validation loss (*best*);
- **Data augmentation:** a lightweight augmentation scheme was applied during training, combining additive noise, amplitude scaling, and temporal shift, and compared against the no-augmentation baseline.

Both factors were crossed at three shot levels, $N \in \{1, 5, 15\}$, with 3 random seeds per combination, for a total of 36 trials. Table IV shows the full results of the experiment.

The effects of checkpoint selection and data augmentation vary across shot levels and are especially unstable at $N{=}1$, where mean F_1 ranges from 0.222 (best checkpoint, no

TABLE IV
CHECKPOINT SELECTION × AUGMENTATION, 3 SHOT LEVELS, 3 SEEDS EACH (36 TRIALS TOTAL)

Shots	Ckpt	Aug	Best F_1	Mean F_1 (std)
1	last	off	0.6667	0.222 (0.314)
1	last	on	0.9983	0.714 (0.216)
1	best	off	0.6667	0.389 (0.283)
1	best	on	0.9772	0.492 (0.399)
5	last	off	0.8119	0.713 (0.070)
5	last	on	0.9917	0.806 (0.144)
5	best	off	0.6667	0.644 (0.026)
5	best	on	0.9917	0.837 (0.114)
15	last	off	0.8186	0.714 (0.074)
15	last	on	0.8108	0.712 (0.070)
15	best	off	0.8848	**0.789 (0.092)**
15	best	on	0.8086	0.712 (0.068)

augmentation) to 0.714 (last checkpoint, with augmentation) across only 3 seeds. Augmentation improves the mean F_1 at every shot level and checkpoint policy tested here except one: at N=5, mean F_1 rises from 0.644/0.713 (best/last, off) to 0.837/0.806 (best/last, on); at N=15, the best-performing combination is instead the *best*-checkpoint without augmentation (mean $F_1 = 0.789$), the only cell in this experiment where augmentation does not help. Given the small seed count (3) and narrow shot-level coverage of this experiment, we treat these results as a directional signal about checkpoint/augmentation effects at very low shot counts, not a definitive characterization

C. Comparative Analysis - Same dataset

As shown, the proposed approach's mean result across seeds falls below both conventional deep learning models from our previous work, despite using less than the full dataset (the training pool at N=300 corresponds to roughly 71% of all 840 recordings, with the remainder held out for validation/testing at the subject level). We report this as evidence of the feasibility of a frozen, language-pretrained backbone on this task under a leakage-safe evaluation, not of superiority over the previous, purpose-built models.

TABLE V
COMPARISON WITH OUR PREVIOUS WORK [3] (MEAN F_1 ACROSS SEEDS).

Model	F_1 *(std)*	*Shots/class*	*Dataset (%)*
Proposed LLM-Based (Exp. 1)	0.8846 (0.1298)	300	≈71%
CNN [3]	0.9801	336	80%
Proposed LLM-Based (Exp. 2)	0.789 (0.092)	15	≈3.6%
LSTM [3]	0.9161	336	80%

The results provide strong evidence of data efficiency and competitive performance, although superiority over previous models is not consistently observed across all random seeds. An additional distinction is that the CNN and LSTM models were developed specifically for ECG classification and trained from scratch, whereas most of the proposed GPT-2 backbone's parameters remain frozen. The results, therefore, suggest that representations learned from a non-medical sequential domain can support ECG classification with only limited adaptation.

D. Comparisons with other State-of-the-Art Methods

Table V provides a reference for the performance of related works. Since these works differ substantially in their data sources and preprocessing pipelines, a direct comparison is not meaningful; nonetheless, the table offers a useful reference point for positioning the proposed approach relative to the state of the art. This result is particularly relevant given that the proposed model processes normalized ECG amplitudes directly, without wavelet filtering, handcrafted descriptors, or an ECG-specific backbone.

However, Table V should be interpreted as a contextual rather than a direct ranking. The studies differ in terms of database composition, number of subjects, signal duration, preprocessing, segmentation procedure, class balance, validation strategy, and whether the partitions are subject-independent. Some methods also classify individual heartbeats or RR intervals, whereas the proposed approach processes fixed-duration ECG windows. A direct benchmark using identical subject-level partitions and evaluation metrics would be required to establish superiority over the state of the art.

TABLE VI
COMPARISON OF RESULTS WITH STATE-OF-THE-ART METHODS FOR CHF CLASSIFICATION

Method	*Acc. (%)*	*Sens. (%)*	*Prec. (%)*	*F1*
[12] (LSTM)	99.86	99.85	99.87	0.9986
[15] (CNN+LSTM)	99.52	99.31	–	0.9894
LLM-Based (Exp. 1)	90.68	–	97.65	0.8846
[14] (CNN)	98.74	–	–	–
[3] CNN	98.21	96.10	100.00	0.9801
[13] (CNN)	97.80	–	–	–
LLM-Based (Exp. 2)	–	–	–	0.789 (0.092)
[3] LSTM	92.26	92.21	91.03	0.9161

V. LIMITATIONS AND FUTURE WORK

A limitation of this study is that each class was captured using a different acquisition setup, which may impact classification performance. It is important to note that every CHF sample originates from BIDMC (natively 250 Hz), while every control sample originates from PTB (natively 1 kHz, downsampled 4×). This means that differences in the acquisition system could potentially confound the classification process.

Another limitation is related to the data collection method. Although Min-Max scaling was fitted separately per class using only the training partition of each seed, avoiding leakage from validation or test data, the two sources employ different acquisition methods, so the resulting amplitude levels may not be directly comparable between classes. Ideally, both classes would be acquired using the same protocol and equipment.

As future work, we intend to include three simple ablation studies: (i) a frozen GPT-2 backbone versus random initialization of the same architecture, (ii) with and without age and gender covariates, and (iii) with and without dual positional encoding. These ablations are motivated by the fact that it is not yet established whether the observed gains stem from the knowledge transferred by GPT-2 or from the remaining components of the pipeline. Conducting these experiments

would make the computational contribution more verifiable without requiring extensive architectural changes.

VI. Conclusion

This work explored the application of a frozen pre-trained GPT-2 backbone to the detection of Congestive Heart Failure from raw ECG signals, following the One-Fits-All framework. Unlike the deep learning methods previously applied to this task, which were purpose-built and trained from scratch on domain-specific data, the proposed approach requires no task-specific architectural design, no signal filtering, and no hand-crafted feature extraction. The frozen backbone processes the raw ECG amplitude directly, with age and gender as the only additional inputs.

The results are modest relative to the full-data baselines. Using a subject-disjoint evaluation protocol designed to avoid patient-identity leakage, the mean result over 10 seeds is $F_1 = 0.8846 \pm 0.1298$ (accuracy $90.68\% \pm 9.67\%$) at the best shot level (300), below both full-data baselines from our previous work (CNN 98.21%, LSTM 92.26%).

The broader implication is architectural. A single frozen backbone, pre-trained exclusively with natural language data, can be repurposed for ECG classification with minimal adaptation. Just a lightweight embedding layer, a classification head, and a handful of trainable normalization parameters. This contrasts sharply with prior CHF detection methods, which require retraining from scratch whenever the dataset, lead configuration, or target pathology changes. Whether this reflects the backbone's pretrained linguistic knowledge or simply the architectural inductive bias of a frozen transformer of this size and depth remains an open question after future study's ablation results, and should be resolved before extending the approach to other cardiac conditions or physiological signals.

Acknowledgments

The authors thank the Computational Intelligence Laboratory - LABIC, Curitiba, Brazil.

References

[1] B. Shahim, C. J. Kapelios, G. Savarese, and L. H. Lund, "Global public health burden of heart failure: An updated review," *Cardiac Failure Review 2023;9:e11.*, 2023. [Online]. Available: https://doi.org/10.15420/cfr.2023.05

[2] T. Mahboob, L. Hernandez, N. Akhtar, and S. K. Ikram, "Detecting congestive heart failure by extracting multimodal features and employing machine learning techniques," *BioMed Research International*, 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7049402/

[3] L. Ribeiro, N. Guerreiro, M. K. Chaabani, L. E. Luiz, A. E. Lazzaretti, and J. P. Teixeira, "Congestive heart failure detection in ecg using lstm and cnn," in *Health Technologies and Demographic Challenges*, P. M. Gaspar, J. M. Cueva Lovelle, C. Mentenegro-Marín, and T. Guarda, Eds. Cham: Springer Nature Switzerland, 2025, pp. 147–156.

[4] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, "One fits all: Power general time series analysis by pretrained lm," *arXiv preprint arXiv:2302.11939*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.11939

[5] V. Guedes, A. Junior, F. Teixeira, J. Fernandes, and J. P. Teixeira, "Long short term memory on chronic laryngitis classification," in *CENTERIS/ProjMAN/HCist 2018 - Procedia Computer Science*, vol. 138, 2018, pp. 250–257.

[6] B. Hou, J. Yang, P. Wang, and R. Yan, "Lstm-based auto-encoder model for ecg arrhythmias classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1232–1240, 2020.

[7] P. M. Rodrigues and J. P. Teixeira, "Classification of electroencephalogram signals using artificial neural networks," in *Proceedings of 3rd International Conference on BioMedical Engineering and Informatics (BMEI'10)*, Yantai, China, 10 2010.

[8] B. Anuradha and V. V. Reddy, "Ann for classification of cardiac arrhythmias," *ARPN Journal of Engineering and Applied Sciences*, vol. 3, no. 3, pp. 1–6, 2008.

[9] R. Costa, T. Winkert, A. Manhães, and J. P. Teixeira, "Qrs peaks, p and t waves identification in ecg," in *CENTERIS/ProjMAN/HCist 2020 - Procedia Computer Science*, vol. 181, 2021, pp. 957–964.

[10] K. F. Hossain, S. A. Kamran, A. Tavakkoli, L. Pan, X. Ma, S. Rajasegarar, and C. Karmaker, "Ecg-adv-gan: Detecting ecg adversarial examples with conditional generative adversarial networks," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021, pp. 50–56.

[11] P. H. Borghi, R. Borges, and J. P. Teixeira, "Atrial fibrillation classification based on mlp networks by extracting jitter and shimmer parameters," in *CENTERIS/ProjMAN/HCist 2020 - Procedia Computer Science*, vol. 181, 2021, pp. 931–939.

[12] A. Darmawahyuni, S. Nurmaini, M. Yuwandini, Muhammad Naufal Rachmatullah, F. Firdaus, and B. Tutuko, "Congestive heart failure waveform classification based on short time-step analysis with recurrent network," *Informatics in Medicine Unlocked*, vol. 21, p. 100441, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820305918

[13] M. Porumb, E. Iadanza, S. Massaro, and L. Pecchia, "A convolutional neural network approach to detect congestive heart failure," *Biomedical Signal Processing and Control*, vol. 55, p. 101597, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809419301776

[14] V. Jahmunah, E. Ng, T. R. San, and U. R. Acharya, "Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using gaborcnn model with ecg signals," *Computers in Biology and Medicine*, vol. 134, p. 104457, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521002511

[15] S. Kusuma and K. Jothi, "Ecg signals-based automated diagnosis of congestive heart failure using deep cnn and lstm architecture," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 247–257, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0208521622000067

[16] H. El-Ghaish and E. Eldele, "Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer," *Biomedical Signal Processing and Control*, vol. 89, p. 105714, 03 2024.

[17] C. Ji, L. Wang, J. Qin, L. Liu, Y. Han, and Z. Wang, "Msgformer: A multi-scale grid transformer network for 12-lead ecg arrhythmia detection," *Biomedical Signal Processing and Control*, vol. 87, p. 105499, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423009321

[18] A. Vaid, J. Jiang, A. Sawant, S. Lerakis, E. Argulian, Y. Ahuja, J. Lampert, A. Charney, H. Greenspan, J. Narula, B. Glicksberg, and G. N. Nadkarni, "A foundational vision transformer improves diagnostic performance for electrocardiograms," *npj Digital Medicine*, vol. 6, no. 1, p. 108, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00840-9

[19] J. Ye, W. Zhang, Z. Li, J. Li, M. Zhao, and F. Tsung, "Medualtime: A dual-adapter language model for medical time series-text multimodal learning," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, ser. IJCAI '25. IJCAI, 2025, p. 880. [Online]. Available: https://www.ijcai.org/proceedings/2025/0880.pdf

[20] M. Cheng, Y. Chen, Q. Liu, Z. Liu, and Y. Luo, "Advancing time series classification with multimodal language modeling," 2024. [Online]. Available: https://arxiv.org/abs/2403.12371

[21] C. Chen, L. Li, M. Beetz, A. Banerjee, R. Gupta, and V. Grau, "Large language model-informed ecg dual attention network for heart failure risk prediction," *IEEE Transactions on Big Data*, vol. 11, no. 3, p. 948–960, 2025. [Online]. Available: http://dx.doi.org/10.1109/TBDATA.2025.3536922

[22] T. Liu, Y. Si, W. Yang, J. Huang, Y. Yu, G. Zhang, and R. Zhou, "Inter-patient congestive heart failure detection using ecg-convolution-vision transformer network," *Sensors*, vol. 22, no. 9, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3283

[23] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Com-

ponents of a new research resource for complex physiologic signals," *Circulation [Online]*, vol. 101, no. 23, pp. e215–e220, 2000.

[24] D. S. Baim, W. S. Colucci, E. S. Monrad, H. S. Smith, R. F. Wright, A. Lanoue, D. F. Gauthier, B. J. Ransil, W. Grossman, and E. Braunwald, "Survival of patients with severe congestive heart failure treated with oral milrinone," *Journal of the American College of Cardiology*, vol. 7, no. 3, pp. 661–670, 1986.

[25] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet," *Biomedizinische Technik / Biomedical Engineering*, vol. 40, no. s1, pp. 317–318, 1995.

Gas Vesicles Image Enhancement using Median Filter in HSV Image Representation

1st Raymundo Cordero Garcia
Electrical Eng. Program
PPGEE - UFMS and
Biomedical Eng. Program /COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
raymundo.garcia@ufms.br

2nd Débora Gama da Silva Mendes
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
mdeboragama@peb.ufrj.br

3rd Isabel Nunes Ramalho da Rocha
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
isabel.nunesrr@peb.ufrj.br

4th Mariana Rocha da Silva
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
mariana.silva@peb.ufrj.br

5th Juliana Oliveira da Silva Fernandes
Biomedical Eng. Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
julianaosf@peb.ufrj.br

6th Wellington Silva Ferreira
Carlos Chagas Filho
Institute of Biophysics
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
wellington.silvaferreira@biof.ufrj.br

7th Beatriz Ferreira de Carvalho Patricio
Post-Graduate Program in Molecular
and Cellular Biology
Laboratory of Pharmaceutical and
Technological Innovation
Federal University of the
State of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
beatriz.patricio@unirio.br

8th Gilberto Weissmuller
Post-Graduate Program in
Biological Sciences
Carlos Chagas Filho
Institute of Biophysics
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
gweissmuler@biof.ufrj.br

9th João Carlos Machado
Biomedical Eng. Program /COPPE
and Post-Graduate Program
in Surgical Sciences
Rio de Janeiro, RJ, Brazil
jcm@peb.ufrj.br

***Abstract*—Different mechanical and acoustic parameters of gas vesicles (GVs) must be analyzed for their application as contrast agent for ultrasound imaging, such as the modulus of elasticity and topogrphic information. Images acquired through Atomic Force Microscopy (AFM) are employed to analyze the GVs elasticity. Those images are affected by noise and other interferences. Hence, different filtering methods were proposed, such as median filtering applied conventionally to the RGB color representation of the image. However, this color representation is not adequate for evaluating useful image representation such as brightness and gray scale. This paper defines a new filtering method based on the application of median filter but applied in the HSV color representation. HSV (hue-saturation-value) gives information about brightness and saturation (related to grayscale) which are useful to identify images problems. Tests performed in experimental images of GVs acquired through AFM prove that the proposed filtering method is an useful alternative in the denoising of images acquired by AFM.**

***Index Terms*—Gas vesicles, Atomic Force Microscopy, HSV color representation, image enhancement, median filter, RGB color representation.**

This research was supported in part by CAPES, CNPq and FAPERJ

I. Introduction

Biomedical imaging examinations are widely used in clinical practice for the assessment of organs and tissues. Among the most commonly used imaging modalities is ultrasound (US), a noninvasive and cost-effective technique that uses nonionizing radiation [1]. US enables the acquisition of real-time images of organs and tissues with high spatial and temporal resolution. Contrast-enhanced ultrasound (CEUS) can be applied to enhance the visualization of specific regions of interest. In CEUS, contrast agents increase the intensity of the echoes generated by the incident US beam within the target region relative to the surrounding tissue, thereby improving the contrast between the region of interest [2]–[4].

Gas vesicles (GVs) have been investigated as a promising alternative to conventional contrast agents used in CEUS. They are easy to produce in laboratory, and they have both structural stability and the capability (due to their nanoscale size) to access extravascular and extratumoral spaces that are inaccessible to conventional microbubble-based contrast agents. Hence, GVs can overcome some limitations of current CEUS techniques based on conventional microbubbles [3], [5].

GVs are subjected to hydrostatic pressure from the surrounding aqueous medium, as well as to internal turgor pressure generated by osmotic forces. GVs collapse occurs when the external pressure exceeds their critical collapse pressure. [5]. Understanding the mechanical properties of GVs is essential for characterizing and controlling their behavior under incident ultrasound. The resulting compression and rarefaction cycles impose mechanical stresses on the GVs that can induce their collapse, thereby limiting their stability and effectiveness as contrast agents.

Analyzing the mechanical properties of complex micro- and nanostructures is possible using Atomic Force Microscopy (AFM) [6]. This technique employs an intermolecular force sensor —consisting of a probe attached to a lever known as a cantilever—to scan the sample pixel by pixel, providing high-resolution qualitative and quantitative data on topography, adhesion, deformation, and sample stiffness.

The scanning process generates images that allow for the correlation of specific topographic regions with their corresponding mechanical responses. However, image quality can be compromised by noise and artifacts that hinder this correlation [7]. Hence, image processing is an essential step, involving tilt correction, line removal, and the application of filters to reduce noise and enhance the reliability of the results.

A image is composed by units called pixels. The color in each pixel can be expressed through different models. The RGB model is the most used representation approach, where the color is expressed as a combination of three basic colors red, green and blue. However, this representation has some limitations such as concepts such a brightness are difficult to interpret in this model. On the other hand, HSV color model gives a representation more similar to the human perception of the color. This color model defines a color using three characteristics: hue (V, the pure color), saturation (S, purity of the color, from gray tone to a lively color) and value (V, brightness concerning the saturation of the color). HSV model is used in many image processing techniques [8]–[11].

There are different kind of filters used in image processing. The median filter allows filtering preserving edges. The pixel in an square region (neighborhood) is substituted by the median of the valued associated to the pixels of the region [12], [13]. The median filter is usually applied into the three component of the RGB color representation of the image.

This paper proposes an image filtering procedure for atomic force microscopy (AFM) images used in the analysis of the mechanical properties of GVs, based on the application of median filtering in the HSV color space. The proposed approach allows analyzing and filtering, independently, the hue, saturation and value properties of each pixel of the image, enabling the attenuation of image imperfections, including noise and pixels with excessively high or low brightness within localized image regions, while preserving the remaining color information. Tests using actual photos about GVs shows that the proposed image filtering method allows improving the quality of images used in the elasticity analysis of GVs.

II. Theoretical Foundations

A. Gas Vesicles

Gas vesicles (GVs) are structures synthesized by microorganisms in the Archaea and Bacteria domains, such as the cyanobacterial strain *Anabaena flos-aquae*, which can be used as contrast agent in contrast-enhanced ultrasound (CEUS) imaging systems. They consist of a hollow, rigid, all-protein envelope composed primarily of the GvpA and GvpC proteins (which give to the GV its shape and structural rigidity, respectively). Their morphology ranges from fusiform to cylindrical, and the interior of GVs is filled with atmospheric air that circulates freely between the internal and external environments [14]–[16].

GVs can collapse under the influence of specific pressures [5]. These pressures result from evolutionary environment adaptations in which microorganisms develop and depend primarily on the cylindrical radius of their structure [17]. Collapse pressure is directly related to the mechanical rigidity of GVs. Hence, determining properties such as the modulus of elasticity, that is, the measure of resistance relative to the degree of deformation, allows a better understanding of the behavior of GVs when exposed to ultrasound waves.

B. Atomic Force Microscopy (AFM)

Mechanical characteristics of the sample—such as maximum contact force, adhesion, deformation, dissipation, and stiffness (e.g., the elastic modulus)—can be obtained using Atomic Force Microscopy (AFM) in PeakForce Tapping Quantitative NanoMechanics (QNM) mode. This mode performs quantitative nanomechanical mapping—specifically, pixel-by-pixel indentation of the sample by the probe—generating force-displacement curves at each pixel. Cantilever deflection determines the force the sample exerts on the probe relative to its deformation (whether reversible or not), following mass-spring dynamics [18].

AFM enables morphological analysis of GVs through topographic imaging. Topography and mechanical properties can be spatially correlated, allowing for the identification of specific structures—such as the cylindrical shape and capsid proteins—which can be examined in detail [19]. The reliability of the results depends on the quality of the images obtained during the scan [7]; excessive noise, artifacts, and other instabilities can hinder this correlation.

AFM image acquisition can involve artifacts stemming primarily from the probe, the piezoelectric scanner, or the feedback system, as well as issues related to surface contamination and electronic noise [20]. Common examples include probe wear and contamination, piezo nonlinearity, improper feedback system gain settings, acoustic noise, and electrical interference. These factors can cause topographic distortions, errors in determining the contact area between the probe and the sample, and alterations to the force-distance curves used to calculate mechanical properties. Consequently, results may be overestimated or underestimated, compromising the correlation between the observed morphology and the sample's mechanical response [21].

Raw data obtained during AFM scanning are rarely evaluated without further data processing. Mechanical and morphological characteristics can be better defined through image processing. Generally, basic leveling routines are applied to remove tilt errors between the microscope axis and the sample plane [22], alongside line subtraction and matrix filtering. These steps minimize distortions and improve the correlation between topography and mechanical maps.

Image processing can be performed using software integrated to the equipment, as well as commercial or free AFM software featuring user-friendly interfaces such as Gwyddion (a modular, cross-platform software for AFM data analysis) [22]. The software performs quantitative processing of the acquired physical data by operating on the original numerical matrix. Key tools include sample tilt removal, median and Gaussian filtering, frequency-domain analysis, and statistical parameter analysis, among others [22].

C. RGB and HSV Color Models

Different models were created to represent the colors that compose an image. The RGB model represents a color as the combination of the three primary colors red (R), green (G) and blue (B), where the intensity of each color goes from 0 to 255 (or from 0 to 1). Thus, a pixel color (C) can be expressed as a vector composed by the intensities of its primary colors:

$$C = [R,\ G,\ B]. \tag{1}$$

However, human sight does not understand some color characteristics such as luminance through these primary colors. HSV model is a better representation of how the human sight perceives colors. This model defines a color in terms of hue (H), saturation (S) and value (V):

- Hue (H): it indicates the true color (red, yellow, green, etc.), represented as an angle between 0^o too 360^o, or equivalently in the range $[0,\ 1]$ or $[0\%,\ 100\%]$.
- Saturation (S): it indicates the amount of gray. A saturation of 0% means pure gray color, while a saturation of 100% means complete pure vibrant color.
- Value (V): it represents the brightness respect of the saturation. The value 0% means black (independently of the Hue), while 100% means a full brightness.

Fig. 1 shows the HSV color representation. One advantage of using HSV color representation is that the luminance can be analyzed independently from chromatic information. Similarly to (1), the HSV representation can be expressed as a vector:

$$C = [H,\ S,\ V]. \tag{2}$$

Let R, G, and B be the intensity of red, green and blue of a pixel, expressed in the range $[0,\ 1]$. The conversion from RGB model to HSV model is performed as follows [24]:

$$C_{max} = max(R, G, B), \tag{3}$$
$$C_{min} = min(R, G, B), \tag{4}$$
$$\delta = C_{max} - C_{min}, \tag{5}$$

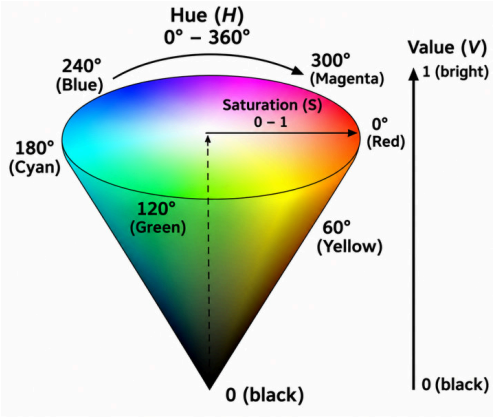


Fig. 1. HSV color map.

$$H = \begin{cases} 0, & \delta = 0; \\ (G-B)/(6\delta), & C_{max} = R \text{ and } C_{min} = B \\ 1 + (R-B)/(6\delta), & C_{max} = G \text{ and } C_{min} = B \\ 2 + (B-R)/(6\delta), & C_{max} = G \text{ and } C_{min} = R \\ 3 + (G-R)/(6\delta), & C_{max} = B \text{ and } C_{min} = R \\ 4 + (R-R)/(6\delta), & C_{max} = B \text{ and } C_{min} = G \\ 5 + (B-G)/(6\delta), & C_{max} = R \text{ and } C_{min} = G, \end{cases} \tag{6}$$

$$S = \begin{cases} 0, & \delta = 0 \\ \delta/C_{max}, & \delta \neq 0, \end{cases} \tag{7}$$

$$V = C_{max}. \tag{8}$$

The results of H, S and V from (6), (7) and (8) are scaled in the range $[0,\ 1]$. The conversion from HSV to RBG representation is done as follows [24]:

$$I = \begin{cases} 0, & 0 \leq 360H < 60 \\ 1, & 60 \leq 360H < 120 \\ 2, & 120 \leq 360H < 180 \\ 3, & 180 \leq 360H < 240 \\ 4, & 240 \leq 360H < 300 \\ 5, & 300 \leq 360H < 360. \end{cases} \tag{9}$$

$$F = 6H - I, \tag{10}$$
$$M = V(1 - S), \tag{11}$$
$$N = V(1 - SF), \tag{12}$$
$$K = V(1 - S(1 - F)), \tag{13}$$

$$[R,\ G,\ B] = \begin{cases} [V,\ V,\ V], & S = 0 \\ [V,\ K,\ M], & I = 0 \\ [N,\ V,\ M], & I = 1 \\ [M,\ V,\ K], & I = 2 \\ [M,\ N,\ V], & I = 3 \\ [K,\ M,\ V], & I = 4 \\ [V,\ M,\ N], & I = 5. \end{cases} \tag{14}$$

D. Median Filter for Images

The median of a dataset is the middle value of the group composed when all the values of the aforementioned dataset are sorted in ascending order.

Let define the $L \times L$ neighborhood of a pixel $P(x, y)$ as the square region with a side length of L pixels, where L is a positive odd number greater than one, and $P(x, y)$ is the center of that region. A median filter for imaging denoising is a non-linear process to suppress impulsive noise while preserving the edges, where the value of the pixel $P(x, y)$ is substituted by the median of the pixels values within its $L \times L$ neighborhood, as illustrated in Fig. 2 [12], [13].

III. Methodology

A. Gas Vesicles

For the cultivation of *Anabaena flos-aquae* (ANA), ASM-1 culture medium is used, and its pH is adjusted to 8.0 after preparation, as described by Gorham et al. [23]. The medium was then autoclaved for subsequent use in the growth of cyanobacteria. The culture was conducted in a room with controlled temperature, aeration, and light conditions, maintained at 24 °C, a 12-hour photoperiod, and a light intensity of 100 μmol photons m^{-2} s^{-1}, for a period of 20 days.

After this period, the culture was transferred to a previously sterilized separation funnel lined on the sides with aluminum foil, allowing light to enter only through the top of the funnel. The system remained in a natural environment for 1 to 3 days, or until a ring of cells formed on the surface. The floating cells were collected in 50 mL Falcon tubes and treated with a lysis solution in a 1:1 (v/v) ratio, consisting of 0.5% Tween 20, 500 mM sorbitol, and 10 mM Tris (pH 7.5), diluted in PBS.

Gas vesicles (GVs) were isolated through three centrifugation steps in Falcon tubes at 400 × g for 24 hours at 4 °C. After centrifugation, the GVs were pooled into a single Falcon tube and stored at 4 °C until use.

B. AFM Image Acquisition

The GVs were initially isolated from the cyanobacterial ANA, following the protocol established by Gorham et al. [23]. Samples with an optical density (OD) of 1.0, as measured by a spectrophotometer (NanoDrop One Microvolume UV-Vis; Thermo Fisher Scientific, Massachusetts, USA) operating at a wavelength of 500 nm, were used.

Purified GVs were diluted 1:10 in distilled water, deposited (150 µL volume) onto a freshly cleaved muscovite mica substrate, and allowed to dry for 24 hours.

Images were acquired using a Bruker Dimension FastScan-Icon Atomic Force Microscope located at the Centro Nacional de Biologia Estrutural e Bioimagem (CENABIO) at the Federal University of Rio de Janeiro. Scanning was performed in air using the PeakForce Quantitative NanoMechanics Mapping mode across four main channels: topography, error, DMT modulus, and adhesion.

During scanning, key parameters were set to a maximum peak force of 300 pN and a scan rate corresponding to a frequency of 0,977 Hz.

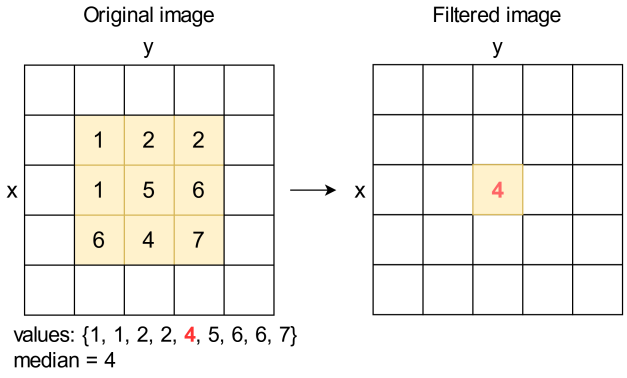


Fig. 2. Median filter for an 3×3 neighborhood.

C. Proposed Filtering Method

An image can be represented as a $M_I \times N_I \times 3$ matrix of pixels, where I_M is the is the number of rows (image height) and N_I is the is the number of columns (image width). Moreover, the number 3 represents the three parameters used to represent the color (R, G, B in RGB system; or H, S, V in HSV system).

The proposed filtering method consists in the application of median filter, with different neighborhood sizes, applied into the HSV representation. Let I the $M_I \times N_I \times 3$ be the matrix that contains the information about hue, saturation and value of each image pixel, obtained by applying the formulas given in (3)–(8) into the RGB color representation of the pixels. Following the matrix notation used in numerical software such as MATLAB and OCTAVE, three $M_I \times N_I$ matrices, denoted as I_H, I_S and I_V, can be defined, as follows:

$$\text{Matrix with Hue information: } I_H = I(:, :, 1), \quad (15)$$

$$\text{Matrix with Saturation information: } I_S = I(:, :, 2), \quad (16)$$

$$\text{Matrix with Value information: } I_V = I(:, :, 3), \quad (17)$$

where $I(:, :, k)$, is a $M_I \times N_I$ matrix composed by the values of the k-th color characteristic of the pixels in all rows and columns of I, where $k = 1$, $k = 2$ and $k = 3$ mean hue, saturation or value, respectively. In the proposed approach, median filters with independent neighborhoods are applied to the matrices $I(:, :, k)$, $k = 1, 2, 3$.

- Hue: $I(:, :, 1)$ gives information about color changes (e.g., edges).
- Saturation: $I(:, :, 2)$ gives information about grayscale in an image.
- Value: $I(:, :, 3)$ gives information about brightness (useful to analyze darker or lighter image artifacts).

Finally, (9)–(14) are used to obtain the RGB color-based image from the filtered hue, saturation, and value information.

IV. Results

The proposed approach was implemented in MATLAB. Two different tests were performed to evaluate the performance of the proposed HSV-based filter.

A. *Test 1*

Fig 3 shows a image of GVs obtained during the process to analyze their elasticity. This image is distorted by horizontal lines and bright points. Figs. 4 and 5 present the RGB and HSV representations of the image used in Test 1. Figs. 4.(b), 4.(c) and 4.(d) show the intensity of the red, green and blue components of the pixels in Fig. 3. Similarly, Figs. 5.(b), 5.(c) and 5.(d) represent the matrices I_H, I_S and I_V that contains the intensities of hue, saturation and value of each pixel in Fig 3, respectively. The noise is more readily identified in the HSV representation than in the RGB representation. Moreover, the hue and saturation representations are noisy, while the value plane is more homogeneous (less noisy).

According to the proposed approach, the following median filters were applied to the matrices I_H, I_S and I_V:

- I_H (hue): median filter with neighborhood $L = 15$.
- I_S (saturation): median filter with neighborhood $L = 31$.
- I_V (value): median filter with neighborhood $L = 9$.

Figs. 6 and 7 show the results obtained by applying a median filter in the the RGB and HSV color representations, respectively. A median filter with $L = 21$ was used in the red, blue and green matrices that compose the RGB image representation. The proposed HSV-based median filtering allows a better noise reduction. Fig. 8 shows the filtered image using the proposed approach, where the horizontal lines and the bright points were significant attenuated.

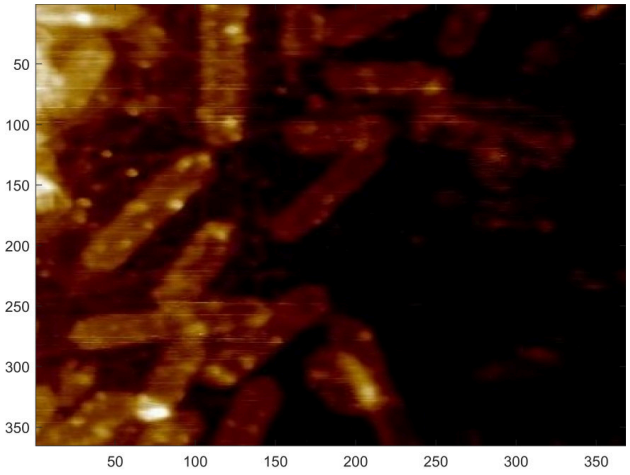


Fig. 3. Original noisy image for test 1.

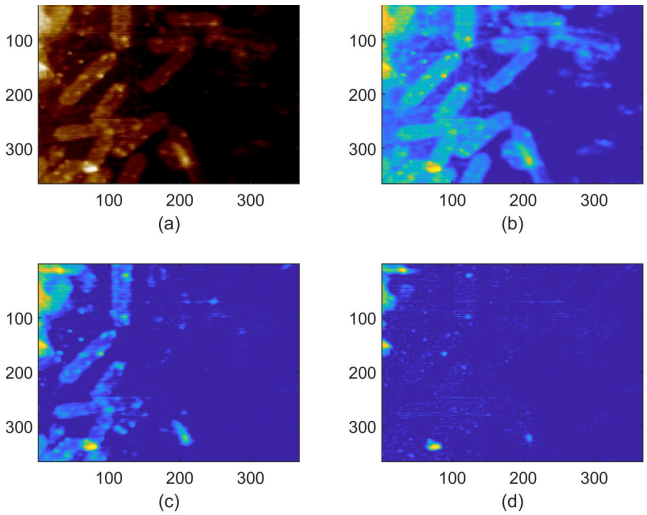


Fig. 4. RGB representation of the noisy image for test 1. a) Original image. b) Red, c) Green, d) Blue.

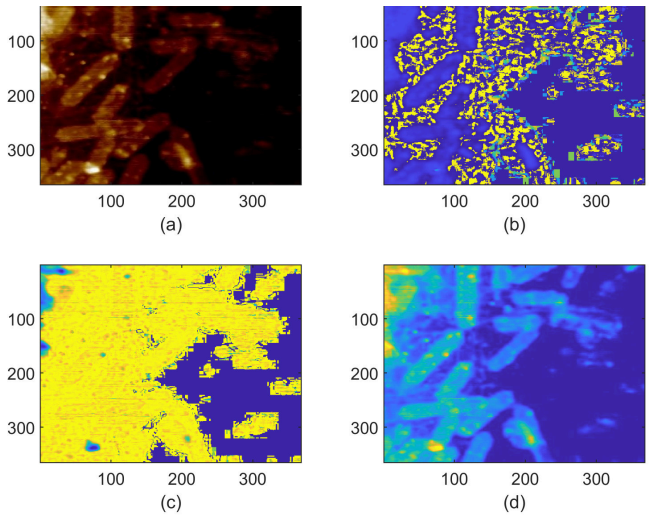


Fig. 5. HSV representation of the noisy image for test 1. a) Original image. b) Hue, c) Saturation, d) Value.

B. *Test 2*

A second test was performed, using the image shown in Fig. 9, obtained through AFM. The color scale in Fig. 9 represents the topography, or height, of the structure. Lighter colors indicate greater height, as this is a topographic image. Thus, the GV is located at a higher elevation relative to the background plane. Consequently, filtering colored noise and other image artifacts is important for accurate topographic analysis of GVs.

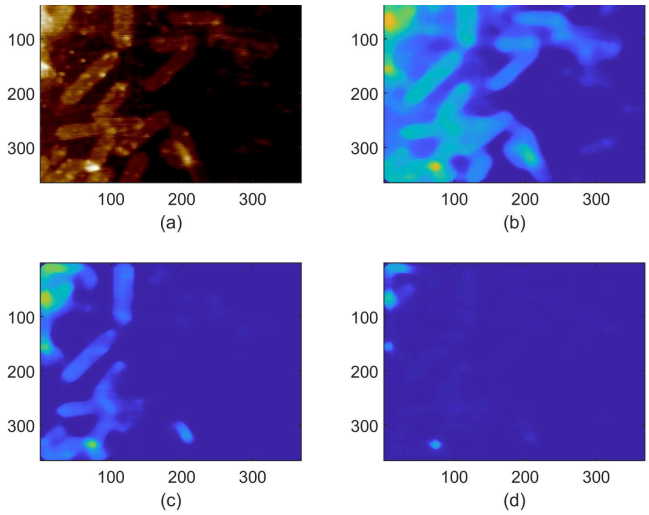


Fig. 6. Image obtained by applying the median filter in the RGB representation. a) Filtered image. b) Red, c) Green, d) Blue.

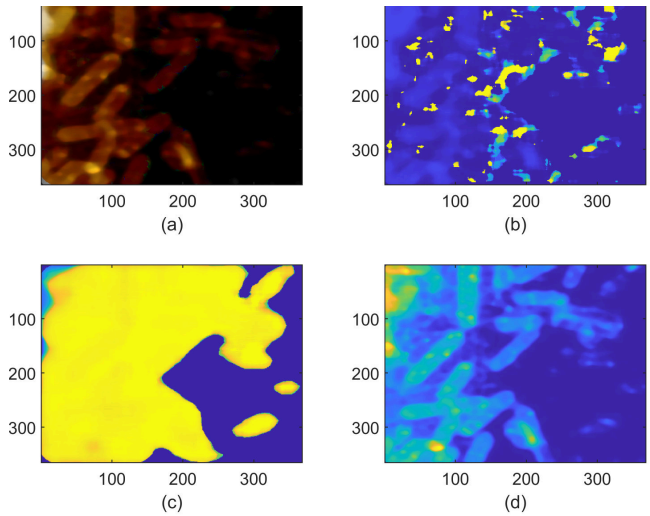


Fig. 7. Image obtained by applying the median filter in the HSV representation. a) Filtered image. b) Hue, c) Saturation, d) Value.

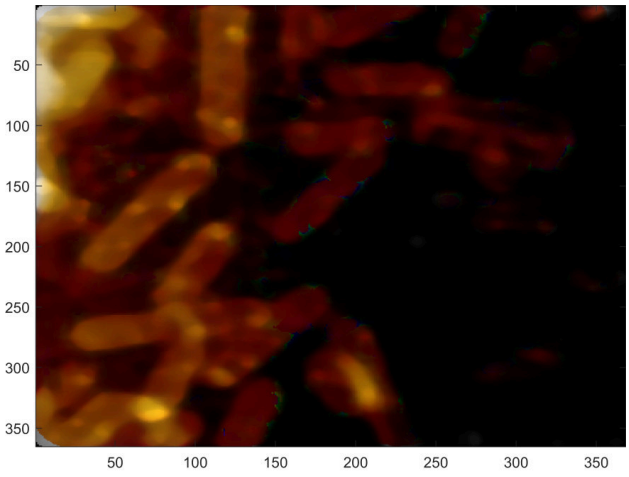


Fig. 8. Filtered image using the proposed approach.

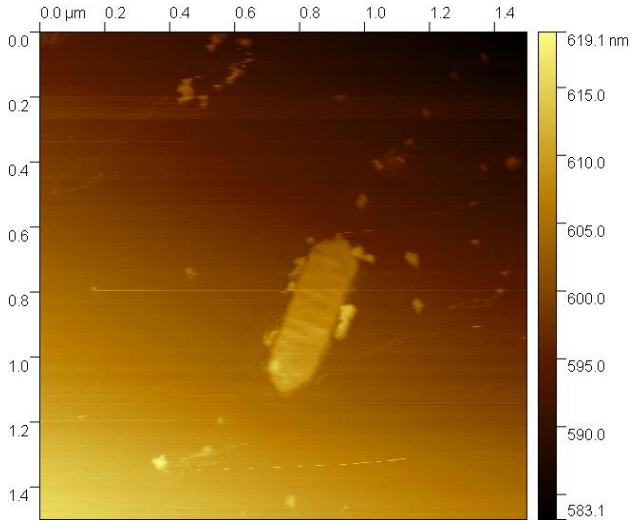


Fig. 9. Original image used in Test 2.

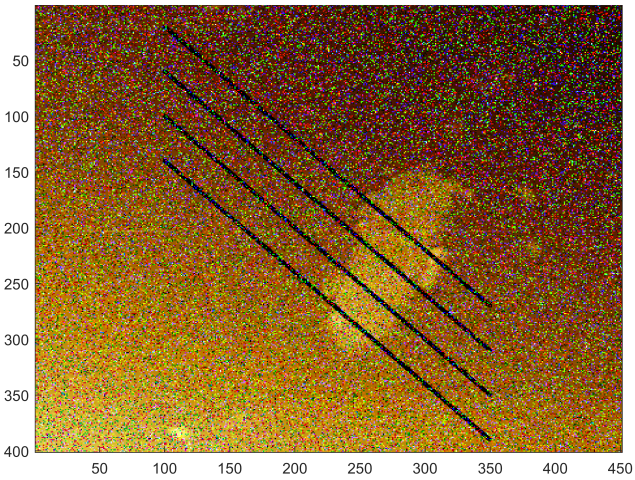


Fig. 10. Distorted image used in Test 2.

To test the proposed filtering process, black diagonal bars and salt-and-pepper noise with a noise density of 0.2 (using the command *imnoise(I2,'salt & pepper', 0.2)*), being *I2* the matrix with the pixels of the image) where added to the image, as shown in Fig. 10.

The structural similarity index (SSIM) was used to measure the similarity between the original image in Fig. 9 with a filtered image. SSIM can be used to compare visual quality based on three components: mean, variance, and cross-correlation [26]. The command *'ssim(Ifiltered, Ioriginal)'* of MATLAB was used to get the SSIM value, between the filtered and the original images (represented by the matrices *Ifiltered* and *Ioriginal*, respectively. Table I shows the SSIM values for different values of L_h, L_s and L_v. The best result was $SSIM = 0.9955$ for $L_h = 11$, $L_s = 11$ and $L_v = 9$.

Fig 11 shows the filtered images obtained through the proposed filter process with the best configuration in Table I ($L_h = 11$, $L_s = 11$ and $L_v = 9$), and using a median filter applied to the RGB representation (with $L = 11$ for the three sets of primary colors). The filtered image using RGB color representation has a $SSIM = 0.98785$. Therefore, the proposed filter method (with $SSIM = 0.9955$) provides a filtered image more similar to the original image, allowing a more accurate topographic analysis of GVs. Moreover, the SSIM values for the proposed HSV-based filtering process in Table I are greater than the SSIM value for the RGB-based filter.

On the other hand, Table I also indicates that increasing the neighborhood size (L) of the median filter will not necessarily improve the image quality. For example, the SSIM value for $L_h = 13$, $L_s = 11$ and $L_v = 15$ is 0.99319, which is less than the the result obtained for $L_h = 11$, $L_s = 11$ and $L_v = 9$ ($SSIM = 0.9955$). Optimization techniques, such as a genetic algorithm, can be used to get the optimal values for the median filter neighborhoods. A genetic algorithm is an optimization method inspired by the principles of biological evolution and natural selection. [27].

TABLE I
$SSIM$ FOR DIFFERENT VALUES OF L_h, L_s AND L_v

L_h	L_s	L_v	SSIM
11	11	15	0.9935
13	11	15	0.99319
11	13	15	0.99292
11	11	17	0.99259
9	11	15	0.99325
7	11	15	0.99062
15	11	15	0.99314
13	9	15	0.99317
11	9	15	0.99358
11	7	15	0.99196
11	9	13	0.99433
11	9	11	0.99532
11	9	9	0.99547
11	9	7	0.99312
11	11	13	0.99403
11	11	11	0.99496
11	11	9	0.9955
11	11	7	0.99281

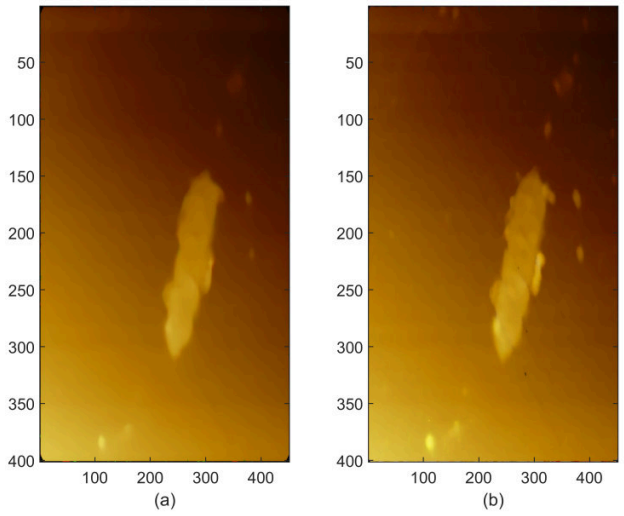


Fig. 11. Filtered images using median filters applied into RGB and HSV color representation: a) Using RGB representation ($SSIM = 0.98785$). b) Using HSV representation ($SSIM = 0.9955$).

V. CONCLUSION

This paper proposed a new image filtering technique for GV analysis based on median filtering applied to the HSV representation of the image. The use of HSV color representation allowed the identification and attenuation of noise and other artifacts in hue, saturation and brightness, which is more intuitive and simple to understand than working with primary colors. The proposed filter has simple mathematical structure, being simple to implement in software (or in embedded processors if necessary). Therefore, the proposed approach could be part of more complex image enhancement software used in the analysis of GVs. The results in Table I demonstrated that the proposed filtering method can provide better results than using a median filter applied to the RGB color representation. As future work, a genetic algorithm (GA) will be used to obtain optimal values of the median filter neighborhoods required in the proposed approach.

ACKNOWLEDGMENT

The authors acknowledge the Ultrasound Laboratory (LUS) and the Biomedical Engineering Program at COPPE/Federal University of Rio de Janeiro, for providing the infrastructure, facilities, and technical support essential to the completion of this study.

REFERENCES

[1] C. Moran and A. Thomson, "Preclinical Ultrasound Imaging—A Review of Techniques and Imaging Applications," *Front. Phys.*, vol. 8, Art. no. 124, 2020.
[2] M. A. Borden and K.-H. Song, "Reverse engineering the ultrasound contrast agent," *Adv. Colloid Interface Sci.*, vol. 262, pp. 39–49, Dec. 2018.
[3] J. Wang *et al.*, "Nanoscale contrast agents: A promising tool for ultrasound imaging and therapy," *Adv. Drug Deliv. Rev.*, vol. 207, Art. no. 115200, 2024.
[4] H. Yusefi and B. Helfield, "Ultrasound contrast imaging: Fundamentals and emerging technology," *Front. Phys.*, vol. 10, Art. no. 791145, 2022.
[5] M. G. Shapiro *et al.*, "Biogenic gas nanostructures as ultrasonic molecular reporters," *Nat. Nanotechnol.*, vol. 9, no. 4, pp. 311–316, 2014.
[6] Y. M. Efremov, T. Okajima, and A. Raman, "Measuring viscoelasticity of soft biological samples using atomic force microscopy," *Soft Matter*, vol. 16, no. 1, pp. 64–81, 2020.
[7] Rico, F. et al. May the Force Be with You! Force–Volume Mapping with Atomic Force Microscopy. *ACS Meas. Sci. Au*, 2021.
[8] J. He, H. Luo, X. Cai, and Z. Gao, "Cross-Modal Fusion With UVI Enhancement for Salient Object Detection in Low-Light ORSI Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, Art. no. 5628916, 2026.
[9] X. Liu, H. Xu, Y. Ju, S. Wang, C. Liu, and L. Chen, "DA-GAN: Dual-Attention GAN for Underwater Image Enhancement With Contrast and Color Correction," *IEEE Trans. Geosci. Remote Sens.*, vol 64, pp. 1–16, Art. no. 4200816, 2026.
[10] A. Agrawal and H. P. Padole, "A Novel Integrated Graph-Based Self-Supervised Image Segmentation With Application to Medical X-Ray Images," *IEEE Sensors J.*, vol. 25, no. 16, pp. 31078–31085, Aug. 2025.
[11] J. Zhang, H. Su, T. Zhang, H. Tian, and B. Fan, "Multi-Scale Fusion Underwater Image Enhancement Based on HSV Color Space Equalization," *Sensors*, vol. 25, Art. no. 2850, 2025.
[12] J. D. Mendiola-Santibañez et al., "Structural Image Recovery Based on Connectivity for Extreme (90%) Salt and Pepper Noise Removal," IEEE Access, vol. 14, pp. 101453-101476, 2026.
[13] M. M. Yahaya, P. Kumam and K. Sitthithakerngkiet, "Denoising Colour Images Corrupted by Salt-and-Pepper Noise: Weighted Mean Filter With Iterative Approach and Deep Learning Modeling," *IEEE Access*, vol. 14, pp. 22661-22677, 2026.
[14] A. E. Walsby, "Gas vesicles," *Microbiol. Rev.*, vol. 58, no. 1, pp. 94–144, 1994.
[15] F. Pfeifer, "Recent advances in the study of gas vesicle proteins and application of gas vesicles in biomedical research," *Life*, vol. 12, no. 9, Art. no. 1455, 2022.
[16] S. T. Huber *et al.*, "Cryo-EM structure of gas vesicles for buoyancy-controlled motility," *Cell*, vol. 186, no. 5, pp. 975–986.e13, 2023.
[17] A. E. Walsby, "The mechanical properties of the *Microcystis* gas vesicle," *J. Gen. Microbiol.*, vol. 137, no. 10, pp. 2401–2408, Oct. 1991.
[18] Y. F. Dufrêne, D. Martínez-Martín, I. Medalsy, D. Alsteens, and D. J. Müller, "Multiparametric imaging of biological systems by force-distance curve–based AFM," *Nat. Methods*, vol. 10, no. 9, pp. 847–854, Sep. 2013.
[19] T. J. McMaster, M. J. Miles, and A. E. Walsby, "Direct observation of protein secondary structure in gas vesicles by atomic force microscopy," *Biophys. J.*, vol. 70, no. 5, pp. 2432–2436, May 1996.
[20] N. Starostina and P. West, "A Guide to AFM Image Artifacts." [Online]. Available: https://www.physik.uni-wuerzburg.de/fileadmin/11000000/03_Studium/03_Master/F-Praktikum/2020/AFM/AFM_Bild-Artefakte.pdf. Accessed: Jul. 22, 2026.
[21] C. Canale *et al.*, "Recognizing and avoiding artifacts in atomic force microscopy imaging," *Methods Mol. Biol.*, vol. 736, pp. 31–43, 2011.

[22] D. Nečas and P. Klapetek, "Gwyddion: An open-source software for SPM data analysis," *Cent. Eur. J. Phys.*, vol. 10, no. 1, pp. 181–188, Feb. 2012.
[23] P. R. Gorham *et al.*, "Isolation and culture of toxic strains of *Anabaena flos-aquae* (Lyngb.) de Bréb," *SIL Proc.*, vol. 15, no. 2, pp. 796–804, Feb. 1964.
[24] V. Chernov, J. Alander, and V. Bochko, "Integer-based accurate conversion between RGB and HSV color spaces," *Comput. Electr. Eng.*, vol. 46, pp. 328–337, Aug. 2015.
[25] T. S. Huang, G. J. Yang, and G. Y. Tang, "A Fast Two-Dimensional Median Filtering Algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 1, pp. 13–18, Feb. 1979.
[26] D. M. Rouse and S. S. Hemami, "Understanding and simplifying the structural similarity metric," 2008 15th IEEE Int. Conf. Image Process., San Diego, CA, USA, 2008, pp. 1188–1191.
[27] T. v. Khanh, L. M. Thanh and N. N. Thuy, "A Feature-Refinement and Similarity-Learning Fusion Framework for Remote Sensing Image Retrieval," *IEEE Access*, vol. 14, pp. 94413–94433, 2026.

Geometric Characterization for Personalized Face-Frame Interface on Assessing Sunglasses UV Blocking

1st Alvaro Costa Jr
Ophthalmic Instrumentation Lab. (LIO)
University of Sao Paulo (USP)
Sao Carlos-SP, Brazil
alvarojunior@usp.br

2nd Pedro da Silva
Ophthalmic Instrumentation Lab. (LIO)
University of Sao Paulo (USP)
Sao Carlos-SP, Brazil
pedrotxs@usp.br

3rd Mauro Masili
Ophthalmic Instrumentation Lab. (LIO)
University of Sao Paulo (USP)
Sao Carlos-SP, Brazil
mmasili@gmail.com

4th Liliane Ventura
Ophthalmic Instrumentation Lab. (LIO)
University of Sao Paulo (USP)
Sao Carlos-SP, Brazil
lilianeventura@usp.br

***Abstract*—Exposure to light and solar radiation can cause cumulative damage to ocular structures, making sunglasses an essential means of protection. However, their effectiveness depends not only on the optical properties of the lenses, but also on the ability of the frame to limit peripheral light entry through gaps between the accessory and the user's face. Current scientific literature predominantly focuses on lens transmittance, with limited quantitative evaluation of frame coverage and facial morphology. To address this gap, this study proposes an experimental methodology based on Digital Image Processing to evaluate the face–frame interface. Frame segmentation employs the YOLO11 model, trained with annotations generated by SAM3. Frontal analysis delineates facial regions using MediaPipe and HSV filtering, while profile analysis employs the YuNet detector with adaptive geometry in the lateral view, isolating exposed skin via bitwise boolean AND operations with the frame mask. The results demonstrate promising frame detection and anatomical mapping. Frontal facial landmarks isolate exposed skin, whereas profile analysis delineates the topology of the unshielded area, revealing spatial variations associated with individual anatomy. The methodology demonstrates potential to identify and quantify the gap dimensions between the frame and the face, establishing the structural baseline to enable future metric conversion to millimeters and subsequent normative comparison with the standardized size-M headform specified in ISO 18526-4:2022.**

***Index Terms*—sunglasses, UV protection, Digital image processing, Facial segmentation, Peripheral radiation**

I. Introduction

Exposure to Ultraviolet Radiation (UVR) constitutes a significant biological hazard to ocular health. The incident levels of UVR are contingent upon environmental variables, specifically altitude, latitude, and ground reflection [1] [2]. These radiation levels are capable of inducing both photochemical and thermogenic damage to various ocular structures. Furthermore, prolonged exposure leads to a cumulative effect, which subsequently manifests as severe pathologies, including but not limited to pterygium, macular degeneration, photokeratitis, and photoretinitis [3] [4].

Biological tissue damage is intrinsically linked to the selective absorption of UVA and UVB radiation, with the cornea and the crystalline lens functioning as natural protective filters [5]. Although the human organism employs physiological defense mechanisms, such as pupillary contraction (known as Miosis) and the squint reflex, these inherent processes are insufficient to confer total ocular protection [6]. Consequently, human intervention, specifically through the implementation of sunglasses, becomes essential.

This intervention serves as a necessary complementary barrier to effectively reduce the overall UVR dose received by the eyes [7].

Despite the established importance of wearing sunglasses, the existing literature presents a limited number of experimental analyses concerning the influence of facial geometry on the efficacy of this form of ocular protection. This gap is relevant because individual facial characteristics and cephalic morphological differences may inherently determine variations in the manner and extent to which solar radiation reaches the human eye [8].

Consequently, given the scarcity of studies investigating radiation penetration through the gaps between sunglass frames and the face under realistic usage conditions, this study proposes an integrated experimental methodology to evaluate and inform users about the true shielding capability of eyewear. In this context, visible light serves as a geometric proxy to map peripheral penetration pathways. Because ultraviolet radiation and visible light share identical spatial propagation trajectories, mapping the gaps through which visible light penetrates provides a direct indicator of anatomical regions

Supported by FAPESP and CNPq

vulnerable to peripheral radiation influx.

To achieve this objective, the proposed framework leverages Digital Image Processing (DIP) techniques for the automated characterization of facial anatomy and frame geometry using synchronized frontal and profile images. The system integrates a controlled-illumination acquisition setup with automated Region of Interest (ROI) detection and image segmentation, enabling individualized characterization of both facial structures and sunglass frames.

A. UVR Exposure Limits

The effective radiant exposure to the unprotected eye within the ultraviolet spectrum (180–400 nm) is restricted to a maximum threshold of 30 J/m² over an eight-hour period. Standard guidelines categorize ultraviolet radiation (UVR) into three distinct bands: UV-C (180–280 nm), UV-B (280–315 nm), and UV-A (315–400 nm). Wavelengths below 180 nm are excluded from hazard assessments, as atmospheric absorption renders their ground-level biological impact negligible. Consequently, while UV-C radiation is almost entirely absorbed by the stratospheric ozone layer, UV-B and UV-A reach the Earth's surface with sufficient intensity to penetrate ocular tissues [9].

Quantitatively, International Commission on Non-Ionizing Radiation Protection (ICNIRP) specifies that the daily effective radiant exposure (H_{eff}), spectrally weighted by the biological hazard function $S(\lambda)$, must not exceed 30 J/m^2 (3.0 mJ/cm^2) for actinic UV radiation (180–400 nm) over an 8-hour period [10]. Furthermore, for the UV-A spectrum (315–400 nm), the unweighted cumulative radiant exposure is capped at 10^4 J/m^2 (1.0 J/cm^2) to mitigate long-term lenticular damage [9]. These quantitative parameters serve as essential safety baselines for evaluating ocular exposure risks and defining sunglass shielding requirements.

In addition to ICNIRP recommendations, other studies provide further insight into UVR exposure thresholds. This is illustrated by the research conducted by [10], which translates normative metrics into practical measurements of radiation doses capable of causing harm. The authors evaluated the effectiveness of various types of protective wear, including sunglasses, caps, and hats, by comparing the exposure of protected and unprotected eyes.

The investigation revealed substantial variation in the efficacy of these devices, demonstrating that diffuse and effective radiation remains a persistent risk. The results indicated that increasing the distance between sunglasses and the forehead by only 6 mm altered the percentage of UVR reaching the eyes, producing values ranging from 3.7% to 44.8%. Consequently, such findings suggest that, in addition to the protective capability of the lenses, geometric factors and frame design also influence the level of protection [10].

B. Ocular UVR Exposure and Attenuation

The primary purpose of sunglasses is to provide visual comfort by attenuating the flux of visible light and acting as filters against ultraviolet radiation (UVR). Ocular protection has traditionally focused on the attenuation of the direct UVR component, largely because most research in the field concentrates on lens transmittance. Consequently, it is equally important to evaluate the impacts of diffuse and reflected radiation. [11].

Effective protection, however, requires attention beyond the filtering capacity of the lenses. The UVR that reaches the eye includes significant indirect components produced by atmospheric scattering and by reflection from surfaces such as the ground, snow, and water [12]. This diffuse and reflected radiation strikes the peripheral regions of the eye and creates penetration pathways that are not attenuated by the lenses.

When the geometric contribution and the design of the frames are considered, the frame can be understood as an additional protective barrier. The main challenge arises from the lateral and superior gaps between the frame and the face. These openings allow peripheral rays to reach the crystalline lens, establishing connections between cortical cataracts and the limbal focus of peripheral radiation, a phenomenon known as the Coroneo Effect [13].

For this reason, sunglasses that lack adequate lateral and superior barriers may pose a risk to ocular health. This demonstrates that, beyond the filtering capacity of the lenses, the design and positioning of the frame are essential for effective protection.

II. MATERIALS AND METHODS

A. Experimental Acquisition System and Setup

For a metrological evaluation, an image acquisition system was developed in a controlled lighting environment, whose physical configuration and experimental setup are illustrated in Fig. 1. The primary light source consisted of a cold white light panel featuring 240 SMD 2835 LEDs, with a color temperature of 6000K, powered by a stabilized power supply to prevent luminosity fluctuations. This 12V LED array was uniformly distributed under a polycarbonate panel. To ensure homogeneous illumination and minimize specular reflections and localized saturation on the lenses of the sunglasses, a high-transmittance white diffuser fabric was positioned over the panel, providing uniform light diffusion throughout the capture area.



(a) subject positioning on the ophthalmic chin rest (lateral view).

(b) orthogonal dual-camera configuration on the optical bench (perspective view).

Fig. 1: Experimental image acquisition setup.

The image acquisition system employed two webcam-type cameras (model WC577), rigidly mounted on an optical bench at a fixed distance of 40 cm from the subject's facial plane. These cameras operated at an effective resolution of 2560 x 1440 pixels and a field of view (FOV) of 110 degrees, with captures performed using the cameras' native operating settings. The optical axes of both cameras were geometrically aligned to ensure acquisition repeatability across different users from standardized viewpoints. Specifically, the frontal camera was positioned at the level of the nasal region, whereas the lateral camera was oriented perpendicularly to the zygomatic region, enabling precise recording of the lateral profile.

To ensure measurement repeatability and minimize the positioning variability inherent to the subject, an adjustable ophthalmic chin rest, originally designed for clinical ophthalmic diagnostic instruments and adapted for the present experimental setup, was incorporated into the optical bench. This device restricts involuntary head movements and maintains a constant position relative to the calibrated focal plane throughout the acquisition.

The images used in this study were acquired with the voluntary participation of members of the Ophthalmic Instrumentation Laboratory research group, where the proposed project was developed. All participants were previously informed about the objectives of the study and provided authorization for the use of their images exclusively for scientific purposes, including the development, validation, and dissemination of the results presented in this paper.

B. Sunglasses Segmentation

Geometric characterization of sunglass frames constitutes a fundamental step in the proposed methodology, as it enables the isolation of the accessory from the remainder of the image and provides an accurate representation of its shape and contours. This information is essential for subsequent stages involving the extraction of geometric parameters and for the analysis of the spatial relationship between the frame and facial anatomy, which will serve as the basis for ultraviolet radiation blocking studies. To automate this process, the SAM3 and YOLO11 tools were employed, whose functions are complementary throughout the construction and application of the segmentation model.

The Segment Anything Model 3 (SAM3) is an instance segmentation architecture developed to perform pixel-level segmentation from various types of prompts, including bounding boxes, points, textual descriptions, and visual examples. Compared with conventional segmentation approaches, the model produces binary masks that describe the contours of the objects of interest with high precision, preserving their geometric representation [14].

In this project, SAM3 was employed in the automatic annotation stage of the dataset. Images containing sunglass frames were processed to generate binary masks separating the frame from the background. These masks were then used to extract frame contours and automatically convert them into the annotation format required by YOLO11, establishing a direct geometric correspondence for supervised training while significantly reducing manual labeling effort.

The YOLO (You Only Look Once) architecture constitutes an approach to real-time object detection in the context of computer vision. Its primary objective is to identify and simultaneously localize multiple objects present in images or video sequences, reconciling high detection accuracy with low inference time. YOLO possesses capabilities beyond object detection, offering native support for instance segmentation, image classification, pose estimation, and oriented bounding box detection [15].

The dataset for training and evaluating the model consisted of 4 voluntary participants, all members of the Ophthalmic Instrumentation Laboratory (LIO) where this research was developed, utilizing 80 distinct sunglass models. This procedure yielded a total of 249 images captured under controlled lighting conditions. To assess the model's generalization capability, the images were divided into proportions of 80% for training (199 images) and 20% for validation (50 images), ensuring the representativeness of facial variations and the wide diversity of frame geometries.

During training, the model receives the images and their respective annotations, learning to automatically localize and segment the sunglass frames from the provided examples. At the end of each epoch, its performance was evaluated using the validation set and the set of weights corresponding to the best result obtained was automatically stored. The final model was saved as the best.pt file, which was subsequently used in the inference stage to perform automatic detection and segmentation of the frames in new images.

Fig. 2 presents the methodological flow employed in the construction of the segmentation model, from the automatic generation of annotations by SAM3 to the obtainment of the trained model (best.pt).

To quantitatively evaluate the segmentation accuracy of the trained YOLO11 model against reference annotations, standard overlap metrics, specifically Intersection over Union (IoU) and the Dice Similarity Coefficient (Dice), were computed on an independent test subset. These widely established metrics quantify spatial agreement and boundary fidelity by evaluating pixel-level overlap between binarized predicted frame masks and ground truth representations.

C. Frontal Facial Segmentation

Segmentation of facial structures constitutes a complementary step to sunglass frame segmentation, enabling the delineation of anatomical regions of interest used in characterizing the interface between the face and the accessory. Such regions can provide geometric information that allows analyzing the proximity between the frame and the facial anatomy, serving as a basis for identifying areas potentially susceptible to ultraviolet radiation incidence.

Automatic facial landmark estimation is a widely employed technique in computer vision for geometrically representing the human face, enabling the localization of characteristic points used in facial analysis applications [16]. In this project,

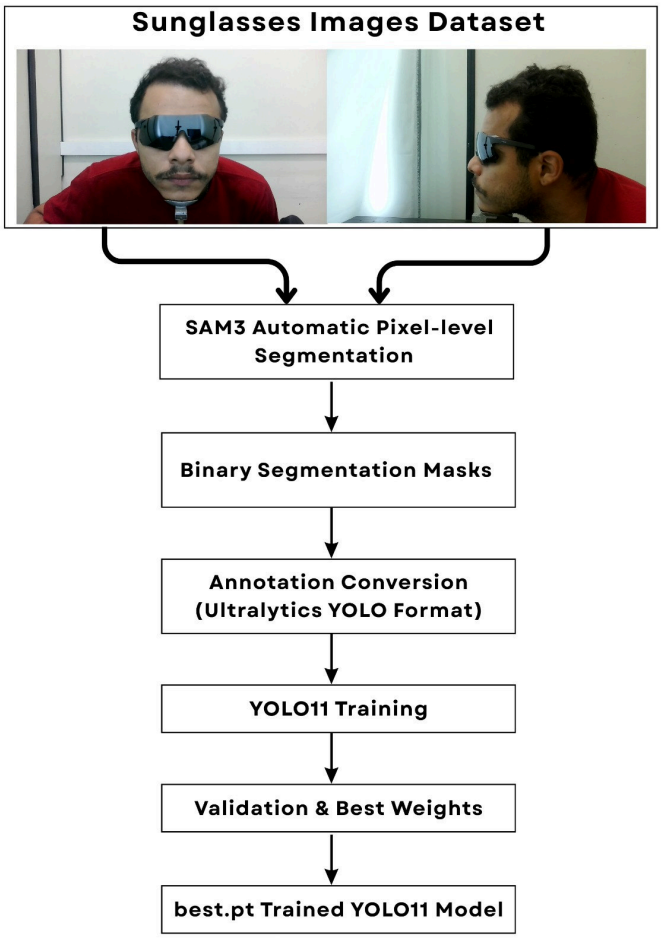


Fig. 2: Proposed SAM3 and YOLO11 segmentation pipeline workflow.

this step was implemented using the MediaPipe Face Landmarker library, which performs automatic detection of these anatomical landmarks and generates the facial mesh employed in defining the regions of interest.

From the selected landmarks, an initial geometric representation of the anatomical regions is constructed using the Convex Hull algorithm, which generates a polygon enclosing the points belonging to each facial structure. For the zygomatic regions, this initial representation is refined through a hybrid strategy based on subtracting the sunglass frame mask, automatically removing the areas covered by the accessory and preserving only the effectively visible facial regions.

Subsequently, the contours of the resulting regions are extracted, and a filter based on the minimum area of the objects is applied to eliminate small spurious regions arising from processing. The final contours and the facial landmarks are then overlaid onto the original image, producing the geometric representation employed in the analyses presented in the results section.

Fig. 3 presents the block diagram of the methodology employed for frontal facial segmentation, from facial landmark detection to the obtainment of the regions of interest used in the geometric characterization.

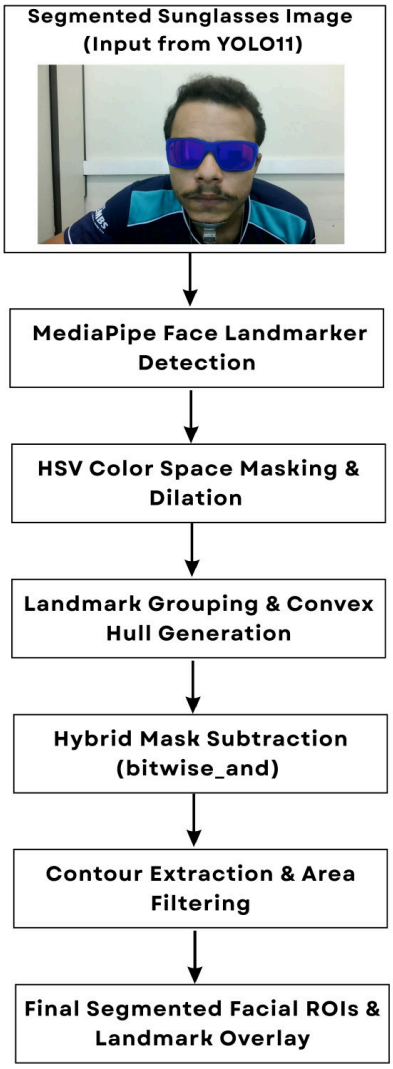


Fig. 3: Block diagram of the proposed frontal facial segmentation pipeline

D. Profile Exposed Facial ROI Segmentation

Although frontal facial segmentation enables the delimitation of different anatomical structures through a dense facial mesh, this approach presents limitations when applied to profile images, in which part of the anatomical landmarks becomes invisible due to facial self-occlusion. Given this limitation, a specific methodology was developed for the profile view, based on the combination of automatically detected anatomical points and adaptive relationships for defining the region of interest.

Face detection was performed using the YuNet Face Detector, which locates the face and estimates the main visible anatomical points, including the eye, nose, and mouth.

From the face bounding box and the coordinates of the detected anatomical points, proportional geometric relationships are calculated and used in constructing the region of

interest. These relationships enable the automatic definition of an adaptive polygon corresponding to the zygomatic region, adjusting its position and dimensions to the individual characteristics observed in each profile image. Subsequently, the image is converted to the HSV color space to retrieve the mask corresponding to the previously segmented sunglass frame, which is then dilated to create a safety margin during processing.

After constructing the initial geometry, the frame mask is subtracted from the region of interest, preserving only the effectively visible facial areas. Subsequently, morphological closing and opening operations are applied using an elliptical structuring element, promoting contour smoothing and reducing irregularities introduced by polygon discretization. Finally, the final contours of the segmented region are extracted, and the region of interest and the detected anatomical points are overlaid onto the original image.

It is worth noting that this profile processing stage specifically targets the delineation of the exposed zygomatic skin area (ROI) directly susceptible to peripheral radiation influx. Rather than directly measuring a single linear gap, the adaptive polygon isolates the unoccluded tissue boundary formed between the lower edge of the frame temple and the cheek contour. This segmented exposed mask acts as a spatial foundation that enables subsequent quantitative extraction of the lateral gap distance.

III. RESULTS

A. Sunglasses Detection and Segmentation Results

The presented results demonstrate that the developed pipeline is not restricted to a single frame shape or to a specific set of facial characteristics. The correct identification of the regions of interest constitutes a fundamental requirement for the subsequent stages of the methodology, since the extraction of geometric information depends directly on the precision with which the frame is segmented relative to the remainder of the image.

The segmentation pipeline was evaluated across the complete independent test set ($n = 50$), achieving a global mean IoU of 0.9044 and a Dice score of 0.9488. Table I highlights a representative subset of 10 evaluation cases to illustrate boundary fidelity across distinct viewing angles and frame geometries, confirming strong spatial accuracy across diverse user poses. Spatial distribution of segmentation fidelity is illustrated in Fig. 4. The pixel classification map shows that true positives, represented by green pixels, dominate the target region. Discrepancies such as false positives (red) and false negatives (blue) are strictly confined to boundary transition zones.

Angle-specific evaluation indicates higher mean IoU values in frontal views (exceeding 0.93) compared to profile acquisitions, where lower values were recorded (such as 0.8031 for sample 8). Across all evaluated perspectives, IoU values remained above 0.80, establishing the segmentation baseline for subsequent facial ROI extraction.

TABLE I: Quantitative evaluation of the YOLO segmentation model on unseen data.

Sample	View	IoU	Dice Score
1	Frontal	0.9509	0.9749
2	Lateral	0.8096	0.8948
3	Lateral	0.8966	0.9455
4	Lateral	0.9189	0.9578
5	Frontal	0.9547	0.9768
6	Frontal	0.9644	0.9819
7	Frontal	0.9444	0.9714
8	Lateral	0.8031	0.8908
9	Frontal	0.9544	0.9767
10	Lateral	0.8472	0.9173
Mean		**0.9044**	**0.9488**

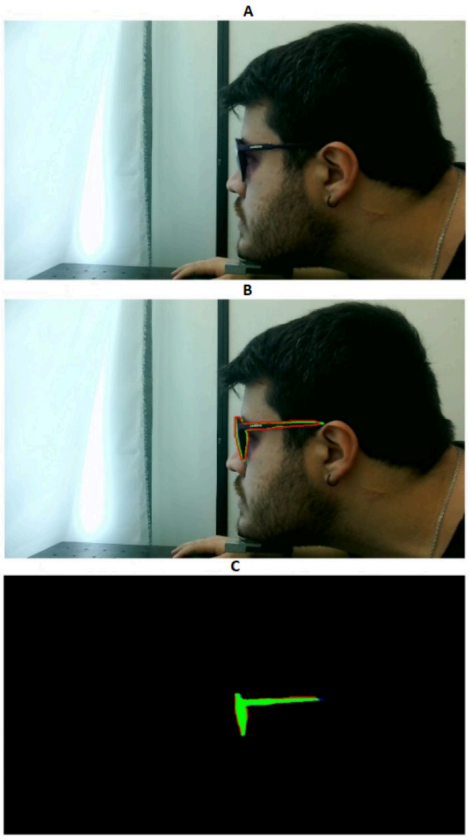


Fig. 4: Qualitative spatial evaluation of segmentation accuracy: (A) original profile acquisition, (B) ground truth (green) and predicted YOLO11 (red) boundary overlay, and (C) pixel classification map highlighting true positives (green), false positives (red), and false negatives (blue).

Although the previous results demonstrate the model capability to detect different frames, an important issue consists of evaluating the behavior of the same frame model when used by different individuals. For this analysis, a round-frame model was selected, characterized by circular and flat lenses with reduced dimensions, which naturally leave uncovered the lateral regions near the eyes.

As observed in Fig. 5, even when using the same frame, differences are verified in the distribution of the spaces existing

between the face and the sunglasses. Such variations may be associated with the facial characteristics of each user, as well as with minor differences in frame positioning during use. Consequently, lateral regions remain more or less exposed depending on the combination between the frame geometry and the facial anatomy.



Fig. 5: Frame-to-face spatial variation assessment using identical round sunglasses across multiple users.

The results presented in Fig. 5 evidence that the geometric characterization of the interface between the frame and the face must be performed for each condition of use. Assuming that the same frame model presents the same fit for all users may lead to inaccurate interpretations, since the distribution of the gaps depends both on the facial anatomy and on the manner in which the accessory is worn. In this way, it becomes possible to individually analyze the regions potentially more susceptible to radiation incidence through the openings existing between the frame and the face.

B. Frontal Facial ROI Segmentation Results

The MediaPipe Face Landmarker detected the anatomical landmarks in the volunteers analyzed. The facial mesh remained stable under different lighting conditions and in the presence of facial hair. From these points, the Convex Hull applied to the nasal region delineated the nose structure using red polygons, without encroaching upon adjacent areas.

The integration of HSV chromatic filtering with a bitwise AND operation cropped the upper cheek boundaries flush with the lower edge of the blue frame (derived from SAM3-YOLO segmentation). As illustrated in Fig. 6, the zygomatic region was effectively restricted to visible skin. However, the lower cheek boundary varied according to individual mandibular anatomy and peripheral landmark distribution. Because the lower ROI contour relies on linear connections between landmarks, minor alignment discrepancies occur relative to the natural curvature of the jawline. Refining this mandibular contouring via non-linear boundary fitting represents an opportunity for future algorithmic improvement.

This approach eliminates accessory overlap over the skin surface, establishing a precise geometric delimitation of the non-occluded skin. Consequently, the extracted topological data provide a promising basis for calculating the area exposed to ultraviolet radiation in subsequent stages.



Fig. 6: Frontal view region of interest (ROI) segmentation results.

C. Profile Facial ROI Segmentation Results

For the profile view analysis, detection of reference points (eye, nose tip, and mouth corner) by the YuNet detector served as anchoring to delimit the lateral zygomatic region, overcoming the self-occlusion limitation inherent to this viewing angle. Delimitation of this area is based on the hypothesis that the cheek constitutes a zone susceptible to peripheral solar radiation incidence. The adaptive polygon adjusted its dimensions to the craniofacial characteristics of each photographed face, maintaining algorithmic stability.

Integration of the sunglass mask via the bitwise AND operation prevented the demarcation from extending over the accessory. As presented in Fig. 7, the upper red boundary precisely follows the lower contour of the blue frame. Although all volunteers wore the same frame model, the results evidence subtle variations in the peripheral gaps between the frame and the anatomy of each face.

The profile segmentation results depicted in Fig. 7 highlight the exposed zygomatic skin area remaining unshielded by the eyewear geometry. The upper boundary of the red polygon strictly conforms to the lower contour of the blue frame mask, while the lower and posterior edges correspond to the adaptive face bounding box proportions.

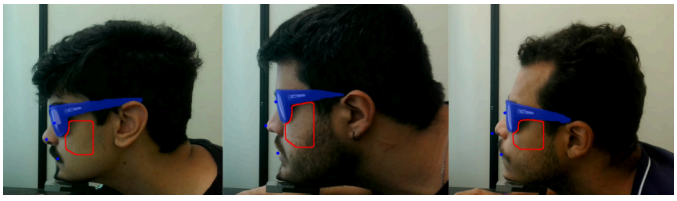


Fig. 7: Profile Facial Segmentation Results Across Multiple Faces

Although this output represents the topological region of exposed skin rather than a direct linear distance measurement, isolating this specific boundary provides the required input for gap quantification. By computing the distance profile across

the top edge of this exposed skin mask and the adjacent frame border, future iterations of the algorithm will extract the physical gap width (in millimeters) along the lateral axis.

IV. Conclusions

This research presented an experimental strategy focused on the geometric analysis of the interface between the user's face and sunglasses. The integration of the SAM3 and YOLO11 models demonstrated promising accuracy in the automated detection and segmentation of frames across distinct poses, supported by IoU metrics that remained consistently above 0.80 even in lateral views. In the frontal view, the methodology enabled the demarcation of critical facial regions, such as the nose and cheeks.

In profile acquisitions, the use of the YuNet detector enabled anchoring and delimitation of the lateral zygomatic region, overcoming the self-occlusion limitations inherent to this viewing angle. The application of a logical bitwise AND operation with the frame mask ensured precise isolation of the exposed skin area without overlap with the accessory, establishing the topological baseline required to quantify the lateral gap width in millimeters in future work.

In continuation of this work, the research progresses toward a methodology designed to spatially map and quantify regions lacking geometric coverage, thereby evaluating the actual blocking capacity of sunglasses. For the experimental evaluation, a single frame model will be analyzed across a sample of approximately five to ten participants exhibiting diverse facial morphological characteristics, alongside a parallel measurement using an ISO 18526-4 standardized headform (size M). This approach will enable the investigation of coverage variability associated with individual anatomical features and establish a direct comparison between standardized normative conditions and actual usage scenarios.

To achieve comprehensive mapping of exposure areas, the experimental apparatus will be expanded with the addition of three cameras positioned to capture the superior and inferior facial regions.

Currently configured as a benchtop platform for metrological research and laboratory validation, the system is projected to evolve into a practical tool suitable for deployment in commercial optical centers, industry trade fairs, public exhibitions, and scientific outreach initiatives as the research matures.

Acknowledgment

The authors gratefully acknowledge the financial support from FAPESP (grant number 2024/22508-0) and the National Council for Scientific and Technological Development (CNPq).

References

[1] D. H. Sliney, "Physical factors in cataractogenesis: ambient ultraviolet radiation and temperature," Invest. Ophthalmol. Vis. Sci., vol. 27, no. 5, pp. 781–790, May 1986.

[2] D. H. Sliney, "UV radiation ocular exposure dosimetry," J. Photochem. Photobiol. B, Biol., vol. 31, no. 1-2, pp. 69–77, Nov. 1995, doi: 10.1016/1011-1344(95)07171-5.

[3] D. G. Pitts, "The human ultraviolet action spectrum," Am. J. Optom. Arch. Am. Acad. Optom., vol. 50, no. 12, pp. 948–960, Dec. 1973, doi: 10.1097/00006324-197312000-00001.

[4] J. P. G. Bergmanson and P. G. Söderberg, "The significance of ultraviolet radiation for eye health," Ophthalmic Physiol. Opt., vol. 15, no. 5, pp. 439–444, Sep. 1995, doi: 10.1016/0275-5408(95)00035-X.

[5] D. Chawda and P. Shinde, "Effects of solar radiation on the eyes," Cureus, vol. 14, no. 10, p. e30861, Oct. 2022, doi: 10.7759/cureus.30861.

[6] S. Mathôt, L. van der Linden, L. Grainger, and J. Theeuwes, "The pupillary light response reveals the focus of covert visual attention," PLoS ONE, vol. 8, no. 10, p. e78168, Oct. 2013, doi: 10.1371/journal.pone.0078168.

[7] C. Backes et al., "Sun exposure to the eyes: predicted UV protection effectiveness of various sunglasses," J. Expo. Sci. Environ. Epidemiol., vol. 29, no. 6, pp. 753–764, Nov. 2019, doi: 10.1038/s41370-018-0087-0.

[8] M. Marro, L. Moccozet, and D. Vernez, "A model of ocular ambient irradiance at any head orientation," *Comput. Biol. Med.*, vol. 179, p. 108903, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108903.

[9] International Commission on Non-Ionizing Radiation Protection (ICNIRP), "Guidelines on limits of exposure to ultraviolet radiation of wavelengths between 180 nm and 400 nm (incoherent optical radiation)," Health Phys., vol. 87, no. 2, pp. 171–186, Aug. 2004, doi: 10.1097/01.HP.0000130453.30811.23.

[10] Y. Deng et al., "Effect of protective measures on eye exposure to solar ultraviolet radiation," Photochem. Photobiol., vol. 97, no. 1, pp. 205–212, Jan. 2021, doi: 10.1111/php.13328.

[11] P. Miguel et al., "Prototype for checking scattered light blocking of sunglasses frame," in Proc. XXIX Brazilian Congr. Biomed. Eng. (CBEB), Ribeirão Preto, Brazil, 2025, pp. 1064–1071, doi: 10.1007/978-3-031-70613-4126.

[12] J. Turner and A. Parisi, "Ultraviolet radiation albedo and reflectance in review: The influence to ultraviolet exposure in occupational settings," Int. J. Environ. Res. Public Health, vol. 15, no. 7, p. 1507, Jul. 2018, doi: 10.3390/ijerph15071507.

[13] M. Masili, F. O. Duarte, and L. Ventura, "Calculation of solar ultraviolet influx in the eye considering the field of view and pupillary dilation due to sunglasses," Sci. Rep., vol. 14, no. 1, p. 6604, Mar. 2024, doi: 10.1038/s41598-024-57121-7.

[14] N. Carion et al., "SAM 3: Segment anything with concepts," unpublished, arXiv:2511.16719, 2026.

[15] G. Liu, K. Wu, W. Lan, and Y. Wu, "YOLO-FDCL: Improved YOLOv8 for driver fatigue detection in complex lighting conditions," *Sensors*, vol. 25, no. 15, Art. no. 4832, Aug. 2025, doi: 10.3390/s25154832.

[16] J. Deng et al., "RetinaFace: Single-shot multi-level face localisation in the wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5203-5212.

Geometry-Guided Deep Learning for Intermediate Projection Synthesis in Digital Breast Tomosynthesis

1st Valentina Perafan Muñoz
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
valper18@usp.br

2nd Jeniffer M. Valencia Ruiz
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
jevalencia84@usp.br

3rd Dyego Alves Pereira
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
dyego.pereira@usp.br

4th Arthur C. Costa
Dept. of Radiology
University of Pennsylvania
Philadelphia, USA
arthur.costa@pennmedicine.upenn.edu

5th Marcelo A. C. Vieira
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
mvieira@sc.usp.br

***Abstract*—Synthesizing non-acquired projections in digital breast tomosynthesis (DBT) could support acquisition protocols with fewer exposures, potentially reducing radiation dose and acquisition time. This work presents a preliminary, projection-domain evaluation of a geometry-guided deep learning architecture for intermediate-projection synthesis. The architecture combines a U-Net that estimates a per-pixel depth map, deterministic differentiable warping based on known acquisition angles and a parallax approximation, and a residual refinement U-Net. Training combines reconstruction, structural similarity, depth regularization, geometric consistency, and gradient-preservation losses. Experiments used 100 OpenVCT breast phantoms, divided into 80 for training, 10 for validation, and 10 for testing. Evaluation included comparison with linear blending and an ablation of the geometric consistency loss at two angular gaps. The original implementation yielded higher mean PSNR and SSIM than linear blending; however, its PSNR was not reproduced with a separately implemented pipeline, preventing a firm conclusion about the consistency of this advantage. In the reported ablation experiments, removing the geometric consistency loss yielded higher mean PSNR and SSIM at both gaps, providing no evidence of a benefit from this loss under the tested conditions. Estimated depth maps lacked physical validation, and mean geometric displacements remained subpixel. These preliminary results motivate further investigation of the architecture as an alternative for DBT projection synthesis, while its specific geometric contribution remains unresolved. Further evaluation should establish reproducibility and assess the impact on tomographic reconstruction before drawing conclusions about potential reductions in radiation dose or acquisition time.**

***Index Terms*—digital breast tomosynthesis, intermediate-projection synthesis, deep learning, acquisition geometry, image interpolation**

This work was supported by the São Paulo Research Foundation (FAPESP; Grants No. 2021/12673-6 and 2026/18409-2), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; Finance Code 001), and the Brazilian National Council for Scientific and Technological Development (CNPq; Grant No. 303251/2025-5).

I. Introduction

Breast cancer remains one of the most frequently diagnosed cancers worldwide, motivating continuous improvements in screening and diagnostic imaging technologies [1]. Digital breast tomosynthesis (DBT) acquires multiple low-dose projections from different angular positions to reconstruct sectional planes of the breast, significantly reducing the anatomical superimposition inherent to conventional two-dimensional mammography [2], [3]. Clinical studies have demonstrated improvements in lesion detection rates and reductions in recall rates compared to standard digital mammography [4].

However, acquiring more projections involves a trade-off among image quality, angular range, and cumulative radiation dose [5]. One strategy to address this consists of acquiring a reduced subset of real projections, computationally synthesizing the intermediate views that were not acquired, and completing the set with these synthetic projections before tomographic reconstruction, thereby preserving angular sampling. Depending on how this protocol is implemented, it could in principle allow a reduction not only in radiation dose but also in acquisition time.

The present work, however, evaluates only the projection synthesis step of this proposal: it does not yet assess the effect of the synthesized projections on tomographic reconstruction, nor does it demonstrate an actual reduction in radiation dose or acquisition time. The application outlined above motivates this line of work, but should be distinguished from what is effectively evaluated in this article, which is restricted to the projection domain; extending the evaluation to reconstruction is left for future work (Section IV-A).

Costa et al. [6] explored this approach by adapting a residual refinement interpolation network (RRIN) [7]—originally de-

veloped for video frame interpolation and trained with a VGG-based perceptual loss—to generate a single intermediate DBT projection from two real neighboring projections. Two usage strategies were investigated: inserting synthetic projections between existing ones to increase angular sampling without additional exposure, and replacing part of the acquired set with synthetic projections to reduce dose and acquisition time—for instance, reconstructing from eight acquired and seven synthesized projections instead of acquiring all 15. Both the synthesized projections and the resulting reconstructions were evaluated on virtual phantoms, a physical anthropomorphic phantom, and clinical cases; the inclusion of synthetic projections improved the contrast-to-noise ratio (CNR) relative to the corresponding reduced sets and helped suppress reconstruction artifacts. However, the synthesized projections also exhibited blurring and a measurable increase in systematic bias relative to the real acquisitions, indicating that the interpolation network did not fully preserve fine structural detail.

The idea of completing a reduced projection set with synthetic projections before reconstruction has therefore already been investigated by Costa et al. [6]. The contribution of the present work lies in studying a different architecture to perform this synthesis step, motivated specifically by the blurring and systematic signal alterations described above.

The RRIN, like other video frame interpolation networks, infers displacement between frames without any explicit notion of acquisition geometry: an optical-flow-based interpolation network estimates the displacement directly from the two input images. In DBT, however, the detector and breast remain stationary while the X-ray source moves along a known trajectory, and internal structures at different depths are displaced by different amounts as the source position changes, following a relationship fully constrained by the known acquisition geometry [9]. This raises the question of whether explicitly incorporating this geometric information can better guide the synthesis process and help preserve such depth-dependent structures, compared to a video-oriented network with no access to it. Rather than conditioning on the angle implicitly, the proposed approach incorporates the acquisition geometry *explicitly* and *differentiably*: a U-Net estimates a per-pixel depth map, the displacements are computed from this depth map and the known acquisition angles via a deterministic geometric deformation, and a second U-Net performs residual refinement, analogous to geometry-integrated novel-view synthesis methods [10]. The formal parallax equation is introduced in Section II.

This work presents a preliminary, projection-domain evaluation of this alternative architecture. We cannot yet claim that it surpasses the RRIN or reduces blurring, as this would require a direct, matched-protocol comparison that was not performed here, since the two studies differ in data, training objective, and evaluation protocol; nor does this work yet assess the resulting effect on tomographic reconstruction, which Costa et al. [6] already investigated for their approach.

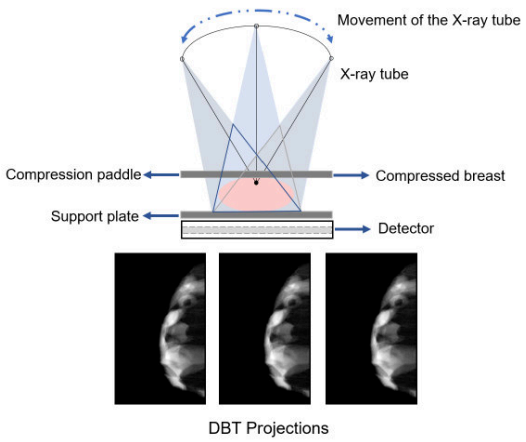


Fig. 1. Acquisition geometry in DBT. The angular movement of the X-ray tube generates projections with depth-dependent parallax offsets along the vertical detector axis. Source: Author's own elaboration.

II. Materials and Methods

A. Dataset and Acquisition Geometry

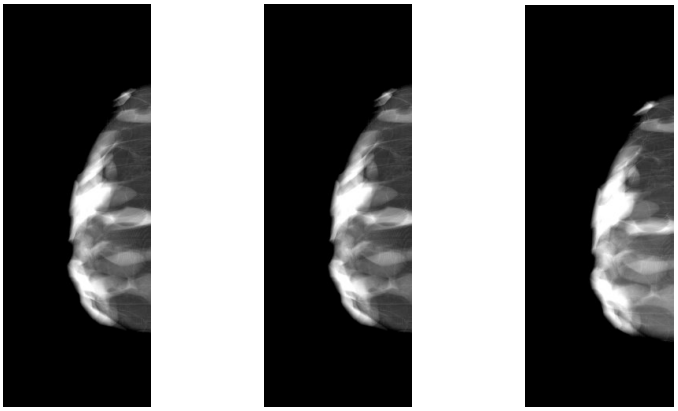
One hundred virtual breast phantoms were generated using OpenVCT [11], simulating the acquisition protocol of the Hologic Selenia Dimensions® system. A virtual clinical trial platform was chosen over a physical DBT system because the acquisition geometry of physical equipment cannot be freely modified, whereas OpenVCT allows generating projections under any specified geometry. This makes it possible to know the exact acquisition angle of every projection and, in turn, to verify directly whether the proposed model is indeed making use of that geometric information, rather than relying on it being approximately recoverable from clinical metadata. Realistic quantum and electronic noise, characteristic of clinical DBT acquisitions, was added to each projection using the noise-insertion model provided by OpenVCT [11]. Each phantom comprises 15 DBT projections in DICOM format with dimensions of 4096×3328 pixels, 14-bit depth, and a pixel spacing of $p = 0.07$ mm (full detector resolution, without binning). Projections are distributed symmetrically over an arc of $\pm 7.49^\circ$ in uniform steps of 1.07°.

The exact acquisition angle θ_i for each projection i was computed from the known X-ray focal spot position (FS_x, FS_z) extracted from simulator metadata as:

$$\theta_i = \operatorname{atan2}(FS_x, FS_z). \tag{1}$$

Angles are reported in degrees for interpretation and converted to radians for trigonometric operations during geometric warping. It was empirically confirmed that parallax displacement occurs predominantly along the vertical axis (rows), consistent with the angular trajectory of the X-ray tube, as illustrated in Fig. 1. The dataset was partitioned into training, validation, and test sets (80/10/10) at the complete phantom level, preventing information leakage.

Fig. 2 shows three DBT projections from the same virtual breast phantom at different angular positions. The visible



(a) Projection 1 (b) Projection 7 (c) Projection 15

Fig. 2. Examples of OpenVCT DBT projections from the same virtual breast phantom at different projection indices. Apparent displacement of anatomical structures illustrates the depth-dependent parallax effect underlying the proposed geometry-guided synthesis model. Source: Author's own elaboration from OpenVCT virtual phantoms [11].

spatial shifts across views are directly related to the depth-dependent parallax expressed in Eq. (2), motivating the explicit use of acquisition geometry during intermediate projection synthesis.

B. Preprocessing, Triplet Formation, and Patch Extraction

Each full-resolution projection was normalized from 16-bit unsigned integer to float32 in $[0, 1]$ using percentile-based intensity scaling (1st–99th percentiles), computed independently from each projection's own pixel values. During training, this includes the target projection P_{target}, which is available in the simulated dataset for supervision. In a real deployment scenario, however, P_{target} is by definition not acquired, so its own percentiles would not be available; in that case, the normalization scale would instead need to be derived from the two acquired neighboring projections (for instance, their average or interpolated percentiles). This practical aspect of the pipeline was not evaluated in the present work. Breast tissue was segmented from the background using Otsu thresholding.

Training samples were organized as ordered triplets $(P_{\text{before}}, P_{\text{target}}, P_{\text{after}})$, where P_{target} is the projection to be synthesized and P_{before}, P_{after} are its two neighboring projections at angular distance gap $\times\, 1.07^\circ$ from the target. For gap = 2, each neighboring projection is located 2.14° from the target,yielding an angular separation of 4.28° between the two input projections. For each phantom with 15 projections and gap = 2, valid target indices range from 2 to 12, resulting in 11 possible triplets per phantom.

Patches of 256×256 pixels were randomly sampled within the segmented breast region, requiring at least 90% tissue coverage. Three valid patches were extracted per triplet, yielding 2,640 training samples, 330 validation samples, and 330 test samples. Quantitative evaluation was performed on 328 test samples because the batch-wise evaluation procedure used batches of eight and discarded the final incomplete batch. As data augmentation, a random vertical flip was applied with probability 0.5, with simultaneous sign inversion of the associated angles—geometrically valid given the vertical direction of parallax.

Fig. 3 illustrates one training triplet with its associated acquisition angles.

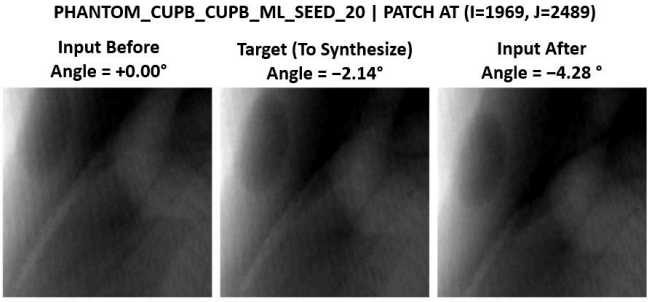


Fig. 3. Training triplet: neighboring input projections ("Input Before" and "Input After") and intermediate target projection ("Target"), with exact acquisition angles incorporated as geometric information. Source: Author's own elaboration from virtual phantoms simulated in OpenVCT.

C. Geometric Formulation

In DBT, the parallax displacement of a structure at depth z above the detector, when transitioning from angle θ_a to angle θ_b, is approximated by [9]:

$$\Delta_{\text{row}} \approx \frac{z}{p} \left(\tan \theta_b - \tan \theta_a \right), \tag{2}$$

where $p = 0.07\,\text{mm}$ is the pixel spacing. This relationship is analogous to the classical shift-and-add DBT reconstruction principle, repurposed here for projection synthesis. The displacement is purely vertical, consistent with the empirically confirmed direction of parallax.

D. Proposed Architecture

The model receives the two neighboring projections and the three acquisition angles associated with the before, after, and target views (θ_{before}, θ_{after}, θ_{target}), with spatial patch dimensions of 256×256 pixels. Stage 1 receives the two neighboring projection images, whereas Stage 3 receives the two warped views and their blend. The acquisition angles are not supplied as additional image channels to either U-Net; they are used in Stage 2 to compute the geometric shift via Eq. (2). The architecture operates in three sequential stages, illustrated in Fig. 4.

Stage 1 — Depth Network (learned): A U-Net [12] with a ResNet-18 encoder [13], randomly initialized (X-ray projections differ substantially from natural ImageNet images), estimates a per-pixel depth map bounded to $[0, 60]$ mm via scaled sigmoid activation, compatible with the simulated compressed breast thickness.

Stage 2 — Geometric Deformation (not learned): The estimated depth map and the exact acquisition angles are combined in (2) to compute a per-pixel vertical shift for each neighboring projection toward the target view. The deformation is implemented as differentiable bilinear sampling,

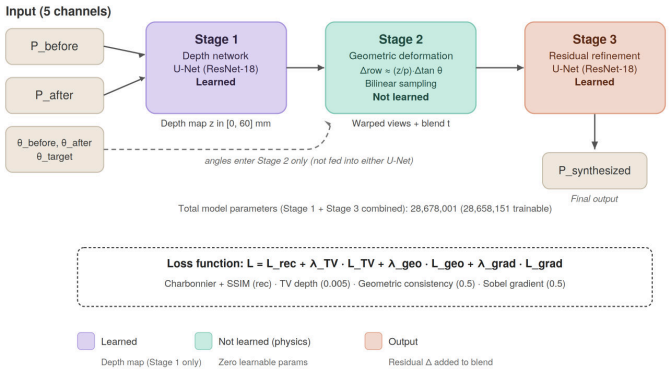


Fig. 4. Proposed three-stage architecture for DBT projection synthesis. Stage 1 (learned): U-Net depth estimator producing a per-pixel depth map. Stage 2 (not learned): deterministic geometric deformation using the DBT parallax equation and differentiable bilinear sampling. Stage 3 (learned): residual refinement U-Net producing the final synthesized projection. Source: Author's own elaboration.

allowing gradients to flow back to the depth network during training. The two warped views are blended by the relative angular position:

$$t = \frac{\theta_{\text{target}} - \theta_{\text{before}}}{\theta_{\text{after}} - \theta_{\text{before}}}. \tag{3}$$

This stage contains *no learnable parameters*.

Stage 3 — Residual Refinement (learned): A second U-Net with a ResNet-18 encoder receives the blended result and both warped views, predicting a residual correction added to the blend to produce the final synthesized projection. The model output includes the synthesized projection, the estimated depth map, and an auxiliary geometric consistency view used only during training. The complete model contains 28,678,001 parameters (28,658,151 trainable).

E. Loss Function

The total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{TV}}\,\mathcal{L}_{\text{TV}} + \lambda_{\text{geo}}\,\mathcal{L}_{\text{geo}} + \lambda_{\text{grad}}\,\mathcal{L}_{\text{grad}}, \tag{4}$$

where weights were determined empirically through preliminary experiments on the validation set, evaluating PSNR over configurations in $[0.001,\ 0.1]$ for λ_{TV} and $[0.1,\ 1.0]$ for λ_{geo} and λ_{grad}.

$\mathcal{L}_{\text{rec}}$ combines Charbonnier loss—a smooth, pixel-wise L1 approximation robust to outliers, widely adopted in video frame interpolation networks [7], here averaged over all pixels—with structural similarity (SSIM) [8]:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{y}_i - y_i)^2 + \varepsilon^2} + 0.2\,(1 - \text{SSIM}(\hat{y},\, y)), \tag{5}$$

where i indexes the N pixels of the patch and $\varepsilon = 10^{-3}$.

$\mathcal{L}_{\text{TV}}$ penalizes high-frequency spatial noise in the depth map, computed as the mean of the absolute horizontal and vertical first-order differences of the estimated depth map, preventing warping artifacts ($\lambda_{\text{TV}} = 0.005$, chosen to regularize without suppressing spatial variation).

$\mathcal{L}_{\text{geo}}$ is the geometric consistency loss ($\lambda_{\text{geo}} = 0.5$): the "before" projection warped toward the "after" view is compared against the real "after" projection using the mean absolute error (L1), providing a spatially meaningful training signal intended to anchor the depth network's output to the acquisition geometry. Since the vertical warp (Section II-D, Stage 2) can require sampling positions outside the 256×256 patch near its top and bottom borders, the sampling row index is clamped to the valid range before bilinear interpolation, which effectively extends the nearest valid row rather than introducing padding or discarding those regions. The individual contribution of the $\mathcal{L}_{\text{geo}}$ term is quantified through an ablation study in Section III-A.

$\mathcal{L}_{\text{grad}}$ is a Sobel-based gradient loss ($\lambda_{\text{grad}} = 0.5$), computed as the mean absolute difference between the Sobel horizontal and vertical edge-magnitude maps of the real and synthesized projections, penalizing blurriness without requiring an external feature extractor network.

No additional normalization (e.g., rescaling each term to a common numerical range) is applied to the four terms in Eq. (4) beyond computing each of them as a mean over its respective per-pixel quantity, as detailed above; the λ weights were tuned empirically at this scale (Section II-E) to balance their relative contributions during training.

F. Experimental Design

The primary experiments were conducted with gap = 2, as defined in Section II-B; the study was later extended to gap = 3 and an additional quality metric (QILV) as part of the ablation study (Section III-A). A separate model was trained independently for each configuration analyzed in this work—one training run per combination of gap (2 or 3) and loss configuration (full loss, or with $\lambda_{\text{geo}} = 0$)—rather than a single shared network applied across gaps. The proposed model (gap = 2, full loss) was compared against a linear blending baseline—a weighted average of the two neighboring projections according to t in (3)—containing no learned components. Evaluation metrics were PSNR and SSIM [8], reported as mean ± standard deviation over 328 test samples, computed from 328 of the 330 available test patches; the batch-wise evaluation procedure discarded the final, incomplete batch of two patches, which was not necessary given that the metrics could equally be computed one patch at a time. Every training run used the Adam optimizer with an initial learning rate of 10^{-4}, batch size of 4, and a maximum of 60 epochs, with ReduceLROnPlateau (patience = 2, factor = 0.5) and EarlyStopping (patience = 5, best weights restored). The model reported in Table I converged at epoch 28; the independently trained instances used in the ablation study (Section III-A) were each trained under the same protocol but converged at their own, generally different, epochs, consistent with each being a separate training run rather than a continuation or reuse of the same weights.

III. Results

Table I presents the quantitative results for gap = 2 over 328 evaluated test samples. The proposed model achieved PSNR = 55.18 ± 7.01 dB and SSIM = 0.9939 ± 0.0052, surpassing the linear blending baseline by $+3.80$ dB in PSNR and $+0.0094$ in SSIM. Only methods evaluated under the same experimental protocol are reported.

TABLE I
IMAGE QUALITY METRICS ON THE TEST SET FOR GAP = 2 (2.14° PER NEIGHBORING PROJECTION RELATIVE TO THE TARGET AND A TOTAL ANGULAR SPAN OF 4.28°). RESULTS ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION OVER 328 EVALUATED SAMPLES.

Method	PSNR (dB)	SSIM
Linear blending baseline	51.38 ± 9.01	0.9845 ± 0.0135
Proposed model	$\mathbf{55.18 \pm 7.01}$	$\mathbf{0.9939 \pm 0.0052}$
Δ: $+3.80$ dB PSNR and $+0.0094$ SSIM.		

In addition to improving the mean PSNR and SSIM, the proposed model reduced the dispersion of the results. The PSNR standard deviation decreased from ± 9.01 dB for the linear blending baseline to ± 7.01 dB for the proposed model. This reduction indicates more consistent performance across the evaluated test samples, particularly in cases for which direct linear interpolation was less effective.

Fig. 5 presents two representative qualitative examples. The columns show the preceding neighboring projection, the corresponding real target, and the synthesized projection. The synthesized images reproduce the main global intensity distributions and tissue patterns observed in their corresponding target projections. Nevertheless, local differences, residual smoothing, and small spatial deviations remain visible.

For the model reported in Table I, the estimated depth values are concentrated within a narrow range of approximately 0–0.1 mm, several orders of magnitude below the physically expected 0–60 mm corresponding to the compressed breast thickness. Visual inspection of these maps, including contrast-enhanced renderings, did not reveal an interpretable anatomical or structural pattern; for this reason, depth map visualizations are not included in this work. The predicted values are therefore considered an internal, non-calibrated geometric representation generated by the depth branch to guide the deformation stage, rather than a physically interpretable depth estimate. This 0–0.1 mm range applies specifically to the model reported in Table I (gap = 2, full loss). The depth range was not consistent across the other training instances analyzed in the ablation study below (Section III-A): for gap = 2, the independently retrained full model exhibited a wider range of approximately 0–0.2 mm, while the ablated (no-$\mathcal{L}_{\text{geo}}$) model exhibited a substantially wider and more variable range, with per-batch minima and maxima observed between approximately 0 and 1.0 mm. The corresponding depth ranges for the two gap = 3 configurations were not systematically recorded. Given this variability, the resulting per-pixel vertical shift (Section III-A), rather than the raw depth magnitude, is used there as the more direct and comparable quantitative characterization of the geometric deformation actually applied in each case.

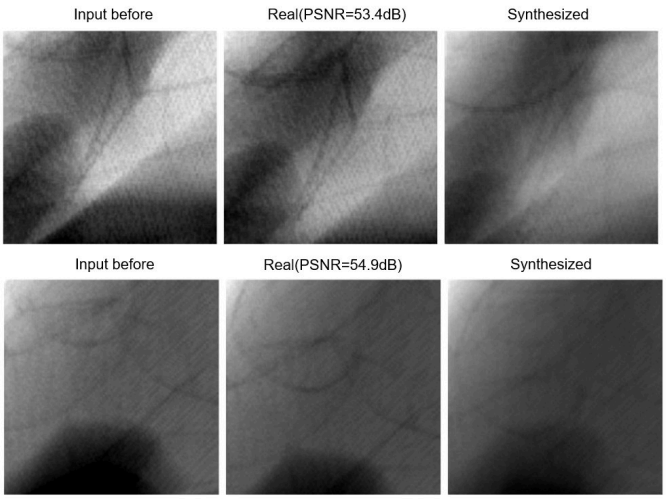


Fig. 5. Qualitative evaluation of intermediate DBT projection synthesis for two representative test cases. From left to right: preceding neighboring projection (Input before), corresponding real target, with the PSNR between the synthesized projection and the real target indicated in the panel header, and the synthesized projection. The synthesized images reproduce the main global intensity distributions and tissue patterns observed in their corresponding target projections; local differences and residual smoothing remain visible. Source: Author's own elaboration.

A. Ablation Study: Contribution of the Geometric Consistency Loss

To directly quantify the contribution of $\mathcal{L}_{\text{geo}}$, a second model was trained under identical conditions but with $\lambda_{\text{geo}} = 0$. This experiment was repeated for gap = 2 and gap = 3 (angular separation of 3.21° per neighboring projection, 6.42° total) to assess whether the contribution of this term scales with the expected parallax magnitude. Results are summarized in Table II. The full-model results for $\text{gap} = 2$ in Tables I and II were obtained using different code implementations. The original PSNR of 55.18 dB was not reproduced in the reconstructed pipeline, which yielded 49.63 dB for the full model reported in Table II. The cause of this discrepancy was not established. Therefore, the two tables should not be interpreted as directly comparable evaluations, and the advantage over linear blending reported in Table I remains specific to the original experiment. The ablation results are reported descriptively, without establishing robustness across implementations or training runs.

Contrary to the initial hypothesis, removing $\mathcal{L}_{\text{geo}}$ did not degrade PSNR or SSIM at either angular separation; the ablated model showed higher mean PSNR and SSIM than the full model at both angular separations tested (gap = 2: -0.74 dB in mean PSNR; gap = 3: -1.56 dB in mean PSNR; see Table II for the corresponding SSIM means). As the 328/264 evaluated patches are drawn from only 10 test phantoms and are therefore not fully independent observations (Section IV-A), these mean differences are reported descrip-

TABLE II

ABLATION OF THE GEOMETRIC CONSISTENCY LOSS ($\mathcal{L}_{\text{GEO}}$) ACROSS TWO ANGULAR SEPARATIONS. "SHIFT" DENOTES THE MEAN ABSOLUTE PER-PIXEL VERTICAL DISPLACEMENT APPLIED BY STAGE 2 (SECTION II-D) WHEN WARPING THE "BEFORE" NEIGHBORING PROJECTION TOWARD THE TARGET VIEW; THE DISPLACEMENT APPLIED TO THE "AFTER" NEIGHBOR IS OF COMPARABLE MAGNITUDE, BY SYMMETRY OF THE GAP-BASED SAMPLING, AND IS NOT REPORTED SEPARATELY.

Gap	Method	PSNR (dB)	SSIM	Shift (px)
2 (n=328)	Full ($\mathcal{L}_{\text{geo}}$)	49.63 ± 4.41	0.9921 ± 0.0039	0.06
2 (n=328)	No $\mathcal{L}_{\text{geo}}$	50.37 ± 2.48	0.9924 ± 0.0036	0.41
3 (n=264)	Full ($\mathcal{L}_{\text{geo}}$)	47.19 ± 5.18	0.9885 ± 0.0070	0.31
3 (n=264)	No $\mathcal{L}_{\text{geo}}$	48.75 ± 2.37	0.9899 ± 0.0048	0.37

tively, without a formal significance test. An analysis of the per-pixel vertical displacement actually applied by Stage 2 (Section II-D) showed that, although this displacement grew with gap as expected from Eq. (2) (mean shift increasing from 0.06 to 0.31 px for the full model), its mean remained sub-pixel in every configuration tested. This mean does not imply that every individual displacement was sub-pixel: for gap = 2, the 95th-percentile and maximum displacement were 0.08 px and 0.09 px for the full model, and 0.43 px and 0.53 px for the ablated model (the corresponding percentiles were not computed for gap = 3). Under such small typical displacements, the warped views fed into Stage 3 are nearly identical to the unwarped inputs (mean absolute difference < 0.0004 on the $[0, 1]$ intensity scale), which is consistent with, but does not by itself establish, the hypothesis that the residual refinement stage accounts for most of the reconstruction quality observed independently of $\mathcal{L}_{\text{geo}}$.

As a consistency check on the units and degree-to-radian and depth-to-pixel conversions used in Eq. (2), a depth of $z = 0.1$ mm at gap = 2 ($\Delta\theta = 2.14°$ between a neighboring projection and the target) yields a predicted shift of $(z/p)\tan(\Delta\theta) \approx (0.1/0.07) \times \tan(2.14°) \approx 0.053$ px. This is consistent, to within the same order of magnitude, with the 0.06 px mean shift observed for the full model at gap = 2 given its ∼0–0.2 mm depth range noted above. Applying the same relation in reverse, the ablated model's larger 0.41 px mean shift at gap = 2 implies a mean depth on the order of 0.4–0.8 mm, consistent with the wider, higher depth range (∼0–1.0 mm) separately observed for that model. This agreement supports that the degree-to-radian conversion and the depth-to-pixel scaling by $1/p$ in the sampling coordinates are implemented correctly. This ablation isolates only the contribution of the $\mathcal{L}_{\text{geo}}$ loss term to training; the geometric deformation module (Stage 2) remains active with or without this loss, since the depth network continues to estimate depth and drive the warp in both configurations. Whether removing the deformation module itself—rather than only the loss term that partly shapes it—would change the results, and whether Stage 3 alone would perform comparably given only the unwarped neighboring projections, were not tested here and are identified as open questions for future work (Section IV-A). As an additional, loss-independent quality metric, the Quality Index based on Local Variance (QILV) [14] was computed for gap = 2, yielding 0.9994 ± 0.0024 for the full model and 0.9997 ± 0.0012 for the ablated model. Both values are close to the maximum of 1, and the small difference between them is in the same direction as the PSNR/SSIM trend above; however, given how close both values already are to 1, this difference should not be interpreted as evidence that either configuration better preserves fine structures, only as being broadly consistent with the direction of the PSNR/SSIM comparison.

IV. DISCUSSION

The proposed model surpassed the linear-blending baseline for gap = 2 (Table I), consistent with the motivating hypothesis that incorporating the acquisition geometry could better guide the synthesis process (Section I) — through a learned depth map, a deterministic parallax-based deformation, and a residual refinement stage. This contrasts with the limitations reported by Costa et al. [6]: their RRIN-based approach, trained without explicit access to the acquisition geometry, exhibited blurring and a measurable increase in systematic bias relative to the real acquisitions (Section I). However, this comparison to the baseline does not by itself establish that the geometric mechanism specifically, as opposed to the additional network capacity introduced by the depth and refinement U-Nets, is responsible for the observed gain; the ablation study in Section III-A showed no measurable benefit from the geometric consistency loss $\mathcal{L}_{\text{geo}}$ specifically, for either gap = 2 or gap = 3. This indicates that this specific loss term does not account for the observed advantage over linear blending; whether that advantage instead reflects the architecture's overall capacity (the depth-driven warping and residual refinement stages acting together), the residual refinement stage alone irrespective of the geometric warp, or some combination of these, was not isolated in this work and remains an open question (Section IV-A).

For reference, Costa et al. [6] reported PSNR = 57.28 dB and SSIM = 0.9981 for RRIN-interpolated projections in their virtual-phantom evaluation. These values are not directly comparable to those in Table I: beyond the noise-free synthetic reference used in that evaluation—versus the simulated projections with realistic quantum and electronic noise used as reference here, generated by OpenVCT and not acquired from patients—the two studies also differ in network architecture, training objective (a VGG-based perceptual loss versus the geometry-aware loss in Eq. (4)), dataset, and patch/evaluation protocol. Consequently, no single factor, including the noise characteristics of the reference, should be assumed to fully account for the numerical gap between the two results.

Regarding the estimated depth values, the raw predicted magnitudes remain far below the physically expected 0–60 mm range corresponding to the compressed breast thickness. Since the estimated depth enters Eq. (2) multiplicatively, a predicted depth close to zero necessarily yields a correspondingly small geometric deformation. Describing these values as an internal,

non-calibrated representation does not fully resolve this discrepancy, since they are not merely an abstract intermediate feature: they are substituted directly into Eq. (2), which interprets them as a physical depth in millimeters to compute the warp. Part of the difficulty may also lie in the depth representation itself, rather than only in training: each pixel of an X-ray projection integrates contributions from structures located at multiple depths along the beam path, so representing this with a single scalar depth per pixel is an approximation that may not admit an unambiguous physical value even for a well-trained network. Visual inspection, including contrast-enhanced renderings, did not reveal a physically interpretable or visually coherent pattern, indicating that the depth branch does not yet learn a calibrated anatomical representation, whether due to insufficient training signal, this representational limitation, or both. The ablation study (Section III-A) further showed that, at these magnitudes, the resulting geometric warp is sub-pixel on average regardless of whether $\mathcal{L}_{\text{geo}}$ is used, consistent with, though not sufficient on its own to establish, the interpretation that Stage 2 contributes only a minor perturbation relative to the original inputs and most of the reconstruction quality is attributable to the residual refinement stage (Stage 3). Stronger depth supervision, for instance using the inter-projection parallax displacement as pseudo-ground-truth, remains a direction worth exploring to anchor the depth branch's output closer to physically meaningful magnitudes; however, given the multi-depth integration described above, it is not guaranteed that such supervision alone would yield anatomically accurate, per-pixel depth estimates, and this should be verified rather than assumed.

The two qualitative examples in Fig. 5 (PSNR of 53.4 dB and 54.9 dB) fall close to the overall test-set mean (55.18 ± 7.01 dB), suggesting these particular cases are not unusually favorable; however, two examples close to the mean do not by themselves characterize the full range of behaviors observed across the test set—in particular, they do not illustrate the higher-dispersion tail reflected in the ± 7.01 dB standard deviation, nor a direct qualitative comparison against the linear-blending baseline at the same locations. In both examples shown, the synthesized projection reproduces the coarse tissue structure and global contrast of the real target; the visible residual smoothing is concentrated around fine, high-frequency structures (thin fibroglandular strands and tissue boundaries), consistent with, though not conclusively demonstrating, the sub-pixel geometric displacement discussed above and the reconstruction burden falling mainly on the residual refinement stage.

A. Limitations

The following limitations should be kept in mind when interpreting the results above:

- **Simulated data only.** All projections come from virtual breast phantoms generated with OpenVCT [11]; no clinical or physical phantom images were used.
- **Evaluation restricted to the 2D projection domain.** The clinically most relevant validation requires assessing the impact on 3D reconstruction quality (e.g., CNR on reconstructed slices), following the methodology of Costa et al. [6]; this was not performed here.
- **Evaluation performed on patches, not full projections.** Quantitative evaluation was carried out on randomly sampled 256×256 patches (Section II-B) rather than on full-resolution (4096×3328) projections; whether the model performs consistently across an entire projection, including larger-scale artifacts or inconsistencies at patch boundaries when patches are reassembled, was not evaluated.
- **Test set limited to 10 phantoms.** The 328 (gap = 2) and 264 (gap = 3) evaluated patches are drawn from only 10 held-out phantoms, so patches sharing a phantom are not statistically independent observations. For this reason, the ablation comparison (Section III-A) is reported descriptively, as mean and standard deviation over patches, without a formal significance test; a phantom-level analysis—computing the mean difference between methods for each of the 10 phantoms individually, then testing at that level—should be performed before drawing stronger conclusions from these averages. Furthermore, comparing many patches produced by two trained networks characterizes the difference between those two specific training runs, but does not by itself substitute for assessing run-to-run variability (e.g., across repeated trainings with different random initializations), which was not evaluated here.
- **Depth map without physical validation.** The estimated depth values, while bounded to the $[0, 60]$ mm range by construction (Section II-D), are concentrated implausibly close to zero rather than spanning the physically expected range for the compressed breast thickness (Section III), and have not been validated against any physical ground truth. It remains unclear whether this stems from insufficient training signal, from representing a multi-depth beam path with a single scalar per pixel, or both.
- **Contribution of the deformation module not yet isolated.** The ablation study (Section III-A) removed only the geometric consistency loss $\mathcal{L}_{\text{geo}}$; the deformation module (Stage 2) remained active in both configurations. Whether removing Stage 2 entirely would change the results was not tested.

V. Conclusions

This work presented a preliminary, projection-domain evaluation of a geometry-guided architecture for DBT intermediate-projection synthesis. As an alternative to the RRIN-based approach investigated by Costa et al. [6], the proposed method combines a learned depth map with known acquisition angles to perform deterministic, differentiable warping, followed by residual refinement. No direct comparison with RRIN was performed.

The original implementation yielded higher mean PSNR and SSIM than linear blending (Table I). However, the full-model results in that comparison and in the ablation study were

obtained using different code implementations, and the original PSNR was not reproduced in the reconstructed pipeline. Consequently, the reported advantage over linear blending should be interpreted as an observation from the original experiment, rather than an established, reproducible gain. The discrepancy cannot be attributed to training variability alone on the available evidence.

In the reported ablation experiments, removing the geometric consistency loss ($\mathcal{L}_{\text{geo}}$) yielded higher mean PSNR and SSIM at both tested angular gaps, and higher mean QILV at $\text{gap} = 2$ (Section III-A). These descriptive results did not establish a benefit from this loss term under the tested conditions. Because the geometric warping module remained active in both configurations, its contribution was not isolated. The lack of physically validated depth estimates and the small mean geometric displacements further limit conclusions about the role of geometric guidance.

Future work should first establish a unified, reproducible implementation and evaluation protocol. Subsequent experiments will assess the contribution of the warping module, investigate depth supervision and calibration, and compare the architecture with RRIN under matched conditions. Evaluation will also extend to clinical images and tomographic reconstructions, comparing reduced acquisition sets, reduced sets completed with synthetic projections, and full reference sets.

These preliminary, projection-domain findings motivate further investigation of the proposed architecture as an alternative for DBT projection synthesis. Its impact on tomographic reconstruction and its potential to support reductions in radiation dose and acquisition time remain to be established.

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] A. Chong *et al.*, "Digital breast tomosynthesis: Concepts and clinical practice," *Radiology*, vol. 292, no. 1, pp. 1–14, 2019.

[3] S. Vedantham *et al.*, "Digital breast tomosynthesis: State of the art," *Radiology*, vol. 277, no. 3, pp. 663–684, 2015, doi: 10.1148/radiol.2015141303.

[4] S. M. Friedewald *et al.*, "Breast cancer screening using tomosynthesis in combination with digital mammography," *JAMA*, vol. 311, no. 24, pp. 2499–2507, 2014, doi: 10.1001/jama.2014.6095.

[5] I. Sechopoulos and C. Ghetti, "Optimization of the acquisition geometry in digital tomosynthesis of the breast," *Medical Physics*, vol. 36, no. 4, pp. 1199–1207, 2009, doi: 10.1118/1.3090889.

[6] A. C. Costa, R. de B. Vimieiro, L. R. Borges, B. Barufaldi, A. D. A. Maidment, and M. A. da C. Vieira, "Assessment of video frame interpolation network to generate digital breast tomosynthesis projections," in *Proc. SPIE*, vol. 12286, p. 122861D, 2022, doi: 10.1117/12.2625748.

[7] H. Li, Y. Yuan, and Q. Wang, "Video frame interpolation via residue refinement," in *Proc. IEEE ICASSP*, pp. 2613–2617, 2020.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[9] R. J. Acciavatti and A. D. A. Maidment, "Proposing an acquisition geometry that optimizes super-resolution in digital breast tomosynthesis," in *Lecture Notes in Computer Science*, vol. 7361, pp. 386–393, 2012, doi: 10.1007/978-3-642-31271-7_50.

[10] D. Liu, F. Fan, and A. Maier, "Novel-view X-ray projection synthesis through geometry-integrated deep learning," *arXiv preprint arXiv:2504.11953*, 2025, doi: 10.48550/arXiv.2504.11953.

[11] B. Barufaldi, D. Higginbotham, P. R. Bakic, and A. D. A. Maidment, "OpenVCT: A GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis," in *Proc. SPIE*, vol. 10573, p. 1057358, 2018, doi: 10.1117/12.2294935.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, pp. 234–241, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, pp. 770–778, 2016.

[14] S. Aja-Fernández, R. San José-Estépar, C. Alberola-López, and C.-F. Westin, "Image quality assessment based on local variance," in *Proc. IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4815–4818, 2006, doi: 10.1109/IEMBS.2006.260053.

Habitats in medical images: the influence of parameter choice on the habitat clustering process

1st Sarah Fernandes de Souza
Federal University of ABC
Santo André, SP, Brazil
ORCID: 0009-0005-4524-7327

2nd Rafael Vaz Pandini
University of São Paulo (FMUSP)
São Paulo, SP, Brazil
ORCID: 0000-0002-8979-9378

3rd Valesca Bizinoto
University of São Paulo (ICESP)
São Paulo, SP, Brazil
ORCID: 0009-0000-9838-5516

4th João Miranda
Mayo Clinic Rochester
Rochester, MN, USA
ORCID: 0000-0003-1273-3108

5th Antonildes N. Assunção Jr.
University of São Paulo (FMUSP)
São Paulo, SP, Brazil
antonildes.junior@hc.fm.usp.br

6th Carla Franco Grego Da Silva
Centro Universitário Das Américas - FAM
São Paulo, SP, Brazil
ORCID: 0009-0005-2923-4059

7th Caio Sergio Rizkallah Nahas
Department of Surgery
University of São Paulo, SP, Brazil
São Paulo, SP, Brazil
ORCID: 0000-0001-8036-512X

8th Natally Horvat
Department of Radiology
Mayo Clinic Rochester, USA
ORCID: 0000-0001-8849-5789

9th John Andrew Sims
Federal University of ABC
Santo André, SP, Brazil
ORCID: 0000-0002-9553-7278

10th Erick Dario León Bueno de Camargo
Federal University of ABC
Santo André, SP, Brazil
ORCID: 0000-0001-6191-0620

***Abstract*—Radiomics extracts quantitative data from medical images and can be used to identify habitats inside tumours, but the lack of reproducibility and of standardisation is still a frequent problem in these studies, which often lack the support needed to reproduce their results. This work implements a habitat clustering algorithm and documents how the sigma of the Gaussian filter, the window size, the radius of the closing structuring element and the step of the sliding window affect the results. The *pipeline* resamples the images to isotropic voxels, selects blocks of voxels through a regular *grid*, extracts 17 first-order and 11 texture (GLCM) features from each block and groups the blocks by K-means into three *clusters*. Thirteen configurations, varying one parameter at a time, were applied to four rectal cancer cases (T2-weighted MRI) and compared with the reference configuration through the adjusted Rand index, the habitat proportions, the visual appearance and the execution time. Larger sigma values produced more homogeneous regions with less detail; larger windows reduced detail and coverage, and one of the cases had no valid supervoxels with the largest window; larger radii favoured the most abundant habitats; and a step larger than the window produced gaps in coverage, although varying the step was what most reduced the execution time. The parameter values are not generalisable, but the qualitative behaviour tends to hold, as it is a geometric effect of the method. The code is publicly available, and the stability and robustness of the features still need to be verified.**

***Index Terms*—habitat imaging, habitat analysis, tumour subregions, radiomics, intratumoral heterogeneity**

Ethics statement: the images used were previously acquired and the study was approved by the Research Ethics Committee of ICESP, under opinion CEP No. 1736/20, approved in June 2020.

I. Introduction

Rectal cancer is a malignant tumour that develops in the rectum, the final portion of the gastrointestinal tract responsible for storing faeces and for evacuation. After acquisition and segmentation, the images must be processed so that the data they carry are turned into information.

Segmentation separates distinct regions of the image, delimiting regions or volumes of interest (ROI and VOI) that help define therapeutic strategies [1], [4], [5]. It can be manual, semi-automatic or automatic, and the latter two reduce the intra- and inter-observer biases inherent to the manual method [1]–[3]. In modalities that produce volumetric data, such as CT, MRI and PET-CT, segmentation can be three-dimensional. This article focuses on magnetic resonance images of rectal cancer, with manual segmentation of the tumour.

After segmentation, filtering prepares the image for the application of other algorithms and is widely used in computer vision to reduce noise [6]. Magnetic resonance images contain Rician noise [7], which tends towards a Gaussian distribution as the signal-to-noise ratio increases; for this reason, Gaussian filters are widely used in this modality, including optimised versions such as the bilateral filter [8]. Filtering also favours the extraction of radiomic features [9], [10]. Another preprocessing stage is the application of structuring elements, for example, to masks, reducing the gaps or holes that hinder the extraction of radiomic features; to this end, the closing

operation is applied, which combines a dilation followed by an erosion [12].

Tumours are heterogeneous and present sub-regions, called habitats, which differ from one another in ways that are not visible to the naked eye [1], [16] and whose identification relies on radiomics [9], [13], [14]. Radiomics is a set of computational methods for extracting quantitative data from medical images, and habitat radiomics identifies these sub-regions of lesions by means of radiomic and clustering methods [15], revealing heterogeneous regions within the segmented area. One of the most widely used algorithms for forming the *clusters* is K-means, an unsupervised clustering method.

However, reproducibility, standardisation and validation are still reported as problems in radiomics studies [9], [14], [17], which also face difficulties in being applied in daily clinical practice. Moreover, the habitat results depend on the parameters and on the pre-processing operations applied to the images.

II. Methodology

A. Materials and Tools

The code was run locally in Python, in VS Code, on a Samsung Galaxy Book 4 notebook. The images were provided by ICESP (Instituto do Câncer de São Paulo). Gemini and Claude assisted in the organisation, the writing and the coding, in a strictly supportive role, with the critical reading of the articles and the final writing remaining the responsibility of the authors. Table I details the tools, the execution environment and the versions of the libraries.

TABLE I
TOOLS AND *software* ENVIRONMENT USED FOR THE DEVELOPMENT OF THE WORK.

Tool / Material	**Description**
Samsung Galaxy Book 4	Intel Core i7-1355U (13$^{\text{th}}$ generation). 512 GB SSD. 16 GB of RAM. Intel Iris Xe graphics.
Operating system	Windows 11, *build* 10.0.26200 (SP0)
VS Code	Version 1.117.0
ITK-SNAP	Version 4.0.2
Python	3.14.3 (MSC v.1944 64 bit / AMD64). Libraries: NumPy 2.4.6; SciPy 1.17.1; scikit-learn 1.9.0; matplotlib 3.10.9; pandas 3.0.3; SimpleITK 2.5.5.

B. Objective

The objective is to evaluate how changes in the window size, in the step, in the size of the closing sphere and in the sigma of the Gaussian filter affect the habitat formation algorithm, making explicit how much the results depend on the pre-processing stages and providing implementation details that frequently do not accompany radiomics studies.

C. Procedures

1) Image acquisition: After approval by the ethics committee, four sets of magnetic resonance images of rectal cancer were selected, all previously segmented and obtained with the same equipment model, the Signa HDx 1.5T (GE Healthcare, Chicago, IL, USA), and the same acquisition protocols. The sequences were T2-weighted (T2WI), with 3 mm thick slices. Further details are given in Table II.

2) Method for the Identification of Tumour Habitats: The method combines the extraction of radiomic features and K-means clustering, which uses these quantitative values to group the voxels. The number of clusters was fixed at $k = 3$ after clustering tests with the `calinski_harabasz` and `davies_bouldin` metrics [18]; all the calculated values for the metrics are given in Table III. This value was also adopted by another study to evaluate tumour sub-regions [1].

TABLE III
CLUSTER EVALUATION METRICS FOR THE DEFINITION OF THE NUMBER OF CLUSTERS. VALUES IN BOLD INDICATE THE BEST RESULT OF EACH METRIC. FOR `DAVIES_BOULDIN`, THE LOWER THE VALUE, THE BETTER THE SEPARATION OF THE GROUPS; FOR `CALINSKI_HARABASZ`, THE HIGHER THE VALUE, THE BETTER THE SEPARATION.

Metric	$k = 2$	$k = 3$	$k = 4$	$k = 5$
`davies_bouldin`	1.84	**1.45**	1.49	1.62
`calinski_harabasz`	131,937	**154,654**	125,050	115,253

The algorithm was organised into a *pipeline* (Fig. 1) with the stages of pre-processing, supervoxel selection, feature extraction, clustering and visualisation.

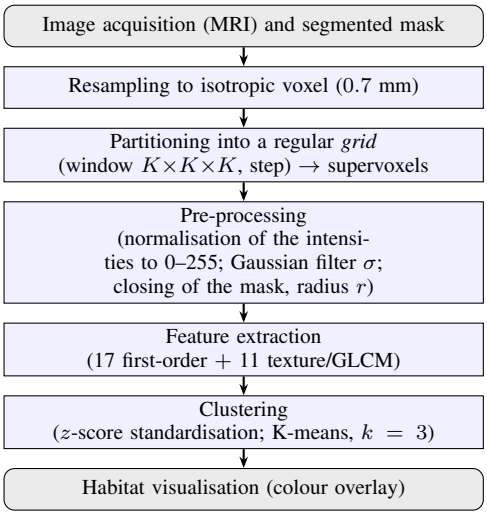


Fig. 1. Stages of the implemented regular-*grid* habitat formation *pipeline*.

In the pre-processing, the image is smoothed by a Gaussian filter (`sigma`) and the mask undergoes morphological closing with a spherical structuring element, being then applied to the image before the feature extraction. Isotropic resampling is also performed, which can be of $1 \times 1 \times 1$ mm [20]–[22]; however, $0.7 \times 0.7 \times 0.7$ mm was chosen so as not to impair the quality of the images, which, being of high resolution, took on a pixelated appearance.

In the supervoxel selection, a cubic $K\times K\times K$ window traverses the volume with a defined step; each block entirely

TABLE II
CHARACTERISTICS OF THE IMAGE SETS, MAGNETIC RESONANCE ACQUISITION PARAMETERS AND PRE-PROCESSING APPLIED.

Item	**Sagittal projection**	**Axial projection**
Acquisition		
Tumour type (histology/ staging)	Rectal cancer; no histological data; locoregional staging.	
Equipment and magnetic field	Signa HDx 1.5 T (GE Healthcare, Chicago, IL, USA).	
Weighting	T2-weighted (T2WI)	
Sequence	FRFSE	FRFSE
TR (ms)	3317	3184
TE (ms)	120	105
Slice thickness (mm)	3	3
Gap	0	0
Matrix	256×256	256×256
Number of signals acquired	6	8
FOV (cm^2)	20–24	16
Bandwidth (kHz)	50	31.25
In-plane spacing	0.3 mm.	
Scanner across cases	Same model and same manufacturer.	
Segmentation and pre-processing		
Origin of the manual mask	Manual delineation in ITK-SNAP, version 4.0.2	
Registration between sequences	Not applied.	
Normalisation of intensities and of the *features*	The intensities are rescaled to the range of 0 to 255 gray levels and the z-score is applied to the extracted features, before the clustering	
Resampling	Resampling to an isotropic spacing of 0.7 mm (SimpleITK), with linear interpolation for the image and nearest neighbour for the mask, on the same grid.	
Lesion volume	Case 1 (32.07 cm³), Case 2 (19.97 cm³), Case 3 (6.23 cm³), Case 4 (4.60 cm³).	

contained within the mask and without null values is considered valid, and the coordinates of its voxels are stored. Only the voxels in the interior of the volume (with a complete $3\times3\times3$ neighbourhood) are considered in the computation of the features, with the border voxels being discarded, as well as the blocks that do not gather at least K^3 non-null voxels.

For each supervoxel, 17 first-order features and 11 texture features derived from the gray-level co-occurrence matrix (GLCM) are computed, detailed in Table IV [9], [19]; the GLCM is obtained in a $3\times3\times3$ neighbourhood around each voxel, with local quantisation into 8 levels and 13 directions. The functions were implemented manually based on the PyRadiomics documentation [11], without the use of the library owing to its lack of maintenance, and part of the construction of the GLCM was based on [19].

In the clustering, the features of the four cases are concatenated, standardised (z-score) and grouped together by K-means with $k = 3$; since the concatenation takes place before the grouping, there is habitat correspondence within the same run, and each label receives a colour. Table V gathers the parameters and the configuration used in the code.

The division of the image into regular *grids* is only a precursor of supervoxels, and many articles do not use this term for these blocks; nevertheless, since they play the same role (in the sense of *oversegmentation*), they will be referred to, for simplicity, as supervoxels throughout the work.

TABLE IV
FIRST- AND SECOND-ORDER FEATURES EXTRACTED FROM EACH SUPERVOXEL.

# First Order	**# Second Order (GLCM)**
1 `mean`	18 `asm`
2 `std_dev`	19 `glcm_entropy`
3 `median`	20 `contrast`
4 `skewness`	21 `homogeneity`
5 `minimum`	22 `correlation`
6 `maximum`	23 `variance`
7 `range`	24 `sum_mean`
8 `kurtosis`	25 `max_probability`
9 `energy`	26 `idm`
10 `MAD_feature`	27 `cluster_tendency`
11 `P10`	28 `dissimilarity`
12 `P25`	
13 `P75`	
14 `P90`	
15 `interquartile_range`	
16 `entropy`	
17 `uniformity`	

3) Varied parameters: The extraction of the *features* and the formation of the habitats were performed 13 times, each with a distinct set of parameters, and one slice

TABLE V
PARAMETERS AND CONFIGURATIONS USED IN THE CODE.

Parameter / Item	Value	Description
Partitioning and pre-processing		
`kernel_size`	3	Supervoxel size ($3 \times 3 \times 3 = 27$ voxels).
`step`	1	Step of the sliding window in the block partitioning (step $<$ *kernel* produces overlap).
`sigma`	0.4	Sigma of the Gaussian filter applied to the image before the extraction.
`closing_radius`	1	Radius of the sphere of the morphological closing of the mask.
`iterations`	1	Number of iterations of the morphological closing.
`novo_spacing`	0.7 mm	Isotropic spacing ($x = y = z$) of the resampling.
`normalizar`	`True`	Rescaling of the intensities to the range of 0 to 255 gray levels.
Feature extraction		
Families extracted	17 + 11	First order (17 features) and GLCM texture (11 features)
`n_levels` (GLCM)	8	Gray levels in the local quantisation of the GLCM.
Clustering		
normalisation	*z-score*	Applied to the features before the clustering (mean = 0, standard deviation = 1).
`n_clusters` (k)	3	Number of habitats.
initialisation	`k-means++`	Default of `scikit-learn`.
`n_init`	10	Number of initialisations (restarts) of K-means.
`max_iter`	300	Maximum number of iterations per K-means initialisation.
`random_state`	42	Seed for reproducibility.

was chosen for the visualisation of each configuration. The reference configuration uses `sigma=0.4`, `kernel=3`, `closing_radius=1` and `step=1`, and only one parameter was varied at a time. Table VI summarises, for each set, the varied parameter, its description and the values used.

TABLE VI
TEST SETS: VARIED PARAMETER. IN EACH SET ONLY ONE PARAMETER IS VARIED; THE OTHERS REMAIN AT THE REFERENCE CONFIGURATION (`SIGMA=0.4`, `KERNEL=3`, `CLOSING_RADIUS=1`, `STEP=1`).

Set	Parameter	Brief description	Values
1	Filtering	Gaussian filter.	sigma = 0.2; 0.4; 0.8; 1
2	Window	Size of the $K \times K \times K$ block in the voxel selection.	3x3x3; 5x5x5; 7x7x7; 9x9x9
3	Structuring element	Radius of the spherical closing element.	radius = 1; 2; 3; 4
4	Step	Displacement of the window.	step = 1; 2; 3; 4

4) Metrics applied: The comparison with the reference configuration was performed with the adjusted Rand index and the variation of volume and proportions and the visual appearance of the habitats were also evaluated.

III. RESULTS

All 13 runs can be seen in Fig. 2. The colours in the images correspond within the same run, because the cases were clustered together, but different runs do not present this correspondence among themselves. In this figure, the visual effect of the parameter variation and its impact on the habitats were evaluated.

Increasing the sigma resulted in more homogeneous areas within each habitat, which captured less detail, and this affected the adjusted Rand index values (Fig. 3). The highest values were obtained for sigma = 0.2 (0.74–0.87) and the lowest for sigma = 1 (0.35–0.61). Varying the window size reduced the detail of the images, but cases 1 and 2 were less affected than the others, probably because they have more voxels. The highest values were obtained for the $5\times5\times5$ window (0.15–0.55) and the lowest for $9\times9\times9$ (0.11–0.32), and case 4 did not present valid supervoxels with this size. Increasing the radius of the closing element prioritised the most abundant habitats, as shown in Fig. 4: as the radius increases, the bars tend to present only two habitats. The highest values were obtained for radius = 2 (0.36–0.58) and the lowest for radius = 4 (0.32–0.54), but the indices were close among the three radii tested, not counting the reference one. In the step variation, the results were good for step = 2 and step = 3, despite the variations in the adjusted Rand index. The drop was more pronounced for step = 4, which, being larger than the default $3\times3\times3$ window, produced gaps in the images, as shown in Fig. 2.

Table VII shows the execution time of the 13 configurations. Varying the step was what most reduced the execution time, and varying the radius of the structuring element was what least affected it. In the window variation, the time peaks at $5\times5\times5$ (1219.3 s) and falls at $7\times7\times7$ and $9\times9\times9$, since larger windows accept fewer blocks.

IV. DISCUSSION

This work presented a tumour habitat formation algorithm based on a regular *grid* of voxels, on the extraction of first-order and texture radiomic features and on K-means clustering, evaluating the effect of four parameters: the sigma of the

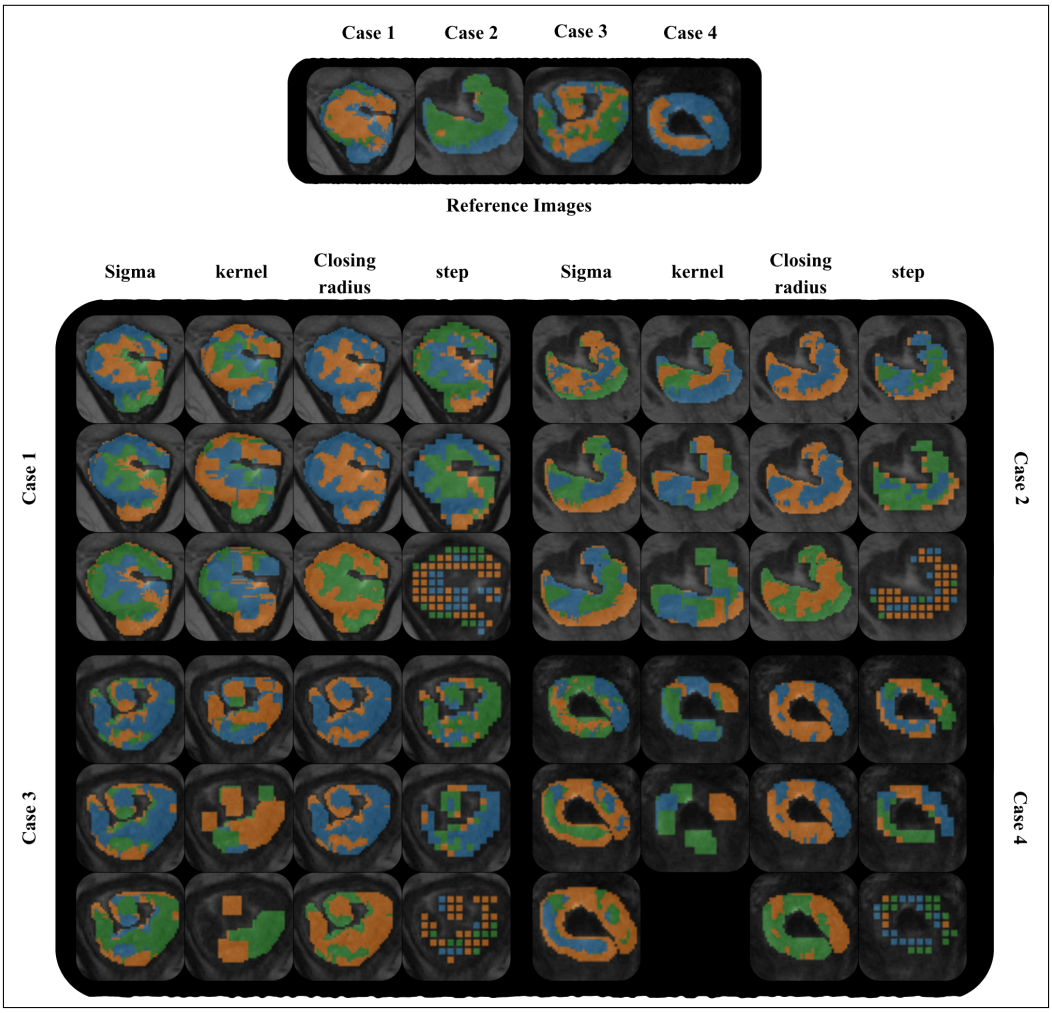


Fig. 2. Visual habitat results for the variations of the sigma, the window, the structuring element radius and the step. The cases are separated into four blocks, and the images at the top correspond to the reference configuration of each case. In each block, the first column shows the variations of the sigma (0.2, 0.4, 0.8 and 1); the second, the window dimensions (5×5×5, 7×7×7 and 9×9×9); the third, the closing radius (2, 3 and 4); and the fourth, the step (2, 3 and 4).

TABLE VII
EXECUTION TIME OF EACH CONFIGURATION. IN EACH SET ONLY ONE PARAMETER IS VARIED RELATIVE TO THE REFERENCE CONFIGURATION (`SIGMA=0.4`, `KERNEL=3`, `CLOSING_RADIUS=1`, `STEP=1`); THE OTHERS REMAIN AT THE REFERENCE VALUES.

Varied parameter	**Value (time in s)**		
Reference	531.6		
Sigma	0.2 (717.4)	0.8 (1286.4)	1 (1707.3)
Window	5 (1219.3)	7 (1013.6)	9 (800.7)
Radius	2 (613.9)	3 (581.8)	4 (667.4)
Step	2 (333.4)	3 (313.0)	4 (243.6)

Gaussian filter, the window size, the radius of the closing structuring element and the step of the sliding window. The tests were performed on four rectal MRI cases from a single institution, acquired with the protocol described in Table II. Since the features of the four cases are concatenated before the clustering, all the cases of a given run share the same centroids, so that no displacement of the centroids occurs from one case to another and the habitats are directly comparable within that run.

A. Filtering

Filtering is necessary to prevent the features from being computed over the noise of the image, but, in excess, it can reduce the information present. Thus, very noisy images may lose diagnostic potential if they cannot be filtered without a loss of relevant information. The results showed this effect, with a slight loss of detail as the sigma increased.

B. Window

The code requires all the values inside the window to be greater than zero (`np.all(current_mask > 0)`), that is, the whole block must lie inside the mask, which has already undergone the closing, in order to avoid gaps. Therefore, the number of accepted blocks decreases as the *kernel* grows, since requiring a large cube (7×7×7) to lie entirely inside the mask is more restrictive than accepting a small one (3×3×3). In very

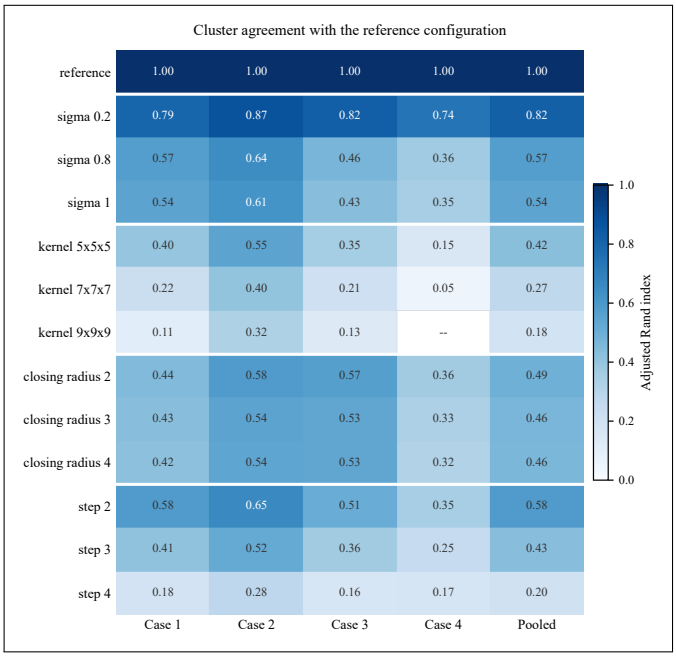


Fig. 3. Adjusted Rand index (ARI) between each configuration and the reference configuration ($\sigma = 0.4$; window $3\times3\times3$; radius 1; step 1). Each row is a configuration and each column is a case; the last column brings the pooled ARI, computed over the voxels of all the cases concatenated. The first row is the reference compared with itself. Cells without a value indicate configurations that did not generate any block in that case. The ARI is invariant to the permutation of the labels, so that the comparison does not depend on which colour was assigned to each habitat.

small masks, large windows may generate few or no blocks, as the images show: larger tumours had enough elements for windows of up to $7\times7\times7$ to return results, which did not occur in the smaller ones. The window size may also interfere with the computation of the features, which must be taken into account, since some slices presented regions that "changed" habitat, something that does not happen biologically.

C. Closing structuring element

The closing fills the holes formed in the mask by non-segmented regions along the volume, allowing more voxels to be processed and avoiding the discarding of groups owing to null values between one slice and its neighbours in the z direction. The images show that the shape of the tumour varies between the slices and, if the mask has many gaps or is very irregular, the operation may not be sufficient to include a good number of voxels in the analysis. In these cases, or if the mask was not well made, it may be worth redoing the segmentation in order to allow the use of smaller structuring elements.

D. Step

The choice of the step must be based mainly on the dimensions of the window *kernel*, in order to avoid gaps in the voxel coverage, and, indirectly, on the size of the tumour. Overlap may occur, since only the central voxels receive the K-means classification and, depending on the step, more than one voxel may receive the classification of the same block.

- Step $<$ *kernel* dimension $\rightarrow$ the blocks overlap, and the last block that passed over a voxel "wins" the overlap, assigning its K-means label to that voxel.
- Step $=$ *kernel* dimension $\rightarrow$ the blocks touch without overlap and without gaps, with perfect coverage in the *array* (each voxel belongs, or belonged, to only one block).
- Step $>$ *kernel* dimension $\rightarrow$ the blocks become separated and gaps appear. In order to use a step larger than K (the dimension of the sliding window), it may be necessary to apply an interpolation, such as *nearest neighbor* or linear [28].

The sigma and the radius preserve the number of voxels evaluated, and their main effect is the loss of information, that is, the reduction of the level of detail of the habitats. The window and the step, on the other hand, reduce the total number of voxels evaluated: increasing the window leaves many portions of the image without coverage, especially in the smaller cases, whereas increasing the step reasonably preserves the composition, although the subsampling may reduce the available information for large values.

E. Execution time

The execution times reported in Table VII are specific to this implementation, but what is interpretable is the relative behaviour between the configurations. The execution time is important because medical images are large and can be difficult to process, and slow algorithms would not be efficient in the medical routine. The parameter that most reduced the time, the step, does so by processing fewer blocks rather than by processing them faster, which is a trade-off between speed and information. The configurations that are cheapest to compute are also the ones that cover less of the tumour. It is not advantageous to sacrifice so much information for speed: some of these methods can be applied, but only up to a certain limit, and, for this reason, when the loss becomes large, it may be worth optimising and parallelising the code or improving the available processing resources.

F. Generalisability

The effects described are limited to these cases and are, initially, not generalisable to other modalities, scanners or sets of pre-processing tools. The performance of prognostic models is sensitive to the *pipeline* and dependent on the sequence used, and MRI intensities are acquired in arbitrary, scanner-dependent units, which prevents the direct comparison of the examinations, even with the same acquisition protocol [24]. Moreover, there are fewer reproducible *features* between different scanners, even with identical acquisition parameters [25], and, although normalisation can be applied in these cases, the normalisation of the features improves reproducibility more than the normalisation of the images [27].

Resampling to isotropic voxels can also improve robustness [26], but above all it corrects the anisotropy, making the parameters (window, step, sigma, radius) comparable in physical units. The variation of the *kernel* dimensions, in turn, may

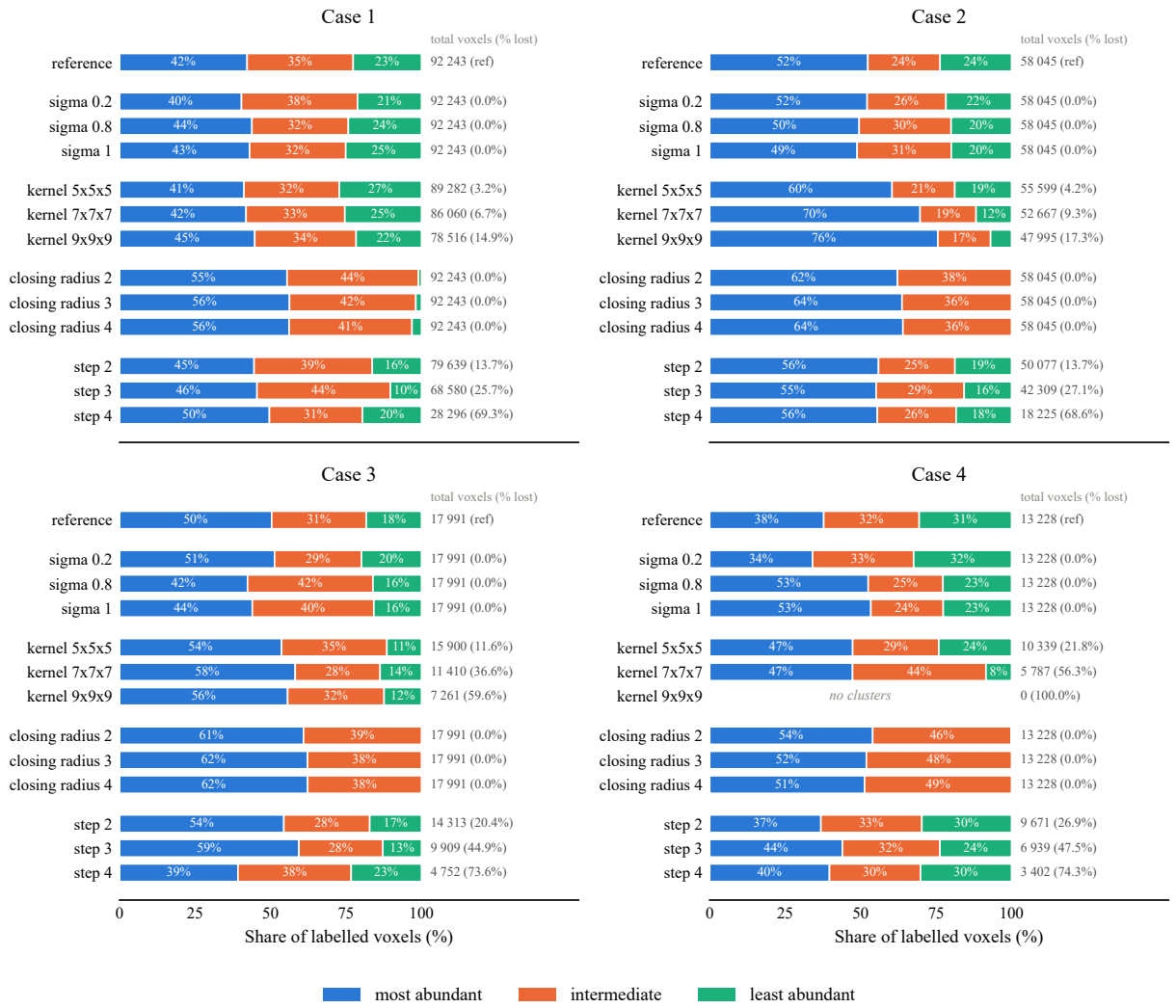


Fig. 4. Proportion of each habitat and total of labelled voxels, per configuration and per case. Each panel corresponds to a case and each bar to a configuration, with the reference at the top and the four parameter families separated by a blank line. Within each bar, the habitats are ordered from the most abundant to the least abundant, and not by label, because the clustering is fitted independently in each run and the labels are not comparable between runs. The number to the right of each bar is the total of labelled voxels in that configuration, followed in parentheses by the percentage of voxels lost relative to the reference configuration of the same case.

reduce the repeatability and the reproducibility of 3D radiomic features [23]. Thus, without testing which features vary in a given context, there is no way to state that they will remain stable in another.

This work can be understood as a basis for future studies and not as a set of recommended values. Although the parameter values are not generalisable, the qualitative behaviour (a larger sigma reduces detail; a larger window reduces coverage in small tumours) tends to hold, as it is a geometric effect of the method rather than a property of the data.

V. Conclusion

This article discussed, in four rectal MRI cases, the effect of the parameter choice on a habitat formation algorithm based on K-means and radiomic features, which employs filtering, a structuring element and a sliding window with a step. Other algorithms and clustering methods can be explored with the same objective: to make accessible the data that support the standardisation and the reproducibility of radiomics studies (a field in which many works present results without the support needed to reproduce them), to improve the method and to guide the choice of parameters by researchers in the same line of research.

The adjusted Rand index is computed over the intersection of the labelled voxels and does not report the portion of the tumour that was left out of the evaluation; the visualisations and the proportion plot complement it. In the variations of sigma and radius, the total of labelled voxels equals that of the reference in all four cases and the comparison covers the complete set; in those of window and step, this total varies, and the index describes the agreement within a subset that may

be small, so that a high value would not guarantee that the division of the tumour as a whole had been maintained. The proportions in Fig. 4, in turn, are ordered by abundance, and not by habitat identity: the habitats are equivalent within a run, but not between runs. Verifying the stability and the robustness of the *features* would be a proposal for future work capable of complementing the data needed to increase the reproducibility of the study.

CODE AVAILABILITY

The developed code and the installation and usage instructions are available at https://github.com/sarahsouza200054/CBEB_2026_Sarah. The MRI images are not part of the repository.

ACKNOWLEDGMENT

The authors thank the funding agencies CAPES and CNPq, as well as the institutions and collaborators who contributed to this work. The translation into English and the foreign-language revision of this manuscript were carried out with the aid of artificial intelligence tools, under the supervision of the authors and with a final revision by them.

REFERENCES

[1] J. Wu, G. Cao, X. Sun, J. Lee, D. L. Rubin, S. Napel, A. W. Kurian, B. L. Daniel, and R. Li, "Intratumoral spatial heterogeneity at perfusion MR imaging predicts recurrence-free survival in locally advanced breast cancer treated with neoadjuvant chemotherapy," *Radiology*, vol. 288, no. 1, pp. 26–35, 2018, doi: 10.1148/radiol.2018172462.

[2] I. Luengo, M. C. Darrow, M. C. Spink, Y. Sun, W. Dai, C. Y. He, W. Chiu, T. Pridmore, A. W. Ashton, E. M. H. Duke, M. Basham, and A. P. French, "SuRVoS: Super-Region Volume Segmentation workbench," *Journal of Structural Biology*, vol. 198, no. 1, pp. 43–53, 2017.

[3] M. Tamajka and W. Benešová, "Supervoxel algorithm for medical image processing," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Melmaruvathur, India, 2017, pp. 3121–3127, doi: 10.1109/ICPCSI.2017.8392300.

[4] B. Irving, J. M. Franklin, B. W. Papież, E. M. Anderson, R. A. Sharma, F. V. Gleeson, M. Brady, and J. A. Schnabel, "Pieces-of-parts for supervoxel segmentation with global context: application to DCE-MRI tumour delineation," *Medical Image Analysis*, vol. 32, pp. 69–83, 2016.

[5] S. Hansen, S. Kuttner, M. Kampffmeyer, T.-V. Markussen, R. Sundset, S. Kjærnes Øen, L. Eikenes, and R. Jenssen, "Unsupervised supervoxel-based lung tumor segmentation across patient scans in hybrid PET/MRI," *Expert Systems With Applications*, vol. 167, p. 114244, 2021.

[6] R. C. Gonzalez and R. E. Woods, *Processamento Digital de Imagens* (in Portuguese), 3rd ed. São Paulo, Brazil: Pearson Prentice Hall, 2010.

[7] S. Aja-Fernández and G. Vegas-Sánchez-Ferrero, *Statistical Analysis of Noise in MRI: Modeling, Filtering and Estimation*. Cham, Switzerland: Springer, 2016.

[8] M. Elhoseny and K. Shankar, "Optimal bilateral filter and convolutional neural network based denoising method of medical image measurements," *Measurement*, vol. 143, pp. 125–135, 2019, doi: 10.1016/j.measurement.2019.04.072.

[9] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. J. R. Cook, "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, Apr. 2020, doi: 10.2967/jnumed.118.222893.

[10] A. S. Kurani, D.-H. Xu, J. D. Furst, and D. S. Raicu, "Co-occurrence matrices for volumetric data," in *7th IASTED Int. Conf. Computer Graphics and Imaging*, Kauai, HI, USA, 2004.

[11] "Pyradiomics documentation." Available: https://pyradiomics.readthedocs.io/en/latest/ (accessed Sep. 10, 2026).

[12] R. A. Lotufo, R. Audigier, A. V. Saúde, and R. C. Machado, "Morphological image processing," in *Microscope Image Processing*, 2nd ed., F. A. Merchant and K. R. Castleman, Eds. London, U.K.: Academic Press, 2023, pp. 75–117, doi: 10.1016/B978-0-12-821049-9.00012-5.

[13] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[14] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into Imaging*, vol. 11, no. 91, pp. 1–16, Aug. 2020, doi: 10.1186/s13244-020-00887-2.

[15] Y. Gao, Z. Li, X. Zhai, G. Zhang, L. Zhang, T. Huang, L. Han, J. Wang, R. Yan, Y. Li, H. Zhao, Q. Zhao, Z. Wei, B. Xie, Y. Sun, J. Zhao, and H. Cui, "MRI-based habitat radiomics combined with vision transformer for identifying vulnerable intracranial atherosclerotic plaques and predicting stroke events: a multicenter, retrospective study," *eClinicalMedicine*, vol. 82, p. 103186, Apr. 2025, doi: 10.1016/j.eclinm.2025.103186.

[16] E. Sala, E. Mema, Y. Himoto, H. Veeraraghavan, J. D. Brenton, A. Snyder, B. Weigelt, and H. A. Vargas, "Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging," *Clinical Radiology*, vol. 72, no. 1, pp. 3–10, Jan. 2017.

[17] C. Charbel, H. C. Kwok, J. Miranda, J. Zheng, M. El Homsi, M. A. El Amine, S. Chhabra, S. Danilova, N. Gangai, I. Petkovska, M. Capanu, R. S. Vanguri, J. Chakraborty, and N. Horvat, "Reliability of rectal MRI radiomic features: comparing rectal MRI radiomic features across reader expertise levels, image segmentation technique, and timing of rectal MRI in patients with locally advanced rectal cancer," *European Journal of Radiology*, 2025, doi: 10.1016/j.ejrad.2025.112019.

[18] scikit-learn, "*sklearn.metrics*: `calinski_harabasz_score`, `davies_bouldin_score`." Available: https://scikit-learn.org/stable/api/sklearn.metrics.html (accessed Jun. 2026).

[19] S. Barburiceanu, R. Terebes, and S. Meza, "3D texture feature extraction and classification using GLCM and LBP-based descriptors," *Applied Sciences*, vol. 11, no. 5, p. 2332, Mar. 2021.

[20] J. Antunes, P. Prasanna, A. Madabhushi, P. Tiwari, and S. Viswanath, "RADIomic spatial textural descriptor (RADISTAT): characterizing intra-tumoral heterogeneity for response and outcome prediction," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham, Switzerland: Springer, 2017, pp. 468–476.

[21] C. Sui, Q. Su, K. Chen, R. Tan, Z. Wang, Z. Liu, W. Xu, and X. Li, "^{18}F-FDG PET/CT-based habitat radiomics combining stacking ensemble learning for predicting prognosis in hepatocellular carcinoma: a multi-center study," *BMC Cancer*, vol. 24, no. 1, p. 1457, 2024.

[22] P.-Q. Wu, F.-L. Guo, J. Wang, Y. Gao, S.-T. Feng, S.-L. Chen, J. Ma, and Y.-B. Liu, "Development and validation of a dynamic contrast-enhanced magnetic resonance imaging-based habitat and peritumoral radiomic model to predict axillary lymph node metastasis in patients with breast cancer: a retrospective study," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 12, p. 8211, 2024.

[23] O. Prior, C. Macarro, V. Navarro, C. Monreal, M. Ligero, A. Garcia-Ruiz, G. Serna, S. Simonetti, I. Braña, M. Vieito, M. Escobar, J. Capdevila, A. T. Byrne, R. Dienstmann, R. Toledo, P. Nuciforo, E. Garralda, F. Grussu, K. Bernatowicz, and R. Perez-Lopez, "Identification of precise 3D CT radiomics for habitat computation by machine learning in cancer," *Radiology: Artificial Intelligence*, vol. 6, no. 2, p. e230118, 2024, doi: 10.1148/ryai.230118.

[24] P. Salome, F. Sforazzini, G. Grugnara, A. Kudak, M. Dostal, C. Herold-Mende, S. Heiland, J. Debus, A. Abdollahi, and M. Knoll, "MR intensity normalization methods impact sequence specific radiomics prognostic model performance in primary and recurrent high-grade glioma," *Cancers*, vol. 15, no. 3, p. 965, 2023, doi: 10.3390/cancers15030965.

[25] H. Mi, M. Yuan, S. Suo, J. Cheng, S. Li, S. Duan, and Q. Lu, "Impact of different scanners and acquisition parameters on robustness of MR radiomics features based on women's cervix," *Scientific Reports*, vol. 10, p. 20407, 2020, doi: 10.1038/s41598-020-76989-0.

[26] H. Um, F. Tixier, D. Bermudez, J. O. Deasy, R. J. Young, and H. Veeraraghavan, "Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets," *Physics in Medicine & Biology*, vol. 64, no. 16, p. 165011, 2019, doi: 10.1088/1361-6560/ab2f44.

[27] V. Giannini, J. Panic, D. Regge, G. Balestra, and S. Rosati, "Could normalization improve robustness of abdominal MRI radiomic features?," *Biomedical Physics & Engineering Express*, vol. 9, no. 5, p. 055002, 2023, doi: 10.1088/2057-1976/ace4ce.

[28] T. M. Lehmann, C. Gönner, and K. Spitzer, "Survey: Interpolation methods in medical image processing," *IEEE Transactions on Medical Imaging*, vol. 18, no. 11, pp. 1049–1075, Nov. 1999, doi: 10.1109/42.816070.

Heart sound detection based on wavelet transform and Shannon energy envelope

João Vitor de Amorim Paula
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brazil
amorimjv20@gmail.com

2nd Renata Valeri de Freitas
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brazil
renata.valfreitas@usp.br

***Abstract*—Valvular heart diseases contribute significantly to global morbidity, making phonocardiography (PCG) an essential non-invasive diagnostic tool. However, the detection of heart sounds (S1 and S2) is frequently hindered by the presence of pathological murmurs and instrumentation noise. This paper proposes an algorithm for the detection of S1 and S2 heart sounds. The method integrates the Discrete Wavelet Transform (DWT) as preprocessing for noise removal, 2nd-order Shannon energy envelope and a dynamic adaptive threshold based on local peaks, followed by postprocessing rules for artifact rejection and physiological segment junction. The algorithm was validated on the public CirCor database using an event-based metric. The results obtained from a total of 787 patients and 31,210 annotated heart sounds demonstrated consistent overall performance, achieving a Sensitivity of 93.48%, Precision of 87.29%, and an F1-Score of 90.28%, when compared against the same algorithm without the preprocessing stage, which yielded gains of 4.01pp in Sensitivity, 2.2pp in Precision and 3.05pp in F1-Score, thereby demonstrating an advantage in its utilization.**

***Index Terms*—Phonocardiogram, wavelet transform, Shannon energy, heart segmentation, heart sounds.**

I. Introduction

Cardiovascular diseases represent the leading cause of mortality worldwide, accounting for an estimated 17.9 million deaths annually according to the World Health Organization (WHO) [1]. Although coronary artery disease is the most common, valvular heart diseases also contribute to global morbidity and mortality. Heart sounds (HS) are vital for early detection and warning of these pathologies, as they contain physiological and pathological information that the electrocardiogram (ECG) often fails to capture, such as murmurs caused by mechanical valve problems [2]. Historically, cardiac auscultation has been the primary technique for this verification; however, its effectiveness is limited by the constraints of human auditory perception and physician subjectivity [2].

To overcome these limitations, digital phonocardiography (PCG) has been adopted as a non-invasive and easily acquired alternative, allowing the recording, graphical visualization, and objective processing of the acoustic signals of the cardiac cycle [2], [3]. However, event detection in PCG faces major challenges: the signals are relatively weak and highly susceptible to environmental noise and instrumentation interference, resulting in noisy data that reduces the accuracy of parameter extraction and increases diagnostic uncertainty [3]. Furthermore, the randomness and variability of cardiovascular diseases lead to enormous complexity and diversity in signal manifestation [3], often masking heart sounds with murmurs characteristic of the pathologies [4].

The cardiac cycle can produce up to four acoustic heart sounds (S1, S2, S3, and S4), which reflect the mechanical and hemodynamic dynamics of the heart [5], [6]. The two main ones, present in a normal cycle, are the first heart sound (S1), generated at the onset of ventricular systole by the closure of the mitral and tricuspid valves, and the second heart sound (S2), which marks the onset of diastole with the closure of the aortic and pulmonary valves. Occasionally, a third heart sound (S3) and a fourth heart sound (S4) can be recorded [5]. S3 occurs during the period of rapid passive ventricular filling in early diastole, being considered physiological in children and young adults, but frequently pathological in adults, associated with cardiac dilation or failure. S4, in turn, occurs in late diastole, immediately before ventricular systole, produced by vibrations in the ventricle when the atria contract; although inaudible S4 vibrations can be recorded by PCG in normal individuals, its audible prominence indicates abnormalities such as reduced ventricular distensibility (stiffness) [5], [6]. The accurate detection and segmentation of S1 and S2 sounds is a fundamental step in any phonocardiography system, as it establishes the temporal reference needed to diagnose and locate these murmurs and extra sounds.

To extract the true heart sounds among various types of murmurs and noise, the literature has adopted a wide range of advanced techniques. Machine learning-based approaches, Hidden Markov Models (HMM), and Deep Learning have been widely explored for PCG signal segmentation [3], [7]. Although they present high accuracy rates, these models often require high computational costs and extensive training databases, which hinders their deployment in portable devices [3] and the broad application of these techniques.

Concurrently, in signal processing methodologies, the wavelet transform has stood out for its ability to efficiently denoise the PCG signal, attenuating murmur energy without compromising the time-frequency characteristics of the S1 and S2 sounds [2], [4]. After this initial filtering, the use

of energy envelopes, such as Shannon energy, is applied to track and enhance heart sounds, isolating them from residual background noise [8], [9]. In this context, classical methods based on static thresholds frequently fail to accurately identify these HS because of the inherent non-stationarity of the signal, intrapatient variability and the presence of artifacts may cause amplitude fluctuations, causing the desired heart sounds to be attenuated or confused with murmurs [9].

Given this scenario, this work proposes a heart sound detection algorithm. The main points of the proposed algorithm are: (1) The use of the wavelet transform for murmur attenuation coupled with the Shannon energy envelope; (2) An adaptive thresholding method based on dynamic peak tracking; and (3) The proposition of an event-based evaluation focused on the midpoint of the heart sounds. The algorithm was validated on the public CirCor database [10].

II. Materials and methods

The detection algorithm proposed in this work was developed using MATLAB [11] and consists of six main stages: pre-processing via wavelet transform, Shannon energy envelope, normalization of the envelope, dynamic thresholding based on local peaks, artifact removal, and the junction of split heart sounds.

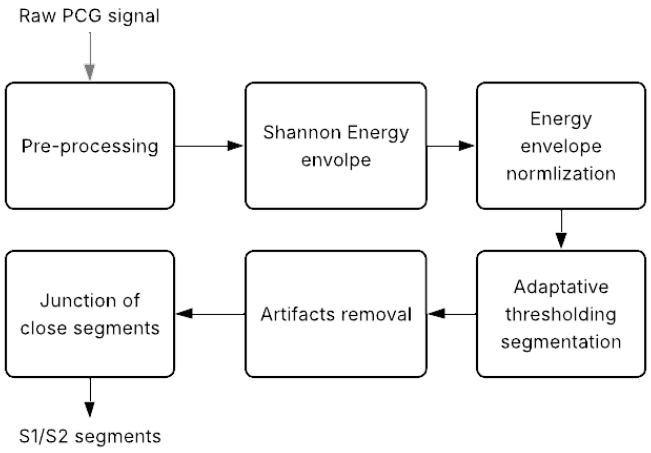


Fig. 1. Algorithm pipeline block diagram

A. Database

The phonocardiogram signals used for the development and validation of the algorithm were extracted from the public CirCor database, which was consolidated through a partnership between the Real Hospital Português de Beneficência and the Hospital Universitário Oswaldo Cruz (HUOC) [10]. This database contains records of patients categorized as "Normal" and "Abnormal", encompassing a great diversity of cases.

Given that the CirCor database focuses on the pediatric population, the age range was classified according to standardized clinical definitions: Neonatal (0 to 27 days), Infant (28 days to 1 year), Child (1 to 11 years), and Adolescent (12 to 18 years). Table I summarizes the characteristics of the database population, detailing weight, height, Body Mass Index, and the distribution between sex and age group.

The patients were designated a category (*Normal* or *Abnormal*) based the pediatric cardiologist's overall assessment of the patient's cardiac condition. This is a diagnosis based on clinical history, physical examination, analog auscultation, echocardiogram, etc. The cardiologists did not have access to the digital auscultations and the murmur gradings when making these diagnoses, and the indicated outcomes do not necessarily imply that the expert had identified a murmur during analog auscultation. The pediatric cardiologists were not the annotators of the digital auscultations.

For this study, we exclusively used records captured at the aortic valve auscultation site. The original sampling rate was maintained at 4000 Hz.

TABLE I
CirCor database patients characteristics

	Min	**Max**	$\mu \pm \sigma$
Weight (kg)	2.3	108.6	23.44 ± 14.7
Height (cm)	35	180	111.08 ± 29.15
BMI	6.18	36.97	17.79 ± 8.89
	Distribution (N and %)		
Sex			
Male	389 (49.43%)		
Female	398 (50.57%)		
Age group			
Neonatal	4 (0.51%)		
Infant	96 (12.2%)		
Child	579 (73.57%)		
Adolescent	60 (7.62%)		
Not informed	48 (6.1%)		
PCG category			
Normal	394 (50.06%)		
Abnormal	393 (49.94%)		

B. Pre-processing

PCG signals are inherently non-stationary and frequently contaminated by low-frequency noise, such as respiratory movements or stethoscope friction, and high-frequency noise, such as electronic interference or ambient noise. Before applying the detection algorithm, the records were submitted to a pre-processing stage to ensure data integrity.

First, each record was temporally segmented. The signal was truncated to maintain exclusively the intervals between the first and last annotations made by the HUOC specialists.

After cropping the validated intervals, the Discrete Wavelet Transform (DWT) was applied for noise suppression (denoising). The signal was decomposed into 8 resolution levels using the 8th-order wavelet of the Daubechies family (db8) [12]. In DWT, the signal is passed through a series of high-pass and low-pass filters, resulting in detail coefficients (D_j) and approximation coefficients (A_j), respectively, which provide a multi-resolution time-frequency analysis.

Given the original sampling rate of 4000 Hz, the Nyquist frequency is 2000 Hz. Through frequency partitioning, the detail coefficients D_5, D_6 and D_7 correspond to the frequency

bands of 15.625 to 125 Hz. Since the predominant components of S1 and S2 heart sounds are concentrated in the frequency range between approximately 30 Hz and 70 Hz [13], the filtered PCG signal was reconstructed exclusively from the sum of these three detail coefficients ($D_5 + D_6 + D_7$), the choice of using the seventh level of detail coefficient is to encompass some low frequency components that are present in some patients HS, it does not usually coincide with the frequency components of undisered noise or murmurs.

Thus, the DWT acts as a band-pass filter, trying to isolate the heart sounds and attenuate murmurs and noise without causing distortion or altering the position of the heart sounds. Finally, the resulting signal was normalized by its absolute maximum value. An example can be seen in Fig. 2.

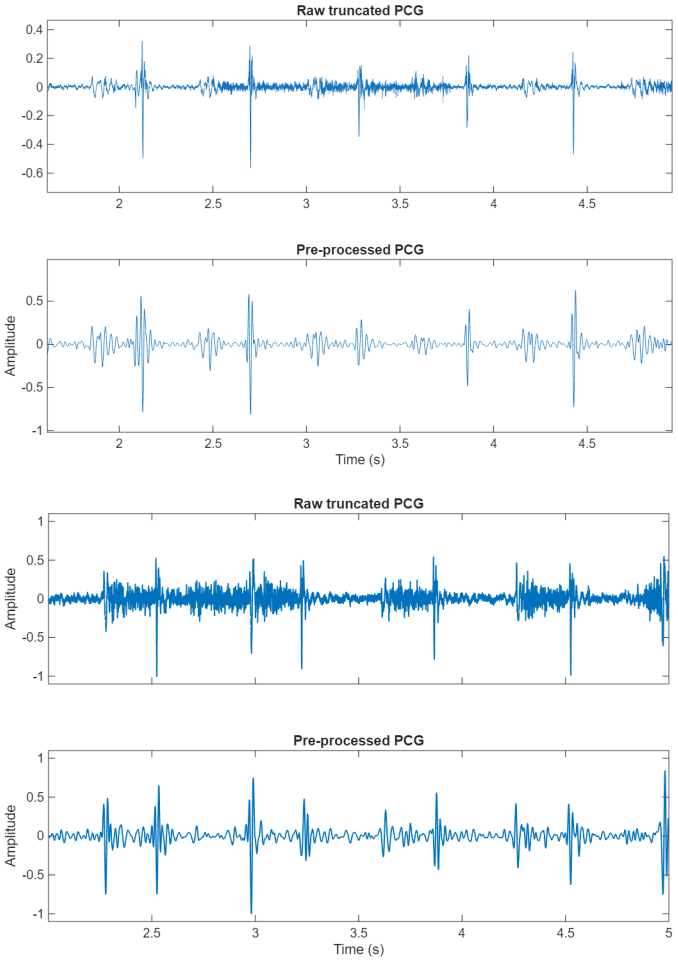


Fig. 2. Comparison of the DWT pre-processing stage. Top: Raw and pre-processed signals for a abnormal patient. Bottom: Raw and pre-processed signals for an abnormal patient with a strong murmur.

C. Shannon energy envelope extraction

Robust detection and segmentation of heart sounds require a representation that highlights short-duration events (S1 and S2) while suppressing background noise and undesired murmurs. For this purpose, the filtered and normalized signal x_i is processed through the 2nd-order Shannon energy envelope. This choice is based on the logarithmic function's ability to emphasize medium-intensity variations, which characterize heart sounds, disproportionately attenuating low-amplitude noise and high-energy murmur interference [9], [14].

The 2nd-order Shannon energy is calculated following Equation 1 [14], [15]:

$$E_s(i) = -\frac{1}{N}\sum_{k=1}^{N} x_k^2 \cdot \log(x_k^2) \tag{1}$$

Where N represents the analysis window size (defined as 80 samples, or 20 milliseconds) and x_k is the filtered PCG signal.

D. Energy envelope normalization

To obtain the final normalized envelope curve, statistical processing is applied to $E_s(i)$ aiming to mitigate the influence of remaining noise and baseline variations. The normalized envelope, $P_a(t)$, is obtained by standardizing the energy samples according to Equation 2 [14], [15]:

$$P_a(t) = \frac{E_s(t) - \mu(E_s(t))}{\sigma(E_s(t))} \tag{2}$$

Where $\mu(E_s(t))$ and $\sigma(E_s(t))$ represent, respectively, the mean and standard deviation of energy within a sliding window of half a second, ensuring the window encompasses only approximately one heart sound at a time. This adaptive normalization is fundamental for the algorithm to maintain detection sensitivity even in signals with temporal variations in gain or acoustic intensity. Figure 3 shows a result of the Shannon energy envelope.

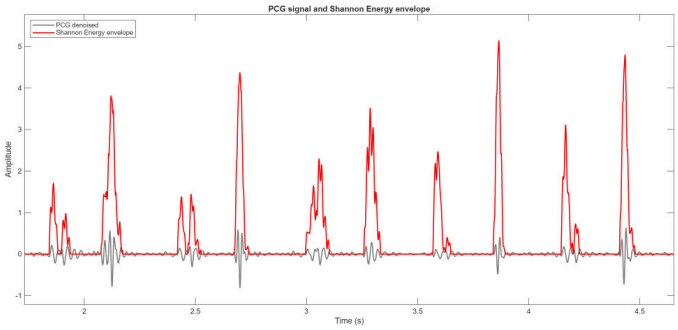


Fig. 3. Processing result after the extraction and normalization of the Shannon energy envelope

E. Adaptive Thresholding and Parameter Optimization

From the normalized envelope $P_a(t)$, the detection threshold must be calculated dynamically. To address limitations that classical methods based os statical thresholds present, the proposed algorithm employs an adaptive threshold based on the maximum energy peak within a 0.5-second sliding window, governed by a percentage factor (α).

To determine the optimal threshold factor and prevent empirical bias, a sweeping optimization was conducted across the training database. The parameter space was evaluated by

varying α from 0.1% to 100% while maintaining the fixed sliding window of 0.5 seconds. The performance trade-off was evaluated using a three-dimensional Precision-Sensitivity-F1 trajectory, which maps the algorithm's performance across the theoretical F1-Score surface, the equations on how Precision, Sensitivity and F1-score are calculated are presented ahead in subsection III-B.

The optimal operating point is visually represented, as illustrated in Fig. 4, by the peak of the trajectory scaling the F1-Score surface. This global maximum was achieved using an α of 10.6% of the local peak value ($Threshold = 0.106 \cdot P_{max,500ms}$). Envelope segments that exceed this dynamically calculated limit are considered initial candidates for S1 and S2 heart sounds. This optimal threshold factor was subsequently fixed for the global validation phase.

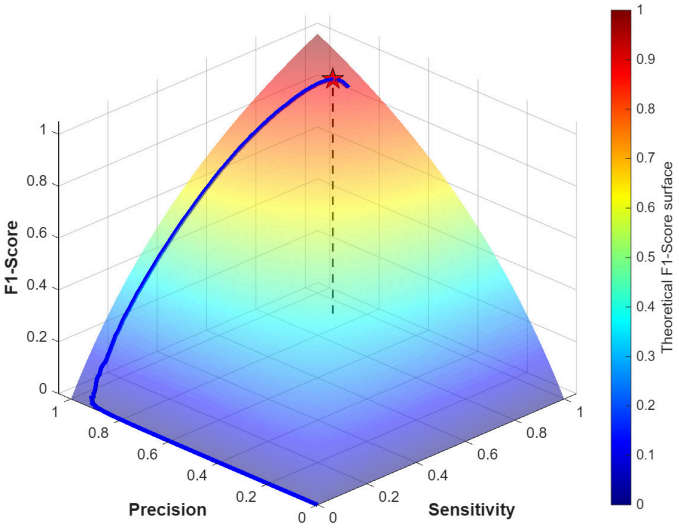


Fig. 4. Three-dimensional Precision-Sensitivity trajectory across the theoretical F1-Score surface for the adaptive threshold optimization. The red marker indicates the optimal threshold parameter ($\alpha = 10.6\%$) that maximizes the global F1-Score.

F. Artifact removal and segment junction

After initial candidates detection, two post-processing procedures are applied to refine the segmentation mask:

1) **Artifact removal:** Events with a duration shorter than 20 ms are classified as non-HS and discarded.
2) **Junction of close segments:** Due to the physiological phenomenon of heart sound splitting, a single heart sound can exhibit two distinct energy peaks separated by a brief interval of silence. According to the temporal stability criteria established by Liang et al. [15], the maximum normal splitting interval between the internal components of a healthy heart sound is approximately 50 ms. Therefore, if the silence interval between two consecutive detected energy peaks is less than 50 ms, they are classified as parts of the same split heart sound and are automatically merged into a single physiological segment.

Figure 5 presents the result of the detection, supression and junction of adjacent signals applied to a record.

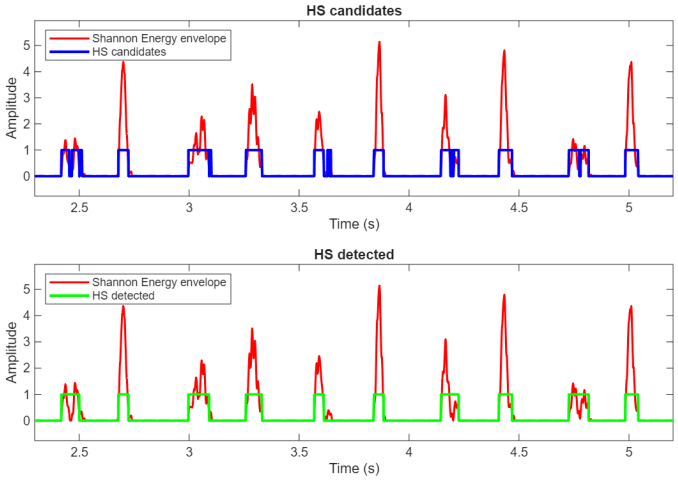


Fig. 5. Initial threshold detection (top) and after HS post-processing procedures (bottom)

III. Results

A. Evaluation criteria

To quantitatively evaluate the performance of the proposed algorithm, an event-based methodology was adopted. Thus, the decision criterion was based on the geometric midpoint of each detected heart sound, which represents the temporal center of the detected segment. This midpoint (n_{mid}) is calculated as the arithmetic mean of the starting (n_{start}) and ending (n_{end}) sample indices of the detected event. Consequently, the classification of each event is established as follows:

- **True positive (TP):** A detected event is classified as TP when its geometric midpoint is contained within the time window of the ground truth annotation (reference annotations by HUOC specialists).
- **False positive (FP):** Occurs when the algorithm identifies an event whose temporal location (midpoint) does not coincide with any reference annotation.
- **False negative (FN):** Occurs when a reference annotation provided by the specialist does not contain any detected heart sound midpoint from the algorithm.

.

B. Algorithm performance

The proposed algorithm was validated using a total sample of 787 patients (394 classified as "Normal" and 393 as "Abnormal") extracted from the public CirCor database [10].

The results of the processing with DWT as preprocessing and with no preprocessing at all are presented in Table II.

It is also possible, using these results, to extract algorithm evaluation metrics such as Sensitivity (Se), Precision (P), and F1-Score ($F1$)

Sensitivity measures the algorithm's ability to correctly detect heart sounds annotated by specialists (minimizing false negatives) and is defined by Equation 3:

TABLE II
HEART SOUNDS DETECTION RESULTS (TP, FN, FP) WITH AND WITHOUT DWT

	TP		FN		FP		Total	
Category	**DWT**	**No DWT**	**DWT**	**No DWT**	**DWT**	**No DWT**	**DWT**	**No DWT**
Clinical condition								
Normal	12564	12137	609	1047	1746	1955	14919	15139
Abnormal	12055	11425	1109	1725	1840	2173	15004	15323
Age group								
Neonate	75	60	1	16	2	15	78	91
Infant	1906	1721	351	532	333	390	2590	2643
Child	18697	17920	1080	1866	2416	2867	22193	22653
Adolescent	2011	1970	157	187	420	432	2588	2589
Not inf.	1930	1891	129	171	415	424	2474	2486

$$Se = \frac{TP}{TP + FN} \tag{3}$$

Precision quantifies the reliability of the detections performed, evaluating the proportion of identified heart sounds that were indeed true (minimizing false positives). Its mathematical formulation is presented in Equation 4:

$$P = \frac{TP}{TP + FP} \tag{4}$$

The F1-Score represents the harmonic mean between sensitivity and precision, providing a single, balanced metric for the global performance of the detection system, as expressed in Equation 5:

$$F1 = 2 \cdot \frac{P \cdot Se}{P + Se} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{5}$$

These were calculated separately for each category, as well as for global results, and tested in a algorithm that had no DWT as pre-processing, and are presented in Table III

TABLE III
HEART SOUNDS DETECTION METRICS WITH AND WITHOUT DWT

	Sens (%)		Prec (%)		F1 (%)	
Category	**DWT**	**No DWT**	**DWT**	**No DWT**	**DWT**	**No DWT**
Clinical condition						
Normal	95.38	92.06	87.80	86.13	91.43	88.99
Abnormal	91.58	86.88	86.76	84.02	89.10	85.43
Age group						
Neonate	98.68	78.95	97.40	80.00	98.04	79.47
Infant	84.45	76.39	85.13	81.53	84.79	78.87
Child	94.54	90.57	88.56	86.21	91.45	88.33
Adolescent	92.76	91.33	82.72	82.01	87.45	86.42
Not inf.	93.73	91.71	82.30	81.68	87.65	86.41
Total	**93.48**	**89.47**	**87.29**	**85.09**	**90.28**	**87.23**

IV. DISCUSSION

The quantitative performance achieved by the proposed algorithm, evidenced by a global F1-Score of 90.28%, a sensitivity of 95.38% for the normal class, and 91.58% for the abnormal class, demonstrates the viability of the approach.

The slight reduction observed in the sensitivity of the abnormal class is an expected behavior from a clinical standpoint. In phonocardiographic records containing murmurs or noise, the amplitude of these acoustic abnormalities can, at specific moments of the cardiac cycle, compete directly with the energy of S1 and S2 heart sounds.

Even so, the algorithm's demonstration of high sensitivity in such a complex pediatric scenario reinforces the quality of the denoising performed by the DWT reconstruction, exclusively using details D_5, D_6 and D_7 of the db8 wavelet

A. Limitations and future works

Despite the good results, the current evaluation methodology and the system architecture present limitations that define clear guidelines for future investigations.

First, the evaluation was conducted using a midpoint event-based metric. While this is a standard approach for initial validation, merely detecting the temporal center of a sound is a relatively simple task for the algorithm and does not guarantee precise morphological boundaries. Second, the reliance on the CirCor database, which is inherently focused on murmur classification rather than temporal heart sound detection, imposes restrictions on validating the exact onset and offset of the S1 and S2 sounds. The improvement of the preprocessing stage, F1 going from 87.23% to 90.28%, could be higher, but the signals from the database had a good Signal-to-Noise-Ratio (SNR).

For future work, we propose expanding the evaluation metrics to strictly measure the temporal boundaries (onset and offset) of the events, rather than just the geometric midpoint. Additionally, future iterations of the algorithm will focus on distinguishing the first heart sound (S1) from the second heart sound (S2). To further optimize the pre-processing stage, rigorous comparative studies with other classes of digital filters will be conducted.

Finally, validating the algorithm demands acquiring or constructing a dedicated database focused exclusively on the precise temporal identification of heart sounds. Overcoming these limitations is essential for future analysis.

Despite the promising results, the current system presents limitations that define guidelines for future investigations.

The primary limitation is that the proposed algorithm can only detect the heart sound position; it is not yet capable of precisely segmenting the start and end of the sounds, nor of distinguishing the first from the second heart sound.

For future work, we propose the development of a more detailed algorithm for the precise segmentation of these heart sounds delimiting their onset and offset and labeling S1 and S2 accordingly, addind to the reliability and checking for the robustness of the proposed algorithm.

REFERENCES

[1] World Health Organization (WHO), "Cardiovascular diseases (CVDs)," WHO Fact Sheets, Jun. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] P. Xiao and K. Wang, "Segmentation of Heart Sound Signals Using Improved Hilbert Transform and Wavelet Packet Transform," *Circuits, Systems, and Signal Processing*, vol. 44, pp. 4752–4773, 2025.

[3] S. Li, F. Li, S. Tang, and W. Xiong, "A Review of Computer-Aided Heart Sound Detection Techniques," *BioMed Research International*, vol. 2020, Article ID 5846191, 2020.

[4] F. Meziani, S. M. Debbal, and A. Atbi, "Analysis of phonocardiogram signals using wavelet transform," *Journal of Medical Engineering & Technology*, vol. 36, no. 6, pp. 283-302, 2012.

[5] M. E. Tavel, *Clinical Phonocardiography and External Pulse Recording*. Chicago: Year Book Medical Publishers, 1967.
[6] A. K. Abbas and R. Bassam, *Phonocardiography Signal Processing*. Morgan & Claypool Publishers, 2009.
[7] Y. Zeng, M. Li, Z. He, and L. Zhou, "Segmentation of Heart Sound Signal Based on Multi-Scale Feature Fusion and Multi-Classification of Congenital Heart Disease," *Bioengineering*, vol. 11, no. 9, p. 876, 2024.
[8] K. A. Babu, B. Ramkumar, and M. S. Manikandan, "Automatic Identification of S1 and S2 Heart Sounds Using Simultaneous PCG and PPG Recordings," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9430-9439, 2018.
[9] N. Giordano and M. Knaflitz, "A Novel Method for Measuring the Timing of Heart Sound Components through Digital Phonocardiography," *Sensors*, vol. 19, no. 8, p. 1868, 2019.
[10] J. Oliveira et al., "The CirCor DigiScope Phonocardiogram Dataset (version 1.0.3)", PhysioNet, 2022. [Online]. Available at: https://doi.org/10.13026/tshs-mw03
[11] The MathWorks Inc. (2025). MATLAB version: 25.2 (R2025b). Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com
[12] A. K. Abbas e R. Bassam, Phonocardiography Signal Processing (Synthesis Lectures on Biomedical Engineering). Cham, Suíça: Springer Nature, 2022, cap. 4, pp. 67–84.
[13] L. Cromwell, F. J. Weibell, e E. A. Pfeiffer, Biomedical Instrumentation and Measurements, 2ª ed. Nova Delhi, Índia: PHI Publication, 2002, cap. 5, pp. 100–104.
[14] M. Saini, "Proposed Algorithm for Implementation of Shannon Energy Envelope for Heart Sound Analysis," *International Journal of Electronics & Communication Technology*, vol. 7, no. 1, 2016.
[15] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," *Proc. Comput. Cardiol.*, pp. 105-108, 1997.

Heart-Evoked Spectral Disturbances in Refractory Epilepsy

Eline Mesquita Melo*, Bryan Lima†, Jéssica Gama*†, Francinaldo Gomes‡, and Antônio Pereira*†
*Graduate Program in Neuroscience and Cell Biology, Federal University of Pará, Belém, Brazil
†Department of Electrical and Biomedical Engineering, Institute of Technology, Federal University of Pará, Belém, Brazil
‡Hospital Ophir Loyola, Belém, Brazil
Corresponding author: Antônio Pereira Jr. (apereira@ufpa.br)

***Abstract*—Heart-evoked spectral disturbances (HESDs) summarize heartbeat-aligned EEG spectral dynamics and provide an exploratory measure of brain–heart interaction. We compared HESDs between nine participants with refractory epilepsy (RE) and seven neurotypical controls (CTRL). Participant-level inference was evaluated across 20 predefined region-by-band comparisons with Benjamini–Hochberg false-discovery-rate correction. Although substantial standardized effect sizes were observed in the RE group, most notably in theta (Hedges' $g = 1.27$–1.68), no region-by-band contrast retained statistical significance at the nominal threshold ($q_{\mathrm{FDR}} \geq 0.116$). Furthermore, none of these salient contrasts exceeded the null distribution obtained by circularly shifting R-peak trains relative to the continuous EEG (all surrogate-adjusted $q = 0.693$), and a dedicated 25–45 Hz sensitivity analysis yielded no FDR-significant regional effect. Quantitative artifact analysis indicated that ICA reduced mean 25–45 Hz ECG–EEG coherence by 57.9%, although residual coupling remained. Conventional time-domain heart-rate variability metrics (RMSSD, SDNN, and pNN50) did not differ significantly between groups. These results characterize peri-R spectral dynamics in refractory epilepsy while establishing an absence of statistical robustness to multiple-comparison correction and a lack of temporal specificity to true cardiac events.**

***Index Terms*—biomedical signal processing, brain–heart interaction, electroencephalography, heart-evoked spectral disturbance, refractory epilepsy**

I. Introduction

Cardiac signals interact continuously with neural activity through autonomic, visceral-afferent, and interoceptive pathways [1], [2]. Cortical responses aligned to individual heartbeats have traditionally been studied using heartbeat-evoked potentials and related time–frequency measures [3], [4]. These signals offer a means of investigating central processing of cardiovascular information, although their low amplitude and temporal proximity to cardiac electrical and mechanical events make artifact control particularly important [4], [5]. Here, we use the term heart-evoked spectral disturbance (HESD) for frequency-specific EEG changes quantified in peri-R segments.

In refractory epilepsy (RE), autonomic dysfunction and altered brain–heart communication have been reported using EEG, heart-rate variability (HRV), and directional coupling measures [6]–[8]. A previous study in refractory epilepsy, for example, identified altered information exchange between cortical and autonomic cardiac dynamics [6], while more recent work in focal epilepsy has demonstrated changes in directional brain–heart information flow surrounding seizures [7]. Nevertheless, the frequency-specific profile and cardiac specificity of peri-R cortical spectral differences remain insufficiently characterized.

The present study compares HESDs between participants with RE and healthy controls. Regional and frequency-specific effects are evaluated using participant-level inference, explicit correction for multiple comparisons, standardized effect sizes and confidence intervals. Because cardiac-field and myogenic artifacts can be particularly problematic in heartbeat-aligned and high-frequency EEG [5], [9], [10], we additionally quantify ECG–EEG contamination, examine the 25–45 Hz range separately, and test temporal specificity using surrogate R-peak trains.

II. Methods

A. Participants and Data Acquisition

EEG and ECG data were obtained from individuals with RE, whose clinical characteristics are summarized in Table I, and neurotypical controls (CTRL). The study was approved by the appropriate institutional research ethics committee (CAAE: 76518417.0.0000.001), and all participants provided written informed consent before data acquisition.

The dataset comprised seven healthy controls (age: 32.0 ± 10.7 years; range: 20–50 years; three female participants) and nine participants with RE (age: 32.8 ± 11.0 years; range: 17–49 years; five female participants). EEG was recorded at 256 Hz, with ECG obtained from a separate lead.

B. R-Peak Detection and HRV Feature Extraction

R-peaks were identified from the CA1+ ECG channel with a minimum interpeak distance of 0.4 s. Candidate event timing was subsequently refined within a ± 10 ms interval using the maximum absolute amplitude of the mean EEG signal. Because this local refinement can increase the temporal alignment of EEG transients, cardiac specificity was not inferred from R-peak alignment alone and was tested independently using circularly shifted event trains.

For the surrogate analysis, each participant's complete R-peak train was displaced relative to the EEG while preserving the number and temporal spacing of events. Standard time-domain HRV measures, including RMSSD, SDNN, and pNN50,

TABLE I
CLINICAL CHARACTERISTICS OF THE PATIENTS INCLUDED IN THE STUDY.

Patient	Age	Sex	Onset of crises (age)	Type of epilepsy / Etiology	Antiseizure medication / dose (mg/day)	Frequency of seizures (per month)	Type of crisis
1	33	M	7	TLE/L	CBZ/1200; PHB/120	12	F
2	45	M	22	TLE/L; Cavernoma	PHB/100; PHT/300	4	F
3	17	F	2	TLE/B; CFD	CBZ/1200; CLM/200; PHB/200	>100	G
4	24	F	12	TLE/L; DNET	PHB/200; PHT/100	5	F
5	45	M	24	TLE/R; Gliosis	CBZ/1200; SV/1500; TPM/200	8	F
6	31	F	14	FLE/B	CLM/20; PHT/300; SV/1500	12	G
7	49	M	45	Unknown	PHT/100	4	G
8	24	F	2	TLE	CBZ/900; PHB/100	12	F
9	27	F	4	TLE/R	SV/1500; CBZ/20	12	F

Notes: M = male; F = female. TLE = temporal lobe epilepsy; CFD = cortical frontal dysplasia; FLE = frontal lobe epilepsy; DNET = dysembryoplastic neuroepithelial tumor. R = right; L = left; B = bilateral. ASM = antiseizure medication. SV = sodium valproate; CBZ = carbamazepine; PHT = phenytoin; PHB = phenobarbital; TPM = topiramate; CLM = clobazam. G = generalized seizure; F = focal seizure.

were computed for each participant according to established HRV definitions [11], [12].

C. EEG Processing and HESD Estimation

Continuous EEG signals were preprocessed using a notch filter to attenuate 60 Hz power-line interference, followed by band-pass filtering from 1 to 40 Hz and average re-referencing. An R-locked cardiac template spanning ± 100 ms around each R-peak was estimated and subtracted prior to HESD estimation. For the bandwise analysis, fourth-order Butterworth filters and Hilbert transforms were applied to delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and nominal gamma (31–45 Hz) bands. Because the continuous signal had undergone 40-Hz low-pass preprocessing, the upper portion of the nominal gamma band was attenuated and high-frequency results were interpreted conservatively.

For each heartbeat, HESD was defined as the difference between the standard deviation of the analytic-signal envelope during the 0 to 600 ms post-R interval and that during the -200 to 0 ms pre-R interval. Heartbeat-level values were averaged within participant and cortical region before group-level inference.

Peri-R time–frequency RE–CTRL difference maps were also calculated for the parietal, temporal, and central regions and are shown in Figs. 2–4. Because inferential testing was performed on participant-level regional summaries rather than on visually selected time–frequency bins, these maps are interpreted descriptively.

D. ECG–EEG Artifact Assessment

Heartbeat-aligned EEG is intrinsically vulnerable to the cardiac field artifact because electrical activity generated by the heart propagates to scalp electrodes [9]. Residual ECG contamination was therefore quantified independently using an ICA-based analysis. Independent component decomposition followed the FastICA framework [13]. ECG–EEG magnitude-squared coherence was calculated before and after ICA in the 25–45 Hz range.

Pre–post changes were summarized at the participant level, and residual post-ICA coherence was also compared between groups. These measurements were treated as quantitative artifact diagnostics rather than as evidence that cardiac or myogenic contamination had been completely removed. This distinction is particularly relevant for high-frequency scalp EEG, in which cranial and ocular muscle activity can substantially contaminate apparent gamma-band activity [10], [14].

E. Statistical Analysis

The participant was the inferential unit. Channels were averaged within participant and cortical region before group comparison. The primary inferential family comprised 20 predefined region-by-band comparisons formed by four regions (frontal, central, parietal, and occipital) and five canonical frequency bands. Benjamini–Hochberg false-discovery-rate correction was applied across all 20 tests [15].

The temporal T7/T8 region was analyzed separately as a five-test secondary family. For regional comparisons, we report raw RE–CTRL mean differences, Hedges' g, participant-level 95% bootstrap confidence intervals based on 10,000 resamples, Welch tests, and exact group-label permutation tests over all 11,440 allocations of 9 of 16 participants to RE.

A separate 25–45 Hz analysis evaluated the robustness of the visually prominent high-frequency regional patterns. Temporal specificity was assessed for the six primary contrasts displaying the largest effect sizes by recomputing the complete HESD extraction after 100 circular shifts of each participant's R-peak train. The same processing and regional summaries were applied to each shifted event train, and the six surrogate tests were corrected together using the Benjamini–Hochberg procedure. HRV measures were compared between groups as exploratory peripheral outcomes.

III. RESULTS

Participant-level analysis suggested observable elevations in HESD across canonical frequency bands in the RE group (Fig. 1-A,B). In the frontal region, the mean RE–CTRL difference in theta was 0.228 μV (95% CI 0.115–0.350 μV; Hedges' $g = 1.51$, 95% CI 1.00–2.64). Substantial point estimates and effect sizes in theta were also present across central ($g = 1.58$), parietal ($g = 1.57$), and occipital ($g = 1.68$) regions, alongside parietal delta ($g = 1.27$) and occipital gamma ($g = 1.38$). Across the complete 20-test primary inferential family, however, no region-by-band contrast retained statistical significance following Benjamini–Hochberg FDR correction at the nominal

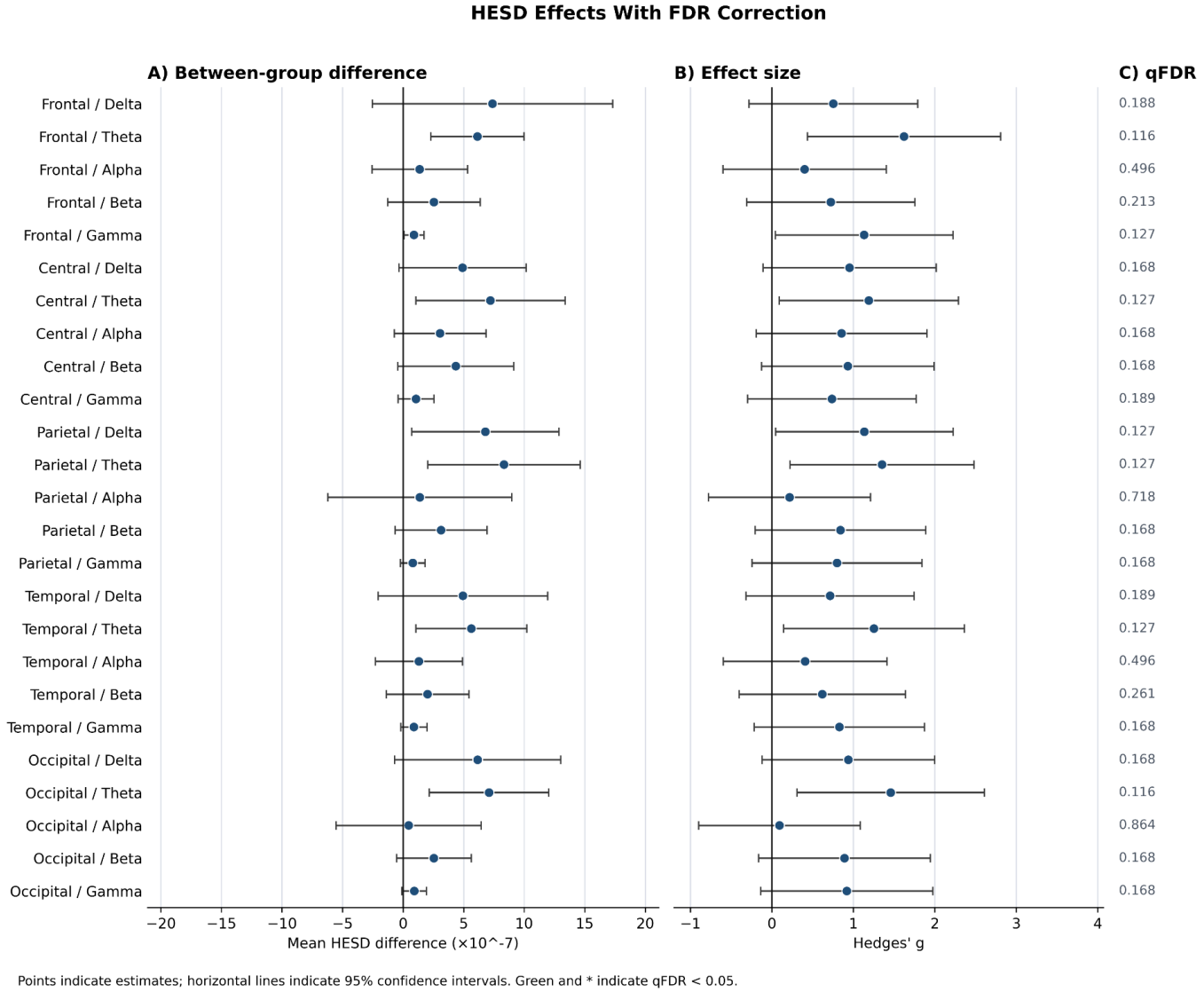


Fig. 1. Participant-level HESD effects with FDR correction. Panel A shows RE–CTRL mean HESD differences with 95% confidence intervals; Panel B shows standardized effect sizes (Hedges' g) with 95% confidence intervals; Panel C reports the corresponding q_{FDR} values.

threshold, yielding minimum values of $q_{\mathrm{FDR}} = 0.116$ for frontal and occipital theta, and $q_{\mathrm{FDR}} \geq 0.127$ across all remaining contrasts (Fig. 1-C). Within the secondary temporal-region family, no frequency band reached statistical significance (minimum $q = 0.127$).

The peri-R time–frequency difference maps showed broad positive RE–CTRL contrasts at approximately 25–45 Hz across parietal, temporal, and central regions (Figs. 2–4). Nevertheless, these descriptive patterns were not corroborated by participant-level high-frequency inference (parietal: $g = 0.80$, $q = 0.178$; temporal: $g = 0.12$, $q = 0.929$; central: $g = 0.19$, $q = 0.793$), and no comparison in the dedicated 25–45 Hz sensitivity analysis survived FDR correction (all $q \geq 0.145$).

Quantitative artifact diagnostics demonstrated that ICA reduced mean 25–45 Hz ECG–EEG coherence from 0.0178 to 0.0075 (a 57.9% reduction; 95% CI 0.0064–0.0146; exact paired $p = 0.00012$), while residual post-ICA coherence did not differ significantly between groups ($q = 0.249$). Notably, in the surrogate control analysis, none of the salient HESD contrasts exceeded the empirical null distribution generated by circular R-peak shifting (all surrogate-adjusted $q = 0.693$), indicating an absence of demonstrable temporal specificity to actual cardiac events under these testing conditions.

Finally, standard peripheral HRV parameters (RMSSD, SDNN, and pNN50) showed no statistically significant differences between RE and CTRL.

IV. Discussion

The participant-level analysis indicated observable exploratory elevations in the derived HESD metric, with the most notable effect sizes concentrated in the theta band across frontal, central, parietal, and occipital regions ($g = 1.27$–1.68) [1], [2].

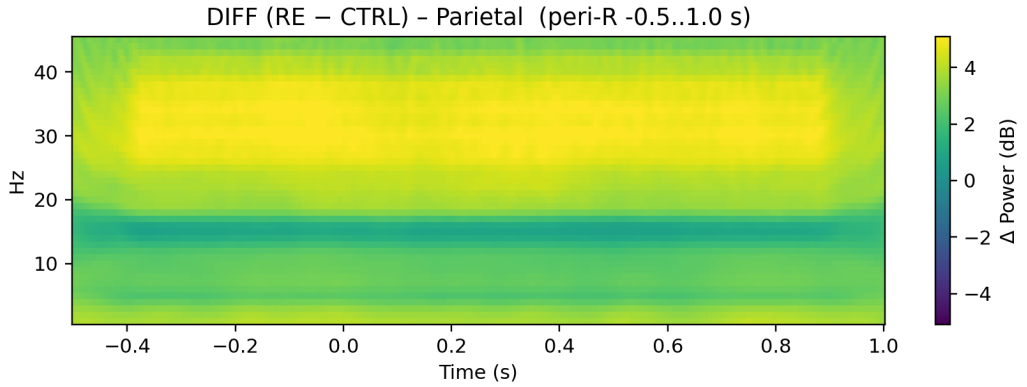


Fig. 2. Descriptive parietal peri-R time–frequency difference map (RE–CTRL). Positive values indicate higher power in RE. The broad high-frequency pattern did not correspond to an FDR-significant participant-level 25–45 Hz effect.

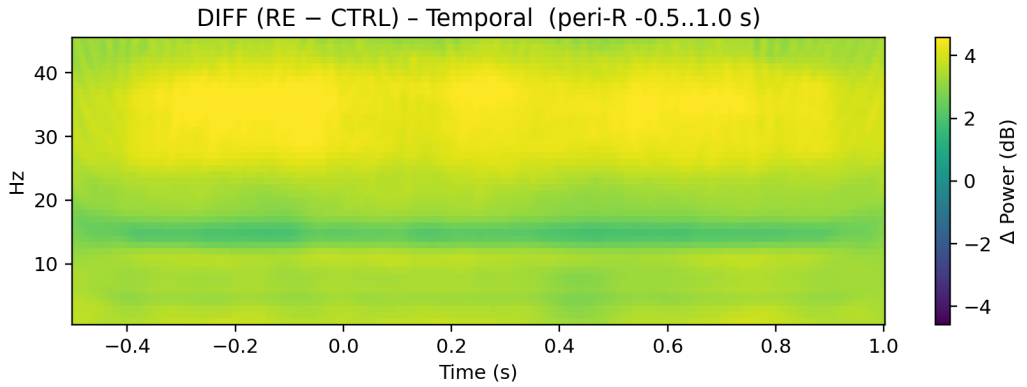


Fig. 3. Descriptive temporal peri-R time–frequency difference map (RE–CTRL). A visually sustained high-frequency group difference is present, but the corresponding participant-level high-frequency comparison did not survive multiple-comparison correction.

However, after strict false discovery rate correction across the 20-test primary inferential family, none of these comparisons maintained statistical significance at the nominal threshold ($q_{\mathrm{FDR}} \geq 0.116$). Similarly, the visually prominent high-frequency contrasts observed in parietal, temporal, and central time–frequency maps were not corroborated by participant-level inference, as no regional comparison survived multiple-comparison adjustment in the dedicated 25–45 Hz sensitivity analysis (all $q \geq 0.145$).

The circular-shift surrogate control analysis provides essential boundaries for interpreting these regional HESD patterns. None of the salient contrasts exceeded the empirical null distribution generated by circularly displacing R-peak event trains relative to the continuous EEG (all surrogate-adjusted $q = 0.693$). Consequently, while the present findings reflect group-level variations in spectral envelope dynamics within peri-R epochs, they do not establish that these perturbations are temporally specific to, or phase-locked with, individual cardiac cycles. In heartbeat-evoked electrophysiology, temporal co-occurrence alone cannot demonstrate a direct neural response to cardiac afference, particularly given the persistence of residual cardiac-field and biomechanical artifacts [3]–[5].

The high-frequency patterns require critical methodological caution. Although independent component analysis substantially attenuated 25–45 Hz ECG–EEG coherence by 57.9% on average, measurable residual coupling persisted across participants, and no regional high-frequency contrast achieved statistical significance. High-frequency scalp EEG is inherently susceptible to cranial and ocular electromyographic contamination as well as residual cardiac-field artifacts [9], [10]. Therefore, the descriptive peri-R time–frequency spectrograms should be considered exploratory representations rather than definitive evidence of cortical gamma-band synchronization or network hyperexcitability.

The low-frequency observations showed greater regional consistency in terms of effect magnitude, with theta-band elevations displaying standardized effect sizes exceeding $g = 1.50$. Cortical slowing and altered frontolimbic oscillatory dynamics are recognized electrophysiological features in focal epilepsy [6], [7]. However, given the lack of temporal specificity confirmed by the surrogate analysis, these observations should be viewed as broad alterations in oscillatory stability during recorded segments rather than genuine heartbeat-evoked interoceptive responses [3], [4].

Standard peripheral HRV metrics (RMSSD, SDNN, and pNN50) did not exhibit statistically significant group differ-

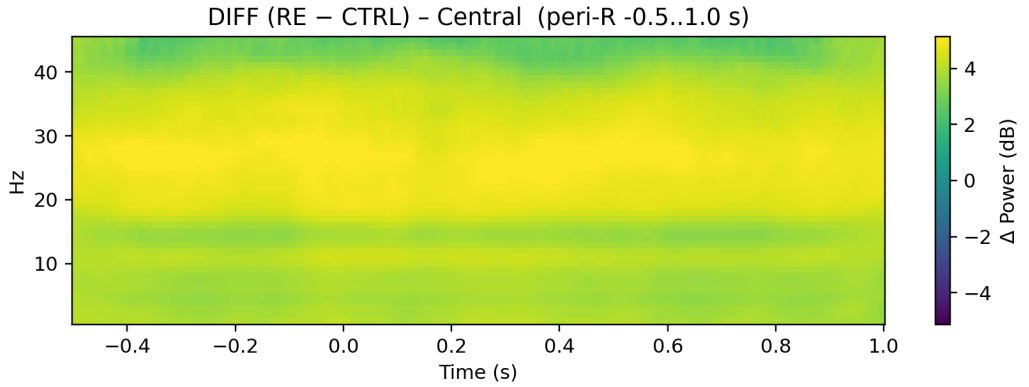


Fig. 4. Descriptive central peri-R time–frequency difference map (RE–CTRL). The visible high-frequency component was not supported by corrected participant-level high-frequency inference.

ences. While previous studies have suggested that central and autonomic interaction metrics may capture subtle regulatory disturbances not readily detected by isolated time-domain HRV parameters [6], [8], the absence of peripheral differences in this cohort must be interpreted with caution. In an exploratory sample, null findings in peripheral autonomic measures may reflect limited statistical power, inter-individual baseline variability, or recording constraints rather than preserved peripheral autonomic integrity [11], [12].

V. Limitations

The conclusions of this study are bounded by some constraints. While non-parametric bootstrap confidence intervals, exact permutation testing, and participant-level inference offer robust protection against inflated false-positive rates, these analytical safeguards cannot compensate for the statistical power constraints imposed by a small sample.

The absence of temporal specificity documented by the circular-shift control analysis constrains physiological interpretation. None of the salient regional HESD contrasts differed significantly from the surrogate null distribution, demonstrating that the observed spectral envelope alterations cannot be definitively resolved as phase-locked or causal cortical responses to discrete cardiac afferent volleys under the present experimental paradigm. Furthermore, the local temporal refinement employed during candidate event detection underscores the necessity of corroborating cardiac specificity through independent surrogate modeling rather than relying on alignment markers alone.

Regarding artifact attenuation, independent component analysis substantially reduced ECG-EEG coherence across the 25–45 Hz band by 57.9%, confirming the efficacy of ICA-based spatial filtering for suppressing volume-conducted cardiac activity in line with standard electrophysiological pipelines. While this filtering strategy effectively isolated neural components from cardiac artifacts, high-frequency scalp recordings remain sensitive to subtle myogenic contributions. Therefore, treating the peri-R time-frequency spectrograms as exploratory representations provides a conservative, methodologically sound framework that complements our participant-level inferential analysis.

Finally, unmodeled clinical heterogeneity presents a potential confounding influence. Key clinical variables documented in Table I, including antiseizure medication regimens, epileptogenic zone localization, duration of epilepsy, seizure burden, and neurobehavioral comorbidities, exert known modulatory effects on central oscillatory dynamics and autonomic tone. Given the cohort size, multivariable adjustment was not statistically tenable, leaving these factors as unmodeled contributors to inter-individual variance that require systematic stratification in larger prospective studies.

VI. Conclusion

In this exploratory investigation, participant-level analysis revealed substantial standardized effect sizes in peri-R spectral dynamics in refractory epilepsy, dominated by elevations in the theta band across frontal, central, parietal, and occipital regions ($g = 1.27$–1.68). However, following Benjamini–Hochberg FDR correction across the complete 20-test primary inferential family, no region-by-band comparison retained statistical significance at the nominal threshold ($q_{\mathrm{FDR}} \geq 0.116$), and conventional time-domain peripheral HRV parameters (RMSSD, SDNN, and pNN50) did not differ significantly between groups. Furthermore, circular-shift surrogate controls demonstrated that the observed spectral envelope perturbations were not temporally specific to true cardiac R-peak timings (all surrogate-adjusted $q = 0.693$).

Although these findings do not establish heartbeat-specific cortical modulation in this cohort, they characterize a distributed peri-R electroencephalographic profile in a well-defined clinical sample. Epilepsy affects approximately 1% of the global population, imposing a considerable burden on patient quality of life, particularly in pharmacoresistant cases where therapeutic options remain constrained. Advancing methodologically rigorous investigations into brain–heart coupling remains imperative to elucidate the neurovisceral mechanisms underlying severe clinical outcomes, such as sudden unexpected death in epilepsy (SUDEP). Ultimately, this work provides

an empirical baseline and underscores the critical need for prospective, large-scale cohorts incorporating surrogate controls and clinical stratification to evaluate the translational utility of central interoceptive metrics.

REFERENCES

[1] E. E. Benarroch, "The central autonomic network: Functional organization, dysfunction, and perspective," *Mayo Clinic Proceedings*, vol. 68, no. 10, pp. 988–1001, 1993.

[2] H. D. Critchley and S. N. Garfinkel, "Interoception and emotion," *Current Opinion in Psychology*, vol. 17, pp. 7–14, 2017.

[3] R. Schandry, B. Sparrer, and R. Weitkunat, "From the heart to the brain: A study of heartbeat contingent scalp potentials," *International Journal of Neuroscience*, vol. 30, no. 4, pp. 261–275, 1986.

[4] H.-D. Park and O. Blanke, "Heartbeat-evoked cortical responses: Underlying mechanisms, functional roles, and methodological considerations," *NeuroImage*, vol. 197, pp. 502–511, 2019.

[5] P. Steinfath, M. Azanova, N. Kapralov, T. Loesche, L. Enk, V. Nikulin, and A. Villringer, "Heartbeat-evoked responses in m/eeg: A systematic review of methods with suggestions for analysis and reporting," *Psychophysiology*, vol. 63, no. 4, p. e70297, 2026.

[6] E. Melo, J. Fiel, R. Milhomens, T. Ribeiro, R. Navegantes, F. Gomes, B. Duarte Gomes, and A. Pereira, "Dynamic coupling between the central and autonomic cardiac nervous systems in patients with refractory epilepsy: A pilot study," *Frontiers in Neurology*, vol. 13, p. 904052, 2022.

[7] L. Frassineti, V. Catrambone, A. Lanatà, and G. Valenza, "Impaired brain-heart axis in focal epilepsy: Alterations in information flow and implications for seizure dynamics," *Network Neuroscience*, vol. 8, no. 2, pp. 541–556, 2024.

[8] G. Costagliola, A. Orsini, M. Coll, R. Brugada, P. Parisi, and P. Striano, "The brain–heart interaction in epilepsy: Implications for diagnosis, therapy, and sudep prevention," *Annals of Clinical and Translational Neurology*, vol. 8, no. 7, pp. 1557–1568, 2021.

[9] G. Dirlich, L. Vogl, M. Plaschke, and F. Strian, "Cardiac field effects on the eeg," *Electroencephalography and Clinical Neurophysiology*, vol. 102, no. 4, pp. 307–315, 1997.

[10] J. F. Hipp and M. Siegel, "Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in eeg," *Frontiers in Human Neuroscience*, vol. 7, p. 338, 2013.

[11] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[12] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability," *Frontiers in Psychology*, vol. 5, p. 1040, 2014.

[13] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[14] S. D. Muthukumaraswamy, "High-frequency brain activity and muscle artifacts in meg/eeg: A review and recommendations," *Frontiers in Human Neuroscience*, vol. 7, p. 138, 2013.

[15] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

Caracterização da saída de sistemas de decodificação da atenção auditiva através de modelos ocultos de Markov

Lucian Soares Schiavon
Programa de Pós-graduação em Engenharia Elétrica
Universidade Federal de Santa Catarina
Florianópolis/SC, Brasil
lucianschiavon@ifsul.edu.br

José Carlos Moreira Bermudez
Programa de Pós-graduação em Engenharia Elétrica
Universidade Federal de Santa Catarina
Florianópolis/SC, Brasil
j.bermudez@ieee.org

Márcio Holsbach Costa
Programa de Pós-graduação em Engenharia Elétrica
Universidade Federal de Santa Catarina
Florianópolis/SC, Brasil
marcio.costa@ufsc.br

***Resumo*—Sistemas de decodificação da atenção auditiva (AAD) buscam identificar o locutor de interesse de um ouvinte na presença de múltiplas fontes sonoras, permitindo o direcionamento espacial de dispositivos de auxílio auditivo. Uma dificuldade existente no projeto e avaliação de sistemas AAD é a limitada disponibilidade de bancos de dados, devido às dificuldades técnicas e ao alto custo de implementação. A construção de um modelo generativo para as saídas de sistemas AAD permite gerar observações sintéticas que preservam as características estatísticas observadas em sujeitos reais, ampliando a disponibilidade de sinais para avaliação experimental. Para isso, este trabalho propõe um modelo estatístico baseado em Modelos Ocultos de Markov (HMM) com distribuições Beta para caracterizar as propriedades estatísticas e temporais dessas saídas. A qualidade dos dados sintéticos gerados pelo modelo é avaliada por meio de testes estatísticos, demonstrando que o modelo preserva as principais características distributivas dos dados originais. A avaliação de algoritmos de pós-processamento utilizando 100 ouvintes sintéticos apresentou menor variabilidade nas métricas de desempenho, como PESQ e STOI, em comparação com a avaliação baseada em apenas 16 participantes reais, indicando que o modelo proposto pode ser uma alternativa para ampliar a validação experimental de sistemas AAD.**

Palavras-chave—Decodificação de atenção auditiva, EEG, modelo oculto de Markov, pós-processamento.

I. Introdução

A capacidade de direcionar a atenção para uma fonte sonora específica em ambientes com múltiplos estímulos acústicos (*cocktail party effect*) é uma das principais características do sistema auditivo humano e desempenha papel fundamental na compreensão da fala em ambientes ruidosos. Entretanto, indivíduos com perda auditiva apresentam dificuldades significativas nessa tarefa, mesmo com o uso de aparelhos auditivos ou implantes cocleares. Isso porque esses dispositivos, embora incorporem algoritmos de amplificação, redução de ruído e separação de fontes, não dispõem de informações sobre qual locutor representa o foco de atenção do usuário [1]. Como consequência, a seleção e o realce do sinal de fala de interesse permanecem como um desafio em cenários com múltiplos falantes.

Nesse contexto, os sistemas de decodificação de atenção auditiva (*Auditory Attention Decoding* - AAD) têm recebido crescente atenção por utilizarem sinais neurais, especialmente o eletroencefalograma (EEG), para estimar o foco atencional do ouvinte e, dessa forma, fornecer informações que permitam o controle de dispositivos auditivos inteligentes. Essa abordagem viabiliza o desenvolvimento de próteses auditivas neurocontroladas (*neuro-steered hearing devices*), capazes de amplificar seletivamente a fala de interesse e atenuar fontes sonoras interferentes [2].

Os métodos de AAD baseados em EEG podem ser classificados em duas abordagens principais: reconstrução do estímulo e classificação direta dos sinais neurais. A primeira baseia-se na reconstrução de características do estímulo auditivo atendido, como o envelope da fala, a partir da atividade neural, seguida da comparação com os estímulos sonoros disponíveis. Embora essa abordagem apresente resultados consistentes, seu desempenho é fortemente dependente do cálculo de correlações com os sinais de áudio de interesse e tende a degradar significativamente em janelas curtas de decisão, limitando sua aplicação em sistemas de processamento em tempo real [2].

Em contrapartida, a classificação direta busca inferir o foco atencional exclusivamente a partir dos sinais neurais, dispensando o conhecimento prévio dos estímulos acústicos durante a inferência. Nos últimos anos, essa abordagem tem recebido crescente atenção devido à incorporação de técnicas de aprendizado profundo, especialmente redes neurais convolucionais (CNNs), recorrentes (RNNs) e, mais recentemente, arquiteturas baseadas em grafos, que exploram de forma mais eficiente as características espaciais e temporais dos sinais de EEG. Em particular, a decodificação da atenção auditiva espacial (*Spatial Auditory Attention Decoding* – SpAAD), que estima a direção da fonte sonora atendida sem utilizar informações do conteúdo da fala, destaca-se como uma alternativa para dispositivos auditivos neurocontrolados, por possibilitar decisões rápidas e reduzir a dependência de informações acústicas externas [3].

Embora os avanços recentes tenham melhorado o desempenho dos algoritmos de AAD, suas predições são normalmente realizadas em janelas temporais curtas de EEG, tornando-as suscetíveis a flutuações e classificações incorretas. Janelas reduzidas são desejáveis por permitirem uma resposta mais rápida às mudanças de atenção do usuário, porém, geralmente, resultam em menor confiabilidade das decisões. Nesse contexto, técnicas de pós-processamento têm sido empregadas para explorar a continuidade temporal do processo atencional, suavizando as predições e reduzindo a influência de erros esporádicos sem comprometer significativamente a capacidade de detectar mudanças reais de atenção. Além de aumentar a robustez das decisões, esses métodos favorecem a implementação de sistemas AAD em tempo real ao proporcionar transições mais estáveis entre estados atencionais e evitar alterações abruptas no ganho dos dispositivos auditivos [4].

Uma das principais limitações da área é a reduzida disponibilidade de bases de dados que reproduzam cenários

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

auditivos realistas com múltiplas fontes sonoras concorrentes e contenham registros neurais adequados para o desenvolvimento e a validação desses sistemas. Além disso, os conjuntos de dados mais amplamente utilizados na literatura contam com um número limitado de participantes, restringindo a diversidade interindividual e a capacidade de generalização dos modelos [5] [6]. Essa limitação afeta não apenas a comparação entre diferentes algoritmos de AAD, mas também a avaliação de técnicas de pós-processamento, cuja eficácia depende da análise de padrões temporais e da variabilidade das decisões produzidas pelos classificadores em um conjunto representativo de usuários.

Neste trabalho, propõe-se um modelo estatístico baseado em Modelos Ocultos de Markov (*Hidden Markov Models* – HMM) para caracterizar temporalmente a saída de sistemas de AAD. O modelo possui dois objetivos principais. O primeiro consiste em possibilitar uma avaliação mais robusta de algoritmos de pós-processamento por meio da geração de ouvintes sintéticos cujas sequências de decisões preservem as características estatísticas observadas em usuários reais, ampliando a variedade de sequências disponível para validação experimental. O segundo objetivo é fornecer uma representação compacta e interpretável do comportamento temporal dos sistemas AAD, permitindo descrever sua dinâmica por meio dos parâmetros estimados do HMM. Para validar a abordagem proposta, foram comparados os desempenhos de um algoritmo AAD com e sem pós-processamento utilizando tanto os participantes reais de uma base de dados quanto os ouvintes sintéticos gerados pelo modelo. A avaliação foi realizada considerando métricas objetivas de qualidade e inteligibilidade da fala, bem como métricas temporais relacionadas à dinâmica das decisões, permitindo analisar o impacto do pós-processamento tanto na qualidade da saída quanto na estabilidade temporal do sistema.

O trabalho foi organizado da seguinte forma: a Seção II apresenta o modelo estatístico que caracteriza a saída de sistemas AAD através de HMM com distribuições Beta. A Seção III detalha a metodologia utilizada, além do banco de dados empregado. A Seção IV apresenta os resultados obtidos e respectiva discussão. As conclusões sobre o trabalho estão na Seção V.

II. Modelo Estatístico da Saída de um Sistema AAD usando HMM com Distribuições Beta

Um modelo oculto de Markov é formalmente definido pelo conjunto de parâmetros $\lambda = \{\pi, A, B\}$, por um conjunto finito de N estados ocultos $S = \{s_1, s_2 \dots s_N\}$ e por uma sequência de T observações $O = \{o_1, o_2 \dots o_T\}$ [7]. A distribuição inicial dos estados é dada por $\pi = \{\pi_1, \pi_2 \dots \pi_N\}$, em que $\pi_i = P(q_1 = s_i)$, sendo q_1 a variável de estado inicial. A matriz de probabilidades de transição entre estados $A = [a_{ij}]$ é definida por

$$a_{ij} = P\left(q_{t+1} = s_j \middle| q_t = s_i\right), \tag{1}$$

descrevendo a probabilidade de transição do estado s_i para o estado s_j no instante t+1, em que q_t representa a variável de estado no instante t. As probabilidades de emissão (ou verossimilhanças das observações) são representadas por B, sendo definidas por

$$b_j(o_t) = P\left(o_t \middle| q_t = s_j\right), \tag{2}$$

que expressa a probabilidade de observar o_t dado que o sistema se encontra no estado s_j.

No contexto da AAD, os algoritmos produzem decisões probabilísticas sequenciais referentes à fonte sonora à qual o indivíduo está direcionando sua atenção em ambientes com múltiplos falantes. Um paradigma experimental amplamente empregado considera dois locutores concorrentes, por exemplo, posicionados à esquerda e à direita do ouvinte. Em cada instante de decisão, o algoritmo estima a probabilidade de que a atenção do usuário esteja direcionada a cada um desses locutores (Fig. 1).

Como os rótulos de referência da direção de atenção são conhecidos nos conjuntos de dados utilizados para treinamento e avaliação, a saída de um sistema AAD pode ser representada por uma matriz $P \in R^{T \times S}$, em que T representa o número de decisões temporais e S o número de participantes. Cada elemento $P_{t,s}$ corresponde à probabilidade de que a decisão no instante t, para o participante s, esteja correta, denotada por p_{correto}.

Nesse contexto, a utilização da probabilidade de classificação correta, p_{correto}, como variável de observação do HMM, mostra-se particularmente adequada, pois fornece uma medida contínua da confiança do decodificador em cada instante de tempo. Diferentemente de decisões binárias, que descartam informações relacionadas ao grau de confiança da classificação, as saídas probabilísticas preservam informações importantes sobre a confiabilidade de cada estimativa produzida pelo sistema AAD, permitindo que o HMM atribua maior ou menor peso às observações de acordo com seu nível de confiabilidade.

Como $p_{\text{correto}} \in [0,1]$, o espaço das observações é naturalmente limitado ao intervalo unitário, o que motiva a utilização da distribuição Beta para modelar as probabilidades de emissão, proporcionando uma forma estatisticamente consistente de incorporar as saídas probabilísticas do sistema AAD ao modelo HMM [8]. Diferentemente da mais empregada distribuição Gaussiana, cujo domínio compreende toda a reta real, a distribuição Beta respeita naturalmente os limites impostos por variáveis probabilísticas, dispensando procedimentos artificiais de truncamento. Essa característica permite interpretar cada estado oculto como um padrão distinto de confiança associado às decisões produzidas pelo sistema AAD.

Considerando essa formulação, a probabilidade de emissão $b_j(o_t)$ associada ao estado s_j é modelada pela função densidade de probabilidade (PDF) de uma distribuição Beta parametrizada pelos parâmetros positivos α_j e β_j. Dessa forma, a verossimilhança de uma observação o_t dado o estado s_j é expressa por

$$b_j(o_t) = f\left(o_t \middle| \alpha_j, \beta_j\right) = \frac{1}{B\left(\alpha_j, \beta_j\right)} o_t^{\alpha_j - 1} (1 - o_t)^{\beta_j - 1}, \tag{3}$$

em que

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{4}$$

representa a função Beta e $\Gamma(\cdot)$ denota a função Gama.

Dada uma sequência de observações O e um conjunto de N estados ocultos, o objetivo consiste em estimar os parâmetros do HMM, incluindo a matriz de probabilidades de

transição A e as probabilidades de emissão B, determinadas pelos parâmetros α_j e β_j das distribuições Beta associadas a cada estado. A estimação desses parâmetros é realizada por meio de um procedimento iterativo de máxima verossimilhança conhecido como algoritmo *forward-backward*, ou algoritmo de Baum-Welch [7]. Esse algoritmo parte de uma estimativa inicial dos parâmetros do modelo e, de forma iterativa, alterna entre a estimação das probabilidades dos estados ocultos e a atualização dos parâmetros do modelo, aumentando progressivamente a verossimilhança da sequência de observações.

Após a estimação dos parâmetros do HMM $\hat{\lambda} = (\hat{\pi}, \hat{A}, \{\hat{\alpha}_j, \hat{\beta}_j\}_{j=1}^{N})$, o modelo pode ser utilizado para gerar sequências sintéticas de probabilidades, provenientes de sistemas AAD, que preservam as características estatísticas e temporais dos dados originais. Em particular, a distribuição inicial dos estados $\hat{\pi}$, a matriz de transição $\hat{A}$ e os parâmetros $(\hat{\alpha}_j, \hat{\beta}_j)$ das distribuições Beta associadas a cada estado definem completamente o processo gerador das observações.

Para construir um conjunto de dados sintético composto por L ouvintes artificiais, são inicialmente simuladas sequências de estados ocultos $\{q_t\}_{t=1}^{T}$. O estado inicial é amostrado de acordo com a distribuição $\hat{\pi}$ e, nos instantes subsequentes, as transições entre estados são geradas segundo as probabilidades definidas em $\hat{A}$. Em cada instante t, dado que o sistema se encontra no estado $q_t = s_j$, uma observação o_t é amostrada da distribuição Beta correspondente. Esse procedimento produz sequências de probabilidades $o_t \in [0,1]$ que reproduzem os níveis de confiança fornecidos pelo sistema AAD original.

Ao repetir esse processo de forma independente para L realizações, obtém-se um conjunto de dados sintético capaz de reproduzir tanto a dinâmica temporal das transições entre estados atencionais quanto as propriedades estatísticas da confiança do decodificador. Esses dados sintéticos são particularmente úteis para a avaliação da robustez de algoritmos, a realização de experimentos controlados e a ampliação de bases de dados experimentais de tamanho reduzido, preservando a estrutura probabilística e temporal aprendida a partir das observações reais produzidas por sistemas AAD.

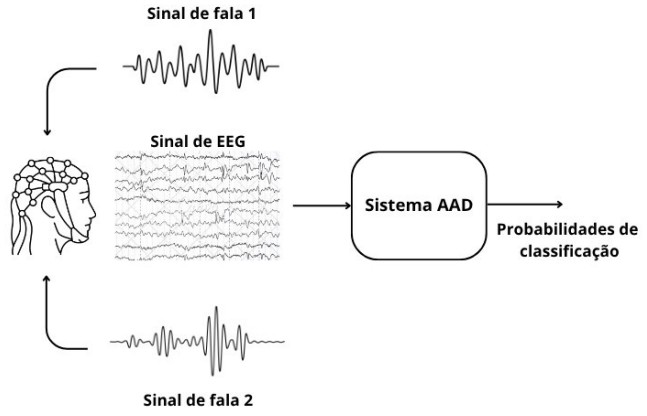


Fig. 1. Diagrama ilustrativo do paradigma experimental utilizado em problema de decodificação de atenção auditiva.

A. Seleção do número ótimo de estados

Para selecionar o número ótimo de estados, os dados de observação foram particionados em conjuntos de treinamento e teste. A seleção do modelo foi realizada utilizando o Critério de Informação Bayesiano (*Bayesian Information Criterion* – BIC) e o Critério Integrado da Verossimilhança Completa (*Integrated Completed Likelihood* – ICL), ambos avaliados sobre o conjunto de treinamento, enquanto a capacidade de generalização foi analisada por meio da log-verossimilhança no conjunto de teste independente [9].

O BIC estabelece um compromisso entre o ajuste do modelo aos dados e sua complexidade, sendo definido por

$$BIC = -2 \cdot \log P(O|\lambda) + p \cdot \log M, \tag{5}$$

em que $P(O|\lambda)$ representa a verossimilhança da sequência de observações do conjunto de treinamento dado o conjunto de parâmetros do HMM $\lambda = \left(\pi, A, \{\alpha_j, \beta_j\}_{j=1}^{N}\right)$, M corresponde ao número total de observações e p representa o número de parâmetros livres do modelo. Para o HMM considerado, com distribuições Beta modelando as probabilidades de emissão, o número de parâmetros é dado por

$$p = (N-1) + N(N-1) + 2N, \tag{6}$$

o qual contempla os parâmetros da distribuição inicial dos estados, da matriz de probabilidades de transição e das distribuições de emissão de cada estado. Valores menores de BIC indicam um melhor equilíbrio entre qualidade de ajuste e complexidade, penalizando modelos excessivamente complexos à medida que o número de estados aumenta.

Com o objetivo de considerar também a qualidade da separação entre os estados ocultos, foi empregado o critério ICL. Esse critério estende o BIC por meio da incorporação de um termo de penalização baseado na entropia da distribuição posterior dos estados:

$$ICL = BIC - 2 \sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_t(k) \log \gamma_t(k), \tag{7}$$

em que $\gamma_t(k) = P(q_t = s_k | O, \lambda)$ representa a probabilidade posterior do sistema estar no estado s_k no instante t. O termo adicional penaliza modelos que apresentam atribuições de estados difusas ou com elevada sobreposição, favorecendo soluções cujos estados sejam bem separados e mais facilmente interpretáveis.

Além desses critérios, a capacidade de generalização de cada modelo candidato foi avaliada por meio do cálculo da log-verossimilhança, $\log P(O|\lambda)$, sobre o conjunto de teste independente. Essa etapa garante que o modelo selecionado não apenas apresente um adequado ajuste aos dados de treinamento, mas também seja capaz de capturar padrões subjacentes que se generalizem para observações não utilizadas durante o processo de estimação.

Por fim, realizou-se uma avaliação qualitativa do modelo aprendido, com o objetivo de verificar sua interpretabilidade e plausibilidade física. Em particular, os parâmetros das distribuições Beta associadas às probabilidades de emissão foram analisados por meio das médias e variâncias por elas induzidas, verificando-se se cada estado oculto representa um regime distinto de confiança do decodificador. Adicionalmente, a estrutura da matriz de probabilidades de transição foi examinada para confirmar a coerência da

dinâmica temporal do modelo, incluindo a presença de estados estáveis, padrões de transição consistentes e a ausência de fragmentação artificial em estados redundantes ou de curta duração.

III. Metodologia

Neste trabalho, é proposto um modelo estatístico baseado em HMM para caracterizar as propriedades estatísticas e temporais da saída de sistemas AAD. Nas subseções seguintes, são apresentados os detalhes do banco de dados e do modelo AAD utilizado, os procedimentos de implementação e treinamento do modelo proposto, bem como as métricas de avaliação e os algoritmos de pós-processamento empregados nos experimentos.

A. Banco de dados e modelo AAD

O modelo de AAD empregado neste trabalho é baseado na abordagem proposta em [3], a qual utiliza uma rede neural convolucional para decodificar a direção espacial da atenção auditiva a partir de sinais de EEG em um cenário com dois locutores. O modelo fornece, para cada janela temporal de decisão, a probabilidade de que o ouvinte esteja direcionando sua atenção ao locutor posicionado à esquerda ou à direita.

Os experimentos foram realizados utilizando uma base de dados pública disponibilizada pela KU Leuven [5]. Os sinais de EEG foram adquiridos de 16 indivíduos com audição normal por meio de um sistema BioSemi ActiveTwo, composto por 64 canais e frequência de amostragem de 8196 Hz. Os estímulos auditivos consistiram em quatro discursos em língua holandesa, narrados por três diferentes locutores do sexo masculino. Cada discurso possui aproximadamente 12 minutos de duração e foi dividido em dois segmentos de aproximadamente 6 minutos. Dessa forma, cada participante (ouvinte) realizou oito ensaios (*trials*) de aproximadamente 6 minutos cada, totalizando 48 minutos de gravação.

Durante cada ensaio, os participantes foram expostos simultaneamente a dois fluxos de fala apresentados com o mesmo nível sonoro, tendo sido instruídos a concentrar sua atenção em apenas um deles, ignorando o outro. Os estímulos foram apresentados de forma dicótica, com um locutor em cada ouvido, ou espacializados por meio de funções de transferência relacionadas à cabeça (*Head-Related Transfer Functions* – HRTFs), simulando fontes sonoras localizadas a ±90° em relação ao ouvinte.

Para evitar vazamento de informação entre os conjuntos de treinamento, validação e teste, neste trabalho tanto o conteúdo dos discursos quanto a identidade dos locutores foram mantidos distintos entre essas partições. Especificamente, a avaliação foi conduzida individualmente para cada participante (*leave-one-subject-out*). As narrativas 1 e 2 foram utilizadas para treinamento e validação, excluindo-se os dados do participante sob avaliação, enquanto as narrativas 3 e 4 desse mesmo participante foram reservadas para teste. Destaca-se que as narrativas 3 e 4 foram narradas pelo mesmo locutor, ao passo que as narrativas 1 e 2 envolvem locutores distintos.

Os sinais de EEG foram segmentados em janelas temporais sobrepostas utilizando uma estratégia de janela deslizante com sobreposição de 50%. Foi considerada a duração da janela de decisão de 1 s.

A saída do modelo AAD consiste em uma sequência de probabilidades que representa, para cada janela de decisão, a probabilidade de que a atenção do ouvinte esteja voltada para o locutor à esquerda ou à direita. Para os 16 participantes, essas probabilidades são organizadas em uma matriz $P \in R^{T\times 16}$, em que T representa o número de janelas temporais e 16 corresponde ao número total de participantes da base de dados. Como o conjunto de teste é composto apenas pelas narrativas 3 e 4, as saídas do sistema AAD correspondem aos ensaios 3, 4, 7 e 8. Essas sequências probabilísticas constituem as observações utilizadas como entrada para o modelo HMM descrito nas seções anteriores. Informações adicionais sobre a arquitetura da CNN, o procedimento de treinamento e a implementação do sistema AAD podem ser encontradas em [10].

B. Inicialização dos parâmetros do HMM

Embora o algoritmo de Baum–Welch possibilite a estimação não supervisionada dos parâmetros de um HMM, seu desempenho depende fortemente da inicialização desses parâmetros [9]. Devido à natureza não convexa da função de verossimilhança, o algoritmo converge garantidamente apenas para um ótimo local. Consequentemente, uma inicialização inadequada pode conduzir a soluções subótimas, convergência lenta ou casos degenerados, como estados com probabilidade de ocupação desprezível ou distribuições de emissão pouco distintas.

Seja o conjunto de parâmetros do HMM denotado por $\lambda = \{\pi, A, B\}$,em que $B = \left\{\alpha_j, \beta_j\right\}_{j=1}^{N}$ representa os parâmetros das distribuições Beta associadas às probabilidades de emissão. A distribuição inicial dos estados, $\pi^{(0)}$, é definida como uniforme, de modo que

$$\pi_j^{(0)} = \frac{1}{N},\ j = 1,\ldots,N, \tag{8}$$

refletindo a ausência de conhecimento prévio que favoreça qualquer estado oculto.

A matriz inicial de probabilidades de transição, $A^{(0)} \in R^{N\times N}$, é construída de forma que cada linha represente uma distribuição de probabilidades válida. Para isso, seus elementos $a_{ij}^{(0)}$ são inicialmente gerados como valores aleatórios positivos e, em seguida, normalizados de forma que

$$\sum_{j=1}^{N} a_{ij}^{(0)} = 1\ ,\ \forall i, \tag{9}$$

garantindo que a soma dos elementos de cada linha seja igual à unidade. Essa estratégia de inicialização confere flexibilidade ao modelo para capturar diferentes dinâmicas temporais, evitando padrões de transição determinísticos ou enviesados.

Para as probabilidades de emissão, cada estado oculto s_j é modelado por uma distribuição Beta parametrizada por (α_j, β_j). A inicialização desses parâmetros é realizada de maneira orientada pelos dados, com o objetivo de promover uma separação significativa entre os estados desde o início do processo de estimação. Especificamente, o conjunto de observações O é particionado em N grupos utilizando N-1 quantis, de forma que cada grupo contenha aproximadamente o mesmo número de amostras. Para cada grupo j, são calculadas a média e a variância amostrais, as quais são

empregadas na estimação inicial dos parâmetros $\alpha_j^{(0)}$ e $\beta_j^{(0)}$ por meio do método dos momentos.

Essa estratégia de inicialização apresenta diversas vantagens. Primeiramente, garante que as distribuições de emissão estejam distribuídas ao longo de todo o intervalo (0,1), reduzindo a sobreposição inicial entre os estados ocultos. Além disso, favorece uma convergência mais rápida do algoritmo de Baum–Welch ao fornecer uma aproximação inicial estruturada da distribuição dos dados. Por fim, reduz a probabilidade de ocorrência de soluções degeneradas, como estados que convergem para distribuições praticamente idênticas ou que passam a representar apenas uma parcela insignificante das observações.

C. Procedimento de treinamento e estabilidade numérica

O treinamento do modelo por meio do algoritmo de Baum–Welch foi realizado com um limite máximo de 50 iterações e um critério de parada antecipada baseado na variação da log-verossimilhança total. Especificamente, o processo iterativo foi encerrado quando o ganho relativo da log-verossimilhança entre duas iterações consecutivas tornou-se inferior a 10^{-4}, assegurando a convergência do algoritmo e evitando iterações desnecessárias.

Cuidados adicionais foram adotados para garantir a estabilidade numérica durante a estimação dos parâmetros de emissão. Como a distribuição Beta é definida no intervalo (0,1), ela não admite observações exatamente iguais a 0 ou 1. Essa limitação decorre da forma de sua função densidade de probabilidade, que pode tornar-se indefinida ou divergir nos extremos do intervalo, dependendo dos valores de seus parâmetros.

Para contornar esse problema, todos os valores das observações foram limitados ao intervalo $[10^{-6}, 1\text{-}10^{-6}]$ por meio de truncamento. De forma análoga, as probabilidades produzidas pelo sistema AAD foram restringidas ao intervalo $[10^{-4}, 1\text{-}10^{-4}]$, evitando valores extremos que poderiam comprometer a estabilidade numérica ou resultar em avaliações degeneradas da função de verossimilhança durante o processo de treinamento.

D. Critérios objetivos de qualidade e inteligibilidade da fala

A qualidade e a inteligibilidade da fala foram avaliadas por meio das métricas de banda larga PESQ (*Perceptual Evaluation of Speech Quality*) [11] e STOI (*Short-Time Objective Intelligibility*) [12], respectivamente.

Devido à longa duração de cada ensaio, os índices PESQ e STOI foram calculados em segmentos de aproximadamente 4 s [13]. Como a saída do sistema AAD consiste em uma estimativa probabilística da direção da atenção auditiva do ouvinte, indicando a probabilidade de atenção direcionada ao locutor à direita ou à esquerda, essas probabilidades foram utilizadas para controlar linearmente o ganho aplicado a cada fluxo de fala. Dessa forma, os valores de PESQ e STOI foram calculados considerando uma combinação ponderada das fontes sonoras, na qual o ganho atribuído a cada locutor é proporcional à probabilidade estimada pelo sistema AAD. Essa abordagem permite avaliar o impacto da incerteza das decisões do AAD na qualidade e inteligibilidade da fala, em vez de considerar apenas decisões binárias de seleção do locutor atendido.

E. Algoritmos de pós-processamento

Foram avaliados dois algoritmos de pós-processamento: o método de controle de ganho (*Gain Control* - GC) proposto em [4] e o rastreador de estados (*State Tracker* - ST) proposto em [14]. O objetivo dessa análise é comparar o desempenho desses métodos em relação à saída direta do sistema AAD e, principalmente, investigar a influência da variabilidade proporcionada pelos ouvintes sintéticos gerados pelo modelo proposto. Para isso, os algoritmos de pós-processamento foram avaliados considerando tanto as saídas originais do AAD obtidas a partir dos participantes reais da base de dados quanto as sequências probabilísticas geradas pelo modelo estatístico proposto.

Além das métricas objetivas de qualidade e inteligibilidade da fala, PESQ e STOI, o desempenho dos métodos também foi analisado por meio de métricas associadas à dinâmica temporal das decisões do classificador, incluindo acurácia, *switch rate*, *false-switch burst* e *dwell time* [10]. Essas métricas permitem avaliar não apenas o desempenho global do sistema, mas também sua estabilidade e capacidade de acompanhar corretamente as mudanças no estado de atenção do usuário. O *switch rate* quantifica a frequência de mudanças incorretas de decisão ao longo do tempo. O *false-switch burst* caracteriza a duração de períodos consecutivos em que o algoritmo permanece em uma decisão incorreta após uma troca indevida, enquanto o *dwell time* representa o tempo durante o qual o sistema permanece corretamente no estado correspondente à atenção verdadeira antes de ocorrer uma transição incorreta.

IV. Resultados e Discussão

Esta seção apresenta os resultados obtidos a partir dos métodos descritos na seção anterior, além da discussão sobre os mesmos.

A. Determinação do número ótimo de estados

O número ótimo de estados ocultos foi determinado por meio de uma análise conjunta do BIC, do ICL e da log-verossimilhança obtida para dados não utilizados no treinamento. O conjunto de observações foi dividido em um conjunto de treinamento e um conjunto de teste, sendo que o conjunto de treinamento compreendeu os ensaios 3, 4 e 7, enquanto o conjunto de teste correspondeu ao ensaio 8. Os critérios BIC e ICL foram calculados utilizando os dados de treinamento, enquanto a capacidade de generalização de cada modelo foi avaliada por meio da log-verossimilhança obtida no conjunto de teste independente.

Foram avaliados diferentes modelos candidatos com número crescente de estados ocultos N, iniciando em $N = 2$. A análise foi conduzida até que não fossem observadas melhorias significativas, caracterizadas por um comportamento de platô ou aumento nos valores de BIC e/ou ICL, juntamente com a saturação da log-verossimilhança calculada no conjunto de teste.

Conforme ilustrado na Fig. 2, observa-se que o ICL apresenta um aumento gradual a partir de $N = 7$, indicando uma redução na separação entre os estados para modelos mais complexos. Simultaneamente, a log-verossimilhança avaliada no conjunto de teste apresenta um comportamento de saturação, indicando que a inclusão de estados adicionais não proporciona melhorias significativas na capacidade de generalização do modelo. Embora o BIC continue apresentando redução para valores maiores de N, a taxa de

diminuição torna-se marginal, indicando um ganho limitado no ajuste do modelo em relação ao aumento da complexidade.

Com base na análise conjunta desses três critérios, o número ótimo de estados foi definido como $N = 8$, proporcionando um equilíbrio adequado entre complexidade do modelo, capacidade de generalização e interpretabilidade, evitando a criação de estados redundantes e o *overfitting*.

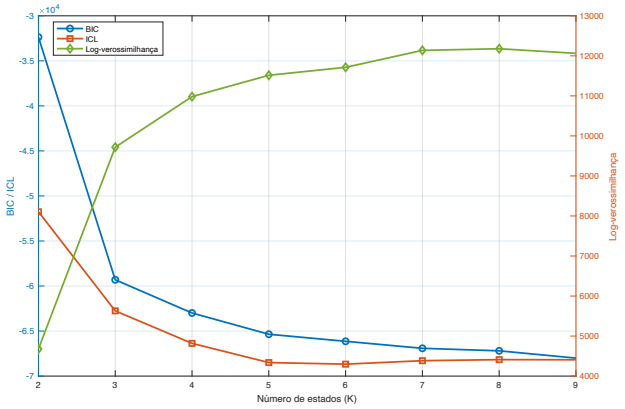


Fig. 2. Determinação do número ótimo de estados ocultos utilizando BIC, ICL e log-verossimilhança.

B. *Validação estatística do modelo*

Foram realizados dois testes estatísticos para avaliar a similaridade entre as distribuições das saídas do sistema AAD obtidas para os 16 participantes da base de dados e para os 100 ouvintes sintéticos gerados pelo modelo proposto: o teste de Kolmogorov–Smirnov para duas amostras (KS) [15] e a distância de Wasserstein de primeira ordem [16].

Em ambos os testes estatísticos, os dados sintéticos foram gerados utilizando o procedimento *leave-one-subject-out*. Para cada um dos 16 ouvintes reais, o modelo HMM foi treinado utilizando os dados dos outros 15 ouvintes. A partir dos parâmetros estimados pelo modelo, foram então gerados os dados sintéticos, que foram comparados, por meio dos testes estatísticos, com os dados do ouvinte real excluído do processo de treinamento. Esse procedimento foi repetido individualmente para cada um dos 16 ouvintes, garantindo que, em cada avaliação, o ouvinte utilizado como referência permanecesse fora do conjunto de dados empregado no treinamento do modelo.

O teste KS quantifica a máxima diferença entre as funções de distribuição acumulada empíricas de duas amostras, enquanto a distância de Wasserstein fornece uma medida da discrepância global entre suas distribuições. Em ambas as análises, foi empregado um procedimento de *bootstrap*, no qual cada participante real foi comparado, em 100 iterações, com ouvintes sintéticos selecionados aleatoriamente.

Os resultados do teste KS, apresentados na Fig. 3, apresentaram estatísticas predominantemente entre 0,13 e 0,27. Como valores menores da estatística KS indicam menores diferenças entre as distribuições acumuladas [15], esses resultados sugerem que as sequências sintéticas apresentam comportamento estatístico semelhante ao observado nos participantes reais, embora algumas diferenças específicas entre indivíduos sejam identificadas.

A distância de Wasserstein de primeira ordem, apresentada na Fig. 4, apresentou valores predominantemente inferiores a 0,15, com média aproximada de 0,11. Considerando que valores menores dessa métrica representam maior proximidade entre as distribuições [16], os resultados indicam que o modelo preserva as características globais das distribuições originais.

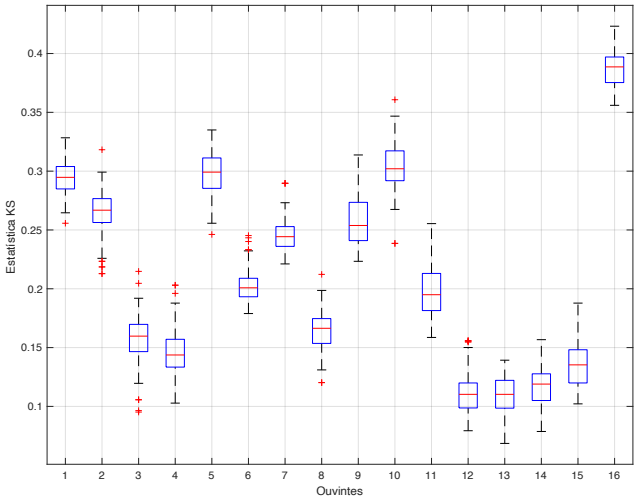


Fig. 3. Distribuição da estatística KS para cada participante real, obtida por *bootstrap* ao comparar as saídas do AAD dos 16 sujeitos da base de dados com os 100 ouvintes sintéticos.

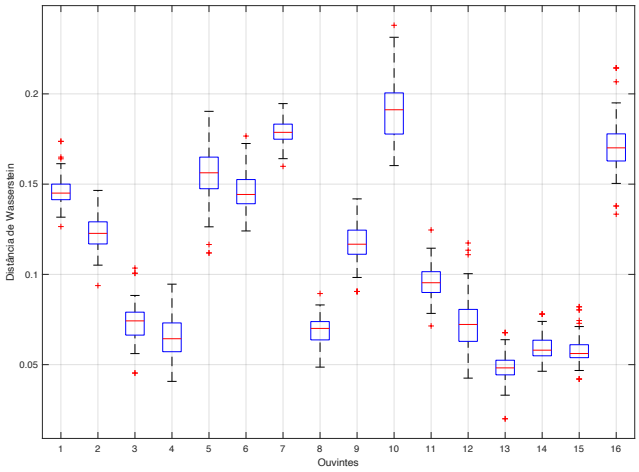


Fig. 4. Distribuição da distância de Wasserstein para cada participante real, obtida por *bootstrap* ao comparar as saídas do AAD dos 16 sujeitos da base de dados com os 100 ouvintes sintéticos.

C. *Comparação das distribuições de PESQ e STOI entre ouvintes reais e sintéticos*

Foram comparados os valores de PESQ e STOI obtidos a partir das saídas do sistema AAD para os 16 ouvintes reais da base de dados e para 100 ouvintes sintéticos gerados pelo modelo proposto, conforme descrito na Seção III-D. As Fig. 5 e Fig. 6 apresentam a comparação global das distribuições dos valores de PESQ e STOI, respectivamente, considerando ambos os grupos de ouvintes. A Fig. 7 apresenta a evolução temporal dessas métricas, mostrando a média e o desvio padrão calculados para cada segmento, permitindo comparar a variabilidade dos resultados entre os ouvintes reais e sintéticos ao longo dos ensaios.

Observa-se que, de forma geral, as distribuições das métricas de qualidade e inteligibilidade da fala são preservadas pelo modelo proposto, indicando que as características estatísticas dos dados reais são adequadamente

reproduzidas pelos ouvintes sintéticos. A análise da evolução temporal das métricas também demonstra comportamento semelhante entre os dois grupos, com valores médios de PESQ e STOI acompanhando a mesma tendência ao longo dos ensaios.

Além disso, nota-se que os ouvintes sintéticos apresentam menor dispersão ao longo dos ensaios, resultando em valores de PESQ e STOI mais estáveis quando comparados aos participantes reais. Por outro lado, os dados reais apresentam maiores oscilações, principalmente nos dois últimos ensaios de teste analisados. Essas variações podem estar associadas a características individuais dos participantes, como diferenças no comportamento de atenção ou na dificuldade de manutenção do foco atencional durante os experimentos.

Apesar dessas diferenças pontuais, os resultados indicam que o modelo proposto é capaz de reproduzir não apenas as distribuições globais de desempenho, mas também o comportamento temporal médio das métricas de qualidade e inteligibilidade obtidas a partir das saídas do sistema AAD.

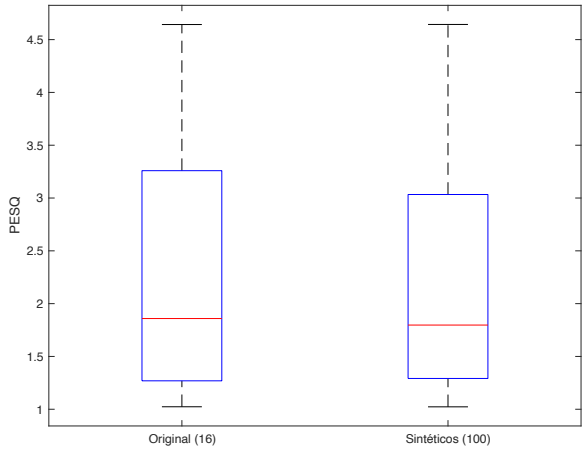


Fig. 5. Diagramas de caixas dos valores de PESQ obtidos para os 16 ouvintes reais e para os 100 ouvintes sintéticos.

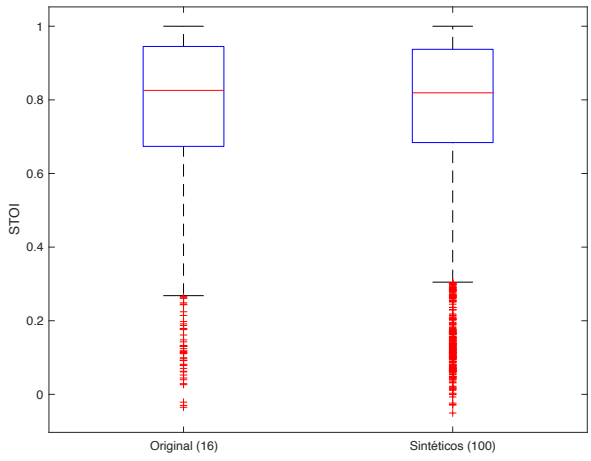


Fig. 6. Diagramas de caixas dos valores de STOI obtidos para os 16 ouvintes reais e para os 100 ouvintes sintéticos.

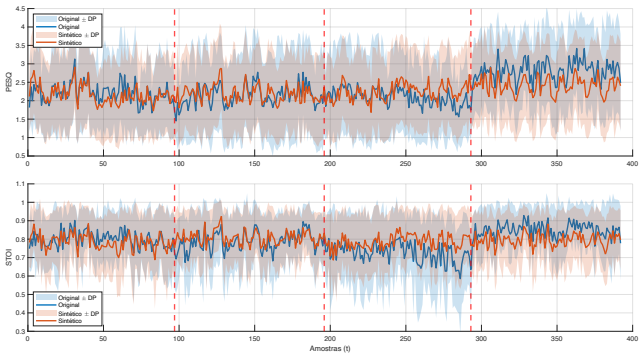


Fig. 7. Evolução temporal dos valores médios de PESQ e STOI para os ouvintes reais e sintéticos. As regiões sombreadas indicam um desvio padrão em torno da média. As linhas tracejadas vermelhas delimitam os quatro ensaios de teste.

D. Avaliação de algoritmos de pós-processamento para AAD utilizando ouvintes reais e sintéticos

Como descrito na Seção III-E, foram avaliados dois algoritmos de pós-processamento propostos na literatura em relação às métricas objetivas e temporais, utilizando como entrada as probabilidades de saída do sistema AAD obtidas tanto para os 16 ouvintes reais da base de dados quanto para os 100 ouvintes sintéticos gerados pelo modelo proposto. Ambos os algoritmos possuem hiperparâmetros responsáveis por controlar a estabilidade temporal das decisões e a velocidade das transições entre estados de atenção. Para cada método, foi realizada uma busca em grade considerando cinco valores para os respectivos hiperparâmetros, sendo apresentados os resultados correspondentes às configurações com melhor desempenho.

Os resultados apresentados nas Tabelas I e II mostram que o algoritmo GC proporciona melhorias consistentes em relação à saída direta do AAD sem pós-processamento, principalmente nas métricas PESQ, STOI, acurácia, *switch rate* e *dwell time*. A redução observada no *false-switch burst* está relacionada ao comportamento esperado dos algoritmos de suavização, uma vez que a maior estabilidade temporal das decisões pode prolongar a permanência em uma classificação incorreta após uma troca equivocada. O algoritmo ST, por sua vez, apresenta ganhos mais limitados, com principalmente nas métricas PESQ, *switch rate* e *dwell time*.

A principal contribuição do modelo proposto é evidenciada pela comparação entre os resultados obtidos com os ouvintes reais e sintéticos. As Tabelas I e II mostram que os valores médios das métricas obtidos com os 100 ouvintes sintéticos permanecem dentro da faixa de variação observada para os 16 ouvintes reais, indicando que o modelo preserva o comportamento estatístico das avaliações de desempenho do sistema AAD e dos algoritmos de pós-processamento.

Além disso, os resultados obtidos com os ouvintes sintéticos apresentam uma redução na variabilidade das métricas, com desvios padrão menores em comparação aos observados para os ouvintes reais. Essa característica representa uma vantagem para a avaliação de algoritmos de AAD e técnicas de pós-processamento, pois permite análises com maior consistência estatística e menor influência de variações individuais presentes em bases de dados com número reduzido de participantes. Dessa forma, o modelo proposto possibilita ampliar a avaliação experimental desses sistemas, mantendo as características estatísticas dos dados originais.

TABELA I. COMPARAÇÃO DAS MÉTRICAS OBJETIVAS E TEMPORAIS OBTIDAS PARA A SAÍDA DIRETA DO AAD E PARA OS ALGORITMOS DE PÓS-PROCESSAMENTO, CONSIDERANDO OS 16 OUVINTES REAIS DA BASE DE DADOS E DIFERENTES CONFIGURAÇÕES DE HIPERPARÂMETROS. *SR*: SWITCH RATE; *FSB*: FALSE-SWITCH BURST; *DT*: DWELL TIME; *GC*: GAIN CONTROL; *ST*: STATE TRACKER.

	Saídas referentes aos 16 ouvintes do banco de dados			
	AAD	***GC (N=3)***	***GC (N=7)***	***ST (K=0,1)***
PESQ	2,32 ± 0,60	3,37 ± 0,59	3,47 ± 0,59	2,58 ± 0,59
STOI	0,79 ± 0,08	0,82 ± 0,11	0,83 ± 0,11	0,79 ± 0,09
Acurácia	0,71 ± 0,12	0,75 ± 0,15	0,76 ± 0,15	0,71 ± 0,12
SR (/min)	22,45 ± 6,04	2,05 ± 0,92	0,89 ± 0,56	17,89 ± 4,99
FSB (s)	1,55 ± 0,51	14,08 ± 8,15	33,19 ± 25,37	1,94 ± 0,71
DT (s)	4,22 ± 2,01	54,03 ± 39,56	119,73 ± 85,57	5,28 ± 2,43

TABELA II. COMPARAÇÃO DAS MÉTRICAS OBJETIVAS E TEMPORAIS OBTIDAS PARA A SAÍDA DIRETA DO AAD E PARA OS ALGORITMOS DE PÓS-PROCESSAMENTO, CONSIDERANDO OS 100 OUVINTES SINTÉTICOS GERADOS PELO MODELO PROPOSTO E DIFERENTES CONFIGURAÇÕES DE HIPERPARÂMETROS. *SR*: SWITCH RATE; *FSB*: FALSE-SWITCH BURST; *DT*: DWELL TIME; *GC*: GAIN CONTROL; *ST*: STATE TRACKER.

	Saídas referentes aos 100 ouvintes sintéticos			
	AAD	***GC (N=3)***	***GC (N=7)***	***ST (K=0,1)***
PESQ	2,25 ± 0,10	3,29 ± 0,12	3,58 ± 0,13	2,47 ± 0,10
STOI	0,79 ± 0,01	0,87 ± 0,02	0,92 ± 0,02	0,79 ± 0,01
Acurácia	0,71 ± 0,02	0,82 ± 0,03	0,91 ± 0,03	0,72 ± 0,02
SR (/min)	23,61 ± 1,22	4,16 ± 0,54	1,60 ± 0,37	18,19 ± 0,97
FSB (s)	1,44 ± 0,07	4,98 ± 0,66	6,49 ± 1,60	1,84 ± 0,10
DT (s)	3,60 ± 0,26	23,17 ± 3,55	64,33 ± 15,15	4,69 ± 0,35

V. CONCLUSÃO

Este trabalho apresentou um modelo estatístico baseado em HMM com distribuições Beta para caracterizar as propriedades estatísticas e temporais das saídas probabilísticas de sistemas de AAD. Ao modelar simultaneamente a evolução temporal dos estados de atenção e a incerteza associada às decisões do AAD, o modelo proposto possibilitou a geração de sinais de observações sintéticas que preservam as principais características observadas nos sujeitos reais.

A similaridade entre as saídas do AAD obtidas para sujeitos reais e sintéticos foi avaliada por meio do teste de Kolmogorov–Smirnov e da distância de Wasserstein, demonstrando que o modelo é capaz de reproduzir a estrutura estatística dos dados originais. Além disso, a avaliação de algoritmos de pós-processamento utilizando ouvintes sintéticos mostrou que os dados gerados mantêm o comportamento observado nos participantes reais, ao mesmo tempo em que reduzem a variabilidade das métricas de desempenho, incluindo PESQ e STOI. No entanto, é importante deixar claro que os dados sintéticos complementam a avaliação dos algoritmos, não substituindo a validação com dados de participantes humanos.

Trabalhos futuros deverão investigar o uso de bases de dados incluindo participantes com deficiência auditiva, bem como avaliar a generalização dos dados sintéticos em dados de saída de diferentes arquiteturas de AAD. Além disso, será importante comparar a abordagem proposta com modelos generativos como *autoencoders* e modelos difusos.

O modelo proposto apresenta-se como uma alternativa para ampliar a avaliação experimental de sistemas AAD em bases de dados com número limitado de participantes. A geração de sinais estatisticamente consistentes de ouvintes sintéticos permite análises mais robustas de algoritmos de pós-processamento e da dinâmica temporal das decisões do AAD.

REFERÊNCIAS

[1] J. O'Sullivan *et al*., "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015, doi: 10.1093/cercor/bht355.

[2] S. Geirnaert *et al*., "EEG-Based Auditory Attention Decoding: Toward Neuro-Steered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, Jul. 2021, doi: 10.1109/MSP.2021.3075931.

[3] S. Vandecappelle, S. Geirnaert, T. Francart, e A. Bertrand, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Neuroscience*, vol. 466, pp. 192–204, Ago. 2021, doi: 10.1016/j.neuroscience.2021.05.013.

[4] N. Heintz, S. Geirnaert, I. Van de Ryck, T. Francart, e A. Bertrand, "Probabilistic Gain Control in a Multi-Speaker Setting Using EEG-Based Auditory Attention Decoding," in *Proc. 32nd European Signal Processing Conference (EUSIPCO)*, Lyon, France, Ago. 2024, pp. 892–896, doi: 10.23919/EUSIPCO63174.2024.10715166.

[5] W. Biesmans, N. Das, T. Francart, e A. Bertrand, "Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, Mai 2017, doi: 10.1109/TNSRE.2016.2571900.

[6] S. A. Fuglsang, D. D. E. Wong, J. Hjortkjær, E. Ceolini, e A. de Cheveigné *et al*., "Dataset of EEG and Audio From a Cocktail Party Paradigm With Competing Speakers," *Zenodo*, Abr. 2018, doi: 10.5281/zenodo.1199011.

[7] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[8] S. L. P. Ferrari e F. Cribari-Neto, "Beta regression for modelling rates and proportions," Journal of Applied Statistics, vol. 31, no. 7, pp. 799–815, 2004, doi: 10.1080/0266476042000214501.

[9] W. Zucchini, I. L. MacDonald, e R. Langrock, Hidden Markov Models for Time Series: An Introduction Using R, 2nd ed. Boca Raton, FL, USA: CRC Press, 2016.

[10] L. S. Schiavon, J. C. M. Bermudez, e M. H. Costa, "Impact of auditory attention decoding accuracy on noise reduction systems for hearing aids," *Biomedical Signal Processing and Control*, vol. 113, part D, art. 109162, 2026, doi: 10.1016/j.bspc.2025.109162.

[11] ITU-T, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," ITU-T Recommendation P.862, Fev. 2001.

[12] C. H. Taal, R. C. Hendriks, R. Heusdens, e J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Set. 2011, doi: 10.1109/TASL.2011.2114881.

[13] ITU-T, "Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P.862.2," ITU-T Recommendation P.862.3, Nov. 2007.

[14] J. Hjortkjær *et al*., "Real-time control of a hearing instrument with EEG-based attention decoding," *Journal of Neural Engineering*, vol. 22, no. 1, Art. no. 016027, 2025, doi: 10.1088/1741-2552/ad867c.

[15] F. J. Massey Jr., "The Kolmogorov-Smirnov Test for Goodness of Fit," Journal of the American Statistical Association, vol. 46, no. 253, pp. 68–78, Mar. 1951, doi: 10.1080/01621459.1951.10500769.

[16] M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in Advances in Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 2013, pp. 2292–23.

Towards Horse Height Estimation from a Single Image Using YOLO Pose and Linear Calibration

Julyana Vieira Cordeiro
Faculdade do Centro Leste
Serra, Brazil
julyanacordeiro@ucl.br

Julia Clabunde Pereira
Faculdade do Centro Leste
Serra, Brazil
juliapereira@ucl.br

Leticia Araujo Silva
Faculdade do Centro Leste
Serra, Brazil
leticia.araujo@ucl.br

***Abstract*—Horse height measurement is important for the contexts of morphometric evaluation, management, selection, and competition. Although conventional measurement with a measuring stick, hippometer, or measuring tape is well established, it depends on direct physical measurement and correct positioning of both the animal and the measuring instrument. This study proposes an image-based methodology for estimating horse height from a single lateral photograph using anatomical landmark detection and pixel-to-centimeter calibration. A You Only Look Once (YOLO) pose model based on the yolo11s-pose architecture was trained to detect two landmarks: the withers and the contact point of the hoof with the ground. The vertical distance between these keypoints was used to estimate height in pixels and then converted into centimeters using simple linear regression. The pose model was trained and validated using 116 lateral horse images, while calibration used 15 images acquired under a standardized protocol with reference heights measured by a leveling hippometer. The evaluation was organized into three stages: YOLO pose validation using 17 images, pixel-to-centimeter calibration using 15 images, and a preliminary complete-pipeline demonstration using two independent images. In validation, the model detected all horses and showed strong agreement between the predicted and reference pixel heights, with an R^2 of 0.9905 and an MAE of 19.61 px. Pixel-to-centimeter calibration achieved an MAE of 2.56 cm and RMSE of 2.96 cm using leave-one-out cross-validation based on reference keypoint annotations, and therefore this error reflects the calibration stage rather than the complete automatic pipeline. In the preliminary complete-pipeline demonstration, height was estimated for two independent images, resulting in a mean absolute error of 2.86 cm and a mean percentage error of 1.74%. The results indicate that the proposed method is a promising single-image alternative for horse height estimation, without the need for depth sensors, multiple viewpoints, or specialized acquisition devices.**

***Index Terms*—Computer vision, equine morphometry, horse height estimation, image-based measurement, pose estimation, YOLO.**

I. Introduction

Measuring the height of a horse may appear to be a simple procedure, but it remains a fundamental part of equine evaluation. Height at the withers is routinely used in morphometric assessment, breed characterization, management, selection, and competition contexts. It also contributes to the description of body proportions and general conformation, supporting decisions related to animal classification, functional aptitude, and breeding programs [1]. For this reason, obtaining reliable height measurements is still an important task in both scientific and practical equine evaluation.

The conventional procedure is based on direct manual measurement using a measuring stick, hippometer, or measuring tape. In formal contexts, the measurement is taken at the highest point of the withers, which is the anatomical reference adopted in official veterinary regulations [2]. Although manual measurement is well established, it requires the evaluator to remain close to the animal and depends on the correct positioning of both the horse and the measuring instrument. In practice, these requirements can make the procedure slower or more difficult when animals are reactive, unfamiliar with the equipment, or evaluated outside controlled facilities.

These practical limitations have motivated the development of image-based and non-contact approaches for animal morphometry. In horses, digital image analysis has been investigated as a tool for estimating body measurements and evaluating conformation from standardized photographs [3]–[5]. Other studies have explored three-dimensional sensing, Light Detection and Ranging (LiDAR), and machine vision systems to obtain equine body measurements with reduced physical interaction between the evaluator and the animal [6], [7]. Together, these works indicate a clear movement toward more objective, repeatable, and less invasive methods for morphometric assessment.

Recent advances in computer vision have further expanded this possibility. Deep learning models can detect animals, localize anatomical landmarks, and support automatic measurement pipelines. In equine applications, target detection and keypoint-based approaches have been proposed for automated body measurement, while horse pose estimation has also emerged as a relevant task for anatomical landmark localization [8], [9]. However, many available approaches are designed for multiple body traits or rely on more complex acquisition setups, such as depth sensors, three-dimensional point clouds, multiple landmarks, or specialized imaging devices.

In this context, a simplified image-based approach focused specifically on height estimation may provide a practical alternative. Since horse height is represented by the vertical relationship between the withers and the ground, this measurement can be adapted to a keypoint-based image analysis problem. By detecting the upper and lower anatomical references in a standardized lateral photograph and applying a calibration strategy, the pixel measurement can be converted into a real-world height estimate. Therefore, this study proposes and eval-

uates an image-based methodology for estimating horse height from a single lateral photograph. The proposed approach uses a You Only Look Once (YOLO) pose model to detect two anatomical landmarks, the withers and the hoof contact point with the ground, calculates the corresponding height in pixels, and applies a linear regression model to convert this measurement into centimeters. By combining anatomical landmark detection, standardized image acquisition, and pixel-to-centimeter calibration, this work aims to provide a practical alternative for estimating horse height from a single lateral image.

II. Materials and Methods

A. Proposed System

This study proposes a methodology for estimating horse height from a single lateral image. Conceptually, the system was structured into two main stages: the automatic detection of anatomical landmarks using a YOLO pose model, and the conversion of the image-based measurement into centimeters. The complete operational pipeline is illustrated in Fig. 1, comprising image acquisition, YOLO-based landmark detection, pixel-height estimation, and regression-based conversion to centimeters.

Initially, the model identifies two points of interest on the animal's body: the withers and the contact point between the most visible front hoof and the ground. The difference between the vertical coordinates of these points provides the horse height in pixels. This measurement is then converted into centimeters using a linear regression model calibrated with images acquired under a standardized protocol and real heights obtained using a leveling hippometer. The following subsections describe each stage in more detail.

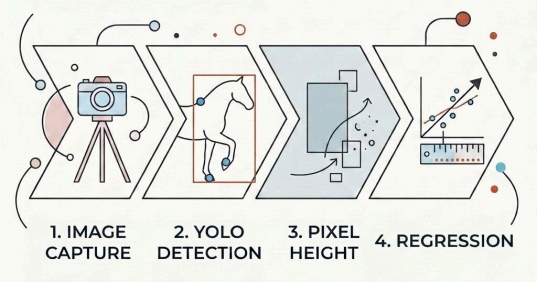


Fig. 1. Overview of the proposed pipeline for horse height estimation, including image acquisition, YOLO-based landmark detection, pixel-height estimation, and regression-based conversion to centimeters.

B. Dataset Description

The images used in this study were organized according to their role in the proposed pipeline. One subset was used for training and validating the YOLO pose model, a second subset was used for calibrating the relationship between pixel-based measurements and real height in centimeters, and two additional images were reserved for a preliminary demonstration of the complete system.

The first subset consisted of 116 lateral-view horse images used for anatomical landmark detection. These images were obtained from a personal catalog and from publicly available online sources licensed under Creative Commons Attribution 4.0 International (CC BY 4.0), selected through a directed search according to visual criteria compatible with the objective of this study. Only images containing a single horse were included, with the animal positioned laterally, visually aligned, and with visible withers and identifiable hooves. This subset was used only for training and validating the landmark detection model, since images obtained from online sources do not necessarily follow a controlled acquisition protocol for metric calibration.

The pose estimation subset was divided into training and validation sets, totaling 99 images for training and 17 images for validation. The split was performed randomly and reproducibly using a fixed seed of 42, with approximately 85% of the images assigned to training and 15% assigned to validation.

The second subset consisted of 15 images acquired specifically for the calibration stage during a classical equestrian competition. These images were obtained under a standardized capture protocol, with the camera positioned 3 m from the horse, in landscape orientation, without digital zoom, with the device positioned at an average height of 1.20 m from the ground and aligned with the lateral plane of the animal. The device used for image acquisition was an iPhone 17 Pro Max.

During the acquisition of the calibration subset, the horses were kept on flat ground and in a quadrupedal standing position, that is, standing still and supported by all four limbs. The real height of each animal was obtained using a leveling hippometer, and this measurement was considered the experimental reference. The images acquired by the authors involved privately owned animals and were obtained with permission from the responsible owners or handlers. Since the procedure was non-invasive and consisted only of photographic recording and reference height measurement, no experimental animal intervention was performed. The use of image-based measurements in this study was supported by previous works that explored digital morphometric measurement, machine vision, and pose estimation approaches for horses and livestock, reinforcing the potential of computer vision as a non-contact alternative for extracting body measurements from images [4], [8], [9].

Finally, two additional images were acquired under the same standardized capture protocol and used only for the preliminary complete-pipeline demonstration. These images were not used for YOLO training, validation, or pixel-to-centimeter calibration.

C. Image Annotation Protocol

All images in the pose estimation subset were manually annotated using the Computer Vision Annotation Tool (CVAT). The annotations were performed by a single annotator to avoid inter-annotator variability. As illustrated in Fig. 2, each image was annotated with a bounding box surrounding the entire horse and two anatomical keypoints: the withers and

the contact point between the most visible front hoof and the ground.



Fig. 2. Manual annotation of the horse bounding box and anatomical keypoints used for height estimation.

The bounding box was included to guide the model in identifying the horse within the image, limiting the region of interest and reducing background interference during keypoint detection. However, it was not directly used to calculate the height in pixels. The height estimation was based only on the anatomical keypoints, which were selected to represent the upper and lower references of the vertical measurement.

The withers were selected as the upper reference point because they correspond to the conventional anatomical landmark used for horse height measurement. According to the veterinary regulations of the Fédération Equestre Internationale (FEI), the measurement is taken at the highest point of the withers, immediately above the spinous process, usually associated with the fifth thoracic vertebra. The same regulations also indicate that the animal should be in a normal standing position, weight-bearing on all four limbs, before measurement [2].

Since the proposed methodology is based on two-dimensional lateral images, the lower reference point was operationally defined as the contact point between the most visible front hoof and the ground. This point represents the lower vertical reference required to estimate the animal's height in the image, aligning the computational measurement with the conventional procedure while adapting it to the visual information available in a lateral photograph. Additionally, the 15 calibration images were annotated following the same criteria to provide reference keypoint coordinates for fitting the pixel-to-centimeter regression model, as described in Section F.

D. YOLO Pose Model Training

For automatic anatomical landmark detection, a YOLO pose estimation model based on the yolo11s-pose architecture was used. This architecture was selected for its ability to simultaneously perform object detection and keypoint localization within the same structure [10], [11]. This characteristic is suitable for the objective of this study, since the estimated height depends on the accurate localization of the withers and the hoof contact point with the ground.

Training was performed using transfer learning from pre-trained weights. The images were processed with an input size of 960 pixels, a batch size of 8, and 150 training epochs. The AdamW optimizer was used, with an initial learning rate of 0.005, a weight decay of 0.0005, and a patience of 50 epochs.

Considering the limited size of the dataset, data augmentation techniques available in the YOLO training framework were applied during training, including variations in scale, translation, rotation, shear, hue, saturation, and brightness, as well as mosaic and mixup in controlled proportions. These strategies were adopted to increase the visual variability presented to the model and reduce the risk of overfitting.

E. Pixel-Based Height Estimation

After training, the model was used to automatically detect the withers and the hoof contact point in lateral horse images. Since the animal's height is represented by a vertical measurement, the initial estimate was calculated using only the y-axis coordinates of the two detected keypoints. The height in pixels was obtained as the absolute difference between the vertical coordinate of the hoof and the vertical coordinate of the withers, as follows:

$$H_{px} = |y_{hoof} - y_{withers}| \tag{1}$$

where H_{px} represents the estimated height in pixels, y_{hoof} corresponds to the vertical coordinate of the hoof contact point with the ground, and $y_{withers}$ corresponds to the vertical coordinate of the withers.

This pixel-based measurement corresponds to an intermediate stage of the system. Although it does not directly represent the unit of interest for the final user, it preserves the logic of conventional height measurement, which also considers the vertical distance between the withers and the ground.

F. Pixel-to-Centimeter Calibration

To convert the estimated height in pixels into centimeters, simple linear regression was used. The calibration was performed using the calibration subset described in Section B, which included images acquired under a standardized capture protocol and associated with real height measurements obtained using a leveling hippometer. For each calibration image, the pixel height was calculated from the reference keypoint annotations, following the same anatomical criteria described in Section C.

This measurement was obtained as the absolute difference between the vertical coordinates of the hoof contact point and the withers, as defined in Section E. The use of reference annotations in this stage allowed the regression model to be fitted from consistent image measurements before being applied to automatically predicted keypoints. The relationship between the image-based measurement and the real height was modeled as:

$$H_{cm} = a \cdot H_{px} + b \tag{2}$$

where H_{cm} represents the estimated height in centimeters, H_{px} corresponds to the height in pixels obtained from the reference keypoint annotations, and a and b are the coefficients fitted by the linear regression model.

G. Evaluation Metrics

The evaluation was organized into three complementary analyses: validation of the YOLO pose model, evaluation of the pixel-to-centimeter calibration, and a preliminary complete-pipeline demonstration. The YOLO pose model was evaluated on the validation set, composed of 17 images not used during weight updating. For each image, the predicted horse instance was matched to the reference annotation using the highest Intersection over Union (IoU) between predicted and annotated bounding boxes, with a minimum IoU threshold of 0.3. This criterion was used only to select the corresponding detected instance before evaluating the anatomical keypoints. The minimum confidence threshold used during validation was 0.05.

Keypoint accuracy was assessed by the Euclidean pixel error between predicted and reference points, calculated separately for the withers and hoof contact point. Pixel-based height was evaluated by comparing the height obtained from predicted keypoints with that obtained from reference annotations, using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination R^2.

The calibration analysis was performed using 15 images acquired under the standardized capture protocol. Pixel heights were obtained from CVAT reference annotations and related to the corresponding hippometer measurements using simple linear regression. Calibration performance was evaluated using leave-one-out cross-validation (LOOCV), with 14 samples used for fitting and one for testing at each iteration. The resulting estimates were compared with the real heights using MAE and RMSE, while R^2 was reported for the regression fitted to the complete calibration subset. This analysis was based on reference keypoint annotations and therefore characterized the pixel-to-centimeter conversion stage.

Finally, the complete system was demonstrated on two additional images not included in the training, validation, or calibration subsets. In this stage, the predicted instance was selected by the highest detection confidence, simulating real use of the system. The pixel height obtained from the YOLO-predicted keypoints was converted into centimeters using the calibrated regression model and compared with the reference height measured using the leveling hippometer.

III. RESULTS AND DISCUSSION

The experimental results are organized according to the main stages of the proposed pipeline: anatomical landmark detection, pixel-to-centimeter calibration, and preliminary complete-pipeline demonstration.

A. YOLO Pose Validation

The YOLO pose model was evaluated on the validation set composed of 17 images that were not used during weight updating. The model detected the horse in all validation images, resulting in a detection rate of 100%. For the pixel-based height estimation, the comparison between the values obtained from the predicted keypoints and those obtained from the reference annotations resulted in an R^2 of 0.9905, RMSE of 24.78 px, and MAE of 19.61 px. When normalized by the reference pixel-height measurements, the mean absolute error corresponded to 3.10%. These results indicate a strong agreement between predicted and reference pixel heights, suggesting that the model was able to preserve the vertical measurement required for subsequent height estimation.

The anatomical landmarks showed different levels of localization error. The MAE was 16.65 px for the withers and 46.72 px for the hoof contact point. The higher error observed for the hoof is consistent with the greater visual variability of the distal limb region, where ground contact, limb overlap, shadow, and background texture may affect point localization. Despite this difference between keypoints, the overall pixel-height error remained moderate relative to the reference pixel-height scale, which is the most relevant factor for the proposed height estimation pipeline.

The mean bounding box IoU was 0.873, indicating that the detected horse instances were well aligned with the reference annotations. However, the bounding box was used only as an auxiliary criterion to match the predicted instance to the annotated animal during validation, and it was not used to calculate height. As shown in Fig. 3, the model localized the withers and hoof contact point and generated a predicted pixel-height measurement close to the reference annotation, illustrating the agreement observed in the quantitative validation results.



Fig. 3. Representative example of YOLO pose prediction and reference annotation, illustrating the procedure used for pixel-height validation.

B. Pixel-to-Centimeter Calibration

After validating the pixel-based measurement, a linear regression model was used to convert height in pixels into

centimeters. The calibration analysis was performed using 15 images acquired under the standardized capture protocol. The reference heights of the animals used for calibration ranged from 151 to 175 cm, with a mean height of 165.53 cm. The fitted regression model was represented by:

$$H_{cm} = 0.0575 \cdot H_{px} + 35.7890, \tag{3}$$

where H_{cm} is the estimated height in centimeters and H_{px} is the height in pixels obtained from the reference anatomical landmarks. In the complete fit, the model achieved an MAE of 2.20 cm, RMSE of 2.55 cm, and R^2 of 0.8195. These results indicate that the pixel-height measurement was able to explain a substantial part of the variation in real height, even though the calibration subset was limited in size.

Leave-one-out cross-validation was used to obtain a more conservative estimate of calibration performance. In this evaluation, the model achieved an MAE of 2.56 cm and RMSE of 2.96 cm. Since this analysis was based on reference keypoint annotations, these values characterize the pixel-to-centimeter calibration stage. The increase in error compared with the complete fit was expected, since each prediction was made using a model trained without the tested sample. Even so, the LOOCV error remained close to the fitting error, suggesting that the relationship between pixel height and real height was reasonably stable under the standardized acquisition protocol. The calibration curve is shown in Fig. 4.

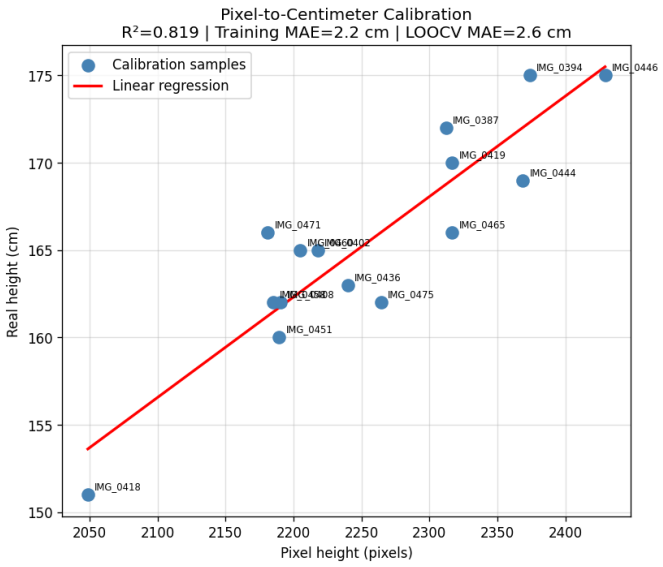


Fig. 4. Linear regression between pixel height and real height in centimeters.

Since the calibration subset included animals between 151 and 175 cm, estimates for horses substantially outside this interval should be interpreted with caution, as they would involve extrapolation beyond the fitted range. Additionally, the fitted regression is specific to the acquisition geometry used in this study, including camera distance, device height, image orientation, and lateral alignment. Changes in these conditions may require a new calibration procedure.

C. Complete Pipeline Demonstration

The complete system was demonstrated using two additional images ($n = 2$) that were not included in the training, validation, or calibration subsets. In this stage, the full automatic pipeline was applied: the YOLO pose model detected the anatomical keypoints, the pixel height was calculated from the predicted coordinates, and the calibrated regression model converted the measurement into centimeters. The results obtained for each image are presented in Table I.

TABLE I
COMPARISON BETWEEN REFERENCE AND ESTIMATED HORSE HEIGHT.

Image	Reference height (cm)	Estimated height (cm)	Absolute error (cm)	Percentage error (%)
IMG_0514	167.00	170.96	3.96	2.37
IMG_0523	160.00	161.76	1.76	1.10

The mean absolute error in this preliminary complete-pipeline demonstration was 2.86 cm, with a mean percentage error of 1.74% and RMSE of 3.06 cm. These values were close to the LOOCV calibration error, indicating that the complete system behaved consistently with the expected performance of the regression model in the two tested cases. However, because this stage was limited to two independent images, these results should be interpreted as an initial demonstration of feasibility rather than a conclusive estimate of system performance. These results suggest that the proposed pipeline can provide useful height estimates when the acquisition protocol is followed, although further testing with a larger number of images is still required.

D. Further Discussion

The results obtained in this study support the feasibility of estimating horse height from a single lateral image using anatomical landmark detection and regression-based calibration. This outcome is aligned with previous studies that have explored image-based morphometric assessment, computer vision, and keypoint detection for equine body measurement [4], [8], [9]. Although these studies differ in acquisition protocols, measured traits, and computational strategies, they reinforce the potential of digital imaging methods as alternatives or complements to traditional morphometric measurement.

When compared with related image-based approaches, the proposed method presented results within a similar order of magnitude, while adopting a simplified height-specific pipeline. Freitag et al. used standardized photographs and ImageJ-based analysis to estimate equine body measurements, reporting strong agreement between image-derived and conventional measurements [3]. More automated approaches based on deep learning and machine vision have reported mean relative errors of 2.98% for withers height and 4.01% for shoulder height in Mongolian horses [7]. Similarly, a method combining target detection and keypoint detection reported mean relative errors of 3.86% for shoulder height,

3.98% for body height, and 2.89% for croup height [8]. In the present study, the YOLO validation resulted in a normalized pixel-height MAE of 3.10%, while the preliminary complete-pipeline demonstration resulted in a mean percentage error of 1.74% in two independent images. These results suggest that the proposed approach is compatible with the performance range reported by related image-based methods, although the complete-pipeline result should be interpreted cautiously due to the limited number of final test images.

A central contribution of the proposed approach is its simplified and height-specific design. Unlike broader morphometric systems, which may require multiple body points, depth information, or more elaborate acquisition setups, the present method estimates height using only two anatomical landmarks: the withers and the hoof contact point. This allows the complete estimation process to be performed from a single lateral RGB image, positioning the method as an intermediate alternative between direct manual measurement and more complex computer vision systems. Compared with approaches that require a physical scale reference in each image, the proposed strategy relies on a fixed acquisition geometry combined with linear calibration. Since fixed distances and spatial references are commonly used in handling, training, and competition settings, the proposed acquisition protocol may be simple to understand and apply in practical equine environments.

The standardized capture protocol was essential for the performance of the calibration model. Since a monocular image does not inherently provide metric scale, controlling distance, camera orientation, zoom, and animal posture helped reduce variation not related to the animal's real height. Therefore, the fitted regression model should be interpreted as specific to the acquisition geometry used in this study, including camera distance, device height, image orientation, and lateral alignment. For this reason, estimates for horses substantially outside the calibration range of 151 to 175 cm should be interpreted with caution, since they would involve extrapolation beyond the fitted data.

Future work should expand both the annotated image dataset used to train the YOLO pose model and the standardized image set used for pixel-to-centimeter calibration. A larger and more diverse dataset may improve landmark detection and calibration robustness across different animal sizes, conformations, coat colors, backgrounds, and acquisition conditions. Additional experiments should also evaluate the influence of camera distance, camera height, lateral alignment, image resolution, lighting conditions, and animal posture on the final height estimate. These experiments may help determine whether a single calibration model is sufficient or whether device-specific or protocol-specific calibration models are required under different acquisition conditions.

Despite these limitations, the results indicate that the proposed method is a promising initial step toward a practical single-image alternative for horse height estimation.

IV. Conclusion

This study presented an image-based methodology for estimating horse height from a single lateral photograph by combining YOLO pose detection, pixel-height calculation from two anatomical landmarks, and linear regression calibration. The results support the feasibility of the proposed approach as an initial proof-of-concept, showing that the pipeline can estimate horse height under a standardized acquisition protocol. The method has potential as a practical single-image alternative that does not require depth sensors, multiple viewpoints, or specialized acquisition devices. However, the complete automatic pipeline was demonstrated on a limited number of independent images, and the reported final estimates should therefore be interpreted as preliminary evidence of feasibility rather than a conclusive measure of system performance.

Future work should expand the annotated dataset used for YOLO training and the calibration set used for pixel-to-centimeter conversion. Additional tests under different acquisition conditions and with a larger number of horses are also required to improve robustness and support broader practical application.

ACKNOWLEDGMENT

The authors would like to thank all those who assisted during image acquisition and reference height measurement.

References

[1] A. Giontella, M. Silvestrelli, A. Cocciolone, C. Pieramati, and F. M. Sarti, "Breeding value estimation based on morphological evaluation of the Maremmano horse population through factor analysis," *Animals*, vol. 14, no. 15, Art. no. 2232, 2024, doi: 10.3390/ani14152232.

[2] Fédération Equestre Internationale, *2025 Veterinary Regulations*, 15th ed., Chapter IX, Articles 1082–1083, effective Jan. 1, 2025.

[3] G. P. Freitag, L. G. F. de Lima, J. A. Jacomini, L. E. Kozicki, and L. B. Ribeiro, "An accurate image analysis method for estimating body measurements in horses," *Journal of Equine Veterinary Science*, vol. 101, Art. no. 103418, 2021, doi: 10.1016/j.jevs.2021.103418.

[4] N. Ünal, Y. Paksoy, and Ö. F. Güngör, "Agreement of the morphometric dimensions determined by traditional and image analysis methods for Arabian and Thoroughbred horses," *Archives of Animal Breeding*, vol. 68, pp. 253–261, 2025, doi: 10.5194/aab-68-253-2025.

[5] M. R. dos Santos, G. Freiberger, F. Bottin, M. Chiocca, A. Zampar, and D. de C. Cucco, "Evaluation of methodologies for equine biometry," *Livestock Science*, vol. 206, pp. 24–27, 2017, doi: 10.1016/j.livsci.2017.10.009.

[6] M. Pérez-Ruiz, D. Tarrat-Martín, M. J. Sánchez-Guerrero, and M. Valera, "Advances in horse morphometric measurements using LiDAR," *Computers and Electronics in Agriculture*, vol. 174, Art. no. 105510, 2020, doi: 10.1016/j.compag.2020.105510.

[7] L. Su, M. Li, Y. Zhang, and Z. Zong, "A measurement method for body parameters of Mongolian horses based on deep learning and machine vision," *Applied Sciences*, vol. 14, no. 13, Art. no. 5655, 2024, doi: 10.3390/app14135655.

[8] L. Su, M. Li, Y. Zhang, Z. Zong, and C. Gong, "Fusion of target and keypoint detection for automated measurement of Mongolian horse body measurements," *Agriculture*, vol. 14, no. 7, Art. no. 1069, 2024, doi: 10.3390/agriculture14071069.

[9] J. Sosa and D. C. Hogg, "A horse with no labels: Self-supervised horse pose estimation from unlabelled images and synthetic prior," in *Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 1041–1048, doi: 10.1109/ICCVW60793.2023.00112.

[10] Ultralytics, "YOLO11," *Ultralytics Docs*, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11. Accessed: Jun. 12, 2026.

[11] Ultralytics, "Pose Estimation," *Ultralytics Docs*, 2024. [Online]. Available: https://docs.ultralytics.com/tasks/pose. Accessed: Jun. 12, 2026.

Experimental Characterization of the Output Stage of a Functional Electrical Stimulator Based on a Howland Current Source

T. P. Silva
Laboratório de Engenharia Biomédica
Centro Universitário FEI
São Bernardo do Campo, Brazil
tpsilva@fei.edu.br

M. C. F. Castro
Laboratório de Engenharia Biomédica
Centro Universitário FEI
São Bernardo do Campo, Brazil
mclaudia@fei.edu.br

***Abstract*—Open-loop Functional Electrical Stimulation (FES) output stages require a characterized relationship between the digital-to-analog converter (DAC) command and output current. This bench study evaluates a Howland current-source stage based on the STIMGRASP architecture. Four separate acquisitions were recorded for each of three nominal resistive loads (1, 2, and 4.7 kΩ), each comprising one ascending DAC sweep. For each load, the first acquisition was used for calibration and model development; the three subsequent acquisitions were held out. Across the nine held-out acquisitions (1683 DAC-bin means; 10 mV bin width), the frozen model gave a pooled bias of 0.200 mA, MAE of 0.529 mA, RMSE of 0.707 mA, and maximum absolute error of 1.815 mA. Mean per-acquisition RMSE was 0.855, 0.389, and 0.514 mA for the 1, 2, and 4.7 kΩ loads, respectively. The repeated acquisitions support acquisition-to-acquisition validation under the same bench setup, while persistent load-dependent bias shows that the single line does not capture all systematic behavior. The nominal 5%-tolerance shunt and loads were measured as 10.2, 984, 1975, and 4592 Ω; these values were used in the analysis. The model remains limited to the evaluated hardware, resistive loads, and frozen load-specific ranges, and embedded use requires explicit command-range checks.**

***Index Terms*—Functional Electrical Stimulation, Howland current source, linearity analysis, open-loop control.**

I. Introduction

Functional Electrical Stimulation (FES) systems electrically activate peripheral nerves in order to assist or restore motor functions in rehabilitation and assistive applications [1], [2]. FES uses electrical pulses to evoke muscle contractions for functional assistance or rehabilitation. Stimulation systems typically allow the configuration of pulse amplitude, width, frequency, and waveform shape. Among these variables, amplitude control is particularly important because it is associated with the recruited neuromuscular response and user comfort [3]. Before any application-level inference, the relationship between the embedded command variable and the output current must be characterized on the bench.

This work is related to the STIMGRASP FES master's thesis developed at Centro Universitário FEI by Barelli, which proposed a portable neuromuscular electrical stimulator for grasp restoration [4]. The original work describes a digitally controlled stimulation architecture based on a modified Howland current source and discusses linear stimulation sweeps and current-controlled stimulation behavior. In that investigation, predictable current control behavior is assumed as part of the stimulation architecture. However, detailed DAC-versus-current characterization curves under different load conditions are not presented, and the behavior implied by the architecture is not quantitatively characterized through correlation, regression, compliance, or load-consistency analyses.

The present work extends the original investigation by experimentally evaluating the output-stage transfer characteristic of the FES circuit. The purpose is not to redesign the STIMGRASP hardware or to provide a clinical validation of the stimulator. Instead, this study focuses on output-stage characterization: the relationship between DAC control voltage and delivered current, the effect of different resistive loads, the identification of a valid compliance region, and the derivation of an embedded DAC estimation model for open-loop operation.

The STIMGRASP circuit does not include current feedback to the microcontroller unit (MCU). Consequently, the output stage operates as an open-loop stimulation system, where the delivered current is not directly monitored during operation. A bench-derived DAC-to-current relation can provide a candidate command mapping only inside its characterized domain. If the load changes substantially, the output stage may reach voltage compliance and the delivered current may be lower than the modeled current. Without current feedback, this condition cannot be detected by the MCU in real time.

The main contributions of this paper are: (i) quantitative bench characterization of the DAC-voltage-to-output-current relationship; (ii) empirical compliance-region analysis based on inferred load voltage; (iii) load-specific and pooled calibration errors; (iv) held-out acquisition validation of the frozen global model; (v) investigation of its structured residuals; and (vi) explicit limits and range-check requirements for any prospective open-loop implementation.

II. Background

A. Howland Current Source and Compliance Voltage

The Howland current source is commonly used when a voltage command must be converted into a load current [5]. Ideally, the output current is determined by the input control voltage and resistor network, and it remains independent of the load impedance. In practice, load-independent behavior is maintained only while the amplifier output can provide the voltage required by the load. This constraint is described by the compliance voltage [6].

For a resistive load, the load voltage can be reconstructed from the measured current as

$$V_{\text{load}} = I_{\text{out}} R_{\text{load}}, \tag{1}$$

where I_{out} is expressed in amperes and R_{load} in ohms. Current-source operation is expected to remain valid inside signed negative and positive compliance boundaries,

$$V_{\text{comp},-} < V_{\text{load}} < V_{\text{comp},+}. \tag{2}$$

Samples near or beyond the available output-voltage range may exhibit compliance-limited behavior, particularly for higher load resistances.

B. Open-Loop DAC Estimation

Because the evaluated FES circuit does not provide real-time current feedback to the MCU, the embedded controller cannot directly regulate the delivered current during stimulation. Instead, current configuration must rely on an experimentally identified DAC-to-current relationship to estimate the DAC voltage required for a desired stimulation current. The experimentally identified direct model is

$$I_{\text{out}} = a V_{\text{DAC}} + b, \tag{3}$$

where V_{DAC} is the DAC control voltage and a and b are regression coefficients.

The embedded DAC estimation equation is obtained by rearranging Equation 3:

$$V_{\text{DAC}} = \frac{I_{\text{target}} - b}{a}. \tag{4}$$

Equation 4 is valid only within the experimentally characterized range. Outside this range, the output current is constrained by the available output voltage and load resistance, and the model may predict a current that cannot be physically delivered once the output stage reaches its compliance limit. Because the present open-loop implementation provides no current or output-voltage feedback to the microcontroller, the firmware cannot detect this condition during operation. It should therefore reject current requests outside the validated range and verify that the resulting DAC command remains within the corresponding characterized interval. These range checks prevent model extrapolation but cannot detect compliance caused by an unknown or time-varying load. Future implementations should incorporate current feedback or compliance monitoring, together with an independent hardware current limit.

III. Materials and Methods

A. Experimental Setup

The evaluated output stage was an FES circuit based on a Howland current source. Dedicated experimental firmware generated an ascending DAC sweep bracketed by initial and final resting intervals; it was not the conventional stimulation firmware and did not generate a therapeutic pulse train. The nominal resting point was approximately 1.65 V, corresponding to approximately zero output current. The DAC was swept from approximately 0 to 3.3 V.

Figure 1 shows the evaluated DAC and Howland-current-source output stage as extracted from the STIMGRASP documentation. It also defines the measurement convention: CH1 is the DAC node, and CH2 is the voltage at the load–shunt junction relative to ground. Thus, $I_{\text{out}} = V_{\text{CH2}}/R_{\text{shunt}}$.

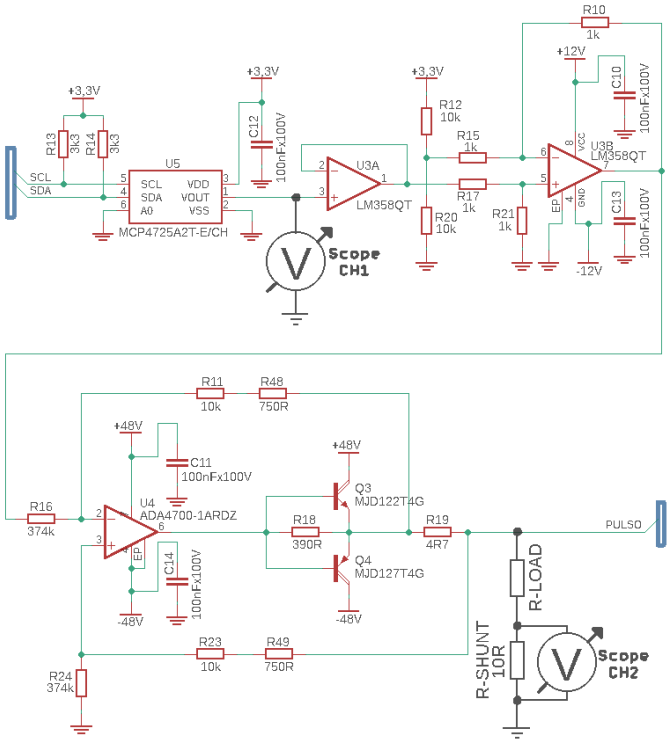


Fig. 1. Evaluated DAC and Howland-current-source output stage.

The 0–3.3 V DAC command is buffered by U3A, centered by U3B about the nominal $V_{\text{ref}} = 1.65$ V, and inverted by the Howland stage, so $I_{\text{out}} \propto -(V_{\text{DAC}} - V_{\text{ref}})$ and a negative slope is expected. The bipolar supplies let commands below and above the midpoint produce positive and negative current, respectively. The former is defined as flow through the load and shunt to ground; the sign of V_{load} therefore denotes current direction.

The nominal 10 Ω shunt and 1, 2, and 4.7 kΩ loads all had a stated tolerance of ±5%. They were measured with a UNIT-T UT89XD multimeter as 10.2, 984, 1975, and 4592 Ω, respectively; all readings were within the stated component tolerance. The measured 10.2 Ω shunt was used to convert

CH2 voltage to current, and the measured load values were used to infer V_{load}. The resistive loads provide controlled bench conditions but are not models of the nonlinear, time-varying, and electrode-dependent skin–electrode interface [7].

B. Data Acquisition

Raw data were recorded with a Tektronix MSO46B oscilloscope in 12 CSV files, corresponding to four separate acquisitions for each of the three loads. Each acquisition contained one complete ascending DAC sweep bracketed by initial and final resting intervals. For each load, the first acquisition was used for calibration and model development, whereas the three subsequent acquisitions were reserved exclusively for held-out testing. No held-out acquisition was used to determine the compliance bounds, select the model form, or estimate the regression coefficients. CH1 measured the DAC voltage and CH2 measured the shunt voltage, as shown in Figure 1. Both channels used DC coupling and 1× probes. The acquisition configuration is summarized in Table I.

TABLE I
MEASUREMENT AND ACQUISITION CONFIGURATION

Parameter	Value
Oscilloscope model	MSO46B
Channels	CH1: DAC; CH2: shunt voltage
Time base	2 s/div
Sampling rate	12.5 kS/s
Configured record length	250 kpoints
Acquisition mode/resolution	Sample, 12 bits
Channel coupling	DC (CH1 and CH2)
Probe attenuation	1× (CH1 and CH2)
Resistance meter	UNIT-T UT89XD
Shunt (nom./meas.)	10/10.2 Ω
Loads (nom./meas.)	1000/984; 2000/1975; 4700/4592 Ω

Figure 2 shows a representative raw recording for the 1 kΩ load. The DAC remains near the nominal zero-current command during the initial and final resting intervals and performs one complete ascending sweep between them. The abrupt DAC transitions delimit the portion subsequently processed as the active sweep segment.

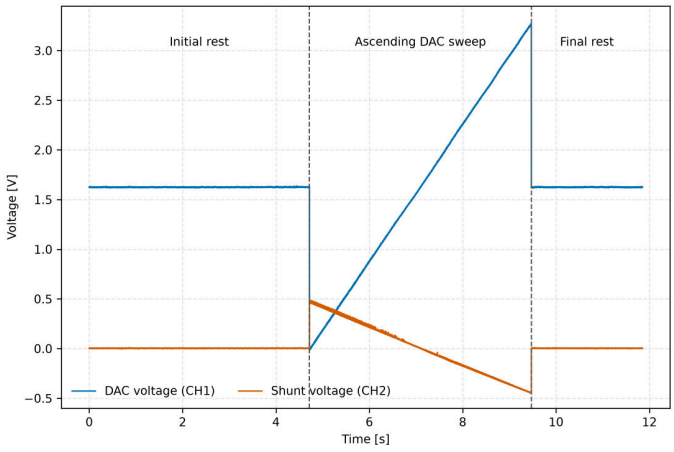


Fig. 2. Representative raw DAC and shunt voltages for the 1 kΩ load.

C. Data Processing

The active sweep segment in each acquisition was delimited by the first and last DAC discontinuities larger than 0.5 V, with a 512-sample (40.96 ms) guard removed at each boundary. The active segments of the calibration acquisitions for 1, 2, and 4.7 kΩ contained 58,386, 58,386, and 58,389 time samples, respectively; those of the nine held-out acquisitions contained 58,395–58,423 samples. Files were processed separately without sample-count equalization. Shunt voltage was converted to output current as $I_{\mathrm{out}} = V_{\mathrm{CH2}}/R_{\mathrm{shunt}}$ using the measured 10.2 Ω value; because the load and shunt were in series, this was also the load current. Output voltage was not measured directly; V_{load} was inferred from Equation 1 using the corresponding measured load resistance.

The DAC axis was rounded to 10 mV bins. For each load and acquisition, the analysis value at a bin was the arithmetic mean of all time-sample currents in that bin. Figure 3 shows these binned output-current values versus DAC voltage for the complete calibration acquisition of each load. A bin was not treated as an independent experimental replicate. The calibration set contained 325 bin means per load before compliance selection. The flattened portions for the larger loads identify output-voltage compliance. Regression used only the selected calibration-bin means; test metrics were first calculated within each complete held-out acquisition and then summarized across the three acquisitions per load.

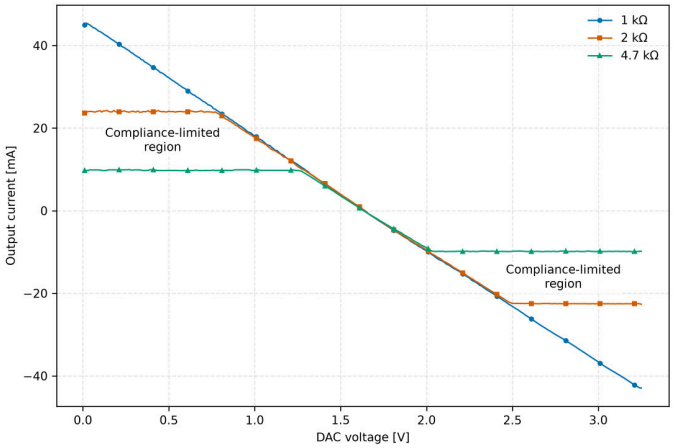


Fig. 3. Binned output current versus DAC voltage.

D. Compliance-Limited Region Selection

For each load, the minimum and maximum binned currents were converted to load-voltage bounds using the measured resistance. The shared empirical selection interval was defined as

$$V_{\mathrm{min,limit}} = \max(V_{\mathrm{min},1k}, V_{\mathrm{min},2k}, V_{\mathrm{min},4.7k}), \qquad (5)$$

and

$$V_{\mathrm{max,limit}} = \min(V_{\mathrm{max},1k}, V_{\mathrm{max},2k}, V_{\mathrm{max},4.7k}). \qquad (6)$$

This produced $-42.257 < V_{\mathrm{load}} < 44.695$ V. These are data-derived selection bounds, not independently measured amplifier compliance ratings. Because plateau noise can place isolated saturated bins just inside a strict voltage boundary, the largest contiguous 10 mV DAC run inside the interval was retained for each load; disconnected bins were rejected. This rule was applied exclusively to the calibration acquisitions, and the resulting load-specific DAC-bin sets were frozen before being applied unchanged to the nine held-out acquisitions. The complete binned calibration data remain available to describe the observed full-range behavior.

After this selection, the retained current-versus-DAC curves were similar but did not have the same DAC support: the larger loads reached compliance earlier.

E. Statistical Analysis and Model Identification

Ordinary least-squares lines were fitted separately to each load and to all retained calibration bin means pooled across loads. For each fit, the slope, intercept, R^2, mean absolute error (MAE), root mean square error (RMSE), and maximum absolute residual were calculated on the fitting data and labeled as calibration metrics. The primary predictive model was the pooled line fitted to the three calibration acquisitions; its coefficients, preprocessing, compliance bounds, and exact load-specific DAC-bin sets were frozen before evaluating the nine held-out acquisitions.

No correlation p-values or point-wise cross-validation were used as evidence of repeatability. Adjacent 10 mV bins from a monotonic DAC sweep are not independent experimental replicates, and a random split would place neighboring bins from the same acquisition in both training and test sets. Instead, each complete acquisition was kept intact. For every held-out acquisition, bias, MAE, RMSE, and maximum absolute error were calculated without refitting; per-load summaries report the mean and sample standard deviation across the three acquisitions. Pooled held-out errors are also given descriptively but reflect the unequal number of valid bins among loads.

For the pooled linear calibration, residual was defined as $e = I_{\mathrm{measured}} - I_{\mathrm{predicted}}$. Bias, MAE, RMSE, maximum absolute error, and lag-1 autocorrelation between residuals in consecutive 10 mV DAC bins were computed separately by load. The pooled line was also compared descriptively with a centered quadratic model and a continuous two-segment linear model with a breakpoint (knot) at 1.65 V. These alternatives were introduced after inspecting the calibration residuals and were not used for the primary held-out claim; choosing among them using the nine tests would require another untouched validation set.

Finally, a calibration sensitivity fit was restricted to the DAC bins represented by all three loads. This exposes the effect of the unequal pooled domains: 1 kΩ contributes 322 bins, whereas 4.7 kΩ contributes only 75. This sensitivity fit was not substituted for the frozen primary equation.

IV. Results

A. Linear Operating Region

The empirical compliance selection retained 322, 164, and 75 bin means for the calibration acquisitions at 1, 2, and 4.7 kΩ, respectively. Table II places the load-specific domains, coefficients, and error measures in one implementation-oriented summary. Each N_{bins} counts 10 mV means from one acquisition, not independent time samples or repeat acquisitions.

B. Linear Regression and Global Model

The frozen pooled calibration had the negative slope expected from the inverting topology shown in Figure 1. Its experimentally identified zero crossing, $-b/a \approx 1.653$ V, was consistent with the nominal 1.65 V reference:

$$I_{\mathrm{out,mA}} = -27.489872 V_{\mathrm{DAC}} + 45.442251. \tag{7}$$

The in-sample $R^2 = 0.999662$ reported in Table II quantifies explained calibration variance; predictive performance was assessed on the held-out acquisitions.

C. Held-Out Acquisition Validation

Table III reports the primary validation of Equation 7. The 1 kΩ tests covered 322 of 322 frozen calibration bins (100.0%), from 0.03 to 3.24 V; all 164 and 75 bins were available for the 2 and 4.7 kΩ tests, respectively. No test coefficient or compliance bound was re-estimated. Across all 1683 held-out bin means, the frozen line gave a bias of 0.2001 mA, MAE of 0.5289 mA, RMSE of 0.7070 mA, and maximum absolute error of 1.8148 mA. Because the pooled result is weighted toward the wider 1 kΩ domain, the per-load and per-acquisition results are the more informative repeatability summary.

Between-acquisition dispersion was small for 1 and 2 kΩ and larger for 4.7 kΩ. More importantly, the repeated-acquisition biases retained the load-dependent signs seen in calibration: positive for 1 kΩ and negative for the larger loads. The held-out RMSE was also higher than the 0.3810 mA pooled calibration RMSE. Thus, the additional data demonstrate stable acquisition-to-acquisition behavior under this setup but confirm systematic model-form/load effects rather than random residual scatter.

Only 75 DAC bins, from 1.26 to 2.00 V, were represented by all three loads. Restricting the sensitivity fit to this common support changed the line to $I_{\mathrm{out,mA}} = -27.0241 V_{\mathrm{DAC}} + 44.3638$ and gave an in-sample RMSE of 0.2444 mA. The coefficient change demonstrates that the pooled result depends on its unequal load composition; the wider pooled range must not be interpreted as support from all loads.

A model-form diagnostic on the same 561 calibration bins reinforced the residual finding. The global linear MAE, RMSE, and maximum error were 0.3241, 0.3810, and 1.1281 mA; corresponding values were 0.1835, 0.2359, and 0.9429 mA for a centered quadratic and 0.1552, 0.2117, and 0.9698 mA for a segmented linear model with a breakpoint at 1.65 V. These alternatives were formulated from the calibration residuals

TABLE II
OPERATING DOMAINS AND IN-SAMPLE LINEAR CALIBRATION METRICS

Load/model	Acq.	N_{bins}	V_{DAC} (V)	I_{out} (mA)	V_{load} (V)	a (mA/V)	b (mA)	R^2	MAE (mA)	RMSE (mA)	Max. (mA)
1 kΩ	1	322	0.03–3.24	−42.882–45.218	−42.196–44.494	−27.5664	45.7041	0.999818	0.2815	0.3461	0.7291
2 kΩ	1	164	0.82–2.45	−21.330–22.610	−42.127–44.655	−27.0874	44.6812	0.999799	0.1505	0.1819	0.4123
4.7 kΩ	1	75	1.26–2.00	−8.989–9.677	−41.276–44.437	−25.6806	42.1314	0.999657	0.0904	0.1030	0.2535
Pooled linear	3[a]	561	load dependent[b]	load dependent[b]	−42.257–44.695[c]	−27.4899	45.4423	0.999662	0.3241	0.3810	1.1281

[a]One calibration source file for each load; all nine held-out files were excluded. [b]The pooled row is the union of the load-specific domains; it is not a domain supported by every load. [c]Empirical signed selection bounds inferred with measured resistances. All errors in this table are evaluated on the calibration bins used to fit the corresponding row.

TABLE III
HELD-OUT ERROR OF THE FROZEN GLOBAL LINEAR MODEL

Load	Test acquisition	N_{bins}	V_{DAC} (V)	Bias (mA)	MAE (mA)	RMSE (mA)	Max. (mA)
1 kΩ	1	322	0.03–3.24	0.5160	0.6311	0.8477	1.7481
1 kΩ	2	322	0.03–3.24	0.4596	0.6305	0.8512	1.7969
1 kΩ	3	322	0.03–3.24	0.4748	0.6459	0.8653	1.8148
1 kΩ	Mean ± SD	3 acquisitions	—	0.483 ± 0.029	0.636 ± 0.009	0.855 ± 0.009	1.8148[a]
2 kΩ	1	164	0.82–2.45	−0.0746	0.3380	0.3756	0.6952
2 kΩ	2	164	0.82–2.45	−0.1368	0.3676	0.3987	0.6710
2 kΩ	3	164	0.82–2.45	−0.1265	0.3670	0.3932	0.6217
2 kΩ	Mean ± SD	3 acquisitions	—	-0.113 ± 0.033	0.358 ± 0.017	0.389 ± 0.012	0.6952[a]
4.7 kΩ	1	75	1.26–2.00	−0.2654	0.3745	0.4359	1.1570
4.7 kΩ	2	75	1.26–2.00	−0.3005	0.4178	0.4814	1.1790
4.7 kΩ	3	75	1.26–2.00	−0.4313	0.5409	0.6239	1.1211
4.7 kΩ	Mean ± SD	3 acquisitions	—	-0.332 ± 0.087	0.444 ± 0.086	0.514 ± 0.098	1.1790[a]
All loads	Pooled bins	1683	Load specific	0.2001	0.5289	0.7070	1.8148

Residual is measured minus predicted current. Per-load bias, MAE, and RMSE summaries are mean ± sample SD across three complete acquisitions; [a] The Max. entry is the worst absolute error among the three acquisitions, not a mean. Pooled metrics use all eligible bin means and therefore inherit unequal load support.

and were not substituted for the prespecified linear model in Table III; selecting one after viewing the held-out results would make those acquisitions a model-selection set rather than an untouched validation set.

D. Residuals and Load Consistency

The pooled model has different calibration error profiles by load. For 1 kΩ, bias was 0.1367 mA, RMSE 0.3788 mA, and maximum error 0.7430 mA. Corresponding values were -0.1030, 0.2828, and 0.5777 mA for 2 kΩ, and -0.3617, 0.5430, and 1.1281 mA for 4.7 kΩ. Adjacent-bin residual autocorrelations were 0.981, 0.981, and 0.996, showing strong persistence along each calibration sweep. These are descriptive diagnostics; no inferential p-values are assigned to adjacent bins.

Figure 4 resolves the previously pooled calibration residuals by load and overlays the three held-out traces for each load. The smooth calibration guides reveal curvature and distinct load-dependent tracks rather than random scatter around zero, while the held-out traces show that the structure persists across acquisitions. In particular, the individual calibration slopes differ from the pooled slope by -0.077, 0.402, and 1.809 mA/V for 1, 2, and 4.7 kΩ, respectively.

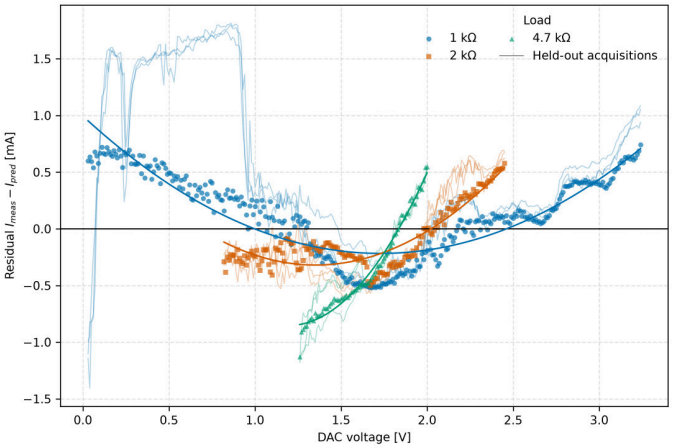


Fig. 4. Residuals of the frozen pooled linear model. Markers and solid guides show calibration residuals; thin light-colored traces show the three held-out acquisitions for each load. Guides are descriptive, not alternative prediction models.

V. DISCUSSION

The DAC-to-current response is nearly linear in a variance-explained sense after empirical compliance selection, but R^2 alone is not an accuracy or repeatability metric for a dense monotonic DAC sweep. The complete held-out acquisitions provide the relevant test: pooled RMSE increased from 0.3810 mA in calibration to 0.707 mA in testing, and the held-out maximum deviation was 1.815 mA. The small within-load SD across three held-out acquisitions supports repeatability under the same setup, but the load-dependent mean errors limit the accuracy of one global line.

The residual structure qualifies the original single-line interpretation. Strong serial persistence, load-dependent bias, visibly curved calibration tracks, and persistence of the bias signs in the repeated acquisitions show that the remaining deviations are systematic. The diagnostic segmented model reduces in-sample RMSE, but it was derived after inspecting the calibration residuals. The present tests therefore validate

the frozen linear model only; using them to promote an alternative would require another untouched dataset for an unbiased final assessment.

The pooled dataset is also unbalanced in DAC support. At the outer voltages, only the 1 kΩ load remains inside the selected interval; all three loads are represented together only from 1.26 to 2.00 V. Consequently, Equation 7 is a candidate calibration for the three observed load-specific domains, not evidence of load invariance over 0–3.3 V. The empirical signed load-voltage bounds are similarly a calibration-set selection rule rather than independently measured compliance specifications.

The nine held-out acquisitions provide acquisition-level validation under the same hardware and measurement chain. Future work should repeat the characterization under varied supply, temperature, and load conditions to quantify the model's robustness beyond the evaluated bench setup.

The resistors characterize the output stage under simple bench conditions; they do not establish performance for stimulation pulses or the electrode–skin interface. A low-cost next experiment should add a documented RC or published electrode–skin equivalent circuit, including topology, measured component values, tolerances, initial capacitor state, and repeated acquisitions using DAC sweeps or stimulation pulses. Until such a load is tested, no inference about clinical current repeatability or outcomes is supported.

Any embedded use of Equation 4 requires finite-value and range checks, rejection to a safe state outside the validated current/DAC domain, and an independent hardware current limit. These checks prevent numerical extrapolation but cannot reveal compliance under an unknown load. Current feedback or compliance monitoring would still be required to detect that condition during operation.

VI. Conclusion

This study provides a reproducible bench calibration and held-out acquisition assessment of the STIMGRASP-related Howland output stage. The first acquisition at each of three nominal resistive loads defined the frozen pooled line; three subsequent acquisitions per load were excluded from fitting. On those nine tests, pooled MAE, RMSE, and maximum absolute error were 0.529, 0.707, and 1.815 mA, respectively. The low within-load dispersion supports repeatability under this bench setup, while the persistent load-dependent residuals show that the global line is not a complete error model.

The evidence validates the frozen equation only for this hardware, the three resistive loads, and their frozen load-specific DAC/current ranges. The tests do not represent the electrode–skin interface, therapeutic pulses, or clinical outcomes; a documented RC/electrode model remains required. Conservative range rejection, compliance awareness, and hardware current limits remain necessary in any embedded implementation.

References

[1] M. A. Khan, H. Fares, H. Ghayvat, I. C. Brunner, S. Puthusserypady, B. Razavi, M. Lansberg, A. Poon, and K. J. Meador, "A systematic review on functional electrical stimulation based rehabilitation systems for upper limb post-stroke recovery," *Frontiers in Neurology*, vol. 14, 2023.

[2] P. H. Peckham and J. S. Knutson, "Functional electrical stimulation for neuromuscular applications," *Annual Review of Biomedical Engineering*, vol. 7, no. 1, pp. 327–360, Aug. 2005.

[3] J. L. V. Luna, M. Krenn, J. A. Cortés, and W. Mayr, "Comparison of current and voltage control techniques for neuromuscular electrical stimulation in the anterior thigh," *Biomedical Engineering / Biomedizinische Technik*, Jan. 2013.

[4] R. G. Barelli, "Stimgrasp: estimulador neuromuscular para restauração de padrões de preensão," Master's thesis, Centro Universitário FEI, 2017.

[5] R. A. Matos and M. C. F. Castro, "Power amplifier topologies for functional electrical stimulation: A comparison between howland current source and wilson current mirror," in *XXIX Brazilian Congress on Biomedical Engineering*, ser. IFMBE Proceedings, vol. 125. Springer Nature Switzerland, 2025, pp. 561–571.

[6] A. J. and I. J. Raglend, "Analytical study of howland current source model," in *2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*. IEEE, 2012, pp. 314–318.

[7] W. G. Besio and A. Prasad, "Analysis of skin-electrode impedance using concentric ring electrode," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2006, pp. 6414–6417.

Hybrid Deep Feature Extraction and Machine Learning for Multiclass Dementia Classification Using EEG Wavelet Images

Anna Beatriz Brito de Souza Almeida
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
anna.britoalmeida@ufpe.br

Théo Saldanha Teixeira
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
theo.saldanhateixeira@ufpe.br

Ana Clara Bezerra do Nascimento
dept.of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
clara.bezerran@ufpe.br

Raylla Sandelly das Montanhas Farias
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
raylla.farias@ufpe.br

Ana Paula S. de Oliveira
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
paula.silvaoliveira@ufpe.br

Clarisse Lins de Lima
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
clarisse.lins@ufpe.br

Maíra Araújo de Santana
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
maira.araujosantana@ufpe.br

Wellington P. dos Santos
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

Giselle M. M. Moreno
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
giselle.moreno@ufpe.br

Juliana C. Gomes Cassemiro
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
juliana.cgomes@ufpe.br

***Abstract*—Dementia represents a major global health challenge, with increasing prevalence driven by population aging and a growing demand for accessible tools for early diagnosis. Electroencephalography (EEG) has emerged as a promising low-cost and non-invasive alternative for cognitive assessment, particularly when combined with machine learning techniques. This study investigates the classification of Healthy Controls (HC), Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD), and Frontotemporal Dementia (FTD) using EEG signals from the Brazilian GeronGnosis database. EEG recordings from 50 older adults were segmented into 2-second windows and transformed into time-frequency representations through Fast Wavelet Transform (FWT) using Daubechies-8 filters. Feature extraction was performed using transfer learning with EfficientNet-B0, generating 1,281-dimensional embeddings (Base 1). Feature selection based on Particle Swarm Optimization (PSO) reduced the dimensionality to 324 attributes (Base 2). Multiple machine learning classifiers, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naive Bayes, and Bayes Network, were evaluated using 10-fold cross-validation with 30 repetitions. In the training phase, RF with 500 trees achieved the highest accuracy on Base 1 (83.43%), while MLP reached 83.57% on Base 2. Independent test-set evaluation showed that MLP achieved the highest accuracy (73.75%) after PSO-based feature selection, whereas RF-500 achieved 68.19% accuracy with substantially lower computational cost, providing a favorable trade-off between predictive performance and efficiency. The proposed approach successfully discriminated among the four cognitive conditions and demonstrated the potential of combining EEG, wavelet-based representations, deep feature extraction, and machine learning for supporting dementia screening in clinical settings. These findings contribute to the development of accessible and scalable decision-support tools tailored to the Brazilian healthcare context.**

***Index Terms*—Electroencephalography, Dementia, Machine Learning, Wavelet Transform**

I. Introduction

Population aging has emerged as one of the most significant demographic phenomena of the 21st century, directly contributing to the increasing prevalence of neurodegenerative diseases. In this context, dementia represents a major public health challenge, currently affecting tens of millions of individuals worldwide, with projections indicating a continuous increase in prevalence by 2050 [1]. In Brazil, millions of people are estimated to be living with dementia, with Alzheimer's disease (AD) accounting for approximately 60–70% of diagnosed cases [2]. Beyond its high prevalence, dementia is associated with progressive loss of functional independence, increased

healthcare costs, and substantial social and economic burdens on patients, families, and healthcare systems.

One of the main clinical challenges associated with dementia syndromes is the early identification and accurate differentiation between healthy aging, Mild Cognitive Impairment (MCI), Alzheimer's disease (AD), and Frontotemporal Dementia (FTD) [3]. MCI represents a critical transitional stage within the continuum of cognitive decline and is associated with an elevated risk of progression to dementia [3], [4]. However, the symptomatic overlap among these conditions often hinders diagnosis based solely on clinical assessment. Although gold-standard diagnostic methods, such as Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), and cerebrospinal fluid (CSF) biomarker analysis, provide high diagnostic sensitivity, their widespread implementation is limited by high costs, invasiveness, and restricted availability in specialized healthcare centers [5], [6].

In response to these limitations, Electroencephalography (EEG) has emerged as a promising alternative for screening and diagnostic support due to its low cost, non-invasive nature, portability, and high temporal resolution [7]. Neurodegenerative processes are reflected in brain electrical activity, producing characteristic spectral alterations associated with signal slowing and reduced complexity of cortical oscillations [8]. Previous studies have demonstrated that abnormal increases in Delta and Theta activity, together with reductions in Alpha power and disruptions in higher-frequency bands, may serve as relevant neurophysiological biomarkers for identifying different stages of cognitive impairment and distinguishing dementia subtypes [9].

The growing availability of Artificial Intelligence (AI) and Machine Learning techniques has significantly expanded the diagnostic potential of EEG. Computational approaches combined with advanced signal processing methods, such as the Wavelet Transform, enable the extraction of time-frequency representations capable of revealing neurophysiological patterns that are difficult to identify through conventional visual inspection [10]. Recent studies have reported promising results in the classification of healthy controls, MCI patients, and individuals with different forms of dementia, achieving high levels of accuracy through the integration of EEG and supervised learning algorithms [4], [11], [12]. Despite these advances, the literature still lacks studies conducted using datasets representative of the Brazilian population, limiting the direct applicability of existing models within the national healthcare context.

In this context, the present study aims to evaluate the performance of supervised machine learning algorithms for the multiclass classification of EEG signals represented in the time-frequency domain using the Wavelet Transform, based on data from the Brazilian GeronGnosis database [9]. Specifically, the proposed framework employs a pre-trained EfficientNet architecture to automatically extract deep features from wavelet-based image representations of EEG signals. These features are then used to train and compare the performance of classical machine learning classifiers, including Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP), among others. Furthermore, Particle Swarm Optimization (PSO) is investigated as a feature selection strategy to reduce dimensionality and identify the most informative feature subset. The proposed approach is evaluated for discriminating healthy controls, patients with Mild Cognitive Impairment, Alzheimer's disease, and Frontotemporal Dementia.

II. Methodology

A. Proposed Method

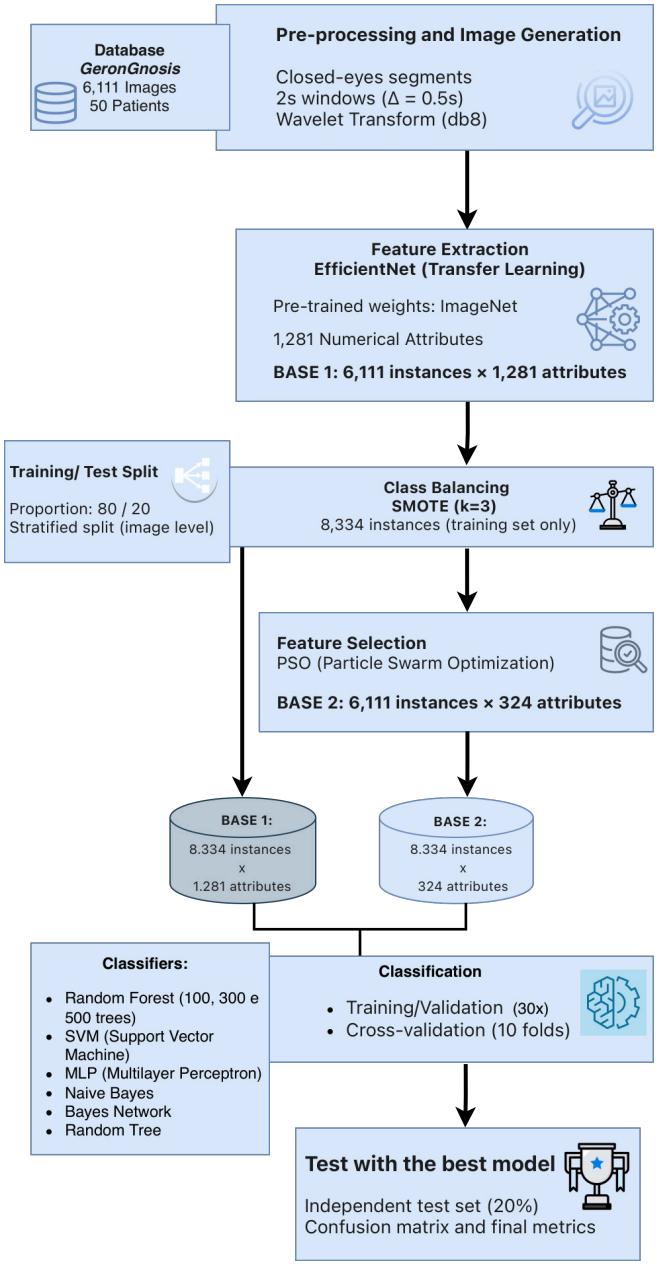


Fig. 1. Methodological workflow: from EEG acquisition to classifier evaluation.

This study proposes a deep learning and machine learning framework for the classification of EEG signals obtained

from the GeronGnosis database. The proposed methodology consists of converting EEG signals into image representations through the Wavelet Transform, followed by feature extraction using the EfficientNet convolutional neural network architecture. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, while Particle Swarm Optimization (PSO) was employed for feature selection and dimensionality reduction.

Several classification algorithms were evaluated, including Random Forest, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naive Bayes, Bayes Network, and Random Tree. The best-performing model was subsequently selected and assessed on an independent test set comprising 20% of the available data in order to obtain the final performance metrics.

The complete methodological workflow, illustrating the EEG signal processing and classification stages, is presented in Figure 1. The following subsections provide a detailed description of each methodological step adopted in this study.

B. GeronGnosis Database

The study used the GeronGnosis database [9], which comprises electroencephalography (EEG) recordings from 50 older adults recruited at the Cognitive Neurology outpatient clinic of the Hospital das Clínicas of the Federal University of Pernambuco (UFPE-EBSERH). The sample was distributed into four diagnostic groups: Alzheimer's Disease (n = 15), Frontotemporal Dementia (n = 5), Mild Cognitive Impairment (n = 10), and Healthy Controls (n = 20). Regarding sex distribution, the sample consisted of 40 women and 10 men, with a mean age of 69.92 ± 7.43 years. Cognitive assessment was performed using the Mini-Mental State Examination (MMSE), yielding an overall mean score of 18.48 ± 5.36 points. The mean scores for each group were: MCI (22.20 ± 4.20), AD (15.84 ± 5.30), FTD (19.50 ± 2.50), and HC (27.10 ± 2.90).

EEG signals were acquired using a Brain Wave III system (Neurovirtual), with 19 active electrodes positioned according to the International 10–20 System [13] (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2), in addition to auricular references (A1 and A2) and a frontal reference electrode. Electrode impedance was maintained below 5 kΩ, with a sampling frequency of 300 Hz and a resolution of 7 μV/mm. The acquisition protocol consisted of 2 minutes of baseline recording, followed by 5 minutes of resting-state EEG with eyes closed and 5 minutes with eyes open.

All procedures were approved by the Research Ethics Committee of UFPE (CAAE 71096223.3.0000.5208), and all participants or their legal representatives provided written informed consent prior to data acquisition.

C. Pre-processing and Image Generation

EEG signals from the GeronGnosis database were preprocessed using the open-source software GNU Octave [9]. Initially, the original recordings in .edf format were converted to .csv files, and only the eyes-closed segments were selected in order to minimize visual interference and standardize the resting-state condition. Temporal segmentation was then performed using 2-second windows with a 0.5-second overlap.

Subsequently, EEG signals were converted into image representations through the Fast Wavelet Transform (FWT), employing Mallat's algorithm with 10 filter banks from the Daubechies 8 (db8) wavelet family. The resulting wavelet coefficients were plotted to generate time-frequency images. Prior to transformation, all EEG channels were concatenated, allowing a single image to be generated for each signal window. This procedure produced image representations capable of capturing the time-frequency characteristics of brain activity [10]. Such an approach enables the application of convolutional neural network (CNN)-based architectures for automatic feature extraction [14].

After image generation, all samples were cataloged in an Attribute-Relation File Format (.arff) file containing the image paths and their corresponding diagnostic labels (Healthy Control, MCI, AD, and FTD). This file was subsequently used as input for the Weka platform, where the feature extraction stage was performed.

For feature extraction (Dataset 1), the EfficientNet-B0 convolutional neural network architecture was employed through the DI4jMlpFilter available in Weka, which enables deep learning-based feature extraction using the DL4J library. EfficientNet belongs to a family of CNN models known for their computational efficiency and scalability, relying on a compound scaling strategy that optimally balances network depth, width, and image resolution [15]. Transfer learning was adopted by utilizing weights pre-trained on the ImageNet dataset, one of the largest publicly available image repositories, containing more than 14 million images distributed across approximately 21,000 categories [16]. This strategy allows the model to leverage visual representations learned from large-scale datasets and adapt them to the specific domain of wavelet images derived from EEG signals.

Feature embeddings were extracted through a Dense Layer positioned after a Max Pooling operation, which preserves the most relevant patterns while reducing spatial dimensionality. This procedure resulted in Dataset 1, comprising 6,111 instances described by 1,281 numerical attributes.

Due to the high dimensionality of Dataset 1, a feature selection stage was performed using Particle Swarm Optimization, generating Dataset 2. PSO is an evolutionary optimization metaheuristic inspired by the collective behavior of bird flocks and fish schools [17]. The algorithm searches for an optimal subset of features by maximizing classifier performance while minimizing redundant and noisy information. As a result, dimensionality was reduced from 1,281 to 324 selected attributes while preserving all 6,111 instances, producing the dataset subsequently used for classification model development.

D. Training/Test Split and Class Balancing

The dataset was partitioned into training and testing subsets using the Resample method [18], following a stratified ratio of 80% for training and 20% for testing. The split was performed at the image level rather than at the subject

level. This approach represents a limitation of the present study, since different image windows derived from the same participant may be present in both the training and testing sets. Consequently, subject-specific patterns may be shared between the two subsets, potentially leading to information leakage and an overestimation of the model's generalization performance. Future studies should therefore adopt subject-level partitioning, ensuring that all image windows from a given participant are assigned exclusively to either the training or testing set.

As described in the previous section, Dataset 1 (features extracted using EfficientNet-B0) resulted in 4,888 instances for training and 1,223 instances for testing, while preserving the original dimensionality of 1,281 attributes. Similarly, Dataset 2 (after feature selection through PSO) contained 4,888 training instances and 1,223 testing instances, with the dimensionality reduced to 324 attributes.

Due to the natural imbalance among diagnostic classes, particularly the Frontotemporal Dementia group, which included the smallest number of participants, the SMOTE was applied exclusively to the training set [19]. Restricting oversampling to the training data prevents data leakage and ensures that model performance is evaluated on real, previously unseen samples during testing.

SMOTE generates synthetic examples for minority classes by interpolating between neighboring instances in the feature space [20]. In this study, a neighborhood size of (k = 3) was adopted for synthetic sample generation. After balancing, the training sets of both Dataset 1 and Dataset 2 reached a total of 8,334 instances, ensuring an approximately uniform representation of all diagnostic classes (Healthy Control, MCI, AD, and FTD) during classifier training.

E. Classifiers and Experimental Setup

To evaluate the performance of machine learning models for the multiclass discrimination of Healthy Controls, Mild Cognitive Impairment, Alzheimer's Disease, and Frontotemporal Dementia based on wavelet images derived from EEG signals, several classifiers implemented in Weka version 3.8.6 were investigated [21]. The evaluated algorithms comprised probabilistic methods (Naive Bayes and Bayes Network), tree-based approaches (Random Tree and Random Forest), artificial neural networks (Multilayer Perceptron), and SVM. For the Random Forest classifier, ensembles consisting of 100, 300, and 500 trees were evaluated. The Multilayer Perceptron employed a single hidden layer with 50 neurons, whereas the SVM classifier was assessed using linear, second-degree polynomial, and Radial Basis Function (RBF) kernels.

Experiments were conducted independently on both generated datasets (Dataset 1 and Dataset 2). Model training and selection were performed using 10-fold cross-validation with 30 repetitions on the balanced training set, providing a robust estimation of classifier performance while reducing the influence of random variations associated with a single data partition.

F. Evaluation Metrics

Upon completion of the training process for both datasets, the results were individually exported to .csv files for subsequent analysis. To assess the performance of each classifier across the two experimental datasets, the following evaluation metrics were considered: Accuracy, Cohen's Kappa coefficient, Sensitivity (Recall), Specificity, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), as described in Table I. In the formulas presented in the Table I, TP denotes True Positives, TN True Negatives, FP False Positives, FN False Negatives, ρ_o the observed agreement, and ρ_e the agreement expected by chance.

TABLE I
PERFORMANCE EVALUATION METRICS OF THE CLASSIFIERS.

Metric	Description	Formula
Accuracy	Proportion of correctly classified instances relative to the total	$\frac{TP + TN}{TP + TN + FP + FN}$
Kappa Index	Agreement measure that accounts for the probability of chance agreement	$\frac{P_o - P_e}{1 - P_e}$
Sensitivity	Model's ability to correctly identify positive cases	$\frac{TP}{TP + FN}$
Specificity	Model's ability to correctly identify negative cases	$\frac{TN}{TN + FP}$
AUC-ROC	Discriminative ability across different thresholds (0 to 1; closer to 1 is excellent)	N/A

G. Best Classifier Selection and Final Testing

Following the evaluation of all classifiers during the training phase, the best-performing model was selected based on its mean Accuracy and Cohen's Kappa coefficient. These metrics were chosen to jointly assess predictive performance and the level of agreement beyond chance, particularly in the context of multiclass classification.

The selected classifier was subsequently applied to the independent test set, corresponding to 20% of the available data. Importantly, the test set retained the original class distribution and contained exclusively original samples; SMOTE was applied only to the training data, ensuring that no synthetic samples were introduced into the final evaluation. Model evaluation was performed using the Explorer interface of Weka, generating the confusion matrix as well as all performance metrics previously described. This final assessment enabled the evaluation of model performance on data not used during model training while preserving the naturally imbalanced class distribution.

III. Results

The experimental results are presented as the mean and standard deviation of the performance metrics described in the metrics section. Table II presents the performance results for the four-class scenario Control, MCI, AD, and FTD using the complete base (Dataset 1), composed of the features extracted directly by EfficientNet without feature selection. The decision-tree-based models, the SVM classifiers, and the MLP stood out, achieving accuracy values equal to or greater than 80% (with the exception of SVM-RBF, which reached 71.93% accuracy). Among all the evaluated models, Random Forest 500 (RF-500) obtained the best classification performance during training, reaching an accuracy of 83.43%, a result that was consistently reflected across all other metrics.

Figure 2 presents the boxplot of the accuracy values obtained by the classifiers evaluated on Dataset 1. In agreement with the results presented in Table II, the Random Forest and SVM with a degree-2 polynomial kernel classifiers presented the highest mean accuracies. Furthermore, both demonstrated low dispersion of results and the absence of outliers, indicating greater stability in performance. It is also observed that the variation ranges of these classifiers overlap, suggesting similar performances when considering the variability of the obtained results.

The RF-300 and RF-500 models were then applied to the independent test set, corresponding to 20% of the data, through the Weka Explorer interface, with the aim of validating the performance obtained in the training stage and supporting the choice of the final classifier. The results confirmed the superior performance of the RF-500 model, which reached an accuracy of 67.70% on the test set. The confusion matrix obtained for this model (Table III) allowed a detailed analysis of the classification patterns. It is observed that the healthy control group presented the highest number of correct classifications, with 416 samples correctly identified, although it showed confusion mainly with the AD (49 samples) and MCI (45 samples) groups. The MCI group also presented good identification capacity, with 161 correct classifications and a lower number of errors distributed among the other classes. For the AD class, 165 correct classifications were obtained; however, a high confusion with the healthy control group was observed, with 128 AD samples classified as control, representing the model's main error pattern. Finally, the FTD group presented high discrimination capacity, with 86 correct classifications and reduced confusion with the other classes (31 classified as control, 3 as MCI, and 7 as AD).

After feature selection via PSO (standard version with 20 iterations), the number of features was substantially reduced. In the multiclass classification (Control, MCI, AD, and FTD), the initial set of 1281 features was reduced to 324. From the analysis of the multiclass classification group's performance, after PSO-based feature selection, through the results in Table IV, it can be observed that the Multilayer Perceptron (MLP), SVM K2, followed by the Random Forest with 500 trees (RF-500) achieved the best training performances, reaching accuracies of 83.57%, 83.13%, and 82.78%, respectively, in addition to obtaining considerable values in most of the other evaluation metrics. Figure 3 displays the box plots of accuracy for all tested models. It is observed that the RF-300, RF-500, SMO K2, and MLP models presented the highest median accuracy values, all greater than approximately 80%, in addition to low dispersion of results across executions. Among them, RF-500 and MLP presented the highest medians, indicating superior performance. In contrast, classifiers such as Naive Bayes, BayesNet, and Random Tree presented lower accuracies, evidencing lower discriminative capacity for the analyzed database. In general, the low amplitude of the boxes and the absence of relevant outliers indicate good stability of the evaluated models.

During the test phase, the MLP continued to demonstrate

TABLE II
Classification result: control, MCI, AD, and FTD for Base 1 (EfficientNet).

Classifiers	Accuracy (%)	Kappa Index	Sensitivity	Specificity	AUC-ROC
Naive Bayes	0.21 ± 0.16	0.06 ± 0.27	0.00 ± 0.06	0.99 ± 0.07	0.50 ± 0.00
Bayes Net	53.54 ± 1.71	0.38 ± 0.02	0.64 ± 0.03	0.70 ± 0.02	0.74 ± 0.02
Random Tree	57.75 ± 1.75	0.44 ± 0.02	0.45 ± 0.04	0.83 ± 0.01	0.64 ± 0.02
RF 100	81.98 ± 1.39	0.76 ± 0.02	0.74 ± 0.03	0.90 ± 0.01	0.92 ± 0.01
RF 300	83.15 ± 1.32	0.78 ± 0.02	0.76 ± 0.03	0.91 ± 0.01	0.93 ± 0.01
RF 500	**83.43 ± 1.31**	**0.78 ± 0.02**	**0.76 ± 0.03**	**0.91 ± 0.01**	**0.93 ± 0.01**
SVM K1	78.30 ± 1.30	0.71 ± 0.02	0.73 ± 0.03	0.87 ± 0.01	0.85 ± 0.01
SVM K2	82.36 ± 1.27	0.76 ± 0.02	0.76 ± 0.03	0.89 ± 0.01	0.87 ± 0.01
SVM RBF	71.93 ± 1.48	0.63 ± 0.02	0.71 ± 0.03	0.83 ± 0.02	0.82 ± 0.01
MLP	80.80 ± 2.12	0.74 ± 0.03	0.71 ± 0.08	0.90 ± 0.03	0.90 ± 0.01

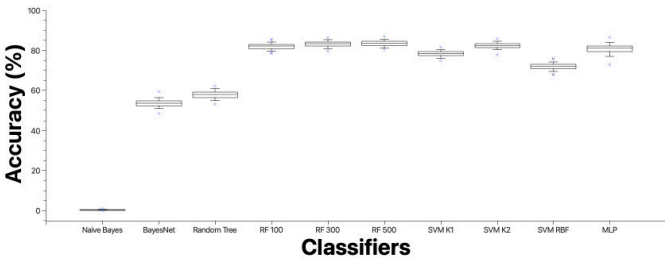


Fig. 2. Box Plot of the classification: control, MCI, AD, and FTD for Base 1 (EfficientNet).

TABLE III
CONFUSION MATRIX: CONTROL, MCI, AD, AND FTD FOR BASE 1 (EFFICIENTNET).

Expected Class	Predicted Class			
	Healthy Control	MCI	AD	FTD
Healthy Control	416	45	49	19
MCI	37	161	18	16
AD	128	24	165	18
FTD	31	3	7	86

the best generalization performance, reaching an accuracy of 73.75% and an AUC-ROC of 0.62. Although the MLP presented the highest test accuracy (73.75%), its high computational cost considerably higher than that of RF-500 during the training phase compromises its viability in clinical applications with limited resources. RF-500, in turn, reached 68.19% accuracy with a significantly lower computational cost, being selected as the reference classifier for Dataset 2 for offering the best balance between predictive performance and computational efficiency. Using the selected classifier, the confusion matrix presented in Table V was obtained. The confusion matrix obtained for reduced dataset, composed of the features selected by PSO, evidences good classifier performance on the independent test set. The healthy control group presented the highest number of correct classifications (417 samples), followed by the MCI (163), AD (166), and FTD (88) groups. Despite the good overall performance, the main source of error was the confusion between individuals with AD and healthy controls, with 124 AD samples classified as control. On the other hand, the FTD class presented greater discrimination capacity, with low confusion relative to the other classes, indicating that the selected features were more effective in distinguishing this group.

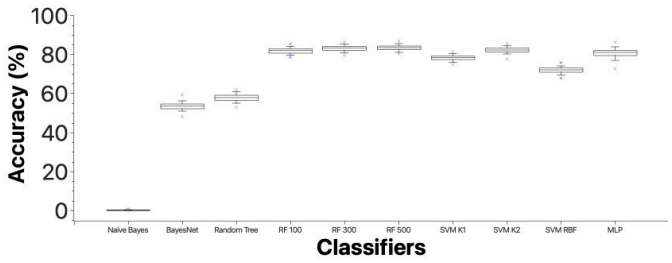


Fig. 3. Box Plot of the classification: control, MCI, AD, and FTD for Base 2 (EfficientNet with PSO).

IV. DISCUSSION

The results obtained demonstrate that the approach based on time-frequency representation of EEG signals, combined with

TABLE IV
CLASSIFICATION RESULT: CONTROL, MCI, AD, AND FTD FOR BASE 2 (EFFICIENTNET WITH PSO).

Classifiers	Accuracy (%)	Kappa Index	Sensitivity	Specificity	AUC-ROC
Naive Bayes	3.08 ± 0.54	0.20 ± 0.10	0.39 ± 0.12	0.79 ± 0.07	0.54 ± 0.01
Bayes Net	53.40 ± 1.70	0.38 ± 0.02	0.67 ± 0.03	0.68 ± 0.02	0.74 ± 0.02
Random Tree	57.36 ± 1.80	0.43 ± 0.02	0.45 ± 0.04	0.83 ± 0.02	0.64 ± 0.02
RF 100	81.26 ± 1.29	0.75 ± 0.02	0.73 ± 0.03	0.90 ± 0.01	0.91 ± 0.01
RF 300	82.50 ± 1.33	0.77 ± 0.02	0.75 ± 0.03	0.91 ± 0.01	0.92 ± 0.01
RF 500	82.78 ± 1.37	0.77 ± 0.02	0.76 ± 0.03	0.91 ± 0.01	0.92 ± 0.01
SVM K1	71.15 ± 1.52	0.62 ± 0.02	0.67 ± 0.03	0.84 ± 0.02	0.82 ± 0.01
SVM K2	83.13 ± 1.27	0.78 ± 0.02	0.74 ± 0.03	0.91 ± 0.01	0.87 ± 0.01
SMO RBF	61.84 ± 1.52	0.49 ± 0.02	0.62 ± 0.03	0.79 ± 0.02	0.76 ± 0.02
MLP	**83.57 ± 1.27**	**0.78 ± 0.02**	**0.73 ± 0.03**	**0.92 ± 0.01**	**0.91 ± 0.01**

TABLE V
CONFUSION MATRIX: CONTROL, MCI, AD, AND FTD FOR BASE 2 (EFFICIENTNET WITH PSO).

Expected Class	Predicted Class			
	Healthy Control	MCI	AD	FTD
Healthy Control	417	42	45	25
MCI	41	163	15	13
AD	124	25	166	20
FTD	29	3	7	88

feature extraction by pre-trained deep convolutional neural networks (EfficientNet) and classification by classical machine learning, is capable of discriminating between four cognitive conditions: Healthy, MCI, AD, and FTD, with competitive performance compared to the literature [1], [3], [22]. RF-500 emerged as the reference classifier in both bases, achieving 67.70% accuracy in Base 1 and 68.19% in Base 2 on the independent test set, with Kappa indices of 0.53 and 0.54, respectively. These values indicate moderate agreement with the actual diagnoses, substantially outperforming simpler models such as Naive Bayes and BayesNet, and demonstrating consistent robustness across the 30 cross-validation repetitions [18]. The performance of Random Forest is consistent with

the findings of Oliveira et al. [1], who reported an accuracy of 98% with RF-500 on the same GeronGnosis database using explicitly extracted EEG signal features. The substantially higher accuracy reported in their study indicates that, for this dataset, explicit signal features provided a more effective representation of the discriminative EEG information than the image embeddings adopted in the present approach. This result also highlights a limitation of the Transfer Learning strategy based on EfficientNet-B0, which was originally pretrained on ImageNet [16]. Although such pretrained networks provide powerful generic feature extractors, the domain mismatch between natural images and EEG time-frequency representations may limit their ability to capture specific spectral and temporal patterns associated with cognitive impairment. These findings suggest that domain-specific EEG features or deep learning models specifically trained or fine-tuned for EEG representations should be further investigated.

In addition, the performance drop observed between the training phase (83.43% in Base 1; 82.78% in Base 2) and the test set (67.70% and 68.19%, respectively) suggests some degree of overfitting to the characteristics of the training data. Despite this reduction, test-set performance remained above the reference threshold for moderate agreement ($\kappa > 0.40$), indicating that the extracted features retained relevant discriminative information beyond the training set [18]. Nevertheless, these results reinforce the need to improve the generalization capability of the proposed approach, particularly through more domain-specific feature representations and validation strategies that better assess performance on previously unseen subjects.

The comparison between Base 1 (without PSO) and Base 2 (with PSO) reveals a positive, albeit modest, impact of feature selection. The reduction from 1,281 to 324 features promoted by the PSO algorithm [9] did not degrade training performance and resulted in a slight improvement in generalization to the test set (67.70% vs. 68.19%), in addition to a substantial reduction in computational cost. These results indicate that PSO was effective in preserving the most informative features, eliminating redundancies without sacrificing performance, which is particularly relevant for applications in systems with limited computational resources, such as the SUS [10].

The per-class analysis reveals that AD was consistently the condition with the highest error rate in both bases (128 and 124 instances misclassified as Healthy, respectively). This finding has direct clinical relevance: in the context of population screening, the incorrect classification of an AD patient as healthy represents a high-impact false negative, as it delays diagnosis and the start of treatment [11]. This limitation is consistent with the literature, which points to the overlap of electrophysiological patterns between AD and normal aging as one of the main diagnostic challenges [3], [12]. In this context, the spectral slowing characteristic of AD marked by the increase in delta and theta bands and the reduction of alpha activity in posterior regions [12] may overlap with normal aging patterns, hindering discrimination in the Wavelet representations.

MCI, by its transitional nature between healthy aging and established dementia [23], also presented confusions distributed with both adjacent classes in the two bases, which is neurologically coherent. It is estimated that 50% to 60% of individuals with MCI progress to AD [22], [23], which implies that part of the sample may already present electrophysiological signatures consistent with early neurodegeneration, while others retain cortical activity profiles closer to healthy aging [3].

FTD demonstrated the best relative performance in both bases, with the highest AUC-ROC values among the pathological classes. This result suggests that the Wavelet representations captured spectral and temporal patterns characteristic of this condition, possibly associated with the predominant impairment of the frontal and temporal regions. Studies indicate that disturbances in the beta and gamma bands provide topographic markers of cortical synchronization useful in discriminating FTD variants [1], which may have contributed to the better discriminative performance of this class. EEG-based FTD classification is also corroborated by Lal et al. [4], who achieved accuracies of 91% to 93% in discriminating between AD, FTD, and healthy controls using SVD entropy associated with the KNN classifier.

From the standpoint of clinical applicability, the proposed approach presents favorable characteristics for integration into diagnostic support systems. The use of EEG as an input modality is non-invasive, low-cost, and widely available in healthcare units, including the Brazilian public system [10], [22]. Furthermore, the use of the GeronGnosis base [9], composed of data from patients in Northeastern Brazil, represents an advance in the validation of computational tools for the national demographic and socioeconomic context, reducing exclusive dependence on international bases.

Among the limitations of the study, the reduced size of the GeronGnosis base stands out (50 participants, only 5 with FTD) [9], which imposes restrictions on the generalization of the models to broader clinical populations. Class imbalance, although mitigated by SMOTE [20] exclusively in the training set, may still influence performance on minority classes in the test set [2]. Additionally, the train-test split was performed at the image level, without separation by subject, which may allow windows from the same patient to simultaneously compose the training and test sets, possibly inflating the obtained metrics and limiting the real estimate of the model's generalization capacity.

In addition, the absence of interpretability analysis of the features extracted by EfficientNet prevents the direct association between regions of the Wavelet images and electrophysiological patterns specific to each condition, limiting the clinical transparency of the model. Future work should prioritize the expansion of the database and external validation in independent populations [1], [22].

V. Conclusion

This work proposed a non-invasive, low-cost computational tool to support the diagnosis of dementias, focusing

on the screening and early differentiation between healthy older adults, individuals with Mild Cognitive Impairment, Alzheimer's Disease, and Frontotemporal Dementia. The central motivation lay in the need to validate accessible and scalable solutions capable of capturing early neurophysiological alterations from the electroencephalogram, a modality compatible with the operational reality of the Brazilian Unified Health System.

To this end, an approach was adopted based on the conversion of EEG signals from the Brazilian GeronGnosis database into time-frequency image representations through the Wavelet Transform, followed by automatic feature extraction with the EfficientNet metwork, and feature selection by Particle Swarm Optimization. Different classical classifiers were compared in a multiclass scenario.

The results demonstrated that the Random Forest with 500 trees achieved accuracy of 67.70% on Dataset 1, while MLP achieved the highest accuracy observed in Dataset 2, reaching 73.75%. Feature selection reduced the dimensionality from 1,281 to 324 feature without degrading performance, with a slight gain in generalization and a reduction in computational cost. The FTD class presented the best relative discriminative performance, while the confusion between AD and healthy controls constituted the model's main predictive limitation.

As future perspectives, we intend to evaluate additional architectures and classifiers that may improve the discrimination between classes, especially between AD and controls. The incorporation of interpretability methods, such as SHAP analysis, is also recommended, in order to associate the extracted features with electrophysiological patterns specific to each condition, expanding the clinical transparency of the model. Finally, the adoption of a multimodal approach is suggested, integrating the clinical and sociodemographic data available in the database with the EEG signals, as well as expansion and external validation in independent populations, consolidating the potential of this tool as clinical decision support in the context of the SUS.

ACKNOWLEDGMENT

We are grateful to the Brazilian Ministry of Health, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES), Fundação de Amparo a Ciencia e Tecnologia do Estado de Pernambuco - Brazil (FACEPE), and the National Council for Scientific and Technological Development - Brazil (CNPq), for the partial financial support.

REFERENCES

[1] E. Nichols, J. D. Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019," *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022.

[2] C. Kilty, S. Cahill, T. Foley, and S. Fox, "Young onset dementia: Implications for employment and finances," *Dementia*, vol. 22, no. 1, pp. 68–84, 2023.

[3] P. M. Rossini, F. Miraglia, and F. Vecchio, "Early dementia diagnosis, mci-to-dementia risk prediction, and the role of machine learning methods for feature extraction from integrated biomarkers, in particular for eeg signal analysis," *Alzheimer's & Dementia*, vol. 18, no. 12, pp. 2699–2706, 2022.

[4] U. Lal, A. V. Chikkankod, and L. Longo, "A comparative study on feature extraction techniques for the discrimination of frontotemporal dementia and alzheimer's disease with electroencephalography in resting-state adults," *Brain sciences*, vol. 14, no. 4, p. 335, 2024.

[5] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish *et al.*, "Nia-aa research framework: toward a biological definition of alzheimer's disease," *Alzheimer's & dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[6] H. Ali, "Ai in neurodegenerative disease research: Early detection, cognitive decline prediction, and brain imaging biomarker identification," *Int J Eng Technol Res Manag*, vol. 6, no. 10, p. 71, 2022.

[7] C. Carrarini, C. Nardulli, L. Titti, F. Iodice, F. Miraglia, F. Vecchio, and P. M. Rossini, "Neuropsychological and electrophysiological measurements for diagnosis and prediction of dementia: a review on machine learning approach," *Ageing Research Reviews*, vol. 100, p. 102417, 2024.

[8] J. Jeong, "Eeg dynamics in patients with alzheimer's disease," *Clinical neurophysiology*, vol. 115, no. 7, pp. 1490–1505, 2004.

[9] A. P. Silva de Oliveira, C. T. de Lima, G. M. de Souza, A. S. de Oliveira Silva, A. L. Queiroz Ribeiro, A. R. Martinez Accioly, C. M. de Lima Simões, M. E. Souza Belmino Lins, Í. A. de Souza, P. M. de Souza Vasconcelos *et al.*, "Gerongnosis: a database for diagnosing dementias using eeg and machine learning," *Research on Biomedical Engineering*, vol. 42, no. 2, p. 29, 2026.

[10] H. U. Amin, A. S. Malik, R. F. Ahmad, N. Badruddin, N. Kamel, M. Hussain, and W.-T. Chooi, "Feature extraction and classification for eeg signals using wavelet transform and machine learning techniques," *Australasian physical & engineering sciences in medicine*, vol. 38, no. 1, pp. 139–149, 2015.

[11] B. Jiao, R. Li, H. Zhou, K. Qing, H. Liu, H. Pan, Y. Lei, W. Fu, X. Wang, X. Xiao *et al.*, "Neural biomarker diagnosis and prediction to mild cognitive impairment and alzheimer's disease using eeg technology," *Alzheimer's research & therapy*, vol. 15, no. 1, p. 32, 2023.

[12] F. Vecchio, F. Miraglia, F. Alù, M. Menna, E. Judica, M. Cotelli, and P. M. Rossini, "Classification of alzheimer's disease with respect to physiological aging with innovative eeg biomarkers in a machine learning implementation," *Journal of Alzheimer's disease*, vol. 75, no. 4, pp. 1253–1261, 2020.

[13] A. P. Gonçalves, C. E. Silvado, I. D. Meira, J. A. Bragatti, L. O. Caboclo, M. B. Guaranha, P. O. da Conceição, P. A. L. de Oliveira, and T. F. Marinho, "Tradução e adaptação para a língua portuguesa–brasil do glossário revisado dos termos mais comumente usados por eletroencefalografistas clínicos e proposta atualizada do formato do laudo de eeg (ifcn revisão 2017)," *Clinical Neurophysiology Practice*, vol. 7, pp. 78–95, 2022.

[14] J. C. Gomes, M. C. A. Rodrigues, and W. P. dos Santos, "Asteri: Image-based representation of eeg signals for motor imagery classification," *Research on Biomedical Engineering*, vol. 38, no. 2, pp. 661–681, 2022.

[15] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[17] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. ieee, 1995, pp. 1942–1948.

[18] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.

[19] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, no. 6, p. 137, 2024.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[22] Brasil and M. da Saúde, "Relatório nacional sobre a demência: Epidemiologia,(re) conhecimento e projeções futuras," 2024.

[23] R. C. Petersen and S. Negash, "Mild cognitive impairment: an overview," *CNS spectrums*, vol. 13, no. 1, pp. 45–53, 2008.

Hybrid Regularization for Faulty-Electrode Electrical Impedance Tomography in Hemorrhagic Stroke

Yigermal Bassie[1], Fernando S. Moura[2], and André K. Takahata[1]
[1]Federal University of ABC, Avenida dos Estados, 5001, Santo André, SP 09210-580, Brazil
[2]Federal University of ABC, Alameda da Universidade, s/n, São Bernardo do Campo, SP 09606-045, Brazil
ORCID: 0000-0003-2238-1724; 0000-0002-0433-1224; 0000-0002-7701-6452
Email:yigermal.bassie@ufabc.edu.br; fernando.moura@ufabc.edu.br; andre.t@ufabc.edu.br

***Abstract*—Electrical impedance tomography is a portable and non-ionizing imaging modality with potential for bedside monitoring of hemorrhagic stroke. Its practical use is limited by the nonlinear, ill-posed inverse problem and by sensitivity to corrupted electrode measurements. This paper presents a finite-element reconstruction framework that combines the complete electrode model with hybrid robust data fidelity and prior regularization. All voltage data used in this study were numerically generated from a two-dimensional hemorrhagic-stroke conductivity phantom on a fine forward mesh and reconstructed using a separate, coarser inverse mesh. The data-fidelity term blends a robust L_1 component, which limits the influence of large residuals, with a quadratic L_2 component, which remains efficient for small approximately Gaussian errors. The prior term similarly balances smooth Tikhonov-type stabilization with sparsity-promoting edge preservation. The resulting nonlinear problem is solved using a Levenberg-Marquardt iteratively reweighted least-squares method. The simulations with a hemorrhagic inclusion and localized faulty-electrode corruption show that robust data weighting improves reconstruction stability, contrast recovery, and spatial overlap compared with purely quadratic fitting. The results indicate that robust data fidelity is the dominant factor for suppressing high-amplitude corrupted-channel residuals, while the prior term controls the trade-off between smoothness and edge preservation.**

***Index Terms*—Electrical impedance tomography, hemorrhagic stroke, faulty electrodes, hybrid regularization**

I. Introduction

Stroke remains a major global health burden and requires rapid, reliable, and accessible diagnostic and monitoring strategies [1]. Hemorrhagic stroke is particularly time sensitive because intracranial bleeding can progress rapidly and can alter the electrical properties of brain tissue. Electrical impedance tomography (EIT) is a promising modality for bedside monitoring in this setting because it is portable, low cost, non-ionizing, and suitable for repeated measurements. In EIT, low-amplitude alternating currents are injected through boundary electrodes, and the resulting boundary voltages are measured. The internal conductivity distribution is then estimated from these current voltage measurements. Previous studies have demonstrated the potential of EIT for brain data acquisition, stroke-oriented monitoring, and conductivity reconstruction [2]–[4]. Classical EIT studies further show that reconstruction quality depends strongly on the electrode model, tissue-property assumptions, measurement noise, and regularization strategy [5]–[7].

Despite these advantages, EIT reconstruction remains a challenging computational problem because the underlying inverse problem is highly nonlinear and ill-posed. Consequently, even small measurement errors in the boundary voltage data may produce large variations in the reconstructed conductivity distribution [6], [7]. A physically realistic forward model is usually obtained using the complete electrode model (CEM), which accounts for finite electrode size, contact impedance, injected currents, and measured electrode voltages [8], [9]. Although the CEM provides a physically realistic measurement model, the associated inverse problem remains sensitive to modeling errors, contact-impedance variations, electrode misplacement, and corrupted measurements. Achieving stable conductivity reconstructions therefore requires both reliable physical modeling and regularization strategies.

Many EIT reconstruction methods are formulated as least-squares data-fitting problems with quadratic Tikhonov regularization [10]–[12]. Such formulations are computationally efficient and generally yield stable reconstructions when the measurement errors are small and approximately Gaussian. However, the L_2 data-fidelity term grows quadratically with the residual magnitude. As a result, a small number of outliers caused by faulty electrodes, motion artifacts, impulsive noise, or contact-impedance changes can dominate the objective function, producing reconstruction artifacts and degrading image quality [13], [14]. Moreover, purely quadratic regularization tends to oversmooth conductivity distributions, which may blur sharp boundaries and reduce the detectability of localized anomalies.

Robust and sparsity-promoting penalties provide an effective alternative to purely quadratic reconstruction models. Unlike the quadratic growth of the L_2 norm, the L_1 norm increases linearly with the residual magnitude, thereby reducing the influence of large residuals and improving robustness when the measured data contain outliers or corrupted channels [15], [16]. When applied to spatial differences, L_1-type regularization can promote localized conductivity variations and help preserve sharp boundaries [17]. These properties are important in EIT because relevant abnormalities may manifest as spatially localized conductivity changes. Nevertheless, purely L_1 formulations are inherently nonsmooth and may lead to numerical

This work was financed by the São Paulo Research Foundation (FAPESP), process numbers 2023/13729-0, 2022/03243-0, 2013/07559-3, and 2021/11905-0.

difficulties during optimization. To address these challenges, hybrid L_1/L_2 formulations offer a practical compromise by combining the numerical stability of quadratic regularization with the robustness and edge-preserving properties of L_1-type penalties.

In this work, a hybrid L_1/L_2 reconstruction framework is developed for EIT in the presence of localized faulty-channel corruption. The proposed formulation combines robust data fidelity and stabilizing prior regularization in a unified nonlinear optimization problem, which is solved using a Levenberg-Marquardt iteratively reweighted least-squares (LM-IRLS) strategy [18], [19]. The LM component improves numerical stability through adaptive damping, whereas the IRLS component handles the L_1-type terms through adaptive weighted L_2 approximations. This study further investigates how different combinations of quadratic and robust penalty terms affect reconstruction accuracy, robustness to corrupted measurements, and preservation of localized conductivity variations.

The rest of this paper is organized as follows. Section II presents the complete electrode model and finite-element discretization. Section III describes the formulation of hybrid objective functions. Section IV derives the LM-IRLS reconstruction algorithm. Section V describes the simulation protocol. Section VI presents the reconstruction results. Section VII discusses the findings and limitations, and Section VIII concludes the paper.

II. Forward Model and Finite-Element Discretization

A. Complete Electrode Model

Let $\Omega \subset \mathbb{R}^2$ denote the imaging domain, and let $\partial\Omega$ represent its boundary. The boundary contains L electrodes, denoted by $E_1, E_2, \ldots, E_L$, through which electrical currents are injected and boundary voltages are measured. For a positive conductivity distribution $\sigma(\mathbf{x}) > 0$, the electric potential field $\Phi(\mathbf{x})$ inside the domain satisfies the conductivity equation

$$\nabla \cdot (\sigma \nabla \Phi) = 0, \qquad \mathbf{x} \in \Omega, \tag{1}$$

which follows from charge conservation under the quasi-static approximation. This forward problem provides a physically realistic representation of electrode-based measurements by accounting for finite electrode size and electrode-skin contact impedance [8], [9].

In the CEM, each electrode E_ℓ is associated with an electrode voltage Φ_ℓ and a contact impedance $z_\ell > 0$. On the electrode surfaces, the boundary conditions are given by

$$\Phi + z_\ell \sigma \frac{\partial \Phi}{\partial \mathbf{n}} = \Phi_\ell, \qquad \mathbf{x} \in E_\ell, \quad \ell = 1, \ldots, L, \tag{2}$$

$$\int_{E_\ell} \sigma \frac{\partial \Phi}{\partial \mathbf{n}} \, \mathrm{d}S = I_\ell, \qquad \ell = 1, \ldots, L, \tag{3}$$

where $\partial\Phi/\partial\mathbf{n}$ denotes the outward normal derivative of the electric potential, and I_ℓ is the current applied through the ℓth electrode. The first condition accounts for the voltage drop across the electrode-tissue contact impedance, whereas the second condition enforces conservation of the total current through each electrode.

On the insulating boundary gaps between electrodes, no current flow is imposed. This boundary condition is expressed as

$$\sigma \frac{\partial \Phi}{\partial \mathbf{n}} = 0, \qquad \mathbf{x} \in \partial\Omega \setminus \bigcup_{\ell=1}^{L} E_\ell. \tag{4}$$

The injected current vector $\mathbf{I} = [I_1, \ldots, I_L]^T$ satisfies the charge-conservation condition,

$$\sum_{\ell=1}^{L} I_\ell = 0. \tag{5}$$

Since the electric potential and electrode voltages are determined only up to an additive constant, a voltage reference condition is required to ensure uniqueness of the forward solution. In this work, the electrode closest to the origin is selected as the ground reference.

B. Nodal-Conductivity Finite-Element Discretization

The domain Ω is discretized with linear triangular finite elements. Let $\{\varphi_i(\mathbf{x})\}_{i=1}^{N_n}$ denote the nodal basis functions, where N_n is the number of mesh nodes. The electric potential and conductivity are approximated by nodal finite-element expansions as

$$\Phi(\mathbf{x}) \approx \Phi_h(\mathbf{x}) = \sum_{i=1}^{N_n} u_i \varphi_i(\mathbf{x}), \tag{6}$$

and

$$\sigma(\mathbf{x}) \approx \sigma_h(\mathbf{x}) = \sum_{a=1}^{N_n} \sigma_a \psi_a(\mathbf{x}). \tag{7}$$

respectively. Here, u_i and $\varphi_i(\mathbf{x})$ denote the unknown nodal electric potential and associated finite-element basis function at node i, whereas σ_a and $\psi_a(\mathbf{x})$ denote the nodal conductivity, and associated basis function at node a. Hence, $\Phi_h(\mathbf{x})$ and $\sigma_h(\mathbf{x})$ represent the finite-element approximations of the electric potential and conductivity fields. Within each triangular element, both quantities are obtained by interpolated from their nodal values.

After assembling the element matrices and using Eq.(2) and Eq.(3), the discrete forward problem can be written as the block linear system

$$\begin{bmatrix} \mathbf{K}(\boldsymbol{\sigma}) + \mathbf{C} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(p)} \\ \boldsymbol{\Phi}^{(p)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{I}^{(p)} \end{bmatrix}, \tag{8}$$

where $\mathbf{u}^{(p)} \in \mathbb{R}^{N_n}$ is the nodal potential vector, $\boldsymbol{\Phi}^{(p)} \in \mathbb{R}^L$ is the electrode-voltage vector, and $\mathbf{I}^{(p)} \in \mathbb{R}^L$ is the applied current vector for the p-th current-injection pattern. Here, the global admittance matrix $\mathbf{K}(\boldsymbol{\sigma})$ is obtained by assembling the local matrices over all elements, with entries

$$K_{ij}(\boldsymbol{\sigma}) = \int_\Omega \sigma_h(\mathbf{x}) \nabla\varphi_i(\mathbf{x}) \cdot \nabla\varphi_j(\mathbf{x}) \, d\mathbf{x}. \tag{9}$$

The electrode contact-impedance matrix is $\mathbf{C}_{ij} = \sum_{\ell=1}^{L} z_\ell^{-1} \int_{E_\ell} \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) \, dS$. The nodal electrode coupling

matrix is $\mathbf{B}_{i\ell} = -z_\ell^{-1} \int_{E_\ell} \varphi_i(\mathbf{x})\, dS$, and the electrode matrix is diagonal, with $\mathbf{D}_{\ell m} = \delta_{\ell m}|E_\ell|/z_\ell$, where z_ℓ is the contact impedance, $|E_\ell|$ is the length of electrode E_ℓ, and $\delta_{\ell m}$ is the Kronecker delta.

For the p-th current-injection pattern and the q-th voltage-measurement pattern, the predicted differential voltage is obtained from the electrode voltages as

$$v_{pq}(\boldsymbol{\sigma}) = \mathbf{m}_q^T \boldsymbol{\Phi}^{(p)}(\boldsymbol{\sigma}), \tag{10}$$

where $\mathbf{m}_q$ defines the q-th differential voltage measurement. Equivalently, if $\mathbf{C}^d$ denotes the measurement matrix that converts electrode voltages into differential measurements, then

$$\mathbf{v}^{(p)}(\boldsymbol{\sigma}) = \mathbf{C}^d \boldsymbol{\Phi}^{(p)}(\boldsymbol{\sigma}). \tag{11}$$

Stacking the voltage measurements from all P current-injection patterns yields the nonlinear forward map

$$\mathbf{v} = \mathbf{F}(\boldsymbol{\sigma}) = \begin{bmatrix} \mathbf{C}^d \boldsymbol{\Phi}^{(1)}(\boldsymbol{\sigma}) \\ \mathbf{C}^d \boldsymbol{\Phi}^{(2)}(\boldsymbol{\sigma}) \\ \vdots \\ \mathbf{C}^d \boldsymbol{\Phi}^{(P)}(\boldsymbol{\sigma}) \end{bmatrix} \in \mathbb{R}^M, \tag{12}$$

where M is the total number of differential voltage measurements. Although $\mathbf{K}(\boldsymbol{\sigma})$ depends linearly on the nodal conductivity values in Eq.(7), the forward map $\mathbf{F}(\boldsymbol{\sigma})$ is nonlinear because the electrode voltages are obtained by solving the conductivity-dependent CEM system in Eq.(8).

C. Adjoint Sensitivity Matrix

The Jacobian $\mathbf{J} = \partial \mathbf{F}/\partial \boldsymbol{\sigma} \in \mathbb{R}^{M \times N}$ relates small conductivity perturbations to voltage perturbations. Since the conductivity is described in Eq.(7) by nodal basis functions, the element stiffness matrix on a triangular element T_e is

$$K_{ij}^{(e)}(\boldsymbol{\sigma}) = \int_{T_e} \sigma_h(\mathbf{x}) \nabla \varphi_i(\mathbf{x}) \cdot \nabla \varphi_j(\mathbf{x})\, d\mathbf{x}. \tag{13}$$

Therefore, the derivative of the local stiffness matrix with respect to σ_a is

$$\frac{\partial K_{ij}^{(e)}}{\partial \sigma_a} = \begin{cases} \displaystyle\int_{T_e} \psi_a(\mathbf{x}) \nabla \varphi_i(\mathbf{x}) \cdot \nabla \varphi_j(\mathbf{x})\, d\mathbf{x}, & a \in \mathcal{N}_e, \\ 0, & a \notin \mathcal{N}_e, \end{cases} \tag{14}$$

where $\mathcal{N}_e$ is the set of global conductivity nodes connected to element T_e. After finite-element assembly, the corresponding global derivative matrix is

$$\frac{\partial \mathbf{K}}{\partial \sigma_a} = \sum_{e:\, a \in \mathcal{N}_e} \mathbf{P}_e^T \left(\frac{\partial \mathbf{K}^{(e)}}{\partial \sigma_a} \right) \mathbf{P}_e, \tag{15}$$

where $\mathbf{P}_e$ is the element-to-global assembly operator. This expression shows that only elements sharing node a contribute to the derivative with respect to σ_a. For the p-th current-injection pattern, Eq. (8) can be written in compact form as

$$\mathbf{A}(\boldsymbol{\sigma}) \mathbf{x}^{(p)} = \mathbf{b}^{(p)}, \tag{16}$$

where $\mathbf{x}^{(p)}$ denotes the full CEM state vector which contains the nodal potentials and electrode potentials, $\mathbf{b}^{(p)}$ contains the imposed current pattern and reference constraints, and $\mathbf{A}(\boldsymbol{\sigma})$ is the assembled CEM system matrix. For fixed contact impedances and fixed electrode geometry, the dependence on the conductivity vector $\boldsymbol{\sigma}$ enters only through the domain stiffness matrix $\mathbf{K}(\boldsymbol{\sigma})$. Differentiating Eq. (16) with respect to the nodal conductivity parameter σ_a gives

$$\mathbf{A}(\boldsymbol{\sigma}) \frac{\partial \mathbf{x}^{(p)}}{\partial \sigma_a} = -\frac{\partial \mathbf{A}}{\partial \sigma_a} \mathbf{x}^{(p)}. \tag{17}$$

The direct computation of Eq.(17) for every conductivity node would require one linear solve per unknown parameter, which becomes computationally expensive for large meshes.

The adjoint method avoids this cost by transferring the sensitivity calculation from the parameter space to the measurement space. Let $\mathbf{c}_m$ be the measurement vector that extracts the m-th voltage datum from the full state vector, so that $F_m^{(p)}(\boldsymbol{\sigma}) = \mathbf{c}_m^T \mathbf{x}^{(p)}$. The corresponding adjoint field $\boldsymbol{\lambda}_m$ is obtained from

$$\mathbf{A}^T(\boldsymbol{\sigma}) \boldsymbol{\lambda}_m = \mathbf{c}_m. \tag{18}$$

Combining Eqs. (17) and (18), the Jacobian entry associated with measurement m, current pattern p, and conductivity node a is

$$J_{ma}^{(p)} = \frac{\partial F_m^{(p)}}{\partial \sigma_a} = -\boldsymbol{\lambda}_m^T \frac{\partial \mathbf{A}}{\partial \sigma_a} \mathbf{x}^{(p)}. \tag{19}$$

Since $\partial \mathbf{A}/\partial \sigma_a$ is nonzero only in the stiffness block, Eq. (19) can also be evaluated using only the nodal-potential parts of the forward and adjoint solutions. For multiple current patterns, (16)-(19) are evaluated for each forward solution and the resulting rows are stacked to form the full sensitivity matrix. In practice, the matrix inverse is not formed explicitly; instead, the same factorization or iterative solver used for the CEM forward problem can be reused to solve the adjoint systems. This formulation is computationally efficient and is consistent with standard finite-element Jacobian calculations for EIT based on the complete electrode model [20].

III. Hybrid L_1/L_2 Reconstruction Model

The reconstruction of the conductivity distribution in EIT is formulated as a nonlinear inverse problem because the relationship between the conductivity distribution and the measured boundary voltages is nonlinear [7], [8]. The objective is to estimate the unknown nodal conductivity vector $\boldsymbol{\sigma}$ from the measured voltage data by minimizing a hybrid objective functional composed of data-fidelity and prior-regularization terms. The proposed hybrid reconstruction objective is denoted by $\mathcal{J}(\boldsymbol{\sigma})$ and is defined as

$$\begin{aligned} \mathcal{J}(\boldsymbol{\sigma}) = \zeta \|\boldsymbol{r}(\boldsymbol{\sigma})\|_1 + \frac{1-\zeta}{2} \|\boldsymbol{r}(\boldsymbol{\sigma})\|_2^2 \\ + \beta \left[\eta \|\boldsymbol{p}(\boldsymbol{\sigma})\|_1 + \frac{1-\eta}{2} \|\boldsymbol{p}(\boldsymbol{\sigma})\|_2^2 \right], \end{aligned} \tag{20}$$

where $\zeta \in [0,1]$ controls the balance between robust L_1 and quadratic L_2 data fidelity, $\eta \in [0,1]$ controls the balance between sparsity-promoting and smooth prior regularization, and $\beta > 0$ is the global regularization parameter. This formulation

provides flexible control over measurement robustness, edge preservation, and reconstruction stability.

The data residual is defined as

$$\boldsymbol{r}(\boldsymbol{\sigma}) = \mathbf{F}(\boldsymbol{\sigma}) - \boldsymbol{\Phi}_{\mathrm{meas}}, \tag{21}$$

where $\mathbf{F}(\boldsymbol{\sigma})$ denotes the nonlinear forward operator that maps the conductivity distribution to the predicted boundary voltage measurements, and $\boldsymbol{\Phi}_{\mathrm{meas}}$ is the measured electrode-voltage vector over all current-injection and voltage-measurement patterns [8].

The prior residual is defined in the transformed roughness domain as

$$\boldsymbol{p}(\boldsymbol{\sigma}) = \boldsymbol{\Lambda}\mathbf{R}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0), \tag{22}$$

where $\mathbf{R}$ is a discrete roughness operator, $\boldsymbol{\Lambda}$ is a diagonal spatial scaling matrix, and $\boldsymbol{\sigma}_0$ is the reference conductivity distribution. In this work, $\mathbf{R}$ acts as a discrete gradient-like operator on the mesh, while the associated quadratic penalty $\mathbf{R}^T\mathbf{R}$ has a graph-Laplacian-like smoothing effect. Therefore, $\boldsymbol{\Lambda}$ acts on the roughness residuals and controls the local strength of the prior penalty over mesh edges. If $\boldsymbol{\Lambda} = c\mathbf{I}$, where $c > 0$ is a scalar constant and $\mathbf{I}$ is the identity matrix, its effect is equivalent to rescaling β. Thus, a nonuniform $\boldsymbol{\Lambda}$ is required to impose spatially varying regularization.

The proposed hybrid formulation provides a continuous transition between classical least-squares reconstruction and robust sparsity-promoting regularization. When $\zeta = 0$, the data-fidelity term reduces to the conventional quadratic least-squares form, which is efficient under approximately Gaussian noise but can be sensitive to outliers [10]. When $\zeta = 1$, the data-fidelity term becomes purely L_1-based, reducing the influence of large residuals and improving robustness to corrupted measurements and faulty electrode channels [16]. Intermediate values of ζ provide a trade-off between statistical efficiency and robustness.

Similarly, the parameter η controls the structure of the prior regularization. Smaller values of η produce smoother, Tikhonov-like reconstructions, whereas larger values promote sparse deviations in the transformed roughness domain and help preserve localized conductivity changes. Such hybrid formulations are useful in ill-posed inverse problems where the measured data may be corrupted and the prior information may be imperfect [6], [14].

The four limiting configurations evaluated in this work have direct interpretations. The pair $(\zeta, \eta) = (0, 0)$ represents quadratic data fidelity with a smooth prior, whereas $(1, 0)$ represents robust data fidelity with a smooth prior. The configuration $(0, 1)$ combines quadratic data fidelity with a sparsity-promoting prior, and $(1, 1)$ corresponds to the fully hybrid formulation, in which robust data fidelity is combined with sparsity-promoting prior regularization.

IV. LM-IRLS Optimization

The objective function in Eq.(20) is nonlinear because of the CEM forward map and is nonsmooth in the limiting L_1 case. To obtain a stable numerical update, the nonlinear forward map is locally linearized and the L_1 terms are approximated by an iteratively reweighted least-squares (IRLS) scheme. At iteration k, the forward map is linearized as

$$\mathbf{F}(\boldsymbol{\sigma}_k + \boldsymbol{\delta}) \approx \mathbf{F}(\boldsymbol{\sigma}_k) + \mathbf{J}_k\boldsymbol{\delta}, \tag{23}$$

where $\mathbf{J}_k$ is the sensitivity matrix and $\boldsymbol{\delta}$ is the conductivity update. The L_1 penalty is approximated by the differentiable function $\phi_\epsilon(x) = \sqrt{x^2 + \epsilon^2}$, which approaches $|x|$ as $\epsilon \to 0$. This approximation gives the diagonal IRLS weighting matrices as

$$\mathbf{W}_{d,k}^{(1)} = \mathrm{diag}\left(\frac{1}{\sqrt{r_{i,k}^2 + \epsilon^2}}\right), \tag{24}$$

$$\mathbf{W}_{p,k}^{(1)} = \mathrm{diag}\left(\frac{1}{\sqrt{p_{j,k}^2 + \epsilon^2}}\right). \tag{25}$$

These weights reduce the influence of large residuals and therefore improve robustness against outliers, corrupted voltage measurements, and faulty-electrode channels [15].

The hybrid objective combines quadratic and robust components through the parameters ζ and η. The corresponding data and prior weighting matrices are

$$\mathbf{W}_{d,k} = \zeta\mathbf{W}_{d,k}^{(1)} + (1 - \zeta)\mathbf{I}, \tag{26}$$

$$\mathbf{W}_{p,k} = \eta\mathbf{W}_{p,k}^{(1)} + (1 - \eta)\mathbf{I}. \tag{27}$$

Here, $\mathbf{W}_{d,k}$ controls the contribution of the voltage residuals, whereas $\mathbf{W}_{p,k}$ controls the contribution of the transformed roughness residuals.

The Levenberg-Marquardt IRLS update is written as

$$\begin{aligned}\left(\mathbf{J}_k^T\mathbf{W}_{d,k}\mathbf{J}_k + \beta\mathbf{R}^T\boldsymbol{\Lambda}^T\mathbf{W}_{p,k}\boldsymbol{\Lambda}\mathbf{R} + \mu_k\mathbf{I}\right)\boldsymbol{\delta}_k \\ = -\left(\mathbf{J}_k^T\mathbf{W}_{d,k}\boldsymbol{r}_k + \beta\mathbf{R}^T\boldsymbol{\Lambda}^T\mathbf{W}_{p,k}\boldsymbol{p}_k\right).\end{aligned} \tag{28}$$

where $\mu_k > 0$ is the Levenberg-Marquardt damping parameter [21]. Larger values of μ_k improve numerical stability and make the update more similar to a gradient-descent step, whereas smaller values approach the Gauss-Newton direction.

After the search direction is computed, a trial conductivity is estimated and projected onto the admissible set:

$$\boldsymbol{\sigma}_{k+1} = \Pi_{\mathcal{C}}\left(\boldsymbol{\sigma}_k + \alpha_k\boldsymbol{\delta}_k\right), \tag{29}$$

where $\alpha_k \in (0, 1]$ is a step length parameter and $\Pi_{\mathcal{C}}$ denotes componentwise projection onto the admissible conductivity set $\mathcal{C} = \{\boldsymbol{\sigma} : \sigma_{\min} \le \sigma_i \le \sigma_{\max}, \quad i = 1, \ldots, N\}$. This projection enforces physically meaningful conductivity bounds after each trial step [19].

The step acceptance is determined from the ratio between the actual and predicted reductions. Defining $\boldsymbol{s}_k = \boldsymbol{\sigma}_k^{\mathrm{trial}} - \boldsymbol{\sigma}_k$, the actual reduction is

$$\Delta\mathcal{J}_k^{\mathrm{act}} = \mathcal{J}_\epsilon(\boldsymbol{\sigma}_k) - \mathcal{J}_\epsilon(\boldsymbol{\sigma}_k^{\mathrm{trial}}), \tag{30}$$

and the quadratic-model predicted reduction is

$$\Delta\mathcal{J}_k^{\mathrm{pred}} = -\boldsymbol{g}_k^T\boldsymbol{s}_k - \frac{1}{2}\boldsymbol{s}_k^T\mathbf{H}_k\boldsymbol{s}_k, \tag{31}$$

where $\boldsymbol{g}_k = \mathbf{J}_k^T \mathbf{W}_{d,k} \boldsymbol{r}_k + \beta \mathbf{R}^T \mathbf{\Lambda}^T \mathbf{W}_{p,k} \boldsymbol{p}_k$, and $\mathbf{H}_k = \mathbf{J}_k^T \mathbf{W}_{d,k} \mathbf{J}_k + \beta \mathbf{R}^T \mathbf{\Lambda}^T \mathbf{W}_{p,k} \mathbf{\Lambda} \mathbf{R} + \mu_k \mathbf{I}$. The acceptance ratio is then $\rho_k = \frac{\Delta \mathcal{J}_k^{\mathrm{act}}}{\Delta \mathcal{J}_k^{\mathrm{pred}}}$. The trial step is accepted when $\rho_k \geq 0.25$ and rejected otherwise. The damping parameter is decreased when $\rho_k > 0.75$ and increased when $\rho_k < 0.25$.

The local linearized reconstruction matrix is approximated by

$$\mathbf{K}_{\mathrm{rec}} = \left[\mathbf{J}^T \mathbf{W}_d \mathbf{J} + \beta \mathbf{R}^T \mathbf{\Lambda}^T \mathbf{W}_p \mathbf{\Lambda} \mathbf{R} + \mu \mathbf{I} \right]^{-1} \mathbf{J}^T \mathbf{W}_d. \quad (32)$$

This matrix maps small voltage perturbations to the corresponding linearized conductivity updates. Therefore, the magnitude $|\mathbf{K}_{\mathrm{rec}}|$ indicates how strongly each measurement channel influences each reconstructed node. In practice, Eq.(32) should be evaluated by solving the associated linear systems rather than by forming an explicit matrix inverse. The reconstruction matrix is useful for identifying high-leverage measurements and for assessing how faulty-electrode corruption affects the inverse mapping.

V. Simulation Setup and Implementation

A. Conductivity Phantom and Meshes

The numerical experiments were conducted on a normalized circular two-dimensional domain with $L = 32$ boundary electrodes. In the normalized coordinate system used for the phantom, the outer circle has radius 1 and therefore diameter 2 normalized length units. The background conductivity was set to $0.17\ \mathrm{S/m}$ and the hemorrhagic inclusion conductivity was set to $0.67\ \mathrm{S/m}$ [22]. The inclusion was centered at $(0.4, 0.4)$ and had radius 0.20, with the center coordinates and radius expressed in the same normalized length units. To reduce inverse-crime effects, synthetic data were generated on a finer forward mesh with 1941 nodes and 3368 triangular elements, whereas reconstruction was performed on a coarser inverse mesh with 1238 nodes and 2218 elements. The CEM contact impedance was fixed at $z_\ell = 0.01$ for all electrodes [8]. In this two-dimensional implementation, this value represents an effective numerical contact impedance scaled by boundary length.

B. Current-Injection, Measurement, and Faulty-Channel Model

Current injection was performed using a skip-4 protocol. For each injection pattern, a current of amplitude $I_0 = 1\ \mathrm{mA}$ was applied between electrode ℓ and electrode $(\ell + 4) \bmod L$, defined as

$$I_{\ell'} = \begin{cases} +I_0, & \ell' = \ell, \\ -I_0, & \ell' = (\ell + 4) \bmod L, \\ 0, & \text{otherwise}, \end{cases} \quad (33)$$

Separated-pair current-injection and electrode-addressing patterns are well established in EIT [23]. Adjacent voltage-difference measurements were computed using electrode pairs with offsets 1 and 2.

The synthetic observed data were generated as

$$\mathbf{\Phi}_{\mathrm{meas}} = \mathbf{\Phi}_{\mathrm{clean}} + \boldsymbol{\varepsilon}_{\mathrm{bg}} + \boldsymbol{\varepsilon}_{\mathrm{fault}}, \quad (34)$$

where $\mathbf{\Phi}_{\mathrm{clean}}$ is the clean simulated voltage vector, $\boldsymbol{\varepsilon}_{\mathrm{bg}}$ denotes zero-mean additive Gaussian noise with standard deviation 0.001, and $\boldsymbol{\varepsilon}_{\mathrm{fault}}$ denotes the larger localized perturbation associated with the faulty electrode channel. Electrode number 23 (E23) was treated as faulty. A sample was marked faulty when E23 participated either in the current-injection pair or in the voltage-measurement pair. Faulty samples received additional Gaussian perturbations with standard deviation 10 times the background noise and a bias equal to 0.3 times the standard deviation of the clean voltage vector.

C. Regularization Parameters and Evaluation Metrics

The hybrid method was evaluated for $\beta \in \{0.01, 0.1, 1.0\}$ and $(\zeta, \eta) \in \{(0,0), (1,0), (0,1), (1,1)\}$. Reconstruction quality was assessed using RMSE, relative error, Dice coefficient, intersection-over-union, peak-conductivity error, and contrast-to-noise ratio. The root-mean-square error (RMSE) was computed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\sigma_i^{\mathrm{true}} - \sigma_i^{\mathrm{rec}})^2}, \quad (35)$$

where $\boldsymbol{\sigma}^{\mathrm{true}}$ and $\boldsymbol{\sigma}^{\mathrm{rec}}$ denote the true and reconstructed conductivity distributions, respectively. The relative error is given by

$$\mathrm{RelErr} = \frac{\|\boldsymbol{\sigma}^{\mathrm{true}} - \boldsymbol{\sigma}^{\mathrm{rec}}\|_2}{\|\boldsymbol{\sigma}^{\mathrm{true}}\|_2}. \quad (36)$$

The peak conductivity error is computed as

$$\mathrm{PeakErr} = \left| \sigma_{\max}^{\mathrm{true}} - \sigma_{\max}^{\mathrm{rec}} \right|. \quad (37)$$

To evaluate inclusion localization and segmentation accuracy, let Ω_{true} and Ω_{rec} denote the true and reconstructed inclusion regions obtained after thresholding. The Dice coefficient and IoU are defined as

$$\mathrm{Dice} = \frac{2|\Omega_{\mathrm{true}} \cap \Omega_{\mathrm{rec}}|}{|\Omega_{\mathrm{true}}| + |\Omega_{\mathrm{rec}}|}, \qquad \mathrm{IoU} = \frac{|\Omega_{\mathrm{true}} \cap \Omega_{\mathrm{rec}}|}{|\Omega_{\mathrm{true}} \cup \Omega_{\mathrm{rec}}|}. \quad (38)$$

The contrast-to-noise ratio (CNR) in decibels is computed as

$$\mathrm{CNR}_{\mathrm{dB}} = 20 \log_{10} \left(\frac{|\mu_{\mathrm{inc}} - \mu_{\mathrm{bg}}|}{\sigma_{\mathrm{bg}}} \right), \quad (39)$$

where μ_{inc} is the mean reconstructed conductivity inside the inclusion, and μ_{bg} and σ_{bg} are the mean and standard deviation of the reconstructed background conductivity, respectively.

VI. Results

A. Faulty Measurement Data

Fig. 1 shows the effect of localized faulty-electrode corruption on the simulated EIT voltage measurements. Fig. 1(a) compares the clean voltage measurements with the measurements affected by the faulty E23 electrode, whereas Fig. 1(b) presents the corresponding residual. The residual is concentrated in specific measurement-index ranges because each index corresponds to a particular current-injection and voltage-measurement configuration. This behavior indicates that the fault is sparse in the measurement domain but produces

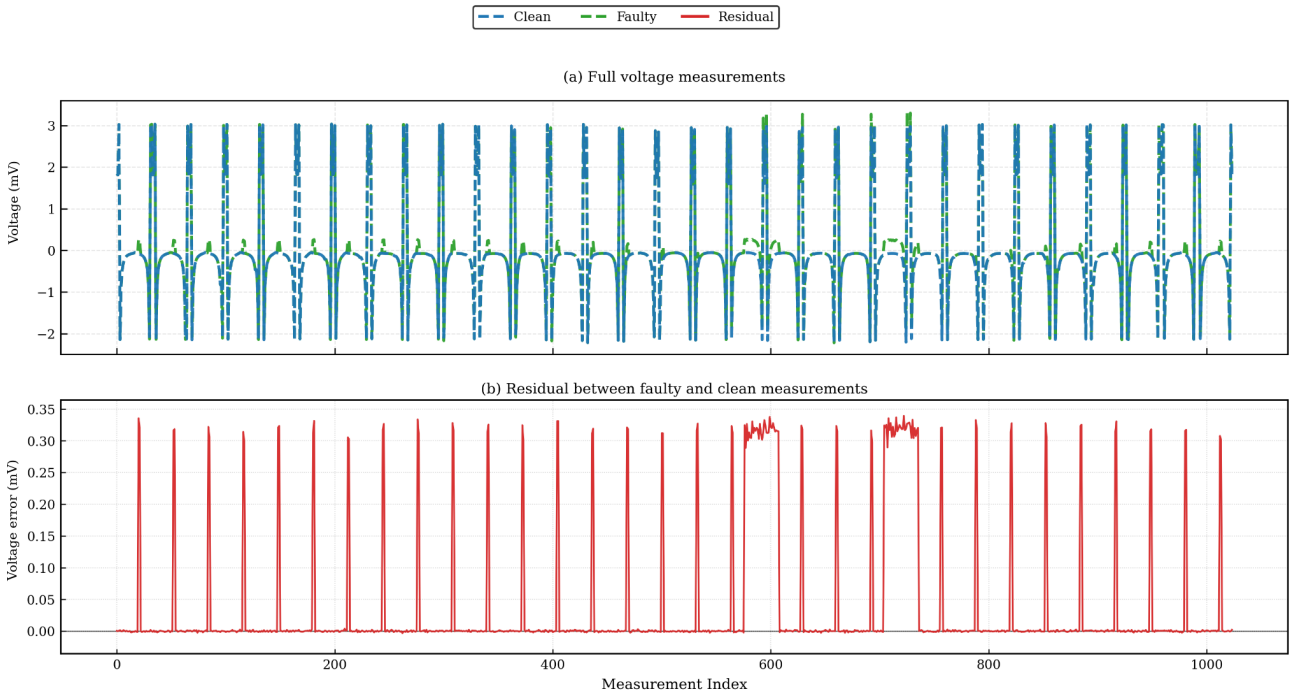


Fig. 1. Effect of faulty-channel corruption on the simulated EIT voltage measurements. (a) compares the clean data with the measurements affected by the faulty E23 electrode. (b) shows the residual between faulty and clean measurements.

localized errors with relatively large amplitudes. The clean and faulty voltage traces exhibit similar overall measurement patterns; which indicates that the fault does not affect all measurements uniformly. However, clear deviations appear at the measurement indices associated with electrode E23. These deviations are more clearly observed in Fig. 1(b), where sharp high-amplitude peaks and localized elevated regions indicate measurement samples affected by the faulty channel. Such sparse, high-amplitude errors can significantly degrade purely quadratic data-fitting methods, since a small number of corrupted residuals may dominate the least-squares objective. This observation supports the use of robust data-fidelity terms and adaptive weighting strategies in the proposed hybrid L_1/L_2 reconstruction framework.

B. Qualitative Reconstructions

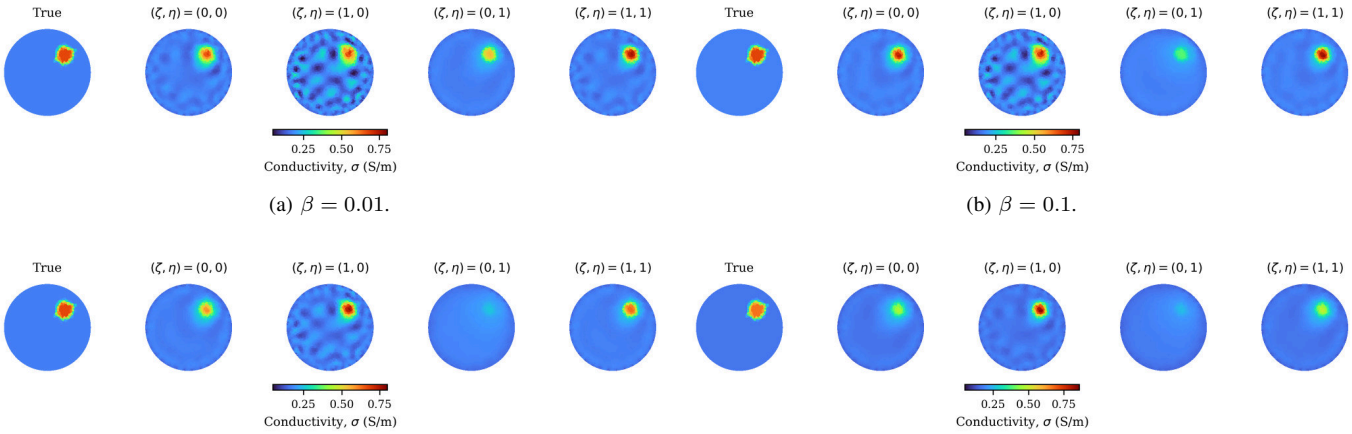
Fig. 2 compares the true conductivity map with reconstructions obtained for the four hybrid settings and different values of β. The quadratic baseline $(0, 0)$ localizes the inclusion but produces diffuse contrast and background artifacts. The robust-data/smooth-prior setting $(1, 0)$ improves inclusion localization and maintains strong threshold-based overlap across regularization levels. The quadratic-data/sparse-prior setting $(0, 1)$ suppresses the inclusion for moderate and large β, showing that prior sparsity alone cannot overcome corrupted data. The fully hybrid setting $(1, 1)$ provides the best overall balance for small to moderate regularization by preserving lesion contrast while reducing background artifacts.

C. Reconstruction Matrix Structure

Fig. 3 shows a representative reconstruction-matrix comparison at $\beta = 0.1$ using $\log_{10}(|\mathbf{K}_{\text{rec}}| + \epsilon_{\log})$, where $\epsilon_{\log}$ is a small positive offset used only to avoid the singularity of $\log_{10}(0)$. The vertical bands correspond to measurement channels with a stronger influence on the reconstructed nodal conductivity values, indicating that the inverse solution is not equally sensitive to all voltage measurements. These sensitivity patterns are determined by the electrode configuration, the skip-4 current-injection protocol, the adjacent voltage-measurement scheme, the adaptive data weights, and the regularization term. In particular, the robust-data configurations modify the relative leverage of inconsistent channels while retaining the broad sensitivity structure of the inverse mapping.

D. Quantitative Performance

Table I summarizes the quantitative reconstruction performance under faulty electrode measurements. The regularization weight was varied over $\beta \in \{0.01, 0.1, 1, 10\}$, and four combinations of the data-fidelity and prior-regularization parameters were evaluated. For $\beta = 0.01$, the fully hybrid setting $(1, 1)$ achieves the lowest RMSE and relative error and the highest Dice, IoU, and CNR. For $\beta = 0.1$, $(1, 1)$ again gives the lowest RMSE and relative error and the highest CNR, whereas $(1, 0)$ gives the highest Dice and IoU. At $\beta = 1$, $(1, 1)$ retains the lowest RMSE, peak error, relative error, and highest CNR, while $(1, 0)$ gives the best spatial overlap. At $\beta = 10$, $(1, 0)$ gives the best RMSE, Dice, IoU, CNR, and relative error. Thus, $(1, 1)$ is the strongest balanced configuration at $\beta = 0.01$-0.1, whereas



(a) $\beta = 0.01$. (b) $\beta = 0.1$.

(c) $\beta = 1$. (d) $\beta = 10$.

Fig. 2. Comparison of the true conductivity distribution and reconstructed images for different regularization parameters β. The background and inclusion conductivities are 0.17 and 0.67 $\mathrm{S/m}$, respectively.

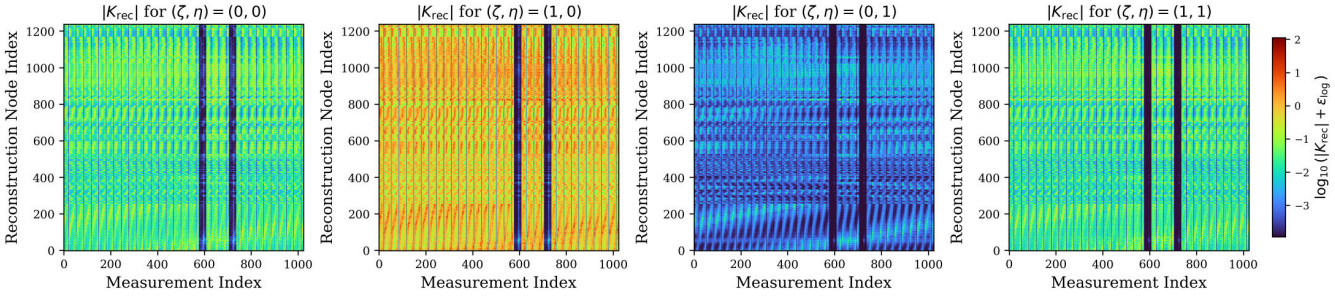


Fig. 3. Representative logarithmic magnitude of the local reconstruction matrix at $\beta = 0.1$ for the four hybrid configurations. The horizontal axis denotes measurement index and the vertical axis denotes reconstruction-node index. Vertical bands indicate measurement channels with relatively high leverage.

$(1,0)$ becomes the most reliable overlap-oriented choice as regularization is strengthened, and it gives the best overall metric profile at $\beta = 10$. This pattern supports the central observation that robust data fidelity is more important than prior sparsity alone when a small subset of measurements is strongly corrupted.

VII. Discussion

The numerical results demonstrate that robust data fidelity is essential for stable EIT reconstruction under localized faulty-electrode corruption generate sparse, high-amplitude voltage residuals. With quadratic data fitting, large faulty-electrode residuals dominate the optimization and may be interpreted as nonphysical conductivity changes. In contrast, the IRLS weights associated with the robust data term reduce the influence of large residuals and allow the reconstruction to rely more strongly on measurements that remain consistent with the CEM forward model.

The prior parameter plays a complementary role. A smooth prior provides conservative stabilization and maintains spatial overlap as β increases, whereas a sparsity-promoting prior can preserve localized contrast when regularization is small. The fully hybrid configuration therefore gives the best balanced performance for small-to-moderate regularization, while the robust-data/smooth-prior setting is preferable when overlap stability is the main objective. These findings are consistent with robust EIT formulations and elastic-net-style regularization in nonlinear inverse problems [11], [12].

The Gaussian-plus-bias perturbation creates finite outliers and is useful for testing the influence function of the data term, but it does not reproduce a completely disconnected electrode. A saturated voltage-sensing channel would likely be strongly downweighted because it generates persistent large residuals. A further limitation is that the quantitative study uses one representative faulty electrode, E23. The present results therefore establish robustness for that configuration but not electrode-position invariance. A second fault location, multiple locations, and hard-disconnection/saturation models should be included in a broader validation study. Other limitations are the two-dimensional circular geometry, fixed contact impedance, synthetic rather than experimentally measured voltages, and one hemorrhage location. Future work should include three-dimensional head models, experimental phantoms, contact-impedance uncertainty, multiple lesions and electrode faults, and adaptive parameter selection.

TABLE I

RECONSTRUCTION PERFORMANCE FOR THE REPRESENTATIVE E23 FAULT. BOLD INDICATES THE BEST VALUE WITHIN EACH β. ↓: LOWER IS BETTER; ↑: HIGHER IS BETTER.

β	(ζ, η)	RMSE↓	Dice↑	IoU↑	PeakErr↓	CNR (dB)↑	RelErr↓
0.01	(0,0)	0.03863	0.8254	0.7027	**0.04370**	21.3942	0.2087
	(1,0)	0.05931	0.8571	0.7500	0.13201	15.8252	0.3205
	(0,1)	0.04505	0.7241	0.5676	0.12345	22.1888	0.2434
	(1,1)	**0.03517**	**0.8955**	**0.8108**	0.10308	**23.7666**	**0.1900**
0.1	(0,0)	0.03835	0.8065	0.6757	**0.07578**	23.2568	0.2072
	(1,0)	0.05158	**0.8955**	**0.8108**	0.12273	17.6375	0.2787
	(0,1)	0.06195	0.0000	0.0000	0.30687	17.4846	0.3347
	(1,1)	**0.03674**	0.8254	0.7027	0.10158	**24.7241**	**0.1985**
1	(0,0)	0.04418	0.7018	0.5405	0.06111	21.3332	0.2387
	(1,0)	0.04181	**0.8788**	**0.7838**	0.12936	20.8353	0.2259
	(0,1)	0.07376	0.0000	0.0000	0.41455	11.1927	0.3985
	(1,1)	**0.03917**	0.7869	0.6486	**0.03083**	**23.8777**	**0.2116**
10	(0,0)	0.05356	0.3913	0.2432	0.19291	18.8774	0.2894
	(1,0)	**0.03830**	**0.8254**	**0.7027**	0.19109	**24.2674**	**0.2069**
	(0,1)	0.07355	0.0000	0.0000	0.41167	11.3983	0.3974
	(1,1)	0.05176	0.4898	0.3243	**0.16954**	19.4407	0.2796

VIII. CONCLUSION

This paper presented a hybrid L_1/L_2 reconstruction framework for hemorrhagic stroke EIT under localized faulty-electrode corruption. The method combines the CEM, a nodal-conductivity finite elements discretization with Gaussian quadrature, hybrid data and prior penalties, and LM-IRLS optimization. For the representative E23 fault, the numerical results indicate that robust data fidelity is essential for suppressing faulty-channel residuals, while the prior controls the trade-off between smoothness, edge preservation, and contrast recovery. The fully hybrid configuration provides the best balanced performance for small-to-moderate regularization, whereas robust data fidelity with a smooth prior remains more stable as regularization becomes stronger. These results are a numerical proof of concept for localized additive corruption and should not be interpreted as experimental validation or as direct evidence for complete electrode disconnection.

REFERENCES

[1] V. L. Feigin, M. Brainin, B. Norrving, S. O. Martins, J. Pandian, P. Lindsay, M. F Grupper, and I. Rautalin, "World stroke organization: global stroke fact sheet 2025," *International Journal of Stroke*, vol. 20, no. 2, pp. 132–144, 2025.

[2] N. Goren, J. Avery, T. Dowrick, E. Mackle, A. Witkowska-Wrobel, D. Werring, and D. Holder, "Multi-frequency electrical impedance tomography and neuroimaging data in stroke patients," *Scientific data*, vol. 5, no. 1, p. 180112, 2018.

[3] T. Ouypornkochagorn, N. Polydorides, and H. McCann, "Towards continuous eit monitoring for hemorrhagic stroke patients," *Frontiers in physiology*, vol. 14, p. 1157371, 2023.

[4] Y. Shi, Y. Lou, M. Wang, S. Zheng, Z. Tian, and F. Fu, "Imaging of conductivity distribution based on a combined reconstruction method in brain electrical impedance tomography." *Inverse Problems & Imaging*, vol. 17, no. 3, 2023.

[5] A. Adler and A. Boyle, "Electrical impedance tomography: Tissue properties to image measures," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2494–2504, 2017.

[6] M. Cheney, D. Isaacson, and J. C. Newell, "Electrical impedance tomography," *SIAM review*, vol. 41, no. 1, pp. 85–101, 1999.

[7] L. Borcea, "Electrical impedance tomography," *Inverse problems*, vol. 18, no. 6, pp. R99–R136, 2002.

[8] K.-S. Cheng, D. Isaacson, J. C. Newell, and D. G. Gisser, "Electrode models for electric current computed tomography," *IEEE Transactions on Biomedical Engineering*, vol. 36, no. 9, pp. 918–924, 1989.

[9] E. Somersalo, M. Cheney, and D. Isaacson, "Existence and uniqueness for electrode models for electric current computed tomography," *SIAM Journal on Applied Mathematics*, vol. 52, no. 4, pp. 1023–1040, 1992.

[10] J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen, "Statistical inversion and monte carlo sampling methods in electrical impedance tomography," *Inverse problems*, vol. 16, no. 5, pp. 1487–1522, 2000.

[11] K. Lee, E. J. Woo, and J. K. Seo, "A fidelity-embedded regularization method for robust electrical impedance tomography," *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 1970–1977, 2017.

[12] J. Wang, B. Han, and W. Wang, "Elastic-net regularization for nonlinear electrical impedance tomography with a splitting approach," *Applicable Analysis*, vol. 98, no. 12, pp. 2201–2217, 2019.

[13] A. Adler, J. H. Arnold, R. Bayford, A. Borsic, B. Brown, P. Dixon, T. J. Faes, I. Frerichs, H. Gagnon, Y. Gärber *et al.*, "Greit: a unified approach to 2d linear eit reconstruction of lung images," *Physiological measurement*, vol. 30, no. 6, pp. S35–S55, 2009.

[14] B. Jin, T. Khan, and P. Maass, "A reconstruction algorithm for electrical impedance tomography based on sparsity regularization," *International Journal for Numerical Methods in Engineering*, vol. 89, no. 3, pp. 337–353, 2012.

[15] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[16] M. Grasmair, M. Haltmeier, and O. Scherzer, "Sparse regularization with lq penalty term," *Inverse Problems*, vol. 24, no. 5, p. 055020, 2008.

[17] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[18] E. H. Bergou, Y. Diouane, and V. Kungurtsev, "Convergence and complexity analysis of a levenberg–marquardt algorithm for inverse problems," *Journal of Optimization Theory and Applications*, vol. 185, no. 3, pp. 927–944, 2020.

[19] N. Marumo, T. Okuno, and A. Takeda, "Majorization-minimization-based levenberg–marquardt method for constrained nonlinear least squares." *Computational Optimization & Applications*, vol. 84, no. 3, 2023.

[20] M. Darbas, J. Heleine, R. Mendoza, and A. C. Velasco, "Sensitivity analysis of the complete electrode model for electrical impedance tomography," *Aims Mathematics*, vol. 6, no. 7, pp. 7333–7366, 2021.

[21] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[22] T. Tang, S. Oh, and R. Sadleir, "A robust current pattern for the detection of intraventricular hemorrhage in neonates using electrical impedance tomography," *Annals of biomedical engineering*, vol. 38, no. 8, pp. 2733–2747, 2010.

[23] M. M. Mellenthin, J. L. Mueller, E. D. L. B. de Camargo, F. S. de Moura, T. B. R. Santos, R. G. Lima, S. J. Hamilton, P. A. Muller, and M. Alsaker, "The ace1 electrical impedance tomography system for thoracic imaging," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 9, pp. 3137–3150, 2018.

Retinal Vessel Segmentation Using a Hybrid Vision Transformer–Convolutional W-Net

Alan Müller
Grad. Prog. in Biomedical Engineering
Universidade de Brasília
Brasília, Brazil
alanmuller@unb.br

Thiago Alves Martins
Grad. Prog. in Medical Sciences
Universidade de Brasília
Brasília, Brazil
martins.thiago@aluno.unb.br

Rodolfo Deusdará
Grad. Prog. in Medical Sciences
Universidade de Brasília
Brasília, Brazil
rodolfo.rodrigues@unb.br

Vinicius de Carvalho Rispoli
Grad. Prog. in Biomedical Engineering
Universidade de Brasília
Brasília, Brazil
vrispoli@unb.br

***Abstract*—Accurate artery–vein segmentation in fundus images is challenging due to thin vessels, crossing ambiguities and strong class imbalance. This work proposes a hybrid ViT–CNN W-Net for simultaneous segmentation of arteries, veins and complete vessel tree. The architecture comprises two consecutive U-shaped modules: the first predicts the three vascular channels, while the second refines artery and vein maps using the vessel-tree output as a structural anchor. Each module combines DINOv3 transformer features, Fidelity-Aware Projection Modules, a full-resolution convolutional stem and a U-Net-like decoder. Low-Rank Adaptation is used to efficiently adapt the transformer attention layers. The proposed model was evaluated on the Fundus-AVSeg dataset using three-fold cross-validation and an independent held-out test set. Five configurations were compared to investigate the effects of transformer backbone size, image enhancement, parameter-efficient adaptation and complete backbone freezing. The best configuration, based on DINOv3-Large with Low-Rank Adaptation and image enhancement, achieved Dice scores of 0.8985, 0.9074 and 0.9397 for arteries, veins and complete vessel tree, respectively, resulting in a mean Dice score of 0.9152. The Base and Large backbones achieved comparable results when adapted with Low-Rank Adaptation, whereas freezing the transformer reduced the mean Dice score to 0.8990. Image enhancement produced only marginal and inconsistent improvements. External validation on the HRF and LES-AV datasets yielded mean Dice scores of 0.7010 and 0.8126, respectively, indicating reduced cross-dataset performance compared with the primary evaluation on Fundus-AVSeg. Nevertheless, qualitative results indicates that the main arterial, venous and vessel-tree structures were consistently recovered, suggesting that differences in image characteristics and annotation protocols may partially account for the quantitative degradation. Overall, these results support the proposed hybrid ViT–CNN W-Net as a feasible framework for joint retinal vessel segmentation and artery–vein differentiation while retaining the flexibility to incorporate alternative transformer backbones.**

***Index Terms*—Retinography, Ophtalmology, Segmentation, U-Net, W-Net, DINOv3, Transfer Learning, LoRA**

I. Introduction

Retinal vessel segmentation is a fundamental step for the development of computational tools for ophthalmic image analysis. The vascular structure observed in color fundus images carries relevant clinical information, since changes in vessel caliber, tortuosity, branching patterns and artery-vein organization may be associated with systemic and ocular conditions such as hypertension, diabetes and cardiovascular diseases [1]–[4]. In this context, accurate vessel segmentation is not only useful as an isolated image-processing task, but also as a prerequisite for downstream diagnostic and decision-support systems.

Convolutional encoder-decoder networks, especially U-Net-based models, have been widely used for biomedical image segmentation [5]. For retinal vessel analysis, cascaded U-shaped architectures have also been explored, forming W-shaped networks in which a first U-Net produces an initial segmentation and a second U-Net refines this output. This strategy is particularly attractive for artery-vein segmentation, since the model must not only identify the vascular tree, but also distinguish visually similar arterial and venous structures. Previous recurrent W-Net approaches have proposed repeated refinement steps to improve segmentation consistency [6], [7]. In previous work, we observed that this type of refinement may improve performance, but its benefit depends strongly on the training regime and on the amount of effective supervision available [8].

More recently, self-supervised vision transformers have become strong feature extractors for medical imaging. Among them, DINOv3 has shown high transferability across visual tasks [9]. In parallel, U-shaped has been proposed as a way of combining frozen DINOv3 representations with U-Net-like decoding for medical image segmentation [10]. In this architecture, intermediate transformer representations are adapted into multi-scale feature maps and used as skip connections in a decoder. However, the original U-shaped formulation was not designed specifically for artery-vein segmentation and its application to fundus imaging was mainly evaluated in optic disc and cup segmentation. Therefore, its use for simultaneous

artery, vein and full vessel-tree segmentation remains an open and relevant question.

In this work, we propose and evaluate a hybrid Vision Transformer–convolutional neural network (ViT–CNN) W-Net for the simultaneous segmentation of retinal arteries, veins and the complete vessel tree. The architecture comprises two consecutive hybrid U-shaped modules. The first module receives the fundus image and produces an initial three-channel prediction, whereas the second receives the sigmoid-normalized outputs of the first stage and refines only the artery and vein maps. The vessel-tree prediction is preserved as a structural anchor, providing anatomical context while allowing the second U-shaped branch to correct artery–vein inconsistencies produced by the initial segmentation stage.

Each hybrid module combines intermediate transformer representations, Fidelity-Aware Projection Modules (FAPMs) [10], a full-resolution convolutional stem and a U-Net-like convolutional decoder. The transformer features are converted into spatial maps and adapted by the FAPMs, which project the high-dimensional representations into stage-specific channel dimensions before constructing the decoder feature pyramid. In parallel, the convolutional stem extracts local information directly from the input and provides a full-resolution skip connection to the decoder. This design combines the semantic representation capacity of a pretrained ViT with the spatial detail supplied by convolutional processing, which is particularly important for segmenting thin retinal vessels. Although the architecture is not restricted to a specific transformer backbone, DINOv3 was adopted in this implementation because of the quality of its visual representations and its demonstrated effectiveness in medical image segmentation [10].

We investigate five experimental configurations using the Fundus-AVSeg dataset [11]. The experiments compare DINOv3-Base and DINOv3-Large backbones, with and without contrast-based image enhancement and evaluate lightweight Low-Rank Adaptation (LoRA) [12] against a completely frozen transformer backbone. LoRA adapts selected attention projections while introducing only a small number of trainable parameters. This strategy is particularly relevant for large pretrained vision encoders, for which full fine-tuning may be computationally expensive and prone to overfitting when only small medical imaging datasets are available.

The main contributions of this work are: (1) the introduction of a hybrid ViT–CNN W-Net for joint artery, vein and vessel-tree segmentation, in which the vessel-tree prediction is preserved as a structural anchor during refinement; (2) the evaluation of LoRA-based transformer adaptation in comparison with a fully frozen backbone; (3) a comparison between DINOv3-Base and DINOv3-Large under the same training and evaluation protocol; and (4) an analysis of the effect of contrast-based fundus image enhancement on segmentation performance. The implementation was developed in PyTorch [13], with sliding-window inference and evaluation supported by MONAI [14] and preprocessing based on scikit-image and CLAHE [15], [16]. Performance was primarily assessed using the Dice–Sørensen coefficient, complemented by sensitivity, specificity, precision, ROC AUC and precision–recall AUC.

II. Methods

A. Study design and Data

This paper evaluated five experiments with different configurations of a W-shaped retinal vessel segmentation model based on DINOv3 encoders and convolutional decoders. Task was defined as a three-channel segmentation problem, with different outputs for arteries, veins and the complete vascular tree. All experiments used the same data split, training protocol and evaluation pipeline, allowing the comparison of the effect of the backbone size, image enhancement, LoRA adaptation and full backbone freezing, under similar conditions.

Fundus-AVSeg [11] was selected because it comprises 100 pairs of fundus images and their corresponding annotations. It is a publicly available dataset specifically designed for the training and evaluation of artificial intelligence models for automatic retinal vessel segmentation. The main reason for selecting this dataset is its size, as other publicly available datasets commonly used for this purpose contain considerably fewer images with artery and vein annotations.

Among the main alternatives are HRF [17], [18], which contains 45 image-mask pairs, RITE [19], with 40 image-mask pairs and LES-AV [20], with only 22 image-mask pairs. Another advantage of Fundus-AVSeg is that its 100 images were acquired from 100 different subjects, as reported in [11]. Therefore, there is no subject overlap within the dataset, eliminating the risk of patient-level data leakage between the training and testing sets.

The fixed random seed 42 was used to split the data into a held-out test set and a training-validation subset. Twenty-five unique images were reserved for final testing and the other 75 images were used in a three-fold cross-validation scheme. In each fold, 50 images were used for training and 25 images for validation. The final test prediction was obtained through the average probability maps produced by all the three fold-specific models.

B. Image and segmentation mask preprocessing

Original images on Fundus-AVseg have two different resolutions: 2656×1992 or 1280×1280. To ensure a standardization, non-square images were cropped on both sides and resized to 1280×1280. Input images were read as RGB fundus images (Fig.1(a)) and normalized to the unit interval. A binary field-of-view (FOV) mask was estimated from the largest connected component of the retinal region and was used to restrict the loss and evaluation metrics to the retinal area. As in ref. [7], original color-coded annotations (as seen in Fig.1(c)) were converted into three binary channels: artery (A) on red channel, vein (V) on green channel and total vessel tree (BV) on blue channel (as seen in Fig.1(d)). Ambiguous or unknown vessel pixels were included in the total vessel channel.

Two preprocessing settings were evaluated: (i) In the non-enhanced setting, the normalized RGB image was used directly; (ii) In the enhanced setting (Fig.1(b)), a contrast-enhancement pipeline was applied into the retinal field of view, based on ref. [7]. This pipeline combines a mild zoom-based background composition, CLAHE in the luminance channel, Gaussian background subtraction and normalization within the FOV. Comparison between enhanced and non-enhanced images was included to assessment, whether this classical preprocessing step improved the proposed model.

C. Model architecture

The proposed model is designed as a hybrid Vision Transformer–convolutional neural network (ViT–CNN) W-Net composed of two consecutive U-shaped modules (Fig. 2). Each module combines a ViT-based encoder, Fidelity-Aware Projection Modules (FAPMs), a shallow convolutional stem and a U-Net-like convolutional decoder. In the present implementation, DINOv3 is adopted as the transformer backbone and adapted using LoRA. The architecture took its inspiration in the U-shaped architecture. However, unlike its original adapter, which uses a Spatial Prior Module to generate multi-scale convolutional priors that interact with the transformer representations, our model employs a simplified full-resolution convolutional stem. The stem provides local spatial information directly to the decoder, while the intermediate DINOv3 representations are independently adapted by the FAPMs and resized to form the remaining decoder skip levels.

The first hybrid U-shaped module receives the fundus image and produces an initial three-channel segmentation map corresponding to arteries, veins and the complete vessel tree, denoted by A_0, V_0 and BV_0, respectively. The second module receives the sigmoid-normalized maps produced by the first stage and refines only the artery and vein predictions. The vessel-tree channel BV_0 is preserved and used as a structural vascular anchor during refinement, since it represents the global vessel topology estimated by the first module. The final prediction therefore combines the refined artery and vein outputs, A_1 and V_1, from the second module with the vessel-tree output BV_0 inherited from the first module. Thus, the first U-shaped branch estimates the overall vascular structure and an initial artery–vein separation, whereas the second branch corrects artery–vein inconsistencies conditioned on the initial predictions, resulting in the W-shaped architecture shown in Fig. 2.

Each U-shaped module receives a three-channel input. The input was then normalized using ImageNet statistics before being passed to the DINOv3 encoder. The encoder extracted intermediate transformer representations from four selected blocks. For the DINOv3-Base experiments, the extracted blocks were 2, 5, 8 and 11. For the DINOv3-Large LoRA experiments, the extracted blocks were 5, 11, 17 and 23.

Since DINOv3 is a Vision Transformer with a patch size of 16 and internal resolution of 512×512, its transformer representations are produced on a patch grid rather than as a natural convolutional feature pyramid. Only the patch tokens were used to construct the spatial feature maps, class and register tokens were not used for this purpose. The selected patch tokens were reshaped into two-dimensional feature maps and then processed by FAPMs. Each FAPM projected a transformer feature map from the DINOv3 embedding dimension to a task-specific number of channels. The resulting feature maps were resized to form a multi-scale representation suitable for the U-Net-like convolutional decoder.

In each U-shaped module, a shallow convolutional stem is applied in parallel to the DINOv3 branch. The stem receives the same ImageNet-normalized three-channel input used by the transformer branch: the RGB fundus image in the base module and the three sigmoid-normalized prediction maps in the refinement module. It is composed of two consecutive 3×3 convolutional layers with stride 1 and padding 1, each followed by batch normalization and ReLU activation. The convolutions do not use bias terms and produce 32 feature channels. Since no pooling or strided convolution is used, the stem preserves the full spatial resolution of the input. Its output is used as the highest-resolution skip connection of the decoder, complementing the lower-resolution DINO-FAPM feature maps and providing local spatial information that may be weakened by the transformer patch representation.

Inside each FAPM, the projection is performed through a combination of a shared low-dimensional basis and a stage-specific basis. The shared basis generates modulation parameters that condition the stage-specific representation in a FiLM-like manner [21]. The modulated features are then refined by a lightweight convolutional block composed of pointwise convolution, batch normalization, ReLU activation, depthwise separable convolution and squeeze-and-excitation recalibration. A shortcut projection is added to preserve information from the modulated representation. In our implementation, the four FAPM stages produced 64, 128, 256 and 512 channels, respectively, using a projection rank of 256.

The decoder follows the standard U-Net principle. Starting from the deepest adapted feature map, transposed convolutions progressively recover spatial resolution. At each level, the upsampled representation is concatenated with the corresponding encoder skip connection and processed by two convolutional layers with batch normalization and ReLU activation. A final 1×1 convolution produces the segmentation logits. In the first U-shaped module, this layer produces three channels, corresponding to arteries, veins and the complete vessel tree. In the second module, it produces two channels, corresponding only to the refined artery and vein predictions.

D. Loss Function

The model was trained using a masked binary cross-entropy loss. Let $\hat{Y}^{(0)} = (\hat{A}_0, \hat{V}_0, \widehat{BV}_0)$ denote the logits produced by the first U-shaped module and let $\hat{Y}^{(1)} = (\hat{A}_1, \hat{V}_1)$ denote the refined artery and vein logits produced by the second U-shaped module. The corresponding ground-truth masks are denoted by A, V and BV.

Because artery and vein labels are not reliable in every pixel, the artery and vein losses were computed only over pixels with

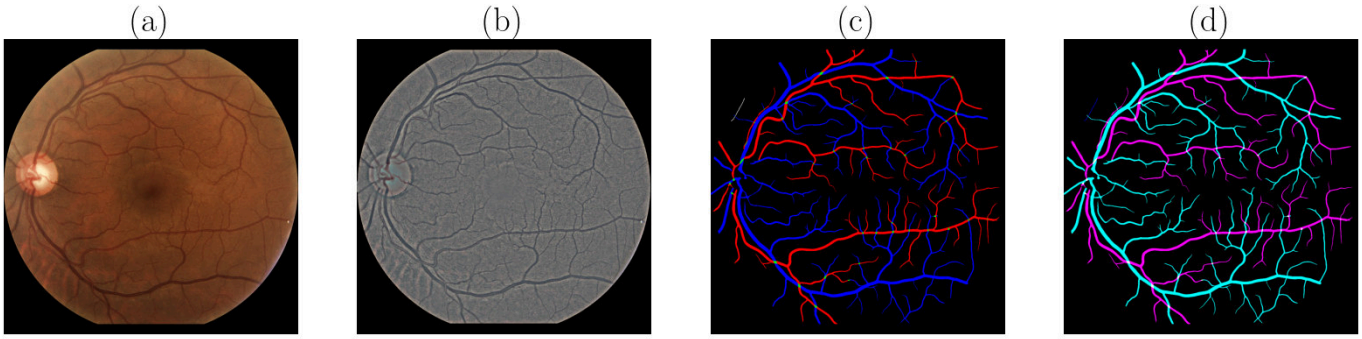


Fig. 1. (a) Raw fundus image; (b) preprocessed and input for the model; (c) original ground truth segmentation mask; (d) adjusted ground truth segmentation mask.

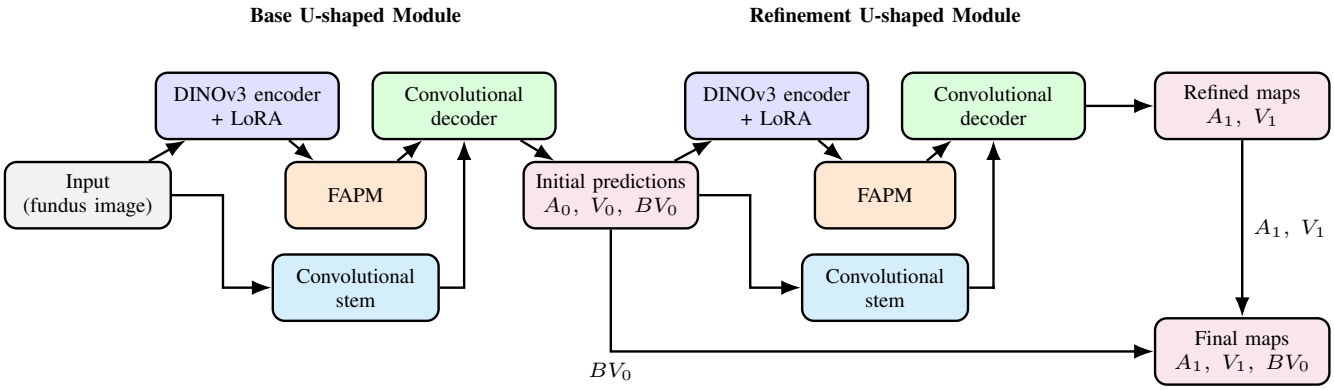


Fig. 2. Schematic representation of the proposed W-shaped architecture. The first U-shaped receives the input fundus image and produces initial predictions for arteries (A_0), veins (V_0) and the complete blood vessel tree (BV_0). The second U-shaped receives these initial maps and refines only the artery and vein predictions. The final blood vessel map remains fixed and is inherited from the first stage.

valid artery-vein annotations. In contrast, the total vessel loss was computed over the complete annotated vessel mask. The masked binary cross-entropy for a predicted logit map $\hat{y}$, target mask y and binary validity mask m was defined as

$$\mathcal{L}_{\mathrm{BCE}}(\hat{y}, y; m) = \frac{\sum_p m_p \,\mathrm{BCEWithLogits}(\hat{y}_p, y_p)}{\sum_p m_p + \epsilon}, \quad (1)$$

where p indexes image pixels and ϵ is a small constant used for numerical stability.

The first module was supervised using the three initial predictions,

$$\begin{aligned}\mathcal{L}_0 &= \mathcal{L}_{\mathrm{BCE}}(\hat{A}_0, A; m_{AV}) + \mathcal{L}_{\mathrm{BCE}}(\hat{V}_0, V; m_{AV}) \\ &+ \mathcal{L}_{\mathrm{BCE}}(\widehat{BV}_0, BV; m_{BV}), \end{aligned} \quad (2)$$

where m_{AV} indicates pixels with valid artery-vein labels and m_{BV} indicates pixels used for vessel supervision. The refinement module was supervised only on the artery and vein outputs,

$$\mathcal{L}_1 = \mathcal{L}_{\mathrm{BCE}}(\hat{A}_1, A; m_{AV}) + \mathcal{L}_{\mathrm{BCE}}(\hat{V}_1, V; m_{AV}). \quad (3)$$

The final loss was defined as

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1. \quad (4)$$

This loss encourages the first module to learn a complete vascular representation while forcing the second module to specialize in artery-vein refinement. Since the blood vessel channel is inherited from the first module, no additional vessel-tree loss is applied to the second module.

E. Backbone adaptation strategies

The five experiments were designed to assess the effects of transformer backbone size, image enhancement and backbone adaptation strategy. Experiments 1 and 2 used DINOv3-Large with LoRA [12], without and with image enhancement, respectively. Experiments 3 and 4 used DINOv3-Base with LoRA, also without and with image enhancement, respectively. Experiment 5 used DINOv3-Base with image enhancement but removed LoRA, keeping the transformer backbone completely frozen.

For the first four experiments, LoRA matrices were injected into the attention projections `q_proj`, `k_proj`, `v_proj` and `o_proj` of the pretrained DINOv3 backbone [12]. The LoRA rank was 16, the scaling parameter was 32 and the dropout probability was 0.05. In these experiments, the task-specific modules, including the stem, feature-adaptation modules, projection layers and decoder, were also trained. The fifth experiment removed LoRA and kept the DINOv3 backbone

completely frozen. In this setting, only the task-specific modules were optimized. This experiment was included to estimate how much performance was provided by the frozen pretrained representation alone, without adapting the transformer attention layers.

F. Training protocol

Models were trained with AdamW using an initial learning rate of 5×10^{-4} and a weight decay of 10^{-2} during the main training stage. During this stage, the learning rate followed a cosine annealing schedule, decreasing from its initial value to a minimum value of 1×10^{-6}. The main training stage was validated every ten epochs and stopped early when validation performance did not improve for 120 epochs, corresponding to 12 consecutive validation checks without improvement. After the main training stage, each fold was fine-tuned from its best validation checkpoint using a fixed lower learning rate of 1×10^{-4}. Fine-tuning was validated every two epochs and stopped early when validation performance did not improve for 60 epochs, corresponding to 30 consecutive validation checks without improvement. All experiments used the same training, validation and test splits.

Training was performed on random image crops of size 896×896 extracted from the original 1280×1280 images and masks. The augmentation pipeline included random color perturbation, random affine transformations, spatial cropping, horizontal and vertical flips and cutout (same as in ref. [7]). Validation and test evaluation were performed without any augmentation.

For this paper, no hyperparameter optimization or ablation techniques were performed, aiming to compare only the performance of the chosen backbones under similar circumstances. No references of similar approach were found on the literature and comparison of results are made to recent techniques using U-Net like architectures. All models were trained in Google Colab using an NVIDIA RTX PRO 6000 GPU with 96 GB of VRAM. Code, models and results are available at: https://github.com/labviaunb/cbeb2026.

G. Evaluation

During evaluation, predictions were generated at the original image resolution using the `sliding_window_inference` function from MONAI [14]. Since the model was trained on cropped image patches, full-resolution fundus images were processed by applying the network over overlapping windows with the same spatial size used during training. MONAI then aggregated the window-level logits to reconstruct the complete segmentation maps for the artery, vein and vessel channels. All quantitative metrics were computed from these reconstructed full-image predictions.

Performance was evaluated on the held-out test set using the ensemble of the three fold-specific models. For each test image, the probability maps from the three models were averaged, with threshold at 0.5, to obtain binary masks. Metrics were computed inside the retinal field of view.

Primary metric is the Dice coefficient, computed separately for artery, vein and total vessel tree. The mean Dice reported in the tables corresponds to the average of these three channel-wise Dice values. Sensitivity, specificity, precision, area under the ROC curve and area under the precision-recall curve were computed for each channel. The final validation Dice is reported as the mean and standard deviation across the three folds after fine-tuning.

III. Results and Discussion

Table I summarizes the test performance of the five experiments. The best mean test Dice was obtained by Exp. 2, which used the DINOv3-Large backbone with LoRA and image enhancement, reaching a mean Dice on the test set of 0.9152. The difference among the four LoRA-based experiments was small: minimum and maximum mean Dice values were 0.9139 and 0.9152. This suggests that, under the current training protocol, both DINOv3-Base and DINOv3-Large provided similarly strong representations when LoRA adaptation was used.

The frozen-backbone experiment showed a clear drop in performance. Exp. 5, using DINOv3-Base without LoRA, achieved a mean Dice of 0.8990 on the test set. This result indicates that the frozen DINO representation alone was useful, but adapting the backbone through LoRA improved segmentation quality, particularly for artery and vein separation.

The effect of image enhancement was not consistent across the two LoRA-adapted backbones. With DINOv3-Large, enhancement slightly increased the mean Dice from 0.9146 in Exp. 1 to 0.9152 in Exp. 2. Furthermore, with DINOv3-Base, enhancement slightly decreased the mean Dice from 0.9142 in Exp. 3 to 0.9139 in Exp. 4. Since the differences are small, they must be interpreted with caution, suggesting that the DINO-based models were already able to extract useful vessel information from the normalized RGB images. In this setting, the contrast-enhancement preprocessing adopted from Morano et al. [6], [7] did not provide a consistent performance advantage.

DINOv3-Large and DINOv3-Base showed very similar performance when LoRA was used. Without image enhancement, Exp. 1 with DINOv3-Large achieved a mean Dice of 0.9146, while Exp. 3 with DINOv3-Base achieved 0.9142. With enhancement, Exp. 2 with DINOv3-Large achieved 0.9152, while Exp. 4 with DINOv3-Base achieved 0.9139. These results indicate that increasing the backbone size from Base to Large did not lead to a large improvement in this dataset and training setting.

Comparison between the LoRA-based Base model and the frozen Base model, in Exp. 4 and Exp. 5, indicates that LoRA adaptation was beneficial. The mean Dice decreased from 0.9139 to 0.8990 when LoRA was removed. The largest drops were observed in the artery and vein channels, which are more complex tasks than total vessel segmentation.

Segmentation of total vessel channel was very consistent in all experiments, even when DINOv3 backbone was frozen. The vessel-tree Dice ranged from 0.9319 to 0.9399, whereas

TABLE I
MAIN TEST RESULTS FOR THE FIVE EXPERIMENTS. THE MEAN DICE IS THE AVERAGE OF ARTERY, VEIN AND TOTAL VESSEL DICE.

Experiment	Mean Dice	Artery Dice	Vein Dice	Vessel Dice	Validation Dice	Validation SD
Exp. 1	0.9146	0.8969	0.9069	**0.9399**	0.9081	0.0042
Exp. 2	**0.9152**	**0.8985**	0.9074	0.9397	**0.9097**	0.0044
Exp. 3	0.9142	0.8982	**0.9075**	0.9369	0.9076	0.0032
Exp. 4	0.9139	0.8978	0.9057	0.9382	0.9088	**0.0029**
Exp. 5	0.8990	0.8763	0.8888	0.9319	0.8937	0.0049

TABLE II
DETAILED TEST METRICS FOR EACH EXPERIMENT AND SEGMENTATION CHANNEL.

Experiment	Channel	Dice	Sensitivity	Specificity	Precision	ROC AUC	PR AUC
Exp. 1	Artery	0.8969	0.8882	0.9947	0.9089	0.9894	0.9515
Exp. 1	Vein	0.9069	0.8955	0.9947	0.9228	0.9925	0.9587
Exp. 1	Vessel	**0.9399**	**0.9337**	0.9929	0.9486	**0.9975**	**0.9871**
Exp. 2	Artery	**0.8985**	0.8900	0.9948	0.9106	**0.9961**	**0.9567**
Exp. 2	Vein	0.9074	0.8935	**0.9949**	**0.9263**	**0.9957**	0.9600
Exp. 2	Vessel	0.9397	0.9326	**0.9930**	**0.9492**	**0.9975**	0.9868
Exp. 3	Artery	0.8982	**0.8903**	0.9947	0.9093	0.9926	0.9558
Exp. 3	Vein	**0.9075**	**0.8971**	0.9946	0.9219	0.9943	**0.9608**
Exp. 3	Vessel	0.9369	0.9322	0.9923	0.9442	0.9973	0.9858
Exp. 4	Artery	0.8978	0.8880	**0.9949**	**0.9112**	0.9948	0.9553
Exp. 4	Vein	0.9057	0.8943	0.9946	0.9220	0.9948	0.9591
Exp. 4	Vessel	0.9382	0.9304	0.9929	0.9486	0.9974	0.9863
Exp. 5	Artery	0.8763	0.8586	0.9943	0.8987	0.9946	0.9443
Exp. 5	Vein	0.8888	0.8771	0.9935	0.9055	0.9948	0.9518
Exp. 5	Vessel	0.9319	0.9244	0.9921	0.9422	0.9967	0.9836

TABLE III
EXTERNAL VALIDATION PERFORMANCE ON THE HRF AND LES-AV DATASETS.

Dataset	Channel	Mean Dice	ROC AUC	PR AUC
HRF	Artery	0.6711 ± 0.1582	0.9371 ± 0.0955	0.7170 ± 0.1654
HRF	Vein	0.7145 ± 0.1608	0.9550 ± 0.0847	0.7930 ± 0.1638
HRF	Vessel	0.7175 ± 0.1630	0.9460 ± 0.0967	0.8006 ± 0.1610
LES-AV	Artery	0.7928 ± 0.0377	0.9876 ± 0.0062	0.8758 ± 0.0367
LES-AV	Vein	0.8120 ± 0.0305	0.9894 ± 0.0061	0.9020 ± 0.0309
LES-AV	Vessel	0.8330 ± 0.0227	0.9866 ± 0.0049	0.9181 ± 0.0197

TABLE IV
CONTEXTUAL COMPARISON WITH THE METRICS REPORTED BY MORANO *et al.* [7]. RESULTS ARE NOT DIRECTLY COMPARABLE SINCE WERE OBTAINED ON DIFFERENT DATASETS.

	Artery		Vein		BV		A/V		BV/BG	
Method	**AUROC**	**AUPR**	**AUROC**	**AUPR**	**AUROC**	**AUPR**	**Sens.**	**Spec.**	**Sens.**	**Spec.**
W-Net (RITE [19])	97.30	83.50	97.89	87.94	98.23	92.66	90.92	92.31	81.85	98.18
RRWNet ($k = 2$) (RITE [19])	98.19	84.93	98.47	89.22	97.98	90.83	86.39	92.26	80.16	98.61
RRWNet ($k = 6$) (RITE [19])	98.55	86.08	98.75	89.64	97.98	90.98	**94.46**	95.38	80.16	98.61
Ours – DINOv3-Large (FundusAV [11])	**99.61**	**95.67**	**99.57**	**96.00**	**99.75**	**98.68**	89.18	**99.49**	**93.26**	**99.30**
Ours – DINOv3-Base (FundusAV [11])	99.48	95.53	99.48	95.91	99.74	98.63	90.12	99.48	93.04	99.29

TABLE V
CONTEXTUAL COMPARISON WITH THE METRICS REPORTED BY ZENG *et al.* [22]. RESULTS ARE NOT DIRECTLY COMPARABLE SINCE WERE OBTAINED ON DIFFERENT DATASETS.

	BV	A/V	
Method	**AUROC**	**Sens.**	**Spec.**
RRWNet (LES-AV [20])	97.33	94.03	96.14
RRWNet (RITE [19])	98.36	96.21	97.20
RRWNet (HRF [17])	98.35	**98.21**	98.33
Ours – DINOv3-Large (FundusAV [11])	**99.75**	89.18	**99.49**
Ours – DINOv3-Base (FundusAV [11])	99.74	90.12	99.48

artery Dice ranged from 0.8763 to 0.8985 and vein Dice ranged from 0.8888 to 0.9075. This behavior is expected because the vessel-tree prediction is generated by the first U-shaped module and preserved as a structural anchor, while the second U-shaped module refines only the artery and vein predictions.

In Table II all the experiment metrics are reported. All experiments achieved high specificity, above 0.992 for all channels, indicating that false positives were relatively controlled. It is very important in retinal vessel segmentation, because most pixels inside the field of view correspond to

background. Sensitivity was lower than specificity, especially in the frozen-backbone experiment, suggesting that missed vessel pixels were main error source. The precision–recall AUC values were also high, particularly for the total vessel channel, where all LoRA-based experiments achieved values above 0.985.

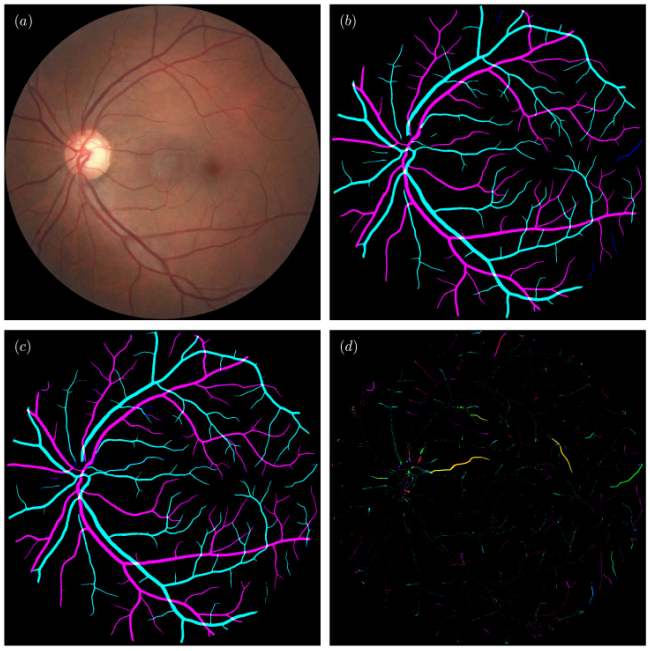


Fig. 3. (a) Raw fundus image from the Fundus-AVSeg dataset [11]; (b) adjusted ground-truth segmentation mask; (c) segmentation predicted by the DINOv3-Large model; and (d) absolute difference between the ground-truth and predicted masks. The predicted segmentation achieved Dice scores of 0.9481 for arteries, 0.9415 for veins and 0.9608 for the complete vessel tree, with a mean Dice score of 0.9501, indicating excellent agreement between the predicted and ground-truth masks.

Figures 3– 5 show comparison between ground truth, prediction and their subtraction. Figure 3 reports test analysis in the same dataset of the training [11], achieving high scores and excellent agreement with ground truth, with mean Dice of 0.9501. On the other hand, Figures 4 and 5 reports the model test in images of the LES-AV [20] and HRF [17], [18] datasets, respectively. These results shows good qualitatively representation of the arteries, veins and vessel tree. In the LES-AV image the mean Dice score reduced to 0.8165, while in the HRF image the mean Dice score reduced to 0.8171.

Table III reports the external validation performance of the proposed model on the HRF and LES-AV datasets. On HRF, the model achieved mean Dice scores of 0.6711, 0.7145 and 0.7175 for arteries, veins and the complete vessel tree, respectively. Higher cross-dataset performance was observed on LES-AV, with corresponding Dice scores of 0.7928, 0.8120 and 0.8330. A similar trend was observed for the ROC AUC and PR AUC metrics, with LES-AV consistently yielding higher values than HRF.

Finally, Tables IV and V provide a contextual comparison with the results reported by Morano *et al.* [7] and Zeng *et al.* [22], respectively, whose W-Net-like models were trained

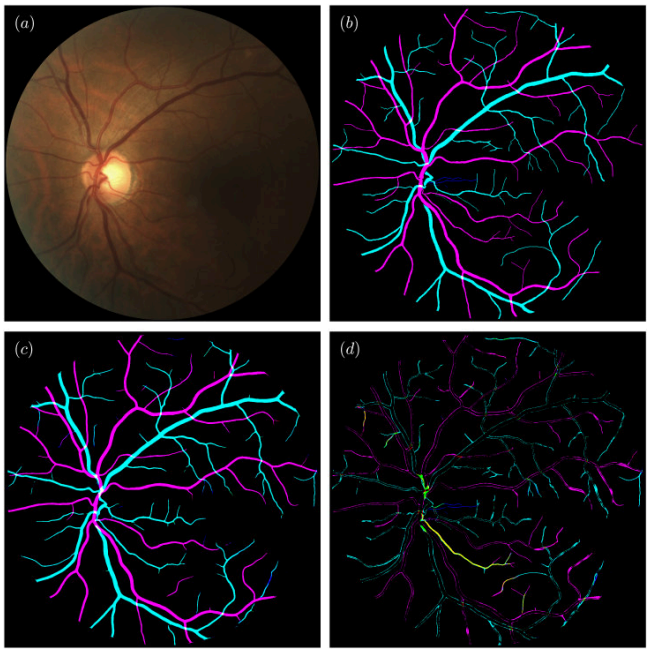


Fig. 4. (a) Raw fundus image from the LES-AV dataset [20]; (b) adjusted ground-truth segmentation mask; (c) segmentation predicted by the DINOv3-Large model; and (d) absolute difference between the ground-truth and predicted masks. The predicted segmentation achieved Dice scores of 0.8158 for arteries, 0.8020 for veins and 0.8317 for the complete vessel tree, with a mean Dice score of 0.8165, indicating good agreement between the predicted and ground-truth masks.

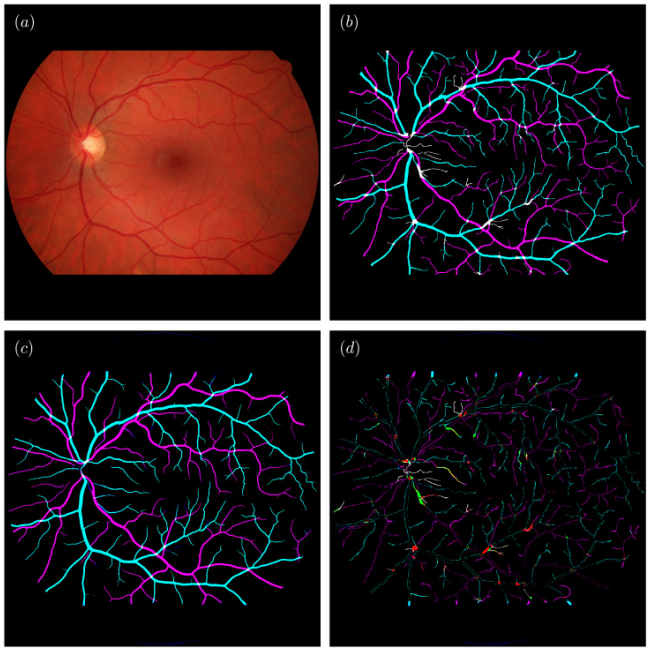


Fig. 5. (a) Raw fundus image from the HRF dataset [17], [18]; (b) adjusted ground-truth segmentation mask; (c) segmentation predicted by the DINOv3-Large model; and (d) absolute difference between the ground-truth and predicted masks. The predicted segmentation achieved Dice scores of 0.7884 for arteries, 0.8276 for veins and 0.8352 for the complete vessel tree, with a mean Dice score of 0.8171, indicating good agreement between the predicted and ground-truth masks.

and evaluated on datasets such as RITE [19], LES-AV [20] and HRF [17], [18]. In contrast, our model was trained on Fundus-AVSeg [11], while Table III reports its external validation on previously unseen HRF and LES-AV images. Therefore, direct comparisons are not appropriate, as differences in training data, annotation protocols, preprocessing and evaluation settings may substantially affect the reported metrics. Then, Tables IV and V should thus be interpreted only as references to the performance range reported in the literature. Despite the lower quantitative performance observed in the external validation, the qualitative examples in Figs. 4 and 5 show that the model still recovered the main arterial, venous and vessel-tree structures in both datasets, with discrepancies likely influenced by differences in image characteristics and annotation criteria. Within these limitations, the results support the use of hybrid ViT–CNN U-shaped modules with parameter-efficient LoRA adaptation as a competitive alternative to fully convolutional W-Net architectures.

IV. Conclusion

This work proposed a hybrid ViT–CNN W-Net for the simultaneous segmentation of retinal arteries, veins and the complete vessel tree. The architecture uses two consecutive U-shaped modules, preserving the vessel-tree prediction from the first stage as a structural anchor for artery–vein refinement in the second stage. The experiments showed that LoRA provides an effective parameter-efficient strategy for adapting the transformer backbone, outperforming the completely frozen configuration. DINOv3-Base and DINOv3-Large produced comparable results, suggesting that increasing backbone size provides limited gains in this setting, while contrast-based image enhancement had only a marginal and inconsistent effect. Overall, the results support the use of hybrid ViT–CNN modules in W-shaped architectures for accurate retinal vessel segmentation and artery–vein differentiation.

The main limitation of this study is that model training and primary evaluation were restricted to Fundus-AVSeg, limiting conclusions about generalization across acquisition devices, populations and annotation protocols. External validation on HRF and LES-AV showed reduced quantitative performance despite good qualitatively vascular predictions, suggesting sensitivity to cross-dataset differences in image characteristics and annotation criteria. Future work should include training and systematic evaluation across multiple datasets, dedicated ablation studies of the proposed architecture, comparisons with full transformer fine-tuning and other parameter-efficient adaptation strategies and assessment of the clinical relevance of the predicted masks through retinal vascular biomarkers.

Acknowledgment

The authors gratefully acknowledges the Brazilian National Council for Scientific and Technological Development (CNPq) for the Technological Development and Innovative Extension Fellowship (Grant No. 380610/2026-4).

References

[1] M. K. Ikram, *et al.*, Retinal vessel diameters and the risk of stroke: the Rotterdam Study, *Neurology*, 66, (2006) 1339–1343.
[2] R. Klein, *et al.*, Retinal vessel caliber measurements: clinical significance, current knowledge and future perspectives, *Archives of Ophthalmology*, 122, (2004) 558–566.
[3] T. Y. Wong, *et al.*, Retinal arteriolar narrowing and risk of coronary heart disease in men and women, *JAMA*, 287, (2002) 1153–1159.
[4] T. Y. Wong, *et al.*, Measures of retinal vessel diameters: associations with cardiovascular disease in the Multi-Ethnic Study of Atherosclerosis (MESA), *Am. J. Ophthalmol.*, 145, (2008) 588–594.
[5] K. Ren, L. Chang, M. Wan, G. Gu, Q. Chen, An Improved U-Net Based Retinal Vessel Image Segmentation Method, *Heliyon*, 8, (2022) e11187.
[6] J. Morano, Á. S. Hervella, J. Novo, J. Rouco, Simultaneous Segmentation and Classification of the Retinal Arteries and Veins from Color Fundus Images, *Artif. Intell. Med.*, 118, (2021) 102116.
[7] J. Morano, G. Aresta, H. Bogunović, RRWNet: Recursive Refinement Network for effective retinal artery/vein segmentation and classification, *Expert Syst. Appl.*, 256, (2024) 124970.
[8] V. de C. Rispoli, *et al.*, Retinal Vessel Segmentation Via Transformer-based Recursively Refined W-Net, *Anais do III Encontro Regional de Matemática Aplicada e Computacional de Santa Catarina (ERMAC-SC)*, (2026).
[9] O. Siméoni, *et al.*, DINOv3, *arXiv*, (2025) 2508.10104.
[10] H. Li, *et al.*, Dino U-Net: Exploiting High-Fidelity Dense Features from Foundation Models for Medical Image Segmentation, *arXiv*, (2026) 2508.20909.
[11] Z. Deng, W. Gao, Z. Gong, R. Gan, L. Chen, S. Zhang, L. Ma, A Fundus Image Dataset for AI-based Artery-Vein Vessel Segmentation, *Sci. Data*, 12, (2025) 1298.
[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, *arXiv*, (2021) 2106.09685.
[13] A. Paszke, *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library, *arXiv*, (2019) 1912.01703.
[14] M. Cardoso, *et al.*, MONAI: An open-source framework for deep learning in healthcare, *arXiv*, (2022) 2211.02701.
[15] S. van der Walt, *et al.*, scikit-image: image processing in Python, *arXiv*, (2014) 1407.6245.
[16] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, K. E. Muller, Contrast-limited adaptive histogram equalization: speed and effectiveness, *Proceedings of the First Conference on Visualization in Biomedical Computing*, Atlanta, GA, USA, (1990) 337–345.
[17] A. Budai, R. Bock, A. Maier, J. Hornegger, G. Michelson, Robust vessel segmentation in fundus images, *International Journal of Biomedical Imaging*, 2013, (2013) 154860.
[18] R. Hemelings, *et al.*, Artery-vein segmentation in fundus images using a fully convolutional network, *Computerized Medical Imaging and Graphics*, 75, (2019) 101636.
[19] Q. Hu, M. D. Abràmoff, M. K. Garvin, Automated separation of binary overlapping trees in low-contrast color retinal images, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Nagoya, Japan, (2013) 436–443.
[20] J. Orlando, J. Barbosa-Breda, K. Van Keer, M. B. Blaschko, P. J. Blanco, C. A. Bulant, LES-AV dataset, *figshare*, (2020).
[21] E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. Courville, FiLM: Visual Reasoning with a General Conditioning Layer, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, (2018) 483.
[22] S. Zeng, C. Lee, K. Li, B. Xie, O. Fu, H. He, L. Zhu, Y. Lu, F. Cheng, Improve retinal artery/vein classification via channel coupling, *Expert Systems with Applications*, 305, (2026) 130795.

Identification of Elevated Cardiac Troponin from 12-Lead ECGs Using Deep Learning and Exploratory Generalization from Myocardial Infarction Models

1st Pedro A. de A. da Silva
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
pedrotatico1321@gmail.com

2nd Villeneve de O. Soares
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
villeneveoliveira@gmail.com

3rd Emanuel T. de A. Silva
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
emanuel.thiago.2018@gmail.com

4th Gabriel A. de Santana
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
andrade.santana@academico.ifpb.edu.br

5th Ana R. O. de Sousa
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
sousa.raquel@academico.ifpb.edu.br

6th Carlos D. M. Regis
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—Cardiac troponin is the reference biomarker for myocardial injury; however, its measurement depends on laboratory analysis, which may delay clinical decision-making in emergency settings. In contrast, the electrocardiogram (ECG) is inexpensive, non-invasive, and immediately available. This study proposes a one-dimensional convolutional neural network (1D-CNN) for the identification of elevated cardiac troponin using standard 12-lead ECG recordings. ECG signals were preprocessed using a 0.67–40 Hz band-pass filter, resampled to 500 Hz, and used to train the proposed model on a private clinical dataset. In addition, an exploratory experiment evaluated a CNN previously trained for myocardial infarction detection on the Physikalisch-Technische Bundesanstalt XL (PTB-XL) database by directly applying it to the private troponin dataset without fine-tuning or domain adaptation. On the validation subset, the proposed model achieved an accuracy of 73.83%, with a sensitivity of 63.6% and a specificity of 75.0% for the positive class; given the low prevalence of elevated troponin in the cohort (11%), the corresponding positive predictive value was 23.3%. The pretrained myocardial infarction model achieved a lower accuracy of 59.61% on the troponin dataset but retained partial discriminative capability. Because myocardial infarction detection and troponin elevation constitute distinct prediction targets, this observation is described as an exploratory internal association rather than as evidence of successful biomarker transfer. Given the limited size of the cohort and the absence of external validation, these findings should be regarded as hypothesis-generating rather than as evidence of clinical applicability.**

***Index Terms*—Electrocardiogram, Convolutional Neural Networks, Cardiac Troponin, Myocardial Infarction.**

I. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for a substantial proportion of hospital admissions and healthcare expenditures [1]–[3]. Among these conditions, myocardial infarction (MI) represents one of the most critical clinical emergencies, requiring rapid diagnosis and treatment to minimize myocardial damage and improve patient outcomes [4], [5]. Early identification of myocardial injury is therefore essential for effective risk stratification and timely therapeutic intervention [5].

The diagnosis of MI relies on a combination of clinical assessment, electrocardiographic findings, and laboratory biomarkers [4], [5]. Among these biomarkers, cardiac troponins are currently considered the reference biomarker for detecting myocardial injury due to their high sensitivity and specificity [4], [6]. Elevated troponin levels are strongly associated with myocardial necrosis and play a central role in contemporary diagnostic guidelines [4].

In clinical practice, physicians often need to make critical decisions while awaiting laboratory troponin results, particularly during the initial evaluation of patients presenting with chest pain. Although high-sensitivity cardiac troponin assays have significantly reduced turnaround times, blood collection, sample processing, and laboratory analysis still introduce delays that may postpone diagnostic confirmation and therapeutic decisions [6], [7]. Furthermore, because cardiac troponins are structural proteins released into the bloodstream after myocardial cell injury, rather than direct markers of cardiac electrical activity, their elevation follows a time-dependent release pattern after symptom onset [4], [8], [9]. In contrast, electrocardiogram (ECG) abnormalities may be absent, transient, or non-specific during the early stages of myocardial infarction, particularly in non-ST-elevation acute coronary syndromes [5], [10]. Consequently, ECG findings and troponin measurements provide complementary information rather than equivalent diagnostic evidence, reinforcing the need for computational approaches capable of identifying latent ECG patterns associated with subsequent troponin el-

evation [8], [11].

Because the ECG is fast, non-invasive, inexpensive, and widely available, it remains the first-line diagnostic test in the evaluation of patients with suspected acute myocardial infarction [5]. Beyond its established role in detecting arrhythmias, conduction abnormalities, ischemia, and ST-segment elevation, the ECG provides real-time information about the heart's electrical activity that may contain subtle signatures of myocardial injury [5], [12]. However, many of these patterns are difficult to recognize through conventional visual interpretation, particularly in patients with non-specific or atypical presentations [13], [14]. This limitation has motivated the development of artificial intelligence methods capable of automatically extracting complex electrophysiological features from raw ECG signals.

Recent advances in deep learning have substantially improved automatic ECG analysis by enabling the extraction of complex electrophysiological patterns directly from raw signals. In the context of myocardial infarction (MI), convolutional neural networks have consistently achieved excellent diagnostic performance. For example, Alghamdi et al. [15] proposed a transfer learning framework based on VGG-Net that achieved accuracies above 99% for MI detection on the PTB database, demonstrating the feasibility of deep learning for automated ECG interpretation . Subsequently, He et al. [16] introduced a lightweight multi-feature branch convolutional network with a lead-attention mechanism capable of exploiting the complementary information from the 12 ECG leads. Their method achieved 99.63% accuracy under intra-patient evaluation and remained competitive under patient-specific settings, while also supporting deployment on resource-constrained devices.

Although deep learning approaches have demonstrated remarkable performance, important methodological challenges remain. Many studies evaluate their models using intra-patient protocols or relatively homogeneous datasets, which may overestimate real-world performance. Addressing this issue, Moghadam and Asl [17] proposed an interpretable approach based on morphological ECG features and random forests, reporting 96.54% accuracy for MI detection and 80.98% accuracy for MI localization under the more clinically relevant inter-patient evaluation protocol.

More recently, deep learning models have been extended beyond conventional ECG interpretation to estimate laboratory biomarkers directly from the electrical activity of the heart. Chaudhari et al. [11] developed a convolutional neural network using more than 64,000 paired ECG–troponin examinations, demonstrating that elevated cardiac troponin I could be identified directly from standard 12-lead ECGs, with an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.783 for clinically relevant troponin thresholds. Their findings suggest that ECG signals contain latent information associated with biomarker-defined myocardial injury that is not readily apparent through conventional visual interpretation . Subsequent studies have further reinforced the potential of AI-based ECG analysis for identifying myocardial injury and supporting early clinical decision-making [18].

Despite the promising results reported in the literature, important challenges remain. Relatively few investigations have explored the relationship between ECG representations learned from myocardial infarction detection tasks and their applicability to related clinical outcomes such as troponin elevation [11], [18]. Understanding whether knowledge acquired from large public ECG repositories can generalize to independent clinical cohorts remains an important research question [19], [20].

In this context, the present study investigates the use of a one-dimensional convolutional neural network (1D-CNN) for the automatic identification of elevated cardiac troponin from standard 12-lead ECG recordings. The proposed approach was evaluated using a private clinical dataset collected in a hospital environment under approval from the local Research Ethics Committee. ECG signals were preprocessed through band-pass filtering and resampling to ensure compatibility with the adopted deep learning architecture.

Additionally, an exploratory generalization analysis was conducted using a model previously trained for myocardial infarction detection on the PTB-XL database [19]. Unlike conventional transfer learning approaches, no fine-tuning or domain adaptation was performed. The model was directly applied to the private clinical dataset labeled according to troponin levels, allowing the investigation of whether ECG representations learned from infarction-related patterns contain information associated with myocardial injury as reflected by troponin elevation.

II. Materials and Methods

A. Private Clinical Dataset

The experiments were conducted using a private clinical dataset collected in a hospital environment. The study was approved by the local Research Ethics Committee (Approval No. 3.493.714; CAAE: 15824819.2.0000.8069), and all participants provided written informed consent before enrollment.

The dataset comprises standard 12-lead electrocardiographic recordings acquired from patients admitted for clinical evaluation, together with the corresponding cardiac troponin measurements obtained from each patient's laboratory records. Each ECG recording consists of 12 simultaneously acquired leads sampled at 1200 Hz, containing 13,440 samples per lead. An example of a raw ECG recording from the private dataset is shown in Fig. 1, illustrating the signal morphology prior to preprocessing.

High-sensitivity cardiac troponin I (hs-cTnI) was measured using an immunometric assay with a detection limit of 0.004 ng/mL. For the purposes of this study, a binary classification task was defined using the assay's diagnostic threshold corresponding to the 99th percentile. Samples with hs-cTnI concentrations ≥ 0.04 ng/mL were labeled as positive, indicating myocardial injury, whereas values < 0.04 ng/mL were labeled as negative. Although patients were clinically classified as having type 1 myocardial infarction (plaque rupture) or type 2 myocardial infarction (oxygen supply–demand

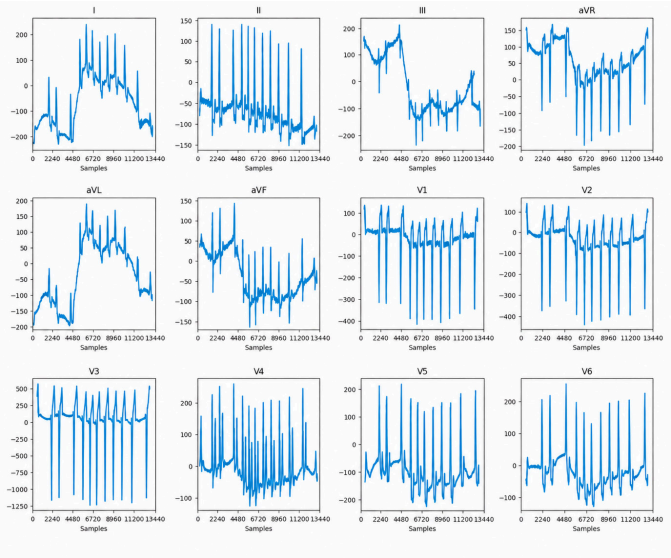


Fig. 1. Example of a raw 12-lead ECG recording from the private clinical dataset before preprocessing. Each recording consists of 12 simultaneously acquired leads sampled at 1200 Hz, containing 13,440 samples per lead.

TABLE I
DISTRIBUTION OF THE PRIVATE CLINICAL DATASET.

Subset	Negative	Positive	Total
Training	214	26	240
Validation	92	11	103
Total	306	37	343

imbalance) according to clinical, laboratory, and imaging criteria, these distinctions were not used as prediction targets in the proposed model.

The private dataset was divided into training and validation subsets at the patient level, ensuring that all recordings from a given patient were assigned to a single subset and that no patient contributed data to both partitions. The training set consisted of 240 ECG recordings, including 214 negative and 26 positive examinations, whereas the validation set comprised 103 recordings, including 92 negative and 11 positive cases. This distribution reveals a marked class imbalance, which reflects the prevalence of patients without elevated troponin in the collected clinical cohort.

The class distribution of the private dataset is summarized in Table I. As can be observed, the dataset is considerably imbalanced, with positive cases representing approximately 11% of the available ECG recordings.

B. PTB-XL Dataset

The second experiment employed the publicly available PTB-XL database [19], one of the largest annotated 12-lead ECG repositories currently available. The database contains 21,837 clinical ECG recordings acquired from 18,885 patients, each consisting of a standard 10-second, 12-lead examination annotated by expert cardiologists according to the SCP-ECG standard. Owing to its size, standardized annotations, and public availability, PTB-XL has become a widely adopted benchmark for developing and evaluating deep learning models for automated ECG analysis.

TABLE II
DISTRIBUTION OF THE PTB-XL DATASET USED FOR MYOCARDIAL INFARCTION CLASSIFICATION.

Subset	Normal	MI	Total
Training	6,351	1,791	8,142
Validation	2,722	768	3,490
Total	9,073	2,559	11,632



Fig. 2. ECG preprocessing pipeline adopted in this study. Raw 12-lead ECG signals sampled at 1200 Hz were filtered using a 0.67–40 Hz band-pass filter, resampled to 500 Hz, normalized to the range $[-1, 1]$, and organized into a 12×5000 input tensor.

For this study, the PTB-XL database was used exclusively to train a CNN for myocardial infarction detection. ECG recordings labeled as *NORM* were assigned to the negative class, whereas recordings corresponding to myocardial infarction (*MI*) were assigned to the positive class. After excluding the remaining diagnostic categories, the resulting dataset comprised 11,632 ECG recordings, including 9,073 normal examinations and 2,559 myocardial infarction cases. The class distribution adopted in this study is summarized in Table II.

C. ECG Preprocessing

The ECG preprocessing pipeline adopted in this study is illustrated in Fig. 2. As shown in the figure, the preprocessing consisted of four sequential stages: band-pass filtering, resampling, amplitude normalization, and tensor organization for CNN input.

Initially, the raw 12-lead ECG recordings, originally sampled at 1200 Hz, were processed using a band-pass filter with cutoff frequencies of 0.67 Hz and 40 Hz. This filtering stage attenuates baseline wander and high-frequency noise while preserving the frequency components that are clinically relevant for cardiac analysis.

Subsequently, the filtered signals were resampled to 500 Hz using polyphase resampling. To ensure a fixed-length input compatible with the PTB-XL dataset and the proposed CNN architecture, the central 5000 samples of each resampled recording were retained, corresponding to approximately 10 s of ECG acquisition. An example of a preprocessed ECG

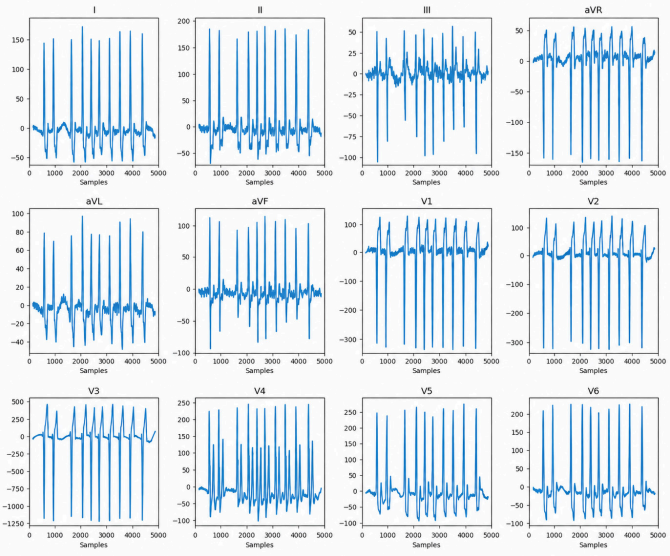


Fig. 3. Example of a 12-lead ECG recording from the private clinical dataset after band-pass filtering (0.67–40 Hz) and resampling to 500 Hz, before amplitude normalization. Each subplot corresponds to one of the 12 simultaneously acquired leads and contains 5000 samples, representing approximately 10 s of ECG acquisition.

recording after band-pass filtering and resampling, but before amplitude normalization, is presented in Fig. 3.

Finally, amplitude normalization was applied independently to each recording by scaling the signal to the interval $[-1, 1]$. After preprocessing, each ECG was represented as a tensor of dimensions 12×5000, corresponding to 12 simultaneously acquired leads with 5000 samples each, which was used as the input to the convolutional neural network. An equivalent preprocessing pipeline was adopted for the PTB-XL dataset. The only difference was that the resampling stage was omitted, since PTB-XL recordings are already provided at a sampling frequency of 500 Hz. Because amplitude normalization was applied on a per-recording basis, no normalization statistics were computed from the private dataset, and no information from the private cohort was used at any stage of the PTB-XL training procedure. Lead ordering followed the standard sequence (I, II, III, aVR, aVL, aVF, V1–V6) in both datasets, and both were represented as 12×5000 tensors corresponding to 10 s of acquisition at 500 Hz. This ensures that the model trained on PTB-XL received inputs of identical dimensionality, lead ordering, and amplitude scale when applied to the private recordings.

D. Convolutional Neural Network

A one-dimensional convolutional neural network (1D-CNN) was designed to process multichannel ECG signals directly, eliminating the need for handcrafted feature extraction. The overall architecture is illustrated in Fig. 4.

The proposed network consists of five one-dimensional convolutional layers with 16, 32, 64, 128, and 256 filters, respectively. Each convolutional layer is followed by Batch Normalization (BatchNorm) and the Rectified Linear Unit

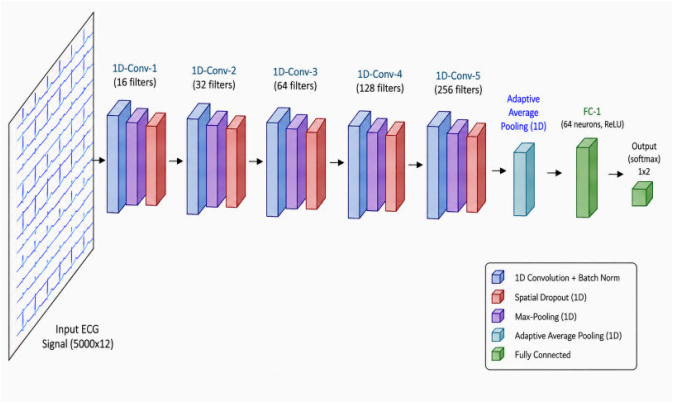


Fig. 4. Proposed one-dimensional convolutional neural network (1D-CNN) architecture for ECG classification.

(ReLU) activation function to improve optimization stability and enhance nonlinear feature learning. A Spatial Dropout layer is applied after each convolutional block to reduce overfitting by randomly suppressing entire feature maps during training. Subsequently, Max-Pooling layers progressively reduce the temporal resolution while preserving the most informative features.

After the final convolutional block, an Adaptive Average Pooling layer aggregates the extracted feature maps into a fixed-length representation, independent of the temporal dimension. This representation is then passed to a fully connected layer with 64 neurons and ReLU activation, followed by a Softmax output layer with two neurons for binary classification.

The resulting architecture is lightweight while maintaining sufficient representational capacity to learn hierarchical temporal features from 12-lead ECG recordings, making it suitable for cardiovascular classification tasks based on ECG signals.

E. Training Configuration

The network was trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16. One-dimensional convolutional kernels of size 4 were used throughout the network, and a Spatial Dropout rate of 0.1 was applied after each convolutional block. To mitigate the effect of class imbalance, the network was optimized using a class-weighted cross-entropy loss in which the weight assigned to each class was inversely proportional to its frequency in the training set. The model trained on the PTB-XL database for myocardial infarction classification was optimized for 50 epochs, whereas the model trained on the private clinical dataset for hs-cTnI classification was optimized for 500 epochs. In both cases, the final model was selected as the checkpoint achieving the lowest validation loss. Predictions were obtained by applying a fixed decision threshold of 0.5 to the Softmax output corresponding to the positive class. All experiments were implemented in PyTorch 2.3.0.

F. Experimental Protocol

Two complementary experiments were performed in this study. The first experiment investigated the generalization capability of ECG representations learned from a clinically related task. For this purpose, the proposed CNN was trained on the PTB-XL database for myocardial infarction detection and subsequently applied directly to the private clinical dataset labeled according to hs-cTnI levels, without performing fine-tuning, domain adaptation, or parameter optimization. Since the target labels differ between the two datasets, this experiment should not be interpreted as a conventional transfer learning evaluation, but rather as an exploratory analysis of whether representations learned for myocardial infarction detection are also informative for identifying elevated troponin levels.

The second experiment consisted of training and validating the proposed CNN directly on the private clinical dataset to identify patients presenting elevated hs-cTnI from standard 12-lead ECG recordings. This experiment was designed to evaluate the predictive performance of a model optimized specifically for the troponin classification task and to compare its performance with that of the exploratory generalization analysis.

G. Performance Evaluation

The predictive performance of the proposed model was assessed using four standard classification metrics: Accuracy, Precision, Recall, and F1-score. Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the overall proportion of correctly classified samples, as expressed in (1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

Precision evaluates the reliability of positive predictions by quantifying the proportion of predicted positive samples that are actually positive, as shown in (2).

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

Recall, also referred to as sensitivity, measures the ability of the model to correctly identify positive cases, as given in (3).

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

The F1-score corresponds to the harmonic mean of Precision and Recall, providing a balanced evaluation metric, particularly in the presence of class imbalance, as formulated in (4).

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

In addition to these quantitative metrics, confusion matrices were analyzed to provide a detailed visualization of the prediction outcomes, highlighting the distribution of true positives, true negatives, false positives, and false negatives. During training, the convergence behavior of the proposed CNN was monitored through the evolution of the training loss and classification accuracy over successive epochs.

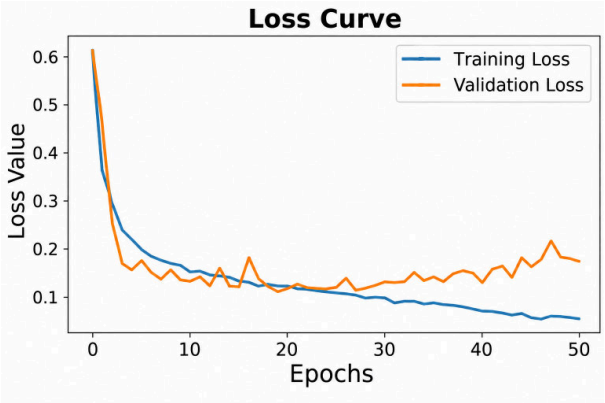


Fig. 5. Training and validation loss during myocardial infarction classification on the PTB-XL dataset.

III. Results and Discussion

This section presents the experimental results in two stages. First, an exploratory analysis evaluates whether ECG representations learned for myocardial infarction detection on the PTB-XL database can generalize to the prediction of elevated hs-cTnI in an independent clinical cohort. Subsequently, the performance of the proposed 1D-CNN trained directly for hs-cTnI prediction on the private clinical dataset is presented and discussed.

A. Exploratory Generalization Analysis Using PTB-XL

Before investigating the relationship between myocardial infarction and troponin elevation, the proposed 1D-CNN was first evaluated on its original task using the PTB-XL database. The model was trained to discriminate myocardial infarction from the normal recordings, and its convergence behavior was monitored through the evolution of the training and validation loss.

Fig. 5 shows the corresponding learning curves. Both training and validation losses decreased during the initial epochs, indicating effective optimization of the network parameters. After approximately 20 epochs, the validation loss stabilized while the training loss continued to decrease gradually, suggesting a slight tendency toward overfitting in the later stages of training. Nevertheless, the validation loss remained stable throughout the optimization process, indicating that the learned representations retained generalization capability on unseen PTB-XL samples.

The quantitative performance obtained on the PTB-XL validation subset is summarized in Table III. The proposed model achieved an overall accuracy of 95.65%, with a precision of 93.91%, recall of 91.64%, and an F1-score of 92.76% for the myocardial infarction class

The confusion matrix obtained on the PTB-XL validation subset is presented in Fig. 6. The model correctly classified

TABLE III
PERFORMANCE OBTAINED FOR MYOCARDIAL INFARCTION CLASSIFICATION ON THE PTB-XL DATASET.

Metric	Value (%)
Accuracy	95.65
Precision (MI class)	93.91
Recall (MI class)	91.64
F1-score (MI class)	92.76

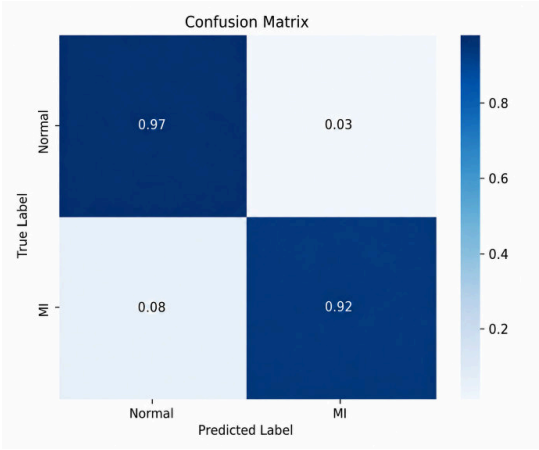


Fig. 6. Confusion matrix obtained for myocardial infarction classification on the PTB-XL validation subset.

TABLE IV
PERFORMANCE OBTAINED BY DIRECTLY APPLYING THE CNN TRAINED ON THE PTB-XL DATASET TO THE PRIVATE TROPONIN DATASET.

Metric	Value (%)
Accuracy	59.61
Macro Precision	55.97
Macro Recall	65.50
Macro F1-score	49.90

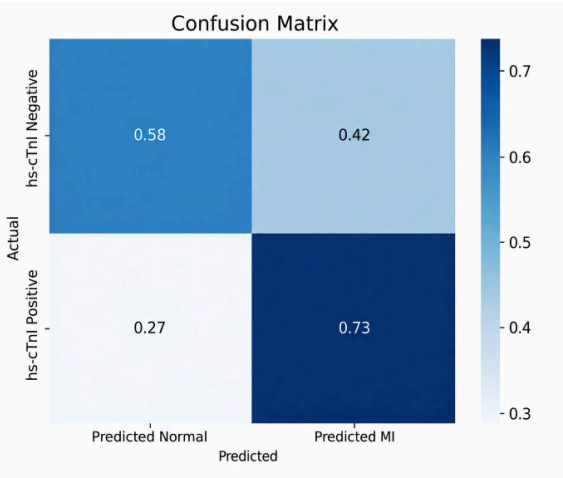


Fig. 7. Confusion matrix obtained by directly applying the CNN trained on PTB-XL to the private clinical dataset labeled according to troponin levels.

approximately 97% of the negative examinations and 92% of the myocardial infarction cases, with false-positive and false-negative rates of only 3% and 8%, respectively. These results confirm that the network learned electrophysiological features associated with myocardial infarction before being evaluated on the independent troponin classification task.

Having established that the proposed CNN effectively learned myocardial infarction-related representations on the PTB-XL dataset, the trained model was subsequently evaluated on the private clinical dataset labeled according to troponin levels, without performing fine-tuning, transfer learning, or any parameter adaptation. This experiment aimed to investigate whether ECG representations learned from myocardial infarction could retain information relevant for identifying elevated troponin in an independent clinical cohort.

The predictive performance is shown in Table IV. The model achieved an overall accuracy of 59.61%, with a macro precision of 55.97%, macro recall of 65.50%, and macro F1-score of 49.90%. As expected, these results are considerably lower than those obtained for myocardial infarction classification on the PTB-XL dataset, reflecting both the domain shift between the public and private datasets and the difference between the prediction tasks.

Fig. 7 presents the corresponding confusion matrix. Since no adaptation was performed after training on the PTB-XL dataset, the observed prediction errors are expected and should be interpreted as a consequence of applying a model optimized for myocardial infarction detection to a different, although clinically related, endpoint.

Although the predictive performance decreased compared with the original myocardial infarction task, the model remained capable of identifying part of the patients presenting elevated troponin. This observation suggests that the representations learned from myocardial infarction are not entirely task-specific and preserve electrophysiological information associated with myocardial injury. However, the performance gap between this experiment and the model trained directly on the private dataset indicates that these representations alone are insufficient for accurate troponin classification without additional adaptation or retraining.

B. Troponin Classification on the Private Dataset

The proposed 1D-CNN was trained using the private clinical dataset to identify patients presenting elevated hs-cTnI directly from 12-lead ECG recordings. Fig. 8 shows the evolution of the training and validation accuracy throughout the optimization process. The validation accuracy increases during the initial epochs and stabilizes after approximately 150 epochs, whereas the training accuracy remains lower and exhibits pronounced oscillation throughout optimization. This pattern is attributable to the small number of positive training examples ($n = 26$) combined with the stochastic regularization applied during training. The reported validation metrics should be interpreted in light of this behavior.

Table V summarizes the performance obtained on the validation subset. The proposed model achieved an overall accuracy of 73.83%. Because the two classes are strongly imbalanced, macro-averaged and positive-class metrics are

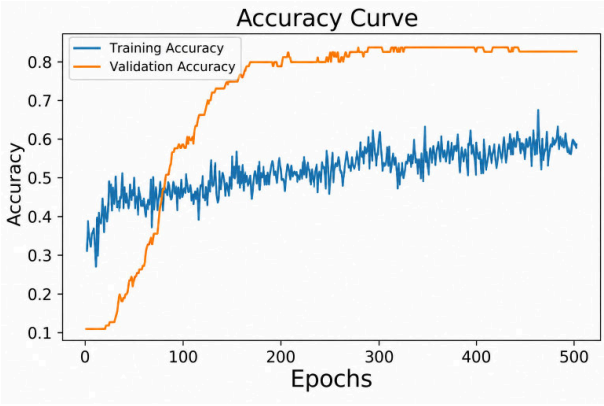


Fig. 8. Training and validation accuracy during the optimization process.

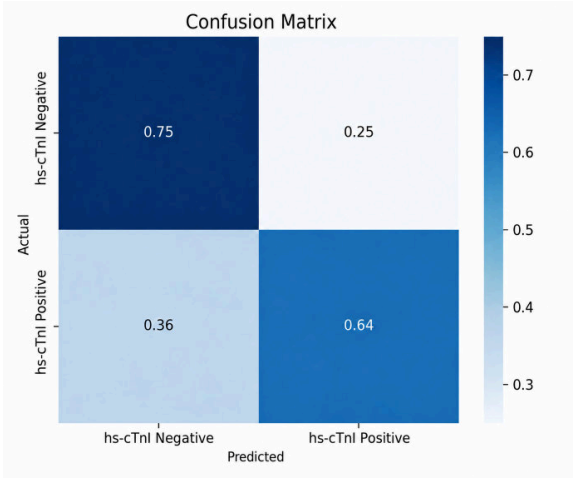


Fig. 9. Confusion matrix of the proposed 1D-CNN for hs-cTnI prediction on the private clinical dataset.

TABLE V
CLASSIFICATION PERFORMANCE OF THE PROPOSED 1D-CNN FOR HS-CTNI PREDICTION ON THE PRIVATE CLINICAL DATASET (VALIDATION SUBSET, $n = 103$; 11 POSITIVE CASES). CONFIDENCE INTERVALS ARE WILSON 95% INTERVALS.

Metric	Value (%)
Accuracy	73.83
Macro Precision	59.01
Macro Recall	69.50
Macro F1-score	58.98
Sensitivity (positive class)	63.6 (35.4–84.8)
Specificity (negative class)	75.0 (65.3–82.7)
Positive predictive value	23.3 (11.8–40.9)
Negative predictive value	94.5 (86.7–97.9)
Positive-class F1-score	34.1

reported separately. The macro-averaged precision, recall, and F1-score were 59.01%, 69.50%, and 58.98%, respectively. For the positive class, which corresponds to elevated hs-cTnI, the model identified 7 of the 11 positive cases (sensitivity 63.6%) while producing 23 false-positive predictions among the 92 negative cases (specificity 75.0%). The resulting positive predictive value was 23.3% and the negative predictive value was 94.5%. This combination of moderate sensitivity and low positive predictive value is a direct consequence of the low prevalence of positive cases in the cohort and should be taken into account when interpreting the overall accuracy.

Fig. 9 presents the confusion matrix obtained on the private dataset. The model correctly classified approximately 75% of the negative cases and 64% of the positive cases. Conversely, 25% of negative examinations were incorrectly classified as positive, while 36% of positive examinations were missed. This behavior is consistent with the relatively lower macro precision reported in Table V, indicating that the classifier tends to generate false-positive predictions while preserving a comparatively higher sensitivity.

The low positive predictive value observed in this experiment is partly a consequence of the class-weighted loss adopted during training, which assigned to the minority class a weight approximately eight times larger than that of the majority class. This choice deliberately favors sensitivity over precision at the fixed 0.5 decision threshold, a trade-off that is appropriate when the intended application is screening rather than diagnostic confirmation, but which necessarily increases the number of false-positive predictions.

From a clinical perspective, this trade-off may be acceptable for screening applications, where failing to identify patients with myocardial injury generally has more serious consequences than requesting additional confirmatory laboratory testing. Considering that the proposed model relies exclusively on ECG recordings, without incorporating demographic information, clinical variables, or laboratory measurements, the obtained results suggest that standard 12-lead ECGs contain discriminative electrophysiological patterns associated with elevated hs-cTnI concentrations. These findings are consistent with recent studies demonstrating that deep learning models can extract latent ECG representations correlated with biochemical markers of myocardial injury [11], [18].

The lower performance observed for hs-cTnI prediction compared with myocardial infarction classification can be explained by the intrinsic differences between the two prediction tasks. In the PTB-XL experiment, the model learned to identify myocardial infarction directly from ECG patterns, whereas in the private dataset the target variable corresponded to a biochemical marker rather than an electrocardiographic diagnosis. Elevated hs-cTnI reflects myocardial cell injury and may occur in patients presenting subtle, non-specific ECG findings, particularly during the early stages of disease or in non-ischemic conditions associated with myocardial injury. Consequently, the relationship between ECG morphology and troponin elevation is inherently more indirect, making the prediction task substantially more challenging.

Interestingly, although the direct application of the PTB-XL model produced inferior performance, the model trained specifically for hs-cTnI substantially improved all evaluation metrics. This result suggests that, although myocardial infarction and troponin elevation share common electrophysiological characteristics, task-specific optimization is required to

fully capture the ECG patterns associated with biochemical evidence of myocardial injury.

Although a direct quantitative comparison is not possible because previous studies primarily reported AUC values rather than accuracy, the present results are consistent with the growing evidence that deep learning models can identify elevated cardiac troponin from standard 12-lead ECG recordings. Chaudhari et al. [11] reported an AUC of 0.783 using more than 64,000 paired ECG–troponin examinations, whereas the present study was conducted on a considerably smaller private clinical dataset. Despite these differences, both studies support the existence of ECG patterns associated with biochemical evidence of myocardial injury.

C. Limitations

Several limitations should be considered when interpreting these results. The private cohort comprised 343 recordings from a single center, of which only 37 were positive, and the evaluation subset contained 11 positive cases. Estimates derived from so few positive examples carry wide confidence intervals, and the reported metrics should be regarded as imprecise. No independent test partition was available, so all reported values correspond to internal validation and are subject to optimistic bias. No external validation on an independent cohort was performed, which limits any assessment of the generalizability of the proposed model.

The binary label was defined by a single hs-cTnI measurement exceeding the 99th-percentile threshold. This definition does not incorporate the rise-and-fall kinetics required by the fourth universal definition of myocardial infarction [4] and therefore captures myocardial injury of any etiology rather than acute myocardial infarction specifically. In addition, type 1 and type 2 myocardial infarction were collapsed into a single positive class despite differing pathophysiological mechanisms and electrocardiographic substrates.

Finally, the model outputs were not calibrated, and no comparison was performed against clinical risk scores or against alternative architectures incorporating attention mechanisms or data augmentation. Such comparisons would be required to establish whether the observed performance derives from the proposed approach or from the general capacity of convolutional architectures applied to ECG signals.

IV. Conclusion

This study presented a one-dimensional convolutional neural network for the identification of elevated cardiac troponin using standard 12-lead ECG recordings. Without relying on handcrafted feature extraction or additional clinical variables, the proposed model achieved an accuracy of 73.83% on the validation subset of a private clinical dataset, with a sensitivity of 63.6% and a positive predictive value of 23.3% for the positive class. These results are consistent with the presence of discriminative information associated with myocardial injury in the ECG, although the small number of positive cases limits the precision of the estimate.

A second exploratory experiment investigated whether representations learned from myocardial infarction detection could be transferred to a related clinical endpoint. A CNN trained on the PTB-XL database for myocardial infarction classification achieved 95.65% accuracy on its original task and, when directly applied to the private troponin dataset without fine-tuning, obtained an accuracy of 59.61%. Although a reduction in predictive performance was observed, the model preserved partial discriminative capability, suggesting that myocardial infarction and troponin elevation share common electrophysiological characteristics encoded in ECG signals.

Overall, within this cohort, the model trained directly for troponin prediction outperformed the model transferred from myocardial infarction classification. Because myocardial infarction detection and troponin elevation constitute distinct prediction targets, the performance obtained on PTB-XL should not be interpreted as evidence of successful biomarker transfer. The observed retention of partial discriminative capability is better described as an exploratory internal association that requires confirmation in larger, externally validated cohorts.

Future work will focus on validating the proposed framework using larger multicenter cohorts, investigating transfer learning and domain adaptation strategies for improving cross-dataset generalization, and incorporating explainable artificial intelligence methods to enhance the clinical interpretability of the learned ECG representations.

References

[1] World Health Organization, "Cardiovascular diseases (cvds)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2025. Accessed: 2026-06-18.

[2] G. A. Roth, G. A. Mensah, C. O. Johnson, *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020.

[3] A. Timmis, P. Vardas, N. Townsend, *et al.*, "European society of cardiology: Cardiovascular disease statistics 2021," *European Heart Journal*, vol. 43, no. 8, pp. 716–799, 2022.

[4] K. Thygesen, J. S. Alpert, A. S. Jaffe, *et al.*, "Fourth universal definition of myocardial infarction (2018)," *Circulation*, vol. 138, no. 20, pp. e618–e651, 2018.

[5] R. A. Byrne, X. Rossello, J. J. Coughlan, *et al.*, "2023 esc guidelines for the management of acute coronary syndromes," *European Heart Journal*, vol. 44, no. 38, pp. 3720–3826, 2023.

[6] Y. Sandoval and A. S. Jaffe, "Type 2 myocardial infarction: Jacc review topic of the week," *Journal of the American College of Cardiology*, vol. 73, no. 14, pp. 1846–1860, 2019.

[7] C. Mueller, "Biomarkers and acute coronary syndromes: An update," *European Heart Journal*, vol. 35, no. 9, pp. 552–556, 2014.

[8] T. Reichlin, W. Hochholzer, S. Bassetti, *et al.*, "Early diagnosis of myocardial infarction with sensitive cardiac troponin assays," *New England Journal of Medicine*, vol. 361, no. 9, pp. 858–867, 2009.

[9] J. W. Pickering, J. M. Young, P. M. George, *et al.*, "Early kinetic profiles of troponin i and t measured by high-sensitivity assays in patients with myocardial infarction," *Clinica Chimica Acta*, vol. 505, pp. 15–25, 2020.

[10] H. Basit, A. Malik, and M. R. Huecker, "Non-st-segment elevation myocardial infarction," 2023. StatPearls [Internet].

[11] G. R. Chaudhari, J. J. Mayfield, J. P. Barrios, S. Abreau, R. Avram, J. E. Olgin, and G. H. Tison, "Deep learning augmented ecg analysis to identify biomarker-defined myocardial injury," *Scientific Reports*, vol. 13, p. 3364, 2023.

[12] A. L. Goldberger, Z. D. Goldberger, and A. Shvilkin, *Goldberger's Clinical Electrocardiography: A Simplified Approach*. Elsevier, 9 ed., 2018.

[13] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, *et al.*, "An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, 2019.

[14] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[15] A. Alghamdi, M. Hammad, H. Ugail, A. Abdel-Raheem, K. Muhammad, H. S. Khalifa, and A. A. Abd El-Latif, "Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14913–14934, 2024.

[16] Z. He, Z. Yuan, P. An, J. Zhao, and B. Du, "Mfb-lann: A lightweight and updatable myocardial infarction diagnosis system based on convolutional neural networks and active learning," *Computer Methods and Programs in Biomedicine*, vol. 210, p. 106379, 2021.

[17] S. R. Moghadam and B. M. Asl, "Automatic diagnosis and localization of myocardial infarction using morphological features of ecg signal," *Biomedical Signal Processing and Control*, vol. 83, p. 104671, 2023.

[18] L. Hilgendorf, P. Petursson, V. Gupta, *et al.*, "Predicting troponin biomarker elevation from electrocardiograms using a deep neural network," *Open Heart*, vol. 11, no. 2, p. e002937, 2024.

[19] P. Wagner, N. Strodthoff, R.-D. Bousseljot, *et al.*, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, p. 154, 2020.

[20] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2021.

Identification of Genotoxic Cells in the Micronucleus Test Using YOLOv7 Neural Networks

Gabriel Cunha Cardoso
CECS
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
g.cunha@aluno.ufabc.edu.br

Carolina Benetti
CECS
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
c.benetti@ufabc.edu.br

Andrea Cecilia Dorion Rodas
CECS
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
andrea.rodas@ufabc.edu.br

Anderson Gabriel Santiago
CECS
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
gabriel.santiago@ufabc.edu.br

***Abstract*—The micronucleus test is one of several recommended tests for assessing genetic damage caused by chemicals under development for use as biomaterials, cosmetics, or drugs. The test involves analyzing cultured cells *in vitro* treated with the compounds and determining the frequency of cells containing one, two, or multiple nuclei. This work presents the implementation of a YOLOv7 neural network architecture for the automatic identification of cells in micronucleus assays, which are used to assess genotoxicity. To facilitate use by non-specialists, two prototypes were developed that directly integrate the original YOLOv7 `detect.py` script without altering its inference logic: (i) a Python/PyQt desktop application featuring a thumbnail gallery, a progress bar, and interactive zoom, converted into a standalone executable using PyInstaller; and (ii) a Flask-based web service capable of handling multiple uploads and allowing users to configure input size and confidence thresholds. On the held-out test set, the model achieved a precision of 0.855, recall of 0.887, and *mAP@0.5* of 0.905 for the three cell classes represented in that split. Neither impurity nor micronucleus instances were present in the test set, preventing their quantitative evaluation; qualitative inference nevertheless showed missed micronucleus detections. Performance tests indicated that CPU-based processing of large batches takes 5 to 10 minutes, highlighting the advantages of using a GPU or optimized model variants. The proposed solutions enhance the accessibility of genotoxicity analyses, offering potential applications in both laboratory settings and remote services.**

***Index Terms*—convolutional neural networks, YOLOv7, genotoxicity, micronucleus assay, computer vision**

I. Introduction

Genotoxicity refers to the ability of chemical or physical agents to cause damage to the genetic material of cells, resulting in gene mutations, chromosomal breaks, or changes in chromosome number. Among the methods used to assess this risk, the *in vitro* micronucleus assay stands out for its simplicity and sensitivity; it is recommended by OECD Guideline 487 [1] and ISO standard 10993-3 [2] for the evaluation of pharmaceuticals, cosmetics, and medical devices. In the classic protocol, cytochalasin B blocks cytokinesis following mitosis, causing cells that have completed nuclear division to remain joined as binucleated cells; The cut off of the test is the binuclear counting once the cytochalasin B is used to keep the cells as binocular. The presence of mononuclear or multinuclear becomes a data which indicates the quality of the cells in the test. The counting of the mucronucleus in binuacleates cells, the rare event in the cells, is sensitive enough to characterize the compound as genotoxics, once the micronucleous is a fragment of the main cell DNA and can be more than one in the same cell, depending on the aggressivity of the chemical compound. [1] (as shown in Fig. 1) — are counted.

The cells observed in the assay fall into five annotation classes — mononuclear, binuclear, multinuclear, impurity, and micronucleus — of which the first four correspond to distinct *biological* cell categories, while *impurity* is a non-cell artifact class (debris, stain residue, or out-of-focus objects) used only to keep the detector from misfiring on non-cellular structures; it carries no biological score. Among the four cell categories, event counting is performed only on binucleated cells, as they confirm that recent cell division has occurred. Furthermore, the ratio of binucleated to mononucleated cells provides a measure of the assay's cytotoxicity and proliferative efficiency [1]. Seager et al. [3] demonstrated the high reproducibility of this assay across human cell lines and highlighted the potential for image-processing-based automation to enhance the efficiency and consistency of genotoxicity assessment, a process currently performed primarily by manual means.

In recent years, convolutional neural networks have achieved high performance in object detection tasks. Modern algorithms such as the R-CNN (*Region-based CNN*) and YOLO (*You Only Look Once*) families have revolutionized the field by enabling the identification and localization of multiple objects with high accuracy; in particular, the YOLO family stands out for performing single-shot detection, unifying class and position prediction into a single inference step. This yields significantly higher speed compared to two-stage approaches without drastically sacrificing precision. YOLOv7 [4] repre-

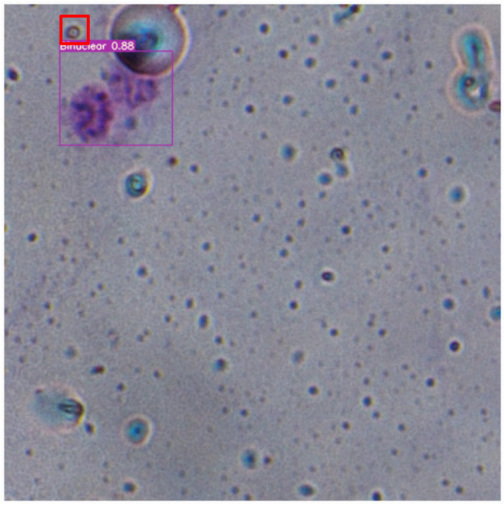


Fig. 1. Binucleated cell with a micronucleus (framed in red).

sents the current state of the art in this family, incorporating techniques such as Extended Efficient Layer Aggregation Networks (E-ELAN) and a "bag-of-freebies" (a set of trainable training adjustments) that boost accuracy without additional inference cost. Its ability to capture patterns at multiple scales makes it well-suited for microscopy images, where cells and micronuclei are small, appear at high densities, and exhibit subtle contrast against the background. Prior to this work, no literature was found specifically on the use of *object-detection* neural networks for cell identification under genotoxic conditions in the micronucleus assay; Section IV situates this work relative to other automated approaches to micronucleus scoring.

Given this scenario, the general objective of this work is to implement and validate a YOLOv7 architecture for the automatic detection and classification of cells in micronucleus assay images. A specific objective is the development and validation of software solutions — a desktop application and a web service — that utilize the trained model to provide accessible, reliable, and efficient interfaces for non-specialist users.

II. Methodology

A. Dataset

The dataset consists of 50 optical microscopy images in TIFF format (2560×1920 pixels), acquired under genotoxicity conditions. Each cell in the images was classified into one of the five classes introduced in Section I (mononuclear, binuclear, multinuclear, impurity, or micronucleus). The micronucleus class was the most significant, as it directly indicates genotoxic damage, yet it was also the rarest by a wide margin: only 2 micronucleus instances were annotated in the original, pre-augmentation dataset, against a much larger pool of mononuclear, binuclear, and multinuclear instances that made up the bulk of the annotated cells.

This extreme imbalance is the central limiting factor discussed throughout the Results and Discussion sections: with only 2 original examples, even a 10× augmentation (Section II-C) yields around 22 micronucleus instances in total, an order of magnitude below the hundreds typically required for a deep detector to learn a reliable visual signature, which is consistent with the model's failure to reach usable confidence for this class (Section III-A).

B. Preprocessing and Annotation

Due to the computational cost imposed by the original resolution, the images were cropped into 512×512-pixel blocks using a Python script developed by the authors; this script also proportionally adjusted the corresponding bounding boxes, resulting in 200 images. The annotations were then manually reviewed using the labelImg tool [5], following the standard YOLOv7 format — class and coordinates (x, y, width, and height) normalized relative to the image size [6] — with classes mapped as 0: Mononuclear, 1: Binuclear, 2: Impurity, 3: Micronuclear, and 4: Multinuclear. Crops containing no annotated cells were discarded, resulting in a final set of 146 images.

C. Balancing and Splitting the Dataset

The Micronucleus class was represented by only two examples in the original dataset, making the direct training of a deep learning model unfeasible [7], [8]. To mitigate this imbalance, data augmentation was applied using the Albumentations library [9] — including horizontal and vertical flips, random brightness and contrast adjustments, rotations, and scaling shifts — generating ten new samples per original image, with bounding boxes automatically adjusted to the applied transformations. The dataset was then split into training (80%), validation (10%), and testing (10%) sets by original source image rather than on the pooled augmented set: all ten augmented variants of a given original micrograph were kept together in a single split, so no augmented copy of a training image could appear in validation or testing. This directly addresses the risk of data leakage between splits, while still maintaining independent subsets for model tuning and evaluation [10].

D. Evaluation Metrics

The detector's performance was monitored using four metrics common to object detection tasks. The box loss, which YOLOv7 calculates using Complete Intersection over Union (CIoU), penalizes errors in predicting the position and size of bounding boxes:

$$\mathrm{CIoU} = 1 - \left(\mathrm{IoU} - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v \right), \qquad (1)$$

where IoU is the ratio of the intersection area to the union area of the predicted and ground-truth boxes, $\rho^2(b, b^{gt})$ is the squared Euclidean distance between their centers, c^2 is the squared diagonal of the smallest rectangle enclosing them, and αv penalizes the difference in the width-to-height aspect ratio between the boxes. Precision measures the proportion of correct detections among all detections made, while recall

TABLE I
TRAINING HYPERPARAMETERS

Hyperparameter	Value
Effective input size (training and evaluation)	544×544 px
Batch size	8
Epochs	180
Initial learning rate (lr_0)	0.01
Momentum	0.937
Weight decay	0.0005
Warm-up epochs	3

measures the proportion of actual objects the model successfully detected:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

where TP, FP, and FN are, respectively, true positives, false positives, and false negatives. The third metric, mean Average Precision with an IoU threshold of 0.5 (*mAP@0.5*), summarizes the average precision across all C classes of the problem:

$$\text{mAP@0.5} = \frac{1}{C} \sum_{i=1}^{C} AP_i^{IoU=0.5}, \tag{3}$$

where AP_i is the area under the precision-recall curve for class i. High values across all three metrics indicate a detector that correctly localizes and classifies cells with few false positives and false negatives.

E. Model Training

The model was trained starting from the `yolov7.pt` pre-trained weights, using the official YOLOv7 implementation [4], [11] on a Google Colaboratory instance equipped with an NVIDIA Tesla T4 GPU — an approach that reduced the duration of each epoch to a few minutes, compared to the tens of hours estimated for an equivalent run on a local CPU. Table I summarizes the key hyperparameters used, which were defined based on the standard YOLOv7 configuration file for training from scratch.

Throughout training, box loss and the metrics described in Section II-D were monitored on the validation split. The command requested 518×518-pixel input, which YOLOv7 automatically adjusted to 544×544 pixels to satisfy the model's 32-pixel maximum stride. The held-out test set was evaluated with `test.py` using the same effective 544×544 resolution, a batch size of 8, a confidence threshold of 0.001, a non-maximum-suppression IoU threshold of 0.65, and the `--task test` option. Thus, the reported validation and test metrics use the same effective input dimensions. The `detect.py` script was used separately, with a confidence threshold of 0.5, only to generate the qualitative examples shown in Fig. 3.

III. RESULTS

A. Detection Performance

Fig. 2 shows the evolution of the metrics throughout training. The training box loss dropped from approximately 0.07 to 0.025 (Fig. 2a), while validation *mAP@0.5* exceeded 0.80 after the 80th epoch and stabilized near 0.85–0.90 (Fig. 2b); validation precision and recall rose rapidly during the first 50 epochs and stabilized between 0.75 and 0.90 (Fig. 2c-d). At the end of 180 epochs, validation precision, recall, *mAP@0.5*, and *mAP@0.5:0.95* were 0.937, 0.756, 0.869, and 0.588, respectively. On the held-out test set, the corresponding aggregate values were 0.855, 0.887, 0.905, and 0.603 for the three cell classes represented in that split. The test result therefore did not show a performance decrease relative to validation, although both subsets were small.

TABLE II
HELD-OUT TEST-SET RESULTS BY REPRESENTED CLASS

Class	Inst.	P	R	mAP@0.5
Mononuclear	5	0.788	0.751	0.747
Binuclear	11	0.901	0.909	0.973
Multinuclear	3	0.876	1.000	0.995
All represented classes	19	0.855	0.887	0.905

Neither the validation nor the test split contained annotated micronucleus instances, so quantitative performance for this class could not be established. Qualitative inference nevertheless produced missed micronucleus detections, which is consistent with the scarcity of original examples for this class (Section II-A) even after data augmentation.

Fig. 3 presents qualitative examples of the detections generated over the test set. Panels (a)–(c) confirm that the model correctly localizes cells even in scenarios involving overlap and low contrast; panel (d) illustrates the model's main failure mode: it correctly detects the parent binucleated cell (magenta box) but does not flag the micronucleus itself (red box, added manually to mark the missed event), which is the direct, visual counterpart of the class-imbalance limitation discussed in Section II-A. Table II summarizes the quantitative test-set results by class.

No impurity or micronucleus instances were present in the test split; therefore, quantitative metrics could not be calculated for these classes. This absence is why Table II reports only the three represented cell classes.

B. Software Prototype

Based on the trained model, two user interfaces were developed (Fig. 4). The desktop application, built with Python/PyQt, incorporates the original `detect.py` script as a subroutine without altering its inference logic; it allows the user to define image size and the confidence threshold, mirroring command-line functionality. An indeterminate progress bar keeps the user informed during inference, and results are displayed in a thumbnail gallery with interactive zooming — a feature particularly useful given the high resolution of microscopy images. These interface choices align with Nielsen's usability heuristics [12] and a clean visual design that avoids irrelevant elements, highlighting only the actual detections. The program was converted into a standalone executable using PyInstaller in "one-folder" mode [13], eliminating the need for the end-user to install dependencies. Meanwhile, the web service — built

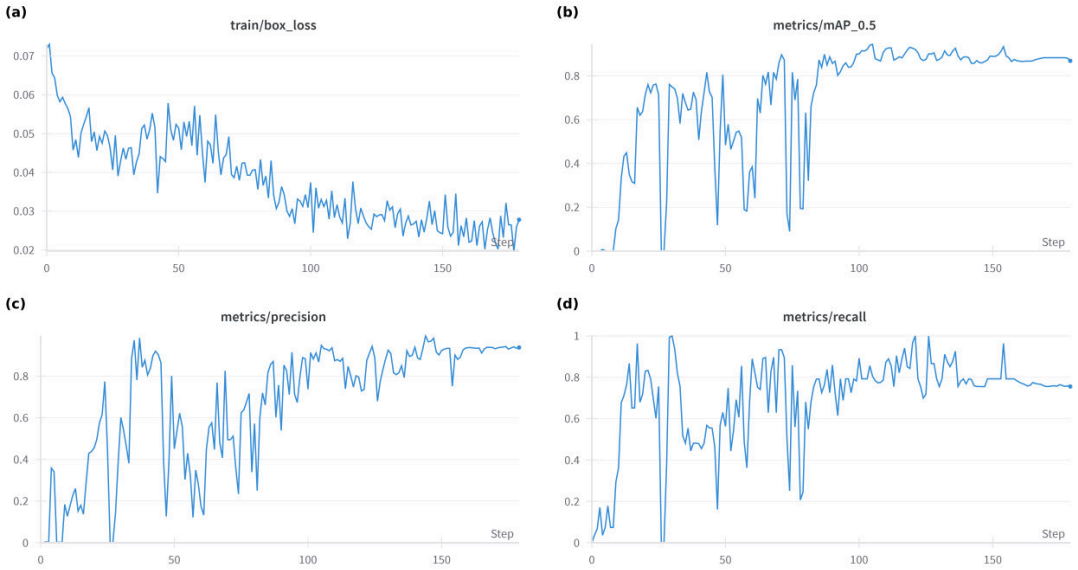


Fig. 2. Training metrics per epoch, tracked on the validation split: (a) *box loss*; (b) *mAP@0.5*; (c) *precision*; (d) *recall*. mAP@0.5 and precision both stabilize above 0.85 by epoch $\sim$100, while recall settles closer to 0.75–0.80, indicating the model trades a small amount of recall for higher precision once training converges.

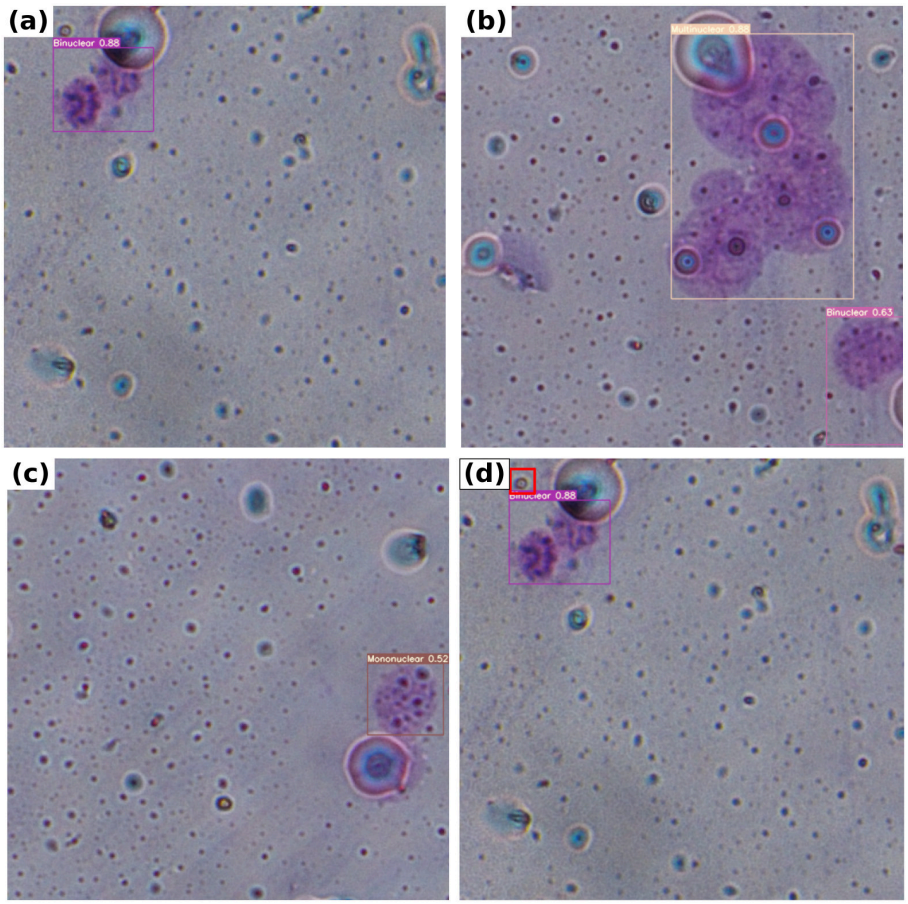


Fig. 3. Qualitative examples of predictions on the test set. (a)–(c) Correct detections of binuclear, multinuclear, and mononuclear cells. (d) A representative failure case: the parent binucleated cell is correctly detected (magenta box, "Binuclear 0.88") but the micronucleus itself is missed by the model (red box, added manually to highlight the miss).

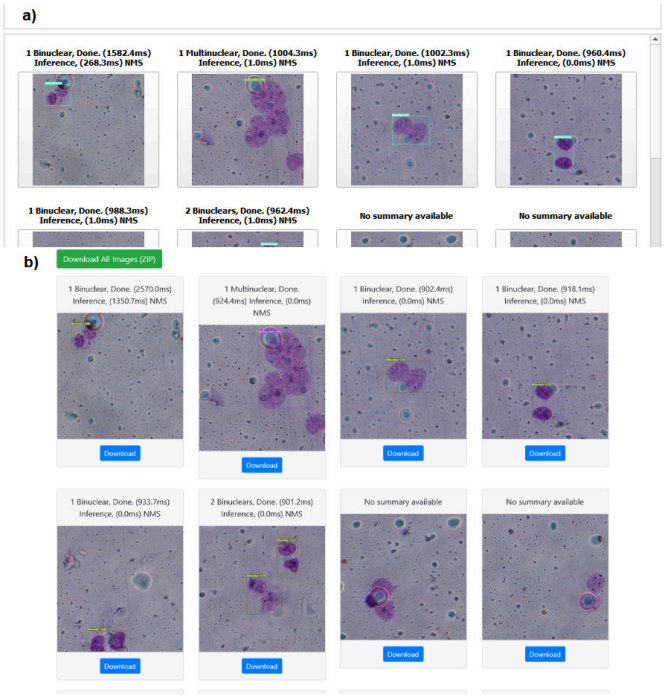


Fig. 4. Detection results as displayed to the end user: (a) desktop application; (b) web service. Both panels show the result gallery produced by the corresponding interface.

using Flask and Jinja2 templates [14] — accepts simultaneous upload of multiple images via a form, supporting multiple files and configurable parameters for input size and confidence threshold, and returns a dynamic HTML page featuring a gallery of annotated images and a batch-download option. Both solutions use the same trained weights and inference logic as the original script, fully preserving the model's accuracy and demonstrating its viability for deployment either locally or as a web-accessible service, as illustrated by the side-by-side detection outputs in Fig. 4.

C. Computational Performance

On hardware without a dedicated GPU, CPU processing proved to be the main bottleneck. In the desktop application, a single image was processed in a few seconds, but the sequential analysis of large batches — such as the dozens of crops obtained from an original image — proved time-consuming: a batch of 18 cropped images, or a single image of 2560×1920 pixels, took 5 to 10 minutes to process, highlighting practical limitations in high-volume data scenarios. In the web service, this limitation is aggravated by the memory and CPU-time restrictions imposed by free or low-cost hosting platforms, such as the one considered for this project (Render), which do not offer GPU acceleration; in practice, deploying the model in this type of environment led to slow responses and instability under large requests. The size of the model itself (tens of millions of parameters and a weights file of hundreds of megabytes) further aggravates load time and memory consumption in these environments, unless paid services with hardware-acceleration support are used.

IV. Discussion

Comparing this detector to other automated approaches to micronucleus scoring helps put its performance in context. Much of the prior automation work has relied on classification-oriented deep learning applied to imaging flow cytometry rather than on object detection over microscopy fields: Wills et al. scaled a CNN-based classifier ("DeepFlow") for the cytokinesis-block micronucleus assay across three independent laboratories, showing that dose-response conclusions matched manual scoring despite differences in imaging site [15]. In the flow citometer technique the cell membrane is broken and only the nucleus with the DNA is analyzed. By this technique, the different classes of cells are not identified and the quality of the cells for the test is not trackable causing misinterpretation of the genotoxicity [1]. More recently, object-detection architectures purpose-built for micronucleus detection have appeared in the literature: Wei et al. proposed STD-YOLOv7, a YOLOv7-based detector combining a CA attention mechanism and a Res-ACmix module specifically tuned for micronucleus morphology, trained on a purpose-built dataset expanded to 13,309 images [16] — roughly three orders of magnitude larger than the two original micronucleus examples available in the present work. This contrast is informative rather than discouraging: it points to dataset scale for the rare class, rather than to a shortcoming of the YOLOv7 architecture itself, as the main gap between the two studies. It reframes the present result: a general-purpose YOLOv7 model, without any micronucleus-specific architectural changes, already reaches full detection performance on the three well-represented cell classes, and closing the remaining gap to specialized micronucleus detectors is plausibly a matter of dataset scale for that one rare class rather than of the underlying detection approach.

The comparison between the two interfaces also highlights a trade-off between accessibility and performance. The desktop version is more appropriate for individual use or in laboratory environments, being able to exploit local hardware resources, including GPUs when available, and mitigating the impact of waiting through continuous visual feedback; the gallery structure, with summarized labels above each thumbnail, facilitated quick interpretation of results, partially compensating for the waiting time inherent to CPU execution. The web version stands out for its convenience in not requiring local installation, but it depends on a more robust infrastructure to achieve acceptable performance in production: without a GPU, the inference throughput drops significantly relative to the hundreds of frames per second reported for YOLOv7 in optimized scenarios [4]. Model optimization techniques — such as quantization or the use of lighter variants, for example YOLOv7-tiny [17] — emerge as natural ways to reduce this bottleneck, as well as the adoption of asynchronous processing queues (e.g., Celery/Redis) in the web back end, in order to avoid timeouts and improve scalability under load.

V. Conclusion

This work implemented and validated a YOLOv7 model for the automatic detection of cells in genotoxicity assays, integrated into two accessible interfaces — desktop and web — that fully preserve the original model's accuracy. The detector demonstrated consistent performance in identifying mononuclear, binuclear, and multinuclear cells but failed to achieve sufficient reliability for detecting the rare micronucleus event; this limitation is attributed to the scarcity of examples of this class in the original dataset, even after data augmentation, and is consistent with the much larger training sets used by specialized micronucleus-detection architectures [16]. Nevertheless, the proposed solutions demonstrate the feasibility of democratizing genotoxicity analysis, with potential applications in laboratories and remote services.

Future work should prioritize expanding the dataset for the micronucleus event — including through machine learning techniques specifically aimed at this rare event — as well as adopting optimized variants of the model (e.g., YOLOv7-tiny, post-training quantization) to enable its execution on the CPU. In the web service, introducing asynchronous processing (e.g., queues with Celery/Redis) would avoid timeouts under load, and migrating the back end from Flask to Django would make it easier to create user accounts and store individualized results. Finally, exposing a RESTful API would allow the detector to be integrated with laboratory management systems and automated image analysis flows, expanding the reach of the solution to third-party batch processing scenarios.

References

[1] "Test no. 487: In vitro mammalian cell micronucleus test," Organisation for Economic Co-operation and Development (OECD), Tech. Rep., 2014.

[2] "ISO 10993-3:2003 - biological evaluation of medical devices – part 3: Tests for genotoxicity, carcinogenicity and reproductive toxicity," International Organization for Standardization, Tech. Rep., 2003.

[3] A. L. Seager, U.-K. Shah, K. Brusehafer, J. Wills, B. Manshian, K. E. Chapman, A. D. Thomas, A. D. Scott, A. T. Doherty, S. H. Doak, G. E. Johnson, and G. J. S. Jenkins, "Recommendations, evaluation and validation of a semi-automated, fluorescent-based scoring protocol for micronucleus testing in human cells," *Mutagenesis*, vol. 29, no. 3, pp. 155–164, 2014.

[4] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *arXiv preprint arXiv:2207.02696*, 2022.

[5] Tzutalin, "labelImg: Graphical image annotation tool," 2015, accessed: Oct. 24, 2024. [Online]. Available: https://github.com/HumanSignal/labelImg

[6] Roboflow, "What is the yolov7 pytorch txt annotation format?" 2022, accessed: Oct. 24, 2024. [Online]. Available: https://roboflow.com/formats/yolov7-pytorch-txt

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[8] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[9] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," p. 125, 2020. [Online]. Available: https://albumentations.ai/

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] WongKinYiu, "YOLOv7 GitHub repository," https://github.com/WongKinYiu/yolov7, 2022, accessed: Apr. 14, 2025.

[12] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, 1994, pp. 152–158.

[13] PyInstaller Development Team, *PyInstaller Documentation (Version 6.13.0)*, https://pyinstaller.org/en/stable/, 2023.

[14] M. Grinberg, "Handling file uploads with Flask," https://blog.miguelgrinberg.com/post/handling-file-uploads-with-flask, 2020.

[15] J. W. Wills, J. R. Verma, B. J. Rees, D. S. G. Harte, Q. Haxhiraj, C. M. Barnes, R. Barnes, M. A. Rodrigues, M. Doan, A. Filby, R. E. Hewitt, C. A. Thornton, J. G. Cronin, J. D. Kenny, R. Buckley, A. M. Lynch, A. E. Carpenter, H. D. Summers, G. E. Johnson, and P. Rees, "Inter-laboratory automation of the in vitro micronucleus assay using imaging flow cytometry and deep learning," *Archives of Toxicology*, vol. 95, no. 9, pp. 3101–3115, 2021.

[16] W. Wei, Y. Leng, L. Cao, and Y. Wang, "STD-YOLOv7: A small target detector for micronucleus based on YOLOv7," *Biomedical Signal Processing and Control*, vol. 106, p. 107810, 2025.

[17] WongKinYiu et al., "Discussion on YOLOv7 model size (issue #1402 on GitHub)," https://github.com/WongKinYiu/yolov7/issues/1402, 2023.

Image-Based Evaluation System for Paraná State Military Physical Fitness Test: 12-Minute Run

Diego Marques De Souza
PPGEB, Universidade Tecnológica Federal do Paraná
Polícia Militar do Paraná, Curitiba, Brazil
ORCID: 0009-0006-5150-3371

Lucas Daniel Ortiz Arias
DAINF/CPGEI, Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0009-0005-3393-4445

Mauro Aloisio Smaniotto
CPGEI, Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0009-0003-9507-2022

Fabio Kurt Schneider
DAELN/CPGEI/PPGEB, Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0001-6916-1361

Julio Cesar Bassan
PPGEB/PPGEF, Universidade Tecnológica Federal do Paraná
Curitiba, Brazil
ORCID: 0000-0002-6910-9077

***Abstract*—This paper presents an image-based system for automatically estimating passage times during the 12-minute running test, a component of the Physical Fitness Test adopted by the Parana State Military Police. The method combines convolution-based image processing and Kalman filter-based multi-object tracking to identify runners and estimate passage times at predefined reference points. Performance was evaluated using 405 repeated passage records from three volunteers across three videos (135 records per runner; 45 per runner in each video). The software was compared with a reference evaluator and with the mean of nine additional human evaluators. Bland–Altman analysis identified statistically significant negative biases in the software comparisons. Inter-rater agreement among all ten evaluators was quantified using a two-way random-effects intra-class correlation coefficient with absolute agreement [ICC(2,1) = 0.99994; 95% CI: 0.99991–0.99996]. These findings characterize agreement within the evaluated recordings and do not establish participant-level validation, test-retest reliability, or performance under different camera and environmental conditions. Further validation is required before routine or official use.**

***Index Terms*—Computer Vision, Image Processing, Multi-Object Tracking, Military Physical Fitness Test, 12-Minute Running Test, Bland–Altman Analysis**

I. Introduction

Indirect estimation of maximal oxygen uptake (VO_2max) is one of the most widely used approaches for assessing aerobic fitness in military forces and sports settings, as it enables the evaluation of cardiorespiratory fitness without the need for complex laboratory equipment. According to the American College of Sports Medicine [1], several field tests, such as running, walking, and progressive exercise protocols, can estimate VO_2max using validated equations that consider time, distance, heart rate, and individual characteristics. The running test developed by Cooper [2] was one of the first protocols to demonstrate a significant correlation between field performance and biomedical variables of interest based on the distance covered in a 12-minute interval. The indirect estimation of VO_2max as a practical alternative for assessing aerobic capacity in large groups is discussed in several studies [1], [3], [4]. Classic studies such as McArdle et al. [3] validated the proposed regression equation, which shows a high correlation with laboratory cardiopulmonary exercise tests. In military settings, the 12-minute run is used as an assessment of aerobic endurance, and recent research has examined its relationship with other measures of physical work capacity in military personnel [12]. Physical performance is also relevant to operational endurance and prolonged effort [5]. Military and security-force personnel are exposed to physically and psychologically demanding occupational conditions, reinforcing the importance of structured training and preparation for operational demands [15]. Due to the high incidence of cardiovascular risk factors in police officers, Adero et al. [6] recommend periodic assessment of health and physical performance. The Paraná State Military Police has more than 20,000 active members, and the 12MR is administered annually to approximately 4,000 to 6,000 members. An image-based system may assist with measurement and recording while reducing manual inconsistencies. Greater standardization and fairness are potential future benefits, but they require validation under routine testing conditions. This work proposes an image-based system that combines omnidirectional image acquisition, digital image processing, and Kalman filter-based tracking to estimate runners' positions and passage times.

II. Materials and Methods

A. System Data Acquisition

The proposed system was designed considering that, in the 12-minute running test, the runner's instantaneous position

is not the primary variable of interest, since performance is evaluated based on the total distance covered at the end of the test. Therefore, digital image processing provides a suitable approach for estimating runners' positions at predefined reference points. Image acquisition was performed using an Insta360 X4 camera with an 8K resolution (7680 $\times$ 3840 pixels) operating at 24 frames per second. This setup enabled continuous visualization of the runners throughout the test, provided that no significant occlusion occurred between participants.

To implement the image processing and runner tracking algorithms, MATLAB® (MathWorks, USA) was used due to its capabilities for numerical computation, data analysis, and graphical visualization. In the current implementation, all video data were processed offline after completion of the running test. Execution time, processing throughput, memory use, and end-to-end latency were not benchmarked; therefore, the present study does not demonstrate real-time capability. Transition to real-time processing would require profiling each stage on the intended hardware, sustaining the required video throughput without unacceptable frame loss or latency, and testing performance with larger groups and more frequent occlusions.

The offline strategy was used for this initial agreement assessment. Three videos were analyzed, each containing the same three runners and 45 passage records per runner, corresponding to 15 reference points over three laps. This yielded 135 passage records per runner and 405 repeated passage records overall. These records are repeated measurements nested within three individuals and should not be interpreted as 405 independent participants.

The images presented in this article depict the authors themselves, who provided written informed consent for publication. A group of three volunteers participated in the data collection procedures. After receiving detailed information about the study protocol, each participant completed the 12-minute running test on a 300-meter track under the supervision of a certified physical education professional. Passage times were recorded at each predefined cone and subsequently analyzed by a panel of evaluators as well as by the proposed image-based system.

Passage times were analyzed by ten evaluators in total: one evaluator designated a priori as the reference based on previous experience with the 12-minute running test protocol and nine additional evaluators. No predefined or formally randomized evaluation order was imposed. Each evaluator reviewed the same videos independently and asynchronously, at a time and location determined by individual availability, without access to measurements recorded by the other evaluators or by the software.For REF $\times$ AVG EVAL comparisons, AVG EVAL denotes the mean of the nine additional evaluators and therefore does not include the reference evaluator.

B. Image Fiducials and Processing

For the application of the system, pre-established reference points (fiducials) were used to determine the relative position of the runners during the race. Orange cones were selected because their color stood out from the runners. The camera was placed in the center of the 300-m track, composed of two straight lines and two arcs. Fiducials were placed 20 m apart, representing 15 reference points around the track.

The processing sequence is shown in Fig. 1. One frame out of every six was analyzed. The current frame was subtracted from the previous frame, a threshold was applied, small objects were removed, connected moving regions were identified, and bounding boxes were generated for object localization.

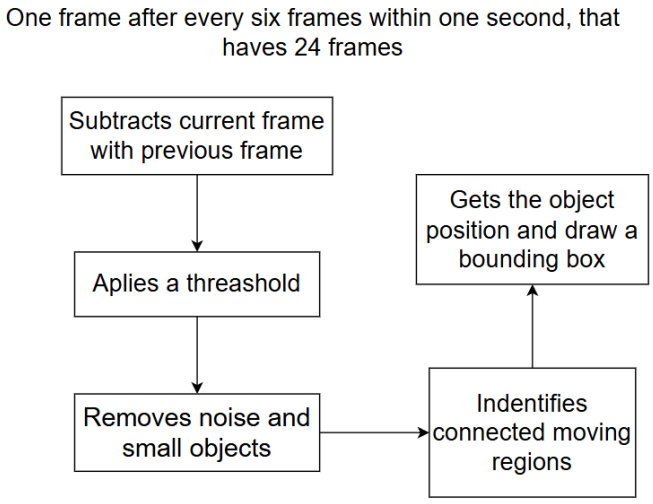


Fig. 1. Flowchart of the proposed image processing pipeline for frame subtraction and object localization.

Frame timestamps were converted to seconds by dividing the frame index by the video frame rate of 24 frames per second. To reduce computational cost, the method processed one out of every six frames, which effectively increased the temporal spacing between analyzed frames from 1/24 s to 0.25 s.

Consequently, consecutive analyzed frames were separated by 0.25 s, limiting the temporal resolution of the passage-time estimates to 0.25 s rather than the original frame-level resolution of approximately 0.042 s. At an average running speed of approximately 3 m/s, a 0.25-s sampling interval corresponds to approximately 0.75 m of distance traveled between consecutive analyzed frames.

Fig. 2 shows the runner and one predefined fiducial marker. Under the evaluated recording conditions, the system enabled localization of the predefined fiducial markers. During initialization, the user selected the position of each cone once. These coordinates were stored and used as fixed spatial references for estimating the runners' positions throughout the test.

The subtraction function highlights the difference between consecutive frames and then applies a predefined threshold. Next, the MATLAB `bwareaopen` function removes objects smaller than 25 pixels. The remaining connected objects are labeled using the `bwlabel` function. Finally, a bounding box is generated for each detected object, and its position is



Fig. 2. Original image acquired during the 12-minute running test showing the runner and one of the predefined fiducial markers (orange cone).

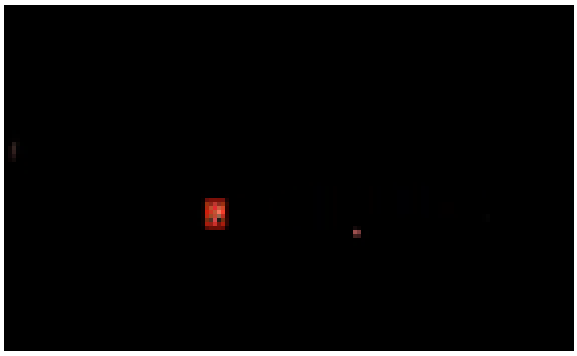


Fig. 3. Processed image showing runner detection after color segmentation.

recorded for subsequent tracking. This technique highlighted movement in all evaluated videos. However, it alone cannot track runners when they overlap and subsequently overtake one another. Therefore, runners wore different colored clothing to enable color-based segmentation and individual identification under the evaluated conditions.

Under the evaluated clothing colors and recording conditions, this approach enabled the runners to be identified individually. After segmentation, the RGB mask isolates the runner from the background, allowing the runner's position to be estimated before the tracking stage, as illustrated in Fig. 3.

C. Running-Track Position Estimation

Once the runner's position has been determined in the image, it is necessary to estimate the corresponding position on the running track. This is accomplished by using the predefined fiducial markers (orange cones) as spatial references. Runner position is most directly associated with a predefined track location when the runner passes in front of a fiducial marker.

The major problem occurs when the camera does not have visibility for every runner. With multiple runners, one may block the view of another, requiring position estimation. This is consistent with common challenges in multi-object tracking, where temporary occlusions prevent direct measurement and predictive filters are used to maintain continuity [11]. Occlusion effects could be minimized by acquiring video from an elevated viewpoint. Considering a camera-to-track distance of approximately 25 m and an average athlete height of 1.80 m, an approximate camera height of 11 m would be required for an additional 5 m of vertical clearance. Since the available recording did not achieve this elevation, position estimates were generated during periods of invisibility.

To address this limitation, a Kalman filter was implemented based on the state-space filtering and prediction framework originally introduced by Kalman [14]. In the proposed system, the filter was employed to infer the runner's position during intervals of lost visual tracking using previously observed motion as a state predictor. This approach allows the estimated state to be propagated when direct visual measurements are temporarily unavailable. Similar predictive filtering strategies have been adopted in athlete-tracking applications [7].

The selected Kalman filter used a constant velocity model, with measurement standard deviations set to [200, 50] and process standard deviations set to [100, 25]. The measurement noise covariance matrix R was diag([40000, 2500]). The initial velocity estimate was [0, 0] because no prior motion information was available before detection. Instantaneous velocity and acceleration estimated from the image sequence were included in the runner state. Acceleration was included to account for changes in apparent image velocity associated with perspective; its effect was not evaluated separately in an ablation analysis. Additional parameters, including permissible blob size, ageing threshold, and non-visible time, must be adapted to camera resolution, camera-to-track distance, group size, and the frequency of occlusions [8].

The current tracking approach has limitations when multiple runners are simultaneously present. During prolonged occlusions, direct visual measurements may become temporarily unavailable, requiring the Kalman filter to propagate the estimated runner state based on previously observed motion. As the duration and complexity of an occlusion increase, uncertainty in the predicted position may also increase. Because only three volunteers were evaluated in the present study, performance under more frequent or prolonged occlusions was not established. Future studies should evaluate larger simultaneous groups and alternative camera configurations, including multiple viewpoints, to determine whether these approaches improve tracking continuity under more complex occlusion conditions.

D. Real Time Processing Limitations.

The proposed solution relies on first segmenting the image into moving and non-moving objects, a stage that can be efficiently parallelized on a GPU, using large baches of images, and therefore contributes positively to real-time feasibility. However, the main bottleneck of the pipeline lies in the current implementation of the Kalman filter. The tracking process and the differentiation between runners depend on the previous state of each track, which inherently imposes sequential dependencies and severely limits parallelization.

Although certain matrix operations within the Kalman filter can, in principle, be offloaded to the GPU, the overall performance gain remains limited. Due to the sequential nature of state updates, frames cannot be processed in large batches; they must be handled individually. This constraint introduces a significant bottleneck related to memory transfer between RAM and GPU, ultimately restricting the scalability of the system and hindering real-time deployment. As a consequence, most of the processing remains confined to the CPU, which is comparatively slow for this type of operation because the workload consists of numerous small state-dependent computations.

Some additional algorithms evaluation is required to explore the potential to reduce the computational cost of the system, these improvement should be a priority for the future development of a a real time processing solution.

E. Validation and Statistical Analysis

Human passage times were determined exclusively from the video recordings. All ten evaluators identified when each runner crossed each predefined reference point. The reference evaluator was compared with the mean of the nine additional evaluators, whereas inter-rater agreement was calculated among all ten evaluators. The dataset contained 405 repeated passage records: 135 per runner and 45 per runner in each of three videos. All runners covered the same distance and performed the same number of repeated passages, ensuring a fully balanced design across participants. Because each individual contributed repeated crossings, the observations were treated as measurements nested within three participants rather than as independent participant-level observations.

All evaluators analyzed the same video recordings and applied the same criteria for identifying passage events; however, viewing equipment and environmental conditions were not formally standardized. Data processing for the proposed software was performed using Python with the NumPy, Pandas, Matplotlib, and Math libraries.

Statistical analyses were performed using R (R Foundation for Statistical Computing, Vienna, Austria). Bland–Altman analysis was used to characterize agreement between the software (SW), the reference evaluator (REF), and the mean of the nine additional evaluators (AVG EVAL) [9], [10].

Inter-rater agreement among all ten evaluators was assessed using a two-way random-effects intraclass correlation coefficient with absolute agreement. Single-measure ICC [ICC(2,1)] and average-measure ICC [ICC(2,10)] were calculated with 95% confidence intervals [16]. Descriptive agreement was also summarized by runner and video using the mean absolute deviation among evaluators and the REF minus AVG EVAL bias with 95% limits of agreement.

Because passage times were cumulative and repeated within only three participants, ICC values were interpreted together with absolute differences and were not treated as participant-level validation.

TABLE I
BLAND–ALTMAN AGREEMENT ANALYSIS

Comparison	Mean bias (s)	95% LoA (s)	p-value
REF–AVG EVAL	-0.71	-1.34 to -0.09	< 0.001
SW–REF	-0.65	-4.32 to 3.02	< 0.001
SW–AVG EVAL	-1.36	-4.91 to 2.19	< 0.001

TABLE II
STRATIFIED DESCRIPTIVE AGREEMENT BY RUNNER AND VIDEO

Group	n	ICC(2,1)	MAD (s) / REF–AVG9 bias (95% LoA), s
Overall	405	0.99994	0.54 / -0.71 (-1.33 to -0.09)
Runner AZL	135	0.99994	0.53 / -0.66 (-1.25 to -0.07)
Runner AMR	135	0.99994	0.53 / -0.71 (-1.30 to -0.12)
Runner LRJ	135	0.99994	0.56 / -0.76 (-1.41 to -0.11)
Video 1	135	0.99993	0.59 / -0.58 (-1.01 to -0.14)
Video 2	135	0.99994	0.51 / -0.75 (-1.49 to -0.02)
Video 3	135	0.99994	0.52 / -0.80 (-1.36 to -0.25)

III. RESULTS

A. Overall Agreement Analysis

Table I summarizes the Bland–Altman agreement analysis for the three comparisons performed in this study.

Inter-rater analysis among all ten evaluators yielded an ICC(2,1) of 0.99994 (95% CI: 0.99991–0.99996) for an individual evaluator and an ICC(2,10) of 0.999994 (95% CI: 0.999991–0.999996) for the mean of ten evaluators. Across 405 repeated passage records, the mean within-passage standard deviation was 0.70 s, the mean absolute deviation from the corresponding ten-evaluator mean was 0.54 s, and the mean range was 2.03 s. The REF minus AVG EVAL bias, with AVG EVAL calculated from the nine additional evaluators, was -0.711 s (95% limits of agreement: -1.326 to -0.095 s).

Participant- and video-stratified values were similar within the evaluated dataset. These summaries remain descriptive because the records are repeated measurements from only three runners; they do not provide independent participant-level validation of the software.

B. Reference Evaluator Versus Average Human Evaluation

The Bland–Altman analysis compared REF with AVG EVAL. Most observations were distributed near the mean difference (Fig. 4). The limits of agreement ranged from -1.34 s to -0.09 s, and the mean difference was statistically different from zero ($p < 0.001$), with a negative bias of approximately -0.71 s. The entirely negative limits of agreement indicate that REF generally recorded lower passage times than AVG EVAL within the evaluated dataset. The relatively narrow interval between the limits describes limited dispersion of the differences around the systematic offset observed between the human evaluations.

C. Software Versus Reference Evaluator

The proposed image-based software (SW) was compared with REF (Fig. 5). Most measurements were located within the 95% limits of agreement. The mean difference was statistically significant ($p < 0.001$), with limits of agreement ranging from

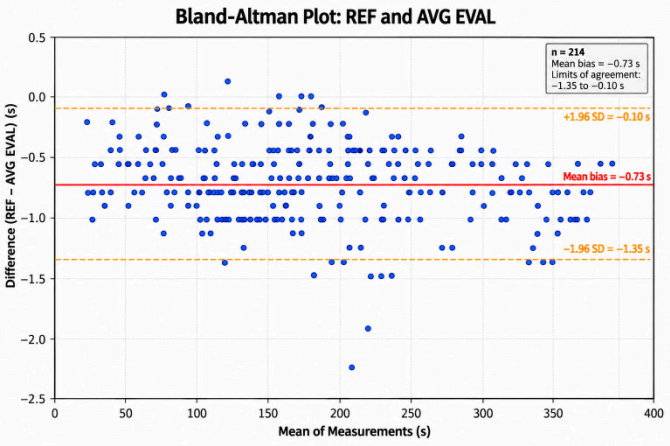


Fig. 4. Bland–Altman plot comparing the reference evaluator (REF) and the average of the evaluators (AVG EVAL).

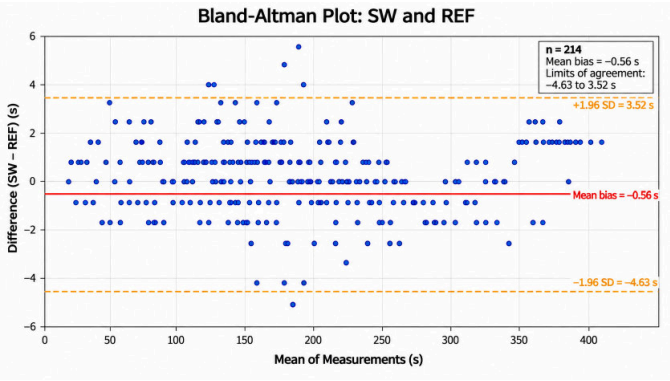


Fig. 5. Bland–Altman plot comparing the proposed software (SW) and the reference evaluator (REF).

-4.32 to 3.02 s. The distribution of the differences extended across zero, indicating that individual software estimates could be either lower or higher than the corresponding REF measurements. The interval between the lower and upper limits of agreement was 7.34 s, indicating greater dispersion than that observed in the REF versus AVG EVAL comparison.

D. Software Versus Average Human Evaluation

The proposed software was compared with AVG EVAL (Fig. 6). Most measurements were located within the 95% limits of agreement. The mean difference was statistically significant ($p < 0.001$), and the negative mean bias was -1.36 s, with 95% limits of agreement from -4.91 to 2.19 s. The distribution of the differences extended across zero, indicating that individual software estimates could be either lower or higher than the corresponding AVG EVAL measurements. The interval between the lower and upper limits of agreement was 7.10 s, showing greater dispersion than that observed between REF and AVG EVAL.

E. Practical Interpretation of the Systematic Bias

Table III converts the absolute mean temporal biases into equivalent running distances at a pace of 5:00 min/km, at

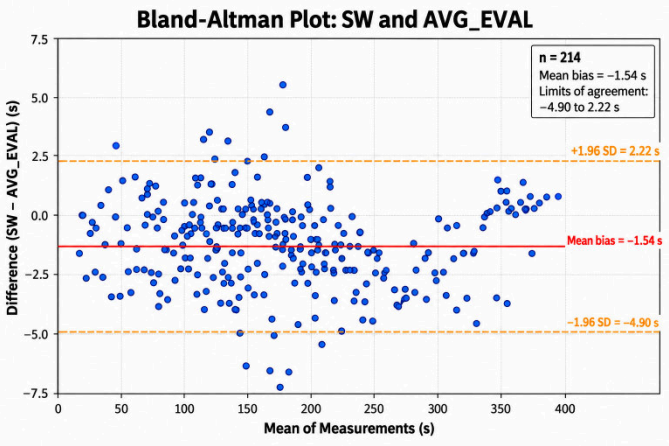


Fig. 6. Bland–Altman plot comparing the proposed software (SW) and the average of the evaluators (AVG EVAL).

TABLE III
PRACTICAL INTERPRETATION OF SYSTEMATIC BIAS AT 5:00 MIN/KM

Comparison	Absolute mean bias (s)	Equivalent distance (m)
REF × AVG EVAL	0.71	2.38
SW × REF	0.65	2.16
SW × AVG EVAL	1.36	4.53

which each second corresponds to approximately 3.33 m traveled. At this pace, the mean biases observed for REF × AVG EVAL, SW × REF, and SW × AVG EVAL corresponded to approximately 2.38 m, 2.16 m, and 4.53 m, respectively. The largest equivalent distance was observed for the SW × AVG EVAL comparison, consistent with its larger absolute mean temporal bias. These values provide a distance-based representation of the systematic differences observed between the compared measurements. The difference between the smallest and largest distance equivalents was approximately 2.37 m. Both comparisons involving the software produced distance equivalents derived directly from the temporal biases observed in the Bland–Altman analyses. Expressing these temporal differences in meters facilitates their interpretation in the context of a running test in which the final performance outcome is determined by the distance covered.

IV. DISCUSSION

The Bland–Altman analyses identified systematic differences between the compared measurements. Because cumulative passage times span a broad range, the very high ICC values should be interpreted together with the absolute differences and the nested repeated-measure structure.

For REF versus AVG EVAL, no visible proportional trend was observed across the range of cumulative passage times. The statistically significant negative bias of approximately -0.71 s indicates a systematic offset rather than an absence of disagreement. Within the evaluated recordings, the reference evaluator showed small absolute differences relative to the mean of the nine additional evaluators. These results support

its use as the reference for the present comparison but do not establish test-retest reliability or reproducibility across sessions.

For SW versus REF, the differences remained approximately constant over cumulative passage time, without a visible increase in variability as the running test progressed. The statistically significant negative bias indicates systematic underestimation by the software rather than absence of disagreement. At a running pace of 5:00 min/km, the limits of agreement from -4.32 to 3.02 s correspond to distance differences of approximately -14.4 to 10.1 m. A passage-time difference of this magnitude could affect estimated final distance and may change the classification of runners whose results are close to a classification threshold. Calibration may reduce the systematic component, but post-calibration agreement was not evaluated.

For SW versus AVG EVAL, the differences likewise did not show a visible increase in variability over cumulative passage time. The negative bias indicates that the software systematically underestimated passage times relative to the evaluator mean. Similar agreement patterns were observed in the SW $\times$ REF and SW $\times$ AVG EVAL comparisons. This similarity describes the behavior of the software within the present dataset; it does not demonstrate independence from the evaluator, reproducibility across sessions, or performance under different camera setups. In sports tracking, computer vision-based systems have also been evaluated through criterion validity and test-retest reliability, providing a methodological basis for assessing both measurement accuracy against a reference method and reproducibility across repeated testing conditions [13].

The absolute mean biases corresponded to estimated positional differences of 2.38 m for REF $\times$ AVG EVAL, 2.16 m for SW $\times$ REF, and 4.53 m for SW $\times$ AVG EVAL at a pace of 5:00 min/km. Although these values represent a small proportion of the total distance covered, measurements near a classification threshold could affect the final distance classification. The practical effect depends on the decision threshold and on where the runner finishes relative to that threshold. Calibration may reduce the systematic component, but this possibility was not tested after correction in the present study.

The distinction between systematic bias and the dispersion represented by the limits of agreement is relevant when interpreting the software performance. A calibration procedure could potentially correct part of the mean systematic offset, but it would not necessarily reduce the variability of individual differences between the software and human evaluations. Therefore, future assessments should examine both the residual bias and the limits of agreement after calibration rather than relying only on changes in the mean difference. This distinction is particularly relevant for measurements obtained near performance classification thresholds, where individual deviations may be more consequential than the average bias. Repeated measurements under independent testing conditions would also help determine whether the observed systematic differences remain stable across sessions.

This study has several limitations. Only three volunteers were included, and each contributed repeated passage records; consequently, the effective participant-level sample remained three despite the 405 recorded crossings. Occlusion frequency and complexity were limited compared with routine tests involving larger groups. The recordings represented a restricted set of camera, lighting, weather, background-motion, and track conditions. Only one camera configuration was examined, without repeated testing sessions or alternative camera positions. Computational execution time, throughput, memory use, and end-to-end latency were not benchmarked, and real-time operation was not demonstrated. Therefore, test-retest reliability, repeatability across sessions, evaluator independence, real-time performance, and performance across different acquisition or environmental conditions were not established. Routine use, fairness, and application in official physical fitness assessments remain potential future applications requiring validation with larger participant samples, larger simultaneous groups, more frequent and prolonged occlusions, repeated sessions, varied environmental conditions, different camera configurations, and explicit computational benchmarking on the intended hardware.

V. Conclusions

The proposed software was evaluated by comparison with a reference evaluator and the mean of nine additional human evaluators using Bland–Altman analysis. Inter-rater agreement was assessed among all ten evaluators. Under the evaluated conditions, the human evaluators showed small absolute differences, while the software exhibited statistically significant negative biases relative to the reference evaluator and the evaluator mean. These biases represent systematic underestimation and should not be interpreted as absence of disagreement.

The observed results indicate that the proposed system was able to generate passage-time estimates that could be systematically compared with human measurements under the evaluated recording conditions. However, the presence of negative biases and the width of the limits of agreement indicate that correction of the mean offset alone would not be sufficient to establish measurement interchangeability. Future validation should therefore assess whether calibration reduces the systematic bias while preserving acceptable limits of agreement across independent participants and repeated sessions. Particular attention should be given to measurements near performance classification thresholds, where individual differences may influence the final distance classification.

The analyses provide an initial assessment based on repeated passage measurements from three participants. They do not establish test-retest reliability, repeatability across sessions, evaluator independence, or reproducibility across camera configurations. The system may have future applications in training, performance monitoring, and official physical fitness assessment. These applications require calibration and validation with larger groups, more frequent and prolonged occlusions, repeated sessions, and different camera positions

and acquisition conditions. Although the method was developed according to the protocol adopted by the Paraná State Military Police, adaptation to other institutions or protocols should be investigated rather than assumed from the present data. Future studies should also establish predefined acceptable measurement-error thresholds according to the operational criteria of the physical fitness test. Such thresholds would allow the observed agreement to be interpreted in relation to its practical effect on final performance classification.

ETHICAL CONSIDERATIONS

This study involved image-based measurements collected during the routine 12-minute run physical fitness test of the Paraná State Military Police (PMPR). No additional running session or physical effort was required specifically for research purposes. Participants performed the test under the standard conditions of their regular physical fitness assessment, and the images were recorded exclusively for subsequent analysis. All participants provided written informed consent for participation in the study and image acquisition.

ACKNOWLEDGMENT

The authors express their gratitude for making the infrastructure of the sports practice facilities of the Paraná Military Police (PMPR) available for this research.

REFERENCES

[1] American College of Sports Medicine, *ACSM's Guidelines for Exercise Testing and Prescription*, 10th ed. Philadelphia, PA, USA: Wolters Kluwer, 2018.

[2] K. H. Cooper, "A means of assessing maximal oxygen intake," *JAMA*, vol. 203, no. 3, pp. 201–204, 1968.

[3] W. D. McArdle, F. I. Katch, and V. L. Katch, *Exercise Physiology: Nutrition, Energy, and Human Performance*, 7th ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2010.

[4] L. A. Léger and J. Lambert, "A maximal multistage 20-m shuttle run test to predict VO_2max," *European Journal of Applied Physiology and Occupational Physiology*, vol. 49, no. 1, pp. 1–12, 1982.

[5] M. L. Foss and S. J. Keteyian, *Fox's Physiological Basis for Exercise and Sport*, 7th ed. New York, NY, USA: McGraw-Hill, 2010.

[6] O. E. Adero, R. D. O. Dev, and M. R. N. W. Wazir, "Dimensions of physical fitness evaluation procedures among police personnel: Systematic literature review," *Int. J. Acad. Res. Prog. Educ. Dev.*, vol. 12, no. 3, pp. 182–192, 2023.

[7] R. Gade and T. B. Moeslund, "Thermal vision for night-time detection and tracking of sports players," *Sensors*, vol. 14, no. 8, pp. 13679–13691, 2014.

[8] P. Rahimian and L. Toka, "Optical tracking in team sports: A survey on player and ball tracking methods in soccer and other team sports," *J. Quant. Anal. Sports*, vol. 18, no. 4, pp. 207–230, 2022.

[9] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[10] J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," *Stat. Methods Med. Res.*, vol. 8, no. 2, pp. 135–160, 1999.

[11] Z. Guan, Z. Wang, G. Zhan, et al., "Multi-object tracking review: Retrospective and emerging trends," *Artif. Intell. Rev.*, vol. 58, Art. no. 235, 2025.

[12] J. Drozd, J. Neubauer, J. Sekanina, and M. Sedlačík, "Bridging the gap: Aligning physical work capacity testing with actual endurance performance in military settings," *Frontiers in Psychology*, vol. 16, Art. no. 1536197, 2025, doi: 10.3389/fpsyg.2025.1536197.

[13] D. Bhardwaj, D. Broadbent, S. A. Feros, and D. B. Dwyer, "Assessing criterion validity and test-retest reliability of a computer vision-based system for tracking player location in cricket," *International Journal of Sports Science & Coaching*, vol. 21, no. 2, 2026.

[14] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960, doi: 10.1115/1.3662552.

[15] M. Sedlacík, R. Soušek, M. Staňková, and J. Cáp, "The role of learning and training in supporting emotional resilience in security forces," *Frontiers in Psychology*, vol. 17, Art. no. 1854418, 2026, doi: 10.3389/fpsyg.2026.1854418.

[16] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979, doi: 10.1037/0033-2909.86.2.420.

Impact of Region of Interest Representation Methods on Ovarian Tumor Classification in Ultrasound Images

Alina Dyakova
Department of Radiotechnical Fundamentals
National Research University "Moscow Power Engineering Institute"
Moscow, Russia
https://orcid.org/0009-0001-5499-0797

Konstantin Boshkov
Department of Radiotechnical Fundamentals
National Research University "Moscow Power Engineering Institute"
Moscow, Russia
https://orcid.org/0009-0006-2473-2937

Luan S. Nascimento
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0000-0001-6829-3258

Kerllon F. Andrade
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0009-0005-4279-4368

Denis Leonov
Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies
Moscow Healthcare Department
Moscow, Russia
https://orcid.org/0000-0003-0916-6552

José Francisco S. Costa-Júnior
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0000-0002-6526-2760

***Abstract*—In medical computer vision, it is commonly assumed that prior extraction of the pathological region improves subsequent classification performance. However, this assumption requires validation in the context of ultrasound imaging. The clinical assessment of ovarian masses relies not only on their internal structure but also on their contours, walls, septa, acoustic features, and surrounding anatomical context. This study aimed to evaluate the impact of different pixel mask utilization strategies on ovarian tumor classification in ultrasound images. Experiments were conducted using the publicly available MMOTU/OTU-2D dataset, comprising 1,202 locally preprocessed images with corresponding masks. Two classification tasks were investigated: a seven-class classification of ovarian mass morphological types and a binary classification distinguishing simple cysts from other lesions. Eight input representation variants were generated from each original image-mask pair: the full frame, a masked crop with a zeroed background, bounding-box crops with margins of 0%, 10%, 30%, 50%, and 80%, and the full frame augmented with the mask as an additional channel. To isolate the effect of input representation, a single EfficientNet-B0 architecture was employed across all primary experiments. Each variant was trained four times using different random seeds. A total of 72 results were obtained, comprising 64 individual training runs and eight ensemble models generated by averaging the predicted probabilities across the eight input representations. In the seven-class task, the full-frame approach achieved a macro-F1 score of 0.4971 ± 0.0185, whereas the masked crop with a zeroed background reached 0.6052 ± 0.0231 (paired difference: +0.1082; 95% CI, +0.0477 to +0.1687; Holm-adjusted p = 0.0537). In the binary classification task, the effect was less pronounced: the full-frame approach yielded a macro-F1 score of 0.6752 ± 0.0266, the masked crop achieved 0.7236 ± 0.0222, and the ensemble reached 0.7728 ± 0.0303. These findings indicate that segmentation masks serve not merely as auxiliary annotations but as mechanisms that redefine the informational framing of the classification task. Consequently, the choice of region-of-interest representation should be considered an independent and critical component of experimental design.**

Keywords*— *Ultrasound, ovarian tumors, image classification, region of interest, MMOTU, OTU-2D, EfficientNet-B0, macro-F1, ensemble learning

I. Introduction

Ultrasound imaging is widely used in gynecology because of its accessibility, noninvasiveness, and ability to provide dynamic assessment of pelvic organs. In clinical practice, standardized approaches such as IOTA and O-RADS are used to evaluate ovarian masses. These systems consider mass morphology, the number of locules, papillary projections, acoustic characteristics, and vascularity [1]–[4]. However, the automated analysis of ultrasound (US) images remains challenging because they are often affected by speckle noise and imaging artifacts [5]–[7], frequently contain embedded measurement calipers and text annotations [8], [9], exhibit variations in scale and anatomy [7], and are acquired under heterogeneous imaging conditions across operators, institutions, and ultrasound systems [5], [7], [9]. These factors become particularly relevant when training neural network models on relatively small medical datasets [1].

For ovarian tumor and cyst classification, the use of prior mass annotations is a natural approach. When a segmentation mask is available, the background can be removed, the lesion can be enlarged, or the model can be explicitly provided with information regarding the location of the region of interest (ROI). At first glance, such strategies are expected to enhance classification performance by reducing the influence of technical artifacts and providing a more focused representation of the pathological region. This rationale is consistent with common practices in medical image analysis, in which localization of the pathological region frequently precedes classification [1].

For ultrasound imaging, however, this rationale is not universally applicable. Clinical assessment relies not only on the internal structure of the mass but also on its contour, wall characteristics, septa, acoustic features, relationship to surrounding tissues, and the scale and position of the lesion within the field of view [2]–[4], [10], [11]. When a mask is

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

applied too restrictively, diagnostically relevant information may be removed together with the background. Consequently, the key question is not whether a mask is necessary, but rather how it should be incorporated into the input representation provided to the classifier.

Previous studies have demonstrated that neural network models can effectively classify ovarian tumors from US images, including comparisons with O-RADS classifications and expert assessments [12], the application of convolutional neural network (CNN) ensembles [13], the classification of multiple histological subtypes [14], and the development of models targeting complex O-RADS categories [15]. However, these studies rarely investigate the representation of the ROI as an independent experimental variable. The present study specifically addresses this intermediate stage between mass annotation and classification.

This work investigates the impact of eight ROI representation strategies on the performance of ovarian tumor classification in US images. The primary hypothesis is that a single pixel mask can give rise to multiple input representations and that these representations are not equivalent from the classifier's perspective, even when the model architecture remains unchanged.

The main contribution of this study lies in the controlled comparison of eight input representation strategies using a single EfficientNet-B0 architecture. In contrast to studies that compare different models, this experimental design isolates the effect of mask utilization. The analysis encompasses two classification tasks: a detailed seven-class classification and a binary classification distinguishing simple cysts from all other masses. Four independent training runs were performed for each representation, ensuring that the results reflect not only peak performance but also the consistency of the observed effects. The findings indicate that, in the seven-class task, ROI-based representations yield higher mean macro-F1 values, whereas their impact is less pronounced in the binary classification task. Moreover, the ensemble combining all eight input representations outperformed every individual variant, suggesting that these representations provide complementary information.

II. Background and Problem Statement

A. ROI Representation Strategies for Ultrasound Classification

In medical image analysis, the ROI can be incorporated in several ways, including as a standalone segmentation task, as a prior localization step, an image-cropping strategy, or supplementary information provided to the classifier [1]. The most restrictive approach retains only the pixels contained within the mask while suppressing all background information. A less restrictive strategy consists of extracting a bounding box around the mask, thereby preserving part of the surrounding anatomical context. Alternatively, the full image can be maintained while incorporating the mask as an additional input channel.

These input representations evaluate distinct hypotheses regarding the information required for classification. A strict crop assumes that the most relevant diagnostic information is confined to the mass itself. A bounding-box crop presumes that both the lesion and its immediate surroundings contribute to classification performance. A full-frame representation supplemented with a mask assumes that the model benefits from access to the entire image while receiving explicit information about the lesion's location. Finally, an ensemble of multiple input representations assumes that different views may provide complementary information. This distinction is particularly important in ovarian ultrasound, as clinical risk assessment systems rely not only on the presence of a mass but also on a range of morphological characteristics that may be partially lost through excessive cropping [2]–[4].

B. MMOTU/OTU-2D Dataset

The MMOTU/OTU-2D dataset is one of the few publicly available datasets for the analysis of ultrasound images of ovarian tumors. It comprises two-dimensional ultrasound (US) images, contrast-enhanced ultrasound (CEUS) images, global class labels, and pixel-level masks [10]. In this study, a locally preprocessed subset of OTU-2D containing 1,202 image-mask pairs was employed. CEUS images were excluded because they represent a different imaging modality.

In the original MMOTU study, the primary emphasis was placed on segmentation and knowledge transfer across imaging modalities [10]. In contrast, the masks in the present work were not used as segmentation targets but rather as mechanisms for generating distinct input representations for subsequent classification. Consequently, the proposed task is related to, yet conceptually distinct from, the original MMOTU framework.

This study does not address the screening task, that is, determining whether an ovarian mass is present in an arbitrary ultrasound image. Instead, it considers a more restricted formulation in which the input consists of preprocessed images from the MMOTU/OTU-2D dataset, for which the presence of an ovarian mass and its corresponding class label are already known.

III. Materials and Methods

A. Data and Splitting

The class distributions and data partitions adopted for the seven-class formulation are presented in Table I. We stratified the official OTU-2D training partition (820 images) into training and validation sets (80:20; random state = 42); the official test partition (382 images) was unchanged. Table II summarizes the composition of the image sets used in the binary formulation. This binary classification task exhibited a marked class imbalance, particularly in the test set, which contained only 19 simple cysts compared with 363 cases belonging to the remaining categories. Consequently, overall accuracy was not considered the primary evaluation metric, as it may provide an overly optimistic assessment of performance under highly imbalanced conditions.

B. Input Image Variants

Eight input representations were generated from each original image and its corresponding mask (Table III, Fig. 1).

The masked crop with a zeroed background was generated by automatically determining the bounding box from the non-zero pixels of the binary lesion mask. Both the original image and the corresponding mask were then cropped using this bounding box. Subsequently, all pixels outside the mask but within the cropped region were set to zero, preserving only the lesion pixels. For the zero-margin variant, the bounding box corresponded exactly to the boundaries of the binary mask, with no additional expansion of the cropped region. When lesions were located near the image boundaries, the crop was constrained by the original image dimensions, and no external padding was applied. For the 10%, 30%, 50%, and 80% margin variants, the bounding box was proportionally expanded around the mask while remaining constrained by the

original image boundaries, with no padding added beyond the image frame. Crops were generated at their original resolution, and resizing to 224 × 224 was performed subsequently in the dataloader. For the validation and test sets, a deterministic Resize(224, 224) transformation was applied, whereas RandomResizedCrop(224, scale = (0.75, 1.0)) was used during training as part of the augmentation pipeline.

TABLE I. DISTRIBUTION OF IMAGES BY CLASS IN THE 7-CLASS FORMULATION

CLASS	TRAINING SET	VALIDATION SET	TEST SET	TOTAL
Endometriotic cyst	181	45	110	336
High-grade serous carcinoma	30	8	15	53
Mucinous cystadenoma	57	14	33	104
Serous cystadenoma	122	31	66	219
Simple cyst	38	9	19	66
Teratoma	182	46	108	336
Theca-cell tumor	46	11	31	88
Total	656	164	382	1202

TABLE II. DISTRIBUTION OF IMAGES IN THE BINARY FORMULATION

CLASS	TRAINING SET	VALIDATION SET	TEST SET	TOTAL
OTHER OVARIAN MASSES	618	155	363	1136
SIMPLE CYST	38	9	19	66
TOTAL	656	164	382	1202

TABLE III. EVALUATED INPUT REPRESENTATION VARIANTS

NO	Input variant	What the model sees	Tested hypothesis
1	Full frame	The entire original US image	Is the whole frame sufficient without explicitly indicating the ROI?
2	Masked, zeroed background	Crop around the mask; pixels outside the exact mask are set to zero	Is the primary signal contained solely within the mass?
3	Bounding box, 0% margin	Bounding box crop around the mask with no extra padding	Is a minimal bounding box around the object sufficient?
4	Bounding box, 10% margin	The mass and a small surrounding area	Does immediate context help?
5	Bounding box, 30% margin	The mass and wider surroundings	Does moderate field-of-view expansion improve performance?
6	Bounding box, 50% margin	Enlarged area around the mass	Is a broader context needed?
7	Bounding box, 80% margin	Large fragment of the original frame	What happens when the expansion approaches the full frame?
8	Full frame + mask	Full frame with the mask as an additional channel	Is explicitly indicating the object's location without removing the background sufficient?

The eight-input ensemble was constructed independently. For each training run, the predicted probabilities produced by the eight models trained on the different input representations were averaged.

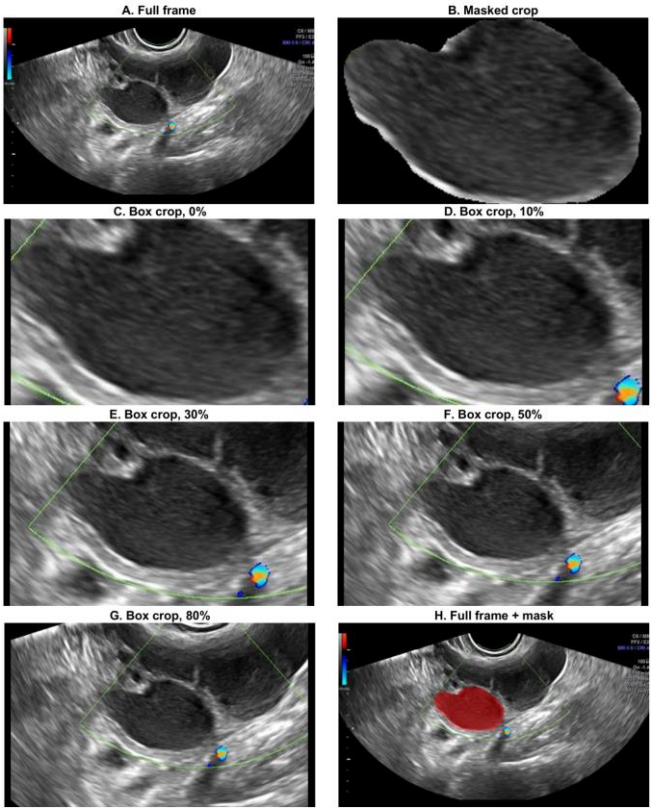


Fig. 1. Eight input representations generated from a single US image and mask: full frame (A), masked crop (B), bounding box crops with various margins (C-G), and full frame with an overlaid mask (H).

C. *Model and Training*

A single EfficientNet-B0 architecture was employed throughout all primary experiments. The EfficientNet family was introduced as a framework for uniformly scaling the depth, width, and input resolution of convolutional neural networks while maintaining an effective balance between predictive performance and computational efficiency [11]. In the present study, EfficientNet-B0 was not selected to identify the optimal architecture, but rather to serve as a fixed baseline classifier, enabling a controlled comparison of input representations under identical modeling conditions.

TABLE IV. MAIN TRAINING AND EVALUATION PARAMETERS

Parameter	Value
Architecture	EfficientNet-B0
Input size	224x224
Number of epochs	25
Number of runs	4
Random seed values	42, 43, 44, 45
Optimizer	AdamW
Learning rate scheduler	CosineAnnealingLR
Model saving criterion	Best val_macro_f1 value
Primary metric	macro-F1 on the test set
Secondary metrics	balanced accuracy, AUROC, overall accuracy
Handling class imbalance	Weighted loss function and balanced image sampling during training

The total volume of the experimental grid is presented in Table V.

TABLE V. VOLUME OF THE EXPERIMENTAL GRID

Task	Input Variants	Runs	Model Training	Ensembles	Total Results
7 CLASSES	8	4	32	4	36
2 CLASSES	8	4	32	4	36
Total	16	4	64	8	72

D. Evaluation Metrics and Statistical Analysis

The primary evaluation metric was macro-F1, as it averages the F1-score across all classes and prevents dominant classes from obscuring errors associated with minority classes. The use of this metric is particularly important in imbalanced classification tasks, in which overall accuracy may overestimate model performance [16].
For a single class:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (1)$$

For a multi-class task:

$$macro - F1 = mean\ F_1 \text{across all classes.} \quad (2)$$

Additionally, balanced accuracy, defined as the mean recall across classes, and AUROC (for the binary formulation) or macro-AUROC (using a one-vs-rest strategy for the seven-class task) were employed, together with confusion matrices, to characterize the distribution of classification errors. All results reported herein correspond to the test set. For each input representation, the mean and standard deviation across four independent runs are presented. In the multiclass setting, AUROC was interpreted as an extension of the area under the ROC curve to multiple classes [17]. All metrics were computed using standard implementations available in the scikit-learn library [18].

Because all input variants used the same four random seeds, comparisons with the full-frame baseline were paired by seed. Mean paired macro-F1 differences were accompanied by two-sided 95% Student-t confidence intervals and evaluated using paired t-tests; within each task, p values for the eight comparisons were adjusted by Holm's procedure. These intervals describe variation across training initializations rather than patient-sampling uncertainty.

IV. RESULTS

A. Seven-Class Classification

In the seven-class task, all variants incorporating a mask or a bounding box surrounding the mask achieved higher mean macro-F1 scores than the full-frame approach (Table VI, Fig. 2). The masked crop had the highest mean macro-F1 among individual representations. The full-frame approach yielded a balanced accuracy of 0.4933 and a macro-AUROC of 0.8338, whereas the masked crop with a zeroed background increased these values to 0.6303 and 0.8883, respectively. The eight-input ensemble further improved performance, achieving a balanced accuracy of 0.6947 and a macro-AUROC of 0.9226. The confusion matrices indicate that this improvement cannot be attributed solely to an increase in correct predictions for the majority classes. Instead, the most significant benefit lies in reducing confusion among morphologically similar categories, although rare classes remain susceptible to misclassification because of the limited number of available samples. Relative to the full-frame baseline, the seed-paired macro-F1 difference was +0.1082 (95% CI, +0.0477 to +0.1687; Holm-adjusted p = 0.0537) for the masked crop and +0.2055 (95% CI, +0.1675 to +0.2434; Holm-adjusted p = 0.0030) for the ensemble. Thus, only the ensemble remained significant at $p < 0.05$ after correction.

B. Binary Classification

In the binary task, the effect of selecting different input representations was less pronounced (Table VII, Fig. 3). The full-frame baseline remained competitive, and the differences between most bounding box variants and the full frame amounted to only a few hundredths in macro-F1.

The masked crop again had the highest mean macro-F1 among individual representations; however, its advantage over the full-frame approach was less pronounced than in the seven-class task. This observation is consistent with the nature of the binary formulation, which represents a less complex classification problem, requiring the model only to distinguish simple cysts from all other masses rather than to differentiate among multiple morphological categories.

TABLE VI. RESULTS OF THE 7-CLASS CLASSIFICATION ON THE TEST SET

Input Variant	Runs	macro-F1, Mean ± SD	Difference from Full Frame
Full frame	4	0.4971 ± 0.0185	-
Masked, zeroed background	4	0.6052 ± 0.0231	+0.1082
Bounding box, 0% margin	4	0.5805 ± 0.0305	+0.0834
Bounding box, 10% margin	4	0.5703 ± 0.0194	+0.0732
Bounding box, 30% margin	4	0.5693 ± 0.0405	+0.0723
Bounding box, 50% margin	4	0.5755 ± 0.0169	+0.0784
Bounding box, 80% margin	4	0.5386 ± 0.0126	+0.0415
Full frame + mask	4	0.5734 ± 0.0290	+0.0764
8-input ensemble	4	0.7025 ± 0.0123	+0.2055

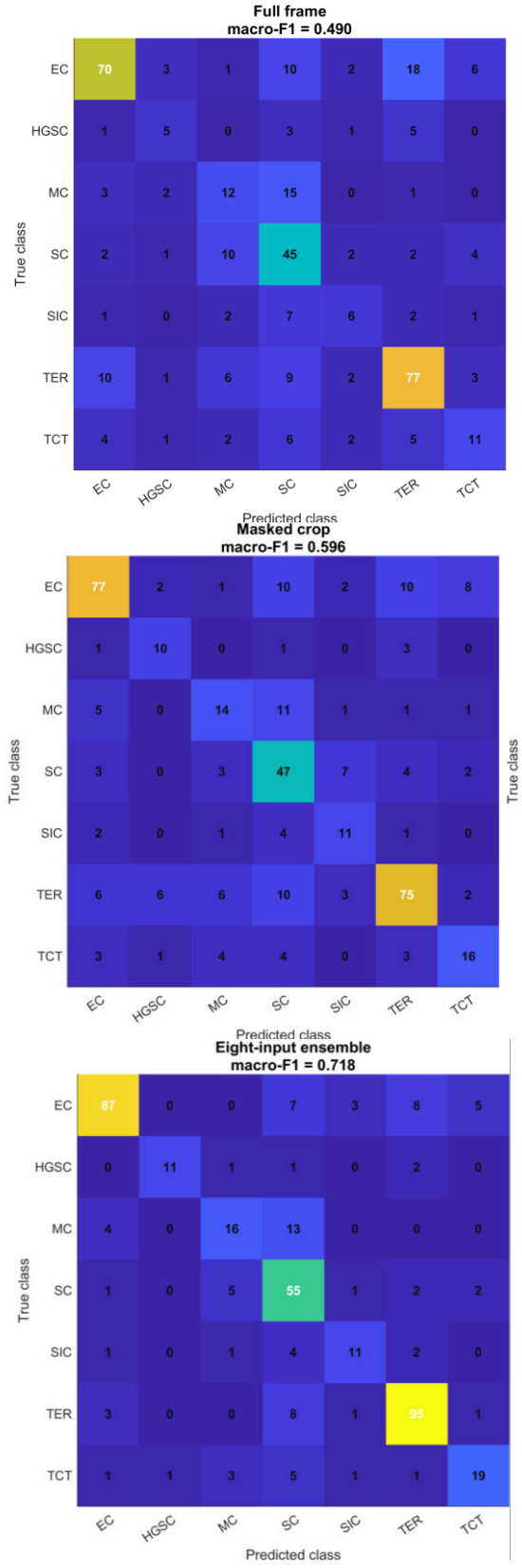


Fig. 2. Confusion matrices for the 7-class task on the same random seed run: full frame, masked crop, and 8-input ensemble. Class abbreviations: EC – endometriotic cyst, HGSC – high-grade serous carcinoma, MC – mucinous cystadenoma, SC – serous cystadenoma, SIC – simple cyst, TER – teratoma, TCT – theca-cell tumor.

TABLE VII. RESULTS OF THE BINARY CLASSIFICATION ON THE TEST SET.

Input Variant	Runs	Macro-F1, Mean ± SD	Difference from Full Frame
Full frame	4	0.6752 ± 0.0266	-
Masked, zeroed background	4	0.7236 ± 0.0222	+0.0484
Bounding box, 0% margin	4	0.6804 ± 0.0241	+0.0053
Bounding box, 10% margin	4	0.6901 ± 0.0210	+0.0150
Bounding box, 30% margin	4	0.6792 ± 0.0360	+0.0041
Bounding box, 50% margin	4	0.6965 ± 0.0681	+0.0213
Bounding box, 80% margin	4	0.6927 ± 0.0365	+0.0176
Full frame + mask	4	0.6925 ± 0.0323	+0.0173
8-input ensemble	4	0.7728 ± 0.0303	+0.0976

In terms of AUROC, the full-frame approach achieved a value of 0.8298, whereas the masked crop reached 0.8733 and the ensemble attained 0.9175. The bounding-box variant with a 50% margin yielded a relatively high mean AUROC of 0.9043 but exhibited substantial variability in macro-F1, with a standard deviation of 0.0681. This finding suggests that its performance was highly dependent on the specific training run and the resulting distribution of predicted probabilities. For macro-F1, the corresponding paired differences were +0.0484 (95% CI, -0.0159 to +0.1127; Holm-adjusted p = 0.6730) for the masked crop and +0.0976 (95% CI, +0.0133 to +0.1820; Holm-adjusted p = 0.2776) for the ensemble; neither remained significant after correction.

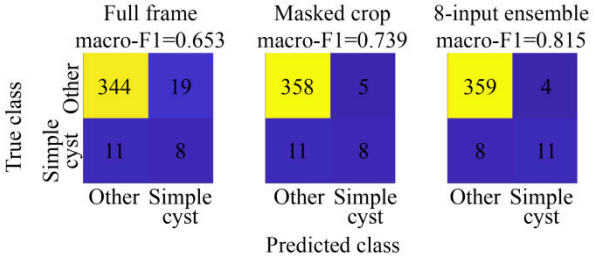


Fig. 3. Confusion matrices for the binary task on the same random seed run: full frame, masked crop, and 8-input ensemble. The "Other" class aggregates all masses except the simple cyst.

C. Input Representation Ensemble

The ensemble had the highest mean macro-F1 in both tasks. In the seven-class task, it increased the macro-F1 score from 0.4971 for the full-frame approach to 0.7025. In the binary task, the macro-F1 score improved from 0.6752 to 0.7728.

These findings indicate that the input representations are not entirely redundant. The full-frame approach preserves the global scene, the masked crop emphasizes lesion morphology, bounding boxes with different margins retain varying amounts of contextual information, and the full frame augmented with a mask explicitly conveys the lesion's location. Averaging the predicted probabilities integrates these complementary sources of information, resulting in improved classification performance.

V. Discussion

The results support the primary hypothesis that ROI representation constitutes an independent factor influencing ultrasound classification performance. The mask should not be regarded merely as an auxiliary segmentation annotation. When incorporated into the classification process, it modifies the distribution of input images, the scale of the lesion, the amount of contextual information available, and the set of features upon which the model can rely.

The results obtained in this study complement existing research on neural network-based ovarian tumor classification. Previous studies by Chen et al.[12], Hsu et al.[13], and Wu et al. [14] have primarily focused on determining whether a model can discriminate between benign and malignant masses or among multiple histological subtypes. In contrast, the present study addresses a different question: given that the mass mask is available, which strategy for incorporating this information leads to a more robust classifier? In this regard, the proposed approach occupies an intermediate methodological position between segmentation and classification.

For the seven-class task, strict localization had the highest mean macro-F1 among individual representations. A plausible explanation is that detailed classification of morphological subtypes requires the model to capture subtle characteristics related to the internal structure and shape of the lesion, whereas the full-frame representation may contain excessive irrelevant information for a lightweight architecture such as EfficientNet-B0. This consideration is particularly important for lesions containing solid mineral inclusions, such as microcalcifications, whose detection in ultrasound imaging depends on specific acoustic patterns that are more effectively characterized within a focused ROI [19]. Under these conditions, suppressing the background and enlarging the lesion provides a substantial advantage. This observation is consistent with the findings reported by Wu et al., who also identified mass annotation and cropping as factors influencing the performance of detailed histological subtype classification [14]. Cropping changes both background content and lesion scale, so the gain cannot be assigned to either mechanism alone.

For the binary classification task, the findings are less definitive. The full-frame representation already contains sufficient information to distinguish simple cysts from other lesions, and many bounding-box variants provide only modest improvements. This observation has important practical implications: results obtained for one classification task cannot be directly generalized to another. Consequently, the choice of input representation should be evaluated independently for each clinical application. This conclusion is consistent with the principles underlying the IOTA and O-RADS frameworks, in which different diagnostic questions rely on distinct sets of features rather than on a universal preprocessing strategy [2]–[4].

Augmenting the full-frame representation with the mask as an additional channel improved performance relative to the standard full-frame approach, although it did not emerge as the best-performing individual variant. This finding suggests that explicitly providing lesion location information does not necessarily ensure that the model will exploit such annotations optimally. In some cases, directly modifying the image representation, for example by enlarging the lesion region and suppressing portions of the background, may constitute a more effective strategy.

The ensemble had the highest mean macro-F1 in both tasks. This pattern suggests that different input representations capture complementary diagnostic information. However, ensemble models are computationally more demanding and more challenging to implement and maintain. Therefore, in practical applications, two distinct scenarios should be considered: the best-performing individual representation as a simple and lightweight solution, and the ensemble as a more accurate yet computationally intensive alternative.

This study has several limitations. First, the experiments were conducted using a single publicly available dataset [10]. Assessing generalizability requires external validation on independent clinical datasets acquired with different ultrasound systems and across multiple institutions. This consideration is particularly important in ultrasound imaging, where image quality and artifact characteristics are strongly influenced by the equipment, acquisition protocol, and operator. Patient identifiers were unavailable in the released index, precluding patient-level splitting and patient-clustered confidence intervals.

Second, the analysis was performed using pre-annotated images in which the presence of a mass was already known. Consequently, the results do not represent the entire clinical workflow, which also includes organ localization, lesion detection, image quality assessment, video analysis, and information integration across multiple image slices.

Third, mask quality was assumed to be an intrinsic characteristic of the dataset. No expert re-annotation was performed; consequently, the results reflect the influence of the available masks rather than that of idealized manual annotations. Clinical translation therefore requires repeating the analysis with automatically predicted masks and explicitly evaluating empty or implausible masks.

Fourth, the binary classification task is highly imbalanced. As a result, overall accuracy may provide a misleading impression of performance, making it necessary to rely on macro-F1, balanced accuracy, AUROC, and confusion matrices for a more comprehensive evaluation [16], [17].

Fifth, EfficientNet-B0 was adopted as a fixed architecture to enable a controlled comparison of input representations [11]. The objective of this study was not to identify the state-of-the-art architecture for the MMOTU/OTU-2D dataset. Although more advanced models may improve absolute performance metrics, they do not diminish the importance of evaluating ROI representation strategies.

Sixth, statistical power was limited by the use of only four random seeds per input variant. Although four runs are common in controlled experiments, larger numbers would be required to detect smaller effect sizes or to establish statistical

significance for individual ROI representations. The ensemble results, however, were robust and statistically significant.

Future studies may also benefit from validation using anatomically realistic three-dimensional printed phantoms, which offer controlled ground-truth conditions for the systematic assessment of ROI preprocessing strategies across different tumor morphologies [20], [21]. Existing ovarian-mass pipelines use segmentation-derived crops or detector-segmenter-classifier cascades [22], [23], but do not isolate ROI representation under fixed classification conditions.

VI. Conclusion

The use of segmentation masks in ovarian tumor classification from ultrasound images should be regarded as an independent experimental hypothesis rather than an inherently beneficial preprocessing strategy. In the seven-class task, the masked crop with a zeroed background increased the macro-F1 score, increasing it from 0.4971 ± 0.0185 to 0.6052 ± 0.0231. In contrast, the effect was less pronounced in the binary task, for which the macro-F1 score increased from 0.6752 ± 0.0266 to 0.7236 ± 0.0222. After correction for multiple comparisons, however, the masked-crop differences were not statistically significant.

The ensemble had the highest mean macro-F1 in both tasks, attaining macro-F1 scores of 0.7025 ± 0.0123 for the seven-class formulation and 0.7728 ± 0.0303 for the binary formulation. Only the seven-class ensemble remained statistically significant after Holm correction. The differences observed between the seven-class and binary tasks suggest that the optimal strategy for ROI utilization depends on the specific clinical application. Before clinical adoption, this approach should undergo external validation using independent datasets, together with a dedicated assessment of mask quality. For practical implementation, we recommend evaluating at least three ROI representations—the full frame, the masked crop, and a bounding box with a 30% margin—and selecting the configuration that yields the best performance for the target task, since the optimal representation is task-dependent.

References

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[2] D. Timmerman *et al.*, "Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group," *BMJ*, vol. 341, no. dec14 1, pp. c6839–c6839, Dec. 2010, doi: 10.1136/bmj.c6839.

[3] B. Van Calster *et al.*, "Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study," *BMJ*, vol. 349, no. oct07 3, pp. g5920–g5920, Oct. 2014, doi: 10.1136/bmj.g5920.

[4] R. F. Andreotti *et al.*, "O-RADS US Risk Stratification and Management System: A Consensus Guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee," *Radiology*, vol. 294, no. 1, pp. 168–185, Jan. 2020, doi: 10.1148/radiol.2019191150.

[5] S. Liu *et al.*, "Deep Learning in Medical Ultrasound Analysis: A Review," *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019, doi: 10.1016/j.eng.2018.11.020.

[6] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, "Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning," *Ultrasound in Medicine & Biology*, vol. 46, no. 5, pp. 1119–1132, May 2020, doi: 10.1016/j.ultrasmedbio.2020.01.001.

[7] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, and S. Moccia, "A review on deep-learning algorithms for fetal ultrasound-image analysis," *Medical Image Analysis*, vol. 83, p. 102629, Jan. 2023, doi: 10.1016/j.media.2022.102629.

[8] K. Mikolaj *et al.*, "Removing confounding information from fetal ultrasound images," 2023, *arXiv*. doi: 10.48550/ARXIV.2303.13918.

[9] A. Bunnell, K. Hung, J. A. Shepherd, and P. Sadowski, "BUSClean: Open-source software for breast ultrasound image pre-processing and knowledge extraction for medical AI," *PLoS ONE*, vol. 19, no. 12, p. e0315434, Dec. 2024, doi: 10.1371/journal.pone.0315434.

[10] Q. Zhao *et al.*, "MMOTU: A Multi-Modality Ovarian Tumor Ultrasound Image Dataset for Unsupervised Cross-Domain Semantic Segmentation," 2022, *arXiv*. doi: 10.48550/ARXIV.2207.06799.

[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019, doi: 10.48550/ARXIV.1905.11946.

[12] H. Chen *et al.*, "Deep Learning Prediction of Ovarian Malignancy at US Compared with O-RADS and Expert Assessment," *Radiology*, vol. 304, no. 1, pp. 106–113, Jul. 2022, doi: 10.1148/radiol.211367.

[13] S.-T. Hsu, Y.-J. Su, C.-H. Hung, M.-J. Chen, C.-H. Lu, and C.-E. Kuo, "Automatic ovarian tumors recognition system based on ensemble convolutional neural network with ultrasound imaging," *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 298, Nov. 2022, doi: 10.1186/s12911-022-02047-6.

[14] M. Wu *et al.*, "Deep convolutional neural networks for multiple histologic types of ovarian tumors classification in ultrasound images," *Front. Oncol.*, vol. 13, p. 1154200, Jun. 2023, doi: 10.3389/fonc.2023.1154200.

[15] W. Xie *et al.*, "Developing a deep learning model for predicting ovarian cancer in Ovarian-Adnexal Reporting and Data System Ultrasound (O-RADS US) Category 4 lesions: A multicenter study," *J Cancer Res Clin Oncol*, vol. 150, no. 7, p. 346, Jul. 2024, doi: 10.1007/s00432-024-05872-6.

[16] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[17] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, Nov. 2001, doi: 10.1023/A:1010920819831.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, pp. 2825–2830, 2011.

[19] D. V. Leonov, N. S. Kulberg, A. I. Gromov, S. P. Morozov, and A. V. Vladzimirskiy, "Diagnostic Mode Detecting Solid Mineral Inclusions in Medical Ultrasound Imaging," *Acoust. Phys.*, vol. 64, no. 5, pp. 624–636, Sep. 2018, doi: 10.1134/S1063771018050068.

[20] V. Grebennikova *et al.*, "Design and validation of a technology for 3D printing training phantoms for ultrasound imaging," *Phys Eng Sci Med*, vol. 49, no. 1, pp. 317–332, Mar. 2026, doi: 10.1007/s13246-025-01670-7.

[21] O. Vlasova *et al.*, "Development of an anthropomorphic ultrasound face phantom for training in cosmetic and aesthetic procedures," Phys. Eng. Sci. Med., 2026. DOI: 10.1007/s13246-026-01763-x

[22] J. Li et al., "A Deep Learning Model System for Diagnosis and Management of Adnexal Masses," Cancers, vol. 14, no. 21, p. 5291, Oct. 2022, doi: 10.3390/cancers14215291.

[23] W.-L. Dai et al., "Development and validation of a deep learning pipeline to diagnose ovarian masses using ultrasound screening: a retrospective multicenter study," EClinicalMedicine, vol. 78, p. 102923, Dec. 2024, doi: 10.1016/j.eclinm.2024.102923.

Inertial Sensor Analysis of Voluntary Stopping and Freezing of Gait in Parkinson's Disease

Lígia Reis Nóbrega
Faculty of Electrical Engineer
Federal University of Uberlândia
Uberlândia, Brazil
https://orcid.org/0000-0001-5169-4573

Luiza Maire David Luiz
Faculty of Physiotherapy
Anhanguera
Uberlândia, Brazil
https://orcid.org/0000-0002-4630-6692

Adriano de Oliveira Andrade
Faculty of Electrical Engineer
Federal University of Uberlândia
Uberlândia, Brazil
https://orcid.org/0000-0002-5689-6606

***Abstract*—Voluntary stopping is a deliberate, self-initiated attempt to stop walking. Voluntary stopping requires actively decelerating forward momentum and stabilizing your center of gravity. Involuntary stopping is the unintentional act of stop walking. In Parkinson's disease (PD), this phenomenon of sudden mobility loss is called Freezing of Gait (FOG). To better characterize the FOG episode and gain a deeper understanding of its underlying nuances, this study aims to compare the voluntary and involuntary stop in people with PD with FOG using inertial sensors. Ten people with PD with history of FOG performed four motor tasks with three sensors attached to the body, two on the pelvis, and one on the shank of the side most affected by PD. Sessions were video-recorded and each event was annotated. For the purpose of this work, the voluntary stop and the FOG episode are the events of interest. Tri-axial accelerometer and gyroscope signals were segmented by event and processed in R; for each sensor–axis–signal combination, amplitude metrics (standard deviation, range, and peak count) were calculated. Voluntary stopping and FOG were compared using linear mixed models, with significance assessed via Wald chi-square tests (Type II) and Benjamini-Hochberg correction for multiple comparisons. Of 108 sensor–axis–feature combinations tested, 95 showed significant differences after correction, with FOG episodes consistently showing greater amplitude than voluntary stops across standard deviation, peak count, and range. These results indicate that FOG episodes involve more irregular movement in the lower limbs than deliberate stopping, consistent with the loss of motor control characteristic of freezing. Quantifying these kinematic differences increases the information about FOG episodes and may inform the development of objective, sensor-based methods for detecting and helping alleviate FOG.**

***Index Terms*—Parkinson's disease, Freezing of gait, inertial measure unit, signal processing, voluntary stopping.**

I. Introduction

Freezing of gait (FOG) is one of the most disabling symptoms of Parkinson's disease (PD), which is a brief episodic absence or marked reduction in stride progression, despite the intention to walk [1]. This sudden inability to begin or continue movement can affect performance in daily activities and subject the individual to frequent falls [2].

There are three types of FOG: (I) Akinetic, the blocked gait, when the feet seem to be glued to the ground, without movement in the trunk and limbs, like complete or partial akinesia; (II) Festination, when the normal gait rhythm changes to a shuffling gait, with faster and shorter steps until the stop; and (III) Trembling place, a local tremor when the lower limbs may show signs of tremor while fixed to the ground, with quick and alternating knee movements [1], [8].

According to Bartels et al. [3], the prevalence of people diagnosed with PD in whom the FOG symptom manifests, ranges from 7% in the early stages to 60% in the severe stages. Morris et al. [4] reported that progressively more people who experience FOG restrict walking and reduce their level of physical activity to avoid triggering the motor disorder.

The risk of fall increases because the most frequent presentation of FOG is not the akinetic type in which there is no movement at all. Video studies did, in fact, show that FOG is commonly associated with an effort to overcome the block, the trembling in place or festination types of FOG. Akinetic freezers who accept the block and wait for spontaneous resolution may be less prone to fall [8].

It is of paramount importance to better characterize the FOG episode and gain a deeper understanding of its underlying nuances because its pathophysiology is still not clear enough [3]. So for means of comparison, the involuntary stop referred in this paper is the FOG episode.

There is an effort to develop some technologies that would be able to remove the person from the freeze state or even avoid the FOG episode to happen. These devices are based on external cues that can bring consciousness to the walking process and improve gait performance [4]–[6]. A critical limitation of these cueing systems, however, is their dependence on accurate real-time FOG detection. Detection algorithms often fail to distinguish a freezing episode from other gait interruptions, such as a voluntary stop, which is a deliberate, self-initiated attempt to stop walking. Improving the specificity of these algorithms therefore requires a clear characterization of how FOG episodes differ kinematically from other stopping behaviors that share superficial similarities.

This study aims to compare the voluntary and involuntary stop in people with PD with FOG using inertial sensors by directly comparing the kinematic signatures of voluntary stops and FOG episodes across Inertial Measure Unit (IMU) based analysis, considering both the OFF and ON medication states.

CNPq (National Council for Scientific and Technological Development), FAPEMIG (Minas Gerais Research Support Foundation), and CAPES (Coordination for the Improvement of Higher Education Personnel).

II. METHOD

The study was conducted according to the guidelines of the Declaration of Helsinki, and all protocols were approved by the Ethics Committee of the Federal University of Uberlândia, Brazil, under protocol number 38885720.3.0000.5152. Informed consent was obtained from all subjects involved in the study. The experiment was performed in a place destined for the clinical care of Parkinson's disease patients.

A. Subjects

Ten PD patients with a history of FOG were enrolled for this study; the number of subjects was based on previous works [9], [10]. The experimental group of the present study, named GFOG+, had six female and four male participants.

Table I shows the clinical information about the participants. The Unified Parkinson's disease Rating Scale from de Movement Disorder Society (MDS-UPDRS) evaluates daily living activities (Part II) and the severity of motor symptoms in PD (Part III) [11]. The motor examination was performed during both medication states. The New FOG Questionnaire is a subjective measure of the severity of FOG [12] and the MiniMental test assesses the cognitive functioning.

TABLE I
CLINICAL CHARACTERISTICS OF STUDY PARTICIPANTS. VALUES ARE PRESENTED AS MEAN ± STANDARD DEVIATION. GFOG+: PD PATIENTS WITH FREEZING OF GAIT.

Characteristics	State	GFOG+
Age (years)	–	60.8 ± 7.48
Time since diagnosis (years)	–	12.5 ± 5.75
MDS-UPDRS Part II	–	17.1 ± 10.78
MDS-UPDRS Part III	OFF	57.5 ± 27.66
	ON	41.7 ± 24.47
New FOG-Q	–	19.2 ± 4.7
MiniMental	–	26.6 ± 2.27

B. Technology

The acquisition of motion data was performed using accelerometer and gyroscope sensors [13]–[15]. The Movement Disorder Monitoring System (NetMD) was used to acquire motion data from participants. NetMD is based on the combined action of an Android mobile phone with three Bluetooth smartwatch devices (Sony Smartwatch 3 SWR50) [17]. The smartwatches were attached to the participant's body using adjustable belts on the pelvis and shank [18]. The X-axis is the horizontal axis (side-to-side axis), the Y-axis is the vertical axis (longitudinal axis), and the Z-axis is the depth axis (anterior-posterior axis). A tripod was used to position a standard cell phone camera (iPhone 12 Pro Max) in the environment in order to record the experiment.

C. Dataset

The participants performed four motor tasks. Three valid trials were collected for each task. The PD participants executed the tasks during the OFF-medication state, 12 hours without taking the levodopa medicine; and during the ON-medication state, 30-50 minutes after taking levodopa. The ON state refers to when medication is effectively controlling symptoms and the OFF state occurs when medication effects wear off, leading to a resurgence of PD symptoms [19].

Task 1 is the Voluntary Stop, a modified version of the Timed Up and Go test, in which the participant stays standing for 10 seconds before turn [7]. Task 2 is the standard Timed Up and Go test (TUG), taken as the golden standard physical mobility task for gait analysis in PD [20].

Task 3 the WalkMaze Test (WMT) and Task 4 is the Cognitive WalkMaze Test (CWMT). Both WMT and CWMT are physical mobility motor tasks developed to trigger FOG in PD [18]. The difference between tasks 3 and 4 is that the CWMT involves a cognitive load, requiring participants to walk while simultaneously performing the Digital Monitoring Task (DMT) [21].

This dataset was generated by the research group and the full experimental protocol was published in 2023 [22].

D. Preprocessing

Data from the three-dimensional accelerometer and three-dimensional gyroscope embedded in each smartwatch was acquired with a sampling frequency of 50 Hz [15]–[17]. The data were stored in the Android mobile phone and transferred to a computer for further processing. The inertial data were resampled at a frequency of 100 Hz using spline interpolation in R to improve the temporal resolution.

E. Data annotation

The inertial signals and the video were synchronized to enable data annotation with the aid of visual information. The videos of the experiment were watched by two specialist professionals and the FOG episodes and the task events were manually annotated using ATLAS, a graphical tool for annotating multimodal data [23]. An example of the identified events for the first task is shown below.

- Task 1 - Voluntary Stop (VS): (VS1) stand up; (VS2) start walking; (VS3) first walk for three meters; (VS4) ten seconds standing; (VS5) 180 degree turn; (VS6) second walk for three meters; (VS7) 180 degree turn to sit; and (VS8) sit down.

F. Events of interest

The final dataset is a data frame with 98,934 observations and 12 variables. The first column, file name, indicates the path to the .csv file containing the raw signal, structured as two variables, time and signal. This .csv file is already separated by sensor position, type and axis. Followed by subject ($1 \rightarrow 10$), state (ON/OFF), task ($1 \rightarrow 4$), trial ($1 \rightarrow 3$), imu id (right pelvis, left pelvis, shank), sensor (acc/gyr), axis (x, y, z), event, that is the windowed event showed in the previous subsection, and window, with the initial and final time of the event annotated. To allow the FOG assessment, there is a column named 'qq', a binary value that assumes 0 when no FOG was registered and 1 if there is a FOG episode in that task event, and a column named fog with the annotation

of initial and final time of FOG episodes, in the existence of one.

The events of interest for this study are the FOG events triggered by the four tasks and the voluntary stop, the forth event of the first task of data collection (VS4), in which the participant stays standing for 10 seconds.

Table II shows the number of observations in the data frame of events of interest. GFOG+OFF is before taking levodopa and GFOG+ON is after.

TABLE II
NUMBER OF EVENTS BY STOP TYPE AND MEDICATION STATE

Stop type	Group	n
Voluntary stop	GFOG+OFF	450
Voluntary stop	GFOG+ON	486
FOG	GFOG+OFF	2622
FOG	GFOG+ON	522

G. Feature Extraction

After visualization of the FOG episode and voluntary stop signals, it was assumed that simple features of amplitude could demonstrate statistically significant difference between the stop types. Three descriptive metrics were computed for each sensor–axis–signal combination: standard deviation, range, and peak count.

The standard deviation of the signal within a given time window is defined as

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{1}$$

where x_i is the i-th sample of the signal within the window, $\bar{x}$ is the mean signal value in that window, and N is the total number of samples in the window. Equation 1 quantifies the overall variability of the signal around its mean, with higher values indicating greater amplitude fluctuation.

The range of the signal within the same window is given by

$$R = x_{max} - x_{min}, \tag{2}$$

where x_{max} and x_{min} are, respectively, the maximum and minimum signal values observed within the window. Equation 2 captures the total spread of the signal, regardless of how the values are distributed between these two extremes.

Finally, the peak count of the signal within the window is defined as

$$P = \sum_{i=2}^{N-1} \mathbb{1}\left(x_i > x_{i-1} \wedge x_i > x_{i+1}\right), \tag{3}$$

where x_i is the signal sample at index i, N is the number of samples in the window, and $\mathbb{1}(\cdot)$ is the indicator function, which returns 1 if the evaluated condition is true, and 0 if it is false. In this case, the evaluated condition is $x_i > x_{i-1} \wedge x_i > x_{i+1}$, i.e., that x_i is a local maximum. Equation 3 reflects the frequency of abrupt oscillatory events within the signal, with higher peak counts indicating more irregular movement.

Together, Equations 1–3 were used to describe the amplitude characteristics of voluntary stops and FOG episodes across all sensor–axis–signal combinations, and were subsequently compared between the two event types using linear mixed models.

H. Statistical Analysis

To compare voluntary stopping and FOG episodes across the extracted amplitude metrics, linear mixed models (LMMs) were fitted separately for each sensor–axis–signal combination using the `lmer` function from the *lme4* package in R. Stop type (voluntary stop or FOG) was included as a fixed effect, and subject was included as a random intercept, i.e., (1 | subject), to account for repeated measures within participants.

The significance of the fixed effect of stop type was assessed using a Type II Wald chi-square test, implemented with the `Anova` function from the *car* package. To control the false discovery rate arising from multiple comparisons across the sensor–axis–feature combinations tested, p-values were adjusted using the Benjamini–Hochberg method. Adjusted p-values below 0.05 were considered statistically significant.

III. RESULTS

Figure 1 shows an example of voluntary stop of Subject 1, the linear acceleration from the accelerometer on the shank on the X axis is represented in milisseconds.

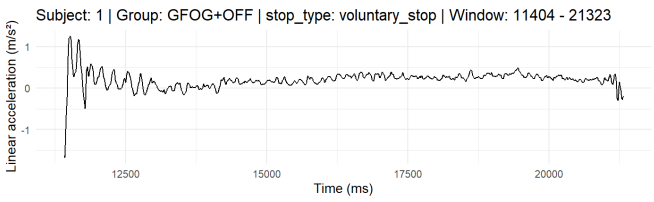


Fig. 1. Linear acceleration on X axis from the right shank of Subject 1.

Figure 2 shows an example of voluntary Stop of Subject 4 during the OFF medication state. The scale of the signal amplitude axis is different from Figure 1.

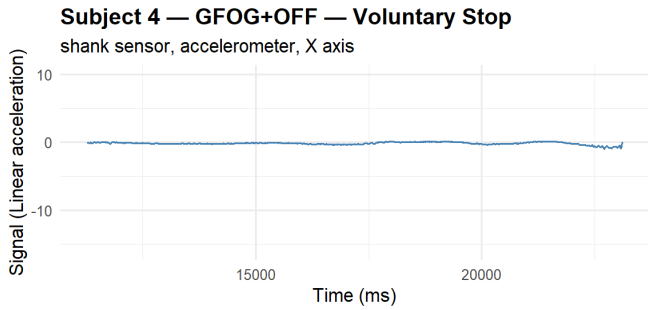


Fig. 2. Voluntary stop of Subject 4. The signal shows the linear acceleration in m/s^2 of the X axis on the shank.

Figure 3 shows an example of FOG episode of Subject 4, that has the akinetic type. This FOG type can manifest as

complete akinesia, in which the entire body is frozen, or partial akinesia, in which only the lower body is frozen.

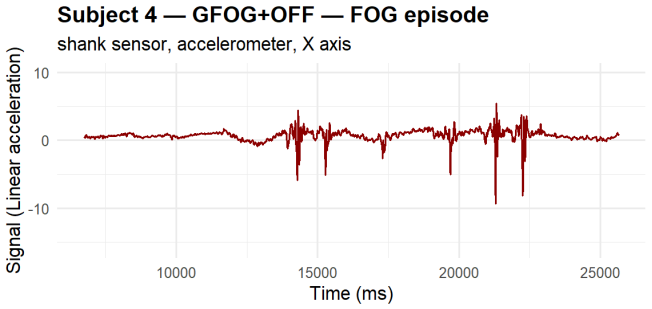


Fig. 3. FOG episode of Subject 4.The signal shows the linear acceleration in m/s^2 of the X axis on the shank.

Even though the walking event is not an event of interest in this paper, we decided to put a signal of example to show the amplitude of walking of Subject 4 during the ON state, in Figure 4, by means of comparison.

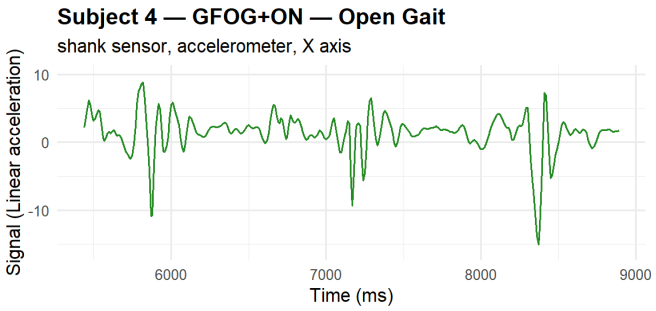


Fig. 4. Behavior of the accelerometer signal during walking in straight line for more than 3 seconds.

Figure 5 shows a comparison between the voluntary and involuntary stop of Subject 9. Both events had similar time duration.

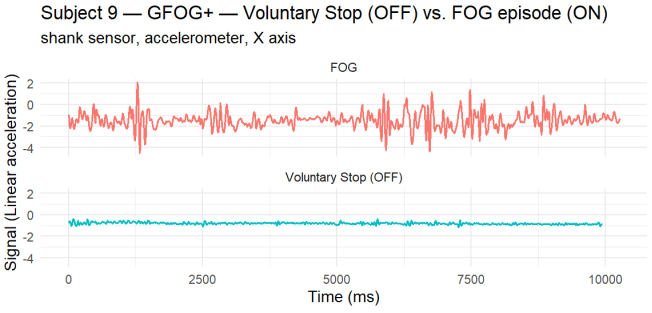


Fig. 5. Involuntary and Voluntary Stop of Subject 9.

Voluntary stopping and involuntary stopping (FOG) were compared using linear mixed models, with stop type as a fixed effect and a random intercept per subject. Significance was assessed via Type II Wald chi-square tests, with Benjamini-Hochberg correction applied across all sensor-axis-feature combinations.

Table III reports the five most significant axis-state combinations for the shank accelerometer, based on the signal standard deviation (sd_val). The estimate is the estimated mean difference in standard deviation between the two conditions (FOG vs. Voluntary Stop). It reflects the effect size of stop type in that model, i.e., how much greater sd_val was during FOG episodes compared to voluntary stops. The p-value column represents the raw test result (Wald chi-square from the lmer model, via car::Anova). The Adj. p-value column shows that same result after applying the Benjamini-Hochberg correction. Each row corresponds to a separate linear mixed model, fitted independently for a given accelerometer axis (X, Y, or Z) and medication state.

TABLE III
TOP 5 SHANK ACCELEROMETER COMBINATIONS BY ADJUSTED P-VALUE (FEATURE: SD_VAL)

Axis	State	Estimate (m/s^2)	p-value	Adj. p-value
X	OFF	1.46	1.45×10^{-15}	2.60×10^{-14}
Z	OFF	1.90	2.02×10^{-15}	3.11×10^{-14}
Z	ON	2.44	2.80×10^{-13}	2.33×10^{-12}
X	ON	1.32	4.24×10^{-13}	3.27×10^{-12}
Y	OFF	1.58	5.40×10^{-12}	2.65×10^{-11}

[a]Shank sensor most affected by Parkinson's disease, accelerometer.

Of the 108 sensor-axis-feature combinations tested, 95 (88.0%) remained statistically significant after correction, all in the direction of FOG episodes exhibiting greater signal amplitude than voluntary stops. Significance rates varied by feature: sd_val showed the highest proportion of significant combinations (34/36), followed by peak (31/36) and range_val (30/36).

Figure 6 shows the boxplots of the sd values for the voluntary stop (blue) and FOG episode (red). The feature exhibited was extracted from the linear acceleration from the accelerometer on the shank. The figure shows the axes X, Y and Z.

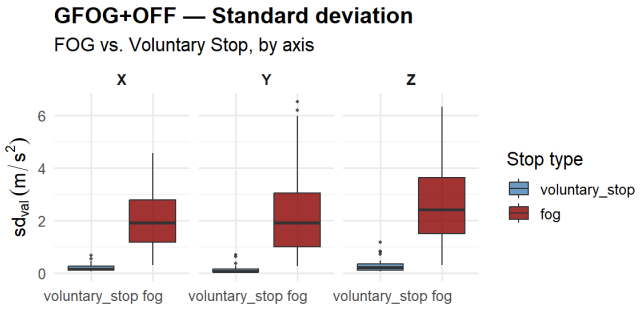


Fig. 6. Distribution of the standard deviation for the shank accelerometer in the OFF medication state. Boxplots compare voluntary stops and FOG episodes on axes X, Y and Z.

An Excel file containing the complete linear mixed model results (estimates, test statistics, raw and adjusted p-values) for all sensor-axis-state-feature combinations is available as Supplementary Material [24].

IV. Discussion

This study demonstrates the potential of inertial measurement units feature extraction and could give insights into the pathophysiology of FOG in PD. Alongside, the findings provide evidence that could inform the detection logic of devices used to aid people with FOG in PD, regardless of medication state.

The inertial sensors allowed to quantify the amplitude of the movements and to better characterize the events. Figure 1 illustrates a typical voluntary stop, in which the deceleration phase is evident between milliseconds 11404 and 13000. Following this interval, it can be inferred that the subject achieved stabilization of the center of gravity.

Figures 1 and 2 show the stability of the signal value, the amplitude of the signal is around 1 m/s^2 during the voluntary stop.

Figures 2, 3, and 4 are all derived from the same subject. Subject 4 presents the rigid-akinetic phenotype of PD and has experienced FOG since the early stages of the disease. The akinetic type is perceived in video recordings because of the lack of movement in the trunk and limbs during the FOG episode.

Figure 2 shows the complete absence of movement during the voluntary stop. Figure 3 illustrates a representative FOG episode from Subject 4 (GFOG+OFF), recorded from the shank accelerometer (X axis) during the dual-task condition. The episode, lasting approximately 20 seconds, is characterized by multiple discrete spikes in signal amplitude. These recurrent spikes likely reflect successive attempts to initiate movement during the FOG episode.

FOG is defined as a brief, episodic absence or marked reduction in stride progression despite the intention to walk. This is evident in Figure 3, where the signal amplitude fails to reach the threshold required to produce effective movement, remaining well below the amplitude range observed during normal walking for the same subject (maximum of 8 m/s^2 and minimum of -15 m/s^2, as shown in Figure 4). It is reasonable to associate the spikes observed at 14,000 ms (-6 m/s^2), 15,500 ms (-6 m/s^2), 21,500 ms (-9.5 m/s^2), and 22,200 ms (-8 m/s^2) in Figure 3 with repeated attempts to initiate walking. This intention to walk has been described in previous literature, but a quantitative demonstration of how it manifests in the signal has, to our knowledge, not yet been shown.

Literature says that the person with an akinetic type of FOG is less prone to fall [1], [8]. However, video recordings of the present study made it noticeable that, after a long episode of akinetic FOG, in which, according to Figure 3 the subject is actively trying to move forward, the person adopts a risky gait pattern characterized by a forward-leaning trunk posture and rapid, shortened steps resembling a light jog. Rapidly the person enters the FOG type II, because this pattern is consistent with festination. After one FOG episode, the whole gait of the subject is compromised, becoming what we call a frozen gait. It differs from the FOG episode that happens briefly and does not prevent starting to walk again. MDS-UPDRS correlates the trouble starting to walk to the need for a cane or walker [11].

In Table III, the correction BH is applied when the original p-value is adjusted for multiple comparisons—in the case of the present study, one model per combination of sensor, sensor placement, axis, and subject medication status.

Table III presents sensor–axis–medication state combinations with the strongest adjusted significance for sd_val at the shank sensor. Most combinations remained highly significant after Benjamini-Hochberg correction (adjusted $p < 10^{-10}$), as it shows, indicating a robust and consistent difference in signal variability between FOG episodes and voluntary stops. Notably, both OFF and ON medication states appear among the top-ranked combinations, suggesting that this discriminative pattern is not restricted to the OFF state, when parkinsonian symptoms are most pronounced. This consistency across medication states reinforces the potential of shank accelerometer for distinguishing FOG from voluntary stopping, regardless of the patient's current dopaminergic status, a relevant consideration for the detection logic of wearable assistive devices.

Figure 5 shows a comparison between a FOG episode and a voluntary stop of similar duration from Subject 9, which motivated the choice of the amplitude-based features extracted in this study. Subject 9 has the trembling place type of FOG, which can be seen in the red signal. The waves are irregular and the signal has intermittent bursts of much higher amplitude than the voluntary stop, which stays nearly flat. Amplitude and frequency vary considerably across the 10-second window of FOG. One may consider that each attempt to break out of the episode has a different intensity, making the entire episode highly heterogeneous.

Figure 6 shows the distribution of sd_val for the shank accelerometer in the OFF-medication state, comparing voluntary stops and FOG episodes across the three axes (X, Y, Z). Across all axes, FOG episodes showed markedly higher sd_val values than voluntary stops. Voluntary stops presented consistently low variability, with narrow interquartile ranges close to zero, reflecting the controlled and stable nature of an intentional stopping movement. In contrast, FOG episodes exhibited substantially wider interquartile ranges and longer whiskers, particularly in the Y and Z axes, indicating greater signal variability during freezing. These results are consistent with the linear mixed-effects model results (Table III), which showed significant differences in sd_val between stop types across all three axes (adjusted $p < 0.001$).

These findings are of particular relevance to better characterize the FOG episode and gain a deeper understanding of its underlying nuances and to optimize real-time systems developed to assist individuals with freezing of gait (FOG) in overcoming or preventing freezing episodes. In particular, these results may support the discrimination of genuine freezing episodes from intentional stopping, thereby contributing to a reduction in false-positive cue activations.

V. Conclusion

It was possible to compare the voluntary and involuntary stop in people with PD with FOG using inertial sensors. Of 108 sensor–axis–feature combinations tested, 95 showed significant differences after correction, with FOG episodes consistently showing greater amplitude than voluntary stops across standard deviation, peak count, and range. These results indicate that FOG episodes involve more irregular movement in the lower limbs than deliberate stopping, consistent with the loss of motor control characteristic of freezing. Quantifying these kinematic differences may inform the development of objective, sensor-based methods for detecting and discriminating FOG from voluntary gait cessation.

Acknowledgment

The present work was carried out with the support of the National Council for Scientific and Technological Development (444437/2024-0, 442150/2023-7, and 405365/2023-3), Coordination for the Improvement of Higher Education Personnel (CAPES), Foundation for Research Support of the State of Minas Gerais (FAPEMIG APQ-05015-24). LRN is a fellow of CNPq (153103/2024-8). AOA is a fellow of CNPq (302942/2022-0).

References

[1] Y. Okuma, "Freezing of gait and falls in Parkinson's disease," J. Parkinsons Dis., vol. 4, no. 2, pp. 255–260, May 2014.

[2] R. C. Moreira, M. B. Zonta, A. P. S. de Araújo, V. L. Israel, and H. A. G. Teive, "Quality of life in Parkinson's disease patients: progression markers of mild to moderate stages," Arq. Neuropsiquiatr., vol. 75, no. 8, pp. 497–502, Aug. 2017.

[3] A. L. Bartels, Y. Balash, T. Gurevich, J. D. Schaafsma, J. M. Hausdorff, and N. Giladi, "Relationship between freezing of gait (FOG) and other features of Parkinson's: FOG is not correlated with bradykinesia," J. Clin. Neurosci., vol. 10, no. 5, pp. 584–588, Sept. 2003.

[4] D. Sweeney, L. R. Quinlan, P. Browne, M. Richardson, P. Meskell, and G. Ólaighin, "A technological review of wearable cueing devices addressing freezing of gait in Parkinson's disease," Sensors, vol. 19, no. 6, p. 1277, Mar. 2019.

[5] E. Ginis, E. Nackaerts, A. Nieuwboer, and E. Heremans, "Cueing for people with Parkinson's disease with freezing of gait: A narrative review of the state-of-the-art and novel perspectives," Ann. Phys. Rehabil. Med., vol. 61, no. 6, pp. 407–413, 2018.

[6] C. Cosentino, M. Putzolu, S. Mezzarobba, M. Cecchella, T. Innocenti, G. Bonassi, A. Botta, G. Lagravinese, L. Avanzino, and E. Pelosin, "One cue does not fit all: A systematic review with meta-analysis of the effectiveness of cueing on freezing of gait in Parkinson's disease," Neurosci. Biobehav. Rev., vol. 150, 2023.

[7] M. E. Morris, F. Huxham, J. McGinley, K. Dodd, and R. Iansek, "The biomechanics and motor control of gait in Parkinson disease," Clin. Biomech., vol. 16, no. 6, pp. 459–470, July 2001.

[8] B. R. Bloem, J. M. Hausdorff, J. E. Visser, and N. Giladi, "Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena," Mov. Disord., vol. 19, no. 8, pp. 871–884, Apr. 2004.

[9] T. Ashfaque Mostafa, S. Soltaninejad, T. L. McIsaac, and I. Cheng, "A comparative study of time frequency representation techniques for freeze of gait detection and prediction," Sensors, vol. 21, no. 19, p. 6446, Sept. 2021.

[10] S. Pardoel, J. Nantel, J. Kofman, and E. D. Lemaire, "Prediction of freezing of gait in Parkinson's disease using unilateral and bilateral plantar-pressure data," Front. Neurol., vol. 13, Apr. 2022.

[11] C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results," Mov. Disord., vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[12] A. Nieuwboer, L. Rochester, T. Herman, W. Vandenberghe, G. E. E. Thomaes, and N. Giladi, "Reliability of the new freezing of gait questionnaire: agreement between patients with Parkinson's disease and their carers," Gait Posture, vol. 30, no. 4, pp. 459–463, Nov. 2009.

[13] J. Zhang, L. Wei, X. Hu, B. Xie, Y. Zhang, G.-R. Wu, and J. Wang, "Akinetic-rigid and tremor-dominant Parkinson's disease patients show different patterns of intrinsic brain activity," Parkinsonism Relat. Disord., vol. 21, no. 1, pp. 23–30, Jan. 2015.

[14] J. O'Day, M. Lee, K. Seagers, S. Hoffman, A. Jih-Schiff, Ł. Kidziński, S. Delp, and H. Bronte-Stewart, "Assessing inertial measurement unit locations for freezing of gait detection and patient preference," J. NeuroEng. Rehabil., vol. 19, no. 1, Feb. 2022.

[15] W. Qi, "Digital biomarkers in Parkinson's disease: a bibliometric analysis and scope review of deep learning models for gait freezing detection," OSF Registries, 2025.

[16] M. Gregorčič and D. Georgiev, "The usefulness of wearable sensors for detecting freezing of gait in Parkinson's disease: a systematic review," Sensors, vol. 25, no. 16, p. 5101, Aug. 2025.

[17] R. López-Blanco, M. A. Velasco, A. Méndez-Guerrero, J. P. Romero, M. D. del Castillo, J. I. Serrano, J. Benito-León, F. Bermejo-Pareja, and E. Rocon, "Essential tremor quantification based on the combined use of a smartphone and a smartwatch: the NetMD study," J. Neurosci. Methods, vol. 303, pp. 95–102, June 2018.

[18] L. R. Nóbrega, E. Rocon, A. A. Pereira, and A. O. Andrade, "A novel physical mobility task to assess freezers in Parkinson's disease," *Healthcare (Basel)*, vol. 11, no. 3, p. 409, Jan. 2023, doi: 10.3390/healthcare11030409.

[19] P.-K. Yang, B. Filtjens, P. Ginis, M. Goris, A. Nieuwboer, M. Gilat, P. Slaets, and B. Vanrumste, "Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops," J. NeuroEng. Rehabil., vol. 21, no. 1, Feb. 2024.

[20] D. Podsiadlo and S. Richardson, "The timed 'up & go': a test of basic functional mobility for frail elderly persons," J. Am. Geriatr. Soc., vol. 39, no. 2, pp. 142–148, Feb. 1991.

[21] E. N. Beck, K. A. Ehgoetz Martens, and Q. J. Almeida, "Freezing of gait in Parkinson's disease: an overload problem?," PLOS ONE, vol. 10, no. 12, Dec. 2015.

[22] L. Nóbrega, "A method to assess freezing of gait in Parkinson's disease with inertial sensors," Ph.D. dissertation, EDUFU - Editora da Universidade Federal de Uberlandia, 2023.

[23] S. Meudt, L. Bigalke, and F. Schwenker, "Atlas - annotation tool using partially supervised learning and multi-view co-learning in human-computer-interaction scenarios," in Proc. 2012 11th Int. Conf. Inf. Sci., Signal Process. Appl. (ISSPA), July 2012, pp. 1309–1312.

[24] L. R. Nóbrega and A. de O. Andrade, "Linear mixed model results - freezing of gait," Zenodo, 2026. doi: 10.5281/zenodo.21674980.

Interpretable Structural-Deviation Descriptors of Osteosarcoma Radiographic Signs via the Universal Image Quality Index

A localization-conditioned feature-characterization study on public radiographs

1st Luiz Felipe Dias da Silva
Dept. of Biomedical Engineering
UFPE
Recife, Brazil
luizfelipe.silva@ufpe.br

2nd Antônio M. G. de Souza
Radiology and Oncology
HCP
Recife, Brazil
amgsouza60@gmail.com

3rd Emery C. Lins
Dept. of Biomedical Engineering
UFPE
Recife, Brazil
emery.lins@ufpe.br

***Abstract*—Osteosarcoma is the most common primary malignant bone tumor in adolescents, and conventional radiography remains the first-line modality in resource-limited settings. Deep learning attains strong classification performance on bone-tumor radiographs but behaves as a black box and demands large annotated datasets and specialized hardware. This work characterizes an interpretable, low-cost feature set derived from classical digital image processing, built around the Universal Image Quality Index (UIQI). The UIQI is a full-reference structural measure that factors into loss of correlation, luminance distortion, and contrast distortion; we repurpose it as a descriptor of the local structural deviation between a lesion region of interest (ROI) and a mirrored intra-image reference region, and we state—as a hypothesis, not as a validated finding—a correspondence between its three components and classical radiological signs of malignancy. To test whether the descriptors carry class-discriminative signal, an interpretable random forest is applied to them. On a binary osteosarcoma-versus-normal-bone task derived from the public multi-center BTXRD dataset (297 osteosarcomas, 1879 normal radiographs; image-level stratified splits of 1740/218/218), and adopting a symmetric extraction that controls a pipeline confound, the UIQI features reach an accuracy of 0.878 ± 0.025 and an area under the receiver operating characteristic curve (AUC-ROC) of 0.872 ± 0.020 under five-fold cross-validation; adding complementary handcrafted descriptors of texture and gradient raises the AUC-ROC to 0.931 ± 0.019 and the F1-score to 0.691 ± 0.079. Three scope conditions bound these numbers: lesion ROIs are placed using the reference segmentation, so performance is conditioned on known lesion location and does not describe an end-to-end automatic system; benign and non-osteosarcoma malignant tumors, which form the hardest differential diagnoses, are excluded; and no external validation, radiologist comparison, calibration analysis, or benchmark against explainable deep models is reported. The contribution should therefore be read as the characterization of an interpretable descriptor set and as the demonstration of an extraction confound, not as a validated diagnostic or triage tool.**

***Index Terms*—osteosarcoma, digital image processing, universal image quality index, structural similarity, radiomics, texture analysis, interpretable features, bone tumor**

This work was supported by the Federal University of Pernambuco (UFPE) and the Cancer Hospital of Pernambuco (HCP).

I. Introduction

Osteosarcoma is the most prevalent primary malignant bone tumor in pediatric and young-adult populations, accounting for about 20% of primary bone sarcomas, with peak incidence during adolescence [5]. Because advances in neoadjuvant chemotherapy and limb-salvage surgery have substantially raised five-year survival, early and accurate diagnosis is decisive for treatment planning and prognosis. Conventional radiography remains the first-line modality for the initial evaluation of bone tumors owing to its wide availability, low cost, and ability to reveal key signs such as bone destruction, periosteal reaction (for example, the sunburst pattern and Codman's triangle), and soft-tissue extension.

Radiographic interpretation of bone tumors is nonetheless difficult. Features that overlap with benign entities such as osteomyelitis, healing fractures, and osteochondroma, subtle early-stage changes, and inter-observer variability between radiologists of differing experience hinder consistent diagnosis. These limitations motivate computer-aided diagnosis (CAD) research aimed at supporting radiologists. In the Brazilian public health system, a shortage of specialized radiologists and geographic disparities in the distribution of expertise amplify these challenges, reinforcing reliance on conventional radiographs as the principal diagnostic tool in primary and secondary care [10].

Deep learning has achieved strong performance in bone-tumor classification [7], [8]. However, such models are data-hungry, computationally demanding, and largely opaque: they report a probability without exposing which radiographic sign drove the decision. Post-hoc explanation methods partially mitigate this, but the case for models that are interpretable by construction in high-stakes clinical settings remains strong [11]. This motivates investigating whether descriptors with explicit physical meaning can encode the information radiologists actually name when reading a bone lesion.

This work therefore studies osteosarcoma radiographs as

a problem of quantifying structural deviation with classical digital image processing (DIP). Our central instrument is the Universal Image Quality Index (UIQI) of Wang and Bovik [1], a full-reference measure used here not to assess image fidelity but repurposed as a structural-deviation descriptor: it compares a suspect lesion ROI with a reference bone region and factors the deviation into three components with distinct physical meanings. The contributions of this paper are:

- A compact, interpretable feature set (20 descriptors) derived entirely from the local UIQI decomposition between a lesion ROI and a mirrored intra-image reference, in which every dimension has an explicit physical meaning.
- The identification, quantification, and control of an ROI-extraction confound that inflates apparent performance by roughly 0.10 AUC-ROC when the two classes follow geometrically distinct extraction paths—a failure mode that, to our knowledge, is not routinely reported in handcrafted bone-radiograph pipelines.
- A hypothesis, stated explicitly as such, mapping the three UIQI components onto named radiological signs, together with a discussion of the sign-level annotations that would be required to test it.

A. Scope and what this paper does not claim

Because the framing of this kind of study is easily overstated, we state the boundaries up front. (i) The lesion ROI is placed using the reference segmentation provided with the dataset; no lesion detector is part of the pipeline. All reported metrics are therefore *conditioned on correct lesion localization* for positive cases and describe the discriminative power of the descriptors, not the performance of an automatic triage system. (ii) The task is binary (osteosarcoma versus radiographically normal bone); benign tumors and non-osteosarcoma malignancies, which in practice constitute the difficult differential, are excluded. (iii) The correspondence between UIQI components and radiological signs is a working hypothesis: the dataset contains no sign-level annotation against which it could be verified. (iv) No external cohort, no reader study, no calibration assessment, and no comparison against convolutional or transformer baselines are reported. Statements about deployment on commodity hardware concern computational cost only, and are not claims of clinical readiness.

II. Related Work

Full-reference image quality metrics were introduced to predict perceived fidelity. The UIQI [1] models distortion as the product of loss of correlation, luminance distortion, and contrast distortion, and inspired a broad family of structural and perceptual metrics, including the Structural Similarity Index [2]. Although designed for quality assessment, these structural measures have been adopted as region-comparison and change-detection descriptors across imaging domains, which motivates their use here as descriptors of the local structural deviation between suspect and reference bone.

Classical radiomic and texture analysis remains widely used in musculoskeletal imaging [3]. Gray-level co-occurrence matrix (GLCM) features capture second-order texture such as trabecular disorganization and matrix heterogeneity, while local binary patterns encode micro-texture robustly to illumination [4].

Prior work on bone-tumor radiographs is predominantly based on deep-learning classifiers, frequently trained on private datasets [8], [9]. Interpretability in that line of work is typically obtained post hoc, through gradient-based saliency such as Grad-CAM [12], or, in transformer architectures [7], [13], through attention-based visualizations. Such maps indicate *where* the model looked, but not *which radiographic property* was abnormal there, and their faithfulness to the underlying decision is itself contested [11]. The present work sits on the opposite side of that trade-off: it computes named quantities directly and accepts a lower ceiling on predictive performance in exchange. We emphasize that we do not benchmark against explainable deep models or vision transformers here. A fair comparison would require training such models under the identical localization-conditioned protocol and image-level splits used below; absent that, this paper makes no claim about the relative accuracy of handcrafted structural descriptors and modern deep architectures, and the interpretability–performance trade-off remains uncharacterized. We flag this comparison as necessary future work in Section VIII.

Reporting for machine-learning studies in medical imaging is increasingly guided by checklists such as CLAIM [15]; Section III-A follows its recommendations on data provenance, case selection, unit of analysis, and partition definition.

III. Materials and Methods

A. Dataset, case selection, and study design

Experiments use the Bone Tumor X-Ray Radiograph Dataset (BTXRD), a three-center collection released in 2024 for bone-tumor analysis [6]. The full release contains 3746 radiographs (1879 radiographically normal, 1525 benign tumor, and 342 malignant tumor), provided as JPEG images in an `images` folder, with segmentation polygons and bounding boxes in LabelMe (`.json`) format in an `Annotations` folder, and image-level labels and clinical metadata consolidated in `dataset.csv`. Each tumor image carries a segmentation polygon and a bounding box; normal images carry no annotation. Acquisition parameters (device, exposure, post-processing) are not standardized across the three contributing centers, and the release does not expose them per image; the heterogeneity is therefore inherited by our pipeline and is not corrected beyond the intensity normalization described in Section III-B. Detailed patient-level characteristics of the source cohorts are given in the original dataset publication [6].

Case and control selection. From the 3746 radiographs we retained the 297 images labeled as osteosarcoma (a subset of the 342 malignant tumors; the remaining 45 form the other-malignant category) and the 1879 normal-bone images, giving 2176 radiographs with an approximate 1:6.3 positive-to-negative ratio (Table I). The 1525 benign tumors and the 45 non-osteosarcoma malignancies were excluded by design,

to isolate the structural contrast between osteosarcoma and normal bone. We stress that this exclusion removes precisely the categories that dominate the clinical differential diagnosis, so the reported separability is an optimistic bound with respect to the multiclass problem a radiologist actually faces.

Unit of analysis and partition unit. The unit of analysis is the radiograph. The public release used here does not provide a patient identifier that would allow multiple radiographs of the same patient (for instance, frontal and lateral views of the same limb) to be grouped; consequently all partitions are performed at the image level, and we cannot rule out that images of the same patient appear in different partitions. This is a source of optimistic bias that we could not eliminate and that patient-level grouping would resolve; it is restated in Section VII.

Partitions. Two protocols are used. Protocol (i) is a single stratified image-level split into 1740 training (237 osteosarcoma / 1503 normal), 218 validation (30 / 188), and 218 held-out test images (30 / 188). Protocol (ii) is stratified five-fold cross-validation over all 2176 images, reporting mean, standard deviation, and an interval across folds, adopted because the number of positives (297) is modest. Feature definitions, classifier hyperparameters, and the decision threshold were fixed using training and validation data only, as detailed in Section III-G; the held-out test set of protocol (i) was evaluated once.

TABLE I
COMPOSITION OF THE BINARY SUBSET EXTRACTED FROM BTXRD

Class	Total	C1	C2	C3	F/L/O
Osteosarcoma	297	215	66	16	187/89/21
Normal bone	1879	–	–	–	–
Total	2176				

C1–C3: acquisition centers; F/L/O: frontal, lateral, and oblique views. Counts are radiographs, not patients; patient-level grouping is not available in the release.

B. Preprocessing

Each radiograph is converted to single-channel intensity and preprocessed to standardize acquisition variability while preserving diagnostic structure: (1) intensity normalization to the range 0–1 by clipping between the first and ninety-ninth percentiles, reducing the influence of outliers; and (2) contrast-limited adaptive histogram equalization (CLAHE, clip limit 0.01) to enhance trabecular and cortical detail. Both operations are applied identically to every image, independently of class and of any label, and their parameters were fixed a priori rather than selected by performance. We note that CLAHE alters local contrast and therefore acts directly on the quantity that the contrast component of the UIQI measures; because it is applied identically to the lesion region and to its reference within the same image, it does not by itself induce a systematic difference between the two, but it does compress the absolute interpretation of the contrast term. Region-of-interest definition follows Section III-E.

C. The Universal Image Quality Index as a structural-deviation descriptor

Let $x = \{x_i\}$ and $y = \{y_i\}$, $i = 1, \ldots, N$, be co-located samples from the reference bone region and the suspect lesion region, respectively. The UIQI [1] is

$$Q = \frac{4\,\sigma_{xy}\,\bar{x}\,\bar{y}}{(\sigma_x^2 + \sigma_y^2)\,(\bar{x}^2 + \bar{y}^2)}, \tag{1}$$

where $\bar{x}, \bar{y}$ are the sample means, σ_x^2, σ_y^2 the variances, and σ_{xy} the covariance. As shown by the authors, (1) factors into three interpretable terms:

$$Q = \underbrace{\frac{\sigma_{xy}}{\sigma_x \sigma_y}}_{Q_c\ \text{(corr.)}} \cdot \underbrace{\frac{2\bar{x}\bar{y}}{\bar{x}^2 + \bar{y}^2}}_{Q_l\ \text{(lum.)}} \cdot \underbrace{\frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}}_{Q_s\ \text{(contr.)}}, \tag{2}$$

with Q in the range $[-1, 1]$ and $Q = 1$ if and only if $x_i = y_i$ for all i. Following [1], Q is not computed globally but with a sliding window of size $B \times B$ (with $B = 8$) over the ROI pair, and the overall index is the mean of the local values, yielding a spatial structural-deviation map $Q(u, v)$ and its component maps $Q_c(u, v)$, $Q_l(u, v)$, $Q_s(u, v)$ (Fig. 1).

D. Hypothesized correspondence with radiographic signs

We propose, as a working hypothesis, that each component tracks a distinct family of radiological findings, as summarized in Table II. A decrease in the correlation term Q_c would be expected where the spatial coherence of the trabecular pattern is broken, as in trabecular disorganization or periosteal reaction. A decrease in the luminance term Q_l would be expected where mean density departs from the reference, as in osteolysis or osteoblastic mineralization. A decrease in the contrast term Q_s would be expected where the local dynamic range changes, as in cortical destruction or a heterogeneous tumor matrix.

This correspondence is stated as a hypothesis and is not verified in this work. The three components are mathematical functions of local means, variances, and covariance, and each of them can also be driven by non-pathological factors: exposure and detector response affect Q_l; soft-tissue superposition, projection obliquity, and CLAHE affect Q_s; noise and mirroring misalignment affect Q_c. Testing the hypothesis requires radiographs annotated at the level of individual signs—for example, expert delineation of osteolytic regions, of periosteal reaction, and of cortical breach—which BTXRD does not provide. Accordingly, throughout this paper we say that a component *may reflect* a sign, and we do not treat a low Q_l as a measurement of osteolysis or a low Q_s as a measurement of cortical destruction.

E. Reference-region definition, symmetric extraction, and confound control

The definition of the two regions is critical and can, if poorly designed, introduce bias. A naive approach extracts the lesion ROI from the mask for tumor images and a generic ROI for normal images; the two classes then follow geometrically distinct extraction paths, and the classifier can learn the

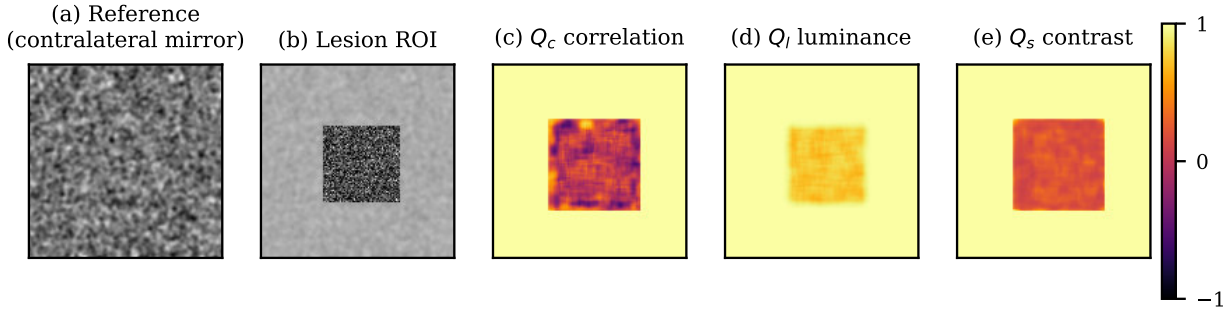


Fig. 1. Synthetic illustration of the UIQI used as a structural-deviation descriptor. (a) Reference region; (b) suspect region, here constructed with a radiolucent area and a heterogeneous matrix; (c)–(e) local maps of the components Qc, Ql, and Qs. This figure is a didactic construction from synthetic data, not a patient radiograph; it illustrates how the decomposition localizes and factors a deviation and is not evidence for the correspondence hypothesized in Table II.

TABLE II
HYPOTHESIZED CORRESPONDENCE BETWEEN UIQI COMPONENTS AND OSTEOSARCOMA RADIOGRAPHIC SIGNS (NOT VALIDATED AGAINST SIGN-LEVEL ANNOTATIONS)

Component	**Image property measured**	**Candidate radiological sign (hypothesis)**
Q_c (correlation)	Structural coherence	Trabecular disorganization; periosteal reaction (sunburst, Codman)
Q_l (luminance)	Mean density	Osteolysis / osteoblastic sclerosis
Q_s (contrast)	Local dynamic range	Cortical destruction; heterogeneous matrix

The middle column states what the term computes; the right column states what we hypothesize it may reflect. Only the middle column is established.

pipeline signature (ROI size, position, overlap) rather than the pathology. To remove this confound we adopt a *symmetric extraction*: every image, normal or tumoral, undergoes the same operation, comparing a square box against the region obtained by reflecting the same box coordinates about the vertical midline of the image. For tumor images the box is positioned from the lesion annotation; for normal images it is positioned at the center of the bone. The only systematic difference remaining between the classes is thus the presence of pathological tissue inside the box. Section V quantifies how much of the apparent performance the naive variant owed to the extraction path.

What the mirrored region is, and when the symmetry assumption fails. Most bone radiographs in BTXRD image a single limb or region, so the reflected box is generally *not* a true contralateral bone. It is a within-image control region acquired under identical exposure, detector, and processing conditions, which is why we call it a *mirrored intra-image reference* rather than a contralateral analog. No anatomical registration, rotation correction, or laterality normalization is applied, and no test verifies that the reflected box actually falls on bone. Four situations degrade the reference: (a) the limb is not centered or is rotated, so the reflected box samples a different anatomical level; (b) the reflected box falls on soft tissue or on background, in which case it is clipped to the image bounds and the resulting deviation reflects the tissue mismatch rather than pathology; (c) the lesion is large or near the midline, so the reflected box partly overlaps the lesion itself and the deviation is attenuated, biasing positives toward the negative class; and (d) periosteal reaction or soft-tissue mass extends beyond the annotated box. Because the identical operation is applied to both classes, this noise is not class-specific by construction and does not reintroduce the extraction confound, but it is not eliminated: it inflates the within-class variance of all descriptors and is a plausible contributor to the limited recall reported in Section V. Quantifying it would require a per-image assessment of reference validity, which we did not perform.

F. Feature set

For each ROI pair we compute the four UIQI maps (Q, Q_c, Q_l, Q_s) and summarize each by five descriptors—mean, minimum, standard deviation, fifth percentile, and altered-area fraction (fraction of pixels below a fixed threshold)—totaling 20 low-dimensional features. The minimum and the fifth percentile capture the worst local point of the suspect region, whereas the altered-area fraction quantifies the extent of the deviation.

Optionally, a complementary set of 24 handcrafted descriptors is extracted from the lesion ROI to capture signals the UIQI does not encode (Fig. 2): five first-order intensity statistics (mean, variance, skewness, kurtosis, entropy); five GLCM texture measures [3] (contrast, correlation, energy, homogeneity, angular second moment) averaged over four orientations; ten bins of the uniform local binary pattern histogram [4] ($P = 8$, $R = 1$); and four gradient and edge descriptors (mean and standard deviation of the Sobel gradient magnitude, its ninety-fifth percentile, and the Canny edge density).

All descriptors are produced by fixed closed-form operators with parameters chosen a priori. No label-driven feature selection, ranking, or dimensionality reduction was performed at any stage, so no information from the validation or test data entered the feature definition.

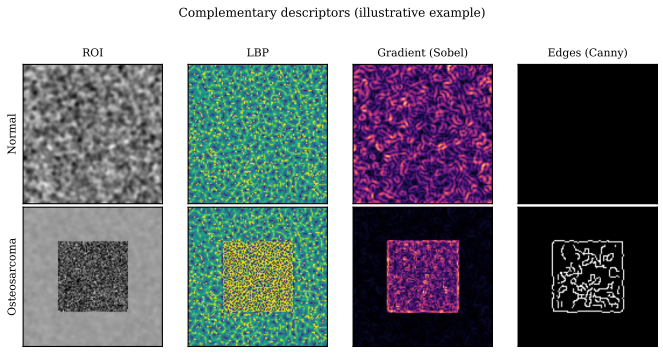


Fig. 2. Complementary descriptors extracted from the ROI, shown for one normal and one osteosarcoma case selected for illustration. Columns: ROI, local binary pattern (LBP) map, Sobel gradient magnitude, and Canny edges. The figure conveys qualitative intuition for what the descriptors encode; being a two-case example, it is not evidence of a consistent group difference. Group-level distributions are given in Fig. 4.

G. Classification protocol

The features (20 from the UIQI, or 44 with the complementary block) are classified by a random forest with 400 trees and balanced class weighting, chosen for robustness on small tabular feature sets and for exposing importance rankings. Tree depth, split criterion, and the remaining hyperparameters are the library defaults. These settings were fixed a priori: no grid search, randomized search, or manual tuning was carried out, so no hyperparameter information leaked from the validation or test partitions. In protocol (i) the decision threshold is chosen by maximizing F1 on the validation set alone and applied unchanged to the held-out test set. In protocol (ii) we report AUC-ROC, which is threshold-independent, together with point metrics computed at the fixed default threshold of 0.5, without per-fold threshold optimization. The classifier serves to test the separability of the descriptors; it is not itself the contribution, and no attempt was made to optimize it.

H. Evaluation metrics

Let TP, TN, FP, and FN denote true and false positive and negative counts for the osteosarcoma class. We report

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \quad \text{Prec} = \frac{TP}{TP+FP}, \tag{3}$$

$$\text{Rec} = \frac{TP}{TP+FN}, \quad \text{Spec} = \frac{TN}{TN+FP}, \tag{4}$$

$$F_1 = \frac{2\cdot\text{Prec}\cdot\text{Rec}}{\text{Prec}+\text{Rec}} = \frac{2TP}{2TP+FP+FN}, \tag{5}$$

together with the threshold-independent AUC-ROC. Given the class imbalance (1:6.3), accuracy is reported for completeness only and is dominated by the negative class; F1, recall, and AUC-ROC carry the informative content.

For protocol (ii), let $m_1, \ldots, m_5$ be the per-fold values of a metric, with mean $\bar{m}$ and standard deviation s. We report $\bar{m} \pm s$ and the interval

$$\bar{m} \pm t_{0.975,\,4}\,\frac{s}{\sqrt{5}} \;=\; \bar{m} \pm 1.24\,s, \tag{6}$$

with $t_{0.975,4} = 2.776$. This interval summarizes fold-to-fold variability of the estimate within this dataset; because the folds are drawn from a single cohort and are not independent samples of the target population, it must not be read as a confidence interval for performance on new institutions or populations.

IV. Experimental Setup

The pipeline is implemented in Python with open-source libraries (NumPy, SciPy, scikit-image, scikit-learn) and runs entirely on the CPU, without a GPU, on a commodity workstation. The UIQI is computed with an 8-by-8 sliding window and the random seed is fixed for reproducibility; library versions and the seed are reported alongside the released code. No GPU-specific or accelerator-specific operation is used at any stage.

V. Results

A. Confound control in extraction

Table III contrasts the naive extraction (distinct geometric paths per class) with the symmetric extraction of Section III-E, on the same held-out test set of 218 images (30 osteosarcoma, 188 normal). The naive variant yields markedly high metrics (AUC-ROC 0.981; F1 0.815), but a large share of that performance is attributable to the pipeline signature: equalizing the extraction operation across classes lowers the AUC-ROC to 0.882 and the F1-score to 0.511 on the identical split, driven mainly by a collapse in recall from 0.733 to 0.400. Specificity is high in both variants, which is expected under a 1:6.3 imbalance and is not by itself informative. We therefore report the symmetric variant as the main result and regard the naive figures as an artifact, included only to document the size of the effect. Given the 30 positives in this split, all point estimates in Table III rest on few events and are correspondingly unstable; the cross-validated results of Table IV are the more reliable summary.

TABLE III
Effect of Confound Control on the Same Held-Out Test Split (218 Images: 30 Osteosarcoma, 188 Normal)

Metric	Naive (biased)	Symmetric
TP / FP / FN / TN	22 / 2 / 8 / 186	12 / 5 / 18 / 183
Accuracy	0.954	0.895
AUC-ROC	0.981	0.882
Precision	0.917	0.706
Recall	0.733	0.400
Specificity	0.989	0.973
F1-score	0.815	0.511

Single split, threshold selected on the validation set only. The naive column is reported to quantify the confound, not as a result of the method.

B. Classification performance and ablation

Table IV reports five-fold cross-validation with symmetric extraction, for the UIQI features alone and for the extended set including the complementary texture and gradient descriptors. The UIQI alone reaches an AUC-ROC of 0.872 ± 0.020 and an F1-score of 0.560 ± 0.024, indicating that the structural

decomposition separates the two classes with a consistent, if moderate, signal. Adding the complementary descriptors raises the AUC-ROC to 0.931 ± 0.019 and the F1-score to 0.691 ± 0.079, with the largest gain in recall (0.570 to 0.681), the metric where the UIQI alone was weakest. The intervals of the two feature sets do not overlap for AUC-ROC or recall, while they do overlap for precision, whose fold-to-fold variability is large. Fig. 3 summarizes the ablation and shows the test confusion matrix.

Two caveats bound the reading of this ablation. First, we did not evaluate the complementary texture and gradient block *in isolation*; consequently these experiments do not establish that the UIQI contributes information beyond what a conventional GLCM/LBP/gradient baseline already provides, and the gain observed here is compatible with the complementary block carrying most of the discriminative weight. An isolated-baseline comparison is required before any claim of complementarity can be made, and we abstain from that claim. Second, the predicted probabilities of the random forest were not calibrated or assessed for calibration; the threshold-dependent metrics therefore describe ranking behaviour at a fixed operating point rather than reliable probability estimates.

TABLE IV
CLASSIFICATION PERFORMANCE AND ABLATION (5-FOLD STRATIFIED CV, SYMMETRIC EXTRACTION, IMAGE-LEVEL FOLDS)

	UIQI (20 feat.)		**UIQI + Compl. (44 feat.)**	
Metric	Mean±SD	95% int.	Mean±SD	95% int.
Accuracy	0.878 ± 0.025	[0.847, 0.909]	$\mathbf{0.914 \pm 0.033}$	[0.873, 0.955]
AUC-ROC	0.872 ± 0.020	[0.847, 0.897]	$\mathbf{0.931 \pm 0.019}$	[0.907, 0.955]
Precision	0.581 ± 0.083	[0.478, 0.684]	$\mathbf{0.716 \pm 0.138}$	[0.545, 0.887]
Recall	0.570 ± 0.117	[0.425, 0.715]	$\mathbf{0.681 \pm 0.041}$	[0.630, 0.732]
F1-score	0.560 ± 0.024	[0.530, 0.590]	$\mathbf{0.691 \pm 0.079}$	[0.593, 0.789]

Intervals are Student-t intervals over the five folds ($\bar{m} \pm 1.24\,s$) and describe fold-to-fold variability within this dataset only. Folds are stratified at the image level; patient-level grouping is not available.

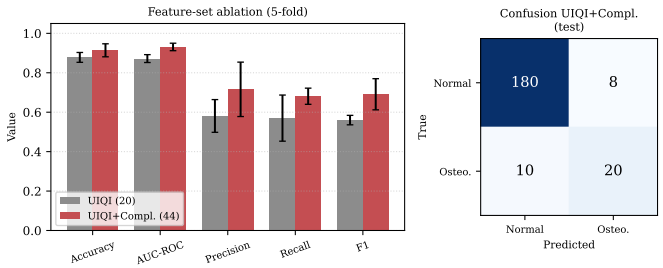


Fig. 3. Left: feature-set ablation (UIQI versus UIQI plus complementary) over five folds, with standard-deviation bars. Right: confusion matrix on the held-out test split for the UIQI plus complementary model with symmetric extraction; counts refer to the 218 test images described in Section III-A.

C. Descriptor distributions and model importances

Fig. 4 shows the class-wise distributions of key descriptors over the 2176 radiographs. The distributions of the minimum values of Q and Q_c differ between classes, with osteosarcoma ROIs concentrating at more pronounced local deviations; osteosarcoma ROIs also show a larger fraction of area with altered luminance and a lower mean contrast term. The direction of these shifts is consistent with what the hypothesis of Table II would predict, but consistency with a hypothesis is not a test of it: the same shifts would arise from exposure differences or from reference-region mismatch, and no sign-level annotation is available to discriminate between these explanations.

The random-forest importances are dominated by the minimum descriptors of the components (Q_{min}, $Q_{l,min}$, $Q_{s,min}$, $Q_{c,min}$), suggesting that the fitted model relies chiefly on the worst local point of the suspect region rather than on region-wide averages. This should be read descriptively and with care. Impurity-based importances are known to be biased toward continuous predictors and to distribute importance arbitrarily among correlated ones [14], and the 20 UIQI descriptors are strongly correlated by construction, since all four maps derive from the same local statistics and the five summaries of each map are functions of the same distribution. The ranking therefore describes one fitted model on this dataset; it does not establish that one radiographic sign is intrinsically more informative than another, and we draw no clinical conclusion from it. Permutation importance with grouped features, or a conditional importance scheme, would be needed for that.

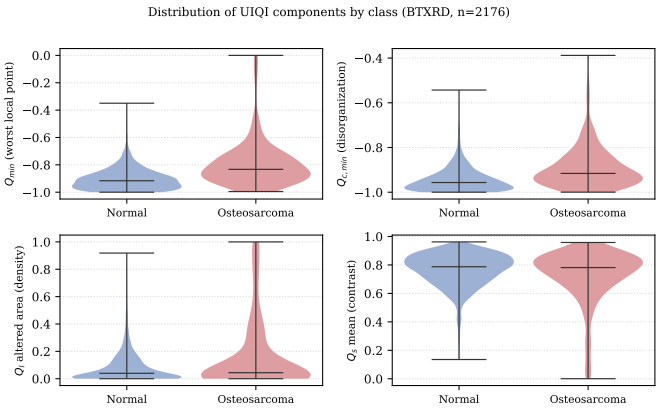


Fig. 4. Class-wise distributions of the UIQI component descriptors (BTXRD, $N = 2176$ radiographs; patient-level grouping unavailable): normal bone (blue) versus osteosarcoma (red).

VI. DISCUSSION

The framework studied here treats osteosarcoma radiographs as a structural-deviation problem and asks whether descriptors with explicit physical meaning retain discriminative signal. Two findings support an affirmative but bounded answer.

The first is methodological and, we argue, the more transferable. When the two classes follow distinct ROI-extraction paths, the resulting classifier attains an AUC-ROC of 0.981—a figure that would look publishable and is largely an artifact of the pipeline geometry. Equalizing the extraction operation across classes removes roughly 0.10 AUC-ROC and 0.30 F1. The confound is subtle because nothing in the code is

obviously wrong: it arises from the fact that positives and negatives are cropped by different rules. Any handcrafted pipeline that uses annotation-derived regions for one class and heuristic regions for the other is exposed to it, and we recommend that such pipelines report the symmetric-extraction control as a matter of course.

The second is that, under the corrected extraction, the UIQI descriptors still separate the classes with an AUC-ROC of about 0.87, rising to about 0.93 with classical texture and gradient descriptors. We interpret this as a localization-conditioned quantity. The lesion ROI for positives is placed from the reference annotation, so the number characterizes the descriptors given a correct region, not a system that must find the region first. A fully automatic pipeline would need a candidate-region detector applied identically to both classes, and its errors would compound with those measured here; the end-to-end figure would be lower, by an amount this study cannot estimate. For the same reason, we avoid the word screening for what is reported: nothing here has been evaluated as a triage instrument.

On interpretability, we deliberately weakened the claim relative to our initial framing. The decomposition of (2) is exact and its terms have unambiguous mathematical meaning; the step from those terms to named clinical phenomena is an inference we have not tested. Equating a low luminance term with osteolysis, or a low contrast term with cortical destruction, would require radiographs annotated per sign, so that one could check whether the component drops where the radiologist marks the sign and stays flat where they do not. That study is feasible with modest annotation effort and is, in our view, the single most valuable next step: it is what would convert a plausible mapping into a validated one, and it is also what would let the method offer per-sign output that a radiologist could audit. Until then, the honest description of the contribution is a compact, physically meaningful feature set with a hypothesis attached.

What remains genuinely favourable is computational cost. The pipeline runs on CPU-only commodity hardware, which matters for research groups and services without accelerators. This is a statement about resources, not about clinical readiness, and it does not offset the absence of external validation, reader comparison, and multiclass evaluation discussed next.

VII. Limitations

We group the limitations by the kind of claim each one restricts.

Task definition. The experiments cover only osteosarcoma versus radiographically normal bone. Benign tumors (1525 images in the source dataset) and non-osteosarcoma malignancies (45 images) were excluded, and these are precisely the categories that generate diagnostic difficulty in practice. The reported separability is therefore an upper bound with respect to the clinically relevant multiclass problem, and it is plausible that descriptors which distinguish osteosarcoma from normal bone perform substantially worse against an aggressive benign lesion or osteomyelitis.

Localization dependence. Lesion ROIs for positives derive from the reference segmentation. No detection stage exists, and the complete diagnostic pipeline was consequently not evaluated.

Reference-region validity. The mirrored intra-image reference assumes approximate symmetry that single-limb radiographs, rotation, and positioning variation frequently violate (Section III-E); large lesions may partly overlap their own reference. No per-image validity check was performed.

Data and partitioning. A single public dataset was used, with no external cohort from independent institutions and no evaluation across distinct acquisition protocols or patient populations. Partitions are at the image level because patient identifiers are unavailable, so same-patient leakage across folds cannot be excluded, and the cross-validation was run once rather than repeated with multiple seeds. With 297 positives, all estimates carry appreciable variance, as the intervals in Table IV show.

Modelling and evaluation. Predicted probabilities were not calibrated or assessed. The complementary descriptor block was not evaluated in isolation, so complementarity with the UIQI is not established. No comparison was made against explainable deep learning or vision transformers under a matched protocol, leaving the interpretability–performance trade-off uncharacterized. Feature importances are impurity-based and affected by correlation among descriptors.

Clinical validation. There is no reader study: the descriptors were not compared with radiologists' assessments, nor were they evaluated as decision support in any workflow. The hypothesized mapping of Table II has no sign-level ground truth behind it. Recall of about 0.68 in the best configuration would in any case be insufficient for autonomous use.

VIII. Future Work

The limitations above define a concrete agenda. (1) Sign-level validation: annotate a subset of radiographs for osteolysis, periosteal reaction, and cortical breach, and test whether each UIQI component responds selectively to its hypothesized sign. (2) Automatic candidate-region detection applied identically to both classes, enabling end-to-end evaluation and removing the localization conditioning. (3) Multiclass extension to benign and other malignant tumors, so that performance is measured against the real differential diagnosis. (4) External multi-center validation on cohorts from independent institutions, with repeated patient-level partitioning, calibration assessment, and confidence intervals derived from that setting. (5) Matched baselines: the texture and gradient block alone, and explainable convolutional and transformer models trained under the identical protocol, to characterize the interpretability–performance trade-off. (6) Prospective clinical evaluation with radiologist comparison, including reader-with-model studies. (7) Hybrid models using UIQI maps as interpretable priors or auxiliary channels.

IX. Conclusion

We presented an interpretable, low-cost digital-image-processing feature set for osteosarcoma radiographs, built

around the Universal Image Quality Index used as a structural-deviation descriptor between a suspect region and a mirrored intra-image reference. The decomposition into correlation, luminance, and contrast provides descriptors whose mathematical meaning is explicit, and we propose—as a hypothesis requiring sign-level annotation to test—that these components track trabecular and periosteal disorganization, density change, and cortical destruction. Under a symmetric ROI extraction that controls an extraction confound whose magnitude we quantify (about 0.10 AUC-ROC and 0.30 F1), the UIQI features separate osteosarcoma from normal bone with an AUC-ROC of 0.872 ± 0.020, and 0.931 ± 0.019 when combined with classical texture and gradient descriptors, under image-level five-fold cross-validation and conditioned on known lesion location. These results characterize a descriptor set and document a confound; they do not describe a triage system. Validation against expert sign-level annotations, an integrated detection stage, multiclass evaluation, and external multi-center testing are required before any clinical claim can be made.

ACKNOWLEDGMENT

The authors thank the creators and curators of the BTXRD dataset and acknowledge the support of the Federal University of Pernambuco (UFPE) and the Cancer Hospital of Pernambuco (HCP).

REFERENCES

[1] Z. Wang and A. C. Bovik, "A universal image quality index," IEEE Signal Process. Lett., vol. 9, no. 3, pp. 81–84, Mar. 2002.
[2] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," Sci. China Inf. Sci., vol. 63, no. 11, art. 211301, 2020.
[3] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging: how-to guide and critical reflection," Insights into Imaging, vol. 11, no. 1, art. 91, 2020.
[4] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: taxonomy and experimental study," Pattern Recognition, vol. 62, pp. 135–160, 2017.
[5] H. C. Beird, S. S. Bielack, A. M. Flanagan, R. Gorlick, et al., "Osteosarcoma," Nature Reviews Disease Primers, vol. 8, art. 77, 2022.
[6] S. Yao et al., "A radiograph dataset for the classification, localization, and segmentation of primary bone tumors," Scientific Data, vol. 11, no. 1, p. 1247, 2024.
[7] F. Shamshad et al., "Transformers in medical imaging: a survey," Medical Image Analysis, vol. 88, p. 102802, 2023.
[8] Y. He, I. Pan, B. Bao, et al., "Deep learning-based classification of primary bone tumors on radiographs," European Radiology, vol. 32, no. 12, pp. 8985–8993, 2022.
[9] E. Claudio, L. Ribeiro, A. Alves, et al., "Classification of bone tumor using convolutional neural network," in IEEE Int. Symp. Computer-Based Medical Systems (CBMS), 2021, pp. 1–6.
[10] "Telehealth initiative to enhance primary care access in Brazil," Journal of Medical Internet Research, vol. 27, no. 1, p. e68434, 2025.
[11] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.
[13] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[14] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," BMC Bioinformatics, vol. 8, art. 25, 2007.
[15] J. Mongan, L. Moy, and C. E. Kahn Jr., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers," Radiology: Artificial Intelligence, vol. 2, no. 2, art. e200029, 2020.

Janelas de Decisão para a Detecção *Online* de Anomalias em EEGs via *One-Class Classification*

Paulo Davi Ramos Albuquerque
Pós-Graduação em Engenharia de Teleinformática (PPGETI)
Universidade Federal do Ceará (UFC), Campus do Pici
Fortaleza, Brasil
paulodralbuquerque@alu.ufc.br

Guilherme de Alencar Barreto
Pós-Graduação em Engenharia de Teleinformática (PPGETI)
Universidade Federal do Ceará (UFC), Campus do Pici
Fortaleza, Brasil
gbarreto@ufc.br

***Abstract*—Neste artigo, são propostas e avaliadas estratégias de decisão baseadas em janelas acumulativas e sua influência no desempenho de modelos de aprendizado de máquina na tarefa de detecção *online* de crises convulsivas. Para esse fim, são testados modelos autocodificadores (*autoencoders*), além de *One-Class SVM* e *Isolation Forest*. Para tanto, foi utilizada a estratégia de treinamento *Open Set* em conjunto com uma abordagem sequencial para emular o cenário *online*. Para a avaliação de desempenho, foram calculadas várias métricas, em especial a sensibilidade e a precisão. Os melhores desempenhos foram obtidos pelos modelos *autoencoders* usando janela acumulativa não causal.**

***Index Terms*—crises epilépticas; eletroencefalograma; detecção *online* de anomalias; janela de decisão.**

I. Introdução

Crises epilépticas são anomalias cerebrais que afetam o sistema nervoso de um indivíduo e fazem com que seus emissores de impulsos elétricos enviem sinais em excesso, provocando convulsões [1]. Esta patologia pode ser diagnosticada através de exames de EEG de escalpo em que eletrodos são colocados no couro cabeludo para capturar e gravar os sinais elétricos do cérebro. No entanto, a análise dos sinais de EEG por um profissional treinado consome muito tempo e pode, eventualmente, induzir a erros de diagnóstico. Por isso, a utilização de técnicas semiautomatizadas com o intuito de auxiliar esses profissionais na correta tomada de decisão é alvo de intensa pesquisa na área de Engenharia Biomédica [2].

Embora seja um tópico que vem sendo pesquisado há anos, percebe-se que grande parte dos trabalhos desenvolvidos visando o uso de técnicas de aprendizado de máquina para detecção de anomalias (*e.g.*, crises epilépticas) ou não abordam ou não dão a devida atenção a questões relacionadas ao uso mais próximo da prática clínica, principalmente o monitoramento em tempo real de pacientes em UTIs, em que um alerta imediato é crucial para a devida conduta médica [3].

Isto posto, neste trabalho investiga-se o desempenho de detectores de anomalias em sinais de EEG quando inseridos em cenários de treinamento sequencial do modelo e o seu subsequente teste em tempo real. Para tal cenário, destaca-se a forte correlação temporal dos dados adquiridos, em virtude da natureza contínua da atividade elétrica cerebral [4]. Tendo isso em mente, serão avaliadas diferentes técnicas de janelamento para tomada de decisão sequencial com o intuito de reduzir a quantidade de alarmes falsos na tarefa de detecção de anomalias em sinais de EEG.

O restante do artigo está organizado da seguinte forma: a Seção II trata dos modelos utilizados na tarefa de detecção de anomalias. A Seção III introduz as técnicas de janelamento de decisões propostas. A metodologia é apresentada na Seção IV, enquanto que os resultados são reportados na Seção V. O artigo é concluído na Seção VI.

II. Modelos de Detecção de Anomalias

Modelos **autocodificadores** são algoritmos de aprendizado utilizados na detecção de anomalias com base no erro de reconstrução dos vetores de atributos através de etapas de codificação e decodificação, conforme ilustrado na Figura 1. Tais detectores são treinados apenas com os vetores de atributos rotulados como negativos (EEG não convulsivo). O erro quadrático de reconstrução correspondente ao k-ésimo vetor de treinamento é dado por $e_k^2 = \|\mathbf{x}_k - \hat{\mathbf{x}}_k\|^2$, $k = 1, \ldots, N$ e $\|.\|$ denota a norma euclidiana do vetor diferença entre o vetor de atributos original $\mathbf{x}_k \in \mathbb{R}^{p\times 1}$ e o vetor reconstruído $\hat{\mathbf{x}}_k \in \mathbb{R}^{p\times 1}$.

Ao final do treinamento, uma etapa adicional usando os próprios dados de treino, mas sem ajuste de parâmetros, faz-se necessária a fim de computar o limiar de decisão L como o percentil $P\%$ dos erros quadráticos de reconstrução obtidos com dados de treino[1]. Sendo A_t a classificação da amostra t, onde os valores -1 e 1 correspondem, respectivamente, a amostras normais e anômalas, a seguinte regra de decisão pode ser usada para fins de detecção de anomalias nos sinais de EEG dos p canais:

$$A_t = \begin{cases} +1, & \text{se} \quad \|\mathbf{x}_{\text{teste}} - \hat{\mathbf{x}}_{\text{teste}}\|^2 > L \\ -1, & \text{caso contrário} \end{cases} \tag{1}$$

ou seja, $\|\mathbf{x}_{\text{teste}} - \hat{\mathbf{x}}_{\text{teste}}\|^2 \leq L$ indica que o vetor de atributos contém trechos não convulsivos dos canais de EEG.

À CAPES (finance code 001), ao CNPq (processo 313251/2023-1) e ao Centro de Referência em Inteligência Artificial - CEREIA, convênios 23067.055633/2021-19 (UFC) / 2020/09706-7 (FAPESP), pelo apoio financeiro.

[1]Isto equivale a encontrar um valor de L para o qual $P\%$ dos valores dos erros de reconstrução no treino estão abaixo desse valor.

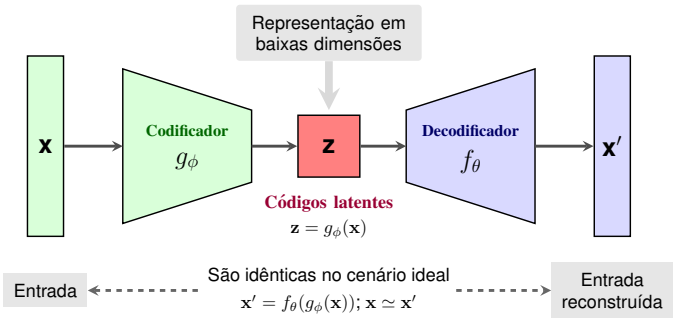


Fig. 1. Arquitetura geral de um modelo autocodificador. Adaptado de [5]. Código vetorial gerado com o auxílio de Inteligência Artificial (Gemini).

Neste artigo, avalia-se um modelo autocodificador linear (*linear autoencoder*, LAE) baseado em Análise de Componentes Principais (PCA) e modelos autocodificadores não lineares (*nonlinear autoencoders*, NLAE) implementados por redes *Multilayer Perceptron* (MLP) com 1 e 3 camadas ocultas.

O modelo ***One-Class Support Vector Machine*** (OCSVM) é uma variação de SVM desenvolvida para lidar com problemas de detecção de anomalias [6]. Seu funcionamento baseia-se na execução, por meio do truque do *kernel*, do produto interno de vetores em um espaço de Hilbert de reprodução [7]. A implementação deste modelo baseou-se na biblioteca `sklearn.svm`[2].

O modelo ***Isolation Forest*** (IF) baseia-se na premissa de que anomalias são observações raras e incomuns no conjunto de dados [8]. Tipicamente, conjuntos de dados de problemas de detecção de anomalias são bastante desbalanceados, contendo uma quantidade significativamente maior de instâncias negativas em relação às positivas (ou anômalas) [9]. A detecção de anomalias via IF é realizada por meio da construção de uma floresta de árvores de decisão onde subconjuntos destes dados são amostrados e particionados aleatoriamente. Instâncias que são isoladas com caminhos curtos nas árvores são mais propensas a serem classificadas como anômalas, enquanto observações negativas exigem mais partições para serem separadas. A aplicação do modelo IF nesta pesquisa fez uso da biblioteca `sklearn.ensemble`[3].

Por fim, também se fez uso de um ***ensemble*** **de modelos detectores** construído com os modelos anteriormente descritos. O comitê utilizou a estratégia de voto majoritário, levando em consideração a classificação realizada por cada modelo de maneira independente.

III. Estratégias de Janelamento da Decisão Propostas

Um dos grandes desafios na utilização de sinais de EEG para a tarefa de detecção de anomalias é a elevada quantidade de disparos errôneos em decorrência da natureza não estacionária e dos artefatos do exame [10]. Por este motivo, neste artigo são propostas e comparadas estratégias de decisão acumulativas dentro de uma janela de tempo a fim de diminuir a ocorrência de falsos positivos.

Janela Acumulativa Causal (JAC): Inspirado por [10], esta estratégia consiste em um pós-processamento baseado na contagem de decisões anômalas ocorrendo dentro de uma janela de comprimento K, mitigando a quantidade de alarmes falsos decorrentes de uma detecção incorreta isolada. Em vez de acionar a detecção de anomalia em um intervalo Δ_t, contam-se as anomalias ocorridas dentro de uma janela de tamanho $K \cdot \Delta_t$. A anomalia só é detectada quando a contagem de anomalias ao final da janela $K \cdot \Delta_t$ for igual a K. Matematicamente, esta regra de decisão é implementada como

$$\text{JAC}_t = \begin{cases} +1, & \text{se } \sum_{j=0}^{K-1} A_{t-j} = K, \\ -1, & \text{caso contrário.} \end{cases} \tag{2}$$

em que A_t está definida na Eq. (1).

Janela Acumulativa Não Causal (JANC): Também foi explorada a implantação de uma estratégia não causal inspirada em [11], porém adaptada para uso no pós-treino do modelo e com percentis. Seja $W = \{A_{t-K}, \ldots, A_{t+K}\}$ a janela móvel de decisões centralizada em A_t, assim o janelamento da decisão é implementado como $\text{JANC}_t = P_q(W)$ tal que $P_q(W)$ representa o percentil q dos valores da janela W. Logo, uma amostra será considerada anômala após a janela de decisão se $(100 - q)\%$ das amostras contidas na janela de comprimento $2K + 1$ também forem desse tipo.

Para viabilizar o uso da JANC em uma aplicação *online*, será empregado um *buffer* de comprimento $2K + 1$, de modo que JANC_t será calculado apenas após a classificação de A_{t+K}.[4] Ademais, serão realizados preenchimentos (*paddings*) de tamanho K e valor 0 no início e no fim do sinal de amostragem.

IV. Materiais e Métodos

O **conjunto de dados** utilizado neste trabalho pertence ao projeto *CHB-MIT Scalp EEG Dataset* [12] e consiste em gravações contínuas de EEG de escalpo, realizadas em 23 pacientes pediátricos no Hospital Infantil de Boston, Massachusetts, EUA, após a suspensão da medicação anticonvulsiva. Entre os pacientes, 6 são do sexo masculino, com idades variando de 3 a 22 anos, e 17 são do sexo feminino, com idades entre 1,5 e 19 anos. O conjunto de dados completo e descompactado corresponde a um total de 42,6 GB de dados. Todos os sinais foram gravados a uma taxa de amostragem de 256Hz, com 16 *bits* de resolução. Os arquivos contêm sinais oriundos de 23 canais de EEG, que foram gravados utilizando o sistema de posicionamento internacional de eletrodos 10-20. A Figura 2 mostra um exemplo de sinal de EEG capturado pelo canal 1 do paciente chb01, destacando os períodos antes (pré-ictal), durante (ictal) e após (pós-ictal) uma crise convulsiva.

Pré-Processamento e Extração de Atributos: Para este trabalho, visando viabilizar o tempo de processamento com-

[2]https://scikit-learn.org/stable/api/sklearn.svm.html

[3]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

[4]Gerando um atraso de K amostras para o início da classificação final.

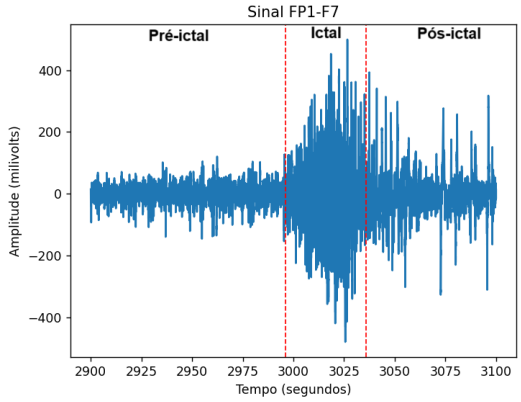


Fig. 2. Sinal extraído do canal 1 do paciente chb01.

putacional nesta etapa exploratória, foram utilizados os arquivos referentes às medições de EEG de um subconjunto de 6 pacientes (chb01, chb03, chb05, chb07, chb14 e chb18). Os arquivos destes pacientes passaram pela etapa de extração de atributos e foram criados novos arquivos com os respectivos rótulos. Os arquivos coletados possuem dimensões $p \times N$, onde p é o número de canais e N o número de amostras contidas no arquivo .edf. Então, são removidos todos os canais que possuam desvio padrão nulo, tendo em vista que estes possuem dados constantes que não auxiliam na detecção das convulsões. Após isso, os sinais são separados em épocas de duração $\Delta_t = 2$s, intervalo de tempo considerado mínimo para determinar se o paciente está ou não tendo uma crise convulsiva [13]. Consequentemente, a uma taxa de amostragem de 256Hz, em $\Delta_t = 2$s, são processadas 512 amostras.

Para cada segmento de duração $\Delta_t = 2$s, os atributos correspondentes a cada canal são extraídos a partir da estimação da densidade espectral de potência (*power spectral density*, PSD) média em 4 bandas de frequência, a saber: faixa Delta ($0,5$Hz $-$ 4Hz), faixa Teta (4Hz $-$ 8Hz), faixa Alfa (8Hz $-$ 12Hz) e faixa Beta (13Hz $-$ 30Hz). A PSD dessas faixas é estimada para cada canal, sendo todo o conjunto de $4p$ atributos agrupado em um único vetor [14]. O vetor gerado é, então, rotulado como positivo (anômalo), recebendo o rótulo numérico $y_t = +1$, se qualquer um dos p canais apresentar pelo menos 1 s de trechos convulsivos no atual intervalo $\Delta_t = 2$s; caso contrário, recebe rótulo negativo ($y_t = -1$). Os segmentos de duração $\Delta_t = 2$s não possuem superposição. Assim, para arquivos de 1 h de duração, são gerados 1800 ($3600/\Delta_t$) vetores de atributos, um por segmento de duração $\Delta_t = 2$s. Os atributos são, enfim, normalizados via *z-score*.

Configuração dos Modelos: Para a parametrização dos classificadores, o modelo autocodificador linear (LAE) foi implementado via PCA com preservação de 99% da variância dos dados originais. Para a definição do limiar de anomalia do modelo LAE, foi escolhido (após experimentação) o percentil 98% dos erros quadráticos de reconstrução obtidos com os dados de treino.

Nas arquiteturas NLAE, o número de neurônios da primeira camada oculta é definido como metade da dimensão do vetor de atributos (arredondado para cima), e as camadas subsequentes reduzem o tamanho da camada anterior pela metade até a camada central. Para as camadas decodificadoras, o tamanho da camada seguinte será o dobro do tamanho da camada atual, resultando num arranjo simétrico do modelo NLAE. O treinamento dos modelos utilizou a função de ativação sigmoide logística e o otimizador Adam com taxa de aprendizado adaptativa (*adaptive learning rate*). Para garantir a convergência e a reprodutibilidade, definiu-se um tamanho de lote (*batch size*) de 64 amostras, máximo de 1000 épocas de treinamento, tolerância de 10^{-5} e semente aleatória (*seed*) fixada em 42. Como técnica de regularização para mitigar o sobreajuste (*overfitting*), empregou-se a parada antecipada (*early stopping*). Assim como no caso do modelo LAE, foi adotado o percentil 98% para determinação do limiar de anomalia.

O modelo OCSVM foi configurado com *kernel* sigmoide e hiperparâmetros $\gamma = 0,01$ e $\nu = 0,1$. O modelo IF empregou os hiperparâmetros `contamination=0,01`, `max_samples=512` e `n_estimators=10`. Para ambos (OCSVM e IF), adotou-se uma adaptação na determinação do limiar: após o treino do modelo, calculou-se o percentil 98% do negativo das distâncias das instâncias ao hiperplano de separação (obtidas através da função *decision_function*), sendo este valor utilizado como limiar de anomalia durante a fase de testes (já que, originalmente, distâncias positivas representam instâncias negativas).

A **metodologia de treino/teste** dos modelos envolve a realização dos experimentos de maneira totalmente sequencial, emulando uma aplicação de uso dos detectores de anomalias em tempo real. Para tal, foi considerada a ordem temporal em que os sinais de EEG de cada paciente estão sendo coletados[5], sendo os modelos treinados com os primeiros N_{files} arquivos sequenciais de um dado paciente e testados com os seus arquivos subsequentes. Cabe ressaltar que a variável N_{files} diz respeito estritamente à quantidade de arquivos de registro (sessões contínuas de EEG) utilizados para o treinamento inicial, e não ao número de pacientes. O valor adequado de N_{files} foi alvo de experimentação e os resultados obtidos serão mostrados na Seção V. Por fim, os modelos são treinados com as primeiras amostras coletadas pelo equipamento e podem ser utilizados para detectar anomalias presentes nos demais sinais em tempo real. Além disso, durante o treinamento fez-se uso da estratégia *Open Set* conforme descrito por [15]. De acordo com esta estratégia, assume-se que todas as primeiras leituras do EEG sejam consideradas negativas, mesmo que algumas (poucas) instâncias positivas estejam presentes durante a fase de treinamento dos modelos detectores.

No tocante às **métricas de desempenho** utilizadas, foram destacadas a sensibilidade (*recall*, R), a precisão (P), a taxa

[5]Consideramos que o paciente está em ambiente de UTI e a coleta está sendo feita no leito onde o paciente se encontra.

TABLE I
RESULTADOS DOS TESTES INICIAIS COM O MODELO LAE.

Modelo	Recall (%)	Precisão (%)	F2-Score	$\mathbf{FAR_{24h}}$
LAE ($N_{\text{files}} = 1$)	98,84	1,36	6,46	7.278,91
LAE ($N_{\text{files}} = 2$)	96,51	2,34	10,67	4.099,01
LAE ($N_{\text{files}} = 3$)	91,45	3,27	14,32	2.431,24
LAE ($N_{\text{files}} = 4$)	90,58	3,48	15,06	2.054,83

de alarmes falsos (FAR) [12] e o F2-Score[6], uma vez que, em conjunto, essas métricas refletem o *trade-off* entre o desempenho do modelo e a frequência de detecções errôneas. Uma atenção especial deve ser atribuída à sensibilidade, uma vez que uma sensibilidade baixa indica que o modelo não está reconhecendo bem amostras positivas (anômalas), o que não é desejável em uma aplicação da área médica. As métricas FAR e F2-Score são definidas como

$$\text{FAR}_{24\text{h}} = \frac{FP \times 86400}{T} \quad \text{e} \quad \text{F2-Score} = \frac{5 \times P \times R}{4R + P}, \quad (3)$$

onde FP é o número de falsos positivos em T segundos.

V. RESULTADOS E DISCUSSÃO

Inicialmente, foram realizados testes com o modelo LAE e dados do paciente CHB01 a fim de investigar a influência do valor de N_{files}[7]. Os resultados desta investigação preliminar estão registrados na Tabela I. Os resultados mostram que o desempenho do modelo melhora à medida que se aumenta o número de arquivos utilizados no treino, o que é relativamente esperado. O importante é que este valor seja o menor possível; logo, não faz sentido avaliar valores de N_{files} maiores que 4. Assim, considerando que o ganho de desempenho ao se aumentar a quantidade de arquivos de 3 para 4 não foi tão grande quanto nos aumentos anteriores e tendo ainda em vista a importância da sensibilidade para o problema em questão, decidiu-se que $N_{\text{files}} = 4$ é o valor adequado para o treinamento dos detectores.

Prosseguiu-se, então, para o refinamento dos hiperparâmetros das técnicas de decisão acumulativas propostas. Para tal, fez-se uso novamente do modelo LAE com dados do paciente CHB01 para definir os tamanhos das janelas e percentis, avaliando-se os valores de $K \in [1, 10]$ para a JAC e as combinações de $K \in [1, 5]$ com $q \in \{25, 50, 75\}$ para a JANC. Os hiperparâmetros ótimos encontrados com o LAE foram então replicados para os demais modelos testados. Reconhece-se que a adoção dos parâmetros otimizados de um modelo linear (LAE) para as demais técnicas e considerando os dados de apenas um paciente constitui uma limitação metodológica do presente estudo, adotada com o intuito prático de reduzir o elevado custo computacional de uma busca exaustiva de hiperparâmetros para cada modelo individualmente. A Tabela II mostra alguns destaques dos resultados obtidos, onde JAC$-K$ representa a janela de decisão causal de tamanho K e JANC$-K/q$ denota a janela não causal de tamanho K e percentil q.

[6]O F2-Score foi escolhido por dar um peso maior à sensibilidade do que o F1-Score.

[7]Número de arquivos para treinamento do detector.

TABLE II
RESULTADOS DO REFINAMENTO DAS TÉCNICAS PROPOSTAS.

Técnica	Recall (%)	Precisão (%)	F2-Score	$\mathbf{FAR_{24h}}$
JAC-2	83,33	11,40	36,84	529,10
JAC-6	60,87	53,16	59,15	43,80
JANC$-3/25$	81,16	40,29	67,47	98,24
JANC$-5/25$	76,09	50,72	69,17	60,37
JANC$-5/50$	92,75	15,90	47,16	400,67
JANC$-5/75$	98,55	6,14	24,58	1.229,82

Os resultados mostram o potencial das decisões acumulativas sobre intervalos maiores em melhorar a precisão do modelo. Dentre os testes com a técnica JAC causal, destacaram-se a JAC-6, com o melhor F2-Score da categoria, e a JAC-2, sendo esta escolhida por ter proporcionado uma melhora considerável na precisão[8] mantendo um valor elevado para o *recall*. Já dentre as técnicas não causais, destacaram-se as janelas com $q = 25$ devido à melhora substancial na precisão, sendo escolhida a JANC$-3/25$ por ter obtido uma melhora considerável na precisão mantendo um *recall* elevado.

Após a definição dos hiperparâmetros, foi realizada uma comparação entre os modelos de detecção avaliados, sem e com a utilização das técnicas de decisão acumulativas propostas. As métricas finais desses experimentos encontram-se nas Tabelas III a V.

Os resultados indicam que modelos de detecção de anomalias que usam janelas de decisão acumulativas sequenciais obtêm bons resultados, sugerindo que seu uso em situações de detecção *online* pode ser viável. Dentre os modelos avaliados, o detector NLAE com 3 camadas ocultas foi o que apresentou um melhor desempenho, uma vez que obteve os maiores valores de *recall* enquanto consistentemente alcançava valores elevados de F2-Score. Além disso, os achados permitem inferir que a aplicação das estratégias propostas impactou positivamente na precisão dos modelos. Paralelamente, destaca-se uma diminuição significativa na quantidade de falsos positivos, alcançando até 96,7% de redução na $\text{FAR}_{24\text{h}}$ no paciente CHB01.

As métricas também demonstram que a tendência decorrente da adição dos filtros (melhora na precisão acompanhada de ligeira piora no *recall*) se mantém mesmo em outros pacientes. Em especial, destaca-se que os baixos valores observados nos resultados do paciente CHB14 sugerem que os modelos possuem mais dificuldades em detectar crises epiléticas nesse paciente, seja por conta de problemas na coleta ou devido a outras condições cerebrais adversas.

Dentre os pós-processamentos, a JANC destaca-se por proporcionar o ganho mais expressivo na precisão dos modelos, alcançando F2-Score de 80,67, precisão de 61,11% e $\text{FAR}_{24\text{h}}$ de 50,56. Contudo, o aumento na precisão impacta negativamente no *recall* do modelo, métrica essencial na área de saúde. Desse modo, ressalta-se a possibilidade de escolher e calibrar o modelo de acordo com as necessidades da aplicação. Do ponto de vista clínico, é imperativo discutir o impacto da latência introduzida pela técnica JANC. Como a estratégia

[8]A diferença de 8% na precisão representa uma redução de 2500 FPs.

TABLE III
RESULTADOS DA DETECÇÃO SEM JANELAS DE DECISÃO ACUMULATIVAS.

CHB01

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
PCA	90,58	3,48	15,06	3472,00
MLP-1camadas	90,58	4,36	18,28	2742,00
MLP-3camadas	96,38	5,35	21,89	2353,00
OC-SVM	91,30	5,29	21,47	2256,00
IF	82,61	7,39	27,21	1429,00
Ensemble	93,48	5,97	23,78	2031,00

CHB03

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
PCA	93,98	1,99	9,17	3842,00
MLP-1camadas	90,36	2,74	12,21	2663,00
MLP-3camadas	89,16	3,72	15,93	1917,00
OC-SVM	49,40	1,57	6,96	2571,00
IF	72,29	6,40	23,62	878,00
Ensemble	87,95	3,76	16,06	1868,00

CHB05

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
PCA	73,94	2,38	10,54	5006,00
MLP-1camadas	66,67	3,83	15,59	2759,00
MLP-3camadas	66,67	4,10	16,45	2573,00
OC-SVM	60,61	5,10	19,07	1862,00
IF	65,45	6,57	23,44	1536,00
Ensemble	66,67	4,65	18,18	2256,00

CHB07

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
PCA	81,10	3,58	15,20	3585,00
MLP-1camadas	77,44	4,03	16,67	3026,00
MLP-3camadas	78,05	4,38	17,89	2793,00
OC-SVM	65,24	5,38	20,23	1882,00
IF	69,51	6,18	22,80	1730,00
Ensemble	76,83	4,57	18,47	2629,00

CHB14

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
PCA	19,67	0,42	1,92	2870,00
MLP-1camadas	26,23	0,91	4,01	1737,00
MLP-3camadas	27,87	1,09	4,70	1548,00
OC-SVM	3,28	0,06	0,26	3607,00
IF	3,28	0,16	0,67	1248,00
Ensemble	16,39	0,74	3,13	1345,00

CHB18

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
PCA	69,94	2,27	10,06	4902,00
MLP-1camadas	67,48	2,55	11,09	4196,00
MLP-3camadas	58,28	3,42	13,85	2683,00
OC-SVM	34,97	2,84	10,71	1952,00
IF	28,83	4,73	14,29	946,00
Ensemble	53,37	3,58	14,11	2345,00

TABLE IV
RESULTADOS DA DETECÇÃO COM ESTRATÉGIA JAC.

CHB01

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	83,33	11,40	36,84	586,99
NLAE-1oculta	84,06	14,48	42,87	449,76
NLAE-3ocultas	92,75	19,08	52,33	356,53
OC-SVM	89,86	25,67	59,90	235,71
IF	71,74	21,81	49,20	233,09
Ensemble	87,68	19,42	51,49	329,61

CHB03

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	90,36	5,92	23,47	840,66
NLAE-1oculta	85,54	8,14	29,49	565,38
NLAE-3ocultas	81,93	10,95	35,68	390,33
OC-SVM	33,73	13,86	26,22	122,82
IF	57,83	18,46	40,54	149,64
Ensemble	81,93	11,28	36,36	377,63

CHB05

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	65,45	5,53	20,67	1264,46
NLAE-1oculta	61,21	9,08	28,50	693,26
NLAE-3ocultas	61,21	9,38	29,07	669,26
OC-SVM	57,58	17,18	39,16	314,06
IF	61,21	14,77	37,57	399,77
Ensemble	61,21	10,40	30,96	596,57

CHB07

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	76,83	9,07	30,81	593,78
NLAE-1oculta	75,00	10,27	33,17	505,39
NLAE-3ocultas	75,00	11,44	35,53	447,57
OC-SVM	68,29	18,15	43,99	237,42
IF	65,85	17,14	41,99	245,41
Ensemble	73,78	11,80	35,99	425,00

CHB14

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	4,92	0,41	1,52	804,00
NLAE-1oculta	13,11	1,74	5,67	494,18
NLAE-3ocultas	14,75	2,36	7,20	405,82
OC-SVM	0,00	0,00	0,00	146,18
IF	0,00	0,00	0,00	303,27
Ensemble	4,92	0,91	2,62	355,64

CHB18

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	61,96	5,79	21,09	1286,51
NLAE-1oculta	60,74	4,56	17,55	1621,85
NLAE-3ocultas	51,53	7,60	23,90	799,96
OC-SVM	30,06	3,83	12,69	962,93
IF	20,25	12,04	17,82	188,82
Ensemble	46,01	8,40	24,27	640,91

necessita de um atraso de K épocas para fechar a janela de decisão centralizada, e sabendo que $\Delta_t = 2$s, um atraso de $K = 3$ resulta em uma latência adicional de 6s para a emissão do alerta. Em ambientes de UTI, um retardo de poucos segundos na detecção automática de uma crise é considerado amplamente aceitável frente ao ganho substancial na redução de falsos alarmes, fator crucial para mitigar o fenômeno de fadiga de alarme (*alarm fatigue*) na equipe médica.

Nessa perspectiva, foram esboçadas curvas *Precision-Recall* [16] para ilustrar o *trade-off* entre tais métricas e analisar os principais pontos de operação dos modelos utilizados. A Figura 3 mostra algumas curvas de destaque nos resultados obtidos para o paciente CHB01. A Figura 3(a) mostra a curva para o modelo NLAE com 3 camadas ocultas, esboçando uma queda brusca na precisão do modelo após uma pequena redução no limiar de anomalia, de modo que a precisão só atinge valores consideráveis quando o *recall* apresenta as menores taxas. Já as Figuras 3(b) e 3(c) revelam que a adição do janelamento acumulativo das decisões torna essas curvas bem mais expressivas, de maneira que se torna viável uma análise de otimização sobre a escolha do limiar de anomalia a fim de encontrar um equilíbrio desejável entre as métricas. Vale salientar que o ponto destacado nas imagens se refere ao limiar sob o qual o modelo apresenta o melhor valor de F2-Score, podendo ser escolhidos diferentes pontos de operação dependendo do contexto de aplicação.

Na sequência, a Figura 4 mostra uma comparação entre as curvas de outros modelos utilizados sobre os dados do paciente CHB01, ressaltando o melhor desempenho do modelo NLAE-3.

TABLE V
RESULTADOS DA DETECÇÃO COM ESTRATÉGIA JANC.

CHB01

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	81,16	40,29	67,47	108,99
NLAE-1oculta	76,81	48,40	68,74	74,19
NLAE-3ocultas	87,68	61,11	80,67	50,56
OC-SVM	81,88	43,30	69,50	97,17
IF	65,22	38,79	57,40	93,24
Ensemble	83,33	54,76	75,46	62,38

CHB03

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	86,75	19,10	50,78	215,28
NLAE-1oculta	83,13	25,75	57,50	140,46
NLAE-3ocultas	79,52	35,68	63,83	84,00
OC-SVM	20,48	48,57	23,16	12,71
IF	43,37	70,59	47,00	10,59
Ensemble	79,52	36,07	64,08	82,58

CHB05

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	56,97	13,51	34,66	412,80
NLAE-1oculta	56,36	20,71	41,93	244,11
NLAE-3ocultas	56,36	20,67	41,89	244,80
OC-SVM	52,12	31,05	45,89	130,97
IF	52,73	26,52	44,03	165,26
Ensemble	55,15	22,14	42,48	219,43

CHB07

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	71,34	25,66	52,61	159,37
NLAE-1oculta	68,29	28,35	53,28	133,05
NLAE-3ocultas	68,29	32,28	55,83	110,48
OC-SVM	63,41	47,06	59,29	55,01
IF	61,59	49,51	58,72	48,42
Ensemble	68,29	33,73	56,68	103,43

CHB14

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	0,00	0,00	0,00	109,09
NLAE-1oculta	0,00	0,00	0,00	60,00
NLAE-3ocultas	0,00	0,00	0,00	29,45
OC-SVM	0,00	0,00	0,00	5,45
IF	0,00	0,00	0,00	18,55
Ensemble	0,00	0,00	0,00	28,36

CHB18

Modelo	Recall	Precisão	F2-Score	$\mathbf{FAR_{24h}}$
LAE	53,99	16,79	37,41	341,61
NLAE-1oculta	58,90	6,45	22,42	1091,42
NLAE-3ocultas	47,85	14,39	32,66	363,55
OC-SVM	22,70	3,72	11,23	750,60
IF	14,11	26,14	15,54	50,93
Ensemble	40,49	16,42	31,31	263,26

Por fim, a Figura 5 esboça uma comparação entre curvas obtidas com o mesmo modelo e filtro para diferentes pacientes, evidenciando um comportamento similar mesmo entre sujeitos distintos.

VI. CONCLUSÕES

Neste trabalho, foram propostas e avaliadas duas estratégias de decisão baseadas em janelas de decisão acumulativas e sua influência no desempenho de modelos de aprendizado de máquina na tarefa de detecção *online* de crises convulsivas em sinais de EEG. Para este fim, foram avaliados modelos autocodificadores, OCSVM e IF. Foi utilizada também a estratégia de treinamento *Open Set* em conjunção com uma abordagem sequencial para emular o treino e teste *online* dos modelos. Os melhores desempenhos foram verificados para o autocodificador não linear com 3 camadas ocultas usando estratégia cumulativa não causal, sendo proposta a utilização de um *buffer* para viabilizar o uso *online* da técnica proposta. Algumas limitações do trabalho devem ser destacadas, como a não utilização de modelos mais complexos de aprendizado de máquina e a subotimização de hiperparâmetros decorrente do uso de apenas um modelo sobre os dados de um único paciente para sua definição. Desse modo, propõe-se a realização de trabalhos futuros com o intuito de realizar experimentos utilizando uma maior diversidade de modelos, bem como processos de otimização de hiperparâmetros mais complexos, a fim de verificar as mudanças nos resultados obtidos com tais abordagens. Como conclusão geral, pode-se afirmar que os resultados obtidos foram promissores, verificando-se como viável o uso dos modelos autocodificadores avaliados para detecção *online* de anomalias em sinais de EEG.

AGRADECIMENTOS

À CAPES (finance code 001), ao CNPq (processo 313251/2023-1) e ao Centro de Referência em Inteligência Artificial - CEREIA, convênios 23067.055633/2021-19 (UFC) / 2020/09706-7 (FAPESP), pelo apoio financeiro.

REFERENCES

[1] M. K. Alharthi, K. M. Moria, D. M. Alghazzawi e H. O. Tayeb, "Epileptic disorder detection of seizures using EEG signals," *Sensors*, v. 22, pp. 6592, 2022.

[2] B. Mouazen, A. Bendaouia, E. H. Abdelwahed e G. De Marco, "Machine learning and clinical EEG data for multiple sclerosis: A systematic review," *Artificial Intelligence in Medicine*, v. 166, pp. 103116, 2025.

[3] A. S. Poncette, C. Spies, L. Mosch, M. Schieler, S. Weber-Carstens, H. Krampe e F. Balzer, "Clinical Requirements of Future Patient Monitoring in the Intensive Care Unit: Qualitative Study," *JMIR Med Inform*, v. 7, no. 2, pp. e13064, 2019.

[4] G. Chakrapani, R. Scholar e D. Deverapalli, "Deep Learning Framework With Integrated Feature Extraction for Detection of Epilepsy and Basal Ganglia Diseases," *IEEE Access*, v. PP, pp. 1-1, 2026.

[5] P. Sarajcev, A. Kunac, G. Petrovic e M. Despalatovic, "Power System Transient Stability Assessment Using Stacked Autoencoder and Voting Ensemble," *Energies*, v. 14, pp. 3148, 2021.

[6] L. G. C. Morais e A. V. da Silva Netto, "Detecção de anomalias na vibração de motores elétricos baseada em OCSVM," em *Simpósio Brasileiro de Automação Inteligente (SBAI'2021)*, v. 1, no. 1, 2021.

[7] J. Pang, X. Pu e L. C. Li, "A hybrid algorithm incorporating vector quantization and one-class support vector machine for industrial anomaly detection," *IEEE Transactions on Industrial Informatics*, v. 18, no. 12, pp. 8786–8796, 2022.

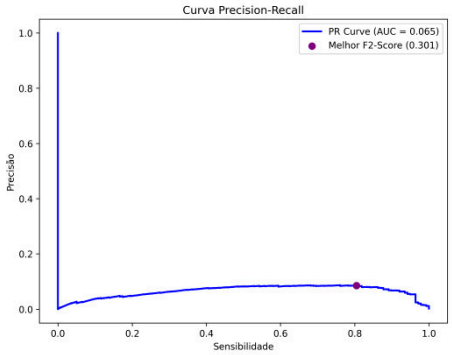
[8] O. Lifandali, N. Abghour e Z. Chiba, "Feature selection using a combination of ant colony optimization and random forest algorithms applied to isolation forest based intrusion detection system," *Procedia Computer Science*, v. 220, pp. 796–805, 2023.

[9] Y. Chabchoub, M. U. Togbe, A. Boly e R. Chiky, "An in-depth study and improvement of isolation forest," *IEEE Access*, v. 10, pp. 10219–10237, 2022.

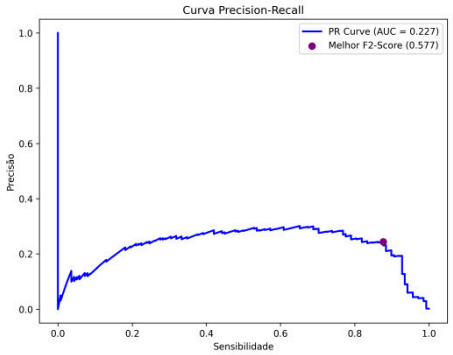
[10] D. Lv, C. Chen, J. Sun, O. W. Samuel e N. Zhang, "EEG-Based Epileptic Seizure Prediction Using Attention-Based Temporal Convolutional Recurrent Neural Networks for Cross-Patient Generalization," *2026 IEEE International Conference on Power, Electronics and Green Energy (ICPEGE)*, pp. 1704-1710, 2026.

[11] S. Mounika e S. R. R., "Improved attention-based PCNN with GhostNet for epilepsy seizure detection using EEG and fMRI modalities: extractive pattern and histogram feature set," *Front. Artif. Intell.*, v. 8, pp. 1679218, 2026.

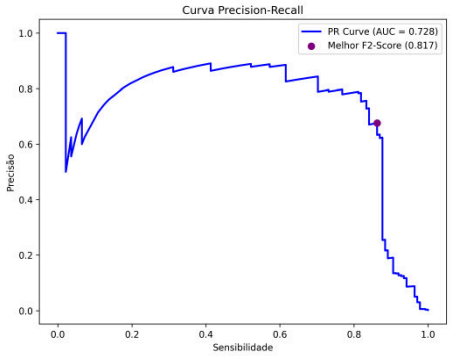
[12] J. Guttag, "CHB-MIT Scalp EEG Database (version 1.0.0)," PhysioNet, 2010. Disponível em: https://doi.org/10.13026/C2K01R.



(a) NLAE3

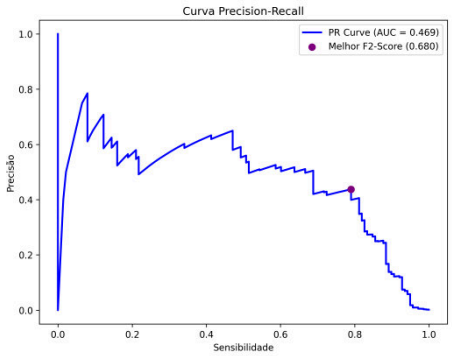


(b) NLAE3/JAC

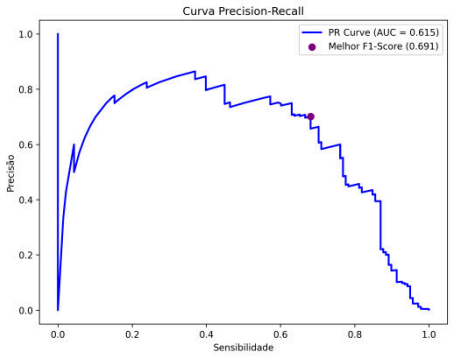


(c) NLAE3/JANC

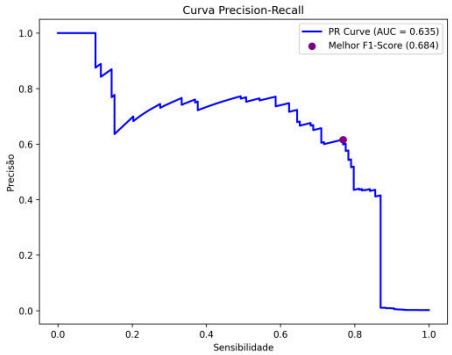
Fig. 3. Curvas *Precision-Recall* do modelo NLAE3.



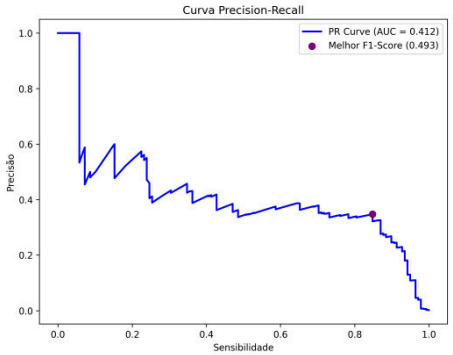
(a) LAE/JANC



(b) NLAE1/JAC

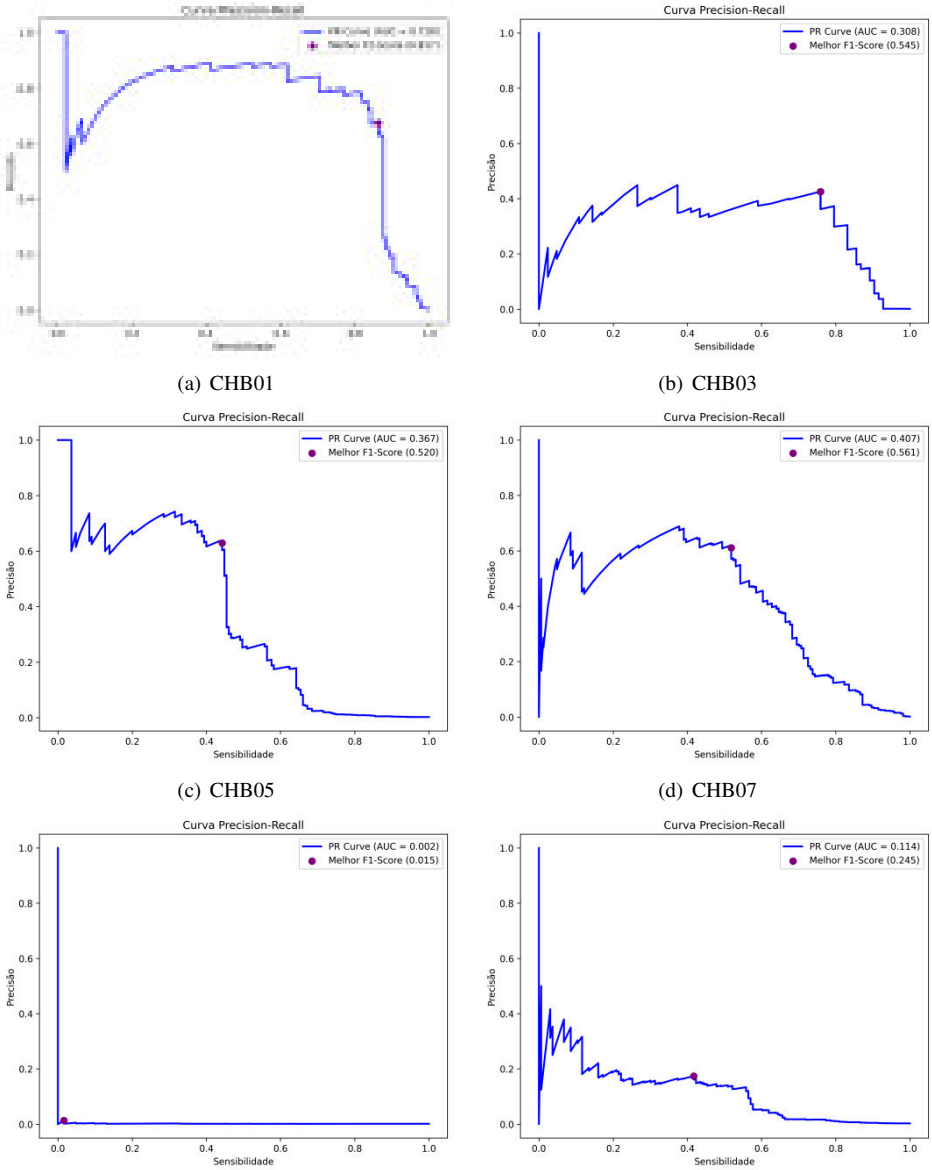


(c) OC-SVM/JANC



(d) IF/JANC

Fig. 4. Curvas *Precision-Recall* dos demais detectores avaliados.



(a) CHB01 (b) CHB03

(c) CHB05 (d) CHB07

(e) CHB14 (f) CHB18

Fig. 5. Curvas *Precision-Recall* do modelo NLAE3/JANC.

[13] A. Shoeb e J. Guttag, "Application of machine learning to epileptic seizure onset detection," em *Proceedings of the 27th International Conference on Machine Learning (ICML'2010)*, Haifa, Israel, pp. 1–8, Junho 2010.

[14] Y. M. Dweiri e T. K. Al-Omary, "Novel ML-based algorithm for detecting seizures from single-channel eeg," *NeuroSci*, v. 5, pp. 59–70, 2024.

[15] A. Brilhador, A. E. Lazzaretti e H. S. Lopes, "A survey on open world learning for image segmentation: Definitions, challenges, and directions," *Learning & Nonlinear Models*, v. 23, no. 1, pp. 33–48, 2025.

[16] T. Saito e M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, v. 10, no. 3, e0118432, 2015.

Korotkoff Sound Filtering Techniques For Exercise Stress Testing And Functional Assessment Protocols

1st Guilherme Leitão
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brazil

2nd Wesley Beccaro
Lab. de Microeletrônica (LME)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brazil

3rd Miguel Arjona Ramírez
Lab. de Processamento de Sinais (LPS)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brazil

4th Renata Valeri de Freitas
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brazil

***Abstract*—The exercise stress test is routinely used for the diagnosis of cardiac conditions, including ischemia and arrhythmias. During the test, blood pressure is commonly measured using the manual auscultatory method, which is susceptible to motion artifacts and background noise. Previous studies have shown that noise can impair Korotkoff sound detection and BP measurement accuracy. However, no database of Korotkoff sounds under treadmill noise is currently available. Therefore, this study proposes a synthetic database of Korotkoff sounds contaminated with treadmill noise and compares different filtering techniques, including neural network based and deterministic methods. Their effectiveness is evaluated based on their ability to recover Korotkoff sounds and restore the performance of an established automatic BP detection algorithm under noisy conditions.**

***Index Terms*—Blood pressure, Korotkoff sounds, noise filtering, automatic methods, exercise stress test.**

I. Introduction

Blood pressure (BP) is one of the main physiological indicators used to assess early cardiovascular risk, and it is crucial for preventing serious complications such as myocardial infarction and stroke [1] , [2]. During stress test, this measure is crucial for monitoring the cardiovascular response to exercise, identifying abnormal BP patterns, and detecting potential signs of ischemia, hypertension, or hemodynamic instability that may not be evident at rest [3] , [4].

Traditionally, BP can be measured using different techniques. The intra-arterial method is considered the gold standard for BP measurement, as it is highly accurate and can provide continuous monitoring. However, because it is an invasive procedure, requiring the insertion of a catheter into an artery, it is impractical in most clinical situations, including during stress tests, and is, therefore, limited to specific hospital settings [1] , [2].

The oscillometric method, widely used in automatic devices, estimates BP from oscillations generated by the brachial artery under cuff compression. However, despite its widespread use under rest conditions, the oscillometric method is generally not recommended for BP measurement during exercise according to major guidelines [5], [6].

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

In contrast, the auscultatory method is considered a clinical reference for BP measurement and can be regarded as a practical gold-standard method in clinical stress tests settings, given the complexity and invasiveness of the intra-arterial approach [1] , [2], [5] . Auscultation is based on the detection of Korotkoff sounds (KS) during cuff deflation by a trained professional, preferably a physician according to medical stress test guidelines [6] .

During gradual cuff deflation, the re-establishment of arterial blood flow produces a characteristic sequence of five Korotkoff phases (I-V): appearance, softening, sharpening, muffling, and disappearance of the sounds. Phase I, marked by the first clear tapping sounds, is conventionally associated with systolic blood pressure (SBP), whereas phase V, defined by complete sound disappearance, is usually associated with diastolic blood pressure (DBP) [1] , [2]. Nevertheless, there is still divergence regarding the definition of DBP, particularly as to whether it should be determined from the muffling of Korotkoff sounds in phase IV or from their complete disappearance in phase V [7], [8], [9].

Hence, despite the importance of the auscultatory method in clinical practice, manual detection of KS is highly vulnerable to operator-dependent factors, even under standard resting clinical conditions. For instance, the auscultatory gap, a transient disappearance of KS during cuff deflation [1], is commonly observed, especially in hypertensive patients, which may lead to underestimation of SBP [10]. In addition, DBP measurement can be particularly challenging due to the attenuation of phase IV KS and the onset of phase V that may be difficult to identify in some patients. In some cases, KS may persist down to very low cuff pressures, or even until 0 mmHg, making the disappearance of blood flow sounds an unreliable marker for DBP estimation [1], [4], [6].

Other studies, have reported that background noise can interfere with the accuracy of BP estimation due to psychoacoustic masking [11]. This challenge is further amplified in exercise stress test, where motion artifacts, footsteps, and treadmill noise can overlap with the frequency and temporal characteristics of KS, making their accurate detection and BP estimation substantially more difficult [3], [4].

In this context, there is growing interest in the development of automatic BP measurement methods that provide greater

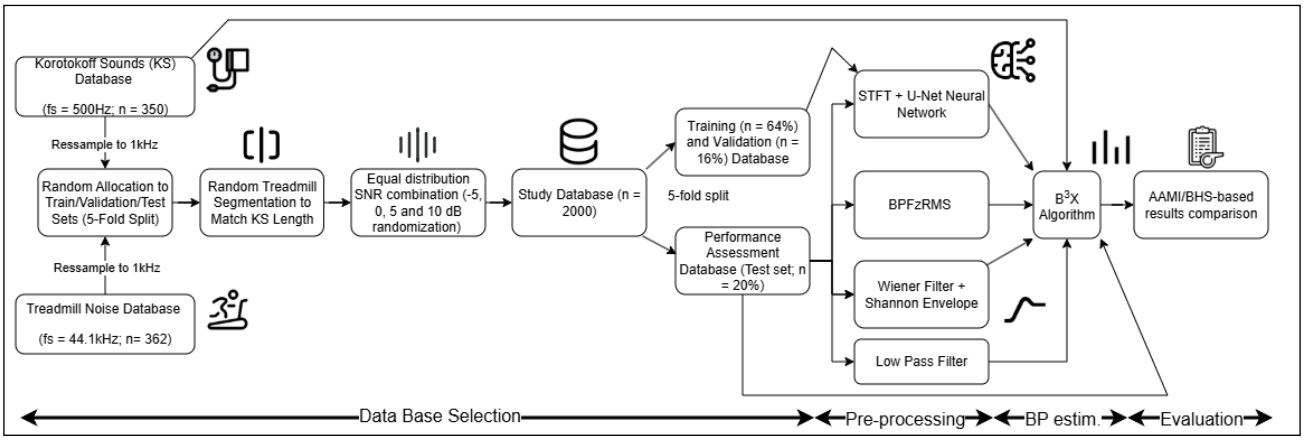


Fig. 1. Methodology overview diagram. The pseudo-algorithm, detailed throughout the Methodology section, summarizes the data selection process for method testing, the preprocessing techniques evaluated in this study, and the assessment of the B³X blood pressure estimation algorithm according to medical standards.

accuracy and clinical safety while facilitating repeated measurements. One example of a promising automatic auscultatory algorithms is the B³X proposed by [12]. Based on the literature review conducted in this work, no studies have evaluated the B³X method under exercise stress test protocols. In this scenario, ensuring robust noise suppression becomes essential not only for B³X but also for any automatic auscultatory BP measurement algorithm [3], [10]. Indeed, one of the main challenges involving auscultatory methods in stress testing lies in separating KS from possible exercise-related noise [4], [13].

II. Objective

The objective of this work is to evaluate different preprocessing filtering techiniques for Korotkoff sounds artificially contaminated with treadmill running noise. The effectiveness of these techniques will be evaluated based on the improvement in the measurement accuracy of both SBP and DBP obtained by B³X algorithm.

III. Methodology

A. Methodological Overview

Fig. 1 presents an overview of the main steps covered throughout this section. The diagram highlights the construction of the study dataset, the preprocessing and corruption of KS signals with treadmill noise, the application of the proposed filtering methods, and the subsequent performance evaluation based on automatic BP estimation and relevant clinical standards.

B. Dataset

Public access databases containing KS are scarce in scientific literature [14]. When considering the context of exercise stress testing, their availability becomes even more limited. Although no databases dedicated to KS recorded during exercise testing were found, it is possible to preliminary combine data from databases containing KS measured at rest with audio recordings of treadmill exercise sounds.

For KS, one of the main reference databases in the literature is the "Auscultatory and Oscillometry BP Data" [15]. This dataset contains 350 records of non-invasive BP measurements, sampled at 500 Hz, collected from human volunteers in a controlled environment, following the recommendations of the ANSI/AAMI SP10 standard. Although participant profiles were not provided in the official database access [15], several studies report the demographic variability of this dataset summarized in Table I , [12], [16], [17], [18]. It should be noted that the 350 recordings were not labeled according to participant identity. Therefore, due to the lack of subject-level information, the recordings were treated as independent and uncorrelated. Each sample consists of audio signals containing KS, the cuff pressure curve throughout the deflation process, expressed in mmHg, the oscillometric signal associated with arterial oscillations, and reference SBP and DBP values. Hereafter, this dataset will be referred to as KS Database.

TABLE I
Demographic and blood pressure characteristics of the KS Database [18].

Parameter	Value
Recordings	350
Participants	155
Male	87 (56.1%)
Female	68 (43.9%)
Age	52.0 ± 21.6 years (23–97)
Age (male)	51.1 ± 20.3 years
Age (female)	53.1 ± 23.4 years
Arm circumference	10–35 cm
Systolic blood pressure	81–191 mmHg
Diastolic blood pressure	37–104 mmHg
Mean arterial pressure	56–118 mmHg

For noise sounds of individuals exercising on treadmills, a recent study [19] consolidated a database containing 40 hours of audio, divided into 363 recordings acquired under different speeds, runners, and gym environments. This dataset was originally recorded with a sample rate of 44.1 kHz. Although this database does not contain treadmill exercise

sounds recorded in clinical settings, its data may serve as an initial approximation for testing noise immunity in the proposed model. Hereafter, this dataset will be referred to as Noise Database or Treadmill Database.

C. 2000 Samples Generation

A database containing 2,000 artificially contaminated KS recordings was generated (this dataset will be referred to as Study Database) with the combination of the two databases previously mentioned in Section III-B. First, all signals were resampled to 1 kHz, since most Korotkoff sound spectral content lies below 250 Hz [11]. Second, treadmill noise recordings were randomly segmented into non-overlapping parts to match the duration of each KS recording. Afterwards, DC offsets were removed from both KS and treadmill noise signals, which were then independently normalized to a common RMS level. The normalized signals were subsequently mixed at controlled SNR levels of -5, 0, 5 and 10 dB and balanced across the generated dataset. Similar amplitude normalization has been applied to phonocardiogram (PCG) signals [20], [21], which share similar acoustic characteristics with KS [7], [22].

For the selection of samples in the Study Database, KS and treadmill recordings, noise segments and SNR levels were randomly selected using independent random generators with fixed seeds to ensure reproducibility. Each KS-treadmill recording pair was used only once to avoid duplication regardless of the SNR or noise segment. However, the same KS recording could be combined with multiple treadmill recordings to increase noise variability and reduce artificial correlations between a singular KS to a singular treadmill sample.

Since one denoising method relies on a neural network baseline (U-net), the study dataset, composed of 2,000 samples, was divided using a group-based 5-fold cross-validation procedure. In each fold, approximately 64% of the samples were assigned to the training subset, 16% to the validation subset, and 20% to the test subset. Because individual KS and treadmill samples were randomly assigned to a specific subset before the database synthesis, no data leakage occurred between subsets. To ensure comparability among all evaluated methods, the remaining methodologies were assessed using the corresponding test subset of each U-Net fold.

D. Noise and Korotkoff signal characterization

To evaluate the ability of the filtering methods to restore the performance of the B³X algorithm under noisy conditions, it is important to characterize both the KS and noise profiles.

An example of a noise segment, a clean KS sample, and a corrupted KS signal with added noise $SNR = 1.0$ dB is shown using a mel-spectrogram representation in Fig. 2. The same samples are also presented in a normalized frequency spectrum in Fig. 2. For bandwidth comparison, all signals were resampled to 1 kHz, since this sampling rate was adopted in the generation of the study dataset.

In the Mel spectrogram, the KS sample exhibits quasi-periodic pulsatile events, with most of its energy concentrated below 50 Hz during silent intervals and transient components extending up to approximately 200 Hz over short time frames. A similar pulsatile pattern is also present in the treadmill noise sample, although its energy is more broadly distributed across the time-frequency domain and presents rapidly varying temporal patterns. As a result, in the noisy KS signal, the treadmill noise masks much of the KS high-frequency content, while the low-frequency pulsatile components remain faintly visible. Indeed, the normalized frequency spectrum clarifies the bandwidth of both KS and noise signals, showing that most of the KS energy is concentrated below 100 Hz, whereas the noise is more prominent between 80 Hz and 450 Hz.

E. Korotoff sounds filtering methods

In this paper, several approaches that were not originally designed to address the noise conditions encountered during exercise stress testing are evaluated according to their ability to preserve the accuracy of automatic BP detection of B³X: Fourth-order 70-Hz Butterworth low-pass filter, Wiener filtering followed by a Shannon envelope [23], band-pass filtering combined with z-score normalization and RMS envelope weighting [18], hereafter referred to as BPFzRMS, and a convolutional U-Net neural network [24].

1) Wiener Filtering and Shannon Envelope: A preprocessing method based on Wiener filtering followed by a Shannon envelope has been proposed to filter KS signals [23]. For Wiener filtering [25], the corrupted signal ($x[n]$) is considered as the sum of the KS ($s[n]$) and a noise component ($w[n]$):

$$x[n] = s[n] + w[n] \tag{1}$$

Assuming that the KS itself is uncorrelated with environmental and patient movement noise, the autocorrelation of the estimated KS signal ($\hat{s}[n]$) is given by:

$$\mathbf{R}_{\hat{s}\hat{s}} = \mathbf{R}_{xx} - \mathbf{R}_{\hat{w}\hat{w}} \tag{2}$$

where $\mathbf{R}_{\hat{w}\hat{w}}$ is the autocorrelation matrix of the estimated noise component, obtained from the first second of the corrupted signal. The coefficients of the Wiener FIR filter, $h[n]$, are obtained by solving the following system of linear equations:

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[M-1] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[M-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[M-1] & r_{xx}[M-2] & \cdots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[M-1] \end{bmatrix} = \begin{bmatrix} r_{\hat{s}\hat{s}}[0] \\ r_{\hat{s}\hat{s}}[1] \\ \vdots \\ r_{\hat{s}\hat{s}}[M-1] \end{bmatrix} \tag{3}$$

Once $h[n]$ is obtained, the estimated KS signal is given by:

$$\hat{s}[n] = h[n] * x[n] \tag{4}$$

where $(*)$ denotes the convolution operator.

The filter length was defined as $M = 0.25\%$ of the Korotkoff signal length, following the procedure adopted in [23]. To avoid excessively long filters and preserve the local temporal characteristics of the KS, M was limited to a maximum duration of 80 ms. In this study, the filter was estimated and applied to the entire signal, without window-based segmentation or temporal overlap [21].

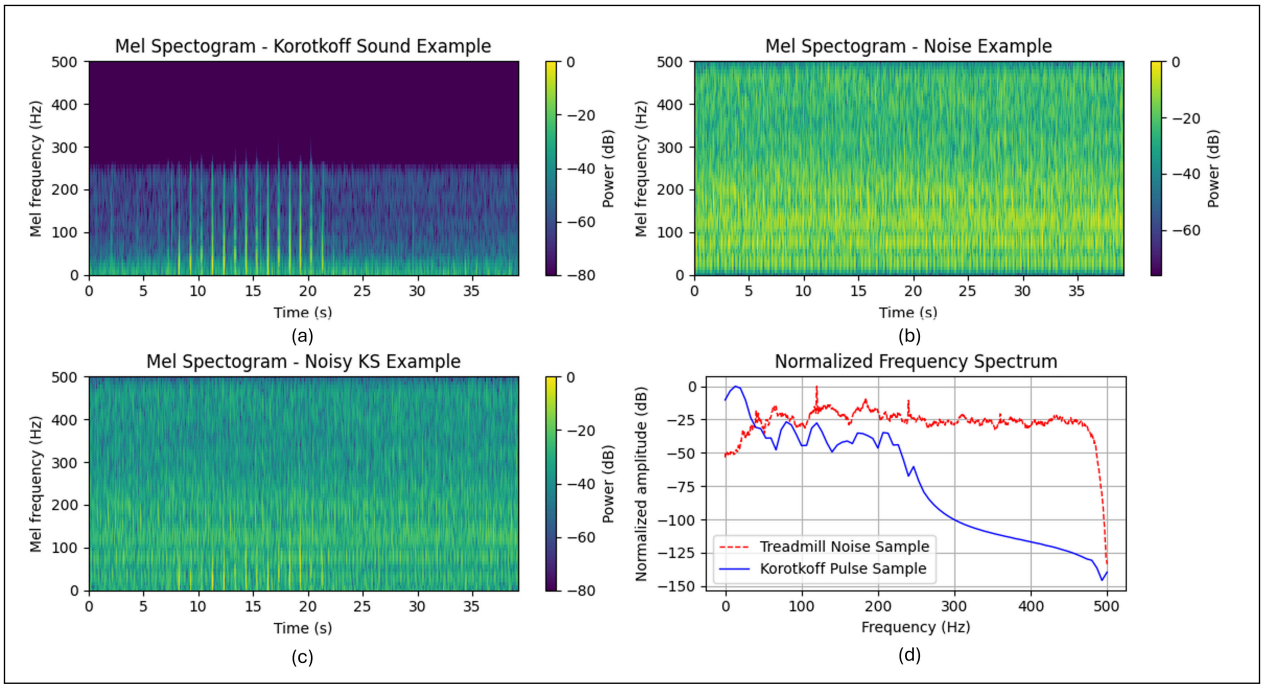


Fig. 2. Korotkoff sound and treadmill noise mel spectrograms. (a) Mel spectrogram of a KS sample (b) Mel spectrogram of a treadmill noise sample (c) Mel spectrogram of the noisy KS signal (d) Normalized frequency spectra of the same KS and noise samples.

For the Shannon envelope, the Korotkoff signal was processed using the Shannon energy envelope, defined as:

$$E[n] = -\frac{1}{N_w}\sum_{i=0}^{N_w-1} z_i^2[n]\log\left(z_i^2[n]\right) \tag{5}$$

where $z_i[n]$ is given by:

$$z_i[n] = \frac{\hat{s}_i[n]}{max|\hat{s}_i[n]|} \tag{6}$$

with the Shannon energy computed using 40 ms windows and 50% overlap. Then, E was normalized using z-score normalization:

$$P[n] = \frac{E[n] - \mu_E}{\sigma_E} \tag{7}$$

Up to this point, the procedure followed the concepts described by [23], [25]. Subsequently, the resulting vector was then used as a mask applied to the original Korotkoff signal:

$$y[n] = (1-\alpha)\hat{s}[n] + \alpha(P[n]\hat{s}[n]) \tag{8}$$

with $\alpha = 0.3$.

2) Band-Pass Filtering Combined with Z-Score Normalization and RMS Envelope Weighting (BPFzRMS): In the present study, isolated high-amplitude artifacts from KS were suppressed. Detected peaks from the absolute KS signal that exceeded twice the amplitude of the second largest peak were zeroed in a ±500 ms window around it. Additionally, boundary peaks located within the first or last 100 ms of the signal were zeroed if they corresponded to the most prominent peak within a 500 ms window. The resulting signal was then filtered using a 20–250 Hz band-pass filter. The method follows with the computation of the RMS energy envelope using a zero-phase moving average filter with a Hamming window [18]:

$$p[n] = x_{\text{filt}}^2[n] \tag{9}$$

$$RMS[n] = \sqrt{\sum_{m=0}^{N_w-1} w[m]p[n-m]} \tag{10}$$

Afterwards, the Korotkoff signals were standardized using z-score normalization, while the RMS envelope was normalized between 0 and 1 in order to preserve regions with higher energy and attenuate regions with lower energy:

$$x_{\text{norm}}[n] = \frac{x_{\text{filt}} - \mu_x}{\sigma_x} \tag{11}$$

$$RMS_{\text{norm}}[n] = \frac{RMS[n] - RMS_{\text{min}}}{RMS_{\text{max}} - RMS_{\text{min}}} \tag{12}$$

Finally, the resulting signal was then weighed by the normalized RMS envelope, so that regions with higher RMS energy were preserved while lower-energy regions were attenuated:

$$g[n] = 0.25 + 0.75RMS_{\text{norm}}[n] \tag{13}$$

$$y[n] = x_{\text{norm}}[n]g[n] \tag{14}$$

3) STFT + U-Net convolutional neural network: Finally, a convolutional U-Net has been proposed for processing phonocardiogram (PCG) signals, i.e., signals that acoustically represent the closure of cardiac valves [24]. Although they are different, KS and PCG signals share similar acoustic

TABLE II
BRITISH HYPERTENSION SOCIETY (BHS) CLASSIFICATION CRITERIA.

Classification	$\lvert Error \rvert$ ≤ 5 mmHg	$\lvert Error \rvert$ ≤ 10 mmHg	$\lvert Error \rvert$ ≤ 15 mmHg
A	$\geq 60\%$	$\geq 85\%$	$\geq 95\%$
B	$\geq 50\%$	$\geq 75\%$	$\geq 90\%$
C	$\geq 40\%$	$\geq 65\%$	$\geq 85\%$
D	$< 40\%$	$< 65\%$	$< 85\%$

characteristics, and this methodology could therefore be tested in this context [7], [22]. This is the only method among those previously described that involves machine learning in the context of noise reduction.

The proposed approach consists of segmenting the signal into 1.5 s frames, computing the short-time Fourier transform (STFT) using an FFT length of 128 and a Hann window with 50% overlap, and separating the transform into real and imaginary channels. The resulting representation is then resized to $(2, 64, 64)$ and used as input to a U-Net for noise reduction [24]. In this study, the proposed architecture was implemented in PyTorch and adapted to process KS signals instead of PCG signals. Unlike the original study, the network was trained for up to 150 epochs with an early-stopping patience of 20 epochs, as well as a learning rate of 0.001.

F. Evaluation Metrics

The Brazilian Guidelines for Exercise Testing do not establish a BP measurement precision and accuracy criteria for automatic devices in clinical exercise testing [6]. Moreover, this lack of standardization is also evident in studies conducted in different international settings [13], [26], [27] that evaluated automated devices for BP measurement during stress testing. In the absence of specific standardized criteria for exercise conditions, these studies resorted to criteria originally established for BP measurement under resting conditions, such as those proposed by the Association for the Advancement of Medical Instrumentation (AAMI) [28] and the British Hypertension Society (BHS) [29]. Thus, this study compared B³X performance across the different preprocessing approaches using the validation criteria established by these organizations.

In 2018, AAMI, ESH, and ISO experts proposed a universal protocol for BP device validation [30]. The consensus retained the ANSI/AAMI/ISO 81060-2 limits of a mean error within ± 5 mmHg and an SD ≤ 8 mmHg for both SBP and DBP, while also considering at least 85% of errors within 10 mmHg acceptable [28], [30]. In contrast, the BHS protocol grades performance from A to D according to the percentage of absolute errors within 5, 10, and 15 mmHg, with all thresholds of a given grade required to be satisfied (Table II) [29].

The AAMI/ESH/ISO protocol also requires a minimum of 85 participants, with a specified distribution across BP ranges, to ensure that the validation sample adequately represents the intended population [30]. As mentioned in Section III-C, the database used in this work satisfies the minimum sample-size requirement. Nonetheless, because the KS database does not provide subject-level labels, only the BP error criteria of these standards were considered (despite previous recommendations emphasizing the importance of population-related requirements for proper validation [31]). Additionally, the use of AAMI and BHS error thresholds alone as performance benchmarks has also been adopted in BP estimation studies [32].

IV. RESULTS

A. Computational Cost

The proposed methodology was conducted on a virtual machine (VM) hosted in Google Colab equipped with a 2.20 GHz Intel(R) Xeon(R) CPU with 12 cores, 167.1 GB of RAM, and an NVIDIA A100-SXM4 GPU with 80 GB of VRAM. All code was developed in Python, with the U-Net architecture and training procedures implemented using the PyTorch framework. The U-Net training took 34 minutes in total, averaging 6.8 ± 1.3 minutes per fold, with 0.38 GB peak GPU memory usage and 0.57 GPU-hours. Each fold used the same 487,226-parameter architecture, which is relatively compact and could potentially be deployed on low to mid-complexity embedded systems after being trained beforehand and used only for inference.

Finally, the mean inference time of each method was measured across all five test folds, including both the filtering procedure and the subsequent B³X BP estimation. Table III summarizes the corresponding computational results.

TABLE III
MEAN INFERENCE COMPUTATIONAL COST ($n = 2000$)

Method	**Mean Inference Time per sample (ms)**
Clean Samples	5.96
Noisy Samples	5.97
Low-pass filter	7.91
Wiener + Shannon	30.58
BPFzRMS	24.25
U-Net	36.30

The low-pass filter achieved the shortest inference time, while all methods remained fast relative to practical BP measurement. The U-Net was the slowest, although its inference time was close to Wiener + Shannon. BPFzRMS was the second fastest, despite increasing the overall B³X estimation time by approximately four times.

B. Performance of the B³X Algorithm in a Noise Context

Initially, the performance of the B³X algorithm was evaluated without any preprocessing filter, in order to assess the robustness of the algorithm to the noise profile proposed. Fig. 3 shows the correlation and Bland-Altman plots comparing the reference and estimated SBP and DBP values. These plots include all samples from the test subsets of the five-fold cross-validation described in Section III-C.

SBP estimation yielded a linear regression of $y = 0.833x + 28.93$, with an R^2 of 0.403, whereas DBP yielded $y = 0.609x + 17.62$, with an R^2 of 0.214. These results indicate limited correlation, influenced by low-SNR samples.

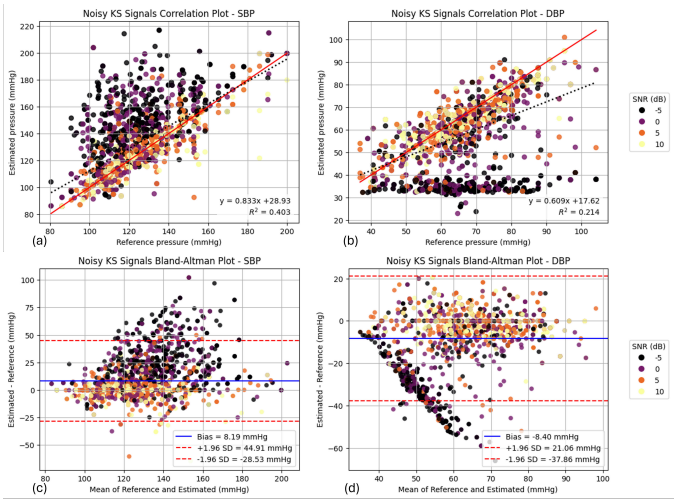


Fig. 3. Correlation plots of B³X-estimated values for (a) SBP and (b) DBP, and Bland–Altman plots for (c) SBP and (d) DBP.

Bland–Altman analysis showed a bias of 8.19 mmHg for SBP, with limits of agreement from -28.53 to 44.91 mmHg, and a bias of -8.40 mmHg for DBP, with limits of agreement from -37.86 to 21.06 mmHg.

The results indicate a lack of robustness of the algorithm to noise, especially at low SNR levels. Although high-SNR samples still show a moderate correlation with the reference values in Fig. 3, low-SNR samples show unacceptable overestimation of SBP and underestimation of DBP according to the AAMI/ESH/ISO criteria. Fig. 4 also highlights the deterioration in algorithm performance as SNR decreases across all five folds. In addition, higher BP values, for both SBP and DBP, were associated with greater estimation errors, regardless of the SNR level.

Table IV shows that B³X meet the AAMI criteria and achieve a BHS grade A classification. However, when applied to noisy KS signals without preprocessing, the B³X algorithm failed to comply with the AAMI requirements under both noise conditions, although it achieved a BHS grade B classification for SNR > 0 dB.

C. BP Estimation Performance Through Filtering Methods

The results presented in Table IV summarize the performance of each preprocessing method across the five folds, reported as the mean $\pm$ standard deviation across folds. The results are stratified by SNR levels of ≤ 0 dB and > 0 dB. Noise-free KS, noisy KS without preprocessing, and each filtering technique are compared according to the corresponding AAMI and BHS validation criteria.

When corrupted KS were filtered, improvements were observed in all evaluated methods, particularly under higher-noise conditions (SNR $\leq$ 0 dB), while gains were more modest under lower-noise conditions (SNR $>$ 0 dB). Both the low-pass filter and BPFzRMS achieved classifications similar to those of the noisy samples. Nevertheless, the low-pass filter enhanced all metrics and achieved a BHS grade B classification for DBP under low-noise conditions. Although BPFzRMS provided greater improvements, particularly at low SNR, it was the only method to yield a worse classification than the noisy samples in one scenario (grade C for SBP at SNR > 0 dB).

Among the evaluated methods, the U-Net achieved the best overall performance for noise suppression in KS signals, meeting the AAMI criteria for both SBP and DBP and being the only method to achieve BHS grade A, although only for SBP under high-SNR conditions. Fig. 4 supports the U-Net results in the AAMI/ESH/ISO $\leq$ 10 mmHg error criterion across all SNR levels, although greater fold-to-fold variability is observed at lower SNRs.

At last, Wiener + Shannon accomplished competitive results to those of the U-Net in DBP, meeting the AAMI criteria for both DBP in both SNR levels and achieving BHS grade B for DBP and for SBP under SNR > 0 dB conditions (despite achieving only grade C for SBP at SNR ≤ 0 dB).

V. Discussion

Across all evaluated methods, performance was consistently better for DBP estimation than for SBP estimation. This difference may be related to the intrinsic characteristics of B³X estimation method, which shows strong effectiveness for DBP estimation, particularly when compared with other BP estimation methods reported in the literature [12]. Indeed, because all filtering methods were evaluated using the same BP estimation algorithm, their measured capability is inherently bounded by the estimation accuracy of B³X.

Fig. 4 shows improved results and reduced fold-to-fold variability with increasing SNR for both SBP and DBP, particularly in the ≤ 10 mmHg metric. This trend may reflect the relative robustness of B³X to noisy samples, although its performance at low SNR remained insufficient to meet the AAMI criteria. Despite the improvements in BP estimation achieved through KS noise filtering, it should be noted that B³X also relies on the oscillometric signal as a complementary input. Under real stress-testing conditions, motion artifacts may degrade not only the KS signal but also the oscillometric signal, potentially reducing the overall performance of the algorithm. Therefore, the apparent robustness of the evaluated methods at low SNR levels may be overestimated when compared with real exercise conditions.

Finally, most methods showed limited improvement under high-noise conditions, particularly for SBP estimation, for which only the U-Net met the AAMI criteria. As discussed in Section IV-A, inference times were sufficiently low for the intended application across methods. Considering its favorable balance among computational cost, relatively compact model size, and results, the U-Net represents a promising preprocessing approach for KS noise filtering to support BP estimation during stress testing. Overall, it achieved the best performance among the evaluated methods, despite not reaching BHS grade A in all scenarios.

TABLE IV

PERFORMANCE COMPARISON GROUPED BY PRESSURE TYPE AND SNR LEVEL.

ME and SD are reported in mmHg. The remaining values (≤ 5, ≤ 10, and ≤ 15) represent mmHg percentages.

SBP, SNR ≤ 0 dB ($n = 1000$)

Model	ME	SD	≤ 5	≤ 10	≤ 15	AAMI	BHS
Clean samples	-1.45 ± 0.85	6.06 ± 0.42	72.50 ± 6.33	91.37 ± 2.44	95.69 ± 1.17	Pass	A
Noisy samples	16.41 ± 2.95	21.47 ± 2.53	27.09 ± 2.59	45.01 ± 6.89	52.46 ± 6.31	Fail	D
Low-pass filter	10.92 ± 2.94	19.77 ± 2.30	37.56 ± 1.99	57.23 ± 4.96	64.52 ± 5.60	Fail	D
Wiener + Shannon	-4.85 ± 0.65	8.69 ± 1.17	53.95 ± 4.14	81.89 ± 2.87	89.06 ± 2.00	Fail	C
BPFzRMS	0.12 ± 1.49	17.12 ± 1.38	37.11 ± 3.60	64.36 ± 3.05	74.72 ± 2.76	Fail	D
U-Net	-2.02 ± 0.99	6.87 ± 0.71	69.77 ± 7.02	88.97 ± 3.17	94.49 ± 1.48	Pass	B

SBP, SNR > 0 dB ($n = 1000$)

Model	ME	SD	≤ 5	≤ 10	≤ 15	AAMI	BHS
Clean samples	-1.21 ± 0.64	5.61 ± 0.26	75.10 ± 7.37	92.05 ± 2.98	97.24 ± 1.36	Pass	A
Noisy samples	0.15 ± 0.47	10.36 ± 1.96	63.33 ± 6.49	84.23 ± 4.11	90.42 ± 1.59	Fail	B
Low-pass filter	-0.43 ± 0.46	8.84 ± 1.05	65.21 ± 6.68	86.97 ± 2.38	92.39 ± 1.28	Fail	B
Wiener + Shannon	-5.28 ± 0.78	8.27 ± 2.01	59.74 ± 4.62	83.70 ± 4.48	90.69 ± 3.22	Fail	B
BPFzRMS	-6.30 ± 0.88	8.95 ± 0.88	53.81 ± 3.71	78.72 ± 4.10	85.97 ± 2.79	Fail	C
U-Net	-1.66 ± 0.77	6.21 ± 0.64	71.21 ± 7.88	91.16 ± 2.66	96.18 ± 1.72	Pass	A

DBP, SNR ≤ 0 dB ($n = 1000$)

Model	ME	SD	≤ 5	≤ 10	≤ 15	AAMI	BHS
Clean samples	-1.05 ± 0.99	6.28 ± 0.75	63.54 ± 6.25	89.03 ± 3.77	95.60 ± 4.21	Pass	A
Noisy samples	-14.81 ± 1.93	17.23 ± 1.49	31.55 ± 3.87	47.69 ± 3.28	56.57 ± 4.68	Fail	D
Low-pass filter	-11.50 ± 2.07	16.14 ± 1.91	38.34 ± 3.86	57.56 ± 4.17	66.65 ± 4.51	Fail	D
Wiener + Shannon	1.03 ± 0.85	6.94 ± 1.20	64.62 ± 3.22	88.42 ± 3.92	94.31 ± 3.33	Pass	B
BPFzRMS	-2.75 ± 0.96	13.09 ± 1.56	53.06 ± 2.60	74.44 ± 3.99	82.37 ± 3.30	Fail	D
U-Net	-1.05 ± 0.77	6.35 ± 0.80	65.17 ± 5.16	89.84 ± 4.47	94.10 ± 4.27	Pass	B

DBP, SNR > 0 dB ($n = 1000$)

Model	ME	SD	≤ 5	≤ 10	≤ 15	AAMI	BHS
Clean samples	-1.16 ± 0.60	6.36 ± 0.57	63.33 ± 3.74	88.41 ± 0.82	95.82 ± 3.56	Pass	A
Noisy samples	-2.15 ± 1.08	8.46 ± 1.38	61.12 ± 4.29	84.74 ± 2.86	91.92 ± 4.77	Fail	B
Low-pass filter	-2.08 ± 0.74	7.72 ± 1.13	61.93 ± 4.26	85.86 ± 3.49	93.43 ± 4.10	Pass	B
Wiener + Shannon	1.58 ± 0.58	6.51 ± 1.06	65.84 ± 4.02	87.15 ± 3.42	94.87 ± 3.86	Pass	B
BPFzRMS	2.28 ± 0.45	6.75 ± 0.93	65.05 ± 3.08	86.30 ± 4.15	93.17 ± 3.53	Pass	B
U-Net	-1.00 ± 0.63	6.26 ± 0.44	65.04 ± 5.57	90.23 ± 1.43	94.88 ± 2.65	Pass	B

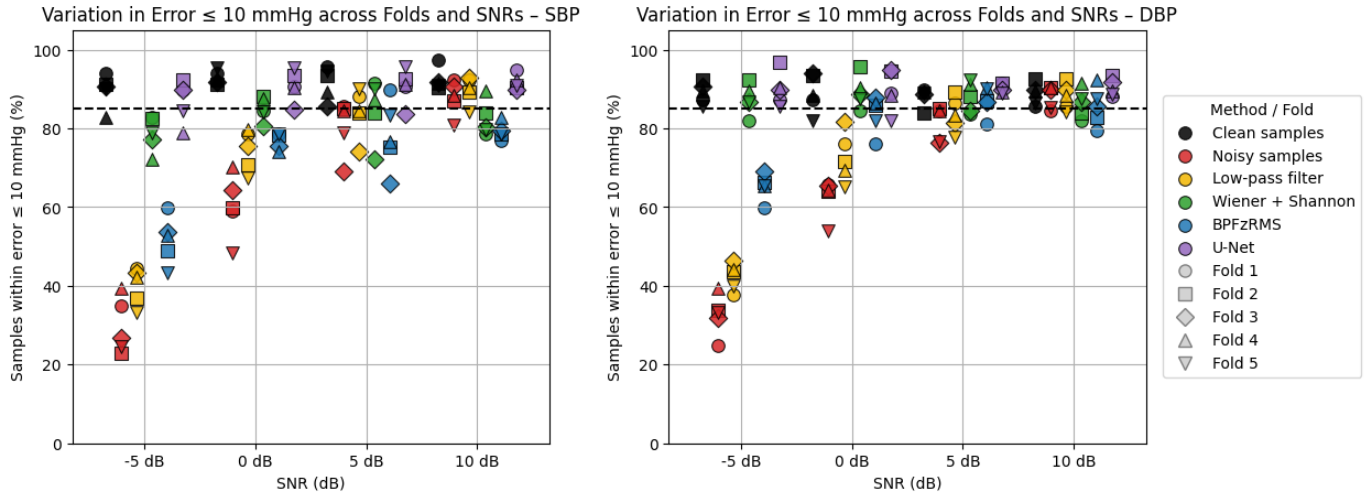


Fig. 4. Variation in the percentage of errors ≤ 10 mmHg across folds and SNR levels for (a) SBP and (b) DBP. The dashed line indicates the 85% threshold considered acceptable according to ANSI/AAMI/ISO.

A. Limitations and Future Work

Despite the promising results on noise filtering for BP detection, this study presents relevant limitations that could be overcome in future work.

First, as mentioned in Subsection III-C, it is important to reinforce the absence, as far as this study could find in the literature, of an available KS signals database in treadmill stress tests or functional assessment protocols. The proposed Study Database is a synthesized database designed to simulate noise conditions BP measurements during exercise. It should be noted that the noise profile of real measurements may present acoustic differences, such as cuff and microphone rubbing against the patient's arm, in addition to treadmill footstep noise captured from a microphone position different from that used in the Noise Database.

Second, since the reference for the KS Database does not relate samples to patient profiles (i.e. age, sex, health condition, etc), future work should rely on a dedicated stress test database with different patient profiles, including KS signals, cuff pressure curves, linkage between each sample and the corresponding de-identified patient profile, and ground truth SBP and DBP collected by physicians. The database should preferably be collected at different medical centers to evaluate the generalizability of the proposed methods.

Third, all filtering methods were tested with the B³X Algorithm. Future work should evaluate these methods with other algorithms under similar noise conditions. Although B³X shows promising results for BP detection, it relies on the oscillometric signal, which may be unreliable during exercise due to motion artifacts. Therefore, future studies should investigate algorithms that do not depend on oscillometric signals and assess the impact of the proposed filtering methods under these conditions.

Finally, sequential combinations of methods, such as U-Net followed by Wiener + Shannon or other filtering pipelines, could also be investigated to improve performance. Future optimization could also help the methods more consistently satisfy the established criteria.

VI. CONCLUSION

This work extends the current literature by introducing a data augmentation pipeline for generating synthetic exercise-

related KS datasets. It also refines existing KS filtering approaches, evaluates these techniques under simulated exercise stress test noise conditions, and demonstrates the feasibility of transferring a U-Net architecture originally developed for PCG denoising to KS filtering. These contributions show that filtering techniques can substantially enhance automatic blood pressure estimation from Korotkoff signals artificially contaminated by treadmill noise. Among the evaluated methods, U-Net achieved the best overall performance, restoring the B³X algorithm to levels close to those obtained with noise-free signals.

REFERENCES

[1] T. G. Pickering et al., "Recommendations for Blood Pressure Measurement in Humans and Experimental Animals: Part 1: Blood Pressure Measurement in Humans: A Statement for Professionals From the Subcommittee of Professional and Public Education of the American Heart Association Council on High Blood Pressure Research," *Circulation*, vol. 111, no. 5, pp. 697–716, Feb. 2005, doi: 10.1161/01.CIR.0000154900.76284.F6.

[2] P. Muntner et al., "Measurement of Blood Pressure in Humans," *Hypertension*, vol. 73, no. 5, pp. e35–e66, May 2019, doi: 10.1161/HYP.0000000000000087.

[3] K. D. Currie, M. G. Schultz, P. J. Millar, and L. S. Pescatello, "The Role of Exercise Blood Pressure in Hypertension: Measurement, Mechanisms, and Management," *Medicine & Science in Sports & Exercise*, vol. 57, no. 2, pp. 425–433, Feb. 2025, doi: 10.1249/MSS.0000000000003569.

[4] J. E. Sharman and A. LaGerche, "Exercise blood pressure: clinical relevance and correct measurement," *Journal of Human Hypertension*, vol. 29, no. 6, pp. 351–358, Jun. 2015, doi: 10.1038/jhh.2014.84.

[5] M. G. Schultz et al., "Exercise and Sports Science Australia (ESSA) - Consensus Statement," *Journal of Science and Medicine in Sport*, vol. 28, no. 8, pp. 605–612, Aug. 2025, doi: 10.1016/j.jsams.2025.05.011.

[6] T. D. Carvalho et al., "Diretriz Brasileira de Ergometria em População Adulta – 2024", Arq. Bras. Cardiol., v. 121, n. 3, p. e20240110, 2024, doi: 10.36660/abc.20240110.

[7] C. F. Babbs, "The origin of Korotkoff sounds and the accuracy of auscultatory blood pressure measurements," *Journal of the American Society of Hypertension*, vol. 9, no. 12, pp. 935–950.e3, Dec. 2015, doi: 10.1016/j.jash.2015.09.011.

[8] A. Shennan, M. Gupta, M. de Swiet, A. Halligan, and D. J. Taylor, "Lack of reproducibility in pregnancy of Korotkoff phase IV as measured by mercury sphygmomanometry," *The Lancet*, vol. 347, no. 8995, pp. 139–142, Jan. 1996, doi: 10.1016/S0140-6736(96)90338-4.

[9] E. T. O'Brien and K. O'Malley, "ABC of Blood Pressure Measurement: The observer," British Medical Journal, vol. 2, no. 6193, pp. 775–776, Sep. 1979, doi: 10.1136/bmj.2.6193.775

[10] R. Shuqi, D. Xiaoyan, F. Yubo, and S. Anqiang, "A review of blood pressure measurement methods based on Korotkoff sounds," *Medical Novel Technology and Devices*, vol. 28, p. 100409, Dec. 2025, doi: 10.1016/j.medntd.2025.100409.

[11] M. Alvarado Alvarez, R. Padwal, and W. Hiebert, "Masking of Korotkoff sounds used in blood pressure measurement through auscultation," *Journal of the Acoustical Society of America*, vol. 153, no. 3, pp. 1496–1505, Mar. 2023, doi: 10.1121/10.0017354.

[12] J. Centracchio, D. De Caro, P. Bifulco, and E. Andreozzi, "B3 X: a novel efficient algorithm for accurate automated auscultatory blood pressure estimation," *Physiological Measurement*, vol. 44, no. 9, p. 095007, Sep. 2023, doi: 10.1088/1361-6579/acf643.

[13] J. D. Cameron, I. Stevenson, E. Reed, B. P. McGrath, A. M. Dart, and B. A. Kingwell, "Accuracy of automated auscultatory blood pressure measurement during supine exercise and treadmill stress electrocardiogram-testing," *Blood Pressure Monitoring*, vol. 9, no. 5, pp. 269–275, Oct. 2004, doi: 10.1097/00126097-200410000-00007.

[14] A. Argha, B. G. Celler, e N. H. Lovell, "Artificial Intelligence Based Blood Pressure Estimation From Auscultatory and Oscillometric Waveforms: A Methodological Review", IEEE Reviews in Biomedical Engineering, v. 15, p. 152–168, 2022, doi: 10.1109/RBME.2020.3040715.

[15] A. Argha, "Auscultatory and Oscillometry Blood Pressure Data," *IEEE DataPort*, May 2020, doi: 10.21227/2J9C-1J21.

[16] A. R. Krishna, R. K. V, P. M. Nabeel, e J. Joseph, "Calibration-Free Diastolic Pressure Estimation Using Brachial Pulse Wave Analysis: Feasibility Study", em 2025 IEEE Medical Measurements & Applications (MeMeA), 2025, p. 1–5. doi: 10.1109/MeMeA65319.2025.11068091.

[17] M. Ezeddin et al., "Oscillometric blood pressure estimation using machine learning-based mapping of waveform features", Biomed. Eng. Lett., v. 15, n. 6, p. 1123–1134, nov. 2025, doi: 10.1007/s13534-025-00496-2.

[18] A. Argha, B. G. Celler, H. Alinejad-Rokny, and N. H. Lovell, "Blood Pressure Estimation From Korotkoff Sound Signals Using an End-to-End Deep-Learning-Based Algorithm," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022, doi: 10.1109/TIM.2022.3217865.

[19] J. Nie et al., "SoundTrack: A Contactless Mobile Solution for Real-time Running Metric Estimation for Treadmill Running in the Wild," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 2, pp. 1–30, Jun. 2025, doi: 10.1145/3729486.

[20] G. Lee e H. Shin, "Multitask learning-based phonocardiogram denoising model for preserving valvular heart disease characteristics", Physiol. Meas., v. 47, n. 7, p. 075017, jul. 2026, doi: 10.1088/1361-6579/ae8caa.

[21] S. K. Ghosh e R. N. Ponnalagu, "Investigation of discrete wavelet transform domain optimal parametric approach for denoising of phonocardiogram signal", J. Mech. Med. Biol., v. 22, n. 06, p. 2250046, ago. 2022, doi: 10.1142/S0219519422500464.

[22] B. Zhu, Z. Zhou, S. Yu, X. Liang, Y. Xie, and Q. Sun, "Review of Phonocardiogram Signal Analysis: Insights from the PhysioNet/CinC Challenge 2016 Database," *Electronics*, vol. 13, no. 16, Art. no. 16, Jan. 2024, doi: 10.3390/electronics13163222.

[23] H. Zhang, R. Wang, H. Zhou, S. Xia, S. Jia, and Y. Wu, "Prediagnosis of Heart Failure (HF) Using Deep Learning and the Korotkoff Sound," *Applied Sciences*, vol. 12, no. 20, p. 10322, Oct. 2022, doi: 10.3390/app122010322.

[24] A. Mukherjee, R. Banerjee, and A. Ghose, "A Novel U-Net Architecture for Denoising of Real-world Noise Corrupted Phonocardiogram Signal," *arXiv preprint arXiv:2310.00216*, Sep. 2023, doi: 10.48550/arXiv.2310.00216.

[25] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, vol. 1. Upper Saddle River, NJ, USA: Prentice Hall, 1993, sec. 12.7, pp. 400–414.

[26] S. E. Griffin, R. A. Robergs, e V. H. Heyward, "Blood pressure measurement during exercise: a review":, Medicine & amp Science in Sports & amp Exercise, v. 29, n. 1, p. 149–159, jan. 1997, doi: 10.1097/00005768-199701000-00022.

[27] T. Wibmer et al., "Blood pressure monitoring during exercise: Comparison of pulse transit time and volume clamp methods", Blood Pressure, v. 24, n. 6, p. 353–360, nov. 2015, doi: 10.3109/08037051.2015.1053253.

[28] W. B. White et al., "National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers," *Hypertension*, vol. 21, no. 4, pp. 504–509, Apr. 1993, doi: 10.1161/01.HYP.21.4.504.

[29] E. O'Brien, B. Waeber, G. Parati, J. Staessen, and M. G. Myers, "Blood pressure measuring devices: recommendations of the European Society of Hypertension," *BMJ*, vol. 322, no. 7285, pp. 531–536, Mar. 2001, doi: 10.1136/bmj.322.7285.531.

[30] G. S. Stergiou et al., "A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement", J Hypertens, v. 36, n. 3, p. 472–478, mar. 2018, doi: 10.1097/HJH.0000000000001634.

[31] Y. Yu e A. Lowe, "Advancements and challenges in blood pressure monitoring using pulse wave propagation: a comprehensive review and ISO 81060-2 based statistical analysis", Hypertens Res, v. 49, n. 7, p. 2156–2180, jul. 2026, doi: 10.1038/s41440-026-02651-3.

[32] J. Long, H. Tang, M. Wang, e Y. Chen, "Multimodal-RTAGNet: a cross-modal attention network with gated fusion for noninvasive and continuous intraoperative blood pressure estimation", Biomedical Signal Processing and Control, v. 123, p. 110492, set. 2026, doi: 10.1016/j.bspc.2026.110492.

Electrocardiogram Analysis from Beat-Level Arrhythmias to WPW Ventricular Pre-Excitation and Glycemic Disorders: A Scoping Review

1st Áris S. Canto
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
aris.canto@ufam.edu.br

2nd Marly G. F. Costa
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
mcosta@ufam.edu.br

3rd Cícero F. F. Costa Filho
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
ccosta@ufam.edu.br

***Abstract*—This scoping review analyzes 16 studies on computational methods for the analysis of cardiac arrhythmias, ventricular pre-excitation associated with Wolff-Parkinson-White syndrome, and glycemic disorders. The corpus was compiled from records identified in the IEEE Xplore, PubMed, and Scopus databases, supplemented by searches using the Consensus tool. The studies were compared regarding datasets, input representations, computational methods, validation protocols, performance, explainability, and limitations. The results demonstrate high performance, particularly on controlled datasets, but also reveal recurring limitations, including beat-level data splitting without patient-level separation, limited external validation, small sample sizes, reliance on pre-trained models, predominantly qualitative explainability assessment, and insufficient control of confounding factors. Regarding ventricular pre-excitation, further research is needed to integrate segmentation, fiducial features, temporal signals, time-frequency representations, and explainability methods with more robust clinical validation. As for glycemic alterations, the findings support the use of the ECG as a complementary marker for hyperglycemia and hypoglycemia, although it does not replace reference glucose measurements. Overall, the review highlights the potential of ECG-derived physiological features for machine learning models aimed at detecting cardiac and metabolic anomalies, while emphasizing the need for more robust validation and clinically grounded assessment.**

***Keywords*—scoping review, cardiac arrhythmias, ventricular pre-excitation, Wolff-Parkinson-White, glycemic abnormalities, explainability.**

I. Introduction

The electrocardiogram (ECG) remains one of the most accessible tools for assessing cardiac electrical activity. The digitization of large clinical datasets and the expansion of wearable sensors have fostered the development of methods capable of detecting peaks, segmenting waves, and classifying rhythms directly from the signal. Convolutional networks, recurrent models, and hybrid architectures have outperformed strictly rule-based approaches in several tasks, particularly when the objective is to recognize complex morphological or temporal abnormalities [1]–[3].

However, improved predictive performance does not eliminate methodological and clinical challenges, because accuracy values near or above 99% may be directly associated with factors such as controlled datasets, beat-level partitions, synthetic balancing, or indirect reuse of the same individuals during model training and testing. In addition, explanations produced by Grad-CAM, SHAP, or LIME are often evaluated only through visual inspection, without measures of fidelity, stability, or clinical agreement. Distinguishing a plausible explanation from one that faithfully reflects the model's decision-making mechanism is particularly important in ECG analysis, where small changes in duration, amplitude, and slope may carry diagnostic significance.

Wolff-Parkinson-White (WPW) syndrome represents a relevant example of this challenge, because its interpretation relies on the joint assessment of the PR interval, the onset of the QRS complex, the delta wave, and ventricular depolarization duration [4]. Specific approaches remain scarce, and general arrhythmia models may exhibit reduced sensitivity when externally evaluated for WPW [5]. In parallel, another line of research uses morphological and autonomic ECG changes to infer prediabetes, hyperglycemia, or hypoglycemia, with the aim of providing non-invasive support for screening and continuous monitoring [6]–[9].

In this context, ventricular pre-excitation associated with Wolff-Parkinson-White (WPW) syndrome and glycemic disorders constitute the two main focuses of this scoping review. Methods for R-peak detection, fiducial point delineation, arrhythmia classification, multimodal ECG representation, and explainable artificial intelligence (XAI) provide the methodological basis for analyzing these topics. This review aims to identify key computational approaches, limitations, and research gaps in beat-level ECG analysis, with an emphasis on temporal, morphological,

and time-frequency representations, multimodal machine learning models, and XAI techniques.

II. Review Methodology

This study consists of a scoping review of computational approaches applied to ECG signal analysis, focusing on Wolff-Parkinson-White (WPW) ventricular pre-excitation syndrome and glycemic disorders. The review followed a predefined methodology comprising the formulation of the research question, search strategy, eligibility criteria, study selection, data extraction, and qualitative synthesis. Study identification, screening, eligibility assessment, and reporting followed the PRISMA-ScR framework [10].

During title and abstract screening, records were excluded if they described purely clinical studies without computational methods, did not analyze ECG signals (e.g., studies based exclusively on PPG), or were unrelated to arrhythmias, WPW syndrome, or glycemic disorders. Following screening and report retrieval, 65 reports underwent full-text eligibility assessment. Studies were eligible for inclusion if they applied computational methods to ECG analysis, provided sufficient methodological information for data extraction and comparison, and addressed topics relevant to the scope of the review. At the full-text stage, reports were excluded if they were restricted to ECG preprocessing, filtering, or signal reconstruction without subsequent computational analysis, focused on specific conditions outside the scope of the review, or provided insufficient methodological detail for analysis and data extraction. From the included studies, information was extracted on datasets, populations, ECG configuration, preprocessing and segmentation, computational models, validation strategies, performance metrics, explainability methods, and reported limitations. The analysis also considered subject-wise data partitioning, sample representativeness, risk of data leakage, external validation, and quantitative assessment of explainability.

The PCC framework (Population/Problem, Concept, and Context) guided the formulation of the research question and search strategy [11]. The main and extended search strings combined terms related to ECG, R-peak detection, beat segmentation, fiducial points, QRS complexes, morphological, temporal, and time-frequency features, machine and deep learning, multimodal approaches, and XAI, with additional terms related to WPW and glycemic disorders. Searches were conducted in IEEE Xplore, Scopus, and PubMed, complemented by a supplementary search using Consensus, and were restricted to studies published between 2021 and 2026. A total of 156 records were identified through IEEE Xplore (n = 43), Scopus (n = 64), PubMed (n = 9), and the supplementary Consensus search (n = 40). After removing 14 duplicates and 35 records for other reasons, 107 records were screened, of which 28 were excluded. Among the 79 reports sought for retrieval, 14 could not be retrieved, leaving 65 reports for eligibility assessment. Of these, 49 were excluded according to the predefined criteria, resulting in 16 studies being included in the review. Figure 1 summarizes the PCC structure, the research question, the search strings, and the PRISMA-ScR study selection flow.

Given the heterogeneity of populations, datasets, ECG representations, computational models, performance metrics, and validation protocols, a qualitative narrative synthesis was performed. The analysis focused on methodological patterns, limitations, research gaps, subject-wise validation, external validation, and the integration of morphological, temporal, time-frequency, multimodal, and explainable approaches. A meta-analysis was not performed because of this methodological heterogeneity. The main characteristics and findings of the included studies are summarized in Table I.

III. Literature Review

A. Studies on Cardiac Arrhythmias and Wolff-Parkinson-White Syndrome

Reliable beat-level ECG analysis depends on accurate R-peak detection, fiducial point delineation, and consistent beat segmentation. These steps are essential for extracting temporal and morphological features from the P wave, QRS complex, and T wave, particularly in conditions such as ventricular pre-excitation. Although deep-learning models have improved ECG classification performance, limited interpretability remains a barrier to clinical application. Therefore, explainability methods should be evaluated not only visually, but also in terms of robustness, reproducibility, and consistency with cardiological knowledge.

Wang, Pan, and Meng [2] directly addressed R-peak detection under noisy conditions using a 1D ResU-Net that combined an encoder-decoder structure, residual blocks, skip connections, and channel attention. The method achieved an F1-score of 99.79% on the MIT-BIH Arrhythmia Database and approximately 99.5% in cross-database evaluations using INCART and the QT Database. Under extreme noise at -6 dB, F1-scores remained between 96.28% and 97.49%. This robustness is favorable for continuous applications, although the 36.4-million-parameter architecture and the potential leakage introduced by overlapping windows require caution.

Prakash et al. [3] extended R-peak alignment to the explicit delineation of the onsets and offsets of the P, QRS, and T waves. They proposed a modified 1D U-Net trained on the Lobachevsky University Database (LUDB) and used it to extract more than 12 temporal and morphological features, which were subsequently classified using a Random Forest. Segmentation accuracies ranged from 99.12% to 99.99%, but the patient-specific beat evaluation was conducted predominantly within the same record, reducing test independence and limiting data generalizability.

Merbouti and Cherifi [4] proposed a seven-stage processing pipeline combining the discrete wavelet transform,

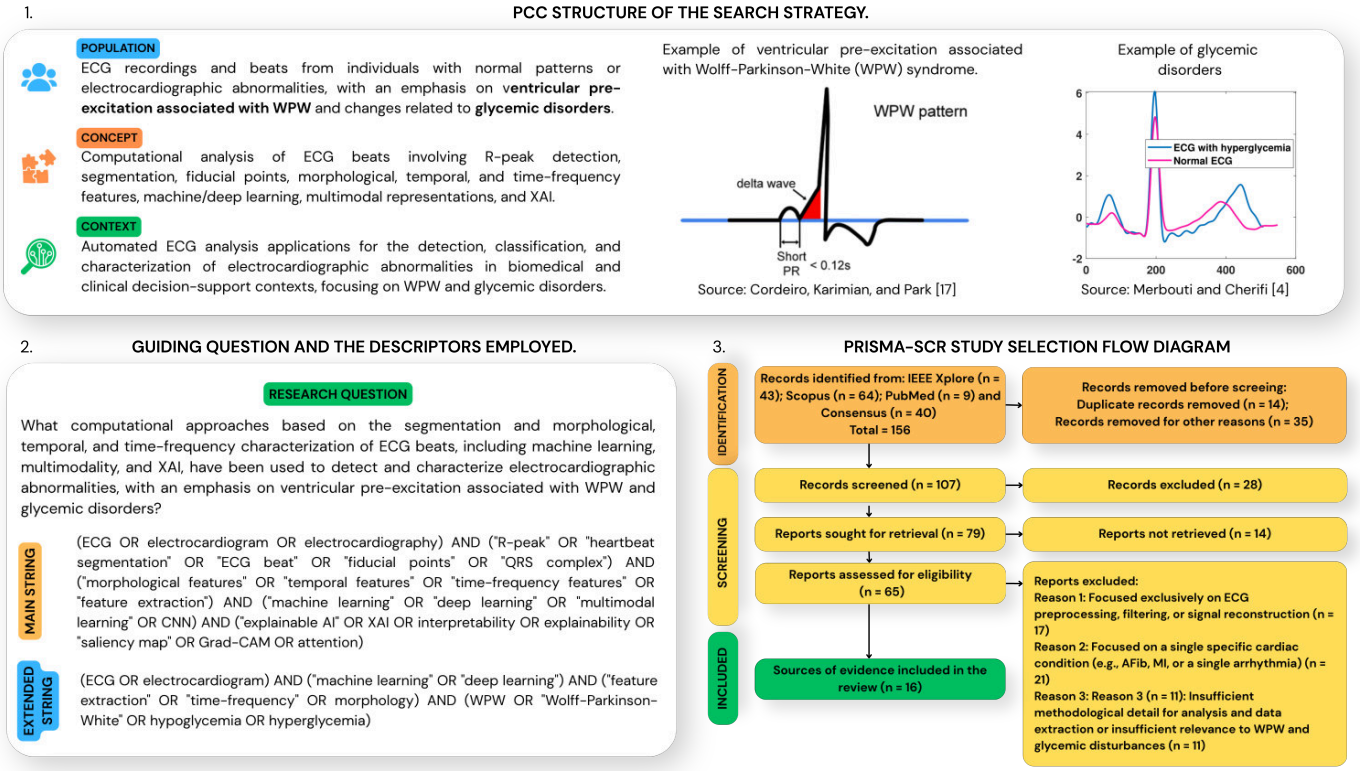


Figure 1. Search and study selection strategy, including the PCC framework with examples of ventricular pre-excitation associated with Wolff–Parkinson–White (WPW) syndrome and glycemic abnormalities, the research question, search strings, and the PRISMA flow for study identification, screening, and inclusion. In the flowchart, AFib refers to atrial fibrillation and MI to myocardial infarction.

adaptive peak scanning, and neural-network, KNN, and Naïve Bayes models. The method explicitly seeks to locate the delta wave and verify criteria associated with ventricular pre-excitation, such as PR-interval shortening and morphological relationships among the P, Q, R, and S waves. Neural networks achieved a mean accuracy of 99.25% and sensitivity of 97.16%. However, the sample was limited to only 21 WPW segments and 101 sinus-rhythm segments. In addition, the absence of conditions with similar electrocardiographic characteristics, such as bundle branch blocks, ventricular hypertrophy, and the Lown-Ganong-Levine pattern, restricted the evaluation of differential diagnoses and prevented a more realistic estimate of the method's specificity for identifying the delta wave and pre-excitation pattern.

Canto et al. [1] used asymmetric 600-ms segments containing 216 samples from lead MLII to preserve the P, QRS, and T waves. The proposed 1D CNN classified ten categories with 99.40% accuracy, 99.40% precision, and 99.32% recall. Grad-CAM was used for explainability and indicated activations consistent with morphological regions, but the use of a single dataset, dependence on annotated peaks, and lack of external validation limit the strength of the study's evidence.

Pantelidis et al. [5] developed ECG-XPLAIM, a multiscale 1D CNN with Inception modules and residual connections. The model was trained on MIMIC-IV and externally validated on PTB-XL. External AUC values ranged from 0.878 to 0.994. For WPW, sensitivity was 0.773, specificity was 0.973, and AUC was 0.895. These results demonstrate that a large-scale architecture can generalize across datasets, but also show that high specificity does not fully compensate for reduced sensitivity under rare conditions.

Singh and Sharma [12] combined a ResNet-inspired CNN, ECANet channel attention, and the K-GradCam method. Adding attention increased the overall F1-score from 97.40% to 98.88%. K-GradCam achieved a Dice coefficient of 0.82 relative to Grad-CAM and 0.71 relative to SHAP, with lower computational cost. However, agreement among methods does not constitute direct evidence of clinical fidelity. The author do not state whether the explanations were compared with the cardiologists annotations.

Neves et al. [13] evaluated LIME, SHAP, and Permutation Sample Importance in k-NN and 1D CNN classifiers. The inter-patient split and comparison with shapelets strengthened the study design. LIME applied to amplitude and derivative representations achieved a Jaccard similarity of 0.88 ± 0.27, whereas SHAP performed close to the random baseline. In a study involving three readers and 20 ECGs, the explanations did not consistently improve di-

agnostic accuracy and received low overall utility ratings. This finding highlights that computational plausibility and clinical utility are distinct criteria.

Sathi et al. [14] introduced federated learning using LSTM, FedAvg, and nine fiducial features. Classification of arrhythmia, ischemia, and healthy condition achieved an accuracy of $92.3 \pm 0.7\%$, an F1-score of 0.91 ± 0.02, and an AUC of 0.99 ± 0.01. SHAP and LIME highlighted P- and R-wave amplitudes, RR, PR, and QT/QTc intervals, and QRS duration. Although subject-wise separation is a strength, each diagnosis was associated with a different dataset, creating a risk that the model learned database-specific characteristics rather than the clinical condition.

Duranay et al. [15] compared 17-layer CNNs trained on the 1D lead-II signal and on a grayscale image generated from the same tracing. The 2D modality achieved 97.44% accuracy and an F1-score of 97.38%, outperforming the 93.07% accuracy and 92.99% F1-score obtained with the 1D modality. MM-GradCAM synchronized temporal and visual explanations, which were qualitatively evaluated by a cardiologist. It should be noted that the image is an alternative representation of the same ECG signal and does not constitute an independent physiological modality. Furthermore, processing this 2D representation of the ECG involves higher computational costs for training and storage than 1D ECG processing.

Alqudah and Alfraihat [16] combined ECG and phonocardiogram signals in a CNN-BiLSTM architecture with Squeeze-and-Excitation blocks, self-attention, and cross-attention. Overall accuracy was 97.0%, with a mean AUC of 0.989. Grad-CAM, SHAP, and Integrated Gradients were evaluated by three cardiologists, with an agreement of $\kappa = 0.82$. Despite this advance, a substantial portion of the public signals was synthetically paired by diagnosis without preserving the electromechanical relationship between the ECG and heart sound of the same individual.

According to the studies reviewed, the robustness of detection and segmentation depends on the accurate identification of ventricular depolarization, the definition of physiologically appropriate windows, and the correct delineation of P and T waves. In cases of ventricular pre-excitation, this delineation is particularly relevant due to the temporal relationship between the P wave, the PR interval, and the onset of the QRS complex. Regarding explainability, aspects such as model fidelity, stability against perturbations, and clinical consistency must be considered. However, the literature still relies predominantly on visual inspection, with limited use of quantitative validation or validation by multiple experts. In multimodal approaches, it is crucial to distinguish the integration of distinct modalities from a simple change in representation. Temporal signals, derived features, and time-frequency representations can provide complementary information, but their incremental contribution and potential redundancy must be quantified. In turn, physiologically distinct modalities, such as ECG and phonocardiography, require simultaneous acquisition or previously validated synchronization in the same patient.

B. Studies on metabolic monitoring and the detection of glycemic disturbances based on ECG

The search for noninvasive methods or approaches complementary to conventional glucose-monitoring systems has motivated studies investigating ECG as an indirect marker of metabolic abnormalities. The central hypothesis is that variations in glucose concentration can affect cardiac electrical conduction, autonomic nervous system regulation, and specific morphological and temporal characteristics of the electrocardiographic signal. These changes may manifest in intervals, waves, amplitudes, and heart-rate-variability patterns, providing potentially useful features for machine-learning models aimed at screening for prediabetes, hyperglycemia, and other glycemic abnormalities.

Among approaches aimed at identifying metabolic abnormalities from ECG, Kulkarni et al. [6] developed DiaBeats, an XGBoost-based model using beats extracted from 12-lead ECGs of 1,262 participants. Each beat was represented by 1,200 features, and the data were divided into training, validation, and test sets in proportions of 85%, 5%, and 10%, respectively. The model achieved an accuracy of 96.8% and an F1-score of 96.6%, while *temperature scaling* reduced the expected calibration error from 10.36% to 6.32%. However, partitioning was performed at the beat level without explicit participant separation, potentially allowing samples from the same individual to appear in different subsets and thereby favoring optimistic performance estimates. In addition, the DISFIN cohort consisted of Sindhi families from Nagpur with a history of type 2 diabetes, with prevalences of approximately 30% for diabetes and 14% for prediabetes. Therefore, although the results indicate potential for discriminating among different glycemic states, population-level screening still requires external validation and subject-independent partitioning.

Koga et al. [7] presented DiaCardia, based on LightGBM and morphological and temporal features extracted from ECG. The model achieved an AUC of 0.851 in internal testing and 0.785 in external validation conducted at another institution using different equipment. The lead-I-only version achieved an AUC of 0.844, with no statistically significant difference from the 12-lead model. Strengths included subject-wise data separation, control of confounding factors, and external validation. However, the positive predictive value of approximately 21% demonstrates that good discriminative performance does not eliminate the influence of prevalence and false positives.

Cordeiro, Karimian, and Park [17] investigated hyperglycemia identification from ECG fiducial changes using a ten-layer fully connected deep neural network. The model extracted 18 features per heartbeat, consisting of distances and slopes between fiducial points, thereby reducing di-

mensionality compared to using the full cardiac cycle. Figure 2 shows examples of heartbeats from individuals with and without hyperglycemia, visually highlighting morphological differences between the two conditions. The dataset included 1,119 participants and 1,963 single-lead ECG recordings, totaling 51,518 segments after preprocessing. In testing, the model achieved an AUC of 94.53%, sensitivity of 87.57%, and specificity of 85.04%, as well as a mean AUC of 93.65% in ten-fold cross-validation. However, participants were not fasting even though hyperglycemia was defined as glucose above 100 mg/dL, which may introduce label uncertainty because of postprandial elevations. In addition, the 80/20 split was performed at the beat level, without an explicit description of participant-wise separation, potentially increasing similarity between training and test sets. The absence of external validation in another population or device further limits the evidence regarding the broad applicability oh the method.

Vyas et al. [8] investigated the complementarity between ECG morphology and inter-beat dynamics for identifying glycemic excursions in ten young adults with type 1 diabetes. They compared morphological information, RR-interval sequences, statistical heart-rate and variability features, and a late-fusion strategy. The combined model achieved the best performance, with a mean AUC of approximately 0.68 for hypoglycemia and 0.70 for hyperglycemia. The use of windows adapted to the RR interval also increased the AUC for hypoglycemia from 52.2% to 56.1% ($p = 0.008$). Morphology contributed more strongly to hypoglycemia, whereas beat timing was more relevant to hyperglycemia, and combining the sources outperformed either representation alone for both tasks. However, evaluation was personalized for each participant without separation between individuals, and the small sample size limits generalization to population-level models. In addition, the higher frequency of hyperglycemic episodes may have favored better performance for that condition.

Islam, Islam, and Hossain Shuvo [18] evaluated Random Forest (RF), Gradient Boosting Machine (GBM), and MLP models for personalized hypoglycemia detection using single-lead ECG from the D1NAMO dataset. Of the nine participants with type 1 diabetes, only four were included because of the availability of hypoglycemic episodes. A total of 74 morphological, interval-based, and HRV features were extracted, 44 of which were related to heart-rate variability. RF and GBM achieved a mean AUC of 0.97 and an F1-score of 0.86; after removing HRV features, the F1-score decreased to 0.74 and 0.73, respectively. SHAP analysis highlighted the PQ interval, QRS duration, QT/QTc parameters, and HRV features. However, the models were trained separately for each participant using an internal 70/30 split and were not validated on unseen individuals. The exclusion of five participants and the lack of external *subject-wise* validation limit the method's generalization.

Dave et al. [9] evaluated a personalized approach for detecting glycemic excursions under free-living conditions using ECG and CGM data from 12 adults with type 1 diabetes monitored for 14 days. Individual models were developed by combining intra-beat morphology and HRV features, together with classifiers trained at different glycemic thresholds. The fusion model achieved an AUC of 0.75 ± 0.10 for hypoglycemia and 0.80 ± 0.06 for hyperglycemia, outperforming the isolated representations. Five-fold cross-validation used one-hour temporal blocks to reduce correlation between training and test sets. Nevertheless, precision for hypoglycemia was only 0.10 ± 0.03, with a high false-positive rate. Event-level evaluation yielded 2.07 false alarms per day, demonstrating that high AUC does not guarantee adequate operational performance and reinforcing the importance of metrics such as precision and false-alert frequency.

The evidence analyzed indicates that ECG-based detection performance for glycemic disturbances is strongly influenced by the study's validation protocol. Results obtained using beat-level data splitting can overestimate model performance compared to subject-level or external validation strategies. Personalized approaches suggest that ECG morphological features and heart rate variability can provide complementary information; however, their generalizability remains limited. These findings underscore the importance of subject-level data splitting, external validation, and the use of performance metrics beyond AUC to ensure a more reliable assessment of model robustness.

IV. Discussion, Trends, and Research Gaps

A. Apparent performance and external validity

The reviewed studies are not directly comparable due to differences in analysis units, class prevalence, number of categories, and data partitioning. Datasets with thousands of heartbeats may still represent few individuals, and heartbeat-level splitting can place data from the same patient in both training and test sets, increasing the risk of information leakage and overly optimistic performance. Patient-level splitting therefore provides a more reliable estimate of generalization, while external validation assesses robustness across populations, institutions, acquisition protocols, and devices. Pantelidis et al. [5], Wang et al. [2], and Koga et al. [7] showed that external evaluation can reveal effects not evident internally, including reduced WPW sensitivity, stable R-peak detection across datasets, and lower AUC in glycemic screening. These findings reinforce the need to consider both partitioning strategy and external validation when interpreting model performance and clinical applicability.

B. Explainability as a validation process

The reviewed literature primarily uses Grad-CAM to identify relevant signal regions, SHAP to estimate feature

Table I

SUMMARY OF THE REVIEWED STUDIES. THE FINAL COLUMN PROVIDES A DESCRIPTIVE METHODOLOGICAL ASSESSMENT EMPHASIZING PARTICIPANT SEPARATION, EXTERNAL VALIDATION, POTENTIAL LEAKAGE, EXPLAINABILITY EVALUATION, AND OTHER STUDY-SPECIFIC LIMITATIONS.

Author	Objective	Methodology	Data and input	Main results	Methodological assessment and limitations
Studies on Cardiac Arrhythmias and Wolff-Parkinson-White Syndrome					
Canto et al. (2025)	Classify ten heartbeat types and explain relevant regions.	1D CNN; asymmetric R-centered window; class weights; Grad-CAM.	MIT-BIH; MLII; 600 ms; 216 samples.	Accuracy 99.40%; precision 99.40%; recall 99.32%.	Single dataset; split not explicitly inter-patient; annotated peaks; no external validation.
Wang et al. (2025)	Detect R peaks under noise and across different databases.	1D ResU-Net; residual blocks; attention; focal loss; *augmentation*.	MIT-BIH, INCART, QT, and Noise Stress Test; 5-s windows.	F1 99.79% on MIT-BIH and approximately 99.5% across databases; F1 >96% at −6 dB.	Large model; overlapping windows; R peak only; no prospective validation.
Prakash et al. (2023)	Delineate P/QRS/T and classify beats using transparent features.	1D U-Net; fiducial features; personalized Random Forest.	LUDB and MIT-BIH; 12 and 2 leads.	Segmentation 99.12-99.99%; class-wise accuracy 93.64-98.98%.	Evaluation partly conducted on records used for adaptation; low precision for rare classes.
Merbouti and Cherifi (2023)	Identify short PR interval and delta wave for WPW.	DWT; adaptive scanning; ANN, KNN, and Naïve Bayes; seven stages.	**122 segments: 21 WPW and 101 normal**; MLII; 10 s.	ANN: accuracy 99.25%; sensitivity 97.16%; mean specificity 74.37%.	Small sample; no differential controls; cumulative errors; minimal external validation.
Pantelidis et al. (2025)	Detect multiple arrhythmias with explainability and external validation.	ECG-XPLAIM; multiscale Inception CNN; residual connections; Grad-CAM.	MIMIC-IV and PTB-XL; 12 leads; 10 s.	External AUC 0.878-0.994; WPW: sens. 0.773, spec. 0.973, AUC 0.895.	Reduced sensitivity in rare classes; qualitative explanations; no calibration/AUPRC.
Singh and Sharma (2022)	Classify six beat types and improve local explanations.	1D CNN-ResNet; ECANet; K-GradCam; SHAP/LIME/Grad-CAM.	MIT-BIH; 259-sample window; SMOTE.	F1 98.88%; K-GradCam/Grad-CAM Dice 0.82.	Beat-level holdout; single dataset; agreement among XAI methods does not prove clinical fidelity.
Neves et al. (2021)	Evaluate model-agnostic explanations for ECG time series.	k-NN and CNN; LIME, SHAP, and PSI; derivative; shapelets; reader study.	MIT-BIH; inter-patient split; 600-ms beats.	Binary F1 90.3%; LIME Jaccard 0.88 ± 0.27.	Low perceived clinical utility; three readers; rare classes; high explanation cost.
Sathi et al. (2025)	Simultaneously classify arrhythmia, ischemia, and normal cardiac function using ECG fiducial markers.	FL-LSTM; FedAvg; nine fiducials; SHAP and LIME.	166 individuals; MIT-BIH, European ST-T, and FANTASIA.	Accuracy $92.3 \pm 0.7\%$; F1 0.91 ± 0.02; AUC 0.99 ± 0.01.	Simulated institutions; class confounded with dataset; SMOTE; no real clinical validation.
Duranay et al. (2026)	Compare 1D ECG and 2D images and fuse explanations.	Two 17-layer CNNs; 1D/2D Grad-CAM; MM-GradCAM.	Chapman-Shaoxing; 10,588 patients; lead II.	1D: F1 92.99%; 2D: F1 97.38%.	Single dataset; image derived from the same signal; higher cost; one cardiologist.
Alqudah and Alfraihat (2026)	Integrate cardiac electrical and mechanical activity.	CNN-BiLSTM-SE; self-attention; cross-attention; SHAP/Grad-CAM/IG.	ECG+PCG; four public datasets and an in-house cohort of 500 individuals.	Accuracy 97.0%; F1 96.7%; AUC 0.989; $\kappa = 0.82$.	Synthetic pairing; only 500 real pairs; no external multimodal validation.
Studies on metabolic monitoring and the detection of glycemic disturbances based on ECG					
Kulkarni et al. (2023)	Screen for prediabetes and diabetes using ECG beats.	DiaBeats; XGBoost; SMOTE; temperature scaling.	1,262 individuals; 10,461 beats; 12-lead ECG.	Accuracy 96.8%; F1 96.6%; calibration error reduced to 6.32%.	Beat-level split; single specific cohort; no external validation.
Koga et al. (2025)	Detect early glycemic abnormalities and evaluate a single-lead approach.	DiaCardia; LightGBM; Optuna; SHAP; confounder control.	16,766 records + 2,456 external individuals; 12 leads/lead I.	AUC 0.851 internal and 0.785 external; lead I: AUC 0.844.	Japanese population; PPV approximately 21%; not evaluated on a real wearable device.
Cordeiro et al. (2021)	Identify hyperglycemia using compact fiducial features.	Ten-layer DNN; 18 PQRST distances and slopes.	1,119 participants; 51,518 beats; single lead.	AUC 94.53%; sens. 87.57%; spec. 85.04%.	Threshold applied without fasting; participant-wise split not reported; no external validation.
Vyas et al. (2023)	Compare morphology and RR dynamics in hypo-/hyperglycemia.	CNN, CNN-LSTM, RR-LSTM, HRV, and late fusion.	Ten adults with type 1 diabetes; ECG and CGM for up to two weeks.	Fusion: AUC approximately 0.68 for hypo- and 0.70 for hyperglycemia.	Personalized models; small sample; no inter-patient validation.
Islam et al. (2025)	Detect hypoglycemia and explain ECG features.	RF, GBM, and MLP; 74 features; HRV; SHAP; LMS filter.	Four D1NAMO patients; single-lead ECG, accelerometry, and CGM.	RF/GBM: AUC 0.97 and F1 0.86; removing HRV reduced F1.	Five patients excluded; random intra-patient split; no external evaluation.
Dave et al. (2024)	Detect hypo- and hyperglycemia under free-living conditions using multi-threshold fusion.	Personalized Random Forest; morphology+HRV; temporal block validation.	12 adults with type 1 diabetes; 14 days; single-lead ECG and CGM.	AUC 0.75 for hypo- and 0.80 for hyperglycemia; hypoglycemia precision 0.10.	Few individuals; calibration required; many false alerts; no external validation.

contributions, and LIME or perturbation-based methods to generate local explanations. However, few studies jointly evaluate fidelity, stability, and correspondence with clinical knowledge. Visually salient regions localized within the QRS complex, for example, may be physiologically plausible without necessarily demonstrating that the model uses the same criteria applied in clinical interpretation. A more rigorous evaluation may combine deletion or insertion tests involving relevant regions, stability analysis under noise or small temporal shifts, expert agreement, and comparison with previously annotated fiducial points. It is also important to analyze explanations associated with classification errors rather than only correct predictions. For WPW, relevance maps can be compared with the PR interval, QRS onset, and the region corresponding to the delta wave. In glycemic studies, the importance assigned to ECG characteristics should be analyzed while considering potential confounders such as heart rate, physical-activity level, acquisition time, food intake, and medication use.

C. Multimodality, redundancy, and computational cost

The reviewed literature demonstrate that combining one-dimensional temporal signals, temporal features, and time-frequency representations expands the ways ECG information is represented, even though these inputs do not necessarily correspond to independent physiological sources. An image or map derived from the ECG itself preserves much of the original information but may also introduce characteristics related to the transformation or rendering process. Similarly, integrating ECG and phonocardiography represents a more direct form of physiological multimodality; however, it requires synchronized acquisition to preserve the relationships between cardiac electrical and mechanical events. Strategies such as late probability fusion, cross-attention, and adaptive weighting can leverage complementary information from different representations. Nevertheless, the contribution of each modality must be verified through ablation studies, alongside assessments of calibration and computational cost. For wearable device applications, factors such as power consumption, inference time, model complexity, and robustness against artifacts caused by motion or temporary signal loss must also be considered.

D. Research gaps in ventricular pre-excitation

Evidence specifically addressing automated WPW detection remains limited. Merbouti and Cherifi [4] evaluated only 21 WPW segments, whereas ECG-XPLAIM [5], despite external validation, achieved an AUC of 0.895 for WPW with sensitivity of 0.773. Together, these findings show both insufficient condition-specific representation and persistent difficulty detecting rare WPW cases under independent validation. Future studies should therefore use larger multicenter cohorts containing WPW and morphologically similar abnormalities, together with annotations of wave boundaries and cardiac intervals. Integrated models may combine robust R-peak detection, P-wave and QRS delineation, temporal signals, morphological and spectral features, and time-frequency representations. Their contribution should be verified through ablation studies, with patient-independent and, whenever possible, cross-database validation. Positive predictive value and calibration should complement sensitivity, specificity, and AUC.

E. Research gaps in glycemic detection

Glycemic-abnormality studies reveal a clear distinction between population-level and personalized models. Personalized approaches may capture subject-specific morphological and autonomic responses but require prior ECG synchronized with a reference glucose measurement, whereas population models must handle greater inter-participant variability and confounders such as age, medication, physical activity, food intake, and acquisition device. Future evaluations should prioritize prolonged ambulatory ECG-CGM recordings and explicitly account for temporal synchronization and possible physiological delays between glucose changes and cardiovascular manifestations. For alert-oriented systems, AUC alone is insufficient. Positive predictive value, false alerts per unit time, detection delay, and probability calibration are required to characterize real-world behavior. This is illustrated by Dave et al. [9], where hypoglycemia discrimination was accompanied by low precision and frequent false alerts. Overall, the evidence supports ECG as a complementary source of information associated with glycemic variation rather than as a replacement for reference glucose measurement, and favors participant-wise and external validation when population-level generalization is intended.

V. Conclusion

This study presented a scoping review based on a previously selected corpus and did not constitute a prospectively registered systematic review. No formal risk-of-bias tool was applied, and exclusion reasons were not quantified by category at all screening stages. In addition, heterogeneity in tasks, populations, datasets, and performance metrics precluded statistical synthesis. Therefore, the analysis emphasized participant independence, external validation, sample representativeness, potential data leakage, and explainability rather than ranking studies solely by reported performance.

The reviewed literature highlights the potential of ECG as a physiological signal for computational applications ranging from arrhythmia detection to the analysis of temporal, morphological, fiducial, and time-frequency features associated with ventricular pre-excitation and glycemic alterations. Regarding interpretability, current evidence remains largely qualitative, highlighting the need for quantitative assessment of explanation fidelity, stability, and clinical consistency, including the analysis of classification

errors. Multimodal and multi-representation approaches may improve discrimination by combining complementary ECG information; however, their benefit should be confirmed through ablation studies, calibration analysis, and assessment of computational cost.

Regarding ventricular pre-excitation associated with Wolff-Parkinson-White (WPW) syndrome, evidence remains limited by small sample sizes, a scarcity of databases with specific labels for the condition, and performance degradation observed during external validation. Future studies should prioritize whenever possible multicenter, publicly available datasets or, alternatively, controlled data collection protocols involving participants with a confirmed WPW diagnosis. Rigorous patient-level data separation, the exploration of complementary ECG representations, and quantitative assessment of model explainability are also recommended.

Regarding glycemic changes, the ECG also shows potential as a complementary physiological marker, although model performance is strongly influenced by data partitioning strategies, population characteristics, and validation protocols. Future studies should adopt controlled acquisition protocols, account for potential confounding factors, and conduct external validation, while ensuring synchronization between ECG signals and reference glucose measurements. Furthermore, metrics of operational relevance such as positive predictive value and the frequency of false alerts should be reported.

Overall, progress in both areas depends on evaluation protocols that ensure participant independence, test for generalizability across external cohorts, and rigorously assess the reliability and clinical consistency of model explanations. Finally, the presented results should be interpreted as a critical synthesis derived from a scoping review, rather than as evidence from a prospectively registered systematic review.

VI. Acknowledgments

This article, developed as part of the Glucose Project, was made possible by Samsung Eletrônica da Amazônia Ltda., with support from SUFRAMA and implementation by the Federal University of Amazonas, a beneficiary of funds originating from Federal Law No. 8,387/1991, with dissemination and publicity in accordance with the provisions of Article 39 of Decree No. 10,521/2020.

In accordance with IEEE policy on the use of AI-generated content, the authors disclose that ChatGPT PRO (OpenAI) was used as a support tool for translation into English, textual organization, and grammar revision across all sections of this manuscript. The tool was not used to generate scientific content, experimental results, figures, or code. All intellectual decisions and full responsibility for the final content remain solely with the authors.

References

[1] A. S. Canto *et al.*, "Cardiac arrhythmia detection based on R-peak centered segments of ECG signals using 1D convolutional neural networks and explanation using Grad-CAM tool," in *Computing in Cardiology*, vol. 52, 2025, doi: 10.22489/CinC.2025.063.

[2] C. Wang, C. Pan, and C. Meng, "Robust R-peak detection in noisy ECG using deep residual U-Net for enhanced cardiac rhythm analysis," *Medicine*, vol. 104, no. 40, Art. no. e44968, 2025, doi: 10.1097/MD.0000000000044968.

[3] A. J. Prakash *et al.*, "A new approach of transparent and explainable artificial intelligence technique for patient-specific ECG beat classification," *IEEE Sensors Letters*, vol. 7, no. 5, Art. no. 5501604, 2023, doi: 10.1109/LSENS.2023.3268677.

[4] M. A. Merbouti and D. Cherifi, "Machine learning based electrocardiogram peaks analyzer for Wolff-Parkinson-White syndrome," *Biomedical Signal Processing and Control*, vol. 86, Art. no. 105302, 2023, doi: 10.1016/j.bspc.2023.105302.

[5] P. Pantelidis *et al.*, "ECG-XPLAIM: eXPlainable locally-adaptive artificial intelligence model for arrhythmia detection from large-scale electrocardiogram data," *Frontiers in Cardiovascular Medicine*, vol. 12, Art. no. 1659971, 2025, doi: 10.3389/fcvm.2025.1659971.

[6] A. R. Kulkarni *et al.*, "Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogram," *BMJ Innovations*, vol. 9, pp. 32-42, 2023, doi: 10.1136/bmjinnov-2021-000759.

[7] D. Koga *et al.*, "Artificial intelligence identifies individuals with prediabetes using single-lead electrocardiograms," *Cardiovascular Diabetology*, vol. 24, Art. no. 415, 2025, doi: 10.1186/s12933-025-02982-4.

[8] K. Vyas *et al.*, "Detection of glycemic excursions using morphological and time-domain ECG features," in *Proc. 2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 2023, pp. 1-4, doi: 10.1109/BSN58485.2023.10331278.

[9] D. Dave, K. Vyas, G. L. Cote, and M. Erraguntla, "Hypoglycemia and hyperglycemia detection using ECG: A multi-threshold based personalized fusion model," *Biomedical Signal Processing and Control*, vol. 96, Art. no. 106569, 2024, doi: 10.1016/j.bspc.2024.106569.

[10] A. C. Tricco et al., "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018, doi: 10.7326/M18-0850.

[11] University of Exeter Library, "Scoping reviews: Creating a search strategy," *University of Exeter LibGuides*, Aug. 6, 2026. [Online]. Available: https://libguides.exeter.ac.uk/c.php?g=702858&p=5054831. [Accessed: Aug. 22, 2026].

[12] P. Singh and A. Sharma, "Interpretation and classification of arrhythmia using deep convolutional network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, Art. no. 2518512, 2022, doi: 10.1109/TIM.2022.3204316.

[13] I. Neves *et al.*, "Interpretable heartbeat classification using local model-agnostic explanations on ECGs," *Computers in Biology and Medicine*, vol. 133, Art. no. 104393, 2021, doi: 10.1016/j.compbiomed.2021.104393.

[14] T. A. Sathi *et al.*, "Explainable federated learning for multi-class heart disease diagnosis via ECG fiducial features," *Diagnostics*, vol. 15, Art. no. 3110, 2025, doi: 10.3390/diagnostics15243110.

[15] F. M. Duranay *et al.*, "MM-GradCAM: An improved multimodal GradCAM method with 1D and 2D ECG data for detection of cardiac arrhythmia," *Scientific Reports*, vol. 16, Art. no. 7919, 2026, doi: 10.1038/s41598-026-38654-w.

[16] A. M. Alqudah and A. Alfraihat, "A multimodal approach for cardiac signals classification using deep learning with explainable AI methods," *Health Information Science and Systems*, vol. 14, Art. no. 5, 2026, doi: 10.1007/s13755-025-00402-1.

[17] R. Cordeiro, N. Karimian, and Y. Park, "Hyperglycemia identification using ECG in deep learning era," *Sensors*, vol. 21, Art. no. 6263, 2021, doi: 10.3390/s21186263.

[18] M. R. Islam, S. K. Islam, and M. M. Hossain Shuvo, "Explainable AI for hypoglycemia detection in type 1 diabetes using single-lead ECG signals," in *Proc. 2025 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2025, pp. 1-6, doi: 10.1109/MEMEA65319.2025.11067978.

Desenvolvimento de uma Plataforma de Baixo Custo para Varredura 2D e Caracterização de Campos Acústicos Ultrassônicos

Ana Laura Waideman de Oliveira
Departamento Acadêmico de Eletrotécnica (DAELT)
Universidade Tecnológica Federal – Paraná (UTFPR)
Curitiba, Brasil
analauraoliveira@alunos.utfpr.edu.br

Ana Clara Annunciato de Oliveira
Departamento Acadêmico de Eletrotécnica (DAELT)
Universidade Tecnológica Federal – Paraná (UTFPR)
Curitiba, Brasil
ana.180106@alunos.utfpr.edu.br

Alexia Marcon Watzlawick
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI-CT)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
awatzlawick@alunos.utfpr.edu.br

Gilson Maekawa Kanashiro
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI-CT)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
gilson.kanashiro@ifpr.edu.br

Joaquim Miguel Maia
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI-CT), DAELN
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
joaquim@utfpr.edu.br

Amauri Amorin Assef
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI-CT), DAELT
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
amauriassef@utfpr.edu.br

***Abstract*—This work presents the development and preliminary performance evaluation of a low-cost platform for ultrasonic transducer characterization and two-dimensional acoustic field mapping. The proposed system integrates a microcontroller-based positioning mechanism, a LabVIEW supervisory interface, and MATLAB routines for signal processing and spectral analysis, supporting manual positioning and autonomous scanning. Experimental evaluation was performed using two single-element ultrasonic transducers with nominal frequencies of 1.6 MHz and 5 MHz, together with a needle hydrophone. The agreement between signals acquired by a digital oscilloscope and those exported through the LabVIEW interface was assessed using the Pearson correlation coefficient (PCC) and normalized root mean square error (NRMSE). Both transducers exhibited high correlation between the acquired and exported signals, while their spectral characteristics were determined from the experimental frequency responses using the -6 dB bandwidth criterion. The autonomous scanning mode was demonstrated over a 20×20 mm region with a 20×20 acquisition grid, enabling the reconstruction of the two-dimensional acoustic response and the estimation of the peak negative pressure. The results demonstrate the feasibility of the proposed low-cost platform for laboratory-based ultrasonic transducer characterization and acoustic field mapping.**

***Index Terms*—ultrasound, acoustic field characterization, hydrophone measurements, ultrasonic transducers, low-cost instrumentation**

I. Introdução

A caracterização espacial e a quantificação dos campos de pressão acústica emitidos por transdutores ultrassônicos são etapas fundamentais para o desenvolvimento, calibração e validação de tecnologias médicas, ensaios não destrutivos e sistemas de instrumentação acústica [1]–[3]. A exatidão e a repetibilidade dessas medições dependem do posicionamento preciso de um sensor de referência, tipicamente um hidrofone, em relação ao emissor acústico [3], [4].

Para atender a esses requisitos, investigações laboratoriais geralmente empregam tanques de água equipados com sistemas computadorizados de varredura tridimensional (3D) [5], [6]. Embora apresentem elevada precisão, essas plataformas são frequentemente limitadas pelo alto custo e pela dependência de soluções proprietárias, o que dificulta a adaptação de software e a integração de novos periféricos. Nesse contexto, plataformas baseadas em microcontroladores e componentes eletromecânicos de baixo custo têm sido propostas como alternativas viáveis [7]. Clinard *et al.* (2022), por exemplo, desenvolveram um tanque de varredura modular baseado em Arduino com desempenho comparável ao de sistemas comerciais [8]. Entretanto, soluções baseadas exclusivamente no controle remoto por computador ainda apresentam limitações nas etapas de calibração e de alinhamento inicial do hidrofone, que requerem ajustes locais e interação em tempo real.

Este trabalho apresenta o desenvolvimento de um sistema de varredura bidimensional (2D) automatizado e de baixo custo, baseado em uma arquitetura de controle híbrida e em equipamentos amplamente disponíveis em laboratórios de instrumentação. O sistema integra uma estrutura mecânica de translação no plano x-z, uma placa microcontroladora, um módulo de acionamento de motores com resolução de micropasso e uma interface supervisória no LabVIEW. A arquitetura proposta combina um modo manual para calibração e alinhamento do hidrofone com um modo autônomo para execução automatizada das rotinas de varredura e aquisição de sinais ultrassônicos.

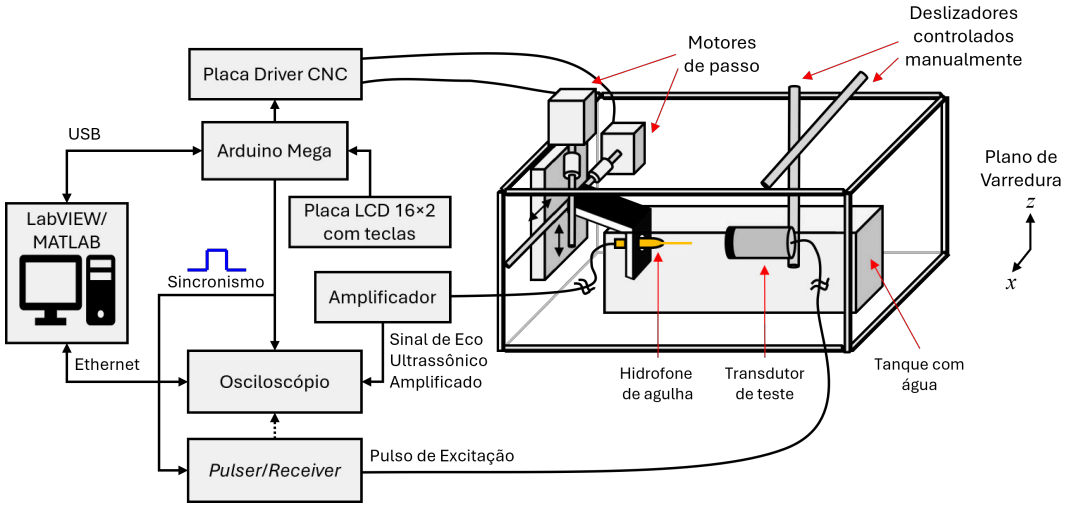


Figura 1. Diagrama de blocos funcional do sistema de varredura: integração de hardware, periféricos e fluxo de comunicação entre o Arduino Mega e o LabVIEW.

II. Materiais e Métodos

Nesta seção, é descrito o sistema de baixo custo para varredura 2D, desenvolvido com foco em aplicações em ambientes laboratoriais de ensino e prototipagem rápida.

A. Arquitetura do Sistema

O sistema proposto para aquisição de sinais ultrassônicos e mapeamento do campo acústico, ilustrado no diagrama em blocos da Fig. 1, é composto por cinco módulos principais: (i) circuito de controle e sincronismo; (ii) módulo de excitação acústica; (iii) módulo de aquisição e digitalização de sinais; (iv) módulo de armazenamento de dados; e (v) plataforma de translação 2D para o posicionamento do hidrofone.

O circuito de controle e sincronismo é constituído por uma placa Arduino Mega 2560 e uma placa CNC Shield V3, mostradas na Fig. 2, além de uma placa Display LCD Keypad Shield 16×2, utilizada para o acionamento local ou automatizado do sistema de movimentação. Esses módulos foram montados em um gabinete com dimensões de 12 × 6 × 4 cm, conforme ilustrado na Fig. 3. Além de controlar dois motores de passo para o alinhamento preciso do hidrofone e do transdutor ao longo do eixo acústico no plano x-z, o microcontrolador gera o sinal de sincronismo, utilizado pelo estágio de excitação do transdutor.

O sistema de atuação mecânica 2D responsável pelo deslocamento do hidrofone no tanque de ensaios acústicos é composto por dois motores de passo NEMA 17, com resolução nativa de 1,8° por passo, o que corresponde a 200 passos por revolução. O acionamento dos motores é realizado por dois módulos DRV8825 instalados na parte superior da CNC Shield V3. Os *drivers* DRV8825 foram configurados para operação em regime de micropasso de 1/32, o que resulta em 6.400 micropassos por revolução do motor NEMA 17 (200 passos por volta). Considerando o fuso roscado TR8×2 com quatro entradas, correspondente a um avanço linear (*lead*) de 8 mm por revolução, o sistema apresenta uma resolução de

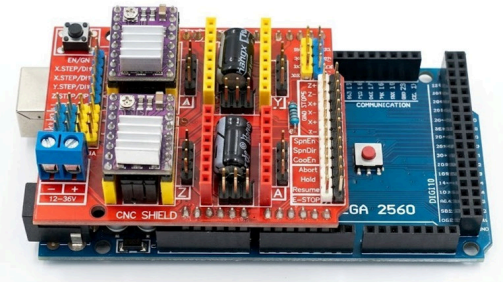


Figura 2. Placa Arduino Mega 2560 conectada à placa LCD Keypad Shield 16×2.



Figura 3. Gabinete do sistema de controle e sincronismo, contendo o Arduino Mega 2560, a placa CNC Shield V3 e o Display LCD Keypad Shield 16×2.

posicionamento de 800 passos por milímetro, equivalente a um deslocamento linear teórico de 1,25 μm por micropasso.

O modo de operação padrão é o manual, no qual o usuário pode selecionar e visualizar os sinais adquiridos pelo hidrofone e, opcionalmente, pelo transdutor. No modo autônomo, a interface envia ao Arduino as coordenadas de posicionamento do hidrofone com base nos parâmetros definidos pelo usuário, permitindo a execução da varredura espacial no plano x–z para a obtenção do mapa de pressão acústica. A etapa de excitação acústica é realizada por um equipamento *pulser/receiver* (P/R)

Olympus 5077PR, responsável pela geração de pulsos elétricos unipolares de alta tensão para excitar o transdutor ultrassônico.

Após o alinhamento inicial no modo manual, um osciloscópio digital Tektronix MDO3014 é utilizado para adquirir o sinal captado pelo hidrofone, previamente condicionado por um estágio externo de amplificação. Opcionalmente, outro canal do osciloscópio pode ser utilizado para visualizar o sinal de eco do transdutor. O osciloscópio realiza a digitalização das formas de onda com taxa de amostragem de 100 MSPS (escala horizontal ajustada para 10 µs/div), fornecendo a resolução temporal necessária para a visualização dos ecos, bem como para a caracterização e análise do campo acústico. Os dados digitalizados são transferidos para uma interface desenvolvida na plataforma LabVIEW versão 2026 Q1, que estabelece comunicação via Ethernet com o osciloscópio. Essa interface permite o armazenamento automático dos dados para posterior processamento e análise.

O sistema de segurança mecânico de calibração é composto por três chaves fim de curso (duas no eixo x e uma no eixo z), conectadas diretamente às entradas do microcontrolador. Esses sensores atuam como mecanismo de proteção contra colisões e delimitam o volume útil de translação do hidrofone, enquanto o transdutor permanece fixo durante as operações de varredura.

A alimentação do Arduino, com tensão de 5 V, é fornecida pela interface USB do computador hospedeiro, enquanto a alimentação do sistema de translação é fornecida por uma fonte chaveada externa de 12 V/20 A, conectada ao barramento de potência do CNC Shield V3. Conforme discutido em [8], o isolamento entre a alimentação de potência e a eletrônica de controle é fundamental para mitigar interferências eletromagnéticas nos sinais adquiridos pelo hidrofone no tanque.

B. Estrutura Mecânica

A plataforma mecânica é composta por uma estrutura construída com perfis comerciais de alumínio V-slot de seção transversal 20 × 20 mm, com dimensões totais de 385 × 240 × 240 mm (C × L × A). Essa estrutura serve de base para os sistemas dos eixos horizontal e vertical, responsáveis pelo deslocamento do hidrofone submerso em água nas respectivas direções. A plataforma mecânica é apresentada na Fig. 4.

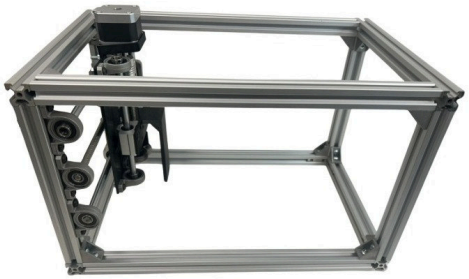


Figura 4. Visão lateral da estrutura mecânica com os sistemas dos eixos horizontal e vertical, além do suporte do hidrofone, já montados.

Cada eixo é composto pelos seguintes elementos: (i) dois eixos lineares retificados de 8 mm; (ii) um fuso roscado TR8×2 com quatro entradas; (iii) cinco mancais de apoio para eixos de 8 mm; (iv) dois *pillow blocks* para eixos de 8 mm; (v) um bloco de porca em POM para fuso TR8; (vi) um motor de passo; (vii) um suporte impresso em 3D para o motor; (viii) um acoplador para eixos de 8/6,35 mm; e (ix) elementos de fixação, como parafusos e porcas.

O motor de passo é acoplado ao fuso roscado por meio do acoplador, sendo responsável, em conjunto com o bloco de POM, pela conversão do movimento rotacional em deslocamento linear. Os *pillow blocks* são montados sobre os eixos retificados, um de cada lado do fuso, proporcionando estabilidade mecânica e mantendo o alinhamento do conjunto durante a movimentação. Esse alinhamento reduz os esforços excessivos sobre o motor de passo e minimiza o risco de travamentos. Os mancais fornecem suporte aos eixos retificados e à extremidade livre do fuso, oposta ao motor. O sistema do eixo vertical ainda possui um componente adicional destinado à fixação do hidrofone. A montagem desse sistema é apresentada em maior detalhe na Fig. 5.

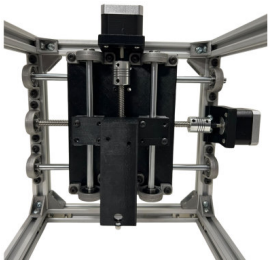


Figura 5. Montagem do eixo vertical fixado ao carro móvel do eixo horizontal.

O sistema do eixo horizontal é fixado diretamente à estrutura principal, enquanto o sistema do eixo vertical é montado sobre o carro móvel do eixo horizontal. Dessa forma, o deslocamento horizontal desloca simultaneamente todo o conjunto vertical, ao passo que o deslocamento vertical atua exclusivamente sobre o hidrofone.

Para garantir a reprodutibilidade do projeto, a lista detalhada de componentes e custos, juntamente com o código-fonte do *firmware* e os arquivos da interface supervisória, está disponível publicamente no repositório do projeto no GitHub: https://github.com/anannunciato/2D-Ultrasonic-Transducer-Characterization-System

C. Implementação do Firmware de Controle e Sincronismo

Integrado ao ambiente supervisório, o *firmware* de controle foi desenvolvido em C/C++ utilizando o Arduino IDE (versão 2.3.8). A função do Arduino no sistema é atuar como um controlador periférico em tempo real, operando de forma subordinada aos comandos locais do usuário no modo manual, realizados por meio do LCD Keypad Shield 16×2, e aos comandos remotos provenientes da interface LabVIEW no modo autônomo, conforme o fluxograma da Fig. 6.

O algoritmo implementado é responsável pela geração dos pulsos de controle dos motores de passo e do sinal

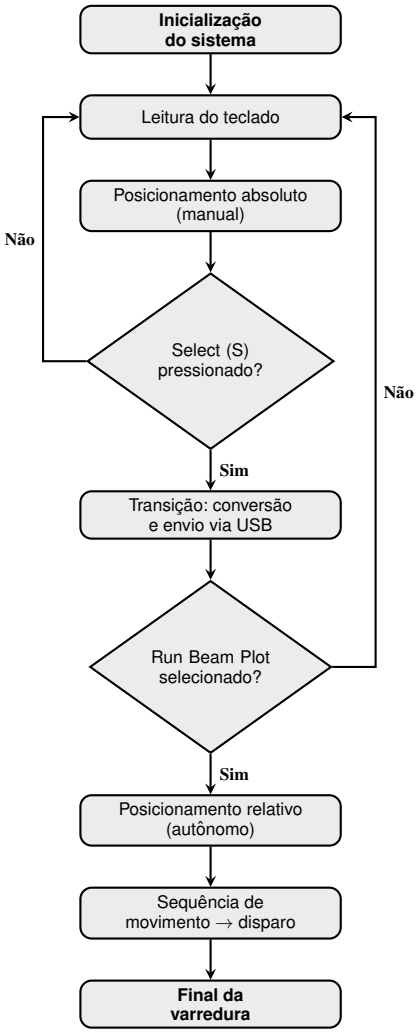


Figura 6. Fluxograma simplificado do *firmware* embarcado, mostrando a transição entre a rotina de posicionamento local e a varredura autônoma coordenada.

de sincronismo do estágio de excitação do transdutor. Para reduzir o risco de perda de passos devido à inércia ou perturbações mecânicas, o código utiliza a biblioteca *AccelStepper*, configurada para implementar perfis de aceleração e desaceleração com execução não bloqueante. Adicionalmente, a biblioteca *LiquidCrystal* é empregada para o gerenciamento da interface e atualização do LCD.

A estrutura do código é dividida em dois estágios operacionais distintos. O primeiro estágio compreende a definição do referencial absoluto de posicionamento, que estabelece a correspondência física, em milímetros (mm), dos eixos x e z no interior do tanque de ensaios acústicos, servindo de base para o posicionamento inicial do hidrofone. O operador interage localmente por meio das teclas integradas ao LCD Keypad Shield 16×2, realizando ajustes incrementais nas coordenadas espaciais por meio das teclas de incremento (▲) e decremento (▼). Os valores são exibidos no visor do LCD com precisão de três casas decimais (000.000 mm). A seleção da casa decimal é realizada por meio das teclas esquerda (◄) e direita (►).

Nessa etapa de movimentação manual, a temporização do sistema é implementada com base na função nativa `millis()`, o que permite a execução não bloqueante do programa. Esse recurso atua como um temporizador interno, possibilitando o controle incremental dos motores de passo, enquanto mantém simultâneas a atualização do LCD e a leitura das teclas. Em paralelo à movimentação dos motores, no modo manual são gerados pulsos de sincronismo externo para o P/R, com taxa de repetição de pulsos (PRF) padrão de 100 Hz. Uma vez definido o ponto inicial de operação (x_0, z_0), o usuário aciona a tecla `SELECT` (*S*) do LCD Keypad Shield 16×2, comando que converte as coordenadas de unidades métricas para passos inteiros equivalentes e envia esses dados ao LabVIEW via USB. Após o recebimento das posições iniciais x_0 e z_0, a interface LabVIEW permite a definição da área de varredura, possibilitando o ajuste da resolução espacial (passos por milímetro), dos limites de deslocamento em cada eixo e da densidade da malha de aquisição para a operação no modo autônomo – valor inicialmente ajustado para 800.

Após o acionamento do botão `Run Beam Plot` na interface supervisória desenvolvida em LabVIEW, o sistema inicia o modo autônomo de varredura, gerando inicialmente uma matriz contendo N pontos de varredura, compreendidos entre as coordenadas (x_0, z_0) e (x_{N-1}, z_{N-1}). Em seguida, as coordenadas absolutas de cada ponto são transmitidas sequencialmente ao microcontrolador, que executa os deslocamentos correspondentes. Ao atingir cada ponto da matriz, o Arduino gera um pulso de disparo para o P/R, garantindo a aquisição sincronizada dos dados acústicos. O padrão de varredura implementado neste trabalho segue uma trajetória em ziguezague, como ilustrado na Fig. 7. Paralelamente às rotinas de movimentação manual e automática, o sistema monitora continuamente o circuito de segurança composto por chaves de fim de curso.

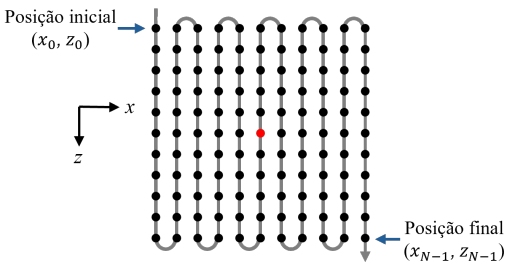


Figura 7. Padrão de varredura ultrassônica no plano x-z com a indicação dos N pontos de varredura. O ponto central da região de interesse é indicado em vermelho.

D. Sistema Supervisório Baseado em LabVIEW

A Fig. 8 apresenta a interface gráfica de usuário desenvolvida para o sistema supervisório baseado em LabVIEW. Nesse exemplo, são apresentados o sinal de eco ultrassônico bruto e seu espectro em frequência, obtido por meio da Transformada Rápida de Fourier (FFT). As principais características da interface para configuração do sistema de aquisição e varredura são descritas na Tabela I.

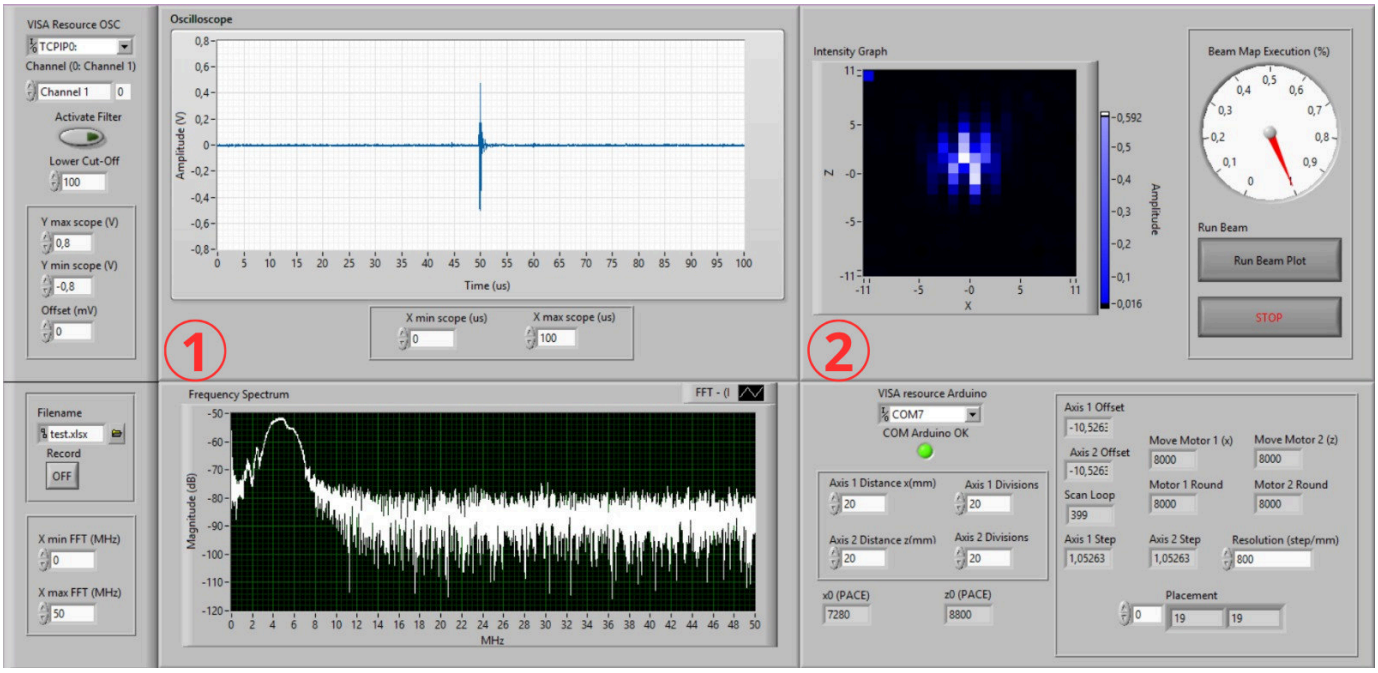


Figura 8. Interface desenvolvida no LabVIEW para visualização e armazenamento das formas de onda ultrassônicas.

Tabela I
PRINCIPAIS CARACTERÍSTICAS DA INTERFACE GRÁFICA DE USUÁRIO DESENVOLVIDA

Módulo	**Funcionalidade**
Aquisição	Visualização da forma de onda e da respectiva FFT Ajuste das escalas de tensão [V], *offset* [V] e tempo [μs] Habilitação do filtro passa-baixa (FPB) Ajuste da faixa de frequência da FFT [MHz] Salvamento do mapa de tensão negativa de pico
Mapa	Ajuste do número de divisões nos eixos x e z Ajuste das dimensões de varredura nos eixos x e z Salvamento do mapa da tensão negativa de pico

O ciclo de aquisição de dados inicia-se com a configuração da interface de comunicação com o instrumento de medição por meio do campo (i) *VISA Resource Osc* (ilustrado operando via protocolo TCP/IP), seguida da seleção do canal analógico do osciloscópio (CH1, CH2, CH3 ou CH4) por meio do controle (ii) *Channel*. Para garantir a integridade dos dados e atenuar componentes de baixa frequência indesejadas, o sistema dispõe de um estágio de condicionamento digital. O operador pode habilitar essa função por meio do botão (iii) *Activate Filter*, definindo a frequência de corte do filtro passa-alta no campo (iv) *Lower Cut-Off*.

Para facilitar a descrição das funcionalidades da interface, a tela apresentada na Fig. 8 foi subdividida em dois painéis. O primeiro painel reúne os recursos destinados ao monitoramento do sinal, realizado simultaneamente nos domínios do tempo e da frequência. No domínio do tempo, a forma de onda adquirida é exibida no gráfico superior, denominado (v) *Oscilloscope*. A janela de visualização pode ser ajustada dinamicamente pelo usuário por meio da configuração dos limites de amplitude nos campos (vi) *Y max scope (V)* e *Y min scope (V)*, bem como da inserção de um *Offset DC*. A base de tempo da aquisição é definida pelos campos (vii) *X min scope (μs)* e *X max scope (μs)*. Em seguida, o software aplica a FFT ao sinal e apresenta seu espectro no domínio da frequência no gráfico inferior (viii) *Frequency Spectrum*. Essa ferramenta permite a análise espectral e a visualização das componentes harmônicas do sinal, sendo a faixa de frequência de interesse ajustável por meio dos limites (ix) *X min FFT (MHz)* e *X max FFT (MHz)*. Por fim, o sistema permite a exportação do sinal adquirido. O usuário define o diretório e o nome do arquivo de saída (no formato de planilha) por meio do campo (x) *Filename* (por exemplo, *teste.xlsx*). A gravação dos dados para posterior análise *offline* é iniciada por meio do botão (xi) *Record*.

No segundo painel, o ciclo operacional inicia-se com a configuração da porta serial no campo (i) *VISA Resource Arduino*, cuja comunicação é confirmada pelo indicador luminoso (ii) *COM Arduino OK*. Inicialmente, o sistema opera no modo manual para o ajuste da posição do hidrofone. Ao pressionar a tecla *S* no LCD Keypad *Shield* 16×2, as coordenadas são transmitidas via USB e preenchidas nos campos (iii) *X inicial (PACE)* e *Z inicial (PACE)*, sendo então assumidas como a origem (x_0, z_0) da malha de varredura automatizada. Com a comunicação estabelecida, os limites geométricos do experimento são definidos pelo usuário por meio dos campos de configuração da área de varredura: (iv) *Axis 1 Distance x (mm)* e *Axis 2 Distance z (mm)*, que especificam as distâncias máximas de translação do hidrofone nos eixos x e z, respectivamente; (v) *Axis 1 Divisions* e

Axis 2 Divisions, que determinam o nível de discretização da malha ao subdividir a área de varredura em intervalos de aquisição; e (vi) *Resolution (steps/mm)*, no qual o usuário informa a resolução de deslocamento, utilizada para calcular a quantidade equivalente de passos dos motores necessária para percorrer a malha.

Após a configuração de todos os parâmetros, o operador aciona o botão (vii) *Run Beam Plot*, alterando o estado do sistema para o modo autônomo de varredura. A partir desse instante, o software inicia a execução da malha de varredura automatizada, enquanto o indicador *Beam Map Execution* exibe seu progresso em percentual.

E. Demonstração Experimental do Sistema de Aquisição

Para a avaliação do funcionamento do sistema de aquisição desenvolvido, foram utilizados dois transdutores ultrassônicos monoelemento, ambos da FUNBEC, com frequências nominais de 1,6 MHz e 5 MHz e diâmetros de 19 mm e 6,35 mm, respectivamente, além de um hidrofone do tipo agulha com diâmetro de 1 mm (modelo 1730, Precision Acoustics). O hidrofone apresenta sensibilidade típica de 991,4 mV/MPa quando utilizado em conjunto com o pré-amplificador submersível PA11029 (Precision Acoustics), cuja saída foi conectada ao sistema de aquisição. O sinal proveniente do hidrofone foi condicionado pelo pré-amplificador e pelo DC Coupler (modelo DCPS180, Precision Acoustics), responsável pela alimentação do pré-amplificador e pelo acoplamento do sinal ao sistema de aquisição. Os transdutores ultrassônicos são apresentados na Fig. 9, enquanto o hidrofone de agulha com o pré-amplificador é apresentado na Fig. 10.



(a) **(b)**

Figura 9. Transdutores ultrassônicos de 1,6 MHz e 5 MHz utilizados nos ensaios: (a) vista lateral e (b) vista frontal.



Figura 10. Hidrofone do tipo agulha de 1 mm de diâmetro, modelo 1730 (Precision Acoustics), utilizado nos ensaios experimentais.

O desempenho do sistema foi inicialmente avaliado por meio da comparação entre a resposta ultrassônica captada pelo hidrofone, adquirida pelo osciloscópio, e o sinal correspondente exportado pelo LabVIEW e processado no MATLAB R2024a. Foram avaliados os parâmetros espectrais do transdutor, bem como a similaridade entre os sinais. A frequência de pico (f_p) foi determinada experimentalmente a partir do espectro de frequência, correspondendo à frequência associada ao valor máximo do espectro. A frequência central (f_c) foi calculada como a média aritmética das frequências de corte inferior (f_L) e superior (f_H), ambas determinadas em -6 dB. As larguras de banda absoluta (BWabs) e relativa (BWrel) foram calculadas pelas Equações (1) e (2), respectivamente.

$$\mathrm{BW}_{\mathrm{abs}} = f_H - f_L \tag{1}$$

$$\mathrm{BW}_{\mathrm{rel}} = \frac{\mathrm{BW}_{\mathrm{abs}}}{f_c} \times 100\% \tag{2}$$

O coeficiente de correlação de Pearson (PCC) foi utilizado para quantificar a similaridade entre os sinais, enquanto o erro quadrático médio normalizado (NRMSE) foi empregado para mensurar o erro entre eles [9], [10]. Valores de NRMSE próximos de 0 (%) e de PCC próximos de 1 (adimensional) indicam elevada concordância entre os resultados. As métricas NRMSE e PCC são definidas, respectivamente, como:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}, \tag{3}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2}}{x_{\max} - x_{\min}} \times 100, \tag{4}$$

em que x_i e y_i representam as amostras dos sinais comparados, $x_{\max}$ e $x_{\min}$ correspondem aos valores máximo e mínimo do sinal de referência, $\bar{x}$ e $\bar{y}$ representam as médias dos respectivos sinais, e N é o número total de amostras.

F. Avaliação Preliminar de Desempenho do Sistema de Varredura

Com o objetivo de avaliar o desempenho e a repetibilidade do sistema de varredura, foram realizadas cinco varreduras mecânicas independentes em uma região de 20×20 mm, utilizando uma malha de 20×20 posições nos eixos x e z. Em cada posição, foi registrada a tensão correspondente ao pico negativo da resposta ultrassônica medida pelo hidrofone, permitindo a construção de um mapa 2D da distribuição espacial. Essa grandeza pode ser posteriormente convertida em pressão por meio da sensibilidade calibrada do hidrofone, possibilitando a determinação da pressão negativa de pico (*Peak Negative Pressure* – PNP) [1], [5].

A repetibilidade espacial do sistema foi avaliada por meio da comparação entre os mapas 2D de pressão acústica negativa de pico obtidos em varreduras sucessivas. Para esse ensaio, o hidrofone foi inicialmente posicionado na região de máxima amplitude acústica gerada pelo transdutor de 5 MHz, a uma

distância de aproximadamente 1 cm de sua superfície. O mesmo transdutor foi então submetido a cinco varreduras independentes, sendo realizado o procedimento de *homing* dos eixos x e z antes do início de cada aquisição. Essa abordagem permite avaliar a consistência do posicionamento do sistema a partir de uma característica espacial do próprio campo acústico, considerando conjuntamente o sistema de translação, o posicionamento do hidrofone e o procedimento de aquisição.

III. Resultados

A Fig. 11 ilustra o arranjo experimental utilizado na demonstração de funcionamento do sistema e na caracterização do transdutor. O arranjo é composto pela estrutura mecânica desenvolvida, um recipiente plástico contendo água e o hidrofone com pré-amplificador, posicionado a 5 mm da face do transdutor. A porção inferior da figura apresenta uma ampliação da região interna do recipiente, destacando o posicionamento relativo entre o hidrofone e o transdutor. O custo total dos componentes para a construção da plataforma de varredura 2D foi de aproximadamente R$ 2.200,00 (USD 425,00, cotação de agosto de 2026), desconsiderando o transdutor ultrassônico, o hidrofone e os equipamentos de instrumentação. A discriminação detalhada dos itens e seus respectivos valores está disponível no repositório do projeto no GitHub, indicado na Seção II-B.

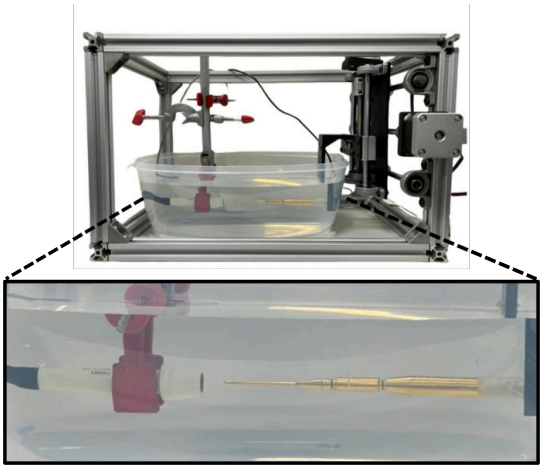


Figura 11. Arranjo experimental utilizado na avaliação do sistema e na caracterização do transdutor. A ampliação na porção inferior da figura destaca o posicionamento do hidrofone a 5 mm da face do transdutor ultrassônico.

A comparação entre os sinais adquiridos pelo osciloscópio e os sinais correspondentes exportados pelo LabVIEW é apresentada na Fig. 12. Para o processamento no MATLAB, os sinais exportados pelo LabVIEW foram submetidos à janela de Hanning antes do cálculo da FFT. As Fig. 12(a) e 12(b) apresentam, respectivamente, os sinais no domínio do tempo e os respectivos espectros de frequência para o transdutor de 1,6 MHz. Observa-se elevada concordância entre os sinais no domínio do tempo, bem como entre os respectivos espectros de frequência, indicando que o procedimento de aquisição e exportação preserva adequadamente as características do sinal ultrassônico. De forma semelhante, as Fig. 12(c) e 12(d) apresentam os resultados no domínio do tempo e da frequência, respectivamente, para o transdutor de 5 MHz, também evidenciando boa concordância entre os sinais obtidos pelo osciloscópio e aqueles exportados pelo LabVIEW.

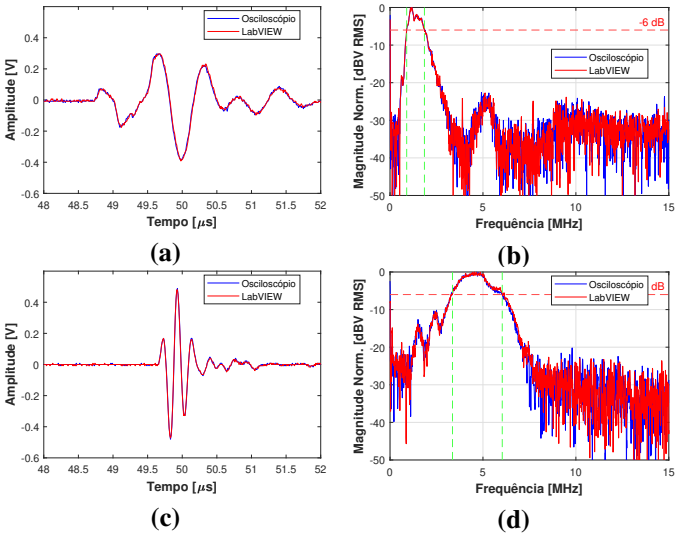


Figura 12. Comparação entre os sinais ultrassônicos adquiridos pelo osciloscópio e os sinais exportados pelo LabVIEW. (a) e (b) Sinais no domínio do tempo e respectivos espectros em frequência para o transdutor de 1,6 MHz, respectivamente. (c) e (d) Sinais no domínio do tempo e respectivos espectros em frequência para o transdutor de 5 MHz, respectivamente.

A Tabela II reúne os resultados quantitativos obtidos nos ensaios experimentais, incluindo as métricas de comparação entre os sinais nos domínios do tempo e os parâmetros espectrais determinados em -6 dB para os dois transdutores apresentados na Fig. 12. Os elevados valores de PCC indicam forte correlação e preservação das características temporais dos sinais. Os valores de NRMSE também indicam boa concordância, sendo influenciados pelas diferenças de amplitude, pois os sinais não foram normalizados.

A Fig. 13 apresenta o mapa 2D da pressão acústica negativa de pico, obtido no modo autônomo e reconstruído por interpolação cúbica com fator de 5. A tensão medida pelo hidrofone foi convertida em pressão utilizando sua sensibilidade calibrada, e o mapa de pressão foi interpolado para melhorar a visualização da distribuição espacial. A tensão negativa de pico máxima medida foi utilizada para estimar a PNP. Observa-se que o sistema foi capaz de reconstruir o perfil espacial da resposta ultrassônica na região avaliada, evidenciando o funcionamento integrado do posicionamento mecânico, da aquisição automática de sinais e do processamento de dados. Considerando a tensão negativa de pico máxima de 0,592 V, medida na saída do pré-amplificador, a PNP máxima estimada foi de 597,14 kPa.

Como complemento à caracterização do sistema, foi realizado um segundo ensaio para avaliar sua repetibilidade espacial, sendo obtidos, em sucessivas varreduras, os respectivos mapas 2D de pressão acústica negativa de pico e

Tabela II
MÉDIA E DESVIO PADRÃO (DP) DOS PARÂMETROS OBTIDOS EM CINCO ENSAIOS EXPERIMENTAIS DE CARACTERIZAÇÃO DOS TRANSDUTORES.

Transdutor	Parâmetro	Média ± DP	Unidade
1,6 MHz	f_L	$0,9 \pm 0,01$	MHz
	f_H	$1,85 \pm 0,01$	MHz
	$\mathrm{BW_{abs}}$	$0,95 \pm 0,01$	MHz
	$\mathrm{BW_{rel}}$	$59,38 \pm 0,44$	%
	f_c	$1,38 \pm 0,01$	MHz
	f_p	$1,14 \pm 0,20$	MHz
	NRMSE	$11,51 \pm 0,54$	%
	PCC	$0,9935 \pm 0,0006$	—
5 MHz	f_L	$3,33 \pm 0,03$	MHz
	f_H	$6,06 \pm 0,06$	MHz
	$\mathrm{BW_{abs}}$	$2,73 \pm 0,06$	MHz
	$\mathrm{BW_{rel}}$	$54,56 \pm 1,15$	%
	f_c	$4,70 \pm 0,04$	MHz
	f_p	$4,54 \pm 0,18$	MHz
	NRMSE	$11,88 \pm 5,92$	%
	PCC	$0,9916 \pm 0,0076$	—

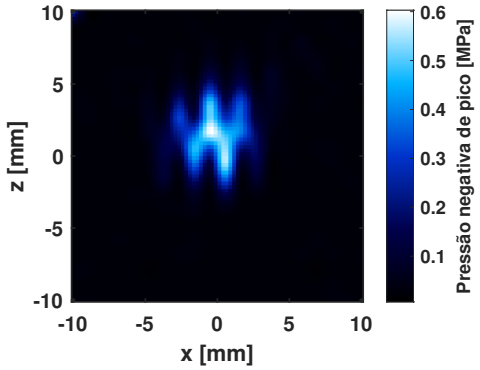


Figura 13. Mapa 2D da pressão acústica negativa de pico obtido pelo hidrofone em uma região de varredura de 20 × 20 mm com 20 × 20 posições de aquisição, após conversão da tensão e interpolação cúbica com fator 5.

as coordenadas $(x_{\max}, z_{\max})$ correspondentes aos pontos de máxima pressão. Entre as cinco varreduras, foram observadas variações máximas de 20 μm nos eixos x e z, caracterizando a repetibilidade espacial do sistema de posicionamento.

Como referência comercial, foi considerado um sistema de varredura de campo acústico da GAMPT Ultrasonic Solutions, composto pelo SF620 (*Sound field scanner 3 dimensions*), SF630 (*Sound field scanning drive with DLL*) e um tanque para medições em água, com custos de €4.201,00, €1.387,00 e €320,00, respectivamente, totalizando €5.908,00. Embora a solução comercial permita varreduras 3D, a plataforma desenvolvida foi projetada para varreduras 2D, atendendo às necessidades deste estudo. A solução proposta apresenta menor custo e integra a movimentação, a aquisição e o processamento de sinais, proporcionando maior flexibilidade para adaptações experimentais. Ressalta-se, contudo, que a comparação considera apenas os custos dos equipamentos, não abrangendo diferenças de precisão, velocidade de varredura e repetibilidade entre as soluções.

IV. CONCLUSÃO

Este trabalho apresentou o desenvolvimento e a avaliação preliminar de uma plataforma de baixo custo para caracterização de transdutores ultrassônicos, integrando um sistema de posicionamento mecânico controlado por microcontrolador, uma interface supervisória em LabVIEW e rotinas de processamento em MATLAB. Os resultados demonstraram o funcionamento adequado dos modos manual e autônomo, bem como a capacidade de aquisição, análise e mapeamento bidimensional da resposta acústica de transdutores ultrassônicos. Como trabalhos futuros, pretende-se aprofundar a avaliação da repetibilidade e da reprodutibilidade das medições nos campos próximo e distante, bem como implementar métricas de intensidade acústica, como a intensidade espacial de pico e temporal média (*Spatial Peak Temporal Average Intensity* – ISPTA), a intensidade espacial de pico e de pulso médio (*Spatial Peak Pulse Average Intensity* – ISPPA) e o índice mecânico (*Mechanical Index* – MI), ampliando as funcionalidades da plataforma para a caracterização de campos ultrassônicos [1]. Adicionalmente, está prevista a extensão da plataforma para uma configuração de varredura 3D, possibilitando a caracterização espacial completa do campo acústico e ampliando sua aplicabilidade.

AGRADECIMENTOS

Os autores agradecem à CAPES, ao CNPq, à Fundação Araucária e à SETI pelo apoio concedido.

REFERÊNCIAS

[1] T. L. Szabo, *Diagnostic Ultrasound Imaging: Inside Out*, 2nd ed. Oxford, U.K.: Academic Press, 2013.

[2] Y. Liu, K. A. Wear, and G. R. Harris, "Uncertainty of high intensity therapeutic ultrasound (HITU) field characterization with hydrophones: Effects of nonlinearity, spatial averaging, and complex sensitivity," *Ultrasound in Medicine & Biology*, vol. 43, no. 10, pp. 2329–2338, 2017.

[3] E. Martin and B. E. Treeby, "Investigation of the repeatability and reproducibility of hydrophone measurements of medical ultrasound fields," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1270–1282, 2019.

[4] R. C. Preston (Ed.), "Output Measurements for Medical Ultrasound," *Springer-Verlag*, London, UK, 1988.

[5] M. E. Schafer and P. A. Lewin, "A computerized system for measuring the acoustic output from diagnostic ultrasound equipment," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 35, no. 2, pp. 102–109, 1988.

[6] E. G. Radulescu et al., "Hydrophone spatial averaging corrections from 1 to 40 MHz," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 48, no. 6, pp. 1575–1580, 2001.

[7] M. R. Lien et al., "A low-cost and open-source platform for automated imaging," *Plant Methods*, vol. 15, article no. 1, 2019.

[8] M. Clinard et al., "Low-cost 3-D hydrophone scanning tank with MATLAB GUI control," *Ultrasound in Medicine & Biology*, vol. 48, no. 1, pp. 157–163, 2022.

[9] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 7th ed. Hoboken, NJ, USA: Wiley, 2017.

[10] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

Low-Cost Wrist Movement Separability Using Two-Channel Surface EMG and Time-Domain Features

Guilherme Felipe Ribeiro
Inatel
Santa Rita do Sapucaí, Brazil
guilherme.felipe@gec.inatel.br

Gabriel S. Carneiro Ribeiro
Inatel
Santa Rita do Sapucaí, Brazil
gabriel.sarlo@geb.inatel.br

Carlos Alberto Ynoguti
Inatel
Santa Rita do Sapucaí, Brazil
ynoguti@inatel.br

Luma R. Borges do Prado
Inatel
Santa Rita do Sapucaí, Brazil
luma.prado@inatel.br

***Abstract*—This work investigates the feasibility of identifying four distinct hand movements — wrist opening, wrist closing, wrist flexion, and wrist extension — from surface electromyographic (sEMG) signals acquired using only two channels positioned on the forearm. The signals were collected with surface sensors connected to an ESP32 microcontroller operating at a sampling rate of 1000 Hz, and subjected to a pre-processing pipeline based on moving average smoothing with window sizes of 5, 10, and 20 samples. A window of 10 samples provided the best trade-off between noise attenuation and preservation of muscular activation patterns. For feature extraction, an interval detection algorithm based on a two-threshold state machine was applied to isolate the contraction segments, after which the mean energy of each channel was computed, yielding a two-dimensional feature vector per repetition. The resulting scatter plot reveals well-separated clusters for wrist flexion and extension, while wrist opening and closing exhibit greater intra-class variability and cluster overlap. Overall, the proposed approach provides preliminary evidence of separability among the analyzed movements using a single time-domain feature and minimal hardware, highlighting its potential for low-cost myoelectric interfaces while leaving automatic classification for future evaluation.**

***Index Terms*—Electromyography, sEMG, Wrist Movement Recognition, Time-Domain Features, Low-Cost Myoelectric Interface, ESP32.**

I. Introduction

The functional loss of the upper limb, whether resulting from amputation, neuromuscular injuries, or degenerative diseases, significantly compromises the autonomy and quality of life of affected individuals, driving growing demand for myoelectric prostheses and human–machine interfaces capable of restoring hand movements in a natural and intuitive manner [1]–[3]. In this context, surface electromyography (sEMG) signals stand out as a non-invasive source of information about the user's motor intention, enabling the control of assistive devices from the electrical activity of the forearm muscles [4], [5].

Despite this potential, the widespread adoption of myoelectric interfaces still faces challenges related to the cost of acquisition systems, the computational complexity of classification algorithms, and the robustness against noise, physiological variability, and embedded hardware constraints [6], [7]. Most high-performance solutions available in the literature assume the use of multiple sEMG channels, specialized hardware, and advanced processing pipelines, which limits their applicability in low-cost scenarios.

Recent studies indicate that the number of acquisition channels directly influences gesture recognition performance: configurations with 8 to 16 channels tend to achieve high accuracy across large movement vocabularies, but at the expense of greater cost and complexity [8]. Conversely, reduced-channel arrangements can maintain competitive performance on smaller gesture sets, provided they are combined with appropriate electrode placement strategies, pre-processing, and feature extraction [9]. In parallel, there is a strong trend toward machine learning and deep learning techniques, including convolutional neural networks, hybrid CNN–LSTM models, and attention mechanisms, which typically demand higher computational resources [10], [11].

In this scenario, approaches based on simple time-domain features — such as mean absolute value (MAV), root mean square (RMS), signal amplitude, and moving average — remain relevant due to their low computational cost and ease of implementation on microcontrollers [12], [13].

Motivated by these considerations, this work investigates the separability of four distinct hand movements — wrist opening, wrist closing, wrist extension, and wrist flexion — from sEMG signals acquired using only two channels positioned on the forearm, relying exclusively on time-domain processing. The main contributions are: (i) the implementation of an analog acquisition and conditioning chain based on sEMG sensors and an ESP32 microcontroller; (ii) the evaluation of the effect of moving average smoothing (with window sizes of 5, 10, and 20 samples) on signal quality; and (iii) the exploratory analysis of mean energy from the two channels as a compact two-dimensional representation of the analyzed movements, supporting the feasibility of simple, replicable, and low-cost myoelectric interfaces.

The authors want to thank Fapemig for partial funding of this research.

II. Background

A. Electromyography (EMG)

Electromyography (EMG) is a well-established technique for assessing the electrical activity associated with skeletal muscle contraction, enabling the inference of motor behavior from the detection of action potentials generated by muscle fibers when activated by the central nervous system [4], [14], [15]. From this perspective, the electromyographic signal constitutes an indirect measure of the neural drive, allowing the study of motor unit recruitment strategies during voluntary and involuntary movements, as well as the functional analysis of muscle groups in dynamic or isometric tasks [4], [16].

EMG signals can be obtained through invasive approaches, such as needle electrodes inserted directly into the muscle, or through non-invasive techniques using electrodes placed on the skin surface [14]. In applications related to rehabilitation, myoelectric prosthesis control, and human–machine interfaces, surface electromyography (sEMG) is the preferred modality due to its safety, reduced user discomfort, and lower cost, making it widely employed in both clinical and research settings [4], [14].

Although a correlation exists between the EMG signal amplitude and the level of muscular activation, this relationship is influenced by multiple factors, including muscle fatigue, subcutaneous tissue thickness, electrode placement and orientation, contact impedance, and inter- and intra-individual variability [4], [16]. Consequently, the quantitative interpretation of the signal requires careful acquisition and processing procedures, particularly in systems that aim to recognize movement patterns from a limited number of channels.

The electromyographic signal is also highly susceptible to noise from various sources, including power line interference, motion artifacts, capacitive coupling with cables and adjacent devices, as well as thermal noise and external electromagnetic interference [6], [17]. Under these conditions, the adoption of analog conditioning and digital filtering strategies to attenuate unwanted signal components while preserving the physiologically relevant frequency band becomes indispensable.

The quality of the acquired sEMG signal critically depends on aspects such as the choice of muscle group, standardization of the inter-electrode distance, skin preparation, and repeatability of electrode placement across sessions [14], [18]. Inadequate placement may result in a low signal-to-noise ratio or in crosstalk — that is, the inadvertent capture of signals from neighboring muscles — thereby compromising physiological interpretation and reducing gesture discriminability [14], [18]. In systems employing few acquisition channels, such as the one proposed in this work, these issues become even more critical, as the amount of information available for movement differentiation is inherently limited [19].

1) Time-domain features: In the context of sEMG-based movement recognition, numerous feature extraction techniques have been proposed, spanning the time domain, frequency domain, and time–frequency representations [12], [20]. Time-domain features such as mean absolute value (MAV), root mean square (RMS), signal amplitude, and waveform length are widely used due to their low computational cost and their ability to effectively represent the intensity of muscular contraction, even in hardware-constrained systems [12], [13].

2) Challenges with a reduced number of channels: The use of a reduced number of sEMG channels introduces specific challenges related to the spatial limitation of the available information and to the overlap of electric fields generated by different muscles [3]. Even small variations in electrode placement between sessions or across users can significantly alter the observed activation patterns, hindering classifier generalization and the stability of amplitude-based decision thresholds [21]. Furthermore, muscle fatigue, variations in applied effort, and task context can modify signal amplitude and morphology, introducing relevant intra-class variability.

On the other hand, recent studies indicate that, when combined with appropriate placement strategies, pre-processing, and feature selection, reduced-channel configurations can maintain competitive performance on limited gesture vocabularies while simultaneously reducing cost, instrumentation complexity, and energy consumption [7]–[9], [22]. This approach is particularly attractive in resource-constrained scenarios, such as low-cost embedded systems.

III. Methodology

The goal of this work is to investigate whether the electromyographic patterns associated with four distinct hand movements (wrist opening, wrist closing, wrist flexion, and wrist extension) exhibit separable activation patterns when represented in a low-dimensional feature space. The analysis is exploratory and is intended to assess the feasibility of using the resulting representation as a basis for future movement classification and myoelectric interface applications. To this end, acquisition and pre-processing procedures were defined to assess the separability of the analyzed movements within a low-dimensional feature space.

This section describes the developed system, justifying the choices made at each stage. The acquisition protocol is presented first, followed by the signal acquisition and conditioning chain and, finally, the feature extraction procedure.

A. System Overview

The overall architecture of the proposed system is presented in Figure 1, which illustrates the complete flow from the acquisition of surface electromyography (sEMG) signals to computational processing and organization of the resulting data. The system is structured into functional blocks comprising: signal acquisition through surface electrodes; analog conditioning and amplification, including filtering for noise and interference reduction; digitization; and processing for digital filtering, feature extraction, and movement analysis.

Each of these stages is described in detail in the following subsections.

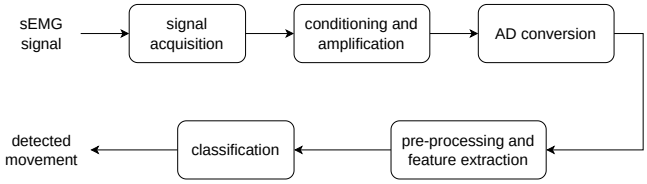


Fig. 1: Block diagram of the developed system.

B. Experimental Protocol

The experiment was conducted with the authors themselves, two healthy male volunteers, both 23 years old, right-handed and physically active (regular physical exercise approximately five times per week), with body mass between 65 kg and 75 kg and no history of fractures or muscular injuries in the upper limbs. Participation was voluntary, and the procedures were explained to the participants prior to the data collection.

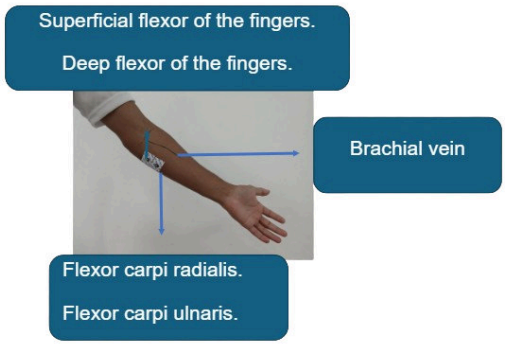
During the experiments, the volunteers remained seated, with the dominant arm supported and relaxed at the beginning of each acquisition. The experimental environment was kept as controlled as possible in order to minimize external interference, postural variations, and muscle fatigue effects over the repetitions.

The surface electrodes were positioned on two specific regions of the forearm, selected based on the muscles predominantly activated during the analyzed movements. Placement was standardized using a previously defined template, ensuring repeatability across acquisitions and across volunteers. Figure 2 presents the adopted positioning from different perspectives, along with the template used to define the electrode fixation points and the distribution of the anatomical references employed.

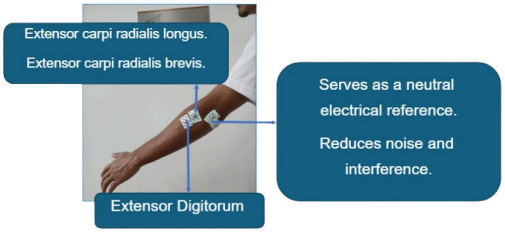
Four distinct forearm and hand movements were analyzed, selected for their functional relevance and because they involve the activation of different muscle groups:

- Wrist opening, associated primarily with the activation of the finger extensor muscles;
- Wrist closing, associated with the activation of the finger flexor muscles;
- Wrist flexion, related predominantly to the flexor carpi radialis muscle;
- Wrist extension, related to the wrist extensor muscles.

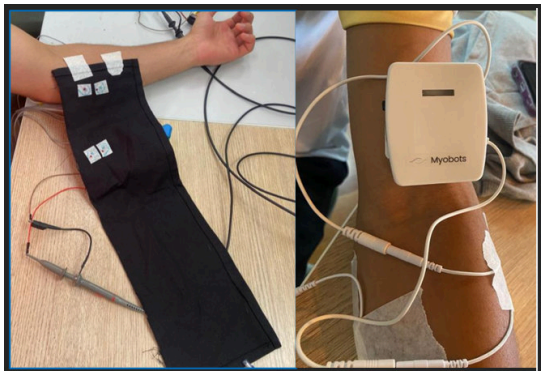
For each movement, 50 repetitions were performed per volunteer, providing a set of samples suitable for the exploratory analysis carried out in this study. Each repetition followed a standardized temporal cycle consisting of 2 seconds of active muscular contraction, corresponding to the execution of the movement, followed by 2 seconds of rest, allowing the return to the baseline state, for a total duration of 4 seconds per repetition. Each acquisition began with the muscle at rest, followed by the execution of the movement during the established contraction time, and ended with the relaxation period. This standardization was adopted to ensure reproducibility of the collections and comparability among the sEMG signals acquired for different movements and volunteers.



(a) Sensor 1 placement.



(b) Sensor 2 placement.



(c) Template for sensor placement.

Fig. 2: Placement details for each of the sensors.

C. Signal Acquisition and Conditioning

1) Signal acquisition: The sEMG signals were acquired using surface sensors designed to capture the electrical activity generated by muscle fibers during contraction. These sensors perform the initial amplification and conditioning of the signal, providing an analog output compatible with the ADC available in the microcontroller.

The system was configured with two acquisition channels, corresponding to two sensors positioned on different regions of the forearm. The use of a reduced number of channels makes it possible to assess the feasibility of distinguishing hand movements from a simplified architecture, aligned with low-cost applications and embedded systems with limited computational resources. The details of the sensor placement are described and justified in Section III-B.

2) Amplification and conditioning: The analog signals from the sensors are routed to a signal conditioning board, responsible for reducing external interference, adjusting the amplitude level, and matching the dynamic range of the signals to the input of the microcontroller's analog-to-digital converter (ADC). This stage contributes to improving the signal-to-noise ratio and to preventing saturation during digitization.

Since sEMG signals are highly susceptible to electrical noise, environmental interference, and amplitude variations arising both from muscular activity and from the limitations of the acquisition system, and considering the restricted dynamic range of the ADC integrated into the ESP32, an analog pre-processing stage was required prior to digitization. For this purpose, a dedicated acquisition board was developed to condition the analog signals, limiting the signal amplitude to a range compatible with the input of the microcontroller's ADC while improving the signal-to-noise ratio.

3) A/D conversion: After analog conditioning, the signals are sent to an ESP32 microcontroller, which performs the analog-to-digital conversion. This device was selected for its processing capability, the availability of multiple ADC channels, and the ease of communication with external computing systems — characteristics suitable for near-real-time biomedical signal acquisition applications.

The sEMG signals were acquired at a sampling rate of 1000 Hz, as this value preserves information relevant to time-domain analysis, to the visualization of contractions, and to the differentiation among the executed movements, while keeping the data volume compatible with near-real-time processing. The acquired data were stored in Comma-Separated Values (CSV) files, which facilitates subsequent organization and processing.

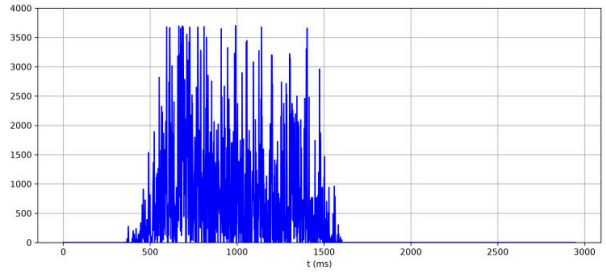
D. Feature Extraction

The objective is to differentiate the four movements described above. Figures 3 to 6 below show the sensor readings for the wrist flexion and extension movements. During wrist extension, the signal from Sensor 1 is considerably larger than the one read by Sensor 2, whereas during wrist flexion the readings are reversed: the signal from Sensor 1 is smaller than the one from Sensor 2. For the wrist opening and closing movements, the signals from the two sensors exhibit approximately the same amplitude.

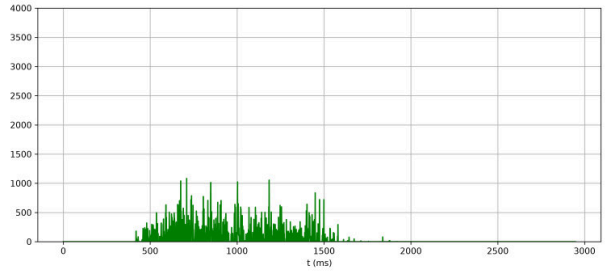
This behavior suggests that the relative activation patterns captured by the two sensors may provide information for distinguishing the analyzed movements. Accordingly, the mean energy calculated independently for each sensor is used to construct a compact two-dimensional feature representation of each repetition.

The computation of this parameter follows the procedure described in Figure 7, detailed below: moving average filtering, analysis interval detection, and mean energy calculation.

1) Moving average filtering: The noisy nature of the signal hinders the detection of the onset and offset instants of muscular activation. To facilitate this task, a moving average filter was applied to smooth the signal:

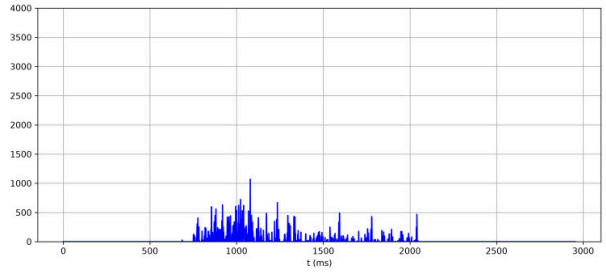


(a) Sensor 1.

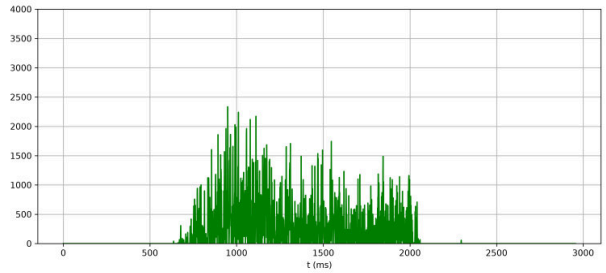


(b) Sensor 2.

Fig. 3: Sensor signals during the wrist flexion movement.

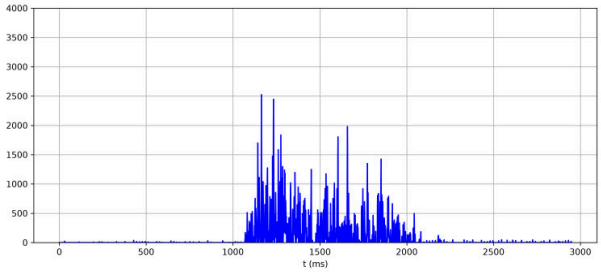


(a) Sensor 1.

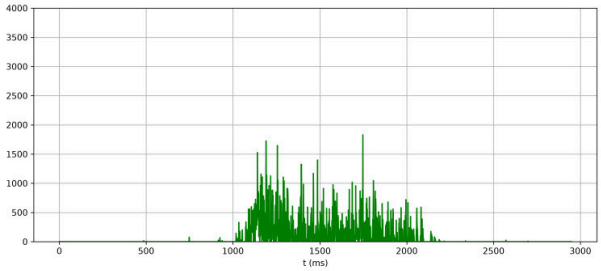


(b) Sensor 2.

Fig. 4: Sensor signals during the wrist extension movement.

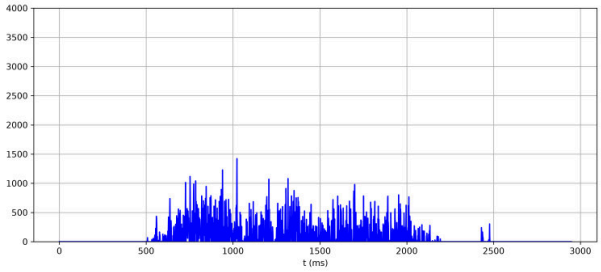


(a) Sensor 1.

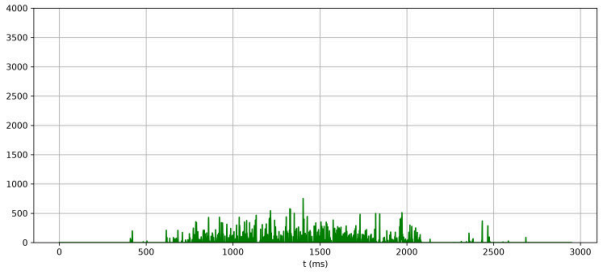


(b) Sensor 2.

Fig. 5: Sensor signals during the wrist opening movement.



(a) Sensor 1.



(b) Sensor 2.

Fig. 6: Sensor signals during the wrist closing movement.

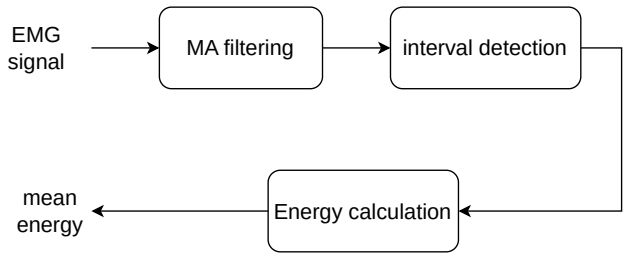


Fig. 7: Block diagram of the feature extraction procedure.

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} x[n-k] \tag{1}$$

where:

- $x[n]$ is the original sEMG signal at the discrete time instant n;
- $y[n]$ is the smoothed signal after applying the filter;
- N is the moving average window size (the number of samples considered in the computation);
- k is the summation index, which iterates over the preceding samples within the window;
- n is the discrete time index of the signal.

Moving average windows of 5, 10, and 20 samples were tested in order to evaluate the impact of the window size on the degree of signal smoothing. The 10-sample window provided the best trade-off between noise reduction and preservation of the signal characteristics, maintaining the integrity of the muscular activation peaks while significantly reducing undesired fluctuations. Figure 8 shows the effect of this filter on the signal.

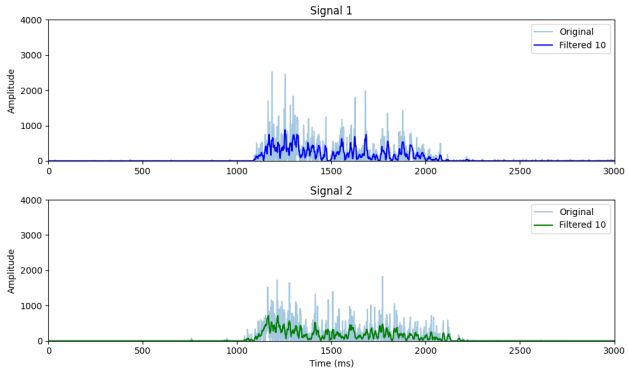


Fig. 8: sEMG signal of the wrist extension movement after appl:ying the moving average filter with a 10-sample window.

2) Analysis interval detection: Defining the analysis interval is important because it discards the regions of the signal where there is no muscular activity (rest regions), thereby allowing an accurate reading of the activation levels for each sensor. The algorithm used for this purpose consists of the following steps:

a) Each signal is divided into 10 ms windows, with 50% overlap between adjacent windows.
b) The energy of each window is computed as the sum of the squared samples of the smoothed signal within that window:

$$E_i = \sum_{n \in W_i} y[n]^2 \tag{2}$$

where E_i is the energy of the i-th window W_i and $y[n]$ is the smoothed signal. The largest value among all windows (E_{max}) is taken as a reference.
c) Two thresholds k_1 and k_2 are defined: k_1 corresponds to 10% of E_{max} and k_2 to 20% of E_{max}.
d) To determine the onset of the interval, a three-state machine is used: 0 (no signal), 1 (onset candidate), and 2 (onset). The system starts in state 0. Beginning at the start of the signal, the energy of each window is monitored. When the energy of a window exceeds k_1, the system transitions to state 1 (onset candidate). From there, two scenarios are possible: if the energy in the subsequent windows remains above k_1 and later exceeds k_2, the onset is set as the first window in which k_2 was reached; if the energy drops below k_1, the system returns to state 0. This mechanism minimizes the occurrence of false positives.
e) The detection of the offset instant is performed analogously, using the same thresholds k_1 and k_2.

For each signal, the onset and offset points of the interval of interest were determined. Figure 9 illustrates this procedure for the wrist extension movement, showing Channel 1 and Channel 2 of the same contraction separately and highlighting the limits used in the mean computation. It can be observed that, although both signals correspond to the same movement, the onset and offset instants may differ between channels, evidencing variations in the intensity of the signal captured by each sensor.

3) Mean energy calculation: Within the interval selected for each acquisition, the mean energy of the signal was computed, resulting in a single scalar value per channel. It is obtained using the following equation:

$$\overline{E} = \frac{1}{M} \sum_{i=1}^{M} E_i \tag{3}$$

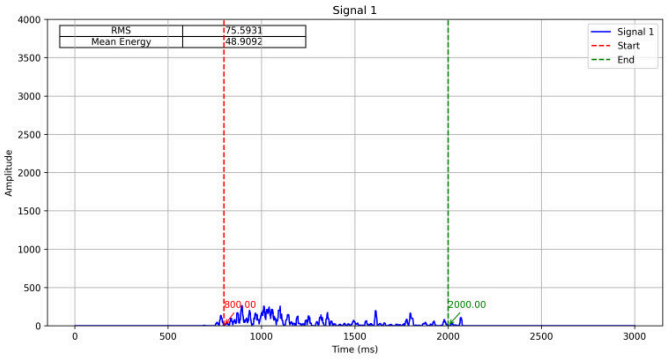
where:

- $\overline{E}$ is the mean energy;
- M is the number of windows within the analysis interval;
- E_i is the energy of the i-th window, as defined in Equation (2).

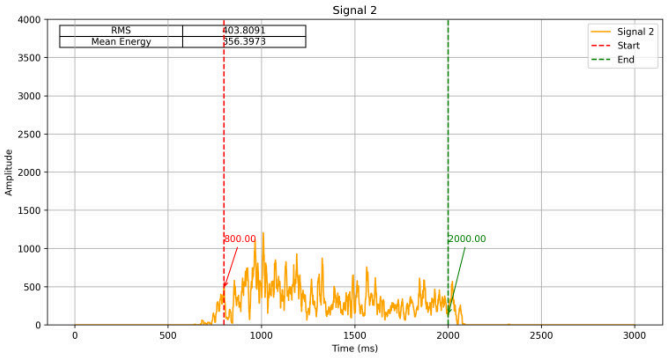
As a result, each executed movement is represented by only two scalar values.

IV. Results and Discussion

To explore the separability of the analyzed movements, the mean energy values obtained in each repetition were arranged in a two-dimensional scatter plot, with Channel 1 on the horizontal axis and Channel 2 on the vertical axis, as shown in Figure 10. The analysis is based on visual inspection of the resulting feature-space distribution and does not constitute a quantitative classification evaluation.



(a) Sensor 1.



(b) Sensor 2.

Fig. 9: Onset and offset instants of the interval of interest produced by the interval detection algorithm for the wrist extension movement.

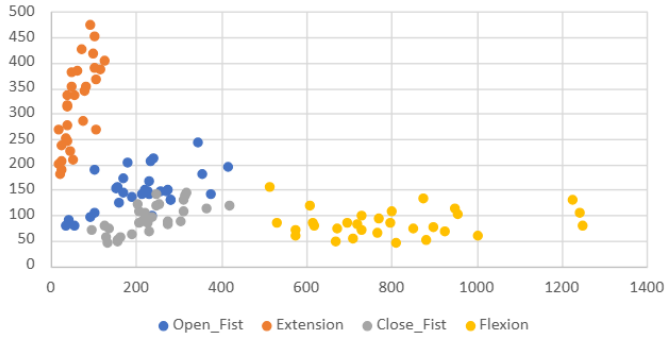


Fig. 10: Scatter plot of the mean energy values of the EMG signals for the four analyzed movements. Each point represents an individual repetition. The colors indicate the different movements: extension (orange), wrist opening (blue), wrist closing (gray), and flexion (yellow).

The formation of clusters associated with each movement can be observed. Extension and flexion occupy weakly

overlapping regions, indicating good separation when both channels are considered. In contrast, wrist opening and wrist closing form closer clusters with a greater tendency to overlap, suggesting more similar mean activation patterns or higher intra-class variability.

Using only two channels and a single time-domain feature (mean energy), the scatter plot provides preliminary evidence of separability among the analyzed movements based on their characteristic activation patterns. Wrist flexion and extension exhibit comparatively distinct regions in the feature space, whereas wrist opening and closing show greater overlap. These observations support the feasibility of the proposed low-dimensional representation as a basis for future automatic classification, but they should not be interpreted as a quantitative measure of classification performance.

These findings should nonetheless be regarded as preliminary, since they were obtained from a small and homogeneous sample of two healthy male volunteers of similar age and build. Consequently, the observed separability cannot be assumed to generalize to a broader population. In particular, no quantitative classification performance was evaluated in the present study. Validation with a larger and more diverse group of participants, together with quantitative classification experiments, is required before drawing more general conclusions.

V. Conclusion

This work addressed the difficulty of acquiring, processing, and interpreting EMG signals in an accessible and reliable manner, with a focus on exploring the separability of different muscular movements using a low-dimensional representation, particularly in applications related to the analysis of upper-limb movements. The variability of EMG signals, the presence of noise, and the need for appropriate processing techniques represent relevant challenges for the correct identification of the muscular patterns associated with each movement.

In this context, an EMG signal acquisition system was developed using the ESP32 microcontroller, operating at the sampling rate defined for the experiment, with the acquired signals stored in CSV files and subsequently processed on a computer. The system incorporated moving average smoothing with windows of different sizes, aimed at reducing noise and improving the representation of muscular activation patterns. The mean energy of each channel was then extracted as the feature used to characterize and compare the signals.

The obtained results showed that the adopted processing techniques provided a clearer representation of the muscular activation patterns. The graphical analysis based on the two-channel feature space revealed that wrist extension and flexion form comparatively well-separated clusters, while wrist opening and closing exhibit greater overlap. Thus, the results provide preliminary evidence that the proposed two-dimensional representation captures movement-dependent activation patterns, while quantitative classification performance remains to be evaluated.

As its main contribution, this work presents a low-cost and relatively simple approach for the acquisition and processing of EMG signals, integrating accessible hardware and efficient computational methods. The resulting two-dimensional feature representation provides an exploratory basis for studying movement-dependent activation patterns and can serve as a starting point for future classification studies and applications in device control, assistive rehabilitation, human–machine interfaces, and academic studies focused on biomedical signals.

As future work, quantitative classification experiments using the existing feature representation are suggested, including the evaluation of suitable classifiers and performance metrics such as accuracy, sensitivity, and confusion matrices. Comparisons with established time-domain features such as mean absolute value (MAV), root mean square (RMS), and waveform length would also help to position the proposed representation relative to classical sEMG descriptors. In addition, increasing the number and diversity of volunteers and the number of analyzed movements is necessary to enhance the robustness and generalizability of the findings. Furthermore, the investigation of additional filtering and feature extraction techniques, along with the integration of the system into real-time actuation devices, could significantly broaden the scope and applications of the project.

References

[1] R. Merletti and P. A. Parker, *Electromyography: Physiology, Engineering, and Non-Invasive Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2004.

[2] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface emg," *Journal of Applied Physiology*, vol. 96, no. 4, pp. 1486–1495, 2004.

[3] E. J. Scheme and K. B. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *Journal of Rehabilitation Research & Development*, vol. 48, no. 6, pp. 643–660, 2011.

[4] C. J. De Luca, "The use of surface electromyography in biomechanics," *Journal of Applied Biomechanics*, vol. 13, no. 2, pp. 135–163, 1997.

[5] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 7, pp. 848–854, 2003.

[6] J. G. Webster, *Medical Instrumentation: Application and Design*, 4th ed. Hoboken, NJ, USA: Wiley, 2010.

[7] S. Muceli and D. Farina, "Simultaneous and proportional estimation of hand kinematics from emg during mirrored movements at multiple degrees-of-freedom," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 3, pp. 371–378, 2012.

[8] E. Campbell, A. Phinyomark, and E. Scheme, "Implications of EMG channel count: Enhancing pattern recognition online prosthetic testing," *Frontiers in Rehabilitation Sciences*, vol. 5, p. 1345364, 2024. [Online]. Available: https://www.frontiersin.org/journals/rehabilitation-sciences/articles/10.3389/fresc.2024.1345364/full

[9] M. A. Ozdemir, D. H. Kisa, O. Guren, and A. Akan, "Classification of hand and wrist movements via surface electromyogram using machine learning approaches," *Biomedical Signal Processing and Control*, 2024, available via PubMed Central, PMC10876538. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10876538/

[10] Y. Çelik and U. Can, "Surface EMG-based hand gesture recognition using a hybrid deep learning model," *Sensors / PubMed Central*, 2025, available via PubMed Central, PMC13075277. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC13075277/

[11] L. Tong, Y. Li, Y. Liang, and C. Wang, "CAM-MR-MS based gesture recognition method using sEMG," *Intelligence and Robotics*, 2025. [Online]. Available: https://www.oaepublish.com/articles/ir.2025.15

[12] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "A novel feature extraction for robust emg pattern recognition," *Journal of Computing*, vol. 1, no. 1, pp. 71–80, 2009.

[13] M. Asif, M. A. Khan, N. Naseer, and M. J. Khan, "Human hand movement classification based on EMG signal using different feature extractors," *Biomedical and Pharmacology Journal*, vol. 17, no. 1, 2024. [Online]. Available: https://biomedpharmajournal.org/vol17no1/human-hand-movement-classification-based-on-emg-signal-using-different-feature-extractor/

[14] R. Merletti and P. A. Parker, *Electromyography: Physiology, Engineering, and Non-Invasive Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2004.

[15] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface emg," *Journal of Applied Physiology*, vol. 96, no. 4, pp. 1486–1495, 2004.

[16] D. Farina, A. Holobar, and R. Merletti, "Decoding the neural drive to muscles from the surface electromyogram," *Clinical Neurophysiology*, vol. 121, no. 10, pp. 1616–1623, 2010.

[17] R. Pallás-Areny and J. G. Webster, "Common mode rejection ratio of differential amplifiers in emg applications," *IEEE Transactions on Biomedical Engineering*, vol. 38, no. 5, pp. 524–529, 1991.

[18] H. J. Hermens *et al.*, "Development of recommendations for semg sensors and sensor placement procedures," *Journal of Electromyography and Kinesiology*, vol. 10, no. 5, pp. 361–374, 2000.

[19] D. Stegeman and H. Hermens, "Standards for surface electromyography: The european project seniam," Roessingh Research and Development, Enschede, Tech. Rep., 2007.

[20] M. A. Oskoei and H. Hu, "Support vector machine-based classification scheme for myoelectric control applied to upper limb," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1956–1965, 2008.

[21] A. J. Young, L. J. Hargrove, and T. A. Kuiken, "Improving myoelectric pattern recognition robustness to electrode shift by changing interelectrode distance and electrode configuration," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 3, pp. 645–652, 2012.

[22] A. Phinyomark, H. Hu, and P. Phukpattaranont, "Application of linear discriminant analysis in emg pattern recognition," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 7, pp. 2096–2104, 2014.

Maximizing EEG task-related information in a Motor Imagery cue-based paradigm

Gabriel Chaves de Melo
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas, Unicamp
Campinas, Brazil
gabrielchaves@alumni.usp.br

Gabriela Castellano
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas, Unicamp
Campinas, Brazil
gabriela@unicamp.br

Arturo Forner-Cordero
Biomechatronics Laboratory
University of São Paulo
São Paulo, Brazil
aforner@usp.br

***Abstract*—Brain-Computer Interfaces (BCIs) based on electroencephalography (EEG) are expected to become an important solution for extreme cases of neuromotor rehabilitation. To do so, it is key to comprehend the characteristics of brain signals elicited in typical BCI paradigms. One of the most popular paradigms for BCI studies is cue-based with visual stimulus indicating the motor imagery (MI) task to be performed. This paradigm is widely adopted to study MI signals and calibrate BCI systems before usage. An important challenge for EEG decoding is the variability of the signal, which is caused, among other reasons, by the inappropriate reference electrode choice. Despite some alternatives to minimize the problem, it remains as an unsolved issue. Recently, a new promising approach to this issue called Inverted Perspective Reference Electrode (IPRE) was proposed. In this paper, we use the IPRE to investigate EEG signals from a publicly available MI-cue-based dataset to find which cortical information dominates the recordings in this paradigm. An IPRE-based score metric is proposed to rank the channels according to the amount of information they contain. Then the 10 channels with highest scores are analyzed along with their signals. Our hypothesis is that either visual or motor cortical activities are more likely triggered by the task paradigm. Therefore, the procedures were applied to α band energy signals corresponding to the trials of left-hand MI. Results showed notable agreement among the channels selected across different subjects, especially the presence of O2 occupying the reference position and the other electrodes distributed over the parieto-occipital cortex, slightly tending to the right motor cortex. The event-related desynchronization (ERD) time-linked to the cue appearance was identified for all subjects and channels, although presenting subject-specificities. The channels and the time-energy features suggest that the ERD corresponds mostly to visual processing with some contribution of the motor cortex due to MI performance.**

***Index Terms*—Electroencephalogrphy (EEG), Motor-Imagery, IPRE, Reference electrode**

I. Introduction

Brain-Computer Interfaces (BCIs) are systems designed to provide a direct communication pathway between the brain and external devices, translating neural signals into computer commands. In the past decades, they have gained attention as a promising alternative for motor rehabilitation due to their encouraging results in such applications [1], [2]. BCIs have become an even more appealing subject when considering the limitations of current therapies towards severely impaired people, such as when a person has only residual motor control over a limb or no control at all [3], [4]. Electroencephalography (EEG) based BCIs are the most popular type of BCI, especially for safety (it is a non-invasive technique) and practical reasons, such as the high temporal resolution of EEG measurements [5]. In this scenario, the analysis of motor imagery (MI) EEG signals is key, given that even in extreme situations patients may still be able to activate their motor cortex by using imagery.

The EEG analysis, however, presents some major challenges, which have been limiting the use of BCIs in actual rehabilitation therapies. Motor imagery signals, in fact, increase these challenges compared to motor execution, for instance. The intra-subject variability of EEG signals is perhaps the major difficulty to assess and has been a research topic from pioneering works related to EEG-based BCIs [6], [7] till the current state-of-the-art [8], [9]. A major part of such variability is caused by the inappropriate choice of the reference electrode [10], [11]. Despite the existence of classical techniques to reduce the reference electrode influence, such as common average reference [12], REST [13], and surface Laplacian [14], [15], there is still no golden standard, and each of these approaches introduces other sources of uncertainties [8]. A different approach is to assume the impossibility to eliminate the reference issue and look at any channel as a bipolar derivation [8]. However, this approach generates a large amount of data, requiring the definition of specific criteria to select subset of channels. A recent approach called Inverted Perspective Reference Electrode (IPRE) [16] suggests that a monopolar set of channels should be viewed as a collection of information from the reference site, and uses PCA to extract the first principal component (PC), in order to maximize the information about the reference in just one signal. In IPRE, the location of interest should be set as the reference and the multi-channel monopolar signals are substituted by the PC.

This research was supported by Coordination for the Improvement of Higher Education Personnel (CAPES), Finance code 001; São Paulo Research Foundation (FAPESP), Grants 2013/07559-3 (BRAINN), 2023/12422-9 and 2024/00998-6 (CRISQuaM); Brazilian National Council for Scientific and Technological Development (CNPq), Grants 306255/2024-3 and 408389/2024-9 (INCT-NeuroComp); Funding Authority for Studies and Projects (FINEP), Grant 2140/22 (DReANNs); and Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001.

In the BCI literature, the offline investigation of MI signals plays an important role in improving the understanding of task-related features [17]. The underlying mechanisms of the brain are complex even during a simple task, such as MI performance. This complexity is due, among other factors, to the variety of stimuli inherent to any experimental setup. These stimuli can originate from the clinical or laboratory environment or from the BCI paradigm itself. When it comes to the BCI paradigm, the use of virtual scenarios with visual cue-based trials is probably the most common strategy for signal acquisition [17]. Therefore, it is important to identify brain activities present in EEG signals in this paradigm, regardless of whether they are directly related to MI or to other stimuli.

In this paper we explored the IPRE approach to rank the channels according to their contribution to the task-related information captured by the IPRE approach in a cue-based paradigm. To rank the channels, we adapted the IPRE strategy to quantify how much each channel (combination of two electrodes) influenced the PCs. To do so, a score was proposed to account for the importance of each component and how much each channel influenced it. With this score, we could rank the channels for each subject. A publicly available dataset [18] with MI signals was used. Because the signal acquisition paradigm was MI cue-based, it was hypothesized that visual potentials related to the cue and motor-related potentials associated with the MI task would be elicited more than other EEG signatures. In both cases, it has been shown that the event-related desynchronization (ERD), which is a band-specific energy reduction time-locked to cortical activation, occurs as an indication of visual processing or motor-imagery execution. The ERDs for these cortical activities have been reported to occur mainly in the α band ($8-12$ Hz) over the occipital cortex for visual processing and over the motor cortex for motor-imagery [19]–[22].

II. Methods

A. EEG Dataset

The publicly available data set from BCI Competition IV (dataset 1, calibration data) [18] has EEG related to MI tasks from seven subjects labeled as A through G. However, subjects C, D, and E correspond to artificially generated signals, so they were not considered in this work. Data were recorded with 59 electrodes according to the $10-20$ extended system distributed over the scalp and more concentrated on the motor cortex. In this work, only MI for the left hand was considered, although the dataset also contained right-hand MI trials. Subjects were oriented to look at a computer screen and execute the MI according to the cues appearing on the screen (arrow pointing left indicated left hand MI). It was a permanent process consisting of 2s of blank screen, 2s with a cross in the center of the screen, 4s with the cross and the cue, and the same sequence repeated multiple times. By the end of data acquisition there were 75 left-hand MI trials. The trial interval with the starting point after the first 1 s of blank screen is schematically shown in Fig. 1.

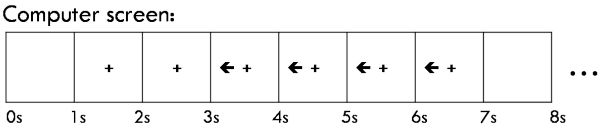


Fig. 1. Sequence of one trial 8s long.

B. EEG Pre-processing

For the entire channel selection procedure and results presentation, only the 4-second interval from $t = 1s$ to $t = 5s$ was considered (see Fig. 1). This 4-second interval was chosen because it contains the most relevant events of the task, which are the cross appearing on the screen as an attention signal, the cue presentation indicating the MI to be executed, and the initiation of the MI activity itself, which likely occurs within two seconds post cue appearance. Because the focus was to find channels with ERD in the α band, signals were band-pass filtered in the $8-12$ Hz range with a 6^{th} order Butterworth filter in both directions.

C. Channel Scoring Pipeline

To perform the channel selection, a procedure consisting of 7 steps took place repeatedly and separately for each subject: 1) Definition of an electrode to be set as the reference; 2) Data re-referencing to the new reference; 3) Transformation of the signals into energy by squaring the samples; 4) Smoothing of the energy signals with a 100 ms sliding window; 5) Trial average calculation separately for each channel; 6) PCA transformation; 7) Score calculation for each combination of channel (pair of electrodes) and PC, which was designated as Total Score.

The procedure was repeated so that all available electrodes were eventually set as the reference in step 1. After all 59 electrodes were used as the reference, all channels (pairs of electrodes) along with their corresponding PC were ranked according to their score. For each subject, the top 10 scores were analyzed to select the channels.

D. IPRE, Channel Impact and Total Score

The IPRE core, as presented in [16], is a way to approach a multichannel monopolar dataset to reduce the issue of the active reference electrode. It consists of setting the electrode of interest as the reference for all the others, opposed to the traditional approach, and calculating the first PC which supposedly contains the maximum information about the reference. In the IPRE, the first PC would then be used as the signal to be analyzed. However, the method proposed in this work does not use the first PC as a substitute for the signals, but instead it quantifies how much each channel contributed to the PC. This channel contribution, called Channel Impact, and the amount of information (variance) of the corresponding PC are used to generate the Total Score for each channel. In the following, details on how the PCA, Channel Impact and Total Score were calculated are presented. Although PCA is a well-known multivariate technique that has been applied

to EEG signals for decades [23]–[28], a brief explanation is given here to provide a more intuitive reasoning for the proposed scoring. Given the EEG data matrix X_{np}, where each row is a time sample and each column is an EEG channel,

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix}$$

we calculate the correlation matrix R_{pp} where each element r_{jk} is given by

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{X_j})(x_{ik} - \bar{X_k})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{X_j})^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{X_k})^2}} \quad (1)$$

where $\bar{X_j}$ and $\bar{X_k}$ are the mean values of columns (channels) j and k, respectively. Next, the eigenvalue problem must be solved for R and the eigenvector $\tilde{a}_i$ must be calculated.

$$|R - \lambda I| = 0 \quad (2)$$

$$\tilde{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix} \quad (3)$$

Finally, the $i-th$ component can be determined with the $i-th$ eigenvector as

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p \quad (4)$$

where $X_1, ..., X_p$ are columns (channels) of a particular re-referenced EEG matrix. The contribution C_i of the $i-th$ PC to the original multichannel data is usually calculated as the percentage of its variance compared to the total variance. However, because we were to compare values from different sets of re-referenced multichannel EEG, the total variance could be seen as an attribute of that reference. Thus, we considered the absolute value of the variance of the first PC instead of its percentage relative to the corresponding total variance. In this case, the contribution was given by:

$$C_i = Var(Y_i) \quad (5)$$

With Y_i, it was possible to quantify the influence of the channel X_j (Channel Impact) on the $i-th$ PC using:

$$Corr(X_jY_i) = r_{X_j.Y_i} = a_{ij}\frac{\sqrt{Var(Y_i)}}{\sqrt{Var(X_j)}} \quad (6)$$

The proposed concept of Total Score TS_{jki} for a channel X_j (which is a combination of electrode j and reference electrode k) associated to a component Y_i was given by Eq. (7):

$$TS_{jki} = C_i r_{X_j.Y_i} \quad (7)$$

Because only the first PC was considered, $i = 1$ for any j and k. Finally, all channels (pairs of electrodes) were ranked according to their TS_{jki} values. The top 10 channels for each subject were selected for analysis.

III. Results

Fig. 2 shows a visual representation of the top 10 channels' positions for each subject. Tab. I informs the 10-20 positions of these same channels, indicating which electrode is used as the reference. With Fig. 3 and Tab. I it can be seen that O2 was the reference electrode in all cases and all the other electrodes were placed in the posterior part of the coronal plane, which means $10-20$ positions between C line (not included) and O line (included). Also, for all four subjects, the closest electrodes to the C line occurred at the right hemisphere.

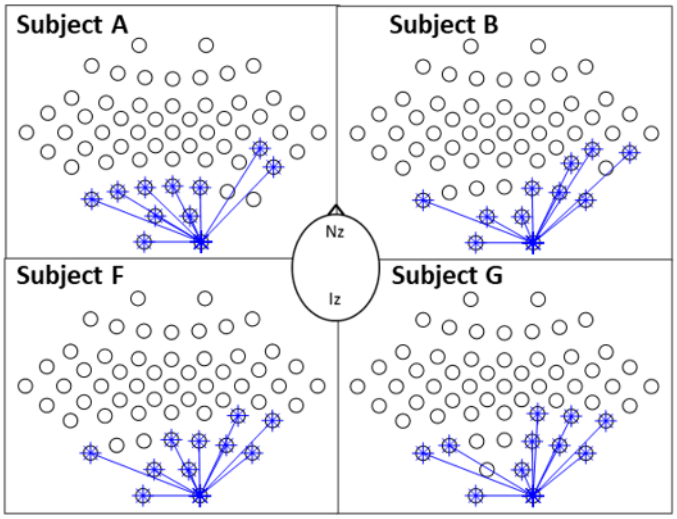


Fig. 2. Visual representations of the top 10 channels for each subject. The lines connect both electrodes within a channel.

Fig. 3 shows the trial-average of the smoothed energy in the α band of all top 10 channels from each subject. A visual inspection allows to find, within each subject, relevant similarities among the signals in the entire interval, given that most of the positive and negative peaks are fairly synchronized, despite magnitude differences. However, only some specific similarities were found among different subjects, which contrasts with the inter-subject similarities between the areas corresponding to the selected channels. For instance, Subjects A and F had similar behaviour in their signals, although for Subject F there was a clear tendency of energy reduction between $t = 2s$ and $t = 3s$ that was not found for Subject A. Subjects B and G, on the other hand, had more unique behaviour and energy levels as well. The time evolution of the energy for Subject G showed a more typical

TABLE I
TOP 10 CHANNELS FOR EACH SUBJECT. "REF" REFERS TO THE ELECTRODES OCCUPYING THE REFERENCE POSITION AND "ELECT" REFERS TO THE CORRESPONDING ELECTRODES ASSOCIATED TO THE HIGHEST SCORES.

Subject A		Subject B		Subject F		Subject G	
Elect.	**Ref.**	**Elect.**	**Ref.**	**Elect.**	**Ref.**	**Elect.**	**Ref.**
PO2	O2	P6	O2	PO2	O2	P6	O2
O1	O2	P4	O2	P6	O2	P4	O2
PO1	O2	O1	O2	P4	O2	O1	O2
P5	O2	PO2	O2	O1	O2	P2	O2
CP6	O2	P2	O2	P2	O2	CP4	O2
Pz	O2	CP4	O2	PO1	O2	PO2	O2
P2	O2	P5	O2	Pz	O2	P3	O2
P1	O2	CCP6	O2	CP4	O2	P5	O2
CCP6	O2	PO1	O2	CP6	O2	CP2	O2
P3	O2	CCP8	O2	P5	O2	CP6	O2

ERD following the cue-appearance, because prior to that there was a 1-second interval with reasonably stable energy levels, matching the requirement for a good ERD baseline [20]. One temporal mark common to all subjects was the negative peak between $t = 3s$ and $t = 4s$, right after the movement cue appearance ($t = 3s$). Moreover, in general there was an event-related desynchronization (ERD) for all subjects, despite their different intensities, initiating approximately at $t = 2s$ (except for Subject G which initiated at $t = 3s$) associated with the negative peak in between $t = 3s$ and $t = 4s$.

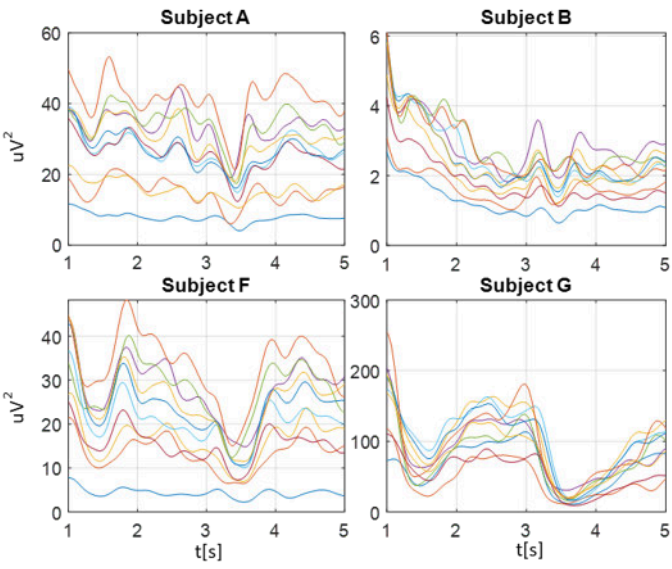


Fig. 3. The trial-average signals from each of the 10 channels of each subject.

IV. DISCUSSION

In this work we proposed a metric using the IPRE to maximize the task-related information in an MI task based on a visual cue paradigm. The procedure allowed to identify channels that carried α band energy changes synchronized to important task events, notably the cue appearance. The selected channels across the four subjects presented notable similarities concerning the cortical regions and shared the same reference electrode.

PCA is a multivariate statistical approach that transforms the original data into uncorrelated components in descending order of variance, i.e., amount of information of the original data. When applied to a multichannel EEG data, this means that these components carry most of the overall information from all channels in the corresponding interval and frequency band. Each set of data in this case is a monopolar set, which means that there are 58 electrodes referenced to the same electrode, making it 58 monopolar channels. In this sense, the IPRE approach [16] argues that the PCA is actually an analysis of the one electrode being used as the reference. In a simple analogy, it relates to 58 points-of-view of one single object (the reference electrode, in this case). This is particularly suitable in this work because the voltage signals are transformed into energy prior to the analysis, therefore their samples are squared, nullifying the effect of which electrode occupies the reference position. The results here presented seem to corroborate this hypothesis, mainly because for each subject all channels had the same reference and at the same time there were notable similarities among the signals.

It has been reported that a simple reference electrode test for the analysis of a certain null-reference candidate location is to change the reference and see if it will produce significant changes in the signal, and if the signals do not change, it is an indication that the locations are adequate for a "silent" reference [11]. This is similar to what was observed in the results if we designated the reference as the "main" electrode and the others as different references for the given "main" electrode. Taking into consideration the presented hypothesis and the results, it seems that the scalp location O2 was the one with most EEG information for all subjects in the MI task protocol in the α band ($8 - 12$ Hz) during the analyzed interval lying in between cross appearance on the screen and 2s post movement cue appearance indicating the left hand MI. This well-founded test and its agreement with the presented results reinforces the validity of these findings.

The similarity regarding the channels selected for the different subjects also deserves great attention, especially because the subject specificity in terms of spatial and temporal features of EEG is well known [8]. It could be expected to find the channels in some large scalp area common to all subjects, but in the presented results they share some very specific similarities that go beyond this. For instance, the electrode O2 was the reference in all 40 channels (10 per subject), and all the other electrodes were in the posterior part of the coronal plane as well. Moreover, whenever an electrode beyond the P line (towards Nz) was selected, it was necessarily at the right hemisphere, in agreement with the contra-lateral activation for left-hand motor tasks. Furthermore, there were 4 channels in common for all subjects (O1-O2, PO2-O2, P5-O2, and P2-O2). Despite these similarities regarding the channels, and some similarities across the signals of the different subjects (e.g. negative peak between $t = 3s$ and $t = 4s$), the signals showed notable subject-specific features, such as different energy range

and overall waveforms. This suggests that even if the cortical processes are similar for all subjects and if they are detected at the same scalp locations, the way they appear in the EEG signals can be considerably different.

Another point that deserves to be noted is the fact that the results suggest that the main information in the selected channels is probably related to visual processing in the occipital cortex. This is in conformity with the experimental paradigm and with the scientific literature [21], [22]. However, we also expected to find movement-related components, which did not seem to be the case. It could be argued that O2 is actually performing the role of a "silent" reference and the other electrodes, closer the the motor cortex and slightly shifted to the contra-lateral hemisphere, are the ones really providing the signals. This is a seductive hypothesis, especially because of the negative peak time-related to the probable initiation of the MI by the subjects, and because it would be a better result at a first sight aiming at a motor rehabilitation BCI. Nevertheless, we believe this is not the case for two reasons: 1) the reference electrode test previously mentioned [11] that puts O2 in the "main" electrode position, and 2) despite the electrodes being slightly closer to the motor cortex, only 10 electrodes out of the top 10 channels for all subjects are actually beyond line P towards the motor cortex. In fact, these 10 electrodes might contain motor-related signals contributing to generate the negative peak in the signals after the cue presentation. This does not mean that there was no significant motor ERD, but only that the visual processing by the occipital cortex carried more information in terms of variance. For instance, maybe it would be possible to find principal components with less variance and with their most relevant channels concentrated in the motor cortex, thus reflecting the motor ERD more purely.

V. Conclusion

In this work we proposed a strategy to find channels that maximize the information from EEG in an MI cue-based paradigm using the IPRE approach. A publicly available dataset was used, with data from four subjects, from which only the left-hand imagery trials were analyzed. A Total Score metric was presented as a criterion to rank the channels separately for each subject. All combinations of pairs of electrodes were considered. Our findings suggested that the electrode position O2 is the one with most task-related information, probably due to visual processing of the stimuli on the computer screen, instead of motor cortex activity. The latter, however, also seems to contribute with ERD information in the hemisphere contra-lateral to the MI. This holds for all subjects. However, although results are interesting, they must be considered with caution due to the limited subject sample. Future studies are needed with a larger group of subjects and diverse paradigms to confirm the presented hypothesis.

Acknowledgment

This research was supported by São Paulo Research Foundation (FAPESP), Grants 2013/07559-3 (BRAINN), 2023/12422-9 and 2024/00998-6 (CRISQuaM); Brazilian National Council for Scientific and Technological Development (CNPq), Grants 306255/2024-3 and 408389/2024-9 (INCT-NeuroComp); Funding Authority for Studies and Projects (FINEP), Grant 2140/22 (DReANNs); and Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001.

References

[1] W. H. Elashmawi, A. Ayman, M. Antoun, H. Mohamed, S. E. Mohamed, H. Amr, Y. Talaat, and A. Ali, "A comprehensive review on brain–computer interface (bci)-based machine and deep learning algorithms for stroke rehabilitation," 7 2024.

[2] J. Cantillo-Negrete, R. I. Carino-Escobar, E. Ortega-Robles, and O. Arias-Carrión, "A comprehensive guide to bci-based stroke neurorehabilitation interventions," *MethodsX*, vol. 11, p. 102452, 12 2023.

[3] Z. Guo, Q. Gao, Y. Jiang, H. Jiang, and N. Jiang, "Therapeutic effects of brain-computer interface on motor recovery of stroke patients: A meta-analysis," *medRxiv*, pp. 1–30, 2023. [Online]. Available: https://doi.org/10.1101/2023.04.11.23288439

[4] I. K. Niazi, M. S. Navid, U. Rashid, I. Amjad, S. Olsen, H. Haavik, G. Alder, N. Kumari, N. Signal, D. Taylor, D. Farina, and M. Jochumsen, "Associative cued asynchronous bci induces cortical plasticity in stroke patients," *Annals of Clinical and Translational Neurology*, vol. 9, pp. 722–733, 2022.

[5] J. Peksa and D. Mamchur, "State-of-the-art on brain-computer interface technology," *Sensors (Basel, Switzerland)*, vol. 23, 2023.

[6] J. J. Vidal, "Toward direct brain-computer communication," *Annual Review of Biophysics and Bioengineering*, vol. 2, pp. 157–180, 1973.

[7] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris, "An eeg-based brain-computer interface for cursor control," *Electroencephalography and Clinical Neurophysiology*, vol. 78, pp. 252–259, 1991.

[8] G. C. de Melo, G. Castellano, and A. Forner-Cordero, "A procedure to minimize eeg variability for bci applications," *Biomedical Signal Processing and Control*, vol. 89, 3 2024.

[9] P. F. G. de Vazquez, C. A. S. Filho, G. C. de Melo, A. Forner-Cordero, and G. Castellano, "Reproducibility analysis of functional connectivity measures for application in motor imagery bcis," *Biomedical Signal Processing and Control*, vol. 85, 8 2023.

[10] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*, 2nd ed. Oxford University Press, 2006.

[11] J. R. Wolpaw and E. W. Wolpaw, *Brain-computer Interfaces: principles and practice*, 2012.

[12] O. Bertrand, F. Perrin, and J. Pernier, "A theoretical justification of the average reference in topographic evoked potential studies," *Electroencephalography and clinical neurophysiology*, pp. 462–464, 1985.

[13] D. Yao, "A method to standardize a reference of scalp eeg recordings to a point at infinity," *Physiological Measurement*, vol. 22, pp. 693–711, 2001.

[14] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and Clinical Neurophysiology*, vol. 72, pp. 184–187, 1989.

[15] J. Kayser and C. E. Tenke, "Issues and considerations for using the scalp surface laplacian in eeg/erp research: A tutorial review," pp. 189–209, 9 2015.

[16] G. C. de Melo, A. Forner-Cordero, and G. Castellano, "The role of the reference electrode in eeg recordings: looking from an inverted perspective," *Biomedical Physics and Engineering Express*, vol. 11, 9 2025.

[17] D. Gwon, K. Won, M. Song, C. S. Nam, S. C. Jun, and M. Ahn, "Review of public motor imagery and execution datasets in brain-computer interfaces," 2023.

[18] B. Blankertz, G. Dornhege, M. Krauledat, K. R. Müller, and G. Curio, "The non-invasive berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, pp. 539–550, 8 2007.

[19] G. Pfurtscheller and A. Aranibar, "Evaluation of event-related desynchronization (erd) preceding and following voluntary self-paced movement 1," Tech. Rep., 1979.

[20] G. Pfurtscheller and F. H. L. D. Silva, "Event-related eeg/meg synchronization and desynchronization: Basic principles," *Clinical Neurophysiology*, vol. 110, pp. 1842–1857, 1999.

[21] I. A. Shevelev, N. B. Kostelianetz, V. M. Kamenkovich, and G. A. Sharaev, "Eeg alpha-wave in the visual cortex: check of the hypothesis of the scanning process," Tech. Rep., 1991.

[22] B. H. Kwon, B. H. Lee, J. H. Cho, and J. H. Jeong, "Decoding visual imagery from eeg signals using visual perception guided network training method," in *International Winter Conference on Brain-Computer Interface, BCI*, vol. 2022-February. Institute of Electrical and Electronics Engineers Inc., 2022.

[23] E. Donchin, "A multivariate approach to the analysis of average evoked potentials," Tech. Rep., 1966.

[24] M. Alessandrini, G. Biagetti, P. Crippa, L. Falaschetti, S. Luzzi, and C. Turchetti, "Eeg-based alzheimer's disease recognition using robust-pca and lstm recurrent neural network," *Sensors*, vol. 22, 5 2022.

[25] G. A. Buzzell, Y. Niu, S. Aviyente, and E. Bernat, "A practical introduction to eeg time-frequency principal components analysis (tf-pca)," *Developmental Cognitive Neuroscience*, vol. 55, 6 2022.

[26] L. Gu, J. Jiang, H. Han, J. Q. Gan, and H. Wang, "Recognition of unilateral lower limb movement based on eeg signals with erp-pca analysis," *Neuroscience Letters*, vol. 800, 3 2023.

[27] T. T. Munia and S. Aviyente, "Multivariate analysis of bivariate phase-amplitude coupling in eeg data using tensor robust pca," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1268–1279, 2021.

[28] F. Scharf, A. Widmann, C. Bonmassar, and N. Wetzel, "A tutorial on the use of temporal principal component analysis in developmental erp research – opportunities and challenges," *Developmental Cognitive Neuroscience*, vol. 54, 4 2022.

Motion-Aware Signal-Level Evaluation of In-Ear Photoplethysmography Using a Commercial Earable

Mateus Armond Freire
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
mateusfreire@alunos.utfpr.edu.br

Lorena Zynger Capaverde
Academic Department of Electrotechnical (DAELT)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
lcapaverde@alunos.utfpr.edu.br

Wemerson Delcio Parreira
Faculty of Electrical Engineering, Polytechnic School
Pontifical Catholic University of Campinas (PUC – Campinas)
Campinas, Brazil
wemerson.delcio@puc-campinas.edu.br

Fabio Pires Itturriet
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
fabioitturriet@utfpr.edu.br

Renata Coelho Borges
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
renatacoelho@utfpr.edu.br

Abstract*—Wearable photoplethysmography (PPG) has enabled continuous cardiovascular monitoring outside clinical environments; however, motion artifacts remain a major source of signal degradation, particularly in consumer-grade earable devices. This study presents a signal-level characterization of in-ear photoplethysmography (earPPG) acquired from a commercial earable and evaluates a lightweight motion-artifact rejection strategy to improve heart rate (HR) estimation. EarPPG and reference electrocardiogram (ECG) signals were simultaneously acquired in a controlled single-subject protocol comprising resting conditions and four motion scenarios: walking, horizontal head movement, vertical head movement, and mastication. Signals were synchronized, filtered, and segmented into analysis windows, while motion contamination was quantified using a moving standard deviation metric with a threshold derived from resting-condition variability. Heart rate estimation performance was evaluated before and after rejecting motion-contaminated windows using mean absolute error (MAE), root mean square error (RMSE), Bland-Altman analysis, and Pearson correlation. The proposed strategy reduced global MAE from 11.67 BPM to 5.97 BPM and RMSE from 19.41 BPM to 10.78 BPM, while reducing estimation bias from -10.11 BPM to -4.87 BPM. The largest improvements were observed during head-motion conditions, with error reductions of 58.47% and 69.89% for horizontal and vertical movements, respectively. These findings demonstrate that lightweight signal-quality screening can substantially improve HR estimation from commercial earPPG recordings acquired under motion, providing a practical foundation for future validation studies involving larger and more diverse populations.

***Index Terms*— earable, photoplethysmography, heart rate estimation, motion artifacts.**

The authors would like to acknowledge the Graduate Program in Biomedical Engineering (PPGEB), Federal University of Technology — Paraná (UTFPR), Fundação Araucária and Pontifical Catholic University of Campinas (PUC-Campinas) for their support.

I. Introduction

The increasing adoption of wearable technologies has expanded the possibilities for continuous physiological monitoring beyond conventional clinical environments [1], [2]. Among the sensing modalities employed in wearable systems, photoplethysmography (PPG) has become one of the most widely adopted techniques due to its low cost, non-invasive operation, and suitability for compact devices [3]. By measuring optical variations associated with pulsatile blood volume changes, PPG enables the estimation of cardiovascular parameters such as heart rate (HR), heart rate variability (HRV), and peripheral vascular dynamics [4], [5].

Despite its widespread use, wearable PPG acquisition remains susceptible to motion artifacts, variations in sensor placement, changes in optical coupling, low perfusion, and environmental interference [6]. These disturbances may distort the waveform morphology, introduce false pulse detections, and degrade the accuracy of downstream physiological estimations. Consequently, robust signal-quality assessment and computationally efficient motion-aware processing remain important challenges in wearable biomedical signal processing.

Recently, earable devices have emerged as a promising platform for physiological sensing by integrating optical and multimodal sensors into consumer electronics such as wireless earbuds and hearing aids [7]. Compared with conventional wrist-worn devices, the auricular region provides a mechanically stable measurement location while remaining comfortable for prolonged daily use. In addition, the widespread adoption of consumer earbuds and hearing aids creates opportunities for

unobtrusive and continuous physiological monitoring using devices that are already incorporated into everyday routines [8]. These characteristics have motivated increasing interest in in-ear photoplethysmography (earPPG) as a practical sensing modality for wearable cardiovascular monitoring.

Although the availability of commercial earable platforms has increased considerably [9], the characteristics of the raw physiological signals provided by these devices remain comparatively less explored than those obtained from dedicated research prototypes. A better understanding of signal behavior under representative operating conditions is essential for the development of reliable signal-processing algorithms and wearable health applications.

A. Related Works on EarPPG and Motion Artifact Assessment

The feasibility of in-ear physiological monitoring has been demonstrated by several studies using dedicated research prototypes and custom-designed sensing platforms. Bubb et al. [10] investigated earable-based physiological sensing and highlighted the suitability of the ear canal for unobtrusive cardiovascular monitoring, emphasizing its mechanical stability and reduced susceptibility to upper-limb motion. Benkert et al. [11] evaluated the influence of different motion conditions on earPPG acquisition, demonstrating that head movements substantially affect signal morphology and heart-rate estimation accuracy.

More recently, research has shifted toward commercially available earable devices. Liu et al. [12] proposed EarFusion, a quality-aware framework that combines in-ear audio and photoplethysmography signals to improve heart-rate estimation by exploiting complementary sensing modalities available in earables. In parallel, Doherty et al. [13] validated the heart-rate measurements provided by AirPods Pro 3 during graded treadmill exercise, demonstrating the growing potential of consumer-grade earbuds as physiological monitoring platforms.

Beyond earable-specific research, motion artifact mitigation has been extensively investigated in wearable PPG systems. Previous studies have proposed adaptive filtering, machine-learning algorithms, and signal-quality assessment techniques to detect or suppress motion-related distortions in photoplethysmographic signals [14], [15]. Among these approaches, lightweight quality assessment methods are particularly attractive for wearable and edge-computing applications because they require limited computational resources while allowing unreliable signal segments to be identified before physiological estimation [16].

Although previous investigations have demonstrated the feasibility of earPPG acquisition, most studies have focused on prototype platforms or on the final performance of physiological estimation algorithms. Comparatively fewer studies have characterized the raw signals provided by commercially available earables under controlled but representative motion conditions. Likewise, the effectiveness of simple signal-quality metrics for identifying motion-contaminated earPPG segments remains insufficiently explored.

In this context, this work presents an exploratory case study based on a single healthy volunteer to characterize raw earPPG signals acquired from a commercially available earable. Rather than providing population-level validation, the objective is to investigate the signal behavior obtained from a consumer-grade platform under controlled motion scenarios. Specifically, this study aims to: (i) characterize raw earPPG signal behavior during different motion conditions; (ii) quantify the impact of motion artifacts on heart-rate estimation; and (iii) evaluate a lightweight motion-aware rejection strategy based on temporal signal variability. The proposed framework provides an initial signal-level assessment that may support the development of more robust physiological monitoring algorithms for future earable applications.

The remainder of this article is organized as follows. Section II introduces the fundamentals of in-ear photoplethysmography, while Section III describes the experimental protocol, data acquisition procedures, signal processing pipeline, motion artifact detection strategy, and evaluation methods. Section IV presents the experimental results obtained under different motion conditions. Section V discusses the implications and limitations of the proposed approach. Finally, Section VI summarizes the main conclusions and future research directions.

II. In-ear Photoplethysmography (earPPG)

Photoplethysmography is a non-invasive optical sensing technique widely used for cardiovascular monitoring by measuring temporal variations in light intensity associated with pulsatile blood volume changes in microvascular tissue [4], [17]. In a typical PPG system, a light-emitting diode (LED) illuminates the tissue while a photodetector measures the intensity of the reflected or transmitted light. The acquired waveform consists of a pulsatile alternating component (AC), produced by cardiac-driven blood volume variations, superimposed on a slowly varying direct component (DC), associated with static tissue properties, venous blood, and sensor-tissue coupling [5].

In earable devices, the PPG sensor is positioned inside the external auditory canal, where the surrounding tissue exhibits favorable vascularization and relatively stable mechanical coupling [18]. Compared with conventional wrist-based configurations, this anatomical location is less affected by upper-limb movements and may provide improved signal stability during daily activities [7], [10]. In the present study, the green optical channel (530 nm) was employed, as this wavelength is widely adopted in wearable PPG systems due to its high sensitivity to pulsatile blood volume variations.

Despite these anatomical advantages, earPPG acquisition remains susceptible to motion-induced disturbances. Head rotations, mastication, variations in insertion depth, changes in contact pressure, and tissue deformation may alter the optical coupling between the sensor and surrounding tissue, producing amplitude fluctuations, baseline variations, and waveform distortions. Such artifacts may compromise pulse detection and, consequently, the estimation of cardiovascular parameters.

For these reasons, reliable physiological monitoring using earPPG requires appropriate signal preprocessing, quality assessment, and artifact mitigation strategies. The methodology adopted in this work focuses on evaluating raw earPPG signals acquired from a commercial earable and investigating whether a lightweight signal-quality metric can identify motion-contaminated segments before heart-rate estimation.

III. Methodology

Figure 1 summarizes the complete processing workflow adopted in this study. Starting from raw signal acquisition, the proposed pipeline performs preprocessing, motion-aware quality assessment, heart rate estimation, and quantitative performance evaluation. Each stage is detailed in the following subsections.

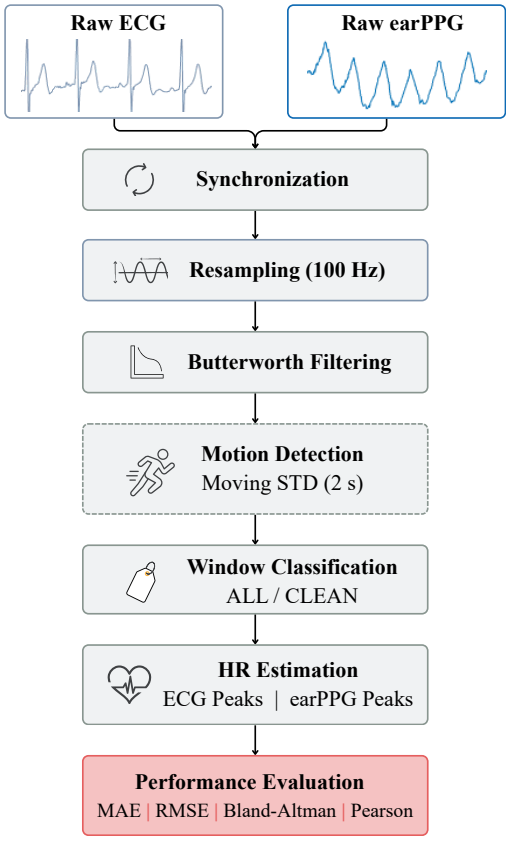


Fig. 1: Overview of the proposed signal processing pipeline for motion-aware heart rate estimation from earPPG signals.

A. Earable Platform

The experimental acquisitions were performed using the OmniBuds Research Platform (OmniBuds Ltd., Cambridge, UK; Figure 2), a sensory earable designed for physiological monitoring and research applications [19]. Originating from Nokia Bell Labs [20], OmniBuds is a commercially available research platform for in-ear physiological monitoring. The device integrates kinetic, acoustic, thermal, and optical sensing modalities, enabling continuous real-time monitoring of multiple vital signs with up to six hours of battery life. For research applications, the platform provides access to raw sensor streams via a dedicated Python API over Bluetooth Low Energy (BLE), allowing unprocessed data to be exported to an external host for offline analysis. This capability was the primary feature exploited in this study.

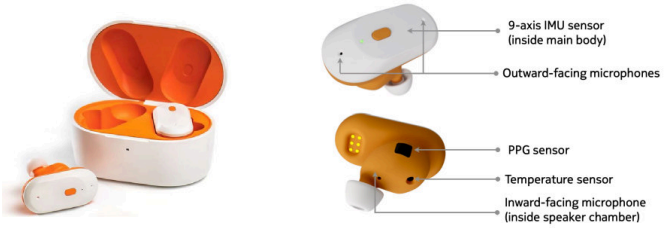


Fig. 2: The OmniBuds Research Platform [20].

B. Participant and Data Acquisition

The acquisitions were conducted with a single healthy adult volunteer in a controlled laboratory environment under constant artificial light. The wearable device was positioned in the right external auditory canal according to manufacturer instructions. Raw PPG signals were exported directly from the device at a sampling frequency of 100 Hz and streamed in real time to a desktop computer for continuous data logging. The green LED channel (530 nm) was employed in this study, and the LED drive current was set to 31 mA, which controls the optical emission intensity.

Recording sessions followed a structured 80-second protocol, divided into a tri-phasic sequence: an initial 30 s rest period, a 20 s movement interval, and a final 30 s recovery rest. This design captured natural variability alongside motion-induced perturbations. The volunteer performed four motion tasks (Figure 3), representing distinct profiles relevant to everyday auricular wearable use:

- Scenario (a): walking at 100 steps per minute (daily physical activity);
- Scenario (b): horizontal head shaking at 70 repetitions per minute (axial head rotation);
- Scenario (c): vertical head shaking at 70 repetitions per minute (sagittal head rotation);
- Scenario (d): mastication through chewing gum (repetitive facial muscle activity).

To ensure temporal consistency and standardized task execution, a metronome was used to pace Scenarios (b)-(d). In total, 20 motion recordings were acquired in addition to baseline resting recordings. Each motion scenario was repeated five times. Simultaneously, a chest-worn heart rate monitor (Polar H10, Polar Electro Oy) was used as a cardiac reference signal [21]. No physiological manipulation, calibration procedures, or signal conditioning were applied during acquisition. The study focused exclusively on signal-level characterization rather than physiological intervention or diagnostic evaluation.

An additional acquisition was performed entirely at rest to validate the earPPG measurements, as illustrated in Figure 4. Figure 4(a) shows synchronized ECG and earPPG recordings

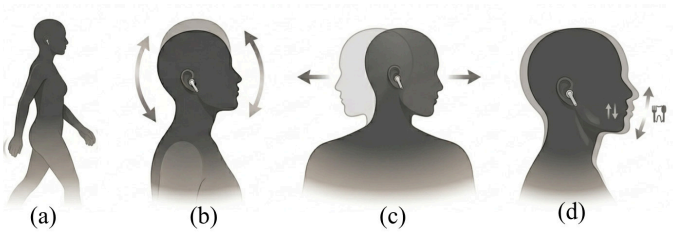
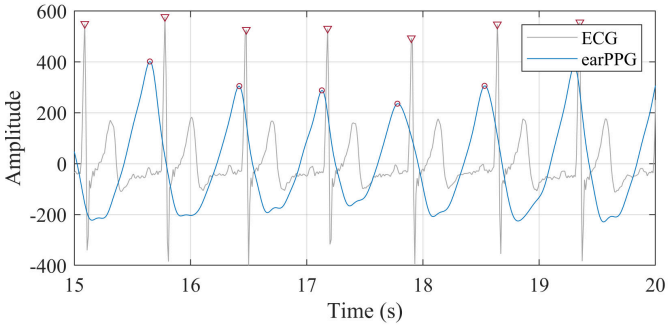
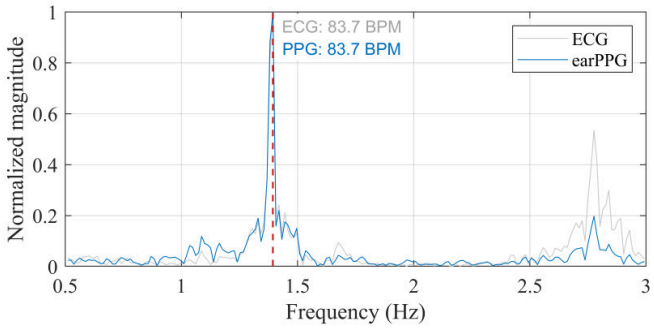


Fig. 3: Representative acquisition scenarios used for structured characterization of in-ear multimodal signals. (a) walking; (b) horizontal head motion; (c) vertical head motion; and (d) chewing.

in the time domain, where the pulsatile earPPG waveform follows the cardiac cycles identified by the ECG R-peaks, despite the expected temporal delay associated with pulse transit time. Figure 4(b) presents the corresponding normalized frequency spectra, showing that both signals exhibit the same dominant frequency component (83.7 BPM), confirming preservation of the cardiac pulsatility in the earPPG signal.



(a) Time-domain ECG and earPPG waveforms.



(b) Corresponding frequency-domain spectra demonstrate preservation of cardiac pulsatility in the in-ear PPG signal.

Fig. 4: Representative synchronized ECG and earPPG recordings acquired during the resting condition.

C. Signal Processing

The raw ECG and earPPG signals were independently acquired and exported as timestamped data streams. Since both sensors operated with distinct internal clocks and potentially different native sampling frequencies, all signals were temporally synchronized prior to analysis. Duplicated timestamps were removed to avoid interpolation inconsistencies. Occasional discontinuities in the earPPG stream caused by Bluetooth packet losses were automatically identified through temporal gap detection. A gap was defined whenever the interval between consecutive PPG timestamps exceeded 1.5 times the nominal sampling interval. Missing segments were reconstructed through linear interpolation using neighboring valid samples.

The common valid time interval between ECG and earPPG recordings was then determined by selecting the overlapping acquisition period shared by both modalities. Subsequently, both signals were resampled onto a uniform temporal grid at 100 Hz using piecewise cubic Hermite interpolation (PCHIP), ensuring synchronized sample-wise alignment while preserving waveform morphology. To eliminate unstable acquisition periods commonly observed at recording initialization and termination, the first and last 10 seconds of each recording were discarded prior to analysis.

Digital filtering was then applied independently to each modality. ECG signals were processed using a fourth-order Butterworth bandpass filter with cutoff frequencies between 0.5 Hz and 40 Hz to preserve the cardiac QRS complex while attenuating baseline wander and high-frequency noise. EarPPG signals were filtered using a fourth-order Butterworth bandpass filter between 0.5 Hz and 8 Hz, preserving pulsatile cardiovascular components while suppressing attenuating baseline fluctuations and high-frequency noise [22].

D. PPG Signal-Quality Screening

Motion contamination was quantified by computing a moving standard deviation (STD) over a 2-second sliding window across the filtered earPPG signal. Let $x[n]$ denote the earPPG signal and L the number of samples corresponding to the 2-second window. The local STD sequence is computed as

$$s[n] = \sqrt{\frac{1}{L-1}\sum_{k=0}^{L-1}\left(x[n-k]-\bar{x}_n\right)^2}, \tag{1}$$

where $\bar{x}_n$ denotes the local mean over the same window. The moving STD characterizes short-term temporal variability in the earPPG waveform, allowing localized increases in signal variance associated with motion artifacts to be identified [23].

The filtered earPPG signal was subsequently segmented into 10-second analysis windows with 50% overlap (i.e., a 5-second hop size). For each analysis window W_i, a scalar motion metric was defined as the average value of the moving STD,

$$M_i = \frac{1}{N}\sum_{n\in W_i} s[n], \tag{2}$$

where N is the number of samples contained in the corresponding analysis window.

The decision threshold was estimated exclusively from the resting-condition recordings as

$$T = \text{median}(M_{\text{rest}}) + 3\ \text{MAD}(M_{\text{rest}}), \tag{3}$$

where M_{rest} denotes the set of motion metric values obtained from resting recordings and MAD is the median absolute deviation [23].

Each analysis window was then classified according to

$$C_i = \begin{cases} \text{CLEAN}, & M_i < T, \\ \text{MOTION}, & M_i \geq T, \end{cases} \tag{4}$$

where C_i denotes the quality label assigned to the i-th analysis window.

Thus, the short-term variability quantified over successive 2-second windows is summarized into a single motion metric for each 10-second analysis window, enabling motion-contaminated segments to be identified prior to heart-rate estimation.

E. Heart Rate Estimation

Heart rate (HR) was independently estimated from both ECG and earPPG signals using peak interval analysis applied within each 10-second analysis window. For ECG processing, cardiac peaks were detected using a peak-finding algorithm based on local prominence, with a minimum peak distance of 0.4 seconds (corresponding to a maximum detectable heart rate of 150 BPM). The prominence threshold was defined as the standard deviation of the ECG signal within each analysis window. RR intervals were computed as the time differences between consecutive detected peaks, and heart rate was then estimated from the average RR interval within each window.

Similarly, systolic peaks were detected in the earPPG signal using the same minimum peak distance criterion, with a prominence threshold defined as the standard deviation of the earPPG signal within each window. Pulse-to-pulse intervals were obtained from consecutive detected peaks and used to compute the corresponding heart rate estimate.

The heart rate estimation error (HRE) for each window was defined as the absolute difference between the earPPG-derived and ECG-derived heart rate values, computed independently for each analysis segment.

F. Performance Evaluation

The accuracy of the earPPG heart rate estimation was evaluated against the reference cardiac signal acquired using a Polar H10 chest strap both before (ALL windows) and after (CLEAN windows) the application of the proposed motion-aware rejection strategy. Performance was quantified using the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), bias, and standard deviation (SD) of the estimation error. While MAE reflects the average absolute difference between earPPGand ECG-derived heart rates, RMSE emphasizes larger deviations associated with motion-induced artifacts. Bias and SD were used to characterize the central tendency and dispersion of the estimation error distribution, respectively. Because all recordings were obtained from a single participant and the analysis windows overlapped, these metrics were treated as descriptive measures of within-session and within-subject performance rather than as population-level estimates.

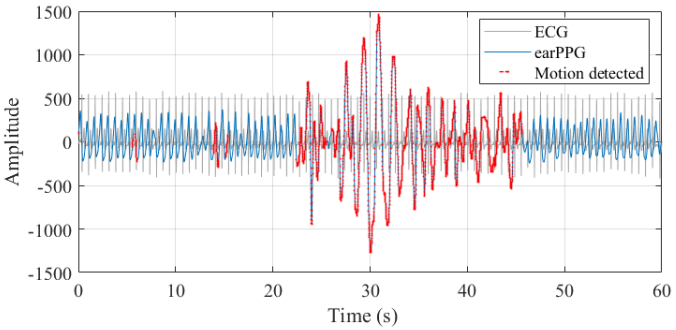
Agreement between both measurement modalities was further assessed through Bland-Altman analysis and the corresponding limits of agreement (LoA) [24]. In addition, the Pearson correlation coefficient was computed between the proposed motion metric and the absolute heart rate estimation error across all analysis windows in order to investigate the relationship between motion intensity and estimation degradation.

G. Ethical Considerations

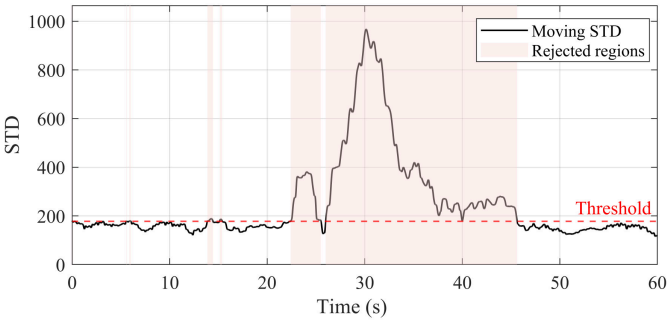
This study was conducted as a self-experiment involving a single participant. The protocol consisted of non-invasive physiological signal acquisition using a commercial earable device and a reference ECG system.

IV. Results

In the evaluated recordings, application of the proposed signal-quality screening reduced the descriptive earPPG heart rate estimation error across the investigated conditions. Figure 5 shows a representative segment containing motion artifacts (Scenario (b)), where increased motion intensity induces substantial amplitude fluctuations in the earPPG waveform, successfully captured by the moving standard deviation metric.



(a) ECG and earPPG signals with motion artifact segments.



(b) Moving STD with rejected regions above the threshold.

Fig. 5: Motion artifact detection in physiological signals.

Table I summarizes the heart rate estimation error obtained before (ALL windows) and after (CLEAN windows) motion-based rejection, together with the percentage of rejected analysis windows. Error reduction was observed for all evaluated

motion conditions. The largest improvements were obtained during head movement conditions, with relative error reductions of 58.47% and 69.89% for horizontal and vertical head motion, respectively. Walking also presented a measurable improvement of 21.98%, while chewing-related motion produced a smaller but consistent reduction of 12.24%. Resting segments already exhibited low estimation error, resulting in only minor improvement after rejection.

TABLE I: Heart rate estimation error before (ALL windows – AW) and after (CLEAN windows – CW) motion rejection.

Condition	AW Error (BPM)	CW Error (BPM)	Improvement (%)	Removed Windows (%)
Baseline	1.79	1.75	2.14	4
Scenario a	13.27	10.36	21.98	20
Scenario b	22.11	9.19	58.47	48
Scenario c	16.89	5.09	69.89	42
Scenario d	4.28	3.76	12.24	44

The percentage of removed windows varied according to the motion profile. Walking produced the lowest rejection rate among dynamic conditions (20%), whereas repetitive head movements and chewing generated higher rejection rates ranging from 42% to 48%. Interestingly, chewing resulted in a relatively high rejection rate (44%) despite a modest reduction in HR estimation error, suggesting that facial movements affect waveform morphology without proportionally degrading pulse interval estimation. These findings indicate that the proposed variability-based metric is sensitive to distinct motion-induced perturbation patterns affecting earPPG acquisition.

Figure 6 presents the distribution of heart rate estimation errors across all motion conditions for both the original (No Rejection) and motion-filtered (Motion-Rejected) cases. Beyond the reduction in average error reported in Table I, the boxplot reveals a consistent decrease in variability, interquartile range, and outlier occurrence after motion rejection. These results indicate that, within the evaluated recordings, the screening strategy was associated with lower descriptive error and reduced error variability after exclusion of low-quality windows.

Global performance metrics demonstrated improved HR estimation performance after motion-aware screening. As summarized in Table II, the Mean Absolute Error (MAE) decreased from 11.67 BPM to 5.97 BPM, while the Root Mean Square Error (RMSE) was reduced from 19.41 BPM to 10.78 BPM after excluding motion-contaminated windows. Likewise, the estimation bias decreased from -10.11 BPM to -4.87 BPM, and the standard deviation of the error decreased from 16.60 BPM to 9.64 BPM. These improvements indicate not only greater accuracy but also a more consistent HR estimation process, with fewer large deviations caused by motion artifacts.

Bland–Altman analysis showed narrower descriptive limits of agreement after motion-aware rejection (Figure 7). The distribution of the differences between earPPG-derived and reference ECG heart rates became more concentrated around the mean bias, with visibly narrower limits of agreement.

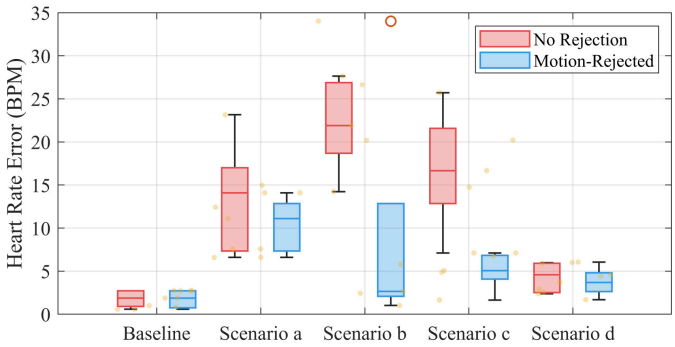


Fig. 6: Distribution of heart rate estimation error across motion conditions for No Rejection and Motion-Rejected cases. Points represent individual analysis windows; markers outside whiskers indicate outliers.

Consistent with the values reported in Table II, the limits of agreement were reduced from -42.65 to 22.43 BPM in the ALL windows condition to -23.77 to 14.03 BPM after rejection. Moreover, the number of extreme disagreement points was substantially reduced, indicating that the proposed screening strategy effectively removed the windows most affected by motion artifacts while preserving the remaining physiological information.

Pearson correlation analysis presented a moderate positive association between the proposed motion metric and absolute heart rate estimation error ($r = 0.342$). Given the single-subject design and overlapping analysis windows, this correlation is interpreted descriptively rather than as population-level statistical evidence. This result suggests that increased temporal variability is associated with higher HR estimation errors, supporting the use of the proposed metric for motion-aware window screening.

V. Discussion

This exploratory study demonstrated that motion-induced variability substantially affects in-ear PPG signals and downstream heart-rate estimation performance in consumer-grade earables. The findings provide an initial feasibility-oriented assessment of a lightweight signal-quality strategy and establish a methodological basis for future validation in larger and more heterogeneous populations.

Motion artifacts substantially degraded earPPG-based heart rate estimation, consistent with previous studies reporting that wearable PPG signals are highly susceptible to motion-induced corruption during ambulatory conditions [14], [15], particularly during dynamic activities such as walking and repetitive head movements. Despite its simplicity, the proposed moving-STD approach was effective in identifying segments associated with increased estimation error without requiring auxiliary inertial sensors or computationally intensive artifact-removal techniques. Because the method relies exclusively on short-term temporal variability, it may offer a computationally lightweight alternative for signal-quality assessment in

TABLE II: Global performance metrics for earPPG-based heart rate estimation (BPM) before and after motion-aware rejection.

Condition	Bias	SD	LoA Lower	LoA Upper	RMSE	MAE
ALL windows	-10.11	16.60	-42.65	22.43	19.41	11.67
CLEAN windows	-4.87	9.64	-23.77	14.03	10.78	5.97

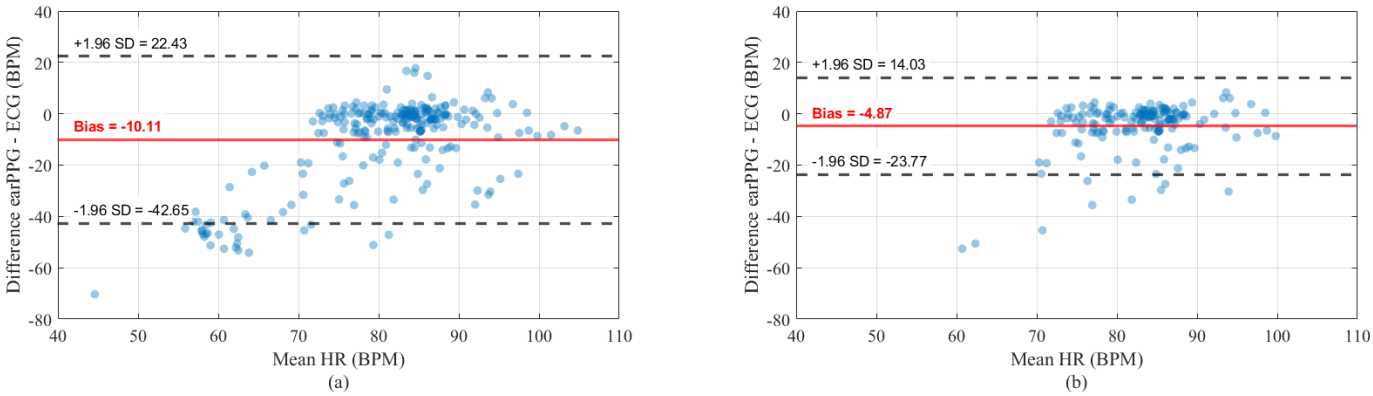


Fig. 7: Bland-Altman analysis comparing earPPG-derived and reference heart rate estimates before (a) and after motion-aware rejection (b).

resource-constrained wearable systems [16]. However, its generalizability and robustness to inter-subject variability remain to be established.

A relevant aspect of the proposed rejection strategy is the trade-off between signal quality improvement and temporal data coverage. Conditions associated with stronger motion contamination, such as repetitive head movements, achieved the largest reductions in heart rate estimation error but also required rejecting a substantial proportion of analysis windows (up to 48%). In contrast, walking exhibited more moderate rejection rates while still producing meaningful error reduction. This behavior suggests that the proposed approach can adaptively preserve reliable physiological segments while discarding highly unstable intervals, which may be advantageous for edge-based remote monitoring systems prioritizing data reliability over continuous transmission.

Interestingly, mastication produced relatively high rejection rates despite comparatively small heart rate estimation errors, suggesting that certain facial-motion patterns substantially alter waveform morphology without fully compromising cardiac periodicity estimation. This behavior is consistent with prior evidence that facial and jaw movements introduce significant biomechanical disturbances in in-ear sensing modalities [25]. This observation reinforces the complexity of motion-related interference in earPPG acquisition and highlights the need for motion-specific artifact characterization strategies.

Considering the growing intersection between audiological care and cognitive health, embedding lightweight physiological-monitoring algorithms directly into hearing aids or consumer-grade true wireless stereo earbuds may enable diverse digital health applications. For the general and working-age population, continuous earPPG monitoring during daily routines could support longitudinal stress assessment and autonomic function monitoring through heart rate variability analytics [26]. For elderly populations, this framework could potentially extend cardiovascular safety monitoring, such as opportunistic screening for asymptomatic atrial fibrillation associated with ischemic stroke risk, without requiring adherence to additional dedicated medical devices.

Although the proposed approach significantly improved agreement with the reference cardiac signal, several limitations must be acknowledged. First, the study was conducted using recordings from a single healthy volunteer, limiting generalizability across different anatomical ear canal structures, skin properties, and cardiovascular conditions. Second, the proposed method performs artifact rejection rather than true artifact correction, meaning that highly corrupted segments are discarded instead of reconstructed. Consequently, excessive rejection rates may reduce temporal data coverage during highly dynamic physical activities. Future work should include validation in larger and more heterogeneous populations, direct comparison with IMU-based artifact-detection strategies, and investigation of interoperability with standardized health data frameworks, further bridging the gap between consumer earables and scalable digital health systems.

VI. Conclusion

This study presented an initial evaluation of a computationally efficient, motion-aware framework for improving heart rate (HR) estimation from consumer-grade earPPG signals. By identifying motion-contaminated segments through temporal variability analysis, the proposed strategy demonstrated, within this single-subject setting, that a lightweight algorithm can substantially improve HR estimation accuracy relative to a ref-

erence cardiac signal without introducing significant computational overhead. The proposed framework additionally reduced the amount of motion-corrupted data requiring downstream transmission and processing, supporting its suitability for low-power edge deployment in wearable telemonitoring systems.

Although restricted to basic HR extraction, these findings highlight the feasibility of performing signal-quality screening directly on resource-constrained wearable devices. Because the current approach relies on rejecting corrupted segments rather than correcting them, future implementations may benefit from integrating the device's embedded inertial sensors (IMU), which were acquired but not fully exploited in the present analysis, to automatically identify and classify specific motion profiles such as walking or mastication.

The combination of optical and inertial sensing modalities provides a promising pathway for adaptive signal-quality assessment and future heart rate variability (HRV) analysis in auricular wearables. The observed signal-quality screening behavior also suggests potential applicability to remote monitoring scenarios; however, computational cost, energy consumption, data-transmission requirements, and clinical performance were not evaluated in the present study. Therefore, broader studies involving multiple participants, independent recording sessions, and real-world monitoring conditions are required to establish the generalizability and practical applicability of the proposed approach.

Artificial Intelligente (AI) Usage Statement

The authors declare the use of a generative AI tool solely for language editing. The AI had no role in data analysis, interpretation, or scientific conclusions. The authors are fully responsible for the content of this manuscript.

References

[1] M. D. Tegegne, S. R. Niakan Kalhori, P. Haas, V. M. Sobotta, J. Warnecke, and T. M. Deserno, "Wearable devices for remote monitoring of chronic diseases: Systematic review," *JMIR mHealth and uHealth*, vol. 14, p. e74071, 2026.

[2] Y. Zhang, X. T. Zheng, X. Zhang, J. Pan, and A. V.-Y. Thean, "Hybrid integration of wearable devices for physiological monitoring," *Chemical Reviews*, vol. 124, no. 18, pp. 10386–10434, 2024.

[3] L. Kane, D. Ball, K. R. Martin, and D. Powell, "Continuous heart rate variability monitoring—understanding patterns of stress and recovery and their relationship with self-reported burnout, resilience and well-being in doctors: a protocol for a sequential explanatory mixed-methods study," *BMJ open*, vol. 15, no. 6, p. e097849, 2025.

[4] A. Jubran, "Pulse oximetry," *Critical Care*, vol. 19, p. 272, 2015.

[5] J. W. Severinghaus, "History and advancements in pulse oximetry," *Respiratory Care*, vol. 52, no. 4, pp. 509–522, 2007.

[6] E. D. Chan, M. M. Chan, and M. M. Chan, "Pulse oximetry: understanding its basic principles facilitates appreciation of its limitations," *Respiratory Medicine*, vol. 107, no. 6, pp. 789–799, 2013.

[7] Y. Du, J. Li, Y. Zhao, S. Tong, L. Wang, and J. Wang, "Health monitoring with earables: A survey," *ACM Transactions on Internet of Things*, vol. 7, no. 2, pp. 1–36, 2026.

[8] L. Cribb, M. Moreno-Betancur, M. P. Pase, R. Wolfe, C. Britt, Z. Zhou, R. C. Shah, G. Rance, K. M. Sheets, T. T.-J. Chong, *et al.*, "Treating hearing loss with hearing aids for the prevention of cognitive decline and dementia," *Neurology*, vol. 106, no. 3, p. e214572, 2026.

[9] C. Hu, Q. Yang, Y. Liu, T. Röddiger, K.-J. Butkow, M. Ciliberto, A. L. Pullin, J. Stuchbury-Wass, M. Hassan, C. Mascolo, *et al.*, "A survey of earable technology: Trends, tools, and the road ahead," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 10, no. 2, pp. 1–41, 2026.

[10] C. A. B. Bubb, M. Weber, N. Kretsch, R. Heim, I. Zellhuber, S. Schmid, *et al.*, "Wearable in-ear pulse oximetry validly measures oxygen saturation between 70% and 100%: A prospective agreement study," *Digital Health*, vol. 9, p. 20552076231211169, 2023.

[11] A. Benkert, J. Bludau, L. Boborzi, S. Prueckner, and R. Schniepp, "Impact of helicopter vibrations on in-ear ppg monitoring for vital signs—mountain rescue technology study (moretech)," *Sensors*, vol. 26, no. 1, p. 324, 2026.

[12] Y. Liu and A. Montanari, "Earfusion: Quality-aware fusion of in-ear audio and photoplethysmography for heart rate monitoring," in *Proceedings of the 3rd ACM International Workshop on Intelligent Acoustic Systems and Applications*, pp. 31–36, 2025.

[13] C. Doherty, D. Burke, O. Mitchell, D. Lawless, J. Brar, M. Fuchs, and R. Lambe, "Accuracy of heart rate measurement using airpods pro 3 during graded treadmill exercise: A laboratory-based validation study," *PLOS Digital Health*, vol. 5, no. 4, p. e0001373, 2026.

[14] S. Ismail, U. Akram, and I. Siddiqi, "Heart rate tracking in photoplethysmography signals affected by motion artifacts: A review," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, p. 5, 2021.

[15] D. Pollreisz and N. TaheriNejad, "Detection and removal of motion artifacts in ppg signals," *Mobile Networks and Applications*, vol. 27, no. 2, pp. 728–738, 2022.

[16] K. T. Putra, A. Z. Arrayyan, N. Hayati, C. Damarjati, A. Bakar, H.-C. Chen, *et al.*, "A review on the application of internet of medical things in wearable personal health monitoring: A cloud-edge artificial intelligence approach," *Ieee Access*, vol. 12, pp. 21437–21452, 2024.

[17] K. Lodewyk, M. Wiebe, L. Dennett, J. Larsson, A. Greenshaw, and J. Hayward, "Wearables research for continuous monitoring of patient outcomes: A scoping review," *PLOS Digital Health*, vol. 4, no. 5, p. e0000860, 2025.

[18] A. Ferlini, A. Montanari, C. Min, H. Li, U. Sassi, and F. Kawsar, "In-ear ppg for vital signs," *IEEE Pervasive Computing*, vol. 21, no. 1, pp. 65–74, 2021.

[19] OmniBuds Ltd., "OmniBuds Research Platform." https://www.omnibuds.tech/solutions/research, 2024. Accessed: April 2026.

[20] A. Montanari, A. Thangarajan, K. Al-Naimi, A. Ferlini, Y. Liu, A. N. Balaji, and F. Kawsar, "Omnibuds: a sensory earable platform for advanced bio-sensing and on-device machine learning," *arXiv preprint arXiv:2410.04775*, 2024.

[21] R. Gilgen-Ammann, T. Schweizer, and T. Wyss, "Rr interval signal quality of a heart rate monitor and an ecg holter at rest and during exercise," *European journal of applied physiology*, vol. 119, no. 7, pp. 1525–1532, 2019.

[22] M. Yalcin and M. E. Latoschik, "End-to-end non-invasive ecg signal generation from ppg signal: a self-supervised learning approach," *Frontiers in Physiology*, vol. 17, p. 1694995, 2026.

[23] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of experimental social psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[24] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[25] A. Montanari, A. Ferlini, A. N. Balaji, C. Mascolo, and F. Kawsar, "Earset: A multi-modal dataset for studying the impact of head and facial movements on in-ear ppg signals," *Scientific data*, vol. 10, no. 1, p. 850, 2023.

[26] S. Massaro, T. Y. Kang Yang, and G. Christopoulos, "Wearable devices methodology: opportunities and challenges in human resource management," *Human Resource Management Journal*, vol. 36, no. 1, pp. 158–177, 2026.

Comparison of Multi-Class and Multi-Label Techniques for Gesture Classification Based on sEMG Signals

1st Gabriel da Rocha Adamoski
Graduate Program in Biomedical Engineering
Federal University of Technology - Paraná (UTFPR)
Curitiba, Brazil

2nd José Jair Alves Mendes Junior
Graduate Program in Electrical and Computer Engineering
Federal University of Technology - Paraná (UTFPR)
Curitiba, Brazil

3rd Daniel Prado Campos
Graduate Program in Biomedical Engineering
Federal University of Technology - Paraná (UTFPR)
Apucarana, Brazil

***Abstract*—Accurate recognition of movement intentions using surface electromyography (sEMG) is crucial for the development of intuitive and functional myoelectric prostheses. This paper presents a comparative study of Multi-Class and Multi-Label classification approaches for recognizing these intentions. Using the NinaPro DB3 database, the performance of the Random Forest algorithm for identifying 8 functional gestures was evaluated. Notably, this study utilizes data from upper-limb amputees, which presents a significantly harder challenge for classification algorithms compared to intact subjects due to altered muscular anatomy and signal variability. The methodology was based on extracting time-domain features from muscular contraction windows. The results indicated the superiority of the holistic Multi-Class approach, which achieved an average accuracy of 66.5%, compared with the anatomical Multi-Label approach. To mitigate spurious posture predictions in the Multi-Label model, which analyzes 7 joint groups in isolation, a biomechanical restriction algorithm based on the Hamming Distance was implemented, raising the system's accuracy from 47.9% (exact match) to 59.3%. The analysis revealed high performance variability among subjects, attributable to physiological factors and inherent sEMG signal degradation, reinforcing the need for locally customized training models for each volunteer. It is concluded that the Multi-Class model presents greater predictive robustness due to its simplified logic.**

***Index Terms*—sEMG, Gesture Classification, Upper Limb Prosthetics, Machine Learning, Random Forest, Multi-Class, Multi-Label.**

I. Introduction

Surface electromyography (sEMG) is used to record muscle electrical activity in clinical and research applications, such as controlling prostheses, exoskeletons, and other assistive devices, primarily due to its non-invasive nature [1] [2] [3]. However, the accuracy and robustness of myoelectric control systems are often impaired by factors such as muscle fatigue, sweating, and electrode placement [4].

To improve integration with prosthetic hardware, Machine Learning algorithms have been employed as powerful alternatives for multichannel sEMG control [5]. Machine Learning is a branch of Artificial Intelligence in which the model is trained to make decisions [6].

Recent comprehensive reviews have mapped the evolution of these algorithms in sEMG pattern recognition, highlighting a significant transition from classical feature engineering to advanced Deep Learning architectures [7], [8]. These surveys emphasize that while traditional shallow classifiers, such as Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDAs), established the foundation for myoelectric control, contemporary research focuses heavily on Convolutional Neural Networks (CNNs) and hybrid models.

However, beyond the choice of the underlying algorithm, the fundamental way the classification problem is structured plays a critical role in prosthesis control. Traditionally, pattern recognition relies on a single-label, or Multi-Class, approach. Previous studies have widely utilized this formulation with sEMG and EEG data [9], [10], treating each movement as a single, holistic entity. While straightforward, this paradigm forces the system to operate within a closed set of predefined gestures, meaning every sample must belong to one unique, isolated class.

In contrast, the Multi-Label approach recognizes that natural hand gestures require the simultaneous activation of multiple muscle groups and motor units. In this paradigm, a single sEMG sample can be associated with a set of independent labels [11]. By decomposing muscle activity into independent, joint-level components, this approach aims for a composite decoding representation that more closely reflects natural biomechanics [12], [13]. For prosthetic control, this offers a significant advantage: the system is not locked into a rigid dictionary of gestures, and if a single target joint is misclassified, the execution of the overall movement is not entirely compromised [11].

To evaluate the practical implications of these structural differences, this work presents a comparative analysis between

This work was supported by a scholarship from Fundação Araucária.

the holistic Multi-Class approach and the anatomically inspired Multi-Label approach for the recognition of movement intentions.

II. Materials and Methods

To effectively evaluate and compare the Multi-Class and Multi-Label classification approaches for movement intention recognition, a structured experimental pipeline, illustrated in Figure 1, was developed. The methodology encompasses four main stages. First, a suitable sEMG database containing data from upper-limb amputees was selected to ensure the study addresses a realistic and challenging clinical scenario. Following data selection, signal processing and the extraction of time-domain features from muscular contraction windows were performed. The functional gestures were then mapped into both the holistic single-label representation and the anatomical multi-label representation. Finally, the Random Forest algorithm was implemented to train both classification models, followed by the application of a biomechanical restriction algorithm to refine the Multi-Label predictions. The following subsections detail each phase of this procedure.

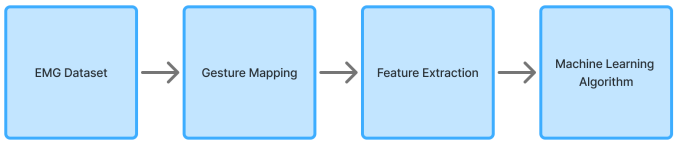


Fig. 1. Methodology flowchart, from the dataset acquisition to the implementation of the Machine Learning Algorithm.

A. Data

The data used in this study were extracted from the NinaPro DB3 database, widely recognized in the literature on myoelectric prosthesis control. This database contains labelled sEMG signals collected from individuals with amputation, in contrast to other databases that include only able-bodied-person data. The samples have diverse characteristics in age, years since amputation, and percentage of remaining forearm [14]. During the acquisition procedure, the subject moved in response to the displayed gesture, while 12 Delsys electrodes recorded sEMG data at 2 kHz [14]. The relevant characteristics of the selected data for our analysis are:

- 11 subjects with upper limb amputation;
- 8 sEMG electrode channels positioned on the forearm. Although DB3 contains 12 electrodes, only the first 8 were selected because they form an equally spaced dense array around the radio-humeral joint, which is the standard configuration for capturing the main synergistic muscle activation patterns while optimizing computational cost and avoiding redundant features from the remaining sparse electrodes;
- 17 total movements recorded in the protocol, of which 8 functional gestures were selected for classification (Figure 2); and
- 6 physical repetitions for each movement, interleaved by neutral rest periods.

In addition to the raw sEMG signals, the original annotation vectors from the database, referred to as *restimulus* (ranging from 0 to 17, indicating the gesture) and *rerepetition* (ranging from 0 to 6, indicating the repetition number), were used as reference to temporally map the onset and offset of each executed muscle contraction.

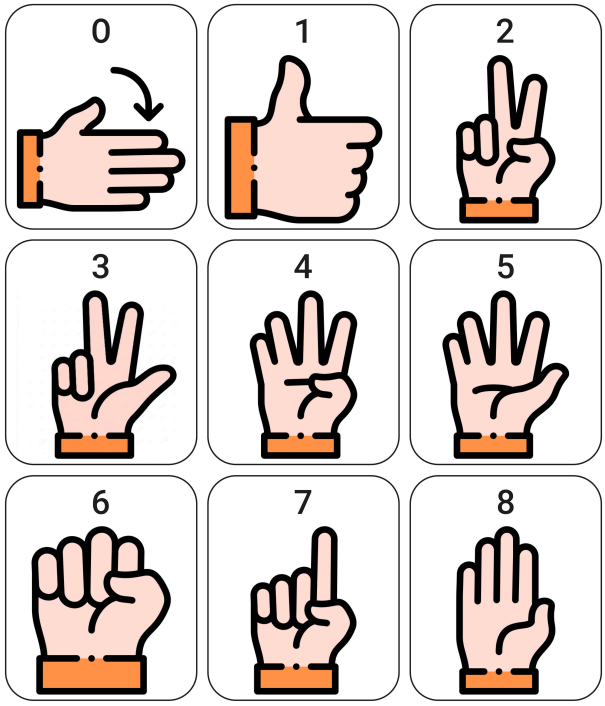


Fig. 2. Gestures of the first eight movements in the Ninapro DB3 and the rest of the gestures labelled.

B. Gesture Mapping

For implementing the Multi-Label classifier, the eight movements extracted from the database were decomposed into discrete joint states. Instead of handling the gesture as a single class, the adopted anatomical model divided the hand into seven independent joint groups: Thumb, Index, Middle, Ring, Little, Fingers, representing combined abduction and adduction, and Carpal, for the carpo-metacarpal movement of the thumb.

Each of these seven joints was mapped to three possible kinematic states for movement execution: Extension (E), Flexion (F), or Neutral (N), as shown in Table I. Thus, the subject's final movement intention is mathematically represented by the simultaneous vector combination of these isolated joints.

C. Feature Extraction

The raw sEMG signal is not directly fed into the classification algorithm. To reduce data size and optimize the

TABLE I
HAND GESTURE CLASSIFICATION, WHERE E IS RELATED TO FINGER EXTENSION, F TO FINGER FLEXION, AND N THE FINGER MAINTAINING AS NEUTRAL POSITION.

Gesture	Description	Thumb	Index	Middle	Ring	Little	Fingers	Carpal
0	Hand at rest	N	N	N	N	N	N	N
1	Thumb extended	E	F	F	F	F	N	E
2	Index and middle extended	F	E	E	F	F	E	F
3	Index, middle and ring extended	E	E	E	F	F	E	E
4	Fingers extended, thumb in palm	F	E	E	E	E	E	E
5	Hand fully open	E	E	E	E	E	E	E
6	Closed fist	F	F	F	F	F	N	N
7	Index finger pointing	E	E	F	F	F	N	F
8	All fingers extended and adducted	E	E	E	E	E	N	E

recognition of muscle activation patterns, the signal is segmented to extract time-domain mathematical features. Instead of using highly redundant overlapping sliding windows, a single continuous processing window was extracted per physical contraction. To ensure temporal consistency and avoid zero-padding, the window size was dynamically defined for each subject based on the duration of their shortest valid contraction execution. From these tailored windows, time-domain mathematical features were extracted. Previous studies [15]–[17] established that features such as Mean Absolute Value (MAV), Root Mean Square (RMS), Waveform Length (WL), and Zero Crossing (ZC) perform robustly with sEMG signals, while other reviews [18] highlight the Teager-Kaiser Energy Operator (TKEO) as a method to enhance gesture classification.

Considering x_i as the amplitude of the sEMG signal at the i-th sample and N as the total number of samples in the analysis window, the following features were extracted individually for each of the eight channels:

- **Mean Absolute Value (MAV):** Provides an estimate of the average muscle contraction amplitude over the time window.

$$MAV = \frac{1}{N}\sum_{i=1}^{N}|x_i|. \tag{1}$$

- **Root Mean Square (RMS):** Models the continuous power of the signal, serving as a robust indicator of motor activation force.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}. \tag{2}$$

- **Waveform Length (WL):** Measures the complexity and cumulative amplitude variation of the signal, capturing information from both amplitude and frequency.

$$WL = \sum_{i=1}^{N-1}|x_{i+1} - x_i|. \tag{3}$$

- **Zero Crossing (ZC):** Counts the number of times the signal amplitude crosses the zero axis (algebraic sign change), serving as an indirect estimate of the signal's frequency properties in the time domain.

$$ZC = \sum_{i=1}^{N-1}\mathbb{I}\{sgn(x_i) \neq sgn(x_{i+1})\}. \tag{4}$$

- **Teager-Kaiser Energy Operator (TKEO):** Computes the instantaneous energy of the signal, highlighting rapid contraction peaks and increasing the distinction between active signal and noise.

$$TKEO = \frac{1}{N-2}\sum_{i=2}^{N-1}(x_i^2 - x_{i-1}\cdot x_{i+1}). \tag{5}$$

By applying these five equations to the eight electrodes, each movement window is mathematically described by a 40-dimensional feature vector (five features $\times$ eight channels), which serves as the definitive input to the Machine Learning model. Although features such as RMS and TKEO have intrinsically different amplitude scales compared to ZC, feature normalization (e.g., Z-score) was deliberately bypassed. This is because the selected classification algorithm, Random Forest, relies on independent threshold-based recursive partitioning, making it strictly scale-invariant and immune to feature magnitude disparities.

D. Multi-Class Classification

The Multi-Class approach represents the traditional method of pattern recognition in myoelectric prostheses, in which each movement intention is treated as a single and global class. In this architecture, the model's objective is to correlate the synergistic muscle activation pattern captured by the electrodes with a single gesture identifier.

The feature matrix X, with dimensions 48×40, acts as the system input. The corresponding target variable is the column vector y, with dimensions 48×1, containing integer labels from one to eight, representing the exact ID of the executed functional gesture.

Among the possible Machine Learning models, Random Forest was chosen due to its robust performance in processing sEMG signals [19]–[21]. As an ensemble classifier, Random Forest operates by constructing a multitude of decision trees during training. This architecture is particularly advantageous for biological signals because it inherently mitigates the

overfitting commonly associated with noisy, high-dimensional data—such as the 40-dimensional feature vectors used in this study.

The algorithm was implemented via *TreeBagger* in MATLAB, configured with an ensemble of 100 decision trees. To ensure reproducibility, default hyperparameters for classification were maintained: the minimum leaf size was set to 1, and the number of predictors sampled at each node split was set to the square root of the total number of features (approximately 6). No artificial data balancing techniques (e.g., SMOTE) were required, as the experimental protocol inherently ensures a balanced class distribution through an equal number of physical repetitions per gesture. The model evaluates all trees simultaneously and uses majority voting to select the final class. This approach has the implicit advantage of learning forearm muscle synergies, as it evaluates the total context of the sEMG signal to determine hand movement as a single block.

E. Multi-Label Classification

In contrast to the holistic Multi-Class model, the Multi-Label approach proposes the isolated prediction of the anatomical states that compose the hand. As defined in the gesture mapping, the desired output is no longer a single scalar identifier but rather a matrix Y with dimensions 48×7. Each of the seven columns of this matrix represents a joint group (Thumb, Index, Middle, Ring, Little, Fingers, and Carpal), and the values contained therein indicate the kinematic state of the joint: extension, flexion, or neutral.

Instead of training a single global classifier, seven independent Random Forest models were instantiated, one for each joint group, all receiving the same normalized input matrix X (48×40) and sharing the same hyperparameters defined in the Multi-Class model.

During the testing phase, the seven models analyze the same time window of the sEMG signal and emit their predictions in parallel, without any communication or weight sharing between them. The final result is a prediction matrix of dimension 1×7, which reconstructs the estimated state of the entire hand at that instant.

F. Error Correction via Hamming Distance

The main limitation of the Multi-Label approach with independent models is the absence of biomechanical constraints. Since there is no communication between the joint classifiers, the system is susceptible to predicting spurious combinations of joint states that do not correspond to any of the eight real gestures mapped in the protocol.

To mitigate this issue and force the model to converge toward anatomically valid postures, a fallback algorithm based on Hamming Distance was implemented. When the system composes the 1×7 prediction vector from the individual model outputs, it simultaneously compares it against a reference lookup table containing the perfect definitions of the eight functional gestures.

The Hamming Distance (D_H) is calculated as the number of positions at which the predicted vector (v_p) differs from the vector of a valid gesture (v_g), expressed by:

$$D_H(v_p, v_g) = \sum_{j=1}^{7} \mathbb{I}\{v_p(j) \neq v_g(j)\} \tag{6}$$

where j represents the joint index and $\mathbb{I}$ is the indicator function that returns "1" in the case of divergence (error) and "0" in the case of agreement.

The algorithm counts the total number of joint errors for each prediction relative to the known classes and replaces the spurious combination with the identifier of the real gesture that presents the smallest D_H, which is the closest gesture. In tie cases, where the predicted vector is equally distant from two or more valid gestures, the algorithm adopts the first-ranked class as a tiebreaker criterion.

G. Model Training and Validation Protocol

To ensure a robust and unbiased evaluation of the classifiers, an intra-subject analysis was adopted. Due to the high inter-subject variability inherent to sEMG signals—caused by differences in muscle anatomy, sensor positioning, and varying levels of amputation—a generalized global model is often ineffective. Therefore, independent Machine Learning models were trained and tested specifically for each of the 11 subjects.

The evaluation procedure followed a Leave-One-Repetition-Out Cross-Validation (LORO-CV) methodology. Because the experimental protocol consists of six independent repetitions per functional gesture, a 6-fold cross-validation strategy was implemented based strictly on these physical repetitions rather than random data splitting.

In each iteration of the cross-validation cycle, five repetitions were aggregated to construct the training set, while the remaining unseen repetition was isolated to form the testing set. This loop was iterated six times until every repetition had served as the test set exactly once. Structuring the validation around full physical repetitions, rather than shuffled time windows, is crucial in sEMG pattern recognition; it actively prevents data leakage between highly correlated adjacent temporal samples and guarantees that the reported accuracy reflects the model's true generalization capacity on novel motor executions.

III. Results and Discussion

This section presents the empirical results obtained from the comparative analysis between the Multi-Class and Multi-Label classification approaches. First, the overall predictive performance of both models is evaluated, highlighting the critical impact of the biomechanical constraint algorithm. Subsequent discussions analyze the underlying causes of the observed accuracy metrics, addressing the inherent biological challenges of myoelectric pattern recognition in the target demographic.

A. Overall Performance: Multi-Class vs. Multi-Label

The model evaluation demonstrated that the traditional Multi-Class approach achieved a higher average accuracy of 66.5%, with a standard deviation of 17.3%. As illustrated in Figure 3, the Multi-Label model in its pure form—which requires an exact match of all seven joint states simultaneously—achieved a lower average accuracy of 47.9%.

The application of biomechanical constraints via the Hamming Distance algorithm proved strictly necessary for the viability of the anatomical approach. Correcting spurious postures raised the average accuracy of the Multi-Label system to 59.3%, with a standard deviation of 16.2%, representing an absolute performance gain of 11.4 percentage points (a 23.8% relative gain) (Figure 3).

To formally evaluate this performance difference, a paired statistical test was conducted. Given the sample size ($n = 11$) and the lack of guaranteed normality, a Wilcoxon signed-rank test was utilized. The test confirmed that the performance difference between the Multi-Class approach and the Multi-Label approach with Hamming distance was statistically significant ($p = 0.007$, effect size $r = 0.81$), substantiating the superiority of the holistic model.

The lower performance of the Multi-Label method compared to the Multi-Class method is primarily justified by the greater mathematical complexity of the problem, which requires the simultaneous convergence of seven independent classifiers rather than predicting a single global label. Furthermore, it is critical to contextualize these absolute accuracy values within the clinical reality of the dataset. Because the signals were acquired from upper-limb amputees, the classification algorithms faced significantly degraded sEMG inputs. Factors such as muscle atrophy, altered anatomical synergies, and variable electrode contact due to residual limb anatomy inherently limit the maximum achievable accuracy when compared to controlled studies conducted on intact subjects. This highlights the severe challenge of translating standard pattern recognition algorithms to realistic clinical scenarios.

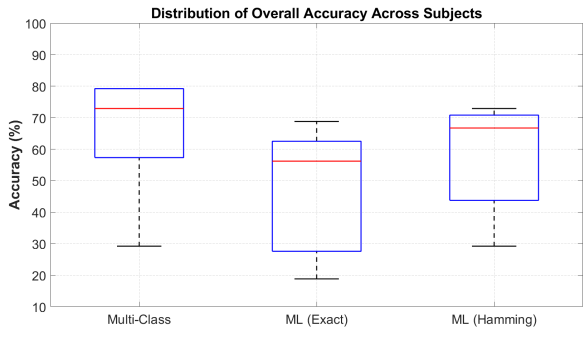


Fig. 3. Distribution of overall sample accuracy. The boxplot highlights the superior performance of the Multi-Class method and the significant performance gain of the Multi-Label model after applying the Hamming Distance correction algorithm.

B. sEMG Signal Variability

A notable characteristic in the results of both approaches is the high standard deviation, driven by a sharp accuracy drop in subjects 05, 07, and 10, as illustrated in Figure 5. In the Multi-Class model, for example, subject 07 recorded only 29.2% accuracy, in stark contrast to subjects that reached nearly 80%.

However, evaluating myoelectric pattern recognition in amputees solely through global accuracy can mask the system's true functional viability. An overall accuracy of 29.2% does not necessarily imply a complete failure of the classifier; rather, it often reflects a severe biological inability to separate specific synergistic patterns due to missing musculature. To properly diagnose the source of these errors, it is necessary to move beyond global accuracy and analyze the specific error distribution. Confusion matrices provide this transparency, allowing for the identification of exactly which gestures lack separable feature representations and which remain functionally robust despite the amputation (Figure 4).

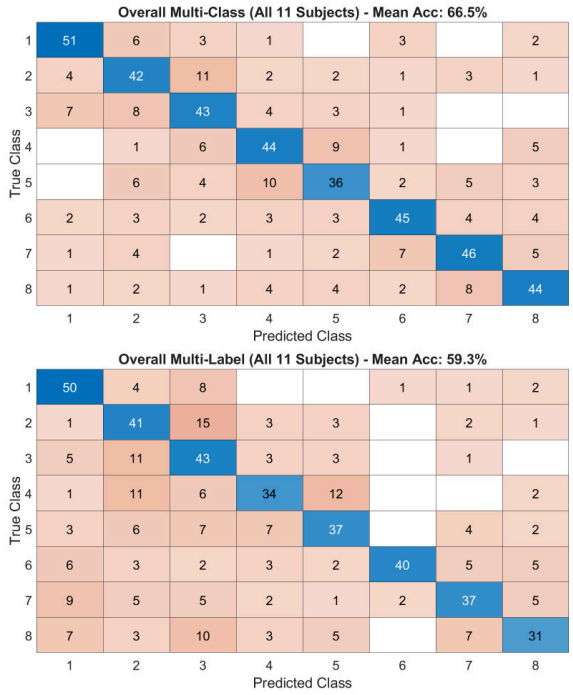


Fig. 4. Overall confusion matrices comparing the Multi-Class and Multi-Label classification approaches. The matrices illustrate the global error distribution across all evaluated subjects.

A direct investigation of the data confirmed that this performance degradation does not stem from flaws in the mapping or algorithm parameterization, but is an inherent consequence of the quality of the sEMG signal captured from these specific individuals. Physiological factors common in amputee databases, such as variations in residual limb volume, skin impedance, or muscle fatigue during the protocol, are reflected in signals with lower SNR, hindering pattern recognition regardless of the classifier architecture.

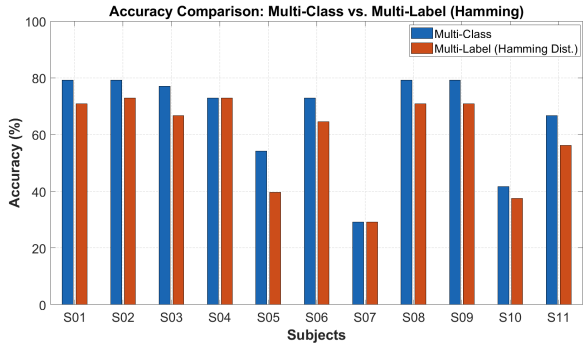


Fig. 5. Direct per-subject accuracy comparison between the Multi-Class and Multi-Label approaches, highlighting the pronounced performance degradation in subjects S05, S07, and S10.

C. Performance Analysis by Joint Group

The main methodological advantage of the Multi-Label approach is the ability to isolate the accuracy rate per anatomical joint. The detailed analysis of the prediction matrix revealed significant discrepancies in the model's ability to identify different components of the hand.

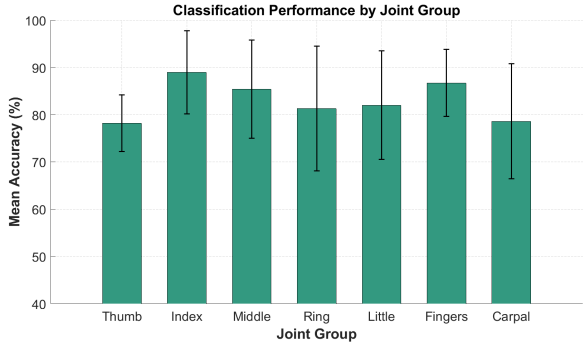


Fig. 6. Average individualized classification accuracy per joint group in the Multi-Label model. Error bars represent the standard deviation across the 11 evaluated subjects.

The movements associated with the Index and Middle fingers presented the highest recognition rates, with average accuracies of 89.0% and 85.4%, respectively. In contrast, the joints that most contributed to the system's overall error were the Thumb and Carpal. The Thumb achieved an average accuracy of 78.2% (with an average of 10.5 errors on incorrectly classified instances), while the Carpal reached 78.6% (with an average of 10.3 errors).

This behavior has biomechanical foundations. Thumb kinematics and radiocarpal movements involve more complex or deep muscle bundles. Surface electrodes positioned along the forearm readily capture the activation of the long flexor and extensor muscles of the central fingers, but present technical limitations in isolating the intricate synergies responsible for carpal rotation and thumb abduction/opposition.

IV. Conclusion

This study presented a comparative analysis between Multi-Class and Multi-Label approaches for classifying movement intentions using sEMG signals from amputee subjects. The results demonstrated that the Multi-Class model achieved better overall performance, benefiting from a simpler, more straightforward predictive architecture. In contrast, the Multi-Label approach, although offering more detailed anatomical mapping, showed greater susceptibility to failures. The requirement to predict multiple joint states in an isolated and simultaneous manner created a broader margin for cumulative errors, which penalized the system's final accuracy even with the application of biomechanical constraint algorithms.

The analysis also highlighted the strong impact of biological and anatomical differences between subjects. Three specific subjects exhibited severe degradation in signal recognition, significantly reducing the study's overall average. This variability, inherent in sEMG data, supports the premise that local classification approaches, where the model is customized and trained individually for each user, make more sense and yield more reliable results than adopting a single global classifier.

Despite the theoretical consistency of the evaluated models, the current methodology presents practical limitations. The deployment of the developed technique in commercial myoelectric prostheses is not ideal due to the extensive processing window required for gesture recognition (which equaled the full duration of the shortest contraction), which introduces unacceptable delays in natural motor control.

Therefore, as proposals for future work, it is suggested to optimize the selection and extraction of mathematical features within the algorithms, as well as the transition to reduced and overlapping time windows. Such improvements are fundamental to balancing the classifier's accuracy with the stringent latency requirements of real-time control of bionic limbs.

References

[1] P. D. H. Bustos, R. R. Serrezuela, A. A. S. Leon, A. E. R. Gomez, and D. M. E. Suaza, "Electromyographic emg signal recognition of hand gestures for multi-class prostheses based on dwt and cnn," in *2024 IEEE VII Congreso Internacional en Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil (AmITIC)*. IEEE, 2024, pp. 1–7.

[2] N. Mehta, B. Islam, and E. A. Clancy, "Machine learning-based estimation of emg baseline noise standard deviation without rest trials," *Journal of Electromyography and Kinesiology*, p. 103097, 2025.

[3] W. N. Abdelrazik, A. El-Bialy, H. A. Ibrahim, and B. A. Hemade, "Enhancing emg-based gesture recognition for prosthetic control: Impact of dimensionality reduction on machine learning classifier," *Sensors and Actuators A: Physical*, p. 117120, 2025.

[4] I. Kyranou, K. Szymaniak, and K. Nazarpour, "Emg dataset for gesture recognition with arm translation," *Scientific data*, vol. 12, no. 1, p. 100, 2025.

[5] B. Abdikenov, D. Zholtayev, K. Suleimenov, N. Assan, K. Ozhikenov, A. Ozhikenova, N. Nadirov, and A. Kapsalyamov, "Emerging frontiers in robotic upper-limb prostheses: Mechanisms, materials, tactile sensors and machine learning-based emg control: A comprehensive review," *Sensors*, vol. 25, no. 13, p. 3892, 2025.

[6] T. Zaim, S. Abdel-Hadi, R. Mahmoud, A. Khandakar, S. M. Rakhtala, and M. E. Chowdhury, "Machine learning-and deep learning-based myoelectric control system for upper limb rehabilitation utilizing eeg and emg signals: a systematic review," *Bioengineering*, vol. 12, no. 2, p. 144, 2025.

[7] L. Xu, W. An, Z. Yang, Y. Li, X. Zhu, and X. Huang, "Progress of intelligent bionic hand based on myoelectric pattern recognition," in *2025 6th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM)*. IEEE, 2025, pp. 466–471.

[8] A. Shinde and V. Shete, "Advancements in myoelectric robotic hands and prosthetic limbs," in *2025 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*. IEEE, 2025, pp. 1–7.

[9] T. Kamali, R. Boostani, and H. Parsaei, "A multi-classifier approach to muap classification for diagnosis of neuromuscular disorders," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 191–200, 2013.

[10] I. Hossain, A. Khosravi, I. Hettiarachchi, and S. Nahavandi, "Multiclass informative instance transfer learning framework for motor imagery-based brain-computer interface," *Computational Intelligence and Neuroscience*, vol. 2018, p. 6323414, 2018.

[11] J. J. A. Mendes Junior, C. E. Pontim, and D. P. Campos, "Multi-label emg classification of isotonic hand movements: a suitable method for robotic prosthesis control," in *Brazilian Congress on Biomedical Engineering*. Cham: Springer, 2020, pp. 1665–1671.

[12] X. Zhang, Z. Lu, C. Fan, Y. Wang, T. Zhang, H. Li, and Q. Tao, "Compound motion decoding based on semg consisting of gestures, wrist angles, and strength," *Frontiers in Neurorobotics*, vol. 16, p. 979949, 2022.

[13] Y. Zou, L. Cheng, L. Han, Z. Li, and L. Song, "Decoding electromyographic signal with multiple labels for hand gesture recognition," *IEEE Signal Processing Letters*, vol. 30, pp. 483–487, 2023.

[14] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific Data*, vol. 1, no. 1, p. 140053, 2014.

[15] A. Phinyomark, R. N. Khushaba, and E. Scheme, "Feature extraction and selection for myoelectric control based on wearable emg sensors," *Sensors*, vol. 18, no. 5, p. 1615, 2018.

[16] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for emg signal classification," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7420–7431, 2012.

[17] I. Mohd Khairuddin, S. N. Sidek, A. P P Abdul Majeed, M. A. Mohd Razman, A. Ahmad Puzi, and H. Md Yusof, "The classification of movement intention through machine learning models: the identification of significant time-domain emg features," *PeerJ Computer Science*, vol. 7, p. e379, 2021.

[18] C. R. Carvalho, J. M. Fernández, A. J. Del-Ama, F. Oliveira Barroso, and J. C. Moreno, "Review of electromyography onset detection methods for real-time control of robotic exoskeletons," *Journal of Neuroengineering and Rehabilitation*, vol. 20, no. 1, p. 141, 2023.

[19] M. M. M. Zawawi, W. M. B. W. Daud, N. Abas, R. Sudirman, M. O. Tokhi, and M. J. Abd Latif, "Integrating random forest, cnn, and ann for emg-based upper limb movement recognition," in *2025 IEEE 13th Conference on Systems, Process & Control (ICSPC)*. IEEE, 2025, pp. 146–151.

[20] S. V. H. Salazar, J. C. Preciado, and S. A. Muñoz, "Comparative analysis of machine learning and deep learning techniques for hand gesture recognition using surface electromyography," in *2024 3rd International Congress of Biomedical Engineering and Bioengineering (CIIBBI)*. IEEE, 2024, pp. 1–6.

[21] X. Jiang, K. Nazarpour, and C. Dai, "Explainable and robust deep forests for emg-force modeling," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 6, pp. 2841–2852, 2023.

Multiclass Skin Cancer Classification Using Deep Learning Models and Grad-CAM-Based Explainability

1st Rubens A. de S. Ferreira
Computational Intelligence Laboratory
Federal University of Rural Semi-Arid
Rio Grande do Norte, Brazil
rubens.ferreira@alunos.ufersa.edu.br

2nd Rosana C. B. Rego
Department of Engineering and Technology
Federal University of Rural Semi-Arid
Rio Grande do Norte, Brazil
rosana.rego@ufersa.edu.br

***Abstract*—Skin cancer is a prevalent and potentially deadly disease that requires early detection for effective treatment. In this study, we propose a multiclass skin cancer classification model using convolutional neural networks (CNNs) and Grad-CAM-based explainability. We utilize the HAM10000 dataset, which contains 10,015 dermatoscopic images of skin lesions categorized into seven classes: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses and intraepithelial carcinoma, vascular lesions, and dermatofibroma. The proposed model achieves an accuracy of 88.42% in classifying these skin lesions, while the Grad-CAM technique provides visual explanations for the model's predictions.**

***Index Terms*—Skin Cancer Classification, Convolutional Neural Networks, Grad-CAM, Explainable AI.**

I. Introduction

Skin cancer is one of the most common types of cancer worldwide, with millions of new cases diagnosed each year [1]. Early detection and accurate classification of skin lesions are crucial for effective treatment and improved patient outcomes [2], [3]. Traditional methods for skin cancer diagnosis rely on visual examination by dermatologists, which can be subjective and prone to errors [4], [5].

Many people suffer from skin cancer, and the number of cases is expected to rise in the coming years due to factors such as increased sun exposure and aging populations [6], [7]. The World Health Organization (WHO) classifies skin cancer into three main types: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma [8]. Melanoma is the most aggressive form of skin cancer and is responsible for the majority of skin cancer-related deaths [8]. Therefore, accurate classification of skin lesions into these categories is essential for guiding treatment decisions and improving patient prognosis.

Previous studies have focused on binary classification, particularly distinguishing between melanoma and non-melanoma [9]–[12]. However, a multiclass classification approach that can differentiate among various types of skin lesions is more clinically relevant and provides more comprehensive diagnostic information.

Artificial intelligence (AI) and machine learning techniques have shown potential in improving the accuracy and efficiency of skin cancer diagnosis [13]–[16]. Convolutional neural networks (CNNs), a class of deep learning models, have been particularly successful in image classification tasks, including medical image analysis [17]. However, the lack of interpretability of CNNs has raised concerns about their adoption in clinical settings. To address this issue, we incorporate the Gradient-weighted Class Activation Mapping (Grad-CAM) technique to provide visual explanations for the model's predictions [18], enhancing the interpretability of the skin cancer classification model.

Motivated by the need for accurate and interpretable skin cancer classification, this study aims to develop a multiclass classification model using CNNs and Grad-CAM-based explainability. We evaluate the performance of the proposed model on the HAM10000 dataset, which contains a diverse set of dermatoscopic images of skin lesions. Moreover, to mitigate the class imbalance present in the dataset, we apply the Deep SMOTE technique [19] to generate synthetic samples for underrepresented classes. The results demonstrate the effectiveness of the model in classifying skin lesions into multiple categories while providing insights into the decision-making process through visual explanations. The main contributions of this study are summarized as follows:

i) Development of a CNN for the multiclass classification of seven skin lesion types using dermoscopic images from the HAM10000 dataset.
ii) Comprehensive evaluation of different data balancing strategies, including conventional oversampling and DeepSMOTE, in combination with DenseNet169, DenseNet201, EfficientNetB4, ensemble learning, and test-time augmentation (TTA) to assess their impact on classification performance.
iii) Integration of the Grad-CAM explainability technique to provide visual interpretations of the CNN predictions.

The rest of the paper is organized as follows: Section II reviews related work in skin cancer classification using machine

learning and deep learning techniques. Section III describes the dataset. Section IV presents the proposed deep learning models and the Grad-CAM explainability technique. Section V presents the results and discussion of the model's performance, while Section VI concludes the study and outlines future research directions.

II. Related Work

Some recent studies have used the HAM10000 dataset for skin cancer classification [20]–[26]. Different approaches have been proposed, including deep learning models, hybrid models, and ensemble methods. Table I summarizes the performance of various models in terms of accuracy, precision, recall, and F1-score, along with their corresponding references.

In [26], the authors proposed a computer-aided diagnosis framework for binary skin lesion classification by combining image preprocessing, data augmentation, transfer learning, and SVM-based classification. The approach employs median filtering, contrast enhancement, and edge detection, followed by fine-tuning of ResNet50 and VGG16 models. To improve performance, the final fully connected layer was replaced with an SVM classifier. In the same way, in [25], the authors proposed a stacked convolutional neural network for binary melanoma classification. The model was trained on the HAM10000 dataset and externally validated on ISIC-2020. Its performance was compared with VGG16, InceptionV3, ResNet50, EfficientNetB0, and traditional machine learning classifiers. The proposed model achieved 96% accuracy and 96% F1-score on HAM10000.

In [23], the networks AlexNet, InceptionV3, and RegNetY-320 were evaluated. RegNetY-320 achieved the best performance, reaching 91% accuracy and 88.1% F1-score, demonstrating the effectiveness of class balancing and deep architectures for skin lesion classification. In a different approach, the authors in [22] applied the DenseNet169 architecture with hair removal using the Dull Razor algorithm. The best performance was achieved by DenseNet169 with undersampling, obtaining 91.2% accuracy and 91.7% F1-score.

In the work [20], the authors proposed a novel deep learning framework called the Dual Optimization Deep Learning Network (DODL net) for skin cancer detection. Dermoscopic images are initially processed using a U-net architecture to segment the specific region of the skin lesion. Dual optimization algorithms, specifically Bacterial Foraging Optimization (BFO) and Particle Swarm Optimization (PSO), are employed to extract features from the segmented images. The DODL net is then trained on the extracted features to classify skin lesions into different categories. The model achieved an accuracy of 98.76%, demonstrating its effectiveness in skin cancer detection.

Differently, in [21], the authors proposed a novel deep learning model called SkinLesNet for the classification of skin lesions and melanoma detection. The model was trained using the Adam optimizer with a learning rate of 0.001 and evaluated on the PAD-UFES-20-Modified, HAM10000, and ISIC2017 datasets. To validate its effectiveness, the authors compared SkinLesNet with well-established architectures such as ResNet50 and VGG16. The proposed model outperformed both reference networks across all evaluated datasets, achieving an accuracy of 96% on PAD-UFES-20-Modified, 90% on HAM10000, and 92% on ISIC2017.

More recently, in [24], the authors proposed SkinWiseNet (SWNet), a deep convolutional neural network designed for binary skin cancer classification. The framework combines feature fusion across multiple datasets, including HAM10000, ISIC2019, ISIC2020, and Melanoma Skin Cancer, to improve generalization and reduce dataset-related bias.

TABLE I
Comparison of Deep Learning Approaches for Skin Cancer Classification

Model	Accuracy	Precision	Recall	F1-Score	Ref.
DODL-Net	98.76	96.02	95.37	95.32	[20]
Multi-layer CNN	90.00	89.00	87.00	85.00	[21]
DenseNet169	91.20	–	–	91.70	[22]
RegNetY-320	91.00	–	–	88.10	[23]
SkinWiseNet (SWNet)	81.93	36.00	55.73	43.79	[24]
Stacked CNN	96.00	97.00	95.00	96.00	[25]
ResNet50 and SVM	98.43	96.31	–	96.98	[26]

Based on the review of related work, it is evident that while significant progress has been made in skin cancer classification using deep learning techniques, there is still a need for models that can effectively handle multiclass classification tasks and provide interpretability. The works [24]–[26] focus primarily on binary classification. The proposed study aims to address these gaps by developing a CNN-based model for multiclass skin cancer classification and integrating Grad-CAM for explainability.

III. Dataset

The dataset used in this study is the HAM10000 dataset, which consists of 10,015 dermatoscopic images of skin lesions [27]. Fig. 1 shows sample images from the dataset. Each image is labeled with its corresponding class.

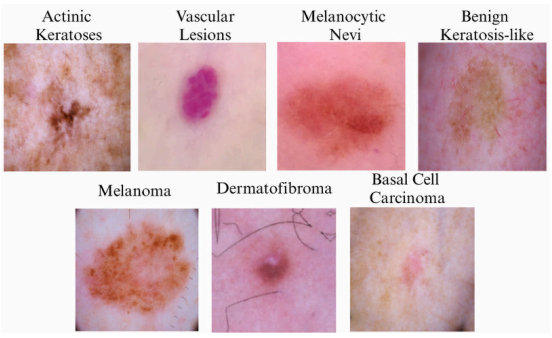


Fig. 1. Sample images from the HAM10000 dataset, showcasing the seven classes of skin lesions.

The dataset includes seven classes of skin lesions: melanocytic nevi (NV), melanoma (MEL), benign keratosis-like lesions (BKL), basal cell carcinoma (BCC), actinic keratoses and intraepithelial carcinoma (AKIEC), vascular lesions (VASC), and dermatofibroma (DF).

The class distribution is imbalanced, with 6,705 images of melanocytic nevi representing the majority class, followed by melanoma with 1,113 images and benign keratosis-like lesions with 1,099 images. The remaining classes are considerably smaller, comprising 514 basal cell carcinoma images, 327 actinic keratoses and intraepithelial carcinoma images, 142 vascular lesion images, and only 115 dermatofibroma images. This pronounced class imbalance poses a significant challenge for multiclass classification, as learning algorithms tend to become biased toward the majority classes, leading to reduced performance on underrepresented categories.

A. Data preprocessing

To mitigate class imbalance, random oversampling [28] and Deep SMOTE (Synthetic Minority Over-sampling Technique) [19] were investigated. Data augmentation strategies, including random rotations, horizontal and vertical flips, were also employed to increase the diversity of the training data and improve the model's generalization capabilities. Fig. 2 presents examples of synthetic images generated using Deep SMOTE through interpolation in the latent space learned by the autoencoder. The generated samples preserve the overall morphological characteristics of the original lesions, including their general shape, size, and color distribution.

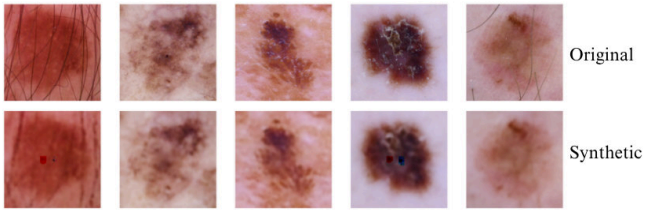


Fig. 2. Samples generated by the Deep SMOTE technique.

Before training and balancing, all images were preprocessed to reduce artifacts and highlight lesion regions. Hair removal was performed using the Dull Razor algorithm [22], which detects hair structures through morphological operations and restores the affected pixels using image inpainting. Fig. 3 illustrates the hair removal results.

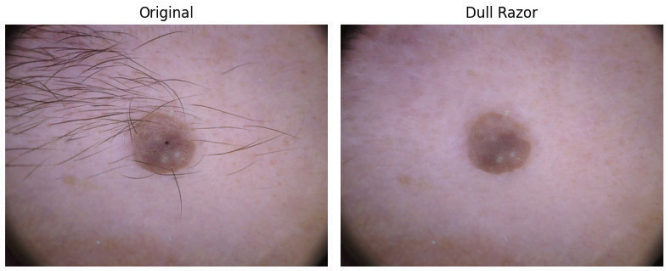


Fig. 3. Example of hair removal algorithm. The left image shows the original input, while the right image illustrates the result after hair removal and cropping around the lesion.

After the Dull Razor algorithm, the images were cropped around the lesion using the provided segmentation masks. The masks were used to localize the lesion region and perform lesion-centered cropping. For each image, the lesion boundaries were identified from the binary mask, and a bounding box with an additional 20-pixel margin was extracted.

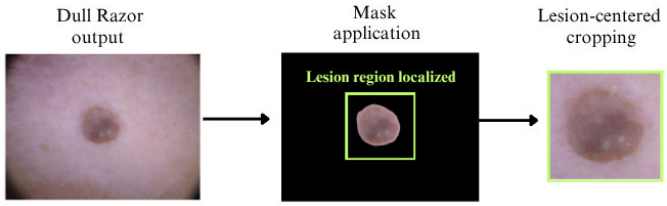


Fig. 4. Preprocessing for lesion-focused classification. The left image shows the Dull Razor result, while the right image illustrates the result after cropping around the lesion using the segmentation mask.

IV. Deep Learning Models

Different CNN architectures were evaluated for multiclass skin lesion classification, including DenseNet169, DenseNet201, and EfficientNetB4. These architectures were selected based on their proven performance in the previous work [22], [25]. The models were initialized with ImageNet pretrained weights and fine-tuned on the HAM10000 dataset. Moreover, to explain the model's predictions, the Grad-CAM technique was applied to generate class activation maps that highlight the regions of the input image that are most influential for the classification decision.

Fig. 5 illustrates the overview of the proposed experimental pipeline for multiclass skin lesion classification. The pipeline consists of five main stages: dataset collection, preprocessing, data handling and balancing, model training, and evaluation with grad-CAM-based explainability. The dataset collection stage involves acquiring dermatoscopic images from the HAM10000 dataset, which contains a diverse set of skin lesions. The preprocessing stage includes hair removal using the Dull Razor algorithm and lesion-focused cropping based on segmentation masks. The data balancing stage explores some strategies, such as oversampling and Deep SMOTE. Finally, the classification stage evaluates three CNN architectures: DenseNet169, DenseNet201, and EfficientNetB4. Ensemble learning and TTA are also explored to improve model robustness and generalization.

The following subsections provide a detailed description of each architecture and the Grad-CAM method.

A. DenseNet169 and DenseNet201

DenseNet169 and DenseNet201 were employed as the primary classification architectures due to their efficient feature reuse and dense connectivity mechanism [29]. Unlike traditional convolutional neural networks, where each layer receives input only from the previous layer, DenseNet connects each layer to all preceding layers, enabling improved gradient propagation and reducing the number of trainable parameters. The output of the (l)-th layer is given by

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \tag{1}$$

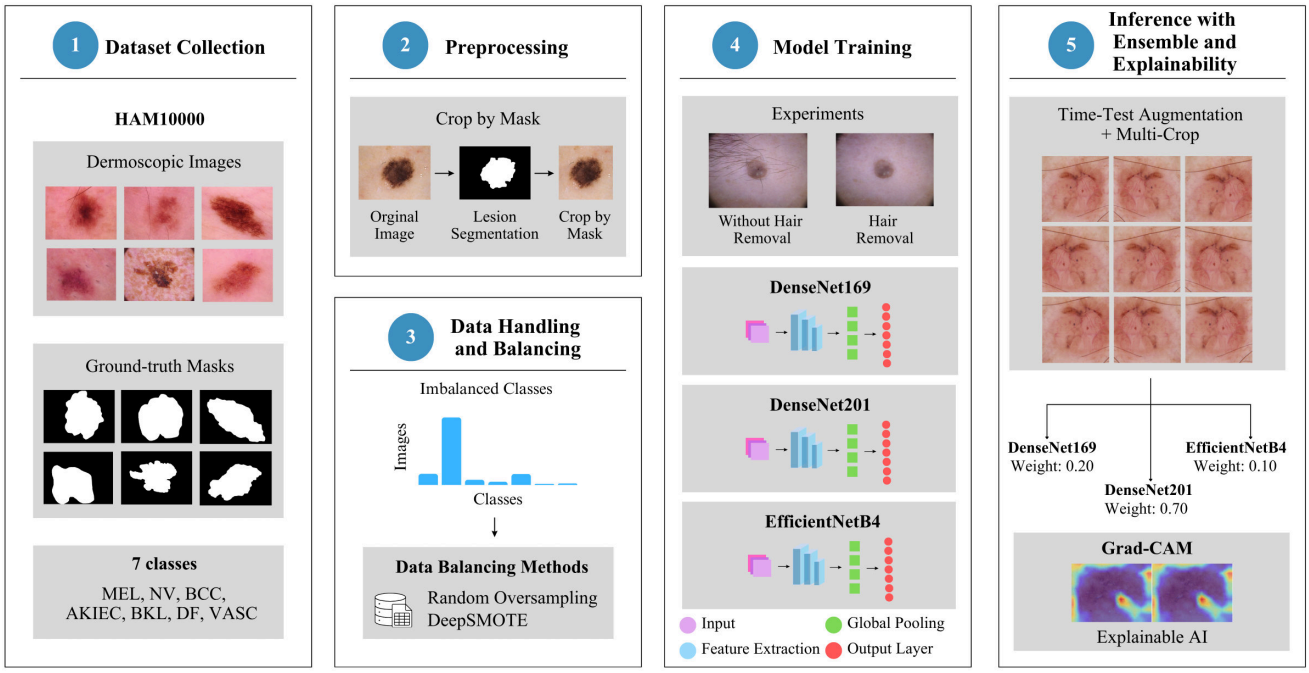


Fig. 5. Overview of the proposed experimental pipeline for multiclass skin lesion classification. The pipeline consists of five main stages: dataset collection, preprocessing, data handling and balancing, model training, and evaluation with grad-CAM-based explainability.

where $[\cdot]$ denotes feature-map concatenation and $H_l(\cdot)$ represents the composite operation of batch normalization, ReLU activation, and convolution.

After the final dense block, global average pooling is applied to obtain a feature vector z, which is passed to a fully connected layer producing the logits

$$o = Wz + b. \tag{2}$$

The class probabilities are computed using the Softmax function

$$P(y = i|X) = \frac{\exp(o_i)}{\sum_{j=1}^{K} \exp(o_j)}, \tag{3}$$

where K is the number of lesion classes. Model training is performed by minimizing the categorical cross-entropy loss

$$L = -\sum_{i=1}^{K} y_i \log\left(P(y = i|X)\right), \tag{4}$$

where y_i is the one-hot encoded ground-truth label. In this work, DenseNet169 and DenseNet201 differ primarily in network depth, containing 169 and 201 layers, respectively. In this work, both models were initialized with ImageNet pretrained weights and fine-tuned for multiclass skin lesion classification.

B. *EfficientNet*

EfficientNetB4 was evaluated as an alternative backbone architecture based on compound scaling [30]. Unlike DenseNet, which emphasizes feature reuse through dense connections, EfficientNet systematically scales network depth, width, and input resolution using a compound coefficient. EfficientNetB4 employs mobile inverted bottleneck convolution (MBConv) blocks combined with squeeze-and-excitation modules to enhance feature representation. The model was initialized with ImageNet pretrained weights and fine-tuned for multiclass skin lesion classification.

C. *Grad-CAM*

The Grad-CAM technique is used to provide visual explanations for the predictions made by the CNN model [18]. Let A^k denote the k-th feature map of the last convolutional layer and y^c the score for class c before the softmax activation. The importance weight of each feature map is obtained by globally averaging the gradients of the target class score with respect to the feature map activations:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{5}$$

where A_{ij}^k represents the activation at spatial location (i, j) of feature map k, and Z corresponds to the total number of spatial locations in the feature map. The Grad-CAM localization map

is then computed as a weighted combination of the feature maps, followed by a ReLU activation:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right), \tag{6}$$

where $L^c_{\text{Grad-CAM}}$ denotes the class activation map associated with class c. The ReLU function removes negative contributions, preserving only the features that positively influence the prediction. The resulting heatmap is subsequently upsampled to the original image size and overlaid on the input image to provide a visual explanation of the model's decision [18].

V. Results and Discussion

The simulation was conducted on a workstation equipped with an NVIDIA GeForce GTX 1660 Ti, 6 GB of VRAM, and 128 GB of RAM. The models were implemented using the TensorFlow and Keras libraries. The training process utilized the Adam optimizer with a learning rate of $1e^{-4}$, a batch size of 32, and early stopping based on validation loss to prevent overfitting.

Experiments were conducted to evaluate different preprocessing, data balancing, and classification strategies for multiclass skin lesion recognition. Fig. 6 illustrates the experimental pipeline, which includes three main stages: preprocessing, data balancing, and classification. The preprocessing stage involved hair removal using the Dull Razor algorithm and lesion-focused cropping based on segmentation masks. The data balancing stage explored three strategies: oversampling, oversampling with the Dull Razor algorithm, and Deep SMOTE. Finally, the classification stage evaluated three CNN architectures: DenseNet169, DenseNet201, and EfficientNetB4. In addition, ensemble learning and TTA were explored to improve model robustness and generalization.

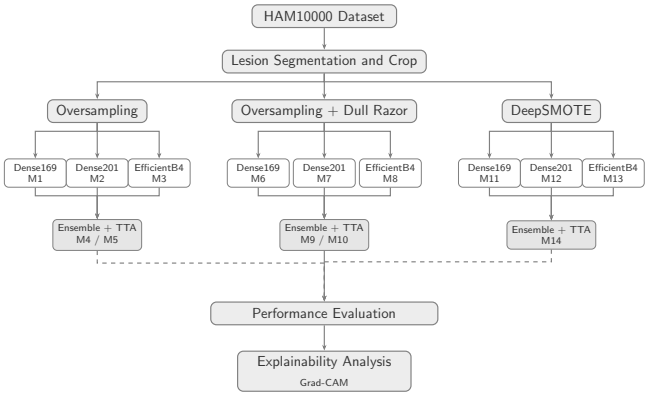


Fig. 6. Experimental pipeline for skin lesion classification, comparing three data augmentation strategies.

Deep SMOTE was implemented using lesion-cropped images resized to (256×256) pixels. The autoencoder was trained to learn a 512-dimensional latent representation of the lesions. The encoded training samples were then balanced by generating synthetic latent vectors through SMOTE-based interpolation between k-nearest neighbors ($k = 5$) within the same class. Finally, the synthetic latent vectors were decoded into dermoscopic images and added to the training set to improve the representation of minority classes.

Fig. 7 illustrates the class distribution of the original training set compared to the distribution after applying Deep SMOTE/Oversampling. The complete Deep SMOTE pipeline, including autoencoder training, latent-space interpolation, image reconstruction, and storage of synthetic samples, required approximately 1,271 minutes of execution time. The generated images were subsequently incorporated into the training set to improve the representation of minority classes.

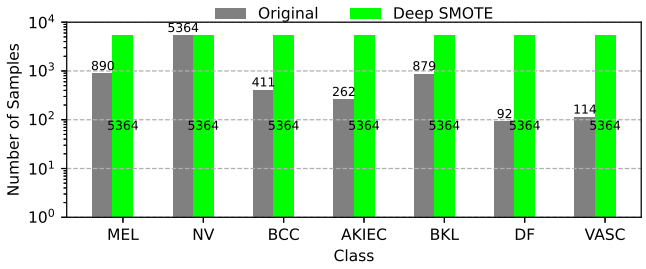


Fig. 7. Class distribution of the original and Deep SMOTE-augmented training sets.

Table II presents the performance comparison of the evaluated models. The results indicate that the ensemble model

TABLE II
Performance comparison of the evaluated models on the HAM10000 dataset.

Model	Accuracy	Precision	Recall	F1-score
M1	0.8512	0.8507	0.8512	0.8507
M2	0.8577	0.8558	0.8577	0.8565
M3	0.8567	0.8553	0.8567	0.8555
M4	0.8742	0.8723	0.8742	0.8728
M5	**0.8842**	**0.8806**	**0.8842**	**0.8815**
M6	0.8228	0.8242	0.8228	0.8231
M7	0.8547	0.8550	0.8547	0.8545
M8	0.8352	0.8373	0.8352	0.8357
M9	0.7923	0.7681	0.5336	0.5506
M10	0.7978	0.8100	0.7978	0.7743
M11	0.8108	0.8023	0.8108	0.8049
M12	0.8152	0.8153	0.8152	0.8137
M13	0.7998	0.7887	0.7998	0.7870
M14	0.8228	0.8198	0.8228	0.8135

M1: DenseNet169 + Oversampling + Crop.
M2: DenseNet201 + Oversampling + Crop.
M3: EfficientNetB4 + Oversampling + Crop.
M4: M1 + M2 + TTA.
M5: M1 + M2 + M3 + TTA.
M6: DenseNet169 + Oversampling + Crop + Dull Razor.
M7: DenseNet201 + Oversampling + Crop + Dull Razor.
M8: EfficientNetB4 + Oversampling + Crop + Dull Razor.
M9: M6 + M7 + TTA.
M10: M6 + M7 + M8 + TTA.
M11: DenseNet169 + DeepSMOTE + Crop.
M12: DenseNet201 + DeepSMOTE + Crop.
M13: EfficientNetB4 + DeepSMOTE + Crop.
M14: M11 + M12 + M13 + TTA.

M5, which combines DenseNet169, DenseNet201, and Effi-

cientNetB4 with oversampling and TTA, achieved the highest accuracy of 88.42% and an F1-score of 88.15%. In contrast, the ensemble model M14, which incorporates Deep SMOTE for data balancing, exhibited a lower accuracy of 82.28% and an F1-score of 81.35%. These findings suggest that conventional oversampling outperformed Deep SMOTE in this multiclass skin lesion classification task.

Moreover, contrary to expectations, the application of hair removal did not improve classification performance. Across all evaluated architectures, the Dull Razor preprocessing step resulted in either negligible gains or a reduction in accuracy and F1-score. One possible explanation is that lesion cropping already minimized the influence of background artifacts, making hair removal redundant. Furthermore, the inpainting procedure used to reconstruct hair-covered regions may have altered subtle dermoscopic structures relevant to lesion discrimination.

Tables III and IV demonstrate that the ensemble composed of DenseNet201, DenseNet169, and EfficientNetB4 trained using conventional oversampling (M5) achieved the best overall performance, reaching an accuracy of 88.42% and an F1-score of 88.15%.

TABLE III
CLASS-WISE PERFORMANCE METRICS FOR M5 MODEL.

Class	Precision	Recall	F1-Score
MEL	0.7853	0.6726	0.7246
NV	0.9302	0.9635	0.9465
BCC	0.7727	0.8252	0.7981
AKIEC	0.6774	0.6462	0.6614
BKL	0.7761	0.7091	0.7411
DF	0.9524	0.8696	0.9091
VASC	0.8966	0.9286	0.9123

TABLE IV
CLASS-WISE PERFORMANCE METRICS FOR MODEL M14.

Class	Precision	Recall	F1-Score
MEL	0.7101	0.4395	0.5429
NV	0.8847	0.9500	0.9162
BCC	0.5620	0.7476	0.6417
AKIEC	0.5119	0.6615	0.5772
BKL	0.7394	0.5545	0.6338
DF	0.8000	0.3478	0.4848
VASC	0.8966	0.9286	0.9123

In contrast, replacing conventional oversampling with Deep SMOTE (M14) reduced the accuracy to 82.28% and the F1-score to 81.35%. The class-wise analysis confirms this trend, as M5 consistently outperformed M14 for most lesion categories, particularly MEL, BCC, AKIEC, BKL, and DF, while both approaches achieved similar performance for VASC. These findings suggest that the synthetic images generated by DeepSMOTE did not provide additional discriminative information for the ensemble. The reason may be that the latent-space interpolation performed by Deep SMOTE introduced samples that did not fully preserve the fine-grained texture, color distribution, and morphological characteristics of real skin lesions. Consequently, the generated samples may have reduced the generalization capability of the network. Conversely, conventional oversampling preserves the original image distribution by replicating authentic training samples, allowing the ensemble to learn more representative visual features and resulting in superior classification performance.

For a more comprehensive evaluation of the model's performance, the confusion matrices were computed for both M5 and M14 models. Fig. 8 presents the normalized confusion matrix for M15, while Fig. 9 shows the normalized confusion matrix for M14.

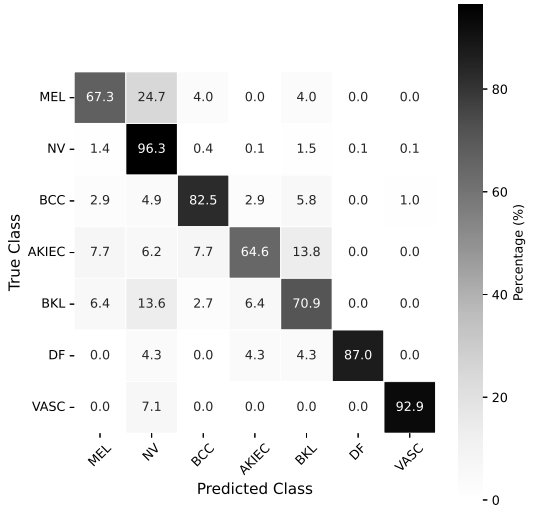


Fig. 8. Normalized confusion matrix for M5 model.

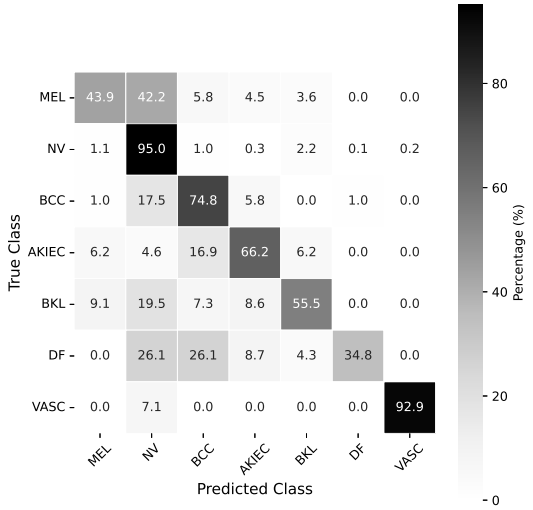


Fig. 9. Normalized confusion matrix for M14 model.

The confusion matrices reveal that M5 achieved higher true positive rates across most classes, particularly for MEL, BCC, AKIEC, and DF, compared to M14. The improved performance of M5 can be attributed to the use of conventional oversampling.

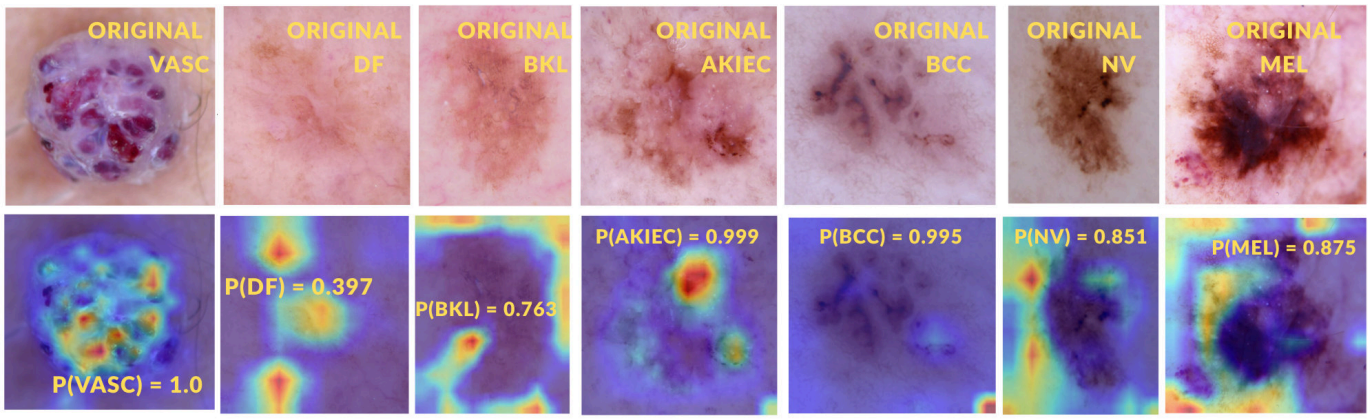


Fig. 10. Grad-CAM visualizations for selected test images. The first column shows the original dermatoscopic images, while the second column presents the corresponding Grad-CAM heatmaps overlaid on the input images.

Fig. 10 presents Grad-CAM visualizations for selected test images with M5. The heatmaps highlight the regions of the lesions that contributed most to the model's predictions.

The Grad-CAM indicates that the model generally focuses on localized image regions that contribute most strongly to the prediction of each skin lesion class. However, the quality of the explanations varies among classes, reflecting differences in lesion morphology and classification confidence. For the VASC example, the model achieved the highest confidence ($P(VASC) = 1.0$), and the activation map concentrated on the vascular structures inside the lesion. This suggests that the network successfully identified the characteristic vascular patterns associated with vascular lesions.

For the DF lesion, the prediction confidence was considerably lower ($P(DF) = 0.397$). The Grad-CAM highlighted only a limited central region, while substantial portions of the lesion received little attention. The BKL case ($P(BKL) = 0.763$) shows activations distributed over several portions of the lesion rather than a single focal region.

For AKIEC ($P(AKIEC) = 0.999$), the activation map is highly localized, with the strongest response centered on the most atypical region of the lesion. Similarly, the BCC example ($P(BCC) = 0.995$) exhibits concentrated activation around irregular pigmented structures, suggesting that the model primarily relied on clinically relevant morphological characteristics instead of background skin regions.

In the NV example ($P(NV) = 0.851$), the network focused mainly on the central pigmented area of the nevus while assigning little importance to the surrounding healthy skin. Finally, for the MEL lesion ($P(MEL) = 0.875$), the activation map covers a large portion of the lesion with stronger emphasis on darker and more irregular regions.

Comparing the best model (M5) with the related works presented in Table I, it is observed that M5 achieved competitive performance; however, it did not surpass the accuracy of the DODL-Net model [20], which achieved 98.76% accuracy. It is important to note that DODL-Net employed a different approach, including dual optimization algorithms and feature extraction techniques, which may have contributed to its superior performance. Nevertheless, the proposed M5 model demonstrated strong performance in multiclass skin lesion classification, achieving an accuracy of 88.42% and an F1-score of 88.15%, while also providing interpretability through Grad-CAM visualizations. The main goal of this study was to develop a multiclass classification with analysis of the impact of different preprocessing, data balancing, and classification strategies, rather than achieving the highest possible accuracy. The results highlight the importance of considering various factors in the design of skin lesion classification models and provide insights for future research in this area.

VI. Conclusion

In this study, we developed a multiclass skin lesion classification model using CNN architectures, including DenseNet169, DenseNet201, and EfficientNetB4. The model was trained and evaluated on the HAM10000 dataset. To address the class imbalance in the dataset, we explored different data balancing strategies, including conventional oversampling and Deep SMOTE. Additionally, we incorporated the Grad-CAM technique to provide visual explanations for the model's predictions, enhancing interpretability.

The experimental results demonstrated that the ensemble model combining DenseNet169, DenseNet201, and EfficientNetB4 with conventional oversampling and TTA achieved the best performance, with an accuracy of 88.42% and an F1-score of 88.15%. In contrast, the use of Deep SMOTE for data balancing resulted in lower performance, suggesting that conventional oversampling was more effective in this multiclass classification task. Furthermore, the application of hair removal using the Dull Razor algorithm did not improve classification performance, indicating that lesion cropping was sufficient to minimize background artifacts. Overall, this study highlights the importance of considering various factors, such as data balancing strategies and preprocessing techniques, in the design of skin lesion classification models. The integration of Grad-CAM provided valuable insights into the model's

decision-making process, contributing to the interpretability of the predictions.

Future research could investigate alternative data augmentation and class-balancing strategies, including more advanced generative approaches, given the inferior performance observed with Deep SMOTE. The removal of the DullRazor step from the pipeline could also be considered, particularly in light of the lack of performance gains and the possibility that the inpainting process may alter or obscure subtle dermatoscopic structures. Furthermore, the development of multimodal models that integrate dermatoscopic images with additional clinical information, such as patient metadata, could be explored to improve the performance and generalization of skin lesion classification models.

ACKNOWLEDGMENT

The authors would like to thank the Federal University of Rural Semi-Arid (UFERSA) for providing the necessary resources and infrastructure to conduct this research.

REFERENCES

[1] The Skin Cancer Foundation, "Skin cancer facts & statistics," https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/, 2026, accessed: 2026-06-07.

[2] N. Hasan, A. Nadaf, M. Imran, U. Jiba, A. Sheikh, W. H. Almalki, S. S. Almujri, Y. H. Mohammed, P. Kesharwani, and F. J. Ahmad, "Skin cancer: understanding the journey of transformation from conventional to advanced treatment approaches," *Molecular cancer*, vol. 22, no. 1, p. 168, 2023.

[3] K. M. Almustafa, "Predictive modeling and optimization in dermatology: Machine learning for skin disease classification," *Computers in Biology and Medicine*, vol. 189, p. 109946, 2025.

[4] J. Dinnes, J. J. Deeks, N. Chuchu, R. N. Matin, K. Y. Wong, R. B. Aldridge, A. Durack, A. Gulati, S. A. Chan, L. Johnston *et al.*, "Visual inspection and dermoscopy, alone or in combination, for diagnosing keratinocyte skin cancers in adults," *Cochrane Database of Systematic Reviews*, no. 12, 2018.

[5] H. D. Heibel, L. Hooey, and C. J. Cockerell, "A review of noninvasive techniques for skin cancer detection in dermatology," *American journal of clinical dermatology*, vol. 21, no. 4, pp. 513–524, 2020.

[6] A. K. Julian, R. A. Ferrer, and F. M. Perna, "Sun protection behavior: health impact, prevalence, correlates and interventions," *Psychology & health*, vol. 38, no. 6, pp. 701–725, 2023.

[7] J. Besch-Stokes, C. M. Brumfiel, M. H. Patel, J. Harvey, J. Montoya, K. J. Severson, H. Cumsky, M. Buras, J. E. G. Fagoaga, C. M. Costello *et al.*, "Skin cancer knowledge, attitudes and sun protection practices in the hispanic population: a cross-sectional survey," *Journal of Racial and Ethnic Health Disparities*, vol. 10, no. 3, pp. 1293–1303, 2023.

[8] World Health Organization, "Radiation: Ultraviolet (uv) radiation and skin cancer," https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer, 2017, accessed: 2026-06-07.

[9] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and deep learning techniques," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27 501–27 524, 2023.

[10] M. Samiei, A. Hassani, S. Sarspy, I. E. Komari, M. Trik, and F. Hassanpour, "Classification of skin cancer stages using a ahp fuzzy technique within the context of big data healthcare," *Journal of Cancer Research and Clinical Oncology*, vol. 149, no. 11, pp. 8743–8757, 2023.

[11] U. K. Lilhore, Y. K. Sharma, S. Simaiya, R. Alroobaea, A. M. Baqasah, M. Alsafyani, and A. Alhazmi, "Skinehdlf a hybrid deep learning approach for accurate skin cancer classification in complex systems," *Scientific Reports*, vol. 15, no. 1, p. 14913, 2025.

[12] N. J. F. Jaber and A. Akbas, "Melanoma skin cancer detection based on deep learning methods and binary harris hawk optimization," *Multimedia Tools and Applications*, vol. 84, no. 22, pp. 25 709–25 722, 2025.

[13] I. Krakowski, J. Kim, Z. R. Cai, R. Daneshjou, J. Lapins, H. Eriksson, A. Lykou, and E. Linos, "Human-ai interaction in skin cancer diagnosis: a systematic review and meta-analysis," *NPJ digital medicine*, vol. 7, no. 1, p. 78, 2024.

[14] G. Brancaccio, A. Balato, J. Malvehy, S. Puig, G. Argenziano, and H. Kittler, "Artificial intelligence in skin cancer diagnosis: a reality check," *Journal of Investigative Dermatology*, vol. 144, no. 3, pp. 492–499, 2024.

[15] C. M. V. Cavalcante, V. A. Almeida, M. Barros, N. Lima, and R. C. B. Rego, "Thyroid syndrome detection using machine learning algorithms: A comparative analysis," in *Proceedings of the XVI Brazilian Conference on Computational Intelligence (CBIC 2023)*. Brazil: Brazilian Computational Intelligence Society (SBIC), 2023.

[16] R. Ferreira and R. Rego, "Rede neural perceptron multicamadas para classificação multiclasse de câncer de pele," *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, vol. 1, no. 9, 2026.

[17] M. M. Musthafa, M. TR, V. K. V, and S. Guluwadi, "Enhanced skin cancer diagnosis using optimized cnn architecture and checkpoints for automated dermatological lesion classification," *BMC Medical Imaging*, vol. 24, no. 1, p. 201, 2024.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[19] D. Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and smote for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, 2023.

[20] E. Gomathi, M. Jayasheela, M. Thamarai, and M. Geetha, "Skin cancer detection using dual optimization based deep learning network," *Biomedical Signal Processing and Control*, vol. 84, p. 104968, 2023.

[21] M. Azeem, K. Kiani, T. Mansouri, and N. Topping, "Skinlesnet: classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network," *Cancers*, vol. 16, no. 1, p. 108, 2023.

[22] H. L. Gururaj, N. Manju, A. Nagarjun, V. M. Aradhya, and F. Flammini, "Deepskin: a deep learning approach for skin cancer classification," *IEEE access*, vol. 11, pp. 50 205–50 214, 2023.

[23] T. M. Alam, K. Shaukat, W. A. Khan, I. A. Hameed, L. A. Almuqren, M. A. Raza, M. Aslam, and S. Luo, "An efficient deep learning-based skin cancer classifier for an imbalanced dataset," *Diagnostics*, vol. 12, no. 9, p. 2115, 2022.

[24] A. A. Abdulredah, M. A. Fadhel, L. Alzubaidi, Y. Duan, M. Kherallah, and F. Charfi, "Towards unbiased skin cancer classification using deep feature fusion," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 48, 2025.

[25] K. T. Ahmed, F. Rustam, A. Mehmood, I. Ashraf, and G. S. Choi, "Predicting skin cancer melanoma using stacked convolutional neural networks model," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 9503–9522, 2024.

[26] W. M. Salamaa and M. H. Aly, "Deep learning design for benign and malignant classification of skin lesions: a new approach," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26 795–26 811, 2021.

[27] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018.

[28] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data mining and knowledge discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[30] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

Multiparametric EEG Analysis in Comatose Patients: A Comparison Between Silence and Musical Stimulation

Arianne Viana Alves da Silva
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-0645-0980

Maria Clara S. S. Surjus
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-9604-8708

Camila Davi Ramos
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-7864-5568

João Batista Destro Filho
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-5306-8058

Abstract*—Monitoring brain activity in comatose patients is a critical challenge in Intensive Care Units (ICUs), particularly within the Brazilian Unified Health System (SUS), where access to advanced neuroimaging is limited. This study proposes a quantitative electroencephalographic (EEG) analysis in 21 comatose patients (Glasgow Coma Scale score = 3) admitted to the ICU of the Hospital de Clínicas of the Federal University of Uberlândia, comparing silence and musical stimulation with Mozart's Sonata K.144 using two power spectral density-based quantifiers: Percentage Power Contribution (PCP) and interhemispheric Coherence (Wilcoxon Signed-Rank Test, = 0.05). Significant differences were identified in 8 of 120 PCP electrode-rhythm combinations (6.67%), predominantly at electrode P3 (left parietal region), and in 24 of 48 Coherence combinations (50.0%), with reductions predominating in both quantifiers. In PCP, the Theta band showed the highest Variation in Percentage (VAP) values (up to 22.73%); in Coherence, Delta showed the highest VAP (up to 23.51%) and Gamma reached significance in 7 of the 8 analyzed electrode pairs. Notably, 11 of the 24 significant Coherence combinations (45.8%) involved high-frequency bands above 30 Hz (Gamma and Super Gamma), with Super Gamma VAP values reaching up to 15.34%, a frequency range rarely investigated in this clinical context. Electrode P3 and the P3–P4 pair exhibited the highest concentration of significant combinations across both quantifiers. These findings suggest that interhemispheric synchrony is more sensitive to structured musical stimulation than local spectral power in this population, and that Coherence may serve as a potential marker of residual cortical responsiveness. The multiparametric, low-cost approach adopted is aligned with the resource constraints of the Brazilian public healthcare context.

***Index Terms*—coma, disorders of consciousness, electroencephalography, interhemispheric coherence, music therapy, power spectral density, quantitative EEG, intensive care unit**

I. Introduction

Monitoring brain activity in critically ill patients is one of the major challenges of clinical practice in Intensive Care Units (ICUs). In this context, Electroencephalography (EEG) stands out as a low-cost, non-invasive, and bedside-applicable tool capable of recording cortical electrical patterns in real-time [1]. Its sensitivity to variations in cerebral metabolism gives EEG an essential role in the stratification of levels of consciousness, especially in cases of profound unconsciousness in which comatose patients remain with closed eyes and without perceptible interaction with the environment [2].

The assessment of the state of consciousness in comatose patients is often performed using clinical scales, such as the Glasgow Coma Scale (GCS). However, these instruments present limitations, particularly in the detection of residual consciousness in apparently unresponsive individuals, and recent studies estimate that approximately 15% of patients clinically classified as unresponsive exhibit patterns of brain activation compatible with some level of consciousness, identifiable only through neurophysiological methods [3]. In this context, the prognostic value of EEG has been consistently demonstrated: Hofmeijer et al. [4] showed that unfavorable EEG patterns within 24 hours predicted poor neurological outcome with 100% specificity in 277 comatose patients after cardiac arrest, while early recovery toward continuous patterns within 12 hours was strongly associated with good outcome; machine learning approaches applied to EEG have further reinforced its value for outcome prediction in comatose patients [5], [6]. Such evidence reinforces the clinical value of EEG in critical care and motivates the investigation of strategies capable of modulating cortical activity in ways detectable by this tool.

In this scenario, music therapy has emerged as a promising non-pharmacological approach for the stimulation of patients with disorders of consciousness (DoC) [7]. Music therapy is particularly suited to this population due to its ability to activate distributed neural circuits, including regions associated with emotion, memory, and attention [8], [9], and

has been shown to modulate functional brain connectivity even in severely impaired states [10]. Studies conducted in clinical settings have demonstrated that musical exposure can induce measurable changes in electroencephalographic activity, with potential effects on clinical prognosis [11]–[14]. Therefore, the combination of quantitative EEG and musical stimulation protocols represents a strategy situated at the intersection of neurophysiology and music therapy, exploring the impact of sound on neurological recovery as a complement to conventional approaches focused on the stabilization of vital systems. This strategy is particularly relevant within the context of the Brazilian Unified Health System (SUS), where access to advanced neuroimaging technologies is often limited by structural and financial constraints, and where rapid and efficient diagnostic approaches are essential due to high patient demand and hospital bed overcrowding. Despite the growing international scientific production in this field, studies employing quantitative EEG analysis under musical stimulation in comatose patients admitted to Brazilian public ICUs remain scarce in the national literature.

In light of this scenario, the present study proposes a quantitative analysis of EEG signals from comatose patients admitted to the Intensive Care Unit of the Hospital de Clínicas of the Federal University of Uberlândia (HC-UFU), with the aim of comparing periods of silence and musical stimulation using Mozart's Sonata K.144. For this purpose, two power spectral density (PSD)-based quantifiers were simultaneously employed: the Percentage Power Contribution (PCP) and interhemispheric Coherence. To the best of our knowledge, this combined approach has not yet been explored in a Brazilian population with this clinical profile.

II. Literature Review

Quantitative EEG (qEEG) spectral analysis stands out as the most widely used approach to measure neural responses to musical stimulation. The index $(\delta + \theta)/(\alpha + \beta)$, defined as the ratio between slow and fast brain waves, has been used as a marker of cerebral alertness, since its reduction indicates a transition from unconscious states to higher cortical activation [11], [15]. In this context, Sun and Chen [15] evaluated 40 patients with traumatic brain injury (TBI) in a coma, exposing the experimental group to music therapy sessions for four weeks. At the end of the intervention, the qEEG index was significantly reduced in the music group compared to the control group ($p = 0.038$), with a significant negative correlation between this index and Glasgow Coma Scale scores, both before ($r = -0.482$, $p = 0.002$) and after treatment ($r = -0.493$, $p = 0.005$). These findings showed that the relative reduction of Delta and Theta waves, accompanied by an increase in Alpha and Beta activity, reflects improved functional brain activity. While this ratio aggregates spectral bands into a single index, the Percentage Power Contribution (PCP) quantifier extends this concept by quantifying the relative contribution of each individual rhythm per electrode, thereby enabling spatially resolved analysis of spectral changes across the entire scalp [16].

A related pattern, linking frequency spectrum reorganization to consciousness recovery, was also reported by Keller and Garbacenkaite [17], who applied neurofeedback based on the θ/β ratio in three patients with unresponsive wakefulness syndrome (UWS) [18], using preferred music as a positive reinforcement stimulus within the neurofeedback protocol. Two of the three patients showed a statistically significant reduction in this ratio over 15 sessions, accompanied by improvements in CRS-R scores in the auditory, motor, and arousal subscales. The authors interpret this decrease as a reflection of a reorganization of the dominant rhythm toward the Alpha band, consistent with partial restoration of cortical circuits.

Beyond the power spectrum, the effect of music on cognitive processing has been investigated through evoked potentials. Castro et al. [19] exposed 13 patients with disorders of consciousness (DoC) to sequences of their own name preceded either by excerpts of preferred music or by neutral spectrally matched noise. The N2/P3 discriminatory response to the subject's own name emerged significantly more frequently in the musical condition, indicating that music promoted a pre-activation of mid-anterior cortical structures that enhanced residual auditory responsiveness. This contextual potentiation effect (in which music acts as a cognitive primer rather than merely an isolated stimulus) suggests that different auditory conditions modulate the organization of brain electrical activity in distinct ways.

The regional distribution of these responses also revealed patterns of functional asymmetry between hemispheres and across brain lobes. Wu et al. [11] analyzed 14 patients with DoC using qEEG during exposure to three auditory stimuli: music, the subject's own name, and white noise, and found that, during musical stimulation, the $(\delta + \theta)/(\alpha + \beta)$ index was consistently higher in parietal and occipital regions than in frontal and temporal regions, in both the DoC and control groups. This topographic gradient suggests that music exerts a preferential influence on areas involved in sensory integration, spatial perception, and higher-order auditory processing. Additionally, in the TBI subgroup, the authors identified a negative correlation between the qEEG index and CRS-R scores ($R = -0.785$, $p = 0.012$), with a positive predictive value of 66.7% for favorable outcomes.

While spectral power indices provide a measure of regional cortical activation, Hu et al. [20] extended this investigation to evoked potentials, evaluating mismatch negativity (MMN, an electrophysiological component associated with automatic auditory discrimination and short-term sensory memory) before and after five minutes of exposure to a symphony with positive emotional associations. Before music exposure, MMN amplitude in patients with DoC was significantly lower than in the control group ($p < 0.001$); after stimulation, this difference was no longer statistically significant, indicating that music increased cortical responsiveness to a level close to that of healthy individuals. Post-musical MMN retesting yielded a negative predictive value of 100%, suggesting that its absence after musical exposure was associated with unfavorable

outcomes in all cases of the study, although this result did not reach statistical significance ($p > 0.05$), likely owing to the limited sample size ($n = 14$). Such evidence reinforces that musical stimulation can reorganize neural synchronization patterns, a process that can be captured not only by power spectrum analysis but also by inter-regional synchrony.

Although the reviewed studies consistently demonstrate that musical stimulation induces measurable changes in the electroencephalographic activity of comatose patients, most are limited to the use of a single spectral quantifier, predominantly indices derived from frequency-band ratios. Complementarily, Ramos et al. [16] and Ramos [21] have documented that interhemispheric Coherence (a measure of the degree of functional coupling between homologous regions of the two hemispheres) shows significant variations across different states of consciousness in Brazilian patients, with reduced values associated with impaired consciousness and potential trajectories of neurological recovery. Furthermore, classical music has been shown to produce measurable behavioral responses in unconscious ICU patients, with Rahgoi et al. [22] reporting significant reductions in pain observational and behavioral scales following exposure to classical music (including Mozart) in 30 unconscious patients ($p < 0.05$), further motivating its use as a structured auditory stimulus in this clinical context. The simultaneous analysis of two independent quantifiers, namely the spectral power distribution per electrode (PCP) and interhemispheric synchrony (Coherence), remains unexplored in this context, as does the application of these methodologies in Brazilian patients admitted to public ICUs. The present study seeks to address this gap by investigating neural responses to musical stimulation through a multiparametric and low-cost approach, suitable for the clinical and healthcare constraints of the Brazilian Unified Health System (SUS).

III. Methodology

A. Data Collection

Electroencephalographic recordings were performed in the Adult ICU of HC-UFU, approved by the UFU Research Ethics Committee (protocol no. 82824017.5.0000.5152). Data were collected during routine care, without interference in therapeutic procedures. The sample consisted of 21 adult patients with Glasgow Coma Scale score = 3. Inclusion criteria were age over 18 years, hemodynamic stability, a clinical indication for EEG, and written informed consent from the legal guardian. Patients requiring high-dose vasopressors or lacking consent were excluded. One additional patient was excluded due to excessive signal artifact. Approximately 62% of participants were male, aged 21–86 years, with predominant diagnoses of traumatic brain injury, hemorrhagic stroke, and subarachnoid hemorrhage. Coma was classified as induced when sedative agents were administered at the time of EEG recording, and as natural otherwise; 5 of 21 patients (24%) met this criterion. All but one patient were under mechanical ventilation (endotracheal tube or tracheostomy); the remaining patient received supplemental oxygen via tracheostomy. Regarding outcome, 8 patients (38%) died (3 due to brain death), 6 (29%) were discharged, 3 (14%) remained hospitalized, and outcome was not registered for 4 (19%).

EEG signals were acquired using the BrainNet BNT-EEG amplifier, with 20 active electrodes positioned according to the international 10-20 system, two auricular reference electrodes, and one cardiac-signal electrode. The sampling frequency was 400 Hz, defined after six months of testing at the Adult ICU of HC-UFU, where higher frequencies resulted in excessive environmental noise, precluding quantitative analysis [16].

The experimental protocol consisted of six auditory stimuli presented sequentially: silence (0–3 min), preferred music (3–6 min), the note A (440 Hz; 6.5–8 min), white noise (10–13 min), Gregorian chant (13.5–16.5 min), and Mozart's Sonata K.144 (17–20 min), interspersed with brief intervals of human conversation. Recordings were performed with patients supine and eyes closed. Stimuli were reproduced through speakers positioned near the patient's head, at intensity adjusted for adequate perception without disturbing other patients. Speakers were used instead of headphones to reduce infection risk from shared devices and avoid additional discomfort in patients already connected to multiple therapeutic devices. The protocol and data processing pipeline are shown in Fig. 1.

B. Data Processing

Following acquisition, two neurologists from the Clinical Neurophysiology unit of HC-UFU selected ten two-second artifact-free epochs by visual inspection, one evaluator per patient across all stimulation conditions; inter-rater agreement could therefore not be calculated. Analyses were conducted on the initial silence period (0–3 min) and the final three minutes of Mozart's Sonata K.144 (17–20 min), the condition established in the group's research protocol [16]. Signals were analyzed using two PSD-based quantifiers: PCP and interhemispheric Coherence.

PCP quantifies the relative contribution of each cerebral rhythm to the total power of the EEG signal at a given electrode [16], calculated according to (1):

$$PCP^{a}_{rhythm,i} = \frac{\displaystyle\int_{f=f_{in}^{rhythm}}^{f_{out}^{rhythm}} |S_{x_{ia}}(f)|^2 \, df}{P_i}, \qquad (1)$$

$$P_i = \int |S_{x_{ia}}(f)|^2 \, df,$$

where i denotes the analyzed epoch; f_{in}^{rhythm} and f_{out}^{rhythm} are the lower and upper frequency boundaries of the rhythm of interest [Hz]; $S_{x_{ia}}(f)$ is the PSD of the EEG signal at electrode x [μV^2/Hz]; and P_i is the total signal power. Both quantifiers were computed using the bilateral auricular reference adopted during acquisition, without re-referencing. The notch filter was applied in the frequency domain, after the Fourier transform and before the power calculation.

The following frequency bands were analyzed: Delta (1–3.5 Hz), Theta (3.5–7.5 Hz), Alpha (7.5–12 Hz), Beta (12–30 Hz), Gamma (30–80 Hz), and Super Gamma (80–100 Hz),

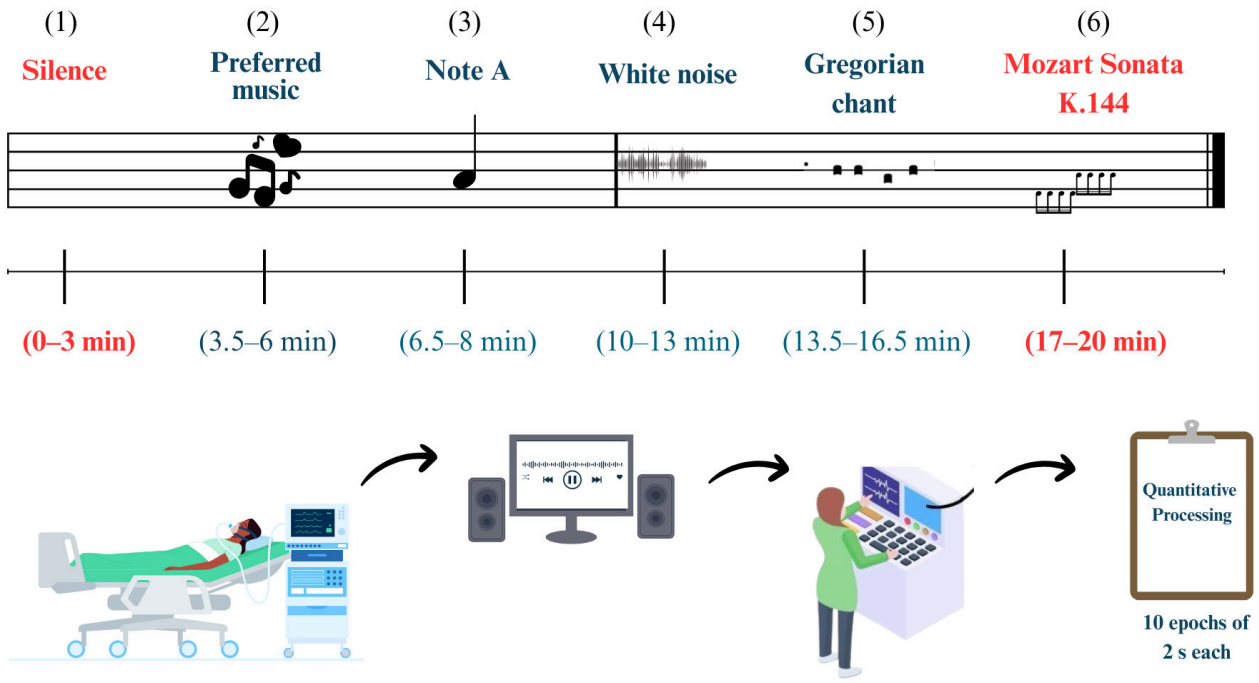


Fig. 1. Stimulation protocol and data processing pipeline. Six auditory conditions were applied sequentially over a total duration of 20 min, separated by 30-second intervals of human conversation sounds: (1) Silence (0–3 min), (2) Preferred music (3.5–6 min), (3) Note A, 440 Hz (6.5–8 min), (4) White noise (10–13 min), (5) Gregorian chant (13.5–16.5 min), and (6) Mozart's Sonata K.144 (17–20 min). Conditions highlighted in red were selected for quantitative analysis. The lower panel illustrates the data processing pipeline: EEG acquisition, auditory stimulation via speakers, visual selection of 10 artifact-free epochs of 2 s each, and quantitative analysis (PCP and Coherence).

across 20 electrodes distributed according to the 10-20 system, yielding a total of $20 \times 6 = 120$ electrode-rhythm combinations.

Coherence assesses the degree of coupling between signals from homologous regions of the right and left hemispheres [1], [16], calculated according to (2):

$$\left|\Gamma_{xy}(e^{j\omega})_i\right|^2 = \frac{\left|S_{xy}(e^{j\omega})\right|^2}{S_x(e^{j\omega}) \cdot S_y(e^{j\omega})}, \quad 0 \leq \left|\Gamma_{xy}(e^{j\omega})_i\right|^2 \leq 1, \tag{2}$$

where S_{xy} is the cross-PSD; S_x and S_y are the individual PSDs of the signals; and i denotes the segment under consideration. Values close to 1 indicate strong interhemispheric synchrony; values close to 0 indicate low coupling between signals.

Eight pairs of homologous electrodes were analyzed: FP1–FP2, F7–F8, F3–F4, T3–T4, C3–C4, T5–T6, P3–P4, and O1–O2, yielding a total of $8 \times 6 = 48$ electrode pair-rhythm combinations.

C. Statistical Analysis

PCP and Coherence results were expressed as median per condition. Comparison between silence and musical stimulation was performed using the Wilcoxon Signed-Rank Test [23], adopting a significance level of 5%. This test was chosen for being appropriate for paired samples from continuous distributions, without assuming data normality, a condition not guaranteed in EEG signals from comatose patients with reduced sample sizes. As each patient served as their own control, this design accounts for inter-subject variability but does not establish whether the observed responses are specific to the comatose state, which would require a non-comatose reference group.

For combinations with statistically significant differences ($p < 0.05$), the Percentage Variation (VAP) was calculated from the medians, according to (3):

$$VAP = \frac{|Md_{music} - Md_{silence}|}{\max(Md_{music}, Md_{silence})} \times 100\%, \tag{3}$$

where Md_{music} and $Md_{silence}$ correspond to the medians obtained under musical stimulation and silence, respectively. As VAP was computed only for combinations already identified as significant, it should be read as a descriptive measure of effect magnitude, not as an independent significance criterion.

IV. Results

A. Percentage Power Contribution (PCP)

The statistical analysis of PCP identified significant differences ($p < 0.05$) in 8 of the 120 electrode-rhythm combinations (6.67%), concentrated in five electrodes: P3, CZ, T5, PZ,

and O2. The P3 electrode, located in the left parietal region, accounted for half of the significant combinations (4 out of 8), with changes observed in the Theta, Alpha, Beta, and Gamma bands. The remaining electrodes (CZ, T5, PZ, and O2) showed significance in only one rhythm each, as detailed in Table I. Of the 8 significant combinations, 6 corresponded to reductions in the median PCP during musical stimulation. Among the significant combinations, the Theta band showed the highest concentration of changes in PCP, appearing in three of the eight significant electrode-rhythm combinations (CZ, P3, and PZ), with VAP values reaching up to 22.73%, the largest observed in this quantifier.

TABLE I
SIGNIFICANT PCP RESULTS ($p < 0.05$; $n = 8/120$; 6.67%)

Electrode	Rhythm	VAP (%)
CZ	Theta	−22.73
T5	Delta	+3.88
P3	Theta	−13.64
P3	Alpha	−17.40
P3	Beta	−15.85
P3	Gamma	−7.80
PZ	Theta	−15.86
O2	Gamma	+9.27

VAP – Percentage variation of the median between silence and music; negative values indicate reduction, positive values indicate increase under musical stimulation.

Fig. 2 illustrates the topographic distribution of the median differences in PCP (ΔPCP = Music − Silence) for the six evaluated brain rhythms. Each map uses its own color scale, allowing identification of the spatial pattern of intra-rhythm variation; comparisons of magnitude between different rhythms should be made based on the absolute values reported in Table I, rather than by the visual intensity of the colors.

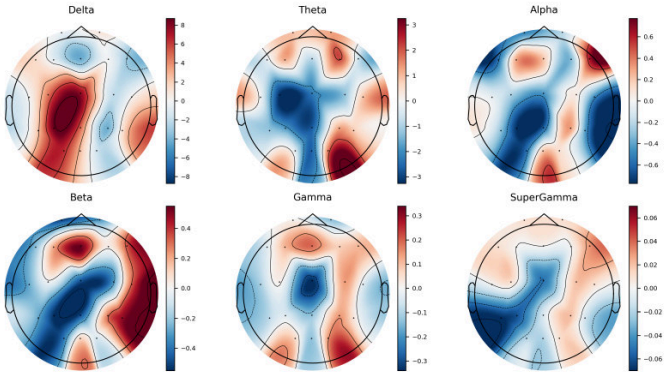


Fig. 2. Topographic distribution of median PCP differences (ΔPCP = Music − Silence) for the six evaluated brain rhythms. Independent color scale per rhythm; red: increase ($\Delta > 0$); blue: decrease ($\Delta < 0$).

B. Interhemispheric Coherence

The statistical analysis of Coherence identified significant differences ($p < 0.05$) in 24 of the 48 electrode-rhythm pair combinations (50.0%). The F7–F8, T3–T4, and P3–P4 electrode pairs concentrated the highest number of significant rhythms, with significance observed in 5 of the 6 evaluated rhythms for each pair. The F3–F4 pair showed significance in 1 rhythm, the FP1–FP2 pair in 1 rhythm, and the C3–C4 and T5–T6 pairs in 2 rhythms each, as detailed in Table II. Of the 24 statistically significant combinations, 16 (66.7%) corresponded to reductions in the median Coherence during musical stimulation, while 8 (33.3%) corresponded to increases. Among the evaluated frequency bands, Gamma showed the broadest spatial distribution, reaching significance in 7 of the 8 analyzed electrode pairs, while Delta exhibited the highest individual VAP value (23.51% in the T3–T4 pair). Of the 24 significant combinations, 11 (45.8%) involved high-frequency bands above 30 Hz, comprising 7 Gamma and 4 Super Gamma combinations, with Super Gamma VAP values reaching up to 15.34% in the O1–O2 pair.

TABLE II
SIGNIFICANT COHERENCE RESULTS ($p < 0.05$; $n = 24/48$; 50.0%)

Pair	Rhythm	VAP (%)
FP1–FP2	Beta	+10.41
F7–F8	Delta	−13.19
F7–F8	Alpha	−20.09
F7–F8	Beta	−8.67
F7–F8	Gamma	−3.92
F7–F8	SG	+11.35
F3–F4	Gamma	−5.58
T3–T4	Delta	+23.51
T3–T4	Theta	−13.45
T3–T4	Beta	−7.75
T3–T4	Gamma	+6.08
T3–T4	SG	+4.67
C3–C4	Beta	−9.36
C3–C4	Gamma	−5.34
T5–T6	Theta	−10.69
T5–T6	Beta	−8.72
P3–P4	Delta	−5.18
P3–P4	Theta	−11.77
P3–P4	Beta	−14.53
P3–P4	Gamma	−10.65
P3–P4	SG	−13.53
O1–O2	Delta	+15.33
O1–O2	Gamma	+12.06
O1–O2	SG	+15.34

SG – Super Gamma. VAP – Percentage variation of the median between silence and music; negative values indicate reduction, positive values indicate increase under musical stimulation.

Fig. 3 presents the spatial distribution of Coherence differences (ΔCoherence = Music − Silence) between electrode pairs for the six evaluated brain rhythms, represented as connection lines whose thickness is proportional to the magnitude of the variation and whose color indicates the direction of the difference.

C. Comparative Analysis

Coherence showed a higher proportion of significant combinations than PCP (50.0% vs. 6.67%), with reductions predominating in both quantifiers: 6 of the 8 significant PCP

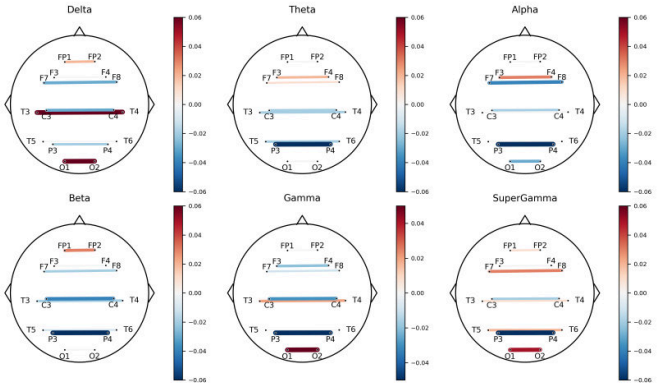


Fig. 3. Interhemispheric Coherence Differences (ΔCoherence = Music − Silence) between electrode pairs for the six evaluated brain rhythms. Color: direction of the difference (red, $\Delta > 0$; blue, $\Delta < 0$). Line thickness: proportional to $|\Delta$Coherence$|$, increasing from the smallest to the largest observed magnitude within each rhythm. Independent color scale for each rhythm.

combinations and 16 of the 24 Coherence combinations. The parietal region emerged as a common point between the two measures, with electrode P3 concentrating most of the significant PCP results and the P3–P4 pair showing significance across five rhythms, all reduced. High-frequency bands above 30 Hz (Gamma and Super Gamma) accounted for 45.8% of the significant Coherence combinations (11 of 24), with Gamma reaching significance in 7 of the 8 electrode pairs and Super Gamma VAP values up to 15.34%.

V. Discussion

The results indicate that musical stimulation with Mozart's Sonata K.144 induced measurable changes in the electroencephalographic activity of comatose patients, detectable through two independent spectral quantifiers. The low proportion of significant combinations in PCP (6.67%), against the high proportion observed in Coherence (50.0%), constitutes the central finding of the present study.

This higher sensitivity of Coherence relative to PCP suggests that musical stimulation preferentially modulates interhemispheric functional organization rather than local spectral power, consistent with the hypothesis that music recruits distributed circuits connecting homologous regions across hemispheres rather than isolated cortical activation [24], [25]. Accordingly, the predominance of Coherence reductions (16 of 24 significant combinations, 66.7%) aligns with Ramos and Destro-Filho [21], who reported that lower interhemispheric coherence characterizes impaired consciousness states in Brazilian patients, and that variations in this parameter during hospitalization may track neurological evolution [16]. Taken together, these results suggest Coherence as a candidate marker of residual cortical responsiveness to structured auditory stimuli, to be further examined in longitudinal designs.

The parietal region was also a spatial point of convergence between the two quantifiers: electrode P3 concentrated four of the eight significant PCP combinations, and the P3–P4 pair showed the most extensive and consistent Coherence reduction, across five rhythms. This converges with Wu et al. [11], who identified the parietal lobe as most influenced by musical stimulation in DOC patients due to its role in multimodal sensory integration and spatial perception, and with Heine et al. [25], who reported increased connectivity in the temporoparietal junction and supramarginal gyrus during music exposure in DOC patients, regions anatomically close to electrodes P3 and P4. As both quantifiers were computed independently, this convergence reduces the likelihood that the pattern reflects a methodological artifact.

In contrast, the low proportion of significant PCP combinations (6.67%) is best read as evidence that local spectral power changes are less sensitive to musical stimulation than interhemispheric synchrony in this population, rather than as an absence of response. Sun and Chen [15] and Wu et al. [11] similarly reported reductions in the $(\delta + \theta)/(\alpha + \beta)$ index under musical stimulation in DOC patients, a pattern partially consistent with the PCP reductions observed here in the Theta band at P3, CZ, and PZ, which further reinforces the spatial specificity of these responses in the parietal region.

Beyond the conventional bands, Gamma activity reached significance in 7 of the 8 analyzed electrode pairs, and Super Gamma combinations showed VAP values up to 15.34%, indicating that musical stimulation also modulates interhemispheric coupling at frequencies rarely investigated in this clinical context. Most studies on musical stimulation in disorders of consciousness restrict spectral analysis to conventional bands below 30 Hz, leaving Gamma and Super Gamma activity largely unexplored [8], [12]. However, the 80–100 Hz range is particularly susceptible to electrical, myogenic, and ventilator-related artifacts in the ICU environment, and no artifact control specific to this frequency range (e.g., line-noise assessment, imaginary coherence, or surrogate data analysis) was applied in the present study; these Gamma and Super Gamma findings should therefore be interpreted with caution until confirmed through such controls.

The F7–F8, T3–T4 and O1–O2 pairs also showed distinct patterns across frequency bands, with predominant Coherence reductions at F7–F8 and T3–T4 and increases at O1–O2, broadly consistent with frontotemporal and occipital connectivity changes reported during music exposure in DOC patients [16], [19], [25]. Nonetheless, these pair-specific patterns should be read as exploratory, given the sample size and the absence of correction for multiple comparisons.

These findings should be interpreted alongside the study's design constraints. The etiological heterogeneity of the sample (traumatic brain injury, stroke, and subarachnoid hemorrhage) and the wide age range (21–86 years), combined with the sample size ($n = 21$), limit generalization beyond the population studied and precluded diagnosis-stratified analysis; future work should confirm these patterns in etiology-stratified cohorts. Likewise, no correlation was tested between EEG findings and clinical outcome or neurological evolution, and the absence of a healthy control group precludes attributing the observed patterns specifically to the comatose state; both aspects were

outside the scope of the present exploratory analysis. The fixed order of stimulus presentation shown in Fig. 1, with silence always first and Mozart's Sonata K.144 last, is a further consideration: the observed differences may partly reflect time-dependent factors such as arousal drift, sedation accumulation, or intervening ICU events, in addition to the auditory content of the stimulus. Randomized or counterbalanced designs with matched non-musical controls would help isolate stimulus-specific effects in future work. Sedation is an additional factor to consider: 5 of the 21 patients received sedative agents at the time of recording, known to alter spectral EEG activity and interhemispheric coupling, and this variability was not statistically controlled for. Exact p-values, effect sizes, and confidence intervals are not reported for individual comparisons; significance is indicated only categorically ($p < 0.05$). Moreover, no correction for multiple comparisons was applied across the 120 PCP and 48 Coherence tests, so the reported significant combinations should not be read as independent evidence, and the present findings are considered exploratory pending replication with corrected significance thresholds and complete reporting of test statistics. Additionally, body temperature, formal auditory assessment, standardized prognostic examinations, and time from injury onset to EEG were not systematically recorded in the original protocol and are noted here rather than treated as covariates. Finally, the analysis was restricted to silence and Mozart's Sonata K.144; incorporating the remaining stimuli from the protocol (preferred music, the note A, white noise, Gregorian chant) will clarify whether the response is specific to structured music or reflects a general auditory response.

Despite these constraints, the multiparametric combination of PCP and Coherence proved feasible using low-cost, bedside-compatible equipment without disrupting routine ICU care, supporting its applicability within SUS-based intensive care units. Also, the convergence of both quantifiers on the parietal region, despite being computed independently, indicates that combining spectral and connectivity measures adds diagnostic value beyond either metric alone. Building on these results, replication in larger, etiology-stratified samples, using randomized or counterbalanced stimulus designs and longitudinal follow-up, would help establish the prognostic value of interhemispheric coherence as a biomarker of cortical responsiveness to musical stimulation in comatose patients.

VI. Conclusion

This study investigated the neural response to musical stimulation with Mozart's Sonata K.144 in comatose ICU patients, through the simultaneous analysis of PCP and interhemispheric Coherence. The sample consisted exclusively of patients with profound coma (Glasgow Coma Scale score of 3), a population that remains underrepresented in the quantitative EEG literature on musical stimulation, particularly in Brazilian public ICUs. Coherence proved substantially more sensitive to musical stimulation than PCP (50.0% vs. 6.67% of significant combinations), suggesting that structured auditory stimuli preferentially modulate interhemispheric functional organization rather than local spectral power. This pattern was reinforced by the predominance of reductions in both quantifiers and by their spatial convergence at the parietal region, particularly at electrode P3 and the P3–P4 pair. Delta and Gamma showed the most pronounced Coherence changes, while Theta showed the highest VAP values in PCP; notably, Coherence also revealed significant Super Gamma-band changes in four homologous electrode pairs, extending the analysis beyond the frequency range typically investigated in this clinical context. Given the sample size and the absence of correction for multiple comparisons, these findings are best regarded as exploratory, warranting confirmation in larger, etiology-stratified cohorts.

Even with these constraints, the approach adopted requires only standard, low-cost EEG equipment and non-pharmacological stimulation, making it directly compatible with the operational and financial constraints of Brazilian public ICUs within the SUS framework. As one of the few studies to jointly apply PCP and Coherence in this population, and the first, to our knowledge, to extend this analysis into the Gamma and Super Gamma bands, it contributes a methodological reference for future quantitative EEG research in resource-constrained critical care settings. Building on these results, future work should incorporate the remaining stimuli from the experimental protocol (preferred music, the note A, white noise, and Gregorian chant) to assess the specificity of the observed responses relative to non-musical auditory controls, and evaluate the prognostic value of Coherence modulation as a biomarker of neurological evolution through longitudinal designs with larger, etiology-stratified samples.

Acknowledgment

The authors thank the Clinical Neurophysiology Unit and Adult ICU of HC-UFU for technical support, with recognition to technician Patrícia Gonçalves da Costa, and neurologists Marcos Campos and Paulo Bettero. The authors also thank Prof. Marina A. Ramos (ICBIM/UFU) for ethical authorization and Prof. Celso Cintra (School of Music, UFU) for generating the audio stimuli. This work was supported by CNPq, FAPEMIG and CAPES through undergraduate research scholarships and graduate research funding.

References

[1] L. Sörnmo and P. Laguna, *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Amsterdam: Academic Press, 2005.

[2] E. F. M. Wijdicks, "Management of the comatose patient," in *Handbook of Clinical Neurology*, vol. 140, pp. 117–129, Amsterdam: Elsevier B.V., 2017.

[3] M. Amiri, S. Achard, O. Bertrand, B. Huber, P. Kahane, A. Mauguiere, G. Nica, A. Perrone-Bertolotti, A. Picard, R. Ryvlin, T. Schmitz, L. Vercueil, and M. Techer, "Multimodal prediction of residual consciousness in the intensive care unit: the CONNECT-ME study," *Brain*, vol. 146, pp. 50–64, Jan. 2023.

[4] J. Hofmeijer, T. M. J. Beernink, F. H. Bosch, A. Beishuizen, M. C. Tjepkema-Cloostermans, and M. J. A. M. van Putten, "Early EEG contributes to multimodal outcome prediction of postanoxic coma," *Neurology*, vol. 85, pp. 137–143, July 2015.

[5] S. Jonas, A. O. Rossetti, M. Oddo, S. Jenni, P. Favaro, and F. Zubler, "EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features," *Human Brain Mapping*, vol. 40, pp. 4606–4617, Nov. 2019.

[6] F. M. Aellen, S. L. Alnes, F. Loosli, A. O. Rossetti, F. Zubler, M. De Lucia, and A. Tzovara, "Auditory stimulation and deep learning predict awakening from coma after cardiac arrest," *Brain*, vol. 146, no. 2, pp. 778–788, 2023.

[7] J. T. Giacino, J. J. Fins, S. Laureys, and N. D. Schiff, "Disorders of consciousness after acquired brain injury: the state of the science," *Nature Reviews Neurology*, vol. 10, pp. 99–114, Feb. 2014.

[8] T. Grimm and G. Kreutz, "Music interventions and music therapy in disorders of consciousness — A systematic review of qualitative research," *The Arts in Psychotherapy*, vol. 74, p. 101782, 2021.

[9] B. Murtaugh, A.-M. Morrissey, S. Fager, H. E. Knight, J. Rushing, and J. Weaver, "Music, occupational, physical, and speech therapy interventions for patients in disorders of consciousness: An umbrella review," *NeuroRehabilitation*, vol. 54, no. 1, pp. 109–127, 2024.

[10] M. Carrière, S. K. Larroque, C. Martial, M. A. Bahri, C. Aubinet, F. Perrin, S. Laureys, and L. Heine, "An echo of consciousness: Brain function during preferred music," *Brain Connectivity*, vol. 10, pp. 385–395, Sept. 2020.

[11] M. Wu, Y. Li, S. Liu, Y. Zhao, L. Yang, C. Yan, and Z. Zhao, "Effect of acoustic stimuli in patients with disorders of consciousness: A quantitative electroencephalography study," *Neural Regeneration Research*, vol. 13, no. 11, pp. 1900–1906, 2018.

[12] G. E. Lancioni, N. N. Singh, M. F. O'Reilly, J. Sigafoos, and L. Desideri, "Music stimulation for people with disorders of consciousness: A scoping review," *Brain Sciences*, vol. 11, no. 7, p. 858, 2021.

[13] H. Yang, W. Huang, W. Wen, S. Long, Y. Zhang, X. Chi, and D. Luo, "The impact of frequency-specific music stimulation on consciousness in patients with disorders of consciousness," *Frontiers in Neurology*, vol. 16, p. 1506261, 2025.

[14] S. Spaccavento, G. Carraturo, E. Brattico, B. Matarrelli, D. Rivolta, F. Montenegro, E. Picciola, N. T. Haumann, K. V. Jespersen, P. Vuust, and E. Losavio, "Musical and electrical stimulation as intervention in disorder of consciousness (DOC) patients: A randomised cross-over trial," *PLoS ONE*, vol. 19, no. 5, p. e0304642, 2024.

[15] J. Sun and W. Chen, "Music therapy for coma patients: preliminary results," *European Review for Medical and Pharmacological Sciences*, vol. 19, no. 7, pp. 1209–1218, 2015.

[16] C. D. Ramos, G. Pereira, A. Rodrigues, M. Campos, J. B. Destro Filho, and G. Alves, "Quantitative analysis of EEG records in distinct states of consciousness: A case study on Brazilian patient," *IEEE Latin America Transactions*, vol. 16, pp. 1878–1883, July 2018.

[17] I. Keller and R. Garbacenkaite, "Neurofeedback in three patients in the state of unresponsive wakefulness," *Applied Psychophysiology and Biofeedback*, vol. 40, pp. 349–356, Dec. 2015.

[18] S. Laureys, G. G. Celesia, F. Cohadon, J. Lavrijsen, J. León-Carrión, W. G. Sannita, and European Task Force on Disorders of Consciousness, "Unresponsive wakefulness syndrome: a new name for the vegetative state or apallic syndrome," *BMC Medicine*, vol. 8, no. 68, pp. 1–4, 2010.

[19] M. Castro, B. Tillmann, J. Luaté, A. Corneyllie, F. Dailler, N. André-Obadia, and F. Perrin, "Boosting cognition with music in patients with disorders of consciousness," *Neurorehabilitation and Neural Repair*, vol. 29, pp. 734–742, Sept. 2015.

[20] Y. Hu, H. Chen, Q. Yin, L. Gao, X. Xu, and Y. He, "Can music influence patients with disorders of consciousness? An event-related potential study," *Frontiers in Neuroscience*, vol. 15, p. 596636, Apr. 2021.

[21] C. D. Ramos and J. B. Destro-Filho, "Measurements of coherence in EEG signal in Brazilian people: A comparison of different consciousness states," *Bioscience Journal*, vol. 37, p. e37008, Jan. 2021.

[22] A. Rahgoi, S. M. Zarankesh, S. S. Kasaeizadeghan, and M. Naghibeiranvand, "Effect of classical music on physiological characteristics and observational and behavioral measures of pain in unconscious patients admitted to intensive care units," *Complementary Medicine Journal*, vol. 10, no. 4, pp. 368–379, 2021.

[23] D. C. Montgomery and G. C. Runger, *Estatística aplicada e probabilidade para engenheiros*. Rio de Janeiro: LTC, 4 ed., 2009.

[24] J. D. Rollnik and E. Altenmüller, "Music in disorders of consciousness," *Frontiers in Neuroscience*, vol. 8, p. 190, July 2014.

[25] L. Heine, M. Castro, C. Martial, B. Tillmann, S. Laureys, and F. Perrin, "Exploration of functional connectivity during preferred music stimulation in patients with disorders of consciousness," *Frontiers in Psychology*, vol. 6, p. 1704, Nov. 2015.

Multisensory prototype for the investigation of sleep-related respiratory events: verification of functional integration and signal processing capability.

Gabriela Moraes
Physiotherapy Postgraduate Program
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0002-6672-3319

Livia Maria Dornelas Gomes
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
livia.dornelas@ufpe.br

Adriano Lourenço
Physiotherapy Postgraduate Program
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-2463-0563

Marcus Vinícius B. Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0003-9751-2124

Bruno Henrique Viana Moreira
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0000-9532-6641

Marco A. B. Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-5143-3696

Abstract*— Obstructive Sleep Apnea Syndrome (OSA) is a severe respiratory disorder, frequently underdiagnosed due to the high cost and complexity of polysomnography, the current diagnostic gold standard. This study aims to develop and evaluate the functional integration and signal processing capability of acoustic, accelerometric, and oximetric data in a multisensory prototype designed to investigate sleep-related respiratory events. The system consists of an electroacoustic transducer (MAX9814), synchronized with wearable accelerometry (MPU-6050) and oximetry (PPG) modules, with MATLAB-based processing for digital filtering and feature extraction (RMS, Spectral Centroid, LF/HF ratio). The proof-of-concept trial, conducted in a laboratory environment (LIHOM-UFPE) through simulated bench tests, demonstrated a pre-configured temporal alignment among the three sensors: a prolonged period of acoustic silence (200–280 s) was followed by a rescue sound of -19 dBFS and by the nadir of simulated desaturation (SpO2 = 91%, t = 291 s), with a delay parameterized according to the lung-to-finger circulation time described in the literature.Spectral analysis via FFT revealed a predominance of energy at 38.82 Hz, within the frequency range reported in the literature for low-frequency snoring components. However, as a single spectral peak from a synthesized stimulus, this result reflects the system's spectral analysis capability rather than obstruction-site classification. The accelerometry module correctly captured the programmed postural transition from supine to lateral decubitus, confirming the sensor's ability to register movement patterns relevant to positional monitoring, without constituting evidence of POSA-related physiological coupling. Built from commercial, low-cost electronic components, the prototype demonstrated technical feasibility as a multisensory acquisition and synchronization platform for future sleep-related respiratory-event research.

Keywords*— *Sleep Apnea Obstructive, Snoring, Accelerometry, Oximetry, Monitoring Physiologic.

I. INTRODUCTION

Sleep-disordered breathing, notably snoring and obstructive sleep apnea syndrome (OSA), represents a growing public health challenge and affects a significant portion of the adult population [1], [2]. Frequently viewed merely as a social nuisance, snoring is the result of airway vibration and acts as an important pathophysiological marker of increased upper airway resistance and changes in the breathing pattern [1], [2]. Apnea, in turn, is a severe chronic condition characterized by repeated interruptions in breathing. Although snoring is relevant to the obstructive pattern, its sensitivity and specificity are limited. Its isolated presence does not necessarily define the diagnosis of OSA [1]-[3].

Studies indicate that snorers undergo changes in respiratory mechanics and airway musculature, especially during Non-Rapid Eye Movement (NREM) sleep. This elevated airway resistance increases the tendency for pharyngeal collapse, which prevents the maintenance of deep sleep stages and results in shorter, poorer-quality rest [1]-[3]. The early detection of obstructive events is clinically crucial, as OSA is associated with severe cardiovascular, metabolic, and neurocognitive complications. Although polysomnography is the diagnostic gold standard, its high operational cost and the requirement for specialized infrastructure limit population access. Consequently, there is a growing demand for alternative, portable, and less invasive methods for screening and nocturnal monitoring, using snoring as a clinical predictor [3], [4].

In the context of developing a prototype based on sensor integration, a wearable device, and IoT (Internet of Things) technology, the acoustic analysis of snoring serves as a useful tool in the functional assessment of the upper airways, especially when analyzing the associations between sound frequencies and specific sites of obstruction, providing support for functional evaluation [5]-[7]. In order to identify obstructive apneic events non-invasively and in a manner adaptable to the home environment, the simultaneous implementation of body movement and peripheral oxygenation sensors expands the possibilities of respiratory monitoring [8], whereby accelerometry applied to the thoracic and abdominal region assists in estimating respiratory dynamics [9]. Additionally, pulse oximetry detects drops in oxygen saturation resulting from reduced airflow, and its integration with acoustic signals can significantly improve the automated identification of obstructive events, reducing dependence on complex laboratory exams such as polysomnography [10], [11].

Therefore, the development of a technology that integrates acoustics, oximetry, and accelerometry emerges as a promising strategy [11]. The fusion of these signals provides a broad view of respiratory dynamics, enabling the detection of obstructive events with high analytical robustness [10] and democratizing the early diagnosis of OSA, especially in regions lacking specialized exams [12].

In this context, this study aims to develop and evaluate the functional integration and signal processing capability of acoustic, accelerometric, and oximetric data in a multisensory prototype designed to investigate sleep-related respiratory events.

II. Material and methods

The proposed device is characterized as a benchtop prototype in the proof-of-concept phase for the continuous and non-invasive monitoring of parameters indicative of Obstructive Sleep Apnea (OSA). Structurally, the system consists of three independent physical modules: acoustic (sound), oximetry (optical plethysmography), and accelerometry (mechanical). The device embodies the IoT technology approach, aiming to minimize the use of wires and integrate the sensors responsible for sound and biomechanical analysis. Its internal structure comprises an independent power supply (battery) with USB cable charging and an SD memory card for data storage, which are later processed and analyzed with the aid of the SISMO UFPE web portal (Sistema de Monitoramento Fisioterapêutico de Pacientes) and MATLAB software, for integrated visualization of the results [13], [14].

A. Processing Unit and Sensors

- *Electroacoustic transducer (ESP32-S3 Super Mini and MAX9814)***:** The module is controlled by the ESP32-S3 Super Mini microcontroller. The choice is due to the presence of internal 12-bit analog-to-digital converters (ADCs), which allow for the precise reading of multiple simultaneous analog channels and high data encapsulation capacity for IoT. Additionally, its compact size, versatility, and high energy efficiency are essential characteristics for continuous monitoring devices [15]. The MAX9814 sensor, responsible for detecting snoring sound episodes, was chosen mainly based on its internal automatic gain control (AGC), which ensures that abrupt variations in acoustic amplitude are properly adjusted and conditioned for the ESP32-S3 analog input, ensuring data integrity by preventing signal saturation [16]. It is important to note that the AGC continuously adjusts the input gain in response to signal amplitude; consequently, the dBFS, RMS, Leq, and threshold-based measurements reported in this study reflect the sensor's post-AGC output and should be interpreted as relative indicators of the processing pipeline's behavior, rather than calibrated sound pressure measurements against a fixed acoustic reference. A calibration procedure was not performed at this stage and is identified as a requirement for future quantitative comparisons;
- *Wearable Peripheral Oximetry Module:* Oxygen saturation is measured non-invasively using a sensor positioned at the fingertip. Its functionality operates through photoplethysmography (PPG), analyzing the difference in red and infrared light absorption by deoxygenated hemoglobin (Hb) and oxyhemoglobin (HbO2), allowing for the continuous calculation of oxygen saturation (SpO2) [14];
- *Wearable Accelerometry Module (MPU-6050):* The sensor captures linear acceleration and angular velocity data. Positioning this sensor on the trunk allows for the evaluation of decubitus (posture) in orthogonal axes (X, Y, Z) and the identification of thoracoabdominal effort associated with respiratory pauses based on chest wall biomechanics [13].

B. Data Processing

- *Hybrid acquisition (Table I):* The acoustic signal is sampled by the ESP32-S3 module via Direct Memory Access (DMA). In the firmware itself, an oversampling technique was implemented, with an initial high-frequency collection rate (20 kHz) followed by a decimation step by a factor of 20. This procedure acts as a moving average digital low-pass filter, attenuating high-frequency noise and adjusting the effective sampling rate to 1 kHz, optimizing the data flow for subsequent export and saving in a .csv file. Simultaneously, the numerical matrices of accelerometry and oximetry are exported from the SISMO cloud platform.

  Each channel was acquired using an independent microcontroller of the same model (ESP32-S3), governed by its own onboard crystal oscillator, with no shared clock or synchronization trigger implemented between channels. Temporal correspondence across channels was instead achieved by manually matching the pre-configured simulated event sequence during data visualization in MATLAB. Accordingly, the reported temporal correlation among channels reflects the precision of this manual matching process rather than a measured synchronization error between independent oscillators. Formal quantification of this error is identified as a requirement for future implementations employing a shared clock reference or hardware-level triggering.
- *Data processing and parameter extraction:* For data processing, the chosen tool was the MATLAB software by MathWorks (WEB version). The raw signals are processed by importing the numerical matrices from the three sources. In MATLAB, the algorithm is programmed to apply low-pass and high-pass filters to identify and attenuate noise, such as sudden movement artifacts and power line interference. This cleaning step is fundamental to enable the temporal alignment of events, allowing the correlation between peaks in the accelerometer signal with drops in the SpO2 curve and variations in snoring amplitude;

TABLE I. Acquisition Parameters and Signal Ranges by Channel

Channel	Signal Type / Origin	Sampling Rate	Range / Units	Applied Filter
Electroacoustic Transducer	Simulated signal (pre-configured sound stimulus), read via MAX9814	20 kHz (initial acquisition), 1 kHz (effective, after decimation by a factor of 20)	dBFS (e.g., -19 dBFS rescue sound; -18 dBFS isolated peak)	DC offset removal; 20–300 Hz band-pass; notch filter; 100-sample windowing with 50% overlap
Wearable Accelerometry Module	Postural movement mechanically reproduced on the bench, read via MPU-6050/ESP32-S3	50 samples per second (Hz) on both axes	Dimensionless amplitude on the X, Y, Z axes (Hz)	Digital moving average filter, 10-sample window
Wearable Peripheral Oximetry Module	Simulated saturation (programmed nadir), obtained via PPG module, exported from the SISMO platform	50 samples per second (Hz)	SpO2 in 80–110% (sensor operating range) Simulated event values ranged from 91% to 97.2% within this scale	Artifact-rejection filter via retention of the last valid value below 90%

- *Filtering of biomechanical signals:* In MATLAB, a digital moving average filter with a 10-sample window is applied to the three orthogonal axes of accelerometry, acting as a linear low-pass filter to attenuate high-frequency noise and tremors. Given the accelerometry sampling rate of 50 Hz, this filter configuration yields an approximate cutoff frequency of 2.2 Hz, effectively attenuating high-frequency artifacts while preserving the lower-frequency postural and thoracoabdominal movement signals of physiological interest;
- *Filtering of oximetry signals:* A logical conditional artifact rejection filter is applied, which retains the last valid physiological value if the reading suddenly drops to values below 90% (characteristic of sensor failure), stabilizing the saturation curve.

  It should be noted that this retention strategy, while stabilizing the visual continuity of the saturation curve, does not distinguish between an artifact-induced drop and a genuine desaturation event below the 90% threshold. As currently implemented, the filter does not generate an explicit flag or annotation when a value is being carried forward rather than newly acquired, which represents a limitation for downstream event detection. A future implementation should log each occurrence of value retention as a distinct output flag rather than allowing it to appear as an unbroken physiological trace;
- *Acoustic signal filtering:* The raw data initially undergo DC offset (continuous level) removal. Filters are implemented for the frequencies of interest. The band-pass filter retains only frequencies between 20 and 300 Hz, eliminating excessively low- or high-frequency noise that does not belong to snoring. The notch filter is used to mitigate electromagnetic interference and artifacts. The resulting signal is normalized and segmented by a 100-sample windowing with 50% overlap;
- *Feature extraction and acoustic analysis:* Through the Fast Fourier Transform (FFT), the frequency spectrum of the entire audio is calculated, allowing the identification of the predominant frequency components. The spectral analysis, illustrated by the FFT, shows that the highest concentration of respiratory sound energy is located in the band below 300 Hz, justifying the digital filter choices. From each window, the algorithm extracts features in the time domain (Respiratory Energy, RMS, Zero-Crossing Rate) and in the frequency domain via Welch's Power Spectral Density (Centroid, Entropy, Rolloff, LF/HF ratio). The detection of apnea/snoring events is based on dynamic thresholds using the signal's median to define the silence baseline, stratifying the intensity of the events into statistical percentile levels (L1, L5, L10) and Equivalent Noise Level (Leq).

C. Operational Protocol

This stage was characterized by the development and verification of the prototype. Given the preliminary nature of the research, experimental validation was conducted in a laboratory environment, through bench tests with simulated sound stimuli, without collecting human biological data, focusing exclusively on the technical functioning of the hardware and the response of the acoustic sensor, with no clinical diagnostic objective. The experimentation was carried out at the Human-Machine Interface Laboratory of the Electronics and Systems Department at UFPE (LIHOM-UFPE). It is noteworthy that the oximetry and accelerometry devices have been previously validated and are functioning, communicating with the monitoring platform [14], [15]. Therefore, the procedure consisted of the functional validation and integration of the MAX9814 sensor into the pre-existing system.

a) Metrics

The metrics analyzed covered the performance of the instrumentation hardware and the accuracy of the detection algorithms. The evaluated parameters included:

1. Hardware and network metrics: Effective sampling frequency, which is the microcontroller's ability to maintain the programmed reading rate (Hz) when integrating the A/D conversion of the audio signal;

2. Acoustic signal quality: Evaluation of the Signal-to-Noise Ratio (SNR) guaranteed by the acoustic sensor's AGC in relation to background environmental noise;

3. Digital processing validation: Accuracy in identifying snoring peaks was ensured by applying an event detection algorithm, which uses dynamic thresholds calculated from the median of the signal baseline. Additionally, the acoustic intensity of the events was stratified in decibels (dBFS) through the Equivalent Noise Level (Leq) and statistical percentiles (L1, L5, and L10).

b) Usage Guidelines

Before starting the test, the responsible technical team considered the main prior precautions necessary to avoid data inadequacy, performing the following instrumental, environmental, and technical checks, respectively:

- *Observation of equipment integrity:* inspecting for possible signs of breakage or mechanical shocks. If there was any damage, the use of the device was to be interrupted until laboratory repair was completed;
- *Minimal use of external sources:* that could cause electromagnetic and light interference during collection, aiming to reduce artifacts in the acquired data;
- *Data verification and formatting:* It is strictly necessary to pay attention to .csv formatting and file-nomenclature standardization, following the model used in the SISMO portal during data acquisition (e.g., proc_20260507170129_(device)_.csv), ensuring compatibility for signal processing in the MATLAB software.

III. Results and Discussion

The interpretation and systematization of the data obtained in the laboratory testing of the device were structured to validate the integrated technical functioning of its modules and fulfill its main purpose: investigating the influence of snoring on obstructive events through the direct correlation between sound characteristics (acoustic module), thoracoabdominal breathing pattern (accelerometry module), and peripheral oxygenation responses (SpO2). For functional and analytical evaluation of the multimodal prototype, the topic was divided into five complementary sections.

A. Validation of the Biomechanical Module in Decubitus Change and Interaction with Oximetry

The prototype's ability to register a controlled simulation of decubitus change and correlate it with ventilatory responses was initially evaluated using the accelerometer module. The reproduction of the postural transition from supine to lateral decubitus was captured by the z-axis of the accelerometer (Fig. 1), through a steep decline in amplitude (t = 0.78 - 1.25 s | max amp = -1.04 | min amp = -0.93). Conversely, a brief reduction (amp = -0.49) was observed around 2 s of testing (t = 1.89 s), which may represent a movement artifact associated with the controlled mechanical movement of the sensor.

As described in the Methods section, a moving average digital filter (10-sample window) was applied in MATLAB over the accelerometry axes to attenuate high-frequency noise and tremors. Fig. 2 shows the difference between raw and filtered data, with a reduction of 0.09 in the maximum amplitude and 0.01 in the minimum (t = 0.78 - 1.25 s | max amp = -0.95 | min amp = -0.92). In the brief reduction period, the filtering effect was slightly more pronounced than in the estimated period for decubitus change, with a difference of 0.14 (t = 1.9 s | amp = -0.35).

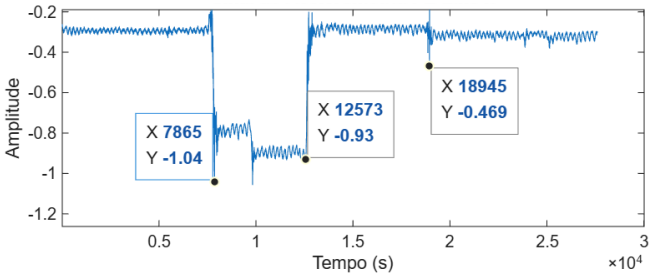


Fig. 1. Representation of thoracoabdominal amplitude alteration of the accelerometer x-axis. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

.

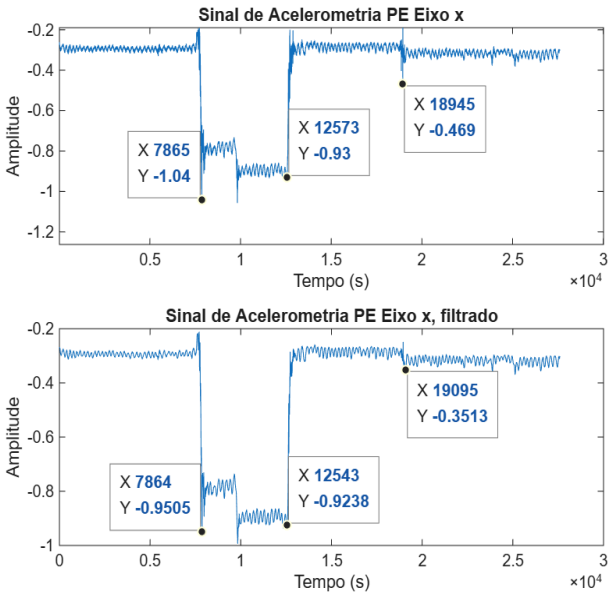


Fig. 2. Raw data versus filtered data on the accelerometry x-axis. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

In Fig. 3, referring to the oximetry signal, an oscillating pattern generated by the oximetry signal during the simulated decubitus trial is demonstrated, evidenced in the demarcated range in the three axes (x-axis: t = 0.78-1.25 s | max amp = -0.95 | min amp = -0.92; y-axis: t = 0.77-1.27 s | max amp = -0.08 | min amp = -0.06) and in the SpO2 curve (z-axis: t = 162-291 s | max SpO2 = 97.2% | min SpO2 = 91%).

This amplitude redistribution pattern falls within the range of postural signatures described in the literature for positional obstructive sleep apnea (POSA), in which supine decubitus favors upper airway collapse due to the action of gravity on the tongue and soft palate, while lateral decubitus tends to reduce airway resistance and the frequency of obstructive events [17]. As the postural transition was mechanically reproduced on the bench rather than performed by a patients, this result demonstrates the accelerometry module's capability to register and temporally align postural changes with the other sensor channels, reinforcing its role as a complementary sensor in multimodal OSA monitoring [9]. However, it does not

constitute physiological evidence of POSA-related airway behavior, which would require validation in human participants.

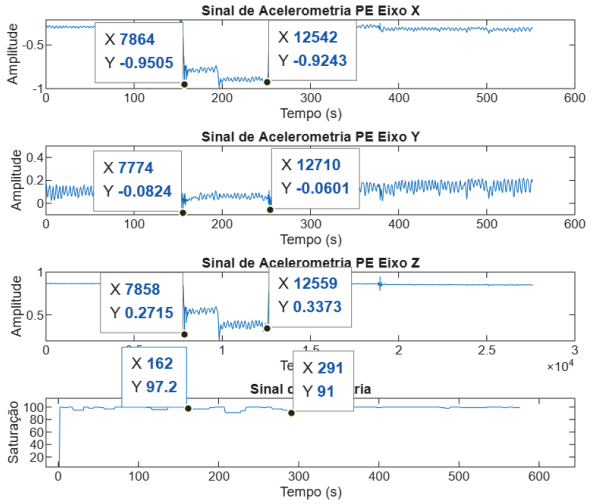


Fig. 3. Thoracoabdominal movement on the accelerometry axes and its association with oximetry. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

B. *Spectral Characterization of Snoring - Fast Fourier Transform (FFT)*

The analysis of energy concentration in the frequency domain, via FFT, revealed a predominance around 38.82 Hz (Fig. 4), characterizing the acoustic signal as low-frequency and compatible with the spectral profile described as a breathing pattern without relevant upper airway alterations in this segment of the recording [18]. This instrumental performance of the prototype aligns with the literature on the acoustic classification of snoring, according to which lower frequency components tend to be associated with higher levels of obstruction (retropalatal/oropharyngeal), while higher frequencies usually reflect obstructions at lower levels of the airway (retrolingual/hypopharyngeal) [18], [7].

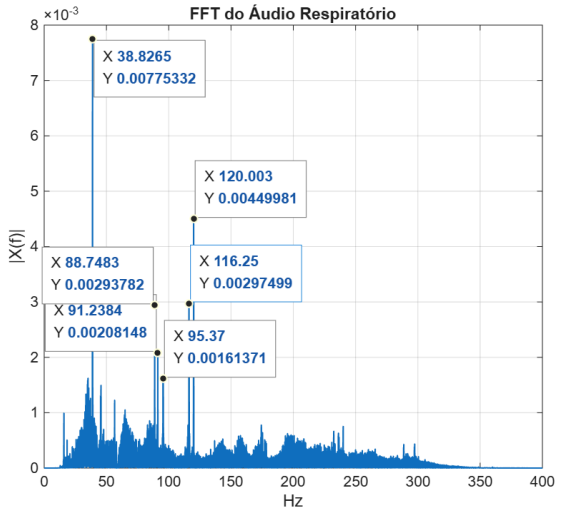


Fig. 4. Fast Fourier Transform (FFT) of Snoring. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

Subsequently, the system demonstrated sensitivity by registering the emergence of irregular frequencies delimited up to the 120 Hz range, showing compatibility with possible superposition of vibration sites.

C. *Extraction of Acoustic Features and Mapping of Obstructive Events*

The evaluation of the prototype over a continuous 665-second window (Fig. 5) certified its capability to extract longitudinal acoustic attributes to identify obstructive breathing disorders. In the interval between 50 s and 440 s, the device's algorithm discriminated a basal, rhythmic, and chronic snoring pattern, evidenced by the persistent RMS amplitude and sharp, cyclic drops in spectral entropy, which mark the periodic transition from chaotic respiratory noise to well-structured and predictable harmonic waves.

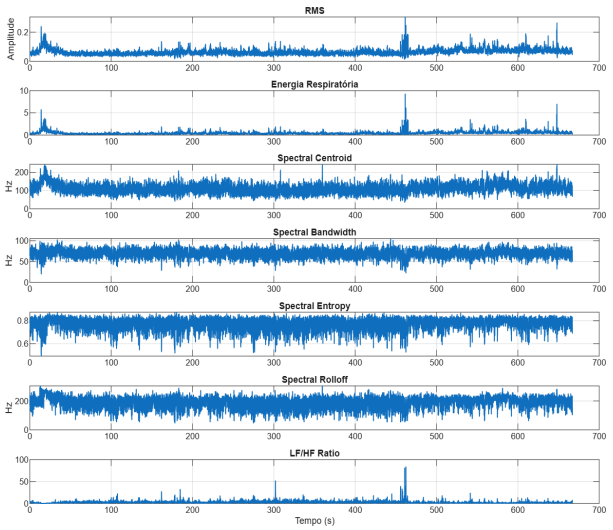


Fig. 5. Respiratory Features regarding snoring amplitude, vibration site, and gravity. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

D. *Stratification of Obstructive Events and Postural Influence Metrics*

The accuracy of the digital processing concerning the detection of respiratory events was evaluated via frequency peaks with a minimum amplitude validation of 35% (Fig. 6). The algorithm identified an outlier in the first 20 seconds of the recording, which artificially elevated the sound frequency ceiling (RMS = 0.21).

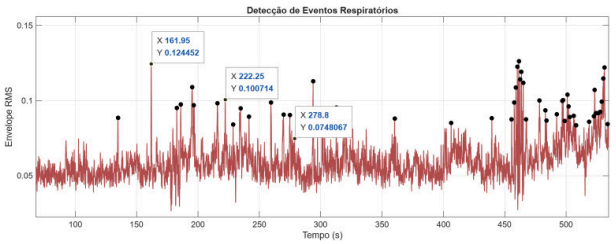


Fig. 6. Detection of Respiratory Events, via frequency peaks based on a minimum amplitude validation of 35%. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

The most critical behavior of the recording was captured abruptly around 450 s, characterized by a significant surge in the RMS envelope and respiratory energy, consistent with the sharp peak in the LF/HF ratio (close to 80). This surge falls within the low-frequency band (20 – 120 Hz) described in the literature as characteristic of gasping/arousal acoustic signatures [19], [20]. As this event was part of the pre-configured simulated sequence, this result demonstrates that the signal-processing pipeline reliably detects programmed amplitude and spectral-ratio transitions. It does not, by itself, validate the acoustic signature of airway

collapse and reopening in a physiological sense, which would require testing with real recordings from human participants.

Highlighting the purpose of integration, the gyroscope analysis in the z-axis (Fig. 7) associated the amplitude limitations and SpO2 oscillations with specific postural changes. The initial RMS peak of 0.12 demarcates the period of transition to lateral decubitus (150 - 250 s). In this interval, a consistent variation of high-pitched sound peaks during the inspiratory phase was recorded (mean RMS = 0.10; t = 222 s), which physiologically show an association with saturation oscillations. The oxygen decline was captured by the plethysmographic module approximately 20 seconds after the acoustic event was captured—an interval compatible with the lung-to-finger circulation time already described in the pulse oximetry literature on sleep apnea [21]. These findings generated by the device suggest the ability to capture patterns compatible with the influence of lateralization, stated in the pathophysiology of POSA [17].

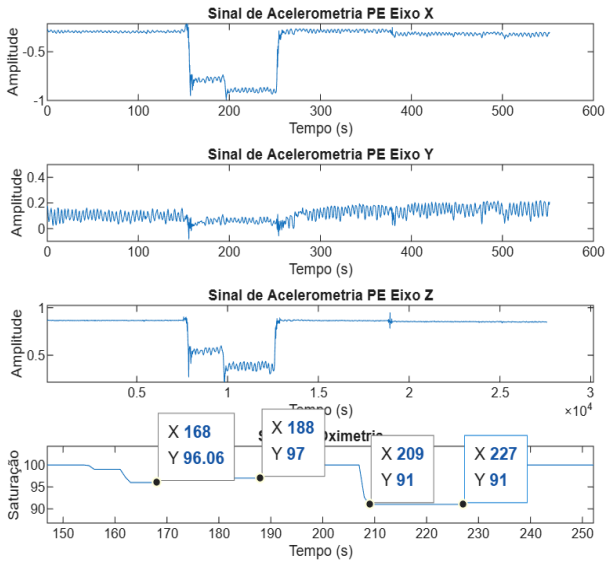


Fig. 7. Postural Alteration Captured by Gyroscope and Associated with Respiratory Events. Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

E. Sound Intensity Weighting (Leq) and Synchronization with Oximetric Profile

To consolidate the validation of the simultaneous and correlated functioning of the sensors, the chronological agreement between the acoustic data and the oximetry profile was evaluated (Fig. 8).

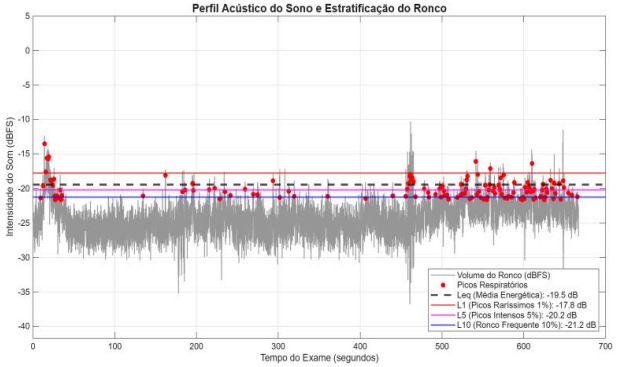


Fig. 8. Peripheral Oximetry Signal and Equivalent Noise Level (Leq) and statistical percentiles (L1, L5 and L10). Source: Data obtained from a bench test with simulated stimuli, processed and statistically analyzed on the MATLAB web portal.

In the first highlighted event, at instant t = 162 s, the emergence of an isolated respiratory peak in the acoustic envelope (approx. -18 dBFS) is observed, coinciding with an oxygen saturation still stable at 97.2%. This record demonstrates the adequate functioning of the transducer in capturing a brief respiratory manifestation, which occurred before there was sufficient time for a relevant drop in peripheral oxygenation. The most relevant instrumental finding regarding the prototype's purpose occurs at time t = 291 s, after a prolonged period of low amplitude in the acoustic signal (t = 200 - 280 s). The device recorded a rescue sound of -19 dBFS, emitted precisely at the moment the oximetry model showed the nadir of a significant desaturation, dropping to 91%.

This temporal agreement was intentionally parameterized in the simulated sequence, based on the physiological delay of approximately 10 to 20 seconds described in the literature between the end of a respiratory event and the detection of desaturation at the peripheral extremity, resulting from the lung-to-finger circulation time and the instrumental latency of the pulse oximeter [21]. Accordingly, this result verifies that the acquisition and visualization pipeline correctly preserved this programmed inter-channel delay - a synchronization and plotting-logic outcome - rather than demonstrating detection of the physiological delay itself, which was not measured but reproduced as a fixed, pre-defined parameter. This integration pattern reflects the timeline structure defined in the pre-configured simulation and demonstrates that the acquisition and synchronization pipeline can reproduce the type of multi-channel event sequence reported in acoustic-oximetric fusion approaches for OSA screening [19], [20]. This result supports the technical viability of the synchronization pipeline, providing a basis for future validation against real physiological delays. Establishing screening or diagnostic performance, however, would require dedicated studies with human participants.

IV. Conclusion

The developed prototype demonstrated satisfactory preliminary functioning during the integration of acoustic, mechanical, and oximetry signals. The system's efficiency was significant due to the precise temporal correlation between the acoustic silence valleys (200 to 280 s), the subsequent rescue sounds with low-frequency signatures and LF/HF ratio peaks close to 80, and the subsequent oxygen desaturations (nadir of 91% | t = 291 s), synchronized to decubitus changes reproduced on the bench and captured by the accelerometer. However, the susceptibility to artifacts in the first 20 seconds - which generated a biased RMS ceiling of 0.21 - evidences the critical need to improve digital filtering algorithms beyond the 10-sample moving average utilized.

From a technical applicability standpoint, it is noteworthy that the prototype was built entirely from commercial, compact, and relatively low-cost electronic components, avoiding the need for the specialized infrastructure and high operational investment required by polysomnography, the diagnostic gold standard [3]. This

characteristic, combined with the simplified operation enabled by IoT technology, supports the technical feasibility of a low-cost, synchronized multisensory acquisition platform as a foundation for future respiratory-event research. Clinical or outpatient applicability, however, has not been demonstrated at this stage and would require validation with human participants and comparison against the diagnostic gold standard.

References

[1] J. B. Skatrud e J. A. Dempsey, "Airway resistance and respiratory muscle function in snorers during NREM sleep", *J. Appl. Physiol.*, vol. 59, n.º 2, pp. 328–335, agosto de 1985. Consult. 2026-06-04. [Em linha]. Disponível: https://doi.org/10.1152/jappl.1985.59.2.328.

[2] M. J. Sateia, "International classification of sleep disorders-third edition", *Chest*, vol. 146, n.º 5, pp. 1387–1394, novembro de 2014. Consult. 2026-06-04. [Online]. Disponível: https://doi.org/10.1378/chest.14-0970.

[3] K. Abu, M. L. Khraiche e J. Amatoury, "Obstructive sleep apnea diagnosis and beyond using portable monitors", *Sleep Med.*, dezembro de 2023. Consult. 2026-06-04. [Online]. Disponível: https://doi.org/10.1016/j.sleep.2023.11.034.

[4] A. Osa-Sanchez, J. Ramos-Martinez-de-Soria, A. Mendez-Zorrilla *et al.*, "Wearable Sensors and Artificial Intelligence for Sleep Apnea Detection: A Systematic Review", *J Med Syst*, v. 49, p. 66, 2025. [Online]. Disponível: https://doi.org/10.1007/s10916-025-02199-8.

[5] L.-A. Lee *et al.*, "Snoring Sounds Predict Obstruction Sites and Surgical Response in Patients with Obstructive Sleep Apnea Hypopnea Syndrome", *Scientific Rep.*, vol. 6, n.º 1, julho de 2016. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.1038/srep30629.

[6] Z. Huang *et al.*, "Prediction of the obstruction sites in the upper airway in sleep-disordered breathing based on snoring sound parameters: a systematic review", *Sleep Med.*, vol. 88, pp. 116–133, dezembro de 2021. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.1016/j.sleep.2021.10.015.

[7] A. Sebastian, P. A. Cistulli, G. Cohen e P. de Chazal, "Association of snoring characteristics with predominant site of collapse of upper airway in obstructive sleep apnea patients", *Sleep*, julho de 2021. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.1093/sleep/zsab176 .

[8] N. T. Tran, H. N. Tran e A. T. Mai, "A wearable device for at-home obstructive sleep apnea assessment: State-of-the-art and research challenges", *Frontiers Neurol.*, vol. 14, fevereiro de 2023. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.3389/fneur.2023.1123227.

[9] Y.-Y. Lin, H.-T. Wu, C.-A. Hsu, P.-C. Huang, Y.-H. Huang e Y.-L. Lo, "Sleep Apnea Detection Based on Thoracic and Abdominal Movement Signals of Wearable Piezoelectric Bands", *IEEE J. Biomed. Health Inform.*, vol. 21, n.º 6, pp. 1533–1545, novembro de 2017. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.1109/jbhi.2016.2636778.

[10] G. Bazoukis *et al.*, "Application of artificial intelligence in the diagnosis of sleep apnea", *J. Clin. Sleep Medicine*, março de 2023. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.5664/jcsm.10532.

[11] D. M. Domingues *et al.*, "Sleep prediction using data from oximeter, accelerometer and snoring for portable monitor obstructive sleep apnea diagnosis", *Scientific Rep.*, vol. 14, n.º 1, outubro de 2024. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.1038/s41598-024-75935-8.

[12] S. Dragic *et al.*, "Obstructive sleep apnea screening among hypertensive patients in low-resource settings", *Cureus*, fevereiro de 2025. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.7759/cureus.79710.

[13] P. F. D. Santana, "Monitoramento do equilíbrio em idosos: acelerômetros como aliados na prática clínica", Dissertação para Mestrado em Engenharia Biomédica, Univ. Fed. Pernambuco, Recife, 2026.

[14] A. D. C. C. Filho, "Dispositivo para monitorização de parâmetros hemodinâmicos de pacientes em reabilitação fisioterapêutica.", Dissertação para Doutorado em Ciências Biológicas, Univ. Fed. Pernambuco, Recife, 2025.

[15] 2026, Espressif Systems. "ESP32-S3-MINI1 and ESP32-S3-MINI1U Series Datasheet.", ESP32-S3. Consult. 2026-05-30. [Online]. Disponível: https://documentation.espressif.com/esp32-s3-mini-1_mini-1u_datasheet_en.pdf.

[16] 2022, MAXIM INTEGRATED. "Microphone amplifier with AGC and low-noise microphone bias", *Analog Devices*. Consult. 2026-05-30. [Online]. Disponível: https://www.analog.com/media/en/technical-documentation/data-sheets/max9814.pdf.

[17] P. R. Srijithesh, R. Aghoram, A. Goel e J. Dhanya, "Positional therapy for obstructive sleep apnoea", *Cochrane Database Systematic Rev.*, maio de 2019. Consult. 2026-07-12. [Online]. Disponível: https://doi.org/10.1002/14651858.cd010990.pub2.

[18] K. Qian *et al.*, "Classification of the excitation location of snore sounds in the upper airway by acoustic multifeature analysis", *IEEE Trans. Biomed. Eng.*, vol. 64, n.º 8, pp. 1731–1741, agosto de 2017. Consult. 2026-06-04. [Online]. Disponível: https://doi.org/10.1109/tbme.2016.2619675.

[19] D. Álvarez *et al.*, "Automated Screening of Children With Obstructive Sleep Apnea Using Nocturnal Oximetry: An Alternative to Respiratory Polygraphy in Unattended Settings", *J. Clin. Sleep Medicine*, vol. 13, n.º 05, pp. 693–702, maio de 2017. Consult. 2026-06-04. [Online]. Disponível: https://doi.org/10.5664/jcsm.6586.

[20] D. Cajal *et al.*, "Obstructive Sleep Apnea Screening by Joint Saturation Signal Analysis and PPG-derived Pulse Rate Oscillations", *IEEE J. Biomed*. Health Inform., pp. 1–11, 2023. Consult. 2026-06-04. [Online]. Disponível: https://doi.org/10.1109/jbhi.2023.3331947.

[21] P. Pahari *et al.*,"Prolonged lung-to-finger circulation time indicates an increased risk of intermittent hypoxaemia in sleep apnoea patients", *ERJ Open Res.*, vol. 10, n.º 4, art. 01051-2023, jul. de 2024. Consult. 2026-06-28. [Online]. Disponível: https://doi.org/10.1183/23120541.01051-2023.

Método Adaptativo Baseado em Variação Espectral para Compressão de Sinais de ECG de Longa Duração

Victor Valadares de Morais
Prog. de Pós-Graduação em Eng. Elétrica
Universidade Federal de Juiz de Fora
Juiz de Fora, Brasil
victor.morais@ufjf.br

Eder Barboza Kapisch
Depto. de Circuitos Elétricos
Universidade Federal de Juiz de Fora
Juiz de Fora, Brasil
eder.kapisch@ufjf.br

Pedro Xavier de Oliveira
Depto. de Elet. e Eng. Biomédica
Universidade Estadual de Campinas
Campinas, Brasil
pedrox@unicamp.br

Carlos Eduardo Batista de Lima
Depto. de Clínica Geral
Universidade Federal do Piauí
Teresina, Brasil
carlos.lima@ufpi.edu.br

***Abstract*—As doenças cardiovasculares lideram a mortalidade global, e o monitoramento eletrocardiográfico (ECG) contínuo é essencial para a detecção precoce de arritmias. Contudo, exames de longa duração, como o Holter, geram volumes de dados que desafiam a capacidade de armazenamento dos dispositivos e sobrecarregam a transmissão sem fio. Este trabalho propõe um algoritmo de compressão de ECG com perdas controladas, baseado na variação do conteúdo espectral obtido por Transformada Rápida de Fourier (FFT), com limiares de tolerância exponenciais adaptativos por componente harmônica, seguido de compressão sem perdas pelo algoritmo LZW. O método foi validado com sinais do banco de arritmias MIT-BIH e avaliado por um cardiologista em protocolo cego, alcançando taxa de compressão média de 154,2 e índices gerais de sensibilidade e especificidade de 91,71% e 99,97%. Nos registros avaliados, o algoritmo reduziu substancialmente o espaço de armazenamento necessário, preservando as características morfológicas e temporais associadas às arritmias estudadas.**

***Keywords*—Compressão de dados, ECG, Variação Espectral**

I. Introdução

Segundo a Organização Mundial da Saúde (OMS), as Doenças Cardiovasculares (DCV) permanecem como a principal causa de morte no mundo, tendo sido associadas a 19,2 milhões de óbitos em 2023 [1, 2], sem perspectiva de melhoria nas projeções para os próximos anos [3]. Sua prevalência cresce acentuadamente com a idade, atingindo mais de 80% da população acima de 80 anos nos Estados Unidos [4]. A elevada incidência está fortemente associada a fatores comportamentais cada vez mais presentes na sociedade contemporânea, tais como sedentarismo, obesidade, tabagismo, consumo excessivo de álcool e má alimentação [5, 6].

Entre as patologias do sistema cardiovascular destacam-se as arritmias cardíacas [7], caracterizadas por alterações na atividade elétrica do coração que modificam o ritmo e a eficácia dos batimentos [8]. Dependendo de sua natureza e gravidade, podem desencadear complicações severas, como acidente vascular cerebral (AVC), infarto agudo do miocárdio, embolia sistêmica e em casos extremos, a morte [9, 10]. Dessa forma, a detecção precoce das arritmias torna-se essencial para possibilitar intervenções antecipadas e reduzir os riscos associados à progressão dessas enfermidades [11].

A identificação das arritmias é realizada por meio do monitoramento dos sinais eletrocardiográficos, utilizando dispositivos como monitores cardíacos, eletrocardiógrafos e sistemas Holter. O Holter é um equipamento usado para registrar continuamente a atividade elétrica do coração por períodos superiores aos exames convencionais, permitindo a detecção de arritmias intermitentes que podem não ser observadas em registros de curta duração. Em geral, o dispositivo armazena os sinais de Eletrocardiograma (ECG) em cartões de memória SD durante intervalos que variam de 2 a 5 dias.

Entretanto, determinadas arritmias podem ocorrer de forma esporádica, apresentando longos intervalos entre os episódios, frequentemente superiores a 120 horas [12, 13]. Nesses casos, torna-se necessário prolongar o período de monitoramento para garantir o registro adequado das alterações rítmicas e morfológicas presentes no traçado eletrocardiográfico. Contudo, a capacidade limitada de memória dos dispositivos pode comprometer a captura e o armazenamento dessas informações de interesse. Além disso, em aplicações que envolvem transmissão sem fio dos sinais, é desejável que os arquivos possuam tamanho reduzido, minimizando o tempo de transmissão e o consumo de recursos do sistema [14].

Para mitigar essas restrições de armazenamento e transmissão, diferentes abordagens de compressão de sinais de ECG têm sido propostas na literatura. Em [15], foram integradas a Transformada Discreta de Fourier (DFT), a transformada *wavelet* e o *Run-Length Encoding* (Código *Run-Length*) (RLE), utilizando a mesma base de dados avaliada no presente estudo, a qual será detalhada na Seção II. Já em [16], o método de compressão sem perdas de sinais eletrocardiográficos proposto, combina um autoencoder convolucional binário (*Binary Convolutional Auto-Encoder*, BCAE) com uma rede de compensação do erro residual (*Residual Error Compensation*, REC). O BCAE converte o sinal em códigos binários, permitindo aumentar a capacidade da camada latente

sem reduzir a taxa de compressão, enquanto a REC corrige as diferenças entre o sinal original e o sinal reconstruído. O trabalho também foi avaliado com a mesma base de dados utilizada no presente projeto. Apesar desses avanços em hibridização e automação de parâmetros, o equilíbrio ideal entre altas taxas de compressão e preservação da morfologia clínica em registros de ultra-longa duração permanece um desafio em aberto.

Diante desse contexto, este trabalho tem como objetivo apresentar um algoritmo capaz de comprimir sinais de ECG de longa duração, focado em minimizar o espaço de armazenamento e, ao mesmo tempo, mantendo uma reconstrução de alta fidelidade. Busca-se preservar as principais características morfológicas e temporais do traçado, de modo a garantir uma análise visual limpa e confiável por parte dos cardiologistas, viabilizando um diagnóstico preciso.

O restante do trabalho está dividido da seguinte forma: a Seção II detalha a arquitetura do sistema de compressão proposto e as métricas de avaliação; a Seção III apresenta a análise de desempenho do algoritmo e a validação clínica; por fim, a Seção IV reúne as conclusões do estudo.

II. Materiais e métodos

A. Sistema de compressão

O método de compressão desenvolvido neste trabalho utiliza a Transformada Rápida de Fourier (FFT) para obter o conteúdo espectral de cada ciclo do sinal de entrada. Cada componente espectral (real e imaginária) de um ciclo é comparada sequencialmente com cada componente correspondente do ciclo subsequente, com o objetivo de armazenar apenas informações relevantes relacionadas às arritmias cardíacas que o paciente possa apresentar, além de reduzir o tamanho do arquivo utilizado para o armazenamento dos dados.

O fluxograma de funcionamento do algoritmo pode ser visto na Figura 1 e os diversos estágios desse sistema são descritos a seguir.

1) Filtragem: Para a remoção dos harmônicos da rede, foram projetados dois filtros digitais *notch* do tipo *butterworth* Resposta ao Impulso Infinita (IIR) de ordem 2, com as frequências de corte localizadas em 60 Hz e 120 Hz, respectivamente. Para a eliminação das variações da linha de base do sinal, um filtro digital IIR passa altas de ordem 2 do tipo *butterworth*, com frequência de corte igual a 0,5 Hz foi projetado.

2) Estimação da frequência cardíaca: A estimação da frequência cardíaca usada neste trabalho foi baseada no Intervalo RR do sinal amostrado [17]. As frequências de cada ciclo do ECG foram utilizadas no processo de Interpolação/Subamostragem do sinal. Inicialmente, para destacar a onda R, foi implementado um filtro digital IIR *butterworth* de ordem 2, com banda passante de 0,5 a 25 Hz, seguido de um filtro digital derivativo de primeira ordem, seguido de um amplificador quadrático, que eleva ao quadrado o valor de todas as amostras da saída do filtro derivativo. Por fim, os valores resultantes de todo o processo, que superaram um limiar pré-estabelecido, foram candidatos a serem pontos do ápice das ondas R. Para garantir a seleção de somente um ponto por ciclo, o estimador foi programado para eleger

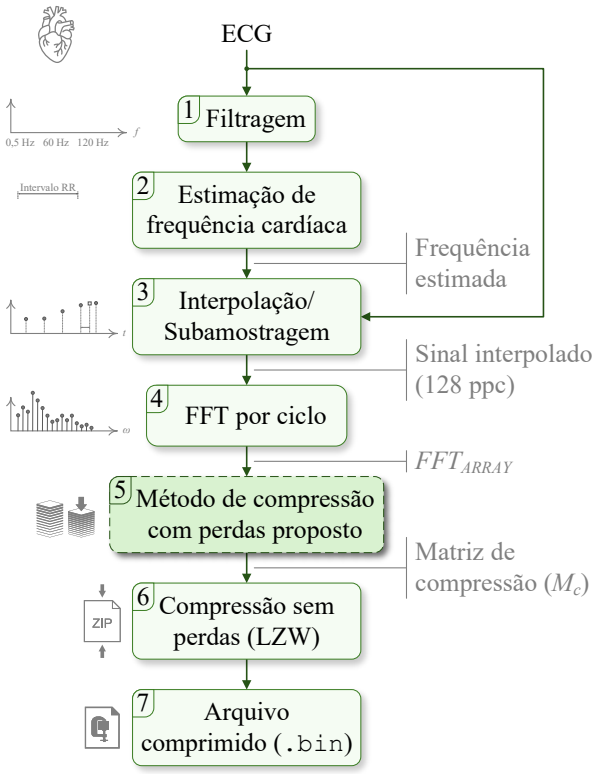


Fig. 1. Fluxograma geral do sistema de compressão de sinais de ECG proposto, ilustrando as etapas sequenciais de pré-processamento, análise espectral e codificação.

apenas um ponto a cada 300 ms. Esse valor foi definido empiricamente, assegurando uma detecção robusta.

3) Interpolação/Subamostragem: O estágio de reamostragem tem a função de normalizar temporalmente os ciclos cardíacos, assegurando que cada ciclo tenha $N = 128$ amostras igualmente espaçadas. Essa padronização dimensional é um requisito matemático e estrutural para garantir o cálculo eficiente da FFT na etapa subsequente. Com base na frequência cardíaca previamente estimada, o algoritmo ajusta dinamicamente a resolução de cada ciclo. Caso o ciclo original contenha mais de 128 amostras, o sinal passa por um processo de subamostragem. Em contrapartida, se o ciclo for mais curto, novos pontos são gerados (interpolação). O método de reamostragem adotado foi a interpolação com polinômio de Lagrange de grau três no domínio do tempo [18]. Essa técnica foi implementada empregando a estrutura de Farrow [19, 20].

4) FFT por ciclo: Após o estágio de reamostragem, o sistema aplica a FFT a cada ciclo. O resultado dessa operação é um conjunto de $K = 128$ números complexos, cada um com sua parte real e imaginária correspondente a cada componente espectral do ciclo em análise. Dessa forma, a saída da FFT, denominada $X[k], k = 0, 1, \ldots, K-1$ contém os K bins resultantes da transformada. Como o sinal de entrada é real, os coeficientes da FFT apresentam simetria conjugada, isto é, $X[k] = X^*[K-k], k = 1, \ldots, K-1$, onde $(\cdot)^*$ denota o conjugado complexo. Consequentemente, o módulo do espectro possui simetria par e a fase possui simetria ímpar, tornando redundante a segunda metade dos coeficientes espectrais [21, 22].

Além disso, na ordenação convencional da FFT, os coefi-

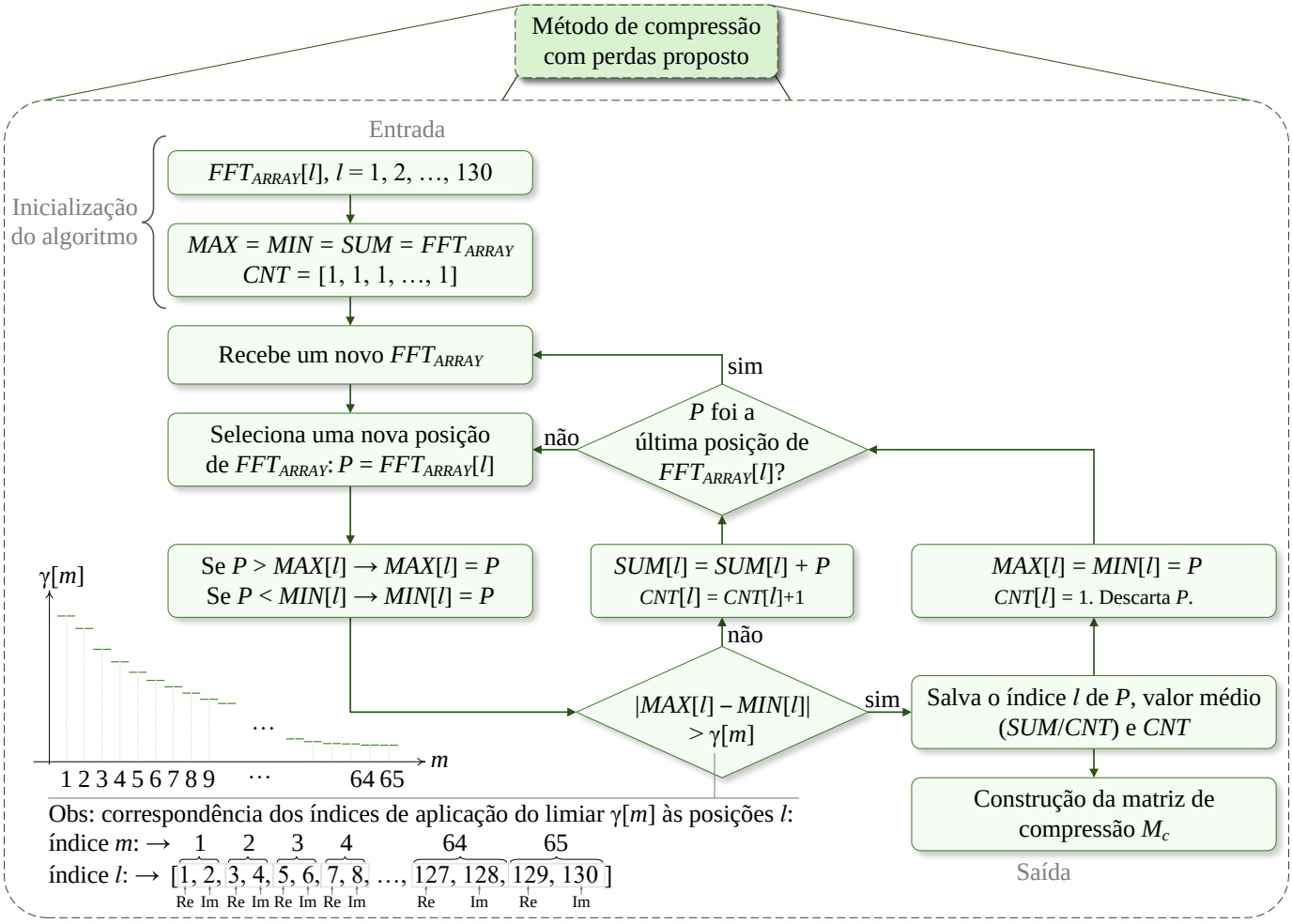


Fig. 2. Fluxograma do método de compressão com perdas proposto.

cientes são dispostos iniciando pelo componente de frequência nula (DC), seguido pelas frequências positivas até a frequência de Nyquist e, posteriormente, pelas frequências negativas em ordem decrescente [23]. Dessa forma, toda a informação espectral de um sinal real pode ser preservada utilizando apenas os primeiros $K/2 + 1$ coeficientes complexos da FFT.

Para a FFT de 128 pontos empregada neste trabalho, isso corresponde a $\frac{K}{2} + 1 = \frac{128}{2} + 1 = 65$ coeficientes complexos. Esses coeficientes são organizados num vetor denominado FFT_{ARRAY} mediante o armazenamento sequencial e intercalado de suas partes real e imaginária. Como resultado, obtém-se um vetor de comprimento $L = 130$, composto por 65 partes reais e 65 partes imaginárias.

O vetor FFT_{ARRAY} atua como o elo de transição dos dados do sistema, sendo transferido para o estágio subsequente, servindo como base para o método de compressão com perdas proposto.

5) Método de compressão com perdas proposto: O diagrama de funcionamento desse método de compressão pode ser visto na Figura 2.

Inicialmente, o conteúdo do vetor FFT_{ARRAY} relativo ao primeiro ciclo é copiado para os vetores MAX, MIN e SUM. Estes, por sua vez, possuem o mesmo comprimento de FFT_{ARRAY}, isto é, $L = 130$ posições, e servem para armazenar, respectivamente, o valor máximo, o valor mínimo e o valor acumulado de cada posição ao longo do tempo. Um quarto vetor, denominado vetor CNT, também de mesmo tamanho, recebe o valor unitário em todas as suas posições. A notação adotada para se referir ao índice destes vetores é l, que varia de 1 a 130.

Com os vetores inicializados, o sistema aguarda um novo FFT_{ARRAY}, correspondente ao próximo ciclo do sinal. A partir disso, todas as amostras desse novo vetor são comparadas com os valores de MIN e MAX, previamente armazenados. Caso o valor presente em alguma posição de índice l do FFT_{ARRAY} seja maior que o valor daquela mesma posição do vetor MAX ou menor que o da posição correspondente do vetor MIN, o MIN ou o MAX é atualizado naquela posição l. Com esses vetores atualizados, é feita uma comparação, para cada posição, do valor absoluto da diferença entre $MIN[l]$ e $MAX[l]$ com um valor de limiar (ou tolerância) específico para cada posição denominado γ. Se essa diferença for menor que a tolerância, o sistema interpreta que não há diferença entre os valores naquela posição e os vetores SUM e CNT sofrem as seguintes alterações: ao vetor SUM, é acumulado o valor do resultado da FFT e ao vetor CNT ocorre o incremento do valor unitário, ambos na posição l em análise. Portanto, SUM armazena o valor acumulado para cada posição dos vetores contendo os espectros dos ciclos do ECG (FFT_{ARRAY}) e o vetor CNT armazena o número de vezes que esse processo se repetiu. Dessa maneira, é possível calcular o valor médio de cada posição l dividindo SUM por CNT, elemento por elemento.

Se o valor absoluto da diferença entre MAX e MIN for

maior que a tolerância, o sistema entende que existe uma variação relevante para aquela posição l do vetor e salva todas as informações necessárias à reconstrução do sinal adquiridas até aquele momento. São elas: a posição em análise l o valor médio daquela posição ($SUM[l]/CNT[l]$) e o valor de CNT. Esses dados formam a matriz de compressão (M_c). Em seguida, o sistema é reiniciado na mesma posição em análise.

Quando todas as posições forem analisadas, o sistema recebe o próximo FFT_{ARRAY} contendo o espectro do próximo ciclo do sinal, repetindo-se todos os passos. O processo se encerra quando todos os ciclos de interesse do sinal de ECG em análise forem processados.

A matriz M_c possui três linhas e J colunas. A primeira linha contém as posições l. A segunda linha contém os valores de CNT de cada posição l da primeira linha. A terceira contém os valores médios também para as mesmas posições.

Em boa parte dos ECGs, as amplitudes das componentes harmônicas diminuem à medida que a ordem harmônica aumenta [24]. Esse comportamento pode ser visto no domínio da frequência. Portanto, para se realizar uma justa comparação, é necessário usar um valor de tolerância γ diferente para cada componente espectral da FFT em vez de um valor único para todas elas. Essa estratégia aumenta a eficácia do método e a qualidade da reconstrução do sinal. A variação dos valores de γ segue uma distribuição exponencial decrescente, pois essa é a função matemática que mais se assemelha ao comportamento dos espectros da maioria dos ECGs. Sua expressão matemática pode ser descrita pela seguinte função:

$$\gamma[m] = T\, m^{-\beta}, \quad 1 \leq m \leq 65, \tag{1}$$

onde T é um fator de ganho, m é o índice da posição do vetor função de limiares $\gamma[m]$ e β é o expoente responsável pela variação exponencial dos valores de tolerância. É importante salientar que uma posição m relaciona-se com a parte real e imaginária de um bin do espectro juntas, enquanto uma posição l corresponde a uma única componente, real ou imaginária. Dessa forma, m varia de 1 a 65, enquanto l varia de 1 a 130. Por causa disso, o mesmo valor de tolerância é aplicado a ambas as partes, real e imaginária, de cada bin do espectro.

6) Compressão sem perdas (Algoritmo LZW): A última etapa do processo de compressão consistiu na aplicação de um método de compressão sem perdas na matriz M_c. Quando todo espaço de memória destinado a M_c é preenchido após sua construção, esses dados passam pelo processo de compressão sem perdas (Algoritmo Lempel-Ziv-Welch (LZW)), disponibilizando um arquivo `.bin`. Por fim, a matriz é reinicializada para uma próxima execução. O algoritmo LZW foi escolhido para esse estágio, pois ele oferece altas taxas de compressão e existem versões *opensource* desse algoritmo disponíveis para os desenvolvedores [25].

7) Arquivo comprimido: Ao final, tem-se um arquivo comprimido no formato `.bin`, que otimiza o espaço de armazenamento. Este arquivo pode ser transmitido para um servidor remoto ou pode ser armazenado em algum dispositivo de memória para posterior reconstrução e investigação de patologias.

B. Reconstrução do sinal

Para reconstruir o sinal, é necessário, primeiramente, realizar a descompressão do arquivo (formato `.bin`), usando o algoritmo LZW e assim, recuperar M_c, a qual possui 3 linhas e J colunas.

Em seguida, os dados de M_c são manipulados buscando montar vários vetores A_i. Cada um desses vetores irá conter os espectros de frequência de cada ciclo do sinal a ser reconstruído, sendo que as componentes reais e imaginárias são armazenadas separadamente, ou seja, uma componente espectral ocupa duas posições de A_i, como feito em FFT_{ARRAY}. Para tal manipulação, são utilizados os valores das posições l, os CNT e valores médios (SUM/CNT). Como mencionado anteriormente, estes valores correspondem a primeira, segunda e terceira linhas de M_c, respectivamente.

Para facilitar o entendimento do processo de reconstrução do sinal a partir da matriz M_c na prática, considere o exemplo didático de uma matriz M_c contendo apenas $J = 4$ colunas, conforme apresentado em (2). O índice das colunas de M_c será denotado por j, que varia de 1 até $J = 4$. Além disso, o comprimento dos vetores A_i é 3.

$$M_c = \begin{array}{r} \\ l \\ \text{CNT} \\ \text{SUM/CNT} \end{array} \begin{array}{c} \begin{array}{cccc} j=1 & j=2 & j=3 & j=4 \end{array} \\ \begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & 1 & 2 & 1 \\ 5 & 4 & 9 & 7 \end{bmatrix} \end{array} \tag{2}$$

De posse dos valores médios e CNT, a reconstrução é feita, inicialmente, analisando o valor da posição l, o valor médio e o CNT da primeira coluna ($j = 1$). Se o valor de CNT for maior que zero, copia-se o valor médio na posição l do vetor A_i. Após isso, um valor unitário é decrementado de CNT da mesma coluna da posição l em análise. Quando o vetor A_i estiver totalmente preenchido, o processo é reiniciado para a criação de um próximo A_i.

Por exemplo, se para a posição $l = 1$ de M_c, o valor médio foi 5 e CNT foi 2, copia-se então o valor médio para a posição $l = 1$ do primeiro vetor ($A_1[1] = 5$) e o valor de CNT é decrementado para 1. O mesmo procedimento é feito de forma subsequente para as próximas colunas $j = 2$ e $j = 3$ de M_c, correspondentes às posições $l = 2$ e $l = 3$, uma vez que o valor de CNT é maior que zero nessas colunas, a fim de completarmos A_1. Após isso, como todas as posições de A_1 foram preenchidas, A_1 está concluído e inicia-se a criação do segundo vetor (A_2). O vetor A_1 completo e a atualização de M_c, podem ser vistos em (3) e (4), respectivamente.

$$A_1 = \begin{array}{c} \begin{array}{ccc} l=1 & l=2 & l=3 \end{array} \\ \begin{bmatrix} 5 & 4 & 9 \end{bmatrix} \end{array} \tag{3}$$

$$M_c = \begin{array}{r} \\ l \\ \text{CNT} \\ \text{SUM/CNT} \end{array} \begin{array}{c} \begin{array}{cccc} j=1 & j=2 & j=3 & j=4 \end{array} \\ \begin{bmatrix} 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 1 \\ 5 & 4 & 9 & 7 \end{bmatrix} \end{array} \tag{4}$$

Realiza-se o mesmo procedimento para a construção do vetor A_2. Porém, neste caso, nota-se que o CNT referente à coluna

$j = 2$ da versão atualizada da matriz M_c é zero. Isso significa que todas as cópias relacionadas ao valor médio daquela coluna já foram feitas. Sendo assim, o algoritmo busca outra coluna que contém a posição $l = 2$ com CNT diferente de zero para utilizar na criação de A_2. Nesse caso, a coluna $j = 4$ de M_c atende a essa condição e o seu valor médio 7 será usado para preencher a posição $l = 2$ de ($A_2[2] = 7$). Os resultados da criação do segundo vetor e a nova atualização de M_c podem ser vistos em (5) e (6), respectivamente.

$$A_2 = \begin{bmatrix} \overset{l=1}{5} & \overset{l=2}{7} & \overset{l=3}{9} \end{bmatrix} \tag{5}$$

$$M_c = \begin{array}{r} \\ l \\ \text{CNT} \\ \text{SUM/CNT} \end{array} \begin{array}{c} \begin{array}{cccc} \overset{j=1}{} & \overset{j=2}{} & \overset{j=3}{} & \overset{j=4}{} \end{array} \\ \begin{bmatrix} 1 & 2 & 3 & 2 \\ 0 & 0 & 0 & 0 \\ 5 & 4 & 9 & 7 \end{bmatrix} \end{array} \tag{6}$$

Como nesse exemplo não há mais colunas em M_c para serem processadas, verifica-se se todos os valores de CNT são nulos (segunda linha de M_c composta apenas de zeros), significando que não há mais vetores A_i a serem criados.

Os últimos estágios do processo de reconstrução consistem na realização da recomposição dos espectros de cada ciclo a partir da propriedade de simetria, aplicação da transformada de Fourier inversa em cada vetor recomposto e concatenação horizontal para obter o sinal reconstruído no domínio do tempo.

C. Custo computacional e requisitos de memória do método

O custo computacional foi avaliado analiticamente pelo número de operações aritméticas reais por ciclo cardíaco, sendo M o número de amostras brutas do ciclo e $N = 128$ o número de amostras após a reamostragem. Todos os estágios apresentam custo linear, exceto a FFT, de custo $O(N \log_2 N)$, resultando em complexidade total $O(M + N \log_2 N)$ por ciclo. A Tabela I detalha a contagem por estágio. Cabe destacar que os 65 valores de $\gamma[m]$ são calculados uma única vez a partir de T e β e armazenados em uma tabela, de modo que o limiar adaptativo não acrescenta custo em tempo de execução. A memória de trabalho totaliza cerca de 22 kB e é constante, independente da duração do registro, pois apenas o arquivo comprimido de saída cresce ao longo da aquisição, o que viabiliza o processamento contínuo de exames de longa duração em dispositivos com memória limitada.

TABELA I
CONTAGEM DE OPERAÇÕES REAIS POR CICLO CARDÍACO (FREQUÊNCIA CARDÍACA DE 75 BPM, $M = 288$, $N = 128$, $L = 130$).

Estágio	**Ordem**	**Mult.**	**Adições**	**%**
Filtragem (3 seções de 2ª ordem)	$O(M)$	4.320	3.456	39,2
Estimação da freq. cardíaca	$O(M)$	1.728	1.728	17,4
Reamostragem (Farrow)	$O(N)$	1.408	1.792	16,2
FFT de 128 pontos	$O(N \log_2 N)$	1.792	2.688	22,6
Comparação espectral e M_c	$O(L)$	–	910	4,6
Total		**9.248**	**10.574**	100

D. Avaliação da qualidade de compressão dos sinais

Para demonstrar o desempenho do método proposto, foram utilizados sinais de ECG do banco de dados de arritmia MIT-BIH, devido à sua ampla utilização em trabalhos na literatura [15, 16] e a presença de diversas arritmias, desde as mais comuns até as mais raras. Além disso, os sinais do banco são contaminados por interferências e outros distúrbios que proporcionam o caráter realístico de aquisição dos ECGs. Foram utilizados 27 registros ambulatoriais, com duração de 30 minutos cada, amostrados a 360 Hz e com resolução de 11 bits [26].

Os registros selecionados contêm as seguintes arritmias: fibrilação atrial (FA), extrassístole ventricular (EV), extrassístole supraventricular (ESV), taquicardia ventricular (TV), taquicardia supraventricular (TSV), síndrome de Wolff-Parkinson-White (WPW) e flutter atrial (FLA), definidas segundo a diretriz da Sociedade Brasileira de Cardiologia [27].

Além disso, eles também apresentam, em sua maioria, um elevado número de eventos com variações rítmicas e morfológicas, artefatos de movimento e presença de ruído intenso em partes dos traçados.

Para avaliar o desempenho de compressão do sistema proposto, foi usada a taxa de compressão (CR) como métrica. Ela é caracterizada pela razão entre o tamanho do sinal original descomprimido e o mesmo sinal, porém na sua forma comprimida. Tal métrica foi calculada individualmente para cada um dos registros utilizados. Além dessas, a média das taxas de compressão de todos os registros utilizados (CR_μ) e o seu respectivo desvio padrão (CR_σ) também foram calculados. Tais métricas podem ser averiguadas em (7) e (8), respectivamente.

$$CR_\mu = \frac{1}{N} \sum_{i=1}^{N} CR_i \tag{7}$$

$$CR_\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(CR_i - CR_\mu\right)^2} \tag{8}$$

As métricas de qualidade de reconstrução usadas neste trabalho são: a diferença percentual da raiz quadrada média (*percentage root-mean-square difference* (PRD)), a média dos valores de PRD (PRD_μ) e o seu desvio padrão (PRD_σ), definidas por (9), (10) e (11), respectivamente.

$$PRD = \sqrt{\frac{\sum_{n=1}^{N_a} [x[n] - x'[n]]^2}{\sum_{n=1}^{N_a} [x[n]]^2}} \times 100\% \tag{9}$$

$$PRD_\mu = \frac{1}{N_r} \sum_{i=1}^{N_r} PRD_i \tag{10}$$

$$PRD_\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(PRD_i - PRD_\mu\right)^2} \tag{11}$$

Sendo CR_i e PRD_i as notações que correspondem aos valores de CR e PRD, respectivamente, para cada sinal, N_r, o número de registros avaliados, N_a, o número de amostras de um sinal, $x[n]$, a n-ésima amostra do sinal original e $x'[n]$, a amostra correspondente do sinal reconstruído.

E. Análise da qualidade de reconstrução dos sinais

Para aumentar a credibilidade do método proposto, foi feita, por um cardiologista com experiência na interpretação de eletrocardiogramas, a inspeção visual dos sinais reconstruídos após o processo de compressão. Os sinais da base de dados foram comprimidos, utilizando $T = 0,03$ e $\beta = 0,1$. Tanto os sinais originais quanto suas versões reconstruídas foram analisados pelo especialista. Os valores de T e β foram obtidos previamente no estudo reportado em [28].

A participação do cardiologista teve caráter complementar à avaliação quantitativa apresentada neste trabalho e não foi considerada, isoladamente, como uma validação clínica do algoritmo. Dessa forma, os resultados obtidos pelo especialista foram analisados em conjunto com as métricas objetivas de desempenho empregadas para avaliar a qualidade dos sinais reconstruídos.

Os critérios para diagnóstico eletrocardiográfico de aspectos morfológicos, intervalos e laudo conclusivo foram baseados na diretriz da Sociedade Brasileira de Cardiologia sobre análise e emissão de laudos eletrocardiográficos [27]. Nesta análise, foram avaliados os parâmetros de regularidade do intervalo RR, batimento a batimento, amplitude e duração do QRS, morfologia das ondas P e T, focando em identificar corretamente todas as arritmias, caso houvesse.

Vale ressaltar que o médico não sabia de nenhuma informação prévia sobre os eventos de anormalidades cardíacas contidos nos sinais. Os diversos parâmetros e registros obtidos da avaliação do cardiologista foram utilizados para identificar os eventos em que o médico relatou a ausência de arritmias no sinal reconstruído, mas que havia no original (Falsos negativos (Fn)), os eventos em que o médico indicou a presença de arritmias no sinal reconstruído, mas que não havia no original (Falsos positivos (Fp)), os eventos em que o cardiologista registrou corretamente a ausência de arritmias em ambos os sinais (Verdadeiros negativos (Vn)) e os eventos em que o cardiologista indicou corretamente a presença de arritmias em ambos os sinais (Verdadeiros positivos (Vp)). Utilizando esses registros, foram feitos os cálculos e análises da sensibilidade (SE) e especificidade (EP) de cada arritmia cardíaca encontrada nos bancos de dados, baseado nas equações (12) e (13).

$$SE = \frac{Vp}{Vp + Fn} \tag{12}$$

$$EP = \frac{Vn}{Vn + Fp} \tag{13}$$

É importante observar que a sensibilidade, neste caso, ilustra a probabilidade do médico identificar no sinal reconstruído a arritmia alvo corretamente. Já a especificidade apresenta a probabilidade do médico identificar, em concordância com o laudo do sinal original, as regiões do sinal reconstruído onde não há a arritmia alvo.

III. Resultados e Discussões

Baseado nas informações da Tabela II e da Tabela III, considerando que existem interferências e artefatos nos sinais originais que podem prejudicar a detecção, pode-se dizer que o método de compressão proposto apresentou boas taxas de compressão e resultados de qualidade de reconstrução satisfatórios. Em relação aos valores de CR, de acordo com a Tabela II, o método alcançou altas taxas de compressão para a maioria dos sinais.

TABELA II
Valores de CR e PRD para cada sinal avaliado. Os valores máximos de cada métrica estão destacados em negrito.

Sinais	CR	PRD (%)
1	115	16,28
2	360	14,54
3	69	17,15
4	**525**	13,13
5	183	8,99
6	141	20,35
7	131	14,02
8	97	13,74
9	169	16,03
10	144	15,10
11	140	15,82
12	295	17,75
13	177	18,37
14	253	13,70
15	140	15,11
16	40	15,85
17	56	13,00
18	56	14,67
19	70	8,43
20	54	15,24
21	136	14,03
22	47	13,19
23	44	15,96
24	98	**26,28**
25	98	14,36
26	102	16,42
27	424	11,10
Média $\mu \rightarrow$	CR_μ = 154,2	PRD_μ = 15,13
Desvio $\sigma \rightarrow$	CR_σ = 120,8	PRD_σ = 3,38

TABELA III
Sensibilidade e especificidade obtida por cada classe de arritmia.

Arritmias	**Sensibilidade (%)**	**Especificidade (%)**
EV	93,54	99,96
ESV	84,54	99,94
TV	100,00	99,81
TSV	81,82	99,70
WPW	100,00	99,91
FA	100,00	97,86
FLA	66,67	100,00
Arritmias Gerais	91,71	99,97

Percebe-se também que o método apresenta, para alguns sinais, valores elevados de PRD. A razão disso é que o sistema de compressão armazena somente as informações relevantes do sinal, logo, involuntariamente, ele filtra boa parte de suas componentes espectrais de alta frequência, resultando em um sinal reconstruído menos ruidoso que o original. Um exemplo desse acontecimento pode ser observado na Figura 3.

Nota-se também que, com exceção da arritmia flutter atrial, o valor do parâmetro sensibilidade foi superior a 80%, atingindo 100% em alguns casos. Esse resultado demonstra que, na maioriados casos, o algoritmo preserva as informações relevantes (manifestações rítmicas e morfológicas das arritmias) dos sinais de ECG originais, nos seus respectivos registros reconstruídos e evidencia que a métrica PRD não foi eficaz, neste caso, para avaliar a qualidade da compressão, devido à

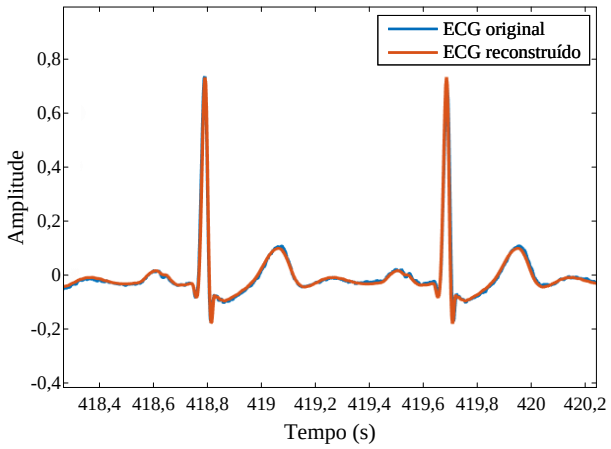


Fig. 3. Um exemplo de um sinal ECG original e reconstruído.

filtragem involuntária realizada pelo sistema, como mencionado anteriormente.

Já o valor da especificidade, por sua vez, foi superior a 97% para todas as arritmias avaliadas, evidenciando que o sistema não introduz alterações significativas nas regiões de normalidade dos sinais originais. Esse comportamento indica que os trechos sem arritmia foram preservados de forma a não gerar discordância relevante entre os laudos emitidos sobre o traçado original e sobre o reconstruído.

Para o caso da FLA, sabe-se que ela possui aspecto serrilhado em seu traçado, sendo usualmente conhecido como "Dente de serrote". No entanto, algumas derivações eletrocardiográficas podem não evidenciar essa morfologia típica. Portanto, existe a possibilidade de alguns sinais enviados para a análise do especialista serem as derivações que não apresentam tal característica dessa arritmia. Esse fato justifica o valor de SE ser um pouco menor e evidencia que o problema, neste caso, não está relacionado com o método de compressão e sim com a derivação em análise.

Considerando-se o grupo genérico de arritmias (Arritmias Gerais) estruturado neste artigo, é possível comparar o presente trabalho com os que são apresentados na literatura em [15] e [16]. A Tabela IV apresenta os resultados de todos os projetos para os seguintes parâmetros: valores de CR e PRD médios (CR_μ, PRD_μ) e máximos (CR_{MAX}, PRD_{MAX}), bem como os de SE e EP. Em relação ao trabalho de [16], as métricas SE e EP não estão presentes na avaliação e os valores de CR_{MAX} e CR_μ obtidos foram inferiores aos do trabalho proposto, com destaque no CR_{MAX} registrado, pelo menos 4 vezes menor que o do presente projeto.

Já em relação ao trabalho de [15], o valor de SE do método proposto foi discretamente inferior, com uma diferença de aproximadamente 4%, enquanto os valores de EP foram similares. Contudo, se também levarmos em consideração os valores de CR_{MAX} e CR_μ obtidos, o método proposto se sobressaiu em ambos, uma vez que o CR_{MAX} registrado neste projeto é mais de 6 vezes maior que o de [15] e o médio é aproximadamente 5 vezes maior. Já em relação aos valores de PRD máximo (PRD_{MAX}) e médio (PRD_μ), é possível notar que o PRD_{MAX} do método proposto é o maior de todos e o PRD_μ é o segundo maior. Isso novamente pode ser justificado pela filtragem involuntária que o método de compressão proposto aplica no sinal ao comprimi-lo. O trabalho de [16] apresenta o menor erro, mas, como não aborda métricas de análise qualitativas, não é possível garantir que os dados reconstruídos preservam as características de interesse. Assim, o ganho expressivo na eficiência de compactação justifica a sutil redução na sensibilidade, dado que o sistema proposto alcança CR_{MAX} e CR_μ maiores que os outros trabalhos abordados.

TABELA IV
COMPARAÇÃO DE DESEMPENHO.

Abordagem	CR_μ	CR_{MAX}	PRD_μ	PRD_{MAX}	SE	EP
Método Proposto	154,2	525	15,13%	26,28%	91,71%	99,97%
Ref. [15]	31,68	76,4	11,15%	31,08%	95,77%	99,18%
Ref. [16]	117,33	117,5	7,76%	22,86%	-	-

Baseando-se nos valores dos parâmetros de sensibilidade e especificidade, pode-se dizer que os resultados das análises dos sinais reconstruídos foram, na maioria das vezes, iguais/semelhantes aos dos sinais originais. Sendo assim, é possível questionar se o uso da métrica PRD na análise de qualidade de reconstrução de ECGs, reflete ou mascara a real situação.

Para verificar o funcionamento do método, uma implementação de referência desenvolvida em MATLAB foi executada em um computador com processador AMD Ryzen 7 7735HS de 3,20 GHz e 16 GB de memória DDR5-4800. O processamento de um registro de 30 minutos demandou aproximadamente 5 s, o que corresponde a 360 vezes o tempo real de aquisição, com consumo de 300 MB de memória de processo. Ambos os valores refletem o ambiente de prototipagem adotado, que carrega o registro integralmente em memória e não foi otimizado, e não a demanda intrínseca do algoritmo, de cerca de 22 kB. Para uma frequência cardíaca de 75 bpm, as $2{,}5 \times 10^4$ operações reais por segundo de sinal estimadas na Seção II-C representam margem superior a três ordens de grandeza sobre microcontroladores de baixo consumo da classe ARM Cortex-M4, corroborando a viabilidade de implementação embarcada.

IV. Conclusão

Este trabalho apresentou um método adaptativo baseado na variação do conteúdo espectral via FFT, associado a limiares de tolerância exponenciais e à codificação LZW, voltado para a compressão eficiente de sinais de ECG de longa duração. A metodologia proposta visa mitigar as restrições de armazenamento e transmissão sem fio em sistemas de monitoramento contínuo de sinais de ECG, como o Holter.

A validação experimental utilizando a base de dados MIT-BIH, complementada por uma avaliação clínica cega conduzida por um cardiologista, indicou o potencial do algoritmo. O método alcançou uma taxa de compressão média ($CR_\mu = 154{,}2$) elevada, preservando os atributos morfológicos e temporais dos traçados nos registros avaliados. Esse desempenho refletiu-se nos elevados índices de sensibilidade (91,71%) e especificidade (99,97%) gerais obtidos na avaliação médica. Embora limitações tenham sido observadas na reconstrução da patologia de flutter

atrial (sensibilidade de 66,67%), não se observou discordância entre os laudos (100% de sensibilidade) para arritmias críticas de caráter intermitente, como a taquicardia ventricular e a Síndrome de Wolff-Parkinson-White (WPW).

Ademais, o estudo levanta uma discussão relevante acerca da métrica PRD, sugerindo que valores elevados de erro quadrático nem sempre se traduzem em perda de qualidade clínica, visto que o processo atua como um filtro atenuador de ruídos de alta frequência.

Como trabalhos futuros, pretende-se investigar a adaptação dinâmica dos parâmetros de ganho (T) e decaimento (β) para otimizar a compressão de espectros complexos como o do flutter atrial, bem como avaliar a viabilidade de implementação do algoritmo em sistemas embarcados de ultra-baixo consumo e processamento em tempo real para dispositivos de monitoramento de próxima geração. Além disso, uma avaliação futura dos sinais envolvendo múltiplos especialistas e protocolos de avaliação clínica padronizados é considerada uma importante extensão deste trabalho.

Agradecimentos

Os autores deste trabalho gostariam de agradecer à Universidade Federal de Juiz de Fora (UFJF), à Universidade Estadual de Campinas (UNICAMP) e à Universidade Federal do Piauí (UFPI) pelo apoio institucional a este estudo. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

References

[1] R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Arrhythmia detection based on time–frequency features of heart rate variability and back-propagation neural network," *Iran Journal of Computer Science*, vol. 2, no. 4, pp. 245–257, 2019.

[2] G. B. of Cardiovascular Diseases and R. . Collaborators, "Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990-2023," *Journal of the American College of Cardiology*, vol. 86, no. 22, pp. 2167–2243, 2025.

[3] K. E. Joynt Maddox, M. S. Elkind, H. J. Aparicio, Y. Commodore-Mensah, S. D. de Ferranti, W. N. Dowd, A. F. Hernandez, O. Khavjou, E. D. Michos, L. Palaniappan *et al.*, "Forecasting the burden of cardiovascular disease and stroke in the united states through 2050—prevalence of risk factors and disease: a presidential advisory from the american heart association," *Circulation*, vol. 150, no. 4, pp. e65–e88, 2024.

[4] S. Kaptoge, L. Pennells, D. De Bacquer, M. T. Cooney, M. Kavousi, G. Stevens, L. M. Riley, S. Savin, T. Khan, S. Altay *et al.*, "World health organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions," *The Lancet global health*, vol. 7, no. 10, pp. e1332–e1345, 2019.

[5] T. Pörschmann, T. Meier, and S. Lorkowski, "Cardiovascular mortality attributable to dietary risk factors in 54 countries in the who european region from 1990 to 2019: an updated systematic analysis of the global burden of disease study," *European journal of preventive cardiology*, vol. 32, no. 16, pp. 1553–1563, 2025.

[6] A. Schmidt-Trucksäss, A. H. Lichtenstein, and R. von Känel, "Lifestyle factors as determinants of atherosclerotic cardiovascular health," *Atherosclerosis*, vol. 395, p. 117577, 2024.

[7] M. Lei, S. C. Salvage, A. P. Jackson, and C. L.-H. Huang, "Cardiac arrhythmogenesis: roles of ion channels and their functional modification," *Frontiers in physiology*, vol. 15, p. 1342761, 2024.

[8] A. K. Nagpal, A. Pundkar, A. Singh, and C. Gadkari, "Cardiac arrhythmias and their management: an in-depth review of current practices and emerging therapies," *Cureus*, vol. 16, no. 8, 2024.

[9] K. C. Roberts-Thomson, D. H. Lau, and P. Sanders, "The diagnosis and management of ventricular arrhythmias," *Nature Reviews Cardiology*, vol. 8, no. 6, pp. 311–321, 2011.

[10] X. Xu, Z. Wang, J. Yang, X. Fan, and Y. Yang, "Burden of cardiac arrhythmias in patients with acute myocardial infarction and their impact on hospitalization outcomes: insights from china acute myocardial infarction (cami) registry," *BMC cardiovascular disorders*, vol. 24, no. 1, p. 218, 2024.

[11] P. Karakasis, P. Theofilis, M. Sagris, K. Pamporis, P. Stachteas, G. Sidiropoulos, P. K. Vlachakis, D. Patoulias, A. P. Antoniadis, and N. Fragakis, "Artificial intelligence in atrial fibrillation: from early detection to precision therapy," *Journal of Clinical Medicine*, vol. 14, no. 8, p. 2627, 2025.

[12] M. D. Solomon, J. Yang, S. H. Sung, M. L. Livingston, G. Sarlas, J. C. Lenane, and A. S. Go, "Incidence and timing of potentially high-risk arrhythmias detected through long term continuous ambulatory electrocardiographic monitoring," *BMC cardiovascular disorders*, vol. 16, no. 1, p. 35, 2016.

[13] R. L. Page, W. E. Wilkinson, W. K. Clair, E. A. McCarthy, and E. Pritchett, "Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia." *Circulation*, vol. 89, no. 1, pp. 224–227, 1994.

[14] T.-H. Tsai and W.-T. Kuo, "An efficient ecg lossless compression system for embedded platforms with telemedicine applications," *IEEE Access*, vol. 6, pp. 42 207–42 215, 2018.

[15] C. K. Jha and M. H. Kolekar, "Tunable q-wavelet based ecg data compression with validation using cardiac arrhythmia patterns," *Biomedical Signal Processing and Control*, vol. 66, p. 102464, 2021.

[16] J. Shi, F. Wang, M. Qin, A. Chen, W. Liu, J. He, H. Wang, S. Chang, and Q. Huang, "New ecg compression method for portable ecg monitoring system merged with binary convolutional auto-encoder and residual error compensation," *Biosensors*, vol. 12, no. 7, 2022. [Online]. Available: https://www.mdpi.com/2079-6374/12/7/524

[17] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.

[18] Y.-M. Chen and C.-C. Chen, "Design of farrow structured variable fractional delay filter for time-varying leo communication channel emulator with srrc communication waveforms," *IEEE Access*, vol. 12, pp. 122 229–122 238, 2024.

[19] C. B. Moler, *Numerical computing with MATLAB*. SIAM, 2004.

[20] N. Bakholdin, V. Chernienko, and S. Bakhurin, "Digital resampler based on farrow filter," in *2025 27th International Conference on Digital Signal Processing and its Applications (DSPA)*. IEEE, 2025, pp. 1–7.

[21] S. K. Salih and M. Hamood, "New algorithm for real-valued fourier transform," *Tikrit Journal of Engineering Sciences*, 2023.

[22] O. Ponomareva, A. Ponomarev, and N. Ponomareva, "Cross complex conjugate symmetry of the coefficients of discrete fourier transforms of real finite discrete one-dimensional and two-dimensional signals," in *27th International Conference on Digital Signal Processing and its Applications (DSPA)*, 2025, pp. 1–6.

[23] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, 2010.

[24] R. M. Rangayyan and S. Krishnan, *Biomedical signal analysis*. John Wiley & Sons, 2024.

[25] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on information theory*, vol. 23, no. 3, pp. 337–343, 1977.

[26] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[27] C. A. Pastore, J. T. d. Pinho, C. Pinho, N. Samesima, H. G. Pereira-Filho, J. C. L. Kruse, A. Paixão, A. R. Pérez-Riera, A. L. Ribeiro, C. A. R. Oliveira *et al.*, "III diretrizes da sociedade brasileira de cardiologia sobre análise e emissão de laudos eletrocardiográficos," *Arquivos Brasileiros de Cardiologia*, vol. 106, no. 4 Suppl 1, pp. 1–23, 2016. [Online]. Available: https://www.scielo.br/j/abc/a/BkbtSMLnNpD5ttRk8ccxh5P/

[28] V. V. de Morais, P. X. de Oliveira, E. B. Kapisch, and A. J. Ferreira, "Spectral variation based method for electrocardiographic signals compression," in *XXVII Brazilian Congress on Biomedical Engineering (CBEB 2020)*, ser. IFMBE Proceedings. Cham: Springer, 2022, pp. 1801–1807.

Neuroimaging, Speech, and Sensors: A Scoping Review of Deep Learning Multimodal Fusion for Cognitive Decline Detection

1st Marcos Augusto de S. Pinto
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
marcos.pinto@ufam.edu.br

2nd Marly Guimarães F. Costa
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
mcosta@ufam.edu.br

3rd Cícero Ferreira F. C. Filho
R&D Center in Electronic and Information Technology
Federal University of Amazonas
Manaus, Brazil
ccosta@ufam.edu.br

***Abstract*—Multimodal deep learning has been increasingly applied to cognitive decline detection, yet fusion strategies across different acquisition settings remain incompletely characterized. This scoping review maps deep learning-based multimodal fusion approaches for the detection and classification of mild cognitive impairment and dementia in studies published between 2021 and 2026. Evidence was identified through bibliographic database searches, retrospective citation searching, and generative artificial intelligence-assisted searches. Post hoc supplementary searches in PubMed and Web of Science were subsequently conducted to assess coverage and selection sensitivity. After screening and categorization, 310 records were organized by primary data source, from which 20 studies were purposively selected for in-depth qualitative synthesis using a reporting-completeness criterion and stratified acquisition-category quotas. The broader bibliographic corpus was strongly dominated by neuroimaging, whereas continuous sensing and eye tracking remained sparsely represented. Among the four of eight neuroimaging studies reporting unimodal baselines, multimodal fusion improved performance in all cases, whereas the two sensor studies allowing direct fusion-gain assessment reported negligible or negative differences. Speech studies were concentrated on related benchmark corpora despite employing diverse fusion mechanisms. Overall, fusion benefits varied across acquisition settings, suggesting that source complementarity may contribute to performance, although the available evidence does not allow its effect to be isolated from other methodological factors.**

***Index Terms*—Alzheimer's disease, cognitive decline, deep learning, multimodal fusion, scoping review**

I. Introduction

Early detection of *Alzheimer's Disease* (AD) and cognitive decline has become an important research direction in artificial intelligence applied to healthcare. Since no single data source captures the full complexity of cognitive decline [1], multimodal deep learning approaches have been increasingly explored to combine complementary information from speech, neuroimaging, and real-world sensors, particularly through attention mechanisms and *Transformer*-based architectures [1], [2].

Despite this progress, the landscape of multimodal fusion for cognitive decline remains incompletely characterized. Existing reviews either provide a broad overview of fusion mechanisms without a clinical focus [2] or are dominated by neuroimaging studies [1], leaving the contribution of speech, continuous sensing, and other acquisition categories unclear. Furthermore, the use of generative AI-assisted search tools for literature identification has not been sufficiently evaluated regarding coverage and reproducibility.

This study maps deep learning-based multimodal fusion approaches for the detection and classification of cognitive decline in studies involving control participants and individuals with *Mild Cognitive Impairment* (MCI) or dementia. Its contributions are threefold: (i) a **quantitative mapping** of modality distributions across recent literature; (ii) a **qualitative synthesis** of fusion strategies, architectures, and methodological limitations; and (iii) an **empirical comparison** between the three original identification routes regarding yield, overlap, and reproducibility, complemented by post hoc analyses of bibliographic database coverage and sensitivity to category-based prioritization.

The findings are interpreted in relation to the meta-analysis by Guo et al. [1], which reported a pooled AUC of 0.924 (95% CI: 0.912–0.936) for *Transformer*-based multimodal models. Whether these conclusions extend beyond the neuroimaging-dominated literature, particularly to speech and continuous sensing, remains an open question addressed by this review.

II. Methodology

This review followed the **PRISMA-ScR** (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews*) guidelines [3]. No protocol was registered for this review. The research question, "What is the current state of the art of deep learning-based multimodal fusion approaches for the detection and classification of cognitive decline?", was formulated using the **PCC** (*Population, Concept, Context*) framework recommended by the *Joanna Briggs Institute* for scoping reviews. The PCC elements were defined as follows: *Population*, adults with MCI, dementia, or Alzheimer's disease, together with cognitively healthy

controls; *Concept*, deep learning based fusion of two or more data modalities for detection, prediction, diagnosis, or classification; and *Context*, any acquisition setting, including clinical, laboratory, and real world environments, in studies published between 2021 and 2026.

A. Information sources and search strategy

Three complementary approaches were used to identify relevant studies, considering publications from **2021 to 2026**.

- **Approach 1: Bibliographic databases.** Searches were performed in the Scopus and IEEE Xplore databases. In Scopus, the query was executed using the `TITLE ABS KEY` field. In IEEE Xplore, the *Command Search* interface was used to search the *Document Title* and *Abstract* fields. In both databases, Boolean queries were developed based on the PCC framework.
- **Approach 2: Retrospective citation searching.** Reference lists of key studies related to the review scope were manually examined using backward citation searching.
- **Approach 3: Generative AI assisted search.** The Elicit and Consensus platforms were queried in natural language using *prompts* derived from the same research question.

The complete Boolean search strings used in Scopus and IEEE Xplore and the natural-language prompts used in Elicit and Consensus are available in the supplementary repository [4].

B. Eligibility criteria

The eligibility criteria were defined before screening. Eligible studies were peer-reviewed primary research articles published in English between 2021 and 2026, with full text available, that employed deep learning to fuse two or more data modalities for the detection, prediction, diagnosis, or classification of cognitive decline, MCI, Alzheimer's disease, or dementia. Reviews, surveys, meta-analyses, editorials, letters, book chapters, conference abstracts without a full paper, posters, extended abstracts, animal-only studies, studies outside the cognitive-decline domain, and studies without multimodal fusion or deep learning were excluded.

C. Operational definition of fusion

The identified modalities were grouped into families according to the nature of the signal: neuroimaging, speech and language, clinical and neuropsychological, omics, neurophysiology, sensors, and behavioral. Based on this taxonomy, two configurations were defined, both of which were considered eligible.

- **Heterogeneous fusion:** the combination of modalities from *two or more* different families, for example neuroimaging and clinical variables, or electroencephalography and magnetic resonance imaging.
- **Homogeneous fusion:** the combination of two or more modalities belonging to the *same* family, for example *sMRI* and *FDG PET*, both from the neuroimaging family, or acoustic signal and transcript, both from the speech and language family.

D. Paper Selection Process

Duplicate records were removed, and title and abstract screening was performed manually according to the predefined eligibility criteria.

Records passing the initial screening were classified according to their primary data source into three acquisition categories: neuroimaging, speech and language, and sensors, with the latter subdivided into physiological sensors, face and gaze, and environment and daily living. Primary data sources were assigned by weighted term occurrence, with matches in the title weighted 3 and matches in the abstract or keywords weighted 1.

Within each category, studies were ranked using an author-defined reporting completeness score calculated from the information available in the title and abstract. The reporting completeness score was calculated according to Equation 1.

$$S = 2M + \min(V, 6) + 3\min(X, 3) + 2L + \min(F, 5) \quad (1)$$

where M represents explicitly named performance metrics, V the number of numerical performance values reported, X the number of interpretability approaches described, L the number of validation strategies reported, and F the number of distinct modalities identified. The score was used exclusively as a prioritization criterion for qualitative synthesis and was not interpreted as a measure of methodological quality or used to determine study eligibility.

Because the categorized corpus was strongly dominated by neuroimaging studies, a purposive stratified quota was applied to preserve coverage of the three acquisition categories in the qualitative synthesis. Twenty studies were therefore selected for in-depth qualitative synthesis: eight neuroimaging-primary, seven speech-primary, and five sensor-primary studies. The quotas were intended to ensure analytical coverage of underrepresented acquisition categories rather than to reproduce the proportional distribution of the retrieved literature. Within each category, studies with higher reporting completeness scores were prioritized for selection.

During full-text charting of the 20 studies selected for qualitative synthesis, methodological characteristics relevant to the review objectives were extracted, including modalities and fusion configuration, validation strategy, participant-level data splitting when reported, external validation, unimodal baselines, quantified fusion gain, performance and agreement metrics, and the use of cognitive scores as model inputs or components of the diagnostic reference standard. These characteristics were used descriptively in the qualitative synthesis and were not used to determine eligibility or derive a methodological quality score.

A formal methodological quality or risk-of-bias assessment was not performed. Consistent with the scope of this review, the objective was to map the available evidence and methodological characteristics rather than to estimate pooled effects or establish clinical recommendations. Critical appraisal is not routinely recommended for scoping reviews and may be

undertaken when required by the specific review purpose [3], [5].

E. Supplementary coverage and sensitivity analysis

After completion of the initial review, supplementary searches were conducted in PubMed and the Web of Science Core Collection to assess whether the use of Scopus and IEEE Xplore as the original bibliographic databases materially affected the coverage and modality distribution observed in the review. These searches were treated as a post hoc coverage and sensitivity analysis rather than as part of the original identification strategy. The same publication window (2021–2026) and conceptual scope of multimodal deep learning for cognitive decline detection and classification were retained.

Records retrieved from the supplementary databases were deduplicated against the original bibliographic corpus and across the two supplementary sources. Records without abstracts and secondary publications were removed before further analysis. An automated rule-based screening procedure was then applied to prioritize records potentially consistent with the review scope. The procedure used predefined search strings and Boolean rules applied to titles and abstracts to identify records mentioning cognitive decline as the predominant clinical domain, at least two distinct modalities, multimodal fusion, deep learning, and a detection or classification objective. This automated step was used only for prioritization within the sensitivity analysis and was not interpreted as a formal eligibility assessment. The reporting completeness framework described above was subsequently reimplemented for the supplementary records using automated metadata detectors. The supplementary analysis was used to examine the robustness of the modality distribution and the effect of the category-based prioritization strategy; it did not retrospectively redefine the original set of 20 studies used for the qualitative synthesis.

The complete PubMed and Web of Science search strings, export details, and sensitivity-analysis worksheet are available in the same supplementary repository.

III. Results

A. Identification results

The bibliographic database search identified **525** records, with 255 retrieved from IEEE Xplore and 270 from Scopus. After removing 33 duplicates, **492** unique records remained for title and abstract screening. Retrospective citation searching identified an additional **14** records, while the generative AI assisted search identified **19** records (15 from Consensus and 4 from Elicit), resulting in a total of **558** identified records.

Records identified through Approach 3 were screened under the same eligibility criteria applied to the other routes. The retrieval strategy and the ranking criteria of these platforms, however, could not be fully replicated, which limits the reproducibility of this identification route.

The supplementary searches retrieved 1,060 records from the Web of Science Core Collection and 580 from PubMed. After restricting the results to the 2021–2026 publication window, removing records already present in the original corpus or duplicated across the supplementary databases, and excluding records without abstracts and secondary publications, 810 records remained available for analysis. A predefined rule-based metadata filter, implemented using search strings and Boolean conditions applied to titles and abstracts, prioritized 287 records as potentially consistent with the review scope. The reporting completeness framework was then applied to the supplementary records, and 98 records reached or exceeded the cutoff defined for their respective acquisition categories. These automated procedures were used only to prioritize records for the sensitivity analysis and did not constitute a formal eligibility decision. Consequently, the supplementary analysis did not redefine or modify the 20 studies selected for the original qualitative synthesis.

B. Distribution of modalities in the corpus

Table I summarizes the frequency of the modality families identified in the 492 consolidated records retrieved through Approach 1. The results reveal a marked imbalance across modalities. Magnetic resonance imaging was identified in 287 records, representing approximately 58% of the corpus, whereas continuous sensing in real world environments, including wearable devices, environmental sensors, and digital biomarkers, was identified in 42 records, representing just over 8%. Behavioral modalities were even less frequent, with face and video appearing in 20 records and eye tracking in only 6.

TABLE I
FREQUENCY OF MODALITY FAMILIES IN THE 492 CONSOLIDATED RECORDS RETRIEVED THROUGH APPROACH 1

Modality	n	**Modality**	n
MRI	287	Sensors / wearables	42
Clinical data	113	Text / language	22
PET	108	CSF biomarkers	20
Speech / audio	75	Face / video	20
Neuropsychological tests	64	Handwriting / drawing	16
EEG / MEG	46	Eye tracking	6
Genetics / omics	44		

A record may mention more than one family, so the counts do not sum to 492.

C. Supplementary coverage analysis

The supplementary Web of Science search reproduced the overall modality distribution observed in the original bibliographic corpus. The rank order of the 13 modality families reported in Table I showed a strong positive correlation with their frequencies in Web of Science (Spearman's $\rho = 0.902$, $p < 0.001$). Neuroimaging modalities remained predominant, whereas sensor-based, behavioral, and eye-tracking modalities remained comparatively less represented. This agreement does not establish exhaustive coverage of the literature, but indicates that the marked imbalance across modality families was robust to the inclusion of an additional bibliographic database.

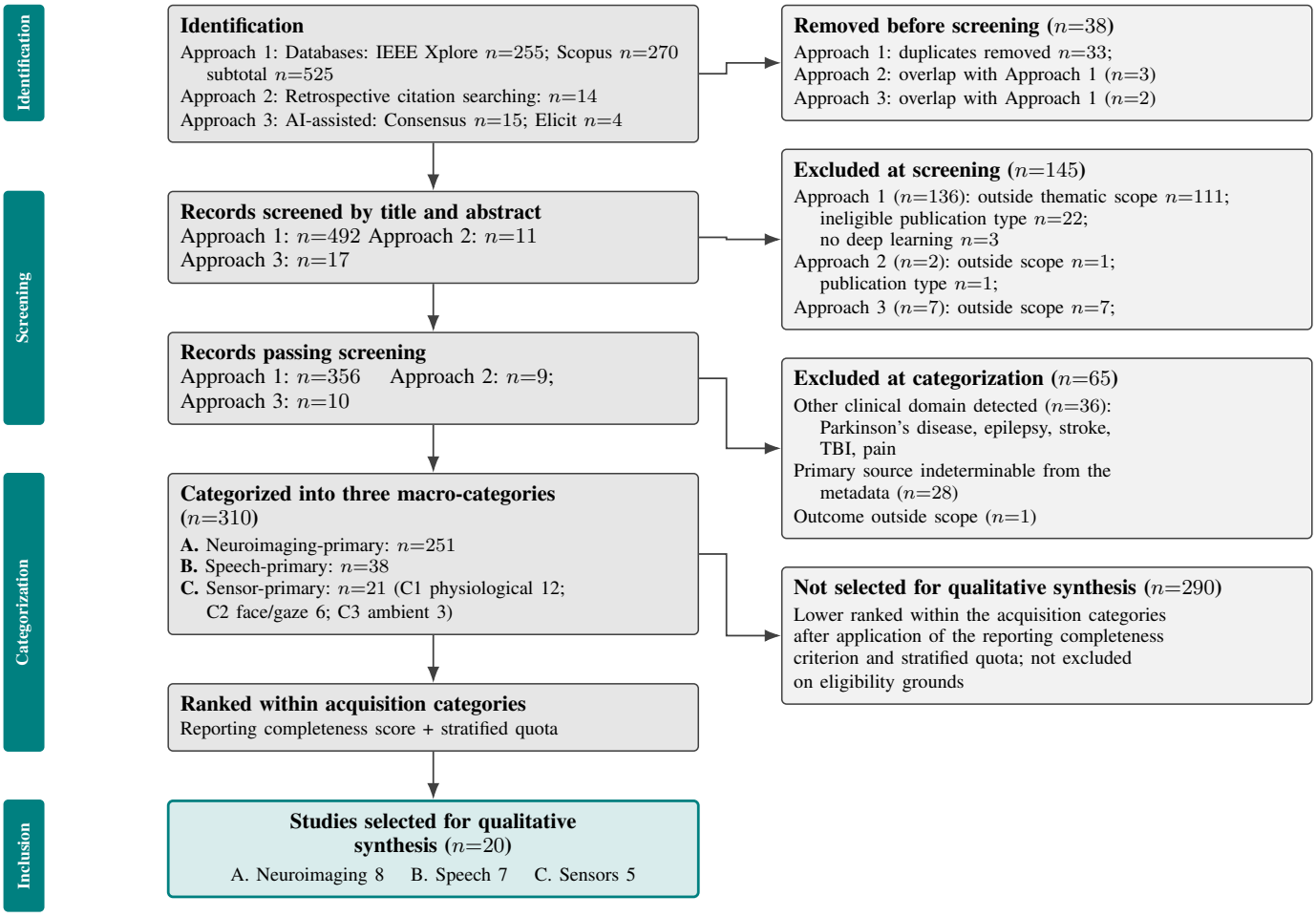


Fig. 1. PRISMA-ScR flow diagram with three original identification pathways. Approach 1 corresponds to the Boolean search of bibliographic databases; Approach 2, to retrospective citation searching; and Approach 3, to AI-assisted search tools. Post hoc PubMed and Web of Science searches were conducted separately for coverage and sensitivity analyses and are therefore not included in the original identification flow.

D. Sensitivity to category-based prioritization

To examine how study composition might differ without category-specific quotas, the supplementary records were ranked globally using the same reporting completeness score. Seventeen records scored above the global selection boundary: 14 neuroimaging-primary and 3 sensor-primary. Seven additional records were tied at the boundary score of 17: 5 neuroimaging-primary and 2 speech-primary. Thus, a 20-study selection based only on reporting completeness would include 15–17 neuroimaging, 0–2 speech, and 3 sensor studies, depending on tie resolution.

This result indicates that the stratified quota substantially changes the composition of the qualitative synthesis. Its purpose was therefore not to reproduce the prevalence of acquisition categories in the literature, but to preserve sufficient representation of speech and sensor studies for qualitative examination. Conclusions regarding the relative frequency of acquisition modalities were consequently based on the broader bibliographic corpus rather than on the quota-selected set of 20 studies.

E. Studies selected for the qualitative synthesis

Table II presents the 20 studies selected for the qualitative synthesis, organized according to their primary data source, together with the fusion configuration and the modalities effectively integrated, as determined through full text extraction.

IV. QUALITATIVE SYNTHESIS

Results across neuroimaging, speech and language, and sensor-based studies should not be interpreted as directly comparable because datasets, diagnostic tasks, sample sizes, performance metrics, and validation strategies differ substantially across studies. Accordingly, fusion benefit was interpreted primarily through within-study comparisons with unimodal baselines when available, rather than through absolute performance comparisons across acquisition categories.

A. Neuroimaging

Neuroimaging was the predominant acquisition category in the broader categorized corpus. The reviewed studies show a clear evolution of fusion strategies, ranging from integration at the image level to the fusion of learned representations and,

TABLE II
STUDIES SELECTED FOR THE QUALITATIVE SYNTHESIS ($n = 20$), ORGANIZED BY PRIMARY DATA SOURCE

#	Study	Year	Config.	Modalities
A. Neuroimaging-primary				
1	Odusami et al. [6]	2023	Homogeneous	MRI + PET (pixel-level)
2	Leng et al. [9]	2023	Homogeneous	sMRI + FDG PET
3	Tang et al. [7]	2024	Homogeneous	sMRI + FDG PET
4	Khan et al. [8]	2024	Homogeneous	MRI + PET
5	Gravina et al. [10]	2024	Homogeneous	MRI + amyloid PET
6	Chen et al. [11]	2025	Heterogeneous	MRI + clinical records
7	Vega-Guzman et al. [12]	2026	Homogeneous	T1 MRI + FDG PET
8	Abarnaswara et al. [13]	2026	Heterogeneous	T1 MRI + FDG PET + clinical (MMSE, MoCA, CDR)
B. Speech-primary				
9	Rohanian et al. [16]	2021	Homogeneous	Acoustic + ASR transcript (disfluencies, pauses)
10	Ilias and Askounis [14]	2022	Homogeneous	Audio (Log Mel) + transcript
11	Sahu et al. [19]	2022	Heterogeneous	Audio + transcript + cognitive scores
12	Lin and Washington [15]	2024	Homogeneous	Audio + transcript + timestamps
13	Luz et al. [17][a]	2024	Homogeneous	Acoustic + linguistic features
14	Gao et al. [18]	2025	Homogeneous	Interpretable linguistic features + pretrained audio embeddings
15	Kumari et al. [20]	2025	Homogeneous	Audio (MFCC, VGGish) + transcript (BERT)
C. Sensor-primary				
16	Ouyang et al. [25]	2024	Heterogeneous	Ambient multimodal sensing (federated)
17	Wu et al. [21]	2025	Heterogeneous	EEG + clinical
18	Seo et al. [24]	2025	Homogeneous	Inertial sensors (phone + watch)
19	Anand and Kumar [22]	2026	Heterogeneous	EEG + structural MRI
20	Syed [23]	2026	Homogeneous	Ambient sensors (gait + sleep)

[a]Treated as a primary study on the grounds that the paper introduces the ADReSS-M dataset and reports an original fusion baseline, rather than synthesizing prior findings. Modalities determined through full text extraction and may differ from those suggested by the abstracts.

more recently, to the integration of information beyond the image itself.

From the pixel-level to the feature *level:* Odusami et al. [6] investigate pixel-level fusion of MRI and PET images using wavelet-based image integration followed by Vision Transformer classification. Evaluated on ADNI, the approach illustrates an early fusion strategy in which multimodal information is combined before feature learning, contrasting with the feature-level methods developed in subsequent studies.

Moving fusion to the *feature level*, Tang et al. [7] extract deep representations from *sMRI* and *PET* images using a 3D CNN and integrate them through an enhanced *Transformer* capable of learning global correlations across the different representations. The architecture performs *cross modal* fusion and incorporates visualization mechanisms to improve interpretability. Evaluated on the ADNI dataset using paired Fluorodeoxyglucose Positron Emission Tomography (FDG PET) and *sMRI* images from the same participant and ten fold cross validation, the model achieved an accuracy of 98.1%, outperforming the unimodal versions by 6.19 percentage points for sMRI and 10.96 percentage points for PET. The left parahippocampal region consistently emerged as one of the most relevant regions for classification.

Khan et al. [8] extend feature-level fusion to early multiclass diagnosis with the *Dual-3DM^{3}AD* model, combining Transformer-based semantic segmentation, multiscale feature extraction, dense feature aggregation, and attention mechanisms over MRI and PET data. The study illustrates the use of increasingly structured representation learning for multimodal multiclass diagnosis, a setting that is inherently more challenging than binary classification.

Leng et al. [9] propose a cross-enhanced feature-level fusion strategy for sMRI and FDG PET, explicitly modeling correlations and complementarity between the two modalities through attention-based interactions. Applied to the joint diagnosis of Alzheimer's disease and subjective memory complaints on ADNI, the study illustrates the transition from simple feature combination toward explicit cross-modal interaction.

A different practical challenge is the presence of incomplete data. Gravina et al. [10] propose a *Multi Input Multi Output* (MI MO) architecture based on a 3D CNN that adapts the training process to the available modalities, allowing inference even when *MRI* or amyloid *PET* images are unavailable. Evaluated on the OASIS 3 dataset for dementia severity estimation, the method outperformed both unimodal models and other fusion approaches. Its main contribution is the ability to handle missing data, a common situation in clinical practice that has received limited attention in previous studies.

Beyond imaging: incorporating clinical variables: Chen et al. [11] extend multimodal fusion beyond imaging by integrating MRI with clinical records in the MMDF model. The architecture combines image and clinical representations through cross-modal feature learning and adaptive feature recalibration, achieving 97.35% accuracy on ADNI.

Vega-Guzman et al. [12] quantify the benefit of fusion using one of the few experimental designs in the corpus that explicitly reported participant-level data splitting. Combining structural magnetic resonance imaging with positron emission tomography produced an improvement of 9.7 percentage points over the best-performing unimodal approach, achieving an accuracy of 0.90 with a standard deviation of 0.03 under ten-fold cross-validation. The data split was explicitly performed at the participant level, a precaution that was rarely reported in the remaining studies and is essential to prevent data leakage when multiple slices or examinations belong to the same individual.

Abarnaswara et al. [13] integrate MRI, FDG PET, and clinical-cognitive variables, achieving 0.921 accuracy, a 6.7 percentage-point gain over the best unimodal model and a 3.4-point gain over image-only fusion. The study also reports Cohen's kappa of 0.902 and quadratic weighted kappa of 0.935. However, the data were curated from unnamed repositories, limiting reproducibility, and the model includes MMSE, MoCA, and CDR scores that also contribute to the diagnostic reference standard, raising a potential circularity concern; Vega-Guzman et al. [12] explicitly excluded such cognitive instruments from model inputs.

B. Speech and language

The speech category is strongly concentrated on the ADReSS corpus and its successors, ADReSSo and ADReSS-M, which share provenance in the Cookie Theft sessions of the Pitt Corpus distributed through DementiaBank; ADReSS-M additionally evaluates cross-lingual generalization using a Greek test set. Although participant sets differ across editions, this shared provenance provides greater internal comparability than in the other acquisition categories while potentially limiting generalizability because of the repeated use of closely related source corpora.

From attention based fusion to pretrained representations: Ilias and Askounis [14] combine speech spectrograms processed by a Vision Transformer with BERT-based transcript representations using attention and gating mechanisms. On ADReSS, their best configuration achieved 90.00% accuracy, an F1 score of 89.94%, and an RMSE of 3.61 for MMSE regression. The study illustrates early end-to-end multimodal fusion of acoustic and linguistic representations, while also highlighting the limitations imposed by small datasets.

Lin and Washington [15] compare textual and acoustic contributions on the Pitt Cookie Theft corpus using text- and audio-based representations under five-fold cross-validation. Textual models dominated performance, particularly after data augmentation, while adding Wav2vec-based audio features produced only marginal improvements. This finding shows that adding a modality does not necessarily provide substantial benefit when the added representation contributes limited complementary information.

Rohanian et al. [16] investigate adaptive gating for combining acoustic features with noisy ASR-derived linguistic information, including disfluencies and pauses. On ADReSSo, the model achieved 84% accuracy and an RMSE of 4.26 for MMSE regression. The results suggest that gating can regulate modality contributions under recognition noise, supporting adaptive rather than fixed fusion strategies.

Luz et al. [17] address cross-lingual generalization in ADReSS-M, evaluating whether models trained in English generalize to Greek. Their multimodal baseline combining acoustic and linguistic information achieved 73.91% accuracy and an RMSE of 4.995 for MMSE regression, with hard fusion outperforming the isolated feature sets. The study extends multimodal speech analysis beyond monolingual evaluation settings.

Gao et al. [18] combine interpretable linguistic features with pretrained audio embeddings for three-class classification of healthy controls, MCI, and dementia, as well as MMSE regression. In the PROCESS Grand Challenge, the approach achieved an F1 score of 0.649 and an RMSE of 2.628, obtaining the highest overall ranking in the competition.

Divergent fusion mechanisms over related corpora: Sahu et al. [19] use late fusion to combine acoustic, linguistic, disfluency, and cognitive information, whereas Kumari et al. [20] apply feature-level fusion of BERT-based text representations and acoustic features. These contrasting integration strategies over closely related speech corpora reinforce that performance differences cannot be attributed to the fusion mechanism alone.

C. Sensors, continuous monitoring, and digital biomarkers

Sensor studies ranged from clinically supervised EEG experiments [21], [22] to wearable and ambient sensing in everyday environments [23]–[25], making this the most heterogeneous acquisition category in the review. Reported sample sizes ranged from 56 to 125 participants, none of the studies included external validation, and only the free-living studies by Seo et al. [24] and Ouyang et al. [25] used prospectively collected data.

Home based acquisition under federated learning: Ouyang et al. [25] propose *ADMarker*, a federated multimodal sensing system for home-based monitoring of Alzheimer's disease that preserves privacy by keeping raw data local. In a four-week study with 91 older adults, the system identified digital biomarkers with up to 93.8% accuracy and detected early-stage Alzheimer's disease with an average accuracy of 88.9%, illustrating the feasibility of multimodal analysis in real-world environments.

Two additional studies illustrate the breadth of the category without introducing new fusion mechanisms. Wu et al. [21] combine electroencephalography with clinical variables into a continuous health score for mild cognitive impairment, whereas Syed [23] applies a *Transformer* to smart home sensors using gait and sleep patterns as behavioral markers.

Protocol realism and evaluation design: Sensor studies differed substantially in acquisition and evaluation design. Seo et al. [24] used 30 days of passive free-living data with a subject-independent split, whereas Anand and Kumar [22] used controlled acquisition with window-level analysis and did not report participant-based grouping. These differences limit direct comparison of performance estimates across studies.

Fusion gain and redundancy between sources: Only two sensor studies reported unimodal baselines allowing fusion gain to be quantified. Anand and Kumar [22] reported a 3.77 percentage-point decrease after combining EEG and MRI, using representations randomly paired across databases without subject-level correspondence. Seo et al. [24] reported an AUC of 0.700 for the best multimodal configuration versus 0.699 for the wrist gyroscope alone. Because the evaluated inertial streams measured the same physical phenomenon at different body locations, the negligible gain is consistent with the possibility of substantial source redundancy.

V. Discussion and Conclusion

This scoping review mapped multimodal deep learning approaches across neuroimaging, speech, and sensor-based settings. Neuroimaging strongly dominated the broader corpus, while the stratified quota intentionally increased the representation of speech and sensor studies in the qualitative synthesis. Accordingly, modality-frequency conclusions were based on the broader corpus, whereas category-specific observations reflect purposive qualitative sampling.

Fusion benefit varies across acquisition contexts

Neuroimaging studies show a progressive evolution from pixel-level fusion [6] to attention-based *cross-modal* fusion [7]–[9], incomplete-data strategies [10], and integration of non-imaging variables [11], [13]. Of the eight neuroimaging studies, only four reported a unimodal baseline. Fusion improved performance in all four. In the three studies reporting directly quantifiable gains over the best-performing unimodal baseline, improvements ranged from 6.2 to 9.7 percentage points [7], [12], [13], while Gravina et al. [10] reported a qualitative improvement over the unimodal models.

Among the sensor studies, only two provided unimodal baselines that allowed the contribution of fusion to be directly assessed. In these studies, fusion produced either a negative or negligible within-study gain: Anand and Kumar [22] reported lower performance after combining EEG and MRI representations, whereas Seo et al. [24] reported an AUC of 0.700 for the best multimodal configuration compared with 0.699 for the best unimodal configuration. Because the studies differ substantially in acquisition protocols, evaluation metrics, and validation procedures, their absolute performance values should not be compared directly. These results therefore provide limited evidence regarding the benefit of fusion in sensor-based settings rather than evidence of systematically lower performance for this acquisition category.

Speech studies point in a similar direction. Despite being the most homogeneous category regarding data provenance, they are highly heterogeneous regarding fusion mechanisms, preventing direct attribution of performance differences to the integration strategy. Moreover, adding audio to textual representations produced only marginal improvements [15].

Returning to the meta-analysis by Guo et al. [1], the pooled AUC of 0.924 (95% CI: 0.912–0.936) is broadly consistent with the strong performance reported in the neuroimaging literature mapped in this review. However, this estimate should be interpreted within the acquisition and evaluation context from which it was derived rather than as a general estimate of the benefit of multimodal fusion across all modality families. The modality distribution reported in Table I further illustrates the predominance of neuroimaging in the broader mapped corpus. Evidence from speech and sensor-based studies remains more heterogeneous in terms of datasets, outcome definitions, metrics, and validation strategies, preventing direct quantitative comparison with pooled neuroimaging estimates.

Source complementarity as a potential factor in fusion benefit

The reviewed studies suggest that the complementarity of the combined sources may influence the benefit obtained from multimodal fusion. For example, the sensor configurations examined by Seo et al. [24] combined signals related to the same underlying physical phenomenon and showed only a negligible gain over the best unimodal configuration, whereas several neuroimaging studies combined structural and metabolic information and reported improvements over unimodal baselines. However, source complementarity or physical independence was not directly manipulated or experimentally compared across the reviewed studies. Differences in datasets, fusion architectures, validation strategies, and acquisition protocols also prevent attribution of the observed fusion gains to this factor alone. Therefore, the pattern identified in this review should be interpreted as a hypothesis-generating observation rather than evidence that source independence is more important than the number of modalities.

Methodological limitations across categories

Three limitations recur throughout the literature. External validation is almost absent, as studies generally evaluate models within partitions of the same cohort. Agreement metrics such as Cohen's kappa are rarely reported, with only one study providing them [13]. Additionally, unimodal baselines are frequently missing, limiting the verification of claimed fusion benefits.

A further concern is the circularity introduced when cognitive screening scores are used as predictors while also contributing to the diagnostic reference standard. Studies addressing this issue reached opposite decisions [12], [13], but the implications remain insufficiently discussed. The field also lacks shared reporting standards, particularly for sensor studies, where metrics, units of analysis, and label definitions vary substantially.

Search strategy as a methodological finding

The identification routes differed in yield and reproducibility. Retrospective citation searching yielded 9 additional records passing screening from 14 tracked references, compared with 10 from 19 AI-assisted records. Although generative AI tools contributed eligible studies, their retrieval and ranking procedures could not be fully reproduced and should therefore be considered complementary to reproducible database searches. The post hoc PubMed and Web of Science analyses further supported the robustness of the observed modality-distribution pattern without establishing exhaustive literature coverage.

Limitations of this review

The categorization of records by primary data source relied on titles, abstracts, and keywords, which does not allow reliable determination of the modalities actually employed. The prioritization of studies for synthesis was based on reporting completeness rather than methodological quality, potentially penalizing studies with limited abstract descriptions. The stratified quota also intentionally altered the category composition of the 20-study qualitative synthesis; therefore, category-specific observations should not be interpreted as proportionally representative of the broader literature. Although the original identification strategy was based on Scopus and IEEE Xplore, supplementary searches in PubMed and Web of Science were subsequently conducted as a post hoc coverage and sensitivity analysis. These additional searches reduced concerns regarding database-specific coverage, but they do not establish exhaustive retrieval of all potentially relevant studies.

Directions for future work

Future priorities differ across categories. Speech studies require stronger unimodal baselines and standardized evaluation protocols before new fusion mechanisms can be meaningfully compared. Sensor research requires public benchmark datasets comparable to those available for neuroimaging, as the only prospective study using realistic protocols and heterogeneous instrumentation [25] is not publicly reproducible. Neuroimaging, in turn, requires external validation across independent cohorts and routine reporting of agreement metrics.

Across all categories, future studies should characterize the complementarity and redundancy of combined sources, report strong unimodal baselines and validation procedures, and clearly distinguish whether cognitive instruments are used as predictors, targets, or components of the diagnostic reference standard. Overall, adding modalities alone does not guarantee improved performance.

ACKNOWLEDGMENT

This article, developed as part of Dementia project, was made possible by Samsung Electronica da Amazonia LTDA, with support from SUFRAMA and implementation by the Federal University of Amazonas, a beneficiary of funds originating from the Federal Law No. (8,387/1991), with its dissemination and publicity in accordance with the provisions of Article 39 of Decree No. 10,521/2020.

In accordance with the IEEE policy on the use of AI-generated content, the authors disclose that ChatGPT (OpenAI) was used as a support tool for translation into English, textual organization, and grammar revision across all sections of this manuscript. The tool was not used to generate scientific content, experimental results, figures, or code. All intellectual decisions and full responsibility for the final content remain solely with the authors.

REFERENCES

[1] H. Guo, Z. Yang, G. Zhang, L. Lv, and X. Zhao, "Meta analysis of the diagnostic efficacy of transformer-based multimodal fusion deep learning models in early Alzheimer's disease," *Frontiers in Neurology*, vol. 16, Art. no. 1641548, 2025.

[2] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal Learning with Transformers: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113–12132, 2023.

[3] A. C. Tricco *et al.*, "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[4] M. A. de Souza Pinto, "Supplementary material: for a scoping review on multimodal deep learning," Zenodo, 2026, doi: 10.5281/zenodo.22694084.

[5] M. D. J. Peters, C. Godfrey, P. McInerney, Z. Munn, A. C. Tricco, and H. Khalil, "Scoping Reviews," in *JBI Manual for Evidence Synthesis*, E. Aromataris, C. Lockwood, K. Porritt, B. Pilla, and Z. Jordan, Eds. JBI, 2024. doi: 10.46658/JBIMES-24-09.

[6] M. Odusami, R. Maskeliūnas, and R. Damaševičius, "Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease," *Electronics*, vol. 12, no. 5, Art. no. 1218, 2023.

[7] Y. Tang, X. Xiong, G. Tong, Y. Yang, and H. Zhang, "Multimodal diagnosis model of Alzheimer's disease based on improved Transformer," *BioMedical Engineering OnLine*, vol. 23, Art. no. 8, 2024.

[8] A. A. Khan, R. K. Mahendran, K. Perumal, and M. Faheem, "Dual-3DM^{3}AD: Mixed Transformer Based Semantic Segmentation and Triplet Pre-Processing for Early Multi-Class Alzheimer's Diagnosis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 696–707, 2024.

[9] Y. Leng *et al.*, "Multimodal cross enhanced fusion network for diagnosis of Alzheimer's disease and subjective memory complaints," *Computers in Biology and Medicine*, vol. 157, Art. no. 106788, 2023.

[10] M. Gravina, A. García-Pedrero, C. Gonzalo-Martín, C. Sansone, and P. Soda, "Multi input-Multi output 3D CNN for dementia severity assessment with incomplete multimodal data," *Artificial Intelligence in Medicine*, vol. 149, Art. no. 102774, 2024.

[11] J. Chen, Y. Wang, A. Zeb, M. D. Suzauddola, and Y. Wen, "Multimodal mixing convolutional neural network and transformer for Alzheimer's disease recognition," *Expert Systems with Applications*, vol. 259, Art. no. 125321, 2025.

[12] S. Vega-Guzman, G. Ramirez-Nava, I. Chairez, and M. Alfaro-Ponce, "Optimizing Alzheimer's diagnosis: Multimodal FDG-PET and MRI feature fusion with explainable deep learning," *Biomedical Signal Processing and Control*, vol. 117, Art. no. 109659, 2026.

[13] R. Abarnaswara, S. Aarthi, V. Aiswarya, S. Dheetchitha, and R. Dhivyashree, "Explainable Multimodal Deep Learning for Early Neurodegenerative Dementia Diagnosis," in *Proc. 9th Int. Conf. Inventive Computation Technologies (ICICT)*, 2026, pp. 1156–1163.

[14] L. Ilias and D. Askounis, "Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts," *Frontiers in Aging Neuroscience*, vol. 14, Art. no. 830943, 2022.

[15] K. Lin and P. Y. Washington, "Multimodal deep learning for dementia classification using text and audio," *Scientific Reports*, vol. 14, Art. no. 13887, 2024.

[16] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs," in *Proc. Interspeech*, 2021, pp. 3820–3824.

[17] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 738–749, 2024.

[18] Y. Gao, L. Guo, and H. Liu, "Leveraging Multimodal Methods and Spontaneous Speech for Alzheimer's Disease Identification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–2, doi: 10.1109/ICASSP49660.2025.10887657.

[19] H. K. Sahu, S. Kumar, S. H. Alsamhi, M. K. Chaube, and E. Curry, "Novel Framework for Alzheimer Early Diagnosis Using Inductive Transfer Learning Techniques," in *Proc. 2nd Int. Conf. Emerging Smart Technologies and Applications (eSmarTA)*, 2022.

[20] M. Kumari, R. Lal, and A. Mehta, "Multimodal NLP-Audio Approach for Early Alzheimer's Detection Using BERT and Acoustic Embeddings," in *Proc. Int. Conf. Computing and Communications (COMPUTINGCON)*, Pune, India, 2025.

[21] X. Wu, Q. Huang, Y. Liu, X. Dong, H. Liu, and D. Wen, "Multimodal Deep Learning-Based EEG Health Scoring Model for Mild Cognitive Impairment," *IFAC-PapersOnLine*, vol. 59, no. 35, pp. 697–702, 2025.

[22] K. Anand and N. Kumar, "Feature-Level Fusion of EEG and MRI for Robust Multimodal Alzheimer's Disease Classification," in *Proc. 4th Int. Conf. Secure Cyber Computing and Communication (ICSCCC)*, 2026, pp. 431–436.

[23] A. A. Syed, "Transformer-Based Early Detection of Mild Cognitive Impairment Using Smart Home Behavioral Sensor Data," in *Proc. Int. Conf. Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)*, 2026.

[24] K. Seo, J. Hwang, K. Kim, H. Argrawal, J. H. Yun, and H. Y. Kim, "End-to-End Classification of Cognitive Impairment Using Daily-Life Gait Data," in *Proc. 47th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025.

[25] X. Ouyang *et al.*, "ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease," in *Proc. 30th Annual Int. Conf. Mobile Computing and Networking (ACM MobiCom)*, 2024, pp. 404–419.

Neurological Prognostication Post-Cardiac Arrest Using Continuous Wavelet Transform and Hybrid Deep Learning

1st Ana Clara Bezerra do Nascimento
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
clara.bezerran@ufpe.br

2nd Anna Beatriz Brito de Souza Almeida
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
anna.britoalmeida@ufpe.br

3rd Maria Luísa B. de L. Barros
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
luisa.brandao@ufpe.br

4th Raylla S. das Montanhas Farias
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
raylla.farias@ufpe.br

5th Wellington P. dos Santos
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

***Abstract*—Neurological prognostication in comatose survivors of cardiac arrest remains a critical challenge in intensive care, where clinical uncertainty often hinders therapeutic decision-making. Although continuous electroencephalography (EEG) is the gold-standard monitoring modality, its visual interpretation is subjective, labor-intensive, and dependent on specialized expertise. This work presents a pilot, hypothesis-generating study of an automated prognostic support pipeline built upon the multicenter I-CARE database, integrating advanced signal processing with hybrid machine learning. The pipeline comprises (i) FIR filtering (0.5–45 Hz), (ii) 1-second temporal segmentation with 25% overlap to preserve slow-wave Delta activity, (iii) time–frequency representation via the Continuous Wavelet Transform (CWT) with a Complex Morlet mother wavelet, and (iv) a hybrid classifier combining ResNet50-based feature extraction with a Random Forest. To prevent data leakage, the dataset was split at the patient level (12 patients for training, 4 for testing). On the independent test set, the model achieved an accuracy of 95.27% (95% Wilson CI: 91.7–97.3%), a Cohen's Kappa of 0.9054, and an Area Under the ROC Curve (AUC) of 0.991. Per-class analysis revealed balanced sensitivity (94.8% for good outcome, 95.7% for poor outcome), with no evidence of class bias. Given that the test partition comprises only four patients, these estimates should be interpreted as a preliminary signal rather than a definitive performance ceiling; nevertheless, the results suggest that latent spectral texture patterns, rather than overt morphological graphoelements, may carry prognostic information, motivating validation on larger, independent multicenter cohorts before any claim of clinical readiness.**

***Index Terms*—Electroencephalography, cardiac arrest, continuous wavelet transform, neurological prognosis, hybrid machine learning.**

I. Introduction

Cardiac arrest (CA) is a high-severity medical emergency that frequently leads to Hypoxic-Ischemic Encephalopathy (HIE) in survivors. The extent of neuronal damage resulting from oxygen deprivation is the primary determinant of functional recovery or progression to permanent vegetative states, making early prognostic evaluation essential for guiding ethical and therapeutic decisions [1], [2]. Current clinical guidelines emphasize the integration of multimodal prognostic markers, including neurological examination, somatosensory evoked potentials, neuroimaging, and continuous electroencephalography (cEEG) [3], [4].

Among these modalities, cEEG is considered the most sensitive for monitoring cortical function in real time. Specific patterns such as background reactivity, burst-suppression, and status epilepticus carry high predictive value [5], [6]. However, continuous interpretation of these records in Intensive Care Units (ICUs) faces significant logistical barriers: visual analysis is intrinsically subjective, susceptible to inter-observer variability (with kappa values as low as 0.40–0.60 in some studies [7]), and impractical for the large data volumes generated 24 hours a day [8].

The need to overcome these limitations has driven the development of quantitative EEG (qEEG). Digital Signal Processing (DSP) techniques allow the decomposition of cerebral signal complexity into objective markers [9]. Traditional Fourier-based approaches (FFT), however, assume signal stationarity and fail to capture transient events that are critical in post-arrest encephalopathy [10]. Wavelet-based time–frequency analysis overcomes this limitation by providing multi-resolution decomposition, simultaneously capturing temporal and spectral dynamics [11], [12].

Recent advances in deep learning have further expanded the diagnostic potential of EEG. Convolutional Neural Networks (CNNs) can automatically extract high-level features from time–frequency representations, eliminating the need for manual feature engineering [13], [14]. However, the application of CNNs to small clinical datasets (typical in neurophysiol-

ogy) poses significant challenges, including overfitting and poor generalization [15]. Transfer learning from pre-trained architectures (e.g., ResNet, EfficientNet) offers a promising solution, leveraging representations learned from large-scale datasets (e.g., ImageNet) and adapting them to the specific domain of biomedical signals [16], [17].

Despite these advances, several gaps remain in the literature: (i) most studies rely on the Discrete Wavelet Transform (DWT), which introduces geometric artifacts that impair feature learning; (ii) the impact of temporal resolution (window size) on prognostic accuracy is underexplored; and (iii) the interpretation of deep learning models in clinical contexts remains opaque, hindering adoption.

This work presents a pilot, hypothesis-generating study of an automated prognostic classification pipeline that addresses these gaps by combining the Continuous Wavelet Transform (CWT) for dense spectral texture extraction with a hybrid machine learning architecture. Given the small cohort size (16 patients), the results are framed as preliminary evidence to be confirmed on larger, independent cohorts, rather than as a claim of clinical-grade performance. The main contributions are:

1) A hybrid pipeline coupling CWT-based scalograms, ResNet50 feature extraction, and Random Forest classification, designed for small-sample clinical datasets and reported with explicit uncertainty quantification;
2) Empirical, exploratory evidence that 1-second windows, rather than the conventional 500 ms, are associated with improved capture of slow Delta-band biomarkers, with a 7.3 percentage-point accuracy gain observed in this cohort;
3) Preliminary evidence that prognostic information may reside in latent spectral textures rather than in visually overt graphoelements, a hypothesis that warrants confirmation with interpretability analysis (e.g., Grad-CAM, SHAP);
4) A patient-level splitting protocol to prevent data leakage, a common methodological flaw in EEG-based prognostic studies, together with an explicit discussion of the residual statistical limitations that patient-level splitting imposes on small cohorts.

II. Background

A. EEG and Post-Cardiac Arrest Prognosis

Neurological prognostication after CA is determinant for clinical management. The EEG consolidates as an essential tool to identify functional patterns associated with recovery or irreversible ischemic injury, such as burst-suppression and status epilepticus [1], [5]. Nevertheless, traditional visual EEG interpretation faces severe limitations due to inter-observer subjectivity and the exhaustion inherent to manual processing of long recordings [7]. Quantitative methods based on DSP, particularly time–frequency wavelet analysis, have become indispensable for extracting latent objective markers, enabling automated decision-support systems that increase diagnostic precision and agility [9], [18].

B. From DWT to CWT

Biomedical signals such as EEG are non-stationary, with spectral content varying dynamically over time. The Continuous Wavelet Transform (CWT) decomposes a signal $x(t)$ by convolving it with a scaled and translated mother wavelet $\psi(t)$:

$$X_{CWT}(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt \qquad (1)$$

where a is the scale parameter (inversely proportional to frequency) and b is the translation parameter. The Complex Morlet wavelet, a sinusoid modulated by a Gaussian, is widely adopted in neurophysiology due to its morphological resemblance to cortical oscillations and its ability to extract both magnitude and phase [19].

During preliminary experiments, the Discrete Wavelet Transform (DWT) produced scalograms with geometric "pyramid" artifacts and artificially filled discontinuities, which impaired neural network feature learning. Migration to the CWT yielded dense, continuous power spectral representations, revealing microstructures essential for class discrimination (Fig. 1).

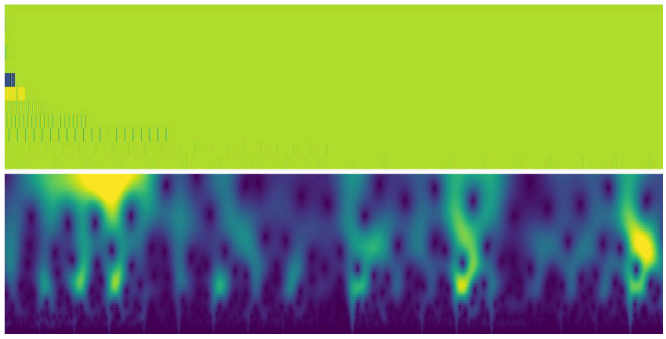


Fig. 1: Visual comparison of wavelet transforms. (A) Artifacts generated by the DWT, discarded from the study. (B) Scalogram generated by the CWT, evidencing the high spectral resolution used in training.

C. Machine Learning for Clinical Decision Support

The high dimensionality of biomedical data frequently exceeds the analytical capacity of traditional statistical methods. Ensemble learners such as Random Forest, proposed by Breiman, aggregate multiple decision trees built on random subsets of data, reducing variance and mitigating overfitting—a common issue when handling extensive feature vectors extracted from images [20]. In imbalanced clinical datasets, the Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic instances of the minority class through linear interpolation between k-nearest neighbors, enriching the decision space and improving classifier sensitivity [21].

III. Methodology

A. Proposed Method: Overview

Fig. 2 summarizes the end-to-end pipeline proposed in this work. Raw multi-channel EEG recordings are first standardized and filtered (Section III-C), then segmented into 1-second

windows with 25% overlap (Section III-D). Each window is converted into a spectral-texture image via the Continuous Wavelet Transform (Section III-E) and passed through a frozen ResNet50 to obtain a 2048-dimensional feature vector (Section III-F.1). SMOTE is applied exclusively to the training partition to balance classes (Section III-F.2), and a Random Forest classifier produces the final good/poor outcome prediction (Section III-F.3). The following subsections detail each stage.

B. Dataset and Patient Selection

Data were obtained from the I-CARE (International Cardiac Arrest REsearch consortium) database, available on PhysioNet, comprising clinical records and continuous EEG from comatose post-CA patients with return of spontaneous circulation (ROSC) [9]. Sixteen patients were selected from five distinct hospital centers, prioritizing diversity of clinical scenarios and temporal homogeneity. The inclusion criteria required recordings performed within the first 24 hours post-ROSC. Table I details the demographic and clinical characteristics of the selected cohort. The rightmost column reports the number of discrete EEG recording files available for each patient within the first 24 hours post-ROSC (not their duration); recording counts vary across patients because the I-CARE acquisition protocol stores continuous monitoring as sequential files whose number depends on hospital-specific acquisition settings and clinical events, rather than on a fixed per-patient schedule. As described below, for every patient the analysis is restricted to 2-minute segments drawn from the first 10 available recordings, so this heterogeneity in file count does not translate into heterogeneity in the amount of data analyzed per patient.

The cohort was intentionally balanced to ensure statistical robustness: 50% male and 50% female; 8 patients with good neurological outcome (Cerebral Performance Category, CPC 1–2) and 8 with poor outcome (CPC 3–5). To prevent data leakage, the dataset was split at the *patient level* rather than the window level: 12 patients (75%) for training and 4 patients (25%) for testing. The split preserved balanced gender distribution and balanced outcomes across both subsets. For each patient, 2-minute segments corresponding to the first 10 available EEG recordings were selected, ensuring temporal standardization within the first 24 hours post-ROSC; because every patient contributes the same number and duration of segments regardless of how many recording files were originally stored, the varying file counts in Table I do not bias the amount of data used per patient.

C. Signal Preprocessing

Given the heterogeneity of data from multiple hospital centers, a rigorous standardization protocol was established:

1) Spatial Standardization and Scale Normalization: Channels were mapped according to the International 10–20 System. To mitigate unit heterogeneity (μV vs. V) across recording devices, a Median Absolute Deviation (MAD) normalization was applied. The median, rather than the mean, was

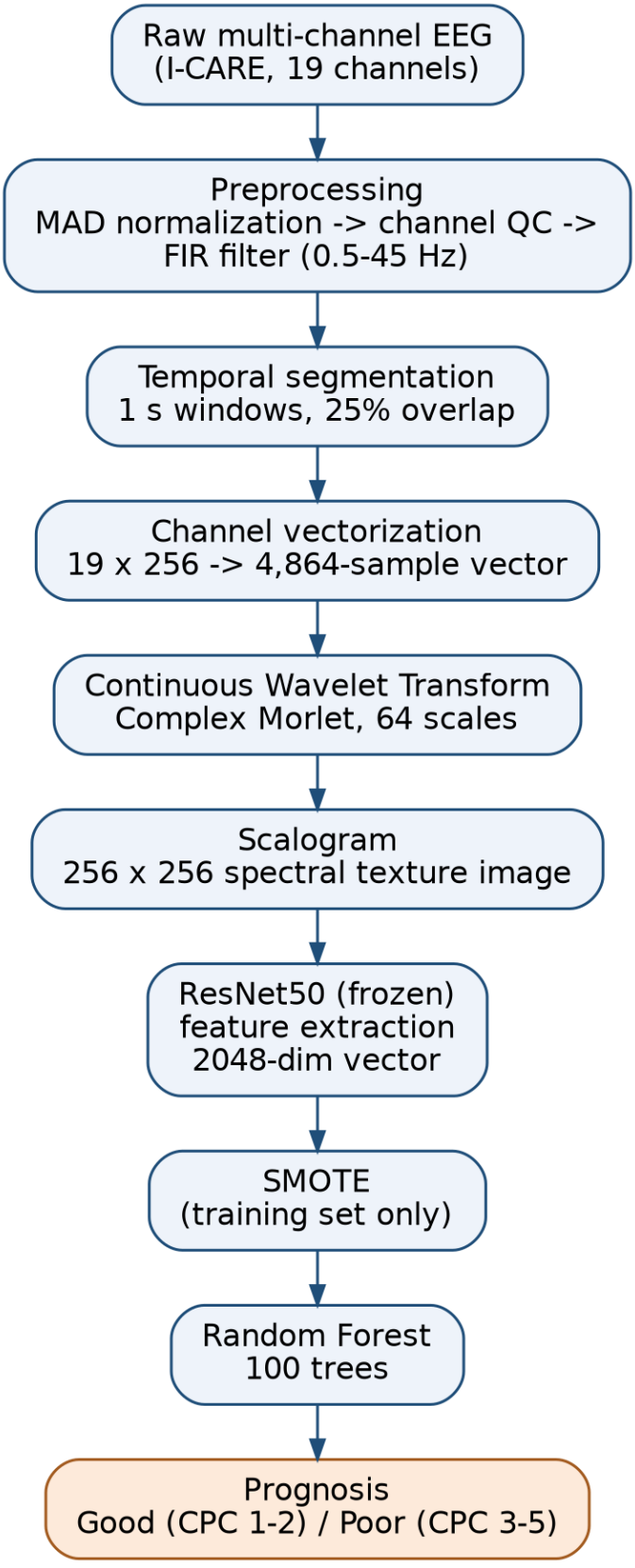


Fig. 2: Overview of the proposed prognostic pipeline, from raw EEG acquisition to the good/poor outcome prediction.

TABLE I: Demographic and Clinical Characteristics of the Selected Cohort (n = 16)

Patient ID	Sex	Outcome	Hospital	Number of EEG recording files (first 24 h)
563	Male	Good	e	16
470	Male	Good	b	19
948	Male	Good	d	20
422	Male	Good	f	39
958	Male	Poor	e	11
702	Male	Poor	b	12
541	Male	Poor	a	11
623	Male	Poor	f	23
913	Female	Good	e	42
487	Female	Good	b	22
543	Female	Good	d	40
843	Female	Good	a	23
588	Female	Poor	b	11
935	Female	Poor	d	20
944	Female	Poor	e	40
577	Female	Poor	f	12

Outcome: Good = CPC 1–2; Poor = CPC 3–5. Hospitals: a, b, d, e, f (multi-center).
The last column is the count of discrete EEG files stored for the patient in the first 24 h post-ROSC, not recording duration.

chosen for its statistical robustness against extreme outliers common in ICU recordings.

2) Channel Quality Validation: Channels with inadequate impedance or connection failure were automatically detected. Channels exhibiting zero standard deviation (flatlines) were discarded to prevent propagation of null coefficients into the wavelet stage.

3) Digital Filtering: A high-order linear-phase FIR filter with passband 0.5–45 Hz was applied. The 0.5 Hz lower cutoff removes baseline wander and low-frequency artifacts (sweat, respiration) while preserving Delta integrity; the 45 Hz upper cutoff attenuates power-line interference (50/60 Hz) and myogenic artifacts without compromising the lower Gamma band.

D. Temporal Segmentation Strategy

Continuous recordings were segmented into 1000 ms (1 s) epochs at $F_s = 256$ Hz, yielding 256 samples per window. The enlargement from the initially tested 500 ms to 1 s was decisive: a 500 ms window imposes a physical limit on the representation of complete slow-wave cycles (< 4 Hz), resulting in spectral information loss and temporal aliasing. The 1 s window ensures faithful representation of Delta oscillations—critical biomarkers in coma and HIE evaluation. A 25% overlap between consecutive segments was applied to maximize data utilization and mitigate boundary information loss, yielding approximately 1,590–1,600 segments per patient.

E. Time–Frequency Transformation

Each 1 s multi-channel window (19 channels $\times$ 256 samples) was transformed into a spectral texture image through three steps:

1) Channel Vectorization: The 19 temporal series were concatenated serially (channel flattening), producing a unified 4,864-sample vector. This choice preserves intra-channel temporal continuity while exposing inter-channel spatial variation in the linear time domain, and avoids the need for a montage-specific multi-channel image encoding, which would require a larger training set to be learned reliably. We acknowledge that concatenation introduces artificial discontinuities at channel boundaries, analogous in principle to the boundary artifacts discussed for the DWT in Section II-B; however, because the CWT is applied to the already-concatenated vector as a single continuous signal, these boundary transitions are smoothed by the wavelet's finite support rather than producing the sharp block artifacts characteristic of DWT decomposition. A systematic comparison against alternative multi-channel encodings (e.g., per-channel scalogram stacking) was not performed in this pilot study and is identified as a direction for future work.

2) CWT Application: The Complex Morlet wavelet (`cmor1.5-1.0`) was applied across 64 scales, covering the entire physiologically relevant frequency range (Delta to Gamma).

3) Scalogram Generation: The magnitude of the complex coefficient matrix was computed (phase discarded) and interpolated to a 256×256 pixel square image. Each EEG segment thus became a spectral texture image serving as high-dimensional input to the deep learning model.

F. Feature Extraction and Classification

The classification pipeline adopted a hybrid Deep Learning / classical machine learning architecture:

1) Feature Extraction with ResNet50: Scalograms were processed by a ResNet50 pre-trained on ImageNet, used in *feature extraction mode*. All convolutional layers were frozen, and the final fully-connected layer (1000 ImageNet classes) was removed. The output of the global average pooling layer—a 2048-dimensional feature vector—served as a compact, high-level representation of each scalogram's spectral texture. This approach was chosen over fine-tuning due to the limited

sample size, as unfreezing convolutional weights on small datasets would inevitably lead to overfitting.

2) Class Balancing (SMOTE): The training set (12 patients) exhibited class imbalance at the window level. SMOTE was applied *exclusively to the training set* to generate synthetic instances of the minority class, preventing classifier bias without introducing leakage into the test set.

3) Random Forest Classifier: The 2048-dimensional feature vectors were classified by a Random Forest configured with 100 decision trees. This ensemble method was selected for its robustness to high dimensionality and its inherent resistance to overfitting through bagging. Model validation followed a stratified 10-fold cross-validation protocol on the training set, with final performance reported on the held-out patient-level test set. This held-out evaluation corresponds to a single train/test split (one round): the 12/4-patient partition was fixed once, and the Random Forest was trained and evaluated on it a single time, rather than being repeated across multiple random splits or resampling runs. Consequently, the point estimates in Table II reflect one realization of the patient-level split; the Wilson confidence interval in Section IV-B quantifies window-level sampling uncertainty within this single split, but does not capture the additional variability that would arise from repeating the split with different patients. Repeating the evaluation across multiple leave-patients-out folds is identified as a priority for future work (Section IV-B and Section V).

IV. Results and Discussion

A. Classifier Performance

The model was evaluated on the independent test set of 4 patients, ensuring that validation data were entirely unseen during training. Table II summarizes the performance metrics.

TABLE II: Random Forest Performance Metrics on Training and Test Sets

Metric	Train	Test	Δ
Accuracy	0.9354	0.9527	+0.0173
Cohen's Kappa	0.8708	0.9054	+0.0346
ROC AUC	0.9850	0.9910	+0.0060

The proximity between training and test metrics indicates a stable operating point rather than gross overfitting, and is consistent with SMOTE mitigating class bias during training. The Cohen's Kappa of 0.9054 on the test set falls in the range conventionally described as almost perfect agreement between automatic prediction and clinical outcome. However, because the test set comprises only four patients, these point estimates carry substantial uncertainty, which we quantify explicitly in Section IV-B. Fig. 3 illustrates the comparison between training and test performance.

B. Statistical Robustness and Effective Sample Size

A window-level accuracy of 95.27% was obtained from 221 correct predictions out of 232 test-set windows (Table III).

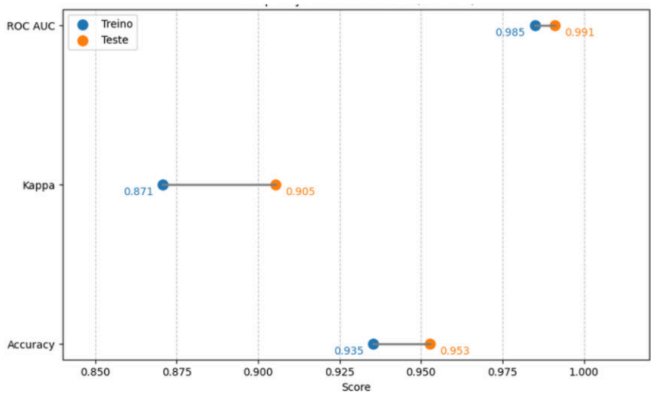


Fig. 3: Comparison of classifier performance metrics (Accuracy, Cohen's Kappa, and ROC AUC) between training and test sets.

Treating these 232 windows as independent Bernoulli trials, a 95% Wilson score confidence interval for accuracy is $[91.7\%, 97.3\%]$. This interval, however, is optimistic: the 232 windows are drawn from only four patients, so predictions within the same patient are statistically dependent (clustered), and the true effective sample size for generalization is closer to four independent units than to 232. Consequently, the reported metrics should be read as a within-cohort estimate of discriminative capacity rather than as a population-level performance guarantee. A leave-one-patient-out evaluation on an expanded cohort, together with patient-level (rather than window-level) confidence intervals – e.g., via a cluster bootstrap resampling patients rather than windows – is identified as a necessary next step before the reported metrics can be interpreted as clinically generalizable, and is planned as immediate future work.

C. Per-Class Analysis

To assess class-specific discrimination capacity, we analyzed the confusion matrix for the test set (Table III). The model demonstrated balanced sensitivity across both classes: 94.8% for good outcome (CPC 1–2) and 95.7% for poor outcome (CPC 3–5). This symmetry is clinically significant, as it indicates that the model does not exhibit bias toward either class—a common pitfall in imbalanced clinical datasets.

TABLE III: Confusion Matrix for the Test Set (4 Patients)

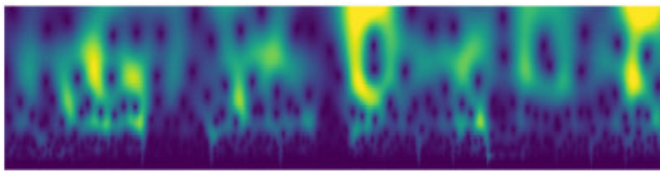
		Predicted Good	Predicted Poor
Actual	Good	110	6
	Poor	5	111

The low number of misclassifications (11 out of 232 total predictions) suggests that the spectral texture features extracted by the CWT + ResNet50 pipeline may capture prognostically relevant markers, though this observation is drawn from a small, four-patient test set and should be treated as preliminary.

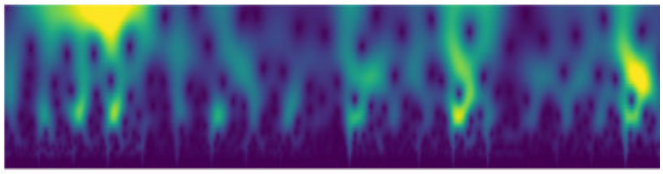
Notably, the model correctly identified 111 out of 116 poor-outcome windows, indicating a favorable sensitivity profile for detecting unfavorable prognoses within this cohort—a property that would be clinically desirable if confirmed on larger samples, since false negatives (misclassifying a poor outcome as good) carry severe ethical implications.

D. Impact of Window Size

A notable finding concerns the influence of temporal resolution on feature extraction quality. Preliminary tests with 500 ms windows yielded an accuracy of 87.9%, compared to 95.3% with 1 s windows—a 7.3 percentage-point difference observed in this cohort. This pattern is consistent with the physiological hypothesis that prognostic markers in critical states reside predominantly in low frequencies (Delta band, 0.5–4 Hz), although, given the small sample, it should be regarded as a motivating observation for a dedicated ablation study on a larger cohort rather than a settled result. A 500 ms window physically constrains the representation of complete slow-wave cycles, which may explain part of the observed gap. By extending the window, the CWT scalogram (Fig. 4) captured greater energy density in these bands, which appears to have helped the ResNet50 identify diffuse slowness patterns associated with unfavorable outcomes in this dataset.



(A)



(B)

Fig. 4: CWT scalograms for (A) favorable outcome and (B) unfavorable outcome. Although the images may present macroscopic similarities to visual inspection, the computational model extracts latent features in the spectral energy distribution that enable correct classification.

E. Spectral Texture vs. Visual Morphology

A common barrier to AI adoption in clinical practice is the expectation of "explicit visual biomarkers" analogous to radiological findings (e.g., a tumor mass). However, qualitative analysis of the generated images reveals that the distinction between "good" and "poor" prognoses does not always manifest through macroscopic morphological events visible to the human specialist.

While traditional EEG visual inspection searches for specific graphoelements (e.g., spikes, sharp waves), the Convolutional Neural Network appears to operate by identifying spectral texture patterns and non-linear correlations distributed throughout the time–frequency image. The observation that the algorithm correctly classifies segments lacking obvious visual abnormalities is consistent with the hypothesis that the model detects micro-alterations in cerebral dynamics—candidate "spectral signatures" that may precede the manifestation of coarse graphic patterns. We emphasize, however, that this interpretation is qualitative and has not yet been verified with a formal interpretability method; it should therefore be read as a hypothesis motivating future work rather than as an established mechanism.

This hypothesis is broadly consistent with recent studies suggesting that deep learning models can identify subtle electrophysiological patterns that escape visual detection [13], [22]. Confirming it, however, requires interpretability methods (e.g., Grad-CAM, SHAP) to bridge the gap between model predictions and clinical understanding—a direction we identify as a priority for future work rather than a claim made by the present study.

F. Comparison with Literature

The performance of the proposed system compares favorably, in point-estimate terms, with recent studies on EEG-based prognostication after cardiac arrest. Tjepkema-Cloostermans et al. [13] achieved an AUC of 0.89 using quantitative EEG features and logistic regression, while Westhall et al. [6] reported a sensitivity of 91% for standardized EEG interpretation. Our model's AUC of 0.991 and balanced sensitivity (94.8% / 95.7%) nominally exceed these benchmarks; however, this comparison must be qualified: the present pilot cohort (16 patients, 4 in the test set) is substantially smaller than those typically used in the cited studies, and metrics obtained from small, single-pipeline cohorts are known to be optimistically biased relative to multi-cohort or externally validated estimates. We therefore present these figures as an encouraging exploratory signal rather than as evidence of outperforming the cited methods in a rigorous head-to-head sense.

However, direct comparison is further limited by differences in datasets, preprocessing pipelines, and evaluation metrics. The I-CARE database used in this study is multi-center and temporally standardized, which enhances generalizability but may differ from other cohorts in terms of patient demographics and recording protocols.

V. Limitations and Future Work

Despite the promising results, this study presents limitations that must be acknowledged. First, the sample size (16 patients) restricts the absolute generalization of findings to heterogeneous clinical scenarios. Multi-center validation on larger cohorts is essential before clinical deployment. Second, the patient-level split, while methodologically rigorous to prevent

data leakage, yields a small test set (4 patients); because test-set windows are clustered within these four patients rather than statistically independent, the effective sample size underlying the reported metrics is much smaller than the raw window count of 232, and the associated uncertainty is larger than a naive window-level confidence interval would suggest (Section IV-B). Moreover, the reported metrics stem from a single train/test split rather than from repeated or leave-patients-out evaluation (Section III-F.3), so they do not yet capture the variability that could arise from a different choice of held-out patients. Third, the current pipeline relies on handcrafted preprocessing choices (filter cutoffs, wavelet parameters, window size) that were tuned empirically on this same small cohort, which raises the possibility of implicit overfitting to cohort-specific characteristics even though the final classifier was not directly tuned on the test patients. Fourth, while the model captures latent spectral textures, a formal interpretability analysis has not been performed—this limits direct association between extracted features and specific electrophysiological patterns, and the discussion in Section IV-D should accordingly be read as hypothesis-generating rather than confirmatory.

Finally, it is crucial to contextualize the scope of this work as a proof-of-concept study. Although the patient-level split prevents data leakage, the small cohort size may limit the statistical robustness and generalizability of the reported benchmarks. Moreover, although the model successfully captures spectral textures, this pilot study does not incorporate formal Explainable AI (XAI) techniques—such as Grad-CAM or SHAP—to interpret its internal decision process. These boundaries were deliberately established to validate the feasibility and low computational cost of the pipeline prior to resource-intensive clinical deployment. Consequently, the current results should be interpreted as a hypothesis-generating signal, rather than as a definitive clinical benchmark.

Future work will focus on: (i) validation on expanded multi-center cohorts; (ii) end-to-end Convolutional Neural Networks that jointly learn preprocessing and classification; (iii) investigation of attention mechanisms and interpretability methods (e.g., Grad-CAM, SHAP) to improve model transparency and highlight the time–frequency regions most influential for prognosis; and (iv) development of a bedside real-time prototype for continuous prognostic monitoring in the ICU.

VI. Conclusion

This pilot study evaluated the feasibility of a clinical decision-support pipeline based on automated EEG signal analysis for post-cardiac arrest prognosis. The proposed approach, combining the time–frequency resolution of the Continuous Wavelet Transform with a Random Forest classifier fed by ResNet50-extracted features, achieved encouraging performance metrics on a small, independent test set (accuracy 95.27%, 95% CI $[91.7\%, 97.3\%]$; AUC 0.991; Kappa 0.9054), offering preliminary support for the hypothesis that reliable prognostic markers can be extracted mathematically even in the absence of overt visual graphoelements. Given the small cohort (16 patients, 4 in the test partition) and the resulting clustering of predictions within few patients, these results should be treated as hypothesis-generating rather than as a validated clinical tool.

A central technical contribution was the preprocessing strategy, specifically the exploratory finding that 1-second segmentation windows outperformed shorter windows in this cohort, consistent with the physiological role of slow-wave (Delta) activity as a candidate biomarker of functional recovery versus irreversible damage. The channel vectorization and concatenation strategy allowed synthesizing the brain's global topography into a single attribute representation, simplifying modeling, though at the acknowledged cost of introducing channel-boundary discontinuities whose impact was not directly quantified.

The hybrid architecture—CWT for dense spectral representation, frozen ResNet50 for high-level feature extraction, and Random Forest for classification—appears well suited to the small-sample regime typical of clinical neurophysiology pilot studies. Before this pipeline can be considered a candidate decision-support tool for intensive care, however, it requires validation on a substantially larger, multi-center, externally held-out cohort, patient-level (rather than window-level) uncertainty quantification, and formal interpretability analysis, all of which we identify as immediate next steps.

Acknowledgment

The authors thank the Universidade Federal de Pernambuco (UFPE) and the Department of Biomedical Engineering for the infrastructure provided. We also extend our gratitude to our classmates for the daily partnership and mutual support throughout this project.

References

[1] E. Niedermeyer and F. H. Lopes da Silva, *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, 7th ed. Oxford: Oxford University Press, 2018.

[2] R. W. Neumar *et al.*, "Post-cardiac arrest syndrome: epidemiology, pathophysiology, treatment, and prognostication," *Circulation*, vol. 118, no. 23, pp. 2452–2483, 2008.

[3] C. Sandroni *et al.*, "Prognostication in comatose survivors of cardiac arrest: An advisory statement from the European Resuscitation Council and the European Society of Intensive Care Medicine," *Intensive Care Medicine*, vol. 40, no. 12, pp. 1766–1780, 2014.

[4] J. C. Griemert *et al.*, "Multimodal neuromonitoring in patients with acute brain injury," *Neurocritical Care*, vol. 28, no. 3, pp. 523–534, 2018.

[5] J. Hofmeijer and M. J. A. M. van Putten, "EEG in postanoxic coma: Prognostic and diagnostic value," *Clinical Neurophysiology*, vol. 127, no. 4, pp. 2047–2055, 2016.

[6] E. Westhall *et al.*, "Standardized EEG interpretation accurately predicts prognosis after cardiac arrest," *Neurology*, vol. 86, no. 16, pp. 1482–1490, 2016.

[7] J. J. Halford *et al.*, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG," *Clinical Neurophysiology*, vol. 126, no. 9, pp. 1661–1669, 2015.

[8] L. J. Hirsch *et al.*, "American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2021 Version," *Journal of Clinical Neurophysiology*, vol. 38, no. 1, pp. 1–29, 2021.

[9] E. Amorim *et al.*, "I-CARE: International Cardiac Arrest REsearch consortium Database (version 2.1)," *PhysioNet*, Jun. 2023. doi: 10.13026/m33r-bj81.

[10] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Burlington: Academic Press, 2009.

[11] C. Torrence and G. P. Compo, “A practical guide to wavelet analysis,” *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998.

[12] A. Subasi, “Automatic recognition of alert, drowsy, and sleeping EEG signals using wavelet packet decomposition and artificial neural networks,” *Expert Systems with Applications*, vol. 37, no. 12, pp. 7955–7962, 2010.

[13] A. Tjepkema-Cloostermans, M. C. Beishuizen, and M. J. A. M. van Putten, “Outcome prediction in postanoxic coma with deep learning,” *Critical Care Medicine*, vol. 47, no. 10, pp. 1424–1432, 2019.

[14] Y. Li *et al.*, “Deep learning for EEG-based brain-computer interfaces: A review,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3366–3385, 2021.

[15] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[18] M. X. Cohen, “A better way to define and describe Morlet wavelets for time-frequency analysis,” *NeuroImage*, vol. 199, pp. 81–86, 2019.

[19] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[21] H. U. Amin *et al.*, “Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques,” *Australasian Physical & Engineering Sciences in Medicine*, vol. 38, no. 1, pp. 139–149, 2015.

[22] B. Jiao *et al.*, “Neural biomarker diagnosis and prediction to mild cognitive impairment and Alzheimer’s disease using EEG technology,” *Alzheimer’s Research & Therapy*, vol. 15, no. 1, p. 32, 2023.

Response to Reviewers

Dear Reviewers,

First, we thank you for your suggestions and time dedicated to reviewing the manuscript "Objective Characterisation of Pronation-Supination Movements in Individuals with Parkinson's Disease Undergoing Pharmacological Treatment" (Paper ID: 285), as well as for the opportunity to submit a revised version of this manuscript to the "Brazilian Congress on Biomedical Engineering".

We have carefully considered all suggestions and revised the manuscript accordingly. Below, we provide a point-by-point response to each comment. For clarity, the reviewers' comments are presented in bold, followed by our responses. Changes made to the manuscript are indicated in red in the annotated version of the article.

Responses to the comments from Reviewer 1

Comment: Gyroscope-based characterization of pronation-supination may complement visual bradykinesia ratings, but nine medicated participants support only a pilot description.

Response: We thank the reviewer for this observation. We agree that the small number of participants limits the generalizability of the results derived from the statistical comparisons. However, the aim of all the proposed analyses was not to provide a clinical validation of gyroscope-based measurements, but rather to present a quantitative characterisation of movement patterns, primarily on an individual basis without relying on measurements from other participants. Rather than focusing exclusively on the average response across participants, this study provides insight into individual response patterns, highlighting the heterogeneity that may be masked by group-level analyses, particularly in small or more homogeneous samples. To clarify this point, we have revised the study objective to more explicitly reflect its focus on characterizing individual response patterns and inter-individual variability.

Comment: Report age, sex, disease stage, affected side, medication dose and timing, MDS-UPDRS scores, sensor placement, sampling, calibration, task instructions, repetitions and signal processing.

Response: We thank the reviewer for this suggestion. We would like to clarify that most of these points had already been addressed in the original version of the manuscript. To provide further detail on the description already provided, we added information regarding the time elapsed between medication administration and assessment, as well as the anatomical region in which the sensors were positioned and the device used to acquire signals, with its settings. Although MDS-UPDRS scores are not reported, scores on the Hoehn and Yahr scale are provided, which reflect the patient's overall clinical status; this scale is not as sensitive to motor fluctuations as the MDS-UPDRS, but it is statistically correlated with it (M. Skorvanek, P. Martinez-Martin, N. Kovacs et al., "Differences in MDS-UPDRS scores based on Hoehn and Yahr stage and disease duration," Movement Disorders Clinical Practice, vol. 4, no. 4, pp. 536-544, 2017). The modifications made to the text are highlighted in red.

Comment: Clarify whether both limbs and repeated cycles were analyzed as independent observations.

Response: For statistical comparisons, the measurements extracted from the cycles (i.e., the peak-to-peak values) were obtained from the same participant for each limb, and these values were averaged within each limb for each participant prior to statistical analysis; this was not necessary for the period, given that this measurement is already an aggregate of the cycles. The aggregated values for the two limbs were then compared using a paired-samples t-test, with each participant serving as the pairing unit. To help readers better understand this in the text, we explain this point in more detail in the "Statistical Analysis" subsection. All modifications made to the text are highlighted in red.

Comment: Use participant-level summaries or mixed models, provide individual trajectories and uncertainty, and relate features to blinded clinical ratings.

Response: We thank the reviewer for these suggestions. As explained in more detail in the response to the previous comment, the statistical analyses already took participant-level summaries into account. In the revised manuscript, we have included effect size estimates for the statistical comparisons, as described in the Methods section, to provide a more informative assessment of the magnitude and uncertainty of the observed effects.

Regarding a correlation between the proposed features and the blinded clinical scores, as the study proposes to characterise the pronation and supination movements in order to better understand the profile of these movements in Parkinson's disease and, when present, to quantify bradykinesia, considering that these movements are typical to evaluate this motor signal, an evaluation of the convergence of the measurements obtained with the inertial sensors and the clinical scores goes beyond the scope of the study, because the results obtained by the different methods would have to be combined in some way and translated to describe only bradykinesia, an aspect that is now discussed in the manuscript.

<u>Comment:</u> Without an OFF-medication or control comparison, observed profiles cannot be attributed to treatment response or disease-specific impairment.

Response: We thank the reviewer for this comment. We acknowledge that assessing these conditions would be beneficial in allowing a comparison of different profiles, and in this regard, we recognize this as one of the limitations of the present study. However, we also acknowledge that, even if these comparisons were included, the findings could not be interpreted as evidence of a response to treatment or as a disease-specific impairment. This is because the effect of medication on each individual's voluntary motor response will also depend on whether the voluntary activity was learned under the influence of the medication or not; and, since this type of movement is routine in the evaluation of these patients, those with a longer duration of diagnosis are likely to be more influenced by the medication in their motor learning as well. Therefore, we hypothesize that medication may contribute to both voluntary and involuntary responses in some cases but not in others, given that even when a person is under the influence of the medication, they exhibit a sustained decline in speed – a very distinct sign of bradykinesia in PD. Thus, the manuscript has been revised, and we now discuss these aspects in more detail.

<u>Comment:</u> The prose smoothly connects sensor features with precise monitoring and individualized therapy, a pattern that may be machine-polished but does not prove AI authorship. Human revision should keep conclusions descriptive, discuss tremor, fatigue and learning effects and identify which metric is genuinely new. The contribution is a compact feature set from one standardized task. Correct statistical units, feature definitions, figures, ethics reporting and the distinction between characterization and validation.

Response: Thank you for all suggestions. We have revised the manuscript to ensure that the interpretation remains descriptive and appropriately distinguishes feature characterization from clinical validation. In addition, the suggested points are best discussed in the manuscript and aligned with the pertinent literature. All changes have been incorporated into the revised manuscript and are highlighted in red.

<u>Responses to the comments from Reviewer 2</u>

Comment: The discussion appropriately relates the observed movement patterns to clinical manifestations such as bradykinesia, motor adaptation, and medication effects, while acknowledging the limitations associated with sample size, kinematic-only assessment, and the absence of physiological measurements. Nevertheless, the manuscript would benefit from a more quantitative interpretation of the results, including a clearer justification of the practical implications of the detected change

points and a deeper comparison with existing movement analysis approaches for Parkinson's disease.

Response: We thank the reviewer for this suggestion. To address these issues, the results pertaining to the points of change were rewritten to provide a more quantitative description of the results. All modifications made to the text are highlighted in red.

In addition, the authors are encouraged to carefully review the manuscript formatting to ensure full compliance with the CBEB template.

Response: We thank the reviewer for this suggestion. The manuscript has been carefully reviewed and reformatted to ensure compliance with the event template.

Observational Analysis of the Performance of Image Reconstruction Algorithm Techniques in Head CT Protocol

Érica Mauricio Policarpo
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
https://orcid.org/0000-0001-9943-5474

Thiago Michael de Brito Farias
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
https://orcid.org/0000-0002-9377-662X

Oswaldo Simões Fontoura
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
fontoura@unifesp.br

Camila Hitomi Murata
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
https://orcid.org/0000-0002-2348-2212

Marcelo Ortiz Ficel
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
mficel@unifesp.br

Kellen A. C. Daros
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
https://orcid.org/0000-0002-0889-2563

Cássio Miri Oliveira
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
cassio.miri@unifesp.br

Arthur Sampaio Buniotti
Radiology Department of Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, Brazil
arthur.buniotti@unifesp.br

***Abstract*—CT exams are widely utilized for diagnosing pathologies due to their rapid acquisition and high image quality. To optimize radiation dose, manufacturers have developed advanced computational image reconstruction algorithms to mitigate the noise inherent in low-dose acquisitions. Following an institutional CT scanner upgrade, this study presents an observational evaluation of FBP, IR, and DLIR reconstruction algorithm techniques using Quality Assurance (QA) phantom. For low-contrast resolution, the FBP - UA kernel - resolved more holes across the three acrylic disks, whereas UB and UC kernels resolved an average of six holes in disk C. The IR resolved an average of six holes in disk C at iDose levels 1 to 6, and iDose level 7 resolved five holes in disk C. Conversely, the DLIR technique (Precise Image) yielded lowest low-contrast detectability, resolving only an average of five holes in disk C. In high-contrast analysis, FBP and DLIR resolved similar results, an average of 6.6 lp/cm, while IR resolving an average of 5.0 lp/cm. DLIR demonstrated superior SNR improvement. The CNR analysis revealed that DLIR induces a substantial loss in image contrast when compared to FBP kernels. No significant CNR variations were observed across IR levels. In conclusion, while DLIR provides unprecedented noise reduction, it may compromise low-contrast detectability in soft-tissue in clinical applications. IR demonstrated an essential digital resource to CT protocols optimization.**

Keywords—computed tomography, image reconstruction, algorithm, low contrast, CNR, SNR

I. INTRODUCTION

Computed tomography (CT) exams are widely utilized for diagnosing pathologies due to their rapid acquisition time and high image quality [1], [2]. Consequently, CT has become the primary contributor to increased population exposure to artificial ionizing radiation worldwide [3]. To optimize radiation doses, manufacturers have invested in advanced computational image reconstruction techniques to mitigate the noise inherent in low-dose acquisitions [4], [5].

While Filtered Back-Projection (FBP) remains the benchmark reconstruction method in CT, it requires higher radiation doses to maintain acceptable image quality [6]. Computational advancements in processing large data volumes have enabled the adoption of innovative reconstruction algorithms [7]. These techniques suppress noise in low-dose images through hybrid approaches, incorporating both Iterative Reconstruction (IR) and Deep Learning Image Reconstruction (DLIR) algorithms [8].

However, depending on the aggressive manipulation of the noise reduction, these algorithms can introduce a "plastic" texture to the images and potentially obscure low-contrast pathologies [5], [9], [10]. Because certain noise-reduction strength levels are operator-dependent, imaging professionals must consider alternative strategies to minimize noise in low-dose protocols [11]. For instance, simply reconstructing the image with a greater slice thickness can reduce noise without compromising critical diagnostic details.

This study presents an observational evaluation of the performance of advanced CT image reconstruction algorithm techniques regarding noise suppression, following institutional head CT protocol, after the upgrade of the CT scanner system.

II. METHODOLOGY

A. Comparative Study Using a QA Phantom

This study was conducted in the Radiology Department of Escola Paulista de Medicina, Universidade Federal de São Paulo. CT images were acquired using a Philips Incisive CT scanner (Philips Medical Systems, Netherlands). Table I presents the reconstruction algorithm techniques available on the Incisive CT system and their main characteristics.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

TABLE I. Characteristics of Incisive CT reconstruction algorithms used in this study

Reconstruction	Characteristics *
Filtered Back-Projection (FBP)	Standard FBP kernel acquisitions: UA smooth on the brain; UB standard preset for head scans; UC brain enhancement, all kernels improve the bone/brain interface, for head scans only [12].
iDose (IR)	Reconstruction techniques are used to reduce noise by implementing levels to determine the amount of noise that is removed from the images (IR). Level 1 (20%) is the least aggressive noise reduction, level 7 (80%) is the most aggressive noise reduction [12].
Precise Image (DLIR)	Precise Image is a reconstruction mode where the system uses a trained deep learning neural network to generate noise reduction images (DLIR) and improve low contrast detectability with reduced dose compared with standard FBP recon mode [12].

* From Philips Incisive CT User Guide [12].

A Philips Quality Assurance (QA) Phantom, model C for AV E1, was used for image acquisition. This methodology, i.e., the use of QA phantoms in clinical settings, is widely described in the literature [13], [14], [15], [16], [17], [18]. These results obtained can be used as image quality baselines and in optimization processes.

The institutional adult routine head CT protocol was selected for this study; Table II lists the acquisition parameters. Three acquisitions were performed, and the reconstruction algorithms applied to the same raw data to allow an unbiased comparison of the results.

TABLE II. Incisive CT parameters for the institutional adult head CT protocol acquisitions.

Exposure Parameters	
Voltage (kV)	120
Current x Time (mAs)	399
FOV*	190
Slice Thickness (mm)	1.00
Pitch	0,5
Collimation	64 x 0.625
Matrix	512 x 512
CTDIvol (mGy)	56.27

* Field of View

The low-contrast and high-contrast resolution, Signal-to-Noise Ratio (SNR) and Contrast-to-Noise Ratio (CNR) parameters were selected because they represent the primary targets for optimization by the reconstruction algorithms.

In this study, the impact of varying radiation doses, for example computed tomography dose index volume (CTDIvol) was not evaluated; only the baseline settings of the institutional adult head CT protocol were considered without dose optimization.

B. Data Analyses

For the qualitative assessments, three independent readers, each with over 10 years of experience, performed unblinded evaluations across all reconstructed images, using a 5-megapixel Barco© medical display, model MDNG-6121.

A qualitative assessment of low-contrast resolution was performed visually. Images were acquired in accordance with the institutional adult routine head CT protocol, using the QA phantom. The low-contrast resolution section of the phantom comprises three disks of varying materials and thicknesses, each containing eight holes of different diameters. Fig. 1 and Table III show the low-contrast test tool region of the phantom, with the disks identified according to their respective materials and thicknesses.

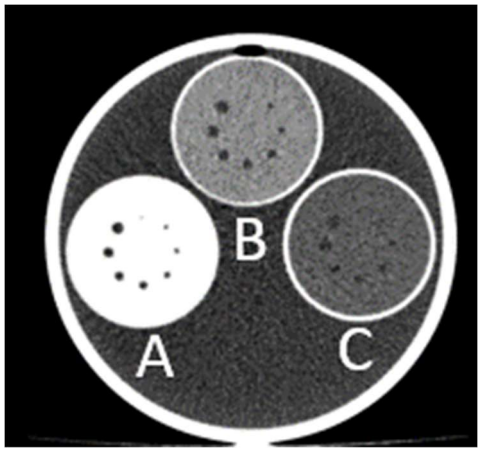


Fig.1. CT image of the QA phantom low-contrast resolution region, categorized by disk thickness.

TABLE III. QA phantom low-contrast test tool groups specifications.

Disk Identification	Material	Thickness
A	Acrylic	1.65
B	Acrylic	0.4
C	Polyester	0.075

A qualitative assessment of high-contrast resolution was performed visually, with the same methodology described for the low-contrast resolution. Images were acquired in the line pairs per centimeter (lp/cm) test tool section of the QA phantom. This assessment identified the last group in which the individual bars remained distinguishable. Fig. 2 presents the group of line pairs on the QA phantom, while Table IV lists the representative resolution values in centimeters.

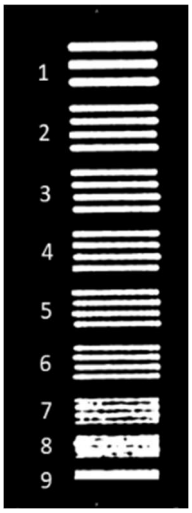


Fig. 2. CT image of the QA phantom high-contrast resolution region, exhibiting the line pairs per centimeters test tool.

TABLE IV. Representative resolution in lp/cm test tool groups

Group	Resolution (lp/cm)
1	5.0
2	6.6
3	7.1
4	7.7
5	8.3
6	9.1
7	11.1
8	12.5
9	14.3

Additionally, a quantitative analysis was conducted by calculating the SNR and CNR, in accordance with the recommendations outlined in the American Association of Physicists in Medicine (AAPM) Task Group 233 (TG-233) report and the American College of Radiology (ACR) CT Quality Control Manual.

The SNR is a measure of the signal level in the presence of noise given by the equation (1):

$$SNR = \frac{MeanValue(ROI)}{SD(ROI)} \quad (1)$$

Where Mean Value and Standard Deviation (SD) are the value of the Hounsfield Unit (HU) in a Region Of Interest (ROI) in the image.

The CNR was calculated as the ratio of Mean Value to SD in the ROI given by the equation (2):

$$CNR = \frac{Mean(Low_Contrast) - Mean(Water)}{SD(Water)} \quad (2)$$

Fig. 3 illustrates the ROI positioned for the quantitative analysis. The disk B of the low-contrast resolution section was selected for this assessment, as its HU values closely approximate those of brain tissue. The ROI had an area of 2.346 cm^2 and comprised approximately 2005 pixels.

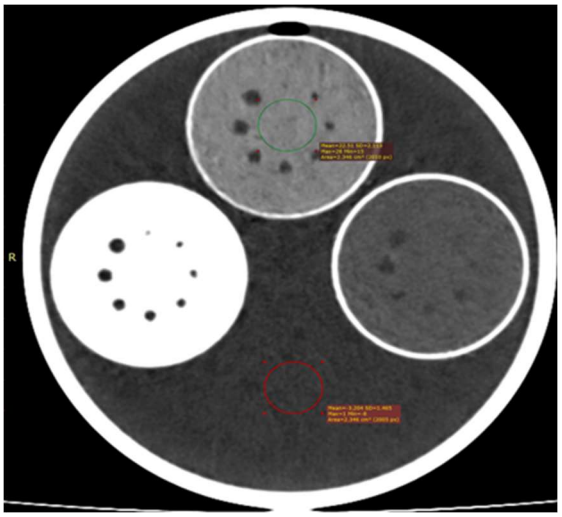


Fig. 3. ROIs location and size for SNR and CNR analysis.

III. Results and Discussion

Following CT image acquisition and the application of the respective image reconstruction algorithm techniques, a noticeable change in the visual perception of the phantom's low-contrast test tools was observed, primarily driven by excessive image noise reduction. However, the opposite occurs with high-contrast visual perception. Fig. 4 and 5 shows the visual appearance of the phantom's low-contrast and high-contrast region after the application of reconstruction algorithm techniques, respectively.

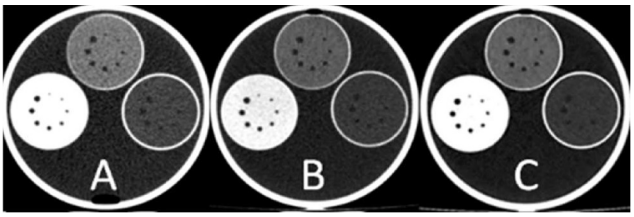


Fig. 4. Low-contrast QA phantom image. In ***A*** UA kernel (FBP), in ***B*** iDose level 7 (IR) and ***C*** Precise Image (DLIR).

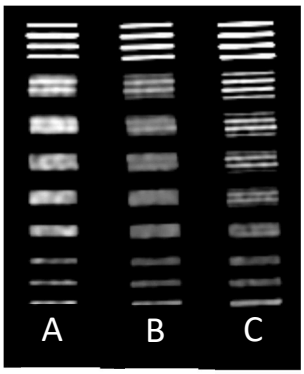


Fig. 5. High-contrast QA Phantom image. In ***A*** UA kernel (FBP), in ***B*** iDose level 7 (IR) and ***C*** Precise Image (DLIR).

A. Low-Contrast analysis

In the visual analysis, all reconstruction algorithms techniques enabled the visualization of all holes across the acrylic disks A and B by the readers. Conversely, the reconstruction algorithms techniques demonstrated random performance to resolve the holes images in the polyester disk C (Table V).

TABLE V. Results of low-contrast visual analysis in polyester disk C, presented as mean ± standard deviation based on the evaluation of three images by three independent readers.

Algorithms	Reader 1	Reader 2	Reader 3
Precise Image	5.33 ± 0.58	5.0 ± 0.0	5.33 ± 0.58
iDose 1	6.0 ± 0.0	5.33 ± 0.58	5.67 ± 0.58
iDose 2	5.67 ± 0.58	6.33 ± 0.58	6.0 ± 0.0
iDose 3	6.33 ± 0.58	5.33 ± 0.58	5.33 ± 0.58
iDose 4	6.67 ± 0.58	6.0 ± 0.0	5.33 ± 0.58
iDose 5	6.0 ± 1.0	5.33 ± 0.58	5.67 ± 0.58
iDose 6	6.33 ± 0.58	6.0 ± 0.0	5.67 ± 0.58
iDose 7	5.33 ± 0.58	5.33 ± 0.58	5.33 ± 0.58
UA Kernel	7.0 ± 0.0	6.0 ± 0.0	5.67 ± 1.15
UB Kernel	6.33 ± 0.58	5.67 ± 0.58	5.67 ± 1.15
UC Kernel	5.67 ± 0.58	6.0 ± 1.0	5.67 ± 1.15

Regarding the IR technique, five to six holes were discernible in disk C at iDose levels 1 to 6. However, at the

most aggressive level (iDose 7), an average of five holes was discernible in the disk C. Also, the DLIR technique (Precise Image) yielded inferior results regarding low-contrast feature detectability in the disk C, an average visualization of five holes.

B. High-Contrast analysis

In the high-contrast resolution analysis, the FBP, IR and DLIR techniques showed similar results (Table VI). The FBP and DLIR technique resolved an average of 6.6 lp/cm, whereas the IR technique resolved 5.0 lp/cm across all iDose levels.

The FBP technique exhibited a gradual, kernel-dependent improvement in high-contrast resolution. The UA kernel resolved an average of 5.0 lp/cm, the UB kernel resolved up to the second group (6.6 lp/cm), and the UC kernel achieved the highest spatial resolution, 7.1 lp/cm.

TABLE VI. Results of visual analysis of high-contrast, presented as mean ± standard deviation based on the evaluation of three images by three independent readers.

Algorithms	Reader 1	Reader 2	Reader 3
Precise Image	6.6 ± 0.0	6.6 ± 0.0	6.6 ± 0.0
iDose 1	5.53 ± 0.92	5.53 ± 0.92	5.53 ± 0.92
iDose 2	5.53 ± 0.92	5.0 ± 0.0	5.0 ± 0.0
iDose 3	6.07 ± 0.92	5.0 ± 0.0	5.53 ± 0.92
iDose 4	5.53 ± 0.92	5.5 ± 0.0	5.0 ± 0.0
iDose 5	5.53 ± 0.92	5.53 ± 0.92	5.0 ± 0.0
iDose 6	5.53 ± 0.92	5.53 ±0.92	5.0 ± 0.0
iDose 7	5.53 ±0.92	5.5 ± 0.87	5.0 ± 0.0
UA Kernel	5.53 ± 0.92	5.0 ± 0.0	5.0 ± 0.0
UB Kernel	6.6 ± 0.0	6.07 ± 0.92	6.07 ± 0.92
UC Kernel	6.77 ± 0.29	6.93 0.29	6.77 ± 0.29

This performance is critical since CT imaging inherently requires a certain level of noise to preserve the boundary differentiation between distinct soft-tissue densities, particularly within the brain parenchyma where structures share highly similar attenuation values.

C. SNR and CNR analysis

In a direct comparison between the reconstruction algorithm techniques, the SNR was determined in the region where the HU was similar to the brain tissue.

In the SNR analysis, the DLIR (Precise Image) demonstrated superior signal improvement (Fig.6), a direct result of the algorithm's noise correction capabilities.

Reconstructions using IR exhibited the expected trend, showing progressive signal enhancement as the iDose level increased.

FBP reconstructions also followed expected patterns: the UA kernel yielded a smoother image; the UC kernel produced the highest noise levels, and the UB kernel, standard for brain imaging, delivered intermediate noise levels between UA and UC.

Regarding CNR analysis, it is noteworthy that although the DLIR method significantly improves the SNR, it induces a substantial loss in image contrast, particularly when compared to the FBP kernels (UA and UC, Fig. 6). While the latter exhibit higher absolute image noise, they preserve greater contrast between structures of similar density, which is often preferable for the clinical evaluation of brain images. This contrast degradation manifests as the characteristic "plastic" artifact that DLIR algorithms can introduce in certain brain scans.

For the IR reconstruction algorithm, no significant variations in image contrast were observed across the different iDose levels.

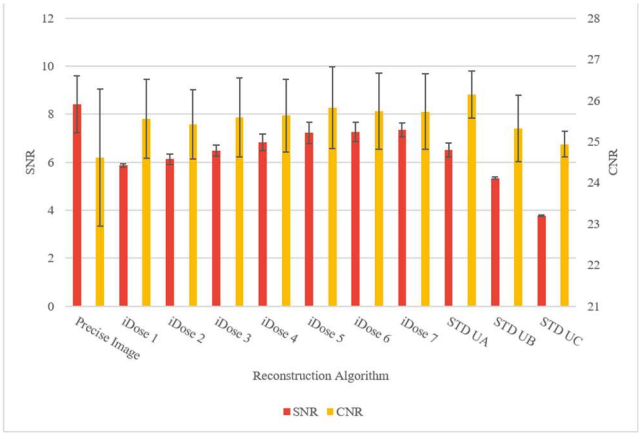


Fig. 6. SNR and CNR through reconstruction algorithm techniques analyzed.

Fig. 6 shows that, for FBP reconstructions, as the CNR is optimized (using sharper kernels), the SNR decreases due to higher noise levels. Conversely, DLIR reconstructions exhibit the opposite performance, enhancing the SNR at the expense of CNR. On the other hand, IR reconstruction techniques attempt to maintain a balance, stabilizing both SNR and CNR across different iterations.

In Fig. 7, the plot demonstrated the influence of the reconstruction algorithm technique in high-contrast resolution, SNR and CNR parameters analyzed in this study.

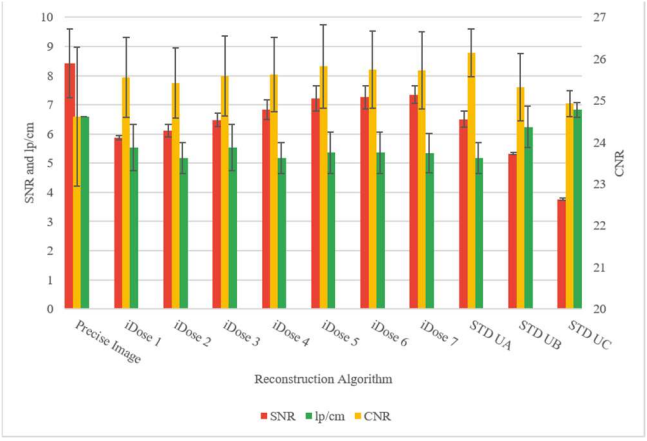


Fig. 7. Results of high-contrast, SNR and CNR through reconstruction techniques analyzed.

The findings of this study highlight the variations that modern reconstruction algorithms can cause in CT images. When significant density differences exist between adjacent objects, the DLIR algorithm (Precise Image) effectively suppresses noise and degrades high and low-contrast

resolution in the image. This phenomenon clinically manifests as a "plastic" artifact. Fig. 8 summarizes the results from qualitative image analysis for high and low-contrast resolution.

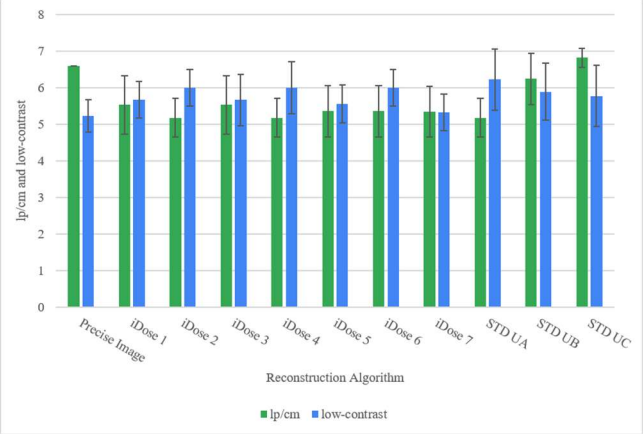


Fig.8. Results from qualitative image analysis for high and low-contrast.

The observed degradation in CNR correlates directly with high-contrast performance. The superior visualization of lp/cm achieved with the UC FBP kernel can be attributed to its better preservation of CNR, despite the higher background noise.

However, modern reconstruction algorithms, such as those based on IR and DLIR, can contribute to the optimization of clinical examinations as they compensate for image noise generated by low doses.

These results suggest that while deep-learning-based reconstructions offer unprecedented noise suppression, traditional FBP kernels remain highly valuable for preserving boundary definitions in low-contrast diagnostic tasks.

IV. Conclusions

This study shows that although DLIR algorithms (e.g., Precise Image) offer superior noise reduction and SNR enhancement, careful consideration is required due to potential low-contrast degradation in structures with similar attenuation.

Iterative reconstruction algorithms like iDose remain essential components for institutional CT protocol optimization.

Further optimization studies are currently underway to establish dose reduction strategies, that balance the noise suppression of IR and DLIR techniques with clinical image quality requirements.

References

[1] S. K. Zhou *et al.*, "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021, doi: 10.1109/JPROC.2021.3054390.

[2] M. Lell and M. Kachelrieß, "Computed Tomography 2.0: New Detector Technology, AI, and Other Developments," *Invest. Radiol.*, vol. 58, no. 8, pp. 587–601, Aug. 2023, doi: 10.1097/RLI.0000000000000995.

[3] N. R. Smoll, Z. Brady, K. J. Scurrah, C. Lee, A. Berrington De González, and J. D. Mathews, "Computed tomography scan radiation and brain cancer incidence," *Neuro-Oncol.*, vol. 25, no. 7, pp. 1368–1376, Jul. 2023, doi: 10.1093/neuonc/noad012.

[4] S. Panda, M. Jangid, and A. Jain, "A Comprehensive Review on the Significance and Impact of Deep Learning in Medical Image Analysis," in *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, Tashkent, Uzbekistan: IEEE, Nov. 2021, pp. 358–366. doi: 10.1109/ICTAI53825.2021.9673189.

[5] A. Sahu *et al.*, "Transforming CT imaging with deep learning: Noise reduction, artifact management, and clinical applications – A comprehensive review," *Eur. J. Radiol. Artif. Intell.*, vol. 4, p. 100042, Dec. 2025, doi: 10.1016/j.ejrai.2025.100042.

[6] X. Tan, X. Liu, K. Xiang, J. Wang, and S. Tan, "Deep Filtered Back Projection for CT Reconstruction," *IEEE Access*, vol. 12, pp. 20962–20972, 2024, doi: 10.1109/ACCESS.2024.3357355.

[7] M. Pula, E. Kucharczyk, A. Zdanowicz, and M. Guzinski, "Image Quality Improvement in Deep Learning Image Reconstruction of Head Computed Tomography Examination," *Tomography*, vol. 9, no. 4, pp. 1485–1493, Aug. 2023, doi: 10.3390/tomography9040118.

[8] L. L. Geyer *et al.*, "State of the Art: Iterative CT Reconstruction Techniques," *Radiology*, vol. 276, no. 2, pp. 339–357, Aug. 2015, doi: 10.1148/radiol.2015132766.

[9] T. Lin *et al.*, "MVSS-PCR: A Multiview Subsampling Approach With Prior-Constrained Reconstruction for Self-Supervised CT Denoising," *IEEE Trans. Instrum. Meas.*, vol. 75, pp. 1–13, 2026, doi: 10.1109/TIM.2026.3671923.

[10] W. Stiller, "Basics of iterative reconstruction methods in computed tomography: A vendor-independent overview," *Eur. J. Radiol.*, vol. 109, pp. 147–154, Dec. 2018, doi: 10.1016/j.ejrad.2018.10.025.

[11] B. Al Mohammad, M. M. Alakhras, and W. Reed, "Assessing the knowledge of CT radiographers regarding how CT parameters affect patient dose and image quality," *Eur. J. Radiol.*, vol. 166, p. 111023, Sep. 2023, doi: 10.1016/j.ejrad.2023.111023.

[12] Philips Healthcare, "User Guide for Philips Incisive CT Scanner." Philips Medical Systems Nederland B.V., 2021. Accessed: Jul. 15, 2026. [Online]. Available: https://www.documents.philips.com/assets/Instruction%20for%20Use/20210811/81380df2c2b5451d9bf5ad81005fd6d7.pdf?feed=ifu_docs_feed

[13] F. Zarb, L. Rainford, and M. F. McEntee, "Image quality assessment tools for optimization of CT images," *Radiography*, vol. 16, no. 2, pp. 147–153, May 2010, doi: 10.1016/j.radi.2009.10.002.

[14] M. Zellner, S. Tschauner, M. S. Weyland, P. E. Hotz, S. Scheidegger, and C. J. Kellenberger, "Low-dose lung CT: Optimizing diagnostic radiation dose – A phantom study," *Eur. J. Radiol. Open*, vol. 13, p. 100614, Dec. 2024, doi: 10.1016/j.ejro.2024.100614.

[15] H. Amasya, Ş. Özel, D. Tunçman, S. Çavdar Karaçam, K. Orhan, and M. Demir, "Comparing Anthropometric and Quality Assurance Phantoms in Quantitative Image Quality Testing for CBCT Imaging," *ADO Klin. Bilim. Derg.*, vol. 14, no. 2, pp. 101–108, May 2025, doi: 10.54617/adoklinikbilimler.1546507.

[16] Thalis L.A. Saint' Yves, Paulo Cesar B. Travassos, Elicardo A.S. Gonçalves, Fernando Mecca A., and Thiago B. Silveira, "Otimização da relação dose x ruído na imagem em protocolos de tomografia computadorizada de crânio pediátrico," *Optimization of the dose x noise in the image on protocols for computed tomography of pediatric head*, vol. 4(1), Revista Brasileira de Física Médica, pp. 11–14, 2010.

[17] M. Mambrini *et al.*, "The importance of routine quality control for reproducible pulmonary measurements by in vivo micro-CT," *Sci. Rep.*, vol. 12, no. 1, p. 9695, Jun. 2022, doi: 10.1038/s41598-022-13477-7.

[18] D. R. D. Pina *et al.*, "Controle de qualidade e dosimetria em equipamentos de tomografia computadorizada," *Radiol. Bras.*, vol. 42, no. 3, pp. 171–177, Jun. 2009, doi: 10.1590/S0100-39842009000300010.

Open-Source Electrooculography Platform for Assistive Double Blink Detection: Design, Implementation, and Validation

1st Mariana Ferraz Marcelino
Department of Electrical and Computer Engineering
Federal Institute of São Paulo
Sertaozinho, SP, Brazil
marianaferrazm505@gmail.com

2nd Pedro do Carmo Sampaio
Department of Electrical and Computer Engineering
Federal Institute of São Paulo
Sertaozinho, SP, Brazil
pedrodocarmo10@gmail.com

3rd Rodrigo Palucci Pantoni
Department of Electrical and Computer Engineering
Federal Institute of São Paulo
Sertaozinho, SP, Brazil
rpantoni@ifsp.edu.br

4th Marcio Bender Machado
Department of Electrical and Computer Engineering
Federal Institute of São Paulo
Sertaozinho, SP, Brazil
marciobma@gmail.com

***Abstract*—This article addresses the challenge of accessibility for people with motor disabilities and highlights the importance of assistive technologies, focusing on human-machine interfaces based on biosignals. It presents the development of an open-source platform for electrooculography (EOG) signal acquisition using the ADS1299-4 chip, aimed at detecting double blinks and converting them into assistive commands. The work fills a gap in the literature by providing detailed hardware schematics, PCB layout, and signal acquisition methodology, supporting scientific reproducibility. A pilot study validated the system, with double-blink events classified by a Support Vector Machine (SVM) model achieving 78% precision, demonstrating feasibility for real-time and embedded applications even in the presence of signal noise. The main contributions are: (1) the creation of an open-source, reproducible EOG acquisition platform; (2) validation for double-blink detection in assistive interfaces; and (3) public release of all materials to foster further research and adaptations, advancing independent interaction for individuals with limited mobility.**

***Index Terms*—Assistive technology, electrooculography (EOG), biosignal acquisition, open-source hardware, double blink detection, human-machine interface, support vector machine (SVM).**

I. Introduction

In a study involving adults from 59 countries, 7.2% of the surveyed population reported having severe or extreme difficulty in moving around [1]. This figure indicates that a considerable portion of society has some type of motor limitation. Furthermore, for many of these individuals, assistance and support are necessary for participation in society [1]. This makes this population highly dependent on family members, preventing both individuals with disabilities and their caregivers from becoming economically active and socially included.

In this regard, assistive technologies are extremely important for reducing this dependence and, consequently, promoting greater autonomy and integration for these individuals. Assistive technology aims precisely to promote autonomy and social interaction through resources, devices, technologies, and services.

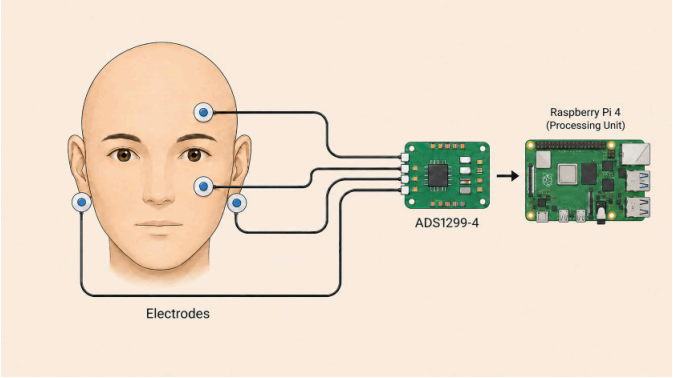


Fig. 1. System overview - Blink acquisition and processing

In this context, assistive human-machine interfaces emerge as solutions that promote interaction with systems and device control for people with physical disabilities. In [2], several studies that use biosignals for conversion into digital commands are presented. In addition, in [3], a similar application is presented, in which double blinks, acquired from EOG signals, and head movements are used to control a robot. In [4], a wearable headband with dry electrodes is presented, and the detection of events such as eye movements and single blinks is investigated.

Regarding the acquisition of EOG signals, [2] shows that most approaches use commercial solutions, while a few are implemented on custom acquisition boards. However, even

in cases where such customization is proposed, few studies presented details of the developed hardware, such as electrical schematics, system layouts, and development guidance, thereby hindering project reproducibility and limiting scientific contributions through improvements, adaptations, and further research based on these circuits.

Considering this aspect, the objective of this work was to develop an open-source, customized acquisition platform using the ADS1299-4 chip to acquire double blinks via electrooculography (EOG) signals. The detected blinks were classified using the Support Vector Machine (SVM) algorithm to be used as commands, expanding the possibilities of interaction for people with physical disabilities and contributing to the reduction of accessibility barriers.

The main contribution of this work is the development and validation of an open-source, EOG-customized acquisition platform, including PCB design, electrical schematics, and signal acquisition methodology, enabling reproducible research in assistive human-machine interfaces, which is available at [5]. The project overview is shown in Fig. 1

The remainder of this paper is organized as follows. The Theoretical Background section presents key concepts and information necessary for a better understanding of the work. The Materials and Methods section describes the entire process carried out to develop the project. In the Results and Discussion section, the results are presented and discussed, including the metrics used to evaluate the proposed solution's performance. Finally, the Conclusion section presents the final considerations based on the results achieved.

II. Theoretical Background

To contextualize the development of this work, this section presents the theoretical background related to the acquisition and use of EOG signals in human-machine interface applications. It addresses biological signals, with emphasis on electrooculography (EOG), the use of double blinks as interaction commands, the main characteristics of the ADS1299-4, and referential and bipolar electrode montages.

A. Electrooculography Signals

Signals originating from physiological activities of the human body are referred to as biosignals and can be measured, processed, and analyzed. Among these signals, bioelectric signals stand out, as they are characterized by voltage or electric current variations generated by biological processes. In this context, the EOG consists of the recording of electrical variations associated with the ocular region, mainly resulting from the potential difference generated by eye movements and blinks.

B. Blinks as Interaction Command

The EOG signal enables the detection of ocular events, such as blinks. These events can then be processed and analyzed to be assigned to specific commands or used as command confirmation mechanisms, thereby increasing safety in human-machine interaction. In a work, for example, blink frequency was associated with specific commands, demonstrating the broad applicability of this type of signal [6].

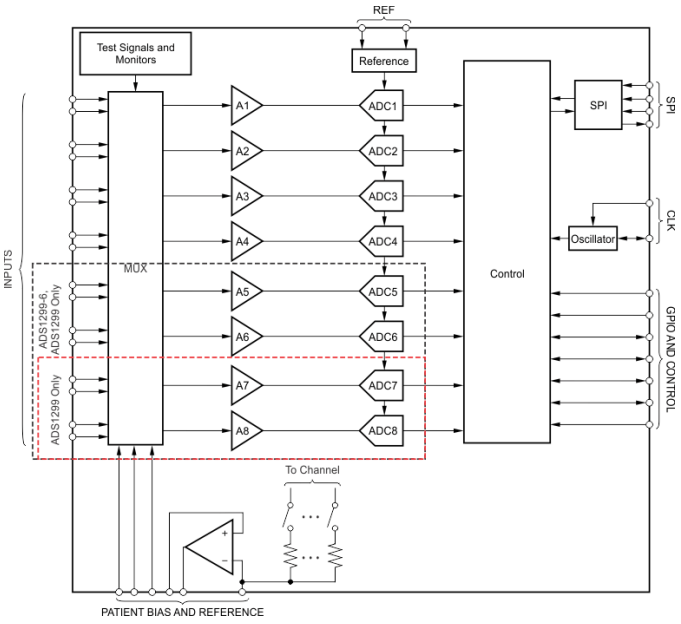


Fig. 2. Block diagram of ADS1299-4 [8]

A blink is an event that can be easily identified in the EOG signal, generally characterized by the presence of a peak or valley in the signal. In the case of a double blink, its main characteristic is the occurrence of two consecutive peaks or two consecutive valleys within a short time interval [7].

C. ADS1299-4

The ADS1299-4 chip is a low-noise, 24-bit analog-to-digital converter with a programmable gain amplifier and SPI communication, designed especially for biosignal acquisition. This chip is available on the market in different versions, which differ only in the number of channels.

Its high resolution, programmable gain amplifier, and low-noise architecture make it suitable for acquiring ECG, EEG, and EOG signals. The Figure 2 shows how the chip operates internally. It includes amplifier circuits, analog-to-digital converters, an internal control module, and a BIAS circuit, which is used to help reduce noise. The signals from the electrodes pass through the analog inputs, are routed by the multiplexers, amplified, and then converted from analog to digital. After this conversion, the data are managed by the internal control block and made available to the external system through SPI communication.

Through this chip, it is possible to develop robust and compact devices, in addition to providing an affordable solution when compared to ready-made systems.

D. Referential and Bipolar Electrode Montages

For biosignal acquisition, two electrode montages stand out for enabling better event detection: the referential montage and the bipolar montage. In the referential montage, the reference electrode is common to all active electrodes, whereas in the

bipolar montage, the potential difference is measured between two electrodes placed in adjacent regions. The referential montage is suitable for analyzing signals with respect to the same reference, while the bipolar montage is suitable for comparing electrical potentials from nearby regions.

III. Materials and Methods

A. Hardware Architecture

First, the system for acquiring and providing EOG signals was developed using the ADS1299-4 chip.

The circuit is composed of two main sections: the analog section, responsible for extracting and amplifying the biosignals; and the digital section, which provides the measurements through the SPI communication protocol. In this sense, the printed circuit board design can be seen in Figure 3.

Part "A" of the figure represents the analog side of the circuit. In this section, it can be observed that each electrode has an associated capacitive low-pass filter, whose purpose is to allow only frequencies below 6.9 kHz to be acquired. Furthermore, an active BIAS reference electrode is also present, whose function is to attenuate noise, such as interference from the power line, for example.

Part "B" of the figure shows the digital region of the system. This section can be connected to a computational device so that the chip can be configured, including parameters such as gain, sampling frequency, among others, and then provide the data according to the defined settings.

It is also worth noting that the two regions have their GND planes interconnected through connector J3. This ensures equipotentiality while maintaining the separation between the digital and analog regions. ***In addition, a current-limited power supply was used, with a maximum current of 4 mA for the analog section and 1 mA for the digital section. All these factors contribute to reducing interference and enhancing user safety.***

Moreover, a symmetrical power supply was adopted for the circuit, and the referential electrode montage was used.

The developed electrical schematic is shown in Figure 4. The electrical schematic was divided into two parts, A and B. Part A corresponds to the electrode connections, initially organized in a referential montage, and also includes analog filters responsible for attenuating noise and limiting the frequency range of the acquired signal. Part B corresponds to the SPI communication between the ADS1299-4 and the control system. In this section, some pins are connected to resistors in order to help stabilize the logic levels, reduce possible interference, and avoid undefined states in certain digital inputs. Unused pins should be kept at a known logic level, when necessary, through pull-up or pull-down resistors, preventing incorrect readings or undesired circuit behavior.

B. Signal Acquisition

Data acquisition was performed using the developed hardware together with a Raspberry Pi 4. Communication between the embedded system and the hardware was carried out through the SPI interface. For this stage, a Python code developed by [9] was used as a basis.

Next, the data collection stage was carried out in order to subsequently evaluate the quality of the signal provided by the circuit.

The selected electrode placement is shown in Figure 5, consisting of two active electrodes, one below and one above the eye, a reference electrode placed on the right earlobe, and a BIAS electrode placed on the left earlobe.

Additionally, the protocol used for this process was as follows: recording of 20 double blinks over 2 minutes; 1 minute in a resting state, with the participant blinking naturally and maintaining a fixed gaze; 10 facial contractions distributed over 1 minute; and, finally, 10 vertical eye movements, upward and downward, also performed over 1 minute. The data were saved in a CSV file.

C. Signal processing

It was necessary to process the stored data in order to make them more suitable for the tests that would be performed subsequently. Therefore, the following steps were carried out: outlier removal; digital filtering using a 0.5 Hz high-pass filter and a 15.0 Hz low-pass filter; windowing with 1-second windows and manual labeling of the windows that contained double blinks. After this process, a new CSV file was generated containing the columns *t_s*, *filtered_signal*, *window* (window with its respective ID) and *classification*.

D. Tests and Validation

A pilot study was conducted with two healthy volunteers to validate the hardware and signal acquisition process. Then, to check the performance of all system, the mean square power of a noise sample and a filtered signal sample was calculated using Equation 1. Subsequently, the average signal-to-noise ratio (SNR) was determined using Equation 2.

$$P = \frac{1}{N}\sum_{n=1}^{N} x^2[n] \tag{1}$$

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right) \tag{2}$$

In addition, a SVM model was trained and tested for double-blink identification, and its performance was evaluated using the metrics of accuracy, recall, and precision, calculated respectively by Equations 3, 4, and 5.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- True Positive (TP): When the model predicts the positive class and the actual class is positive.

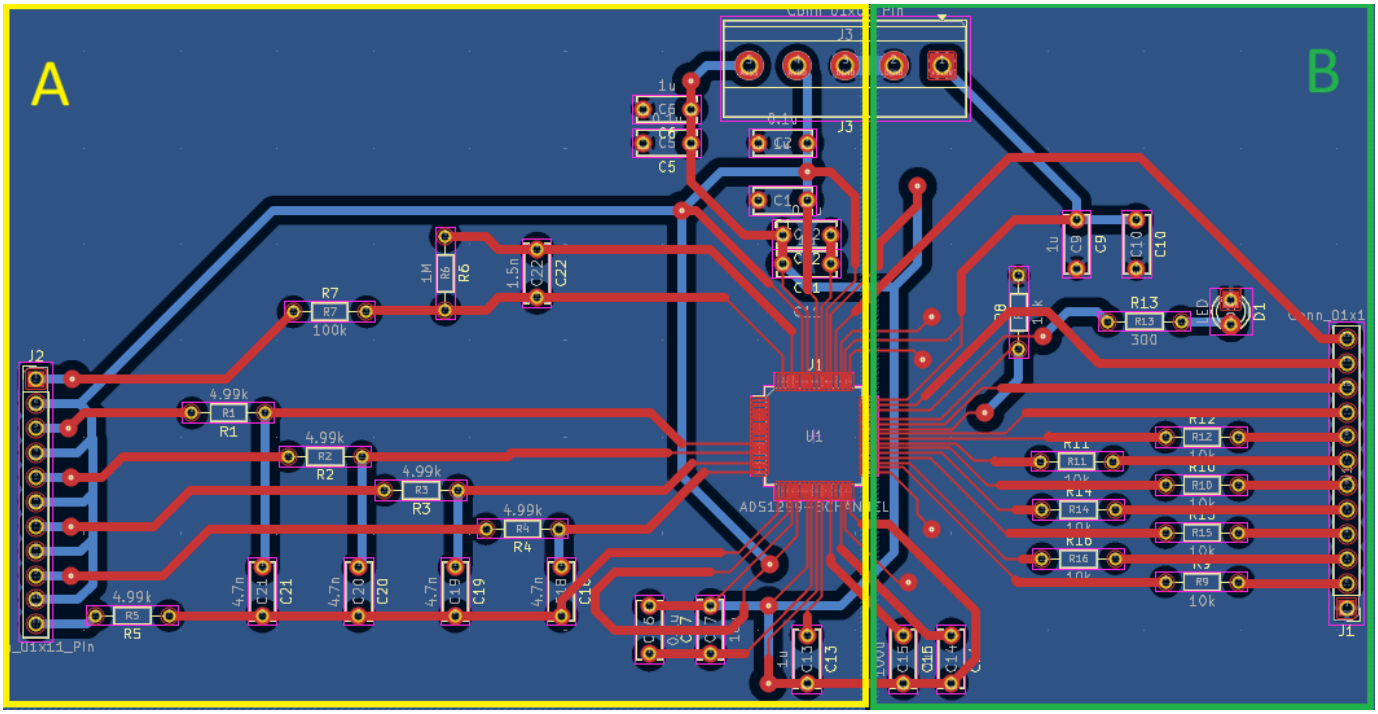


Fig. 3. Layout of the printed circuit board

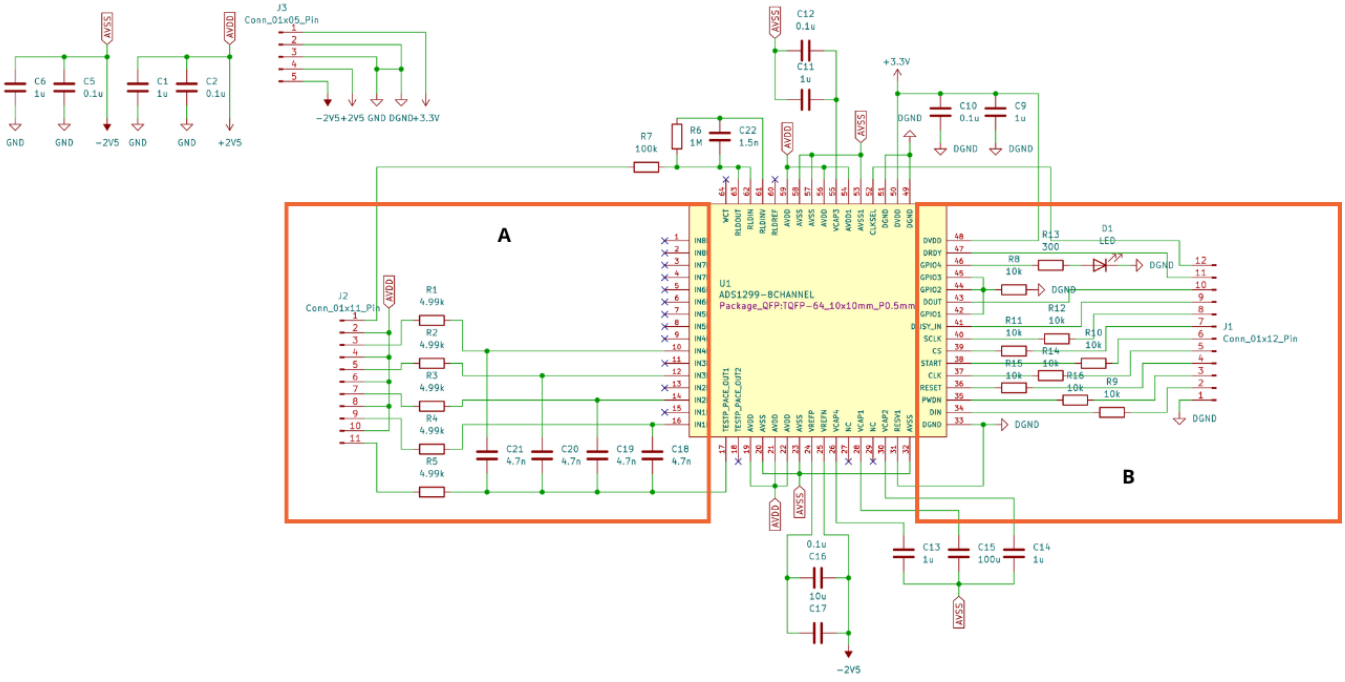


Fig. 4. System circuit schematic.

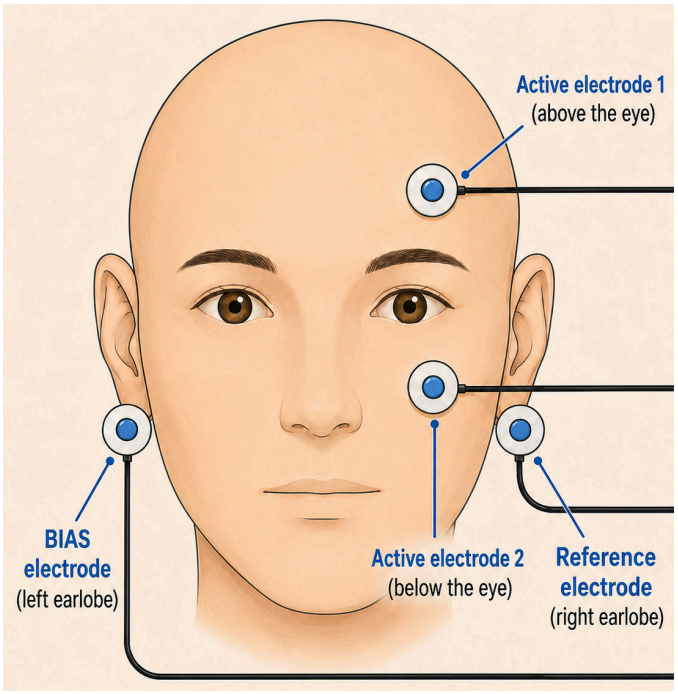


Fig. 5. Positioning of the electrodes.

- True Negative (TN): When the model predicts the negative class and the actual class is negative.
- False Positive (FP): When the model predicts a positive class but the actual class is negative.
- False Negative (FN): When the model predicts the negative class and the actual class is positive.

The SVM was selected because it has demonstrated good performance in EOG classification tasks even with relatively small datasets. In addition, it has parameters that allow control over the balance between the separation margin and classification errors. This test was essential to verify whether the developed solution is capable of extracting relevant information from the acquired signals.

Before training the model, the data were normalized using the StandardScaler function in Python. And after that, for the SVM model, the following settings were used: RBF kernel, a C value of 20, gamma set to 'scale', class_weight set to 'balanced', and random_state set to 42. The RBF kernel enables the separation of data that is not linearly separable. The C value was chosen to strike a balance between the width of the separation margin and the tolerance for misclassifications, while setting gamma to 'scale' allows this parameter to be calculated automatically. Additionally, 'balanced' class_weight was used to balance the classes. The training was conducted using 25% for testing and 75% for training.

It is also important to highlight that, for SVM training, the following metrics had to be extracted from each signal window: mean, minimum value, peak-to-peak amplitude, skewness, kurtosis, maximum absolute difference, and median crossing rate. **All the implemented code can be accessed at [5]**

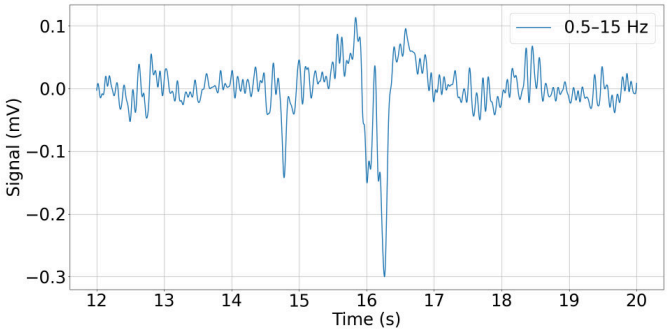


Fig. 6. EOG signal containing a double blink.

IV. Results and Discussion

Based on the developed hardware, it was possible to acquire EOG signals. However, it was necessary to adapt the electrode montage, since the referential configuration did not prove suitable for the application. Therefore, to overcome this issue, a bipolar montage was used by subtracting the signals from active electrodes 1 and 2 using software, rather than performing the subtraction analogically through the circuit.

To evaluate the developed hardware, the SNR was calculated. The system presented an SNR of -4.54 dB, indicating the presence of considerable noise levels in the acquired signal, possibly due to the prototype nature of the project. This value indicates that the noise power is greater than that of the useful signal. This noise caused some distortion in the signal, even after the application of digital filtering. However, the essential characteristics for double-blink detection were preserved, such as the presence of two consecutive peaks or valleys, as shown in Figure 6.

The model was trained and tested using data from two participants. Initially, the model was trained using data from participant 1 and tested using data from participant 2. The procedure was then reversed, with participant 2 data used for training and participant 1 data used for testing. This approach allowed the model's ability to generalize to individuals whose data were not included in the training set to be evaluated. Additionally, an experiment was conducted in which data from both participants were used for both training and testing the model.

The SVM machine learning model achieved the performance results shown in Table I, II e III. From these results, the following metrics were extracted: accuracy, precision, and recall, as presented in Table IV, V e VI.

TABLE I
Confusion Matrix for Training with Participant 1's Data and Testing with Participant 2's Data

	Predicted Class	
Actual Class	**No blink**	**Double blink**
No blink	250	12
Double blink	10	18

TABLE II
CONFUSION MATRIX FOR TRAINING WITH PARTICIPANT 2'S DATA AND TESTING WITH PARTICIPANT 1'S DATA

	Predicted Class	
Actual Class	**No blink**	**Double blink**
No blink	252	11
Double blink	7	20

TABLE III
CONFUSION MATRIX FOR TRAINING AND TESTING WITH DATA FROM PARTICIPANTS 1 AND 2

	Predicted Class	
Actual Class	**No blink**	**Double blink**
No blink	127	4
Double blink	0	14

According to Tables I and II, the model showed limited performance in generalizing double-blink detection across different participants, indicating the need to expand the dataset with signals acquired from a larger number of individuals. The corresponding performance metrics are presented in Tables IV and V. In contrast, Tables III and VI show that the model achieved better classification performance when data from both participants were combined and subsequently divided into training and test sets, reaching a precision of 78%. However, this experiment does not assess generalization to unseen participants, since data from both individuals were represented in the dataset used for training and testing. Therefore, these results highlight the need for future studies involving a larger and more representative number of participants to enable a more robust evaluation of the generalization capability of the developed system.

Furthermore, the model was deployed on a Raspberry Pi 4 and achieved a prediction execution time of 7.11 ms. The low processing time highlights the feasibility of its use in real-time applications, especially in embedded systems aimed at human-machine interaction.

A. Limitations

Despite the positive results, the project presents some limitations, such as the reduced number of participants used in the system tests, who, moreover, did not have physical disabilities, as well as possible limitations related to the developed circuit,

TABLE IV
SVM MODEL PERFORMANCE METRICS REGARDING TABLE I

Model	Accuracy	Precision	Recall
SVM	0.92	0.60	0.64

TABLE V
SVM MODEL PERFORMANCE METRICS REGARDING TABLE II

Model	Accuracy	Precision	Recall
SVM	0.93	0.64	0.74

TABLE VI
SVM MODEL PERFORMANCE METRICS REGARDING TABLE III

Model	Accuracy	Precision	Recall
SVM	0.97	0.78	1

including the electrode montage configuration and the choice of a symmetrical power supply.

The reduced number of participants prevents the generalization of the obtained results. Additionally, the participants do not directly represent the target audience of the project, which may present specific difficulties when using the system that were not observed in the evaluated individuals.

The ADS1299-4 chip enables differential signal acquisition through the positive and negative inputs of its channels. This feature could have been explored in the project through a bipolar montage implemented directly in the hardware, instead of performing signal subtraction in software. Such an approach could contribute to the rejection of common-mode noise and simplify the signal processing stage.

In addition to this hardware-level bipolar montage, other measures could further reduce the noise levels observed in the system. The absence of a dedicated notch filter for rejecting power-line frequency interference (50/60 Hz) represents another opportunity for improvement, through an analog notch stage. Furthermore, although the analog and digital ground planes were interconnected through a single point at connector J3, improvements in board routing, with greater isolation between digital and analog traces, could further reduce the coupling between the two sections of the circuit.

Moreover, the circuit was powered using a symmetrical power supply. However, an asymmetrical power supply could be investigated in future versions, provided that the appropriate biasing conditions and input range of the ADS1299-4 are respected. This alternative could simplify the solution by reducing the number of power supplies required for circuit operation.

V. CONCLUSION

This work presented the design, implementation, and preliminary validation of an open-source EOG acquisition platform for assistive human-machine interfaces. The main contribution of this work lies in the open-source hardware itself — the detailed schematics, PCB layout, and signal acquisition methodology — which addresses a recognized gap in the literature regarding reproducibility of custom EOG acquisition circuits and provides a foundation upon which other researchers can build, adapt, and improve.*

The SVM-based classification results, achieving a precision of 78% under the most favorable experimental condition, should be interpreted as an initial indication that the developed hardware is capable of capturing signal features relevant to double-blink detection, rather than as conclusive evidence of the system's readiness for real-world

assistive applications. The limited number of participants, the absence of individuals with motor disabilities among the evaluated subjects. Similarly, the negative SNR obtained indicates that noise levels remain a considerable challenge for the current hardware revision, even though the essential morphological features of double-blink events were still discernible after filtering.*

Future work should therefore prioritize expanding the dataset to include a larger and more representative population, including individuals with motor disabilities; implementing hardware-level improvements to reduce noise, such as a differential bipolar montage, dedicated notch filtering, and improved analog-digital isolation; and conducting more rigorous cross-subject validation to better assess the system's generalization capability. While the results obtained so far are encouraging as a proof of concept, further development and more extensive validation are necessary before the platform can be considered a robust or reliable solution for assistive device interaction.*

REFERENCES

[1] World Health Organization and World Bank, *World Report on Disability*. Geneva, Switzerland: World Health Organization, 2011. [Online]. Available: https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/world-report-on-disability. Accessed: Jun. 9, 2026.

[2] L. Tao, H. Huang, C. Chen, L. Feijs, J. Hu, and W. Chen, "Review of electrooculography-based human-computer interaction: recent technologies, challenges and future trends," *Connected Health and Telemedicine*, vol. 2, no. 3, Sep. 2023, doi: 10.20517/chatmed.2023.05.

[3] F. Wang, X. Li, and J. Pan, "A human-machine interface based on an EOG and a gyroscope for humanoid robot control and its application to home services," *Journal of Healthcare Engineering*, vol. 2022, Art. no. 1650387, pp. 1–14, 2022, doi: 10.1155/2022/1650387.

[4] S. Ban, Y. J. Lee, S. Kwon, Y.-S. Kim, J. W. Chang, J.-H. Kim, and W.-H. Yeo, "Soft wireless headband bioelectronics and electrooculography for persistent human–machine interfaces," *ACS Applied Electronic Materials*, vol. 5, no. 2, pp. 877–886, 2023, doi: 10.1021/acsaelm.2c01436.

[5] M. F. Marcelino and P. do C. Sampaio "EOG-Project", GitHub repository, 2026. [Online]. Available: https://github.com/FerrazMariana/EOG-Project. Accessed: Jun. 24, 2026.

[6] R. Zhang, S. He, X. Yang, X. Wang, K. Li, Q. Huang, Z. Yu, X. Zhang, D. Tang, and Y. Li, "An EOG-based human–machine interface to control a smart home environment for patients with severe spinal cord injuries," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 89–100, Jan. 2019, doi: 10.1109/TBME.2018.2834555.

[7] A. J. Molina-Cantero, C. Lebrato-Vázquez, M. Merino-Monge, R. Quesada-Tabares, J. A. Castro-García, and I. M. Gómez-González, "Communication technologies based on voluntary blinks: assessment and design," *IEEE Access*, vol. 7, pp. 70770–70798, 2019, doi: 10.1109/ACCESS.2019.2919324.

[8] Texas Instruments, *ADS1299, ADS1299-4, ADS1299-6: Low-noise, 4-, 6-, 8-channel, 24-bit, analog-to-digital converter for EEG and biopotential measurements*, Data sheet SBAS499C, revised Jan. 2017. [Online]. Available: https://www.ti.com/lit/ds/symlink/ads1299.pdf. Accessed: Jun. 11, 2026.

[9] F. Simard, "RaspberryPiADS1299", GitHub repository, 2017. [Online]. Available: https://github.com/wjcroft/RaspberryPiADS1299. Accessed: Dec. 15, 2025.

Phonocardiographic Signal Analysis for Elevated Blood Pressure Classification

Gustavo Esmanhotto Bareta
Academic Department of Informatics (DAINF)
Federal University of Technology — Paraná (UTFPR)
Curitiba, Paraná, Brazil
bareta@alunos.utfpr.edu.br

Douglas Henz
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology — Paraná (UTFPR)
Curitiba, Paraná, Brazil
douglashenz@alunos.utfpr.edu.br

Wemerson Delcio Parreira
Faculty of Electrical Engineering, Polytechnic School
Pontifical Catholic University of Campinas (PUC-Campinas)
Campinas, São Paulo, Brazil
wemerson.delcio@puc-campinas.edu.br

Renata Coelho Borges
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology — Paraná (UTFPR)
Curitiba, Paraná, Brazil
renatacoelho@utfpr.edu.br

José Jair A. Mendes Júnior
Graduate Program in Electrical and Industrial Informatics Engineering (CPGEI), Federal University of Technology — Paraná (UTFPR), Curitiba, Paraná, Brazil
jjjunior@utfpr.edu.br

Fábio Pires Itturriet
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology — Paraná (UTFPR)
Curitiba, Paraná, Brazil
fabioitturriet@utfpr.edu.br

***Abstract*—This study investigates the potential of phonocardiogram (PCG) signals to support the classification of elevated blood pressure using supervised machine learning. Heart sound recordings from 78 participants were segmented into S1, systole, S2, and diastole using a Logistic Regression–Hidden Semi-Markov Model (LR-HSMM). Temporal, spectral, and statistical features were extracted with the TSFRESH library, followed by Correlation-based Feature Selection (CFS) to identify the most relevant features. For the binary classification task, participants whose reference blood pressure measurements fell within the pre-hypertensive or hypertensive ranges were grouped into an elevated blood pressure class and compared with normotensive participants. Multiple classifiers were evaluated using three PCG representations: the complete recording, fused segment-specific features, and individual cardiac cycle segments. Each representation was combined with demographic variables comprising age, weight, height, and sex. Among the evaluated segment-specific configurations, S2 features combined with demographic variables yielded the highest observed F1-score of 85.25% with Logistic Regression. These findings suggest that the combination of S2-derived and demographic features contains information associated with elevated blood pressure in the evaluated dataset. Despite the limited sample size and single-dataset evaluation, these exploratory results support further investigation of PCG-based methods for potential use in non-invasive cardiovascular risk screening.**

***Index Terms*—Elevated blood pressure, phonocardiography, segment-specific analysis, machine learning**

I. Introduction

Hypertension (HTN) remains a major global public health challenge and a leading modifiable risk factor for cardiovascular diseases (CVDs) [1]. According to the World Health Organization (WHO), CVDs account for approximately 32% of deaths worldwide [2]. Recent estimates indicate that approximately 1.4 billion adults were living with hypertension in 2024, while around 44% were unaware of their condition, partly because hypertension is frequently asymptomatic [3].

Early identification of altered blood pressure enables timely clinical follow-up and lifestyle interventions that may reduce the risk of cardiovascular complications [4]. Conventional blood pressure assessment relies primarily on cuff-based measurements performed in clinical settings, as well as on ambulatory blood pressure monitoring (ABPM) and home blood pressure monitoring (HBPM) [5]. Although these methods are clinically established, they provide intermittent measurements and may cause discomfort, particularly during repeated or nocturnal assessments. Consequently, blood pressure variations occurring between measurements may remain unobserved [6].

The 2025 Brazilian Hypertension Guideline classifies office blood pressure measurements with systolic blood pressure (SBP) between 120 and 139 mmHg and/or diastolic blood pressure (DBP) between 80 and 89 mmHg as pre-hypertension, whereas measurements with SBP $\geq$ 140 mmHg and/or DBP $\geq$ 90 mmHg fall within the hypertensive range [7]. The guideline also emphasizes out-of-office blood pressure assessment and identifies cuffless and wearable technologies as promising approaches for more frequent and comfortable monitoring.

Within this context, digital phonocardiography (PCG) emerges as a promising complementary modality for non-invasive cardiovascular monitoring. Unlike oscillometric methods, PCG acquisition does not require cuff inflation or arterial

compression, enabling heart sounds to be recorded with minimal physical interference. This characteristic may facilitate repeated measurements and future integration into wearable devices, with potentially less disruption to daily activities and sleep [8].

The physiological rationale for investigating this technology lies in the relationship between heart sounds and cardiovascular hemodynamics. The cardiac cycle is divided into two main phases: systole and diastole. Systole corresponds to ventricular contraction and begins near the first heart sound (S1), which is primarily associated with the closure of the mitral and tricuspid valves. Diastole represents ventricular relaxation and begins near the second heart sound (S2), which is primarily associated with the closure of the aortic and pulmonary valves [9], [10]. The literature suggests that the timing, intensity, and spectral characteristics of heart sounds may be associated with arterial pressure variations and other hemodynamic changes [11], [12].

The complex temporal and spectral information contained in PCG recordings represents a promising opportunity for the development of machine learning techniques for blood pressure estimation and classification. These techniques can identify acoustic patterns associated with blood pressure variations and support the development of non-invasive screening approaches [13]–[15]. However, the discriminative contribution of individual cardiac cycle segments to elevated blood pressure classification remains insufficiently explored.

The present study compares complete, fused, and segment-specific PCG representations for elevated blood pressure classification using a public dataset published in 2025 as an exploratory proof of concept. Automatic cardiac cycle segmentation is used to isolate and individually analyze S1, systole, S2, and diastole, enabling comparison of classification performance across segment-specific representations. A large set of temporal, spectral, and statistical features is subsequently extracted to train supervised machine learning models. For the binary classification task, participants whose reference blood pressure measurements fell within the pre-hypertensive or hypertensive ranges were grouped into a single target class, hereafter referred to as elevated blood pressure, and compared with normotensive participants. Given the limited sample size and controlled acquisition conditions, this study is intended as an exploratory step toward future non-invasive approaches to blood pressure screening.

II. Related Works

In recent years, several studies have investigated the relationship between phonocardiographic (PCG) signals and blood pressure, seeking less invasive alternatives for continuous or intermittent blood pressure estimation. Overall, these works suggest that acoustic information from the cardiac cycle contains relevant physiological patterns associated with hemodynamic variations.

One of the early relevant approaches was proposed in [13], where only the second heart sound (S2) was used for blood pressure estimation. The method extracts frequency-domain features using FFT and employs an SVM regression model. Although moderate correlations between acoustic features and blood pressure were reported, the study did not compare S2 with other cardiac cycle segments, preventing an assessment of their relative contributions.

In [14], the authors adopt a deep learning-based approach combining Mel-frequency cepstral coefficients (MFCCs), deltas, and delta-deltas with demographic information. The architecture employs convolutional neural networks with Squeeze-and-Excitation attention blocks, achieving improved performance in systolic and diastolic blood pressure regression. However, as is common in deep learning-based approaches, the physiological interpretability of the learned representations remains limited.

To exploit complementary mechanical information, a multimodal approach combining PCG and ballistocardiogram (BCG) signals was proposed in [16], with temporal intervals extracted between mechanical cardiac events. These intervals showed strong correlations with systolic blood pressure and were used as input to Random Forest-based regression models. Despite leveraging multiple modalities, the method still relies on handcrafted interval definitions and does not investigate the contribution of individual cardiac cycle segments.

More recently, a real-time blood pressure monitoring framework based on nonlinear relationships between heart sound features and blood pressure was proposed in [17]. The method employs multifeature heart sound analysis combined with an improved Whale Optimization Algorithm (T-WOA) to estimate systolic and diastolic blood pressure continuously during surgery. Clinical validation demonstrated promising accuracy for intraoperative monitoring. Nevertheless, the study focuses on the regression of absolute blood pressure values and does not investigate the discriminative contribution of individual cardiac cycle segments or compare complete, fused, and segment-specific PCG representations for elevated blood pressure classification.

Unlike the aforementioned studies, the present work does not aim to estimate absolute blood pressure values. Instead, it frames blood pressure assessment as a binary classification task, distinguishing normotensive participants from those with elevated blood pressure. Moreover, the proposed framework automatically segments the cardiac cycle into S1, systole, S2, and diastole, enabling comparison of the segment-specific configurations. Complete, fused, and segment-specific PCG representations are then evaluated using multiple supervised machine learning models.

Table I summarizes and compares the main characteristics of the reviewed works with the proposed approach.

III. Materials and Methods

The proposed framework consists of six sequential stages: (i) PCG dataset acquisition, (ii) signal preprocessing, (iii) cardiac cycle segmentation, (iv) signal reconstruction, (v) feature extraction and selection, and (vi) machine learning-based elevated blood pressure classification. An overview of the complete pipeline is presented in Figure 1, followed by a detailed description of each stage.

TABLE I
COMPARISON OF RELATED STUDIES AND THE PROPOSED APPROACH.

Work/Year	Signals	Features	Model	Task
[13]/2015	PCG (S2 only)	Frequency-domain features (FFT of S2)	SVM regression	Continuous estimation of SBP, DBP, MAP
[14]/2025	PCG + metadata	MFCCs, deltas, and delta-deltas (feature maps)	CNN with Squeeze-and-Excitation blocks	SBP and DBP regression
[16]/2022	PCG + BCG	Temporal intervals between cardiac events (e.g., S1–J, J–S2)	Random Forest regression	SBP estimation
[17]/2026	PCG	Multifeature PCG analysis with nonlinear separation parameters	T-WOA optimized regression model	Real-time estimation of SBP, DBP, and blood pressure variations during surgery
This study	PCG + metadata	Time-, frequency-, statistical-, and derived-domain features from cardiac cycle segments (S1, systole, S2, diastole)	Supervised models (LR, SVM, RF, XGBoost, GB, KNN, MLP)	Binary classification (normotensive vs. elevated blood pressure)

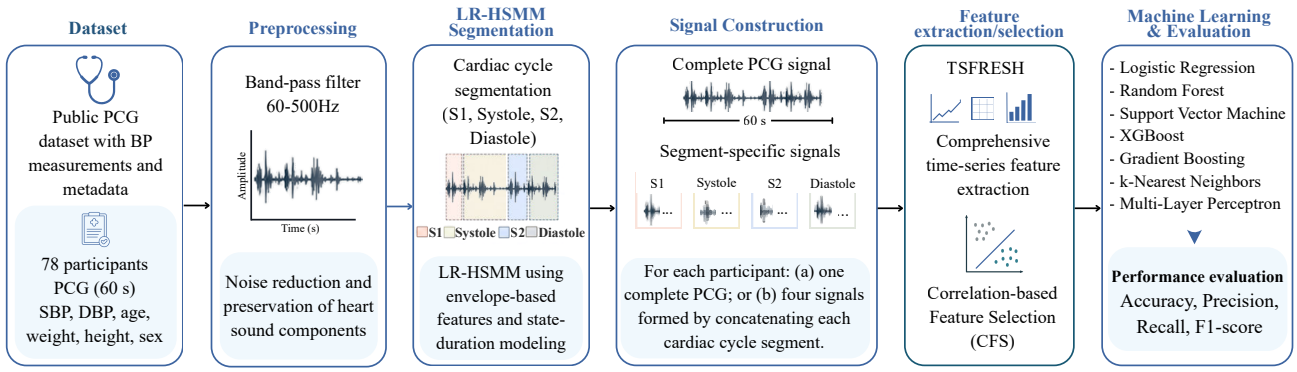


Fig. 1. Pipeline of the proposed PCG-based elevated blood pressure classification framework.

A. Dataset and Class Definition

This study used the publicly available *Phonocardiogram Signal Dataset Paired with Blood Pressure Measurements*, providing one 60-s PCG recording and one retained SBP/DBP pair for each of 78 participants under controlled conditions [18]. BP and PCG were acquired simultaneously, with acquisition repeated only when anomalies were detected. The dataset also includes pulse rate, age, weight, height, and sex. Participants were 21 to 66 years old. No demographic values were missing, and no imputation was required.

Before data acquisition, participants remained seated for 5 min in a quiet, temperature-controlled room. Blood pressure was measured using a NISSEI DSK-1031 automated monitor with the cuff positioned on the upper left arm. Simultaneously, PCG signals were recorded for 60 s using a ThinkLabs One digital stethoscope positioned at the pulmonic auscultation area, located at the second left intercostal space near the sternum. The stethoscope had a frequency response of 20–2000 Hz.

For the binary classification task, participants were categorized using the reference SBP and DBP measurements available in the dataset. Those with SBP below 120 mmHg and DBP below 80 mmHg were assigned to the normotensive class. Participants in the pre-hypertensive range (SBP 120–139 mmHg and/or DBP 80–89 mmHg) and hypertensive range (SBP $\geq$ 140 mmHg and/or DBP $\geq$ 90 mmHg) were combined into a single elevated blood pressure class to reduce class fragmentation in the limited dataset. The final dataset comprised 38 normotensive participants and 40 participants in the elevated blood pressure class. These labels represent blood pressure ranges derived from the available reference measurements and should not be interpreted as independent clinical diagnoses of hypertension.

B. Signal Preprocessing and Segment Construction

For each participant, the 60-s band-pass-filtered PCG recording provided in the dataset was analyzed. The dataset authors applied a band-pass filter between 60 and 500 Hz to attenuate background noise and retain the principal frequency components of the first (S1) and second (S2) heart sounds [18].

Cardiac cycle segmentation was performed using the Logistic Regression–Hidden Semi-Markov Model (LR-HSMM) proposed in [19]. This method combines the temporal modeling capabilities of a HSMM, which explicitly accounts for the expected duration of each cardiac state (S1, systole, S2, and diastole), with logistic regression to estimate the emission probabilities of each hidden state from features extracted from the PCG signal. Unlike conventional HMM-based approaches, which typically model state emissions using Gaussian distributions, logistic regression provides discriminative estimates for distinguishing cardiac states.

Figure 2 illustrates an example of a PCG signal segmented using the LR-HSMM algorithm. The method employs four envelope-based features — homomorphic, Hilbert, wavelet, and Power Spectral Density (PSD) envelopes — to characterize the signal. The most likely sequence of cardiac states is then estimated using an extended version of the Viterbi algorithm capable of handling recordings that do not begin

precisely at the start of a cardiac cycle. The combination of temporal modeling, explicit state-duration information, and discriminative emission estimation supports the segmentation of S1 and S2 in noisy recordings, as reported in the original study.

Two signal representations were investigated. In the first approach, the complete PCG recording was used without modification, preserving the natural temporal structure of the cardiac cycle. The second approach aimed to evaluate whether individual cardiac cycle segments provide distinct information for the classification of elevated blood pressure. After segmenting each recording into S1, systole, S2, and diastole, four segment-specific signals were reconstructed for each participant by sequentially concatenating all occurrences of the same segment type. Consequently, each participant was represented by four additional recordings: one containing only S1 segments, one containing only systolic segments, one containing only S2 segments, and one containing only diastolic segments. This strategy enabled the independent evaluation of each cardiac cycle segment and the investigation of whether specific segments contain more relevant information for elevated blood pressure classification.

C. Feature Extraction

Features were extracted using TSFRESH (Time Series Feature Extraction Based on Scalable Hypothesis Tests) [20], which computes statistical, temporal, spectral, autocorrelation, entropy, wavelet, and peak-related descriptors. In total, 778 PCG-derived features were independently extracted from the complete recording and from each segment-specific signal reconstructed from S1, systole, S2, and diastole.

In an additional feature-level fusion experiment, the feature vectors extracted from S1, systole, S2, and diastole were concatenated before feature selection and classifier training. For all evaluated representations, the PCG-derived features were combined with age, weight, height, and sex.

D. Feature Selection

Correlation-based Feature Selection (CFS) was used to reduce dimensionality and redundancy in the feature vectors [21]. As a filter-based method, CFS favors subsets containing features that are strongly associated with the target class while exhibiting low interfeature correlation. A best-first search strategy was employed to identify the subset with the highest merit.

CFS was applied independently to the complete-PCG, segment-specific, and feature-fusion representations. Within each cross-validation iteration, CFS was fitted using only the training folds, and the resulting feature subset was subsequently applied to the held-out test fold. This procedure prevented information from the test fold from influencing feature selection.

E. Model Training

Machine learning methods are well suited to PCG analysis because they can capture complex temporal, spectral, and nonlinear patterns in heart sounds. Accordingly, this study evaluated a set of supervised classifiers commonly applied to biomedical signal classification.

Logistic Regression (LR) is a probabilistic linear classifier widely used for binary classification tasks. Due to its interpretability, computational efficiency, and competitive performance in problems with approximately linear decision boundaries, it is commonly employed as a baseline model in biomedical signal classification [22].

Random Forest (RF) is an algorithm based on multiple decision trees combined through the bagging technique. This strategy reduces unwanted variance and helps mitigate overfitting, while also enabling the capture of nonlinear relationships among variables [23]. Its stability and good generalization ability make it widely used in biomedical tasks.

Support Vector Machine (SVM) models with radial basis function (RBF) and polynomial (Poly) kernels were included due to their ability to project data into higher-dimensional spaces, enabling the definition of nonlinear decision boundaries [24]. SVMs are known for their consistent performance on high-dimensional datasets and for their ability to maximize the margin between classes, a property frequently exploited in physiological signal analysis.

XGBoost is one of the most efficient implementations of gradient boosting methods based on decision trees. It builds models sequentially, where each new tree corrects the residual errors of the previous ones, enabling the capture of complex patterns and feature interactions [25].

K-Nearest Neighbors (KNN) classifies a sample based on the majority class among its nearest neighbors. Although simple, the method often performs well in scenarios where data exhibits well-defined cluster structures [26].

Gradient Boosting (GB) also follows the boosting principle, combining weak learners iteratively to progressively reduce error [27]. Like XGBoost, it performs well on tabular datasets and is effective in capturing complex relationships among features derived from physiological signals.

Finally, the Multilayer Perceptron (MLP) is an artificial neural network architecture composed of multiple fully connected layers [28]. Its flexibility allows it to model highly nonlinear relationships, making it suitable for datasets with a large number of extracted features, such as biomedical signals.

Models were implemented using scikit-learn, except XGBoost, which used the XGBoost library. Default settings were retained, except for LR and MLP, whose maximum iteration limits were set to 1,000 and 500, respectively, and for the SVM models, which used RBF and third-degree polynomial kernels.

F. Evaluation Metrics

The evaluation metrics used in this work were accuracy, precision, recall, F1-score, and confusion matrices, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The elevated blood pressure class was considered the positive class. In this screening-oriented classification task, false-negative predictions are particularly

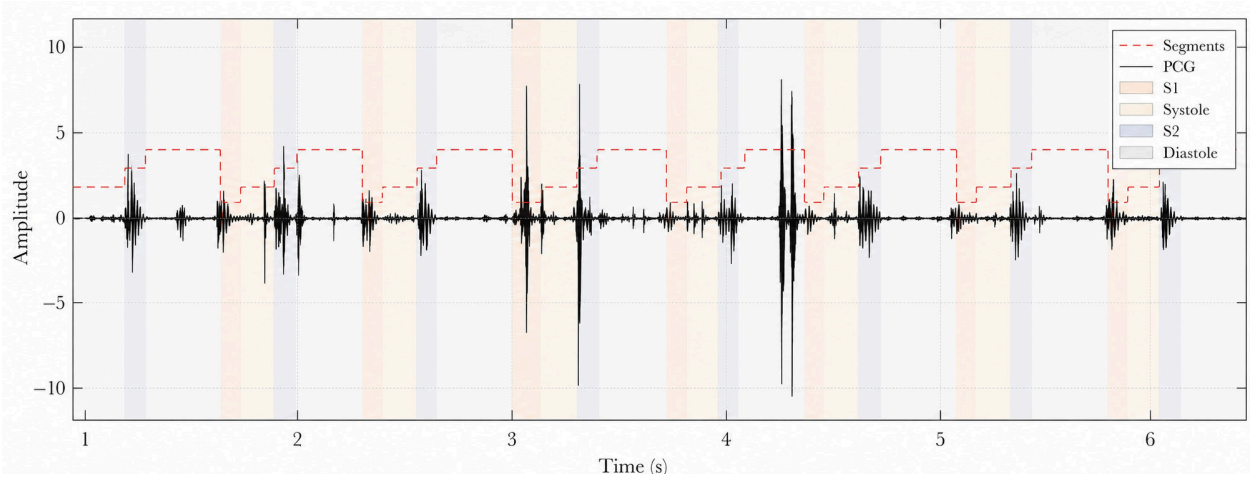


Fig. 2. Example of a phonocardiogram segmentation using the algorithm proposed in [19].

relevant because they correspond to participants with elevated blood pressure measurements incorrectly classified as normotensive. Therefore, precision, recall, and F1-score complement accuracy by providing a more detailed assessment of classification performance. Confusion matrices were normalized by the total number of predictions.

Model performance was evaluated using stratified 10-fold cross-validation. The participants were divided into ten approximately equal folds while preserving the proportion of normotensive and elevated blood pressure samples. In each iteration, nine folds were used for training and the remaining fold was used for testing. This process was repeated until each fold had been used once as the test set. The reported metrics correspond to the average values obtained across the ten folds.

IV. Results

Three experimental scenarios were evaluated. In the first scenario, features were extracted from the complete 60-s PCG recording without cardiac cycle segmentation, establishing the baseline representation. In the second scenario, the feature vectors extracted from S1, systole, S2, and diastole were concatenated into a single representation before feature selection and classification. In the third scenario, the features extracted from each cardiac cycle segment were evaluated independently, resulting in four segment-specific classification experiments. In all three scenarios, the PCG-derived features were combined with age, weight, height, and sex.

A. Complete-PCG Baseline

To establish a baseline performance, full PCG signals were initially used without segmentation into specific cardiac cycle segments. The processing followed the proposed pipeline, starting with feature extraction using the tsfresh package, which generates a large and diverse set of time-domain and spectral descriptors. This process resulted in 778 features extracted from the audio signals.

These features were then concatenated with demographic information provided by the dataset authors: age, weight, height, and sex, resulting in a total of 782 initial features. The inclusion of these physiological attributes is justified by their known relationship with hemodynamic parameters, potentially complementing patterns extracted directly from the cardiac signal.

In the next stage, feature selection was applied using the Correlation-based Feature Selection (CFS) method, whose objective is to identify feature subsets with high correlation to the target class and low redundancy among features. After this process, the feature set was reduced to 18 features, representing less than 3% of the original total, which substantially reduces the dimensionality of the problem and mitigates the risk of overfitting.

These selected features mainly include information related to participant weight, spectral coefficients derived from the Fourier Transform (phase and imaginary components), autoregressive model-derived measures, temporal statistics such as the median, autocorrelation-based attributes, temporal interaction features, and Continuous Wavelet Transform coefficients at different scales. This set suggests that both cardiac signal characteristics and participant-specific information have discriminative potential for blood pressure-related classification.

With the final feature set defined, model training was performed. All classifiers were trained using the same preprocessed data, ensuring comparability across results. Performance metrics indicated that the best performance in this stage was achieved by the Logistic Regression model, which obtained an F1-score of 62.29%, as shown in Table II. The corresponding confusion matrix is presented in Figure 3.

B. Feature-Level Fusion

In the second scenario, the feature vectors independently extracted from S1, systole, S2, and diastole were concatenated into a single representation. Since TSFRESH extracted 778 features from each cardiac cycle segment, the concatenated representation contained 3,112 PCG-derived features. Age, weight, height, and sex were subsequently added, resulting in 3,116 initial attributes per participant.

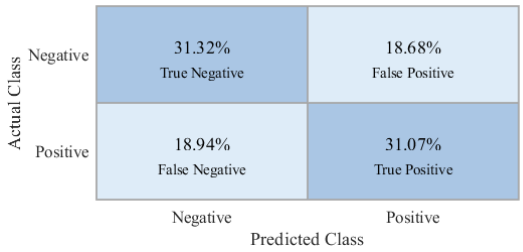


Fig. 3. Confusion matrix of the LR model without segmentation.

TABLE II
MODEL PERFORMANCE USING THE COMPLETE PCG RECORDING.

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Logistic Regression	62.39	62.29	62.45	62.13
Random Forest	59.77	58.52	59.89	59.15
SVM (RBF)	58.35	57.89	58.53	57.26
SVM (Poly)	52.06	52.02	52.06	51.99
XGBoost	53.15	53.41	53.11	53.71
KNN	51.03	50.79	51.04	50.54
GB	51.05	50.72	51.06	50.39
MLP	59.44	58.96	59.66	58.28
Average (μ)	55.75	55.65	55.82	55.48

CFS was applied to the expanded feature representation to reduce dimensionality and feature redundancy, retaining 42 attributes for classification. Table III presents the performance obtained by the evaluated classifiers.

Logistic Regression achieved the highest performance in this scenario, with an accuracy of 78.11%, an F1-score of 78.00%, a precision of 78.41%, and a recall of 77.59%. The MLP obtained the second-highest F1-score, reaching 77.14%, followed by SVM-RBF with 73.28%. The mean F1-score across all classifiers was 70.33%.

The normalized confusion matrix of the Logistic Regression model is presented in Fig. 4. True-negative and true-positive predictions accounted for 39.30% and 38.80% of all predictions, respectively, whereas false-positive and false-negative predictions accounted for 10.70% and 11.20%. Compared with the complete-PCG baseline, feature-level fusion increased the highest F1-score from 62.29% to 78.00%, representing a 15.71 percentage-point improvement.

TABLE III
MODEL PERFORMANCE USING CONCATENATED SEGMENT-SPECIFIC FEATURES.

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Logistic Regression	78.11	78.00	78.41	77.59
Random Forest	67.88	67.77	68.00	67.54
SVM (RBF)	73.38	73.28	73.56	73.00
SVM (Poly)	69.75	69.62	69.91	69.33
XGBoost	62.74	63.16	62.45	63.89
KNN	73.45	73.07	74.12	72.05
GB	60.72	60.63	60.76	60.51
MLP	77.10	77.14	77.01	77.27
Average (μ)	70.39	70.33	70.40	70.15

C. Segment-Specific Evaluation

In the third scenario, the signals reconstructed from S1, systole, S2, and diastole were evaluated independently. For each cardiac cycle segment, TSFRESH extracted 778 PCG-derived features, which were combined with age, weight, height, and sex, resulting in 782 initial attributes per representation. CFS was then independently applied to each segment-specific feature set.

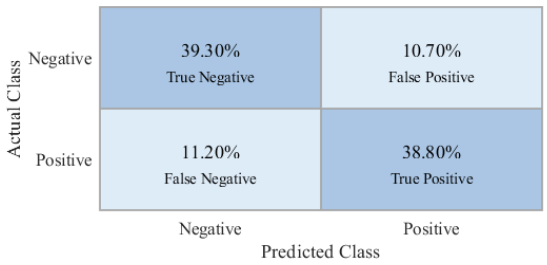


Fig. 4. Confusion matrix of the LR classifier using concatenated features from the S1, systole, S2, and diastole segments.

Tables IV–VII present the classification results obtained for S1, systole, S2, and diastole, respectively. All segment-specific representations achieved higher average F1-scores than the complete-PCG baseline. The average F1-scores were 64.63% for S1, 71.32% for systole, 77.01% for S2, and 64.91% for diastole, compared with 55.58% for the complete recording.

For S1, Logistic Regression achieved the highest F1-score of 69.78%. For systole, the best result was obtained by the MLP, with an F1-score of 77.79%. For diastole, the MLP achieved the highest F1-score of 70.03%. The configuration combining S2 features with demographic variables yielded the highest observed classification performance. Logistic Regression achieved an accuracy of 85.46%, an F1-score of 85.25%, a precision of 86.48%, and a recall of 84.06%. The MLP and SVM-RBF also achieved high performance using S2, with F1-scores of 84.45% and 82.00%, respectively. Overall, the best segment-specific result exceeded the complete-PCG baseline by 22.96 percentage points and the feature-level fusion result by 7.25 percentage points in terms of F1-score.

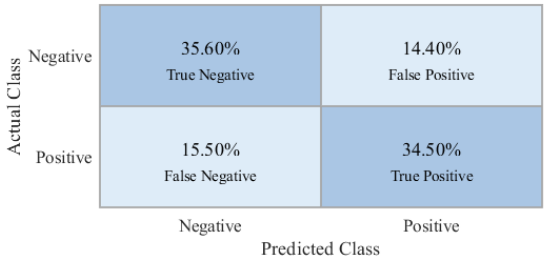


Fig. 5. Confusion matrix for the LR model using the S1 segment.

The confusion matrices of the best-performing classifier for each cardiac cycle segment are shown in Figs. 5–8. The configuration combining S2 features with demographic variables showed the highest concentration of predictions along the main diagonal, with lower false-positive and false-negative proportions than the other segment-specific representations. The systolic representation also produced a comparatively

TABLE IV
MODEL PERFORMANCE USING THE S1 SEGMENT

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Logistic Regression	70.11	69.78	70.56	69.02
Random Forest	67.62	67.30	67.98	66.63
SVM (RBF)	66.80	66.39	67.23	65.58
SVM (Poly)	61.96	61.95	61.96	61.93
XGBoost	61.99	61.83	62.09	61.57
KNN	61.35	60.99	61.56	60.43
GB	60.14	59.46	60.49	58.46
MLP	69.28	69.30	69.25	69.34
Average (μ)	64.91	64.63	65.39	64.37

TABLE V
MODEL PERFORMANCE USING THE SYSTOLIC SEGMENT

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Logistic Regression	76.46	76.29	76.85	75.75
Random Forest	72.90	72.68	73.27	72.10
SVM (RBF)	73.60	73.31	74.13	72.51
SVM (Poly)	69.00	68.77	69.28	68.28
XGBoost	68.39	68.10	68.73	67.49
KNN	69.47	69.27	69.74	68.80
GB	64.46	64.36	64.54	64.18
MLP	77.96	77.79	78.39	77.21
Average (μ)	71.03	70.82	71.36	70.29

TABLE VI
MODEL PERFORMANCE USING THE S2 SEGMENT

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Logistic Regression	85.46	85.25	86.48	84.06
Random Forest	75.10	75.09	75.11	75.07
SVM (RBF)	82.20	82.00	82.93	81.09
SVM (Poly)	75.77	75.68	75.97	75.39
XGBoost	70.12	69.63	70.78	68.52
KNN	75.10	75.01	75.28	74.74
GB	69.16	68.97	69.41	68.53
MLP	84.51	84.45	84.78	84.12
Average (μ)	77.93	77.76	78.22	76.94

TABLE VII
MODEL PERFORMANCE USING THE DIASTOLIC SEGMENT

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Logistic Regression	69.50	69.21	69.89	68.54
Random Forest	68.83	68.46	69.28	67.66
SVM (RBF)	67.68	67.29	68.10	66.50
SVM (Poly)	60.93	60.96	60.92	61.00
XGBoost	61.46	61.30	61.56	61.05
KNN	62.38	61.97	62.65	61.31
GB	60.11	60.06	60.13	59.99
MLP	69.82	70.03	69.56	70.50
Average (μ)	65.09	64.91	65.26	64.57

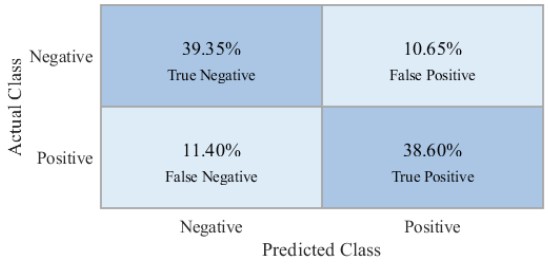


Fig. 6. Confusion matrix of the MLP classifier using the systole segment.

balanced distribution of correct classifications, whereas S1 and diastole presented higher proportions of misclassified samples. These findings are consistent with the performance metrics and show that the configuration combining S2 features with demographic variables provided the best balance between identifying normotensive participants and participants with elevated blood pressure.

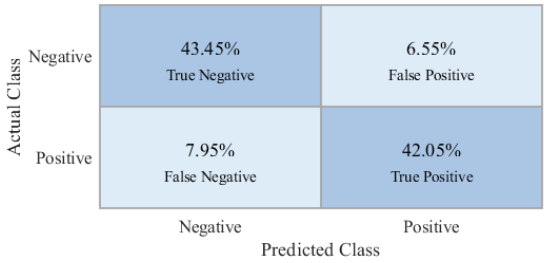


Fig. 7. Confusion matrix of the LR classifier using the S2 segment.

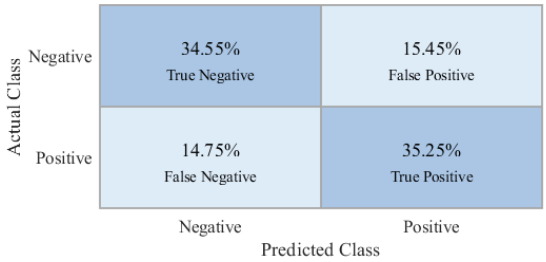


Fig. 8. Confusion matrix of the MLP classifier using the diastole segment.

V. DISCUSSION

This study compared complete, fused, and segment-specific PCG representations, each combined with age, weight, height, and sex. The highest observed F1-scores were 62.29% for complete-PCG and 78.00% for feature-level fusion. Among the segment-specific configurations, S2 features combined with demographic variables yielded the highest observed F1-score of 85.25% with Logistic Regression and the highest mean F1-score across classifiers. Because multiple representations and classifiers were compared using the same cross-validation results, these comparisons are exploratory and the highest observed performance may be optimistic.

One possible physiological explanation for the higher observed performance of the configuration based on S2 features and demographic variables is that S2 is primarily associated with the closure of the aortic and pulmonary valves, whose acoustic characteristics may reflect variations in arterial pressure gradients. This interpretation is consistent with previous studies that identified relationships between S2-derived characteristics and blood pressure variations [11]–[13].

Although feature-level fusion outperformed the complete-PCG configuration, it did not surpass the configuration combining S2 features with demographic variables. Within the evaluated models, adding features from S1, systole, and diastole did not improve performance over the combined S2 and demographic feature set. The additional features may have introduced redundant or less discriminative variability, even after CFS. Further analyses are required to determine the separate contributions of PCG features and demographic variables.

Logistic Regression achieved the best result in the complete-PCG, feature-fusion, and S2 experiments, while more complex models did not consistently improve performance. Considering the limited sample size and the dimensionality reduction performed by CFS, the selected feature space may have favored a comparatively simple decision boundary. This result also suggests that increasing model complexity does not necessarily improve classification when the number of participants is limited.

Several limitations should be considered. The dataset contains only 78 participants and was acquired under controlled conditions using a single PCG device and auscultation location, which restricts the generalization of the findings to other populations and acquisition environments. In addition, the classes were defined from the available reference blood pressure measurements and represent measurement ranges rather than independent clinical diagnoses of hypertension. All experimental scenarios included age, weight, height, and sex. Therefore, we did not separately quantify the contributions of PCG-derived features and participant-related variables. Finally, LR-HSMM segmentation was not evaluated against manual annotations or an independent reference.

VI. CONCLUSION

This study compared complete, fused, and segment-specific PCG representations, all combined with demographic variables, for elevated blood pressure classification. The configuration combining S2 features with demographic variables yielded the highest observed performance, while feature-level fusion did not improve upon it. These findings suggest that PCG information associated with blood pressure ranges may vary across cardiac cycle phases.

If externally validated, segment-specific PCG analysis could support preliminary screening followed by confirmatory cuff-based measurement. However, the limited sample and single controlled dataset currently preclude clinical generalization.

REFERENCES

[1] G. Gao, Z. Chen, G. Yan, and M. Bao, "Impact of hypertensive heart disease, risk factors, and age-period-cohort models across 204 nations and regions from 1990 to 2019: a global perspective from the 2019 global burden of disease study," *Frontiers in Cardiovascular Medicine*, vol. 11, 2024.

[2] W. H. Organization, "Cardiovascular diseases (cvds)." https://www.who.int/health-topics/cardiovascular-diseases. Acesso em: 18/11/2025.

[3] W. H. Organization, "Hypertension." https://www.who.int/news-room/fact-sheets/detail/hypertension. Acesso em: 18/11/2025.

[4] M. Slama, D. Susic, and E. D. Frohlich, "Prevention of hypertension," *Current Opinion in Cardiology*, vol. 17, pp. 531–536, Sept. 2002.

[5] W. K. S. Barroso *et al.*, "Diretrizes brasileiras de hipertensão arterial – 2020." Sociedade Brasileira de Cardiologia, 2020.

[6] M. R. Pioli, A. M. Ritter, A. P. de Faria, and R. Modolo, "White coat syndrome and its variations: differences and clinical impact," *Integrative Blood Pressure Control*, vol. 11, pp. 73–79, 2018.

[7] A. A. Brandão *et al.*, "Diretriz brasileira de hipertensão arterial – 2025," 2025. Brazilian Guidelines of Hypertension – 2025.

[8] K. Kario *et al.*, "The promise and illusion of continuous, cuffless blood pressure monitoring," *Current Hypertension Reports*, vol. 25, pp. 363–370, 2023.

[9] M. F. Braggion-Santos *et al.*, *Physiology, Heart Sounds*. Treasure Island (FL): StatPearls Publishing, 2023.

[10] H. K. Walker, W. D. Hall, and J. W. Hurst, eds., *Clinical Methods: The History, Physical, and Laboratory Examinations*. Boston: Butterworths, 3 ed., 1990. Available from: https://www.ncbi.nlm.nih.gov/books/NBK341/.

[11] X.-Y. Zhang, E. MacPherson, and Y.-T. Zhang, "Relations between the timing of the second heart sound and aortic blood pressure," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 4, pp. 1291–1297, 2008.

[12] A. Castro, S. S. Mattos, and M. T. Coimbra, "Noninvasive blood pressure and the second heart sound analysis," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5494–5497, IEEE, 2014.

[13] Y.-H. Peng, P.-C. Kuo, and Y.-T. Zhang, "Cuffless and continuous blood pressure estimation from the heart sound signals," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6058–6061, 2015.

[14] T. Jumphoo *et al.*, "Cuffless blood pressure estimation from phonocardiogram signals using deep learning with adaptive feature recalibration," *Symmetry*, vol. 17, no. 11, 2025. Art. no. 1943.

[15] W. Chen *et al.*, "Deep learning for heart sound analysis: A literature review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1419, 2021.

[16] R. Gonzalez-Landaeta, B. Ramirez, and J. Mejia, "Estimation of systolic blood pressure by random forest using heart sounds and a ballistocardiogram," *Scientific Reports*, vol. 12, no. 17196, 2022.

[17] Y. Jiang, Y. Chen, and H. Tang, "Real-time intraoperative blood pressure monitoring via heart sound," *Biomedical Signal Processing and Control*, vol. 112, p. 108922, 2026.

[18] K. Kokkhunthod, W. Pathonsuwan, T. Jumphoo, A. Rattanasak, P. Nimkuntod, P. Tongdee, P. Uthansakul, and M. Uthansakul, "Phonocardiogram signal dataset paired with blood pressure measurements," *Data in Brief*, vol. 59, 2025. Art. no. 111433.

[19] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-hsmm-based heart sound segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 822–832, 2016.

[20] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[21] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," pp. 359–366, 01 2000.

[22] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Hoboken, New Jersey: John Wiley & Sons, 3 ed., 2013.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[25] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.

[26] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[27] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 10 2001.

[28] S. Haykin, *Neural Networks and Learning Machines*. Pearson, 3 ed., 2009.

Automated sEMG Signal Processing for Monitoring Motor Rehabilitation in Parkinson's Disease: A Development-Stage Software Platform

Ernesto Yail Pichs Magdariaga[1], Alberto López Delis[1*], Luanne Cardoso Mendes[2], Ariana Moura Cabral[2], Camille Marques Alves[2], Adson Ferreira da Rocha[3], Adriano de Oliveira Andrade[2]

[1]Medical Biophysics Center, University of Oriente, Santiago de Cuba, Cuba

*Corresponding author: lopez.delis69@gmail.com

[2]Centre for Innovation and Technology Assessment in Health (NIATS), Faculty of Electrical Engineering, Federal University of Uberlândia, Uberlândia, Brazil

[3]Department of Electrical Engineering, University of Brasília, Brasília, Brazil

Abstract—Objective and reproducible tools are required to monitor motor rehabilitation in Parkinson's disease (PD) beyond subjective clinical scales. This paper presents a fully automated MATLAB-based software platform for processing surface electromyography (sEMG) signals acquired during serious-gamebased and conventional rehabilitation sessions. The platform integrates four functional modules: data input and loading, signal processing and feature extraction, visualization and graphical user interface (GUI), and statistical analysis with automated reporting. From the raw sEMG recordings of five upper-limb muscles, the system automatically applies a sequential preprocessing pipeline, extracts time-domain, spectral, and nonlinear features, detects muscle activation onset/offset events, and performs paired statistical testing with automated normality assessment via the Shapiro-Wilk test. A preliminary validation was performed on a cohort of fifteen PD patients who completed twelve sessions with the serious game *RehaBEElitation*, plus five patients undergoing conventional physiotherapy. Seven statistically significant intragroup improvements ($p < 0.05$, uncorrected for multiple comparisons) were detected in the experimental group, including increases in task-specific peak amplitude, root mean square (RMS) amplitude, Katz fractal dimension, logarithm of the TeagerKaiser energy operator (log-TKEO), muscle activation rate, and inter-muscular amplitude ratios. The platform constitutes a reproducible, clinician-oriented solution for objective sEMGbased evaluation of PD rehabilitation and is extensible to other neuromuscular disorders. The results presented are preliminary and reflect a development-stage tool requiring further validation with larger samples and intergroup analysis.

Index Terms—Parkinson's disease, surface electromyography, rehabilitation monitoring, serious games, signal processing, MATLAB, automated platform, preliminary validation

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that compromises motor function, manifesting with resting tremor, rigidity, bradykinesia, and postural instability [1], [2]. It is the second most prevalent neurodegenerative disease and the one with the fastest global expansion; projections suggest that the number of affected individuals will exceed 25 million by 2050 [3]. Current pharmacological and surgical treatments—including levodopa administration and deep brain stimulation—provide only symptomatic relief without halting disease progression. Motor rehabilitation therefore emerges as an essential strategy to preserve patient autonomy and quality of life [4].

Serious games and virtual environments have gained relevance as innovative rehabilitation resources, promoting active patient participation and facilitating task-oriented motor training [5], [6]. However, the evaluation of these interventions still depends predominantly on subjective clinical scales—such as the MDS-UPDRS or timed functional tests—which may fail to detect subtle motor changes. More than 70% of rehabilitation studies in PD rely exclusively on such conventional assessments [4], [7].

Surface electromyography (sEMG) offers a noninvasive means of recording neuromuscular activity, enabling objective characterisation of atypical activation patterns in PD [8], [9]. Despite this potential, fully automated computational tools that integrate sEMG processing into serious-game-based rehabilitation protocols for PD remain scarce [7]. Most existing platforms rely on traditional clinical scales, use only kinematic or inertial data, or require manual offline processing of sEMG signals, limiting clinical applicability and reproducibility.

This paper describes the design, functional capabilities, and preliminary validation of a MATLAB-based software

#Restrita

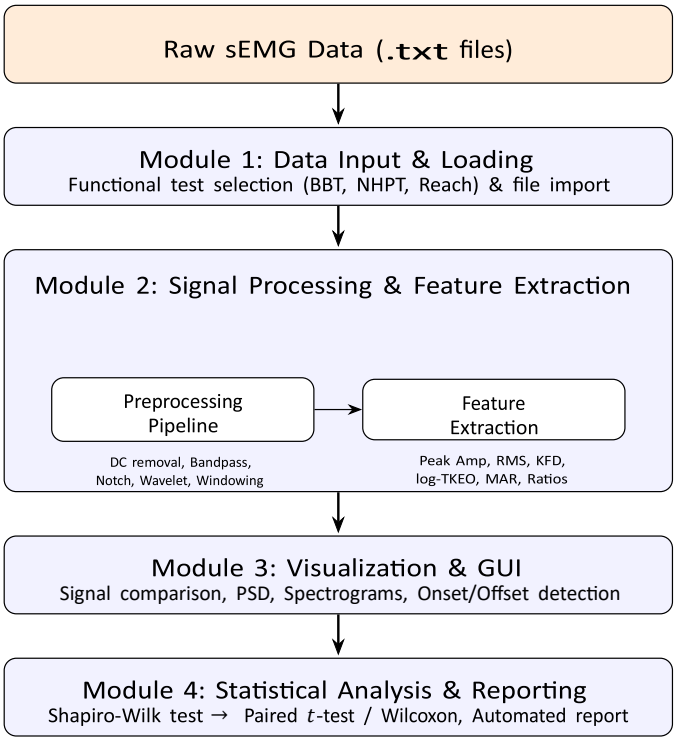


Fig. 1. General workflow and functional diagram of the MATLAB platform.

platform that addresses these gaps. The system performs endto-end sEMG analysis—from raw data import to automated statistical reporting—supporting objective evaluation of PD rehabilitation outcomes. We emphasize that the tool is presented as a development-stage platform; the validation results reported herein are preliminary and intended to demonstrate the tool's functionality and potential, not to establish definitive clinical efficacy.

II. Software Platform Architecture

The platform is organised into four tightly integrated functional modules that operate in a hierarchical manner (Fig. 1): (1) Data Input and Loading, (2) Signal Processing and Feature Extraction, (3) Visualization and GUI, and (4) Statistical Analysis and Reporting. This modular architecture allows the system to function both as an analytical instrument for single-subject exploratory analysis and as a rigorous evaluative platform for group-based clinical research.

A. Module 1: Data Input and Loading

EMG data are provided to the system through a dedicated graphical user interface. The user selects the functional test of interest—Box and Block Test (BBT), Nine Hole Peg Test (NHPT), or reach-and-grasp task—from a dropdown menu in the main application window. A file-selection dialog allows browsing the local file system to import the corresponding .txt files containing raw sEMG signal values. Preintervention (first session) and post-intervention (last session) recordings are loaded separately for each participant. Upon import, signals from all five monitored muscles (Flexor Carpi Radialis, Extensor Carpi Ulnaris, Biceps Brachii, Triceps Brachii, and Anterior Deltoid) automatically enter the complete preprocessing pipeline described in Module 2.

B. Module 2: Signal Processing and Feature Extraction

1) Preprocessing Pipeline: The preprocessing pipeline operates automatically upon data loading and comprises five sequential stages (Fig. 2):

1) DC Offset Removal. The mean of the signal is subtracted, centering it around zero and correcting baseline shifts caused by electrode coupling or impedance variability [10].
2) Bandpass Filtering. A ninth-order IIR Butterworth filter with bidirectional zero-phase processing is applied at 20–350 Hz, preserving muscle-frequency components with a flat passband and suppressing low-frequency motion artifacts and high-frequency noise without phase distortion [11].
3) Notch Filtering. An IIR notch filter centered at 60 Hz and its harmonics (120 Hz, 180 Hz, ...) eliminates power-line interference without affecting the remaining signal spectrum [12].
4) Wavelet-Based Adaptive Denoising. A fourth-order symmetric wavelet with six decomposition levels is applied. Soft thresholding using the BayesShrink method calculates adaptive thresholds to suppress noise without affecting transient muscle patterns [13].
5) Hamming Windowing. A Hamming window of the same length as the signal attenuates edge discontinuities, reducing spectral leakage and improving the accuracy of subsequent time-frequency analyses [14].

All processing parameters are predefined and applied uniformly across all recordings to ensure full reproducibility.

2) Feature Extraction: Following preprocessing, the system automatically extracts the following multi-domain features from each of the five muscle channels:

a) Task-Specific Peak Amplitude: Important methodological note: The metric referred to as "Maximum Voluntary Contraction" (MVC) in this platform is defined as the peak amplitude of the rectified signal during the functional task execution, providing a task-specific index of peak neuromuscular recruitment:

$$\text{Peak} = \max\left|x(t)\right| \tag{1}$$

This is not a true standardized MVC (which would require a maximal isometric contraction against resistance), but rather a task-specific peak amplitude that reflects the maximum activation during the functional tasks (BBT, NHPT, or reachand-grasp). As such, it is better characterized as a taskspecific peak recruitment index. This definition was chosen to avoid the

#Restrita

practical challenges of obtaining standardized MVC measurements in the clinical setting (specific equipment, patient fatigue, compliance issues) and to maintain the fully automated, task-integrated nature of the platform. In PD, decreased peak amplitude values reflect central and peripheral muscle weakness associated with bradykinesia and

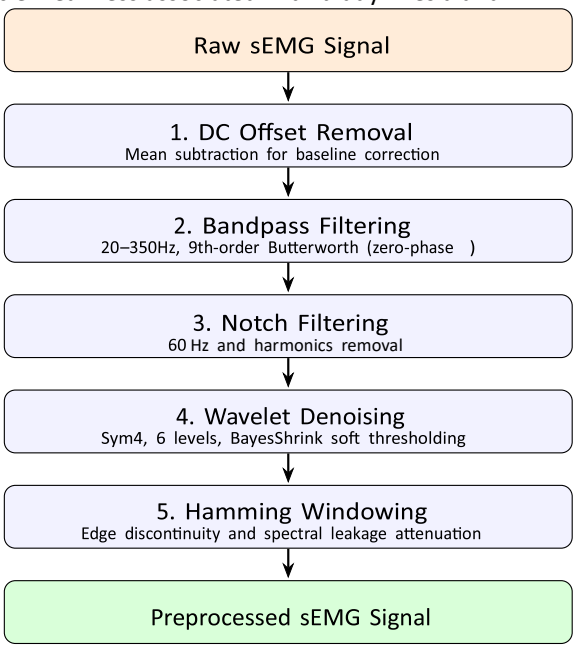


Fig. 2. Sequential preprocessing pipeline for sEMG signals.

rigidity, serving as biomarkers of disease severity [16]. We acknowledge that this metric is sensitive to task execution and may include transient artifacts; future iterations of the platform will incorporate a standardized MVC normalization protocol.

b) Root Mean Square (RMS): Computed over overlapping windows of 200–500 ms:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}, \tag{2}$$

where N is the number of samples per window. Mean and peak RMS values are extracted. Lower RMS in PD patients reflects reduced motor-unit recruitment and firing associated with bradykinesia [9], [15].

c) Kurtosis: A nonlinear descriptor that quantifies the peakedness and tail behaviour of the amplitude distribution, useful for identifying tremor-related transients and rigidity patterns:

$$\kappa = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right]^2}, \tag{3}$$

where μ is the window mean. High kurtosis indicates burst-like or tremor-related activation; low values reflect more uniform neuromuscular control [17], [18].

d) Central Frequency Variance: The centroidal frequency in the i-th window is:

$$f_{c_i} = \frac{\sum_{k=1}^{M} f_k\, P_i(f_k)}{\sum_{k=1}^{M} P_i(f_k)}, \tag{4}$$

where $P_i(f_k)$ is the power spectral density at frequency f_k. The variance across W windows,

$$\mathrm{Var}_{f_c} = \frac{1}{W}\sum_{i=1}^{W}\left(f_{c_i} - \bar{f}_c\right)^2, \tag{5}$$

quantifies motor-unit recruitment stability. Higher variance in PD reflects unstable, erratic activation related to bradykinesia and tremor [19], [20].

e) Katz's Fractal Dimension (KFD): Quantifies the complexity of the EMG waveform:

$$D_{\mathrm{Katz}} = \frac{\log_{10}(n)}{\log_{10}(n) + \log_{10}(d/L)}, \tag{6}$$

where n is the number of signal points, L is the total arc length, and d is the maximum distance from the first point. Lower KFD in PD indicates reduced complexity due to aberrant motor-unit recruitment [21], [22].

f) Logarithm of the Teager-Kaiser Energy Operator (logTKEO): The TKEO emphasises instantaneous energy changes:

$$\Psi[x(n)] = x(n)^2 - x(n-1)x(n+1). \tag{7}$$

Its logarithm,

$$\log\left(\Psi[x(n)] + \varepsilon\right), \quad \varepsilon = 10^{-10}, \tag{8}$$

compresses instantaneous energy dynamics and improves robustness against noise artefacts [23]. In PD, attenuated logTKEO profiles are biomarkers of bradykinesia and altered motoneuron synchronisation [24].

g) Muscle Activation Rate (MAR): The duty cycle above 10% of the task-specific peak amplitude:

$$\mathrm{MAR} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(|x_i| - 0.1 \cdot \mathrm{Peak}) \times 100, \tag{9}$$

where $\mathrm{H}(\cdot)$ is the Heaviside step function. Reduced MAR in PD indicates briefer, more sporadic activations [25]. We note that the 10% threshold is conventional but may not be optimal for

all patients; this is a limitation of the current implementation that will be addressed through adaptive thresholding in future versions.

*h) Inter-Muscular Amplitude and Temporal Ratios:*Pairwise amplitude ratios (based on the task-specific peak
amplitude),

$$R_{ab} = \frac{\text{Peak}_a}{\text{Peak}_b}, \quad (10)$$

characterise the relative dominance of each muscle during a task (Table I). Relative Temporal Activation Metrics (RTAM), derived from onset/offset detection, quantify changes in muscle synergies and co-contraction patterns [26].

TABLE I
INTER-MUSCULAR AMPLITUDE RATIO MATRIX

	M1	M2	M3	M4	M5
M1	–	R_{12}	R_{13}	R_{14}	R_{15}
M2	R_{21}	–	R_{23}	R_{24}	R_{25}
M3	R_{31}	R_{32}	–	R_{34}	R_{35}
M4	R_{41}	R_{42}	R_{43}	–	R_{45}
M5	R_{51}	R_{52}	R_{53}	R_{54}	–

M1: Flexor Carpi Radialis; M2: Extensor Carpi Ulnaris; M3: Biceps Brachii; M4: Triceps Brachii; M5: Anterior Deltoid.

C. Module 3: Visualization and Graphical User Interface

Upon completion of preprocessing, a dedicated comparison window opens automatically. For each selected muscle, the GUI simultaneously displays:

- Raw and filtered waveforms of both sessions (superimposed or side-by-side);
- Rectified signals and linear envelopes;
- Power spectral density (PSD) curves;
- Time-frequency spectrograms computed via short-time Fourier transform (STFT).

A single click on the *Show Statistics* button computes and displays a comprehensive comparative table containing all extracted features for each session and muscle. All values can be exported in tabular form (copy-to-clipboard or .csv) for external processing.

1) Onset/Offset Detection Submodule: Accessible from the comparison window, this submodule applies the TKEO followed by adaptive thresholding and morphological validation to all five channels simultaneously (Fig. 3). The detection algorithm proceeds as follows:

1) Compute the TKEO envelope $\Psi[x(n)]$.
2) Apply a moving-average filter of length L:

$$y(n) = \frac{1}{L}\sum_{k=0}^{L-1}\Psi[x(n-k)] \quad . \quad (11)$$

3) Estimate background noise standard deviation σ from non-activation segments.
4) Define an adaptive activation threshold:

$$U = \mu_{\text{noise}} + M_\sigma \sigma, \quad (12)$$

where M_σ is a user-tunable multiplier. 5) Detect candidate onset/offset events where $y(n)$ crosses U.
6) Apply morphological validation to reject spurious bursts shorter than a minimum duration.

Detected onsets (green vertical lines) and offsets (red vertical lines) are overlaid on the filtered signals, enabling immediate visual inspection of co-contraction changes and sequential activation patterns between sessions. Detection parameters (window size, noise samples, threshold multiplier, minimum burst duration) can be fine-tuned interactively in real time. Important: The user-adjustable parameters are provided

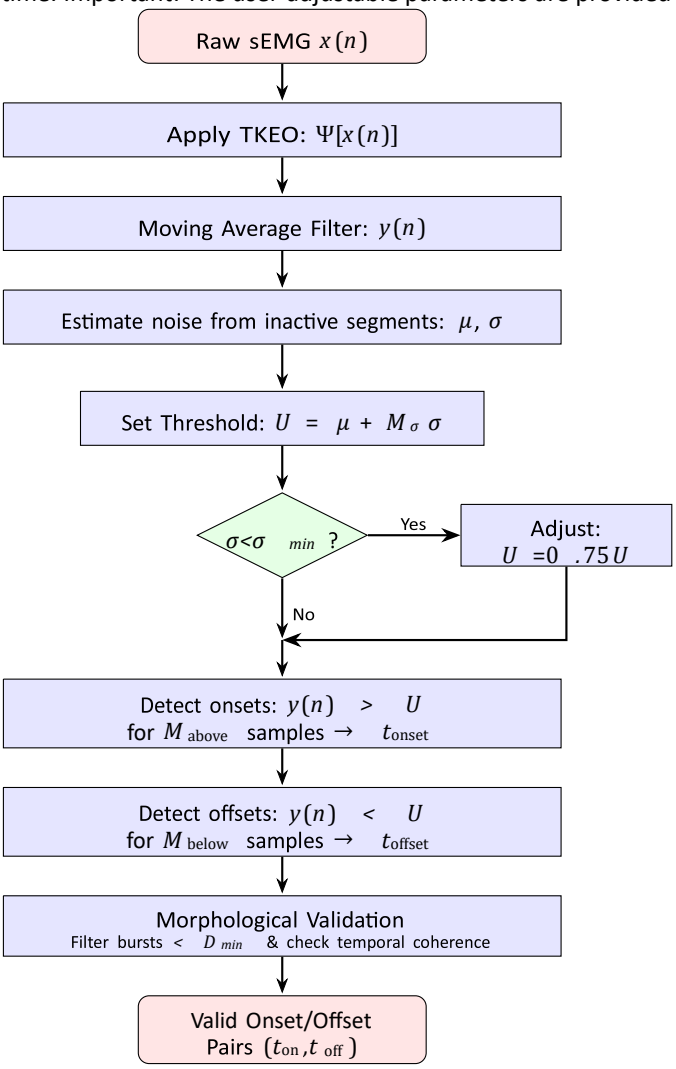


Fig. 3. Automated onset/offset detection algorithm flowchart.

#Restrita

for exploratory analysis and to accommodate signal variability; however, the platform also supports fixed, predefined parameters for automated batch processing. The current implementation has not been formally validated against manual annotation or reference methods, and this validation is planned as future work.

D. Module 4: Statistical Analysis and Reporting

The statistical module performs automatic paired-samples comparison of all features extracted from pre- and postintervention sessions across all monitored muscles and all subjects. The normality of the paired differences is assessed automatically using the Shapiro-Wilk test. Based on this outcome:

- If normality is not rejected, a parametric paired t-test is applied.
- Otherwise, the non-parametric Wilcoxon signed-rank test is used.

This automated decision logic is critical because sEMGderived variables frequently present non-normal distributions due to the inherent variability of neuromuscular activation [27]. Results are displayed as significance tables (p-values and test type) and can be exported to .csv for downstream reporting. We acknowledge that the current implementation does not apply correction for multiple comparisons (e.g., Bonferroni), and does not report effect sizes or confidence intervals. These are planned enhancements for future versions.

III. Platform Validation

A. Study Design and Participants

The platform was validated on a cohort of twenty PD patients: fifteen in the experimental group (EG) who completed twelve serious-game rehabilitation sessions with *RehaBEElitation*, and five in the control group (CG) who underwent conventional physiotherapy (twelve sessions each). Groups were matched by age and sex. Inclusion criteria required confirmed PD diagnosis, Hoehn & Yahr scale stages I–III, MMSE score above 22, and absence of severe sensory or other neurological impairments. All participants provided written informed consent (Ethics Committee protocol 43229921.8.0000.5152 Opinion Number: 8.262.209 (date: 05/03/2026)).

We acknowledge that the sample size is small and unbalanced (EG: n=15, CG: n=5), which limits the statistical power and generalizability of the findings. The results reported are intended as preliminary evidence of platform functionality rather than definitive clinical validation.

B. Data Acquisition

Surface EMG was recorded from five upper-limb muscles per participant (Flexor Carpi Radialis, Extensor Carpi Ulnaris, Biceps Brachii, Triceps Brachii, and Anterior Deltoid) during three standardised functional tasks: BBT, NHPT, and a reachand-grasp task—administered before and after the intervention protocol. Electrode placement followed established guidelines with 2 cm inter-electrode distance and a reference electrode on the acromion of the scapula.

C. Results

The platform processed all recordings automatically without any manual intervention at the signal processing or statistical stages. In the EG, seven statistically significant improvements ($p < 0.05$, uncorrected for multiple comparisons) were identified following the *RehaBEElitation* intervention:

- Significant increase in task-specific peak amplitude and RMS amplitude in the Biceps Brachii, indicating enhanced motor-unit recruitment.
- Increase in Katz Fractal Dimension in forearm muscles, reflecting greater neuromuscular complexity and more diverse activation patterns.
- Increase in log-TKEO, indicating improved instantaneous energy of muscle contractions.
- Increase in Muscle Activation Rate, reflecting longer and more consistent activation above threshold.
- Significant changes in inter-muscular amplitude ratios, suggesting improved coordination and redistribution of force across the monitored muscle group.

Due to the unbalanced sample size (EG: $n = 15$; CG: $n = 5$), analysis was restricted to within-group comparisons. The small CG precluded robust inter-group statistical evaluation; future studies should include larger, balanced cohorts. The lack of intergroup comparison means that the observed changes cannot be definitively attributed to the intervention versus spontaneous recovery or placebo effects.

IV. Discussion

A. Platform Functionality

The results demonstrate that the proposed platform can reliably detect subtle neuromuscular adaptations in PD patients following a serious-game intervention, overcoming the limitations of subjective clinical scales. The fully automated preprocessing, feature extraction, onset/offset detection, and statistical reporting pipeline eliminates operator-dependent variability, ensuring reproducibility across clinical sites.

The multi-domain feature set (time, spectral, and nonlinear) captures complementary aspects of motor recovery. Taskspecific peak amplitude and RMS inform on changes in peak and sustained neuromuscular recruitment; KFD and kurtosis quantify alterations in signal complexity and burst irregularity; log-TKEO highlights changes in instantaneous

#Restrita

energy; and MAR together with RTAM provide information on activation efficiency and inter-muscular coordination. This comprehensive characterisation exceeds the sensitivity of isolated metrics used in prior sEMG-based PD studies [15], [18].

The adaptive normality assessment (Shapiro-Wilk → *t*test or Wilcoxon) is a methodologically sound design choice, given that sEMG features regularly depart from Gaussianity in clinical populations [27]. The interactive onset/offset detection submodule further extends clinical utility by enabling real-time exploration of co-contraction dynamics and muscle synergy changes—phenomena increasingly recognised as key markers of rehabilitation-induced motor reorganisation [26].

B. Limitations and Future Work

Statistical Limitations: The current study has several important limitations. First, the small and unbalanced sample (EG: n=15, CG: n=5) severely limits statistical power and precludes meaningful intergroup comparisons. The reported *p*-values are uncorrected for multiple comparisons, increasing the risk of Type I errors. Effect sizes and confidence intervals were not computed, which are essential for clinical interpretation. Future studies with larger, balanced cohorts and appropriate statistical corrections (e.g., Bonferroni, HolmBonferroni, or FDR) are required to validate these preliminary findings.

Signal Processing Considerations: The Hamming window is applied to the entire signal, which may alter amplitude-based measures (RMS and peak). This is a recognized limitation of the current implementation; we note that for our preliminary analyses, the windowing was applied consistently across all signals, and the primary focus was on relative changes between sessions rather than absolute values. Future versions of the platform will implement windowing only for spectral analysis, preserving raw amplitude for temporal metrics.

The task-specific peak amplitude metric (referred to as MVC in the platform) is not a standardized MVC measurement but rather the peak activation during the functional task. This is a pragmatic choice to maintain automation and avoid the need for separate MVC recording sessions. However, this metric is sensitive to task execution and may include transient artifacts. Future work will incorporate a standardized MVC normalization protocol for more robust comparisons.

The 10% threshold for MAR is conventional but may not be optimal for all patients or muscles. Future versions will explore adaptive thresholding methods to improve comparability across sessions.

The onset/offset detection algorithm has not been formally validated against manual annotation or reference methods. Future work will include systematic validation with ground-truth data and comparison with established detection algorithms.

Performance and Reproducibility: The current paper does not report quantitative performance metrics such as detection accuracy, processing time per subject, failure rate, or testretest reliability. Formal validation with SNR or RMSE metrics for the wavelet denoising stage has not yet been performed. These are critical for clinical adoption and will be addressed in future work. We will also conduct comparative studies with established sEMG processing tools (e.g., EMGlab, OT Bioelettronica software) to benchmark the platform's performance.

Clinical Validation: While the platform demonstrates potential for objective assessment, the current evidence does not establish clinical validity, superior sensitivity to traditional scales, or definitive efficacy of the intervention. The preliminary nature of these results must be emphasized; the primary contribution of this paper is the description of a novel, automated software platform, and the validation is illustrative rather than confirmatory. Future studies with larger samples, intergroup comparisons, and correlation with clinical scales (MDS-UPDRS) are required to establish clinical utility.

Mitigation of Limitations: We are actively working on the following enhancements:

- Expansion of sample size through ongoing data collection.
- Implementation of intergroup analysis with properly matched controls.
- Incorporation of multiple comparison corrections and effect size reporting.
- Formal validation of onset/offset detection against manual annotation.
- Quantitative evaluation of denoising performance using SNR and RMSE metrics.
- Benchmarking against established sEMG processing tools.
- Development of standardized MVC normalization protocols.

V. Conclusions

This paper presented a modular, fully automated MATLABbased software platform for objective sEMG analysis in Parkinson's disease rehabilitation. The system integrates data loading, a five-stage preprocessing pipeline, multi-domain feature extraction (task-specific peak amplitude, RMS, kurtosis, central frequency variance, KFD, log-TKEO, MAR, and intermuscular ratios), interactive visualization with onset/offset detection, and automated paired statistical testing into a single, clinician-oriented workflow.

Preliminary validation on a cohort of PD patients confirmed the platform's sensitivity, detecting seven significant intragroup neuromuscular improvements following twelve sessions of serious-game rehabilitation ($p < 0.05$, uncorrected for multiple comparisons). The platform is available as open-source software [30] and represents a reproducible foundation

#Restrita

for future large-scale clinical trials aimed at definitively evaluating the efficacy of game-based rehabilitation in PD.

We emphasize that the results presented are preliminary and reflect a development-stage tool. The primary contribution of this work is the platform architecture and automation capabilities, not definitive clinical validation. Substantial future work is required, including larger sample sizes, intergroup comparisons, multiple comparison corrections, effect size reporting, onset/offset validation, SNR quantification, and clinical validation against established scales.

Acknowledgment

The present work was carried out with the support of the National Council for Scientific and Technological Development (444437/2024-0, 442150/2023-7, and 405365/2023-3), Coordination for the Improvement of Higher Education Personnel (CAPES), Foundation for Research Support of the State of Minas Gerais (FAPEMIG APQ-05015-24). AOA (302942/2022-0) and CMA (153088/2024-9) are CNPq fellows. The funders had no role in study design, data collection, analysis, decision to publish, or manuscript preparation. No generative artificial intelligence was used in any part of this manuscript.

References

[1] S. Sveinbjornsdottir, "The clinical symptoms of Parkinson's disease," *J. Neurochem.*, vol. 139, no. S1, pp. 318–324, 2016.

[2] E. R. Dorsey, T. Sherer, M. S. Okun, and B. R. Bloem, "The emerging evidence of the Parkinson pandemic," *J. Parkinson's Dis.*, vol. 8, no. S1, pp. S3–S8, 2018.

[3] Y. Ben-Shlomo *et al.*, "The epidemiology of Parkinson's disease," *The Lancet*, vol. 403, no. 10423, pp. 283–292, 2024.

[4] K. Dockx *et al.*, "Virtual reality for rehabilitation in Parkinson's disease," *Cochrane Database Syst. Rev.*, no. 12, p. CD010760, 2016.

[5] I. Quirosa Galán *et al.*, "Video games in rehabilitation programs for people with Parkinson's disease: A systematic review," *Appl. Sci.*, vol. 15, no. 1, p. 311, 2024.

[6] M. Tarousi *et al.*, "Serious games for Parkinson's disease management as implemented in PROPHETIC platform," *Health Inform. J.*, vol. 27, no. 2, p. 14604582211011231, 2021.

[7] A. Garcia Agundez and C. Eickhoff, "Towards objective quantification of hand tremors and bradykinesia using contactless sensors: A systematic review," *Front. Aging Neurosci.*, vol. 13, p. 716102, 2021.

[8] U. Kleinholdermann, M. Wullstein, and D. Pedrosa, "Prediction of motor UPDRS scores in patients with Parkinson's disease using surface electromyography," *Clin. Neurophysiol.*, vol. 132, no. 7, pp. 1708–1713, 2021.

[9] M. Nakamori *et al.*, "High density surface electromyography evaluation of the suprahyoid muscles in patients with Parkinson's disease," *J. Neurol. Sci.*, vol. 466, p. 123266, 2024.

[10] R. Merletti and P. Parker, Eds., *Electromyography: Physiology, Engineering, and Noninvasive Applications*. Wiley, 2004.

[11] M. Davoudi *et al.*, "Are electromyography data a fingerprint for patients with cerebral palsy?" *Appl. Sci.*, vol. 15, no. 2, p. 766, 2025.

[12] J. Chen *et al.*, "Reducing power line interference from sEMG signals based on synchrosqueezed wavelet transform," *Sensors*, vol. 23, no. 11, p. 5182, 2023.

[13] C. Ouyang, T. Liang, and Z. Li, "An improved wavelet threshold denoising approach for surface electromyography signal," *EURASIP J. Adv. Signal Process.*, vol. 2023, p. 104, 2023.

[14] Z. Zhang *et al.*, "Hardware and software design and implementation of surface EMG-based gesture recognition and control system," *Electronics*, vol. 13, no. 2, p. 454, 2024.

[15] A. Y. Meigal *et al.*, "Non-linear EMG parameters for differential and early diagnostics of Parkinson's disease," *Front. Neurol.*, vol. 4, p. 135, 2013.

[16] A. Monte *et al.*, "Muscle shape changes in Parkinson's disease impair function during rapid contractions," *Acta Physiol.*, vol. 238, no. 1, p. e13957, 2023.

[17] S. M. Rissanen *et al.*, "Signal features of surface electromyography in advanced Parkinson's disease during different settings of deep brain stimulation," *Clin. Neurophysiol.*, vol. 126, no. 12, pp. 2290–2298, 2015.

[18] S. Rissanen *et al.*, "Ambulatory surface electromyography with accelerometry for evaluating daily motor fluctuations in Parkinson's disease," *MDS Abstracts*, 2021.

[19] R. H. Chowdhury *et al.*, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, no. 9, pp. 12431–12466, 2013.

[20] S. Farashi, "Distinguishing between Parkinson's disease patients and healthy individuals using a comprehensive set of time, frequency and time-frequency features," *Biomed. Signal Process. Control*, vol. 62, p. 102132, 2020.

[21] K. Zhao *et al.*, "Fractal characteristics based motor dyskinesia assessment," *Biomed. Signal Process. Control*, vol. 68, p. 102707, 2021.

[22] L. Livi, A. Sadeghian, and H. Sadeghian, "Discrimination and characterization of Parkinsonian rest tremors by analyzing long-term correlations and multifractal signatures," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 11, pp. 2243–2249, 2016.

[23] S. Solnik *et al.*, "Teager-Kaiser energy operator signal conditioning improves EMG onset detection," *Eur. J. Appl. Physiol.*, vol. 110, no. 3, pp. 489–498, 2010.

[24] S. Micera *et al.*, "Improving detection of muscle activation intervals," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 6, pp. 38–46, 2001.

[25] C. S. Klein *et al.*, "Sensitivity of 24-h EMG duration and intensity in the human vastus lateralis muscle to threshold changes," *J. Appl. Physiol.*, vol. 108, no. 3, pp. 655–661, 2010.

[26] E. Lecce *et al.*, "Higher dominant muscle strength is mediated by motor unit discharge rates and proportion of common synaptic inputs," *Sci. Rep.*, vol. 15, no. 1, p. 8269, 2025.

[27] C. Starbuck, "Analysis of differences," in *The Fundamentals of People Analytics*. Springer, 2023, pp. 149–179.

[28] N. A. Kelly *et al.*, "Novel, high-intensity exercise prescription improves muscle mass, mitochondrial function, and physical capacity in individuals with Parkinson's disease," *J. Appl. Physiol.*, vol. 116, no. 5, pp. 582– 592, 2014.

[29] L. Roeder *et al.*, "Effects of resistance training on measures of muscular strength in people with Parkinson's disease: A systematic review and meta-analysis," *PLOS ONE*, vol. 10, no. 7, p. e0132135, 2015.

[30] A. Lopez-Delis and Y. P. M. Ernesto, "Tool for the evaluation of a rehabilitation protocol for Parkinson's disease," Zenodo, 2026. [Online]. Available: https://doi.org/10.5281/zenodo.18262481

[31] I. Campanini *et al.*, "Surface EMG in clinical assessment and neurorehabilitation: Barriers limiting its use," *Front. Neurol.*, vol. 11, p. 934, 2020.

[32] R. P. Bremm *et al.*, "Sensor-based quantification of MDS-UPDRS III subitems in Parkinson's disease using machine learning," *Sensors*, vol. 24, no. 7, p. 2195, 2024.

#Restrita

Classification of Epileptic States from EEG Signals Using Wavelet Transform and Deep Neural Networks

1st Théo S. Teixeira
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
theo.saldanhateixeira@ufpe.br

2nd Juliana A. S. Costa
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
juliana.ascosta@ufpe.br

3rd Déborah E. L. Alcantara
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
deborah.evelyna@ufpe.br

4th Gabriel L. dos Santos
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
gabriel.limasantos@ufpe.br

5th Sandrelly Mirely da Silva Santos
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
sandrelly.santos@ufpe.br

6th Wellington Pinheiro dos Santos
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

***Abstract*—This work presents the development of a digital signal processing and machine learning system for epileptic seizure detection. The primary objective is to differentiate between interictal and ictal states using electroencephalogram (EEG) recordings from the Siena Scalp EEG Database. The proposed methodology employs a processing pipeline comprising channel selection, filtering, temporal window segmentation, and conversion of these segments into time-frequency representations via the Continuous Wavelet Transform (CWT). The approach frames the problem as a computer vision task, wherein the generated time-frequency images (scalograms) are fed into a pre-trained Convolutional Neural Network (CNN) for deep feature extraction. Final classification is carried out through a comparative evaluation of decision tree-based models. Preliminary results indicate that the image-based representation of EEG signals reveals distinct activity patterns across frequency bands, demonstrating the applicability of machine learning to this classification task. The CWT yielded time-resolved frequency information both during and outside seizure episodes, while spatial interpolation of electrode data enabled the generation of scalp topographic maps for each frequency band.**

***Index Terms*—EEG; Epilepsy; Wavelet Transform; Convolutional Neural Networks**

I. Introduction

Epilepsy is one of the most prevalent neurological conditions worldwide, affecting approximately 50 million people globally, according to the World Health Organization (WHO) [1]. This syndrome significantly impacts patients' quality of life, potentially leading to physical, psychological, and social implications. Therefore, early detection, continuous monitoring and management, and accurate classification of epileptic states—such as interictal and ictal—are fundamental steps for therapeutic planning and seizure prediction [2]. In this context, the development of supportive technologies becomes essential to advance clinical decision-making and patient care.

Currently, the electroencephalogram (EEG) has become established as the primary non-invasive tool for studying, diagnosing, and monitoring brain activity in real time, offering high temporal resolution and sensitivity to neurological alterations associated with epilepsy and other seizure disorders [3]. Notably, EEG plays a critical role in identifying the epileptogenic zone, which may vary significantly across patients [4]. However, manual analysis of EEG recordings is extremely labor-intensive, subjective, and prone to human error—particularly in long-term monitoring scenarios. Furthermore, epileptic patterns often manifest subtly and exhibit high clinical variability among individuals [5]. This transient and non-stationary nature of the biological signal challenges traditional methods based on spectral analysis (such as the Fourier Transform) or purely time-domain features, as these approaches fail to capture localized frequency changes occurring at the precise onset of a seizure [6].

Therefore, this work proposes the development and evaluation of an automated system for classifying epileptic states from EEG signals. The approach integrates the Wavelet Transform for multiresolution signal analysis, capturing both abrupt transitions and subtle variations in brain activity [7]. This transform was selected to simultaneously capture the time and frequency characteristics of ictal, interictal, and transitional events, as EEG signals are inherently non-stationary and tracking state dynamics is a key objective of this study [2]. This time-frequency feature extraction step serves as the foundation for deep learning models, specifically designed to identify complex, nonlinear patterns that are often imperceptible to conventional analysis. The integration of these techniques aims not only to automate detection but also to enhance accuracy and reliability in distinguishing between ictal and interictal periods, ultimately contributing to more precise and accessible

brain activity monitoring.

II. Methodology

This study employs a series of steps to process neural signals, enabling the classification of epileptic phases using neural networks. These procedures are described in the flowchart shown in Fig. 1

A. Dataset Selection

This study utilized the Siena Scalp EEG Database, publicly available on the PhysioNet platform [8]. The dataset comprises electroencephalogram (EEG) recordings from 14 patients (9 male and 5 female), aged 20 to 71 years, acquired at the Neurology and Neurophysiology Unit of the University of Siena. Signals were recorded at a sampling rate of 512 Hz and include expert annotations marking seizure onset and offset.

The sample exhibits significant clinical heterogeneity, with patients presenting different seizure types, including focal impaired awareness seizures (FIAS), focal aware seizures (FAS), and those evolving to bilateral tonic-clonic seizures. Additionally, variability is observed regarding the localization of epileptic activity (temporal region) and lateralization—previously diagnosed as right hemisphere, left hemisphere, or bilateral involvement [9].

To ensure that the machine learning model was exposed to broad physiological and clinical variability while minimizing sampling bias, patient allocation to training and testing sets was performed following stringent criteria. The training set comprised six patients, encompassing both sexes, diverse age ranges, multiple seizure types, and left, right, and bilateral lateralization. The test set included three patients, selected to maintain comparable heterogeneity in terms of sex, age, and clinical manifestations. With a total of 9 out of the 14 patients included in this split, the adopted division enables a more robust and clinically representative validation of the proposed approach without requiring the use of the entire dataset [8].

B. Preprocessing

Following patient selection, the preprocessing stage was initiated. Initially, only channels corresponding to the International 10-20 system for EEG electrode placement were retained, excluding spurious or unannotated channels [10]. Subsequently, the spatial signals underwent digital filtering using a Finite Impulse Response (FIR) bandpass filter. The FIR topology was selected due to its linear phase property, which prevents time-domain distortions and preserves the original morphology of brain waves—a critical requirement for subsequent feature extraction via the Continuous Wavelet Transform (CWT). The passband was set between 0.5 Hz and 45 Hz, using the firwin filtering method from the MNE-Python library to calculate the FIR filter coefficients based on the sampling band; the order of the filter was not fixed, as it depends on the sampling frequency of each of the processed signals. The data were restricted to the physiologically relevant EEG frequency range while simultaneously attenuating baseline drift and high-frequency noise [4].

To characterize the non-stationary dynamics of cerebral activity, the filtered spatial signals were segmented into short temporal windows of 500 ms. This window duration was selected to provide high temporal resolution, facilitating the analysis of low- and mid-frequency EEG components, albeit with partial representation of slower frequencies (< 2 Hz). Additionally, a 25% overlap was applied between consecutive windows. This strategy aims to mitigate edge artifacts inherent to frequency-domain transforms and ensure uninterrupted capture of information across brain state transitions [11] and [6].

Given that continuous EEG recordings exhibit severe class imbalance—with the interictal state representing the vast majority of the data—an extraction strategy based on regions of interest (ROIs) was adopted. Processing was restricted to approximately 100-second epochs aligned with each epileptic event. This interval was evenly split into 50 s of ictal activity (a value reflecting the average seizure duration in the selected cohort) and the 50 s immediately preceding clinical onset (interictal activity, with emphasis on the pre-ictal period) [12]. For seizures lasting less than 50 s, the actual ictal duration is retained; for longer events, only the initial 50 s are used. This constraint not only ensures representative class balancing for training the final classifier but also drastically reduces the computational cost associated with generating multi-channel images from the complete original recordings [13].

The entire methodology described herein was implemented in Python within the Google Colab environment, with access to the database files provided via Google Drive. The MNE-Python library (version 1.11.0) [14] was employed for reading EDF-format files, selecting and renaming channels according to the 10-20 system, applying the spatial montage, and performing FIR bandpass filtering. Numerical array manipulation, including the computation of windowing indices and segment extraction, was carried out using the NumPy library (version 2.0.2) [15].

C. Wavelet Transform

A key methodological contribution of the proposed approach lies in the signal processing stage, which is based on the Continuous Wavelet Transform (CWT) using the Morlet mother wavelet. This wavelet was selected due to its ability to provide an effective balance between time and frequency resolution, as well as its high morphological similarity to the sinusoidal oscillations typically observed in neural signals [16].

Specifically, the CWT was applied independently to each one-dimensional EEG channel within each already-isolated 500 ms temporal window, generating a matrix of time-frequency coefficients for each segment. For each temporal window, the spectral energy obtained from the CWT was integrated and averaged within the frequency ranges corresponding to four clinically relevant EEG bands: Delta ($\leq 3,5$ Hz), Theta (4–7,5 Hz), Alpha (8–12 Hz), and Beta (14–30 Hz).

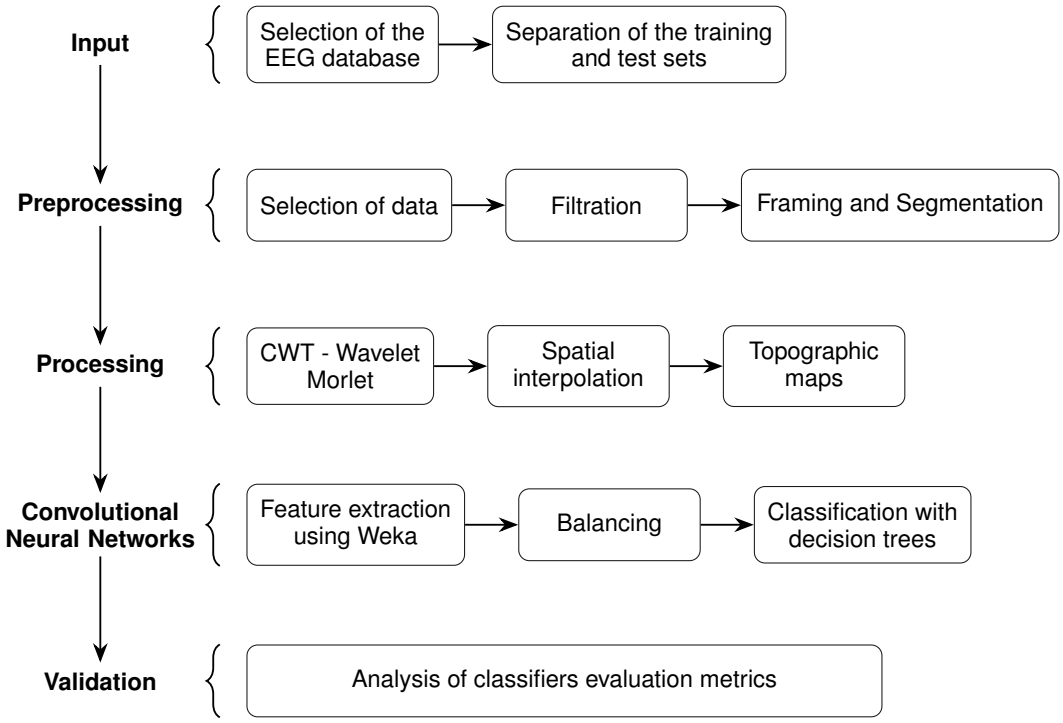


Fig. 1. Detailed flowchart of the proposed methodology for processing and classifying EEG signals.

To convert the channel-wise energy values into a continuous two-dimensional spatial representation, the three-dimensional coordinates of the EEG electrodes, positioned according to the international 10–20 system, were projected onto a two-dimensional plane. Subsequently, a spline-based spatial interpolation algorithm was applied to estimate neural activity in the unsampled regions between electrodes, producing a dense pixel matrix for each of the four frequency bands [17].

This procedure generated four independent topographic maps for each temporal window. These maps were then combined into a single EEG-derived topographic image through horizontal concatenation, resulting in a unified visual representation with a resolution of 564 × 2416 pixels (height × width). Such a representation enables deep learning architectures, such as ResNet, to extract complex patterns based on the simultaneous analysis of spatial and spectral correlations [17] [18].

Additionally, during the colormap generation process, a logarithmic transformation was applied to the spectral power values (ranging from $0.01\,\mu V$ to $100\,\mu V$) prior to normalization into the RGB color space. This procedure was essential for accommodating the high dynamic range of EEG signals, allowing low-amplitude fluctuations to remain visible while preventing clipping and information loss caused by large-amplitude variations. Consequently, the visual representation preserved the temporal evolution of the signal across its full amplitude range while maintaining the fidelity of the underlying neurophysiological information [19].

The rendering of topographic maps was performed using the Matplotlib library, both directly and as the underlying rendering engine for MNE-Python's plotting functions [14]. Logarithmic amplitude normalization was applied via the LogNorm class from the matplotlib.colors module (version 3.10.0) [20], ensuring consistent color mapping across the full dynamic range of EEG amplitudes. Finally, the four frequency-band maps generated for each temporal window were spatially concatenated into a single composite image using the Pillow (PIL) library (version 11.3.0) [21], producing the multichannel representations subsequently used as input to the deep learning model.

D. Machine Learning and Weka

The data mining and machine learning stage was entirely conducted using the Weka platform (Waikato Environment for Knowledge Analysis) [22]. Initially, the generated topographic representations were structured into .arff files, organized into training and testing sets. For feature extraction, the ResNet-50 architecture [23]—a Convolutional Neural Network (CNN) pre-trained with ImageNet dataset—was employed to identify complex visual, spatial, and spectral patterns within each temporal window.

During dataset organization, a natural class imbalance was observed. This occurred because several recordings contained seizures with clinical durations shorter than the standard 50-second window, resulting in an underrepresentation of ictal instances relative to interictal ones. To address this asymmetry in the training set and prevent classifier bias toward the majority class, the SMOTE algorithm (Synthetic Minority Over-sampling Technique) was applied [24]. This synthetic oversampling technique goes beyond simple data duplication;

it operates in feature space by interpolating values between neighboring instances of the minority class, thereby generating new, representative, and consistent samples that support robust model learning.

Finally, the classification of deep features was performed using decision tree-based supervised learning models. These algorithms were selected due to their low complexity for good computational performance. The evaluated classifiers included: Random Forest (with the number of trees ranging from 100 to 500 in increments of 100), Random Tree, and J48. These models were evaluated using cross-validation, which was applied exclusively to the training set for hyperparameter tuning. The test set was strictly isolated at the subject level, ensuring that no data from the test subjects were included in the training or validation sets. This approach prevents data leakage and guarantees an unbiased assessment of the model's generalization capability across different individuals [25].

The hyperparameter regarding the number of decision trees in the Random Forest ensemble was varied across 100, 200, 300, 400, and 500, in addition to the evaluation of Random Tree and J48 algorithms. The quantitative selection of this parameter is substantiated by stability tests to identify an optimal convergence point for minimizing generalization error and model variance, ensuring high accuracy without incurring computational saturation or overfitting [26].

These algorithms were evaluated and benchmarked to determine the most effective approach for the classification of ictal (seizure) and interictal (non-seizure) periods in patients.

III. Results and Discussion

Following the previously described methodology, Wavelet transform images were generated to spatially represent the intensity of brain activity, segmented into four distinct frequency bands: delta (1 to 3.5 Hz), theta (4 to 7.5 Hz), alpha (8 to 12 Hz) and beta (13 to 30 Hz). Approximately 266 images were obtained for each examination of the selected patients, comprising both ictal and interictal activity phases. These images were constructed as topographical maps, where color scales indicate the intensity of cerebral activity within specific regions [19]. The resulting multichannel images, characterized by the concatenation of these frequency bands, are illustrated in Fig. 2 and Fig. 3 below:

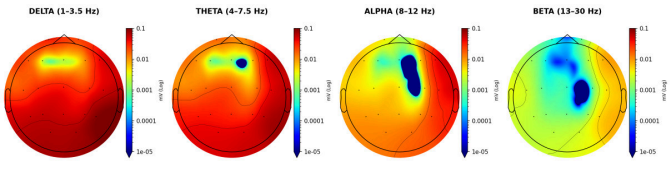


Fig. 2. Topographical representation of ictal activity.

From the processed images, an .arff dataset was compiled, comprising 1,729 interictal and 1,441 ictal instances. To address the class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied at a 20% rate, resulting in a balanced distribution of 1,729 samples for each class. Regarding the training set, following the same experimental pipeline, a final count of 1,333 instances per class was obtained.

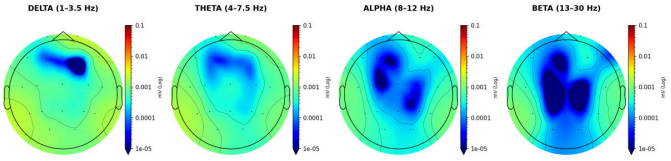


Fig. 3. Topographical representation of interictal activity.

Following feature extraction using the ResNet-50 architecture, a total of 2,048 features were yielded for both the training and testing datasets. Subsequently, the upcoming decision tree-based classifiers were evaluated: Random Forest (with the number of trees ranging from 100 to 500), Random Tree, and the J48 algorithm.

Through 10-fold cross-validation, it was observed that the highest performance metrics—including accuracy, sensitivity, specificity, Kappa index, and Area Under the ROC Curve (AUC)—were achieved by the Random Forest classifiers with 300, 400 and 500 trees, as illustrated in Tab. I. Given the highly similar performance across these models, multiple comparisons were performed using the ROC area as the primary metric, as it was the only parameter exhibiting variance among the three. Statistical analysis revealed that the three classifiers are practically equivalent, as the p-value for the significance of the difference was p = 1.0. Consequently, the model with the lowest computational overhead, the Random Forest with 300 trees, was selected as the optimal choice.

Performance metrics obtained via 10-fold cross-validation for the evaluated classifiers are presented in Table I, while the corresponding multiple comparison analysis is shown in Table II. Furthermore, the results of the selected RF 300 classifier, when applied to the independent test set, are summarized in Table III, with its respective confusion matrix illustrated in Table IV.

An analysis of the images generated for each patient reveals a high degree of correspondence with their clinical characteristics. Specifically, patients with diagnosed left-side lateralization exhibited heightened activity in the left regions of the generated topographical maps. For instance, the maps for Patient 5, who has a diagnosed epileptic focus in the left hemisphere, reflect this localization as shown in Fig. 4. Conversely, Patient 17, whose seizures predominantly affect the right hemisphere, presents images characterized by right-side activity, as illustrated in Fig. 5.

Regarding the machine learning performance metrics, the results demonstrated significant robustness, with consistent values indicating a low-bias classification. Furthermore, the model exhibited high discriminative power, effectively distinguishing between the ictal and interictal classes.

The confusion matrix further reveals that the model achieves higher classification accuracy for interictal activity compared to ictal events. Despite this variance, a promising application

TABLE I
PERFORMANCE METRICS OBTAINED THROUGH 10-FOLD CROSS-VALIDATION FOR ALL EVALUATED CLASSIFIERS

Group	Classifiers	Accuracy	Kappa	Sensitivity	Specificity	ROC Area
Mean	RF 100	0.875	0.751	0.875	0.875	0.941
	RF 200	0.877	0.753	0.877	0.877	0.942
	RF 300	0.878	0.756	0.878	0.878	0.942
	RF 400	0.878	0.756	0.878	0.878	0.942
	RF 500	0.878	0.756	0.878	0.878	0.943
	Random Tree	0.798	0.596	0.798	0.798	0.798
	J48	0.814	0.628	0.814	0.814	0.801
Standard Deviation	RF 100	0.185	0.0370	0.0185	0.0185	0.0130
	RF 200	0.187	0.0374	0.0187	0.0187	0.0124
	RF 300	0.177	0.0355	0.0177	0.0177	0.0123
	RF 400	0.177	0.0353	0.0177	0.0177	0.0124
	RF 500	0.177	0.0355	0.0177	0.0177	0.0123
	Random Tree	0.206	0.0413	0.0206	0.0206	0.0206
	J48	0.187	0.0374	0.0187	0.0187	0.0231

TABLE II
STATISTICAL MULTIPLE COMPARISON ANALYSIS AMONG RF 300, 400, AND 500 CLASSIFIERS USING ROC AREA

Classifier 1	Classifier 2	W	p
RF 300	RF 400	0.114	1
RF 300	RF 500	0.1192	1
RF 400	RF 500	0.0553	1

TABLE III
RESULTS OF THE SUPPLIED TEST SET (RF300)

Accuracy	Sensitivity	Specificity	AUC
0.821	0.821	0.822	0.865

TABLE IV
CONFUSION MATRIX OF THE RF 300 CLASSIFIER FOR THE TEST SET

	Predicted Ictal	Predicted Interictal
Actual Ictal	70.2% (936)	29.8% (397)
Actual Interictal	5.9% (78)	94.1% (1243)

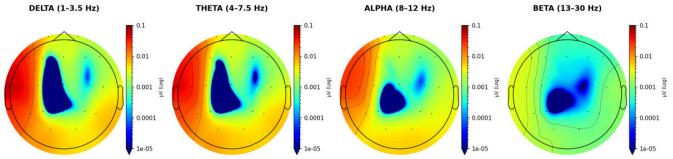


Fig. 4. Ictal activity mapping of a subject with a diagnosed left-hemisphere epileptic focus.

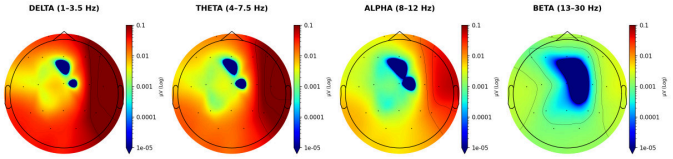


Fig. 5. Ictal activity mapping of a subject with a diagnosed right-hemisphere epileptic focus.

of these results lies in seizure prediction, which can provide patients with enhanced safety, improved quality of life, and the potential for early therapeutic intervention.

Furthermore, the integration of machine learning techniques with high-performance computing enables the development of personalized models tailored to the specific neurophysiological characteristics of each patient. This approach not only enhances classification performance but also facilitates integration into real-time monitoring and alert systems for both patients and caregivers.

Despite the promising performance achieved by the proposed framework, some methodological aspects could be further investigated. Future work should improve: the use of a window duration greater than 500 ms to include the minimum frequencies of the delta band, without compromising the number of frames available to construct the dataset used in Weka; the normalization of the global EEG signal voltage scale for each patient to obtain more precise results; and the use of feature extraction strategies specifically adapted to EEG spatial and spectral representations. In particular, fine-tuning the ResNet-50 network using EEG-derived topographic maps could allow the learned representations to become more sensitive to neurophysiological patterns that may not be adequately captured by features learned exclusively from natural images. In addition, comparative experiments involving ImageNet-pre-trained models, EEG-specific architectures, and domain-adapted convolutional neural networks could help determine the most appropriate feature representation for epileptic-state classification.

IV. CONCLUSION

This study presented a hybrid architecture for classifying epileptic states based on EEG signals, using topographic mapping derived from the Continuous Wavelet Transform (CWT), pre-trained deep neural networks for image feature extraction, and a Random Forest classifier. Through topographic mapping,

it was possible to visualize brain activity in specific time windows across various frequency bands. This provided an informative and well-structured topographic representation for application in deep learning applications.

The quantitative and qualitative results demonstrate that the hybrid architecture proposed in this study, which integrates topographic mapping using wavelet transforms with pre-trained neural networks for image feature extraction, combined with a Random Forest classifier comprising 300 trees, achieves significant performance. Promising results were obtained, including an accuracy of 82.1% and an area under the ROC curve of 0.87, indicating a balanced overall discriminatory capacity in distinguishing between ictal and interictal states. Notably, the interictal state achieved a higher detection rate of approximately 94.1%. In summary, the results indicate that the proposed framework is a promising approach to automatic identification of epileptic states from EEG signals, demonstrating the feasibility of combining wavelet-based topographic mapping, deep feature extraction, and machine learning techniques for EEG analysis.

Although the proposed framework achieved promising results, its evaluation was restricted to a single public dataset. Future work should investigate its performance on larger and more diverse EEG databases, as well as explore architectures specifically designed for biomedical signals. Such efforts may contribute to improving model generalization and facilitating its future translation into clinical decision-support systems.

REFERENCES

[1] W. H. Organization *et al.*, "Epilepsy: a public health imperative," 2019.

[2] A. Kumar, E. T. Ighodaro, and S. Sharma, "Focal impaired awareness seizure," Treasure Island, FL, USA, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK519030/

[3] S. J. M. Smith, "Eeg in the diagnosis, classification, and management of patients with epilepsy," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 2, pp. ii2–ii7, 2005.

[4] S. Noachtar and J. Rémi, "The role of eeg in epilepsy: A critical review," *Epilepsy & Behavior*, vol. 15, no. 1, pp. 22–33, 2009.

[5] U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated eeg analysis of epilepsy: A review," *Knowledge-Based Systems*, vol. 45, pp. 147–165, 2013.

[6] A. Subasi, "Eeg signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084–1093, 2007.

[7] M. R. Arab, A. A. Suratgar, and A. R. Ashtiani, "Electroencephalogram signals processing for topographic brain mapping and epilepsies classification," *Computers in Biology and Medicine*, vol. 40, no. 9, pp. 733–739, 2010.

[8] P. Detti, G. Vatti, and G. Z. Lara, "Siena scalp eeg database (version 1.0.0)," 2020. [Online]. Available: https://physionet.org/content/siena-scalp-eeg/1.0.0/

[9] P. Arpaia *et al.*, "Assessing the role of eeg biosignal preprocessing to enhance multiscale fuzzy entropy in alzheimer's disease detection," *Biosensors*, vol. 15, no. 6, p. 374, 2025.

[10] H. H. Jasper, "The ten-twenty electrode system of the international federation," *Electroencephalography and Clinical Neurophysiology*, vol. 10, pp. 371–375, 1958.

[11] S. Sanei and J. A. Chambers, *EEG signal processing*. John Wiley & Sons, 2013.

[12] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction—ready for a new era," *Nature Reviews Neurology*, vol. 14, no. 10, pp. 618–630, 2018.

[13] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 5, pp. 804–813, 2019.

[14] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen *et al.*, "Meg and eeg data analysis with mne-python," *Frontiers in Neuroinformatics*, vol. 7, p. 267, 2013.

[15] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[16] D. Zhang, L. Yao, X. Chen, S. Wang, J. Chang, and K. Fang, "Learning spatial–spectral–temporal eeg features with recurrent 3d convolutional neural networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 1, pp. 31–42, 2019.

[17] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from eeg with deep recurrent-convolutional neural networks," in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: https://arxiv.org/abs/1511.06448

[18] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for eeg motor imagery classification based on multi-layer cnns feature extraction," *Future Generation Computer Systems*, vol. 90, pp. 276–283, 2019.

[19] B. Direito, C. Teixeira, B. Ribeiro, M. Castelo-Branco, F. Sales, and A. Dourado, "Modeling epileptic brain states using eeg spectral analysis and topographic mapping," *Journal of Neuroscience Methods*, vol. 210, no. 2, pp. 220–229, 2012.

[20] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[21] A. Clark *et al.*, "Pillow (pil fork) documentation," *readthedocs*, 2015.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[25] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad *et al.*, "Epileptic seizures detection using deep learning techniques: a review," *International journal of environmental research and public health*, vol. 18, no. 11, p. 5780, 2021.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

Pix2Pix GAN-Based Translation of Dermoscopic Images into Clinical Skin Lesion Images: VGG16-Guided Losses and FID Evaluation

1st Villeneve de O. Soares
Digital Signal Processing Group – GPDS
Federal Institute of Paraíba – IFPB
João Pessoa, Brazil
villeneveoliveira@gmail.com

2nd Pedro A. de A. da Silva
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
pedrotatico1321@gmail.com

3rd Emanuel T. de A. Silva
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
emanuel.thiago.2018@gmail.com

4th Gabriel A. de Santana
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
andrade.santana@academico.ifpb.edu.br

5th Carlos D. M. Regis
Digital Signal Processing Group - GPDS
Federal Institute of Paraíba - IFPB
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

6th Leonardo V. Batista
Digital Video Applications Laboratory – LAVID
Federal University of Paraíba - UFPB
João Pessoa, Brazil
leonardo@ci.ufpb.br

***Abstract*—The simultaneous availability of clinical and dermoscopic skin lesion images is relevant for multimodal dermatological analysis, but paired datasets remain limited. This work investigates the translation of dermoscopic images into clinical skin lesion images using a Pix2Pix-based generative adversarial framework trained on paired samples from the MILK10k dataset. The proposed model uses an encoder-residual-decoder generator and a PatchGAN-like discriminator, combined with adversarial supervision and VGG16-based perceptual losses extracted from different feature layers. The generated images were evaluated using the Frechet Inception Distance to quantify their distributional similarity to real clinical images. The results show that perceptual losses based on earlier VGG16 layers achieved lower FID values than the evaluated MAE baseline when compared with a pixel-wise MAE baseline. The best configuration, vgg16.10L2, achieved the lowest FID score of 67.47, followed by vgg16.L2 with an FID of 69.74, while the MAE baseline obtained an FID of 82.50. Qualitative analysis indicated that the proposed approach can generate plausible clinical images and showed visually plausible lesion appearance in successful cases, although failure cases revealed limitations related to lesion morphology, localization, and semantic consistency. These findings suggest that feature-level supervision is a promising strategy for dermoscopic-to-clinical image translation, while also highlighting the need for clinically oriented evaluation metrics in future studies.**

***Index Terms*—Dermoscopic images, clinical skin lesion images, image-to-image translation, generative adversarial networks, Pix2Pix, perceptual loss, VGG16, Frechet Inception Distance, medical image synthesis, dermatology.**

I. Introduction

Skin cancer is among the common neoplasms worldwide, and early detection is closely associated with better clinical outcomes and reduced patient morbidity [1]. In dermatological practice, imaging methods support the visual assessment of lesion morphology before histopathological confirmation [2]. In recent years, deep learning techniques have been explored in dermatology for tasks such as classification, segmentation, and computer-aided diagnosis, with studies reporting specialist-level performance in specific diagnostic scenarios [3]–[5].

Clinical and dermoscopic images are two used modalities in skin lesion assessment. Clinical images capture the macroscopic appearance of the lesion and its surrounding tissue, resembling the visual inspection performed during routine examination. Dermoscopic images, in turn, are acquired using optical systems that reduce surface reflection and reveal pigmentary structures, vascular patterns, and other subsurface details that are not visible to the naked eye [2]. These modalities may provide complementary information for dermatological assessment, motivating the development of computational methods capable of exploiting both image domains [5].

Despite this complementarity, the availability of paired clinical and dermoscopic images remains limited when compared with datasets focused on a single modality, especially dermoscopy [5], [6]. This limitation constrains the development of multimodal models, image fusion strategies, and supervised translation methods between dermatological imaging domains. Recent multimodal initiatives, such as MILK10k, provide paired clinical and dermoscopic images and enable the investigation of learning-based approaches for cross-modality image synthesis in skin lesion analysis [7].

In this context, image-to-image translation has emerged as a suitable framework for learning mappings between different visual domains [8], [9]. GAN-based approaches are particularly relevant for this task, since they combine a generative model with an adversarial training strategy to synthesize images with realistic visual characteristics [10]. Pix2Pix introduced a supervised conditional GAN formulation for paired image translation [8], whereas CycleGAN extended this idea to unpaired domains through cycle-consistency constraints [9]. These approaches have also been adapted to medical

imaging, where preserving anatomically or clinically structures is essential during the synthesis process [11].

However, translating dermoscopic images into clinical images remains challenging due to differences in illumination, texture, field of view, scale, hair visibility, and anatomical context [5]. In addition, losses based only on pixel-wise differences may produce overly smoothed outputs and fail to preserve perceptually details [12]. Perceptual losses address this limitation by comparing feature representations extracted from pretrained convolutional neural networks, allowing the optimization process to consider image characteristics at different levels of abstraction [12], [13].

Therefore, this work investigates dermoscopic-to-clinical image translation using a Pix2Pix-based conditional GAN trained on paired multimodal skin lesion images. The method incorporates a VGG16-based perceptual loss, and different feature extraction layers are evaluated to analyze how distinct representation levels influence the quality of synthesized clinical images. The generated images are quantitatively assessed using the Fréchet Inception Distance (FID), providing a comparative analysis of the effect of perceptual representations on the translation from dermoscopic to clinical images [14].

II. Methods

A. Dataset and Preprocessing

MILK10k is a large-scale multimodal skin lesion dataset developed to advance artificial intelligence research in dermatology. It contains 10,480 images corresponding to 5,240 skin lesions, where each lesion is represented by a paired clinical close-up photograph and a dermoscopic image [7].

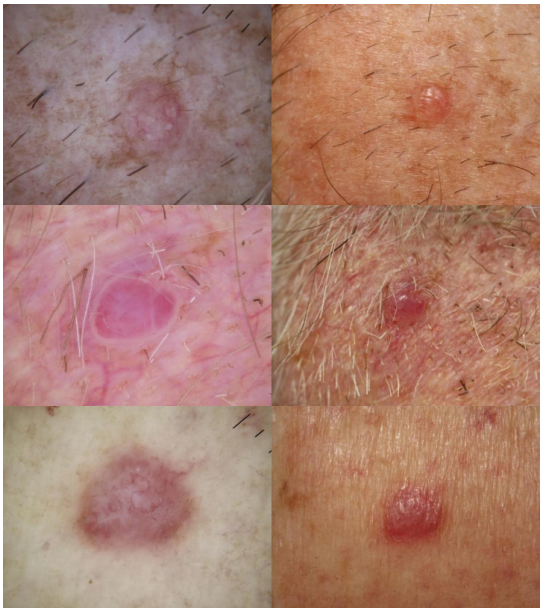


Fig. 1. Representative samples from the dataset. The left column contains dermoscopic images of skin lesions, and the right column shows the corresponding clinical images of the same lesions.

The dataset was collected from five international centers and includes metadata such as patient age, sex, skin tone, anatomical site, and diagnostic information. Importantly, more than 95% of the lesions were confirmed through histopathological examination, providing a reliable ground truth for machine learning studies [7].

Figure 1 presents examples from the MILK10k dataset. Each pair consists of a dermoscopic image and its corresponding clinical photograph acquired from the same skin lesion.

All images in the MILK10k dataset were originally stored in a $4:3$ aspect ratio. During preprocessing, each image was resized while preserving its aspect ratio such that the smallest spatial dimension measured 142 pixels. A random crop of size (128×128) was extracted, producing RGB images with final dimensions of $(128 \times 128 \times 3)$. In addition to standardizing the input size, the random cropping operation acted as a lightweight data augmentation strategy, introducing small spatial variations that helped mitigate overfitting during training. Finally, pixel intensities, originally represented in the $([0, 255])$ range, were normalized to have zero mean and unit standard deviation.

B. Proposed Dermoscopic-to-Clinical Translation Framework

Figure 2 presents the generator and discriminator architectures used in this work. The overall framework follows the generative adversarial learning paradigm originally introduced in [10], in which a generator learns to synthesize images while a discriminator is trained to distinguish generated samples from real ones. In this work, the generator is designed for image-to-image translation, following the same general task formulation explored in conditional adversarial models such as Pix2Pix [8], CycleGAN [9], and medical image translation approaches such as MedGAN [11].

The generator follows an encoder–residual–decoder design. The encoder receives an RGB image and progressively reduces its spatial resolution through convolutional blocks with stride 2. Each block is composed of a convolutional layer followed by instance normalization and ReLU activation. This structure is closely related to the convolutional encoders commonly adopted in image-to-image translation networks, where spatial downsampling is used to map the input image into a compact feature representation before reconstruction [8], [9]. This stage maps the input image to a latent representation with increased feature depth, allowing the network to extract structural and textural information from the source domain.

The latent stage is composed of five residual convolutional blocks. The use of residual blocks is consistent with architectures adopted in neural image transformation and translation models, where residual mappings are used to refine feature representations while preserving information through identity connections [9], [12]. Each residual block applies a sequence of convolution, normalization, and nonlinear activation operations, while preserving an identity shortcut connection between the block input and output. The residual output is obtained through an element-wise summation, which helps preserve low-level information and improves gradient propagation during training. In this work, the residual blocks operate at the

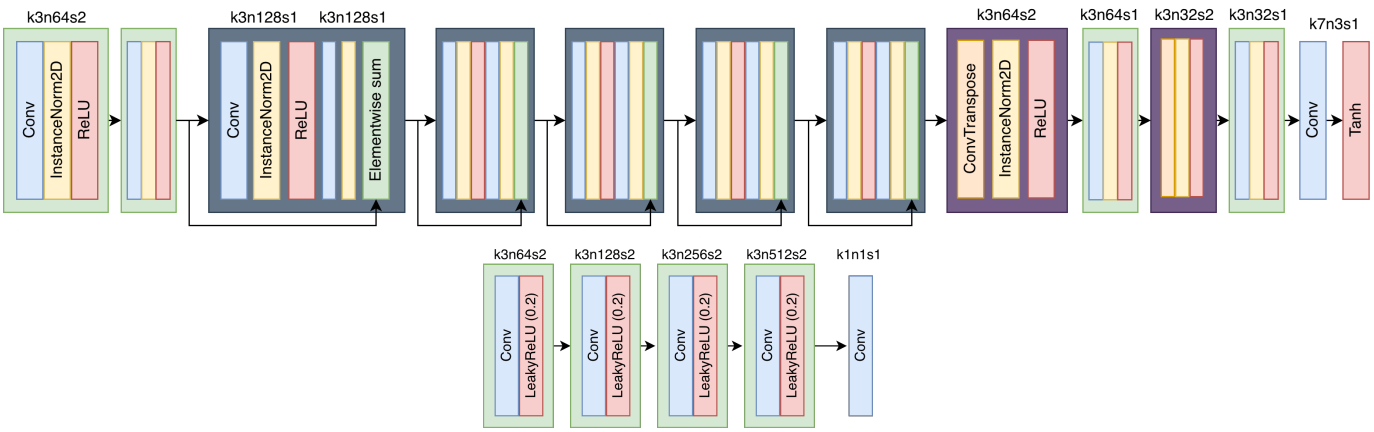


Fig. 2. Block diagram of the proposed model architecture. The upper diagram represents the generator architecture, while the lower diagram represents the discriminator architecture.

same spatial resolution and channel depth, refining the encoded representation before reconstruction.

The decoder reconstructs the output image from the latent representation using transposed convolutional layers for spatial upsampling. After each upsampling operation, convolutional blocks are used to refine the generated feature maps. This encoder–decoder organization is related to previous image-to-image translation architectures, in which the decoder progressively restores the spatial resolution of the feature maps to generate an image in the target domain [8], [9]. The final layer consists of a convolution followed by a hyperbolic tangent activation function, producing a three-channel RGB image with normalized pixel values. Overall, the generator is designed to transform the input image into the target domain while preserving anatomical and visual structures.

The discriminator follows a convolutional PatchGAN-like architecture, similar to the discriminator design used in conditional image-to-image translation models [8]. It receives an RGB image and progressively downsamples it through a sequence of convolutional layers with increasing numbers of feature channels. LeakyReLU activations are used after each intermediate convolution to improve gradient flow and allow nonzero responses for negative activations. The final convolutional layer maps the extracted features to a single-channel output, representing local real/fake decisions over image patches rather than a single global classification score. This design encourages the generator to synthesize realistic local textures and spatial details [5], [11].

C. Objective Functions and Training Strategy

The discriminator loss, defined in Eq. (1), follows the Least Squares GAN (LSGAN) formulation [15], encouraging the discriminator to assign values close to one for real clinical images y and zero for generated images $G(x)$.

$$\mathcal{L}_D = \mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[(D(y)-1)^2 + D(G(x))^2\right] \tag{1}$$

The adversarial loss of the generator, given by Eq. (2), encourages the generated images to be classified as real by the discriminator. In addition, the perceptual loss in Eq. (3) computes the mean absolute error (MAE) between the feature maps extracted from the ℓ-th layer of a pretrained VGG16 network for the real and generated images, where C_ℓ, H_ℓ and W_ℓ denote the number of channels, height, and width of the corresponding feature maps, respectively [14].

$$\mathcal{L}_{AdvG} = \mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[(D(G(x))-1)^2\right] \tag{2}$$

$$\mathcal{L}_{VGG}^{(\ell)} = \mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[\frac{1}{C_\ell H_\ell W_\ell}\left\|\phi_\ell(y)-\phi_\ell(G(x))\right\|_1\right] \tag{3}$$

Finally, the generator is optimized by minimizing the weighted sum of the adversarial and perceptual losses, as defined in Eq. (4), where λ controls the contribution of the perceptual term.

$$\mathcal{L}_G = \mathcal{L}_{AdvG} + \lambda \mathcal{L}_{VGG}^{(\ell)} \tag{4}$$

Differently from the original Pix2Pix formulation [8], the discriminator was not conditioned on the input dermoscopic image. This simplification was adopted to keep the experimental comparison focused on the perceptual loss configurations. Instead, it received only clinical-domain images, either real samples y or generated samples $G(x)$. The training flow is represented in Fig. 3.

The model was trained for 1000 epochs using a mini-batch size of 40. Each epoch consisted of 131 optimization steps, corresponding to approximately $131 \times 40 = 5240$ image pairs processed per epoch. Therefore, over the complete training procedure, the model processed approximately $5240 \times 1000 = 5{,}240{,}000$ image-pair presentations.

D. Evaluation Metrics

The quality of the generated clinical images was quantitatively evaluated using the Fréchet Inception Distance (FID). FID is used to assess the similarity between the distribution of real images and the distribution of generated images in a deep feature space [16].

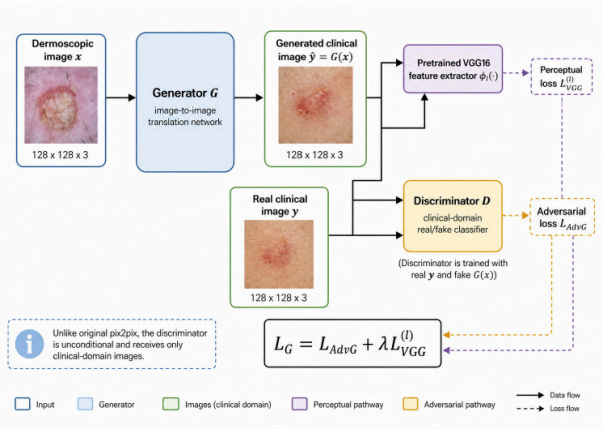


Fig. 3. Block diagram of the proposed generator training scheme for dermoscopic-to-clinical image translation.

Instead of comparing images directly at the pixel level, FID extracts feature representations from a pretrained Inception network and models the real and generated image sets as multivariate Gaussian distributions [16]. Lower FID values indicate that the generated images are closer to the real images in terms of feature statistics, suggesting better visual realism and distributional similarity.

Let (μ_r, Σ_r) denote the mean vector and covariance matrix of the feature representations extracted from real clinical images, and let (μ_g, Σ_g) denote the corresponding statistics extracted from generated clinical images. The FID is defined as

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \quad (5)$$

where $\|\cdot\|_2$ is the Euclidean norm and $\mathrm{Tr}(\cdot)$ denotes the trace operator. In this work, FID was used to compare the generated clinical-domain images against the corresponding distribution of real clinical images, providing an objective measure of how closely each trained model reproduced the target image domain.

III. Results

The notation used to identify each loss configuration indicates the type of reconstruction or perceptual constraint adopted during training. The term *MAE* denotes the pixel-wise mean absolute error baseline. Configurations identified as *vgg16* use perceptual losses computed from feature maps extracted by a pretrained VGG16 network. The suffixes *L2*, *L9*, and *L16* indicate the VGG16 layer used to extract the feature representation. When the prefix *10* is included, as in *vgg16.10L2*, the corresponding perceptual loss term is multiplied by a weight of 10 in the generator objective. Therefore, *vgg16.10L2* denotes a VGG16-based perceptual loss extracted from layer 2 with a weighting factor of 10.

The quantitative results obtained for the evaluated loss configurations are presented in Table I. Among the evaluated configurations, the best result was obtained by the model trained with the *vgg16.10L2* loss configuration, which achieved an FID of 67.47. The second-best result was obtained with *vgg16.L2*, with an FID of 69.74. These results suggest that using perceptual information extracted from earlier VGG16 layers contributed to the generation of images with greater distributional similarity to the target clinical domain.

TABLE I
FID scores obtained for the different loss configurations evaluated in the dermoscopic-to-clinical image translation task. Lower FID values indicate higher similarity between the generated and ground-truth clinical image distributions.

Loss configuration	FID
10MAE	82.50
vgg16.L2	69.74
vgg16.10L2	**67.47**
vgg16.L9	86.35
vgg16.10L9	75.06
vgg16.L16	77.55
vgg16.10L16	76.17

The baseline configuration based on the MAE loss achieved an FID of 82.50, indicating lower performance when compared with the best VGG16-based configurations. This behavior is consistent with the visual limitations often observed in pixel-wise reconstruction losses, which tend to favor average solutions and may reduce the preservation of perceptually details. In contrast, the use of perceptual losses based on VGG16 feature maps allowed the generator to compare images in a learned feature space, improving the visual correspondence between generated and real clinical images.

Figure 4 shows examples of successful dermoscopic-to-clinical image translation. In these cases, the generated images preserve aspects of the corresponding ground-truth clinical images, including the global skin appearance, lesion location, and local color patterns. Although differences in texture and sharpness can still be observed across the generated outputs, the models were generally able to synthesize plausible clinical images from the dermoscopic inputs. The results also show that different VGG16 layers influence the generated appearance in distinct ways. Configurations based on shallower layers, particularly *vgg16.L2* and *vgg16.10L2*, tended to produce images with better global consistency and fewer severe structural artifacts.

However, the qualitative analysis also revealed failure cases, as shown in Figure 5. In these examples, the generated images present deviations from the corresponding ground-truth clinical images. The observed errors include changes in lesion morphology, incorrect lesion localization, hallucinated pigmented structures, excessive texture smoothing, and inconsistent skin color.

The configurations associated with deeper VGG16 layers did not necessarily improve the synthesis quality. In particular, *vgg16.L9* produced the highest FID value among the evaluated models, reaching 86.35, while *vgg16.L16* and *vgg16.10L16* achieved FID values of 77.55 and 76.17, respectively. This suggests that features extracted from deeper layers may emphasize more abstract semantic representations while being

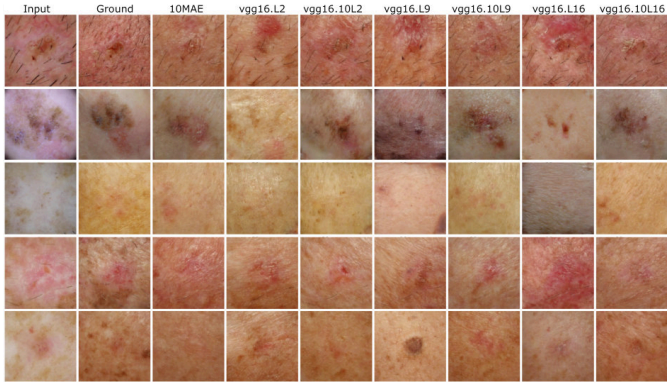


Fig. 4. Examples of successful dermoscopic-to-clinical image translation. Each row presents the input dermoscopic image, the ground-truth clinical image, and the images generated using different MAE and VGG16-based perceptual loss configurations. The selected cases show good preservation of lesion location, skin texture, and global clinical appearance across the synthesized outputs.

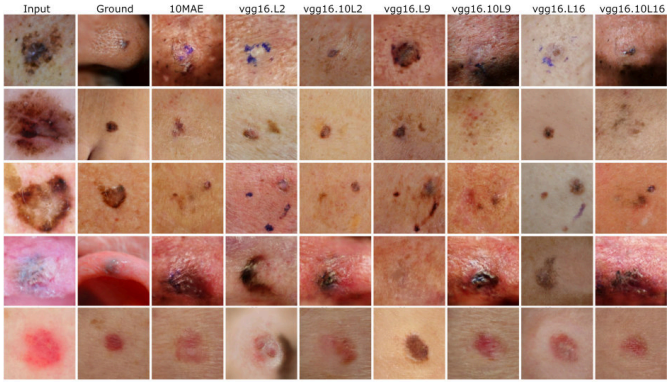


Fig. 5. Examples of failure cases in dermoscopic-to-clinical image translation. Each row presents the input dermoscopic image, the ground-truth clinical image, and outputs generated under different MAE and VGG16-based perceptual loss configurations. The selected cases highlight limitations such as lesion distortion, incorrect localization, hallucinated visual patterns, and loss of correspondence with the target clinical image.

less effective at preserving low-level visual characteristics such as texture, color distribution, and fine lesion boundaries. Since the target task requires the generation of visually realistic clinical images with coherent local appearance, the use of earlier perceptual representations appears to be more suitable.

Overall, the results indicate that the inclusion of VGG16-based perceptual losses improved the dermoscopic-to-clinical translation performance when compared with the MAE-based configuration alone. The best quantitative performance was achieved by *vgg16.10L2*, while the qualitative examples suggest that this configuration provides a better balance between visual realism, lesion preservation, and skin texture consistency. Nevertheless, the failure cases highlight that the proposed approach is still limited in scenarios involving modality discrepancies, complex lesion structures, or differences in field of view between dermoscopic and clinical images.

IV. Limitations

This study has several limitations. First, the evaluation is based primarily on FID and qualitative inspection; therefore, clinical usefulness, diagnostic consistency, and preservation of clinically relevant morphology were not directly validated. Second, no dermatologist-based evaluation or downstream diagnostic task was performed. Third, the study does not provide cross-dataset validation or comparison with contemporary image-to-image translation architectures. Fourth, the isolated effect of the simplified discriminator was not evaluated through an ablation study.

Finally, although successful cases showed plausible outputs, failure cases included changes in lesion morphology, localization, color, and hallucinated structures, indicating that the generated images should not be interpreted as clinically validated evidence.

V. Conclusion

The present study was conducted using the MILK10k dataset and does not evaluate cross-dataset generalization. Therefore, the reported results should not be interpreted as evidence of robustness across different acquisition domains or clinical settings. The objective of this study is not to establish a comprehensive benchmark among image-to-image translation architectures, but to investigate the effect of different perceptual loss configurations within a Pix2Pix-based dermoscopic-to-clinical translation framework.

This work investigated the generation of clinical skin lesion images from dermoscopic images using a Pix2Pix-based generative adversarial framework. The proposed approach explored the use of perceptual losses extracted from different layers of a pretrained VGG16 network, aiming to evaluate how distinct levels of visual representation influence the quality of the synthesized images. The results indicate that the choice of the perceptual layer affects both the visual appearance of the generated samples and the quantitative performance measured by FID, reinforcing the relevance of feature-level supervision in medical image-to-image translation tasks.

Future work should investigate architectural modifications that increase the generator's ability to perform semantic transformations. In the current formulation, allowing the generator to operate at a lower spatial resolution, such as a 16×16 latent representation, may provide greater capacity to modify global structures and contextual attributes of the image, rather than focusing predominantly on local texture reconstruction. This modification could improve the model's ability to synthesize clinical images whose overall organization is more consistent with the target domain.

Another important direction is to approximate the model more closely to the original Pix2Pix formulation by providing the real input image to the discriminator. In this configuration, the discriminator would evaluate not only whether the generated clinical image appears realistic, but also whether it is conditionally coherent with the corresponding dermoscopic image. This could encourage the generator to preserve lesion-

specific characteristics more consistently during the translation process.

In addition, future evaluations should go beyond general image quality metrics such as FID. Although FID is useful for estimating distributional similarity, it does not directly measure whether clinically relevant attributes were preserved. For this reason, a promising direction is to use a neural network trained specifically for dermatological pathology classification to verify whether the diagnostic semantics of the lesion remain consistent before and after image translation. Such an analysis would provide a more clinically meaningful assessment of the generated images.

Other future improvements include evaluating larger and more diverse datasets, performing cross-dataset validation, comparing the proposed model with alternative architectures such as CycleGAN, attention-based GANs, and diffusion models, and incorporating additional metrics such as LPIPS, SSIM, or task-oriented classification consistency. Furthermore, qualitative assessment by dermatology specialists could provide valuable insight into the clinical plausibility and potential utility of the synthesized images. These extensions may contribute to the development of more reliable multimodal image synthesis systems for dermatological applications.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] A. A. Marghoob, J. Malvehy, and R. P. Braun, eds., *Atlas of Dermoscopy*. CRC Press, 2 ed., 2012.

[3] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] Y. Liu *et al.*, "Deep learning in dermatology: opportunities and challenges," *Journal of the American Academy of Dermatology*, 2020.

[5] A. Bissoto, E. Valle, and S. Avila, "Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review," *Computerized Medical Imaging and Graphics*, 2021.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018.

[7] T. Philipp, A. N. Bengü, R. Cliff, R. Veronica, T. Verche, W. Jochen, W. A. Katharina, M. Christoph, K. Nicholas, H. Allan, *et al.*, "Milk10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators," *Journal of Investigative Dermatology*, 2025.

[8] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[9] J.-Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[11] K. Armanious *et al.*, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, 2020.

[12] J. Johnson *et al.*, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[13] R. Zhang *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," *CVPR*, 2018.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2794–2802, 2017.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

Potenciais biomarcadores de ressonância magnética funcional no comprometimento cognitivo leve e na doença de Alzheimer: uma revisão de escopo

Iasmim Cruz[†]
Programa de Pós-Graduação em
Engenharia Biomédica – PPGEB
Faculdade de Ciências e
Tecnologias em Engenharia – FCTE
Universidade de Brasília – UnB
[†]E-mail: nadinidacruz@gmail.com

Cristiano Jacques Miosso
Programa de Pós-Graduação em
Engenharia Biomédica – PPGEB
Faculdade de Ciências e
Tecnologias em Engenharia – FCTE
Universidade de Brasília – UnB

***Resumo*—A doença de Alzheimer (DA) é uma condição neurodegenerativa progressiva associada a alterações estruturais, metabólicas e funcionais. A ressonância magnética funcional (RMf) permite investigar, de maneira não invasiva, a atividade e a conectividade de redes envolvidas no comprometimento cognitivo leve (CCL) e na DA. Esta revisão de escopo teve como objetivo mapear achados de RMf investigados como potenciais biomarcadores nessas condições, caracterizando populações, paradigmas de aquisição, métodos de análise, regiões e redes estudadas e lacunas para validação clínica. Foram consultadas as bases SciELO, PubMed e Google Acadêmico, com a última busca em 20/03/2026. Os cinco estudos primários clínicos totalizaram 451 participantes declarados, embora possa haver sobreposição entre amostras provenientes das mesmas bases de dados. Os achados mapeados concentraram-se na rede de modo padrão, no hipocampo, em circuitos subcorticais e em redes do tronco encefálico e do cerebelo. A literatura descreve tanto aumento quanto redução da conectividade no CCL e predomínio de hipoconectividade na DA, mas a heterogeneidade de populações, protocolos e métodos analíticos impede interpretar esse comportamento como trajetória universal. Os resultados referem-se predominantemente a diferenças entre grupos e devem ser considerados candidatos a biomarcadores, ainda sem validação suficiente para emprego diagnóstico individual. Estudos multicêntricos, longitudinais e padronizados devem estabelecer reprodutibilidade e utilidade clínica.**

***Index Terms*—doença de Alzheimer, comprometimento cognitivo leve, imagem por ressonância magnética funcional, conectividade funcional, biomarcadores**

I. Introdução

As demências são doenças neurodegenerativas progressivas que comprometem a cognição e a funcionalidade [1]. A doença de Alzheimer (DA) destaca-se como a forma mais prevalente, sendo caracterizada pela perda neuronal, redução sináptica e atrofia cerebral [2].

Clinicamente, a DA envolve comprometimento da memória, alterações comportamentais e declínio das funções executivas [3]. A progressão da doença ocorre em estágios, desde alterações iniciais até incapacidade severa [4].

Do ponto de vista fisiopatológico, a DA está associada ao acúmulo de placas de beta-amiloide e à formação de emaranhados neurofibrilares de proteína tau, resultando em disfunção sináptica e morte neuronal [5]. Esses processos iniciam-se décadas antes da manifestação clínica [6].

O comprometimento cognitivo leve corresponde a uma síndrome clínica em que o desempenho cognitivo se encontra abaixo do esperado, sem perda funcional suficiente para caracterizar demência. Pode ser classificado em apresentações amnésticas ou não amnésticas, de domínio único ou de múltiplos domínios [7], [8]. O CCL não possui necessariamente etiologia da doença de Alzheimer: parte dos indivíduos permanece estável, retorna a um desempenho considerado normal ou progride para outras condições. Ainda assim, apresentações amnésticas estão associadas a maior probabilidade de progressão para demência por DA [9]. Essa heterogeneidade torna indispensável informar os critérios clínicos e a eventual confirmação etiológica adotados em cada estudo de neuroimagem.

Também existem apresentações atípicas da DA, com predomínio de alterações linguísticas, visuoespaciais, comportamentais ou executivas. Entre elas estão a variante logopênica da afasia progressiva primária, a atrofia cortical posterior e a variante comportamental/disexecutiva [10]. A inclusão dessas apresentações sem estratificação pode contribuir para a variabilidade dos padrões relatados.

O crescimento da população idosa e a elevada dependência do Sistema Único de Saúde (SUS) reforçam a relevância de estratégias diagnósticas reprodutíveis e acessíveis no Brasil [11]–[14]. Biomarcadores moleculares obtidos por tomografia por emissão de pósitrons (PET) podem contribuir para a caracterização da patologia amiloide e tau, mas seu acesso permanece limitado em muitos contextos. A escolha dos exames deve ser compreendida como parte de uma avaliação integrada, e não como substituição isolada da avaliação clínica e neuropsicológica [15], [16].

Nesse contexto, a ressonância magnética (RM) assume papel estratégico por sua disponibilidade e ausência de radiação ionizante, fornecendo informações estruturais, funcionais e

metabólicas do sistema nervoso central [17].

Entre suas modalidades, a ressonância magnética funcional (RMf) destaca-se pela possibilidade de investigar a conectividade funcional cerebral de maneira não invasiva, permitindo a análise integrada de redes neurais envolvidas nos processos cognitivos afetados pela DA [6], [17].

A RMf baseada no contraste dependente do nível de oxigenação sanguínea (BOLD) fornece uma medida indireta da atividade neural, mediada pelo acoplamento neurovascular. Na aquisição em repouso, correlações temporais espontâneas do sinal permitem estimar a conectividade funcional entre regiões. Alterações na rede de modo padrão (RMP) e em outras redes de larga escala têm sido descritas no CCL e na DA, embora sua direção e magnitude variem conforme a população, o estágio clínico e o método analítico [6], [18].

Tradicionalmente, os estudos de neuroimagem na doença concentraram-se no lobo temporal medial, particularmente no hipocampo e no córtex entorrinal. Entretanto, evidências recentes indicam que a DA não deve ser compreendida exclusivamente como uma doença cortical ou hipocampal, mas como uma desordem de redes distribuídas [17], [19].

Alterações na RMP, nos circuitos talamocorticais, em núcleos monoaminérgicos do tronco encefálico e em regiões cerebelares associadas à cognição têm sido investigadas em estudos de RMf em repouso [20]–[23].

A relevância dessa abordagem é ampliada pelo fato de que a aquisição em repouso dispensa a execução de tarefas cognitivas complexas, permitindo sua aplicação em indivíduos com diferentes graus de comprometimento cognitivo [17]. Estudos investigam se a conectividade pode revelar alterações nas fases iniciais, mas sua utilidade diagnóstica individual permanece incerta [18], [19].

Diante desse cenário, esta revisão de escopo tem como objetivo mapear os achados de RMf investigados como potenciais biomarcadores no CCL e na DA, caracterizando as populações estudadas, os paradigmas de aquisição, os métodos de análise, as redes e regiões envolvidas, as direções das alterações funcionais e as principais lacunas para sua validação clínica. Secundariamente, discutem-se os requisitos que ainda limitam sua potencial aplicabilidade no contexto da saúde pública brasileira.

II. Metodologia

Foi conduzida uma revisão de escopo orientada pelas recomendações metodológicas do JBI e relatada conforme a extensão PRISMA para revisões de escopo (PRISMA-ScR) [24], [25]. O processo de identificação, seleção e elegibilidade dos estudos foi estruturado com base nas recomendações da declaração PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), utilizando seus princípios como guia para a organização e transparência das etapas da revisão, sem adoção integral de todos os seus itens e procedimentos.

A pergunta e os critérios de elegibilidade foram organizados segundo a estratégia População–Conceito–Contexto (PCC). A população compreendeu adultos cognitivamente saudáveis, com CCL ou com DA; o conceito correspondeu a medidas obtidas por RMf/BOLD investigadas como achados funcionais ou potenciais biomarcadores; e o contexto incluiu estudos clínicos e de pesquisa sem restrição geográfica. A pergunta principal foi: "Quais achados de ressonância magnética funcional têm sido investigados como potenciais biomarcadores em indivíduos com comprometimento cognitivo leve e doença de Alzheimer?"

Foram elegíveis estudos sobre a progressão da doença, com seres humanos na faixa etária a partir de 60 anos de idade, que utilizaram RMf em repouso ou com tarefa e incluíram participantes com CCL ou DA, com ou sem comparação com controles cognitivamente saudáveis. Estudos multimodais foram considerados apenas quando os resultados de RMf puderam ser extraídos separadamente. Foram excluídos trabalhos exclusivamente baseados em RM estrutural, difusão, espectroscopia, arterial spin labeling (ASL), PET ou SPECT, sem componente de RMf. Foram aceitas revisões sistemáticas e revisões da literatura. Estudos de aprendizado de máquina foram incluídos como fontes de evidências experimentais sobre biomarcadores.

Os estudos selecionaram participantes que atenderam aos critérios do National Institute on Aging–Alzheimer's Association (NIA-AA), com base em testes de rastreio cognitivo, como a Avaliação Cognitiva de Montreal (MoCA) e o Mini-Exame do Estado Mental. Alguns estudos relataram a investigação de biomarcadores no líquido cefalorraquidiano, como Aβ1–42, t-tau e p-tau, como ocorreu no estudo de Tang et al. [23].

As buscas foram realizadas em bases como PubMed e SciELO e complementadas pelo Google Acadêmico, com filtro para um período de 5 a 10 anos, dependendo da disponibilidade de temas recuperados, utilizando os termos de busca: Neuroimage AND biomarkers AND magnetic Alzheimer, functional magnetic resonance AND Alzheimer disease, cerebellar alzheimer's disease MRI. A busca compreendeu publicações entre 2016 e 2026, nos idiomas inglês e português, e foi atualizada pela última vez em 20/03/2026.

A estratégia combinou termos controlados e palavras livres relacionadas à DA, ao CCL e à RMf. Na PubMed, foram feitas buscas nos campos título e resumo, com os operadores AND e OR nas seguintes expressões: fMRI OR functional magnetic resonance AND Alzheimer's disease, (Alzheimer's disease) AND (functional MRI), cerebellum MRI AND alzheimer disease, com filtro de tempo para 10 anos. A consulta sobre barreira hematoencefálica foi retirada por não corresponder diretamente ao conceito desta revisão.

Os registros recuperados foram exportados para o gerenciador de referências Mendeley, no qual foram removidas as duplicatas. A seleção, realizada pela autora principal, ocorreu em duas etapas: triagem de títulos e resumos e avaliação dos textos completos. As dúvidas de elegibilidade foram discutidas com o segundo autor, e os motivos de exclusão na etapa de texto completo foram registrados.

Elaborou-se formulário de extração com autores, ano, país, desenho, grupos, tamanho amostral, critérios diagnósticos,

Tabela I. Estratégias documentadas pertinentes ao recorte funcional. Os totais exibidos pelos buscadores não foram utilizados como números de registros triados.

Fonte	Expressão e condição documental
PubMed	(Alzheimer's disease) AND (functional MRI)
PubMed	fMRI OR functional magnetic resonance AND Alz.heimer's disease. Grafia e ausência de agrupamento preservadas do documento de origem.
PubMed	cerebellum MRI AND alzheimer disease
PubMed, estratégia detalhada	`("Alzheimer Disease"[MeSH Terms] OR Alzheimer*[Title/Abstract]) AND ("functional magnetic resonance imaging"[Title/Abstract] OR fMRI[Title/Abstract] OR ((MRI[Title/Abstract] OR "magnetic resonance"[Title/Abstract]) AND ("resting-state"[Title/Abstract] OR "resting state"[Title/Abstract] OR "functional connectivity"[Title/Abstract]))) AND ("2016/01/01"[Date - Publication] : "2026/08/22"[Date - Publication]) AND (english[Language] OR portuguese[Language] OR spanish[Language])`
Google Acadêmico	Neuroimage AND biomarkers AND magnetic Alzheimer
Google Acadêmico	functional magnetic resonance AND Alzheimer disease
Google Acadêmico	cerebellar alzheimer's disease MRI
SciELO	Expressão integral e data específica não documentadas; consulta contextual, sem contribuição identificada para a síntese de resultados.

aquisição, paradigma de repouso ou tarefa, método analítico, região ou rede, comparação, direção do achado, associação clínica, desempenho diagnóstico e limitações.

O formulário, redigido pelos autores, buscou contemplar as regiões encefálicas descritas na literatura, o tipo de achado em RMf, o paradigma, isto é, se o exame foi realizado de forma dinâmica ou em repouso e quais interpretações os achados sugerem de acordo com a progressão da doença. Além disso, os achados foram classificados conforme sua relevância para o CCL e a DA.

Os estudos foram agrupados por paradigma, população, rede ou região e desfecho. A síntese foi descritiva, com contagens por categoria, sem metanálise.

III. Resultados

A. Seleção e características das fontes

Entre os estudos incluídos, 1 avaliou o CCL, 5 contemplaram DA e CCL e 8 artigos abordaram aspectos da DA. Quanto ao paradigma, 10 artigos mencionaram ressonância magnética funcional em repouso. A modalidade tem ganhado visibilidade por possibilitar o estudo das redes sem necessidade de esforço físico ou cognitivo do paciente. A síntese também informa a distribuição por método analítico, tamanho amostral e presença de confirmação da patologia da doença de Alzheimer. Foram selecionados estudos que abordassem a doença de Alzheimer e seus biomarcadores, contemplando participantes classificados como comprometimento cognitivo leve (CCL) ou doença de Alzheimer (DA). Além disso, os estudos elegíveis deveriam apresentar a descrição dos biomarcadores investigados e discutir a interpretação de seus achados, incluindo sua relevância e seu significado no contexto fisiopatológico da doença. A conferência das citações no texto de resultados e na tabela do manuscrito identificou 12 fontes únicas: seis publicações primárias e seis revisões. Entre as primárias,

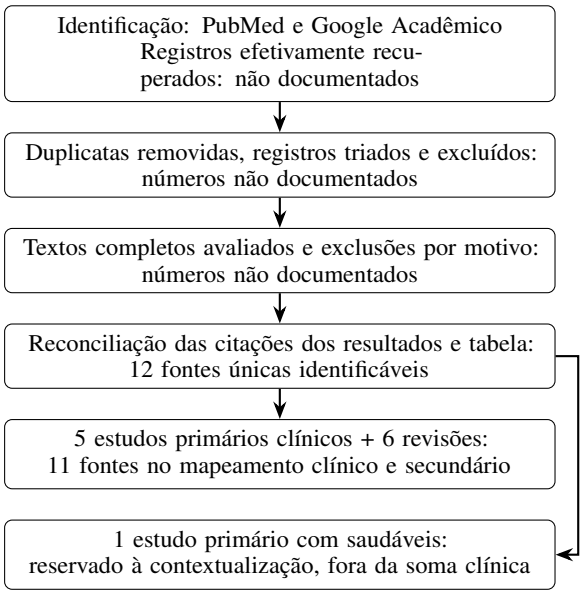


Figura 1. Fluxo retrospectivo de documentação e reconciliação. As contagens finais descrevem fontes identificáveis no manuscrito, não uma reprodução integral da seleção histórica. Não se inferiram exclusões a partir da diferença para totais anteriormente declarados.

cinco investigaram grupos clínicos [20], [22], [23], [26], [27]; a sexta avaliou adultos saudáveis e foi reservada à contextualização [28]. Assim, o mapeamento clínico e sua síntese secundária compreendem 11 fontes, discriminadas nas Tabelas II e IV. A Figura 1 distingue essa reconciliação do processo histórico de busca.

Os tamanhos explicitados na Tabela II somam 451 participantes entre estudos: $95+206+65+61+24$. Esse total combina os tamanhos declarados nas etapas disponíveis de

Tabela II. Estudos primários clínicos: amostras, diagnóstico e caracterização metodológica. C: controles; NC: não confirmado nas fontes acessíveis à conferência; LCR: líquido cefalorraquidiano.

Estudo	Desenho e amostra declarada	Diagnóstico e biomarcadores	Aquisição e análise
Wu et al. (2022) [20]	Estudo observacional comparativo; 48 DA + 47 C = 95 recrutados.	Avaliação de múltiplos domínios cognitivos; critérios diagnósticos detalhados e confirmação amiloide/tau: NC.	RMf; sementes em subdivisões talâmicas, correlações cognitivas e classificador SVM linear. Campo, TR, TE, resolução e esquema de validação: NC.
Serra et al. (2018) [22]	Comparação entre grupos; 169 pacientes com DA ou CCL amnéstico devido à DA + 37 C = 206. Divisão dos 169 por diagnóstico: NC.	Avaliação cognitiva e neuropsiquiátrica; confirmação por biomarcadores e critérios completos: NC.	RMf em repouso, 3 T; conectividade da área tegmental ventral e locus coeruleus com o restante do cérebro. TR, TE e resolução: NC.
Tang et al. (2021) [23]	Transversal; 27 DA + 25 CCL + 13 C = 65.	NIA-AA 2011; avaliação neuropsicológica; análise de LCR relatada nos pacientes. Coleta não foi equiparada a positividade universal.	Siemens 3 T, GRE-EPI; TR/TE = 2000/30 ms; 240 volumes; cortes de 4 mm; olhos fechados. DPABI, sementes Crus II e lóbulo IX bilaterais.
Liu et al. (2018) [26]	Análise original de dados ADNI; 64 iniciais; 61 analisados: 17 DA + 25 CCL + 19 C, após exclusões por movimento.	Grupos ADNI, MMSE e CDR descritos; positividade amiloide/tau como requisito: NC.	Philips 3 T, EPI; TR/TE = 3000/30 ms; 140 volumes; olhos abertos. Índice de lateralidade intrínseca (iLI), com 256 pares de regiões.
Klaassens et al. (2019) [27]	Estudo farmacológico; 12 DA + 12 C = 24 incluídos.	DA leve; pareamento por idade e sexo. Critérios completos e confirmação etiológica: NC.	RMf em repouso sob citalopram e galantamina; efeito farmacológico sobre conectividade. Parâmetros de aquisição e número utilizável por condição: NC.

cada publicação, utilizando a amostra pós-exclusão de Liu et al.; não representa uma amostra combinada uniformemente pós-controle de qualidade nem indivíduos comprovadamente distintos. Os 193 adultos saudáveis de Saiz-Masvidal et al. não integram essa soma clínica [28]. As revisões não possuem amostra primária própria atribuível a este artigo.

B. Evidências clínicas primárias

Os estudos investigam desfechos diferentes, resumidos na Tabela III. Wu et al. identificaram redução de conectividade das subdivisões intralaminar e medial do tálamo com o precuneus esquerdo. A associação com MMSE foi relatada para a conexão do núcleo medial, não indistintamente para todas as conexões talâmicas [20].

Serra et al. examinaram diferenças de conectividade da área tegmental ventral (ATV) e do locus coeruleus. Os contrastes clínicos sustentam envolvimento mesencefálico, mas não equivalem a medição de dopamina ou a demonstração longitudinal de conversão. O resultado negativo do locus coeruleus deve ser interpretado com as limitações espaciais reconhecidas pelos autores [22].

Tang et al. investigaram conexões de sementes cerebelares e encontraram diferenças dependentes da região e do grupo. A descrição preserva essa especificidade, sem atribuir aumento ou redução ao cerebelo inteiro [23]. Liu et al. avaliaram lateralização funcional global; o iLI não é uma medida volumétrica hipocampal e seus resultados não foram convertidos em perda anatômica localizada [26].

Klaassens et al. estudaram modulação farmacológica da conectividade. Seu desenho permite discutir responsividade de sistemas neuromodulatórios, mas deve permanecer separado de comparações observacionais basais e de estudos de acurácia diagnóstica [27].

C. Evidências secundárias e contextualização

As seis revisões têm funções distintas (Tabela IV). A literatura secundária descreve alterações da RMP e de circuitos relacionados à memória, além da heterogeneidade dos métodos e das populações. Os relatos de redução posterior e de aumentos regionais não foram contados como replicações independentes quando provenientes de revisões potencialmente sobrepostas [17]–[19].

Paradigmas com tarefa são discutidos nas revisões, mas os cinco estudos clínicos primários identificados neste conjunto não constituem uma amostra de experimentos de codificação e recuperação de memória. Não se atribuiu a Liu et al. um paradigma com tarefa. Também não se presumiu que toda alteração de ativação em tarefa correspondesse a alteração equivalente de conectividade em repouso [6], [17].

A revisão de Patthy et al. discute mecanismos monoaminérgicos, mas não permite atribuir à RMf a mensuração direta da disponibilidade de transportadores [21]. O estudo de Saiz-Masvidal et al., realizado em adultos saudáveis, foi utilizado apenas como contexto de mapeamento multimodal e não como demonstração de alteração clínica na DA [28].

D. RMf em repouso

Os achados de conectividade funcional concentraram-se na RMP, que inclui o córtex cingulado posterior, o precuneus e o córtex pré-frontal medial. Reduções de conectividade nessas regiões foram recorrentes em grupos com DA [17], [18]. Em grupos com CCL, foram descritos tanto aumentos quanto reduções regionais, comportamento interpretado em parte da literatura como possível reorganização compensatória [17], [19]. Como a maior parte dessas comparações é transversal, os dados não estabelecem uma trajetória universal de hiperconectividade seguida por hipoconectividade.

Tabela III. Correspondência entre evidência primária, achado funcional e limite interpretativo. A direção refere-se à comparação original, não a uma trajetória universal.

Fonte	Achado ou desfecho	Interpretação delimitada
Wu et al. [20]	Redução da conectividade tálamo–precuneus nos pacientes; classificação por SVM explorada.	Associação cognitiva e discriminação na amostra não demonstram utilidade clínica externa.
Serra et al. [22]	Desconexão da ATV predominantemente parietal no CCL amnéstico e em nós posteriores da RMP na DA; sem evidência de desconexão do locus coeruleus.	Comparação entre grupos, sem acompanhamento da conversão em 24 meses. Ausência de diferença não comprova integridade do núcleo.
Tang et al. [23]	Padrões distintos de conectividade de Crus II e lóbulo IX, com aumentos e reduções conforme a conexão e o contraste.	Evidência regional cerebelocortical; não comprova compensação ou progressão intraindividual.
Liu et al. [26]	Mudanças no padrão de lateralização funcional entre controles, CCL e DA.	Métrica global de conectividade em repouso; não corresponde à ativação hipocampal durante tarefa nem a uma medida estrutural.
Klaassens et al. [27]	Resposta da conectividade a intervenções colinérgica e serotoninérgica.	Evidência farmacológica; não demonstra redução direta de transportadores por RMf.

Tabela IV. Fontes secundárias identificadas nos resultados e na tabela original. Nenhuma acrescenta participantes primários próprios à soma clínica.

Fonte	Tipo de publicação	Função na síntese e limite
Forouzannezhad et al. (2019) [6]	Revisão de aplicações e métodos	Contextualiza paradigmas e processamento de RMf; não constitui uma coorte clínica adicional.
Zhang et al. (2023) [17]	Revisão sobre patogênese e imagem	Contextualiza achados de diferentes modalidades; resultados estruturais não foram convertidos em BOLD.
Ibrahim et al. (2021) [18]	Revisão sistemática	Sintetiza investigação diagnóstica por RMf em repouso; os estudos que a compõem não foram automaticamente incorporados como novas inclusões.
Jalilianhasanpour et al. (2019) [19]	Revisão	Contextualiza conectividade em doenças neurodegenerativas; associações genéticas não demonstram especificidade diagnóstica.
Patthy et al. (2021) [21]	Revisão neuropatológica e mecanística	Fundamenta o interesse por núcleos monoaminérgicos; não constitui evidência primária de redução rafe–precuneus por BOLD em pacientes.
Seoane et al. (2024) [29]	Revisão sistemática e metanálise	Amplia o mapeamento subcortical da RMP; sobreposição de estudos exige cautela ao interpretar convergência.

Alterações funcionais envolvendo hipocampo, giro parahipocampal e córtex cingulado posterior também foram relatadas [17], [18]. A redução de conectividade entre componentes posteriores e temporais mediais da RMP foi descrita tanto no CCL amnéstico quanto na DA. Em alguns contrastes, aumentos em regiões frontais coexistiram com reduções posteriores, mostrando que a direção do efeito depende da conexão e da métrica analisadas. O estudo de Liu et al. avaliou lateralização da conectividade funcional em repouso e encontrou mudanças nos padrões de lateralização entre controles, CCL e DA [26]. Medidas anatômicas de assimetria ou perda de substância cinzenta pertencem à RM estrutural e foram excluídas dos resultados BOLD.

Estudos de estruturas subcorticais relataram alterações de conectividade em subdivisões talâmicas, incluindo núcleos intralaminares e conexões com o precuneus, algumas associadas ao desempenho cognitivo [20]. A especificidade espacial desses resultados depende, contudo, do método de segmentação e da resolução empregada.

A área tegmental ventral e outros núcleos do tronco encefálico também foram investigados [21], [22]. Em estudo longitudinal com participantes com CCL atribuído à DA, aqueles que converteram para demência em 24 meses apresentaram, no início do seguimento, menor conectividade dirigida da área tegmental ventral com o córtex cingulado posterior e o córtex pré-central [22]. Esse resultado é relevante para o prognóstico, mas não autoriza generalizar a redução para todas as projeções dopaminérgicas ou para todos os indivíduos com CCL.

Quanto ao núcleo da rafe dorsal, estudos neuropatológicos, farmacológicos e mapas de conectividade em adultos saudáveis oferecem plausibilidade para o envolvimento de sistemas serotoninérgicos na DA [21], [27], [28]. Entretanto, a evidência para uma redução específica e reprodutível da conexão rafe–precuneus como marcador diagnóstico ainda é limitada. Em razão do pequeno volume desses núcleos e da resolução espacial convencional da RMf, os achados do tronco encefálico exigem cautela e replicação externa.

No cerebelo, alterações envolvendo Crus I, Crus II e lóbulo IX foram descritas no CCL e na DA [23], [29]. No estudo de Tang et al., o Crus II esquerdo apresentou diferenças de conectividade com tálamo, caudado e regiões frontais, temporais, límbicas e parietais; parte dessas alterações esteve associada ao desempenho cognitivo [23]. O lóbulo IX, por sua vez, participa da representação cerebelar da RMP; foram relatadas alterações em suas conexões com regiões parietais próximas ao precuneus [23], [29]. As direções dos efeitos variaram conforme a região-semente e o contraste, o que

impede resumir o envolvimento cerebelar como aumento ou redução global. Estudos exclusivamente volumétricos não foram considerados evidência funcional direta.

Paradigmas de codificação e recuperação da memória revelaram diferenças de ativação hipocampal entre controles, CCL e DA em ressonância magnética com tarefa [17]. No CCL, a literatura descreve aumento de ativação durante alguns paradigmas de codificação e redução durante a recuperação, enquanto grupos com DA apresentam, com maior frequência, menor recrutamento hipocampal. A aparente hiperativação foi interpretada como possível compensação, mas também pode refletir dificuldade da tarefa, desempenho desigual ou menor eficiência neural. Esses achados requerem interpretação separada da conectividade em repouso, pois dependem do paradigma, desempenho e contraste.

A Tabela V sintetiza os principais padrões funcionais por estágio e sistema cerebral. A versão detalhada, com todos os achados e limites interpretativos, está disponível na tabela suplementar online.

IV. Discussão

O mapeamento indica que a literatura de RMf no CCL e na DA se concentra na conectividade da RMP e de regiões relacionadas à memória. Alterações no córtex cingulado posterior, no precuneus, no córtex pré-frontal medial e no sistema temporal medial aparecem de forma recorrente [17], [18]. Isso sustenta a compreensão da DA como condição que envolve redes distribuídas, e não apenas o hipocampo. Contudo, a recorrência em comparações entre grupos não constitui validação diagnóstica individual.

O córtex cingulado posterior e o precuneus ocupam posição central na integração da RMP e se relacionam a processos de memória episódica e cognição autorreferencial. Sua desconexão do hipocampo e do giro parahipocampal pode indicar menor eficiência de comunicação entre os subsistemas posterior e temporal medial. O aumento da conectividade em regiões frontais, quando observado, pode representar recrutamento adicional diante da disfunção posterior; alternativamente, pode refletir perda de especialização e segregação entre redes. Essas interpretações não são mutuamente exclusivas e exigem associação longitudinal com desempenho cognitivo e carga patológica.

A coexistência de aumentos e reduções de conectividade, sobretudo no CCL, pode refletir diferenças de estágio, seleção de amostras, reserva cognitiva, carga neuropatológica e processamento dos dados. A interpretação da hiperconectividade como compensação é biologicamente plausível, mas permanece uma hipótese quando não há acompanhamento longitudinal ou relação direta com manutenção do desempenho. Da mesma forma, estudos transversais não demonstram que todos os indivíduos percorram uma sequência de hiperconectividade seguida por hipoconectividade [17], [19].

O envolvimento do tálamo, da área tegmental ventral e de outros núcleos do tronco encefálico amplia o modelo centrado em regiões corticais e hipocampais [20]–[22]. Entretanto, a segmentação de núcleos pequenos é vulnerável à resolução espacial, ao movimento, ao registro anatômico e aos efeitos de volume parcial. A evidência neuropatológica e a conectividade observada em adultos saudáveis oferecem plausibilidade biológica, mas não validam, isoladamente, uma conexão específica como biomarcador da DA.

A área tegmental ventral origina projeções dopaminérgicas mesocorticolímbicas relacionadas à memória, à motivação e ao comportamento. Alterações desse sistema podem contribuir para sintomas como apatia, anedonia e redução da motivação, além do declínio cognitivo [30]. O núcleo da rafe dorsal, por sua vez, constitui uma importante origem de projeções serotoninérgicas para regiões límbicas e corticais; sua disfunção tem sido relacionada a sintomas neuropsiquiátricos e distúrbios do sono [31]. A passagem da plausibilidade mecanística para um biomarcador requer demonstração independente de especificidade, repetibilidade e valor prognóstico.

As alterações descritas em Crus II e no lóbulo IX apoiam a participação do cerebelo em redes cognitivas [23], [29], [32]. A direção dos efeitos, porém, depende da região-semente e do contraste analisado. Por isso, expressões como "envolvimento cerebelar" são mais adequadas do que uma afirmação de aumento ou redução global da conectividade.

Esse conjunto de achados é compatível com a compreensão contemporânea de que regiões cerebelares posteriores participam de circuitos associados à memória de trabalho, linguagem, atenção e regulação afetiva [32]. O interesse no cerebelo amplia o campo para além das regiões corticais tradicionalmente enfatizadas, mas ainda há menos estudos e maior diversidade de sementes e métodos do que na investigação da RMP. Assim, os resultados cerebelares devem ser apresentados como uma frente promissora de pesquisa, e não como marcador consolidado.

Os estudos incluídos também devem ser interpretados à luz da heterogeneidade clínica. O CCL não possui necessariamente etiologia da doença de Alzheimer e pode permanecer estável ou decorrer de outras condições [7], [9]. Além disso, apresentações atípicas da DA podem produzir perfis funcionais diferentes da forma amnéstica [10]. Estudos futuros devem estratificar participantes por fenótipo e por evidência de patologia amiloide e tau.

A heterogeneidade metodológica é igualmente relevante. Análises baseadas em sementes dependem da definição da região de interesse; decomposição por componentes independentes, métricas de amplitude local e análises de grafos descrevem aspectos diferentes da função cerebral. Movimento de cabeça, duração da aquisição, estado de vigília, correção do sinal global, atrofia e fatores vasculares podem modificar as estimativas de conectividade. Um mapa de evidências deve tornar essas diferenças visíveis, evitando agrupar desfechos distintos como "conectividade alterada".

Para uso clínico, um potencial biomarcador deve demonstrar repetibilidade, sensibilidade, especificidade, validação externa e benefício adicional em relação à avaliação clínica, neuropsicológica e aos biomarcadores disponíveis. A RMf não utiliza radiação ionizante, mas requer aquisição padronizada, controle rigoroso de movimento, processamento especializado e validação multicêntrica. Assim, a maior disponibilidade de

Tabela V. Versão resumida dos principais achados funcionais mapeados no CCL e na DA. A tabela completa, com a discriminação de regiões e comparações, está disponível online. CCL: comprometimento cognitivo leve; CCLa: comprometimento cognitivo leve amnéstico; DA: doença de Alzheimer; RMP: rede de modo padrão; ATV: área tegmental ventral; CCP: córtex cingulado posterior.

Estrutura ou rede	Achado resumido	Estágio e técnica	Síntese clínica e limite	Fontes
Hipocampo e RMP	Na tarefa, menor ativação na recuperação e maior na codificação; em repouso, redução da conectividade CCP–parahipocampo e alteração da lateralização hipocampal.	CCL; RMf em tarefa e repouso (BOLD)	Padrão dependente da etapa da memória e da métrica; aumentos não comprovam compensação.	[17], [18], [26]
Redes cerebelares	Aumentos e reduções envolvendo tálamo, regiões frontoparietais, Crus II e lóbulo IX.	CCL/CCLa; RMf em repouso (BOLD)	Reorganização regional heterogênea, sem trajetória ou mecanismo compensatório demonstrado.	[23]
RMP e lobo temporal medial	Predomínio de redução da conectividade no CCP, precuneus, hipocampo e córtex entorrinal; aumentos frontais ou durante a codificação em alguns contrastes.	DA; RMf em repouso e com tarefa (BOLD)	Desorganização de redes de memória; direção variável por conexão, sem inferência causal ou diagnóstica individual.	[17], [18]
Redes corticais e lateralização	Alterações sensoriomotoras, visuais e auditivas, interrupção parahipocampal e redução da assimetria funcional.	DA; RMf em repouso e tarefa (BOLD)	Distribuição ampla e dependente do método; achados morfológicos não equivalem a medidas BOLD.	[6], [26]
Tálamo–precuneus	Redução da conectividade de subdivisões talâmicas, incluindo núcleos intralaminares, com o precuneus.	DA; RMf em repouso (BOLD)	Associação com pior desempenho cognitivo, ainda sem validação clínica externa.	[20]
Núcleos do tronco encefálico	Redução de conexões da ATV com redes parietais/RMP e associações neuropsiquiátricas; alteração da conexão rafe–precuneus descrita; ausência de diferença significativa no locus coeruleus.	CCL e DA; RMf em repouso (BOLD)	Evidência regional; a resolução e os efeitos de volume parcial limitam a análise de núcleos pequenos, e a RMf não mede neurotransmissores diretamente.	[21], [22]
Crus II e lóbulo IX	Redução de conexões cerebelares com regiões frontais, temporais, límbicas, parietais e subcorticais.	DA; RMf em repouso (BOLD)	Comprometimento cerebelocortical específico por semente; não indica redução global do cerebelo.	[23]

equipamentos de RM em comparação com PET não permite concluir, por si só, que a RMf constitua alternativa acessível ou custo-efetiva no SUS.

No contexto brasileiro, a elevada dependência do SUS e a distribuição desigual de recursos diagnósticos justificam investigar métodos potencialmente escaláveis. Entretanto, a eventual incorporação da RMf dependerá não apenas da disponibilidade do equipamento, mas também do tempo de aquisição, da capacidade de processamento, do treinamento das equipes, do controle de qualidade e da demonstração de benefício clínico. Nesse estágio, sua contribuição mais bem sustentada está na compreensão fisiopatológica e na pesquisa de candidatos a biomarcadores complementares.

V. Conclusão

A literatura mapeada indica que alterações funcionais na RMP, no hipocampo, em circuitos subcorticais e em redes do tronco encefálico e do cerebelo são investigadas como potenciais biomarcadores no CCL e na DA. Entretanto, a heterogeneidade das populações, dos protocolos e dos métodos analíticos, aliada à limitada validação longitudinal e externa, ainda restringe sua aplicação clínica individual. Os achados devem, portanto, ser compreendidos como candidatos a biomarcadores em pesquisa. Estudos futuros devem priorizar protocolos padronizados, coortes multicêntricas, confirmação etiológica e avaliação do valor incremental da RMf em relação aos métodos diagnósticos existentes.

Referências

[1] D. F. Matos, W. S. Paz *et al.*, "Caracterização epidemiológica da mortalidade por Alzheimer no Brasil entre 2010 a 2019," *Research, Society and Development*, 2021, acesso em: 28 nov. 2023.

[2] Alzheimer's Association, "Dentro do cérebro: Uma viagem interativa," 2023, acesso em: 23 nov. 2023. [Online]. Available: https://www.alz.org/br/passeio-pelo-c%C3%A9rebro-interativo.asp

[3] A. Sereniki and M. A. B. F. Vital, "A doença de Alzheimer: aspectos fisiopatológicos e farmacológicos," *Revista de Psiquiatria do Rio Grande do Sul*, vol. 30, no. 1, pp. 1–17, 2008, acesso em: 20 nov. 2023. [Online]. Available: https://www.scielo.br/j/rprs/a/LNQzKPVKxLSsjbTnBCps4XM/?lang=pt&format=pdf

[4] Sociedade Brasileira de Geriatria e Gerontologia. (2021) Dia nacional de conscientização da doença de Alzheimer. [Online]. Available: https://sbgg.org.br/dia-nacional-de-conscientizacao-da-doenca-de-alzheimer/

[5] A. P. R. Machado, I. O. Carvalho, and H. M. d. Rocha Sobrinho, "Neuroinflamação na doença de Alzheimer," *Revista Brasileira Militar de Ciências*, vol. 6, no. 14, pp. 30–38, 2020, acesso em: 20 nov. 2023. [Online]. Available: https://rbmc.emnuvens.com.br/rbmc/article/view/33/27

[6] P. Forouzannezhad, A. Abbaspour *et al.*, "A survey on applications and analysis methods of functional magnetic resonance imaging for Alzheimer's disease," *Journal of Neuroscience Methods*, vol. 317, pp. 121–140, 4 2019.

[7] A. B. Q. Steiner, A. F. Jacinto *et al.*, "Mild cognitive impairment and progression to dementia of Alzheimer's disease," *Revista da Associação Médica Brasileira*, vol. 63, pp. 651–655, 7 2017.

[8] M. Martorelli, L. Hartle *et al.*, "Diagnostic accuracy of early cognitive indicators in mild cognitive impairment," *Dementia & Neuropsychologia*, vol. 14, pp. 358–365, 12 2020.

[9] A. A. T. Monfared, M. J. Byrnes *et al.*, "Alzheimer's disease: Epidemiology and clinical progression," *Neurology and Therapy*, vol. 11, pp. 553–569, 6 2022.

[10] L. P. Schilling, M. L. F. Balthazar *et al.*, "Diagnóstico da doença de Alzheimer: recomendações do Departamento Científico de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia," 2022, acesso em: 24/06/2025.

[11] Instituto de Estudos de Saúde Suplementar, "Panorama dos idosos beneficiários de planos de saúde no Brasil – março de 2020," Instituto de Estudos de Saúde Suplementar (IESS), São Paulo, Brasil, Tech. Rep., May 2020, acesso em: 24 de nov. 2023. [Online]. Available: https://www2.iess.org.br/cms/rep/panorama_dos_idosos.pdf

[12] Instituto Brasileiro de Geografia e Estatística. (2023, Oct.) Censo 2022: número de pessoas com 65 anos ou mais de idade cresceu 57,4% em 12 anos. Instituto Brasileiro de Geografia e Estatística. Acesso em: 24 de nov. 2023. [Online]. Available:

https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/38186-censo-2022-numero-de-pessoas-com-65-anos-ou-mais-de-idade-cresceu-57-4-em-12-anos

[13] Fundação Oswaldo Cruz. (2026, May) Fiocruz apresenta estudo sobre envelhecimento e lança painel inédito de indicadores. Fundação Oswaldo Cruz. Disponível em: https://fiocruz.br/noticia/2026/05/fiocruz-apresenta-estudo-sobre-envelhecimento-e-lanca-painel-inedito-de-indicadores.

[14] Estudo Longitudinal da Saúde dos Idosos Brasileiros. (2026) Painel de indicadores ELSI-Brasil. Fundação Oswaldo Cruz. [Online]. Available: https://elsi.cpqrr.fiocruz.br/indicadores-lista/

[15] A. F. Almeida, L. C. Silva *et al.*, "Doença de Alzheimer: aspectos clínicos, diagnósticos e terapêuticos na atualidade," *Revista Eletrônica Acervo Saúde*, vol. 25, p. e19816, 2025. [Online]. Available: https://acervomais.com.br/index.php/saude/article/view/19816

[16] Agência Nacional de Saúde Suplementar, "Anexo II: Diretrizes de utilização para cobertura dos procedimentos na saúde suplementar," Agência Nacional de Saúde Suplementar, Rio de Janeiro, Brasil, Tech. Rep., 2021, resolução Normativa n. 465/2021 e atualizações posteriores. Acesso em: 24 jun. 2026. [Online]. Available: https://www.gov.br/ans

[17] Y. Zhang, C. Zhang *et al.*, "Alzheimer's disease: Potential pathogenesis and imaging findings," *VIEW*, vol. 4, 10 2023.

[18] B. Ibrahim, S. Suppiah *et al.*, "Diagnostic power of resting-state fMRI for detection of network connectivity in Alzheimer's disease and mild cognitive impairment: A systematic review," *Human Brain Mapping*, vol. 42, pp. 2941–2968, 6 2021.

[19] R. Jalilianhasanpour, E. Beheshtian *et al.*, "Functional connectivity in neurodegenerative disorders," *Topics in Magnetic Resonance Imaging*, vol. 28, pp. 317–324, 12 2019.

[20] Y. Wu, X. Wu *et al.*, "Abnormal functional connectivity of thalamic subdivisions in Alzheimer's disease: A functional magnetic resonance imaging study," *Neuroscience*, vol. 496, pp. 73–82, 8 2022.

[21] Á. Patthy, J. Murai *et al.*, "Neuropathology of the brainstem to mechanistically understand and to treat Alzheimer's disease," *Journal of Clinical Medicine*, vol. 10, p. 1555, 4 2021.

[22] L. Serra, M. D'Amelio *et al.*, "In-vivo mapping of brainstem nuclei functional connectivity disruption in Alzheimer's disease." *Neurobiology of Aging*, 2018, acesso em: 12/04/2026.

[23] F. Tang, D. Zhu *et al.*, "Differences in cerebellar functional connectivity between mild cognitive impairment and Alzheimer's disease: A seed-based approach," *Frontiers in Neurology*, 2021.

[24] A. C. Tricco, E. Lillie *et al.*, "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[25] M. D. J. Peters, C. Marnie *et al.*, "Updated methodological guidance for the conduct of scoping reviews," *JBI Evidence Synthesis*, 2020.

[26] H. Liu, L. Zhang *et al.*, "Changes in brain lateralization in patients with mild cognitive impairment and Alzheimer's disease: A resting-state functional magnetic resonance study from Alzheimer's Disease Neuroimaging Initiative," *Frontiers in Neurology*, vol. 9, 2 2018.

[27] B. L. Klaassens, J. M. van Gerven *et al.*, "Cholinergic and serotonergic modulation of resting state functional brain connectivity in Alzheimer's disease," *NeuroImage*, vol. 199, pp. 143–152, 10 2019.

[28] C. Saiz-Masvidal, V. D. la Peña-Arteaga *et al.*, "Mapping cross-modal functional connectivity of major neurotransmitter systems in the human brain," *Brain Structure and Function*, vol. 230, p. 137, 8 2025.

[29] S. Seoane, M. van den Heuvel *et al.*, "The subcortical default mode network and Alzheimer's disease: a systematic review and meta-analysis," *Brain Communications*, vol. 6, 3 2024.

[30] P. Krashia, E. Spoleti, and M. D'Amelio, "The VTA dopaminergic system as diagnostic and therapeutical target for Alzheimer's disease," *Frontiers in Psychiatry*, vol. 13, 10 2022.

[31] S. Saggu, A. Bai *et al.*, "Monoamine alterations in Alzheimer's disease and their implications in comorbid neuropsychiatric symptoms," *GeroScience*, vol. 47, pp. 457–482, 9 2024.

[32] J. D. Schmahmann, X. Guell *et al.*, "The theory and neuroscience of cerebellar cognition," *Annual Review of Neuroscience*, vol. 42, pp. 337–364, 7 2019.

A Comparative Study of Pre-trained Convolutional Neural Networks for Microcalcification Detection in Digital Mammography

Clara da Silva e Oliveira
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0003-1746-3055

Bruno Augusto Machado Goes
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0001-6815-5525

Marcos Vinicius Santos Coelho
Faculty of Electrical Engineering(FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0009-0007-3051-5768

Samuel de Lima Pereira
Institute of Physics (INFIS)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0009-0005-8931-7593

Ana Claudia Patrocinio
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Pedro Cunha Carneiro
Faculty of Electrical Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

***Abstract*—Microcalcifications are relevant mammographic findings whose identification can be challenging because of their small size and low local contrast. This study evaluated and compared three convolutional neural network architectures pre-trained on ImageNet (VGG19, InceptionV3, and ResNet50V2) for binary classification of mammographic regions of interest according to the presence or absence of microcalcifications. A balanced dataset containing 5,378 ROIs derived from the INbreast database was used, comprising 2,689 samples from each class. The dataset was divided into training, validation, and test subsets using an 80/10/10 ratio. The models were trained using data augmentation and a two-stage transfer-learning strategy, in which the model was trained before fine-tuning the last 20 backbone layers. Performance was evaluated on a test set containing 538 ROIs. VGG19 achieved the highest overall performance, with an accuracy of 92.75%, sensitivity of 91.45%, specificity of 94.05%, F1-score of 92.66%, and AUC of 0.9708. ResNet50V2 achieved an accuracy of 88.85% and an AUC of 0.9344, whereas InceptionV3 achieved an accuracy of 86.25% and an AUC of 0.9305. Under the evaluated conditions, VGG19 provided the best balance between positive-class detection and overall discrimination.**

***Index Terms*—Breast cancer, convolutional neural networks, deep learning, mammography, microcalcifications, transfer learning.**

I. Introduction

Breast cancer still ranks high among public health issues, mainly because it remains very common and because late diagnosis directly affects the prognosis. For this reason, early detection is essential for improving clinical outcomes. In this context, mammography plays a central role, as it is the most widely used screening method and allows the identification of breast changes even before obvious clinical manifestations appear [1]. Among the most relevant mammographic findings, microcalcifications stand out as potential indicators of malignancy. Microcalcifications are small calcium deposits in breast tissue that may be associated with benign or malignant lesions, depending on their morphological characteristics and distribution. Traditionally, microcalcification analysis is performed by experienced radiologists, a process that requires time and expertise and is subject to interobserver variability [2]. Furthermore, mammography depends not only on the image itself but also on the experience of the interpreter. Interpretation can vary among observers, and this point becomes even more important when dealing with subtle findings. In such situations, computational tools that help highlight relevant patterns become highly useful, primarily as support for the specialist and not as substitute for medical analysis [3]. In this scenario, artificial intelligence methods have increasingly been investigated within medical image analysis. Among these computational approaches, deep learning frameworks, particularly convolutional neural networks (CNNs), distinguish themselves by automatically extracting hierarchical feature representations directly from raw pixel data, eliminating the dependency on operator-dependent and handcrafted features. This sub-surface feature learning is very useful in digital mammography, where microcalcification detection relies on capturing highly subtle variations in local contrast, granular texture, and complex spatial distributions that are difficult to formalize into static, rule-based algorithms [3]. Another important factor is that, in medical applications, massive image datasets are not commonly available. For this reason, the use of pre-trained models with transfer learning has become a very

useful strategy. Instead of starting training from scratch, one benefits from an architecture that has already been trained on a large dataset and adapts that knowledge to a specific task. In practice, this helps make the process more feasible and often improves the model's performance on smaller, more specialized datasets [4]. In recent years, several studies have demonstrated the application of convolutional neural networks in mammography for segmentation, detection, and classification. Even so, there are still differences among published studies, particularly regarding the type of image used, preprocessing methods, and chosen models. Therefore, it still makes sense to compare well-established architectures when seeking to understand which one performs best on a specific problem, such as distinguishing regions with microcalcifications from regions without microcalcifications [5]. To facilitate microcalcification detection, this study aims to evaluate and compare the performance of three pre-trained convolutional neural network (CNN) architectures for binary classification of digital mammographic regions of interest (ROIs) containing or lacking microcalcifications. Using transfer learning and data augmentation strategies on data derived from the public INbreast database, this research explores how residual connections, multi-scale feature extraction, and architectural depth impact the model's ability to capture subtle contrast and textural patterns in a small medical dataset. This study compares VGG19, ResNet50V2, and InceptionV3 for the binary classification of mammographic ROIs with and without microcalcifications. To improve the fairness of the comparison, all architectures were evaluated using the same dataset split, batch size, optimizer, learning rates, classification head, data-augmentation strategy, and fine-tuning procedure. Performance was assessed using accuracy, sensitivity, specificity, precision, F1-score, and AUC.

II. Methods

A. Dataset

The images used in this study consist of regions of interest (ROIs) extracted from digital mammograms provided by the public database InBreast [6], divided into two classes as seen in Table I:

TABLE I
Distribution and Categorization of the Extracted Mammographic Regions of Interest

Class	Description	Quantity
With Microcalcification	ROIs containing microcalcifications (suspicious or benign)	2689 samples
Without Microcalcification	ROIs without microcalcifications (normal breast tissue)	2689 samples

In total, 5,378 ROIs with size of 100×100 pixels were provided in a 16-bit format (pixel values ranging from 0 to 4,095), preserving the original radiographic resolution. To ensure compatibility with the pre-trained models (which expect RGB images), the following preprocessing methods were applied:

1) Conversion to 3 channels (simulating RGB images)
2) Normalization of pixels to the range [0, 1] (division by 4,095)
3) Resizing pixel spatial resolution to the CNN model input

The 100×100 pixel size is the most appropriately area to represent a breast region containing microcalcifications (usually smaller than 1 millimeter), allowing visualization of local anatomical context without including unnecessary areas. However, the regions of interest were resized to 224×224 pixel size for input compatibility with the ResNet50V2 and VGG19 models. On the InceptionV3 model, the spatial resolution was resized to 299×299, as necessary for the input size.

B. Data Division

The data were divided into training, validation, and test sets using an 80/10/10 ratio, as shown in Table II, ensuring that all models were evaluated on the same test set.

TABLE II
Data Partitioning Strategy for Model Training, Validation, and Test

Set	Proportion	Quantity	Purpose
Training	80%	4302 ROIs	Model weight tuning
Validation	10%	538 ROIs	Hyperparameter tuning and early stopping
Test	10%	538 ROIs	Final performance evaluation

The dataset was stratified to maintain class proportion across the splits, and a fixed random seed was used to ensure the reproducibility of the experiments.

C. Architectures of the Evaluated Models

Three pre-trained architectures widely used in the medical computer vision literature – VGG19, InceptionV3 and ResNet50V2 – were evaluated. These models were pre-trained on ImageNet [7] [8], an image dataset with over a million training images and 1,000 object classes. The original ImageNet classification head was removed from each architecture and replaced with the same custom binary classification head. Table III summarizes the main structural characteristics and the input configuration adopted for each backbone.

1) VGG19 (Visual Geometry Group - 19 layers):

VGG19 is a sequential convolutional neural network developed by the Visual Geometry Group at the University of Oxford. As summarized in Table III, it comprises 16 convolutional layers organized into five blocks. The architecture applies successive 3×3 convolutional filters to progressively extract image features, with max-pooling operations used between blocks to reduce spatial resolution. The simplicity and uniformity of its architecture have made VGG19 a widely used benchmark model [9]. In the original ImageNet architecture, the convolutional backbone is followed by three fully connected layers. However, these original classification layers

TABLE III
STRUCTURAL CHARACTERISTICS AND FEATURES OF THE EVALUATED CNN BACKBONES

Characteristic	**VGG19**	**InceptionV3**	**ResNet50V2**
Main architectural design	Sequential convolutional blocks	Parallel multi-scale convolutional modules	Residual bottleneck blocks with skip connections
Structural organization	16 convolutional layers organized into five blocks	11 Inception modules with parallel processing paths	16 residual bottleneck blocks
Main convolutional operations	Successive 3×3 convolutions	Parallel and factorized convolutions	1×1 and 3×3 bottleneck convolutions
Distinctive mechanism	Uniform sequential feature extraction	Multi-scale feature extraction	Residual learning with identity mappings
Approximate backbone parameters	20.0 million	21.8 million	23.6 million
Backbone output channels	512	2048	2048
Input size used in this study	$224 \times 224 \times 3$	$299 \times 299 \times 3$	$224 \times 224 \times 3$
Custom classification head	Global average pooling, dropout (0.3), and sigmoid output	Global average pooling, dropout (0.3), and sigmoid output	Global average pooling, dropout (0.3), and sigmoid output

were not used in the present study. Instead, the convolutional backbone was connected to the common classification head described in Table III, consisting of global average pooling, dropout with a rate of 0.3, and a sigmoid output unit. The original 100×100-pixel mammographic ROIs were resized to 224×224 pixels and replicated across three channels before being provided to VGG19. The model was initialized with ImageNet weights and subsequently adapted to the microcalcification classification task using fine-tuning described in Section II-D [9].

2) InceptionV3:

The InceptionV3 architecture focuses on increasing computational efficiency and network width, performing convolutions of different sizes (1x1 and 3x3) and pooling operations in parallel. This allows the network to capture features at multiple scales within the same module. InceptionV3 differs from the original Inception concept by transforming a 5×5 convolution into two stacked 3×3 convolutions to reduce the number of parameters and computational cost, together with batch normalization to stabilize and accelerate training. The architecture is relatively deep but computationally more efficient than VGG19, requiring a minimum input size of 75×75 pixels and being optimized for an input of 299×299 pixels. This architectural optimization makes InceptionV3 a strong candidate for clinical environments where computational bandwidth and memory efficiency are critical operational constraints [10]. InceptionV3 architecture is summarized in Table III. The original mammographic ROIs were resized to 299×299 pixels and converted to three channels before being provided to the backbone. The model was initialized with ImageNet weights and fine-tuned for the classification of ROIs according to the presence of microcalcifications [10].

3) ResNet50V2 (Residual Network - 50 layers):

ResNet50V2 is an evolution of the original residual network (ResNet), proposed to facilitate the training of deep neural networks through residual connections, also known as skip connections. This approach allows gradients to propagate directly through the network during training, mitigating the vanishing gradient problem common in deep architectures. ResNet50V2 introduces a pre-activation design, in which batch normalization and ReLU activation are applied before each convolution rather than after. This configuration, with its 50 weighted layers, employs "bottleneck" blocks that are computationally more efficient, making the network faster than other architectures [11]. Table III presents the architecture of ResNet50V2. Each bottleneck block combines 1×1 convolutions, used to reduce or restore channel dimensionality, with 3×3 convolutions for spatial feature extraction. The mammographic ROIs were resized to 224×224 pixels, converted to three channels, and provided to the backbone initialized with ImageNet weights [11].

D. Transfer Learning

To initialize the network with a feature representations, a transfer learning approach was adopted [12]. The backbone architectures were pre-trained on ImageNet [7], a large-scale natural image dataset. This initialization provides a advantage by using general visual patterns learned from millions of images [13], thereby accelerating convergence and mitigating the risk of overfitting [14], especially given the relatively limited size of medical imaging datasets. To fine-tune the models, training was carried out in two stages. In the first stage, the backbone was kept fully frozen and the model was trained for 5 epochs, using a learning rate of 1×10^{-3}. In the second stage, the last 20 layers of the backbone were unfrozen and retrained with a reduced learning rate of 1×10^{-5}, allowing these layers to adapt to the specific characteristics of the dataset while minimizing the risk of disrupting the pre-trained weights.

E. Data Augmentation for Mammograms

To artificially expand the training set and enhance the model's generalization capability, a set of data augmentation strategies was employed [15]. These transformations were

selected to preserve the clinical integrity of the mammographic images while simulating realistic variations encountered in routine screening [16]. The following geometric augmentations were applied, inspired by techniques commonly used in deep learning [17], [18]:

- **Rotation (±20 degrees):** Accounts for minor variations in breast positioning and compression during image acquisition [19].
- **Horizontal and Vertical Shifts (±20%):** Simulates small displacements of the region of interest (ROI) relative to the image frame.
- **Zoom (±20%):** Introduces variability in the magnification scale, compensating for differences in imaging equipment and patient anatomy.
- **Horizontal Flip (Enabled):** Exploits the bilateral symmetry between left and right breasts, allowing the model to learn invariant features from both sides.

Transformations that could distort the morphological characteristics of microcalcifications, such as shearing, elastic deformations, or non-rigid warping, were intentionally omitted. Preserving the shape, size, and spatial distribution of microcalcifications is important in medical analysis, as these features are often indicators of malignancy in mammographic findings [20]. Therefore, all augmentations were constrained to rigid and affine transformations that maintain pixel intensity relationships and structural fidelity.

F. Hyperparameters

The same training configuration was applied to VGG19, ResNet50V2, and InceptionV3. The hyperparameters were empirically selected using the training and validation subsets, without using information from the test set. The Adam optimizer was adopted because of its adaptive learning-rate mechanism and suitability for stochastic optimization [21]. The main training and fine-tuning parameters are summarized in Table IV.

TABLE IV
MAIN TRAINING AND FINE-TUNING PARAMETERS

Parameter	Value	Purpose
Batch size	16	Common batch size for all models
Optimizer	Adam; weight decay: 1×10^{-4}	Optimization and regularization
Initial training	5 epochs; LR: 1×10^{-3}	Training with the backbone frozen
Fine-tuning	Last 20 layers; LR: 1×10^{-5}	Adaptation of high-level features
Early stopping	Validation AUC; patience: 5	Prevention of unnecessary training
LR scheduler	Factor: 0.5; patience: 3	Learning-rate reduction after a plateau

Binary cross-entropy was used as the loss function. The custom classification head consisted of a global average pooling layer, followed by a dropout layer with a rate of 0.3, and a single sigmoid output unit. Dropout and weight decay were adopted to reduce the risk of overfitting during training [22]. A threshold of 0.5 was applied to convert the sigmoid probabilities into binary predictions. EarlyStopping and ReduceLROnPlateau were applied only during the second training stage, with both callbacks monitoring validation AUC. EarlyStopping terminated training after five consecutive epochs without improvement, limiting unnecessary training and reducing the risk of overfitting [22]. ReduceLROnPlateau reduced the learning rate by a factor of 0.5 after three consecutive epochs without improvement, allowing smaller updates to the model weights after the validation metric reached a plateau.

G. Evaluation Metrics

The evaluation metrics employed to assess the performance of the models are summarized in Table V. Accuracy measures the overall proportion of correct diagnoses, while sensitivity and specificity evaluate the model's ability to detect true microcalcifications and correctly identify normal tissue, respectively. Precision assesses the reliability of positive predictions, and the F1 score was adopted as a balanced measure between precision and recall, particularly relevant for handling potential class imbalance [23]. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was employed to evaluate the overall discriminating capacity of the models across all classification thresholds [24].

TABLE V
DEFINITIONS OF THE BINARY CLASSIFICATION PERFORMANCE METRICS

Metric	Formula	Interpretation
Accuracy	(TP+TN)/Total	Overall proportion of correct predictions
Sensitivity (Recall)	TP/(TP+FN)	Ability to detect true microcalcifications
Specificity	TN/(TN+FP)	Ability to identify normal tissue
Precision	TP/(TP+FP)	Probability that a finding is truly a microcalcification
F1 Score	$2 \times (P \times R)/(P + R)$	Balance between precision and recall
AUC-ROC	Area under ROC curve	Overall discriminative capacity of the test

Where: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, P = Precision and R = Recall. High sensitivity indicates a low rate of false negatives, which is critical to avoid undetected cancer cases, while high specificity indicates a low rate of false positives, which is essential to avoid unnecessary biopsies and patient recall. The F1-Score provides a balanced measure when both precision and recall are equally important for the clinical application.

III. RESULTS

The classification performance of each evaluated CNN architecture was assessed on the independent test set. The metrics acquired from these experiments, encompassing overall classification accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC, are consolidated and cross-compared in Table VI.

TABLE VI
CLASSIFICATION PERFORMANCE COMPARISON ACROSS THE THREE CNN ARCHITECTURES

Metrics	ResNet50V2	InceptionV3	VGG19
Accuracy (%)	88.85	86.25	92.75
Sensitivity (Recall) (%)	82.53	81.04	91.45
Specificity (%)	95.17	91.45	94.05
Precision (%)	94.47	90.46	93.89
F1-Score (%)	88.09	85.49	92.66
AUC (%)	93.44	93.05	97.08

A. Analysis by Model

1) VGG19: VGG19 produced the best overall results among the three architectures. It achieved the highest sensitivity (91.45%), detecting approximately 91 out of 100 microcalcifications, along with the highest AUC (97.08%) and F1-score (92.66%), reflecting a balanced relationship between precision and recall. During fine-tuning, the learning rate was reduced to 5×10^{-6} at epoch 28 after validation AUC plateaued, and early stopping was triggered at epoch 38. The binary confusion matrix for the VGG19 model is shown in Fig. 1. The model correctly identified 246 true positives and 253 true negatives, with only 16 false positives and 23 false negatives.

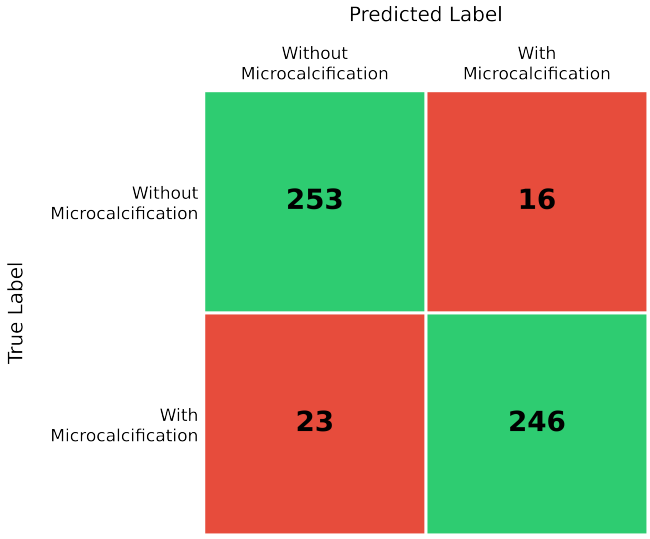


Fig. 1. Binary Confusion Matrix for VGG19

2) ResNet50V2: ResNet50V2 obtained the highest specificity among the three architectures (95.17%), indicating a low rate of false positives, and the highest precision (94.47%), although both metrics are close to VGG19 results. The confusion matrix in Fig. 2 shows 222 true positives, 256 true negatives, 13 false positives, and 47 false negatives, corresponding to a sensitivity of 82.53%. The learning rate was reduced to 5×10^{-6} at epoch 30, and early stopping was triggered at epoch 34.

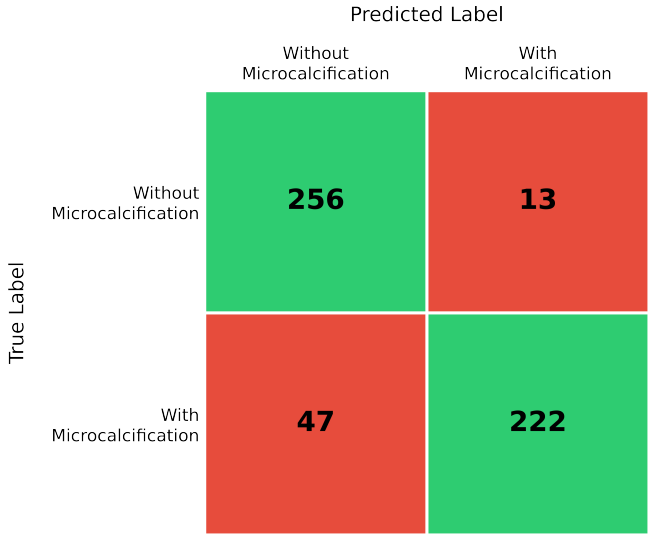


Fig. 2. Binary Confusion Matrix for ResNet50V2

3) Inception V3: InceptionV3 achieved the lowest results among the three architectures, although some metrics are close to ResNet50V2 performance. During fine-tuning, the learning rate was reduced to 5×10^{-6} at epoch 45, and early stopping was triggered at epoch 48. It obtained a sensitivity of 81.04%, specificity of 91.45%, and precision of 90.46%. Fig. 3 displays the confusion matrix for InceptionV3, indicating that the model correctly identified 218 true positives and 246 true negatives, but produced 23 false positives and 51 false negatives, the highest false-negative count among the three models.

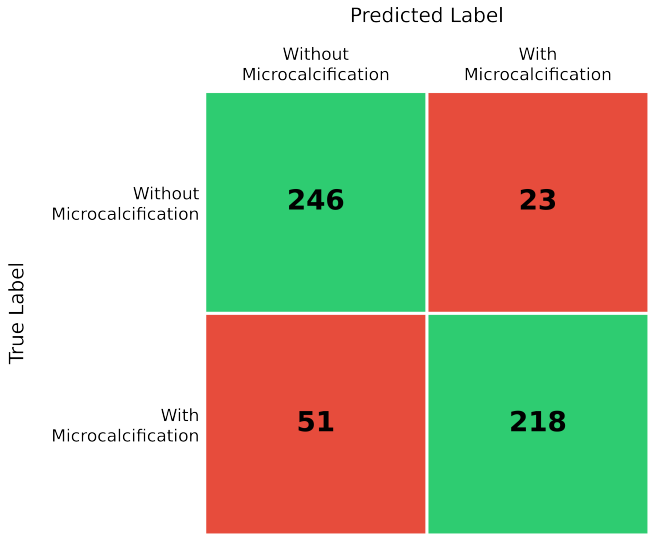


Fig. 3. Binary Confusion Matrix for InceptionV3

Fig. 4 provides a comparative performance analysis of the three CNN architectures evaluated for microcalcification detection. The multi-bar chart compares Accuracy, Recall, and

F1-Score, showing the superior performance of the VGG19 model.

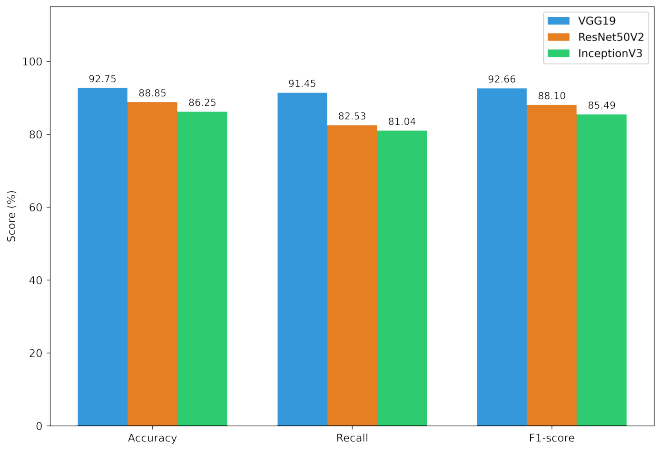


Fig. 4. Visual comparison of key metrics among models for the detection of microcalcifications

Fig. 5 presents the ROC curves for the three architectures. VGG19 (AUC = 0.97) maintains the highest true positive rate across nearly all false positive rate thresholds. ResNet50V2 and InceptionV3 achieve nearly identical overall performance (AUC = 0.93 for both), with their curves closely overlapping throughout most of the plot, but ResNet50V2 shows a minimal advantage over InceptionV3 at lower false positive rates. As the false positive rate increases, however, the two curves progressively converge, indicating that the two architectures achieve comparable discriminative performance at higher decision thresholds.

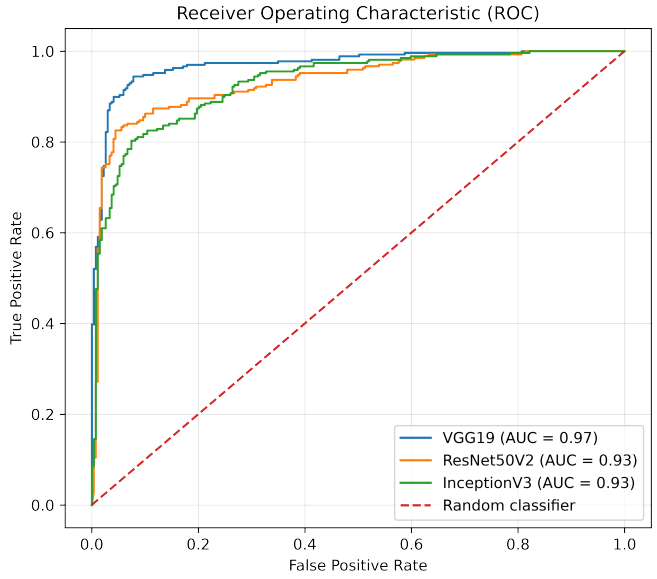


Fig. 5. Visual comparison of ROC curves across the CNN models

IV. Discussion

The experimental results support the premise of this study, that the architecture of a pre-trained convolutional neural network influences its capacity to capture the subtle contrast, morphology, and textural patterns that characterize mammographic microcalcifications. Among the three models evaluated, VGG19 achieved the best overall performance, with an accuracy of 92.75%, an F1-score of 92.66%, and an AUC of 0.971. Although VGG19 is the oldest and structurally simplest of the three architectures, its performance may be explained by the interaction between its architecture and the fine-tuning strategy employed. VGG19's uniform stack of small 3×3 convolutional filters allows the network to progressively build fine-grained, hierarchical representations of localized textural patterns such as microcalcifications. Unfreezing the last 20 layers during the second training stage, with a reduced learning rate of 1×10^{-5}, likely allowed these late-stage filters to adapt closely to the specific texture and contrast signatures of microcalcifications. The confusion matrix for VGG19 (Fig. 1) reveals that the model misclassified 39 ROIs, comprising 16 false positives and 23 false negatives. When the model indicates a microcalcification, there is a 93.89% precision probability that the finding is correct, supporting its potential as a candidate for microcalcification detection workflows. ResNet50V2 also achieved well-balanced results, obtaining the highest specificity (95.17%) and precision (94.47%) among the three architectures, indicating a low rate of false positives. This behavior is consistent with the stabilizing effect of its identity skip connections, which allow high-frequency features to propagate directly across deeper layers and mitigate the vanishing gradient phenomenon common in deep sequential networks. The model achieved an overall accuracy of 88.85% and an AUC of 0.934, with a sensitivity of 82.53%, correctly detecting 222 of the 269 microcalcifications present in the test set while missing 47. Although its sensitivity trails VGG19's, its low false-positive rate makes it a viable choice in clinical contexts where minimizing false positives is prioritized over maximizing detection rate. InceptionV3 achieved the lowest overall performance among the three architectures, although its results remain promising, with an AUC of 0.931. The network obtained a sensitivity of 81.04%, capturing 218 of the 269 true microcalcifications in the test set, alongside a specificity of 91.45% and a precision of 90.46%. This relatively lower performance may stem from the parallel 1×1, 3×3, and 5×5 filters within its Inception modules, which concatenate features after each block. This design may make the network more prone to mistaking dense parenchymal tissue patterns for microcalcifications, contributing to its comparatively higher false-negative count (51) relative to VGG19 and ResNet50V2. It is also possible that InceptionV3's greater architectural complexity required more extensive fine-tuning of its multi-branch modules than the fixed 20-layer unfreezing strategy allowed, limiting its ability to fully adapt its feature extractors to this specific classification task. Even so, InceptionV3's overall discriminative capacity remains meaningful, suggesting it retains value as a possible classification tool, albeit with a higher expected rate of missed detections than the other two architectures evaluated. Taken together, these results indicate a clear trade-off among the three architectures rather than a

single unambiguous winner. VGG19 offers the greatest overall discriminative capacity and the most balanced error profile, making it the most promising candidate where maximizing detection accuracy is the priority. ResNet50V2, by contrast, offers conservative and precise predictions, which may be preferable in workflows where reducing false positives is prioritized over sensitivity. InceptionV3, while the weakest of the three in this study, still performed well and may benefit from a fine-tuning strategy tailored more specifically to its multi-branch structure in future work. Furthermore, the relatively small test set size (538 ROIs) limits the extent to which these findings can be generalized without further validation on larger dataset.

V. Conclusion

This study compared convolutional neural networks with transfer learning for binary classification of microcalcifications in digital mammography. By evaluating the performance of three distinct architectures (VGG19, InceptionV3, and ResNet50V2) on a dataset extracted from the INbreast database, the results indicate that VGG19 outperforms the other models on this specific microcalcification classification task, achieving an accuracy of 92.75% and an AUC of 0.971. The findings offer insights into the potential deployment of deep learning models as microcalcification detection support tools. ResNet50V2's high specificity and precision suggest it may be particularly well suited to minimizing false positives, while VGG19's better overall sensitivity and discriminative capacity make it the most promising candidate among the three for maximizing detection of true microcalcifications. InceptionV3, while achieving the weakest results of the three, still demonstrated meaningful results, indicating that all three architectures retain some value as classification algorithms. Future research will focus on evaluating these architectures against larger and imbalanced databases, exploring advanced attention mechanisms to further improve classification performance, and developing an end-to-end microcalcification detection algorithm for full-field mammography.

References

[1] J. Subelack, R. Morant, M. Blum, A. Gräwingholt, J. Vogel, A. Geissler, and D. Ehlig, "Retrospective evaluation of interval breast cancer screening mammograms by radiologists and ai," *European Radiology*, vol. 36, no. 2, pp. 824–836, 2026.

[2] J. Rani, J. Singh, and J. Virmani, "Deep learning based segmentation methods applied to ddsm images: A review," *Archives of Computational Methods in Engineering*, vol. 32, 2025.

[3] A. M. M. da Silva, A. C. Patrocínio, and H. Schiabel, "Processamento e análise de imagens médicas," *Revista Brasileira de Física Médica*, vol. 13, no. 1, pp. 34–48, 2019.

[4] M. M. Emam, E. H. Houssein, N. Abdelsamee, M. Alohali *et al.*, "Breast cancer diagnosis using optimized deep convolutional neural network based on transfer learning technique and improved coati optimization algorithm," *Expert Systems with Applications*, vol. 255, p. 124581, 2024.

[5] F. A. Zeiser *et al.*, "Segmentation of masses on mammograms using data augmentation and deep learning," *Journal of Digital Imaging*, vol. 33, pp. 858–868, 2020.

[6] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Miami, FL, USA: IEEE, Jun. 2009, pp. 248–255.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015, arXiv:1409.1556.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2818–2826.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*. Springer, 2016, arXiv:1603.05027.

[12] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, Montreal, Canada, 2014, pp. 3320–3328.

[14] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, J. Wei, and K. Cha, "Mass detection in mammograms using transfer learning and deep convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 10, pp. 2296–2307, 2016.

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, Jul. 2019.

[16] Q. Zhou, M. S. Hosseini, M. J. H. H. van der Maaten, D. Behnami, R. Wang, and H. R. Tizhoosh, "Augmentation of mammography images: A study on deep learning performance," *IEEE Access*, vol. 8, pp. 78 615–78 626, 2020.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, originally presented at NeurIPS 2012.

[18] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv Preprint*, Dec. 2017, arXiv:1712.04621.

[19] M. Elter, J. H. Oehl, T. Wittenberg, and C. Roth, "Computer-aided detection of microcalcifications in mammograms," *IEEE Transactions on Medical Imaging*, vol. 28, no. 11, pp. 1793–1804, 2009.

[20] J. Wang, K. Chen, X. Chen, and X. Wang, "Morphological features of microcalcifications for breast cancer diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1363–1372, 2016.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015, arXiv:1412.6980.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[23] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[24] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

Prediction of Hospital Readmission Within 30 days After Discharge with Wearable Temporal Signals

Fábio Arthur Soares Araujo
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
fabioaraujo@ufam.edu.br

Pedro Lucas Bezerra Mendes
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
pedro.mendes@icomp.ufam.edu.br

Marly Guimarães Fernandes Costa
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
mcosta@ufam.edu.br

Francisco de Assis Pereira Januário
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
franciscojanuario@ufam.edu.br

Robson Luis Oliveira de Amorim
Getúlio Vargas University Hospital
Federal University of Amazonas
Manaus, Brazil
amorim.robson@gmail.com

Maria Elizete de Almeida
Getúlio Vargas University Hospital
Federal University of Amazonas
Manaus, Brazil
elizetemanaus@ufam.edu.br

Cícero F. F. Costa Filho
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
ccosta@ufam.edu.br

***Abstract*— Hospital readmission within 30 days of discharge is a significant challenge to healthcare systems, because it is associated with worse patient outcomes and higher costs. In this work, we investigated the prediction of early hospital readmission using clinical variables and temporal features extracted from wearable accelerometry signals. The experiments were carried out on the Bosch Dataset after preprocessing steps to retain patients with sufficient number of valid measurements. Different configurations were evaluated, including progressive feature combinations and scenarios with and without feature selection. The models were trained and evaluated using five outer folds and two inner folds approach, which ensured that hyperparameter tuning and performance estimation were independent. Nine machine learning classifiers were assessed with hyperparameter tuning performed using Grid Search and Optuna. The best result was obtained with SVM with Feature Set 4, Optuna and Pearson-based feature selection, with a mean F1-score of 90.00% and a average accuracy of 93.33% over the five outer folds. These findings suggest potential value in combining temporal accelerometry descriptors and patient-related information, although the small cohort and substantial between-fold variability should be interpreted as exploratory internal validation.**

Keywords*— *hospital readmission prediction, wearable devices, machine learning, temporal feature extraction, feature selection, hyperparameter optimization, nested cross-validation.

I. INTRODUCTION

Hospital readmission is commonly used as an indicator of healthcare quality and is also associated with avoidable costs, especially during the transition from hospital to home. Large studies in the United States have shown that many readmissions occur within 30 days after discharge and are often associated with clinical complications, including in surgical patients [1]. Due to the clinical and economic impact of readmission, programs such as the Hospital Readmissions Reduction Program have increased interest in methods that can identify high-risk patients before adverse outcomes occur [2].

However, predicting readmission is still difficult as it is related to the severity of diseases, comorbidities, care transitions, and patient-specific characteristics. Previous research has demonstrated that models based on routine clinical and social variables often have only moderate predictive performance [3]. Structured scores such as the LACE index and the HOSPITAL score are popular as they are easy to use but they rely mostly on information available at discharge and frequently show poor discrimination across clinical settings [4], [5].

With the increasing availability of digital health data, machine learning models based on electronic health records (EHR) and claims databases have been investigated to improve readmission prediction. Some studies reported that large-scale feature extraction from EMR data can improve performance compared with more traditional approaches, although important challenges related to model generalization and real-world implementation still remain [6]. Similar results have been reported for chronic diseases such as COPD, where combining knowledge-driven and data-driven features improved the area under the receive operating characteristic curve (ROC), while more complex deep learning models did not consistently outperform strong traditional baselines [7].

A limitation of many existing models is that they estimate readmission risk using only information collected at or before discharge. In practice, however, patient recovery continues after leaving the hospital and is influenced by physiological and behavioral changes over time. In this context, remote patient monitoring and wearable devices provide an important source of information because they allow the collection of daily-life signals that are usually not present in hospital records. Recent randomized evidence suggests that post-discharge features related to activity and sleep can improve the prediction of 30-day readmission or death [8].

Recent studies have also explored temporal and behavioral patterns extracted from wearable data. Continuous multi-sensor

monitoring combined with machine learning has identified early signs of impending hospitalization in heart failure patients several days in advance [9]. Consumer wearable data have also shown that disruptions in circadian and biobehavioral rhythms during perioperative stages may be associated with readmission risk [10]. Moreover, wearable-derived descriptors related to variability and signal complexity, including entropy-based measures from heart rate and step-count series, have improved readmission prediction when combined with clinical variables [11], [12]. Probabilistic deep models have further been proposed to estimate how readmission risk evolves day by day after discharge, rather than representing it as a fixed score [13].

Wearable data are increasingly used for readmission prediction. However, the impact of alternative temporal representations of repeated in-hospital accelerometry observations is little studied. This study thus investigates if the combination of independent hourly descriptors, gradually cumulative windows and sequential measurement pairs might improve 30-day readmission prediction. The main contributions are: (i) a multiscale temporal representation for accelerometry signals; (ii) a progressive analysis of feature sets designed to assess the impact of cumulative and contextual information; and (iii) a leakage-aware patient-level nested cross-validation protocol where preprocessing and hyperparameter tuning are confined to the training patients.

II. Materials and Methods

A. Materials

The study used the Bosch Dataset [14], whose data were collected at Robert Bosch Hospital, located in Stuttgart, Germany, throughout 2016. The dataset is publicly available on the PhysioNet platform, version 1.0, published in 2020.

According to the dataset documentation, the cohort is composed of data from 58 hospitalized adult patients admitted for different clinical conditions, including acute infections, chronic obstructive pulmonary disease exacerbations, heart failure, and oncological complications; however, data from only 47 patients were made available, where 16 patients (34,0%), experienced hospital readmission within 30 days, while 31 patients (66.0%), were not readmitted during this period.

B. Methods

During preprocessing, an eligibility procedure was applied to ensure that the final cohort contained patients with sufficient temporal information for feature extraction and longitudinal analysis. Only data sequences containing at least 8 hours of valid data were considered for this study. In addition, only patients with at least two valid data sequences were included, since the proposed strategy was designed to analyze temporal changes between different observation moments rather than isolated records.

These criteria were adopted to ensure that each patient has enough longitudinal information to support the temporal strategy. Patients with only one measurement could not provide consecutive observations, while measurements shorter than 8 hours would not satisfy the window defined for feature generation. After applying these criteria, the final dataset was reduced to 20 patients with a total of 76 valid data sequences.

Within this subset, each patient has an average of 3.8 valid data sequences, with a standard deviation of 2.19, indicating moderate heterogeneity in the number of records per individual. Regarding the clinical outcome, 8 patients experienced hospital readmission, while 12 patients were not readmitted, resulting in a relatively balanced class distribution.

Additional signal-cleaning procedures were not applied, since, according to the dataset documentation, the main preprocessing and signal validation had already been performed during database construction. The preprocessing adopted focuses only on sample selection, temporal standardization, and signal preparation for feature extraction.

Every data sequence was standardized to an 8-hour observation window. From this time window, two complementary temporal representations were defined. The first one was based on individual intervals, in which the 8-hour observation period was divided into eight independent hourly segments, as illustrated in Fig. 1. This method focuses on local information and short-term dynamics.

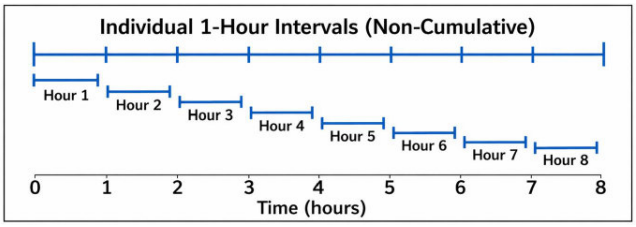


Fig. 1. Individual Hourly Method

The second representation focuses on cumulative strategy, as illustrated in Fig. 2. In this case, features were progressively calculated over the temporal windows starting from the last hour up to the 8 hours period. This method focuses on temporal information that captures changes over the time.

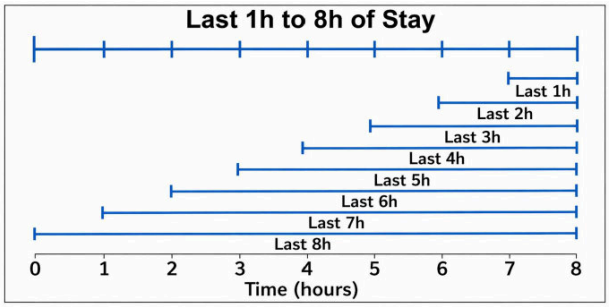


Fig. 2. Cumulative feature Methods

Taken together, these two representations were treated as complementary. While the hourly representation captures immediate and localized fluctuations, the cumulative representation provides a longitudinal view of motor behavior throughout the observation period. This combination was adopted to preserve both short-term variability and progressive temporal evolution in the extracted signal descriptors.

As shown in Table I, a diverse set of time-series derived attributes were extracted from raw accelerometry data, following the general rationale of large-scale time-series feature extraction methods [15]. The features were calculated using 1-hour time windows. Considering 8 hours of monitoring per data sequence, 8 values were generated for each feature, preserving the temporal dynamics of the signals [16].

TABLE I. FEATURES EXTRACTED FROM RAW ACCELEROMETER DATA

PARAMETER	MEASURE
MOVEMENT INTENSITY	MAGNITUDE OF THE ACCELEROMETER
AVERAGED ACCELERATION	STATISTIC CLASSIC TIME DOMAIN
SKEWNESS	KURTOSIS
ENERGY	ROOT MEAN SQUARE
AVERAGED ROTATION ENERGY	SIGNAL MAGNITUDE AREA
AVERAGED VELOCITY	FIRST ORDER DERIVATIVE
SECOND ORDER DERIVATIVE	ZERO CROSSING RATE
CORRELATION BETWEEN AXIS	MEAN CROSSING RATE
SPECTRAL ENTROPY	EIGENVALUES OF THE COVARIANCE MATRIX
AVERAGED ROTATION ANGLES	DOMINANT FREQUENCY

To assess the impact of richer input representations on model performance, four nested feature sets were developed. For each feature set, variables from the previous set were kept and a further source of information was added. This incremental architecture permitted the separate investigation of the contribution of basic statistics, signal dynamics, cumulative temporal information, and patient-related data. Table II summarizes the composition of the four feature sets.

Feature Set 1 was the most basic representation of the movement behavior. Feature Set 2 further developed this representation with descriptors relating to signal dynamics, oscillatory patterns, the intensity of movement and the inter-axis correlations of the accelerometer. Feature Set 3 incorporated cumulative temporal information to account for the gradual changes during the eight-hour period of observation. Finally, Feature Set 4 was the most complete setup by integrating wearable-derived features and patient-related information.

To incorporate temporal information without using sequential deep architectures, the model´s input was constructed by pairs of consecutive measurements, as illustrated in Fig. 3. For each patient with N valid data sequences, N−1 pairs were generated by combining each data sequence with its immediate successor. In each pair, one column represented the previous data sequence and the other, the current data sequence. By doing this, the transitions between consecutive states were preserved.

For labeling, the positive labels were assigned only to the final data sequence pair of readmitted patients. All other pairs, including prior pairs from patients that were readmitted, were labeled as negative. The pair-wise training labels were not identical to the patient-level clinical outcome: a negative label for an earlier pair did not mean that the patient stayed free of readmission for the entire follow-up period. Predictions were made for pairs of measurements, but inner and outer cross-validation partitions were generated based on patient identity so that all pairs from the same patient remained in the same partition.

In the outer-test evaluation, the selected pairs were based on the observed outcome on the patient level. All available measurement pairs with negative reference labels for patients discharged without readmission within 30 days were evaluated. For patients readmitted, only the most recent pair was labeled as positive; earlier pairs of these patients were excluded from test scoring. Confusion-matrix counts and test metrics were computed on these selected pairs. Hence, each readmitted patient results in one positive evaluation pair, while each non-readmitted patient can result in several negative evaluation pairs. This retrospective selection based on the outcome should be differentiated from the assignment of a single prediction for each patient.

TABLE II. COMPOSITION OF THE FEATURE SETS

Feature set	Composition	Number of features
Feature Set 1	Basic statistical descriptors extracted from the accelerometer magnitude, including mean, median, maximum, minimum, and standard deviation.	80
Feature Set 2	Feature Set 1 combined with additional time-domain, frequency-domain, statistical, and inter-axis descriptors, including signal energy, RMS, derivatives, crossing rates, spectral entropy, dominant frequency, movement intensity, and correlations between axes.	496
Feature Set 3	Feature Set 2 combined with cumulative descriptors progressively calculated over observation windows ranging from 1 to 8 hours.	992

Feature Set 4	Feature Set 3 combined with demographic, clinical, and contextual information, including age, sex, anthropometric characteristics, lifestyle variables, medication use, length of hospital stay, housing status, and comorbidity index.	1006

In addition to the experiments reported on all available variables, the performance of reduced feature sets was examined using feature selection based on Pearson correlation. For each feature set, candidate variables were ranked according to the absolute value of Pearson correlation with the pair-level training target and the top 50 ranked variables were retained.

Feature selection was performed inside the nested cross-validation procedure. The correlations were calculated during hyperparameter optimization only with the inner-training partition. The selection resulting from this was then applied unchanged to the corresponding inner-validation partition. The hyperparameters were chosen and the Pearson ranking was computed again on all outer-training patients and the resulting subset was directly applied to the outer-test patients. Thus, the inner-validation and outer-test results were not used for feature selection to assess those partitions.

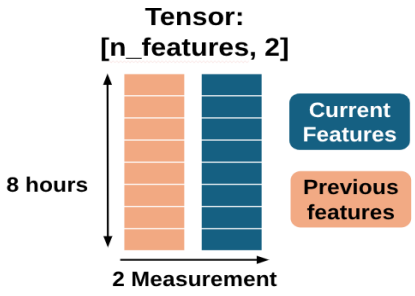


Fig. 3. Model input

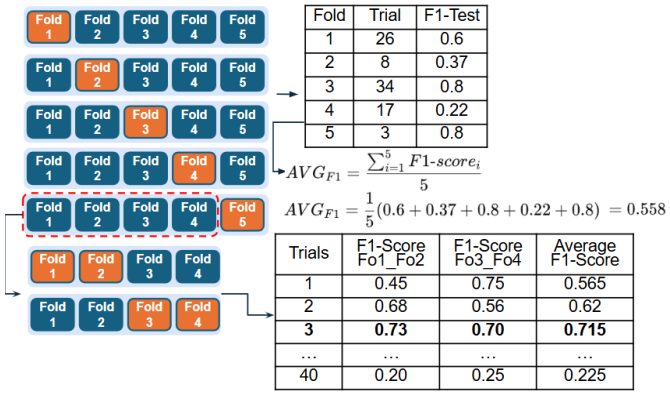
This procedure reduced the input dimensionality and allowed the comparison of complete and reduced feature sets. We conducted four experimental setups by testing each feature set with and without feature selection, using Grid Search and Optuna. As feature selection was done within each training partition separately, the selected variables might be different across folds.

To examine various learning algorithms for hospital readmission prediction, nine machine learning classifiers were assessed. Also, several combinations of hyperparameters were evaluated for each classifier. The purpose of this stage was to determine which arrangement produces best results. Since the F1-score is more informative for imbalanced data, it was chosen as the optimization target.

For the optimization of the models, each classifier was tuned with the main hyper-parameters associated to its learning mechanism. We optimized tree-based models, Random Forest, Decision Tree, Gradient Boosting and Extra Trees by varying parameters such as number of estimators, tree depth, split and leaf criteria, feature selection strategy and, where applicable, learning rate and subsampling. KNN: number of neighbors, weighting strategy, distance metric and p value were evaluated. Logistic Regression and SVM were tuned by regularization and kernel related parameters, and Naive Bayes were tuned by variance smoothing. Finally, the SGD was tuned by modifying the loss function, penalty, alpha, learning rate and initial learning rate. The proposed tuning strategy allows a comparison between different families of models in their most relevant configurations.

Model development and performance estimation were performed using nested cross-validation with five outer and two inner folds [17] as shown in Fig. 4. The outer loop was used to estimate the performance on held-out patients and the inner loop was used to select hyperparameters. This protocol was applied to all the proposed feature sets, with and without feature selection, offering a common evaluation framework to the different input representations and experimental settings.

In the outer loop patients were divided into five folds. For each iteration, one-fold was left out as the test set and the remaining four folds were used for hyperparameter optimization and subsequent fitting the model. All consecutive measurements of the same patient were put into the same fold. Thus, no patient contributed measurement pairs to both the model-development and outer-test partitions in the same iteration.



Fold	Trial	F1-Test
1	26	0.6
2	8	0.37
3	34	0.8
4	17	0.22
5	3	0.8

Trials	F1-Score Fo1_Fo2	F1-Score Fo3_Fo4	Average F1-Score
1	0.45	0.75	0.565
2	0.68	0.56	0.62
3	**0.73**	**0.70**	**0.715**
…	…	…	…
40	0.20	0.25	0.225

Fig. 4. Training Methodology

In each outer loop the patients of the four model development folds were further split into 2 inner folds. This division was also performed at the patient level to keep all the measurement pairs from the same individual together. The two inner folds alternated between training and validation so that each hyperparameter configuration was evaluated without using the outer-test patients.

Several hyperparameter settings were tested in the inner loop. The model was trained two times for each configuration, by swapping the roles of the two inner folds. Each training run was evaluated on a validation F1-score and the two scores were averaged. The configurations were then ranked according to

this mean validation F1-score and the one with the highest value was selected for the particular outer iteration.

All preprocessing operations that required estimation from the data were fit only on the training partition used for each run. In the case of Pearson-based feature selection experiments, correlations between the candidate variables and the training target were computed only with the inner-training data. The 50 variables with the highest absolute correlations were kept and this selection was applied unchanged to the corresponding inner-validation data. The ranking was recomputed for each inner-training split instead of being computed once before cross-validation. Therefore, the chosen variables can differ between the two inner training runs as well as between different outer iterations.

Once the optimal hyperparameter setting was identified, the whole preprocessing and classification pipeline was retrained on all patients of the four outer-training folds. For feature selection configurations, Pearson correlations were recomputed only on this complete outer-training partition, and its top 50 ranked variables were kept. And, any other training-dependent preprocessing operations and the classifier were fit on the outer-training data only. This pipeline was then applied to the outer test patients without any modification. Their predictor values and outcomes were not used for fitting of preprocessing operations, feature selection, and hyperparameter optimization.

This procedure was repeated until all five outer folds had used once for testing. Performance was summarized by the mean and standard deviation of the five outer-test F1 scores. The nesting design decoupled the choice of hyperparameters from the estimation of outer-test performance. We performed partitioning at the patient level, ensuring no overlap between individuals in the partitions, and preprocessing and feature selection were done on the fit from the corresponding training data alone, so that held-out observations did not affect these operations.

The same framework was repeated for all feature sets, that is, with and without feature selection, allowing performance comparison to be attributed to the input representation rather than to changes in the protocol.

III. Results and Discussion

Figure 5 shows the F1-scores obtained across the four feature sets using Grid Search and Optuna, with and without Pearson-based feature selection. Points indicate the average F1-score over the 5 outer-test folds, error bars indicate the standard deviation over folds. These summaries report the average performance and the variability between folds, not the confidence intervals.

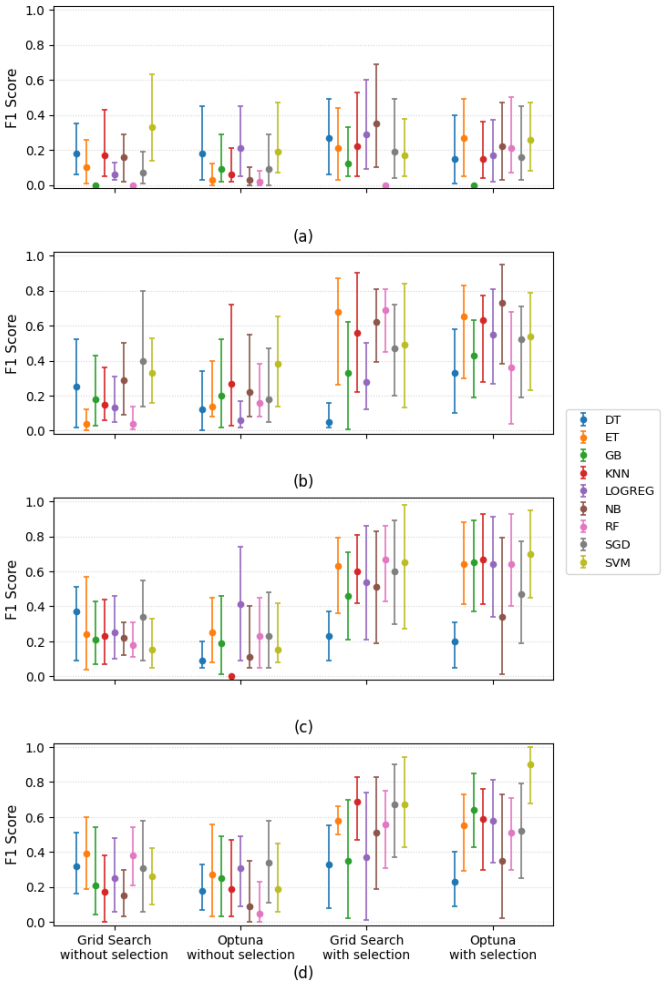


Fig. 5. Mean F1-score of the evaluated classifiers across the four feature sets under different optimization and feature selection settings: (a) feature set 1, (b) feature set 2, (c) feature set 3, and (d) feature set 4.

When Pearson-based feature selection was employed, there was a major improvement. The configurations that used selected features performed better than those that used the all features. This effect became clearer as the number of features in the sets augment, which suggests this selection cut down redundancy. The results also showed an improvement with greater input sets. As shown, feature set 1 had the lowest overall F1-score. This means that basic statistical descriptors were not enough to describe patterns related to readmission. In feature set 2, the performance was more competitive, which suggests that temporal, frequency-domain, and augmented statistical descriptors made a better representation of patient behavior.

The best results were observed in feature sets 3 and 4. These sets included cumulative temporal information, which let the models see changes over time. The best overall performance came from feature set 4, which also included demographic and contextual variables. Therefore, combining wearable descriptors with patient information provided a more robust representation for predicting hospital readmission up to 30 days.

When we look at the optimization strategies, both Grid Search and Optuna perform well. But the best overall results came from using Optuna with feature set 4 and Pearson-based feature selection. The best performance was obtained with SVM achieving, an F1-score of 90.0%. This result shows how important it is using richer temporal representations along with dimensionality reduction and hyperparameter optimization.

Table III shows the performance of the SVM using Feature Set 4, Optuna, and Pearson-based feature selection. The model achieved mean F1-score of 90.00% and mean accuracy of 93.33%. The mean sensitivity, precision and specificity were 100%, 86.67% and 92.00% respectively. Means and sample standard deviations were computed across the five outer-test folds with equal weighting for each fold.

TABLE III. PERFORMANCE OF BEST SVM WITH OPTUNA, AND PEARSON-BASED FEATURE SELECTION.

Metric	Mean Value	Standard Deviation
Accuracy	0.9333	0.1491
F1-score	0.9000	0.2236
Precision	0.8667	0.2981
Sensitivity	1.0000	0.0000
Specificity	0.9200	0.1789

A comparison with the literature was included as a contextual reference because the study used the same original data source. Nevertheless, a direct numerical comparison is not appropriate. The studies differ in prediction tasks, input modalities, model families, and evaluation design. Ali et al [18]. jointly addressed readmission state and length of stay using a multitask deep learning framework and multimodal physical-activity sensory data, whereas the present study focuses on temporal accelerometry representations and patient-level nested evaluation for 30-day readmission.

These results should be seen as exploratory internal validation. The cohort was small, consisting of only 20 patients (8 of whom were readmitted) and the candidate feature sets contained as many as 1006 variables. The small number of independent patients and the large dimensionality constitute a serious risk of overfitting and inaccurate performance estimates. The reported standard deviations describe the variance across folds, and do not replace confidence intervals. Also, the observed sensitivity of 1.0000 does not guarantee that all future readmissions will be captured.

IV. CONCLUSION

This work investigated the prediction of 30-day hospital readmission based on clinical variables and features extracted from accelerometry data. The results showed that feature selection contributed to reducing redundancy in the input set and improving the generalization capability of the models.

The use of Optuna made it possible to obtain more refined models than those previously adjusted with Grid Search, due to its automatic hyperparameter selection strategy, with the SVM classifier standing out by achieving an F1-score of 90.00% and greater consistency throughout the iterations. This result highlights the potential of the proposed strategy for building more robust and effective models in the context of readmission prediction.

ACKNOWLEDGMENT

This article, developed as part of the R&D project on AI models for predicting readmission based on biomarkers from wearable devices, was made possible by Samsung Electronica da Amazonia LTDA, with support from SUFRAMA and implementation by the Federal University of Amazonas, a beneficiary of funds originating from the Information Technology Law for the Western Amazon (Federal Law No. 8,387/1991), with its dissemination and publicity in accordance with the provisions of Article 39 of Decree No. 10,521/2020.

REFERENCES

[1] S. F. Jencks, M. V. Williams, and E. A. Coleman, "Rehospitalizations among patients in the Medicare fee-for-service program," New England Journal of Medicine, vol. 360, no. 14, pp. 1418–1428, Apr. 2009, doi: 10.1056/NEJMsa0803563.

[2] Centers for Medicare & Medicaid Services (CMS), "Hospital Readmissions Reduction Program (HRRP)," CMS.gov. [Online]. Available: https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp. Accessed: Feb. 23, 2026.

[3] O. Hasan et al., "Hospital readmission in general medicine patients: A prediction model," Journal of General Internal Medicine, vol. 25, no. 3, pp. 211–219, Mar. 2010, doi: 10.1007/s11606-009-1196-1.

[4] C. van Walraven, I. A. Dhalla, C. Bell, P. C. Austin, A. J. Forster, J. Stiell, and P. A. Rochon, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community," CMAJ, vol. 182, no. 6, pp. 551–557, Apr. 2010, doi: 10.1503/cmaj.091117.

[5] J. Donzé, D. Aujesky, D. Williams, and J. L. Schnipper, "Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model," JAMA Internal Medicine, vol. 173, no. 8, pp. 632–638, 2013, doi: 10.1001/jamainternmed.2013.3023.

[6] K. Shameer et al., "Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using Mount Sinai heart failure cohort," Pacific Symposium on Biocomputing, vol. 22, pp. 276–287, 2016.

[7] X. Min, B. Yu, and F. Wang, "Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on COPD," Scientific Reports, vol. 9, art. no. 2362, 2019, doi: 10.1038/s41598-019-39071-y.

[8] M. S. Patel et al., "Using remotely monitored patient activity patterns after hospital discharge to predict 30 day hospital readmission: a randomized trial," Scientific Reports, vol. 13, art. no. 8258, 2023, doi: 10.1038/s41598-023-35201-9.

[9] J. Stehlik et al., "Continuous wearable monitoring analytics predict heart failure hospitalization: The LINK-HF multicenter study," Circulation: Heart Failure, vol. 13, p. e006513, 2020, doi: 10.1161/CIRCHEARTFAILURE.119.006513.

[10] A. Doryab, A. K. Dey, G. Kao, and C. Low, "Modeling biobehavioral rhythms with passive sensing in the wild: A case study to predict readmission risk after pancreatic surgery," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 3, no. 1, art. 8, Mar. 2019.

[11] V. Nagarajan, S. P. Shashikumar, A. Malhotra, S. Nemati, and G. Wardi, "Impact of wearable device data and multi-scale entropy analysis on improving hospital readmission prediction," Journal of the American Medical Informatics Association, vol. 31, no. 11, pp. 2679–2688, 2024, doi: 10.1093/jamia/ocae242.

[12] H. H. Yhdego et al., "Prediction of unplanned hospital readmission using clinical and longitudinal wearable sensor features," medRxiv, preprint, Apr. 2023, doi: 10.1101/2023.04.10.23288371.

[13] C. Qian, P. Leelaprachakul, M. Landers, C. Low, A. K. Dey, and A. Doryab, "Prediction of hospital readmission from longitudinal mobile data streams," Sensors, vol. 21, no. 22, art. 7510, 2021, doi: 10.3390/s21227510.

[14] Schricker S, Schmid N, Schanz M, Kimmel M, Alscher M D. In-hospital physical activity measured with a new Bosch accelerometer sensor system (version 1.0). PhysioNet. 2020. RRID:SCR_007345. Available from: https://doi.org/10.13026/yrz5-5909

[15] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," arXiv preprint arXiv:1610.07717, 2016.

[16] K. D. Gong, "Identifying signals of ICU readmissions and post-discharge mortality from physiological time series data using machine learning," Ph.D. dissertation, Johns Hopkins University, 2021.

[17] D. Berrar, “Cross-validation,” preprint submitted to Encyclopedia of Bioinformatics and Computational Biology, 2nd ed., Elsevier, 2024.

[18] S. Ali, S. El-Sappagh, F. Ali, M. Imran and T. Abuhmed, "Multitask Deep Learning for Cost-Effective Prediction of Patient's Length of Stay and Readmission State Using Multimodal Physical Activity Sensory Data," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 12, pp. 5793-5804, Dec. 2022, doi: 10.1109/JBHI.2022.3202178.

Preliminary results of a large dataset of biosignals for training of control algorithms of upper-limb prostheses

1st Julia Martins Frejat
Engenharia Mecatronica
Escola Politecnica da USP
Sao Paulo, Brazil
juliafrejat@usp.br

2nd Chi-Nan Pai
Engenharia Mecatronica
Escola Politecnica da USP
Sao Paulo, Brazil
chinan.pai@usp.br

***Abstract*—The amputation of the upper limbs represents a significant clinical and social challenge, which requires assistive technologies that restore functionality in an intuitive and accessible manner. Although myoelectric prostheses controlled by surface electromyography (sEMG) are the current standard, their performance is still limited by signal variability and the scarcity of robust public datasets for training control algorithms. Widely used datasets such as Ninapro, CapgMyo, and BioPatRec present limitations including small data volume, scarcity of participants with amputations, and lack of standardized acquisition protocols, which hinder their comparison or integration. This study proposes the development of a biosignal acquisition protocol aimed at the development of natural control systems for upper limb prostheses, especially those in development in our group, with more than 10 degrees of freedom and suitable for amputation at the wrist level. The methodology includes the collection of signals from volunteers using an eight-channel wireless sEMG system. Data are acquired during the execution of a set of movements and rest with repeated trials, and the experimental setup aims to minimize artifacts and variability in the data through the electrode placement, the use of wireless equipment and a support structure, and the repetition of movements with rest periods in synchronization with video demonstrations. Although the experimental setup is designed to accommodate individuals with amputations in future data collection, this manuscript describes the preliminary results validated with a sample of 10 able-bodied participants. In the future, the complete dataset, following further expansion, will be made publicly available to support research on upper-limb prosthetics.**

***Index Terms*—Upper-limb prosthesis, surface electromyography, data acquisition protocol, EMG control.**

I. Introduction

Loss of an upper limb can significantly affect an individual's daily life, affecting mobility and balance, compromising the ability to interact with the environment, and introducing psychological and social barriers associated with limb difference [1]. The demand for assistive technologies can be evaluated by the prevalence of this challenge. In 2021, an estimated 140 million people worldwide were living with amputations [2], and approximately 1,500 upper-limb amputations are recorded annually in Brazil through the Unified Health System (SUS) [3].

Prostheses can restore motor function or address the appearance of a missing limb, and their capabilities are continuously improving. However, while complex bionic hand and arm systems are commercially available [4]–[10], their interfaces still present limitations that hinder their use. The cost of these devices is high, the deployment in embedded devices can be restrictive, and control strategies are often limited to a small set of gestures or parameters and are not intuitive, requiring extensive training and cognitive effort from the user [11].

As a control principle, surface electromyography (sEMG) has become the standard. This method is non-invasive, using electrodes placed on the skin to detect biopotential differences caused by ionic currents flowing in the body. These currents are associated with the electrical activity of tissues such as muscles and nerves [12]. In the context of amputations, the residual limb can still receive signals from the nervous system related to the intention to move the amputated limb, and their spatial distribution across the skin contains functional information [13]. Thus, sEMG can be used to decode the user's intention and control a prosthetic device, and there are many studies and commercial prostheses that use this approach. However, this method has limitations that restrict its applicability and performance.

Frequently, surface electrodes are positioned on the forearm to target muscles that control hand and wrist movements, limiting their use to amputations that retain that part of the anatomy [14]. However, over 30% of individuals with upper-limb amputation do not retain the complete forearm musculature [15], which limits the number of electrodes that can be placed on the residual limb and the amount of information that can be captured.

This approach also suffers from the low amplitude of the sEMG signal and its susceptibility to noise and artifacts, which can be caused by several factors such as skin impedance,

This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - grant #409948/2022-5, by the São Paulo Research Foundation (FAPESP) - grant #2024/01468-0 and #2022/06183-9, by the Dean's Office for Research and Innovation of the University of Sao Paulo, by the Endowment Fund Amigos da Poli, and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

sweat, imprecise electrode placement, crosstalk between adjacent muscles, variations in contraction force, and fatigue. These parameters can vary between individuals and across different sessions for the same individual, which can compromise the performance of the control system [16]. Pattern recognition and machine learning techniques have been used to address these issues, but they require representative training data and robust validation to ensure adequate performance in real-world scenarios [17].

An obstacle to advancing sEMG approaches is the lack of available labeled data. Although known databases exist, including Ninapro, BioPatRec, CAPG-MYO, and HD-FW KIN [18]–[22], they present limitations including small sample sizes, relatively few participants with amputations, and the execution of a single repetition per gesture. This prevents control algorithms from being able to generalize between individuals and across different sessions. Additionally, differences in the hardware and acquisition protocols used for each database hinder any comparison or integration [23], [24]. Moreover, these datasets focus on electrodes placed on the forearm, with Ninapro being the only one to include targeting two specific arm muscles and the hand. With such a limited number of signals from the arm, systems produced from the data may not be capable of producing a functional set of gestures for individuals with amputations at higher levels.

The choice of the number of electrodes is also a factor in the design of an sEMG interface. Currently available databases often employ multi-channel setups to capture detailed muscle activity, ranging from 8 to 16 channels. Although these configurations improve pattern recognition accuracy, they increase hardware complexity and computational demand, which can limit their viability. This is reflected in commercial prostheses, which were observed to use 2 to 8 channels. Alternatively, a single-channel system was shown to be capable of identifying finger flexions by mapping the magnitude and density of signal peaks to different muscle layers [25]. However, the performance of this approach can be restrictive when applied to a larger set of gestures, requiring feature reduction methods [26].

In our laboratory, we are developing upper limb prostheses with more than 10 degrees of freedom [27]. One of the prototypes, as shown in Fig. 1, has the capability to perform the main power and precision grips. We intend to control these prostheses using the biosignals generated by muscles and nerves. To solve the problems described earlier, we are building a large dataset to train signal processing algorithms to extract the motion intention of the user. This paper presents the preliminary results of this dataset. The proposed method uses an 8-channel wireless sEMG system to acquire biosignals during the execution of 17 functional gestures. As dexterous prosthetic hand control demands richer information than conventional sEMG alone, this study opted for the placement of electrodes on the arm and forearm without targeting specific muscles. This approach aims to capture information reflecting both neural drive and muscle activation, enabling more natural and robust decoding of movement intention and extending the applicability of the dataset to cases in which a limited number of muscles are available.

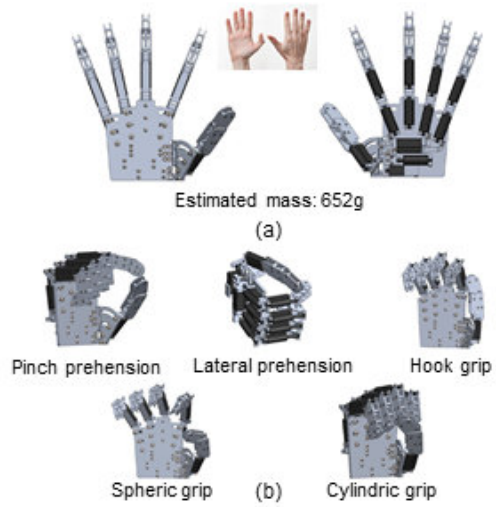


Fig. 1. One of the prototypes of upper limb prostheses under development in our laboratory. (a) Its ventral and dorsal view, (b) the main possible prehensions and grasps

Through the standardization of factors such as electrode placement, rest position, and movement repetition cycles, this work proposes a reproducible data acquisition protocol and introduces a dataset applicable to users with different levels of upper-limb amputations. These contributions aim to support the development of more robust, intuitive, and accessible prosthetic control systems.

II. Methods

A. Experimental Setup

This study uses a sEMG system configured with 8 wireless probes (FREEEMG 1000, BTS Bioengineering S.p.A., Brazil). The use of a wireless system was motivated by electrical safety, achieving electrical isolation from the power supply and thus reducing the risk of electrical shock. Another advantage of this system is the mobility available to participants due to the absence of cables between the probes and the acquisition device, eliminating forces acting on the probes that could compromise adequate positioning and signal acquisition. Additionally, the absence of wires allows movements that better represent gestures performed in daily life.

The eight channels (Ch 1 - Ch 8) were divided into two groups of four, one group (G1) positioned on the arm and the other (G2) on the forearm, as shown in Fig. 2. In the future, in cases of volunteers with amputation, where the forearm is not available, only group G1 will be used.

Aiming to further reduce artifacts from undesired muscle tension, the patients were instructed to remain seated in a comfortable position, with the upper limbs supported on the arm of an armchair. An armchair with adjustable height of the seat and armrests was used to ensure that participants could maintain a relaxed posture independently of their height. As

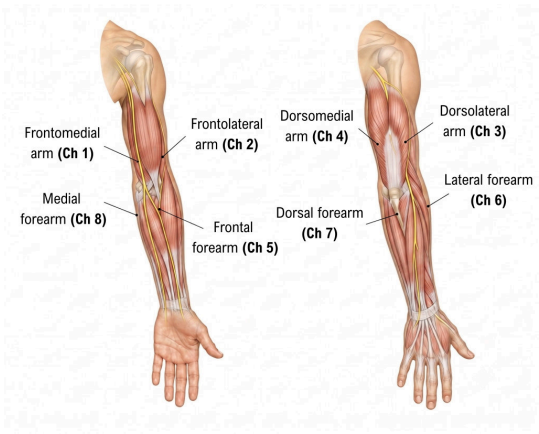


Fig. 2. Target probe placement on the upper limb for biosignal acquisition. The eight measurement sites are organized into two sets of four: group G1 (Ch 1 – Ch 4) targeting the upper arm, and group G2 (Ch 5 – Ch 8) targeting the forearm. The left and right illustrations depict the anterior and posterior views, respectively.

shown in Fig. 3, a chair armrest too low or too high may cause unwanted muscle tension on the back or shoulder of the volunteer.

To avoid forcing the electrodes against the skin due to contact with the chair's armrest, a support structure was developed and manufactured using 3D printing technology. The structure can be mounted on the chair's armrest to elevate the limb and support the hand and forearm in a neutral position, as shown in Fig. 3, preventing pressure on the probes located at the dorsal forearm position. This is defined as the rest position and used as a baseline for gesture executions.

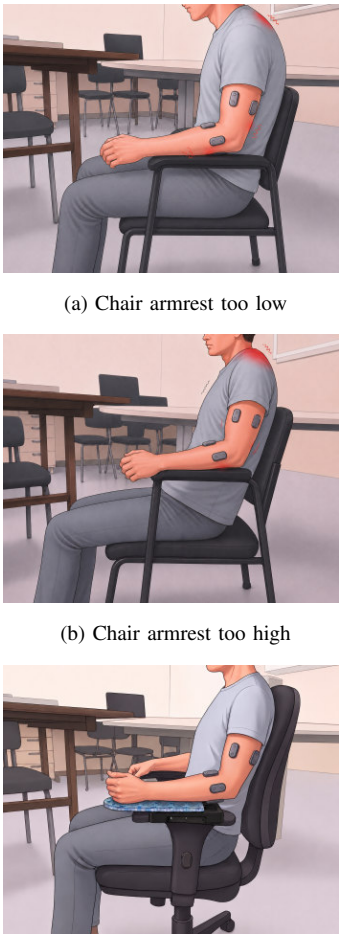
The biosignals are captured using disposable Ag/AgCl electrodes, which are attached to the probes and placed on the skin. Such electrodes are widely used in clinical and research applications due to their low contact impedance, high potential stability, and good adherence to the skin surface, minimizing variability [28]. To minimize the impedance between the skin and the electrodes, the skin was prepared by cleaning the area with 70% isopropyl alcohol.

The data are transmitted to a computer and recorded using the BTS EMGAnalyzer software (BTS Bioengineering S.p.A., Brazil). The software allows for real-time visualization of the signals and data storage, and it is directly compatible with the acquisition system.

B. Participants

The experimental protocol was approved by a research ethics committee, ensured that all procedures were in accordance with the standards for human subjects research, and obtained informed consent from all participants. For this paper, we recruited 10 adult volunteers without upper-limb amputation to evaluate our methodology.

To investigate factors that may introduce variability in sEMG signals, the protocol was designed to document demographic characteristics such as age, sex, height, weight, and dominant hand. These allow for the calculation of the body



(a) Chair armrest too low

(b) Chair armrest too high

(c) Support structure maintaining a neutral posture

Fig. 3. Comparison of forearm support conditions. (a) A low chair armrest causes the participant to lean forward, increasing muscle tension and compressing the dorsal forearm probe. (b) A high chair armrest requires shoulder elevation, increasing muscle tension while the probe remains under pressure. (c) The support structure maintains a neutral forearm posture and prevents pressure on the probes.

mass index (BMI), for example, which may negatively affect the quality of the sEMG signal [29].

III. Execution Procedure

The protocol includes the execution of a set of 17 movements from a baseline rest condition, listed in Table I. The gestures include hand poses and grasp patterns that are functional for daily activities.

To maintain consistency in the execution of the movements, a video demonstration of each gesture was presented to the participants prior to data collection. During execution, the participants were instructed to perform the movement in synchronization with the video, following the same speed and range of motion. For each gesture, participants completed a sequence of 3 repeated trials following the workflow shown in Fig. 4.

The protocol was designed to include rest periods of 5 seconds between each movement to minimize fatigue, which can alter the properties of the biosignal and degrade the quality

TABLE I
MOVEMENTS INCLUDED IN THE EXPERIMENTAL PROTOCOL

ID	Movement
M1	Thumb Abduction/Adduction
M2	Thumb Flexion/Extension
M3	Thumb Opposition/Reposition
M4	Metacarpophalangeal (MCP) Flexion/Extension[a]
M5	Interphalangeal (IP) Flexion/Extension[a]
M6	One (Thumbs Up)
M7	One (Pointing)
M8	Two (V Sign)
M9	Three
M10	Four
M11	Five (Open Hand)
M12	Closed Fist
M13	Pinch Grasp
M14	Lateral Grasp
M15	Spherical Grasp
M16	Cylindrical Grasp
M17	Hook Grasp

[a]For M4 and M5, movements were performed with all four fingers simultaneously and with each finger individually.

of the collected data. Although a 3-second rest period was used in other studies [19], an extended interval was chosen so that it would be longer than the maximum duration of a gesture (4 seconds, including the transition from the rest position to the gesture and back). This ensures that the participant has sufficient time to return to a relaxed state before executing the next movement.

IV. RESULTS

Table II summarizes the baseline characteristics collected for the 10 participants, providing a characterization of the recruited sample.

TABLE II
CHARACTERISTICS OF THE SAMPLE

Characteristic	Total ($N = 10$)
Age (years)	22.0 ± 2.18
Height (cm)	170.3 ± 6.36
Weight (kg)	70.8 ± 14.49
Female	4
Male	6
Right-hand dominant	8
Left-hand dominant	2

Values are mean $\pm$ SD or count (n).

To verify that the experimental setup was effective in minimizing noise in the acquired data, the average Root Mean Square (RMS) amplitude of the signals was calculated for a 10-second interval in the baseline rest position with and without the support structure. The data represent the averaged values across 2 recording sessions from 2 participants, and the results are shown in Table III.

The introduction of the support structure resulted in a clear overall reduction in baseline noise, decreasing the average RMS across the active channels by $39.23\,\%$ (from $12.46\,\mu\text{V}$ to $7.57\,\mu\text{V}$). This behavior was most pronounced in Channel 7, which is the probe located on the dorsal forearm and

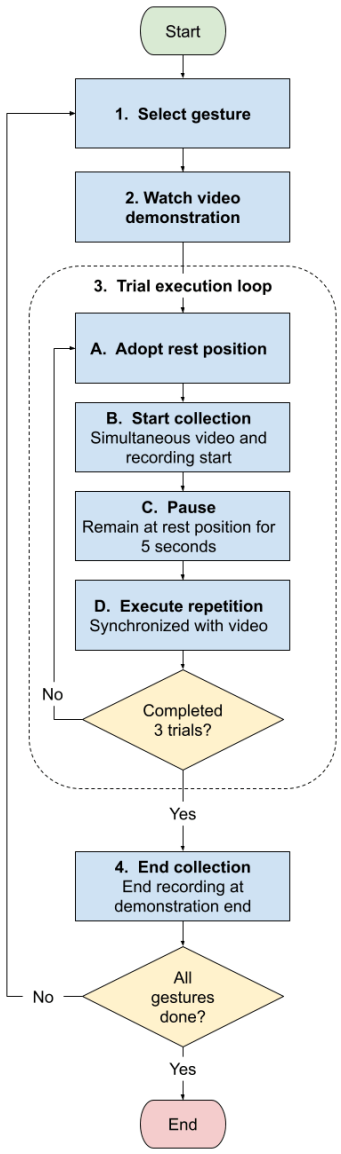


Fig. 4. Workflow of the experimental protocol for biosignal acquisition.

is highly susceptible to artifacts caused by contact with the support chair. Implementing the support structure stabilized this interface, reducing Channel 7's average RMS noise by over $50\,\%$ (averaging a drop from $27.92\,\mu\text{V}$ to $11.32\,\mu\text{V}$).

Noise was minimized in nearly all instances, with 7 out of the 8 evaluated channels showing a direct reduction in baseline RMS. Minor fluctuations were isolated to Channels 4 and 5, which likely reflect small movements during resting. Overall, these results demonstrate that the support structure successfully improves the quality of the collected signals by reducing baseline noise.

To further evaluate how this stabilization impacts control signals, the Signal-to-Noise Ratio (SNR) during active move-

TABLE III
EMG BASELINE NOISE COMPARISON WITH AND WITHOUT MECHANICAL SUPPORT STRUCTURE.

	RMS Noise (μV)		Peak-to-Peak (μV)	
Channel	**No Support**	**Support**	**No Support**	**Support**
Ch 1	6.48	6.27	66.35	62.65
Ch 2	4.46	3.72	44.55	34.45
Ch 3	28.41	11.56	292.10	106.15
Ch 4	14.92	12.78	164.90	124.55
Ch 5	5.14	5.66	46.55	48.95
Ch 6	6.32	4.92	58.85	44.05
Ch 7	27.92	11.32	392.35	160.90
Ch 8	6.03	4.34	72.35	46.75
Mean	**12.46**	**7.57**	**142.25**	**78.56**
Reduction		**-39.23%**		**-44.78%**

ment tasks was assessed using the data from the 2 participants across their 2 respective recording sessions. To obtain the hypothetical SNR without the support (as shown in Table IV), it was assumed that the support structure does not alter the amplitude of the signals captured during gesture execution. Under this assumption, the hypothetical SNR was calculated by comparing the RMS amplitude of the active movement signals (measured with the support structure) against the baseline noise floor measured without the support structure. The actual SNR with support was calculated using the baseline noise floor measured with the structure. The results shown in Table IV indicate that the SNR improved across all evaluated sessions when the support structure was used, with an average gain of $+3.79\,\mathrm{dB}$ across both sessions.

TABLE IV
IMPACT OF THE SUPPORT STRUCTURE ON ACTIVE MOVEMENT SIGNAL-TO-NOISE RATIO (SNR).

Data Group	**Hypothetical SNR (No Support)**	**Actual SNR (With Support)**	**Average SNR Gain**
Trial 1	-2.81 dB	-1.86 dB	+0.94 dB
Trial 2	-3.37 dB	3.26 dB	+6.63 dB
Overall	**-3.09 dB**	**0.70 dB**	**+3.79 dB**

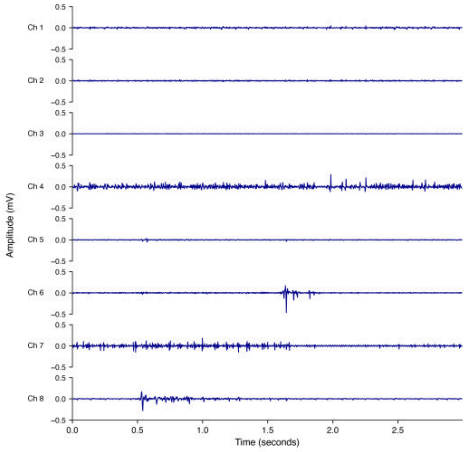
The execution of the movements was also evaluated to verify that the protocol was effective in capturing distinct biosignals for each gesture. Fig. 5 and Fig. 6 illustrate movements M7 (One – Pointing) and M16 (Cylindrical Grasp), respectively, and present the raw signals from a single execution of each gesture from the rest position, demonstrating that the protocol captures distinct activation patterns.

V. DISCUSSION

The experimental protocol proposed in this study addresses gaps in current sEMG datasets by prioritizing standardization while considering anatomical variability. The distribution of probes across the arm and forearm is intentionally designed to support control strategies applicable to different amputation levels where not enough residual musculature is available for conventional electrode placement. By capturing signals from multiple locations without relying on specific muscle targets, the dataset can be used with a broad range of control



(a) One (pointing) gesture

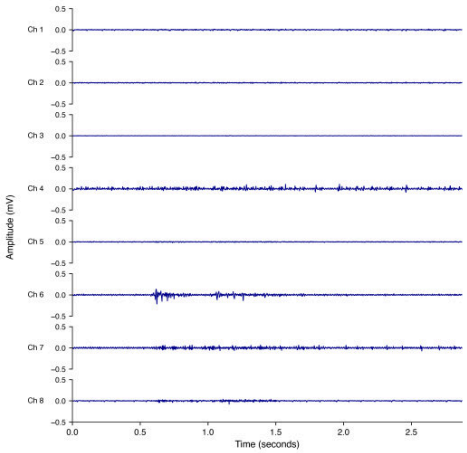


(b) Raw biosignals

Fig. 5. M7 (One – Pointing) gesture and the corresponding raw biosignals recorded by the eight probes.



(a) Cylindrical grasp



(b) Raw biosignals

Fig. 6. M16 (Cylindrical Grasp) gesture and the corresponding raw biosignals recorded by the eight probes.

algorithms for pattern recognition and still provide information for gesture classification.

Design choices such as the use of a wireless system, the design of a support structure, the video demonstrations, and the inclusion of rest periods between movements were made to enhance the quality of the collected data. The wireless system and support structure allow for more natural movements and avoid artifacts caused by cable movement, pressure on the electrodes, and muscle tension due to the need to maintain the limb position. Additionally, to avoid undesired variability in the data, video demonstrations are introduced to ensure that participants perform the movements consistently, and rest periods between movements are included to reduce fatigue-related changes in the sEMG signal.

Furthermore, by not using a muscle-specific approach, this work aims to reduce the cognitive load on users, who often must train the isolated contraction of specific muscles to operate a prosthesis. Instead, the protocol allows for the capture of activity from natural movements, potentially leading to more intuitive control strategies that are not limited to the activation of a restricted set of muscles and thus can be more easily generalized to users with different amputation levels.

VI. Conclusion and Future Work

This study developed a standardized protocol for biosignal acquisition to support the development of control systems for upper-limb prostheses. The protocol includes the use of a wireless system, a support structure, video demonstrations, and repetition cycles with rest periods to limit variability in the data. The protocol was developed to capture signals from a set of 17 functional movements and a rest state, with electrode placement that does not target specific muscles, designed to allow data collection from individuals with different amputation levels by isolating specific channel groups.

A sample of 10 able-bodied participants was used to validate the changes introduced in the protocol, and more data will be collected to constitute the dataset. We started the collection of biosignals with participants without amputation because the upper-limb prostheses being developed in our laboratory are suitable for amputations at the wrist level, and we assume that there will be little difference in biosignals between an individual without amputation and with amputation at the wrist level, considering that the location of the probes is around the elbow, far from the amputation site.

Future work will focus on expanding the data with a larger and more diverse sample, specifically including volunteers with upper-limb amputations to validate the protocol's adaptability and improve generalizability. Upon completion of the ongoing data collection, the full dataset will be made publicly available. Moreover, it will be used to develop and validate control algorithms for upper-limb prostheses, to improve the intuitiveness and expand the applicability of these systems.

Acknowledgment

The authors would like to thank the volunteer subjects who agreed to participate in this study.

References

[1] H. Shahsavari, P. Matourypour, S. Ghiyasvandian, A. Ghorbani, F. Bakhshi, M. Mahmoudi, and M. Golestannejad, "Upper limb amputation; care needs for reintegration to life: An integrative review," *International Journal of Orthopaedic and Trauma Nursing*, vol. 38, p. 100773, Aug. 2020.

[2] Institute for Health Metrics and Evaluation. Who rehabilitation need estimator. [Online]. Available: https://vizhub.healthdata.org/rehabilitation

[3] BRASIL. Ministério da Saúde. DATASUS. (2026) Tabnet. [Online]. Available: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/

[4] M. Bionics. The luke arm. [Online]. Available: https://mobiusbionics.com/

[5] B. Labs. Adam's hand. [Online]. Available: https://bionitlabs.com/it/tecnici-ortopedici/

[6] Össur. Prosthetic hands and digits for upper limb difference. [Online]. Available: https://www.ossur.com/en-us/prosthetics/products/hands

[7] O. Bionics. The hero arm overview. [Online]. Available: https://openbionics.com/en/hero-arm-overview/

[8] PSYONIC. Humans. [Online]. Available: https://www.psyonic.io/ability-hand

[9] G. Motorica. Single grip bionic hand — indy. [Online]. Available: https://global.motorica.org/prosthetics/upper-limb/indy-hand

[10] S. Group. Espire pro elbow. [Online]. Available: https://www.steepergroup.com/prosthetics/upper-limb-prosthetics/wrists/espire-pro-elbow/

[11] T. Kulkarni and R. Uddanwadiker, "Overview: Mechanism and control of a prosthetic arm," *Molecular & Cellular Biomechanics*, vol. 12, no. 3, p. 147–195, 2015.

[12] R. Merletti and G. Cerone, "Tutorial. Surface EMG detection, conditioning and pre-processing: Best practices," *Journal of Electromyography and Kinesiology*, vol. 54, p. 102440, Oct. 2020.

[13] D. Farina, F. Negro, M. Gazzoni, and R. M. Enoka, "Detecting the Unique Representation of Motor-Unit Action Potentials in the Surface Electromyogram," *Journal of Neurophysiology*, vol. 100, no. 3, pp. 1223–1233, Sep. 2008.

[14] D. Segura, E. Romero, V. E. Abarca, and D. A. Elias, "Upper limb prostheses by the level of amputation: A systematic review," *Prosthesis*, vol. 6, no. 2, p. 277–300, Apr. 2024.

[15] F. Cordella, A. L. Ciancio, R. Sacchetti, A. Davalli, A. G. Cutti, E. Guglielmelli, and L. Zollo, "Literature Review on Needs of Upper Limb Prosthesis Users," *Frontiers in Neuroscience*, vol. 10, May 2016.

[16] S. Benatti, E. Donati, A. Moin, M. Zanghieri, M. Orlandi, A. Burrello, F. Artoni, S. Micera, L. Benini, and J. M. Rabaey, "Emg acquisition and processing for hand movement decoding on embedded systems: State of the art and challenges," *Proceedings of the IEEE*, vol. 113, no. 3, p. 256–286, Mar. 2025.

[17] B. Abdikenov, D. Zholtayev, K. Suleimenov, N. Assan, K. Ozhikenov, A. Ozhikenova, N. Nadirov, and A. Kapsalyamov, "Emerging frontiers in robotic upper-limb prostheses: Mechanisms, materials, tactile sensors and machine learning-based emg control: A comprehensive review," *Sensors*, vol. 25, no. 13, p. 3892, Jan. 2025.

[18] M. Atzori, A. Gijsberts, S. Heynen, A.-G. M. Hager, O. Deriaz, P. Van Der Smagt, C. Castellini, B. Caputo, and H. Muller, "Building the Ninapro database: A resource for the biorobotics community," in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. Rome, Italy: IEEE, Jun. 2012, pp. 1258–1265.

[19] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific Data*, vol. 1, no. 1, p. 140053, Dec. 2014.

[20] M. Ortiz-Catalan, R. Brånemark, and B. Håkansson, "Biopatrec: A modular research platform for the control of artificial limbs based on pattern recognition algorithms," *Source Code for Biology and Medicine*, vol. 8, p. 11, Apr. 2013.

[21] Q. Dai, X. Li, W. Geng, W. Jin, and X. Liang, "Capg-myo: A muscle-computer interface supporting user-defined gesture recognition," in *Proceedings of the 9th International Conference on Computer and Communications Management*, ser. ICCCM '21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, p. 52–58. [Online]. Available: https://dl.acm.org/doi/10.1145/3479162.3479170

[22] W. Guo, Z. Zhao, Z. Zhou, Y. Fang, Y. Yu, and X. Sheng, "Hand kinematics, high-density sEMG comprising forearm and far-field potentials for motion intent recognition," *Scientific Data*, vol. 12, no. 1, p. 445, Mar. 2025.

[23] H. Le, M. In Het Panhuis, G. M. Spinks, and G. Alici, "The effect of dataset size on emg gesture recognition under diverse limb positions," in *2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, Sep. 2024, p. 303–308, iSSN: 2155-1782. [Online]. Available: https://ieeexplore.ieee.org/document/10719858/

[24] E. Farago, D. MacIsaac, M. Suk, and A. D. C. Chan, "A review of techniques for surface electromyography signal quality analysis," *IEEE Reviews in Biomedical Engineering*, vol. 16, p. 472–486, 2023.

[25] D. K. Kumar, S. Poosapadi Arjunan, and V. P. Singh, "Towards identification of finger flexions using single channel surface electromyography – able bodied and amputee subjects," *Journal of NeuroEngineering and Rehabilitation*, vol. 10, no. 1, p. 50, Jun. 2013.

[26] M. Arozi, W. Caesarendra, M. Ariyanto, M. Munadi, J. D. Setiawan, and A. Glowacz, "Pattern Recognition of Single-Channel sEMG Signal Using PCA and ANN Method to Classify Nine Hand Movements," *Symmetry*, vol. 12, no. 4, p. 541, Apr. 2020.

[27] D. Pascotto Costa and C. N. Pai, "A novel design of robotic hand with linear motors," in *Proceedings of the 26th International Congress of Mechanical Engineering*. ABCM, 2021. [Online]. Available: http://abcm.org.br/anais-de-eventos/COB2021/0570

[28] A. Albulbul, "Evaluating major electrode types for idle biological signal measurements for modern medical technology," *Bioengineering*, vol. 3, no. 3, p. 20, Aug. 2016.

[29] M. Atzori, A. Gijsberts, I. Kuzborskij, S. Elsig, A.-G. Mittaz Hager, O. Deriaz, C. Castellini, H. Müller, and B. Caputo, "Characterization of a Benchmark Database for Myoelectric Movement Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 1, pp. 73–83, Jan. 2015.

Preprocessing and Physiological Model Fitting of Smartphone-Based Pupillary Light Reflex Recordings

Jian L. B. Veras
Electronics Engineering Division
Aeronautics Institute of Technology
São José dos Campos, BR
jianlbv@gmail.com

João Gabriel B. de Souza
Electronics Engineering Division
Aeronautics Institute of Technology
São José dos Campos, BR
joao.souza.102297@ga.ita.br

Marcus H. V. Júnior
Electronics Engineering Division
Aeronautics Institute of Technology
São José dos Campos, BR
marcus@ita.br

Monica M. S. Matsumoto
Electronics Engineering Division
Aeronautics Institute of Technology
São José dos Campos, BR
matsumoto@ita.br

***Abstract*—The pupillary light response (PLR) is the autonomic iris response to light and enables noninvasive observation of pupil diameter changes. Dynamic PLR responses were acquired with a smartphone-based system during a clinical study of depression treatment in elderly. We proposed a preprocessing and model-fitting pipeline for the acquired PLR signals. 1438 videos from 62 patients were analyzed, considering acquisitions from both eyes across different treatment weeks. The recordings were intrinsically noisy and presented common challenges such as missing samples, blink-related artifacts, and tracking errors. We proposed a data quality assessment that retained 1069 videos from 60 patients. Outliers were removed using the Hampel method, short intervals of missing samples were interpolated according to their duration, and complete recordings were excluded considering the amount of missing values thresholds. The resulting curves were fit to a second-order nonlinear PLR model using Nelder-Mead optimization. We analyzed different weights for the light stimulation region in the cost function. Patient medians of root mean squared error (RMSE) were compared using two-sided Wilcoxon signed-rank tests with Holm correction. Median stimulation RMSE decreased from 0.180 to 0.084 and 0.052, whereas global RMSE increased from 0.097 to 0.109 and 0.188. Early accommodation error also increased. All nine comparisons were significant (adjusted p ¡ 0.001). However, 99.8% of runs reached the 300-iteration limit; numerically valid fits were retained regardless of convergence. Under the same configuration, increasing the stimulation weight improved local agreement at the expense of other regions, and did not provide a general improvement in PLR fitting.**

***Index Terms*—pupillary light reflex, pupillometry, preprocessing, Nelder-Mead, physiological modeling, smartphone-based acquisition**

I. Introduction

The pupillary light reflex (PLR) is the constriction and subsequent dilation of the pupil in response to light stimulation. This response is mediated by the antagonistic action of the iris sphincter and dilator muscles, which are modulated by the parasympathetic and sympathetic branches of the autonomic nervous system. Because of this relationship, PLR analysis has been explored as a noninvasive approach to assess autonomic function and neurological or psychiatric conditions [4], [5], [8].

Previous work developed a low-cost smartphone-based pupillometer and a computer vision pipeline to extract pupil size from PLR videos [3]. In that study, the authors used *YOLOv4* to estimate pupil radius and fitted the resulting pupilograms to a second-order nonlinear physiological model [3], [5]. The initial dataset included 80 recordings from five patients, collected as part of a clinical study involving older adults with depression [1]–[3].

The present work extends this previous effort by evaluating a larger set of clinical PLR recordings and addressing the methodological challenges that arise before biomarker analysis. This paper investigates the preprocessing steps used to prepare pupilograms for model fitting and the limitations of fitting a physiological PLR model using Nelder-Mead optimization with a weighted cost function [3], [10]. This focus is relevant because noisy pupil-size data, missing samples, blinks, saccades, and tracking errors can distort the estimated curve and lead to unstable or physiologically implausible model parameters [4], [7]–[9].

The main contributions of this work are: (i) a regional preprocessing strategy for smartphone-based PLR recordings; (ii) a data quality assessment of 1438 clinical videos; (iii) a segment-specific exclusion criterion based on missing samples after preprocessing; and (iv) an evaluation of how different weights in the fitting cost function affect global and regional fitting errors.

II. Methods and Materials

A. Data Source and Acquisition Protocol

The analyzed data were obtained from PLR videos acquired using the smartphone-based pupillometer developed in previous work [3]. The recordings were collected from 62 patients diagnosed with depression as part of a clinical research protocol conducted from January 2019 to March 2023 at the Institute of Psychiatry, Hospital das Clínicas FMUSP, University of São Paulo [1], [2]. Participants were between 60 and 87 years old, with a mean age of 67.5 years. The sample included xx men and yy women.

Recordings were planned at three time points: before the first theta-burst stimulation session, after the first session, and four weeks after treatment initiation, following the treatment session scheduled for that week. Before each set of recordings, participants underwent a 5 min period of adaptation to darkness in a dark room. At each time point, recordings were planned for both eyes under four auditory conditions: no audio, neutral audio, pleasant audio, and unpleasant audio. Each video lasted 20 s and was acquired at 20 Hz, yielding 400 samples per recording. The first 5 s were recorded in darkness to estimate the baseline pupil diameter. This interval was followed by 1 s of smartphone light stimulation to induce pupil constriction. The remaining 14 s were recorded in darkness to capture pupil redilation and recovery [3]. At 20 Hz, consecutive samples are separated by 50 ms. This sampling rate has been described as sufficient for comparisons of pupil size between experimental conditions in cognitive pupillometry, although higher sampling rates are desirable for measures such as maximum constriction velocity [7].

A total of 1438 videos were analyzed. 369 videos were discarded after preprocessing and exclusion based on recording quality, leaving 1069 videos from 60 patients. Two patients had no recordings that met the quality criteria. Across the original cohort of 62 patients, the mean number of videos per patient decreased from 23.2 (SD = 2.4) before filtering to 17.2 (SD = 6.4) after filtering. The latter calculation included the two patients with no retained recordings.

B. PLR Segmentation

Each recording was divided into four temporal regions according to the expected physiological behavior of the PLR curve and the acquisition protocol. The segmentation was also used to guide outlier removal, missing values assessment, exclusion criteria, and regional fitting evaluation. Each recording was divided into four regions: pupil baseline, frames 0 to 5.00 s; light stimulation, from 5.00 s to 6.00 s; early accommodation, from 6.00 s to 12.00 s; and final accommodation, from 12.00 s to 20.00 s. The boundary between early and final accommodation corresponded to 6 s after light offset, or 12 s after the start of the recording. This region defines the PLR feature named post-illumination pupil response (PIPR), quantified approximately six seconds after light offset in studies of depressive disorders [13].

C. Data Treatment Pipeline

Fitting nonlinear physiological models described by ordinary differential equations (ODEs) can be challenging when the objective function contains multiple local minima. Different parameter sets may produce similar fitted curves. Local derivative-free algorithms such as Nelder-Mead may converge to suboptimal solutions, particularly when the available data are noisy or incomplete [6]. The preprocessing pipeline was therefore designed to address two issues relevant to subsequent model fitting: outliers that could disproportionately influence the cost function and missing samples in regions containing rapid changes in pupil size. This approach follows pupillometry guidelines that recommend identifying invalid samples and addressing missing data and artifacts before analysis [7]–[9].

Missing samples reduce the information available for fitting and can leave some portions of the response less represented in the objective function. We therefore quantified the proportion of missing samples in each recording and in each region described in Section II-B, defined as the number of missing samples over the expected number of samples for the period. This assessment characterized the temporal distribution of missing data, while the duration of each consecutive gap was evaluated to determine its eligibility for interpolation.

Outliers can disproportionately influence the cost function. Following recommendations to identify invalid samples before interpolation [7], [8], we applied the Hampel method to the complete recording. This method evaluates the deviation of each sample from the local median using the median absolute deviation (MAD) as a measure of local variability [12]. The implementation was configured with `window_size = 5` and `n_sigma = 2.0`. Samples identified as outliers were marked as missing for subsequent gap filling.

Following outlier removal, gaps were filled according to the number of consecutive missing samples. Linear and cubic spline interpolation have both been described as approaches to reconstruct short missing intervals in pupillometry [7], [8]. Gap duration was defined as the number of missing samples divided by the sampling rate. Gaps of one or two samples, corresponding to up to 100 ms, were filled using linear interpolation. Gaps of three to five samples, corresponding to more than 100 ms and up to 250 ms, were filled using cubic spline interpolation. Both methods required two consecutive valid samples immediately before and two immediately after the gap. Neighbor eligibility was assessed using the missing-data mask before interpolation, so interpolated samples could not support the filling of another gap. No extrapolation was performed. Gaps longer than five samples or without the required neighbors were left unfilled. The 250-ms cutoff was informed by a configuration reported by Kret and Sjak-Shie [8].

D. Recording Exclusion Criteria

Previous studies have discussed the use of missing-data thresholds to exclude recordings with insufficient usable data [4], [8], [9]. After outlier removal and gap filling, the proportion of remaining missing samples was evaluated separately

for each PLR region. A stricter threshold was applied to the light stimulation region because it was shorter and had a lower proportion of missing samples in the raw recordings. A recording was excluded if the proportion of missing samples exceeded 10% during light stimulation or 30% in any of the other three regions.

E. Trial Aggregation

After recording exclusion, recordings from the same patient, eye, and acquisition time point were aggregated across auditory conditions by computing the mean of the available values at each corresponding frame. This strategy was intended to attenuate variability unrelated to the light stimulus, following the principle of averaging responses aligned to a common event [9]. Frames with no available values across the contributing recordings remained missing. Auditory conditions were not compared in the present analysis; each resulting curve represented the average response across the available conditions. This provided 328 PLR signals for 60 patients, with a mena number of 5.5 (SD=0.9) signals per patient.

F. Physiological PLR Model

We fitted the data to a second-order nonlinear physiological model proposed for transient PLR responses induced by short light flashes [5]. This model represents the iris dynamics using viscoelastic properties and autonomic inputs associated with parasympathetic and sympathetic modulation. We cosnidered this model for the present study because it provides a compact physiological description of PLR dynamics while remaining feasible for numerical fitting.

The model can be written in simplified form as a second-order differential equation:

$$\frac{d^2r}{dt^2}+D\frac{dr}{dt}+K_c(r-l_{0c})^2+K_d(r-L_{0d})^2 = F_s(t)-F_p(t)+P_0, \tag{1}$$

where r represents pupil radius, D is the damping term, K_c and K_d are elastic constants, l_{0c} and L_{0d} are resting-related parameters, $F_p(t)$ and $F_s(t)$ represent parasympathetic and sympathetic inputs, and P_0 is a static force term [5].

G. Nelder-Mead Optimization

The model parameters were estimated using the Nelder–Mead algorithm, a derivative-free simplex method for function minimization [10]. This algorithm was retained from the previous smartphone pupillometer implementation [3]. The same initial parameter vector and optimization settings were used for every mean curve and weight condition, as summarized in Table I. Each fit used a single initialization, and sensitivity to alternative initial values was not evaluated.

Before optimization, each mean curve was smoothed using a Savitzky-Golay filter with a window of five samples and a polynomial order of three. Remaining missing samples were temporarily filled to permit filtering: interior gaps were filled by linear interpolation, and missing values at the recording boundaries were filled using the nearest available value. The missing-data mask obtained before this temporary filling was then restored. These positions were excluded from both the error sums and the corresponding sample counts in the objective function. Values reconstructed during the earlier preprocessing stage remained eligible for inclusion.

Let I denote the set of sample indices with available values in the input mean curve before temporary filling, and let I_L denote the subset within the light stimulation interval. For either set A, the mean squared error was defined as

$$\mathrm{MSE}_A(\theta) = \frac{1}{|A|}\sum_{i\in A}\left[\tilde{y}_i - \hat{y}_i(\theta)\right]^2, \tag{2}$$

where $\tilde{y}_i$ is the smoothed reference value and $\hat{y}_i(\theta)$ is the model output for parameter vector θ. The objective function was

$$J(\theta) = 10^{-4}\left[\mathrm{MSE}_I(\theta) + (w-1)\mathrm{MSE}_{I_L}(\theta)\right], \tag{3}$$

where w was set to 1, 2, or 10. The factor 10^{-4} corresponds to dividing each residual by 100 before squaring, as implemented in the optimization routine. With $w = 1$, the objective reduces to the scaled global mean squared error. Larger values of w increase the contribution of the additional error term for the light stimulation interval. These conditions were compared to assess how emphasizing this interval affected global and regional fitting errors.

The model equations were integrated using `odeint`, which uses the LSODA solver, with the default integration tolerances of the recorded SciPy version. Parameter admissibility was checked using the restrictions listed in Table I. Optimization was limited to 300 iterations. Convergence required both the parameter tolerance of 10^{-4} and the objective-function tolerance of 10^{-6} to be satisfied, as assessed from the spread of the simplex vertices and their objective values in the SciPy implementation. Termination at the iteration limit was recorded separately from successful convergence.

For each fit, we recorded the optimizer termination status, number of iterations, number of objective-function evaluations, final integration status, and compliance with the implemented parameter restrictions. Failed integrations during objective-function evaluations were also counted, and distances from the final parameter values to the applicable limits were stored. Numerical convergence was distinguished from successful integration and parameter admissibility. These diagnostics characterize numerical behavior and do not establish the physiological accuracy of the fitted parameters.

H. Fitting Evaluation and Statistical Analysis

Fitting performance was evaluated using global and regional root mean squared error (RMSE), comparing the model output with the preprocessed mean curve before Savitzky-Golay smoothing. Remaining missing samples were excluded, whereas samples reconstructed during preprocessing were retained. Numerical validity required successful final integration, compliance with the implemented parameter restrictions, and

TABLE I
PREPROCESSING SETTINGS, INITIAL PARAMETER VALUES, AND OPTIMIZATION CRITERIA.

Model parameter	**Initial value**	**Implemented restriction**
r_0	3.00	No explicit bound
l_{0d}	3.59	$l_{0d} \geq 0$
l_{0c}	0.92	$l_{0c} \geq 0$
k_d	1.29	$k_d \geq 0$
k_c	0.059	$k_c \geq 0$
D	3.78	No explicit bound
f_p	12.73	No explicit bound
f_s	9.10	No explicit bound
p_0	-0.736	No explicit bound
t_p	0.26	$0 < t_p < 1$ s
t_s	0.64	$0 < t_s < 1$ s
Processing setting	**Value**	**Details**
Sampling rate	20 Hz	400 samples per recording
Hampel parameters	`window_size = 5; n_sigma = 2.0`	Applied to the complete recording; detected outliers marked as missing
Linear interpolation	1–2 missing samples	Gap duration up to 100 ms
Cubic spline interpolation	3–5 missing samples	Gap duration greater than 100 ms and up to 250 ms
Interpolation support	Two valid samples on each side	Consecutive samples immediately adjacent to the gap; eligibility assessed before filling; no extrapolation
Recording exclusion	$> 10\%$ missing during light stimulation; $> 30\%$ in any other region	Assessed after outlier removal and gap filling
Savitzky–Golay filter	Window: 5 samples; polynomial order: 3	SciPy default edge mode, `interp`
Residual scaling	Division by 100	Equivalent to multiplying the combined MSE terms by 10^{-4}
Light stimulation weights	1, 2, 10	Same initialization and optimization settings for all weights
Optimization setting	**Value**	**Details**
Algorithm	Nelder–Mead	Same initialization per curve and weight; default SciPy initial simplex
Initial model state	$r(0) = r_0$; $\dot{r}(0) = 0$	r_0 estimated during optimization
ODE integration	`odeint` / LSODA	Default SciPy integration tolerances
Maximum iterations	300	`maxiter`
Parameter tolerance	10^{-4}	`xatol`
Objective-function tolerance	10^{-6}	`fatol`, applied to the scaled objective
Software versions	SciPy: 1.17.1; `hampel`: 1.0.2	Versions used for the reported analysis

finite RMSE values. Optimizer convergence was assessed separately, and reaching the iteration limit did not by itself invalidate a fit. Failure and convergence rates were summarized by weight and patient, together with the number of eligible curves per patient [16].

Patients were eligible if at least one of their mean curves produced valid, converged fits under all three weights. For these patients, the primary analysis also included additional curves with numerically valid fits under all three weights, even when the optimizer reached the iteration limit. For each patient and weight, RMSE was summarized by the median over the same eligible curves, preserving comparisons within patients and avoiding treatment of repeated curves as independent observations.

Patient medians were compared using two-sided Wilcoxon signed-rank tests for weights 1 versus 2, 1 versus 10, and 2 versus 10 [14]. Comparisons covered global, light stimulation, and early accommodation RMSE. Holm correction was applied jointly to these nine comparisons at a significance level of 0.05 [15]. The tests assume approximately symmetric paired differences. Patient medians were summarized using the median and interquartile range. Pairwise comparisons were reported with rank-biserial effect sizes and Holm-adjusted p-values. The number of patients included in the analysis was also reported.

III. RESULTS AND DISCUSSION

A total of 1438 PLR videos from 62 patients were analyzed. After data treatment and quality assessment, 369 videos were discarded (25.7%), leaving 1069 videos from 60 patients. Two patients had all their recordings excluded. Across the original 62 patients, the mean number of videos decreased from 23.2 (SD = 2.4) before filtering to 17.2 (SD = 6.4) afterward, including patients with no remaining recordings. The greater dispersion in the number of retained recordings indicates that exclusion affected some patients more than others.

Fig. 1 shows the distribution of the proportion of missing samples across recordings, separately for each PLR region, before and after data treatment. The light stimulation interval presented a lower proportion of missing samples, whereas early accommodation showed greater missingness. Improved pupil detection under illumination and artifacts associated with blinking or tracking instability are possible explanations for

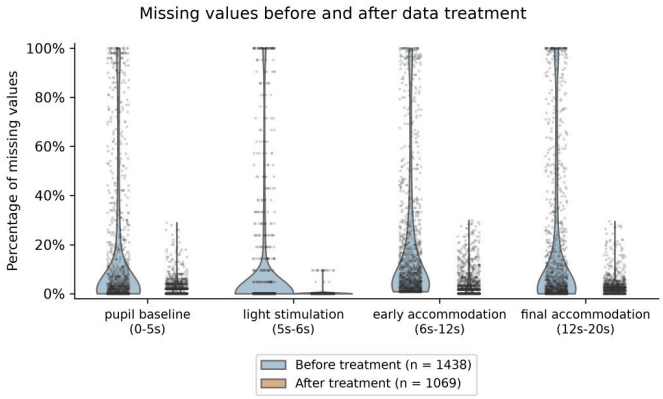


Fig. 1. Distribution of the proportion of missing samples per recording in the four PLR regions, before and after data treatment.

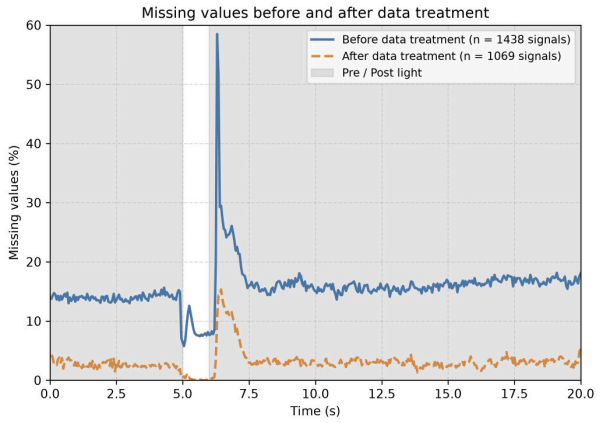


Fig. 2. Percentage of recordings with a missing sample at each time point, before and after data treatment.

these differences, but their contributions were not directly assessed. Some recordings contained few or no valid samples within individual regions, supporting the use of regional quality assessment.

The temporal distribution of missing data was also assessed by computing the percentage of recordings with a missing value at each frame, as shown in Fig. 2. Missing samples were more frequent near stimulus onset and during early accommodation, indicating that data availability varied throughout the acquisition protocol. Interpolation reduced missingness in eligible intervals, but the resulting values were reconstructed estimates. These temporal patterns alone do not establish the cause of data loss.

The retained recordings yielded 328 mean curves after trial aggregation. Each curve was fitted using weights of 1, 2, and 10, resulting in 984 optimization runs. Final ODE integration was successful in all runs. However, only one run under weight 1 and one under weight 2 satisfied the optimizer convergence criteria before reaching the iteration limit; none did so under weight 10. The remaining 982 runs (99.8%) reached the limit of 300 iterations. Visual inspection of the objective-function histories showed little further reduction in cost near

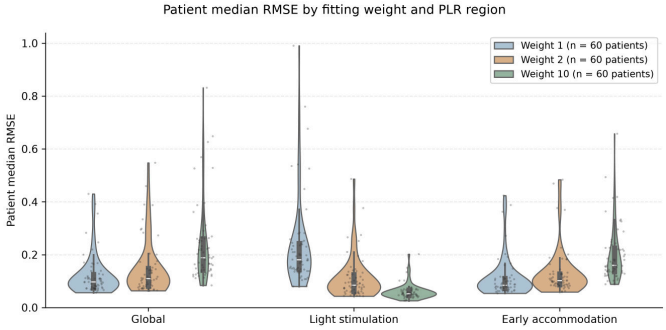


Fig. 3. Distributions of patient median RMSE for the global curve,light stimulation, and early accommodation under weights 1, 2, and 10. Each point represents one patient. Internal boxplots summarize the distributions, and the same eligible curves were used across weights within each patient.

the iteration limit, despite the unmet convergence criteria. Fits without optimizer convergence were retained when their final integration was successful, the implemented parameter restrictions were satisfied, and the required RMSE values were finite under all three weights. The resulting comparisons included 60 patients.

We assessed how the weight assigned to the light stimulation interval affected global RMSE, light stimulation RMSE, and early accommodation RMSE. Fig. 3 shows the distributions of the patient medians for these measures. With weight 1, the median RMSE during light stimulation was 0.180, exceeding the global median of 0.097. Thus, under the unweighted objective, the stimulation interval showed greater error than the recording as a whole.

Increasing the stimulation weight reduced the median RMSE within that interval to 0.084 and 0.052 for weights 2 and 10, respectively. However, the global median RMSE increased from 0.097 to 0.109 and 0.188, while the early accommodation median increased from 0.085 to 0.102 and 0.160. The interquartile range also increased with weight for global and early accommodation RMSE, whereas it decreased for light stimulation RMSE.

All nine paired comparisons were statistically significant after Holm correction (Table II). Weight 10 produced higher patient median RMSE than weights 1 and 2 for both the global and early accommodation measures in all 60 patients. Conversely, weights 2 and 10 produced lower light stimulation RMSE than weight 1 in every patient. These consistent directions yielded rank-biserial correlations of $+1$ or -1, indicating unanimity in the direction of the paired differences rather than a proportional change in error.

Most fits reached the iteration limit, so these findings describe performance with the selected initialization and a maximum of 300 iterations. Apparent plateaus in the objective function can coexist with unmet convergence criteria. Initialization, parameter scaling, objective-function geometry, and recording characteristics may contribute to this behavior, but their effects were not assessed separately. The observed errors therefore reflect the combined effects of preprocessing, the

TABLE II
PAIRED COMPARISONS OF PATIENT MEDIAN RMSE USING THE WILCOXON SIGNED-RANK TEST. ALL COMPARISONS INCLUDED 60 PATIENTS.

RMSE measure	Weights	r_{rb}	p_{Holm}
Global	1 vs. 2	$+0.674$	1.114×10^{-5}
	1 vs. 10	$+1.000$	1.467×10^{-10}
	2 vs. 10	$+1.000$	1.467×10^{-10}
Light stimulation	1 vs. 2	-1.000	1.467×10^{-10}
	1 vs. 10	-1.000	1.467×10^{-10}
	2 vs. 10	-0.969	1.976×10^{-10}
Early accommodation	1 vs. 2	$+0.666$	1.114×10^{-5}
	1 vs. 10	$+1.000$	1.467×10^{-10}
	2 vs. 10	$+1.000$	1.467×10^{-10}

r_{rb}: matched-pairs rank-biserial correlation. Differences were calculated as the higher weight minus the lower weight; positive values indicate higher RMSE under the higher weight. p_{Holm}: p-value under Holm correction. Holm correction was applied jointly to the nine comparisons.

model, and optimizer configuration.

The model did not explicitly account for variation in eye position. Furthermore, physiological parameter accuracy was not established through comparison with a reference pupillometer or repeatability analysis. Lower fitting error should therefore be interpreted as closer agreement with the processed recordings, without establishing the suitability of the estimated parameters as biomarkers.

IV. CONCLUSION

This work evaluated a data treatment and model fitting pipeline for smartphone-based PLR recordings acquired in a clinical context. Of 1438 videos, 369 were excluded using regional quality criteria, leaving 1069 recordings from 60 patients. Averaging recordings across auditory conditions within each patient, eye, and acquisition time point yielded 328 mean curves for fitting.

Motivated by the suggestion that greater emphasis on the light stimulation interval could improve PLR fitting [3], this study examined the consequences of increasing its weight in the objective function. Under the evaluated settings, this strategy reduced stimulation error but increased global and early accommodation error, despite the lower proportion of missing samples during stimulation. These findings do not support increasing the stimulation weight alone as a general solution for improving the complete PLR fit. Agreement outside this interval also matters when the fitted curve is used to estimate baseline pupil diameter or PIPR, since errors in the corresponding regions may affect these measurements.

The fact that 99.8% of optimization runs reached the iteration limit constrains the interpretation of these findings. Only one initialization was evaluated, and the effects of alternative initial values, parameter scaling, and stopping criteria were not examined. The observed lack of convergence may reflect a combination of these choices and the characteristics of the model and recordings. This study therefore does not establish whether Nelder-Mead is suitable for this problem under other configurations; it characterizes the fitting errors obtained with the selected initialization and computational budget.

Future work should evaluate these optimization choices jointly with the weighting strategy, including alternative optimization methods and robust loss functions. Comparison with reference measurements and repeatability assessment are also needed to establish the physiological accuracy and stability of the estimated parameters before biomarker analysis.

ACKNOWLEDGMENT

The authors thank the University of Sao Paulo research team and all collaborators involved in the clinical data collection and previous smartphone-based pupillometer work.

REFERENCES

[1] Valiengo L, Pinto BS, Marinho KAP, Santos LA, Tort LC, Benatti RG, Teixeira BB, Miranda CS, Cardeal HB, Suen PJC, Loureiro JC, Vaughan RAR, Dini Mattar RAMPF, Lessa M, Oliveira PS, Silva VA, Gattaz WF, Brunoni AR, Forlenza OV. Treatment of depression in the elderly with repetitive transcranial magnetic stimulation using theta-burst stimulation: Study protocol for a randomized, double-blind, controlled trial. Front Hum Neurosci. 2022 Aug 31;16:941981. doi: 10.3389/fnhum.2022.941981. PMID: 36118977; PMCID: PMC9471379.

[2] L. da Costa Lane Valiengo et al., "Bilateral theta-burst stimulation of dorsolateral prefrontal cortex regions for late-life depression: A randomized, sham-controlled clinical trial," *American Journal of Psychiatry*, early access, Mar. 11, 2026, doi: 10.1176/appi.ajp.20250537.

[3] D. H. V. Barroso, "Smartphone-based pupillometer design and automatic PLR curve identification," M.S. dissertation, Instituto Tecnologico de Aeronautica, Sao Jose dos Campos, Brazil, 2021.

[4] H. M. Pinheiro and R. M. da Costa, "Pupillary light reflex as a diagnostic aid from computational viewpoint: A systematic literature review," *Journal of Biomedical Informatics*, vol. 117, Art. no. 103757, May 2021, doi: 10.1016/j.jbi.2021.103757.

[5] X. Fan and G. Yao, "Modeling transient pupillary light reflex induced by a short light flash," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 36–42, Jan. 2011.

[6] A. Abdullah, S. Deris, S. Anwar, and S. N. V. Arjunan, "An evolutionary firefly algorithm for the estimation of nonlinear biological model parameters," *PLOS ONE*, vol. 8, no. 3, Art. no. e56310, 2013.

[7] S. Mathot and A. Vilotijevic, "Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis," *Behavior Research Methods*, vol. 55, pp. 3055–3077, 2023.

[8] M. E. Kret and E. E. Sjak-Shie, "Preprocessing pupil size data: Guidelines and code," *Behavior Research Methods*, vol. 51, pp. 1336–1342, 2019.

[9] M. B. Winn, D. Wendt, T. Koelewijn, and S. E. Kuchinsky, "Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started," *Trends in Hearing*, vol. 22, pp. 1–32, 2018.

[10] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[11] R. R. Barton and J. S. Ivey, Jr., "Nelder-Mead simplex modifications for simulation optimization," *Management Science*, vol. 42, no. 7, pp. 954–973, 1996.

[12] Roos-Hoefgeest Toribio M, Garnung Menéndez A, Roos-Hoefgeest Toribio S, Álvarez García I. A Novel Approach to Speed Up Hampel Filter for Outlier Detection. Sensors (Basel). 2025 May 25;25(11):3319. doi: 10.3390/s25113319. PMID: 40968853; PMCID: PMC12157161.

[13] K. A. Roecklein et al., "Melanopsin-driven pupil response in summer and winter in unipolar seasonal affective disorder," *Journal of Affective Disorders*, vol. 291, pp. 93–101, 2021.

[14] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, Dec. 1945.

[15] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[16] Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ. 1990 Jan 27;300(6719):230-5. doi: 10.1136/bmj.300.6719.230. PMID: 2106931; PMCID: PMC1662068.

Prototype Vibration Stimulator for Evaluating Somatosensory Nerve Fibers: Application in Charcot-Marie-Tooth Disease

Ricardo de Souza Toscano
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ricardo.toscano@peb.ufrj.br

Liene Duarte Silva
Clementino Fraga Filho University Hospital
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
liene.duarte@gmail.com

Fernando Augusto N. Castro Pinto
Mechanical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
fcpinto@mecanica.coppe.ufrj.br

Tiago Paulino dos Santos
Mechanical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
tiagopaulino@mecanica.coppe.ufrj.br

Carlos Julio Tierra-Criollo
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
carjulio@peb.ufrj.br

Abstract*—The assessment of the Vibration Perception Threshold (VPT) is an important quantitative method for analyzing the function of peripheral Aβ nerve fibers. Changes in this sensitivity threshold are indicators for a possible diagnosis of peripheral neuropathies, such as the hereditary Charcot-Marie-Tooth (CMT) disease. The objective of this study is to develop a vibratory stimulation prototype using a state-of-the-art technological approach that combines a Linear Resonant Actuator (LRA) vibration motor with the physical quantity of acceleration to quantify the VPT. The use of this physical quantity differs from the parameter adopted by more traditional commercial devices, which use, for example, the physical quantity of displacement. The device consists of a microcontroller connected to a haptic driver, which controls the vibration amplitude of the LRA motor, which has a resonance frequency of 235 Hz, in a gradual (step-by-step) and progressive manner. To characterize the prototype's tuning curve, measurements of the RMS acceleration (m/s²) produced at each vibration step were taken using a laser vibrometer. The results demonstrated that the prototype operates with an RMS acceleration ranging from 7.655 m/s² to 29.993 m/s². During system validation tests, a harmonic distortion of 3.32% was also observed, a value that falls within regulatory standards (ISO 13091-1). In the evaluations of the VPT in the distal segment of the lower limbs, there was a statistically significant difference between the groups (p = 0.0016), with higher VPT values for the group of patients with CMT, reflecting a possible impairment of peripheral somatosensory nerve fibers, specifically Aβ fibers, suggesting a potential of the prototype in clinical evaluation.

Keywords— Vibratory stimulation, peripheral somatosensory nerve fibers, vibration perception threshold (VPT), Charcot-Marie-Tooth.

I. INTRODUCTION

The somatosensory system, also known as the general somatic sensory system, is responsible for perceiving sensory information detected throughout the body, as distinct from the sensory signals associated with the special senses, such as vision, hearing, smell, taste, and balance [1].

The somatosensory system encompasses three main functions: proprioception, exteroception, and interoception [2, p. 475]. Proprioception refers to the perception of the body's own position and movement; exteroception is characterized by the perception of stimuli arising from the direct interaction of the external environment with the body; and interoception corresponds to the perception of the functioning of organ systems and the body's internal states.

The somatosensory system is composed of somatosensory neurons, and since they transmit information to the central nervous system (CNS), they are called afferent fibers or somatosensory peripheral nerve fibers [3].

Afferent fibers differ in terms of the presence or absence of a myelin sheath and are classified, respectively, as myelinated or unmyelinated [4]. These fibers also vary in diameter and are classified as thin fibers (type C – unmyelinated and type Aδ – myelinated) and thick fibers (type Aβ – myelinated and type Aα – myelinated, groups Ia and Ib). It should be noted that there are also efferent fibers, which connect the CNS to muscles and glands, designated as Aα [2, p. 770].

For action potentials to be generated in response to physical or physiological stimuli, the process of sensory transduction must occur, in which different forms of energy are converted into electrical signals in the nervous system. In the somatosensory system, mechanical stimuli are detected by specialized receptors called cutaneous mechanoreceptors, which are responsible for the perception of pressure, vibration, touch, stretch (tissue extension), and body position (proprioception). Among the main cutaneous mechanoreceptors are Merkel discs, Meissner corpuscles, Ruffini corpuscles, and Pacini corpuscles, each of which exhibits specific characteristics of adaptation and sensitivity when subjected to different mechanical stimuli [5]. Pacini corpuscles exhibit high sensitivity to high-frequency vibrations, in the approximate range of 40 to 400 Hz [6].

The perception of mechanical stimuli on the skin is mediated by myelinated Aβ nerve fibers. Diseases that more specifically affect myelinated fibers, such as Charcot-Marie-Tooth disease (CMT), cause significant impairment of functions related to the perception of vibratory sensation, especially in the early stages of the disease, when more severe motor symptoms are not yet present. Thus, the assessment of vibratory sensation serves as an important clinical diagnostic method in the functional evaluation of somatosensory peripheral nerve fibers [7].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

One parameter used in the assessment of vibration sensitivity is the Vibration Perception Threshold (VPT), defined as the lowest intensity of a vibratory stimulus that an individual can perceive. Devices for quantitatively assessing the VPT in clinical trials have shown greater reliability compared to tuning fork tests [8].

In this regard, this study presents the development and characterization of a vibratory stimulator designed for the quantitative assessment of vibratory sensitivity. The proposal adopts a more recent technological approach, which combines the use of a Linear Resonant Actuator (LRA) vibration motor with the physical quantity RMS acceleration (a_{RMS} in m/s²) to quantify VPT, thus differing from more traditional commercial systems, which use, for example, the physical quantity of displacement (µm) as a parameter for quantifying VPT [8], [9].

A preliminary study evaluating the VPT in patients with the hereditary disease Charcot-Marie-Tooth (CMT) is presented.

II. Materials and Methods

A. Vibration System Architecture

The architecture of the vibration stimulation system was designed around three key components: an Arduino Uno R3 microcontroller, a DRV2605L haptic driver, and a Linear Resonant Actuator (LRA), as shown in Fig. 1.

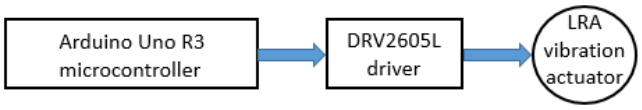


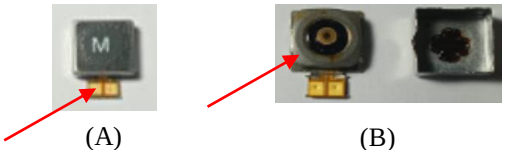
Fig. 1. Block diagram of the implemented system, consisting of an Arduino Uno R3 microcontroller, a DRV2605L haptic driver, and an LRA-type vibration actuator.

The Arduino Uno R3 microcontroller, based on the ATmega328P, controls the system by executing code written in C++. This code allows for 128 vibration amplitude steps (from 0 to 127). Parameters such as the duration of the vibration stimulus at each step, the interval between stimuli, the initial step, and the step increment can be configured. Communication with the haptic driver is carried out via the I²C interface.

The DRV2605L haptic driver, developed by Texas Instruments, is a low-voltage integrated circuit that provides a library of haptic effects, offering capabilities for controlling and driving vibratory actuators [10]. In this study, the DRV2605L is responsible for driving the Linear Resonant Actuator (LRA), enabling control of the vibration amplitude at each step, according to the parameters defined by the microcontroller.

Vibration actuators are electromechanical devices capable of converting electrical signals into mechanical vibrations; they are typically used in haptic interfaces found in various electronic devices, such as cell phones, video game controllers, and medical equipment designed for tactile stimulation [11], [12].

Among vibratory actuators, LRA-type devices stand out for their precision and efficiency in generating controlled stimuli. These actuators generate vibration when subjected to an alternating voltage through the linear displacement of an internal mass, driven by the magnetic field of a coil and restored by an elastic system [13]. Fig. 2 shows the LRA motor used in the prototype.



(A) (B)

Fig. 2. LRA vibration motor. In (A), its external configuration is shown, with the terminals for soldering the conductors (red arrow). In (B), the internal structure of the LRA motor is shown, highlighting (red arrow) the oscillating mass, which is responsible for the linear vibrational motion without any rotational component.

LRA-type motors exhibit maximum vibration amplitude when excited by a periodic signal at their resonance frequency (f_R), defined as the natural frequency at which the system exhibits the greatest dynamic response [14].

This frequency can be described by a mass–spring system model, as shown in Equation (1) [15].

$$f_R = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \quad (1)$$

Wherein:

- f_R is the resonance frequency of the LRA motor, in Hz;
- k is the spring constant, in N/m;
- m is the effective mass of the vibrating system, in kg.

Fig. 3 illustrates the (typical) frequency response of an LRA vibration motor, showing a peak in RMS acceleration amplitude (a_{RMS}), in m/s², at the resonance frequency (f_R) of the LRA motor used in the study (235 Hz).

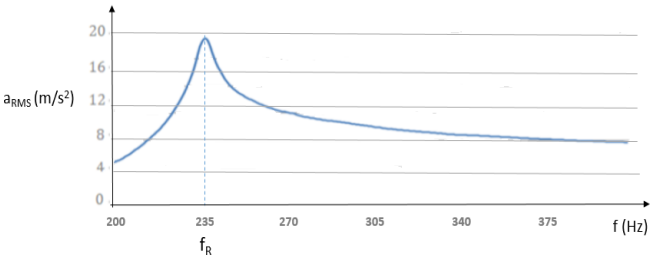


Fig. 3. RMS acceleration (a_{RMS}) response curve, in m/s², at the resonance frequency of an LRA vibration motor (adapted from [16]).

B. Step adjustment curve vs RMS acceleration

Measurements of the vibrations produced by the LRA motor at each step were taken using a laser vibrometer (PDV100 – Polytec) and a 01dB – NETDB multichannel analyzer. Fig. 4 shows the equipment used.



(A) (B)

Fig. 4. Equipment used to measure the calibration curve: (A) laser vibrometer (PDV100 – Polytec); (B) 01dB multichannel analyzer – NETDB.

The PDV100 is a high-precision device that enables the non-contact measurement of mechanical vibrations on surfaces, using laser technology to obtain vibration velocity values [17], [18].

The 01dB – NETDB multichannel analyzer is a data acquisition system used for acoustic and vibration applications, offering spectral analysis and signal processing capabilities via the dBFA Suite software [19].

On the PDV100, the following settings were used for the experiment.

- Peak scale range of 100 mm/s;
- Low-pass filter with a cutoff frequency of 1 kHz;
- High-pass filter with a cutoff frequency of 100 Hz.

The following parameters were configured in the NETDB software:

- Acquisition time of 10 seconds per step;
- 1-second post-trigger for stabilization purposes.

In the experiment, a silicone hand, supported and stabilized on a base, was used to simulate realistic contact conditions between the LRA motor and the index finger of the hand. The LRA motor was attached to the distal pad of the index finger, as shown in Fig. 5 (highlighted with a red arrow).



Fig. 5. Experimental setup for measuring the adjustment of the step curve vs RMS acceleration.

With this setup, a C++ language code was implemented to generate individual vibrations (per step), with intensities ranging from 0 to 127. Following data acquisition by PDV-100, post-processing was performed using the NETDB software, including the following actions:

- Application of a bandpass filter from 200 to 270 Hz, with a slope of 24 dB/eighth (4th order), centered on the resonance frequency of the LRA motor (235 Hz);
- Derivation of the velocity signal to obtain the acceleration (a) at each step;
- Calculation of the effective acceleration value (a_{RMS}) at each step, in m/s^2.

C. Experimental Protocol for Assessing the Vibration Perception Threshold (VPT)

Experimental evaluations were conducted on human participants to assess their perception of the vibrational stimulus generated by the developed system. The participants were divided into two distinct groups, as classified below:

- Group 1 – Control group, consisting of healthy individuals with no symptoms of neuropathy;
- Group 2 – Individuals diagnosed with CMT-related peripheral neuropathy.

The study was conducted in accordance with the ethical guidelines for research involving human subjects (Brazilian CNS Resolution 466/12). The protocol was approved by the Research Ethics Committee of the School of Dentistry – Federal University of Rio de Janeiro (UFRJ) under CAAE 95585626.0.0000.0268 on March 27, 2026, with written informed consent obtained from all participants.

Group 1 consisted of nine participants. The inclusion criteria were: being between 18 and 75 years of age, with no history of peripheral neuropathies. The exclusion criteria were: use of antidepressants or sedatives that impair attention and/or the locomotor and somatosensory systems; or having any type of neurodegenerative disease.

Group 2 consisted of three participants. The inclusion criteria were: being between 18 and 75 years of age and having a diagnosis of CMT; and being under medical care for this hereditary disease. The exclusion criteria were: individuals with CMT who had cognitive decline, dementia, cerebellar ataxia, or other peripheral neuropathies, such as those caused by diabetes.

The participants were seated in an air-conditioned room with a temperature of approximately 25°C (± 1°C) and dim lighting; they were instructed to remain relaxed throughout the assessment period.

The assessments were performed on the upper limbs (UL) and lower limbs (LL), ipsilaterally, on the left side. In the UL, the pulp of the distal phalanx of the index finger was used to attach the LRA motor. In the LL, the LRA motor was attached to the proximal phalanx of the hallux (big toe), on the dorsal side.

The procedure for evaluating the Vibration Perception Threshold (VPT) consisted of the following steps:

- Step 1: measure the temperature of the limb to be evaluated using a digital infrared thermometer (FLUKE 62 MAX+);
- Step 2: position of the LRA motor at the body points used in the study;
- Step 3: generation of vibratory stimuli (four events), with gradual increases in vibration amplitude (m/s^2), until the participant signals perception of the stimulus by pressing a stop button;
- Step 4: calculation of the arithmetic mean of the values obtained in the four events, which determines the participant's VPT value.

Fig. 6 shows the setup as it was assembled during the evaluations of one participant.

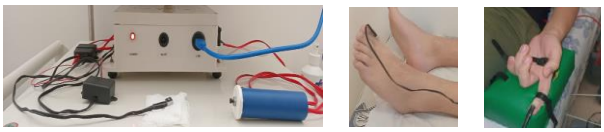


Fig. 6. Experimental setup assembled and evaluations.

III. RESULTS

A. *Prototype adjustment curve*

For each step, the a_{RMS} values were collected and tabulated in m/s². Table 1 shows the a_{RMS} values obtained, with an initial value of 7.655 m/s² (step 24) and a final value of 29.993 m/s² (step 127).

TABLE 1

a_{RMS} values at each step

Steps	a_{RMS}	33	10.316	67	19.267	101	26.065
0	0.000	34	10.316	68	19.261	102	26.195
1	0.000	35	10.360	69	19.229	103	26.160
2	0.000	36	10.706	70	19.629	104	26.109
3	0.000	37	11.592	71	20.108	105	26.897
4	0.000	38	11.542	72	20.102	106	26.977
5	0.000	39	11.495	73	20.098	107	27.049
6	0.000	40	11.542	74	20.192	108	26.921
7	0.000	41	12.746	75	21.032	109	27.575
8	0.000	42	12.810	76	21.194	110	27.739
9	0.000	43	12.745	77	21.127	111	27.738
10	0.000	44	12.758	78	21.125	112	27.665
11	0.000	45	13.904	79	21.304	113	27.898
12	0.000	46	13.980	80	21.964	114	28.468
13	0.000	47	13.917	81	21.932	115	28.632
14	0.000	48	13.932	82	21.887	116	28.478
15	0.000	49	14.644	83	22.497	117	28.439
16	0.000	50	15.053	84	22.795	118	29.311
17	0.000	51	15.092	85	22.786	119	29.025
18	0.000	52	15.062	86	22.750	120	28.860
19	0.000	53	15.007	87	22.799	121	29.054
20	0.000	54	16.189	88	23.629	122	29.456
21	0.000	55	16.080	89	23.660	123	29.550
22	0.000	56	16.181	90	23.807	124	29.387
23	0.000	57	16.069	91	23.727	125	29.445
24	7.655	58	17.206	92	23.909	126	29.484
25	7.660	59	17.267	93	24.625	127	29.993
26	7.659	60	17.125	94	24.554		
27	7.660	61	17.173	95	24.401		
28	8.782	62	18.079	96	25.117		
29	9.022	63	18.134	97	25.355		
30	9.056	64	18.225	98	25.388		
31	8.986	65	18.086	99	25.359		
32	10.289	66	18.299	100	25.438		

Based on the acceleration (a_{RMS}) values presented in Table 1, a curve showing the RMS acceleration (a_{RMS}) as a function of step increments was generated using GNU Octave software. A gradual and progressive increase in RMS acceleration is observed as the number of steps increases, as illustrated in Fig. 7.

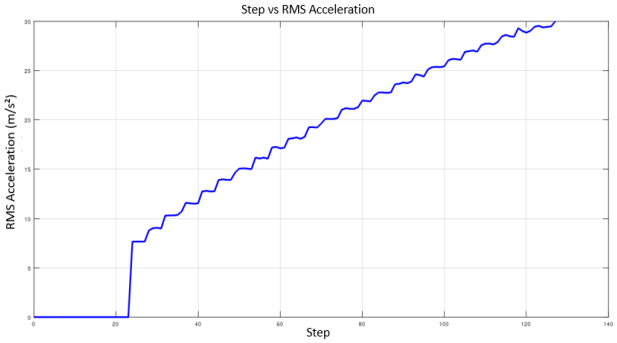
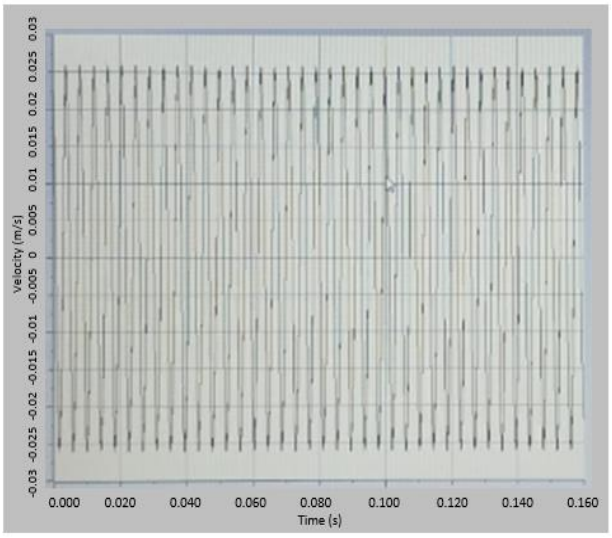


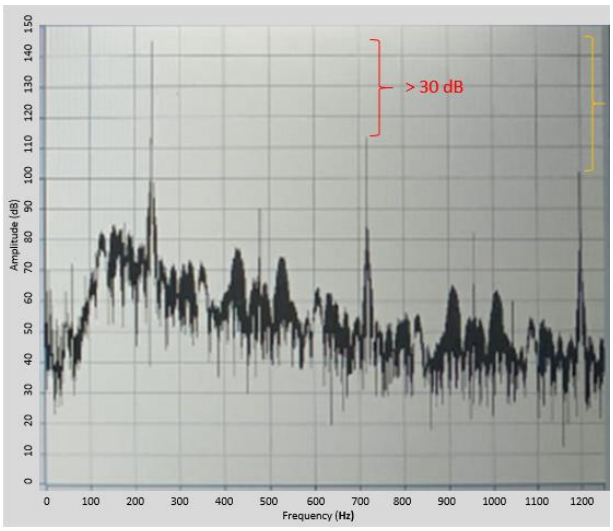
Fig. 7. RMS Acceleration (a_{RMS}) Curve as a Function of Step Increments for the LRA Motor.

As can be seen in Table 1 and Fig. 7, the vibration begins at step 24. This behavior is associated with an adjustment made to a specific parameter in the microcontroller code, which limits the amplitude of the LRA motor's excitation signal. This adjustment made it possible to minimize the presence of harmonic distortions, contributing to an improvement in the quality of the vibration signal, but causing the vibration to be generated only starting at step 24.

Fig. 8 (A) illustrates the vibration signal corresponding to step 111 in the time-domain, as recorded by the PDV-100. The signal shows distortion. Fig. 8 (B) shows the frequency-domain analysis of the signal, where the odd harmonics are more evident. There is a difference of more than 30 dB (highlighted in red) between the amplitudes of the fundamental (235 Hz) and the 3rd harmonic (at 705 Hz). This corresponds, considering 30 dB on the linear scale, to an amplitude of the fundamental component that is 31.62 times greater than that of the 3rd harmonic and, in terms of power, to a difference of approximately 1,000 times. It can also be observed in (B) that there is a difference of more than 40 dB (highlighted in orange) between the amplitudes of the fundamental (235 Hz) and the 5th harmonic (1,175 Hz). This latter frequency lies outside the response band of Pacini corpuscles.



(A)



(B)

Fig. 8. In (A), analysis in the time-domain, and in (B), in the frequency-domain.

The ISO 13091-1 standard (Mechanical vibration - Vibrotactile perception thresholds for the assessment of nerve dysfunction), establishes the purity criteria for the vibrational

stimulus wave. According to the technical literature, the guidelines of this standard require that a stimulation system have a Total Harmonic Distortion (THD) within the following parameters [20]:

- < 30% from 4 Hz to 50 Hz;
- < 10% from 63 Hz to 250 Hz.

The calculated THD value for the stimulus at the fundamental frequency of the LRA motor (235 Hz), considering the 3rd and 5th harmonics as the main distortion components (705 Hz and 1,175 Hz) was 3.32%. This THD value falls well within the limits specified in ISO 13091-1, demonstrating that the developed system complies with the required international limits.

B. Assessment of the Vibration Perception Threshold (VPT)

Table 2 shows the VPT values obtained in the assessment of Group 1—Control. It can be seen that all participants (9 in total) reported perceiving vibration at step 24, where the RMS acceleration (a_{RMS}) was 7.655 m/s², in both the lower and upper limbs.

TABLE 2

Values of Group 1 - Control

ID	Gender	Age (years)	Foot VPT (m/s^2)	Hand VPT (m/s^2)
1	M	24	7.655	7.655
2	F	21	7.655	7.655
3	M	24	7.655	7.655
4	M	29	7.655	7.655
5	M	22	7.655	7.655
6	M	29	7.655	7.655
7	M	22	7.655	7.655
8	F	34	7.655	7.655
9	F	25	7.655	7.655
Median	—	24	7.655	7.655
Minimum	—	21	7.655	7.655
Maximum	—	34	7.655	7.655

Table 3 presents the data collected during the vibration assessment of Group 2, participants with CMT-related peripheral neuropathy, comprising a total of three participants.

TABLE 3

Values of Group 2 – CMT holders

ID	Gender	Age (years)	Foot VPT (m/s^2)	Hand VPT (m/s^2)
10	M	53	27.739	7.655
11	F	46	29.993	29.993
12	M	18	10.005	7.655
Median	—	46	27.739	7.655
Minimum	—	18	10.005	7.655
Maximum	—	53	29.993	29.993

The VPT values for Group 2 – patients with CMT, were higher than those for Group 1 – the control group. Participant No. 10 presented a typical phenotype in the lower extremities, with pes cavus, hammertoes, and distal muscle atrophy (peroneal and tibialis anterior), along with difficulties with walking, balance, and postural stability, symptoms characteristic of CMT-related peripheral neuropathy [21]. In the upper extremities, there were no visible symptoms of CMT. Participant No. 11 exhibited the same symptoms in the lower extremities as those described for Participant No. 10, with even more visual evidence. In addition, impairment of the intrinsic hand muscles, characteristic of CMT, was observed in the upper extremities, consistent with the degree of disease progression [22]. Participant No. 12 did not exhibit symptoms characteristic of CMT in either the lower or upper extremities.

For the statistical comparison between the groups (Control vs. CMT), the non-parametric Wilcoxon-Mann-Whitney test was applied. In the evaluation of the vibration perception threshold in the lower limbs (foot), inferential statistical analysis indicated a significant difference between the Control group and the CMT group, presenting a large effect size ($W = 0$; $p = 0.0016$; Cliff's Delta $\delta = 1.00$; 95% CI = [0.816; 1.000]). In contrast, for the upper limbs (hand), no statistically significant difference was observed between the groups, with a small to moderate effect size ($W = 9$; $p = 0.124$; Cliff's Delta $\delta = 0.33$; 95% CI = [-0.391; 0.803]).

IV. Discussion

The study presented the implementation of a prototype vibratory stimulator designed to assess the sensitivity of peripheral Aβ nerve fibers by quantifying VPT. The device produces vibratory stimuli in a gradual and progressive manner, employing a state-of-the-art technological approach that combines the use of an LRA motor with the physical quantity RMS acceleration (a_{RMS}), expressed in m/s², to quantify VPT. This methodology differs from more traditional commercial systems, which typically operate using displacement, in µm, as the unit of measurement for vibration. From the perspective of the prototype's functional reliability, the Total Harmonic Distortion (THD) was only 3.32%, with the 3rd and 5th harmonics being the main distortion components, indicating a strong predominance of the fundamental component at 235 Hz. This level of spectral purity meets the harmonic distortion criterion established by the ISO 13091-1 standard, although full compliance with the standard depends on other technical parameters, such as the incorporation of contact force control systems. Although experimental validation with the non-contact laser vibrometer (Polytec PDV100) ensures adjustment accuracy and the reliability of the generated stimulus, the formal calculation of the expanded measurement uncertainty of the entire instrumental system constitutes a step to be addressed in future work. In the VPT assessments, the results showed a statistically significant difference in the distal segment of the lower limbs ($p = 0.0016$), which was not observed in the upper limbs ($p = 0.124$), suggesting a possible trend associated with the length-dependent pattern of CMT neuropathy [23]. However, a central limitation of the current device is the floor effect observed in the Control Group, given that all healthy participants detected the vibration at the minimum level of the prototype (7.655 m/s²). This indicates that the true threshold for this population may lie below the current dynamic range of the hardware, and this value should be interpreted as an operational limit of the system, rather

than an absolute normative threshold of sensitivity. Furthermore, the detection of altered thresholds in the asymptomatic participant (No. 12) represents a preliminary exploratory observation. Although it suggests the potential of the device for early screening of somatosensory changes, the confirmation of a diagnostic capability at a subclinical stage requires cross-validation with reference examinations (such as electroneuromyography) and longitudinal follow-up. The chosen frequency of 235 Hz, which preferentially stimulates Pacinian corpuscles, limits the evaluation of other mechanoreceptors that may be compromised by different types of peripheral neuropathies. Future work will address multifrequency stimulation. Additionally, subsequent developments will focus on operation at lower acceleration levels to minimize the floor effect, contact force control, the incorporation of new algorithms into the firmware to enhance measurement rigor, measurement uncertainty evaluation, performance comparison with consolidated commercial devices, and the execution of studies with larger samples.

V. Conclusion

The implemented prototype is a promising tool for the quantitative assessment of vibratory sensitivity, especially in conditions that cause peripheral neuropathies. Although the study presents a preliminary investigation with a reduced number of participants in the CMT peripheral neuropathy group (n=3), the results obtained indicated a possible differentiation when compared to the healthy group, suggesting a potential of the system for clinical application. Additional tests, considering a significantly larger sample, must be performed for the purpose of full validation of the produced equipment.

THANKS

The authors thank the study participants for their cooperation and willingness to take part in the assessments. This work was supported by the following Brazilian research funding agencies: CAPES/PRONEX, through the Biomedical Engineering Program at the Alberto Luiz Coimbra Institute for Graduate Studies and Engineering Research (COPPE); and the National Council for Scientific and Technological Development (CNPq).

References

[1] A. C. Guyton and J. E. Hall, Textbook of Medical Physiology, 12th ed. Philadelphia, PA, USA: Saunders Elsevier, 2011.

[2] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum and A. J. Hudspeth, Principles of Neural Science, 5th ed. New York, NY, USA: McGraw-Hill, 2014.

[3] M. F. Bear, B. W. Connors, and M. A. Paradiso, Neuroscience: Exploring the Brain, 4th ed. Philadelphia, PA, USA: Wolters Kluwer, 2016.

[4] M. F. Bear, B. W. Connors and M. A. Paradiso, Neuroscience: Exploring the Brain, 3rd ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2007.

[5] S. E. Mulroney and A. K. Myers, Netter's Essential Physiology, Philadelphia, PA, USA: Saunders Elsevier, 2009.

[6] R. S. Johansson and J. R. Flanagan, “Coding and use of tactile signals from the fingertips in object manipulation tasks,” Nature Reviews Neuroscience, vol. 10, pp. 345–359, May 2009.

[7] H. M. Canelas, "Sensibilidade vibratória: valor semiótico em algumas afecções do sistema nervoso" [Vibratory sensitivity: Semiological value in some affections of the nervous system], Arquivos de Neuro-Psiquiatria, vol. 16, no. 4, pp. 275–352, 1958.

[8] M. Gao, X. Yun and T. Zhang, “VSA-3000: A quantitative vibration sensation testing device for patients with central nervous system injury,” Frontiers in Neurology, vol. 11, art. 936, 2020, doi: 10.3389/fneur.2020.00936.

[9] SBMEDIC Electronics, “Vibrameter® Type IV: For an early diagnosis of nerve damage in neurology, occupational medicine and diabetic care,” Solna, Sweden, Technical Brochure, Sep. 12, 2008. [Online]. Available: https://www.somedicprod.se.

[10] Texas Instruments, “DRV2605L 2- to 5.2-V Haptic Driver for LRA and ERM with Effect Library and Smart-Loop Architecture,” SLOS854D, May 2014, revised Mar. 2018. [Online]. Available: https://www.ti.com/lit/ds/symlink/drv2605l.pdf

[11] H. Culbertson, S. B. Schorr, and A. M. Okamura, “Haptics: The present and future of artificial touch sensations,” Annual Review of Control, Robotics, and Autonomous Systems, vol. 1, pp. 12.1–12.25, 2018. doi: 10.1146/annurev-control-060117-105043.

[12] S. Choi and K. J. Kuchenbecker, “Vibrotactile display: Perception, technology, and applications,” Proceedings of the IEEE, vol. 101, no. 9, pp. 2093–2104, Sep. 2013.

[13] Texas Instruments, “How does a linear resonant actuator work?,” SSZTAP9, Oct. 2016. [Online]. Available: https://www.ti.com/lit/ta/ssztap9/ssztap9.pdf.

[14] Texas Instruments, “LRA actuators: How to move them?,” Application Report SLOA209, 2014. [Online]. Available: https://www.ti.com/lit/an/sloa209/sloa209.pdf

[15] S. Liu, Y. Kao, and G. Feng, “Linear resonator actuator-constructed wearable haptic system with the application of converting remote grinding force to vibratory sensation,” Actuators, vol. 12, no. 9, 2023. doi: 10.3390/act12090359.

[16] Vybronics, “VLV101040A Linear Vibration Motor – Product Specification,” Rev. A/3, Aug. 2022. [Online]. Available: www.vybronics.com

[17] POLYTEC GmbH, “Portable Digital Vibrometer PDV-100: User Manual,” Waldbronn, Germany, 2011.

[18] POLYTEC GmbH, “Portable Digital Vibrometer PDV-100: Technical Data,” Waldbronn, Germany, 2007.

[19] 01dB-Metravib, “NetdB DAQ12: Product Information Sheet,” Limonest, France, 2006.

[20] Human Factors Research Unit, "HVLab Vibrotactile Perception Meter brochure," Institute of Sound and Vibration Research, University of Southampton, Southampton, UK, Tech. Rep. Issue 2.0, Aug. 2014.

[21] M. M. Dudziec, L. E. Lee, C. Massey, D. T. Meng, M. Skorupinska, M. Laurá, M. M. Reilly, and G. M. Ramdharry, "Home-based multi-sensory and proximal strengthening program to improve balance in Charcot–Marie–Tooth disease Type 1A: A proof of concept study," *Muscle Nerve*, vol. 69, no. 3, pp. 354–361, Mar. 2024, doi: 10.1002/mus.28032.

[22] F. Albuquerque, D. Cunha, A. C. Rodrigues *et al.*, "The Impact of a Late Diagnosis: A Case of Charcot-Marie-Tooth Type 1," *Cureus*, vol. 15, no. 1, p. e33727, Jan. 2023, doi: 10.7759/cureus.33727.

[23] SZIGETI, K.; LUPSKI, J. R. Charcot–Marie–Tooth disease. European Journal of Human Genetics (2009) 17, 703 – 710. Doi: 10.1038/ ejhg.2009.31, 2009.

Radiomic Evaluation of Additive Manufacturing Materials for Optimizing Anthropomorphic Mammography Phantoms

Gabriel Henrique Silva Cardoso
Faculty of Eletrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0009-0008-2130-4886

Paloma Anselmo Rosa
Faculty of Eletrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0009-0002-0181-1481

Enzo Fernandes Rodrigues
Faculty of Eletrical Engineering
Federal University of Uberlândia?
Uberlândia, Brazil
0009-0002-2864-3107

Pedro Cunha Carneiro
Faculty of Eletrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0000-0002-0120-5273

Ana Claudia Patrocinio
Faculty of Eletrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
0000-0001-9376-7689

***Abstract*—Early detection is essential for reducing breast cancer mortality, driving ongoing research to improve lesion diagnosis in mammograms. However, the development of CAD and CADx systems is hindered by a lack of large, standardized, and clinically annotated datasets. Anthropomorphic breast phantoms provide a standardized alternative by simulating realistic breast structures. This study evaluates three 3D-printed materials—polylactic acid (PLA), thermoplastic polyurethane (TPU), and photopolymer resin—for their visual and textural equivalence to actual breast lesions. Test structures with varying geometries and infill densities were embedded in beeswax plates, which simulate adipose tissue. Mammograms from 11 configurations were analyzed using radiomic texture features to assess the effects of material, shape, and size. Results showed that TPU exhibited the most pronounced deviations from the PLA reference, particularly in Contrast, Sum Variance, IMC1, IMC2, and MCC. Additionally, geometric shape and size marked textural discrepancies in the radiomic responses, specifically affecting Sum Variance, Inverse Difference Moment, Difference Variance, and Entropy. Ultimately, TPU's variable radiographic response makes it unsuitable for simulating glandular tissues and lesions. Photopolymer resin showed high similarity to PLA, but PLA remains the most suitable material for simulating lesions, optimizing the development of anthropomorphic phantoms.**

***Index Terms*—mammography, breast cancer, plates, Additive manufacturing,**

I. Introduction

Breast cancer is a leading cause of death among women and one of the most prevalent malignant neoplasms worldwide [1], [2]. Early detection is crucial, as treatment in the initial stages halts disease progression, prevents metastasis, and significantly improves patient survival [3], [4]. Accordingly, the World Health Organization (WHO) recognizes mammography as the gold standard for population-based screening in asymptomatic women [5].

To enhance image interpretation efficiency and accuracy, ongoing research focuses on developing Computer-Aided Detection (CAD) and Diagnosis (CADx) systems [6]. However, developing and validating these algorithms requires large datasets of mammographic images, which poses a significant challenge. Private databases are often restricted to their owners, while public repositories frequently suffer from bureaucratic access, unstandardized file formats, and inconsistent image acquisition protocols. Crucially, these public datasets often lack the physical and structural diversity necessary to train robust, modern models [7]–[9]. Given this scenario, image sets based on radiographic exposure of breast phantoms, prostheses specially designed to simulate real breasts in various aspects, have proven to be a promising alternative to replace or complement real-image databases [10]. These phantoms can be physical or virtual [11], and their use presents notable advantages: precise knowledge of the structural attributes of the exposed phantom, the ability to control and vary the location and morphology of lesions, repeatability in image acquisition, generation of large volumes of data without exposing patients to additional radiation doses, and elimination of the need for medical reports or biopsies for result validation [12]–[14]. These phantoms are used not only in research but also for calibration, quality assurance, and accuracy assessment of images acquired through mammography [15]. However, the use of commercial phantoms still face barriers regarding accessibility, cost, and fidelity to actual breast tissue, particularly for lesions mimicking glandular tissue [16]. This fidelity issue arises because the radiological response of a phantom depends on the attenuation and texture of the materials used, which in commercial phantoms are fixed and often not representative of breast tissue. Additive manufacturing (3D printing) directly addresses this limitation: by selecting the printing material

and controlling the infill pattern and percentage, the resulting attenuation and texture can be adjusted to approximate those of breast tissue. It also offers lower cost and greater ease of customization compared to commercial phantoms, making it an accessible route for producing phantoms that more closely reproduce breast tissue. As a result, diverse tumor models varying in size, shape, density, and morphology can be created. Recent studies have begun to explore this avenue; for instance, Pinheiro [17] investigated materials and printing parameters specifically for radiological applications. In this context, optimizing anthropomorphic phantoms through careful material selection and quantitative radiological analysis is essential for advancing breast imaging research. This study aims to evaluate 3D-printed materials for the fabrication of new plates to be integrated into a phantom currently under development by the research group [18]. To assess this, the materials were evaluated through visual inspection and quantitative radiomic analysis of mammograms acquired from printed test objects.

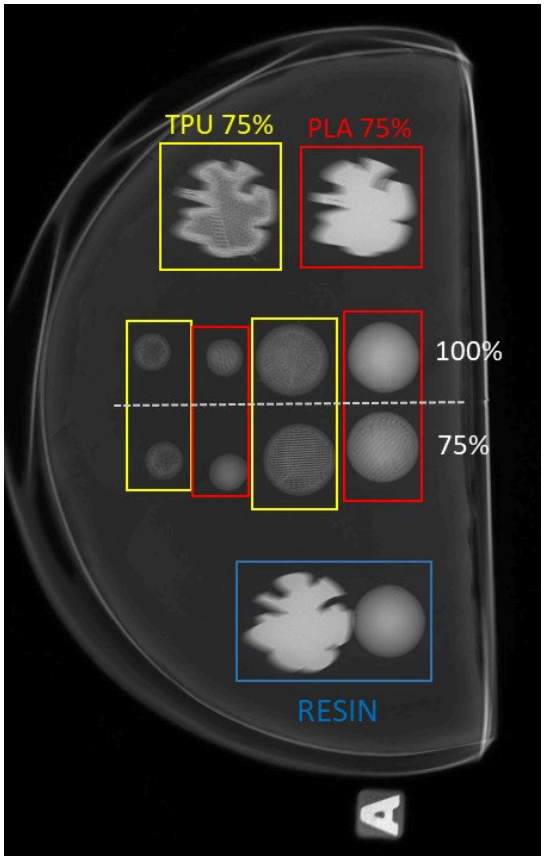


Fig. 1. Different structures used to simulate nodules with varying materials, geometries, and infill percentages.

II. METHODS

It is important to emphasize this study was conducted to evaluate the radiographic behavior of materials manufactured using additive manufacturing for application in anthropomorphic breast phantom plates. Test objects of various shapes and sizes were 3D-printed using three different materials and varying infill configurations. Mammograms of these objects were then acquired and processed to extract radiomic features, enabling quantitative texture analysis and comparison of the materials' radiographic behavior.

A. Design and Fabrication of the Test Objects

The structures simulating different nodules were manufactured using 3D printing with polylactic acid (PLA), thermoplastic polyurethane (TPU), and photopolymer resin. Components with different geometries, including circular and irregular shapes in both large and small sizes, as well as internal infill percentages of 75% and 100%, were produced to evaluate the influence of these variables on the radiographic behavior of the materials. These structures can be seen in Fig.1

The test objects were positioned on plates designed to simulate adipose and fibroglandular tissues. These background plates comprise: (1) an adipose-equivalent base, for which both beeswax and paraffin gel were evaluated as candidate materials [ref]; and (2) a higher-density mixture of beeswax, fiberglass, and polyvinyl chloride (PVC) film, replicating fibroglandular tissue [ref]. Originally developed for the BeeMamma phantom [ref], these plates provide varying X-ray attenuation properties to reproduce the radiographic characteristics of glandular breast tissue. A total of eleven mammographic images were analyzed. Eleven images corresponded to the experimental configurations (0A–10A), comprising different combinations of printing materials, filling formulations, and lesion-simulating structures. One additional image was acquired using the AEC (Automatic Exposure Control),that being configuration 7E, while another corresponded to configuration 10E, consisting of a phantom without lateral supports, containing a resin block and additional lesion-simulating structures. The experimental configurations are summarized in Table.1.

TABLE I
EXPERIMENTAL PHANTOM CONFIGURATIONS USED FOR MAMMOGRAPHIC IMAGE ACQUISITION

Config.	Experimental Configuration
0A	PLA, TPU and resin structures (75% and 100% infill) embedded in beeswax.
1A	Large PLA, TPU and resin structures embedded in beeswax.
2A	Small PLA, TPU and resin structures embedded in beeswax.
3A	Small PLA, TPU and resin structures embedded in beeswax and dental wax.
4A	Comparison of beeswax, aged beeswax and beeswax with paraffin.
5A	Beeswax and dental wax reinforced with fiberglass and 6 g of PVC.
6A	Beeswax and dental wax reinforced with fiberglass and 12 g of PVC.
7A	Beeswax and dental wax reinforced with fiberglass and 18 g of PVC.
7E	Same configuration as 7A acquired using the AOP DOSE exposure mode.
8A	Beeswax with fiberglass and 6 g PVC, beeswax with paraffin lesions, and dental wax with paraffin lesions.
10E	Modified phantom without lateral supports, containing a resin block, simulated microcalcifications and a gold enamel inclusion.

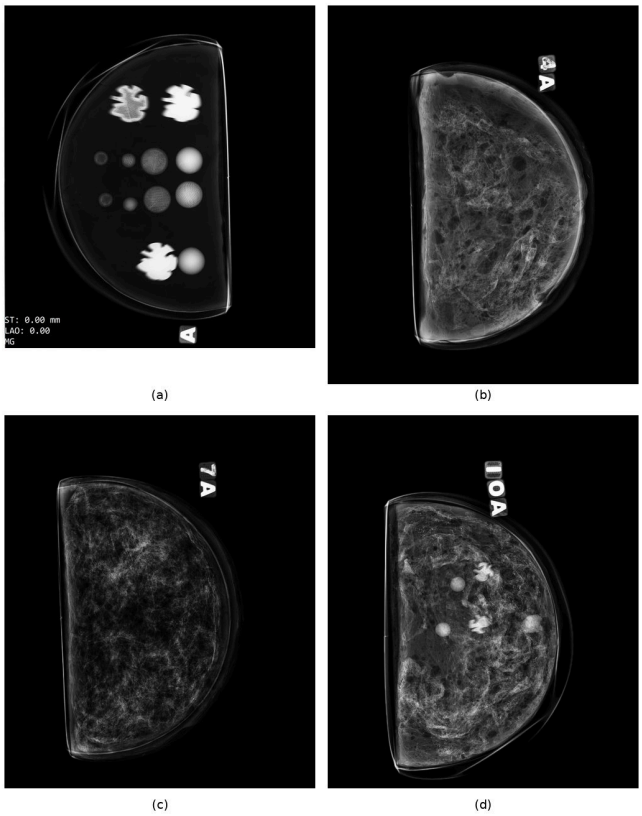


Fig. 2. Representative mammographic images acquired from different anthropomorphic phantom configurations: (a) configuration 0A, (b) configuration 4A, (c) configuration 7A, and (d) configuration 10A.

B. Image acquisition

A total of 11 images, representing various combinations of materials and infill configurations, were acquired using a full-field digital mammography system, model Senographe Crystal, in Automatic Exposure Control (AEC) mode. The specific configuration for each image is detailed in Table 1. Figure 1 presents four representative mammograms (configurations 0A, 4A, 7A, and 10A), selected to illustrate key stages of testing and optimization for the BeeMamma phantom.

C. Radiomic Feature Extraction

Quantitative characterization was performed by extracting radiomic texture features from the mammographic images acquired for the different phantom configurations. The analysis considered the distribution of gray-level values among pixels within regions of interest (ROIs) defined for each object and each material. This distribution allows the identification of texture patterns associated with the radiographic characteristics of the different materials evaluated. The comparison was performed using three complementary approaches: (i) visual assessment of the mammographic images; (ii) analysis of normalized Haralick feature curves (including skewness and kurtosis) with PLA as reference; and (iii) comparison of the Gaussian distribution curves of each feature. The visual assessment consisted of inspecting the mammographic images for structure visibility, edge definition, and texture realism, comparing each material against PLA as the reference. For texture characterization, Haralick descriptors were used [19], obtained from the Gray-Level Co-occurrence Matrix (GLCM), which describes the spatial relationship between the gray levels of image pixels, allowing the extraction of several statistical measures. A total of 14 Haralick features were extracted from each ROI, providing a comprehensive characterization of the texture properties of the mammographic images. These features, derived from the GLCM, describe different aspects of the spatial organization of grayscale values and are briefly described as follows [11]:

- **Angular Second Moment (ASM):** Measures texture homogeneity; higher values indicate more uniform textures.
- **Contrast:** Estimates local intensity variation by quantifying differences between gray levels of adjacent pixels; lower values correspond to smoother regions.
- **Correlation:** Assesses the linear dependency between gray levels of neighboring pixels, revealing periodic patterns or smooth areas.
- **Variance:** Quantifies the dispersion of gray-level intensities, reflecting image contrast.
- **Inverse Difference Moment (IDM):** Evaluates local homogeneity, being sensitive to subtle textural changes.
- **Sum Average:** Represents the mean of the gray-level distribution.
- **Sum Variance:** Indicates the variability of the gray-level distribution.
- **Sum Entropy:** Measures the information content associated with the sum of gray-level pairs.
- **Entropy:** Quantifies the randomness or disorder in the

gray-level distribution, with higher values indicating more complex textures.

- **Difference Variance:** Measures the variability of gray-level differences, centered at zero.
- **Difference Entropy:** Quantifies the disorder or randomness among pixel differences.
- **Information Measure of Correlation 1 (IMC1):** Quantifies correlation by analyzing the entropy of individual elements and their paired combinations.
- **Information Measure of Correlation 2 (IMC2):** Assesses correlation based on the similarity between entropies.
- **Maximal Correlation Coefficient (MCC):** Indicates the heterogeneity of the gray-level distribution.

In addition to the GLCM-based features, two first-order statistical measures were computed from the gray-level intensity distribution within each ROI [20]:

- **Skewness:** Measures the asymmetry of the gray-level intensity distribution. Positive skewness indicates a tail toward higher intensities, while negative skewness indicates a tail toward lower intensities.
- **Kurtosis:** Quantifies the "tailedness" or peakedness of the intensity distribution. Higher kurtosis values indicate a distribution with heavier tails and a sharper peak, suggesting the presence of outliers or extreme intensity values.

These descriptors collectively evaluate textural properties such as homogeneity, contrast, correlation, entropy, heterogeneity, and intensity distribution shape, allowing a quantitative comparison of the radiographic behavior of the materials under investigation.

The feature extraction was performed using a MATLAB script. The GLCM was computed considering a pixel distance of one and four orientations (0°, 45°, 90° and 135°). For each feature, the final value was obtained by calculating the arithmetic mean of the values computed for the four analyzed orientations.

D. Statistical analysis

The extracted radiomic features enabled the analysis of the texture characteristics of the different phantom configurations, as well as the identification of the materials with radiographic behavior most similar to glandular breast tissue. Initially, a descriptive analysis of the features obtained for each experimental configuration was performed.

Subsequently, Gaussian curves were used to evaluate the distribution of radiomic features among the different configurations. For this purpose, normal distributions were constructed based on the normalized mean and standard deviation values of each feature, considering both the material type and the lesion shape represented. The overlap between the curves was used as an indicator of similarity among the evaluated materials and the nodule geometries employed in the manufacture of the test structures, allowing the identification of the features that exhibited the most stable behavior.

Finally, a comparative analysis was performed to evaluate the radiomic response of the different materials across all experimental configurations. For this purpose, the value of each Haralick feature extracted from the TPU and photopolymer resin structures was divided by the corresponding feature value extracted from the PLA structures, which was adopted as the reference material.

PLA was chosen as the baseline material for the comparative analysis for three reasons. First, within the context of the phantom currently under development by the research group [18], PLA was previously employed as the primary material for fabricating tumor-simulating structures. further supported by a previous study [17], which demonstrated that PLA is suitable for simulating both circumscribed and spiculated nodules in mammographic phantoms, validating its use through comparison of mass attenuation coefficients with invasive ductal carcinoma across the mammographic energy range (10–30 keV). The phantom developed in that study, named Cleopatra, employed PLA to simulate nodules and was validated using 14 Haralick texture descriptors, showing good detectability of test structures. Additionally, radiographic results obtained from PLA demonstrated excellent visibility and realistic texture in mammographic images, further supporting its suitability for simulating biological lesions. Second, PLA is widely available, cost-effective, and well-characterized in the additive manufacturing literature, with studies confirming that its X-ray attenuation properties are compatible with breast tissue equivalents under clinical mammography [**?**]. Third, by adopting PLA as the reference, the influence of absolute intensity variations is eliminated, and the relative differences in texture properties between PLA and the other materials (TPU and photopolymer resin) can be directly assessed. This normalization approach enables a fair and consistent comparison of the radiographic behavior of the three materials, highlighting which features exhibit the greatest similarity or divergence relative to the established PLA baseline. Mathematically, for each feature f and material $m \in \{\text{TPU}, \text{Resin}\}$, the normalized value R_m was calculated as:

$$R_m = \frac{V_m}{V_{\text{PLA}}} \tag{1}$$

where V_m is the raw feature value for material m and V_{PLA} is the corresponding value for PLA.

The resulting normalized values were plotted using line charts, with each material represented by a separate line and each point corresponding to a specific Haralick feature. This visualization allowed the identification of which characteristics exhibited the greatest similarity or divergence between materials relative to the PLA baseline.

III. Results

The radiomic analysis revealed distinct texture patterns among the three 3D-printed materials and also indicated alterations in radiomic features due to the shape used in the lesion representation. As a first approach, the normalized line chart using PLA as the reference was analyzed. This chart

provides a visual comparison of how each Haralick feature behaves in TPU and photopolymer resin relative to the PLA baseline.

The Fig.3 presents the normalized Haralick features for TPU and photopolymer resin relative to PLA. As can be observed, TPU exhibits the most pronounced deviations from the PLA baseline, particularly for the Contrast, MeanSum, VarSumQ, VarSum and IMC1. The photopolymer resin presents a lower variation in values compared to TPU, while still exhibiting comparable differences in certain features such as Uniformity (Angular Second Moment), Inverse Difference Moment (IDM) and MCC. The observed differences in skewness and kurtosis across the three materials underscore their sensitivity to material composition and structural characteristics, as they are heavily influenced by the shape of the nodules in each test configuration.

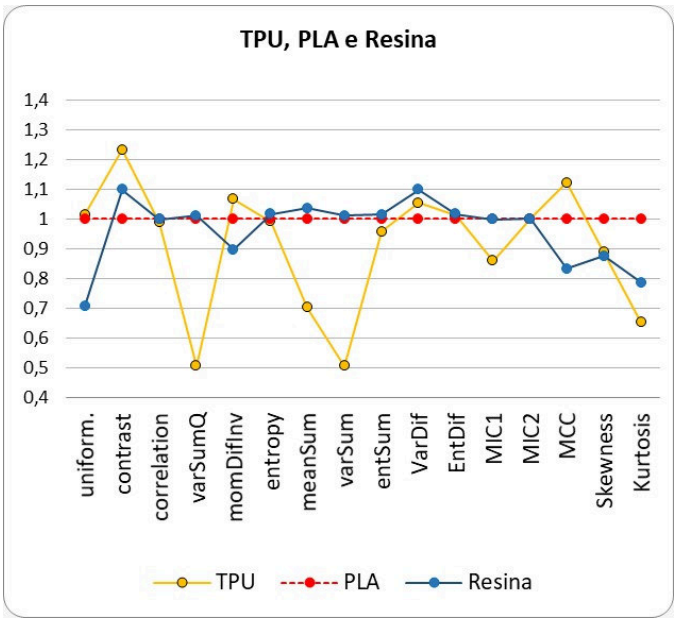


Fig. 3. Normalized Haralick features for TPU and photopolymer resin relative to PLA (reference).

The Gaussian curve analysis further supported these findings. Among the evaluated materials, the Information Measures of Correlation (IMC1 and IMC2) exhibited the most pronounced differences between the Gaussian curves, particularly when comparing TPU to both PLA and the photopolymer resin, as can be observed in Figures 4 and 5, shown below.

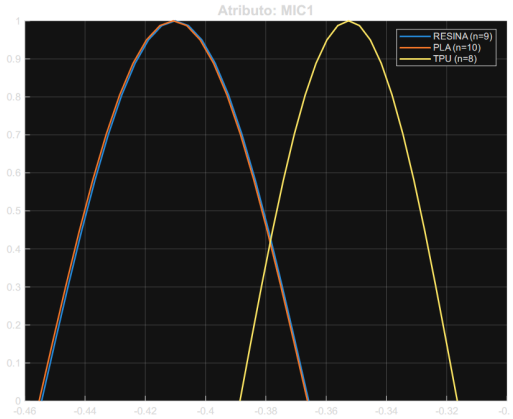


Fig. 4. Gaussian curve overlap for IMC1 across TPU (yellow), PLA (red), and photopolymer resin (blue).

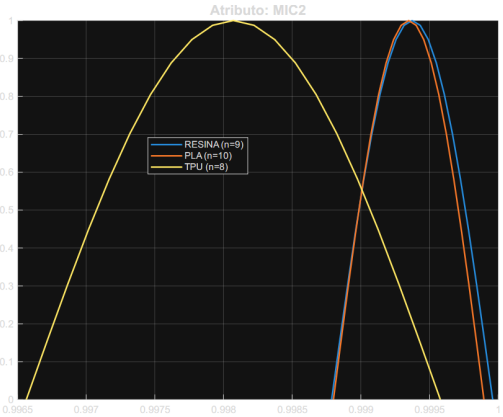


Fig. 5. Gaussian curve overlap for IMC2 across TPU (yellow), PLA (red), and photopolymer resin (blue).

The radiomic analysis revealed that TPU exhibits the most pronounced deviations from the PLA baseline, particularly in features related to texture heterogeneity, such as Contrast, Sum Variance, IMC1, IMC2, and MCC. These differences indicate that the radiographic texture response of TPU differs considerably from that of PLA, which has been previously supported as a suitable material for simulating tumorous structures in the phantom under development [18]. The observed discrepancies can be attributed to the intrinsic physical properties of TPU. Unlike PLA, which is rigid and provides consistent internal structures during 3D printing, TPU is an elastomeric material with high flexibility. This characteristic leads to less precise control over infill geometry and surface finish during additive manufacturing, resulting in greater textural heterogeneity. The Gaussian curves, analyzed according to the shapes of the printed lesions, further demonstrate that both the size and the regularity of the lesion-simulating structures have a considerable influence on the radiomic response of the acquired images. The attributes that most clearly demonstrate the influence of size and shape regularity are presented in the Gaussian curves of Figs. 6–9. These include the contrast difference between regular and irregular shapes, as well as

the Inverse Difference Moment, with skewness and kurtosis highlighting the differentiation of objects by size.

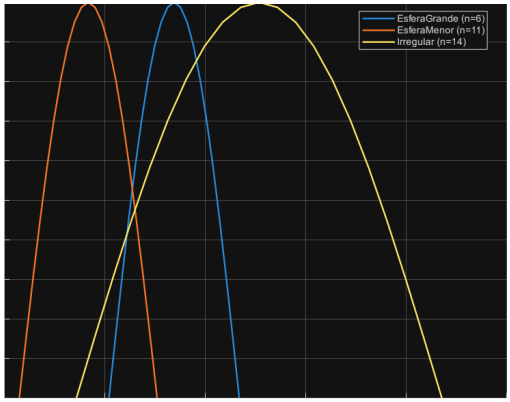


Fig. 6. Gaussian curve overlap for Sum Variance across big spheres (blue), small spheres (red), and irregular lesions (yellow).

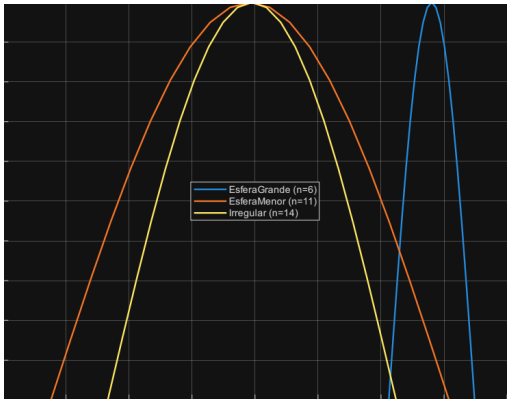


Fig. 7. Gaussian curve overlap for Inverse Difference Moment across big spheres (blue), small spheres (red), and irregular lesions (yellow).

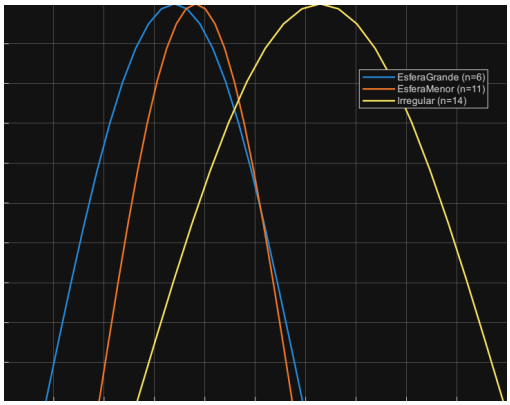


Fig. 8. Gaussian curve overlap for Difference Variance across big spheres (blue), small spheres (red), and irregular lesions (yellow).

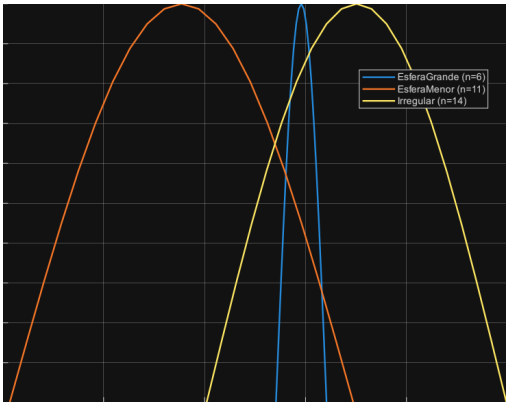


Fig. 9. Gaussian curve overlap for Entropy across big spheres (blue), small spheres (red), and irregular lesions (yellow).

IV. Conclusion

This study evaluated PLA, TPU, and photopolymer resin for application in anthropomorphic breast phantoms via radiomic analysis of mammographic images. The comparison was performed through three approaches: visual assessment; normalized Haralick feature curves, including skewness and kurtosis, referenced to PLA; and differences in the distribution of the Gaussian curves for each feature. TPU exhibited the most pronounced deviations from PLA in all three analyses, particularly in texture features such as Contrast, IMC1, IMC2, and MCC, indicating a distinct and variable radiographic response. Gaussian curve analysis of the different shapes and infill patterns revealed that these parameters influence the radiomic features, with Sum Variance, Inverse Difference Moment, Difference Variance, and Entropy displaying the most divergent distributions. Conversely, photopolymer resin demonstrated greater radiomic similarity to PLA, suggesting it is a viable alternative. Overall, PLA remains the preferred material for simulating tumor structures, whereas TPU showed both textural discrepancies and poor visual quality. Notably, the TPU exemplar differed markedly from the PLA exemplar even when printed from the same geometry. Future studies should explore additional materials and validate them against clinical images to ensure their efficacy in mammography research.

References

[1] World Health Organization (WHO), “Global Breast Cancer Initiative (GBCI),” Geneva: WHO, 2021. [Online]. Available: https://www.who.int/initiatives/global-breast-cancer-initiative

[2] Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA), “Estimativa 2026: incidência de câncer no Brasil,” Rio de Janeiro: INCA, 2026. [Online]. Available: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa/introducao

[3] L. M. Rosa and V. Radünz, “Taxa de sobrevida na mulher com câncer de mama: estudo de revisão,” Texto & Contexto Enfermagem, vol. 21, no. 4, 2012.

[4] O. Amato, V. Guarneri, and F. Girardi, “Epidemiology trends and progress in breast cancer survival: earlier diagnosis, new therapeutics,” Current Opinion in Oncology, vol. 35, no. 6, pp. 612–619, 2023.

[5] World Health Organization (WHO), “WHO Position Paper on Mammography Screening,” Geneva: WHO, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK269545/

[6] J. Katzen and K. Dodelzon, “A review of computer aided detection in mammography,” Clinical Imaging, vol. 52, pp. 305–309, 2018, doi: 10.1016/j.clinimag.2018.08.014.

[7] The Cancer Imaging Archive, “CBIS-DDSM: Curated Breast Imaging Subset of DDSM,” [Online]. Available: https://wiki.cancerimagingarchive.net/

[8] “Systematic review on dataset diversity for medical imaging applications,” [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0899707124002997

[9] “RSNA Breast Cancer Dataset,” Radiology: Artificial Intelligence. [Online]. Available: https://pubs.rsna.org/doi/10.1148/ryai.250375

[10] H. Schiabel *et al.*, “Analytical software for visual and quantitative comparison of real mammography and phantom breast cancer imaging,” Academia Oncology, 2025. [Online]. Available: https://doi.org/10.20935/AcadOnco7529

[11] E. Sizikova *et al.*, “Knowledge-based in silico models and dataset for the comparative evaluation of mammography AI,” Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS), 2023.

[12] Y. A. Vasilev *et al.*, “Anthropomorphic breast phantoms for radiology imaging: A review,” Digital Diagnostics, vol. 4, no. 4, pp. 569–592, 2023, doi: 10.17816/DD623341.

[13] A.-K. Carton *et al.*, “Development of a physical 3D anthropomorphic breast phantom,” Medical Physics, vol. 38, no. 2, pp. 891–896, 2011, doi: 10.1118/1.3533896.

[14] M. A. Z. Sousa *et al.*, “A comparison of methods to evaluate gray scale response of tomosynthesis systems using a software breast phantom,” Proceedings of SPIE Medical Imaging, vol. 10136, 2017, doi: 10.1117/12.2255554.

[15] J. Bermúdez *et al.*, “3D phantom for image quality assessment of mammography systems,” Physics in Medicine & Biology, vol. 68, no. 20, p. 205019, 2023.

[16] R. M. V. Rocha, “Desenvolvimento de objetos simuladores para imagens de pulmão e mama utilizando técnicas de impressão 3D,” M.Sc. dissertation, Instituto de Física, Universidade de São Paulo, São Paulo, Brazil, 2024.

[17] M. S. D. Pinheiro, “Análise de materiais e desenvolvimento de objeto simulador mamográfico,” M.Sc. dissertation, Universidade Federal de Uberlândia, Uberlândia, Brazil, 2024, doi: 10.14393/ufu.di.2024.735.

[18] M. C. M. dos Santos, ”Desenvolvimento e aprimoramento de Phantoms antropomórficos mamográficos,” M.S. thesis, Programa de Pós-graduação em Engenharia Biomédica, Universidade Federal de Uberlândia, Uberlândia, Brazil, 2025, doi: 10.14393/ufu.di.2026.5550.

[19] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[20] M. N. Gurcan, Y. Yardimci, A. E. Cetin, and R. Ansari, “Detection of microcalcifications in mammograms using higher order statistics,” IEEE Signal Process. Lett., vol. 4, no. 8, pp. 213–216, Aug. 1997, doi: 10.1109/97.611922.

[21] A. Belarra et al., “Experimental determination of effective X-ray attenuation coefficients of 3D-printed materials under clinical mammography spectra,” Front. Bioeng. Biotechnol., vol. 13, 2026, doi: 10.3389/fbioe.2025.1719551.

Reconhecimento de padrões vocais associados a estresse/emoções negativas em Comunicação Ruidosa via Adaptação Adversarial de Domínio

1st Neander Danubio Marinho Andrade
Programa de Pós Graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
neander.eng@alu.ufc.br

2nd Alberto Matheus de Norões Milfont Carvalho
Departamento da Indústria (DEIND)
Instituto Federal do Ceará (IFCE)
Fortaleza, Brasil
alberto.matheus.noroes60@aluno.ifce.edu.br

3rd José Neco Saulo Capistrano
Programa de Pós Graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
jose.neco@exemplo.com

4th Victor Hugo Costa de Albuquerque
Departamento de Engenharia de Teleinformática (DETI)
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
victor.albuquerque@ieee.org

***Abstract*—Speech foundation models have commonly been used in *Speech Emotion Recognition* (SER) tasks. However, as mostly researches focus on controlled datasets, evaluating the efficiency of those architectures in real world environments is fundamental. This paper focus on investigating the *domain shift* and channel noise impact on the automatic classification of vocal stress in public security communication. By using the foundation model Wav2Vec2/XLS-R, has been analyzed the performance degradation of trained classifiers in clean lab environment (RAVDESS and SAVEE) when exposed to a real emergency calls dataset, EmerCall-VoiceStress (EC-VS). To counter this problem, it's proposed a adversarial architecture based on DANN Networks, composed by a *Gradient Reversal Layer* (GRL) for minimax optimization. The results show that studio trained models suffer from severe collapses of generalization because of the audio noise and compression, proving that the acoustic invariance is not native for those models. The proposed adversarial adaptation approach mitigated the collapse, elevating the *Macro F1-Score* in the operational noisy scenario to 0.7238), suggesting that the alignment of local distribution efficacy for trial critics infrastructure.**

***Index Terms*—Biomarcadores de voz, Estresse vocal, Wav2Vec2/XLS-R, Deslocamento de domínio, Segurança pública.**

I. Introdução

O avanço recente de pesquisas em *Machine Learning* (ML) e *Deep Learning* (DL) impulsionou o desenvolvimento de modelos cada vez mais eficientes e robustos para aplicações em diversas áreas, como detecção e reconhecimento em imagens, processamento de sinais e análise de áudio. Entre essas aplicações, destaca-se o reconhecimento de emoções na fala (Speech Emotion Recognition – SER), cujo objetivo é identificar estados emocionais a partir das características presentes na voz humana.

A voz pode ser representada por um sinal acústico *s=s(t)*. Onde s(t) é amplitude do sinal para determinado instante t. Esses sinais possuem propriedades que carregam informações relevantes sobre o estado emocional do falante. Para facilitar a análise, é comum representar o sinal no domínio da frequência. Com isso, é possível extrair características como frequência fundamental, energia e conteúdo espectral. Permitindo identificar padrões espectrais que podem ser explorados por modelos de aprendizado de máquina.

Diversos estudos investigam o desempenho de diferentes arquiteturas nessa tarefa. [3] realiza uma análise comparativa entre *Support Vector Machine (SVM)*, *Classification And Regression Tree (CART)* e *Long-Short Term Memory (LSTM)* para reconhecimento de emoções utilizando uma base de dados, composta por amostras rotuladas em categorias como neutralidade, raiva, tristeza e felicidade. Em todos os experimentos, os autores empregam um estágio de pré-processamento para extração dos coeficientes cepstrais em escala Mel (Mel-Frequency Cepstral Coefficients - MFCC). De forma semelhante, [19] propõe um modelo para classificação de choros de recém-nascidos, destacando as dificuldades de obtenção de bases de dados em ambientes reais e o impacto da presença de ruídos na qualidade das amostras e, consequentemente, no desempenho dos classificadores.

Apesar desses avanços, a maioria dos modelos é treinada e avaliada em bases de dados coletadas sob condições laboratoriais controladas, nas quais fatores como ruído ambiente, distância do microfone e qualidade de gravação são cuidadosamente monitorados. Essa condição difere substancialmente dos cenários operacionais encontrados em aplicações reais. A presença de ruídos intensos, múltiplos interlocutores, interferências de comunicação e sons urbanos aumenta significativamente a variabilidade acústica, degradando as fronteiras de decisão aprendidas durante o treinamento.

Esse problema torna-se particularmente relevante em aplicações voltadas à segurança pública. Policiais, bombeiros e

demais agentes de emergência estão constantemente expostos a situações de elevado estresse físico e emocional, sendo a voz um potencial indicador de seu estado psicofisiológico. Entretanto, as comunicações desses profissionais ocorrem, predominantemente, por meio de rádios transmissores sujeitos a severas degradações de canal, ruídos impulsivos e interferências eletromagnéticas, dificultando a aplicação direta de modelos treinados em ambientes laboratoriais.

A Figura 1 ilustra esse cenário ao comparar espectrogramas de sinais obtidos em condições controladas com sinais provenientes de comunicações reais submetidas a ruído de canal, bem como o efeito da aplicação de técnicas clássicas de processamento digital de sinais na redução dessas degradações.

No contexto de SER, modelos auto-supervisionados, como Wav2Vec2 e XLS-R, têm demonstrado elevada capacidade de extrair representações profundas da fala, alcançando resultados expressivos na identificação de emoções primárias, como felicidade, raiva e medo. Diante desse contexto, este trabalho investiga sistematicamente o impacto das discrepâncias de domínio provocadas por ruídos e degradações de canal na tarefa de reconhecimento de padrões vocais associados a estresse/emoções negativas. Para isso, são avaliados três cenários de referência: (i) um modelo clássico baseado em estatísticas de MFCCs e SVM; (ii) modelos *XLS-R* e *HuBERT* utilizando transferência direta de domínio; e (iii) um pipeline híbrido que combina Subtração Espectral Clássica com modelos *XLS-R* e *HuBERT*.

Com base nas limitações observadas nesses *benchmarks*, propõe-se uma arquitetura baseada em Redes Neurais Adversariais de Domínio (*Domain-Adversarial Neural Network – DANN*), acoplada ao modelo Wav2Vec2/XLS-R. A proposta busca aprender representações latentes invariantes ao domínio de aquisição, reduzindo os efeitos do ruído diretamente no espaço de características e promovendo um sistema de reconhecimento de padrões vocais associados a estresse/emoções negativas mais robusto.

II. Trabalhos Relacionados

A área de Reconhecimento de Emoções na Fala (SER-*Speech Emotion Recognition*) passou de modelos estatísticos rasos para arquiteturas fundacionais de grande escala. Para contextualizar essa evolução, esta seção percorre desde os algoritmos clássicos de classificação até o aprendizado autossupervisionado e as técnicas de adaptação de domínio.

Nos primeiros avanços da área, classificadores como as Redes de Vetores de Suporte (SVM) ganharam espaço ao delimitar hiperplanos de margem máxima para separar classes [1]. Essa abordagem, contudo, dependia do enquadramento manual de atributos acústicos — a exemplo de MFCCs e variações de pitch [2]. Embora métodos baseados em árvores (como CART) e as próprias SVMs apresentem boa estabilidade, a literatura aponta que arquiteturas recorrentes do tipo LSTM lidam melhor com a dinâmica temporal intrínseca ao sinal de áudio, alcançando maior precisão [3].

Com a ascensão do Aprendizado Profundo, o foco migrou para a extração automática de feições. O uso de Redes Neurais Convolucionais (CNN) aplicadas a espectrogramas permitiu capturar padrões hierárquicos do áudio sem engenharia de recursos manual, superando os baselines anteriores [4]. Esse avanço refletiu-se também na visão computacional: modelos como a VGGNet atingiram marcas expressivas na detecção de expressões faciais [5]–[7], enquanto variações de DNNs e CNNs vêm sendo empregadas na saúde mental para mapear quadros de ansiedade e depressão [8].

Posteriormente, a introdução dos Transformers [9] estabeleceu um novo patamar ao modelar dependências de longa distância com mais eficiência que as redes recorrentes. No áudio, o framework wav2vec 2.0 consolidou o pretreinamento autossupervisionado direto do sinal bruto, reduzindo a dependência de grandes volumes de dados rotulados [10]. Analisando a representação interna dessas redes, Pepino et al. [11] observaram que a informação emocional se concentra nas camadas intermediárias, o que justifica a fusão ponderada de representações para aprimorar o desempenho em SER.

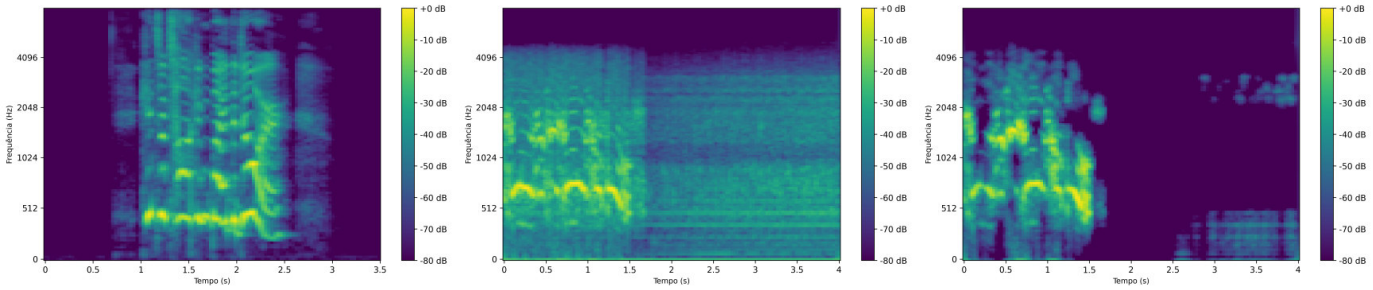
A capacidade de generalização expandiu-se com modelos multilíngues como o XLS-R, treinado em 128 idiomas [12], e com o HuBERT, que emprega agrupamento para inferir unidades ocultas da fala [13]. Estudos comparativos indicam que arquiteturas como XLS-R e Whisper retêm características do locutor de maneiras distintas ao longo de suas camadas [14], sustentando aplicações que vão da síntese de voz a sistemas de diálogo complexos [15].

Apesar desses avanços, a implantação em cenários reais esbarra no deslocamento de domínio (*domain shift*), recorrente em canais de emergência. Para atenuar essa variação acústica, abordagens adversariais como a DANN utilizam classificadores que forçam a rede a extrair atributos invariantes ao ambiente [16]. Em complemento, Hsu et al. [17] mostraram que incorporar áudio não rotulado do domínio-alvo no pré-treino reduz em até 73% a perda de desempenho causada pelo ruído.

Em cenários com restrição de acesso aos dados de origem, estratégias de adaptação contrastiva (como a ECAN) preservam a coerência emocional local e global [18]. Por fim, o monitoramento contínuo de estados emocionais é crítico em ambientes ocupacionais de alto risco, nos quais o estresse afeta a tomada de decisão [20]. Em vez de estimativas estáticas, o uso de Transformers para acompanhar a transição temporal do estresse permite capturar a evolução do estado do indivíduo [20]. Tais sistemas auxiliam desde diagnósticos terapêuticos multimodais [21] até a proteção de profissionais de segurança em situações extremas, contexto que fundamenta o modelo proposto neste trabalho.

III. Materiais e Métodos

Os conjuntos de dados comumente utilizados no treinamento de modelos de *DL* e *ML* são construídos sob condições controladas para garantir que suas amostras se aproximem de cenários ideais. Por exemplo, os sinais de áudio do *dataset* RAVDESS possuem uma duração fixa de 4 segundos, iniciam e terminam invariavelmente com pausas de fala e apresentam ausência de ruídos intensos. O conjunto de dados EmergCall-VoiceStress-(EC-VS) foi construído com base em amostras



(a) Classe *Fear* (RAVDESS). (b) Classe *Stress* (EC-VS Ruidoso). (c) Classe *Stress* (EC-VS Filtrado via DSP).

Fig. 1: Comparativo tempo-frequência do impacto do ruído de canal e da filtragem analítica no espectro da fala.

de áudios em ambientes operacionais de segurança pública. Há a presença de diferentes tipos de ruído, como sirenes de viaturas, estática de rádio, ruído em virtude trânsito etc. Inicialmente, foi realizada uma análise comparativa entre diferentes abordagens para reconhecimento de emoções na fala (SER), com o objetivo de avaliar seu desempenho no conjunto de dados EC-VS. Para isso, foram definidos cinco procedimentos representativos da literatura. Os procedimentos do ceneário 1 empregas um classificador SVM utilizando coeficientes cepstrais em frequência Mel (MFCC) como descritores acústicos. Os procedimentos do cenário 2 utilizas modelos auto-supervisionados pré-treinados (HuBERT e XLS-R) para extração de representações latentes da fala. Nos procedimentos do cenário 3, os mesmos modelos são empregados após a aplicação da técnica de subtração espectral e Digital Signal Processing - (DSP), com o objetivo de investigar o impacto da redução de ruído sobre o desempenho dos classificadores. O último procedimento correspondem à arquitetura proposta baseada em adaptação de domínio adversarial, sendo utilizados para comparação com as abordagens convencionais.

A fim de reduzir o efeito da aleatoriedade inerente ao processo de treinamento, cada experimento foi executado utilizando cinco sementes aleatórias (*seeds*), correspondentes aos valores 42, 123, 777, 999 e 2024. Os resultados apresentados correspondem à média e ao desvio padrão obtidos entre essas execuções, proporcionando uma avaliação estatisticamente mais representativa do desempenho dos modelos. Em todos os experimentos de referência, os modelos foram treinados durante cinco épocas (*epochs*). Os resultados são mostrados nas Tabelas I e II.

O modelo de extração de características acústicas adotado foi o **Wav2Vec2/XLS-R (300M)**, pré-treinado em 436 mil horas de fala multi-idioma. O sinal de áudio bruto foi remodelado nativamente para 16 kHz e truncado em janelas máximas de 4 segundos (64.000 amostras). As camadas convolucionais internas do *feature encoder* foram congeladas (requires_grad = False), atuando estritamente como extratoras de tensores latentes para classificação lineares. Também fora usado o HuBERT Base (LS-960), pré-treinado de forma auto-supervisionada em aproximadamente 960 horas do corpus LibriSpeech.

A. Mapeamento de classes

Devido à indisponibilidade de conjuntos de dados públicos contendo amostras de fala de indivíduos diagnosticados por profissionais de saúde com estresse ocupacional, foi necessário estabelecer um mapeamento entre as emoções presentes nos conjuntos de dados RAVDESS e SAVEE e as classes de interesse deste trabalho.

No conjunto de dados RAVDESS, as emoções *Neutral* e *Calm* foram agrupadas e mapeadas para a classe *Normal/Calm*, enquanto as emoções *Fearful* e *Angry* foram agrupadas e mapeadas para a classe *Stress/Fear*. De forma análoga, no conjunto de dados SAVEE, a emoção *Neutral* foi associada à classe *Normal/Calm*, ao passo que as emoções *Anger* e *Fear* foram agrupadas e mapeadas para a classe *Stress/Fear*.

Esse mapeamento teve como objetivo aproximar emoções negativas frequentemente associadas a estados de tensão e sofrimento psicológico da classe *Stress/Fear*, preservando as emoções neutras e de calma na classe *Normal/Calm*. Embora essa simplificação não represente um diagnóstico clínico de estresse ocupacional, ela permite investigar a viabilidade do uso de modelos de aprendizado profundo na identificação de padrões vocais potencialmente relacionados a estados de estresse em um cenário de classificação binária.

B. Datasets

Três bases de dados com características acústicas distintas foram empregadas neste trabalho:

- **RAVDESS:** Base de dados audiovisual de fala e canção emocional validada academicamente. Foram extraídas amostras de falas limpas gravadas em estúdio de alta fidelidade, mapeando as classes neutra (*Normal*) e medo (*Stress*).
- **SAVEE:** Banco de dados focado em expressões emocionais masculinas gravadas sob condições controladas de laboratório. Utilizou-se um *script* de filtragem para extrair as amostras com base na nomenclatura dos arquivos, separando expressões neutras (*Normal*) de manifestações de medo (*Stress*).
- **EmergCall-VoiceStress (EC-VS):** Conjunto de dados do mundo real contendo 591 amostras de áudio provenientes de chamadas de emergência autênticas de segurança

pública. Inicialmente havia 90 amostras para a classe de Stress/Fear e 501 para classe Normal. Este conjunto caracteriza-se pela presença severa de ruídos de fundo, estática de rádio, compressão telefônica e estresse psicológico operacional real. Todas as amostras foram devidamente rotuladas analisando seu conteúdo, contexto emocional, tom de voz, presença ou não de ruídos indicadores de disparos de armas de fogo, sineres ou qualquer outro elementos que influenciem no comportamento e elementos prosódicos dos falantes. Para extrair amostras com duração de 4s, utilizou-se um script em linguagem *Python*. Os nomes das amostras contém a *string* 'sample' seguido de _ e um número que indica a sequência de extração da amostra, seguida de _ e um valor numérico, 1 para amostras com classificadas contendo estresse e 0 caso contrário. As amostras possuem o seguinte formato: *sample_XXXX_X.wav*). Este será o conjunto de teste para verificarmos a generalização do modelo proposto e realização dos teste t e McNemar.

C. Balanceamento e Data Augmentation dos Domínios

Para mitigar o desbalanceamento severo de classes e prevenir vieses algorítmicos durante o treinamento e alinhamento de domínios, aplicaram-se técnicas de aumento de dados (*Data Augmentation*) via *software* utilizando a biblioteca *librosa*.

Nos domínios de origem (RAVDESS e SAVEE), o procedimento foi executado estritamente sobre a partição de treinamento ($80/20$), garantindo a total blindagem dos conjuntos de validação. Após o procedimento, o conjunto RAVDESS alcançou 960 instâncias (480 por classe) e o SAVEE totalizou 360 amostras (180 por classe).

No domínio alvo (EC-VS), observou-se uma forte assimetria nativa entre as classes. Para evitar que a disparidade amostral comprometesse a representação do espaço de características do domínio alvo durante o alinhamento adverso-dominial, realizou-se a equalização *a priori* da classe minoritária. O processo identificou automaticamente a classe minoritária e gerou amostras complementares em disco até paridade com a classe majoritária, consolidando o conjunto EC-VS em 1.002 instâncias (501 da classe Normal e 501 da classe *Stress*).

Para a geração das novas amostras em todos os conjuntos, alternaram-se duas técnicas de processamento de sinais de áudio:

1) **Deslocamento de Frequência (*Pitch Shifting*):** com variação aleatória de passos entre $[-2{,}0; -1{,}2]$ e $[1{,}2; 2{,}0]$;
2) **Variação de Velocidade (*Time Stretching*):** com fatores de escala aleatórios entre $[0{,}85; 0{,}95]$ e $[1{,}1; 1{,}25]$.

A reprodutibilidade ponta a ponta do *pipeline* de geração foi assegurada pela fixação da semente pseudoaleatória global (seed $= 42$).

D. Primeiro Cenário - coeficientes MFCC e SVM

O primeiro procedimento abordou a engenharia clássica de características acústicas através da extração manual de Coeficientes Cepstrais de Frequência Mel (MFCCs). Os vetores resultantes, contendo as médias e desvios padrões dos coeficientes, alimentaram um classificador SVM de margens suaves com *kernel* de Função de Base Radial (RBF). Os modelos SVM foram treinados em cada conjunto (RAVDESS e SAVEE), utilizando a estratégia de validação cruzada 5-fold e avaliados diretamente no EC-VS. Os resultados obtidos para ambos os cenários *cross-domain* são apresentados na Tabela I.

TABLE I: Desempenho da Validação Cruzada Stratified 5-Fold na Transferência de Domínio para EC-VS no cenário 1.

Dataset	Acurácia Média	F1-score macro médio
RAVDESS	$0{,}4636 \pm 0{,}0633$	$0{,}4145 \pm 0{,}0281$
SAVEE	$0{,}2244 \pm 0{,}0041$	$0{,}2207 \pm 0{,}0047$

E. Segundo cenário - XLS-R e HuBERT padrões

Os procedimentos a seguir verificaram as performances de generalização dos modelos fundacionais XLS-R (facebook/wav2vec2-xls-r-300m) e HuBERT, treinados em RAVDESS e SAVEE, aplicados ao EC-VS sem qualquer processamento ou filtragem de ruídos nos amostras de entrada. Neste cenário, o extrator de características convolucional nativo do modelo foi congelado e as camadas do *Transformer* foram refinadas através de *fine-tuning*. Os resultado são apesentado na tabela III.

F. Terceiro cenário - XLS-R e HuBERT com subtração espectral

Estes Procedimentos avaliaram o impacto do tratamento clássico de sinais no nível de entrada (*input-level adaptation*). Para este fim, os áudios ruidosos do EC-VS foram submetidos a um algoritmo de redução de ruído por Subtração Espectral Clássica (*Spectral Gating*), utilizando os primeiros 0.3 segundos de sinal para a estimativa e atenuação do ruído estacionário de fundo. O sinal tratado foi então injetado nos modelos fundacional XLS-R e HuBERT (previamente tunado nos RAVDESS, SAVEE) para quantificar o ganho obtido pela filtragem analítica isolada. O resultado desse experimento pode ser visualizado na tabela III.

G. Arquitetura proposta - Rede Domain-Adversarial Neural Network (DANN)

Por fim, definiu-se e implementou-se a Arquitetura proposta, composta pelo modelo fundacional XLS-R acoplado a uma Rede DANN atuando diretamente no espaço latente. Para a obtenção de um extrator de características invariante ao canal e ao ruído, a função de perda total do sistema foi modelada pela soma ponderada da perda do classificador de tarefa e da perda do classificador de domínio, otimizadas de forma *minimax* por meio de uma Camada de Reversão de Gradiente (*Gradient Reversal Layer* – GRL). Desta forma, o modelo é forçado a aprender representações que retêm o conteúdo emocional da fala, enquanto ignora ativamente as assinaturas acústicas dos canais ruidosos do EC-VS. Vale salientar que os sinais de entradas também foram submetidos a processamento de sinais com subtração espectral. Durante o passo de otimização, o classificador de emoção opera de forma

TABLE II: Desempenho de treinamento de HuBERT e XLS-R (Média ± desvio padrão).

Cenário	Modelo	Métricas	RAVDESS	SAVEE
2	HuBERT	Accuracy	0.9865 ± 0.0047	0.9917 ± 0.0124
		Precision	0.9866 ± 0.0046	0.9920 ± 0.0118
		Recall	0.9866 ± 0.0047	0.9917 ± 0.0124
		Macro F1	0.9865 ± 0.0047	0.9917 ± 0.0124
		Loss	0.0748 ± 0.0194	0.0506 ± 0.0582
	XLS-R	Accuracy	0.9740 ± 0.0037	0.9972 ± 0.0062
		Precision	0.9743 ± 0.0037	0.9973 ± 0.0060
		Recall	0.9741 ± 0.0037	0.9972 ± 0.0062
		Macro F1	0.9740 ± 0.0037	0.9972 ± 0.0062
		Loss	0.1029 ± 0.0188	0.0867 ± 0.0355
3	HuBERT	Accuracy	0.9938 ± 0.0044	0.9694 ± 0.0317
		Precision	0.9978 ± 0.0048	0.9726 ± 0.0323
		Recall	0.9892 ± 0.0076	0.9667 ± 0.0456
		Macro F1	0.9937 ± 0.0044	0.9694 ± 0.0317
		Loss	0.0306 ± 0.0170	0.1296 ± 0.1265
	XLS-R	Accuracy	0.9927 ± 0.0029	0.9667 ± 0.0320
		Precision	1.0000 ± 0.0000	1.0000 ± 0.0000
		Recall	0.9849 ± 0.0059	0.9333 ± 0.0639
		Macro F1	0.9927 ± 0.0029	0.9666 ± 0.0321
		Loss	0.0484 ± 0.0166	0.2071 ± 0.1109

TABLE III: Resultados Comparativos dos Desempenhos Cross-Domain dos modelos treinados nos conjuntos SAVEE/RAVDESS aplicado ao EC-VS

Procedimento	Base	Acc.	Macro F1	Normal/Calm			Stress/Fear		
				Prec.	Rec.	F1	Prec.	Rec.	F1
P1 (SVM + MFCC)	SAVEE	0.2301	0.2279	0.8966	0.1038	0.1860	0.1576	0.9333	0.2697
	RAVDESS	0.4636	0.4145	0.8387	0.5709	0.6793	0.1400	0.3889	0.2059
P2 (XLS-R)	SAVEE	0.2707	0.2699	1.000	0.140	0.245	0.173	1.000	0.295
	RAVDESS	0.1777	0.1649	0.941	0.032	0.062	0.155	0.989	0.268
P2 (HuBERT)	SAVEE	0.2420	0.2390	1.000	0.106	0.191	0.167	1.000	0.287
	RAVDESS	0.5212	0.4935	0.978	0.445	0.612	0.234	0.944	0.375
P3 (XLS-R)	SAVEE	0.6068	0.5862	0.7066	0.3653	0.4816	0.5720	0.8483	0.6833
	RAVDESS	0.5499	0.4381	0.9630	0.1038	0.1874	0.5264	0.9960	0.6888
P3 (HuBERT)	SAVEE	0.5569	0.5000	0.6748	0.2196	0.3313	0.5340	0.8942	0.6687
	RAVDESS	0.4910	0.4472	0.4795	0.2096	0.2917	0.4943	0.7725	0.6028

supervisionada utilizando as amostras rotuladas do RAVDESS. Simultaneamente, o classificador de domínio atua sobre os dados do EC-VS sem processar seus rótulos de estresse, configurando uma estratégia de adaptação de domínio não-supervisionada (*Unsupervised Domain Adaptation* – UDA). A inferência final do modelo proposto foi validada de maneira holística nos domínios RAVDESS, SAVEE e EC-VS. Os resultados são apresentados na Tabela V e na Figura 2.

IV. Resultados e Discussão

A. Análise do Espaço Latente via t-SNE

Antes de avaliar o desempenho linear dos classificadores, investigou-se a distribuição geométrica das características acústicas extraídas pelo modelo Wav2Vec2/XLS-R. Para tanto, os tensores de alta dimensão gerados pela última camada oculta do *encoder* foram projetados em um espaço bidimensional utilizando a técnica de redução de dimensionalidade não-linear t-SNE (*t-Distributed Stochastic Neighbor Embedding*). O resultado dessa projeção está ilustrado na Fig. 3.

A análise visual da Fig. 3 revela um comportamento crítico para a compreensão dos colapsos de generalização: as amostras

TABLE IV: Resumo de performance do modelo proposto na tarefa de predição nos conjuntos de dados

Domínio (Métricas Globais)	Precision	Recall	F1-Score	support
RAVDESS (Acc: 0.9979 — Macro F1: 0.9979)				
Normal / Calm	0.9959	1.0000	0.9979	480
Stress / Fear	1.0000	0.9958	0.9979	480
SAVEE (Acc: 0.8717 — Macro F1: 0.8382)				
Normal / Calm	0.7639	0.9889	0.8620	180
Stress / Fear	0.9843	0.6944	0.8143	180
EC-VS (Acc: 0.7295 — Macro F1: 0.7238)				
Normal / Calm	0.6780	**0.8743**	**0.7637**	501
Stress / Fear	**0.8230**	0.5848	0.6838	501

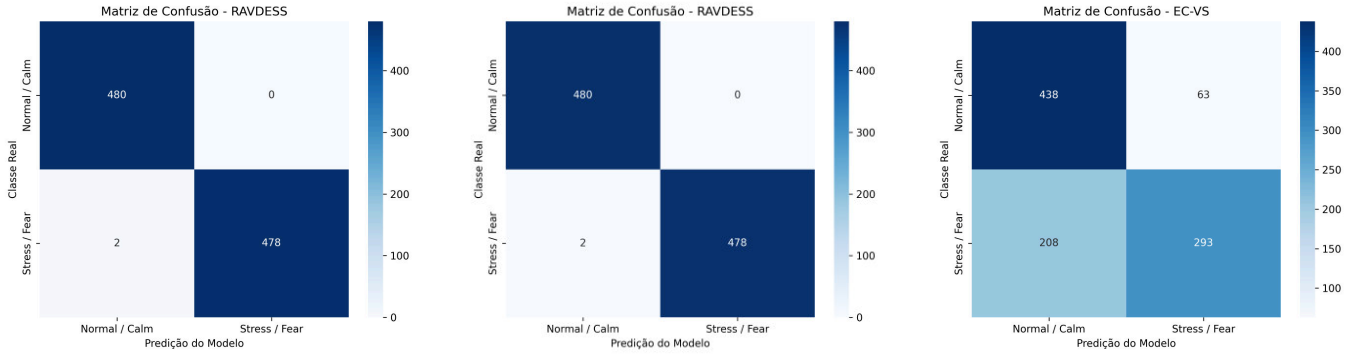


Fig. 2: Matrizes de confusão obtidas como resultados da predição do modelo proposto aplicada aos conjuntos RAVDESS, SAVEE e EC-VS.

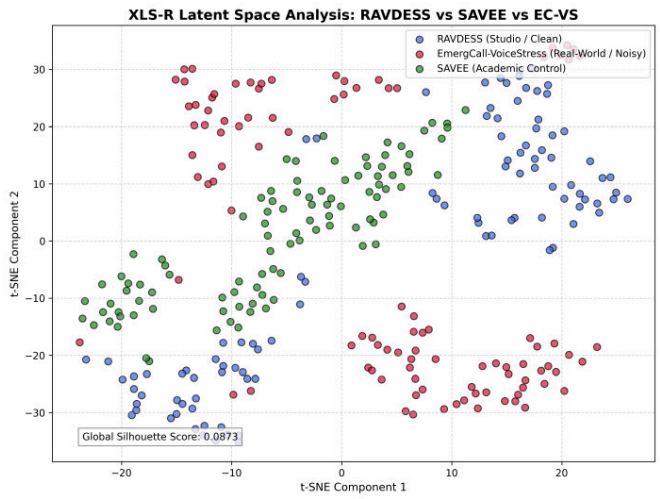


Fig. 3: Projeção t-SNE dos tensores latentes extraídos pelo XLS-R para os datasets RAVDESS, SAVEE e EC-VS. Observa-se a tendência de agrupamento de amostras em razão da qualidade do sinal em detrimento da separação semântica das classes emocionais.

não se agrupam com base nas classes emocionais alvo (*Normal* vs. *Stress*), mas sim de forma estritamente segregada por *dataset* de origem. Os clusters do RAVDESS e do SAVEE (ambos gravados em ambientes controlados de laboratório) apresentam fronteiras bem definidas e isoladas do cluster pertencente ao EmergCall-VoiceStress (EC-VS).

Essa separação baseada no domínio físico e de captação é quantificada pelo Escore de Silhueta Global, que computou um valor de apenas **0,0873**. Esse índice próximo de zero confirma estatisticamente a sobreposição quase completa das distribuições de estresse e neutralidade entre domínios distintos. O modelo fundacional, em seu estado pré-treinado, demonstra maior sensibilidade às variações físicas do meio de transmissão (ruído de rádio, resposta de frequência do microfone e compressão de codec) do que aos biomarcadores prosódicos e dinâmicos do estresse vocal humano. Essa distorção geométrica no espaço de Hilbert da rede justifica e antecipa os colapsos observados nos testes cruzados convencionais.

O cálculo do Maximum Mean Discrepancy (MMD) após o treinamento adversarial da DANN resultou em um valor de **0,5259**. Embora teoricamente o alinhamento perfeito tenda a zero, na adaptação de domínio aplicada à fala operável, a convergência para um valor intermediário reflete o equilíbrio ótimo do jogo minimax. Este índice demonstra que a rede mitigou as discrepâncias severas causadas pela estática de rádio e codecs de compressão do domínio alvo (EC-VS), preservando simultaneamente a separação geométrica necessária para a discriminação dos biomarcadores prosódicos de estresse, o que é empiricamente validado pelo expressivo Macro F1-Score de 85,87% obtido no cenário real.

B. Procedimentos

A tabela III mostra que as abordagens com SVM, XLS-R, e XLS-R com subtração espectral obtiveram resultados

similares. E os resultados sugerem que todos os modelos tendem a encontrar similaridades presentes em sinais ruidosos com características de stress/Fear presentes nas amostras. Os modelos tendem classificar os sinais ruidosos inerentes aos ambientes de comunicação policial operacional

C. Resultados da Arquitetura DANN

Conforme apresentado na Tabela IV, o modelo proposto atingiu uma precisão de 82,30% para a Classe *Stress/Fear*, demonstrando indícios de reconhecimento dos elementos componentes do fala apesar da presença de ruído característico de ambientes operacionais policiais. Além disso, o modelo alcançou uma acurácia global de 72,95% no conjunto de teste.

D. Teste t Pareado e Bootstrap

Com o objetivo de fundamentar a significância estatística da diferença de desempenho entre a arquitetura proposta (DANN) e o modelo *baseline* (XLS-R com DSP), aplicou-se o Teste t de Student pareado sobre a taxa de erro individual por instância e a técnica de *Bootstrap* sobre a métrica F1-Macro.

O Teste t pareado avaliou a inferência de ambas as arquiteturas para cada uma das $N = 1002$ amostras contidas no conjunto de teste EC-VS (composto por 501 amostras para cada classe). Para cada instância i, atribuiu-se o valor 1 em caso de erro de classificação e 0 em caso de acerto, obtendo-se os vetores binários de erro para cada modelo. A diferença pareada por amostra d_i é calculada conforme a Equação 1:

$$d_i = x_{\text{DANN},i} - x_{\text{DSP},i} \tag{1}$$

em que $d_i = 0$ representa empate no resultado da predição (ambos os modelos acertaram ou ambos erraram), $d_i = -1$ denota acerto exclusivo do modelo DANN e $d_i = 1$ indica acerto exclusivo do modelo DSP.

A estatística do Teste t pareado é calculada por meio da Equação 2:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{N}}} \tag{2}$$

onde $\bar{d}$ é a diferença média observada nas predições (Equação 3), s_d é o desvio-padrão amostral das diferenças (Equação 4) e N é o tamanho total da amostra pareada:

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{3}$$

$$s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2} \tag{4}$$

Com $N = 1002$ observações, o teste dispõe de $df = N - 1 = 1001$ graus de liberdade. A aplicação do teste paramétrico é respaldada pelo Teorema do Limite Central, o qual assegura que a distribuição da média amostral das diferenças ($\bar{d}$) aproxima-se de uma distribuição normal para amostras de grande porte ($N \gg 30$). Dito isto, definem-se as seguintes hipóteses estatísticas para o Teste t pareado sobre a taxa de erro:

- H_0: $\bar{d} = 0$ — não existe diferença estatisticamente significante entre a taxa de erro média dos dois modelos no conjunto de teste;
- H_1: $\bar{d} \neq 0$ — existe diferença estatisticamente significante entre a taxa de erro média dos dois modelos no conjunto de teste.

Para o procedimento de *Bootstrap*, realizou-se a reamostragem aleatória com reposição sobre o conjunto de teste EC-VS, repetida 1.000 vezes, utilizando a métrica F1-Macro e intervalo de confiança de 95%. Os resultados detalhados dos testes e da reamostragem são apresentados na Tabela V.

Os resultados evidenciam a superioridade da abordagem de adaptação adverso-dominial (DANN) em relação aos*baselines* do cenário 3. Conforme apresentado na Tabela V, o modelo DANN alcançou uma acurácia de $72{,}95\%$ e um F1-Macro de $0{,}7238$, superando o modelo DSP, que registrou $60{,}68\%$ de acurácia e $0{,}5824$ de F1-Macro — representando um ganho absoluto de $+14{,}14\%$ no F1-Macro.

A análise de reamostragem por *Bootstrap* (1.000 iterações) confirmou a consistência desse ganho, resultando em um Intervalo de Confiança a 95% para a diferença de desempenho de $[0{,}0992;\ 0{,}1789]$. Como o intervalo de confiança da diferença está estritamente acima de zero, a superioridade da arquitetura proposta é confirmada.

Para validar formalmente a significância estatística sobre as 1.002 instâncias de teste do EC-VS, foram aplicados o Teste de McNemar e o Teste t de Student pareado sobre a taxa de erro individual por amostra. Ambos os testes rejeitaram a hipótese nula (H_0) de equivalência entre os modelos com elevada significância estatística ($\chi^2 = 33{,}9043$, $p < 0{,}001$; $t = -5{,}9711$, $p < 0{,}001$). Esses resultados comprovam que a arquitetura proposta inferiu de forma significativamente mais robusta no conjunto de teste EC-VS.

V. Conclusão

Os resultados obtidos no conjunto de teste EC-VS (F1-Macro de $72{,}38\%$), combinados com os resultados da validação estatística (*Bootstrap*, Teste de McNemar e Teste t pareado), sugerem que a abordagem de adaptação adverso-dominial proposta apresenta indícios de ser uma solução mais promissora e robusta para o Reconhecimento de Emoções na Fala (SER) em ambientes operacionais ruidosos do que o emprego de modelos fundacionais submetidos exclusivamente ao processamento digital de sinais (DSP).

VI. Limitações

Para este trabalho utilizamos apenas um conjunto de dados, a saber, EmergCall-VoiceStress (EC-VS). Que é conjunto com gravações de comunicação de rádio em ambiente operacional de segurança pública. Conjuntos de dados dessa natureza são bastante restritos. Portanto, a obtenção de múltiplos conjuntos de dados desse domínio é inerentemente rigorosa e restrita. Dada dificuldade de obtenção, que decorre de questões de legislação e ética, de um conjunto de dados com amostras reais rotulados por profissionais de saúde para casos de estresse.

TABLE V: Resultados Comparativos do Desempenho Cross-Domain no dataset EC-VS

Modelo	**Acurácia**	**F1-Macro**	**IC 95% (F1-Macro)***	**McNemar** (χ^2)	**Teste** t (t)	p**-valor**
XLS-R/DSP	0,6068	0,5824	[0,5509; 0,6144]	—	—	—
DANN/DSP (Proposto)	**0,7295**	**0,7238**	**[0,6946; 0,7496]**	**33,9043**	**-5,9711**	$< 0,001$

*Intervalo de Confiança estimado via Bootstrap (1.000 reamostragens).

Fizemos a aproximação para a classe de stress com as classe de *fearful* (medo) e *angry*(raiva) que estão presentes nos conjuntos RAVDESS e *Fear* e *Anger* no SAVEE. Para EC-VS, a rotulação deu-se manualmente, ouvindo cada amostra e observando a presença da quantidade de interlocutores, de ruído causado por sirenes, ruídos inerentes ao trânsito, de ruídos ocasionados por disparos por arma de fogo e de gritos ou choros. A presença de disparos de arma de fogo e gritos atua como um forte indicador contextual de estresse agudo, justificando a associação semântica entre as classes.

REFERENCES

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] A. A. Yehya and F. M. Ramo, "The intelligent recognition of speech emotions: Survey study," *Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)*, vol. 17, no. 2, pp. 1–10, 2023.

[3] P. Sonu, V. Babu, and M. Anuj, "Diverse machine learning models for speech emotion recognition," in *Proc. 2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2023, pp. 1–6.

[4] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Continuous speech emotion recognition with convolutional neural networks," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 14–24, Feb. 2020.

[5] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," *arXiv preprint arXiv:2105.03588*, May 2021.

[6] S. A. Hussain, A. Sharma, N. Y. Reddy, S. Sharma, J. Srivardhan, and S. Gochhait, "Automated emotion recognition from facial expressions using convolutional neural network," in *Proc. 2024 5th International Conference on Data Analytics for Business and Industry (ICDABI)*, Sakhir, Bahrain, Oct. 2024, pp. 601–606.

[7] C. D. Wu and L. H. Chen, "Facial emotion recognition using deep learning," *The University of Texas at Austin*, Tech. Rep., 2019.

[8] N. Nigar, "Speech emotion recognition using convolutional neural network and its use case in digital healthcare," Master's thesis, Technische Universität Hamburg (TUHH), Jun. 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, Dec. 2017.

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12449–12460, 2020.

[11] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech 2021*, Brno, Czechia, Aug.-Sep. 2021, pp. 3405–3409.

[12] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, Nov. 2021.

[13] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, Oct. 2021.

[14] L. Stuhlmann and M. Saxer, "Evaluating the effectiveness of transformer layers in Wav2Vec 2.0, XLS-R, and Whisper for speaker identification tasks," *ZHAW School of Engineering*, Tech. Rep., 2024.

[15] S. Latif, A. Zaidi, H. Cuayáhuitl, F. Shamshad, M. Shoukat, M. Usama, and J. Qadir, "Transformers in speech processing: A survey," *Queensland University of Technology*, arXiv preprint arXiv:2303.11602, 2023.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2130, Apr. 2016.

[17] W. N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, Sep. 2021.

[18] Y. Zhao, J. Wang, C. Lu, S. Li, B. W. Schuller, Y. Zong, and W. Zheng, "Emotion-aware contrastive adaptation network for source-free cross-corpus speech emotion recognition," in *Proc. ICASSP 2024*, 2024.

[19] H. Chaudhari, A. J. Shah, and H. A. Patil, "Cross lingual speech representation for infant cry classification," in *Proc. 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Gandhinagar, India, Dec. 2024, pp. 2154–2161.

[20] V. Lall and Y. Liu, "Dynamic stress detection: A study of temporal progression modelling of stress in speech," *Singapore Maritime Institute (SMI)*, Tech. Rep., 2025.

[21] P. Lestari, D. Pranata, R. Wulandari, and A. Santoso, "Neural networks for emotion recognition in therapy settings," *Universitas Indonesia*, Tech. Rep., Dec. 2017.

Recurrence Quantification Analysis of Resting-State EEG in Post-Stroke Rehabilitation with tDCS and Extended Reality

Hellen Cristina da Costa
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas (UNICAMP)
Campinas, SP, Brasil
hellencdacosta@gmail.com

Bruna M. Carlos
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, SP, Brasil
bruna.carlos@usp.br

Cassio V. Ruas
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas (UNICAMP)
Campinas, SP, Brasil
cassiovruas@gmail.com

Pedro Felipe G. de Vazquez
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas (UNICAMP)
Campinas, SP, Brasil
p263293@dac.unicamp.br

Alexandre F. Brandão
Escola Politécnica
Pontifícia Universidade Católica de Campinas (PUCC)
Campinas, SP, Brasil
alexandre.brandao@puc-campinas.edu.br

Gabriela Castellano
Instituto de Física Gleb Wataghin
Universidade Estadual de Campinas (UNICAMP)
Campinas, SP, Brasil
gabriela@unicamp.br

***Abstract*—Stroke is a leading cause of long-term disability, and there is growing interest in combining transcranial direct current stimulation (tDCS) with extended reality (XR) to enhance motor recovery. Most electroencephalography (EEG) evaluations of such protocols rely on linear measures, which may overlook the non-linear, non-stationary nature of brain activity. This work applies Recurrence Quantification Analysis (RQA) to eyes-closed resting-state EEG from chronic stroke survivors undergoing a ten-session rehabilitation program, comparing a real tDCS group of ten patients with a sham group of eight patients. Broadband (1–50 Hz) motor-cortex signals were embedded in phase space without selective narrowband filtering, and determinism, laminarity, and diagonal-line entropy were computed per channel and averaged within the affected and non-affected hemispheres. The change between baseline and post-rehabilitation was contrasted between groups with non-parametric tests and false discovery rate correction. Determinism and laminarity in the affected hemisphere increased significantly in the real tDCS group while decreasing in the sham group; the remaining comparisons followed the same opposite-direction trend at borderline significance. These findings are consistent with changes toward more deterministic recurrence patterns after real stimulation, and suggest that RQA is a promising complementary tool for assessing XR- and tDCS-based stroke rehabilitation.**

***Index Terms*—recurrence quantification analysis, electroencephalography, stroke rehabilitation, transcranial direct current stimulation, extended reality**

I. Introduction

Stroke is one of the leading causes of physical disability in adults and the leading cause of premature death in Brazil [1]. The damage to descending neural pathways caused by stroke impairs neuromotor control and leads to functional and structural alterations, resulting in chronic disability and loss of autonomy in performing daily activities [2]–[4]. Persistent muscle weakness and paresis arise, to a large extent, from deficits in voluntary muscle activation, a condition that represents one of the main causes of neurological disability after stroke [5], [6]. Despite advances in rehabilitation strategies, approximately 80% of patients continue to present limitations when treated only with traditional therapeutic approaches [7], [8], which motivates the search for new interventions.

In this context, several neurotechnologies have been incorporated into motor rehabilitation programs. Among them, extended reality (XR) applications stand out by providing engaging and safe environments for exercise practice, as does transcranial direct current stimulation (tDCS), a non-invasive technique that modulates cortical excitability through the direct application of low-intensity electrical current to target brain areas [9]–[11]. Recent studies indicate that the combination of XR and tDCS may enhance the recovery of post-stroke patients [9], [12]. In particular, investigations conducted by our group, based on the same protocol whose data are analyzed in the present work, found greater muscle activation during hip flexion of the affected limb and improvement in the Timed Up and Go test after XR+tDCS rehabilitation [13].

Traditionally, the effects of these interventions are assessed by means of clinical and functional scales. More recently,

electroencephalography (EEG) has stood out as a tool for assessing neurophysiological correlates, by combining high temporal resolution, low cost, and portability, features that allow the reorganization of neural circuits to be recorded at the same site where the intervention is performed [11], [13]. In [14], the authors found increased EEG interhemispheric symmetry and functional connectivity in the alpha band after XR+tDCS rehabilitation [14]. The EEG analyses in these studies, however, have relied predominantly on linear measures, such as spectral, symmetry, and connectivity indices. Since the EEG signal is complex, non-stationary, and nonlinear in nature, such measures may fail to capture relevant aspects of brain dynamics. Complex systems approaches have proven valuable across a wide range of domains, from climate and Earth system dynamics [15], [16] to the near-Earth electromagnetic environment [17], providing a general and physically grounded framework for characterizing nonlinear, non-stationary processes [18]. In this scenario, approaches based on dynamical systems emerge as a robust quantitative alternative for modeling the stability, variability, and complexity of brain activity over time [19].

Recurrence Quantification Analysis (RQA) is one such approach. Grounded in recurrence plots [20], which identify when a system revisits similar states in phase space, RQA quantifies the structures formed by the diagonal and vertical lines of these plots through a set of complexity measures [21]. Although the recurrence rate (RR) expresses the overall density of recurrent states, the present work focuses on the structural measures derived from the line distributions: determinism (DET), associated with the predictability of the system [22]; laminarity (LAM), related to the occurrence of intermittent or laminar states [23], [24]; and the Shannon entropy of the diagonal line lengths (ENTR), which quantifies the complexity of the recurrence patterns [22], [25]. RQA is particularly suitable for the analysis of short, noisy, and non-stationary signals such as EEG [26], [27], and has been successfully employed to characterize transitions and dynamical brain states, including the distinction between normal and epileptiform states [28]–[31].

Despite the potential of RQA to reveal aspects of brain dynamics not accessible to linear analyses, its application to the assessment of post-stroke rehabilitation protocols based on XR and tDCS is still little explored. Therefore, the aim of this work is to apply RQA to resting-state (eyes-closed) EEG signals from post-stroke patients undergoing a rehabilitation program that combines XR and tDCS, contrasting a group that received real stimulation with a sham group. The analysis focuses on the motor cortex channels, organized by hemisphere (affected and non-affected), and quantifies, for the DET, LAM, and ENTR metrics, the change between the beginning (pre-rehabilitation) and the end (post-rehabilitation) of the program. The central question is whether this pre- to post-rehabilitation change differs between the real and sham groups, that is, whether RQA is able to discriminate the neurodynamic effects of real stimulation. An overview of the complete analysis workflow, from the experimental protocol and EEG preprocessing to the RQA pipeline, is presented in Fig. 1.

TABLE I
PATIENT CHARACTERISTICS. SUBJECTS S01–S10 RECEIVED ACTIVE TDCS; SUBJECTS S11–S18 RECEIVED SHAM STIMULATION.

Subject	Age	Sex	Affected hemisphere	Months post-stroke
S01	36	F	D	29
S02	65	M	D	51
S03	64	F	D	72
S04	60	M	D	26
S05	61	F	E	23
S06	66	M	D	8
S07	67	M	E	17
S08	77	M	E	49
S09	63	F	D	76
S10	43	M	D	6
S11	68	F	D	7
S12	49	F	D	23
S13	52	F	D	24
S14	62	F	D	12
S15	51	F	E	13
S16	86	M	D	75
S17	52	M	E	20
S18	33	M	E	7

II. MATERIALS AND METHODS

A. Experimental protocol and data acquisition

The data analyzed in this work were collected in a clinical trial of XR combined with tDCS for chronic stroke rehabilitation [13], [14]. The study was approved by the Research Ethics Committee of the University of Campinas (CEP 5404-CAAE 35771314.4.0000.5404) and conducted in accordance with the Declaration of Helsinki and the CONSORT guidelines [32]. All participants provided written consent before entering the study. Eighteen chronic stroke survivors were allocated into two parallel groups in a block randomized controlled design: ten patients received real anodal tDCS (tDCS group) and eight received sham stimulation (Sham group). Both groups performed the same rehabilitation program. The participants and the remaining testers and instructors were blinded to the stimulation condition and group allocations. Table I summarizes the demographic and clinical characteristics of the participants, including age, sex, affected hemisphere, and time since stroke. All 18 participants included in this study completed all rehabilitation sessions. Clinical and functional assessments were performed in sessions 1, 5, and 10 of the rehabilitation program, in addition to the resting-state EEG recordings analyzed in this work. These included the Fugl-Meyer Assessment for upper- and lower-limb motor function [33], the Mini-Mental State Examination [34], the Berg Balance Scale [35], the Functional Independence Measure [36], the Timed Up and Go test [37], and the 10-m walk test, following the protocol described in [13].

The rehabilitation program comprised ten sessions of 30 min, carried out daily (five times a week) over two weeks, organized into two blocks of five sessions separated by a 48-h interval. In each session, participants interacted with the BRAINN_XR software [38], positioned 2 m from a large

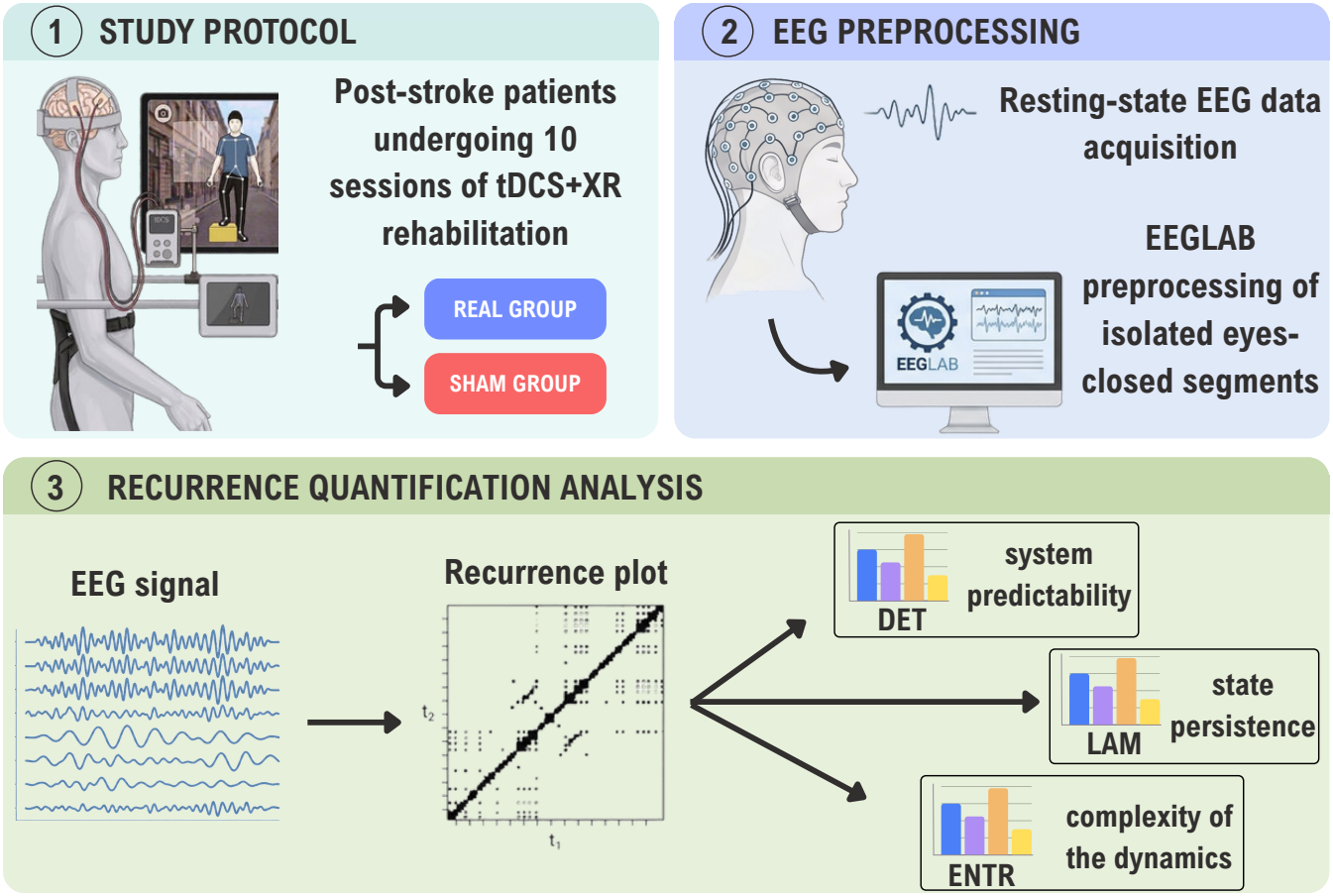


Fig. 1. Overview of the analysis workflow. (1) Study protocol: post-stroke patients underwent ten sessions of combined tDCS and XR rehabilitation and were assigned to a real-stimulation group or a sham group. (2) EEG preprocessing: resting-state signals were acquired and the eyes-closed segments were isolated and preprocessed in EEGLAB. (3) Recurrence Quantification Analysis: each EEG channel was reconstructed in phase space to build its recurrence plot, from which the determinism (DET), laminarity (LAM), and entropy (ENTR) metrics were extracted, quantifying system predictability, state persistence, and the complexity of the dynamics, respectively.

screen that displayed their body image. The protocol included upper-limb exercises (a jigsaw puzzle task, ePuzzle, performed while sitting, standing, and standing on an unstable surface) and lower-limb exercises (stepping over virtual obstacles and stationary walking through an interactive street-view environment, eMaps), emphasizing both the paretic and non-paretic limbs [39].

During the exercises, the tDCS group received low-intensity anodal stimulation (2 mA) delivered for the full duration of each session (30 min), with the anode placed over the ipsilesional primary motor cortex (M1; C3 or C4) and the cathode over the contralesional supraorbital area (Fp1 or Fp2). For the Sham group, the current was ramped up to 2 mA over the first 30 s and ramped down shortly after, remaining off for the rest of the session, thus reproducing the initial sensation of stimulation without delivering an effective dose.

Resting-state EEG was recorded before the tDCS intervention, in session 1 (baseline) and session 10 (post-rehabilitation). Signals were acquired at 256 Hz with 16 monopolar electrodes positioned according to the 10/20 system (F1, F2, Cz, AFz, F5, F6, C1, C2, C3, C4, C5, C6, P1, P2, P5, P6), as shown in Fig. 2, with the reference electrode on the left mastoid and the ground electrode on the right mastoid. The resting-state paradigm consisted of 64 s with eyes open followed by 64 s with eyes closed, beginning with a short transient interval and separated by a second transient interval between the two conditions. This work used the eyes-closed segment, which favors more stable and reproducible alpha oscillations and better test–retest reliability. The motor-cortex channels (C1–C6) were the focus of the present analysis.

B. Data analysis

1) Preprocessing and segmentation: EEG signals were preprocessed in EEGLAB. The first second of each recording was discarded to remove acquisition transients, and data were band-pass filtered (FIR, 1–50 Hz). For the resting-state segments, artifacts were removed using Artifact Subspace Reconstruction (ASR, Clean Rawdata plugin), followed by removal of power-line noise (60 Hz) with Zapline-plus [40]. Signals were then re-referenced to a robust common average reference (rCAR), with no channels excluded from the average computation, to correct for the electrical asymmetry

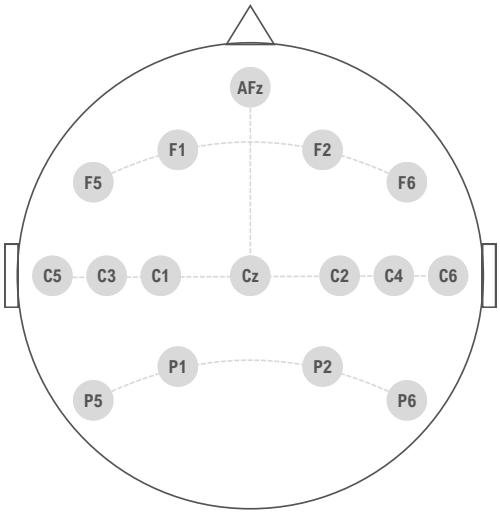


Fig. 2. Spatial layout of the 16 monopolar electrodes positioned according to the 10/20 system. The reference electrode was placed on the left mastoid and the ground on the right mastoid. The motor-cortex channels analyzed in this work (C1–C6) lie along the central row.

introduced by the original left-mastoid reference. No channels were rejected in their entirety. Independent component analysis (ICA) was applied to the data. Artifactual components were initially flagged using the ICLabel plugin and subsequently removed after manual confirmation based on visual inspection of component topography, time courses, and power spectra.

From each recording, the eyes-closed resting-state segment was extracted, corresponding to the final 64 s of the acquisition. To ensure anatomical comparability across patients, a left–right channel flip was applied whenever the lesion was on the left hemisphere, so that the right side consistently corresponds to the affected hemisphere for all participants. The six motor channels were then grouped into the affected (C2, C4, C6) and non-affected (C1, C3, C5) hemispheres. All analyses compared the baseline condition (session 1) with the post-rehabilitation condition (session 10).

2) Recurrence Quantification Analysis: A dynamical system exhibits recurrence when its trajectory returns to a neighborhood of states previously visited in phase space. This property is captured by the recurrence matrix [20], [21]:

$$\mathbf{R}_{ij}(\epsilon) = \Theta(\epsilon - \|\vec{x}_i - \vec{x}_j\|) \qquad i, j = 1, ..., N \tag{1}$$

where $\vec{x}_i$ are the phase-space vectors reconstructed by time-delay embedding, Θ is the Heaviside step function, ε is a threshold distance, and N is the number of states. The graphical representation of this matrix is the recurrence plot (RP), whose small-scale structures, formed by diagonal and vertical lines, reveal the nature of the underlying process: deterministic systems produce long diagonal lines, whereas stochastic processes yield isolated, scattered points (Fig. 3).

From the RP, a set of complexity measures was computed [21], [22], [23]. Determinism (DET) quantifies the predictability of the system; laminarity (LAM) quantifies the occurrence

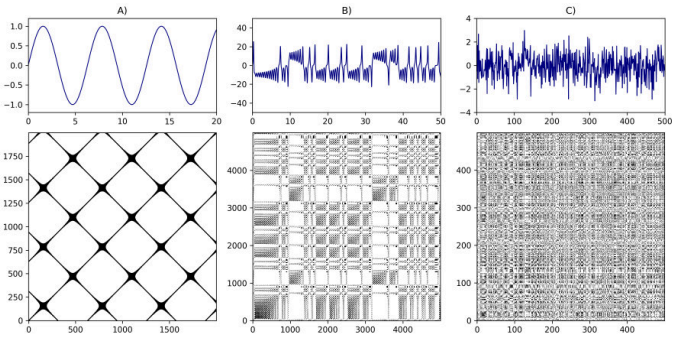


Fig. 3. Time series (top) and corresponding recurrence plots (bottom) for three types of dynamical processes: (A) periodic (sine), (B) chaotic (Lorenz attractor), and (C) stochastic (white noise).

of laminar (intermittent) states; and the Shannon entropy of the diagonal-line lengths (ENTR) quantifies the complexity of the recurrence patterns:

$$DET = \frac{\sum_{l=l_{\min}}^{N} l\, P(l)}{\sum_{l=1}^{N} l\, P(l)}, \tag{2}$$

$$LAM = \frac{\sum_{v=v_{\min}}^{N} v\, P(v)}{\sum_{v=1}^{N} v\, P(v)}, \tag{3}$$

$$ENTR = -\sum_{l=l_{\min}}^{N} p(l)\, \ln p(l). \tag{4}$$

where $P(l)$ and $P(v)$ are the histograms of diagonal and vertical line lengths, respectively, and $p(l)$ is the probability distribution of the diagonal-line lengths. No selective narrow-band filtering (e.g., isolating a single frequency band such as alpha) was applied; signals retained their broadband structure within the 1–50 Hz range used in preprocessing, preserving the joint contribution of multiple frequency components to the analysis. The phase space was reconstructed via time-delay embedding of the EEG signals. To guarantee topological comparability across subjects and groups reconstruction parameters were kept constant. The embedding dimension was set to m = 6, guided by evaluations using the False Nearest Neighbors (FNN) algorithm. For the time delay, we selected τ = 10 samples (approx. 39 ms at 256 Hz) based strictly on the physiological properties of eyes-closed resting-state EEG; this value corresponds to a quarter-cycle of the dominant alpha-band oscillations (10 Hz), ensuring optimal temporal decoupling. The recurrence threshold ε was adapted per channel, from the distribution of pairwise distances, so as to target a recurrence rate of approximately 5%, preventing matrix saturation and controlling for signal amplitude variations. The metrics were computed per channel and then averaged (arithmetic mean) within each hemisphere, rather than from a global signal. The RQA pipeline was implemented in Julia (using the RecurrenceAnalysis.jl and MAT.jl packages).

3) Statistical analysis: Given the small sample size and the non-normal, highly variable nature of the data, all comparisons

relied on non-parametric tests (implemented in Python with scipy). For each patient and metric, the change between baseline (session 1) and post-rehabilitation (session 10) was computed as $\Delta = \text{post} - \text{baseline}$.

The primary inferential comparison was performed between groups: for each metric and hemisphere, the distribution of Δ in the tDCS group was contrasted with that of the Sham group using the Mann–Whitney U test, assessing whether the pre- to post-rehabilitation change differed between real and sham stimulation. As supporting evidence, within-group changes were evaluated with the Wilcoxon signed-rank test (baseline vs. post-rehabilitation) for each group separately. Effect sizes were estimated as $r = |Z|/\sqrt{N}$ (large, $r > 0.5$; medium, $0.3 \leq r \leq 0.5$; small, $r < 0.3$).

To account for multiple comparisons, p-values were corrected with the Benjamini–Hochberg false discovery rate (FDR) procedure, applied separately within each family of tests. The between-group family comprised the six Mann–Whitney comparisons (three metrics $\times$ two hemispheres), and the within-group Wilcoxon tests of the tDCS and Sham groups formed two additional families of the same size. Corrected values are reported as q-values, with significance set at $q < 0.05$. Results were summarized at the hemisphere level (affected and non-affected).

III. Results and Discussion

The RQA metrics were compared between baseline (session 1) and post-rehabilitation (session 10) at the hemisphere level, for the affected and non-affected hemispheres. After FDR correction, statistically significant between-group differences were found for determinism (DET) and laminarity (LAM) in the affected hemisphere; diagonal-line entropy (ENTR) in the affected hemisphere and all three metrics in the non-affected hemisphere reached only borderline significance (q = 0.056). Across all six hemisphere-level comparisons, however, a consistent pattern was observed: the real tDCS group tended to increase the metrics from baseline to post-rehabilitation, whereas the Sham group tended to decrease them, so that the two groups followed opposite trajectories irrespective of whether the contrast crossed the corrected threshold (Figs. 4 and 5).

Between-group changes (Δ = post $-$ baseline) were the primary inferential comparison (Mann–Whitney U, Benjamini–Hochberg FDR; Fig. 4). In the affected hemisphere, both DET and LAM differed significantly between groups (q = 0.044, r = 0.59; uncorrected p = 0.015), with the real group increasing and the Sham group decreasing. The corresponding effect for ENTR in the affected hemisphere was in the same direction but did not survive correction (q = 0.056, r = 0.52). In the non-affected hemisphere, the three metrics followed the same real-increase / sham-decrease pattern, yet all remained just above threshold after correction (DET, r = 0.48; LAM and ENTR, r = 0.46; q = 0.056 for the three).

Higher DET and LAM reflect, respectively, a larger proportion of recurrence points organized into diagonal structures (more predictable, regular trajectories in phase space) and a greater occurrence of laminar, persistent states; ENTR captures the complexity of these recurrence structures [21], [22], [23]. The increases observed in the real tDCS group are consistent with an increase in recurrence-based determinism and laminarity after rehabilitation, a signature that did not emerge, and was often reversed, in the Sham group. This stabilization was most clearly expressed in the affected (ipsilesional) hemisphere, where the DET and LAM changes were robust to correction, whereas the non-affected hemisphere displayed the same directional trend at borderline significance, suggesting a predominantly ipsilesional, but possibly bilateral, pattern of

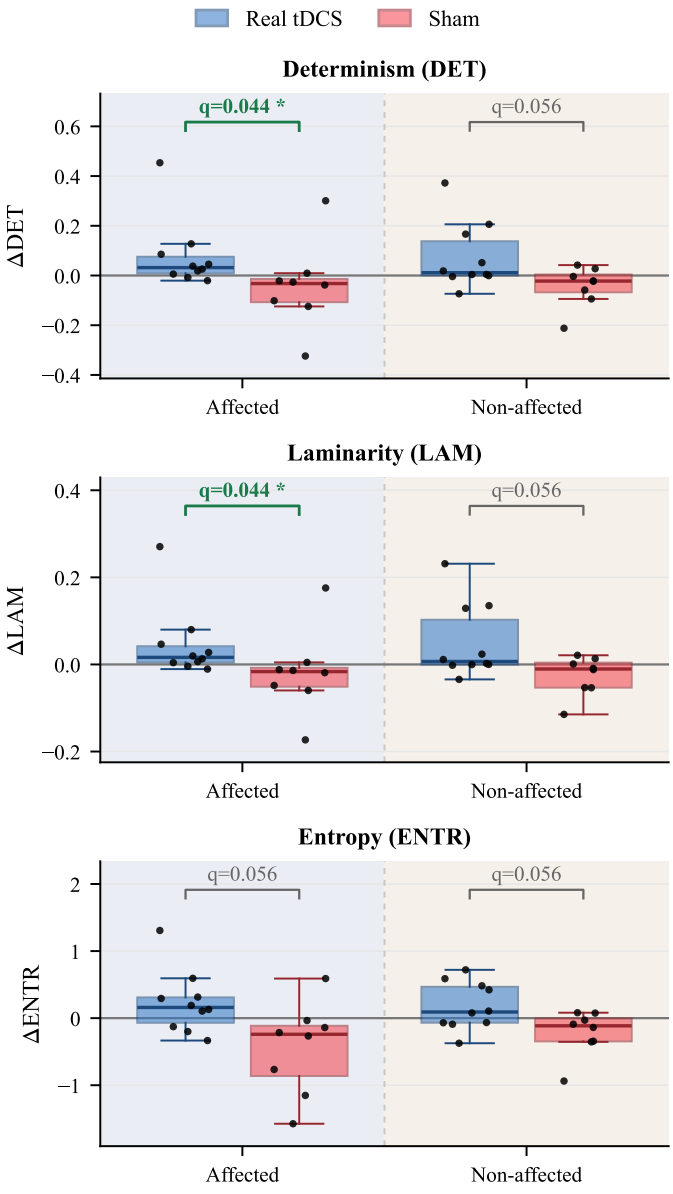


Fig. 4. Per-patient change (Δ = S10 $-$ S1) in the RQA metrics (DET, LAM, ENTR) for the real tDCS and Sham groups, in the affected and non-affected hemispheres. Boxes show the distribution of Δ and points are individual patients; the horizontal line marks no change. Between-group contrasts (Mann–Whitney U, FDR-corrected q-values) are indicated above each pair, and † marks a within-group change (Wilcoxon, $q < 0.05$).

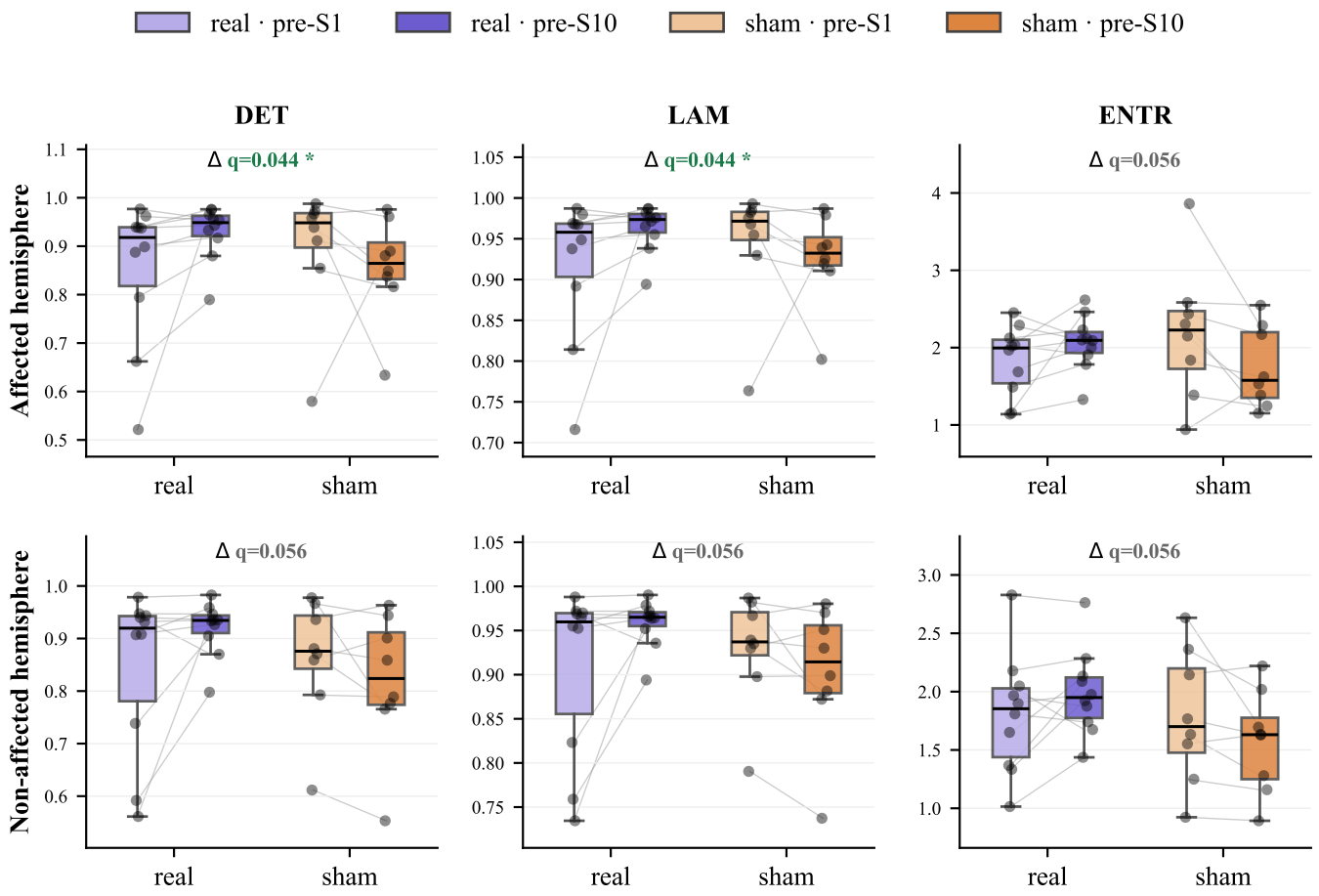


Fig. 5. RQA metrics at baseline (S1) and post-rehabilitation (S10) for the real tDCS and Sham groups, organized by hemisphere (rows) and metric (columns). Thin lines connect the two sessions of each patient. The between-group comparison of the pre- to post-rehabilitation change (Δ; Mann–Whitney U, FDR q-value) is shown at the top of each panel.

change of cortical dynamics.

These findings are consistent with, and complement, previous analyses of the same dataset, in which the combination of XR and tDCS was associated with increased interhemispheric symmetry and alpha-band functional connectivity in chronic stroke patients [11], [14], [41]. Whereas those descriptors are essentially linear, the present results show that a nonlinear, dynamical-systems approach captures convergent changes in cortical activity after real stimulation, which may reflect underlying neural reorganization. A possible interpretation, consistent with proposed mechanisms of tDCS-induced excitability changes, is that anodal stimulation over the ipsilesional M1 may facilitate corticomotor excitability and neuroplasticity [9], [10], [42]; the same-direction trend observed in the non-affected hemisphere is, in turn, compatible with the role attributed to the contralesional hemisphere in post-stroke recovery.

Finally, the convergence between the present nonlinear results and previously reported linear effects, the robustness of the affected-hemisphere DET and LAM effects, and the consistent opposite-direction trajectories of the two groups across all comparisons, together suggest that RQA is a promising complementary tool for characterizing the neurodynamic effects of XR- and tDCS-based rehabilitation in stroke survivors, and motivate its validation in larger cohorts.

This study has some limitations. The main one is the small sample size. Indeed, recruitment of stroke survivors to perform a two-week rehabilitation program is difficult due to their depending on other people (who usually work, and have to miss work) to bring them to the sessions. Also, it is difficult to recruit a population with the exact same lesion characteristics and exact same motor impairments. Although the inclusion and exclusion criteria intended to form a homogeneous sample, this is not always possible. This pilot, exploratory scope also constrained the analysis in other respects: only resting-state EEG was considered, without task-related recordings or follow-up beyond the ten-session program; the comparison was restricted to three RQA metrics, without benchmarking against other nonlinear complexity measures; phase-space reconstruction parameters were fixed rather than tested through a formal sensitivity analysis; RQA metrics were not directly correlated with clinical outcome scores; and hemisphere-level,

non-parametric group comparisons were used in place of participant-level mixed-effects modeling, given the limited statistical power at this sample size. Studies with larger cohorts, incorporating these extensions, are necessary to validate and extend these findings.

IV. Conclusion

This work applied Recurrence Quantification Analysis to eyes-closed resting-state EEG to characterize the neurodynamic effects of a combined XR and tDCS rehabilitation program in chronic stroke survivors, contrasting a real-stimulation group with a sham group. Determinism, laminarity, and diagonal-line entropy were computed per channel, averaged within the affected and non-affected hemispheres, and their pre- to post-rehabilitation change was compared between groups with non-parametric tests and FDR correction.

The real tDCS group exhibited a significant increase in determinism and laminarity in the affected hemisphere, whereas the Sham group decreased; the remaining comparisons followed the same opposite-direction trend at borderline significance. This convergent pattern points to a transition toward more stable, deterministic, and structured cortical dynamics after real stimulation, expressed most clearly in the ipsilesional hemisphere but possibly bilateral.

Taken together with the linear effects previously reported for the same dataset, these results indicate that RQA captures nonlinear aspects of changes is cortical activity not accessible to linear measures, and is therefore a promising complementary tool for assessing XR- and tDCS-based stroke rehabilitation. Validation in larger cohorts remains necessary to confirm and generalize these findings.

Acknowledgment

This research was supported by São Paulo Research Foundation (FAPESP), Grants 2026/01790-5 (HCC), 2020/16571-0 (BMC), 2022/08057-0 (CVR), 2013/07559-3 (BRAINN), 2024/00998-6 (CRISQuaM); Brazilian National Council for Scientific and Technological Development (CNPq), Grants 306255/2024-3 and 408389/2024-9 (INCT-NeuroComp); Funding Authority for Studies and Projects (FINEP), Grant 2140/22; and Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001. The authors also thank Saulo Feitosa, Márcio Vinícius Silva, Larissa L. Pontes, Jayne Silvestre and Sara R. M. Almeida for their assistance in the data collection.

References

[1] M. Corti, C. Patten, and W. Triggs, "Repetitive transcranial magnetic stimulation of motor cortex after stroke: A focused review," Am. J. Phys. Med. Rehabil., vol. 91, no. 3, pp. 254–270, 2012.

[2] V. Azzollini, S. Dalise, and C. Chisari, "How does stroke affect skeletal muscle? State of the art and rehabilitation perspective," Front. Neurol., vol. 12, p. 797559, 2021.

[3] A. Andrews and R. Bohannon, "Decreased shoulder range of motion on paretic side after stroke," Phys. Ther., vol. 69, no. 9, pp. 768–772, 1989.

[4] Y. Borisova and R. Bohannon, "Positioning to prevent or reduce shoulder range of motion impairments after stroke: a meta-analysis," Clin. Rehabil., vol. 23, no. 6, pp. 681–686, 2009.

[5] X. Hu, A. K. Suresh, W. Z. Rymer, and N. L. Suresh, "Assessing altered motor unit recruitment patterns in paretic muscles of stroke survivors using surface electromyography," J. Neural Eng., vol. 12, no. 6, p. 066001, 2015.

[6] V. Hachinski et al., "Stroke: Working toward a prioritized world agenda," Stroke, vol. 30, pp. 127–147, 2010.

[7] E. H. Hoyer and P. A. Celnik, "Understanding and enhancing motor recovery after stroke using transcranial magnetic stimulation," Restor. Neurol. Neurosci., vol. 29, no. 6, pp. 395–409, 2011.

[8] S. Wist, J. Clivaz, and M. Sattelmayer, "Muscle strengthening for hemiparesis after stroke: A meta-analysis," Ann. Phys. Rehabil. Med., 2016.

[9] R. Llorens et al., "Effectiveness of a combined transcranial direct current stimulation and virtual reality-based intervention on upper limb function in chronic individuals post-stroke with persistent severe hemiparesis: a randomized controlled trial," J. NeuroEng. Rehabil., vol. 18, no. 1, p. 108, 2021.

[10] X. Yao et al., "Effects of transcranial direct current stimulation with virtual reality on upper limb function in patients with ischemic stroke: a randomized controlled trial," J. NeuroEng. Rehabil., vol. 17, no. 1, p. 73, 2020.

[11] B. M. Carlos et al., "Graph network and symmetry analysis after combined XR and tDCS in stroke rehabilitation," Biomed. Signal Process. Control, vol. 96, p. 106499, 2024.

[12] J. Meng et al., "Transcranial direct current stimulation with virtual reality versus virtual reality alone for upper extremity rehabilitation in stroke: A meta-analysis," Heliyon, vol. 9, no. 1, p. e12695, 2023.

[13] C. V. Ruas et al., "The effects of transcranial direct current stimulation during extended reality exercises for cortical, neuromuscular, and clinical recovery of stroke survivors," Neural Plast., vol. 2025, p. 5688648, 2025.

[14] B. M. Carlos et al., "Effects in brain symmetry and functional connectivity of tDCS combined with XR in chronic stroke," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 33, pp. 4558–4568, 2025.

[15] N. Boers, J. Kurths, and N. Marwan, "Complex systems approaches for Earth system data analysis," J. Phys. Complex., vol. 2, no. 1, p. 011001, 2021.

[16] S. Gupta et al., "Perspectives on the importance of complex systems in understanding our climate and climate change—The Nobel Prize in Physics 2021," Chaos, vol. 32, no. 5, p. 052102, 2022.

[17] G. Balasis et al., "Complex systems methods characterizing nonlinear processes in the near-Earth electromagnetic environment: Recent advances and open challenges," Space Sci. Rev., vol. 219, no. 5, p. 38, 2023.

[18] J. Kwapień and S. Drożdż, "Physical approach to complex systems," Phys. Rep., vol. 515, no. 3, pp. 115–226, 2012.

[19] U. R. Acharya et al., "Non-linear analysis of EEG signals at various sleep stages," Comput. Methods Programs Biomed., vol. 80, no. 1, pp. 37–45, 2005.

[20] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," Europhys. Lett., vol. 4, no. 9, p. 973, 1987.

[21] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," Phys. Rep., vol. 438, no. 5–6, pp. 237–329, 2007.

[22] C. L. Webber Jr. and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," J. Appl. Physiol., vol. 76, no. 2, pp. 965–973, 1994.

[23] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths, "Recurrence-plot-based measures of complexity and their application to heart-rate-variability data," Phys. Rev. E, vol. 66, no. 2, p. 026702, 2002.

[24] N. Marwan and A. Meinke, "Extended recurrence plot analysis and its application to ERP data," Int. J. Bifurc. Chaos, vol. 14, no. 2, pp. 761–771, 2004.

[25] U. R. Acharya, Y. Hagiwara, S. N. Deshpande, S. Suren, J. E. W. Koh, S. L. Oh, N. Arunkumar, E. J. Ciaccio, and C. M. Lim, "Characterization of focal EEG signals: A review," Future Gener. Comput. Syst., vol. 91, pp. 290–299, 2019.

[26] M. A. Lopes et al., "Recurrence quantification analysis of dynamic brain networks," Eur. J. Neurosci., vol. 53, no. 4, pp. 1040–1059, 2021.

[27] T. L. Prado et al., "Spatial recurrence analysis: A sensitive and fast detection tool in digital mammography," Chaos, vol. 24, no. 1, p. 013106, 2014.

[28] N. Thomasson, T. J. Hoeppner, C. L. Webber Jr., and J. P. Zbilut, "Recurrence quantification in epileptic EEGs," Phys. Lett. A, vol. 279, no. 1, pp. 94–101, 2001.

[29] U. R. Acharya et al., “Automated EEG analysis of epilepsy: A review,” Knowl.-Based Syst., vol. 45, pp. 147–165, 2013.
[30] A. Babloyantz, “Evidence for slow brain waves: a dynamical approach,” Electroencephalogr. Clin. Neurophysiol., vol. 78, no. 5, pp. 402–405, 1991.
[31] B. Akbarian and A. Erfanian, “Automatic seizure detection based on nonlinear dynamical analysis of EEG signals and mutual information,” Basic Clin. Neurosci., vol. 9, no. 4, 2018.
[32] K. F. Schulz et al., “CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials,” BMC Med., vol. 8, no. 1, p. 18, 2010.
[33] A. R. Fugl-Meyer et al., “The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance,” Scand. J. Rehabil. Med., vol. 7, no. 1, pp. 13–31, 1975.
[34] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “‘Mini-mental state’: A practical method for grading the cognitive state of patients for the clinician,” J. Psychiatr. Res., vol. 12, no. 3, pp. 189–198, 1975.
[35] S. T. Miyamoto et al., “Brazilian version of the Berg Balance Scale,” Braz. J. Med. Biol. Res., vol. 37, pp. 1411–1421, 2004.
[36] M. Riberto et al., “Validação da versão brasileira da Medida de Independência Funcional,” Acta Fisiatr., vol. 11, pp. 72–76, 2004.
[37] A. Costa et al., “Instrumentos utilizados no Brasil para avaliação da mobilidade física como fator preditor de quedas em adultos,” Ciênc. Cuid. Saúde, vol. 10, pp. 401–407, 2011.
[38] A. F. Brandão, D. R. C. Dias, G. Castellano, N. A. Parizotto, and L. C. Trevelin, “RehabGesture: An alternative tool for measuring human movement,” Telemed. e-Health, vol. 22, no. 7, pp. 584–589, 2016.
[39] L. G. S. Rodrigues, D. R. C. Dias, M. P. Guimarães, A. F. Brandão, L. Rocha, R. L. Iope, and J. R. F. Brega, “Supervised classification of motor-rehabilitation body movements with RGB cameras and pose tracking data,” J. Interact. Syst., vol. 13, pp. 221–231, 2022.
[40] A. de Cheveigné, “ZapLine: A simple and effective method to remove power line artifacts,” Neuroimage, vol. 207, p. 116356, 2020.
[41] S. Eldeeb et al., “EEG-based functional connectivity to analyze motor recovery after stroke: A pilot study,” Biomed. Signal Process. Control, 2019.
[42] V. Di Lazzaro and J. C. Rothwell, “Corticospinal activity evoked and modulated by non-invasive stimulation of the intact human motor cortex,” J. Physiol., vol. 592, no. 19, pp. 4115–4128, 2014.

Assessment of Remote Photoplethysmography Algorithms for Vital Signs Estimation in Intensive Care Unit Patients: A Pilot Study

Sophia Celine Rafael Alves Pereira
Biomedical Engineering Laboratory
University of São Paulo
São Paulo, Brazil
sophiaceline@usp.br

Paolo José Cesare Biselli
University Hospital
University of São Paulo
São Paulo, Brazil
paolobiselli@usp.br

Felipe Fava de Lima
Biomedical Engineering Laboratory
University of São Paulo
São Paulo, Brazil
felipefavadelima@usp.br

***Abstract*—Vital sign monitoring is essential for evaluating the physiological state of a patient and is typically performed using contact-based medical devices. Nonetheless, non-invasive monitoring approaches offer relevant benefits in both clinical and non-clinical settings by increasing patient comfort, reducing contact-associated risks, and providing greater flexibility of application. In this context, remote photoplethysmography (rPPG) has emerged as a promising technique for estimating physiological parameters such as heart rate, respiratory rate, oxygen saturation and blood pressure by detecting cyclical variations of the cardiovascular system which can be extracted from subtle changes in the color intensity of pixels from the skin region captured by video. Although several methods have been proposed, including unsupervised approaches and deep learning–based techniques, most studies are conducted under controlled conditions and involve predominantly healthy individuals. In view of this limitation, this work aims to assess the performance of the main rPPG algorithms using video recordings of patients in a public Intensive Care Unit (ICU) which is part of the Brazilian Unified Health System (*Sistema Único de Saúde* - SUS). In this preliminary study, the dataset is composed of video recordings acquired using an industrial video camera and reference signals collected directly from bedside multiparameter monitors of an initial cohort of three patients. The results aim to assess the methodological feasibility of evaluating rPPG methods under realistic intensive care conditions of an under-represented population.**

***Index Terms*—remote photoplethysmography, intensive care unit, vital signs**

I. Introduction

Vital sign monitoring plays a significant role in the evaluation of the physiological state of an individual and helps to prevent the deterioration of health conditions [1]. However, the constraints imposed by continuous monitoring using wired devices in direct contact with the body may cause discomfort, particularly during long periods. In fact, these constraints can be considered one of the main stressors for patients in Intensive Care Units (ICUs) [2].

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and by Amigos da Poli endowment fund under CARE 4.0 project.

Given the challenges associated with vital sign monitoring using contact-based medical devices, camera-based physiological monitoring has emerged as a reliable alternative, offering advantages in terms of cost-effectiveness and ease of deployment [3]. Most of these approaches are based on remote photoplethysmography (rPPG), a technique that aims to capture cyclic cardiovascular variations from subtle temporal changes in the intensity of skin color components caused by blood flow. In this process, imaging devices capture the light that has propagated through skin tissues and is subsequently re-emitted [4].

Remote photoplethysmography signal extraction can be performed either under ambient light or controlled illumination sources, using one or more imaging devices for video acquisition. The typical workflow, as shown in Fig. 1, involves the detection and tracking of a region of interest (ROI), from which a raw signal is obtained through the analysis of the color channels of the pixels in each video frame. This signal is subsequently filtered and processed to reconstruct the blood volume pulse (BVP) waveform, also denoted as the rPPG signal, and its properties can then be exploited to estimate physiological parameters such as heart rate (HR) and respiratory rate (RR). Nonetheless, signal quality is influenced by lighting variations, subject movement, ROI definition and detection and other factors regarding the acquisition method such as video compression or even physiological aspects as skin tone [5], [6].

Numerous algorithms have been proposed for rPPG signal extraction. However, the literature has predominantly focused on controlled laboratory environments and healthy participants [7], [8]. Furthermore, the datasets most commonly used for the development and evaluation of these methods underrepresent older adults and individuals with cardiovascular conditions [9], despite the fact that this technology is intended to capture physiological variations that may support the detection and monitoring of health conditions to which these populations are particularly susceptible. Clinical studies remain comparatively scarce and are concentrated primarily in China, the United States and European countries [10]. In

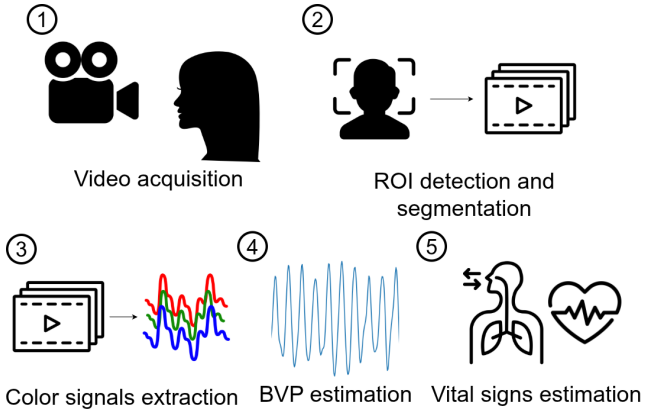


Fig. 1. Remote photoplethysmography-based vital signs estimation workflow. 1) Videos of the subject's face are acquired under natural or artificial lighting. 2) ROI is detected and segmented from video frames. 3) Image color channels are processed for rPPG signal extraction. 4) Final rPPG signal is obtained, corresponding to the estimated BVP waveform. 5) Vital signs such as heart rate and respiratory rate are estimated based on the obtained BVP.

view of these limitations, this pilot study aims to evaluate the methodological feasibility of assessing the main unsupervised and deep learning-based rPPG approaches for HR and RR estimation in an ICU of a hospital within the Brazilian Unified Health System (*Sistema Único de Saúde* - SUS), which is a previously unexplored context, by establishing a video-processing pipeline and an initial data-collection to provide a preliminary qualitative assessment of these algorithms.

II. Related work

Most rPPG signal extraction algorithms are based either on unsupervised methods or on deep learning (DL) approaches. Unsupervised methods estimate physiological signals through color-based analyses of input video data, whereas DL-based approaches include both end-to-end models trained to infer the rPPG signal or physiological parameters directly from video, and hybrid frameworks that integrate traditional signal-processing pipelines with DL components for feature extraction at one or more stages of the estimation process [11].

A. Unsupervised methods

One of the first known rPPG solutions performed a spectral analysis based on the Fast Fourier Transform (FFT) of signals obtained from the spatial mean of pixels of the face region of individuals recorded by an RGB camera. This method resulted in the estimation of HR values and also proposed that the green channel has predominant influence over the rPPG signal [12]. Early rPPG-based systems relied primarily on algorithms and mathematical models designed to process video data acquired from conventional cameras and extract the physiological signal of interest while minimizing the influence of noise. Within the approaches proposed for this purpose, signal separation techniques such as Independent Component Analysis (ICA) and Principal Component Analysis (PCA) became particularly prominent [13], [14]. These methods aim to decompose the observed color variations into underlying components, enabling the isolation of physiological signals associated with cardiovascular activity from motion artifacts, illumination changes, and other confounding factors.

Subsequent model-based approaches primarily rely on modeling color vectors of different components to extract pulse-related information from signals captured by RGB cameras. Among them is the chrominance-based (CHROM) technique, which aims to attenuate the effects of specular reflections and other noise artifacts common to the red, green, and blue channels by projecting the signals onto two orthogonal chrominance components [15]. Similarly, the blood volume pulse signature method, also called PBV, is based on the hypothesis that the different absorption spectra of arterial blood and the cutaneous tissue cause variations in a characteristic vector of the normalized RGB space [16]. In contrast, the Spatial Subspace Rotation (2SR) method defines a subspace from the distribution of skin pixels in the RGB color space, such that pulse-related color variations are represented as rotations within this subspace [17]. Comparably, the Projection Orthogonal to Skin (POS) method proposes a plane orthogonal to the temporally normalized skin-tone vector, in which the specular and diffuse components of reflected light have distinct distributions and can, therefore, be disaggregated [18].

The Local Group Invariance (LGI) method enhances robustness to facial motion and illumination changes by projecting RGB signals onto a subspace orthogonal to dominant interference components. The transformed signal is subsequently modeled as a stochastic oscillatory process that captures the quasi-periodic and non-stationary nature of cardiac activity, enabling heart rate estimation [19]. Lastly, the Orthogonal Matrix Image Transformation (OMIT) method uses QR decomposition to transform RGB signals into an orthogonal basis, from which a projection matrix is derived to extract the BVP waveform from the input color signals [20].

B. Deep-learning-based methods

Deep learning is based on the use of deep neural networks to analyze and identify complex patterns in data. Specifically, Convolutional Neural Networks (CNNs) have been widely used to identify spatial patterns in image and video processing tasks. In fact, in the context of remote photoplethysmography, the DeepPhys model comprises a motion model and an appearance model based on convolutional attention networks. The motion model leverages an attention mechanism to learn motion representations from normalized frame differences, whereas the appearance model uses attention to emphasize skin regions with stronger physiological signals, improving BVP and respiration signal estimation [21]. Similarly, the Multi-Task Temporal Shift Convolutional Attention Network (TSCAN) model uses temporal shift modules such that two-dimensional tensors are shifted through the temporal dimension in order to extract sequential information between multiple frames at the same time that a multi-task mechanism shares pulse and respiratory representations for simultaneous HR and RR estimation [22]. Other solutions comprised three-dimensional CNNs to simultaneously identify temporal and

spatial features. The PhysNet model, for instance, proposes a 3D CNN version capable of extracting the rPPG signal directly from face videos without preprocessing steps [23].

Later applications proposed the use of Transformer architectures to extract the rPPG signal. The PhysFormer model employs the use of a Visual Transformer (ViT) to explore the long-range spatio-temporal relationship for rPPG signal estimation. This solution leverages temporal differences to guide global attention mechanisms and refine local spatio-temporal representations for noise robust signal extraction [24]. The EfficientPhys model employs a Swin Transformer architecture, which is a two-dimensional ViT featuring a shifted-window mechanism which improves computational efficiency. This architecture is combined with a temporal shift module to facilitate the extraction of temporal features, allowing reliable cardiac pulse measurement [25].

III. Materials and Methods

This study aims to assess the methodological feasibility of evaluating the performance of the main rPPG algorithms and models under the context of intensive care. This preliminary work is based on an initial cohort of three videos recorded from patients admitted in the ICU of the University Hospital of the University of São Paulo (HU-USP), which is part of SUS. The workflow comprises the recording of face videos of the patients and corresponding physiological ground truth signals from multiparameter bedside monitors, video processing for rPPG signal extraction, estimation of heart rate and respiratory rate and results evaluation. This work was approved by HU-USP Ethics Committee (*Comitê de Ética em Pesquisa* - CEP) under Certificate of Submission for Ethical Review (*Certificado de Apresentação para Apreciação Ética* - CAAE) No. 86650625.3.0000.0076. All data were acquired with explicit permission of all participants or their legal representatives.

A. Data acquisition

Data were acquired by recording face videos of patients admitted in the ICU under ambient lighting conditions, with no restrictions imposed to the patient during monitoring and no impact on clinical conduct. Initially, after the inclusion of the patient in the study, the recording positioning was defined based on patient disposition inside the room, such that if the patient was lying down in the hospital bed, the camera was positioned at the end of the bed at a distance of approximately 170 cm from the face ROI, whereas if the patient was sitting down in the available chair, the camera was positioned at a distance of approximately 190 cm.

Before each recording, lighting intensity was registered near the face region of the patient with a digital lux meter (Fluke 941, Fluke Corporation, USA). Skin-tone was also measured according to the Fitzpatrick scale using a digital phototype analyzer (Skin Up Beauty Devices, Brazil) positioned in the forehead. Afterwards, videos of approximately two minutes were recorded with an industrial VCXG.2-57C video camera (Baumer, Switzerland) with Ethernet communication to a laptop, configured to capture 25 frames per second with BGR8 pixel format. For each subject, the exposure time, gain, focus and sensor aperture were configured in order to minimize pixel saturation and low illuminance regions caused by underexposure, considering that the exposure time is limited by frame rate. Videos were recorded in uncompressed AVI format with planar YUV 4:2:0 color representation. Natural ICU monitoring conditions were maintained, including patient natural movements, lighting variations and other influencing factors related to the clinical setting.

Physiological ground truth signals were acquired simultaneous to video recordings directly from the bedside monitors (BSM-3000 Series, Nihon Kohden, Japan). The electrocardiogram (ECG) was recorded with a sample rate of 250 Hz, respiratory signal was recorded with sample rate of 125 Hz and the photoplethysmography (PPG) signal was recorded at 62,5 Hz.

B. Data processing

In order to assess different rPPG signal extraction methods in the acquired data, recorded videos were processed in an analysis pipeline implemented based on the open source toolbox rPPG-Toolbox [26]. This platform not only supports the most widely used rPPG model-based algorithms and state-of-the-art deep learning models for training and testing in public datasets, but also supports the addition of new algorithms and customized datasets. Therefore, this work proposes the assessment of the methods shown in Table I in the acquired data using the aforementioned toolbox for rPPG signal extraction.

For each input video, face detection was performed using the Yolo5Face face detector every 25 frames and the resulting bounding box was defined as the ROI [27]. All frames were cropped according to the detected ROI and, specifically for unsupervised algorithms, the spatial mean of the ROI pixels was used as input. For the supervised algorithms, the models trained on the PURE [28] dataset and on the UBFC-rPPG [29] dataset were used.

After video processing, the resulting BVP waveform obtained from each method was filtered between 0.6–3.3 Hz using a band-pass first order Butterworth filter and used for heart rate estimation through spectral analysis performed on 15-second non-overlapping windows. For each window, the power spectral density (PSD) was estimated using the periodogram method, using a rectangular window and the FFT length zero-padded to the next power of two. The frequency corresponding to the maximum PSD value was identified as the dominant spectral peak and taken as the HR estimate. As a second analysis, in order to obtain a global representation of the distribution of spectral energy over the entire recording, alongside a time-frequency analysis to point out temporal variations of the spectral content during acquisition, the PSD estimate and the spectrogram of the entire signal were computed. This PSD was obtained using the periodogram method, with the FFT length zero-padded to the next power of two and a rectangular window. The spectrogram was computed using the Short-Time Fourier Transform (STFT) with 256-sample

TABLE I
EVALUATED RPPG METHODS

Method	Strategy
GREEN [12]	Green channel intensity analysis
ICA [13]	Blind Source Separation
CHROM [15]	Chrominance-based
PBV [16]	Blood volume pulse signature
POS [18]	Projection in the subspace orthogonal to skin tone vector
LGI [19]	Spatial variations modelling for movement robustness
OMIT [20]	QR decomposition
DeepPhys [21]	2D CNN with attention mechanism
TSCAN [22]	2D CNN with temporal shift modules
PhysNet [23]	3D CNN
PhysFormer [24]	Visual Transformer
EfficientPhys [25]	Swin Transformer

windows, 50% overlap, and an FFT length zero-padded to twice the next power of two.

The respiratory rate was estimated by extracting the three types of respiratory-induced variations in the obtained BVP signal namely the baseline wander (BW), the amplitude modulation (AM), and the frequency modulation (FM). Firstly, the BW component was obtained using a discrete sequence formed by the mean between the amplitudes of each trough and its subsequent peak. The AM component was demodulated using the discrete sequence formed by the amplitude differences between each trough and its subsequent peak. The demodulation of the FM component was obtained using the discrete sequence formed by the intervals of consecutive peaks. Subsequently, based on the algorithm proposed by Karlen *et al.* [30], the discrete sequences respective to the AM and BW components were resampled in a 4 Hz uniform temporal grid using linear interpolation, whereas the discrete sequence respective to the FM component was resampled to a 4 Hz uniform temporal grid using the Berger algorithm [31]. Finally, RR corresponding to each component was estimated though spectral analysis performed in 30-second non-overlapping windows. For each window, the dominant spectral peak within the 4–65 rpm frequency range identified in the PSD was taken as the RR estimate. In this case, the PSD was obtained using the periodogram method, with the FFT length zero-padded to the next power of two and a rectangular window. Additionally, in order to enhance RR estimation, a fusion strategy was performed, using the mean of the three values obtained from each component as the final estimate [30], [32], [33].

Ground truth values for HR were acquired from the Lead II ECG signal using the HeartPy toolkit [34] considering 15-second windows. The photoplethysmogram acquired by the pulse oximeter was used to evaluate the waveform estimated by each rPPG method based on the Pearson correlation coefficient computed after temporal alignment using normalized cross-correlation. Lastly, the ground truth respiratory rate was obtained by spectral analysis of the ground truth respiratory signal performed in 30-second windows. In this case, for each window, the PSD was estimated using the periodogram method with the FFT length zero-padded to the next power of two and the dominant spectral peak within 4 and 65 rpm was considered as the ground truth respiratory rate.

The HR and RR estimation results were evaluated by the Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). For additional HR estimation evaluation, the Signal-to-Noise Ratio (SNR) was also obtained by the ratio between the spectral energy associated to the frequency of the ground truth HR value and its first harmonic and the remaining spectral energy, as shown:

$$SNR = 10\log_{10}\left(\frac{\sum_{f=f_{min}}^{f_{max}}\left(U_t(f)S(f)\right)^2}{\sum_{f=f_{min}}^{f_{max}}\left((1-U_t(f))S(f)\right)^2}\right) \quad (1)$$

where S refers to the rPPG signal frequency spectrum, f is the frequency in bpm, $f_{min} = 36$ and $f_{max} = 198$ are respectively the maximum and minimum frequency values considered in bpm and $U_t(f)$ represents a binary mask centered in the fundamental frequency associated to the ground truth HR value and its first harmonic. This mask assumes a unit value in the frequency values of interest, within a tolerance of ±0.1 Hz around these frequencies, and zero elsewhere in the spectrum.

IV. RESULTS

Data were acquired from a total of three patients admitted in the ICU and the acquisition parameters are shown in Table II. Patient 1 was heavily sedated and was receiving respiratory support, whereas patient 2 was active and seated on the chair available in the hospital room. As for the data acquired from patient 3, the pulse oximeter was not connected during recording, therefore waveform comparison between the PPG and rPPG signals could not be obtained. The waveform obtained from each rPPG method for patients 1 and 2 was compared to the ground-truth PPG signal. Pearson correlation coefficients computed after temporal alignment using the lag that maximized the normalized cross-correlation are presented in Table III.

The rPPG methods which resulted in the best Pearson correlation coefficients among the supervised methods were the TSCAN model trained on the UBFC-rPPG dataset for patient 1 and the same model trained on the PURE dataset for patient 2. The signals obtained from unsupervised methods which resulted in the best Pearson correlation coefficients were extracted using the GREEN method for patient 1 and the POS method for patient 2. These signals compared to the respective ground-truth PPG signals are shown in Fig. 2.

A. Heart rate estimation

Table IV reports the MAE, RMSE, MAPE, and SNR of HR estimates derived from rPPG signals, using ECG-based HR as ground truth. The unsupervised method which resulted in the lowest MAE for patient 1 was the CHROM method. As for patient 2, CHROM, ICA and POS resulted in the lowest error values. Comparatively, LGI and POS resulted in the lowest MAE value for HR estimates related to patient 3. In

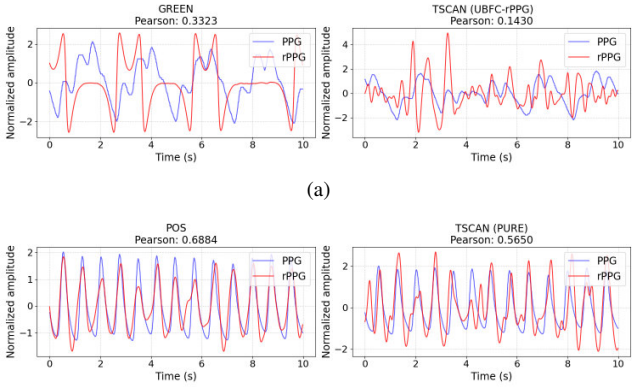
TABLE II
ACQUISITION PARAMETERS FOR EACH PATIENT.

Parameter	Patient 1	Patient 2	Patient 3
Skin type	III	IV	II
Illuminance (lux)	880	850	1080
Camera distance to face (cm)	170	190	175
Gain	1.0	1.0	1.5
Resolution	864×900	2000×2028	1408×1120
Frame rate (fps)	25	25	25
Exposure time (ms)	38000	28000	38000
Pixel format	BGR8	BGR8	BGR8
Color representation	YUV 4:2:0	YUV 4:2:0	YUV 4:2:0
Observation	Intubated patient	Patient sitting and talking	Patient lying and active

TABLE III
PEARSON CORRELATION COEFFICIENTS BETWEEN THE RPPG SIGNALS AND THE GROUND-TRUTH PPG SIGNALS FOR PATIENTS 1 AND 2 AFTER TEMPORAL ALIGNMENT USING NORMALIZED CROSS-CORRELATION.

Method	Patient 1	Patient 2
Unsupervised Methods		
CHROM	0.0991	0.5754
GREEN	**0.3323**	0.1350
ICA	0.1114	0.6783
LGI	0.1305	0.3575
OMIT	0.1336	0.3629
PBV	0.1252	0.5288
POS	0.1184	**0.6884**
Supervised Methods		
DeepPhys (PURE)	0.1070	0.4276
DeepPhys (UBFC)	0.1061	0.4992
EfficientPhys (PURE)	0.0944	0.4576
EfficientPhys (UBFC)	0.0949	0.4342
PhysFormer (PURE)	0.0740	0.4547
PhysFormer (UBFC)	0.0602	0.4141
PhysNet (UBFC)	0.0636	0.3151
PhysNet (PURE)	0.0425	0.4367
TSCAN (PURE)	0.1309	**0.5650**
TSCAN (UBFC)	**0.1430**	0.4505

Terms in parenthesis indicate the dataset on which the models were trained.



(a)

(b)

Fig. 2. Comparison of rPPG signals (red) with highest linear correlation to ground-truth PPG (blue) after temporal alignment using normalized cross-correlation for (a) patient 1 and (b) patient 2.

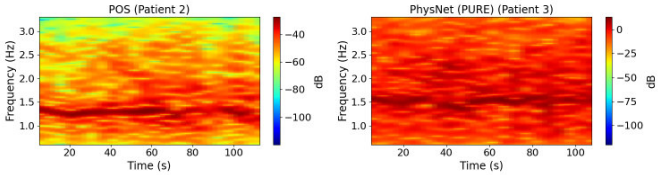


Fig. 3. Spectrograms of the signals which achieved the best heart rate estimation error results. On the left, the spectrogram of the signal extracted with the POS method from the video of patient 2. On the right, the spectrogram of the signal extracted with the PhysNet model trained on the PURE dataset from the video of patient 3.

addition, the supervised methods with the best HR estimates error results were the PhysFormer model for patient 1, TSCAN for patient 2 and PhysNet for patient 3, all trained on the PURE dataset. Regarding the SNR metric, the signal extracted with the POS method showed the best SNR value for patients 2 and 3, whereas the signal extracted with PhysFormer trained on PURE resulted in the best SNR for patient 1.

Fig. 3 shows the spectrograms of the signals that achieved the best HR estimates error results among both types of methods. In both spectrograms, it is possible to distinguishably identify a region with greater spectral density throughout the signal which, in this case, was analyzed in its entirety. Similarly, the corresponding PSD for each signal is presented in Fig. 4 where, in both cases, the dominant frequency can be identified as a distinguishable peak, based on which the HR estimation can be performed. Additionally, the mean absolute error per window for unsupervised methods and for supervised methods for patients 2 and 3 is shown in Fig. 5. It can be observed that, for each patient, both categories of methods exhibit some common error peaks, such as window 5 of the signal for patient 2, which corresponds to the highest error for both types of algorithm. Likewise, window index 0 of the same signal yields the lowest MAE for both method categories. Furthermore, unsupervised methods exhibit lower MAE than supervised methods in most windows for both patients.

B. Respiratory rate estimation

Table V reports the MAE, RMSE and MAPE of respiratory rate estimates derived from rPPG signals, in comparison to estimates obtained from the reference respiratory signal. Among the unsupervised methods, POS and GREEN achieved the best

TABLE IV
HEART RATE ESTIMATION METRICS OBTAINED BY EACH RPPG METHOD USING ECG-DERIVED HEART RATE AS GROUND TRUTH.

Method	Patient 1				Patient 2				Patient 3			
	MAE	RMSE	MAPE	SNR	MAE	RMSE	MAPE	SNR	MAE	RMSE	MAPE	SNR
Unsupervised Methods												
CHROM	**16.93**	**19.37**	**16.12**	-9.80	**0.89**	**1.06**	**1.15**	0.14	3.65	6.34	4.22	0.07
GREEN	25.89	26.13	24.75	-9.40	11.60	14.99	14.93	-8.39	24.60	27.50	27.02	-10.80
ICA	39.70	41.61	38.24	-11.06	**0.89**	**1.06**	**1.15**	2.66	18.56	21.59	20.30	-7.68
LGI	26.31	26.50	25.15	-9.52	4.06	5.49	5.20	-3.92	**2.59**	**4.99**	**3.02**	-2.51
OMIT	26.31	26.50	25.15	-9.54	4.06	5.49	5.20	-3.79	4.49	7.99	5.11	-4.03
PBV	26.31	26.50	25.15	-9.53	1.29	1.51	1.67	-0.45	2.80	5.05	3.25	-1.41
POS	28.82	34.73	27.83	-9.74	**0.89**	**1.06**	**1.15**	**2.83**	**2.59**	**4.99**	**3.02**	**1.51**
Supervised Methods												
PhysFormer (PURE)	**16.97**	**26.69**	**16.43**	**-6.07**	5.68	9.53	7.21	-3.85	8.25	15.92	9.08	-2.56
PhysFormer (UBFC)	40.54	41.43	38.67	-7.96	3.09	5.87	3.91	-2.61	8.53	14.37	9.42	-1.96
EfficientPhys (PURE)	45.47	46.69	43.43	-10.40	6.44	10.87	8.18	-3.10	6.82	13.82	7.52	**-0.87**
EfficientPhys (UBFC)	48.49	49.84	46.33	-8.49	8.27	12.05	10.52	-3.72	6.82	13.82	7.52	-1.61
PhysNet (PURE)	33.15	36.16	31.72	-11.06	5.09	8.34	6.61	-2.80	**1.20**	**1.87**	**1.38**	-1.66
PhysNet (UBFC)	26.20	29.30	25.10	-8.59	7.78	11.48	9.89	-6.01	17.86	23.15	19.59	-5.36
TSCAN (PURE)	50.16	53.07	47.87	-8.56	**1.42**	**2.05**	**1.82**	**-1.25**	17.62	24.35	19.47	-4.96
TSCAN (UBFC)	33.15	36.16	31.72	-9.69	3.93	8.05	4.96	-2.87	16.61	22.73	18.59	-3.96
DeepPhys (PURE)	51.00	51.35	48.77	-11.68	13.56	19.26	17.37	-3.50	29.52	34.72	32.99	-6.25
DeepPhys (UBFC)	50.16	51.83	47.81	-10.24	10.79	16.05	13.74	-2.71	18.72	25.62	21.06	-3.07

Terms in parenthesis indicate the dataset on which the models were trained. MAE and RMSE are in bpm, MAPE is in % and SNR is in dB.

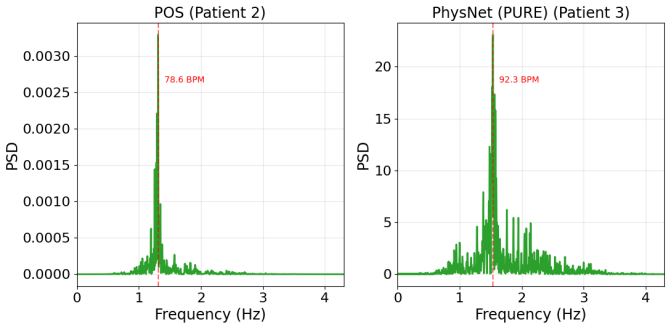


Fig. 4. PSD of the signals which achieved the best heart rate estimation error results. On the left, the PSD of the signal extracted with the POS method from the video of patient 2. On the right, the PSD of the signal extracted with the PhysNet model trained on the PURE dataset from the video of patient 3.

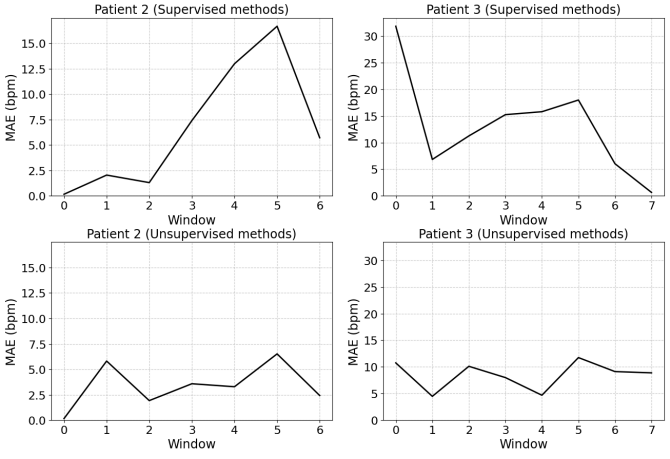


Fig. 5. MAE per window for unsupervised methods (below) and for supervised methods (above) for patients 2 (left) and 3 (right). It can be observed that, for each patient, both categories of methods exhibit some common error peaks and minima, suggesting similar patterns of error variation.

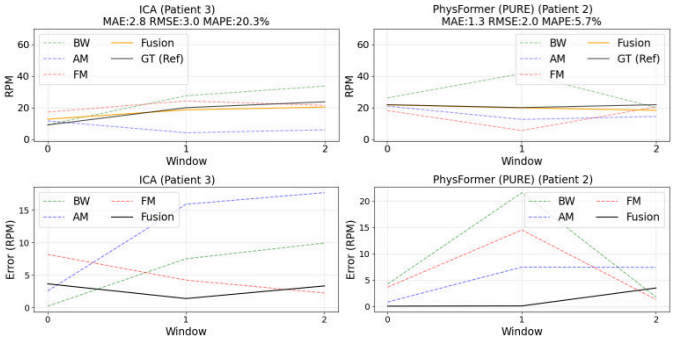


Fig. 6. Respiratory rate estimates (above) and the corresponding MAE per window (below) for the methods which achieved the best performance among unsupervised and supervised approaches: ICA for patient 3 (left) and PhysFormer trained on PURE for patient 2 (right). The values obtained from the BW, AM and FM components are shown alongside the estimates obtained from the fusion strategy. It can be observed how estimates obtained from the fusion strategy result in a lower MAE when compared to the estimates obtained from the respiratory components alone for most windows.

performance for patient 1, PBV for patient 2, and CHROM for patient 3. For the supervised approaches, PhysNet trained on the PURE dataset yielded the lowest MAE for patient 1, while PhysFormer trained on the same dataset achieved the best performance for patient 2, and EfficientPhys trained on PURE produced the best results for patient 3.

Fig. 6 presents the error per window and the RR estimates per window obtained from the AM, BW and FM components and the fusion between them, for the methods which achieved the best performance among unsupervised and supervised approaches across all patients. The figure shows how estimates obtained from the fusion strategy result in a lower MAE when compared to the estimates obtained from the respiratory components alone.

TABLE V
RESPIRATORY RATE ESTIMATION METRICS OBTAINED BY EACH RPPG METHOD.

Method	Patient 1			Patient 2			Patient 3		
	MAE	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**
Unsupervised Methods									
CHROM	12.09	12.27	47.17	5.95	7.06	27.25	**4.10**	**5.10**	**19.65**
GREEN	**8.34**	**8.98**	**32.55**	9.07	9.53	42.94	7.85	9.29	37.60
ICA	8.55	8.88	33.36	5.23	6.23	23.85	5.14	5.58	28.07
LGI	11.68	11.81	45.55	5.05	6.28	23.51	9.93	10.13	61.44
OMIT	11.68	11.75	45.55	6.99	8.27	33.20	9.31	9.60	59.57
PBV	10.43	10.43	40.67	**4.18**	**4.48**	**19.71**	10.14	11.51	51.74
POS	**8.34**	**8.98**	**32.55**	5.02	5.57	23.16	8.89	10.32	43.85
Supervised Methods									
DeepPhys (PURE)	9.59	10.05	37.42	6.36	7.43	29.23	8.68	9.41	52.43
DeepPhys (UBFC)	10.63	12.56	41.49	2.76	3.00	12.73	9.10	9.59	53.13
EfficientPhys (PURE)	11.88	12.09	46.36	**1.15**	1.58	5.62	**2.64**	**3.14**	**12.57**
EfficientPhys (UBFC)	7.51	7.66	29.30	3.77	4.66	18.16	6.87	8.49	31.70
PhysFormer (PURE)	9.59	10.18	37.42	**1.15**	**1.38**	**5.27**	6.81	7.37	38.95
PhysFormer (UBFC)	9.80	10.55	38.23	3.56	4.29	16.27	6.50	7.57	46.30
PhysNet (PURE)	**5.84**	**7.28**	**22.79**	5.26	6.55	24.63	7.43	7.62	52.17
PhysNet (UBFC)	11.68	12.12	45.55	6.06	6.84	27.90	7.01	7.14	49.47
TSCAN (PURE)	9.80	10.35	38.23	7.27	9.25	33.90	8.89	9.75	54.55
TSCAN (UBFC)	6.68	7.53	26.04	8.45	8.92	39.06	7.85	8.76	43.49

Terms in parenthesis indicate the dataset on which the models were trained. MAE and RMSE are in rpm, MAPE is in % and SNR is in dB.

V. DISCUSSION AND CONCLUSION

The extraction of remote photoplethysmography signals from videos can be based on unsupervised algorithms or on DL-based models. These approaches result in a blood volume pulse waveform that shows significant similarity to the conventional photoplethysmographic signal. The BVP obtained from rPPG methods can be used to estimate heart rate and respiratory rate. Nonetheless, illumination and movement artifacts can greatly impact on signal quality, as could be observed by comparing HR estimates per window, in which windows with greater error values can be related to abrupt patient movement or expressive lighting variations, affecting signal extraction for all methods similarly. Moreover, video compression artifacts can also negatively impact the SNR, which motivated the use of an industrial video camera in this work. In the same way, camera exposure time and image resolution can influence the obtained results and the different values set for each patient recording was observed as one of the limitations of this work.

Definition and detection of the ROI also play an important role in remote vital signs estimation, which can be observed by the results obtained for patient 1. The rPPG signal extracted for this patient showed low linear correlation to the ground truth PPG signal and presented the highest error metrics for HR and RR estimates. This can be due to the fact that the face detector failed to identify any face in the video frames, possibly as a result of the occlusion of the patient's face caused by the mechanical ventilator tubes. As a result, frames were introduced as input to the rPPG algorithms without any ROI segmentation, leading to SNR degradation. Moreover, it is noted that the reference PPG signal was different from the conventional expected waveform, which can be due to physiological factors related to the clinical condition of the patient and a further analysis is necessary to clarify the influence of this factor over the remotely extracted signal. Additionally, specifically for patient 3, although the RR error metrics were similar to those of the other patients, the PSD of the ground-truth respiratory signal revealed that multiple frequency components exhibited comparable power levels, resulting in an ambiguous dominant frequency estimate and, therefore, an ambiguous ground-truth RR value, which can be due to bad signal quality.

Finally, the error metrics for HR and RR estimates for all patients show that evaluating different methods for rPPG signal extraction in an ICU within SUS depends on an effective face detection algorithm in order to account for face occlusion cases common to this context, alongside an evaluation of the quality of reference signals so that comparisons to the remotely extracted signal yield reliable results. Furthermore, albeit the small initial cohort, the dataset used represents a clinical environment and patient population not yet explored in the rPPG literature. In particular, SUS serves a large and socioeconomically and demographically diverse population in Brazil, which may provide an interesting setting for investigating the robustness and capacity of generalization of rPPG methods beyond the controlled environments and populations commonly represented in existing datasets. Overall, these findings demonstrate the feasibility of such evaluations and provide methodological insights for future data acquisition, supporting larger-scale assessments of rPPG performance and applicability in this context.

VI. FUTURE WORK

The main limitation of this work is the limited data. For this reason, future work will include the acquisition of more videos from the ICU with standardized image resolution and exposure time, including diverse lighting conditions and patient clinical and demographic profiles representative of the Brazilian public

health context, providing analyses of the assessed algorithms by patient skin type, age, gender or clinical condition. Furthermore, future work will also include making the additional data acquired in this setting available as an open-source dataset, in order to contribute for data diversity in future rPPG research. Lastly, future video-processing will consider a more robust face detector, in order to address face occlusions caused by medical devices commonly used in intensive care settings, such as mechanical ventilators.

ACKNOWLEDGMENT

The authors thank the University Hospital of the University of São Paulo for all the support regarding data acquisition.

REFERENCES

[1] M. Elliott and A. Coventry, "Critical care: the eight vital signs of patient monitoring," *British Journal of Nursing*, vol. 21, no. 10, pp. 621–625, 2012.

[2] J. Ede, S. Vollam, J. L. Darbyshire, O. Gibson, L. Tarassenko, and P. Watkinson, "Non-contact vital sign monitoring of patients in an intensive care unit: A human factors analysis of staff expectations," *Applied Ergonomics*, vol. 90, p. 103149, 2021.

[3] F.-T.-Z. Khanam, A. Al-Naji, and J. Chahl, "Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review," *Applied Sciences*, vol. 9, no. 20, p. 4474, 2019.

[4] V. Selvaraju, N. Spicher, J. Wang, N. Ganapathy, J. M. Warnecke, S. Leonhardt, R. Swaminathan, and T. M. Deserno, "Continuous monitoring of vital signs using cameras: A systematic review," *Sensors*, vol. 22, no. 11, p. 4097, 2022.

[5] D. J. McDuff, J. R. Estepp, A. M. Piasecki, and E. B. Blackford, "A survey of remote optical photoplethysmographic imaging methods," in *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2015, pp. 6398–6404.

[6] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 284–285.

[7] M. J. Bautista, M. Kowal, D. G. Cave, C. Downey, and D. G. Jayne, "Clinical applications of contactless photoplethysmography for monitoring in adults: A systematic review and meta-analysis," *Journal of Clinical and Translational Science*, vol. 7, no. 1, p. e129, 2023.

[8] M. Harford, J. Catherall, S. Gerry, J. D. Young, and P. Watkinson, "Availability and performance of image-based, non-contact methods of monitoring heart rate, blood pressure, respiratory rate, and oxygen saturation: a systematic review," *Physiological measurement*, vol. 40, no. 6, p. 06TR01, 2019.

[9] S. S. Khan and N. Mostafa, "Representation biases in remote photoplethysmography datasets–a narrative review," *Applied Computing and Informatics*, pp. 1–16, 2025.

[10] R. J. Lee, S. Sivakumar, and K. H. Lim, "Review on remote heart rate measurements using photoplethysmography," *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 44 699–44 728, 2024.

[11] W. Chen, Z. Yi, L. J. R. Lim, R. Q. R. Lim, A. Zhang, Z. Qian, J. Huang, J. He, and B. Liu, "Deep learning and remote photoplethysmography powered advancements in contactless physiological measurement," *Frontiers in bioengineering and biotechnology*, vol. 12, p. 1420100, 2024.

[12] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[13] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[14] M. Lewandowska and J. Nowak, "Measuring pulse rate with a webcam," *Journal of Medical Imaging and Health Informatics*, vol. 2, no. 1, pp. 87–92, 2012.

[15] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE transactions on biomedical engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[16] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, no. 9, pp. 1913–1926, 2014.

[17] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE transactions on biomedical engineering*, vol. 63, no. 9, pp. 1974–1984, 2015.

[18] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[19] C. S. Pilz, S. Zaunseder, J. Krajewski, and V. Blazek, "Local group invariance for heart rate estimation from face videos in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1254–1262.

[20] C. A. Casado and M. B. López, "Face2ppg: An unsupervised pipeline for blood volume pulse extraction from faces," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 11, pp. 5530–5541, 2023.

[21] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 349–365.

[22] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 400–19 411, 2020.

[23] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *arXiv preprint arXiv:1905.02419*, 2019.

[24] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4186–4196.

[25] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 5008–5017.

[26] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, Y. Wang, S. Sengupta, S. Patel, and D. McDuff, "rppg-toolbox: Deep remote ppg toolbox," *arXiv preprint arXiv:2210.00716*, 2022.

[27] D. Qi, W. Tan, Q. Yao, and J. Liu, "Yolo5face: Why reinventing a face detector," in *European Conference on Computer Vision*. Springer, 2022, pp. 228–244.

[28] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE, 2014, pp. 1056–1062.

[29] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern recognition letters*, vol. 124, pp. 82–90, 2019.

[30] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on biomedical engineering*, vol. 60, no. 7, pp. 1946–1953, 2013.

[31] R. D. Berger, S. Akselrod, D. Gordon, and R. J. Cohen, "An efficient algorithm for spectral analysis of heart rate variability," *IEEE Transactions on biomedical engineering*, no. 9, pp. 900–904, 1986.

[32] P. H. Charlton, T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale, and P. J. Watkinson, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological measurement*, vol. 37, no. 4, pp. 610–626, 2016.

[33] D. Luguern, S. Perche, Y. Benezeth, V. Moser, L. A. Dunbar, F. Braun, A. Lemkaddem, K. Nakamura, R. Gomez, and J. Dubois, "An assessment of algorithms to estimate respiratory rate from the remote photoplethysmogram," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 304–305.

[34] P. van Gent, H. Farah, N. Nes, and B. Arem, "Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data," in *Proceedings of the 6th HUMANIST Conference*, 2018.

Impacto do *Slicing Aided Hyper Inference* (SAHI) na detecção de lesões de retinopatia diabética usando YOLO

Cezário Luiz Mourão C. Mesquita
dept. engenharia de teleinformática
Universidade Federal do Ceará
cezariomourao@alu.ufc.br

José Valclemir Rodrigues da Silva
dept. engenharia de teleinformática
Universidade Federal do Ceará
valclemir15@hotmail.com

Paulo César Cortez
dept. engenharia de teleinformática
Universidade Federal do Ceará
cortez@lesc.ufc.br

***Resumo*—A retinopatia diabética (RD) é uma das principais causas de perda de visão evitável no mundo e se manifesta, em seus estágios iniciais, por lesões de fundo de olho como microaneurismas, exsudatos duros, exsudatos moles e hemorragias. Sua detecção automatizada é desafiadora, pois as imagens têm alta resolução e as lesões ocupam pouquíssimos *pixels*, sendo os microaneurismas, sinal mais precoce da doença, os mais difíceis. Este trabalho investiga, de forma controlada e estratificada por tipo de lesão, se a inferência assistida por fracionamento de imagem *Slicing Aided Hyper Inference* (SAHI) melhora a detecção de lesões de RD quando aplicada a um detector YOLO moderno, no conjunto de dados público DDR. Sob validação cruzada repetida e estratificada por lesão (5 *folds* $\times$ 3 repetições) e protocolo unificado de avaliação (*mean Average Precision*, mAP, a IoU 0,1 e 0,5), comparam-se três configurações: (A) treino e inferência em imagem cheia; (B) SAHI apenas na inferência; e (C) SAHI também no ajuste fino. A Config. B *reduziu* significativamente o desempenho global em relação ao *baseline* YOLOv8m ($p < 0{,}01$, Wilcoxon pareado), enquanto a Config. C aproximou o mAP global do *baseline*, sem diferença detectável sob o teste corrigido para a não-independência dos *folds*. O ganho concentrou-se nos microaneurismas: a Config. C elevou significativamente o AP@0,1 ($p_{\mathrm{Holm}} = 0{,}009$; $r = 0{,}82$) e aumentou o AP@0,5 de 0,059 para 0,111, ao custo de piora nas lesões maiores e difusas. Conclui-se que, para detectores YOLO modernos em alta resolução, o ganho do SAHI é localizado: não melhora o mAP global, mostra-se eficaz na detecção de microaneurismas no DDR.**

***Palavras-chave*—retinopatia diabética, detecção de lesões, YOLO, SAHI.**

I. Introdução

A retinopatia diabética (RD) é uma complicação microvascular da retina e uma das principais causas de perda irreversível de visão em pessoas com diabetes [1], [2]. No Brasil, mais de 13 milhões de pessoas vivem com diabetes, cerca de 6,9% da população [3], e a RD atinge entre 24% e 39% desse grupo [4]. Quando diagnosticada e tratada precocemente, o risco de cegueira pode ser reduzido para menos de 5% [4].

A RD pode ser caracterizada pela presença de determinados tipos de lesões na retina, tais como exsudatos duros (EX), exsudatos moles ou algodonosos (SE), microaneurismas (MA) e hemorragias (HE) [2]. Os microaneurismas, em particular, costumam ser as primeiras alterações visíveis associadas à doença e, ao mesmo tempo, as mais difíceis de detectar, por ocuparem pouquíssimos *pixels* nas imagens de fundo de olho. A identificação dessas lesões em fases iniciais é essencial para prevenir a progressão da doença a estágios mais graves e a consequente perda de visão.

Os exames que identificam essas lesões dependem de médicos especializados, cuja disponibilidade é limitada, o que pode retardar o diagnóstico, sobretudo em regiões remotas e de difícil acesso [5]. A análise humana está ainda sujeita a fatores subjetivos, como cansaço e nível de experiência do especialista [6], o que compromete a precisão justamente nos estágios iniciais, quando as alterações são sutis. Nesse cenário, a detecção automática por aprendizado profundo tem sido proposta como ferramenta de apoio ao especialista: detectar e classificar cada lesão, e não apenas graduar a imagem, oferece informação para o diagnóstico precoce, de forma mais rápida e menos suscetível à subjetividade do avaliador.

Automatizar essa tarefa, contudo, permanece um problema em aberto. No conjunto de dados público DDR (*Dataset for Diabetic Retinopathy*) [1], um dos maiores *benchmarks* da área, os detectores genéricos avaliados pelos próprios autores obtiveram mAP inferior a $0{,}01$ a IoU 0,5, com AP de microaneurismas abaixo de 1% [1], [7]. Esse desempenho reduzido decorre do tamanho e do formato das lesões, do ruído e do contraste das imagens, do reduzido número de exemplos anotados e da dificuldade dos algoritmos de aprendizado profundo em detectar objetos muito pequenos [5].

Para mitigar o problema dos objetos minúsculos, a literatura recente tem explorado estratégias diversas: realçar seletivamente as regiões mais informativas da imagem [8], [9], ampliar artificialmente sua resolução [10] ou modelar relações entre regiões distantes, em vez de apenas vizinhanças locais [11]. Uma alternativa independente da arquitetura do detector consiste em fracionar a imagem em janelas sobrepostas (*tiling*/*slicing*), de modo que cada janela seja processada em resolução efetiva mais alta. O *framework* SAHI (*Slicing Aided Hyper Inference*) [12] implementa essa ideia de forma

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob o Processo nº 315447/2023-0, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Código de Financiamento 001, pelo apoio a esta pesquisa.

genérica, aplicável na inferência e, opcionalmente, também no ajuste fino do detector. No contexto do DDR, trabalhos anteriores demonstraram que o fracionamento eleva o desempenho da detecção de lesões, seja por *tiling* convencional [5], seja pelo SAHI [7]. Entretanto, esses estudos empregaram arquiteturas YOLO de gerações anteriores (YOLOv4/v5, YOLOR-CSP), permanecendo uma lacuna de investigação do SAHI combinado a detectores YOLO mais recentes, sob um protocolo de comparação controlado e com análise estratificada por tipo de lesão.

Diante dessa lacuna, este trabalho avalia, de forma controlada e estratificada por tipo de lesão, se o SAHI melhora a detecção de lesões de RD quando aplicado a um detector YOLO moderno (YOLOv8m), em comparação com a inferência em imagem completa, utilizando o conjunto de dados DDR. As principais contribuições deste trabalho são:

- o emprego de um detector YOLO moderno (YOLOv8m) combinado ao SAHI no DDR;
- uma comparação *controlada* entre YOLO e YOLO+SAHI sob um protocolo unificado de pré-processamento, resolução, validação cruzada e avaliação, que isola o efeito do *slicing* (mesmo detector, mesmos *folds*);
- uma análise *estratificada por classe* de lesão (MA, EX, SE e HE), com teste estatístico pareado, que evidencia o *trade-off* do *slicing*, ganho nas menores lesões (microaneurismas) e perda nas maiores e mais difusas (exsudatos moles e hemorragias), demonstrando que seu benefício sobre um YOLO moderno em alta resolução é condicional à classe de lesão e ao regime de treino.

O restante do artigo apresenta a fundamentação teórica (Seção II), os trabalhos relacionados (Seção III), os materiais e métodos (Seção IV), os resultados e a discussão (Seção V) e as considerações finais (Seção VI).

II. Fundamentação Teórica

A. Detecção de objetos

A detecção de objetos consiste em delimitar cada instância presente na imagem por uma *bounding box* (caixa retangular que circunscreve o objeto) e atribuir-lhe uma classe [13]. Os detectores baseados em aprendizado profundo dividem-se em detectores de dois estágios, que primeiro propõem regiões de interesse e depois as classificam, como o Faster R-CNN [14], e detectores de estágio único, que tratam a detecção como um problema de regressão e predizem simultaneamente as caixas e as probabilidades de classe sobre a imagem completa, como o SSD [15] e a família YOLO (*You Only Look Once*) [16]. A principal vantagem dos detectores de estágio único é a velocidade de inferência, o que os torna adequados a aplicações em tempo real e a *hardware* de menor custo.

B. A arquitetura YOLOv8

Neste trabalho adota-se o YOLOv8 [17], na variante `m` (*medium*). Trata-se de um detector *anchor-free* com *backbone* do tipo CSPDarknet, *neck* PAN-FPN e cabeça de detecção desacoplada, que realiza a predição em três escalas (P3/P4/P5) com cabeças especializadas em objetos pequenos, médios e grandes, característica relevante para a detecção de lesões de tamanhos variados na retina. A variante `m` equilibra capacidade e custo computacional, sendo compatível com o treino em uma única GPU.

C. Slicing Aided Hyper Inference (SAHI)

O SAHI, proposto por Akyon *et al.* [12], é uma solução de código aberto que fornece um processo genérico de inferência e ajuste fino auxiliado por fracionamento. A imagem de entrada é dividida em janelas (*tiles*) sobrepostas; a detecção é realizada em cada janela e as predições são reagrupadas na imagem original por pós-processamento (por exemplo, supressão não máxima gulosa, *GREEDYNMM*). Ao processar regiões menores em resolução efetiva mais alta, o método amplia a visibilidade de objetos pequenos, sendo promissor para o problema dos microaneurismas. Pode ser aplicado apenas na inferência (*sliced inference*) ou também no treinamento (*slicing-aided fine-tuning*), regime no qual a literatura reporta ganhos maiores [12].

III. Trabalhos Relacionados

A. Evolução da família YOLO

A família YOLO evoluiu por sucessivas versões (YOLOv3, YOLOv4, YOLOv5, YOLOR, YOLOv7, YOLOv8, YOLOv11, entre outras), incorporando aprimoramentos de *backbone*, *neck* e *head*, além de técnicas de aumento de dados e funções de perda mais robustas. Essa evolução importa aqui porque os trabalhos que aplicaram *slicing* à detecção de lesões de RD empregaram gerações anteriores da família [5], [7], cujas limitações na detecção de objetos pequenos podem ter sido atenuadas nas versões recentes.

B. Detecção de lesões de RD no DDR

No estudo que apresentou o conjunto de dados, Li *et al.* [1] avaliaram detectores genéricos (SSD, YOLO e Faster R-CNN) e demonstraram a dificuldade central da tarefa: enquanto a classificação alcançou acurácia da ordem de 0,83 com a arquitetura SE-BN-Inception, os modelos de segmentação e de detecção tiveram desempenho fraco, com AP de microaneurismas abaixo de 1%, o que evidencia a necessidade de aprimorar o reconhecimento de microlesões para viabilizar a aplicação clínica. Alyoubi *et al.* [18] combinaram uma rede de classificação (CNN512) a um YOLOv3 para localização de lesões, obtendo mAP de 0,216 na detecção no DDR.

Uma linha de trabalhos correlatos explorou progressivamente a família YOLO nessa base. Santos *et al.* [19] propuseram uma abordagem baseada em YOLOv4 e, depois [5], um método baseado em YOLOv5 com pré-processamento, aumento de dados e *Tiling*, alcançando mAP de 0,2630 na validação e 0,1540 no teste do DDR (IoU 0,5). Em seguida, Pereira *et al.* [7] combinaram a arquitetura YOLOR-CSP ao *framework* SAHI, atingindo mAP de 0,3808 na *validação* do DDR (IoU 0,5), mas de apenas 0,2225 no *teste*; essa diferença de 0,16 no mesmo modelo ilustra o quanto números de um único *split* de validação podem ser otimistas.

C. Lacuna de pesquisa

Como a combinação YOLO+SAHI já foi aplicada ao DDR [7], a comparação direta "YOLO vs. YOLO+SAHI" não constitui, por si só, uma novidade. O que permanece em aberto, e é o objeto deste trabalho, é saber se o ganho atribuído ao *slicing* persiste sobre um detector YOLO moderno, quando medido sob protocolo unificado, sem o viés de *split* único e de forma estratificada por tipo de lesão.

IV. Materiais e Métodos

A. Visão geral do método

O fluxo experimental é ilustrado na Fig. 1. Não há aquisição própria de imagens: utilizam-se as 757 imagens do conjunto público DDR [1] que possuem anotações de lesão em nível de *bounding box* (Seção IV-B). Essas anotações, distribuídas em formato Pascal VOC, são convertidas para o formato YOLO, com deduplicação das caixas repetidas e unificação dos três *splits* oficiais em um único *pool* (Seção IV-C), sobre o qual se aplica validação cruzada repetida e estratificada pelo vetor de presença de lesão, tendo a imagem como unidade de divisão (Seção IV-E). Em cada *fold*, o mesmo detector YOLOv8m é treinado sob dois domínios de entrada, imagem completa em alta resolução (Configs. A e B) e *tiles* de 512×512 (Config. C, *slicing-aided fine-tuning*), com hiperparâmetros fixos, de modo que a única variável experimental seja o fracionamento (Seção IV-F). A inferência ocorre sobre a imagem completa (Config. A) ou de forma fracionada com o SAHI (Configs. B e C), caso em que as predições de cada *tile* são reagrupadas nas coordenadas da imagem original por supressão não máxima gulosa. Por fim, as predições são confrontadas com o *Ground Truth* pelo avaliador COCO, calculando-se o mAP a IoU 0,1 e 0,5, global e por classe (Seção IV-G); como as três configurações compartilham os mesmos *folds*, os resultados são pareados e comparados por testes estatísticos pareados (Seção IV-H).

B. Conjunto de dados

Os experimentos utilizam o conjunto de dados público DDR [1], também chamado OIA-DDR, composto por 13.673 imagens de fundo de olho de 9.598 pacientes, coletadas em 147 hospitais de 23 províncias da China; destas, apenas 757 possuem anotações de detecção em nível de *bounding box*, sendo este o subconjunto empregado neste trabalho. As quatro classes consideradas são microaneurismas (MA), exsudatos duros (EX), exsudatos moles (SE) e hemorragias (HE). As imagens têm alta resolução e tamanhos variáveis (da ordem de 1088×1920 a 3456×5184 *pixels*), enquanto as lesões ocupam regiões diminutas, o que constitui a principal motivação para investigar o *slicing*. As anotações são bastante desbalanceadas, com SE presente em apenas 239 imagens e 1.293 instâncias, contra 23.659 instâncias de EX, conforme resumido na Tabela I. A densidade das microlesões e a grade de *tiles* do SAHI são ilustradas, sobre uma imagem real do DDR, na Fig. 2.

Tabela I
Distribuição das anotações de detecção do DDR (subconjunto de 757 imagens anotadas), após a deduplicação descrita na Seção IV-C.

Lesão	Imagens com a lesão	Instâncias	% instâncias
MA (microaneurismas)	570	10.380	21,7%
EX (exsudatos duros)	486	23.659	49,4%
SE (exsudatos moles)	239	1.293	2,7%
HE (hemorragias)	601	12.538	26,2%
Total	757	47.870	100%

Tabela II
Configurações experimentais avaliadas.

Config.	Treino	Inferência	Propósito
A	imagem cheia	imagem cheia	*baseline* sem *slicing*
B	imagem cheia	*sliced* (SAHI)	efeito do *slicing* só na inferência
C	*tiles* (*slicing-aided FT*)	*sliced* (SAHI)	regime completo de *slicing*

C. Conversão de anotações para o formato YOLO

O DDR fornece as *bounding boxes* oficiais em arquivos Pascal VOC (XML), organizadas nos *splits* `train`/`valid`/`test`. A conversão adotada parseia esses XML e escreve as caixas no formato YOLO (`classe cx cy w h`, com coordenadas normalizadas em $[0, 1]$), unindo os três *splits* em um único *pool* de 757 imagens, de modo que a divisão entre treino, validação e teste passa a ser feita pela validação cruzada (Seção IV-E). Verificou-se que os XML contêm caixas duplicadas exatas (fator próximo de $1{,}7\times$); aplica-se, portanto, deduplicação por caixa, resultando nas 47.870 instâncias da Tabela I. Após a conversão, as caixas foram validadas visualmente sobre uma amostra de imagens, para descartar erros de eixo ou de normalização. Na avaliação, adotou-se um lado mínimo de caixa de 2 *pixels*, o que evita áreas nulas no reagrupamento geométrico do SAHI.

D. Arquitetura de experimentos

Para isolar o efeito do *slicing*, três configurações são avaliadas sob o mesmo protocolo, conforme a Tabela II.

A configuração A é o *baseline* (sem *slicing*); a B isola o efeito do SAHI aplicado apenas na inferência, sobre o *mesmo* modelo treinado em imagem completa; e a C corresponde ao regime completo, em que o modelo também é treinado sobre *tiles*. As Configs. A e B derivam de um *único* treino por *fold*, avaliado de duas formas (imagem cheia e SAHI), o que torna a comparação A vs. B um teste exato do efeito da inferência fracionada.

E. Validação cruzada

Adota-se validação cruzada repetida e estratificada, com a imagem como unidade de divisão (as 757 anotadas), de modo que uma mesma imagem nunca apareça em mais de um

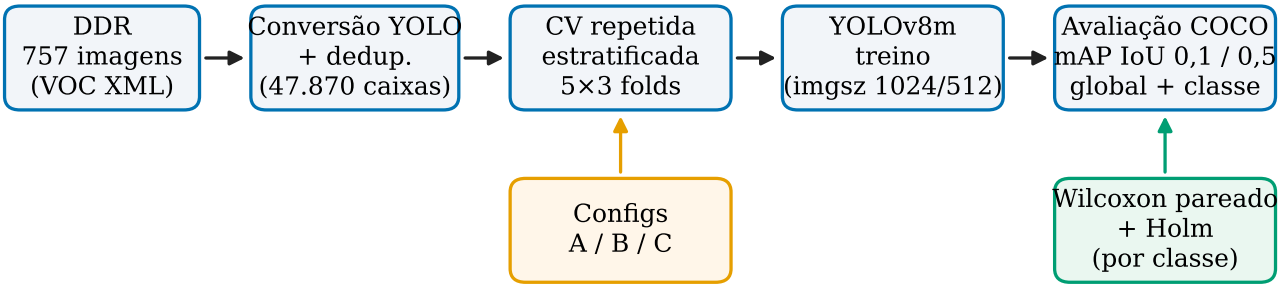


Figura 1. Visão geral do *pipeline* proposto: conversão das anotações VOC do DDR para o formato YOLO (com deduplicação), validação cruzada repetida e estratificada por lesão, treino do YOLOv8m, avaliação COCO (mAP a IoU 0,1 e 0,5, global e por classe) e teste estatístico pareado. As três configurações (A/B/C) compartilham os mesmos *folds*.

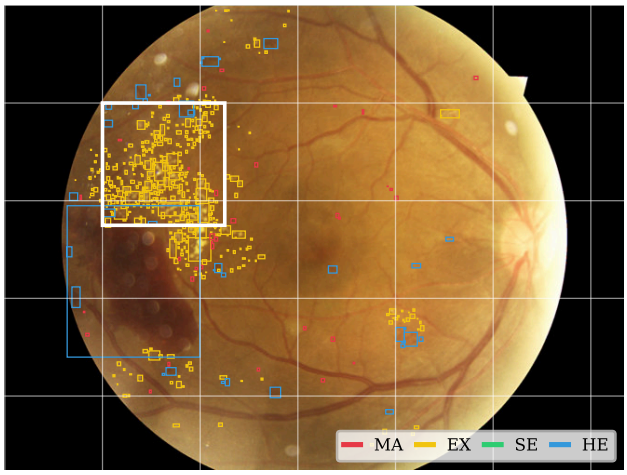


Figura 2. Imagem do DDR com as *bounding boxes* do *Ground Truth* (cores por classe) e a grade de *tiles* do SAHI (512 × 512, sobreposição 0,2); em destaque, um *tile*. A alta densidade de microlesões diminutas, muitas menores que a espessura da linha das caixas, motiva a inferência assistida por fracionamento.

subconjunto, evitando vazamento de dados. Como o DDR não disponibiliza identificadores de paciente para o subconjunto anotado, a partição é por imagem, não por paciente: imagens de um mesmo paciente podem, em princípio, ocorrer em *folds* distintos. Estimativas de partição única são instáveis em conjuntos pequenos, e a repetição reduz essa variância ao promediar múltiplas partições [20], [21]; no próprio DDR, Pereira *et al.* [7] reportam diferença de 0,16 de mAP entre validação e teste para o mesmo modelo, magnitude superior a qualquer efeito que se pretenda medir entre configurações.

Utiliza-se `RepeatedMultilabelStratifiedKFold` (5 *folds* × 3 repetições = 15 avaliações), estratificado pelo vetor binário de presença de lesão (tem MA / tem EX / tem SE / tem HE) de cada imagem [22]. Essa estratégia é essencial dado o desbalanceamento, sobretudo para SE, presente em 239 das 757 imagens: sob partição aleatória, os *folds* teriam contagens díspares dessa classe, inflando a variância de seu AP. Em cada *fold* externo, o conjunto de treino é novamente dividido de forma estratificada, reservando-se 80% para o ajuste dos pesos e 20% para a validação interna do *early-stopping*.

Como detector, hiperparâmetros (Tabela III), *seed*, *folds* e protocolo de avaliação são compartilhados, qualquer diferença entre A, B e C é atribuível ao fracionamento. O desempenho é reportado como média ± desvio-padrão do mAP entre os 15 *folds*.

F. Modelos e hiperparâmetros

O detector YOLOv8m foi implementado com a biblioteca `ultralytics` e o *slicing* com a biblioteca `sahi`, em Python. No *baseline* (Configs. A/B), as imagens são processadas em resolução elevada (lado maior redimensionado para 1024), a fim de preservar as lesões pequenas; nas inferências com SAHI, utilizam-se *tiles* de 512 × 512 com sobreposição de 0,2. Na Config. C, o modelo é treinado diretamente sobre *tiles* de 512 *px*, com *batch* maior, dado o menor consumo de memória por *tile*; treino e validação interna operam no domínio de *tiles*, para que o *early-stopping* reflita o mesmo domínio da inferência SAHI. Para preservar a morfologia das microlesões, emprega-se aumento de dados moderado e desativa-se o *mosaic*, que tende a prejudicar objetos minúsculos. A Tabela III lista os hiperparâmetros finais.

G. Métricas e protocolo de avaliação

Para comparar os modelos utiliza-se a *Intersection over Union* (IoU), que mede a sobreposição entre a *bounding box* predita e a do *Ground Truth*, conforme a Eq. (1):

$$\mathrm{IoU} = \frac{\mathrm{Area}\,(gt \cap pd)}{\mathrm{Area}\,(gt \cup pd)}, \tag{1}$$

em que gt é a caixa do *Ground Truth* e pd a caixa predita. A *Average Precision* (AP) é a área sob a curva de Precisão×Revocação de cada classe, e o *mean Average Precision* (mAP), a média das AP entre as classes, conforme a Eq. (2):

Tabela III
HIPERPARÂMETROS FIXOS (IDÊNTICOS ENTRE CONFIGURAÇÕES, SALVO OS *overrides* DE TREINO DA CONFIG. C).

Parâmetro	**Valor**
Detector / variante	YOLOv8m (`ultralytics`)
Otimizador / lr_0	SGD / 0,01
Resolução de treino (A/B)	1024 (lado maior)
Resolução de treino (C)	*tiles* 512
Batch (A/B | C)	8 | 16
Épocas (teto) / *patience*	100 / 20 (*early-stopping*)
Mosaic	desativado
SAHI *tile* / *overlap*	512×512 / 0,2
SAHI pós-proc. / *conf*	*GREEDYNMM* / 0,1
Seed	42

Tabela IV
DESEMPENHO GLOBAL (MAP, MÉDIA $\pm$ DESVIO ENTRE OS 15 *folds*) E TESTES PAREADOS. p_{Holm}: WILCOXON COM CORREÇÃO DE HOLM; r: TAMANHO DE EFEITO; p_{NB}: TESTE-T CORRIGIDO DE NADEAU–BENGIO. MELHOR VALOR EM **NEGRITO**.

Config.	**mAP@0,1**	**mAP@0,5**
A (*baseline*)	**0,300 $\pm$ 0,019**	**0,189 $\pm$ 0,015**
B (SAHI na inferência)	0,245 $\pm$ 0,033	0,095 $\pm$ 0,019
C (SAHI + *fine-tuning*)	0,290 $\pm$ 0,016	0,159 $\pm$ 0,012
Comparações pareadas (mAP@0,1):		
A vs. B	A>B; p_{Holm}=0,002, r=0,88, p_{NB}=0,001	
A vs. C	A>C; p_{Holm}=0,005, r=0,72, p_{NB}=0,125	
B vs. C	C>B; p_{Holm}=0,002, r=0,87, p_{NB}=0,013	

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i. \tag{2}$$

O mAP é calculado com o avaliador COCO (`pycocotools`) a IoU 0,1, compatível com o protocolo de Li *et al.* [1], e também a IoU 0,5, de forma global e por classe, sempre sobre o mesmo conjunto de teste de cada *fold*; dado o grande número de lesões por imagem, o limite de detecções por imagem foi ampliado (*maxDets* até 1000). Classes ausentes do *Ground Truth* de um *fold* são excluídas da média daquele *fold*, seguindo o padrão COCO.

H. Análise estatística

As três configurações são comparadas duas a duas, nos pares A vs. B, A vs. C e B vs. C, sob a hipótese nula bilateral de ausência de diferença de desempenho. A unidade de observação é o *fold*: como as três configurações são avaliadas exatamente sobre os mesmos 15 conjuntos de teste, cada uma produz um vetor de 15 valores, pareados elemento a elemento. Para um par (X, Y), define-se o vetor de diferenças

$$d_i = m_i^X - m_i^Y, \qquad i = 1, \ldots, 15, \tag{3}$$

em que m_i é a métrica obtida no *fold* i. O pareamento elimina a variabilidade atribuível à partição, isolando o efeito do *slicing*. O procedimento é aplicado ao mAP global (três comparações) e ao AP de cada lesão (doze comparações: 3 pares $\times$ 4 classes), sempre sobre o desempenho a IoU 0,1.

Cada vetor de diferenças é submetido a dois testes complementares. O primeiro é o de Wilcoxon pareado (*signed-rank*) [23], escolhido por não pressupor normalidade das diferenças, suposição pouco verificável com apenas 15 observações; dele se reporta o valor-p e o tamanho de efeito $r = |Z|/\sqrt{N}$, com $N = 15$. Esse teste, contudo, pressupõe observações independentes, condição que a validação cruzada repetida não satisfaz, pois *folds* de repetições distintas compartilham as mesmas imagens. Tal dependência subestima a variância das diferenças e tende a produzir valores-p artificialmente baixos, e não existe variante de Wilcoxon que a corrija. Aplica-se, por isso, um segundo teste: o teste-t pareado com a correção de variância de Nadeau–Bengio [24], que adiciona ao erro-padrão um termo proporcional à razão $n_{\mathrm{teste}}/n_{\mathrm{treino}} = 0{,}25$ de cada *fold*.

Os testes cumprem papéis distintos: Wilcoxon dispensa a normalidade, mas ignora a dependência entre *folds*; Nadeau–Bengio corrige essa dependência, ao custo de assumir diferenças aproximadamente normais. Considera-se robusta apenas a diferença em que ambos concordam; quando divergem, adota-se a conclusão do teste mais conservador, situação que ocorre em A vs. C (Seção V-A). Como análise complementar, dado o poder estatístico limitado de $n = 15$, a comparação é repetida tendo a imagem como unidade de observação: o AP é calculado por imagem e, como cada uma das 757 compõe o conjunto de teste exatamente uma vez por repetição, obtêm-se 757 observações pareadas, submetidas ao mesmo teste de Wilcoxon. Por fim, aplica-se a correção de Holm–Bonferroni separadamente em cada família de hipóteses, com $\alpha = 0{,}05$, denotando-se por p_{Holm} o valor-p corrigido.

V. RESULTADOS E DISCUSSÃO

A. Desempenho global

O desempenho global das três configurações, expresso pelo mAP médio entre os 15 *folds*, é apresentado na Tabela IV e na Fig. 3. O *baseline* em imagem completa (Config. A) obteve os maiores valores em ambos os limiares. A aplicação do SAHI restrita à inferência (Config. B) resultou em desempenho inferior, contrariando os ganhos reportados na literatura para arquiteturas anteriores [7], [12]. O regime completo, com ajuste fino no domínio de *tiles* (Config. C), atenuou essa degradação e aproximou-se do *baseline* a IoU 0,1, embora permanecesse abaixo dele a IoU 0,5.

Os testes pareados sustentam essa leitura. A diferença entre as Configs. A e B é significativa e com tamanho de efeito elevado ($p_{\mathrm{Holm}} = 0{,}002$; $r = 0{,}88$), resultado corroborado pela análise complementar por imagem ($p_{\mathrm{Holm}} = 0{,}003$) e pelo teste de Nadeau–Bengio ($p_{\mathrm{NB}} = 0{,}001$): a inferência fracionada, aplicada isoladamente sobre um modelo treinado em imagem completa, degrada o desempenho global. A diferença entre as Configs. B e C também é significativa ($p_{\mathrm{Holm}} = 0{,}002$; $p_{\mathrm{NB}} = 0{,}013$), o que evidencia que o ajuste fino no domínio de *tiles* é condição necessária para compensar essa perda.

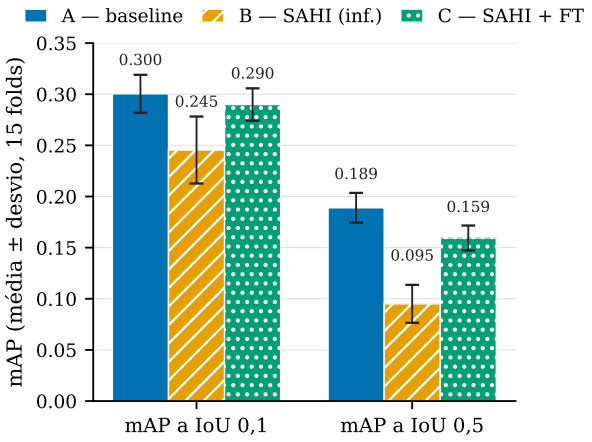


Figura 3. mAP global por configuração a IoU 0,1 e 0,5 (média ± desvio entre os 15 *folds*).

Já a comparação entre as Configs. A e C depende do teste adotado: a diferença é significativa sob Wilcoxon ($p_{\text{Holm}} = 0{,}005$), mas não sob Nadeau–Bengio ($p_{\text{NB}} = 0{,}125$), que ajusta a variância para a não-independência entre os *folds*. Pelo critério mais conservador, não há evidência suficiente para rejeitar a hipótese de ausência de diferença de mAP global entre elas. O *slicing* não produziu, portanto, nesta configuração experimental, o ganho global descrito em trabalhos com gerações anteriores de YOLO [5], [7].

B. Desempenho por classe

A estratificação por tipo de lesão revela comportamento distinto do agregado: os efeitos do *slicing* têm sinal e magnitude dependentes da classe, sendo positivos nas lesões de menor extensão espacial, conforme a Tabela V e as Figs. 4 e 5.

Os microaneurismas constituem a única classe em que uma configuração com *slicing* supera o *baseline* de forma estatisticamente significativa. A Config. C elevou o AP@0,1 de 0,177 para 0,216, incremento absoluto de 0,039 e relativo de 22% ($p_{\text{Holm}} = 0{,}009$; $r = 0{,}82$), e, a IoU 0,5, de 0,059 para 0,111. A Config. B também superou o *baseline* nessa classe (0,189 a IoU 0,1), sem atingir significância após a correção de Holm–Bonferroni. O efeito é consistente entre partições: a Config. C supera o *baseline* em 13 dos 15 *folds* (Fig. 6).

Nos exsudatos duros, o maior AP@0,1 coube à Config. B (0,404 contra 0,348 do *baseline*; $p_{\text{Holm}} = 0{,}008$); a IoU 0,5, porém, o *baseline* mantém-se superior (0,181 contra 0,108), o que sugere que o fracionamento aumenta a taxa de detecção dessa classe em detrimento da precisão de localização das caixas. Exsudatos moles e hemorragias apresentaram redução de AP sob ambas as configurações com *slicing*, comportamento analisado na Seção V-C.

C. O trade-off do slicing

A contrapartida do ganho nos microaneurismas é a perda nas lesões maiores e difusas. Exsudatos moles (SE) e hemorragias (HE), que frequentemente excedem ou cruzam as dimensões de um *tile* de 512 *px*, são fragmentados, e o reagrupamento por pós-processamento não os reconstrói integralmente. A inspeção qualitativa das predições corrobora o mecanismo, com caixas espúrias e fragmentadas concentradas nas junções dos *tiles*. O efeito é acentuado na Config. B: o AP@0,1 de SE cai de 0,354 para 0,178 ($\Delta = -0{,}176$) e o de HE cai 0,112, o que explica isoladamente a queda do mAP global dessa configuração. O ajuste fino da Config. C atenua o dano (SE recupera para 0,329; HE para 0,255), pois o modelo aprende a operar no domínio de *tiles*, mas ainda não iguala o *baseline* nessas classes (Fig. 5).

Esse *trade-off* explica por que o SAHI não melhora o mAP *global*: sendo uma média entre classes, o mAP dilui o ganho nos microaneurismas na perda em SE e HE. Na prática, o valor do *slicing* depende da métrica priorizada: quando o AP de microaneurismas é o critério de interesse, a Config. C supera o *baseline*; para uma detecção equilibrada das quatro classes, o *baseline* em alta resolução é mais robusto, um achado metodológico, não uma recomendação clínica, que exigiria validação adicional.

Em análise exploratória *post-hoc* (fora do protocolo predefinido e sem teste estatístico), a reavaliação dos pesos da Config. C com fusão IOU (em vez da IOS, padrão do SAHI) no *GREEDYNMM* eleva o mAP@0,5 de 0,159 para 0,183, próximo do *baseline* (0,189), preservando o AP@0,5 de MA ($\approx$ 0,11), indício de que a fusão imperfeita de caixas nas junções, mais que erros de detecção, responde por parte da perda a IoU alto.

O *slicing* também tem custo de inferência: cronometrado o teste de um *fold* (145 imagens; em média 34,9 *tiles* por imagem a 512/0,2), a Config. B custa 11,4× e a C 5,0× o *baseline*; a diferença entre ambas decorre da resolução de entrada herdada do treino (1024 vs. 512 *px*). Como tempos absolutos dependem do *hardware*, importam as razões: um sobrecusto de cerca de uma ordem de grandeza, a ponderar na viabilidade de rastreamento em larga escala.

D. Comparação com a literatura

Os resultados são posicionados frente à literatura do DDR na Tabela VI, com todos os valores a IoU 0,5. A comparação direta exige cautela: os trabalhos anteriores reportam um único

Tabela V
AP POR CLASSE (MÉDIA ENTRE OS 15 *folds*), A IOU 0,1 E 0,5. EM **NEGRITO**, O MELHOR VALOR POR CLASSE/IOU. †: DIFERENÇA SIGNIFICATIVA VS. A ($p_{\text{Holm}} < 0{,}05$, WILCOXON NA FAMÍLIA POR CLASSE). OS TESTES DE SIGNIFICÂNCIA FORAM CONDUZIDOS SOBRE O AP A IOU 0,1.

IoU	Config.	MA	EX	SE	HE
0,1	A	0,177	0,348	**0,354**	**0,323**
	B	0,189	**0,404**†	0,178†	0,210†
	C	**0,216**†	0,360	0,329†	0,255†
0,5	A	0,059	**0,181**	**0,308**	**0,208**
	B	0,032	0,108	0,150	0,091
	C	**0,111**	0,132	0,258	0,137

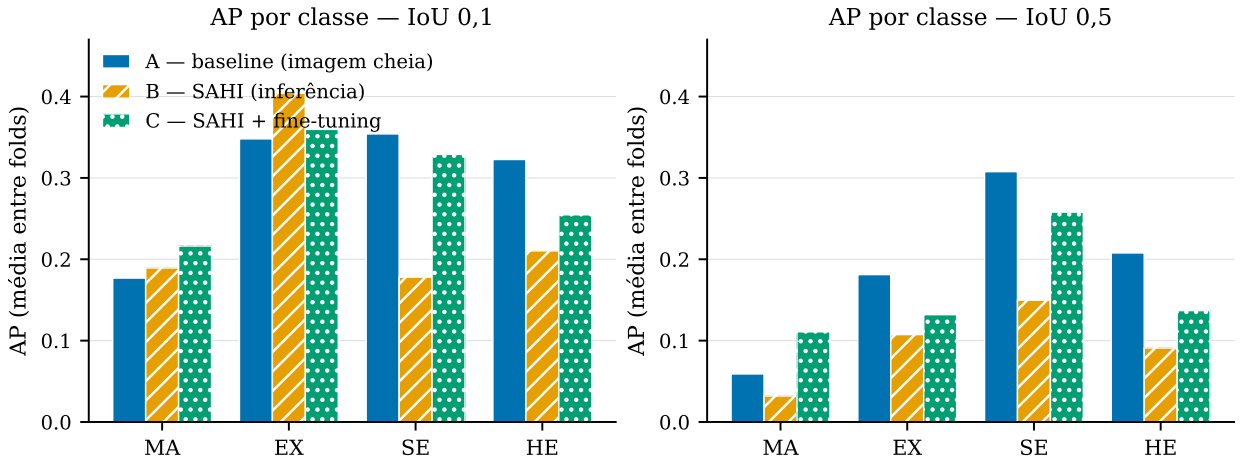


Figura 4. AP por classe de lesão (média entre os 15 *folds*), a IoU 0,1 (esquerda) e 0,5 (direita). O *slicing* beneficia os microaneurismas (MA) e, na inferência, os exsudatos duros (EX), mas prejudica os exsudatos moles (SE) e as hemorragias (HE), que são lesões maiores e mais difusas.

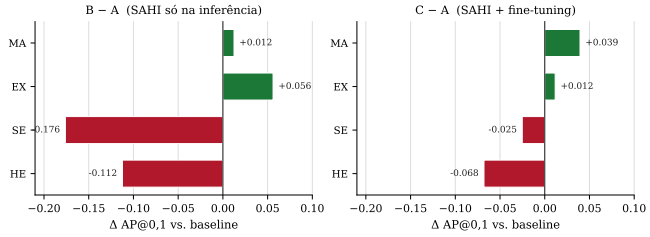


Figura 5. Ganho de AP@0,1 de cada configuração de *slicing* em relação ao *baseline* (A), por classe. Verde: ganho; vermelho: perda.

Tabela VI
COMPARAÇÃO COM A LITERATURA NO DDR (MAP A IOU 0,5). "VAL."/"TESTE" INDICAM O SUBCONJUNTO; "CV" INDICA VALIDAÇÃO CRUZADA REPETIDA (MÉDIA DE 15 *folds*). PROTOCOLOS, RESOLUÇÕES E ARTIFÍCIOS DE TREINO DIFEREM ENTRE ESTUDOS E NÃO SÃO DIRETAMENTE COMPARÁVEIS.

Trabalho / método	**Protocolo**	**mAP@0,5**
Li *et al.* [1] (SSD/YOLO)	teste	$< 0{,}01$
Alyoubi *et al.* [18] (CNN512+YOLOv3)	n. i.	0,216
Santos *et al.* [5] (YOLOv5+*Tiling*)	val.	0,263
Santos *et al.* [5] (YOLOv5+*Tiling*)	teste	0,154
Pereira *et al.* [7] (YOLOR-CSP+SAHI)	val.	0,381
Pereira *et al.* [7] (YOLOR-CSP+SAHI)	teste	0,222
Este trabalho, A (*baseline*)	CV	0,189
Este trabalho, C (SAHI+FT)	CV	0,159
Este trabalho, A (*split* oficial)	teste	0,113

split (tipicamente a *validação*), enquanto aqui se reporta a média de uma validação cruzada repetida, protocolo mais conservador. Para uma comparação justa, avaliou-se também a Config. A no *split* oficial do DDR (383/149/225), o mesmo usado por [5], [7]: nele o *baseline* obtém mAP@0,5 de 0,113, valor coerente com a faixa de teste dos trabalhos correlatos (0,154 e 0,222).

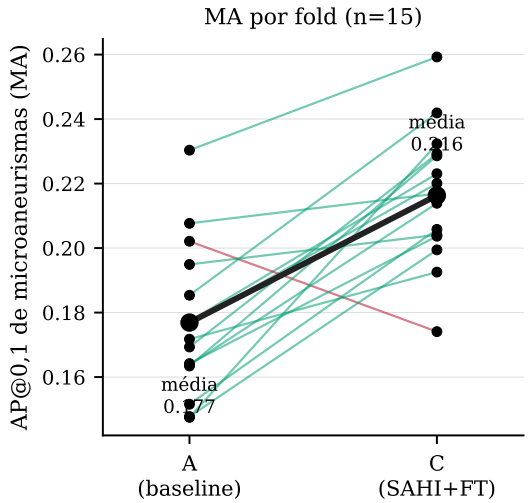


Figura 6. AP@0,1 de microaneurismas (MA), por *fold* ($n = 15$): *baseline* (A) vs. Config. C. As linhas verdes indicam *folds* em que a Config. C supera o *baseline* (13 de 15); a linha preta liga as médias (0,177 → 0,216).

Não se pretende, aqui, competir por desempenho absoluto: os métodos da literatura empregam artifícios adicionais (recorte do fundo, pré-treino em COCO, milhares de épocas, ajuste de âncoras, aumento de dados agressivo) que este estudo deliberadamente evita para *isolar* o efeito do *slicing*. O ponto é metodológico: a diferença entre validação e teste observada em [7] (0,381 vs. 0,222) reforça que números de *split* único devem ser lidos com cautela, e que a validação cruzada oferece estimativa mais fiel do desempenho esperado.

E. Limitações

Este estudo apresenta limitações. Primeiro, os hiperparâmetros foram fixados (sem busca), privilegiando a comparabilidade sobre o desempenho absoluto; uma busca dedicada poderia alterar as magnitudes, mas dificilmente inverteria o

trade-off por classe, de natureza geométrica. Segundo, avaliou-se um único detector (YOLOv8m) e uma única configuração de fracionamento (512/0,2); *tiles* menores poderiam favorecer os microaneurismas ao custo de agravar a fragmentação, e a extensão do *trade-off* a outras arquiteturas (YOLO recentes, *Transformers*) permanece em aberto. Terceiro, a avaliação restringe-se ao DDR, com generalização a outras bases, equipamentos e populações (por exemplo, IDRiD [25]) a verificar; e, sem identificadores de paciente, a partição é por imagem (Seção IV-E). Por fim, o escopo é a acurácia de detecção: desfechos em nível de paciente (graduação de RD, sensibilidade de triagem) não foram avaliados.

VI. Considerações Finais

Este trabalho avaliou, de forma controlada e estratificada por lesão, se a inferência assistida por SAHI melhora a detecção de lesões de RD no DDR quando aplicada a um detector YOLO moderno (YOLOv8m). A resposta é matizada. *Globalmente*, o SAHI não melhora o mAP: a inferência fracionada sobre um modelo treinado em imagem cheia (Config. B) degrada significativamente o desempenho, e mesmo o regime completo com ajuste fino (Config. C) não apresenta diferença detectável em relação ao *baseline*. *Por classe*, entretanto, o *slicing* cumpre sua promessa exatamente na lesão mais precoce e crítica: a Config. C melhora de forma significativa e consistente a detecção de microaneurismas, ao custo de piora nos exsudatos moles e nas hemorragias, que o fracionamento tende a fragmentar, *trade-off* que o mAP, por ser uma média entre classes, oculta.

A principal mensagem é que, em detectores YOLO modernos operando em alta resolução, o benefício do SAHI é *condicional*: localizado nas menores lesões e dependente de ajuste fino em *tiles*, e não o ganho global sugerido por trabalhos com arquiteturas anteriores. O resultado deve ser lido como evidência metodológica sobre o detector no DDR; uma recomendação clínica exigiria validação externa e desfechos em nível de paciente, fora do escopo deste estudo.

Como trabalhos futuros, pretende-se: (i) varrer o tamanho de *tile* e a sobreposição, buscando o equilíbrio do *trade-off*; (ii) explorar estratégias de fusão de caixas nas bordas dos *tiles* para reduzir a perda em lesões grandes; (iii) validar *cross-dataset*, treinando no DDR e avaliando em IDRiD [25] e em bases brasileiras; (iv) contrastar o *slicing* com arquiteturas mais recentes de YOLO e com *Transformers* de detecção; (v) aprofundar a análise dos modos de falha com técnicas de explicabilidade; e (vi) integrar a detecção a um *pipeline* de triagem com desfechos em nível de paciente.

Referências

[1] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019.

[2] A. A. Alghadyan, "Diabetic retinopathy – An update," *Saudi Journal of Ophthalmology*, vol. 25, no. 2, pp. 99–111, 2011.

[3] Ministério da Saúde, Governo Federal do Brasil, "Diabetes," 2024. [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/diabetes. Acesso em: 3 set. 2024.

[4] Ministério da Saúde, Comissão Nacional de Incorporação de Tecnologias no SUS (CONITEC), *Protocolo Clínico e Diretrizes Terapêuticas – Retinopatia Diabética (Versão Resumida)*, 2021. [Online]. Available: https://www.gov.br/conitec/pt-br/midias/protocolos/resumidos/PCDTResumidoRetinopatiaDiabtica.pdf

[5] C. Santos, M. Aguiar, D. Welfer, and B. Belloni, "A new approach for detecting fundus lesions using image processing and deep neural network architecture based on YOLO model," *Sensors*, vol. 22, no. 17, p. 6441, 2022.

[6] D. F. D. Assis, P. C. Cortez, and P. Motta, "Métodos de classificação de glaucoma em imagens do disco óptico," in *Anais do 14 Congresso Brasileiro de Inteligência Computacional*, Curitiba, 2019, pp. 1–8.

[7] A. Pereira, C. Santos, M. Aguiar, D. Welfer, M. Dias, and M. Ribeiro, "Detection of retinal microlesions through YOLOR-CSP architecture and image slicing with the SAHI algorithm," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. on Computer Vision (ECCV)*, 2018, pp. 3–19.

[9] A. Jithesh, C. Karumuri, V. K. R. Kotha, M. Doddapuneni, and T. Jeong, "Diabetic retinopathy lesion segmentation through attention mechanisms," in *Proc. IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2025.

[10] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9725–9734.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. European Conf. on Computer Vision (ECCV)*, 2020, pp. 213–229.

[12] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 966–970.

[13] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. European Conf. on Computer Vision (ECCV)*, 2016, pp. 21–37.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[17] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[18] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, p. 3704, 2021.

[19] C. Santos, M. S. De Aguiar, D. Welfer, and B. Belloni, "Deep neural network model based on one-stage detector for identifying fundus lesions," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.

[21] X. Bouthillier *et al.*, "Accounting for variance in machine learning benchmarks," in *Proc. Machine Learning and Systems (MLSys)*, vol. 3, 2021, pp. 747–769.

[22] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2011, pp. 145–158.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[24] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[25] P. Porwal *et al.*, "IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge," *Medical Image Analysis*, vol. 59, 101561, 2020.

Riemannian geometry-based classification of bidirectional movement intention from EEG in reaching tasks

Caio de Lima Pagano de Figueiredo
Department of Biomedical Engineering
Albert Einstein Israeli Faculty of Health Sciences
São Paulo, Brazil
caiofigueiredo@figa.fit

Jean Mendes Nascimento
Department of Mechanical Engineering and Mechanical Systems
University of São Paulo
São Paulo, Brazil
jeanmendes@usp.br

Arturo Forner Cordero
Department of Mechanical Engineering and Mechanical Systems
University of São Paulo
São Paulo, Brazil
aforner@usp.br

Bruna Mezzari Carlos
Department of Mechanical Engineering and Mechanical Systems
University of São Paulo
São Paulo, Brazil
bruna.carlos@usp.br

***Abstract*—Robotic devices, such as exoskeletons and robotic arms, have proven instrumental in upper limb rehabilitation and assistance in activities of daily living. These devices can be controlled by electroencephalographic (EEG) signals that detect the user's movement intention. However, EEG signals are susceptible to noise, positioning artifacts, and interference, posing significant challenges for real-time classification. This work presents the development and validation of an EEG-based classification pipeline for bidirectional (left and right) movement intention detection in robotic arm reaching tasks. The proposed system employs Riemannian geometry-based feature extraction combined with Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to process and classify EEG signals into left and right movement intentions. The pipeline was trained and validated on data acquired from five subjects using cross-validation, achieving a mean accuracy of approximately 79% over all subjects, with high stability between classes. These results show a performance comparable to existing literature while requiring fewer computational resources and EEG channels. Future work will focus on expanding the subject pool, increasing the number of movement classes, integrating complementary sensing modalities such as electromyography (EMG) and Inertial Measurement Units (IMUs), and validating the system under true real-time conditions.**

***Index Terms*—Robotic devices, Electroencephalography, BCI, Machine learning**

I. Introduction

Robotic devices, such as exoskeletons and robotic arms, have been proven to be useful in upper limb rehabilitation and assistance in activities of daily living (ADL) [1]. A significant advantage of these systems is the ability to monitor user electrophysiological and biomechanical variables that can be used to infer motor intentions. Detecting the user's intention is paramount for the robot to assist the movement seamlessly and smoothly. In this context, electroencephalographic (EEG) signals are widely used, as they capture the electrical brain activity preceding physical movement [2], [3].

However, the interfaces between the human body and assistive devices still present challenges. The major obstacle facing assistive robotic devices (ARDs) is how to control these devices without cognitively overloading the user. The control interface must be intuitive, allowing users to operate ARDs naturally [4]. For robotic arms, which are versatile machines capable of assisting individuals lacking motor control, achieving seamless coordination requires robust control and real-time classification algorithms.

One of the most frequently performed ADLs is the task of grasping an object. This task requires coordination among multiple joints, along with anticipatory planning by the central nervous system [5]. This complexity becomes a problem in exoskeleton control, as signals from several different channels must be processed simultaneously [6]. Therefore, a solid technique for validating control algorithms is the use of reaching tasks, where the robotic arm must be moved to a certain target within a defined space, with its precision and speed serving as performance metrics. These tasks challenge the algorithm's ability to process multiple sources of information at the same time and produce a real-time response [7].

Although EEG-based human movement intention prediction

(MIP) is promising, its practical application is still limited by the intrinsic variability of the electroencephalographic signal and the need for complex and time-consuming calibration for each user [8], leading to challenges in real-time acquisition, detection and classification. Furthermore, the movement of electrodes during use and the susceptibility to biological artifacts can alter the characteristics of the recorded signal and negatively impact classification [9], while practical barriers — such as comfort and setup difficulty — continue to hinder continuous signal acquisition and the daily adoption of these technologies [10].

A number of systems capable of real-time EEG interpretation have been reported in the literature. For instance, Meng et al. [11] developed a non-invasive Brain-Machine Interface system that enabled individuals to control a robotic arm in three dimensions in real time, allowing the user to perform complex target-reaching tasks. Ding et al. [12] trained a neural network model capable of controlling finger movements in robotic arms, achieving an accuracy of 80% in two-finger movements, and 60% in three-finger movements.

Given the existing challenges in EEG signal processing, the present work develops a classification pipeline for bidirectional movement intention in complex target-reaching tasks. We implemented processing routines to extract key features, trained machine learning models to classify bidirectional movement intention, and validated the trained model using pre-recorded signals.

II. Methodology

A. Data Generation

The data were obtained from five subjects performing an experimental task that used a robotic arm to reach specific targets. To control the arm movement, we used a methodology developed at the University of São Paulo (USP) that employs two Inertial Measurement Units (IMUs)—one attached to the subject's forearm and another to their wrist. The IMUs were calibrated and operated by sending the real-time position of the user's arm and hand to the robotic arm, which replicated the movement. Before each task, the subject maintained a resting state for 60 seconds with their eyes open to establish a baseline, though these data were not used in the present analysis and is reserved for future work. The experiment is shown in Figure 1 and is structured as follows: The subject started in the center, and a target was shown on the monitor, the subject had to then move the arm and stop at the correct target, repeated 96 times.

The raw signal was saved in a .poly5 file - a binary data format developed by the company TMSi, for use with the REFA amplifier - which contained the data for 32 EEG channels as well as 4 EMG channels, and sampled signals at 2048 Hz. For each experiment, the data from the first 24 EEG channels according to the 10-20 system (Fp1, Fp2, Fz, FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6) were used. Data from 5 healthy adult subjects were obtained, each participating in one session. The protocol was approved by the university

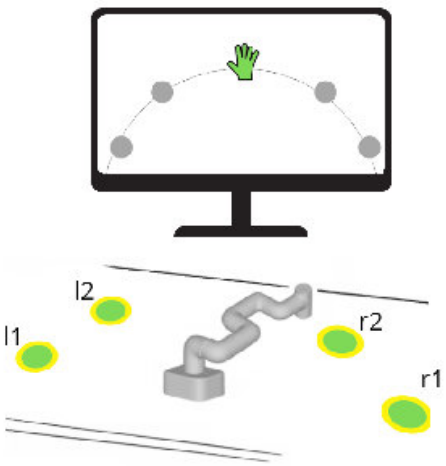


Fig. 1: Diagram of the task used in the project

Hospital Ethics Committee (CAAE: 90632625.7.0000.0076), and all participants provided written informed consent before entering the study.

B. Algorithm Development

The classification algorithm was developed in Python, using primarily Numpy, Pandas, Scikit-learn, Pyriemann, and MNE. The code can be divided into two parts: Processing and classification. The processing stage was based on extracting and cleaning the raw signal from the EEG, as well as extracting discriminative features for the classifier. The classification stage involved training the classifier on the extracted features, and evaluating its accuracy.

For noise removal, a notch filter was used to remove powerline (60 Hz) noise, in addition to a fourth-order Butterworth bandpass filter to retain frequencies in EEG range (0.5-100 Hz). Lastly, the DC component of the original signal was removed using mean subtraction, where the per-channel temporal mean was computed and subtracted from each channel independently. The filters were applied both forwards and backwards, eliminating lag. It's worth noting that for real-time implementations, causal variants of those filters will have to be used. To simplify the implementation, the preprocessing utilities provided by SciPy were used. Once the data was cleaned, the algorithm aligned the trigger onset time with the EEG capture time and the description of the cue provided to the subject. From the alignment, the algorithm extracted two-second windows, starting 0.5 seconds after the trigger, to extract the features.

With the data prepared, the Floating Sequential Forward Feature Selector (FSFFS) technique was applied to select the most important channels for the model. This technique evaluates the full feature extraction and classification pipeline on each channel individually, selecting the one that yields the highest accuracy. Next, it tests the combination of this channel with all the others in the same manner and selects the most accurate combination. After each addition, the algorithm

tries removing previously selected channels, and evaluating the cross-validated accuracy. This process is repeated until the channel selection threshold is reached, which in this project was set to 10.

The processing pipeline involved calculating the covariance matrix across the selected channels, followed by a projection of this matrix onto a tangent space on the Riemannian manifold. With the Riemannian vectors, PCA was performed to reduce its dimensions to 15, followed by a Linear Discriminant Analysis (LDA) classifier.

C. Validation

The trained models were validated using techniques such as cross-validation, with performance assessed through accuracy, recall, precision, F1 score, and confusion matrices. Cross-validation involved dividing the dataset for each subject into k equal-sized subsets using k-fold cross-validation with k = 5. The model was iteratively trained on all but one of these subsets and tested on the remaining subset. This approach allows assessment of how well the model generalizes unseen data.

To measure classification performance, the project focused on four key metrics. Accuracy represents the overall proportion of correct predictions made by the model. Recall measures the proportion of actual positive cases the model correctly identified. Precision evaluates the proportion of positive predictions the model made that were actually correct. Finally, the F1 Score serves as the harmonic mean of Precision and Recall, providing a balanced metric for the model's performance.

In order to display these metrics, a confusion matrix was used. A confusion matrix divides the data points into four cells, according to the correspondence between their true and predicted labels, creating True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These metrics are formally defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

In addition to per-class metrics, statistical methods were also employed in order to ascertain that the models were truly performing better than just chance. To evaluate whether or not the model would produce correct predictions, the binomial test was used to determine whether or not the models were predicting with an accuracy greater than chance. A model was considered above chance if the calculated p-value was below the threshold of $\alpha = 0.05$. The hypotheses considered for the binomial test were as follows, following a one-sided approach:

H_0 : $\pi \leq 0.5$:The model predicts at or below chance.

H_1 : $\pi > 0.5$:The model predicts above chance.

Because a finite number of trials allows the classifier to exceed the chance level by luck, an adjusted chance level was computed as the upper 95% confidence bound on the accuracy of a random classifier over n trials:

$$\text{Adj. Chance} = p_0 + z_{0.95}\sqrt{\frac{p_0(1-p_0)}{n}}$$

where $p_0 = 0.5$ is the chance rate and $z_{0.95} = 1.645$ is the one-sided 95% normal quantile.

Alongside the binomial test, the Wilcoxon signed-rank test was used to determine whether subjects as a group performed above chance. A one-sample Student's t-test was not chosen because the small sample size does not allow the assumption of normally distributed accuracies. The hypotheses for the Wilcoxon test were as follows, also following a one-sided approach:

H_0 : The distribution of the subject accuracies are centered at chance: Subject accuracies are symmetric around chance.

H_1 : The distribution of the subject accuracies are shifted toward positive values: subjects as a group perform above chance.

III. Results

Having processed data from 5 different subjects, the results from their respective models are as follows: Figures 2 to 6 contain the confusion matrices for subjects 1 through 5, respectively. Table I complements that with the computed metrics for all subjects, as well as averages for each metric. Table II and figure 7 contain data on which of the 24 available channels the classifier selected for each subject, according to the 10-20 system.

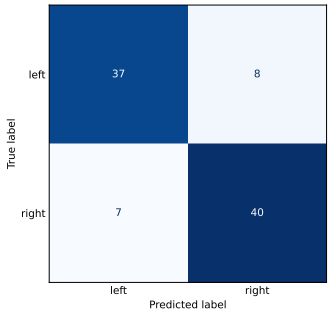


Fig. 2: Confusion Matrix for Subject 1

TABLE I: Classification Reports for All Subjects

		Left			Right		
Subject	**Acc.**	**P**	**R**	**F1**	**P**	**R**	**F1**
Subject 1	0.8370	0.8409	0.8222	0.8315	0.8333	0.8511	0.8421
Subject 2	0.8018	0.8200	0.7593	0.7885	0.7869	0.8421	0.8136
Subject 3	0.7789	0.7955	0.7447	0.7692	0.7647	0.8125	0.7879
Subject 4	0.7553	0.7556	0.7391	0.7473	0.7551	0.7708	0.7629
Subject 5	0.7849	0.7500	0.8478	0.7959	0.8293	0.7234	0.7727
Average	**0.7916**	**0.7924**	**0.7826**	**0.7865**	**0.7939**	**0.8000**	**0.7958**

P: Precision; R: Recall; F1: F1-Score; Acc.: Accuracy

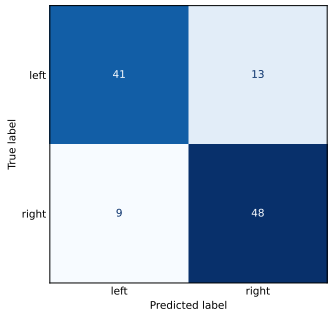


Fig. 3: Confusion Matrix for Subject 2

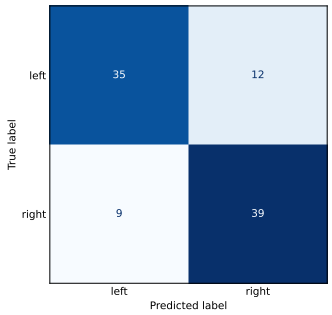


Fig. 4: Confusion Matrix for Subject 3

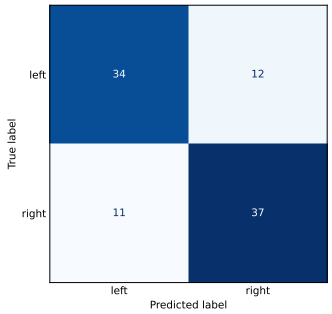


Fig. 5: Confusion Matrix for Subject 4

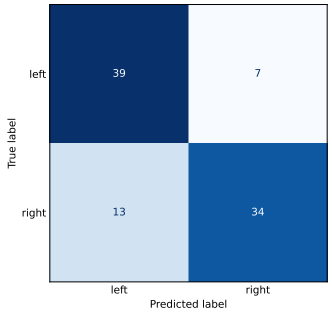


Fig. 6: Confusion Matrix for Subject 5

The matrices and Table I show that the models from all subjects behaved similarly, achieving similar levels of accuracy, precision and recall for both left and right classes, with a mean accuracy of 79%. This is also supported by the matrices themselves, which show a clear line along the main diagonal, indicating that the model made more correct predictions than incorrect ones. The one small exception is Subject 5 (Figure 6), which shows a notable spike of erroneous left predictions, resulting in the lowest left precision of all models (75%). The best performing model, Subject 1 (Figure 2), made only 15 incorrect predictions, having the highest-performing metrics out of all tested models.

TABLE II: FSFFS-Selected EEG Channels per Subject

Subject	**Selected Channels**
Subject 1	FC6, C2, CP5, Fp1, FC3, C3, FCz, FC4, CP1, C6
Subject 2	FC5, C2, C6, FC3, Fp1, CP4, Fp2, FC4, FC6, CP6
Subject 3	FC6, FC4, FC3, FC5, FC1, FCz, Cz, Fz, C1, CP1
Subject 4	CP3, FC5, C2, Fp1, Cz, FC2, C3, FC4, Fp2, FC6
Subject 5	Fz, FC6, FC5, C1, CP5, FCz, CPz, FC3, CP4, Cz

Channels listed in FSFFS selection order. All 24 available electrodes spanned Fp1–CP6 in the 10–20 system.

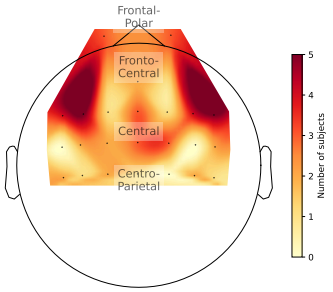


Fig. 7: Channel Selection Frequency Across Subjects

Table II shows that most models selected similar channels. Channel FC6 appears across all subjects and is frequently the first channel FSFFS selected, with recurring presence of channels C2, C3, and CP1. Figure 7 supports this, showing strong selection in the frontocentral region, with some centroparietal activity. Pre-frontal channels also show moderate selection. Temporal and occipital electrodes were not included in the 24-channel setup and thus are absent from the map.

TABLE III: Per subject adjusted chance level and binomial test p-value.

Subject	n	Adj. chance	p
Subject 1	92	0.5857	1.6×10^{-11}
Subject 2	111	0.5781	4.8×10^{-11}
Subject 3	95	0.5844	2.1×10^{-8}
Subject 4	94	0.5848	3.6×10^{-7}
Subject 5	93	0.5853	1.5×10^{-8}
Average	**97**	**0.58366**	

TABLE IV: Inter-subject summary of cross-validated accuracy.

Statistic	Value
Subjects (N)	5
Mean accuracy	0.7916
SD	0.0303
SEM	0.0136
95% CI	[0.7539, 0.8293]
Median accuracy	0.7849
Range	0.7553–0.8370
Wilcoxon W	15
Wilcoxon p (one-sided)	0.031

Table III shows the results of the binomial tests applied to the models, in which they all perform above chance ($p < 0.05$). Table III also shows the adjusted chance threshold for a given subject's model, showing that all models reach accuracies higher than those adjusted by the sample sizes. Table IV shows the per-subject variation, with a standard deviation of 3.0% and standard error of the mean of 1.4%. Table IV also shows the results of the Wilcoxon test, which returned the smallest possible p value for 5 subjects: 0.031, meaning that the group performs above chance.

IV. Discussion

The results demonstrate that the models have a classification precision and stability between classes consistent with what would be required for user intention and control of assistive devices. The models were able to correctly classify a movement across most trials, with little bias towards any given class. Furthermore, the results indicate that it is possible to attain high precision in EEG processing without relying on resource-intensive filtering like Laplacian filters or Independent Component Analysis (ICA).

In addition, the statistical reports also confirm the results shown by Table I. The low probabilities calculated by the binomial test show that the chance of models reaching those accuracies by luck is extremely low, which means that there must be some pattern that the model is correctly picking up on, and predicting. This is further reinforced by the adjusted chance values, which show that every model still performs above chance against the upper 95% bound.

The inter-subject metrics tell a similar story. A low standard deviation alongside a low standard error of the mean show the models don't diverge too much from the mean, and stay consistent in their predictions. The Wilcoxon test also helped prove this, as all 5 subjects had positive signed differences, resulting in a W value of 15, and the lowest possible value for p: 0.031. This means that there is significant evidence at $\alpha = 0.05$ against the per-subject accuracies being symmetrically distributed about the chance level.

Furthermore, the channels selected by the models align with previous results reported in the literature. The frontocentral and central regions are established as sites for motor-related cortical activity [13], and the channels chosen by the models were, for the most part, all part of these areas. The exceptions are part of the centro-parietal region and the frontal polar, in which the former helps to integrate somatosensory stimuli with motor signals [14], and the latter related to visual processing [15]. It's worth noting that frontal polar channels have also been related to eye blinking artifacts, which could be impacting the model's performance.

The classification accuracies achieved are also consistent with comparable work in the literature. For instance, Ding et al. [12] also developed a model capable of decoding movements using EEG signals. Their model was also capable of decoding more precise movements, such as individual fingers, in real time. However, their model differed significantly from ours. They used a deep neural network as their classifier, trained on a larger cohort of 21 healthy adult subjects. Our model achieved similar results, using a less computationally expensive approach and fewer subjects, albeit for less precise intention monitoring.

Meng et al. [11] also developed an EEG classifier for a robotic arm, employing an autoregressive model trained on data from 13 healthy individuals across 62 channels. They were also successful in using their model to drive the arm in reaching tasks, achieving accuracies of around 80%. Our model uses far fewer channels to classify movements, while still maintaining comparable accuracy.

However, this processing pipeline was only tested with 5 different subjects, all with comparable characteristics, in a controlled environment. This means that while our model performs well under controlled testing conditions, it might not achieve similar results when deployed in real-time conditions. This could be attributed to several factors inherent to real-time use — such as signal latency, data loss, subject movement, and environmental noise — that do not affect pre-recorded signals.

Furthermore, the small sample size means that this pipeline was not tested on a sufficiently diverse population. While the model performed well on the limited subset evaluated, it might not generalize well to people of a different ethnicity, geographical location, or age. Additionally, the model's classification was binary, limited to distinguishing left and right movements.

The models proposed by Meng et al. [11] and Ding et al. [12] incorporated a greater number of movement classes, making them more useful in real-world applications. While binary classification served as a strong proof of concept, expanding the number of classes is necessary for broader clinical applicability.

Therefore, the next steps of this research project are to increase the number of movement classes, to broaden the range of movements the classifier can analyze, and to move toward a model capable of continuous movement decoding. This will create a more robust model and increase its usefulness in real-world scenarios.

Real-time deployment remains the most immediate challenge. This will require measuring the latency and response time in each step of the pipeline and developing functionality for real-time monitoring and data delivery. This would allow the pipeline to use EEG signals to control devices like robotic arms, expanding its applicability in clinical settings.

Furthermore, to increase the reliability of the model, the subject pool would be expanded, testing the model by incorporating more diverse datasets encompassing subjects of varying ages and neurological profiles. This will assess whether our classifier can accurately and consistently detect movement patterns from a broader and more diverse population, rather than a narrow, homogeneous subgroup.

Lastly, the signals from complementary modalities such as EMG and inertial sensors could be merged with the EEG data and integrated into the model, providing the model with richer, multimodal input. While EEG captures neural activity from cortical sources, EMG captures peripheral muscle activation, providing richer contextual information than cortical signals alone. Inertial sensors would further complement both modalities by capturing kinematic data that reflects the physical execution of the intended movement, thereby enriching the spatiotemporal patterns available for model training.

V. Conclusion

This work presented a pipeline for the processing and classification of EEG signals for bidirectional movement intention detection in robotic arm reaching tasks. Using a Riemannian geometry-based feature extraction approach combined with Linear Discriminant Analysis, the proposed system achieved a mean accuracy of 79.2% across five subjects, with high stability between classes. These results can be compared with those reported in the literature [12] [11], while requiring significantly fewer computational resources, fewer EEG channels, and a smaller subject cohort for training.

Nevertheless, several limitations remain to be addressed before this pipeline can be considered for real-world deployment. The model was validated exclusively on pre-recorded data from a small, homogeneous subject pool, and its performance under true real-time conditions, where signal latency, noise, and subject movement are unavoidable, remains to be evaluated. Additionally, the binary nature of the classifier restricts its utility across broader assistive robotics scenarios.

Future work will focus on expanding the number of movement classes, implementing and benchmarking real-time signal acquisition and classification, broadening the subject pool to improve generalizability, and incorporating complementary sensing modalities such as EMG and inertial sensors. Taken together, the results presented here establish a promising foundation to help improve the efficiency of real-time EEG-based robotic control systems.

References

[1] C. Tang, Z. Xu, E. Occhipinti, W. Yi, M. Xu, S. Kumar, G. Virk, S. Gao, and L. G. Occhipinti, "From brain to movement: Wearables-based motion intention prediction across the human nervous system," *Nano Energy*, vol. 115, pp. 108 712–108 712, 10 2023.

[2] E. Lew, "Detection of self-paced reaching movement intention from eeg signals," *Frontiers in Neuroengineering*, vol. 5, 2012.

[3] M. H. Uribe, C. R. Vazquez, and J. M. Antelis, "A fast emg-based algorithm for upper-limb motion intention detection by using levant's differentiators," *IEEE Access*, vol. 10, pp. 111 623–111 635, 2022.

[4] B. Yang, J. Huang, X. Chen, C. Xiong, and Y. Hasegawa, "Supernumerary robotic limbs: A review and future outlook," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, pp. 623–639, 08 2021.

[5] E. Otaka, Y. Otaka, S. Kasuga, A. Nishimoto, K. Yamazaki, M. Kawakami, J. Ushiba, and M. Liu, "Clinical usefulness and validity of robotic measures of reaching movement in hemiparetic stroke patients," *Journal of NeuroEngineering and Rehabilitation*, vol. 12, 08 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4533801/

[6] B. Chen, Y. Zhou, C. Chen, Z. Sayeed, J. Hu, J. Qi, T. Frush, H. Goitz, J. Hovorka, M. Cheng, and C. Palacio, "Volitional control of upper-limb exoskeleton empowered by emg sensors and machine learning computing," *Array*, vol. 17, p. 100277, 03 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005623000024

[7] E. Trigili, L. Grazi, S. Crea, A. Accogli, J. Carpaneto, S. Micera, N. Vitiello, and A. Panarese, "Detection of movement onset using emg signals for upper-limb exoskeletons in reaching tasks," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, 03 2019.

[8] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, p. 031005, 04 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aab2f2/meta

[9] J. A. Urigüen and B. Garcia-Zapirain, "Eeg artifact removal-state-of-the-art and guidelines," *Journal of neural engineering*, vol. 12, p. 031001, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25834104

[10] R. Portillo-Lara, B. Tahirbegi, C. A. R. Chapman, J. A. Goding, and R. A. Green, "Mind the gap: State-of-the-art technologies and applications for eeg-based brain–computer interfaces," *APL Bioengineering*, vol. 5, p. 031507, 09 2021.

[11] J. Meng, S. Zhang, A. Bekyo, J. Olsoe, B. Baxter, and B. He, "Non-invasive electroencephalogram based control of a robotic arm for reach and grasp tasks," *Scientific Reports*, vol. 6, 12 2016.

[12] Y. Ding, C. Udompanyawit, Y. Zhang, and B. He, "Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level," *Nature Communications*, vol. 16, 06 2025. [Online]. Available: https://www.nature.com/articles/s41467-025-61064-x

[13] Óscar Wladimir Gómez-Morales, D. F. Collazos-Huertas, A. M. Álvarez Meza, and C. G. Castellanos-Dominguez, "Eeg signal prediction for motor imagery classification in brain–computer interfaces," *Sensors*, vol. 25, pp. 2259–2259, 04 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11991189/

[14] L. Fogassi and G. Luppino, "Motor functions of the parietal lobe," *Current opinion in neurobiology*, vol. 15, pp. 626–31, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16271458

[15] S. Ghafoori, A. Cetera, A. Rabiee, M. H. Farhadi, R. Singh, M. Furmanek, Y. Shahriari, and R. Abiri, "Cross-frequency bispectral eeg analysis of reach-to-grasp planning and execution," *Computers in Biology and Medicine*, vol. 213, p. 111791, 02 2026.

Risk Classification for Sudden Cardiac Death Using Single-Lead ECG

Ana V. M. Inocêncio Lopes
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
vitoria.morais@ufpe.br

Naelso Alves Cunha
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
naelso.cunha@ufpe.br

Gilson José Alves
Electrical Department
Federal Institute of Pernambuco
Recife, Brazil
gilsonj.alves@gmail.com

Emmanuel Andrade de Barros Santos
Control and Automation Department
University of Pernambuco
Recife, Brazil
emmanuel.andrade@upe.br

Diogo Roberto Raposo de Freitas
Electronics Department
Federal University of Pernambuco
Recife, Brazil
diogo@poli.br

Marco Aurélio Benedetti Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
marco.rodrigues@ufpe.br

***Abstract*— Cardiovascular diseases (CVDs) remain the leading cause of death worldwide in both developed and developing countries. A significant portion of these deaths is associated with sudden cardiac death (SCD), an unexpected event frequently triggered by cardiac arrhythmias that compromise the heart's ability to pump blood effectively. The electrocardiogram (ECG) is the most widely used non-invasive tool to evaluate the functional status of the heart, as it records the electrical activity generated by the depolarization and repolarization of the cardiac chambers. The objective of this work is to propose and evaluate a parameter-optimized 1D-CNN, seeking to enable classification with reduced patient-dependency bias. To measure its true efficiency and generalization capacity, the model is contrasted with established approaches in literature. Data from 19 patients from the same database were used, with 15 assigned for training and 4 for validation. The data were digitally processed using a digital filter, subsequently segmented into 5-second intervals, and normalized. Intervals considered normal and high-risk were separated based on database annotations. The data was used to train a one-dimensional convolutional neural network with four convolutional layers, a Global Average Pooling (GAP) layer, a dense layer, and an output layer. The results yielded an overall accuracy of 96.69%, sensitivity for high-risk findings of 94.31%, sensitivity for normal findings of 97.91%, specificity of 97.39%, and an Area Under the Curve (AUC) of 0.9894. The results demonstrate the system's applicability for sudden death risk detection, serving as a potential auxiliary tool in clinical diagnosis.**

Keywords*— *One-dimensional convolutional neural networks, 1D CNN, sudden cardiac death, classification

I. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide in both developed and developing countries. Projections indicate that the global number of deaths associated with CVDs will continue to increase significantly over the coming decades, representing a critical challenge for public healthcare systems [1], [2]. A relevant portion of these deaths is associated with sudden cardiac death (SCD), an unexpected event often triggered by cardiac arrhythmias that compromise the heart's ability to pump blood effectively [3], [4]. In this scenario, early identification of abnormal cardiac rhythms is essential, not only to support clinical decision-making but also to enable preventive strategies in high-risk patients. Therefore, continuous and non-invasive monitoring of cardiac activity plays a central role in modern cardiology and biomedical engineering [5].

The electrocardiogram (ECG) is the most widely used non-invasive tool to evaluate the functional status of the heart, as it records the electrical activity generated by the depolarization and repolarization of the cardiac chambers [5], [6]. The ECG signal is quasi-periodic, and each heartbeat is characterized by a sequence of distinct waveforms — the P wave, the QRS complex, and the T wave — whose morphology, amplitude, and timing contain diagnostic information about the underlying cardiac condition [6], [7]. However, reliable identification of abnormal cardiac rhythms is challenging due to the variety of noise sources and high morphological variability of the ECG across different individuals [8], [9]. Although clinical multi-lead ECG systems provide comprehensive spatial data, single-lead ECG configurations offer a low-cost, broadly accessible alternative suitable for long-term monitoring and wearable device applications [10].

A wide range of techniques has been proposed in the literature for ECG wave delineation, feature extraction, and arrhythmia classification. Traditional approaches frequently utilize the Pan-Tompkins algorithm [11] for QRS complex detection, wavelet-based transforms for joint time-frequency representation [12],[13],[14], or classical machine learning classifiers (such as Support Vector Machines) [15]. However, these traditional processing pipelines are highly sensitive to noise and rely heavily on the precise detection of fiducial points, which often fails in real clinical environments due to inter- and intra-patient variability [8]. Recently, Deep Learning models, particularly 1D Convolutional Neural Networks (1D-CNNs), have emerged as a powerful alternative by performing automated end-to-end feature extraction directly from raw or filtered signals, bypassing the need for manual wave segmentation [16].

Wang et al. [17] proposed an early prediction method for sudden cardiac death risk using a Nested LSTM combined with wavelet denoising and phase space reconstruction to predict the continuous temporal trajectory of ECG signals. The work of Telangore et al. [18] proposed a multimodal framework for early SCD prediction, using a fusion of 1D-CNN networks combined with LSTM for raw ECG data extraction, 2D-CNN networks to analyze Hilbert-Huang spectra (HHT), and Vision Transformer (ViT) models with

self-attention applied to 2D wavelet scalograms. The same study validated its model through a 4-fold cross-validation scheme. In a recent study within time-series processing, Khan et al. [19] demonstrated the feasibility of using 1D Convolutional Neural Networks operating directly on raw ECG signals. The authors proposed a deep residual architecture (ResNet) focused on multi-class classification of five arrhythmia types from the MIT-BIH database.

In this context, the objective of this work is to develop and evaluate a highly regularized, parameter-efficient 1D-CNN for patient-independent classification of sudden cardiac death (SCD) risk using single-lead, low-resolution ECG signals. Unlike conventional deep learning approaches that suffer from overfitting or evaluate performance using patient-overlapping data splits, the proposed methodology focuses on rigorous patient-independent validation to ensure true clinical generalization. The network architecture incorporates Global Average Pooling (GAP) and biologically aligned data handling to mitigate overfitting. Signals from a public database are used to validate the proposed model, demonstrating high classification accuracy on previously unseen patients while maintaining minimal computational load, suitable for resource-constrained devices.

II. BACKGROUND

A. Electrocardiogram

While the cardiac impulse travels through the heart, electrical currents propagate to the surrounding tissues, and a small portion reaches the body surface [20]. The electrocardiogram (ECG) signal results from the polarization and depolarization of the internal cardiac chambers. Its acquisition is performed via electrodes placed on the skin surface. The ECG is an essential tool for providing information on the functional status of the heart.

Through the electrocardiogram, it is possible to identify how the cardiac impulse circulates across the muscle surface and whether contractions (systoles) and relaxations (diastoles) were performed properly. The ECG examination can inform on the physical condition of the heart muscle simply by capturing the signal via surface electrodes. The simplicity of the examination and the depth of information it provides reinforce its importance in cardiological diagnosis [22].

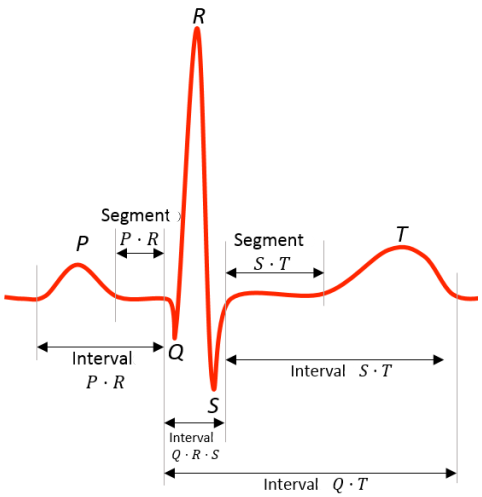


Fig. 1. Hearth waves and segments. Source: Adapted from [21].

The electrocardiogram (ECG) graphically translates the mechanical movements of the heart through three main phases (Fig. 1): the P wave records the electrical activation that causes the atria (upper chambers) to contract to expel blood; the QRS complex signals the rapid and vigorous contraction of the ventricles (lower chambers) combined with atrial relaxation; and the T wave represents the relaxation and recovery period of the ventricles. Thus, each oscillation on the graph acts as a continuous visual map of cardiac muscle exertion and rest [5].

B. Sudden Cardiac Death

Maintaining a healthy individual also requires a healthy heart. It is the electrical activity of the heart that provides a normal cardiac rhythm. Irregular electrical activities or arrhythmias are caused by cardiovascular system problems, which can accelerate the heart (tachycardia) or cause it to beat too slowly (bradycardia). Additionally, another type of arrhythmia is fibrillation, caused by a chaotic and irregular cardiac rhythm. Ventricular fibrillation and ventricular tachycardia are the main causes of sudden cardiac death, requiring immediate medical attention [23].

Sudden cardiac death refers to an abrupt loss of heart function, leading to loss of consciousness and cessation of breathing. SCD is one of the leading causes of natural death worldwide [24]. The patient must be diagnosed before reaching states requiring emergency intervention, which is possible through cardiac signal monitoring and patient risk classification.

C. Limitations of Traditional Clinical Assessment

As mentioned earlier, immediate medical intervention is required for a patient to have a chance of survival. The problem is that physicians currently attempt to predict this by measuring static intervals (such as QRS duration or QT interval). Manual measurements are accurate considering the professional's level of expertise; however, arrhythmic conditions appear silently and unexpectedly [25]. Diagnosis may be straightforward for healthcare professionals, but static measurements fail to capture the short-term dynamics of the heart. Traditional clinical assessment requires constant evaluation depending on the patient. Therefore, equipment such as a low-cost Holter monitor, which uses processing to inform physicians of cardiac events and provide quantified data along with a high SCD risk diagnosis, could assist and streamline medical practice.

D. Deep Learning Approach

A CNN, or Convolutional Neural Network, is a deep neural network model with multiple stacked layers designed to mimic neuronal activity in the human brain [26]. The CNN feature extraction process involves several hidden layers, including convolutional layers, Rectified Linear Unit (ReLU) activation functions, and pooling layers [27]. The CNN classification process includes a fully connected layer and an activation function that produces the classification output.

The typical structure of a 1D CNN is shown in Fig. 2, and it is composed of three important layers: 1D convolutional layers, pooling layers, and fully connected layers [28].

A one-dimensional convolutional neural network (1D-CNN) analyzes the signal continuously from start to finish. It does not need to measure where a wave starts or ends; the training process adjusts the weights so that the output layer correctly classifies the signal, thereby learning to correlate

subtle, dynamic morphological variations that go unnoticed by the human eye.

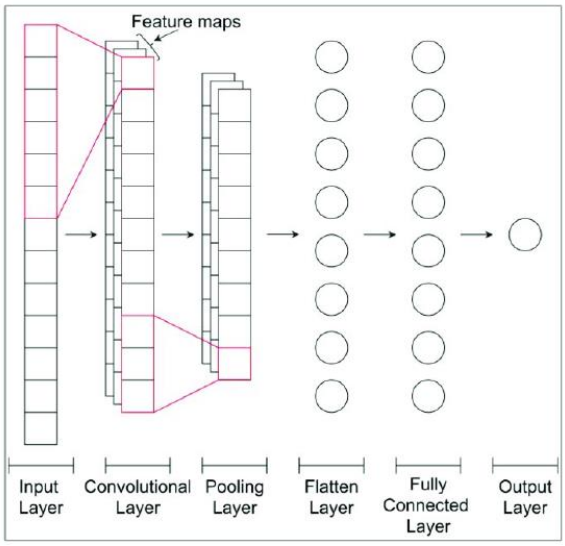


Fig. 2. One-dimensional convolutional neural network (1D CNN) architecture. Source: Guessoum at al. [28].

III. Methodology

The proposed system architecture for early sudden cardiac death risk detection was divided into digital signal preprocessing, neural network construction and training, and structural validation. For the development of this work, the Python programming language was used, along with essential libraries for algorithm development, such as PyTorch (an open-source machine learning framework) and NumPy (focused on array and matrix operations) [29].

A. Database

For algorithm development, training, and validation, the Sudden Cardiac Death Holter Database from the PhysioNet platform was used [30], [31]. The database consists of 23 long-term Holter recordings corresponding to 23 patients who exhibited sustained ventricular tachyarrhythmias, most of whom progressed to cardiac arrest during monitoring. ECG signals were acquired via two leads at a sampling rate of 250 Hz and 12-bit resolution, covering a dynamic range of ±10 mV. Recordings vary in duration between approximately 4 and 25 hours, with a predominance of recordings close to 24 hours. In this study, only one ECG lead was used — the Modified Lead II with electrodes positioned on the chest in areas of lower muscle movement — characterizing a single-lead ECG approach.

Nineteen patients were selected from the database. To prevent data leakage, a inter-patient splitting protocol was implemented. Of the 19 patients, 12 were assigned to neural network training, 3 were isolated for the validation set, used for early stopping monitoring and hyperparameter tuning during the training phase. Four patients were completely held out to form the final testing set.

Segmentation into 5-second windows was based on the annotation files (.atr and .ari) provided by the PhysioNet platform [30][31]. Segments displaying baseline sinus rhythm without tachyarrhythmias (annotation N) were labeled as normal rhythm (control). The high-risk class comprised windows containing confirmed episodes of Ventricular Tachycardia (VT), and Ventricular Fibrillation (VF), as well as ECG time windows corresponding to the 5-minute interval immediately preceding the onset of these fatal events. This formulation allows evaluating the network's capacity to recognize pre-arrhythmic and acute arrhythmic electrocardiographic alterations associated with Sudden Cardiac Death (SCD).

B. Signal preprocessing

To ensure ECG signal quality and eliminate noise that could compromise network learning, raw data underwent preprocessing. A 4th-order Butterworth bandpass filter was applied to the signal, constraining it between cutoff frequencies of 0.5 Hz and 100 Hz. This removed electromyographic noise and respiration baseline wander, making the signal suitable for the neural network.

Following filtering, the signal was segmented into 5-second time windows, totaling 360 normal time windows and 180 high-SCD-risk windows per patient.

Finally, to homogenize signal amplitudes across different patients and electrodes, segment normalization was performed using Z-score standardization, adjusting values based on standard deviations from the mean to contain input data within a uniform distribution scale.

C. Proposed 1D-CNN Architecture

The use of convolutional neural networks (CNNs) has demonstrated effectiveness in both image and signal processing, transforming biomedical signal processing. CNNs have achieved success ranging from X-ray and MRI image processing to ECG signal classification [32].

Classification of ECG segments was performed via a 1D Convolutional Neural Network (1D-CNN) architecture. Choosing 1D convolutions is justified by their intrinsic ability to process time series, extracting local morphological patterns directly from the signal without requiring manual feature engineering [33].

The model in Fig. 3 adopts an end-to-end learning approach structured into three stages: feature extraction, temporal aggregation, and classification. The first stage consists of four sequential convolutional blocks in a hierarchical structure. The initial layers use larger receptive fields (wider kernels) combined with Max-Pooling to capture general heart rhythm shapes and reduce temporal dimensionality. As the signal progresses through the network, kernel sizes decrease while channel depth increases, allowing representation of more complex and refined morphological features. All blocks share Batch Normalization for training stability and ReLU non-linear activation.

Next, a Global Average Pooling (GAP) layer is applied to synthesize feature maps into a single-dimensional vector, drastically mitigating the risk of overfitting. Finally, this vector is processed by a densely connected layer — regularized by Dropout — which directs learned representations to the linear output layer for binary classification between Normal Rhythm and High Risk.

It is worth noting that simpler structures were attempted prior to this final architecture; however, modifications were required to improve final classification performance and accuracy.

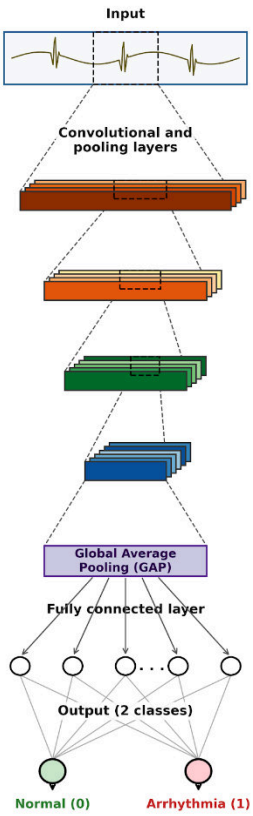


Fig. 3. Structure of the 1D-CNN used. Source: the authors (2025).

D. *Training and Evaluation Protocol*

1D-CNN training was conducted using the Categorical Cross-Entropy loss function from the PyTorch library, ideal for the proposed classification scenario. Network weights were updated using the Adam optimizer with an initial learning rate of 0.001 and batch size set to 32. The learning process extended over 80 epochs, incorporating early stopping monitored via validation loss to prevent overfitting. To measure predictive performance on the independent test set (4 patients unseen during training), Accuracy, Sensitivity (Recall), Specificity, F1-Score, and Area Under the ROC Curve (AUC-ROC) were calculated, ensuring a multidimensional and statistically solid evaluation.

To evaluate the impact of inter-patient variability and address the temporal dependence between multiple segments extracted from the same subject, the model performance was analyzed both in aggregate and on a per-patient basis. For this analysis, the following parameters were calculated for each patient: total segments, sensitivity for sudden death, specificity for normal status, and accuracy.

IV. Results

Performance of the proposed 1D-CNN architecture was evaluated using the independent test set comprising ECG recordings from 4 patients not presented to the network during training. This set generated a total of 2,208 continuous 5-second time windows (1,452 normal and 756 high-risk). It should be noted that although the primary evaluation unit is the time segment to enable continuous signal analysis, these samples correspond to clustered observations originating from the 4 tested individuals. Figure 4 displays the signal from Record No. 30, containing segments considered normal and high risk.

The architecture achieved an overall accuracy of 96.69%, with a sensitivity of 94.31% for the high-risk class and a specificity of 97.93% for the normal class. For a robust, clinically relevant analysis, performance was measured across multiple metrics, consolidated in Table I.

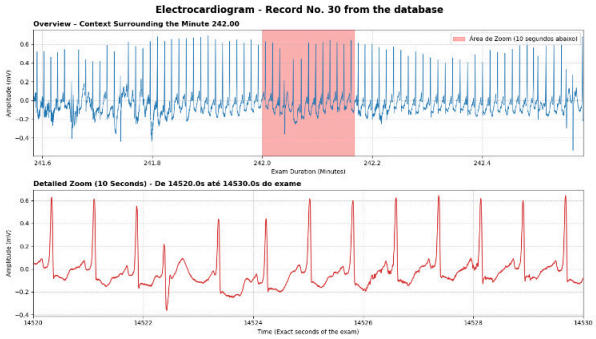


Fig. 4. Structure of the 1D-CNN used. Source: the authors (2025).

The sensitivity achieved by the model was 94.31% and minimizes false negatives, an extremely dangerous occurrence during screening of patients at risk for sudden death. The specificity of 97.93% demonstrates that the model is robust against false alarms, correctly identifying periods of stable normal rhythm.

TABLE I. Quantitative performance of the proposed 1D-CNN on the independent test set.

	Precision	Recall	F1-score	support
Normal	0.9706	0.9793	0.9750	1452
Risk/Arrythmia	0.9596	0.9431	0.9513	756
Acuracy			0.9669	2208
Macro avg	0.9651	0.9612	0.9631	2208
Weighted avg	0.9669	0.9669	0.9669	2208
Sensitivity (Risk Detection): 94.31%				
Specificity (Normal Detection Rate): 97.93%				

Source: the authors (2025).

ROC curve analysis (Fig. 5) revealed an Area Under the Curve (AUC) of 0.9894, indicating that the network maintains a highly stable and reliable decision threshold regardless of individual morphological variations. Model convergence occurred stably during training, reaching early stopping at epoch 70 without signs of overfitting.

Detailed breakdowns of correct and incorrect classifications can be observed in Table II, expressing the data for each test patient. In addition, Figure 6 represents the confusion matrix showing the total distribution regarding the patients as a test set. The model correctly classified 1,422 normal rhythm segments as control, with only 30 false positives. For the high-risk class, the network accurately identified 713 pathological segments. Notably, only 43 false negatives occurred (risk cases misclassified as normal), reinforcing the clinical safety of the proposed methodology for automated SCD screening. With these findings, sensitivity reaches 94.31% for high-risk findings and 97.91% for normal findings.

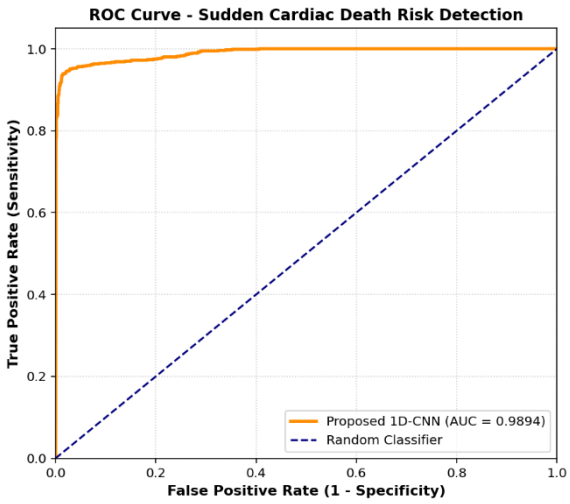


Fig. 5. ROC Curve. Source: the authors (2025)

TABLE II. INDIVIDUAL CLASSIFICATION PERFORMANCE ACROSS INDEPENDENT TEST PATIENTS EVALUATED.

Patient ID	Total segments	Sensitivity	Specificity	Accuracy
P01	540	96.00%	96.47%	96.30%
P02	576	89.77%	99.00%	96.18%
P03	552	97.03%	97.43%	97.28%
P04	540	93.82%	98.62%	97.04%
Total / Overall	2,208	94.31%	97.93%	96.69%

Although overall accuracy reached 96.69%, this was obtained under an independent patient-wise validation protocol, which is more rigorous than random segment splitting. Thus, reported performance reflects the model's ability to generalize to unobserved individuals, aligning experimental setup with real clinical application.

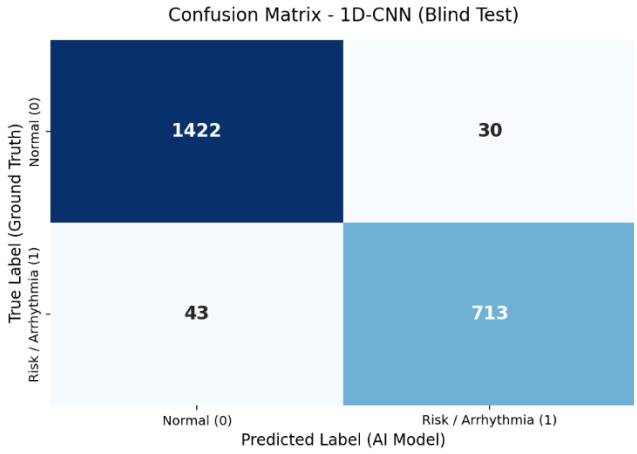


Fig. 6. 1D-CNN confusion matrix. Source: the authors (2025)

V. DISCUSSION

This study aims to construct an artificial neural network for identifying high risk of sudden cardiac death. The use of ECG for identifying SCD parameters is well documented in the literature [18], [19].

Electrocardiogram monitoring has increased substantially in recent years due to its ability to provide an overview of mechanical heart function using low-cost portable devices outside hospital settings. However, its application for sudden death risk detection remains limited due to the wide variety of cardiac alterations indicating sudden death, such as T-wave alternans, ST-segment duration, and late potentials [34].

Findings from this study align with Wang et al. [17], who predicted sudden death on the same database using a Nested LSTM combined with wavelet denoising. Similarly, Telangore et al. [18] proposed a fusion of 1D-CNN and LSTM networks for raw ECG extraction and SCD prediction. Collectively, these studies and our findings reinforce the growing trend toward artificial neural networks for sudden cardiac death prediction, serving as diagnostic support.

When comparing the findings with the literature is not simple, because there are differences in protocols, tasks and architectures. Khan et al. [19] reported a sensitivity of 92,41% using a 1D-CNNs directly on raw ECG signals using a deep residual architecture (ResNet) for multi-class classification of five arrhythmia types from the MIT-BIH database. His sensitivity is lower than the related in this work, but his evaluation is more complex because of the arrythmia classification. The performance achieved in this work highlights the efficacy of the proposed 1D-CNN for the specific binary screening task of SCD risk within the Sudden Cardiac Death Holter Database.

Some limitations should be acknowledged. First, the dataset used is restricted to 19 patients available on a public database and four patients were used to the blind test. Future work will focus on incorporating Group K-fold cross-validation and evaluating the architecture on external ECG databases with different recording equipment to ensure broader clinical generalization.

Increasing the number of patients and databases would add robustness to the neural network. This study focuses primarily on binary classification between "normal" and "high risk", limiting classification regarding specific cardiac arrhythmias. Besides thar, future studies should expand validation to different databases and analyze the model's predictive capacity at time windows progressively further removed from the fatal event (e.g., 60, 30, and 5 minutes before cardiac arrest).

In summary, findings demonstrate that the proposed neural network represents a promising alternative for detecting high-risk SCD ECG signals, serving as a potential auxiliary tool in clinical diagnosis.

VI. CONCLUSIONS

This work presented a robust and computationally efficient approach based on One-Dimensional Convolutional Neural Networks (1D-CNN) for early detection and risk classification of Sudden Cardiac Death (SCD) using single-lead ECG signals. Unlike traditional clinical approaches based on manual static metrics, the proposed model performs end-to-end learning, capturing complex dynamic morphological variations directly from filtered signals.

By adopting a rigorous inter-patient split validation protocol, the system was validated on individuals completely unseen by the network, overcoming common data leakage bias and accurately simulating a real hospital screening scenario.

On the independent test set, the architecture achieved an overall accuracy of 96.69%, with a sensitivity of 94.31% (a critical factor in mitigating false negatives in high-risk conditions), a specificity of 97.93%, and an AUC-ROC of 0.9894.

Future work includes validating architecture generalization on larger and more diverse clinical databases, including different heart disease etiologies. Additionally, an artificial intelligence model will be developed to indicate the probability of a patient being high-risk while specifying heart rhythm characteristics. In parallel, a low-cost Holter device will be developed to transmit captured signals to an online portal.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] D. Mozaffarian et al., "Heart Disease and Stroke Statistics — 2015 Update," Circulation, vol. 131, no. 4, Jan. 2015.

[2] S. Raj, K. Maurya, and K. C. Ray, "A knowledge-based real time embedded platform for arrhythmia beat classification," Biomedical Engineering Letters, vol. 5, no. 4, pp. 271–280, 2015.

[3] N. S. Priya and R. Balakrishnan, "A novel algorithm for sudden cardiac death risk estimation using LabVIEW," in Proc. Int. Conf. on Recent Trends in Information Technology (ICRTIT), IEEE, Jun. 2011.

[4] T. W. Ow et al., "SoC-based design of arrhythmia detector," in Proc. 2nd Int. Conf. on Electronic Design (ICED), IEEE, Aug. 2014.

[5] Bronzino, J.D. (Ed.). Biomedical Engineering Fundamentals (1st ed.). CRC Press. 2006. https://doi.org/10.1201/9781420003857

[6] R. M. Rangayyan, Biomedical Signal Analysis, 2nd ed., vol. 1. Hoboken: John Wiley & Sons, 2015.

[7] A. Kumar, D. Berwal, and Y. Kumar, "Design of High-Performance ECG Detector for Implantable Cardiac Pacemaker Systems using Biorthogonal Wavelet Transform," Circuits, Systems, and Signal Processing, vol. 37, no. 9, pp. 3995–4014, 2018.

[8] Y. Ma et al., "Novel Real-Time FPGA-Based R-Wave Detection Using Lifting Wavelet," Circuits, Systems, and Signal Processing, vol. 35, no. 1, pp. 281–299, 2016.

[9] B. U. Köhler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," IEEE Engineering in Medicine and Biology Magazine, vol. 21, no. 1, pp. 42–57, 2002.

[10] Wołk K, Niklewski J, Tatara MS, Kopczyński M. Multimodal Wearable Biosensing and Edge AI for Personalized Health: A Comprehensive Review. *Electronics*. 2026; 15(14):3237. https://doi.org/10.3390/electronics15143237

[11] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[12] D. Boutana, M. Benidir, and F. Bouaziz, "Multiresolution wavelet-based QRS complex detection algorithm suited to several abnormal morphologies," IET Signal Processing, vol. 8, no. 7, pp. 774–782, Sep. 2014.

[13] E. M. El Hassan and M. Karim, "An FPGA-based implementation of a pre-processing stage for ECG signal analysis using DWT," in Proc. 2nd World Conf. on Complex Systems (WCCS), 2014.

[14] Yun, D., Lee, HC., Jung, CW. *et al.* Robust R-peak detection in an electrocardiogram with stationary wavelet transformation and separable convolution. *Sci Rep* **12**, 19638 (2022). https://doi.org/10.1038/s41598-022-19495-9

[15] P. N. Ranaware and R. A. Deshpande, "Detection of arrhythmia based on discrete wavelet transform using artificial neural network and support vector machine," in Proc. Int. Conf. on Communication and Signal Processing (ICCSP), IEEE, Apr. 2016.

[16] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019 Jan;25(1):65-69. doi: 10.1038/s41591-018-0268-3. Epub 2019 Jan 7. Erratum in: Nat Med. 2019 Mar;25(3):530. doi: 10.1038/s41591-019-0359-9. PMID: 30617320; PMCID: PMC6784839.

[17] Wang K, Zhang K, Liu B, Chen W, Han M. Early prediction of sudden cardiac death risk with Nested LSTM based on electrocardiogram sequential features. BMC Med Inform Decis Mak. 2024 Apr 10;24(1):94. doi: 10.1186/s12911-024-02493-4. PMID: 38600479; PMCID: PMC11005267.

[18] Hardik Telangore, Victor Azad, Manish Sharma, Ankit Bhurane, Ru San Tan, U. Rajendra Acharya, Early prediction of sudden cardiac death using multimodal fusion of ECG Features extracted from Hilbert–Huang and wavelet transforms with explainable vision transformer and CNN models, Computer Methods and Programs in Biomedicine, Volume 257, 2024, 108455, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2024.108455

[19] Fahad Khan, Xiaojun Yu, Zhaohui Yuan, Atiq ur Rehman. ECG classification using 1-D convolutional deep residual neural network

[20] Frownfelter, Donna L, and Elizabeth Dean. Cardiovascular and Pulmonary Physical Therapy : Evidence to Practice. 5th ed. St. Louis, Mo: Elsevier/Mosby, 2012.

[21] E. C. S. Dantas, C. D. M. Regis e L. G. Caldeira. Amostragem compressiva em sinais de ECG e imagens digitais. XXIV Congresso Brasileiro de Engenharia Biomédica, 2014.

[22] Namjouyan K, Sejdić E, Link MS, Pelliccia A, Glicksberg B, Trayanova N, Krittanawong C. Precision Medicine for Electrocardiogram Interpretation: Clinical Relevance, Challenges, and Advances. Rev Cardiovasc Med. 2025 Dec 24;26(12):47007. doi: 10.31083/RCM47007. PMID: 41524056; PMCID: PMC12781007.

[23] de Luna AB, Coumel P, Leclercq JF (1989) Ambulatory sudden cardiac death: mechanisms of production of fatal arrhythmia on the basis of data from 157 cases. American Heart Journal 117(1):151. https://doi.org/10.1016/0002-8703(89)90670-4

[24] Wong CX, Brown A, Lau D, Chugh SS, Albert C, Kalman J et al (2019) Epidemiology of sudden cardiac death: global and regional perspectives. Heart Lung Circ 28(1):6. https://doi.org/10.1016/j.hlc.2018.08. 026

[25] Viskin S, Rosovski U, Sands AJ, Chen E, Kistler PM, Kalman JM, Rodriguez Chavez L, Iturralde Torres P, Cruz F FE, Centurión OA, Fujiki A, Maury P, Chen X, Krahn AD, Roithinger F, Zhang L, Vincent GM, Zeltser D. Inaccurate electrocardiographic interpretation of long QT: the majority of physicians cannot recognize a long QT when they see one. Heart Rhythm. 2005 Jun;2(6):569-74. doi: https://doi.org/10.1016/j.hrthm.2005.02.011. PMID: 15922261.

[26] Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2016; ISBN 0262337371.

[27] Banerjee, I.; Ling, Y.; Chen, M.C.; Hasan, S.A.; Langlotz, C.P.; Moradzadeh, N.; Chapman, B.; Amrhein, T.; Mong, D.; Rubin, D.L. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif. Intell. Med. 2019, 97, 79–88.

[28] Guessoum, S.; Belda, S.; Ferrandiz, J.M.; Modiri, S.; Raut, S.; Dhar, S.; Heinkelmann, R.; Schuh, H. The Short-Term Prediction of Length of Day Using 1D Convolutional Neural Networks (1D CNN). *Sensors* **2022**, *22*, 9517. https://doi.org/10.3390/s22239517

[29] Imambi, S., Prakash, K.B., Kanagachidambaresan, G.R. (2021). PyTorch. In: Prakash, K.B., Kanagachidambaresan, G.R. (eds) Programming with TensorFlow. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-57077-4_10

[30] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

[31] Greenwald SD. Development and analysis of a ventricular fibrillation detector. M.S. thesis, MIT Dept. of Electrical Engineering and Computer Science, 1986.

[32] Abdullah, Fatima Z, Abadian A, Mejorada CGS, Ruiz MJT, Téllez RQ. Integration of Machine Learning Techniques in ECG-Based Multiclass Arrhythmia Classification with Explainability Analysis. *Biosensors*. 2026; 16(6):326. https://doi.org/10.3390/bios16060326

[33] J. Malik, O. C. Devecioglu, S. Kiranyaz, T. Ince and M. Gabbouj, "Real-Time Patient-Specific ECG Classification by 1D Self-Operational Neural Networks," in IEEE Transactions on Biomedical

Engineering, vol. 69, no. 5, pp. 1788-1801, May 2022, doi: https://doi.org/10.1109/TBME.2021.3135622.

[34] D. Lai, Y. Zhang, X. Zhang, Y. Su and M. B. Bin Heyat, "An Automated Strategy for Early Risk Identification of Sudden Cardiac Death by Using Machine Learning Approach on Measurable Arrhythmic Risk Markers," in IEEE Access, vol. 7, pp. 94701-94716, 2019, doi: https://doi.org/10.1109/ACCESS.2019.2925847.

Robotic Prosthesis Control using Linear Discriminant Analysis of Residual Electromyographic Signals

Carneiro, H. A. B.
Federal University of Pará (UFPA)
Belém, Pará, Brazil
heloisacarneiro70@gmail.com

Sousa, M. L. L.
Federal University of Pará (UFPA)
Belém, Pará, Brazil
maria.lisboa.sousa22@gmail.com

Farias, E. A. S.
Federal University of Pará (UFPA)
Belém, Pará, Brazil
eduafsantos@gmail.com

Sales, I. Q.
Federal University of Pará (UFPA)
Belém, Pará, Brazil
isasales40@gmail.com

Fonseca, M. C. P.
Federal University of Pará (UFPA)
Belém, Pará, Brazil
conceicao@ufpa.br

Miziara, I. M.
Federal University of Pará (UFPA)
Belém, Pará, Brazil
immiziara@ufpa.br

***Abstract*—The widespread adoption of myoelectric upper-limb prostheses is limited by the complexity, accessibility, and control requirements of commercially available devices, factors that contribute to prosthesis abandonment. This work presents an accessible myoelectric prosthesis prototype built by 3D printing from an open-source design, using readily available materials and electronic components, and conceived as an experimental platform for the development of prosthetic systems for individuals with upper-limb amputation. The control strategy accommodates different muscle groups, including residual musculature, provided that training enables the generation of distinguishable electromyographic activation patterns, with preference given to residual muscles naturally associated with the movements previously performed by the amputated limb. Surface electromyographic (EMG) signals were collected from four volunteers to investigate the identification of muscle activation patterns associated with prosthetic hand opening and closing. A real-time acquisition, processing, and classification framework was developed based on time-domain EMG features and Linear Discriminant Analysis (LDA), with performance evaluated through Leave-One-File-Out (LOFO) and Leave-One-Subject-Out (LOSO) validation. LOFO achieved an overall accuracy of 89.9%, with recall of 81.4%, precision of 81.9%, and an F1-score of 81.7%, whereas LOSO resulted in an overall accuracy of 89.4%, with recall of 79.0%, precision of 82.1%, and an F1-score of 80.5%. The proposed control methodology was experimentally integrated with a 3D-printed, servomotor-actuated robotic hand, which reproduced the opening and closing commands identified from the EMG signals. Although the results in able-bodied volunteers were promising, validation with amputee participants, assessment of residual muscle recruitment under realistic prosthetic conditions, and embedded implementations of the control system remain as future work.**

***Index Terms*—Myoelectric prosthesis, surface electromyography, pattern recognition, linear discriminant analysis, 3D printing, assistive technology.**

I. Introduction

Over the past decades, upper-limb prostheses have become an important resource for the rehabilitation of individuals with limb loss, contributing to the recovery of functional independence in activities of daily living. In Brazil, the demand for this type of assistive technology is substantial: according to the 2010 Demographic Census, more than 470,000 Brazilians reported some form of amputation, of whom approximately 94,000 had upper-limb amputations [1]. Despite this demand, access to functional upper-limb prosthetic devices remains limited. Within the Brazilian Unified Health System (SUS), prosthetic care is provided through national policies addressing orthoses, prostheses, and mobility aids [2]. Nevertheless, limitations related to availability, technological complexity, maintenance, adaptation, and access to more advanced prosthetic systems still restrict their use [3], [4]. Prosthesis abandonment is also an important challenge, and previous studies have associated device discontinuation with discomfort, excessive weight, limited functionality, muscle fatigue, and difficulties in prosthesis control [5], [6]. Myoelectric prostheses use electromyographic (EMG) signals generated by muscle contractions to produce commands for prosthetic movements, potentially providing a more direct relationship between muscle activation and device operation than passive or purely mechanical systems [7], [8]. However, this control approach depends on the availability of muscles capable of producing distinguishable EMG patterns and on the user's ability to activate them consistently, requirements that are particularly relevant in individuals with upper-limb amputation, whose remaining musculature may vary considerably in anatomical and functional characteristics. In this context, the present study focuses on the development of a myoelectric prosthesis prototype based on an open-source robotic hand manufactured by 3D printing. The open-source design, readily available electronic components, and relatively simple actuation architecture were adopted to reduce the technological and manufacturing complexity of the prototype while increasing its accessibility and reproducibility. EMG signals were acquired from two antagonistic forearm muscles involved in finger movements: the flexor digitorum superficialis (FDS), associated with finger flexion, and the extensor digitorum (ED), associated with finger extension. These muscles served as a methodological model for investigating whether antagonistic activation patterns could be identified and classified to control the opening and closing commands of the robotic

hand; in future applications with amputee participants, the same strategy is intended to be applied to functional residual muscles associated with the original movements of the missing hand. Therefore, the objective of this study was to develop a myoelectric upper-limb prosthesis prototype by fabricating a 3D-printed robotic hand based on an open-source design, developing a system for the acquisition and conditioning of surface EMG signals, and implementing real-time EMG signal processing and classification to identify muscle activation patterns associated with hand opening and closing commands. Classification performance and robotic prototype operation were experimentally assessed in four non-amputee volunteers as an initial investigation of the proposed strategy, which does not constitute clinical validation of the prosthetic system. Future studies involving individuals with upper-limb amputation are required to determine its applicability under real-world prosthetic-use conditions.

II. Methodology

The development of the prototype followed sequential stages: mechanical fabrication, signal acquisition hardware, experimental data collection, signal processing, classification, and robotic actuation. Four healthy volunteers without upper-limb amputation were recruited, and surface EMG signals were acquired from the FDS and ED muscles while participants performed predefined hand opening and closing tasks. The acquired signals were represented by time-domain features and classified using an LDA classifier, whose performance and generalization were assessed using LOFO and LOSO cross-validation, and the classification output was applied to the robotic hand to convert the detected EMG patterns into opening and closing commands of the prototype.

A. Development of the EMG acquisition circuit

An analog electronic circuit dedicated to the acquisition and conditioning of electromyographic signals was designed and implemented using a modular architecture composed of three printed circuit boards (PCBs): a motherboard, which incorporates the active reference circuit shared by the system, and two identical daughterboards, each responsible for the complete signal-conditioning chain of one independent acquisition channel. This architecture enables the simultaneous acquisition of EMG signals from two forearm muscles, providing complementary information on muscle activation and improving the discrimination between rest and contraction states. Each

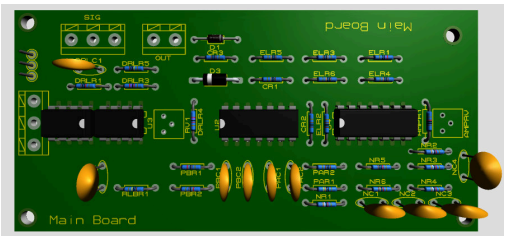


Figure 1. EMG printed circuit board.

acquisition channel incorporates an AD620 instrumentation amplifier, selected for its high input impedance and high common-mode rejection ratio (CMRR), which are particularly important for acquiring low-amplitude biopotentials while minimizing external electromagnetic interference. Gain adjustments based on a fixed resistor and a trimmer accommodate the variability of the EMG signal amplitude across subjects and electrode–skin contact impedances, and a Driven Right Leg (DRL) active feedback circuit feeds the inverted common-mode component back to the subject through a dedicated reference electrode, actively attenuating 60 Hz power-line interference [**?**]. Following the initial amplification stage, the signal passes through a second-order Butterworth band-pass filter with cutoff frequencies of 20 Hz and 250 Hz, which preserves the spectral content of interest of the EMG signal while attenuating motion artifacts and high-frequency noise; the 250 Hz low-pass cutoff also serves as the antialiasing filter for the subsequent analog-to-digital conversion. A second amplification stage is subsequently applied to maximize the effective resolution of the analog-to-digital converter (ADC). Since the EMG signal is bipolar, whereas the ESP32 ADC operates exclusively with positive input voltages, a level-shifting circuit was incorporated to adjust the signal to the ADC input range, establishing a 1.5 V reference around which the conditioned signal oscillates. The conditioned signals were then digitized by the ESP32 internal 12-bit ADC, providing 4096 quantization levels, at a sampling frequency of 1000 Hz for each EMG channel, and a protection circuit prevents overvoltage at the ESP32 analog inputs.

B. Construction and assembly of the robotic prosthesis

The mechanical prototype consisted of a humanoid robotic hand fabricated by additive manufacturing (3D printing) from an open-source design [9]. This choice enables unrestricted reproduction of the mechanical structure, facilitating its use in research and in the development of accessible assistive technologies, while additive manufacturing reduces fabrication complexity, enables device customization, and allows the production of lightweight prosthetic structures at lower manufacturing cost. The assembly process included the installation of the phalanges, the articulation of the interphalangeal joints, and the routing of the actuation tendons through the internal channels provided in the original mechanical design. Finger actuation follows a tendon-driven mechanism that reproduces the operating principle of the extrinsic musculature of the human hand. Each finger is actuated by a 0.40 mm diameter nylon tendon routed through the internal channels of the phalanges and connected to its corresponding actuator, while finger extension is achieved passively by elastic elements positioned on the opposite side of each finger, which act as restoring springs. This antagonistic configuration enables bidirectional finger motion—active flexion and passive extension—using a single actuator per finger. Actuation was performed by five SG90 servomotors, one assigned to each finger, selected for their compact dimensions, affordability, and pulse-width modulation (PWM) control interface, which is directly compatible with the digital outputs of the ESP32. To transmit motion from each

servomotor to its corresponding tendon, a smooth pulley was designed in Autodesk Fusion and fabricated by 3D printing: mounted on the servomotor shaft, it secures the tendon end and converts the servo angular displacement into linear tendon traction, producing finger flexion during rotation in the closing direction, whereas rotation in the opposite direction releases the tendon and allows the elastic element to restore finger extension.

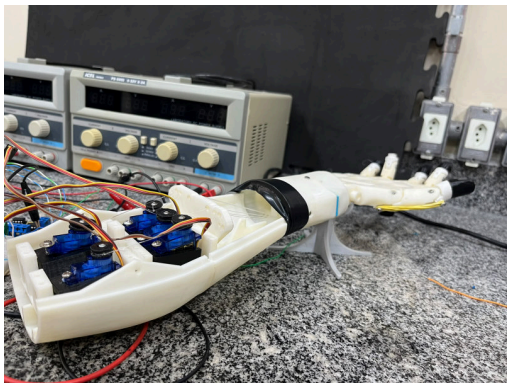


Figure 2. Assembled humanoid robotic prosthesis.

C. Power supply and servomotor actuation

The electrical architecture used for powering and controlling the prototype is shown in Fig. 3. The signals from the two EMG acquisition channels were connected to the ESP32 analog inputs, while the microcontroller communicated with the computer through the USB interface and generated the PWM signals responsible for controlling the five servomotors, configured at a frequency of 50 Hz with pulse widths ranging from 500 μs to 2400 μs. An angular position of 180° corresponded to the fully open hand, whereas 0° corresponded to the fully closed hand. The same angular command was sent to all servomotors, producing synchronized opening and closing movements according to the state identified by the classification system. The servomotors were powered independently from the ESP32 by a regulated 5 V supply capable of providing at least 3 A, since preliminary tests showed that the power supplied through the development board was insufficient for the simultaneous actuation of all five servomotors, resulting in unstable system operation. A 1000 μF electrolytic capacitor was connected in parallel with this supply to reduce voltage fluctuations during transient current demands. The EMG acquisition boards were powered by a separate 9 V supply, while the ESP32 was powered through the USB connection, and the grounds of all supplies were connected to a common point in a star-ground configuration, as illustrated in Fig. 3.

D. EMG acquisition protocol and electrode placement

The data collection procedures were conducted after approval by the Human Research Ethics Committee of the João de Barros Barreto University Hospital Complex at the Federal University of Pará, and the experiments were carried out at eBio, the Biomedical Engineering Laboratory Complex of

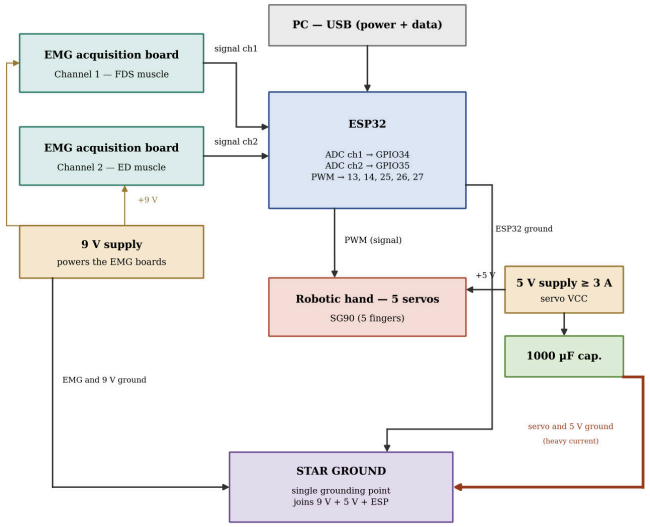


Figure 3. Electrical architecture of the EMG-controlled robotic hand system.

the School of Electrical and Biomedical Engineering at the Federal University of Pará. Four healthy adult volunteers over 18 years of age, without upper-limb amputation and without self-reported conditions that could interfere with the execution of the proposed motor tasks, were recruited for the study. Before data collection, each participant received information about the objectives of the study and the procedures involved in the experimental protocol, and those who agreed to participate voluntarily signed the Informed Consent Form prior to the beginning of the experiments. During data acquisition, participants were seated in a comfortable chair in an environment with controlled temperature and were instructed to remain as relaxed as possible between the proposed tasks. Personal information and the identity of all volunteers were kept confidential throughout the study, and the collected data were handled in a manner that preserved participant anonymity. The first acquisition channel was positioned over the FDS, on the anterior aspect of the forearm, whereas the second channel was positioned over the ED, on the posterior aspect. Because these muscles perform antagonistic functions in finger flexion and extension, respectively, they produce distinct activation patterns, which favors discrimination between hand opening and closing movements [10]. The surface electrodes were positioned over the muscle bellies, aligned parallel to the orientation of the muscle fibers; before electrode placement, the skin was cleaned to reduce electrode–skin contact impedance, and disposable electrodes were replaced at each acquisition session. Electromyographic signals were acquired at a sampling frequency of 1000 Hz, selected according to the Nyquist criterion to ensure a sampling rate greater than twice the highest frequency of interest, defined as 250 Hz by the analog low-pass filter cutoff, providing an adequate representation of the EMG spectral content while reducing the risk of aliasing

during analog-to-digital conversion. In addition to the two EMG acquisition channels, a third channel recorded the state of a light-emitting diode (LED) used as a synchronization signal: LED activation indicated the period during which the volunteer was instructed to perform the muscle contraction, allowing contraction and rest intervals to be automatically identified during subsequent signal processing. The experimental protocol began with a 30 s resting period for signal stabilization and baseline EMG recording, followed by 10 acquisition cycles, each consisting of 5 s of voluntary muscle contraction, indicated by LED activation, and 10 s of rest, during which the LED remained off, as illustrated in Fig. 4.

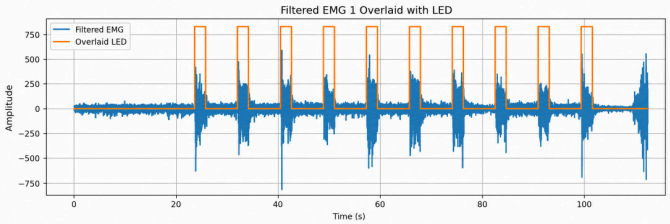


Figure 4. EMG acquisition protocol showing the initial 30 s baseline followed by 10 cycles of 5 s contraction and 10 s rest, synchronized by the LED signal.

E. Preprocessing of the electromyographic signals

Signal preprocessing began with the removal of the direct current (DC) component through subtraction of the mean value of each signal, centering the waveform around zero. A fourth-order Butterworth digital band-pass filter with cutoff frequencies of 20 Hz and 200 Hz was then applied to restrict the analysis to the frequency band most relevant for the discrimination of contraction patterns, followed by notch filters centered at 60 Hz, 120 Hz, and 180 Hz, each with a quality factor of 30, to attenuate power-line interference and its harmonics. The filtered signals were segmented into sliding windows of 250 ms (250 samples), with a window shift of 50 ms (50 samples) and an overlap of 80%. This configuration provides improved temporal resolution while preserving a sufficient number of samples for reliable feature extraction. The LED channel was used exclusively as a temporal synchronization marker for the automatic labeling of the EMG windows. For each window, the mean value of the LED signal was computed: windows with a mean value greater than or equal to 0.8 were labeled as muscle contraction, indicating that the LED remained active during nearly the entire window, whereas windows with a mean value lower than or equal to 0.2 were labeled as rest, corresponding to periods in which the LED remained off. Windows with intermediate mean values were discarded because they contained transitions between the rest and contraction states.

F. Feature extraction and classification

Six time-domain features were extracted from each EMG window: root mean square (RMS), mean absolute value (MAV), variance (VAR), waveform length (WL), zero crossings (ZC), and slope sign changes (SSC), as defined in [10]. Since two EMG acquisition channels were employed, each window was represented by a 12-dimensional feature vector. To reduce the influence of residual noise, the thresholds for the computation of ZC and SSC were defined adaptively for each acquisition channel as three times the standard deviation of the corresponding filtered signal. The features from both channels were then concatenated into a single feature vector and standardized using the mean and standard deviation computed exclusively from the training set. Classification between the rest and muscle contraction classes was performed using LDA, which determines a linear projection that maximizes class separability, with the decision function given by

$$s(\mathbf{z}) = \mathbf{w}^T \mathbf{z} + b. \tag{1}$$

Because of the class imbalance, balanced prior probabilities were adopted to reduce the tendency of the classifier to favor the majority class. Two cross-validation strategies were employed. In LOFO validation, one complete acquisition file was held out for testing at each iteration, while the remaining files were used for training, allowing the classifier performance to be assessed on recordings or acquisition sessions that did not participate in the respective training. In LOSO validation, all records belonging to one participant were completely excluded from training and used as the test set at each iteration, with the process repeated until each participant had been used once as the test set, directly evaluating the ability of the model to recognize the rest and contraction states in an individual whose data were not used during training. In both strategies, feature standardization and LDA training were performed again at each iteration, using exclusively the data belonging to the training set. Performance was quantified by means of accuracy, recall, precision, F1-score, and the confusion matrix.

G. Real-time classification and control

To further assess the real-time behavior of the classification system, additional tests were performed using an acquisition file held out from training, following the LOFO strategy. The recorded EMG signals were processed as a simulated data stream using 250 ms windows updated every 50 ms, with classification stability improved through majority voting over five consecutive predictions. Processing latency was measured for each classification window as the interval between the start of signal processing and completion of the LDA prediction, using a high-resolution timer. Mean, median, standard deviation, minimum, maximum, and 95th and 99th percentile latency values were subsequently calculated. Classification errors and correct detections were evaluated by comparing the classifier output with the LED synchronization channel, used as the reference signal. Predictions were categorized as true positive, true negative, false positive, or false negative. In addition, contraction and rest transitions were analyzed as events, allowing the calculation of the detection rate, missed events, and isolated false-positive activations.

III. Results

A. Analysis of the RMS of the EMG channels

The RMS quantifies the signal energy within each analysis window and is widely used to estimate the level of muscle activation, with lower values typically associated with resting conditions and pronounced increases indicating voluntary muscle contractions. The EMG1 channel exhibits higher activation peaks and greater dynamic variation during contraction periods, reaching values close to 200, whereas EMG2 also presents clear increases during muscle activation, although with lower amplitudes, reaching approximately 120. These differences are expected and may be attributed to factors such as electrode placement, the anatomical characteristics of the monitored muscles, and variations in muscle recruitment between the two acquisition sites. During resting periods, the RMS of EMG1 remains predominantly between 10 and 25, whereas EMG2 exhibits slightly higher baseline values, generally ranging from 20 to 40, a behavior that may be associated with residual electronic noise, low-level involuntary muscle activity, or differences in the baseline activation of the monitored muscles. Despite these differences, both channels exhibit synchronous increases during the contraction intervals, demonstrating a clear temporal correspondence between muscle activation and the increase in RMS, which provides a reliable representation of muscle activation and contributes to the discrimination between rest and contraction states.

B. Classification performance under LOFO and LOSO validation

The ability of the classifier to maintain its performance across acquisition records not used during training was evaluated using the LOFO strategy, in which each complete recording was used once as the test set while the remaining recordings were used for training. Across nine valid folds, the classifier achieved an overall accuracy of 89.9%, recall of 81.4%, precision of 81.9%, and an F1-score of 81.7%. The accumulated confusion matrix showed 37,611 rest windows and 12,546 contraction windows correctly classified, whereas 2,765 rest windows were incorrectly classified as contraction and 2,870 contraction windows were classified as rest. Inter-subject generalization was evaluated using the LOSO strategy, in which the recordings associated with one participant were used exclusively for testing while the remaining data were used for training. The LOSO evaluation achieved an overall accuracy of 89.4%, recall of 79.0%, precision of 82.1%, and an F1-score of 80.5%. The accuracies obtained in the three validation folds were 86.6%, 93.0%, and 88.8%, indicating relatively consistent performance across the evaluated subjects. The accumulated confusion matrix contained 37,722 correctly classified rest windows and 12,176 correctly identified contraction windows, with 2,654 false-positive and 3,240 false-negative classifications. The results of both validation strategies are summarized in Table I.

Table I
Classification performance under LOFO and LOSO validation.

Metric	LOFO	LOSO
Accuracy	0.899	0.894
Recall	0.814	0.790
Precision	0.819	0.821
F1-score	0.817	0.805

C. Temporal Evaluation of Contraction and Rest Commands

The temporal performance of the system was further evaluated at the event level, using the LED signal as a reference to determine the onset and end of each contraction period. A total of 30 contraction-onset events (ON) and 30 return-to-rest events (OFF) were analyzed, resulting in 60 expected transitions. All 30 ON events were correctly identified, corresponding to a detection rate of 100% with no missed contraction events, which indicates high sensitivity of the system in detecting the onset of muscle activation. For the OFF events, corresponding to the transition from contraction to rest, 12 of the 30 events were correctly detected, whereas 18 were not recognized according to the adopted temporal criterion, resulting in a detection rate of 40%. Considering the ON and OFF transitions together, 42 of the 60 expected events were correctly identified, corresponding to an overall detection rate of 70%. No isolated false activations were observed during the analyzed period, resulting in a rate of 0 false positives per minute. The results reveal an asymmetric behavior of the system in identifying state transitions: while contraction-onset detection showed consistent performance, recognition of the transition back to rest was less effective, indicating the need to optimize the logic used to confirm the OFF state.

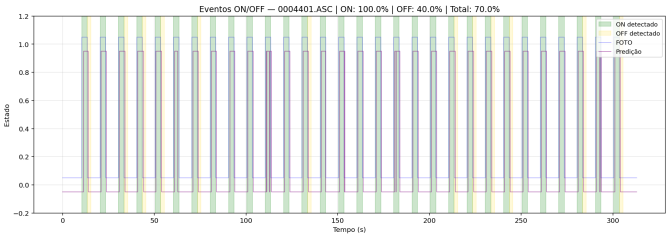


Figure 5. ON/OFF event detection during the temporal evaluation of contraction and rest commands.

D. Computational Latency of Real-Time Processing

The computational latency of the processing and classification pipeline was evaluated over 6,255 consecutive windows, yielding a mean processing time of 3.3 ms, a median of 2.8 ms, and a standard deviation of 1.4 ms, with minimum and maximum recorded times of 2.4 ms and 45.1 ms, respectively, and 95th and 99th percentiles of 5.5 ms and 7.6 ms. The distribution of the latency values is illustrated in Fig. 6. The proximity between the mean and median values, together with the 95th percentile of 5.5 ms, indicates that the vast majority of windows were processed within a few milliseconds, despite

the occurrence of isolated higher-latency peaks. Considering that each analysis window had a duration of 250 ms, the computational processing time was substantially shorter than the duration of the window itself, indicating that the feature extraction, standardization, and LDA classification stages impose a computational burden compatible with real-time operation. However, the reported values correspond specifically to the computational latency of the algorithm and not to the total delay between the onset of muscle contraction and the complete execution of the prosthetic movement, which also depends on acquisition-window filling, temporal confirmation logic, serial communication, and servomotor actuation time.

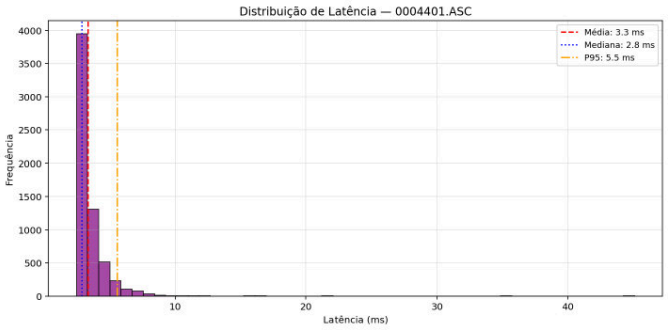


Figure 6. Distribution of computational latency during real-time processing.

E. Validation on the Humanoid Prosthesis

As the final experimental validation stage, the proposed control system was implemented on the humanoid robotic prosthesis, which was operated in real time using the electromyographic signals acquired from the user. During the experiments, the prosthesis consistently executed the opening and closing movements corresponding to the muscle activation patterns identified by the classifier, demonstrating the practical applicability of the proposed methodology. A short delay was observed between the user's voluntary muscle contraction and the execution of the corresponding prosthetic movement, primarily associated with the cumulative processing time required for signal acquisition, preprocessing, feature extraction, classification, temporal confirmation, serial communication, and servomotor actuation. Despite this delay, the system operated reliably throughout the experiments, with stable command recognition and consistent prosthesis actuation, experimentally demonstrating the feasibility of the proposed approach for real-time control of humanoid myoelectric prostheses.

IV. Discussion

A. Classification performance and inter-subject generalization

The results obtained in this study indicate that a relatively compact myoelectric control architecture, based on two EMG channels, time-domain features, and LDA, can discriminate between rest and muscle contraction while maintaining relatively low computational complexity. The use of time-domain descriptors for myoelectric pattern recognition is well established in the literature: features related to signal amplitude, temporal variability, zero crossings, and changes in signal slope provide useful representations of muscle activation while requiring limited computational resources [10], [11], characteristics that make them particularly suitable for real-time prosthetic control systems and future embedded implementations [7], [12]. The LOFO and LOSO evaluations (Table I) achieved overall accuracies of 89.9% and 89.4%, respectively, indicating that the classifier maintained its ability to distinguish between contraction and rest across complete acquisition records and subjects not included during training. The small difference between the two validation strategies suggests that relevant characteristics of the EMG activation patterns were preserved under inter-subject evaluation, although the reduction in recall relative to overall accuracy indicates that identifying the contraction class remains more challenging than identifying rest. This behavior may be related to inter-individual differences in anatomy, muscle recruitment strategies, electrode positioning, contraction intensity, subcutaneous tissue characteristics, and physiological variability [6], [7], and Fougner et al. [13] demonstrated that variations in limb position can substantially affect EMG classification performance. Therefore, although the LOSO results are promising, future applications may benefit from short subject-specific calibration procedures or adaptive strategies capable of compensating for differences in EMG feature distributions between users. These values are within the range reported in previous studies employing time-domain features and linear classification methods, although direct comparisons should be made cautiously because of differences in acquisition protocols, number of channels, movement classes, participant characteristics, and validation strategies. Wang et al. [14], for example, employed time-domain EMG features and LDA and reported classification accuracies between approximately 96.9% and 98.4%, depending on the feature combination, whereas Koppolu and Chemmangat [15] reported an LDA accuracy of 93.63% after artifact reduction in a multiclass protocol involving three acquisition channels. The lower accuracy observed in the present study compared with some highly controlled offline investigations should not necessarily be interpreted as inadequate classifier performance, since evaluations in which training and test windows originate from the same acquisition session generally represent a less demanding condition than evaluations involving complete independent recordings or previously unseen subjects [6], [12]. Therefore, the LOFO and LOSO results provide a more informative estimate of the generalization capability of the proposed classifier.

B. Decision threshold and number of channels

An additional analysis investigated the influence of the LDA decision threshold on the balance between contraction detection and false-positive classifications. Increasing the decision offset progressively increased the recall of the contraction class, although this was accompanied by a reduction in precision. The highest F1-score was obtained using an offset of +1.00, at which recall increased to 88.9%, with a precision of 77.6%, an F1-score of 82.9%, and an accuracy of 89.9%. For prosthetic

control, this balance is particularly relevant because failing to detect an intended contraction and generating an unintended command may have different functional consequences; the selected operating point favored increased sensitivity to muscle activation while maintaining an overall accuracy close to 90%, although the optimal decision threshold may depend on the intended prosthetic application and should be further investigated under real-time functional conditions. Another relevant aspect of the proposed architecture is that classification was performed using only two EMG acquisition channels. Classical multifunction myoelectric control approaches frequently employ four or more recording channels [11], and although high-density EMG systems can provide additional spatial information, they require more electrodes, processing resources, and more complex electrode placement procedures. Daley et al. [16], for example, reported high-density EMG classification accuracies ranging from approximately 85% to 98% in normally limbed and transradial amputee participants, also suggesting that useful performance could be maintained with a reduced number of appropriately positioned electrodes, and Li and Kuiken [17] similarly demonstrated that residual forearm EMG contains relevant information for recognizing basic wrist and hand movements. In the present study, the use of two acquisition channels represents a deliberate compromise between classification performance and system complexity, consistent with the objective of developing a more accessible prototype. The choice of LDA follows the same principle: linear classifiers remain attractive for prosthetic applications because of their relatively low memory requirements, rapid inference, and straightforward implementation [11], [12], making the method suitable for future investigation in embedded platforms.

C. Real-time processing and event-based evaluation

The computational latency measurements indicate that the processing and classification stages impose a relatively small computational burden, as reported in the Results section. Previous studies have emphasized that real-time myoelectric control requires a compromise between classification performance and controller response time: Smith et al. [18] identified window durations between approximately 150 and 250 ms as an appropriate compromise for pattern-recognition-based myoelectric control. However, the mean computational latency of 3.3 ms should not be interpreted as the complete response time of the prosthetic system, since the total delay between muscle activation and mechanical movement also includes the time required to fill the analysis window, temporal confirmation of successive classifications, serial communication, and servomotor actuation [12], [19]. The event-based analysis provided complementary information regarding the temporal behavior of the classifier, revealing that the system was more effective in identifying the onset of muscle activation (100% ON detection) than the subsequent return to rest (40% OFF detection). One possible explanation is the temporal decision strategy adopted to increase command stability: strategies based on repeated classifications or majority voting can reduce isolated misclassifications, but may also delay state transitions when muscle relaxation occurs gradually, a trade-off between stability and responsiveness also described in previous studies of myoelectric control [20]. Importantly, no isolated false-positive activation was observed during the analyzed experiment, corresponding to 0 false-positive events per minute; although encouraging, this result should be interpreted as preliminary, and longer experiments involving more diverse motor tasks will be required to determine the frequency of unintended activations under conditions closer to actual prosthetic use. The lower performance observed for OFF transitions indicates that the decision logic used to identify the return-to-rest state should be further optimized, and independent confirmation criteria for contraction onset and relaxation may represent a useful strategy for future implementations.

D. Limitations and future application in individuals with amputation

An important limitation of the present study is that the experimental data were obtained from participants without upper-limb amputation. Although the prototype was conceived for future application in individuals with transradial limb loss, residual musculature may differ from intact musculature in terms of anatomy, available muscle volume, recruitment patterns, signal amplitude, and possible electrode positions. Maintaining a correspondence between the muscle activation used to generate a command and the intended prosthetic movement is conceptually relevant because it may reduce the need to associate an unrelated muscular contraction with a specific prosthetic action, and previous studies have demonstrated that residual forearm musculature can provide useful information for pattern-recognition-based prosthetic control [16], [17]; however, the present results do not demonstrate that the same electrode positions or activation patterns will be directly transferable to individuals with transradial amputation, and potential benefits such as reduced muscle fatigue, lower cognitive demand, or shorter training time should be considered hypotheses for future investigation rather than demonstrated advantages of the current prototype. Consequently, the current results should be interpreted as a preliminary assessment of the technical feasibility of the proposed acquisition and classification architecture rather than as clinical validation of a prosthetic system. Additional limitations include the relatively small participant sample, the binary classification between contraction and rest, and the controlled conditions under which the signals were acquired, since real-world prosthetic use introduces additional sources of variability, including changes in limb position, electrode displacement, muscle fatigue, changes in contraction intensity, and long-term physiological adaptation [6], [13]. Future studies should therefore include participants with upper-limb amputation, evaluate residual-muscle training and subject-specific calibration, investigate the system under different limb positions and electrode-placement conditions, optimize the detection of transitions from contraction to rest, and subsequently evaluate the implementation of the classification architecture in embedded systems.

V. Conclusion

This work presented the development and experimental assessment of an accessible myoelectric upper-limb prosthesis prototype based on a 3D-printed open-source robotic hand, integrating two-channel EMG acquisition and conditioning, time-domain feature extraction, LDA classification, and real-time robotic actuation. Overall accuracies of 89.9% and 89.4% were achieved under LOFO and LOSO validation, respectively (Table I), with decision-threshold adjustment further improving contraction detection (F1-score of 82.9% at an offset of +1.00). The event-based analysis showed reliable detection of contraction onset (100% of ON events), whereas the detection of the return to rest (40%) indicates that the transition logic should be further optimized. The computational processing stage presented a mean latency of 3.3 ms over 6,255 evaluated windows, compatible with real-time operation, although these values represent only the algorithmic latency and not the complete response time between muscle activation and prosthetic movement. Future studies should evaluate the prototype in individuals with upper-limb amputation, particularly by identifying and training residual muscles associated, whenever possible, with the original movements of the missing hand. Further investigations should also include a larger participant sample, optimization of contraction-to-rest detection, evaluation under different limb positions and electrode-placement conditions, and subsequent implementation of the control architecture in embedded systems.

References

[1] Instituto Brasileiro de Geografia e Estatística, *Censo Demográfico 2010: Características gerais da população, religião e pessoas com deficiência*. Rio de Janeiro, Brasil: IBGE, 2012.

[2] Brasil, Ministério da Saúde, *Guia para prescrição, concessão, adaptação e manutenção de órteses, próteses e meios auxiliares de locomoção*. Brasília, DF, Brasil: Ministério da Saúde, 2021.

[3] São Paulo, Secretaria de Estado dos Direitos da Pessoa com Deficiência, "Relatório sobre tecnologia assistiva e concessão de órteses e próteses no estado de são paulo," Secretaria de Estado dos Direitos da Pessoa com Deficiência, São Paulo, SP, Brasil, Tech. Rep., 2016.

[4] L. Barbieri, R. G. Tait, and S. Gandolfi, "Dispensação de órteses, próteses e meios auxiliares de locomoção no sistema Único de saúde," *Revista de Saúde Pública*, vol. 52, 2018.

[5] E. A. Biddiss and T. T. Chau, "Upper-limb prosthetics: Critical factors in device abandonment," *American Journal of Physical Medicine & Rehabilitation*, vol. 86, no. 12, pp. 977–987, Dec. 2007.

[6] I. Kyranou, S. Vijayakumar, and M. S. Erden, "Causes of performance degradation in non-invasive electromyographic pattern recognition in upper limb prostheses," *Frontiers in Neurorobotics*, vol. 12, p. 58, 2018.

[7] D. Farina and O. C. Aszmann, "Bionic limbs: Clinical reality and academic promises," *Science Translational Medicine*, vol. 6, no. 257, p. 257ps12, 2014.

[8] D. Farina, I. Vujaklija, R. Brånemark, A. M. J. Bull, H. Dietl, B. Graimann, L. J. Hargrove, K.-P. Hoffmann, H. Huang, T. Ingvarsson, B. Janusson, K. Kristjánsson, T. Kuiken, K. Micera, T. Stieglitz, A. Sturma, S. Tyler, R. F. Weir, and O. C. Aszmann, "Toward higher-performance bionic limbs for wider clinical use," *Nature Biomedical Engineering*, vol. 7, no. 4, pp. 473–485, 2023.

[9] G. Langevin, "Inmoov: Open-source 3d printed life-size robot," https://inmoov.fr/, 2012, accessed: 2026-07-30.

[10] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 40, no. 1, pp. 82–94, 1993.

[11] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 7, pp. 848–854, 2003.

[12] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *Journal of Rehabilitation Research and Development*, vol. 48, no. 6, pp. 643–660, 2011.

[13] A. Fougner, E. Scheme, A. D. C. Chan, K. Englehart, and O. Stavdahl, "Resolving the limb position effect in myoelectric pattern recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 6, pp. 644–651, 2011.

[14] Y. Wang, Y. Tian, J. Zhu, H. She, H. Yokoi, Y. Jiang, and Q. Huang, "A study on the classification effect of semg signals in different vibration environments based on the lda algorithm," *Sensors*, vol. 21, no. 18, p. 6234, 2021.

[15] P. K. Koppolu and K. Chemmangat, "A novel procedure to automate the removal of pli and motion artifacts using mode decomposition to enhance pattern recognition of semg signals for myoelectric control of prosthesis," *Biomedical Physics & Engineering Express*, vol. 10, no. 6, p. 065013, 2024.

[16] H. Daley, K. Englehart, L. Hargrove, and U. Kuruganti, "High density electromyography data of normally limbed and transradial amputee subjects for multifunction prosthetic control," *Journal of Electromyography and Kinesiology*, vol. 22, no. 3, pp. 478–484, 2012.

[17] G. Li and T. A. Kuiken, "Emg pattern recognition control of multifunctional prostheses by transradial amputees," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 6914–6917.

[18] L. H. Smith, L. J. Hargrove, B. A. Lock, and T. A. Kuiken, "Determining the optimal window length for pattern recognition-based myoelectric control: Balancing the competing effects of classification error and controller delay," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 2, pp. 186–192, 2011.

[19] S. Abbaspour, A. Naber, M. Ortiz-Catalan, H. GholamHosseini, and M. Linden, "Real-time and offline evaluation of myoelectric pattern recognition for the decoding of hand movements," *Sensors*, vol. 21, no. 16, p. 5677, 2021.

[20] A. H. Al-Timemy, G. Bugmann, and J. Escudero, "Adaptive windowing framework for surface electromyogram-based pattern recognition system for transradial amputees," *Sensors*, vol. 18, no. 8, p. 2402, 2018.

Segmentação de Cárie Dentária por Severidade em Radiografias Panorâmicas com YOLO11n-seg e *Slicing Aided Hyper Inference*

José Valclemir R. dos Santos
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
Email: valclemir15@hotmail.com

Cezário Luiz Mourão Carvalhêdo Mesquita
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
Email: cezariomourao@alu.ufc.br

Paulo César Cortez
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
Email: cortez@lesc.ufc.br

Bruno Riccelli dos Santos Silva
Universidade Federal do Ceará
Crateús, Brasil
Email: bruno@lesc.ufc.br

Francisco Rubens de Dutra Lima
Universidade Federal do Ceará
Crateús, Brasil
Email: franciscorubensdutra@alu.ufc.br

Amanda da Cruz Borges
Discente do curso de odontologia
Universidade Christus
Fortaleza, Brasil
Email: amandaborge13@gmail.com

Luéline Paiva Elias
Centro Universitário Maurício de Nassau
Fortaleza, Brasil
Email: luelinelias@gmail.com

***Resumo*—A segmentação automática de cárie em radiografias panorâmicas é dificultada pelo baixo contraste e pelo tamanho diminuto das lesões: em média uma cárie ocupa cerca de 1,5‰ da imagem, e a cárie superficial menos de 0,5‰. O *benchmark* público do conjunto DC1000 trata a cárie como tarefa binária, ignorando a severidade — dimensão codificada no próprio dataset e clinicamente decisiva (esmalte → dentina → polpa). Este trabalho avalia se o fatiamento (*Slicing-Aided Fine-Tuning*, SF, e *Slicing Aided Hyper Inference*, SAHI) melhora a segmentação *por severidade* (superficial/média/profunda) com YOLO11n-seg, e se o ganho é maior na cárie superficial. Na validação cruzada de 5 *folds*, o SF+SAHI com *tiles* de 384 px elevou o Dice macro de 0,308 para 0,415 ($\Delta = +0{,}108$) e, na cárie superficial, de 0,092 para 0,265 ($\Delta = +0{,}173$; *recall* de 0,097 para 0,323). Os cinco *folds* melhoraram em todas as classes (Wilcoxon unilateral exato, $p = 0{,}031$; indicador descritivo de consistência direcional). A ablação indica que o ganho depende da combinação (SAHI só na inferência: 0,152; SF sem fatiar a inferência: 0,291). O ganho de sensibilidade veio acompanhado de mais falsos positivos superficiais (0,526→1,591 instâncias/imagem), o que restringe o sistema a apoio ao especialista, e não a uso autônomo.**

***Palavras-chave*—**cárie dentária; segmentação de imagem; radiografia panorâmica; YOLO; SAHI; aprendizado profundo.

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 315447/2023-0 and Coordination of Superior Level Staff Improvement (CAPES)-Finance Code 001 for the support to this research. The authors declare no conflict of interest; the DC1000 dataset is public (CC BY 4.0). Generative AI tools assisted with text writing and revision, with the authors fully responsible for the conception, implementation, analyses, and conclusions.

I. Introdução

A CÁRIE dentária é uma das doenças crônicas mais prevalentes do mundo. Estima-se que a cárie não tratada afete mais de um terço da população global e que, em crianças, seja a condição crônica mais comum, atingindo cerca de 514 milhões de indivíduos [1]–[3]. Trata-se de um processo progressivo: iniciando-se como desmineralização do esmalte, a lesão avança para a dentina e, se não interrompida, atinge a polpa, podendo causar dor, infecção e perda do dente [3], [4]. Por isso, a detecção precoce e o estadiamento corretos são decisivos: o estágio da lesão determina a conduta, da remineralização não invasiva, nas lesões de esmalte, à intervenção restauradora ou endodôntica, nas lesões profundas [5], [6].

O diagnóstico convencional, baseado em inspeção visual-tátil e leitura radiográfica, é limitado por interpretação subjetiva, variabilidade entre examinadores e baixa sensibilidade às lesões iniciais de esmalte [1], [4], [7]. Esse cenário motivou a adoção de aprendizado profundo em odontologia, com redes convolucionais aplicadas à detecção e ao diagnóstico de cárie em diferentes modalidades [7], [8]. Entre elas, a radiografia **panorâmica** destaca-se como método extraoral pela baixa dose de radiação, rapidez, conforto e cobertura de toda a arcada, sendo especialmente útil em pacientes pediátricos [4].

A automação em panorâmicas, contudo, é desafiadora. A região cariada cobre apenas uma fração minúscula da imagem — em média $\sim$1,5‰ da área, e a cárie superficial menos de 0,5‰ [9] — o que aproxima a tarefa do problema de *detecção de objetos pequenos*. Soma-se o baixo contraste: a sensibilidade da leitura manual para cárie inicial varia de apenas 0,37 a 0,69 [10], com elevada taxa de falsos-negativos nas lesões incipientes — justamente o regime em que a automação poderia agregar mais valor clínico.

No conjunto público **DC1000** [9], [11], o *benchmark* mais recente avaliou doze arquiteturas de ponta (CNNs, *Transformers* e modelos *Mamba*) sob protocolo uniforme e concluiu que arquiteturas convolucionais superam as demais, com a DoubleU-Net [12] atingindo Dice de 0,7345 [1]. Esse e os demais trabalhos no DC1000, entretanto, tratam a cárie como tarefa *binária* (lesão *vs.* fundo), ignorando a *severidade* — apesar de o próprio dataset codificá-la nas máscaras. Persiste, assim, uma lacuna: avaliar a segmentação *por severidade* no DC1000, atacando o gargalo das lesões pequenas e de baixo contraste apontado por [1] como limite de desempenho.

Este estudo contribui ao formular a severidade como segmentação por instância no DC1000 e ao isolar, por ablação, a interação entre treino fatiado e inferência SAHI.

II. Objetivos e Hipótese

A. Objetivo geral

Avaliar o impacto do fatiamento (*Slicing-Aided Fine-Tuning* e *Slicing Aided Hyper Inference*) na segmentação de cárie dentária *por severidade* com YOLO11n-seg sobre o conjunto DC1000, sob validação cruzada e análise estatística.

B. Objetivos específicos

- Converter as máscaras de severidade do DC1000 para o formato de segmentação por instância e treinar um *baseline* YOLO11n-seg sobre a imagem inteira;
- Implementar o *Slicing-Aided Fine-Tuning* e a inferência fatiada (SAHI), variando o tamanho do *tile* em $\{640, 512, 384\}$ px;
- Decompor, no *tile* selecionado, os efeitos de SF e SAHI em quatro braços: *baseline*, *baseline*+SAHI, SF com inferência inteira e SF+SAHI;
- Avaliar a segmentação *por classe de severidade* com **Dice** (por classe, macro e binário), **precisão**, ***recall*** e $\mathbf{F_2}$ da superficial, **mIoU** binário e falsos positivos por imagem;
- Comparar *baseline* e configurações SF+SAHI por validação cruzada de 5 *folds* e teste de Wilcoxon pareado;
- Posicionar o resultado frente à literatura do DC1000 em colapso binário.

C. Hipótese

O fatiamento, ao recortar e reescalar *tiles*, amplia a área relativa de pixels das lesões e, portanto, beneficia *mais* a cárie superficial (pequena, baixo contraste) do que a profunda (maior em área), reduzindo o *gap* de desempenho entre as classes de severidade.

III. Fundamentação Teórica

Em visão computacional, o **aprendizado profundo** [13] domina a segmentação. A **segmentação por instância** distingue objetos individuais da mesma classe, ao contrário da semântica, que apenas rotula pixels [14]; para cárie por severidade isso é natural, pois cada lesão é um objeto cujo rótulo é o grau. Entre os detectores, os de **estágio único** (família **YOLO** [15]–[17]) predizem caixas e classes em uma passagem, com melhor relação velocidade/precisão que os de dois estágios (ex. Faster R-CNN [18]); a variante YOLO-seg acrescenta um ramo de protótipos de máscara à cabeça.

Detectores treinados em baixa resolução só reconhecem objetos acima de uma fração mínima da imagem (critérios tipo DORI, $\sim$10% da altura). Em panorâmicas de alta resolução com lesões diminutas, o **fatiamento** (*slicing*) recorta a imagem em *tiles* reescalados, elevando a área relativa do alvo e recuperando a detectabilidade, ao custo de processar múltiplas fatias [19].

IV. Trabalhos Relacionados

Segmentação binária no DC1000. O DC1000 foi introduzido por [9] e detalhado por [11]; o método que o acompanha, MLUA, é semissupervisionado: combina um esquema *teacher–student* com estimativa de incerteza em múltiplos níveis para aproveitar imagens não rotuladas e recorta o treino em quadrados de 384 px para ampliar a área de *foreground*, reportando Dice entre 0,614 (265 *slices*) e 0,711 (530 *slices*) [9]. Posteriormente, [1] estabeleceram o *benchmark* de referência, treinando doze arquiteturas (CNNs, *Transformers* e modelos *Mamba*) sob protocolo idêntico; a CNN DoubleU-Net [12] venceu (Dice 0,7345; mIoU 0,5978), e os autores observaram que a maior complexidade arquitetural não superou as redes convolucionais — um alto poder discriminativo não se traduzindo necessariamente em boa delimitação de máscara — atribuindo o teto de desempenho ao dataset pequeno, ao desbalanceamento de classes e às lesões sutis de baixo contraste. Ambos os trabalhos, contudo, colapsam a cárie em uma única classe.

Cárie multiestágio. O CariesNet [6] segmenta cárie em três estágios com um decodificador parcial e um módulo de atenção axial em escala completa para refinar lesões pequenas, atingindo Dice de 0,936 — porém em *dataset próprio* de lesões mais salientes, o que torna o valor não comparável ao DC1000. [20] combinam YOLOv5 e *Attention* U-Net [22] para detecção e segmentação por esmalte/dentina/polpa. [21] graduam cárie superficial/média/profunda em panorâmicas segmentando esmalte, dentina e polpa e intersectando essas regiões anatômicas com a lesão, porém em base própria. Na vertente de *detecção*, [10] propõem um *Transformer* para cárie inicial (graus D1–D4, *mAP* 0,873). Tais trabalhos confirmam a relevância da severidade, mas não a avaliam no DC1000.

Fatiamento para objetos pequenos. O SAHI [19] melhora a detecção de objetos pequenos sem alterar a arquitetura, atuando no treino (SF) e na inferência. O próprio MLUA recorta o treino em quadrados de 384 px para elevar a área de *foreground* de $\sim$1,5‰ para 11,79% [9] — evidência de que o

fatiamento ataca o núcleo do problema. A Tabela I resume a lacuna operacional avaliada aqui: combinar, no DC1000, severidade, validação cruzada e fatiamento.

V. Materiais e Métodos

A Figura 1 resume o *pipeline*. As subseções a seguir detalham o conjunto de dados, a conversão das anotações, o modelo YOLO11n-seg, o fatiamento SF/SAHI, as métricas, a validação cruzada, o teste de Wilcoxon e os hiperparâmetros.

O experimento segue quatro etapas encadeadas: (1) treina-se um *baseline* sobre a imagem inteira; (2) aplica-se o *Slicing-Aided Fine-Tuning*, varrendo os tamanhos de *tile* $\{640, 512, 384\}$ px com inferência SAHI, para identificar a escala mais promissora; (3) no *tile* selecionado, avaliam-se os quatro braços *baseline*, *baseline*+SAHI, SF com inferência inteira e SF+SAHI; e (4) submete-se cada configuração à validação cruzada de 5 *folds*, registrando as métricas por classe e por imagem. As comparações pareadas quantificam a consistência direcional entre *folds*. A grade de *tiles* permanece exploratória; a ablação evita atribuir ao pacote completo um efeito que pertença exclusivamente ao treino ou à inferência.

A. Conjunto de dados DC1000

O DC1000 [9], [11] deriva do *Children's Dental Panoramic Radiographs Dataset* e é um dos poucos conjuntos públicos de panorâmicas com máscaras de cárie — e o primeiro amplamente utilizado para a tarefa. A base-mãe agrega subconjuntos de segmentação de estrutura dentária e de detecção de doenças, com pares imagem/máscara; o recorte de interesse aqui é o de cárie anotada por grau de severidade, em nível de pixel por dentistas experientes [11]; o nome DC1000 provém do *benchmark* de [9]. Do release ($\sim$597 panorâmicas), pareamos 593 imagens com máscara de cárie válida em escala de cinza (2943×1435 px), das quais 2 sem lesão. Cada lesão recebe rótulo de severidade, codificado nas máscaras pelos tons de cinza $\{102, 153, 255\}$, correspondentes a *superficial*, *média* e *profunda*. A Tabela II resume a distribuição das instâncias.

A análise de tamanho fundamenta a hipótese: pela convenção COCO de objeto pequeno (área $< (32\,\text{px})^2$), **99,2%** das lesões superficiais e 90,3% das médias são "pequenas", contra apenas 36,9% das profundas — estas predominantemente médias/grandes. É esse predomínio de lesões diminutas nas classes superficial e média que o fatiamento procura mitigar.

B. Pré-processamento e conversão para YOLO-seg

O pré-processamento segue: (i) leitura da máscara de severidade e mapeamento de cada tom de cinza à sua classe; (ii) extração de componentes conexos por classe, descartando-se os de área inferior a 10 px como ruído; (iii) aproximação poligonal dos contornos; e (iv) escrita no formato de segmentação por instância do YOLO (uma linha por instância: classe seguida dos vértices normalizados do polígono). O *split* é estratificado por classe, preservando a proporção de severidades em treino e validação — cuidado relevante dado o desbalanceamento de área entre as classes. O descarte de componentes com área $<$10 px removeu apenas 20 anotações (0,44%; 17 na superficial, todas $\leq$9 px) — componentes extremamente pequenos, de impacto quantitativamente limitado sobre o conjunto de anotações.

C. Modelo base: YOLO11n-seg

Adotamos a **YOLO11n-seg** [16], [17], variante de segmentação por instância da família YOLO, na configuração *nano* (a mais leve). Diferentemente da segmentação semântica, que apenas rotula cada pixel, a segmentação *por instância* distingue objetos individuais — aqui, cada lesão é uma instância cuja classe é o seu grau de severidade. A rede é composta por: (i) um *backbone* convolucional com blocos *C3k2* (variante eficiente de CSP), agregação multiescala por *SPPF* e atenção espacial *C2PSA*; (ii) um *neck* do tipo *PAN-FPN* [23] que funde características em múltiplas escalas; e (iii) uma cabeça desacoplada, *anchor-free*, que prediz caixas (via *Distribution Focal Loss*) e classes, acrescida de um ramo de *protótipos de máscara* no estilo YOLACT [14]: a máscara de cada instância resulta da combinação linear de k protótipos ponderados pelos coeficientes preditos. A função de perda combina termos de caixa (*CIoU* + DFL), de classificação (entropia cruzada binária) e de máscara. A escolha do *backbone nano* privilegia custo computacional, adequado a um cenário de apoio ao diagnóstico.

D. Slicing-Aided Fine-Tuning *e SAHI*

O **SAHI** [19], [24] ataca o problema de objetos pequenos em duas frentes independentes e combináveis. No **treino** (*Slicing-Aided Fine-Tuning*), as imagens são recortadas em *tiles* sobrepostos de $S \times S$ px, com $S \in \{384, 512, 640\}$ e sobreposição de 20% entre *tiles* adjacentes; cada *tile* é reescalado para a resolução de entrada da rede, de modo que a área relativa de pixels de uma lesão pequena cresce substancialmente. *Tiles* positivos (com lesão) são mantidos e uma amostra de *tiles* vazios é incluída para regular falsos-positivos. Na **inferência** (*Slicing Aided Hyper Inference*), a imagem completa é fatiada com o mesmo S e sobreposição; cada *tile* é processado independentemente; e as predições são remapeadas para as coordenadas originais e fundidas por *Non-Maximum Suppression* (NMS) com casamento por IoU e limiar de casamento 0,5 (no ponto de operação conf $=$ 0,25). A biblioteca SAHI oferece ainda a fusão por *merging* (NMM e sua variante gulosa GREEDYNMM), que combina — em vez de descartar — detecções sobrepostas; adotamos o NMS por casamento de IoU, mais conservador. A ablação usa o *tile* de 384 px e a sobreposição de 0,20, mantendo o mesmo ponto de operação em todos os braços (Tabela III). Ela é operacional: o SF altera tanto a escala quanto a distribuição de exemplos de treino; por isso a interpretação do efeito de treino fatiado deve considerar também o orçamento adicional de *fine-tuning*.

A sobreposição de 20% entre *tiles* reduz o risco de uma lesão ser cortada na borda de uma fatia e, com isso, perdida ou fragmentada. O mecanismo explica por que o fatiamento favorece justamente a cárie pequena: detectores treinados em baixa resolução só reconhecem objetos acima de uma fração

Tabela I
POSICIONAMENTO FRENTE À LITERATURA. SEV.: SEGMENTAÇÃO POR SEVERIDADE; CV: VALIDAÇÃO CRUZADA k-FOLD; FATIAM.: *slicing*/SAHI.

Trabalho	Dataset	Tarefa	Sev.	CV	Fatiam.
MLUA [9]	DC1000	Seg. bin.	–	–	✓ (treino)
Ghimire *et al.* [1]	DC1000	Seg. bin.	–	–	–
CariesNet [6]	próprio	Seg. 3 est.	✓	–	–
Wang & Li [10]	próprio	Det. D1–D4	✓	–	–
Pornprasertsuk *et al.* [20]	próprio	Det.+Seg.	✓	–	–
Este trabalho	**DC1000**	**Seg. severidade**	✓	✓	✓

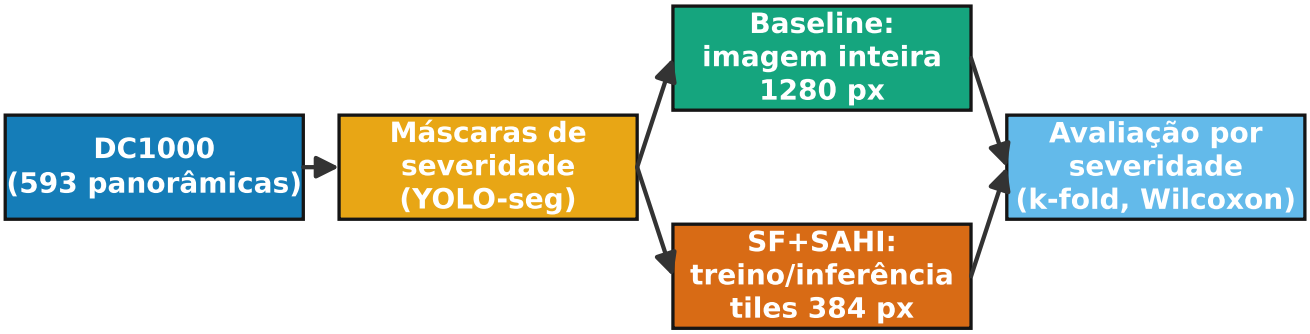


Figura 1. *Pipeline* experimental: do DC1000 anotado por severidade à avaliação estatística. O *baseline* processa a imagem inteira; o ramo SF+SAHI treina e infere por *tiles*.

Tabela II
CARACTERÍSTICAS DO DC1000 E PARTIÇÕES DA VALIDAÇÃO CRUZADA. AS CONTAGENS DE LESÕES DO *Total* SÃO DO CONJUNTO COMPLETO; TREINO E VALIDAÇÃO MOSTRAM MÉDIAS POR *fold*.

	Imagens	Superf.	Média	Profunda
Treino (por *fold*)	473–475	≈1784	≈1018	≈799
Validação (por *fold*)	118–120	≈446	≈255	≈199
Total	593	2230	1273	997

Tabela III
DESENHO DA ABLAÇÃO NO *tile* DE 384 PX.

Braço	Treino	Inferência
Baseline	imagem inteira	imagem inteira (1280 px)
Baseline+SAHI	imagem inteira	SAHI (384 px)
SF	*tiles* de 384 px	imagem inteira (1280 px)
SF+SAHI	*tiles* de 384 px	SAHI (384 px)

mínima da imagem, e o reescalonamento do *tile* eleva a lesão a essa faixa. As configurações são denotadas `sf`S`_o0.2`.

E. Métricas de avaliação

A avaliação é *pixel-level*, calculada *por classe de severidade* e também em macro-média e na variante *binária* (cárie *vs.* fundo). Sejam, por classe, TP, FP e FN o número de pixels verdadeiros-positivos, falsos-positivos e falsos-negativos. Reportamos [6], [9]:

$$\mathrm{Dice} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \quad \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2}$$

$$F_2 = \frac{5\,P \cdot R}{4\,P + R}. \tag{3}$$

O Dice (equivalente ao F_1 na segmentação) mede a sobreposição com a referência; a IoU é mais severa, penalizando mais as diferenças de borda. Adotamos o Dice como métrica primária e destacamos o *recall* (sensibilidade): em cárie incipiente, o falso-negativo é clinicamente mais grave que o falso-positivo, lógica que o F_2 formaliza ao ponderar R acima de P. Deliberadamente não reportamos acurácia e especificidade: como o fundo domina a imagem (a cárie ocupa ∼1,5‰), ambas excedem 99% mesmo para predições pobres, sendo pouco informativas. As métricas são calculadas *por classe* e agregadas por macro-média (média simples entre as três severidades), de modo que a classe profunda — maior em área — não domine o indicador global. As contagens de TP, FP e FN são acumuladas no nível de pixel, e a métrica de cada radiografia é calculada antes da média sobre o conjunto de validação (convenção por imagem), evitando que imagens com lesões extensas dominem o valor agregado. No corpo reportamos Dice (por classe, macro e binário), precisão/*recall*/F_2 da superficial e mIoU binário; não usamos *mAP* de máscara, mantendo o foco *pixel-level*. Além de FP em pixels, contabilizamos componentes preditos sem casamento um-para-um com componente de referência a IoU de máscara $\geq 0{,}5$. Reportamos esses falsos positivos

por imagem, juntamente com precisão e *recall* da superficial. Quanto a classes ausentes: a média *por classe* inclui apenas as radiografias que contêm aquela severidade (convenção *m*); imagens sem a classe não entram em sua média. No binário, todas as radiografias contam — uma imagem sem cárie e sem predição recebe Dice $= 1$, sendo penalizada apenas na presença de falso-positivo. Para os indicadores de FP por imagem, contudo, todas as radiografias entram no denominador, inclusive as sem a severidade em questão.

F. Validação cruzada k-fold

Empregamos validação cruzada estratificada de $k = 5$ ***folds***. O conjunto é dividido em k subconjuntos mutuamente exclusivos de tamanho aproximadamente igual; em cada uma das k iterações, um *fold* serve de validação e os $k-1$ restantes, de treino, até que cada *fold* tenha sido validado exatamente uma vez. A divisão é feita **por radiografia**, não por *tile*: os *folds* são formados a partir das imagens inteiras e, *somente depois*, os *tiles* de cada imagem são gerados *dentro* do seu *fold*; assim, nenhuma fatia de uma mesma radiografia aparece em treino e validação ao mesmo tempo, evitando **vazamento de dados**. A estratificação usa a **classe de severidade dominante** de cada imagem (a mais frequente, com desempate por área), preservando aproximadamente a proporção de severidades em cada *fold*. De fato, a distribuição da classe dominante no conjunto de validação é praticamente idêntica entre os cinco *folds* (≈64 imagens superficiais, 27 médias e 27 profundas por *fold*), confirmando que a estratificação funcionou. As duas radiografias sem lesão não formam um estrato de cinco partes; elas são distribuídas deterministicamente em dois *folds* de validação, e essa assimetria é registrada nos manifestos das partições. O desempenho é agregado pela média e desvio-padrão entre *folds*,

$$\mu = \frac{1}{k}\sum_{i=1}^{k} x_i, \qquad \sigma = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(x_i-\mu)^2}, \tag{4}$$

onde x_i é a métrica no *fold* i. A escolha $k = 5$ equilibra viés e variância da estimativa [25] em um conjunto pequeno e de centro único, mantendo tratável o custo da grade principal (20 treinamentos: 5 *folds* × *baseline* mais três modelos SF). A ablação acrescenta inferências, mas não novos treinamentos. A configuração recomendada é aquela com o maior ganho médio de Dice na cárie superficial *sem* deteriorar as demais classes, entre as consistentes. Esta varredura de *tile* é **exploratória**: como as três escalas são comparadas e a melhor é destacada no *mesmo* esquema de *folds*, há viés de seleção, que um conjunto de teste externo deveria confirmar.

G. Teste estatístico de Wilcoxon

Para comparar cada braço da ablação e cada configuração SF+SAHI com o *baseline* usamos o teste de **Wilcoxon de postos sinalizados**, não paramétrico e adequado a amostras pareadas que não pressupõem normalidade — caso das métricas obtidas sobre os *mesmos folds*. Para cada *fold* i calcula-se a diferença d_i entre a métrica das duas configurações; ordenam-se as diferenças por $|d_i|$, atribuindo-se postos; somam-se os postos das diferenças positivas (R^+) e negativas (R^-), e a estatística é $T = \min\{R^+, R^-\}$. Adotamos a hipótese nula H_0 de mediana das diferenças igual a zero, alternativa *unilateral* (melhora direcional), método exato e $\alpha = 0{,}05$. Duas ressalvas são tratadas explicitamente. Primeiro, com $n = 5$ o menor valor-p unilateral alcançável é $1/2^5 = 0{,}031$ (todos os *folds* na mesma direção); nesse regime o teste degenera, essencialmente, no *teste de sinais*. Segundo, não aplicamos correção para múltiplas comparações. Além disso, os conjuntos de treino dos *folds* se sobrepõem e não constituem repetições independentes. Por isso tratamos os valores-p como indicadores *descritivos* de consistência direcional e enfatizamos os **tamanhos de efeito** (Δ), a precisão e os falsos positivos por imagem.

H. Hiperparâmetros e ambiente

A Tabela IV resume o protocolo de treino, comum a todas as configurações. A grade principal compreende 20 treinamentos; o controle de imagem inteira em 1920 px acrescenta 5, totalizando 25, cada um executado em subprocesso isolado para controle de memória; o *pipeline* foi implementado em Python com Ultralytics, SAHI e scikit-learn, e executado em GPU. A geração das partições usou semente 42; os treinamentos usaram semente 42+índice do *fold*. Uma auditoria dos *checkpoints* confirmou que a seleção automática do otimizador resolveu para **AdamW** com taxa de aprendizado efetiva $\approx 1{,}43 \times 10^{-3}$ em *todos* os braços, com 85–100 épocas efetivas por treino; o *weight decay* efetivo foi 0,000375 no *baseline* e 0,0015 nos braços SF (escalonamento da biblioteca conforme o *batch*), e as atualizações diferem pela maior contagem de *tiles* — ambos decorrentes da configuração, não da escolha do otimizador. O artefato de execução `fase3_environment.json` registra as versões efetivas de Python, PyTorch/CUDA, Ultralytics, SAHI, scikit-learn e GPU; os manifestos `fase3_fold*_split.json` preservam a lista de validação de cada *fold*. Esses arquivos, os CSVs por *fold* e o código serão depositados em repositório público com DOI após a aceitação.

VI. Resultados

Varredura de *tile*. A Tabela V apresenta o Dice por classe (média de 5 *folds*). Todas as configurações SF+SAHI superam o *baseline* na cárie superficial e em macro-média. O *tile* menor (384 px) é o melhor em superficial, média e macro; na cárie profunda, o *tile* de 512 px leva ligeira vantagem.

Ablação de SF e SAHI. A Tabela VI separa os dois componentes no *tile* de 384 px. SAHI aplicado somente ao *baseline* reduz o Dice macro de 0,308 para 0,152. O SF com inferência inteira não melhora o *baseline* em macro (0,291). Em contraste, SF+SAHI alcança Dice macro de 0,415 e Dice superficial de 0,265. A combinação, portanto, é o elemento associado ao ganho; ela também eleva a precisão e o *recall*, mas aumenta os falsos positivos por imagem.

Tabela IV
PROTOCOLO DE TREINO E INFERÊNCIA.

Parâmetro	Valor
Modelo	YOLO11n-seg
Semente (*splits*)	42
Semente (treino)	42+índice do *fold*
Épocas (máx.)	100
Early stopping (paciência)	30
Otimizador	AdamW (seleção automática)
Resolução — *baseline*	1280 px (imagem inteira)
Resolução — SF	*tiles* $S \in \{384, 512, 640\}$ px
Sobreposição de *tiles*	0,20
Augmentation	*flip* horiz. (0,5); rotação $\pm 5°$; *jitter* de brilho (HSV-V 0,2)
Confiança (operação)	0,25
Fusão SAHI	NMS, casamento por IoU
Folds (k)	5 (estratificado)

Tabela V
DICE POR CLASSE DE SEVERIDADE (MÉDIA DE 5 *folds*). MELHOR POR COLUNA EM **NEGRITO**.

Configuração	Superf.	Média	Profunda	Macro
Baseline	0,092	0,261	0,571	0,308
`sf640_o0.2`	0,261	0,341	0,556	0,386
`sf512_o0.2`	0,256	0,347	**0,627**	0,410
`sf384_o0.2`	**0,265**	**0,358**	0,623	**0,415**

Controle de resolução. Como o *baseline* reduz a imagem a 1280 px enquanto o SF opera sobre recortes *nativos*, o fatiamento poderia estar confundido com maior resolução efetiva. Treinamos então um *baseline* de imagem inteira em **alta resolução (1920 px)**, mesmos *folds* e protocolo (Tabela VI). A resolução recupera *parte* do ganho (Dice superficial 0,092 → 0,168; *recall* → 0,180; $p = 0{,}031$), mas o SF+SAHI-384 ainda o supera em Dice superficial (+0,098), macro (+0,050) e *recall* (+0,143), cinco *folds* na mesma direção ($p = 0{,}031$); só na profunda empatam ($\Delta = +0{,}004$; $p = 0{,}5$). O ganho, portanto, não é explicado apenas pela resolução efetiva, sugerindo uma contribuição adicional do fatiamento.

Consistência direcional (grade completa). A Tabela VII traz, para cada configuração, o tamanho de efeito (Δ) por métrica. A `sf384_o0.2` é a única que melhora *todas* as classes com os cinco *folds* na mesma direção ($p = 0{,}031$), inclusive na profunda. Em `sf512` também há melhora direcional na profunda, enquanto em `sf640` a profunda chega a piorar ($\Delta = -0{,}015$).

***Baseline vs.* SF+SAHI-384.** A Figura 2 e a Tabela VI detalham a comparação com a vencedora. O Dice superficial quase triplica (0,092 → 0,265), a precisão aumenta de 0,113 para 0,286, o *recall* de 0,097 para 0,323 e o F_2 superficial de 0,092 para 0,283. Em contrapartida, os FP superficiais por imagem aumentam de 0,526 para 1,591.

Estabilidade entre *folds*. A `sf384_o0.2` é também a mais *estável*: o menor desvio-padrão do Dice superficial entre *folds* (0,009, contra 0,032 da `sf512` e 0,014 do *baseline*), combinando maior ganho médio à menor variância — o ganho não depende de um *fold* particularmente favorável.

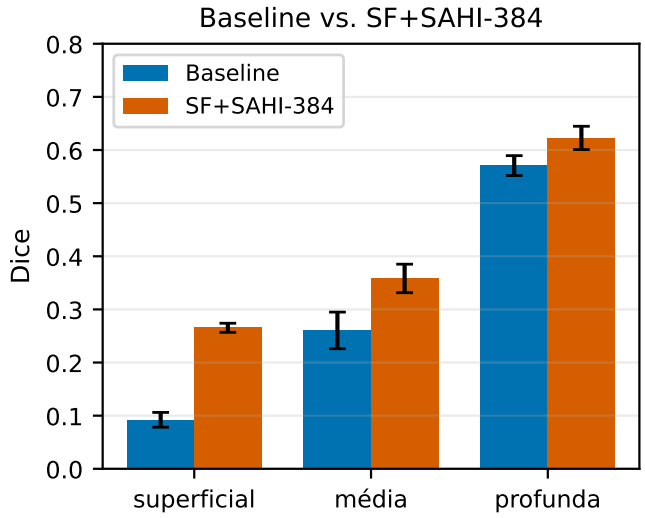


Figura 2. Dice por classe: *baseline vs.* `sf384_o0.2`. Barras de erro = desvio-padrão entre os 5 *folds*.

Posicionamento binário. *Uma ressalva precede a comparação:* o colapso da nossa saída de segmentação por *instância* em uma máscara semântica binária, somado a diferenças de arquitetura, resolução e *split*, faz de qualquer comparação com a literatura um *contexto*, não um confronto direto. Feita a ressalva, a Tabela VIII situa nosso Dice binário (0,608), próximo ao do MLUA em 265 *slices* (0,614) e abaixo da DoubleU-Net (0,735).

Falsos positivos e custo. A Figura 3 e a Tabela VI mostram que o ganho de sensibilidade não é gratuito: os FP superficiais passam de 0,526 para 1,591 instâncias por imagem e os binários de 1,246 para 2,713. Quanto ao custo, a inferência fatiada processa $\lceil W/S \rceil \times \lceil H/S \rceil$ fatias por imagem: *tiles* menores (384 px) implicam mais fatias e maior tempo — *trade-off* entre sensibilidade e custo a considerar em uso clínico.

VII. DISCUSSÃO

A ablação identifica uma interação entre SF e SAHI. O ganho do SF+SAHI-384 concentrou-se na cárie superficial (Δ Dice $= +0{,}173$; Δ recall $= +0{,}226$), acima da profunda (+0,052), e ocorreu na mesma direção nos cinco *folds*. Entretanto, o resultado não pode ser atribuído a qualquer componente isoladamente: SAHI sobre o *baseline* reduz o Dice macro para 0,152, e SF com inferência inteira chega a 0,291, abaixo do *baseline* (0,308). A melhoria requer o alinhamento entre a escala vista no *fine-tuning* e a escala da inferência fatiada. Isso é coerente com o mecanismo proposto para objetos pequenos [19] e com o fato de 99,2% das lesões superficiais serem objetos "pequenos".

O piso baixo do *baseline* não é um artefato. O Dice superficial do *baseline* (0,092) é muito baixo, mas a ablação mostra que não decorre apenas do *fine-tuning* extra (SF-inteira atinge só 0,098) nem apenas da resolução (o *baseline*

Tabela VI
ABLAÇÃO DE SF, SAHI E CONTROLE DE RESOLUÇÃO (*baseline* 1920 PX), MÉDIA DE 5 *folds*. FP: INSTÂNCIAS PREDITAS SEM CASAMENTO POR IOU DE MÁSCARA $\geq 0{,}5$; MENOR É MELHOR.

Configuração	Dice sup.	Precisão sup.	Recall sup.	FP sup./img.	Dice macro	Dice bin.	FP bin./img.
Baseline (1280 px)	0,092	0,113	0,097	0,526	0,308	0,536	1,246
Baseline (1920 px)	0,168	0,205	0,180	0,966	0,366	0,594	1,549
Baseline+SAHI	0,023	0,055	0,017	0,125	0,152	0,284	2,162
SF (inteira)	0,098	0,114	0,107	0,644	0,291	0,520	1,232
SF+SAHI	**0,265**	**0,286**	**0,323**	1,591	**0,415**	**0,608**	2,713

Tabela VII
TAMANHO DE EFEITO (Δ DICE; MÉDIA DE 5 *folds*) DE CADA SF *vs. baseline*. TODA MELHORA DIRECIONAL CONSISTENTE (5/5 *folds*) CORRESPONDE A $p = 0{,}031$ (WILCOXON UNILATERAL, DESCRITIVO — SEÇÃO V-G); $\Delta < 0$ INDICA PIORA. "REC_S": *recall* SUPERFICIAL; "BIN": BINÁRIO.

Config.	Superf.	Média	Profunda	Macro	rec_S	bin
`sf640`	+0,168	+0,081	−0,015	+0,078	+0,234	−0,008
`sf512`	+0,164	+0,087	+0,056	+0,102	+0,203	+0,060
`sf384`	+0,173	+0,098	+0,052	+0,108	+0,226	+0,072

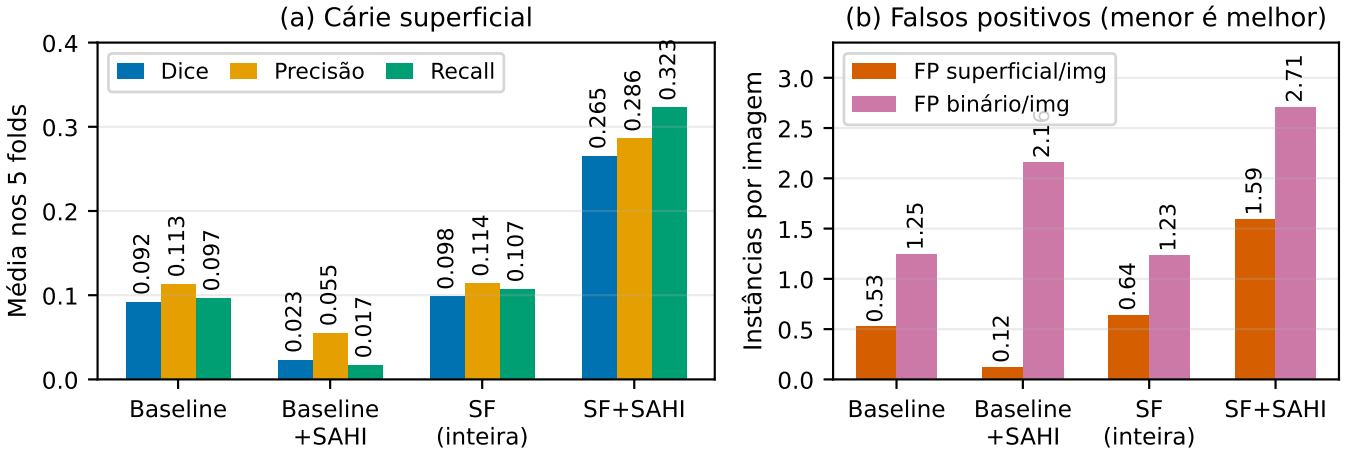


Figura 3. Ablação de SF e SAHI no protocolo de cinco *folds*. (a) Dice, precisão e *recall* da cárie superficial; maior é melhor. (b) falsos positivos por imagem; menor é melhor. O ganho da configuração SF+SAHI vem acompanhado de maior número de falsos positivos.

Tabela VIII
POSICIONAMENTO NO DC1000 (BINÁRIO). **CONTEXTO**, NÃO COMPARAÇÃO DIRETA.

Método	Dice (bin.)	mIoU (bin.)
Baseline (nosso)	0,536	0,406
SF-384+SAHI (nosso)	0,608	0,472
MLUA 265 *slices* [9]	0,614	–
MLUA 530 *slices* [9]	0,711	–
DoubleU-Net [1], [12]	0,735	0,598

a 1920 px chega a 0,168): expressa a dificuldade intrínseca de segmentar lesões diminutas e de baixo contraste. Uma atribuição causal estrita ao treino fatiado ainda exigiria igual orçamento de atualizações em imagem inteira; os valores absolutos permanecem modestos.

O mecanismo, porém, não é estritamente monótono — uma questão em aberto. O *tile* de 384 px maximiza a superficial, a média e o macro, enquanto 512 px leva vantagem na profunda (classe majoritariamente média/grande). Pelo argumento puro de área relativa, espera-se que *tiles* menores ajudem *menos* as classes maiores — o que não se observa de forma limpa. Não temos, com os dados atuais, explicação fechada para essa não-monotonicidade; ela permanece uma questão em aberto, cuja investigação exige isolar a densidade de falsos-positivos por fatia e o efeito da fusão das predições entre *tiles*. Registramos, assim, que o argumento de área relativa explica o ganho dominante na superficial, mas não esgota o comportamento por classe.

***Trade-off* de falsos positivos.** A melhor configuração melhorou todas as severidades em Dice e elevou a precisão superficial (0,113 → 0,286), mas também triplicou os FP superficiais por imagem (0,526 → 1,591) e mais que dobrou os FP binários (1,246 → 2,713). O ganho não decorre, portanto, de uma simples troca de *recall* por precisão; ele vem com maior atividade preditiva, que ainda exige triagem pelo especialista.

Posicionamento. Nosso Dice binário (0,608) é compatível com o MLUA em 265 *slices* (0,614) e inferior à DoubleU-Net (0,735) — esperado, pois colapsamos uma saída por instância em máscara semântica, além de diferenças de arquitetura e *split*. Ademais, nosso próprio *baseline* (0,536) está $\sim$0,20 abaixo da DoubleU-Net, de modo que o $+0{,}072$ do SF+SAHI é melhora *relativa* controlada sobre o mesmo *baseline*, e não aproximação do estado da arte binário — que uma rede-base mais forte provavelmente já alcançaria. A contribuição principal não é superar o *benchmark* binário, mas avaliar a severidade — informação que orienta a conduta [5], [6] e ausente nos trabalhos no DC1000.

Análise estatística. Diferentemente dos trabalhos no DC1000, usamos validação cruzada estratificada, comparação pareada e tamanhos de efeito; reportar os Δ ao lado dos valores-p evita a leitura binária "significativo/não" que o pequeno n tornaria frágil.

Implicações clínicas. Segmentar *por severidade* distingue estágios de conduta distinta — da remineralização, na superficial, à restauração ou endodontia, na profunda. O ganho de *recall* na superficial ($0{,}097 \rightarrow 0{,}323$) é relevante, pois é na fase incipiente que a prevenção é mais eficaz e a leitura manual mais falha [10]. Ainda assim, os valores modestos e o aumento de FP posicionam a ferramenta como *apoio* ao especialista, não substituto.

Limitações. (i) Com $n = 5$ *folds* o poder estatístico é baixo (o teste degenera no de sinais); enfatizamos tamanhos de efeito. (ii) Métricas em um único ponto de operação (conf $= 0{,}25$). (iii) DC1000 é de centro único. (iv) Os valores absolutos de Dice são modestos. (v) As classes derivam da intensidade das máscaras, sem validação histológica. (vi) Usamos o *backbone* mais leve (YOLO11n); variantes maiores podem alterar os absolutos. (vii) A escolha do *tile* e o teste usam o mesmo esquema de *folds*, sem *hold-out* — viés otimista que um conjunto externo deveria confirmar. (viii) O SF recebe *fine-tuning* adicional; um controle de igual orçamento de atualizações em imagem inteira reforçaria a atribuição causal ao treino fatiado.

VIII. Conclusão

Avaliamos a segmentação de cárie *por severidade* no DC1000. No protocolo de cinco *folds*, SF+SAHI com *tiles* de 384 px foi a configuração mais consistente com YOLO11n-seg, elevando o Dice macro de 0,308 para 0,415 e o superficial de 0,092 para 0,265 (*recall* $0{,}097 \rightarrow 0{,}323$). A ablação indica que o ganho depende da combinação entre treino e inferência fatiados, e um controle em alta resolução (1920 px) mostra que ele **não** se reduz à resolução efetiva. O custo foi o aumento dos falsos positivos ($0{,}526 \rightarrow 1{,}591$/imagem), que restringe o uso a apoio ao especialista. Próximas etapas: teste externo, mais centros, validação histológica, varredura do limiar e *backbones* mais robustos.

Referências

[1] A. Ghimire *et al.*, "When CNNs outperform transformers and mambas: Revisiting deep architectures for dental caries segmentation," *arXiv:2511.14860*, 2025, doi: 10.48550/arXiv.2511.14860.

[2] N. J. Kassebaum *et al.*, "Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015," *J. Dent. Res.*, vol. 96, no. 4, pp. 380–387, 2017, doi: 10.1177/0022034517693566.

[3] N. B. Pitts *et al.*, "Dental caries," *Nat. Rev. Dis. Primers*, vol. 3, art. 17030, 2017, doi: 10.1038/nrdp.2017.30.

[4] E. Asci *et al.*, "A deep learning approach to automatic tooth caries segmentation in panoramic radiographs of children in primary dentition, mixed dentition, and permanent dentition," *Children*, vol. 11, no. 6, art. 690, 2024, doi: 10.3390/children11060690.

[5] A. I. Ismail *et al.*, "The International Caries Detection and Assessment System (ICDAS): An integrated system for measuring dental caries," *Community Dent. Oral Epidemiol.*, vol. 35, no. 3, pp. 170–178, 2007, doi: 10.1111/j.1600-0528.2007.00347.x.

[6] H. Zhu *et al.*, "CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image," *Neural Comput. Appl.*, vol. 35, pp. 16051–16059, 2023, doi: 10.1007/s00521-021-06684-2.

[7] F. Schwendicke, T. Golla, M. Dreher, and J. Krois, "Convolutional neural networks for dental image diagnostics: A scoping review," *J. Dent.*, vol. 91, art. 103226, 2019, doi: 10.1016/j.jdent.2019.103226.

[8] J.-H. Lee *et al.*, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm," *J. Dent.*, vol. 77, pp. 106–111, 2018, doi: 10.1016/j.jdent.2018.07.015.

[9] X. Wang *et al.*, "Multi-level uncertainty aware learning for semi-supervised dental panoramic caries segmentation," *Neurocomputing*, vol. 540, art. 126208, 2023, doi: 10.1016/j.neucom.2023.03.069.

[10] L. Wang and Z. Li, "Transformer-based intelligent detection model for early dental caries in panoramic radiographs," *Sci. Rep.*, vol. 16, no. 1, art. 3507, 2026, doi: 10.1038/s41598-025-33391-y.

[11] Y. Zhang, F. Ye, L. Chen *et al.*, "Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection," *Scientific Data*, vol. 10, art. 380, 2023, doi: 10.1038/s41597-023-02237-5.

[12] D. Jha *et al.*, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE CBMS*, 2020, pp. 558–564, doi: 10.1109/CBMS49503.2020.00111.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.

[14] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF ICCV*, 2019, pp. 9157–9166.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE CVPR*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[16] G. Jocher and J. Qiu, "Ultralytics YOLO11," ver. 11.0.0, 2024.

[17] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv:2410.17725*, 2024, doi: 10.48550/arXiv.2410.17725.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[19] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE ICIP*, 2022, pp. 966–970, doi: 10.1109/ICIP46576.2022.9897990.

[20] S. Pornprasertsuk-Damrongsri *et al.*, "Clinical application of deep learning for enhanced multistage caries detection in panoramic radiographs," *Sci. Rep.*, vol. 15, art. 33491, 2025, doi: 10.1038/s41598-025-16591-4.

[21] Q. Chen, J. Huang, H. Zhu *et al.*, "Automatic and visualized grading of dental caries using deep learning on panoramic radiographs," *Multimedia Tools Appl.*, vol. 82, pp. 23709–23734, 2023, doi: 10.1007/s11042-022-14089-z.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[23] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE CVPR*, 2017, pp. 2117–2125.

[24] F. C. Akyon *et al.*, "SAHI: A lightweight vision library for large scale object detection and instance segmentation," Zenodo, 2021, doi: 10.5281/zenodo.5718950.

[25] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, 1995, pp. 1137–1145.

Sensitivity Analysis between qEEG Metrics and Longitudinal Clinical Scores in Post-Stroke Patients

1st Guilherme Tavares dos Santos
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
gtavares@fei.edu.br

2nd Maria Claudia F. Castro
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
mclaudia@fei.edu.br

***Abstract*—Quantitative EEG (qEEG) indices have been proposed as early biomarkers of post-stroke neuroplasticity, on the premise that synaptic reorganization precedes the functional gains captured by clinical scales used in neurorehabilitation. This work proposes and demonstrates a clinimetric sensitivity-analysis framework that jointly evaluates the reliability and responsiveness of candidate qEEG biomarkers against the Fugl-Meyer Assessment, scoring each on four mutually non-redundant, equally weighted core criteria—reliability (ICC), signal above measurement noise (MDC_{95}), responsiveness (SRM), and clinical discrimination (AUC-ROC)—while SEM, paired Cohen's d and clinical coupling (Spearman's ρ) are reported without being scored. To overcome the limited availability of patient-level raw EEG in the source cohort, a hybrid dataset was constructed by pairing published longitudinal clinical scores for proportional and poor recoverers with group-conditioned, physiologically clamped Gaussian Monte Carlo simulations of PSD-derived qEEG trajectories (DAR, DTABR, BSI, pdBSI) at the initial (2 weeks) and final (3 months) post-stroke periods, parameterized from the stroke EEG literature. Applied to two proportional-recovery subgroups with high and low clinical variation, the framework identified DTABR and BSI as the most informative biomarkers, distinguishing statistically significant from clinically meaningful candidates even when clinical improvement was minimal or imperceptible. These results show how the framework separates reliable, responsive signals from measurement noise, motivating its application to neurorehabilitation protocols. By conceptualizing the brain as a dynamic target and therapy as a responsive system, this framework may substantially improve functional recovery.**

***Index Terms*—Clinimetrics, Fugl-Meyer Assessment, Monte Carlo simulation, Post Stroke neurorehabilitation, Quantitative EEG, Sensitivity analysis**

I. Introduction

Sensitivity Analysis (SA), or uncertainty analysis, examines how uncertainty in a model's outputs can be traced back to different sources of variability in its inputs. This field focuses on understanding how uncertainties in the inputs influence the outputs [1]. To tackle this problem, researchers employ a range of methodologies, each bringing distinct assumptions and limitations. These approaches yield conclusions that vary in interpretability and detail, all aimed at answering a central question: How do different sources of uncertainty or variations in model inputs affect output uncertainty? Sensitivity Analysis has been applied in patient-centered neurorehabilitation programs designed to optimize motor and cognitive recovery. Sensitivity Analysis provides a structured framework for assessing how variations in clinical assumptions, patient characteristics, or therapy intensity influence functional outcomes [2]. By assessing the distinct impact of each variable on outcomes, SA reveals which factors most strongly influence patient improvement, rather than presuming a uniform response to standard protocols. This insight is essential for crafting truly personalized rehabilitation strategies, especially for complex conditions like stroke [3], spinal cord injury [4], and multiple sclerosis [2]. As a result, SA stands as a cornerstone of methodological rigor in neurorehabilitation research, particularly as studies frequently shift their focus across diverse metrics and encounter related challenges, including:

- Patient heterogeneity—including differences in severity, chronicity, and lesion location [5]—and high dropout rates;
- The use of multiple clinical outcome scales (e.g., Fugl-Meyer Assessment (FMA), Barthel, ARAT, FIM, WISCI-II) [6];
- Lack of blinding in therapist-delivered interventions [7].

Therefore, SA helps ensure that conclusions remain robust despite these specific challenges [8].

A. Kinds of Sensitivity Analysis in Neurorehabilitation

SA application in neurorehabilitation encompasses a set of complementary methods used to test whether study findings remain robust across parameter values and analytic choices. Some of the principal methods are:

- Reliability, which evaluates how much of the observed score variance is due to true differences between real outcomes vs. measurement error [9];
- Responsiveness Analysis, which evaluates the instrument's ability to detect change over time, particularly clinically important change [10].

Together, these two properties define the clinimetric quality this work evaluates: the ability of a specific metric or instrument to sensitively detect true differences or changes in patients receiving treatment [11]. A third quantity, clinical coupling—the correlation between a quantitative metric and an established clinical scale—is customarily invoked as evidence of concurrent validity and is reported here as a companion

measure, but it is not part of the scored core. Seven statistics are computed in total. Four of them—ICC, MDC_{95}, SRM and AUC-ROC—constitute the scored core of the framework; the remaining three—SEM, paired Cohen's d and Spearman's ρ—are reported for completeness and carry no weight in the composite score, for the scale-invariance and redundancy reasons detailed in Section II-B, item (4). They are defined as follows:

1) Reliability:

- The Intraclass Correlation Coefficient (ICC), defined in Eq.1, quantifies the consistency of repeated measurements within the same individual. A high ICC indicates that a scale can reliably reproduce the same value across repeated assessments, effectively distinguishing true between-subject variance from measurement error [12]. If the ICC falls below the commonly accepted threshold (typically $\geq$ 0.7), distinguishing real signal changes from measurement instability becomes challenging. The closer the ICC is to 1, the more consistently the metric produces reliable results. Interpretation: $|\text{ICC}| \geq 0.90$ = excellent; $|\text{ICC}| \geq 0.70$ = acceptable; $|\text{ICC}| \geq 0.50$ = moderate; $|\text{ICC}| < 0.50$ = poor.
- The Standard Error of Measurement (SEM), defined in Eq.2, quantifies measurement precision by representing the 'noise band' surrounding an individual score [13]. SEM depends on both the standard deviation (SD) of measurements and the reliability (ICC). Higher reliability (ICC approaching 1) or lower SD yields a smaller SEM, indicating greater measurement precision. *Reported, not scored* (Section II-B, item (4)); its information enters the framework through MDC_{95} (Eq.3);
- Minimal Detectable Change (MDC_{95}), defined in Eq.3, is the absolute threshold above which an individual's change in score has a $\leq 5\%$ probability of being due to measurement error. Changes smaller than MDC_{95} are likely the result of measurement error rather than genuine change [14]. Lower MDC_{95} values reflect higher measurement sensitivity.

2) Responsiveness:

- The Standardized Response Mean (SRM), presented in Eq.4, is a widely used measure of internal responsiveness in rehabilitation. SRM captures both the magnitude of change and its consistency across patients in a single index. A higher absolute SRM value ($|\text{SRM}|$) indicates greater ability to detect true changes, effectively distinguishing genuine recovery signals from measurement noise [15]. Interpretation: $|\text{SRM}| \geq 0.8$ = large; ≥ 0.5 = moderate; ≥ 0.2 = small change.
- Paired Cohen's d, described in Eq.5, facilitates comparison of change magnitude across metrics with different scales by expressing changes in units of the baseline standard deviation. The absolute value of Cohen's d ($|d|$) quantifies the change relative to data variability, offering a standardized metric for mean differences [16]. Higher absolute values indicate a greater difference between the initial (T_0) and final (T_n) stages of longitudinal recovery. *Reported, not scored*: d shares its numerator with the SRM, and the MDC_{95} criterion is a strictly more demanding form of the same comparison (Section II-B, item (4)).
- The Area Under the Receiver Operating Characteristic Curve (AUC-ROC), defined in Eq.6, measures the ability of $|\Delta\text{metric}|$ (e.g., changes in a quantitative electroencephalogram metric) to distinguish clinical responders—those with improvement above the median on the clinical scale ($|\Delta\text{FMA}|$)—from non-responders. AUC-ROC translates statistical responsiveness into clinical discriminative utility, providing the essential link between the metric and diagnostic or prognostic decision-making. Interpretation: AUC-ROC = 0.5 indicates random discrimination; ≥ 0.7 is acceptable; ≥ 0.8 is considered good for detecting improvement above the median of ΔFMA [17].

3) Clinical coupling (reported, not scored):

- Spearman's rank correlation coefficient (ρ), described in Eq.7, is a non-parametric statistic that quantifies the strength and direction of a monotonic association between two ranked variables, such as a qEEG metric and a clinical scale [18]. *Reported, not scored*: it enters the composite score with $w_\rho = 0$ (Section II-B, item (4)).

$$\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} \tag{1}$$

$$\text{SEM} = \text{SD} \times \sqrt{1 - \text{ICC}} \tag{2}$$

$$\text{MDC}_{95} = 1.96 \times \text{SEM} \times \sqrt{2} \tag{3}$$

$$\text{SRM} = \frac{\overline{\Delta}}{\text{SD}(\Delta)} \tag{4}$$

$$\text{Cohen}'\text{s } d = \frac{\overline{\Delta}}{\text{SD}(T_0)} \tag{5}$$

$$\text{AUC} = \frac{U_+}{n_+\, n_-} = \frac{\displaystyle\sum_{i:\, y_i=1} r_i - \frac{n_+(n_+ + 1)}{2}}{n_+\, n_-} \tag{6}$$

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\,(n^2 - 1)} \tag{7}$$

Equations for the evaluation metrics are as follows: ICC (Eq. 1) represents the intraclass correlation coefficient, quantifying the proportion of total variance attributable to between-subject differences, where $\sigma^2_{\text{between}}$ is the between-subject variance and σ^2_{within} is the within-subject (residual) variance. SEM (Eq. 2), MDC_{95} (Eq. 3), SRM (Eq. 4), and paired Cohen's d (Eq. 5) assess measurement error and responsiveness. Here, $\overline{\Delta}$ denotes the mean change between time points, SD(Δ) the standard deviation of that change, SD(T_0) the baseline standard deviation, and SD the relevant sample standard deviation. AUC (Eq. 6) refers to the area under the ROC curve, computed through equivalence to the Mann–Whitney U statistic: r_i is the rank of score x_i among all samples (with ties resolved by midranks), U_+ is the number of correctly ranked positive–negative pairs, and n_+ and n_- are the counts of positive ($y_i = 1$) and negative

samples, respectively. Spearman's ρ (Eq. 7) uses d_i, the difference between paired ranks, to quantify monotonic association.

B. Electroencephalogram in Neurorehabilitation

Quantitative Electroencephalogram (qEEG) metrics offer objective neurophysiological biomarkers for diagnosis, prognosis, and monitoring of recovery in stroke and other neurological disorders [19]. These metrics are increasingly employed to track recovery and guide interventions, providing reliable indicators of neurophysiological change. In many cases, qEEG metrics are more sensitive than traditional clinical scales. Commonly used qEEG metrics include:

- Power Spectral Density (PSD) serves as the basis for most derived qEEG indices, quantifying EEG power within standard frequency bands: delta (δ, 0.5–4 Hz), theta (ϑ, 4–8 Hz), alpha (α, 8–13 Hz), beta (β, 13–30 Hz), and gamma (γ, >30 Hz), as defined in Eq.8. It is treated in this work as the substrate of the candidate biomarkers rather than as a candidate in its own right;
- Delta/Alpha Ratio (DAR), defined in Eq.9, is the ratio of delta (δ) to alpha (α) band power. An elevated DAR is associated with increased cortical dysfunction [20];
- Delta/Theta to Alpha/Beta Ratio (DTABR), defined in Eq.10, is the ratio of combined δ and ϑ band power to combined α and β band power. Higher DTABR reflects a shift toward slow-wave activity and is correlated with greater cortical dysfunction; one of the most robust qEEG markers of impaired Stroke recovery [20].
- Brain Symmetry Index (BSI), defined in Eq.11, quantifies interhemispheric spectral asymmetry and is widely used for acute stroke monitoring [21];
- Pairwise-Derived BSI (pdBSI), defined in Eq. 12, quantifies asymmetry across all symmetric EEG channel pairs and frequency bands. Unlike traditional BSI, pdBSI is computed dynamically over time windows and for each symmetric pair individually, without prior averaging across hemispheres [22].

$$PSD(f) = \frac{1}{N \cdot f_s} \left| \sum_{t=0}^{N-1} x(t)\, w(t)\, e^{-j2\pi f t / f_s} \right|^2 \tag{8}$$

$$\mathrm{DAR} = \frac{\int_{0.5}^{4} PSD(f)\, df}{\int_{8}^{13} PSD(f)\, df} \tag{9}$$

$$\mathrm{DTABR} = \frac{\int_{0.5}^{4} PSD(f)\, df + \int_{4}^{8} PSD(f)\, df}{\int_{8}^{13} PSD(f)\, df + \int_{13}^{30} PSD(f)\, df} \tag{10}$$

$$\mathrm{BSI} = \frac{1}{N \cdot M} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| \frac{R_{ij} - L_{ij}}{R_{ij} + L_{ij}} \right| \tag{11}$$

$$\mathrm{pdBSI} = \frac{1}{N \cdot M} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| \frac{R_{ij}(f) - L_{ij}(f)}{R_{ij}(f) + L_{ij}(f)} \right| \tag{12}$$

qEEG biomarkers are derived from the EEG signal $x(t)$. The Power Spectral Density (PSD, Eq. 8) is computed using a windowed ($w(t)$) discrete Fourier transform over N samples at a sampling frequency f_s, providing the foundation for band-power metrics. The Delta/Alpha Ratio (DAR, Eq. 9) and Delta+Theta/Alpha+Beta Ratio (DTABR, Eq. 10) capture the balance of slow versus fast oscillatory power, with higher values reflecting increased cortical dysfunction. The Brain Symmetry Index (BSI, Eq. 11) and its pairwise-derived spectral variant (pdBSI, Eq. 12) assess average normalized power asymmetry between homologous right (R_{ij}) and left (L_{ij}) hemisphere electrodes across M channel pairs and N frequency bins, where higher values signal greater interhemispheric asymmetry. Collectively, these indices are among the strongest qEEG predictors of stroke recovery.

C. The Importance of Multimodal Integration of qEEG and Clinical Scales in Stroke Neurorehabilitation

In neurorehabilitation, relying solely on either qEEG or clinical scales is insufficient. Clinical scales reflect the end result of motor and cognitive function but frequently overlook subclinical brain reorganization. Conversely, qEEG uncovers the neurophysiological substrates of recovery—such as cortical excitability, interhemispheric balance, and network reorganization [19]—yet lacks direct correspondence to functional outcomes. Integrating these modalities enables earlier and more accurate prognosis, personalized therapy, objective recovery monitoring, and deeper mechanistic understanding [20]. qEEG-derived metrics also reduce operator-dependent variability, converting subjective clinical assessments into reproducible biomarkers ideal for clinical trials and longitudinal studies [23]—mechanistic distinctions that clinical indices alone cannot provide [24]. As neurorehabilitation advances toward greater precision and individualization, the convergence of qEEG biomarkers and clinical scales is shaping the next generation of personalized neurorestoration. In this paradigm, therapy intensity, modality, and timing are continuously tailored to each patient's evolving neurophysiological profile [25]. This work proposes a multimodal assessment framework designed to bridge the gap between neurophysiological mechanisms (what is happening in the brain) and functional outcomes (what the patient can do).

II. Methodology

A. Hybrid Solution: Real and Simulated Data

To illustrate the application of Sensitivity Analysis (SA) in neurorehabilitation, clinical assessment data from [26] were analyzed. This dataset includes longitudinal FMA scores recorded at 2 weeks (T_0) and 3 months (T_1) post-stroke, allowing comparison between proportional (PROP) and poor (POOR) recovery groups in motor rehabilitation programs (see Fig.1). However, only aggregated FMA scores for the overall cohort or specific subgroups were available, and no raw EEG data or channel-wise PSD signals were accessible [26]. This lack of raw data precluded recalculation of qEEG metrics

for individual frequency bands or sensor regions[1]. To address this, Gaussian Monte Carlo simulations were performed to estimate the expected evolution of brain physiological patterns for each subgroup, drawing on established findings from the qEEG literature [27]. Value assignment was guided by theoretical markers of post-stroke cortical slowing: poorer outcomes are typically associated with higher relative energy in slow bands (delta/theta), reduced balance with fast bands (alpha/beta), and greater interhemispheric asymmetry [28]. Informed by this and related studies [29], the simulations enforced separation between groups at T_0 and distinct trajectories through T_1—improvement in PROP patients, and maintenance or worsening in POOR patients. Group-conditioned Gaussian sampling with clamp truncation [30] ensured physiological plausibility and avoided unrealistic outliers. As a result, all longitudinal qEEG variations in the analysis reflect simulated group trajectories rather than direct patient measurements.

$$X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{13}$$

$$\tilde{X}_0 = \mathrm{clip}(X_0, L_0, U_0) \tag{14}$$

$$\varepsilon \sim \mathcal{N}(\mu_\Delta, \sigma_\Delta^2) \tag{15}$$

$$\tilde{X}_1 = \mathrm{clip}(\tilde{X}_0 + \varepsilon,\ L_1, U_1) \tag{16}$$

$$\Delta X = \tilde{X}_1 - \tilde{X}_0 \tag{17}$$

Gaussian Monte Carlo simulations with clipping. Baseline values X_0 are drawn from a normal distribution (Eq.13) and clipped to the bounds $[L_0, U_0]$ to obtain $\tilde{X}_0$ (Eq.14). A random increment ε is sampled and added (Eq.15), with the result clipped to $[L_1, U_1]$ to yield the follow-up value $\tilde{X}_1$ (Eq.16). The simulated change is $\Delta X = \tilde{X}_1 - \tilde{X}_0$ (Eq.17).

B. Implementation of the qEEG–FMA Clinimetrics Sensitivity Analysis Framework

In clinical research, significant clinical coupling—most often quantified by Spearman correlation—between a quantitative metric and a validated clinical scale provides evidence of concurrent validity for the quantitative metric [31]. This relationship addresses a central question: Does the metric's signal change alongside the patient's clinical status? Yet, concurrent validity alone is not enough. The metric must also exhibit reliability—consistently yielding stable values when no real change occurs—and responsiveness—the ability to accurately detect true change when it happens. Without these properties, observed clinical coupling may be confounded by measurement noise, ceiling or floor effects (when scales cannot detect further improvement), or random fluctuations between therapeutic sessions [10]. This rigor is particularly crucial in post-stroke recovery. The proportional recovery rule and the ceiling effect of the FMA scale—which can mask subclinical improvements—require that a candidate qEEG biomarker both minimizes measurement error (Is the signal stable when no true change occurs?) and sensitively tracks clinical outcomes (Does the signal change when genuine recovery happens?) — otherwise, the metric risks statistical significance (demonstrating clinical coupling) without actual clinical relevance. In other words, a metric with high correlation may still be noisy, suffer from ceiling or floor effects, or fail to capture meaningful longitudinal change [32].

The proposed pipeline (Fig. 2) systematically evaluates whether candidate qEEG biomarkers are both reliable and responsive indicators of motor recovery, indexed by the FMA.

(1) Inputs: The power spectral density (PSD, Eq. 8) is simulated at two time points (T_0 and T_1), in parallel with the FMA clinical scale, and the four candidate qEEG biomarkers scored by the framework—DAR, DTABR, BSI, and pdBSI (Eqs. 9–12)—are derived from it. The PSD is therefore an input to the pipeline rather than a scored candidate: being a frequency- and channel-resolved function, it admits no scalar summary that is not itself a band-power or asymmetry index already represented among the four derived metrics.

(2) Longitudinal Change: Within-subject change scores ($\Delta = T_1 - T_0$) are calculated for each biomarker. ΔFMA is then used to dichotomize patients into responders and non-responders using a median split.

(3) Reliability, Responsiveness, and Clinical Coupling Metrics: Each biomarker is characterized in three domains, and within each domain a distinction is drawn between the statistics that are *scored* and those that are *reported only* (see item (4)): - Reliability/Measurement Error: ICC (scored); the minimal detectable change at 95% confidence, MDC_{95} (scored, through the comparison $|\overline{\Delta}| \geq \mathrm{MDC}_{95}$); and SEM (reported only). - Responsiveness: the standardized response mean SRM (scored); the AUC-ROC, which discriminates responders from non-responders (scored); and paired Cohen's d (reported only). - Clinical Coupling: Spearman's correlation ρ between each Δ-metric and ΔFMA, reported for descriptive purposes only and entering the framework with a weight of exactly zero ($w_\rho = 0$): *sensitivity is not correlation*, and a static, cross-sectional association must not compensate for a failure in the longitudinal domain the framework interrogates (see item (4)).

(4) Decision Gate / Scoring: Each biomarker earns one point per core criterion met, yielding a composite score $S \in \{0, \ldots, 4\}$ formally defined in Eq. 18, where c_k denotes the k-th criterion and $\mathbf{1}[\cdot]$ the indicator function. The four core criteria carry equal unit weight ($w_k = 1$ for all k) and are deliberately chosen so that each interrogates a distinct property and none is an algebraic consequence of the others: (c_1) $\mathrm{ICC} \geq 0.70$—is the metric stable?; (c_2) $|\overline{\Delta}| \geq \mathrm{MDC}_{95}$—does the observed change exceed the metric's own measurement noise?; (c_3) $|\mathrm{SRM}| \geq 0.80$—is the change large relative to its dispersion across patients?; and (c_4) $\mathrm{AUC\text{-}ROC} \geq 0.70$—does the metric separate responders from non-responders? All four are invariant under a rescaling of the biomarker—in c_1 and c_3 the scale factor cancels between numerator and denominator, in c_2 it multiplies both sides of the inequality, and c_4 is rank-based—so the same thresholds apply to bounded asymmetry indices (BSI, pdBSI) and to unbounded power ratios (DAR, DTABR) alike. The practical import is clearest

[1]A formal data request was sent to the corresponding authors, but no response was received.

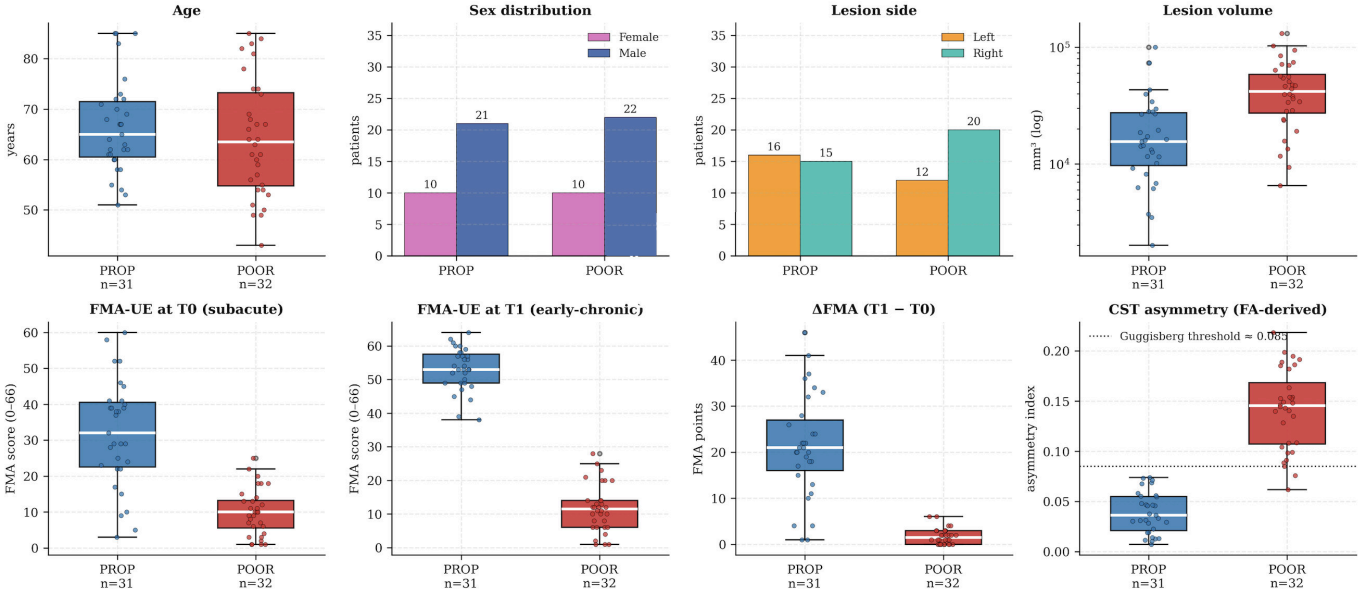


Fig. 1: Based on article [26] metadata of 63 patients. Age 65.3 ∓ 10.4 years (range 43–85y). Sex: 20 females, 43 males. Lesion side: 28 left, 35 right. T_0 = 20 ∓ 4 days post Stroke, T_1 = 89 ∓ 7 days post Stroke. FMA T_0 20.7 ∓ 15.7, FMA T_1 32.1 ∓ 21.7. Corticospinal Tract (CST) asymmetry and lesion volume data were not considered for simulations.

by counterexample: DTABR in Subgroup I (SEM = 0.285) would fail an absolute SEM < 0.05 test, yet expressing the same trajectories in a unit ten times smaller turns that failure into a pass (0.0285) while leaving all four core criteria unchanged—a verdict flipped by a choice of unit measures scale, not precision. The bias is systematic here: the bounded asymmetry indices carry SEMs in the hundredths, the unbounded power ratios in the tenths, so an absolute cutoff would reward the former for being small, not for being reproducible. Two statistics that are commonly reported in clinimetric work are computed here (Tables I and II) but deliberately excluded from S. The SEM is excluded on two grounds: it carries the native units of the biomarker, so an absolute threshold is not comparable across metrics whose scales differ by an order of magnitude; and, by Eq. 2, $\mathrm{SEM} = \mathrm{SD}\sqrt{1-\mathrm{ICC}}$, so at a fixed SD it is a deterministic function of c_1. Paired Cohen's d is excluded for a stronger reason: under the framework's own definitions it is not an independent test at all. Substituting Eq. 2 into Eq. 3 gives $\mathrm{MDC}_{95} = 1.96\sqrt{2}\,\mathrm{SD}\sqrt{1-\mathrm{ICC}}$; dividing both sides of c_2 by the baseline standard deviation $\mathrm{SD}(T_0)$ that normalizes Eq. 5 then yields

$$|\overline{\Delta}| \geq \mathrm{MDC}_{95} \iff \underbrace{\frac{|\overline{\Delta}|}{\mathrm{SD}(T_0)}}_{|d|} \geq 1.96\sqrt{2}\,\sqrt{1-\mathrm{ICC}}\,.$$

Criterion c_2 and a threshold of the form $|d| \geq 0.80$ are therefore two statements about the same quantity, differing only in the threshold—fixed in one case, reliability-adjusted in the other. Since $1.96\sqrt{2}\,\sqrt{1-\mathrm{ICC}} > 0.80$ for every $\mathrm{ICC} < 0.917$, and all reliabilities observed here fall below that value, c_2 is strictly more demanding: satisfying it already guarantees the fixed-threshold form, so scoring both would award two points for a single comparison. (The identity assumes the SD of Eq. 2 is the baseline SD of Eq. 5; with a pooled estimate it is approximate, but the dependence remains.) Spearman's ρ is likewise excluded, entering Eq. 18 with $w_\rho = 0$. A high-scoring biomarker is thus simultaneously reliable, responsive, and clinically discriminative, irrespective of its clinical coupling.

$$S = \sum_{k=1}^{4} w_k\,\mathbf{1}[c_k] \;+\; w_\rho\,|\rho(\Delta\mathrm{qEEG}, \Delta\mathrm{FMA})|\,, \qquad (18)$$

where $w_k = 1$ for all four core criteria $c_1, \ldots, c_4$ and $w_\rho = 0$. The second term is retained in the formulation only to make explicit that clinical coupling is measured and reported, but does not contribute to S.

(5) Output: biomarkers are classified by their composite score S as Strong (4), Good (3), Intermediate (2), or Weak (0–1), and ranked by S alone. Because $w_\rho = 0$, clinical coupling does not act as a tie-breaker: biomarkers sharing the same S are reported as tied and are distinguished qualitatively by which domain—reliability or responsiveness—drives their score.

III. Results

A. Applying the Clinimetrics Sensitivity Analysis Framework to Simulated Data

The results and statistical evaluation of longitudinal changes were analyzed for the following two PROP subgroups:

- Subgroup I with a high clinical variation: 10 patients with $20 \leq \Delta\mathrm{FMA} \leq 25$, 70% right-sided lesions and 80% males;

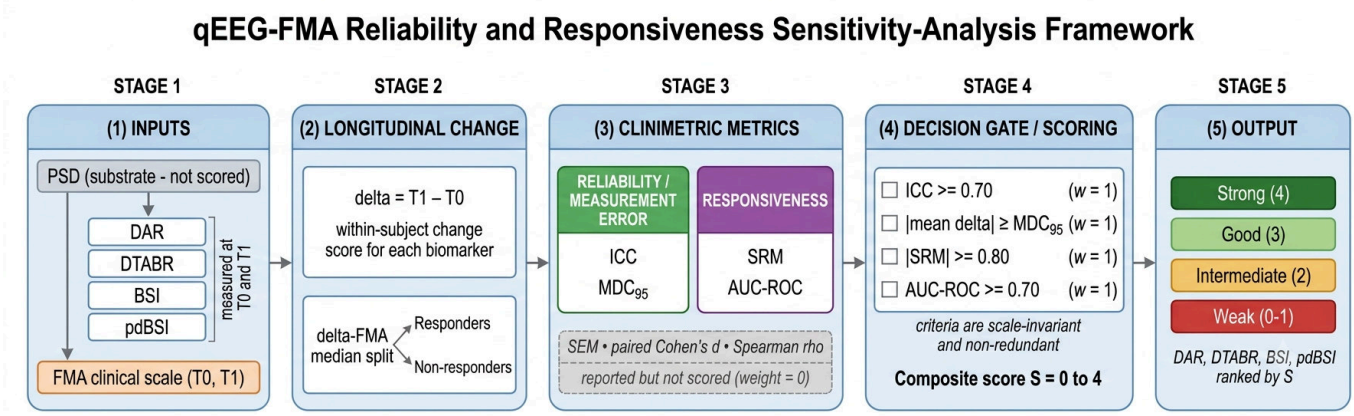


Fig. 2: **Overview of the qEEG–FMA reliability and responsiveness sensitivity-analysis framework.** The pipeline scores candidate qEEG biomarkers according to how reliably and responsively they track motor recovery, indexed by the FMA. **(1) Inputs:** the power spectral density (PSD) is the common substrate of the analysis and is not itself scored; the four candidate biomarkers derived from it—DAR, DTABR, BSI, and pdBSI—are obtained at two time points (T_0 and T_1), in parallel with the FMA clinical scale. **(2) Longitudinal change:** within-subject change scores ($\Delta = T_1 - T_0$) are derived for every biomarker, and ΔFMA is used to dichotomize patients into responders versus non-responders by a median split. **(3) Clinimetric metrics:** the scored statistics are grouped into two domains—reliability/measurement error (ICC and the minimal detectable change at 95% confidence, MDC_{95}) and responsiveness (SRM and the AUC-ROC discriminating responders from non-responders). The dashed box holds the three statistics that are computed and reported but carry no weight in the score—SEM, paired Cohen's d and Spearman's ρ—for the scale-invariance and redundancy reasons detailed in Section II-B, item (4). **(4) Decision gate / scoring:** each biomarker earns one point per criterion met, the four mutually non-redundant criteria carrying equal unit weight ($w_k = 1$): $\mathrm{ICC} \geq 0.70$; $|\overline{\Delta}| \geq \mathrm{MDC}_{95}$; $|\mathrm{SRM}| \geq 0.80$; and $\mathrm{AUC\text{-}ROC} \geq 0.70$, yielding a composite score $S \in \{0, \ldots, 4\}$ as defined in Eq. 18. A strong biomarker must be simultaneously reliable, responsive, and clinically discriminative. **(5) Output:** the four derived qEEG biomarkers are ranked by S alone and color-coded as Strong (4), Good (3), Intermediate (2), or Weak (0–1).

- Subgroup II with a low clinical variation: 7 patients with $0 \leq \Delta\mathrm{FMA} \leq 15$ (low or imperceptible clinical improvement), with 75% left-sided lesions and 50% females.

These subgroups were selected to demonstrate that qEEG metrics can reveal neurophysiological improvements not only in patients with pronounced clinical gains (high clinical variation), but also in those with minimal or imperceptible clinical improvement. In both cases, reliability and responsiveness measures support the sensitivity of qEEG metrics, as shown in Tables I and II.

In the first subgroup, the BSI emerged as the most reliable qEEG metric, exhibiting the highest |ICC| (0.906), lowest |SEM| (0.015), and lowest MDC_{95} (0.044), despite showing negligible clinical coupling (low $|\rho|$ = 0.219). BSI satisfied both reliability criteria (c_1 and c_2) but failed both responsiveness criteria: its |SRM| of 0.144 is an order of magnitude below the 0.80 threshold and its AUC-ROC of 0.60 does not reach 0.70. It therefore met 2 of the 4 core criteria (Score: 2/4), classifying it as Intermediate — reliable, but without a longitudinal signal to be reliable about. DTABR was the most responsive metric, with the highest |SRM| (2.740) and an excellent AUC-ROC (0.96), making it the best qEEG metric for distinguishing patients who improved above the median ΔFMA. It met three of the four core criteria—ICC $\geq$ 0.70 (0.792), |SRM| $\geq$ 0.80 (2.740) and AUC-ROC $\geq$ 0.70 (0.96)—failing only on $|\overline{\Delta}| \geq \mathrm{MDC}_{95}$. DTABR therefore scored 3/4 (Good) in this subgroup: strongly responsive and discriminative, but with a mean change that does not yet clear its own measurement noise at this sample size. The pdBSI matched DTABR at 3/4 (Good) by a complementary route: it combined an excellent ICC (0.865) with a mean change exceeding its MDC_{95} (0.050) and a large SRM (0.927), failing only the discrimination criterion (AUC-ROC 0.64). The two Good biomarkers of this subgroup therefore rest on different evidence—pdBSI on reliability plus change magnitude, DTABR on change magnitude plus discrimination—and neither dominates the other. BSI and DAR both scored 2/4 (Intermediate), Table I. For Subgroup II, BSI once again emerged as the most reliable scale, exhibiting the highest |ICC| (0.899), low |SEM| (0.209), and the lowest MDC_{95} (0.039), along with moderate clinical coupling $|\rho|$ (0.447). BSI also proved to be the most responsive metric, with the highest |SRM| (1.758) and an acceptable AUC-ROC (0.75). These values indicate good discriminatory power and responsiveness within this subgroup. BSI was consequently the only biomarker in the study to satisfy all four core criteria (4/4, Strong), being simultaneously reliable, above its own measurement noise, responsive, and discriminative. DTABR followed at 3/4 (Good), missing only

TABLE I: Sensitivity metrics for the first subgroup with high clinical variation

	Responsiveness			Reliability			Clinical coupling	
Metric	SRM	Cohen's d	AUC-ROC	ICC	SEM	MDC_{95}	$\rho\times\Delta$FMA	Score
pdBSI	-0.927	-0.361	0.64	0.865	0.024	0.050	0.094	3/4
DTABR	-2.740	-0.628	0.96	0.792	0.285	0.292	-0.813	3/4
BSI	-0.144	-0.069	0.60	0.906	0.015	0.044	-0.219	2/4
DAR	-1.118	-0.440	0.64	0.843	0.370	0.731	-0.219	2/4

First PROP subgroup — 10 patients with ΔFMA between 20 and 25. DTABR emerged as the most responsive scale, effectively detecting change between T_0 and T_1 and discriminating patient responses ($|\mathrm{SRM}| = 2.740$, AUC-ROC = 0.96); its clinical coupling was also the strongest of the panel ($\rho = -0.813$), although this does not contribute to its score. BSI proved to be the most reliable metric—demonstrating stability and sensitivity to subtle variations (|ICC| = 0.906, SEM near zero, MDC_{95} = 0.044, the lowest of all metrics)—but failed both responsiveness criteria. Scoring follows the four-criterion core and bands of Section II-B, item (4). In this subgroup: pdBSI and DTABR are Good (3/4) — the former on reliability plus change magnitude, the latter on change magnitude plus discrimination — while BSI and DAR are Intermediate (2/4). The SEM, Cohen's d and $\rho\times\Delta$FMA columns are reported for completeness but are not scored criteria (Section II-B, item (4)); in particular, $w_\rho = 0$.

TABLE II: Sensitivity metrics for the second subgroup with low clinical variation

	Responsiveness			Reliability			Clinical coupling	
Metric	SRM	Cohen's d	AUC-ROC	ICC	SEM	MDC_{95}	$\rho\times\Delta$FMA	Score
BSI	-1.758	-0.534	0.75	0.899	0.209	0.039	-0.447	4/4
DTABR	-1.547	-0.562	0.75	0.846	0.029	0.412	-0.447	3/4
DAR	-0.981	-0.672	0.00	0.756	0.354	1.031	-0.894	2/4
pdBSI	-1.118	-1.215	0.25	0.033	0.024	0.053	0.447	1/4

Second PROP subgroup — 7 patients with ΔFMA between 0 and 15 (imperceptible or low clinical improvement). BSI was the only biomarker to satisfy all four core criteria (4/4, Strong), standing out as both the most responsive (|SRM| = 1.758, AUC-ROC = 0.75) and the most reliable (ICC = 0.899, MDC_{95} = 0.039) metric in this subgroup. DTABR followed at 3/4 (Good), missing only the $|\overline{\Delta}| \geq MDC_{95}$ criterion. DAR was Intermediate (2/4) and pdBSI Weak (1/4), the latter driven by an ICC of 0.033 that places it below the reliability threshold by a wide margin. The SEM, Cohen's d and $\rho\times\Delta$FMA columns are reported for completeness but are not scored criteria (Section II-B, item (4)); in particular, $w_\rho = 0$.

$|\overline{\Delta}| \geq \mathrm{MDC}_{95}$ — the same criterion it missed in Subgroup I. DAR was Intermediate (2/4) and pdBSI Weak (1/4); the latter is a marked reversal from Subgroup I and is driven by an ICC of 0.033, indicating that under low clinical variation the pairwise-derived index carried essentially no reproducible between-subject signal. See Table II.

IV. Discussion

The proposed pipeline offered a robust, systematic framework for identifying the most informative qEEG metrics, thereby enhancing the rigor of biomarker selection and validation. All evaluated metrics performed in agreement with expectations from the literature in our simulations [20, 27, 33]. Both DAR and DTABR—the ratios of slow to fast cortical activity—showed elevated values and power spectra shifted toward slower rhythms, indicating reduced cortical activation, thalamo-cortical deafferentation, and functional disconnection. These patterns mirror those observed in stroke patients with minimal or no recovery. Likewise, BSI and pdBSI, which quantify interhemispheric power asymmetry, also showed higher values, a feature frequently associated with poorer motor recovery. In this study, BSI and DTABR emerged as the most sensitive qEEG metrics for tracking patient progress during neurorehabilitation, independent of gender or lesion laterality. Our findings demonstrate that qEEG scales can reliably detect early progression, contributing to more accurate prognoses. High-reliability measures, such as ICC, consistently distinguished true between-subject variance from measurement error, while responsiveness metrics, such as SRM, captured both the magnitude and consistency of change across patients. Notably, these strengths were established independently of clinical coupling, which by construction carries no weight in the composite score and is reported only as descriptive evidence of concurrent validity. Applying this framework to longitudinal data from stroke and traumatic brain injury cohorts could further demonstrate its utility for personalized treatment. Integrating these qEEG metrics with clinical scales enables early detection and supports individualized, objective monitoring of neuroplastic changes throughout recovery. Ultimately, conceptualizing the brain as a dynamic, adaptable target—and designing therapy as a responsive, stage-adaptive system—offers substantial promise for advancing functional recovery. Future research should explore qEEG metrics beyond traditional spectral densities, such as Functional Connectivity (FC) [34]. By quantifying synchronization between brain regions, FC may offer richer prognostic insights and help guide the development of more tailored therapeutic strategies. Additionally, adopting advanced simulation models—such as neural mass models [35]—that more accurately represent the neurophysiological dynamics of post-stroke patients in neurorehabilitation could uncover further valuable qEEG biomarkers. Beyond that, incorporating additional statistical metrics—such as the Jacobson–Truax index, which identifies individual changes that surpass measurement error, and the Rasch/IRT proxy Z score, an exploratory indicator of signal-to-noise ratio—could also further strengthen sensitivity analyses in neurorehabilitation [36]. Two design choices deserve emphasis: every core criterion is invariant under a rescaling of the biomarker, and the criteria were screened for algebraic dependence before being given equal weight (Section II-B, item (4)); excluding SEM, Cohen's d and ρ from S is what makes the equal weighting defensible. A residual limitation is that c_2 depends on the measurement-error estimate propagated from the ICC, so biomarkers with poorly estimated reliability inherit that uncertainty. It is important to highlight that this work centers on sensitivity analysis as a method for validating biomarkers of post-stroke neural reorganization and recovery. The primary challenge addressed is measurement error—which sensitivity analysis directly quantifies—rather than establishing causal relationships. Responsiveness and reliability are essential prerequisites for subsequent causal modeling, as they ensure a metric reliably captures the relevant signal. Causal modeling should follow as a next step, ideally within a larger, prospectively designed interventional study.

V. Conclusion

This work presented a sensitivity analysis framework that reliably identified DTABR and BSI as the most informative qEEG biomarkers for tracking post-stroke motor recovery: DTABR was the only biomarker to reach Good in both

subgroups, and BSI the only one to satisfy all four core criteria in either. By jointly assessing reliability and responsiveness—with clinical coupling reported as unweighted descriptive evidence—the approach distinguished robust signals from noise and revealed neurophysiological improvements that traditional scales might miss. Although based on simulated data, this framework highlights the importance of reliability and responsiveness in biomarker validation and can be extended to real datasets, connectivity metrics, and more advanced models for personalized neurorehabilitation.

VI. Acknowledgment

We gratefully acknowledge [26] for making the clinical data available, which was essential to the success of this study.

References

[1] B.E. Shepherd et al. "Global Sensitivity Analysis. The Primer by". In: *Biometrics* 65 (2009), pp. 1311–1312. DOI: 10.1111/j.1541-0420.2009.01343.

[2] R.B. Price and R. Duman. "Neuroplasticity in cognitive and psychological mechanisms of depression: an integrative model". In: *Mol Psychiatry* (Mar. 25, 2020). DOI: 10.1038/s41380-019-0615-x..

[3] J.J. Deeks et al. "Analysing data and undertaking meta-analyses. Chapter 10: Analysing data and undertaking meta-analyses". In: *Cochrane Handbook for Systematic Reviews of Interventions* (2019). Ed. by J.P.T. Higgins et al.

[4] J. Li et al. "Parameter identification and sensitivity analysis of a lower-limb musculoskeletal model". In: *Front. Bioeng. Biotechnol* 13.1566381 (2025). DOI: 10.3389/fbioe.2025.1566381..

[5] M. Oberholzer and R.M. Müri. "Neurorehabilitation of Traumatic Brain Injury (TBI): A Clinical Review". In: *Med Sci* (Mar. 18, 2019). DOI: 10.3390/medsci7030047..

[6] Jeremy Charles Hobart. "Measuring disability in neurological rehabilitation : psychometric evaluation of two outcome measures". PhD thesis. London School of Hygiene & Tropical Medicine, 1999. DOI: 10.17037/PUBS.00682304..

[7] J.P. Friedberg, Natarajan Lipsitz SR, and S. "Challenges and recommendations for blinding in behavioral interventions illustrated using a case study of a behavioral intervention to lower blood pressure". In: *Patient Educ Couns* Jan;78(1):5-11 (2010). DOI: 10.1016/j.pec.2009.04.009..

[8] M.H. Liang, A.H. Fossel, and M.G. Larson. "Comparisons of five health status instruments for orthopedic evaluation". In: *Med Care* (July 28, 1990). DOI: 10.1097/00005650-199007000-00008..

[9] L.B. Mokkink et al. "The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes". In: *J Clin Epidemiol* Jul;63(7):737-45 (2010). DOI: 10.1016/j.jclinepi.2010.02.006..

[10] J.A. Husted et al. "Methods for assessing responsiveness: a critical review and recommendations". In: *J Clin Epidemiol* (May 5, 2000). DOI: 10.1016/s0895-4356(99)00206-1..

[11] L.B. Mokkink, H.C.W. Vet, and C.A.C. Prinsen. "COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures". In: *Qual Life Res* 27 (2018), pp. 1171–1179. DOI: 10.1007/s11136-017-1765-4..

[12] J.M. Cassidy et al. "Low-Frequency Oscillations Are a Biomarker of Injury and Recovery After Stroke". In: *Stroke* (May 5, 2020). DOI: 10.1161/StrokeAHA.120.028932..

[13] Vet HC et al. "When to use agreement versus reliability measures". In: *J Clin Epidemiol* Oct;59(10):1033-9 (2006). DOI: 10.1016/j.jclinepi.2005.10.015..

[14] Stephen M. Haley and Maria A. Fragala-Pinkham. "Interpreting Change Scores of Tests and Measures Used in Physical Therapy". In: *Physical Therapy* 86.ue 5 (May 1, 2006), pp. 735–743. DOI: 10.1093/ptj/86.5.735..

[15] P.W. Stratford, J.M. Binkley, and D.L. Riddle. "Health status measures: strategies and analytic methods for assessing change scores". In: *Phys Ther* Oct;76(10):1109-23 (1996). DOI: 10.1093/ptj/76.10.1109..

[16] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 1988. DOI: 10.4324/9780203771587..

[17] P.W. Stratford, D.L. Riddle, and Epub. "When minimal detectable change exceeds a diagnostic test-based threshold change value for an outcome measure: resolving the conflict". In: *Phys Ther* Oct;92(10):1338-47 (2012). DOI: 10.2522/ptj.20120002..

[18] M.M. Mukaka. "Statistics corner: A guide to appropriate use of correlation coefficient in medical research". In: *Malawi Med J* (2012).

[19] Jinru Feng, Weili Jia, and Zixiao Li. "Electroencephalography: A valuable tool for assessing motor impairment and recovery post-Stroke". In: *Journal of Neuroscience Methods* 422 (2025), p. 110518. DOI: 10.1016/j.jneumeth.2025.110518.

[20] S. Finnigan, M.J. Putten, and Epub. "EEG in ischaemic Stroke: quantitative EEG can uniquely inform (sub-)acute prognoses and clinical management". In: *Clin Neurophysiol* Jan;124(1):10-9 (2013). DOI: 10.1016/j.clinph.2012.07.003..

[21] Putten MJ et al. "A brain symmetry index (BSI) for online EEG monitoring in carotid endarterectomy". In: *Clin Neurophysiol* (May 5, 2004). DOI: 10.1016/j.clinph.2003.12.002..

[22] Putten MJ and Epub. "The revised brain symmetry index". In: *Clin Neurophysiol* v;118(11):2362-7 (2007). DOI: 10.1016/j.clinph.2007.07.019..

[23] H.M. Schambra et al. "Differential PostStroke Motor Recovery in an Arm Versus Hand Muscle in the Absence of Motor Evoked Potentials". In: *Neurorehabil Neural Repair* (July 7, 2019). DOI: 10.1177/1545968319850138..

[24] A.R. Carter, G.L. Shulman, and M. Corbetta. "Why use a connectivity-based approach to study Stroke and recovery of function?" In: *Neuroimage* 1;62(4):2271-80 (Oct. 2012). DOI: 10.1016/j.neuroimage.2012.02.070..

[25] L.A. Boyd et al. "Biomarkers of Stroke recovery: Consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable". In: *Int J Stroke* (July 12, 2017). DOI: 10.1177/1747493017714176..

[26] A.G. Guggisberg et al. "Longitudinal Structural and Functional Differences Between Proportional and Poor Motor Recovery After Stroke". In: *Neurorehabil Neural Repair* Dec;31(12):1029-1041 (2017). DOI: 10.1177/1545968317740634..

[27] Putten MJ, Tavy DL, and Epub. "Continuous quantitative EEG monitoring in hemispheric Stroke patients using the brain symmetry index". In: *Stroke* v;35(11):2489-92 (2004). DOI: 10.1161/01.STR.0000144649.49861.1d..

[28] E. Schleiger et al. "Frontal EEG Delta/Alpha ratio and screening for post-Stroke cognitive deficits: the power of four electrodes". In: *Int J Psychophysiol* Oct;94(1):19-24 (2014). DOI: 10.1016/j.ijpsycho.2014.06.012..

[29] R.V. Sheorajpanday et al. "Quantitative EEG in ischemic Stroke: correlation with functional status after 6 months". In: *Clin Neurophysiol* (May 5, 2011). DOI: 10.1016/j.clinph.2010.07.028..

[30] B. Lapeyre. *Introduction to Monte-Carlo Methods*. 2019. URL: https://cermics.enpc.fr/~bl/Luxembourg/monte-carlo/lecture-1.pdf (visited on 06/05/2026).

[31] D.L. Streiner, G.R. Norman, and J. Cairney. *Health Measurement Scales: A practical guide to their development and use*. OUP Oxford, 2024. ISBN: 9780192696311. URL: https://books.google.com.br/books?id=pb30EAAAQBAJ.

[32] G. Kwakkel et al. "Standardized measurement of quality of upper limb movement after Stroke: Consensus-based core recommendations from the Second Stroke Recovery and Rehabilitation Roundtable". In: *Int J Stroke* Oct;14(8):783-791 (2019). DOI: 10.1177/1747493019873519..

[33] E. Cuspineda et al. "QEEG prognostic value in acute Stroke". In: *Clin EEG Neurosci* (July 3, 2007). DOI: 10.1177/155005940703800312..

[34] E. Napoli et al. "Biomarkers of stroke recovery using EEG-based resting-state functional connectivity: a systematic review". In: *J NeuroEng Rehabil* 23.142 (2026). DOI: 10.1186/s12984-026-01928-5..

[35] Elif Köksal-Ersöz et al. "eCOALIA: Neocortical neural mass model for simulating electroencephalographic signals". In: *SoftwareX* 28 (2024), p. 101924. DOI: 10.1016/j.softx.2024.101924.

[36] A. Tennant and P.G. Conaghan. "The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?" In: *Arthritis Rheum* Dec 15;57(8):1358-62 (2007). DOI: 10.1002/art.23108..

Sparse Time-Frequency Characterization of Ischemic ST-Segment Morphology in ECG Using Matching Pursuit Decomposition

Elzbieta Olejarczyk
Nalecz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland; *Ariadna Std*., Lublin, Poland
eolejarczyk@ibib.waw.pl

Elzbieta Raus-Jarzabek
Department of Biocybernetics and Biomedical Engineering,
AGH University of Krakow, Krakow, Poland
raus@agh.edu.pl

Owahidur Rahman
Department of Biocybernetics and Biomedical Engineering,
AGH University of Krakow, Krakow, Poland; Ariadna Std., Lublin, Poland
rowahidur@agh.edu.pl

Abstract*—Objective: Electrocardiographic ST-segment analysis is important for the characterization of myocardial ischemia, but amplitude-based measurements provide limited information about the time-frequency organization of repolarization morphology. This study investigates Matching Pursuit (MP) decomposition as a sparse representation of controlled ST-segment modifications in synthetic ECG signals. Methods: Synthetic lead-II ECG signals with ST-segment deviations from −0.8 to +0.8 mV were generated and contaminated with additive white Gaussian noise at multiple signal-to-noise ratio levels. Signals were decomposed using a Gabor-dictionary MP approach, and energy-, frequency-, scale, and sparse-decomposition descriptors were extracted. Kruskal–Wallis testing, regression analysis, two-way ANOVA with interaction effects, effect-size analysis, correlation analysis, and principal component analysis were applied. Results: MP consistently represented QRS morphology using short-scale, higher-frequency atoms, whereas ST morphology was represented by longer-scale, low-frequency atoms. Energy-related descriptors showed strong nonlinear dependence on ST deviation magnitude, with quadratic R^2=0.971 for total ST energy. Mean ST frequency and the number of ST-related atoms within the first ten iterations showed moderate linear relationships with signed ST deviation (R^2=0.354 and 0.223, respectively). For the first ST iteration, the signed linear relationship was negligible (R^2=0.001); however, regression against absolute ST deviation yielded R^2=0.572, indicating a relationship primarily with ST magnitude rather than polarity. Among detected ST atoms, the corresponding timing relationship reached R^2=0.654. The largest noise effect was observed for mean ST scale (η^2=0.133). Conclusion: Under the controlled synthetic conditions investigated, MP provides complementary descriptors of ST morphology beyond a single amplitude measure. The results support further investigation of MP-based sparse descriptors using clinical ECG recordings and systematic comparison with alternative time-frequency methods.

Keywords—electrocardiogram, ST-segment deviation, myocardial ischemia, Matching Pursuit, sparse representation, time–frequency analysis, ECG morphology

I. Introduction

Electrocardiographic (ECG) analysis remains one of the principal non-invasive tools for detection and assessment of myocardial ischemia and acute coronary syndromes [1–4]. In clinical practice, ischemic abnormalities are primarily identified through changes in ST-segment morphology, including ST elevation, ST depression, and repolarization disturbances [1, 3, 5, 6]. Conventional ECG interpretation is largely based on amplitude measurements relative to the isoelectric line and visual assessment of waveform morphology [2, 3, 7]. However, these approaches may provide limited characterization of complex repolarization dynamics, particularly in noisy recordings and subtle ischemic conditions [8-9].

Time–frequency analysis methods offer an attractive alternative for investigation of nonstationary biomedical signals such as ECG, EEG, and MEG [10–13]. Among these approaches, Matching Pursuit (MP) decomposition introduced by Mallat and Zhang [10] enables adaptive sparse representation of signals using abundant time–frequency dictionaries. In MP decomposition, a signal is iteratively approximated using elementary waveforms ("atoms") selected to maximize correlation with the residual signal. This adaptive strategy provides high temporal and spectral resolution while preserving signal interpretability [10, 14].

Matching Pursuit and related sparse decomposition techniques have previously demonstrated utility in biomedical signal analysis, including EEG time–frequency characterization [11–13], ECG compression and QRS detection [15–16], and adaptive decomposition of cardiovascular signals [17]. Recent developments in optimized dictionary structures and GPU-accelerated implementations further improved computational efficiency and representation quality of MP-based approaches [18–19].

Sparse decomposition methods may be particularly suitable for ECG analysis because ventricular depolarization and repolarization exhibit substantially different temporal and spectral characteristics. The QRS complex corresponds to rapid synchronized electrical activation, producing transient high-frequency structures, whereas the ST segment reflects slower plateau-phase ventricular repolarization processes [20]. Consequently, pathological ischemic changes may alter not only ECG amplitudes but also the sparse compressibility structure and spectral organization of repolarization components within adaptive time–frequency space.

Previous studies on ECG morphology modelling have suggested that sparse signal representations may provide complementary information beyond conventional amplitude-based descriptors [21]. However, the application of MP decomposition specifically for characterization of ischemic ST-segment morphology and its robustness under varying noise conditions remains insufficiently explored.

Therefore, the aim of the present study was to investigate sparse time–frequency representation of ischemic ST-segment abnormalities using MP decomposition with Gabor dictionaries. Synthetic ECG signals with controlled ST-segment deviations and additive noise contamination were analysed to evaluate the behaviour of MP-derived descriptors

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

related to energy distribution, spectral organization, temporal scale, and sparse decomposition structure. Particular attention was devoted to the relationship between ischemic morphology, sparse representability, and robustness to additive noise contamination.

II. Materials and Methods

A. ECG Signal Generation

Synthetic ECG signals were generated using the MPS450 biomedical signal simulator (Fluke Biomedical, USA). Lead II ECG recordings were acquired for five-minute intervals at a sampling frequency of 1000 Hz. ST-segment deviations were systematically varied from −0.8 mV to +0.8 mV in steps of 0.1 mV, resulting in 17 distinct ST levels.

To evaluate robustness of the proposed approach under realistic acquisition conditions, additive white Gaussian noise was introduced into the ECG signals. The signal-to-noise ratio (SNR) varied from 0 dB to 20 dB in steps of 2 dB, representing a range from highly noisy to relatively clean signal conditions.

The use of synthetic data allowed precise control over ST amplitude while maintaining consistent waveform morphology. All signals were acquired using the ADS1298 analog front-end (Texas Instruments, USA), ensuring high-resolution digitization suitable for time–frequency analysis.

B. Matching Pursuit Decomposition

Matching Pursuit (MP) decomposes signals into a set of elementary waveforms (atoms) characterized by parameters such as temporal location, energy, amplitude, frequency, phase, and scale, reflecting temporal duration. Gabor atoms are commonly employed due to their optimal time–frequency localization, making them particularly suitable for modelling SCG and ECG components.

In general, a waveform from a time–frequency dictionary can be expressed as a translation (u), dilation (s), and modulation (ω) of a window function g(t):

$$g_{\gamma=(s,u,\omega)}(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) \cdot e^{i\omega t} \quad (1)$$

In this study, the dictionary was restricted to Gabor functions of the form:

$$g(t) = A \cdot exp\left(-\frac{(t-t_0)^2}{2s^2}\right) cos(2\pi f(t-t_0) + \varphi) \quad (2)$$

where A denotes amplitude, f frequency, φ phase, t_0 the temporal position of the atom, and s the scale parameter.

The MP decomposition was performed using the SVAROG software environment. Only the first 15 MP iterations were analysed in order to focus on the dominant structures contributing to ECG morphology. The following parameters were extracted: energy, amplitude, frequency, phase, and scale, for each atom.

C. ECG components definition based on atom classification

Following decomposition, atoms were grouped according to their frequency and scale to distinguish between QRS complexes and ST segments. The QRS complex was identified as a high-frequency (8–30 Hz), short-duration (scale ≤ 0.17) component, while the ST segment was associated with low-frequency (< 8 Hz), longer-duration atoms (scale ≥ 0.8).

D. MP-derived descriptors

For each ECG recording, several MP-derived descriptors characterizing ventricular repolarization were extracted. The analysed features included: total ST energy (E_{ST}), total QRS energy (E_{QRS}), ST-to-QRS energy ratio ($R_{ST/QRS}$), iteration number of the first detected ST-related atom, number of ST atoms among the first ten MP iterations, mean ST atom frequency, mean ST atom scale, relative ST energy contribution, i.e., the ratio of the ST energy to the total energy expressed as the sum of the squares of the amplitudes of all atoms. Feature extraction was performed independently for all combinations of ST deviation and noise level. The analysis therefore captured both morphological and noise-related variability of sparse ECG representation.

E. Statistical Analysis

To investigate the joint influence of ischemic morphology and signal quality, a two-factor statistical design was applied. ST-segment deviation level and additive noise level were treated as independent factors in a two-way ANOVA model including interaction effects. Effect sizes were quantified using eta-squared statistics. Since several MP-derived descriptors violated normality assumptions, complementary nonparametric Kruskal–Wallis analysis was additionally performed.

Linear regression analysis was used to quantify directional relationships between MP-derived descriptors and signed ST-segment deviation levels. Because several energy-related descriptors exhibited partially symmetric behaviour for ST elevation and ST depression conditions, additional quadratic regression models were evaluated to investigate potential nonlinear U-shaped dependence on ST deviation.

Robustness to additive noise contamination was assessed using Spearman correlation analysis across SNR conditions.

To investigate multivariate organization of ECG morphology within the MP feature space, principal component analysis (PCA) and correlation analysis were performed using the extracted MP-derived descriptors.

Fig. 1 demonstrates the experimental workflow.

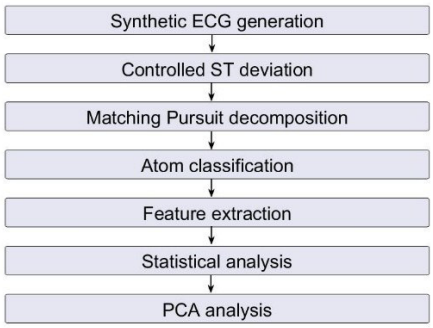


Fig. 1. Experimental workflow of ECG acquisition, Matching Pursuit decomposition, atom classification, feature extraction, and multivariate statistical analysis performed across ST-segment and noise conditions.

III. Results

A. Sparse Representation of ECG Morphology

Matching Pursuit decomposition produced sparse time-frequency representations across the investigated ST-segment

and noise conditions. Representative examples of synthetic ECG recordings for selected ST levels and additive noise conditions are shown in Fig. 2, whereas the adaptive time–frequency decompositions are presented in Fig. 3. QRS complexes were predominantly represented by short-scale, higher-frequency atoms, whereas ST-segment morphology was represented by multiple longer-scale, low-frequency atoms (Fig. 3).

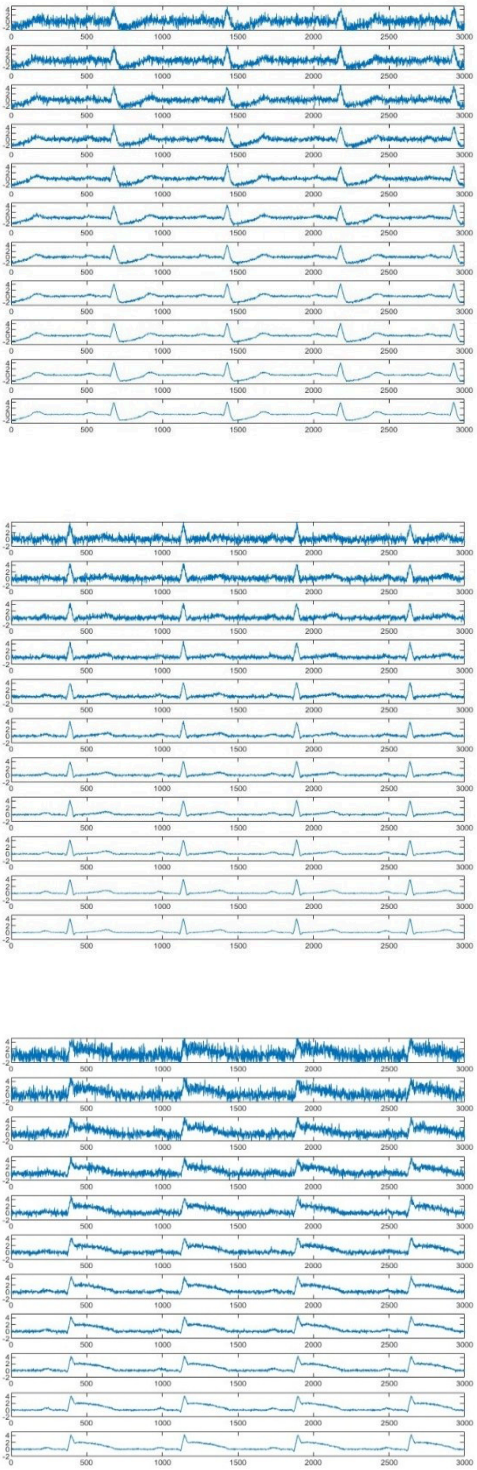


Fig. 2. Representative synthetic ECG recordings for selected ST-segment deviation levels: -0.5 mV, 0 mV, +0.5 mV, and additive noise conditions. Examples illustrate progressive changes in ventricular repolarization morphology under varying signal-to-noise ratios.

According to the atom classification criteria summarized in Table I, QRS-related atoms were typically characterized by frequencies within 8 – 30 Hz and short temporal scales ($\leqslant$ 0.17), while ST-related atoms occupied lower frequency ranges (<8 Hz) with substantially larger scales ($\geqslant$0.8).

This separation was preserved across the investigated noise conditions, supporting the use of the selected frequency- and scale-based criteria for distinguishing QRS- and ST-related components.

The resulting atom-level representation provided complementary information on energy, frequency, temporal scale, and decomposition order. These properties formed the basis for the subsequent ST-morphology descriptors.

TABLE I. DEFINITION OF ECG COMPONENTS.

ECG component	Frequency [Hz]	scale
ST	< 8	≥ 0.8
QRS	8–30	≤ 0.17

B. Statistical Dependence of MP Features on ST Deviation

Kruskal–Wallis analysis demonstrated statistically significant differences among ST levels for all analysed MP-derived descriptors ($p_{KW} < 0.001$; Table II). However, the corresponding regression results demonstrated substantial differences in explanatory strength. This distinction is important because statistical significance indicates that differences among experimental conditions are unlikely to be attributable to random variation alone, whereas R^2 quantifies the proportion of variance explained by a particular regression model.

TABLE II. LINEAR STATISTICAL ANALYSIS OF MP-DERIVED ECG FEATURES (KRUSKAL-WALLIS P-VALUES, NOISE CORRELATIONS, LINEAR REGRESSION R^2).

Feature	p_{KW}	ρ noise	p noise corr	R^2	R^2 adj
E_{ST}	< 0,001	0,016	0,841	0,014	0,008
E_{QRS}	< 0,001	0,004	0,960	0,008	0,002
$R_{ST/QRS}$	< 0,001	0,005	0,949	0,006	0,000
1st ST iteration	< 0,001	-0,037	0,628	0,001	-0,005
$N_{ST\ TOP10}$	< 0,001	0,062	0,421	0,223	0,218
mean ST frequency	< 0,001	0,110	0,154	0,354	0,351
mean ST scale	< 0,001	0,189	0,014	0,147	0,142
ST energy ratio	< 0,001	0,058	0,455	0,046	0,040

For signed ST deviation, the strongest linear relationships were obtained for mean ST atom frequency and the number of ST-related atoms detected within the first ten MP iterations (Table II and Fig. 4). Mean ST frequency achieved R^2=0.354 (adjusted R^2=0.351), whereas $N_{ST\ TOP10}$ achieved R^2=0.223 (adjusted R^2=0.218). These results indicate moderate linear dependence on signed ST deviation.

Energy-related descriptors showed weak linear relationships with signed ST level. For example, total ST energy yielded R^2=0.014 (Table II). However, these descriptors exhibited pronounced nonlinear dependence when elevation and depression were considered together. Quadratic

regression produced R^2=0.971 for total ST energy, R^2=0.686 for the ST energy ratio, and R^2=0.555 for $R_{ST/QRS}$ (Table III, Fig. 4). The nonlinear behaviour therefore reflects the approximately magnitude-dependent response of the sparse energy distribution to both positive and negative ST deviations.

Representative feature trends across ST levels and noise conditions are summarized in Fig. 5, whereas correlations between features are visualized in the heatmap representation shown in Fig. 6.

TABLE III. NONLINEAR DEPENDENCE OF ENERGY-BASED MP DESCRIPTORS ON ST-SEGMENT DEVIATION.

Feature	p-value	p ST^2	R^2	R^2 adj
E_{ST}	< 0,001	< 0,001	0,971	0,971
E_{QRS}	< 0,001	< 0,001	0,629	0,624
$R_{ST/QRS}$	< 0,001	< 0,001	0,555	0,550
ST energy ratio	< 0,001	< 0,001	0,686	0,682

C. First ST-Atom Timing

The first-ST-atom descriptor (Fig. 7) requires more careful interpretation than the other sparsity-related measures.

For the complete dataset, linear regression against signed ST deviation yielded only R^2=0.001 (adjusted R^2=−0.005, p=0.634) (Table IV). Thus, although the descriptor differed significantly among ST conditions in the nonparametric analysis, a simple linear model did not explain a meaningful proportion of its variance.

Additional analysis demonstrated that the descriptor was substantially more closely related to ST deviation magnitude. Regression against absolute ST deviation produced R^2=0.572 (adjusted R^2=0.569, $p=9.62\times10^{-33}$) (Table IV). A quadratic model using signed ST deviation also showed significant nonlinear dependence (R^2=0.371, adjusted R^2=0.363; quadratic term $p=1.80\times10^{-18}$) (Table IV). However, the absolute ST deviation model provided the stronger explanatory relationship.

The first-ST-atom timing was also examined only among recordings in which ST-related atoms were detected. In this subset (N=43), the timing showed a strong linear relationship with signed ST deviation (R^2=0.654, adjusted R^2=0.646, $p=5.29\times10^{-11}$) (Table IV). These findings suggest that first-ST-atom timing should not be interpreted as a universal linear marker of ST polarity. Rather, it appears to reflect the magnitude and detectability of the ST-related morphology, with the relationship becoming more pronounced when ST-related atoms are actually present.

TABLE IV. REGRESSION ANALYSIS OF FIRST ST-ATOM ITERATION.

Model	R^2	R^2 adj	p
Signed ST linear	0.001	−0.005	0.634
Absolute ST linear	0.572	0.569	<0.001
Signed ST quadratic	0.371	0.363	< 0.001
Detected ST atoms only	0.654	0.646	<0.001

D. Effects of Noise Contamination

Two-way ANOVA demonstrated statistically significant effects of ST deviation, noise level, and their interaction for all analysed descriptors (p<0.001; Table V). Nevertheless, the corresponding effect sizes differed substantially.

The largest ST-related effects were observed for total ST energy (η^2 ST = 0.986), first ST iteration (η^2 ST = 0.941), ST energy ratio (η^2 ST = 0.729), and mean ST frequency (η^2 ST = 0.615) (Table V).

The strongest interaction effect was observed for mean ST scale (η^2 interaction = 0.401), followed by $N_{ST\ TOP10}$ (η^2 interaction = 0.370). $R_{ST/QRS}$ and mean ST frequency showed equal interaction effect sizes (η^2 interaction = 0.357) and should therefore be regarded as tied rather than ranked sequentially (Table V).

The largest main effect of noise was observed for mean ST scale (η^2 noise = 0.133). In comparison, the noise effect sizes for total ST energy, first ST iteration, $N_{ST\ TOP10}$, and mean ST frequency were considerably smaller (Table V).

The sensitivity of mean ST scale to noise is consistent with the sequential nature of MP decomposition. Because the next atom is selected from the residual after subtraction of preceding atoms, additive noise can alter the residual structure and consequently affect the balance between shorter- and longer-scale atoms selected to represent low-frequency ST morphology. Mean scale may therefore be more sensitive to noise than descriptors based primarily on the presence or ordering of dominant ST-related atoms.

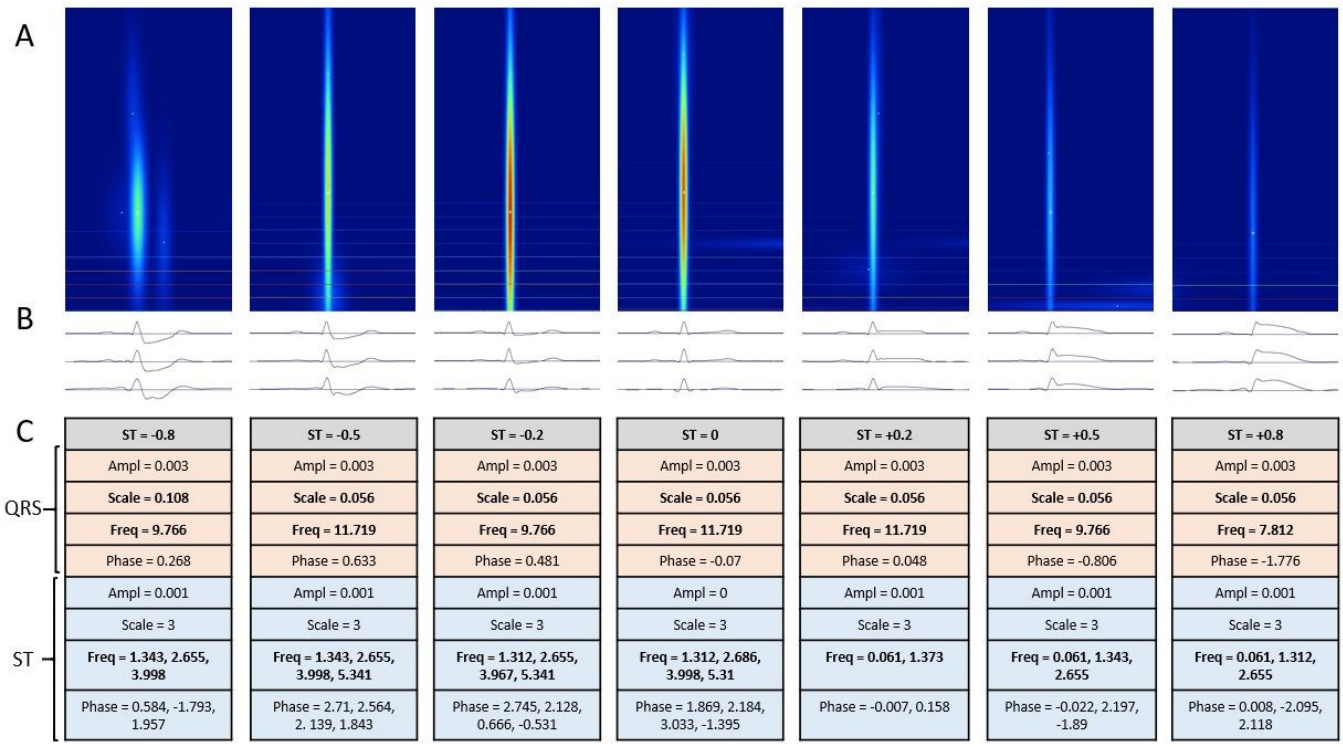
TABLE V. EFFECTS OF ST DEVIATION, NOISE LEVEL, AND THEIR INTERACTION ON MP-DERIVED FEATURES.

Feature	p ST	p noise	p inter	η^2 ST	η^2 noise	η^2 inter
E_{ST}	< 0,001	< 0,001	< 0,001	0,986	0,000	0,014
E_{QRS}	< 0,001	< 0,001	< 0,001	0,852	0,007	0,141
$R_{ST/QRS}$	< 0,001	< 0,001	< 0,001	0,613	0,030	0,357
1st ST iteration	< 0,001	< 0,001	< 0,001	0,941	0,003	0,056
N_{ST} TOP10	< 0,001	< 0,001	< 0,001	0,605	0,026	0,370
mean ST frequency	< 0,001	< 0,001	< 0,001	0,615	0,027	0,357
mean ST scale	< 0,001	< 0,001	< 0,001	0,466	0,133	0,401
ST energy ratio	< 0,001	< 0,001	< 0,001	0,729	0,013	0,258

E. Multivariate Feature-Space Structure

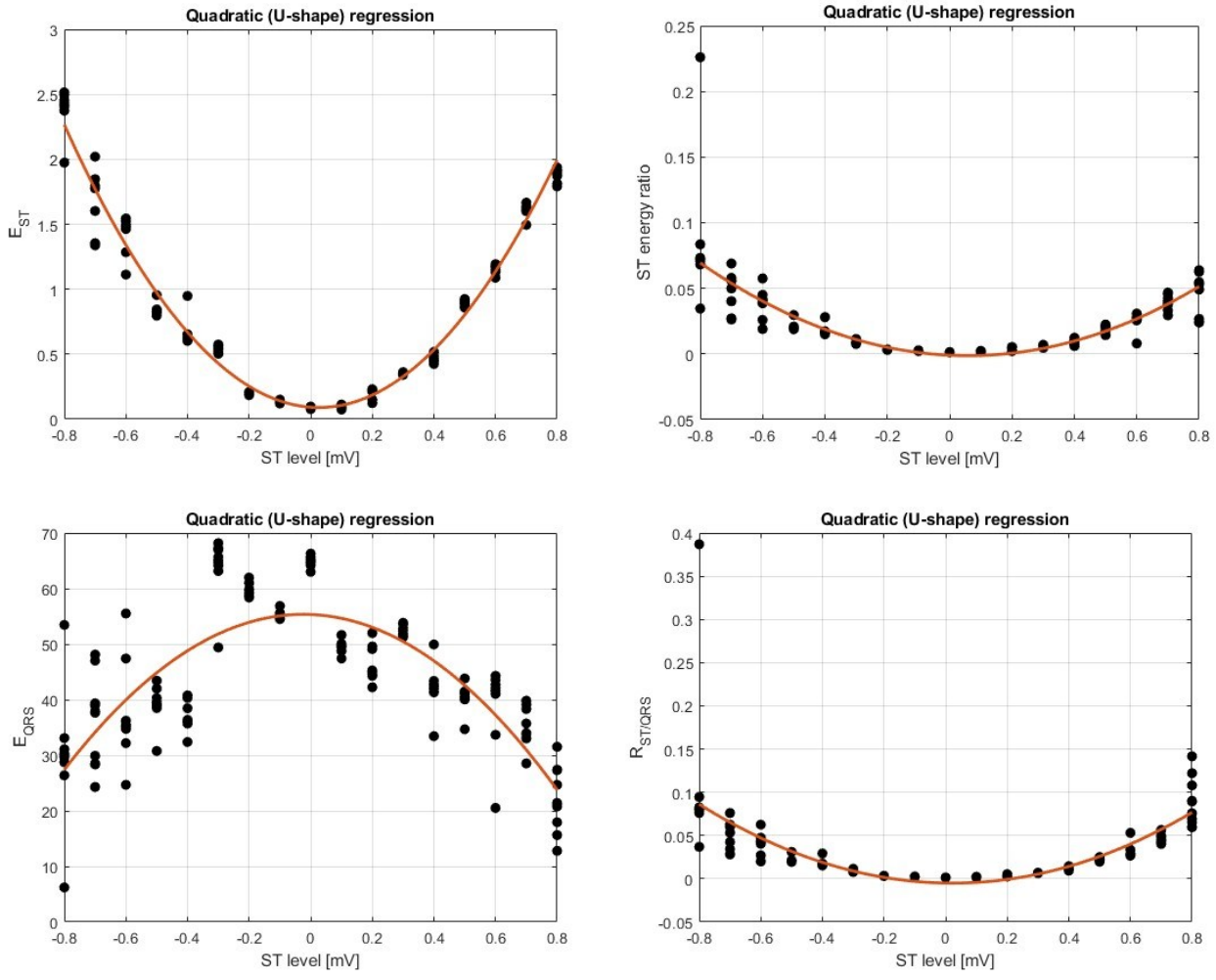
Principal component analysis demonstrated partial organization of the recordings according to ST condition in the multivariate MP feature space (Fig. 8). Separation between ST conditions remained visible across a range of noise levels, although the organization became less distinct under stronger contamination.

Correlation analysis indicated complementary relationships among energy-, frequency-, scale-, and sparsity-related descriptors (Fig. 6). In particular, the energy-based descriptors captured magnitude-related changes, whereas frequency- and sparsity-related descriptors provided additional information concerning the spectral and temporal organization of the sparse representation.



	ST = -0.8	ST = -0.5	ST = -0.2	ST = 0	ST = +0.2	ST = +0.5	ST = +0.8
QRS	Ampl = 0.003	Ampl = 0.003	Ampl = 0.003	Ampl = 0.003	Ampl = 0.003	Ampl = 0.003	Ampl = 0.003
	Scale = 0.108	**Scale = 0.056**	**Scale = 0.056**	**Scale = 0.056**	**Scale = 0.056**	**Scale = 0.056**	**Scale = 0.056**
	Freq = 9.766	**Freq = 11.719**	**Freq = 9.766**	**Freq = 11.719**	**Freq = 11.719**	**Freq = 9.766**	**Freq = 7.812**
	Phase = 0.268	Phase = 0.633	Phase = 0.481	Phase = -0.07	Phase = 0.048	Phase = -0.806	Phase = -1.776
ST	Ampl = 0.001	Ampl = 0.001	Ampl = 0.001	Ampl = 0	Ampl = 0.001	Ampl = 0.001	Ampl = 0.001
	Scale = 3	Scale = 3	Scale = 3	Scale = 3	Scale = 3	Scale = 3	Scale = 3
	Freq = 1.343, 2.655, 3.998	**Freq = 1.343, 2.655, 3.998, 5.341**	**Freq = 1.312, 2.655, 3.967, 5.341**	**Freq = 1.312, 2.686, 3.998, 5.31**	**Freq = 0.061, 1.373**	**Freq = 0.061, 1.343, 2.655**	**Freq = 0.061, 1.312, 2.655**
	Phase = 0.584, -1.793, 1.957	Phase = 2.71, 2.564, 2. 139, 1.843	Phase = 2.745, 2.128, 0.666, -0.531	Phase = 1.869, 2.184, 3.033, -1.395	Phase = -0.007, 0.158	Phase = -0.022, 2.197, -1.89	Phase = 0.008, -2.095, 2.118

Fig. 3. A. Time–frequency energy maps obtained from MP decomposition of ECG for representative ST-segment deviations. B. Original ECG signals, reconstructed waveforms and selected atoms, described by parameters (amplitude, scale, frequency and phase) provided in the corresponding tables (C).



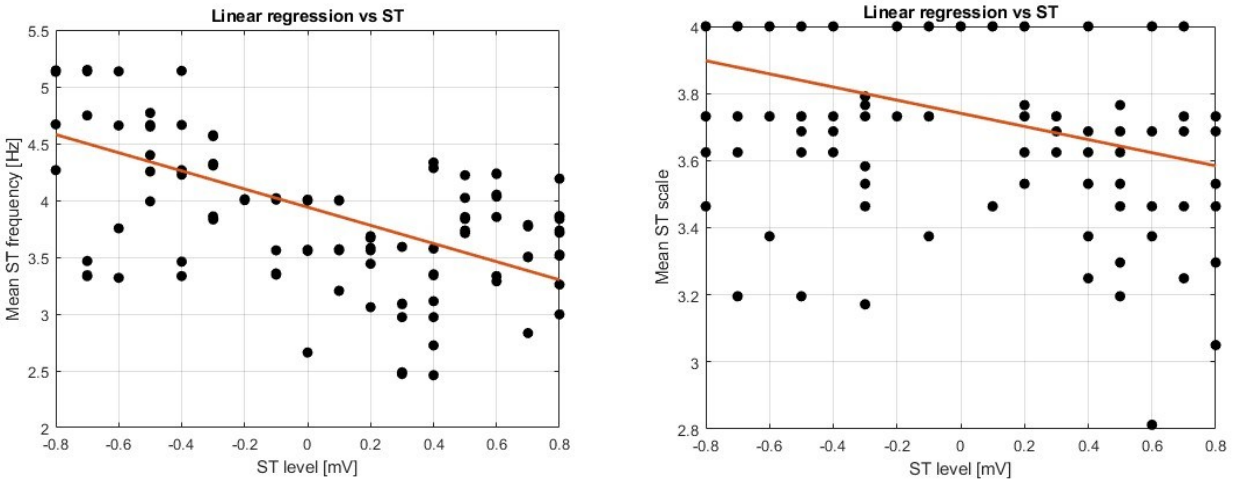


Fig. 4. Dependence of selected MP-derived ECG features on ST-segment deviation level. Linear regression curves for mean ST frequency and mean ST scale, and quadratic (U-shape) regression curves for energy-related features were shown in red.

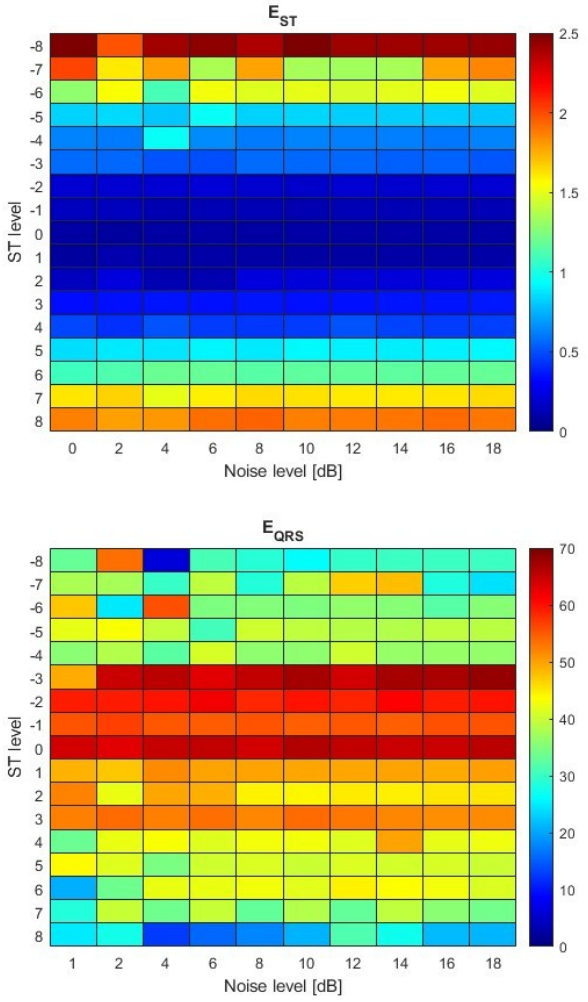


Fig. 5. Heatmap representation of MP-derived features across ST-segment levels and additive noise conditions.

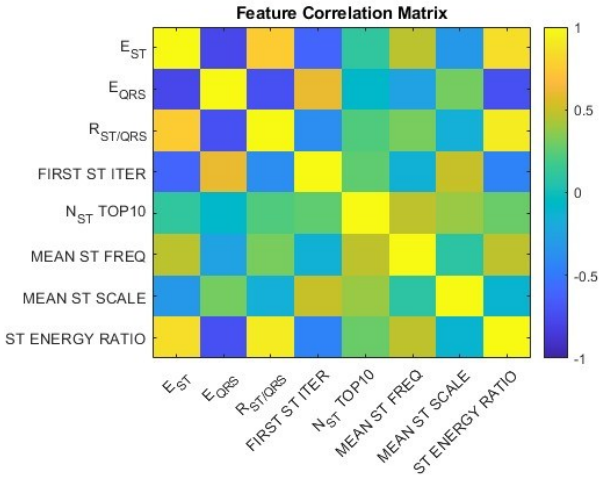


Fig. 6. Correlation matrix of MP-derived ECG descriptors showing relationships between energy-, frequency-, scale-, and sparsity-related features.

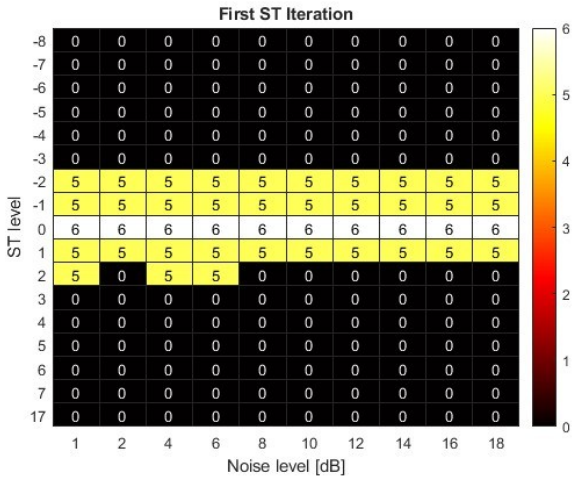


Fig. 7. Iteration order of ST-related atoms as a function of ST-segment deviation and noise level.

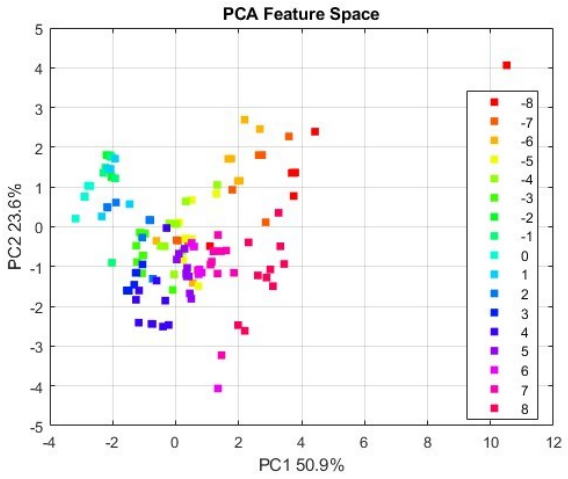


Fig. 8. Principal component analysis (PCA) of the multivariate MP feature space. Clustering according to ST-segment condition is partially preserved across varying noise levels. PC1: E_{ST}, PC2: E_{QRS}.

IV. Discussion

The present study demonstrates that Matching Pursuit provides a sparse and physiologically interpretable representation of controlled ST-segment morphology in synthetic ECG signals. An important observation is that MP does not produce a single descriptor that uniformly represents ST abnormalities. Instead, different descriptors capture different properties of the altered morphology, including energy redistribution, spectral organization, temporal scale, and the order in which ST-related structures become prominent during sparse decomposition. The ability of MP to represent signals through a small number of adaptive time–frequency atoms was originally established by Mallat and Zhang [10], while subsequent biomedical applications demonstrated the usefulness of this representation for resolving transient signal components [11]–[13].

The decomposition showed a consistent distinction between depolarization and repolarization structures. QRS complexes were predominantly represented by short-scale, higher-frequency atoms, whereas ST morphology required multiple longer-scale, low-frequency atoms. This behaviour is compatible with the substantially different temporal characteristics of ventricular depolarization and repolarization. It also illustrates an important property of the MP representation: individual atoms retain explicit information about amplitude, energy, frequency, scale, phase, and temporal location, allowing descriptors to be constructed from several complementary aspects of the decomposition. The use of Gabor-type dictionaries is particularly appropriate for such analysis because they provide joint temporal and spectral localization [10], while previous ECG studies have demonstrated the applicability of MP to QRS detection and ECG decomposition [15], [16]. Sparse ECG decompositions have also previously been investigated for separating cardiac components and representing ventricular and atrial activity [20].

The energy-related descriptors demonstrated the strongest dependence on ST deviation magnitude. Although their linear relationships with signed ST level were weak, quadratic models provided excellent fits, particularly for total ST energy (R^2=0.971). This behaviour is expected for descriptors that quantify the amount of energy assigned to ST-related atoms because both ST elevation and ST depression represent deviations from the reference morphology. Consequently, an energy-based descriptor can respond strongly to the magnitude of the perturbation while remaining relatively insensitive to its sign. The results therefore indicate that energy descriptors should be interpreted primarily as measures of ST-abnormality magnitude, rather than as direct indicators of ST elevation versus depression. This interpretation is consistent with previous work showing that sparse ECG representations can provide information complementary to conventional waveform measurements [20], [23].

Frequency-related descriptors provided complementary information. Mean ST atom frequency demonstrated the strongest linear dependence among the analysed MP descriptors (R^2=0.354), while also showing a relatively small noise effect. This suggests that changes in the spectral organization of the ST-related representation may contain information about the direction or structural characteristics of the imposed ST modification that is not captured by energy alone. Time–frequency analysis is particularly relevant to ECG because the signal contains transient and nonstationary components whose spectral characteristics change over time [10]–[13]. The present findings therefore support the use of MP-derived frequency descriptors as complementary descriptors of ST morphology, rather than as replacements for conventional ST-amplitude measurements.

The interpretation of sparsity-related descriptors requires greater caution. The number of ST-related atoms within the first ten iterations demonstrated a moderate linear relationship with signed ST deviation (R^2=0.223). In contrast, the first-ST-atom timing had essentially no explanatory power in a simple signed linear model (R^2=0.001). Importantly, the additional analysis showed a much stronger relationship with absolute ST deviation (R^2=0.572) and a significant nonlinear quadratic dependence. Furthermore, among recordings in which ST-related atoms were actually detected, the timing relationship reached R^2=0.654. These results indicate that the earlier appearance of ST-related atoms should be interpreted as an observed property of the sparse representation associated mainly with the magnitude and detectability of ST alteration, rather than as evidence of a strong monotonic relationship with signed ST polarity. Previous studies of sparse ECG representations have similarly demonstrated that the ordering and characteristics of selected atoms can provide information about ECG morphology beyond conventional waveform amplitudes [15], [20], [23].

This distinction also illustrates an important statistical issue. All analysed descriptors were statistically significant in the Kruskal–Wallis analysis, but statistical significance did not imply equivalent explanatory strength. The first-ST-atom descriptor provides the clearest example: $p_{KW} < 0.001$ coexisted with a signed linear R^2 of only 0.001. Conversely, total ST energy showed both statistical significance and extremely strong nonlinear explanatory power. Therefore, the interpretation of MP descriptors should consider statistical significance together with effect size and model explanatory power. This distinction is particularly important when developing quantitative ECG descriptors, because a statistically detectable difference between experimental conditions does not necessarily imply clinically or practically useful predictive performance.

The noise analysis further demonstrated that statistical significance should not be interpreted as evidence of large practical sensitivity. Although noise produced statistically significant effects for all descriptors, most noise effect sizes were relatively small. Mean ST scale was the notable

exception, with η^2 noise = 0.133. This sensitivity may arise because scale directly describes the temporal extent of individual atoms. Additive noise modifies the residual signal used by the greedy MP selection process, potentially changing the relative selection of shorter- and longer-scale low-frequency atoms. Mean ST scale should therefore be interpreted cautiously under low-SNR conditions. Future implementations may investigate noise-aware dictionaries or aggregation strategies designed to stabilize scale-related descriptors. The importance of dictionary structure and envelope selection for MP performance has previously been demonstrated by Rozanski [18], providing additional motivation for investigating ECG-specific dictionary configurations.

Advantages and scope of the MP representation

An important advantage of MP for the present application is that it provides a multidimensional sparse representation rather than a single scalar measurement. Conventional ST amplitude primarily quantifies the magnitude of voltage displacement relative to a reference level. In contrast, the MP decomposition provides simultaneous information about energy, frequency, scale, temporal location, phase, and decomposition order. This allows several complementary descriptors of ST morphology to be derived from the same representation. The theoretical basis for such adaptive sparse decomposition was established by Mallat and Zhang [10], while the use of optimized Gabor dictionaries and practical MP implementations has subsequently been developed for biomedical signal analysis [11]–[13], [18], [19].

These properties make MP potentially useful for subsequent automated ECG analysis, particularly when the objective is to characterize morphology rather than simply measure amplitude. Previous studies have demonstrated applications of MP to ECG compression, QRS detection, and sparse representation of cardiac activity [15], [16], [20], [23]. However, the present study does not demonstrate that MP is superior to wavelet transforms, conventional ST-amplitude analysis, or other time-frequency techniques. A direct quantitative benchmark was outside the scope of the present controlled representation study. Such a comparison should be performed under identical signal and noise conditions and should evaluate sensitivity to ST deviation, robustness to noise, temporal localization, feature stability, and computational requirements.

Computational considerations

The MP decomposition used in this study was performed using the SVAROG software environment, and only the first 15 iterations were retained for feature extraction. Because the decomposition was performed externally, a directly comparable execution-time measurement was not obtained within the present MATLAB analysis pipeline. Consequently, no unsupported claim regarding real-time performance is made.

The restriction to the first 15 iterations nevertheless limits the amount of the decomposition required for the analysed descriptors. In addition, recent developments in optimized MP implementations, including the GPU-accelerated empi implementation, provide opportunities for reducing computational burden [18], [19]. A dedicated computational study using the same hardware and implementation for MP and alternative time-frequency methods would be required to establish real-time feasibility quantitatively.

Limitations

Several limitations should be considered. First, the study was performed exclusively using synthetic ECG signals with controlled ST morphology and additive white Gaussian noise. This design provides excellent control over the imposed ST deviation and noise level, but it does not reproduce the nonlinear dynamics, inter-subject variability, anatomical differences, respiration-related modulation, baseline wander, electrode-motion artefacts, conduction abnormalities, or other sources of variability present in clinical recordings. Previous ECG signal-processing literature emphasizes the complexity of real ECG morphology and the importance of robust processing under noisy clinical conditions [21], [24-25]. Second, the analysis was performed using a single ECG lead. Multi-lead recordings may provide additional spatial information about repolarization abnormalities. Third, the present work investigates representation characteristics rather than clinical diagnostic performance. The imposed ST deviation should therefore be regarded as a controlled morphological modification of the simulated ECG waveform, rather than as a direct clinical representation of myocardial ischemia. Studies of subtle ECG changes associated with acute coronary syndromes indicate that clinically relevant repolarization abnormalities can involve features beyond simple ST amplitude [22], but the present experiment was not designed to reproduce such clinical complexity. Fourth, the present study did not quantitatively compare MP with alternative time-frequency or feature-extraction techniques. Such benchmarking, together with validation on clinical databases, is an important direction for future research.

Overall, the findings support the use of MP as a candidate representation for further investigation of ECG repolarization morphology, while the synthetic and controlled nature of the present experiment requires that the results be interpreted as methodological evidence rather than clinical validation.

V. CONCLUSION

This study investigated Matching Pursuit decomposition for sparse characterization of controlled ST-segment modifications in synthetic ECG signals. The results demonstrate that MP provides complementary descriptors of ST morphology, including energy-, frequency-, scale-, and sparse-decomposition characteristics.

Energy-related descriptors primarily reflected the magnitude of ST deviation and showed strong nonlinear relationships, with quadratic R^2 reaching 0.971 for total ST energy. Mean ST frequency and the number of ST-related atoms in the first ten iterations provided complementary information related to the signed ST modification. The first-ST-atom timing required more cautious interpretation: its signed linear relationship was negligible (R^2=0.001), whereas its relationship with absolute ST deviation was considerably stronger (R^2=0.572), and the relationship was stronger when analysis was restricted to conditions in which ST-related atoms were detected.

The analysis also demonstrated that most MP descriptors were relatively insensitive to additive noise in terms of effect size, whereas mean ST scale was the most noise-sensitive descriptor (η^2 noise = 0.133). These findings suggest that different MP descriptors have different levels of robustness and should therefore be selected according to the intended application.

The present results are limited to controlled synthetic ECG signals and should not be interpreted as evidence of clinical diagnostic performance. Quantitative comparison with conventional ST-amplitude and wavelet-based approaches, validation using clinical multi-lead ECG databases, and systematic assessment of computational requirements are important next steps. Nevertheless, the findings demonstrate that MP can provide a physiologically interpretable sparse representation of controlled ST morphology and support further development of ECG-oriented dictionaries and MP-derived descriptors for clinical ECG analysis.

ACKNOWLEDGMENT

This work was partially funded by the European project PROVIDE ("Prediction and prevention of cardiovascular diseases in pre- and type 2 diabetes", Grant Agreement no. 101128983, call topic EU4H-2022-PJ-1). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] G. W. Reed, J. E. Rossi, and C. P. Cannon, "Acute myocardial infarction," *Lancet*, vol. 389, pp. 197–210, 2017.

[2] G. S. Wagner *et al.*, "AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part VI: Acute ischemia/infarction," *J. Amer. Coll. Cardiol.*, vol. 53, no. 11, pp. 1003–1011, 2009.

[3] K. Thygesen *et al.*, "Fourth universal definition of myocardial infarction," *Circulation*, vol. 138, pp. e618–e651, 2018.

[4] B. Ibanez *et al.*, "2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation," *Eur. Heart J.*, vol. 39, no. 2, pp. 119–177, 2018.

[5] M. J. Junttila *et al.*, "Clinical significance of variants of J-points and J-waves: Early repolarization patterns and risk," *Eur. Heart J.*, vol. 33, pp. 2639–2643, 2012.

[6] R. A. Byrne *et al.*, "2023 ESC Guidelines for the management of acute coronary syndromes," *Eur. Heart J.*, vol. 44, no. 38, pp. 3720–3826, 2023.

[7] X. Zhao *et al.*, "Reliable detection of myocardial ischemia using machine learning based on temporal-spatial characteristics of electrocardiogram and vectorcardiogram," *Front. Physiol.*, vol. 13, Art. no. 854191, 2022.

[8] S. T. Duca *et al.*, "T-wave analysis on the 24 h Holter ECG monitoring as a predictive assessment of major adverse cardiovascular events in patients with myocardial infarction: A literature review and future perspectives," *Life*, vol. 13, Art. no. 1155, 2023.

[9] C. R. Vázquez-Seisdedos, J. E. Neto, E. J. Marañón Reyes *et al.*, "New approach for T-wave end detection on electrocardiogram: Performance in noisy conditions," *BioMed. Eng. OnLine*, vol. 10, Art. no. 77, 2011.

[10] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[11] P. J. Durka, K. J. Blinowska, and J. Zygierewicz, "Application of the matching pursuit method for the time-frequency analysis of EEG signals," *Med. Biol. Eng. Comput.*, vol. 38, no. 4, pp. 402–409, 2000.

[12] K. J. Blinowska and P. J. Durka, "Unbiased high resolution method of EEG analysis in time-frequency space," *Acta Neurobiol. Exp.*, vol. 61, pp. 157–174, 2001.

[13] R. Kus, P. Rozanski, and P. Durka, "Multivariate matching pursuit in optimal Gabor dictionaries: Theory and software with interface for EEG/MEG via Svarog," *BioMed. Eng. OnLine*, vol. 12, Art. no. 94, 2013.

[14] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Systems and Computers*, vol. 1, pp. 40–44, 1993.

[15] C. Hernando-Ramiro, L. Lovisolo, F. Cruz-Roldan, and M. Blanco-Velasco, "Matching pursuit decomposition on electrocardiograms for joint compression and QRS detection," *Circuits Syst. Signal Process.*, vol. 38, pp. 2653–2676, 2019.

[16] M. O. Rahman, P. Augustyniak, and E. Olejarczyk, "QRS detection based on the matching pursuit algorithm," in *Proc. Joint 20th Nordic-Baltic Conf. Biomed. Eng. and 24th Polish Conf. Biocybernetics Biomed. Eng.*, 2025.

[17] S. Baliarsingh and M. Mohanty, "Electrocardiogram compression using adaptive decomposition techniques for smart healthcare system," *Cureus J. Eng.*, vol. 2, Art. no. es44388-024-02145-3, 2025.

[18] P. Rozanski, "Effects of envelope and dictionary structure on the performance of matching pursuit," *IET Signal Process.*, vol. 14, no. 2, pp. 89–96, 2020.

[19] P. T. Rozanski, "empi: GPU-accelerated matching pursuit with continuous dictionaries," *ACM Trans. Math. Softw.*, vol. 50, no. 3, Art. no. 17, 2024.

[20] O. D. Escoda *et al.*, "Ventricular and atrial activity estimation through sparse ECG signal decompositions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 2, pp. II-–II-, 2006.

[21] L. Sörnmo and P. Laguna, "Electrocardiogram (ECG) signal processing," in *Wiley Encyclopedia of Biomedical Engineering*, M. Akay, Ed. Hoboken, NJ, USA: Wiley, 2006.

[22] R. Gunaseelan *et al.*, "Early/subtle electrocardiography features of acute coronary syndrome and ST-segment elevation myocardial infarction," *J. Emerg. Trauma Shock*, vol. 15, no. 1, pp. 66–69, 2022.

[23] S. Raj and K. Ch. Ray, "Sparse representation of ECG signals for automated recognition of cardiac arrhythmias," *Expert Syst. Appl.*, vol. 105, pp. 49–64, 2018.

[24] N. Adam, D. Valko, Z. Balogh *et al.*, "Comparative evaluation of filtration techniques for ECG signal denoising with emphasis on stationary wavelet transform," *Sci. Rep.*, vol. 15, Art. no. 42514, 2025.

[25] E. Olejarczyk and C. Massaroni, "Comparison of machine learning models in nonlinear and stochastic signal classification," *Appl. Sci.*, vol. 15, Art. no. 11226, 2025.

Spatiotemporal Organization of Atrial Fibrillation via Intracardiac and Surface Recordings: An Exploratory Case Series

Jâmily K. M. Soares
Program in Telecommun. Eng. (PPGET)
IFCE
Fortaleza, Brazil
https://orcid.org/0009-0006-8970-7855

Lucas Abdalah
Dept. of Teleinformatics Eng. (DETI)
UFC
Fortaleza, Brazil
https://orcid.org/0000-0002-7461-801X

Prof. Dr. Walter da C. F. Júnior
Dept. of Teleinformatics Eng. (DETI)
UFC
Fortaleza, Brazil
https://orcid.org/0000-0002-0821-3460

***Abstract*—Atrial fibrillation (AF) is the most prevalent sustained arrhythmia in clinical practice, and characterizing its substrate dynamics remains a central challenge in electrophysiology. This study presents an exploratory framework evaluating recurrence percentage (Rec%) in local atrial electrograms (EGMs) alongside surface electrocardiogram (ECG) complexity. To isolate atrial activity, Adaptive Singular Value Cancellation (ASVC) was applied to remove ventricular components. The pipeline was evaluated on a subset of a publicly accessible Intracardiac Atrial Fibrillation Database ($n = 3$ patients, comprising 12 synchronized mapping points across four anatomical sites: AFW, IVC, SVC, and TVA). Morphological recurrence in EGMs was compared against surface Sample Entropy (SampEn). In this exploratory cohort, an aggregated inverse correlation (Pearson $r = -0.38$, $p < 0.001$, $R^2 = 0.148$) was observed between intracardiac Rec% and surface SampEn. These preliminary findings suggest that surface complexity metrics may reflect intracardiac substrate regularity, as well the potential of pairing both approaches, establishing a reproducible computational baseline for future investigation in larger, clinically stratified cohorts.**

***Index Terms*—Atrial fibrillation, recurrence analysis, biomedical signal processing, electrocardiogram, catheter ablation.**

I. Introduction

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia. It is characterised by chaotic and disorganised atrial activation, which impairs cardiac mechanical function [1]. As well as reducing patients' quality of life, AF is associated with an increased risk of stroke, cardiac insufficiency and mortality [2]. It accounts for a growing number of hospitalisations, particularly in an ageing population [2].

Treatment for AF has evolved from pharmacological therapies towards more invasive interventions, such as catheter ablation (CA), which aims to isolate or eliminate the tissue responsible for sustaining the arrhythmia. However, precisely identifying ablation targets remains a considerable challenge. During electrophysiological mapping, complex fractionated atrial electrograms (CFAEs) are often encountered and may indicate regions of critical slow conduction or re-entrant wave pivots [3]. However, the visual interpretation of these signals is subjective and does not consistently correlate with procedural success.

Conventional mapping techniques using dominant frequency (DF) or CFAE detection often fail to distinguish between potentials representing the arrhythmia's source (e.g. rotor cores) and those resulting from passive conduction or noise [4]. In persistent AF, rotor cores may appear outside the pulmonary veins, which substantially reduces the success rate of conventional ablation procedures.

To overcome these limitations, nonlinear dynamical system analysis offers a promising approach. Recurrence analysis (RA) and recurrence percentage (Rec%) allow visualisation and quantification of state recurrences in a phase-space trajectory. When applied to biomedical signals, these methods can measure the stability and morphological regularity of electrical activations over time, independently of signal complexity in the time or frequency domain [6].

This work presents a reproducible computational framework to quantify local atrial activation regularity using recurrence percentage (Rec$\%$) and assesses its covariation with non-invasive surface ECG complexity (SampEn) following adaptive ventricular cancellation. Rather than claiming definitive clinical guidance, this exploratory study investigates whether localized intracardiac morphological regularity is reflected on the thoracic surface within a preliminary case series ($n = 3$), establishing a baseline methodology for future longitudinal and patient-specific validation.

II. Methodology

A. Dataset

The Intracardiac Atrial Fibrillation Database (IAFDB), publicly available on PhysioNet, was used in this study. To evaluate the feasibility of the proposed signal processing framework, the analysis pipeline was applied to an exploratory subset of $n = 3$ patients (IAF1, IAF2, and IAF3), comprising both paroxysmal and persistent AF subtypes. Each patient possesses recordings acquired from four standard catheter positions: against the atrial free wall (AFW), near the inferior vena cava (IVC), near the superior vena cava (SVC), and near

the tricuspid valve annulus (TVA), totaling 12 synchronized spatial mapping points.

Each record contains simultaneously acquired multi-channel signals sampled at 1 kHz:

- Intracardiac EGM: five bipolar channels from a decapolar catheter (CS12, CS34, CS56, CS78, CS90).
- Surface ECG: four standard leads (I, II, V1, aVF).

Raw data were stored in WFDB format and loaded using the `rdsamp` function from the WFDB MATLAB Toolbox. Table I summarises the main dataset characteristics.

TABLE I
DATASET CHARACTERISTICS

Parameter	Description
Patients	3 (IAF1, IAF2, and IAF3)
AF type	Paroxysmal and persistent
Catheter positions	AFW, IVC, SVC, TVA
EGM channels	CS12, CS34, CS56, CS78, CS90
ECG leads	I, II, V1, aVF
Sampling frequency	1000 Hz
Segment duration	16.77 s

B. EGM Preprocessing

The raw intracardiac EGM signals were bandpass filtered using a 4th-order Butterworth filter with cut-off frequencies of 30 and 300 Hz. This passband enhances local atrial activations while attenuating baseline drift and high-frequency noise. The filtered signals were subsequently divided into non-overlapping segments for recurrence analysis.

C. Ventricular Activity Cancellation (ASVC)

Surface ECG recordings are dominated by ventricular activity (QRST complexes), which must be removed before atrial activity (AA) can be analysed. The Adaptive Singular Value Cancellation (ASVC) method [9] was employed.

Initially, ECG signals were bandpass filtered between 0.5 and 40 Hz using a 3rd-order Butterworth filter. R-peaks were then detected using the Pan–Tompkins algorithm [7], which combines bandpass filtering, derivative filtering, squaring, moving-window integration, and adaptive thresholding. Peak locations were refined through a local maximum search within a ± 50 ms window.

For each detected beat, a QRST segment extending from $-0.3RR_{\min}$ to $+0.7RR_{\min}$ relative to the R-peak was extracted. The resulting beat matrix was decomposed using Singular Value Decomposition (SVD),

$$X = USV^T \tag{1}$$

where the first singular vector represents the dominant ventricular morphology. For each beat k, an adaptive scaling coefficient was estimated according to the minimum mean-square error criterion,

$$\alpha_k = \frac{x_k^T u_1}{u_1^T u_1} \tag{2}$$

and the atrial activity estimate was obtained as

$$\hat{x}_{AA,k} = x_k - \alpha_k u_1. \tag{3}$$

To minimise boundary artefacts, Gaussian windowing and interpolation were applied to the cancelled segments, followed by low-pass filtering and DC removal. Signal quality was verified through Welch power spectral density estimation. A dominant frequency of 4.88 Hz, within the physiological AF range (3–12 Hz), confirmed preservation of atrial dynamics after cancellation.

D. Recurrence Analysis and Rec% Calculation

For each EGM segment, atrial activation peaks were detected using a threshold equal to 1.2 times the signal RMS value and a minimum inter-peak distance of 50 ms. The same procedure was applied to the AA signal extracted from the ECG after ASVC.

For each pair of detected activations (i, j), a 100 ms window centred on each peak was extracted and zero-mean normalised. Morphological similarity was quantified using the maximum normalised cross-correlation over lag,

$$RM(i,j) = \max_{\tau} \left[\frac{(\tilde{x}_i \star \tilde{x}_j)(\tau)}{\|\tilde{x}_i\| \|\tilde{x}_j\|} \right]. \tag{4}$$

The resulting $N \times N$ matrix constitutes the recurrence map (RM). A correlation threshold of 0.8 was applied to define recurrence. The recurrence percentage (Rec%) was computed as

$$\mathrm{Rec}\% = \max_j \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[RM(i,j) > 0.8] \right) \times 100, \tag{5}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function.

A Rec% close to 100% indicates highly organised and periodic activation patterns, whereas lower values (approximately 35–50%) are characteristic of chaotic atrial fibrillation activity. Intermediate values around 80–90% suggest substantial electrophysiological reorganisation during ablation.

E. Temporal Classification

To enhance the temporal resolution of the dynamics, Rec% was computed continuously over successive non-overlapping 5 s segments, providing a denser time series to track substrate stabilization. The noise level was estimated as

$$\sigma = \mathrm{median}\left(|\mathrm{Rec}\%_{i+1} - \mathrm{Rec}\%_i| \right). \tag{6}$$

A Savitzky–Golay filter with a 9-point window and second-order polynomial approximation was applied to smooth the data. Based on a threshold of 4σ, channels were classified into three patterns:

- Type I (Step): abrupt increase greater than 4σ within a 5-minute interval.
- Type II (Gradual): cumulative increase greater than 4σ from beginning to end without a sudden step.
- Type III (No Response): absence of significant increase in Rec%.

F. Cross-Validation: EGM Rec% vs. ECG SampEn

To evaluate the feasibility of non-invasive monitoring, intracardiac Rec% values were compared with ECG complexity quantified through Sample Entropy (SampEn) across the pooled $n = 12$ synchronized spatial mapping pairs. SampEn was calculated for non-overlapping 1.5-second ECG windows using embedding dimension $m = 2$ and tolerance $r = 0.2\sigma_{\text{segment}}$.

Temporal synchronisation was achieved by averaging all SampEn values contained within each corresponding EGM segment, generating matched pairs $(Rec\%_k, \text{SampEn}_k)$.

The association between both measures was evaluated using Pearson's correlation coefficient (r), Spearman's rank correlation coefficient (ρ), and the coefficient of determination (R^2).

Because this exploratory study evaluates a pilot subset ($n = 3$ subjects, 12 spatial mapping pairs), the pooled correlation analysis serves primarily as an aggregate descriptive assessment of trends across recorded sites. We explicitly acknowledge that pooling intra-subject spatial points across time windows does not satisfy strict sample independence assumptions; therefore, the resulting correlation metrics (r, R^2, p-value) should be interpreted as proof-of-concept indicators of signal covariation rather than population-wide inferential statistics.

III. Results

A. Signal Preprocessing Quality

The bandpass-filtered EGM signal (30–300 Hz) exhibited well-defined atrial activation peaks with stable baseline and low-amplitude background noise, consistent with high-quality intracardiac recordings (Fig. 1). Peak amplitudes ranged between -0.5 mV and $+0.5$ mV, and the morphological consistency of successive activations over the analysed 10-second window suggested moderate-to-high atrial organisation at the selected catheter position (AFW).

For the surface ECG, the ASVC pipeline successfully removed ventricular interference. The dominant frequency (DF) of the extracted atrial activity signal was 4.88 Hz, which falls within the physiological range of AF (3–12 Hz), confirming that the processing preserved the intrinsic atrial dynamics without introducing spectral distortion. The Pan–Tompkins algorithm reliably detected R-peaks across the full 120-second recording, maintaining stable adaptive thresholds even in the presence of baseline variation.

B. Signal and Recurrence Metrics

Table II summarises the acquisition parameters, recurrence metrics, and statistical results obtained for the representative case IAF1-AFW.

C. Rec% Temporal Dynamics

The temporal evolution of Rec% during the ablation procedure showed increasing atrial organisation over time. Although the raw Rec% values exhibited abrupt fluctuations attributable to local tissue dynamics and catheter repositioning between recording segments, the Savitzky–Golay smoothed trend (9-point window, second-order polynomial) revealed a progressive increase in atrial electrical organisation, rising from approximately 55% in the early segments to a final value of 83.87% in the last analysed segment (Fig. 2).

TABLE II
Signal and Recurrence Metrics for Patient IAF1 at Position AFW

Parameter	Value
Catheter position	AFW (atrial free wall)
EGM channels	CS12, CS34, CS56, CS78, CS90
Sampling frequency	1000 Hz
Segment duration	5.0 s
EGM filter	30–300 Hz Butterworth (4th order)
ECG filter	0.5–40 Hz Butterworth (3rd order)
Dominant frequency	4.88 Hz
Recurrence threshold	$r > 0.8$
Final Rec%	83.87%
Pearson r	-0.38
p-value	< 0.001
R^2	0.148
n (position pairs)	12
Classification	Type III (No Response)

According to the 4σ classification criterion, where $\sigma = \text{median}|Rec\%_{i+1} - Rec\%_i|$, this channel was classified as *Type III (No Response)*. Although the increase was physiologically meaningful, it did not reach the abrupt transition threshold required to characterise a Type I response nor the sustained global increase associated with a Type II response. This finding is consistent with a gradual sub-threshold reorganisation of the atrial substrate at the free-wall position, where an arrhythmogenic driver is unlikely to be anchored.

D. Recurrence Map Analysis

The $N \times N$ recurrence map computed for the final segment is shown in Fig. 3. The map is predominantly composed of high-similarity regions (red-orange tones, correlation $r > 0.8$), with symmetric patterns and repetitive blocks along both axes. Low-recurrence regions ($r < 0.5$) are sparse, indicating minimal morphological variability between successive atrial complexes.

The Rec% value of 83.87% reflects a high degree of electrical regularity, approaching the values expected for organised rhythms such as atrial flutter (Rec% $\approx$ 95–100%) and substantially exceeding the recurrence levels typically observed during active AF (Rec% $\approx$ 35–50%). This confirms that the atrial substrate at this position had undergone significant electrophysiological reorganisation by the end of the recording period.

E. EGM vs. ECG Multi-Patient Exploratory Analysis

To assess the potential correspondence between invasive and non-invasive measures, cross-correlation analysis was carried out on the pooled data from the $n = 12$ synchronized spatial mapping points. The aggregate analysis revealed an inverse relationship between both metrics, with a Pearson correlation coefficient $r = -0.38$ ($p < 0.001$), coefficient of determination $R^2 = 0.148$, and Spearman $\rho = -0.35$ (Fig. 4).

The pooled analysis yielded an inverse correlation of $r = -0.38$ with a coefficient of determination $R^2 = 0.148$. This

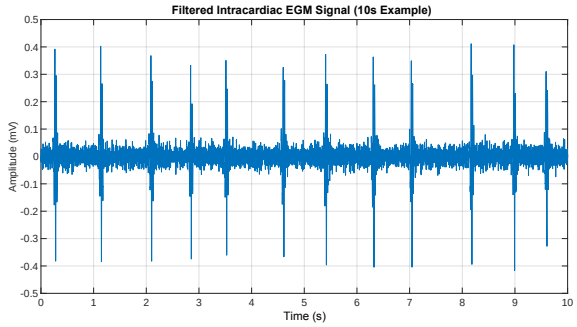


Fig. 1. Bandpass-filtered intracardiac EGM signal (30–300 Hz) for patient IAF1 at the AFW position, showing well-defined atrial activation peaks over a 10-second window.

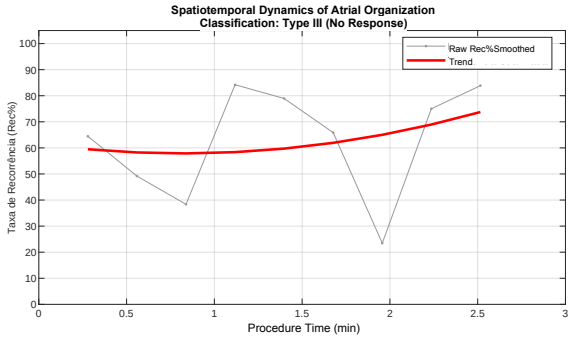


Fig. 2. Temporal evolution of Rec% for the IAF1-AFW channel. Raw values (dashed) and Savitzky–Golay smoothed trend (solid) show a progressive increase from approximately 55% to 83.87%.

indicates that roughly 15% of the variance observed in surface SampEn is linearly associated with local intracardiac Rec%. This moderate strength is physically expected: surface ECG leads integrate electrical wavefronts across the entire atrial syncytium through the thoracic volume conductor, whereas the bipolar EGM captures only highly localized near-field activation. Far from indicating a failure of the metric, this divergence confirms that surface dynamics reflect a global superposition of multiple localized substrate regions.

Although the nominal p-value indicates statistical significance within the pooled data points, these figures reflect an exploratory proof-of-concept across 12 paired recordings from three patients rather than population-wide inference. Patient IAF1-AFW was selected as the representative case because it encompasses the complete analysis pipeline and yields a physiologically interpretable Rec% value of 83.87%, providing a clear contrast with the recurrence levels characteristic of active AF.

IV. Discussion

The results demonstrate that Rec%, derived from recurrence analysis of intracardiac electrograms (EGMs), provides a quantitative and physiologically interpretable measure of atrial substrate organisation during catheter ablation. The progressive increase in Rec% observed in the representative case (IAF1-AFW), from approximately 55% in the initial segments to 83.87% in the final segment, is consistent with the gradual electrophysiological reorganisation expected as ablation modifies the arrhythmogenic substrate.

Physiologically, the inverse correlation between EGM Rec% and ECG SampEn is explained by the physics of volume conduction. During active, chaotic AF, multiple wandering wavelets and unstable rotors propagate randomly through the atria. This localized desynchronization generates highly fractionated electrograms (CFAEs) and, when projected onto the body surface, manifests as highly irregular and complex fibrillatory waves (resulting in low Rec% and high SampEn). Conversely, when catheter ablation successfully terminates or anchors these rapid drivers, the localized atrial tissue undergoes electrophysiological synchronization. The resulting regular, repetitive activation wavefronts are projected through the thoracic volume conductor, translating into highly regular and uniform f-waves on the chest surface, which significantly lowers the SampEn.

It is crucial to clarify that the current framework quantifies local morphological regularity rather than directional activation vectors or driver causality. Consequently, high Rec% values or decreased surface entropy cannot be interpreted as direct identification or spatial localization of rotor cores. A localized increase in Rec% simply reflects that the underlying tissue is undergoing repetitive, morphologically stable activations—which may arise either from proximity to an anchored driver or from regular passive activation by distant sources.

Compared to traditional spectral metrics such as Dominant Frequency (DF) and classical CFAE mean-interval algorithms, the proposed recurrence formulation presents distinct theoretical advantages. While DF mapping relies heavily on Fourier assumptions that are susceptible to harmonic distortion, spectral leakage, and amplitude variations, Rec% relies purely on morphological cross-correlation thresholds, making it robust against amplitude scaling and baseline wander. Unlike subjective CFAE fractionation criteria, Rec% yields a bounded, continuous metric (0–100%) directly grounded in non-linear dynamical stability.

From a computational perspective, the pipeline relies on low-order linear filtering, standard Pan-Tompkins detection, rank-1 SVD projection for ventricular cancellation, and pairwise cross-correlation over short sliding windows (5 s). Given that both SVD truncation and signal correlation have standard polynomial complexities that scale efficiently with window length, the pipeline is well-suited for near real-time execution on standard clinical workstations without requiring high-performance computing clusters.

The temporal classification framework also provides clinically relevant descriptive information. In the representative case, the AFW channel was classified as Type III (No Response), indicating that the observed increase in Rec% did not exceed the 4σ threshold required for a Type I or Type II response. Type I patterns, characterised by abrupt increases in Rec%, are of particular interest as markers of localized

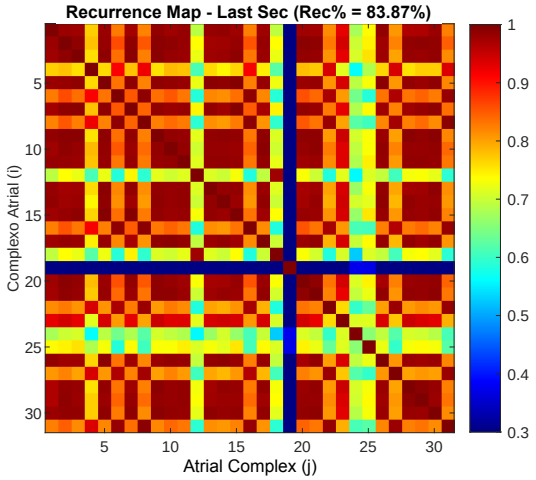


Fig. 3. Recurrence map (RM) computed for the final EGM segment of patient IAF1-AFW. High-similarity regions (warm tones, $r > 0.8$) dominate, indicating a Rec% of 83.87%.

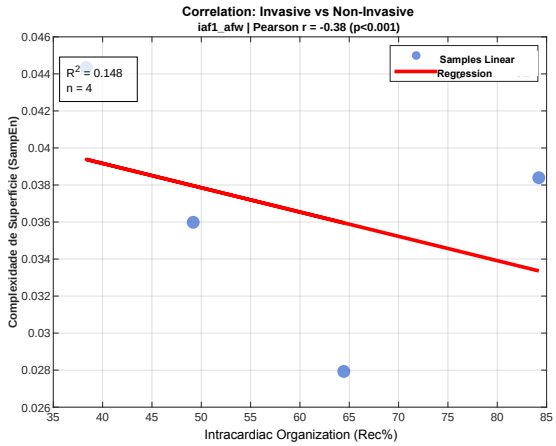


Fig. 4. Scatter plot of intracardiac Rec% versus surface ECG SampEn across 12 exploratory position pairs. Note: Data points represent pooled spatial positions; axes labels correspond to intracardiac recurrence (%) and surface sample entropy.

substrate reorganization, whereas Type II patterns suggest gradual stabilization and Type III patterns indicate persistent irregularity.

Compared with the Granger Causality (GC) framework proposed by Alcaine et al. [8], which evaluates temporal predictability and directional interactions among EGM channels, Rec% quantifies intra-channel morphological consistency. Thus, the two approaches are complementary: combining recurrence-based metrics with directional causality measures represents a logical pathway to capture both morphological morphology and propagation networks.

This study has clear limitations that define its exploratory scope. Foremost is the restricted cohort size ($n = 3$ patients, 12 recording positions), which precludes broad clinical generalizations and prevents patient-stratified hierarchical modeling (e.g., repeated-measures correlation or mixed-effects models). Second, the current dataset does not include long-term clinical ablation outcomes or post-procedure arrhythmia recurrence data, preventing any direct claim of therapeutic prognostic value. Future work will apply this pipeline across the complete 8-patient IAFDB cohort and clinical registries using repeated-measures statistical models to confirm inter-patient robustness.

V. Conclusion

This exploratory case series presented a computational framework for the spatiotemporal characterisation of atrial fibrillation using intracardiac electrograms and surface ECGs. Recurrence percentage (Rec%) proved to be a viable, bounded metric to track local morphological regularity during catheter ablation. An aggregate inverse association between intracardiac Rec% and surface SampEn was documented, physically consistent with thoracic volume conduction phenomena. While the current sample size does not allow for inferring procedural outcomes or pinpointing causal drivers, the proposed pipeline establishes a reproducible baseline for non-invasive substrate evaluation to be substantiated in larger, prospectively designed cohorts.

References

[1] P. Kirchhof et al., "2016 ESC Guidelines for the management of atrial fibrillation," Eur. Heart J., vol. 37, no. 38, pp. 2893–2962, 2016.

[2] L. Mainardi, S. Cerutti, and L. Sörnmo, Understanding Atrial Fibrillation: The Signal Processing Contribution. Morgan & Claypool, 2008.

[3] K. Nademanee, "Mapping of complex fractionated atrial electrograms as target sites for AF ablation," in Proc. 33rd IEEE EMBS, 2011, pp. 5539–5542.

[4] P. B. Tabereaux, D. J. Dosdall, and R. E. Ideker, "Mechanisms of VF maintenance: wandering wavelets, mother rotors, or foci," Heart Rhythm, vol. 6, no. 3, pp. 405–415, 2009.

[5] V. Ravikumar and E. G. Tolkacheva, "Optimizing multiscale entropy approach for rotor core identification using simulated intracardiac electrograms," in Proc. 42nd IEEE EMBC, 2020, pp. 414–417.

[6] S. Iravanian and J. J. Langberg, "Spatiotemporal organization during ablation of persistent atrial fibrillation," Heart Rhythm, vol. 12, no. 9, pp. 1937–1944, 2015.

[7] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Trans. Biomed. Eng., vol. BME-32, no. 3, pp. 230–236, 1985.

[8] A. Alcaine et al., "A multi-variate predictability framework to assess invasive cardiac activity and interactions during atrial fibrillation," IEEE Trans. Biomed. Eng., vol. 64, no. 5, pp. 1157–1168, 2017.

[9] R. Alcaraz and J. J. Rieta, "Adaptive singular value cancellation of ventricular activity in single-lead AF electrocardiograms," Physiol. Meas., vol. 29, no. 11, pp. 1351–1369, 2008.

[10] M. A. Martins et al., Clínica Médica, Vol. 2: Doenças Cardiovasculares. Manole, 2016.

Spectral Residual Analysis and Feature Map Visualization of Deep Learning Denoisers for Digital Mammography

1st Dyego A. Pereira
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
dyego.pereira@usp.br

2nd Renann F. Brandão
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
renannbrandao@usp.br

3rd Bruno Barufaldi
Department of Radiology
University of Pennsylvania
Philadelphia, USA
bruno.barufaldi@pennmedicine.upenn.edu

4th Andrew D. A. Maidment
Department of Radiology
University of Pennsylvania
Philadelphia, USA
andrew.maidment@pennmedicine.upenn.edu

5th Marcelo A. C. Vieira
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
mvieira@sc.usp.br

***Abstract*—Deep learning denoisers have shown promising results for digital mammography, but their effects on noise texture and internal representations remain unclear. By combining local or contextual information, these networks may alter noise texture and introduce spatially structured residuals, potentially affecting image appearance and the visibility of subtle findings. This study investigates how seven pretrained convolutional and transformer-based denoisers modify the spectral properties of denoised images and removed components, while examining their internal representations through feature map analysis. Realistic mammograms were generated using a virtual clinical trial framework and degraded with heteroscedastic noise. Spatial-frequency behavior was assessed using power spectral density (PSD), with mean squared error (MSE) computed over low-, mid-, and high-frequency bands. The denoisers showed different trade-offs between noise suppression and structure preservation. Models with stronger apparent noise suppression produced removed-component spectra with greater low-frequency dominance, whereas less aggressive models preserved more high-frequency content but left more visible noise. Restormer achieved the lowest low- and mid-frequency errors for the denoised-image PSD, while DRUNet achieved the lowest high-frequency error. For the removed-component PSD, DRUNet, SCUNet, and GSDRUNet showed the closest spectral agreement with the inserted noise, depending on the frequency band. Feature maps revealed strategies ranging from image reconstruction through abstract latent representations to direct residual noise modeling. These results suggest that denoising performance and structural preservation are closely related to the representations learned by each network. PSD and feature map analyses provide complementary insights into how pretrained denoisers process mammograms. The source code is available at https://github.com/LAVI-USP/CBEB_2026_RESIDUAL_FEATURE.git.**

***Index Terms*—Digital mammography, deep learning denoising, explainability, feature map analysis, power spectral density.**

This work was supported by the São Paulo Research Foundation (FAPESP; Grants No. 2021/12673-6 and 2026/18409-2), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; Finance Code 001), and the Brazilian National Council for Scientific and Technological Development (CNPq; Grant No. 303251/2025-5).

I. Introduction

Breast cancer remains a major public health concern and is among the most common causes of cancer-related mortality in women worldwide. Excluding non-melanoma skin cancers, it represented the most commonly diagnosed malignancy in 2020, accounting for approximately 11.7% of all newly reported cancer cases [1]. Numerous studies have demonstrated that routine screening programs can substantially decrease breast cancer mortality rates, particularly among women at elevated risk [2]. In this context, medical imaging plays a crucial role in early diagnosis and clinical management.

Digital mammography (DM) is currently one of the most widely adopted imaging modalities for breast cancer screening and detection. Since DM employs ionizing radiation, radiation exposure must be minimized according to the principle of keeping doses as low as reasonably achievable [3]. Nevertheless, reducing the radiation dose generally leads to increased image noise, which may compromise the visibility of subtle lesions, including microcalcifications. Consequently, there is an inherent balance between patient safety and image quality.

To address this challenge, image denoising techniques have been extensively investigated as a means of enhancing DM images while maintaining diagnostically important anatomical information [4]. Recent advances in deep learning (DL) have led to the development of numerous pretrained denoising models that achieve state-of-the-art performance on natural image restoration tasks [5]–[11]. Recently, [12] investigated the pretrained version of these models in mammography denoising within a physics-integrated model-based deep learning framework. Although promising results were obtained in terms of image quality and microcalcification detectability, the impact of spatial correlation in denoised images produced by DL methods remains unexplored. Spatially correlated noise is an important image-quality degradation in digital mammography

because it can reduce radiologists' ability to detect breast lesions, particularly subtle findings such as microcalcifications [13].

Although pretrained deep learning denoisers have recently been investigated for digital mammography, their impact on the spatial-frequency properties of restored images remains insufficiently understood. Since denoising networks combine information from neighboring pixels or broader contextual regions, they may alter noise texture and introduce spatially structured residuals. These changes can affect the visual appearance of mammographic images and may potentially influence the detectability of subtle findings, such as microcalcifications. Moreover, it remains unclear how different denoising architectures modify the spectral characteristics of the removed component and how their internal representations are related to noise suppression and structural preservation. In this work, we address this gap by combining power spectral density analysis of denoised images and removed components with feature map visualization of pretrained convolutional and transformer-based denoisers applied to digital mammography.

II. Materials and Methods

A. Image Dataset

Virtual clinical trials (VCTs) provide a framework for performing in silico experiments in medical imaging. In mammography, these simulations enable the generation of realistic breast images while allowing precise control over anatomical and acquisition parameters. In this study, a synthetic dataset was generated using a breast imaging VCT platform capable of modeling the entire imaging process, including breast anatomy generation, image acquisition, and image formation. Realistic digital mammography (DM) projections were produced using the GPU-based OpenVCT simulation framework [14].

Four anatomical scenarios were considered, each characterized by distinct tissue composition and compression conditions. For each anatomy, ten independent noise realizations were generated, yielding 40 noisy images. The corresponding four noise-free images were retained as ground truth, resulting in a total of 44 images. Since noise was added after the generation of the ideal mammograms, paired noisy and ground-truth images were available for quantitative evaluation. The characteristics of the simulated dataset are summarized in Table I.

Mammograms were synthesized by adding a Gaussian approximation of heteroscedastic Poisson–Gaussian noise to noise-free images, which is valid under high photon statistics [15]. The standard dose for the Selenia Dimensions mammography system (Hologic, Bedford, MA) was adopted. Let $y(i,j)$ denote the ground-truth image and $z(i,j)$ the corresponding noisy observation. The degradation process is described by

$$z(i,j) = y(i,j) + n(i,j), \qquad n(i,j) \sim \mathcal{N}(0, Var(i,j)), \tag{1}$$

where $n(i,j)$ represents the noise term, including quantum and electronic noise, modeled as a Gaussian random variable with zero mean and variance $Var(i,j)$.

TABLE I: Simulated phantom dataset and anatomy details.

Parameter	Value
System	Hologic
Resolution	4096×3328
Volume	450 mL
Percent Dense	25
Skin Thickness	0.15 cm
Deformation Volume	200–400 mL
Number of Anatomies	4
Number of Noise Realizations	10
Number of Ground Truth Images	4
Total Number of Images	44

Although the underlying acquisition noise is Poisson–Gaussian, under high photon statistics it can be approximated as a zero-mean Gaussian noise term with signal-dependent variance. The noise variance depends on both local signal intensity and detector position, capturing the heteroscedastic behavior of mammographic acquisitions. It is expressed as

$$Var(i,j) = y(i,j)\xi_q(i,j) + \xi_e^2, \tag{2}$$

where $\xi_q(i,j)$ represents the spatially varying quantum gain map and ξ_e denotes the standard deviation of the electronic noise component. These parameters were estimated from measurements obtained using a Selenia Dimensions mammography system (Hologic, Bedford, MA) [4]. The spatially varying $\xi_q(i,j)$ map had a mean value of 0.115, whereas ξ_e = 1.15.

B. Deep-Learning Denoisers

Seven deep learning-based (DL) denoising methods were evaluated for the restoration of VCT-generated images corrupted by heteroscedastic Poisson–Gaussian noise. These included four convolutional neural network (CNN)–based models (DRUNet, DnCNN, DSCcP, and GSDRUNet) and three hybrid CNN–Transformer architectures (Restormer, SCUNet, and SwinIR). Table II summarizes the evaluated DL approaches, including each model's architecture, whether it uses an encoder–decoder architecture and the number of channels. The table also presents whether the model requires the noise standard deviation σ as input for denoising. For models requiring a noise-level parameter, a single scalar σ was estimated independently for each input image using the median absolute deviation method and provided to the network during inference. The mean estimated value across the dataset was $\sigma = 8.57$. In this study, all methods were employed in their pretrained, off-the-shelf configurations as implemented in the DeepInv Python library [16]. For models where the grayscale (1-channel) version was not available, the grayscale channel was replicated across all three channels, and the mean was computed over the three channels after denoising. This effectively introduces an additional averaging operation, but it represents the most straightforward strategy for adapting 3-channel architectures to grayscale medical imaging. No variance-stabilizing transformation (VST) was applied before inference, since its impact under similar imaging conditions was previously evaluated in [12].

TABLE II: Denoising methods evaluated in this study.

Model	Ch.	Architecture	Enc.-Dec.	Uses σ	Params.
DRUNet	1	CNN	Yes	Yes	32.7M
Restormer	1	CNN + Transformer	Yes	No	25.31M
SwinIR	1	CNN + Transformer	No	Yes	11.8M
DnCNN	1	CNN	No	No	0.56M
SCUNet	3	CNN + Transformer	Yes	Yes	18M
DSCcP	3	CNN	No	Yes	70K
GSDRUNet	1	CNN	Yes	Yes	17M

C. Spectral Analysis

Let $y(i,j)$ denote the ground-truth image, $z(i,j)$ the corresponding noisy image, and $\hat{y}(i,j)$ the denoised image. Since the inserted noise is known in the simulated dataset, it was defined as

$$n(i,j) = z(i,j) - y(i,j), \tag{3}$$

whereas the image component removed by each denoiser was defined as

$$\hat{n}(i,j) = z(i,j) - \hat{y}(i,j). \tag{4}$$

Thus, $n(i,j)$ represents the noise expected to be removed, while $\hat{n}(i,j)$ represents the component effectively removed by the denoising network.

The two-dimensional power spectral density (PSD) was computed for the denoised images, the inserted noise, and the removed components as

$$PSD(f_x, f_y) = \frac{1}{MN} \left| \mathcal{F}\{x(i,j)\} \right|^2, \tag{5}$$

where $x(i,j)$ represents either $\hat{y}(i,j)$, $n(i,j)$, or $\hat{n}(i,j)$, $\mathcal{F}\{\cdot\}$ denotes the two-dimensional Fourier transform, and M and N are the image dimensions. The resulting PSD was radially averaged as

$$\overline{PSD}(k) = \frac{1}{N_k} \sum_{(f_x,f_y)\in\mathcal{R}_k} PSD(f_x, f_y), \tag{6}$$

where $\mathcal{R}_k$ denotes the set of frequency samples within the k-th radial frequency bin and N_k is the number of samples in that bin. The radial PSDs were then averaged across all L image realizations,

$$\left\langle \overline{PSD}(k) \right\rangle = \frac{1}{L} \sum_{l=1}^{L} \overline{PSD}_l(k), \tag{7}$$

yielding a representative radial spectrum for each denoising method.

Two complementary spectral comparisons were performed. First, the PSD of each denoised image $\hat{y}(i,j)$ was compared with the PSD of the corresponding ground-truth image $y(i,j)$, allowing the preservation of image spatial-frequency content to be assessed. Second, the PSD of the removed component $\hat{n}(i,j)$ was compared with the PSD of the inserted noise $n(i,j)$. This comparison evaluates whether the component removed by each denoiser has spectral characteristics similar to the known noise process. If a denoiser removes little noise, $\hat{n}(i,j)$ is expected to deviate from $n(i,j)$. Conversely, if the denoiser primarily removes the inserted noise, the PSD of $\hat{n}(i,j)$ should be closer to the PSD of $n(i,j)$. Deviations between these spectra may indicate underfiltering, overfiltering, or removal of image structures whose spectral content overlaps with noise.

In addition to the qualitative comparison of the radial PSD curves, the mean squared error (MSE) was computed between reference and estimated radially averaged PSDs. For the denoised-image analysis, the reference spectrum was obtained from the ground-truth image. For the removed-component analysis, the reference spectrum was obtained from the inserted noise. The PSD curves were divided into low- (0–1 mm^{-1}), mid- (1–2 mm^{-1}), and high-frequency (> 2 mm^{-1}) bands to evaluate the frequency-dependent behavior of each denoising method. These bands were selected to summarize the main low-, mid-, and high-frequency trends observed in the PSD curves.

D. Interpretability

To provide insights into the internal representations learned by each denoising network, a feature map visualization analysis was performed. For each pretrained model, all convolutional layers were first identified and ordered according to their position in the network architecture. Since visualizing every feature map from every layer would be impractical in a two-column paper, 16 convolutional layers were uniformly sampled across the network depth, allowing the evolution of learned representations to be inspected from early to late processing stages.

For a given input mammogram, the activations of the selected convolutional layers were extracted using forward hooks during inference. Let $\mathbf{A}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote the output tensor of the l-th convolutional layer, where C_l is the number of feature channels and $H_l \times W_l$ corresponds to the spatial dimensions of the activation map. The energy associated with the c-th feature map was computed as

$$E_c^{(l)} = \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \left(A_c^{(l)}(i,j) \right)^2. \tag{8}$$

The feature map with the highest energy was then selected for visualization,

$$c^* = \arg\max_c E_c^{(l)}, \tag{9}$$

and the corresponding activation map

$$F^{(l)} = A_{c^*}^{(l)} \tag{10}$$

was retained as a representative feature map for that layer.

The use of the highest-energy feature map was motivated by the assumption that channels with larger activation energy tend to contain the most prominent responses generated by the network for a given input image. This strategy avoids arbitrary channel selection while providing a compact visualization of the dominant structures being processed throughout the network.

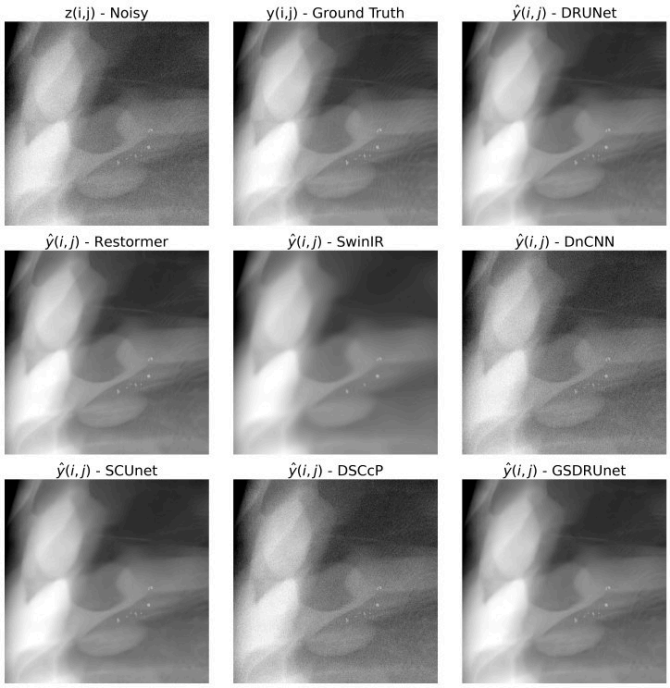


Fig. 1: Representative regions of interest (ROIs) from a synthetic mammogram: noisy (z), ground truth (y), and denoised with the deep learning denoisers, $\hat{y}$. All ROIs are displayed using the same intensity window.

III. Results and Discussion

A. Denoising

Fig. 1 shows the regions of interest in the images denoised by all deep learning denoisers. The images indicate that DnCNN and DSCcP leave more visible residual noise. SwinIR produces a visibly smoother image, whereas the remaining denoisers achieve a better trade-off between noise suppression and image blurring. These differences likely arise, at least in part, from the distinct noise distributions and noise-level parameterizations used during training. SwinIR was trained primarily for higher noise levels ($\sigma > 15$), which may explain its more aggressive denoising behavior. In contrast, DnCNN and DSCcP provide weaker apparent noise suppression under the considered noise conditions.

B. Spectral Analysis

Fig. 2 shows the PSD of the denoised images. In the PSD domain, SwinIR, DRUNet, GSDRUNet, and SCUNet exhibit the strongest attenuation of medium-to-high spatial frequencies, indicating aggressive noise suppression at the expense of fine-texture preservation. DnCNN and DSCcP remain farther from the ground-truth spectrum, suggesting weaker noise suppression and higher residual high-frequency content. Restormer retains the highest high-frequency PSD among the denoisers, consistent with greater preservation of high-frequency content, which may include both anatomical detail and residual noise, compared with the other methods. Thus, the figure reveals a continuum between noise suppression and detail preservation.

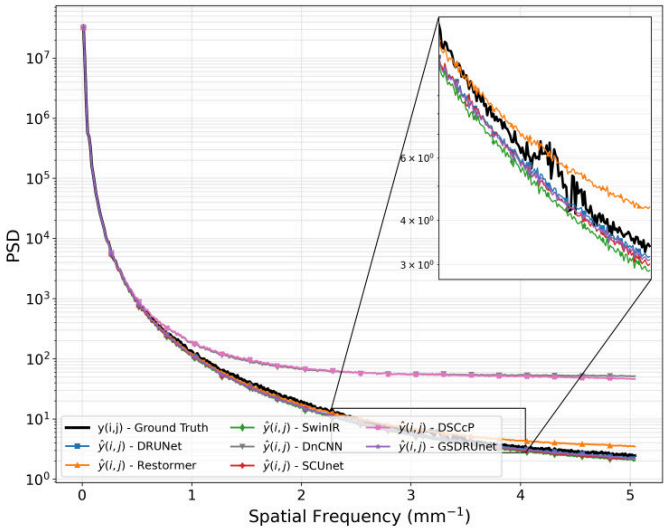


Fig. 2: Power Spectral Density (PSD) for the ground-truth and all the denoised images.

Additionally, Fig. 3 shows the removed-component PSD. For the models with weaker apparent noise suppression, such as DnCNN and DSCcP, the PSD of the removed component remains farther from the reference inserted-noise PSD, indicating that only a limited fraction of the noise was removed. Conversely, the remaining models exhibit smaller deviations from the inserted-noise PSD, suggesting that the removed component is more spectrally similar to the known noise process. However, SwinIR and SCUNet show a stronger low-frequency component in the removed PSD, which may be associated with stronger smoothing and partial removal of low-frequency image structures. DRUNet, Restormer, and GSDRUNet show a more moderate low-frequency deviation, suggesting a different trade-off between noise removal and detail preservation.

This behavior may be related to the spatially structured operations used by denoising networks, including convolutional filtering, attention mechanisms, and reconstruction constraints, which can transform initially white or weakly correlated noise into structured removed components. Consequently, the removed component $\hat{n}(i,j)$ may exhibit spatially structured patterns rather than purely random noise. In the frequency domain, this behavior is reflected by deviations of the removed-component PSD from the reference inserted-noise PSD. Such deviations may indicate underfiltering, overfiltering, or removal of anatomical structures whose spectral content overlaps with that of the inserted noise.

Table III shows the MSE for the low-, mid-, and high-frequency bands of the radial PSD of the denoised images. In the low-frequency band, Restormer exhibited the greatest similarity to the ground-truth PSD, followed by DSCcP and DnCNN. Interestingly, the noisy image exhibited a lower low-frequency PSD error than several denoising methods. This does not indicate superior overall image quality, since the inserted noise contributes predominantly to higher spatial frequencies. DnCNN and DSCcP introduce relatively small mod-

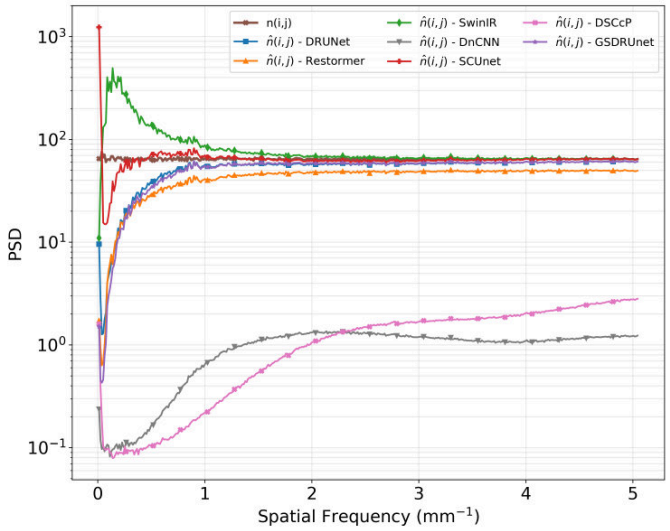


Fig. 3: Power spectral density (PSD) of the inserted noise $n(i,j)$ and of the removed component $\hat{n}(i,j)$ for each denoising method.

ifications in the low-frequency content and therefore remain close to the ground-truth PSD in this band, despite leaving substantial visible noise. Conversely, SwinIR, SCUNet, and DRUNet produce larger low-frequency modifications, resulting in higher errors. Since anatomical signal dominates at low spatial frequencies, these deviations may reflect partial alteration of image structures rather than noise removal alone. Therefore, low-frequency PSD similarity should not be interpreted in isolation as a measure of restoration quality. For the mid-frequency band, Restormer achieved the lowest error, followed by DRUNet, GSDRUNet, and SCUNet. For the high-frequency band, DRUNet achieved the lowest error, followed by SCUNet, GSDRUNet, and SwinIR.

TABLE III: Mean squared error (MSE) computed over the low- (indicated as LF), mid- (indicated as MF), and high-frequency (indicated as HF) ranges of the radial mean PSD of the denoised images. Lower values indicate better agreement with the ground-truth image PSD.

Denoiser	MSE - LF	MSE - MF	MSE - HF
DRUNet	4.20×10^6	60.13	**0.568**
Restormer	$\mathbf{4.43 \times 10^5}$	**26.33**	1.146
SwinIR	1.29×10^7	101.62	1.000
SCUNet	1.92×10^9	73.07	0.658
GSDRUNet	7.14×10^6	65.70	0.676
DSCcP	5.20×10^5	3069.17	1997.59
DnCNN	7.52×10^5	2571.74	2345.50
Noisy	1.57×10^6	4148.38	4121.92

Table IV reports the MSE between the PSD of the inserted noise and the PSD of the component removed by each denoiser. Lower values indicate that the removed component is spectrally closer to the known inserted noise. DRUNet achieved the lowest error in the low-frequency band, whereas SCUNet achieved the lowest errors in the mid- and high-frequency bands. However, Fig. 3 shows that SCUNet and SwinIR remove relatively strong low-frequency components, which is consistent with the smoother appearance observed in Fig. 1. GSDRUNet and DRUNet showed intermediate behavior, suggesting a more balanced trade-off between noise removal and structure preservation. DSCcP and DnCNN showed the highest errors in the mid- and high-frequency bands, consistent with visible residual noise and weaker apparent noise suppression.

TABLE IV: Mean squared error (MSE) computed over the low- (indicated as LF), mid- (indicated as MF), and high-frequency (indicated as HF) ranges between the PSD of the inserted noise and the PSD of the component removed by each denoiser. Lower values indicate greater spectral similarity between the removed component and the known inserted noise.

Denoiser	MSE - LF	MSE - MF	MSE - HF
DRUNet	**1250.99**	59.58	20.12
Restormer	1652.49	375.57	226.14
SwinIR	19574.67	108.90	1.95
SCUNet	18153.65	**3.57**	**1.65**
GSDRUNet	1412.34	53.72	16.66
DSCcP	4107.69	4079.09	3784.47
DnCNN	4095.06	4016.76	3953.37

C. Feature Maps

1) DRUNet: Fig. 4 illustrates representative feature maps extracted from DRUNet, which contains 61 convolutional layers. In the initial layers (0–20), the feature maps closely resemble the input, preserving low-level features such as intensity variations, edges, and anatomical boundaries while retaining visible noise. In the intermediate layers (24–44), the representations become more abstract, emphasizing anatomical structures and spatial patterns while suppressing stochastic noise. In the deeper layers (48–60), the network progressively reconstructs the image, producing smoother feature maps with reduced noise while preserving major anatomical structures and microcalcifications. Overall, DRUNet performs denoising through hierarchical representation learning, progressively extracting low-level features, enhancing structural information, and reconstructing the image while balancing noise suppression and preservation of clinically relevant details. This behavior is consistent with the PSD analysis.

2) Restormer: Fig. 5 illustrates representative feature maps extracted from Restormer, which contains 274 convolutional layers. In the initial layers (0–54), the feature maps preserve low-level features such as intensity variations, edges, and anatomical structures. In the intermediate and deep layers (72–218), the representations become more abstract, emphasizing tissue patterns, microcalcifications, and structural information while suppressing stochastic noise. Unlike DRUNet, Restormer maintains abstract and texture-like representations throughout its deeper layers. Even in the final stages (236–254), anatomical structures remain embedded within textured activations, while the last feature map (273) still exhibits noise-like patterns. Overall, these feature maps suggest that Restormer combines local feature extraction with broader contextual modeling, maintaining rich intermediate

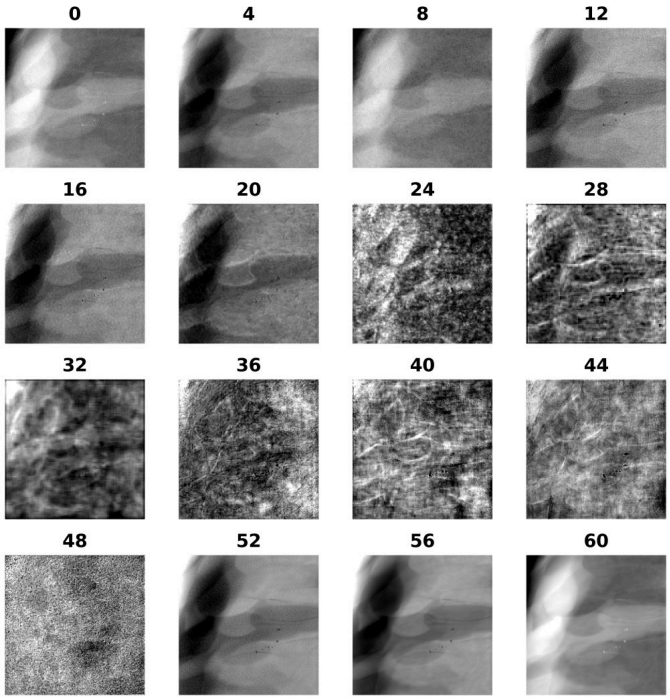


Fig. 4: DRUNet feature maps along its convolutional layers.

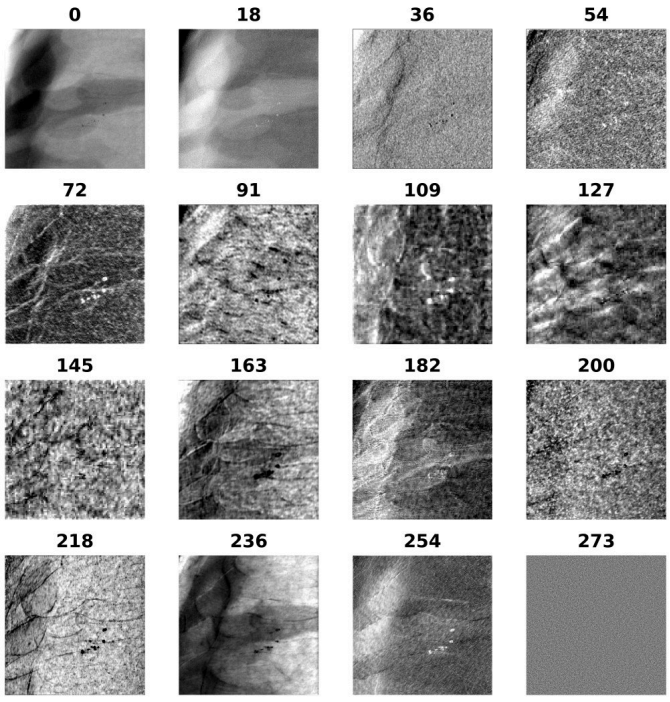


Fig. 5: Restormer feature maps along its convolutional layers.

representations that may contribute to the preservation of fine anatomical details and textures.

3) SCUNet: SCUNet contains 117 convolutional layers, and the representative feature maps in Fig. 6 show a progressive transformation of the mammographic image. In the initial layers (0–23), the feature maps closely resemble the input, preserving anatomical structures, intensity distributions,

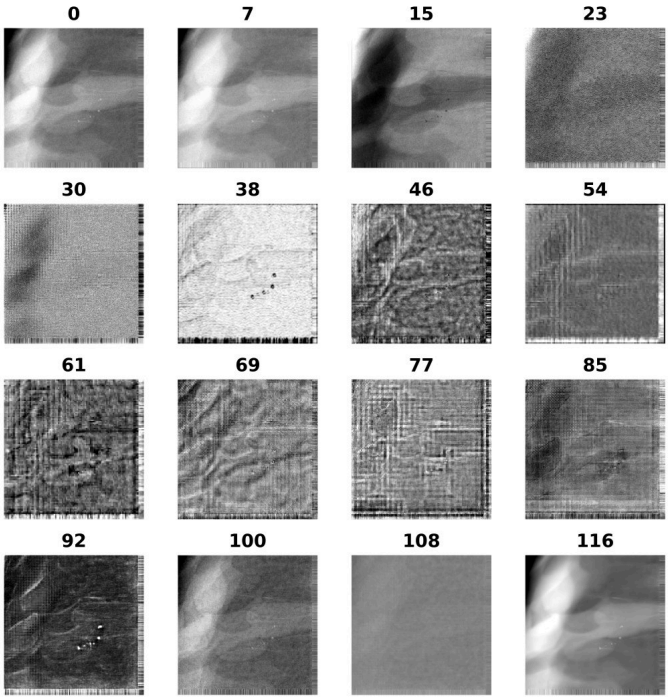


Fig. 6: SCUNet feature maps along its convolutional layers.

and visible noise, indicating low-level feature extraction. In the intermediate layers (30–85), the representations become more abstract, with texture-like patterns, enhanced edges, and high-frequency responses, while emphasizing tissue boundaries and microcalcifications. In the final layers (100–116), the feature maps gradually recover image-like representations while suppressing the previously observed abstract patterns. The final layer (116) reconstructs a smooth image while preserving major anatomical structures and microcalcifications. Overall, SCUNet follows an encoder–decoder-like strategy, progressively transforming low-level features into abstract representations and reconstructing them into a denoised output.

4) SwinIR: SwinIR contains only 9 convolutional layers, with most of its representational capacity concentrated in transformer-based self-attention blocks. The feature maps in Fig. 7 show a markedly different behavior from conventional CNN denoisers. Only layers 0 and 7 retain an image-like appearance, preserving anatomical structures such as tissue boundaries and microcalcifications, whereas the remaining maps (1–6 and 8) exhibit abstract texture patterns, edge responses, and high-frequency activations. This suggests that denoising is primarily performed through contextual representations learned by the attention modules rather than progressive convolutional filtering. The image-like representation at layer 7 indicates the reconstruction stage, where transformer features are projected back into the image domain. Overall, SwinIR relies predominantly on attention-based latent representations, with its convolutional layers mainly supporting feature embedding and image reconstruction.

5) DnCNN: DnCNN contains 20 convolutional layers, and its feature maps reveal a representation strategy distinct from encoder–decoder and transformer-based denoisers. Rather than

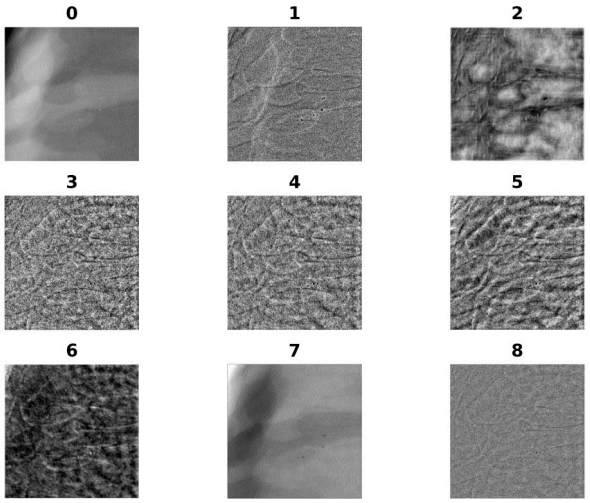


Fig. 7: SwinIR feature maps along its convolutional layers.

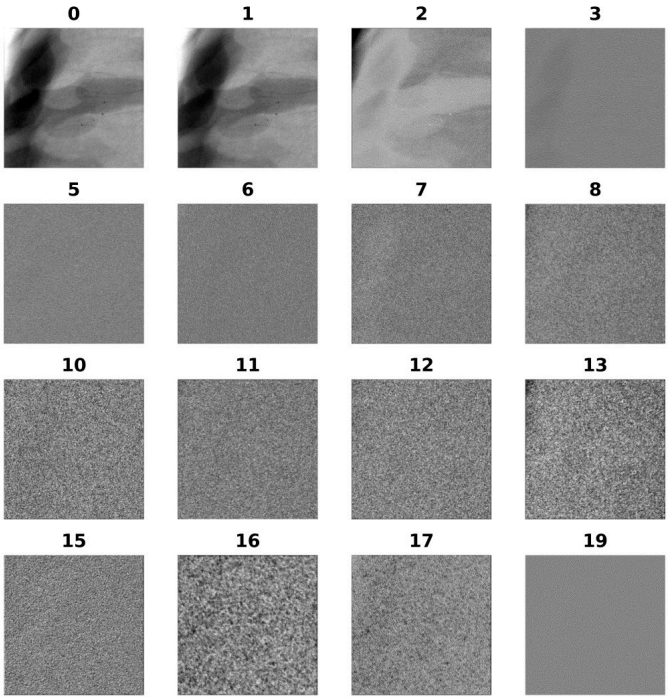


Fig. 8: DnCNN feature maps along its convolutional layers.

progressively encoding and reconstructing the image, DnCNN is explicitly trained to estimate the noise residual. As shown in Fig. 8, the first layers (0–2) preserve image-like representations containing anatomical structures and microcalcifications, while the subsequent layers rapidly suppress image content and become dominated by noise-related patterns. In the intermediate layers (3–17), the feature maps mainly exhibit stochastic textures with little anatomical information, indicating that the network focuses on modeling noise characteristics. In the final layers (17–19), the residual noise is estimated and subsequently subtracted from the noisy input to produce the denoised image. Overall, the feature maps are consistent with DnCNN's residual-learning strategy, in which denoising is performed through noise estimation rather than direct image reconstruction.

6) DSCcP: Fig. 9 shows the feature maps extracted from DSCcP, which contains 20 convolutional layers. Unlike encoder–decoder architectures, DSCcP preserves the same spatial resolution throughout its layers. The early and intermediate layers (0–13) show low-level representations of the noisy input image, including contrast-inverted patterns and texture-like responses. Layer 2 appears to emphasize dense breast structures and the microcalcification cluster. In the final convolutional layers (15–19), the network shows both nearly uniform responses and low-level noise-like textures. Overall, these feature maps suggest that DSCcP performs denoising mainly through low-level texture representations rather than through a hierarchical reconstruction process.

7) GSDRUNet: Fig. 10 illustrates representative feature maps extracted from GSDRUNet, which contains a total of 33 convolutional layers. In the initial layers (0–6), the feature maps correspond to noisy low-level representations of the input image. In the intermediate layers (8–29), the

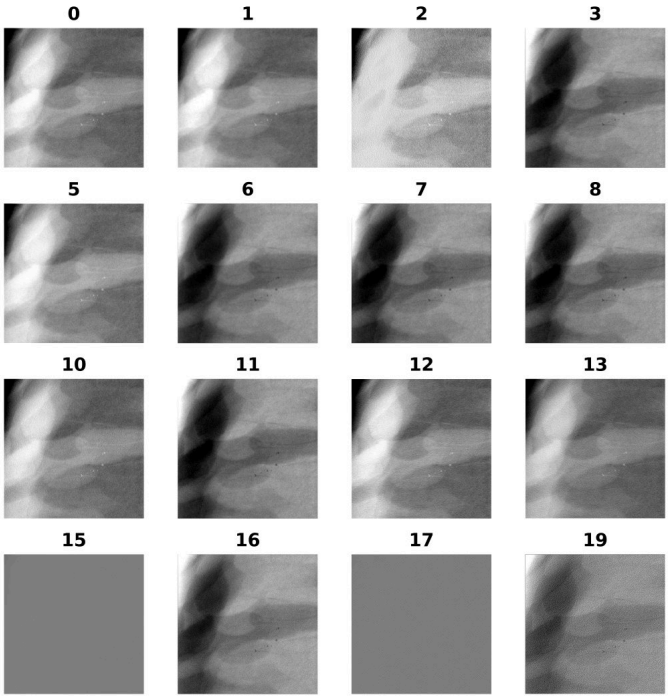


Fig. 9: DSCcP feature maps along its convolutional layers.

representations become increasingly abstract and are dominated by texture-like and noise-related patterns. In the final layer (32), an image-like denoised representation is recovered. Overall, the feature maps suggest that GSDRUNet follows a hierarchical representation strategy similar to DRUNet, progressively transforming low-level image features into abstract representations before reconstructing the denoised output.

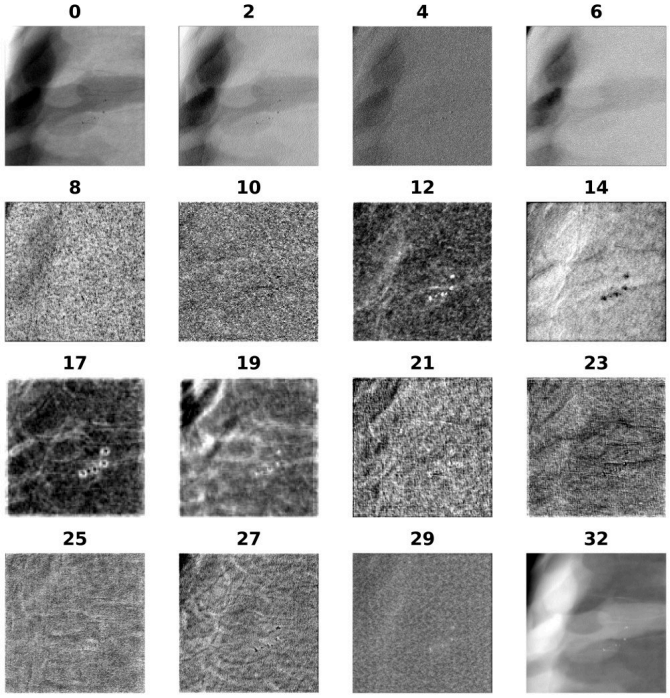


Fig. 10: GSDRUNet feature maps along its convolutional layers.

IV. Conclusion

This work investigated how pretrained deep learning denoisers affect the spectral properties of denoised images and removed components in digital mammography through PSD and feature map analyses. The results showed that some models removed more low-frequency content, whereas others provided weaker suppression in these frequency ranges, highlighting a trade-off between noise removal and mammographic detail preservation. Feature map analysis further suggested that different architectures adopt distinct denoising strategies: encoder–decoder networks progressively transform and reconstruct image representations, transformer-based models rely on attention-driven latent texture representations, and residual-learning networks primarily model noise-like components. These findings indicate that denoising behavior is closely related to the internal representations learned by each architecture and that PSD and feature map analyses provide complementary tools for understanding how deep learning denoisers process mammographic images.

V. Limitations and Future Work

The limitations of this study are outlined as follows. The experiments were conducted using simulated images, which may not fully represent the anatomical variability, acquisition conditions, and noise characteristics of real clinical mammograms. Furthermore, the evaluated denoisers were used in their pretrained, off-the-shelf configurations and were not specifically trained for mammography, while the DeepInverse implementations may differ from the original models. Although PSD analysis provides valuable insights into the frequency-domain behavior, it does not directly assess diagnostic performance or distinguish noise removal from the inadvertent removal of anatomical structures. Finally, the feature map analysis was qualitative and limited to the highest-energy feature map from each sampled layer. Future work should include evaluations on clinical datasets, task-based observer studies, and quantitative explainability methods.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] P. R. Eby, S. Ghate, and R. Hooley, "The benefits of early detection: evidence from modern international mammography service screening programs," *Journal of Breast Imaging*, vol. 4, no. 4, pp. 346–356, 2022.

[3] W. H. Organization *et al.*, "International basic safety standards for protection against ionizing radiation and for the safety of radiation sources," 1996.

[4] L. R. Borges, P. R. Bakic, A. Foi, A. D. Maidment, and M. A. Vieira, "Pipeline for effective denoising of digital mammography and digital breast tomosynthesis," in *Medical Imaging 2017: Physics of Medical Imaging*, vol. 10132. SPIE, 2017, pp. 36–46.

[5] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6360–6376, 2021.

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[7] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[9] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, D.-P. Fan, R. Timofte, and L. V. Gool, "Practical blind image denoising via swin-conv-unet and data synthesis," *Machine Intelligence Research*, vol. 20, no. 6, pp. 822–836, 2023.

[10] H. T. V. Le, A. Repetti, and N. Pustelnik, "Unfolded proximal neural networks for robust image gaussian denoising," *IEEE Transactions on Image Processing*, 2024.

[11] S. Hurault, A. Leclaire, and N. Papadakis, "Gradient step denoiser for convergent plug-and-play," *arXiv preprint arXiv:2110.03220*, 2021.

[12] D. A. Pereira, R. F. Brandão, L. R. Borges, B. Barufaldi, A. D. A. Maidment, and M. A. C. Vieira, "Benchmarking pretrained machine learning denoisers for low-dose digital mammography: The impact of physics-guided model-based integration," in *Proceedings of the International Workshop on Breast Imaging (IWBI)*, 2026, in press.

[13] J. Boita, R. E. van Engen, A. Mackenzie, A. Tingberg, H. Bosmans, A. Bolejko, S. Zackrisson, M. G. Wallis, D. M. Ikeda, C. Van Ongeval *et al.*, "How does image quality affect radiologists' perceived ability for image interpretation and lesion detection in digital mammography?" *European radiology*, vol. 31, no. 7, pp. 5335–5343, 2021.

[14] B. Barufaldi, D. Higginbotham, P. R. Bakic, and A. D. A. Maidment, "OpenVCT: a GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis," in *Medical Imaging 2018: Physics of Medical Imaging*, J. Y. Lo, T. G. Schmidt, and G.-H. Chen, Eds., vol. 10573, International Society for Optics and Photonics. SPIE, 2018, p. 1057358.

[15] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *IEEE transactions on image processing*, vol. 17, no. 10, pp. 1737–1754, 2008.

[16] J. Tachella, M. Terris, S. Hurault, A. Wang, and e. a. Davy, "Deepinverse: A python package for solving imaging inverse problems with deep learning," *Journal of Open Source Software*, vol. 10, no. 115, p. 8923, 2025. [Online]. Available: https://doi.org/10.21105/joss.08923

Stability of Microstates and Machine Learning in Resting-State Electroencephalography

1st Marcos José Alves de Barros Monteiro
Graduate Program in Teleinformatics Engineering (PPGETI)
Federal University of Ceará (UFC)
Fortaleza/CE, Brazil
marcosmonteiro.ufc@gmail.com

2nd Arthur Guilherme da Frota Brito Machado
Graduate Program in Teleinformatics Engineering (PPGETI)
Federal University of Ceará (UFC)
Fortaleza/CE, Brazil
arthur.frota@dataqore.com.br

3rd Caio dos Santos Nascimento
Department of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza/CE, Brazil
caio.santos@alu.ufc.br

4th Victor Hugo C. de Albuquerque
Department of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza/CE, Brazil
victor.albuquerque@lesc.ufc.br

***Abstract*—Microstate analysis has emerged as a useful framework for characterizing the spatiotemporal structure of resting-state electroencephalography, yet the computational choices involved in microstate extraction remain a major source of methodological variability. This work presents a systematic evaluation of how preprocessing decisions, clustering algorithms, and model-order selection influence both the stability of microstate representations and the performance of machine learning models. Using a reproducible pipeline applied to a large clinically annotated EEG dataset, we compare multiple clustering approaches and vary the number of microstates to quantify their effects on cluster validity, temporal descriptors, and classification accuracy. The results show that moderate model orders produce more coherent and reliable microstate features, whereas excessively low or high values lead to fragmentation and reduced discriminability. Among clustering methods, hierarchical and centroid-based algorithms yield the most stable topographic maps, which translate into improved performance for ensemble and kernel-based classifiers. Overall, the study demonstrates that methodological rigor in microstate extraction is essential for achieving robust EEG classification and provides practical guidelines for designing reproducible pipelines in computational neurodiagnostics.**

***Index Terms*—Electroencephalography (EEG), microstates, clustering algorithms, machine learning, neurological disorder detection**

I. Introduction

Neurological disorders affect nearly 3 billion people ($\sim$1 in 3) and have driven an increase of $\approx 18\%$ in combined disability and premature mortality in recent assessments [1]. According to the World Health Organization (WHO), nearly 70% of cases occur in low- and middle-income countries and account for around nine million deaths each year; the leading contributors to disability-adjusted life years (DALYs) are stroke (42.2%), migraine (16.3%), dementia (10.4%), meningitis (7.9%), and epilepsy (5.0%) [2]. In the Region of the Americas, 2019 estimates indicate on the order of 533,000 deaths (about 60% among women) and marked between-country heterogeneity in mortality and disability, Brazil ranks highest in years lived with disability (YLDs), while age-standardized mortality ranges from 47.4 per 100,000 (United States) down to 6.6 per 100,000 (Venezuela) [3].

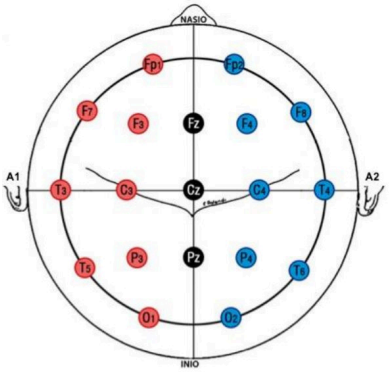


Fig. 1. Illustration of a standard EEG cap with electrodes placed according to the 10–20 international system. The montage highlights the spatial distribution used for recording scalp electrical activity.

A. Electroencephalogram and Microstates

Electroencephalography (EEG) is a noninvasive technique that measures voltage fluctuations on the scalp arising from synchronized postsynaptic activity of cortical pyramidal neurons (Fig. 1). Its millisecond-scale temporal resolution enables direct observation of fast neural dynamics that are not captured by structural or hemodynamic imaging modalities, making EEG particularly useful for studying fast neural processes and large-scale functional organization. Despite its advantages, routine EEG interpretation is often hindered by low signal-to-noise ratios, nonstationarity, inter-subject variability, and contamination from ocular, muscular, and environmental sources, which complicate the reliable extraction of physiologically meaningful features [2].

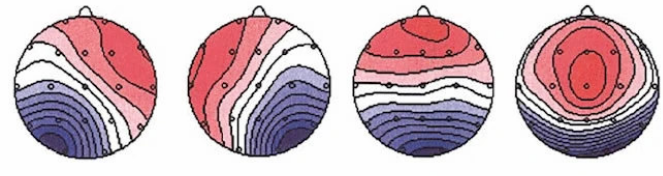


Fig. 2. Canonical EEG microstate maps (A–D).

EEG microstate analysis provides a compact representation of these large-scale neural dynamics by segmenting the continuous signal into brief, quasi-stable topographies that recur over time. These microstates, typically lasting 50–120 ms, are frequently interpreted as elementary building blocks of spontaneous cognitive processing, and their temporal characteristics (duration, occurrence rate, fractional occupancy, and transition structure) are known to exhibit characteristic differences across a range of physiological and pathological conditions [4]. However, microstate-based inference remains highly sensitive to methodological decisions involving preprocessing pipelines, referencing approaches, artifact handling, map normalization, clustering algorithms, and, critically, the choice of the number of microstates. Although the canonical four-class solution (A-D, Fig. 2) is widely used, empirical evidence consistently demonstrates that the optimal model order depends on cohort characteristics, recording conditions, and preprocessing strategies [5]. Common selection criteria, including global explained variance, cross-validation, information-theoretic measures, and cluster-validity indices, often yield divergent estimates, highlighting the need for principled, data-driven approaches rather than fixed heuristic conventions [6], [7].

B. Machine Learning Techniques for EEG Microstates

Machine learning has become increasingly central to EEG analysis and has gained particular relevance in microstate research, where models must capture complex spatiotemporal patterns, offering tools capable of modeling nonlinear relationships, extracting high-dimensional structure, and improving diagnostic accuracy beyond handcrafted descriptors. Supervised approaches, including artificial neural networks, convolutional neural networks, support vector machines, and ensemble-based classifiers, have demonstrated strong performance in differentiating healthy from pathological EEG patterns, detecting epileptiform discharges, and identifying candidate biomarkers for neurological and psychiatric disorders [8], [9]. Their ability to learn discriminative representations directly from raw or minimally processed signals makes them particularly attractive in settings where subtle spatiotemporal signatures carry diagnostic relevance.

Unsupervised methods provide complementary advantages by uncovering organization within EEG activity without requiring labeled data, a property that aligns naturally with the principles of microstate analysis. Algorithms such as k-means, DBSCAN, and graph-based clustering have been used to infer latent state structure, segment recurring topographies, and characterize cohort heterogeneity [10], [11]. Yet, these methods remain sensitive to preprocessing choices, noise levels, and, critically, assumptions about the number of clusters or density structure, limitations shared with traditional microstate pipelines. These considerations mean that the value of machine learning in microstate research is determined as much by the surrounding methodology as by the algorithms themselves. Careful choices in preprocessing, feature construction, and model-order selection are often what separate stable, interpretable findings from results that fail to generalize. When these elements are handled consistently and transparently, machine learning and microstate analysis complement each other well, offering a flexible framework for probing large-scale neural dynamics. Without such methodological discipline, however, the apparent power of these tools can be misleading, underscoring the importance of reproducible and well-justified analytical pipelines.

C. Motivation

Although EEG microstates and modern machine learning techniques offer promising avenues for characterizing large-scale brain dynamics, their practical use remains limited by methodological instability and lack of consensus. A key unresolved issue is the determination of the optimal number of microstates: the choice of model order substantially alters temporal metrics, transition structure, and classification performance, yet commonly adopted heuristics often yield inconsistent or contradictory results across datasets, cohorts, and preprocessing pipelines [12]. This variability hampers interpretability, reduces reproducibility, and complicates clinical translation, especially in heterogeneous populations where neural patterns may deviate from canonical assumptions.

At the same time, machine learning classifiers and clustering algorithms many of the same sensitivities. Their outputs depend strongly on data normalization, artifact removal, feature scaling, and initialization strategies, as well as on the quality of the microstate representations they consume [13]. Without systematic evaluation, it becomes unclear whether observed improvements stem from genuine neurophysiological differences or from arbitrary modeling choices. These challenges highlight a critical gap in the literature: the need for an integrated, data-driven framework that (i) rigorously estimates the optimal number of microstates, (ii) evaluates how this choice propagates through clustering and classification pipelines, and (iii) establishes methodological guidelines to support reproducible EEG microstate research.

D. Contributions

This work addresses the methodological instability surrounding EEG microstate extraction and its impact on machine

learning analysis. The main contributions are summarized as follows:

- We develop a reproducible preprocessing and segmentation pipeline for EEG microstate extraction, consolidating best practices in filtering, referencing, artifact handling, and map normalization.
- We conduct a systematic, data-driven evaluation of model-order selection, comparing multiple criteria, including cluster-validity indices, cross-validation metrics, and information-theoretic measures to estimate the optimal number of microstates.
- We analyze how different microstate model orders propagate through unsupervised and supervised machine learning algorithms, assessing robustness, sensitivity, and overall classification performance.
- We provide a comparative study of multiple clustering methods (k-means, modified variants, AAHC, k-medoids, DBSCAN) in the context of EEG microstate segmentation, highlighting strengths, limitations, and convergence patterns.
- We establish practical recommendations and methodological guidelines to support reproducible research in microstate modeling and its application to neurological disorder assessment.

II. Material and Methods

A. Dataset

This study uses the Temple University Abnormal EEG Corpus (TUAB), a publicly available subset of the Temple University EEG Corpus (TUEG). TUAB contains clinically annotated scalp EEG recordings labeled as normal or abnormal, selected to provide a demographically balanced dataset while minimizing confounding factors. All recordings are distributed in EDF format and were originally acquired at the Temple University Hospital (TUH) Department of Neurology. The dataset follows HIPAA de-identification guidelines, with all personal identifiers removed prior to release.

TUAB comprises 2,993 EEG sessions from 2,329 unique subjects (52.1% female), with a mean age of 48.6 $\pm$ 17.9 years. The dataset is nearly balanced, containing 50.8% normal and 49.2% abnormal recordings. The official split includes 2,717 recordings for training and 276 for evaluation and is detailed by class in Table I, which reports non-pathological and pathological counts for each subset. Each session typically includes 21 EEG channels, most commonly sampled at 250 Hz. Recordings shorter than 15 minutes were excluded during dataset construction, and abnormal sessions were validated through concordance between automated detectors and clinical reports, achieving approximately 97–100% agreement between automated labels and clinical reports.

All recordings used in this work were obtained from the official TUH distribution portal and processed exclusively for research purposes. Although TUAB is already fully de-identified, all procedures adhere to the ethical guidelines established by the original data providers. Because the dataset is publicly released for research under a Temple University institutional review board (IRB) protocol, no additional ethical approval was required for secondary analysis.

TABLE I
Number of recordings and patients in the TUAB dataset used in previous machine learning studies

Set	Non-Pathological Recordings	Pathological Recordings	Total Recordings
Training	1371	1346	2717
Evaluation	150	126	276
Total	**1521**	**1472**	**2993**

Adapted from [14].

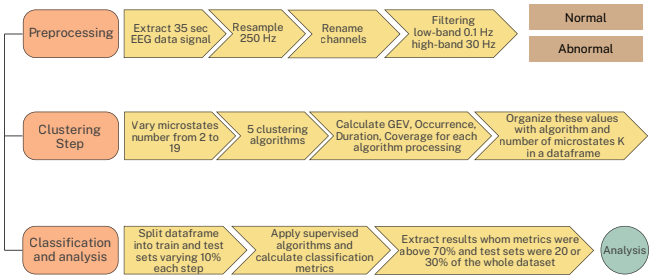


Fig. 3. Experimental workflow.

B. Workflow Overview

The experimental pipeline consists of four sequential stages (Fig. 3). First, EEG recordings from the TUAB dataset were preprocessed to standardize montage, sampling rate, and signal quality.

Second, microstate segmentation was performed using multiple clustering algorithms over a range of model orders, producing a structured set of microstate descriptors.

Third, clustering solutions were compared using standard validity indices.

Finally, the resulting microstate-feature tables were used to train several supervised learning models under 5-fold cross-validation.

Together, these stages form a unified pipeline designed to evaluate how different preprocessing, clustering strategies, and microstate model orders influence classification performance.

C. Preprocessing

All EEG recordings were preprocessed through a reproducible pipeline designed to standardize signal quality across subjects and sessions. For each TUAB recording, a fixed 35-second segment was extracted (from 175 s to 210 s), ensuring uniform temporal length while avoiding periods commonly

affected by recording artifacts or clinical interventions. Signals originally sampled at heterogeneous rates (250, 256, 400, or 512 Hz) were resampled to 250 Hz to ensure consistency across all analyses.

Channel names were then mapped to a unified 10–20 nomenclature using the correspondence defined in Table II, and recordings missing required channels were discarded. A 0.1–30 Hz band-pass filter was applied to retain the conventional EEG frequency bands associated with microstate dynamics while suppressing slow drifts and high-frequency muscular noise. After filtering, all signals were referenced to the average reference montage, a standard choice in microstate research due to its stability and topographic interpretability.

TABLE II
EEG CHANNEL NAME MAPPING USED IN PREPROCESSING

Original Channel Name	Renamed Channel
EEG FP1-REF	FP1
EEG FP2-REF	FP2
EEG F3-REF	F3
EEG F4-REF	F4
EEG C3-REF	C3
EEG C4-REF	C4
EEG P3-REF	P3
EEG P4-REF	P4
EEG O1-REF	O1
EEG O2-REF	O2
EEG F7-REF	F7
EEG F8-REF	F8
EEG T3-REF	T3
EEG T4-REF	T4
EEG T5-REF	T5
EEG T6-REF	T6
EEG A1-REF	A1
EEG A2-REF	A2
EEG FZ-REF	FZ
EEG CZ-REF	CZ
EEG PZ-REF	PZ

Source: the authors.

All preprocessing steps were executed in Python 3.11 using the MNE library, which provides robust and validated implementations of EEG transformation routines. To ensure the reproducibility of every data manipulation procedure, all stochastic operations were controlled using a fixed random seed (42). The resulting preprocessed signals served as the standardized input for the microstate clustering stage.

The preprocessing parameters were selected to remain consistent with common practices reported in the EEG microstate literature. The use of a 0.1–30 Hz band-pass filter preserves the frequency components most frequently associated with resting-state microstate dynamics while reducing slow drifts and high-frequency artifacts. Similarly, average referencing was adopted because it provides stable topographic representations and facilitates comparisons across subjects and recording sessions. The 35-second segment length was chosen to ensure a standardized analysis window for all recordings while maintaining a sufficient number of microstate transitions for reliable descriptor estimation.

D. Clustering

Microstate extraction was performed by applying multiple clustering algorithms to the preprocessed EEG topographies. For each recording, the Global Field Power (GFP) peaks were first identified and used as candidate topographic maps, following standard practice in microstate analysis. These maps were then submitted to five clustering algorithms: K-means, Modified K-means, K-medoids, Adaptive AAHC, and DBSCAN. This combination includes both centroid-based and density-based approaches, enabling a broad evaluation of methodological variability in microstate segmentation.

To systematically assess model-order effects, the number of microstates was varied from $K = 2$ to $K = 19$ for all algorithms except DBSCAN, whose number of clusters is data-driven. For each clustering solution, a set of microstate descriptors was computed, including global explained variance (GEV), occurrence, duration, coverage. These descriptors quantify spatial alignment, temporal stability, and dynamical structure, forming the feature space used in the subsequent classification stage.

Cluster quality was further assessed using the validity indices Silhouette Score, Calinski–Harabasz Index, Davies–Bouldin Index, and Dunn Index allowing comparison across clustering algorithms and microstate model orders.

E. Classification

Following the clustering stage, an intermediate feature dataset was constructed for supervised learning. For each recording, and for every combination of clustering algorithm and microstate model order, four microstate descriptors were extracted: global explained variance (GEV), mean occurrence, mean duration, and coverage. Each row of the resulting table corresponds to a single recording, and each column corresponds to one of these descriptors, yielding a structured representation of the temporal and dynamical properties of the microstate solution.

This feature table served as the input to the classification stage, in which several machine learning models were trained to distinguish normal from abnormal EEG recordings. The evaluated classifiers included Logistic Regression, Support Vector Machines (SVM), Random Forests, Decision Trees, k-Nearest Neighbors (KNN), Naive Bayes, Gradient Boosting, XGBoost, and the Optimum-Path Forest (OPF) classifier; for OPF, multiple distance functions available in OPF programming package were tested to account for different geometries in feature space.

All models were trained and assessed using 5-fold cross-validation. Performance was quantified using accuracy, precision, recall, and F1-score, with confusion matrices computed to inspect error distribution. Hyperparameters were kept at default settings unless required for numerical stability, ensuring fair comparison across methods.

By training classifiers on microstate descriptors derived under different clustering strategies and model orders, this stage provides a systematic evaluation of how upstream preprocessing and segmentation choices propagate into diagnostic performance.

Although classifier performance was assessed using multiple evaluation metrics under cross-validation, the primary objective of this work was not to identify a single best-performing model. Instead, the classification stage was used as a controlled framework to investigate how methodological choices made during microstate extraction influence downstream predictive performance. Therefore, differences among classifiers should be interpreted mainly as indicators of the quality and stability of the resulting microstate representations.

III. Discussion

The results demonstrate that methodological choices in microstate extraction substantially influence both the structure of the derived microstate representations and the classification of normal versus abnormal EEG recordings. Varying the number of microstates produced consistent changes in temporal descriptors, with lower model orders yielding more stable and interpretable dynamics, whereas higher values of K introduced fragmentation and reduced global explained variance (GEV). This behavior was observed across all clustering algorithms but was more pronounced in centroid-based methods such as K-means and Modified K-means.

The observed differences among clustering algorithms can be partially explained by their underlying assumptions. Hierarchical approaches such as AAHC tend to preserve topographic organization throughout the clustering process, whereas centroid-based methods favor compact cluster representations. In contrast, density-based methods are more sensitive to the distribution of GFP peak topographies, which may lead to greater variability in the resulting microstate structures. These characteristics help explain the differences observed in cluster validity measures and classification performance.

To illustrate these effects quantitatively, Table III summarizes representative cluster validity metrics obtained with the Modified K-Means algorithm across a range of model orders. The selected values show the trade-off between compactness, separation, and stability as K increases, supporting the observed pattern in which low to moderate model orders yield more coherent microstate structure, while larger K values introduce fragmentation and reduced descriptor reliability.

TABLE III
Cluster validity metrics for the Modified K-Means algorithm comparing different K values between 2 and 19

Microstates	Silhouette	Calinski–Harabasz	Davies–Bouldin	Dunn
3	-0.0091	142.32	10.92	0.00276
4	-0.0297	149.95	11.43	0.00120
5	-0.0411	**151.56**	11.28	**0.00132**
14	-0.0789	138.47	**9.96**	0.00025
16	-0.0824	138.32	9.87	0.00026

Source: the authors.

Comparisons among clustering strategies revealed systematic differences in topographic stability and descriptor reliability. AAHC generally produced compact and well-separated clusters, reflected in higher Silhouette and Calinski–Harabasz scores, while DBSCAN showed strong dependence on density parameters and resulted in variable cluster counts. K-medoids offered robust performance in terms of temporal descriptors but exhibited lower GEV for larger values of K. These findings indicate that algorithmic assumptions directly shape the microstate solutions and should be considered when interpreting large-scale neural dynamics.

TABLE IV
Average cluster validity metrics across clustering algorithms comparing different K values between 2 and 19

Algorithm	Silhouette	Calinski–Harabasz	Davies–Bouldin	GEV
K-Means	-0.045	148.2	11.30	0.76
Modified K-Means	-0.038	150.4	10.95	0.79
K-Medoids	-0.022	141.0	12.10	0.72
AAHC	**0.012**	**162.8**	**9.84**	**0.83**
DBSCAN	-0.091	118.5	14.22	0.67

Source: the authors.

The intermediate feature dataset constructed from microstate descriptors enabled a structured evaluation of how these methodological variations propagate into supervised learning. Classification performance showed clear sensitivity to both the clustering method and the selected model order. Moderate values of K produced more discriminative feature spaces, supporting the idea that excessively coarse or overly fragmented microstate partitions reduce the separability of normal and abnormal EEG recordings. Among the classifiers, ensemble-based and kernel-based models generally achieved higher accuracy and F1-scores, whereas distance-based methods were more sensitive to descriptor scaling and class boundaries. Among all OPF configurations tested, the highest performance was obtained using the Euclidean distance, which consistently outperformed Manhattan, Chebyshev, and Minkowski variants in terms of accuracy, recall, and F1-score.

TABLE V
Classification performance across machine learning models (5-fold cross-validation)

Classifier	Accuracy	Precision	Recall	F1
Logistic Regression	0.71	0.70	0.69	0.69
SVM (RBF)	**0.78**	0.77	0.76	**0.76**
Random Forest	0.75	**0.79**	0.71	0.74
KNN (k=5)	0.69	0.67	0.66	0.66
Naive Bayes	0.64	0.62	0.63	0.62
Gradient Boosting	0.76	0.75	0.74	0.74
XGBoost	0.77	0.76	**0.77**	0.76
OPF (euclidean distance)	0.74	0.73	0.73	0.73

Source: the authors.

Taken together, these results reinforce the importance of methodological rigor in EEG microstate research. Small vari-

ations in preprocessing, clustering strategy, and model-order selection propagate into measurable differences in classification performance, underscoring the risks of relying on fixed heuristics or canonical parameter choices. The findings support the need for systematic, data-driven evaluation when designing microstate pipelines, particularly in diagnostic settings in which reproducibility and interpretability are essential.

Some limitations should be considered when interpreting the results. The analysis was conducted using a single public dataset and focused on resting-state EEG recordings acquired under specific recording conditions. Although TUAB provides a large and clinically relevant cohort, performance estimates may vary when different populations, acquisition protocols, or clinical settings are considered. In addition, the present study emphasizes methodological comparisons rather than neurophysiological interpretation of individual microstate classes. Future investigations may explore these aspects in greater detail.

IV. Conclusion

This study examined how methodological choices in EEG microstate extraction, including preprocessing decisions, clustering algorithms, and model-order selection, shape both the structure of microstate representations and the performance of classifiers. The results show that variations in number of microstates K, as well as differences among clustering strategies, produce measurable effects on temporal descriptors, cluster validity, and ultimately on the ability of machine learning models to discriminate normal from abnormal EEG recordings. Moderate values of K yielded the most stable and discriminative feature spaces, while excessively low or high model orders reduced descriptor reliability and degraded classification performance, suggesting that moderate model orders, particularly those close to the traditional four-microstate solution, provide a good balance between stability, interpretability, and classification performance in the evaluated dataset..

Among the evaluated clustering methods, AAHC and Modified K-Means produced the most coherent microstate structures, which translated into superior performance when combined with contemporary ensemble and kernel-based classifiers such as Gradient Boosting, XGBoost, and SVM. These findings highlight the importance of systematic, data-driven parameter exploration in microstate research, since small methodological variations can propagate into substantial differences in diagnostic accuracy.

The study also reinforces the need for reproducible and transparent microstate pipelines. By providing a controlled experimental design and a comprehensive comparison across clustering and classification techniques, this work contributes practical insights for researchers developing microstate-based analytical tools, particularly in clinical contexts where interpretability and robustness are essential.

Beyond classification performance, the findings may also be relevant for the design of EEG-based decision-support systems. More stable microstate representations can contribute to the development of reliable computational tools for screening, prioritization, and complementary assessment of neurological abnormalities, particularly in large-scale clinical environments where reproducibility is essential.

Future work may extend this analysis by exploring larger EEG datasets, testing additional clustering formulations, integrating deep learning models for microstate representation learning, and evaluating adaptive or subject-specific model-order selection procedures. Such directions may further advance the reliability and applicability of microstate-based approaches in computational neurodiagnostics.

References

[1] GBD 2021 Nervous System Disorders Collaborators, "Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," Lancet Neurology, 2024.

[2] World Health Organization (WHO), Intersectoral Global Action Plan on Epilepsy and Other Neurological Disorders 2022–2031 (IGAP), 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240076624. Accessed: Nov. 10, 2025.

[3] Pan American Health Organization (PAHO), The Burden of Neurological Conditions in the Region of the Americas, 2000–2019, 2021. [Online]. Available: https://www.paho.org/en/enlace/burden-neurological-conditions. Accessed: Nov. 10, 2025.

[4] K. Fatima, A. M. Mehendale, and H. Reddy, "Young-Onset Dementia and Neurodegenerative Disorders of the Young With an Emphasis on Clinical Manifestations," Cureus, vol. 14, no. 10, pp. e30025, 2022. doi: 10.7759/cureus.30025.

[5] C. Sun and C. Mou, "Survey on the research direction of EEG-based signal processing," Frontiers in Neuroscience, vol. 17, 2023.

[6] W. Hu, Z. Zhang, L. Zhang, G. Huang, L. Li, and Z. Liang, "Microstate detection in naturalistic electroencephalography data: A systematic comparison of topographical clustering strategies on an emotional database," *Frontiers in Neuroscience*, vol. 16, Art. no. 812624, Feb. 2022, doi: 10.3389/fnins.2022.812624.

[7] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. 4, Art. no. 1457619, Sep. 2024, doi: 10.3389/fbinf.2024.1457619.

[8] Q. Ma, M. Wang, L. Hu, L. Zhang, and Z. Hua, "A Novel Recurrent Neural Network to Classify EEG Signals for Customers' Decision-Making Behavior Prediction in Brand Extension Scenario," Frontiers in Human Neuroscience, vol. 15, 2021.

[9] W. Chen et al., "An automated detection of epileptic seizures EEG using CNN classifier based on feature fusion with high accuracy," BMC Medical Informatics and Decision Making, vol. 23, no. 1, p. 96, 2023.

[10] T. Y. Wen and S. A. Mohd Aris, "Hybrid Approach of EEG Stress Level Classification Using K-Means Clustering and Support Vector Machine," IEEE Access, vol. 10, pp. 18370–18379, 2022.

[11] Y. Du, G. Li, M. Wu, and F. Chen, "Unsupervised Multivariate Feature-Based Adaptive Clustering Analysis of Epileptic EEG Signals," Brain Sciences, vol. 14, no. 4, p. 342, 2024.

[12] H. Ren, X. Ran, M. Qiu, S. Lv, J. Wang, C. Wang, Y. Xu, Z. Gao, W. Ren, X. Zhou, J. Mu, Y. Yu, and Z. Zhao, "Abnormal nonlinear features of EEG microstate sequence in obsessive-compulsive disorder," *BMC Psychiatry*, vol. 24, no. 1, Art. no. 881, Dec. 2024, doi: 10.1186/s12888-024-06334-6.

[13] M. J. Antony et al., "Classification of EEG Using Adaptive SVM Classifier with CSP and Online Recursive Independent Component Analysis," Sensors, vol. 22, no. 19, 2022.

[14] A.-K. Kiessner, R. T. Schirrmeister, L. A. W. Gemein, J. Boedecker, and T. Ball, "An extended clinical EEG dataset with 15,300 automatically labelled recordings for pathology decoding," *NeuroImage: Clinical*, vol. 39, p. 103482, 2023. doi: 10.1016/j.nicl.2023.103482.

Study of Interpretable Deep Neural Networks in Brain-Computer Interfaces

1st Helena Guachalla de Andrade
School of Electrical and Computer Eng.
Universidade Estadual de Campinas
Campinas, Brazil
https://orcid.org/0009-0006-9861-4693

2nd Natalia Gil Canto
School of Electrical and Computer Eng.
Universidade Estadual de Campinas
Campinas, Brazil
https://orcid.org/0000-0002-1451-7452

3rd Levy Boccato
School of Electrical and Computer Eng.
Universidade Estadual de Campinas
Campinas, Brazil
https://orcid.org/0000-0001-9319-9829

***Abstract*—The application of deep learning models in brain-computer interfaces (BCIs) has increased in the last years, with many convolutional neural networks (CNNs) being developed to exploit both the temporal and spatial information contained in the electroencephalograms (EEGs). In this scenario, it is important to find models that not only have good classification performance, but also offer a certain level of interpretability for their results. The EEGNet model and its variants present such a feature, conciliating strong performance with an architectural design that allows the inspection of electrode importance and the analysis of the learned kernels in terms of frequency filters. Interestingly, another family of models emerged after the proposal of SincNet, which adopts sinc-based convolution kernels to learn a filterbank of bandpass filters with a reduced number of parameters. Thus, this work brings together seven models from these families with interpretable elements and establishes a comparison in the context of BCI systems based on SSVEP and motor imagery. Two databases were explored, and the average accuracy across subjects was evaluated. In addition, given the interpretable profile of the models, we also analyzed the learned filters (both temporal and spatial).**

***Index Terms*—Brain-computer interface, electroencephalography, deep learning, spatio-temporal processing, frequency filtering**

I. Introduction

The advancement of artificial intelligence technologies has driven the development of methods for analyzing complex biological signals, such as electroencephalograms (EEGs), which are widely used in medical and experimental applications, e.g. epilepsy diagnosis, sleep studies, and brain-computer interfaces (BCIs). In this context, deep learning techniques [1] have been widely explored [2]–[4] and present notable results in different scenarios: detection of seizures [5], motor rehabilitation [6], and mental workload monitoring [7]. Convolutional neural networks (CNNs) are predominant in this field, and proved to be suitable techniques for processing brain signals [8].

BCIs based on EEGs aim at establishing an alternative path to communicate human and machine, which can have a key role in the development of new assistive technologies, especially for people with disabilities. Deep learning methods have been explored in BCI systems based on different paradigms, such as motor imagery (MI) [9]–[12] and steady-state visually evoked potentials (SSVEP) [3], [13], [14], showing promising results. Interestingly, there has been an increasing interest not only in performance, but also in discussing how, and possibly why, the models carry out their decision-making process.

In this context, the implementation of interpretable models for decoding EEGs within BCI systems is fundamental to building trustworthy and helpful technology. In light of this, there are works in the BCI literature that incorporate interpretable components into deep models. In particular, EEGNet [15] can be considered a well-established reference, whose architecture explores different convolutional operations, e.g., depthwise and separable convolutions, to create temporal and spatial filtering layers.

Besides the performance, the authors highlight how the learned filters were connected to known neurophysiological phenomena, based on the analysis of the output spatial filters and the weights of the convolutional kernels.

Since then, a family of CNN models has emerged, taking inspiration from EEGNet, among which we mention EEG-TCNet [16], EEGITNet [17] and EEGNex [18]. All of these variations contain similar layers that allow some degree of interpretation, both in terms of frequency filtering and electrode (spatial) relevance.

Another model that was conceived with a strong appeal to interpretable layers was SincNet [19]. Originally applied to audio processing, this model employs a special first convolution layer whose kernels are designed according to a $sinc(\cdot)$ function, in order to create a set of band-pass filters parameterized by the cutoff frequencies, which are the only learnable elements of the layer.

As a consequence, the number of learnable parameters is reduced, and the role played by the learned kernels can be easily evaluated in the frequency domain. In fact, the prior knowledge imposed by the filters shape implies a constructive bias that gives enough flexibility to the filterbank to adapt to the data. Thus, the generated filter maps tend to be more intelligible and reflect the importance given to specific frequency bands.

This study was financed, in part, by the São Paulo Research Foundation (FAPESP, Process Number 2024/07783-5), as part of the BI0S - Brazilian Institute of Data Science (2020/09838-0).

This study was financed in part by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001.

In [10], the sinc-convolutions proposed in SincNet were explored in the context of EEGs and BCI systems, leading to the SincShallowNet model, which also rely upon depthwise convolutions to extract temporal and spatial features of the data.

The interpretability brought by SincNet motivated the development of more advanced designs for the band-pass filters, which form a second family of models, including SincSquareNet [20], PF-Net [21] and IIRI-Net [22]. All these variants adopt a specially-tailored first convolution layer that creates multiple frequency-band filters with simplified parameterizations, but present an architectural limitation: they can process univariate (single channel) temporal signals.

Having this scenario in view, in this work we propose to systematically study and compare several models from EEGNet [15] and SincNet [19] families in the context of BCI systems, evaluating not only the classification performance, but also the attained filters (temporal and/or spatial), highlighting the brain regions that were most activated based on the learned weights. Here, the concept of interpretability is understood in the restricted sense of inspectability: both the EEGNet and the SincNet families have initial layers that can act as frequency filters, which can be inspected in the frequency domain and may be connected with the nature of the data and of the task, as well as spatial filtering layers, which may reveal the electrode relevance that was learned by these models.

In order to explore more recent variations of SincNet [19] for EEG processing, we extend the PF-Net [21] and IIRI-Net [22] models to the multivariate scenario, making an adaptation similar to that employed in SincShallowNet [10]. To the best of our knowledge, these variations have not been applied to BCI-EEG systems yet.

In the end, a total of seven models were tested, considering both SSVEP and MI paradigms of BCIs, allowing a more comprehensive perception of the potential advantages of each design.

The sequence of the paper is organized as follows: Section II reviews the relevant literature related to EEGNet [15] and SincNet [19] families. Section III details the methodology, the datasets, and the experimental setup used in this study. Section IV presents the findings, followed by a discussion of their implications. Finally, Section V concludes the paper and outlines directions for future research.

II. Theoretical Background

The model named SincNet [19] was originally built for audio signal processing and the $sinc(\cdot)$ based convolutions extract features from raw data. The main idea consists in modeling each convolution kernel as a band-pass filter in the frequency domain, by means of a difference between two lowpass filters, as in:

$$G[f, f_1, f_2] = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right), \qquad (1)$$

where $rect(a) = 1,\ \text{for}\ |a| < 1/2$ and 0 otherwise.

Applying the inverse Fourier transform, the corresponding impulse response of each filter is as follows:

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \qquad (2)$$

In practice, a truncated version of $g[n]$ is employed by means of windowing:

$$g_w[n, f_1, f_2] = g[n, f_1, f_2] \cdot w[n], \qquad (3)$$

where a Hamming window was chosen by [19] to mitigate the ripples in the passband and improve the attenuation in the stopband.

After the customized sinc-based convolutional layer, the SincNet architecture contains standard layers, with pooling, layer normalization, Leaky ReLU nonlinearity, dropout, a few convolutional layers and, finally, softmax [19].

Aiming to benefit from the interpretable characteristic present in SincNet, some works have adapted the model in multiple contexts, including scenarios with learnable and fixed parameters [23], as a feature extractor combined with modules that learn a sparse representation of the filtered data [12], and applications to high-dimensional EEG signals [24], and complex raw radar data [25]. In this way, a family of models based on SincNet has been introduced in the literature over the years and some of them are highlighted in this work.

The first extension is the SincShallowNet [10], a more lightweight version of the SincNet [19] applied to EEG signal processing. This approach is based on a three-block method: in the first block it applies temporal filtering via sinc-based convolutions, and spatial filtering with depthwise convolutional layers, aiming to reduce the number of parameters of the network and increase its interpretability; the second block performs a temporal pooling stage of the first block feature maps; the third block includes fully-connected layers to perform classification. Fig. 1 illustrates the first block of this architecture, highlighting how it adapts the approach for multivariate inputs. Specifically, the network input is structured as a 2D array of size $C \times T$, representing a single trial where C denotes the number of EEG channels and T represents the number of time samples.

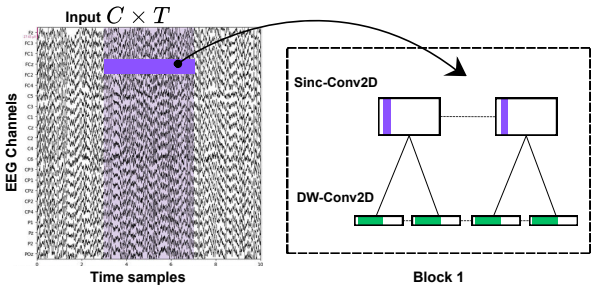


Fig. 1. Representation of the first block of SincShallowNet and its input format. Adapted from [10].

By doing that, SincShallowNet allows direct interpretation of the learned spectral–spatial features through kernel inspection, preserving an important characteristic of the SincNet

model, the interpretability, while training faster. SincShallowNet has also been applied in multiple contexts recently, e.g. mental healthcare [27] and drowsiness detection [28].

The second model is PF-Net [21], which adds higher flexibility for the input filters by parameterizing the cutoff limits of the band-pass filters based on the $sinc(\cdot)$ function. The approach models the frequency response by dividing it into small, learnable line segments defined by adjustable deformation points across low and high frequencies, which can be achieved through the reconstruction of the function g so that g contains more learnable variables. The frequency domain characteristics of each segment are defined below:

$$G[f, f_k, f_{k+1}] = \frac{h_{k+1} - h_k}{f_{k+1} - f_k}(f - f_k) + h_k, \tag{4}$$

where f_k and f_{k+1} are the low and high frequencies of each segment, with f within $[f_k, f_{k+1}]$, and h_k and h_{k+1} are their frequency amplitude intensity. All f and h are learnable.

In the time domain, the expression for the filtering frequency band from PF-Net can be defined as:

$$g[f, f_{beg}, f_{end}] = \sum_{k=beg}^{end-1} g[n, f_k, f_{k+1}]. \tag{5}$$

As a result, PF-Net accelerates convergence compared to standard CNNs and outperforms SincNet [19] due to its ability to extract more detailed features from the signal spectrum.

Also part of the SincNet family, we highlight IIRI-Net [22], a CNN network that was defined for audio processing as well and implements infinite impulse response (IIR) inspired filterbanks that are meaningful and controllable. In the first convolutional layer, it uses second order filters with IIR instead of truncated $sinc(\cdot)$ filters. The filters of this model have the shape of stable resonators and allow the adjustment of the bandwidth (2σ) and the center frequency (w_0), as in:

$$h[n] = \frac{\sin[(n+1)\omega_0]}{\sin(\omega_0)} e^{-n\sigma} u[n] \tag{6}$$

In this scenario, the only parameters required for learning are ω_0 and σ, which represent the location of the poles of the reference IIR filter. The learned filter is a band-pass filter that amplifies the frequencies around the center frequency, ω_0, and attenuates other frequencies. Additionally, as in SincNet [19], the gain of the filter is not a learning parameter because the subsequent layers are responsible for the importance of each filter's output.

To handle multivariate signals such as EEGs, both PF-Net [21] and IIRI-Net [22] required adaptation. These modifications are detailed in Section III (Methodology).

Alternatively, EEGNet [15] was conceived for BCI applications, and its first two layers operate with a direct signal processing meaning. A 2D convolution applies a fixed number of kernels along the time axis, each spanning a single channel and shared across all electrodes, so that every kernel acts as a temporal filter whose frequency response, obtained from the Fourier transform of its weights, reveals the bands emphasized during training. A depthwise convolution then operates along the channel axis, learning a fixed number of spatial filters per temporal filter, each assigning a coefficient to every electrode. A separable convolution and a fully-connected classifier complete the architecture, summarized in Fig. 2. Both temporal and spatial filtering stages are therefore inspectable by construction.

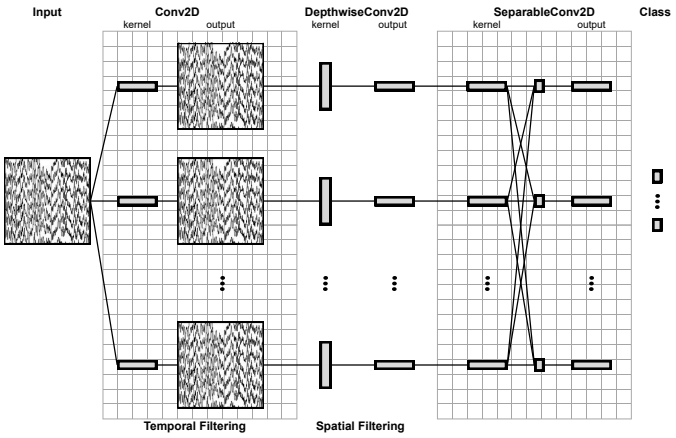


Fig. 2. EEGNet architecture. Adapted from [15].

EEGNet has become a significant baseline model to the BCI literature, being utilized as a benchmark in several subsequent works, and has also inspired the proposal of modifications of its core elements. These models were selected based on their potential to bring better classification performance while maintaining the interpretable characteristic of EEGNet.

The first EEGNet variant we can highlight is EEG-TCNet [16], which follows an usual structure for EEG data (convolutional layers for spatial-temporal feature extraction). This neural network differs by the implementation of TCN (Temporal Convolutional Network) module right after the depthwise and separable convolution modules and before entering a fully-connected network, aiming to identify temporal dependencies in motor imagery signals. Because of its limited computational complexity, it can be easily used in small devices. In the last years, this approach has been adapted with self-attention blocks to build a temporal convolutional transformer [29] and also adapted to learn more signal features through multiple convolutional kernels in parallel [26].

To work on the lack of interpretability of deep learning models in the context of motor imagery classification, EEG-ITNet [17] was proposed. This approach introduces an inception block in the beginning of the architecture, followed by a temporal convolutional (TC) block similar to the one in EEG-TCNet [16], a dimension reduction (DR) block and a classification block. Interestingly, this work explores an extra dimension in the inception block to achieve interpretability, i.e., the inception block consists of 2D convolutional layers applied over the time axis of multichannel EEG with different kernel sizes. These initial 2D layers act as frequency filters and, taking the Fourier transform of the convolutional kernels, they reveal the sub-bands selected during training. This model

has also been explored recently alongside data augmentation [30] and regularization [31] techniques.

The third variation selected in this work is EEGNex [18], which presents some structural changes in comparison with EEGNet, such as doubling the convolutional layer to capture more spatial patterns, and increasing the temporal learning by replacing the separable convolution with two standard convolutional blocks. As a result, EEGNex is an adaptable architecture that can be explored in different domains and applications, being considered recently as a state-of-the-art method for EEG classification and a widely explored model [32], [33].

All these variants were originally validated on motor imagery data, namely the BCI Competition [36] and High-Gamma [35] datasets, and none was assessed on SSVEP. It is possible to observe, therefore, a growing interest from the community in models that offer greater transparency regarding how brain data are processed, while also achieving high performance in classification tasks. Nevertheless, there is still room for studies that explore the application and analysis of interpretable deep architectures for BCI systems, including models not yet employed in the BCI context. We aim to reduce this gap by carrying out this comparative study with the seven aforementioned models, as well as by considering both MI and SSVEP BCI paradigms.

III. Methodology

We have selected EEG datasets based on SSVEP and MI paradigms, respectively: the Benchmark [34] and BCI Competition IV-2a [36] datasets.

The Benchmark dataset comprises recordings from 35 subjects. Each individual participated in six distinct sessions, with 40 stimuli each. On the monitor, a digital keyboard displayed each key flickering at a different frequency. For each stimulus, corresponding to the keys on the keyboard, a red square signaled the target to the individual. Thus, a cue appeared for half a second, followed by five consecutive seconds of stimulation, with a half-second pause before the next target. For each session, there were a few minutes of rest.

To capture the signal, 64 electrodes were positioned over the scalp following the 10-20 standard, with six samples collected for each key (corresponding to the sessions performed).

The BCI Competition IV-2a dataset comprises recordings from 9 subjects, with two sessions per participant, each consisting of six runs with 48 trials (12 per class), totaling 288 trials per session. The classes correspond to motor imagery tasks of the left hand, right hand, feet, and tongue.

At the beginning of each session, a 5-minute recording was performed to estimate EOG influence: two minutes with eyes open and fixed, one minute with eyes closed, and one minute with eye movements. Each trial began with a fixed cross on a black screen and an acoustic warning. After two seconds, an arrow indicated the task (up, down, left, or right) for 1.25 seconds, followed by an execution period with the central cross for an additional four seconds.

The first stage of data treatment involved splitting the available data in such a way that there was no leakage between training and test sets.

In the Benchmark dataset, as in the original paper, we have considered 9 electrodes from the occipital and parietal regions of the brain: Pz, PO5, PO3, POz, PO4, PO6, O1, Oz and O2. Also, from the 40 stimuli frequencies available, we have selected all integer frequencies from 8 Hz to 15 Hz, thus establishing a scenario with 8 classes in the BCI-SSVEP system. The proportion of trials was 3:1:2, i.e. 3 trials for training, 1 for validation and 2 for testing. The training data is segmented in 2s windows, with an overlap of 1s, while the test data remains raw. During the training process, we have used samples from all subjects.

In the BCI dataset, as we have 2 different sessions, one is fully used for testing and the other is split in half to create training and validation sets. All 22 electrodes were explored, and windowing process was not performed.

The data was pre-processed following the procedure adopted in [9], [17], i.e. (*i*) resampling to 128 Hz; (*ii*) z-Score normalization for training and validation sets; (*iii*) a band-pass filtering with a 3rd-order Butterworth filter to preserve the frequency band from 4 Hz to 38 Hz. We have only considered data from the same subject in the training/test stage.

All EEGNet models were accessed through BrainDecode [35], as well as SincShallowNet [10]. The other models from SincNet family, originally defined for audio processing, have the limitation of one channel for input data, which is not suited for EEG data. Based on that, we adapted these models by mimicking the structure of SincShallowNet, i.e., creating multiple filters for each channel. Thus, the models from SincNet family used in the experiments were SincShallowNet and, also, the adapted and renamed PFShallowNet and IIRIShallowNet.

This adaptation yields a controlled comparison within the SincNet family: the three networks share the same spatial filtering, pooling and classification blocks, differing solely in the first-layer kernels, i.e., truncated $sinc(\cdot)$ filters (SincShallowNet), piecewise-linear segments (PFShallowNet) and second-order IIR resonators (IIRIShallowNet). Hence, performance differences between them derive from the filterbank profile rather than from architectural capacity, which can be perceived through inspection of the learned frequency responses.

The performance of each model was evaluated by accuracy per subject, and the average of five independent runs was considered. Additionally, the temporal and/or spatial features learned by the interpretable layers were examined.

IV. Results

In this section, we present the experimental results obtained by the EEGNet- and SincNet-based models.

A. SSVEP

Fig. 3 exhibits the boxplots of the accuracy performance obtained by EEGNet and SincNet variants in the SSVEP dataset, whereas Table I presents the performance by individuals. In

bold are the best individual performances among the models of the EEGNet family, while highlighted in blue are the best performances obtained by a model of the SincNet family. The maximum accuracy per individual is further marked with an asterisk.

It is possible to notice that among all trained models, EEGNet presents the best performance, followed by EEGNex and EEG-ITNet by a maximum margin of 1.28 percentage points. We can also observe a notable performance gap occurred between subjects 16 and 32: while the latter was easily classified by most models, none exceeded 55% accuracy for the former. This is likely due to unique inter-subject variability and differing individual brain profiles, which can be recognized in Fig. 4 by comparing the frequency spectra for electrode Oz EEGs for stimuli frequencies of 10 Hz, 12 Hz and 15 Hz.

With respect to the SincNet family, there is a significant performance gap between the models, being the SincShallowNet the option that reaches the best average accuracy, approximately 3 percentage points worse than EEGNet. The more sophisticated bandpass filter designs from PF-Net and IIRI-Net were not beneficial in terms of final classification in this dataset.

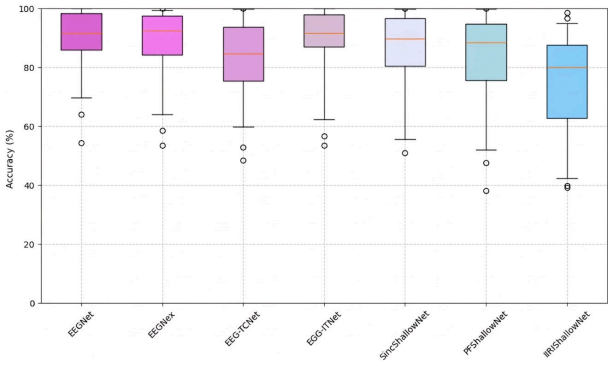


Fig. 3. Accuracy comparison by technique in the SSVEP paradigm.

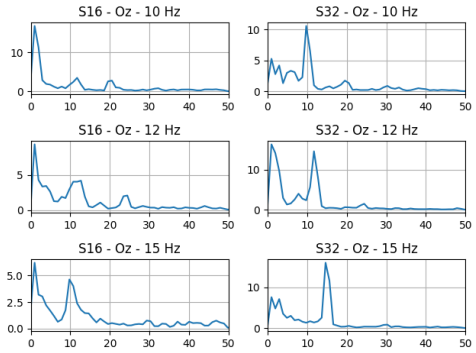


Fig. 4. Comparison between subjects 16 and 32 for different stimulus frequencies from electrode Oz.

Focusing on the temporal and spatial filters learned by EEGNet, Figs. 5 and 6 present four frequency spectrum corresponding to learned filters, in which we can observe the most relevant frequency band after model training and, based on the topoplots, the relative contribution from each electrode for the classification.

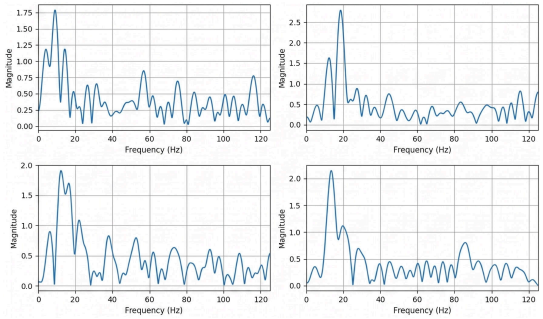


Fig. 5. Temporal filters learned by EEGNet for SSVEP.

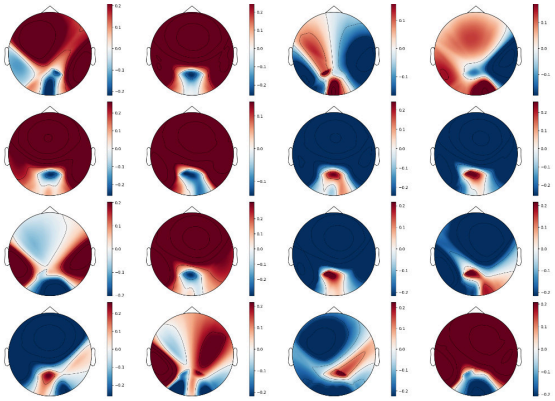


Fig. 6. Spatial filters learned by EEGNet for SSVEP.

It is important to observe that, in general, the spatial filters attribute larger coefficients to the occipital region than to the other regions of the brain, which is consistent with the visual nature of the SSVEP task. Besides that, we can notice that the frequency response of the temporal filters tend to prioritize regions related to stimulation frequencies (or their harmonics).

Another relevant remark is that the temporal filters can be quite different across the techniques, while the resulting spatial filters do not present clear changes. For example, Fig. 7 illustrates the temporal filters learned by the model with the worst average classification accuracy in the SSVEP paradigm. Clearly, the IIRIShallowNet model have not focused on the stimulation frequencies nor in their harmonics, as occurred with EEGNet.

B. MI

Fig. 8 exhibits the boxplots of the accuracies obtained by EEGNet and SincNet variants in the BCI Competition IV-2a dataset, while Table II separates the performance through the nine subjects.

Some remarks can be drawn from Table II. First, the performances we obtained with EEGNet and its variants are lower than those reported in [18]. However, it is worth mentioning that in the original works of EEGNex [18] and EEG-ITNet

TABLE I
ACCURACY (%) PER SUBJECT IN THE SSVEP PARADIGM.

Subject	EEGNet	EEGNex	EEG-TCNet	EEG-ITNet	SincShallowNet	PFShallowNet	IIRIShallowNet
1	**97.50***	95.62	92.19	94.37	**95.94**	92.81	86.87
2	**86.87***	79.69	78.44	86.56	**81.25**	73.75	77.81
3	**99.06***	97.50	93.13	97.19	**96.88**	92.19	89.69
4	**69.69***	64.06	59.69	62.19	67.50	**68.44**	56.25
5	90.00	92.50	79.69	**94.06***	**92.19**	87.50	80.00
6	90.62	94.69	89.06	**95.63***	**88.75**	85.63	80.00
7	**97.50***	**97.50***	95.00	97.19	**91.56**	90.63	82.19
8	93.75	**96.57***	85.62	94.38	80.31	**85.00**	70.00
9	**87.50**	**87.50**	84.38	86.25	**89.68***	88.44	77.19
10	96.88	96.56	90.94	**97.50***	93.75	**95.00**	85.94
11	72.19	**74.06***	62.81	68.12	50.94	**51.88**	39.06
12	94.06	88.75	91.87	**97.50***	**93.13**	84.07	80.00
13	**90.00***	86.88	75.00	88.13	**73.75**	70.94	55.94
14	**90.31***	84.37	70.00	88.13	**82.50**	77.19	58.75
15	**98.44***	**98.44***	92.50	**98.44***	92.19	**94.38**	80.94
16	**54.38**	53.44	48.44	53.44	**55.62***	38.12	39.69
17	89.69	87.19	84.69	**91.56***	**84.37**	79.69	78.44
18	85.00	83.75	77.50	**90.00***	**85.94**	80.00	75.00
19	79.69	**82.19***	64.69	80.00	**72.81**	63.12	55.00
20	**90.94**	89.06	81.88	90.31	**91.56***	90.31	82.81
21	**92.50***	88.75	75.31	87.81	88.75	**90.00**	75.62
22	**100.00***	97.19	**100.00***	**100.00***	**97.50**	**97.50**	88.13
23	78.44	58.44	**80.94***	70.00	**77.81**	69.06	50.62
24	**99.69***	98.44	**99.69***	98.44	**99.69***	99.38	96.56
25	98.13	**99.06***	97.82	98.44	**96.88**	95.63	83.12
26	**100.00***	98.75	99.06	99.38	**100.00***	**100.00***	90.94
27	98.75	**99.38***	94.38	98.75	98.13	**98.44**	89.06
28	91.56	**93.44**	84.06	91.25	**96.25***	89.69	89.06
29	79.69	84.06	74.38	**87.50***	**73.44**	58.44	59.06
30	85.00	**88.44***	75.31	75.94	**80.62**	80.00	66.25
31	**100.00***	**100.00***	**100.00***	**100.00***	99.38	**99.69**	95.00
32	**100.00***	**100.00***	99.69	**100.00***	**100.00***	**100.00***	98.44
33	**64.06***	**64.06***	52.81	56.56	**55.62**	47.50	42.19
34	**100.00***	97.82	97.50	99.38	98.13	**99.69**	93.75
35	**94.06***	93.12	91.56	91.56	89.69	**92.19**	77.50
Mean	**89.60***	88.32	83.43	88.74	**86.07**	83.04	75.05
Std	11.10	12.27	13.84	12.68	13.26	16.05	16.56
Parameters	3704	55 592	4040	8506	20 296	20 456	20 360

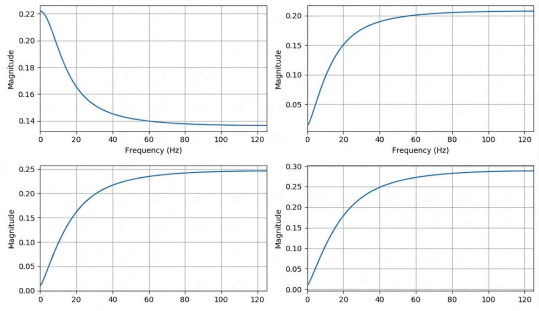


Fig. 7. Temporal filters learned by IIRIShallowNet for SSVEP.

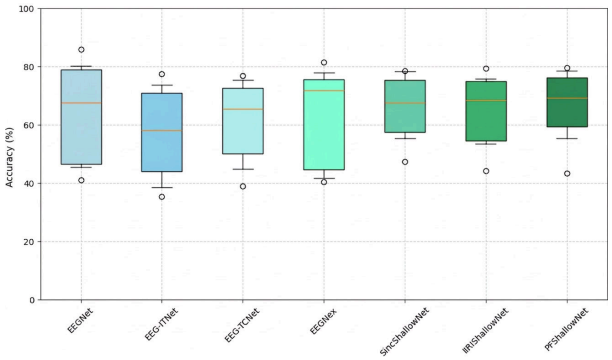


Fig. 8. Accuracy comparison by technique in the MI paradigm.

[17], a specialized initialization scheme was employed, in which the test samples participate in the parameter initial setting. Hence, there is some leakage between training and testing, so that the generalization performance may be overestimated. Second, the best average accuracy was attained by PFShallowNet, which obtained the top accuracy for five out of nine subjects. Regarding the number of trainable parameters, within the EEGNet family, the best average accuracy was reached by EEGNet, with 3 704 parameters in SSVEP and 2 484 in MI, whereas within the SincNet family the best models were SincShallowNet and PFShallowNet, with 20 296 and 11 812 parameters, respectively. All seven architectures remain lightweight, with fewer than 56 000 parameters. The learned temporal filters in Fig. 9 present a bandpass shape

TABLE II
ACCURACY (%) PER SUBJECT IN THE MI PARADIGM.

Subject	EEGNet	EEGNex	EEG-TCNet	EEG-ITNet	SincShallowNet	PFShallowNet	IIRIShallowNet
1	**75.35**	71.74	67.92	57.99	75.00	**75.49***	74.24
2	46.46	40.28	**53.13**	44.03	55.35	**55.42***	53.40
3	**85.97***	81.39	72.71	73.68	78.33	78.61	**79.31**
4	**54.24**	46.67	50.00	49.51	57.36	**59.44***	54.58
5	41.11	**41.60**	38.96	38.40	58.61	**60.21***	56.60
6	**45.42**	44.51	44.72	35.28	**47.29***	43.40	44.17
7	67.64	**72.36***	65.49	70.90	67.50	**69.17**	68.33
8	**79.03**	75.63	76.74	67.71	78.54	**79.65***	74.93
9	**80.28***	77.85	75.42	77.50	75.42	**76.25**	75.76
Mean	**63.94**	61.33	60.56	57.22	65.93	**66.40***	64.59
Std	17.28	17.46	14.11	16.00	11.59	12.49	12.56
Parameters	2484	55 972	4196	5324	11 652	11 812	11 716

with an emphasis on subregions within the alpha (8–13 Hz) and beta (13–30 Hz) bands. Motor imagery is associated with a reduction in power within the alpha (8–13 Hz) and beta (13–30 Hz) bands, a phenomenon known as event-related desynchronization (ERD), which typically occurs in the sensorimotor cortex contralateral to the imagined limb [17]. Several trained temporal filters, particularly from the EEGNet family, exhibit this effect through a decrease in magnitude within these specific frequency bands.

The spatial filters shown in Fig. 10 highlight sensor and motor areas of the brain, around the center of the scalp. The activations correspond to contralateral motor regions, which is consistent with the imagery of right- or left-hand movements. The filters of this section are from the training data of subject 3, in which we have the best performance.

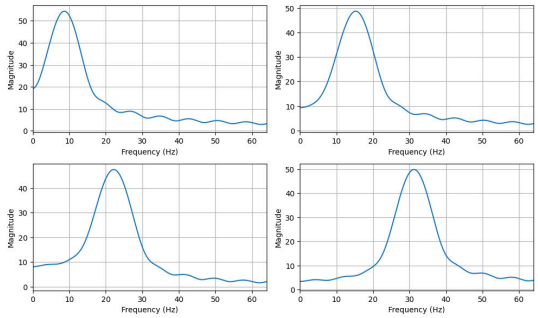


Fig. 9. Temporal filters learned by PFShallowNet for MI.

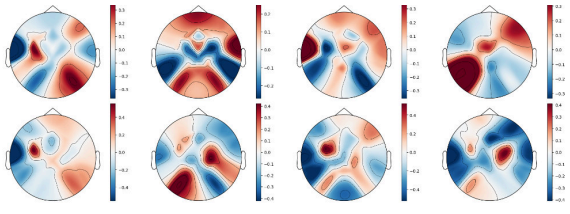


Fig. 10. Spatial filters learned by PFShallowNet for MI.

Similarly to the SSVEP context, in the MI paradigm the temporal filters also exhibit an important contrast across techniques compared to the spatial filters. Fig. 11 displays a subset of filters from the model with the lowest average classification accuracy in MI, illustrating a clear failure to exhibit the distinct bandpass characteristic observed in Fig. 9. As the SincNet variants differ only in the first-layer kernels, inspecting them indicates that the filters should match the spectral signature of the paradigm.

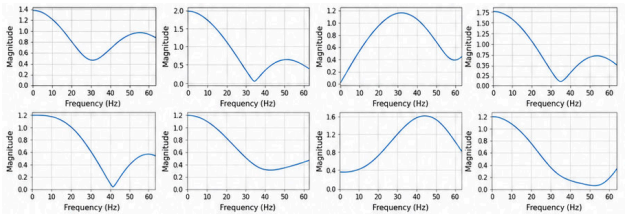


Fig. 11. Subset of temporal filters learned by EEG-ITNet for MI.

V. CONCLUSION

This study evaluated the performance and interpretability of different neural network models in different EEG paradigms. We have identified in the literature approaches that incorporate interpretable elements in two families of models: EEGNet, and its variants EEG-TCNet, EEG-ITNet and EEGNex; and SincNet, with the variants SincShallowNet, PF-Net and IIRI-Net. In the latter case, we have implemented extensions of the original models of PF-Net and IIRI-Net in order to allow the processing of multivariate signals, which we defined as PFShallowNet and IIRIShallowNet.

It was considered the scenarios of MI, already explored in EEG-ITNet [17] and EEGNex [18], and SSVEP, which is less considered in this context. Besides the observation of the performance of the models in terms of accuracy, we also have given special attention to how they learn and represent relevant neural information, a delicate matter when dealing with human data and using AI for decision-making processes.

On the one hand, for SSVEP classification, EEGNet and its variants performed well, though the original architecture achieved the highest average accuracy. Within the SincNet family, SincShallowNet yielded the best results. On the other hand, SincNet-based models surpassed EEGNet and its variants in the MI paradigm, where PFShallowNet yielded the top average accuracy.

The learned filters were directly inspectable in both paradigms. In the SSVEP context, temporal filters emphasized target frequencies, while spatial filters assigned larger coefficients over the occipital cortex, consistent with visual tasks. For MI, the temporal filters formed distinct bandpass shapes, while spatial activations concentrated over contralateral motor regions, reflecting the patterns expected during hand motor imagery.

For future works, we plan to extend the experiments to the cross-subject scenario and validate the generalizability of the PFShallowNet and IIRIShallowNet architectures by extending them to MI datasets that feature a larger number of subjects and greater inter-subject variability. We also intend to complement the present analysis with statistical comparisons between models and quantitative measures of the stability of the inspected filters.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[2] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers," Journal of Neural Engineering, vol. 18, no. 3, p. 031002, May 2021.

[3] D. Xu, F. Tang, Y. Li, Q. Zhang, and X. Feng, "An analysis of deep learning models in SSVEP-based BCI: A survey," Brain Sciences, vol. 13, no. 3, p. 483, Mar. 2023.

[4] X. Gu et al., "EEG-Based Brain-Computer Interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 18, no. 5, pp. 1645–1666, Sept.-Oct. 2021.

[5] A. Emami, N. Kunii, T. Matsuo, T. Shinozaki, K. Kawai, and H. Takahashi, "Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images," NeuroImage: Clinical, vol. 22, 2019, p. 101684.

[6] A. Donati et al., "Long-term training with a brain-machine interface-based gait protocol induces partial neurological recovery in paraplegic patients," Scientific Reports, vol. 6, p. 30383, Aug. 2016.

[7] Z. Yin and J. Zhang, "Cross-session classification of mental workload levels using EEG and an adaptive deep learning model," Biomedical Signal Processing and Control, vol. 33, pp. 30–47, Mar. 2017.

[8] H. Alwasiti, M. Yusoff, and K. Raza, "Motor imagery classification for brain computer interface using deep metric learning," IEEE Access, vol. 8, 2020, pp. 109949–109963.

[9] N. Sharma et al., "Recent trends in EEG-Based motor imagery signal analysis and recognition: A comprehensive review," IEEE Access, vol. 11, 2023, pp. 10600–10623.

[10] D. Borra, S. Fantozzi, and E. Magosso, "Interpretable and lightweight convolutional neural network for EEG decoding: application to movement execution and imagination," Neural Netw., vol. 129, pp. 55–74, Sept. 2020.

[11] T. Izzuddin, N. Safri, and M. A. Othman, "Compact convolutional neural network (CNN) based on SincNet for end-to-end motor imagery decoding and analysis," Biocybernet. Biomed. Eng., vol. 41, no. 3, pp. 1041–1056, July–Sept. 2021.

[12] C. Liu et al., "SincNet-based hybrid neural network for motor imagery EEG decoding," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 30, pp. 1234–1244, May 2022.

[13] M. Attia, I. Hettiarachchi, S. Mohamed, M. Hossny, and S. Nahavandi, "A frequency domain classifier of steady-state visual evoked potentials using deep separable convolutional neural networks," in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Miyazaki, Japan, 2018, pp. 2134–2139.

[14] N. Aznan, S. Bonner, J. Connolly, N. Al Moubayed, and T. Breckon, "On the classification of SSVEP-based dry-EEG signals via convolutional neural networks," in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Miyazaki, Japan, 2018, pp. 3726–3731.

[15] V. Lawhern, A. Solon, N. Waytowich, S. Gordon, C. Hung, and B. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," Journal of Neural Engineering, vol. 15, no. 5, p. 056013, Jul. 2018.

[16] T. Ingolfsson et al., "EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC), 2020, pp. 2958–2965.

[17] A. Salami, J. Andreu-Perez, and H. Gillmeister, "EEG-ITNet: An explainable inception temporal convolutional network for motor imagery classification," IEEE Access, vol. 10, 2022, pp. 36672–36685.

[18] X. Chen, X. Teng, H. Chen, Y. Pan, and P. Geyer, "Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNex," Biomedical Signal Processing and Control, vol. 87, p. 105475, Jan. 2024.

[19] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with SincNet," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2019, pp. 7335–7339.

[20] D. Saritha, A. Kirupakaran, R. Laskar, and M. Choudhury, "SincSquareNet: Deep neural network-based speaker identification for raw speech," in Advances in Distributed Computing and Machine Learning, Springer, 2023, pp. 477–486.

[21] W. Li, Z. Tan, Z. Xia, D. Wu, and J. Ning, "PF-Net: Personalized filter for speaker recognition from raw waveform," in Proc. Int. Conf. Adv. Data Min. Appl. (ADMA), Springer, 2023, pp. 642–656.

[22] H. Fayyazi and Y. Shekofteh, "IIRI-Net: An interpretable convolutional front-end inspired by IIR filters for speaker identification," Neurocomputing, vol. 558, p. 126767, Nov. 2023.

[23] S. H. Kim, A. H. Jo, and K. C. Kwak, "SincNet-based speaker identification for robotic environments with varying human–robot interaction distance," Electronics, vol. 13, no. 23, Art. no. 4836, 2024.

[24] H. Zeng et al., "EEG emotion classification using an improved SincNet-based deep learning model," Brain Sci., vol. 9, no. 11, art. no. 326, 2019.

[25] S. Biswas, C. O. Ayna, S. Z. Gurbuz and A. C. Gurbuz, "CV-SincNet: Learning complex sinc filters from raw radar data for computationally efficient human motion recognition," in IEEE Transactions on Radar Systems, vol. 1, 2023, pp. 493-504.

[26] H. Liu and Y. Hu, "EEG emotion signal recognition method based on improved EEG-TCNet," in Proceedings of the 2025 3rd International Conference on Internet of Things and Cloud Computing Technology, 2025, pp. 302–307.

[27] R. Sakurai et al., "End-to-End machine learning for depressive state classification via EEG and fNIRS," 2026, *arXiv:2606.11555*.

[28] M. Malik, P. Sharma, G. K. Punj *et al.*, "Multi-body sensor based drowsiness detection using convolutional programmed transfer VGG-16 neural network with automatic driving mode conversion," Sci. Rep., vol. 15, art. no. 8838, 2025.

[29] A. H. P. Nguyen, O. Oyefisayo, M. A. Pfeffer, and S. H. Ling, "EEG-TCNTransformer: A temporal convolutional transformer for motor imagery brain–computer interfaces," Signals, vol. 5, no. 3, 2024, pp. 605–632.

[30] M. K. Titkanlou, E. Monjezi, and R. Moucek, "A type of EEG-ITNet for motor imagery EEG signal classification," in Proc. 17th Int. Joint Conf. Biomed. Eng. Syst. Technol. (BIOSTEC), vol. 2, 2024, pp. 257–262.

[31] P. S. Nzakuna, V. Paciello, V. Gallo, A. Lay-Ekuakille, and A. K. Lusala, "FT-DropBlock: A novel approach for spatiotemporal regularization in EEG-based convolutional neural networks," in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC), Chemnitz, Germany, 2025, pp. 1–6.

[32] F. Tian, L. Zhang, Q. Pan, J. Liu, Q. Zhao, and B. Hu,"MAMILS: A memory-aware multiobjective scheduler for real-time embedded EEG depression diagnosis," IEEE Trans. Parallel Distrib. Syst., vol. 37, no. 3, pp. 600–617, Mar. 2026.

[33] Y. Zhang et al., "Dse-net based aphasia patient identification and semantic decoding from EEG signals," Biomed. Signal Process. Control, vol. 119, Art. no. 109835, 2026.

[34] Y. Wang, X. Chen, X. Gao, and S. Gao, "A benchmark dataset for SSVEP-based brain-computer interfaces," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 10, 2017, pp. 1746–1754.

[35] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Human Brain Mapping, vol. 38, no. 11, 2017, pp. 5391–5420.

[36] C. Brunner, R. Leeb, and G. Müller-Putz, "BCI Competition 2008 – Graz data set a," 2008.

tDCS stimulation effects in post-stroke rehabilitation: A preliminary EEG LORETA study

Eliana M. Santos
COENS-PINHAIS
Instituto Federal do Paraná – IFPR
Pinhais, Brazil
0000-0003-0824-802X

Bruna M. Carlos
DRCC-IFGW
Universidade Estadual de Campinas – UNICAMP
Campinas, Brazil
0000-0002-9020-4992

Pedro F. G. de Vazques
DRCC-IFGW
Universidade Estadual de Campinas – UNICAMP
Campinas, Brazil
0000-0003-3088-1315

Saulo S. Feitosa
DRCC-IFGW
Universidade Estadual de Campinas – UNICAMP
Campinas, Brazil
sauloffeitosa@gmail.com

Gabriel C. de Melo
DRCC-IFGW
Universidade Estadual de Campinas – UNICAMP
Campinas, Brazil
0000-0002-4360-7617

Cássio V. Ruas
DRCC-IFGW
Universidade Estadual de Campinas – UNICAMP
Campinas, Brazil
0000-0002-6876-7190

Gabriela Castellano
DRCC-IFGW
Universidade Estadual de Campinas – UNICAMP
Campinas, Brazil
0000-0002-5927-487X

***Abstract*— Stroke is one of the leading causes of motor disability, and recovery depends on mechanisms of cortical plasticity. Transcranial direct current stimulation (tDCS) has been used as a complementary approach to promote functional reorganization, and electroencephalography (EEG) has been applied to investigate this reorganization. However, most works using EEG analyze the data in the sensors' space. The aim of this study was to investigate the effect of a tDCS-based rehabilitation program in functional reorganization in the EEG sources' space, using LORETA software. More specifically, EEG band power in the sources' space was evaluated during motor imagery tasks. For this, 17 post-stroke subjects were analyzed, divided into two groups: 8 in the active group and 9 in the sham group. The active group received anodal tDCS on the lesioned hemisphere and the sham group received sham tDCS stimulation for 10 daily sessions during two weeks. Statistical analysis identified nine Brodmann areas (BAs) with significant variations, among which BA3 and BA6 stood out. Overall, the active group showed more consistent clinical improvement accompanied by balanced increases in the slow bands (delta, theta, and alpha), suggesting a more stable pattern of cortical reorganization. The sham group also exhibited elevated values in these bands, but inflated by extreme responses, reflecting greater heterogeneity. Fast bands (beta2 and gamma1) remained close to stability in both groups, indicating that the main differences between them occurred in the slow bands. These findings reinforce that the clinical gains observed in the active group are associated with more balanced functional reorganization, whereas the sham group presented accentuated variability and less predictable patterns.**

Keywords*— *Motor Imagery; tDCS; sLORETA; Cortical Reorganization; Neurorehabilitation.

I. INTRODUCTION

Stroke is one of the leading causes of motor disability, and recovery depends on mechanisms of cortical reorganization [1]. Motor imagery is recognized as a strategy capable of activating cortical networks involved in movement, generating consistent neural patterns [2]. In addition to its use in brain–computer interface (BCI) systems, where it demonstrates applicability in different contexts [3], recent studies have also employed it as an assessment task in post-stroke patients. For example, [4] analyzed brain connectivity during motor imagery and movement execution, showing that it can reveal functional reorganization patterns crucial for monitoring clinical evolution. Similarly, [5] highlighted that EEG paradigms based on motor imagery are valuable tools for assessing motor deficits and tracking recovery, reinforcing its relevance as an experimental evaluation approach.

Transcranial direct current stimulation (tDCS) has been established as a non-invasive technique capable of modulating cortical excitability, influencing synaptic plasticity, and enhancing functional recovery [6], [7]. Studies such as those by [7], [8] demonstrated that tDCS, already established as a non-invasive technique capable of modulating cortical excitability and influencing synaptic plasticity, can have its effects assessed with motor imagery, contributing to the understanding of functional reorganization mechanisms and enhancing recovery. Clinical studies indicate that anodal stimulation over motor regions can promote the reorganization of circuits involved in movement planning and execution, showing promising results in post-stroke rehabilitation. Several studies have sought to understand the effects of these interventions through different analytical approaches, such as the analysis of sensorimotor band power and electroencephalogaphy (EEG) classification [6], systematic mapping of functional outcomes in post-stroke patients [7], and the evaluation of functional connectivity and brain symmetry in protocols combining tDCS with XR [9]. Among them, sLORETA (Standardized Low Resolution Electromagnetic Tomography), proposed by [10], [11], is a method for estimating the localization of cortical sources from EEG, capable of identifying functional reorganization patterns and providing a spatial view of spectral modulation. This type of approach has already been applied to investigate neural changes associated with electrical stimulation in frontoparietal networks [12], highlighting the potential of source localization methods to contribute to the understanding of plasticity mechanisms. Given this scenario, the present study conducted a preliminary analysis of the effects of tDCS using LORETA software to investigate

cortical reorganization patterns during motor imagery and provide additional evidence on the neurophysiological mechanisms underlying neurorehabilitation patterns.

II. Methodology

Fig. 1. presents the methodological pipeline adopted in this study, from the initial EEG signal processing to the clinical correlation. The diagram summarizes the main stages: (A) Dataset, (B) Signal Processing, (C) LORETA software (Spectral Source Transformation, statistical analysis, and selection of regions of interest), (D) Percentage Change Analysis, and (E) Clinical Correlation.

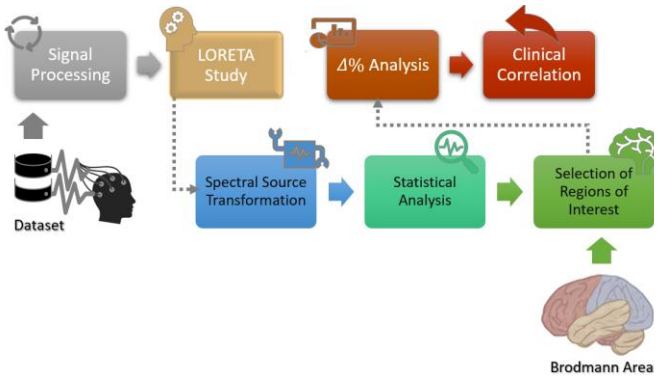


Fig. 1 - Methodological workflow of the study, from signal processing to clinical correlation.

We implemented the sLORETA technique using the LORETA software. Hereafter, 'LORETA' refers to the software, whereas 'sLORETA' denotes the localization method.

A. Dataset

The experimental protocol adopted in this study, combining transcranial direct current stimulation (tDCS) applied to the lesioned hemisphere (C3 or C4) with extended reality tasks. The dataset is composed of 17 patients with chronic stroke (mean 30 ± 23.3 months post-stroke; 9 women and 8 men; mean age 58.9 ± 12.7 years, range 33–86). All had a single stroke more than six months prior, were older than 18 years, and had no history of seizures, pacemaker, or metallic implants. The study adopted a single-blind design, with IDs 1–10 in the active tDCS group and IDs 11–20 in the sham group. In the active group (n=8), mean age was 60.2 ± 12.0 years, mean time post-stroke 35.7 ± 25.1 months, with 6 cases of right hemisphere lesions and 2 cases of left hemisphere lesions. In the sham group (n=9), mean age was 57.6 ± 14.0 years, mean time post-stroke 24.2 ± 21.0 months, with 5 cases of right hemisphere lesions and 4 cases of left hemisphere lesions.

The rehabilitation protocol consisted of 10 consecutive sessions, with upper and lower limb exercises in extended reality (GestureCollection, BRAINN_XR[13]. Tasks included puzzle-solving with reaching movements, stationary walking with virtual obstacles and environment based on Google Maps. Neuromodulation was applied using anodal tDCS (MicroEstim, NKL, Brazil) with 2 mA, 30-s ramp-up/down, for 30 minutes per session, during the extended reality exercises. Electrodes (35 cm², saline-soaked sponges) were positioned over the ipsilesional primary motor cortex (C3 or C4) and contralateral supraorbital region (Fp1 or Fp2). In the sham group, current was turned off after 1 min. EEG was recorded in the first and last session during motor imagery (MI) tasks, before rehabilitation, with 16 dry electrodes (g.USBamp, g.tec, Austria), positioned according to the 10–20 system and referenced to the left mastoid. Channels included F1, F2, Cz, AFz, F5, F6, C1, C2, C3, C4, C5, C6, P1, P2, P5, P6; sampling frequency 256 Hz. Each MI recording was composed of 8 trials (total duration of 128 s) consisting of: a cross appeared for 2 s, followed by an instruction (hand/foot, left/right) for 6 s and 8 s of rest before the next trial. Clinical assessment was performed before the first and after the last session, using the upper extremities Fugl-Meyer assessment (UE-FMA) (0–66), lower extremities FMA (LE-FMA) (0–34), Berg Balance Scale (0–56), and the 10-meter walk test.

B. Signal Processing

EEG signals were preprocessed using EEGLAB and MATLAB. Filtering was performed in two steps: a 4th-order Butterworth high-pass filter with a cutoff at 1 Hz to remove slow drifts and movement artifacts, and a 4th-order Butterworth low-pass filter with a cutoff at 40 Hz to reject high-frequency components and residual noise around 50–60 Hz. After filtering, signals were re-referenced to the average of all channels (average reference). Experimental events were adjusted to mark the onset of motor imagery instructions, considering a 2-second delay after the fixation marker. Independent Component Analysis (ICA) was applied using the runica algorithm (extended mode), a method already consolidated in the literature EEG [14]. Subsequently, automatic classification of independent components was performed with ICLabel (v1.7), which assigns probabilities to categories such as Brain, Eye, Muscle, and Line Noise. Automatic removal was restricted to components classified as Eye (probability > 0.90) and Line Noise (> 0.95), while Muscle components (> 0.85) were flagged for manual inspection. This procedure proved effective: ocular artifacts were identified and removed without compromising motor signals, and most subjects required little or no correction, indicating good data quality during acquisition [15].

C. LORETA

LORETA (Low Resolution Electromagnetic Tomography) is a functional neuroimaging technique developed to address the EEG inverse problem by estimating cortical current density distributions from scalp-recorded potentials. The method restricts the solution space to cortical gray matter and applies spatial smoothing to ensure continuity across neighboring voxels [11]. The forward problem can be expressed as:

$$b = Ax + n \tag{1}$$

where b represents the EEG signals measured at the electrodes, A is the lead field (gain) matrix linking cortical sources to sensors, x denotes the source activations over time, and n is the noise term. The inverse problem consists of estimating x from b. Since the number of dipoles far exceeds the number of electrodes, the solution is underdetermined. LORETA addresses this by applying Tikhonov regularization, formulated as:

$$\min_x \| Ax - b \|^2 + \lambda \| Wx \|^2 \tag{2}$$

where λ is a smoothing parameter and W is the regularization operator encoding prior information about the solution. The analytical solution is given by:

$$x^* = A^T (AA^T + \lambda W)^{-1} b \quad (3)$$

This transformation provides an estimate of cortical current densities, imposing spatial smoothness and favoring superficial dipoles. In LORETA, the regularization operator is based on a weighted Laplacian of the sources, ensuring spatial continuity and reducing the complexity of the inverse problem [11]. Thus, scalp potentials are converted into three-dimensional maps of neuronal activity, grouped into regions of interest (ROIs).

Previously cleaned and epoched signals were organized into two sessions (S1 and S10 tDCS), considering only subjects within each group (active or sham). Trials were classified according to the experimental condition (Left or Right MI), and each segment was extracted individually. So as to have the lesions corresponding always to the same side (right), for left hemisphere lesion subjects channel flipping of symmetric pairs (F1/F2, F5/F6, C1/C2, C3/C4, C5/C6, P1/P2, P5/P6) was applied, while midline channels (Cz and AFz) remained unchanged. Signals were then segmented into 2-second windows with 50% overlap, ensuring greater stability for spectral analysis. Each window was saved in text format (.txt), structured by subject, group, and experimental condition, forming the dataset prepared for LORETA analysis. From this organization, the next step consisted of spectral analysis, aimed at characterizing the frequency distribution of the signals across cortical regions.

Spectral Source Transformation

In LORETA, the brain is discretized into thousands of voxels, each representing a potential source of electrical current in three-dimensional coordinates (x, y, z). The lead field matrix A connects each voxel to the electrode measurements, and the inverse solution estimates the current density at each voxel. This discretization allows scalp-recorded signals to be transformed into three-dimensional maps of neuronal activity [16], [17].

1) EEG to Cross-Spectrum: EEG segments were transformed into cross-spectral matrices between channels. Formally, for two signals $xi(t)$ and $xj(t)$, the cross-spectrum is obtained by the Fourier transform of their cross-correlation function:

$$C_{ij}(f) = F\{R_{ij}(\tau)\} \quad (4)$$

where $R_{ij}(\tau)$ is the cross-correlation between channels i and j. This formulation is widely used in EEG spectral analysis [18].

2) Cross spectra to sLORETA: The cross-spectral matrices were then projected into the cortical source space using the sLORETA[10] inverse solution. In this phase, the inverse transformation matrix G^+ is applied to the spectral data:

$$S(f) = G^+ . C(f) \quad (5)$$

where $S(f)$ represents the spectral distribution of current densities across cortical voxels [19].

3) sLORETA to ROIs: In this process, voxel-level estimates are aggregated, and the spectral activity in each ROI is calculated as the mean of voxel estimates within the region:

$$ROI_k(f) = 1/N_k \sum_{v \in ROIk} S_v(f) \quad (6)$$

where v denotes each voxel belonging to ROI_k, $S_v(f)$ is the spectral current density estimated for voxel v at frequency f, and N_k is the total number of voxels in ROI_k. This procedure reduces data complexity and facilitates statistical comparisons across cortical regions [11], [17], [19]. All source localization procedures were performed using the standard sLORETA implementation, which relies on the default MNI head model, predefined lead field matrix, standard source grid, and default regularization parameters[10].

Statistical Analysis

We used the statistical module of the LORETA, which implements the Statistical non-Parametric Mapping (SnPM) methodology based on non-parametric randomization tests and correction for multiple comparisons [20], [21]. Analyses included between-session comparisons (session 1 × session 10), and between-group comparisons (active × sham), enabling evaluation of longitudinal and differential effects of the intervention. LORETA estimates the cortical origin of signals recorded by scalp electrodes and associates activation patterns with specific anatomical regions [11], [19]. For each frequency band (delta: 1–4 Hz; theta: 4–8 Hz; alpha: 8–12 Hz; low beta: 13–20 Hz; high beta: 21–30 Hz; gamma: 31–45 Hz [22], the software computes t-values derived from voxel-wise t-Student tests [23]. These values are linked to MNI coordinates, from which the software identifies possible Brodmann areas at different neighborhood distances (0 mm, 5 mm, 7 mm, 20 mm), allowing interpretation of the cortical regions involved in each comparison. The statistical process consisted of extracting t-values for each voxel and frequency band, defining the threshold of significance according to the number of subjects and the adopted significance level, and then comparing each value with this threshold. Values above the threshold were considered significant, those close to the threshold were classified as trends, and those far below were deemed non-significant. Results were organized into tables containing the t-value, statistical status, corresponding Brodmann area, and MNI coordinates. Finally, only regions with significant values were selected as regions of interest (ROIs) for subsequent analysis. Thus, the LORETA statistical analysis allowed the identification of the most relevant cortical activation peaks, ensuring that the interpretation of results was based on robust evidence corrected for multiple comparisons.

Selection of Regions of Interest

Based on the Brodmann areas (BA) identified in the statistical analysis, regions of interest were defined and linked to the corresponding frequency bands and EEG channels, establishing a direct relationship between the cortical estimates obtained by LORETA and the electrode recordings. The

regions identified included the postcentral gyrus (BA 2 and BA 3), the frontal/precentral region (BA 6), the middle frontal gyrus (BA 8 and BA 9), the insula (BA 13), the occipital cortex (BA 18), and multimodal parietal regions such as the angular gyrus (BA 39) and the supramarginal gyrus (BA 40). In total, 9 regions of interest were selected to guide the subsequent stages of investigation.

D. Percentage Change Analysis

Percentage variations were calculated by comparing mean values obtained at session 1 (S1) and session 10 (S10) in both the active and sham groups, according to the following formula:

$$Variation\ (\Delta\%) = \frac{S_{10}-S_1}{S_1}\ X\ 100 \qquad (7)$$

This comparison was chosen to evaluate longitudinal evolution throughout the protocol, thus reflecting the accumulated effects of the intervention. The analysis highlighted two regions of interest, BA3 and BA6, both showing relevant percentage changes across sessions. Each region of interest presented specific values for each subject and frequency band, reinforcing the independence of the findings. However, the analyses focused on BA6 due to its functional relevance for motor and premotor functions, which are central to the objectives of the study. BA6 encompasses premotor and supplementary motor regions, recognized as fundamental for movement preparation, execution, and motor imagery — processes directly modulated by tDCS [24], [25].

E. Clinical Correlation

In the final stage of the protocol, correlations were performed between the neurophysiological indicators obtained through LORETA and the clinical measures collected across sessions. The objective was to determine whether the percentage variations observed in the regions of interest, particularly BA6, were associated with changes in patients' functional outcomes. Pearson's correlation coefficient was applied as the standard method for assessing linear associations between continuous variables [26]. Despite the small sample size, this choice is justified by the widespread use of Pearson's coefficient in exploratory studies of neurophysiology and rehabilitation, where it provides interpretable estimates of the strength of association between continuous variables even in limited samples [27], [28]. This approach is consistent with previous studies investigating the relationship between cortical modulation induced by tDCS and clinical outcomes in post-stroke patients [9], reinforcing its applicability in rehabilitation contexts.

III. Results

Fig. 2 illustrates the mean percentage variations (Δ%) between session 1 and session 10 in the active and sham groups, evidencing cortical reorganization [24], [29] and particularly highlighting BA3 and BA6, which showed the most consistent changes. The analyses focused on BA6 detailed in Fig. 3, which presents percentage variations across spectral bands (δ, θ, α, β_1, β_2, γ_1). In the active group, the largest variations occurred in slow bands (δ, θ, α), confirming the predominance of effects in slower rhythms [22], [30]. In the sham group, more consistent variations were observed in δ, θ, α, and β_1, suggesting that, while less homogeneous, these bands also contributed to cortical reorganization. In both groups, the fast bands (β_2, γ_1) showed reduced variations.

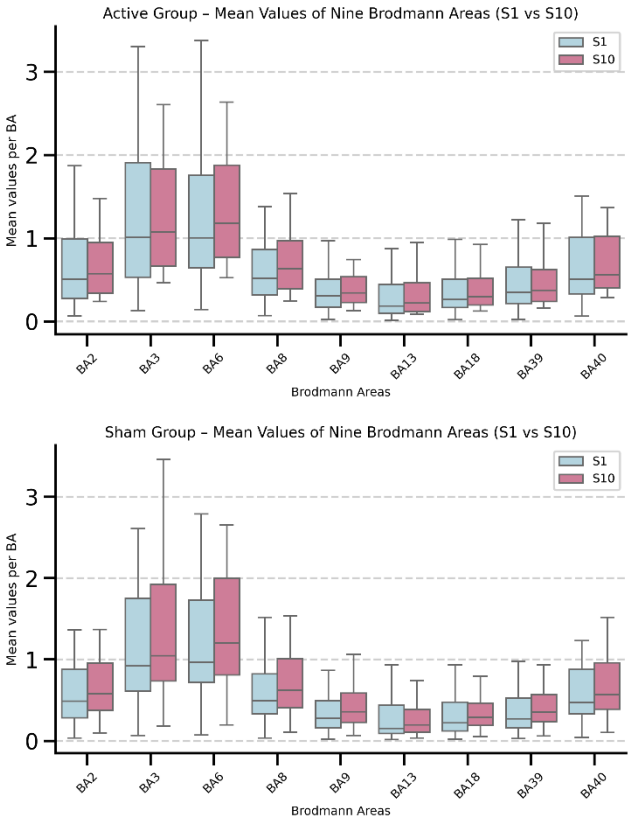


Fig. 2 - Mean values of EEG broadband (all frequency bands) spectral changes at S1 and S10 for Active and Sham groups across all investigated Brodmann areas.

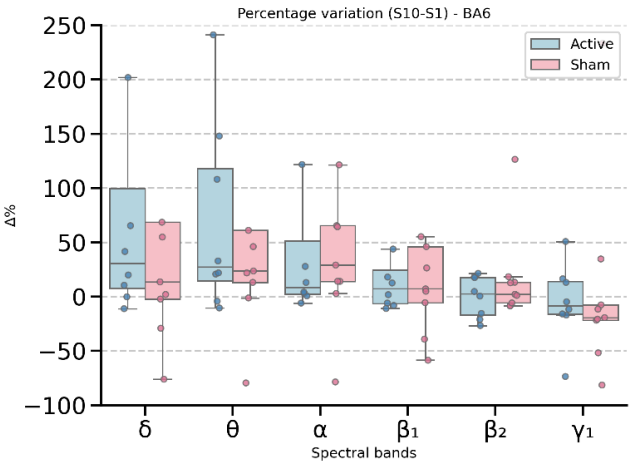


Fig. 3 - Percentage variation (Δ%) in BA6 across spectral bands for Active and Sham groups.

A total of 14 clinical metrics were initially evaluated, including FM upper limb (fm_sup_tot, fm_sup_66), FM lower limb (fm_inf_tot, fm_inf_34), FM total (fm_tot_100), Mini-Mental, MIF, Berg Balance Scale, TUG, TUG_cog, and gait parameters (time in seconds, number of steps, velocity, and cadence). Among these 14 metrics, eight showed significant

correlations with EEG spectral bands, as represented in the heatmap (Fig. 5): (1) Berg Balance Scale, (2) FM lower limb – 34 items (FM_inf_34), (3) FM lower limb – total (FM_inf_tot), (4) FM upper limb – 66 items (FM_sup_66), (5) FM total – 100 items (FM_tot_100), (6) MIF, (7) TUG, and (8) TUG_cog.

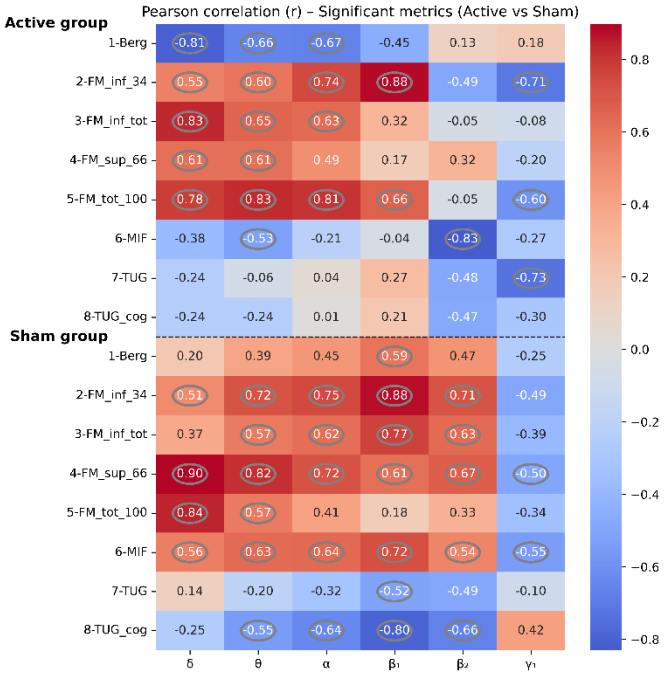


Fig. 4 - Pearson correlations (r) between significant metrics and the spectral bands in the Active and Sham groups. Dashed horizontal line separates the band pairs and gray circles highlight correlations that met the criteria for significance: sufficient strength ($|r| \geqslant 0.5$) and statistical evidence ($p < 0.05$).

Correlations between EEG and clinical metrics were calculated using the Pearson method [31], and only statistically significant results ($|r| \geqslant 0.5$, $p<0.05$) are highlighted. In the active group, significant correlations are more intense in δ, θ, and α, whereas in the sham group they are more homogeneously distributed, with greater concentration in β_1 and γ_1. The patterns reinforce that the clinical improvements observed in the active group are associated with cortical reorganization compatible with synchronization in slow rhythms [22], [32], [33]. Table 1 presents the individual and mean percentage variation values of the eight significant metrics in the active and sham groups, allowing visualization of the magnitude of clinical and functional changes. In the active group, consistent gains were observed, especially in motor metrics, with averages above 8%, and while some subjects showed neutral results in certain metrics (s1, s6, s10), these cases were less frequent, suggesting a more uniform progression.

Table 1 - Individual and mean percentage variation values of clinical assessments (Active vs Sham groups). (1) Berg Balance Scale, (2) FM lower limb – 34 items (FM_inf_34), (3) FM lower limb – total (FM_inf_tot), (4) FM upper limb – 66 items (FM_sup_66), (5) FM total – 100 items (FM_tot_100), (6) MIF, (7) TUG, and (8) TUG_cog. M: mean, SD: standard deviation.

Clinical outcomes (%)								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Active Group								
s1	0.0	0.0	18.0	35.3	16.2	0.0	-23.4	-30.2
s3	16.2	0.0	0.0	11.1	4.0	2.9	1.7	-4.9
s4	13.6	5.0	1.4	13.6	9.5	0.0	-5.3	-21.7
s6	22.2	0.0	0.0	9.8	5.4	1.3	-3.3	14.7
s7	7.8	3.0	7.5	0.0	1.0	3.5	-13.3	-12.5
s8	11.1	11.5	3.9	11.4	11.4	3.5	3.8	-2.9
s9	7.7	6.8	5.1	1.8	3.4	0.9	-31.3	-17.7
s10	1.9	25.0	11.1	18.2	20.6	1.7	0.0	0.0
M	**10.1**	**6.4**	**5.9**	**12.6**	**9.0**	**1.7**	**-8.91**	**-9.4**
± SD	**±7.6**	**±8.6**	**±6.6**	**±11.6**	**±6.5**	**±1.3**	**±12.1**	**±15.9**
Sham Group								
s11	6.5	13.8	6.3	5.4	8.2	4.3	-20.0	0.0
s12	10.0	0.0	4.8	0.0	0.0	0.0	-30.3	7.4
s13	8.5	15.4	3.1	0.0	6.7	0.0	-3.9	7.3
s14	22.7	38.1	19.7	25.8	30.8	2.7	-22.7	-29.6
s15	25.0	57.1	22.0	40.0	47.1	4.5	-39.5	-55.4
s17	25.0	7.4	9.4	1.9	3.8	1.0	-7.1	13.0
s18	0.0	0.0	0.0	33.3	104.1	1.8	17.0	31.5
s19	3.6	0.0	2.5	0.0	0.0	0.9	-25.6	-7.9
s20	0.0	0.0	0.0	0.0	0.0	0.0	-23.0	-15.4
M	**11.3**	**14.7**	**7.5**	**11.8**	**22.3**	**1.7**	**-17.2**	**-5.4**
± SD	**±9.6**	**±18.7**	**±7.6**	**±14.7**	**±33.6**	**±1.5**	**±17.6**	**±21.4**

In the sham group, the results were more heterogeneous. Expressive variations were observed in FM total (+22.30%) and relevant reductions in TUG (–17.23%), along with a significant correlation in TUG_cog (Fig. 5). However, a greater number of subjects presented more than one null value (s12, s13, s18, s19, s20), in contrast to the active group. Thus, the comparison between groups shows that, in addition to more homogeneous gains, the active group presented fewer neutral results compared to the sham group, highlighting the greater impact of the intervention on clinical progression [34], [35].

Table 2 presents the percentage variations of spectral energy (Δ%) across six frequency bands (δ, θ, α, β_1, β_2, γ_1) for the subjects in the active and sham groups and the mean ± standard deviation (SD) across subjects. Some subjects presented extreme percentage variations (>600–1000%), driven by very low baseline values. In the Active group, subject s10 showed unusually high increases across several bands, while in the Sham group subjects s15 and s18 presented extreme variations. These outliers inflated group averages and standard deviations, and results should therefore be interpreted with caution.

Table 2 - EEG spectral changes (Δ%) for BA6 (Equation 7) in Active and Sham groups.

EEG spectral changes for BA6 - Active vs Sham groups (Δ%)						
Active group						
	δ	**θ**	**α**	**β_1**	**β_2**	**γ_1**
s1	201.8	147.9	121.7	-5.9	18.2	16.4
s3	-0.2	-4.2	13.0	1.8	0.5	-11.7
s4	19.9	108.0	2.9	43.8	21.4	-17.0
s6	10.5	20.8	0.7	-10.9	4.7	13.1
s7	41.5	21.8	4.2	18.0	-26.7	-4.8
s8	-11.1	-10.5	-6.0	-8.1	-15.5	-16.0
s9	65.3	32.9	27.8	12.6	17.5	50.8
s10	267.9	241.1	309.7	408.4	-21.0	-73.6
M	86.6	69.0	60.5	57.0	0.9	5.7
± SD	± 97.5	± 87.2	± 104.3	± 139.0	± 17.5	± 45.2
Sham group						
s11	13.5	23.6	29.1	55.2	-8.2	-7.6
s12	-76.1	-79.5	-78.6	-58.6	18.3	233.2
s13	-2.3	21.8	14.0	7.1	-8.6	34.7
s14	68.5	46.1	14.1	46.1	2.1	-51.8
s15	836.3	1092.5	658.5	360.9	126.4	-81.5

EEG spectral changes for BA6 - Active vs Sham groups (Δ%)						
Active group						
s17	54.9	61.1	65.3	26.4	1.1	-21.8
s18	694.2	377.7	121.4	-39.2	13.2	-20.8
s19	2.0	13.1	64.3	4.9	12.7	-19.4
s20	13.5	23.6	29.1	55.2	-8.2	-7.6
M	173.5	172.8	98.8	44.3	17.9	8.9
± SD	± 298.0	± 356.0	± 212.0	± 134.0	± 42.0	± 95.0

On average, there was a large increase in band power for the lower frequency bands, namely, δ, θ, α and β_1, after the intervention. This increase was larger for the sham group. On the other hand, the large SDs found for both groups corroborate the high inter-subject variability observed in EEG measurements. Looking at results individually, in the active group subjects s1, s4, s6, and s7 showed partial clinical improvement accompanied by increases in δ, θ, and α, suggesting cortical reorganization compatible with functional evolution [22], [32]. Subject s9 stood out by presenting broad clinical improvement and consistent increases across all bands, representing a profile of global evolution aligned between clinical and EEG findings [33]. Subject s8 presented clinical improvement accompanied by reductions in all bands, characterizing a dissociation between clinical evolution and cortical activity [22]. Subject s10 was an outlier, with relevant clinical gains and extreme increases in δ, θ, α, and β_1, accompanied by reductions in γ_1, suggesting cortical reorganization that was intense and unstable [24]. In the sham group, subjects s14 and s15 showed robust clinical improvement accompanied by increases in several EEG bands. Subject s14 presented a mixed profile, with gains in δ, θ, α, and β_1, along with a marked reduction in γ_1 [22]. Subject s15 stood out as an extreme outlier, with extreme increases in δ, θ, α, β_1, and β_2, accompanied by a reduction in γ_1, suggesting intense cortical reorganization [9]. Subject s18 showed marked clinical improvement in motor metrics and marked increases in δ, θ, and α, while also showing reductions in β_1 and γ_1, configuring a mixed profile with strong enhancement of slow bands [29]. Subject s20 maintained clinical stability with improvement in TUG and TUG_cog, and in the EEG increases were observed in δ, θ, α, and β_1, accompanied by reductions in β_2 and γ_1, suggesting a mixed profile in which functional gains coexist with partial cortical reorganization [2]. To illustrate the relationship between EEG spectral variations and clinical outcomes at the individual level, Fig. 6 presents representative subjects from the active and sham groups, showing side-by-side changes in spectral bands and the eight clinical metrics that reached significance. This visualization highlights the heterogeneity of responses and complements the correlation patterns described in Fig. 5.

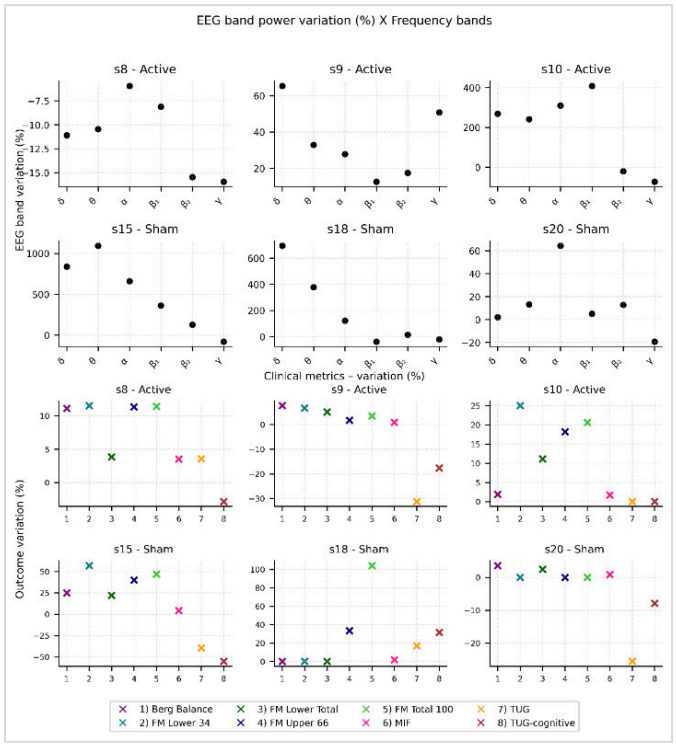


Fig. 5 - Representative subjects from the active and sham groups, illustrating EEG spectral band percentage variations alongside significant clinical metrics percentage variation (Equation 7).

In the active group, s8 showed clinical improvement with reductions across all bands, s9 presented global improvement with consistent increases, and s10 appeared as an outlier with extreme slow band increases. In the sham group, s15 exhibited extreme responses across several bands, s18 showed partial improvement with slow band increases and fast band reductions, and s20 maintained clinical stability with partial reorganization in slow bands.

IV. Discussion

The use of LORETA allowed us to identify patterns of cortical reorganization across sessions, providing a spatial view of the spectral modulation induced by the intervention. In previous work [36], LORETA was shown to reveal cortical generators in specific neural contexts. In the present study, it evidenced increases in slow rhythms (delta, theta, and alpha) as well as increases in fast bands (high beta and low gamma), directly linking the method to the reorganization patterns observed. [12] applied LORETA in electrical stimulation protocols and also observed consistent changes in the beta and gamma bands, reinforcing that this technique is sensitive to detecting cortical reorganization induced by neuromodulation. Another study using LORETA [2] also highlighted the BA6 region, reinforcing its relevance in cortical reorganization processes.

Beyond findings with LORETA, several studies have specifically investigated the association between tDCS and motor imagery. [6] demonstrated that anodal stimulation applied to the motor cortex and cerebellum improves motor imagery classification in EEG, modulating specifically the alpha and beta bands. This result suggests that tDCS enhances the capacity of cortical networks to sustain motor patterns during mental practice, in line with the effects observed in our subjects. [34] conducted a comprehensive mapping of evidence on tDCS in post-stroke motor function, including more than 170 studies, concluding that the technique is safe and potentially effective for

motor recovery, while emphasizing the need for higher-quality clinical trials with larger samples.

The convergence between our findings and previous studies strengthens the interpretation that anodal stimulation not only modulates cortical activity in key regions such as BA6, but also enhances mechanisms already described in motor imagery protocols, expanding the capacity for functional reorganization in post-stroke subjects [2], [6], [7], [12], [36]. The results obtained highlight the potential of tDCS combined with extended reality as a complementary strategy in post-stroke rehabilitation. The identification of early and sustained reorganization suggests that electrical stimulation protocols may promote functional plasticity [1], increasing the capacity for motor recovery in subjects with sequelae.

However, increases in slow bands and in alpha may be considered relevant markers of recovery. In the active group, these patterns were more evenly distributed among subjects, whereas in the sham group they relied mainly on extreme responses from a few individuals.

Compared to the active group, the sham group presented higher means in almost all metrics, except for metric (4) FM upper limb – 66 and metric (8) TUG_cog, and equivalent performance in metric (6) MIF (Table 1). This result, however, should be interpreted with caution, as it may be associated with the higher number of null values observed rather than reflecting clinical superiority. In addition, the greater variability found in the sham group contrasts with the more homogeneous progression verified in the active group, later related to the correlation patterns evidenced in the heatmap (Fig. 4). This difference reinforces the interpretation that tDCS combined with extended reality favors a more consistent and homogeneous reorganization. Nevertheless, some limitations must be acknowledged: the small sample size, the constraints of source localization with a reduced EEG montage (16 electrodes), the absence of direct behavioral measures of motor performance, and the restriction to electrophysiological analyses limit the generalization of the findings. In addition, clinical heterogeneity among subjects may have influenced response patterns. The correlational analyses should also be interpreted with caution, given the small sample sizes and the large number of tested combinations. It should be noted, however, that the statistical routines implemented in LORETA include corrections for multiple voxel-wise comparisons through nonparametric randomization tests, which control for the risk of false positives when testing thousands of voxels simultaneously [10]. The findings showed that the clinical superiority of the active group was not clearly demonstrated, as the sham group presented equal or higher averages in several outcomes. However, these results must be interpreted in light of the very small sample sizes (8 subjects in the active group and 9 in the sham group), which increase the impact of extreme values. In fact, the sham outcomes were accompanied by markedly greater variability, suggesting heterogeneous responses, whereas the active group showed more consistent neurophysiological changes in BA6 spectral activity. This difference reinforces the interpretation that the intervention favors a more stable reorganization pattern, even if clinical superiority was not evident in this sample. Future studies with larger samples, integration of functional measures, and multimodal protocols and the application of robust statistical approaches, including formal outlier inspection and repeated-measures models, may confirm and expand the clinical applicability of the results presented here.

ACKNOWLEDGMENT

This research was supported by São Paulo Research Foundation (FAPESP), Grants 2013/07559-3 (BRAINN), 2022/08057-0, 2023/12422-9, 2024/00998-6 (CRISQuaM), 2025/10963-8; Brazilian National Council for Scientific and Technological Development (CNPq), Grants 306255/2024-3 and 408389/2024-9 (INCT-NeuroComp); Funding Authority for Studies and Projects (FINEP), Grant 2140/22; and Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001. The authors also thank Saulo Feitosa, Márcio Vinícius Silva, Larissa L. Pontes, Jayne Silvestre and Sara R. M. Almeida for their assistance in the data collection.

REFERENCES

[1] T. H. Murphy and D. Corbett, “Plasticity during stroke recovery: from synapse to behaviour,” *Nat. Rev. Neurosci. 2009 1012*, vol. 10, no. 12, pp. 861–872, Nov. 2009, doi: 10.1038/nrn2735.

[2] A. M. Cebolla, E. Palmero-Soler, A. Leroy, and G. Cheron, “EEG spectral generators involved in motor imagery: A swLORETA study,” *Front. Psychol.*, vol. 8, no. DEC, p. 284479, Dec. 2017, doi: 10.3389/FPSYG.2017.02133/FULL.

[3] G. Pfurtscheller and C. Neuper, “Motor imagery and direct brain-computer communication,” *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001, doi: 10.1109/5.939829.

[4] D. Guo *et al.*, “Variation in brain connectivity during motor imagery and motor execution in stroke patients based on electroencephalography,” *Front. Neurosci.*, vol. 18, p. 1330280, Feb. 2024, doi: 10.3389/FNINS.2024.1330280/TEXT.

[5] J. Feng, W. Jia, and Z. Li, “Electroencephalography: A valuable tool for assessing motor impairment and recovery post-stroke,” *J. Neurosci. Methods*, vol. 422, p. 110518, Oct. 2025, doi: 10.1016/J.JNEUMETH.2025.110518.

[6] I. N. Angulo-Sherman, M. Rodríguez-Ugarte, N. Sciacca, E. Iáñez, and J. M. Azorín, “Effect of tDCS stimulation of motor cortex and cerebellum on EEG classification of motor imagery and sensorimotor band power,” *J. Neuroeng. Rehabil.*, vol. 14, no. 1, p. 31, Apr. 2017, doi: 10.1186/S12984-017-0242-1.

[7] Y. Qin, J. Xu, and S. S. M. Ng, “Effects of transcranial direct current stimulation (tDCS) on motor function among people with stroke: evidence mapping,” *Syst. Rev. 2025 141*, vol. 14, no. 1, pp. 60-, Mar. 2025, doi: 10.1186/S13643-025-02795-2.

[8] I. N. Angulo-Sherman, M. Rodríguez-Ugarte, N. Sciacca, E. Iáñez, and J. M. Azorín, “Effect of tDCS stimulation of motor cortex and cerebellum on EEG classification of motor imagery and sensorimotor band power,” *J. NeuroEngineering Rehabil. 2017 141*, vol. 14, no. 1, pp. 31-, Apr. 2017, doi: 10.1186/S12984-017-0242-1.

[9] B. M. Carlos *et al.*, “Effects in Brain Symmetry and Functional Connectivity of tDCS Combined With XR in Chronic Stroke,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 33, pp. 4558–4568, 2025, doi: 10.1109/TNSRE.2025.3630557.

[10] R. D. Pascual-Marqui, “Standardized low resolution brain electromagnetic tomography (sLORETA): technical details Running title: sLORETA,” *Exp. Clin. Pharmacol.*, pp. 5–12, 2002, Accessed: Jan. 11, 2020. [Online]. Available: http://www.keyinst.unizh.ch/loreta.htm

[11] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, “Low

resolution electromagnetic tomography: a new method for localizing electrical activity in the brain," *Int. J. Psychophysiol.*, vol. 18, no. 1, pp. 49–65, 1994, doi: 10.1016/0167-8760(84)90014-X.

[12] T.-W. Lee and G. Tramontano, "Neural Changes to Transcranial Alternating Current Stimulation in the Gamma Range Over the Left Frontoparietal Network: A Preliminary eLORETA EEG Study," *medRxiv*, p. 2025.01.24.25321077, Jan. 2025, doi: 10.1101/2025.01.24.25321077.

[13] A. Fonseca Brandão *et al.*, "GestureCollection for Motor and Cognitive Stimuli: Virtual Reality and e-Health prospects," *J. Heal. Informatics*, vol. 10, no. 1, Feb. 2018, Accessed: Jun. 28, 2026. [Online]. Available: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/544

[14] S. Makeig, T.-P. Jung, A. J. Bell, and T. J. Sejnowski, "Independent Component Analysis of Electroencephalographic Data," *Adv. Neural Inf. Process. Syst.*, vol. 8, 1995.

[15] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website," *Neuroimage*, vol. 198, pp. 181–197, Sep. 2019, doi: 10.1016/J.NEUROIMAGE.2019.05.026.

[16] R. D. Pascual-Marqui *et al.*, "Low resolution brain electromagnetic tomography (LORETA) functional imaging in acute, neuroleptic-naive, first-episode, productive schizophrenia," *Psychiatry Res. - Neuroimaging*, vol. 90, no. 3, pp. 169–179, Jun. 1999, doi: 10.1016/S0925-4927(99)00013-X.

[17] S. Dattola, F. C. Morabito, N. Mammone, and F. La Foresta, "Findings about LORETA Applied to High-Density EEG—A Review," *Electron. 2020, Vol. 9, Page 660*, vol. 9, no. 4, p. 660, Apr. 2020, doi: 10.3390/ELECTRONICS9040660.

[18] P. L. Nunez and R. Srinivasan, "Electric Fields of the Brain: The neurophysics of EEG," *Electr. Fields Brain neurophysics EEG*, pp. 1–611, May 2006, doi: 10.1093/ACPROF:OSO/9780195050387.001.0001.

[19] R. D. Pascual-Marqui, "Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: exact, zero error localization," *Llinas*, Oct. 2007, Accessed: Jun. 13, 2026. [Online]. Available: https://arxiv.org/pdf/0710.3341

[20] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Hum. Brain Mapp.*, vol. 15, no. 1, pp. 1–25, 2002, doi: 10.1002/HBM.1058.

[21] G. E. Thomas, P. H. Westfall, and S. S. Young, "Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment," *J. R. Stat. Soc. Ser. D Stat.*, vol. 43, no. 2, pp. 347–348, Jun. 1994, doi: 10.2307/2348369.

[22] E. Başar, "Brain oscillations in neuropsychiatric disease," *Dialogues Clin. Neurosci.*, vol. 15, no. 3, pp. 291–300, 2013, doi: 10.31887/DCNS.2013.15.3/EBASAR.

[23] STUDENT, "THE PROBABLE ERROR OF A MEAN," *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908, doi: 10.1093/BIOMET/6.1.1.

[24] C. J. Stagg and M. A. Nitsche, "Physiological basis of transcranial direct current stimulation," *Neuroscientist*, vol. 17, no. 1, pp. 37–53, Feb. 2011, doi: 10.1177/1073858410386614.

[25] G. Papitto, A. D. Friederici, and E. Zaccarella, "Distinct neural mechanisms for action access and execution in the human brain: insights from an fMRI study," *Cereb. Cortex (New York, NY)*, vol. 34, no. 4, p. bhae163, Apr. 2024, doi: 10.1093/CERCOR/BHAE163.

[26] H. Jin *et al.*, "Is Pearson's correlation coefficient enough for functional connectivity in fMRI?," *Imaging Neurosci.*, vol. 3, p. 2025, Dec. 2025, doi: 10.1162/IMAG.A.1052.

[27] P. Schober and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.

[28] D. G. Bonett and T. A. Wright, "Sample size requirements for estimating Pearson, Kendall and Spearman correlations," *Psychometrika*, vol. 65, no. 1, pp. 23–28, 2000, doi: 10.1007/BF02294183/METRICS.

[29] A. G. Guggisberg, P. J. Koch, F. C. Hummel, and C. M. Buetefisch, "Brain networks and their relevance for stroke rehabilitation," *Clin. Neurophysiol.*, vol. 130, no. 7, p. 1098, Jul. 2019, doi: 10.1016/J.CLINPH.2019.04.004.

[30] A. J. Shackman, B. W. McMenamin, J. S. Maxwell, L. L. Greischar, and R. J. Davidson, "Identifying Robust and Sensitive Frequency Bands for Interrogating Neural Oscillations," *Neuroimage*, vol. 51, no. 4, p. 1319, Jul. 2010, doi: 10.1016/J.NEUROIMAGE.2010.03.037.

[31] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Springer Top. Signal Process.*, vol. 2, pp. 1–4, 2009, doi: 10.1007/978-3-642-00296-0_5/SAVE-RESEARCH.

[32] W. Klimesch, "Alpha-band oscillations, attention, and controlled access to stored information," *Trends Cogn. Sci.*, vol. 16, no. 12, pp. 606–617, Dec. 2012, doi: 10.1016/j.tics.2012.10.007.

[33] G. Pfurtscheller and F. H. Lopes Da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, 1999, doi: 10.1016/S1388-2457(99)00141-8.

[34] W. Chen *et al.*, "A Novel Deep Learning Model for Motor Imagery Classification in Brain–Computer Interfaces," *Inf.*, vol. 16, no. 7, Jul. 2025, doi: 10.3390/INFO16070582.

[35] Z. Bai, K. N. K. Fong, J. J. Zhang, J. Chan, and K. H. Ting, "Immediate and long-term effects of BCI-based rehabilitation of the upper extremity after stroke: A systematic review and meta-analysis," *J. Neuroeng. Rehabil.*, vol. 17, no. 1, pp. 1–20, Apr. 2020, doi: 10.1186/S12984-020-00686-2/FIGURES/7.

[36] E. M. Santos, R. San-Martin, and F. J. Fraga, "Comparison of LORETA and CSP for Brain-Computer Interface Applications," *18th IEEE Int. Multi-Conference Syst. Signals Devices, SSD 2021*, pp. 817–822, Mar. 2021, doi: 10.1109/SSD52085.2021.9429518.

Theta/Alpha Ratio and Auditory Stimulus Type Modulate Alpha-Power Lateralization During Spatial Attention

Luciana Rita Nicacio
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
0009-0009-7675-5382

Ana Paula de Souza
Departamento de Eletrônica e Biomédica
Centro Federal de Educação Tecnológica de Minas Gerais
Belo Horizonte, Brazil
0000-0002-5835-0754

Leonardo Bonato Felix
Núcleo Interdisciplinar de Análise de Sinais
Universidade Federal de Viçosa
Viçosa, Brazil
0000-0002-6184-2354

Eduardo M. A. M. Mendes
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
0000-0002-3267-3862

***Abstract*—Auditory selective attention enables listeners to focus on a target sound embedded in a complex acoustic mixture. When competing sound sources are spatially separated, auditory spatial attention modulates neural activity, particularly alpha-band oscillations. Alpha power typically increases in the hemisphere ipsilateral to the attended stimulus and decreases in the contralateral hemisphere. This study investigated whether alpha-power lateralization could be elicited by monotonic sinusoidal and musical stimuli and whether this lateralization differed between participant groups stratified according to their theta/alpha ratio (TAR). Twenty-two participants were stratified into lower-TAR (Group I) and higher-TAR (Group II) groups. A significant three-way interaction among Group, Stimulus, and Hemisphere was observed, indicating that the pattern of alpha power across hemispheres differed between groups and stimulus types. Group I exhibited significant alpha-power lateralization during auditory stimulation period, but only under musical stimulation. In contrast, no significant alpha-power lateralization was observed in Group II under either stimulus condition. These findings indicate that alpha-power lateralization during auditory spatial attention depends on the combination of participant TAR profile and stimulus type, with musical stimulation eliciting a more pronounced lateralized response than sinusoidal stimulation. Thus, the alpha-power lateralization induced by musical stimulation may provide a promising control signal for auditory brain-computer interfaces.**

***Index Terms*—auditory spatial attention, alpha-power lateralization, electroencephalography (EEG), theta/alpha ratio, musical stimulation.**

I. Introduction

The ability to selectively attend to a target sound while ignoring competing auditory stimuli, known as the cocktail party effect, has been widely used to investigate the neural mechanisms underlying top-down auditory attention [1], [2]. Experimental paradigms investigating auditory attention are often conducted in sound-attenuated rooms with participants' eyes closed, a setting particularly suitable for auditory brain-computer interfaces (BCIs) [3]. However, maintaining wakefulness under these conditions is challenging, making it essential to monitor participants' level of alertness. Although behavioral measures, such as reaction time or target detection, are commonly used for this purpose [4], [5], requiring a behavioral response may influence the neural activity of interest by engaging additional cognitive processes, including motor preparation and response planning [6]. Furthermore, behavioral-response-free paradigms are essential for the development of auditory BCIs for individuals with complete locked-in syndrome (CLIS), whose communication relies exclusively on neural activity generated during attentional processing [7].

Wakefulness refers to the physiological state in which an individual is conscious and able to appropriately perceive and respond to external stimuli while performing cognitive and motor tasks [8]. Reduced wakefulness can impair sustained attention and consequently degrade the performance of BCI systems [9]. Evidence indicates that signs of drowsiness can emerge rapidly during resting-state EEG recordings. For example, [10] reported that approximately one-third of participants exhibited reduced alertness after only three minutes of eyes-closed recording. Despite extensive evidence showing that attentional processes are strongly influenced by the level of alertness [8], [11], [12], fluctuations in vigilance are still frequently overlooked in cognitive neuroscience studies. Consequently, EEG trials acquired during periods of reduced alertness are often included in analyses of attention and other cognitive functions [13]. Therefore, identifying episodes of

This study was financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brazil (Finance Code 001).

reduced wakefulness before subsequent analyses is essential to exclude unsuitable trials and improve the reliability of the findings.

In this context, the theta/alpha ratio (TAR) has been proposed as a neurophysiological indicator associated with drowsiness [4], [5], [14], [15], particularly when assessed over the occipital cortex [16]. Theta (4–8 Hz) and alpha (8–13 Hz) oscillations can be readily measured at the scalp using electroencephalography (EEG), and EEG recordings acquired under eyes-closed conditions in awake individuals generally exhibit prominent alpha activity over the occipital cortex [17]. Previous studies have shown that during the transition from wakefulness to sleep, or sleep onset period (SOP), theta activity generally becomes more prominent, whereas alpha activity decreases, resulting in higher TAR values [5]. Thus, higher TAR values may reflect changes in the EEG spectral profile associated with the SOP.

At the same time, alpha activity is closely involved in the allocation of auditory spatial attention, making its lateralization particularly relevant to selective listening. Previous studies have consistently reported alpha-power lateralization during listening tasks, with alpha power increasing in the cortical hemisphere ipsilateral to the attended auditory hemifield and decreasing in the contralateral hemisphere [18]–[20]. During selective auditory spatial attention, two distinct mechanisms contribute to attentional selection: target selection and distractor suppression. Reference [21] demonstrated that both mechanisms are associated with alpha-power lateralization and that their neural signatures are dissociable and uncorrelated, suggesting that distractor suppression can occur independently of target selection. In addition, previous studies have shown that alpha-power lateralization can emerge even before the onset of auditory stimulation, during the interval between the presentation of the spatial cue and stimulus onset [18], [20]–[24]. Together, these findings support alpha-band lateralization as a reliable neural signature of both the anticipatory deployment and sustained maintenance of auditory spatial attention.

The magnitude of alpha lateralization, however, depends on the characteristics of the auditory stimuli. When sounds are presented through headphones without natural spatial cues, such as interaural time differences (ITDs) and interaural level differences (ILDs), alpha lateralization tends to be substantially weaker [23], [24]. Moreover, most previous studies investigating alpha lateralization have employed speech-based stimuli, such as spoken syllables or numbers [18], [20], [22], [23], whereas alternative classes of auditory stimuli remain comparatively underexplored.

Among the available alternatives, musical stimulation has the potential to provide a more comfortable and engaging experimental experience. Compared with the conventional sinusoidal amplitude-modulated (AM) stimuli commonly used to evoke 40-Hz auditory steady-state responses (ASSRs) in auditory BCI research [19], [25], musical stimuli may reduce the physical and psychological burden on participants while preserving the inherently pleasant experience of listening to music [26], [27]. Besides improving participant acceptance, musical stimulation may also help maintain alertness throughout the recording session, as its melodic structure allows participants to follow the auditory stream and remain engaged with the task.

Given the association between drowsiness and changes in theta and alpha oscillations, this study uses the theta/alpha ratio to stratify participants into higher- and lower-TAR groups. It then investigates whether dichotic stimulation with both musical and monotonic (single-carrier sinusoidal) stimuli induces alpha-power lateralization in each group. By assessing alpha-power lateralization across these conditions, this study provides further insight into the robustness of alpha-power lateralization as a neural marker and its potential application in auditory BCIs [19].

II. Materials and Methods

A. Participants

A total of 22 volunteers participated in this study (9 females, mean age = 29.3 years, SD = 6.4 years). All participants self-reported normal hearing. EEG data were acquired in a soundproof cabin while participants were seated comfortably and instructed to keep their eyes closed throughout the recordings. The study protocol received approval from the Research Ethics Committee of the Universidade Federal de Viçosa (CEP/UFV: 3.773.292). All participants gave informed consent before participation.

B. Experimental Paradigm

The experiment started with the completion of a questionnaire containing binary questions, such as “At this moment, are you wearing a watch?”, “Do you consider yourself tall?”, and “Did you drink coffee today?”. Although each question required a binary decision, participants could choose among three response options: “Yes”, “No”, or “?”, where the question mark indicated either uncertainty or a preference not to answer. Questions marked with “?” were not considered in the subsequent experimental procedures. After completing the questionnaire, participants proceeded with the experimental protocol, as shown in Fig. 1. For each participant, the questions were randomly selected from the questionnaire and consisted of five questions that had been answered with “Yes” and five that had been answered with “No”. Before the presentation of the questions used in the attention task, a 34-s calibration trial was recorded to obtain baseline EEG activity. During this phase, the two auditory stimuli were simultaneously presented through earphones, and participants were instructed not to attend to either stimulus. To maintain a neutral attentional state, participants were asked to mentally sing the Brazilian National Anthem throughout the calibration period.

EEG acquisition was performed on a trial-by-trial basis rather than continuously across the entire experiment. For each trial, recording started at least 1 s before the onset of the auditory stimulation. Each participant completed a total of eleven EEG recordings: one 34 s calibration trial without selective auditory attention and ten attention trials, each lasting

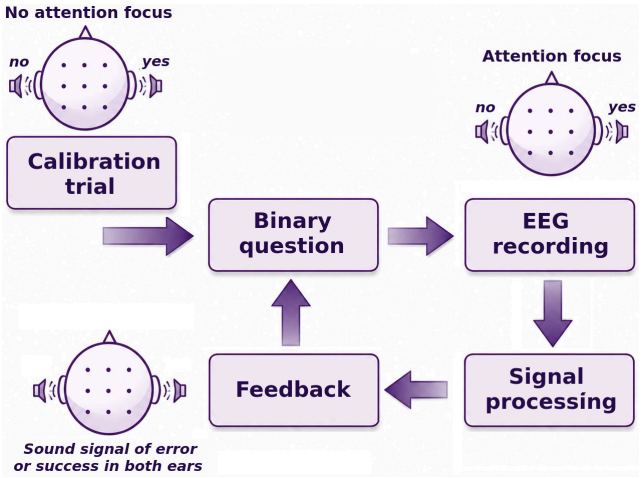


Fig. 1. Experimental paradigm. The experiment began with a 34 s calibration recording, during which the auditory stimuli were delivered through insert earphones. During this period, participants were instructed not to attend to either stimulus. After the calibration trial, a binary question requiring a "Yes" or "No" response was presented, immediately followed by the simultaneous dichotic delivery of two auditory stimuli. Depending on the experimental condition, the stimuli consisted of either sinusoidal tones or musical instrument sounds, both of which were amplitude-modulated. The stimulus associated with the "No" response was always presented to the left ear, whereas the "Yes" stimulus was presented to the right ear. EEG activity was recorded throughout the 5.1 s stimulation interval while participants selectively directed their attention toward the stimulus corresponding to their intended answer. At the end of each trial, the recorded EEG was processed to identify the attended stimulus, and auditory feedback was provided to inform participants whether the decoded response matched their intended answer. A 5 s resting interval was then introduced before the onset of the next question.

5.1 s, in which participants selectively attended to the stimulus corresponding to their intended response.

The experimental protocol was originally developed to investigate the modulation of ASSRs by auditory spatial attention for the implementation of an online binary communication system. However, because the proposed system achieved limited classification performance, the recorded data were subsequently reanalyzed offline to investigate whether participants' reduced wakefulness could have contributed to the lower decoding accuracy. Based on this analysis, the scope of the study was expanded to evaluate the effects of auditory spatial attention on alpha-band activity, with particular emphasis on alpha-power lateralization induced by spatially separated auditory stimuli. Consequently, the feedback provided to participants during the experiment was based on the comparison between their questionnaire responses and the EEG classifications generated by the ASSR-based modular classifier proposed by [25]. Accordingly, all analyses reported in the present study were conducted as a post hoc analysis of the previously acquired EEG dataset. To ensure transparency and reproducibility, the raw EEG recordings are publicly available on GitHub at https://github.com/Luciana-Nicacio/eeg_dataset.

C. Auditory Stimulation

Binary response options were encoded using amplitude-modulated (AM) auditory stimuli. The experiment consisted of two blocks of 10 questions, each employing a different type of carrier signal. In the first block, pure sinusoidal tones served as carriers, with frequencies of 500 Hz and 4 kHz corresponding to the "No" and "Yes" responses, respectively. In the second block, musical instrument sounds were used as carriers, consisting of a piano melody for "No" and a harmonica melody for "Yes". These audio files were obtained from the Internet Archive (https://archive.org/) and, like the sinusoidal stimuli, were sampled at 44.1 kHz. To avoid possible order effects, the presentation sequence of the two stimulus blocks was counterbalanced across participants.

The carrier signals were amplitude-modulated at 31.13 Hz for the "No" stimulus and 39.36 Hz for the "Yes" stimulus. These modulation frequencies were adopted according to the protocol proposed by [25]. The sound intensity of all stimuli was calibrated to 70 dB SPL (sound pressure level), ensuring comfortable listening while remaining well above the ambient noise level. Audio files containing the amplitude-modulated stimuli are available on GitHub at https://github.com/Luciana-Nicacio/Stimulation-audio-files.

D. EEG Recording and Processing

Signal acquisition was performed using a BrainNet BNT-36 amplifier (EMSA Equipamentos Médicos, Brazil) and 22 Ag/AgCl scalp electrodes positioned according to the international 10–20 system. The ground electrode was placed at Fpz, and Cz served as the reference electrode. Data were acquired with analog band-pass filtering between 1 and 100 Hz and digitized at a nominal sampling rate of 600 Hz (601.5 Hz according to the manufacturer's specifications). The amplifier's hardware notch filter was intentionally disabled, and power-line interference at 60 Hz and its harmonics was removed during offline processing using a digital infinite impulse response (IIR) notch filter.

Subsequent offline analyses were performed in Python (version 3.10.12) using the MNE-Python toolbox (version 1.9.0) [28].

1) TAR-Based Grouping: Participants were stratified into two groups according to the occurrence of higher TAR values across the attention trials, based on the hypothesis that higher TAR values may be associated with lower levels of wakefulness. To extract spectral features from the EEG signals, the theta (4–8 Hz) and alpha (8–13 Hz) rhythms were first isolated using digital band-pass filtering with the *mne.filter.filter_data* function from the MNE-Python package and its default settings. Subsequently, the amplitude envelope of each frequency band was derived from the magnitude of the analytic signal obtained using the Hilbert transform. Instantaneous power in the three occipital channels (O_1, O_2, and O_z) was estimated as the squared amplitude envelope, $P(t) = A(t)^2$, and mean power was calculated over time for each channel. The TAR was then computed for each channel as the ratio of theta to alpha mean power, and the final TAR was obtained by averaging the ratios across the three occipital channels. This procedure was applied to the 20 attention trials of all 22 participants, comprising 10 trials under

musical stimulation and 10 under sinusoidal stimulation. Four TAR thresholds were considered separately (1.00, 0.75, 0.60, and 0.50), resulting in four distinct participant groupings. For each threshold, participants who exhibited a TAR greater than the respective threshold in more than 5 of the 20 attention trials ($> 25\%$) were assigned to Group II, hereafter referred to as the higher-TAR group. The remaining participants were assigned to Group I, referred to as the lower-TAR group. For each TAR threshold, subsequent analyses examined the main effects and interactions of Group, Stimulation, and Hemisphere using the alpha-power attentional modulation index (AMI_α), as described below.

2) Alpha-Power Lateralization: At this stage, the EEG recordings were re-referenced to the average of the two mastoid channels. The alpha-band-filtered EEG signals were then normalized on a trial-by-trial basis using z-scores computed from the mean and standard deviation calculated across all EEG channels. Subsequently, the envelope of the normalized EEG signal was extracted for two time windows: the pre-stimulus interval (-0.5 to 0 s) and the auditory stimulation interval (0 to 5.1 s). Alpha power in each interval was estimated as the mean squared amplitude envelope within the corresponding time window. The alpha power values were then averaged across the five "No" trials and the five "Yes" trials for each participant. Finally, the AMI_α was computed for each participant and channel following the approach proposed by [18]. This metric quantified the hemispheric lateralization of alpha power induced by auditory spatial attention for each analysis period and was calculated as follows:

$$AMI_\alpha = \frac{X_{att_{No}} - X_{att_{Yes}}}{X_{att_{No}} + X_{att_{Yes}}} \tag{1}$$

where X represents the mean alpha power measured while selectively attending to either the "No" or the "Yes" stimulus. Accordingly, positive AMI_α values indicate higher alpha power during attention to the "No" stimulus, whereas negative values indicate higher alpha power during attention to the "Yes" stimulus.

To quantify the degree of alpha-power lateralization, the mean AMI_α was calculated across all electrodes within the left hemisphere (LH) and right hemisphere (RH).

3) Statistical Analysis: To evaluate the effects of Group, Stimulus, and Hemisphere on the AMI_α, separate linear mixed-effects models (MixedLM) were fitted for the two time windows. Statistical analyses were performed using the *statsmodels* package in Python. In both models, Group (between-subjects factor), Stimulus (within-subjects factor: musical vs. sinusoidal), and Hemisphere (within-subjects factor: left vs. right), together with all corresponding two-way and three-way interactions, were included as fixed effects. Participant ID was included as a random intercept to account for the repeated-measures design. Residual normality (Shapiro–Wilk test) and homoscedasticity (Levene's test) were assessed. When these assumptions were satisfied, the AMI_α values were analyzed on their original continuous scale. Sphericity was considered automatically satisfied because all within-subjects factors had only two levels ($\varepsilon = 1.0$). When the normality or homoscedasticity assumptions were violated, a robust non-parametric rank-transformation procedure was applied to the AMI_α values before model fitting. The fixed effects were assessed using Wald χ^2 tests at a 5% significance level. When a significant three-way interaction among Group, Stimulus, and Hemisphere was identified, post-hoc analyses were performed to further characterize the interaction by comparing mean AMI_α between the left and right hemispheres within each group and stimulation type. The normality of the differences in mean AMI_α between hemispheres was first assessed using the Shapiro–Wilk test. If the distribution of these differences did not deviate significantly from normality, paired *t*-tests were used to compare mean AMI_α between hemispheres. Otherwise, the Wilcoxon signed-rank test was applied to the paired observations.

III. Results

A. Participant Grouping

Descriptive statistics for TAR across the musical and sinusoidal stimulation conditions, along with the percentage of trials exceeding four distinct TAR thresholds (1.00, 0.75, 0.60, and 0.50), are summarized in Table I. The number of participants assigned to each group across the four TAR-based groupings is presented in Table II.

B. Alpha-Power Lateralization: Main Effects and Interactions

The effects of the main factors and their interactions were evaluated separately for each EEG time window: the pre-stimulus interval (-0.5 to 0 s) and the auditory stimulation interval (0 to 5.1 s), as described in the following subsections.

1) Before Stimulation Onset: For this period, the assumptions of normality and homoscedasticity were satisfied for all groupings. The linear mixed-effects model (MixedLM, Wald test) revealed a significant main effect of Stimulus in groupings 1 ($\chi^2 = 7.20$, $p = 0.0073$), 2 ($\chi^2 = 10.06$, $p = 0.0015$), and 3 ($\chi^2 = 6.96$, $p = 0.0083$), but not in grouping 4 ($\chi^2 = 3.28$, $p = 0.070$). No significant main effects of Group or Hemisphere were found at any threshold ($p > 0.05$). Likewise, none of the two-way or three-way interactions reached statistical significance.

2) During Auditory Stimulation: The Shapiro-Wilk test indicated a violation of the normality assumption ($p < 0.0001$), requiring a robust rank-transformed linear mixed-effects model for all four grouping configurations. A significant main effect of Hemisphere was observed in groupings 2 ($\chi^2 = 4.28$, $p = 0.039$) and 3 ($\chi^2 = 4.62$, $p = 0.032$), whereas grouping 1 showed a marginal effect ($\chi^2 = 3.63$, $p = 0.057$). A significant two-way Group $\times$ Hemisphere interaction was consistently observed in groupings 1 ($\chi^2 = 3.98$, $p = 0.046$), 2 ($\chi^2 = 4.02$, $p = 0.045$), and 3 ($\chi^2 = 3.91$, $p = 0.048$). In addition, significant Stimulus $\times$ Hemisphere interactions were observed in groupings 2 ($\chi^2 = 3.86$, $p = 0.049$) and 3 ($\chi^2 = 3.93$, $p = 0.047$), together with significant three-way Group $\times$ Stimulus $\times$ Hemisphere interactions ($\chi^2 = 4.55$, $p = 0.033$ and $\chi^2 = 3.89$, $p = 0.048$, respectively). No

TABLE I
TAR DESCRIPTIVE STATISTICS AND PERCENTAGE OF TRIALS EXCEEDING EACH OF FOUR TAR THRESHOLDS ACROSS THE 20 ATTENTION TRIALS PER PARTICIPANT

Participant	Musical (Mean ± SD)	Sinusoidal (Mean ± SD)	TAR > 1.00 (%)	TAR > 0.75 (%)	TAR > 0.60 (%)	TAR > 0.50 (%)
ID01	0.50 ± 0.09	0.40 ± 0.10	0	0	5	30
ID02	0.11 ± 0.03	0.12 ± 0.02	0	0	0	0
ID03	0.12 ± 0.03	0.37 ± 0.64	5	5	5	5
ID04	0.18 ± 0.07	0.36 ± 0.33	5	5	5	5
ID05	0.70 ± 0.19	0.60 ± 0.18	0	35	50	80
ID06	0.34 ± 0.09	0.36 ± 0.05	0	0	0	0
ID07	0.76 ± 0.48	0.85 ± 0.44	40	55	55	55
ID08	0.51 ± 0.16	0.60 ± 0.17	0	20	35	50
ID09	0.25 ± 0.06	0.21 ± 0.03	0	0	0	0
ID10	0.71 ± 0.31	0.79 ± 0.15	10	45	70	90
ID11	1.00 ± 0.22	1.05 ± 0.33	60	85	90	95
ID12	0.57 ± 0.23	0.41 ± 0.07	5	5	10	35
ID13	0.26 ± 0.04	0.38 ± 0.17	0	0	5	10
ID14	0.31 ± 0.17	0.49 ± 0.19	0	10	15	20
ID15	0.70 ± 0.83	1.17 ± 0.69	55	55	60	60
ID16	1.59 ± 0.82	0.82 ± 0.32	35	75	85	90
ID17	0.62 ± 0.09	0.51 ± 0.09	0	0	40	70
ID18	0.67 ± 0.15	0.84 ± 0.19	20	40	70	95
ID19	1.44 ± 0.39	1.12 ± 0.30	85	95	100	100
ID20	1.49 ± 0.32	1.33 ± 0.17	100	100	100	100
ID21	0.09 ± 0.02	0.13 ± 0.03	0	0	0	0
ID22	0.59 ± 0.29	0.50 ± 0.41	10	25	40	50

TABLE II
NUMBER OF PARTICIPANTS IN EACH GROUP ACCORDING TO THE ADOPTED TAR THRESHOLD

Grouping	Group I	Group II
1: TAR > 1.00	16	6
2: TAR > 0.75	13	9
3: TAR > 0.60	10	12
4: TAR > 0.50	8	14

main effects or interactions reached statistical significance in grouping 4.

The three-way interaction patterns for the grouping 3 (TAR > 0.60) are illustrated in Fig. 2. The interaction plots revealed distinct patterns of hemispheric alpha-power modulation between groups. In Group I, musical stimulation was associated with higher AMI_α ranks over the left hemisphere, whereas sinusoidal stimulation showed higher ranks over the right hemisphere. In Group II, this pattern was reversed, with higher AMI_α ranks for musical stimulation over the right hemisphere and for sinusoidal stimulation over the left hemisphere. Thus, the relationship between stimulus type and hemispheric alpha-power lateralization differed between the two groups.

Post hoc pairwise comparisons using paired t-tests (with normality confirmed by Shapiro–Wilk tests, $p > 0.12$ across conditions) further clarified the factors underlying the three-way interaction during stimulus presentation. For grouping 3, Group I showed a statistically significant interhemispheric difference in alpha-power lateralization during musical stimulation, with higher values over the left hemisphere than over the right hemisphere (Fig. 3A; $p = 0.035$). In contrast, no significant interhemispheric difference was observed in Group I under sinusoidal stimulation (Fig. 3B; $p = 0.43$). Group II showed no significant interhemispheric differences under either musical (Fig. 3C; $p = 0.42$) or sinusoidal (Fig. 3D; $p = 0.85$) stimulation. Notably, the same pattern was also observed for grouping 2, with a significant interhemispheric difference in Group I during musical stimulation ($p = 0.049$).

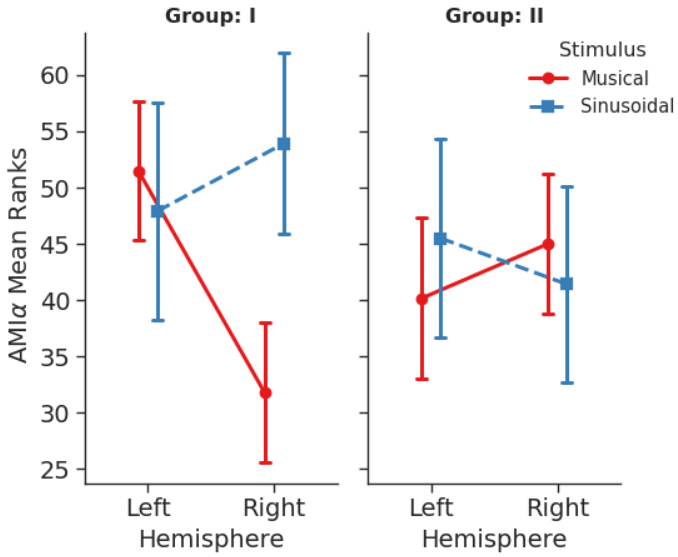


Fig. 2. Interaction profile of alpha power asymmetry (AMI_α mean ranks) during the stimulus presentation interval (0–5.1 s) for grouping 3. Point plots depict mean ranks across hemispheres (Left vs. Right) for musical (red solid line, circles) and sinusoidal (blue dashed line, squares) stimuli, faceted by group profile (Group I, $n = 10$; Group II, $n = 12$). Error bars represent standard errors of the mean (±SE).

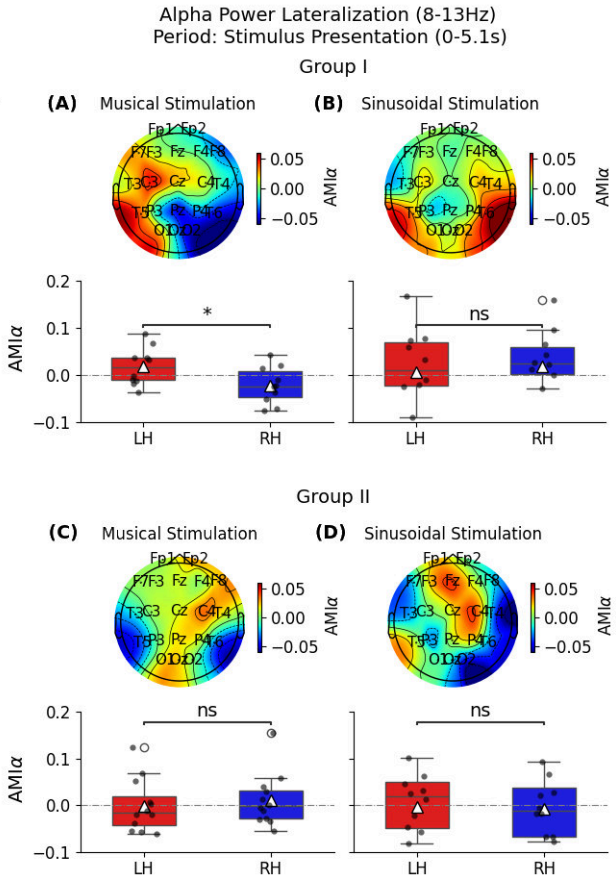


Fig. 3. Topographic maps and boxplots of alpha-power lateralization AMI_α during the stimulus presentation interval (0–5.1 s) for the grouping 3 under musical and sinusoidal stimulation. Panels (A) and (B) correspond to Group I, whereas panels (C) and (D) correspond to Group II. The topographic maps display the mean AMI_α across participants. The boxplots summarize the mean AMI_α for the left (LH) and right (RH) hemispheres. White triangles indicate the mean value, gray circles represent individual participants, and the horizontal dashed line denotes $AMI_\alpha = 0$. Statistical significance between hemispheres is indicated above each boxplot ($^*p < 0.05$, $^{**}p < 0.01$, $^{n.s.}p > 0.05$).

IV. Discussion

The present study investigated alpha-power lateralization associated with auditory spatial attention across individuals stratified according to distinct TAR thresholds during musical and sinusoidal auditory stimulation. Participants were stratified according to their TAR levels into Group I (lower-TAR) and Group II (higher-TAR) using four different grouping criteria. Although the literature does not establish a universal TAR threshold for reliably distinguishing an individual's level of alertness, two of the four thresholds used for participant stratification yielded a significant three-way interaction among Group, Stimulus, and Hemisphere during the stimulation period.

Post hoc analyses further showed that a significant alpha-power lateralization effect was observed only in Group I during musical stimulation. Alpha-power lateralization associated with auditory spatial attention exhibited the expected opposite pattern of modulation, with positive AMI_α values over the left hemisphere and negative values over the right hemisphere. Consistent with previous studies, alpha-power lateralization was primarily evident over the posterior cortex [18]–[20]. In contrast, no significant interhemispheric differences were observed in Group II under either stimulation condition, suggesting that alpha-power lateralization was less evident among participants with higher TAR values. This finding is in line with previous studies reporting a strong association between higher TAR values and reduced wakefulness or increased sleep propensity [4], [5], [15], which may have affected the participants' ability to sustain attention during the experiment.

These findings highlight the importance of assessing wakefulness in EEG experiments involving prolonged eye closure and the absence of behavioral measures of alertness. Failure to account for differences in wakefulness may affect the interpretation of neural activity by obscuring attention-related neural signatures. Therefore, assessing wakefulness before including EEG recordings in subsequent analyses may help reduce variability associated with vigilance and improve the reliability of attention-related EEG measures.

The absence of significant alpha-power lateralization during sinusoidal stimulation may reflect the limited acoustic information available to sustain auditory spatial attention. Unlike the musical stimuli, the sinusoidal stimuli consisted of amplitude-modulated pure tones in which the sinusoidal carrier had a constant amplitude, thus providing little acoustic information to sustain auditory spatial attention. Furthermore, because both stimulus types were delivered through headphones in a dichotic listening paradigm, they did not provide the natural spatial cues generated by acoustic sources surrounding the listener, including interaural time differences (ITDs), interaural level differences (ILDs), and pinna-related spectral cues. These cues are normally available in everyday listening environments, where a single sound source reaches both ears, and contribute to the cocktail party effect. As discussed by [23], these spatial cues are important for generating realistic spatial auditory percepts and engaging cortical networks involved in auditory spatial attention. Their findings showed that attentional modulation of parietal alpha activity was strongest when natural spatial cues derived from individualized head-related transfer functions (HRTFs) were available. In contrast, diotic stimulation based on isolated spatial cues, such as ITDs and ILDs, produced weaker and less consistent alpha-power lateralization, suggesting that these cues alone may provide less effective engagement of the neural mechanisms underlying auditory spatial attention.

In this context, the findings reported by [19] corroborate those of [23]. Although their study also employed sinusoidally amplitude-modulated stimuli, they observed alpha-power lateralization over parieto-occipital regions. However, unlike the present study, their stimuli were presented through loudspeakers positioned at azimuths of $\pm 45°$. Thus, their experimental setup provided spatial information associated with the physical location of the sound sources, as well as acoustic cues generated by the head and pinnae. The

observation of alpha lateralization under these conditions suggests that sinusoidal AM stimuli can elicit attentional lateralization when presented with richer spatial information. Together with the findings of [23], these results suggest that the absence of significant alpha-power lateralization during sinusoidal stimulation in the present study may be related not only to the limited spectrotemporal structure of the stimuli but also to the restricted spatial information available during dichotic headphone presentation.

Based on the hypothesis that the dynamic acoustic features of musical stimuli may facilitate the sustained allocation of auditory attention during dichotic stimulation, our findings can be interpreted in the context of previous evidence showing that temporally structured auditory signals enhance attentional modulation. This interpretation is supported by the studies of [18], [22], which successfully demonstrated alpha-power lateralization during dichotic stimulation without individualized spatial cues. In these studies, the target stimuli consisted of spoken numbers, whose naturally varying temporal envelopes provided low-frequency temporal structures synchronized with speech dynamics. Together, these findings indicate that, beyond spatial information, temporally structured auditory features contribute to the engagement of top-down attentional control mechanisms.

Reference [24] investigated the influence of nonspatial acoustic features on auditory spatial attention using diotic stimulation with simulated spatial cues generated through interaural time differences. Three temporally regular auditory stimuli were designed to be perceived as originating from different spatial locations (left, center, and right) by applying ITDs of -100, 0, and +100 μs, respectively. Their findings showed that parietal alpha lateralization reflected the spatial cue associated with the target stimulus before stimulus onset, indicating an initial allocation of attention toward the expected auditory location. However, after stimulus presentation began, alpha lateralization was no longer observed when the target and distractor stimuli had large pitch separation, whereas it remained present when pitch differences were smaller. The authors suggested that large pitch differences between target and distractor stimuli may reduce the reliance on sustained spatial attention during auditory selection. In contrast, [23] demonstrated that parieto-occipital alpha lateralization varied monotonically as the direction of attention shifted from far left (-600 μs) to far right (+600 μs). Therefore, the lower spatial salience of the cues employed by [24] may have limited the recruitment of spatial attention mechanisms and contributed to the attenuation of sustained alpha lateralization during stimulation. These findings suggest that, in paradigms based on diotic stimulation where pitch differences between the target and distractor stimuli are larger, greater ITDs may be required to robustly engage the neural mechanisms underlying sustained auditory spatial attention and induce alpha-power lateralization.

Finally, an interesting finding of this study was the significant main effect of Stimulus observed during the pre-stimulus period for groupings 1, 2, and 3. Because no auditory stimulation was presented during this interval, this effect cannot be attributed to the immediate processing of the auditory stimuli themselves. Instead, it may reflect anticipatory or preparatory neural activity associated with knowledge of the upcoming stimulus type. Before the attention trials, participants had already completed the calibration trial and were therefore familiar with the stimulus type that would be presented in each subsequent trial. This prior knowledge may have allowed participants to anticipate the upcoming auditory stimulation and engage stimulus-specific preparatory processes before stimulus onset.

A. Limitations and Future Work

This study has some limitations that should be addressed in future investigations. Participant grouping relied on the same TAR thresholds across all participants, whereas future studies should consider individualized thresholds established from a baseline recording obtained under confirmed wakefulness conditions to reduce the potential for misclassification. Because eye closure was not objectively monitored, higher TAR values observed in the present study may have been influenced by periods of eye opening, which are known to reduce occipital alpha-power amplitude. Therefore, elevated TAR values cannot be attributed exclusively to possible reductions in wakefulness or drowsiness. Moreover, approximately 10% of healthy individuals do not exhibit prominent alpha rhythms even under eyes-closed conditions [17]. Therefore, future studies should assess alpha activity under both eyes-open and eyes-closed conditions before the experiment to identify individuals who do not exhibit the typical increase in occipital alpha-power amplitude associated with eye closure. Additionally, experiments without ASSR-based feedback should be conducted to determine whether alpha lateralization remains stable in the absence of potential influences from classifier outcomes. The use of more realistic spatial stimulation based on head-related transfer functions (HRTFs) may also enhance auditory spatial attention and elicit stronger alpha-power lateralization. Furthermore, although the present study evaluated the conventional alpha band (8–13 Hz), future analyses should investigate individualized alpha sub-bands, as [20] demonstrated that attentional modulation was more accurately captured by analyzing a narrow frequency band centered on each participant's individual parietal alpha peak rather than the entire alpha range. In addition, previous studies investigating the transition from wakefulness to sleep using the TAR have employed narrower frequency bands within the theta and alpha ranges, suggesting that more narrowly defined frequency bands may provide greater sensitivity to changes in wakefulness [4], [14]. Considering the dynamic temporal structure of musical stimuli, future studies should also examine whether alpha lateralization synchronizes with the envelope of the attended stimulus, as previously observed for speech-based attention [18], [22]. Finally, analyses using shorter temporal windows and restricting the analysis to centro-parieto-occipital regions may help identify the temporal periods and cortical areas associated with stronger attentional alpha-power modulation.

V. Conclusion

This study demonstrated that TAR-based participant stratification and stimulus type were associated with differences in alpha-power lateralization during auditory spatial attention. Significant lateralization was observed only in Group I (lower-TAR) during musical stimulation, suggesting that this type of stimulation may have contributed to sustained attentional engagement throughout the experiment. Future studies combining temporally dynamic auditory stimuli, including musical stimuli, with richer spatial cues provided by individualized HRTFs may further enhance alpha-power lateralization and strengthen its potential as a neural marker of auditory spatial attention. Building on this approach, a classifier for decoding the direction of attention based on alpha-power lateralization could be implemented and integrated with ASSR-based classification to improve the performance and reliability of auditory BCIs. Together, these complementary approaches could support the development of hybrid auditory BCIs for binary communication in patients with CLIS.

Acknowledgment

The authors acknowledge the financial support provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brazil (Finance Code 001).

References

[1] A. Bednar and E. C. Lalor, "Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG," *Neuroimage*, vol. 205, p. 116283, 2020.

[2] S. Cai, H. Zhu, T. Schultz, and H. Li, "EEG-based Auditory Attention Detection in Cocktail Party Environment," *APSIPA Trans. Signal Inf. Process.*, vol. 12, 2023.

[3] D. Hübner, A. Schall, N. Prange, and M. Tangermann, "Eyes-closed increases the usability of brain-computer interfaces based on auditory event-related potentials," *Front. Hum. Neurosci.*, vol. 12, p. 391, 2018.

[4] I. M. Comsa, T. A. Bekinschtein, and S. Chennu, "Transient topographical dynamics of the electroencephalogram predict brain connectivity and behavioural responsiveness during drowsiness," *Brain Topogr.*, vol. 32, pp. 315–331, 2019.

[5] M. Strauss, J. D. Sitt, L. Naccache, and F. Raimondo, "Predicting the loss of responsiveness when falling asleep in humans," *NeuroImage*, vol. 251, p. 119003, 2022.

[6] D. J. Gale, C. N. Areshenkoff, C. Honda, I. S. Johnsrude, J. R. Flanagan, and J. P. Gallivan, "Motor planning modulates neural activity patterns in early human auditory cortex," *Cereb. Cortex*, vol. 31(6), pp. 2952–2967, 2021.

[7] P. Séguin, E. Maby, M. Fouillen, A. Otman, J. Luauté, P. Giraux, D. Morlet, and J. Mattout, "The challenge of controlling an auditory BCI in the case of severe motor disability," *J. Neuroeng. Rehabil.*, vol. 21(1), p. 9, 2024.

[8] V. P. B. Venkat and S. Chinara, "Automatic classification methods for detecting drowsiness using wavelet packet transform extracted time-domain features from single-channel EEG signal," *J. Neurosci. Methods*, vol. 347, p. 108927, 2021.

[9] B. Oken, T. Memmott, B. Eddy, J. Wiedrick, and M. Fried-Oken, "Vigilance state fluctuations and performance using brain-computer interface for communication," *Brain-Comput. Interfaces*, vol. 5(4), pp. 146–156, 2018.

[10] E. Tagliazucchi and H. Laufs, "Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep," *Neuron*, vol. 82(3), pp. 695–708, 2014.

[11] E. Wiggins, M. Mottarella, K. Good, S. Eggleston, and C. Stevens, "24-hour sleep deprivation impairs early attentional modulation of neural processing: An event-related brain potential study," *Neurosci. Lett.*, vol. 677, pp. 32–36, 2018.

[12] C. Lacaux, M. Strauss, T. A. Bekinschtein, and D. Oudiette, "Embracing sleep-onset complexity," *Trends Neurosci.*, vol. 47(4), pp. 273–288, 2024.

[13] S. R. Jagannathan, A. Ezquerro-Nassar, B. Jachs, O. V. Pustovaya, C. A. Bareham, and T. A. Bekinschtein, "Tracking wakefulness as it fades: Micro-measures of alertness," *NeuroImage*, vol. 176, pp. 138–151, 2018.

[14] C. A. Bareham, T. Manly, O. V. Pustovaya, S. K. Scott, and T. A. Bekinschtein, "Losing the left side of the world: rightward shift in human spatial attention with sleep onset," *Sci. Rep.*, vol. 4, p. 5092, 2014.

[15] T. Andrillon, J. Taillard, and M. Strauss, "Sleepiness and the transition from wakefulness to sleep," *Neurophysiol. Clin.*, vol. 54(2), p. 102954, 2024.

[16] K. Susmáková and A. Krakovská, "Discrimination ability of individual measures used in sleep stages classification," *Artif Intell Med*, vol. 44(3), pp. 261–77, 2008.

[17] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications," *Am. Acad. Sleep Med.*, 2012.

[18] M. Wöstmann, B. Herrmann, B. Maess, and J. Obleser, "Spatiotemporal dynamics of auditory attention synchronize with speech," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113(14), p. 3873–3878, 2016.

[19] M. Spüler and S. Kurek, "Alpha-band lateralization during auditory selective attention for brain-computer interface control," *Brain-Comput. Interfaces*, vol. 5(1), pp. 23–29, 2018.

[20] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, vol. 207, p. 116360, 2020.

[21] M. Wöstmann, M. Alavash, and J. Obleser, "Alpha oscillations in the human brain implement distractor suppression independent of target selection," *J. Neurosci.*, vol. 39(49), pp. 9797–9805, 2019.

[22] M. Wöstmann, B. Maess, and J. Obleser, "Orienting auditory attention in time: Lateralized alpha power reflects spatio-temporal filtering," *NeuroImage*, vol. 228, p. 117711, 2020.

[23] Y. Deng, I. Choi, B. Shinn-Cunningham, and R. Baumgartner, "Impoverished auditory cues limit engagement of brain networks controlling spatial selective attention," *NeuroImage*, vol. 202, p. 116151, 2019.

[24] L. M. Bonacci, S. Bressler, and B. G. Shinn-Cunningham, "Nonspatial features reduce the reliance on sustained spatial auditory attention," *Ear Hear.*, vol. 41(6), pp. 1635–1647, 2020.

[25] A. P. de Souza, Q. B. Soares, L. B. Felix, and E. M. A. M. Mendes, "Classification of auditory selective attention using spatial coherence and modular attention index," *Comput. Methods Programs Biomed.*, vol. 166, pp. 107–113, 2018.

[26] J. Heo, H. J. Baek, S. Hong, M. H. Chang, J. S. Lee, and K. S. Park, "Music and natural sounds in an auditory steady-state response based brain-computer interface to increase user acceptance," *Comput. Biol. Med.*, vol. 84, pp. 45–52, 2017.

[27] Y. Yokota, K. Tanaka, M. Chang, Y. Naruse, Y. Imamura, and S. Fujii, "Gamma music: a new acoustic stimulus for gamma-frequency auditory steady-state response," *Front. Hum. Neurosci.*, vol. 17, p. 1287018, 2024.

[28] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Front. Neurosci.*, vol. 7, 2013.

Theta-Band Modulation in Autism Spectrum Disorder: Exploratory Holo-Hilbert Maps and Complementary Relative-Power Analysis

Helder Luiz de Souza da Silva Filho
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
helder3345@gmail.com

Suzana Cescon de Souza
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
suzanacescon@ufpa.br

Maria Fernanda de Leão Fava
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
mariaflfava@gmail.com

Kamilla Paixão Castro
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
kamilla.castro@itec.ufpa.br

Antônio Pereira Junior
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
apereira@ufpa.br

***Abstract*—This exploratory study examined parietal EEG activity before and after a visual cancellation task in six participants with autism spectrum disorder (ASD) and six controls. EEG was acquired at 256 Hz for 180 s per resting condition, and P3 and P4 channels were analyzed. Holo-Hilbert spectral analysis (HHSA) maps were used to describe carrier-frequency and amplitude-modulation patterns; visual map differences were not treated as evidence of neural adaptation. A complementary analysis estimated relative theta power, defined as 4–8 Hz power divided by 1–45 Hz power, from nonoverlapping 4-s epochs (45 epochs per recording). A linear mixed-effects model included group, period, channel, their interactions, and a participant-specific random intercept. Channel-specific pre-to-post change contrasts were tested with Holm correction. The ASD-minus-control difference in change was +0.0050 at P3 (adjusted $p = 0.8906$) and -0.0710 at P4 (unadjusted 95% CI, -0.1329 to -0.0091; adjusted $p = 0.0491$). However, the three-way group-by-period-by-channel interaction was not significant ($p = 0.1018$). Due to small sample size and incomplete clinical characterization, the findings are preliminary and do not establish a robust group-specific task response or a validated functional marker.**

***Index Terms*—EEG, autism spectrum disorder, theta band, Holo-Hilbert spectral analysis, visual attention, mixed-effects model.**

I. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition involving differences in social communication and restricted or repetitive patterns of behavior [1]–[3]. EEG allows noninvasive assessment of brain electrical activity and has been used to investigate atypical brain dynamics in ASD [4], [5]. Theta activity has been associated with cognitive control, sustained attention, and performance monitoring, especially in midfrontal EEG studies [6]–[9]. However, this literature does not establish that parietal theta changes at a single electrode reflect a specific cognitive mechanism.

This study analyzed EEG recorded at P3 and P4 during resting periods before and after a visual cancellation task, a paradigm used to assess visual search and attentional performance [10]–[14]. These electrodes sample left and right parietal scalp regions according to the international 10–20 system [15]. Although parietal regions are involved in attention networks [16], [17], scalp location alone does not identify neural generators or demonstrate lateralized attentional function. Prior studies have reported atypical theta-band responses and attentional-network differences in ASD [18]–[21].

Holo-Hilbert spectral analysis (HHSA) represents carrier and amplitude-modulation frequencies through adaptive decompositions, making it suitable for nonlinear and nonstationary signals [22], [23]. In this work, HHSA maps were retained for descriptive inspection, while a complementary Welch-based relative theta analysis provided a reproducible quantitative outcome [24]. The objective was to estimate whether pre-to-post changes in relative theta power differed between ASD and control groups and whether this difference varied between P3 and P4. Given the limited sample and unresolved data-quality issues, the analysis was treated as exploratory.

II. Theoretical Background

A. Theta Activity and Parietal Scalp Signals

Theta activity motivates cognitive investigation, but evidence from midfrontal cognitive-control studies does not establish a specific function for parietal theta changes [6]–[9]. EEG studies in ASD have reported atypical oscillatory activity, altered connectivity, and differences in attentional processing [4], [5], [18]–[21]. P3 and P4 therefore denote scalp recording sites rather than isolated neural generators [15]. Consequently, this study does not directly measure neural flexibility, adaptive reorganization, or physiological recovery.

B. Holo-Hilbert Spectral Analysis

HHSA extends adaptive signal decomposition to represent carrier and amplitude-modulation frequencies [22], [23]. In empirical mode decomposition, the signal is expressed as

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t), \tag{1}$$

where $c_i(t)$ are intrinsic mode functions and $r_n(t)$ is the residual [22]. An IMF can be represented as

$$c_i(t) = a_i(t)\cos\phi_i(t), \qquad f_i(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi_i(t)}{\mathrm{d}t}. \tag{2}$$

Here, f_i is expressed in hertz. A further decomposition of the envelopes characterizes modulation frequencies, yielding a representation denoted here by $H(f_c, f_m)$, where f_c and f_m are carrier and modulation frequencies. The original HHSA region of interest comprised $4 \le f_c \le 8$ Hz and $0.5 \le f_m \le 4$ Hz. This differs from the complementary relative-power outcome, which integrates a conventional power spectral density over frequency [24] and is not interchangeable with HHSA energy.

III. Materials and Methods

A. Participants and Experimental Protocol

The sample comprised 12 participants: six with ASD and six controls. The source manuscript reports that ASD diagnoses were established by specialized professionals using DSM-5 criteria [2]. Controls reportedly had no history of neurological or psychiatric disorders or psychoactive medication use. Reported exclusion criteria included epilepsy, brain injury, and medical conditions that could compromise EEG interpretation. The study reports written informed consent and approval by the Research Ethics Committee of the Federal University of Pará, under CAAE 56365722.9.0000.0018.

Participants completed a computerized visual cancellation task on a 21.5-inch monitor in a quiet room. Visual cancellation tasks are used to assess visual search and attentional performance [10]–[14]. Each display contained 64 stimuli in an 8×8 grid, with 16 targets and 48 distractors, following the automated cancellation-task design described by Cascaes et al. [11]. Task duration was fixed at 7 min 30 s. The planned sequence included approximately 3 min of initial rest, task execution, and approximately 3 min of final rest, totaling approximately 13 min 30 s. Only the pre-task and post-task resting periods were analyzed, as summarized in Fig. 1. The original protocol diagram labels both resting periods as eyes open; this condition requires confirmation against acquisition records.

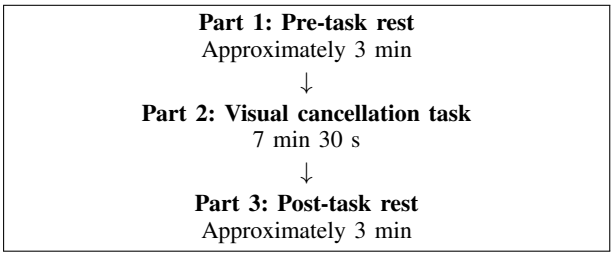


Fig. 1. Planned experimental sequence. Only Parts 1 and 3 were analyzed. The diagram does not depict measured physiological recovery.

B. Acquisition and Relative Theta Power

EEG was acquired with the NeuroMap system (Neurotec, Brazil), using 21 channels in the international 10–20 arrangement [15], 16-bit resolution, bilateral mastoid referencing, and a sampling frequency of 256 Hz. The available analysis used separately exported P3 and P4 MATLAB files for each resting period. All recordings corresponded to 180 s of effective acquisition, yielding 45 nonoverlapping 4-s epochs for every participant, period, and channel.

The quantitative reanalysis read the numeric signal stored in the variable `data`. The source manuscript describes visual inspection, 0.5–40 Hz band-pass filtering, and independent component analysis before channel extraction; however, these steps were not implemented by the supplied Python reanalysis and could not be verified from the available processing records.

For each 4-s epoch, the mean was removed and a power spectral density was computed using `scipy.signal.welch` [24], with a Hann window, full-epoch segment length, and no additional detrending. The Fourier spacing was 0.25 Hz. The epoch-level outcome was

$$R_{\theta,e} = \frac{\int_4^8 S_e(f)\,\mathrm{d}f}{\int_1^{45} S_e(f)\,\mathrm{d}f}, \tag{3}$$

where $S_e(f)$ is the epoch power spectral density. Integrals were approximated by the trapezoidal rule. The participant-period-channel outcome was the arithmetic mean of valid epoch-level ratios. This dimensionless fraction was not normalized by group maximum, baseline, image color scale, signal amplitude, or z-score, and should not be described as HHSA energy normalization.

The code excluded epochs with nonfinite samples and omitted ratios when total power was nonpositive. It also allowed amplitude-based rejection, but no amplitude threshold was supplied because the signal unit and an appropriate threshold were not documented. The saved analysis records contained zero rejected epochs. These checks establish numerical eligibility, not freedom from ocular, muscular, or movement artifacts.

C. HHSA Maps and Parameter Definitions

The original manuscript reports customized MATLAB processing based on HHSA, including IMF decomposition, Hilbert analysis, envelope decomposition, and Gaussian smoothing for visualization [22], [23]. The supplied figures show carrier frequency on the horizontal axis and modulation frequency on the vertical axis, but the executable MATLAB pipeline and numerical HHSA arrays were not included. Therefore, HHSA maps are retained for descriptive inspection only.

The available description does not specify the exact EMD implementation, sifting threshold, maximum number of iterations or modes, boundary extension, IMF-selection rule, or reproducible HHSA normalization formula. Embedded peak annotations are not treated as participant-level measurements, and color intensities are not compared quantitatively because units, logarithmic transformation, normalization, and aggregation order remain unverified.

For future verification, let H_{jk} be a nonnegative, time-aggregated spectral density evaluated at carrier/modulation bin centers $(f_{c,j}, f_{m,k})$. For the region R defined by carrier frequencies 4–8 Hz and modulation frequencies 0.5–4 Hz, a grid-based peak can be defined as

$$(\hat{j}, \hat{k}) = \arg \max_{(j,k)\in R} H_{jk}, \qquad (\hat{f}_c, \hat{f}_m) = (f_{c,\hat{j}}, f_{m,\hat{k}}). \quad (4)$$

If H_{jk} is a density per unit carrier and modulation frequency, regional energy can be specified as

$$E_R = \sum_{(j,k)\in R} H_{jk} \Delta f_{c,j} \Delta f_{m,k}. \quad (5)$$

These definitions require numerical arrays with documented grids, units, weighting, boundary conventions, and normalization. Logarithmic displays, smoothing, interpolation, and raster rendering may alter apparent peaks; therefore, visual displacements remain descriptive.

Image resolution and analytical resolution are distinct. Coarse binning, interpolation, Gaussian smoothing, and raster rendering can each affect the apparent location and intensity of spectral peaks. The 0.25-Hz Fourier spacing of the complementary Welch-based calculation does not specify HHSA resolution or guarantee 0.25-Hz resolving power for the HHSA maps. Therefore, decimal precision in the original visual annotations should not be interpreted as demonstrated frequency accuracy.

D. Complementary Statistical Analysis

The analysis included 48 participant-level observations: 12 participants × two periods × two channels. Epochs estimated each observation and were not treated as independent participants. A linear mixed-effects model was fitted by restricted maximum likelihood to account for repeated measurements, with a random intercept for participant and fixed effects for group, period, channel, all two-way interactions, and the three-way interaction [25]. Reference categories were control, pre-task, and P3. Optimization used the Powell method, and the reported model converged.

Table I
RELATIVE THETA POWER BY GROUP, PERIOD, AND CHANNEL.

Ch.	Group	Pre-task	Post-task	Change
P3	Control	0.1471 ± 0.0347	0.1568 ± 0.0527	+0.0097
P3	ASD	0.1123 ± 0.0594	0.1269 ± 0.0850	+0.0147
P4	Control	0.1361 ± 0.0277	0.2170 ± 0.0633	+0.0808
P4	ASD	0.1376 ± 0.0694	0.1399 ± 0.0581	+0.0023

For each channel c, the contrast of interest was

$$\delta_c = (\mathrm{ASD}_{\mathrm{post},c} - \mathrm{ASD}_{\mathrm{pre},c}) - (\mathrm{Control}_{\mathrm{post},c} - \mathrm{Control}_{\mathrm{pre},c}). \quad (6)$$

Contrasts were derived from the fixed effects and covariance matrix. Two-sided asymptotic Wald z tests and unadjusted 95% confidence intervals were calculated. Holm correction was applied across the two channel-specific contrasts [26]. These confidence intervals are pointwise, not simultaneous intervals adjusted for the two comparisons. Effect sizes are reported as differences in relative power on the original scale. The three-way interaction evaluated whether the between-group difference in change differed between channels. No finite-sample or small-sample correction was applied.

IV. RESULTS

A. Descriptive HHSA Maps

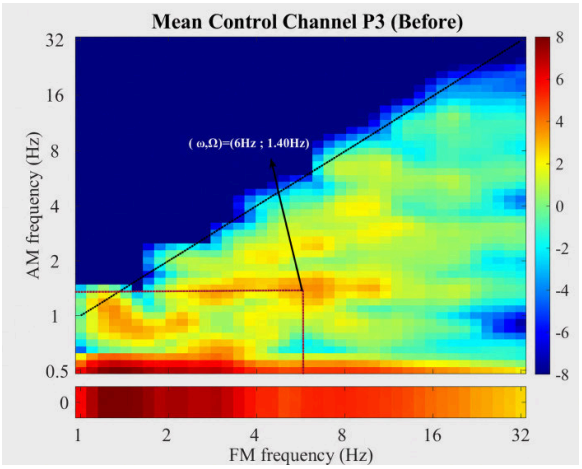
The HHSA maps display spectral-intensity distributions and annotated carrier/modulation coordinates. These annotations describe selected positions on group-labeled images and do not establish a reproducible peak-selection rule or a statistically supported group difference. Shifts in annotated coordinates are not termed desynchronization or adaptive reorganization. The maps also include frequencies beyond the theta region, so not every visible feature represents theta activity. The original within-group HHSA significance claims and Cohen's $d = 0.50$ are not retained because participant-level inputs, channel-specific tests, and uncertainty estimates could not be verified. Significance in one group and nonsignificance in another would not, by itself, demonstrate a difference in task response.

B. Relative Theta Power and Change Contrasts

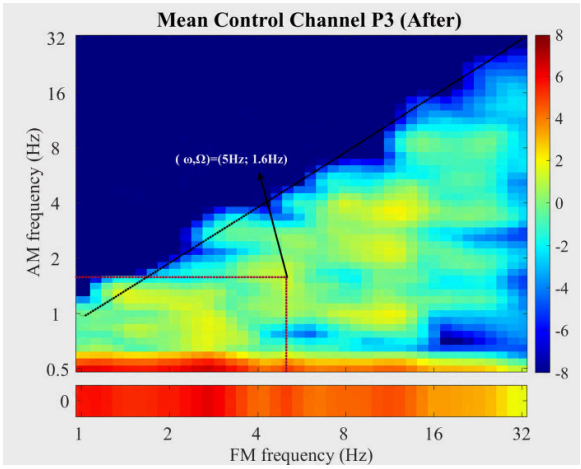
Table I reports participant-level means and standard deviations. Relative theta power increased descriptively in each group-by-channel cell, with the largest mean increase in control P4. These summaries refer to relative power, not absolute theta power or HHSA map energy.

The channel-specific difference in change was not significant at P3 but reached the Holm-adjusted threshold at P4 (Table II). At P4, the negative contrast indicates a smaller mean pre-to-post increase in relative theta power in ASD than in controls. The unadjusted test yielded $p = 0.0246$, with a pointwise 95% CI from -0.1329 to -0.0091. After correction, the P4 contrast remained significant ($p_{\mathrm{Holm}} = 0.0491$), whereas the P3 contrast was not significant ($p_{\mathrm{Holm}} = 0.8906$).

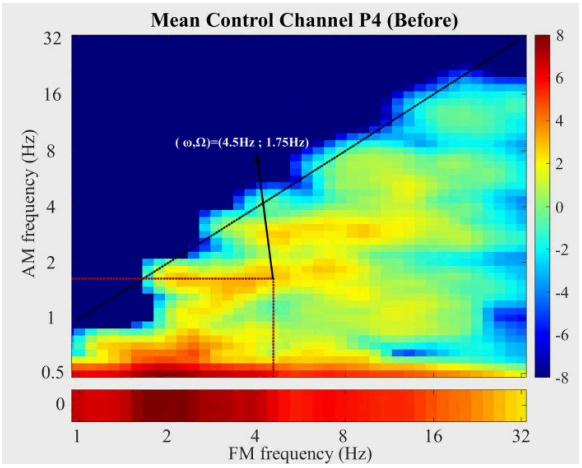
The group-by-period-by-channel interaction was not significant (Wald $\chi^2(1) = 2.6769$, $p = 0.1018$). Thus, although the



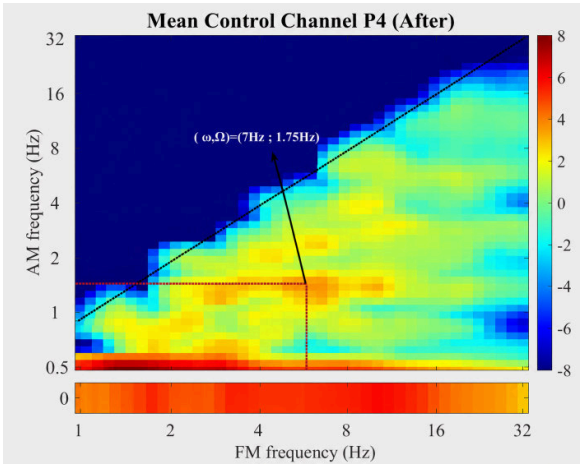
(a) Pre-task (P3)



(b) Post-task (P3)



(c) Pre-task (P4)



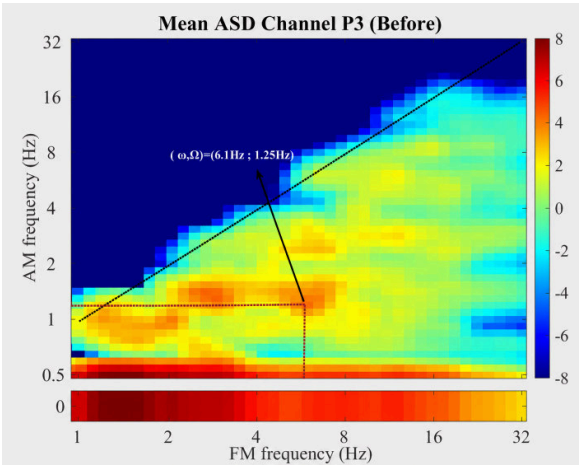
(d) Post-task (P4)

Fig. 2. Control HHSA maps retained from the original project as descriptive group-labeled maps. Panel annotations are original visual selections, not validated peak estimates; aggregation method, color normalization, and numerical values remain unverified.

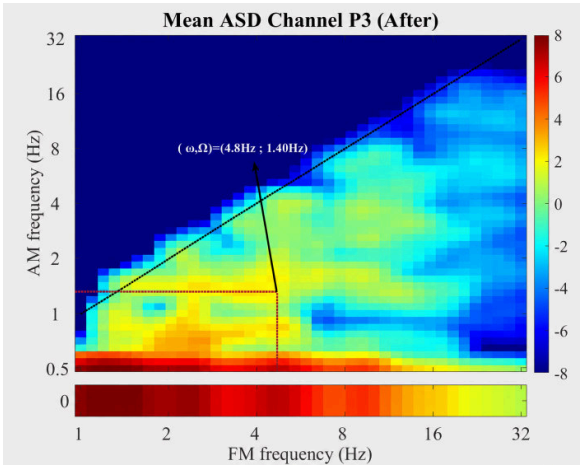
Table II
MODEL-BASED DIFFERENCE IN CHANGE, ASD MINUS CONTROL.

Ch.	δ_c	SE	z	95% CI	Holm p
P3	+0.0050	0.0361	+0.1376	-0.0658 to +0.0757	0.8906
P4	-0.0710	0.0316	-2.2480	-0.1329 to -0.0091	0.0491

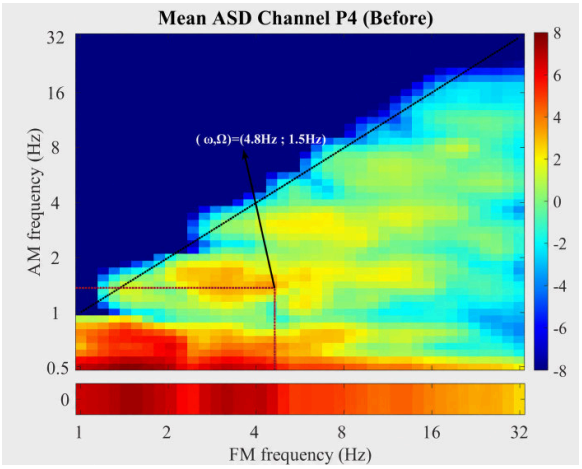
P4 contrast reached the corrected threshold, the interaction did not show that the between-group difference in change differed significantly between P3 and P4.



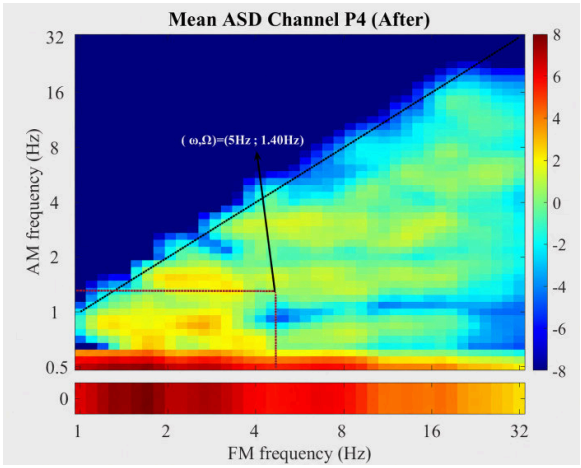
(a) Pre-task (P3)



(b) Post-task (P3)



(c) Pre-task (P4)



(d) Post-task (P4)

Fig. 3. ASD HHSA maps retained from the original project as descriptive group-labeled maps. Panel annotations are original visual selections, not validated peak estimates; aggregation method, color normalization, and numerical values remain unverified.

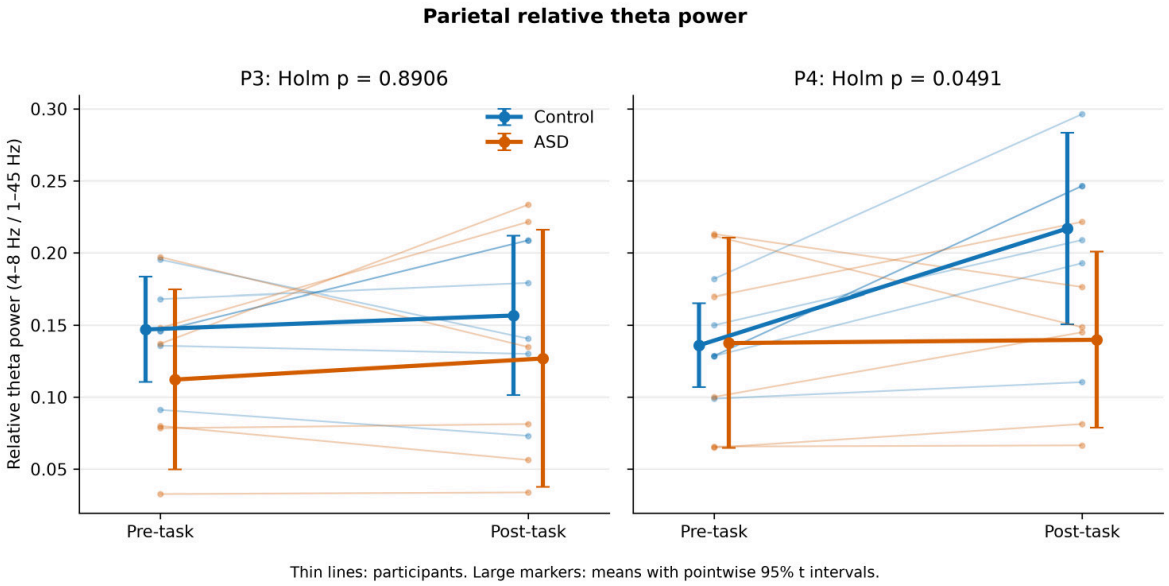


Fig. 4. Relative theta power at P3 and P4 before and after the visual attention task. Thin lines connect observations from the same participant; large markers show group means with pointwise 95% descriptive intervals.

V. Discussion

The quantitative analysis showed increased post-task relative theta power in control P4 and little average change in ASD P4. This observation is discussed in light of literature relating theta activity to attentional and cognitive-control processes [6]–[9]. The P4 contrast reached the Holm-adjusted threshold, whereas the P3 contrast did not. However, the nonsignificant group-by-period-by-channel interaction indicates no evidence that the between-group difference in change varied significantly between P3 and P4. Therefore, the P4 result should be interpreted cautiously and confirmed in a larger independent sample.

The HHSA maps and the relative-power analysis address different quantities. A change in HHSA peak location or intensity need not correspond to a change in band-integrated relative power. Relative theta power may also change through variation in the denominator, so an increase in this ratio does not establish an increase in absolute theta activity. The available analysis does not resolve which spectral components account for the observed ratios.

Interpretations involving neural flexibility, recovery efficiency, or hemisphere-specific attentional mechanisms exceed the available measurements, although attention-network models and ASD studies provide useful context [16], [17], [19]–[21]. Two scalp electrodes do not establish source localization, and this analysis did not quantify connectivity, cross-frequency coupling, or association with behavioral performance. Without a no-task condition, changes between resting periods cannot be uniquely attributed to the visual attention task.

A. Comparison with Conventional Spectral Analysis

HHSA represents carrier and envelope-modulation frequencies jointly [22], [23]; Welch's method estimates power spectral density over frequency [24]. The present Welch-based relative-theta analysis provides a reproducible quantitative reference but not a direct validation of HHSA, because relative band power and HHSA modulation-region energy are different outcomes. A direct benchmark was not possible because the numerical HHSA arrays and executable pipeline were unavailable.

Future comparisons should apply HHSA and conventional methods to the same verified, duration-matched, artifact-controlled segments. Shared outcomes would require compatible normalization before agreement and repeatability are assessed. Synthetic AM–FM signals could estimate bias and robustness, while EEG comparisons should evaluate reproducibility and uncertainty rather than rank methods by their smallest p value. No classification or independent validation was conducted; therefore, the present findings do not establish a biomarker.

B. Limitations and Next Steps

The sample included only six participants per group, and individual values may influence group summaries and contrasts. Missing age, sex, cognitive-level, ASD-severity, and medication information prevents assessment of group comparability and confounding. The bounded relative-power outcome also requires evaluation of model fit, and the random-intercept model imposes a common participant-level covariance contribution across periods and channels. Residual diagnostics and sensitivity to alternative covariance assumptions have not yet been documented.

Artifact handling and upstream preprocessing remain incompletely documented. A verified pipeline should report retained duration, exclusions, signal units, filtering, spectral normalization, residual diagnostics, and sensitivity to alternative model or normalization assumptions. Future work should use larger characterized samples, documented behavioral outcomes, consistent resting-state conditions, reproducible HHSA outputs, and independent validation where relevant.

VI. Conclusion

HHSA provided descriptive representations of carrier and modulation frequencies in parietal EEG before and after a visual attention task. Visual peak differences alone do not demonstrate adaptive reorganization, neural flexibility, or physiological recovery. In the complementary relative-theta analysis, the between-group difference in pre-to-post change at P4 reached the Holm-adjusted threshold, whereas the P3 contrast was not significant. However, the group-by-period-by-channel interaction was not significant, and the small sample and methodological limitations warrant cautious interpretation. These findings remain preliminary and exploratory. Larger independent studies with reproducible HHSA outcomes and fuller participant characterization are required before theta modulation can be proposed as a functional marker.

References

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, 2018. https://doi.org/10.1016/S0140-6736(18)31129-2

[2] American Psychiatric Association, *Manual Diagnóstico e Estatístico de Transtornos Mentais: DSM-5*. Porto Alegre, Brazil: Artmed, 2013.

[3] R. Marotta et al., "Neurodevelopmental disorders: a review on the epidemiology, clinical features, and neuroimaging findings in autism spectrum disorder," *Frontiers in Pediatrics*, vol. 8, p. 377, 2020.

[4] R. Coben, A. R. Clarke, W. Hudspeth, and R. J. Barry, "EEG power and coherence in autistic spectrum disorder," *Clinical Neurophysiology*, vol. 119, no. 5, pp. 1002–1009, 2008.

[5] A. G. Nicotera et al., "Abnormal EEG complexity in autism spectrum disorder: a developmental perspective," *Brain Topography*, vol. 32, no. 6, pp. 935–947, 2019.

[6] J. F. Cavanagh and M. J. Frank, "Frontal theta as a mechanism for cognitive control," *Trends in Cognitive Sciences*, vol. 18, no. 8, pp. 414–421, 2014. https://doi.org/10.1016/j.tics.2014.04.012

[7] D. J. Mitchell, N. McNaughton, D. Flanagan, and I. J. Kirk, "Frontal-midline theta from the perspective of hippocampal theta," *Progress in Neurobiology*, vol. 86, no. 3, pp. 156–185, 2008.

[8] C.-H. Chang et al., "Evaluating attentional control mechanisms through theta oscillations during visuospatial tasks," *Neuropsychologia*, vol. 168, p. 108190, 2022.

[9] C.-H. Juan et al., "Revealing the role of theta oscillations in sustained attention using EEG during cancellation tasks," *Biological Psychology*, vol. 162, p. 108114, 2021.

[10] C. O. Cardoso, O. F. A. Bueno, and J. C. da Costa, "Cancelamento de estímulos visuais: análise de estratégias de busca em adultos jovens," *Psicologia: Reflexão e Crítica*, vol. 24, no. 3, pp. 482–490, 2011.

[11] A. Cascaes et al., "Cars' test: An automated visual cancellation task for assessment of attention and visual search strategies," *Applied Neuropsychology: Adult*, vol. 27, no. 5, pp. 436–445, 2020.

[12] D. Kaur, T. V. Chacko, and S. Abraham, "Cancellation tests: A review and proposal for a new classification," *Applied Neuropsychology: Adult*, vol. 28, no. 3, pp. 299–310, 2021.

[13] T. Manly et al., "The test of everyday attention," *Journal of the International Neuropsychological Society*, vol. 9, no. 6, pp. 982–986, 2003.

[14] L. Diller and Y. Ben-Yishay, *Neuropsychological Rehabilitation*. New York, NY, USA: Wiley-Interscience, 1974.

[15] H. H. Jasper, "The ten-twenty electrode system of the International Federation," *Electroencephalography and Clinical Neurophysiology*, 1958.

[16] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[17] M. I. Posner and S. E. Petersen, "The attention system of the human brain," *Annual Review of Neuroscience*, vol. 13, pp. 25–42, 1990.

[18] G. Akdeniz, "Delta and theta band power alterations during face and face pareidolia perception in children with autism spectrum disorder: An electroencephalographic analysis," *Medicina*, vol. 61, no. 4, p. 754, 2025.

[19] J. Larraín-Valenzuela et al., "Theta and alpha oscillations during the processing of facial expressions in children with autism spectrum disorder," *Biological Psychology*, vol. 123, pp. 106–114, 2017.

[20] B. Keehn, R.-A. Müller, and J. Townsend, "A brief overview of research on attention in autism spectrum disorders," *Journal of Learning Disabilities*, vol. 46, no. 4, pp. 327–341, 2013.

[21] B. Keehn et al., "Neural abnormalities in the brain attention network in autism," *Brain Research*, vol. 1512, pp. 72–81, 2013.

[22] N. E. Huang et al., "On Holo-Hilbert spectral analysis: a full informational spectral representation for nonlinear and non-stationary data," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, article 20150206, 2016. https://doi.org/10.1098/rsta.2015.0206

[23] N. E. Huang, Z. Wu, and X. Long, "Holo-Hilbert spectral analysis: a new method for analyzing nonlinear and nonstationary data," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, p. 20150206, 2016.

[24] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967. https://doi.org/10.1109/TAU.1967.1161901

[25] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. New York, NY, USA: Springer, 2000.

[26] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. https://www.jstor.org/stable/4615733

Time-Domain Speech Signal Enhancement with Impulsive Noise Using U-Net: A Preliminary Step Toward Future Assessment of Auditory Attention in Individuals with Central Auditory Processing Disorder

1st Janara Hellen Marques da Silva
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
0009-0003-3862-0184

2nd Cristiano Jacques Miosso
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
0000-0002-7605-6305

3rd Gabriella Gomes dos Santos
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
0009-0002-0169-6874

Abstract*—Deep learning-based speech enhancement has emerged as an alternative to classical noise reduction techniques, particularly in non-stationary acoustic scenarios. This study proposes a supervised approach based on a one-dimensional U-Net architecture, applied directly to the raw waveform, aimed at attenuating impulsive noise in speech signals. To prevent information leakage across experimental subsets, the LibriSpeech corpus was partitioned by speaker prior to segmentation, yielding 176 speakers for training, 38 for validation, and 37 for testing, with no overlap between groups. From this partitioning, 16,000 four-second segments were generated at a 16 kHz sampling rate—12,000 for training, 2,000 for validation, and 2,000 for final evaluation.Eighteen independent recordings of impulsive noise from door knocks and impacts, released under a Creative Commons Zero license, were distributed independently across the training, validation, and test subsets. Noisy signals were synthesized under four controlled input signal-to-noise ratio conditions (-5, 0, 5, and 10 dB), and model performance was assessed using four metrics: SNR, Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Short-Time Objective Intelligibility (STOI), and Perceptual Evaluation of Speech Quality (PESQ).On the independent test set of 2,000 samples, the average SNR improvement reached 3.21 dB, with a positive gain observed in 99.9% of the evaluated samples. SI-SDR increased by an average of 3.00 dB. STOI and PESQ, in turn, showed average variations of 0.0076 and 0.0394, respectively—values indicating comparatively modest effects on intelligibility and perceptual quality relative to the energy-based gains. These findings suggest that, although the U-Net effectively reduced impulsive noise energy under the tested conditions, the corresponding perceptual gains were considerably more limited. This work constitutes a preliminary step toward a broader research effort, which is expected to include perceptual evaluation and testing with individuals diagnosed with Central Auditory Processing Disorder.

Keywords: speech enhancement; U-Net; deep learning; impulsive noise; SNR; SI-SDR; STOI; PESQ.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Finance Code 001.

I. Introduction

Speech enhancement (SE) aims to recover signals degraded by environmental noise and acoustic distortions, with the goal of improving their quality and intelligibility. These techniques are applied in telecommunications systems, automatic speech recognition, hearing aids, and human-machine interfaces [1]–[3].

Traditionally, noise reduction relied on methods such as Wiener filtering, spectral subtraction, and least-squares estimation. However, these approaches have limitations when dealing with non-stationary noise and variable acoustic conditions. With advances in deep learning, neural networks have begun to deliver superior results in speech enhancement tasks, operating in the time domain, the frequency domain, or in time-frequency representationsa [1], [4].

Among the architectures employed are Convolutional Neural Networks, Recurrent Neural Networks, and encoder-decoder models. Subsequently, attention mechanisms, multistage structures, Transformers, and diffusion models were incorporated to improve the representation of the temporal and spectral dependencies of the signals [1], [4]. Adversarial approaches, such as Generative Adversarial Networks (GANs), have also been used to generate signals with greater naturalness and perceptual quality [5].

In GANs, a generator produces enhanced signals, while a discriminator seeks to distinguish real samples from the

generated ones. Recent variants incorporate attention mechanisms and discriminative latent representations to improve speech reconstruction, especially under low signal-to-noise ratio conditions [3], [6]. However, these architectures require the simultaneous training of competing networks, thereby increasing computational complexity.

Encoder-decoder models with shortcut connections, such as U-Net and Wave-U-Net, offer a simpler alternative, as they preserve relevant temporal information and require fewer components to be optimized simultaneously [7], [8]. In light of this, this study proposes a one-dimensional U-Net architecture, trained in a supervised manner directly in the time domain, to evaluate its ability to reduce impulsive noise in speech signals.

The motivation for this study stems from the difficulties faced by people with Central Auditory Processing Disorder (CAPD), a condition that can impair speech comprehension, especially in the presence of competing or degraded acoustic stimuli. Everyday noises, such as computer keyboards, barking, the movement of objects, and background speech, can compete with the target voice and impair the identification and comprehension of the spoken message. This effect is particularly relevant in educational and professional settings, where background noise can impair speech perception, auditory attention, and the ability to follow instructions [9]–[11].

In this context, this study constitutes a preliminary stage of an investigation focused on individuals with CP. In this first stage, we develop and objectively evaluate a system for reducing impulsive noise in speech signals. Subsequently, we intend to evaluate the impact of this processing on speech intelligibility and comprehension by people with CP, as well as to investigate the incorporation of a discriminator for comparison with an equivalent adversarial architecture

II. Materials and Methods

This study proposes the development of a deep learning-based speech enhancement system using a one-dimensional U-Net architecture applied directly to audio signals in the time domain. The choice of this approach was motivated by recent studies demonstrating the feasibility of encoder-decoder models with skip connections for noise reduction and speech enhancement tasks, especially when applied directly to the signal's waveform [7], [12].

A. Base de Dados

The database for this study was constructed from clean speech signals extracted from the LibriSpeech corpus, a widely used reference in speech processing and automatic speech recognition research. Derived from public-domain audiobooks, the corpus organizes its signals by speaker, chapter, and reading segment—a structure that allows for controlling the identity of the individuals present in each experimental subset. All files are available in a single channel, with a sampling rate of 16 kHz [13].

The *train-clean-100* subset was selected, containing approximately 100 hours of speech, in which 251 distinct speakers were identified. The recordings of these speakers formed the basis for the training, validation, and test datasets used in this study.

Data partitioning was performed at the speaker level, prior to any signal segmentation, ensuring independence among the subsets. A fixed pseudorandom seed (42) was adopted to ensure the reproducibility of the procedure. The 251 speakers were distributed into three mutually exclusive groups: 176 for training, 38 for validation, and 37 for testing, corresponding to approximately 70.1%, 15.1%, and 14.7% of the total, respectively.

Once this division was complete, the speaker identifiers were automatically verified, confirming the absence of overlap among the three sets. The purpose of this step was to ensure that the model was evaluated exclusively with voices not used during training.

In terms of files, the training set comprised 20.001 recordings, corresponding to approximately 70.37 hours of speech. The validation set totaled 4.400 files, amounting to about 15.52 hours, while the test set comprised 4,138 files, equivalent to approximately 14.70 hours. In all cases, the original sampling rate of 16 kHz was maintained.

The recordings were then segmented into fixed 4-second windows, each containing 64.000 samples. This standardization ensured a uniform time scale for the network inputs and enabled batch processing during training. This procedure yielded 16.000 clean speech segments, of which 12.000 were allocated for training, 2,000 for validation, and 2.000 for testing

Segments were selected based on a criterion of balance across speakers, ensuring that individuals with a larger number of recordings did not contribute disproportionately to the composition of the dataset. Thus, each speaker contributed approximately 68 to 69 segments to the training set, 52 to 53 to the validation set, and 54 to 55 to the test set.

Throughout all stages of processing—including segmentation, mixture generation, and neural network processing—the signals remained in a single channel at their original frequency of 16 kHz. Thus, each input provided to the model was represented by 64.000 time points.

Noisy signals were generated independently for each subset. Each set received only the noise sources previously reserved for its respective experimental stage, preserving the separation both between speakers and between noise recordings.

Ultimately, the dataset used in the study consisted of 16.000 clean–noisy signal pairs, distributed among training, validation, and testing without any overlap between speakers or noise sources. This organization allows for a more reliable assessment of the model's generalization ability when faced with voices and impulsive events not observed during training.

B. Impulsive Noise Dataset

o construct the degraded signals, 18 independent recordings of impulsive noise were used, including knocks, impacts, and events associated with doors, obtained from the Freesound repository under a Creative Commons Zero (CC0) license. The use of multiple recordings was intended to introduce variability

in aspects such as intensity, duration, spectral content, and temporal distribution of the events. Impulsive noises generally exhibit non-stationary behavior, characterized by abrupt variations in energy over short intervals, which poses particular challenges for speech enhancement systems [14].

These recordings were previously divided into three mutually exclusive groups, corresponding to the training, validation, and test sets. Six noise sources were reserved for each stage, with no file sharing between the subsets. Thus, the final evaluation was performed using impulse events derived from recordings that were completely unobserved during model training. Before generating the mixtures, all noise material was standardized to a single channel and a sampling rate of 16 kHz, ensuring compatibility with the speech signals used in this study. The original recordings were kept separately, while only their standardized versions were used in the experimental stages.

When constructing the noise windows, care was taken to ensure that each segment contained at least one impulsive event. For recordings lasting four seconds or less, the signal was inserted in its entirety into a four-second window, with variable temporal positioning. For recordings longer than this interval, regions of interest were identified based on the signal's energy envelope, and four-second windows centered on the detected events were then extracted

During the construction of the noise windows, care was taken to ensure that each segment contained at least one impulsive event. For recordings with a duration equal to or shorter than four seconds, the signal was inserted in its entirety into a four-second window, with variable temporal positioning. For recordings longer than this interval, regions of interest were identified based on the signal energy envelope, and four-second windows centered on the detected events were then extracted.

After removing duplicate excerpts and confirming the presence of at least one impulsive event in each segment, 58 distinct noise windows were obtained. Of these, 30 were assigned to the training set, 7 to the validation set, and 21 to the test set.

The use of independent noise sources across the three subsets, combined with the requirement that all segments contain impulsive events, was adopted to reduce the possibility of memorizing specific characteristics of the recordings. This strategy also enables a more rigorous assessment of the model's ability to generalize to previously unseen impulsive events belonging to the same acoustic class.

To ensure the traceability and reproducibility of the experimental dataset, the 18 impulsive-noise recordings used in this study were made available in the public repository associated with this work. The files were preserved with identifiers that allow them to be linked to their respective source recordings on Freesound. The repository also presents the distribution of the noise sources among the training, validation, and test subsets, preserving a record of the recordings effectively used in the experiments. The files can be accessed at GitHub.

C. Generation of Noisy Signals

Each noisy signal was obtained by combining a clean speech segment with an impulsive-noise window belonging to the same experimental subset. Thus, signals assigned to the training set were combined exclusively with noise recordings reserved for training, and the same procedure was adopted for the validation and test sets, with no crossover between the sources used at different stages. This organization preserved the independence among the subsets during the mixture-generation process as well.

The signals were combined directly in the time domain. For each clean speech segment $x(n)$, an impulsive-noise window $r(n)$ was selected, both with a duration of four seconds and a sampling rate of 16 kHz. Before summing the signals, the noise amplitude was adjusted to produce a predefined input signal-to-noise ratio. This control was performed based on the RMS (*Root Mean Square*) values of the speech and noise signals, according to the conventional SNR formulation commonly employed in speech processing and enhancement [15].

Four input SNR conditions were considered: -5, 0, 5, and 10 dB.

These levels represent different degrees of acoustic degradation. At -5 dB, the energy contribution of the noise is greater than that of the speech signal, representing the highest-interference condition among those evaluated. As the SNR increases to 0, 5, and 10 dB, the relative contribution of the noise to the mixture is progressively reduced.

The four conditions were evenly distributed across all experimental subsets. In the training set, composed of 12,000 segments, 3,000 mixtures were generated for each SNR level. In the validation and test sets, each containing 2,000 segments, 500 mixtures were generated per condition. Thus, all degradation levels remained equally represented both during model training and in the final evaluation.

The relationship between speech and noise was controlled based on the RMS value of each signal. For a discrete signal $x(n)$, composed of N samples, the RMS is defined as

$$\mathrm{RMS}(x) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x^2(n)}.$$

Considering the clean speech signal $x(n)$ and the already scaled noise $r_s(n)$, the signal-to-noise ratio in decibels is expressed as

$$\mathrm{SNR}_{\mathrm{dB}} = 20\log_{10}\left(\frac{\mathrm{RMS}(x)}{\mathrm{RMS}(r_s)}\right).$$

Based on the previously established target SNR, the desired RMS value of the noise is determined by

$$\mathrm{RMS}_{r,\mathrm{desejado}} = \frac{\mathrm{RMS}(x)}{10^{\mathrm{SNR}_{\mathrm{alvo}}/20}}.$$

Next, the scaling factor α required to adjust the amplitude of the original noise window is calculated as

$$\alpha = \frac{\mathrm{RMS}_{r,\mathrm{desejado}}}{\mathrm{RMS}(r)}.$$

The scaled noise is then obtained as $r_s(n) = \alpha r(n)$, while the final noisy signal is produced by the sum $y(n) = x(n) + r_s(n)$.

This procedure allows direct control of the relationship between the RMS amplitudes of the speech and noise signals in each segment, reproducing the four SNR conditions established for the experiment.

After generating each mixture, the maximum absolute amplitude value of the signal was checked. Whenever this value exceeded 0.99, the same gain factor was applied to the clean signal, the scaled noise, and the corresponding mixture. Since all three signals were multiplied by the same factor, digital clipping was avoided without modifying the relative relationship between speech and noise and, consequently, without altering the previously defined SNR.

Once the mixtures had been generated, an automatic audit of the dataset was performed. The number of pairs produced in each subset, the distribution of the four SNR levels, the correspondence between clean and noisy signals, the presence of non-numeric or infinite values, possible clipping occurrences, and the difference between the target SNR and the SNR effectively obtained in each mixture were verified.

The audit confirmed the generation of the expected 16,000 pairs, with 12,000 assigned to the training set, 2,000 to the validation set, and 2,000 to the test set, all with a balanced distribution across the four SNR levels. The mean absolute error between the target SNR and the obtained SNR was practically zero, while the maximum error remained on the order of 3×10^{-6} dB. No *NaN* (*Not a Number*) values, infinite values, or remaining clipping cases were identified in the final signals.

In 4,757 of the 16,000 generated pairs, the application of the common gain factor was required for amplitude control. Because this adjustment was applied equally to the clean signal, the scaled noise, and the corresponding mixture, the SNR established for each segment was preserved.

Finally, a detailed record of each mixture was maintained, including the segment identifier, experimental subset, noise source and noise window used, target SNR, effectively obtained SNR, scaling factor applied to the noise, and any gain factor used for amplitude control. This record ensures sample traceability and the reproducibility of the dataset generation process.

D. Signal Preprocessing

Each pair of clean and noisy speech signals maintained temporal alignment, a single channel, a sampling rate of 16 kHz, and a fixed duration of 4 seconds, corresponding to 64,000 samples per segment.

The U-Net received the noisy signal $y(n)$ as input, while the clean signal $x(n)$ was used as the supervised reference. Before being provided to the network, the segments were arranged with dimensions of $64{,}000 \times 1$, corresponding to the temporal axis and the single audio channel.

Both training and inference were performed directly in the time domain, without prior conversion to spectral representations. This approach is consistent with Wave-U-Net-type architectures, which apply one-dimensional convolutions directly to the signal waveform [7], [8].

Throughout the entire processing pipeline, the three subsets remained strictly separated. The 12,000 training pairs were used to update the network parameters, while the 2,000 validation pairs were used to monitor performance during training and guide model selection. The 2,000 test pairs, in turn, were reserved exclusively for the final evaluation.

E. Neural Network Architecture

The adopted architecture was based on a one-dimensional U-Net, composed of four encoding levels, an intermediate region (*bottleneck*), and four symmetric decoding levels. Architectures of this type have been employed in speech enhancement tasks because they enable the combination of representations obtained at different temporal resolutions through skip connections (*skip connections*) [7], [8].

In the encoding path, the four levels used 16, 32, 64, and 128 convolutional filters, respectively, while the *bottleneck* consisted of 256 filters. Each convolutional block contained two one-dimensional convolutional layers with a kernel size of 15 and *same* padding, followed by batch normalization (*batch normalization*) and a *Leaky ReLU* activation function.

Throughout the encoder, the temporal resolution was progressively reduced using *max pooling* layers with a factor of 2. In the decoding path, this dimension was gradually restored through upsampling (*upsampling*) operations, also with a factor of 2. The skip connections concatenated the representations produced at each encoder level with the corresponding decoder layers, favoring the preservation of temporal information during signal reconstruction.

The output layer consisted of a one-dimensional convolution with a kernel size of 1, a single filter, and a hyperbolic tangent (*tanh*) activation function, responsible for generating the enhanced speech signal from the noisy input. In total, the architecture contained 3,275,457 parameters, of which 3,272,513 were trainable and 2,944 were non-trainable.

F. Model Training

The model was trained in a supervised manner, using the noisy signal as the network input and the corresponding clean signal as the reference. The Adam optimizer was employed for parameter optimization, with an initial learning rate of 10^{-4}.

As the loss function, a combination of Mean Absolute Error (MAE) and Mean Squared Error (MSE) was adopted, defined as

$$\mathcal{L} = \mathrm{MAE} + 0.5\,\mathrm{MSE}.$$

This combination was used to simultaneously account for the absolute deviations between the estimated signal and the clean reference signal while assigning a greater penalty to errors of larger magnitude.

Training was conducted with a batch size of 4 and a maximum limit of 50 epochs. During this process, three control mechanisms were employed. *ReduceLROnPlateau* was used to reduce the learning rate when the validation loss stopped

improving; *Early Stopping* allowed training to be interrupted in the absence of further progress on the validation set; and *Model Checkpoint* was used to preserve the weights associated with the best validation performance.

The 12,000 pairs belonging to the training set were used to update the network parameters, whereas the 2,000 validation pairs were used to monitor performance during training and guide model selection. The 2,000 pairs belonging to the test set remained isolated throughout this entire process and were used only during the final evaluation stage.

Training was stopped at epoch 20 according to the early-stopping criterion. At the end of the process, the weights corresponding to epoch 8, at which the best performance on the validation set was obtained, were restored.

The entire experiment was implemented in Python using TensorFlow 2.21 in a WSL 2 environment, with acceleration provided by an NVIDIA GeForce RTX 3050 Laptop GPU with 6 GB of memory.

G. Performance Evaluation

The final evaluation was performed on the 2,000 samples belonging to the test set, which were kept separate throughout all training and validation stages. To analyze the model performance, four complementary metrics were used: *Signal-to-Noise Ratio* (SNR), *Scale-Invariant Signal-to-Distortion Ratio* (SI-SDR), *Short-Time Objective Intelligibility* (STOI), and *Perceptual Evaluation of Speech Quality* (PESQ).

SNR was employed to quantify the relationship between the energy of the clean reference signal and the energy of the residual error. Considering a reference signal $x(n)$ and an evaluated signal $y(n)$, the metric was calculated as

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} [x(n) - y(n)]^2} \right).$$

The metric was calculated before and after processing by the U-Net, with the improvement defined as $\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{after}} - \mathrm{SNR}_{\mathrm{before}}$. Positive values of ΔSNR indicate a reduction in the residual error relative to the clean reference signal [15].

SI-SDR was used as a complementary measure of distortion between the estimated signal and the reference signal. Unlike metrics that are directly sensitive to absolute amplitude, SI-SDR compensates for global scale differences between the signals, allowing a more specific assessment of the distortion introduced during processing. This characteristic makes the metric particularly suitable for signal separation and enhancement tasks [16].

To estimate objective speech intelligibility, STOI was employed, which compares short-time temporal characteristics between the processed signal and its corresponding clean reference. The metric typically produces values ranging from 0 to 1, with values closer to 1 associated with higher estimated intelligibility [17].

Finally, PESQ was used as an objective measure of perceptual quality. This metric compares the processed signal with the clean reference signal using a perceptual model developed to represent aspects related to human perception of speech quality, with higher values associated with better perceived quality [18].

All four metrics were calculated before and after U-Net processing. In addition to the overall analysis of the 2,000 samples in the test set, the results were also evaluated separately for each input SNR condition, corresponding to -5, 0, 5, and 10 dB. This analysis made it possible to observe the model behavior under different levels of acoustic degradation.

III. Results

Training was stopped at epoch 20 according to the *Early Stopping* criterion. The best performance on the validation set was achieved at epoch 8, whose weights were restored for the final evaluation. As shown in Fig. 1, the training loss decreased continuously, whereas the validation loss remained relatively stable after the first epochs, indicating stabilization of the model performance on the validation set.

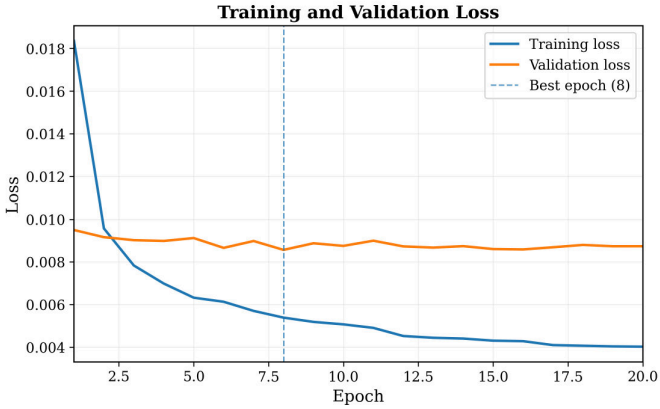


Fig. 1. Evolution of the loss function on the training and validation sets during U-Net optimization. The dashed line indicates the epoch with the best performance on the validation set.

A. Overall Performance on the Test Set

The final evaluation was performed on all 2,000 samples in the test set, with no observations excluded. Overall, the U-Net showed improvements across all four evaluated metrics. The mean gain was 3.21 dB in SNR and 3.00 dB in SI-SDR, whereas STOI and PESQ showed more modest mean variations of 0.0076 and 0.0394, respectively. In addition, 99.9% of the samples exhibited $\Delta\mathrm{SNR} > 0$. The mean values before and after processing are presented in TableI.

Tabela I
Overall performance on the test set.

Metric	Noisy	Enhanced	Change
SNR (dB)	2.500	5.713	+3.213
SI-SDR (dB)	2.092	5.094	+3.003
STOI	0.8580	0.8655	+0.0076
PESQ	1.6581	1.6975	+0.0394

The largest gains were observed for the SNR and SI-SDR metrics, whereas STOI and PESQ showed improvements of smaller magnitude.

B. Performance as a Function of Input SNR

When the 2.000 test samples were stratified according to the input SNR condition, with 500 samples at each level, the highest mean SNR gains were observed at 0 dB and -5 dB, with improvements of 3.53 dB and 3.49 dB, respectively. This was followed by a mean gain of 3.17 dB for the 5 dB condition and, finally, 2.66 dB at 10 dB, corresponding to the smallest improvement among the four evaluated conditions.

SI-SDR showed a similar trend, with mean gains ranging from 2.58 to 3.30 dB. STOI and PESQ also exhibited positive mean variations across all four conditions, although with smaller magnitudes. Regarding the proportion of samples with $\Delta\text{SNR} > 0$, the model achieved 100% in the -5, 0, and 5 dB conditions, with a slight reduction to 99.6% at 10 dB. Table II details these results according to the input condition.

Tabela II
U-NET PERFORMANCE AS A FUNCTION OF INPUT SNR.

SNR	n	ΔSNR	95% CI	ΔSI-SDR	ΔSTOI	ΔPESQ
-5 dB	500	3.490	[3.283; 3.697]	3.091	0.0075	0.0330
0 dB	500	3.530	[3.334; 3.727]	3.295	0.0091	0.0526
5 dB	500	3.173	[3.002; 3.344]	3.043	0.0093	0.0571
10 dB	500	2.661	[2.505; 2.816]	2.583	0.0044	0.0147

A progressive reduction in the mean SNR gain can be observed under less degraded input conditions, with the smallest relative benefit occurring at 10 dB.

C. Distribution of SNR Improvement

Fig. 2 illustrates the distribution of ΔSNR across the four evaluated input conditions. In all cases, most samples showed a positive gain after processing by the U-Net. In the -5, 0, and 5 dB conditions, the distributions remained entirely above zero, whereas at 10 dB only a small fraction of negative results was observed. In addition, a trend toward lower central values of ΔSNR can be noted as the input SNR increases, suggesting that the relative gain provided by the model is smaller under conditions in which the signal already presents a lower level of degradation.

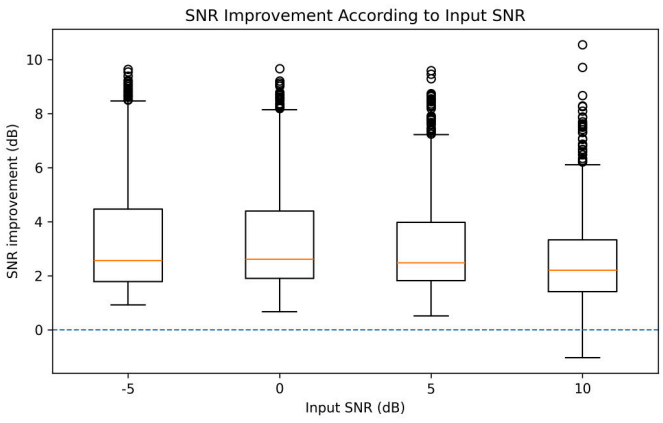


Fig. 2. Distribution of SNR improvement (ΔSNR) under the input conditions of -5, 0, 5, and 10 dB. Each group contains 500 samples from the test set. The dashed horizontal line indicates $\Delta\text{SNR} = 0$, corresponding to the absence of variation after processing.

D. Magnitude of the Observed Effects

The paired analysis between the signals before and after processing indicated large effects for SNR and SI-SDR, with Cohen's d_z values of 1.51 and 1.38, respectively. For STOI, a moderate effect was observed ($d_z = 0.58$), whereas PESQ showed a small effect ($d_z = 0.18$).

For the inferential analysis, paired tests were applied to the values obtained before and after processing, using the paired t-test and, complementarily, the Wilcoxon signed-rank test. The overall differences were statistically significant for all four metrics ($p < 0.001$). However, under the 10 dB input condition, the variation in PESQ did not show statistically significant evidence according to the Wilcoxon test ($p = 0.779$).

These results indicate that the most pronounced changes occurred in SNR and SI-SDR, whereas the effects on STOI and, particularly, PESQ were of smaller magnitude.

E. Qualitative Analysis of the Signals

Additionally, a qualitative analysis was carried out by contrasting the clean reference signal, the noisy input signal, and the corresponding U-Net output. Fig. 3 presents the spectrograms of the three signals for sample `test_001863`, all represented using the same magnitude scale to facilitate visual comparison. It can be observed that the network output preserves a large part of the spectrotemporal structure of the speech signal while attenuating the components associated with impulsive interference.

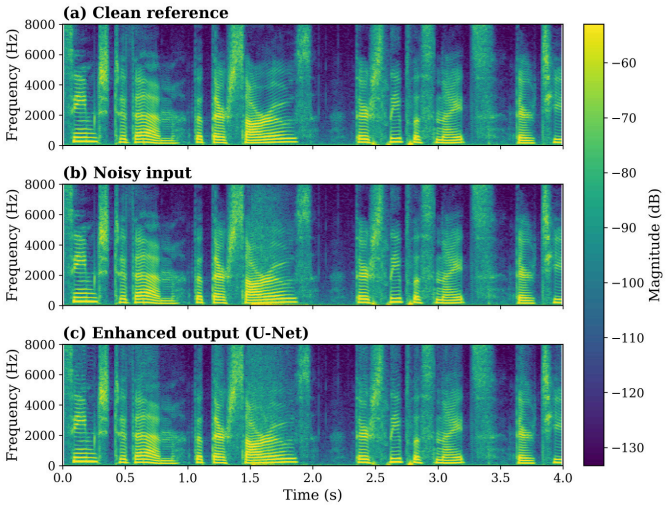


Fig. 3. Spectrogram comparison for sample `test_001863`: (a) clean reference signal, (b) noisy input signal, and (c) signal enhanced by the U-Net. All three spectrograms use the same magnitude scale to allow direct comparison.

Fig. 4 presents the corresponding comparison in the time domain. Panel (a) shows the complete signals over the four-second interval, whereas panel (b) presents a magnified region to facilitate visualization of local differences among the three waveforms. In the enlarged interval, the noisy signal exhibits local deviations from the clean reference, while the U-Net output follows the reference waveform more closely in several regions.

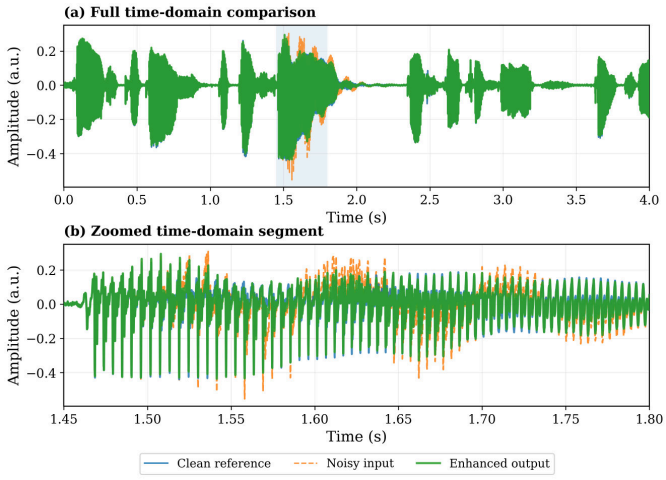


Fig. 4. Time-domain comparison for sample `test_001863`. (a) Complete signals over the 4-s segment, with the selected region highlighted. (b) Enlarged view of the highlighted interval, showing the clean signal, the noisy signal, and the output enhanced by the U-Net.

These figures are illustrative and qualitative in nature, whereas the conclusions regarding model performance are based on the aggregate analysis of the 2.000 samples in the test set.

F. Audio Example

To complement the quantitative evaluation and visual analysis, audio examples containing the clean reference signal, the noisy input signal, and the output produced by the U-Net were made available. To facilitate the auditory perception of the impulsive interference and its attenuation after processing, a sample from the test set with an input SNR of -5 dB and positive performance across the evaluated metrics was selected. The audio files are available in the public repository associated with this work, as presented in Table III.

Tabela III
AUDIO EXAMPLE AVAILABLE IN THE PUBLIC REPOSITORY ASSOCIATED WITH THIS WORK.

Signal	**Link**
Clean reference	Listen to audio
Noisy	Listen to audio
Enhanced by the U-Net	Listen to audio

IV. Discussion and Limitations

Taken together, the results indicate a consistent reduction in the error present in signals degraded by impulsive noise through the use of the one-dimensional U-Net. The mean improvements of 3.21 dB in SNR and 3.00 dB in SI-SDR, together with a positive SNR gain in 99.9% of the 2,000 test samples, demonstrate stable model performance under the evaluated conditions. Notably, the largest gains occurred under the most adverse conditions, -5 and 0 dB, whereas at 10 dB the mean benefit was more modest, suggesting a smaller margin for improvement when the input signal already presents a lower relative noise level.

Although STOI and PESQ also showed mean improvements, their variations were of smaller magnitude. Consistently, the effect size was large for SNR and SI-SDR, moderate for STOI, and small for PESQ, indicating that a reduction in error energy does not necessarily imply a proportional gain in intelligibility or objective perceptual quality. The 10 dB condition illustrates this behavior, since the variation in PESQ was not statistically significant in this case, reinforcing the importance of interpreting the metrics jointly.

From a methodological perspective, the organization of the experimental dataset, with speakers and noise sources kept independent across the training, validation, and test sets, together with the use of four controlled SNR conditions and the requirement that each contaminated signal contain at least one impulsive event, contributed to a more rigorous assessment of the network's generalization capability within the proposed scenario.

Despite the results obtained, some limitations should be considered. The scope of the study was restricted to impulsive noises associated with knocks, impacts, and door-related events, which prevents direct extrapolation of the observed performance to continuous noise or other acoustic scenarios. In addition, the LibriSpeech corpus consists of recordings derived from audiobooks, a condition that does not represent the full variability found in real-world communication situations. It is also important to consider that STOI and PESQ are objective metrics and therefore do not replace perceptual evaluations conducted with human listeners.

Finally, the system has not yet been directly evaluated in individuals with Central Auditory Processing Disorder (CAPD). Therefore, the results presented here exclusively reflect the objective performance of the method under the analyzed experimental conditions, and no clinical benefit or effective improvement in speech comprehension for individuals with CAPD can be inferred at this stage.

V. Conclusion

This study evaluated a one-dimensional U-Net applied to the enhancement of speech signals degraded by impulsive noise, while maintaining independent speakers and noise sources across the training, validation, and test sets. Across the 2,000 test samples, the model achieved mean gains of 3.21 dB in SNR and 3.00 dB in SI-SDR, with a positive SNR improvement in 99.9% of the cases. STOI and PESQ, on the other hand, although also presenting positive mean variations, showed considerably smaller gains, indicating that the error reduction observed through SNR and SI-SDR does not necessarily translate into a proportional perceptual improvement. Therefore, the results demonstrate the network's ability to attenuate impulsive noise under the evaluated experimental conditions, without, however, supporting claims of superiority over other architectures or allowing any anticipated clinical benefit to be inferred.

VI. Acknowledgments

The authors would like to express their gratitude for the guidance, availability, and fundamental contributions to the development of this work. They also thank their parents, who have always been present and supportive, especially Mr. Edvaldo Gomes and Mrs. Vera Lúcia, as well as their sisters, who have consistently provided affection and encouragement throughout this journey, especially Nayara Kellen. The authors also extend their thanks to their partners for their understanding, encouragement, and support during the different stages of this work.

The authors also acknowledge the Graduate Program in Biomedical Engineering at the University of Brasília (PPGEB/UnB) for its academic support and encouragement of research, as well as the Coordination for the Improvement of Higher Education Personnel (CAPES) for supporting academic training and scientific development.

Finally, the authors thank both longtime and new friends who were present throughout this journey, offering support, motivation, and words of encouragement during the most challenging moments.

In summary, the preparation of this article included the use of generative artificial intelligence tools for formatting-related tasks, including the organization of equations in LaTeX, the structuring of bibliographic entries in the `bibitem` format, and the positioning of figures within the text. The technical content, results, analyses, and conclusions are entirely the responsibility of the authors.

References

[1] X. Lin, Y. Zhang and S. Wang, "Mixed T-domain and TF-domain Magnitude and Phase Representations for GAN-based Speech Enhancement," *Scientific Reports*, vol. 14, art. 17698, 2024, doi: 10.1038/s41598-024-68708-w.

[2] H. R. Guimarães, J. Su, R. Kumar, T. H. Falk and Z. Jin, "DiTSE: High-Fidelity Generative Speech Enhancement via Latent Diffusion Transformers," *arXiv preprint arXiv:2504.09381*, 2025.

[3] S. S. Shetu, E. A. P. Habets and A. Brendel, "Leveraging Discriminative Latent Representations for Conditioning GAN-Based Speech Enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, 2026, doi: 10.1109/TASLPRO.2026.3677639.

[4] H. Zhao and N. Madhu, "Study of Lightweight Transformer Architectures for Single-Channel Speech Enhancement," in *Proc. 33rd European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, 2025, pp. 101–105, doi: 10.23919/EUSIPCO63237.2025.11226622.

[5] S. Pascual, A. Bonafonte and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3642–3646, doi: 10.21437/Interspeech.2017-1428.

[6] H. Chen and X. Zhang, "CGA-MGAN: Metric GAN Based on Convolution-Augmented Gated Attention for Speech Enhancement," *Entropy*, vol. 25, no. 4, art. 628, 2023, doi: 10.3390/e25040628.

[7] C. Macartney and T. Weyde, "Improved Speech Enhancement with the Wave-U-Net," *arXiv preprint arXiv:1811.11307*, 2018.

[8] D. Stoller, S. Ewert and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 334–340.

[9] American Speech-Language-Hearing Association, "Central Auditory Processing Disorder," *ASHA Practice Portal*. [Online]. Available: https://www.asha.org/Practice-Portal/Clinical-Topics/Central-Auditory-Processing-Disorder/.

[10] M. Klatte, T. Lachmann, and M. Meis, "Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting," *Noise & Health*, vol. 12, no. 49, pp. 270–282, 2010, doi: 10.4103/1463-1741.70506.

[11] M. Sharma, J. Purdy, and P. Kelly, "Effect of speech in noise training in the auditory and cognitive skills in children with auditory processing disorders," *International Journal of Pediatric Otorhinolaryngology*, vol. 146, art. 110735, 2021, doi: 10.1016/j.ijporl.2021.110735.

[12] S. Yecchuri and S. D. Vanambathina, "Sub-convolutional U-Net with Transformer Attention Network for End-to-End Single-Channel Speech Enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 8, pp. 1–18, 2024, doi: 10.1186/s13636-024-00331-z.

[13] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964. Available: https://www.openslr.org/12.

[14] C. Medina and R. Coelho, "Impulsive noise detection for intelligibility and quality improvement of speech enhancement methods applied in time-domain," *arXiv preprint arXiv:1910.02710*, 2019. Available: https://arxiv.org/abs/1910.02710.

[15] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013, ISBN: 978-1-4665-0422-6, doi: 10.1201/b14529.

[16] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-Baked or Well Done?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 626–630, doi: 10.1109/ICASSP.2019.8683855.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011, doi: 10.1109/TASL.2011.2114881.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)–A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2001, pp. 749–752, doi: 10.1109/ICASSP.2001.941023.

Transfer Learning for Dental Pathology Detection in Panoramic Radiographs: YOLOv11 vs. Zero-Shot Grounding DINO

1st Ernane Ferreira Rocha Junior
Dept. of Computer Engineering and Automation
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
ernane.ferreira.100@ufrn.edu.br

2nd Ignacio Sánchez-Gendriz
Northeast Strategic Technologies Center
CETENE / MCTI
Recife, PE, Brazil
ignacio.gendriz@cetene.gov.br

3rd Yoandris González Sánchez
Fundación Odontológica Social Luis Seiquer
Seville, Spain
yoandris.gonzalez81@gmail.com

4th Luiz Affonso Guedes
Dept. of Computer Engineering and Automation
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
affonso@dca.ufrn.br

Abstract*—Panoramic dental radiography is the most common supplementary imaging exam in dental practice, yet automated detection systems for it require annotated training data that is scarce and expensive to obtain. Zero-shot vision-language models offer an annotation-free path, but how accurately they localize pathologies on domain-specific images is not well characterized. This work compares Grounding DINO, a widely used open-vocabulary object detector, against a YOLOv11m model fine-tuned on 474 of the 678 annotated disease images from the DentexChallenge 2023 public dataset (950 after augmentation), targeting three pathologies: Caries, Periapical lesion, and Impacted tooth. On the DentexChallenge test split ($n = 103$), Grounding DINO achieved mAP@50 = 0.00 for all three classes across a range of prompt formulations, model sizes, and IoU thresholds down to 0.10, while the fine-tuned YOLOv11m reached mAP@50 = 0.557 (mAP@50:95 = 0.361), with Impacted tooth at 0.931 and Caries at 0.551. A three-stage pipeline was then applied to 50 externally sourced OPGs: YOLOv11 detection (466 total detections, mean 9.3 per image), spontaneous recall evaluation against dentist-written clinical descriptions (mean SR = 89.7% across 34 evaluable images), and LLM-based structured pre-clinical report generation via Gemini 2.5 Flash (BERTScore F1 = 0.779 raw, -0.309 rescaled, on 50 reports). An input ablation isolating the contribution of the image versus the detection summary to report text found that neither input alone was clearly superior on text-similarity metrics, a result we attribute to the brevity of the reference descriptions rather than to the image being uninformative. The results show that the evaluated zero-shot detector (Grounding DINO Tiny) is insufficient for clinical OPG analysis at standard IoU thresholds, and that supervised fine-tuning on a small public dental dataset produces a usable detection baseline.

***Index Terms*—panoramic radiography, dental pathology detection, YOLOv11, Grounding DINO, zero-shot detection, transfer learning, DentexChallenge, automated report generation**

I. Introduction

Panoramic dental radiography (orthopantomography, OPG) produces a single wide-field exposure covering the entire dentition, mandible, maxilla, and surrounding structures. It is the standard first-line imaging request in dental clinical practice, used for screening caries, periapical lesions, bone loss, impacted teeth, retained roots, and orthodontic findings [1]. Systematic interpretation takes time and requires experience, and access to dental radiologists varies considerably across healthcare settings.

Automated detection systems based on convolutional neural networks and, more recently, transformer architectures have shown promising results on specific dental imaging tasks: Tuzoff et al. [2] applied Faster R-CNN to tooth detection and numbering in OPGs, Ekert et al. [3] used a U-Net for periapical lesion detection, and Alqhtani et al. [4] benchmarked multiple YOLO variants against 17 fine-grained dental anomalies, reporting mAP@50 between 0.54 and 0.78 depending on pathology. All these approaches depend on large, fully annotated bounding box datasets, which take considerable clinical expertise to produce.

A possible path around that bottleneck comes from vision-language models (VLMs) trained at internet scale. These models learn open-vocabulary visual representations that can be queried at inference time using text prompts, without any task-specific fine-tuning [5]. Grounding DINO [6], presented at ECCV 2024, extends the DINO transformer architecture with grounded pre-training and reported 52.5 AP on COCO zero-shot at publication. Deploying such a model directly on OPGs, prompted with disease names and without annotation effort, is an appealing idea.

This work tests whether that idea holds in practice. Specifically: at a standard IoU threshold of 0.5, does Grounding DINO detect Caries, Periapical lesion, and Impacted tooth at a useful level of precision? If not, what does supervised fine-tuning on the largest available public dental dataset achieve on the same task, and with what training resources?

Grounding DINO achieved mAP@50 = 0.00 on all three classes. YOLOv11m fine-tuned on 474 DentexChallenge training images (950 after augmentation) reached mAP@50 = 0.557 overall, with per-class results that reflect training data imbalance directly. A complete end-to-end pipeline then applied those detections to the 50 external OPGs and fed the output to a multimodal LLM to generate structured pre-clinical reports, evaluated against dentist-written reference descriptions using standard NLP metrics. Code and data splits are publicly available at https://github.com/ErnaneJ/orthopantomography.

II. Related Work

A. Supervised Dental Pathology Detection

Convolutional detectors applied to OPGs generally report mAP in the 0.50–0.80 range for well-represented classes. Alqhtani et al. [4] found that performance varies sharply by pathology, with visually distinctive findings such as impactions and large restorations outperforming subtle ones such as early caries and periodontal bone loss by 20–30 mAP points within the same model, a per-class pattern that also appears in our results. The DentexChallenge 2023 [7], [8], a MICCAI workshop challenge, provided a public benchmark with disease-level bounding box annotations across three pathology classes, used here as both training source and evaluation benchmark.

B. Zero-Shot and Open-Vocabulary Detection

Grounding DINO [6] uses a Swin Transformer image encoder paired with a BERT-based text encoder, fused through a feature enhancer, to produce open-vocabulary bounding box predictions from text prompts. Prior work applying CLIP [5] to zero-shot classification in medical imaging found that general-purpose vision-language representations transfer to clinical domains for coarse classification, but noted performance gaps relative to supervised models [9]; localization at IoU ≥ 0.5 is substantially harder than classification, and our experiments confirm a larger gap here.

C. Automated Radiology Report Generation

Large multimodal language models have been applied to chest radiograph reporting [10], with evaluation using BLEU [11], ROUGE-L [12], and BERTScore [13]. BERTScore F1, computed from contextual token embeddings, correlates more reliably with human quality ratings than n-gram metrics when reference and hypothesis texts differ substantially in length and style [13]. That is the case in our Stage 3 evaluation. Dasanayaka et al. [14] applied a multimodal LLM to OPG report generation and question answering; the 50 externally sourced OPGs evaluated in our Stage 3 (Section III-A) are drawn from the public Apache License 2.0 subset of their DPT Image and Caption Dataset [15]. Our task decomposition and evaluation protocol still differ from theirs (three separate stages of detection, recall, and report generation, evaluated against independently produced dentist descriptions, versus their single end-to-end captioning model evaluated against the dataset's original captions), so results are not directly comparable, even though the source imagery overlaps. Balel et al. [16] trained a detector on 30,954 panoramic radiographs, converted the output to structured JSON, and generated reports with several locally hosted and commercial LLMs evaluated by specialists; the commercial models tested, including a Gemini variant, hallucinated findings in every report reviewed. That result, at almost two orders of magnitude more training images than used here, cautions against treating LLM-generated dental reports as reliable without the specialist review step we identify as a limitation and future work below.

III. Materials and Methods

Fig. 1 summarizes the three-stage pipeline described in this section: YOLOv11 detection (Stage 1, Section III-B), spontaneous recall evaluation against dentist descriptions (Stage 2, Section III-C), and LLM-based report generation (Stage 3, Section III-D).

A. Datasets

DentexChallenge 2023 (training and evaluation). The DentexChallenge 2023 quadrant-enumeration-disease subset [7], [8] is the largest publicly available OPG dataset with instance-level disease annotations (CC0 license). The full release contains 1,154 images; of these, 678 contain at least one bounding box annotation for the four disease categories used here: Caries, Deep Caries (merged with Caries in preprocessing), Periapical Lesion, and Impacted tooth. The remaining images have only quadrant or tooth-enumeration labels and were excluded from training and evaluation. The class distribution is skewed: across the 678 annotated images there are 2,767 Caries instances, 604 Impacted tooth instances, and 158 Periapical lesion instances (3,529 total). Images were split 70/15/15 into train ($n = 474$), validation ($n = 101$), and test ($n = 103$) sets using stratified random sampling (seed = 42). Per-split instance counts (Caries / Periapical lesion / Impacted tooth) are 1,934 / 111 / 458 in train, 408 / 25 / 77 in validation, and 425 / 22 / 69 in test. Images containing Periapical lesion or Impacted tooth annotations in the training split received two additional augmented copies each (Section III-B), yielding 950 effective training images and per-class training totals of 3,942 / 333 / 1,374 after augmentation.

External OPG dataset (pipeline evaluation). Fifty panoramic radiograph images (JPEG format, mean resolution approximately 2976×1488 pixels) were drawn from the publicly released subset of the DPT Image and Caption Dataset accompanying Dasanayaka et al. [14], [15], made available under the Apache License 2.0. The original authors report obtaining ethical clearance and informed consent for this data from Teaching Hospital Peradeniya and the Postgraduate Institute of Science, University of Peradeniya (Sri Lanka), with personally identifiable information removed prior to release (see Ethical Statement). Independent written clinical descriptions for these 50 images were subsequently produced by a trained dentist, co-author Y. González Sánchez, at the Fundación Odontológica Social Luis Seiquer (Seville, Spain); thirteen of these additionally have an audio recording of the verbal report. These independently produced

descriptions, not the original dataset captions, serve as the reference text for the Stage 2 and Stage 3 evaluations below. Bounding box annotations are not available for this dataset; it was used exclusively for the Stage 2 recall and Stage 3 report generation evaluations.

B. Stage 1: Detection Model

Preprocessing and augmentation. Images were resized to 640×640 pixels using letterbox padding. Training augmentation included CLAHE (clip limit 3.0) for X-ray contrast normalization, horizontal flip ($p = 0.5$), random brightness and contrast perturbation (± 0.2, $p = 0.5$), Gaussian noise ($p = 0.3$), and affine transformations (scale 0.9–1.1, translate $\pm 5\%$, rotate ± 5, $p = 0.4$), implemented with Albumentations. Color augmentations (HSV hue and saturation) were disabled because OPGs are grayscale. Images containing Periapical lesion or Impacted tooth annotations received two additional augmented copies each, tripling their effective training frequency.

YOLOv11m fine-tuning. YOLOv11m (medium variant, approximately 20M parameters) was initialized from COCO-pretrained weights and fine-tuned on the DentexChallenge training split using the Ultralytics framework [17]. Hyperparameters: AdamW optimizer, initial learning rate 10^{-3}, linear decay to 1% over training, batch size 8, image size 640×640, up to 100 epochs with early stopping (patience = 20). Training ran on an Apple M5 MacBook Air (16 GB unified memory) using the Metal Performance Shaders backend. The model reached best validation mAP@50 = 0.547 at epoch 27 and stopped at epoch 41. Table I results (Section IV-A) follow the standard Ultralytics mAP evaluation protocol, which sweeps a low confidence threshold (≥ 0.001) to construct the full precision-recall curve; the operational threshold used for the qualitative Stage 1 output on the external OPGs (Section IV-B) and for the fixed-decision comparison against Grounding DINO in Fig. 2 is ≥ 0.25, which trades recall for precision at deployment time. These are two different, individually standard use cases for the same trained model and should not be read as inconsistent thresholds for the same number.

YOLOv11n comparison. A smaller YOLOv11n variant (nano, 2.62M parameters, 181 layers, 6.6 GFLOPs) was trained under the same protocol for 50 epochs, reaching best validation mAP@50 = 0.484 at epoch 32. The gap to YOLOv11m's 0.547 reflects the capacity difference between the two variants; YOLOv11m was used for all downstream pipeline stages, and YOLOv11n is reported here only as a lightweight-deployment reference point.

Grounding DINO zero-shot baseline. Grounding DINO Tiny (`IDEA-Research/grounding-dino-tiny`, Hugging Face Transformers) was evaluated on the same test split without fine-tuning. The three class names were used as text prompts. Detection thresholds: box confidence ≥ 0.25, text score ≥ 0.20. Predictions were matched to ground-truth boxes at IoU ≥ 0.5, consistent with COCO evaluation. Section V-A reports a broader ablation over prompt formulations, model size (Tiny vs. Base), and IoU threshold to test whether this single configuration understates the model's true performance.

C. Stage 2: Spontaneous Recall Evaluation

The 50 external OPGs were processed with the fine-tuned YOLOv11m. For each image, the corresponding dentist description was parsed to identify mentions of the three YOLO-trained classes through substring matching. The Spontaneous Recall (SR) score for image i is:

$$SR(i) = \frac{|\hat{C}(i) \cap C^*(i)|}{|C^*(i)|} \tag{1}$$

where $\hat{C}(i)$ is the set of classes detected by YOLOv11 and $C^*(i)$ is the set of classes mentioned in the dentist description, both restricted to the three trained classes. The detector processes each image with a fixed confidence threshold of 0.25 and receives no information from the reference text. Images for which no YOLO-trained class appeared in the description were excluded from SR statistics (16 images), leaving 34 evaluable images.

Mentions of $C^*(i)$ are extracted from the free-text description by substring matching against a keyword-to-class alias dictionary, written in English to match the dentist descriptions. The dictionary has 31 distinct alias keys mapping to 24 clinical classes; 8 keys affect the SR metric reported here: *caries* and *carious* $\rightarrow$ Caries; *periapical* $\rightarrow$ Periapical lesion; *impacted* and *impaction* $\rightarrow$ Impacted tooth. A manual audit of the 20 images initially excluded found three misspellings the original dictionary missed, *carie*, *careis*, and *imapcted*, all present verbatim in the dentist's free-text notes; these were added as additional aliases to Caries and Impacted tooth respectively, which is why the excluded count above is 16 rather than 20. The remaining 23 keys match disease and treatment terms outside the three-class detection scope and do not influence SR.

D. Stage 3: Automated Report Generation

Structured pre-clinical reports were generated for all 50 external OPGs using Gemini 2.5 Flash via the OpenRouter API. Each request contained: (1) the OPG image in base64 encoding, and (2) a summary of detections from Stage 1 (class, confidence, and spatial location for the top 15 detections). The reference clinical description was not included in the prompt. The model was instructed to produce a five-section structured report using FDI tooth notation and to report only findings visible in the image. All requests used the OpenRouter-routed `google/gemini-2.5-flash` alias, queried in June and August 2026; OpenRouter does not expose a fixed dated snapshot for this alias, so results reflect whichever backend revision Google had live at request time, a reproducibility constraint of API-hosted models that we did not control.

Generated reports were compared to dentist-written reference descriptions using BLEU-4 [11], ROUGE-L [12], and BERTScore F1 [13] with RoBERTa-large embeddings.

To isolate the contribution of each input modality, the same 50 cases were regenerated under two additional conditions

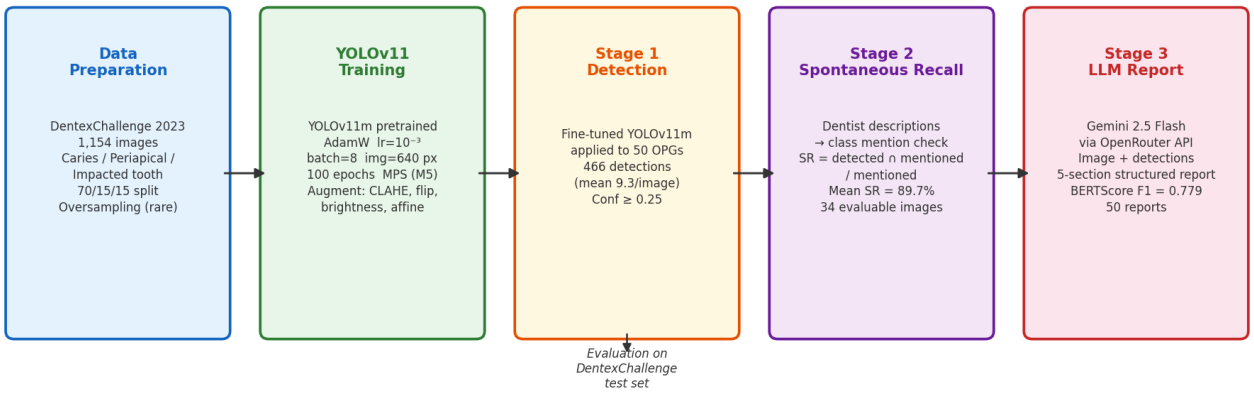


Fig. 1. Three-stage pipeline: YOLOv11 detection (Stage 1), spontaneous recall evaluation against dentist descriptions (Stage 2), and LLM-based report generation via Gemini 2.5 Flash (Stage 3).

with the LLM, prompt structure, and evaluation held fixed: *image-only* (the OPG image with no detection summary) and *detections-only* (the Stage 1 detection summary as text with no image). Results are reported below (Section IV-D) alongside the default *both* configuration described above.

IV. Results

A. Zero-Shot vs. Fine-Tuned Detection

Table I shows detection results on the DentexChallenge 2023 test split ($n = 103$ images), the only evaluation set with ground-truth bounding boxes.

TABLE I
DETECTION PERFORMANCE ON DENTEXCHALLENGE 2023 TEST SET (IOU = 0.5)

Model	Car.	Peri.	Imp.	mAP@50
GDINO-Tiny (zero-shot)	0.000	0.000	0.000	0.000
YOLOv11m (fine-tuned)	0.551	0.189	0.931	0.557

Car. = Caries; Peri. = Periapical; Imp. = Impacted. YOLOv11m evaluated at conf $\geq$ 0.001 (standard Ultralytics protocol, full PR curve); GDINO-Tiny at conf $\geq$ 0.25, text $\geq$ 0.20 (Section III-B). Per-class mAP@50:95, P/R, and bootstrap CI reported in text below.

Grounding DINO produced bounding box outputs when queried with the three class names, but none reached IoU $\geq$ 0.5 with any ground-truth box across the test split. The model generates text-anchored candidate regions that do not align closely enough with annotated pathology boundaries for clinical localization; Section V-A shows this holds across prompt formulations, model size, and looser IoU thresholds.

The fine-tuned YOLOv11m reached mAP@50 = 0.557 (bootstrap 95% CI [0.501, 0.645], Section IV; mAP@50:95 = 0.361), with Precision = 0.583 and Recall = 0.550 at the confidence ($\approx$0.30) that maximizes mean F1 across classes on the confidence sweep, not at the fixed conf $\geq$ 0.25 deployment threshold used elsewhere in this paper. Fig. 2 shows the per-class comparison. Per-class results track the training data distribution: Impacted tooth (AP@50 = 0.931) has a distinctive radiographic appearance and strong training coverage relative to its class size; Caries (AP@50 = 0.551) is the most prevalent pathology but also the most morphologically variable; Periapical lesion (AP@50 = 0.189) had only 158 annotated training instances.

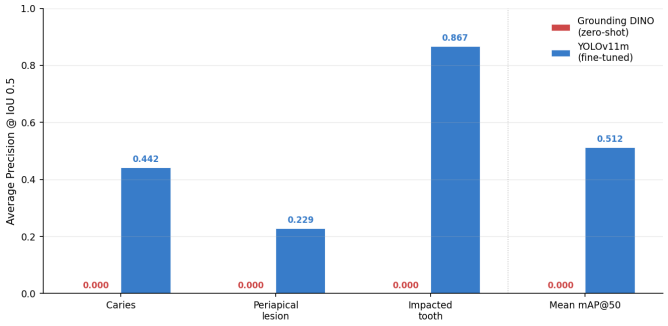


Fig. 2. Per-class AP@50 for Grounding DINO (zero-shot) and YOLOv11m (fine-tuned) on the DentexChallenge test set, both evaluated at the fixed conf $\geq$ 0.25 operational decision threshold with a custom 11-point AP implementation (Section III-B); distinct from the confidence-sweep values in Table I.

B. Detection on the External OPGs (Stage 1)

Applied to the 50 external OPGs, the fine-tuned YOLOv11m produced 466 total detections (mean 9.3 ± 3.2 per image, range 3–20). By class: Caries 387 (83.0%), Impacted tooth 59 (12.7%), Periapical lesion 20 (4.3%). This distribution is close to the class balance of the training data itself (78.4% Caries, 17.1% Impacted tooth, 4.5% Periapical lesion of 3,529 annotated instances, Section III-A), which is the more direct explanation than a general claim about prevalence in unrelated clinical populations; the low Periapical count also follows from the model's weaker AP for that class.

Fig. 3 shows a representative Stage 1 output on one of the external OPGs. Multiple Caries detections appear bilaterally in the posterior sectors, a single Impacted tooth is flagged

in the upper left third molar region, and a single Periapical lesion is flagged at low confidence (0.31) in an adjacent lower posterior region, consistent with that class's overall lower AP@50. The color scheme follows the application convention: red for disease-category findings (Caries, Periapical lesion) and green for tooth-status findings (Impacted tooth).

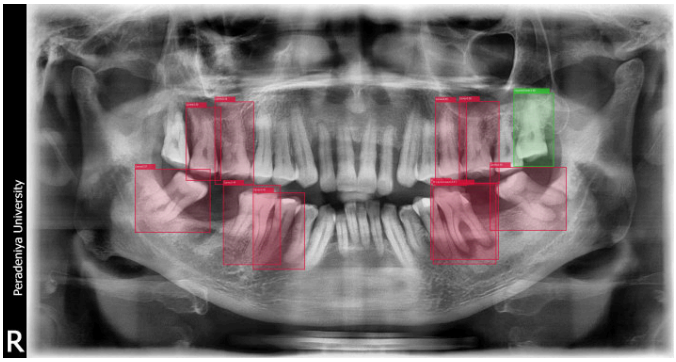


Fig. 3. Representative Stage 1 detection output on one of the external OPGs. Red bounding boxes indicate Caries and Periapical lesion detections (disease-category findings, bilateral posterior sectors); the green bounding box marks an Impacted tooth in the upper left third molar region. Confidence scores are displayed on each label; the Periapical lesion box has the lowest confidence (0.31), consistent with that class's overall lower AP@50.

C. Spontaneous Recall (Stage 2)

Of the 50 external OPGs, 34 had dentist descriptions mentioning at least one YOLO-trained class. Mean SR across those 34 images was 89.7% (SD = 29.2%; median = 1.000). Table II shows the distribution.

TABLE II
STAGE 2 SPONTANEOUS RECALL DISTRIBUTION ($n = 34$ EVALUABLE IMAGES)

SR band	n	%
SR = 0.0	3	8.8
$0 < \text{SR} < 0.5$	0	0.0
$0.5 \leq \text{SR} < 0.8$	1	2.9
$\text{SR} \geq 0.8$	30	88.2

Wilson 95% CI for SR $\geq$ 0.8: [73.4%, 95.3%].

Two of the three images with SR = 0.0 (Image_23, Image_25) mention only Impacted tooth as the testable class, and YOLOv11m returned no Impacted tooth detection for either despite that class's high aggregate AP@50 = 0.931 on DentexChallenge; the third (Image_48) mentions only Periapical lesion, consistent with that class's low AP@50 = 0.189. The first two cases show that a class's strong aggregate benchmark performance does not guarantee detection on every individual image. Given the small evaluable sample ($n = 34$), the Wilson 95% CI for the high-recall proportion is wide ([73.4%, 95.3%]); the point estimate of 88.2% should be read with that uncertainty in mind.

One case illustrates SR = 1.0 concretely: the dentist's description mentioned a vertically impacted lower right third molar (tooth 48, FDI notation), and the model detected an Impacted tooth in the same image. YOLOv11m also detected Caries in this image, but Caries was not among the classes the dentist mentioned and is therefore not evaluated by the SR metric for this image.

D. Report Generation Quality (Stage 3)

Gemini 2.5 Flash generated structured five-section preclinical reports for all 50 images without error (136,257 input tokens, 26,057 output tokens). Table III shows NLP metrics against dentist-written reference descriptions.

TABLE III
NLP EVALUATION OF GENERATED REPORTS VS. DENTIST REFERENCE DESCRIPTIONS ($n = 50$)

Metric	Mean	SD	Embeddings
BLEU-4	0.0010	0.0008	n-gram
ROUGE-L	0.0315	0.0177	n-gram
BERTScore F1 (raw)	0.7792	0.0129	RoBERTa-large
BERTScore F1 (rescaled)	-0.3085	0.0767	RoBERTa-large

Rescaled BERTScore [13] is normalized against a general-domain baseline distribution; negative values reflect that structured, domain-specific report pairs sit outside that baseline's typical range and are not by themselves evidence of poor semantic match (see text). Model: `roberta-large`, 17 layers, no IDF weighting (hash: `roberta-large_L17_no-idf_version=0.3.12(hug_trans=5.10.2)`).

Low BLEU-4 and ROUGE-L values are expected: generated reports average approximately 300 words across five structured sections while dentist references average approximately 45 words of free text, so exact n-gram overlap is low regardless of semantic accuracy. Raw BERTScore F1 = 0.779 alone would suggest substantial shared semantic content, but this needs a control: to test whether BERTScore discriminates a correct report from an incorrect one here, rather than saturating near a floor for any English text pair in this domain, we recomputed it with each report paired against a mismatched reference description (a fixed derangement of the 50 cases, seed = 42, so no case is paired with its own reference). The mismatched-pair mean was 0.7777 raw (-0.317 rescaled) against 0.7792 raw (-0.309 rescaled) for correctly matched pairs; the two distributions overlap almost completely (SD $\approx$ 0.013–0.014 raw for both), so neither raw nor rescaled BERTScore separates correct from randomly wrong pairings on this sample. This confirms, rather than merely anticipates, the concern raised about this metric: neither variant is evidence of clinical accuracy, or even of same-versus-different-image correctness, for these short technical texts. The concrete example below carries more evidentiary weight than the aggregate BERTScore number.

A concrete example illustrates this gap: for Image_00, the generated report identified impacted teeth 18 and 38 with correct angulation descriptions, listed multiple carious lesions in FDI notation with severity estimates, and recommended appropriate follow-up actions. The dentist's reference note for the same image read: "Grossly carious 48 and deep caries in 36 with pulp exposure. Septic root of 25." BLEU and ROUGE scores for this pair are near zero because the texts differ in structure and length; BERTScore F1 for this pair was 0.783

raw (-0.286 rescaled), reflecting that both texts describe the same pathological findings.

Table IV compares the default *both* configuration against the *image-only* and *detections-only* conditions described in Section III-D. Detections-only scored marginally higher than both other conditions on ROUGE-L and BERTScore F1 (raw and rescaled), while image-only scored lowest on every metric. BLEU-4 stayed near zero across all three and is not informative here for the reason given above. This pattern is discussed in Section V.

TABLE IV
INPUT ABLATION FOR STAGE 3 REPORT GENERATION ($n = 50$ PER CONDITION)

Condition	BLEU-4	ROUGE-L	BERT (raw)	BERT (resc.)
Image-only	0.0010	0.0304	0.7761	-0.3269
Detections-only	0.0015	0.0335	0.7956	-0.2113
Both (default)	0.0010	0.0315	0.7792	-0.3085

V. DISCUSSION

A. Localization Failure in Zero-Shot Detection

Grounding DINO is not blind to dental anatomy: when prompted with class names, it generates candidate boxes in plausible image regions. The failure is at localization precision, not semantic recognition. Standard mAP evaluation counts a detection as correct only if $\text{IoU}(\hat{b}, b^*) \geq 0.5$. For small lesions like periapical radiolucencies (typically a few millimeters in diameter at clinical scale), a box that roughly surrounds the affected region but is 30% too large fails this threshold entirely. Grounding DINO was trained to ground language to visual regions from natural image datasets, where objects are much larger and more salient than millimeter-scale dental pathologies. That spatial precision is simply not in what the model learned.

To test whether this failure is an artifact of prompt or model size, we ran the Tiny model under three prompt formulations (bare class names, descriptive phrases, full noun phrases) and added the Base model under the class-name prompt as a model-size check, each of these 4 prompt/model configurations evaluated at three IoU thresholds (0.50, 0.25, 0.10), for 12 conditions in total. mAP stayed at or near zero across all of them: the highest value observed at any condition, including IoU $\geq$ 0.10, was 0.015 (Tiny, class-name prompt, Periapical lesion); most conditions scored exactly 0.000 even at IoU $\geq$ 0.10, and Base scored no higher than 0.005 at any threshold, performing no better than Tiny. This rules out prompt engineering and model size as explanations: candidate regions do not consistently land inside the correct anatomical area at all, not merely outside the IoU $\geq$ 0.5 boundary.

We also tested whether the fixed box confidence threshold (0.25) was filtering out correctly located but low-confidence boxes, re-running the Tiny model with the class-name prompt at confidence thresholds 0.05, 0.10, 0.15, and 0.25, crossed with the same three IoU thresholds. At the loosest IoU ($\geq$ 0.10), lowering confidence from 0.25 to 0.05 nearly tripled mAP (0.015 $\rightarrow$ 0.041); at the clinically relevant IoU $\geq$ 0.5, mAP never exceeded 0.007 at any confidence tested. Confidence threshold matters only when the acceptance criterion already rewards approximately correct boxes; it does not rescue performance at the standard localization threshold, confirming the failure is in where boxes are placed, not how many are filtered before scoring.

In practice, this means that Grounding DINO, as evaluated here across both size variants and multiple prompt formulations, is not usable as a detection component in any pipeline that requires IoU-based localization on panoramic dental radiographs, even at thresholds well below the clinical standard. Whether other zero-shot or foundation detection architectures, or in-context/few-shot prompting strategies not tested here, would behave differently remains an open question. The result should be read as evidence against this specific family of models on this task, not as a general claim about zero-shot detection.

B. Transfer Learning from a Small Public Dataset

The mAP@50 = 0.557 from 474 training images (950 after augmentation) supports two conclusions. First, transfer from ImageNet/COCO pre-training generalizes to the X-ray domain at this dataset size when augmentation matches the modality (CLAHE, disabled color perturbation, oversampling of rare classes), consistent with recent evidence that YOLO-family transfer learning remains competitive with newer architectures on dental panoramic radiographs at small dataset scale [18]. Second, Impacted tooth at AP@50 = 0.931, a class among the most commonly missed findings in routine OPG review, is a promising starting point for computer-aided screening, pending clinical validation on larger and more diverse populations.

C. Spontaneous Recall as a Proxy Metric

The SR metric defined in (1) measures only whether the detector identifies at least one instance of each class the dentist mentioned: it does not measure spatial precision, does not penalize detections of classes not mentioned in the description (which may be genuine incidental findings or false positives), and does not validate bounding box accuracy. An image where the dentist wrote "caries present" and the detector produced ten spatially incorrect Caries boxes still scores SR = 1.0 for that class. SR is thus a class-level recall proxy, useful for characterizing pipeline behavior on the external OPGs where no bounding-box ground truth exists, but a necessary rather than sufficient condition for clinical utility; the high mean SR (89.7%) is consistent with the model's detection rates on DentexChallenge and should not be read as a clinical accuracy figure.

D. Periapical Lesion: A Data Problem

Periapical lesion AP@50 = 0.189, well below Caries (0.551) and Impacted tooth (0.931), follows partly from having only 111 annotated training instances (333 after augmentation), a third of the next-smallest class; periapical radiolucencies

are also small (2–10 mm), subtle in early stages, and easily confused with anatomical structures or noise, so some of the gap may reflect a harder visual task rather than data volume alone. This experiment cannot separate the two factors, since scaling annotated data while holding architecture fixed was not tested.

Asif and Khan [18] trained YOLO26l-seg on a Roboflow-processed DENTEX subset and reported mAP@50 = 0.591 across four disease classes, close to our 0.557 across three, but argue visual distinctiveness, not annotated volume, is the stronger per-class driver on this dataset. Our results do not settle this: Impacted tooth here has both the second-highest instance count and the most distinctive appearance, so volume and architecture are confounded in the same direction as in their setup. Separating them would require training on volume-matched subsets, which we did not do; we frame the data-volume account as a plausible partial explanation, not a demonstrated one.

Improving this class would require more training data, targeted augmentation, or semi-supervised learning from unlabeled OPGs, ideally evaluated against an architecture ablation to isolate how much of the gap is data-driven versus intrinsic to the detector.

E. Image and Detection Contributions to Report Text

Table IV shows that detections-only text matched the dentist's reference description at least as well as, and on ROUGE-L and BERTScore slightly better than, the full image-plus-detections configuration used elsewhere in this paper; image-only scored lowest on every metric. Taken at face value this could be misread as evidence that the image contributes nothing once a detection summary is available, but we attribute it to a property of the reference texts rather than the image: dentist reference descriptions average approximately 45 words and list only the findings judged clinically salient, so a report instructed to describe only the detection list stays closer in content and length to that short reference. The full *both* condition gives the model room to describe additional visually plausible findings (e.g., restorations or edentulous spaces) outside the three trained classes and outside what the reference mentions, which lowers overlap with a short reference without indicating a worse report.

This does not show that the image is unnecessary. It shows that BLEU-4, ROUGE-L, and BERTScore, which all score similarity to a short reference text, cannot by themselves settle whether visual grounding improves report quality, since a terser, more constrained output tends to score better against a terse reference regardless of clinical completeness. A more informative test would compare the three conditions against dentist judgment of completeness and correctness directly, i.e., the clinical review step identified as a limitation above. Practically, a text-only deployment tier, using the detection summary without re-transmitting the image to the LLM, is worth considering for cost and latency, provided a clinical reviewer confirms it does not omit visually apparent findings outside the detector's three classes.

F. Web Prototype

To assess compatibility with clinical workflows, a proof-of-concept web application was built on the three-stage pipeline: a FastAPI backend runs Stage 1 detection via YOLOv11m, calls Stage 3 Gemini 2.5 Flash, and stores the annotated image and report; a React frontend overlays detection boxes with a raw/annotated toggle, lists findings by class, displays the five-section report in tabbed form, and accepts voice input for clinician corrections.

End-to-end latency from image upload to a fully rendered annotated radiograph and structured pre-clinical report was approximately 15 seconds on an Apple M5 MacBook Air with a standard broadband connection to the OpenRouter API. The dominant cost is the Gemini API call, not the local YOLO inference, which completes in under one second.

The prototype confirms that the computational requirements of the pipeline are compatible with commodity hardware: no discrete or datacenter GPU is required at inference time, and YOLO inference runs locally via Apple's Metal Performance Shaders backend if the pipeline runs offline. This is a proof of concept for interactive deployment, not a clinically validated tool; a production system would require prospective clinical evaluation, regulatory compliance, and server-side inference at scale, outside the scope of this work.

G. Limitations

Four constraints limit these conclusions. First, quantitative detection evaluation relies on a single public benchmark (DentexChallenge 2023); performance on other equipment, populations, or annotation protocols may differ. Second, Stage 2 SR and Stage 3 NLP metrics are proxy evaluations on the external OPGs without bounding box ground truth: SR measures class-level agreement, not spatial precision or false-positive rate; BERTScore measures text similarity, not clinical correctness, and our random-pairing control (Section IV-D) shows it barely distinguishes a correct reference from a mismatched one here, so it carries little independent weight. The same limitation applies to the image-vs-detections ablation (Section IV-D): it measures similarity to a short reference, not clinical completeness, so it cannot alone determine whether the image is dispensable. Third, the generated reports were not reviewed by clinicians; a blinded expert evaluation is planned as immediate future work before any clinical utility claim can be made. Fourth, training and inference ran on an Apple M5 MPS backend, consumer hardware; throughput figures are a proof-of-concept reference, not a production benchmark.

VI. Conclusion

Grounding DINO, tested without fine-tuning across three prompt formulations and two model sizes, achieved mAP@50 = 0.00 (and near-zero even at IoU $\geq$ 0.10) on the Dentex-Challenge 2023 test set for all three target classes because its localization precision is insufficient for millimeter-scale dental pathologies. YOLOv11m fine-tuned on 474 annotated training images (950 after augmentation) from the same public dataset reached mAP@50 = 0.557 overall (95% CI [0.501, 0.645]),

with Impacted tooth at 0.931, Caries at 0.551, and Periapical lesion at 0.189. Applied to 50 externally sourced OPGs without bounding box ground truth, the pipeline produced a mean spontaneous recall of 89.7% (class-level agreement with dentist descriptions, Wilson 95% CI [73.4%, 95.3%] for the high-recall proportion) and structured pre-clinical reports with BERTScore F1 = 0.779 raw (-0.309 rescaled; semantic text similarity against reference notes, though a random-pairing control found this metric barely separates correct from mismatched report-reference pairs on this sample); all figures are exploratory proxies pending clinical validation.

Grounding DINO, evaluated zero-shot across both size variants and multiple prompts, was not suitable for clinical dental detection here, even at IoU thresholds well below the clinical standard; this should not be generalized to all zero-shot or foundation models without further evaluation. Supervised fine-tuning on a small public dataset, with augmentation tailored to the modality, produces a viable detection baseline. A React/FastAPI web prototype of the complete pipeline delivers a structured pre-clinical report from image upload in approximately 15 seconds on commodity hardware, with no discrete or datacenter GPU required at inference time.

Future work will focus on scaling the annotated training set, exploring semi-supervised approaches for Periapical lesion, extending detection to the full 31-class clinical vocabulary, and obtaining clinical validation of the generated reports by specialist dentists.

Ethical Statement

The DentexChallenge 2023 benchmark used for Stage 1 training and evaluation is released under a CC0 license and contains no identifiable patient information. The 50 external OPGs used for the Stage 2 and Stage 3 evaluations are drawn from the publicly available subset of the DPT Image and Caption Dataset [15] accompanying Dasanayaka et al. [14]. The original authors report that ethical clearance for this data was obtained from the relevant hospital ethics committees under international ethical standards, that informed consent was obtained from all subjects, and that the released images were de-identified prior to publication; the dataset is distributed under the Apache License 2.0. No additional patient contact, data collection, or identifiable information handling was performed by the present authors; our use is limited to secondary analysis of this already-published, de-identified, licensed dataset, together with independently authored clinical descriptions (see Section III-A). The authors declare no conflicts of interest.

Acknowledgment

The authors thank the DentexChallenge 2023 organizers for making the dataset publicly available under a CC0 license, and Dasanayaka et al. [14], [15] for releasing the DPT Image and Caption Dataset under the Apache License 2.0. Independent clinical descriptions and audio recordings for the 50 external evaluation images were produced by co-author Y. González Sánchez at the Fundación Odontológica Social Luis Seiquer (Seville, Spain). Code and evaluation artifacts are publicly available at https://github.com/ErnaneJ/orthopantomography.

References

[1] J. B. Ludlow, L. E. Davies-Ludlow, S. L. Brooks, and W. B. Howerton, "Dosimetry of 3 CBCT devices for oral and maxillofacial radiology: cone beam CT vs. panoramic X-ray," *Dentomaxillofacial Radiology*, vol. 35, no. 4, pp. 219–228, 2006.

[2] D. V. Tuzoff *et al.*, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," *Dentomaxillofacial Radiology*, vol. 48, no. 4, 2019.

[3] T. Ekert *et al.*, "Deep learning for the radiographic detection of apical lesions," *Journal of Endodontics*, vol. 45, no. 7, pp. 917–922, 2019.

[4] N. Alqhtani *et al.*, "Detection accuracy of an AI platform for dental treatment features on panoramic radiographs," *Scientific Reports*, vol. 15, 2025.

[5] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[6] S. Liu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, arXiv:2303.05499.

[7] "DentexChallenge 2023: MICCAI workshop on dental enumeration and diagnosis on panoramic X-rays," Available: https://dentex.grand-challenge.org/, 2023, CC0 license, accessed June 2026.

[8] I. E. Hamamci *et al.*, "DENTEX: Dental enumeration and tooth pathosis detection benchmark for panoramic X-ray," arXiv:2305.19112, 2023.

[9] M. Zhang *et al.*, "Exploring zero-shot anomaly detection with CLIP in medical imaging," arXiv:2411.09310, 2024.

[10] J. Liu *et al.*, "Automated structured radiology report generation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, arXiv:2505.24223.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[12] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020, arXiv:1904.09675.

[14] C. Dasanayaka, K. Dandeniya, M. B. Dissanayake, C. Gunasena, and R. Jayasinghe, "Multimodal AI and large language models for orthopantomography radiology report generation and Q&A," *Applied System Innovation*, vol. 8, no. 2, p. 39, 2025.

[15] C. Dasanayaka, "DPT image caption dataset," Available: https://huggingface.co/datasets/LexiconShiftInnovations/DPT_Image_Caption_Dataset, 2024, Apache License 2.0, accessed 15 May 2024.

[16] Y. Balel *et al.*, "A novel hybrid large language model approach for reporting panoramic radiographs and performance comparison with current large language models," *Journal of Imaging Informatics in Medicine*, 2026.

[17] G. Jocher and J. Qiu, "Ultralytics YOLO11," Available: https://github.com/ultralytics/ultralytics, 2024, MIT license.

[18] K. A. Asif and R. A. Khan, "Dental panoramic radiograph analysis using YOLO26: From tooth detection to disease diagnosis," arXiv:2604.16231, 2026.

Analysis and comparison of TRCA and CCA spatial filters for SSVEP-based BCIs

Caio Amaral
Department of Computer Engineering and Automation
School of Electrical and Computer Engineering
University of Campinas (UNICAMP)
Campinas, SP, Brazil
c165444@dac.unicamp.br

Rafael Ferrari
Department of Computer Engineering and Automation
School of Electrical and Computer Engineering
University of Campinas (UNICAMP)
Campinas, SP, Brazil
rafaelf@unicamp.br

***Abstract*—Steady-state visual evoked potential (SSVEP)-based brain-computer interfaces (BCIs) rely on spatial filtering algorithms to suppress background noise and interference. Traditionally, the literature evaluates the standard formulation of canonical correlation analysis (CCA) as a calibration-free, unsupervised method, contrasting it with the supervised task-related component analysis (TRCA). To establish a methodologically equitable comparison, this study formulates a supervised version of CCA and investigates the collinearity between its spatial weight vectors and those extracted by TRCA. Additionally, we analyze the structural impact of deploying empirical biological templates versus synthetic sinusoids within a unified filter-bank, multidimensional ensemble projection pipeline. Utilizing two public benchmark datasets, our spatial convergence analysis reveals that the electrode weight vectors optimized by supervised CCA and TRCA reach a highly collinear alignment for window size exceeding 0.7 s. In the decoding stage, a cross-substitution ablation study considering both single-band and multi-band filter-bank architectures demonstrates that configuring the ensemble supervised CCA with biological reference templates yields decoding accuracies and information transfer rates equivalent to those of standard ensemble TRCA, while outperforming traditional synthetic sinusoid configurations. These findings indicate that the performance superiority typically attributed to TRCA in the literature stems primarily from the structural integration of empirical biological templates, filter-bank and ensemble techniques, rather than an intrinsic mathematical advantage of its underlying spatial optimization objective function.**

***Index Terms*—Brain-computer interface, SSVEP, Spatial filter, Canonical Correlation Analysis (CCA), Task Related Component Analysis (TRCA).**

I. Introduction

Brain Computer Interfaces based on Steady State Visual Evoked Potentials (SSVEP) stand out in neuroengineering for providing high Information Transfer Rates (ITR) [1], [2]. However, the low Signal-to-Noise Ratio (SNR) inherent to the Electroencephalogram (EEG) requires spatial filtering algorithms. [3].

Historically, Canonical Correlation Analysis (CCA) is a recognized method that stands out because it does not require user calibration and relies on finding linear correlations between EEG channels and synthetic sine-cosine templates [4]. In contrast, Task Related Component Analysis (TRCA) has been consolidated as an approach that extracts filters maximizing signal reproducibility based on the covariance calibration data. When researchers evaluate these two methods in standard configuration, the literature shows TRCA outperforming CCA in classification accuracy [5].

TRCA implementations are typically evaluated as part of a comprehensive signal processing pipeline, incorporating sub-band decomposition (Filter Bank) [7] and multiple spatiotemporal projections (Ensemble) [3], [5]. These architectural enhancements, however, are mathematically independent of the underlying spatial filter. They can be integrated into a filter built upon linear correlation with synthetic references (CCA) as well as to a filter optimizing empirical signal reproducibility (TRCA).

In this context, the present study has a twofold objective. First, we conduct a similarity analysis to evaluate the equivalence between the spatial filters optimized by TRCA and supervised CCA. Second, we perform an ablation study, evaluating the various combinations of these spatial filters when integrated with complementary structural techniques, specifically Filter Bank and Ensemble methods. Through this approach, we aim to elucidate whether the overall classification performance depends predominantly on the specific mathematical formulation of the spatial filters, or if it is driven by the complementary structural signal processing techniques.

II. SSVEP Classification Pipeline

A general classification framework for SSVEP based BCIs is depicted in Figure 1. Classification methods based on CCA and TRCA can be described within this general framework. We consider a system designed to recognize N_c distinct visual stimuli, where each stimulus represents a unique target class $c \in \{1, 2, \ldots, N_c\}$. The classification task consists of inferring which of these N_c targets the user was focusing on during a given EEG time window.

The processing flow begins by separating both the input test signal $\mathbf{X} \in \mathbb{R}^{C \times N_s}$ (where C is the number of channels and

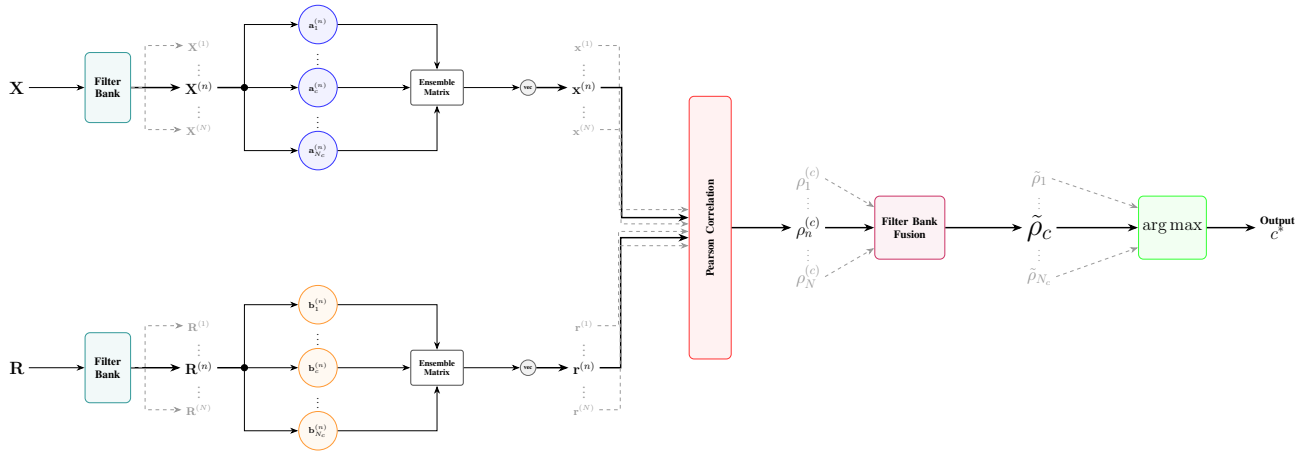


Fig. 1. Spatiotemporal processing pipeline diagram. The top path processes the test EEG signal $\mathbf{X}$, while the bottom path processes the reference template $\mathbf{R}$. Nodes $\mathbf{a}_c^{(n)}$ and $\mathbf{b}_c^{(n)}$ represent the class-specific spatial filters for the c-th stimulus.

N_s is the number of samples) and a class reference template $\mathbf{R}$ into N distinct sub-bands via Filter Bank. Depending on the adopted decoding paradigm, this reference $\mathbf{R}$ can represent either a synthetic sinusoidal matrix or an empirical biological mean. For each sub-band $n \in \{1, \ldots, N\}$, the multichannel data are projected through the class-specific spatial filters $\mathbf{a}_c^{(n)}$ and $\mathbf{b}_c^{(n)}$ and organized into their respective Ensemble matrices.

These multidimensional projections are then flattened by the vectorization (vec) operators to generate the structured vectors $\mathbf{x}^{(n)}$ and $\mathbf{r}^{(c,n)}$ for a evaluated candidate class c. The similarity between these vectorized test and template responses is quantified through Pearson correlation, yielding an individual coefficient $\rho_c^{(n)}$ for each sub-band and class combination.

The Filter Bank Fusion block then aggregates these sub-band metrics for each class independently, producing a single fused similarity score $\tilde{\rho}_c$. In the final stage of the pipeline, the system evaluates the entire set of accumulated class scores $[\tilde{\rho}_1, \tilde{\rho}_2, \ldots, \tilde{\rho}_{N_c}]$ through an $\arg\max$ operation. The target class that maximizes this final fused correlation coefficient is selected as the system decision C^*:

$$C^* = \arg\max_c \tilde{\rho}_c \tag{1}$$

A. Filter Bank

To effectively capture the harmonic components of SSVEP signals, we implement a Filter Bank (FB) architecture. Originally proposed for the FB-CCA method [7] and later adapted for FB-TRCA [5], this approach overcomes a fundamental limitation in standard feature extraction.

When a subject gazes at a flickering target, the visual cortex responds not only at the fundamental stimulation frequency (f), but also at its harmonic integer multiples. While incorporating these upper harmonics into the analysis is essential, their signal-to-noise ratio decays sharply at higher frequencies. Consequently, extracting harmonic features through a single, broad frequency band is inefficient, as these upper components are easily masked by background noise [7].

Within the decoding pipeline depicted in Figure 1, the Filter Bank serves as the primary preprocessing layer. Given an input multichannel signal $\mathbf{Z} \in \mathbb{R}^{C \times N_s}$, representing either the test trial $\mathbf{X}$ or the reference template $\mathbf{R}$, the architecture decomposes it into N sub-band components. For the n-th sub-band ($n \in \{1, 2, \ldots, N\}$), the filter outputs a band-limited signal $\mathbf{Z}^{(n)} \in \mathbb{R}^{C \times N_s}$, retaining the same spatial and temporal dimensions.

Following the design established in [7], the first sub-band covers the fundamental frequency and all subsequent harmonics, the second skips the fundamental to start at the first harmonic and captures everything above it, and so on. This separation forces the target detection algorithms to operate strictly within the spectral regions where the harmonic energy is concentrated.

For each of these N sub-bands, the system independently calculates a similarity coefficient, $\rho_c^{(n)}$, representing the correlation strength for that specific frequency band and candidate class c. To reach a consolidated similarity measure, these individual sub-band coefficients are merged using the linear fusion rule [5]

$$\tilde{\rho}_c = \sum_{n=1}^{N} w_n \cdot \text{sign}(\rho_c^{(n)}) \cdot \left(\rho_c^{(n)}\right)^2, \tag{2}$$

where w_n acts as a scaling factor for each sub-band coefficient, while the sign$(\cdot)$ operator preserves the algebraic sign of the correlation. Squaring $\rho_c^{(n)}$ highlights the peaks associated with the true target class [5].

Because higher-order harmonic components carry less spectral energy, treating every sub-band equally is suboptimal. Therefore, the weights w_n decay mathematically as the sub-band index increases, adopting the empirical formulation proposed in [7]

$$w_n = n^{-a} + b. \tag{3}$$

Following established BCI literature, our experiments evaluate $N = 5$ sub-bands, with the weighting parameters locked at $a = 1.25$ and $b = 0.25$ [7].

B. Ensemble

To boost classification performance, the ensemble technique projects the signals into a multidimensional subspace generated by the optimized spatial filters of all available classes [3], [5]. As illustrated by the parallel spatial filter groups in Figure 1, this method accommodates generalized filter pairs: vectors $\mathbf{a}_c^{(n)}$ applied to the incoming test signal, and vectors $\mathbf{b}_c^{(n)}$ applied to the reference template. We build the respective ensemble matrices, $\mathbf{W}_A^{(n)}$ and $\mathbf{W}_B^{(n)}$, by stacking these individual filter vectors obtained across all N_c classes (stimuli)

$$\begin{aligned}\mathbf{W}_A^{(n)} &= [\mathbf{a}_1^{(n)}, \mathbf{a}_2^{(n)}, \ldots, \mathbf{a}_{N_c}^{(n)}],\\ \mathbf{W}_B^{(n)} &= [\mathbf{b}_1^{(n)}, \mathbf{b}_2^{(n)}, \ldots, \mathbf{b}_{N_c}^{(n)}].\end{aligned} \tag{4}$$

The diagram inf Figure 1 highlights a symmetric processing route for this projection stage. Along the top pathway, the raw test signal $\mathbf{X}^{(n)}$ is mapped directly into its subspace, yielding the feature matrix

$$\tilde{\mathbf{X}}^{(n)} = (\mathbf{W}_A^{(n)})^T \mathbf{X}^{(n)} \in \mathbb{R}^{Nc \times N_s}. \tag{5}$$

The system processes a reference signal in the same way for each target class c along the bottom pathway. This architecture is independent of the nature of the template. The reference $\mathbf{R}^{(c)}$ can be an empirical biological mean derived from calibration data (as in TRCA), a synthetic sinusoidal matrix (as CCA) or any other template signal. The template is projected as

$$\tilde{\mathbf{R}}^{(c,n)} = (\mathbf{W}_B^{(n)})^T \mathbf{R}^{(c,n)} \in \mathbb{R}^{Nc \times N_s}. \tag{6}$$

At this point, the projections $\tilde{\mathbf{X}}^{(n)}$ and $\tilde{\mathbf{R}}^{(c,n)}$ remain two-dimensional spatiotemporal responses. To calculate the Pearson correlation, the system must evaluate both dimensions simultaneously by applying a vectorization operator [3], visually denoted by the vec$(\cdot)$ nodes in Figure 1. This operator flattens the projected matrices by sequentially stacking their columns, converting the two-dimensional response into a single column vector $\mathbf{v} \in \mathbb{R}^{(Nc \cdot N_s) \times 1}$,

$$\mathbf{x}^{(n)} = \text{vec}(\tilde{\mathbf{X}}^{(n)}), \quad \mathbf{r}^{(c,n)} = \text{vec}(\tilde{\mathbf{R}}^{(c,n)}). \tag{7}$$

Finally, we quantify the similarity between the test signal and class c within sub-band n using the Pearson correlation coefficient ($\rho_c^{(n)}$). Mathematically, this metric corresponds to the ratio between the sample covariance of the flattened vectors and the product of their standard deviations [5]

$$\begin{aligned}\rho_c^{(n)} &= \frac{\text{cov}\left(\mathbf{x}^{(n)}, \mathbf{r}^{(c,n)}\right)}{\sigma_{\mathbf{x}^{(n)}}\, \sigma_{\mathbf{r}^{(c,n)}}}\\ &= \frac{\sum_{i=1}^{N_c \cdot N_s} \left(x_i^{(n)} - \bar{x}^{(n)}\right)\left(r_i^{(c,n)} - \bar{r}^{(c,n)}\right)}{\sqrt{\sum_{i=1}^{N_c \cdot N_s} \left(x_i^{(n)} - \bar{x}^{(n)}\right)^2} \sqrt{\sum_{i=1}^{N_c \cdot N_s} \left(r_i^{(c,n)} - \bar{r}^{(c,n)}\right)^2}},\end{aligned} \tag{8}$$

where $x_i^{(n)}$ and $r_i^{(c,n)}$ represent the i-th elements of the flattened vectors $\mathbf{x}^{(n)}$ and $\mathbf{r}^{(c,n)}$, while $\bar{\mathbf{x}}^{(n)}$ and $\bar{\mathbf{r}}^{(c,n)}$ represent their respective sample means.

Once these coefficients are extracted across all N sub-bands for a specific class c, the Filter Bank Fusion module consolidates them into a single aggregated score $\tilde{\rho}_c$. Finally, the target class C^* that yields the maximum aggregated correlation is selected according to (1).

III. Design of the Spatial Filters

The CCA and the TRCA are two methodologies that can be used to design spatial filters for extracting the evoked potential information from the EEG signals. The following sections describe these two techniques.

A. Canonical Correlation Analysis (CCA)

CCA is a multivariate statistical method that evaluates the linear relationship between two sets of random variables, making it effective for SSVEP decoding [4]. In its original formulation, the algorithm aims to find an optimal pair of linear combinations that maximize the correlation between the recorded EEG signal $\mathbf{X}$ and a set of synthetic reference signals. For any given stimulus class c, the artificial reference matrix $\mathbf{Y}_c \in \mathbb{R}^{2N_h \times N_s}$ is defined as

$$\mathbf{Y}_c = \begin{bmatrix} \sin(2\pi f_c t) \\ \cos(2\pi f_c t) \\ \vdots \\ \sin(2\pi N_h f_c t) \\ \cos(2\pi N_h f_c t) \end{bmatrix}, \tag{9}$$

where f_c denotes the fundamental stimulation frequency of class c, $(N_h - 1)$ represents the number of considered harmonics, and $t = \left[\frac{1}{f_s}, \frac{2}{f_s}, \ldots, \frac{N_s}{f_s}\right]$ is the time vector dependent on the sampling frequency f_s.

The objective is to determine the spatial weight vectors $\mathbf{w}_X$ and $\mathbf{w}_Y$ that maximize the Pearson correlation between the canonical variables $\mathbf{w}_X^T\mathbf{X}$ and $\mathbf{w}_Y^T\mathbf{Y}$ [4],

$$(\mathbf{w}_X, \mathbf{w}_Y) = \arg\max_{\mathbf{u}_X, \mathbf{u}_Y} \frac{\mathbf{u}_X^T \mathbf{X}\mathbf{Y}^T \mathbf{u}_Y}{\sqrt{\mathbf{u}_X^T \mathbf{X}\mathbf{X}^T \mathbf{u}_X \cdot \mathbf{u}_Y^T \mathbf{Y}\mathbf{Y}^T \mathbf{u}_Y}}. \tag{10}$$

In [4], this optimization problem must be solved for each class in every incoming window of EEG, without relying on labeled calibration data. The system computes a unique set of filters for each target frequency, and the class that achieves the highest maximum canonical correlation is selected as the intended target.

However, the CCA can be used to optimize the spatial filters in a different way, exploring labeled calibration data. In this setup, which will be called supervised CCA hereinafter, instead of solving the optimization problem during online testing, the spatial filters are computed during a dedicated training stage using labeled calibration trials, i.e., trials in which the target stimulus is known.

This supervised framework maps directly onto the modular architecture illustrated in Figure 1. During calibration, the

training trials corresponding to class c are used to solve (10). This optimization determines $\mathbf{w}_X$, which corresponds directly to the spatial filter $\mathbf{a}_c^{(n)}$ operating along the test pathway, and $\mathbf{w}_Y$, which corresponds directly to the filter $\mathbf{b}_c^{(n)}$ operating along the template path in Figure 1.

During the online phase, the spatial filters are not modified and are used to process all the EEG test windows. As shown in Figure 1, the incoming test signal $\mathbf{X}$ is routed through the top pathway and projected through the pre-optimized spatial filters $\mathbf{a}_c^{(n)}$. Along the bottom pathway, the synthetic sine-cosine references take the place of the template matrix ($\mathbf{R} = \mathbf{Y}_c$) and are projected through their respective harmonic weights $\mathbf{b}_c^{(n)}$. These precomputed filters extract the canonical variables for each sub-band n prior to the vectorization (vec) and Pearson correlation stages, providing a subspace that isolates the target SSVEP responses.

B. Task-Related Component Analysis (TRCA)

TRCA is a spatial filtering method designed to extract robust signal components by maximizing reproducibility across multiple trials of the same task [5]. Rather than looking for correlations with synthetic templates, TRCA focuses entirely on the underlying biological consistency of the brain responses [5]. The algorithm seeks an optimal spatial filter vector $\mathbf{w}$ that maximizes the similarity of the filtered EEG signals across trials of a given stimulus class.

This optimization objective is mathematically formulated as the maximization of the ratio between the inter-trial covariance and the total intra-trial variance. The optimal weight vector $\mathbf{w}$ is computed through the following formulation

$$\mathbf{w} = \arg\max_{\mathbf{u}} \frac{\mathbf{u}^T \mathbf{S} \mathbf{u}}{\mathbf{u}^T \mathbf{Q} \mathbf{u}}, \tag{11}$$

where the matrix $\mathbf{S}$ isolates the task-related neural activity by accumulating the cross-covariance between all possible combinations of distinct calibration trials for a specific class. In contrast, the matrix $\mathbf{Q}$ captures the total intra-trial covariance across the concatenated EEG dataset, serving as a spatial regularizer that accounts for both the targeted evoked potentials and the background noise. This optimization problem is solved by transforming it into a generalized eigenvalue problem

$$\mathbf{S}\mathbf{u} = \lambda \mathbf{Q}\mathbf{u}, \tag{12}$$

where the optimal spatial filter corresponds to the eigenvector associated with the largest eigenvalue λ.

The optimization process in TRCA extracts a single spatial filter vector $\mathbf{w}$ for each class. This single filter is applied identically to both the online test signal and the reference template.

This symmetric deployment is visualized within the modular architecture of Figure 1. When evaluating a target class c at a specific sub-band n, the precomputed TRCA spatial filter is mapped directly onto both paths of the pipeline, meaning that $\mathbf{a}_c^{(n)} = \mathbf{b}_c^{(n)} = \mathbf{w}$. Along the top pathway, the incoming test signal $\mathbf{X}$ is projected through the spatial weights $\mathbf{a}_c^{(n)}$, and the template signal $\mathbf{R}$ takes the form of the mean of the training trials ($\bar{\mathbf{X}}_c$) and is projected through the identical filter weights $\mathbf{b}_c^{(n)}$ [5].

IV. Methodology

A. Databases

This study utilized two well-known public SSVEP-BCI datasets, which differ in their level of experimental control and the volume of data available for calibration. The first is the Benchmark Dataset [6], recorded in a controlled laboratory environment. The dataset contains EEG signals from 35 subjects operating a speller with 40 visual stimuli ($N_c = 40$) structured under the Joint Frequency-Phase Modulation (JFPM) technique. The stimulation frequencies ranged from 8.0 Hz to 15.8 Hz (increments of 0.2 Hz), with phases ranging from 0 to 1.5π radians (increments of 0.5π). The protocol consisted of 6 independent recording blocks, resulting in 240 trials per individual (6 trials for each stimulus), with each trial comprising 0.5 s of pre-stimulus and 5.0 s of active visual stimulation. The second is the BETA Database (BEnchmark database Towards BCI Application) [8]. Designed to simulate a practical real-world scenario, this dataset was recorded outside rigorously isolated environments, introducing higher levels of background noise. Although it shares the exact same 40-target JFPM visual configuration as the previous dataset, BETA expands the sample to 70 subjects and imposes a greater challenge on the classifiers, since the subjects performed only 4 independent recording blocks (totaling 160 trials). Furthermore, the visual stimulation time is reduced and variable, being 2.0 s for subjects S1 to S15, and 3.0 s for subjects S16 to S70 (both preceded by 0.5 s of pre-stimulus).

To evaluate the generalization performance of the algorithms, a leave-one-block-out cross-validation scheme was adopted across all subjects, ensuring that no participants were excluded from the analysis. For each trial designated as evaluation data, all spatial filter optimization stages, template generation, and parameter estimations were performed exclusively using the remaining independent calibration blocks. Under this separation, no test trial information was incorporated into the training phase.

B. Preprocessing

The analysis focused on 9 electrodes distributed over the parietal and occipital regions of the scalp (Pz, PO5, PO3, POz, PO4, PO6, O1, Oz, O2), as they capture the largest portion of the visual evoked response energy [5]. The EEG signal, originally sampled at 1000 Hz, was down-sampled to 250 Hz and subjected to digital band-pass filtering between 7 Hz and 90 Hz to attenuate baseline drift and high-frequency noise inherent to the acquisition [5]. Additionally, a 50 Hz notch filter was applied to eliminate power-line interference.

According to the literature [5], [6], we considered a physiological visual latency of approximately 140 ms. Therefore, by adding this biological delay to the 0.5 s static pre-stimulus period of each trial, the effective start of the feature extraction and analysis window for all algorithms was set at 0.64 s.

V. Experimental Evaluation

The experimental evaluation was conducted across windows size ranging from 0.1 s to 2.5 s, comprising two complementary experiments. The first experiment, the spatial equivalence analysis, assesses whether the spatial filters optimized by supervised CCA and TRCA achieve directional collinearity as the window size increases. The second experiment consists of a cross-substitution ablation study designed to decouple the influence of the spatial filter formulation from the complementary structural techniques (Filter Bank and Ensemble), comparing accuracy and Information Transfer Rate (ITR).

A. Spatial Equivalence Analysis

The first experiment evaluates the spatial equivalence between the filters across distinct stimulation window lengths. For each considered window size and stimulus class $c \in \{1, \ldots, 40\}$, the spatial filters $\mathbf{w}_{cca}$ (specifically the EEG spatial weight vector $\mathbf{w}_X$, obtained via supervised CCA (10)) and $\mathbf{w}_{trca}$ (obtained via TRCA (11), $\mathbf{w}$) were independently computed using calibration data. The implementation utilized a broadband configuration, omitting Filter Bank decomposition ($N = 1$ sub-band), with $N_h = 5$ harmonics for the CCA reference signals. To verify how the spatial filters generated by both methods compare as the window length increases, we employ the normalized inner product, or directional cosine, as a metric of collinearity. To ensure the measure reflects the directional alignment regardless of the filter's magnitude, both weight vectors $\mathbf{w}_{cca}$ and $\mathbf{w}_{trca}$ are normalized to unit norm within the calculation. For any given stimulus class c, the similarity metric S_c is defined as:

$$S_c = \frac{|\mathbf{w}_{cca}^T \mathbf{w}_{trca}|}{\|\mathbf{w}_{cca}\| \|\mathbf{w}_{trca}\|}. \tag{13}$$

The absolute value in this formulation is mathematically justified by a global polarity ambiguity inherent in the classification stage. As illustrated in the general processing flow of Figure 1, the target identification relies on the Pearson correlation coefficient. Because Pearson correlation is immune to sign inversions ($\mathbf{w} \to -\mathbf{w}$), two weight vectors pointing in diametrically opposed directions may produce functionally identical projections. This property holds as long as the consistency of the polarity is preserved between the test and template pathways of the pipeline. By omitting the algebraic sign, the index $S_c \in [0, 1]$ strictly measures the collinearity of the electrode weights across the spatial filter vectors, evaluating their directional alignment without being influenced by arbitrary signal inversions. Consequently, as S_c approaches 1, it serves as a direct indicator of equivalence between the two methodologies. The final spatial similarity metric, denoted as $\bar{S}_c$, was computed by averaging the individual coefficients across all N_c stimulus classes and M evaluated subjects:

$$\bar{S}_c = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{N_c} \sum_{i=1}^{N_c} S_{i,m} \right) \tag{14}$$

where $S_{i,m}$ represents the spatial similarity index obtained for the m-th participant and the i-th stimulus class.

B. Ablation Study

Supported by our unified classification architecture, we conducted an ablation study focusing on the cross-substitution of spatial filters. The objective was to investigate the classification performance when substituting the native spatial filter of one technique with the projection weights derived from another, thereby examining whether the spatial filters optimized by CCA and TRCA exhibit operational equivalence.

This systematic evaluation was first applied to the unsupervised CCA, supervised CCA, and TRCA algorithms, all implemented in a broadband configuration without Filter Bank decomposition or Ensemble architectures, and subsequently extended to incorporate these stages to isolate the individual contribution of each processing component.

The ablation study evaluates the following configurations, defined by their spatial projection matrices $\mathbf{W}_A$ and $\mathbf{W}_B$ and reference template $\mathbf{R}$ within the framework of Figure 1:

- *CCA (Unsupervised)* [4]: The filter bank and the ensemble mapping are bypassed, i.e., $N = 1$, $\mathbf{W}_A = \mathbf{a}_c^{(1)}$, $\mathbf{W}_B = \mathbf{b}_c^{(1)}$ and $\tilde{\rho}_c = \rho_c$. The system computes the spatial filter $\mathbf{a}_c^{(1)} = \mathbf{w}_X$ and the reference weight vector $\mathbf{b}_c^{(1)} = \mathbf{w}_Y$ dynamically for each EEG window solving (10), considering the template matrix $\mathbf{R} = \mathbf{Y}_c$.
- *Supervised CCA*: Uses the same configuration as the unsupervised version: $N = 1$, $\mathbf{W}_A = \mathbf{a}_c^{(1)} = \mathbf{w}_X$, $\mathbf{W}_B = \mathbf{b}_c^{(1)} = \mathbf{w}_Y$ and $\tilde{\rho}_c = \rho_c$. However, $\mathbf{a}_c^{(1)} = \mathbf{w}_X$ and $\mathbf{b}_c^{(1)} = \mathbf{w}_Y$ are computed using labeled calibration data and remain fixed during testing.
- *TRCA* [5]: The filter bank and the ensemble mapping are bypassed ($N = 1$, $\tilde{\rho}_c = \rho_c$). It uses the empirical biological mean through the template pathway ($\mathbf{R} = \bar{\mathbf{X}}_c$), which is computed by averaging the N_t available calibration trials for class c:

$$\bar{\mathbf{X}}_c = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{X}_{c,i} \tag{15}$$

  where $\mathbf{X}_{c,i}$ represents the i-th training trial. It applies symmetric spatial filtering, setting $\mathbf{W}_A = \mathbf{a}_c^{(1)} = \mathbf{w}$ and $\mathbf{W}_B = \mathbf{b}_c^{(1)} = \mathbf{w}$, where $\mathbf{w}$ is obtained solving (11).
- *TRCA with CCA Filter*: Uses the same configuration as standard TRCA ($N = 1$, $\tilde{\rho}_c = \rho_c$, template $\mathbf{R} = \bar{\mathbf{X}}_c$). However, it adopts the precomputed supervised CCA spatial filter, setting $\mathbf{W}_A = \mathbf{a}_c^{(1)} = \mathbf{w}_X$ and $\mathbf{W}_B = \mathbf{b}_c^{(1)} = \mathbf{w}_X$ instead of the native TRCA spatial filter.
- *Ensemble TRCA (eTRCA)* [5]: Operates in a broadband configuration ($N = 1$, $\tilde{\rho}_c = \rho_c$) with biological templates ($\mathbf{R} = \bar{\mathbf{X}}_c$). The ensemble mapping is activated across all N_c classes by stacking the optimized TRCA spatial filters to form projection matrices: $\mathbf{W}_A = \mathbf{W}_B = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{N_c}]$.
- *eTRCA with CCA Filter*: Uses the same configuration as eTRCA ($N = 1$, $\tilde{\rho}_c = \rho_c$, template $\mathbf{R} = \bar{\mathbf{X}}_c$). However, the projection matrices are constructed using the precomputed supervised CCA spatial filters: $\mathbf{W}_A = \mathbf{W}_B = [\mathbf{w}_X^{(1)}, \mathbf{w}_X^{(2)}, \ldots, \mathbf{w}_X^{(N_c)}]$.

- *eFB-TRCA* [5]: Enables filter-bank decomposition ($N = 5$) and the sub-band fusion block. For each sub-band $n \in \{1, \ldots, 5\}$, the test signal $\mathbf{X}^{(n)}$ and biological template $\mathbf{R}^{(n)} = \bar{\mathbf{X}}_c^{(n)}$ are projected through sub-band specific ensemble matrices $\mathbf{W}_A^{(n)} = \mathbf{W}_B^{(n)} = [\mathbf{w}_1^{(n)}, \ldots, \mathbf{w}_{N_c}^{(n)}]$. The final score $\tilde{\rho}_c$ aggregates the sub-band Pearson correlations $\rho_c^{(n)}$.
- *eFB-TRCA with CCA Filter*: Uses the same configuration as eFB-TRCA ($N = 5$, template $\mathbf{R}^{(n)} = \bar{\mathbf{X}}_c^{(n)}$). However, the sub-band projection matrices are constructed using supervised CCA spatial filters calibrated independently for each frequency band: $\mathbf{W}_A^{(n)} = \mathbf{W}_B^{(n)} = [\mathbf{w}_X^{(1,n)}, \ldots, \mathbf{w}_X^{(N_c,n)}]$.

With the exception of the unsupervised CCA, to evaluate these configurations, we adopted a Leave-One-Block-Out cross-validation strategy. For each subject, one block of trials was held out as the test dataset, while the remaining blocks served as the training data. During the training phase, the offline pipeline is activated to calibrate the spatial filter vectors ($\mathbf{a}_c^{(n)}$ and $\mathbf{b}_c^{(n)}$) and to calculate the biological template matrices ($\bar{\mathbf{X}}_c$). During the online testing phase, the system classifies the held-out trials using the pre-optimized weights, and the average decoding accuracy and Information Transfer Rate across all subjects is reported as the final performance metric.

VI. Results and Discussion

A. Similarity of the Spatial Filters

As illustrated in Figure 2 for the Benchmark dataset, the evolution of the similarity ($\bar{S}_c$) between the spatial filters optimized by TRCA and supervised CCA revealed a convergent behavior.

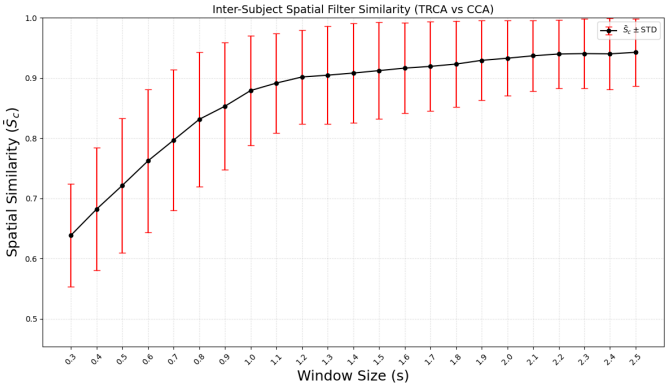


Fig. 2. Effect of window size on the spatial similarity ($\bar{S}_c$) between the filters generated by TRCA and supervised CCA.

For short stimulation windows, a low spatial correlation was observed between the weight vectors. A possible hypothesis for this divergence lies in the dynamics of the visual onset response. In shorter windows, the initial transient accounts for a significant proportion of the visual evoked potential. Because TRCA relies on empirical biological templates ($\bar{\mathbf{X}}_c$), its optimization naturally captures the dynamic of this early transient. In contrast, supervised CCA projects the data against synthetic sinusoids with constant amplitude ($\mathbf{Y}_c$), which fail to model the transient response, forcing the filter to align with the sustained steady-state oscillation. Consequently, for short window size, the algorithms obtain filters with low spatial collinearity.

However, as the window size increases, the spatial similarity increases monotonically with the S_c index stabilizing at high concordance values for a window size exceeding 0.7 s. Aligning with our dynamics hypothesis, a plausible explanation is that the sustained steady-state oscillation progressively accounts for a larger fraction of the total signal energy over a longer period of time. Once the initial transient settles, the empirical covariance captured by TRCA becomes governed by the stationary periodic rhythm. Under this interpretation, because such rhythmic activity aligns with the synthetic sinusoids of supervised CCA, both methods appear to converge toward extracting functionally equivalent spatial filters of the underlying neural response.

B. Classification Performance and Information Transfer Rate

To evaluate how the observed collinearity of the weight vectors translates into classification performance, we now analyze the classification accuracy and ITR provided by the ablation configurations.

The classification outcomes of these cross-substitution pipelines are detailed in Figures 3 and 4, which plot the classification accuracy curves for both the Benchmark and BETA datasets.

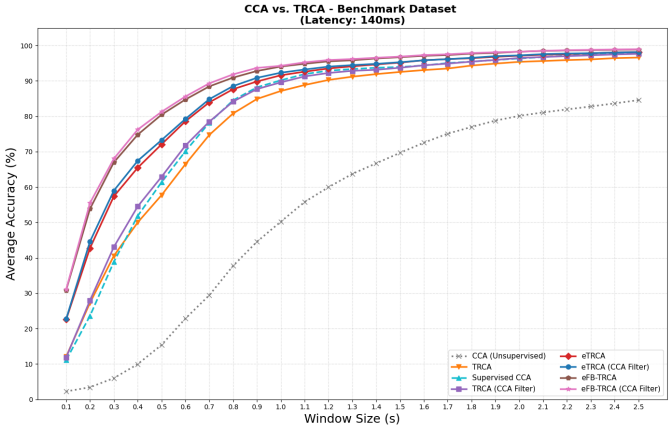


Fig. 3. Classification accuracy for the Benchmark dataset.

In the Benchmark dataset, characterized by a high SNR and the availability of 5 blocks for calibration (one block is used for test), the overall classification accuracy is high across all window lengths compared to the BETA dataset. For configurations without Filter Bank and Ensemble, *Supervised CCA* outperformed *TRCA*, indicating that synthetic reference sinusoids yielded higher decoding accuracy than the empirical mean, while *Unsupervised CCA* resulted in the lowest accuracy. Overall, configurations without Filter Bank and Ensemble presented lower decoding accuracy across

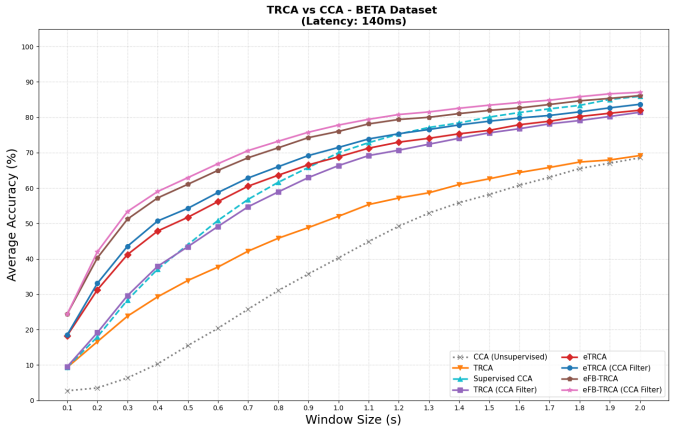


Fig. 4. Classification accuracy for the BETA dataset.

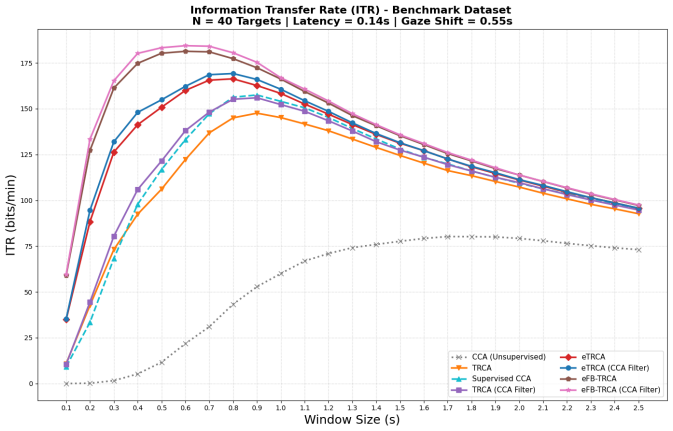


Fig. 5. Information Transfer Rate for the Benchmark dataset.

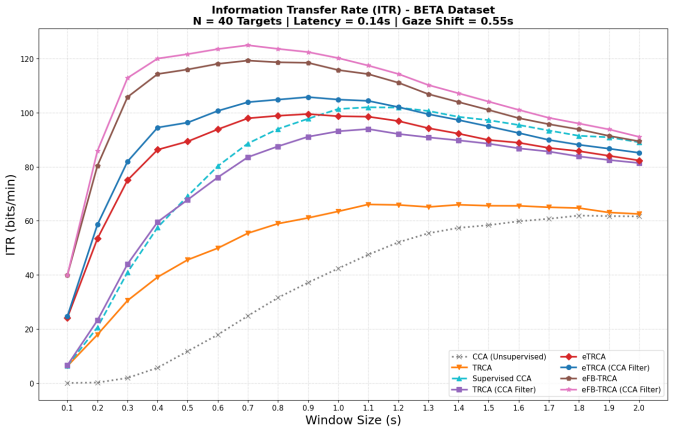


Fig. 6. Information Transfer Rate for the BETA dataset.

all window sizes than those incorporating these modules. Regarding the cross-substitution analysis, the native methods and their counterparts operating with the CCA filter exhibit a consistent behavior. The performance curves of the TRCA variations and those adapted with the CCA filter are almost superimposed. This behavior extends to the Filter Bank and Ensemble architectures, where eFB-TRCA and the eFB-TRCA with the substituted filter exhibit nearly identical performance, supporting the spatial convergence hypothesis ($S_c \rightarrow 1$).

In the BETA dataset, characterized by a less controlled environment with lower SNR and the availability of 3 blocks for calibration, the overall classification accuracy is reduced across all window lengths compared to the Benchmark dataset. For configurations without Filter Bank and Ensemble, *Supervised CCA* outperformed *TRCA*, while *Unsupervised CCA* resulted in the lowest accuracy. Overall, configurations without Filter Bank and Ensemble presented lower decoding accuracy across all window lengths than those incorporating these modules, with this performance gap widening under low-SNR conditions. Regarding the cross-substitution analysis, the native methods and their counterparts operating with the CCA filter exhibit a consistent behavior. The performance curves of the TRCA variations and those adapted with the CCA filter remain closely aligned along. This parity extends to the Filter Bank and Ensemble architectures, where eFB-TRCA and the eFB-TRCA with the substituted filter exhibit nearly identical performance.

To further evaluate the practical communication efficiency of the BCI system, the ITR was calculated following the standard evaluation protocol described in [6], assuming a gaze shifting time of 0.5 s. Figure 5 and Figure 6 illustrates the ITR curves for both datasets.

As shown in Figures 5 and 6, the maximum ITR is achieved at intermediate window lengths, typically before the accuracy curve reaches its saturation point. Consistent with the accuracy results, the Filter Bank and Ensemble architectures, particularly eFB-TRCA, achieve higher peak ITRs compared to the baseline methods in both controlled and noisy environments. Moreover, the substitution of the native TRCA spatial filter with the supervised CCA filters maintains the communication rates. The similar ITR profiles between the native and cross-substituted methods reinforce the spatial convergence and seems to not compromise the overall throughput of the system.

VII. Conclusion

The spatial convergence analysis showed collinearity ($S_c \rightarrow 1$) between supervised CCA and TRCA weight vectors for windows exceeding 0.7 s. However, the ablation study revealed equivalent decoding performance across all window lengths, suggesting the hypothesis that filter bank decomposition, ensemble, and template matching compensate for the spatial weight divergence observed at shorter windows. Further investigation is required to fully explain this mechanism.

Consequently, these findings suggest that the performance gap reported in literature between CCA and TRCA arises primarily from the classification structure rather than the criteria used to design the spatial filters. When deployed within identical frameworks with equivalent reference signals, TRCA and CCA appear to yield comparable decoding outcomes.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] Y. Wang, X. Gao, B. Hong, C. Jia, and S. Gao, "Brain-computer interfaces based on visual evoked potentials," *IEEE Engineering in Medicine and Biology Magazine*, vol. 27, no. 5, pp. 64–71, 2008.

[2] F. B. Vialatte, M. Maurice, J. Dauwels, and A. Cichocki, "Steady-state visually evoked potentials: focus on essential paradigms and future perspectives," *Progress in Neurobiology*, vol. 90, no. 4, pp. 418–438, 2010.

[3] M. Nakanishi, Y. Wang, Y.-T. Wang, and T.-P. Jung, "A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials," *PLoS One*, vol. 10, no. 10, p. e0140703, 2015.

[4] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2610–2614, 2006.

[5] M. Nakanishi, Y. Wang, X. Chen, Y.-T. Wang, X. Gao, and T.-P. Jung, "Enhancing detection of SSVEPs for a high-speed brain-computer interface based on task-related component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 104–112, 2018.

[6] Y. Wang, X. Chen, X. Gao, and S. Gao, "A benchmark dataset for SSVEP-based brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1746–1752, 2016.

[7] X. Chen, Y. Wang, S. Gao, T.-P. Jung, and X. Gao, "Filter bank canonical correlation analysis for steady-state visual evoked potential-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046008, 2015.

[8] B. Liu, X. Huang, Y. Wang, X. Chen, and X. Gao, "BETA: A large benchmark database toward SSVEP-BCI application," *Frontiers in Neuroscience*, vol. 14, p. 627, 2020.

Triagem Multitarefa Explicável de Retinopatia Diabética e Glaucoma em Fundo de Olho sob Incerteza e Deslocamento de Domínio

1st Carlos Alberto da Silva Brasil
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
carlos.brasil@alu.ufc.br

2nd Paulo Cesar Cortez
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
cortez@lesc.ufc.br

3rd Victor Hugo Costa de Albuquerque
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
victor.albuquerque@ieee.org

***Resumo*—Este estudo avaliou um pipeline multitarefa de triagem conjunta para retinopatia diabética referenciável e glaucoma em imagens de fundo de olho que combina EfficientNet-B0 com análises de discriminação, calibração, incerteza preditiva, limiares operacionais e explicabilidade post-hoc. O objetivo foi caracterizar, em validação interna e avaliação externa, o comportamento dessas propriedades preditivas sob deslocamento de domínio, delimitando o sistema como uma estrutura experimental de avaliação, e não como ferramenta clinicamente pronta para uso. As principais constatações foram: desempenho discriminatório elevado para retinopatia diabética referenciável na validação interna histórica do APTOS, com AUC de 0,9791 (IC95%: 0,9704–0,9865), cuja interpretação requer cautela devido à identificação retrospectiva de duplicatas exatas entre treinamento e validação, e redução no EyePACS, com AUC de 0,7181 (IC95%: 0,7051–0,7310); maior instabilidade para glaucoma, com AUC de 0,8060 (IC95%: 0,7237–0,8801) no REFUGE-CLAHE, 0,5525 (IC95%: 0,4382–0,6625) no ACRIMA-CLAHE e 0,5949 (IC95%: 0,5834–0,6058) no EyePACS-AIROGS-light-V2-CLAHE; efeitos do CLAHE dependentes do domínio, sem eliminação da baixa especificidade observada na avaliação externa; comportamento tarefa-dependente de MC Dropout e Temperature Scaling; e explicações Grad-CAM/SLIC mais concentradas em predições corretas e mais difusas em erros ou casos incertos.**

***Palavras-chave*—retinopatia diabética, glaucoma, fundo de olho, aprendizado multitarefa, validação externa, calibração, incerteza, explicabilidade.**

I. Introdução

Retinopatia diabética e glaucoma estão entre as causas relevantes de perda visual evitável, e ambas podem ser rastreadas por fotografias de fundo de olho [7], [9], [13]. Em oftalmologia, porém, modelos de deep learning frequentemente perdem acurácia quando avaliados em bases externas [14], [26], devido a variações de aquisição, população, prevalência, resolução e critérios de rotulagem [22]. Esse cenário justifica investigar abordagens multitarefa, pois a triagem clínica pode exigir a avaliação de mais de uma patologia na mesma imagem [4].

Este manuscrito avalia um pipeline multitarefa baseado em EfficientNet-B0 [12], com cabeças separadas para retinopatia diabética e glaucoma acopladas ao mesmo extrator de características [11]. Foram analisados MC Dropout [20], Temperature Scaling [8], CLAHE apenas para glaucoma [17], limiares operacionais empíricos e interpretabilidade post-hoc com Grad-CAM [23] e superpixels SLIC [24], em validação interna e avaliação externa, sem assumir prontidão clínica.

As contribuições são: i) avaliar um pipeline multitarefa auditável; ii) comparar o desempenho em validação interna e avaliação externa em bases públicas; iii) examinar calibração, incerteza, limiares e deslocamento de domínio; e iv) analisar erros com Grad-CAM e SLIC. O foco é caracterizar a confiabilidade além da AUC [15].

II. Materiais e Métodos

A. Delineamento do Estudo

Este estudo retrospectivo avaliou um sistema multitarefa aplicado a imagens coloridas de fundo de olho. O sistema gera duas saídas: glaucoma, tratado como classificação binária, e retinopatia diabética, inicialmente mapeada em cinco classes e posteriormente binarizada em referível ou não referível.

$$\left(p_i^{DR}, p_i^{G}\right) = h_\theta\left(x_i\right) = \left(h_{DR}\left(f_\theta\left(x_i\right)\right), h_G\left(f_\theta\left(x_i\right)\right)\right). \quad (1)$$

As funções h_{DR} e h_G correspondem, respectivamente, às cabeças de classificação para retinopatia diabética e glaucoma. Os métodos foram avaliados em conjuntos internos, para experimentos controlados, e em conjuntos externos, para examinar generalização de domínio. Para auditoria e reprodutibilidade, registraram-se as predições do modelo, os limiares adotados, as medidas de calibração, as estimativas de incerteza e as métricas de cada configuração experimental.

B. Modelo Multitarefa

A configuração experimental multitask_real_v2 foi adotada como base. A EfficientNet-B0 foi escolhida por sua reprodutibilidade e baixo custo computacional [12]. O classificador original foi substituído por duas cabeças: uma de cinco classes para retinopatia diabética e outra binária para glaucoma. Uma camada Dropout com probabilidade de 0,3 foi posicionada antes dessas cabeças, permitindo regularização e posterior inferência com MC Dropout.

Para retinopatia diabética, a cabeça de classificação em cinco classes produziu logits $z_{i,k}^{DR}$, com $k \in \{0, 1, 2, 3, 4\}$. Antes da etapa de calibração, as probabilidades de classe correspondentes são dadas pela função softmax:

$$p_{i,k}^{DR} = \frac{\exp\left(z_{i,k}^{DR}\right)}{\sum_{c=0}^{4} \exp\left(z_{i,c}^{DR}\right)},\ k \in \{0,1,2,3,4\}. \quad (2)$$

Para a tarefa de glaucoma, a cabeça binária do modelo produziu um logit z_i^G:

$$p_i^G = \sigma\left(z_i^G\right) = \frac{1}{1+\exp\left(-z_i^G\right)}. \quad (3)$$

A probabilidade de retinopatia diabética referenciável foi obtida pela soma das probabilidades associadas aos graus 2 a 4. No pipeline final, tais probabilidades correspondem às saídas resultantes do escalonamento de temperatura (Temperature Scaling), descrito na sequência:

$$p_i^{RDR}(T_{DR}) = \sum_{k=2}^{4} p_{i,k}^{DR}(T_{DR}). \quad (4)$$

As decisões binárias de rastreamento foram obtidas mediante a aplicação de limiares específicos para cada tarefa, definidos como:

$$\widehat{y}_i^{RDR} = I\left(p_i^{RDR}(T_{DR}) \geq \tau_{DR}\right),\ \widehat{y}_i^{G} = I\left(p_i^{G} \geq \tau_{G}\right), \quad (5)$$

em que $I(\cdot)$ denota a função indicadora, τ_{DR} corresponde ao limiar adotado para a retinopatia diabética referenciável e τ_G, ao limiar adotado para o glaucoma. A decisão conjunta de encaminhamento foi definida pela união lógica das saídas específicas de cada tarefa, expressa por:

$$\widehat{y}_i^{Joint} = \widehat{y}_i^{RDR} \vee \widehat{y}_i^{G}. \quad (6)$$

Assim, a imagem foi encaminhada quando pelo menos uma das saídas específicas foi positiva.

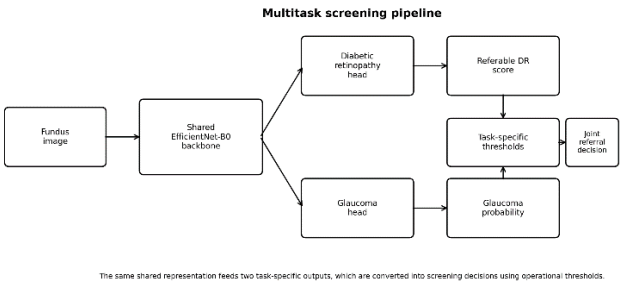


Fig. 1. Uma visão geral da arquitetura do modelo multitarefa proposto. Uma rede de backbone EfficientNet-B0 é empregada e seus mapas de características finais são enviados para duas cabeças específicas para cada tarefa (retinopatia diabética e glaucoma), seguidas pelos limiares operacionais e pela decisão conjunta de encaminhamento.

A proveniência documental da configuração `multitask_real_v2` foi reconstruída a partir dos artefatos experimentais preservados. Ela manteve as 3.662 imagens do APTOS de `multitask_real_v1`, redistribuídas por particionamento estratificado 80/20 (semente 42) em 2.929 imagens de treinamento e 733 de validação interna. Como os registros não documentam uma busca formal de hiperparâmetros, `multitask_real_v2` é tratada como configuração experimental consolidada, e não como modelo vencedor. As análises estatísticas utilizaram predições individuais congeladas preservadas, que reproduzem as estimativas pontuais originalmente reportadas. Não foi possível verificar documentalmente uma correspondência unívoca entre essas predições e um arquivo de checkpoint atualmente preservado; assim, a reprodutibilidade dos resultados aqui reportada refere-se à recomputação das métricas a partir das predições congeladas, e não à regeneração dessas predições a partir dos pesos do modelo.

C. Conjuntos de Dados e Coortes de Avaliação

Este estudo utilizou bases de dados públicas tanto para a validação interna, compreendendo APTOS [1] e REFUGE [10], quanto para a avaliação externa, compreendendo EyePACS [5], ACRIMA [3] e EyePACS-AIROGS-light-V2 [2]. Para a retinopatia diabética, os graus 0–1 foram agrupados na categoria não referível e os graus 2–4 na categoria referível, ao passo que o glaucoma foi tratado como tarefa de classificação binária. A Tabela I resume a composição das coortes resultantes.

Tabela I
COMPOSIÇÃO DOS CONJUNTOS DE DADOS UTILIZADOS NA AVALIAÇÃO CONSOLIDADA

Tarefa	Dataset	Função	Formato do rótulo	n
Glaucoma	REFUGE	Treinamento	Binário	400
	REFUGE	Validação interna/desenvolvimento	Binário	400
	ACRIMA	Teste externo	Binário	106
	EyePACS-AIROGS-light-V2	Teste externo	Binário	9,540
Retinopatia diabética	APTOS	Treinamento	Grau de RD em cinco classes	2,929
	APTOS	Validação interna/desenvolvimento	Grau de RD em cinco classes	733
	EyePACS	Teste externo	Grau de RD em cinco classes	10,906
Total				25,014

EyePACS, ACRIMA e EyePACS-AIROGS-light-V2 foram tratados como conjuntos de teste externo para a avaliação de generalização sob deslocamento de domínio. Os limiares operacionais aplicados nessas coortes foram definidos a partir dos dados de validação interna e posteriormente mantidos fixos na avaliação externa. Dessa forma, os resultados externos são apresentados como estimativas de transferência do pipeline para domínios distintos, e não como desempenho ajustado especificamente a essas coortes.

Uma auditoria retrospectiva das partições identificou comportamentos distintos entre as tarefas. Para glaucoma, não foram observadas duplicatas exatas entre as partições internas REFUGE nem entre REFUGE e as coortes externas avaliadas. Para retinopatia diabética, APTOS e EyePACS permaneceram separados no nível de duplicatas exatas; entretanto, foram identificados 41 hashes SHA-256 compartilhados entre as partições de treinamento e validação interna do APTOS, envolvendo 83 registros. Assim, não se pode assumir independência completa em nível de imagem entre essas duas partições, e as métricas internas de desenvolvimento para retinopatia diabética devem ser interpretadas com essa limitação. A independência em nível de paciente não pôde ser verificada uniformemente, devido à ausência de identificadores de paciente nas bases utilizadas.

Não houve conjunto interno de teste independente utilizado para a avaliação final de nenhuma das duas tarefas. Para a retinopatia diabética, o conjunto APTOS foi particionado em 2.929 imagens destinadas ao treinamento e 733 imagens destinadas à validação interna/desenvolvimento, sem a

definição de uma partição adicional de teste interno. Para glaucoma, as mesmas 400 imagens do REFUGE-CLAHE classificadas como validação interna/desenvolvimento foram utilizadas tanto para examinar os pontos de operação de τ_G quanto para estimar o desempenho interno dependente de limiar apresentado neste manuscrito. Consequentemente, essas métricas devem ser interpretadas como resultados de desenvolvimento, e não como estimativas obtidas em um teste interno independente, podendo refletir otimismo associado ao uso do mesmo conjunto na definição do ponto de operação. A avaliação das coortes externas mantidas separadas das decisões de desenvolvimento constitui a principal evidência de generalização do pipeline.

D. Pré-processamento

Todas as imagens de fundo de olho foram redimensionadas para 224×224 pixels, convertidas em tensores e normalizadas segundo os valores de média e desvio-padrão do ImageNet, em conformidade com a distribuição de entrada requerida pelo EfficientNet-B0 [12], sendo cada canal normalizado, após o redimensionamento, conforme a expressão a seguir:

$$\tilde{x}_{i,c} = \frac{x_{i,c} - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\}, \tag{7}$$

em que $x_{i,c}$ denota a intensidade da imagem no canal c, μ_c corresponde à média do ImageNet e σ_c, ao desvio-padrão correspondente.

Para o ramo de glaucoma, o CLAHE [17] foi avaliado apenas em inferência, sem retreinamento ou alteração dos parâmetros do modelo. As cópias processadas por CLAHE foram geradas no canal de luminância do espaço LAB, com limite de corte de 2.0 e grade de blocos 8×8:

$$\begin{aligned} L_i^{'} &= \text{CLAHE}\left(L_i; \alpha = 2.0, g = 8 \times 8\right), \\ I_i^{'} &= \text{RGB}\left(L_i^{'}, a_i, b_i\right). \end{aligned} \tag{8}$$

Aqui, L_i, a_i e b_i denotam os canais LAB da imagem i; α é o limite de corte; g define a grade de blocos; e I_i' é a imagem RGB resultante. O CLAHE foi usado apenas na avaliação de glaucoma. Para retinopatia diabética, manteve-se o fluxo padrão de pré-processamento.

E. Avaliação de Duplicatas e Vazamento de Dados

Foi realizada uma auditoria retrospectiva de possível vazamento de dados com base em identificadores de amostra e hashes SHA-256 dos arquivos [16]. No APTOS, essa auditoria identificou 41 hashes compartilhados entre treinamento e validação interna, envolvendo 83 registros, enquanto não foram observadas colisões SHA-256 entre APTOS e EyePACS. Para glaucoma, não foram observadas duplicatas exatas entre as partições internas do REFUGE nem entre REFUGE e as coortes externas avaliadas. Essa verificação detecta identidade binária dos arquivos, mas não exclui versões visualmente equivalentes obtidas por recorte, redimensionamento ou recompressão; adicionalmente, a independência em nível de paciente não pôde ser verificada de forma uniforme por ausência de identificadores apropriados.

F. Métricas de Avaliação

A capacidade discriminatória foi avaliada pela AUC [6], e a calibração pelo Brier score e pelo Expected Calibration Error (ECE) [8], [18], este último calculado com $M = 10$ intervalos igualmente espaçados. O desempenho nos pontos operacionais foi caracterizado por sensibilidade, especificidade, acurácia, PPV, NPV, F1-score, acurácia balanceada, falsos positivos e falsos negativos, sendo a acurácia balanceada definida como a média entre sensibilidade e especificidade. Para retinopatia diabética, os graus 0–1 foram considerados não referíveis e os graus 2–4 referíveis; o glaucoma foi tratado como classificação binária.

G. Calibração com Temperature Scaling

O Temperature Scaling [8] foi aplicado como pós-processamento específico por tarefa, sem retreinamento dos pesos do modelo. Para retinopatia diabética, a temperatura escalar previamente determinada na validação interna do APTOS ($n = 733$), $T_{DR} = 0.9358$, foi aplicada aos logits da cabeça multiclasse antes da função softmax:

$$p_{i,k}^{DR}(T_{DR}) = \frac{\exp\left(z_{i,k}^{DR}/T_{DR}\right)}{\sum_{j=0}^{4}\exp\left(z_{i,j}^{DR}/T_{DR}\right)}. \tag{9}$$

A temperatura foi mantida no pipeline final de retinopatia diabética por reduzir o ECE na validação interna e foi aplicada sem novo ajuste à coorte externa EyePACS. O escore de retinopatia diabética referível utilizado nas análises binárias foi calculado a partir dessas probabilidades calibradas como $p_i^{RDR}(T_{DR}) = p_{i,2}^{DR}(T_{DR}) + p_{i,3}^{DR}(T_{DR}) + p_{i,4}^{DR}(T_{DR})$.

Para glaucoma, o Temperature Scaling foi examinado separadamente em uma análise auxiliar de calibração utilizando o REFUGE sem CLAHE (raw). Como Brier score e ECE pioraram após a transformação, nenhuma temperatura de glaucoma foi incorporada ao pipeline final. Os registros preservados documentam os valores de temperatura e as comparações antes/depois, mas não preservam informação suficiente para reconstruir o procedimento numérico utilizado originalmente para determinar as temperaturas; por isso, nenhum algoritmo específico de otimização é atribuído neste manuscrito.

H. Inferência Estatística

Para as principais métricas de discriminação, calibração e desempenho operacional, foram calculados intervalos de confiança de 95% por bootstrap não paramétrico em nível de imagem, com 10.000 reamostragens e semente fixa 42. Em cada combinação de tarefa e conjunto de dados, as imagens foram reamostradas com reposição, e AUC, sensibilidade, especificidade, F1-score, Brier score e ECE foram recomputados. O ECE foi recalculado em cada réplica mantendo $M = 10$ intervalos igualmente espaçados. Réplicas contendo apenas uma classe foram excluídas do cálculo da AUC. PPV e NPV foram mantidos como estimativas pontuais. Cada réplica foi obtida recalculando a métrica θ:

$$\widehat{\theta}^{*(b)} = T\left(D^{*(b)}\right), \; b = 1, \ldots, B, \tag{10}$$

em que $D^{*(b)}$ representa o b-ésimo conjunto de dados bootstrap, $T(\cdot)$ corresponde à função da métrica considerada

e $B = 10\,000$. O intervalo de confiança bootstrap percentílico foi definido, por sua vez, como:

$$\mathrm{CI}_{95\%}(\theta) = \left[Q_{0.025}\left(\widehat{\theta}^{*}\right), Q_{0.975}\left(\widehat{\theta}^{*}\right)\right], \qquad (11)$$

em que $Q_q(\cdot)$ denota o q-quantil empírico da distribuição bootstrap [21]. O teste de McNemar foi mantido exclusivamente para comparações pareadas auxiliares selecionadas, ao passo que as análises de DeLong e de Holm-Bonferroni não foram preservadas como resultados primários.

I. Análise de Limiar

O limiar $\tau_G = 0.10$ foi preservado como ponto operacional congelado da avaliação principal de glaucoma, tendo sido conduzida, adicionalmente, uma análise exploratória de sensibilidade ao ponto de operação no conjunto REFUGE-CLAHE [10]. Dentre os pontos examinados nessa análise, $\tau_G = 0.249$ foi selecionado internamente como ponto balanceado, por apresentar a maior acurácia balanceada no REFUGE-CLAHE, enquanto $\tau_G = 0.300$ foi mantido como alternativa conservadora destinada a ilustrar a redução de falsos positivos. O ponto balanceado exploratório foi definido, assim, como:

$$\tau_G^{*} = \arg\max_{\tau \in \mathcal{T}} \frac{1}{2}\left[\text{Sensitivity}(\tau) + \text{Specificity}(\tau)\right], \qquad (12)$$

em que $\mathcal{T}$ denota os pontos de operação examinados no REFUGE-CLAHE. Para cada ponto, foram calculados sensibilidade, especificidade, acurácia balanceada, falsos positivos e falsos negativos. Os IC95% de sensibilidade, especificidade e acurácia balanceada na avaliação externa foram estimados pelo mesmo bootstrap não paramétrico em nível de imagem, com $B = 10\,000$ e semente 42. ACRIMA-CLAHE e EyePACS-AIROGS-light-V2-CLAHE foram usados apenas para avaliação externa, sem participação na escolha dos pontos exploratórios. Em particular, os resultados externos não foram utilizados para redefinir retrospectivamente nem o limiar principal de 0.10 nem o ponto balanceado selecionado internamente.

J. Experimentos de Glaucoma em Tarefa Única

Também foram avaliados modelos de tarefa única para glaucoma, a fim de verificar se as limitações observadas decorriam do desenho multitarefa [3], [9]. Cinco configurações foram testadas: CLAHE com entropia cruzada binária (224 × 224), CLAHE com focal loss (224 × 224), CLAHE com entropia cruzada binária (384 × 384), CLAHE com corte central de 80% e entropia cruzada binária (384 × 384), e CLAHE com corte central de 80% e entropia cruzada binária (512 × 512). Esses testes avaliaram o efeito da normalização de contraste, da focal loss, do aumento de resolução e da centralização do disco óptico na classificação de glaucoma. A entropia cruzada binária foi definida por:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(p_i\right) + \left(1 - y_i\right)\log\left(1 - p_i\right)\right]. \qquad (13)$$

O modelo de focal loss foi escrito como [25]:

$$L_{\text{focal}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_t\left(1 - p_{t,i}\right)^{\gamma}\log\left(p_{t,i}\right), \qquad (14)$$

onde $p_{t,i} = p_i$ quando $y_i = 1$ e $p_{t,i} = 1 - p_i$ quando $y_i = 0$. Os parâmetros α_t e γ controlam, respectivamente, o balanceamento entre classes e a redução de peso de exemplos já bem classificados. Devido ao tamanho limitado e ao desequilíbrio do conjunto de dados, os modelos de tarefa única foram tratados como configurações comparativas, não como classificadores definitivos.

K. Análise de Erros

Falsos positivos e falsos negativos foram avaliados quanto a contraste, iluminação global, foco, grau de zoom, obstruções visuais, artefatos e visibilidade da anatomia relevante. Como informações clínicas adicionais não estavam disponíveis, os erros foram descritos como visuais ou relacionados à qualidade da imagem, e não como categorias clínicas de erro.

L. Estimativa de Incerteza com MC Dropout

A incerteza preditiva foi estimada por MC Dropout com $T = 10$ passagens estocásticas durante a inferência, sem retreinamento [20]. Foram calculados média de predição, desvio-padrão preditivo e entropia preditiva para predições corretas, incorretas, falsos positivos e falsos negativos. Para cada imagem, a média e o desvio-padrão preditivos foram definidos como:

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p_i^{(t)}, \qquad (15)$$

$$s_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(p_i^{(t)} - \bar{p}_i\right)^2}, \qquad (16)$$

em que $p_i^{(t)}$ é a probabilidade prevista na t-ésima passagem estocástica. Para a predição binária de glaucoma, a entropia preditiva foi calculada como:

$$H_i^{G} = -\bar{p}_i^{G}\log\left(\bar{p}_i^{G}\right) - \left(1 - \bar{p}_i^{G}\right)\log\left(1 - \bar{p}_i^{G}\right). \qquad (17)$$

Para a retinopatia diabética, a entropia preditiva multiclasse foi calculada a partir das probabilidades médias de classe obtidas em múltiplas passagens estocásticas:

$$H_i^{DR} = -\sum_{k=0}^{4}\bar{p}_{i,k}^{DR}\log\left(\bar{p}_{i,k}^{DR}\right). \qquad (18)$$

A incerteza foi caracterizada, para o glaucoma, pela probabilidade média de glaucoma, pelo desvio-padrão entre as passagens estocásticas e pela entropia preditiva binária associada a essa predição; de modo análogo, para a retinopatia diabética, foi avaliada pelo escore médio de retinopatia diabética referenciável, pelo desvio-padrão entre as passagens estocásticas e pela entropia preditiva multiclasse. Tais medidas foram derivadas diretamente das predições estocásticas, sem qualquer alteração dos parâmetros ou dos limiares empregados na predição.

M. Avaliação de Eficiência Computacional

A eficiência computacional foi avaliada em duas frentes: a sobrecarga associada às 10 passagens estocásticas do MC Dropout e a comparação entre o backbone multitarefa compartilhado e dois modelos EfficientNet-B0 independentes. As medições foram conduzidas com 400 imagens do conjunto REFUGE-CLAHE, batch size de 16, entradas de 224 × 224 pixels e uma GPU NVIDIA A100 80GB PCIe. Esse benchmark teve finalidade exclusivamente computacional, não tendo sido utilizado para recalcular os resultados clínicos ou probabilísticos.

III. Resultados

A. Desempenho Geral do Sistema Multitarefa Final

O sistema multitarefa apresentou comportamento dependente da tarefa. Nas predições preservadas, a retinopatia diabética apresentou AUC de 0,9791 (IC95%: 0,9704–0,9865) na validação interna do APTOS, estimativa que deve ser interpretada à luz da sobreposição por duplicatas exatas descrita nos Métodos, e AUC de 0,7181 (IC95%: 0,7051–0,7310) no EyePACS externo; o glaucoma foi mais sensível ao deslocamento de domínio. No REFUGE-CLAHE, a AUC do glaucoma foi 0,8060 (IC95%: 0,7237–0,8801), enquanto no ACRIMA-CLAHE foi 0,5525 (IC95%: 0,4382–0,6625) e no EyePACS-AIROGS-light-V2-CLAHE, 0,5949 (IC95%: 0,5834–0,6058). O intervalo observado no ACRIMA-CLAHE evidencia maior incerteza nessa coorte externa de menor tamanho amostral e inclui 0,50, sendo portanto compatível com discriminação ao nível do acaso nessa coorte. Nos limiares originais, o glaucoma no EyePACS-AIROGS-light-V2-CLAHE apresentou alta sensibilidade e baixa especificidade, enquanto a retinopatia diabética apresentou alto desempenho no APTOS e, no teste externo, maior especificidade que sensibilidade.

Tabela II
Desempenho probabilístico do sistema multitarefa final nos conjuntos de avaliação internos e externos. Os valores entre parênteses representam IC95% bootstrap percentílicos.

Conjunto de dados	n	AUC (IC95%)	Brier (IC95%)	ECE (IC95%)
Glaucoma – REFUGE-CLAHE	400	0.8060 (0.7237–0.8801)	0.0841 (0.0680–0.1012)	0.0824 (0.0647–0.1127)
Glaucoma – ACRIMA-CLAHE	106	0.5525 (0.4382–0.6625)	0.4389 (0.3580–0.5198)	0.4410 (0.3564–0.5423)
Glaucoma – EyePACS-AIROGS-light-V2-CLAHE	9,540	0.5949 (0.5834–0.6058)	0.2868 (0.2819–0.2917)	0.1987 (0.1890–0.2088)
Diabetic ret. – APTOS	733	0.9791 (0.9704–0.9865)	0.0550 (0.0450–0.0657)	0.0254 (0.0223–0.0468)
Diabetic ret. – EyePACS	10,906	0.7181 (0.7051–0.7310)	0.1472 (0.1415–0.1529)	0.1048 (0.0984–0.1121)

Tabela III
Desempenho clínico-operacional do sistema multitarefa final nos limiares operacionais originais

Conjunto de dados	n	Sens. (IC95%)	Spec. (IC95%)	F1 (IC95%)
Glaucoma – REFUGE-CLAHE	400	0.8250 (0.6999–0.9388)	0.5722 (0.5205–0.6236)	0.2907 (0.2115–0.3684)
Glaucoma – ACRIMA-CLAHE	106	0.3833 (0.2586–0.5088)	0.6522 (0.5102–0.7857)	0.4646 (0.3333–0.5794)
Glaucoma – EyePACS-AIROGS-light-V2-CLAHE	9,540	0.9356 (0.9286–0.9424)	0.1050 (0.0964–0.1137)	0.6611 (0.6519–0.6702)
Diabetic ret. – APTOS	733	0.9597 (0.9355–0.9806)	0.8920 (0.8614–0.9195)	0.9065 (0.8813–0.9290)
Diabetic ret. – EyePACS	10,906	0.3439 (0.3234–0.3645)	0.9236 (0.9180–0.9292)	0.4112 (0.3903–0.4314)

Aplicado ao problema do glaucoma, o CLAHE trouxe resultados de discriminação e calibração que variaram conforme o domínio avaliado. No REFUGE, por exemplo, chegou-se a uma AUC de 0,8060 (IC95%: 0,7237–0,8801), com Brier score de 0,0841 e ECE de 0,0824, enquanto no ACRIMA a AUC caiu para 0,5525 (IC95%: 0,4382–0,6625), acompanhada de Brier de 0,4389 e ECE de 0,4410, e no EyePACS-AIROGS-light-V2 registrou-se AUC de 0,5949 (IC95%: 0,5834–0,6058), com Brier de 0,2868 e ECE de 0,1987; dessa forma, ficou claro que os efeitos do CLAHE dependeram do domínio em que o modelo foi testado, e que a perda de desempenho verificada na avaliação externa não foi resolvida por essa técnica.

B. Pontos Operacionais de Limiar para Glaucoma

Como análise exploratória da sensibilidade ao ponto de operação, este estudo avaliou retrospectivamente a aplicação de limiares alternativos ao EyePACS-AIROGS-light-V2-CLAHE sem reajuste a partir da coorte externa. O objetivo foi caracterizar o trade-off operacional sob $\tau_G = 0.249$, selecionado internamente no REFUGE-CLAHE como ponto balanceado, e $\tau_G = 0.300$, adotado como alternativa conservadora, mantendo-se $\tau_G = 0.10$ como limiar principal congelado. Como apresentado na Tabela IV, o aumento do limiar reduziu falsos positivos e elevou a especificidade, à custa de menor sensibilidade e maior número de falsos negativos. Esses achados evidenciam o trade-off operacional esperado em triagem, sem redefinição retrospectiva dos pontos de operação a partir da coorte externa.

Tabela IV
Pontos Operacionais para Glaucoma no EyePACS-AIROGS-light-V2 usando a configuração multitarefa com CLAHE. Os valores entre parênteses representam IC95% bootstrap percentílicos.

Ponto operacional	Sens. (IC95%)	Spec. (IC95%)	BAcc (IC95%)	FP	FN
$\tau = 0.100$ (principal)	0.9356 (0.9286–0.9424)	0.1050 (0.0964–0.1137)	0.5203 (0.5147–0.5258)	4,269	307
$\tau = 0.249$ (balanceado interno)	0.6721 (0.6588–0.6854)	0.4662 (0.4520–0.4803)	0.5692 (0.5596–0.5785)	2,546	1,564
$\tau = 0.300$ (conservador)	0.5551 (0.5411–0.5691)	0.5857 (0.5714–0.5995)	0.5704 (0.5602–0.5801)	1,976	2,122

C. Robustez Externa Além da AUC

Além da redução de AUC na avaliação externa, o ramo de glaucoma exibiu calibração inferior e acentuada dependência do limiar operacional. A elevação desse limiar diminuiu os falsos positivos, porém incrementou os falsos negativos. O ajuste, desse modo, redistribuiu a fragilidade sob deslocamento de domínio, sem removê-la.

D. Experimentos Multitarefa Versus Tarefa Única e Baselines

No experimento exploratório conduzido no REFUGE-CLAHE, a configuração multitarefa alcançou AUC de 0,8060, ao passo que as configurações de tarefa única examinadas registraram AUCs situadas entre 0,4619 (CLAHE com focal loss em 224 × 224) e 0,5610 (CLAHE com recorte central de 80% em 512 × 512). Essa comparação deve ser lida como evidência experimental restrita às configurações avaliadas, e não como demonstração geral de superioridade da abordagem multitarefa. O desbalanceamento e a escassez crítica de dados de glaucoma constituem os fatores que atribuem caráter estritamente preliminar a esses achados, cuja consolidação científica demandará a inclusão de um baseline para retinopatia diabética e o desenvolvimento de modelos independentes com backbones separados em ablações futuras.

E. Resultados de MC Dropout para Glaucoma

A utilidade do MC Dropout na triagem de glaucoma variou entre as coortes de forma consistente com sensibilidade ao deslocamento de domínio. No REFUGE-CLAHE, um desvio-padrão médio mais elevado (0,0539 versus 0,0218 nos acertos) e uma maior entropia (0,6017 contra 0,2456), com um pico nítido nos falsos positivos (DP de 0,0578 e

entropia de 0,6443), foram os comportamentos apresentados pelos erros. Uma clara distinção entre falsos positivos e falsos negativos (DP de 0,0429 vs 0,0176; entropia de 0,6581 vs 0,1936) também foi estabelecida pela incerteza no cenário do ACRIMA-CLAHE. Em contrapartida, uma dissolução quase completa dessa separação foi o que ocorreu no EyePACS-AIROGS-light-V2-CLAHE, ambiente em que uma entropia de 0,5559 nos acertos e de 0,5638 nos erros acabou sendo quantificada.

$$\Delta U = \mathrm{E}\left[U_i \mid \widehat{y}_i \neq y_i\right] - \mathrm{E}\left[U_i \mid \widehat{y}_i = y_i\right], \tag{19}$$

em que U_i representa a métrica de incerteza, como entropia ou desvio-padrão preditivo. Valores positivos e maiores de ΔU indicam maior separação entre predições incorretas e corretas.

F. Resultados de MC Dropout para Retinopatia Diabética

Uma separação de acertos e erros mais consistente do que a observada no glaucoma sugeriu maior capacidade informativa do MC Dropout para a retinopatia diabética nas coortes avaliadas, inclusive no cenário de avaliação externa. Maior entropia média (1,0161 vs 0,5263 nas corretas) e maior desvio-padrão preditivo (0,0442 vs 0,0122) caracterizaram as predições incorretas no APTOS, ao passo que aumentos acentuados na entropia (de 0,3996 para 0,7373) e no desvio-padrão (de 0,0147 para 0,0295) marcaram o comportamento do modelo no EyePACS. Por fim, os maiores níveis de incerteza do sistema acabaram se concentrando nos falsos positivos, registrando marcas de entropia de 1,0587 no APTOS e 1,2566 no EyePACS. Para que a predição seletiva fosse avaliada [19], foram mantidas apenas as amostras cujo nível de incerteza ficou abaixo de um ponto de corte q:

$$R(q) = \{i : U_i \leq q\}, \tag{20}$$

e a acurácia dos casos retidos foi calculada como:

$$\mathrm{Acc}_{\mathrm{retained}}(q) = \frac{1}{|R(q)|} \sum_{i \in R(q)} I\left(\widehat{y}_i = y_i\right). \tag{21}$$

Um salto expressivo na acurácia retida de 0,9195 para 1,0000 no APTOS e de 0,8191 para 0,8951–0,9013 no EyePACS foi provocado pela redução da cobertura de 100% para 50%. Em contrapartida, uma evolução bem mais contida caracterizou o quadro de glaucoma, no qual uma elevação na acurácia retida de 0,8925 para 0,9650 foi obtida no REFUGE-CLAHE, ao passo que ganhos meramente marginais ou instáveis foram apresentados por ACRIMA-CLAHE e EyePACS-AIROGS-light-V2-CLAHE. Assim, esses achados sugerem maior utilidade da predição seletiva para a retinopatia diabética do que para o glaucoma nas coortes avaliadas, sem estabelecer superioridade geral fora desses cenários experimentais.

G. Análise Explicável de Erro e Incerteza Baseada em Superpixels Sob Desvio de Domínio Externo

A análise XAI-SLIC combinou mapas Grad-CAM e superpixels SLIC em 381 imagens revisadas manualmente das bases APTOS, EyePACS, REFUGE-CLAHE, ACRIMA-CLAHE e EyePACS-AIROGS-light-V2-CLAHE. As amostras foram agrupadas, conforme disponibilidade de dados, em predições corretas de baixa incerteza, corretas de alta incerteza, falsos positivos e falsos negativos. Foram avaliadas entropia explicativa normalizada, área percentual de alta atribuição, razão de atribuição nos 10 principais superpixels e métricas complementares.

Atribuições mais concentradas foram mantidas pelas predições corretas com baixa incerteza, ao passo que explicações visivelmente mais difusas foram apresentadas por cenários de maior incerteza. No âmbito das bases APTOS e EyePACS, uma redução expressiva na razão Top-10 (passando de 0,602 para 0,230 e de 0,724 para 0,266, respectivamente) combinada a um salto na entropia (de 0,856 para 0,960 e de 0,769 para 0,951, respectivamente) foram as dinâmicas registradas. A reprodução desse mesmo comportamento caracterizou a análise do glaucoma (Tabela V), cenário no qual declínios acentuados na razão Top-10 foram observados no REFUGE-CLAHE (de 0,593 para 0,207), no ACRIMA-CLAHE (de 0,595 para 0,281) e no EyePACS-AIROGS-light-V2-CLAHE (de 0,467 para 0,238), de forma simultânea à estabilização da entropia em patamares próximos a 0,94–0,95.

A análise XAI-SLIC deve ser interpretada como avaliação post-hoc da consistência espacial das atribuições, não como validação clínica. Máscaras de lesão, mapeamento do disco óptico e validação por especialistas não foram realizados.

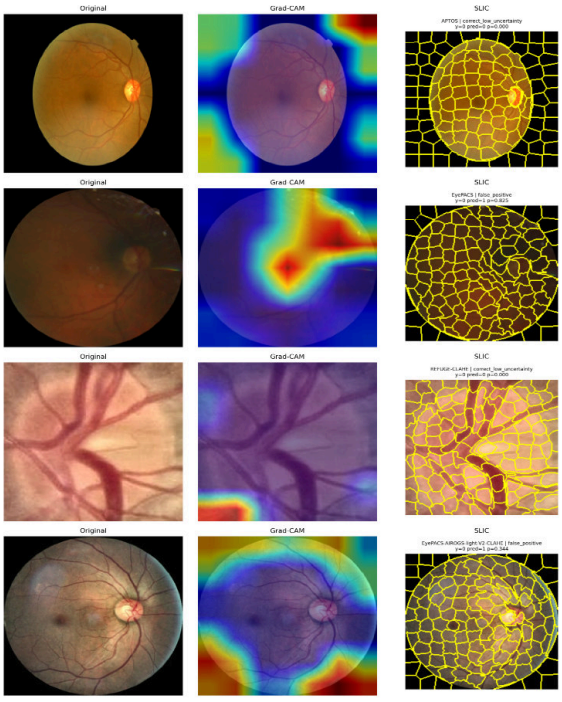


Fig. 2. Explicações representativas por Grad-CAM e SLIC. Imagens originais, sobreposições Grad-CAM e mapas de superpixels SLIC ilustram os padrões de atribuição espacial em casos de triagem por imagens de fundo de olho.

As atribuições mais densas ocorreram nas predições corretas de baixa incerteza; erros e casos incertos produziram explicações mais difusas. O efeito foi maior na retinopatia diabética (EyePACS, $\Delta = 0,458$) do que no glaucoma sob deslocamento externo (EyePACS-AIROGS-light-

Tabela V
MÉTRICAS DE EXPLICABILIDADE EM DIFERENTES DESFECHOS DE PREDIÇÃO

Conjunto de dados	Baixo	Alto	Δ	Padrão
DR–APTOS	0.602	0.230	0.372	Forte
DR–EyePACS	0.724	0.266	0.458	Forte
G–REFUGE-CLAHE	0.593	0.207	0.386	Bem definido
G–ACRIMA-CLAHE	0.595	0.281	0.314	Heterogêneo
G–EyePACS-V2-CLAHE	0.467	0.238	0.229	Mais fraco

V2, $\Delta = 0,229$), sugerindo menor estabilidade espacial das explicações em cenários de domain shift.

H. Temperature Scaling

O Temperature Scaling teve efeito dependente da tarefa. Para retinopatia diabética, $T_{DR} = 0,9358$ preservou a acurácia em 0,7804, praticamente não alterou a AUC OVR multiclasse (0,9240 para 0,9239) e reduziu o ECE de 0,0379 para 0,0276; por isso, foi mantido. Para glaucoma, na análise auxiliar REFUGE raw, $T = 1,2739$ preservou a AUC em 0,7581, mas piorou Brier (0,0881 para 0,0900) e ECE (0,0587 para 0,0741), sendo rejeitado. Esses valores de calibração são distintos das métricas operacionais do pipeline principal.

I. Resumo da Ablação em Nível de Componentes

A presente análise de ablação caracterizou efeitos dependentes da tarefa entre os componentes do modelo multitarefa. No ramo de glaucoma, a aplicação de CLAHE elevou a AUC, ao passo que o deslocamento de limiares voltado à mitigação de falsos positivos culminou em incremento correspondente de falsos negativos. O MC Dropout apresentou capacidade variável de separar acertos e erros e perdeu utilidade sob deslocamento externo acentuado; em contrapartida, o Temperature Scaling reduziu o ECE na validação interna da retinopatia diabética, mas piorou Brier score e ECE na análise auxiliar de glaucoma. Tais resultados evidenciam que as saídas do modelo multitarefa demandam a implementação de estratégias de pós-processamento específicas, condicionadas ao escopo individual de cada tarefa, em vez da adoção de uma configuração uniforme aplicada indistintamente.

J. Eficiência Computacional em Comparação com Modelos Independentes

No benchmark com 400 imagens REFUGE-CLAHE em GPU A100, o modelo multitarefa obteve 0,4661 ms/imagem e 30,99 MB, contra 0,9458 ms/imagem e 61,93 MB para dois EfficientNet-B0 sequenciais. Isso correspondeu a 2,03$\times$ menor latência e 49,96% menos armazenamento, ganhos de engenharia que não implicam superioridade clínica.

IV. DISCUSSÃO

Os resultados evidenciam heterogeneidade entre tarefas e domínios. A retinopatia diabética apresentou melhor discriminação relativa, enquanto o glaucoma foi mais vulnerável à transferência; no ACRIMA-CLAHE, por exemplo, o IC95% da AUC (0,4382–0,6625) inclui 0,50. Limiares, MC Dropout, Temperature Scaling e Grad-CAM/SLIC também responderam de modo dependente da tarefa e do domínio. A vantagem computacional do backbone compartilhado deve ser interpretada apenas como ganho de engenharia, não como evidência de superioridade clínica.

O pipeline deve, portanto, ser entendido como protótipo de pesquisa auditável, com métricas recomputáveis a partir das predições preservadas. Controle de qualidade, segmentação do disco óptico, adaptação de domínio, coortes mais representativas e validação prospectiva permanecem prioridades.

V. LIMITAÇÕES

O glaucoma permaneceu sensível ao deslocamento de domínio, e os modelos de tarefa única devem ser considerados exploratórios. MC Dropout utilizou apenas dez passagens, enquanto Grad-CAM/SLIC não contou com segmentação anatômica nem validação por especialistas.

Não houve teste interno independente. Para retinopatia diabética, a auditoria identificou 41 hashes SHA-256 compartilhados entre treinamento e validação do APTOS, envolvendo 83 registros; portanto, a estimativa interna pode ser otimista e deve ser interpretada como resultado de desenvolvimento. Para glaucoma, as mesmas 400 imagens REFUGE-CLAHE foram usadas no exame dos pontos de operação e na estimação das métricas internas dependentes de limiar.

A ACRIMA-CLAHE é pequena ($n = 106$), o que amplia a incerteza. Além disso, o bootstrap foi realizado em nível de imagem; no EyePACS, a possível presença de dois olhos do mesmo paciente pode violar independência e estreitar os IC95%, pois não havia identificadores suficientes para reamostragem agrupada por paciente.

VI. CONCLUSÃO

O pipeline multitarefa mostrou melhor robustez relativa para retinopatia diabética e maior vulnerabilidade externa para glaucoma. A elevada estimativa interna do APTOS deve ser interpretada com cautela devido à sobreposição retrospectivamente identificada entre treinamento e validação, enquanto o EyePACS fornece a principal evidência externa para essa tarefa. CLAHE, limiares, calibração e incerteza apresentaram efeitos dependentes do domínio. O sistema deve, portanto, ser considerado um protótipo experimental auditável, não uma ferramenta clínica validada. Trabalhos futuros devem priorizar glaucoma, adaptação de domínio e validação prospectiva em nível de paciente.

REFERÊNCIAS

[1] Asia Pacific Tele-Ophthalmology Society (APTOS), "APTOS 2019 Blindness Detection," Kaggle Competition, 2019. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed: Apr. 22, 2026.

[2] C. de Vente et al., "AIROGS: Artificial Intelligence for ROBust Glaucoma Screening Challenge," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 542–557, Jan. 2024, doi: 10.1109/TMI.2023.3313786.

[3] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "CNNs for automatic glaucoma assessment using fundus images: an extensive validation," *BioMedical Engineering OnLine*, vol. 18, no. 1, p. 29, Mar. 2019, doi: 10.1186/s12938-019-0649-y.

[4] R. Tan et al., "Evaluating the outcome of screening for glaucoma using colour fundus photography-based referral criteria in a teleophthalmology screening programme for diabetic retinopathy," *British Journal of Ophthalmology*, vol. 108, no. 7, pp. 933–939, Jul. 2024, doi: 10.1136/BJO-2023-323339.

[5] EyePACS, "Diabetic Retinopathy Detection," Kaggle Competition, 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed: Apr. 22, 2026.

[6] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[7] V. Gulshan et al., “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.

[9] Z. Li et al., “Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on colour fundus photographs,” *Ophthalmology*, vol. 125, pp. 1199–1206, 2018.

[10] J. I. Orlando et al., “REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs,” *Medical Image Analysis*, vol. 59, p. 101570, Jan. 2020, doi: 10.1016/j.media.2019.101570.

[11] S.-L. Yi, X.-L. Yang, T.-W. Wang, F.-R. She, X. Xiong, and J.-F. He, “Diabetic retinopathy diagnosis based on RA-EfficientNet,” *Applied Sciences*, vol. 11, no. 22, Art. no. 11035, 2021, doi: 10.3390/app112211035.

[12] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[13] B. A. Cooper, “Ten key elements of a diabetes-related eye examination,” in *A Practical Guide to Diabetes-Related Eye Care*. Arlington, VA, USA: American Diabetes Association, Jul. 2022, doi: 10.2337/db20223-4.

[14] D. S. W. Ting et al., “Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes,” *JAMA*, vol. 318, pp. 2211–2223, 2017.

[15] J. Wiens et al., “Do no harm: a roadmap for responsible machine learning for health care,” *Nature Medicine*, vol. 25, pp. 1337–1340, 2019.

[16] M. J. Willemink et al., “Preparing medical imaging data for machine learning,” *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020, doi: 10.1148/radiol.2020192224.

[17] K. Zuiderveld, “Contrast limited adaptive histogram equalization,” in *Graphics Gems IV*, P. S. Heckbert, Ed. Academic Press, 1994, pp. 474–485.

[18] M. Minderer et al., “Revisiting the calibration of modern neural networks,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 15682–15694.

[19] Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4878–4887.

[20] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2016, pp. 1050–1059.

[21] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.

[22] M. Kamnitsas et al., “Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,” in *Proc. Int. Conf. Information Processing in Medical Imaging (IPMI)*, 2017, pp. 597–609.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012, doi: 10.1109/TPAMI.2012.120.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858825.

[26] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

Um sistema baseado em CNN 3D para diferenciação entre esquizofrenia, transtorno bipolar e controles saudáveis com base em ressonância magnética.

Rafael Lorenzo Baima Lamar Araujo
Laboratório de Processamento da Informação Biológica
Universidade Federal do Maranhão
São Luís,Brasil
rafael.lorenzo@discente.ufma.br

Allan Kardec Duailibe Barros Filho
Coordenação do curso de Engenharia Elétrica
Universidade Federal do Maranhão
São Luís,Brasil
allan.kardec@ufma.br

Priscila Lima Rocha
Departamento de Eletroeletrônica
Instituto Federal do Maranhão
São Luís,Brasil
priscila.rocha@ifma.edu.br

Thaiane Coelho dos Santos
Laboratório de Processamento da Informação Biológica
Universidade Federal do Maranhão
São Luís,Brasil
thaiane.coelho@discente.ufma.br

Gilberto Sousa Alves
Serviço de psiquiatria
Hospital Nina Rodrigues
São Luís,Brasil
Gilberto.sousa@ufma.br

Diego Dutra Sampaio
Laboratório de Processamento da Informação Biológica
Universidade Federal do Maranhão
São Luís,Brasil
dd.sampaio@discente.ufma.br

Resumo— **Este artigo descreve a implementação de uma arquitetura de redes neurais convolucionais 3D com o objetivo de classificar pacientes nas categorias controle, esquizofrenia e transtorno bipolar, com base em características extraídas de dados de MRI. A arquitetura utilizada consiste em uma rede neural híbrida multimodal do tipo multi-input com fusão tardia, combinando uma CNN 3D para extração de características espaciais ao longo da sequência dos vídeos com uma rede MLP para processamento de atributos tabulares. O modelo apresentou acurácia de 61,9% (IC95%: 42,9%–81,0%), acurácia balanceada de 61,9% (IC95%: 42,9%–81,0%), precisão macro de 61,3% (IC95%: 38,3%–83,3%), sensibilidade macro de 61,9% (IC95%: 42,9%–81,0%) e F1-score macro de 60,7% (IC95%: 39,5%–80,4%). A AUC macro, calculada pela estratégia One-vs-Rest, foi de 0,738 (IC95%: 0,588–0,878). Os resultados indicam capacidade discriminativa inicial entre as classes avaliadas.**

Palavras-chaves: **Redes Neurais Convolucionais 3D, Ressonância Magnética, Classificador, Esquizofrenia, Transtorno Bipolar**

I. INTRODUÇÃO

A esquizofrenia e o transtorno bipolar estão entre os transtornos psiquiátricos mais complexos, pois embora clinicamente parecidos possuem características sintomáticas distintas [1]. O transtorno bipolar caracteriza-se principalmente por acentuadas oscilações de humor e níveis de energia, as quais podem comprometer significativamente o desempenho nas atividades cotidianas. Em contraste, na esquizofrenia, há presença de distorções perceptivas que podem dificultar a distinção entre realidade e irrealidade [1]. Além disso, indivíduos com esquizofrenia podem apresentar prejuízos nos processos cognitivos e nas interações sociais [1].

Nas últimas décadas, a neuroimagem tem proporcionado uma perspectiva promissora e não invasiva para o estudo do cérebro humano. Técnicas baseadas em ressonância magnética, como a ressonância magnética estrutural (sMRI), a ressonância magnética funcional (fMRI) e a ressonância magnética por difusão (dMRI), possibilitam a obtenção de informações espacialmente detalhadas acerca da estrutura e do funcionamento cerebral, além de permitirem a análise da conectividade funcional e estrutural. O avanço dessas metodologias tem ampliado significativamente a compreensão sobre o cérebro humano e fortalecido as perspectivas de pesquisadores dedicados a investigar um dos sistemas mais complexos conhecidos, o funcionamento cerebral [2].

Os sistemas de compreensão de imagens baseados em técnicas de aprendizado de máquina (Machine Learning – ML) avançaram na área de neuroimagem para distúrbios cerebrais [3]. Entre as abordagens mais utilizadas, destacam-se algoritmos como árvores de decisão, métodos de agrupamento (clustering), máquinas de vetores de suporte (SVM), k-vizinhos mais próximos (k-NN), máquinas de Boltzmann restritas (RBMs) e florestas aleatórias (Random Forests). Um dos principais desafios na aplicação dessas técnicas reside na etapa de extração de características discriminantes, que em muitos casos, depende de conhecimento prévio da pessoa que está manuseando, especialmente em aplicações relacionadas à análise e compreensão de imagens [3].

Nesse contexto, modelos baseados em aprendizado profundo têm ganhado destaque atualmente, em particular as redes neurais convolucionais (Convolutional Neural Networks – CNN). Essas arquiteturas são capazes de aprender automaticamente representações relevantes a partir dos dados, eliminando, em grande medida, essa etapa de extração manual de características. As CNNs são compostas por camadas convolucionais formadas por filtros que atuam na detecção e extração de padrões essenciais, contribuindo para uma interpretação mais eficiente e precisa dessas imagens médicas [3].

Estudos anteriores utilizaram técnicas de aprendizado profundo a partir de imagens de ressonância magnética [4] [7] e utilizaram redes neurais para analisar padrões de ciclos temporais a partir de rsfMRI [5]. Apesar desses avanços, a distinção entre dois distúrbios permanece desafiadora [6].

Nesse estudo, foi desenvolvido um modelo de rede neural convolucional tridimensional (CNN 3D) com o objetivo de classificar pacientes em três categorias distintas: indivíduos com esquizofrenia, transtorno bipolar e controles saudáveis. A escolha por uma abordagem tridimensional justifica-se pela natureza volumétrica das imagens de ressonância magnética, permitindo a exploração mais completa das informações espaciais presentes nos dados cerebrais.

II. MATERIAIS E MÉTODOS

A. Base de dados e aquisição de imagens

Todos os dados foram disponibilizados pelo Hospital Nina Rodrigues. O recrutamento ocorreu entre julho de 2017 e abril de 2019, no serviço ambulatório de humor e psicose do Hospital Nina Rodrigues. A amostra foi composta especificamente por indivíduos de etnia afrodescendente (negros e pardos) e indígenas. Os diagnósticos foram estabelecidos de acordo com os critérios do Manual Diagnóstico e Estatístico de Transtornos Mentais, 5ª edição (DSM-5), da Associação Americana de Psiquiatria. [8]. A Entrevista Diagnóstica para Psicose e Transtornos Afetivos (Di-PAD), baseada nos algoritmos do DSM-IV, foi utilizada para estabelecer o diagnóstico de ESQ e TB, além de avaliar sintomas depressivos e maníacos [8].

A idade dos participantes variou entre 22 e 31 anos, com mediana de idade de 25 anos entre os 36 homens e 28 mulheres, contudo o sexo não foi incluído como covariável no modelo estatístico principal. Em relação à composição étnica, 49 participantes foram classificados como pardos, 14 como negros e um como indígena. Os critérios de exclusão incluíram casos de transtornos orgânicos como tumor cerebral, AIDS, trauma craniano ou infecções no sistema nervoso central, casos de deficiência intelectual ou dependência química foram retirados também [8] [9].

O conjunto de dados inclui imagens de ressonância magnética estrutural no formato NIfTI (Neuroimaging Informatics Technology Initiative) de 20 pacientes bipolares, 23 pacientes esquizofrênicos e 21 indivíduos saudáveis. Cada participante contribuiu, em média, com 160 imagens, abrangendo cortes coronais, sagitais e axiais do volume cerebral. O software de processamento utilizado foi o FreeSurfer 7.4.1 com o atlas Desikan-Killiany que segmenta o córtex em 68 regiões de interesse [9].

No total, foram selecionadas 10.240 imagens de 64 sujeitos. Todas as imagens foram analisadas por profissionais do hospital para eliminar aquelas com movimento excessivo ou ruídos que pudessem comprometer a eficácia do algoritmo de classificação.

B. Comitê de Ética

A utilização dos dados empregados neste estudo foi aprovada pelo Comitê de Ética em Pesquisa do Hospital Nina Rodrigues, sob o Certificado de Apresentação para Apreciação Ética (CAAE) nº 45.604.215.3.1001.5162 e seguiu os protocolos estabelecidos pela declaração de Helsinque, em conformidade com a Resolução nº 466, de 12 de dezembro de 2012, do Conselho Nacional de Saúde. Todos os participantes assinaram o termo de consentimento livre e Esclarecido (TCLE) após explicação completa, com autorização adicional por escrito de seus familiares. As imagens de ressonância magnética foram disponibilizadas pelo Hospital Nina Rodrigues exclusivamente para fins de pesquisa científica e previamente anonimizadas, garantindo a confidencialidade e a privacidade dos participantes

C. Aquisição e pré – processamento das imagens

Todas as imagens no formato NIfTI foram submetidas a uma etapa de pré-processamento e avaliadas manualmente pelos profissionais por meio do software FSL (FMRIB Software Library), desenvolvido pelo Oxford Centre for Functional MRI of the Brain (FMRIB). Esse conjunto de ferramentas é utilizado principalmente em pesquisas neurocientíficas para o processamento e análise de dados de ressonância magnética [15].

Inicialmente, para a correção de artefatos e distorções decorrentes do movimento dos pacientes durante a aquisição das imagens, empregou-se a ferramenta MCFLIRT do FSL. Esse método realiza o registro e o alinhamento das imagens em um espaço comum, garantindo maior consistência dos dados e aumentando a precisão das análises subsequentes. Em seguida, procedeu-se à remoção do crânio por meio do algoritmo BET (Brain Extraction Tool), responsável por segmentar e extrair o tecido cerebral, eliminando estruturas não cerebrais. Após essa etapa, foi realizada uma inspeção visual por especialistas do hospital, com o objetivo de assegurar a qualidade da extração e identificar possíveis falhas, tais como remoção inadequada de tecidos, erros operacionais ou saídas corrompidas.

Posteriormente, realizou-se o registro linear das imagens utilizando a ferramenta FLIRT (FMRIB's Linear Image Registration Tool). Esse processo envolveu a normalização das imagens de cada paciente para o espaço anatômico padrão MNI152, com resolução de 1 mm^3, por meio de transformações lineares, incluindo rotações, translações e escalonamentos [16]. Tal procedimento possibilitou a padronização dos dados, permitindo comparações consistentes entre diferentes indivíduos.

Em seguida, foi aplicada a ferramenta FNIRT para a realização do registro não linear, com o objetivo de ajustar variações anatômicas individuais e corrigir distorções mais complexas, especialmente na região da substância cinzenta, no espaço MNI152. Essa etapa contribuiu para o aumento da precisão nas análises interindividuais.

Assim, foram gerados dois conjuntos de neuroimagens para cada paciente. O primeiro conjunto, mostrado na Figura 1, consiste em imagens normalizadas para 1 mm3 no espaço anatômico MNI152. O segundo conjunto, representado na Figura 2, foi obtido a partir de imagens pré-processadas com a aplicação da máscara da substância cinzenta, também no padrão MNI152 utilizado o FNIRT, com o objetivo de realçar a região da substância cinzenta do cérebro.

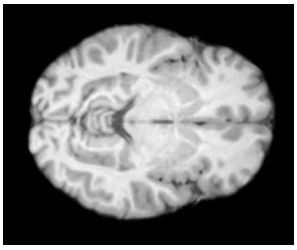


Figura 1- Exemplo de imagem de ressonância magnética normalizadas para 1 mm3 no espaço anatômico MNI152.



Figura 2- Exemplo de máscara de substancia cinzenta no espaço MNI152.

D. Extração de características

A etapa de extração de características foi realizado com o intuito de transformar os dados volumétricos de neuroimagem em representações numéricas compactas e discriminativas, adequadas para tarefas de classificação supervisionada.

Inicialmente, os volumes de imagens de ressonância magnética foram organizados como sequências tridimensionais $A \in \mathbb{R}^{T \times H \times W}$, onde T representa o número de cortes (frames), e $H \times W$ correspondem às dimensões espaciais de cada imagem. Paralelamente, foi considerada uma máscara cerebral $B \in \mathbb{R}^{T \times H \times W}$, responsável por delimitar regiões de interesse.

Como etapa fundamental de padronização, aplicou-se a normalização do tipo *z-score* sobre os valores de intensidade da imagem, considerando apenas os voxels pertencentes à região cerebral utilizando a máscara como guia. Essa transformação é definida pela equação 1:

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

Em que x representa a intensidade original do voxel, μ a média e σ o desvio padrão calculados exclusivamente sobre os voxels da máscara cerebral. Após a normalização, foram extraídas características estatísticas ao longo da sequência de cortes, considerando apenas os voxels pertencentes ao cérebro. Para cada frame T, foi definida uma região de interesse $\Omega = \{(i,j) | B\ (i,j) > \tau\}$, sendo τ um limiar. A média de intensidade em cada frame T é dada pela equação 2:

$$\mu_T = \frac{1}{|\Omega_T|} \sum_{(i,j \in \Omega_T)} A_T(i,j) \qquad (2)$$

Para cada frame, foram calculadas as seguintes medidas: média, desvio padrão, mediana, intervalo interquartil, desvio absoluto mediano, assimetria e curtose. Essas medidas permitem capturar propriedades centrais, dispersão e forma da distribuição das intensidades dos voxels cerebrais. Além das estatísticas por frame, foram calculadas características baseadas na variação temporal entre cortes consecutivos, definidas pela equação 3:

$$\Delta A_T = A_T + 1 - A_T \qquad (3)$$

A partir dessas diferenças, foram extraídas novamente as mesmas estatísticas, permitindo modelar mudanças estruturais ao longo do eixo volumétrico.

Após a extração inicial, os dados foram particionados em conjuntos de treinamento e teste de forma independente por participante, garantindo que um mesmo paciente não estivesse presente em ambos os conjuntos. A etapa de seleção de características foi então realizada exclusivamente sobre o conjunto de treinamento com o objetivo de reduzir redundâncias com objetivo de melhorar a capacidade de generalização do modelo. Primeiramente, todas as características foram padronizadas via z-score, garantindo comparabilidade entre diferentes escalas. Em seguida, aplicou-se o teste ANOVA (*F-test*) para avaliar a relevância estatística de cada característica em relação às classes, selecionando aquelas com valor de $p < 0.05$.

E. Particionamento dos Dados

Os dados foram divididos por paciente, cada participante foi representado uma sequência tridimensional contendo 160 cortes axiais, onde cada corte da seção transversal foi convertido em um frame, e esses frames foram combinados em um vídeo de 16 segundos de duração e com dimensões de 96x96 pixels.

A proporção adotada para a divisão dos dados foi de 70% dos pacientes para treinamento e 30% dos pacientes para teste, onde a separação foi realizada de forma aleatória, não sfoi utilizada a validação cruzada, pois essa estratégia foi adotada considerando o custo computacional da arquitetura 3D proposta e a possibilidade, por conta do tamanho reduzido da amostra, uma única partição pode produzir estimativas dependentes da composição do conjunto de teste.

Para evitar vazamento nenhuma imagem de um paciente presente no conjunto de treinamento foi incluída no conjunto de teste, essa separação foi feita utilizando o identificador do participante como variável de agrupamento no algoritmo GroupShuffleSplit. Assim, embora cada participante tenha contribuído com aproximadamente 160 fatias para a representação volumétrica, essas fatias não foram consideradas observações independentes durante a divisão dos dados.

Não houve aumento da base de dados por meio de técnicas como deslocamento ou ruído, uma vez que cada participante é representado por um volume completo de 160 cortes axiais.

F. Entradas na rede neural

Para entrada da rede neural, medições do número de cortes no eixo Z, correspondendo ao topo do crânio, e das coordenadas XY do mesencéfalo foram realizadas a partir da visualização axial de cada imagem como demonstrado na figura 3.

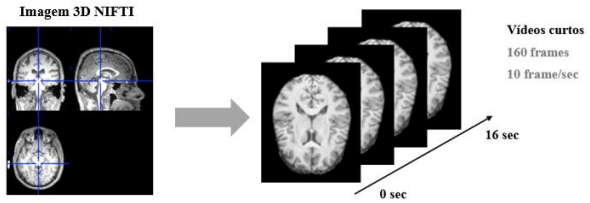


Figura 3-Processo de conversão de uma imagem cerebral 3D NIfTI em vídeos curtos de slices axiais.

A ressonância magnética estrutural utilizada não corresponde a um vídeo real, pois os cortes adquiridos representam diferentes posições anatômicas do cérebro. A utilização dos vídeos ocorreu como estratégia para preservar a sequência espacial dos cortes axiais do volume cerebral. A escolha dessa abordagem também ocorreu por se assemelhar a abordagem clínica de interpretação dessas imagens pelos médicos.

G. Fusão tardia

As caraterísticas tabulares são processadas por uma MLP que gera uma representação vetorial, paralelamente a CNN 3D processa todos os mapas de características e calcula a media dos valores daquele mapa e os transformar em um vetor de caraterísticas que representa as informações aprendidas pela rede de forma compacta.

Os dois vetores são concatenados para gerar um vetor de características multimodal sendo encaminhado para as camadas seguintes que irão produzir as probabilidades de pertencimento a uma das classes propostas como demonstrado pela figura 4.

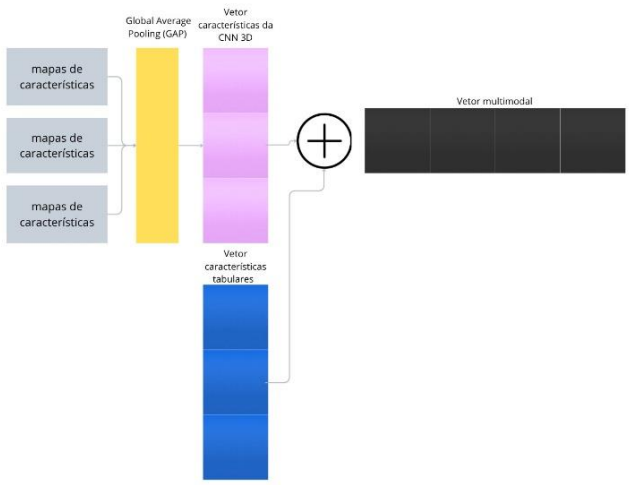


Figura 4- Esquema da fusão tardia da rede.

H. Rede neural 3D

A arquitetura da rede neural consiste em três blocos convolucionais 3D sequenciais. O primeiro bloco da rede é composto por camadas convolucionais tridimensionais com 24 filtros e kernels de dimensão 3×3×3 para extração de padrões locais iniciais; em seguida, são aplicadas operações de normalização e funções de ativação do tipo ReLU, introduzindo não-linearidade e estabilizando o processo de treinamento. É realizado um MaxPool3D para reduzir a resolução espacial (altura e largura) pela metade em cada uma das medidas, mas mantendo a dimensão temporal *T* inalterada.

O segundo bloco convolucional segue estrutura semelhante, porém com aumento do número de filtros para 48, permitindo detectar padrões mais complexos e combinações de características do bloco anterior.

No terceiro bloco, o número de filtros é expandido para 72, permitindo a extração de representações ainda mais abstratas e o pooling reduz a dimensão temporal, dimensão altura e dimensão largura para diminuir risco de overfitting.

Após os três blocos, a rede aplica uma Global Average Pooling 3D, que reduz toda a informação do espaço temporal em um vetor de 72 características. Por fim, uma camada densa com ativação ReLU processa essas características, preparando-as para a concatenação com as features tabulares e a classificação final.

Paralelamente ao processamento das imagens pela CNN 3D, as características tabulares selecionadas são inseridas em uma camada de normalização e posteriormente processadas por uma camada densa de 64 neurônios com ativação ReLU, seguida de Dropout de 0,15. As saídas dos dois ramos (convolucional e de características) são concatenadas e encaminhadas para ás camadas finais de classificação. O otimizador utilizado foi o AdamW com taxa de aprendizado de 3×10^{-4}, associado à função de perda Focal Loss que foi utilizado para lidar com possíveis desbalanceamentos entre as classes. A arquitetura da rede neural e representada pela figura 5 com as camadas de convolução e entrada das caraterísticas para a decisão.

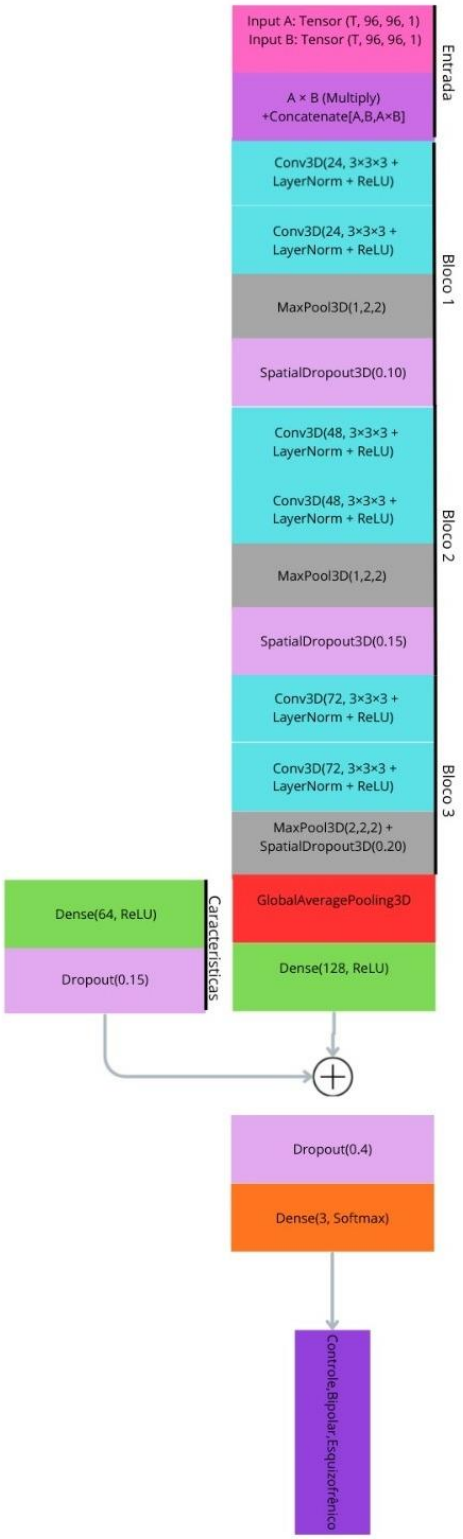


Figura 5-Arquitetura da rede neural 3D.

I. Treinamento e seleção de hiperparâmetros

Os hiperparâmetros foram primordialmente escolhidos com base na literatura utilizada em tarefas similares. Consecutivamente, a partir de experimentos exploratórios e

ajustes iterativos do modelo e considerando o desempenho observado no conjunto de teste se chegou nos parâmetros da rede observados na tabela 1. Os experimentos foram executados utilizando o ambiente Google Colab com aceleração por GPU, com implementação em Python utilizando o framework TensorFlow/Keras. Foi utilizado uma GPU NVIDIA Tesla T4 com 16 GB de memória dedicada.

Tabela 1: Parâmetros da rede

Parâmetro	Valor
Tensor de entrada de cada vídeo	160 × 96 × 96 × 1
Número de pacientes (Treino/Teste)	64 (43/21)
Filtros Conv3D	24,48,72
Otimizador	AdamW
Número máximo de épocas	100
Dropout espacial	0.10;0.15; 0.20
Taxa de aprendizado inicial	3×10^{-4}
Função de perda	Focal Loss ($\gamma = 2,0$)

III. Resultados e discursões

A. *Metricas de avaliação por nivel de participante*

A métricas foram avaliadas considerando cada participante como uma unidade. A acurácia calcula a taxa de acertos do modelo dividido pelo total de casos, a acurácia balanceada é calculada pela média do recall de cada classe. A Precisão macro corresponde à média aritmética das precisões obtidas individualmente para cada classe, atribuindo o mesmo peso a todas elas. A sensibilidade corresponde à média dos valores de recall das três classes. O F1 combina precisão e sensibilidade por meio da média harmônica dessas duas medidas, enquanto o F1-score macro corresponde à média dos F1-scores das classes [10]. Os resultados do modelo estão descritos na tabela 2.

Tabela 2: Resultados nas métricas.

Métricas do teste em nível de participante	
Acurácia	0,6190
Acurácia Balanceada	0,6190
Precisão macro	0,6127
Sensibilidade macro	0,6190
F1 score macro	0,6071

A acurácia está acima de um chute aleatório, contudo ainda não é o bastante para apresentar uma boa capacidade de generalização do modelo, o valor semelhante de sensibilidade macro e da acurácia balanceada se dá principalmente porque o conjunto teste possui numero de participantes em cada classe sendo sete pacientes para cada classe durante o teste.

A dificuldade em identificar os participantes em suas respectivas classes e demonstrando principalmente pelo baixo valor da sensibilidade, indicado que um parcela dos participantes não foi categorizada de forma correta. O F1 e a precisão demonstram uma dificuldade em atribuir os pacientes em suas categorias corretas e um desempenho limitado no equilíbrio e evitar classificações incorretas.

B. *Intervalo de confiança de 95% e desempenho por classe*

O intervalo de confiança é uma faixa de valores que demonstra a incerteza a estimativa de um parâmetro. O nível de 95% foi adotado por ser o mais utilizado em trabalhos semelhantes. A amplitude do intervalo de confiança mede o grau de incerteza do resultado final, para garantir os 95% os limites de 42.9% a 81% foram estipulados como intervalos de incertezas. Os resultados da tabela 3 foram obtidos pelo método de bootstrap estratificado por classe demonstrando o resultado do modelo nesses intervalos.

Tabela 3: Estimas das métricas de avaliação

Métricas	Estimativa	IC95% Inferior	IC95% Superior
Acurácia	0,619	0,429	0,810
Acurácia Balanceada	0,619	0,429	0,810
Precisão macro	0,613	0,383	0,833
Sensibilidade macro	0,619	0,429	0,810
F1 score macro	0,607	0,395	0,804
AUC macro	0,738	0,588	0,878

Os intervalos de confiança apresentam amplitude considerável, refletindo a incerteza associada ao tamanho reduzido da amostra. A acurácia e a sensibilidade macro apresentaram IC95% entre 0,429 e 0,810, enquanto a precisão macro variou entre 0,383 e 0,833. Esses resultados indicam que as estimativas apresentam elevada variabilidade e elevada oscilação entre os valores.

O melhor resultado do sistema se mostra no seu valor de AUC macro, que foi calculado utilizando por meio do One-vs-Rest, que não se intercepta pela limites do acaso mostrando que o algoritmo desenvolveu capacidade relevante de discriminação entre as classes exibindo uma consistência maior que as outras métricas.

C. *Curva ROC*

A curva ROC é usada para avaliar a capacidade discriminativa de um teste e comparar o desempenho de dois ou mais testes . Também é usado para selecionar um valor de corte ótimo para determinar a presença ou ausência de uma doença.[11].

Quanto mais a curva ROC se aproxima do canto superior esquerdo, melhor é a qualidade do teste quanto à sua capacidade para discriminar os grupos. No centro, a linha de referência diagonal da curva ROC representa uma região de completa aleatoriedade do teste, mostrando um modelo incapaz de classificar tanto doentes como saudáveis [12].

Uma das métricas de avaliação é o AUC (Área sobre a curva) que demonstra a capacidade do modelo de diferenciar as classes. Um valor AUC próximo de um 1 indica um bom poder discriminatório, enquanto um valor próximo de 0.5 indica uma escolha aleatória. A curva ROC do modelo obtido é representado pela figura 6, onde a classe controle é representada pela linha de cor azul, a bipolar pela linha de cor

laranja e esquizofrênico pela linha de cor verde e a reta de referência do chute aleatório é a de cor vermelha.

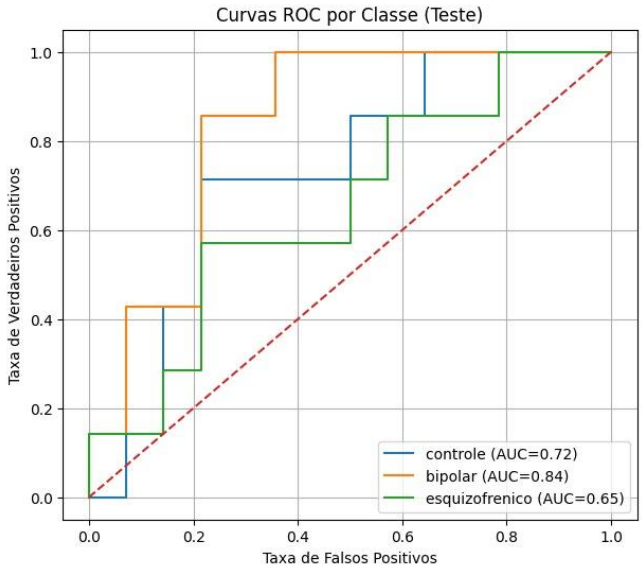


Figura 6- Curva ROC do sistema.

A classe controle apresentou o valor de AUC sendo 0.72 demonstrando uma capacidade de diferenciação do modelo entre pacientes saudáveis em relação aos portadores de transtornos psiquiátricos, com a curva demonstrando elevada sensibilidade mesmo para baixos níveis de falso positivo.

Por outro lado, a classe esquizofrenia obteve AUC de 0,65 sendo o pior de todas as classes avaliadas, enquanto a classe bipolar apresentou o maior desempenho, com AUC de 0,84. Uma das causas desse menor desempenho é a existência de semelhanças clínicas e neurobiológicas entre esses transtornos como, por exemplo, o compartilhamento de características sintomatológicas relacionados [14]. Além disso, pode ocorrer sobreposição nos circuitos cerebrais afetados, o que torna mais difícil a diferenciação, pois os modelos precisam detectar diferenças finas, o que reduz a acurácia.

A limitação do tamanho da base de pacientes também demonstra impacto no resultado, pois em aprendizado profundo, conjuntos de dados reduzidos podem dificultar a generalização do modelo e limitar sua capacidade de aprender características discriminativas mais robustas para cada classe. Além disso modelos treinados com diferentes quantidades de paciente podem favorecer classes mais representadas.

O modelo alcançou um AUC médio de 0.7381, todos os valores de AUC ficaram acima da linha de referência (AUC = 0.50), confirmando que o modelo possui poder preditivo capaz de produzir classificações corretas com destaque a classe bipolar que se saiu superior as outras nos testes.

D. Matriz de confusão

A Matriz de confusão apresenta a distribuição entre as classes reais e as classes previstas pelo modelo [13]. As linhas representam as classes reais, enquanto as colunas representam as classes previstas pelo modelo. A diagonal principal, apresenta os acertos do modelo em prever corretamente a classificação do indivíduo. A matriz do conjunto de teste do modelo gerado é representado pela figura 7.

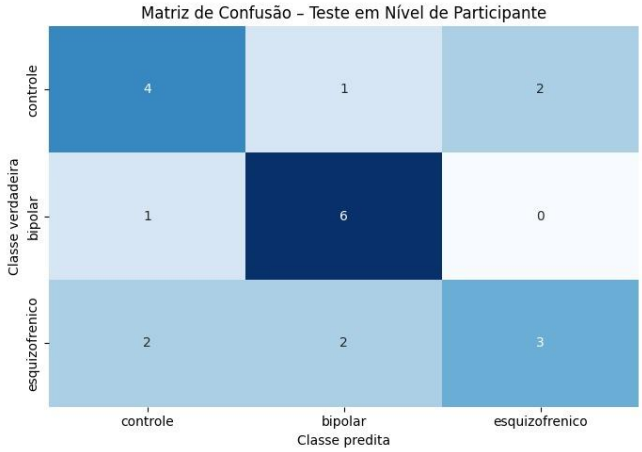


Figura 7- Matriz de confusão do teste.

O modelo previu a classe controle categorizando errado uma de seus participantes para bipolar e dois para esquizofrênico de forma incorreta formando uma sensibilidade para a classe de 57.14%.

Em contrapartida, para a classe esquizofrênico, das sete amostras reais, o modelo acertou apenas três com sensibilidade de 42,86% indicando uma dificuldade em capturar padrões discriminativos adequados para essa categoria e alinhando-se com os resultados obtidos na análise das curvas ROC, onde a classe esquizofrênico, apresentou o menor AUC.

Para a classe de pacientes com bipolar, nenhum deles foi classificado como esquizofrênico e um de seus participantes foram classificados incorretamente como da categoria controle, com uma sensibilidade de 85,71% demonstrando um desempenho superior a toda as outras classes. Demonstrando que o modelo teve facilidade para diferenciar a classe bipolar das outras duas, onde a classe esquizofrênica apresentou erros entre controle e bipolar demonstrando que engano não se concentra em uma classe especifica

IV. Conclusão

Como conclusão do estudo, foi possível observar que há capacidade discriminativa limitada utilizando a rede neural híbrida multimodal e os resultados de experimento exploratório com um conjunto de dados locais. Com esse estudo, se espera contribuir para os estudos sobre a capacidade de modelos de aprendizado profundo em identificar padrões cerebrais associados a transtornos psiquiátricos complexos.

Espera-se que este trabalho utilizado para novas pesquisas que utilizam técnicas de aprendizado profundo, pois os resultados apontam para um potencial das redes neurais convolucionais 3D como ferramenta auxiliar para identificação de transtornos mentais, incentivando o desenvolvimento de modelos mais robustos, interpretáveis e com potencial para ferramenta experimental para estudos futuros.

Portanto, diante dos resultados modelo de rede neural proposto para a base de dados particular, pretende-se aplicar e avaliar o modelo em outras bases de dados que contenham

as mesmas classes e aplicação de técnicas mais avançadas para obter resultados mais expressivos.

V. REFERENCIAS

[1] T. E. Goldberg, “Some fairly obvious distinctions between schizophrenia and bipolar disorder,” Schizophrenia Research, vol. 39, no. 2, pp. 127–132, Sep. 1999, doi: 10.1016/s0920-9964(99)00111-5.

[2] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, “Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls,” NeuroImage, vol. 145, no. Pt B, pp. 137–165, Jan. 2017, doi: 10.1016/j.neuroimage.2016.02.079.

[3] D. R. Sarvamangala and R. V. Kulkarni, “Convolutional neural networks in medical image understanding: a survey,” Evolutionary Intelligence, vol. 15, no. 1, pp. 1–22, Jan. 2021, doi: 10.1007/s12065-020-00540-3.

[4] M. Hu, K. Sim, J. H. Zhou, X. Jiang and C. Guan, "Brain MRI-based 3D Convolutional Neural Networks for Classification of Schizophrenia and Controls," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 1742-1745, doi: 10.1109/EMBC44109.2020.9176610.

[5] D. Janeva et al., "A System for Differentiation of Schizophrenia and Bipolar Disorder based on rsfMRI," 2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP), Ohrid, North Macedonia, 2023, pp. 1-5, doi: 10.1109/IWSSIP58668.2023.10180306.

[6] R. A. McCutcheon, T. Reis Marques, and O. D. Howes, “Schizophrenia—An overview,” JAMA Psychiatry, vol. 77, no. 2, pp. 201–210, Feb. 2020, doi: 10.1001/jamapsychiatry.2019.3360.

[7] M. F. Green, “Cognitive impairment and functional outcome in schizophrenia and bipolar disorder,” The Journal of Clinical Psychiatry, vol. 67, no. 10, p. e12, Oct. 2006, doi: 10.4088/jcp.1006e12.

[8] M. B. L. De Freitas et al., “Resting-state fMRI is associated with trauma experiences, mood and psychosis in Afro-descendants with bipolar disorder and schizophrenia,” Psychiatry Research Neuroimaging, vol. 340, p. 111766, Dec. 2023, doi: 10.1016/j.pscychresns.2023.111766.

[9] R. K. P. Silva et al., “Cortical thickness alterations in afro-descendants with schizophrenia and bipolar disorder: An exploratory analysis,” Psychiatry Research Neuroimaging, vol. 355, p. 112094, Nov. 2025, doi: 10.1016/j.pscychresns.2025.112094.

[10] R. Diallo, C. Edalo, and O. O. Awe, “Machine Learning Evaluation of Imbalanced Health data: A comparative analysis of Balanced Accuracy, MCC, and F1 Score,” in STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, 2024, pp. 283–312. doi: 10.1007/978-3-031-72215-8_12.

[11] F. S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” Korean Journal of Anesthesiology, vol. 75, no. 1, pp. 25–36, Jan. 2022, doi: 10.4097/kja.21209.

[12] T. C. F. Polo and H. A. Miot, “Aplicações da curva ROC em estudos clínicos e experimentais,” Jornal Vascular Brasileiro, vol. 19, p. e20200186, Jan. 2020, doi: 10.1590/1677-5449.200186..

[13] F. C. Soczek and R. Orlovski, “Mineração de dados: conceitos e aplicação de algoritmos em uma base de dados na área da saúde,” Revista Científica Semana Acadêmica, no. 50, Mar. 2014. [Online]. Available:https://semanaacademica.org.br/system/files/artigos/mineracao_de_dados_conceitos_e_aplicacao_de_algoritmos_em_uma_base_de_dados_na_area_da_saude.pdf

[14] C. Daban et al., “Specificity of Cognitive Deficits in Bipolar Disorder versus Schizophrenia,” Psychotherapy and Psychosomatics, vol. 75, no. 2, pp. 72–84, Jan. 2006, doi: 10.1159/000090891.

[15] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, “FSL,” NeuroImage, vol. 62, no. 2, pp. 782–790, Aug. 2012, doi: 10.1016/j.neuroimage.2011.09.015.

[16] A. Giff et al., “Spatial normalization discrepancies between native and MNI152 brain template scans in gamma ventral capsulotomy patients,” Psychiatry Research Neuroimaging, vol. 329, p. 111595, Jan. 2023, doi: 10.1016/j.pscychresns.2023.111595.

Variance Stabilization of Unprocessed Digital Breast Tomosynthesis Projections Using a Residual Neural Network: A Preliminary Evaluation

Jeniffer M. Valencia Ruiz
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
jevalencia84@usp.br

Valentina Perafan Muñoz
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
valper18@usp.br

Dyego Alves Pereira
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
dyego.pereira@usp.br

Renann F. Brandão
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
renannbrandao@usp.br

Marcelo A. C. Vieira
São Carlos School of Engineering
University of São Paulo
São Carlos, Brazil
mvieira@sc.usp.br

***Abstract*—Raw digital breast tomosynthesis (DBT) projections exhibit noise whose variance depends on both signal intensity and detector position. Analytical stabilization using the Generalized Anscombe Transform (GAT) requires a spatial quantum-gain map and an electronic-noise parameter, which may be unavailable without access to equipment calibration data. This preliminary study investigates whether a residual convolutional neural network can approximate the analytical transformation directly from noisy projections, without receiving these parameters at inference. The network was trained on 700 OpenVCT-simulated patches from 70 phantoms and tested on 150 patches from 15 independent phantoms. A single spatially varying gain map and a fixed electronic-noise level were used throughout; therefore, evaluation addressed new anatomies under the same calibration configuration. Wavelet-based estimation yielded an output noise standard deviation of 0.897 ± 0.044, compared with the target value of 1.0. Monte Carlo evaluation using 20 noise realizations for each of 75 test patches yielded a mean stabilization error of 50.45%, with a standard deviation of 14.32% across the 15 phantom means and a 95% confidence interval of $[43.20\%, 57.70\%]$. Patch-level error was strongly correlated with local quantum gain ($r = 0.782$). The analytical GAT yielded an error of 0.095% under the same evaluation procedure, supporting the interpretation that the network's larger error reflects limitations of the learned transformation. The network partially approximated the GAT but did not achieve uniform variance stabilization under the evaluated conditions. Future work will train and evaluate the network across different gain maps and electronic-noise levels to assess generalization beyond the calibration configuration used here.**

***Index Terms*—digital breast tomosynthesis, variance stabilization, generalized Anscombe transform, residual neural network, heteroscedastic noise, virtual clinical trials.**

This work was supported by the São Paulo Research Foundation (FAPESP; Grants No. 2021/12673-6 and 2026/18409-2), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; Finance Code 001), and the Brazilian National Council for Scientific and Technological Development (CNPq; Grant No. 303251/2025-5).

I. Introduction

Digital breast tomosynthesis (DBT) complements conventional digital mammography by reducing tissue-overlap effects, reconstructing a 3D volume from multiple projections acquired over a limited angular arc [1], [4].

In raw ("for processing") DBT projections, noise is heteroscedastic: it arises predominantly from the quantum nature of X-ray radiation and follows an affine variance model, where each pixel's variance depends linearly on the underlying signal and on the detector's spatially varying quantum gain, plus a constant electronic-noise component [1], [2].

Variance-stabilizing transforms (VSTs), such as the Generalized Anscombe Transform (GAT), convert this signal-dependent noise into approximately signal-independent noise, facilitating denoising. However, accurate GAT application requires the detector's quantum-gain map and electronic-noise parameter, estimable when calibration images are available [1], [2] but often unknown for images from unfamiliar equipment. This motivates the present work: investigating whether a network can stabilize a raw DBT projection from the image alone, without the gain map or electronic-noise parameter at inference time.

Brandão et al. [3] addressed this with an iterative method that implicitly estimates these parameters from a single projection (errors of 1.16–7.69% across three systems), but requires per-image optimization at inference. Convolutional networks have also been used to learn stabilizing transforms directly: Zhang et al. [5] proposed VST-Net for Poisson denoising; Eckert et al. [6] used the Anscombe transform for physics-driven augmentation in mammography; and in DBT specifically, Vimieiro et al. [4] applied CycleGAN in a variance-stabilized domain (an unpaired approach with no pixel-wise fidelity guar-

antee), while Araújo et al. [7] and Gao et al. [8] denoised DBT projections and reconstructed slices, respectively, without an explicit variance-stabilization step and with strong dependence on dose/sample size. More recently, Herbreteau and Unser [11] proposed Noise2VST, a self-supervised, model-free VST, though untested on medical X-ray images and still requiring per-image optimization. Mota et al. [12] instead used a 2D total variation filter on reconstructed DBT slices, without a learned, data-driven transformation.

This work reports a *first stage* of a broader investigation into calibration-free variance stabilization for DBT: a residual neural network (ResNet) is trained and evaluated using a single, fixed quantum-gain map and electronic-noise level, simulated via OpenVCT [9] to keep calibration exactly known. Under this fixed configuration, we evaluate only whether the network generalizes to new, unseen anatomies; generalization across different gain maps, noise levels, or acquisition systems is not claimed and is left for future work (Section V).

II. MATERIALS AND METHODS

Fig. 1 summarizes the main steps of the methodology described in this section, from the generation of synthetic noisy projections to the evaluation metrics used to assess variance stabilization performance.

A. Noise Model and Synthetic Data Generation

Images were obtained from the 100-Phantom DBT Hologic dataset, generated by the OpenVCT platform for mammography and DBT simulation [9]. For each phantom, the central projection of the acquisition arc, corresponding to index 7, was used, as it presents the least beam obliquity and most adequately represents the average signal-to-noise ratio of the dataset.

The adopted noise model follows the affine variance model described by Borges et al. [1], [2], in which the observed pixel value $z(i,j)$, conditioned on the underlying noise-free signal $y(i,j)$, presents variance:

$$\mathrm{Var}\{z(i,j) \mid y(i,j)\} = \alpha(i,j)\, y(i,j) + \sigma_e^2 \tag{1}$$

where $\alpha(i,j)$ represents the spatially varying quantum gain of the detector, and σ_e corresponds to the electronic noise standard deviation, assumed constant across the detector. In this work, a single spatially-varying gain map was used throughout training and evaluation, with values ranging between 0.130 and 0.365 over the full detector and between 0.131 and 0.182 within the breast region (Section III-D); the electronic-noise level used was $\sigma_e = 1.3743$ ADU. The noise parameters were estimated following the calibration protocol described by Borges et al. [10].

Both $z(i,j)$ and $y(i,j)$ are expressed in raw detector counts (ADU, analog-to-digital units), following the vendor's linearized pixel values prior to any offset correction. Consequently, since $\mathrm{Var}\{z(i,j) \mid y(i,j)\}$ has units of ADU^2 and $y(i,j)$ is expressed in ADU, the quantum gain $\alpha(i,j)$ is expressed in ADU, and the electronic noise standard deviation σ_e is expressed in ADU as well.

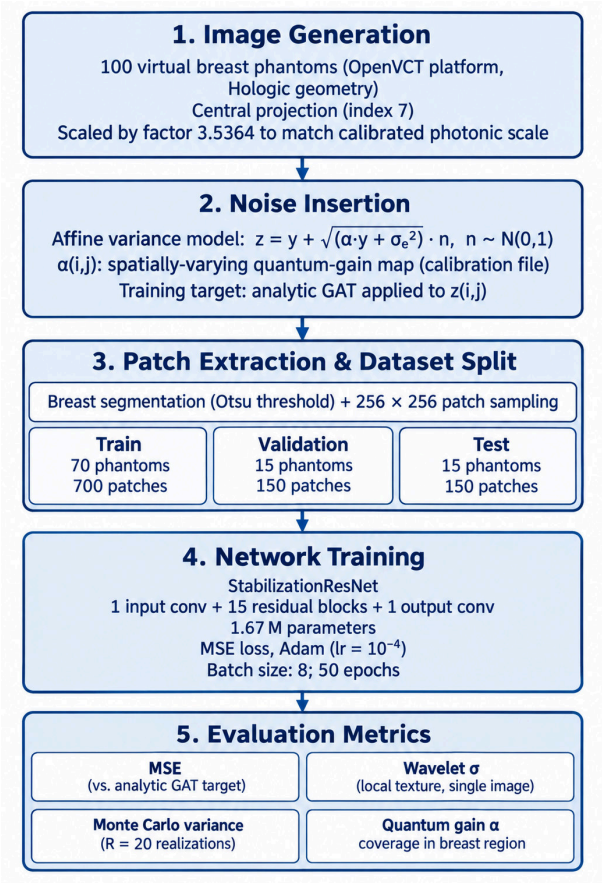


Fig. 1. Overview of the methodology: synthetic image generation, heteroscedastic noise insertion, patch extraction with phantom-level dataset split, network training, and evaluation metrics.

Since the signal level of the synthetic image derived from VCT does not correspond to the signal level of a clinical image, a scaling factor procedure was performed to match the signal levels. Thus, a multiplicative scale factor of 3.5364 was computed from a sample of 20 phantoms and applied to all noise-free projections (y) before noise simulation. Heteroscedastic noise was then generated as:

$$\begin{aligned} z(i,j) &= y(i,j) + \sigma(i,j)\,\eta(i,j), \\ \sigma^2(i,j) &= \alpha(i,j) y(i,j) + \sigma_e^2, \\ \eta(i,j) &\sim \mathcal{N}(0,1). \end{aligned} \tag{2}$$

where $\sigma(i,j)$ denotes the signal- and spatial-dependent noise standard deviation, and $\eta(i,j)$ represents a zero-mean, unit-variance Gaussian noise field.

It is important to clarify the scope of this simulation: the gain map $\alpha(i,j)$ used here already varies *spatially* across the detector (Eq. 1); what was *not* varied was the calibration configuration itself between simulations — the same gain map and the same electronic-noise level σ_e were used for every phantom, in training, validation, and test. Consequently, evaluating the network on the 15 test phantoms (Section III) assesses its behavior on new anatomies *within this fixed*

calibration configuration; it does not assess its ability to operate with a gain map or noise level unseen during training. Extending the training data to span multiple gain maps and noise levels is discussed in Section V.

B. Breast Segmentation and Patch Extraction

The breast region was segmented using Otsu thresholding with threshold inversion to separate the breast from the background. From the binary mask, 10 patches of 256×256 pixels were randomly extracted per phantom, fully contained within the segmented region, allowing up to 1000 attempts per patch. The 100 phantoms were randomly divided — with a fixed seed — into 70 for training, 15 for validation, and 15 for testing at the phantom level, avoiding anatomy sharing between sets. All 100 phantoms successfully contributed patches, yielding 700 training, 150 validation, and 150 test patches, for a total of 1000 patches. Fig. 2 illustrates breast segmentation and patch extraction for a representative phantom.

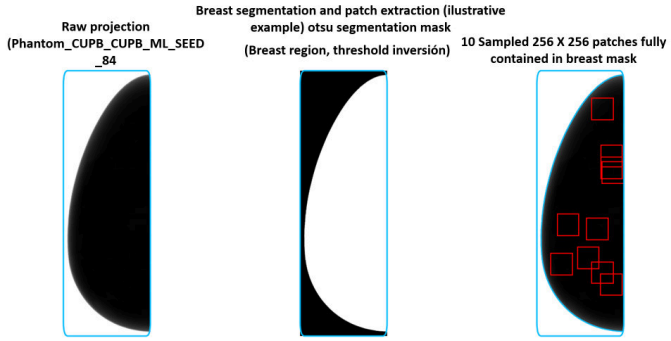


Fig. 2. Illustration of the breast segmentation and patch extraction process for a representative phantom. From left to right: raw DBT projection, Otsu segmentation mask (breast region, threshold inversion), and the 10 sampled 256×256 patches fully contained within the segmented breast mask.

C. Training Target: Generalized Anscombe Transform

The training target was defined as the GAT applied directly to the simulated noisy projection $z(i,j)$:

$$\hat{z}(i,j) = \frac{2}{\alpha(i,j)} \sqrt{\max\left(0,\ \alpha(i,j)\, z(i,j) + \frac{3}{8}\alpha^2(i,j) + \sigma_e^2\right)} \tag{3}$$

This expression is algebraically equivalent to the classical GAT formulation described by Borges et al. [1], assuming $\lambda(i,j) \equiv \alpha(i,j)$ and $\tau = 0$, which is valid for the virtual phantoms used. Thus, the training target corresponds to the analytical VST applied directly to the noisy projection, rather than to a noise-free image.

1) Two-step interpretation of the analytical GAT: The analytical stabilization procedure summarized in Eq. (3) can be understood as two successive operations, and it is useful to make this explicit, since the network is trained to approximate their joint effect. First, the noisy projection is normalized by the spatial gain map. With a null offset ($\tau = 0$, Eq. (1)), define

$$u(i,j) = \frac{z(i,j)}{\alpha(i,j)}. \tag{4}$$

After this normalization, the conditional variance becomes

$$\mathrm{Var}\{u(i,j) \mid y(i,j)\} = \frac{y(i,j)}{\alpha(i,j)} + \frac{\sigma_e^2}{\alpha^2(i,j)}. \tag{5}$$

Dividing by $\alpha(i,j)$ makes the quantum-gain term unitary with respect to the normalized signal, but it does *not* eliminate the spatial dependence of the noise: the electronic-noise term is rescaled as well, with its standard deviation becoming $\sigma_e/\alpha(i,j)$ and its variance $\sigma_e^2/\alpha^2(i,j)$. Second, the GAT is applied to $u(i,j)$ with unit gain, using the rescaled electronic-noise variance from Eq. (5):

$$\hat{z}(i,j) = 2\sqrt{\max\left(0,\ u(i,j) + \frac{3}{8} + \frac{\sigma_e^2}{\alpha^2(i,j)}\right)}. \tag{6}$$

Substituting $u(i,j) = z(i,j)/\alpha(i,j)$ from Eq. (4) into Eq. (6) and factoring $1/\alpha^2(i,j)$ out of the square root recovers exactly Eq. (3). This confirms that the two-step procedure (spatial-gain normalization followed by a unit-gain GAT with rescaled electronic noise) is algebraically identical to the single-step formula already used as the training target; the network's task is therefore to learn this joint transformation — reducing the dependence of the variance on both signal and detector position — directly from the noisy image, without being given $\alpha(i,j)$ or σ_e explicitly at inference time. We verified that the noise-simulation code (Eq. (2)) and the target-generation code (Eq. (3)) apply this electronic-noise rescaling consistently.

D. Network Architecture

The proposed architecture, named *StabilizationResNet*, is a fully convolutional residual network composed of three main stages: an input layer consisting of a 3×3 convolution mapping one input channel to 64 feature channels, followed by a ReLU activation; 15 residual blocks, each composed of three 3×3 convolutional layers with 64 channels, interleaved with Batch Normalization and ReLU activations, and connected through residual skip connections; and an output layer consisting of a 1×1 convolution mapping 64 feature channels to one output channel, with linear activation. The network contains a total of 1,668,225 trainable parameters. Table I summarizes the main hyperparameters used during training.

E. Normalization and Training

Input images (z) and targets ($\hat{z}$) were normalized independently by z-score standardization using training-set statistics, with clipping at ±5 standard deviations, to account for their different dynamic ranges. Averaged over the 150 test patches, this clipping affected 0.68% of input pixels and 0.73% of target pixels, confirming that, on average, the normalization range preserves the vast majority of the signal distribution. However, this average conceals substantial variability across individual patches: as discussed in Section III-B, the patches with the highest MSE show clipping fractions between 7% and 19%, an order of magnitude above this average. Training used MSE loss, Adam optimizer (lr $= 1 \times 10^{-4}$), batch size 8, and 50 epochs on Google Colab GPU. The model with lowest validation loss was selected.

TABLE I
NETWORK ARCHITECTURE AND TRAINING HYPERPARAMETERS.

Parameter	Value
Input	$256 \times 256 \times 1$ normalized raw patch
Initial layer	3×3 convolution, $1 \rightarrow 64$ channels, ReLU
Residual blocks	15 blocks, 64 channels each
Final layer	1×1 convolution, $64 \rightarrow 1$ channel, linear activation
Trainable parameters	1,668,225
Loss function	MSE
Optimizer	Adam, $lr = 1 \times 10^{-4}$
Batch size	8
Epochs	50
Input normalization, z	$\mu = 586.14$, $\sigma = 84.05$
Target normalization, $\hat{z}$	$\mu = 128.61$, $\sigma = 8.19$
Data split	700 / 150 / 150 patches from 70 / 15 / 15 phantoms

F. Evaluation Metrics

Performance was evaluated through: (i) MSE between the denormalized network output and the analytic GAT target; (ii) local residual noise standard deviation, $\hat{\sigma}$, estimated using a wavelet-based estimator (median absolute deviation of the finest-scale wavelet detail coefficients, following Donoho and Johnstone [13]), with a target value of 1.0; and (iii) Monte Carlo validation with $R = 20$ independent noise realizations per patch, computing the empirical pointwise variance (sample variance, unbiased estimator, ddof $= 1$) across realizations — the physically correct criterion for variance stabilization. For each patch, the stabilization error is defined as

$$E_{\%} = 100 \times \left| \overline{\widehat{\mathrm{Var}}_R[\cdot]} - 1 \right|, \tag{7}$$

where $\widehat{\mathrm{Var}}_R[\cdot]$ is the unbiased sample variance computed pixelwise across the $R = 20$ realizations, and $\overline{(\cdot)}$ denotes the spatial average of this pointwise variance over all pixels of the patch.

Note that the wavelet-based $\hat{\sigma}$ measures the local texture of a single output image, while the Monte Carlo variance measures reproducibility across independent noise instances, which are fundamentally different assessments [2]. As a control, the same Monte Carlo procedure (identical patches and identical noise realizations) was also applied to the analytic GAT itself, providing a reference for the error attributable to the evaluation procedure as opposed to the network. Additionally, quantum gain, α, coverage within breast patches was analyzed.

III. RESULTS

A. Training

The network was trained for 50 epochs using the training and validation sets. The final model was selected based on the lowest validation loss, achieved at epoch 47, with a validation loss of 0.079018, as shown in Fig. 3. The training curve shows a consistent decrease in both training and validation losses, with validation oscillations between epochs 25 and 40, which may be attributed to the small batch size and limited dataset size.

B. Quantitative Evaluation on Test Set

Table II summarizes the metrics obtained from the 150 test patches. The MSE distribution was strongly asymmetric, with a mean of 9.92, a median of 5.49, and a standard deviation of 21.00: a small number of patches with markedly elevated error increased the mean substantially above the median.

TABLE II
METRICS COMPUTED OVER THE 150 TEST PATCHES.

Metric	Mean	Median	Std. dev.	Min.	Max.
MSE relative to analytic target	9.92	5.49	21.00	0.55	194.43
Estimated σ in network output	0.897	–	0.044	0.670	0.973
Estimated σ in analytic target	0.999	–	0.008	0.973	1.020

Regarding the estimated σ, the network output presented a mean value of 0.897 ± 0.044, while the analytic GAT target presented $\sigma \approx 0.999 \pm 0.008$, confirming that the analytic GAT correctly stabilizes the variance close to unit value under this pointwise, single-image metric. The network systematically underestimates this parameter, although only 1 out of 150 patches, corresponding to 0.7%, presented a deviation $|\sigma - 1| > 0.3$. As discussed below, this pointwise metric substantially understates the network's stabilization error relative to the reproducibility-based Monte Carlo assessment (Section III-C).

Visual inspection of representative test patches, shown in Fig. 4, reveals over-smoothing in the network output relative to the analytic target, suggesting partial loss of fine noise texture in some patches. Table III presents the five test patches with the highest MSE values, consolidating patch identification, error, local gain, spatial location, and normalization-clipping fraction in a single table.

To assess whether patches with elevated MSE are concentrated near the detector edges, where the quantum gain α is higher, as could be expected from the acquisition geometry, a spatial audit was performed on the full test set (150 patches), recording, for every patch, the coordinates, local mean α, distance to the detector edge, and internal signal statistics. Across all 150 test patches, MSE showed essentially no linear association with local α ($r = -0.019$) or with distance to the detector edge ($r = 0.119$).

Contrary to an edge/high-α hypothesis, none of the five highest-MSE patches are close to the detector edge, and their local gain (mean $\alpha \approx 0.136$ ADU) is close to the mean α of the 20 lowest-MSE patches ($\alpha \approx 0.137$ ADU) — local gain and edge proximity do not distinguish the worst patches from the rest. Instead, the strongest distinguishing factor is the internal signal variability of the patch, computed in the analytic-GAT (dimensionless) domain: the five highest-MSE patches had a mean internal standard deviation of 21.03 and a mean dynamic range of 119.98, versus 3.26 and 20.28, respectively, for the 20 lowest-MSE patches — roughly a sixfold difference.

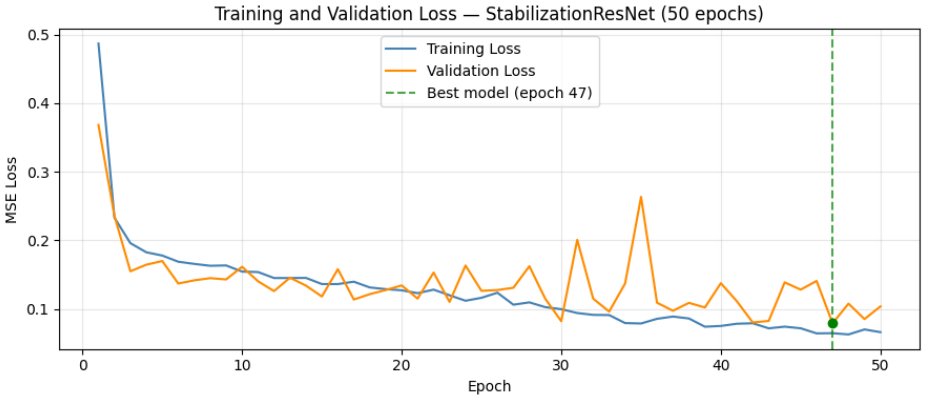


Fig. 3. Training and validation loss curves for the proposed StabilizationResNet model. The lowest validation loss was achieved at epoch 47.

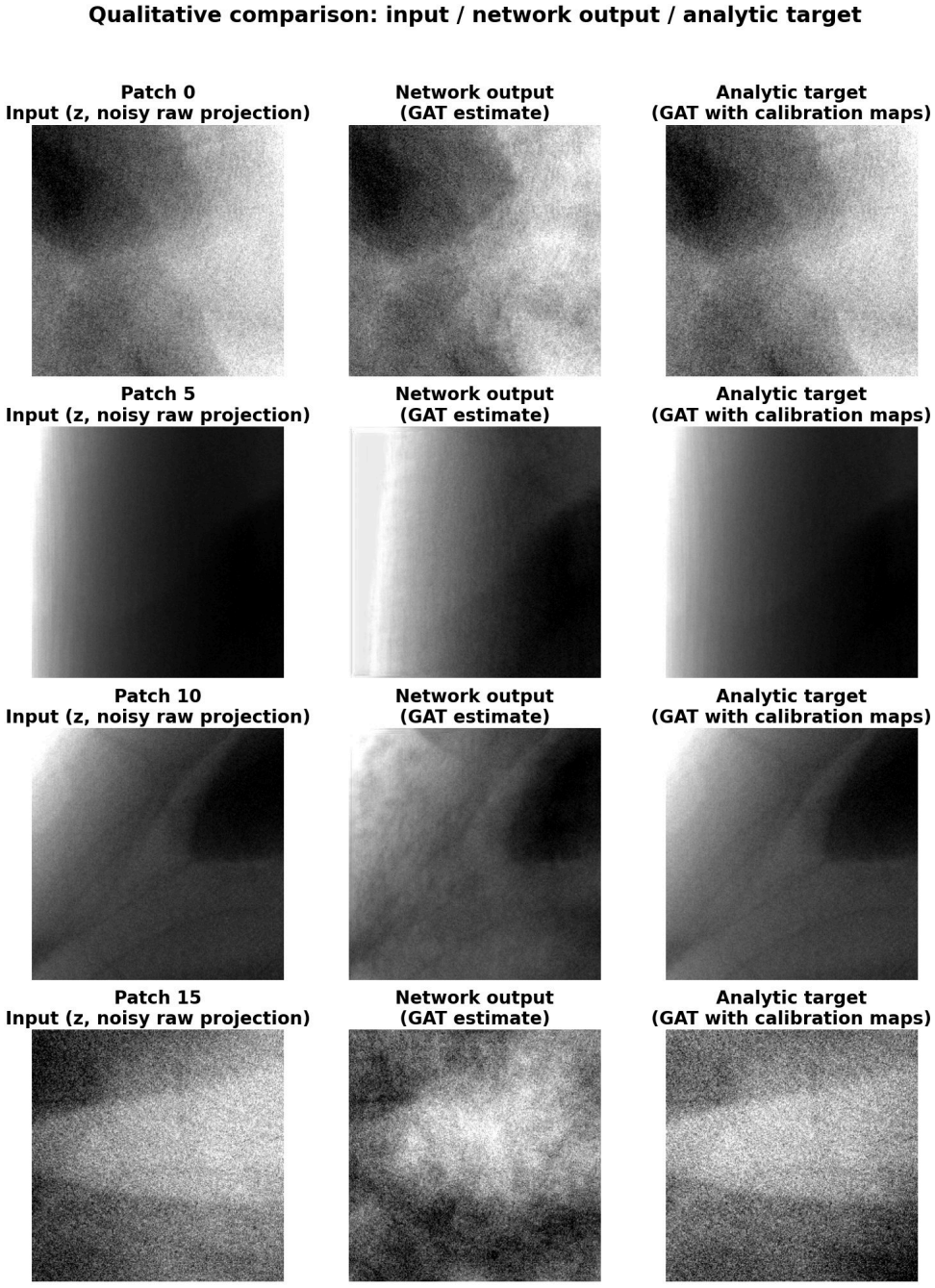


Fig. 4. Qualitative comparison: noisy input (left), network output — GAT estimate (center), analytic target — GAT with calibration maps (right), for patches 0, 5, 10, and 15. Patch 0 (top) represents a typical case; Patch 5 illustrates a case of pronounced over-smoothing.

TABLE III
FIVE HIGHEST-MSE TEST PATCHES: IDENTIFICATION, ERROR, LOCAL GAIN, POSITION, DISTANCE TO DETECTOR EDGE, AND NORMALIZATION-CLIPPING FRACTION (INPUT / TARGET).

Patch	Phantom	MSE	Mean α (ADU)	Position (i, j)	Dist. to edge	Clip. %
18	CUPB_ML_seed_28	194.43	0.135	(1490, 2522)	550 px	15.96 / 19.47
137	CUPB_ML_seed_92	120.95	0.133	(1678, 2470)	602 px	10.09 / 13.81
5	CUPB_ML_seed_56	106.54	0.132	(2176, 2377)	695 px	11.02 / 14.45
122	CUPB_ML_seed_21	53.55	0.146	(1057, 2731)	341 px	8.38 / 7.79
133	CUPB_ML_seed_92	43.63	0.135	(2520, 2386)	686 px	7.28 / 9.82

The worst patches also show elevated clipping fractions (Table III, 7–19%, against a $0.68\%/0.73\%$ average over all 150 test patches, Section II-E); however, wide dynamic range alone does not directly explain this effect, since the fraction of clipped pixels also depends on the mean signal level of the patch relative to the $\pm 5\sigma$ normalization threshold, in both the input and the target. Reported MSE values were computed against the original (unclipped) analytic GAT target (Y_{test}), not against a clipped version of it; the clipping described here therefore affects only the normalized network *input* and *training target*, not the reference against which test-time MSE is evaluated.

Visual inspection of the highest-MSE patches revealed that they correspond to regions with sharp anatomical transitions (near the skin line), where the network output shows a locally saturated boundary artifact not present in the analytic GAT target. One plausible explanation is that the network's Batch Normalization layers, whose batch-level statistics may become poorly calibrated for patches whose local dynamic range substantially exceeds the typical training-batch range, contribute to this artifact; however, this explanation has not been directly tested in this work and should be regarded as a hypothesis, not a demonstrated cause. The measured association between clipping fraction and MSE (Table III), by contrast, is a direct empirical finding. This observation motivates future work on normalization strategies less sensitive to per-sample dynamic range (e.g., instance or group normalization) to mitigate this specific failure mode.

Inference time was benchmarked on a Tesla T4 GPU and, for reference, on CPU. A single 256×256 patch required 65.75 ± 0.44 ms on GPU (mean $\pm$ std over 100 repetitions, after warm-up) and 2719 ± 365 ms on CPU (20 repetitions). Stabilizing a full DBT projection (4096×3328 pixels, tiled into 208 non-overlapping patches) required approximately 13.7 s in total on GPU. The proposed network requires a single forward pass per patch, with no per-image optimization at inference time, unlike the iterative method of Brandão et al. [3], which repeats a per-block noise-variance estimation step until convergence for each new acquisition; a direct, head-to-head runtime comparison against that method was not performed in this work, so no quantitative speed advantage is claimed.

C. Monte Carlo Variance Stabilization Evaluation

Table IV presents the Monte Carlo evaluation results for the network and, as a control, for the analytic GAT evaluated under the identical procedure (same 75 patches, same $R = 20$ noise realizations per patch). For both the network and the analytic-GAT control, the reported mean, standard deviation, and 95% CI are computed from the 15 per-phantom mean errors (one mean per test phantom, averaging its 5 patches), not from the 75 individual patches, since five patches originate from each phantom and therefore share anatomy and are not statistically independent; the patch-level standard deviation was 28.93% for the network. To account for the grouping of patches within phantoms, the statistics reported in Table IV were calculated from the 15 per-phantom mean errors. Note also that Eq. (7) evaluates the spatial *average* of the pointwise variance over each patch; it does not verify whether that variance is spatially uniform within the patch.

TABLE IV
MONTE CARLO EVALUATION RESULTS ($R = 20$, 75 PATCHES, 15 TEST PHANTOMS). CONFIDENCE INTERVALS ARE CLUSTER-ROBUST ACROSS PHANTOMS ($n = 15$).

Metric	Network	Analytic GAT (control)
Mean stabilization error	$50.45\% \pm 14.32\%$	$0.095\% \pm 0.023\%$
95% CI (cluster-robust)	$[43.20\%, 57.70\%]$	$[0.083\%, 0.107\%]$
Mean pointwise variance	1.503 (target: 1.000)	1.000 (target: 1.000)
Correlation α vs. error (%), r	0.782	0.028
Correlation y vs. error (%), r	0.325	–

The analytic GAT control confirms that the Monte Carlo procedure itself is sound: applied to the known analytic transformation, it recovers a pointwise variance of 1.000 with negligible error (0.095%) and no dependence on α ($r = 0.028$). The network's error is therefore not an artifact of the evaluation methodology, but reflects a genuine limitation of the learned transformation: a mean stabilization error of $50.45\% \pm 14.32\%$ (95% CI $[43.20\%, 57.70\%]$) and a mean pointwise variance of 1.503, substantially above the ideal value of 1.000. This is a markedly larger error than suggested by the pointwise wavelet-based $\hat{\sigma}$ metric of Section III-B, underscoring the importance of the Monte Carlo criterion. Figure 5 provides a visual summary: the stabilization error distribution shows substantial variability among patches, and the strong association between error and α ($r = 0.782$) indicates that patches with higher local detector gain tend to present larger stabilization errors.

D. Quantum Gain Coverage in the Breast Region

Table V and Fig. 6 present the α coverage analysis. The full detector operating range spans $[0.130, 0.365]$. However, within the segmented breast region, α varies only between 0.131 and 0.182, covering approximately 21.5% of the total system range. Training patches of 256×256 pixels, restricted to the breast region, present an even narrower range of 0.132–0.168, with a mean value of 0.145.

It is important to note that this narrow coverage arises from the breast's position relative to a single, fixed gain map used throughout this study (Section II-A), not from a variation in the gain map itself.

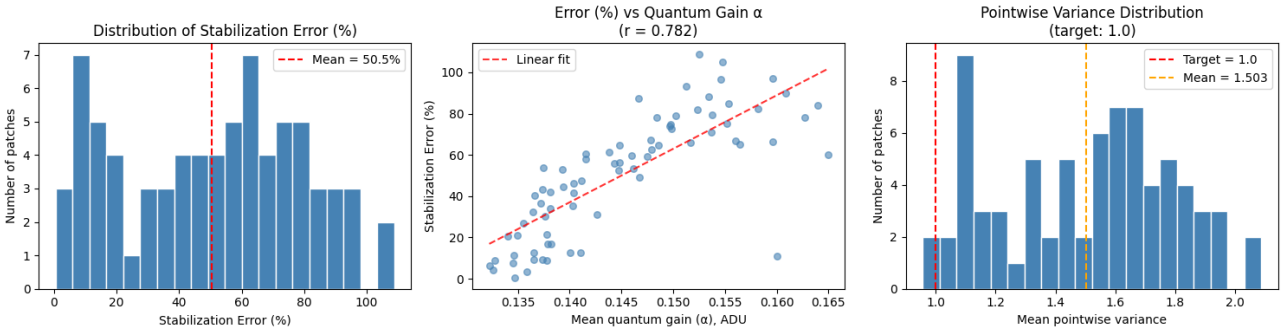


Fig. 5. Monte Carlo results for the network (75 patches, $R = 20$ realizations): left, distribution of stabilization error (%); center, stabilization error versus mean quantum gain α ($r = 0.782$, linear fit in red); right, pointwise variance distribution, with mean equal to 1.503 and target equal to 1.0.

Whether this restricted coverage is a primary cause of the network's stabilization error (Section III-C), and whether training with multiple gain maps spanning a wider α range would improve generalization, remain hypotheses to be tested in future work (Section V); the present results do not establish a causal relationship, only the co-occurrence of narrow gain coverage and a strong error–α correlation within this single-configuration dataset.

TABLE V
QUANTUM GAIN (α) COVERAGE BY REGION.

Region	α **range (ADU)**	**Mean** α **(ADU)**
Training patches (256 × 256)	0.1318–0.1683	0.1445
Full breast region	0.1309–0.1816	0.1469
Full detector	0.1300–0.3653	–

IV. DISCUSSION

Training a compact residual network (1.67 million parameters) to *partially* approximate the GAT without explicit calibration at inference is feasible, but this preliminary, single-gain-map study sits between two established strategies: the precise, calibration-dependent framework of Borges et al. [1], [2], and the calibration-free, per-image iterative method of Brandão et al. [3], who reported errors of 1.16–7.69% in their own published experiments (not evaluated here). Against those published figures, the proposed network showed a much larger Monte Carlo error (50.45%, Table IV); this is not a head-to-head comparison, but it shows that, at this stage, the approach trades accuracy for single-pass inference rather than matching [3]'s accuracy. It also differs from VST-Net [5] (globally parameterized noise, no spatial gain dependence) and Noise2VST [11] (self-supervised but still per-image optimization). Inference took 65.75 ms per patch (∼13.7 s per full projection) on a Tesla T4 GPU, avoiding per-block iterative convergence; no direct runtime comparison against [3] was performed, so this is a qualitative architectural difference, not a measured speed advantage.

A key finding is the gap between pointwise and Monte Carlo metrics: the wavelet-based $\hat{\sigma} \approx 0.897$ (Table II) reflects local texture and can be confused with true variance reduction [2], whereas the Monte Carlo pointwise variance (1.503) measures reproducibility across realizations — the physically correct criterion — revealing a much larger gap from ideal. The analytic-GAT control (error 0.095%, variance 1.000, Section III-C) confirms this gap is a genuine network limitation, not an evaluation artifact; both metrics are complementary.

The strong error–α correlation ($r = 0.782$), together with the narrow α coverage of the breast region under this single gain map (21.5% of the detector range), is consistent with a sampling-limitation hypothesis, but only as a co-occurrence within one configuration, not a controlled result; isolating this from other causes (architecture, normalization, training-set size) requires training across multiple gain maps and noise levels (Section V). Geometric augmentation (rotations/flips) was also tried but did not improve performance and was not used; the cause was not confirmed.

This study has several limitations. First, all experiments used a single central angular projection per phantom; other angles were not evaluated. Second, and most importantly, a single quantum-gain map and electronic-noise level ($\sigma_e = 1.3743$ ADU) were used throughout, so results characterize generalization to new anatomies *under this fixed calibration*, not across gain maps, noise levels, or systems. Third, the error–α correlation (Sections III-C and IV) is an association within this single-configuration dataset, not experimentally isolated from other factors. Fourth, no real clinical or physical-phantom validation was performed; all results are OpenVCT-simulated.

V. CONCLUSION

This work presented a preliminary, first-stage evaluation of a ResNet-based approach to estimate the Generalized Anscombe Transform directly from simulated noisy raw DBT projections, without requiring the quantum-gain map or the electronic-noise parameter as input during inference. Under a single, fixed calibration configuration (one gain map, one noise level), the network partially approximated the analytical transformation for new, unseen anatomies. Pointwise evaluation showed a mean wavelet-based $\sigma = 0.897$, with only 0.7% of patches presenting a large pointwise deviation.

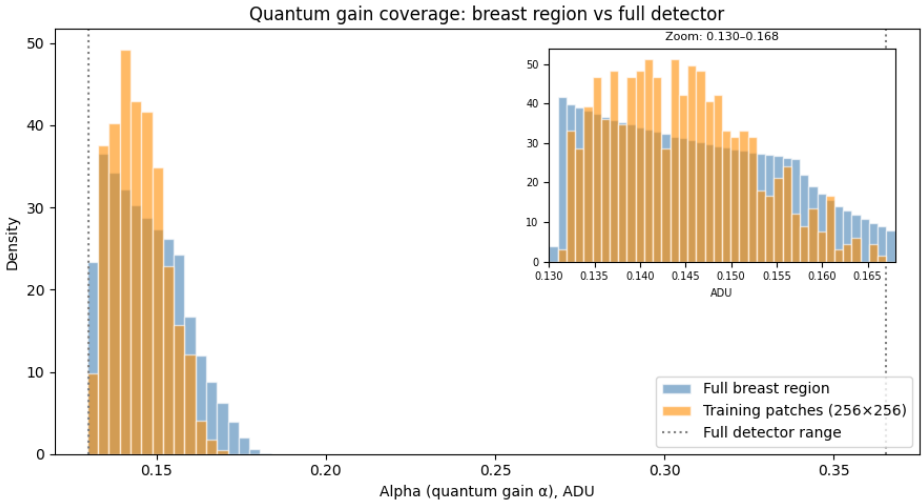


Fig. 6. Quantum gain (α, in ADU) density distribution for the full breast region and training patches of 256×256 pixels. Dotted lines indicate the full detector operating range $[0.130, 0.365]$ ADU. The breast region covers only 21.5% of the total detector range, for the single gain map used in this study.

However, Monte Carlo validation, using an analytic-GAT control to confirm the soundness of the procedure, revealed a substantially larger and non-uniform stabilization error (mean 50.45%, 95% CI $[43.20\%, 57.70\%]$ across 15 independent test phantoms), strongly associated with local quantum gain ($r = 0.782$). The network has therefore not yet achieved uniform variance stabilization under the evaluated conditions, and generalization to gain maps, noise levels, or acquisition systems other than the one used here was not demonstrated.

These findings reinforce the importance of combining pointwise metrics with Monte Carlo validation, and of using an analytic control to distinguish genuine model limitations from artifacts of the evaluation procedure, when assessing learned variance-stabilizing transformations. Future work will use the OpenVCT platform to generate training data spanning multiple spatially-varying quantum-gain maps and electronic-noise levels, and will evaluate the network on combinations of these parameters not seen during training, while preserving the phantom-level separation between training and test anatomies already used in this study. This will allow the sampling hypothesis raised in Section IV to be tested directly. Subsequently, transferability to real clinical or physical-phantom acquisitions will need to be assessed. Additional future directions include investigating alternative network architectures or normalization strategies (e.g., instance or group normalization, motivated by the clipping-related artifact discussed in Section III-B) to improve variance stabilization while preserving the calibration-free nature of the proposed approach.

REFERENCES

[1] L. R. Borges, L. Azzari, P. R. Bakic, A. D. A. Maidment, M. A. C. Vieira, and A. Foi, "Restoration of low-dose digital breast tomosynthesis," *Meas. Sci. Technol.*, vol. 29, no. 6, p. 064003, 2018.

[2] L. R. Borges, P. R. Bakic, A. Foi, A. D. A. Maidment, and M. A. C. Vieira, "Pipeline for effective denoising of digital mammography and digital breast tomosynthesis," in *Proc. SPIE Medical Imaging 2017*, vol. 10132, p. 1013206, 2017.

[3] R. F. Brandão, L. R. Borges, B. Barufaldi, T. L. Vent, R. F. Caron, B. B. Oliveira, A. D. A. Maidment, and M. A. C. Vieira, "Iterative method to achieve noise variance stabilization in single raw digital breast tomosynthesis," in *Proc. SPIE Medical Imaging 2021*, vol. 11595, p. 115954R, 2021.

[4] R. B. Vimieiro, L. R. Borges, R. F. Caron, B. Barufaldi, A. D. A. Maidment, G. Wang, and M. A. C. Vieira, "Suppressing noise correlation in digital breast tomosynthesis using convolutional neural network and virtual clinical trials," in *Proc. SPIE IWBI*, vol. 12286, p. 122861B, 2022.

[5] M. Zhang, F. Zhang, Q. Liu, and S. Wang, "VST-Net: Variance-stabilizing transformation inspired network for Poisson denoising," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 12–22, 2019.

[6] D. Eckert, S. Vesal, L. Ritschl, S. Kappler, and A. Maier, "Deep learning-based denoising of mammographic images using physics-driven data augmentation," in *Bildverarbeitung für die Medizin 2020*, Informatik aktuell, Springer Vieweg, Wiesbaden, 2020, pp. 94–100, doi: 10.1007/978-3-658-29267-6_21.

[7] D. M. N. de Araújo, D. H. P. Salvadeo, and D. D. de Paula, "Denoising digital breast tomosynthesis projections using convolutional neural networks," in *Proc. SPIE Medical Imaging 2021*, vol. 11596, p. 115961L, 2021.

[8] M. Gao, J. A. Fessler, and H.-P. Chan, "Deep convolutional neural network with adversarial training for denoising digital breast tomosynthesis images," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1805–1816, Jul. 2021, doi: 10.1109/TMI.2021.3066896.

[9] B. Barufaldi, P. R. Bakic, D. Higginbotham, and A. D. A. Maidment, "OpenVCT: a GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis," in *Proc. SPIE Medical Imaging 2018*, vol. 10573, p. 1057358, 2018, doi: 10.1117/12.2294935.

[10] L. R. Borges, B. Barufaldi, R. F. Caron, P. R. Bakic, A. Foi, A. D. A. Maidment, and M. A. C. Vieira, "Noise models for virtual clinical trials of digital breast tomosynthesis," *Med. Phys.*, vol. 46, no. 6, pp. 2683–2689, 2019.

[11] S. Herbreteau and M. Unser, "Self-calibrated variance-stabilizing transformations for real-world image denoising," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2025, pp. 10496–10506, doi: 10.1109/ICCV51701.2025.00977.

[12] A. M. Mota, J. Mendes, and N. Matela, "Digital breast tomosynthesis: Towards dose reduction through image quality improvement," *J. Imaging*, vol. 9, no. 6, p. 119, 2023, doi: 10.3390/jimaging9060119.

[13] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

Binary Classification of Venous Thrombosis in Ultrasound Images Using Convolutional Neural Networks and Explainable AI Methods

1st Miqueias Nilton de Araujo
Computer Engineering
Federal University of Rural Semi-Arid
Rio Grande do Norte, Brazil
miqueias.araujo@alunos.ufersa.edu.br

2nd Rubens A. de S. Ferreira
Computational Intelligence Laboratory
Federal University of Rural Semi-Arid
Rio Grande do Norte, Brazil
rubens.ferreira@alunos.ufersa.edu.br

3rd Rosana C. B. Rego
Department of Engineering and Technology
Federal University of Rural Semi-Arid
Rio Grande do Norte, Brazil
rosana.rego@ufersa.edu.br

***Abstract*—The early detection of venous thrombosis is important for preventing complications and improving patient outcomes. This study implements a Convolutional Neural Network (CNN) for binary classification of venous thrombosis in ultrasound images using a dataset of 1000 images. The model achieved 77% accuracy, with a recall of 0.65, specificity of 0.89, and negative predictive value of 0.718. At the selected operating point, 35 of 100 positive cases were not detected, while 35 of 124 negative predictions were false. These results indicate an exploratory-stage model and should not be interpreted as evidence of clinical readiness. To improve interpretability, Grad-CAM and LIME were applied to identify image regions contributing to the predictions. LIME produced more localized explanations than Grad-CAM, highlighting its potential for interpreting CNN-based classifications. Further validation using larger, independent, patient-level datasets is required before clinical application.**

***Index Terms*—Deep Learning, Explainable AI, Ultrasound Imaging, Venous Thrombosis.**

I. Introduction

Venous thrombosis is a medical condition characterized by the formation of blood clots within the veins, which can lead to serious complications such as pulmonary embolism and post-thrombotic syndrome [1], [2]. Early detection and accurate diagnosis of venous thrombosis are crucial for effective treatment and prevention of adverse outcomes. Ultrasound imaging is a widely used non-invasive technique for diagnosing venous thrombosis, providing real-time visualization of the veins and blood flow [3], [4].

The interpretation of ultrasound images for the diagnosis of venous thrombosis can be challenging due to the variability in image quality, anatomical structures, and the subtlety of thrombotic features. Traditional methods for analyzing ultrasound images rely heavily on the expertise of clinicians, which can lead to inter-observer variability and potential misdiagnosis [5], [6].

In recent years, Convolutional Neural Networks (CNNs) have shown great promise in medical image analysis, including the classification of ultrasound images for various conditions [7]–[10]. However, the black-box nature of CNNs can limit their adoption in clinical settings, as clinicians may be hesitant to trust predictions without understanding the underlying reasoning. To address this issue, explainable AI (Artificial Intelligence) methods have been developed to provide insights into the decision-making process of machine learning models. Some popular explainable AI methods for deep learning models include Grad-CAM (Gradient-weighted Class Activation Mapping) [11] and LIME (Local Interpretable Model-agnostic Explanations) [12], which can help visualize the regions of the input image that contribute most to the model's predictions.

There have been several studies that have explored the use of machine learning and deep learning techniques for the diagnosis of venous thrombosis [13]–[16]. However, there is still a need for further research to improve the accuracy and interpretability of these models, particularly in the context of ultrasound imaging. For instance, [13] proposed a machine learning-based system for the early diagnosis of Deep Vein Thrombosis (DVT). The study evaluated six classifiers, such as Decision Trees (DT), Extra Trees (ET), K-Nearest Neighbors (KNN), Random Forests (RF), Support Vector Machines (SVM), and Multi-Layer Perceptron Neural Networks (MLP-NN), using clinical features derived from the Wells criteria, age, and gender. Furthermore, the authors investigated the deployment of the models on a Raspberry Pi 4 edge device to develop a portable and low-cost diagnostic support system, with KNN achieving the highest diagnostic performance.

Differently, in the work [14], the authors employed the RetinaNet object detection architecture with ResNet50 and ResNet152 backbones for the detection of iliofemoral deep vein thrombosis (DVT) in lower-extremity computed tomography angiography (LECTA) images. Experimental results demonstrated that the proposed three-slice approach improved detection performance, achieving a sensitivity of 0.843 and a mean average precision (mAP) of 0.807 with the ResNet50 backbone, highlighting the effectiveness of incorporating clinically relevant contextual information into deep learning-based DVT detection.

In the manuscript [15], the authors proposed a novel Deep R-Belief Network for the automated diagnosis of DVT from

Duplex Ultrasound (DUS) images. Their methodology consists of a multi-stage pipeline that includes image denoising using a Gaussian filter, contrast enhancement through Contrast Limited Adaptive Histogram Equalization (CLAHE), and edge segmentation using a fuzzy-based thresholding algorithm to address intensity variations and ambiguous boundaries. Experimental results demonstrated the effectiveness of the proposed approach, achieving an accuracy of 98.63%, precision of 97.44%, specificity of 97.93%, recall of 97.55%, and F1-score of 98.13%.

In the study [16], the authors proposed a two-stage cascaded segmentation model with integrated attention mechanisms designed to handle multimodal inputs from CT, MRI, and ultrasound for automated detection of lower extremity deep vein thrombosis. The framework was evaluated on a dataset of 322 patients with lower extremity DVT covering CT, MRI, and ultrasound modalities.

Different from the aforementioned studies that do not focus on the interpretability of the model's predictions, the proposed work focuses on the binary classification of venous thrombosis in ultrasound images using a CNN architecture with emphasis on the interpretability of the model's predictions through the application of explainable AI methods, Grad-CAM, and LIME. Therefore, the main objective of this study is to develop a CNN-based model for the binary classification of venous thrombosis in ultrasound images and to provide insights into the model's decision-making process using explainable AI techniques.

This work is organized as follows: Section II describes the dataset and the methodology used for training the CNN and applying explainable AI methods. Section III presents the results of the classification performance and the insights gained from the explainable AI. Finally, Section IV concludes the study and discusses potential future directions for research in this area.

II. Methodology

This section describes the dataset used for training and evaluating the CNN model, the architecture of the CNN, and the explainable AI methods applied to interpret the model's predictions.

A. Dataset

The dataset used in this study is composed of ultrasound images labeled for the presence or absence of venous thrombosis. The images were extracted from venous ultrasound examinations and were derived from the ThrombUS+ dataset [4], [17], [18], a collection of ultrasound examinations acquired within the Horizon Europe ThrombUS+ project and distributed through Zenodo in DICOM format. The frame-level images employed in this work were downloaded from the publicly available Kaggle repository, Thrombus vs Non-Thrombus Dataset Images [19]. The repository contains ultrasound image frames from the ThrombUS+ dataset extracted from venous ultrasound examinations and labeled according to the presence or absence of thrombosis. The ultrasound images correspond to four anatomical regions of the lower limb veins: Popliteal Vein (PV), Common Femoral Vein (CFV), Great Saphenous Vein (GSV), and Femoral Vein (FV). Fig. 1 illustrates the anatomical locations of these veins in the lower limb. For more details about the dataset, please refer to the work [4].

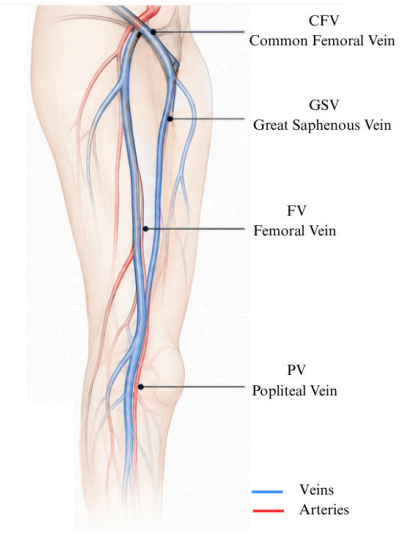


Fig. 1. Diagram of lower limb veins.

Fig. 2 shows examples of ultrasound images from the dataset used for training and evaluating the implemented method. The images in the dataset exhibit variations in terms of quality, contrast, and anatomical structures, which pose challenges for accurate classification.

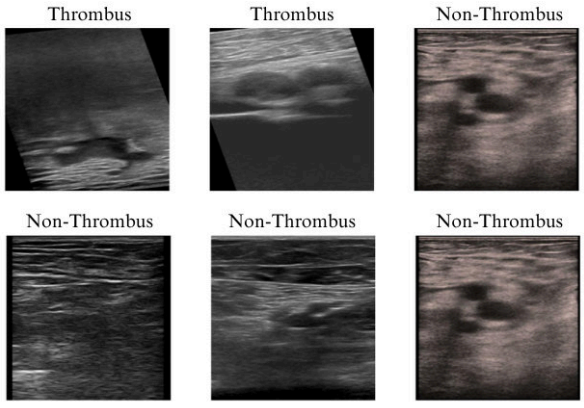


Fig. 2. Example of ultrasound images from the dataset used for training and evaluating the proposed method.

The selected dataset consists of a total of 1000 ultrasound images, with 800 images initially allocated for training and 200 images reserved for testing. The original dataset does not provide patient-level identifiers or metadata indicating whether the images correspond to individual patient examinations or to frames extracted from ultrasound examinations. Therefore, it was not possible to verify whether multiple images originated from the same patient. From the 800 images allocated to the

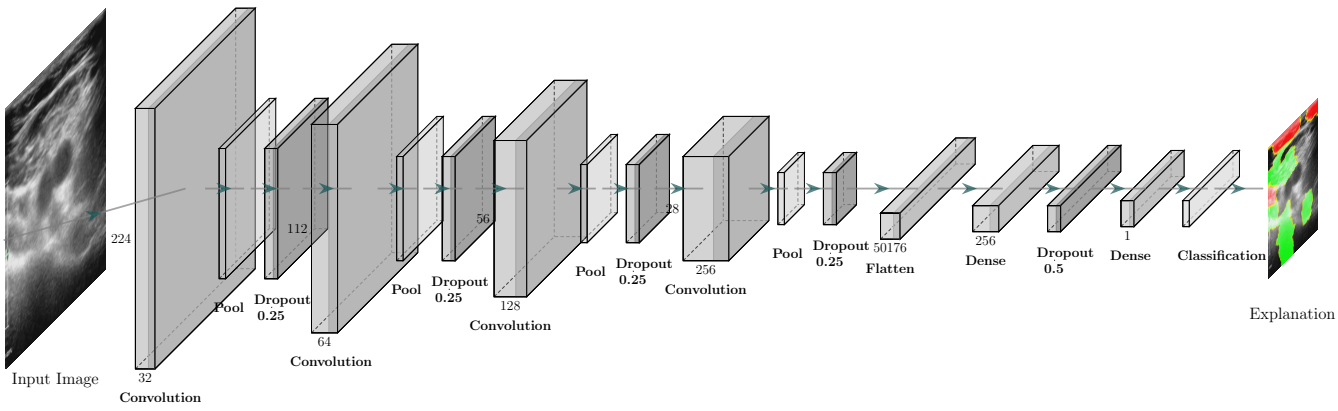


Fig. 3. Architecture of the proposed method for binary classification of venous thrombosis in ultrasound images.

training set, 20% were further separated for validation. Each image has a resolution of 224×224 pixels. The images were preprocessed by normalizing the pixel values to the range [0.0, 1.0].

B. Convolutional Neural Network

A CNN was implemented for the binary classification of venous ultrasound images into thrombus and non-thrombus categories. Fig. 3 illustrates the architecture of the proposed method. The model receives the ultrasound images as input and outputs a binary classification indicating the presence or absence of venous thrombosis. Moreover, after the classification, we apply explainable AI methods to interpret the model's decisions and identify the regions of the ultrasound images that contribute most to the classification results.

The architecture receives RGB images resized to 224×224 pixels as input and consists of four convolutional blocks. Each block comprises a 3×3 convolutional layer with ReLU activation and same padding, followed by batch normalization, a 2×2 max-pooling layer, and a dropout layer. The number of filters increases progressively from 32 to 64, 128, and 256, enabling the extraction of increasingly complex spatial features while reducing the spatial dimensions of the feature maps. The final layer consists of a single neuron with a sigmoid activation function given by:

$$\hat{y} = \frac{1}{1 + e^{-z}}, \tag{1}$$

where z is the output of the last fully connected layer. The sigmoid activation function maps the output to a probability value between 0 and 1. The output $\hat{y}$ represents the predicted probability of the positive class (thrombosis) for a given input image.

The network was trained using the Adam optimizer with a learning rate of 1×10^{-4} and the binary cross-entropy loss function given by:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{2}$$

where N is the number of samples, y_i is the true label for the i-th sample, and $\hat{y}_i$ is the predicted probability of the positive class for the i-th sample. The model was trained for 100 epochs with a batch size of 32. To prevent overfitting, early stopping was implemented with a patience of 15 epochs.

C. Explainable AI methods

To interpret the model's decisions and identify the regions of the ultrasound images that contribute most to the classification results, we applied two explainable AI methods: Grad-CAM and LIME. The following subsections describe each method in detail.

1) GRAD-CAM: is an explainability technique that highlights the image regions that most influence a CNN's prediction. Given an input image and a target class c, Grad-CAM computes the importance of each feature map in the last convolutional layer by analyzing the gradients of the class score with respect to the feature activations [11].

Let $A^k \in \mathbb{R}^{H \times W}$ denote the k-th feature map of the final convolutional layer, and let y^c represent the score associated with class c. The importance weight α_k^c of feature map A^k is obtained through global average pooling of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{3}$$

where $Z = H \times W$ corresponds to the total number of spatial locations in the feature map. The weight α_k^c measures how strongly the feature map A^k contributes to the prediction of class c. Positive values indicate that increasing the activation of the feature map tends to increase the class score.

The class-discriminative localization map is then computed as a weighted combination of all feature maps:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right), \tag{4}$$

where ReLU removes negative contributions and retains only the features that positively influence the prediction [11].

The resulting heatmap $L^c_{\text{Grad-CAM}}$ is subsequently resized to the original image dimensions and overlaid on the input image. Regions with higher activation values correspond to areas that contributed most strongly to the network's decision for class c, providing visual evidence of the model's reasoning process.

2) LIME: is an explainability technique that explains individual predictions by approximating the behavior of a complex model with a simpler and interpretable model [12]. Unlike Grad-CAM, LIME is model-agnostic and can be applied to any classifier without requiring access to its internal architecture or gradients.

Given an input image x and a trained classifier $f(\cdot)$, LIME first partitions the image into a set of interpretable components, called superpixels. Let

$$x' = [x'_1, x'_2, \ldots, x'_M] \tag{5}$$

denote the binary representation of the image, where $x'_m = 1$ indicates that the m-th superpixel is present and $x'_m = 0$ indicates that it is masked. A set of perturbed samples is generated by randomly turning superpixels on or off. For each perturbed image z', the original classifier produces a prediction $f(z')$.

LIME then assigns a proximity weight to each perturbed sample according to its similarity to the original image:

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right), \tag{6}$$

where $D(x, z)$ is a distance metric between the original image and the perturbed sample, and σ controls the neighborhood size. The explanation is obtained by fitting a simple interpretable model $g \in G$, typically a sparse linear model, that minimizes the objective

$$\xi(x) = \arg\min_{g \in G} \left[L(f, g, \pi_x) + \Omega(g)\right], \tag{7}$$

where $L(f, g, \pi_x)$ measures how closely the surrogate model approximates the predictions of the original classifier in the local neighborhood of x, and $\Omega(g)$ penalizes model complexity to maintain interpretability.

For a linear surrogate model, the explanation takes the form

$$g(z') = w_0 + \sum_{m=1}^{M} w_m z'_m, \tag{8}$$

where w_m quantifies the contribution of the m-th superpixel to the prediction. Positive coefficients indicate regions that support the predicted class, whereas negative coefficients indicate regions that oppose it [12].

Although both Grad-CAM and LIME aim to explain CNN predictions, they operate differently. Grad-CAM exploits the internal gradients and feature maps of a CNN to generate a class-discriminative heatmap. Consequently, it requires access to the network architecture and differentiable layers. In contrast, LIME treats the model as a black box and explains predictions by observing how the output changes when interpretable image regions are perturbed [20]. LIME estimates feature importance by fitting a local surrogate model whose coefficients directly indicate the contribution of each superpixel.

III. Results and Discussion

The proposed CNN model was evaluated on an independent test set comprising 200 ultrasound images using an classification threshold of 0.756. The classification threshold of 0.756 was selected using the validation set by maximizing the Youden index, defined as $J = \text{Sensitivity} + \text{Specificity} - 1$. Thus, the selected threshold represents a statistical operating point that balances sensitivity and specificity. Performance metrics were estimated together with 95% confidence intervals (CI) obtained through bootstrap resampling with 2,000 iterations.

Table I summarizes the performance of the CNN model on the test set, including sensitivity (recall), specificity, precision, negative predictive value (NPV), Area Under the Receiver Operating Characteristic Curve (AUC-ROC), AUC-PR Area Under the Precision-Recall Curve (AUC-PR), Matthews Correlation Coefficient (MCC), F1-score, positive likelihood ratio (LR+), negative likelihood ratio (LR–), and accuracy.

TABLE I
Performance of the CNN Model on the Test Set ($n = 200$)

Metric	Value	95% CI
Recall	0.6500	[0.5556, 0.7423]
Specificity	0.8900	[0.8229, 0.9490]
Precision	0.8553	[0.7733, 0.9325]
NPV	0.7177	[0.6339, 0.7953]
AUC-ROC	0.8608	[0.8041, 0.9114]
AUC-PR	0.8572	[0.7852, 0.9195]
MCC	0.5563	[0.4321, 0.6649]
F1-Score	0.7386	–
LR+	5.9091	–
LR–	0.3933	–
Brier Score	0.1794	–
Accuracy	0.7700	–

The MCC value of 0.556 further suggests a moderate positive correlation between the predicted and true labels, supporting the model's overall discriminative capability. The positive predictive value was 85.5%, meaning that when the model predicts thrombosis, the prediction is correct in most cases. Conversely, the NPV was 71.8%, indicating that approximately seven out of ten negative predictions correspond to true non-thrombotic examinations.

Fig. 4 shows the calibration curve, which compares the predicted probabilities of thrombosis with the observed frequencies. The reliability analysis revealed moderate calibration errors with an expected calibration error (ECE) of 0.171, indicating that the predicted probabilities do not perfectly reflect the true occurrence frequencies. The calibration assessment is further supported by the Brier Score of 0.179, which indicates moderate probabilistic accuracy. Nevertheless, the confidence distributions show a reasonable separation between thrombus and non-thrombus cases, suggesting that the CNN achieves modest discriminative ability.

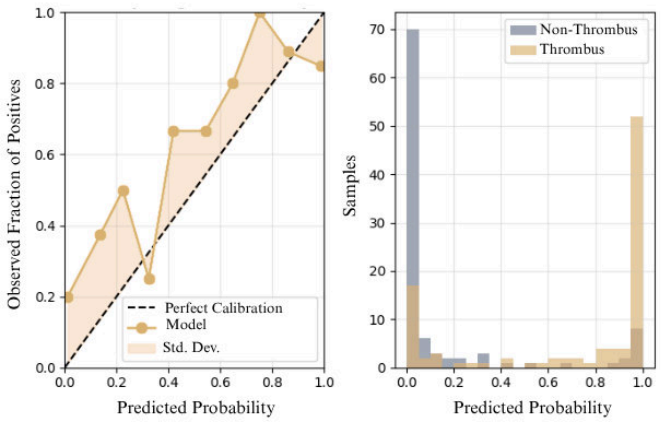


Fig. 4. Calibration curve with ECE of 0.171 and confidence distributions achieved by the CNN model.

Fig. 5 illustrates the AUC-ROC and AUC-PR curves. The AUC-ROC of 0.8608 indicates reasonable overall performance, while the AUC-PR of 0.8572 suggests that the model maintains good precision across a range of recall values.

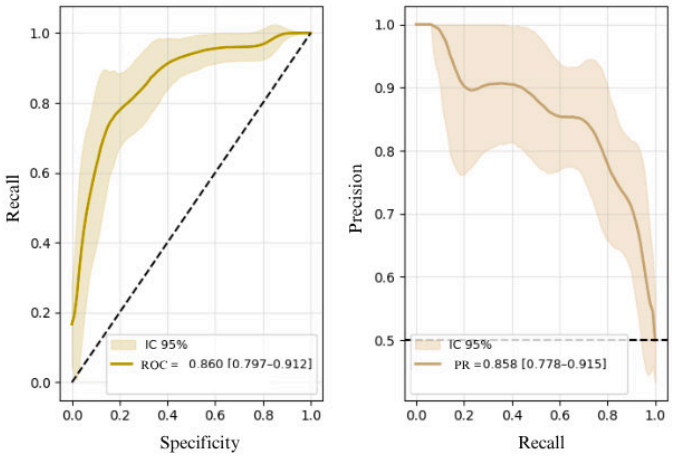


Fig. 5. AUC-ROC and AUC-PR curves for the CNN model on the test set.

Fig. 6 shows the confusion matrix. The model correctly classified 65 thrombotic cases and 89 non-thrombotic cases, while it misclassified 11 non-thrombotic cases as thrombotic and 35 thrombotic cases as non-thrombotic.

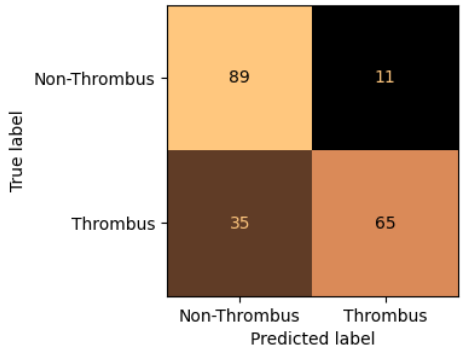


Fig. 6. Confusion matrix for the CNN model on the test set.

To explain the model's decisions and identify the regions of the ultrasound images that contribute most to the classification results, we applied the Grad-CAM and LIME techniques. Fig. 7 shows an example of a Grad-CAM visualization for a correctly classified thrombotic case. The heatmap indicates that the model focuses on specific regions of the ultrasound image. These regions may be associated with visual characteristics of venous thrombosis. In this example, the model predicted the presence of thrombus with $P(Y) = 0.984$, where Y denotes the presence of thrombus.

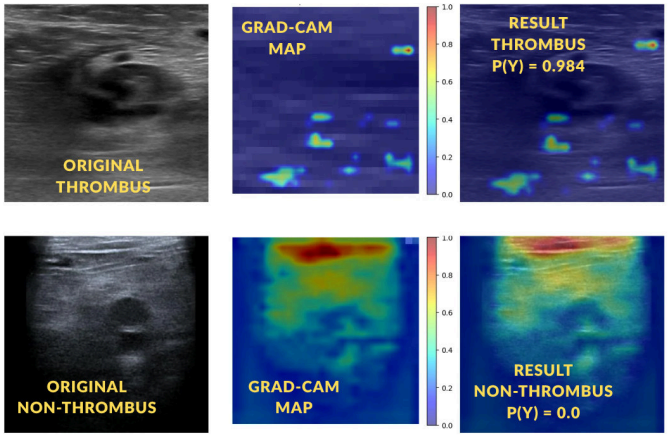


Fig. 7. Grad-CAM visualization of the proposed method. The heatmap highlights the regions that contribute most to the classification decision.

Fig. 8 presents representative misclassification cases together with their Grad-CAM visualizations. In the false-positive example, the network incorrectly classified a non-thrombotic image as thrombus with $P(Y) = 0.893$. The Grad-CAM visualization shows attention distributed across multiple heterogeneous regions rather than a clearly localized vascular structure.

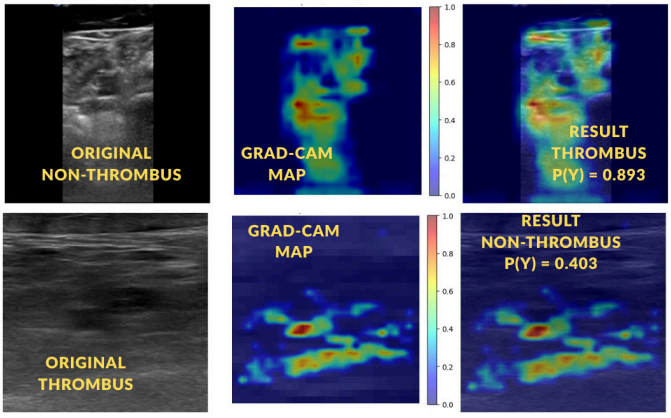


Fig. 8. Grad-CAM visualization with false positive and false negative cases. The heatmaps indicate the regions of the ultrasound images that influenced the model's incorrect predictions.

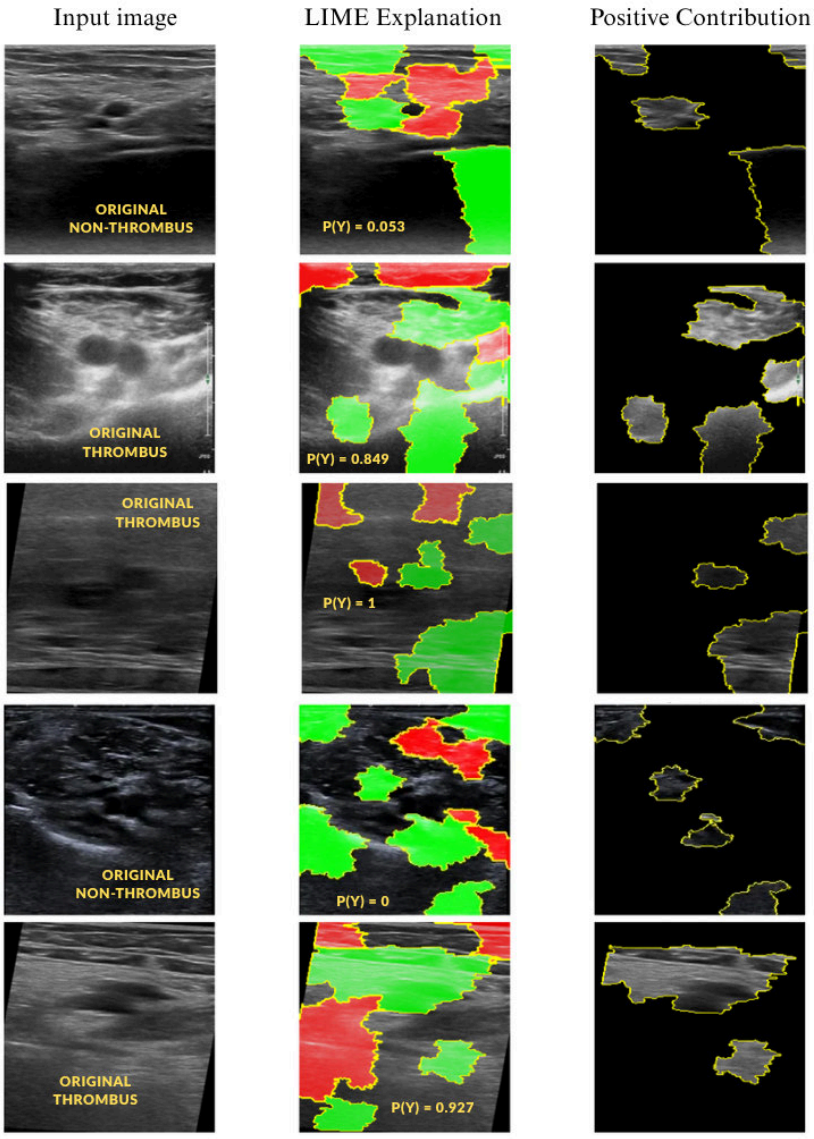


Fig. 9. LIME visualization of the proposed method. The highlighted areas indicate the regions of the ultrasound image that are most influential in the model's prediction.

Conversely, in the false-negative example, the model failed to assign sufficient importance to the thrombus region with only $P(Y) = 0.403$, resulting in a non-thrombus prediction. The Grad-CAM maps indicate that the CNN relies on broader contextual ultrasound patterns, which may contribute to classification errors when these patterns resemble features associated with thrombosis.

Fig. 9 depicts representative LIME explanations for correctly classified thrombus and non-thrombus ultrasound images.

Green regions indicate positive contributions toward the thrombus class, whereas red regions represent features associated with a prediction against the thrombus class. In the thrombus cases, positive superpixels were frequently observed around vascular structures and regions with visual patterns associated with thrombosis. Conversely, non-thrombus cases generally showed less positive evidence for the thrombus class and larger regions associated with predictions against it. These observations provide a qualitative indication of the image regions associated with the model's predictions.

Although the exact anatomical correspondence between the highlighted superpixels and thrombus locations cannot be confirmed without expert annotations, the LIME results suggest that localized image features may be associated with the CNN's predictions. In the examples analyzed, LIME produced more focused explanations than the Grad-CAM visualizations, which often showed broader activation regions. This difference is reported as an exploratory qualitative observation and should not be interpreted as evidence of the superiority of LIME over Grad-CAM. A quantitative assessment involving expert evaluation, anatomical annotations, or explanation-fidelity metrics

would be required to establish a definitive comparison.

IV. Conclusion

This study presented a CNN-based approach for the binary classification of venous thrombosis in ultrasound images, achieving an accuracy of 77% and an AUC-ROC of 0.86 on the test set. Beyond predictive performance, the study investigated the interpretability of the model through two widely used explainable artificial intelligence techniques. The explainability analysis revealed differences between the methods. Grad-CAM provided a global visualization of the regions influencing the network's predictions, but the resulting activation maps were often diffuse and extended beyond the most relevant anatomical structures. In contrast, LIME produced visually more localized maps in the examples examined by highlighting specific image regions that positively or negatively contributed to the classification outcome. The qualitative analysis suggested that LIME offered more focused explanations, facilitating the identification of discriminative image patterns associated with thrombus and non-thrombus cases. Overall, the results demonstrate that combining CNNs with explainable AI techniques can improve the transparency of thrombosis detection systems.

Although the proposed CNN demonstrated good discriminative ability, the calibration analysis revealed that the predicted probabilities do not perfectly reflect the true occurrence frequencies of thrombosis. Therefore, future studies will investigate probability calibration approaches, such as temperature scaling and isotonic regression, to improve the reliability of confidence estimates while preserving classification performance. Furthermore, since the current approach analyzes individual ultrasound frames independently, future work will explore temporal deep learning models capable of capturing the sequential dependencies between consecutive frames. Incorporating temporal information from ultrasound videos may provide richer contextual information and improve the robustness and accuracy of thrombosis detection, particularly in challenging cases where diagnostic evidence is distributed across multiple frames.

References

[1] A. S. Wolberg, F. R. Rosendaal, J. I. Weitz, I. H. Jaffer, G. Agnelli, T. Baglin, and N. Mackman, "Venous thrombosis," *Nature reviews Disease primers*, vol. 1, no. 1, p. 15006, 2015.

[2] M. Stubbs, M. Mouyis, and M. Thomas, "Deep vein thrombosis," *Bmj*, vol. 360, 2018.

[3] M. Didaskalou, G. Ioannakis, E. Kaldoudi, and G. Drosatos, "Deep learning approaches for thrombosis detection and risk assessment via ultrasound imaging: A scoping review," *Ultrasound in Medicine & Biology*, 2025.

[4] S. Didaskalou, N. Portokallidis, K. Tzatzimaki, M. Liapi, N. Moustakidis, T. Moustakidis, N. Balciuniene, A. Macas, R. Kijauskas, A. Aladaitis *et al.*, "A dataset for benchmarking machine learning models for autonomous deep vein thrombosis detection based on compression ultrasound videos." in *BIOSTEC (4)*, 2026, pp. 853–862.

[5] S. Goodacre, F. Sampson, S. Thomas, E. van Beek, and A. Sutton, "Systematic review and meta-analysis of the diagnostic accuracy of ultrasonography for deep vein thrombosis," *BMC medical imaging*, vol. 5, no. 1, p. 6, 2005.

[6] F. Akram, B. E. Fan, C. W. Tan, W. C. Teoh, P. Prandoni, and E. S. Yap, "The clinical application of venous ultrasound in diagnosis and follow-up of lower extremity deep vein thrombosis (dvt): a case-based discussion," *Thrombosis Journal*, vol. 21, no. 1, p. 110, 2023.

[7] C. Wei, J. Wang, P. Yu, A. Li, Z. Xiong, Z. Yuan, L. Yu, and J. Luo, "Comparison of different machine learning classification models for predicting deep vein thrombosis in lower extremity fractures," *Scientific Reports*, vol. 14, no. 1, p. 6901, 2024.

[8] R. Arun, B. R. C. Joseph, B. Muthukumar, and A. Ahilan, "Deep vein thrombosis detection via combination of neural networks," *Biomedical Signal Processing and Control*, vol. 100, p. 106972, 2025.

[9] J. F. do Nascimento Filho, "A deep learning approach for deep-vein thrombosis detection in ultrasound videos." in *BIOSTEC (3)*, 2026, pp. 419–424.

[10] W. Qiu, P. Cui, S. Li, Z. Tang, J. Chen, J. Wang, and Y. Li, "Machine learning models predict risk of lower extremity deep vein thrombosis in hospitalized patients with spontaneous intracerebral hemorrhage," *Scientific Reports*, vol. 15, no. 1, p. 24932, 2025.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[13] E. E. Contreras-Luján, E. E. García-Guerrero, O. R. López-Bonilla, E. Tlelo-Cuautle, D. López-Mancilla, and E. Inzunza-González, "Evaluation of machine learning algorithms for early diagnosis of deep venous thrombosis," *Mathematical and Computational Applications*, vol. 27, no. 2, p. 24, 2022.

[14] J. W. Seo, S. Park, Y. J. Kim, J. H. Hwang, S. H. Yu, J. H. Kim, and K. G. Kim, "Artificial intelligence-based iliofemoral deep venous thrombosis detection using a clinical approach," *Scientific Reports*, vol. 13, no. 1, p. 967, 2023.

[15] R. Arun, B. Rogers Cross Joseph, B. Muthukumar, and A. Ahilan, "Deep vein thrombosis detection via combination of neural networks," *Biomedical Signal Processing and Control*, vol. 100, p. 106972, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809424010309

[16] H. Wang, Y. Jing, H. Li, Y. Zhang, C. Chang, and G. Shi, "Multimodal deep learning with attention mechanisms for automated detection of lower extremity deep vein thrombosis: Integrating ultrasound, ct, and mri," *Biomedical Signal Processing and Control*, vol. 119, p. 109749, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809426003034

[17] "Thrombus+ training dataset," 2025. [Online]. Available: https://zenodo.org/records/17659415

[18] S. Didaskalou, N. Portokallidis, and E. Kaldoudi, "Thrombus+ ultrasound dataset #1 - compression ultrasound videos - training set," Nov. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.17659415

[19] E. Kaldoudi, "Classification of thrombus/non-thrombus-lowerlimb," 2026. [Online]. Available: https://www.kaggle.com/dsv/16888108

[20] Y. H. Chel and L. L. Poh, "Brain tumor classification in mri: Insights from lime and grad-cam explainable ai techniques," *IEEE Access*, 2025.

X-Pan: Building, Documenting and Validating an OPG Dataset

Nicolas B. Sousa
Department of Electronics - DAELN
Department of Informatics - DAINF
Federal University of Technology - Paraná
Curitiba, Brazil
nicolasbarbierisousa@gmail.com

Gustavo B. Borba
Department of Electronics - DAELN
Graduate School on Biomedical Engineering - PPGEB
Federal University of Technology - Paraná
Curitiba, Brazil
gustavobborba@utfpr.edu.br

Sérgio Lúcio P. de C. Lopes
Department of Diagnostics and Surgery
São Paulo State University
São José dos Campos, Brazil
sergio.lopes@unesp.br

Ana Amélia Barbieri
Department of Social Dentistry and Pediatric Clinic
São Paulo State University
São José dos Campos, Brazil
a.barbieri@unesp.br

***Abstract*—X-Pan, a dataset of 1,566 de-identified panoramic radiographs (OPGs), is presented and tested for forensic dental age and sex estimation. Each instance contains associated tabular demographic and clinical attributes. The dataset was curated through an eligibility-driven sampling process from institutional records, with expert-supported labeling and structured documentation following the Datasheets for Datasets framework. Preprocessing includes privacy-preserving de-identification via manual cropping to remove identifying non-anatomical content, followed by dataset cleaning and label assignment according to predefined guidelines. To validate its usability, the dataset was employed in two baseline experiments: age estimation applying publicly available zero-shot weights, which achieved an MAE of 1.04 years (model 912, ages 1–24), and sex estimation fine-tuning three ImageNet-pretrained backbones, which achieved a best F1-score of 78.08% and AUC of 0.841 with ResNet50. To support reproducibility, dataset releases follow semantic versioning and all materials — including model weights, code, and the complete datasheet — are available at the project repository.**

***Index Terms*—orthopantomography dataset, forensic odontology, age estimation, sex estimation**

I. Introduction

Age estimation in living or deceased individuals is a core issue in forensic science, with applications ranging from criminal and civil proceedings to the identification of disaster victims and undocumented individuals. Forensic odontology plays a key role in this process because teeth are among the most durable tissues in the human body, remaining preserved even under extreme conditions such as charring, and because dental development follows a predictable chronology that enables age estimation from the formation and mineralization of crowns and roots visible in radiographic examinations [1]–[3].

Among radiographic modalities, panoramic radiographs (orthopantomographs, OPGs) are particularly suited to this task because a single image provides a comprehensive view of all teeth, the mandible, the maxilla, and the temporomandibular joint, permitting simultaneous analysis of eruption chronology and corono-radicular mineralization across the full dentition [4]. As a result, OPGs are the predominant image type in machine learning studies on forensic age estimation; alternatives such as cone-beam computed tomography are less common, in part due to their higher cost and more limited clinical availability [5].

The advancement of artificial intelligence (AI), and deep learning in particular, is closely tied to data availability [6], making large, well-documented datasets a prerequisite for effective models [7]. AI systems have been applied to forensic dentistry for over 20 years, with an approximately fivefold increase in publications across the periods 2011–2015 (2 articles), 2016–2020 (11 articles), and 2021–2023 (49 articles) [8, p. 89, 90]. Of the studies reviewed by Anees et al. [8, p. 89, 90], 46.27% (31 articles) focus on age estimation, while the remaining literature addresses human identification and estimates related to sex, dental characteristics, and implants.

However, the reproducibility of this body of work is limited. Representative examples of studies that use panoramic radiographs for machine learning tasks and provide limited or no access to their source code and model weights, and no instructions on how to obtain their datasets include [9]–[13]. Furthermore, none of the analyzed studies combine a detailed dataset description with instructions for acquiring the data for subsequent research.

To address this gap, this paper presents X-Pan, an annotated dataset of 1,566 OPGs (ethics committee approval: protocol No. 7.560.465; Instituto de Ciência e Tecnologia de São José dos Campos) documented following the Datasheets for Datasets framework [14], which includes extensive in-

formation on the dataset's motivation, composition, collection, preprocessing, cleaning, labeling, usage, distribution, and maintenance. These details will be summarized in the coming sections. To validate the dataset's usability, two baseline experiments are reported: age estimation using publicly available model weights from Heinrich et al. [15] and Koch et al. [16], and sex estimation through fine-tuning of three ImageNet-pretrained backbones.

II. Motivation

X-Pan was created primarily to support the development of deep learning models for automated dental age estimation, addressing the reproducibility gap identified in the Introduction: the lack of documented, accessible OPG datasets for subsequent research. Its construction was also designed to support a wider range of dental imaging studies, including sex estimation and population analyses. Detection of clinical interventions is a longer-term goal contingent on targeted expansion of the relevant classes (Section VI).

For the development of effective models in these tasks, large, diverse, and high-quality data are essential [7]. X-Pan aims to contribute to this need by providing a structured, standardized dataset with documented metadata, expert-supported labeling, and semantic versioning to facilitate reproducibility and reuse.

The dataset was created by the authors of this paper, affiliated with the Departments of Diagnostics and Surgery and of Social Dentistry and Pediatric Clinic at São Paulo State University, and the Departments of Electronics and of Informatics at the Federal University of Technology – Paraná. No grant was involved in its creation.

III. Composition

The dataset is composed of 1,566 panoramic radiographs, each corresponding to a single patient. Each instance is accompanied by the tabular attributes described in Table I, stored in a CSV file. The primary targets for supervised learning tasks are the `age` and `sex` fields.

Fig. 1 shows the age pyramid, stratifying the age distribution by sex. The dataset contains significantly more images from minors than from adults, a skew that prospective users should account for when designing experiments.

Table II shows the sex distribution. Sex is nearly balanced, with male patients accounting for 50.57% of instances.

The acquisition device is also recorded for each image (Table I): 1,091 instances (69.67%) captured with OP300 and 475 (30.33%) with Orthophos XG.

Table III shows the distribution of categorical attributes. The *pathol* attribute comprises the categories bone fracture, furcation lesion, endodontic-periodontal lesion (EPL), bone pathology, periapical lesion, and none. The *findings* attribute comprises the categories anomaly, foreign body, calcification, and none. For *braces*, *retainer*, and *orthog*, the allowed values are S (superior arch), I (inferior arch), SI (both arches), and N (none). The combined value SI is used in place of separate S and I entries to avoid ordering ambiguity in multi-value fields.

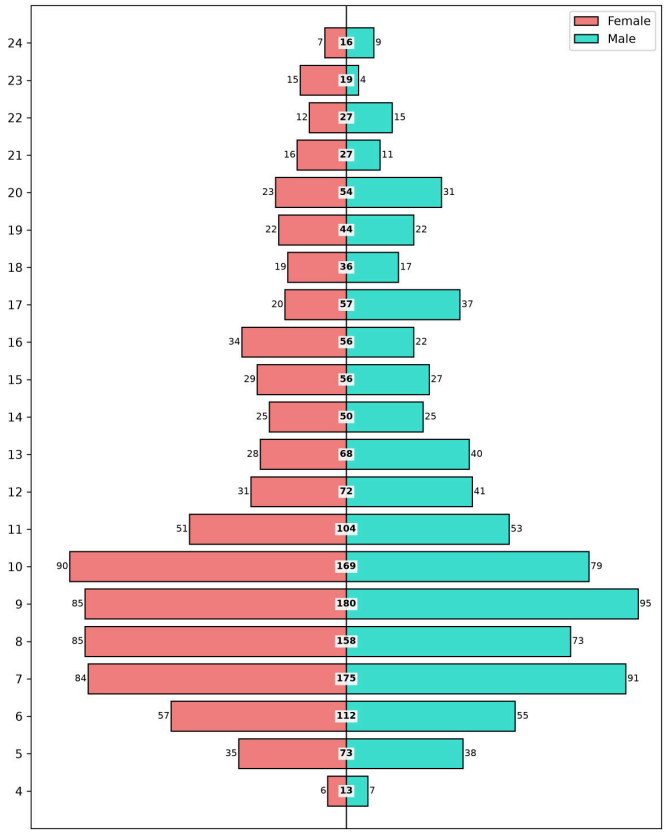


Fig. 1. Age distribution stratified by sex (pink: female; teal: male). Horizontal axis: number of images; vertical axis: age in years. The centered number represents the total number of images for the corresponding age.

Tables IV and V show the distributions for the numeric and pathology/findings attributes, respectively.

During the early phases of data collection, panoramic radiographs presenting orthodontic appliances (e.g., braces and retainers), foreign bodies, calcifications, signs of endodontic treatment, orthognathic surgery-related features, pathological findings, or absence of teeth were excluded. These criteria were later revised to admit the aforementioned cases, thereby increasing clinical diversity. Consequently, the prevalence of these characteristics in the final dataset is relatively low, reflecting their partial exclusion during the initial collection phase.

After revision, the remaining criteria were that images were required to contain age and sex records, be sharp and display adequate image quality, and allow complete visualization of the dental and osseous structures under study. Only images from patients ranging from 4 to 24 years of age were collected. All retained instances are complete – no metadata fields are missing.

The dataset is a non-probabilistic, consecutive sample drawn from a larger institutional database. A minimum sample size was determined using the standard finite-population proportion estimator:

TABLE I
ATTRIBUTES STORED IN THE X-PAN CSV METADATA FILE FOR EACH INSTANCE, INCLUDING THEIR DATA TYPES, ALLOWED VALUES, AND DESCRIPTIONS.

Field	Type	Allowed values / units	Description
image_name	string	filename / ID	Associates CSV row with image file
device	categorical	OP300, Orthophos XG	Radiograph acquisition device
month_of_capture	date	MM/YYYY; e.g., 04/2010	Month and year the radiograph was captured
age	integer	months	Patient's age in months
sex	categorical	M (male) , F (female)	Patient's biological sex
braces	categorical	S (superior), I (inferior), SI (both), N (none)	Presence/location of braces
retainer	categorical	S (superior), I (inferior), SI (both), N (none)	Presence/location of retainer
orthog	categorical	S (superior), I (inferior), SI (both), N (none)	Presence/location of orthognathic surgery indicators
endo	integer	0,1,2,...	Number of endodontic treatments visible
implan	integer	0,1,2,...	Number of dental implants visible
pathol	list / categorical	comma-separated terms	Pathologies present (e.g., periapical_lesion)
findings	list / categorical	comma-separated terms	Radiologist findings (e.g., foreign_body)

TABLE II
DISTRIBUTION OF PATIENT SEX

Sex	Count	%
Male	792	50.57
Female	774	49.43

TABLE III
DISTRIBUTION OF CATEGORICAL ATTRIBUTES. "SUPERIOR" AND "INFERIOR" REFER TO THE JAW IN WHICH THE APPLIANCE OR FEATURE IS PRESENT. "BOTH" CORRESPONDS TO THE "SI" VALUE IN THE CSV. "NONE" INDICATES ABSENCE.

	Superior		Inferior		Both		None	
Attribute	Count	%	Count	%	Count	%	Count	%
Braces	29	1.9	15	1.0	86	5.5	1,436	91.7
Retainer	0	0	52	3.4	4	0.3	1,510	96.4
Orthog	0	0	0	0	0	0	1,566	100.0

$$n = \frac{Z^2\, p(1-p)}{c^2} \tag{1}$$

where n is the target sample size, Z is the normal quantile associated with the specified confidence level, p is the anticipated population proportion, and c is the tolerable margin of error in proportion units. Applied to the Brazilian population aged 5 to 24 years (57,266,806 inhabitants [17]), with a 99% confidence level, a 5% margin of error, and $p = 0.5$ (maximizing the Bernoulli variance and yielding the most conservative estimate), the formula gives a minimum of 664 images. The present dataset contains 1,553 instances within this age range, exceeding that threshold by a factor of approximately 2.4. Representativeness relative to the broader source population was not formally validated beyond this criterion.

Each image is treated as an independent instance: there is no longitudinal connection between records, and each patient appears only once. The dataset is self-contained and does not depend on external resources.

All image-derived attributes (`braces`, `retainer`, `orthog`, `endo`, `implan`, `pathol`, `findings`) were independently labeled by three annotators following a written guideline, with the final value for each field determined by majority vote among the three assessments. No train-test-validation split is currently recommended by the authors.

All personally identifying information was previously anonymized, making it impossible to directly identify individuals included in the dataset.

The original radiographs constitute medical records protected by patient confidentiality. They were obtained for diagnostic and clinical purposes, archived in an institutional database authorized for scientific use, and de-identified prior to researcher access. Participants authorized use of their data by signing an Informed Consent Form. Additional data protection measures include restricted access control and removal of all direct identifiers. Given that no direct identifiers remain and the dataset is not linked to any external database, re-identification of individuals is considered infeasible; authorized users are nonetheless contractually prohibited from attempting it.

The dataset contains medical radiographic images, which may be distressing in a clinical context but do not depict graphic content.

IV. COLLECTION PROCESS

A. Data source and setting

Panoramic radiographs and associated clinical records were obtained directly from the partner institution's radiographic database, where they are stored in a Picture Archiving and Communication System (PACS). Images were not sourced from public websites or third parties. Patient age was derived from dates of birth and exam dates extracted from both physical and digital medical records, and was verified against source documents by the research team.

B. Time frame and sampling

Dataset construction took place between 2025 and 2026, through retrospective selection of images from a larger institutional archive whose records predate this period, spanning April 2010 to December 2025 (month and year of capture are recorded in the metadata). A non-probabilistic convenience/consecutive sampling was used: images were selected

TABLE IV
DISTRIBUTION OF NUMERICAL ATTRIBUTES

	0		1		2		≥ 3	
Attribute	Count	%	Count	%	Count	%	Count	%
Endodontic treatments	1,539	98.3	22	1.4	4	0.2	1	0.1
Dental implants	1,565	99.9	1	0.1	0	0	0	0

TABLE V
DISTRIBUTION OF PATHOLOGY AND FINDINGS CATEGORIES

Attribute	Category	Count	%
Pathol	Bone fracture	0	0
	Furcation lesion	0	0
	EPL	0	0
	Bone pathology	0	0
	Periapical lesion	1	0.1
	None	1,565	99.9
Findings	Anomaly	2	0.1
	Foreign body	0	0
	Calcification	0	0
	None	1564	99.9

based on availability and eligibility per inclusion/exclusion criteria, as detailed in Sec. III, with the aim of preserving demographic and radiographic variability (age ranges, biological sex, imaging patterns).

C. Data types and capture

Each instance comprises a panoramic radiograph and associated demographic metadata (biological sex, chronological age) recorded at patient intake prior to imaging. Radiographs were captured using one of two devices: Orthophos XG (Sirona Dental Systems, Germany) or OP300 (Instrumentarium, Finland). The specific capture device is stored as image metadata (Table I), and its distribution is described in Section III.

D. Data provenance, verification, and access

Clinical records and images originate from the institution's database to which the research group has authorized scientific access. Extracted metadata were verified by the research team for consistency and eligibility against study criteria. No remuneration beyond regular institutional salaries was paid to students, radiologists, or staff involved in data collection.

E. Ethics, consent, and anonymization

The study received approval from the Research Ethics Committees of both participating institutions — the Institute of Science and Technology of São Paulo State University (ICT/UNESP) and the Federal University of Technology – Paraná (UTFPR). Approval documents and committee reference numbers are archived and available upon institutional or editorial request. Patients were informed, as part of routine clinical care, that their clinical data and imaging exams may be used for academic and scientific purposes, and provided written consent via an Informed Consent Form. All data were de-identified prior to any experiments. Minimization measures were applied, including removal of direct identifiers and restricted access control.

F. Withdrawal and retraction

Once de-identified and incorporated into the dataset, images cannot be retroactively removed through any technical mechanism. However, patients may oppose future use of their data for research in accordance with applicable institutional policies.

V. PREPROCESSING, CLEANING, AND LABELING

Patient-identifying regions were removed by cropping, performed manually using Microsoft Photos (Windows 11). Images with missing age or sex metadata were excluded. Labels correspond to the attributes described in Table I.

The original unprocessed images and raw metadata were retained for traceability, stored separately on an access-restricted device.

Image-derived labels were assigned by majority-vote consensus among three independent annotators following a predefined written guideline, as described in Sec. III.

All scripted processing steps are archived in the project repository at https://github.com/Nickalverst/X-Pan with versioned filenames; all software and tool versions are recorded.

VI. USAGE

X-Pan has been used in the two baseline validation experiments reported in this paper: age estimation using publicly available model weights, and sex estimation through fine-tuning of three ImageNet-pretrained backbones (Section IX). All subsequent papers and systems that make use of X-Pan will be compiled, along with the up-to-date datasheet and the present paper, in the project repository.

The dataset is suitable for tasks inherent to panoramic radiographic analysis [18], including age and sex estimation and population-level dental imaging studies. Detection of clinical interventions (e.g., endodontic treatments, orthodontic appliances) is not currently supported by the dataset's composition; see the limitation below.

Several compositional factors may affect the performance and generalizability of models trained on X-Pan. First, all images originate from a single Brazilian institutional database and were captured using one of two specific devices; models may therefore underperform when applied to images from other institutions, regions, or equipment [19]–[21]. Second, sex distribution is nearly balanced (50.57% male), but age distribution is not: the dataset contains significantly more images of minors than of adults, as shown in Fig. 1. Models

trained on X-Pan should be expected to perform less reliably in adult age ranges. Third, clinical findings such as pathologies, endodontic treatments, and orthodontic appliances have low prevalence due to their partial exclusion during early collection. The current composition does not support such studies, which remain future work pending targeted class expansion.

The dataset should not be used for clinical diagnosis without independent validation on the target population and imaging equipment. Any attempt to re-identify individuals from the dataset is prohibited under the terms of use, which are available as an appendix to the datasheet.

VII. Distribution

The dataset is attainable by third parties exclusively for academic, scientific, or technological research, upon formal request to the dataset custodian (a.barbieri@unesp.br) and subsequent approval by the ethics committee. Access requires submission of signed agreement with the dataset's terms of use, confidentiality, and data protection requirements. Access requests may be submitted from the date of publication of this paper.

When approved, data will be shared via electronic transfer; no unrestricted public release is planned. The dataset is not issued under an open public license and has no associated access fees. A terms-of-use document governing authorized access is available through the project repository. The dataset currently has no DOI.

No third-party intellectual property restrictions are known to apply to the de-identified images. The original radiographs are held by the source institution, which has authorized their use for scientific purposes as described in Section IV.

VIII. Maintenance

The dataset is hosted and maintained by the lead researcher for programming and computational development (nicolasbarbierisousa@gmail.com), under supervision of the project research team. Dataset-related enquiries and access requests may also be directed to the dataset curator (a.barbieri@unesp.br). Data storage takes place in a controlled-access environment with security measures appropriate for anonymized radiographic data. There is currently no erratum.

A. Updates and versioning

The dataset will be reviewed and updated semiannually to add eligible new instances and, where necessary, correct labeling errors, standardize data, or remove inconsistent records. All changes will be documented and version-controlled; previous versions of the documentation will be retained and remain accessible through the project repository so that published experiments can be reproduced. Releases follow semantic versioning (MAJOR.MINOR.PATCH), where MAJOR denotes breaking changes, MINOR denotes backward-compatible additions, and PATCH denotes backward-compatible corrections. Updates and version history will be communicated via the repository.

No fixed data retention period was communicated to participants at the time of consent. Data will be retained indefinitely under the controlled-access conditions described in Section VII, subject to any future changes in applicable legislation or institutional policy.

B. Contributions and curation

There is no public contribution pipeline at present. External researchers may propose contributions by contacting the dataset curator at a.barbieri@unesp.br. Proposed additions will be reviewed by the research team for methodological compatibility, data quality, image standardization, and ethical compliance before incorporation. Accepted contributions will be integrated, documented under the appropriate version tag, and announced through the project repository.

IX. Validation

To illustrate the complete pipeline described in this paper, Fig. 2 shows an overview of the process, from data acquisition to validation results. The experiments and their results are described below.

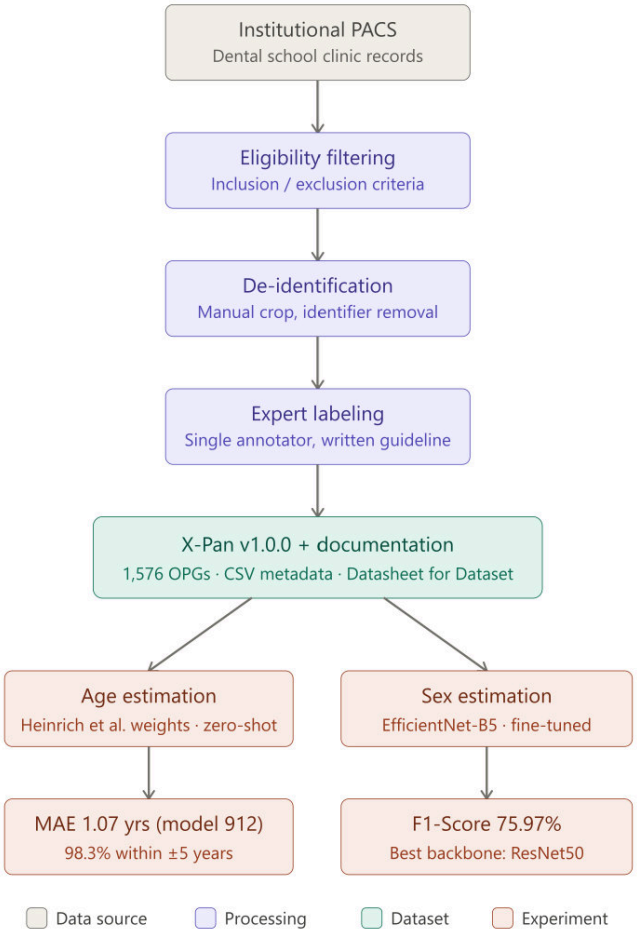


Fig. 2. Dataset construction and validation pipeline, from institutional data acquisition to the X-Pan v1.0.0 release and baseline validation experiments.

To demonstrate the usability of X-Pan with machine learning algorithms, two experiments were carried out. The first consisted of applying pre-trained weights from Heinrich et al. [15] (model 2216, for ages 2–89 years) and Koch et al. [16] (model 912, for ages 1–24 years) to estimate subject age. The second focused on sex estimation, in which three ImageNet-pretrained backbones — EfficientNet-B0, EfficientNet-B5, and ResNet50 — were fine-tuned on X-Pan as baseline experiments, using [22] as a reference for methodology and comparison.

A. Age estimation

For age estimation, the code from [15] was adapted to preprocess JPG instead of DICOM images. Both models were applied to all 1,566 instances. The results are shown in Table VI.

Model 912 achieved a 0.57-year (approx. 35.4%) lower overall MAE than model 2216, which is expected given that its training population (ages 1–24 years) closely matches the demographic distribution of X-Pan.

Grad-CAM activations represent the regions that most influenced the model's inference and should not be interpreted as evidence that a single anatomical structure determined the prediction.

Grad-CAMs generated by model 2216 are shown in Fig. 3. In panel A (ages 4–9), primary activations are anatomically compatible with the right maxillary tuberosity and posterior alveolar process in the area of upper first permanent molar root formation, the right mandibular ramus and mandibular angle, the left coronoid process, and the left mandibular ramus in the region corresponding to developing lower premolars. Secondary activations in the nasal cavity and maxillary sinuses may reflect craniofacial growth patterns, while notable activations anatomically compatible with the cervical vertebrae suggest the model draws on skeletal maturation cues beyond dental development alone. In panel B (ages 10–19), activations shift toward the maxillary tuberosity and the second permanent molar and upper third molar germ region, the right mandibular ramus and retromolar area corresponding to the developing lower third molar, the mandibular angle, and the left mandibular incisure; discrete activations persist along the mandibular body and in regions compatible with the cervical vertebrae and skull base. In panel C (ages 20–24), the pattern is more diffuse: primary foci are the right lower premolar region and the nasal cavity floor, with discrete activations near the lower third molar roots and mandibular body, and small activations near the mandibular condyles and superior image edges possibly related to the skull base or structural superposition. This reduced specificity is consistent with the diminishing discriminability of radiographic age markers in adulthood and with the wider confidence intervals observed for this group in Table VI.

B. Sex estimation

EfficientNet-B5, EfficientNet-B0 [23], and ResNet50 [24] backbones pretrained on ImageNet were fine-tuned on X-Pan

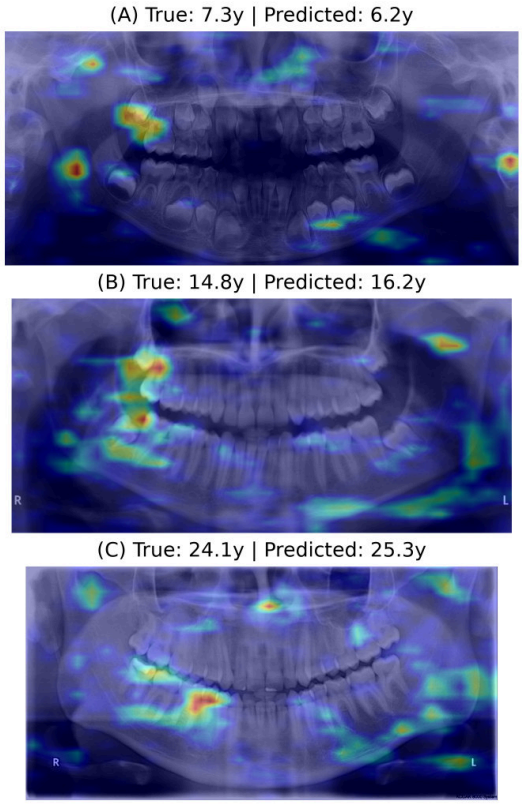


Fig. 3. Grad-CAM overlays produced by model 2216 for representative OPGs from three age bins: (A) 4–9 years, (B) 10–19 years, (C) 20–24 years. Each image was selected as the instance closest to the median absolute error within its bin. Warmer colours indicate regions of higher model attention.

using binary cross-entropy loss, AdamW optimization with learning rate of 1×10^{-4} and cosine annealing scheduling over a maximum of 12 epochs with early stopping (patience = 5). Training used 80% of the dataset and validation the remaining 20%. This constitutes a baseline experiment intended to demonstrate dataset usability rather than a reproduction of Bu et al. [22], whose dual-stream TResNet+CoAtNet-4 architecture and ROI preprocessing pipeline were not implemented. A faithful reproduction of their architecture is left for future work.

All three backbones were evaluated on accuracy, precision, recall, F1-score, and AUC. The results are shown in Table VII. ResNet50 achieved the highest F1-score (78.08%) and AUC of 0.841, surpassing EfficientNet-B0 (the second-best backbone) by 3.48 percentage points in F1. These results are substantially below Bu et al.'s reported accuracy of 86.79%, which is expected given the considerably smaller training set (1,253 versus 8,562 images) and simpler single-backbone architecture employed here.

Grad-CAMs generated by ResNet50 are shown in Fig. 4. Overall, activations are concentrated in the anterior dentoalveolar region of the mandible, with secondary activations in the posterior mandible and cranial regions. In panel A (correctly classified female), activation is anatomically compatible

TABLE VI

RESULTS OF APPLYING THE MODELS WEIGHTS PROVIDED BY [15] AND [16]. THE MEAN ABSOLUTE ERROR (MAE) GIVES THE MAGNITUDE OF THE DEVIATION BETWEEN THE ACTUAL AGE AND THE PREDICTED AGE. THE MAE WAS REPORTED ALONG WITH THE STANDARD DEVIATION (SD), STANDARD ERROR (SE), AND 95% CONFIDENCE INTERVAL (CI). THE SUCCESS RATE DESCRIBES THE PERCENTAGE OF OPGS IN WHICH THE AGE PREDICTION WAS CORRECT, WITH AN ACCEPTABLE ERROR RANGE OF $\pm$ 5, AND $\pm$ 10 YEARS.

		Absolute error (years)				Success rate (%)		
Model	**Age (years)**	**OPGs**	**MAE**	**SD**	**SE**	**95% CI**	**±5 years**	**±10 years**
912	4–9	554	0.72	0.94	0.04	0.64-0.80	99.6	99.6
	10–19	668	1.25	1.48	0.06	1.14-1.36	97.8	99.6
	20–24	143	1.40	1.40	0.12	1.17-1.63	98.6	99.3
	All	**1,566**	**1.04**	**1.29**	**0.03**	**0.97-1.10**	**98.6**	**99.6**
2216	4–9	554	1.38	1.33	0.06	1.27-1.49	98.9	99.6
	10–19	668	1.75	1.80	0.07	1.61-1.89	96.3	99.3
	20–24	143	1.83	2.00	0.17	1.51-2.16	97.2	98.6
	All	**1,566**	**1.61**	**1.64**	**0.04**	**1.53-1.69**	**97.4**	**99.4**

TABLE VII

SEX ESTIMATION RESULTS BY BACKBONE ON THE VALIDATION SET (20% STRATIFIED SPLIT).

Backbone-model	Precision (%)	Recall (%)	Accuracy (%)	F1 (%)	AUC (%)
EfficientNetB0 [23]	76.32	72.96	74.84	74.60	81.39
EfficientNetB5 [23]	73.03	69.81	71.66	71.38	81.17
ResNet50 [24]	74.71	81.76	76.75	78.08	84.10

with the right upper and lower molar region. In panel B (correctly classified female), activation concentrates on the right mandibular angle and the region corresponding to the lower third molar. In panels D and E (correctly classified males), the model consistently attends to the anterior region: panel D shows primary activation over the upper and lower incisors and the anterior mandibular and maxillary alveolar process, with secondary activation in the left lower premolar region; panel E shows activation over the upper and lower incisors and lower canines. This anterior dentoalveolar focus is consistent with the known sexual dimorphism of the anterior mandibular arch and alveolar morphology [22]. In panel C (misclassified female), activation concentrates on the upper and lower incisor region and the anterior mandibular alveolar process, as well as bilaterally in the lower canine and premolar region; the anatomically coherent but ambiguous pattern suggests morphological overlap between this individual's anterior arch and male-typical features. Panel F (misclassified male) shows activation in the left mandibular ramus and coronoid process, the right lower molar region, and areas compatible with the skull base, rather than the discriminative anterior dentoalveolar region; this failure to localise to the key features is consistent with the misclassification. Taken together, the Grad-CAM patterns reveal a more anatomically coherent and consistent representation than previously reported, with the anterior dentoalveolar region emerging as the primary discriminative area.

X. FUTURE WORK

Several directions for extending X-Pan and the baseline experiments presented here are identified.

Regarding the dataset, the most pressing limitation is the age imbalance: future collection cycles should prioritise adult images to improve coverage of the 20–89 year range. Expanding the dataset to include OPGs from multiple institutions and imaging devices would also improve generalizability and reduce the domain shift risk described in Section VI, but is among the most challenging advancements proposed.

For age estimation, the baseline experiments reported here applied pre-trained weights in a zero-shot setting. Fine-tuning, transfer learning, or training models directly on X-Pan are natural next steps.

For sex estimation, a faithful reproduction of the dual-stream TResNet+CoAtNet-4 architecture proposed by Bu et al. [22] is left for future work, as mentioned in Section IX.

Finally, detection of clinical interventions remains future work for X-Pan, contingent on targeted collection to increase the prevalence of pathologies, implants, and endodontic treatments (Section VI). Population-level analyses, already supported by the current composition, remain unexplored.

XI. HARDWARE AND TOOLS

All experiments were conducted on an NVIDIA GeForce RTX 3060 GPU under Linux (kernel 6.6.87.2-microsoft-standard-WSL2). Age estimation used TensorFlow 2.21.0 with CUDA 12.5.1 and cuDNN 9.0. Sex estimation used PyTorch 2.4.1 with CUDA 12.1 and cuDNN 9.1.0.

ACKNOWLEDGMENT

To the UTFPR Voluntary Scientific Initiation Program (Edital PROPPG 02/2025).

REFERENCES

[1] N. K. Agrawal, L. Hackman, and S. Dahal, "Dental age assessment using demirjian's eight teeth method and willems method in a tertiary hospital," *Journal of Nepal Medical Association*, vol. 56, no. 214, pp. 912–916, 2018.

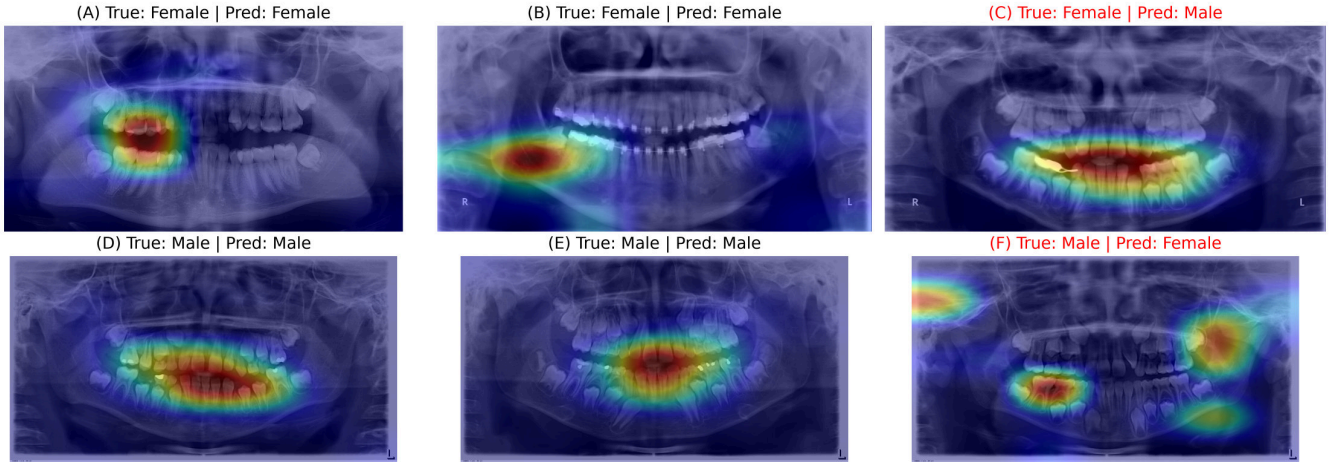


Fig. 4. Grad-CAM overlays generated by the ResNet50 backbone for three representative female (A–C) and three representative male (D–F) OPGs from the validation set. Each image was selected as the instance closest to the median absolute error within its sex group; panel C and panel F show misclassified examples (titles in red). Warmer colours indicate regions of higher model attention.

[2] I. Bedek, J. Dumančić, T. Lauc, M. Marušić, and I. Čuković Bagić, "New model for dental age estimation: Willems method applied on fewer than seven mandibular teeth," *International Journal of Legal Medicine*, 2019.

[3] C. B. R. B. Soares, J. N. Figueiroa, R. M. X. Dantas, L. M. Kurita, A. A. Pontual, and F. Ramos-Perez, "Evaluation of third molar development in the estimation of chronological age," *Forensic Science International*, 2015.

[4] F. Sharifonnasabi, N. Z. Jhanjhi, J. John, P. Obeidy, S. S. Band, H. Alinejad-Rokny, and M. Baz, "Hybrid hcnn-knn model enhances age estimation accuracy in orthopantomography," *Frontiers in Public Health*, vol. 10, p. 879418, 2022. [Online]. Available: https://doi.org/10.3389/fpubh.2022.879418

[5] R. Saric, J. Kevric, N. Hadziabdic, A. Osmanovic, M. Kadic, M. Saracevic, D. Jokic, and V. Rajs, "Dental age assessment based on cbct images using machine learning algorithms," *Forensic Science International*, vol. 334, p. 111245, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073822000755

[6] T. B. Ludermir, "Inteligência artificial e aprendizado de máquina: estado atual e tendências," *Estudos Avançados*, vol. 35, no. 101, 2021.

[7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, 2019.

[8] W. Anees, F. Riaz, and A. Franco, "Use of artificial intelligence in forensic dentistry: A review of literature," *Revista Brasileira de Odontologia Legal*, vol. 11, no. 2, 2024.

[9] M. Zaborowicz, K. Zaborowicz, B. Biedziak, and T. Garbowski, "Deep learning neural modelling as a precise method in the assessment of the chronological age of children and adolescents using tooth and bone parameters," *Sensors (Basel, Switzerland)*, vol. 22, no. 2, 2022.

[10] M. Xu, Y. Wu, Z. Xu, P. Ding, H. Bai, and X. Deng, "Robust automated teeth identification from dental radiographs using deep learning," *Journal of Dentistry*, vol. 136, p. 104607, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0300571223001938

[11] V. Patil, J. Saxena, R. Vineetha, R. Paul, D. K. Shetty, S. Sharma, K. Smriti, D. K. Singhal, and N. Naik, "Age assessment through root lengths of mandibular second and third permanent molars using machine learning and artificial neural networks," *Journal of Imaging*, vol. 9, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2313-433X/9/2/33

[12] A. Heinrich, F. Güttler, S. Wendt, S. Schenkl, M. Hubig, R. Wagner, G. Mall, and U. Teichgräber, "Forensic odontology: Automatic identification of persons comparing antemortem and postmortem panoramic radiographs using computer vision," *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 190, 06 2018.

[13] J. De Tobel, P. Radesh, D. Vandermeulen, and P. Thevissen, "An automated technique to stage lower third molar development on panoramic radiographs for age estimation: A pilot study," *The Journal of forensic odonto-stomatology*, vol. 2, 12 2017.

[14] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford, "Datasheets for Datasets v7," Link to publication 2020-08-06, 2018. [Online]. Available: https://arxiv.org/abs/1803.09010

[15] A. Heinrich, "Accelerating computer vision-based human identification through the integration of deep learning-based age estimation from 2 to 89 years," *Scientific Reports*, vol. 14, no. 1, p. 4195, Feb 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-54877-1

[16] R. M. Koch, H.-J. Mentzel, and A. Heinrich, "Deep learning for forensic age estimation using orthopantomograms in children, adolescents, and young adults," *European Radiology*, vol. 35, no. 7, pp. 4191–4202, Jul 2025. [Online]. Available: https://doi.org/10.1007/s00330-025-11373-y

[17] IBGE - Instituto Brasileiro de Geografia e Estatística, "Indicadores — panorama do censo demográfico 2022: Pirâmide etária (brasil)," Painel interativo, 2022, acessado em: 2026-06-26. [Online]. Available: https://censo2022.ibge.gov.br/panorama/indicadores.html?localidade=BR

[18] S. M. Mallya and E. W. N. Lam, Eds., *White and Pharoah's Oral Radiology: Principles and Interpretation*, 8th ed. St. Louis, MO: Elsevier, 2019.

[19] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, p. e1002683, nov 2018.

[20] E. H. P. Pooch, P. Ballester, and R. C. Barros, "Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification," in *Thoracic Image Analysis*, J. Petersen, R. San José Estépar, A. Schmidt-Richberg, S. Gerard, B. Lassen-Schmidt, C. Jacobs, R. Beichel, and K. Mori, Eds. Cham: Springer International Publishing, 2020, pp. 74–83.

[21] L. Yao, J. Prosky, B. Covington, and K. Lyman, "A strong baseline for domain adaptation and generalization in medical imaging," *CoRR*, vol. abs/1904.01638, 2019. [Online]. Available: http://arxiv.org/abs/1904.01638

[22] W. qing Bu, Y. xin Guo, D. Zhang, S. yi Du, M. qi Han, Z. xuan Wu, Y. Tang, T. Chen, Y. cheng Guo, and H. tian Meng, "Automatic sex estimation using deep convolutional neural network based on orthopantomogram images," *Forensic Science International*, vol. 348, p. 111704, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073823001548

[23] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

YOLO11 and YOLO26 Deep Learning Models for Diabetic Retinopathy Lesion Segmentation: A Comparative Analysis

1st Carlos V. G. Moura
Teleinformatics Engineering Department (DETI)
Federal University of Ceara (UFC)
Fortaleza, Brazil
carlos.victor@alu.ufc.br

2nd Paulo César Cortez
Teleinformatics Engineering Department (DETI)
Federal University of Ceara (UFC)
Fortaleza, Brazil
cortez@lesc.ufc.br

3rd Bruno R. S. Silva
Federal University of Ceara (UFC)
Crateús, Brazil
bruno@lesc.ufc.br

4th Vinícius M. Filgueiras
Teleinformatics Engineering Department (DETI)
Federal University of Ceara (UFC)
Fortaleza, Brazil
viniciusmffilgueiras@gmail.com

***Abstract*—Diabetic Retinopathy (DR) is an ocular disease caused by diabetes mellitus. In its later stages, DR can cause permanent damage to visual acuity, and an early diagnosis is essential to prevent vision loss. DR can be detected by analyzing a fundoscopy image, where specialists look for characteristic lesions in the retina, but ophthalmologists face several difficulties in the manual diagnosis of the disease. Computer-aided diagnosis assistance is a viable method to help experts overcome these difficulties. Therefor, this work evaluates and compares the YOLO11 and YOLO26 deep learning models in the task of segmenting DR lesions in fundoscopy images. The DDR dataset was used for training, validation and testing, and the IDRiD dataset was used as an external testing set to evaluate the models' generalization capabilities. Both dice score, intersection-over-union and inference time metrics were used to quantitatively analyze the results. The YOLO26 model obtained the best overall results, achieving 44.06% mean dice, 30.11% mean intersection-over-union and 49.88 ms mean inference time on the DDR dataset, as well as 51.78% dice, 42.66% intersection-over-union and 61.98 ms inference time on the IDRiD dataset. Microaneurysms were by far the hardest lesion for all models to segment. The models obtained higher dice and intersection-over-union metrics in the IDRiD dataset compared to DDR, even though IDRiD was used only for external testing, indicating that IDRiD is an easier dataset for segmentation an that the models exhibit good cross-dataset generalization capabilities. Thus, the proposed method has potential to assist ophthalmologists in DR screening and diagnosis, contributing as an accessible and efficient approach to DR screening.**

***Index Terms*—diabetic retinopathy, image segmentation, YOLO11, YOLO26, deep learning**

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 315447/2023-0 and the Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) under Grant No. BMD-0008-00514.01.40/26 for the support to this research.

I. Introduction

Diabetic Retinopathy (DR) is a set of retinal and vitreous changes caused by diabetes, which in its later stages can cause permanent damage to the patient's vision. In early stages, DR is asymptomatic and usually causes no noticeable visual changes. However, as the disease progresses, DR can lead to loss of sight. Moreover, the progression of DR indicates a higher risk of vascular diseases, such as strokes, kidney problems and cardiovascular complications [1].

Among the most clinically significant DR lesions are microaneurysms, hemorrhages, hard exudates and soft exudates. Microaneurysms (MAs) are the earliest clinically detectable signs of DR and appear as small, round, red dots. Hemorrhages (HEs) occur when damaged retinal blood vessels rupture and appear as dot or blot red spots. Hard Exudates (EXs) are protein deposits that accumulate in the retina and appear as yellowish-white spots with well-defined margins. Soft Exudates (SEs) are retinal lesions caused by accumulation of axoplasmic material and appear as white, fluffy spots [2].

Due to loss of vision often being a late symptom of DR, many patients remain undiagnosed even when the disease is already causing severe retinal damage. DR can be detected by analyzing a fundoscopy screening image, where the presence of characteristic lesions indicate the disease. Fundus imaging is an initial and commonly employed diagnostic method due to being non-invasive, cost-efficient and portable [3].

According to the International Diabetes Federation [4], in 2024, it is estimated that 589 million adults aged 20-79 were living with diabetes, and the total number of people living with diabetes is projected to reach 853 million by 2050. Moreover, in the same year, Brazil was the 6th country with the most adults with diabetes.

For these reasons, regular ophthalmologic monitoring of people with diabetes is essential. However, ophthalmologists face several difficulties in the manual diagnosis of DR, such as a large number of retinal fundus images to be analyzed, which can cause abnormalities to be misinterpreted due to fatigue [5]. Furthermore, morphological changes in the retina are often subtle and small, making the analysis process time-consuming, and may cause observer variability [6].

Computer-aided diagnosis assistance to an expert is a viable method to help ophthalmologists overcome these daily difficulties, making fundoscopy image analysis process faster and DR diagnostic more accurate. By providing an initial assessment of the progression of the disease, the computational system enables the ophthalmologist to spend more time analyzing more severe cases, as well as to give faster diagnostics.

You Only Look Once (YOLO) from Ultralytics is a deep learning model for fast computer vision which provides a framework for several computer vision tasks, such as classification, object detection, segmentation and others. The YOLO11 version of YOLO has shown promising results in medical image classification [7], and the YOLO26 version outperforms YOLO11 in benchmark tests [8]. Therefore, this work applies the YOLO11 and YOLO26 deep learning models for the task of segmenting DR lesions.

II. Related Works

Automatic DR lesion segmentation has attracted considerable attention due to its potential at assisting ophthalmologists in early diagnosis and monitoring. Unlike image level DR classification, lesion segmentation aims to localize pathological regions at the pixel level, providing greater interpretability and enabling quantitative assessment of lesion areas.

Herrero *et at.* [9] proposed a deep decoder-focused U-Net model for segmenting DR lesions. By allocating more layers to the decoder, the model is able to gradually restore spatial details lost during down-sampling, mitigating over-compression and and enhancing fine-grained feature preservation. With this method, training on both the DDR and IDRiD datasets, a mean dice score of 52.24% and a mean intersection-over-union of 36.96% was obtained on the DDR test set, as well as 69.93% and 57.31% on the IDRiD test set, respectively.

Zhang, Xu and Zhai [10] proposed a method based on Fusion DenseNet and U-Net. By replacing the encoder part of U-Net with DenseNet and introducing Omni-Dimensional Dynamic Convolution to replace traditional convolutions, the model's adaptability and boundary recognition capabilities are improved. With this method, training on both the DDR and IDRiD datasets, a mean dice score of 46.03% and a mean intersection-over-union of 30.41% was obtained on the DDR test set, as well as 65.68% and 50.48% on the IDRiD test set, respectively.

Sivaz and Aykut [11] proposed a mamba-based dual-decoder adversarial network. A dual-decoder structure ensures consistency between predictions from the full-resolution and downscaled images, and two adversarial discriminators are employed to reduce feature distribution mismatches and align segmentation predictions across different scales. With this method, training on both the DDR and IDRiD datasets, a mean intersection-over-union of 33.85% was obtained on the DDR test set, as well as 53.08% on the IDRiD test set.

To our best knowledge, there has been no work in the literature using YOLO11 or YOLO26 for DR lesion segmentation. However, studies that used YOLO11 and YOLO26 for other segmentation tasks, including medical imaging, obtained satisfactory results. Ikram [12] used YOLO11 to brain tumor segmentation in MRI images, obtaining a mean average precision 91%. Nizar [13] used YOLO26 for multi-class fetal head ultrasound segmentation, obtaining a mean dice score of 92.53%. Thus, YOLO11 and YOLO26 show great potential in image segmentation, and testing its capabilities in further undeveloped areas is of utmost importance.

Therefore, motivated by the yet unexplored potential of YOLO11 and YOLO26 for DR lesion segmentation, this work proposed a comparative analysis of YOLO11 and YOLO26 with the goal of identifying the most suitable model for a high-precision and cost-efficient automated screening system.

III. Materials and Methods

For this study, the dataset images were first split into training, validation and test sets. Among the 757 DDR images, 383 were used for training, 149 for validation and 225 for testing. This split is the standard recommended by DDR creators, to facilitate comparisons between different experiments. Moreover, all 81 IDRiD images were used for testing, to evaluate the generalization capabilities of the models. Then, the models were trained with the training and validation sets. After training, the models' predictions on the DDR and IDRiD test sets were taken, and the evaluation metrics were calculated. This work's overall workflow diagram is presented in Figure 1.

All the model training and evaluation process was executed on a Windows 10 computer with the following hardware: AMD Ryzen 5 7600 6-core processor, 16 GB of RAM, NVIDIA GeForce GTX 1650 (4 GB) graphics card and 447 GB SSD Kingston SA400S37480G storage device.

During the training, the standard data augmentation implemented in the Ultralytics framework was used, which includes several methods such as hue, saturation and brightness adjustments, geometric transformations, and random erasing. This way, the relatively low number of images and class imbalance could not severely impair model performance.

Ultralytics provides YOLO segmentation models pretrained in the ImageNet dataset. When performing further training on pretrained models, the backbone and neck of the models have their weight values fixed, such that only the head of the architecture uptades during the training. This way, the model can inherit the feature extraction capabilities of the previous training, saving computational costs. This work used the pretrained YOLO11 and YOLO26 provided by Ultralytics.

A. Datasets

Two datasets were used in this study, both widely used in the literature for DR lesion segmentation: The Diabetic

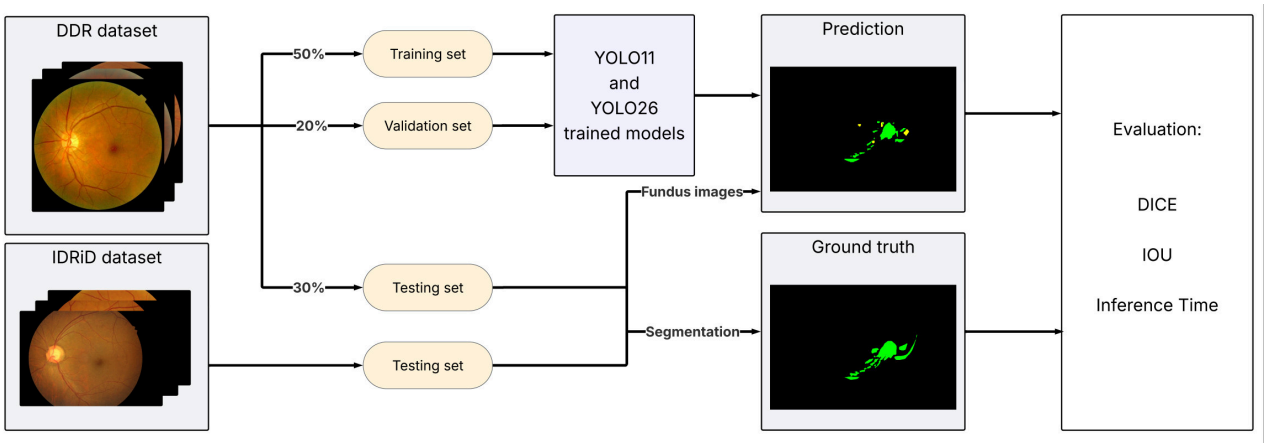


Fig. 1: Diagram of overall workflow process.

Retinopathy Database (DDR) for model training and testing and the Indian Diabetic Retinopathy Image Dataset (IDRiD) for external validation of the models.

1) DDR: The DDR [14] segmentation dataset contains 757 images obtained from hospitals in China. Image sizes range from 1088 $\times$ 1920 to 3456 $\times$ 5184 pixels. From the 757 total images, 601 contain HEs, 570 contain MAs, 486 contain SEs and 239 contain EXs. The original train, validation and test sets division proposed from the dataset creators was used in this study.

2) IDRiD: The IDRiD [15] dataset contains 81 images obtained from indian hospitals. All images are sized at 4288 $\times$ 2848 pixels. From the 81 total images, 80 images contain HEs, 40 SEs and all images contain MAs and EXs.

B. YOLO11

YOLO11 [16], released on september 10, 2024, is one of the most recent versions of the YOLO family of models, focused on optimizing performance and computational efficiency. One of the main innovations of this model is the C3k2 block, which performs several simultaneous convolutions that reduce the number of feature maps and concatenates the outputs, increasing the number of channels while the resolution remains constant, allowing for the capture of more complex features. YOLO11 also introduces the C2PSA block, which performs convolutions as well as employ an attention module and a fedd forward network, allowing the model to focus on the most relevant information of the image [18].

C. YOLO26

YOLO26 [17], released on january 14, 2026, is the current most recent version of the YOLO family of models. It introduces end-to-end inference, a lighter detection head, and task-specific heads for detection, segmentation, pose estimation, classification, and oriented detection. These changes make YOLO26 faster and more accurate than its predecessors, as well as more specialized in each of its available tasks.

D. Evaluation Metrics

To quantitatively analyze the experimental results, the DICE coefficient (DICE), intersection-over-union (IOU) and inference time (IT) performance metrics were adopted to evaluate the segmenting of DR lesions. The metrics of each prediction were calculated, and the mean and 95% confidence interval were taken for each model to analyze model performance. To calculate the metrics, the true positive (TP), false positive (FP) and false negative (FN) area of each lesion in each image was first calculated. The true positive refers to the intersection of the true lesion area and predicted lesion area, the false positive refers to the intersection of the true background area and predicted lesion area and the false negative refers to the intersection of the true lesion area and predicted background area. The relations of the DICE and IOU metrics to TP, FP and FN are shown in equations (1) and (2).

$$DICE = \frac{2TP}{2TP + FP + FN}, \tag{1}$$

$$IOU = \frac{TP}{TP + FP + FN}, \tag{2}$$

The IOU metric provides an intuitive assessment of the segmentation performance, as it directly compares the number of overlapping pixels between the ground truth and the predicted segmentation. The DICE metric, on the other hand, doubles the weight of true positives, effectively reducing the penalty of the model's FP and FN predictions.

IV. Results and Discussion

On the DDR dataset, YOLO26 consistently outperformed YOLO11 across all lesion categories. The mean DICE increased from 38.76% to 44.06%, an absolute improvement of 5.30%, while the mean IOU improved from 26.24% to 30.11%, an improvement of 3.87%. The models achieved the highest segmentation performance on EX lesions, with DICE scores of 49.83% for YOLO11 and 55.51% for YOLO26, followed by SEs, HEs and MAs. Among all lesion types,

TABLE I: Segmentation results on DDR

	YOLO11			YOLO26		
Lesion	DICE (%)	IOU (%)	IT (ms)	DICE (%)	IOU (%)	IT (ms)
Microaneurysm	21.66±17.42	14.55±15.83	-	24.52±14.67	18.50±16.21	-
Hemorrhage	37.24±20.15	25.83±18.74	-	44.18±12.92	28.35±14.56	-
Soft Exudate	46.31±18.68	30.46±21.37	-	52.03±17.43	35.17±14.85	-
Hard Exudate	49.83±22.61	34.12±19.29	-	55.51±14.74	38.42±15.93	-
Mean	38.76	26.24	114.04±32.18	44.06	30.11	49.88±23.72

TABLE II: Segmentation results on IDRiD

	YOLO11			YOLO26		
Lesion	DICE (%)	IOU (%)	IT (ms)	DICE (%)	IOU (%)	IT (ms)
Microaneurysm	27.93±12.37	22.90±11.82	-	35.73±9.64	27.38±12.51	-
Hemorrhage	43.18±13.25	38.41±10.47	-	51.12±12.73	43.27±8.91	-
Soft Exudate	50.65±9.76	42.57±12.34	-	58.74±11.28	48.15±14.62	-
Hard Exudate	54.76±12.83	46.32±15.57	-	61.53±7.45	51.84±10.36	-
Mean	44.13	37.55	120.34±40.42	51.78	42.66	61.98±16.64

MAs remained the most challeging to segment, obtaining the lowest DICE and IOU scores. In addition tot he accuracy improvementes, YOLO26 also substantially reduced the average inference time from 114.04 ms to 49.88 ms, more than halving the overall inference time. The segmentation results on the DDR test set are presented in Table I.

A similar trend was observed on the IDRiD external dataset, where YOLO26 achieved higher performance for every lesion class. The mean DICE increased from 44.13% to 51.78%, while the mean IOU increased from 37.55% to 42.66%, corresponding to absolute gains of 7.65% and 5.11% respectively. As in the DDR dataset, EXs exhibited the highest segmentation accuracy, reaching a DICE of 61.53%, whereas MAs remained the most difficult lesion to detect, with a DICE of 35.73%. YOLO26 also, again, demonstrated a considerably lower inference time, decreasing the average processing time from 120.34 ms to 61.98 ms, an improvement of 48.5%. The segmentation results on the IDRiD external test set are presented in Table II.

Although the models were trained exclusively on the DDR dataset and the IDRiD datset was used solely for external testing, the segmentation performance on IDRiD was consistently higher across all lesion categories. This behavior suggests that, despite serving as an unseen external dataset, IDRiD may present more favorable imaging characteristics for lesion segmentation, such as higher image quality, more consistent acquisition conditions, improved lesion visibility, or more homogeneous annotations. In contrast, the larger and more heterogeneous DDR dataset contains a wider range of lesion appearances, making it easier for the models to confuse lesions with one another, and making it a more challenging benchmark for pixel-level segmentation. For example, Figure 2 shows a case in which the original image from the DDR test set contained only an HE lesion, however, both models mistook part of the HE area for an SE lesion, and the YOLO11 model also segmented a small area as EX. These findings could also indicate that the trained models exhibit good cross-dataset generalization capabilities, since the learned parameters were able to produce good segmentations in very different scenarios, but further testing in more datasets and scenarios would be necessary to ascertain these capabilities. Moreover, these findings indicate that ophthalmologists segment DR lesions in standardized ways even in very different locations and clinical settings, since models only trained on DDR could also segment

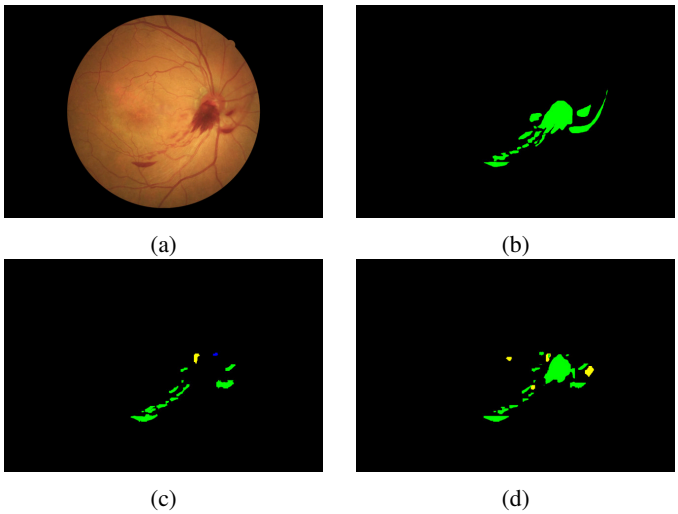


Fig. 2: Segmentation result sample on the DDR test set. The masks are presented using a lesion-specific color map: blue (EX), green (HE), and yellow (SE). (a) Original image. (b) Ground truth. (c) YOLO11. (d) YOLO26.

lesions in IDRiD.

Even though the IDRiD dataset was used solely for testing, both models obtained overall smaller 95% confidence intervals for all lesions in IDRiD compared to DDR, which further supports the possibility of IDRiD having more favorable imaging characteristics for consistent lesion segmentation. Moreover, MAs presented the relatively widest confidence interval in both datasets between all lesions, which indicates that, in addition to being the most challenging lesion to segment, MAs also present the most inconsistent segmentation results. This behavior could be likely related to the small size and low contrast of MAs compared with the more visually distinguishable exudative lesions. Furthermore, the confidence intervals of all lesions were considerably high, at a minimum of 8.91% YOLO26 HE IOU in IDRiD, which suggests that variability in image acquisition, lesion visibility or annotation method could be leading to inconsistent segmentations between images.

Overall, the experimental results show that YOLO26 provides consistent improvements over YOLO11 in both segmentation accuracy and computational efficiency. The largest performance gains were observed for HEs and exudates, while MAs continued to present the greatest segmentation challenge due to their small size and subtle appearance on the retina. YOLO26, compared to state-of-the-art methods, also obtained competitive results in the DDR dataset. Moreover, the relative drop in performance in the IDRiD dataset was comparatively small compared to state-of-the-art methods, even though other methods mostly used IDRiD for testing and training, instead of only for external testing. Furthermore, the low inference time achieved by YOLO26 indicates that the improved segmentation performance is accompanied by faster execution, making it more suitable for real-time or large-scale DR screening applications, as well as using the models on low computational power scenarios.

V. Conclusion

This study proposes a comparative analysis between YOLO11 and YOLO26 deep learning models in the task of segmenting diabetic retinopathy lesions in fundus images. The dice, intersection-over-union and inference time metrics were used to quantitatively evaluate the models. The DDR public dataset was used for training and validation, and the IDRiD dataset for external testing, allowing for an evaluation of the models' generalization capabilities.

Overall, the YOLO26 model obtained the best results in all lesions as well as in both datasets, achieving 44.06% mean dice, 30.11% mean intersection- over-union and 49.88 ms mean inference time on the DDR dataset, as well as 51.78% mean dice, 42.66% mean intersection-over-union and 61.98 ms mean inference time on the IDRiD dataset. Microaneurysms were by far the hardest lesion for all models to segment, such that the YOLO26 model obtained 24,52% dice and 18.50% intersection-over-union segmenting the lesion on the DDR dataset.

The results indicate that YOLO26 offers great potential in medical image processing applications, especially in large-scale and real-time scenarios, as it obtained competitive results to state-of-the-art methods. Moreover, both models demonstrated potential to have good generalization capabilities to unseen scenarios, but further research in different datasets and scenarios is necessary to ascertain these capabilities. Therefore, the presented method could help assist ophthalmologists in screening and diagnosing diabetic retinopathy by segmenting the characteristic lesions of the disease with high speed and moderate accuracy, especially in hardware-constrained scenarios.

Future works could use additional methods, such as more specialized data augmentation methods, to improve segmentation performance on difficult lesions, such as microaneurysms. Other datasets could also be used to further improve model generalization capabilities. Furthermore, the YOLO26 architecture could also be modified to improve feature extraction and learning capabilities on fundoscopy images and diabetic retinopathy lesions.

Acknowledgment

This work was made possible with the financial support of the Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) and with the support of the Laboratory of Computer Systems Engineering (LESC) - UFC.

References

[1] W. L. Alyoubi, M. F. Abulkhair, W. M. Shalash. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors, v. 21, n. 11, p. 3704, 2021.

[2] D. G. Cogan, D. Toussaint, T. Kuwabara. Retinal Vascular Patterns: IV. Diabetic Retinopathy. Arch Ophthalmol. 1961;66(3):366–378. doi:10.1001/archopht.1961.00960010368014

[3] K. R. Khaderi, K. A. Ahmed, J. L. Berry, L. T. Labriola, R. Cornwell. Retinal imaging modalities: advantages and limitations for clinical practice. Retinal Physician, 2011.

[4] International Diabetes Federation. IDF Diabetes Atlas. 11th ed. Brussels: International Diabetes Federation, 2025.

[5] J. Kaur, D. Mittal, R. Singla. Diabetic retinopathy diagnosis through computer-aided fundus image analysis: a review. Archives of Computational Methods in Engineering, 2021.

[6] M. R. K. Mookiah, U. R. Acharya, C. K. Chua, C. M. Lim, E. Y. K. Ng, A. Laude. Computer-aided diagnosis of diabetic retinopathy: a review. Computers in Biology and Medicine, v. 43, n. 12, p. 2136–2155, 2013.

[7] R. A. Diptho, S. Basak. Enhancing dermatological diagnosis through medical image analysis: how effective is YOLO11 compared to leading CNN models? NDT, v. 3, p. 11, 2025.

[8] G. Jocher, J. Qiu, A. Chaurasia. Ultralytics YOLO: version 8.0.0 [software]. Ultralytics, 2023. Available at: https://github.com/ultralytics/ultralytics. License: AGPL-3.0. Accessed on: 15 Jun. 2026.

[9] M. Herrero-Tudela, R. Romero-Oraá, G. C. Gutiérrez-Tobal, et al. DDFU-Net: A Deep Decoder-Focused U-Net Model for Retinal Lesion Segmentation. Annals of Biomedical Engineering, 2026. doi:10.1007/s10439-026-04032-w.

[10] Y. Zhang, Y. Xu, J. Zhai. Diabetic Retinopathy Image Segmentation Method Based on Fusion DenseNet and U-Net Network. Engineering Letters, v. 33, n. 2, 2025.

[11] O. Sivaz, M. Aykut. Mamba-based dual-decoder adversarial network for multi-scale diabetic retinopathy lesion segmentation. Neurocomputing, v. 686, p. 133736, 2026. doi:10.1016/j.neucom.2026.133736.

[12] C. Ikram. From detection to diagnosis: An advanced transfer learning pipeline using YOLO11 with morphological post-processing for brain tumor analysis for MRI images. Journal of Imaging, v. 11, n. 8, p. 282, 2025.

[13] M. F. Nizar. Domain-Guided YOLO26 with Composite BCE-Dice-Lovász Loss for Multi-Class Fetal Head Ultrasound Segmentation. arXiv preprint arXiv:2603.26755, 2026.
[14] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, H. Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences, v. 501, p. 511–522, 2019.
[15] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau. Indian Diabetic Retinopathy Image Dataset (IDRiD). IEEE Dataport, 2018.
[16] G. Jocher and J. Qiu. Ultralytics yolo11, 2024.
[17] G. Jocher, J. Qiu, M. Liu, S. Lyu, F. C. Akyon, M. E. Kalfaoglu. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. arXiv, 2026. doi:10.48550/arXiv.2606.03748. Available at: https://arxiv.org/abs/2606.03748.
[18] P. Hidayatullah, N. Syakrani, M. R. Sholahuddin, T. Gelar, R. Tubagus. YOLOv8 to YOLO11: a comprehensive architecture in-depth comparative review. arXiv, 2025.

Low-Cost YOLOv8-Based Detection of Hematological Elements in Canine and Feline Blood Smears

Maria Clara Alvarenga de Souza[#]
Institute of Science and Technology
Federal University of São Paulo
(UNIFESP)
São José dos Campos, Brazil
clara.alvarenga@unifesp.br

Joao Victor de Carvalho Rangel[#]
Institute of Science and Technology
Federal University of São Paulo
(UNIFESP)
São José dos Campos, Brazil
jrangel12@unifesp.br

Thiago Martini Pereira[*]
Institute of Science and Technology
Federal University of São Paulo
(UNIFESP)
São José dos Campos, Brazil
t.pereira@unifesp.br

Henrique Alves de Amorim
Institute of Science and Technology
Federal University of São Paulo
(UNIFESP)
São José dos Campos, Brazil
henrique.amorim@unifesp.br

Fabio Henrique Miguel Jardini[**]
University of the Paraiba Valley
UNIVAP
São José dos Campos, Brazil
fjardini@yahoo.com

Matheus Cardoso Moraes[**]
Institute of Science and Technology
Federal University of São Paulo
(UNIFESP)
São José dos Campos, Brazil
matheus.moraes@unifesp.br

*Corresponding author: t.pereira@unifesp.br ; [#]Contribute Equally; [**]Joint Co-lead

***Abstract*—**Microscopic examination of stained blood smears remains the reference method for detecting morphological abnormalities of leukocytes, erythrocytes, and platelets in veterinary hematology, but it is labor-intensive and expertise-dependent. This study evaluated the feasibility of a low-cost, YOLOv8-based deep learning framework for automated detection and classification of hematological elements in canine and feline blood smears acquired with conventional optical microscopy. A total of 8,855 hematological elements across thirteen classes were manually annotated and evaluated using blood-smear-level five-fold cross-validation. The model achieved a mean accuracy of 70.5%, mean specificity of 96.7%, and mean macro-averaged F1-score of 0.325. Performance was strongly class-dependent: segmented neutrophils (F1 = 0.914), eosinophils (F1 = 0.830), and lymphocytes (F1 = 0.762) were detected reliably, whereas band neutrophils, toxic neutrophils, and several rare classes performed poorly. Error analysis indicated that object-versus-background discrimination, rather than confusion between cell types, was the dominant source of error. These findings support the feasibility of conventional-microscopy-based AI screening as a triage aid for veterinary hematology, though current sensitivity is insufficient for autonomous diagnostic use.

***Index Terms*—**artificial intelligence, deep learning, YOLOv8, object detection, veterinary hematology, blood smear, canine, feline, class imbalance.

I. INTRODUCTION

Hematological examination is among the most frequently requested diagnostic procedures in veterinary medicine, playing a fundamental role in the evaluation of infectious, inflammatory, parasitic, immune-mediated, and hematopoietic disorders[1]. The complete blood count provides quantitative and qualitative information on erythrocytes, leukocytes, and platelets, and the interpretation of specific elements (segmented and band neutrophils, lymphocytes, eosinophils, monocytes, and platelet abnormalities) frequently informs clinical decision-making and therapeutic monitoring in canine and feline patients[1].

Although automated hematology analyzers have substantially improved laboratory efficiency, important limitations remain regarding morphological assessment. Impedance and flow-cytometry-based systems estimate cellular counts with high precision but are frequently unable to reliably identify subtle morphological abnormalities, atypical or reactive leukocyte populations, platelet aggregates, or hemoparasites [2]–[4]. Microscopic examination of stained blood smears therefore continues to be considered an essential, reference-standard component of veterinary hematological evaluation, despite being labor-intensive and dependent on operator expertise[5], [6]. Increased band-neutrophil counts and toxic morphological change are historically used as indicators of a left shift and disease severity, although both findings carry well-documented inter-observer and analytic variability even among trained personnel [2], [7], [8]; platelet aggregates and related sample-handling artifacts can likewise complicate manual and automated platelet assessment alike [3], [9].

Convolutional neural networks and one-stage object-detection architectures of the You Only Look Once (YOLO) family [10], [11] have achieved promising results for the detection and classification of human leukocytes, including hybrid YOLO-classification pipelines and vision-transformer combinations [12], single-shot detector and YOLO comparisons trained on thousands of annotated leukocytes [13], convolutional networks trained via transfer learning [14], and networks augmented with dual-attention mechanisms [15] or applied to immature-versus-mature neutrophil discrimination [16]. A global scoping review covering 136 primary studies on machine- and deep-learning-based white blood cell classification confirmed a clear shift toward convolutional architectures over the past decade, alongside a persistent dependency of model accuracy on dataset size and balance [17]. Almost all of this evidence base, however, concerns human hematology under comparatively curated imaging conditions.

In veterinary medicine, AI-assisted hematology remains comparatively unexplored. A recent multicenter evaluation of a commercial point-of-care platform integrating convolutional neural network algorithms for canine and feline blood smear screening reported leukocyte differential agreement above 90% with board-certified clinical pathologists, illustrating the potential clinical value of such systems under controlled, well-resourced conditions [18]. However, most veterinary clinics

and small diagnostic laboratories lack access to such proprietary platforms or dedicated digital pathology scanners, motivating interest in low-cost approaches that pair conventional optical microscopes with affordable digital cameras.

From a computer-vision standpoint, blood smear analysis combines challenges rarely encountered jointly in other detection domains: individual hematological elements occupy a very small fraction of total image area, the morphological boundary between adjacent classes (e.g., band versus segmented neutrophils) can be genuinely continuous rather than discrete, and class prevalence in a clinical sample is, by definition, highly imbalanced rather than curated for machine learning. Strategies such as targeted data augmentation and synthetic oversampling have shown promise for related small-sample, imbalanced cell-classification problems [19], but their effectiveness for multi-class hematological detection in non-curated veterinary samples remains unestablished.

Outside hematology, YOLO-based detectors have been applied across a markedly diverse range of medical-imaging domains, illustrating both the maturity and cross-specialty transferability of the architecture: polyp detection in colonoscopy [20], dermoscopic and smartphone-based skin/nail lesion analysis [21], recognition of anatomical landmarks and instruments in laparoscopic surgery [22], detection of diabetic-retinopathy-related fundus lesions [23], multi-scale detection of pediatric wrist fractures in radiographs [24], grading of rib fractures on computed tomography [25], and identification of helminth eggs in stool microscopy [26], the latter a particularly close analogue, given that it likewise applies YOLO-based detection to small, morphologically variable biological objects in conventional brightfield microscopy images. Despite this breadth of cross-domain adoption, a systematic application of YOLO-based detection to the simultaneous, multi-class identification of hematological elements directly from canine and feline blood smears has not been identified in the reviewed literature.

The aim of this study was therefore to evaluate the feasibility of a low-cost, AI-assisted hematological screening approach based on conventional microscopy and a YOLOv8 object detection framework [10], [11], trained and evaluated on canine and feline blood smear images containing thirteen classes of hematological elements, using blood-smear-level five-fold cross-validation to characterize its strengths and limitations under realistic, imbalanced clinical conditions.

II. MATERIALS AND METHODS

Fig. 1 summarizes the overall experimental workflow, comprising sample collection, blood smear preparation, conventional-microscopy image acquisition, expert annotation, and YOLOv8-based training and evaluation, each described in detail in the subsections below.

A. Ethics Statement

All blood samples were residual material obtained during routine diagnostic hematological procedures performed for clinical purposes, independent of and prior to this research; no animals were sedated, sampled, or otherwise manipulated specifically for this study.

B. Dataset Acquisition

Peripheral blood smears from canine and feline patients were obtained through a collaboration with Diagnovet, a veterinary diagnostic laboratory in São José dos Campos, Brazil; case selection and identification validation were supervised by Fábio H. M. Jardini, a veterinary clinical pathologist and the laboratory's technical director.

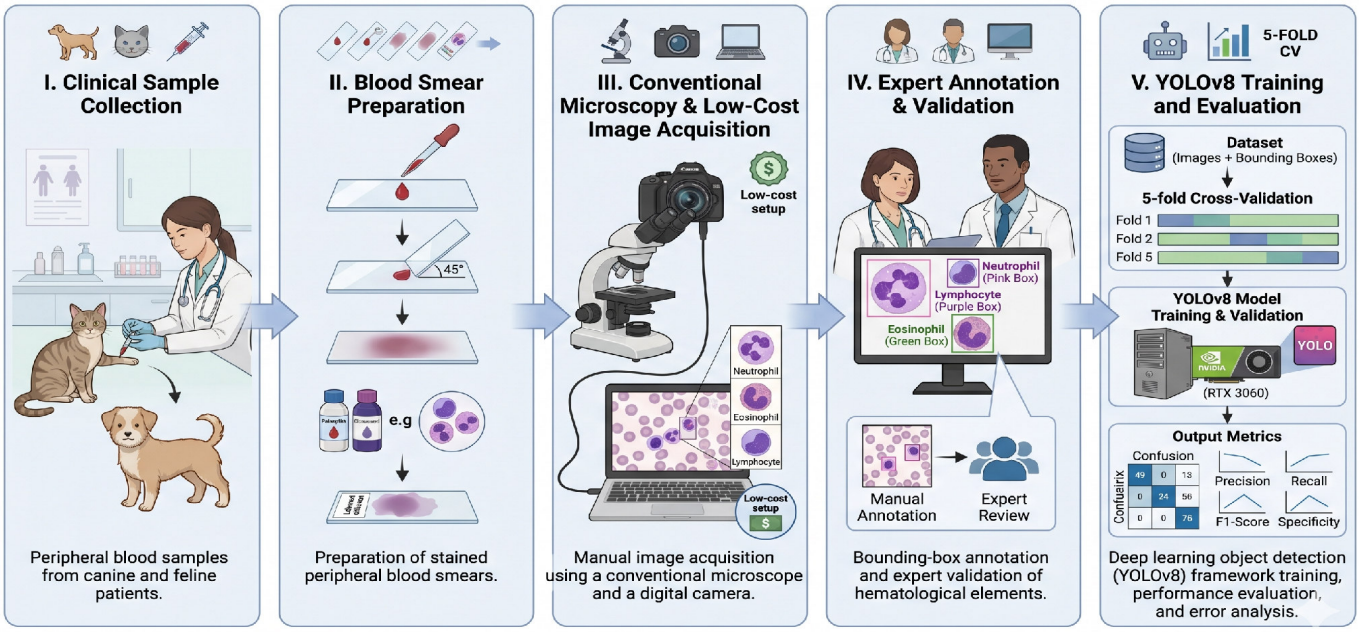


Fig. 1. Overview of the experimental workflow: peripheral blood smear sample collection, slide preparation, conventional optical microscopy and image acquisition, expert annotation of hematological elements, and YOLOv8-based model training and evaluation.

A total of 35 blood smear slides were analyzed, divided among 15 female dogs, 17 male dogs, 2 female cats, and 1 male cats. From a morphological and physiological standpoint, combining canine and feline samples within a single dataset is highly justified due to the structural analogy of their hematological elements. From a computer vision perspective, however, this deliberate pooling intentionally increases biological variability, presenting a more rigorous and realistic multi-species detection challenge for the model while simulating a diverse, low-cost clinical screening environment. Samples were processed using standard hematological protocols and stained with a panoptic (Romanowsky-type) protocol. Digital images were obtained by manual photography of stained smears under conventional optical microscopy: the responsible researcher visually scanned each slide and selectively photographed fields containing elements of interest, rather than exhaustively digitizing whole slides; reported class frequencies therefore describe the photographed fields rather than a uniform random sample of whole-slide cellularity. The images were obtained using a Canon Rebel T3i camera attached to a Motic BA210E microscope equipped with a 100x oil immersion objective and 10x eyepieces, yielding a total magnification of 1000x. Although the camera setup is capable of capturing 16-bit raw data, images were processed and evaluated as standard 8-bit JPEG files to maintain alignment with the computational restrictions of a low-cost diagnostic workflow.

C. Image Annotation

All images were manually annotated with bounding boxes by a veterinary researcher with clinical-pathology experience, with labels subsequently reviewed and validated by an experienced veterinary clinical pathologist. Thirteen classes were defined: segmented, band, and toxic neutrophils, eosinophils, lymphocytes, reactive lymphocytes, monocytes, an aggregate class, macroplatelets, Howell-Jolly-body-containing erythrocytes, Barr bodies, and background.

The final composed dataset comprised 8,855 annotated hematological elements (Table I), with marked natural class imbalance reflecting the prevalence of structures encountered during routine clinical examination. Segmented neutrophils were most prevalent (40.6%), followed by the aggregate class (16.0%, see flagged note in Section II-C), background (13.9%), and lymphocytes (11.0%). Three classes were represented by thirteen or fewer combined ground-truth instances (Howell-Jolly-body erythrocytes, macroplatelets); Barr bodies and toxic neutrophils, by contrast, had moderate instance counts (81 and 70, respectively), a distinction that becomes relevant in Section IV, since these two classes did not perform as their instance counts alone would predict.

D. YOLOv8 Architecture and Training

Object detection and instance segmentation were performed using the YOLOv8 architecture [10], [11], a one-stage framework that enables simultaneous localization and multi-class classification [10], [11]. This architecture was selected for its high computational efficiency and proven robustness across a diverse range of biomedical microscopy applications [12], [13], [15], [17], [20]–[26]. To evaluate the feasibility of the proposed approach under conditions realistically available to small veterinary diagnostic laboratories rather than dedicated research computing clusters [1], training was conducted on a deliberately modest, non-specialized workstation equipped with an Intel Core i7-4770 CPU (3.4–3.9 GHz, 8 threads), 32 GB DDR3 RAM, a 2 TB SSD, and a single NVIDIA GeForce RTX 3060 GPU with 12 GB GDDR6 [1].The pipeline was implemented using the default Ultralytics framework without architectural modifications, utilizing the nano variant for segmentation (YOLOv8n-seg) initialized with pretrained weights (yolov8n-seg.pt) [1], [11]. Training was embedded within a 5-fold cross-validation scheme and executed over 50 epochs per fold, using a batch size of 8 and an input image resolution of 640 x 640 pixels [1]. The optimization process was governed by the Ultralytics automatic configuration, which automatically selected the AdamW optimizer with a learning rate of 0.000625 and a momentum of 0.9 [1]. To optimize computational efficiency and prevent overfitting, automatic mixed precision (AMP) training was enabled, and early stopping was enforced with a patience threshold set to 10 epochs [1].

TABLE I. DISTRIBUTION OF ANNOTATED HEMATOLOGICAL ELEMENTS (N = 8,855)

Class	n	%
Segmented neutrophil	3,592	40.6
Aggregate *	1,416	16.0
Background	1,234	13.9
Lymphocyte	970	11.0
Eosinophil	597	6.7
Band neutrophil	339	3.8
Reactive lymphocyte	317	3.6
Monocyte	223	2.5
Barr body	81	0.9
Toxic neutrophil	70	0.8
Other rare classes†	13	0.15

**Biological identity pending confirmation (Sec. II-C). †Howell-Jolly-body erythrocytes (n=10), macroplatelets (n=2), abnormal erythrocytes (n=1), combined for brevity; class-wise totals differ from the nominal dataset total by <0.1% due to rounding.*

E. Cross-Validation and Evaluation Metrics

A five-fold cross-validation strategy, partitioned at the blood-smear level to prevent data leakage between training and validation partitions, was used: all elements from a given smear were assigned to a single fold (fold sizes: 1,854; 1,696; 2,133; 1,248; 1,924 elements). For each fold, precision, recall (sensitivity), specificity, and F1-score were computed per class from the resulting confusion matrix as:

Precision = TP/(TP+FP);

Recall = TP/(TP+FN);

Specificity = TN/(TN+FP);

F1 = 2 x (Precision x Recall)/(Precision+Recall),

where TP, FP, TN, and FN denote, respectively, true positives, false positives, true negatives, and false negatives for a given class. Macro-averaging across the thirteen classes was used to give equal weight to rare and frequent classes rather than allowing performance to be dominated by the majority class. Results are reported throughout as mean ± SD across the five folds; no formal inferential statistical testing across classes or folds was performed, and reported differences between classes or folds should accordingly be interpreted descriptively.

III. RESULTS

A. Dataset Composition

The resulting class-imbalance ratio showed in Table I (most-frequent class, segmented neutrophil, n = 3,592, versus least-frequent class) exceeds 3,500:1, substantially wider than the imbalance typically reported in curated human-hematology benchmarks and a likely contributor to the macro-averaged performance gap discussed in Section IV-A.

B. Cross-Validation Performance

Across folds, the model achieved mean accuracy 70.5% (range 64.3–73.9%), mean specificity 96.7% (range 96.4–96.9%), mean macro-averaged precision 31.4%, mean macro recall 34.2%, and mean macro F1 0.325 (range 0.281–0.361). The remarkably narrow range of macro specificity across folds indicates that the model consistently rejected non-target regions regardless of fold composition, whereas macro precision and recall, and therefore F1, were comparatively more sensitive to which blood smears were assigned to a given fold. Table II summarizes macro-level and representative class-wise performance; full results for every class are reported in the corresponding extended manuscript.

TABLE II. PERFORMANCE BY CLASS (MEAN ± SD ACROSS 5 FOLDS)

Class	Recall	Prec.	Spec.	F1
Overall (macro, 13 cls.)	.342±.038	.314±.024	.967±.002	.325±.029
Segmented neutrophil	**.923±.035**	**.906±.024**	**.951±.010**	**.914±.027**
Eosinophil	**.856±.023**	**.777±.097**	**.988±.004**	**.830±.062**
Lymphocyte	**.782±.108**	**.767±.112**	**.952±.019**	**.762±.042**
Aggregate	.571±.168	.436±.215	.937±.022	.468±.213
Monocyte	.474±.153	.449±.089	.975±.008	.455±.110
Reactive lymphocyte	.308±.107	.292±.055	.984±.002	.297±.081
Band neutrophil	.242±.353	.145±.226	.994±.005	.177±.277
Toxic neutrophil	.119±.108	.118±.091	.998±.001	.103±.067
Barr body	.000	.000	.998±.001	.000
Other rare classes†	.000	.000	.999–.000	.000

†Howell-Jolly body, macroplatelet, and abnormal erythrocyte classes, combined; "Overall" row reflects macro-averaging across all 13 originally annotated classes (Sec. III-A), not only the rows shown.

C. Error Analysis

The aggregated confusion matrices (Figs. 2-3) show that most classification errors involved object-versus-background discrimination rather than confusion among morphologically distinct cell types. Aggregates were correctly identified in only 56.9% of instances, with 42.4% misclassified as background and, conversely, 41.0% of true background misclassified as aggregates, the single largest error pathway in the matrix. Howell-Jolly-body-containing erythrocytes (n = 10) were misclassified as background in 100% of instances. Band neutrophils were correctly classified in only 4.4% of instances, with 81.7% misclassified as segmented neutrophils, mirroring the well-documented morphological continuum and inter-observer variability between these two forms [7], [8]. Toxic neutrophils were similarly unstable, correctly identified in only 8.6% of cases and otherwise distributed across background (21.4%), segmented neutrophils (18.6%), monocytes (17.1%), lymphocytes (11.4%), and eosinophils (10.0%); reactive lymphocytes were correctly classified in 30.9% of instances, most often confused with conventional lymphocytes (45.4%), a confusion that was symmetric (11.0% of true lymphocytes were misclassified as reactive lymphocytes). In contrast, segmented neutrophils and eosinophils showed minimal leakage to other classes (91.8% and 85.6% correctly classified, respectively), with residual error directed mostly toward background rather than toward another cellular class.

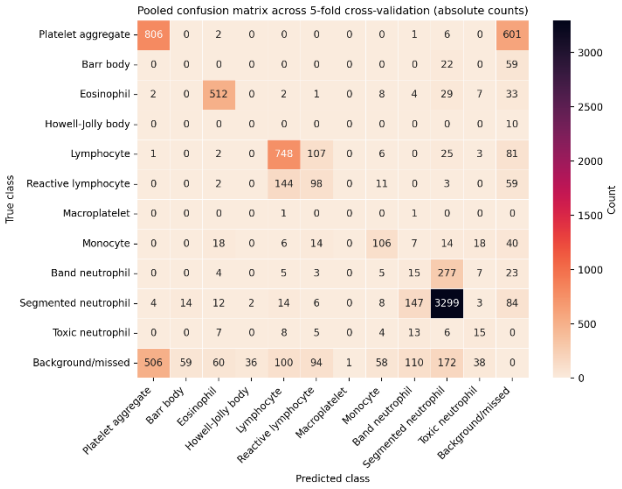


Fig. 2. Aggregated confusion matrix across the five cross-validation folds, absolute instance counts.

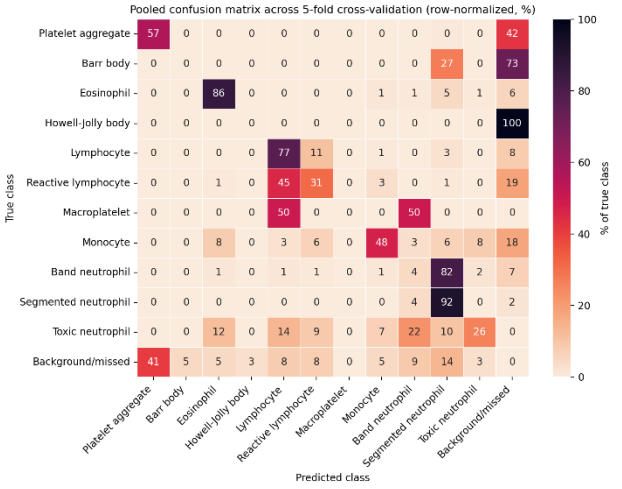


Fig. 3. Aggregated confusion matrix across the five cross-validation folds, row-normalized (each row sums to 100% of ground-truth instances of that class).

D. Variability and Qualitative Assessment

Class-level variability (Fig. 4) exceeded fold-level variability, indicating that class identity, rather than the specific smears assigned to a fold, was the dominant determinant of performance. Classes with very few instances showed essentially no dispersion because performance was near zero in every fold, whereas intermediate-frequency classes such as aggregates and band neutrophils showed the widest dispersion, consistent with their large standard deviations in Table II.

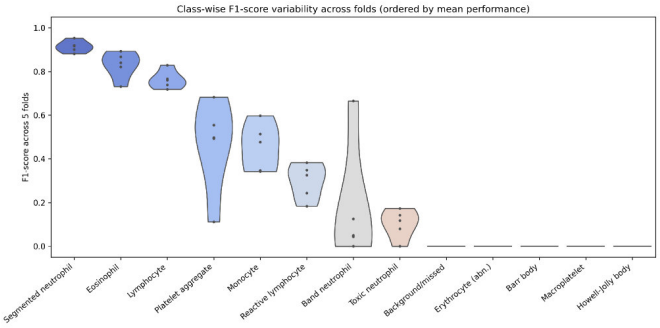


Fig. 4. Distribution of per-class of F1-score across the five cross-validation folds, illustrating heterogeneity of detection performance among the thirteen hematological classes.

A qualitative comparison between ground-truth annotations and model predictions on representative validation fields (Fig. 5) corroborated these quantitative patterns: dominant, morphologically distinct classes such as segmented neutrophils and eosinophils were consistently and accurately localized and classified, whereas aggregates and band neutrophils showed visible instances of missed detection or misclassification as segmented neutrophils, and the rarest classes were largely absent from the model's predictions even when present in the ground-truth annotation.

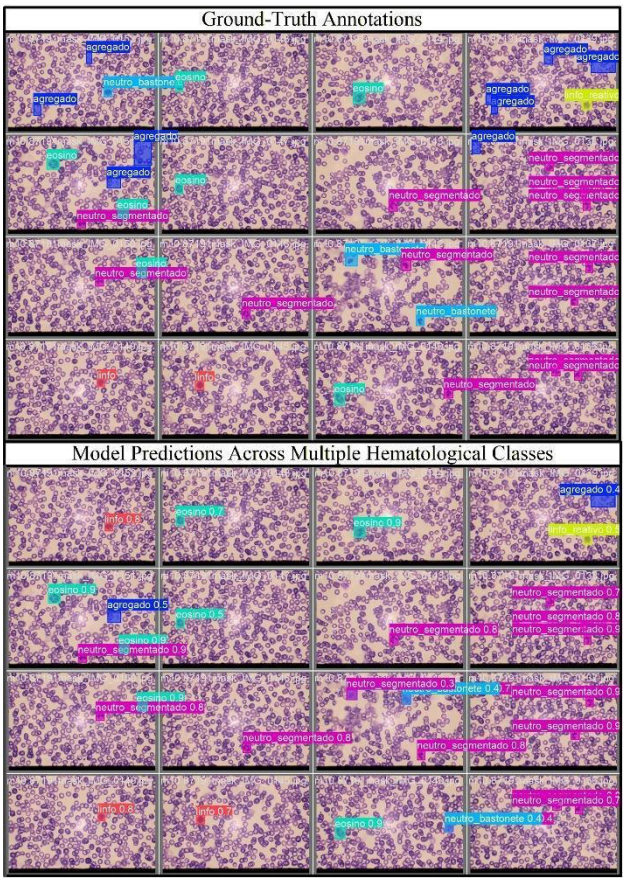


Fig. 5. Representative validation image fields comparing ground-truth annotations with model predictions across multiple hematological classes.

E. Synthesis of Quantitative Findings

Considered jointly, Tables I-II and Figs. 2-5 point to a consistent overall pattern: detection performance was strongly and predictably associated with the number of available training instances for a given class, with a secondary, independent contribution from the intrinsic morphological distinctiveness of that class. Segmented neutrophils and eosinophils combined high instance counts with comparatively distinctive morphology and accordingly achieved the highest and most stable F1-scores. Band and toxic neutrophils, despite moderate instance counts, achieved markedly lower performance than instance count alone would predict, reflecting their well-documented morphological continuity with segmented neutrophils. The aggregate class , despite being the second most frequent class overall, showed the largest variance in performance across folds, consistent with an amorphous and visually heterogeneous appearance. This pattern suggests that future efforts to improve performance should be differentially prioritized: increasing instance counts is likely to yield the greatest benefit for the rarest classes, whereas band/toxic neutrophils and the aggregate class may additionally require dedicated feature-engineering strategies targeting fine-grained morphological discrimination rather than instance count alone.

IV. DISCUSSION

A. Performance Patterns and Comparison with Prior Work

Strong, consistent performance for segmented neutrophils, eosinophils, and lymphocytes (together about 58% of annotated elements) likely reflects their relatively large size, distinctive staining characteristics, and low intra-class morphological variability, consistent with prior human-hematology reports [12], [13], [16]. The overall macro-averaged performance achieved here (F1 = 0.325) is considerably lower than the accuracies typically reported for human leukocyte detection, several exceeding F1 = 0.90 [12]–[16], [27]; this gap likely reflects a smaller, more class-imbalanced dataset (8,855 elements across thirteen classes, several with fewer than one hundred instances, versus datasets of several thousand to tens of thousands of relatively balanced images in many human-hematology studies), non-standardized conventional-microscopy acquisition rather than calibrated digital pathology scanners, and a broader taxonomy spanning leukocyte, platelet, and erythrocyte abnormalities plus an explicit background class. A global scoping review of 136 white-blood-cell-classification studies similarly concluded that accuracy in this domain remains strongly dataset-size- and balance-dependent [17], consistent with the present finding rather than representing an isolated negative result. The only directly comparable veterinary system, a commercial point-of-care CNN platform, reported >90% agreement with clinical pathologists [18], but was developed with substantially larger curated datasets and dedicated capture hardware integrated into the point-of-care device, unlike the explicitly low-cost, conventional-microscopy design objective adopted here.

The cross-domain YOLO literature surveyed in Section I [20]–[26], [28] offers a useful frame for interpreting these results. Several of these applications, including colonic polyp detection [20], helminth-egg identification in stool microscopy

[26], and pediatric wrist-fracture detection [24], likewise involve small, morphologically variable targets embedded in heterogeneous backgrounds, and generally report higher detection metrics than obtained here for the rarest hematological classes; in most cases, however, those studies were trained on datasets that were either explicitly curated for class balance or several-fold larger than the 8,855-element dataset used in the present study. This reinforces the broader conclusion of the scoping review by Asghar et al. [17]: across medical-imaging domains, YOLO-family detectors are architecturally capable of strong small-object performance, but realizing that capability for any individual class is conditional on adequate, reasonably balanced training data for that class-a condition that routine veterinary clinical sampling, by its nature, does not guarantee.

B. Sources of Error

A key finding was that the majority of misclassifications involved confusion between a true hematological element and background, rather than confusion between two distinct cell types: the aggregate-background error pathway alone accounted for the single largest source of error in the entire confusion matrix, and toxic neutrophils were likewise most frequently confused with background rather than with any single alternative leukocyte class. This suggests that the principal bottleneck lies in the object-detection (localization, object-versus-background discrimination) stage rather than in downstream classification of already-localized elements. Classes with ten or fewer combined instances (Howell-Jolly bodies, macroplatelets, abnormal erythrocytes) showed near-zero F1 in every fold, an expected consequence of insufficient training examples rather than fundamental model incapacity [19]. Band and toxic neutrophils, despite moderate instance counts (n = 339 and n = 70), also performed poorly, reflecting the genuinely continuous, subjective morphological boundary with segmented neutrophils that drives substantial inter-observer disagreement even among trained morphologists [7], [8]. Barr bodies (n = 81) likewise achieved near-zero F1 despite an instance count comparable to toxic neutrophils, a pattern that, unlike the truly rare classes above, is not fully explained by data scarcity alone and may instead reflect intrinsically low visual distinctiveness of this structure under conventional microscopy; this discrepancy was identified during internal stats-checking of the present draft and is highlighted here as a result that should be discussed explicitly rather than grouped with the data-scarcity explanation used for the other rare classes. If the “agregado” class is confirmed to represent platelet aggregates, it would constitute an inherently ambiguous detection target: amorphous, variable in density, and prone to blend with debris or out-of-focus regions, consistent with the well-documented phenomenon of in vitro platelet clumping that complicates both manual and automated platelet assessment [3], [9].

C. Clinical Implications

Taken together, these findings suggest that a low-cost, YOLOv8-based framework operating on conventional microscopy images can already provide clinically meaningful support for detecting the most common hematological elements (segmented neutrophils, eosinophils, and lymphocytes), which together represent the majority of diagnostically relevant decisions in routine canine and feline blood smear evaluation. However, the framework cannot presently be relied upon for independent detection of band neutrophils, toxic neutrophils, the aggregate class, or rare erythrocyte and platelet abnormalities, several of which carry direct clinical significance for recognizing inflammatory left shift, sepsis-associated toxic change, pseudothrombocytopenia, or regenerative anemia [2], [3], [7], [9]. Given the consistently high specificity observed across classes (96.4–96.9% macro-averaged), the framework is best positioned, at its current stage, as a triage aid flagging fields or smears for focused human review rather than as a stand-alone diagnostic substitute for microscopic evaluation by a trained veterinary professional.

In practical terms, a triage-oriented deployment could flag image fields or whole smears containing predicted segmented neutrophils, eosinophils, or lymphocytes for streamlined confirmation, while explicitly routing fields with predicted or suspected band/toxic neutrophils, aggregates, or rare erythrocyte/platelet findings for mandatory full manual review rather than automated acceptance or rejection. Such a tiered workflow would concentrate clinician attention on the classes where the present model is least reliable, rather than asking the model to make an autonomous call on findings most likely to be clinically decisive, such as toxic change or platelet aggregation.

D. Limitations and Future Directions

This study has several limitations that should be considered when interpreting its findings. First, the dataset was derived from approximately 50 blood smear slides obtained from a single veterinary diagnostic laboratory using a single conventional-microscopy acquisition setup; this limited and geographically concentrated sampling may restrict generalizability to other laboratories, staining protocols, microscope models, or camera systems, and external multi-center validation was not performed.

Second, several clinically relevant classes (Howell-Jolly bodies, macroplatelets, abnormal erythrocytes) were represented by ten or fewer ground-truth instances; the near-zero performance observed for these classes reflects insufficient training examples rather than fundamental model incapacity, and corresponding macro-averaged metrics should be interpreted with this caveat in mind. Barr bodies, with a moderate instance count (n = 81) but likewise near-zero F1, are a partial exception to this explanation, as discussed in Section IV-B.

Third, image annotation was performed by a single primary veterinary researcher with subsequent review by one experienced clinical pathologist; a fully independent multi-rater design with formal inter-observer agreement statistics was not implemented, which limits the ability to disentangle model error from possible annotation variability, particularly for morphologically ambiguous classes such as band and toxic neutrophils.

Fourth, several acquisition and training specifications flagged in Section II are not yet finalized in this draft; their absence does not affect the validity of the reported cross-

validation metrics, which were computed directly from model outputs, but does limit independent reproducibility until completed. Fifth, only descriptive statistics (mean ± SD across folds) were used; no formal inferential testing was performed, so claims of statistically significant differences between classes or folds cannot be made from this study alone. Finally, this study evaluated detection performance in isolation from downstream clinical workflows; the impact of integrating this framework into actual veterinary practice was not assessed.

Because the dominant error pathway was object-versus-background discrimination rather than inter-class confusion, improving detector sensitivity for small, low-contrast, or morphologically diffuse objects is likely to yield disproportionate gains relative to classification-only refinements. Targeted oversampling or synthetic augmentation for rare classes [19], explicit ordinal/hierarchical modeling of continuous morphological distinctions (e.g., band-to-segmented neutrophil, lymphocyte-to-reactive-lymphocyte), and prospective multi-laboratory validation with varied microscopy and staining equipment represent concrete next steps toward clinical deployment.

V. CONCLUSION

This study evaluated the feasibility of a low-cost, YOLOv8-based framework for automated detection and classification of thirteen classes of hematological elements in canine and feline blood smears acquired with conventional optical microscopy. Using blood-smear-level five-fold cross-validation on 8,855 annotated elements, the framework achieved mean accuracy 70.5% and mean specificity 96.7%, with reliable detection of the most clinically common cell types but substantially reduced sensitivity for band/toxic neutrophils, the aggregate class , Barr bodies, and rare erythrocyte/platelet abnormalities. Error analysis indicated that object-versus-background discrimination, rather than inter-class confusion, was the dominant source of error.

These findings support the feasibility of affordable, conventional-microscopy-based deep learning as a veterinary hematology screening aid, though current performance does not support fully autonomous diagnostic use. Given that most prior YOLO-based medical-imaging successes [20]–[26], [28] relied on larger or better-balanced datasets than the one used here, the present results should be read as a proof-of-concept lower bound rather than an upper bound on what this class of architecture can achieve in veterinary hematology. Future work should prioritize expanding underrepresented classes through multi-center data collection, refining small-object detection, and prospective clinical validation to determine the framework's practical value as a triage tool within real-world veterinary diagnostic workflows.

The main contributions of this study are threefold: (1) a quantitative feasibility assessment of YOLOv8-based detection across thirteen classes of canine and feline hematological elements using images acquired with conventional, non-specialized optical microscopy; (2) a blood-smear-level five-fold cross-validation protocol that avoids data leakage between training and evaluation partitions, providing a more conservative and clinically realistic performance estimate than image-level splitting; and (3) an explicit, class-by-class error analysis identifying object-versus-background discrimination-rather than inter-class confusion-as the dominant error source, which directly informs where future modeling effort should be concentrated.

ACKNOWLEDGMENT

This study was supported by the Brazilian Ministry of Health (MS) and the National Program for Genomics and Precision Health (Genomas Brasil); by the National Council for Scientific and Technological Development (CNPq), grants 406947/2023-6 and 443992/2023-1; and by the Coordination for the Improvement of Higher Education Personnel (CAPES). The authors declare no conflict of interest.

Data Availability: The annotated dataset and analysis code generated and/or analyzed during this study are not publicly available at the time of this draft but may be made available by the corresponding author upon reasonable request, subject to institutional approval.

During the preparation of this work, the authors used AI-assisted technologies solely for spelling and grammatical correction of the text, and to assist in the design of the illustrative methodological diagram (Figure 1). After using these tools, the authors reviewed, edited, and verified the final content, and take full responsibility for the academic integrity and accuracy of the scientific work.

REFERENCES

[1] N. Lotfalizadeh and S. Nazifi, "Clinical applications of blood cell ratios in veterinary hematology: From diagnosis to prognosis," Vet. Anim. Sci., vol. 32, art. 100622, 2026, doi: 10.1016/j.vas.2026.100622.

[2] N. I. Stacy, C. Hollinger, J. E. Arnold, C. Cray, H. Pendl, P. J. Nelson, and J. W. Harvey, "Left shift and toxic change in heterophils and neutrophils of non-mammalian vertebrates: A comparative review, image atlas, and practical considerations," Vet. Clin. Pathol., vol. 51, no. 1, pp. 18–44, 2022, doi: 10.1111/vcp.13117.

[3] H. E. Lunde, A. N. Hjelmtvedt, and E. K. Amundsen, "The diagnostic accuracy of Sysmex XN for identification of pseudothrombocytopenia using various thresholds for definition of platelet aggregation," Int. J. Lab. Hematol., vol. 44, no. 5, pp. 854–860, 2022, doi: 10.1111/ijlh.13920.

[4] P. Schuff-Werner, P. Kohlschein, A. Maroz, J. Linssen, K. Dreißiger, and C. Burstein, "Performance of the XN-2000 WPC channel-flagging to differentiate reactive and neoplastic leukocytosis," Clin. Chem. Lab. Med., vol. 54, no. 9, pp. 1503–1510, 2016, doi: 10.1515/cclm-2015-1105.

[5] H. Marahrens, F. Freise, F. Kiene, M. Ganter, and M. G. Wagener, "Results of digitised blood smear differentiations by veterinary students using item analysis," Sci. Rep., vol. 15, no. 1, art. 5947, 2025, doi: 10.1038/s41598-024-84881-4.

[6] A. Kratz, S.-H. Lee, G. Zini, J. A. Riedl, M. Hur, and S. Machin, "Digital morphology analyzers in hematology: ICSH review and recommendations," Int. J. Lab. Hematol., vol. 41, no. 4, pp. 437–447, 2019, doi: 10.1111/ijlh.13042.

[7] S. Paltrinieri and E. Talon, "Analytic variability in the enumeration of neutrophil subpopulations in canine blood," Vet. Clin. Pathol., vol. 46, no. 4, pp. 551–557, 2017, doi: 10.1111/vcp.12530.

[8] J. Langenstein, N. Bauer, and A. Moritz, "Extreme neutrophilia in cats – aetiology and prognosis," Tierarztl. Prax. Ausg. K Kleintiere Heimtiere, vol. 43, no. 5, pp. 323–330, 2015, doi: 10.15654/TPK-150134.

[9] A. M. Engelmann, P. A. Veleda, C. B. E. Mello, L. S. Dutra, T. R. Mann, A. Bueno, A. Krause, and C. M. de Andrade, "Amikacin prevents platelet aggregation in feline venous blood samples," Vet. Clin. Pathol., vol. 51, no. 1, pp. 51–56, 2022, doi: 10.1111/vcp.13074.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.

[11] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Ultralytics, version 8.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[12] S. A. Tarimo, M.-A. Jang, E. E. Ngasa, H. B. Shin, H. Shin, and J. Woo, "WBC YOLO-ViT: 2 way – 2 stage white blood cell detection and classification with a combination of YOLOv5 and vision transformer," Comput. Biol. Med., vol. 169, art. 107875, 2024, doi: 10.1016/j.compbiomed.2023.107875.

[13] Q. Wang, S. Bi, M. Sun, Y. Wang, D. Wang, and S. Yang, "Deep learning approach to peripheral leukocyte recognition," PLoS One, vol. 14, no. 6, art. e0218808, 2019, doi: 10.1371/journal.pone.0218808.

[14] A. I. Shahin, Y. Guo, K. M. Amin, and A. A. Sharawi, "White blood cells identification system based on convolutional deep neural learning networks," Comput. Methods Programs Biomed., vol. 168, pp. 69–80, 2017, doi: 10.1016/j.cmpb.2017.11.015.

[15] S. Khan, M. Sajjad, N. Abbas, J. Escorcia-Gutierrez, M. Gamarra, and K. Muhammad, "Efficient leukocytes detection and classification in microscopic blood images using convolutional neural network coupled with a dual attention network," Comput. Biol. Med., vol. 174, art. 108146, 2024, doi: 10.1016/j.compbiomed.2024.108146.

[16] T.-R. Tseng and H.-M. Huang, "Classification of peripheral blood neutrophils using deep learning," Cytometry A, vol. 103, no. 4, pp. 295–303, 2022, doi: 10.1002/cyto.a.24698.

[17] R. Asghar, S. Kumar, A. Shaukat, and P. Hynds, "Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review," PLoS One, vol. 19, no. 6, art. e0292026, 2024, doi: 10.1371/journal.pone.0292026.

[18] E. Morissette, C. D. Penn, R. A. Hall Sedlak, A. J. Rhodes, D. S. Tippetts, M. Loenser, and R. Goldstein, "Point-of-care platform integrated with deep-learning, convolutional neural network algorithms effectively evaluates canine and feline peripheral blood smears," Am. J. Vet. Res., pp. 1–11, 2024, doi: 10.2460/ajvr.24.08.0226.

[19] S. Yu, S. Zhang, B. Wang, H. Dun, L. Xu, X. Huang, E. Shi, and X. Feng, "Generative adversarial network based data augmentation to improve cervical cell classification model," Math. Biosci. Eng., vol. 18, no. 2, pp. 1740–1752, 2021, doi: 10.3934/mbe.2021090.

[20] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, "An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets," Comput. Biol. Med., vol. 141, art. 105031, 2021, doi: 10.1016/j.compbiomed.2021.105031.

[21] W. Mio et al., "Establishment of an intelligent analysis system for clinical image features of melanonychia based on deep learning image segmentation," Comput. Med. Imaging Graph., vol. 123, art. 102543, 2025, doi: 10.1016/j.compmedimag.2025.102543.

[22] A. Zygomalas, D. Kalles, N. Katsiakis, A. Anastasopoulos, and G. Skroubis, "Artificial intelligence assisted recognition of anatomical landmarks and laparoscopic instruments in transabdominal preperitoneal inguinal hernia repair," Surg. Innov., vol. 31, no. 2, pp. 178–184, 2024, doi: 10.1177/15533506241226502.

[23] A. R. Wahab Sait, "A lightweight diabetic retinopathy detection model using a deep-learning technique," Diagnostics (Basel), vol. 13, no. 19, art. 3120, 2023, doi: 10.3390/diagnostics13193120.

[24] M. Li et al., "MFSD-YOLO: A multi-scale feature detection network for pediatric wrist abnormalities in radiographic images," PLoS One, vol. 21, no. 2, art. e0340408, 2026, doi: 10.1371/journal.pone.0340408.

[25] T. Li, M. Liao, Y. Fu, F. Zhang, L. Shen, J. Che, S. Wu, J. Liu, W. Wu, P. He, Q. Xu, and Y. Wu, "Intelligent detection and grading diagnosis of fresh rib fractures based on deep learning," BMC Med. Imaging, vol. 25, no. 1, art. 98, 2025, doi: 10.1186/s12880-025-01641-0.

[26] M. A. Mohammed, E. T. Anzaku, P. K. Ward, B. Levecke, J. Krishnamoorthy, W. De Neve, and S. Van Hoecke, "Detecting soil-transmitted helminth and Schistosoma mansoni eggs in Kato-Katz stool smear microscopy images: A comprehensive in- and out-of-distribution evaluation of YOLOv7 variants," PLoS Negl. Trop. Dis., vol. 19, no. 7, art. e0013234, 2025, doi: 10.1371/journal.pntd.0013234.

[27] S. Choudhary et al., "Advancing blood cell detection and classification: performance evaluation of modern deep learning models," BMC Med. Inform. Decis. Mak., vol. 25, no. 1, art. 207, 2025, doi: 10.1186/s12911-025-03027-2.

[28] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," Comput. Biol. Med., vol. 121, art. 103792, 2020, doi: 10.1016/j.compbiomed.2020.103792.

Evolução do Processo de Desenvolvimento de Rodas para Cadeiras de Rodas Infantis via Impressão 3D

1st Brenda Gomes Oliveira
Laboratório de Estudos e Inventividades em Tecnologia Assistiva (Lab. E.I.T.A.)
Universidade Federal de Goiás(UFG)
Goiânia, Brasil
brendabrenda@discente.ufg.br

2nd Ana Carolina Ramos Nunes
Laboratório de Estudos e Inventividades em Tecnologia Assistiva (Lab. E.I.T.A.)
Universidade Federal de Goiás(UFG)
Goiânia, Brasil
anaramos2@discente.ufg.br

3rd Pedro Henrique Gonçalves
Laboratório de Estudos e Inventividades em Tecnologia Assistiva (Lab. E.I.T.A.)
Universidade Federal de Goiás(UFG)
Goiânia, Brasil
pedrogoncalves@ufg.br

***Resumo*—A tecnologia assistiva desempenha um papel fundamental na promoção da mobilidade, da autonomia e da qualidade de vida das pessoas com deficiência, sendo as cadeiras de rodas um dos principais recursos utilizados para esse fim. Entretanto, o elevado custo de componentes comerciais e a necessidade de adaptação às demandas específicas dos usuários representam desafios para sua viabilidade. Nesse contexto, o presente trabalho teve como objetivo desenvolver e aperfeiçoar uma roda para cadeira de rodas por meio da tecnologia de impressão 3D, visando à obtenção de uma solução funcional e de baixo custo. Inicialmente, foram avaliadas rodas comerciais de etileno-vinil-acetato (EVA) e rodas pneumáticas com câmara de ar, cujas limitações econômicas e de manutenção motivaram o desenvolvimento de um modelo próprio. O processo de prototipagem envolveu sucessivas modificações estruturais, passando de modelos segmentados em oito e quatro partes até a obtenção de uma roda em peça única, além da implementação de um aro de propulsão para maior conforto e ergonomia. Para a fabricação dos protótipos, foram utilizados filamentos de poliuretano termoplástico (TPU), acrilonitrila butadieno-estireno (ABS) e polietileno tereftalato glicol (PETG) reforçado com fibra de carbono, selecionados em função de suas propriedades mecânicas e funcionais. Os resultados demonstram a viabilidade técnica de um processo iterativo de desenvolvimento baseado em manufatura aditiva para a fabricação de rodas personalizadas de baixo custo, evidenciando o potencial dessa abordagem para reduzir a dependência de componentes comerciais e ampliar a flexibilidade no projeto de cadeiras de rodas infantis.**

***Palavras-chave*—tecnologia assistiva; cadeira de rodas; rodas; impressão 3D.**

I. INTRODUÇÃO

De acordo com a Organização Mundial da Saúde (OMS), a tecnologia assistiva engloba produtos, sistemas e serviços destinados a manter ou melhorar a funcionalidade e a independência dos indivíduos. Esses recursos, como cadeiras de rodas, desempenham um papel fundamental na promoção da autonomia, da inclusão social e da qualidade de vida, especialmente para pessoas com deficiência. Apesar dos benefícios reconhecidos, a OMS destaca que o acesso aos produtos assistivos ainda é limitado em diversas regiões, evidenciando a necessidade de ampliar sua disponibilidade e adoção em larga escala [1].

As cadeiras de rodas desempenham um papel essencial na reabilitação, promovendo a mobilidade, a autonomia e a participação social dos usuários. Além disso, a capacidade de utilizá-las está diretamente relacionada à inclusão e à melhoria da qualidade de vida [2].

Nesse contexto, o Laboratório xx, da Universidade YY, atua no desenvolvimento de soluções inovadoras em tecnologia assistiva, buscando criar dispositivos que atendam às necessidades específicas dos usuários por meio da integração entre pesquisa, ensino e extensão. O laboratório tem como missão desenvolver recursos acessíveis que promovam funcionalidade, autonomia e inclusão, aproximando o conhecimento científico das demandas reais da comunidade.

Entre os projetos desenvolvidos pelo laboratório, destaca-se a construção de cadeiras de rodas infantis destinadas a crianças de um a cinco anos de idade. Inicialmente, a estrutura desse equipamento era confeccionada em MDF (placa de fibra de média densidade); posteriormente, o projeto evoluiu para a utilização de chapas de tampinhas plásticas recicladas, proporcionando maior resistência, durabilidade e segurança ao equipamento. A proposta inicial consistia na utilização de rodas comerciais para a fabricação de bicicletas infantis, adaptando-as para a aplicação nas cadeiras de rodas. Contudo, durante o

desenvolvimento, observou-se que essa alternativa apresenta limitações relacionadas ao desempenho, à adaptação e ao custo, motivando a busca por uma solução própria.

Nesse contexto, a impressão 3D possibilita o desenvolvimento de dispositivos personalizados, adequados às necessidades individuais dos usuários de menor custo [3]. Dessa forma, iniciou-se o desenvolvimento de uma roda específica para essas cadeiras de mobilidade infantis utilizando manufatura aditiva por impressão 3D, visando obter um componente funcional, resistente, de baixo custo e passível de personalização conforme as necessidades dos usuários.

Embora a manufatura aditiva esteja sendo amplamente empregada no desenvolvimento de dispositivos de tecnologia assistiva, ainda não há estudos que descrevam processos sistemáticos para o desenvolvimento de rodas destinadas especificamente a cadeiras de rodas infantis. Considerando que esse componente influencia diretamente a mobilidade, a estabilidade, a ergonomia e o desempenho global do equipamento, sua otimização representa um desafio relevante para a engenharia aplicada à tecnologia assistiva.

Este trabalho tem como objetivo desenvolver um processo iterativo de projeto para rodas personalizadas destinadas a cadeiras de rodas infantis utilizando manufatura aditiva, descrevendo as sucessivas modificações estruturais implementadas durante o desenvolvimento e sua contribuição para o aprimoramento funcional do componente.

II. METODOLOGIA

Inicialmente, foram adquiridas rodas de etileno-vinil-acetato (EVA) comercialmente disponíveis, com aro 12, como apresentado na figura 1, utilizadas nas cadeiras de rodas produzidas pelo laboratório. Entretanto, verificou-se a necessidade de rodas com aro 14, ilustrado na figura 2, para atender adequadamente crianças maiores. O custo aproximado deste novo modelo, cerca de R$120,00 por unidade, tornava a aquisição economicamente inviável.



Fig. 1. Roda de EVA comercialmente disponível.

Foram testadas rodas pneumáticas de borracha com câmara de ar, compatíveis com o aro 14, conforme a imagem 3 mostra o pneu. No entanto, além do custo adicional referente às câmaras de ar (aproximadamente R$15,00 por unidade), a necessidade de manutenção periódica decorrente da perda de pressão comprometeu sua viabilidade de uso.



Fig. 2. Roda de Aro 14.



Fig. 3. Pneu de borracha, comercialmente disponível.

Diante dessas limitações, o desenvolvimento foi conduzido segundo uma abordagem iterativa de engenharia, na qual cada protótipo foi analisado quanto às limitações observadas durante sua utilização. Os resultados dessa avaliação orientaram sucessivos ciclos de redesenho, seleção de materiais e aperfeiçoamento estrutural, até a obtenção da configuração final, expressa pela figura 4. Em um primeiro momento, devido às limitações dimensionais da impressora utilizada (Creality K1 Max), o aro da roda foi projetado em quatro segmentos interligados, como observado na figura 5, produzidos em filamento polietileno tereftalato glicol (PETG) reforçado com fibra de carbono (PETG+CF). Embora funcional, essa configuração segmentada apresentava complexidade de montagem e compromete a integridade estrutural da peça.

A divisão do aro em segmentos representou uma solução transitória imposta pelas limitações dimensionais da impressora disponível naquele momento. Entretanto, a presença de múltiplas interfaces de união aumentava a complexidade da montagem e introduziu potenciais pontos de concentração de tensões mecânicas.

Com a aquisição da Creality Ender 5 Max, tornou possível a impressão do aro em peça única, como demonstra a figura 6, eliminando as junções entre segmentos e conferindo maior robustez ao componente.



Fig. 4. Modelo 3D da roda completa.

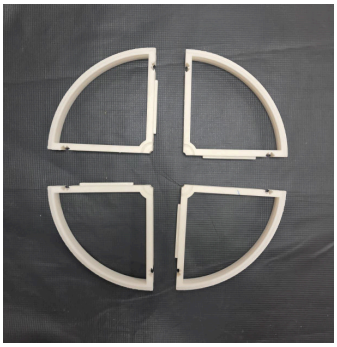


Fig. 5. Aro de impressão 3D dividida em 4 partes. Posteriormente, com a aquisição da impressora 3D.

O pneu foi projetado em oito segmentos independentes, indicado pela figura 7, acoplados posteriormente à estrutura da roda e impressos a partir do filamento de poliuretano termoplástico (TPU), material selecionado por sua elevada flexibilidade e capacidade de absorção de impactos. Adicionalmente, foi incorporado um aro de propulsão, também impresso em PETG reforçado com fibra de carbono, destinado a facilitar o manuseio da cadeira e proporcionar maior conforto ao usuário. A roda completa é resultado da montagem integrada desses três componentes produzidos separadamente, como ilustrado na figura 8.

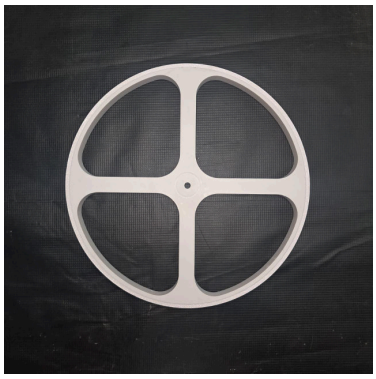


Fig. 6. Aro de impressão 3D peça única.



Fig. 7. Segmento de pneu sendo impresso. Pneu composto por oito partes.



Fig. 8. Roda completa, com aro de quatro partes, pneu e aro de propulsão

A. Materiais Utilizados

As rodas constituem um dos principais componentes das cadeiras de rodas, influenciando diretamente a mobilidade, a estabilidade, a manobrabilidade e o conforto dos usuários. Dessa forma, características como dimensões, materiais e configuração podem impactar significativamente o desempenho funcional da cadeira de rodas e, consequentemente, a independência e a participação social dos indivíduos [4].

O acrilonitrila butadieno-estireno (ABS) foi utilizado inicialmente na impressão dos aros devido à sua boa resistência ao impacto, rigidez e estabilidade dimensional. Além disso, apresenta boa processabilidade e durabilidade, características que favorecem a fabricação de componentes estruturais submetidos a esforços mecânicos. [6] Para a fabricação dos protótipos utilizou-se o filamento Hyper ABS na cor branca, com diâmetro de 1,75 mm e massa líquida de 1,0 kg. O processo de impressão foi realizado em temperaturas variando entre 230 °C e 270 °C, conforme as especificações do fabricante.

O PETG reforçado com fibras de carbono (PETG+CF) teve sua utilização concentrada nos componentes estruturais da roda devido à necessidade de elevada rigidez e estabilidade dimensional, características importantes para suportar os esforços mecânicos gerados durante o deslocamento da cadeira e, posteriormente, como substituto do ABS, dada a sua elevada resistência mecânica, maior rigidez e boa estabilidade térmica. Além disso, mantém boa

resistência ao impacto, tornando-se adequado para a fabricação de componentes submetidos a cargas e esforços mecânicos [7]. Dessa forma, na fabricação dos componentes da roda utilizou-se filamento PETG reforçado com fibra de carbono (PETG+CF), da marca Apex 3D (Carbon), na cor branca. O material apresenta diâmetro de 1,75 mm e foi processado em temperaturas de extrusão entre 260 °C e 300 °C, com temperatura da mesa aquecida variando entre 40 °C e 90 °C. O filamento foi fornecido em bobinas com massa líquida de 1 kg.

O poliuretano termoplástico (TPU) foi empregado exclusivamente na região correspondente ao pneu, explorando suas propriedades elastoméricas para promover absorção de impactos, aumento da aderência e maior conforto durante a movimentação, resistência ao desgaste e capacidade além de resistência ao desgaste. Além disso, apresenta alta resistência mecânica e boa durabilidade, características que favorecem sua aplicação em componentes sujeitos a cargas e atritos, como rodas e elementos de amortecimento [5]. Foi utilizado o filamento Rapid TPU 95A, da marca ELEGOO, na cor preta, com diâmetro de 1,75 mm e massa líquida de 1,0 kg. O material foi processado a uma temperatura de impressão de 200 °C a 230 °C, sendo recomendada a secagem prévia a 70 °C por 6 horas, conforme especificações do fabricante.

B. Aspectos éticos

A pesquisa foi submetida e aprovada pelo Comitê de Ética em Pesquisa pela Universidade Federal de Goiás, sob o parecer nº7.688.714, respeitando os princípios éticos aplicáveis às pesquisas envolvendo seres humanos.

III. RESULTADOS

O processo de desenvolvimento resultou em uma evolução progressiva das rodas infantis produzidas no laboratório, passando de soluções comerciais para um produto produzido internamente por meio da manufatura aditiva, à medida que surgiam limitações a cada alternativa avaliada.

A adoção da manufatura aditiva permitiu superar as limitações observadas nas rodas comerciais. A primeira versão, segmentada em quatro partes, embora funcional, apresentava maior complexidade na montagem e limitações estruturais. Além da obtenção do aro em peça única, a substituição do filamento ABS pelo PETG+CF também contribuiu para a resistência estrutural ao componente.

O aro em peça única consome 180g de filamento, dos quais 30g correspondem ao material de suporte, que são removidos após a impressão. Os oito segmentos de pneu consomem 24 g cada, de modo que a peça completa resulta em 192g. Por fim, o aro de propulsão consome aproximadamente 204g (sendo 15g de material de suporte). Portanto, a roda final completa utiliza, em média, 576g de filamento.

O tempo total de impressão de uma roda completa é de aproximadamente 18 horas e 40 minutos, distribuídas entre seus três componentes: a roda demanda 8 horas, o conjunto do pneu (oito segmentos), 6 horas e o aro de propulsão, 4 horas e 40 minutos. Os custos estimados de produção por roda, considerando o preço de mercado dos filamentos utilizados e a quantidade de material consumido, estão detalhados na Tabela 1.

TABELA I
Estimativa de custo de produção.

Estrutura	Filamento utilizado	Valor do filamento (R$/Kg)	Quantidade (g)	Custo estimado (R$)
Aro da roda	PETG+CF	189,00	180 g	28,35
Pneu	TPU	169,00	192 g	32,45
Aro de propulsão	PETG+CF	189,00	204 g	38,55
TOTAL	-	-	576 g	105,02

Observa-se que aproximadamente 60% do custo total se concentra nos componentes estruturais produzidos em PETG reforçado com fibra de carbono, enquanto o pneu em TPU representa cerca de um terço do custo final. Essa distribuição evidencia que futuras otimizações do projeto poderão se concentrar principalmente na redução do consumo de material estrutural.

Além da vantagem econômica, a solução desenvolvida apresentou menor valor de produção comparado às alternativas comerciais. O custo estimado para a fabricação de uma roda foi de R$105,02, totalizando R$210,04 para um par. Em comparação, as rodas comerciais apresentaram custo aproximado de R$120,00 por unidade, acrescido de R$15,00 referente à câmara de ar, resultando em aproximadamente R$270,00 para o par. Dessa forma, a solução desenvolvida representa uma redução de aproximadamente 22,2% no custo do par de rodas. Além da vantagem econômica, o modelo apresenta versatilidade dimensional, podendo ser redimensionado digitalmente para produzir rodas maiores ou menores conforme o porte físico da criança, sem a necessidade de novos moldes ou da aquisição de componentes adicionais.

As rodas na versão final foram aplicadas nas cadeiras de rodas infantis e submetidas a uso real por crianças de um a cinco anos de idade [Fig. 9]. Até o momento, os componentes apresentaram desempenho satisfatório, sem registro de falhas

estruturais ou necessidade de manutenção.

IV. DISCUSSÃO

O principal resultado deste estudo foi a demonstração da viabilidade de um processo iterativo de desenvolvimento aplicado à fabricação de rodas para cadeiras de rodas infantis por manufatura aditiva. O desenvolvimento foi conduzido de forma iterativa, com sucessivas modificações estruturais orientadas pelas limitações identificadas em cada etapa, contribuindo para o aprimoramento progressivo do componente. Esses aspectos reforçam um dos principais desafios apontados pela literatura: ampliar o acesso a tecnologias assistivas por meio de soluções acessíveis e adaptáveis [1,3].



Fig. 9. Criança testando a cadeira de rodas com a Roda completa, composta por um aro de quatro partes, pneu e aro de propulsão.

A evolução dos protótipos evidenciou a importância do processo de desenvolvimento iterativo. A impressão do aro em peça única eliminou as junções presentes na versão segmentada, reduzindo possíveis pontos de falha e aumentando a resistência estrutural do componente. Além disso, a substituição do filamento ABS pelo PETG reforçado com fibra de carbono proporcionou maior rigidez e resistência mecânica ao aro, enquanto o TPU apresentou desempenho compatível com a função de pneu, devido à sua flexibilidade e capacidade de absorção de impactos. A combinação desses materiais contribuiu para a obtenção de uma roda mais robusta e adequada à aplicação.

Outro aspecto relevante foi a possibilidade de personalização proporcionada pela manufatura aditiva. O modelo digital pode ser facilmente redimensionado para diferentes tamanhos de roda, permitindo sua adaptação às necessidades de crianças com diferentes características físicas, sem a necessidade de novos moldes ou componentes específicos. A possibilidade de redimensionamento digital do modelo amplia significativamente seu potencial de aplicação clínica, permitindo adequações rápidas às diferentes faixas etárias sem necessidade de fabricação de novos moldes ou aquisição de componentes específicos. Essa característica representa uma vantagem particularmente relevante em dispositivos destinados ao público pediátrico, cujas necessidades dimensionais se modificam rapidamente em função do crescimento. [8]

A aplicação das rodas nas cadeiras de rodas infantis demonstrou desempenho satisfatório durante o uso, sem ocorrência de falhas estruturais. Entretanto, o estudo atual apresenta como limitação a ausência de ensaios mecânicos padronizados. Dessa forma, estudos futuros são necessários para avaliar parâmetros como resistência mecânica, durabilidade, fadiga, desgaste e desempenho em diferentes condições de uso, além de comparações com rodas comerciais. Essas avaliações poderão fornecer evidências mais robustas sobre a confiabilidade e a aplicabilidade da solução desenvolvida.

De modo geral, os resultados reforçam o potencial da impressão 3D para o desenvolvimento de dispositivos de tecnologia assistiva de baixo custo, favorecendo a ampliação do acesso a recursos que promovam a mobilidade, autonomia e inclusão social.

Durante o desenvolvimento do protótipo, foram identificadas algumas dificuldades relacionadas às limitações dimensionais das impressoras 3D disponíveis. Ainda é necessário, trabalhos futuros, para realizar ensaios mecânicos padronizados para avaliar a resistência, a durabilidade e o desgaste dos componentes, bem como acompanhar o desempenho das rodas por períodos mais prolongados e em diferentes condições de uso.

V. CONCLUSÃO

O presente estudo demonstrou a viabilidade técnica de um processo iterativo de desenvolvimento de rodas para cadeiras de rodas infantis utilizando manufatura aditiva. A evolução sucessiva dos protótipos permitiu superar limitações identificadas nas soluções anteriores, resultando em um componente funcional, personalizável e economicamente competitivo. Mais do que o desenvolvimento de uma nova roda, o estudo estabelece um modelo de projeto potencialmente replicável para fabricação de componentes personalizados destinados a dispositivos de tecnologia assistiva.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] World Health Organization. Assistive technology [Internet]. Geneva: World Health Organization; 2025. Available from: https://www.who.int/teams/health-product-and-policy-standards/assistive-and-medical-technology/assistive-technology

[2] Ferretti EC, Curi HT, Andrade LF, Cooper RA, Soárez PC de. Conceptual mapping proposed to comprehend the effect of wheelchair mobility on social participation and quality of life: a systematic review*. Disability and Rehabilitation: Assistive Technology. 2022 Oct 19;19(3):1–17. doi: 10.1080/17483107.2022.2126904

[3] Rasmussen KAM, Schwartz JK, Aylmer K, Keahi C, Said A, Ballard DH, et al. A systematic review of 3D-printed assistive technology rehabilitation interventions. Assistive technology : the official journal of RESNA [Internet]. 2026 May [cited 2026 July 30];1–12. Available from: https://pubmed.ncbi.nlm.nih.gov/41642098/ . doi: 10.1080/10400435.2026.2619686.

[4] Medola FO, Elui VMC, Santana C da S, Fortulan CA. Aspects of Manual Wheelchair Configuration Affecting Mobility: A Review. Journal of Physical Therapy Science [Internet]. 2014 Feb 1 [cited 2026 July 30];26(2):313–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3944313/ . doi: 10.1589/jpts.26.313.

[5] Liang S, Huang M, Jiang D, Chen J, Hu L, Chen J, et al. 3D Printing of TPU-Liquid Metal Composite Inks for the Preparation of Flexible Sensing Electronics. ChemistryOpen [Internet]. 2024 Sept [cited 2026 July 30];13(9):e202300301. Available from: https://pubmed.ncbi.nlm.nih.gov/38666528/ .doi: 10.1002/open.202300301.

[6] Campo EA. Seleção de materiais poliméricos: como selecionar propriedades de projeto a partir de diferentes normas. Norwich (NY): William Andrew; 2008.

[7] Hozdić E, Hozdić E. Comparative Analysis of the Influence of Mineral Engine Oil on the Mechanical Parameters of FDM 3D-Printed PLA, PLA+CF, PETG, and PETG+CF Materials. Materials (Basel, Switzerland) [Internet]. 2023 Spring [cited 2026 July 30];16(18):6342. Available from: https://pubmed.ncbi.nlm.nih.gov/37763618. doi: 10.3390/ma16186342.

[8] World Health Organization (WHO), “Weight-for-age: Birth to 5 years,” Child Growth Standards, 2006.

A Brief Review of Control Strategies for Functional Electrical Stimulation Cycling

Caio Sudário Alves
Department of Electrical Engineering
University of Brasília (UnB)
Brasília, DF, Brazil
ORCID: 0009-0009-2926-5372

João Luiz Quagliotti Durigan
Department of Physiotherapy
University of Brasília (UnB)
Brasília, DF, Brazil
ORCID: 0000-0002-7511-5289

Roberto de Souza Baptista
Department of Electrical Engineering
University of Brasília (UnB)
Brasília, DF, Brazil
ORCID: 0000-0002-0228-6669

Abstract—Functional Electrical Stimulation (FES) cycling uses electrically induced muscle contractions to produce coordinated pedaling movements and support lower-limb rehabilitation. Since the neuromuscular response is a nonlinear, time-varying system affected by muscle fatigue and highly dependent on individual physiological characteristics, developing an efficient control system remains a challenging task. This paper presents a brief review of the main control strategies applied to FES cycling. The methods are organized into three groups: classical feedback and proportional-integral-derivative control; adaptive, optimized, and model-based control; and intelligent and learning-based control, including fuzzy logic, neural networks, and deep reinforcement learning. Their main advantages, limitations, implementation requirements, and potential roles are qualitatively compared. Conventional PID controllers remain attractive because of their simplicity and low computational cost, whereas adaptive and model-based strategies offer greater personalization at the cost of increased complexity. Fuzzy logic and learning-based methods provide alternatives for handling nonlinear behavior and variability. The reviewed literature suggests that hybrid architectures combining classical feedback with adaptive or intelligent methods represent a promising direction for reliable and personalized FES cycling.

Keywords—Functional electrical stimulation, FES cycling, spinal cord injury, rehabilitation, closed-loop control, PID control, adaptive control.

I. INTRODUCTION

Functional Electrical Stimulation (FES) is a rehabilitation and assistance technique in which electrical pulses are applied to peripheral nerves to activate weakened or paralyzed muscles. When applied to the lower limbs, FES can produce functional movements such as standing, walking, and pedaling. In FES cycling, the stimulation of muscle groups, such as the quadriceps and hamstrings, is coordinated with the crank position to generate cyclic pedaling motion. Fig. 1 illustrates a typical FES cycling configuration.

This approach provides lower-limb exercise and has been studied in individuals with neurological impairments, particularly those resulting from stroke and spinal cord injury [1], [2]. Beyond restoring movement, FES cycling has been associated with improved lower-limb muscle health in adults with spinal cord injury, with lower certainty of evidence for gains in power output and aerobic fitness, and inconsistent findings for bone and cardiovascular outcomes [3]. Because these adaptations depend on sustained and repeated training, the duration and consistency of each session are clinically relevant, which motivates control strategies capable of maintaining stable pedaling as the muscles fatigue.



Fig. 1. FES cycling training session performed using a recumbent tricycle mounted on a stationary trainer. Surface electrodes positioned over the thigh muscles deliver stimulation synchronized with crank motion, producing cyclic pedaling. Orthopedic braces stabilize the lower limbs, constrain their movement primarily to the sagittal plane, and secure the feet to the pedals. The system also includes an LCD display, control buttons mounted on the handlebars, and a rear compartment housing the batteries, electrical stimulator, embedded computer, and sensor receiver.

Despite its rehabilitation potential, control loop design for FES cycling encounters significant barriers related to the time-varying parameters of muscles and fatigue. Many systems apply predefined stimulation patterns based on crank angle or trial-and-error procedures. These open-loop approaches are relatively simple, but fail to automatically compensate for changes in muscle response during exercise. The relationship between electrical stimulation and the resulting movement is non-linear and time-varying, while muscle fatigue, spasticity, electrode placement, external disturbances, and inter-subject variability can considerably affect cadence and performance during exercise [1], [4], [5].

Closed-loop control seeks to overcome some of these limitations through the use of feedback from encoders installed on the crank, inertial measurement units (IMUs), pedal force sensors, or mechanomyographic sensors. The measured movement is compared with a desired reference, such as cadence, crank velocity, or joint trajectory, and the stimulation parameters are adjusted according to the resulting error. Pulse width, current amplitude, and stimulation frequency are among the parameters that can be modified in real time to regulate movement and maintain the desired exercise condition [1], [4], [5].

Various control methods have been investigated for this purpose. Conventional proportional-integral-derivative (PID) control is widely used due to its simple structure, interpretability, and low computational cost, both in terms of

algorithm and hardware. However, its fixed parameters can become inadequate as fatigue increases [2].

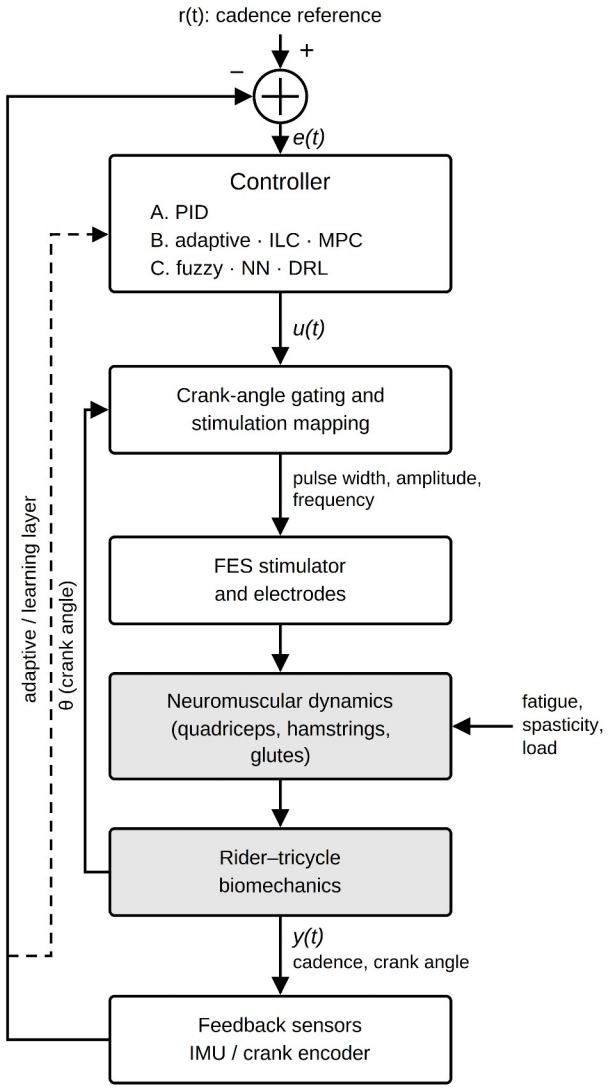


Fig. 2. Closed-loop architecture of an FES cycling system. Labels A, B, and C inside the controller block correspond to the three control families reviewed in Section II. Shaded blocks represent the physiological and mechanical plant. The dashed path represents an optional adaptation or learning mechanism that may update controller parameters, internal models, or stimulation policies depending on the adopted strategy.

Adaptive and optimized controllers seek to update stimulation commands or controller gains during operation, whereas model-based approaches use mathematical representations of the neuromuscular and biomechanical systems to predict their responses. Model Predictive Control (MPC) can additionally optimize future control actions while considering physiological and operational constraints. However, most available evidence on MPC pertains to FES-assisted motor tasks in general, rather than FES cycling specifically [6].

Intelligent control strategies offer alternative ways to deal with non-linearities and rider-specific behaviors. Fuzzy logic-based controllers represent the relationship between tracking errors and stimulation commands through linguistic rules, without requiring an exact mathematical model [7], [8]. Neural networks can learn non-linear mappings from experimental data and can be used to estimate muscle responses or adjust controller parameters [2], [5]. Another technique, deep reinforcement learning, has been studied as a method to learn stimulation parameters capable of improving cadence tracking while limiting unnecessary stimulation [9].

This paper presents a brief narrative review of representative control strategies applied to FES cycling. The selected studies were used to illustrate the main characteristics of each control family, rather than to perform a systematic comparison or meta-analysis. The search and selection procedure is described in Section II. Unlike existing reviews in [1], [2], and [6], this work organizes the strategies according to the point where each acts within the closed-loop architecture and presents the experimental validation associated with each, so that algorithm development and clinical evidence can be evaluated separately.

II. REVIEW METHODOLOGY

The aim was not to provide an exhaustive coverage of the field, but rather to select representative studies illustrating each control family and the level of experimental evidence associated with each. Therefore, the study was conducted as a narrative review with its search procedure documented.

A. Search Strategy

The literature search combined keyword searches with an analysis of the references from the articles identified. In addition, references that appeared in common among different studies were examined in order to identify other relevant works for the review. The search was conducted between April and June 2026 and considered studies published from 2016 onwards. Keyword searches were conducted in Google Scholar using at least four similar queries relating FES cycling to different control families, including, for example: "functional electrical stimulation"; "cycling closed-loop control"; "cycling closed-loop control" with bang-bang control; "FES cycling PID"; "FES cycling adaptive" OR "iterative learning" OR "model predictive" control; and "FES cycling fuzzy" OR "neural network" OR "reinforcement learning" control. The results of each search were analyzed in order of relevance. For the more specific queries, approximately the first 30 results were assessed, whereas for the broader queries, the screening extended to more than 100 results, which were analyzed in blocks. For these broader searches, the screening was progressively expanded until the additional results no longer yielded studies that met the inclusion criteria. Since Google Scholar provides only an estimated total number of results, the number of results actually screened in each search was recorded instead. Initially, the studies were assessed based on their title and abstract. Studies considered potentially relevant were subsequently examined based on their abstract, methodology, and conclusions. When a study remained consistent with the established criteria, the full text was then read and analyzed.

The keyword search was also complemented by citation network analysis using ResearchRabbit, based on four articles considered relevant to the topic [2], [3], [13], [16]. The references cited by these articles, the studies that cited them, and other related studies suggested by the platform were analyzed. The studies identified through this process were also checked in Google Scholar and evaluated according to the same eligibility criteria.

This step was intended to identify studies that were not found through the keyword searches, particularly studies that presented control strategies related to FES cycling but did not use the term "FES cycling" in the title or abstract.

B. Inclusion and Exclusion Criteria

Studies were included if they described a closed-loop control strategy applied to FES cycling or to an FES-assisted motor task directly related to cadence regulation. The studies were also required to describe the controller structure and the modulated stimulation parameter, as well as report quantitative performance results obtained through simulation, in able-bodied participants, or in participants with neurological impairment. Studies conducted by research teams participating in competitions, particularly those conducted following cycling-related competitions such as Cybathlon, were also considered. Studies were excluded if they did not describe a control law, focused exclusively on physiological or clinical outcomes, or did not report quantitative performance measures such as cadence, speed, or time. Extended abstracts and non-peer-reviewed materials were also excluded. Only studies published in English were included.

C. Selection Procedure

The studies were initially selected based on a review of their titles and abstracts. After this stage, 39 papers were selected for full-text reading and assessment according to the eligibility criteria. In the end, 16 studies were selected because they presented results considered relevant to the control families discussed in Section III. When selecting the studies, preference was given to papers that presented quantitative results, clearly described the stimulation parameters used, and included experimental validation. More recent studies were also preferred when papers with similar characteristics were available. References [10] and [11] were not part of the selected studies and were used only as references for the formal description of PID control.

D. Limitations of the Review

This review has some limitations. The search was conducted mainly through Google Scholar and complemented by citation network analysis. The number of results analyzed varied depending on the search, and no systematic or preregistered review protocol, such as PRISMA, was used. In addition, the studies were selected by a single author, with some guidance from the advisor, without a third independent assessment or a formal risk-of-bias analysis. A meta-analysis was also not performed. Therefore, the possibility of selection bias cannot be completely ruled out. For this reason, the comparison presented in Section IV should be understood as a qualitative analysis of representative studies rather than as evidence that one particular control strategy is superior to the others.

III. CONTROL STRATEGIES

Fig. 2 shows a representative closed-loop architecture adopted in FES cycling systems. The controller generates a control command based on the measured error. This command is synchronized with the crank angle and converted into stimulation parameters applied to the quadriceps and hamstrings, while the resulting movement is measured and fed back to the comparator. Muscle fatigue acts as a disturbance within the control loop rather than on the reference signal, as commonly represented in the literature. Therefore, this muscular fatigue helps explain why fixed controller parameters may become inadequate over the course of a training session. Accordingly, the reviewed strategies differ mainly in the implementation of the controller block and, in some methods, in the presence of adaptation or learning mechanisms capable of modifying gains, internal models, or stimulation parameters. The three control families discussed in the following subsections are indicated within the controller block: classical feedback and PID control (A); adaptive, optimized, and model-based control (B); and intelligent and learning-based control (C).

A. Classical Feedback and PID Control

The proportional-integral-derivative (PID) controller is one of the most popular feedback control techniques in engineering. The controller combines information related to the present, the past, and the expected trend of the tracking error through the proportional, integral, and derivative terms, respectively [1], [2]. The control signal can be represented as:

$$u(t) = K_p e(t) + K_i \int e(\tau)\, d\tau + K_d \frac{de(t)}{dt} \quad (1)$$

In (1) $e(t) = r(t) - y(t)$ is the difference between the desired reference $r(t)$and the measured output of the system $y(t)$ [10], [11].

In FES cycling, the reference signal may represent the desired pedaling cadence range, pedal speed, knee angle, or joint trajectory. The measured variable is obtained through sensors such as encoders or inertial measurement units (IMUs). Based on the difference between the desired and measured values, the PID controller modifies an electrical stimulation parameter, usually pulse width, current, or stimulation amplitude. For this purpose, it is necessary to select which parameter will be modulated while the others remain fixed. When the measured cadence is below the reference value, the controller increases the stimulation command to generate a stronger muscle contraction. Conversely, when the cadence exceeds the desired value, the stimulation command is reduced. This feedback structure allows the system to continuously regulate stimulation rather than applying a predefined and constant stimulation pattern [2], [4], [12].

The use of feedback is important in FES cycling because the muscular response cannot be clearly and accurately predicted solely from the applied stimulation signal. Open-loop systems generally use stimulation patterns determined through clinical experience, visual observation, or trial-and-error procedures. However, their performance may deteriorate when the muscular response changes due to fatigue, external disturbances, electrode positioning, or differences among users.

Closed-loop PID control seeks to partially compensate for these variations by continuously monitoring the movement and correcting the stimulation parameters according to the observed error. Closed-loop feedback can improve cadence regulation and trajectory tracking compared with fixed-stimulation approaches, although its performance de-

pends on the controller design, tuning procedure, and experimental conditions [2], [4], [12].

Fonseca et al. evaluated cadence tracking and disturbance rejection in a participant with complete spinal cord injury using an FES-controlled tricycle. The system modulated the stimulation pulse width according to cadence feedback obtained from an IMU. The control strategy was able to track a varying cadence reference when the cadence remained above approximately 25 rpm using quadriceps stimulation alone. However, disturbance loads reduced the cadence, and the addition of hamstring and gluteal stimulation was associated with increased oscillations and spasticity. Because the investigation was a single-participant case study, its results should not be generalized to the entire population of FES-cycling users [12]. The same tricycle platform, in a mobile configuration and with the same stimulation and crank-sensing architecture, was later used in the Cybathlon 2016 FES bike race, where pilot-defined stimulation intensity was adopted in place of automatic cadence regulation [13]. This contrast illustrates that closed-loop cadence control and practical operation by the user may impose competing requirements.

Despite these advantages, conventional PID control presents important limitations in FES cycling. The relationship between electrical stimulation and the resulting muscle force is nonlinear, time-varying, and strongly dependent on the physiological condition of the user. Muscle fatigue progressively modifies this relationship during exercise. This means that a set of fixed controller gains may provide satisfactory performance at the beginning of a training session but become inadequate over time. Because FES cycling involves a complex biological system, several dynamic factors affect pedaling performance, with fatigue being one of the most important [1], [2], [5]. Other difficulties include overshoot, long settling times, and sensitivity to disturbances. Individual differences in muscle recruitment, electrode positioning, neurological impairment, and fatigue resistance also reduce the possibility of using the same fixed PID gains for all participants. Consequently, conventional PID controllers generally require experimental tuning for each user and operating condition. Although fixed-gain PID control may outperform open-loop stimulation, it may not ensure consistent performance throughout prolonged rehabilitation sessions [2], [4], [11], [12].

Therefore, PID control remains an important reference strategy for FES cycling because of its simplicity, interpretability, low implementation cost, and established use in feedback systems. However, its performance when used alone may be limited under nonlinear and time-varying operating conditions. Consequently, conventional PID is particularly useful as a baseline controller, whereas adaptive or hybrid PID structures appear more suitable for personalized and long-duration FES-cycling applications.

B. Adaptive, Optimized, and Model-Based Control

Muscle response varies among FES users; therefore, adaptive control is a promising tool for FES cycling. Adaptive strategies use feedback information to update the stimulation command or controller parameters according to the observed system response. Variables such as cadence, crank angle, resistance, and tracking error can be used to adjust the stimulation intensity through pulse width, current amplitude, or stimulation frequency. This allows the controller to compensate, at least partially, for changes caused by muscle fatigue [2].

The repetitive and predictable nature of pedaling also makes FES cycling suitable for iterative learning control (ILC) and repetitive learning control (RLC). Both use information recorded during previous cycles or trials to correct the stimulation command applied in subsequent cycles, so that systematic tracking errors are progressively reduced without requiring the complete system dynamics to be known in advance. The two formulations differ in their operating assumptions. ILC is applied to tasks performed over a finite interval, with the initial conditions reset at the beginning of each trial, whereas RLC is intended for continuous operation without resetting [14].

Ghanbari et al. applied ILC to a motorized FES cycle-rider system within a passivity-based framework. The closed-loop switched system was shown to be output strictly passive, and an Arimoto-type learning update law was then used to guarantee convergence of the cadence tracking error in the L2 sense. In experiments with seven able-bodied participants stimulated only over the quadriceps, the mean cadence tracking error was 0.00 ± 2.47 rpm for a 50 rpm reference. The authors ascribed the remaining error to unaccounted factors like electromechanical delay and muscle fatigue, noting that participants with neurological conditions had not yet been assessed [15].

Duenas et al. developed a switched RLC in which stimulation was distributed among six lower-limb muscle groups according to the crank angle, while an electric motor provided assistance in the regions of low torque transfer. Asymptotic cadence tracking was established through a Lyapunov-based analysis for nonsmooth switched systems. The controller was evaluated in five able-bodied participants and three participants with neurological conditions. Trials performed with the learning feedforward term active yielded a lower root-mean-squared cadence error than trials in which it was disabled, with mean values of 3.68 rpm against 4.20 rpm for the able-bodied group and 3.17 rpm against 3.49 rpm for the impaired group [14].

These findings imply that cadence control in FES cycling can be enhanced by learning-based feedforward terms. Two restrictions should be taken into account, though. First, the reported tracking performance cannot be solely attributable to electrical stimulation because the learning-based controller in both cases worked in tandem with an electric motor. Second, whereas muscle fatigue is a nonperiodic disturbance that builds up during the session and is not immediately addressed by the learning term, learning control makes up for the periodic component of the system dynamics [14], [15]. Therefore, the consistency of the pedaling task and the accuracy of the feedback readings determine how effective these strategies are [2].

Optimization and adaptation are related but are not identical. Optimized PID structures represent another approach for improving conventional feedback control. Optimization searches for controller parameters or stimulation patterns that minimize a predefined performance criterion, whereas adaptation modifies these parameters in response to changes observed during operation [5].

Backpropagation (BP) is an algorithm used to train a neural network by adjusting the weights between its layers according to the error between the obtained and desired outputs. In a BP-PID controller, the three neurons in the output layer correspond to the proportional, integral, and derivative

gains, respectively. Therefore, the learning capability of the neural network allows the values of Kp, Ki, and Kd to be updated according to the system behavior. In the complete structure, Particle Swarm Optimization (PSO) improves the search and optimization process, the BP neural network adjusts the controller parameters, and the PID controller generates the command responsible for modifying the electrical stimulation pulse width [5].

Model-based control uses mathematical representations to predict how stimulation parameters affect muscle activation, joint torque, lower-limb movement, and user performance. Patient-specific characteristics, such as limb dimensions, range of motion, and nonlinear force–length and force–velocity relationships, can also be incorporated into a biomechanical model [5], [6]. An important example of an optimized and model-based approach is Model Predictive Control (MPC). MPC calculates stimulation commands that minimize a cost function while satisfying constraints related to stimulation intensity, joint movement, muscle activation, or the contribution of assistive devices. To do so, the controller predicts the future response of the system over a finite horizon. When new measurements become available, the optimization process is repeated, but only the first calculated control command is applied [6].

However, the evaluated studies primarily comprised movements with a limited number of degrees of freedom and covered a variety of upper- and lower-limb FES tasks. As a result, the evidence was not limited to cycling. Therefore, rather than being a proven cycling solution, MPC should be considered a promising model-based approach for future FES-cycling systems [6].

The challenges of finding precise patient-specific models, the computational costs of real-time optimization, and differences between theoretical predictions and actual muscular responses are the primary drawbacks of optimal and model-based controllers [6]. In general, model-based, optimal, and adaptive techniques are more robust and personalized than fixed-gain control. However, dependable sensors, computationally effective algorithms, and models that can be updated in response to changes in the user's muscular reaction are necessary for their practical use in FES cycling [1], [2], [5], [6].

C. Intelligent and Data-Driven Control

Fuzzy Logic Control (FLC) represents the relationship between system inputs and stimulation commands through membership functions and rules expressed in linguistic terms. The inputs commonly include the tracking error and its rate of change, while the output may represent a variation in stimulation amplitude or pulse width. The fuzzy inference mechanism combines the activated rules and converts the resulting fuzzy output into a numerical stimulation command [2], [7], [8].

Watanabe and Tadano developed a continuous-time fuzzy FES controller that used knee-angle error and its derivative to modify the stimulation pulse amplitude. The controller was compared with a PID controller in knee-extension experiments involving three healthy participants. Although the fuzzy controller generally presented favorable tracking performance, the study did not directly test FES cycling. It was conducted as a preliminary evaluation before the controller's intended application to cycling-speed regulation [7].

A fuzzy controller was studied in FES cycling by Ahmad et al. The system used phase-angle information and adjusted the stimulation pulse width between 50 and 200 μs while maintaining a frequency of 35 Hz. Three healthy participants and one with post-stroke hemiplegia participated in the experiments. The hemiplegic individual had a higher fatigue index than the healthy participants, despite the controller producing comparatively minimal trajectory errors. Further validation and patient-specific adaptation are required, as evidenced by the small sample size and the inclusion of only one participant with neurological impairment [8].

Neural networks provide a different approach by learning nonlinear mappings between measured variables, stimulation inputs, and the resulting movement. Once trained, a neural network can estimate stimulation parameters, predict muscular responses, or adjust the gains of another controller, such as a BP-PID controller. Therefore, neural networks can improve personalization and reduce dependence on fixed controller parameters. However, their performance depends on the quantity and quality of the training data, and their computational requirements may hinder real-time implementation [2], [5].

Deep Reinforcement Learning (DRL) allows the controller to learn a stimulation policy through interaction with a simulated or physical environment. In reinforcement learning, the controller applies a control action after receiving information about the current state. Depending on the quality of the resulting movement, the controller receives a numerical reward. The controller learns to create stimulation patterns that enhance cadence tracking while reducing unnecessary stimulation by optimizing the total reward [9].

Wannawas et al. developed a neuromechanics-based DRL controller trained using a personalized musculoskeletal model that incorporated a model of muscle fatigue. Two agents were employed: a Starter agent initiated the cycling movement, whereas a Tracker agent regulated the desired cadence. In simulation, the DRL controller achieved lower tracking errors and shorter response times than the implemented PID and fuzzy controllers. However, the DRL controller had direct access to simulated fatigue information that was unavailable to the conventional controllers, which limits the fairness of a direct comparison [9]. Moreover, the complete adaptive controller was not transferred to real-time patient operation, and the authors stated that real-world fatigue estimation and the amount of required training data remained important limitations. As a result, while the results show viability, they do not yet represent thorough clinical validation [9].

Overall, fuzzy logic is a comparatively straightforward and comprehensible method for managing nonlinear behavior without the need for a precise system description [2], [7], [8]. Increased learning and personalization capabilities, such as the ability to automatically recognize stimulation patterns and adjust to muscle fatigue, are provided by neural networks and reinforcement learning [2], [5], [9]. However, these advantages are accompanied by greater computational complexity, dependence on training data or simulation models, and the need for more rigorous clinical validation [1], [6], [8], [9]. Therefore, intelligent and learning-based methods are promising for personalized FES cycling, but their reliability and safety must be demonstrated through more robust studies.

D. Fatigue Measurement and Estimation

Muscle fatigue appears throughout this review as

one of the main factors that reduce FES cycling performance during a session. Therefore, it is important to examine how fatigue was measured or estimated in the reviewed studies. In some cases, fatigue is only compensated for by the controller, while in others it is estimated or explicitly included in the model. Overall, three approaches were identified.

The first, and most common, approach is not to measure fatigue directly. In this case, fatigue is indirectly compensated for through the cadence error: as the muscle response decreases, cadence drops, the tracking error increases, and the controller increases the stimulation. This approach was used in [4], [7], [12], [14], and [15]. In these studies, fatigue is recognized as a limitation and, in [14] and [15], is described as a non-periodic disturbance that cannot be compensated for by the learning term. However, no fatigue-related variable is directly included in the control loop. An intermediate case is reported in [4], in which mechanomyographic sensors were used to record muscle activity, but their signals were not incorporated into the closed-loop control algorithm due to real-time processing limitations. This suggests that the limitation was more related to implementation than to the availability of sensing technology.

The second approach uses fatigue indicators derived from signals already available in the control system [8], [5]. These indicators are useful, but they are still based on tracking error or task interruption rather than on a direct physiological measure of muscle state.

The third approach considers fatigue as an explicit state of the system. In the deep reinforcement learning (DRL)-based controller presented in [9], the personalized musculoskeletal model included a fatigue model, allowing its state to be directly used by the learning agent. However, this state was simulated rather than experimentally measured. In addition, it was not available to the PID and fuzzy controllers used for comparison, which limits the equivalence between the evaluated strategies.

Outside closed-loop FES cycling control, muscle fatigue has also been directly assessed using force-based metrics. Time to fatigue and the force–time integral were used as measures, and the authors observed that the advantage of distributed stimulation decreased significantly at high stimulation intensities [16]. These results were obtained under isometric and open-loop conditions, showing that fatigue reduction in FES has mainly been addressed through the stimulation strategy rather than directly through the controller. Overall, none of the reviewed FES cycling studies used a physiological indicator of muscle fatigue as a feedback signal in closed loop. Fatigue is frequently identified as one of the main disturbances affecting the system, but it remains either unmeasured or estimated from the same error signal already used by the controller. This is an important limitation because a controller that only compensates for a decrease in cadence cannot distinguish whether it was caused by muscle fatigue or by an external load. Therefore, it cannot adapt the stimulation specifically to the muscle's fatigue state.

The three categories mentioned above, fixed structure and fixed gains, explicit adaptation or prediction, and data-driven learning, address the same control problem from different angles, but the studies supporting them were conducted under markedly different conditions. Table I summarizes the experimental characterization of each representative study, including the number and profile of participants, stimulated muscles and stimulation parameters, reported error metrics, the presence of motorized assistance, and level of evidence associated with each result. The heterogeneity shown in Table I is substantial, and the reported errors are not directly comparable across studies, since they were obtained under different references, modulated variables, and populations. The following section compares the strategies in terms of the trade-offs that determine their applicability.

Table I—Experimental characterization of the reviewed FES control studies

Study	Participants	Muscles / FES parameters	Error metric (reported result)	Motor assist	Level of evidence
Fonseca [12]	1 male, complete SCI T9	Quadriceps (± hamstrings, glutei); 40 mA, 50 Hz; PW 0–500 μs	Absolute mean error 1.9–10 % (quadriceps only)	No	Single-participant case study
Woods [4]	1 able-bodied	Quadriceps, hamstrings, glutei; 40 Hz; PW 200–250 μs; % current modulated	No convergence above 60 rpm	No	Proof of concept, able-bodied
Ghanbari [15]	7 able-bodied	Quadriceps only; 60 Hz; 90 mA; PW modulated	0.00 ± 2.47 rpm at 50 rpm	Yes	Controlled exp., AB, motor-assisted
Duenas [14]	5 able-bodied + 3 neurological	6 muscle groups; 60 Hz; 70–90 mA; PW modulated	RMS 3.68 vs 4.20 rpm (AB); 3.17 vs 3.49 (NC)	Yes	Controlled exp., AB & neuro., motor-assisted
Sun [5]	6: 1 healthy, 3 SCI, 2 stroke	6: 1 healthy, 3 SCI, 2 stroke	RMSE not consistently reduced; gain in cycling time	No	Small mixed clinical sample
Watanabe [7]	3 able-bodied	Rectus femoris, vastus lateralis; 30 Hz; PW 300 μs; amplitude modulated	MAE lower than PID ($p < 0.05$) 0.05)	No	Preliminary, non-cycling task
Ahmad [8]	3 able-bodied + 1 stroke	Quadriceps only; 35 Hz; PW 50–200 μs	Trajectory error 2.69–3.46°; fatigue index 0.046–0.230	No	Small sample, one patient
Wannawas [9]	None (simulation)	Simulated model	Lower RMSE than PID and fuzzy, with access to fatigue state	N/A	Simulation only

SMA: single muscle activation; MMA: multiple muscle activation; PW: pulse width. Values are reported as in the original publications. Watanabe and Tadano [7] evaluated knee extension rather than cycling. The number of participants reported in the text of [5] differs from the value given in its Table 1.

Table II—Proposed comparison criteria for FES cycling controllers

Criterion	Operational definition	Unit
Cadence tracking error	RMS of the difference between reference and measured cadence over a declared steady-state window; the reference value and window must be reported	rpm (and % of reference)
Response time	Time from a step change in the reference until the cadence enters and remains within a declared tolerance band	s
Mechanical output	Integral of crank torque times angular velocity over the session; mean power when session durations differ	J (or W)
Stimulation charge	Sum over channels of current × pulse width × frequency × active time; normalized by mechanical work to express stimulation cost	mC (and mC/J)
Fatigue index	Decline of a declared physiological or mechanical quantity between the initial and final portions of the session, or time until that quantity falls below a stated fraction of its initial value	dimensionless (or s)
Computational cost	Controller execution time per iteration relative to the control period, together with the hardware class used	ms / % of period
Safety	Declared current and pulse-width limits, number of saturation events, and adverse occurrences such as spasms, discomfort, or session interruption	counts

IV. COMPARATIVE DISCUSSION

The reviewed strategies differ mainly in how they trade simplicity and interpretability against adaptability. Conventional PID and fuzzy controllers remain attractive when low implementation cost and transparent behavior are priorities, whereas adaptive, learning-based, and model-based approaches offer greater personalization at the cost of modeling effort, data requirements, or computational load [1], [2], [5], [6], [12].

The experimental basis of the reviewed strategies is markedly uneven. Among the representative studies selected in this review, direct experimental evidence for fixed-gain PID includes small-scale investigations [12], whereas iterative and repetitive learning controllers have been evaluated in both able-bodied participants and individuals with neurological impairments [14], [15]. The PSO-BP-PID approach reported in [5] was experimentally evaluated with six participants, although the sample remained small. In contrast, the DRL approach in [9] was evaluated primarily in simulation, while the available MPC evidence is dominated by non-cycling FES tasks [6]. Therefore, experimental maturity and algorithmic sophistication remain uneven across the reviewed control strategies.

The reviewed evidence does not indicate a universally superior control strategy for FES cycling. The rehabilitation goal, available sensors, processing power, necessary degree of customization, and safety restrictions all influence the best option.

The development of hybrid architectures that combine adaptive, model-based, or learning-based methods with the reliability and interpretability of conventional feedback control represents a promising future direction. A logical division of tasks is suggested by the contrast between periodic and nonperiodic disturbances, where an adaptive term deals with the progressive effect of fatigue and a learning term compensates for cycle-to-cycle repeating dynamics [14], [15]. Standardized performance measurements, patient-specific models, reliable fatigue-related feedback, and validation with larger and more diverse populations are all necessary for further advancement [1], [2], [5], [9].

The main limitation of the comparison presented above is not the lack of results, but the absence of a common evaluation framework. The reviewed studies use different reference signals, control variables, session durations, and error metrics, making direct comparisons difficult.

Rather than proposing a new control structure, we suggest that the architecture in Fig. 2 be adopted as a reference instrumentation frame, in which four measurement points are defined explicitly. At the controller output, the stimulation parameters should be recorded for each channel, allowing the total delivered electrical charge to be calculated. At the stimulator–electrode interface, fixed and modulated parameters should be reported, making it clear which variable is used for control. At the plant output, kinematic and, when available, kinetic measures should be recorded, including crank angle, cadence, torque, and mechanical power. When motorized assistance is used, its contribution should also be measured separately from the muscle contribution.

As shown in Table I, the two studies that used learning-based feedforward terms [14], [15] included motorized assistance, but neither reported cadence tracking with the motor disabled. Separating these contributions does not require additional hardware, only independent recording of the two control inputs.

Table II summarizes seven quantities that can be used to characterize FES cycling controllers. The goal is not to establish fixed target values, which depend on the rehabilitation objective and the study population, but to facilitate comparisons across different studies. The stimulation charge normalized by the mechanical work produced expresses how much stimulation was required per unit of useful output, linking tracking accuracy to its physiological cost. Thus, reducing cadence error by increasing stimulation is not necessarily equivalent to achieving the same error with a lower delivered electrical charge. Safety should also be reported more consistently, since previous studies have described spasticity [12], discomfort limiting session duration [14], and stimulation levels limited by participant tolerance [4].

Finally, studies should report a minimum set of information, including the number and clinical profile of participants, the stimulated muscle groups, fixed and modulated stimulation parameters, the presence and contribution of motorized assistance, whether fatigue-related information was used in the control loop, and the order in which the experimental conditions were tested.

V. CONCLUSION

FES cycling is a promising rehabilitation technique, but its performance is strongly affected by muscle fatigue, nonlinear neuromuscular behavior, external disturbances, and user individuality. This review discussed three main groups of control methods: classical PID control; adaptive, optimized, and model-based control, including iterative and repetitive learning; and intelligent and learning-based control.

Conventional PID remains an important reference because of its simplicity, but fixed gains may not provide consistent performance when the muscular response changes over time. Adaptive, optimized, model-based, and learning-based approaches offer greater personalization at the cost of additional modeling, data, or computational requirements. A promising direction is therefore not the complete replacement of classical controllers, but their integration with these methods [2], [5], [9], [14], [15]. The existing evidence, however, remains heterogeneous: several advanced approaches have been evaluated primarily in simulations, in healthy participants, or in small clinical samples, and the strategies with the strongest theoretical guarantees are not those with the broadest experimental support. Progress therefore depends less on additional controller formulations than on the evidence required to choose among those that already exist. Four gaps follow directly from this review. First, no reviewed study used a physiological indicator of muscle fatigue as a closed-loop feedback signal, although fatigue is consistently identified as the dominant disturbance. Second, the studies reporting the strongest theoretical guarantees employed motorized assistance without isolating the contribution of stimulation. Third, the comparison favoring learning-based control was obtained in simulation and under unequal information conditions. Fourth, reported errors are not commensurable across studies. Addressing these gaps requires the reporting conventions proposed in Section IV, comparisons under equivalent information and assistance conditions, and validation in larger and clinically representative samples.

Acknowledgment: The authors thank Estevão Lopes for participating in the FES cycling session shown in Fig. 1 and the EMA/LARA team at the University of Brasília for providing the experimental infrastructure and technical support. During the preparation of this manuscript, the authors used Claude (Anthropic) to assist in translating text originally written in Portuguese into English and for minor language editing, including grammar and syntax corrections. All AI-assisted text was reviewed by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

[1] L. Chaikho, E. Clark, and M. Raison, "Transcutaneous functional electrical stimulation controlled by a system of sensors for the lower limbs: A systematic review," Sensors, vol. 22, no. 24, Art. no. 9812, Dec. 2022, doi: 10.3390/s22249812.

[2] G. Wang, Y. Zhao, M. Sun, W. Dong, and S. Zhu, "Different control methods for FES cycling: Review," in *Proc. IEEE 24th Int. Conf. Softw. Qual., Rel., Secur. Companion (QRS-C)*, Cambridge, U.K., Jul. 2024, pp. 759–766, doi: 10.1109/QRS-C63300.2024.00101.

[3] J. W. van der Scheer, V. L. Goosey-Tolfrey, S. E. Valentino, G. M. Davis, and C. H. Ho, "Functional electrical stimulation cycling exercise after spinal cord injury: A systematic review of health and fitness-related outcomes," *J. NeuroEng. Rehabil.*, vol. 18, Art. no. 99, Jun. 2021, doi: 10.1186/s12984-021-00882-8.

[4] B. Woods, M. Subramanian, A. Shafti, and A. A. Faisal, "Mechanomyography based closed-loop functional electrical stimulation cycling system," in Proc. 7th IEEE Int. Conf. Biomed. Robot. Biomechatronics (BioRob), Enschede, The Netherlands, Aug. 2018, pp. 179–184, doi: 10.1109/BIOROB.2018.8487941.

[5] M. Sun et al., "Modeling and control of functional electrical stimulation cycling training system," Sci. Rep., vol. 15, Art. no. 6452, Feb. 2025, doi: 10.1038/s41598-024-84046-3.

[6] K. Co, M. Begon, F. Bailly, and F. Moissenet, "Optimal control-driven functional electrical stimulation: A PRISMA-ScR scoping review," Comput. Biol. Med., vol. 199, Art. no. 111311, Nov. 2025, doi: 10.1016/j.compbiomed.2025.111311.

[7] T. Watanabe and T. Tadano, "Design of closed-loop fuzzy FES controller and tests in controlling knee extension movements," IEICE Trans. Inf. Syst., vol. E100-D, no. 9, pp. 2261–2264, Sep. 2017, doi: 10.1587/transinf.2017EDL8061.

[8] M. K. I. Ahmad, A. U. B. Shamsudin, Z. A. Soomro, R. B. Abdul Rahim, B. S. K. K. Ibrahim, and M. S. Huq, "Closed-loop functional electrical stimulation (FES)-cycling rehabilitation with phase control fuzzy logic for fatigue reduction control strategies for stroke patients," Sinergi, vol. 28, no. 1, pp. 63–74, Feb. 2024, doi: 10.22441/sinergi.2024.1.007.

[9] N. Wannawas, M. Subramanian, and A. A. Faisal, "Neuromechanics-based deep reinforcement learning of neurostimulation control in FES cycling," in Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER), Virtual Event, Italy, May 2021, pp. 381–384, doi: 10.1109/NER49283.2021.9441354.

[10] L. Guo, "Feedback and uncertainty: Some basic problems and results," Annu. Rev. Control, vol. 49, pp. 27–36, 2020, doi: 10.1016/j.arcontrol.2020.04.001.

[11] Q. A. Mumuni, A. I. Olayiwola-Mumuni, and A. A. Yussouff, "The advent of the proportional integral derivative controller: A review," J. Adv. Eng. Technol., vol. 2, no. 10, pp. 5–22, 2023, doi: 10.24412/2181-1431-2023-2-5-22.

[12] L. O. da Fonseca, A. P. L. Bó, J. A. Guimarães, M. E. Gutierrez, and E. Fachin-Martins, "Cadence tracking and disturbance rejection in functional electrical stimulation cycling for paraplegic subjects: A case study," Artif. Organs, vol. 41, no. 11, pp. E184–E195, Nov. 2017, doi: 10.1111/aor.13055.

[13] A. P. L. Bó, L. O. da Fonseca, J. A. Guimarães, E. Fachin-Martins, M. E. G. Paredes, G. A. Brindeiro, A. C. C. de Sousa, M. C. N. Dorado, and F. M. Ramos, "A novel system for cycling using electrical stimulation for individuals with paraplegia, and preparation for Cybathlon 2016," IEEE Robot. Autom. Mag., vol. 24, no. 4, pp. 58–65, Dec. 2017, doi: 10.1109/MRA.2017.2751660.

[14] V. H. Duenas, C. A. Cousin, A. Parikh, P. Freeborn, E. J. Fox, and W. E. Dixon, "Motorized and functional electrical stimulation induced cycling via switched repetitive learning control," IEEE Trans. Control Syst. Technol., vol. 27, no. 4, pp. 1468–1479, Jul. 2019, doi: 10.1109/TCST.2018.2827334.

[15] V. Ghanbari, V. H. Duenas, P. J. Antsaklis, and W. E. Dixon, "Passivity-based iterative learning control for cycling induced by functional electrical stimulation with electric motor assistance," IEEE Trans. Control Syst. Technol., vol. 27, no. 5, pp. 2287–2294, Sep. 2019, doi: 10.1109/TCST.2018.2854773.

[16] E. Jafari, P. Kajganic, V. Bergeron, J. Di Marco, A. Metani, and L. Popovic-Maneski, "Efficacy of high- versus moderate-intensity spatially distributed sequential stimulation in subjects with spinal cord injury: An isometric study," J. NeuroEng. Rehabil., vol. 22, Art. no. 65, 2025, doi: 10.1186/s12984-025-01567-2.

A Computer Vision and Non-Immersive Virtual Reality Platform for Automated Functional Assessment During Hospital Cardiorespiratory Rehabilitation

1. Fernando Vinicius Teixeira Guimarães Pereira
Post Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
ft.fernandopereira@gmail.com

2. Luis Guilherme Silva Rodrigues
Interfaces and Visualization Laboratory
São Paulo State University (UNESP)
Bauru, Brazil
guilherme.rodrigues@unesp.br

3. Roberta Bordignon Rodrigues Machado
Post Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
robertabordignon_@hotmail.com

4. Thamires dos Santos Novais
Multidisciplinary Residency Program in Clinical Emergencies and Trauma at the University of Santo Amaro. São Paulo, Brazil.
tamiresnovais05@gmail.com

5. Patricia Lira dos Santos
Multidisciplinary Residency Program in Clinical Emergencies and Trauma at the University of Santo Amaro, São Paulo, Brazil
ptysantos22@gmail.com

6. Alessandro Márcio Hakme da Silva
Post Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
alessandro.silva@ub.edu.br

7. Alexandre Fonseca Brandão
Laboratory of Health Informatics and Innovation
Pontifical Catholic University of Campinas (PUC-Campinas)
Campinas, Brazil
alexandre.brandao@puc-campinas.edu.br

8. Marcela Maria Carvalho da Silva Zorrer Franco
Post Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
marcela.franco@ub.edu.br

***Abstract*—** Functional assessment is an essential component of hospital cardiorespiratory rehabilitation; however, conventional assessment methods still rely on manual repetition counting, external timing devices, and direct supervision by healthcare professionals, limiting workflow efficiency and assessment standardization. This study presents the design, development, and preliminary technical verification of EX-REAB, a modular computational platform that integrates computer vision and non-immersive virtual reality for automated functional assessment in hospital rehabilitation. The platform was developed through an iterative user-centered software engineering process involving physiotherapists, biomedical engineers, computer scientists, and software developers. EX-REAB incorporates MediaPipe Pose-based markerless motion capture, automated movement recognition, standardized execution of the 2-Minute Step Test and the 30-Second Sit-to-Stand Test, real-time performance feedback, and structured clinical data management. The resulting architecture comprises six interconnected functional layers, enabling automatic motion analysis, computation of kinematic variables, and export of standardized performance metrics for subsequent clinical and research applications. The modular architecture also supports future integration with physiological monitoring systems, wearable sensors, sleep assessment technologies, artificial intelligence algorithms, and clinical decision-support tools. EX-REAB establishes a scalable computational framework for automated functional assessment and provides the technological foundation for next-generation intelligent rehabilitation systems in hospital environments.

***Keywords*—** *Computer Vision, Non-Immersive Virtual Reality, Functional Assessment, Cardiorespiratory Rehabilitation, MediaPipe Pose, Digital Health.*

I. Introduction

Patients hospitalized with cardiorespiratory disorders frequently experience a rapid decline in physical function due to the combined effects of the underlying disease, prolonged bed rest, systemic inflammation, and physical inactivity during hospitalization. This functional deterioration is associated with reduced exercise capacity, muscle weakness, prolonged hospital stays, increased dependency, and poorer clinical outcomes after discharge. Consequently, early mobilization and cardiorespiratory rehabilitation have become

fundamental components of hospital care and are strongly recommended by national and international clinical guidelines for medically stable patients because they preserve functional capacity, reduce hospital length of stay, decrease complications, and improve functional independence and quality of life [1]-[3].

Despite the well-established benefits of early rehabilitation, its implementation remains limited in routine clinical practice. Previous studies have reported that fewer than 10% of critically ill patients are mobilized beyond the bedside [4]. Multiple barriers contribute to this scenario, including patient-related factors such as low motivation, anxiety, and clinical instability; organizational barriers including limited staffing and insufficient professional training; and hospital infrastructure constraints. Moreover, rehabilitation programs still rely predominantly on conventional functional assessments based on manual repetition counting, external timing devices, and continuous supervision by healthcare professionals. These limitations increase the evaluator`s workload, introduce measurement variability, reduce protocol standardization, and hinder the personalization of rehabilitation according to each patient's functional status [5].

Recent advances in computer vision, artificial intelligence, and virtual reality have created new opportunities to develop intelligent rehabilitation technologies that automate functional assessment through markerless motion capture and real-time biomechanical analysis. Among these technologies, MediaPipe Pose enables accurate estimation of human body posture from conventional RGB cameras, eliminating the need for wearable sensors or optical marker systems. Nevertheless, few digital platforms integrate computer vision, non-immersive virtual reality, automated performance analysis, and structured clinical data management into a unified solution specifically designed for hospital cardiorespiratory rehabilitation. To the best of our knowledge, no previous platform has combined markerless computer vision, non-immersive virtual reality, automated functional assessment, and structured clinical data management into a single modular architecture specifically designed for hospital cardiorespiratory rehabilitation.

To address this technological gap, we developed EX-READ, a modular digital platform that combines computer vision and non-immersive virtual reality to automate functional assessment during hospital cardiorespiratory rehabilitation. The platform integrates standardized functional assessment protocols, real-time motion capture using MediaPipe Pose, automatic performance analysis, and structured clinical data management within a scalable architecture designed for future integration with physiological monitoring systems, wearable sensors, artificial intelligence algorithms, and clinical decision-support tools.

This study presents the design, development, modular architecture of the computational platform, and preliminary technical verification of EX-READ as a computer vision-based platform for automated functional assessment in hospital cardiorespiratory rehabilitation.

II. Materials and Methods

A. Study Design

This study is a technological development study focused on the design, implementation, and preliminary technical verification of EX-READ, a non-immersive virtual reality modular computational platform developed to support functional assessment during hospital cardiorespiratory rehabilitation. EX-READ implements two standardized functional assessment protocols through a modular execution engine capable of automatic motion processing, real-time performance analysis, and structured clinical data management. The modular computational platform was conceived as the first technological module of a broader digital rehabilitation ecosystem, enabling future integration with physiological monitoring systems, wearable sensors, artificial intelligence algorithms, and clinical decision-support tools.

B. Modular computational platform Development Process

The development of EX-READ was carried out by a multidisciplinary team comprising physiotherapists, biomedical engineers, computer scientists, and software developers, following an iterative, incremental software development process. The workflow included five sequential stages: (i) requirements analysis; (ii) system architecture design; (iii) modular computational platform implementation; (iv) integration and internal verification; and (v) iterative refinement of the final application.

The modular computational platform was developed using an iterative user-centered approach, enabling continuous refinement of the application across successive development cycles.

Initially, clinical and technical requirements were established based on current scientific evidence, hospital rehabilitation workflows, and the practical experience of healthcare professionals involved in cardiorespiratory rehabilitation. Functional requirements included patient registration, execution of standardized functional assessment protocols, automatic performance recording, audiovisual feedback, data storage, and report generation. Non-functional requirements included usability, reliability, interoperability, scalability, modular computational platform maintainability, and information security.

C. Motion Capture and Tracking Layer

The Motion Capture and Tracking Layer is responsible for acquiring, detecting, and processing human body movements throughout the functional assessment. This layer establishes the communication between the participant's physical movements and the computational modules of EX-READ, enabling automated assessment without the need for wearable sensors, external markers, or manual repetition counting.

Human motion capture is performed using MediaPipe Pose (Google LLC) [6], a real-time computer vision framework that estimates three-dimensional human body posture from RGB images. The framework detects 33 anatomical landmarks distributed across the head, trunk, upper

limbs, and lower limbs, allowing continuous tracking of body segments during exercise execution.

During each assessment, image frames are continuously acquired by a conventional RGB camera positioned in front of the participant. Each frame is processed in real time by the MediaPipe Pose model [7], which identifies the anatomical landmarks and estimates their spatial coordinates. These data are subsequently transferred to the EX-READ processing pipeline, where they are used to monitor body movement throughout the assessment.

To improve robustness, the tracking layer monitors confidence and temporal consistency between consecutive frames. The default confidence parameter is 0.5; frames with confidence ≤0.5 are ignored for joint-angle calculation and repetition-state transitions until confidence exceeds 0.5. Robustness under hospital conditions, including occlusion, lighting, clothing, medical devices, use of support, and out-of-plane movement, remains to be evaluated.

The extracted kinematic information is subsequently transmitted to the Functional Assessment Core Engine, where the anatomical landmarks are analyzed according to the selected assessment protocol. The processing pipeline automatically identifies movement cycles, estimates joint displacement, recognizes valid repetitions, and generates the performance variables required for each functional test.

The modular implementation of the Motion Capture and Tracking Layer enables the future incorporation of alternative motion-sensing technologies, including depth cameras, inertial measurement units (IMUs), wearable sensors, and multimodal physiological monitoring devices, without requiring modifications to the remaining modular computational platform architecture.

Based on these specifications, the modular computational platform architecture was designed using a modular structure to facilitate future expansion and integration with additional technological resources.

D. Internal Verification and Modular Computational Platform Refinement

Following platform implementation, the first functional prototype of EX-READ underwent internal technical verification by the multidisciplinary development team. This stage aimed to verify platform functionality, operational stability, interface consistency, navigation workflow, and correct execution of the programmed protocol logic under simulated operating conditions. No hospitalized participants were included, and the automated outputs were not compared with measurements obtained by an independent physiotherapist or a reference motion-capture system. Therefore, this stage was not designed to establish measurement error, agreement, repeatability, or clinical validity and should not be interpreted as a clinical or measurement-validation study.

The verification process comprised systematic execution of all modular computational platform functionalities, including configuration of assessment protocols, user interaction, motion tracking, timing control, automatic repetition counting, real-time audiovisual feedback, and data recording. Each software module was inspected and subsequently evaluated after full system integration to ensure proper communication among components and consistent behaviour of the modular computational platform during complete assessment sessions.

Based on the issues identified during internal verification, two iterative refinement cycles were performed. These refinement stages included optimizing the graphical user interface, improving navigation flow, adjusting motion-tracking parameters, refining exercise execution logic, enhancing timing synchronization, and optimizing real-time visual feedback. The modular computational platform was re-evaluated after each refinement cycle to verify the effectiveness of the implemented modifications and to ensure system stability.

This process produced the final functional version of EX-READ, providing a stable, standardized, and modular computational framework for functional assessment using non-immersive virtual reality. This version constitutes the technological basis for subsequent usability, feasibility, and clinical validation studies.

III. RESULTS

A. EX-READ Modular computational platform Architecture

The final version of EX-READ incorporates a complete computational framework for automated functional assessment, including patient registration, RGB camera configuration, MediaPipe Pose-based motion capture, automatic repetition counting, real-time timing control, joint angle estimation, standardized execution of the 2-Minute Step Test and 30-Second Sit-to-Stand Test, structured clinical data management, automatic export of assessment results to Microsoft Excel®, and generation of clinical performance reports. Together, these functionalities establish a comprehensive modular platform capable of supporting standardized functional assessment in hospital rehabilitation. (Fig. 1).

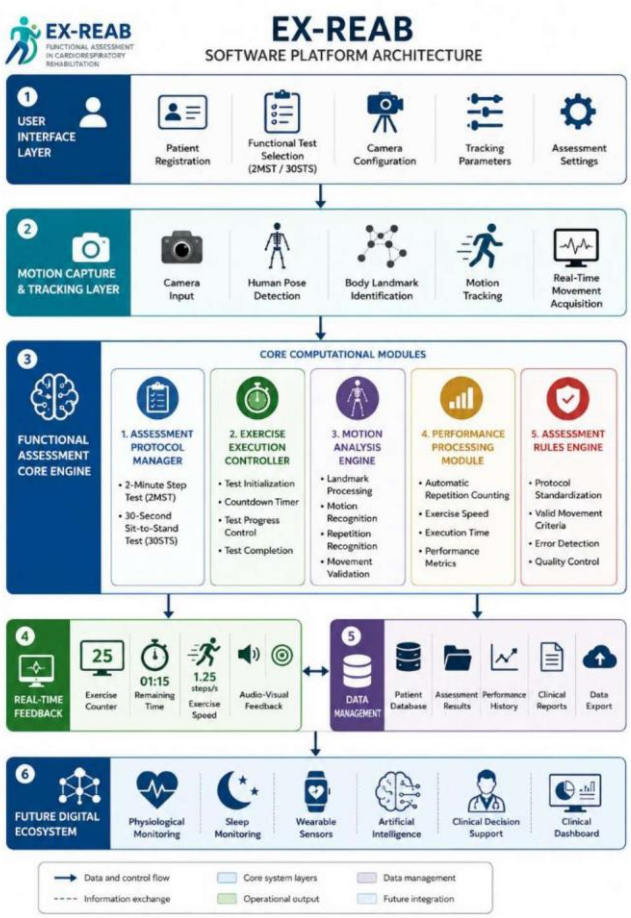


Fig. 1A. Overall architecture of the EX-READ modular computational platform. The system was designed as a modular digital platform comprising six functional layers: User Interface, Motion Capture and Tracking, Functional Assessment Core Engine, Real-Time Feedback, Data Management, and a Future Digital Ecosystem. The modular architecture enables standardized functional assessment through non-immersive virtual reality while supporting future integration with physiological monitoring systems, wearable sensors, artificial intelligence algorithms, and clinical decision-support tools.

The modular computational platform architecture was designed following a modular approach, enabling independent development and maintenance of each functional component while preserving communication among all modular computational platform layers. This architecture also allows future incorporation of additional digital health technologies without requiring substantial structural modifications (table I).

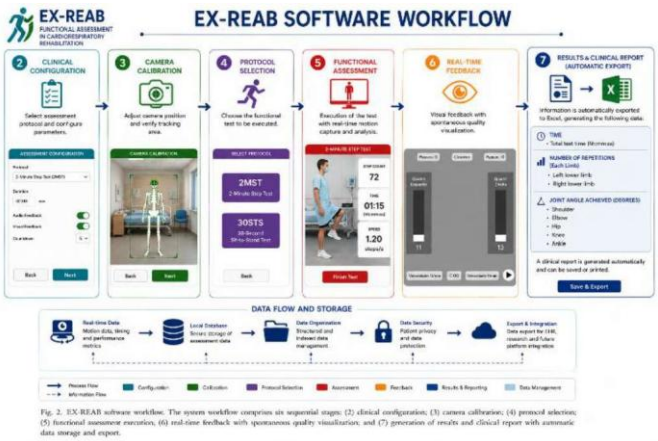
TABLE I. COMPONENTS AND FUNCTIONALS

Module	***Main Function***	***Output***
User Interface Layer	Patient registration, protocol selection, camera configuration, assessment setup	Configured assessment session
Motion Capture and Tracking Layer	RGB image acquisition, MediaPipe Pose processing, detection of 33 anatomical landmarks	Real-time body landmark coordinates
Motion Analysis Pipeline	Joint angle estimation, movement trajectory analysis, biomechanical threshold evaluation	Movement recognition and validated repetitions
Functional Assessment Core Engine	Protocol management, execution control, performance processing, assessment rules	Standardized functional assessment metrics
Real-Time Feedback Layer	Real-time visualization of timing, repetitions, movement status, and protocol progression	Immediate visual feedback to clinician and participant
Data Management Layer	Data Management Layer	Data Management Layer
Clinical Report Module	Automatic generation of standardized assessment reports	Clinical performance report
Future Digital Ecosystem	Integration with wearable sensors, physiological monitoring, sleep assessment, artificial intelligence, and clinical decision-support systems	Scalable digital rehabilitation ecosystem

Abbreviations: RGB, Red-Green-Blue; AI, Artificial Intelligence; EX-READ, Computer Vision and Non-Immersive Virtual Reality Platform for Automated Functional Assessment.

B. User Interface Layer

The user interface layer provides the interaction between healthcare professionals and the modular computational platform (Fig. 2). Before each assessment, users can configure the functional protocol, select the assessment duration, adjust camera parameters, and define tracking settings through an intuitive graphical interface.



The information is automatically exported to excel:total assessment time; number of repetitions; lower-limb movement counts; joint angle measurements; protocol identification; performance metrics generated during assessment.

C. Motion Analysis Pipeline (Fig. 3 and 4)

After landmark detection, the EX-REAB motion analysis pipeline processes MediaPipe Pose coordinates to identify movement cycles according to each functional protocol. For the 30STS, shoulder, hip, and knee landmarks are used to calculate the hip angle; for the 2MST, hip, knee, and ankle landmarks are used to calculate the knee angle. Repetitions are validated by protocol-specific state-transition rules derived from the literature [8], [9]. In the 2MST, the states are resting (<130°), transition (100°-130°), and reached (<100°), with a new step counted only after return to resting. In the 30STS, the states are resting (<120°), transition (120°-150°), and reached (>150°). Incomplete sequences are not counted, postural compensations are not independently classified, and threshold sensitivity was not evaluated at this development stage.

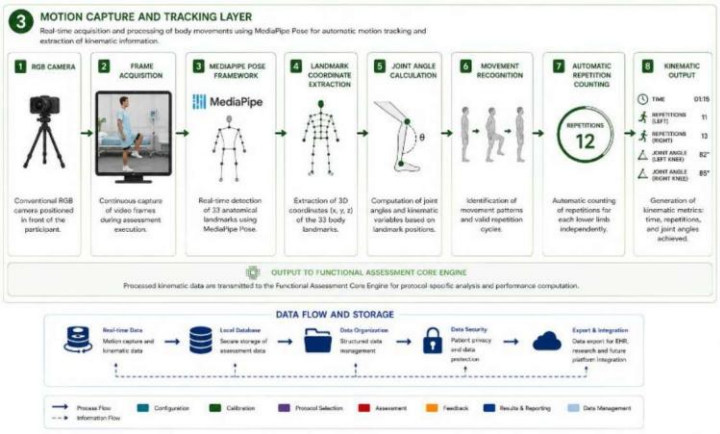


Fig. 3. Motion Capture and Tracking Layer workflow. The layer comprises eight sequential processes: (1) RGB camera; (2) frame acquisition; (3) MediaPipe Pose framework; (4) landmark coordinate extraction; (5) joint angle calculation; (6) movement recognition; (7) automatic repetition counting; and (8) generation of kinematic output. The processed data are transmitted to the Functional Assessment Core Engine for protocol-specific analysis and performance computation.

D. Functional Assessment Core Engine

The Functional Assessment Core Engine represents the central computational module of the EX-REAB platform, coordinating all processes required for automated functional assessment. This module receives the kinematic information generated by the Motion Capture and Tracking Layer and executes the computational routines responsible for protocol management, movement interpretation, performance analysis, and generation of assessment outcomes.

The Core Engine operates through five integrated computational components: Assessment Protocol Manager, Exercise Execution Controller, Motion Analysis Engine, Performance Processing Module, and Assessment Rules Engine (Fig. 3). Together, these components ensure standardized execution of functional assessment protocols while maintaining synchronization among all modular computational platform layers.

The Assessment Protocol Manager initializes the selected functional protocol, configures protocol-specific parameters, and controls the execution logic of the 2-Minute Step Test (2MST) and the 30-Second Sit-to-Stand Test (30STS). Once the protocol is initiated, the Exercise Execution Controller supervises test progression by managing countdown timing, exercise duration, protocol transitions, and test completion according to predefined assessment rules.

The Motion Analysis Engine continuously processes the anatomical landmarks provided by MediaPipe Pose to identify movement cycles, recognize valid repetitions, and

verify whether the participant's movements satisfy the biomechanical criteria established for each assessment protocol. Subsequently, the Performance Processing Module automatically computes clinically relevant variables, including repetition count, execution time, movement cadence, and protocol-specific performance indicators.

Finally, the Assessment Rules Engine applies predefined decision rules to ensure protocol standardization, identify execution errors, verify movement quality, and maintain consistency throughout the assessment. This modular organization enables independent maintenance and future expansion of each computational component while preserving seamless communication with the remaining modular computational platform layers.

E. Real-Time Feedback

The Real-Time Feedback Layer was designed to provide continuous visual information to both healthcare professionals and participants throughout the assessment process. By continuously receiving processed data from the Functional Assessment Core Engine, the system updates performance indicators instantaneously, allowing immediate monitoring of exercise execution.

During assessment, the graphical interface dynamically presents protocol progression, elapsed time, repetition count, exercise cadence, and task status. In addition, visual indicators representing lower-limb movement are continuously updated according to the detected joint displacement, enabling immediate identification of movement symmetry and execution quality.

This real-time interaction facilitates participant engagement while allowing clinicians to supervise protocol execution without manual counting or external timing devices. The feedback architecture may reduce operator workload and support standardized assessment procedures; however, these effects were not measured in the present study.

The modular implementation of the feedback layer enables future incorporation of additional visualization resources, including biomechanical indicators, physiological variables, performance dashboards, and adaptive feedback strategies based on artificial intelligence algorithms.

F. Data Management

The Data Management Layer is responsible for automatic storage, organization, and export of all information generated during functional assessment. Throughout test execution, performance variables produced by the Motion Analysis Pipeline and Functional Assessment Core Engine are continuously transferred to the local database without requiring user intervention.

Following completion of each assessment, EX-REAB automatically generates a structured dataset containing protocol identification, total assessment time, number of repetitions performed, lower-limb movement metrics, joint angle measurements, and protocol-specific performance indicators. These data are automatically exported to Microsoft Excel®, allowing subsequent statistical analysis and long-term monitoring of patient progression.

The generated datasets provide standardized quantitative information that can be integrated into clinical records, research databases, and future digital health platforms. The modular architecture also supports future interoperability with electronic health records (EHRs), cloud-based repositories, wearable monitoring systems, and clinical decision-support platforms.

G. Modular computational platform Capabilities

The final functional version of EX-REAB provides an implemented computational platform capable of performing the complete automated functional-assessment workflow, from patient registration and protocol configuration to motion capture, movement recognition, real-time performance analysis, data storage, and automatic export of clinical results. These functionalities demonstrate technical implementation and establish the technological basis for subsequent usability, feasibility, and clinical validation studies.

IV. DISCUSSION

The present study describes the design and development of EX-REAB, a modular digital platform that integrates computer vision, non-immersive virtual reality, automated motion analysis, and structured clinical data management for functional assessment during hospital cardiorespiratory rehabilitation. To the best of our knowledge, this is one of the first platforms specifically designed to combine these technologies within a unified architecture for hospital rehabilitation.

Beyond its technical implementation, EX-REAB was conceived to address practical barriers that limit routine functional assessment in hospital rehabilitation (Table II). EX-REAB may simplify workflow and reduce manual tasks, but effects on workload, staffing requirements, and evaluator variability were not measured in the present study. Contactless RGB-based capture may also offer potential biosafety advantages. Automatic export of assessment time, repetition count, joint angles, and performance metrics may facilitate clinical documentation, longitudinal follow-up, and future integration with digital health systems, although their validity and clinical impact require formal evaluation.

TABLE II. Potential Advantages of EX-READ Compared with Conventional Functional Assessment*

Conventional assessment	***EX-READ platform***
Manual repetition counting	Automatic repetition counting
External stopwatch	Automatic timing
Visual observation only	Computer vision-based motion analysis
Subjective movement evaluation	Algorithm-derived kinematic analysis*
Greater evaluator workload	Potential reduction in manual workload*
Possible need for two professionals in some situations	Potential to support assessment by a single trained evaluator*
Contact with shared assessment equipment	Contactless assessment using RGB camera
Manual data recording	Automatic data export to Excel

**Potential advantages were not measured and require confirmation in future clinical validation studies.

Another important contribution of EX-READ is its modular computational platform architecture, which was intentionally designed to support future technological expansion. Unlike standalone rehabilitation applications, the platform allows independent incorporation of additional computational modules without requiring substantial modifications to the existing framework. This design facilitates future integration with wearable physiological sensors, sleep monitoring systems, artificial intelligence algorithms, electronic health records, and clinical decision-support tools, supporting the evolution of EX-READ into a comprehensive digital ecosystem for precision rehabilitation.

The present work focused on modular computational platform development and preliminary technical verification and therefore does not establish validity, reliability, measurement error, agreement, repeatability, usability, or clinical effectiveness. Future studies should compare EX-READ-derived repetition counts, execution times, and joint-angle measurements with simultaneous measurements obtained by trained physiotherapists and/or a reference motion-capture system in participants with different mobility levels, including analyses of absolute error, agreement, and repeatability. Future validation should also evaluate threshold sensitivity and performance under representative hospital conditions.

In addition, future versions of EX-READ will incorporate continuous physiological monitoring, sleep assessment, wearable sensing technologies, and artificial intelligence-based predictive models to support personalized rehabilitation strategies in hospital environments.

Therefore, EX-READ establishes the technological foundation for intelligent and scalable digital rehabilitation systems and for future clinically validated applications in hospital rehabilitation settings.

V. CONCLUSION

EX-READ provides a computational framework for automated functional assessment in hospital rehabilitation by combining markerless computer vision, non-immersive virtual reality, standardized computational workflows, and structured digital data management. The present study demonstrates technological development and preliminary technical verification under simulated conditions. Clinical validity, reliability, error, agreement, repeatability, usability, and workflows, and structured digital data management within a unified modular architecture. This technological foundation has the potential to support the next generation of intelligent rehabilitation systems that deliver scalable, objective, and data-driven physiotherapy in hospital settings.

REFERENCES

[1] C. J. Tipping et al., "The effects of active mobilisation and rehabilitation in ICU on mortality and function: A systematic review," Intensive Care Med., vol. 43, no. 2, pp. 171–183, Feb. 2017.

[2] J. W. Devlin et al., "Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep e825 disruption in adult patients in the ICU," Crit. Care Med., vol. 46, no. 9, pp. –e873, Sep. 2018.

[3] C. L. Hodgson et al., "Expert consensus and recommendations on safety criteria for active mobilization of mechanically ventilated critically ill adults," Crit. Care, vol. 26, no. 1, pp. 1–18, 2022.

[4] E. E. Aquim et al., "Brazilian guidelines for early mobilization in intensive care unit," Rev. Bras. Ter. Intensiva, vol. 31, no. 4, pp. 434–443, 2019.

[5] A. R. Lira Ribeiro et al., "Barreiras à mobilização precoce percebidas pela equipe multiprofissional de uma unidade de terapia intensiva," Fisioter. Bras., vol. 25, no. 3, pp. 1456–1468, 2024.

[6] Google, "MediaPipe Pose Landmarker," Google for Developers. [Online]. Available:

https://developers.google.com/edge/mediapipe/solutions/vision/pose_landmarker. [Accessed: Jul. 15, 2026].

[7] V. Bazarevsky et al., "BlazePose: On-device real-time body pose tracking," arXiv:2006.10204, 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[8] R. W. Bohannon and R. H. Crouch, "Two-Minute Step Test of Exercise Capacity: Systematic Review of Procedures, Performance, and Clinimetric Properties," J Geriatr Phys Ther, vol. 42, no. 2, pp. 105-112, 2019, doi:10.1519/JPT.0000000000000164.

[9] A. A. Morita et al., "Best protocol for the sit-to-stand test in subjects with COPD," Respir Care, vol. 63, no. 8, pp. 1040-1049, 2018, doi:10.4187/respcare.05100.

A Multimodal Assistive Interface for Wheelchair Control Using Electrooculography, Capacitive and Inertial Sensing: Initial Development and Pilot Functional Evaluation

1st Souza, A. O.
Faculty of Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Brazil
arthursouzaoliveira250@gmail.com

2nd Sousa, M. L. L.
Faculty of Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Brazil
maria.lisboa.sousa22@gmail.com

3rd Sales I. Q.
Faculty of Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Brazil
isasales40@gmail.com

4th Figueiredo, L. E. T. C.
Faculty of Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Brazil
cassianolynnda@gmail.com

5th Fonseca, M. C. P.
Faculty of Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Brazil
conceicao@ufpa.br

6th Miziara, I. M.
Faculty of Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Brazil
immiziara@ufpa.br

***Abstract*—Independent mobility is essential for individuals with severe motor impairments, although conventional wheelchair interfaces may not be suitable for users with limited upper-limb control. This work presents the initial development and pilot functional evaluation of a multimodal interface for assistive wheelchair control using electrooculography (EOG), capacitive sensing, and inertial sensing. EOG signals generate left and right commands, the capacitive sensor provides a forward command, and an inertial measurement unit monitors lateral neck inclination. Two ESP32 microcontrollers communicate through MQTT over Wi-Fi, and the received commands are converted into motor actuation through an H-bridge driver. Preliminary EOG acquisition and functional tests were performed by one of the authors as an initial pilot evaluation of the proposed architecture. The results were qualitatively analyzed to verify the functional integration and interaction among the sensing, processing, communication, and actuation modules. The preliminary findings support the feasibility of the proposed proof-of-concept architecture under controlled bench conditions, while quantitative evaluation of classification performance, communication characteristics, and system robustness is planned as part of future experimental stages. The present study does not constitute quantitative validation of wheelchair performance, usability, or safety.**

***Index Terms*—Assistive Technology, Electrooculography (EOG), Human–Machine Interface, Capacitive Sensing, Inertial Sensing**

I. Introduction

Mobility is a fundamental aspect of autonomy, social participation, and quality of life for people with motor disabilities. Partial or complete loss of locomotion compromises daily activities, reduces functional independence, and may lead to significant physical, emotional, and social impacts. This condition is observed in individuals with paraplegia, quadriplegia, spinal cord injuries, neuromuscular diseases, and other neuromotor disorders, which may present different degrees of impairment affecting upper- and lower-limb control [1], [3], [4].

In many of these cases, powered wheelchairs represent an important solution for increasing user independence. However, conventional control systems, typically based on joysticks, are not always suitable for individuals with severe upper-limb impairments. Users with reduced muscle strength, impaired fine motor coordination, or limited voluntary movements may experience difficulties operating these devices accurately and safely, highlighting the need for alternative control interfaces [2].

Given these limitations, several studies have investigated human–machine interfaces capable of exploiting preserved motor or physiological functions. Among these alternatives, electrooculography (EOG) has attracted considerable attention because it enables the detection of eye movements based on variations in the corneo-retinal potential. Since ocular control may remain preserved even in individuals with severe motor impairments, EOG represents a promising approach for generating commands in assistive mobility systems [3], [5].

Despite the potential of ocular control, relying on a single source of information may introduce challenges related to reliability, user adaptation, and operational safety. Involuntary eye movements, acquisition noise, blinks, and signal fluctuations may hinder command interpretation. Furthermore, users may exhibit different preserved motor abilities, making it advantageous to combine multiple input modalities. In this context, multimodal approaches have been investigated to integrate different signals and sensors into a unified control

strategy [2].

Residual body movements and complementary sensors can further expand the interaction possibilities of assistive systems. Capacitive sensors can serve as low-effort activation interfaces, while inertial sensors can contribute to postural monitoring and the detection of unsuitable operating conditions. Integrating these sources of information allows command generation and monitoring functions to be distributed across multiple inputs, resulting in an interface that is more adaptable to the user's functional capabilities.

Based on this motivation, this work presents the initial development and pilot functional evaluation of a multimodal human–machine interface for assistive wheelchair control. The proposed architecture integrates EOG for directional commands, capacitive sensing for forward movement, and inertial sensing for monitoring lateral neck inclination. The primary objective of this initial stage is to verify the functional integration of the sensing, processing, communication, and motor-actuation modules and to assess qualitatively whether the proposed interaction strategy operates as intended. Therefore, the present study is intentionally framed as a preliminary proof-of-concept investigation. Quantitative evaluation of classification performance, false activations, response time, communication reliability, and user-related performance is reserved for subsequent experimental stages and future work.

II. Materials and Methods

The proposed system was developed to integrate different sources of information related to the user's intentions and the operating conditions of the device, supporting command generation and postural monitoring within the prototype. The overall architecture of the proposed system is illustrated in Fig. 1.

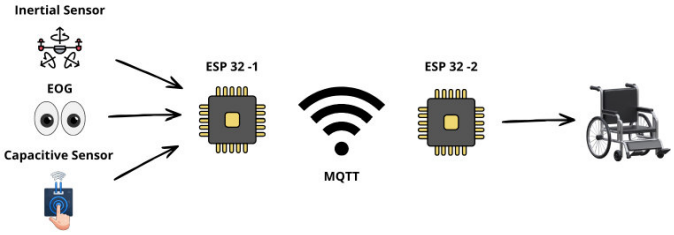


Fig. 1. Project schematic.

As illustrated in the schematic, the sensors acquire the information required for system operation. The acquired signals are transmitted to an ESP32 microcontroller, where data processing is performed, while a second ESP32 receives the generated commands and controls the actuators. Communication between the microcontrollers is established through the Message Queuing Telemetry Transport (MQTT) protocol over a Wi-Fi network.

The proposed architecture combines three input modalities. Horizontal eye movements generate LEFT and RIGHT commands, capacitive activation generates the FORWARD command, and the inertial sensor monitors lateral neck inclination. When no valid movement command is detected, or when the predefined interruption condition is reached, the system returns to the STOP state. The resulting command is then transmitted to the motor-control module.

A. Pilot Evaluation Scope

The experiments reported in this study correspond to an initial pilot stage of the proposed system development. The primary purpose of this stage was to verify the functional integration of the sensing, signal-processing, communication, and motor-actuation modules under controlled bench conditions. Accordingly, the evaluation was exploratory and qualitative rather than a quantitative performance assessment.

The preliminary EOG acquisition and classification tests were conducted with one author-participant, without external volunteers. The experiments were intended to verify whether intentional horizontal eye movements produced distinguishable signal patterns and whether these patterns resulted in the expected state transitions. Similarly, the multimodal integration, MQTT communication, and motor-actuation stages were evaluated by observing whether the implemented commands were correctly processed, transmitted, and translated into the expected prototype responses.

No statistical inference or generalized performance claim is intended from these experiments. The observations obtained in this phase were used to verify the expected qualitative behavior of the system and to identify aspects that require further investigation. Quantitative characterization, including classification accuracy, false-activation rate, response time, communication latency, packet loss, inertial-estimation error, and evaluation with multiple participants, is planned for subsequent stages of the project.

Table I summarizes the specific questions addressed by each pilot evaluation and distinguishes them from the quantitative questions reserved for future work.

B. Inertial Sensor

The MPU-9250 module, which integrates a three-axis accelerometer and a three-axis gyroscope, was employed to monitor the user's lateral neck inclination during prototype operation. The inertial sensor enables the detection of postures that exceed the predefined operational threshold, as illustrated in Fig. 2.

Figure 2 illustrates the inclination range adopted for user postural monitoring. Based on preliminary experiments, an angular threshold of $\pm 50^{\circ}$ relative to the neutral neck position was adopted as an operational criterion for the prototype. When the measured inclination remains within this range, movement commands are enabled; otherwise, motor actuation is interrupted. This threshold was selected for functional pilot testing and was not derived from a clinical or ergonomic safety assessment.

Data acquisition from the inertial sensor was performed through the Inter-Integrated Circuit (I^2C) communication protocol using an ESP32 microcontroller. Before operation, the gyroscope was calibrated to compensate for measurement

TABLE I
SCOPE OF THE PILOT FUNCTIONAL EVALUATION.

Module	Pilot question	Evidence collected	Not evaluated quantitatively
EOG	Whether horizontal gaze produced the expected signal patterns and state transitions.	Representative EOG trace and observed state-machine transitions.	Accuracy, sensitivity, false activations, robustness and generalization.
Capacitive sensor	Whether activation generated the intended forward command.	Qualitative activation and command response.	Repeatability, sensitivity and environmental robustness.
IMU	Whether inclination outside the prototype range triggered interruption logic.	Observed threshold-dependent state response.	Angular accuracy, drift and clinical/ergonomic validity.
MQTT	Whether commands could be transmitted and interpreted by the actuation module.	Published and received command messages.	Latency, jitter, packet loss, and security performance.
Motors	Whether received commands were mapped to the intended motor actions.	Bench observation of motor-command mapping.	Response time, trajectory, load behavior and navigation performance.

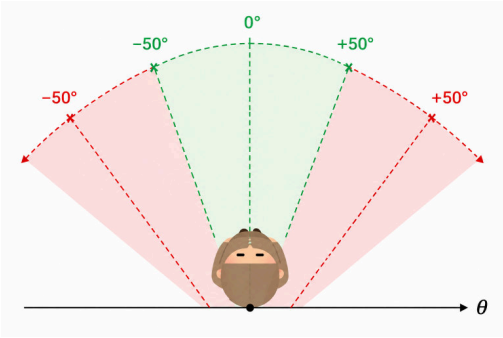


Fig. 2. Lateral neck inclination adopted for postural monitoring.

offsets, since inertial sensors based on Microelectromechanical Systems (MEMS) technology are susceptible to systematic errors, measurement noise, and angular drift, particularly when orientation is estimated through numerical integration [4], [5].

During the calibration procedure, the sensor was maintained at rest, and the average of 500 samples was computed to estimate the initial measurement offset [4]. After offset compensation, the angular velocity measurements were sampled at approximately 60 Hz and numerically integrated by the ESP32 to estimate the lateral neck inclination. The estimated inclination was then compared with the predefined operational threshold. Because the estimate is obtained by direct gyroscope integration, cumulative drift may occur over time and was not quantitatively characterized in the present pilot study.

C. Capacitive Sensor

The capacitive sensor was employed as an interface to detect the user's intention to initiate forward motion. In assistive mobility systems, individuals with paraplegia, quadriplegia, or other motor impairments may experience difficulties operating conventional interfaces, such as joysticks, thereby motivating the adoption of alternative activation methods that require minimal physical effort [2], [6].

Its operating principle is based on variations in capacitance produced by the proximity or contact of the human body with the sensor-sensitive surface. These variations are interpreted by the embedded system as an activation signal, enabling the detection of a voluntary command without requiring significant mechanical force [7], [8].

In the proposed architecture, the capacitive sensor is not used as a general wheelchair enable signal and does not interfere with the directional commands generated by the EOG module. Instead, it provides an independent command associated exclusively with forward movement. Thus, sensor activation is interpreted by the system as the user's intention to move the wheelchair forward.

Accordingly, each sensing modality performs a specific function within the multimodal interface. The capacitive sensor generates the forward command, the EOG module identifies the lateral directional commands, and the inertial measurement unit (IMU) monitors the user's postural inclination. This functional separation enables the different sensing modalities to operate in a complementary manner, increasing the flexibility of the proposed control strategy.

The sensor position should be adjusted according to the user's functional capabilities. For individuals with paraplegia, activation can be performed using body regions with greater preserved motor function, such as the hands or arms. For users with quadriplegia, the sensor may be positioned near the head, shoulder, chin, or another region with preserved voluntary control [9].

D. EOG

Figure 3 illustrates the functional architecture of the electrooculography (EOG) signal acquisition and conditioning module developed in this work. The module converts the electrical potentials generated by eye movements into signals compatible with embedded processing, allowing their use as one of the input modalities of the proposed multimodal assistive interface.

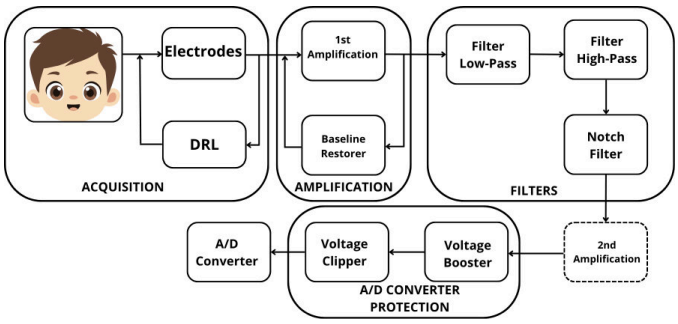


Fig. 3. Functional diagram of the EOG signal acquisition and conditioning module.

Electrooculography was employed because it enables the detection of horizontal eye movements through variations in the corneo-retinal potential. In this configuration, rightward and leftward eye movements produce voltage deflections with opposite polarities, allowing these events to be associated with directional commands. This characteristic makes EOG a suitable alternative for human–machine interfaces intended for users with severe motor impairments, since ocular control is frequently preserved even in conditions involving substantial upper-limb limitations [10], [11], [13], [14].

EOG signals were acquired using surface electrodes positioned around the eyes according to the direction of the ocular movements to be monitored. Because EOG signals exhibit low amplitudes and are susceptible to external interference, the acquired signal was initially processed by a differential amplification stage combined with a *Driven Right Leg* (DRL) active reference circuit. This configuration improves common-mode rejection and stabilizes the electrical reference during signal acquisition [10], [12].

Subsequently, the signal was processed by a baseline restoration stage and active filters. The baseline restorer compensates for slow DC-level variations, whereas second-order high-pass and low-pass filters implemented using the Sallen–Key topology define the frequency band of interest while preserving the components associated with voluntary eye movements. Additionally, a Notch filter was employed to attenuate the 60 Hz power-line interference commonly observed in biopotential acquisition systems.

The conditioned signal was then subjected to a second amplification stage to increase the amplitude of the useful EOG components. Before analog-to-digital conversion, a protection circuit shifted the signal to a positive voltage range and limited its amplitude to levels compatible with the analog input of the microcontroller, thereby protecting the embedded acquisition input.

For the prototype implementation, a printed circuit board (PCB) was designed and fabricated. Figure 4 presents the developed PCB for the EOG signal acquisition module.

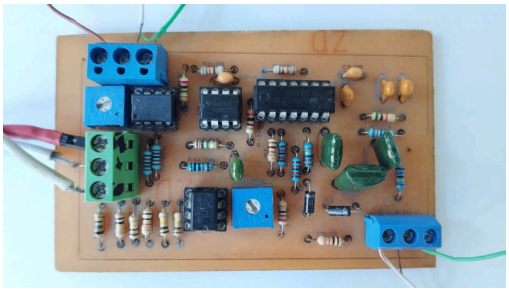


Fig. 4. Printed circuit board developed for the EOG signal acquisition and conditioning module.

For the acquisition of horizontal eye movements, surface electrodes were positioned in the periocular region to capture the potential variations associated with lateral gaze displacement [10]. The electrodes were placed near the outer canthi of the eyes, together with a reference electrode to stabilize the measurements. This configuration was adopted because it provides greater sensitivity to horizontal eye movements, which constitute the directional commands employed by the proposed assistive interface [11].

Figure 5 illustrates the electrode placement adopted during the acquisition experiments.

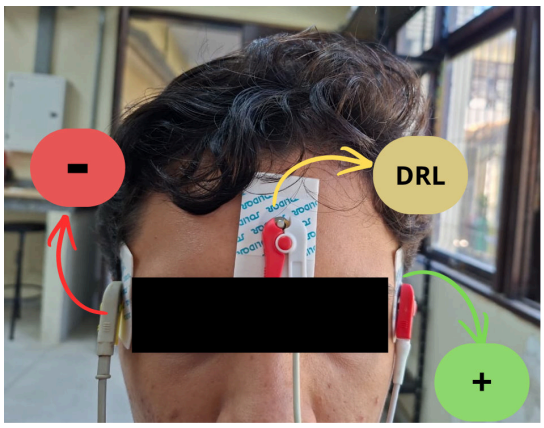


Fig. 5. Electrode placement adopted for horizontal EOG signal acquisition.

The conditioned EOG signal was digitized at approximately 60 Hz using an ADC with 12-bit resolution. The acquisition chain used an amplifier gain of approximately 66, high-pass and low-pass cutoff frequencies of 0.5 Hz and 12 Hz, respectively, and surface Ag/AgCl electrodes. Electrode–skin impedance was not quantitatively measured during the experiments. These parameters are reported to improve reproducibility of the prototype implementation, while the present study does not establish whether they are optimal for classification performance across users.

Preliminary EOG acquisition tests were performed by one of the authors of this study, with no external volunteers involved in the reported experiments. These tests constituted an initial pilot evaluation intended to verify the functionality of the EOG acquisition chain and the response of the state-machine-based classification algorithm. During the tests, the participant performed intentional leftward and rightward eye movements while the EOG signal and the corresponding state transitions were observed.

At this stage, the experiments were exploratory and were not designed for statistical characterization. The number of repetitions was therefore not systematically recorded for quantitative analysis. The observations obtained in this phase were used to verify the expected qualitative behavior of the signal and classification states and to identify aspects of the system that require further investigation.

The reported pilot tests were performed by one of the authors, and no external volunteers were recruited. The experiment was limited to preliminary prototype development and functional verification and did not involve clinical assessment. The authors recognize that the use or publication of identifiable participant images must comply with the applicable institutional and event requirements; therefore, any image used in

the final submission should be accompanied by the appropriate authorization or documentation when required.

E. State-machine-based EOG Classification Algorithm

After signal acquisition and analog conditioning, the digitized electrooculography (EOG) signal was processed by the ESP32 to identify horizontal eye movements and convert them into discrete navigation commands. A state-machine-based approach was adopted because EOG signals are characterized by temporal patterns rather than isolated voltage values.

Initially, the filtered signal was centered with respect to the baseline estimated during the calibration stage. During this procedure, the user maintained the gaze in the neutral position, allowing the reference value b_0 to be estimated. The centered signal was defined as

$$x[n] = v_f[n] - b_0, \tag{1}$$

where $v_f[n]$ denotes the filtered EOG signal and $x[n]$ represents the centered signal. This operation shifts the baseline to approximately zero, facilitating the identification of positive and negative deflections associated with eye movements.

EOG events were identified by analyzing the temporal evolution of the centered signal. To this end, the variation between consecutive samples was computed as

$$\Delta x[n] = x[n] - x[n-1], \tag{2}$$

This variation allows the detection of changes in the signal trend. A transition from an increasing to a decreasing trend characterizes a positive peak, whereas a transition from a decreasing to an increasing trend characterizes a negative peak. Consequently, eye-movement detection depends not only on the instantaneous signal amplitude but also on its temporal behavior.

Figure 6 presents the flowchart of the proposed classification algorithm. The process begins with EOG signal acquisition, followed by digital filtering, baseline centering, and validation of candidate peaks. Based on the polarity of the first detected peak, the movement is classified as either left or right. Subsequently, the algorithm searches for a peak with opposite polarity, which is interpreted as the return of the gaze to the neutral position.

The adopted strategy assumes that each horizontal eye movement consists of two sequential events: gaze displacement and gaze recentering. Accordingly, the first valid peak represents the user's directional intention. A positive peak is interpreted as a left command, whereas a negative peak corresponds to a right command. This convention reflects the electrode configuration adopted in the experiments and can be reversed if the signal polarity is inverted.

After detecting the first peak, the algorithm does not immediately accept a new command. Instead, the state machine enters a waiting state in which it searches for a peak of opposite polarity corresponding to gaze recentering. Once this second peak is detected, the state returns to the neutral condition. This strategy prevents the natural return of the eyes from being incorrectly interpreted as a new directional command.

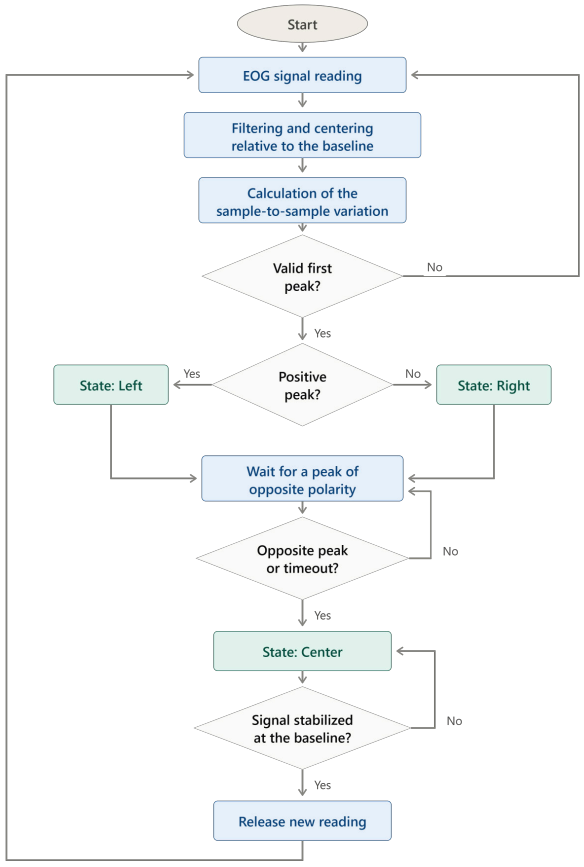


Fig. 6. Flowchart of the state-machine-based EOG classification algorithm.

To reduce false detections caused by noise or small signal fluctuations, amplitude thresholds were incorporated into the algorithm. The directional threshold, denoted by L_{dir}, validates the deflection associated with leftward or rightward gaze displacement, whereas the recentering threshold, denoted by L_{rec}, validates the opposite-polarity deflection corresponding to gaze recentering. Consequently, only significant signal variations are considered during the decision process.

The directional and recentering thresholds were selected empirically during preliminary pilot tests performed by one of the authors. The adopted values were $L_{dir} = 0.40$ and $L_{rec} = 0.15$. These thresholds correspond to the directional gaze displacement threshold and the opposite-polarity gaze recentering threshold, respectively. They were used for functional prototype testing and should not be interpreted as universally validated physiological limits.

Finally, after returning to the center state, the algorithm waits until the signal stabilizes around the baseline before enabling a new detection cycle. This condition reduces the likelihood that residual oscillations produced by the eye movement itself will be interpreted as additional commands.

F. Communication Between Modules Using MQTT

Communication between the embedded modules was required due to the functional partitioning of the system into two ESP32 microcontrollers. The first microcontroller is responsible for sensor data acquisition and processing, whereas the second controls the actuators. This partitioning improves the organization of the prototype architecture by isolating the sensing module from the power stage while facilitating the integration between signal processing and actuation.

Communication between the two modules was established using the MQTT (*Message Queuing Telemetry Transport*) protocol over a Wi-Fi network. MQTT adopts a publish/subscribe communication model in which clients exchange messages organized in topics through a server referred to as a *broker*. This architecture reduces the coupling between devices while requiring low computational and communication overhead, making it widely adopted in embedded systems and Internet of Things (IoT) applications [15]–[17].

In the proposed architecture, both ESP32 microcontrollers operate as MQTT clients. The acquisition module acts as the *publisher*, acquiring and processing the signals from the EOG module, the capacitive sensor, and the inertial measurement unit (IMU). Based on these inputs, the system generates a discrete command representing the current state of the interface.

Figure 7 illustrates the MQTT communication architecture adopted in the proposed system. The acquisition ESP32 processes the sensor data and publishes the generated command to the topic *prototype/wheelchair/command*. The MQTT *broker* forwards this message to the second ESP32, which operates as a *subscriber*. The actuation module then interprets the received command and generates the control signals required by the H-bridge and the drive motors.

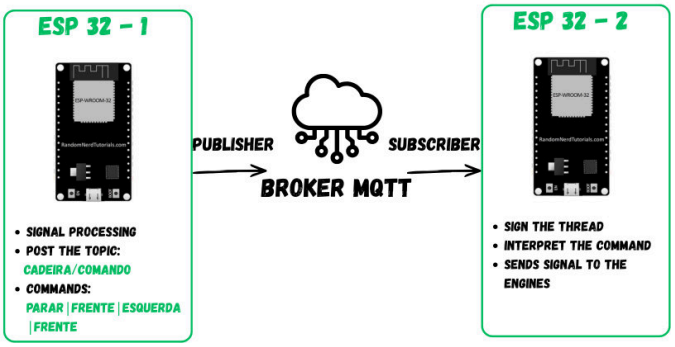


Fig. 7. MQTT communication architecture between the embedded modules.

The processing performed by the acquisition ESP32 results in one of four discrete motor commands: *STOP*, *FORWARD*, *LEFT*, or *RIGHT*. The *FORWARD* command is generated by activation of the capacitive sensor, whereas the *LEFT* and *RIGHT* commands are determined from the EOG signal classification. The *STOP* command is issued when no valid movement command is detected or when the IMU indicates that the predefined operational inclination range has been exceeded. The *CENTER* state used internally by the EOG classifier is not treated as a motor command. Simultaneous conflicting activation of the capacitive and EOG inputs was not systematically evaluated in the present pilot study.

The MQTT Quality of Service (QoS) level 0 was selected because it minimizes communication overhead and latency by eliminating message acknowledgment requirements [15]. Moreover, messages were configured as non-retained, preventing previously transmitted commands from being executed after a possible reconnection.

As a communication protection mechanism, the acquisition ESP32 periodically retransmits the current command, thereby acting as a heartbeat signal for the communication link. At the actuation module, if no valid message is received within the predefined timeout interval implemented in the firmware, the motors are automatically disabled. This behavior provides a functional interruption response to loss of command messages; however, its timing performance and safety effectiveness were not quantitatively validated in this pilot stage.

III. Results and Discussion

The experimental sequence was intentionally organized as a staged functional verification of the prototype. First, the EOG acquisition and state-machine behavior were examined; next, the three sensing modalities were integrated; subsequently, command transmission through MQTT was observed; and finally, the received commands were mapped to motor actuation. This sequence was selected to verify the main functional links of the proposed architecture before undertaking a structured quantitative evaluation.

The results presented in this section correspond to the initial pilot stage of the project and were therefore evaluated qualitatively. The purpose of the experiments was to verify the functional integration of the proposed sensing, processing, communication, and actuation chain under controlled bench conditions. The results should not be interpreted as quantitative estimates of classification accuracy, communication reliability, usability, or wheelchair safety.

A. Preliminary Qualitative Evaluation of Eye Movements

The initial functional evaluation was conducted as a qualitative pilot assessment of the EOG acquisition and classification stages. The purpose of this experiment was not to estimate classification accuracy, but rather to verify whether the proposed acquisition and state-machine strategy produced the expected responses to intentional horizontal eye movements.

After signal conditioning and digitization, lateral gaze displacements produced characteristic voltage variations, consisting of an initial peak associated with gaze displacement followed by an opposite-polarity deflection corresponding to gaze recentering. Figure 8 illustrates a representative EOG signal and the corresponding state transitions observed during the pilot test.

Under the tested condition, the algorithm generated the expected left, right, and center states according to the temporal

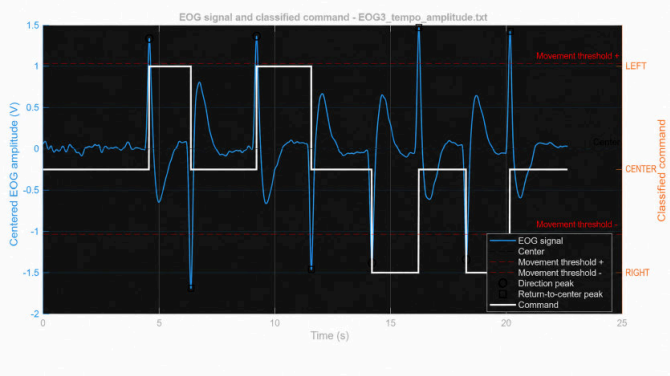


Fig. 8. Representative EOG signal and corresponding state-machine transitions during the preliminary functional test.

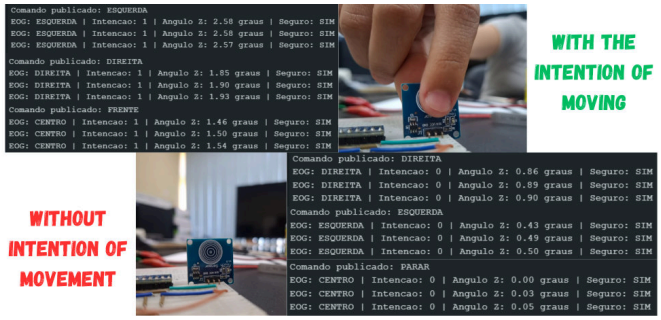


Fig. 9. ESP32 serial monitor during multimodal sensor integration, displaying the estimated IMU angle, the capacitive sensor state and the directional command identified from the EOG signal. In the monitor, *Comando publicado* corresponds to *Published command*, *Intencao* to *Intention*, *Angulo Z* to *Z angle*, and *graus* to *degrees*. The label *Seguro: SIM* indicates that the measured inclination is within the operational range adopted in the firmware. The commands *ESQUERDA*, *DIREITA*, *FRENTE*, *CENTRO*, and *PARAR* correspond to *LEFT*, *RIGHT*, *FORWARD*, *CENTER*, and *STOP*, respectively.

sequence of detected events. This observation provides preliminary qualitative evidence that the proposed state-machine strategy can distinguish the intended movement sequence under controlled conditions. However, because the present evaluation was exploratory and involved a single author-participant, these observations should not be interpreted as estimates of classification accuracy, robustness, or generalization to other users.

A quantitative evaluation involving repeated and labeled trials will be conducted in a subsequent stage of the project to characterize classification accuracy, false activations, response time, and robustness under different operating conditions.

B. Multimodal Sensor Integration

Following the functional evaluation of eye-movement identification, the sensing modalities were integrated to form the proposed multimodal interface. At this stage, the embedded system simultaneously processed three sources of information: the directional commands obtained from the EOG, the forward command generated by the capacitive sensor, and the postural inclination estimated by the inertial measurement unit (IMU).

Figure 9 shows the serial monitor during prototype operation, displaying the combined information processed by the embedded system. The monitored variables include the inclination angle estimated by the IMU, the activation state of the capacitive sensor, and the directional command identified from the EOG signal. The displayed data show concurrent processing of the three sensing modalities during the functional bench test.

The multimodal integration stage was evaluated qualitatively as part of the pilot development phase. The purpose was to verify whether the three sensing modalities could be acquired and processed concurrently within the embedded architecture and whether their respective functions remained distinguishable during operation.

During the experiments, the capacitive sensor operated as an independent input associated with forward movement. When activated, the embedded system interpreted this event as a forward navigation command. The EOG module remained responsible for the lateral directional commands, and the observed states were processed concurrently with the capacitive input during the bench tests.

Simultaneously, the IMU continuously estimated the user's postural inclination. Consequently, the proposed multimodal interface combines sensing modalities with distinct and complementary functions: the capacitive sensor generates the forward command, the EOG provides the lateral directional commands, and the IMU performs postural monitoring.

The serial-monitor observations show that the three sensing modalities were acquired and processed simultaneously by the ESP32. This stage demonstrates the functional integration of command generation and postural monitoring within a single embedded architecture, enabling subsequent evaluation of MQTT communication and motor actuation. It does not, however, establish quantitative performance under repeated or conflicting multimodal inputs.

C. MQTT Communication

The MQTT communication stage was evaluated qualitatively during the pilot phase to verify the functional transmission of commands between the acquisition and actuation modules. Communication between the modules was established using a public HiveMQ server operating as the *broker*. The ESP32 responsible for sensor acquisition and processing acted as the *publisher*, transmitting the commands generated by the multimodal interface to the topic *prototype/wheelchair/command*. The second ESP32, responsible for motor actuation, operated as the *subscriber*, remaining subscribed to the same topic to receive and interpret the transmitted commands.

Figure 10 shows the MQTT communication interface during the experiments, including the active connection to the *broker*, the subscribed topic, the QoS level set to 0, and the exchanged messages.

During the bench experiments, the published messages were received by the actuation module and translated into the corresponding motor commands. The *FORWARD*, *LEFT*, *RIGHT*,

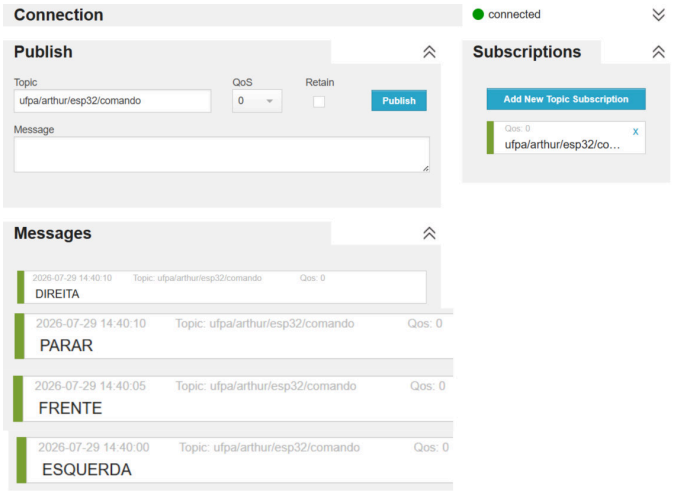


Fig. 10. MQTT communication interface showing QoS 0, active connection to the broker, and published navigation commands during the functional test.

and *STOP* messages produced the command interpretations defined in the firmware under the tested conditions.

The QoS 0 level was adopted because it minimizes communication overhead by eliminating message acknowledgment requirements [15]. Because QoS 0 does not guarantee message delivery, the present implementation should be interpreted as a proof-of-concept communication strategy. Latency, jitter, packet loss, reconnection time, and long-term communication reliability were not quantitatively evaluated in the pilot stage. These parameters will be characterized in subsequent experimental work.

D. Motor Actuation

Following the functional evaluation of MQTT communication, the motor actuation stage was evaluated qualitatively under bench conditions. The purpose of this stage was to verify whether the commands generated by the multimodal interface and transmitted through MQTT were correctly converted into the corresponding H-bridge control signals and motor responses.

Figure 11 illustrates the experimental setup adopted for motor actuation. The actuation ESP32 was connected to the H-bridge inputs, whereas the motors were connected to the corresponding outputs. Based on the received commands, the prototype executed the movements defined by the proposed multimodal control strategy. These tests were performed in a bench configuration and did not include a powered wheelchair under user load.

During the bench tests, the *FORWARD* command, generated when the capacitive sensor was activated and the IMU remained within the adopted operational range, resulted in simultaneous activation of both motors. For the *LEFT* command, the left motor remained disabled while the right motor was activated. Conversely, for the *RIGHT* command, the right motor remained disabled while the left motor was activated.

Whenever no valid command was detected or the IMU identified an inclination outside the adopted operational range, both motors remained disabled according to the implemented interruption logic. This behavior demonstrates the functional response of the motor-control stage to the predefined prototype conditions but does not constitute validation of wheelchair safety.

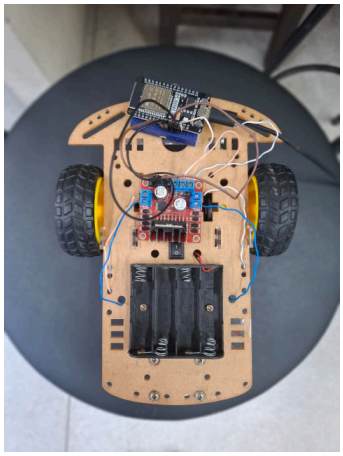


Fig. 11. Experimental setup used for functional bench evaluation of motor actuation through the ESP32 and the H-bridge.

Table II summarizes the motor response associated with each command observed during the experiments.

TABLE II
FUNCTIONAL MOTOR RESPONSE ACCORDING TO THE RECEIVED COMMAND.

Command	Left motor	Right motor	Movement
FORWARD	On	On	Forward movement
LEFT	Off	On	Left turn
RIGHT	On	Off	Right turn
STOP	Off	Off	Stop

Overall, the bench observations indicate that the actuation module converted the digital commands generated by the multimodal interface into the expected motor actions under the tested conditions. These observations demonstrate functional integration between sensing, communication, and actuation, but do not provide quantitative information regarding motor response time, acceleration, trajectory tracking, maneuverability, energy consumption, or performance under wheelchair load.

E. Discussion

The results obtained in this study should be interpreted within the scope of an initial pilot development stage. The primary objective was to verify the functional integration of the proposed multimodal architecture rather than to establish quantitative performance or clinical effectiveness.

Within this scope, the experiments demonstrated that the EOG acquisition and state-machine classification produced the expected qualitative responses to intentional horizontal eye movements, that the capacitive and inertial modalities could be processed concurrently with the EOG input, and that the generated commands could be transmitted through MQTT and

converted into the expected motor-control responses under bench conditions.

The obtained observations are consistent with previous studies that investigated alternative interfaces for assistive mobility, particularly considering the limitations of conventional joystick-based control for users with severe motor impairments [2]. Previous EOG-based approaches have demonstrated the feasibility of wheelchair control through eye movements [13], [14]; however, most of these systems rely on a single input modality for command generation.

The proposed methodology differs by adopting a multimodal interface in which complementary sensing modalities are integrated within a single embedded architecture. Rather than replacing unimodal approaches, the system distributes command generation and postural monitoring across different inputs. This organization is intended to support adaptation to users with different preserved motor abilities, although usability and comparative performance were not evaluated in the present pilot stage.

The present study has important limitations that are directly related to its exploratory scope. The EOG experiment involved only one author-participant and was not designed for statistical assessment of classification performance. Accuracy, confusion matrices, false-activation rates, and robustness against blinking or other physiological artifacts were not evaluated. In addition, gyroscope drift, communication latency, jitter, packet loss, reconnection behavior, and inertial-estimation error were not quantitatively characterized. The $\pm 50^{\circ}$ inclination limit was adopted only as an operational prototype threshold, and the experiments were performed under bench conditions rather than with a powered wheelchair in realistic navigation scenarios.

The simultaneous occurrence of potentially conflicting commands from the capacitive and EOG modalities was also not systematically evaluated. This aspect should be addressed in future experiments because multimodal systems require explicit command-priority and conflict-management rules under concurrent user inputs.

The integration of signal acquisition, embedded processing, MQTT communication, and motor actuation also aligns the proposed architecture with developments in embedded systems and Internet of Things (IoT) applications, where modularity, scalability, and low communication overhead are relevant design considerations [15]–[17]. Nevertheless, the use of wireless communication and a public broker should be considered a characteristic of the current proof-of-concept implementation rather than a validated architecture for safety-critical wheelchair operation.

F. Pilot Study Limitations and Future Quantitative Evaluation

The present study represents an initial pilot stage in the development of the proposed multimodal assistive interface. Accordingly, the experimental results should be interpreted as preliminary qualitative evidence of functional integration rather than as quantitative validation of system performance.

The main limitation of this stage is the absence of a systematic quantitative evaluation. The EOG experiments involved a single author-participant, and no statistical analysis of classification performance was performed. Similarly, communication latency, jitter, packet loss, reconnection behavior, inertial-estimation error, false activations, response time, and long-term system reliability were not quantitatively characterized. The motor-actuation experiments were conducted under bench conditions and did not reproduce the mechanical and environmental conditions of a powered wheelchair operating under user load.

The simultaneous occurrence of potentially conflicting commands from the capacitive and EOG modalities was also not systematically evaluated. This aspect should be addressed in future experiments because multimodal systems require explicit command-priority and conflict-management rules under concurrent user inputs. Debounce behavior, command duration, release conditions, and emergency-stop behavior likewise require dedicated evaluation.

The next experimental stage will quantify the performance of each functional link identified in the present pilot study. For the EOG modality, the planned protocol will include repeated labeled trials with predefined ground-truth commands and multiple participants, enabling the calculation of accuracy, sensitivity by direction, confusion matrices, false activations per unit time, missed commands, and response time. The influence of blinking, spontaneous eye movements, and signal artifacts will also be investigated.

For the inertial modality, future experiments will characterize angular drift and estimation error over time and will investigate the relationship between the selected operational threshold and user posture. If necessary, sensor-fusion methods combining gyroscope and accelerometer information will be evaluated. The adopted $\pm 50^{\circ}$ threshold will be treated as an experimentally adjustable parameter rather than as a clinically or ergonomically validated safety limit.

For the communication layer, future tests will measure end-to-end latency, jitter, packet loss, reconnection behavior, and effective stop time after communication interruption. Because the present prototype uses QoS 0 and a public broker only for proof-of-concept bench testing, future implementations will investigate local or authenticated communication, command expiration, sequence control, and additional fail-safe mechanisms appropriate to a motorized assistive device.

Finally, controlled experiments will progressively move from laboratory characterization to multi-participant evaluation and, subsequently, to tests of the integrated system under realistic wheelchair operating conditions. This staged approach is intended to provide increasingly strong evidence regarding robustness, usability, practical applicability, and safety without attributing conclusions to the current pilot study that the available data cannot support.

IV. Conclusion

This work presented the initial development and pilot functional evaluation of a multimodal human–machine interface for assistive wheelchair control based on electrooculography,

capacitive sensing, inertial sensing, MQTT communication, and H-bridge motor actuation.

The preliminary experiments demonstrated the functional integration of the sensing, processing, communication, and motor-actuation modules under controlled bench conditions. The results obtained in this stage were intentionally analyzed qualitatively, with the primary purpose of verifying the expected operation of the proposed architecture and identifying aspects requiring further experimental investigation.

Because the present study represents an initial pilot stage, the results should not be interpreted as quantitative validation of classification performance, usability, wheelchair navigation, or safety. The EOG evaluation involved a single author-participant, while communication performance, false-activation rates, response time, inertial-estimation error, and motor performance were not systematically quantified.

Future work will focus on the quantitative characterization of the proposed system through repeated labeled experiments, multiple participants, objective classification and false-activation metrics, response-time measurements, communication-performance analysis, inertial-estimation characterization, and controlled experiments under progressively more realistic wheelchair operating conditions. This staged approach is intended to progressively evaluate the robustness, usability, and practical applicability of the proposed multimodal interface.

REFERENCES

[1] World Health Organization, *International Perspectives on Spinal Cord Injury*. Geneva, Switzerland: WHO Press, 2013.

[2] L. Fehr, W. E. Langbein, and S. B. Skaar, "Adequacy of power wheelchair control interfaces for persons with severe disabilities: A clinical survey," *Journal of Rehabilitation Research and Development*, vol. 37, no. 3, pp. 353–360, 2000.

[3] S. C. Kirshblum, S. P. Burns, F. Biering-Sorensen, W. Donovan, D. Graves, M. Jha, L. Jones, M. Johansen, A. Mulcahey, M. Schmidt-Read, and K. Waring, "International standards for neurological classification of spinal cord injury (Revised 2011)," *The Journal of Spinal Cord Medicine*, vol. 34, no. 6, pp. 535–546, 2011.

[4] J. R. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. New York, NY, USA: Oxford University Press, 2012.

[5] O. J. Woodman, *An Introduction to Inertial Navigation*. Cambridge, U.K.: University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-696, 2007.

[6] K. Kaneswaran and K. Arshak, "Capacitive interfaces for navigation of electric powered wheelchairs," in *13th International Conference on Biomedical Engineering*, C. T. Lim and J. C. H. Goh, Eds. Berlin, Heidelberg, Germany: Springer, 2009, pp. 1153–1157, doi: 10.1007/978-3-540-92841-6_283.

[7] S. Bian, M. Liu, B. Zhou, P. Lukowicz, and M. Magno, "Body-area capacitive or electric field sensing for human activity recognition and human-computer interaction: A comprehensive survey," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, Art. no. 4, pp. 1–49, Mar. 2024, doi: 10.1145/3643555.

[8] T. G. Zimmerman, J. R. Smith, J. A. Paradiso, D. Allport, and N. Gershenfeld, "Applying electric field sensing to human-computer interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO, USA, 1995, pp. 280–287, doi: 10.1145/223904.223940.

[9] P. Dorrington, C. Wilkinson, L. Tasker, and A. Walters, "User-centered design method for the design of assistive switch devices to improve user experience, accessibility, and independence," *Journal of Usability Studies*, vol. 11, no. 2, pp. 66–82, 2016.

[10] J. G. Webster, Ed., *Medical Instrumentation: Application and Design*, 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2010.

[11] C. Belkhiria and A. Peysakhovich, "EOG-based human–computer interface: 2000–2020 review," *Sensors*, vol. 22, no. 13, Art. no. 4914, 2022, doi: 10.3390/s22134914.

[12] E. M. Spinelli, N. H. Martinez, and M. A. Mayosky, "A single-supply biopotential amplifier," *Medical Engineering & Physics*, vol. 23, no. 3, pp. 235–238, 2001, doi: 10.1016/S1350-4533(01)00040-6.

[13] R. Barea, L. Boquete, M. Mazo, and E. López, "System for assisted mobility using eye movements based on electrooculography," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 4, pp. 209–218, Dec. 2002, doi: 10.1109/TNSRE.2002.806829.

[14] Q. Huang, Z. Gu, S. He, and Z. L. Yu, "An EOG-based human–machine interface for wheelchair control," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 2023–2032, Sept. 2018, doi: 10.1109/TBME.2017.2732479.

[15] A. Banks and R. Gupta, Eds., "MQTT Version 3.1.1," OASIS Standard, Oct. 2014. [Online]. Available: https://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html

[16] D. Thangavel, X. Ma, A. Valera, H.-X. Tan, and C. K.-Y. Tan, "Performance evaluation of MQTT and CoAP via a common middleware," in *Proc. IEEE 9th Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, Singapore, 2014, pp. 1–6, doi: 10.1109/ISSNIP.2014.6827678.

[17] E. Al-Masri, K. R. Kalyanam, J. Batts, J. Kim, S. Singh, T. Vo, and C. Yan, "Investigating messaging protocols for the Internet of Things (IoT)," *IEEE Access*, vol. 8, pp. 94880–94911, 2020, doi: 10.1109/ACCESS.2020.2993363.

A Non-Immersive Virtual Tour System Integrated with a Smart Cycle Ergometer for Chronic Respiratory Disease Rehabilitation

Yago Azevedo Fontele
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza(UNIFOR)
Fortaleza, Brasil
yagofontele@unifor.br

Guilherme Garcia Monteiro
Vice-Reitoria de Pesquisa (VRP)
Universidade de Fortaleza(UNIFOR)
Fortaleza-Brasil
guigarciamonteiro@unifor.br

João Victor Rodrigues Barreto
Vice-Reitoria de Pesquisa (VRP)
Universidade de Fortaleza(UNIFOR)
Fortaleza-Brasil
joao.barreto@unifor.br

Valentim Araújo de Fontenele Aragão
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
valentim@unifor.br

Ester Firmino Martins
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
esterfirmino@unifor.br

Felipe Cassiano Barbosa
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
felipecassianobarbosa@hotmail.com

Natália Bitar da Cunha Olegario
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
nataliabitar@unifor.br

Jorge Luis Pereira Monteiro
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
jorgeluizpereira@unifor.br

Jonatas Pereira da Silva
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
jonatassilva@unifor.br

Guilherme Pinheiro Ferreira da Silva
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
guilhermepinheiro87@unifor.br

José Fernando Rodrigues Ferreira Neto
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
guilhermepinheiro87@unifor.br

Joel Sotero da Cunha Neto
Vice-Reitoria de Pesquisa(VRP)
Universidade de Fortaleza (UNIFOR)
Fortaleza-Brasil
joelsotero@unifor.br

***Abstract*—This study presents the design and proof-of-concept evaluation of a non-immersive virtual tour system integrated with a smart cycle ergometer for pulmonary rehabilitation of patients with chronic respiratory diseases (CRDs). The platform combines a retrofitted ergometer, a wireless pulse oximeter, and an ESP32-based embedded system to acquire mechanical (speed, distance, power) and physiological (heart rate, oxygen saturation) data, synchronizing them with a first-person virtual route displayed on an external screen. A dual Bluetooth architecture (BLE and HM-10) supports simultaneous acquisition and transmission, complemented by a fault-tolerance mechanism that preserves mechanical data streaming during sensor disconnections. A technical evaluation with healthy volunteers assessed data continuity, communication stability, and recovery after simulated sensor failures, showing reliable multimodal acquisition and stable synchronization. As an early-stage proof of concept with healthy participants, no clinical claims are made, but the results establish a technical foundation for future clinical validation of interactive rehabilitation systems for CRD patients.**

***Index Terms*—IoT, cyclo ergometer, proof-of-concept, pulmonary rehabilitation**

I. INTRODUCTION

Chronic respiratory diseases (CRDs), including chronic obstructive pulmonary disease (COPD), interstitial lung disease, bronchiectasis, and chronic asthma, represent leading causes of death and disability worldwide [1]. COPD alone is recognized as the third leading cause of death globally, with approximately 24% of patients dying within five years of diagnosis [2], [3]. These conditions progressively impair lung function, exercise tolerance, and quality of life, creating a substantial and growing healthcare burden [2], [4], [3]. This scenario has also motivated international initiatives focused on digital health solutions for chronic respiratory conditions. One example is MULTIPULM, a Horizon Europe-funded project that aims to implement an integrated digital care ecosystem for multimorbid patients with chronic respiratory conditions in Brazil, Serbia, and Türkiye [5]. Although the system proposed in this study was developed independently from MULTIPULM, part of the research group involved in this article also participates in the consortium, reinforcing the alignment of this work with broader international efforts in

remote monitoring, patient engagement, and digital support for respiratory care.

Pulmonary rehabilitation (PR) has been established as a cornerstone in the management of CRDs. There is robust evidence that well-structured PR programs incorporating aerobic exercise, muscular strength training, and respiratory techniques significantly improve exercise capacity, reduce dyspnea, and enhance health-related quality of life [6], [2], [1], [3]. Specifically, exercise modalities such as cycling on stationary ergometers have been widely employed as effective tools for endurance training within PR programs [7], [1], [8]. However, despite these well-documented benefits, PR programs face persistent challenges related to patient accessibility, adherence, and motivation, which remain major barriers to successful outcomes [1], [3]. Fear of activity-triggered breathlessness, monotony of conventional exercise routines, and limited access to rehabilitation facilities contribute to low participation and high dropout rates [4], [3].

In recent years, virtual reality (VR) technology has emerged as a promising complementary approach to address these challenges. VR can be broadly classified into two categories: immersive VR, which utilizes head-mounted displays (HMDs) to fully envelop the user in a virtual world, and non-immersive VR, which presents virtual environments on external displays such as computer screens or television monitors without fully isolating the user from the real world [9], [10], [11]. Both modalities offer the capacity to create engaging environments that can shift patients' attention away from negative sensations such as fatigue and dyspnea during physical activity [6], [2], [4]. Studies have demonstrated that VR-based interventions can enhance patient motivation, increase exercise enjoyment, and improve treatment compliance [2], [9], [4]. The integration of VR with exercise training has shown significant improvements in exercise tolerance, lung function parameters such as FEV1/FVC, and psychological outcomes including reduced anxiety and depression in patients with COPD [2], [3].

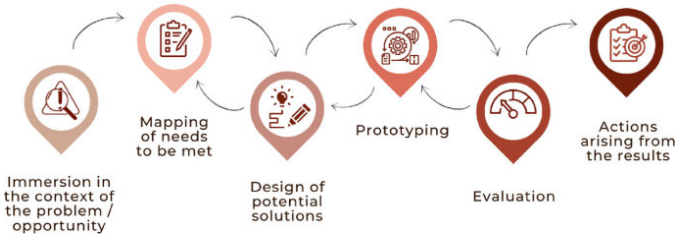
While immersive VR systems using HMDs have received considerable research attention, non-immersive approaches present distinct advantages for clinical rehabilitation populations. Immersive VR headsets generate noticeable heat during use, which, paired with physical exertion, can quickly become uncomfortable for participants [12]. Furthermore, immersive VR carries a significantly greater risk of simulator sickness, including nausea, headaches, and dizziness, compared to non-immersive alternatives [13], which may be particularly problematic for patients with CRDs who already experience respiratory distress. Non-immersive VR systems, which display virtual environments on television screens or projected displays positioned in front of the exercising patient, have demonstrated comparable benefits in terms of engagement and distraction while avoiding these adverse effects [10], [14], [15]. Historically, the combination of cycle ergometers with non-immersive virtual environments displayed on monitors has been explored since the mid-1990s, when early proof-of-concept systems used screens positioned in front of stationary bikes to present interactive virtual scenarios [10]. More recently, post-COVID-19 rehabilitation programs have successfully employed cycle ergometers synchronized with projected video simulations of outdoor cycling displayed on 75-inch television screens, with patients rating this modality among the highest-valued rehabilitation components [14], [7].

Despite these advances, the integration of smart cycle ergometers with virtual tour experiences specifically designed for non-immersive pulmonary rehabilitation of patients with CRDs remains underexplored. Most existing VR systems have focused on upper limb recovery, balance, or cognitive training, with cycling-based cardiorespiratory training receiving comparatively less attention [15]. This article presents the development and evaluation of a proof-of-concept system integrating a smart cycle ergometer with a non-immersive virtual tour application, displayed on a television screen and synchronized with the user's pedaling cadence. By combining real-time physiological monitoring, adaptive exercise parameters, and interactive virtual tours, the proposed architecture aims to mitigate well-documented drawbacks of head-mounted devices—such as cybersickness and thermal discomfort—while providing an accessible setup for pulmonary care. Therefore, this study focuses on the design, implementation, and engineering validation of the solution architecture, evaluating communication stability and integration among its core components.

II. METHODOLOGY

The project, from its conception to the production of this paper, was grounded in the Interdisciplinary Method for Technological Research Development with a Focus on Health (MIDTs), proposed by Vasconcelos et al. (2026). This methodology consists of six cyclical and iterative development steps, as described in Figure 1.

Fig. 1. MIDTs simplified flow



Source: Vasconcelos et al. (2026).

A. Problem Understanding

The first stage aimed to understand the rehabilitation context and identify opportunities for technological support in the care of patients with chronic respiratory diseases. This stage was based on discussions with physiotherapy professionals and a review of the literature on pulmonary rehabilitation, cycle ergometer-based exercise, patient adherence, and the use of virtual reality resources in rehabilitation.

The analysis indicated that stationary cycling is a widely used strategy in pulmonary rehabilitation because it supports controlled aerobic training and can be adapted to different functional capacities. However, conventional cycling sessions

may become repetitive and poorly engaging, which can negatively affect adherence and continuity of rehabilitation. Therefore, the use of a non-immersive virtual environment synchronized with pedaling cadence was defined as a strategy to increase engagement without exposing users to the discomfort and cybersickness risks associated with head-mounted displays.

In addition, the problem understanding stage highlighted the need for real-time physiological monitoring during exercise, especially for patients with respiratory limitations. Thus, the proposed system was designed to combine exercise interaction, visual feedback, and the acquisition of physiological and mechanical variables relevant to pulmonary rehabilitation.

B. Requirements

After defining the problem, the project requirements were established from three complementary perspectives: clinical monitoring, hardware integration, and interactive feedback. From the clinical perspective, the system should monitor variables commonly used during supervised exercise sessions, including peripheral oxygen saturation, heart rate, pedaling speed, exercise time, and estimated power.

From the hardware perspective, the system should acquire data from both the cycle ergometer and the pulse oximeter, process these signals through an embedded system, and transmit them wirelessly to the interactive interface. The prototype should also include local data storage to preserve session records and support future analysis.

From the software perspective, the interface should display the virtual tour and real-time exercise data in a clear and minimally disruptive way. The visual environment should progress according to the user's pedaling cadence, creating a direct relationship between physical effort and movement through the virtual route. The interface should also display physiological and mechanical indicators, allowing the user or supervising professional to monitor the activity without interrupting the exercise.

Based on these requirements, the proof of concept focused on four core functions: acquisition of physiological data, estimation of mechanical variables, wireless communication between hardware and interface, and synchronization between pedaling cadence and virtual tour progression.

C. System Design

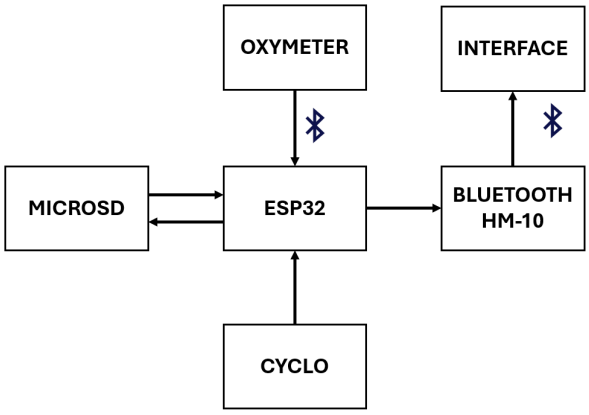
The system architecture was designed to integrate three main components: an adapted cycle ergometer, a pulse oximeter for physiological monitoring, and a web-based non-immersive virtual tour interface. The ESP32 microcontroller was selected as the embedded processing unit due to its low cost, compact size, processing capacity, and native support for Bluetooth Low Energy communication.

A central design requirement was the need to handle two simultaneous Bluetooth-based communication flows. The first flow involved receiving physiological data from the pulse oximeter through BLE, including heart rate and peripheral oxygen saturation. The second flow involved transmitting the processed mechanical and physiological data from the embedded system to the web interface. To support this dual communication architecture, an additional HM-10 Bluetooth module was incorporated into the system. In this configuration, the ESP32 was responsible for communicating with the pulse oximeter via BLE, while the HM-10 module provided a dedicated Bluetooth communication channel between the embedded system and the interface.

The mechanical data acquisition subsystem was designed to estimate pedaling speed and power. A magnetic sensor was used to detect pedaling pulses and estimate angular velocity, serving as the basis for calculating speed and generated power. The sensor was also calibrated using a tachometer to improve the accuracy of the estimates. Torque was estimated through testing, and power was derived indirectly from the relationship between torque and speed. A low-cost methodology was employed to calculate these parameters [16].

The data flow was organized so that mechanical variables captured from the cycle ergometer and physiological variables received from the oximeter could be processed by the embedded system and transmitted to the web interface through the HM-10 communication channel. This architecture separated the acquisition and transmission tasks, reducing communication conflicts and supporting real-time visualization during exercise. The architecture also included a MicroSD module for local backup storage, allowing session data to be preserved even in the event of communication instability. The conceptual architecture of the proposed system is presented in Figure 2.

Fig. 2. System architecture and data flow between the cycle ergometer, oximeter, embedded system, HM-10 module, and web interface.



D. Prototype Development

The prototype was developed by integrating the embedded hardware, firmware, and interactive web interface. In the hardware module, the pulse oximeter was connected to the ESP32 through Bluetooth Low Energy to acquire oxygen saturation and heart rate data. The cycle ergometer was adapted with

a magnetic sensor to acquire pedaling information, allowing the estimation of speed and power. The power estimation method was based on a previous study developed by the research group, in which pedaling power was calculated from the relationship between angular velocity and torque in an adapted cycle ergometer [16].

The firmware was implemented to process the mechanical data from the cycle ergometer and combine them with the physiological data received from the oximeter. These data were organized into structured packets and transmitted to the web interface through the HM-10 Bluetooth module. A printed circuit board was also designed to organize the components, improve connection stability, and support more reliable testing.

The interface was developed using HTML5, CSS3, and JavaScript. Communication between the hardware and the browser-based interface was implemented using the Web Bluetooth API, enabling the direct reception of Bluetooth data packets by the web application. This approach was selected to reduce dependence on local network infrastructure and to minimize latency fluctuations that could affect the synchronization between pedaling and the visual experience.

The interactive interface was implemented to display a first-person non-immersive virtual tour and real-time indicators of exercise performance and physiological status. The estimated speed was used to control the progression of the virtual route, while power, oxygen saturation, heart rate, and elapsed time were organized as visual indicators on the exercise screen. Initially, two virtual routes were developed: one representing Fortaleza, Brazil, where the system was developed, and another representing Novi Sad, Serbia, a partner country in the MULTIPULM project. These routes were selected to demonstrate the flexibility of the system in incorporating different geographic and contextual scenarios into the rehabilitation experience. The developed interface and its operation are presented in the Results section.

E. Evaluation

The technical evaluation was conducted as a proof-of-concept test with six healthy participants, with a mean age of 21 ± 2 years. The evaluation aimed to verify the feasibility of the integrated system in terms of data acquisition, real-time visualization, communication stability, and fault tolerance, without assessing clinical or therapeutic outcomes.

The web-based interface was developed as a localhost application, operating exclusively within a controlled local environment without external network dependencies. This architecture allowed for direct and predictable communication between the hardware and the interface throughout all testing sessions.

Each participant performed a standardized five-minute session on the adapted cycle ergometer while using a fingertip pulse oximeter, following a predetermined and identical virtual route across all participants. During the session, all participants were instructed to exercise at their maximum sustainable effort to ensure consistent exercise intensity throughout the testing protocol. Participants were informed of this objective before each session, and the interface continuously displayed their current speed to enable real-time monitoring of their performance. Since the objective was to evaluate the technical behavior of the system under standardized conditions, the route was not considered as an experimental variable, but rather as a controlled element to ensure reproducibility across all testing sessions.

During the sessions, the system collected mechanical variables from the cycle ergometer, including elapsed time, distance, speed, and estimated power, as well as physiological variables from the pulse oximeter, including peripheral oxygen saturation and heart rate. These data were transmitted to the web interface and displayed in real time. The oximeter readings were also manually checked throughout the tests and compared with the values shown on the platform to verify consistency between the source device and the interface.

Beyond the participant trials, three independent 15-minute tests were conducted strictly to monitor the hardware-interface connection stability and quantify data packet loss. Furthermore, the recorded sessions were analyzed for signal continuity, the presence of missing values, and possible communication interruptions between the cycle ergometer, the oximeter, and the interface.

Finally, an additional fault-tolerance test was performed by temporarily removing the pulse oximeter from the participant's finger and then repositioning it. This procedure simulated a transient disconnection of the physiological signal and was used to verify whether the system could detect the interruption, preserve the mechanical data stream, and automatically resume heart rate and oxygen saturation monitoring after reconnection without requiring a new session.

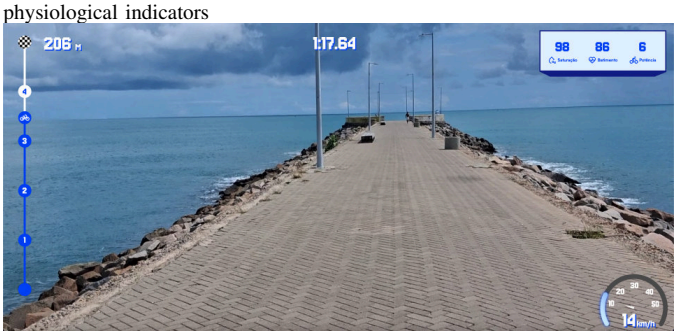
III. RESULTS

A. Presentation of the Integration

The proposed prototype is an integrated non-immersive virtual tour system connected to an adapted cycle ergometer and a pulse oximeter. It enables users to pedal while following a first-person virtual route displayed on an external screen, with the progression of the route synchronized with the pedaling speed estimated from the cycle ergometer.

Figure 3 presents the developed interface during the virtual tour. The central area of the screen displays the first-person route, while real-time indicators are arranged around the visual field to support continuous monitoring without interrupting the activity. The interface displays mechanical parameters, including elapsed time, estimated speed, distance, and pedaling power, together with physiological indicators, namely peripheral oxygen saturation (SpO_2) and heart rate. These elements satisfy the functional requirements previously established for the prototype, particularly the simultaneous visualization of exercise performance and physiological status.

Fig. 3. Non-immersive virtual tour interface with real-time mechanical and physiological indicators



Two initial virtual routes were implemented: one representing Brazil, where the system was developed, and another representing Serbia, a partner country in the MULTIPULM project. The interface also includes a random route selection option. These alternatives demonstrate the flexibility of the system to incorporate different geographic contexts into the rehabilitation experience. To ensure reliability during these sessions, the interface includes an alert system (Figure 4) that notifies users when sensor disconnections occur, allowing participants to remain aware of system status without interrupting their exercise.

Fig. 4. Non-immersive virtual tour interface with real-time mechanical and physiological indicators with component disconnection alerts



When a sensor disconnection occurs, the interface displays three warning indicators in the upper-right corner. Each orange icon signals the unavailability of a specific parameter: Saturação (SpO_2), Batimento (heart rate), and Potência (power output). When this happens, the system continues to record mechanical data normally. Physiological measurements are marked with a sentinel value (-1) until the sensor automatically reconnects.

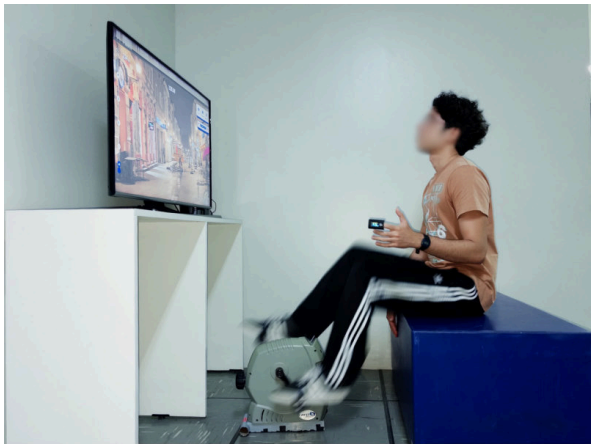
Figure 5 shows the prototype during the proof-of-concept evaluation. Panel (a) presents a focused side view of the participant pedaling while wearing the pulse oximeter, whereas panel (b) shows a broader view of the system and the interaction with non-immersive VR. Together, these views illustrate the intended use scenario, in which the participant remains physically engaged in the cycling activity while receiving real-time visual feedback from the simulated route.

The evaluation aimed to verify whether the prototype could operate as an integrated system under realistic conditions. The following analyses address real-time data acquisition, synchronization between mechanical and physiological measurements, communication stability, consistency of the acquired signals, and system behavior during a simulated temporary disconnection of the pulse oximeter.

Fig. 5. Prototype operation during the proof-of-concept test: (a) side view of the participant using the adapted cycle ergometer; (b) broader view showing the participant's interaction with the virtual tour interface.



(a)



(b)

B. Real-Time Data Acquisition and Synchronization

During the testing sessions, the system continuously recorded telemetry data from both the cycle ergometer and the pulse oximeter, enabling a practical evaluation of the integration between the hardware modules and the web interface. Table I presents a representative subset from one session on the Brazil route, which illustrates the structure of the transmitted data packets, while the final row summarizes the average values computed over the complete session.

TABLE I
REAL-TIME DATA COLLECTION SAMPLE AND AVERAGES

Timestamp (ISO 8601)	**Uptime** (mm:ss.cs)	**Distance** (m)	**Speed** (km/h)	**SpO_2** (%)	**Heart Rate** (bpm)	**Power** (W)
2026-07-09T21:21:18.387Z	00:00.12	0	4.31	99	79	8.38
2026-07-09T21:21:18.917Z	00:00.72	0	5.78	99	79	11.25
2026-07-09T21:21:19.428Z	00:01.20	1	6.50	98	80	12.65
2026-07-09T21:21:19.927Z	00:01.68	2	6.86	98	80	13.34
—	**Session Avg**	**781 (final)**	**9.90**	**98.83**	**88.79**	**19.25**

[a]Excerpt of sequential, non-duplicated transmissions taken from the initial moments of the collection; averages computed over the complete 5-minute session (*valentim.csv*, n = 2501).

In addition to the sample presented above, the overall characteristics of the test sessions were quantified to evaluate the system under continuous operation. Table II summarizes the global and communication-related metrics observed across all five-minute sessions performed during the proof-of-concept evaluation, providing an overview of the robustness of the data acquisition pipeline.

TABLE II
QUANTITATIVE SUMMARY OF THE TEST SESSIONS

Metric	**Result**
Selected Routes	*brazil* and *servia*
Total Duration (per session)	5 minutes (5:00.00)
Local Packaging Cadence	∼120 ms (interface sampling tick)
BLE Telemetry Update Interval	∼508 ms (min. 468, max. 561 ms)
Total Data Records	12,503 records across 6 sessions (∼2,500 each)
Missing Physiological Values	288 records (∼2.3% of samples)
Successful Data Capture Rate	∼97.7% (mechanical stream: 100%)

Analysis of Simultaneous Recording

Across the six complete sessions, mechanical (speed, distance, and power) and physiological (oxygen saturation and heart rate) measurements were consistently recorded within the same integrated data packet, using a local sampling cadence of approximately 120 ms. Because both data streams are timestamped and merged before transmission to the interface, no temporal drift or synchronization offset was observed between them. These results demonstrate that the embedded architecture can process data from the cycle ergometer and the pulse oximeter concurrently, without one acquisition stream blocking or delaying the other.

Analysis of Transmission and Telemetry Gaps

Unlike the mechanical measurements, which were successfully captured in 100% of the records, the physiological measurements (SpO_2 and heart rate) exhibited brief, isolated gaps in two of the six sessions, accounting for approximately 2.3% of the total samples. These interruptions appeared as short sequences of zero-valued samples, consistent with transient signal loss from the fingertip pulse oximeter rather than a system-level failure, since the mechanical data stream remained uninterrupted throughout the same intervals. This behavior anticipates the fault-tolerance mechanism discussed later in the Reliability of the Data subsection, in which the application represents the temporary absence of physiological data using a sentinel value while preserving uninterrupted acquisition of the mechanical measurements.

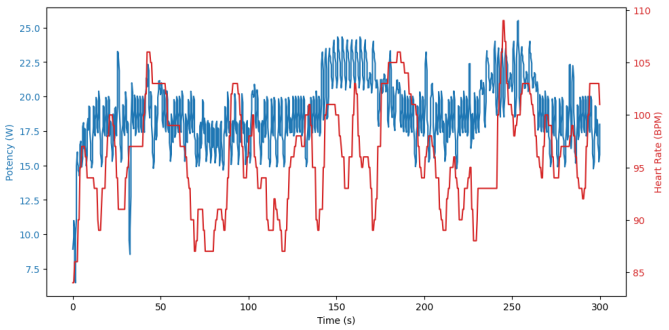
Regarding communication timing, the local packaging interval remained stable at approximately 120 ms, whereas the BLE telemetry link updated every 508 ms on average (ranging from 468 to 561 ms), consistent with the communication stability analysis presented later in this section. No unexpected disconnections, packet losses, or processing delays capable of compromising the real-time visual feedback loop were observed beyond the isolated pulse oximeter dropouts described above. Overall, the system achieved a physiological data capture rate of approximately 97.7%, while maintaining uninterrupted acquisition of the mechanical data stream throughout all test sessions.

C. Consistency of Physiological and Mechanical Data

In addition to confirming the simultaneous acquisition of mechanical and physiological measurements, it is also important to evaluate whether the recorded signals exhibited behavior consistent with the exercise being performed. Figure 6 illustrates the relationship between the cycle ergometer's estimated power output and the participant's heart rate during a representative session. As expected, heart rate increased during periods of greater mechanical effort and decreased as the exercise intensity was reduced.

As the evaluation involved healthy volunteers performing a short, non-standardized exercise protocol, these observations should be interpreted solely as evidence that the proposed system can continuously acquire, synchronize, and display mechanical and physiological signals during exercise. They are not intended to demonstrate clinical effectiveness or improvements in cardiovascular performance.

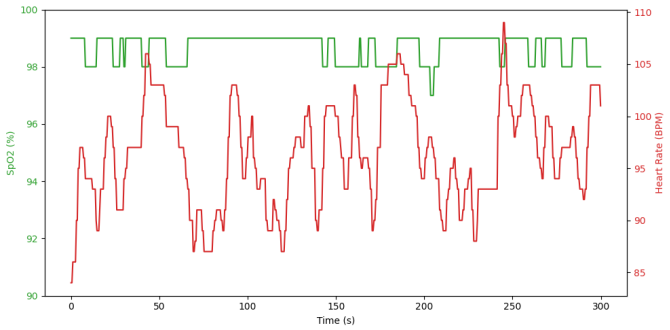
Fig. 6. Mechanical Effort vs. Physiological Response



To further demonstrate the continuity of the physiological data stream, Figure 7 presents the temporal evolution of SpO_2 and heart rate for a representative participant throughout the entire session. Oxygen saturation remained within a stable range during exercise, whereas heart rate varied continuously in response to pedaling effort, without abrupt jumps, prolonged flat segments, or discontinuities that would suggest stalled or

repeated measurements. These results indicate that the interface successfully captured and displayed the natural temporal variation of both physiological signals, rather than freezing or interpolating values during data acquisition.

Fig. 7. Temporal Evolution of SpO_2 and Heart Rate (BPM)



To ensure the reliability of the data displayed on the web interface, the researchers periodically compared these values with the pulse oximeter throughout the testing sessions. These manual comparisons consistently confirmed agreement between the displayed and source measurements in real time, with no discrepancies beyond the expected latency of the BLE communication. This evaluation demonstrates the technical fidelity of the data transmission pipeline; however, no claims are made regarding the clinical significance of the observed physiological responses, since the participants were healthy volunteers performing a short proof-of-concept evaluation rather than a clinical rehabilitation protocol.

D. BLE Communication Latency

During the independent 15-minute extended stability tests, the system maintained a highly stable telemetry stream, successfully receiving 99.73% of the expected data packets (representing a minimal average packet loss of only 0.27%). Notably, two of the three extended sessions exhibited absolute continuity with 0% packet loss, with all minor disconnections isolated to a single session.

To evaluate the stability of the wireless communication, the inter-packet interval of the BLE telemetry was analyzed for the six participants who completed a full session. Table III summarizes the mean, standard deviation, minimum, and maximum latency observed for each participant, together with the aggregated results across all sessions.

The packets were received within a narrow interval around the nominal BLE update period of 500 ms for most of the participants, with maximum latencies not exceeding 561 ms and standard deviations below 15 ms. These results indicate a highly stable communication cadence, with no other evidences of packet loss or connection instability during normal operation apart from a single latency spike of 1030 ms, observed during DA's session. As discussed in the following subsection, this event coincided with the intentional temporary removal of the pulse oximeter and therefore does not represent a communication failure under normal operating conditions.

TABLE III
BLE Inter-Packet Latency by Participant

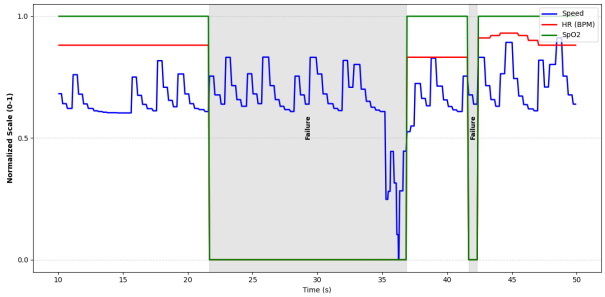
Participant	Packets	Mean (ms)	Std. Dev. (ms)	Min (ms)	Max (ms)
AN	591	508.6	11.1	459	551
ES	591	509.1	13.7	470	555
JV	590	509.1	13.8	471	541
DA	589	509.3	26.2	470	1030
VA	591	508.0	13.4	470	550
YA	591	507.9	14.1	468	561
Aggregated	**3537**	**508.7**	**16.1**	**459**	**1030**

The participants's names were ommited for privacy purposes.

E. Fault Tolerance During Simulated Oximeter Disconnection

To evaluate the fault tolerance of the proposed architecture, the pulse oximeter was intentionally removed from the participant's fingertip during one session and subsequently repositioned, simulating a temporary loss of the physiological signal. As illustrated in Figure 8, the application maintained normal operation throughout the event. As soon as valid oximeter readings became unavailable, the system automatically assigned a reserved sentinel value (-1) to the physiological variables while continuing to acquire and display the mechanical measurements (speed, distance, and power) without interruption or reduction in the sampling rate. Once the pulse oximeter was repositioned, SpO_2 and heart rate measurements resumed automatically, demonstrating that the system could recover autonomously without requiring user intervention or a restart of the exercise session.

Fig. 8. Behavior of Speed, HR, and SpO_2 During a Simulated Connection Failure



These results demonstrate that the proposed architecture remains operational during temporary physiological signal loss, an important requirement for pulmonary rehabilitation systems. As patients with chronic respiratory diseases frequently experience dyspnea and early fatigue [2], [4], unexpected monitoring interface interruptions could negatively affect exercise continuity and patient adherence. The adopted fault-tolerance strategy isolates communication failures by assigning a sentinel value (-1) to unavailable physiological measurements while preserving acquisition and visualization of the mechanical data stream. Consequently, the rehabilitation session proceeds normally until the physiological signal is restored, avoiding unnecessary exercise interruption due to temporary sensor disconnections.

It should be emphasized that the present evaluation focused exclusively on validating the technical reliability and integration of the proposed system using healthy participants. Clinical aspects, including adaptive adjustment of exercise intensity based on physiological feedback, patient-reported usability, comfort, and rehabilitation outcomes, were beyond the scope of this proof-of-concept study and should be investigated in future clinical trials involving individuals with chronic respiratory diseases.

IV. CONCLUSION

This study presents the design and proof-of-concept evaluation of an integrated system for pulmonary rehabilitation that combines an adapted cycle ergometer, a pulse oximeter, and a non-immersive virtual tour interface. By integrating mechanical and physiological monitoring into a single platform, the proposed architecture enables synchronized data acquisition and real-time visualization, providing an interactive environment to support rehabilitation sessions.

The technical evaluation demonstrated that the system could reliably acquire, synchronize, and transmit multimodal data throughout exercise while maintaining stable BLE communication and continuous visual feedback. Furthermore, the proposed fault-tolerance mechanism preserved uninterrupted acquisition and visualization of the mechanical data stream during temporary pulse oximeter disconnections and automatically resumed physiological monitoring once the sensor was repositioned, without requiring user intervention. These results demonstrate the technical feasibility and robustness of the proposed architecture, establishing a solid foundation for future pulmonary rehabilitation systems that combine real-time physiological monitoring with interactive virtual environments.

This study, however, represents a technical proof of concept. The evaluation was conducted exclusively with healthy volunteers and was not intended to assess clinical efficacy, patient adherence, or rehabilitation outcomes. Therefore, no conclusions can be drawn regarding the therapeutic benefits of the proposed system for individuals with chronic respiratory diseases.

Future work will focus on integrating cloud-based data storage to enable longitudinal monitoring, implementing remote supervision tools for physiotherapists, calibrating and comparing the system against standard cycle ergometers and oximeters, and conducting clinical studies involving patients with chronic respiratory diseases. These studies will investigate usability, patient engagement, safety, and the efficacy of the proposed system in real-world pulmonary rehabilitation programs.

ACKNOWLEDGMENT

The team would like to thank the University of Fortaleza (UNIFOR) and the Vortex Laboratory for the opportunity, and especially professors Natália Bitar da Cunha Olegario and Joel Sotero da Cunha Neto, as well as Jonatas Silva Pereira, for supervising the development of this project.

REFERENCES

[1] S. A. Robinson and M. L. Moy, "Promoting exercise training remotely," *Life*, vol. 12, no. 2, p. 262, 2022.

[2] Q. Wei *et al.*, "Synergistic effects of virtual reality and traditional treatment methods in the management of chronic obstructive pulmonary disease: a systematic review and meta-analysis of randomized controlled trials," *Peerj*, vol. 13, p. e20047, 2025.

[3] I. Patsaki, V. Avgeri, T. Rigoulia, T. Zekis, G. A. Koumantakis, and E. Grammatopoulou, "Benefits from incorporating virtual reality in pulmonary rehabilitation of copd patients. a systematic review and meta-analysis," 2023.

[4] C. Condon, W. T. Lam, C. Mosley, and S. Gough, "A systematic review and meta-analysis of the effectiveness of virtual reality as an exercise intervention for individuals with a respiratory condition," *Advances in Simulation*, vol. 5, no. 1, 2020.

[5] MULTIPULM, "Integrated care digital-based interventions for multimorbid patients suffering from chronic respiratory conditions in brazil, serbia and türkiye," 2026, accessed: May 17, 2026. [Online]. Available: https://www.multipulm.eu/

[6] S. Rutkowski, K. Bogacz, A. Rutkowska, J. Szczegielniak, and R. Casaburi, "Inpatient post-covid-19 rehabilitation program featuring virtual reality—preliminary results of randomized controlled trial," *Frontiers in Public Health*, vol. 11, 2023.

[7] M. Migała *et al.*, "Expectations of patients recovering from sars-cov-2 towards new forms of pulmonary rehabilitation," *International Journal of Environmental Research and Public Health*, vol. 20, no. 1, p. 104, 2022.

[8] S. Rutkowski, K. Bogacz, O. Czech, A. Rutkowska, and J. Szczegielniak, "Effectiveness of an inpatient virtual reality-based pulmonary rehabilitation program among covid-19 patients on symptoms of anxiety, depression and quality of life: Preliminary results from a randomized controlled trial," *International Journal of Environmental Research and Public Health*, vol. 19, no. 24, p. 16980, 2022.

[9] K. Kawaguchi *et al.*, "Effects of different visual flow velocities on psychophysiological responses during virtual reality cycling," *Cureus*, 2024.

[10] E. R. Høeg, "Just like riding a bike: Exploring the motivating capabilities, contextual barriers and facilitators of virtual reality and cycle ergometers for the rehabilitation of older adults in clinical practice."

[11] V. Colombo *et al.*, "Rehabilitation of post-covid patients: A virtual reality home-based intervention including cardio-respiratory fitness training," 2022, pp. 3–17.

[12] S. L. Finnegan *et al.*, "Breathlessness in a virtual world: An experimental paradigm testing how discrepancy between vr visual gradients and pedal resistance during stationary cycling affects breathlessness perception," 2022.

[13] L. R. Biller *et al.*, "Incongruent virtual reality attenuates breathlessness and leg fatigue during stationary cycling," 2024.

[14] A. d. S. Costa *et al.*, "Revive-hf: Rehabilitation with immersive virtual reality and exercise in hospitalized patients with heart failure - a randomized controlled trial protocol," 2025.

[15] V. Colombo *et al.*, "Rehabilitation of post-covid patients: A virtual reality home-based intervention including cardio-respiratory fitness training," 2022, pp. 3–17.

[16] Y. A. Fontele, I. S. L. Waki, J. S. Pereira, J. S. Cunha Neto, and N. B. C. Olegário, "Método para cálculo da potência de um cicloergômetro," in *Proceedings of the XXXI Encontro de Iniciação à Pesquisa*. Fortaleza, Brazil: Universidade de Fortaleza, 2025. [Online]. Available: https://uol.unifor.br/oul/conteudosite/?uuid=3C2D9FFEB6C899B51741E6FAECD8492233092801

[17] J. E. de Vasconcelos Filho, J. S. da Cunha Neto, and J. F. Ferreira, *MIDTs: Interdisciplinary Method for Technological Research Development with a Focus on Health*. Cham: Springer Nature Switzerland, 2026, pp. 131–143. [Online]. Available: https://doi.org/10.1007/978-3-031-89820-4_10

A Virtual Simulation System to Support Wheelchair Prescription Concerning Patient Environment

Victor de Mendonça Ramos
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, São Paulo
ORCID: 0009-0002-8758-4549

Regina Célia Coelho
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, São Paulo
ORCID: 0000-0002-4428-9745

Caio Cigagna de Godoy
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0000-0002-6548-8153

Tiffany Mariano
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0009-0001-5166-4478

Fabrício Morais Muller
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0009-0004-4937-3155

Eliana Chaves Ferretti
Instituto de Saúde e Sociedade
Universidade Federal de São Paulo (UNIFESP)
Santos, Brasil
ORCID: 0000-0002-7605-828X

Vagner Rogério dos Santos
Escola Paulista de Medicina
Universidade Federal de São Paulo (UNIFESP)
São Paulo, Brasil
ORCID: 0000-0002-6034-6658

Carlos Marcelo Gurjão de Godoy
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0000-0001-8846-3242

***Abstract*— Wheelchair prescription depends on functional, postural, and environmental analysis. A recurrent difficulty in this process is anticipating whether a wheelchair will be compatible with the patient's home before the device is delivered or adapted. This paper presents a visual-interactive prototype developed in Unreal Engine 5 to support environmental analysis during wheelchair prescription planning. The system allows physiotherapists to create simplified residential layouts, insert and reposition objects, switch between patient-centred and caregiver/conductor views, and navigate the environment using a simulated wheelchair. The prototype integrates wheelchair control, environment editing, object manipulation, and an inventory of furniture and structural elements. Initial requirements were guided by physiotherapy collaborators in the project team, and preliminary functional verification focused on navigation, object positioning, view switching, collision behavior, and interaction consistency. The current version does not replace clinical assessment and has not yet been validated in real homes or with patients. Rather, it demonstrates preliminary technical feasibility as a visual support tool for testing spatial configurations and observing possible environmental constraints during prescription planning.**

***Keywords*—** Wheelchair Prescription, Virtual Reality, Unreal Engine 5, Rehabilitation Engineering, Assistive Technology

I. INTRODUCTION

Wheelchair prescription is a multidisciplinary process that considers the user's physical condition, posture, mobility needs, and daily environment. The choice of an inadequate wheelchair can reduce comfort, independence, and accessibility, and may also contribute to secondary problems related to posture and mobility use [1-3]. For this reason, the prescription process cannot be limited to the device itself. It also requires attention to how the wheelchair will be used in the places where the patient performs daily activities.

A practical challenge in this process is the relation between the wheelchair and the patient's home. Door widths, corridors, furniture placement, turning areas, and circulation spaces can determine whether a wheelchair is usable in daily routines. These conditions are usually assessed through measurements, photographs, visits, or patient reports. Although these procedures are useful, they are mostly static. They may indicate that a corridor has a certain width, but they do not necessarily show how a wheelchair behaves while turning, entering a room, or avoiding furniture [2].

Virtual reality and real-time simulation have been used in rehabilitation and assistive technology because they allow interaction with controlled and repeatable scenarios [4], [5]. Wheelchair simulators have also been proposed for navigation training, driving assessment, and evaluation of user

This work was developed using resources obtained from the Financiadora de Estudos e Projetos (FINEP) – Ref. Finep nº 2174/22.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

performance [6-8]. These studies are important because they show that virtual environments can reproduce mobility tasks without exposing users to the same risks found in real environments.

However, many wheelchair simulation systems focus on the wheelchair user as the main participant. Their usual goal is to train driving skills, measure performance, or study interaction in immersive environments. The problem addressed in this work addresses a different problem. During prescription planning, the physiotherapist needs to reason about the relation between the device and the patient environment before final decisions are made. In this context, the system must support layout construction, object manipulation, and quick testing of different spatial configurations.

II. RELATED WORKS

Virtual environments have been widely explored for wheelchair training and mobility assessment. Vailland et al. [6] evaluated a VR-based power wheelchair simulator with regular users, while Bigras et al. [7] studied augmented feedback during powered wheelchair simulator training. These works show that simulation can be useful for safe practice and for analyzing aspects of wheelchair performance.

Other studies focused on realism, usability, and user experience. Riera et al. [8] developed a haptic virtual walker for wheelchair users in immersive environments. Chaar and Archambault [9] evaluated a low-cost virtual reality simulator for manual wheelchair training and emphasized usability issues that must be considered in this type of system. Salgado et al. [10] presented a multimodal dataset from an immersive wheelchair simulator, including physiological and questionnaire-based data. These works indicate that interaction quality and user comfort are relevant aspects when virtual simulators are applied to rehabilitation.

Some studies also move closer to accessibility evaluation. Monteiro et al. [11] proposed a virtual wheelchair accessibility evaluation tool capable of capturing trajectory and time metrics. Systematic reviews have discussed the use of virtual reality to complement wheelchair skills training and broader rehabilitation activities [12], [13]. These studies point to the value of simulation, but also show that validation protocols and clinical use cases are still developing.

The present work uses this literature as a technical reference, but targets a different use case. The objective is not to train the wheelchair user or to measure driving performance as the main outcome. The focus is to provide the physiotherapist with an editable environment that can represent the patient's home and support prescription planning. This difference guided the selection of the implemented features, especially the editing mode, inventory system, user-configurable obstacles, and the transition between patient-centered and caregiver/conductor views.

III. MATERIALS AND METHODS

A. Functional Requirements and Scope

The system was developed in Unreal Engine 5.5.1. The engine was selected because it supports real-time rendering, collision handling, view switching, and visual scripting through Blueprints. GitHub was used for version control and organization of implementation tasks. Blender was also used during asset preparation and 3D modeling activities.

The initial functional requirements were defined through iterative discussions within the development team and with the physiotherapy collaborators involved in the project. These discussions were used to identify the main environmental and interaction features considered important for the initial prototype.The requirements included the presence of walls, doors, windows, tables, chairs, cabinets, and a user-configurable object that could represent obstacles or other elements found in a patient's home.

Another requirement was the possibility of switching between the view of the wheelchair user and the view of the person conducting or assisting the patient. In the implementation, this mechanism was created using camera resources available in Unreal Engine 5, but in the clinical description of the system it is treated as user-type view switching. This distinction is important because the purpose is not to describe camera technology itself, but to provide different viewpoints for navigation, inspection, and discussion of spatial configurations.

The current version was defined as a visual-interactive prototype. It does not yet incorporate formal accessibility standards as automatic constraints, and the virtual scenarios were not built from real residential measurements. Therefore, the prototype should be understood as a platform for testing the interaction workflow and the feasibility of environmental simulation, not as a validated accessibility measurement tool.

B. Development Process and System Architecture

The development process followed an iterative approach. First, the project focused on studying wheelchair prescription, rehabilitation environments, and accessibility requirements. In parallel, Unreal Engine resources were studied, including interaction systems, view control, collision behavior, and possibilities for future VR integration.

The prototype is organized into four main modules: the virtual environment module, the wheelchair simulation module, the environment editing module, and the inventory module. The virtual environment represents the residential layout. The wheelchair module controls navigation and basic collision with objects. The editing module allows objects and structural elements to be modified during execution. The inventory module stores and instantiates furniture and other elements used to build the simulated home.

The system was planned to work in two basic modes: navigation and editing. In navigation mode, the user controls the wheelchair and observes its interaction with the environment. In editing mode, the user changes the scene by inserting, moving, rotating, or resizing objects. The same

environment can be adjusted and tested repeatedly. Fig. 1 summarizes the main modules and the interaction flow.

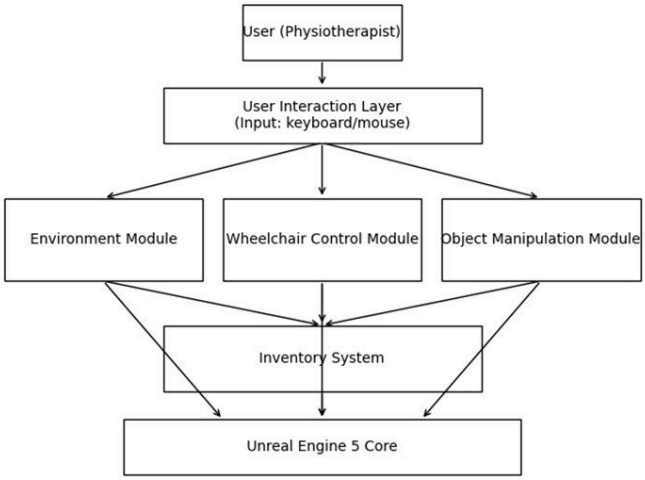


Fig. 1. Overall architecture of the system and interaction between implemented modules .

C. Wheelchair Control and Interaction Model

The wheelchair model is currently controlled by keyboard commands. The W key moves the wheelchair forward, the S key moves it backward, the A key rotates it to the left, and the D key rotates it to the right. This control strategy was adopted to provide a simple and reproducible way to test navigation through the virtual layouts during the prototype stage.

Basic collision detection was implemented for interactions with walls and objects inserted in the scene, including chairs, tables, cabinets, and other elements available through the inventory. This allows the user to observe whether the wheelchair can pass through simplified circulation areas or whether an object blocks the route. However, the current movement model is simplified and does not reproduce the full behavior of a real wheelchair.

At this stage, the system does not include a biomechanical model of the user, a detailed kinematic model of the wheelchair, propulsion effort estimation, automatic turning radius calculation, trajectory analysis, task completion time, or minimum passage width measurement. These functions are planned as future improvements and are necessary before the prototype can be used as an objective accessibility evaluation tool.

D. Virtual Environment and Editing Tools

A virtual indoor environment was created to represent rooms, circulation areas, walls, and furniture placement zones. The environment was intentionally simplified in this version because the main objective was to test the interaction workflow and not to produce a final architectural rendering. The prototype also includes doors, windows, furniture elements, and configurable blocks that can be used to approximate obstacles or structural components of a residence.

After an object is inserted into the scene, the user can reposition it using manipulation controls. As shown in Fig. 2, the object can be translated in different directions and rotated without recreating it. This is useful for testing small layout changes, such as repositioning furniture or changing object orientation to improve wheelchair circulation in the simulated scene.

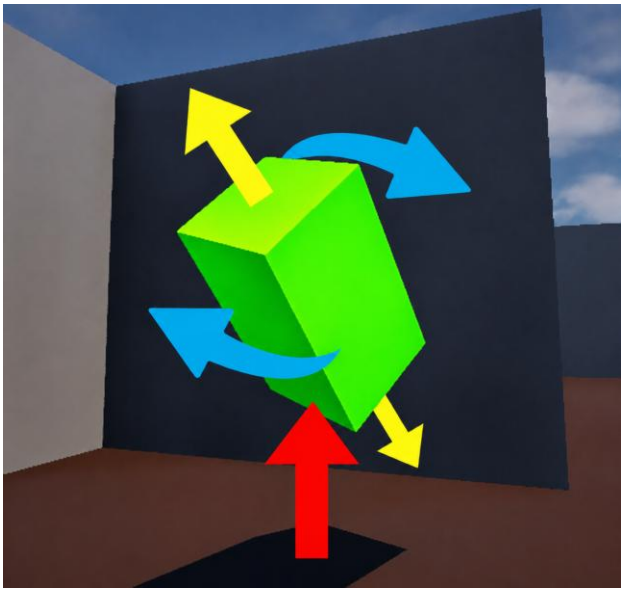


Fig. 2. Object translation and rotation controls used during environment editing.

A modular object generation mechanism was also implemented, allowing the user to create configurable blocks and structural elements at runtime. The purpose is to enable simplified construction of patient-specific layouts without requiring the entire scene to be modeled in advance. Fig. 3 illustrates the object generation and positioning tools available in the system.



Fig. 3. Object generation and positioning tools available within the system.

E. Inventory System and User-Type Views

The inventory system was developed to simplify the insertion of objects into the virtual environment. It contains furniture elements such as tables, chairs, and cabinets, and it can be expanded with new elements in future versions. The user can select an object, instantiate it in the scene, and then adjust its position through the editing tools.

The inventory and editing systems were designed to work together so that the user can insert an object, test wheelchair circulation, change the object position, and test the route again without leaving the simulation. In addition, the user-type view switching mechanism provides both a patient-

centered view and a caregiver/conductor view. These views support inspection of the scene from different perspectives and were included based on feedback from the physiotherapy collaborators

IV. RESULTS

The prototype integrated environment creation, object manipulation, inventory management, user-type view switching, and wheelchair navigation in the same environment. Fig. 4 shows the simplified residential layout used to test the prototype. This scene was not modeled from a real patient's house and does not represent validated residential dimensions. Its purpose is to demonstrate the interaction flow and the main functions implemented in the system.

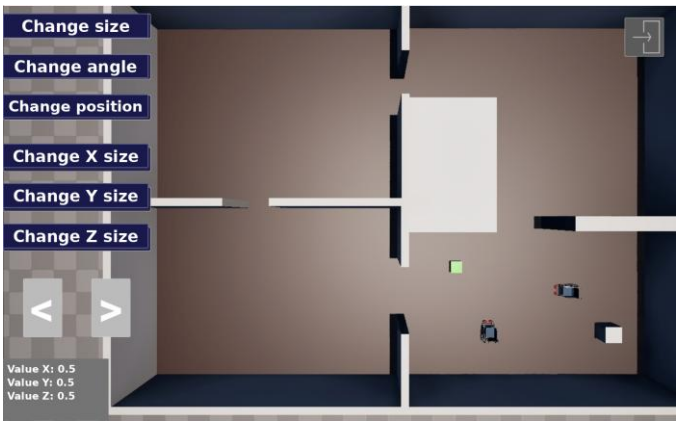


Fig. 4. Virtual environment representing an overview of walls and floor plan of a house developed for the simulation platform.

The view switching mechanism allowed transition between the view of the person conducting or assisting the wheelchair user and the patient-centered view, as shown in Fig. 5. This function was implemented to support both navigation and scene inspection. It does not represent a clinical outcome measure, but it helps the user observe the layout from different perspectives during planning.

Fig. 6 shows the inventory interface and examples of available objects. The implemented inventory allows furniture to be inserted into the scene and adjusted through the editing tools. The current inventory is limited, but it is sufficient for testing the basic workflow of object insertion, repositioning, and repeated navigation.

As a whole, the prototype allows users to create simplified residential layouts, insert furniture from an inventory, move and rotate objects, switch user-type views, and navigate the environment using a wheelchair model. Basic structures such as rooms, walls, corridors, and circulation areas can be modified during execution, enabling rapid layout changes without rebuilding the scene.

The current version does not reproduce all physical aspects of a real wheelchair or generate clinical metrics automatically. Table I summarizes what was functionally verified and what remains outside the current scope.



Fig. 5. Camera switching between patient-centered and companion views.

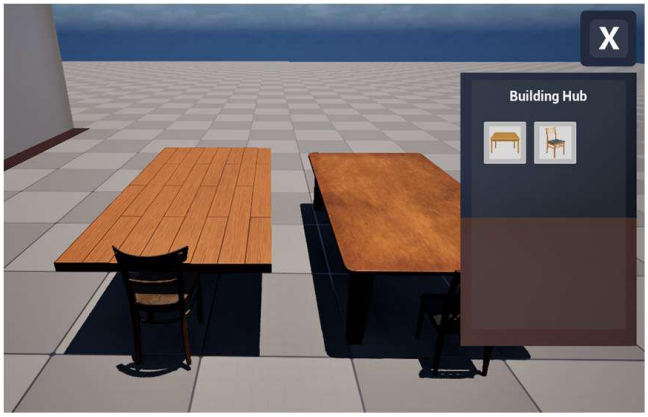


Fig. 6. Inventory interface and examples of available objects

TABLE I. FUNCTIONAL VERIFICATION AND CURRENT SCOPE

Function	*Current implementation*	*Current limitation*
Residential layout	Simplified rooms, walls, doors, windows, corridors, and circulation areas.	No geometric validation with real homes yet.
Object manipulation	Translation, rotation, resizing, and repositioning during execution.	No automatic measurement of distances or clearances yet.
Inventory	Insertion of tables, chairs, cabinets, and structural elements.	Limited object categories.
Wheelchair navigation	Keyboard control (W/A/S/D) with basic collision.	Simplified dynamics; no biomechanical model.
User-type views	Switching between patient-centered and caregiver/conductor views.	Not yet immersive VR.
Accessibility metrics	Not implemented in the current version.	Turning radius, trajectory, time, effort, and passage width remain future work.

V. DISCUSSION

The prototype demonstrates preliminary technical feasibility for a visual-interactive environment aimed at wheelchair prescription planning. Its main purpose is to allow the professional to modify a simplified residential layout and immediately observe how a wheelchair model behaves in that updated space. This is different from claiming clinical validation or objective accessibility assessment, which were not performed in this stage.

Compared with wheelchair simulators focused on user training and driving performance [6], [7], [9], the proposed system gives more emphasis to environmental modification by the clinician. Monteiro et al. [11] presented an accessibility evaluation tool that captures trajectory and time metrics, whereas the current prototype prioritizes scene editing, object inventory, and view switching. Therefore, the differential is not the use of Unreal Engine 5 alone, but the combination of these functions in the same planning workflow.

The participation of physiotherapy collaborators in the project helped define initial requirements, such as the need to represent household elements, model obstacles, and switch between patient-centered and caregiver/conductor views. This involvement supported the design of the prototype, but it should not be confused with formal clinical validation. The system still needs usability testing with physiotherapists, comparison with conventional assessment procedures, and testing based on real residential measurements.

Several limitations remain. The virtual environment is simplified and was not geometrically validated against real homes. The current wheelchair model uses simplified movement and basic collision only. It does not calculate turning radius, minimum passage width, trajectory, time, maneuvering effort, or collision metrics. It also does not include anthropometric models of patients, biomechanical simulation, different commercial wheelchair models, or complex architectural adaptations frequently found in real residences.

Table II compares the main characteristics of the proposed system with those of related wheelchair simulation studies. In this comparison, “VR” refers to immersive virtual reality as described in the corresponding studies. The proposed prototype, in contrast, currently operates as a non-immersive interactive 3D environment. The term “Prescription Planning Support” in the table refers only to visual support for environmental simulation and planning and does not indicate clinically validated decision-making support. Although the current prototype cannot yet determine whether a specific wheelchair model should be prescribed, it supports the identification of potential spatial constraints through visualization.

TABLE II. COMPARISON BETWEEN RELATED WORKS AND THE PROPOSED SYSTEM

Work	*VR Environment*	*Environment Customization*	*Wheelchair Simulation*	*Clinical Prescription Support*
Yang et al. [6]	Yes	No	Yes	No
Bigras et al. [7]	Yes	No	Yes	No
Riera et al. [8]	Yes	No	Yes	No
Chaar and Archambault [9]	Yes	Limited	Yes	No
Monteiro et al. [11]	Yes	Limited	Yes	Partial
Proposed System	Not yet	Yes	Yes	Visual planning only

VI. CONCLUSION

This paper presented a virtual simulation prototype developed in Unreal Engine 5 for environmental analysis during wheelchair prescription planning. The current version allows the user to edit simplified home layouts, manipulate objects, insert elements from an inventory, switch between patient-centered and caregiver/conductor views, and navigate with a simulated wheelchair.

The study demonstrates preliminary technical feasibility of integrating these functions in the same interactive environment. The system is not intended to replace clinical assessment and should not yet be interpreted as a validated prescription tool. Its current contribution is to provide a visual

environment in which spatial configurations can be created, modified, and inspected during planning.

The main lessons from development are that real-time editing, reusable object inventories, and user-type view switching are relevant functions for this type of application. At the same time, the work shows important research gaps: the need for quantitative accessibility metrics, geometric validation with real homes, more realistic wheelchair dynamics, and formal usability studies with physiotherapists.

Future work will focus on enhancing wheelchair movement, adding measures such as turning radius, trajectory, time, collisions, and minimum passage width, expanding the inventory, and testing the system with real residential scenarios. Further studies could also examine how each system module contributes to the overall functionality of the platform, including through controlled ablation experiments. Other possibilities include integration with spatial scanning, mobile devices, immersive virtual reality, augmented reality, and automatic accessibility analysis methods. These steps are necessary before evaluating the clinical utility of the platform in actual prescription workflows.

REFERENCES

[1] F. Routhier, C. Vincent, J. Desrosiers, and S. Nadeau, "Mobility of wheelchair users: A proposed performance assessment framework," Disability and Rehabilitation, vol. 25, no. 1, pp. 19–34, 2003.

[2] World Health Organization, Guidelines on the Provision of Manual Wheelchairs in Less-Resourced Settings. Geneva, Switzerland: WHO Press, 2008.

[3] L. Fehr, W. E. Langbein, and S. B. Skaar, "Adequacy of power wheelchair control interfaces for persons with severe disabilities: A clinical survey," Journal of Rehabilitation Research and Development, vol. 37, no. 3, pp. 353–360, 2000.

[4] P. L. Weiss and N. Katz, "The potential of virtual reality for rehabilitation," Journal of Rehabilitation Research and Development, vol. 41, no. 5, pp. vii–x, 2004.

[5] D. Levac, S. Glegg, H. Colquhoun, P. Miller, and F. Noubary, "Virtual reality and active video game-based practice, learning needs, and preferences: A Cross-Canada survey of physical therapists and occupational therapists," Games for Health Journal, vol. 6, no. 4, pp. 217–228, 2017.

[6] G. Vailland, D. Louise, F. Pasteau, F. Nouviale, B. Fraudet, E. Leblong, M. Babel, and V. Gouranton, "VR based power wheelchair simulator: Usability evaluation through a clinically validated task with regular users," in 2021 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2021, pp. 420–427.

[7] C. Bigras, S. Bouchard, and A. Leroux, "Augmented feedback for powered wheelchair training in a virtual environment," Journal of NeuroEngineering and Rehabilitation, vol. 16, no. 1, pp. 1–13, 2019.

[8] J. V. Riera, B. Palma, P. Casanova-Salas, M. Pérez, J. Gimeno, and M. Fernandez, "Development and evaluation of a haptic virtual walker for wheelchair users in immersive VR environments," Applied Sciences, vol. 15, no. 1, Art. no. 23, 2025.

[9] F. Chaar and P. S. Archambault, "Usability of a virtual reality manual wheelchair simulator," Disability and Rehabilitation: Assistive Technology, vol. 18, no. 8, pp. 1489–1499, 2023.

[10] D. P. Salgado, S. Fallon, Y. Qiao, and E. Naves, "WheelSimPhysio-2023 dataset: Physiological and questionnaire-based dataset of immersive multisensory wheelchair simulator from 58 participants," Data in Brief, vol. 54, Art. no. 110535, 2024.

[11] E. P. Monteiro, A. Campos,M. Gitirana and R. Oviedo, "Virtual Wheelchair Accessibility Evaluation Tool for Virtual Environments: Design and Preliminary Tests" in Proc. Symposium on Virtual and Augmented Reality (SVR), 2023, pp. 1–5.

[12] C. Zorzi, A. Covaci, and G. Marcelli, "Using virtual reality to complement traditional wheelchair skills training: A systematic review," Applied Sciences, vol. 15, no. 1, Art. no. 187, 2025.

[13] Q. Wang, M. Li, Y. Zhang, and H. Liu, "Effects of virtual reality-based rehabilitation on motor function: A systematic review and meta-analysis," Journal of NeuroEngineering and Rehabilitation, vol. 21, no. 1, Art. no. 45, 2024.

A Wearable Edge-Computing HD sEMG Platform for Assistive Applications: Technology Readiness and Laboratory-Scale Reproduction Cost

Ricardo Gonçalves Molinari
Center for Biomedical Engineering
University of Campinas
Campinas, Brazil
molinari@unicamp.br

Valeria Avilés-Carrillo
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
v222717@dac.unicamp.br

Cristian David Guerrero-Mendez
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
c197597@dac.unicamp.br

Natanael Pereira Batista
Center for Biomedical Engineering
University of Campinas
Campinas, Brazil
natanael@unicamp.br

Mário Sérgio Maduro Santana
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
m223615@dac.unicamp.br

Leonardo Abdala Elias
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
leoelias@unicamp.br

***Abstract*—We retrospectively assessed the technology readiness and laboratory-scale reproduction cost of a 128-channel heterogeneous wearable edge-computing platform that integrates high-density surface electromyography (HD sEMG) acquisition, flexible embedded processing, and autonomous on-device myoelectric control. The Technology Readiness Level (TRL) assessment covered hardware, software, and functional integration. Offline integrity checks found no missing or empty samples in the analyzed recordings. The platform exhibited acquisition-to-command latencies ranging from 63.3 ± 1.0 to 219.1 ± 4.5 ms. Across three full-load discharge tests, the mean runtime was 11 h 45 min. The accumulated evidence included an estimated 321 h of end-to-end myoelectric control exposure with physical and virtual terminal devices, including pilot demonstrations involving two participants with transradial amputations. Collectively, the evidence supported a TRL 6 classification. With reusable specialized infrastructure amortized across 25 units, the estimated laboratory-scale reproduction cost was USD 5,754.41 per unit, 20.2–94.7 % lower than the 2026 reference acquisition prices of selected commercial and modular systems offering at least 128 channels. Together, the technology readiness assessment and quantified reproduction cost provide a technical and economic basis for further validation in assistive technology and motor rehabilitation.**

***Index Terms*—high-density surface electromyography, motor rehabilitation, myoelectric control, edge-computing, technology readiness level, reproduction cost, assistive technology.**

The Brazilian Public Ministry of Labor (MPT) supported the work and funded postdoctoral fellowships for Ricardo G. Molinari under Contract No. 002118.2019 and for Natanael P. Batista under Contract No. 39433-26. The Coordination for the Improvement of Higher Education Personnel (CAPES) supported Valeria Avilés-Carrillo through the Academic Excellence Program (PROEX), Process No. 88887.842775/2023-00. The São Paulo Research Foundation (FAPESP) funded Cristian D. Guerrero-Mendez through PhD Scholarship No. 2025/03022-2. The Brazilian Innovation Agency (FINEP) supported Mário S. M. Santana under Agreement No. 01.24.0110.00 (Ref. 2140/22). The National Council for Scientific and Technological Development (CNPq) supported Leonardo A. Elias through Research Productivity Grant No. 316320/2023-4.

I. Introduction

Upper-limb functional loss resulting from amputation, neurological injury, or musculoskeletal disease compromises autonomy, activities of daily living, and social interaction. Studies of individuals with spinal cord injury report restrictions in autonomy in 72–80 % of evaluated participants [1]. Unequal access to assistive technologies across Latin America further underscores the need for adaptable, economically accessible rehabilitation systems [2].

Myoelectric interfaces can capture residual motor intention and translate it into commands to assistive and rehabilitation systems [3], [4], including prostheses [5], exoskeletons [6], visuomotor feedback [7], and virtual environments for motor training [8]. High-density surface electromyography (HD sEMG) further expands the capabilities of myoelectric interfaces by increasing the spatial resolution of acquisition and facilitating the decoding of hand motor intention [9].

Despite the potential of HD sEMG-based interfaces, their progression from laboratory prototypes to assistive and rehabilitation applications requires architectures that combine high-channel-count acquisition, computational flexibility, and autonomous wearable operation. Existing platforms often trade one requirement for another: microcontroller-centered systems face throughput and processing constraints, field-programmable gate array (FPGA)-centered systems increase development complexity, and distributed architectures depend on external computing resources and connectivity [10]. These limitations, together with high acquisition costs, proprietary architectures, and limited reporting on technology readiness and reproduction costs, hinder broader research and translational use [2].

The platform evaluated here addresses this gap through

a heterogeneous edge-computing architecture in which programmable logic performs time-critical 128-channel acquisition, while Arm Cortex-A53 cores running Linux/PYNQ execute myoelectric decoding, system control, and user interaction. Acquisition, processing, and command generation, therefore, remain on-device [10]. Building on the platform's previous technical characterization, this study evaluates its Technology Readiness Level (TRL) across hardware, software, and functional integration and quantifies the resources required to reproduce one functional unit at laboratory scale. The main contribution is the joint technical-economic characterization of the platform, linking accumulated system-level evidence to explicit readiness criteria and a transparent cost model encompassing components, labor, services, consumables, and reusable infrastructure. The combined perspective provides a structured basis for planning subsequent validation, laboratory replication, technology transfer, and translation to assistive and rehabilitation applications [11].

II. Methods

A. Technology Readiness Assessment

We assessed the platform's technology readiness using Mankins's TRL framework [12], [13]. The system under assessment comprised the entire wearable embedded chain, from 128-channel HD sEMG acquisition to on-device myoelectric command generation. The assessment covered three dimensions: hardware, including a ZU3EG embedded processing module, two analog front ends for HD sEMG acquisition, wireless communication, and battery power; software, encompassing the functions responsible for real-time myoelectric control; and functional integration, defined as the coordinated operation of the hardware and software with physical or virtual terminal devices. Prostheses and other external interfaces served only as terminal devices for functional demonstration, remained outside the platform scope, and received no separate TRL classification. In this study, Technology Readiness Assessment (TRA) refers to the structured process of mapping the available technical and functional evidence to the definitions and criteria of the TRL framework. The authors retrospectively mapped the available evidence to each assessed dimension, assigned the corresponding readiness levels, and resolved disagreements by consensus. For the global classification, we applied a conservative TRA rule: the lowest level across the three dimensions and, in cases of ambiguity, the lower level [13].

The assessment focused on TRL 4–6 because earlier work had documented evidence consistent with the lower levels [10], whereas progression to higher levels requires demonstration in an operational environment and, at later stages, real-world use. We defined supervised clinical use and unsupervised home use as operational environments [13]. We required three system-level criteria for TRL 6: an integrated prototype configuration, end-to-end operation, and demonstration in a relevant environment. Supervised laboratory protocols met the relevant-environment criterion when the battery-powered wearable system operated end-to-end during dynamic tasks involving body movement and interaction with physical or virtual terminal devices. Evaluation segments included movements, object interactions, and task states not contained in the brief calibration sequence. We regarded these conditions as functionally representative laboratory constraints, not as operational or clinical use. The two participants with transradial amputations provided additional target-population evidence but were neither treated as a sufficient sample for clinical inference nor used to claim an operational environment.

The Research Ethics Committee of the University of Campinas (UNICAMP) approved the studies involving human participants under CAAE Nos. 80299124.2.1001.5404 (October 24, 2024) and 59961616.8.0000.5404 (February 26, 2025). The ongoing studies described below were conducted under CAAE No. 90477225.9.0000.5404 (August 5, 2025). All procedures involving human participants were conducted in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

The ongoing evidence comprised visuomotor-tracking, physical and virtual manual-dexterity, and multimodal acquisition studies. The first two used the same core platform and one-degree-of-freedom proportional decoder, with approximately 100-ms command updates and per-session calibration based on 20 s of dynamic data, to control physical or virtual terminal devices during tracking, manipulation, the Box and Block Test (BBT), and tasks inspired by the Southampton Hand Assessment Procedure (SHAP); the manual-dexterity study included two participants with transradial amputations. The multimodal study synchronized myoelectric, kinematic, and visual signals without command generation and therefore provided subsystem-level evidence. Only sessions completed by June 17, 2026, were included; further protocol details are provided in the Supplementary Material [14].

For the reported counts, one "study enrollment" was defined as the inclusion of one individual in one study; therefore, enrollment counts did not necessarily represent unique participants. End-to-end exposure included acquisition, processing, command generation, and interaction with a physical or virtual terminal device. Because continuous device logs were not available for every session, cumulative exposure was estimated from the number of recorded sessions and the nominal duration of each protocol; the reported hours therefore represent approximate experimental exposure rather than precisely logged operating time.

B. Reproduction Cost Estimation

We estimated the cost of reproducing one functional platform unit at laboratory scale. We defined total cost as the sum of unit-specific materials and components (C_{mat}); reproduction-related services, labor, and consumables (C_{serv}); and an allocated share of reusable specialized infrastructure (C_{infra}/N), as shown in (1). The term C_{infra} includes bench equipment for soldering, printed circuit board rework, inspection, electrical validation, and 3D printing. The term N denotes the number of platform units across which we amortized infrastructure costs. The deterministic sensitivity analysis

varied C_{mat} by ± 30 % to examine favorable and adverse scenarios involving prices, exchange rates, freight, taxes, and material losses; varied C_{serv} by ± 20 %; and evaluated $N = 5$, 25, and 125.

$$C_{\mathrm{total}}(N) = C_{\mathrm{mat}} + C_{\mathrm{serv}} + \frac{C_{\mathrm{infra}}}{N} \tag{1}$$

We decomposed materials and components according to (2). The term C_{PCB} denotes printed circuit boards; C_{comp}, front-end electronic components; C_{mod}, commercial modules; C_{conn}, connectors, cables, and interfaces; and C_{power}, the power and charging system. Documented unit prices were used as the base inputs. Freight, taxes, and import charges were retained when included in the source price but were not added separately when unavailable; the potential effect of these procurement conditions was examined within the ± 30 % material-cost scenarios.

$$C_{\mathrm{mat}} = C_{\mathrm{PCB}} + C_{\mathrm{comp}} + C_{\mathrm{mod}} + C_{\mathrm{conn}} + C_{\mathrm{power}} \tag{2}$$

We descriptively compared the platform with commercial and modular systems offering 128 channels or more. We expressed acquisition costs in 2026 US dollars (USD) and adopted June 17, 2026, as the economic base date. We updated historical US-dollar-denominated prices using the US Consumer Price Index for All Urban Consumers (CPI-U) [15] and historical euro-denominated prices using the Harmonised Index of Consumer Prices (HICP) for the euro area [16]. We converted Brazilian-real-denominated (BRL) values using an exchange rate of BRL 5.05/USD and euro-denominated (EUR) values using the cross-rate implied by BRL 5.88/EUR and BRL 5.05/USD. We did not assess functional or regulatory equivalence among systems. When applicable, we reported quantitative results as mean $\pm$ standard deviation. The Supplementary Material [14] provides the input data, economic assumptions, and calculation records supporting the cost analysis.

III. Results

A. Technical and Functional Evidence of the Platform

The technical evidence supported three defining system-level capabilities of the platform: reliable high-density acquisition, embedded real-time processing, and autonomous wearable operation. As previously reported [10], the acquisition subsystem continuously acquired signals from 128 channels at 2052.52 Hz per channel. Offline integrity checks found that the analyzed recordings contained the number of samples expected from their duration and sampling rate, with no empty sample values. The signal-to-noise ratio (SNR) was 13.93 ± 7.51 dB. The processing module achieved 8.5 billion floating-point operations per second in the LINPACK benchmark. Previously published characterization reported mean acquisition-to-command latencies ranging from 63.30 ± 1.01 ms for a 30-ms acquisition window to 219.11 ± 4.47 ms for a 150-ms window, with the reported standard deviations representing timing jitter. This metric includes acquisition-window duration, embedded processing, and controller area network (CAN) command transmission [10].

TABLE I
Retrospective evidence-to-criterion mapping supporting the TRL 6 classification.

Dimension: Hardware	**TRL 6**
Integrated prototype: Wearable integration of 128-channel acquisition, processing, communication, and battery power.	
End-to-end operation: Continuous hardware operation supporting acquisition, embedded processing, command generation, and communication with physical or virtual terminal devices during task execution.	
Relevant environment: Dynamic tasks involving user movement and interaction with physical or virtual terminal devices.	
Dimension: Software	**TRL 6**
Integrated prototype: Concurrent acquisition, processing, and myoelectric command generation on the embedded platform.	
End-to-end operation: Continuous data flow from acquisition to control of the terminal device through a common software core.	
Relevant environment: Real-time decoding of motor intention during dynamic tasks and interaction with a terminal device.	
Dimension: Functional integration	**TRL 6**
Integrated prototype: Functional coupling of the platform to multiple physical and virtual terminal devices while preserving the core architecture.	
End-to-end operation: Continuous myoelectric control of physical and virtual terminal devices for approximately 321 h of estimated cumulative exposure.	
Relevant environment: Demonstrations with physical and virtual terminal devices, including approximately 10.8 h with two participants with transradial amputations.	
Global TRL: 6 (lowest level across dimensions)	

Note: Because multimodal acquisition included only signal acquisition and synchronization, we treated it solely as subsystem evidence.

Autonomous operation was characterized across three independent battery units, with a mean power consumption of 6.02 ± 0.11 W and a mean runtime of 11 h 45 min under continuous full-load operation until cutoff by the battery management system (BMS) [10]. The power-subsystem architecture, monitoring conditions, and charge–discharge characterization are detailed elsewhere [17]. In the visuomotor feedback tasks, the estimated motor intention reached a peak normalized cross-correlation of 0.80 ± 0.16 with the reference trajectory [10]. Across all end-to-end scenarios, the acquisition-to-command chain remained on-device; only the experimental task and the physical or virtual terminal device varied.

Our joint assessment of hardware, software, and functional integration classified the platform as TRL 6 (Table I). The evidence base combined the previously published initial validation with the ongoing studies described in Methods. The visuomotor-tracking and manual-dexterity studies contributed end-to-end evidence, whereas the multimodal study contributed subsystem-level evidence. The manual-dexterity study included pilot demonstrations with two participants with transradial amputations (Fig. 1).

Across the studies, we counted 67 study enrollments and 136 sessions, corresponding to an estimated 359 h of cumulative use (Table II). End-to-end myoelectric control exposure was calculated as 84 h of initial validation plus 180 h of

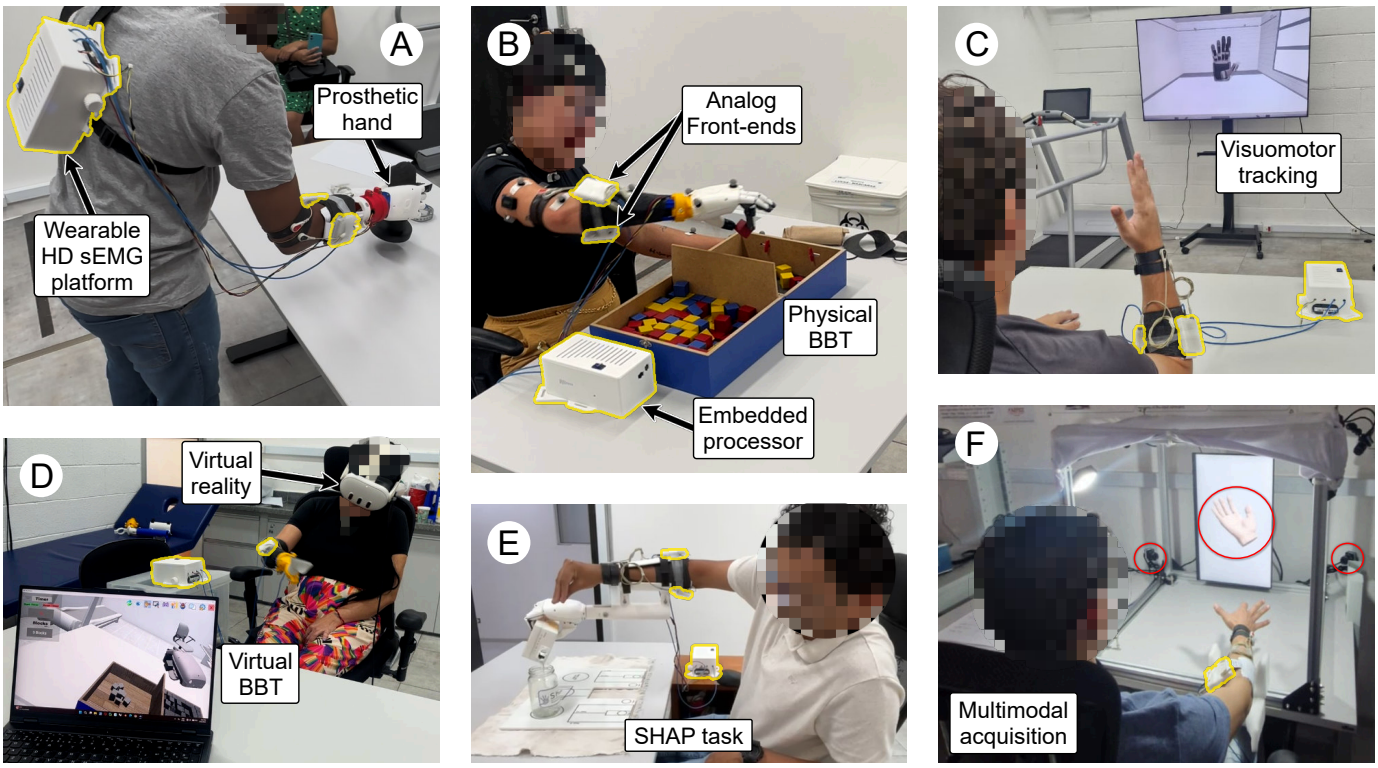


Fig. 1. Functional evidence of the platform in representative scenarios. Panels A, B, D, and E depict tasks included in pilot demonstrations with two participants with transradial amputations. **(A)** Wearable use with a myoelectric prosthesis during a manipulation task. **(B)** Physical Box and Block Test (BBT) using a prosthetic hand with simultaneous kinematic recording. **(C)** Real-time virtual-hand control during visuomotor tracking. **(D)** Virtual BBT with arm tracking and myoelectric hand control. **(E)** Prosthetic-hand control during a Southampton Hand Assessment Procedure (SHAP)-inspired task. **(F)** Synchronized acquisition of myoelectric signals, markerless kinematic data, and a visual reference.

TABLE II
ESTIMATED CUMULATIVE EXPERIMENTAL EXPOSURE OF THE PLATFORM ACROSS USAGE SCENARIOS.

Scenario	Enrollments	Sessions	Exposure
Initial validation†	21	21	84 h
Visuomotor tracking‡	15	60	180 h
Physical/virtual BBT and SHAP‡	8*	32	57 h
Multimodal acquisition‡	23	23	38 h
Total	**67**	**136**	**359 h**

† Published data [10]; ‡ ongoing studies described in the Supplementary Material [14]; * includes two participants with transradial amputations. BBT = *Box and Block Test*; SHAP = *Southampton Hand Assessment Procedure*.

visuomotor tracking and 57 h of physical/virtual BBT and SHAP-inspired tasks, totaling approximately 321 h. A further 38 h of multimodal acquisition was counted only as subsystem evidence. The visuomotor-tracking and manual-dexterity cohorts did not overlap. However, we could not determine whether participants in either cohort also took part in the initial validation or multimodal acquisition studies. As a result, we could not calculate the total number of unique participants.

Descriptive and exploratory pilot demonstrations involving two participants with right transradial amputations of the dominant upper limb accounted for approximately 10.8 h of use. Neither participant had previous experience using a prosthesis. During the physical and virtual BBT and the task inspired by the Southampton Hand Assessment Procedure (SHAP) [18], both used a 3D-printed socket coupled to the open-source Brunel Hand 2.0 prosthetic hand. From the first to the fourth session, participant 1 improved from 5 to 12 blocks in the physical BBT (+140.0 %), from 8.6 ± 1.1 to 13.8 ± 1.1 blocks in the virtual BBT (+60.5 %), and from 14.56 to 35.78 points in the SHAP-inspired task (+145.7 %). For participant 2, the corresponding values increased from 8 to 9 blocks (+12.5 %), from 9.6 ± 1.7 to 12.6 ± 1.1 blocks (+31.3 %), and from 39.22 to 43.98 points (+12.1 %). These demonstrations provided evidence of integrated platform operation with participants from the target population and contributed to the technology readiness assessment.

B. Laboratory Reproduction Cost Estimate

In the reference scenario ($N = 25$), we estimated the laboratory-scale reproduction cost of one platform unit at $C_{\text{total}}(25) =$ USD 5,754.41 (Table III); materials and components accounted for 62.49 % of the total. In the sensitivity analysis, varying C_{mat} by ± 30 % changed total cost by ± 18.75 %, whereas varying C_{serv} by ± 20 % changed total

TABLE III
LABORATORY-SCALE REPRODUCTION COST OF ONE PLATFORM UNIT IN THE REFERENCE SCENARIO ($N = 25$).

Category	Estimated cost (2026 USD)
Printed circuit boards (C_{PCB})	110.50
Front-end electronic components (C_{comp})	1,281.60
Commercial modules (C_{mod})	691.93
Connectors, cables, and interfaces (C_{conn})	820.02
Power and charging system (C_{power})	691.65
Materials and components (C_{mat})	**3,595.70**
Electronic assembly and electrical validation	1,542.57
3D-printed enclosure and associated labor	386.14
Services, labor, and consumables (C_{serv})	**1,928.71**
Specialized infrastructure amortized over 25 units ($C_{\mathrm{infra}}/25$)	230.00
Total reproduction cost (C_{total})	**5,754.41**

TABLE IV
DETERMINISTIC SENSITIVITY ANALYSIS OF THE REPRODUCTION COST.

Scenario	Total cost (USD)	Variation
Reference: $N = 25$	5,754.41	Reference
C_{mat}: -30 %	4,675.70	-18.75 %
C_{mat}: $+30$ %	6,833.12	$+18.75$ %
C_{serv}: -20 %	5,368.67	-6.70 %
C_{serv}: $+20$ %	6,140.15	$+6.70$ %
$N = 5$	6,674.41	$+15.99$ %
$N = 125$	5,570.41	-3.20 %
Combined adverse*	8,138.86	$+41.44$ %

* A 30 % increase in C_{mat}, a 20 % increase in C_{serv}, and $N = 5$. For the individual analyses, we changed only the indicated parameter.

cost by ± 6.70 %. Changing the amortization scale to $N = 5$ or $N = 125$ changed total cost by $+15.99$ % and -3.20 %, respectively. The combined adverse scenario increased the cost to USD 8,138.86, 41.44 % above the reference scenario (Table IV). After expressing all values in 2026 US dollars, the $N = 25$ estimate was 20.2–94.7 % lower than the acquisition costs of commercial and modular systems offering 128 channels or more (Table V).

TABLE V
COST COMPARISON OF COMMERCIAL OR MODULAR MULTICHANNEL MYOELECTRIC/BIOPOTENTIAL ACQUISITION SYSTEMS AND THE PROPOSED PLATFORM.

System	Cost (2026 USD)
Proposed platform ($N = 25$)	**5,754.41**
Open Ephys Acquisition Board, 128 channels (two 64-channel modules) [19]	7,213.19
Intan RHD Recording System, 128 channels [20]	13,160.00
g.HIamp 256 [21]	45,471.65
Quattrocento 128+256 [22]	44,524.25
BioSemi ActiveTwo/ActiveThree 128+8 [23]	52,396.04
actiCHamp 128/160 [24]	108,162.20

Note: We adjusted historical prices using indices compatible with the source currency (CPI-U for USD and the euro-area HICP for EUR) and converted current prices using the adopted exchange rates. The platform value represents laboratory-scale reproduction at $N = 25$; differences in purpose and regulatory status preclude direct equivalence with the listed systems.

IV. DISCUSSION

A. Technology Readiness and System-Level Significance

Applying the retrospective criteria, we classified the platform as TRL 6 because it progressed beyond validation of isolated subsystems and operated as an integrated system in a relevant experimental environment [12], [13]. Although laboratory supervision distinguished these protocols from operational use, the platform operated under functionally representative constraints for the intended application: continuous battery-powered wearable operation, body movement, interaction with objects and physical or virtual terminal devices, and, in some demonstrations, involvement of participants from the target population. In the end-to-end demonstrations, signal acquisition, processing, and command generation remained embedded and did not depend on bench equipment or external computing resources; external systems were used only to implement the experimental task or to serve as the terminal device. The TRL 6 classification, therefore, reflects the maturity of the integrated system demonstrated under these laboratory conditions.

The previously reported processing and command-transmission time of 33.3–69.1 ms remained within the range associated with favorable myoelectric control performance [10], [25]. The previously characterized mean runtime of 11 h 45 min across three full-load discharge tests supported prolonged experimental protocols without external power [10], [17]. Pilot demonstrations with two participants with transradial amputations provided additional evidence of integrated use in the target population. However, the small exploratory sample does not support inferences regarding robustness across users and use contexts or clinical effectiveness. The lack of demonstration in an operational environment, particularly in supervised clinical or unsupervised home use, precludes classification as TRL 7. The available evidence also does not support claims of regulatory qualification or commercial availability.

B. Reproduction Cost and Translational Potential

The commercial-system comparison was secondary and descriptive: systems were selected according to channel count and the availability of price evidence, without requiring equivalence in intended use, electrodes, amplifiers, software, support, or certification. After adjustment to 2026 US dollars, the platform's estimated laboratory-scale reproduction cost was lower than the acquisition costs of all selected commercial and modular systems offering at least 128 channels. Because the estimate covers an integrated wearable edge-computing platform, including 128-channel acquisition, embedded processing, battery-powered operation, and control of terminal devices, rather than acquisition hardware alone, it may indicate lower barriers to laboratory replication and multicenter validation [10]. The estimate may support planning

for laboratory replication and subsequent economic evaluation, including future assessments of comparative cost-effectiveness and budget impact.

C. Lessons Learned

The assessment highlighted two methodological considerations relevant to the development of modular wearable myoelectric systems. First, explicit definition of the system boundary ensured that readiness evidence and the resulting classification referred consistently to the wearable acquisition, processing, and control platform, while physical and virtual terminal devices remained external use-case interfaces. Second, establishing TRL milestones prospectively can make readiness assessment an integral part of the development process rather than solely a retrospective classification exercise, providing a structured basis for design decisions, verification planning, configuration control, and evidence generation across successive stages of maturity.

D. Limitations

Our cost estimate covers the laboratory-scale reproduction of one unit, not industrial production or a commercial product with regulatory qualification. We excluded certification, testing for compliance with applicable standards, regulatory documentation, support, warranty, logistics, commercial margins, and post-deployment activities. The reference estimate assumes successful reproduction of one functional unit and does not model manufacturing yield, failed boards beyond the inspection and correction labor already included, repeated calibration, software maintenance, replacement parts, field servicing, or lifecycle support. These omissions distinguish the reported laboratory reproduction cost from industrial production and commercial lifecycle cost. The estimate also depends on commercially available components and reflects the prices, sourcing conditions, exchange rates, freight, and taxes adopted for the reference scenario; costs may therefore vary across countries and procurement contexts.

The economic comparison provides only an order-of-magnitude reference. Differences in intended purpose, certification, support, product finish, functional integration, and cost composition, along with inconsistent inclusion of taxes, freight, import charges, maintenance, and training, prevent direct inferences of functional equivalence or economic superiority.

The TRA was conducted retrospectively by consensus among the platform developers, using evidence from studies that were not originally designed around prospectively specified readiness criteria. Consequently, the classification depends on the retrospective interpretation of the available evidence and applies to the evaluated platform configuration under supervised laboratory conditions. As an ordinal framework, TRL summarizes readiness at a given stage but does not fully capture iterative or nonlinear development trajectories. Readiness should therefore be reassessed as the platform configuration and intended-use context evolve.

Cumulative exposure was estimated from session counts and nominal protocol durations because continuous device logs were not available for every experiment. In addition, some participants may have enrolled in more than one study, and the number of unique individuals could not be reconstructed. The exposure totals therefore characterize the accumulated use of the platform, not the duration of a controlled reliability test or the number of unique users.

E. Next Steps

Future work should first expand evaluation in participants with transradial amputations and subsequently extend it to populations with spinal cord injury, stroke, and neuromuscular diseases under conditions progressively closer to real-world use [1], [26], [4]. The ongoing extension from one to four degrees of freedom should be evaluated for decoding performance, end-to-end latency, robustness, and usability. Progression from TRL 6 to TRL 7 will require a stable, reproducible prototype configuration and demonstration in an operational environment, such as supervised clinical use or unsupervised home use.

Further translation will require a clearly defined intended use and corresponding regulatory pathway, supported by application-specific evidence of performance and usability. Progression toward scalable manufacturing will also depend on qualified suppliers, controlled production processes, product traceability, and maintenance planning. Adoption within public healthcare systems would further require evidence of clinical effectiveness, comparative cost-effectiveness, and budget impact, together with assessment of implementation requirements, workforce training, and long-term service sustainability.

V. Conclusion

The retrospective assessment supported a TRL 6 classification for the 128-channel wearable edge-computing HD sEMG platform, and its laboratory-scale reproduction cost was estimated at USD 5,754.41 per unit. Evidence accumulated across physical and virtual tasks demonstrated the coordinated operation of high-density acquisition, embedded myoelectric processing, battery power, and terminal-device control under representative laboratory conditions. By combining technology readiness assessment with a transparent reproduction cost model, this study establishes a system-level reference for platform replication and continued development toward deployment in assistive and rehabilitation applications.

Acknowledgments

The authors thank Flavio Renato Santos, Nelio Hilton Gava, and Glaudson Frade Assunção of the Main Design Lab of the Center for Biomedical Engineering at the University of Campinas for technical support in the design and fabrication of the platform and prosthetic-hand prototypes.

REFERENCES

[1] L. Lili, K. S. Sunnerhagen, T. Rekand, and M. A. Murphy, "Participation and autonomy, independence in activities of daily living and upper extremity functioning in individuals with spinal cord injury," *Scientific Reports*, vol. 14, 12 2024.

[2] R. A. M. Collado, Y. H. Lee, and J. B. Kim, "Access to assistive technology and related services in latin american countries: A systematic review," pp. 37–46, 2026.

[3] D. S. Oliveira, M. Ponfick, D. I. Braun, M. Osswald, M. Sierotowicz, S. Chatterjee, D. Weber, B. Eskofier, C. Castellini, D. Farina, T. M. Kinfe, and A. D. Vecchio, "A direct spinal cord-computer interface enables the control of the paralysed hand in spinal cord injury," *Brain*, vol. 147, pp. 3583–3595, 10 2024.

[4] C. O'Connell, M. Guo, B. Soucy, M. Calder, J. Sparks, and S. Plamondon, "All hands on deck: The multidisciplinary rehabilitation assessment and management of hand function in persons with neuromuscular disorders," pp. 869–888, 5 2025.

[5] N. Jiang, C. Chen, J. He, J. Meng, L. Pan, S. Su, and X. Zhu, "Bio-robotics research for non-invasive myoelectric neural interfaces for upper-limb prosthetic control: a 10-year perspective review," *National Science Review*, vol. 10, 4 2023. [Online]. Available: https://academic.oup.com/nsr/article/doi/10.1093/nsr/nwad048/7057876

[6] X. Peng, S. Li, and L. Stirling, "Improving Complex Task Performance in Powered Upper Limb Exoskeletons With Adaptive Proportional Myoelectric Control for User Motor Strategy Tracking," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4655–4662, 2024.

[7] A. Sturma, L. A. Hruby, C. Prahm, J. A. Mayer, and O. C. Aszmann, "Rehabilitation of upper extremity nerve injuries using surface emg biofeedback: Protocols for clinical application," *Frontiers in Neuroscience*, vol. 12, 2018.

[8] Y. Xiao, Y. Gao, H. Bai, G. Song, H. Wang, J. S. Rao, A. Hao, X. Li, and J. Zheng, "Impact of an upper limb motion-driven virtual rehabilitation system on residual motor function in patients with complete spinal cord injury: a pilot study," *Journal of NeuroEngineering and Rehabilitation*, vol. 22, 12 2025.

[9] R. C. Sîmpetru, D. I. Braun, A. U. Simon, M. März, V. Cnejevici, D. S. de Oliveira, N. Weber, J. Walter, J. Franke, D. Höglinger, C. Prahm, M. Ponfick, and A. D. Vecchio, "Myogestic: Emg interfacing framework for decoding multiple spared motor dimensions in individuals with neural lesions," *Science Advances*, vol. 11, pp. 1–14, 2025.

[10] R. G. Molinari, V. Avilés-Carrillo, G. A. G. de Villa, and L. A. Elias, "A Wearable Platform for Real-Time Control of a Prosthetic Hand by High-Density EMG," *IEEE Access*, vol. 14, pp. 36 851–36 868, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/11421324/

[11] N. N. Mayrink, L. Alcoforado, E. J. V. de Oliveira, V. G. B. Junior, F. S. Toscas, E. Aragão, F. Fernandes, L. P. C. F. Alves, and R. A. Valentim, "Innovation in medical devices for the SUS: a translational pathway to support brazilian researchers," *Research on Biomedical Engineering*, vol. 41, 12 2025.

[12] J. C. Mankins, "Technology readiness levels: A white paper," National Aeronautics and Space Administration (NASA), Office of Space Access and Technology, Washington, DC, USA, White Paper, 1995.

[13] J. C. Mankins, "Technology readiness assessments: A retrospective," *Acta Astronautica*, vol. 65, pp. 1216–1223, 11 2009.

[14] R. G. Molinari, V. Avilés-Carrillo, C. D. Guerrero-Mendez, N. P. Batista, M. S. M. Santana, and L. A. Elias, "Supplementary material for a wearable edge HD sEMG platform for assistive applications: Technology readiness and laboratory-scale reproduction cost," 2026. [Online]. Available: https://doi.org/10.5281/zenodo.21342559

[15] U.S. Bureau of Labor Statistics, "CPI-U: All Items in U.S. City Average, Series CUUR0000SA0," https://data.bls.gov/timeseries/CUUR0000SA0, 2026, accessed: June 17, 2026.

[16] Eurostat, "Harmonised Index of Consumer Prices: Monthly Data (Index, 2015 = 100)," https://ec.europa.eu/eurostat/databrowser/view/prc_hicp_midx/default/table, 2026, euro area, all items; values for July 2020 and May 2026. Accessed June 17, 2026.

[17] R. G. Molinari, "Desenvolvimento e análise de uma plataforma baseada em sinais mioelétricos de alta densidade para controle em tempo real de próteses de mão," Doctoral dissertation, Universidade Estadual de Campinas, Campinas, SP, Brazil, 2025, faculdade de Engenharia Elétrica e de Computação; Graduate Program in Electrical Engineering; Biomedical Engineering concentration.

[18] C. M. Light, P. H. Chappell, and P. J. Kyberd, "Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: Normative data, reliability, and validity," *Archives of Physical Medicine and Rehabilitation*, vol. 83, pp. 776–783, 2002.

[19] Open Ephys, "Acquisition Board, 3rd Generation," https://open-ephys.org/acq-board, 2026, accessed: June 17, 2026.

[20] Intan Technologies, "Orders and Pricing," https://intantech.com/pricing.html, 2026, accessed: June 17, 2026.

[21] K. S. Mannatunga, S. H. Md Ali, M. L. Crespo, A. Cicuttin, and J. T. G. Samarawickrama, "High Performance 128-Channel Acquisition System for Electrophysiological Signals," *IEEE Access*, 2020, https://doi.org/10.1109/ACCESS.2020.3007082.

[22] Texas A&M Univ., "Purchase Order AB0575227: Quattrocento 128+256," https://spar.tamu.edu/Transparency/GetFile?dir=PoDocs&id=8772, 2020, accessed: June 17, 2026.

[23] BioSemi, "System Prices," https://www.biosemi.com/faq/prices.htm, 2026, accessed: June 17, 2026.

[24] P. Ledwidge, J. Foust, and A. Ramsey, "Recommendations for developing an EEG laboratory at a primarily undergraduate institution," *Journal of Undergraduate Neuroscience Education*, vol. 17, no. 1, pp. A10–A19, Dec. 2018, published December 15, 2018. PMID: 30618494; PMCID: PMC6312138. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6312138/

[25] T. R. Farrell and R. F. Weir, "The optimal controller delay for myoelectric prostheses," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 1, pp. 111–118, mar 2007.

[26] A. Hutchison, K. D'Cruz, P. Ross, and S. Anderson, "Exploring the barriers and facilitators to community reintegration for adults following traumatic upper limb amputation: a mixed methods systematic review," *Disability and Rehabilitation*, vol. 46, pp. 1471–1484, 2024.

ACL-Digital Assist: E-Health Platform for Knee Rehabilitation Tracking

Lucie Emilie Reichardt
ESIGELEC
Rouen, France
lucie.reichardt@groupe-esigelec.org

Akim Lamtamri
ESIGELEC
Rouen, France
akim.lamtamri@groupe-esigelec.org

Danilo Gondo Chinen
DAINF/DAELN
Universidade Tecnologica Federal do Paraná - UTFPR
Curitiba, Brazil
chinen@alunos.utfpr.edu.br

Fabio Kurt Schneider
DAELN/CPGEI
Universidade Tecnologica Federal do Paraná - UTFPR
Curitiba, Brazil
fabioks@professores.utfpr.edu.br

Mauricio Kugler
IRSEEM UR4353
ESIGELEC, Université Rouen Normandie
Rouen, France
mauricio.kugler@esigelec.fr

Pascal Falla
ESIGELEC
Rouen, France
Pascal.Falla@esigelec.fr

Benoît Decoux
ESIGELEC
Rouen, France
benoit.decoux@esigelec.fr

***Abstract*—Anterior Cruciate Ligament (ACL) reconstruction requires rigorous rehabilitation to restore knee functionality. Clinical monitoring relies on assessing joint range of motion and movement symmetry, yet practitioners often lack affordable, continuous tracking tools outside biomechanical laboratories. This paper presents ACL-Digital Assist, a low-cost, centralized e-health ecosystem designed for the longitudinal monitoring of lower limb rehabilitation. Built on a cross-platform framework, the architecture integrates a patient-facing mobile application and a web-based clinical dashboard, synchronized through a secure cloud backend utilizing strict row-level security. By leveraging the embedded inertial sensors of a standard smartphone secured to the patient distal tibia, the platform captures kinematic telemetry at 20 Hz. The system introduces distinct protocols to evaluate active mobility, passive joint limits, and isometric endurance, while computing a normalized acceleration variation score to detect physiological hesitations. To validate the technical feasibility and data pipeline integrity, a preliminary evaluation was conducted on healthy volunteers within a laboratory setting. The results indicate the system's capability to capture repeatable kinematic data, manage high-frequency telemetry via JSON storage arrays, and automatically generate standardized clinical reports, highlighting its promising potential to optimize physiotherapy workflows.**

***Index Terms*—Anterior cruciate ligament, biomechanical stability, cloud computing, mobile health, range of motion, row-level security, sensor telemetry.**

I. Introduction

Rupture of the Anterior Cruciate Ligament (ACL) is one of the most prevalent musculoskeletal injuries, frequently requiring surgical reconstruction to restore joint function and prevent long-term complications such as osteoarthritis [1], [2]. The rehabilitation process often spans several months, with the goal of restoring full joint functionality and preventing reinjury. Recovery relies heavily on achieving symmetrical knee mechanics, quantified in clinical settings using the Limb Symmetry Index (LSI). To ensure a safe return to sports, continuous tracking of the knee Range of Motion (ROM), movement stability, and static endurance is very important. Recent studies have demonstrated that mobile health-based interventions and smartphone self-rehabilitation applications can effectively complement standard care, significantly improving early functional outcomes, reducing pain, and increasing adherence in post-operative patients [3]–[5].

In current practice, longitudinal monitoring suffers from a technological divide. Precise biomechanical assessments rely on computerized isokinetic dynamometers or optical motion-capture systems. While highly accurate, they are exceedingly expensive and restricted to specialized facilities. Conversely, everyday clinical monitoring often depends on manual goniometers, which can exhibit clinically relevant measurement error and inter-operator variability [6], [7]. Traditional manual methods still lack automated historical logging, and only capture static angular boundaries rather than dynamic movement quality. In contrast, self-measurement of knee ROM using smartphone applications has proven to be a feasible and highly reliable alternative [8].

To help address this gap, this paper introduces ACL-Digital Assist, a low-cost e-health Proof-of-Concept designed to assist in monitoring post-ACL reconstruction rehabilitation. The architecture comprises two software applications sharing a unified, centralized cloud backend: a mobile application for telemetry acquisition and a web dashboard for healthcare professionals. By leveraging the embedded inertial sensors of a standard smartphone securely strapped to the distal tibia, the system extracts kinematic data to evaluate three clinical modalities: active movement, passive guided ROM, and isometric endurance. To address the need for qualitative movement analysis, the platform computes a normalized

stability score based on angular acceleration variations. To validate the technical feasibility and data pipeline integrity, a proof-of-concept (PoC) evaluation was conducted on healthy volunteers.

The remainder of this paper is organized as follows. Section II details the dual-application system architecture, kinematic processing, and clinical protocols. Section III presents the preliminary technical validation alongside the automated clinical reporting features. Finally, Section IV concludes the paper and discusses future work, including the integration of computer vision pipelines.

II. SYSTEM ARCHITECTURE AND BIOMECHANICAL METHODOLOGY

The system architecture is illustrated in a block diagram (Fig. 1), which details the interaction between the front-end data acquisition tools and the backend cloud infrastructure.

A. Mobile Application and Sensor Data Acquisition



Fig. 2. Experimental setup ensuring geometric accuracy.

The patient-facing application was developed using the React Native framework[1], a cross-platform mobile development environment, combined with Expo tools[2] to access native hardware capabilities. Real-time kinematic data is acquired via the Expo DeviceMotion API. Instead of extracting raw uncalibrated gyroscope data, the system leverages the native sensor fusion engine of the mobile operating system (OS) to obtain pre-filtered Euler angles, specifically the rotation around the X-axis (*DeviceMotionMeasurement.rotation.beta*). This design choice significantly reduces the computational load on the mobile CPU while providing stable kinematic readings.

Inertial Measurement Units (IMUs) have shown good accuracy in sagittal-plane knee motion, offering a reliable and accessible alternative to complex motion-capture systems for longitudinal tracking [9], [10]. The mobile device is firmly secured to the distal tibia using an elastic armband (as illustrated in Fig. 2, displaying the leg setup). To guarantee accuracy, strict starting postures are enforced: seated with thighs perpendicular to the torso for extension exercises, and standing straight for flexion exercises.

Data sampling operates at a sampling frequency f_s of 20 Hz utilizing the smartphone's embedded Inertial Measurement Unit (IMU). This rate of 20 Hz was selected as a practical compromise between mobile-device constraints and the need to capture low-frequency rehabilitation motions, as biomechanical studies indicate that the knee joint primarily exhibits low-frequency dynamic responses during movement [19]. To maintain acquisition stability, the telemetry strictly operates in the application foreground, preventing OS background sleep modes. Furthermore, the architecture mandates an active Wi-Fi connection, as data is directly streamed to the cloud without local offline caching. To enhance clinical usability, a "Monitor Mode" was implemented, allowing the physiotherapist to control the exercise flow and view real-time feedback on a secondary synchronized device (tablet or smartphone, as depicted in Fig. 3).

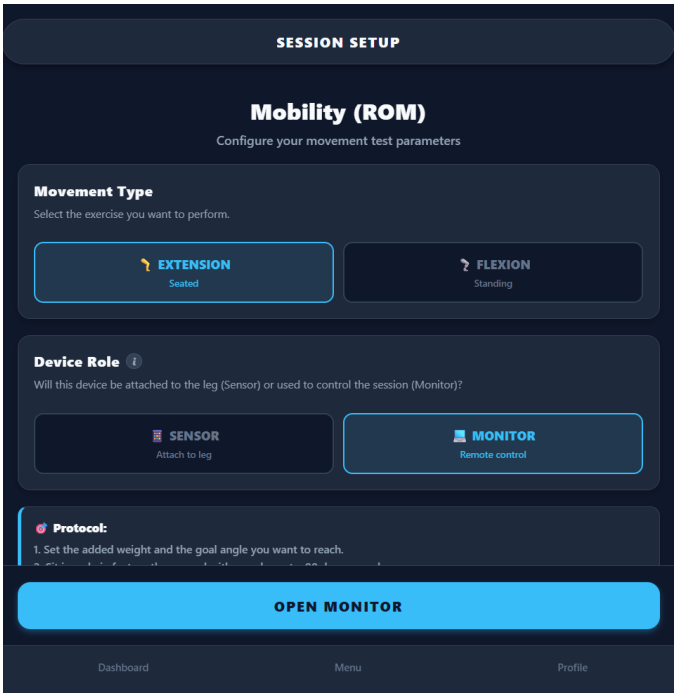


Fig. 3. The Monitor Mode interface displayed on a secondary synchronized device, allowing the physiotherapist to control the exercise state machine and monitor real-time kinematic feedback.

B. Kinematic Processing and Repetition State Machine

Let θ_{rest} define the resting baseline angle, θ_{th1} the initial trigger threshold angle, and θ_{th2} the validation threshold angle, with Δt representing the movement duration.

The angular data is first converted from radians to degrees. To map the telemetry to a clinically relevant plane, the algorithm references an θ_{rest} absolute vertical baseline. For instance, a patient standing with the leg straight registers at an

[1]https://reactnative.dev/, accessed July 2026.
[2]https://docs.expo.dev/, accessed July 2026.

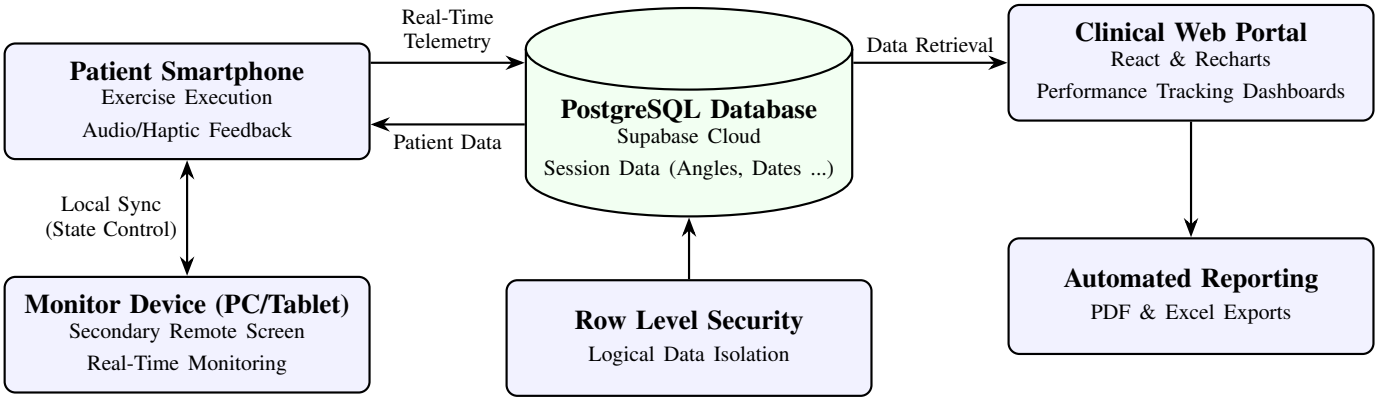


Fig. 1. System architecture of ACL-Digital Assist. The patient smartphone acts both as an exercise execution device and a personal dashboard retrieving data from the Supabase cloud. The physiotherapist controls the flow locally via a secondary monitor device (PC or tablet). The web portal accesses the cloud for clinical analytics and performance tracking.

θ_{rest} typically equal to $90°$; flexing the knee backwards until the lower leg is parallel to the floor approaches $0°$, and further flexion yields negative values. This mathematical continuity natively supports measuring a Range of Motion exceeding $90°$.

A dynamic state machine governs repetition counting. A repetition is initialized when the knee angle drops below an θ_{th1} threshold (e.g., $65°$) and concludes when returning past an θ_{th2} threshold (e.g., $70°$). Based on empirical tests conducted on healthy subjects, an invalidation criterion is applied: repetitions are only logged if the total movement duration Δt is superior or equal to 500 ms, effectively filtering out transient noise.

C. Multi-Modal Exercise Protocols

The platform supports three distinct clinical protocols:

- Extension Exercises (Fig. 4): This module evaluates knee extension. In the Active Dynamic mode, the subject performs five voluntary repetitions per limb, with the system logging the peak Range of Motion (ROM) for each repetition. The protocol mandates testing the healthy leg first to ensure a natural rest period during the arm-band transfer, preventing central nervous system fatigue crossover. In the Passive ROM mode, the patient remains relaxed while the practitioner manually guides the limb to its maximum extension, with data acquisition controlled via the monitor device's state machine to prevent parasitic noise during setup.
- Flexion Exercises (Fig. 5): This module evaluates knee flexion. Similar to the extension protocol, it supports two modalities: Active Dynamic repetitions, where the subject performs five voluntary movements to evaluate the peak flexion ROM, and Passive ROM, where the practitioner guides the limb to its maximum amplitude. Both modes maintain the same rigorous safety protocols, including initial testing of the healthy limb and the use of the remote state machine for passive sessions to ensure accurate, artifact-free data logging.
- Isometric Hold and Endurance (Fig. 6): The patient maintains the limb at a target angle. An endurance timer runs as long as the angle remains within a $\pm 10°$ tolerance window, which absorbs natural anatomical variances such as the posterior tibial slope. Real-time haptic, audio, and visual feedback guide the patient. A 2-second grace period accommodates minor postural adjustments and is currently included in the final endurance tally.

At the conclusion of any session, the application prompts the patient to input a self-reported pain score and optional clinical comments.



(a) (b)

Fig. 4. Extension exercise protocol. (a) Initial baseline posture. (b) Execution of the voluntary movement to evaluate the peak extension Range of Motion (ROM) in the active dynamic modality.



(a) (b)

Fig. 5. Flexion exercise protocol. (a) Initial baseline posture. (b) Execution of the voluntary movement to evaluate the peak flexion Range of Motion (ROM) in the active dynamic modality.



Fig. 6. Isometric hold and endurance protocol. The patient maintains the limb at a specific target angle, guided by real-time haptic, audio, and visual feedback within a defined tolerance window.

D. Biomechanical Assessment Metrics

Limb Symmetry Index (LSI): For dynamic exercises, the system computes the ROM as $90° - \theta_{\text{maxAngle}}$. The LSI is then derived from the average ROM of both legs:

$$\text{LSI}_{\text{dynamic}} = \left(\frac{x}{y}\right) \times 100\,, \tag{1}$$

where x denotes the average ROM of the rehabilitated limb, and y denotes the average ROM of the healthy limb, yielding a result expressed as a percentage. Longitudinal tracking models indicate that the uninvolved limb's function also fluctuates during rehabilitation, underscoring the necessity of continuously monitoring both limbs to derive an accurate LSI rather than relying on a static baseline [11]. Furthermore, while LSI is a standard metric, relying solely on it for return-to-sport clearance has shown poor discriminatory ability for preventing secondary injuries, highlighting the need for the multi-modal assessments provided by this platform [12]. To prevent mathematical instability or division by zero in cases of failed extension, the algorithm enforces a safety constraint: if healthyAvgROM $\leq 5°$, the dynamic LSI is defaulted to 0.

Biomechanical Stability Score: To quantify hesitations, an acceleration-based stability score is computed. The angular velocity v_i is defined by $\frac{\theta_i - \theta_{i-1}}{\Delta t}$, where Δt is the sampling interval. Variations in acceleration are accumulated $(\sum |v_i - v_{i-1}|)$ and divided by the absolute angular distance $(\sum |\theta_i - \theta_{i-1}|)$. The result is normalized into a 0 to 100 scale. While small interval variations may occur due to OS processing, the normalization over the total movement minimizes their impact.

E. Cloud Infrastructure

The ecosystem utilizes a centralized Supabase[3] cloud backend powered by an open-source PostgreSQL[4] database. To prevent high-frequency telemetry from causing database bloat, the 20 Hz raw data points are compressed and stored as JSON[5] (JavaScript Object Notation, a lightweight, text-based data interchange format) arrays, within single session records. Privacy is enforced natively via logical Row Level Security (RLS) policies, restricting data access based on authenticated user identities.

F. Clinical Analytics and Dashboard Reporting

To transform raw telemetry into actionable insights, a dedicated web-based dashboard processes the backend data. The system executes temporal aggregations, grouping post-surgical data by elapsed weeks (e.g., "W+2", "W+4") to allow for cohort-independent comparisons regardless of the actual surgery date. The dashboard calculates weekly averages for the LSI, ROM, hold durations, and stability scores while automatically filtering null values. To monitor patient safety, a clinical alert engine flags sessions where the self-reported pain score exceeds 7/10. The analytics engine also splits exercises into "mobility" and "strength" categories using case-insensitive filtering, normalizing movement angles to overlay curves coherently for visual comparison between the healthy and rehabilitated leg.

G. Hybrid Approach

A primary limitation encountered with the current single-sensor IMU solution is its inability to detect potential postural compensations made by the patient during exercises. Literature shows that ACL patients may show compensatory trunk/proximal strategies, such as forward trunk leaning, ipsilateral trunk lean, and altered hip kinematics to compensate for knee deficits [13]–[15]. Moreover, reduced degrees-of-freedom IMU setups are particularly susceptible to measurement errors induced by frontal plane misalignments [16]. These compensations can artificially alter the recorded angles, thereby potentially impacting the accuracy of the clinical results. While various tracks are possible to address this issue, the chosen solution, which will form the core of future work, is the combination of our current system with an image-based approach. The objective will be to film the performed exercises to extract information regarding the correct or incorrect execution of the movement. The subsequent section will discuss how this visual computation-based approach translates into our future developments.

[3]https://supabase.com/, accessed July 2026.

[4]https://www.postgresql.org/, accessed July 2026.

[5]ECMA International, Standard ECMA-404. https://www.json.org/, accessed July 2026.

III. PRELIMINARY RESULTS AND DISCUSSION

A. Technical Feasibility and Pipeline Validation

To validate the data pipeline, state machine responsiveness, and cloud synchronization, a preliminary Proof-of-Concept (PoC) evaluation was conducted on a limited number of healthy volunteers. The objective was strictly to validate the technical integration, rather than providing clinical diagnostic validation.

Due to the lack of access to standard calibrated goniometers or optical motion-capture systems, an exploratory concurrent assessment of the sensor logic was performed using the MoveNet Thunder deep learning model. This model was selected as a comparative baseline due to its established performance in remote knee Range of Motion (ROM) tracking. [17]. However, it must be strongly emphasized that this vision-based tool does not constitute a sufficient clinical reference standard, as both the smartphone IMU and the 2D camera model are inherently affected by placement variations and two-dimensional geometric assumptions. The smartphone was tested alongside a synchronized webcam setup under controlled lighting conditions during 10 dynamic repetitions of both knee extension and knee flexion.

To quantify the system's precision, the Mean Absolute Error (MAE) and the Standard Deviation (SD) of the error were calculated in degrees. The MAE serves as an indicator of the system's trueness, while the SD evaluates its repeatability across multiple trials. The results are summarized in Table I.

TABLE I
KINEMATIC VALIDATION: SMARTPHONE IMU (GOOGLE PIXEL 8) VS. COMPUTER VISION

Movement	Reps	MAE	SD
Knee Extension	10	1.52°	$\pm 0.66^\circ$
Knee Flexion	10	2.67°	$\pm 1.34^\circ$

The exploratory data provided an MAE of 1.52° (SD $\pm 0.66^\circ$) for extension and 2.67° (SD $\pm 1.34^\circ$) for flexion. While these figures demonstrate measurement repeatability, supporting the system's potential for longitudinal patient tracking, they must be interpreted with caution. Given the nature of this PoC, claims regarding absolute angle agreement, clinical validity, or the complete removal of inter-operator variability cannot be established. The proposed system primarily serves as a technological foundation designed to inspire future scalable e-health solutions, rather than acting as a definitive clinical assessment tool at this stage.

B. Clinical Analytics and Dashboard Reporting

The web-based dashboard seamlessly retrieves metrics from the Supabase backend, utilizing the Recharts library[6] to render responsive time-series graphs of the ROM, LSI, and Stability Scores (as seen in the dashboard interface example in Fig. 7). The platform successfully aggregates metrics into standardized PDF and Excel exports. These reports combine the objective telemetry with the patient-reported pain scores collected at the end of each session. Furthermore, external applied weights are successfully registered, providing critical context to the physiotherapist regarding the kinetic load associated with the recorded kinematic performance.

[6]https://recharts.org/, accessed July 2026.

C. Limitations and Future Work

As a PoC, the current system exhibits several limitations to be addressed in the near future. First, the evaluation was conducted exclusively on healthy volunteers. As a result, the system assumes a baseline capacity to perform movements with a minimum range of motion. Its performance remains to be evaluated in patients recovering from ACL reconstruction, who often present different movement patterns, pain levels, and compensatory strategies. Second, by relying on a single IMU attached to the tibia, the kinematic model assumes strict movement within the 2D sagittal plane. This design inherently ignores complex 3D joint mechanics, such as the screw-home mechanism, and limits the ability to detect proximal postural compensations, such as pelvic tilting. Third, the system assumes optimal patient compliance to reach true muscular failure during isometric tests. Finally, before deploying the platform in real healthcare settings, the security framework needs to be expanded. While Row-Level Security (RLS) is a good starting point for data isolation, future versions will build on this foundation. Next steps include adding robust authentication, end-to-end encryption, and clear governance rules for data anonymization to meet LGPD and GDPR data retention standards.

To address unconscious postural compensations (e.g., pelvic tilting or trunk leaning) that single-sensor setups cannot inherently detect, future work will focus on integrating a supplementary Computer Vision Pipeline. An image-based approach allows for more precise insights on the patient's posture. With the use of a webcam and pose-detection deep learning models, such as MoveNet Thunder, valuable information on the subject's positioning, such as shoulder and hip alignment, knee height variation and hip-knee-heel angle, can be utilized to verify the execution's correctness. The model receives an image (a frame captured by the webcam) as input and outputs 17 body keypoints (y-coordinate, x-coordinate and confidence score). By processing such points provided by the model, the algorithm can go through a checklist of conditions in order to determine whether the subject is in an appropriate posture:

- The subject must be seated. This may be checked by a geometric heuristic to verify a quasi-horizontal orientation of the femur: the vertical distance between the hip and knee is less than half of their horizontal distance (The thigh segment is lying horizontally). Similarly, the same result may be reached by monitoring the thigh's angle with the x-axis.
- In order to properly assess the subject's posture, the subject must be positioned in a side-view (sagittal) orientation relative to the camera, such that the leg executing the exercise is facing the camera.

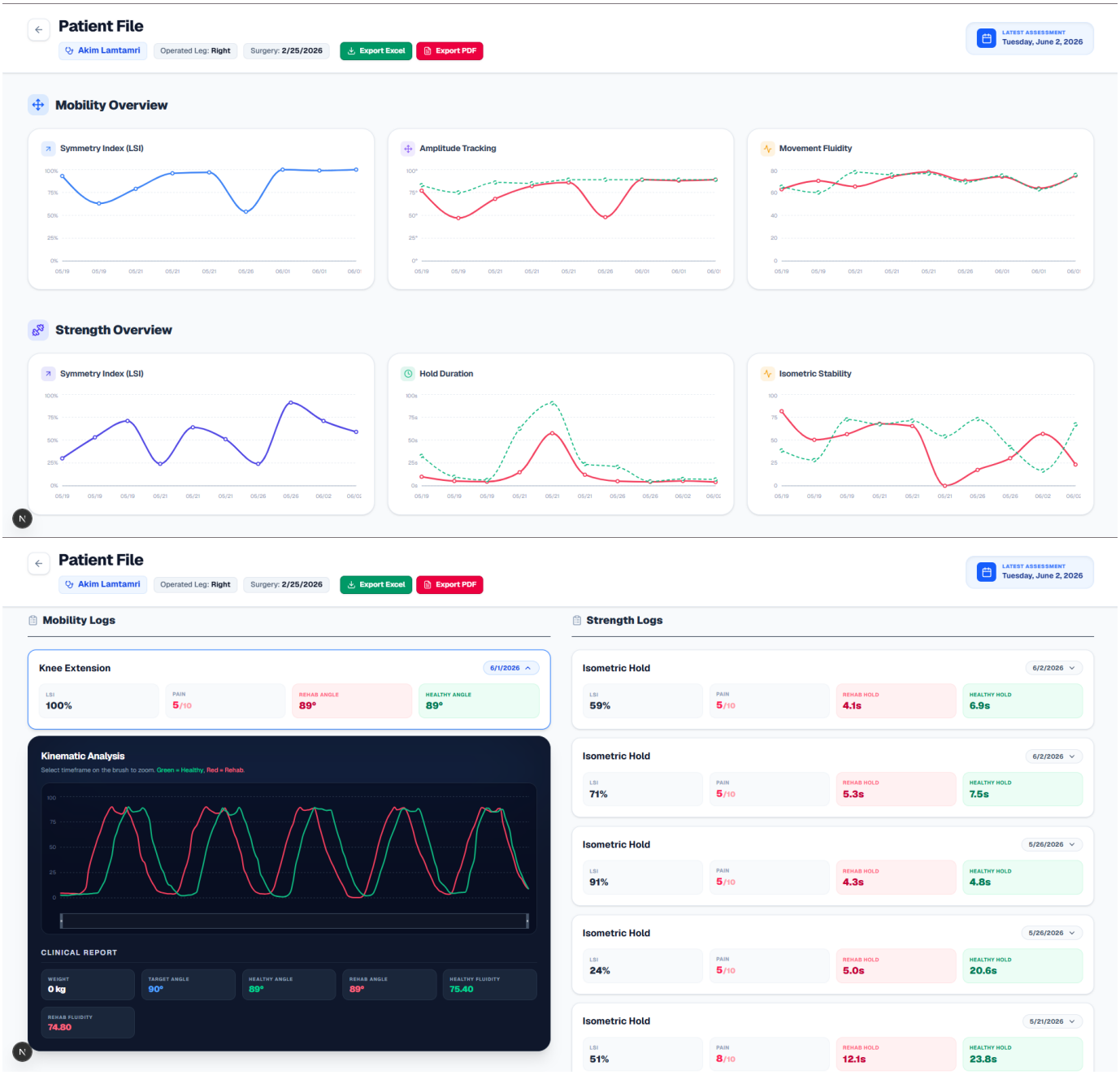


Fig. 7. The clinical web dashboard interface utilizing the Recharts library to render responsive, time-series graphs of the ROM, Limb Symmetry Index (LSI), and Stability Scores for longitudinal patient tracking.

- The subject's leg which is doing the exercise must be facing the camera. Given that the subject is expected to be seated, for a right-facing subject, this condition may be verified by checking if the x-coordinate of the knee is greater than the hip. Similarly, for a left-facing subjects, the x-coordinate of the hip must be greater than the knee.
- In order to account for trunk leaning, the subject must remain with their shoulders and hips aligned. This may be ensured by monitoring the angle between the y-axis and the segment created by the hip and shoulder.
- In order to account for pelvic tilting, the subject's hip's position must not vary greatly during the exercise. This may be ensured by monitoring the distance between the hip's initial position at the beginning of a repetition and the hip's current position.

Finally, the image-based approach must take into consideration frames in which the model was unable to correctly assess the subject's pose. Such errors in the model's predictions may be intensified in low-light or low-resolution images [18]. To enhance reliability against noisy data, frames with

low-confidence keypoints may be filtered out. Moreover, the software must ignore sharp changes in the subject's position. For instance, a variation of 60° in the subject's hip-knee-heel angle in a single frame is unreasonable and should, therefore, not be taken into consideration.

IV. CONCLUSION

This paper presented ACL-Digital Assist, an e-health platform designed to offer a more accessible alternative between expensive biomechanical laboratories and simple manual measurements. By using a centralized database, the system combines a patient mobile application, a dual-device monitor mode, and a web dashboard for clinicians. Laboratory tests on healthy subjects showed that the system can successfully record movements at 20 Hz, handle data using JSON arrays, and calculate recovery metrics like the Limb Symmetry Index and a Stability Score. While relying on a single smartphone sensor for 2D tracking naturally limits the system's precision, the automatic data collection and clinical reports show promising potential for improving physiotherapy workflows. To prepare the system for real clinical use, future work must focus on usability tests with both clinicians and patients. These first tests will help prepare for larger studies to evaluate patient engagement, how reliable the platform is at home, and its actual impact on recovery. Ultimately, by showing that standard smartphone IMUs can be used for this task, this work aims to encourage more accessible research on low-cost patient tracking solutions.

ACKNOWLEDGMENT

The authors thank the research laboratory facilities at the Federal Technological University of Paraná (UTFPR) in Curitiba, Brazil, for providing the software engineering environment necessary for the development of this project.

REFERENCES

[1] S. R. Filbay and H. Grindem, "Evidence-based recommendations for the management of anterior cruciate ligament (ACL) rupture," *Best Practice and Research Clinical Rheumatology*, vol. 33, issue 1, pp. 33–47, 2019.

[2] B. Tamalet and P. Rochcongar, "Épidémiologie et prévention de la rupture du ligament croisé antérieur du genou," *Revue du Rhumatisme Monographies*, vol. 83, no. 2, pp. 103–107, 2016.

[3] Y. Guo, D. Li, Y.-B. Wu, X. Sun, X.-Y. Sun, and Y.-P. Yang, "Mobile health-based home rehabilitation education improving early outcomes after anterior cruciate ligament reconstruction: A randomized controlled clinical trial," *Front. Public Health*, vol. 10, 2023.

[4] Foissey C, Thaunat M, Bondoux L, Sonnery-Cottet B, Fayard JM, "Self-rehabilitation guided by a mobile application after anterior cruciate ligament reconstruction leads to improved early motion and less pain," *Arthrosc. Sports Med. Rehabil.*, vol. 3, no. 5, pp. e1457–e1464, 2021.

[5] P-H Bauwens, J-M Fayard, M. Tatar, H. Abid, B. Freychet, B. Sonnery-Cottet, et al., "Evaluation of a smartphone application for self-rehabilitation after anterior cruciate ligament reconstruction during a COVID-19 lockdown," *Orthopaedics and Traumatology: Surgery and Research*, vol. 109, no. 1, article 103342, 2023.

[6] M. Svensson, V. Lind, and M. Löfgren Harringe, "Measurement of knee joint range of motion with a digital goniometer: A reliability study," *Physiother. Res. Int.*, vol. 24, no. 2, 2019.

[7] S. Kiatkulanusorn, N. Luangpon, "Analysis of the concurrent validity and reliability of five common clinical goniometric devices," *Sci. Rep.*, vol. 13, 20931, 2023.

[8] Chew ES, Woon E-L, Miao-Shi Low JK, Haseler LJ, Ismail I, Alif MB, et al., "Feasibility, reliability and validity of self-measurement of knee range-of-motion using an accelerometer-based smartphone application by patients with total knee arthroplasty," *PLOS One*, vol. 19, no. 10, e0307219, 2024.

[9] T. Jaysrichai, A. Suputtitada, and W. Khovidhungij, "Mobile sensor application for kinematic detection of the knees," *Ann. Rehabil. Med.*, vol. 39, no. 4, pp. 599–608, 2015.

[10] S. Krishnakumar, B.-J. F. van Beijnum, C. T. M. Baten, P. H. Veltink, and J. H. Buurke, "Estimation of kinetics using IMUs to monitor and aid in clinical decision-making during ACL rehabilitation: A systematic review," *Sensors*, vol. 24, no. 7, art. no. 2163, 2024.

[11] Rohman E, Steubs JT, Tompkins M, "Changes in involved and uninvolved limb function during rehabilitation after anterior cruciate ligament reconstruction: implications for Limb Symmetry Index measures," *Am. J. Sports Med.*, vol. 43, no. 6, pp. 1391–1398, 2015.

[12] R. Simonsson, A. Sundberg, R. Piussi, "Questioning the rules of engagement: A critical analysis of the use of limb symmetry index for safe return to sport after anterior cruciate ligament reconstruction," *Br. J. Sports Med.*, vol. 59, no. 6, pp. 376–384, 2025.

[13] L. Labanca, L. Laudani, P. P. Mariani, and A. Macaluso, "Postural adjustments following ACL rupture and reconstruction: A longitudinal study," *Int. J. Sports Med.*, vol. 39, no. 7, pp. 549–554, 2018.

[14] Noehren B, Abraham A, Curry M, Johnson D, Ireland ML, "Evaluation of proximal joint kinematics and muscle strength following ACL reconstruction surgery in female athletes," *J. Orthop. Res.*, vol.32, no. 10, pp. 1305-1310, 2015.

[15] Shi D, Li N, Wang Y, Jiang S, Lin J, Zhu W, "Gait modification strategies in trunk over right stance phase in patients with right anterior cruciate ligament deficiency," *Gait Posture*, vol.46, pp.63-68, 2016.

[16] A-L. Hollingworth, J-B. Hutchinson, J-R. Bechtel, R-M. Chapman, "Knee deep in Error: Unraveling the accuracy limitations of consumer-grade wearable motion sensors with limited degrees of freedom computing knee joint

angles," *Computers in Biology and Medicine*, vol.193, 2025.

[17] K. Rajandaran, J. C. M. Than, J. Xiao, and P. H. H. Then, "Comparative Study of Pose Estimation Techniques for Remote Range of Motion Monitoring in Healthcare," *2026 IEEE 16th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia, 2026, pp. 1-6, doi: 10.1109/ISCAIE68866.2026.11576317

[18] H. Y. Alvarez, "Validity of the MoveNet deep learning model for joint angle estimation," M.S. thesis, San Francisco State University, San Francisco, CA, USA, 2023.

[19] M. MUNERA RAMIREZ, "Analyse vibro-biomécanique et dynamique en sport/santé. Cas du cyclisme," Ph.D. dissertation, 2014.

Aplicação da Manufatura Avançada para a Confecção de Prótese Facial Auricular

1st Maria Eduarda Nunes Badauy
Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. E.I.T.A.)
Univesidade Federal de Goiás (UFG)
Anápolis, Goiás
mariabadauy@discente.ufg.br

2nd Pedro Henrique Gonçalves
Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. E.I.T.A.)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás
pedrogoncalves@ufg.br

Resumo—Este trabalho apresenta o desenvolvimento e a avaliação experimental de um fluxo de fabricação digital para prótese facial auricular utilizando escaneamento tridimensional, modelagem CAD, manufatura aditiva e moldagem em silicone. O estudo teve caráter exploratório e laboratorial, com cinco ensaios consecutivos, cada condição executada uma única vez (n=1), destinados a investigar a influência do material do molde e da formulação do silicone sobre a obtenção do protótipo. Os resultados indicaram incompatibilidade entre o molde produzido em resina fotopolimerizável e o silicone empregado, enquanto o molde em ABS, após acabamento superficial, apresentou comportamento mais favorável à cura. A exploração preliminar de ferramentas baseadas em inteligência artificial indicou limitações na reconstrução da geometria auricular e não foi utilizada como fonte de dados para a fabricação final. Os achados demonstram viabilidade técnica nas condições laboratoriais avaliadas, mas não constituem avaliação experimental clínica.

Palavras-Chave—Manufatura Avançada. Próteses Faciais. Impressão 3D. Silicone. Tecnologia Assistiva.

I. INTRODUÇÃO

Atualmente pacientes com câncer que passaram por mutilações faciais, devido a abscisão cirúrgica de tumores, acidentes ou malformações congênitas que necessitam de próteses faciais como dispositivo de reabilitação funcional e psicossocial, buscam seu tratamento fora da região Goiana se locomovendo para o Sul e Sudeste devido a escassez de mão de obra qualificada em Goiás, onde não se tem documentado locais especializados, públicos ou privados com tal serviço.

Enquanto isso, o Paraná se consolida como referência na área, por meio do ambulatório de Prótese Facial Reconstrutiva do Hospital de Reabilitação, idealizado pela Dra. Roberta Zanicotti. Projeto esse que passou por implementação no SUS (Sistema Único de Saúde) em 2020, operando em Curitiba. Desde sua criação já foram efetuados 1.500 atendimentos de pacientes multilados faciais (nem todos eletivos ao uso do dispositivo), sendo entregues 200 próteses. Dentro do projeto mencionado, cerca de 70% dos pacientes tiveram suas mutilações causadas pelo câncer e 30% por traumas em acidentes.[2]

Esses dados, apresentados em matéria ao ANDIFES (Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior) em janeiro de 2026, demonstram a demanda crescente por esse dispositivo e sua eficácia ao tratamento de reabilitação dos usuários, o que levanta a problemática dessa polarização de mão de obra no Brasil nos dias de hoje. Ademais consultórios particulares cobram cerca de 5.000 a 10.000 reais pelas próteses, o que desde 2020 está sendo entregue de forma gratuita via SUS pelo projeto supracitado, porém com aplicação somente no Sul do país.

Segundo dados da secretaria de estado da saúde de Goiás, estima-se que em 2025 houve cerca de 13.281 novos casos de câncer e em 2026 cerca de 553, desses 553 sendo 105 de pele, uma das tipologias de câncer causadora das mutilações faciais, isto reforça a necessidade de profissionais especializados e qualificados na área de confecção de próteses faciais no estado de Goiás. [9]

As próteses faciais configuram-se como importantes Tecnologias Assistivas (TA) destinadas à reabilitação de pacientes multilados faciais. Essas TAs desempenham um papel funcional e fisiológico quanto à oclusão de cavidades expostas (Figura 1). Para além da reabilitação funcional do paciente, esses dispositivos são ferramentas psicossociais, as quais ajudam na reintegração social do usuário, devolvendo sua autoestima e minimizando o sentimento de rejeição.

Possuindo assim, uma carga emocional e identitária no papel da reinserção em suas atividades diárias comunitárias, como o trabalho. Dentro desse contexto, nota-se a importância da reintegração social do paciente por meio das próteses faciais atuando ativamente em seu estado psicológico e em sua qualidade de vida.[12]



Fig. 1—Exemplos de próteses faciais utilizadas na reabilitação de pacientes com mutilações faciais.
Fonte: adaptado de [4], [10].

Contudo a reabilitação do paciente não se dá somente pelo dispositivo e seu uso cotidiano, para a recapacitação mais eficaz do paciente é necessário uma equipe transdisciplinar que cuide dos aspectos psicológicos e dos processos de desenvolvimento da prótese, podendo incluir profissionais de diversas áreas a depender do caso [8]. Os profissionais de design entram nessa cadeia interdisciplinar no que tange o desenvolvimento das próteses visando um modelo de produção com maior efetividade e menor gasto de energia.

O processo convencional de fabricação envolve moldagem física da região afetada, obtenção de modelos em gesso, escultura manual, confecção do molde, processamento do silicone, acabamento e adaptação. Essas etapas dependem de habilidade técnica e podem demandar sucessivas intervenções manuais [7], [12].

Entre as limitações do método convencional estão o desconforto durante a moldagem, a dependência da habilidade do profissional, o desperdício de materiais e a dificuldade de reproduzir digitalmente modificações realizadas durante o processo. Por isso, fluxos baseados em escaneamento e manufatura aditiva vêm sendo investigados como alternativas para etapas da fabricação [3], [5], [7].

A manufatura avançada permite integrar aquisição tridimensional, modelagem computacional e impressão 3D, criando possibilidades de armazenamento digital, ajuste do modelo antes da fabricação e produção de moldes personalizados. Neste contexto, este trabalho investigou experimentalmente um fluxo digital para fabricação de prótese facial auricular, com foco na aquisição da geometria, modelagem do molde, manufatura aditiva e compatibilidade entre o material do molde e o silicone utilizado. O estudo também explora, de forma preliminar, ferramentas de inteligência artificial para reconstrução tridimensional, sem utilizá-las como evidência clínica.

O objetivo é desenvolver e avaliar experimentalmente um fluxo de fabricação digital, identificar limitações técnicas e validar sua eficácia. O estudo não pretende validar o uso clínico da prótese nem avaliar conforto, retenção, durabilidade ou satisfação de usuários.

II. METODOLOGIA

Este estudo consiste em um desenvolvimento tecnológico com avaliação experimental de um fluxo digital para prótese facial auricular. O trabalho foi conduzido em ambiente laboratorial e não incluiu instalação de prótese em pacientes e avaliação clínica.

O desenvolvimento foi orientado por princípios do Design Centrado no Usuário (UCD), especialmente quanto à redução da invasividade da moldagem e à consideração de requisitos de personalização. Entretanto, não houve avaliação formal com usuários finais; portanto, o estudo não deve ser interpretado como avaliação experimental de UCD.



Fig. 2—Fluxo comparativo entre o método convencional e o fluxo digital investigado.
Fonte: imagem gerada por IA, utilizada apenas como representação esquemática do fluxo; não constitui evidência clínica.

O fluxo compreendeu: (1) aquisição da geometria por escaneamento 3D; (2) tratamento e reconstrução da malha; (3) modelagem CAD do molde negativo; (4) fabricação do molde por manufatura aditiva; (5) preparação do silicone; e (6) moldagem e avaliação experimental.

A. Aquisição de dados (escaneamento 3D)

A primeira etapa consistirá no escaneamento tridimensional da orelha de uma paciente hipotética, empregando o Revopoint Scanner 3D Range 2, com o propósito de avaliar a produção de uma prótese auricular. Por integrar a classificação das próteses faciais, esse dispositivo foi selecionado por apresentar menor complexidade de confecção, permitindo a validação inicial do método e dos materiais empregados em sua fabricação.

Cabe ressaltar que a paciente hipotética selecionada não apresentará mutilações faciais. Além da justificativa anteriormente apresentada para a escolha da prótese auricular, a seleção desse modelo fundamenta-se no fato de que pacientes oncológicos frequentemente apresentam perda unilateral da orelha. Dessa maneira, em uma aplicação futura com um caso real, seria possível realizar a digitalização da orelha saudável e utilizar seu modelo tridimensional como referência para o espelhamento no lado afetado, dando prosseguimento às etapas subsequentes.

B. Modelagem Computacional (CAD)

Após a aquisição dos dados tridimensionais, a malha resultante será processada no Blender 5.1, visando à eliminação de ruídos e à correção de eventuais falhas geométricas. Fotografias serão empregadas exclusivamente como referências visuais para a reconstrução de detalhes anatômicos.

Com a superfície digital devidamente tratada, o arquivo será exportado no formato STL e posteriormente encaminhado ao Fusion 360, no qual será convertido em corpo sólido. Nessa etapa, serão executadas operações booleanas, estabelecida a linha de partição e desenvolvida a geometria do molde negativo. O Meshmixer poderá ser empregado, quando necessário, para intervenções complementares na superfície digital.

C. Manufatura Aditiva (Impressão 3D)

Para a preparação do arquivo correspondente ao molde negativo destinado à impressão tridimensional, será selecionado um dos softwares de fatiamento disponíveis, LycheeSlicer ou o software de fatiamento da Creality, de acordo com a tecnologia de fabricação empregada.

A versão inicial será confeccionada em resina utilizando a impressora Creality LD-002R, com o material KEDAGBR/F1 Branco. Para a fabricação dos moldes pela tecnologia FDM, será empregado filamento ABS. O pós-processamento desse material exigirá acabamento superficial com acetona. Considerando que esse procedimento pode modificar a textura da superfície, o componente será submetido à inspeção visual antes da etapa de moldagem.

D. Pigmentação e aplicação do silicone ao molde

A moldagem será conduzida sob diferentes condições experimentais envolvendo o silicone, contemplando combinações com e sem catalisador e com ou sem pigmentação intrínseca, conforme cada ensaio. A incorporação do pigmento ocorrerá previamente à introdução do silicone no molde.

Inicialmente, o material pigmentado será distribuído sobre a superfície interna do molde, que em seguida, será fechado e o restante de silicone será introduzido por meio de uma seringa, possibilitando a formação da peça.

Cada uma das cinco condições experimentais será executada uma única vez (n=1). O propósito dessa etapa é identificar possíveis incompatibilidades entre os materiais e orientar o refinamento iterativo do fluxo de fabricação, não sendo objetivo realizar inferência estatística. Serão avaliados os seguintes aspectos: ocorrência de cura completa ou incompleta, presença de vazamentos, facilidade de desmoldagem, preservação da integridade geométrica, qualidade da superfície e necessidade de ajustes no modelo CAD.

III.RESULTADOS

O desenvolvimento ocorreu de maneira iterativa, de modo que os resultados obtidos em cada ensaio serviram como referência para a definição da condição subsequente. Considerando que cada condição experimental foi realizada uma única vez (n=1), os achados apresentados possuem caráter exclusivamente descritivo.

Durante a aquisição tridimensional com o Revopoint Range 2, foram identificadas falhas nas regiões posteriores e nas áreas de maior reentrância da orelha. A digitalização demandou múltiplas tentativas, seguida de tratamento e correção da malha obtida. Também foram testados marcadores fotorrefletores como recurso auxiliar à captura; entretanto, essa estratégia não proporcionou melhoria suficiente na qualidade dos dados para justificar sua incorporação ao fluxo metodológico definitivo. Figura 3

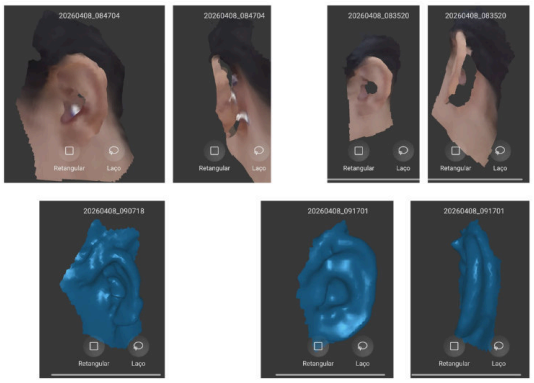


Fig. 3—Modelos 3D obtidos durante a aquisição da anatomia auricular.
Fonte: Acervo dos autores.

Como uma tentativa de mitigar tais problemas enfrentados na obtenção da malha gerada por escaneamento 3D, ferramentas baseadas em inteligência artificial foram avaliadas apenas de forma exploratória. Algumas produziram distorções da geometria auricular; outras geraram modelos visualmente plausíveis, mas apresentaram restrições de exportação. Por essas limitações, os modelos de IA não foram utilizados para a fabricação final nem considerados evidência de precisão anatômica.

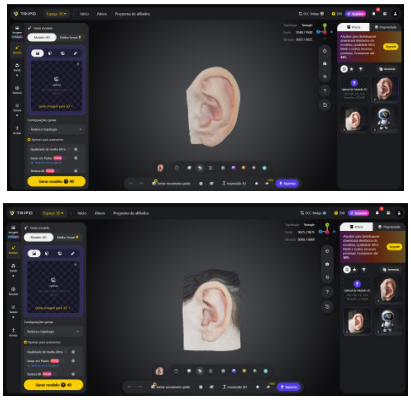


Fig. 4—Exemplo de modelo tridimensional gerado por ferramenta de IA.
Fonte: Acervo dos autores; ferramenta de IA utilizada apenas para exploração metodológica.

Devido às limitações encontradas no uso da inteligência artificial quanto à exportação de um modelo com qualidade adequada, optou-se pela utilização do arquivo obtido por escaneamento. Como complemento aos dados tridimensionais, foi registrada uma fotografia da orelha da paciente hipotética com um smartphone Redmi Note 13, a qual foi empregada como referência visual durante a etapa de modelagem. A partir desse referencial, iniciou-se o tratamento, a modelagem e o refinamento do arquivo STL selecionado no Blender 5.1. Figura 5

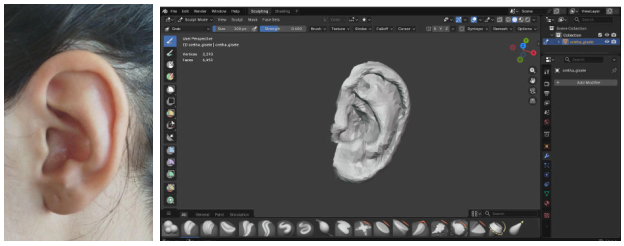
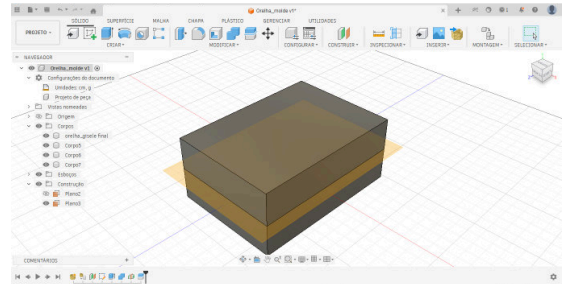


Fig. 5—Referência fotográfica e modelo digital utilizado na reconstrução anatômica.
Fonte: Acervo dos autores.

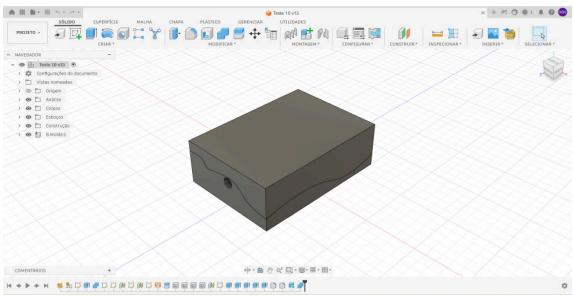
A malha foi exportada do Blender 5.1 no formato STL e, posteriormente, importada para um novo projeto no Fusion 360. Inicialmente, empregou-se a ferramenta de edição direta de malha para realizar os ajustes necessários e transformar a estrutura, inicialmente caracterizada apenas como uma "casca", em um volume preenchido. Na sequência, o modelo digital foi convertido em um corpo sólido.

Para a elaboração do molde, foi criado um esboço retangular, posteriormente submetido à extrusão de modo a envolver completamente a geometria da orelha. Nessa operação, a extrusão foi configurada como "Novo Corpo", evitando sua união indesejada à peça original. Em seguida, utilizou-se o comando Combinar, estabelecendo o retângulo como corpo de destino e a orelha como ferramenta. A operação selecionada foi Corte, mantendo-se habilitada a opção "Manter Ferramentas".

Na etapa seguinte, foi criado um plano de construção destinado à delimitação da linha de partição do molde. A partir desse plano, aplicou-se o comando Divisão de Corpo, obtendo-se duas metades distintas do conjunto (Figura 6-A). Por fim, foram incorporados os elementos necessários ao funcionamento do molde, incluindo os canais para injeção do silicone e os sistemas de encaixe e travamento entre as partes (Figura 6-B).



.A



.B

Fig. 6- Print tirado da área de trabalho do software Fusion 360 da modelagem final VF com acabamento do molde em negativo
Fonte: Acervo do autor

Após o tratamento da malha, o modelo foi refinado no Blender 5.1 e convertido em sólido no Fusion 360. Foram identificadas dificuldades na operação booleana quando havia descontinuidades na malha, exigindo correções adicionais. A linha de partição do molde foi modificada de um corte plano (Figura 6-A) para uma geometria acompanhando a anatomia, melhorando o encaixe entre as partes. Figura 6-B

Em seguida, para o primeiro teste de impressão do molde foi selecionada a impressão em Resina devido ao seu acabamento, porém cabe ressaltar que o molde de resina interfere diretamente na cura do silicone que precisa de catalisador, acredita-se que seja devido às propriedades da resina, impedindo a saída da umidade de dentro do molde, impedindo assim a cura completa do silicone. Figura 7

Para a impressão em Resina será usado a impressora Creality LD-002R, usando os seguintes parâmetros para resina no software de fatiamento:

- marca 3DLAB;
- nome Standard White;
- espessura 50um (oficial).

Porém a resina utilizada para o molde será a da marca KED, a KEDAGBR/F1 BRANCO.

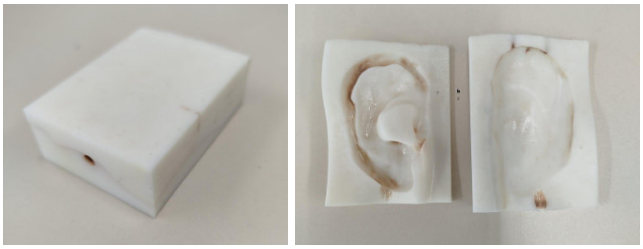


Fig. 7—Molde negativo produzido em resina fotopolimerizável.
Fonte: Acervo dos autores.

No primeiro ensaio, o molde foi impresso em resina fotopolimerizável. Observou-se cura incompleta do silicone, com superfície viscosa e úmida em regiões de contato com o molde. O resultado inviabilizou a obtenção da peça e motivou a substituição do material do molde. Figura 8

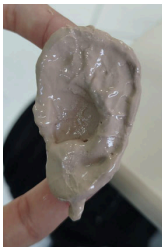


Fig. 8—Resultado do ensaio 1.
Fonte: Acervo dos autores.

Conforme mencionado anteriormente, a resina interfere no processo de catalização do silicone, comprometendo o resultado. Diante da inviabilidade técnica observada com o molde de resina, confeccionou-se um novo componente utilizando o filamento ABS (Acrilonitrila Butadieno Estireno), com o propósito de viabilizar a cura adequada do silicone. Figura 9

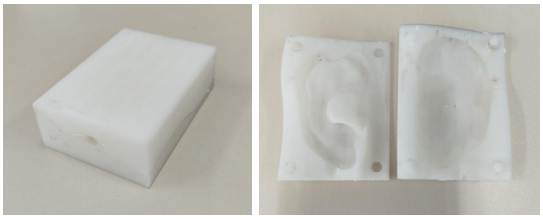


Fig. 9—Molde negativo produzido em ABS.
Fonte: Acervo dos autores.

No segundo ensaio, foi utilizado molde em ABS. O comportamento do silicone foi melhor, mas a peça apresentou cura apenas parcial e textura pegajosa. Esse resultado indicou que a condição química do silicone e/ou do sistema de catalisação também influenciava a polimerização. Figura 10

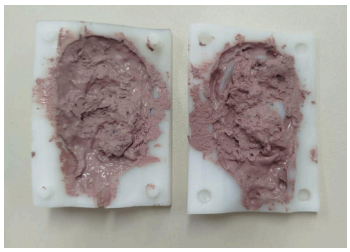


Fig. 10—Resultado do ensaio 2.
Fonte: Acervo dos autores.

Após as intercorrências registradas nos dois ensaios anteriores, conduziu-se um terceiro teste utilizando o mesmo molde em filamento ABS, porém alterando as variáveis do processo: o silicone foi aplicado sem a adição de pigmentos e utilizou-se uma marca distinta, cuja formulação não necessita de catalisador. Este procedimento obteve êxito no que diz respeito à cura do material (Figura 11). Todavia, identificou-se a necessidade de retornar à etapa de modelagem digital para reajustar os furos de injeção do silicone. Essa revisão foi necessária uma vez que ocorreu o escoamento indesejado do material por essa abertura logo após o preenchimento do molde.



Fig. 11—Resultado do ensaio 3.
Fonte: Acervo dos autores.

Após a implementação dos ajustes e a avaliação experimental do terceiro ensaio, realizou-se um novo teste incorporando a pigmentação ao silicone sob a condição da que havia apresentado melhor cura. Foi obtida a peça com o resultado mais satisfatório dentre os experimentos conduzidos anteriormente apresentando cura homogênea e reprodução geral da geometria do molde, porém ainda possui necessidade de ajustes no que tange a mitigação de bolhas que geram falhas na prótese. Figura 12



Fig. 12—Resultado do ensaio 4.
Fonte: Acervo dos autores.

Por fim, realizou-se um ensaio adicional (o quinto ensaio) no presente trabalho, visando ao ajuste cromático e à eliminação de bolhas de ar (Figura 13). Contudo, alguns ajustes serão necessários, como: controle bolhas de ar para evitar vazios estruturais e garantir a integridade da peça; aprimoramento da mistura para a reprodução fidedigna do tom de pele do usuário; intensificação do acabamento com acetona no molde impresso, visando reduzir a rugosidade superficial e, consequentemente, minimizar a transferência das linhas de impressão para a textura da peça final; e inclusão de uma camada envoltória à estrutura auricular projetada, com o intuito de proporcionar uma interface de encaixe e fixação mais precisa à face do paciente.



Fig. 13—Resultado do ensaio 5.
Fonte: Acervo dos autores.

Os cinco ensaios são sintetizados na Tabela 1. A comparação evidencia que a substituição do molde de resina por ABS e a alteração da formulação do silicone foram as principais modificações associadas à melhora do resultado experimental.

Ensaio	**Material do molde**	**Condição do silicone**	**Resultado observado**	**Réplica / conduta**
Teste 1	Resina fotopolimerizável (SLA)	Silicone + catalisador + pigmento	Cura incompleta; superfície viscosa	n=1; substituir molde
Teste 2	ABS (FDM)	Silicone + catalisador + pigmento	Cura parcial; textura pegajosa	n=1; alterar condição química

Teste 3	ABS (FDM)	Silicone sem catalisador e sem pigmento	Cura satisfatória; vazamentos	n=1; ajustar CAD/vedação
Teste 4	ABS (FDM)	Silicone pigmentado	Melhor resultado; cura homogênea	n=1; refinar acabamento
Teste 5	ABS (FDM)	Silicone pigmentado; ajuste cromático	Persistência de bolhas e limitações de cor/textura	n=1; requer novos testes

TABELA 1—Síntese dos ensaios experimentais.
Fonte: Dados do estudo.

Em conjunto, os resultados indicam viabilidade técnica preliminar do fluxo em ambiente laboratorial, especialmente quanto à utilização do ABS como material de molde nas condições avaliadas.

IV. DISCUSSÃO

Os resultados indicam que a principal contribuição do estudo é a identificação experimental da influência do material do molde sobre a cura do silicone. A incompatibilidade observada no molde de resina fotopolimerizável evidencia que a escolha do material do molde deve considerar não apenas acabamento e resolução de impressão, mas também sua compatibilidade química com o elastômero.

A cura incompleta observada no primeiro ensaio pode estar relacionada à presença de resíduos de fotoiniciadores ou monômeros não reagidos na superfície de peças produzidas por fotopolimerização, especialmente quando o pós-processamento não elimina completamente esses componentes. O resultado, contudo, não permite determinar o mecanismo químico específico, pois não foram realizados ensaios laboratoriais de caracterização de resíduos.

Os moldes em ABS apresentaram comportamento mais favorável nas condições avaliadas. O acabamento com acetona reduziu irregularidades superficiais, mas esse procedimento também altera a superfície do polímero e exige controle para evitar deformações.

As dificuldades de escaneamento demonstram que a qualidade da malha depende da geometria anatômica, da aquisição e do processamento posterior. A exploração de IA não permitiu substituir o escaneamento, pois as reconstruções não foram submetidas a análise quantitativa de desvio. Assim, os resultados de IA devem ser interpretados somente como observação preliminar.

Sob a perspectiva da Engenharia Biomédica, o estudo oferece uma contribuição prática ao documentar um problema de compatibilidade entre materiais e uma alternativa de molde que apresentou melhor comportamento nas condições testadas. Estudos futuros devem quantificar desvios geométricos entre o modelo original e a peça, realizar avaliação profissional cega, comparar custos e tempos de fabricação e investigar biocompatibilidade, estabilidade de cor, retenção, resistência e durabilidade antes de qualquer uso clínico.

Embora o fluxo proposto integre as etapas de escaneamento tridimensional, modelagem CAD e manufatura aditiva, este estudo não teve como objetivo estabelecer uma comparação quantitativa com o método convencional. A proposta concentrou-se na validação da viabilidade do método desenvolvido como alternativa para a produção de próteses faciais, buscando verificar sua aplicabilidade e identificar aspectos que possam ser aprimorados em etapas futuras.

Outra limitação é o número de réplicas. Cada condição foi executada uma única vez (n=1), o que impede conclusões sobre variabilidade do processo. Os resultados devem, portanto, ser considerados exploratórios e úteis para direcionar experimentos subsequentes.

Este estudo desenvolveu e avaliou experimentalmente, em ambiente laboratorial, um fluxo digital para a fabricação de uma prótese facial auricular. Os resultados permitiram identificar limitações no escaneamento e na modelagem e, principalmente, uma diferença de comportamento entre moldes em resina fotopolimerizável e ABS durante a moldagem em silicone.

Nas condições investigadas, o molde em resina apresentou cura incompleta do silicone, enquanto o ABS apresentou comportamento mais favorável após acabamento superficial. A melhor condição experimental foi obtida após ajustes na formulação do silicone e incorporação de pigmentação, embora tenham permanecido limitações de bolhas, cor e acabamento superficial.

Os resultados não permitem afirmar que o fluxo esteja validado ou pronto para aplicação clínica, pois cada condição foi testada uma única vez (n=1) e não foram realizadas avaliações quantitativas de precisão, repetibilidade, custo, tempo, propriedades mecânicas, biocompatibilidade, retenção, durabilidade ou satisfação do usuário.

Somente após essas etapas deverá ser considerada a validação clínica do fluxo.

AGRADECIMENTOS

Os autores agradecem pelo apoio no desenvolvimento deste trabalho à Universidade Federal de Goiás (UFG)— Faculdade de Artes Visuais (FAV) e ao Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. E.I.T.A.).

REFERÊNCIAS

[1] L. D. B. Alves, A. C. S. Menezes, C. S. Boasquevisque, and Stahel-Lage, "Próteses bucomaxilofaciais na reabilitação estético-funcional de pacientes oncológicos," Rev. Nav.

Odontol., vol. 49, no. 1, pp. 27–35, 2022. [Online]. Available: https://portaldeperiodicos.marinha.mil.br/index.php/odontoclinica/article/view/2946/2871

[2] ANDIFES, "Paraná tem prótese facial pelo SUS que simula até as manchas da pele," Jan. 5, 2026. [Online]. Available: https://www.andifes.org.br/2026/01/05/parana-tem-protese-facial-pelo-sus-que-simula-ate-as-manchas-da-pele/

[3] B. O. Artoli, M. C. O. Carvalho, and M. E. Kunkel, "Manufatura Digital de Próteses Auriculares: Uma Revisão de Escopo," Rev. Saúde Coletiva, no. 101, vol. 16, 2025. [Online]. Available: https://revistasaudecoletiva.com.br/index.php/saudecoletiva/article/view/3637/4967

[4] A. C. Buco, "Fotografia de Adriana Corsetti Buco." [Online]. Available: https://adrianacorsettibuco.com.br/quem-sou/

[5] B. N. L. Cabral, D. P. Costa, R. G. Grempel, and A. Paiva Neto, "Protocolo de aquisição de imagens 3D por smartphone para criação de um banco virtual de modelos nasais utilizados na reabilitação estético-funcional de pacientes mutilados," Rev. Saúde Digital Tecnol. Educ., vol. 8, no. 1, pp. 21–39, Mar. 2023. [Online]. Available: http://periodicos.ufc.br/resdite/article/index

[6] L. O. Dias, V. M. S. Ribeiro, and M. F. Haddad, "Pigmentação em Prótese Maxilofacial – Revisão de Literatura," Arch. Health Invest., vol. 12, no. 1, pp. 126–133, 2023, doi:10.21270/archi.v12i1.5529.

[7] F. M. Franco, "Uso de modelagem 3D e manufatura aditiva no auxílio à fabricação de próteses faciais de silicone," Bachelor's thesis, Universidade Tecnológica Federal do Paraná, Curitiba, Brazil, 2021. [Online]. Available: https://repositorio.utfpr.edu.br/jspui/bitstream/1/28728/1/fabricacaoprotesefacial.pdf

[8] D. M. Galavoti and V. H. A. Ferreira, "Reabilitação estético-funcional com as próteses faciais," Rev. Cient. Unilago, vol. 1, no. 2, 2025. [Online]. Available: https://revistas.unilago.edu.br/index.php/revista-cientifica/article/view/1164

[9] Secretaria de Estado da Saúde de Goiás, "Painel de Oncologia," Goiânia, GO, Brazil, 2024. [Online]. Available: https://indicadores.saude.go.gov.br/public/oncologia.html

[10] Horb Odontologia, "Procedimento odontológico/Equipamentos." [Online]. Available: https://horbodontologia.com.br/odonto8.php

[11] F. O. Medola and L. C. Paschoarelli, Tecnologia Assistiva: Desenvolvimento e Aplicação, 1st ed. Bauru, SP, Brazil: Canal6 Editora, 2018.

[12] SALAZAR-GAMARRA, Rodrigo *et al*. Introdução à metodologia "Mais Identidade": próteses faciais 3D com a utilização de tecnologias acessíveis para pacientes sobreviventes de câncer no rosto. *In*: SANTOS, Emanuela Carla dos (org.). Comunicação Científica e Técnica em Odontologia 2. Ponta Grossa, PR: Atena Editora, 2019. cap. 20, p. 251-272. DOI: https://doi.org/10.22533/at.ed.26519290313. Available: https://www.researchgate.net/publication/332212528_INTRODUCAO_A_METODOLOGIA_MAIS_IDENTIDADE_PROTESES_FACIAIS_3D_COM_A_UTILIZACAO_DE_TECNOLOGIAS_ACESSIVEIS_PARA_PACIENTES_SOBREVIVENTES_DE_CANCER_NO_ROSTO.

[13] A. J. Severino, Metodologia do Trabalho Científico, 1st ed. São Paulo, SP, Brazil: Cortez Editora, 2013.

[14] P. L. C. Silva, "Aplicação da impressão 3D e reconstrução digital no desenvolvimento de próteses maxilofaciais," M.S. thesis, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, 2019. [Online]. Available: https://repositorio.ufmg.br/items/baa6f096-eff5-499a-b1fa-baff9b81470d/full

[15] Universidade de São Paulo, "Banco de dados criado na FORP armazena análises faciais do brasileiro." [Online]. Available: https://www5.usp.br/noticias/saude-2/banco-de-dados-criado-na-forp-armazena-analises-faciais-do-brasileiro/

An HD sEMG-Based Virtual Reality Box and Blocks Test for Hand Dexterity Assessment

Mário Sérgio Maduro Santana
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
m223615@dac.unicamp.br

Ricardo Gonçalves Molinari
Center for Biomedical Engineering
University of Campinas
Campinas, Brazil
molinari@unicamp.br

Natanael Pereira Batista
Center for Biomedical Engineering
University of Campinas
Campinas, Brazil
natanael@unicamp.br

Leonardo Abdala Elias
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
leoelias@unicamp.br

***Abstract*—The relationship between manual performance in myoelectrically controlled virtual reality (VR) environments and equivalent real-world tasks is currently poorly understood. The present study investigated the correlation between scores on a conventional physical Box and Blocks Test (BBT) and on the same test performed using a VR setup. In the VR setup, the BBT was implemented as a virtual reality environment, in which virtual hand opening and closing were controlled through a proportional control strategy based on high-density surface electromyography (HD sEMG), while hand position was tracked using the Meta Quest controller. Ten healthy participants completed the BBT under conventional physical and VR setups. Spearman's rank correlation revealed a significant positive association between performance under the two experimental conditions ($\rho_{spearman} = 0.80$; 95% confidence interval (CI): $[0.27, 0.97]$; $p - value = 0.003$). The mean BBT scores under the conventional physical and VR setups were 52.7 (standard deviation (SD) = 5.17) and 9.8 (SD = 3.49) blocks, respectively. The difference between the mean scores of the two conditions was 42.9 blocks (SD = 3.67; 95% CI: [40.8, 45.1]; $p - value = 0.001$). These results indicate that the proposed VR setup may be able to differentiate between varying levels of manual dexterity. Nonetheless, larger sample sizes are necessary to establish a more precise estimate of the association and to further validate the proposed VR setup as a viable alternative for hand dexterity assessment.**

***Index Terms*—Virtual Reality, Box and Blocks Test, High-Density Surface Electromyography, Myoelectric Control, Manual Dexterity.**

I. Introduction

It remains unclear to what extent virtual reality (VR) environments controlled by myoelectric signals can be used as an alternative to physical hand dexterity tests. Differences between physical and virtual environments, such as discrepancies between simulated and real-world physics [1], potential discomfort associated with VR headset use [2], and difficulties with initial contact with myoelectric control [3], [4], may directly influence the execution and performance of the proposed tasks.

The Box and Blocks Test (BBT) is widely used to assess gross manual dexterity in upper-limb motor rehabilitation [5], [6]. More recently, virtual implementations of the BBT have been explored due to their potential to support remote and home-based assessment, provide automatically computed quantitative performance measures, and be incorporated into virtual rehabilitation systems [7]. Everard et al. [7] investigated the relationship between the physical BBT and a VR version without myoelectric control, in which participants manipulated virtual blocks using VR controllers through button presses. The authors reported strong positive correlations between the two environments in both healthy participants and individuals with stroke. Furthermore, Hashim et al. [8] investigated the relationship between performance in a virtual myoelectrically controlled version of the BBT and its physical counterpart in healthy participants, reporting a weak positive correlation between the two conditions. However, the methodology of the latter study does not clearly describe how the correlation analysis was performed. Since each participant completed multiple repetitions of the task, it remains unclear whether the reported correlation was computed from independent observations or from repeated measurements within the same individuals. In addition, Hashim et al. employed a conventional dual-site proportional myoelectric control strategy based on antagonistic forearm muscle activation, rather than an HD sEMG-based proportional motor-intention estimation approach.

Investigating the relationship between performance on the conventional physical Box and Blocks Test (BBT) and on a virtual reality implementation using proportional myoelectric control based on high-density surface electromyography (HD sEMG) still represents a gap in the literature. If a strong

Ricardo Molinari was supported by the Brazilian Public Ministry of Labor under Contract No. 002118.2019. Leonardo A. Elias is a CNPq Research Productivity Fellow under Process No. 316320/2023-4. Mário Santana was supported by FINEP under Agreement No. 01.24.0110.00 (Ref. 2140/22). Natanael Batista was supported by the Brazilian Public Ministry of Labor under Contract No. 39433-26.

correlation is observed between scores across the respective environments, such an association would suggest that the proposed VR interface can consistently reflect differences in manual dexterity, even if its scores do not directly reproduce those obtained in the conventional physical BBT. Such a relationship could support the use of the proposed interface as an alternative approach for assessing hand dexterity in clinical populations.

Therefore, the present study aimed to investigate the correlation between the performance of healthy participants on the conventional physical BBT and on a VR implementation of the BBT using proportional myoelectric control based on HD sEMG [9].

II. Methodology

The experimental protocol adopted in the present study is part of a broader research project involving experiments with myoelectric prostheses. However, the prosthesis-related data are outside the scope of the present study and will be analyzed in future work.

The two experimental conditions analyzed in the present study are referred to as the *conventional physical setup*, corresponding to the standard administration of the BBT, and the *VR setup*, in which participants performed a VR implementation of the BBT using a virtual hand whose opening and closing were controlled by proportional myoelectric control based on HD sEMG, while hand position was tracked using the Meta Quest controller. In addition to these two setups, the experimental protocol included a third setup, referred to as the *myoelectric prosthetic setup*, in which the conventional physical BBT was performed using the myoelectric prosthetic system. As stated above, data obtained under the myoelectric prosthetic setup are outside the scope of the present study and will not be included in the analyses.

A. Participants

Ten healthy right-handed participants (age: 31.2 ± 10.2 years) were recruited for the experiment. Inclusion criteria were as follows: absence of upper-limb musculoskeletal injuries and no history of neurological, musculoskeletal, autoimmune, or metabolic diseases. All participants signed an informed consent form prior to participating in the study. The experimental protocol was approved by the Research Ethics Committee of the University of Campinas (UNICAMP), Brazil (CAAE: 90477225.9.0000.5404).

B. Hardware Setup

The developed interface was executed on a Lenovo Legion Pro 5i (Lenovo, Beijing, China) computer equipped with an Intel Core i9-14900HX processor, 32 GB of RAM, and an NVIDIA RTX 4070 GPU with 8 GB of dedicated memory, together with a Meta Quest 3 (Meta, CA, USA) mixed reality headset. HD sEMG signals were acquired at a sampling rate of 2052.52 Hz using a wearable system developed in our laboratory [9].

C. HD sEMG Electrode Placement

Before experiments, the participant's forearm skin was prepared with a light abrasion and alcohol cleaning to reduce skin-electrode impedance. Two 8×8 electrode grids with a 10-mm inter-electrode distance (GR10MM0808, Bioelettronica, Torino, Italy) were positioned in the proximal forearm over the finger flexor and extensor muscles according to the recommendations of Zipp [10]. Specifically, the proximal border of the electrode array placed over the flexor muscles was positioned approximately at 25% of the distance from the medial epicondyle to the anterior wrist crease, measured from the medial epicondyle, and adjusted by palpation to align with the muscle belly. Likewise, the proximal border of the electrode array placed over the extensor muscles was positioned approximately at 25% of the distance between the lateral epicondyle and the ulnar styloid process, measured from the lateral epicondyle, and similarly adjusted by palpation to align with the extensor muscle belly.

D. Myoelectric Control Strategy

To enable proportional myoelectric control, the HD sEMG signals acquired in real time were processed to continuously estimate the participant's movement intention. The control algorithm is fully described in [9], following a synergy-based method [11]. Briefly, root mean square (RMS) features extracted from the HD sEMG recordings were decomposed using non-negative matrix factorization (NMF) into components associated with opposite finger movement directions (flexion and extension). A proportional movement-intention signal was then obtained from these components. The resulting estimate was used to generate proportional control commands to VR hand or a prosthetic device.

Before performing the experimental tasks, the model was calibrated individually for each participant. For the purposes of the present study, participants interacted with the blocks using a three-finger pinch. Therefore, only this gesture was used during model calibration and subsequent task execution. The calibration procedure consisted of the following steps:

1) Initially, the participant performed successive opening and closing movements of a three-finger pinch in synchrony with the movements of a prosthetic or VR hand for 25s while recording HD sEMG signals.
2) Next, the acquired signals were used to estimate the parameters of the myoelectric control model. The model was evaluated based on two criteria: the maximum cross-correlation between the estimated movement-intention signal and the corresponding reference signal, and a brief practical control test performed by the participant. In general, maximum cross-correlation values above 0.5 were considered indicative of satisfactory control performance.
3) If both the correlation metric and the practical control test yielded satisfactory results, the experimental protocol continued. Otherwise, the previous steps were repeated until a suitable model was obtained. If satisfactory performance could not be achieved after multiple

attempts, the experiment was conducted using the best-performing model obtained up to that point.

In the present study, the calibration procedure required a mean of 3.9 attempts per participant ($SD = 2.42$).

E. Box and Blocks Test

The BBT is a functional test to assess the gross manual dexterity [6]. The test consists of a box divided into two compartments by a central barrier, containing wooden blocks on only one side. The participant must transfer, one block at a time, as many blocks as possible from one compartment to the other within 60 seconds. Performance was quantified by the total number of blocks correctly transferred during this time interval. Throughout this paper, the terms *performance*, *score*, and *number of blocks transferred* are used interchangeably to refer to the participant's result on the BBT.

In the present study, the BBT test was performed under three experimental conditions (see Figure 1): the conventional physical setup, the myoelectric prosthetic setup, and the VR setup. In all conditions, the participant remained seated in front of the box positioned on a table.

In the conventional physical and myoeletric prosthetic setups,the BBT box measured approximately 53 cm $\times$ 25.5 cm $\times$ 8.5 cm and was divided into two compartments by a 15.3-cm-high partition. The blocks were cubic, measuring 2.5 cm $\times$ 2.5 cm $\times$ 2.5 cm. These dimensions are close to those reported by Mathiowetz et al. [6]. In the VR setup, the entire BBT environment, including the box, partition, and virtual blocks, was proportionally scaled according to the corresponding real-world dimensions.

F. Virtual Reality Environment

The VR environment developed consists of a replica of the laboratory room where the experiments were conducted. The interface was developed in Unity (Unity Technologies, San Francisco, CA, USA) and designed to support two users simultaneously: the participant immersed in the VR environment using the Meta Quest 3 headset, and the experimenter responsible for monitoring and configuring the test via an external computer. Figure 2 illustrates the developed VR version of the BBT.

During the experiment, the experimenter was able to view the environment as seen by the participant in real time, as well as to make necessary adjustments to the interface and experimental parameters. These adjustments included positioning the participant within the virtual scene and adjusting their height.

G. Experimental Protocol

The experiment was conducted in a single session. Before the experiment, the participant was screened for eligibility and demographics were collected. After these initial steps, the HD sEMG electrode grids were positioned on the participant as described in Subsection II-C. Once the grids were secured and connected to the HD sEMG acquisition system, the experiment started.

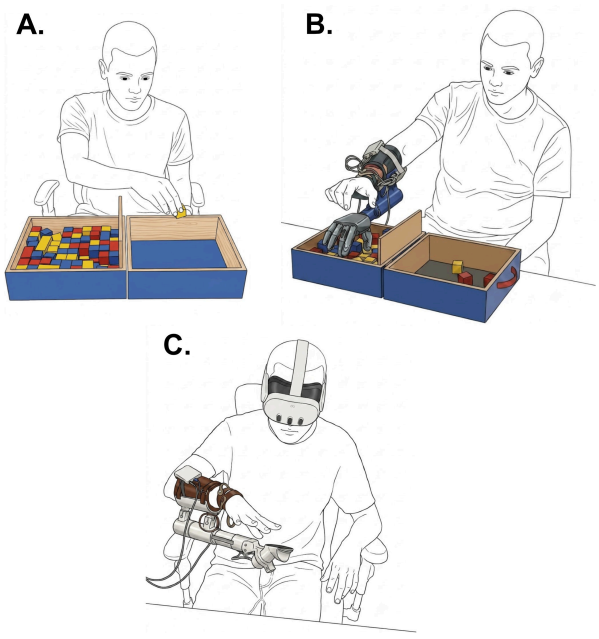


Fig. 1. Experimental conditions of the Box and Blocks Test: (A) conventional physical setup, (B) myoelectric prosthetic setup, used for model calibration and not included in the analyses presented in this study, and (C) VR setup, consisting of a virtual reality implementation of the BBT, in which virtual hand opening and closing were driven by proportional myoelectric control based on high-density surface electromyography (HD sEMG), while hand position was tracked using the Meta Quest controller.

First, the participant performed the BBT under the conventional physical setup. Before the test, the participant received instructions regarding the test procedure and completed a brief familiarization task by transferring five blocks from one compartment of the box to the other. After the familiarization phase, the full test was administered once, and the participant's score was recorded.

After completing the conventional physical BBT, the participant performed the BBT under the myoelectric prosthetic setup. As previously noted, the data obtained under the prosthetic condition are not included in the analyses presented in the present work and will be reported in future studies. For this stage, a socket that attaches the prosthesis to the participant's arm was used. In addition, as described in Section II-D, the myoelectric control model was calibrated using the prosthetic hand as the reference hand. Once again, the participant received instructions regarding the test procedure and completed a shortened version of the task, transferring five blocks between the box compartments to become familiar with both the prosthetic system and the task dynamics. After this adaptation phase, the complete test was performed once, and the score was recorded.

Finally, the participant performed the BBT under the VR setup. For this purpose, the same myoelectric control model previously calibrated for the prosthetic condition was used, along with a socket similar to that used in that condition, which secured the Meta Quest 3 controller to the participant's arm. Once again, the participant received instructions regarding

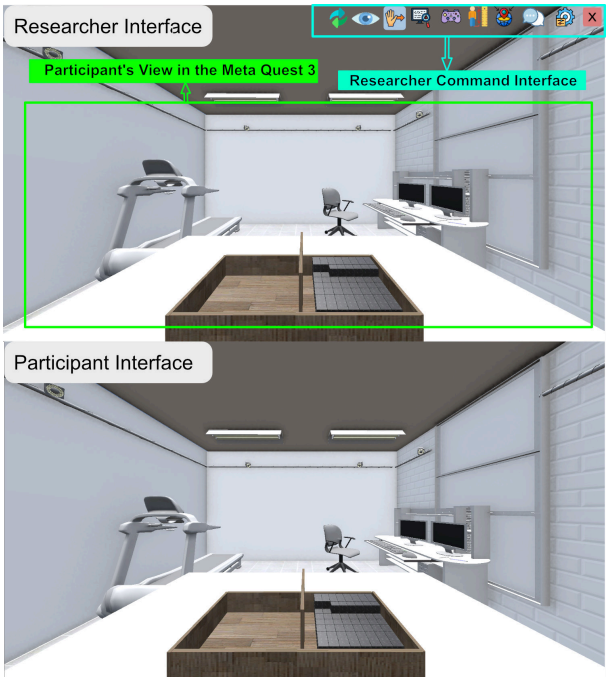


Fig. 2. Interfaces of the VR setup developed in Unity. The upper portion displays the researcher interface, containing controls for scene configuration, while the lower portion displays the participant interface used to interact with the virtual environment.

the test procedure and completed an adaptation phase by transferring five blocks between the box compartments in the VR environment. Subsequently, the complete VR BBT was performed once, and their performance was recorded.

In cases where the model calibrated for the prosthetic condition did not provide satisfactory control performance in the VR setup, an additional calibration procedure was performed following the protocol described in Section II-D, using the virtual hand as the reference. The latter condition occurred only for one participant, who was unable to successfully perform the task using the prosthetic system. After recalibration, satisfactory control performance was achieved in the VR environment.

It is worth noting that no BBT trials performed under the conventional physical and VR setups were excluded from the analyses.

H. Data Analysis

1) Correlation Analysis: The performance of each participant was represented by an ordered pair corresponding to the results obtained under the VR and physical setups, respectively. The relationship between VR and physical performance was evaluated using Spearman's correlation coefficient ($\rho_{spearman}$). For this analysis, the correlation coefficient, the p-value associated with Spearman's test, and the corresponding 95% confidence interval were estimated, with the latter obtained using the percentile bootstrap method [12] with $20,000$ paired resamples. The number of resamples was selected after preliminary analyses indicated that further increases produced only negligible changes in the estimated confidence intervals. Since the study hypothesis assumed that participants with better performance in VR would tend to exhibit better performance with their physical hand, a one-tailed test was used to evaluate the presence of a positive correlation between the conditions ($H_0 : \rho_{spearman} \leq 0$ and $H_1 : \rho_{spearman} > 0$). A statistical significance level of 5% was adopted.

2) Difference Between Mean Scores of the Environments: The difference between the mean scores obtained under the conventional physical and VR setups was evaluated. The p-value of this difference was obtained using an exact paired permutation test based on all possible sign permutations of the paired differences [13], whereas the 95% confidence interval was estimated using the percentile bootstrap method [12] with 20,000 resamples. Similarly to the correlation analysis, preliminary analyses indicated that increasing the number of resamples beyond 20,000 did not change the estimated confidence intervals. Since the study hypothesis assumed that performance in the conventional physical setup is superior to that observed in VR, a one-tailed test was adopted for the comparison between means ($H_0 : \mu_{Physical} - \mu_{VR} \leq 0$ and $H_1 : \mu_{Physical} - \mu_{VR} > 0$). A significance level of 5% was adopted.

III. Results

A positive Spearman's correlation was observed between the performance under the VR and conventional physical setups ($\rho_{spearman} = 0.80$; 95% confidence interval (CI): $[0.27,\ 0.97]$; $p - value = 0.003$). The relationship between the scores obtained under the two conditions is presented in Figure 3.

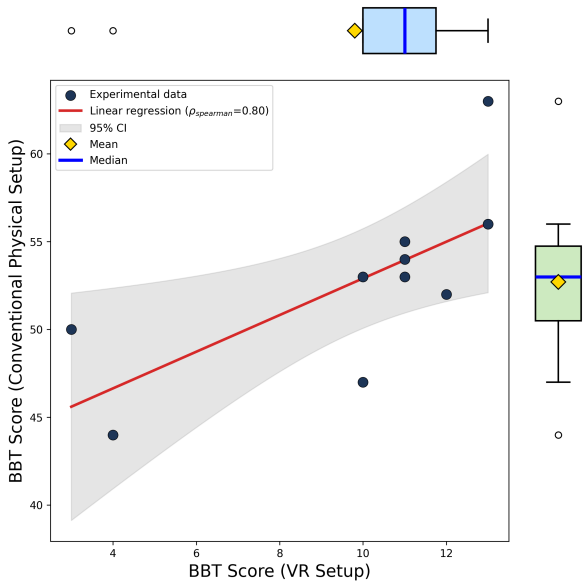


Fig. 3. Relationship between the BBT scores obtained under the VR and conventional physical setups. The solid line represents the linear fit to the data, and the shaded region corresponds to the 95% confidence interval of the fitted line.

In addition to analyzing the relationship between performances obtained under the two experimental conditions, the difference between the mean BBT scores obtained under the conventional physical and VR setups was also evaluated. The mean scores obtained under the conventional physical and VR setups were 52.7 (standard deviation (SD) = 5.17) and 9.8 (SD = 3.49) blocks, respectively. The difference between the mean scores of the two conditions was 42.9 blocks (SD = 3.67; 95% CI: [40.8, 45.1]), and this difference was statistically significant ($p-value = 0.001$). The distribution of the scores under each condition is shown in Figure 4.

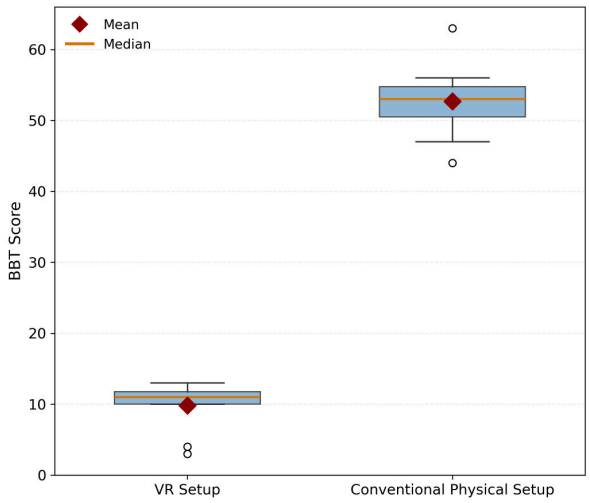


Fig. 4. Distribution of BBT scores under the conventional physical and VR setups. Diamonds represent the sample means, and horizontal lines represent the median values.

IV. Discussion

The results of the present study showed a positive correlation between BBT performance under the conventional physical and VR setups, suggesting that participants who perform better under one condition tend to perform better under the other. However, the wide confidence interval for the correlation coefficient suggests that the magnitude of this association should still be interpreted with caution, given the small sample size ($N = 10$). The observed positive association between performance under the conventional physical and VR setups is consistent with previous findings in the literature. Everard et al. [7] also reported positive correlations between performance on the conventional BBT and its VR counterpart, even in the absence of myoelectric control. Similarly, Hashim et al. [8] observed a positive association between performance on the original test and its VR version with myoelectric control. However, the latter study did not employ an HD sEMG-based motor-intention estimation approach. Also, in [8], the description of the procedure used to calculate the correlation coefficient was not sufficiently detailed to allow a clear interpretation of the independence of the data points included in the analysis. Therefore, evidence regarding the association between performance under conventional physical conditions and VR tasks controlled through myoelectric interfaces remains limited in the literature, and the present study provides additional evidence to address this gap.

Regarding the BBT scores, participants transferred significantly fewer blocks under the VR setup than under the conventional physical setup. One possible explanation for this difference is that, unlike the conventional physical setup, the VR setup required participants to control the virtual hand through an HD sEMG-based myoelectric interface, simultaneously performing the functional task and managing a control strategy with which they had little familiarity. This interpretation is supported by Tabor et al. [3], who demonstrated that myoelectric control learning is a gradual and complex process, with consistent improvements occurring across successive training sessions and no evidence of an early performance plateau. Their findings suggest that performance measured during the initial stages of myoelectric training may underestimate an individual's true capability after sufficient practice. Therefore, the participants' limited experience with myoelectric control likely contributed to the lower scores observed under the VR setup compared with the conventional physical setup.

Additionally, it should be considered that participants have extensive experience using their hands in activities of daily living, making superior performance under the conventional physical setup expected when compared with performance achieved using a newly learned myoelectric control strategy. In agreement with this perspective, Daud et al. [14] state that activities of daily living, such as writing, tying shoelaces, and buttoning a shirt, contribute to the development of fine motor skills and manual dexterity. Therefore, the participants' extensive prior experience performing these tasks with their physical hands may have contributed to the superior performance observed under the conventional physical setup.

Factors inherent to VR environments may also have contributed to the observed reduction in performance. Such factors include the absence of natural tactile feedback [7], differences in spatial perception [15], and discomfort associated with VR headset use [2]. Evidence supporting this interpretation is provided by Everard et al. [7], who also observed lower performance in a virtual version of the BBT that did not employ myoelectric control. Therefore, because reduced performance was observed in two interfaces using different control strategies, the present results suggest that this phenomenon is not solely due to the limitations of HD sEMG-based myoelectric control but also reflects, at least in part, challenges inherent to interaction within VR environments.

Taken together, the results suggest that the proposed VR setup has the potential to effectively differentiate between levels of manual dexterity, despite not perfectly mimicking performance under the conventional physical setup. Although a significant difference in mean scores was observed between the conventional physical and VR setups, performance across both modalities was positively correlated. This indicates that, despite the systematic difference in absolute scores between the two setups, participants who performed better under the conventional physical setup also tended to perform better

under the VR setup. Therefore, the utility of the VR setup for assessing manual dexterity does not necessarily depend on reproducing the absolute scores obtained under the conventional physical setup, but rather on its ability to distinguish different levels of manual dexterity.

However, the findings of the present study should be interpreted in light of some additional limitations. First, the different experimental setups were performed in a fixed order (conventional, prosthetic, and VR setups, respectively), which may have introduced learning and fatigue effects during the tests in subsequent setups. In addition, as previously discussed, the relatively small sample size ($N = 10$) limits the precision of the estimates obtained, as evidenced by the wide 95% confidence intervals. The present study also did not investigate the intra-subject repeatability or test-retest reliability of the proposed VR setup, nor did it assess potential learning and adaptation to its HD-sEMG-based myoelectric control strategy.

Future studies should address these limitations by using larger sample sizes and randomized or counterbalanced experimental orders to minimize potential effects associated with learning and fatigue and to obtain more precise estimates with narrower confidence intervals. Furthermore, protocols involving multiple sessions and repeated assessments should be adopted to investigate the intra-subject repeatability and test-retest reliability of the proposed VR setup, as well as potential learning and adaptation effects associated with repeated use of the HD-sEMG-based myoelectric control strategy. Beyond these considerations, future work could explore alternative control strategies that increase the number of degrees of freedom of the virtual hand, enabling more complex and versatile hand movements.

V. Conclusion

The present study examined the correlation between performance on the BBT under the conventional physical and VR setups. The results showed a positive correlation between performance under the two experimental conditions, suggesting that participants who performed better under the conventional physical setup also tended to perform better under the VR setup, despite differences in absolute scores between the two setups. These findings contribute to the understanding of manual dexterity assessment in virtual reality using proportional HD sEMG-based control and suggest that the proposed VR setup may have potential for future use in assessment and rehabilitation training applications. Future studies with larger sample sizes are necessary to confirm the observed association and to further validate the proposed VR setup.

Acknowledgment

The authors would like to thank Mr. Flavio Renato Santos and Mr. Nelio Hilton Gava from the Main Design Lab, Center for Biomedical Engineering (CEB), UNICAMP, for their technical support in manufacturing the sockets used in this study and in fabricating and maintaining the prosthetic system.

References

[1] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 737–744.

[2] E. Chang, H. T. Kim, and B. Yoo, "Virtual reality sickness: A review of causes and measurements," *International Journal of Human–Computer Interaction*, vol. 36, no. 17, pp. 1658–1682, 2020. [Online]. Available: https://doi.org/10.1080/10447318.2020.1778351

[3] A. Tabor, S. Bateman, and E. Scheme, "Evaluation of myoelectric control learning using multi-session game-based training," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, pp. 1680–1689, 9 2018.

[4] C. Igual, L. A. Pardo, J. M. Hahne, and J. Igual, "Myoelectric control for upper limb prostheses," 11 2019.

[5] L. Santisteban, M. Térémetz, J. P. Bleton, J. C. Baron, M. A. Maier, and P. G. Lindberg, "Upper limb outcome measures used in stroke rehabilitation studies: A systematic literature review," *PLoS ONE*, vol. 11, 5 2016.

[6] V. Mathiowetz, G. Volland, N. Kashman, and K. Weber, "Adult norms for the box and block test of manual dexterity," *American Journal of Occupational Therapy*, vol. 39, no. 6, pp. 386–391, 1985.

[7] G. Everard, Y. Otmane-Tolba, Z. Rosselli, T. Pellissier, K. Ajana, S. Dehem, E. Auvinet, M. G. Edwards, J. Lebleu, and T. Lejeune, "Concurrent validity of an immersive virtual reality version of the box and block test to assess manual dexterity among patients with stroke," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, 12 2022.

[8] N. A. Hashim, N. A. A. Razak, and N. A. A. Osman, "Comparison of conventional and virtual reality box and blocks tests in upper limb amputees: A case-control study," *IEEE Access*, vol. 9, pp. 76 983–76 990, 2021.

[9] R. G. Molinari, V. Avilés-Carrillo, G. A. G. de Villa, and L. A. Elias, "A wearable platform for real-time control of a prosthetic hand by high-density emg," *IEEE Access*, vol. 14, pp. 36 851–36 868, 2026.

[10] P. Zipp, "Recommendations for the standardization of lead positions in surface electromyography*," Tech. Rep., 1982.

[11] N. Jiang, K. B. Englehart, and P. A. Parker, "Extracting simultaneous and proportional neural control information for multiple-dof prostheses from the surface electromyographic signal," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1070–1080, 2009.

[12] P. Bruce, A. Bruce, and P. Gedeck, *Estatística Prática para Cientistas de Dados: +50 Conceitos Essenciais Usando R e Python*, 2nd ed. Rio de Janeiro, Brazil: Alta Books, 2025.

[13] P. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer, 2005.

[14] A. Z. Che Daud, M. K. Yau, F. Barnett, J. Judd, R. E. Jones, and R. F. Muhammad Nawawi, "Integration of occupation based intervention in hand injury rehabilitation: A randomized controlled trial," *Journal of Hand Therapy*, vol. 29, no. 1, pp. 30–40, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089411301500174X

[15] D. J. Harris, G. Buckingham, M. R. Wilson, and S. J. Vine, "Virtually the same? how impaired sensory information in virtual reality may disrupt vision for action," pp. 2761–2766, 11 2019.

An Integrated Platform for Alternating Peripheral Electrical Nerve Stimulation with Application to Parkinson's Disease Tremor

Danilo Andrés Molina-Vidal
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0003-4420-4585

Pedro Sebastian Silverio Cevallos
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0009-0002-5712-4699

Evelyn Juliette Pesántez-Vera
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0009-0009-0019-6713

Danielle Martins da Silva Liquori
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0009-0002-1804-3490

Carlos Julio Tierra-Criollo
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0003-4880-9186

***Abstract*—Noninvasive peripheral electrical stimulation has been investigated as an alternative approach for managing tremor in Parkinson's disease, particularly due to the limitations of prolonged pharmacological treatment and the invasiveness of surgical treatments. The aim of this work was to develop and implement an integrated platform for alternating peripheral electrical nerve stimulation with real-time tremor monitoring, enabling the evaluation of different waveforms and individualized stimulation parameters. The system consisted of a dual-channel electrical stimulator, a hand-worn wireless triaxial accelerometry module, and graphical user interfaces for stimulation control, signal acquisition, visualization, and analysis. A protocol was implemented in the user interface to apply patterned bursts of either biphasic pulse or sinusoidal electrical stimulation. The median and radial nerves were stimulated alternately at the patient's tremor frequency. A pilot study was conducted with ten patients with Parkinson's disease (PD), assigned to either the biphasic pulse stimulation group (n = 5) or the sinusoidal stimulation group (n = 5). The stimulator interface allowed configuration of waveform type, amplitude, frequency, and burst frequency. The monitoring interface enabled real-time acquisition and analysis of acceleration signals. In the pilot study, both biphasic pulse and sinusoidal stimuli produced a median tremor reduction of -29.66% and -17.25%, respectively. The integrated platform was shown to be flexible and functional for alternating peripheral electrical nerve stimulation and real-time tremor monitoring. The system supports the investigation of different protocols and waveforms, although it currently operates in an open-loop configuration; future work should integrate stimulation control with the accelerometry system to implement closed-loop strategies.**

***Index Terms*—Electrical stimulation, neurostimulation, Parkinson's Disease, TAPS.**

I. Introduction

Noninvasive electrical stimulation represents a well-established therapeutic modality in clinical settings, encompasses diverse techniques and anatomical targets, including transcranial, neuromuscular, and peripheral nerve stimulation (PNS) [1]. In particular, PNS is utilized across a wide spectrum of disorders, including: neurological and movement disorders (attenuation of pathological tremors in Parkinson's Disease and Essential Tremor, spasticity management, and facilitation of neural plasticity following stroke) [2], [3]; acute and chronic pain conditions (modulation of nociceptive and neuropathic pain through spinal gate mechanisms and endocrine opioid release) [4], [5]; autonomic and circulatory dysfunctions (modulation of peripheral blood flow, acceleration of tissue healing, and regulation of sympathovagal tone via transcutaneous vagus nerve stimulation) [6], [7].

Parkinson's disease (PD) is a progressive neurodegenerative disorder and one of the most prevalent worldwide, affecting approximately 6.1 million people [8]. PD is characterized by the loss of dopaminergic neurons in the substantia nigra [9], [10], causing motor symptoms such as tremor [11], bradykinesia [12], and gait disturbances [13]. Tremor is one of the most prevalent manifestations in patients with PD, significantly impacting daily activities and quality of life. Pharmacological treatment with levodopa and other dopaminergic agents is a common approach for managing motor symptoms in DP. However, the prolonged use of these medications can lead to relevant side effects, such as motor fluctuations, dyskinesias, and other complications that may limit their long-term effectiveness [14], [15].

Non-pharmacological alternatives based on electrical stimulation, such as Deep Brain Stimulation (DBS), have been explored as complementary methods for tremor management. DBS applies electrical pulses to specific subcortical nuclei for tremor reduction. Despite its clinical efficacy, DBS is an invasive intervention associated with surgical risks [16]. Consequently, non-invasive neuromodulation techniques have

emerged as promising alternatives for tremor management.

Among the non-invasive alternatives, peripheral electrical stimulation such as Functional Electrical Stimulation (FES) and Transcutaneous Electrical Nerve Stimulation (TENS) has been investigated. FES applies electrical pulses to evoke muscle contractions, generating compensatory forces that counteract tremor-related oscillations [17]. In contrast, TENS delivers sub-motor-threshold stimulation (electrical pulses) to peripheral afferent nerves, primarily activating thick sensory fibers [18], and is typically applied over the radial and/or median nerves at the wrist [19]. Although the underlying mechanism through which TENS attenuates tremor remains elusive, it is widely hypothesized that the stimulation of peripheral sensory fibers modulates pathological oscillations within the central nervous system [20]. In patients with PD, these oscillations may be associated with abnormally synchronized neural activity. Abnormally synchronized neuronal firing in structures such as the basal ganglia and thalamus may occur at a frequency close to that of the clinically observed tremor [21]. Computational studies of coordinated reset (CR) neuromodulation [21], [22] showed that sequential phase-resetting stimuli, delivered with coordinated delays corresponding to fractions of the characteristic oscillation period (e.g., a fraction of the Parkinsonian tremor period), can separate a pathologically synchronized neuronal population into phase-shifted subgroups. This phase-shifted can be implemented through direct electrical stimulation of neurons' somata or indirectly through synaptically mediated stimulation, such as non-invasive sensory stimulation [22].

Studies have evaluated the wrist-worn Cala ONE stimulator device (Cala Health, Inc., San Mateo, CA, EUA) in patients with Essential Tremor (ET) [23] and more recently in DP [24], reporting promising reductions of tremor. The device uses Transcutaneous Afferent Patterned Stimulation (TAPS), a neuromodulation technique that delivers patterned bursts of electrical biphasic pulse stimulation alternately to the median and radial nerves at the patient's tremor frequency (e.g., 5 Hz). Stimulation parameters for tremor reduction, including intensity, frequency, duration, and waveform, remain insufficiently defined in the literature [17].

Pulsatile electrical stimulation is widely utilized to artificially elicit tactile responses, recruit intermediate nerve fibers ($A\beta$) [25], [26]. However, elevating the stimulation intensity subsequently recruits small-diameter fibers ($A\delta$ and C) [27]. These limitations suggest that more selective engagement of somatosensory pathways could potentially yield superior therapeutic effects on motor symptoms. Among the alternative stimulation waveforms, sinusoidal electrical stimulation has been shown to provide more selective activation of sensory nerve fibers [28], [29]. Within this context, we proposed to develop an integrated platform to investigate Transcutaneous Somatosensory Sine-Wave Electrical Stimulation (tSSES) —a modality employing continuous sinusoidal waveforms to target specific afferent fiber populations [28]–[30] -recently emerged as a promising alternative for tremor management in Parkinson's disease [31].

II. Materials and Methods

The proposed system is composed of three main components: A) a dual-channel electrical stimulator; B) a hand-worn wireless accelerometry module; and C) a user interface for control and signal monitoring. The general structure is shown in Fig. 1, and each subsystem is described in this section.

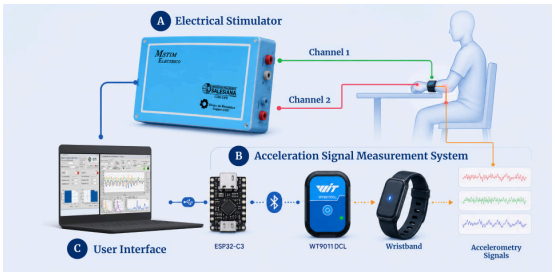


Fig. 1. General system structure. A) Dual-channel electrical stimulator to arbitrary waveform; B) Wireless accelerometry module; and C) User interface for stimulation control and signal monitoring.

A. Electrical Stimulation

The stimulation module is based on the MESTIM-ELECT stimulator [32], powered by a 12-V battery supply. Each channel includes a constant-current source capable of delivering up to 10 mA. The stimulator provides two fully independent channels, each capable of generating arbitrary waveforms (sinusoidal, square, or custom).

Stimulation parameters -including waveform, amplitude, and channel activation- are sent through a LabVIEW interface and transmitted to the stimulator via Ethernet communication. This architecture enables the configuration of diverse protocols for the user.

B. Accelerometer System

The tremor monitoring subsystem is based on the BWT901CL-BT50 module (WitMotion Shenzhen Co., Ltd.), which integrates a 9-axis IMU (accelerometer, gyroscope, and magnetometer) with onboard Kalman filtering for real-time motion estimation. The module supports Bluetooth Low Energy (BLE) communication and sampling rates up to 200 Hz; for this study, the sampling frequency was set to 50 Hz.

A XIAO-ESP32-C3 microcontroller (Seeed Studio) was used as the BLE receiver. It establishes a wireless link with the IMU and forwards the decoded motion data to the computer via serial communication. This setup enables continuous, low-latency acquisition of tremor signals during stimulation.

C. User Interface

The user interface originally developed for the MESTIM-ELECT stimulator [32] was expanded to support novel stimulation paradigms. Currently, a stimulation protocol based on TAPS, designed for tremor suppression in Parkinson's disease, was incorporated. In this protocol, each channel alternately delivers electrical stimulus to two different peripheral nerves. The alternation rate is determined by a user-configurable burst frequency (F_{b}) parameter, allowing the system to be

adapted to different experimental protocols as well as to the individual needs of each patient. Actually, two stimulation options are available: a charge-balanced biphasic square wave and a sinusoidal waveform.

The biphasic square waveform consisted of a 150 Hz waveform with a positive phase (300 μs), an interphase interval of 50 μs, and a negative phase (300 μs), following stimulation parameters commonly reported in studies involving peripheral nerve stimulation for ET and PD [23], [33]. The novel protocol utilizing a sinusoidal signal was proposed. For both waveform types, the stimulation amplitude and the frequency parameters can be configured by the user through the user interface. The user can also define a burst frequency (F_{b}) that controls the temporal alternation between the two independent output channels of the stimulator. Each channel remains active for a period equal to $1/(2 \times F_{\mathrm{b}})$, while the other channel remains inactive.

D. Graphical User Interface Customized for monitoring tremors in Parkinson's disease

A custom Graphical User Interface (GUI) was developed to monitor, record, and analyze real-time acceleration signals from Parkinson's disease patients in both time and frequency domains. The interface establishes a serial communication protocol with the microcontroller ESP32-C3 to receive multi-axis kinematic data, which is simultaneously logged into structured text (.txt) files for offline analysis.

The GUI processes the incoming data utilizing a sliding-window architecture, where both the window length and the sampling frequency are fully configurable. To isolate pathological tremors from voluntary movements and constant gravitational acceleration components, the raw signals are filtered using a 4th-order Butterworth filter. This filtering stage allows the user to select between low-pass, high-pass, band-pass, or band-stop topologies, with a predefined default band-pass range from 3 to 12 Hz. The GIU performs concurrent spectral and temporal analyses on the segmented data blocks.

Power Spectral Density (PSD) Estimation: Spectral analysis (PSD) is conducted by applying a Hanning window to each signal segment. To ensure statistical stability and smooth transient noise, a running exponential average of five consecutive PSD windows was implemented. This averaging method gives greater weight to recent PSD estimates than to older ones, enabling rapid adaptation to recent tremor variations. To provide an omnidirectional representation of tremor power independently of sensor orientation, the individual PSDs from the orthogonal axes (X, Y, and Z) are combined to construct a joint Random Vector PSD [34]. Within the computed PSDs, two primary metrics are extracted: first, a peak-detection algorithm identifies the dominant frequency; second, the ratio between the power of the dominant spectral peak and the total power within the filtered frequency band (3–12 Hz) is calculated. This ratio provides a quantitative criterion for detecting the presence of tremor [35], [36].

Envelope Extraction: Temporal amplitude variations are tracked via the analytical signal obtained through the Hilbert Transform, computed over the same operational sliding windows. The signal envelope is defined as the magnitude of this analytical signal, which is subsequently averaged to a single representative value per window. These windowed values are utilized to compute a cumulative acceleration magnitude vector, which is plotted in real-time against the time domain to monitor changes in movement amplitude.

E. Pilot Experimental Protocol

A pilot study was conducted with ten male patients diagnosed with Parkinson's disease and presenting clinically evident tremor. The patients were between 58 and 75 years of age (66 $\pm$ 5.8 years) and were classified as stages 1 to 3 according to the Hoehn and Yahr scale. Patients with Parkinson's disease-related dementia, diabetes, a history of upper-limb injuries or implanted biomedical devices, such as cardiac pacemakers or Deep Brain Stimulation (DBS) systems, were excluded. The protocol was approved by the Research Ethics Committee (CAAE 58744222.2.1001.5257), and all patients provided written informed consent.

The study was performed in a temperature-controlled room (24 $\pm$ 1 °C) at the Neurology Service of the Hospital Universitário Clementino Fraga Filho (HUCFF), Federal University of Rio de Janeiro (UFRJ). Patients remained seated with their forearms supported on a table, and the upper limb most affected by tremor was selected. After the skin was cleaned, bipolar gel electrodes were positioned over the median and radial nerves, and a triaxial accelerometer was also attached to the dorsum of the same hand.

Subsequently, a postural tremor task based on the Movement Disorders Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS Part III) was performed. During the task, participants maintained the upper limb in the required posture for 20 s while acceleration signals were recorded. The dominant tremor frequency was identified as the frequency component with the highest spectral amplitude during the postural task.

Patients were randomly assigned to either the biphasic stimulation group (n = 5) or the sinusoidal stimulation group (n = 5). The stimulation waveform frequency was set to 150 Hz for biphasic stimulation and 3 kHz for sinusoidal stimulation. Before the intervention, the sensory threshold (ST) was determined separately for the median and radial nerves according to the procedure described in [29]. For each nerve, the stimulation amplitude was gradually increased from the respective ST until the highest comfortable and well-tolerated amplitude was reached, and lower sub-motor threshold.

Electrical stimulation was applied for 40 min, with stimulation alternating (burst frequency = initial postural frequency) between the median and radial nerves according to the previously determined dominant tremor frequency. Immediately after the stimulation, the postural task was performed again following the same procedure used in the pre-stimulation assessment.

F. Processing

Accelerometer signals were saved for offline analysis and processed in MATLAB (MathWorks, Inc., Natick, USA). A 4th-order Butterworth high-pass filter at 0.5 Hz was first applied to remove low-frequency drift. Subsequently, the Power Spectral Density (PSD) was estimated using Welch's method.

The relative percentage reduction was calculated with respect to the pre-stimulation period, and results are presented as medians with the quartile range [Q1–Q3], corresponding to the 25th (Q1) and 75th (Q3) percentiles. For full experimental pilot study and analysis details, see the companion manuscript entitled "Effects of Peripheral Electrical Stimulation Waveforms on Tremor in Parkinson's Disease: A Pilot Study" submitted to the XXX Brazilian Congress of Biomedical Engineering (CBEB 2026) [37].

III. Results

A. Electrical Stimulator User Interface

Fig. 2 presents the user interface implemented for setting the stimulation protocol. The interface includes the protocol stimulation (section 1), the configuration of the burst frequency (F_b) to control the alternation between the two channels (section 2), and the independent adjustment of waveform type, frequency, and amplitude for each channel. It also displays the waveforms generated by both channels (section 3). Channel 1 is active during the first half of the burst period, whereas Channel 2 is activated during the second half, producing an alternating stimulation pattern. Fig. 2.A shows the charge-balanced biphasic waveform configured at 150 Hz, while Fig. 2.B shows the sinusoidal waveform configured at 3 kHz. The enlarged waveform segments illustrate the morphology of each stimulation signal.

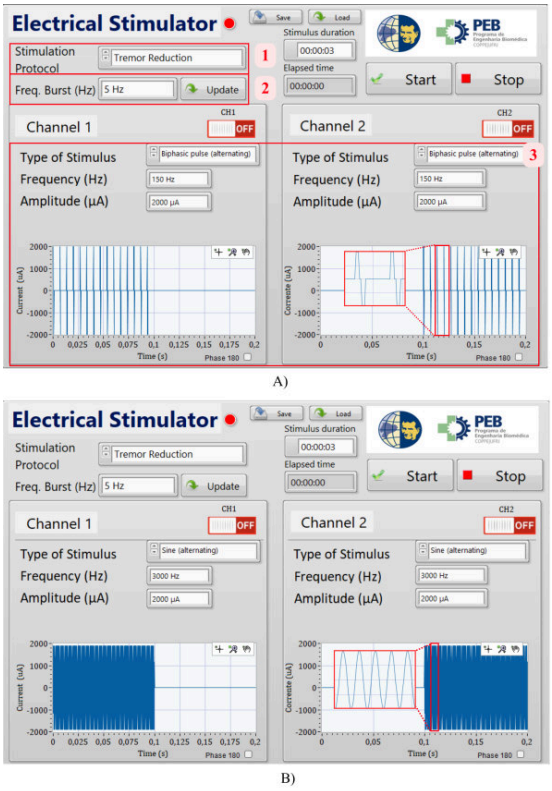


Fig. 2. User interface of the MESTIM-ELECT electrical stimulator. A) Interface configured with a charge-balanced biphasic square waveform (150 Hz); B) Interface configured with a sinusoidal waveform (3 kHz). Both configurations use the protocol designed for tremor suppression in Parkinson's disease (section 1) and include burst frequency (F_b) adjustment (section 2) to control the alternation between stimulation channels.

B. Application for monitoring tremor in Parkinson's disease

The custom graphical user interface (GUI) was successfully implemented for the real-time acquisition, visualization, and processing of triaxial acceleration signals. As shown in Fig. 3, the interface is divided into functional modules that support experimental control and simultaneous time- and frequency-domain analysis.

- *Section A* contains the main acquisition settings. It allows the operator to select and connect the acquisition device through the designated serial COM port, stop data acquisition, and configure the parameters used for spectral analysis, including the number of samples per window, window overlap, and sampling frequency.
- *Section B* integrates the controls required to monitor the experimental protocol. The tool enables operators to select a specific MDS-UPDRS movement, insert a tracking marker, and record execution time.
- *Section C* contains signal-processing controls. It allows operators to enable or disable the digital filter, select the filter type, and define the lower and upper cutoff frequencies. Users can also select the window function for power spectral density (PSD) estimation. In addition, the PSD reset control clears the accumulated spectral estimate at the start of a new task or recording stage.
- *Section D* includes a configurable PSD threshold for tremor validation and visual indicators (LEDs) that show whether the criterion is met for acceleration axis or the three-dimensional Random Vector [34].

Sections E, F, and H display the acquired and processed signals.

- *Section E* presents the acceleration signals in the time domain.
- *Section F* displays the PSD estimates and dominant frequencies for the X, Y, and Z axes. Visual indicators identify the axis with the maximum spectral peak.
- *Section G* displays the PSD of the three-dimensional Random Vector, together with its peak PSD value.
- Finally, *Section H* displays the time-domain envelope of the triaxial acceleration signals, with event markers as dashed vertical lines.

C. Pilot Study

This section presents an example of the results obtained with the Integrated Platform for Alternating Peripheral Electrical Nerve Stimulation. The complete findings of the pilot study are reported in [37].

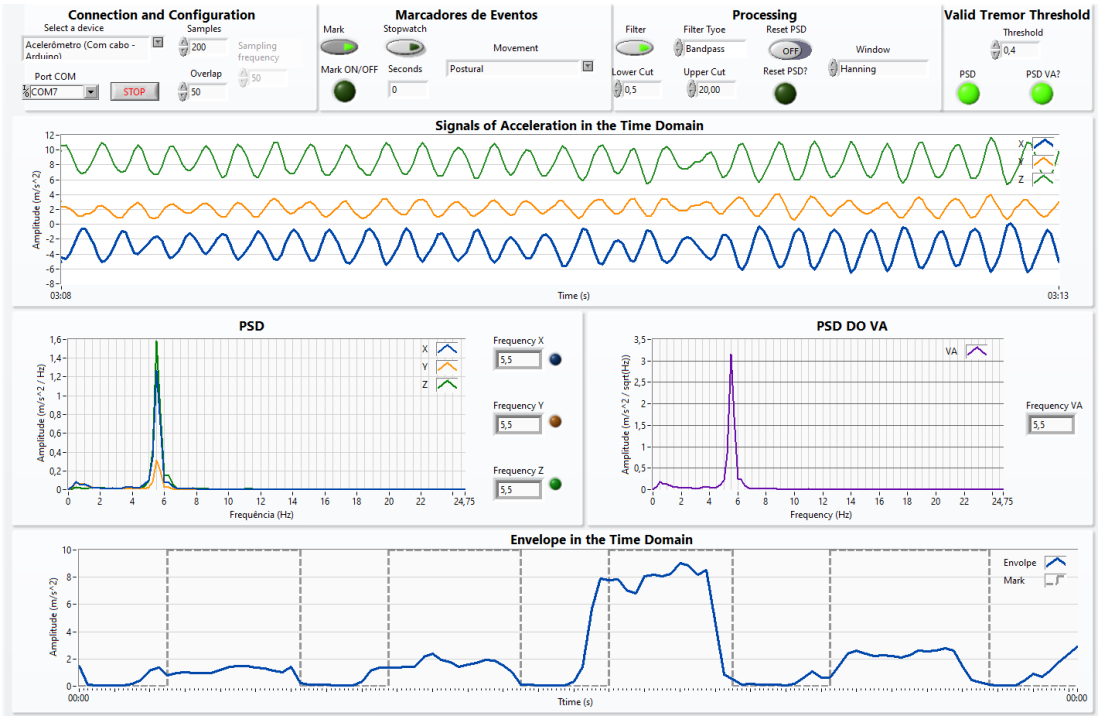


Fig. 3. Graphical user interface developed for real-time acquisition, monitoring, and analysis of tremor-related acceleration signals. A) Device connection and acquisition settings; B) MDS-UPDRS movement selection, event-marker controls, and task timing; C) Digital filtering and PSD configuration; D) Tremor validation threshold setting; E) Time-domain acceleration signals for the X, Y, and Z axes; F) PSD estimates of the three acceleration axes and their dominant frequencies G) PSD of the three-dimensional Random Vector and its dominant frequency; H) Time-domain signal envelope with MDS-UPDRS movement markers.

Fig. 4 shows the PSD of the postural tremor recorded pre- and post-stimulation for both the biphasic pulse and sinusoidal protocols. In both protocols, the dominant tremor peak power decreased post-stimulation: biphasic pulse stimulation produced a 60% reduction in peak tremor power, while sinusoidal stimulation produced a 57% reduction. In addition to changes in amplitude, the biphasic pulse protocol did not produce a change in the dominant tremor frequency, which remained at 2.5 Hz during both the pre to post stimulation; similarly, in the sinusoidal protocol, the dominant frequency also remained unchanged.

Post-stimulation, the median relative reduction across participants was -29.66% $[-71.58; -13.34]$ for the biphasic-pulse protocol and -17.25% $[-20.92; 10.86]$ for the sinusoidal protocol.

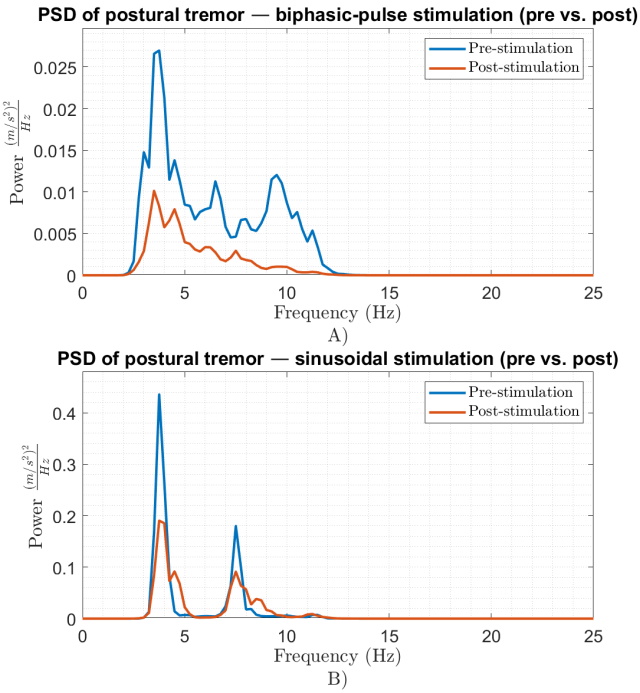


Fig. 4. Power spectral density (PSD) of postural tremor recorded pre- and post-stimulation. A) Patient A — PSD pre and post biphasic square stimulation, and B) Patient B — PSD pre and post sinusoidal stimulation.

IV. Discussions and Conclusions

In the present work, an Integrated Platform for Alternating Peripheral Nerve Stimulation was developed and implemented, combining a programmable dual-channel electrical stimulator, wireless triaxial accelerometry, and dedicated graphical interfaces for stimulation control, signal monitoring, and analysis. Its contribution is the implementation of a flexible experimental framework that accommodates alternative stimulation paradigms while providing a platform for the exploration of new waveforms. This is relevant because key aspects of peripheral electrical stimulation for Parkinson's disease, including

the site of stimulation, stimulation protocols, amplitude, and waveform design, have yet to be comprehensively evaluated, warranting further investigation [17]. Compared with commercial wrist-worn devices for specific clinical applications, such as gammaCore and Cala One, the present platform allows user-defined adjustment of waveform type, amplitude, stimulation frequency, burst frequency, channel alternation, and session duration. This configurability enabled the implementation of a charge-balanced biphasic pulse stimulus based on Transcutaneous Afferent Patterned Stimulation (TAPS) like stimulation reported in Essential Tremor and Parkinson's disease studies [23], [24]. In addition, the platform demonstrated the feasibility of implementing sinusoidal stimulation as an alternative paradigm for investigating somatosensory afferent neuromodulation. Sinusoidal stimuli are particularly relevant because current perception threshold (CPT) studies indicate that stimulation frequency can influence the activation of specific sensory fiber populations. For example, high-frequency stimuli (2-3 kHz) preferentially activate large-diameter, $A\beta$ fibers.

Another aspect of the system is the use of tremor information to individualize the stimulation protocol before the intervention. As indicated by computational simulations in [21] and [22], phase coordinated stimulation can promote desynchronization by dividing a synchronized neuronal population into multiple subpopulations (M), each receiving sequential stimuli with delays equal to a fraction of the pathological tremor oscillation. In these computational studies, M = 3 and 4 were more efficiently than M = 2 to disrupt pathological synchrony by creating several out of phase subgroups. In our practical implementation, stimulation was applied through two peripheral nerves, producing M = 2 subpopulations, similar to the configuration used in the Cala ONE [33]. Even with two stimulation subpopulations, the system still suggests tremor reduction.

This theoretical concept is also supported by studies in patients showing that knowledge of individual tremor characteristics can be relevant for defining stimulation parameters. Pascual-Valdunciel et al. [38] reported that transcutaneous spinal cord stimulation delivered at the individual tremor frequency reduced tremor amplitude by approximately 25% during the latter half of stimulation and produced a greater reduction than stimulation delivered at 21 Hz. Similarly, in patients with Parkinson's disease, Arruda et al. [39] showed that median nerve stimulation phase-locked to limb acceleration could either attenuate or amplify tremor depending on the phase at which stimulation was delivered. Together, these findings indicate that the motor response to electrical stimulation depends not only on the stimulation parameters themselves, but also on their temporal relationship with the ongoing tremor oscillation.

The burst frequency was defined based on the dominant initial postural tremor frequency estimated using the accelerometry module, allowing the stimulation parameters to be adapted to each participant. These features highlight the platform's flexibility and its potential to be adapted for a wide range of applications in research, diagnosis, and treatment.

For example, Sigurdsson et al. proposed transcutaneous cervical vagus nerve stimulation (tcVNS) as a potential therapeutic approach for gait rehabilitation and for non-motor symptoms, including cognitive impairment, depression, anxiety, gastrointestinal dysfunction, and autonomic disturbances [40]. Mondal et al. (2024) [41] subsequently evaluated tcVNS using the commercial gammaCore device, which delivers 1 ms of 5 kHz sinusoidal waveform repeated every 40 ms (25 Hz). After one month of treatment, the authors reported an increase in gait velocity of 16% (p = 0.003) and step length of 11% (p = 0.007), as well as a reduction in step time of 16% (p = 0.003), compared with pre-treatment values [41].

Beyond Parkinson's disease, one emerging application is bimodal neuromodulation combining peripheral electrical stimulation with auditory stimulation for tinnitus rehabilitation [42]. This type of paradigm could potentially be investigated by integrating our platform with devices of auditory stimulation.

Despite its flexibility, the proposed platform currently operates as an open-loop system, in which the stimulation parameters are individualized before treatment but remain fixed throughout the stimulation session. In contrast, recently developed closed-loop systems, such as the Felix NeuroAI Wristband for Essential Tremor [43], continuously monitor tremor characteristics through triaxial accelerometry and an artificial intelligence-based algorithm to continuously characterize the tremor pattern and automatically adjust the stimulation current, allowing the intervention to respond to fluctuations in tremor over time. Future work should therefore focus on integrating the MESTIM-ELECT user interface with the accelerometry system to implement a closed-loop control strategy.

Although the pilot study included five participants in each group, its results provide an initial functional validation of the complete platform during a real experimental protocol involving peripheral nerve stimulation and tremor monitoring. In the illustrative cases, both biphasic and sinusoidal stimulation were followed by reductions of -29.66% and -17.25%, respectively. These preliminary findings are consistent with the broader literature suggesting that peripheral stimulation may attenuate pathological tremor in some patients with Parkinson's disease. For example, similar results were reported by Brillman et al., who evaluated four weeks of home-based biphasic TAPS therapy in patients with Parkinson's disease and found a median 64% reduction in postural tremor power ($p < 0.001$); 79% of participants achieved at least a 50% reduction, and 78% reported that relief persisted after stimulation for a median of 60 minutes. However, the present study was limited to the assessment of pre and immediate post-stimulation responses following a single stimulation session and to the analysis of postural tremor alone. The absence of a control or sham condition prevents the observed changes from being specifically attributed to the stimulation. In addition, the small sample size limited the statistical power to detect differences between stimulation conditions. Given the marked inter- and intra-individual variability characteristic of Parkinsonian tremor, future studies should preferably adopt a crossover

design in which the same participants receive the different stimulation conditions in separate sessions, thereby allowing within-subject comparisons and reducing the influence of inter-individual variability. Future studies should therefore evaluate tremor changes over longer post-stimulation periods and include different motor tasks to determine whether the response to peripheral nerve stimulation is task-dependent and to provide a broader characterization of its effects on motor performance.

In conclusion, this work presented an integrated and flexible platform for alternating peripheral electrical nerve stimulation with real-time tremor monitoring. The platform provides a configurable experimental tool for evaluating different stimulation waveforms, individualized alternation frequencies with an interface for the visualization and analysis of accelerometer signals. Furthermore, its modular architecture enables future expansion to other experimental paradigms requiring peripheral electrical stimulation.

ACKNOWLEDGMENT

This research received financial support from the National Council for Scientific and Technological Development, Brazil (CNPq – grant 315019/2021-2), CAPES, Financiadora de Estudos e Projetos (FINEP – grant 01.24.0122.00); and Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro, Brazil (FAPERJ - grants E-26/204.393/2024).

REFERENCES

[1] P. Rossini, D. Burke, R. Chen, L. Cohen, Z. Daskalakis, R. Di-Iorio, V. Di-Lazzaro, F. Ferreri, P. Fitzgerald, M. George, M. Hallett, J. Lefaucheur, B. Langguth, H. Matsumoto, C. Miniussi, M. Nitsche, A. Pascual-Leone, W. Paulus, S. Rossi, J. Rothwell, H. Siebner, Y. Ugawa, V. Walsh, and U. Ziemann, "Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. an updated report from an i.f.c.n. committee." *Clinical Neurophysiology*, vol. 126, no. 6, pp. 1071–1107, 2015.

[2] A. Pascual-Valdunciel, A. Rajagopal, J. Pons, and S. Delp, "Non-invasive electrical stimulation of peripheral nerves for the management of tremor," *Journal of the Neurological Sciences*, vol. 435, p. 120195, 2022.

[3] M. Amat, M. García-Quintana, M. Vilchez-Barrera, A. Báez-Suárez, F. Molina-Cedrés, R. Arteaga-Ortiz, D. Alamo-Arce, and R. Medina-Ramirez, "Non-invasive electrotherapy in the rehabilitation of motor sequelae and spasticity following stroke: A systematic review," *Journal of Clinical Medicine*, vol. 15, no. 8, p. 3085, 2026.

[4] C. Li, R. Xu, Y. Qiu, W. Sun, and T. Yang, "Transcutaneous electrical nerve stimulation and opioid analgesia: Exploring a potential synergistic relationship," *Current Opinion in Pharmacology*, vol. 87, p. 102606, 2026.

[5] C. Vance, D. Dailey, R. Chimenti, B. Van-Gorp, L. Crofford, and K. Sluka, "Using tens for pain control: Update on the state of the evidence," *Medicina*, vol. 58, no. 10, p. 1332, 2022.

[6] A. Gauthey, S. Morra, P. van-de Borne, D. Deriaz, N. Maes, and J. le-Polain-de Waroux, "Sympathetic effect of auricular transcutaneous vagus nerve stimulation on healthy subjects: A crossover controlled clinical trial comparing vagally mediated and active control stimulation using microneurography," *Frontiers in Physiology*, vol. 11, p. 599896, 2020.

[7] L. Uehara, J. Corrêa, R. Ritti, P. Leite, D. de Faria, K. Pacheco-Barrios, L. Castelo-Branco, F. Fregni, and F. Corrêa, "Transcutaneous auricular vagus nerve stimulation effects on inflammatory markers and clinical evolution of patients with covid-19: a pilot randomized clinical trial." *Expert Review of Medical Devices*, vol. 19, no. 11, pp. 915–920, 2022.

[8] B. Bloem, M. Okun, and C. Klein, "Parkinson's disease," *Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[9] A. Hughes, S. Daniel, and A. Lees, "Improved accuracy of clinical diagnosis of lewy body parkinson's disease," *Neurology*, vol. 57, no. 8, pp. 1497–1499, 2001.

[10] W. Gibb and A. Lees, "The relevance of the lewy body to the pathogenesis of idiopathic parkinson's disease," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 51, no. 6, pp. 745–752, 1988.

[11] M. Dirkx and M. Bologna, "The pathophysiology of parkinson's disease tremor," *Journal of the Neurological Sciences*, vol. 435, p. 120196, 2022.

[12] M. Bologna, A. Guerra, G. Paparella, L. Giordo, D. Alunni-Fegatelli, A. Vestri, J. Rothwell, and A. Berardelli, "Neurophysiological correlates of bradykinesia in parkinson's disease," *Brain*, vol. 141, no. 8, pp. 2432–2444, 2018.

[13] N. Giladi, J. Balash, and J. M. Hausdorff, "Gait disturbances in parkinson's disease," in *Mapping the Progress of Alzheimer's and Parkinson's Disease*, ser. Advances in Behavioral Biology, Y. Mizuno, A. Fisher, and I. Hanin, Eds. Boston, MA: Springer, 2002, vol. 51, pp. 329–335.

[14] R. Cacabelos, "Parkinson's disease: From pathogenesis to pharmacogenomics," *International Journal of Molecular Sciences*, vol. 18, no. 3, p. 551, 2017.

[15] S. M. Jankovic J, "Medical management of levodopa-associated motor complications in patients with parkinson's disease," *CNS Drugs*, vol. 21, no. 8, pp. 677–692, 2007.

[16] J. M. d. Cunha and E. C. d. Siqueira, "O papel da neurocirurgia na doença de parkinson: revisão de literatura," *Revista de Medicina*, vol. 99, no. 1, pp. 66–75, 2020.

[17] L. Meng, M. Jin, X. Zhu, and D. Ming, "Peripherical electrical stimulation for parkinsonian tremor: A systematic review," *Frontiers in Aging Neuroscience*, vol. 14, p. 795454, 2022.

[18] I. Jones and M. Johnson, "Transcutaneous electrical nerve stimulation," *Continuing Education in Anaesthesia, Critical Care and Pain*, vol. 9, no. 4, pp. 130–135, 2009.

[19] A. Pascual-Valdunciel, G. W. Hoo, S. Avrillon, F. O. Barroso, J. G. Goldman, J. C. Hernandez-Pavon, and J. L. Pons, "Peripheral electrical stimulation to reduce pathological tremor: a review," *Journal of neuroengineering and rehabilitation*, vol. 18, no. 1, p. 33, 2021.

[20] A. Yadav and M. Nicolelis, "Electrical stimulation of the dorsal columns of the spinal cord for parkinson's disease," *Movement Disorders*, vol. 32, no. 6, pp. 820–832, 2017.

[21] P. Tass, "A model of desynchronizing deep brain stimulation with a demand-controlled coordinated reset of neural subpopulations," *Biological cybernetics*, vol. 89, pp. 81–8, 09 2003.

[22] O. V. Popovych and P. A. Tass, "Desynchronizing electrical and sensory coordinated reset neuromodulation," *Frontiers in Human Neuroscience*, vol. 6, 2012. [Online]. Available: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2012.00058

[23] R. Pahwa, R. Dhall, J. Ostrem, R. Gwinn, K. Lyons, S. Ro, C. Dietiker, N. Luthra, P. Chidester, S. Hamner, E. Ross, and S. Delp, "An acute randomized controlled trial of noninvasive peripheral nerve stimulation in essential tremor," *Neuromodulation: Technology at the Neural Interface*, vol. 22, no. 5, pp. 537–545, 2019.

[24] S. Brillman, P. Khemani, S. H. Isaacson, R. Pahwa, R. Deshpande, V. Zraick, A. Rajagopal, D. Khosla, and K. H. Rosenbluth, "Non-invasive transcutaneous afferent patterned stimulation therapy offers action tremor relief in parkinson's disease," *Tremor and Other Hyperkinetic Movements*, vol. 13, no. 1, pp. 1–11, 2023, art. no. 25.

[25] J. Ara, S. H. Hwang, T. Song, and G. Khang, "Effects of the stimulus parameters on the tactile sensations elicited by single-channel transcutaneous electrical stimulation," *International Journal of Precision Engineering and Manufacturing*, vol. 15, pp. 305–313, 2014.

[26] B. Geng, K. Yoshida, L. Petrini, and W. Jensen, "Evaluation of sensation evoked by electrocutaneous stimulation on forearm in nondisabled subjects," *Journal of Rehabilitation Research Development*, vol. 49, no. 2, pp. 297–308, 2012.

[27] J. P. Reilly, *Applied Bioelectricity: From Electrical Stimulation to Electropathology*. New York, NY: Springer, 1998.

[28] L. P. Volpi, "Efeitos dos estímulos elétricos senoidais na resposta induzida cerebral: estudo da seletividade das fibras sensitivas," Dissertação de Mestrado em Engenharia Biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2017.

[29] H. Resende-Martins, R. Zanetti, C. Cardoso-dos Santos, G. Mastrocola-Manzano, and C. J. Tierra-Criollo, "Current perception threshold and reaction time in the assessment of sensory peripheral nerve fibers through sinusoidal electrical stimulation at different frequencies," *Revista Brasileira de Engenharia Biomédica*, vol. 29, pp. 278–285, 2013.

[30] J. Katims, D. Long, and L. Ng, "Transcutaneous nerve stimulation. frequency and waveform specificity in humans," *Applied Neurophysiology*, vol. 49, no. 1-2, pp. 86–91, 1986.

[31] J. J. Fernández-Pérez, J. Avendaño-Coy, D. Serrano-Muñoz, F. Oliveira Barroso, C. Montero-Pardo, B. López-Moreno, A. Lerín-Calvo, J. P. Romero Muñoz, and J. Gómez-Soriano, "Transcutaneous kilohertz high-frequency alternating current at 10 khz for upper-limb tremor in people with parkinson's disease: A double-blind, randomized, crossover study," *Journal of Clinical Medicine*, vol. 13, no. 24, p. 7566, 2024.

[32] D. A. Molina-Vidal, P. Cevallos Larrea, L. Guambaña Calle, D. Liquori, and C. J. Tierra-Criollo, "Design and evaluation of a prototype of dual channel electrical stimulator: Application in the assessment of current perception thresholds to multiple stimuli," in *Smart Technologies, Systems and Applications: SmartTech-IC 2022*, ser. Communications in Computer and Information Science, vol. 1705. Cham: Springer, 2023, pp. 342–355.

[33] J. Yu, A. Rajagopal, J. Syrkin-Nikolau, S. Shin, K. Rosenbluth, D. Khosla, E. Ross, and S. Delp, "Transcutaneous afferent patterned stimulation therapy reduces hand tremor for one hour in essential tremor patients," *Frontiers in Neuroscience*, vol. 14, p. 530300, 2020.

[34] P. S. S. Cevallos, D. M. da Silva-Liquori, and C. J. Tierra-Criollo, "Estimador da frequência de tremor em parkinsonianos com vetor aleatório de acelerometria," 2025.

[35] H. Dai, P. Zhang, and T. C. Lueth, "Quantitative assessment of parkinsonian tremor based on an inertial measurement unit," *Sensors*, vol. 15, no. 10, pp. 25 055–25 071, 2015. [Online]. Available: https://doi.org/10.3390/s151025055

[36] F. Luft, S. Sharifi, W. Mugge, A. C. Schouten, L. J. Bour, A.-F. van Rootselaar, P. H. Veltink, and T. Heida, "A power spectral density-based method to detect tremor and tremor intermittency in movement disorders," *Sensors*, vol. 19, no. 19, p. 4301, 2019. [Online]. Available: https://doi.org/10.3390/s19194301

[37] P. Silverio, P. S., E. Pesántez-Vera, D. M. d. S. Liquori, D. A. Molina-Vidal, and C. J. Tierra-Criollo, "Effects of peripheral electrical stimulation waveforms on tremor in parkinson's disease: A pilot study," 2026, manuscript submitted to the XXX Brazilian Congress of Biomedical Engineering (CBEB 2026), unpublished.

[38] A. Pascual-Valdunciel, J. Ibáñez, L. Rocchi, J. Song, J. C. Rothwell, K. P. Bhatia, D. Farina, and A. Latorre, "Frequency-selective suppression of essential tremor via transcutaneous spinal cord stimulation," *Movement Disorders*, vol. 39, no. 10, pp. 1817–1828, 2024. [Online]. Available: https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.29966

[39] B. Silveira de Arruda, C. Reis, J. Sermon, A. Pogosyan, P. Brown, and H. Cagnan, "Identifying and modulating distinct tremor states through peripheral nerve stimulation in parkinsonian rest tremor," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, 12 2021.

[40] H. P. Sigurdsson, R. Raw, H. Hunter, M. R. Baker, J. P. Taylor, L. Rochester, and A. J. Yarnal, "Noninvasive vagus nerve stimulation in parkinson's disease: current status and future prospects," *Expert review of medical devices*, vol. 18, no. 10, pp. 971–984, 2021.

[41] B. Mondal, S. Choudhury, R. Banerjee, A. Roy, K. Chatterjee, P. Basu, R. Singh, S. Halder, S. Shubham, S. Baker, M. Baker, and H. Kumar, "Effects of non-invasive vagus nerve stimulation on clinical symptoms and molecular biomarkers in parkinson's disease," *Frontiers in Aging Neuroscience*, vol. 15, p. 1331575, 2024.

[42] E. E. McMahan and H. H. Lim, "Retrospective chart review demonstrating effectiveness of bimodal neuromodulation for tinnitus treatment in a clinical setting," *Communications Medicine*, vol. 5, no. 1, p. 112, 2025. [Online]. Available: https://doi.org/10.1038/s43856-025-00837-3

[43] R. r. Dewey, S. Isaacson, R. J. Dewey, S. Betté, K. Lyons, Z. Yang, A. Nguyen, Q. Zhao, Z. Zhang, and R. Pahwa, "A pilot study of ai-controlled transcutaneous peripheral nerve stimulation for essential tremor," *Tremor and Other Hyperkinetic Movements*, vol. 15, p. 10, 2025.

App My Pill: Design and Case Demonstration of a Low-Data Reference-Based System for Loose Pill Identification

1st Thainara Alves Gouvea
Laboratory of Bionanotechnology
Institute of Research and Development
University of Vale do Paraiba
Sao Jose dos Campos, Brazil
thaigouvea20@gmail.com

2nd Luara Teodoro Lima
Laboratory of Bionanotechnology
Institute of Research and Development
University of Vale do Paraiba
Sao Jose dos Campos, Brazil
luaratlima@icloud.com

3rd Lara Silva Reboucas
Laboratory of Bionanotechnology
Institute of Research and Development
University of Vale do Paraiba
Sao Jose dos Campos, Brazil
larasilvasouzareboucas@hotmail.com

4th Gustavo Emmanuel Cereza Oliveira
Universidade Anhanguera de Sao Jose dos Campos
Sao Jose dos Campos, Brazil
contato.elcereza@gmail.com

5th Maiara Castilho
Laboratory of Bionanotechnology
Institute of Research and Development
University of Vale do Paraiba
Sao Jose dos Campos, Brazil
mcastilho@univap.br

6th Filipe Loyola Lopes
Centro de Tecnologia da Informacao Renato Archer (CTI)
Campinas, Brazil
flopes@cti.gov.br

***Abstract*—Loose tablets and capsules separated from their original packaging are difficult to identify because many pharmaceutical presentations share similar colors, shapes, and dimensions. This paper presents App My Pill, a low-data, reference-based computer vision prototype designed to rank possible matches for a loose solid oral dosage form from images of its two faces. The pipeline combines central-region segmentation, Lab color contrast, edge detection, morphological processing, GrabCut refinement, optional manual region selection, MobileNetV2 visual embeddings, imprint-text similarity, geometric descriptors, color features, and optional metric scale from an ArUco marker. A dynamic fusion strategy reduces the influence of physical descriptors when the crop heuristic is poor and applies a rule-based decision and abstention mechanism based on total compatibility, candidate margin, visual evidence, and the internal crop heuristic. The prototype intentionally reports a ranked hypothesis rather than a medical identification and abstains when the predefined criteria are not met. The current database contains two exact presentations: simethicone 40 mg, Neo Quimica, white tablet, and paracetamol 750 mg, Prati, each with 150 reference images per face. For each presentation and each face, 100 images were acquired under the same lighting setup and 50 additional images were acquired in different environments and lighting conditions. All reference and query images were acquired with a Samsung Galaxy A51, and the demonstrated simethicone query used files distinct from the reference images. In this two-class case demonstration, simethicone was ranked first with a compatibility index of 82.5%, compared with 65.3% for paracetamol. The internal crop heuristic was $Q_h = 0.977$. Because the no-imprint acceptance threshold was 86%, the system returned an unconfirmed hypothesis. This single case demonstrates execution of the proposed workflow but does not estimate accuracy, sensitivity, specificity, generalization, robustness, or clinical safety.**

***Index Terms*—computer vision, pill identification, medication safety, low-shot recognition, image segmentation, MobileNetV2, pharmaceutical informatics**

I. Introduction

Medication-related harm remains a major global patient-safety concern. The World Health Organization (WHO) updated its systematic review of preventable medication-related harm in 2024 and continues to promote system-level interventions through the Medication Without Harm initiative [1], [2]. Although medication errors span prescribing, dispensing, administration, monitoring, and transitions of care, one practical problem occurs when a tablet or capsule is separated from its labeled container. In this situation, patients, caregivers, pharmacists, and health professionals may need to distinguish among visually similar products without reliable contextual information.

The visual identification problem is intrinsically fine-grained. Numerous products may be white and round, while the same active ingredient may appear in different doses, manufacturers, colors, coatings, or imprint patterns. Conversely, different active ingredients may share nearly identical gross appearance. Regulations for solid oral dosage forms in the United States explicitly recognize that reliable identification depends on the combination of imprint, size, shape, and color, rather than on a single visual property [5]. Therefore, an automated system should treat each exact pharmaceutical presentation as a separate class and should integrate complementary evidence.

This U.S. regulatory context should not be directly generalized to Brazil. The Brazilian Health Regulatory Agency (Agencia Nacional de Vigilancia Sanitaria, Anvisa) establishes national medication-labeling requirements through RDC No. 768/2022 and provides official consultation resources for registered medicines, labels, and package inserts [3], [4].

Accordingly, visual features used by App My Pill are treated as retrieval cues only, and any practical identification must still be confirmed using original packaging, official regulatory information, manufacturer information, or a qualified health professional.

Computer vision has been applied to pill recognition through handcrafted descriptors, convolutional neural networks (CNNs), metric learning, retrieval systems, object detectors, and multimodal architectures [6]–[10], [13], [14]. However, real-world deployment remains challenging. Consumer photographs differ from curated reference images in illumination, viewpoint, background, scale, occlusion, reflections, device characteristics, and image quality. Low-shot benchmarks such as ePillID demonstrate the difficulty of distinguishing thousands of appearance classes with very few reference examples [10]. Recent work also emphasizes that segmentation quality and cross-domain distribution shift may remain limiting even when semantic classification appears strong [15], [16].

This study proposes App My Pill, a proof-of-concept system intended for low-data prototyping. Instead of training a new classifier for each medication class, the system builds a reference database from a small number of images and performs similarity-based retrieval. The current contribution is fourfold: 1) a two-sided reference-matching architecture combining deep visual embeddings with physical and textual cues; 2) a segmentation pipeline with both automatic and manual fallback modes; 3) an explicit abstention mechanism that separates a ranked hypothesis from a confirmed output; and 4) a transparent preliminary case demonstration in which all score components can be reproduced from the configured weights.

The prototype is not intended to authorize medication ingestion, replace a pharmacist, or provide a clinical diagnosis. Its current purpose is to investigate technical feasibility and define a validation pathway for a future assistive system.

II. Related Work

A. Pill Recognition and Retrieval

Early approaches relied on shape, color, texture, and imprint descriptors, whereas later studies adopted CNNs to learn discriminative representations. Wong *et al.* developed a fine-grained deep convolutional approach for unconstrained mobile images [7]. MobileDeepPill combined multiple CNN branches to capture pill shape, color, and imprint while reducing the model footprint for mobile use [8]. Delgado *et al.* demonstrated a medication-identification service based on reference and consumer images [9].

The ePillID benchmark formalized pill identification as a low-shot, fine-grained retrieval problem with two appearance classes per pill, corresponding to its two faces [10]. This formulation is directly relevant to loose pills because the two sides may contain different imprints or surface characteristics. Subsequent studies evaluated drug recognition in hospital workflows, retrieval systems, and multinational databases [11]–[14]. Across these studies, high reported performance is commonly associated with a particular dataset and acquisition protocol; it should not be assumed to transfer unchanged to a new population of products or consumer devices.

B. Segmentation and Domain Shift

Pill recognition requires both localization and classification. Background leakage, shadows, fingers, packaging, and fiducial markers can distort geometric and color measurements. GrabCut provides an iterative graph-cut method that combines edge and color information for foreground extraction with limited user interaction [18]. Fiducial markers such as ArUco provide a robust reference for scale and camera geometry [19].

A 2025 review identified dataset scarcity, visually confusing classes, imaging variability, and deployment robustness as continuing challenges in pill recognition [6]. The 2026 MEDISEG dataset expanded attention toward instance segmentation in cluttered and overlapping medication scenes [15], while deployment-oriented few-shot evaluation showed that localization and recall can deteriorate under domain shift even when class adaptation is rapid [16]. These findings motivate the explicit measurement of segmentation quality and the use of abstention in the present prototype.

III. Materials and Methods

A. System Overview

The system was implemented as a desktop proof of concept in Python 3.14 using OpenCV 5.0, NumPy, an ONNX version of MobileNetV2, and a Tkinter graphical interface. The processing pipeline is summarized in Fig. 1. The user may acquire images through a camera or select stored files. Two images are required, corresponding to face A and face B. The sides may be automatically matched in either direct or swapped order, reducing dependence on how the user labels the faces.

Each database class represents an exact presentation rather than only an active ingredient. The metadata structure includes a class identifier, display name, active ingredient, dose strength, manufacturer, dosage form, expected imprint on each face, expected color, and expected shape. This design follows the principle that product identification depends on the joint physical presentation [5], [20].

B. Reference Database and Image Acquisition

Reference images are organized by class and face. Multiple photographs of each face can be stored, allowing the reference representation to include modest variations in rotation and acquisition. The current preliminary database contains two exact presentations: Simethicone 40 mg—Neo Quimica, white tablet, and Paracetamol 750 mg—Prati, tablet. For both presentations, 150 reference images were acquired for face A and 150 for face B. For each presentation and each face, 100 images were obtained under the same lighting setup and 50 additional images were obtained in different environments and lighting conditions. All reference photographs and the demonstrated query were acquired with a Samsung Galaxy A51. The two simethicone query images were separate files and were not reused as reference images. However, acquisition and parameter development occurred iteratively, without prospectively separated

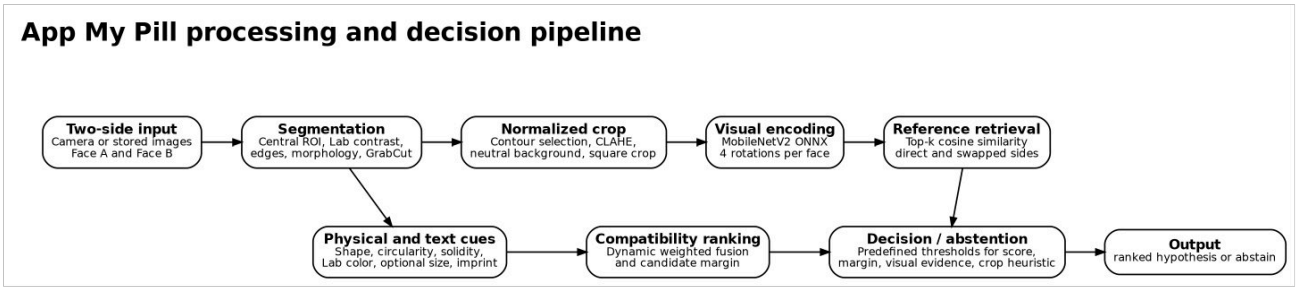


Fig. 1. Processing and decision pipeline. Two face images are segmented and normalized before visual embedding extraction. Deep visual similarity is fused with available physical and imprint evidence. A predefined decision and abstention rule returns either a ranked hypothesis or an abstention; it is not a clinically validated safety mechanism.

development, validation, and test sessions. Consequently, the demonstrated query is independent at the file level but cannot be considered an independent test set. Because data collection was still in progress, the reference set was intentionally treated as a proof-of-concept database rather than a statistically powered dataset.

Recommended acquisition conditions include a matte, uniform, high-contrast background; a perpendicular camera; limited glare; and one pill centered in the acquisition region. An ArUco marker from the DICT_4X4_50 dictionary, identifier 23, with a nominal printed side length of 20 mm, can be placed in the same plane as the pill. When detected, the mean marker-side length in pixels is divided by 20 mm to estimate pixels per millimeter. Size is omitted from the compatibility calculation when no valid scale is available.

C. Automatic and Manual Segmentation

The automatic pipeline first defines a central region of interest (ROI), corresponding to 76% of the image dimensions for still-image processing and 62% for the live camera guide. Marker polygons are expanded and masked so that the fiducial pattern does not compete with the pill contour.

The image is converted to the CIE Lab color space. Border pixels are used to estimate the median background color, and a per-pixel Euclidean distance from the background is calculated. Three candidate foreground signals are combined: 1) Otsu thresholding of the normalized Lab-distance image; 2) a robust threshold based on the median and median absolute deviation; and 3) Canny edges followed by region filling. Morphological opening, closing, and median filtering suppress isolated noise and bridge small contour gaps.

GrabCut refinement is initialized from the combined mask and the central ROI [18]. Candidate contours are rejected when their area is outside configured ROI-relative bounds. Remaining contours are scored using centrality, area plausibility, compactness, solidity, color contrast, and whether they touch the ROI boundary. The highest-scoring contour is filled and morphologically smoothed.

Automatic segmentation can fail when illumination is nonuniform, the object has low contrast, or the selected area contains a hand or large shadow. Therefore, the graphical interface includes manual ROI selection. In this mode, the user draws a rectangle around the pill; GrabCut is applied inside the selection, everything outside it is discarded, and the isolated object is placed on a uniform background before the normal processing pipeline resumes.

D. Internal Crop Heuristic

An internal crop heuristic $Q_h \in [0, 1]$ is calculated from contour centrality, solidity, compactness, foreground-background contrast, and boundary contact. This value is an engineering heuristic and not a percentage of segmentation quality, a Dice coefficient, or an intersection-over-union score because pixel-level ground-truth masks were not available. It serves two roles: warning the user about visibly poor crops and reducing the influence of physical descriptors when the segmentation is unreliable.

The quality modulation factor is

$$q = \mathrm{clip}\left(\frac{Q_h - 0.25}{0.65}, 0, 1\right). \tag{1}$$

Thus, physical measurements receive their full configured weight only when the internal crop heuristic is sufficiently high.

E. Crop Normalization and Visual Embeddings

The selected contour is cropped with a 16% margin around its bounding box, padded to a square, resized to 320×320 pixels, and placed on a neutral background. Contrast-limited adaptive histogram equalization is applied to the Lab lightness channel to reduce illumination variation while preserving chromatic channels.

Visual representations are extracted using MobileNetV2, a lightweight CNN based on inverted residuals and linear bottlenecks [17]. The ONNX network receives 224×224 RGB images normalized by ImageNet channel statistics. The output vector is flattened and L_2 normalized. Four embeddings are generated per face using rotations of $0°$, $90°$, $180°$, and $270°$, reducing sensitivity to in-plane orientation without training a new model.

For each query rotation, cosine similarity is calculated against the reference embeddings. The median of the strongest $k = 4$

reference similarities is retained. The two-face visual score for candidate c is

$$V_c = \max\left[\frac{v^{(c)}_{A,A} + v^{(c)}_{B,B}}{2}, \frac{v^{(c)}_{A,B} + v^{(c)}_{B,A}}{2}\right], \tag{2}$$

where the first term represents direct face correspondence and the second represents swapped correspondence.

F. Physical and Imprint Features

Shape is represented by aspect ratio, circularity, and solidity. Candidate similarity is calculated against the median and median absolute deviation of the reference class. Color is measured inside an eroded mask to reduce contamination from shadows and background pixels. The Lab a and b channels provide chromatic evidence, while relative lightness compares the pill interior with a local background ring. Optional width and height are estimated when the ArUco marker is available.

The physical score is a weighted mean over the available physical components,

$$P_c = \frac{\sum_{j\in\mathcal{A}} w_j s_{j,c}}{\sum_{j\in\mathcal{A}} w_j}, \tag{3}$$

where $\mathcal{A}$ contains only the available components and the nominal weights are 0.48 for shape, 0.42 for color, and 0.10 for size. Therefore, unavailable size is excluded from both numerator and denominator and does not impose an artificial 50% penalty.

Imprints are currently entered by the user rather than recognized automatically. Text is uppercased, accents and non-alphanumeric characters are removed, and visually confusable groups such as 0/O, 1/I/L, 2/Z, 5/S, and 8/B are normalized before fuzzy string comparison. The direct and swapped face assignments are evaluated, and the larger two-face imprint score is retained. If both the observed and expected imprints are blank, the interface may display agreement; however, the imprint component is not included in the total score unless the user actually enters an imprint. Automatic optical character recognition remains future work.

G. Dynamic Compatibility Index

The compatibility index for candidate c combines visual, imprint, and physical evidence:

$$S_c = w_v(Q_h)V_c + w_i I_c + w_p(Q_h)P_c. \tag{4}$$

When an imprint is supplied, the nominal weights are 0.55 for visual evidence, 0.30 for imprint, and 0.15 for physical evidence. The physical weight becomes $0.15q$, and its unused portion is transferred to the visual component. Without an imprint, the nominal weights are 0.78 visual and 0.22 physical; the physical weight becomes $0.22q$, and $w_v = 1 - w_p$. This design prevents uncertain contours from producing excessive influence through color or shape.

The fusion weights and decision thresholds used in the present prototype were selected heuristically during iterative engineering development. They were not optimized, calibrated, or statistically derived from an independent validation set, and complete independence between parameter selection and previously observed development behavior cannot be guaranteed. Accordingly, these values are operating rules for the prototype and must not be interpreted as clinically validated thresholds or as controlling a known false-accept rate.

The compatibility index is a deterministic similarity score, not a calibrated posterior probability. Modern neural-network scores can be poorly calibrated even when classification performance is high [21]; consequently, percentages shown by the interface are explicitly labeled as compatibility rather than confidence or probability.

H. Decision and Abstention Rule and User Interface

Candidates are ordered by S_c. Let $\Delta = S_{(1)} - S_{(2)}$ denote the margin between the first and second candidates. Without an imprint, the current prototype accepts a possible match only when $S_{(1)} \geq 0.86$, $\Delta \geq 0.075$, $V_{(1)} \geq 0.82$, and $Q_h \geq 0.55$. With an imprint, the corresponding thresholds are 0.80, 0.045, 0.74, and 0.55. At least two classes must exist in the reference database before any result can be accepted. These configured thresholds implement the intended abstention behavior but have not been validated as a clinical safety mechanism.

When one or more conditions fail, the system still displays the highest-ranked candidate but labels it as an unconfirmed hypothesis. The graphical interface supports camera capture, stored images, face-specific manual cropping, optional imprint fields, ranked candidate details, an annotated image with contour and label, and text-to-speech output. Every output includes a warning not to use the result as the sole basis for medication ingestion. When metric size is unavailable, the revised interface displays "not evaluated" rather than the neutral 50% placeholder used in the development build shown in Fig. 2.

IV. Case Demonstration

A. Two-Class Query

A query was performed using two images of a Simethicone 40 mg—Neo Quimica white tablet. The two query files were distinct from the reference images. Manual cropping was applied to both faces. No imprint was entered. The system compared the query with the two presentations in the database and produced the results in Table I. The internal crop heuristic was $Q_h = 0.977$, corresponding to $q = 1$ after clipping in (1). This value is not a validated segmentation-accuracy percentage.

TABLE I
Two-Class Candidate Ranking in the Case Demonstration

Component	Simethicone 40 mg	Paracetamol 750 mg
Visual similarity V_c	0.860	0.728
Shape similarity H_c	0.719	0.214
Color similarity C_c	0.682	0.582
Size similarity Z_c	unavailable	unavailable
Internal crop heuristic Q_h	0.977	0.977
Compatibility index S_c	**0.825**	0.653
Rank	**1**	2

Because size was unavailable, the physical score was renormalized over shape and color. For simethicone,

$$P_{\text{sim}} = \frac{0.48(0.719) + 0.42(0.682)}{0.90} = 0.702, \tag{5}$$

and

$$S_{\text{sim}} = 0.78(0.860) + 0.22(0.702) = 0.825. \tag{6}$$

For paracetamol, the same calculation produced approximately 0.653. The first-to-second margin was 0.172, exceeding the no-imprint margin requirement of 0.075. The visual and internal-crop-heuristic requirements were also met. However, the winning total score remained below the no-imprint threshold of 0.86; therefore, the system correctly returned "principal hypothesis: Simethicone 40 mg—Neo Quimica" while retaining the "not confirmed" warning.

B. Interpretation of the Preliminary Scores

The visual embedding provided the strongest discriminatory signal, with an 86.0% visual similarity for simethicone versus 72.8% for paracetamol. The shape and color components also favored simethicone. The candidate margin shows that the expected class was ranked above the only alternative in this specific example. Nevertheless, this case cannot be interpreted as classification accuracy, sensitivity, specificity, or clinical performance. It is a single demonstration in a closed two-class database and does not assess unknown pills, additional manufacturers, damaged products, lighting variation, or independent users.

The abstention is therefore a desirable result rather than a failure. A system designed for medication support should avoid transforming a plausible match into an apparently definitive statement. This aligns with broader recommendations that health-related AI systems should prioritize safety, transparency, human oversight, and accountability [22].

V. Discussion

The preliminary experiment demonstrates that a hybrid, reference-based approach can provide a meaningful ranking with very limited class-specific data. This differs from a conventional supervised classifier that requires retraining when a new presentation is added. In the current design, a new class can be enrolled by adding metadata and reference images, after which embeddings and physical features are appended to the reference database.

The current reference set contains the same number of images for both enrolled presentations (300 images per class, 150 per face). Even with this balanced reference count, the present case does not quantify how reference density would affect performance as the database expands. In addition, no baseline or ablation experiments were performed; the contribution of two-face matching, physical cues, and crop-quality weighting therefore remains to be measured in an independent study.

The use of two faces is particularly important. Imprints, score lines, coatings, and logos may differ across sides, and a one-sided query can discard useful evidence. Direct and swapped matching also reduces user-interface error because the user does not need to know which side was labeled A during enrollment. This design is consistent with low-shot pill-recognition datasets that treat the two sides as separate appearance classes [10].

The case also illustrates why the compatibility value should not be described as a probability. The result is generated by a configured fusion rule, and its numerical scale depends on reference density, feature distributions, and threshold choices. Before deployment, the score would need calibration on a representative validation set, including known classes and unknown products. Calibration metrics and reliability diagrams should complement top-k classification results [21].

Segmentation remains the most immediate engineering limitation. The internal crop heuristic reached $Q_h = 0.977$ in the presented case, yet visual inspection still showed that the contour included some surrounding background. This discrepancy occurs because the indicator measures internal geometric and contrast heuristics rather than overlap with a human-annotated ground truth and therefore must not be interpreted as 97.7% segmentation accuracy. Future segmentation evaluation should use pixel-level masks and report intersection over union and Dice coefficient. The recent MEDISEG work is relevant because it provides instance masks under clutter, overlap, varied backgrounds, and reflective conditions [15].

The current physical features are interpretable and useful for error analysis. For example, the paracetamol candidate received a markedly lower shape similarity than simethicone. However, physical descriptors can also amplify acquisition bias: Lab color changes under illumination and white balance, apparent dimensions depend on perspective, and shape descriptors change when the contour includes shadows. The dynamic weighting mechanism mitigates but does not eliminate these risks.

The current system requires user-entered imprint text. This approach avoids premature dependence on OCR, but it limits usability. Pill imprints can be embossed, debossed, curved, partially occluded, low-contrast, or rotated. Future work should evaluate geometry-aware OCR and should preserve uncertainty rather than forcing a text string. The regulatory rationale for imprint use remains strong because the combination of imprint, size, shape, and color is designed to support unique identification [5].

For the Brazilian context, the FDA imprint framework is used only as international background. Anvisa regulations and official consultation systems govern Brazilian medication labeling and access to registered product information [3], [4]. Therefore, a future Brazilian deployment must preserve confirmation through packaging, official Anvisa information, manufacturer data, or a qualified health professional.

VI. Limitations and Risk Considerations

The present study has several important limitations.

First, the database includes only two enrolled presentations, with 300 reference images for each class. A two-class ranking is much easier than retrieval among hundreds or thousands of visually similar products. Second, the result is based on one

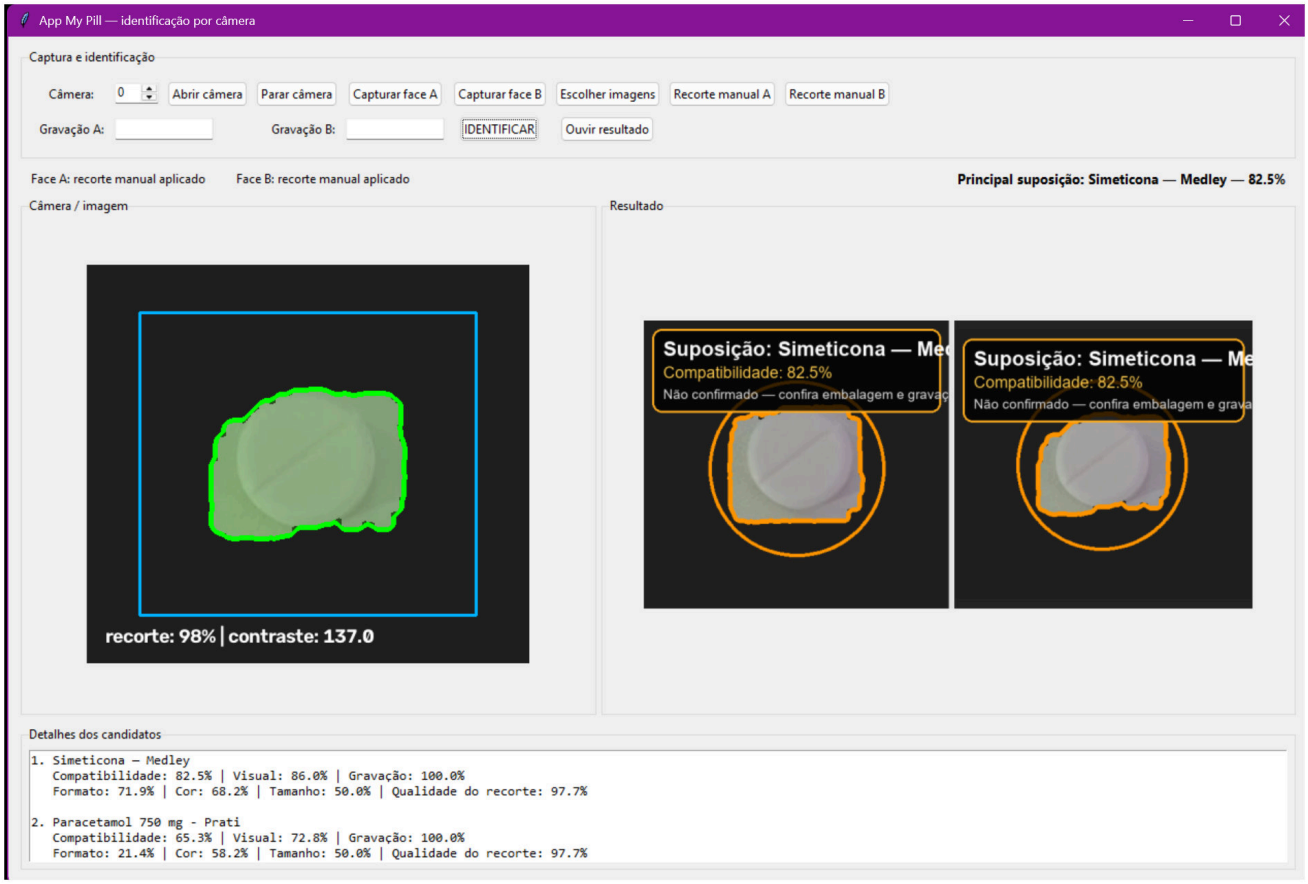


Fig. 2. Prototype interface during the simethicone case demonstration. The expected class was ranked first at 82.5%, but the result remained unconfirmed because the predefined no-imprint acceptance threshold was not reached. The interface is shown in Portuguese, the language used during prototype development. The screenshot corresponds to the development build, in which unavailable size was displayed as a neutral 50%; the revised interface displays this field as "not evaluated."

query case and cannot support any statistical performance claim. Third, all images were acquired with a Samsung Galaxy A51, and although the query files were distinct from the reference files, development, validation, and test acquisitions were not prospectively separated by session. Fourth, the internal crop heuristic has not been validated against annotated masks. Fifth, size information was unavailable in the demonstrated case; the neutral 50% placeholder visible in the development-build screenshot is not used in the compatibility calculation and has been replaced by "not evaluated" in the revised interface.

Sixth, the prototype operates in a closed-set manner: every query is ranked against enrolled classes, even when the true product is absent. Although threshold-based abstention reduces risk, robust open-set recognition requires dedicated unknown-class evaluation. Seventh, the MobileNetV2 representation was pretrained on natural images rather than pharmaceutical imagery and was not fine-tuned on a representative pill dataset. Eighth, color and shape alone cannot guarantee unique identification. Ninth, product appearance can change with manufacturer, market, dose, coating, storage, damage, or counterfeit status.

Finally, the software is a research prototype and has not undergone clinical evaluation, usability validation, cybersecurity assessment, regulatory classification, or quality-system development. WHO guidance emphasizes that AI for health should embed ethics and human rights throughout design and deployment [22]. Accordingly, any future practical version should remain an assistive tool, clearly communicate uncertainty, protect user data, preserve traceability, and require confirmation through original packaging, an authoritative database, or a qualified health professional.

VII. Proposed Validation Protocol

A subsequent study should separate enrollment, validation, and test images by acquisition session and, where possible, by device and user. The number of classes should be expanded to include visually confusable presentations, multiple manufacturers, multiple doses, and negative or unknown samples. Baselines should include visual-only, physical-only, and one-face retrieval, while ablations should remove two-face matching, physical cues, and crop-quality weighting one component at a time. Table II summarizes the proposed evaluation framework.

Thresholds should be chosen on the validation set according to a safety objective rather than maximum overall accuracy. For example, the system may prioritize a low false-accept

TABLE II
PROPOSED EVALUATION FRAMEWORK FOR THE NEXT DEVELOPMENT PHASE

Domain	Primary Measures	Recommended Stress Conditions
Segmentation	Intersection over union, Dice coefficient, boundary error, failure rate	Low contrast, hard shadows, reflective coatings, hands, clutter, partial occlusion, different backgrounds
Recognition	Top-1 and top-3 accuracy, macro precision, recall, F1 score, confusion matrix	Similar color/shape classes, different doses, manufacturers, front/back swaps, damaged imprints
Open-set safety	Unknown-detection AUROC, false-accept rate, false-reject rate, risk-coverage curve	Non-enrolled pills, vitamins, candy, fragments, counterfeit-like samples
Calibration	Expected calibration error, Brier score, reliability diagram	Device shift, illumination shift, background shift, reference-count variation
Robustness	Performance stratified by device, user, orientation, scale, and lighting	Independent smartphones, camera resolutions, indoor/outdoor illumination, motion blur
Baselines and ablations	Same recognition and risk-coverage measures	Visual-only, physical-only, one-face, two-face, full fusion; remove physical cues, swapped matching, or crop-quality weighting
Usability	Task completion, capture failure, time, comprehension of warnings, trust calibration	Patients, caregivers, pharmacists, older adults, users with visual or motor limitations
Efficiency	End-to-end latency, memory, CPU use, database growth	Offline CPU inference, increasing class count, mobile-device deployment

rate and tolerate more abstentions. External testing should then be conducted without changing the selected thresholds. Recent evidence on few-shot pill recognition under domain shift reinforces the need for deployment-oriented testing rather than only within-dataset evaluation [16].

Source code, YAML configuration files, model-identification information for MobileNetV2-10 ONNX, and reproducibility instructions are retained by the authors and can be provided with the research materials. The current image collection was produced during prototype development and is not presented as a public benchmark dataset.

VIII. FUTURE WORK

The next technical phase will focus on five areas. First, the reference database will be expanded and versioned with standardized metadata. Second, segmentation will be evaluated against human masks and may be replaced or augmented by a lightweight learned model. Third, automatic imprint extraction will be investigated using orientation-aware OCR. Fourth, the compatibility score will be calibrated and extended with explicit open-set detection. Fifth, the desktop proof of concept will be translated into a mobile architecture with offline inference, accessibility functions, and auditable result logs.

A clinically relevant study will require collaboration with pharmacists and medication-safety specialists. Ground-truth labels should originate from verified packaging, regulatory records, or manufacturer data. The evaluation should also consider the difference between identifying an active ingredient and uniquely identifying a presentation. The latter requires strength, manufacturer or distributor, and physical appearance information, as reflected in imprint regulation [5].

IX. CONCLUSION

App My Pill demonstrates the feasibility of a low-data, reference-based pipeline for preliminary loose-pill matching. The system combines two-sided MobileNetV2 embeddings, physical descriptors, optional imprint evidence, internal-crop-heuristic-aware weighting, and a predefined decision and abstention rule. In the initial two-class case, simethicone was ranked above paracetamol with a 17.2 percentage-point margin, but the system abstained because the total compatibility did not meet the no-imprint threshold. This behavior illustrates the intended distinction between a useful hypothesis and an unsafe assertion.

The current case demonstration does not establish accuracy or clinical utility. The principal contribution is the transparent architecture and the definition of a validation pathway. Expansion to representative datasets, open-set evaluation, calibrated scores, ground-truth segmentation, device-shift testing, and professional confirmation will be essential before the approach can support real medication-safety workflows.

ACKNOWLEDGMENT

The authors acknowledge the institutional environments that support the development of interdisciplinary projects in biomedical engineering, computer vision, and pharmaceutical safety.

REFERENCES

[1] World Health Organization, *Global Burden of Preventable Medication-Related Harm in Health Care: A Systematic Review*. Geneva, Switzerland: WHO, 2024, ISBN 978-92-4-008888-7.

[2] World Health Organization, *Medication Without Harm: Policy Brief*. Geneva, Switzerland: WHO, 2024, ISBN 978-92-4-006276-4.

[3] Agencia Nacional de Vigilancia Sanitaria, "Bulas, rotulos e nome comercial," Anvisa, Brasilia, Brazil. [Online]. Available: https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/bulas-e-rotulos. Accessed: Aug. 24, 2026.

[4] Agencia Nacional de Vigilancia Sanitaria, "Sistema de Consultas," Anvisa, Brasilia, Brazil. [Online]. Available: https://www.gov.br/anvisa/pt-br/sistemas/sistema-de-consultas. Accessed: Aug. 24, 2026.

[5] U.S. Food and Drug Administration, "Imprinting of solid oral dosage form drug products for human use," *Code of Federal Regulations*, Title 21, Part 206, current electronic edition.

[6] L. N. T. My, V.-T. Le, T. Vo, and V. T. Hoang, "A comprehensive review of pill image recognition," *Comput. Mater. Contin.*, vol. 82, no. 3, pp. 3693–3740, 2025, doi: 10.32604/cmc.2025.060793.

[7] Y. F. Wong, H. T. Ng, K. Y. Leung, K. Y. Chan, S. Y. Chan, and C. C. Loy, "Development of fine-grained pill identification algorithm using deep convolutional network," *J. Biomed. Inform.*, vol. 74, pp. 130–136, 2017, doi: 10.1016/j.jbi.2017.09.005.

[8] X. Zeng, K. Cao, and M. Zhang, "MobileDeepPill: A small-footprint mobile deep learning system for recognizing unconstrained pill images," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2017, pp. 56–67, doi: 10.1145/3081333.3081336.

[9] N. Larios Delgado, N. Usuyama, A. K. Hall, R. J. Hazen, M. Ma, S. Sahu, C. Lundin, and J. Lundin, "Fast and accurate medication identification," *npj Digit. Med.*, vol. 2, art. no. 10, 2019, doi: 10.1038/s41746-019-0086-0.

[10] N. Usuyama, N. Larios Delgado, A. K. Hall, and J. Lundin, "ePillID dataset: A low-shot fine-grained benchmark for pill identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 910–911.

[11] H.-W. Ting, S.-L. Chung, C.-F. Chen, H.-Y. Chiu, and Y.-W. Hsieh, "A drug identification model developed using deep learning technologies: Experience of a medical center in Taiwan," *BMC Health Serv. Res.*, vol. 20, art. no. 312, 2020, doi: 10.1186/s12913-020-05166-w.

[12] C. A. Lester, J. Li, Y. Ding, B. Rowell, J. Yang, and R. A. Kontar, "Performance evaluation of a prescription medication image classification model: An observational cohort," *npj Digit. Med.*, vol. 4, art. no. 118, 2021, doi: 10.1038/s41746-021-00483-8.

[13] J. Heo, Y. Kang, S. Lee, D.-H. Jeong, and K.-M. Kim, "An accurate deep learning-based system for automatic pill identification: Model development and validation," *J. Med. Internet Res.*, vol. 25, e41043, 2023, doi: 10.2196/41043.

[14] K. Al-Hussaeni, I. Karamitsos, E. Adewumi, and R. M. Amawi, "CNN-based pill image recognition for retrieval systems," *Appl. Sci.*, vol. 13, no. 8, art. no. 5050, 2023, doi: 10.3390/app13085050.

[15] W. I. Chu, S. Hirani, G. Tarroni, and L. Li, "A dataset of medication images with instance segmentation masks for preventing adverse drug events," arXiv:2603.10825, 2026, doi: 10.48550/arXiv.2603.10825.

[16] W. I. Chu, G. Tarroni, and L. Li, "Evaluating few-shot pill recognition under visual domain shift," arXiv:2603.10833, 2026, doi: 10.48550/arXiv.2603.10833.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[18] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004, doi: 10.1145/1015706.1015720.

[19] S. Garrido-Jurado, R. Munoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marin-Jimenez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognit.*, vol. 47, no. 6, pp. 2280–2292, 2014, doi: 10.1016/j.patcog.2014.01.005.

[20] U.S. Food and Drug Administration, *Size, Shape, and Other Physical Attributes of Generic Tablets and Capsules: Guidance for Industry*. Silver Spring, MD, USA: FDA, 2015.

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1321–1330.

[22] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva, Switzerland: WHO, 2021, ISBN 978-92-4-002920-0.

Evaluation of the use of assistive cutlery in an institutional setting: a comparison between perceptions of users with Parkinson's disease and multiple disabilities

1st Kelly Cristine Milanez
Federal University of Technology – Paraná
Curitiba, PR, Brazil
0009-0008-8612-1888

2nd Giullia Valentini Dancini de Oliveira
Colégio Bom Jesus
PIBIC-EM/CNPq
Curitiba, PR, Brazil
giulliaoliveira2009@gmail.com

3rd Henrique Cunha Carvalho
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná
Campo Mourão, PR, Brazil
0000-0003-0806-4132

4th Cristhiane Gonçalves
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná
Curitiba, PR, Brazil
0000-0002-3894-2971

5th Frieda Saicla Barros
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná
Curitiba, PR, Brazil
0000-0002-1645-1111

***Abstract*—Independent feeding is a fundamental activity of daily living associated with autonomy, dignity, and social participation. Motor and cognitive impairments may affect utensil handling and functional performance during meals, particularly in individuals with multiple disabilities. This study aimed to evaluate the interaction with assistive cutlery among institutionalized adults with multiple motor and cognitive disabilities during real-life feeding situations and to discuss the findings in relation to previous literature. A qualitative, descriptive, and observational study was conducted with 16 adults selected from an initial group of 30 individuals based on their ability to use the spoon independently, sufficiently understand the proposed activity, and agree to participate. Two assistive spoon models, with stainless-steel and polypropylene handles, were evaluated during usual meals. Observations included grasp patterns, food spillage, need for assistance, adaptation, sensory aspects, and material preferences. Most participants accepted using both assistive spoon models and showed different grasp and adaptation strategies. Some participants appeared to show better food control, although this was not quantitatively measured. Fourteen of the 16 participants (87.5%) preferred the polypropylene-handled spoon, mainly due to thermal and tactile comfort and familiarity. The findings highlight the importance of considering motor, cognitive, sensory, and contextual factors when evaluating assistive technologies under real-life conditions.**

***Index Terms*—Assistive technology, feeding, multiple disabilities, rehabilitation, user-centered design.**

I. Introduction

The ability to eat independently plays an important role in maintaining autonomy, dignity, and social participation in daily life. For individuals with motor impairments, tremor, rigidity, and coordination difficulties may affect utensil handling and compromise performance during meals.

In individuals with Parkinson's disease, Pagnussat et al. [1] identified difficulties related to food spillage and grasp instability, as well as preferences for specific utensil characteristics, including texture, weight, and shape. Other studies have proposed different solutions to support feeding, including intelligent systems and utensils designed for movement stabilization and tremor compensation [2-6].

Despite these advances, studies conducted during real-life feeding situations involving individuals with multiple disabilities remain less frequent. In this population, motor and cognitive impairments may coexist and influence utensil handling, acceptance, and adaptation. The institutional setting also presents specific characteristics related to feeding routines, different levels of assistance, and everyday interaction with support resources.

In this context, observing the use of assistive cutlery in the usual feeding environment may complement assessments performed under controlled conditions or based only on users' perceptions. In addition to performance during handling, grasp patterns, need for assistance, material preferences, and individual responses may contribute to understanding the interaction between the user and the product.

II. Methodology

This qualitative, descriptive, and observational study was conducted at the Pequeno Cotolengo Health Complex, a

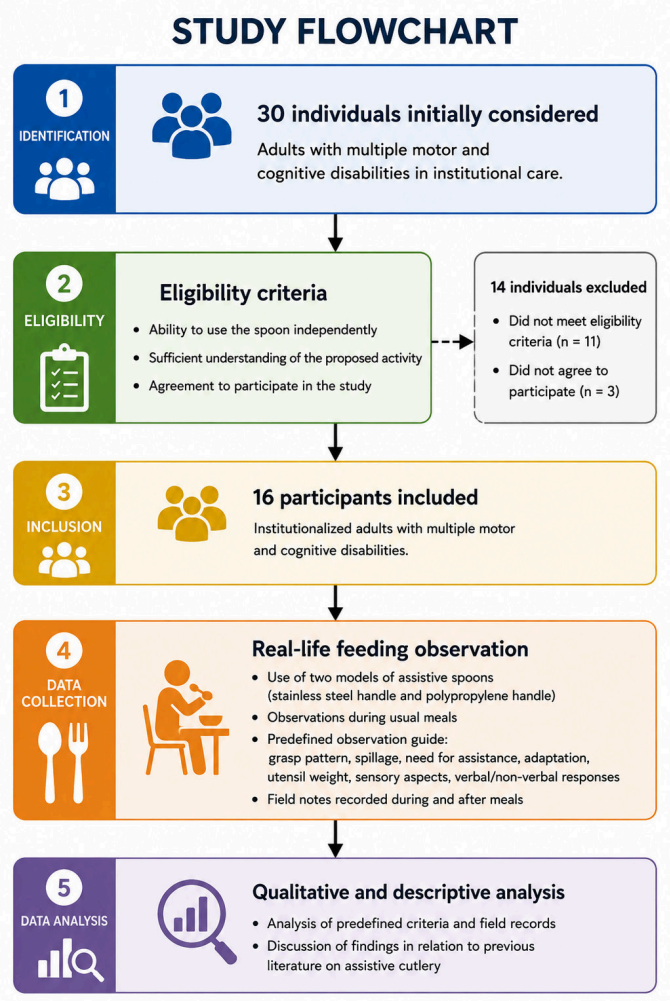


Fig. 1. Study flowchart.

long-term care institution for individuals with multiple disabilities. The study evaluated participants' interaction with assistive cutlery during real-life feeding situations in their usual environment. The study was approved by the Research Ethics Committee involving human subjects under protocol no. 7,752,165 and CAAE no. 90038025.2.0000.0177.

A. Participants and Procedures

Initially, 30 institutionalized adults with multiple motor and cognitive disabilities and different levels of functional impairment were considered for participation. The final sample included 16 participants who were able to use the spoon independently, sufficiently understand the proposed activity, and agreed to participate. The participants presented heterogeneous motor and cognitive conditions, including different degrees of upper-limb impairment, altered grasp patterns and, in some cases, dysphagia. Verbal communication was not required for participation. Given the heterogeneity of the sample, the study did not aim to establish associations between the findings and specific clinical diagnoses.



Fig. 2. Assistive spoon models evaluated in the study.

Two models of assistive spoons developed by a Brazilian company were evaluated: one made entirely of stainless steel with an adapted handle, and another consisting of a stainless-steel bowl and a polypropylene handle (Fig. 2). The spoon was selected because it can be used with different food consistencies commonly present in the institutional routine and provides greater food containment during transport to the mouth.

The utensils were presented as a new assistive cutlery option, without formal training for their use. Observations were conducted during usual meals in the participants' everyday environment and were performed by the same researcher. The use of a single observer resulted from the operational conditions of the study, in which including and training external observers was not feasible. Maintaining the same researcher allowed the same observation criteria to be applied throughout data collection, although it does not eliminate the subjectivity inherent to direct observation and prevented the assessment of inter-observer agreement.

A predefined observation guide included grasp pattern, preference between handle materials, food spillage and possible changes compared with usual feeding behavior, need for assistance, perception of utensil weight and, for participants able to communicate verbally, impressions and suggestions for product modifications. The main observational domains are summarized in Fig. 3.

Notes were recorded during the observations and completed immediately after each meal. Spontaneous manifestations were also documented, including individual adaptation strategies, behavioral and non-verbal responses, and preferences not initially included in the observation guide. Some observations were subsequently discussed with members of the multidisciplinary team and caregivers familiar with the participants' routines, helping to contextualize their usual feeding behavior.

B. Data Analysis

Data were analyzed qualitatively and descriptively, considering both the predefined observation criteria and aspects identified in the field records. The analysis focused on grasp patterns, food spillage, need for assistance, adaptation to motor

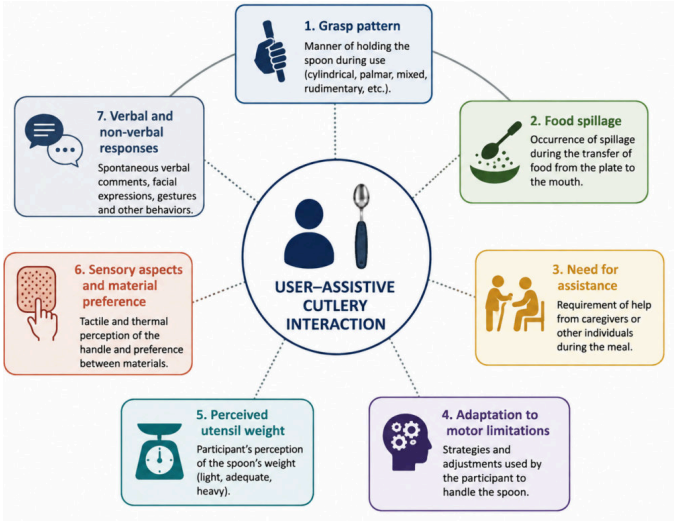


Fig. 3. User–assistive cutlery interaction model and main observational domains.

limitations, perception of utensil weight, material preferences, and individual responses during use.

The findings were discussed in relation to previous literature on assistive cutlery, including studies involving individuals with Parkinson's disease. However, differences in diagnosis, motor and cognitive impairment, functional independence, and context of use were considered in the interpretation. Therefore, the literature was used to identify points of convergence and contrast rather than as a direct comparison between equivalent populations.

III. Results and Discussion

Overall, most participants accepted using both assistive spoon models and showed different adaptation strategies during use. The heterogeneity of the sample was particularly evident in grasp patterns, sensory responses, and the way each participant interacted with the utensils during usual meals. The main observational findings and their relationship with aspects previously reported by Pagnussat et al. [1] are summarized in Table I.

Regarding food spillage, some participants who usually experienced food loss while moving the spoon from the plate to the mouth appeared to show better food control when using the assistive spoons. However, spillage episodes were not quantified, and no instruments were used to measure tremor or motor stability. Therefore, these findings should be interpreted as functional changes observed under real-life conditions and do not allow an objective reduction in food spillage or improvement in motor stability to be established. Previous studies have used different strategies to evaluate utensils designed to compensate for tremor during feeding [2]–[4], reinforcing the importance of objective measures in future assessments.

Variability in grasp patterns was also relevant. Participants with upper-limb impairments or deformities used different strategies to handle the spoons, including cylindrical and more

TABLE I
Observational Findings in Relation to Previous Literature

Aspect	**Pagnussat et al. [1]**	**Present Study**
Food spillage	Reported as a difficulty by users	Apparent improvement in food control in some participants; not quantitatively measured
Grasp	Difficulties in utensil handling	Different grasp and adaptation strategies, including cylindrical and rudimentary patterns
Motor aspects	Tremor affected utensil handling	Different motor limitations were observed; tremor and stability were not objectively measured
Assistance	Not the main focus	Generally similar to the assistance usually required during meals
Design and material	Texture, weight, size, and depth influenced preferences	14 of 16 participants (87.5%) preferred the polypropylene handle; thermal and tactile comfort and familiarity were relevant
User profile	Individuals with Parkinson's disease	Institutionalized adults with multiple motor and cognitive disabilities
Dysphagia	Not the study focus	Present in part of the sample; not evaluated as an outcome

rudimentary grasp patterns. These adaptations indicate that utensil use does not depend exclusively on a conventional grasp and reinforce the importance of considering individual characteristics when developing assistive technologies for neurologically heterogeneous populations [5].

Most participants maintained a level of assistance similar to that usually required during meals. Since this aspect was not assessed using a standardized scale, a quantitative gain in autonomy resulting from utensil use cannot be established.

Fourteen of the 16 participants (87.5%) preferred the spoon with the polypropylene handle. Perceived weight, handle shape, and ease of grasping may also have contributed to this preference. Pagnussat et al. [1] identified texture, size, weight, and utensil depth as relevant characteristics for individuals with Parkinson's disease, indicating that user experience is influenced by both functional and sensory characteristics of the product.

Cognitive factors also appeared to influence interaction. Although most participants accepted trying the utensils, one participant showed lower adherence, which, within the observed context, was associated with difficulty accepting changes in the usual feeding routine. Since no specific cognitive assessment instruments were applied, this relationship should be interpreted as a contextual observation rather than a demonstrated association.

The presence of dysphagia in some participants should also be considered. Feeding performance in this population involves factors beyond utensil manipulation, including eating pace, food consistency, swallowing safety, and need for super-

vision. Therefore, changes observed in grasp or food transport should not be interpreted as improvements in swallowing safety.

Comparison with Pagnussat et al. [1] requires caution because the two studies involved populations with different clinical and functional characteristics. While their study focused on individuals with Parkinson's disease, the present sample consisted of institutionalized adults with multiple motor and cognitive disabilities. The previous findings were therefore used to identify points of convergence and contrast rather than to establish direct equivalence between the populations.

Recent studies also illustrate the value of combining user perceptions with objective measures. Rosa et al. [6], for example, evaluated assistive cutlery using structured satisfaction assessment, video recordings, and kinematic analysis. In contrast, the present study prioritized spontaneous interaction under real-life institutional conditions. These approaches may be complementary: direct observation allowed individual adaptation strategies, non-verbal responses, and preferences to be identified, while standardized instruments and quantitative measures may improve the objectivity and comparability of future assessments.

A. Limitations, Lessons Learned, and Research Gaps

The main limitations of this study include the small and heterogeneous sample from a single institution, the use of a single observer, and the absence of standardized instruments and quantitative performance measures. Although maintaining the same observer provided a consistent observational perspective throughout data collection, inter-observer agreement could not be assessed. Meal duration, number of spillage episodes, grasp efficiency, and perceived effort were also not quantified. These limitations prevent conclusions regarding the clinical effectiveness of the utensils and restrict the generalizability of the findings.

Despite these limitations, the study provided relevant lessons from real-life use. Interaction with the utensils appeared to involve not only motor limitations but also cognitive and sensory factors, familiarity with the materials, and individual grasp strategies. Direct observation was particularly useful for participants with limited verbal communication, since spontaneous choices, behavioral responses, and individual adaptations complemented the predefined observation guide.

Future studies should include larger and multicenter samples, characterization according to motor and cognitive profiles, validated usability and satisfaction instruments, and quantitative performance measures. Longitudinal assessments and comparisons with other assistive cutlery models may also help determine whether acceptance and preferences are maintained over time. Combining real-life observation with objective measures, as demonstrated in recent assistive cutlery research [6], may improve the comparability of findings and support the development of solutions better suited to users' needs.

IV. Conclusion

This study evaluated the interaction of institutionalized adults with multiple motor and cognitive disabilities with two models of assistive spoons under real-life feeding conditions. Participants showed different grasp and adaptation strategies, and most accepted using both assistive spoon models. Fourteen of the 16 participants (87.5%) preferred the polypropylene-handled model, mainly in relation to thermal and tactile comfort and familiarity with the material.

The findings indicate that motor, cognitive, sensory, and contextual factors should be considered when evaluating and developing assistive technologies for feeding. Due to the qualitative and exploratory nature of the study, the results do not establish the clinical effectiveness of the utensils but provide information from real-life use that may support future studies using standardized instruments and objective performance measures.

Acknowledgment

The authors would like to thank the Universidade Tecnológica Federal do Paraná (UTFPR – Curitiba) for providing the resources necessary for the development of this project. The authors also acknowledge the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for supporting the Junior Scientific Initiation Scholarship (PIBIC-EM), and Fundação Araucária and the Superintendência de Ciência, Tecnologia e Ensino Superior (SETI) for providing conference grants to H. C. C. and F. S. B. The assistive cutlery used in this study was donated by Tramontina S.A.

References

[1] A. S. Pagnussat, A. S. do Pinho, C. Pinto, T. C. da Rosa, M. Moscovich, C. de Sousa Andrade, and Y. A. Chen, "How do people with Parkinson's disease perceive challenges in handling cutlery?—A mixed study," *Disability and Rehabilitation: Assistive Technology*, vol. 20, no. 5, pp. 1453–1459, 2025, doi: 10.1080/17483107.2025.2459324.

[2] F. Talaei and S. M. Kargar, "Design and fabrication of a device for reducing hand tremor in Parkinson patients during eating," *Journal of Medical Signals & Sensors*, vol. 13, no. 1, pp. 21–28, 2023, doi: 10.4103/jmss.jmss_116_21.

[3] P. A. Zartika, M. A. Muslim, and E. Yudaningtyas, "Spoon stabilization for essential tremor patients using PID control optimized by PSO," *KINETIK*, vol. 10, no. 4, pp. 443–454, 2025, doi: 10.22219/kinetik.v10i4.2272.

[4] M. Dubé, T. Laliberté, V. H. Flamand, and A. Campeau-Lecours, "Development of a new and mechanically intelligent anti-tremor utensil," *Applied Engineering*, vol. 8, no. 1, pp. 47–60, 2024, doi: 10.11648/j.ae.20240801.15.

[5] M. Bonanno, B. Saracino, I. Ciancarelli, G. Panza, A. Manuli, G. Morone, and R. S. Calabrò, "Assistive technologies for individuals with a disability from a neurological condition: A narrative review on the multimodal integration," *Healthcare*, vol. 13, no. 13, Art. no. 1580, 2025, doi: 10.3390/healthcare13131580.

[6] T. C. da Rosa *et al.*, "Usability of assistive cutlery for people with Parkinson's disease: A user-centred design prototype," *Disability and Rehabilitation: Assistive Technology*, pp. 1–10, 2026, doi: 10.1080/17483107.2026.2621201.

Assistive System for Signed Alphanumeric Spelling: a communicative mediation

1st Victor de Andrade
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
victor.andr12@gmail.com

2nd Lucas C. Carmona
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
carmonalucas98@gmail.com

3rd Gustavo B. Pires
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
gubrait@hotmail.com

4thMaria Claudia F. Castro
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
mclaudia@fei.edu.br

***Abstract*—Brazilian Sign Language (LIBRAS) is a vital means of communication for the Deaf community, with signed alphanumeric spelling serving as a fundamental resource for spelling out proper names, acronyms, and specific terms. However, few accessible assistive technology solutions are designed to physically reproduce the alphanumeric characters used in this context. This paper presents the development of an assistive upper-limb system for signed alphanumeric spelling, based on a 3D-printed robotic hand controlled by an ESP32 microcontroller and operated via a mobile app over Bluetooth. The system was developed using the open-source DexHand V1.0 platform, which was adapted to simplify the mechanics, reduce costs, and improve reproducibility. The results demonstrated the feasibility of reproducing signed alphanumeric spelling, highlighting the solution's potential as an assistive, educational, and research tool.**

***Index Terms*—Brazilian Sign Language - LIBRAS, assistive robotics, signed alphanumeric spelling, 3D printing, embedded systems.**

I. Introduction

Communication is a key element for social inclusion and access to knowledge. In this context, LIBRAS plays a fundamental role for the Brazilian Deaf community, as it is legally recognized as a means of communication and expression. Among its linguistic features, signed alphanumeric spelling stands out; it represents letters and numbers through specific hand configurations [1, 2].

LIBRAS is a full natural language rather than a simple gestural code: linguistic descriptions identify five parameters that combine to form each sign, namely hand configuration, place of articulation, movement, palm orientation, and non-manual markers such as facial expressions, all of which contribute to meaning [2]. Signed alphanumeric spelling relies mainly on hand configuration and is used chiefly to spell proper names, acronyms, and terms without a conventional sign [1]. Accordingly, the system presented here does not aim to replace LIBRAS communication or to reproduce the language in full; it targets a delimited, hand-configuration-based subset of the language, intended as a technological resource for teaching, demonstration, and research on signed alphanumeric spelling.

Visual aids, repetition, and imitation activities play a central role in instruction, helping the learner perceive hand configuration, articulation point, movement, and palm orientation—elements whose correct articulation defines the meaning of the sign [3, 4]. From a pedagogical standpoint, this approximates a multimodal reinforcement approach: the student observes, imitates, corrects, and repeats until the gesture is stabilized—a process particularly useful for signs that differ only by slight formal variations.

Although digital applications for teaching LIBRAS exist, solutions that can physically reproduce the gestures associated with signed alphanumeric spelling remain limited.

Tantum 1 [5] is an assistive robotic hand developed to support tactile sign interpretation for deafblind individuals. [6] details the specifications of the initial Tantum 1 prototype, featuring 15 servomotors controlled by an Arduino Mega microcontroller, along with a cloud service that converted text into movement instructions for the servomotors to reproduce American Sign Language signs.

The Sign-Representing Robotic Hand (MAR2S), developed by [7], is based on the open-source InMoov model and utilizes five servomotors controlled by a Raspberry Pi 3 Model B microcontroller.

Given this context, this study proposes developing a robotic assistive system capable of reproducing alphanumeric characters using the manual alphabet of LIBRAS, leveraging accessible technologies, an open architecture, and low implementation costs.

II. Materials and Methods

The developed system integrates three main subsystems: (1) a 3D-printed five-finger robotic upper limb based on the open-source DexHand V1.0 platform [8]; (2) an embedded control system based on the ESP32-WROOM-32D microcontroller with Bluetooth Low Energy (BLE); and (3) an iOS mobile

application developed in Flutter that communicates wirelessly with the microcontroller via the GATT protocol. Control flows from the mobile application to the microcontroller over BLE, which then drives all servomotors. Fig. 1 shows the system diagram, whereas Fig. 2 details system architecture.

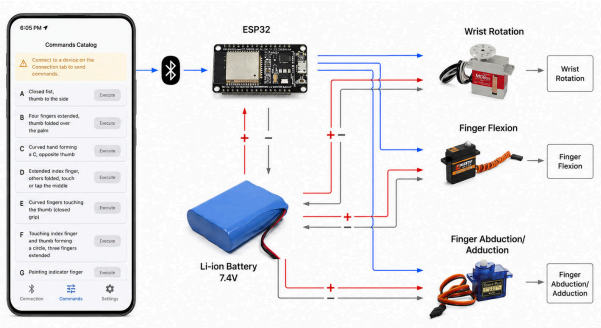


Fig. 1. System diagram.

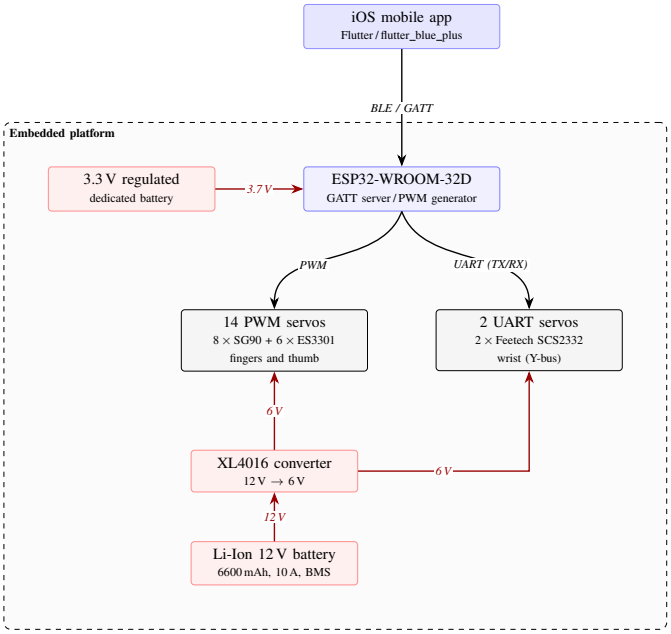


Fig. 2. System architecture.

A. Mechanical Platform: DexHand V1.0

The DexHand V1.0 [8] is an open-source, 3D-printed robotic hand employing up to 19 servomotors, nylon strings, and return springs distributed across five independent fingers and a wrist. Each main finger—index, middle, ring, and little—is driven by three servos: two TowerPro SG90 mounted at the metacarpal level for abduction/adduction, actuating a nylon string through a bearing in the proximal phalanx; and one Emax ES3301 housed in the forearm for flexion, via a string through internal channels across all three phalanges—extension is recovered by return springs. The thumb attaches to a support piece at the base of the middle finger and uses four ES3301s in the forearm for adduction, abduction, opposition, and flexion. Two Feetech SCS2332 in the forearm drive wrist flexion/extension and abduction/adduction. The overall mechanical structure is illustrated in Fig. 3 and the specifications and comparison of the three servo models are shown in Table I.

The body parts are fabricated using additive manufacturing (3D printing) with polylactic acid (PLA), the most common material for low-cost functional prototypes. Compared to acrylonitrile butadiene styrene (ABS), PLA offers better dimensional accuracy due to its lower coefficient of thermal expansion, which is critical for the tight tolerances required by finger joints and servo mounting cavities.

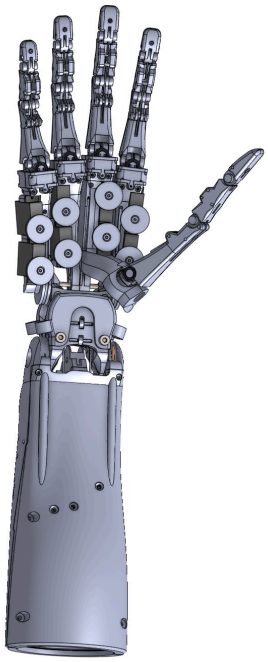


Fig. 3. CAD model: DexHand V 1.0 [8].

TABLE I
SERVO MOTOR SPECIFICATIONS.

Feature	SG90	ES3301	SCS2332
Type	Analog	Analog	Smart digital
Control	PWM	PWM	Serial
Torque	1.8kg·cm	2.2kg·cm	4.5kg·cm
Weight	9g	10.6g	20g
Gears	Plastic	Eng. plastic	Metal
Range	180°	180°	300°

B. Electronic Hardware

The hardware is based on the ESP32-WROOM-32D microcontroller, which simultaneously manages all actuators. Table II lists the main system components.

The servo power supply uses a 12V Li-Ion battery pack, stepped down to 6V by the XL4016 DC-DC converter, and is compatible with all three servo types. The ESP32 is

TABLE II
ELECTRONIC SYSTEM COMPONENTS.

Component	Qty.
ESP32-WROOM-32D microcontroller	1
TowerPro SG90	8
Emax ES3301	6
Feetech SCS2332	2
Li-Ion battery 12V, 6600mAh	1
XL4016 DC-DC step-down (12V → 6V)	1
3.7V battery (ESP32)	1
Custom PCB board	1
1 kΩ UART conditioning resistor	1
4700 μF electrolytic capacitor	1
100 nF electrolytic capacitor	5
AWG20 power wiring / AWG24 signal wiring	–

powered by a dedicated 3.7V rechargeable battery through its internal 3.3V regulator, isolating the logic supply from servo current transients. Of the 16 servomotors, 14 (8×SG90 and 6×ES3301) are controlled individually via PWM signals generated by ESP32 GPIO pins. The two Feetech SCS2332 wrist servos use half-duplex UART communication (GPIO16 RX, GPIO17 TX) on a shared Y-bus; each servo has a unique internal ID allowing independent addressing on the same bus. A 1kΩ resistor on the UART line limits current and improves signal integrity. A 4700μF bulk capacitor at the power input and five 100nF local capacitors distributed near the servo connectors reduce voltage drops and noise during simultaneous actuation.

Two changes distinguish this implementation from the original DexHand V1.0. First, motion commands in the original design are issued either by a radio receiver or by a computer running computer-vision hand tracking, relayed over serial to an Arduino Mega 2560 Pro; in this work, that microcontroller and its external vision/radio peripherals were replaced by a single ESP32-WROOM-32D, which provides native BLE connectivity and is driven directly by the mobile application (Section II-C). Second, the two ES3301 servos originally assigned to thumb abduction/adduction were removed: this motion can be reproduced through the combination of flexion, extension, and opposition without loss of the alphanumeric signs implemented, reducing the active actuator count from 19 to 16 while the 18-channel command protocol retains two reserved channels for compatibility with the complete architecture.

C. Firmware

The firmware (C++, Arduino framework) configures the ESP32 as a BLE GATT peripheral with a control service and a write characteristic, both of which share fixed UUIDs with the mobile application. The central concept of the command protocol is an 18-element array of joint angles (in degrees) that fully encodes a hand configuration. The array is structured in fixed groups: indices 0–11 encode the four fingers in triplets (pitch, yaw, and tip flexion for index, middle, ring, and little, respectively); indices 12–15 encode the thumb (pitch, yaw, tip, roll); indices 16–17 encode the wrist (pitch, yaw). This ordering is summarized in Table III.

TABLE III
DATA PACKET STRUCTURE AND ACTUATOR MAPPING.

Bytes	Group / DOF	Servos
0–2	Index: pitch, yaw, tip	2 SG90 + 1 ES3301
3–5	Middle: pitch, yaw, tip	2 SG90 + 1 ES3301
6–8	Ring: pitch, yaw, tip	2 SG90 + 1 ES3301
9–11	Little: pitch, yaw, tip	2 SG90 + 1 ES3301
12–15	Thumb: pitch, yaw, tip, roll	4 ES3301 (2 active)
16–17	Wrist: pitch, yaw	2 SCS2332
18	Checksum (mod-256 sum)	—

Table IV shows the configuration vectors for two sign examples. In vector **A**, all four finger triplets share `pitch = 25°` and `tip = 180°` (fully curled); index 12 sets thumb `pitch = 90°` (vertically extended); indices 16–17 keep the wrist neutral. Letter **B** contrasts with all fingers extended (`tip = 20°`) and the thumb repositioned (`yaw = 120°`, `tip = 100°`). Defining or modifying a sign requires only editing its array—no firmware changes are needed.

TABLE IV
CONFIGURATION VECTORS FOR TWO SIGN EXAMPLES.

Sign	18-angle vector (degrees)
A	`[25, 0, 180, 25, 0, 180, 25, 0, 180, 25, 0, 180, 90, 0, 0, 0, 0, 0]`
B	`[0, 0, 20, 0, 0, 20, 0, 0, 20, 0, 0, 20, 90, 120, 100, 0, 0, 0]`

Before transmission, each angle is encoded into a single byte to comply with the BLE MTU limit (Eq. 1), and a checksum byte is appended, resulting in a 19-byte BLE packet. Upon reception, the firmware validates the length and checksum, then decodes each byte.

The value 127 maps to 0°, enabling positive and negative angles. The decoded array is then distributed to `Finger`, `Thumb`, and `Wrist` objects, which apply per-joint angular limits—preventing out-of-range commands from mechanically overloading the structure—and direction correction before translating the target angles into PWM pulses (14 SG90/ES3301 servos) or UART position commands (2 SCS2332 wrist servos) via `updateHand()`. On disconnection or timeout, all joints return to the neutral reference posture, and BLE advertising resumes.

$$b = \mathrm{clamp}\left(\mathrm{round}\left(\theta \cdot \frac{256}{360} + 127\right), 0, 255\right) \qquad (1)$$

D. Command Catalog

Static signs use a single 18-element array, whereas dynamic signs (H, J, K, X, Z) add a wrist waypoint sequence (pitch, yaw, dwell time) transmitted as successive packets after the handshape is set. The mobile application, developed in Flutter for iOS using the *flutter_blue_plus* BLE library, provides three

tabs: (1) connection management (BLE scan, device selection, status); (2) command catalog—36 signs (26 letters + digits 0–9) with one-tap execution, disabled when disconnected; and (3) settings (pairing, version info). Commands are dispatched as a timed packet sequence—fingers first, then wrist and thumb—to distribute the electrical demand.

III. Results and Discussion

A. Built Prototype

The constructed prototype (Fig. 4) integrates 16 servomotors (8 SG90, 6 ES3301, 2 SCS2332) within a PLA forearm shell and is controlled via BLE from an iOS device. The 18-channel protocol retains two reserved channels for the removed thumb servos, preserving future compatibility with the complete DexHand architecture. Electrical characterization at 6.6V revealed a resting draw of ≈300mA, ≈600mA during single-finger movements, and peaks of ≈6A under simultaneous multi-servo actuation—confirming the relevance of the bulk decoupling capacitor and the sequential actuation strategy. Given the 6600mAh servo battery capacity, these currents place the estimated autonomy between ≈11h of continuous single-finger actuation (≈600mA) and ≈22h of idle operation at the neutral posture (≈300mA); the 6A peaks are transient and therefore do not represent a sustained draw, but they set the instantaneous demand that the power stage must support.

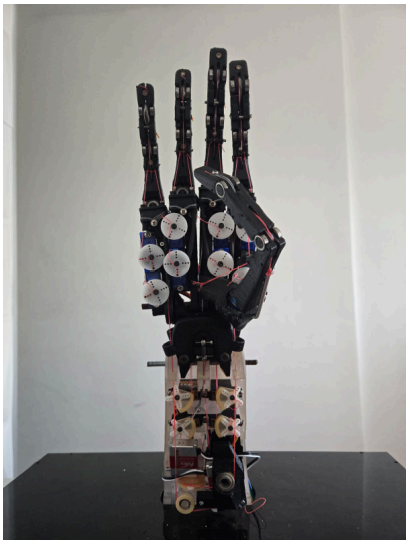


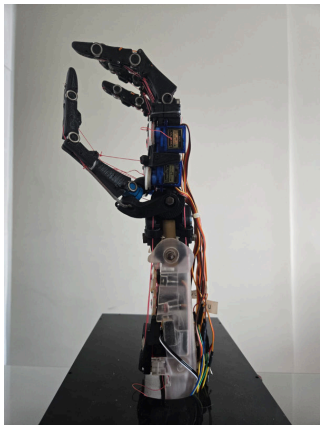
Fig. 4. Prototype developed for LIBRAS alphanumeric character reproduction.

B. Implemented Signed Alphanumeric Spelling

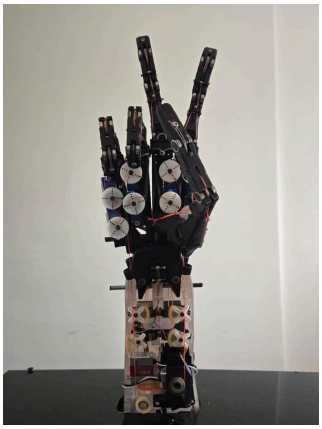
All 36 signed alphanumeric spelling (26 letters + digits 0–9) were implemented and evaluated by visual inspection against the expected manual alphabet configuration. Table V summarizes the results; Fig. 5 illustrates two examples—letter C (arc formed by all fingers with the thumb opposing) and letter V (index and middle extended and separated)—demonstrating the system's ability to produce distinct hand configurations.

TABLE V
Signed alphanumeric spelling reproduction quality.

Quality	Signs
Excellent	A, B, C, E, O, U, V; digits 1–4
Good	D, F, G, I, L, P, R, S, T, W, Y; digit 0; digits 5–9; H, J, K, X, Z
Limited	M, N, Q (require forearm rotation)



(a) Letter C



(b) Letter V

Fig. 5. Examples of signed alphanumeric spelling reproduction.

Dynamic signs (H, J, K, X, Z) were implemented via the multi-waypoint wrist-trajectory approach, yielding recognizable letter strokes within 0.5–1.2 s after handshape stabilization. The checksum mechanism ensured communication reliability: every command applied to the hardware matched exactly what the application sent, so all observed deviations between intended and reproduced postures are attributable to mechanical factors.

C. Mobile Application

The application (Fig. 6) performed reliably across all testing sessions: BLE scanning and connection were stable, and the command catalog provided immediate one-tap access to all 36 signs while correctly disabling execution when disconnected. Static postures are assumed in 2–4 s; the sequential packet dispatch (fingers → wrist → thumb, with configurable inter-step delays) attenuated current peaks and maintained system stability. All timing parameters are software-defined, enabling speed adjustment without hardware modifications.

D. Limitations

Primary constraints are mechanical: nylon-string friction limits effective torque delivery to the finger joints, and string breakage occurred recurrently in the little finger, ring finger, and thumb during testing. Progressive loosening of string tension requires session-by-session recalibration, compromising repeatability. The SG90 servos' limited torque amplifies these effects. Signs M, N, and Q additionally require forearm pronation, which the static forearm shell cannot reproduce: because the forearm is fixed, the fingertips would instead face

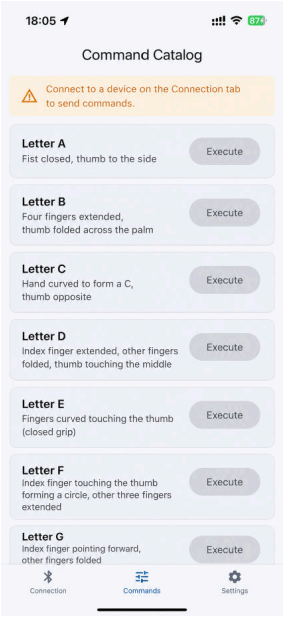


Fig. 6. Mobile application — command catalog tab.

downward, so these three signs could only be approximated in a static posture. Weight and dimensions, while compatible with the demonstrator scope, reduce portability for prolonged interaction scenarios. Future work should address string routing, lower-friction guides, higher-tensile strings, and user studies with the target population.

IV. Conclusion

This work presented the development of an upper-limb assistive system for communicational education, based on a low-cost, 3D-printed robotic platform. The integration of the mechanical structure, embedded system, and mobile application enabled the reproduction of signs corresponding to signed alphanumeric spelling, demonstrating the technical feasibility of the proposal.

The results indicate that open-architecture assistive devices can contribute to future research in accessibility, rehabilitation, and assistive communication, expanding the possibilities for interaction between users and robotic systems.

Beyond its technological contribution, the project highlights the potential of assistive technologies applied to communication and social inclusion. The concepts regarding LIBRAS and human-machine interfaces provide a foundation for future system enhancements, including myoelectric control, movement customization, and an expanded sign repertoire.

References

[1] Fernando César Capovilla and Walkíria Duarte Raphael. *Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira*. São Paulo: EdUSP, 2001.

[2] Ronice Müller de Quadros and Lodenir Becker Karnopp. *Língua de sinais brasileira: estudos linguísticos*. Porto Alegre: Artmed, 2004.

[3] IFMT. *Fundamentos da Língua Brasileira de Sinais*. Cuiabá: IFMT – Instituto Federal de Mato Grosso, 2019.

[4] Ygor Corrêa and Carina Rebello Cruz. *Língua Brasileira de Sinais e Tecnologias Digitais*. Porto Alegre: Penso, 2019.

[5] Tatum Robotics. *Tatum Robotics*. https://tatumrobotics.com/. Acesso em: 1 out. 2025. 2025.

[6] S. T. Johnson et al. "An adaptive, affordable, open-source robotic hand for deaf and deaf-blind communication using tactile American Sign Language". In: *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Guadalajara: IEEE, 2021. DOI: 10.1109/EMBC46164.2021.9630332.

[7] Everson de Souza Silva and Nanau Lopes da Silva. "MAR2S: mão robótica representativa de sinais para auxiliar no aprendizado de LIBRAS". Acesso em: 19 out. 2025. Trabalho de Conclusão de Curso (Bacharelado em Engenharia da Computação). Curitiba: Universidade Tecnológica Federal do Paraná, 2022. URL: https://repositorio.utfpr.edu.br/jspui/handle/1/28884.

[8] DexHand. *DexHand V1.0 — Open-source Dexterous Robotic Hand*. https://www.dexhand.org/. Accessed: Oct. 2025. 2025.

Automatic Transcription of Monophonic Recorder Audio to Braille Music Notation Using Classical Signal Processing Methods

Samuel Sales Furtado
Department of Teleinformatics Engineering
Federal University of Ceará (UFC)
Fortaleza, Brazil
salessamuel@alu.ufc.br

Edilson Rocha Porfírio Filho
Department of Teleinformatics Engineering
Federal University of Ceará (UFC)
Fortaleza, Brazil
edilson.filho@ufc.br

***Abstract*—Braille music notation allows people with visual impairment to read and study music autonomously. However, its production depends almost exclusively on manual transcription or on tools that require an already-digitized score, excluding musicians who learn by ear – such as the recorder students at the Instituto de Cegos do Ceará. This work develops and evaluates a pipeline for automatic transcription of monophonic soprano recorder melodies, from audio directly to braille music notation following the New International Manual of Braille Music Notation (NIMBN) standard, without any supervised training stage, combining pitch detection via pYIN, rhythmic segmentation and quantization based on MIDI discretization, and symbolic conversion to braille cells. Evaluation was conducted on 30 synthetic melodies (613 notes) with ground truth (GT) derived from MIDI files and eight real recorder recordings (329 notes), comparing performance against Basic Pitch, a lightweight neural transcription model, as a baseline. The proposed system achieved an onset F1-score of 94.6% on the synthetic dataset and 93.6% on the real recordings; a calibrated Basic Pitch baseline reached a higher onset F1 (96.8%) but a substantially worse Symbol Error Rate (SER) on the resulting braille output (36.4% vs. 8.6%). These values were obtained without a validation set independent of the test set; the comparison should be read as a measure of fit on the evaluated sets, not of generalization. The braille encoding table was verified cell by cell against the official manual, achieving 100% conformance for notes and rests. For this niche instrument, the results indicate that a classical, signal-driven approach is a viable alternative to a generic neural model when no large-scale annotated data is available, addressing a gap between automatic music transcription and braille music notation identified by a systematic literature review (Section II).**

***Index Terms*—automatic music transcription, braille music notation, recorder, pYIN, accessibility**

I. Introduction

Braille music notation is the tactile music notation system created by Louis Braille as a natural extension of his reading system [1], standardized internationally by the *New International Manual of Braille Music Notation* (NIMBN) [1]. It allows people with visual impairment to read, write and study music autonomously. In Brazil, institutions such as the Instituto Benjamin Constant and the Instituto de Cegos do Ceará (ICC) use braille music notation as their main tool for musical inclusion [2] [3].

Despite its importance, producing braille scores faces a persistent bottleneck: the conversion of music to this format is mostly a manual process, requiring professionals with dual expertise in conventional notation and braille music notation. Available tools such as Goodfeel, BrailleMUSE and Musibraille automate only the final conversion step and require a digitized score in MusicXML as input [4] [5], which presupposes that the music already exists in visual notation, transcribed by a sighted person. For visually impaired musicians who learn by ear or work without digitized scores, this conversion chain never closes.

Automatic Music Transcription (AMT) is the research area that investigates computational methods to convert audio signals into symbolic music representations, such as scores or MIDI files [6]. The field has advanced markedly for polyphonic instruments such as the piano, driven by deep learning architectures [7]. To assess whether an audio-to-braille system already exists, we conducted a systematic literature review across IEEE Xplore, Scopus and ACM Digital Library, combining terms for "automatic music transcription"/"audio-to-score" with terms for the target instruments and for braille music notation. Of 1,020 papers identified, 198 were accepted after title/abstract and full-text screening against explicit inclusion criteria (peer-reviewed AMT systems with a working implementation) and exclusion criteria (surveys without a proposed method, non-audio symbolic-to-symbolic conversion only). None of the 198 accepted papers performs the complete chain from musical audio to braille music notation, and none addresses the recorder specifically. This gap represents a concrete research opportunity with potential impact on the musical accessibility of people with visual impairment.

The contribution of this paper is threefold: (i) a training-free, audio-to-braille transcription pipeline for a niche monophonic instrument, the first, to the authors' knowledge, to close the chain from raw audio to symbolic braille output without an in-

This paper is a condensed, English-language derivation of the first author's undergraduate thesis (Trabalho de Conclusão de Curso), Department of Teleinformatics Engineering, Federal University of Ceará (UFC), 2026, advised by the second author.

termediate digitized-score step; (ii) an end-to-end quantitative evaluation, including a symbol-level metric (SER) computed on the actual braille output rather than only on an intermediate MIDI representation; and (iii) a cell-by-cell verification of the braille encoder against the official NIMBN manual, with 100% conformance for notes and rests.

A. Motivation

The soprano recorder is widely adopted in Brazilian music education, present in both elementary schools and specialized conservatories [8] [9]. Its popularity stems from its low cost, the absence of mechanical parts, and fingering that can be learned within a few classes. At the ICC, founded in Fortaleza in 1943 and a pioneer in the education of visually impaired people in the state [3], the recorder is taught regularly, making the demand for braille material for this instrument concrete and immediate.

From a technical standpoint, the recorder combines favorable characteristics for an AMT system. Being monophonic, it eliminates the multi-pitch separation problem that is the main difficulty of transcription systems for piano and ensemble instruments [6]. Its signal also has a well-defined fundamental frequency and a fixed melodic range (C5 to D7 on the soprano recorder), which restricts the pitch detector's search space and simplifies note quantization.

The literature review found AMT work for other monophonic, niche wind instruments – the sopele, a traditional Croatian flute [10], and the diatonic harmonica [11] – showing that instrument-adapted classical approaches are viable and competitive even without large annotated datasets. However, no identified work addresses the recorder specifically, and none proposes braille music notation as the transcription system's output. This work positions itself at that intersection.

B. Objectives

The general objective is to develop and evaluate a system for automatic transcription of soprano recorder melodies, from audio, to braille music notation. The system implements a classical signal-processing pipeline – pitch detection via pYIN, segmentation and rhythmic quantization, and symbolic conversion to braille – without any supervised training stage, and its technical feasibility is verified through quantitative metrics on synthetic and real recordings. Specific objectives are: (1) a systematic literature review on the AMT state of the art; (2) a pitch detection and note segmentation module based on pYIN [12]; (3) a converter from note sequences to braille music notation following the NIMBN standard [1]; (4) an evaluation set of 30 synthetic melodies with automatic ground truth and real recorder recordings provided by the ICC; (5) a quantitative evaluation against the Basic Pitch neural model [13] as a baseline.

II. Related Work

The AMT literature is dominated by deep neural systems developed and evaluated mostly for piano, with datasets of hundreds of hours of annotated audio [7] [6]. This concentration motivates the empirical comparison, in this work, with a general-purpose neural model (Basic Pitch [13]) as a direct test of a domain-mismatch hypothesis.

For niche, monophonic wind instruments, Skoki et al. [10] transcribed the sopele using a supervised model trained on a small, purpose-built dataset, while Lins et al. [11] transcribed the diatonic harmonica using a classical approach without large-scale training. Marták, Hu and Widmer [14] systematically show that general-purpose neural transcription models degrade when recording conditions or musical style diverge from their training corpus, a domain-mismatch hypothesis tested empirically in Section IV.

On the braille side, existing tools (Goodfeel, BrailleMUSE, Musibraille, FreeDots) and academic systems such as Goto et al. [5] all convert from an already-digitized score (MusicXML) to braille notation; none starts from raw audio. The gap between audio-based AMT and braille music notation output, identified by the systematic review underlying this work, remains, to the authors' knowledge, unaddressed in the literature.

TABLE I
pYIN Parameters

Parameter	Value
`fmin` / `fmax`	523.3 Hz (C5) / 2,349 Hz (D7)
`frame_length`	2048 samples ($\sim$93 ms)
`hop_length`	512 samples ($\sim$23 ms)
`voiced_threshold`	0.10
Sample rate	22,050 Hz

III. Methodology

The proposed system is a modular AMT pipeline for soprano recorder with braille music notation output (Fig. 1), composed of three sequential modules, implemented in Python 3.11 using `librosa` [15] and `pretty_midi` [16].

A. Module 1: Pitch Detection (pYIN)

Fundamental frequency (f_0) is estimated frame-by-frame with the pYIN algorithm [12] via `librosa.pyin`. Key parameters, calibrated to the recorder's physical range, are listed in Table I. Audio is pre-processed by resampling to 22,050 Hz, trimming silence at the edges (-30 dB threshold), and amplitude normalization.

B. Module 2: Note Segmentation and Rhythmic Quantization

f_0 is discretized to the nearest integer MIDI number per voiced frame, after 3-frame median smoothing. A new note starts on an *unvoiced→voiced* transition or on a confirmed MIDI change (persisting for $\geq$ 2 frames), which avoids false onsets from attack transients. Since sustained repeated notes of equal pitch produce no MIDI change and no voiced/unvoiced transition, a dedicated onset-detection step is applied to segments longer than 1.3 beats: an onset novelty

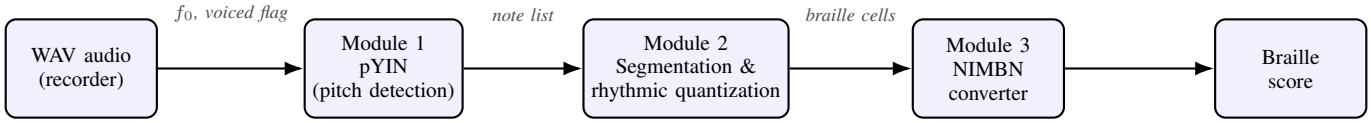


Fig. 1. Overview of the proposed pipeline: pitch detection (pYIN), note segmentation and rhythmic quantization, and conversion to braille music notation (NIMBN standard).

function (`librosa`) is peak-picked with a calibrated threshold ($\delta = 0.55$), and candidates are validated against rhythmic-grid proximity and energy-drop criteria before a segment is split. Note durations are quantized to $\{0.5, 1.0, 2.0, 4.0\}$ beats (eighth, quarter, half, whole note) using tempo estimated by `librosa.beat_track` [15].

C. Module 3: Braille Music Notation Converter

The converter implements the subset of the NIMBN standard [1] sufficient to represent diatonic monophonic melodies with basic rhythm: notes C5–D7, the four durations above and their rests, octave signs, treble clef, and 4/4 and 3/4 time signatures. The cell-encoding table was verified cell-by-cell against the official manual, reaching 100% conformance for notes and rests after correcting an initial implementation error. Output is a Unicode string using the Braille Patterns block (U+2800–U+28FF). This conformance check verifies that the encoder correctly implements the programmed NIMBN rules for the supported elements; it does not, by itself, demonstrate that the resulting score is musically correct, legible, or usable by a visually impaired reader – a distinct question this work does not evaluate and leaves to the formal legibility validation with braille music readers proposed in Section V.

D. Evaluation Set and Baseline

The evaluation set combines two subsets. The *synthetic subset* comprises 30 monophonic soprano recorder melodies (613 notes) – scales, technical exercises, and folk/classical pieces – synthesized via FluidSynth with a General MIDI *Recorder* soundfont; ground truth (GT) was extracted automatically from the source MIDI files. The *real subset* comprises 8 recordings (329 notes) made by ICC students with a smartphone in an untreated acoustic environment; ground truth was derived from reference scores and pYIN-based spectral analysis. Note identity in all eight real-subset ground truths has since been independently verified by a music teacher at the ICC, who transcribed each recording by ear from the printed score without access to the pipeline's output; this cross-check corrected two notes omitted from one recording.

As a baseline, Basic Pitch [13] is a lightweight, instrument-agnostic neural AMT model whose documented training corpus does not include wind instruments. It was evaluated in three configurations – default parameters, frequency range restricted to the recorder's register, and a grid-search-calibrated configuration (`onset_threshold`, `frame_threshold`; 16 combinations) – giving both methods a broadly comparable degree of calibration effort. A third Basic Pitch parameter controlling note-length granularity (`minimum_note_length`) was left at its library default and not included in the grid search, so this comparison should be read as approximately, not strictly, equivalent in calibration effort.

TABLE II
CONSOLIDATED METRICS – SYNTHETIC DATASET ($n = 30$ MELODIES, 613 NOTES)

Metric	Value
Onset F1 (mir_eval, 50 ms tol.)	**94.6%**
Temporal pitch accuracy	94.0%
Positional pitch accuracy	80.6%
Note Error Rate (NER)	5.4%
Symbol Error Rate (SER)	**8.6%**
Notes detected / ground truth	593 / 613

Four metrics were used: onset F1-score (`mir_eval` [17], 50 ms / $\pm$50-cent tolerance), positional pitch accuracy (index-aligned, more demanding for braille readability), Note Error Rate (NER $= 1 - F_1$), and Symbol Error Rate (SER), the edit distance between predicted and reference braille cell sequences, adopted as the most relevant metric since it directly reflects the pipeline's final product.

IV. RESULTS AND DISCUSSION

A. Synthetic Dataset

Table II summarizes results on the 30 synthetic melodies. The pipeline detected 593 of 613 expected notes, with onset F1 of 94.6% and SER of 8.6%. Ten melodies (mostly scales and stepwise-motion pieces) reached F1 = 100%; the worst case, *jingle_bells_frag* (F1 = 64.9%), contains long runs of fast repeated notes.

Failure analysis converges on a single root cause: the difficulty of delimiting consecutive notes of equal pitch, which surfaces both as under-segmentation (merged repeated notes, e.g. *parabéns*: 25→22 notes) and over-segmentation (a single sustained note split in two, e.g. *intervalo_oitava*: 9→11 notes) – both visible in the Δ column of Table III. This explains the gap between temporal (94.0%) and positional (80.6%) pitch accuracy: any count mismatch shifts every subsequent note's index, cascading the penalty through the rest of the sequence.

Fig. 2 shows the full distribution: melodies with exclusively stepwise motion cluster at F1 = 100%, while the worst performers share long runs of repeated notes at moderate-to-fast tempo. The dashed line in Fig. 2 marks the unweighted mean across the 30 melodies (94.6%), consistent with the note-weighted onset F1 reported in Table II.

TABLE III
PER-MELODY RESULTS ($n = 30$ MELODIES; EXCERPT: 3 BEST, 4 WORST BY ONSET F1)

Melody	GT	Pred	Δ	F1	SER
escala_do_maior	16	16	0	100.0%	0.0%
london_bridge	24	24	0	100.0%	0.0%
minueto_sol_bach	30	30	0	100.0%	0.0%
intervalo_segunda	16	16	0	87.5%	0.0%
mary_had_a_little_lamb	26	22	−4	87.5%	25.9%
intervalo_oitava	9	11	+2	80.0%	22.2%
jingle_bells_frag	24	13	−11	64.9%	64.0%

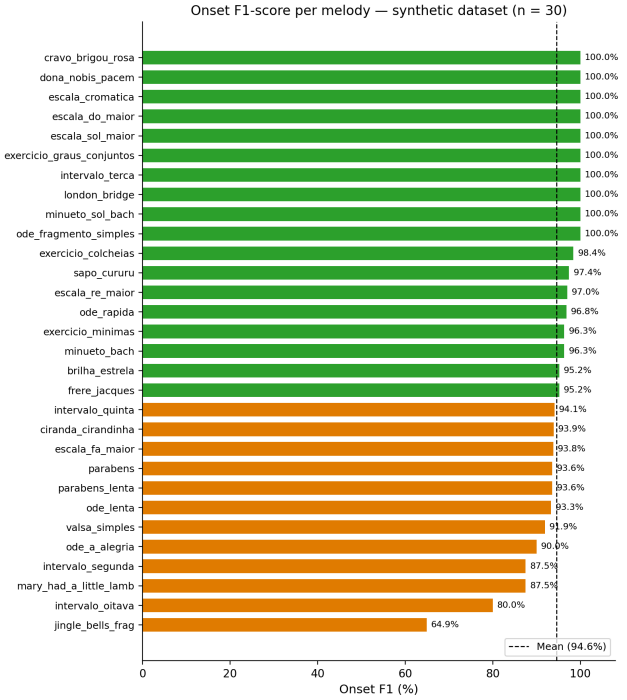


Fig. 2. Onset F1-score per melody, synthetic dataset, sorted in descending order.

B. Comparison with Basic Pitch

Table IV compares the two methods. In its default configuration, Basic Pitch detected 1,111 notes for 613 expected (1.81×). Restricting only the frequency range barely changed this (1,109 notes), ruling out simple out-of-range detections as the cause. Grid-search calibration of confidence thresholds corrected most of the excess (644 notes, 1.05×) and raised onset F1 to 96.8%, above the proposed pipeline (94.6%) on that single metric. However, SER tells a different story: even calibrated, Basic Pitch's SER (36.4%) remains over four times worse than pYIN's (8.6%), showing that domain mismatch persists at a finer transcription layer not captured by onset F1 – more than a third of the resulting braille cells would still require manual correction, against 8.6% for the proposed pipeline. This is consistent with Marták, Hu and Widmer's [14] observation that general-purpose neural models degrade when recording conditions or instrument diverge from their training

TABLE IV
PYIN (PROPOSED) VS. BASIC PITCH (BASELINE) – SYNTHETIC DATASET ($n = 30$ MELODIES, GT = 613 NOTES)

Method	Onset F1	SER	Notes	Notes/GT
pYIN (proposed)	94.6%	**8.6%**	593	0.97×
Basic Pitch – default	71.7%	107.7%	1,111	1.81×
Basic Pitch – restr. range	71.6%	107.7%	1,109	1.81×
Basic Pitch – calibrated	**96.8%**	36.4%	644	1.05×

TABLE V
CONSOLIDATED METRICS – REAL RECORDINGS ($n = 8$, 329 NOTES)

Metric	Value
Onset F1	**93.6%**
Temporal pitch accuracy	99.2%
Positional pitch accuracy	40.8%
Note Error Rate (NER)	6.4%
Symbol Error Rate (SER)	17.3%
Notes detected / ground truth	369 / 329 (1.12×)

corpus, and supports the choice of a classical, instrument-adapted approach for niche instruments lacking large annotated datasets. We note that SER is also the metric on which pYIN's advantage over calibrated Basic Pitch is most pronounced among those reported in Table IV; we treat SER as the most relevant metric because it corresponds directly to the pipeline's final output (Section III-D), not because of this coincidence, but the coincidence itself is a limitation to bear in mind when interpreting the comparison (see Section IV-E).

C. Real Recordings

On real recordings (Table V), onset F1 (93.6%) is close to the synthetic result (94.6%), indicating consistent behavior once the synthetic dataset's register was corrected to match the recorder's real range. This correction was informed by comparing the synthetic dataset's pitch range against the real recordings' observed range, which revealed a register mismatch in the original synthetic set. Positional pitch accuracy drops markedly (40.8% vs. 80.6%), driven by systematic over-segmentation present in all eight recordings (6.5%–25.0% extra notes, Table VI) combined with longer average sequences, which amplifies the cascading-index-error effect described above. SER (17.3%) is adopted as the most informative metric for practical readability, since it captures both pitch and count errors proportionally rather than under an all-or-nothing criterion. The best individual result was *serra_serrador_fl2* (F1 = 96.3%, SER = 11.1%), a rhythmically simple piece; the worst was *trem_maluco_fl1* (F1 = 88.9%, SER = 29.2%), composed entirely of long notes, which are more prone to the over-segmentation described above.

D. Positioning Relative to the Literature

The obtained F1 (94.6%/93.6%) is in the same order of magnitude as prior work on training-free transcription of niche wind instruments. Mauch and Dixon [12], evaluating pYIN on synthesized singing from the RWC Music Database, report recall/precision between 97.7% and 98.7%, but computed

TABLE VI
PER-RECORDING RESULTS – REAL AUDIO ($n = 8$)

Recording	GT	Pred	F1	SER
asa_branca	104	120	91.1%	18.4%
cancao_ninar	46	49	92.6%	18.2%
morgen_kommt	48	54	94.1%	18.4%
trem_maluco_fl1	16	20	88.9%	29.2%
trem_maluco_fl2	17	19	94.4%	16.7%
ode_a_alegria	49	54	95.1%	17.7%
serra_serrador_fl1	23	25	95.8%	8.6%
serra_serrador_fl2	26	28	96.3%	11.1%

frame-by-frame (5.8 ms windows, one-semitone tolerance) rather than at note/onset level. Lins et al. [11], transcribing diatonic harmonica, reported a mean F-measure of 68.1% for multi-pitch detection, also frame-based (9.5 ms grid). In both cases the unit of analysis is more permissive than the note-level, 50 ms-tolerance onset F1 used here, since a frame can be counted correct even when note-level segmentation is wrong. None of the three evaluations share metric, dataset, or tolerance, so these figures indicate that the proposed pipeline's performance sits within the same magnitude range as the niche-instrument literature, not evidence of superiority.

E. Threats to Validity

The following limitations are grouped by validity type, following standard practice for empirical evaluations of signal-processing systems.

Statistical conclusion validity. The two thresholds calibrated empirically (`voiced_threshold`, onset-detection δ) were tuned on the same 30 melodies later reported as the final synthetic result, without an independent holdout set. A symmetric limitation applies to the calibrated Basic Pitch baseline, whose grid search used the same 30 melodies. Table VII shows that `voiced_threshold` degrades smoothly across a $6\times$ wider range than typically tested (F1: 94.6%→93.4% from 0.05 to 0.30), and δ varies within 3 percentage points of F1 across its tested range; the pipeline does not depend on a narrow, fragile optimum. Reported metrics are unweighted means (30 melodies / 8 recordings) with no significance test or confidence interval. The real subset (329 notes across only six distinct pieces, two contributing two voices each) is too small to support claims of variance beyond the sample evaluated. The comparison in Section IV-B also rests on a single neural baseline (Basic Pitch); the claim that the proposed classical approach is more reliable than general-purpose neural models for this niche instrument should not be read as established beyond that one baseline, evaluated on the calibration-leakage-affected sample described above.

Construct validity. Real-recording ground truth was derived via pYIN-based spectral analysis, the same algorithm under evaluation, calibrated to the audio's estimated BPM. Note identity in all eight ground truths has since been independently verified by a music teacher at the ICC, who transcribed each recording by ear from the printed reference score without access to the pipeline's output; this cross-check

TABLE VII
THRESHOLD SENSITIVITY (ONSET F1 / SER), SYNTHETIC DATASET

`voiced_thr.`	F1 / SER	δ (onset)	F1 / SER
0.05	94.6% / 8.6%	0.45	93.3% / 12.9%
0.10	**94.6% / 8.6%**	**0.55**	**94.6% / 8.6%**
0.20	94.3% / 9.4%	0.60	95.4% / 6.5%
0.30	93.4% / 10.3%	0.70	95.1% / 7.5%

corrected two notes omitted from one recording (previously undercounted in *Canção de Ninar*). Onset/offset instants were not independently re-measured in any of the eight recordings – the reported values remain pYIN-derived in all cases. The same teacher did informally confirm, while listening to each recording alongside her transcription, that the rhythmic pacing sounded consistent with the audio; this is a qualitative, ear-based impression, not a precise, sample-level independent annotation of onset/offset instants, so it attenuates but does not eliminate the circularity described above for timing information, and the real-audio evaluation still cannot be treated as fully independent of the evaluated technique on that dimension. Tempo (BPM) was also supplied from ground truth rather than estimated autonomously in every reported result, which directly affects rhythmic quantization and therefore SER. A fully autonomous deployment – the work's actual motivating use case, where no reference score exists – may yield different SER values. NER and SER are also reported only in aggregate; we did not decompose either metric into insertion, deletion, substitution, duration-specific, and rhythm-specific error components. Such a decomposition would isolate the contribution of the repeated-note segmentation bottleneck (Section IV-A) from other error sources more precisely than the aggregate figures allow, and is left to future work.

External validity. Both datasets cover mostly elementary didactic repertoire (scales, exercises, simple folk melodies), performed by students rather than professional musicians, reflecting the proficiency level at the partner institution at collection time rather than a deliberate sampling of difficulty levels. The pipeline was not tested on intermediate/advanced repertoire, other recorder registers (alto, tenor, bass), or other instruments, and performer skill is confounded with pipeline error in the reported real-audio results.

V. CONCLUSION

This work developed and evaluated a training-free pipeline combining pYIN-based pitch detection, MIDI-discretization segmentation with localized repeated-note detection, and NIMBN-standard braille conversion, achieving onset F1 of 94.6% (synthetic) and 93.6% (real recordings), and outperforming a calibrated general-purpose neural baseline on SER (8.6% vs. 36.4%, Section IV-B). The braille cell-encoding module reached full conformance with the official manual, as detailed in Section III-C. The main remaining limitation is delimiting consecutive notes of equal pitch, particularly in fast passages with repeated eighth notes.

These results support the three contributions outlined in Section I. For this niche instrument, our results suggest that a classical, instrument-adapted method remains competitive with – and in symbol-level accuracy, more reliable than – the single general-purpose neural baseline tested here, even after calibration; whether this pattern generalizes to other niche instruments or other neural baselines is left to future work. From an engineering standpoint, integrating braille conversion directly into the transcription pipeline eliminates the intermediate MusicXML digitization step required by existing tools (Goodfeel, BrailleMUSE, Musibraille), directly addressing the literature gap identified in Section II.

Addressing the repeated-note segmentation bottleneck is prioritized ahead of extending the pipeline to other instruments. Two directions appear most promising, both preserving the pipeline's training-free character by refining decision rules within the existing signal-processing framework rather than introducing a supervised model: (a) replacing the current hard energy-drop threshold with a probabilistic decision – e.g., a duration-aware Hidden Markov Model over inter-onset intervals, conditioned on estimated tempo, so a split is accepted only when jointly consistent with the rhythmic grid and a learned prior over note-duration sequences, rather than by a single onset-novelty peak; and (b) combining multiple onset-detection functions (spectral flux, phase deviation, and the energy-envelope criterion already used) to reduce dependence on any single acoustic cue, since the failure mode identified in Section IV-A suggests the energy-drop criterion alone under- or over-triggers depending on tempo and dynamics.

Future work includes: (1) the segmentation refinements above, together with dynamic calibration of the energy-drop threshold conditioned on estimated tempo; (2) extension to other monophonic instruments and recorder registers; (3) expansion of the real dataset with professional performers, extending the independent verification described in Section IV-E from note identity to onset/offset instants (not yet covered), and applying a development/test split or music-level cross-validation to obtain generalization estimates free of the calibration leakage also disclosed in Section IV-E; (4) comparison against additional AMT baselines beyond Basic Pitch, checking whether the SER advantage reported in Section IV-B holds on a fully independent test set; (5) formal validation with visually impaired users of the ICC, assessing braille-score legibility, the number of manual corrections required, time to understand or perform the transcribed melody, and perceived usefulness – since the 100% NIMBN conformance reported in Section III-C establishes only that the encoder is correctly implemented, not that its output is musically legible or useful in practice; and (6) public release of the dataset, source code, and evaluation scripts.

Acknowledgment

The authors thank the Instituto de Cegos do Ceará for providing the recordings and reference scores that made the real-audio evaluation possible.

References

[1] B. Krolick, *New International Manual of Braille Music Notation*. World Blind Union, 1996.

[2] Instituto Benjamin Constant, *Musicografia Braille*. Rio de Janeiro: IBC, [n.d.]. [Online]. Available: http://www.ibc.gov.br. Accessed: Jun. 2026.

[3] Instituto de Cegos do Ceará, *Quem somos*. Fortaleza, 2024. [Online]. Available: https://institutodecegosdoceara.com.br/quem-somos/. Accessed: Jun. 2026.

[4] UK Association for Accessible Formats (UKAAF), *Introduction to automated braille music transcription*. London: UKAAF, 2024. [Online]. Available: https://www.ukaaf.org/standards/introduction-to-automated-braille-music-transcription/. Accessed: Jun. 2026.

[5] D. Goto, T. Gotoh, R. Minamikawa-Tachino, and N. Tamura, "A Transcription System from MusicXML Format to Braille Music Notation," *EURASIP J. Advances in Signal Processing*, vol. 2007, art. 42498, 2007, doi: 10.1155/2007/42498.

[6] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic Music Transcription: An Overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[7] C. Hawthorne et al., "Onsets and Frames: Dual-Objective Piano Transcription," in *Proc. Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, 2018, pp. 50–57.

[8] T. N. Taets, G. G. de C. Taets, and M. J. S. da C. Lins, "Ensino coletivo de flauta doce na educação básica," *Revista Brasileira de Educação Básica*, 2017. [Online]. Available: https://rbeducacaobasica.com.br/2017/10/26/ensino-coletivo-de-flauta-doce-na-educacao-basica/. Accessed: Jun. 2026.

[9] F. D. da Silva, "O ensino de flauta doce na escola pública," *Revista Brasileira de Educação Básica*, 2018. [Online]. Available: https://rbeducacaobasica.com.br/2018/10/03/o-ensino-de-flauta-doce/. Accessed: Jun. 2026.

[10] A. Skoki, S. Ljubic, J. Lerga, and I. Štajduhar, "Automatic music transcription for traditional woodwind instruments sopele," *Pattern Recognition Letters*, vol. 128, pp. 340–347, 2019.

[11] F. Lins, M. O. Johann, E. Benetos, and R. Schramm, "Automatic Transcription of Diatonic Harmonica Recordings," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, 2019.

[12] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 659–663, doi: 10.1109/ICASSP.2014.6853678.

[13] R. M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert, "A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 781–785.

[14] L. S. Marták, P. Hu, and G. Widmer, "Sound and music biases in deep music transcription models: a systematic analysis," *J. Audio, Speech, and Music Processing*, 2025, doi: 10.1186/s13636-025-00428-z.

[15] B. McFee et al., "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf. (SciPy)*, Austin, 2015, pp. 18–24.

[16] C. Raffel and D. P. W. Ellis, "Intuitive Analysis, Creation and Manipulation of MIDI Data with pretty_midi," in *ISMIR Late Breaking Demos*, 2014.

[17] C. Raffel et al., "mir_eval: A Transparent Implementation of Common MIR Metrics," in *Proc. Int. Society for Music Information Retrieval Conf. (ISMIR)*, Taipei, 2014, pp. 367–372.

Capturing Functional Upper-Limb Movements Using Low-Cost IMUs

Rui Gonçalves de Oliveira Júnior
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
ruioliveira@discente.ufg.br

João Vitor Menegon Xavier
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
joaoxavier@discente.ufg.br

Lucas Matsuno da Costa
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
lucasmatsuno@discente.ufg.br

Ligyane Hellen Tobias Lopes
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
ligyanelopes@discente.ufg.br

Marco Antonio Assfalk de Oliveira
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
assfalk@ufg.br

Jalusa Andreia Storch Díaz
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
jastorch@ufg.br

Aldo André Díaz Salazar
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
aldo.diaz@ufg.br

Brenda Gomes Oliveira
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
brendabrenda2@discente.ufg.br

***Abstract*—Objective monitoring of upper-limb rehabilitation after stroke typically relies on expensive motion-capture systems confined to laboratories. This work describes the operation of a low-cost inertial capture system, composed of an ESP32 microcontroller and BNO085 sensors, that acquires the orientation of the limb segments, transmits the data over Wi-Fi, and processes them through a biomechanical modeling pipeline in OpenSim. The system reproduces the patient's pose in three dimensions and simultaneously estimates the shoulder, elbow, and forearm joint angles during functional movements, making it possible to measure joint range of motion objectively and replay the same movement as many times as needed, for educational and therapeutic purposes. Using only two sensors, placed on the arm and forearm, the prototype, with a materials-only cost of approximately BRL 255, demonstrates the technical feasibility of an affordable tool for kinematic assessment in rehabilitation; the accuracy of the estimated angles against a reference and the clinical validity with patients remain to be established.**

***Index Terms*—motion capture, inertial sensors, IMU, functional movements, assistive technologies, rehabilitation, stroke, biomechanics, OpenSim**

I. Introduction

Stroke is one of the leading causes of motor disability worldwide, frequently impairing upper-limb function [1]. Recovery is a prolonged process involving many physical-therapy sessions based on movement repetition. This process imposes a physical burden on the healthcare professional and, when the frequency of care is low, may delay the patient's return to functional independence.

There is, therefore, room for systems that increase the frequency of exercises and provide objective movement measurements without adding to the therapist's physical burden. Motion capture with inertial sensors (IMUs) is well suited to this context, as it does not require cameras, is not affected by occlusion, and uses portable, inexpensive sensors. Established commercial inertial systems [2], however, are costly, which makes their routine use in public health services unfeasible.

Moreover, there is still no consensus in the literature on the use of IMUs for analyzing upper-limb functionality, particularly regarding the number of sensors required and the anatomical locations where they should be positioned. This work therefore also aims to contribute to that debate.

The clinical focus is on capturing functional movements, that is, task-oriented gestures from daily life, such as bringing a cup to the mouth. By reproducing these movements in three dimensions and quantifying their joint angles, the system allows the physical therapist to visualize and replay a recording repeatedly, measure joint range of motion without a goniometer, and objectively track the patient's progress. These characteristics place this work within the scope of assistive rehabilitation technologies and were outlined in collaboration with a physical-therapy research group.

This work describes the operation of this low-cost inertial capture system for the upper limbs, from data acquisition by the sensors to the estimation of joint angles through biomechanical modeling, with an emphasis on its application to rehabilitation.

II. Related Work

Beyond its acute impact, stroke is a leading cause of long-term disability, imposing a major burden on patients, families, and society [1]. Rehabilitation after a stroke is an iterative process involving specialized physical-therapy assessment and training, which are often limited by the scarce resources of health and rehabilitation centers. Low-cost IMUs have the potential to assess and monitor patients both inside and outside

clinical settings, allowing a more detailed evaluation of the impairment and enabling the individualization of rehabilitation therapies [3].

Reference motion-capture systems are based on multiple cameras or infrared tracking with markers. Although they offer high accuracy, they are costly and require a controlled environment, falling outside the scope of routine use in public physical-therapy services. For this reason, the state of the art considered here is restricted to low-cost, low-complexity solutions compatible with the project's budget.

Within this scope, two approaches stand out. The first is capture with inertial sensors (IMUs), in which segment orientation is estimated by fusing accelerometer, gyroscope, and magnetometer data; it is portable, is not affected by occlusion, and requires no cameras. The second is vision-based pose estimation with a single camera, in which algorithms extract body posture directly from images [4]; it is affordable and requires no instrumentation on the patient, but it depends on lighting and framing conditions and suffers from occlusions.

The proposed system falls under the first approach, prioritizing portability and freedom from occlusion. Vision-based pose estimation is considered a complementary resource for future work, and not a direct competitor, given the adopted low-cost scope.

Regarding validation, IMU-based upper-limb systems are usually assessed against optical motion capture [3], [5]. For single-joint shoulder and elbow tasks, root-mean-square errors below 4° have been reported for the angles mainly involved in each movement [5], with larger errors in the remaining degrees of freedom, which are more sensitive to soft-tissue artifacts and magnetic disturbances. The present work does not yet include such a comparison, which is discussed as a limitation in Section IV.

III. The Proposed System

As an assistive technology, the system was designed for a simple usage flow by the physical therapist: the professional attaches two sensors to the patient's limb, performs a brief calibration, and asks the patient to perform the functional movement; the system then records the execution and, at the end, provides the three-dimensional animation and the joint angles. No step requires technical knowledge of instrumentation, and the digital record makes it possible to revisit the capture at any time.

Technically, the system is organized into three stages: embedded inertial acquisition, wireless transmission, and biomechanical processing. The ESP32 reads the sensors, preprocesses the orientations, and sends them over Wi-Fi to a server, which applies them to a musculoskeletal model in OpenSim to reproduce the patient's pose and compute joint kinematics. Fig. 1 shows the embedded unit of the prototype.

A. Acquisition: ESP32 and BNO085 Sensors

The embedded unit uses an ESP32 microcontroller, chosen for its low cost, native Wi-Fi connectivity, and dual-core processor running at about 240 MHz, capable of reading multiple sensors in real time. Its role in the project is to acquire, preprocess, and transmit the IMU data.

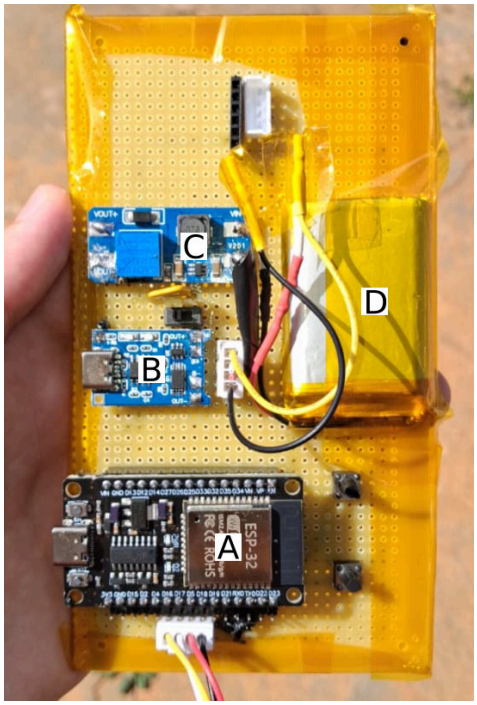


Figure 1. Prototype of the embedded unit, on perfboard: ESP32 (A), TP4056 charging module (B), MT3608 boost converter (C), and 3.7 V LiPo battery (D).

Segment orientation is measured by BNO085 sensors, which integrate an accelerometer, a gyroscope, and a magnetometer. The readings from these sensors are processed by a proprietary on-chip sensor-fusion algorithm, which directly provides the orientation of each segment as a quaternion. Communication with the ESP32 is carried out over an I^2C bus. Two sensors are used, attached with velcro straps to the right upper limb: one on the forearm and one on the arm, associated with the two I^2C addresses available on the component (Fig. 2). This limit of two addresses is the main challenge for scaling up the number of sensors, and it can be circumvented by signal routing (an I^2C multiplexer or multiple buses).

B. Firmware and Communication

The ESP32 firmware continuously reads the two sensors and converts the sensor's orientation reference frame to the frame adopted by OpenSim. At the start of the capture, the system performs a calibration that zeroes the initial horizontal orientation (yaw), removing the dependence on cardinal direction without altering the sensor's operating mode. During the capture, the orientations are stored on the device in a standardized file, while a live summary is streamed over Wi-Fi (UDP) to the server for monitoring; at the end of the capture, the file is uploaded to the server for processing. The nominal sampling rate is 50 Hz; the measured effective rate ranged from 30 to 36 Hz, because the firmware schedules each sample with a non-blocking 20 ms timer and, whenever an iteration

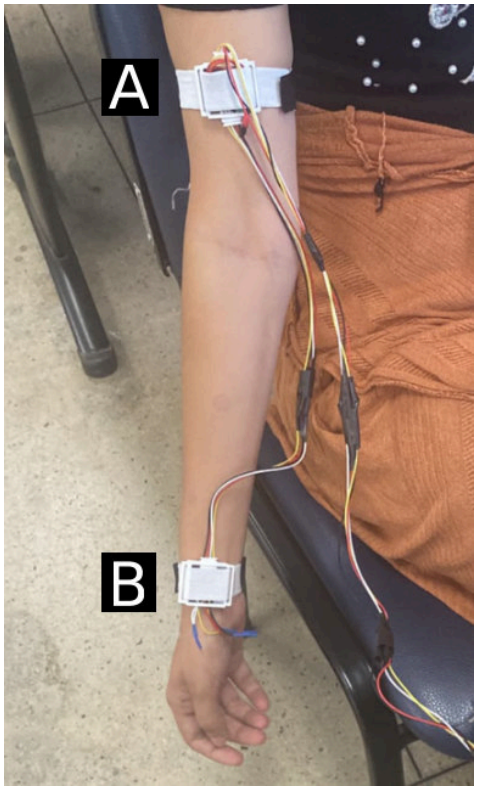


Figure 2. Wearable system positioned on the right upper limb, with the sensors attached with velcro straps to the arm (A) and the forearm (B).

of the main loop takes longer than that — I^2C polling of the two sensors, writing to the flash file system, and servicing the Wi-Fi stack — the missed interval is not recovered. For the wrist during activities of daily living, the dominant frequency is about 1 Hz and most of the spectral energy lies below 5 Hz [6], and 6 Hz is the usual upper bound adopted for voluntary human movement [7]; an effective rate of 30 Hz is therefore still 2.5 times the Nyquist rate for a 6 Hz bandwidth. Each sample is stamped with the ESP32 clock at the moment of acquisition and stored on the device, so the actual, irregular intervals are preserved in the recorded file and used by the inverse-kinematics step rather than a uniform 50 Hz being assumed.

C. Biomechanical Modeling in OpenSim

OpenSim is free software for biomechanical simulation of human movement; it relies on musculoskeletal models composed of bones, joints, and anatomical degrees of freedom [8]. The processing pipeline occurs in three steps. First, a calibration aligns the user's initial posture with the virtual skeletal model, registering the sensors to their respective segments. Next, orientation-based inverse kinematics [9] estimates, for each sample, the joint angles that best explain the measured orientations, mirroring the patient's pose. Finally, the resulting angles are exported in standardized biomechanical-analysis formats and visualized in three dimensions. The coordinates analyzed were shoulder flexion, adduction, and rotation; elbow flexion; and forearm pronation–supination. The functional movements evaluated — bringing a cup to the mouth, combing the hair, and waving — are relevant to activities of daily living, particularly in the context of post-stroke upper-limb rehabilitation. Fig. 3 illustrates the musculoskeletal model reproducing the captured pose, with the right upper-limb segments highlighted.

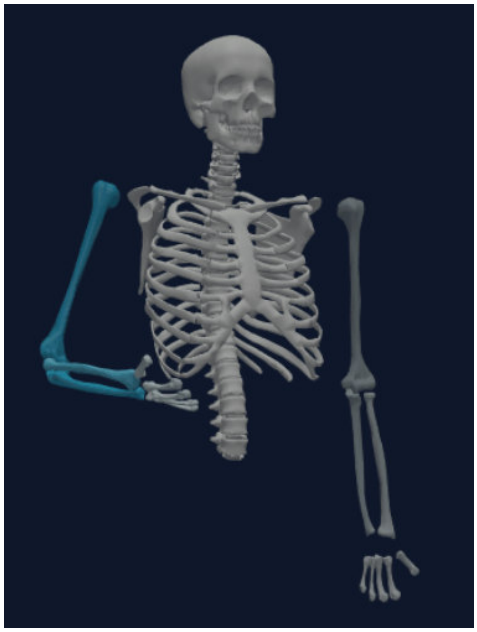


Figure 3. Musculoskeletal model in OpenSim reproducing the captured pose; the arm and forearm segments of the right upper limb tracked by the sensors are highlighted.

D. Web Interface

A web interface, served by the same server, provides the physical therapist with a single screen to operate the system and monitor the capture in real time. The data relayed by the server reach the browser through Server-Sent Events (SSE), so no software installation is required on the client side.

Each capture is stored as a session, listed in the interface together with the files produced by the pipeline — the calibrated model, the calibration and motion files, the inverse-kinematics results and orientation errors, and the plot and animation of the movement — which can be revisited at any time (Fig. 4). This organization keeps a persistent record of every acquisition, supporting the objective follow-up of the patient across sessions.

The interface also displays the connection status of the device and, for each of the two IMUs, the current roll, pitch, and yaw angles, an anatomical pose indicator with a qualitative fit score, and live plots of the three angles over time (Fig. 5, top). A control panel (Fig. 5, bottom) exposes the operations of the workflow — calibrate, record, and stop — together with the configuration of the OpenSim model, the heading axis, the sensor-to-OpenSim rotation, the inverse-kinematics weight, and the arm and forearm segment lengths. A real-time log reports each step, from the calibration pose to the

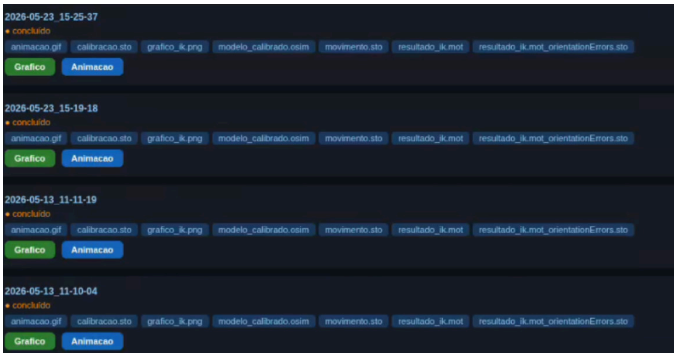


Figure 4. Session list in the web interface: each completed capture keeps the files produced by the OpenSim pipeline (calibrated model, calibration and motion files, inverse-kinematics results and orientation errors, plot, and animation).

triggering of the OpenSim pipeline on the server. This interface consolidates the acquisition, calibration, and processing steps into a workflow that requires no technical knowledge of instrumentation from the operator.

IV. Results and Discussion

The results reported in this section correspond to a functional demonstration of the instrument, in which the authors themselves operated the prototype to exercise the complete acquisition-to-visualization pipeline; no study with participants was conducted, and the estimated angles were not compared with a reference measurement. The system reconstructs the movement in three dimensions and provides the joint angles simultaneously (Figs. 3 and 6). Because the capture is recorded digitally, the same movement can be replayed as many times as needed, which favors educational use, training, and therapeutic purposes, while the direct reading of the joint angles dispenses with the manual goniometer and yields an objective record of the patient's progress without adding to the therapist's physical burden. Unlike the manual goniometer, which provides single-joint measurements, the system quantifies multiple joints simultaneously. Whether these measurements can support the assessment of movement quality or the detection of compensations is a question of clinical validity that is not answered by the present results, which establish the technical feasibility of the low-cost pipeline but not its measurement accuracy.

The estimates are sensitive to error sources inherent to inertial technology, especially magnetic disturbances and sensor-to-segment misalignment. In addition, because no sensor is placed on the chest, the shoulder angles are computed assuming a stationary chest, so any chest motion or compensation is reflected as error in these angles; an additional sensor would be required. In patients, who frequently recruit trunk compensations, this error is expected to be larger than in the demonstration captures, which were performed seated with the back supported. The anatomical constraints of the musculoskeletal model help keep the angles within physiologically plausible ranges, but they do not eliminate the cause of the error. Moreover, the OpenSim model must be appropriately customized for each patient. For assistive use, this reinforces the need for a standardized collection protocol that fixes the initial posture, the positioning of the limbs, and the placement of the sensors; even small postural variations affect the captured movement and, therefore, the repeatability of the measurements. The definition of this protocol is being carried out together with the partner physical-therapy group, so as to ensure that the measurements are clinically useful and comparable across sessions.

A. Limitations

The main limitation of the present results is the absence of a quantitative accuracy assessment: the angles estimated by the system were not compared with a reference such as a universal goniometer, calibrated video, or optical motion capture, and the repeatability across executions was not quantified. The internal orientation error reported by the inverse-kinematics solver (Fig. 4) indicates how well the model fits the measured orientations, but it is not a measure of accuracy against the true joint angle. Root-mean-square errors below 4° against optical motion capture have been reported for commercial inertial systems in single-joint shoulder and elbow tasks [5], but these values cannot be assumed for the present low-cost prototype. This assessment, using MAE and RMSE per joint and the variability of the range-of-motion peaks across repetitions, is the immediate next step, and until it is completed the system should be regarded as a technically feasible low-cost pipeline rather than a validated measurement instrument.

B. Cost

The total cost of the prototype, considering only the materials (components) and excluding labor or development, was approximately BRL 255, a value considerably lower than that of commercial inertial systems, which cost between 4 and 31 times more per sensor. This is, however, a simplified comparison, since dividing the resale price by the number of sensors is not entirely fair; moreover, many commercial systems require recurring license payments. The comparison therefore reveals only the order of magnitude of the cost difference. The low cost and the use of the open-source OpenSim API and GUI package, as well as commercial off-the-shelf (COTS) components, are central to the proposal, since they enable the local and large-scale use and maintenance of the tool in physical therapy.

V. Conclusion

This work described the operation of a low-cost inertial motion-capture system for the upper limbs, from acquisition with the BNO085 sensors and the ESP32 to the estimation of joint angles through biomechanical modeling in OpenSim. The system reproduces the patient's pose in three dimensions and provides objective, simultaneous measurements of the shoulder, elbow, and forearm, with the potential to support rehabilitation monitoring without adding to the therapist's physical burden.



Figure 5. Web interface for capture control and real-time monitoring. Top: monitoring dashboard with connection status, current roll/pitch/yaw angles of the two IMUs, anatomical pose indicators, and live angle plots. Bottom: control and calibration panel, with the calibrate/record/stop operations, OpenSim model and parameter configuration, and the real-time log.

The results demonstrate the technical feasibility of the proposed approach: with only two sensors and a total material cost of approximately BRL 255, based on commercial off-the-shelf components and the open-source OpenSim package, it provides a kinematic assessment tool aimed at the same purpose as systems costing several times more. The complete pipeline was exercised by the authors operating the prototype, confirming its end-to-end functionality. These results do not establish measurement accuracy or clinical validity: the angles were not compared with a reference and the captures assumed a stationary thorax. Quantitative validation against a goniometer or calibrated video, followed by validation with patients in a real clinical scenario, is planned as future work, to be conducted upon approval by a research ethics committee.

The results presented are preliminary and restricted to what was actually observed in the captures performed. Future work includes the definition of a unified collection protocol, covering posture, limb positioning, sensor attachment, and the number of sensors. Regarding this last point, reported configurations range from two to ten sensors, and establishing the number that is sufficient for functional assessment remains an open question that we intend to investigate. In particular, we intend to evaluate the addition of a third sensor on the scapular girdle so as to jointly assess scapular, shoulder, and elbow motion and thus better cover the dynamics of the upper limb, removing the stationary-thorax assumption. Future work

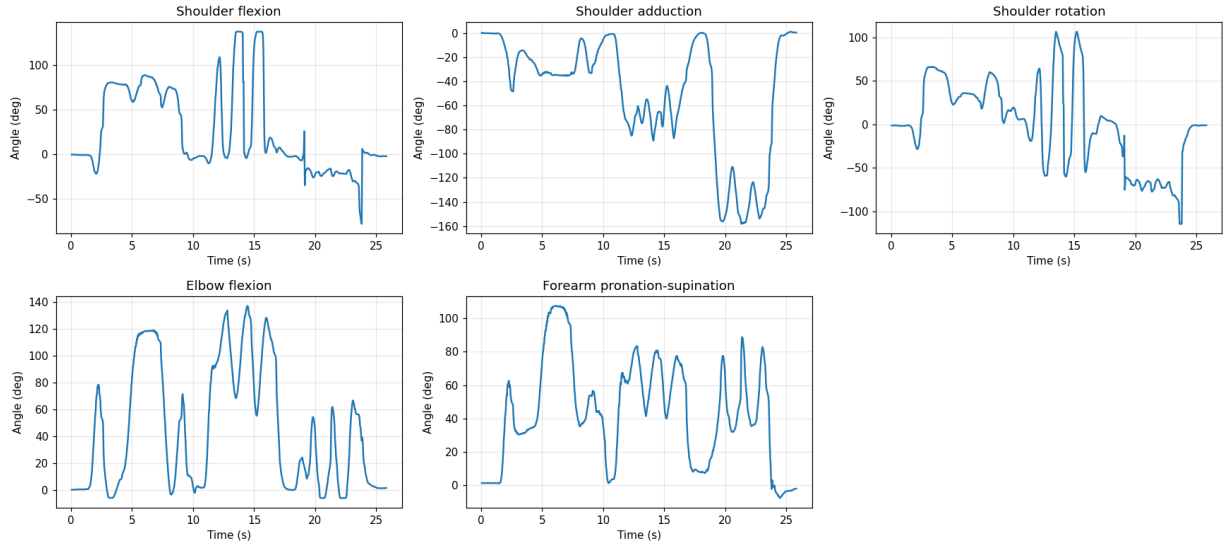


Figure 6. Joint angles estimated over three functional movements performed in sequence: bringing a cup to the mouth, combing the hair, and waving.

also includes the segmentation of movements into phases and the temporal normalization of the series to a common time base by resampling/interpolation [10], to allow comparison between executions of different durations. Finally, it includes comparison with quality-assessment methods from the literature, the use of complementary means such as computer vision, and the application of movement clustering and classification methods, such as K-means, to organize and compare large sets of recorded functional movements. A key next step is to validate the estimates against reference techniques, such as a universal goniometer and calibrated video in a first stage, and marker-based optical motion capture (the gold standard) and established inertial systems (e.g., Xsens MVN [2]) subsequently, quantifying the estimation error (MAE/RMSE per joint) and the repeatability between executions; the use of digital twins is envisioned as a future application of the tool, not as a comparison reference.

References

[1] V. L. Feigin et al., "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," Lancet Neurol., vol. 20, no. 10, pp. 795–820, 2021.

[2] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors," Xsens Technologies B.V., Enschede, The Netherlands, Tech. Rep., 2009. [Online]. Available: https://www.semanticscholar.org/paper/ffd2416057dc96d69ab6ed047910606f270039be

[3] T. Unger et al., "Upper limb movement quality measures: comparing IMUs and optical motion capture in stroke patients performing a drinking task," Front. Digit. Health, vol. 6, Art. no. 1359776, 2024.

[4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, 2021.

[5] A. Cutti, A. Giovanardi, L. Rocchi, A. Davalli, and R. Sacchetti, "Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors," Med. Biol. Eng. Comput., vol. 46, no. 2, pp. 169–178, 2008.

[6] K. A. Mann, F. W. Werner, and A. K. Palmer, "Frequency spectrum analysis of wrist motion for activities of daily living," J. Orthop. Res., vol. 7, no. 2, pp. 304–306, 1989.

[7] D. A. Winter, Biomechanics and Motor Control of Human Movement, 4th ed. Hoboken, NJ, USA: Wiley, 2009.

[8] A. Seth et al., "OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement," PLoS Comput. Biol., vol. 14, no. 7, Art. no. e1006223, 2018.

[9] M. Al Borno et al., "OpenSense: An open-source toolbox for inertial-measurement-unit-based measurement of lower extremity kinematics over long durations," J. NeuroEng. Rehabil., vol. 19, no. 1, Art. no. 22, 2022.

[10] N. E. Helwig, S. Hong, E. T. Hsiao-Wecksler, and J. D. Polk, "Methods to temporally align gait cycle data," J. Biomech., vol. 44, no. 3, pp. 561–566, 2011.

Clinical Applicability of Wearable Accelerometers for Functional Assessment of Older Adults in a Home Environment

Amanda M.C. Perez
Department of Physical Therapy
Federal University of Pernambuco
Recife, Brazil
amanda.mariac@ufpe.br

Caroline C. B Souza
Department of Physical Therapy
Federal University of Pernambuco
Recife, Brazil
caroline.bsouza@ufpe.br

Alysson L. P. C. da Silva
Department of Electronics and Systems
Federal University of Pernambuco
Recife, Brazil
alysson.lucas@ufpe.br

Naelso A. Cunha
Department of Electronics and Systems
Federal University of Pernambuco
Recife, Brazil
naelso.cunha@ufpe.br

Malki-çedheq B.C. da Silva
Department of Electronics and Systems
Federal University of Pernambuco
Recife, Brazil
malki-cedheq.benjamim@ufpe.br

Marco A. B. Rodrigues
Department of Electronics and Systems
Federal University of Pernambuco
Recife, Brazil
marco.rodrigues@ufpe.br

***Abstract*—Aging is associated with progressive changes in functional capacity, particularly in balance, mobility, and muscle strength, increasing the risk of falls and dependency. In this context, wearable sensors, such as accelerometers, have emerged as a promising alternative for the objective assessment of human movement. The aim of this study was to analyze the clinical applicability of these devices in the functional assessment of older adults in a home environment. This was a pilot, observational, and cross-sectional study involving individuals aged 60 years or older. Clinical tests, including the Timed Up and Go (TUG) test and the Medical Research Council (MRC) scale, were administered for screening purposes, and accelerometer data were analyzed during the performance of functional tasks. The wearable accelerometers successfully acquired kinematic signals during all functional tasks, allowing the identification of differences in movement symmetry, oscillatory patterns, and motor strategies among participants. It is concluded that wearable sensors have potential clinical applicability in the functional assessment of older adults in a home environment.**

Keywords— rehabilitation; physical therapy; older adults; monitoring; accelerometer data; functional assessment

I. Introduction

Population aging represents one of the major contemporary challenges in public health, particularly due to the increasing prevalence of chronic conditions and functional decline [1], [2]. Among the most common changes observed in older adults are balance deficits, reduced muscle strength, slower gait, and impaired functional mobility, all of which are directly associated with an increased risk of falls, hospitalizations, and loss of independence [3].

Home-based assessment has become increasingly relevant because it allows the evaluation of older adults under their habitual environmental conditions, potentially providing greater ecological validity than assessments performed exclusively in clinical settings.

In physical therapy practice, functional assessment represents a fundamental step in identifying limitations, establishing therapeutic interventions, and monitoring clinical progress. Traditionally, this assessment is performed using validated clinical instruments, such as the Timed Up and Go (TUG) test and muscle strength scales, which are widely used due to their practicality and low cost [4]. However, these methods have limitations, as they rely partly on the examiner's subjective observation and provide limited information regarding the quality of the movement performed [5].

In this context, the incorporation of technologies such as portable devices has expanded the possibilities for functional assessment [6]. Wearable sensors, particularly accelerometers, enable the real-time measurement of quantitative variables related to human movement, providing objective data on acceleration, symmetry, body sway, and motor performance during functional tasks such as upper limb elevation, cycling on a cycle ergometer, and gait. Thus, these technologies may contribute to more accurate, sensitive, and reproducible assessments in the older adult population [6], [4], [7] [8].

Despite the growing use of wearable sensors in human movement analysis, studies investigating the clinical applicability of accelerometer-based assessment in older adults assessed in a home environment remain scarce. Therefore, the present study aimed to analyze the clinical applicability of wearable sensors in the functional assessment of older adults in a home environment.

II. Materials and Methods

A. Study Design

This was a pilot study of technological applicability with a cross-sectional design and an observational approach applied to functional analysis using wearable sensors. The study was approved by the Human Research Ethics Committee of the Health Sciences Center at the Federal University of Pernambuco under approval number 5.135.351.

B. Volunteers

The sample consisted of five older adults aged 60 years or older, recruited by convenience sampling. The inclusion criteria comprised individuals of both sexes, aged 60 years or older, with preserved cognitive function as assessed by the Mini-Mental State Examination (MMSE), and with internet access at home. Individuals younger than 60 years, those with cognitive impairment identified by an MMSE score below 24 points, those with uncontrolled hypertension, and those who.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

used assistive walking devices that could interfere with locomotion and the use of the accelerometers were excluded.

C. Assessment Procedures

Data collection was conducted in the participants' homes to reproduce habitual functional conditions and enhance the clinical applicability of the assessment, in the city of Recife, Pernambuco, and its metropolitan region.

Initially, participants underwent sociodemographic and clinical characterization. Subsequently, functional assessment was performed using the Timed Up and Go (TUG) test to evaluate mobility and fall risk, the Medical Research Council (MRC) scale to assess muscle strength, and the Mini-Mental State Examination (MMSE).

Five wearable sensor modules equipped with MPU-9052 accelerometers were used. One sensor was secured on the dorsal surface of the right wrist and one on the dorsal surface of the left wrist. One sensor was positioned over the right lateral malleolus and one over the left lateral malleolus. The fifth sensor was placed on the anterior midline of the trunk, immediately inferior to the sternum. Each accelerometer recorded three orthogonal acceleration signals (X, Y, and Z), corresponding to the anteroposterior, mediolateral, and vertical directions, respectively. The acceleration signals were acquired at a sampling frequency of 50 Hz using an accelerometer measurement range of ±2 g.

Participants performed stationary walking, upper limb abduction, and cycling on a cycle ergometer for 30 seconds to 1 minute at a comfortable, self-selected pace. During these activities, accelerometer signals were continuously acquired. The accelerometers were positioned on the participants' ankles, wrists, and trunk, as shown in Fig. 1.

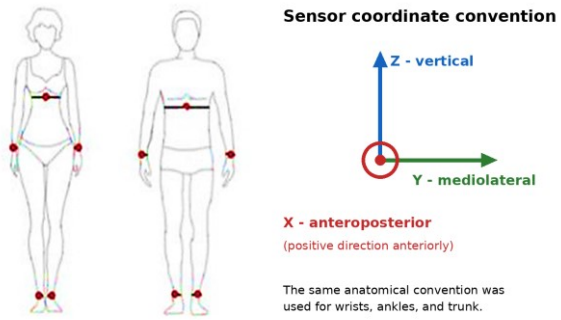


Fig. 1. Anatomical placement of the five wearable modules: dorsal surface of each wrist, over each lateral malleolus, and on the anterior midline of the trunk immediately inferior to the sternum. The anatomical coordinate convention was X = anteroposterior, Y = mediolateral, and Z = vertical.

D. Data Acquisition System and Signal Transmission

The data acquisition system was developed by the Biomedical Engineering Research Group (GPEB) at the Human-Machine Interface Laboratory (LIHOM), Federal University of Pernambuco (UFPE). It comprised five wearable inertial measurement modules, a fixed acquisition module based on a Raspberry Pi 4 Model B 8GB, and a web-based portal for data storage and monitoring.

Each wearable module incorporated an MPU-9052 inertial sensor with a triaxial accelerometer and a triaxial gyroscope. The accelerometer measurement range was configured to ±2 g and the gyroscope range to ±250°/s. Acceleration and angular velocity signals were sampled at 50 Hz, corresponding to a sampling interval of 20 ms.

Data transmission from the wearable modules was performed over a Wi-Fi network using the Message Queuing Telemetry Transport (MQTT) protocol. Each module published its inertial measurements within the same MQTT communication infrastructure. Acquisition of the five modules was initiated by a common trigger: a command published by the web-based portal to a dedicated MQTT topic and received by all modules. Upon receipt of this command, the modules initiated data acquisition, establishing a common temporal reference for the beginning of each recording. The acquired measurements were then transmitted to the remote server for real-time visualization and storage through the web portal.

E. Analyzed Variables

The acceleration signals obtained from the accelerometers during the execution of the proposed functional tasks were analyzed. Based on the collected data, the extraction of kinematic parameters related to linear acceleration, directions, and movement patterns was performed. The analyzed axes corresponded to anteroposterior (X), lateral (Y), and vertical (Z) displacements.

The signals were stored in the cloud for subsequent processing and analysis. Video recordings were also used as a complementary resource to verify movement execution and correlate it with the obtained kinematic data.

Although all participants completed the full functional protocol, a random selection procedure was used solely to determine the participant, functional task, and accelerometer axis to be presented as representative examples in the manuscript. All eligible participants, exercises, and axes were included in the selection process, and the final examples were determined by random draw to minimize selection bias.

No filtering was applied to the acceleration or angular velocity signals during the acquisition process. The signals were retained in their original acquired form to avoid introducing additional signal-processing effects that could modify their temporal or amplitude characteristics.

F. Statistical Analyses

The data were analyzed using descriptive statistics, applying measures of central tendency and dispersion, including mean and standard deviation for numerical variables, as well as absolute and relative frequencies for sample characterization. Graphical analyses of the accelerometer signals obtained during the functional tasks were also performed.

III. Results

A. Sample Characterization

Five eligible older adults were recruited for participation in the study, including three females and two males. The purpose is to emphasize the clinical applicability of the protocol using wearable sensors.

The sociodemographic and clinical characteristics of the sample are presented in TABLE I, as well as the categorical variables presented in TABLE II.

TABLE I. Sociodemographic and clinical characteristics of the presented volunteers.

ID	Age	Gender	Laterality	Occupation
NMJS	64	F	Right	Public servant
NMRN	62	M	Right	Public servant
ZCP	69	F	Right	Nutritionist
MMS	61	F	Right	Homemaker
LAS	67	M	Right	Retired

TABLE II. Categorical variables of the presented volunteers.

Variable	n (%)
Female	3 (60%)
Male	2 (40%)
Right	5 (100%)

B. Applicability of Wearable Sensors

The wearable sensors demonstrated adequate operational applicability, enabling the acquisition of accelerometer signals during all stages of the functional tests performed on the analyzed participants. No relevant complications related to device use were observed during data collection, such as participant discomfort, movement restriction, sensor displacement, or signal loss, demonstrating the feasibility of using these devices for functional assessment.

C. Analysis of Movement Signals

The obtained recordings enabled an objective analysis of movement, revealing individual variations in motor performance, differences in kinematic patterns, and the presence of asymmetries during the execution of the proposed functional tasks. The analyses presented below correspond to representative examples of the accelerometer signals obtained during the functional protocols performed by the participants.

The volunteer NMJS presented a regular movement pattern in the right arm during abduction, with mild oscillations and reduced amplitude at the end of the task. In the right leg, during stationary gait, repetitive peaks compatible with foot contact with the ground and the swing phase were observed.

The trunk maintained a stable signal with low variation. Quantitatively, the normalized RMS (nRMS) values were 0.454 for the right wrist X-axis, 0.511 for the right ankle Y-axis, and 0.391 for the trunk Z-axis, with corresponding normalized peak-to-peak amplitudes (nP2P) of 6.162, 4.893, and 3.479, respectively.

These accelerometer signals were consistent with the regular movement pattern and differences in signal amplitude observed during the functional tasks, as shown in Fig. 2.

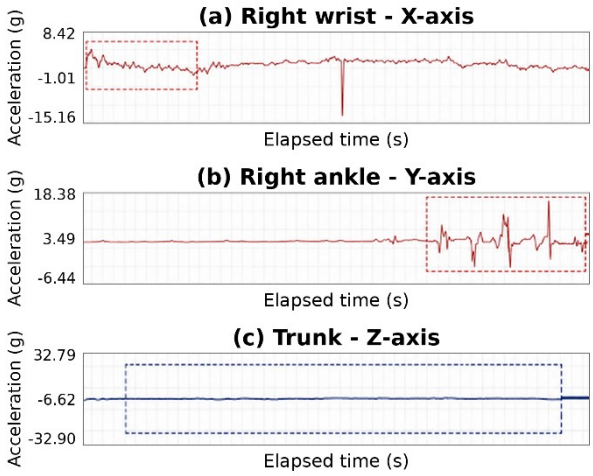


Fig. 2. Upper limb abduction and stationary gait exercises showing the X-axis of the right wrist, Y-axis of the right ankle, and Z-axis of the trunk.

Note: G (Acceleration); S (Time in Seconds)

In the accelerometer signal analysis of the left arm horizontal abduction exercise performed by volunteer NMRN, a pattern of regular medium-amplitude oscillations was observed on the X-axis, suggesting consistent movement execution. On the Y-axis, greater signal variability was identified compared with the other axes, while the Z-axis showed a regular acceleration pattern throughout the task. The quantitative metrics were consistent with these observations: nRMS was 0.215 on the X-axis, 0.559 on the Y-axis, and 0.215 on the Z-axis, while nP2P was 1.793, 2.794, and 1.495, respectively (Fig. 3).

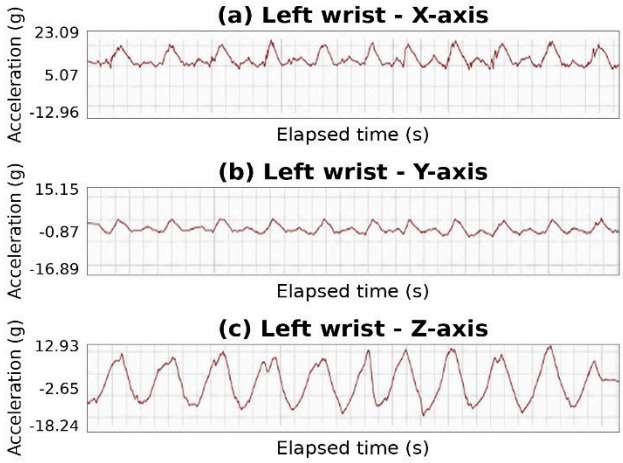


Fig. 3. Accelerometer signal analysis of volunteer NMRN – Horizontal Abduction Exercise (Left Arm).

In volunteer ZCP, a comparison of the accelerometer signals between the right and left lower limbs was performed. In the right leg, the X-axis showed higher-amplitude oscillations, while the Y-axis presented more defined and consistent peaks. On the Z-axis, oscillations of greater magnitude were observed compared with the contralateral limb.

In the left leg, the X-axis presented smooth oscillations and a regular cyclic pattern. On the Y-axis, slightly more pronounced variations were observed throughout the task, and on the Z-axis, lower-magnitude peaks were identified compared with the right leg. Quantitatively, bilateral symmetry indices based on nRMS were 17.6% for the X-axis, 34.4% for the Y-axis, and 5.7% for the Z-axis. Overall, both limbs presented cyclic behavior compatible with the proposed activity, while the quantitative metrics confirmed differences in signal amplitude between limbs (Fig. 4).

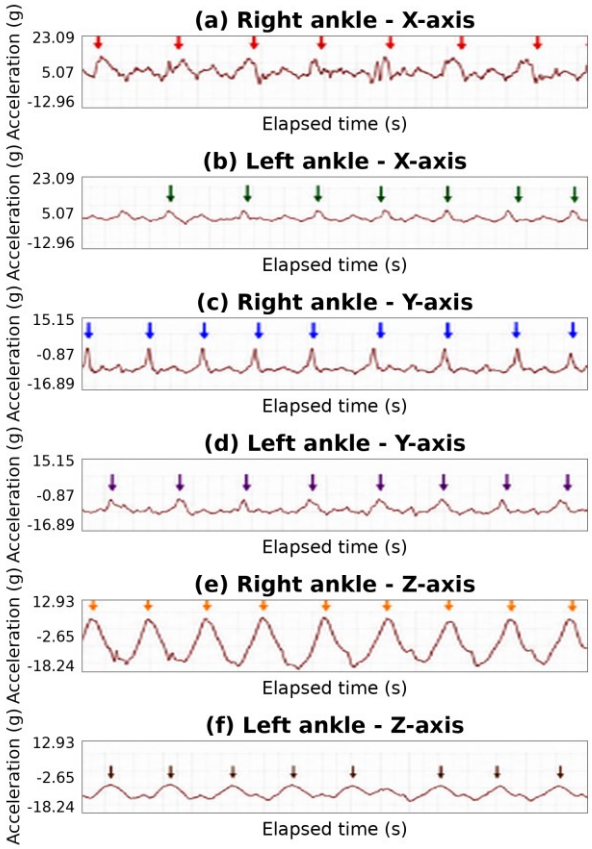


Fig. 4. Stationary gait exercise, right and left legs of volunteer ZCP, along the X, Y, and Z axes.

The analysis of the upper limb abduction movement of volunteer MMS showed that the left arm presented a more regular signal, with low amplitude and absence of pronounced peaks. In contrast, the right arm exhibited greater accelerometer signal variability and higher-amplitude oscillations. On the Y-axis, nRMS was 0.086 for the left wrist and 0.133 for the right wrist, with nP2P values of 0.521 and 0.786, respectively, resulting in a bilateral symmetry index of 43.3%. These quantitative differences were consistent with the observed difference in movement magnitude between limbs (Fig. 5).

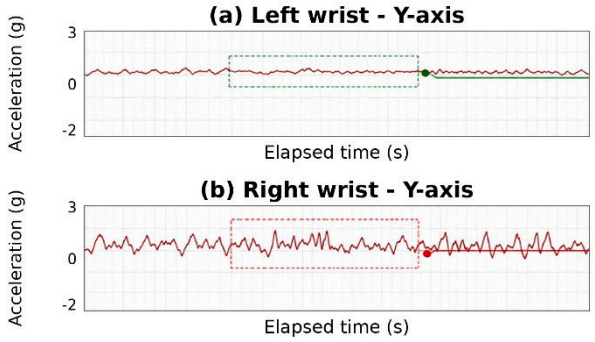


Fig. 5. Analysis of upper limb abduction movement (Y-axis) in volunteer MMS.

In the accelerometer signal analysis of walking along the Y-axis in volunteer LAS, alternating and subsequent patterns between the lower limbs were demonstrated, indicating an alternating pattern during gait execution. Stable temporal patterns were observed without abrupt variations or marked irregularities in the analyzed signals. Quantitatively, nRMS was 0.279 for the left ankle and 0.446 for the right ankle, while nP2P was 3.597 and 3.950, respectively, resulting in a bilateral symmetry index of 46.0%.

The small variations identified in the signals were compatible with postural adjustments, while the smooth transitions between steps indicated a regular and consistent temporal pattern during movement execution. The quantitative metrics additionally demonstrated a difference in acceleration amplitude between the lower limbs, as shown in Fig. 6.

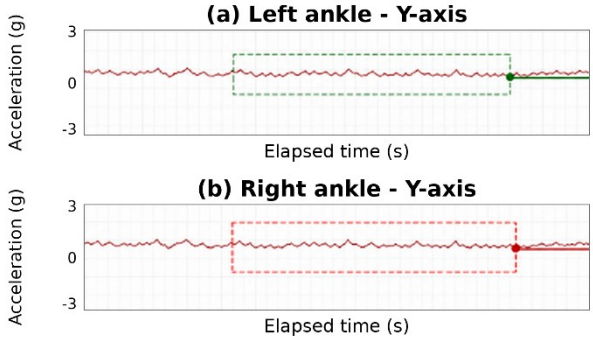


Fig. 6. Gait analysis – Accelerometer (Y-axis) in volunteer LAS.

Across the selected signals, normalized RMS ranged from 0.086 to 0.559 n.u. Bilateral symmetry indices ranged from 5.7% to 46.0%. In Table III, nRMS denotes normalized dynamic RMS, nP2P denotes normalized peak-to-peak amplitude, SI denotes the bilateral symmetry index calculated from the relative difference between left and right nRMS values, and L/R denotes left/right. An SI closer to 0% indicates more similar bilateral RMS amplitude; it does not by itself define normality, impairment, or unequal weight bearing.

TABLE III. Quantitative metrics of accelerometer signals during functional tasks.

Signal	nRMS (n.u.)	nP2P (n.u.)	SI (%)
NMRN wrist-X	0.215	1.793	-
NMRN wrist-Y	0.559	2.794	-
NMRN wrist-Z	0.215	1.495	-
ZCP ankles-X (L/R)	L 0.275 R 0.328	L 1.870 R 2.158	17.6
ZCP ankles-Y (L/R)	L 0.216 R 0.305	L 3.196 R 3.953	34.4
ZCP ankles-Z (L/R)	L 0.294 R 0.277	L 3.650 R 3.953	5.7
LAS ankles-Y* (L/R)	L 0.279 R 0.446	L 3.597 R 3.950	46.0*
NMJS right wrist-X	0.454	6.162	-
NMJS right ankle-Y	0.511	4.893	-
NMJS trunk-Z	0.391	3.479	-
MMS wrists-Y (L/R)	L 0.086 R 0.133	L 0.521 R 0.786	43.3

IV. Discussion

The results demonstrated that wearable sensors presented relevant clinical applicability in the functional assessment of older adults, enabling the objective analysis of parameters related to movement, kinematic patterns, and motor strategies. Accelerometer signals allowed the identification of differences in kinematic patterns, asymmetries, and variations in movement strategies during the functional tasks performed. The quantitative metrics, including nRMS, nP2P, and bilateral symmetry indices, complemented the signal-pattern observations and may assist physical therapists in identifying individual functional patterns and guiding more specific therapeutic interventions.

Michaelsen et al. [8] identified, using accelerometer signals, slower movements and longer deceleration times in individuals with hemiparesis. Similarly, the volunteers in this study presented differences between dominant and non-dominant limbs regarding movement regularity, amplitude, and impact, reinforcing the potential of accelerometer signals for functional kinematic analyses [9].

The asymmetries observed between the right and left legs in this study, during the analysis of volunteer ZCP, demonstrated differences in signal amplitude between limbs, with bilateral symmetry indices of 17.6%, 34.4%, and 5.7% for the X, Y, and Z axes, respectively. These findings corroborate previous studies that associate accelerometer oscillations with individual gait patterns and neuromuscular adjustments for maintaining postural stability [10], [11], [12], [13].

The analysis of acceleration signals across different axes enables the identification of variations in body movement during the execution of functional tasks, allowing the observation of anteroposterior, lateral, and vertical oscillations potentially related to motor control and functional trategies adopted by older adults, reinforcing the importance and applicability of accelerometer-based assessment in the functional and otor evaluation of this population.

Thus, the findings of this research corroborate the analyzed studies by indicating that differences in limb acceleration can be objectively characterized using accelerometer signals and quantitative metrics. Furthermore, they reinforce that accelerometers are useful tools for assessing motor functions and activities in different populations and environments.

The limitations of the study include the small sample size, characteristic of a pilot study, the absence of a control group and inferential statistical analyses, as well as the inherent variability of the home environment during data collection. However, the study demonstrates the feasibility and potential clinical applicability of accelerometers for the functional assessment of older adults in a home environment.

V. Conclusion

This study assessed the applicability of wearable accelerometers for functional evaluation of older adults in home settings, supporting their potential as a tool for ecologically valid, remote functional monitoring in this population.

References

[1] DOS SANTOS, C. S. et al. Capacidade funcional de idosos acompanhados pela Fisioterapia de uma Unidade de Atenção Primária à Saúde. Revista FisiSenectus, v. 7, n. 2, p. 23-38, 2019.

[2] OLIVEIRA, B. L. C. A. DE; PINHEIRO, A. K. B. Mudanças nos comportamentos de saúde em idosos brasileiros: dados das Pesquisa Nacional de Saúde 2013 e 2019. Ciência & Saúde Coletiva, v. 28, n. 11, p. 3111–3122, nov. 2023.

[3] NONATO DE LIMA, D. et al. Envelhecimento e Resistência Muscular: Impacto na Capacidade Funcional. Fiep Bulletin - online, v. 94, n. 1, p. 237–247, 30 jan. 2024.

[4] PEREZ, A. M. C. Interface homem-máquina: uma ferramenta de auxílio na reabilitação da marcha e do equilíbrio em pacientes pós-acidente vascular encefálico. Dissertação de mestrado - Universidade Federal de Pernambuco: [s.n.]2020.

[5] PILLON, B. P. C. Desenvolvimento de um jogo digital em realidade virtual para a reabilitação virtual do público sênior. Tese - Universidade Federal do Rio Grande do Sul: [s.n.] 2021.

[6] NOVELETTO, F.; BERTEMES, P.; HOUNSELL, M.S.; SOARES, A.V. Biomedical Control Interface for a Physical Rehabilitation Serious Game. Biomedical Control Interface, v. 14, n. 1, p. 38-44, 2016.

[7] GODFREY, A Wearables for independent living in older adults: Gait and falls. Maturitas, v. 100, p. 16–26, 2017.

[8] SILVA, M. B. C. Plataforma para Telereabilitação Motora: Auxílio de Tratamento de Pacientes Pós-Covid 19. Tese de doutorado - Universidade Federal de Pernambuco: [s.n.]2025.

[9] MICHAELSEN, S. M. et al. Usando um acelerômetro para analisar o movimento de alcance e preensão após o AVE. Motriz, v. 19, n. 4 p. x. 2013.

[10] BEDNARSKI , F. Análise Da Eficácia Do Uso De Acelerômetro Para Monitorar Os Movimentos Do Corpo Durante Exercícios Em Equipamentos Ao Ar Livre. dissertação—Universidade Estadual do Oeste do Paraná,: Programa de Pós-Graduação em Ciências Aplicadas à Saúde. Paraná 2024.

[11] SOUZA, M. O. DE et al. Impactos da COVID-19 na aptidão cardiorrespiratória: exercícios funcionais e atividade física. Revista Brasileira de Atividade Física & Saúde, v. 25, p. 1–5, 31 dez. 2020.

[12] OLIVEIRA, H. N. Compressão de Sinais Biomédicos. Dissertação de Mestrado (Engenharia Biomédica) Universidade Federal da Paraíba - UFPB - João Pessoa-PB, p. 12, 2019.

[13] M. SCHWENK, J. et al. Wearable sensor-based in-home assessment of gait, balance, and physical activity for discrimination of frailty status: baseline results of the Arizona frailty cohort study. *Gerontology. v.* 61, p. 258–267, n. 3, 2015.

Combined effects of brain-computer interface and virtual reality game with transcranial direct current stimulation in stroke patients

Pedro Felipe Giarusso de Vazquez
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
p263293@dac.unicamp.br

Karoliny Lisandra Teixeira Cruz
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
karolfisio6@gmail.com

Luara Inocêncio Pereira Silva
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
luarainocencio27@gmail.com

Eduardo Abreu da Silva
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
e247129@dac.unicamp.br

Gabriel Chaves de Melo
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
chaves1135@gmail.com

Amilton Lamas
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
amiltonlamas@gmail.com

Thiago Bulhões da Silva Costa
Escola Paulista de MedicinaUniversidade Federal de São Paulo – UNIFESP, São Paulo, Brazil
thiago.bulhoes@unifesp.br

Gabriela Castellano*
Institute of Physics Gleb Wataghin
University of Campinas - UNICAMP
Campinas, Brazil
gabriela@unicamp.br

***Abstract*—Motor imagery (MI) brain-computer interface (BCI), virtual reality (VR), and transcranial direct current stimulation (tDCS) have been investigated as complementary approaches for post-stroke motor rehabilitation. However, EEG longitudinal changes due to these approaches are still poorly understood. In this work longitudinal changes in motor-imagery spectral modulation were analyzed during a 10-session rehabilitation protocol combining MI-BCI neurofeedback, VR-based motor training, and anodal tDCS. Relative power modulation (RPM) during MI with respect to rest was computed per frequency band, hemisphere, and acquisition (pre- and post-intervention), and changes between the first and final sessions were assessed with Wilcoxon signed-rank tests under Bonferroni correction. The most consistent descriptive pattern was observed during post-intervention affected-hand MI in the lesioned hemisphere. All 10 participants exhibited a negative longitudinal change in the relative power modulation (RPM) associated with the β MI, while 9 of 10 showed the same direction in the θ rhythm. These exploratory findings are compatible with stronger event-related desynchronization (ERD) at the end of the protocol and motivate confirmation in larger, controlled studies.**

Keywords—brain-computer interfaces, virtual reality, transcranial direct current stimulation, stroke, rehabilitation

I. Introduction

Persistent upper-limb impairment is one of the most challenging consequences of stroke, particularly when residual voluntary movement is insufficient to support intensive task-oriented practice. In this context, neurorehabilitation approaches that engage motor-related brain activity without requiring complete movement execution, are especially relevant. Motor imagery (MI), defined as the mental simulation of an action without physically executing it, recruits premotor, parietal, and sensorimotor regions that partially overlap with those activated during actual movement [1]. MI can provide a pathway to activate motor representations in individuals with impaired physical movement.

Brain-computer interfaces (BCIs) are systems that translate neurophysiological activity into commands to control an external device. Electroencephalography-based brain-computer interfaces (EEG-BCIs) allow for the detection of MI-related neural activity and its conversion into real-time feedback. During movement and MI, sensorimotor rhythms commonly exhibit event-related desynchronization (ERD), characterized by a reduction in oscillatory power relative to a resting reference, particularly in the μ (8-12 Hz) and β (13-30 Hz) frequency bands. Nevertheless, some studies have shown event-related synchronization (ERS), i.e., band power increases, related to MI [2], [3]. These responses are characterized by frequency and spatial patterns and should be interpreted as changes in oscillatory organization rather than as direct measurements of cortical excitation or inhibition [4]. Thus, sensorimotor rhythms can be used as BCI markers.

The most important point about using MI-BCI in neurorehabilitation is that training can establish a closed sensorimotor loop between movement intention, brain activity, and an observable consequence. Controlled studies have reported motor improvements following BCI-assisted training in both subacute and chronic stroke populations [5], [6].

Nevertheless, EEG responses are heterogeneous after stroke. The magnitude and hemispheric distribution of ERD/ERS may depend on lesion location, impairment severity, task demands, and the participant's ability to generate a consistent MI response. MI-BCI training may be complemented by interventions that provide embodied motor practice and modulate cortical responsiveness. Virtual reality (VR) enables oriented exercises with multisensory feedback, which can be performed while employing transcranial direct current stimulation (tDCS). tDCS consists of applying a weak (1~2mA) electrical current through the scalp to modulate

This research was supported by Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) – Finance Code 001; São Paulo Research Foundation (FAPESP), Grants 2025/10963-8 (PFGV), 2025/11139-7 (KTC), 2024/13566-7 (EA), 2023/12422-9 (GCM), 2013/07559-3 (BRAINN), 2024/00998-6 (CRISQuaM); Brazilian National Council for Scientific and Technological Development (CNPq), Grants 306255/2024-3 and 408389/2024-9 (INCT-NeuroComp); and Funding Authority for Studies and Projects (FINEP), Grant 2140/22.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

cortical excitability. Rather than directly eliciting action potentials, the resulting electric field is able to produce subthreshold shifts in neuronal membrane polarization, thereby altering the likelihood of neuronal populations responding to ongoing synaptic input. Nevertheless, the physiological response remains incompletely characterized and varies considerably across individuals and stimulation protocols [7].

Fatigue may also influence the participant's ability to sustain MI and operate a BCI. Relative β power has been associated with mental fatigue and MI-BCI performance in stroke survivors, suggesting that changes in oscillatory modulation should be interpreted alongside clinical and subjective outcomes [8].

Previous studies have investigated MI-BCI with VR feedback and tDCS as post-stroke rehabilitation strategies, either separately or in partial combinations. However, less is known about how MI power modulation changes longitudinally when those approaches are integrated in a single study protocol. Therefore, this present study examined task and hemisphere-specific changes in RS EEG power during affected and non-affected hand MI through a 10-session intervention combining MI-BCI neurofeedback, VR motor training, and anodal tDCS in individuals with chronic stroke. Relative power modulation was examined during affected- and non-affected hand MI across lesioned and non-lesioned hemispheres. The relative predominance of each hemisphere. Changes from the first to the final session were evaluated separately before and after the intervention, and exploratory correlations were performed between EEG changes and clinical outcomes. This approach aimed to determine whether repeated training was accompanied by frequency-specific changes in MI-related modulation.

Thus, the novelty of this study lies not in the isolated use of BCI, VR, or tDCS, but in the longitudinal characterization of frequency, task, and hemisphere specific EEG modulation during a multimodal MI-BCI + VR + tDCS rehabilitation protocol.

II. Methods

A. Subjects

Ten adults with chronic stroke (mean 50±29 months post-stroke) were recruited in this study (mean age 46±9 years, range 34-68 years, 7 females). All subjects had a stroke episode more than six months before our clinical intervention, had no history of seizures, no pacemaker, or other metallic implants, and had not removed any part of the brain tissue. Their demographic and clinical characteristics are summarized in Table I.

A minimal voluntary motor activity in the affected upper limb was required to ensure that the subjects could engage with the proposed VR motor tasks.. This was operationally defined as I observable attempt or voluntary muscle contraction during hand opening or closing and/or the ability to initiatIan active reaching movement with the paretic arm.

Participants underwent 10 rehabilitation sessions combining MI-BCI neurofeedback, VR-based upper-limb training, and anodal tDCS. The project was approved by the ethics committee of Universidade Estadual de Campinas (CAAE 80319924.7.0000.5404) and all participants provided written consent before entering the study.

Table I
Patients Characteristics

Subject	Age	Sex	Affected Hemisphere	Months post-stroke
KG	49	F	Right	21
AV	37	F	Left	28
RN	43	M	Right	50
MM	44	F	Right	60
JG	45	F	Right	48
TC	34	F	Right	60
CN	41	F	Right	60
CS	50	F	Left	120
PC	68	M	Left	44
AS	53	M	Right	16

Subject codes represent anonymized participant identifiers.
F-Female; M-Male

B. Pre- and Post-Intervention Clinical Assessment

Participants were evaluated using clinical tests immediately before the first rehabilitation session and right after the last session. Both evaluations were conducted by the same physical therapist. Before enrollment, participants completed a medical screening questionnaire and the Mini-Mental State Examination (MMSE) to document medication use, verify eligibility according to inclusion and exclusion criteria, and assess preservation of cognitive function after stroke.

Clinical assessments were performed during the first and tenth sessions. These included manual dexterity tests, using the Nine-Hole Peg Test [9] and the Box and Block Test [10], followed by sensory assessments including tactile sensitivity with a monofilament, thermal sensitivity, pain/deep pressure sensitivity, stereognosis, and vibration sensitivity using a tuning fork. Muscle strength was then assessed using a hydraulic Jamar hand dynamometer for handgrip strength [11], as well as isometric elbow and wrist flexion and extension strength tests.

Finally, clinical scales were applied, including the upper-limb Fugl-Meyer Assessment [12], Modified Rankin Scale, Barthel Index [13], and Modified Ashworth Scale[14]. Participants also completed fatigue questionnaires, including the Modified Fatigue Impact Scale (MFIS) [15] and Fatigue Severity Scale (FSS)[16], to evaluate the impact of fatigue on daily activities.

C. Rehabilitation Protocol

As mentioned, participants completed 10 rehabilitation sessions conducted on consecutive weekdays, with no sessions scheduled on weekends. The first and final sessions lasted approximately 3-3.5 h because they included additional assessments, whereas Sessions 2-9 lasted approximately 1.5 h. Before the first acquisition, the MI task and the sequence of procedures were explained. Participants were seated comfortably with their forearms and hands supported on a pillow placed across their lap. They were instructed to remain relaxed and to perform kinesthetic MI of repeatedly opening and closing the cued hand without producing overt movement. In addition to visualizing the movement, they were encouraged to imagine the sensory and proprioceptive sensations associated with the movement.

The EEG cap was positioned while these instructions were provided. Participants were informed that EEG activity and BCI classification performance naturally vary across trials and that inconsistent neurofeedback did not necessarily indicate incorrect task execution. A brief practice period was conducted to ensure that the task and experimental cues were understood.

The acquisition sequence is illustrated in Figure 1. The trial structure was then presented to the participant: 6 s of rest, followed by a 2 s of a directional cue and 6 s of the MI task. Left- and right-hand trials were presented in a pseudorandomized order while maintaining an equal number of trials for each hand.

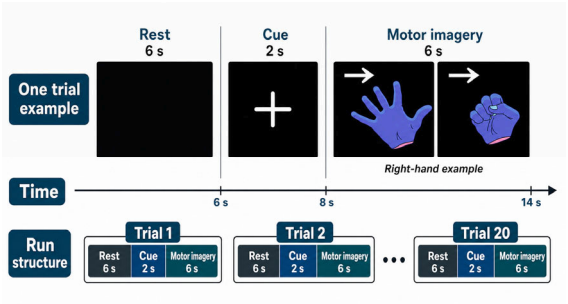


Figure 2. MI-BCI run structure. A neurofeedback run is illustrated in the example. Calibration runs had less trials (10 for the initial calibration and 6 trials for the pre and post tDCS calibrations).

The first session began with a calibration acquisition comprising 10 trials, with five trials assigned to each hand. This initial extended calibration was required to establish the participant-specific BCI model. From Session 2 onward, the data obtained in the preceding session, combined with a brief six-trial training, was used as the initial model of the day. This was followed by two neurofeedback runs comprising 20 trials each, equally divided between left- and right-hand imagery. A 3-min resting interval was provided between runs. Following the pre-intervention EEG assessment, the EEG equipment was removed, and the participant proceeded to the VR and tDCS stage. For this stage, the participant was seated in a chair with armrests while the tDCS electrodes were positioned. Instructions concerning the VR task and the movements required to control the game were provided during preparation. A brief guided familiarization was conducted using the VR headset, while the headset display was mirrored on a television to allow the research team to monitor and explain the upcoming activities.

Participants then completed 20 minutes of VR-based motor training combined with anodal tDCS. After completing the VR+tDCS intervention, participants were escorted back to the EEG room. The EEG equipment was repositioned, and a brief six-trial recalibration acquisition was performed to update the BCI model following the intervention. Two additional 20-trial neurofeedback runs were then completed, separated by a 3-min resting interval.

D. EEG Acquisition

EEG signals were acquired using a g.Nautilus PRO system (g.tec, Austria) with 32 hybrid electrodes operated in dry mode throughout all experimental sessions. The electrodes were arranged according to the montage illustrated in Figure 2, based on the 10-10 international system. The reference and ground electrodes were positioned over the right and left preauricular regions, respectively.

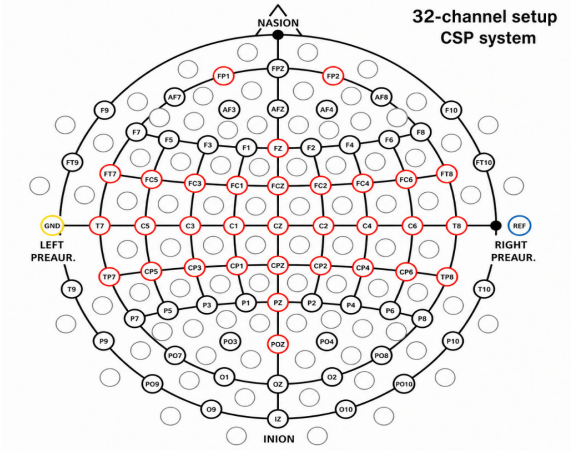


Figure 1. Electrodes used in the EEG acquisition. Red: Active electrodes; Yellow: Ground electrode; Blue: Reference electrode.

Signals were sampled at 250 Hz and recorded using the amplifier's analog band-pass filters between 0.5 and 100 Hz. Communication between the amplifier and the acquisition computer was implemented through the g.tec API, 2024 release, in MATLAB R2023a (MathWorks). Event markers identifying rest, directional cues, and MI periods were recorded synchronously with the EEG signal.

Because impedance measurements obtained from dry electrodes may not be a reliable indicator of electrode-skin contact, signal quality was primarily evaluated through visual inspection of the continuous EEG traces. Before the acquisition, participants were asked to alternate between eyes-open and eyes-closed periods and to briefly clench the jaw to generate chewing artifacts. Channels exhibiting unstable, saturated, or non-physiological signals were adjusted before data collection.

The same signal-quality procedure was repeated after the VR+tDCS intervention. Particular attention was given to amplitude attenuation or possible electrical bridging caused by residual saline from the tDCS electrodes. When these effects were identified, the scalp was dried. EEG acquisition was initiated only after the signals were considered adequate.

E. Online MI-BCI Neurofeedback

The online MI-BCI system was implemented in MATLAB R2023a using the g.tec API and Psychtoolbox. Following the amplifier-level filtering described in Section D, signals from all 32 EEG channels shown in Figure 2 were used for online processing. Before each training or neurofeedback acquisition, the amplifier streamed data for 10 s while the participant remained at rest. These initial samples were discarded to reduce the influence of transient amplifier instability.

Each neurofeedback run comprised 20 trials, equally divided between left- and right-hand MI. Trial order was pseudorandomized while allowing no more than two consecutive trials from the same class. Each trial comprised 6 s of rest, a 2-s preparatory fixation cue, and 6 s of kinesthetic motor imagery. At MI onset, a directional arrow

and an animation of the corresponding hand opening and closing were presented.

For classifier training, the 6-s MI periods were labeled as left-hand class (0) or right-hand class (1) and divided into 2-s windows with a 1-s step, corresponding to 50% overlap. Each window was processed using a bank of six causal Chebyshev Type II band-pass filters: 8-12, 12-16, 16-20, 20-24, 24-28, 28-32 Hz.

For each frequency sub-band, a normalized covariance matrix was calculated for every temporal window:

$$C_i = \frac{X_i X_i^T}{tr(X_i X_i^T)}$$

where X_i represents the multichannel EEG data from window i. Covariance matrices were then averaged separately for the left- and right-hand classes.

To improve numerical stability, each class covariance matrix was regularized using shrinkage toward a scaled identity matrix:

$$\tilde{C}_k = (1-\lambda)C_k + \lambda \frac{tr(C_k)}{p} I$$

where p is the number of EEG channels and $\lambda = 0.3$.

Common Spatial Patterns (CSP) filters were estimated by whitening the composite covariance matrix and solving the corresponding eigenvalue decomposition [17]. The spatial filters were ordered according to their eigenvalues, with filters at opposite extremes maximizing variance for one MI class while minimizing it for the other. Two spatial-filter pairs were retained from each frequency sub-band: the first two and last two CSP components. The variance of each spatially filtered component was calculated and logarithmically transformed. This produced four log-variance features per sub-band and 24 features in total. All features were retained; no additional feature-selection procedure was applied.

Before classification, features were standardized using the mean and standard deviation estimated from the training data. A linear support vector machine (SVM) was trained to discriminate left- from right-hand MI. Posterior probability calibration was applied using Platt scaling [18] when at least 30 training windows were available.

F. VR-Based Motor Training Combined with tDCS Neuromodulation

A custom virtual-reality (VR) game was developed to provide an engaging environment for repetitive upper-limb motor training. The application was developed using Unity and presented through Meta Quest 3. Hand and upper-limb movements were captured using the hand-tracking system illustrated in Figure 3.

Before training, the affected side was selected in the game settings according to the participant's paretic side. An individual calibration procedure was then performed to estimate the participant's minimum and maximum active movement amplitudes. These values were used to adapt the game's interaction thresholds and visual feedback to the participant's possible motor range. The calibration assessed opening and closing hand movements.

While the tDCS electrodes were positioned, participants received an initial explanation of the virtual environment and its objectives. They were then allowed to briefly explore the game before receiving additional task-specific instructions.



Figure 3. Participant exploring the VR-based motor training game in Session 1.

The primary task consisted of reaching toward a virtual "magical seed", grasping it, and transporting it to a vase located at the center of the virtual environment. Seeds appeared pseudo-randomly at left, upper-central, or right spatial locations, encouraging reaching in different directions. When the seed entered the vase's target area, the vase became active. Participants were then instructed to open the affected hand. A virtual flower emerged from the vase, with its growth determined by the hand-opening amplitude relative to the calibration range. Flower type was selected randomly from a predefined pool. Participants subsequently closed their hand before beginning the next reaching cycle. Thus, each cycle combined reaching, object transport, hand opening, and hand closing. The number of completed movement cycles and successfully bloomed flowers was displayed within the participant's field of view. Game difficulty increased progressively by introducing multiple colored seeds. At this stage, the vase indicated a target color, and participants were required to select the seed with the corresponding color.

Participants completed 20 minutes of VR-based upper-limb training. The game recorded movement repetitions, flower count, and hand tracking coordinates.

VR-based motor training was performed concurrently with anodal transcranial direct current stimulation (tDCS). Stimulation was delivered for 20 minutes using a MicroEstim tDCS stimulator (NKL, Brazil) at an intensity of 1 mA, with 30-s ramp-up and ramp-down periods. The anode was positioned over the ipsilesional primary motor cortex at C3 or C4, according to the international 10–20 system, while the return electrode was placed over the contralateral supraorbital region at Fp2 or Fp1, respectively. Anode and cathode consisted of 35-cm² electrodes inserted into saline-soaked sponge pockets. The sponges were evenly soaked without oversaturation to prevent saline leakage and maintain stable skin contact throughout stimulation. Electrode–skin impedance was monitored and maintained below 40 kΩ.

Stimulation began simultaneously with game onset and continued during the VR exercise. Participants were asked about discomfort, itching, burning, headache, or other sensations. No independent effects of VR or tDCS were estimated for this group of participants.

G. Offline EEG Preprocessing and Signal Processing

Offline EEG preprocessing was performed in MATLAB 2023a using EEGLAB [19], whereas subsequent spectral analyses were conducted in Python. Continuous EEG

recordings were imported separately for each participant, session, and run. Event markers corresponding to rest, cue, and MI were reconstructed from the recorded trial-order vectors. Electrode coordinates were assigned according to the 32-channel montage illustrated in Figure 2. The initial 6 s rest period of the first trial was excluded because it overlapped were excluded from baseline-dependent analyses. Line-noise contamination was attenuated using Zapline-plus [20], and the continuous signals were band-pass filtered between 1 and 50 Hz. Transient high-amplitude artifacts were attenuated using artifact subspace reconstruction, with a burst criterion of 20 and burst rejection disabled [21]. Automatic channel-rejection criteria were disabled. Signals were then re-referenced to the common average. Extended Infomax independent component analysis was performed [22]. Independent components were classified using ICLabel [23]. Components were removed when the probability assigned to muscle, eye, cardiac, line-noise, or channel-noise activity was at least 0.70. After preprocessing, the continuous recordings were segmented according to the experimental structure into 6 s rest periods, 2 s cue, and 6 s MI.

For offline analysis, each complete 14 s trial was filtered using a fourth-order Butterworth filter applied in the forward and reverse directions in the δ (1-4 Hz), θ (4-8 Hz), μ (8-12 Hz), and β (13-31 Hz) bands. μ and β were examined as motor-related frequency bands, whereas δ and θ were included as secondary exploratory analyses. Band-specific power was estimated by squaring the filtered signal and applying a 250 ms moving-average smoothing window. For each trial, electrode, and frequency band, baseline power (P_{REST}) was defined as the mean power across the corresponding 6 s rest period, whereas MI power (P_{MI}) was calculated as the mean power across the 6 s imagery period. MI-related relative power modulation (RPM) was estimated following the ERD/ERS quantification proposed by Pfurtscheller and Lopes da Silva [4]:

$$RPM = 100 \times \frac{P_{MI} - P_{REST}}{P_{REST}}$$

Under this convention, negative values indicate ERD, whereas positive values indicate ERS. Feature-level observations were excluded when baseline, MI power, or RPM values were identified as outliers according to either the 1.5 interquartile-range criterion or a four-median-absolute-deviation criterion, calculated within each participant, session, run, frequency band, and electrode. Three task scenarios were examined: all MI trials combined, affected-hand MI trials, and non-affected-hand MI trials. Trial-level values were first averaged for each electrode and run. These run-level estimates were subsequently averaged across Runs 1 and 2 to obtain the pre-intervention (pre-tDCS+VR) value and across Runs 3 and 4 to obtain the post-intervention (post-tDCS+VR) value. This aggregation was performed separately for each participant, session, task scenario, frequency band, and hemisphere, with estimates additionally averaged across the five sensorimotor electrodes assigned to each hemisphere: C3-C4, FC3-FC4, C5-C6, C1-C2, and CP3-CP4. Data from participants with left-hemisphere lesions were logically mirrored so that the same canonical side consistently represented the non-lesioned hemisphere across participants.

Longitudinal changes Δ in RPM were calculated as the difference between the final assessment and Session 1. For with the filtering transient response. The acquisition routine was subsequently corrected for two participants (PC and AS), allowing filter stabilization to occur before experimental data recording; therefore, their initial rest periods were retained. Consequently, trials without a valid preceding rest baseline

most participants, the final session corresponds to Session 10. Because of participant availability, subjects AV, KG, and PC completed their final assessment during Session 9.

H. Statistical Analysis

Longitudinal changes were calculated as $\Delta = \text{Final session} - \text{First session}$. For each RPM combination, two-sided Wilcoxon signed-rank test was utilized to assess whether the participant-level longitudinal changes differed from zero. Bonferroni correction across all 32 RPM tests (two task scenarios, four frequency bands, two hemispheres, two intervention conditions – pre and post) was applied. The resulting adjusted significance threshold was $\alpha = 0.05/32 = 0.0015625$.

I. Exploratory correlations with clinical scales

Exploratory Spearman correlations were calculated between eight longitudinal clinical outcomes and 32 EEG features, resulting in 256 comparisons. Clinical changes comprised grip strength, elbow flexion and extension strength, Fugl-Meyer uppe-limb motor score change, and FSS.

III. Results

Longitudinal changes calculation ($\Delta = \text{Final session} - \text{First session}$) was performed separately for the pre- and post-intervention acquisitions, task scenarios, frequency bands, and hemispheres. For RPM, negative Δ values indicate that RPM became more negative in the final session, corresponding to stronger ERD or weaker ERS under the adopted convention.

After Bonferroni correction, no longitudinal RPM comparison reached statistical significance. However, a consistent descriptive RPM pattern was observed in the post-intervention acquisition during affected-hand MI, illustrated in Figure 4. Specifically, in the β band, all 10 participants exhibited negative ΔRPM values, with a median change of -2.57 percentage points (mean = –3.89; uncorrected p = 0.00195). Thus, RPM became more negative in the final session, indicating a stronger β ERD or weaker β ERS. No comparably consistent β band effect was observed in the pre-intervention acquisition. A second descriptive pattern was found in the θ band for the same task, hemisphere, and intervention condition. Nine of the 10 participants exhibited negative longitudinal changes, with a median ΔRPM of –9.39 percentage points (p = 0.00586). Regarding associations with clinical and functional assessments, the strongest descriptive correlations in sensorimotor band are shown in Figure 5 to illustrate possible trends. Greater non-affected-side elbow-flexion strength gains were associated with more positive β ΔRPM in the affected hemisphere during pre-intervention non-affected-hand MI (ρ = 0.88, uncorrected p = 0.0008). Greater affected-side elbow-flexion strength gains were associated with more positive μ ΔRPM in the non-affected hemisphere during pre-intervention non-affected-hand MI (ρ = 0.83, uncorrected p = 0.0029), and with more positive β ΔRPM in the affected hemisphere during pre-intervention

affected-hand MI (ρ = 0.77, uncorrected p = 0.0092). Greater affected-side elbow-extension strength gains were associated with more negative β ΔRPM in the non-affected hemisphere during post-intervention non-affected-hand MI (ρ = -0.76, uncorrected p = 0.0111). Greater FSS improvement, reflecting reduced fatigue severity, was associated with more negative β ΔRPM in the non-affected hemisphere during post-intervention affected-hand MI (ρ = -0.79, uncorrected p = 0.0119). Finally, greater non-affected-side elbow-flexion gains were associated with more positive μ ΔRPM in the non-affected hemisphere during pre-intervention non-affected-hand MI (ρ = 0.75, uncorrected p = 0.0133). None of these associations survived correction for multiple comparisons; therefore, they should be interpreted as exploratory.

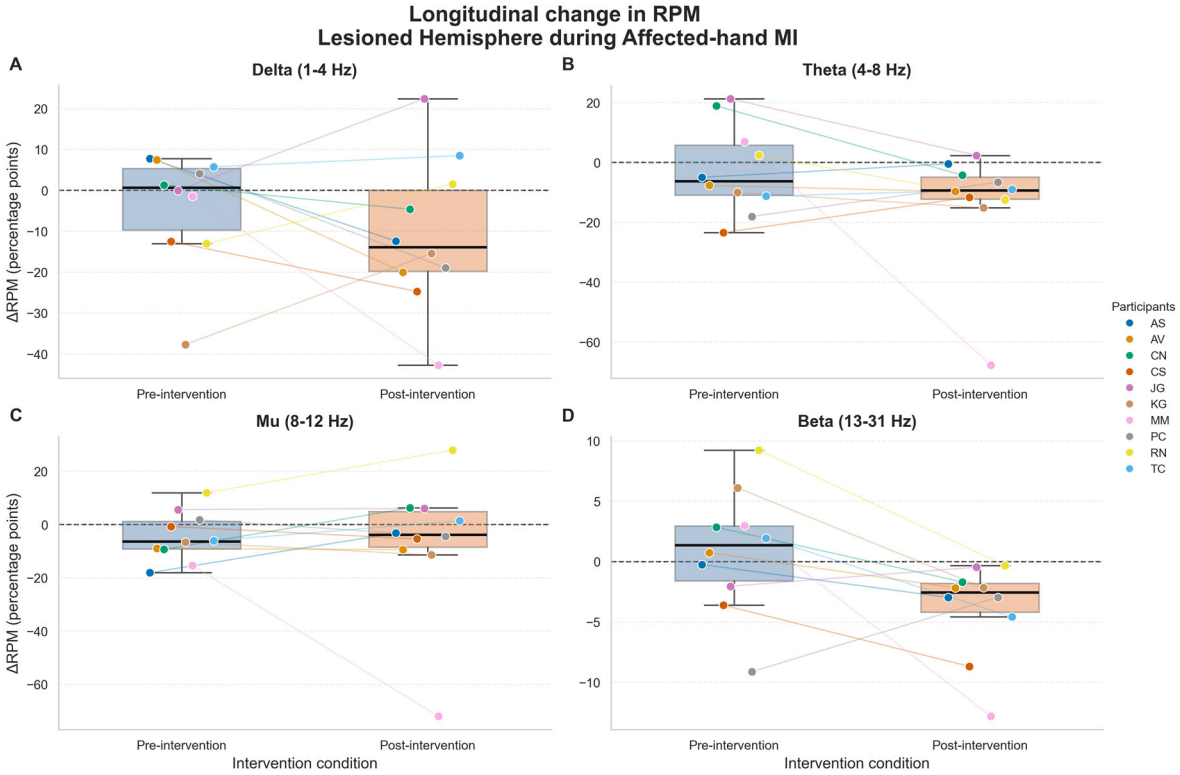


Figure 4. Longitudinal changes in relative power modulation (RPM) during affected-hand MI over the lesioned hemisphere. Values represent ΔRPM = final session – first session and are shown separately for the pre- and post-intervention acquisitions in the δ(A), θ(B), μ (C), and β (D). Boxes summarize the participant distribution, colored points represent individual participants, and connected lines link the two intervention conditions.

IV. Discussion

The present study investigated longitudinal changes in MI-related EEG activity following a 10-session multimodal rehabilitation protocol combining MI-BCI neurofeedback, VR-based motor training, and anodal tDCS.

Two frequency-specific trends were particularly consistent during post-intervention affected-hand MI over the lesioned hemisphere: all participants exhibited negative longitudinal β ΔRPM values, while 9/10 showed the same direction in θ. Under the adopted RPM convention, more negative values are compatible with stronger ERD or weaker ERS during MI relative to rest [4]. However, because RPM is normalized to resting power, this pattern cannot be attributed exclusively to reduced MI-period power; changes in resting power may also have contributed. No RPM comparison remained statistically significant after Bonferroni correction for multiple testing. Thus, these results are better described as longitudinal changes in task-related power modulation rather than direct evidence of improved motor recovery.

The β-band pattern is physiologically plausible because sensorimotor β desynchronization is commonly associated with motor preparation, execution, and imagery, and MI-related desynchronization has been associated with corticospinal excitability [24]. It is also broadly compatible with studies showing that BCI-based motor training can promote functional improvement after stroke [5], [6], [25]. However, stronger ERD should not be equated automatically with improved function or neural recovery. Chen and colleagues found that relationships between motor impairment and ERD/ERS depended on task and hemisphere; during MI, Fugl-Meyer scores were associated with contralesional ERS rather than simply with stronger ERD [26]. Accordingly, the present β result is better interpreted as altered task-related modulation than as restored motor activity.

The θ-band finding should be interpreted more cautiously. θ activity is less established as a direct sensorimotor recovery marker and may reflect attention, cognitive control, action-feedback integration, or coordination within distributed brain networks. In MI-BCI paradigms, θ coupling has been associated with the ability to modulate β rhythms and with sensory feedback modality [27], while slow cortical rhythms have also been implicated in MI-based BCI control [28].

Therefore, the θ pattern may indicate altered engagement with MI and multisensory feedback, but it should remain an exploratory observation.

The exploratory EEG-clinical correlations provide preliminary context but do not support physiological or clinical conclusions. The strongest uncorrected associations involved elbow flexion, elbow extension, and fatigue; none survived correction for the 256 exploratory comparisons, and the smallest Bonferroni-adjusted p-value in the entire family was 0.21. They are therefore reported as descriptive observations only, and the associations in Figure 5 were selected solely as the largest uncorrected coefficients within the sensorimotor bands.

The decision to examine imagined hand and hemisphere separately follows previous evidence that MI after stroke engages distributed, task-dependent sensorimotor networks. De Vico Fallani et al. showed that affected-hand MI after unilateral stroke altered β-band network topology compared with unaffected-hand MI, and that sensorimotor function was predicted by affected-hemisphere connectivity [29]. Similarly, Mansour et al. reported that both ipsilesional and contralesional hemispheres can express ERD/ERS during impaired hand MI, with contralesional signals being especially relevant in more impaired patients [30]. These studies motivated the factorial structure of the present analysis, but they cannot be used to interpret the individual associations reported previously. Our results identify candidate relationships for future studies with larger samples.

These findings may also have limited but useful implications for future adaptive closed-loop neurorehabilitation systems. Although variance- or covariance-based features generally provide stronger MI-BCI classification performance, longitudinal band-specific RPM profiles may provide complementary information about each participant's ability to modulate task- and hemisphere-specific EEG activity across sessions. If replicated, such measures could help monitor interhemispheric modulation over time, guide recalibration, or identify participants who may require alternative MI strategies. This application remains speculative and requires validation in larger controlled studies.

Clinical plausibility is reported by randomized studies reporting benefits from combining VR with tDCS [31], [32]. Nevertheless, the present design cannot determine whether the observed EEG patterns originated from MI-BCI practice, VR motor exercise, tDCS, repeated task exposure, or their interaction at this moment. A sham group is being formed at the present moment. The small sample size is an important limitation of this study. Although n=10 is comparable to several exploratory EEG-BCI neurorehabilitation studies involving repeated sessions and multimodal assessments, it limits statistical power and increases uncertainty about the effect size of our intervention. For this reason, this study should be interpreted as an exploratory analysis. Regarding the intervention, the mechanisms of tDCS are not yet fully understood. The results depend on several anatomical parameters, such as the thickness of the skull and cerebrospinal fluid layer. Consequently, some techniques may need to be individualized.

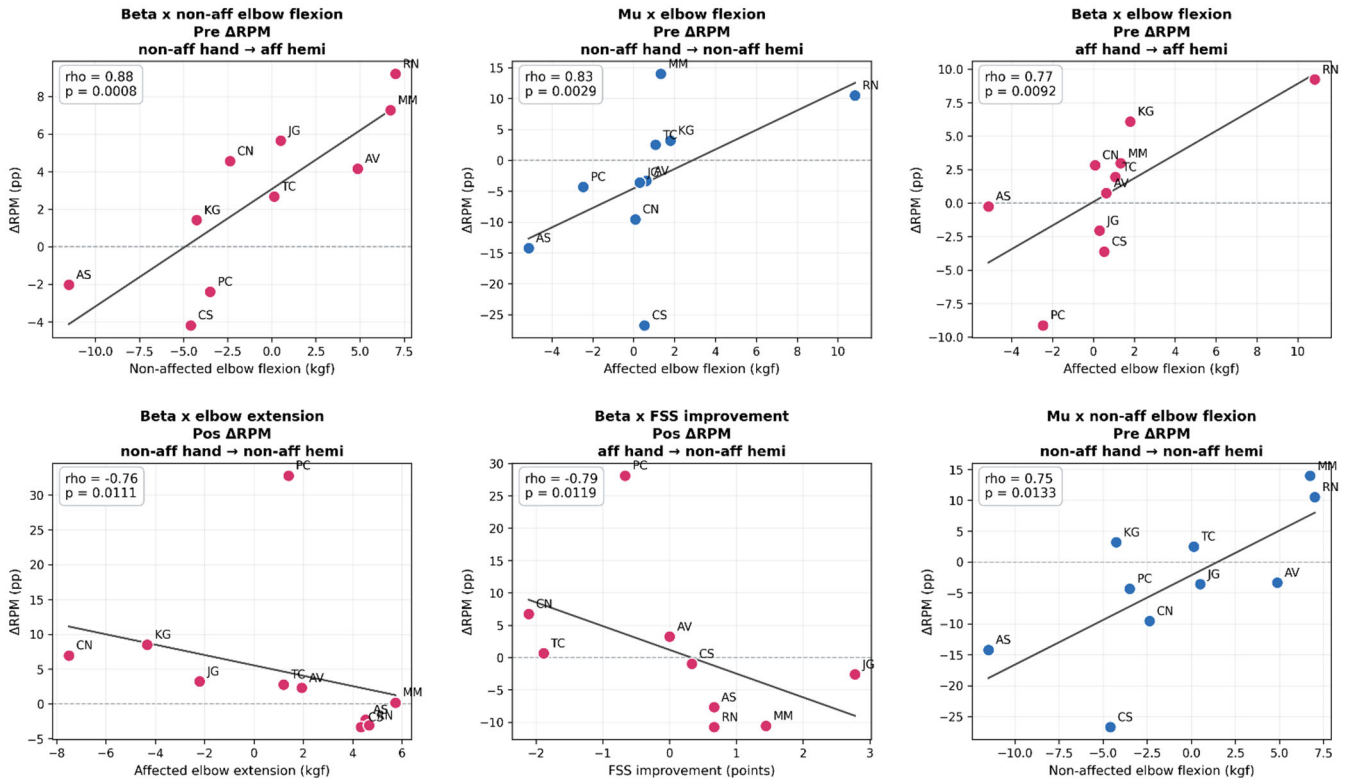


Figure 5. Scatterplots show the strongest uncorrected Spearman correlations between ΔRPM and changes in strength or fatigue. ΔRPM was calculated as final session minus first session and is shown for the indicated acquisition condition, MI task, and hemisphere. Positive clinical change indicates strength gain, whereas positive FSS improvement indicates reduced fatigue severity. These associations did not survive correction for multiple comparisons and should be interpreted as exploratory.

V. Conclusion

This exploratory study investigated longitudinal changes in motor-imagery-related EEG activity during a 10-session intervention combining MI-BCI neurofeedback, VR-based motor training, and anodal tDCS in individuals with chronic stroke. The most consistent descriptive finding was a shift toward more negative RPM during affected-hand MI in the lesioned hemisphere, particularly in the β and θ bands. This pattern is compatible with stronger ERD in the final session, although it does not independently demonstrate improved MI ability or motor recovery. Larger controlled studies are needed to determine whether these changes are reproducible, clinically meaningful, and attributable to specific components of the combined intervention.

Although variance or covariance features have generally achieved stronger MI-BCI classification performance, the present exploratory findings suggest that band-specific power modulation may provide complementary information for adaptive neurorehabilitation. If replicated, individualized β and θ-band RPM profiles could help identify participants who consistently modulate task and hemisphere specific activity across sessions.

Acknowledgments

The authors would like to thank Julia Macedo, Kevin Souza, Ana Lívia Barreto, Bianca Souza, Samira Alves, Maria Alves and Lia Menezes for their assistance with data collection and pilot procedures.

References

[1] S. Hétu et al., "The neural network of motor imagery: An ALE meta-analysis," 2013.

[2] G. Pfurtscheller et al., "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," Neuroimage, vol. 31, no. 1, 2006.

[3] C. S. Nam et al., "Movement imagery-related lateralization of ERD/ERS: Motor-imagery duration effects," Clin. Neurophysiol., vol. 122, no. 3, 2011.

[4] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," Clin. Neurophysiol., 1999.

[5] A. Ramos-Murguialday et al., "Brain-machine interface in chronic stroke rehabilitation," Ann. Neurol., vol. 74, no. 1, 2013.

[6] F. Pichiorri et al., "Brain-computer interface boosts motor imagery practice during stroke recovery," Ann. Neurol., vol. 77, no. 5, 2015.

[7] S. Wiethoff, M. Hamada, and J. C. Rothwell, "Variability in response to tDCS of the motor cortex," Brain Stimul., vol. 7, no. 3, 2014.

[8] R. Foong et al., "EEG-based MI-BCI, visual feedback, and mental fatigue in upper-limb stroke rehabilitation," IEEE Trans. Biomed. Eng., vol. 67, no. 3, 2020.

[9] N. Watanabe et al., "Reliability of the Modified Nine Hole Peg Test in stroke," Prog. Rehabil. Med., vol. 7, 2022.

[10] G. Everard et al., "Immersive VR Box and Block Test validity in stroke," J. Neuroeng. Rehabil., vol. 19, no. 1, 2022.

[11] I. Gatt et al., "The Takei Handheld Dynamometer," Hand, vol. 13, no. 3, 2018.

[12] J. Sanford et al., "Reliability of the Fugl-Meyer assessment after stroke," Phys. Ther., vol. 73, no. 7, 1993.

[13] D. Barthel and F. Mahoney, "Functional evaluation: The Barthel Index," Md. State Med. J., vol. 14, 1965.

[14] R. W. Bohannon and M. B. Smith, "Interrater reliability of a modified Ashworth scale," Phys. Ther., vol. 67, no. 2, 1987.

[15] J. D. Fisk et al., "Measuring the functional impact of fatigue," Clin. Infect. Dis., vol. 18, 1994.

[16] O. Ozyemisci-Taskiran et al., "Validity and reliability of fatigue severity scale in stroke," Top. Stroke Rehabil., vol. 26, no. 2, 2019.

[17] K. K. Ang et al., "Filter Bank Common Spatial Pattern in BCI," Proc. IJCNN, 2008.

[18] J. Platt, "Probabilistic outputs for support vector machines," Adv. Large Margin Classifiers, vol. 10, no. 3, 1999.

[19] A. Delorme and S. Makeig, "EEGLAB," J. Neurosci. Methods, vol. 134, no. 1, 2004.

[20] M. Klug and N. A. Kloosterman, "Zapline-plus," Hum. Brain Mapp., vol. 43, no. 9, 2022.

[21] T. R. Mullen et al., "Real-time neuroimaging and cognitive monitoring using wearable dry EEG," IEEE Trans. Biomed. Eng., vol. 62, no. 11, 2015.

[22] T. W. Lee, M. Girolami, and T. J. Sejnowski, "Extended Infomax ICA," Neural Comput., vol. 11, no. 2, 1999.

[23] L. Pion-Tonachini et al., "ICLabel," Neuroimage, vol. 198, 2019.

[24] K. Takahashi et al., "Corticospinal excitability and MI-related EEG desynchronization," J. Neuroeng. Rehabil., vol. 15, no. 1, 2018.

[25] A. Biasiucci et al., "Brain-actuated FES elicits lasting arm motor recovery after stroke," Nat. Commun., vol. 9, no. 1, 2018.

[26] S. Chen et al., "Sensorimotor rhythm and upper-limb motor impairment in stroke," Clin. EEG Neurosci., vol. 53, no. 3, 2022.

[27] M. Vukelić and A. Gharabaghi, "MI oscillatory entrainment is modulated by feedback modality," Neuroimage, vol. 111, 2015.

[28] J. A. Barios et al., "Slow cortical rhythms during MI-based BMI control," Int. J. Neural Syst., vol. 29, no. 5, 2019.

[29] F. De Vico Fallani et al., "Functional brain networks during motor imagery after stroke," Neuroimage, vol. 83, 2013.

[30] S. Mansour et al., "Contralesional hemisphere control of BCI in stroke survivors," Sci. Rep., vol. 12, no. 1, 2022.

[31] X. Yao et al., "tDCS with VR for upper-limb function after ischemic stroke," J. Neuroeng. Rehabil., vol. 17, no. 1, 2020.

[32] R. Llorens et al., "tDCS and VR intervention for upper-limb function after chronic stroke," J. Neuroeng. Rehabil., vol. 18, no. 1, 2021.

Structural and Mechanical Assessment of Glass and Carbon Fiber-Reinforced Composite Laminates for Wheelchair Seat Applications

Camila A. Diniz
Lightweight Structures Laboratory (LEL)
Institute for Technological Research (IPT)
São José dos Campos, Brazil
0009-0001-8269-6229

Paul L. Ragueneau
Lightweight Structures Laboratory (LEL)
Institute for Technological Research (IPT)
São José dos Campos, Brazil
0009-0008-9617-254X

Jefferson P. Nascimento
Lightweight Structures Laboratory (LEL)
Institute for Technological Research (IPT)
São José dos Campos, Brazil
0000-0002-8756-011X

Mário H. F. Batalha
Lightweight Structures Laboratory (LEL)
Institute for Technological Research (IPT)
São José dos Campos, Brazil
0000-0003-0053-2667

Alessandro Guimarães
Lightweight Structures Laboratory (LEL)
Institute for Technological Research (IPT)
São José dos Campos, Brazil
0000-0002-1459-6058

***Abstract*— The advancement of assistive technologies has increased the demand for lightweight and structurally efficient materials for rehabilitation devices. In this context, this study presents a comparative investigation of carbon fiber- and glass fiber-reinforced composite laminates for rigid wheelchair seat applications. Mechanical characterization was performed through tensile, compressive, and shear testing, and the resulting material properties were integrated into an analytical design approach based on Classical Laminate Theory (CLT) and further assessed through finite element analysis (FEA). The analytical and numerical results indicated that structural stiffness, rather than material strength, governed the seat design under the prescribed static loading conditions. Both laminate configurations satisfied the adopted deflection criterion and exhibited Tsai-Wu failure indices below the failure threshold, suggesting satisfactory structural performance for the loading scenario considered. The carbon fiber-reinforced laminate achieved the lowest maximum deflection (1.99 mm), together with reduced laminate thickness and lower structural mass, whereas the glass fiber-reinforced laminate represented a more cost-effective alternative based on estimated raw material costs while also satisfying the adopted design criterion. The main contribution of this work is a structured comparison between two composite laminate systems through an integrated experimental-CLT-FEA framework for preliminary structural assessment of wheelchair seat designs. Overall, the results indicate that both laminate configurations are promising candidates for the static loading scenario analyzed. Further studies involving full-scale seat testing, realistic fixation conditions, fatigue, impact, seat-cushion interaction, and user-centered evaluations are necessary before practical implementation can be assessed.**

Keywords—composite material, wheelchair seat, carbon fiber, glass fiber.

I. INTRODUCTION

In recent years, the use of composite materials for wheelchair components has increased significantly. This growing interest is primarily driven by their unique mechanical and physical properties. Composites consist of a matrix phase and a reinforcement phase that are engineered to provide high stiffness, an excellent strength-to-weight ratio, corrosion resistance, and fatigue resistance. Owing to these characteristics, they are particularly well suited for lightweight structural applications, including high-performance wheelchair components. This trend is consistent with international guidelines that promote the development of lighter, more durable assistive devices tailored to users' specific needs [1-5].

Among the various wheelchair components that can benefit from the use of composite materials, the seat plays a fundamental role in providing postural support, user comfort, and functional performance. A rigid seat base improves seating stability by providing adequate support for the wheelchair cushion, thereby promoting a more stable sitting posture and enhancing user safety and mobility [6], [7]. In addition, seating support characteristics directly influence interface pressure, shear forces, and user posture, factors that are associated with pressure injury risk and long-term seating performance [7]. Recent investigations have demonstrated that seat support conditions and seat-base deformation may significantly affect shear loading and sliding behavior during wheelchair use, highlighting the importance of structurally stable seating systems in rehabilitation technologies [8]. Consequently, the development of lightweight and structurally efficient composite seats represents a promising approach for improving seating support while simultaneously reducing the mass of assistive mobility devices.

Although composite materials have been previously investigated for wheelchair structures and lightweight mobility devices, studies focusing specifically on rigid wheelchair seat structures remain limited, particularly regarding the comparative structural performance of different laminate systems under equivalent loading conditions [1], [2]. Furthermore, few studies integrate experimentally determined material properties with analytical laminate design and finite element analysis within a reproducible workflow tailored to assistive technology applications. To address this gap, the present study investigates the use of carbon fiber- and glass fiber-reinforced composite laminates for rigid wheelchair seat applications. The mechanical properties of both laminate systems were determined through tensile, compressive, and shear testing and subsequently incorporated into analytical calculations based on Classical Laminate Theory (CLT) and finite element analyses (FEA) to predict seat deflection, evaluate structural performance, and assess failure indices under prescribed static loading conditions. In addition, laminate thickness, structural mass, and estimated raw material cost were compared to examine the trade-offs between structural efficiency and economic considerations. The main contribution of this work is a structured comparison between two composite laminate families through an integrated experimental-CLT-FEA methodology for the preliminary structural assessment of wheelchair seat designs.

II. Materials and Methods

A. Composite Materials and Laminate Fabrication

To evaluate the influence of the reinforcement material on the structural performance of the seat, two bidirectional woven prepregs were selected: carbon fiber-reinforced (CCS 200) and glass fiber-reinforced (WR 0200), both supplied by the Brazilian manufacturer Texiglass. These materials were selected because they represent two widely used composite reinforcement systems with distinct stiffness-to-weight and cost-performance characteristics. The carbon fiber fabric had an areal weight of 364 g/m², whereas the glass fiber fabric had an areal weight of 200 g/m². Both prepregs employed an epoxy resin matrix.

For the experimental characterization, composite laminates consisting of 12 fabric plies arranged in alternating 0°/90° orientations were manufactured. Following laminate lay-up, vacuum bagging was applied, and the laminates were cured in an oven using a thermal cycle refined through differential scanning calorimetry (DSC) analyses performed to evaluate the curing behavior of the epoxy matrix. Based on these results, a curing cycle of 1 h at 130 °C followed by 4 h at 150 °C was adopted to ensure an adequate degree of cure of the composite laminates. Fig. 1 illustrates the vacuum bag assembly and the laminates positioned inside the oven during processing.

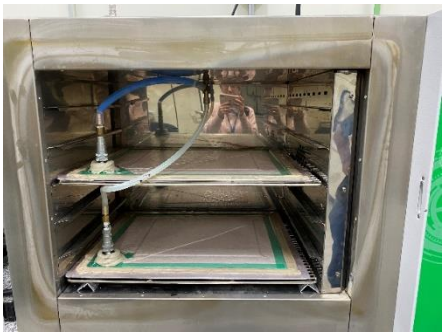


Fig. 1. Laminates positioned inside the oven for the curing cycle.

Mechanical test specimens were subsequently machined from the cured laminates. For the structural analyses, the wheelchair seat was represented by an idealized square composite plate with side dimensions of 400 mm. The plate dimensions were defined based on the seat dimensions of the Dellamed D400 wheelchair, which was used as a reference geometry for the study. This simplified representation was adopted to enable a consistent comparison between the laminate configurations and to ensure compatibility between the analytical and finite element models employed in the structural assessment. The laminate thickness was treated as a design variable and defined according to the number of plies required to satisfy the adopted maximum deflection criterion of 2.5 mm, which was established as a design requirement for the static loading condition investigated in this study.

B. Tensile, Compression, and Shear Testing

Mechanical characterization was performed using an Instron 5982 universal testing machine equipped with a 100 kN load cell. The testing machine and experimental setup are presented in Fig. 2.



Fig. 2. Instron machine used in mechanical tests.

Tensile, compression, and shear tests were conducted according to standards published by ASTM International, including ASTM D3039, ASTM D6641, and ASTM D3518, respectively. All tests were performed on both carbon fiber- and glass fiber-reinforced composite laminates. Five specimens were tested for each material and loading condition.

For the tensile and compression tests, specimens were machined along the 0° and 90° fiber directions, whereas the shear tests employed specimens with a ±45° fiber orientation. The tensile and shear specimens measured 250 × 25 mm, while the compression specimens measured 140 × 13 mm. At least three specimens from each group were instrumented with strain

gauges for strain measurement during testing, enabling the determination of the corresponding elastic properties.

Tensile and shear tests were conducted under displacement control at a crosshead rate of 2 mm/min. Compression tests were performed at a crosshead rate of 0.5 mm/min using a compression fixture, in which both specimen ends were clamped over a length of approximately 63.5 mm prior to testing, in accordance with ASTM D6641.

The experimental results obtained from these tests were subsequently used to determine the elastic properties and strength parameters required for the laminate analyses based on CLT and the Tsai-Wu failure criterion.

C. Classical Laminate Theory and Tsai-Wu Failure Criterion

The mechanical behavior of the laminate was evaluated using CLT, which was employed to determine the equivalent stiffness of the laminate from the elastic properties and stacking sequence of the individual plies. Based on this formulation, the extensional (A), extension–bending coupling (B), and bending (D) stiffness matrices were obtained, as expressed in (1) [9], [10].

$$\begin{Bmatrix} N \\ M \end{Bmatrix} = \begin{bmatrix} A & B \\ B & D \end{bmatrix} \begin{Bmatrix} \varepsilon^0 \\ \kappa \end{Bmatrix} \tag{1}$$

where N and M are the in-plane force and moment resultant vectors, respectively, ε^0 is the mid-plane strain vector, and κ is the laminate curvature vector.

For symmetric laminates, the coupling matrix (B) is equal to zero, allowing the bending stiffness matrix (D) to govern the plate bending behavior through the transformation of the material mechanical properties according to the fiber orientation of each ply. Therefore, the laminate bending stiffness can be determined using (2) [9].

$$D_{ij} = \frac{1}{3}\sum_{k=1}^{N} \bar{Q}_{ij}^{(k)} \left(z_{k+1}^3 - z_k^3\right) \tag{2}$$

where N is the total number of plies, and (k) and (k+1) denote the coordinates of the upper and lower surfaces of the *k*-th ply.

The bending stiffness coefficients were subsequently incorporated into an analytical plate-bending model based on the Lévy solution to estimate the deflection of the wheelchair seat. Due to the mathematical limitations of the analytical formulation, the plate was modeled considering two opposite simply supported edges, providing a conservative approximation of the boundary conditions adopted for the structural assessment. In contrast, the finite element model incorporated the idealized restraints considered in the seat design, allowing the structural response to be evaluated under the prescribed numerical boundary conditions. Consequently, the analytical solution was primarily employed as a preliminary design and verification tool, whereas the finite element model was used for a more detailed assessment of the structural response of the idealized seat configuration investigated in this study. Ply stresses were then obtained from the CLT strain–stress relations.

Failure was evaluated using the Tsai–Wu criterion, which predicts failure when the value of (3) is equal to or greater than 1 [11].

$$F_1\sigma_1 + F_2\sigma_2 + F_{11}{\sigma_1}^2 + F_{22}{\sigma_2}^2 + F_{66}{\tau_{12}}^2 + 2F_{12}\sigma_1\sigma_2 \geq 1 \tag{3}$$

where σ_1 and σ_2 are the stresses in the material principal directions 1 and 2, respectively, and τ_{12} is the in-plane shear stress in the 1–2 coordinate system. The coefficients F_{ij} were calculated from the tensile, compressive, and shear strengths of the material, as defined in (4) [11], [12].

$$F_1 = \frac{1}{X_t} - \frac{1}{X_c}, \quad F_2 = \frac{1}{Y_t} - \frac{1}{Y_c}, \quad F_{11} = \frac{1}{X_t X_c},$$
$$F_{22} = \frac{1}{Y_t Y_c} \quad e \quad F_{66} = \frac{1}{S^2} \tag{4}$$

where X and Y denote the material strengths in the x- and y-directions, respectively, and the subscripts (t) and (c) refer to tensile and compressive strengths. The interaction coefficient F_{12} was calculated according to the procedure proposed in the literature [12].

The criterion was employed to assess whether the predicted stress state of the laminate remained below the failure threshold under the prescribed loading conditions. No progressive damage analysis or ply-by-ply failure propagation modeling was considered in the present study.

D. Numerical Modeling using the Finite Element Method

A finite element model was developed in ANSYS® Parametric Design Language (APDL) 2024 to simulate the structural behavior of the wheelchair seat and to complement the analytical predictions obtained from Classical Laminate Theory (CLT). The composite plate was discretized using SHELL281 elements with an initial element size of 10 mm. A mesh convergence study was performed by reducing the element size to 5 mm to ensure numerical accuracy and mesh-independent results.

The plate geometry was defined as 400 × 400 mm, with thickness varying according to the reinforcement material and laminate configuration. Two opposite edges were fully clamped, whereas the remaining edges were free. The numerical model incorporated idealized constraints intended to represent the attachment of the seat to the wheelchair structure. However, the actual stiffness of the fixation system may vary depending on frame geometry, fastening method, manufacturing tolerances, and assembly conditions. Experimental validation of a full-scale seat mounted on a wheelchair frame is therefore necessary to quantify the influence of these factors on the structural response.

A design load of 1800 N was uniformly distributed over the entire plate surface, as shown in Fig. 3. This load was obtained by applying a safety factor of 1.5 to a nominal service load of 1200 N, corresponding to an adult wheelchair user with a body mass of approximately 120 kg. The adopted loading condition was intended to represent a conservative static loading scenario for the preliminary structural assessment of the wheelchair seat.

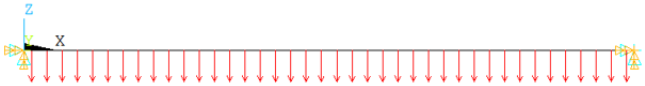


Fig. 3. Boundary conditions for the wheelchair seat.

The finite element models were used to evaluate the structural response of the laminates in terms of maximum

deflection and maximum Tsai–Wu failure index under the prescribed static loading condition. The Tsai-Wu criterion was employed solely to verify whether the predicted stress state remained below the adopted failure threshold, without considering progressive damage evolution, delamination mechanisms, or fatigue effects.

The numerical simulations were based on material properties experimentally determined through tensile, compression, and shear tests performed at the Lightweight Structures Laboratory of the Institute for Technological Research (IPT).

To promote computational reproducibility, the complete ANSYS APDL script used in this study is publicly available in an open-access Zenodo repository [13], enabling the reproduction of the finite element analyses presented in this work.

III. Results and Discussion

A. Mechanical Properties of the Composite Laminates

The tensile, compression, and shear tests conducted on the glass fiber- and carbon fiber-reinforced composites allowed the mechanical properties required for the laminate analyses to be determined. The corresponding results are summarized in Tables I and II.

TABLE I. Mechanical Properties of the WR-200-RE-02B Glass-Fiber-Reforced Laminate.

Property	Symbol	Method	Unity	Value
0° Tensile Strength	F1t	ASTM D3039	MPa (Ksi)	427.521 (62.0)
90° Tensile Strength	F2t	ASTM D3039	MPa (Ksi)	467.771 (67.8)
0° Tensile Modulus	E1t	ASTM D3039	GPa (Msi)	30.722 (4.5)
90° Tensile Modulus	E2t	ASTM D3039	GPa (Msi)	32.08 (4.6)
0° Compressive Strength	F1c	ASTM D6641	MPa (Ksi)	318.271 (46.2)
90° Compressive Strength	F2c	ASTM D6641	MPa (Ksi)	404.945 (58.7)
0° Compressive Modulus	E1c	ASTM D6641	GPa (Msi)	30.604 (4.4)
90° Compressive Modulus	E2c	ASTM D6641	GPa (Msi)	33.428 (4.8)
In-plane Shear Strength 5%	F12	ASTM D3518	MPa (Ksi)	51.996 (7.5)
In-plane Shear Modulus	G12	ASTM D3518	GPa (Msi)	3.865 (0.56)
Poisson's Ratio	v12	ASTM D3039	---	0.131
Poisson's Ratio	v21	ASTM D3039	---	0.135
Cured Ply Thickness	t cured	---	mm (inches)	0.133 (0.0052)

TABLE II. Mechanical Properties of the CCS-200-RE-02B Carbon-Fiber Reforced Laminate.

Property	Symbol	Method	Unity	Value
0° Tensile Strength	F1t	ASTM D3039	MPa (Ksi)	649.397 (94.2)
90° Tensile Strength	F2t	ASTM D3039	MPa (Ksi)	537.167 (77.9)
0° Tensile Modulus	E1t	ASTM D3039	GPa (Msi)	57.279 (8.3)
90° Tensile Modulus	E2t	ASTM D3039	GPa (Msi)	52.087 (7.5)
0° Compressive Strength	F1c	ASTM D6641	MPa (Ksi)	390.499 (56.6)
90° Compressive Strength	F2c	ASTM D6641	MPa (Ksi)	425.673 (61.7)
0° Compressive Modulus	E1c	ASTM D6641	GPa (Msi)	50.285 (7.3)
90° Compressive Modulus	E2c	ASTM D6641	GPa (Msi)	46.634 (6.8)
In-plane Shear Strength 5%	F12	ASTM D3518	MPa (Ksi)	62.625 (9.1)
In-plane Shear Modulus	G12	ASTM D3518	GPa (Msi)	3.692 (0.54)
Poisson's Ratio	v12	ASTM D3039	---	0.07
Poisson's Ratio	v21	ASTM D3039	---	0.044
Cured Ply Thickness	t cured	---	mm (inches)	0.244 (0.0096)

The results indicate that the carbon fiber-reinforced composite exhibits higher stiffness under both tensile and compressive loading conditions. In tension, the tensile modulus reached 57.28 GPa (0°) and 52.09 GPa (90°), while in compression it reached 50.28 GPa (0°) and 46.63 GPa (90°). These values were substantially higher than those obtained for the glass fiber-reinforced composite, which presented tensile moduli of 30.72 GPa (0°) and 32.08 GPa (90°), and compressive moduli of 30.60 GPa (0°) and 33.43 GPa (90°).

In addition, the cured ply thickness of the carbon fiber prepreg (0.24 mm) was approximately 85% greater than that of the glass fiber prepreg (0.13 mm). This characteristic is relevant for laminate design because the flexural stiffness of composite laminates depends not only on the elastic properties of the constituent materials but also on the laminate thickness and ply distribution through the thickness.

The glass fiber-reinforced composite exhibited lower elastic moduli and therefore greater deformability under mechanical loading. Furthermore, the average Poisson's ratio measured for the glass fiber laminate (0.13) was higher than that of the carbon fiber laminate (0.05), indicating a greater tendency for transverse deformation under load. These differences in elastic behavior are expected to influence the bending stiffness and structural response of the laminate configurations evaluated in the subsequent analytical and numerical analyses.

B. Structural Analysis of the Composite Wheelchair Seat

For the calculation of seat deflection in the laminated composite structure, a maximum allowable deflection of 2.5 mm was adopted as a design criterion. This value was selected to limit excessive seat deformation and maintain adequate stiffness of the seating support under the prescribed static loading condition. Although no specific standard prescribing this limit was considered in the present study, minimizing seat deformation may help preserve the mechanical stability of the seating system, since previous studies have shown that seat support conditions and seat-base deformation can influence shear forces, sliding behavior, and seating posture in wheelchair users [7], [8]. The laminate configurations, including the number of plies and laminate thickness, together with the calculated bending stiffness coefficients and the corresponding analytically predicted seat deflections, are summarized in Table III.

TABLE III. Laminate Configuration, Bending Stiffness Coefficients and Analytically Deflection for Carbon and Glass Fiber-Reinforced Composite Materials.

Fiber - reinforced	Number of plies	Thickness (mm)	D11	D22	D12	D66	Maximum deflection (mm)
Carbon	18	4.40	773936.00	773936.01	36704.62	36500.96	2.13
Glass	38	5.05	340821.71	340821.71	45209.54	41578.99	2.46

The analytical results indicated that both laminate configurations satisfied the adopted maximum deflection criterion of 2.5 mm. The experimentally determined elastic properties incorporated into the CLT calculations revealed substantial differences in laminate bending stiffness between the two reinforcement systems. The carbon fiber-reinforced laminate exhibited higher bending stiffness coefficients, particularly for the D_{11} and D_{22} terms, reflecting the higher elastic moduli measured during mechanical characterization. Consequently, the carbon fiber laminate achieved the lowest

predicted maximum deflection (2.13 mm), despite having a smaller thickness of 4.40 mm and fewer plies (18) than the glass fiber laminate, which required 38 plies and a thickness of 5.05 mm to achieve a predicted deflection of 2.46 mm. These results indicate that the higher stiffness of the carbon fiber reinforcement enabled the adopted deflection criterion to be satisfied with a thinner and lighter laminate configuration.

It should be noted that the analytical predictions were obtained using a simplified plate-bending model based on the Lévy solution and therefore represent preliminary estimates of the structural response. Because the analytical formulation employed a conservative approximation of the boundary conditions, the calculated deflections should be interpreted primarily as a design-oriented assessment of the laminate behavior under the prescribed static load.

C. Numerical Evaluation of Stress, Deflection, and Tsai–Wu Failure Index

The first principal stress distribution obtained from the finite element analysis exhibited maximum values near the clamped edges and minimum values in the central region for both laminates, as depicted in Fig. 4.

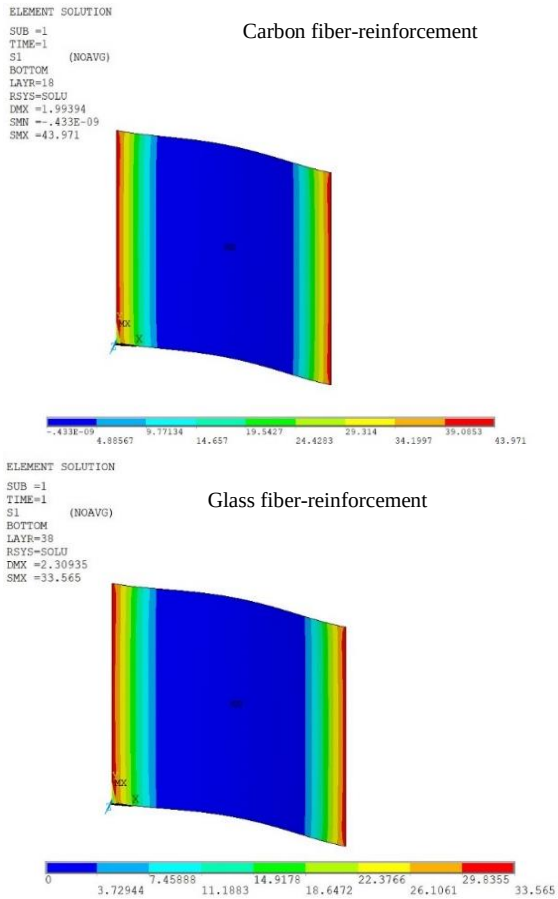


Fig. 4. First Principal Stress distribution for the carbon fiber- and glass fiber-reinforced laminates, respectively.

This behavior is characteristic of rectangular plates subjected to uniformly distributed transverse loading, where fully constrained boundaries generate higher bending moments and consequently higher stress concentrations near the supports. The symmetric stress distribution confirms that the structural response was dominated by bending.

The finite element model predicted maximum deflections of 1.99 mm and 2.31 mm for the carbon and glass laminates, respectively, corresponding to reductions of approximately 7% and 6% compared with the analytical predictions. The finite element results predicted slightly lower deflections than those obtained analytically. This difference is expected because the analytical solution was based on a simplified boundary condition with opposite simply supported edges, whereas the finite element model included additional constraints associated with the seat attachment configuration. Since increased boundary restraint leads to higher structural stiffness and reduced displacement, the analytical predictions can be regarded as conservative estimates of the laminate response under the prescribed loading conditions. The corresponding maximum deflection contours predicted for the carbon fiber- and glass fiber-reinforced laminate configurations are presented in Fig. 5.

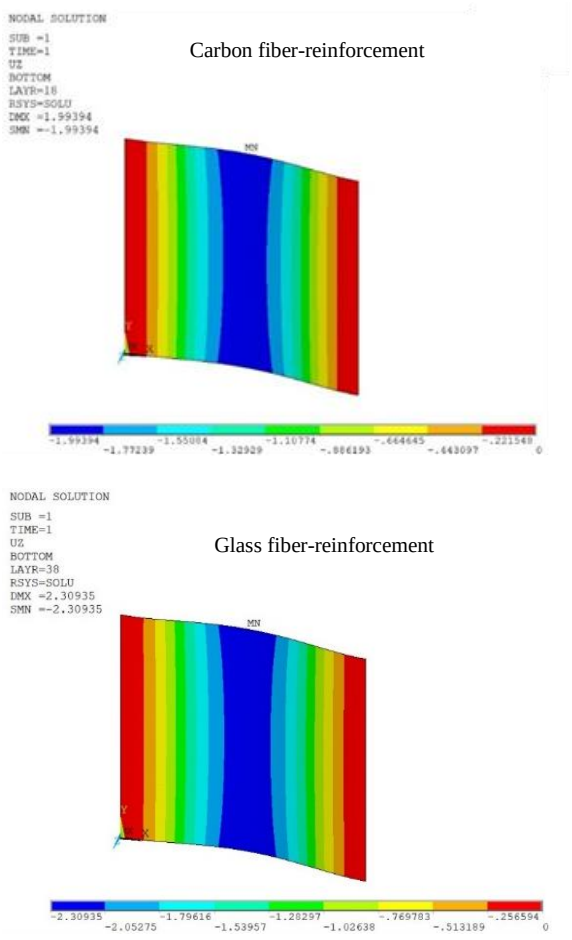


Fig. 5. Maximum deflection contours for the carbon fiber- and glass fiber-reinforced laminates, respectively.

These numerical results are consistent with the bending stiffness coefficients obtained from the Classical Laminate Theory analysis, particularly the D_{11} and D_{22} terms. The higher stiffness experimentally measured for the carbon fiber-reinforced laminate resulted in larger laminate bending stiffness

coefficients and consequently lower deflection values in both the analytical and numerical models.

Although deflection was adopted as the primary design criterion, the Tsai-Wu failure criterion was also evaluated through finite element analysis. The corresponding maximum Tsai-Wu failure indices for the carbon fiber- and glass fiber-reinforced laminates are presented in Fig. 6.

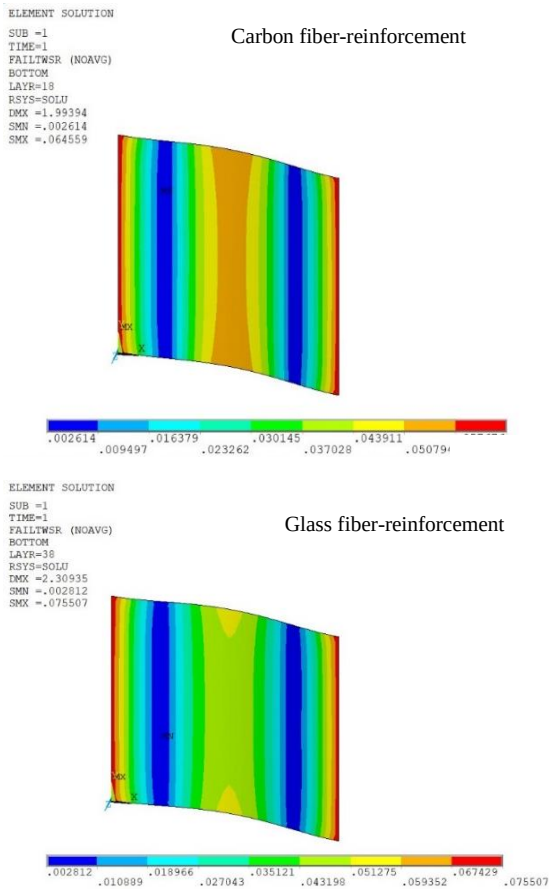


Fig. 6. Tsai-Wu failure index distribution for the carbon fiber- and glass fiber-reinforced laminates, respectively.

As shown in Fig. 6, the maximum Tsai-Wu indices obtained for the carbon (0.065) and glass (0.075) fiber-reinforced laminates remained well below the critical value of 1.0. These results indicate that, for the prescribed static loading condition and within the assumptions adopted by the numerical model, the predicted stress states remained below the failure threshold defined by the Tsai-Wu criterion. It should be noted that the present analysis was limited to the evaluation of the laminate response through the maximum Tsai-Wu index and did not include progressive damage modeling or ply-by-ply failure analysis. Consequently, it was not possible to determine the sequence of potential ply failures or assess damage evolution mechanisms.

Accordingly, the numerical results suggest that the stress levels developed in both laminate configurations remained within the limits defined by the adopted failure criterion for the prescribed loading condition.

The results presented in this study should be interpreted in light of several limitations. The structural assessment was performed using an idealized plate geometry subjected to a prescribed static load, and the complete wheelchair seat was not manufactured or experimentally tested. Consequently, aspects such as fatigue behavior, impact resistance, damage accumulation, long-term durability, and the effects of manufacturing-related imperfections were not evaluated. Furthermore, the interaction between the seat, cushion, and user, as well as realistic pressure distributions and dynamic loading conditions, were not considered in the present analysis.

Although the numerical model included idealized constraints intended to represent the attachment of the seat to the wheelchair frame, variations in fixation stiffness and assembly conditions may influence the actual structural response and should be investigated experimentally in future studies.

Together with the deflection analysis, these results indicate that both laminate configurations satisfied the structural design criteria established for the prescribed static loading scenario and support their preliminary structural assessment within the assumptions adopted in the present study.

D. Technical and Economic Assessment of Carbon Fiber- and Glass Fiber-Reinforced Composites for Seat Applications

From a structural perspective, the results indicate that both carbon fiber- and glass fiber-reinforced laminate configurations satisfied the design criteria established for the prescribed static loading condition. The carbon fiber laminate required a thickness of 4.40 mm, whereas the glass fiber laminate required a thickness of 5.05 mm to satisfy the adopted maximum deflection limit. These differences reflect the higher stiffness of the carbon fiber reinforcement and its influence on the laminate bending response.

In the present study, the cost of carbon fiber prepreg was estimated at R$ 376.40/m², whereas glass fiber prepreg was estimated at R$ 73.84/m². These values were based on supplier quotations obtained in mid-2025 and exclude applicable taxes. Therefore, the reported costs should be regarded as indicative values, as they may vary according to market conditions, supplier availability, and economic factors.

Based on the seat area of 0.16 m², the required number of plies for each laminate configuration, and the corresponding mass per ply, the direct raw material cost and final seat mass were estimated for each reinforcement system. The resulting values are summarized in Table IV.

TABLE IV. COST AND MASS OF CARBON FIBER- AND GLASS FIBER-REINFORCED USED IN A WHEELCHAIR SEAT.

Reinforcement Material	Number of plies	Direct raw material cost	Seat mass (kg)
Carbon	18	R$ 1,545.87	0.977
Glass	38	R$ 477.83	1.545

While the carbon fiber-reinforced laminate exhibited higher specific stiffness and enabled reductions in laminate thickness and structural mass, its estimated direct raw material cost was

substantially higher than that of the glass fiber-reinforced laminate. Based solely on the prepreg material costs considered in this study, the carbon fiber laminate was approximately 223.5% more expensive than the glass fiber laminate. In contrast, the carbon fiber configuration resulted in a 36.7% reduction in seat mass, highlighting the trade-off between lightweight design and material cost.

It should be noted that this economic comparison reflects only the estimated direct raw material cost and does not include manufacturing-related expenses such as material waste, vacuum bagging consumables, curing energy, labor, tooling, machining, assembly, or quality-control procedures. Consequently, the results should not be interpreted as representative of the final manufacturing cost of a complete wheelchair seat.

For the static loading scenario investigated, both laminate configurations satisfied the adopted structural design criterion and remained below the Tsai-Wu failure threshold. The carbon fiber laminate provided the lowest deflection and structural mass, whereas the glass fiber laminate represented a lower-cost material alternative. These results illustrate the trade-off between structural efficiency and material cost that must be considered during the preliminary design of composite wheelchair seating systems.

Although the reduced mass achieved with the carbon fiber laminate may be advantageous in applications where lightweight components are desired, the present study did not assess factors such as propulsion efficiency, user effort, comfort, durability, or clinical performance. Therefore, the potential functional implications of weight reduction should be investigated through future experimental and user-centered studies.

IV. Conclusion

This study compared carbon fiber- and glass fiber-reinforced composite laminates for wheelchair seat applications through an integrated methodology involving mechanical characterization, analytical modeling based on CLT, and FEA. The results indicated that structural stiffness, rather than material strength, governed the response of the proposed seat configuration under the prescribed static loading conditions. Both laminate configurations satisfied the adopted deflection criterion and exhibited Tsai-Wu failure indices below the adopted failure threshold for the loading scenario considered.

The analytical and numerical approaches showed good agreement regarding the structural behavior of the investigated laminates. Slightly lower deflections were obtained from the finite element analyses, which is consistent with the different boundary conditions adopted in each approach. While the analytical Lévy solution required a simplified plate-support condition and therefore provided a conservative estimate of displacement, the finite element model incorporated the idealized boundary restraints adopted for the seat design. Consequently, the numerical analyses provided a more detailed assessment of the structural response of the idealized seat configuration investigated in this study.

It should be emphasized that the adopted maximum deflection limit of 2.5 mm represents a design requirement established for the present study and should not be interpreted as a standardized acceptance criterion for wheelchair seating systems. Similarly, the Tsai-Wu failure indices obtained in the numerical analyses indicate only that the predicted stress states remained below the failure threshold defined by the adopted criterion under the prescribed static loading condition. These results do not provide information regarding fatigue resistance, damage progression, delamination behavior, impact resistance, or long-term durability.

The carbon fiber-reinforced laminate achieved the lowest structural mass and deflection, reflecting the higher specific stiffness of carbon fibers and allowing the adopted design criterion to be satisfied with fewer plies and reduced laminate thickness. In contrast, the glass fiber-reinforced laminate represented a lower-cost alternative when considering direct raw material costs, while also satisfying the structural requirements established for the analyzed loading condition. These findings highlight the trade-off between structural efficiency and material cost that must be considered during the preliminary design of composite seating structures for assistive technology applications.

From an assistive technology perspective, the reduction in structural mass achieved with the carbon fiber laminate may be advantageous because lighter wheelchair components are generally associated with reduced inertial loads during handling, transportation, and propulsion. However, the present study did not evaluate clinical outcomes, user effort, comfort, propulsion efficiency, seating biomechanics, or functional performance. Therefore, any potential benefits associated with mass reduction should be confirmed through experimental and user-centered investigations.

The principal methodological contribution of this work lies in the structured comparison of two composite laminate systems using experimentally determined material properties integrated with CLT-based analytical calculations, failure assessment, and finite element analysis within a reproducible design framework. The main lessons learned from this study are that: (*i*) laminate stiffness was the governing design parameter under the analyzed loading conditions; (*ii*) carbon fiber reinforcement enabled substantial mass reduction with fewer plies and lower laminate thickness; (*iii*) glass fiber reinforcement remained a competitive lower-cost alternative from a raw-material perspective; and (*iv*) the combined use of experimental characterization, CLT, and FEA provided consistent support for the preliminary structural assessment of composite wheelchair seat concepts.

The results should be interpreted as a preliminary structural assessment of two laminate configurations under an idealized static loading scenario. Important research gaps remain and should be addressed in future work, including full-scale seat manufacturing and experimental testing, validation of realistic seat-to-frame fixation conditions, fatigue and impact evaluation, investigation of seat-cushion-user interaction, assessment of realistic pressure distributions, and studies involving comfort, postural support, and functional performance with wheelchair users. Within these limitations, the methodology presented provides a useful foundation for future structural optimization and development of composite wheelchair seating systems.

ACKNOWLEDGMENT

This study was financed in part by the São Paulo Research Foundation (FAPESP), Brasil. Grants No. 2025/11908-0 and No. 2024/01120-4.

REFERENCES

[1] F. Teodorescu, C. Dumitru, G. Stanca, V. Avramescu, and R. M. Nita, “Static behaviour of an advanced ultra-light sandwich composite structure for a wheelchair,” In: proceedings of the 1st wseas international conference on finite differences-finite elements-finite volumes-boundary elements, pp. 39–44, 2008.

[2] W. S. Chong, M. Y. Shin, and C. H. Yu, “Structural analysis of carbon composite frame for foldable electric wheelchair development”, Journal of Mechanics in Medicine and Biology, 19(07), 1940045, 2019.

[3] S. Merai, D. Shah, B. Trivedi, P. Joshi, and S. Kushwah, “A study and design of standing wheelchair”, Materials Today: Proceedings, vol. 65, pp. 3787-3792, 2022.

[4] World Health Organization, Wheelchair provision guidelines. Geneva, Switzerland: World Health Organization, 2023. [Online}. Available: https://www.ispoint.org/wp-content/uploads/2023/06/WHO-Wheelchair-provision-guidelines-2023.pdf

[5] E. J. Barbero, Introduction to composite materials design, 2nd ed. Boca Raton, FL, USA: CRC Press, 2010.

[6] A. M. Cook, and J. M. Polgar, Assistive technologies: principles and practice, 4th ed. St. Louis, MO, USA: Elsevier, 2015.

[7] K. Sasaki et al. “Effects of wheelchair seat sagging on seat interface pressure and shear, and its relationship with changes in sitting posture”, biomechanics, 5(2), 41, 2025.

[8] Z. Huang, J. Cui, Y. Wang, and S. Yu, "Improving wheelchair user sitting posture to alleviate lumbar fatigue: a study utilizing sEMG and pressure sensors", Frontiers in Neuroscience, 18, 2024.

[9] J. N. Reddy, Theory and analysis of elastic plates and shells, 2nd ed. Boca Raton, FL, USA: CRC Press, 2007.

[10] A. K. Kaw, Mechanics of composite materials, 2nd ed. Boca Raton, FL, USA: CRC press, 2005.

[11] L. P. Kollar, and G. S. Springer, Mechanics of composite structures. Cambrigde, U.K.: Cambridge University Press, 2003.

[12] L. A. Carlsson, D. F. Adams, and R. B. Pipes, Experimental characterization of advanced composite materials, 4th ed. Boca raton, FL, USA: CRC press, 2014.

[13] C. A. Diniz, P. L. Ragueneau, J. P. Nascimento, M. H. F. Batalha, and A. Guimarães, A*NSYS APDL script for structural analysis of fiber-reinforced composite wheelchair seats*, Version 1.0, Zenodo, Jun. 2026. Doi: 10.5281/zenodo.20798832.

Cultura Maker Aplicada à Tecnologia Assistiva: Desenvolvimento de um Processo de Fabricação Digital para Bancos Terapêuticos Destinados à Reabilitação Pediátrica

1st Geovane Umbelino Marques
Laboratório de Estudos Inventivos em Tecnologias Assistivas (Lab EITA)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
geovanemarques@ufg.br

2nd Pedro Henrique Gonçalves
Laboratório de Estudos Inventivos em Tecnologias Assistivas (Lab EITA)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
pedrogoncalves@ufg.br

***Resumo*—O acesso às tecnologias assistivas permanece limitado, especialmente em contextos de vulnerabilidade social, devido aos elevados custos e à baixa disponibilidade de dispositivos adaptados. Nesse cenário, este trabalho teve como objetivo desenvolver um conjunto modular de bancos terapêuticos para crianças com paralisia cerebral, utilizando princípios da cultura maker e processos de fabricação digital. A estimativa dos custos diretos de materiais e serviços resultou em valor inferior à média dos conjuntos comerciais utilizados como referência. O equipamento foi aplicado preliminarmente em ambientes de pesquisa e atendimento terapêutico, permitindo observar sua utilização em atividades funcionais. Os resultados demonstram a viabilidade técnica do processo de desenvolvimento e fabricação digital proposto, evidenciando seu potencial para produção local de dispositivos terapêuticos em ambientes makers.**

***Palavras-chave*—Bancos terapêuticos, tecnologia assistiva, cultura maker; fabricação digital, paralisia cerebral.**

I. Introdução

A tecnologia assistiva (TA), compreendida como o conjunto de produtos, recursos e serviços que promovem funcionalidade, independência e participação de pessoas com deficiência [1], desempenha papel fundamental na melhoria da qualidade de vida e da inclusão social. Entretanto, o acesso a esses recursos permanece restrito, sobretudo em contextos de vulnerabilidade social, devido ao elevado custo, à limitada disponibilidade e à escassa personalização dos dispositivos. Estima-se que mais de 2,5 bilhões de pessoas necessitem de pelo menos um produto de tecnologia assistiva, sendo que uma parcela significativa ainda não possui acesso adequado, especialmente em países de baixa e média renda [2].

Entre os grupos que mais dependem desses recursos estão as crianças com deficiência, especialmente aquelas com paralisia cerebral, cuja condição demanda o uso contínuo de dispositivos de posicionamento, mobilidade e comunicação para favorecer o desenvolvimento funcional, a participação social e a autonomia ao longo do crescimento. A paralisia cerebral (PC) é definida como uma condição do neurodesenvolvimento de início precoce e caráter permanente, que acompanha o indivíduo ao longo de toda a vida. Caracteriza-se por limitações na atividade decorrentes de distúrbios do desenvolvimento do movimento e da postura, manifestando-se por alterações motoras como espasticidade, distonia, coreoatetose e/ou ataxia [3]-[5].

Entre os recursos utilizados na reabilitação dessas crianças destacam-se os dispositivos de posicionamento, que contribuem para o alinhamento postural, o controle de tronco e a realização de atividades terapêuticas e funcionais. Entretanto, esses equipamentos frequentemente apresentam elevado custo e limitada disponibilidade, dificultando seu acesso em serviços públicos de saúde e por famílias de baixa renda.

Nesse sentido, a cultura maker, entendida como um movimento fundamentado na aprendizagem pela prática, na colaboração e no compartilhamento de conhecimentos [6], aliada às tecnologias de fabricação digital — conjunto de processos de manufatura controlados por computador que transformam modelos digitais em objetos físicos [7] — emerge como uma alternativa promissora para o desenvolvimento de tecnologias assistivas acessíveis, personalizadas e reproduzíveis. Essas abordagens possibilitam a fabricação local de dispositivos, contribuindo para a redução dos custos de produção, a adaptação às necessidades específicas dos usuários e a ampliação do acesso aos recursos assistivos.

Embora o uso da impressão 3D e de outras tecnologias de fabricação digital tenha crescido significativamente na área da tecnologia assistiva, ainda são escassos estudos que descrevam o desenvolvimento e a avaliação de dispositivos terapêuticos de baixo custo voltados especificamente à reabilitação de crianças com paralisia cerebral, considerando sua viabilidade técnica e seu potencial de replicação em contextos de recursos limitados.

Diante desse cenário, este trabalho tem como objetivo desenvolver um conjunto modular de bancos terapêuticos por meio de um processo estruturado de fabricação digital baseado nos princípios da cultura maker e realizar uma avaliação preliminar de sua aplicação em ambientes de reabilitação.

II. Bancos Terapêuticos na Reabilitação de Crianças com Paralisia Cerebral

O conjunto de bancos terapêuticos (Fig. 1) consiste em um equipamento modular destinado à realização de atividades terapêuticas voltadas ao desenvolvimento do equilíbrio, do controle postural, do fortalecimento muscular e das habilidades motoras funcionais. Composto por módulos de diferentes alturas, o equipamento possibilita a execução de exercícios como subir e descer degraus, sentar, levantar e deslocar-se entre diferentes níveis de apoio, permitindo adequar as atividades às capacidades funcionais e à evolução clínica de cada criança. Em geral, esses conjuntos são comercializados em kits compostos por três a seis módulos que se encaixam uns dentro dos outros (Fig. 2), característica que facilita seu armazenamento, transporte e organização em ambientes terapêuticos.

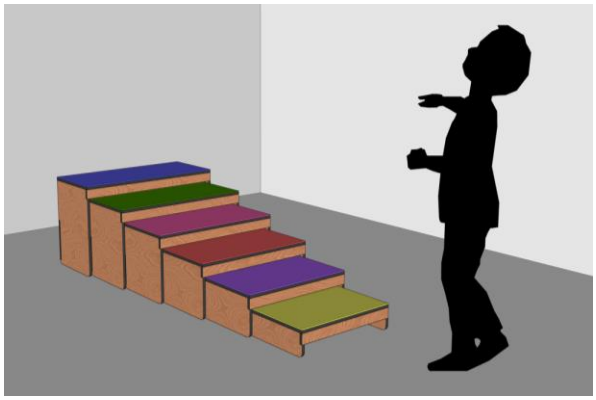


Fig. 1. Conjunto de bancos terapêuticos.

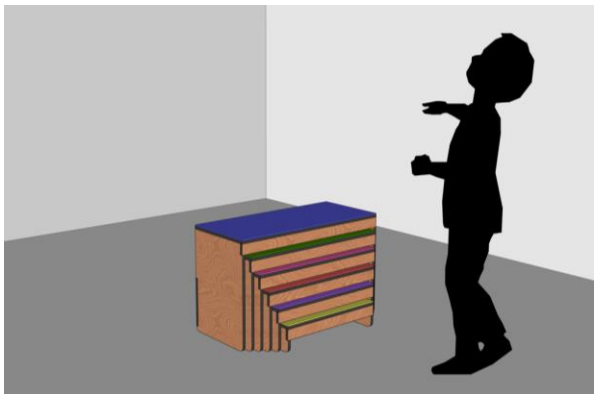


Fig. 2. Conjunto de bancos terapêuticos recolhidos.

Em um estudo piloto com oito crianças com paralisia cerebral, os autores [8] avaliaram um programa de treinamento funcional orientado à tarefa que incluía exercícios como subida e descida de degraus, step-ups, step-downs, sentar e levantar, caminhada em rampas e escadas. Os autores observaram melhora significativa da força muscular, do desempenho funcional, da marcha e da capacidade de realizar o movimento de sentar e levantar após quatro semanas de treinamento.

Considerando que o conjunto de bancos terapêuticos desenvolvido neste estudo pode permitir a execução de atividades semelhantes às descritas em [8], seu uso apresenta potencial para apoiar programas de treinamento funcional voltados ao desenvolvimento das habilidades motoras de crianças com paralisia cerebral. Contudo, ressalta-se que a aplicação do equipamento em ambientes de atendimento, pesquisa ou domiciliares ainda requer estudos específicos que avaliem sua eficácia, segurança e aplicabilidade em diferentes contextos de uso.

III. Concepção, Métodos e Materiais

O projeto foi concebido a partir da análise de conjuntos de bancos terapêuticos disponíveis comercialmente e do levantamento de requisitos identificados durante a prática clínica da fisioterapia pediátrica no Laboratório Inventivo em Tecnologias Assistivas (Lab EITA), da Universidade Federal de Goiás (UFG). A proposta buscou preservar as funcionalidades dos equipamentos convencionais, incorporando soluções voltadas à redução dos custos de produção, à simplificação do processo de fabricação e à possibilidade de replicação em laboratórios makers, em consonância com os princípios da fabricação digital e da cultura maker.

Como requisito inicial de projeto, optou-se pela utilização de chapas de MDF (*Medium Density Fiberboard*), em razão de sua ampla disponibilidade comercial, baixo custo e compatibilidade com processos de fabricação digital, especialmente o corte a laser. O material apresenta características favoráveis ao processamento por corte a laser, possibilitando a obtenção de peças com precisão dimensional e bom acabamento superficial quando empregados parâmetros adequados de processamento [9]. Além disso, sua superfície homogênea e a facilidade de colagem e acabamento favorecem a fabricação de componentes com encaixes precisos, contribuindo para a simplificação da montagem e para a reprodução do equipamento em diferentes laboratórios makers.

Em comparação com processos construtivos que empregam materiais como compensado e perfis metálicos, o uso do MDF associado ao corte a laser permite simplificar determinadas etapas de fabricação, dispensando operações como soldagem e reduzindo a necessidade de usinagem especializada. Essa característica favorece a produção das peças em ambientes de fabricação digital com infraestrutura relativamente simples.

Além do MDF, foram empregados outros materiais no processo de fabricação, cujas especificações são apresentadas na Tabela I. Entre os insumos utilizados estão folhas de etileno-acetato de vinila (EVA) com 4 mm de espessura, adesivo à base de acetato de polivinila (PVA), adesivo instantâneo à base de cianoacrilato e parafusos para madeira. O EVA foi aplicado como revestimento das superfícies superiores dos módulos e fixado ao MDF por meio do adesivo à base de cianoacrilato, enquanto o adesivo PVA e os parafusos foram empregados na união e fixação dos componentes estruturais.

Materiais Empregados		
Material	***Especificação***	***Finalidade***
Chapa de MDF	12mm	Fabricação dos seis módulos
Folha de EVA	4mm	Revestimento
Cola Branca PVA Extra	500g	Colagem das peças de MDF
Cola Cianoacrilato	100g	Colagem do EVA nos Módulos
Parafusos para Madeira	3,5mm x 25 mm	Fixação das peças de MDF

TABELA I

A. Modelagem Digital

A modelagem tridimensional do conjunto de bancos terapêuticos foi desenvolvida no SketchUp para Web, plataforma on-line de modelagem tridimensional com acesso gratuito,

utilizada para a definição da geometria, das dimensões e dos encaixes entre os módulos. Como requisito inicial de projeto, estabeleceu-se a utilização de chapas de MDF com 12 mm de espessura, em razão de sua compatibilidade com o processo de corte a laser e de sua disponibilidade comercial.

A partir dessa restrição dimensional, todas as peças foram projetadas considerando a espessura do material, garantindo o correto encaixe dos componentes e a estabilidade estrutural do conjunto. O ambiente CAD (Computer Aided Design) possibilitou ainda a realização de ajustes dimensionais, a verificação das interferências entre as peças e a visualização do conjunto antes da fabricação física.

Os módulos foram dimensionados de forma progressiva, sendo o primeiro com 10,5 cm de altura e os demais acrescidos de incrementos de 6,5 cm até o sexto módulo (Fig. 3). A definição desses valores teve como referência a experiência prática desenvolvida no Lab EITA em atividades de avaliação funcional e reabilitação pediátrica, considerando a necessidade de oferecer diferentes níveis de altura para adaptação das atividades às capacidades funcionais das crianças. O incremento de 6,5 cm foi adotado como parâmetro de projeto para estabelecer diferenças graduais entre os níveis de apoio, permitindo a progressão das atividades sem produzir variações excessivas de demanda entre módulos consecutivos.

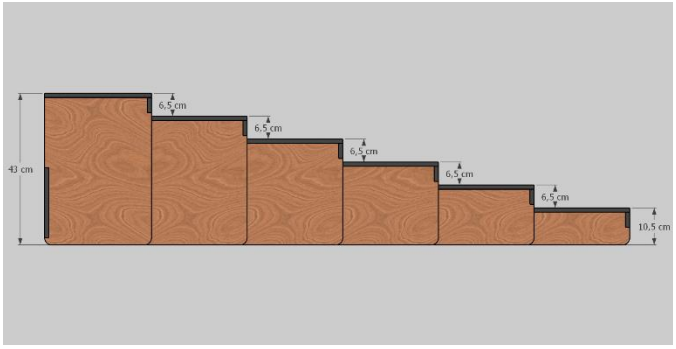


Fig. 3. Vista lateral do conjunto de bancos terapêuticos.

Consequentemente, os incrementos sucessivos de 6,5 cm resultaram em alturas de 17 cm, 23,5 cm, 30 cm, 36,5 cm e 43 cm (Fig. 4). A progressão estabelecida permite selecionar diferentes alturas de apoio de acordo com a atividade terapêutica e as características funcionais de cada criança, além de possibilitar a combinação dos módulos em diferentes configurações.

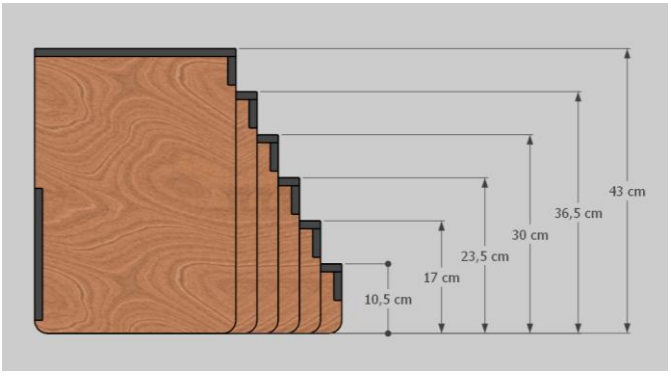


Fig. 4. Vista lateral do conjunto de bancos terapêuticos recolhidos.

Visando ao encaixe entre os módulos, cada banco foi projetado com um comprimento 5 cm inferior ao do módulo imediatamente superior, permitindo que todos fossem armazenados de forma compacta quando não estivessem em uso. Essa solução reduz o espaço necessário para armazenamento, facilita o transporte e a organização do equipamento em ambientes terapêuticos, além de contribuir para sua praticidade durante a rotina clínica.

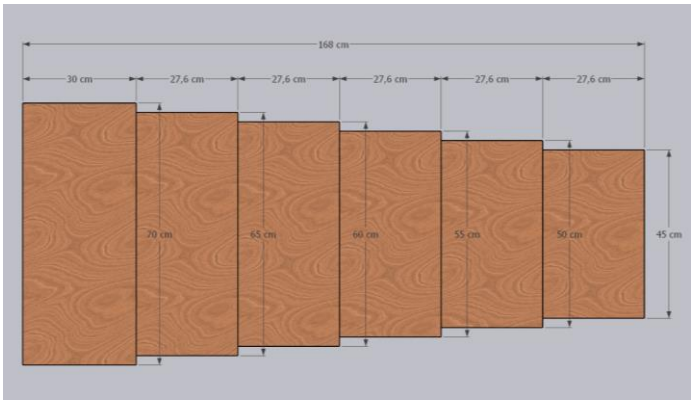


Fig. 5. Vista superior do conjunto de bancos terapêuticos.

Para facilitar o manuseio do conjunto e evitar a separação completa de seus componentes durante a retirada para uso, foram incorporadas abas de retenção em todos os módulos. Esses dispositivos atuam como limitadores mecânicos de deslocamento, impedindo o desprendimento completo entre as peças durante a abertura do conjunto (Fig. 6). Dessa forma, mantém-se o alinhamento lateral dos bancos, facilitando sua organização para utilização. Adicionalmente, a aba frontal do primeiro módulo foi dimensionada para funcionar como um puxador, permitindo a abertura e a organização do conjunto de forma prática.

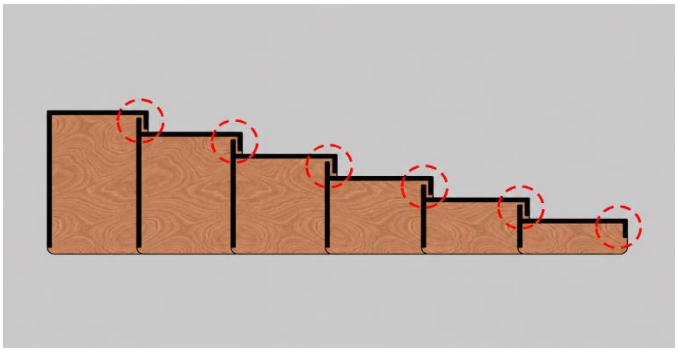


Fig. 6. Corte longitudinal do conjunto de bancos terapêuticos.

Com o objetivo de facilitar o deslocamento do equipamento, o projeto foi concebido de modo que apenas as laterais dos bancos permanecessem em contato com o piso. As extremidades inferiores dessas peças receberam cantos arredondados, permitindo que o conjunto transponha pequenas irregularidades superficiais com menor resistência ao arraste. Adicionalmente, a travessa posterior foi posicionada acima do nível do piso, reduzindo a área de contato durante o deslocamento e, consequentemente, o esforço necessário para a abertura e organização do conjunto. Seus cantos superiores receberam acabamento curvo, minimizando a possibilidade de interferência entre os módulos durante sua movimentação e evitando o travamento entre as peças. (Fig. 7).

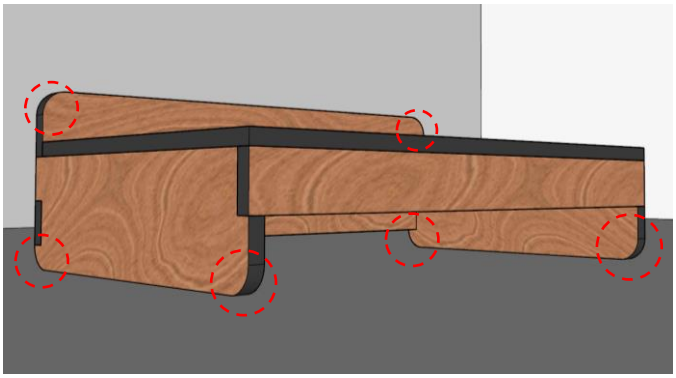


Fig. 7. Vista em perspectiva do primeiro módulo do conjunto.

As peças foram projetadas com encaixes intertravados do tipo dedo (*finger joints*), cuja geometria complementar favorece o acoplamento e o posicionamento entre as peças (Fig. 8). Em processos de fabricação digital, a precisão geométrica das superfícies de contato constitui um aspecto relevante para o adequado ajuste das juntas, uma vez que desvios podem resultar em folgas, interferências ou dificuldades de montagem [11]. A geometria desses encaixes atua como um sistema de orientação, restringindo combinações incompatíveis entre as peças e reduzindo a possibilidade de erros durante a montagem. Essa solução torna o processo construtivo mais intuitivo, contribuindo para que o equipamento possa ser montado por diferentes usuários, mesmo sem experiência prévia.

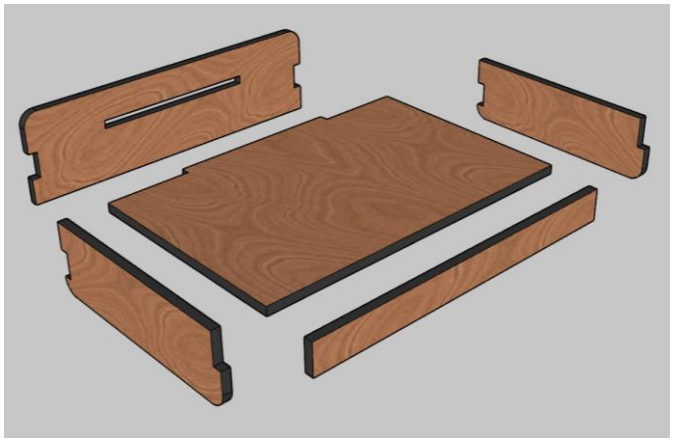


Fig. 8. Diagrama do primeiro módulo do conjunto.

De modo geral, as decisões de projeto priorizaram a simplicidade construtiva, a funcionalidade e a replicabilidade do equipamento, princípios alinhados à cultura maker e à fabricação digital. A utilização da modelagem tridimensional possibilitou validar virtualmente as soluções propostas antes da fabricação, reduzindo a necessidade de retrabalhos e contribuindo para a obtenção de um equipamento modular passível de reprodução.

B. Planificação e Organização das Peças

Em razão de incompatibilidades entre o formato de saída do SketchUp para Web e a interpretação dos arquivos pelo software RDWorks, tornou-se necessária uma etapa intermediária de revisão e compatibilização no AutoCAD. Nessa etapa, foram verificadas as dimensões, os encaixes e a disposição das peças no plano de corte, além da organização dos componentes para melhor aproveitamento da chapa de MDF. Posteriormente, o plano de corte foi exportado em formato PDF e importado para o RDWorks para configuração dos parâmetros operacionais e envio à máquina de corte a laser.

Cada módulo do conjunto foi concebido com cinco peças: duas laterais, uma plataforma superior, uma travessa frontal e uma travessa posterior (Fig. 9).

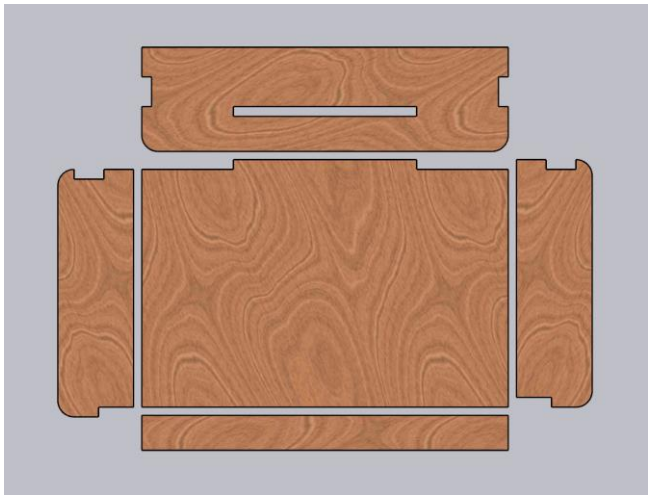


Fig. 9. Módulo planificado.

Uma vez realizada a planificação dos seis módulos e sua disposição em uma área correspondente a uma chapa padrão de MDF (2,75 m × 1,85 m), verificou-se que a fabricação de um conjunto completo de bancos terapêuticos consome aproximadamente 75% da área útil de uma única chapa (Fig. 10). Esse aproveitamento permite estimar que, com três chapas de MDF, seja possível fabricar quatro conjuntos completos de bancos, evidenciando a eficiência do plano de corte e o reduzido desperdício de material.

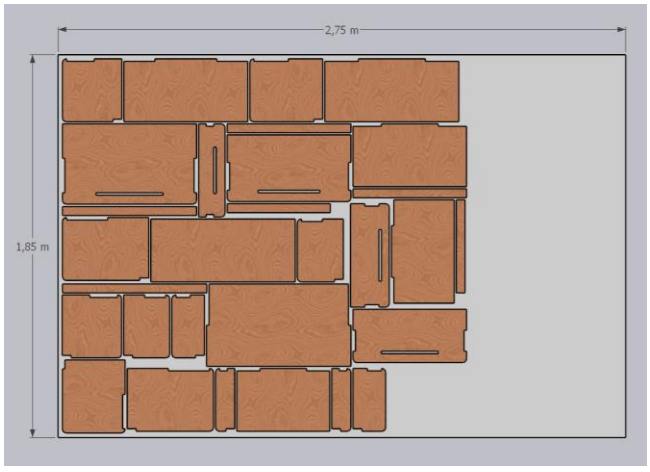


Fig. 10. Bancos terapêuticos planificados na área da chapa de MDF.

Após a exportação do modelo tridimensional das peças planificadas em formato STL, o arquivo foi convertido para o formato DXF por meio da plataforma online de acesso gratuito GroupDocs. Posteriormente, o arquivo DXF foi aberto no software AutoCAD, utilizando-se uma licença educacional, para a remoção de linhas excedentes, a conferência final das dimensões e a geração do arquivo em formato PDF.

Na impossibilidade de obtenção de uma licença educacional do AutoCAD, o software de código aberto LibreCAD pode ser

empregado como alternativa para essa etapa. Dessa forma, o emprego de ferramentas alternativas permite a reprodução do fluxo de trabalho sem a necessidade de aquisição de licença comercial. Em seguida, o arquivo em formato PDF foi importado para o software RDWorks, no qual foram definidos os parâmetros operacionais da máquina de corte a laser, incluindo potência, velocidade e sequência de corte.

A fabricação e montagem dos componentes foi realizada nas instalações do IPE Lab (Laboratório de Ideias, Prototipagem e Empreendedorismo), localizado no Parque Tecnológico Samambaia da Universidade Federal de Goiás (UFG). O laboratório disponibiliza infraestrutura voltada ao desenvolvimento de protótipos e processos de fabricação digital, atendendo demandas de projetos acadêmicos, empresas e inventores independentes, constituindo um ambiente adequado para o desenvolvimento e a validação de tecnologias [10].

O corte das peças foi realizado em uma máquina de corte a laser Delta CNC (Fig. 11), modelo 1310, com potência de 130 W e área útil de corte de 1,30 m × 1,00 m. Em razão das dimensões da área útil do equipamento serem inferiores às da chapa padrão de MDF (2,75 m × 1,85 m), a chapa foi previamente seccionada em quatro partes para adequação ao processo de fabricação. Dentre essas partes, apenas três foram utilizadas na fabricação dos módulos. Aplicando uma velocidade de 5 mm/s e com a potência do laser em 100%, o corte de todas as peças levou aproximadamente 3h para ser concluído.



Fig. 11. Delta CNC, modelo 1310.

Após o corte, os componentes foram separados e agrupados de acordo com o módulo correspondente, evitando a mistura de peças com dimensões distintas. A montagem foi realizada de forma progressiva, inicialmente com o posicionamento de dois componentes por meio dos encaixes previamente projetados. Em seguida, aplicou-se cola PVA nas superfícies de união e realizou-se a fixação com parafusos para madeira.

O procedimento foi repetido sequencialmente, incorporando-se uma peça por vez até a conclusão de cada módulo. A união progressiva dos componentes proporcionou maior estabilidade durante a montagem, facilitando o manuseio e o posicionamento das peças subsequentes. Após a montagem dos seis módulos, obteve-se o conjunto completo, cuja configuração final correspondeu à geometria e às dimensões definidas durante a etapa de modelagem digital (Fig. 12).


Fig. 12. Bancos terapêuticos montados.

Entre os componentes do conjunto, o módulo de maior altura foi considerado potencialmente mais suscetível aos efeitos de carregamentos excêntricos e esforços laterais, em razão de sua maior dimensão vertical. Para essa verificação, foi aplicada uma carga estática correspondente ao peso corporal de um indivíduo adulto com massa de 103 kg, aproximadamente centralizada sobre a superfície superior do módulo. Durante o procedimento, não foram observadas avarias aparentes, desprendimento das uniões ou deformações visíveis (Fig. 13).

Esse procedimento permitiu uma observação inicial da integridade estrutural do módulo sob carregamento estático, mas não caracteriza um ensaio mecânico padronizado, nem permite determinar sua capacidade máxima de carga, resistência a carregamentos excêntricos, estabilidade ou durabilidade. Dessa forma, a avaliação sistemática desses parâmetros constitui uma etapa necessária para estudos futuros, especialmente considerando a utilização do equipamento em atividades de reabilitação pediátrica de longo prazo.


Fig. 13. Aplicação estática de carga sob o módulo mais alto.

Como etapa final do processo de fabricação, as superfícies superiores de cada módulo foram revestidas com folhas de EVA de 4 mm de espessura. Posteriormente, foram aplicadas pinturas e elementos gráficos de temática infantil, conferindo identidade visual e caráter lúdico aos conjuntos produzidos. A personalização visual não seguiu um padrão único, resultando em diferentes composições gráficas entre os equipamentos fabricados (Fig. 14-15).



Fig. 14. Bancos terapêuticos mantidos no Lab EITA.



Fig. 15. Bancos terapêuticos mantidos no Lab EITA.

De modo geral, o processo de desenvolvimento permitiu transformar o modelo digital em um conjunto físico correspondente às características geométricas e construtivas definidas durante a etapa de projeto. A integração entre modelagem tridimensional, planificação, otimização do plano de corte e fabricação digital possibilitou a produção dos seis módulos com aproveitamento eficiente da matéria-prima e emprego de processos construtivos de relativa simplicidade. As soluções incorporadas ao projeto, como os encaixes intertravados, o sistema de retenção e a independência entre os módulos, contribuíram para facilitar a montagem, o manuseio e a organização do equipamento. Dessa forma, o processo adotado demonstrou a viabilidade de fabricação do conjunto por meio da infraestrutura de fabricação digital utilizada, permitindo sua posterior aplicação em atividades de avaliação funcional e práticas terapêuticas.

C. Custo Estimado de Materiais e Serviços Diretamente Empregados na Fabricação

Além das vantagens relacionadas à fabricação digital, a utilização do MDF mostrou-se economicamente viável. De acordo com os valores de referência do Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil (SINAPI), adotado como parâmetro oficial de custos da construção civil brasileira [12], uma chapa de MDF cru com espessura de 12 mm e dimensões de 2,75 m × 1,85 m apresentava custo aproximado de R$ 188,50, considerando a tabela de referência de maio de 2026. Como o plano de corte desenvolvido neste trabalho possibilita a fabricação de um conjunto completo utilizando cerca de 75% da área útil da chapa, estima-se que o custo da principal matéria-prima corresponda a aproximadamente R$ 142,00 por conjunto.

Como o Lab EITA é um projeto de pesquisa vinculado à Universidade Federal de Goiás (UFG), o custo do corte a laser foi calculado com base na tarifa subsidiada prevista na política institucional de utilização do IPE Lab para projetos de pesquisa vinculados à UFG e a outras instituições de ensino superior. Assim, o valor empregado neste estudo foi de R$ 270,00, correspondente à tarifa de R$ 1,50 por minuto de utilização da máquina. Para usuários externos não contemplados por essa política, a tarifa vigente é de R$ 2,50 por minuto.

O custo do revestimento em EVA foi estimado a partir da média de preços obtida junto a três fornecedores nacionais de materiais para artesanato e revestimento. Para o revestimento das superfícies superiores dos módulos foram utilizados aproximadamente 1,30 m² de EVA com espessura de 4 mm. Considerando um valor médio de mercado de R$ 33,00/m², estimou-se um custo de aproximadamente R$ 43,00 para esse material. Considerando os custos dos demais materiais e serviços empregados no processo de fabricação, o custo total estimado para a produção de um conjunto de bancos terapêuticos foi de aproximadamente R$ 530,00, conforme apresentado na Tabela II.

A montagem do conjunto foi realizada pelo próprio pesquisador, com duração aproximada de três horas. Embora esse período corresponda a um esforço de mão de obra, não foi atribuído um custo monetário a essa etapa, uma vez que a proposta do projeto está fundamentada na possibilidade de reprodução do conjunto por laboratórios makers, instituições de pesquisa e demais organizações que atuem no apoio à reabilitação de crianças com paralisia cerebral. Dessa forma, foram considerados na estimativa os custos diretamente relacionados aos materiais e serviços empregados na fabricação, sem incorporar remuneração de mão de obra, considerando a possibilidade de que a montagem seja realizada de forma voluntária ou como parte das atividades institucionais desses espaços.

Da mesma forma, os custos relacionados à pintura e aos materiais empregados na personalização visual dos bancos não foram incluídos na estimativa, uma vez que esses elementos não foram considerados componentes essenciais à funcionalidade estrutural e terapêutica do conjunto. Além disso, não foi adotado um padrão único de pintura, resultando em variações nos materiais e nas quantidades empregados em cada equipamento. Assim, esses elementos foram considerados etapas complementares e facultativas do processo de fabricação, não sendo incorporados à composição dos custos apresentada neste estudo.

Custos de Produção		
Material	***Metragem Utilizada***	***Valor***
Chapa de MDF	3,82 m²	R$ 142,00
Corte do MDF	Não se aplica	R$ 270,00
Cola Branca PVA Extra	Não se aplica	R$ 30,00
Folha de EVA	1,30 m²	R$ 43,00
Parafusos para Madeira	Não se aplica	R$15,00
Cola de Cianoacrilato	Não se aplica	R$ 30,00
VALOR TOTAL		**R$ 530,00**

TABELA II

Como parâmetro de comparação econômica, foram considerados os valores de aquisição de conjuntos comerciais compostos por seis bancos terapêuticos, confeccionados em diferentes materiais: compensado, com valor de R$ 1.200,00; e estrutura metálica, com valor de R$ 1.830,00. A média dos preços dos dois conjuntos comerciais foi de R$ 1.515,00, enquanto o conjunto desenvolvido neste estudo apresentou custo estimado de

fabricação de aproximadamente R$ 530,00. A diferença corresponde a R$ 985,00, representando um valor aproximadamente 65% inferior à média dos preços comerciais utilizados como referência. Esses resultados indicam o potencial da fabricação digital e da infraestrutura compartilhada de laboratórios makers para a redução dos custos associados à produção de equipamentos terapêuticos.

Ressalta-se, entretanto, que os valores comparados correspondem a categorias econômicas distintas, uma vez que o valor do equipamento desenvolvido representa o custo estimado de fabricação, enquanto os valores dos conjuntos comerciais correspondem aos preços de aquisição de produtos acabados, que podem incorporar despesas com mão de obra, logística, tributos, custos indiretos e margens comerciais. Dessa forma, a comparação possui caráter referencial e busca evidenciar o potencial de acessibilidade econômica do processo produtivo proposto, não constituindo uma comparação direta entre custos industriais equivalentes.

IV. Aplicação do Conjunto de Bancos Terapêuticos

O conjunto de bancos terapêuticos foi aplicado em situações reais de uso em três ambientes de atendimento e pesquisa localizados em Goiânia, Goiás: o Laboratório de Estudos Inventivos em Tecnologias Assistivas (Lab EITA), da Universidade Federal de Goiás (UFG); o Instituto dos Raros Christiane Toledo (IRCT); e a Associação Crianças de Microcefalia por Zika Vírus e Outras Deficiências (AMizika). O objetivo dessa etapa não foi avaliar a eficácia clínica do equipamento, mas verificar sua aplicabilidade prática, estabilidade durante o uso e potencial de incorporação à rotina de atendimento das instituições participantes.

O estudo foi submetido e aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Goiás (UFG), sob CAAE nº 87957025.7.0000.5083 e parecer nº 7.688.714. A participação das crianças nas atividades ocorreu mediante assinatura do Termo de Consentimento Livre e Esclarecido (TCLE) pelos respectivos responsáveis legais e, quando aplicável, do Termo de Assentimento pelos participantes. A utilização das imagens apresentadas neste trabalho foi autorizada pelos responsáveis legais, em conformidade com os procedimentos éticos aprovados no estudo.

No Lab EITA, o conjunto foi utilizado como recurso complementar durante a avaliação funcional das crianças, permitindo aos profissionais observar seu desempenho em diferentes tarefas motoras. As atividades realizadas nos módulos contribuíram para a observação de aspectos relacionados ao controle postural, equilíbrio e mobilidade, fornecendo informações complementares ao processo de avaliação para indicação e desenvolvimento de dispositivos de tecnologia assistiva, como o parapodium.

Já no Instituto dos Raros Christiane Toledo (IRCT) e na Associação Crianças de Microcefalia por Zika Vírus e Outras Deficiências (AMizika), o conjunto foi empregado como recurso de apoio às atividades terapêuticas desenvolvidas pelos profissionais. Os bancos foram utilizados tanto na composição de circuitos de atividades (Fig. 16) quanto em exercícios isolados (Fig. 17), nos quais apenas um ou alguns módulos eram empregados. Por serem independentes, os módulos podem ser organizados em diferentes configurações, permitindo aos profissionais adaptar sua disposição às atividades propostas e às capacidades funcionais de cada criança.



Fig. 16. Atividade terapêutica desenvolvida no IRCT.



Fig. 17. Atividade terapêutica desenvolvida no IRCT.

Os três ambientes mencionados utilizam o conjunto de bancos terapêuticos em suas atividades há aproximadamente quatro meses e, até o momento de elaboração deste artigo, não foram relatadas avarias nos módulos. Nesse período, o Lab EITA utilizou o equipamento com 10 crianças durante atividades de avaliação funcional. No IRCT, aproximadamente 240 crianças fizeram uso de um ou mais módulos do conjunto em atividades terapêuticas, enquanto na AMizika esse número foi de 35 crianças. Embora o período de acompanhamento e a ausência de ensaios mecânicos específicos não permitam estabelecer conclusões definitivas quanto à resistência e à durabilidade do equipamento, a manutenção da integridade física dos módulos durante o período de uso constitui um indicativo preliminar da adequação da solução construtiva adotada.

V. Considerações Finais

O desenvolvimento do conjunto de bancos terapêuticos demonstrou o potencial da cultura maker e da fabricação digital

como estratégias aplicáveis à produção de tecnologias assistivas destinadas ao contexto da reabilitação pediátrica. A utilização da modelagem digital associada ao corte a laser possibilitou a fabricação de um equipamento modular a partir de materiais de ampla disponibilidade comercial e de processos construtivos de relativa simplicidade. As soluções incorporadas ao projeto, como os módulos independentes com alturas progressivas, o sistema de recolhimento dos módulos para armazenamento e as abas de retenção, contribuíram para facilitar o manuseio, a organização e a adaptação do conjunto a diferentes atividades.

Sob o aspecto econômico, o custo estimado de fabricação de aproximadamente R$ 530,00 indica uma possibilidade de produção economicamente mais acessível quando comparado aos valores de aquisição dos conjuntos comerciais adotados como referência. Embora a comparação não represente uma equivalência direta entre custos de fabricação e preços de produtos comerciais acabados, a diferença observada evidencia o potencial da fabricação local e da infraestrutura compartilhada de laboratórios makers para reduzir os recursos financeiros necessários à obtenção desse tipo de equipamento.

A aplicação exploratória dos conjuntos no Lab EITA, no IRCT e na AMizika permitiu observar diferentes possibilidades de utilização em contextos reais. A utilização dos conjuntos durante aproximadamente quatro meses sem relatos de avarias constitui um indicativo preliminar de manutenção da integridade física dos módulos nas condições de uso observadas. Entretanto, o período de acompanhamento ainda é reduzido e a ausência de danos visíveis não permite estabelecer conclusões definitivas sobre a resistência mecânica, a durabilidade ou a segurança do equipamento em longo prazo.

Embora a verificação preliminar realizada não tenha identificado falhas aparentes sob a carga aplicada, a ausência de ensaios mecânicos padronizados constitui uma limitação deste estudo. Nesse sentido, estudos futuros deverão incluir ensaios mecânicos controlados para avaliação da capacidade de carga dos módulos, considerando cargas estáticas, dinâmicas e excêntricas, bem como a realização de testes cíclicos capazes de simular o uso repetitivo do equipamento ao longo do tempo. A verificação empírica realizada com a aplicação estática da massa corporal de um indivíduo de 103 kg sobre o módulo de maior altura constitui apenas uma observação preliminar e não substitui ensaios padronizados para determinação dos limites estruturais e dos fatores de segurança.

Outro aspecto a ser investigado refere-se ao comportamento do MDF diante das condições ambientais presentes nos locais de utilização. Por se tratar de um material derivado da madeira, torna-se relevante avaliar os efeitos da umidade, do contato frequente com o piso e dos procedimentos de higienização sobre a integridade dos módulos. Estudos posteriores poderão analisar a aplicação de seladores, revestimentos ou outros sistemas de proteção superficial capazes de reduzir a absorção de umidade e facilitar a limpeza do equipamento, sem comprometer sua acessibilidade econômica e sua possibilidade de fabricação em laboratórios makers.

Na versão atual, os arquivos digitais ainda não são disponibilizados publicamente, constituindo essa uma limitação quanto à reprodução imediata do equipamento por terceiros. Como perspectiva de continuidade do trabalho, após uma avaliação mais abrangente do equipamento, pretende-se disponibilizar os arquivos digitais do projeto em diretórios públicos voltados ao compartilhamento de projetos de fabricação digital, como o Vetor Maker. Essa medida deverá contribuir para ampliar a replicabilidade do conjunto, permitindo que outros laboratórios, instituições e usuários interessados tenham acesso aos arquivos necessários para sua reprodução e eventual adaptação às suas necessidades.

Mais do que apresentar um conjunto de bancos terapêuticos de menor custo, este trabalho demonstra a viabilidade de um processo estruturado de desenvolvimento baseado em fabricação digital e princípios da cultura maker. A integração entre levantamento de requisitos, modelagem CAD, otimização do plano de corte, fabricação e aplicação em ambientes de atendimento estabelece um fluxo potencialmente reproduzível por laboratórios universitários, serviços de reabilitação e espaços makers, contribuindo para o desenvolvimento de tecnologias assistivas personalizáveis e economicamente acessíveis.

REFERÊNCIAS

[1] Brasil. Subsecretaria Nacional de Promoção dos Direitos da Pessoa com Deficiência. Comitê de Ajudas Técnicas, Tecnologia Assistiva. Brasília, DF, Brazil: CORDE, 2009, 138 p.

[2] World Health Organization and United Nations Children·s Fund, Global report on assistive technology. Geneva, Switzerland: WHO, 2022.

[3] M. Bax, ·Definition of cerebral palsy,· Developmental Medicine and Child Neurology, vol. 6, no. 3, pp. 295·297, 1964.

[4] P. Rosenbaum et al., ·A report: the definition and classification of cerebral palsy April 2006, Developmental Medicine & Child Neurology Supplement, vol. 109, pp. 8·14, 2007, doi: 10.1111/j.1469-8749.2007.tb12610.x.

[5] B. Dan et al., ·Proposed updated description of cerebral palsy,· Developmental Medicine & Child Neurology, vol. 67, no. 6, pp. 700·709, 2025, doi: 10.1111/dmcn.16274.

[6] D. Dougherty, ·The Maker Movement,· Innovations: Technology, Governance, Globalization, vol. 7, no. 3, pp. 11·14, 2012.

[7] N. Gershenfeld, Fab: The Coming Revolution on Your Desktop: From Personal Computers to Personal Fabrication. New York, NY, USA: Basic Books, 2005.

[8] S. W. Blundell, R. B. Shepherd, C. M. Dean, R. D. Adams, and B. M. Cahill, ·Functional strength training in cerebral palsy: a pilot study of a group circuit training class for children aged 4·8 years,· Clinical Rehabilitation, vol. 17, pp. 48·57, 2003.

[9] H. A. Eltawahni, A. G. Olabi, and K. Y. Benyounis, ·Investigating the CO· laser cutting parameters of MDF wood composite material,· Optics & Laser Technology, vol. 43, no. 3, pp. 648·659, 2011.

[10] Universidade Federal de Goiás, Portaria no. 2774, de 18 de setembro de 2020. Goiânia, Brazil: UFG, 2020. [Online]. Available: https://files.cercomp.ufg.br/weby/up/1229/o/Regulamento_Rede_IPElab-VF-Assin%282%29.pdf. Accessed: Jul. 27, 2026.

[11] A. Ganeshan et al., ·MiGumi: making tightly coupled integral joints millable,· ACM Transactions on Graphics, vol. 44, no. 6, Art. no. 193, Dec. 2025, doi: 10.1145/3763304.

[12] Caixa Econômica Federal and Instituto Brasileiro de Geografia e Estatística, Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil (SINAPI), reference May 2026. Brasília, Brazil, 2026. [Online]. Available: https://www.caixa.gov.br/poder-publico/modernizacao gestao/sinapi/Paginas/default.aspx. Accessed: Jul. 9, 2026.

DCV: dispositivo de comunicação vibrotátil personalizável para pessoas surdas e surdocegas

L.C Cantieri[1], L.O Martinz[4], F.P Cantieri[3], G. Arruda[3], R. F. Rohrich[4], J.C. Bassan[1] and A.R Cantieri[2]

[1] Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Engenharia Biomédica, Curitiba, Brasil
[2] Instituto Federal do Paraná/ Campus Pinhais, Laboratório de Robótica e Computação Aplicadas, Pinhais, Brasil
[3] Escola Superior de Educação Física, Universidade de Pernambuco, Recife, Brasil
[4] Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Engenharia Engenharia Elétrica e Informática Industrial, Curitiba, Brasil

Abstract— **This article presents the development of a customizable vibrotactile communication device (DCV) designed to support deaf and deafblind individuals. The device receives commands from a mobile application via Wi-Fi and converts them into vibration patterns that can be freely configured by the user, who assigns each pattern a specific meaning. The DCV consists of an ESP32-C3 Mini board, an ERM (Eccentric Rotating Mass) vibration motor, a removable battery, and a 3D-printed case, and is attached to the body using adhesive Kinesiotape. Preliminary laboratory tests were conducted to evaluate participants' recognition of vibrotactile commands. The results showed a recognition accuracy of 93.33%, indicating that participants were able to correctly associate most vibration patterns with the corresponding commands after a brief familiarization period. The proposed device represents a low-cost, portable, and customizable platform with the potential to contribute to the development of accessible communication solutions and to support future research on vibrotactile interaction for deaf and deafblind individuals.**

Keywords— **Assistive technology; vibrotactile communication; deafblindness; deafness; haptic feedback; wearable device.**

I. INTRODUÇÃO

A comunicação constitui uma necessidade constante nas relações humanas, em diversos aspectos de suas vidas. Para pessoas surdas e surdocegas, entretanto, barreiras comunicacionais podem dificultar interações cotidianas e trocas mínimas necessárias em diferentes contextos (1).

A perda auditiva representa um desafio significativo em diferentes contextos da sociedade, podendo dificultar o acesso à informação, a interação com outros indivíduos, a realização de atividades cotidianas, assim como reduzir sua autonomia em situações que dependem de comunicação rápida e clara (2).

Caracterizada pela ausência simultânea da visão e da audição, a surdocegueira impõe desafios particulares, uma vez que estratégias utilizadas por pessoas com apenas uma deficiência sensorial nem sempre podem ser empregadas da mesma forma. Nesse contexto, a língua de sinais tátil destaca-se como uma das formas de comunicação mais utilizadas por pessoas surdocegas, permitindo a percepção de sinais por meio do contato físico e do tato (3).

Tecnologias assistivas voltadas à deficiência auditiva podem ser classificadas em cinco categorias principais: conversão de fala em texto, conversão de texto em fala, sistemas de exibição de informações, mecanismos de feedback háptico ou visual e recursos de suporte para dispositivos móveis (4). Para pessoas com surdocegueira, tecnologias baseadas em feedback háptico estão entre as mais utilizadas (3).

Entre os sistemas hápticos, a tecnologia vibrotátil tem sido utilizada em dispositivos assistivos voltados a pessoas surdas e surdocegas como uma forma alternativa de transmissão de informações. Nesses recursos, sinais, mensagens e alertas podem ser convertidos em estímulos percebidos pela pele, permitindo sua aplicação em dispositivos que convertem texto em fala, fornecem alertas de segurança e oferecem suporte à comunicação, entre outros usos (5).

Entre as tecnologias assistivas voltadas a pessoas surdas, os dispositivos vestíveis baseados em vibração podem ser utilizados para sinalizar informações de forma tátil ao usuário. Em um estudo com um dispositivo vestível testado em pessoas surdas, observou-se a transmissão de sons importantes do ambiente por meio de vibrações, permitindo seu reconhecimento em tempo real (6).

A vibração já vem sendo explorada em recursos portáteis e acessíveis para transmissão de alertas relevantes a pessoas surdas. Nesse contexto, sistemas de sinalização vibratória em dispositivos móveis têm sido investigados para a emissão de alertas de emergência, demonstrando o potencial da vibração como recurso assistivo em situações cotidianas que exigem percepção rápida e segura de informações (7).

Esse potencial se estende também à percepção musical. (8) demonstraram, por meio de registros eletroencefalográficos, que estímulos vibrotáteis aplicados nas mãos de pessoas com perda auditiva severa foram capazes de ativar as mesmas regiões cerebrais envolvidas na percepção emocional da música em ouvintes, como o córtex auditivo e o córtex temporal superior. Esses achados indicam que a vibração não representa apenas uma alternativa sensorial, mas um canal capaz de acionar respostas neurais semelhantes às produzidas pelo som.

Apesar dos avanços observados no desenvolvimento de dispositivos vibrotáteis assistivos, a possibilidade de personalização dos padrões vibratórios pelo próprio usuário ainda é pouco explorada na literatura. De modo geral, os dispositivos desenvolvidos até o momento operam com padrões vibratórios pré-definidos, voltados a funções específicas, como alertas sonoros, navegação ou comunicação em contextos delimitados. Como consequência, o usuário possui poucas possibilidades de adaptar os estímulos às suas necessidades, preferências e contexto de utilização. Flores Ramones e del-Rio-Guerra destacam que a literatura ainda carece de dispositivos hápticos configuráveis, capazes de permitir ajustes em parâmetros como intensidade da vibração, limiar de detecção sonora, seleção dos sons monitorados e padrões vibratórios, proporcionando uma experiência personalizada de acordo com os requisitos de cada usuário (5).

O presente artigo propõe a utilização de um dispositivo vibrotátil com capacidade de modulação de vibrações para uso personalizado no cotidiano de pessoas surdas e surdocegas. Os usuários podem escolher padrões de vibração e correlacioná-los com os significados de comunicação desejados.

O restante deste artigo está organizado da seguinte forma: na Seção II descrevemos o desenvolvimento do dispositivo, detalhando suas especificidades; na Seção III demonstramos o funcionamento do aplicativo e a forma como pode ser programado; na Seção IV apresentamos o posicionamento do dispositivo no corpo; na Seção V são apresentadas algumas possibilidades de uso; na Seção VI são apresentados os testes preliminares em ambiente de laboratório e as análises dos resultados obtidos; na Seção VII é realizada a discussão dos resultados; por fim, na Seção VIII são apresentadas as conclusões e os trabalhos futuros.

II. Descrição do Dispositivo de Comunicação Vibrotátil

O dispositivo de comunicação vibrotátil (DCV) foi desenvolvido com foco na portabilidade, leveza e simplicidade construtiva, de forma que pudesse ser utilizado em diferentes contextos sem exigir adaptações estruturais. Essas diretrizes orientaram diretamente a escolha dos componentes e o design do conjunto, resultando em um sistema compacto e de baixo custo, com potencial de uso amplo por pessoas surdas e surdocegas em situações cotidianas.

Os componentes eletrônicos do DCV são apresentados na Tabela 1 e consistem em uma placa ESP32-C3 Mini com conectividade Wi-Fi integrada, um motor de vibração do tipo ERM versão mini, uma bateria removível de 200 mAh e 3,7 V com autonomia aproximada de duas horas, e um LED indicador de funcionamento.

Table 1: Componentes Eletrônicos do DCV

Componente	**Características**
Placa ESP32-C3 Mini	Placa compacta e programável com conectividade Wi-Fi integrada, utilizada para controlar os sinais enviados aos motores.
Bateria removível	Capacidade: 200 mAh, tensão: 3,7V.
Motor de vibração ERM	Utilizado para geração dos estímulos vibrotáteis.
LED vermelho	Indicador de funcionamento, sinaliza que o dispositivo está ligado e conectado ao aplicativo.

O motor de vibração do tipo ERM (Eccentric Rotating Mass) foi escolhido por sua simplicidade estrutural, baixo custo e facilidade de integração ao conjunto (9). A potência do motor oferece uma janela maior de ajuste de intensidade que permite ao usuário encontrar o nível de sensibilidade mais adequado à sua percepção. Além disso, motores ERM produzem forças de cisalhamento lateral na pele que, combinadas com vibrações verticais, evocam sensações táteis robustas mesmo em regiões com menor densidade de mecanorreceptores (9).

A placa ESP32-C3 Mini é responsável pelo controle e pela transmissão dos sinais do DCV. Ela opera por meio de uma rede Wi-Fi própria, gerada pelo próprio dispositivo, à qual o aplicativo de controle se conecta diretamente. A estimativa de alcance, avaliada nos testes preliminares, é de até 10 metros em ambientes internos com paredes entre o emissor e o receptor, de até 50 metros em ambientes internos sem

barreiras e 100 metros em ambientes abertos. A autonomia em uso varia conforme o número de sinais enviados, uma vez que o maior consumo é referente ao motor vibratório. Em estado de repouso, apenas com a placa de controle ativa, o dispositivo apresentou uma autonomia média de mais de 24 horas sem necessidade de recarga.

O Sistema foi concebido buscando garantir mobilidade e liberdade de movimento ao usuário. Seu baixo consumo energético e suas dimensões reduzidas de 22,5 x 18 mm contribuem para tornar o dispositivo compacto e portátil. O case do DCV foi projetado em software de modelagem tridimensional e impresso em 3D, com dimensões de 7,0 x 4,0 x 2,0 cm e bordas arredondadas para conforto no contato com a pele. Uma divisória interna separa o motor de vibração dos demais circuitos eletrônicos conforme ilustrado na Figura 1. Para fixação ao corpo, foi utilizada uma bandagem elástica adesiva do tipo Kinesiotape, escolhida por sua aderência e compatibilidade com uso direto na pele.



Fig. 1: Projeto 3D do case do DCV: visão geral do conjunto (superior esquerda), visão externa fechada (superior direita), visão interna mostrando a divisória (inferior esquerda) e visão da base (inferior direita). Fonte: A autora.

O conjunto de componentes escolhidos reflete diretamente os princípios que orientaram o desenvolvimento do DCV, compondo um sistema funcional, compacto e de baixo custo, com potencial de uso em diferentes contextos cotidianos.

III. Aplicativo de Controle

Uma das características centrais do DCV é a possibilidade de personalização dos padrões vibratórios pelo próprio usuário, que pode configurar cada sinal de acordo com suas necessidades e atribuir a ele o significado que desejar.

O DCV é controlado por um aplicativo móvel desenvolvido para o sistema Android, que se comunica com o dispositivo por meio de rede Wi-Fi. A interface conta com oito botões programáveis, por meio dos quais o usuário define os parâmetros de cada sinal vibratório, incluindo quantidade de pulsos, duração e intensidade da vibração. O usuário pode escolher quais botões estão ativos ou não, caso não necessite de um número menor de comandos, diminuindo assim a possibilidade de erro de acionamento durante o uso.

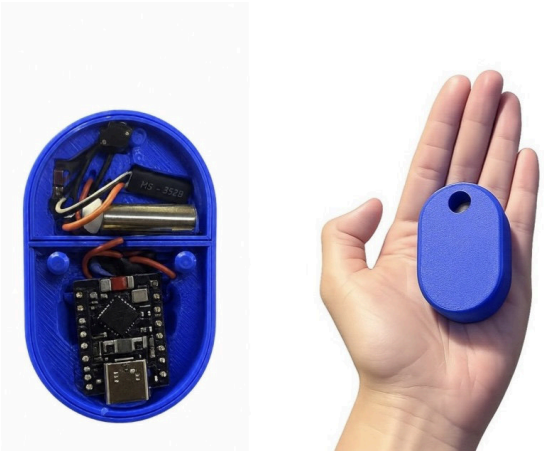


Fig. 2: Dispositivo de Comunicação Vibrotátil (DCV), apresentando o circuito eletrônico interno (à esquerda) e o dispositivo posicionado sobre a palma da mão para referência de suas dimensões (à direita).

Todos os botões podem ser configurados individualmente, permitindo a criação de padrões vibratórios distintos. A configuração individual de cada botão é realizada em uma aba específica no aplicativo. Para cada botão o usuário pode programar a emissão de um a três sinais vibratórios em cada execução, cada padrão composto por uma combinação de vibrações ”curtas”, ”médias” e ”longas”, o que permite uma grande flexibilidade na criação dos sinais de comunicação desejados. É possível também utilizaer um número menor de sinais programados em cada botão, utilizando-se na configuração do parâmetro ”zero”, que transmite um sinal nulo naquela combinação.

Os tempos de vibração ”longo”, ”médio” e ”curto” são configurados independentemente através de três *sliders*, e são os mesmos para todos os botões ativos no aplicativo. Isso permite a programação de uma grande gama de padrões, progamados de acordo com a necessidade da utilização ou experimento proposto. As figuras 3 e 4 apresentam as telas dos botões para envio dos comandos, a configuração dos tempos e potência e a tela da programação dos padrões, respectivamente.

Ao pressionar o botão, o padrão programado é transmitido ao dispositivo em aproximadamente 4 milissegundos, garantindo uma resposta praticamente imediata. Dessa forma,

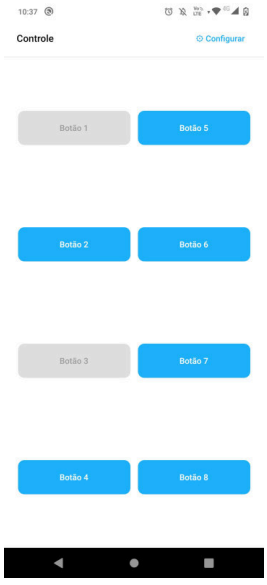


Fig. 3: Interface do aplicativo de controle do DCV, contendo cinco botões programáveis para envio dos comandos vibratórios e controles para ajuste do tempo e da intensidade da vibração.

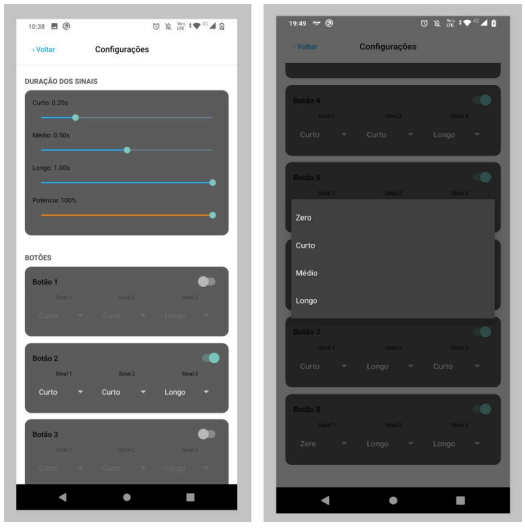


Fig. 4: Tela de configuração dos padrões de cada botão: ”Zero”, ”Curto”, ”Médio” e ”Longo” e tela de configuração dos tempos e potência.

o usuário pode associar cada padrão vibratório a um significado próprio, como uma comunicação previamente combinada entre as pessoas envolvidas.

Por meio do aplicativo, é possível conectar mais de um DCV simultaneamente, permitindo que a mesma mensagem vibrotátil seja transmitida para diferentes usuários. Essa característica amplia o potencial de aplicação do sistema em contextos coletivos, como atividades educacionais, esportivas, treinamentos, deslocamentos guiados ou situações em que um mesmo comando precisa ser recebido por várias pessoas de forma discreta e sincronizada.

IV. Posicionamento do Dispositivo

No trabalho apresentado em (10) avaliou-se a sensibilidade vibrotátil em 16 posições corporais distintas. A Figura 5 apresenta um mapeamento de sensibilidade para o corpo humano. A escala de cores indica áreas de alta sensibilidade (verde) e sensibilidade média a baixa (amarelo), enquanto os marcadores em roxo exemplificam possíveis locais de implementação do DCV.

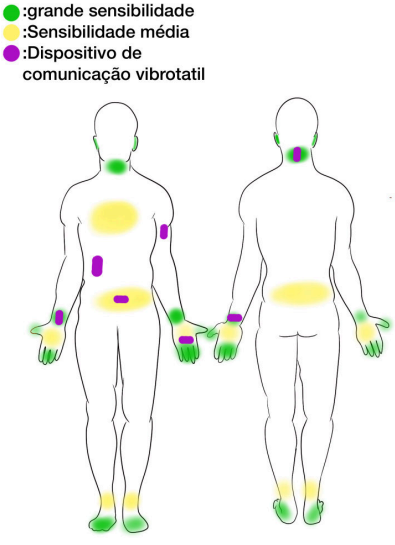


Fig. 5: Mapeamento de sensibilidade vibrotátil para o corpo humano. Áreas em verde indicam alta sensibilidade, amarelo indicam sensibilidade média a baixa, e marcadores em roxo exemplificam possíveis locais de implementação do DCV. Fonte: adaptado de (10).

O posicionamento do dispositivo no corpo é personalizado de acordo com a preferência do usuário. O DCV pode ser acoplado em qualquer região do corpo por meio de Kinesiotape, uma bandagem elástica adesiva que se adapta a diferentes superfícies e oferece fixação confortável e segura, permitindo ao usuário escolher o local que melhor se adequa às suas necessidades e sensibilidade individual.

A sensibilidade vibrotátil pode ser estudada e mapeada; apesar disso, a percepção da vibração varia de forma individual, assim como o conforto e a preferência pelo local de uso. Portanto, mesmo que algumas regiões sejam

indicadas como mais sensíveis, o usuário pode ter uma sensibilidade que considere melhor para o dispositivo, ou sentir mais conforto de uso. Assim, o DCV pode ter sua localização personalizada para o usuário.

V. Possibilidades de Aplicação do DCV

O DCV foi desenvolvido para se adaptar a diferentes necessidades e contextos de uso, considerando que a comunicação pode mudar conforme a situação, o ambiente e a preferência de cada usuário. A personalização dos padrões vibratórios permite que o dispositivo transmita mensagens simples e específicas, como instruções de "inicie" e "pare", alertas do tipo "preste atenção em mim" e respostas como "sim" e "não".

Em sua versão atual, como protótipo funcional, a aplicação tanto em pesquisa quanto para uso cotidiano já é viável e acessível, uma vez que sua construção é bastante simples tanto em relação à eletrônica quanto em relação ao design 3D. Um arquivo público online está sendo elaborado para disponibilizar os arquivos STL, o código do ESP, o aplicativo, circuito esquemático e manuais de montagem para permitir sua reprodução pelo público interessado.

Os padrões vibratórios podem ser personalizados de acordo com a necessidade de cada usuário. Por exemplo, uma pessoa surdocega acompanhada por alguém que não conhece Braille ou outra forma específica de comunicação pode combinar diferentes padrões vibratórios para transmitir informações simples do dia a dia. Um padrão pode indicar que a pessoa acabou de entrar no ambiente, outro pode informar que ela está saindo, enquanto um terceiro pode identificar quem está presente, como "é a Maria". Dessa forma, mesmo sem utilizar uma linguagem formal ou conhecimentos prévios, torna-se possível estabelecer uma comunicação simples e útil para situações cotidianas.

Em situações nas quais o usuário precisa manter a atenção em uma atividade principal, sem deixar de receber comunicações secundárias, o dispositivo pode ser utilizado como um sistema de alertas discretos. Por exemplo, durante uma aula, um estudante surdo pode estar concentrado na escrita de um exercício enquanto o professor deseja chamar sua atenção. Em vez de caminhar até a carteira ou tocar no aluno, o professor pode enviar um padrão vibratório previamente combinado com o significado "olhe para mim". Em algumas situações, essa abordagem pode evitar que o estudante receba uma forma de chamada diferente da utilizada com os demais colegas, favorecendo uma interação mais natural e integrada ao ambiente da sala de aula.

Durante uma atividade física, o dispositivo também pode ajudar o treinador a transmitir comandos simples. Em uma aula coletiva, uma pessoa surda pode estar pulando corda, um movimento que exige manter o olhar fixo para conservar o ritmo. Nessa situação, olhar para o professor pode prejudicar a execução do exercício. Com o DCV, comandos como "iniciar" e "parar" podem ser enviados por meio de padrões vibratórios, sem interromper o movimento. Para pessoas surdocegas, essa forma de comunicação se torna ainda mais importante, pois não depende da audição nem da visão.

Situação semelhante pode ser observada durante a execução de uma flexão de braço. Nesse exercício, o praticante permanece próximo ao chão e com o olhar voltado para a frente, o que dificulta acompanhar a contagem visual realizada pelo treinador. Para pessoas surdocegas, essa barreira é ainda maior, pois a contagem não pode ser percebida nem pela audição nem pela visão. Nesse caso, o DCV pode transmitir as contagens por meio de vibrações, permitindo que a informação seja recebida pelo tato.

Os exemplos apresentados ilustram apenas algumas das possibilidades de aplicação do DCV. O dispositivo estabelece um canal de comunicação intuitivo, rápido e prático para informações simples, sem exigir conhecimento prévio de linguagens específicas como Braille ou língua de sinais. Qualquer pessoa pode utilizá-lo, criando pontes de comunicação acessíveis e abrindo possibilidades para pessoas surdas e surdocegas se comunicarem com qualquer outra pessoa ao seu redor.

A. Aplicação do Dispositivo em Pesquisas Aplicadas em Comunicação Vibrotátil

Uma das dificuldades encontradas pelos pesquisadores da área de comunicação e dispositivos assistivos é justamente a disponibilidade e o custo para aquisição de equipamentos que permitam a construção do arranjo experimental. Não é fácil encontrar dispositivos desenvolvidos especificamente para a aplicação neste tipo de pesquisa.

Uma alternativa é o grupo de pesquisa desenvolver as suas próprias ferramentas, tarefa não trivial e que consome muito tempo. Dessa forma, demonstra-se que o DCV também pode ser utilizado como ferramenta de apoio em pesquisas desenvolvidas na área da comunicação vibrotátil para pessoas surdas e surdocegas. Por ser um dispositivo funcional e configurável, ele permite que pesquisadores investiguem diferentes aplicações da comunicação tátil sem a necessidade de desenvolver um novo protótipo desde o início, concentrando seus esforços no planejamento experimental e na análise dos resultados.

Entre as possibilidades de aplicação estão pesquisas sobre percepção vibrotátil, sensibilidade em diferentes regiões do corpo, posicionamento do dispositivo, reconhecimento

de padrões vibratórios e comparação entre diferentes estratégias de comunicação. O dispositivo também pode servir como base experimental para estudos envolvendo códigos, sequências e alfabetos vibrotáteis, contribuindo para o desenvolvimento de novas formas de comunicação para pessoas surdas e surdocegas.

Além disso, o DCV pode ser empregado em estudos preliminares para avaliar a viabilidade de novas aplicações da comunicação vibrotátil em diferentes contextos. Mesmo quando o objetivo final da pesquisa for o desenvolvimento de outro dispositivo, o DCV pode ser utilizado como plataforma experimental para validar conceitos, testar hipóteses e orientar o desenvolvimento de novas tecnologias assistivas.

VI. Testes Preliminares em Ambiente de Laboratório

Os testes preliminares foram conduzidos em ambiente laboratorial controlado para explorar, de forma inicial, se a comunicação vibrotátil seria funcional. Essa etapa exploratória buscou observar, em um primeiro momento, se os sinais vibratórios poderiam ser percebidos e diferenciados pelos participantes, servindo como investigação inicial sobre a viabilidade da tecnologia. Todos os participantes eram pesquisadores do laboratório.

Participaram do teste três voluntários adultos ouvintes, sem familiarização prévia com o dispositivo vibrotátil. Durante a realização das atividades, os participantes utilizaram fones de ouvido reproduzindo música em volume elevado, enquanto o ambiente de teste permaneceu silencioso e com mínima interferência externa.

Antes do início da coleta, foi realizado um período de treinamento com duração aproximada de cinco minutos. Nessa etapa, cada participante teve contato com os quatro padrões vibratórios utilizados no teste e recebeu a explicação correspondente a cada movimento. Esse treinamento teve curta duração e foi utilizado apenas para familiarização inicial com o dispositivo e com os comandos, permitindo que os voluntários reconhecessem previamente a relação entre cada estímulo e a ação corporal esperada.

Table 2: Padrões vibrotáteis utilizados no teste preliminar.

Código	Padrão vibratório	Movimento
V1	Uma vibração curta	Levantar braço direito
V2	Duas vibrações curtas	Levantar braço esquerdo
V3	Uma vibração longa	Dar um passo para frente
V4	Duas vibrações longas	Dar um passo para trás

Para a realização do teste, o dispositivo foi posicionado na região superior e externa do braço, próximo ao músculo deltóide. Esse local foi escolhido por oferecer boa área de contato para fixação, permitir a percepção clara dos estímulos vibratórios e apresentar baixa interferência na execução dos movimentos solicitados. Além disso, essa posição favoreceu a estabilidade do equipamento durante as respostas motoras dos participantes.

O aplicador permaneceu com o dispositivo móvel posicionado atrás de um anteparo, impedindo que o participante visualizasse o acionamento dos comandos no aplicativo. Os sinais vibratórios foram enviados de forma aleatória entre os quatro padrões estabelecidos, evitando uma sequência previsível de estímulos. Cada participante realizou três séries de vinte comandos, totalizando 60 movimentos por voluntário e 180 movimentos no conjunto do teste.

A. Resultado e Análise

Os resultados dos testes preliminares foram analisados a partir das respostas motoras realizadas pelos três voluntários durante a execução dos comandos vibrotáteis. Ao todo, foram avaliados 180 movimentos, correspondentes às três séries de vinte comandos realizadas por cada participante. Para esta análise, considerou-se como acerto a execução do movimento previamente associado ao padrão vibratório enviado, enquanto o erro foi registrado quando não houve resposta correta ao comando.

De modo geral, os participantes apresentaram bom desempenho na identificação dos comandos vibrotáteis. Dos 180 movimentos realizados, 168 foram executados corretamente e 12 foram classificados como erro, correspondendo a 93,33% de acertos e 6,67% de erros. Esse resultado demonstra que, mesmo após um curto período de familiarização inicial e sendo este o primeiro contato dos voluntários com o dispositivo, a maior parte dos estímulos vibratórios foi percebida e associada corretamente aos movimentos previamente estabelecidos (Figura 6).

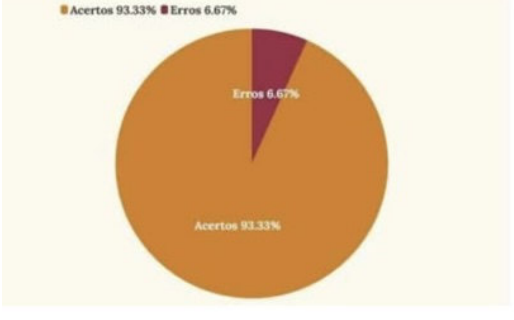


Fig. 6: Distribuição percentual de acertos e erros obtidos no teste de reconhecimento dos padrões vibratórios.

A análise por padrão vibratório permitiu observar

diferenças no reconhecimento dos estímulos. O comando V1 apresentou 49 acertos e 6 erros, enquanto o V2 registrou 38 acertos e 5 erros. O V3 obteve o melhor desempenho entre os padrões analisados, com 42 acertos e nenhum erro registrado. Já o V4 apresentou 39 acertos e apenas 1 erro. Esses dados indicam que todos os padrões foram reconhecidos pela maioria dos participantes, com destaque para os sinais longos, que apresentaram menor ocorrência de erros em comparação aos sinais curtos(Figura 7).



Fig. 7: Distribuição do número de acertos e erros para cada padrão vibratório avaliado (V1 a V4).

Além dos dados quantitativos, foi realizada uma análise visual da interação dos participantes com o dispositivo durante o teste. De acordo com a observação dos aplicadores, o equipamento permaneceu estável no local de fixação e não apresentou sinais aparentes de desconforto ou interferência na execução dos movimentos solicitados. Os participantes conseguiram responder aos estímulos durante a prática e demonstraram rápida assimilação da relação entre os padrões vibratórios e os movimentos correspondentes. Esse comportamento, associado ao elevado número de acertos, reforça a viabilidade inicial da comunicação vibrotátil para a transmissão de comandos simples em ambiente controlado.

VII. Discussão

Os resultados obtidos nos testes preliminares indicam a viabilidade inicial do DCV como recurso de comunicação vibrotátil simples. Mesmo com um curto período de familiarização e sendo este o primeiro contato dos voluntários com o dispositivo, os participantes conseguiram associar a maior parte dos padrões vibratórios aos movimentos previamente combinados. Esse desempenho sugere que estímulos vibrotáteis podem ser assimilados de maneira rápida quando organizados em comandos simples, permitindo que a vibração funcione como um canal alternativo para a transmissão de informações básicas.

Além dos resultados funcionais observados, o DCV apresenta características que reforçam sua proposta como dispositivo portátil e acessível. O sistema foi desenvolvido a partir de um conjunto compacto, com componentes simples, de baixo custo e disponíveis no mercado, associados a um aplicativo próprio de controle. Essa estrutura permite uma lógica de funcionamento direta e adaptável, sem exigir equipamentos complexos ou alto investimento para sua produção.

Outro aspecto relevante do DCV é sua versatilidade para atender diferentes necessidades comunicacionais de pessoas surdas e surdocegas. Por ser portátil, vestível e controlado por aplicativo, o dispositivo pode ser utilizado em situações cotidianas que exigem sinais rápidos e objetivos, como alertas, chamadas de atenção, orientações simples e comunicações previamente combinadas. Dessa forma, o DCV não se limita a uma função única, mas se apresenta como um recurso de comunicação tátil ajustável ao contexto e à necessidade do usuário.

O principal diferencial do DCV está na personalização dos padrões vibratórios e de seus significados. Diferentemente de tecnologias vibrotáteis assistivas com padrões previamente definidos, o dispositivo permite configurar intensidade, duração e quantidade de pulsos pelo aplicativo, associando cada sinal a uma mensagem combinada pelo usuário. Essa característica dialoga com a necessidade de dispositivos vibrotáteis mais programáveis e adaptáveis (5), permitindo que a vibração funcione como uma linguagem tátil simples, personalizada e compartilhada em diferentes situações do cotidiano.

Nesse sentido, os resultados preliminares e as características do DCV convergem para uma mesma compreensão: a vibração pode ser utilizada como uma forma simples de comunicação tátil quando organizada em padrões reconhecíveis e adaptáveis ao usuário, especialmente para apoiar pessoas surdas e surdocegas em situações que exigem comunicação rápida e acessível.

VIII. Conclusões e trabalhos futuros

Este artigo apresentou o desenvolvimento de um dispositivo de comunicação vibrotátil voltado ao apoio comunicacional de pessoas surdas e surdocegas. A construção do DCV mostrou-se relevante por reunir portabilidade, simplicidade estrutural, baixo custo dos componentes e possibilidade de controle por aplicativo móvel. Essas características aproximam o dispositivo de uma proposta de tecnologia assistiva acessível, pensada para uso prático e adaptável a diferentes contextos.

Os testes preliminares indicaram que o dispositivo

apresenta potencial para a transmissão de comandos simples por meio da vibração. Mesmo com um curto período de familiarização, os participantes conseguiram associar a maior parte dos padrões vibratórios aos movimentos previamente combinados, sugerindo que a comunicação vibrotátil pode ser compreendida de forma rápida quando organizada em sinais objetivos e reconhecíveis. Dessa forma, o DCV demonstra potencial como recurso auxiliar para comunicações rápidas, discretas e personalizáveis no cotidiano.

Como trabalhos futuros, pretende-se ampliar os testes com um número maior de participantes, incluindo pessoas surdas e surdocegas, a fim de avaliar o uso do dispositivo em situações mais próximas da realidade cotidiana. Também são previstos aprimoramentos no aplicativo e na estrutura física do DCV, buscando melhorar sua usabilidade, resistência, autonomia e possibilidades de personalização. Em etapas posteriores, espera-se disponibilizar o dispositivo ao público, contribuindo para ampliar o acesso a tecnologias assistivas simples, funcionais e socialmente relevantes.

IX. Agradecimentos

A autora agradece à Universidade Tecnológica Federal do Paraná, instituição onde a pesquisa relacionada ao dispositivo de comunicação vibrotátil teve início. Agradece também ao Laboratório de Robótica e Computação Aplicadas do Instituto Federal do Paraná – Campus Pinhais, pela implementação do DCV utilizado neste trabalho. Estende-se o agradecimento aos pesquisadores, colaboradores e integrantes do grupo de pesquisa envolvidos no desenvolvimento do projeto, pelo apoio técnico e científico ao longo da realização deste estudo.

X. Cumprimento dos requisitos éticos

A. *Conflito de interesses*

Os autores declaram não haver conflito de interesses no desenvolvimento desse trabalho.

B. *Declaração de direitos humanos e animais*

Os procedimentos realizados neste trabalho seguiram os padrões éticos do Comitê Ético da UFES e da Declaração de Helsinque de 1975.

References

1. Sorgini Francesca, Cali'o Renato, Carrozza Maria Chiara, Oddo Calogero Maria. Haptic-assistive technologies for audition and vision sensory disabilities *Disability and Rehabilitation: Assistive Technology.* 2017.
2. Muñoz-Baell Irma M., Ruiz Maria Teresa. Empowering the Deaf. Let the Deaf Be Deaf *Journal of Epidemiology and Community Health.* 2000;54:40–44.
3. Dyzel Vernandi, Oostrom-Calo Rony, Worm Mijke, Sterkenburg Paula S.. Assistive Technology to Promote Communication and Social Interaction for People With Deafblindness: A Systematic Review *Frontiers in Education.* 2020;5:578389.
4. Kbar Ghassan, Bhatia Akshay, Abidi Mustufa Haider, Alsharawy Ibraheem. Assistive Technologies for Hearing and Speaking Impaired People: A Survey *Disability and Rehabilitation: Assistive Technology.* 2016.
5. Ramones Alejandro Flores, Rio-Guerra Marta Sylvia. Recent Developments in Haptic Devices Designed for Hearing-Impaired People: A Literature Review *Sensors.* 2023;23:2968.
6. Yaganoglu Mete, Kose Cemal. Real-Time Detection of Important Sounds with a Wearable Vibration Based Device for Hearing-Impaired People *Electronics.* 2018;7:50.
7. Harkins Judith, Tucker Paula E., Williams Norman, Sauro Jeff. Vibration Signaling in Mobile Devices for Emergency Alerting: A Study With Deaf Evaluators *Journal of Deaf Studies and Deaf Education.* 2010;15:327–335.
8. Lucía María J., Revuelta Pablo, García Álvaro, et al. Vibrotactile Captioning of Musical Effects in Audio-Visual Media as an Alternative for Deaf and Hard of Hearing People: An EEG Study *IEEE Access.* 2020;8:190873–190881.
9. Guo Xu, Avila Raudel, Wu Ziqi, et al. Stimulation Modalities in Wearable Haptic Systems: Single-Mode Feedback to Multiphysics Actuation *Research.* 2026;9:1192.
10. Dim Nem Khan, Ren Xiangshi. Investigation of Suitable Body Parts for Wearable Vibration Feedback in Walking Navigation *International Journal of Human-Computer Studies.* 2017;107:1–34.

Author: Letícia Christine Cantieri
Institute: Universidade Tecnológica Federal do Paraná
Street: Av. Sete de Setembro, 3165
City: Curitiba
Country: Brasil
Email: lechris.c@gmail.com

Desenvolvimento de Dispositivos Assistivos de Baixo Custo para Apoio à Escrita de Crianças com Paralisia Cerebral: Relato de Dois Casos

Pedro Victor da Costa Santos*, Gabrielly Blasques Oliveira Albuquerque*,
Pedro Henrique Sousa de Oliveira Xavier‡, Izabella Caroline Pereira da Silva Moura*,
José Lucas da Silveira Oliveira*, Lucas Sousa Linhares‡, Luiz Lampreia Neto*,
Letícia Thaynna Xavier Araujo Oliveira de Azevedo*, Angelo Gabriel Santos Vieira‡,
Andrezza Cristina da Silva Barros Souza‡, Danilo Alves Pinto Nagem*,
Luciana de Andrade Mendes*, Linda Carter Souza da Silva†
*Departamento de Engenharia Biomédica, Centro de Tecnologia, Universidade Federal do Rio Grande do Norte, Natal, Brasil
Email: {pedro.costa.709, izabella.moura.096, gabrielly.blasques.091, lucas.oliveira.016,
luiz.lampreia.016, thaynna.araujo.016}@ufrn.edu.br,
{danilo.nagem, luciana.mendes}@ufrn.br
†Núcleo de Educação da Infância (NEI), Universidade Federal do Rio Grande do Norte, Natal, Brasil
Email: linda.carter@nei.ufrn.br
‡Escola de Ciências e Tecnologia, Universidade Federal do Rio Grande do Norte, Natal, Brasil
Email: {pedro.xavier.713, angelo.santos.124, lucas.linhares.045}@ufrn.edu.br, andrezza.souza@ufrn.br

***Resumo*—Cerebral palsy (CP) impairs motor coordination and can limit children's participation in writing activities, creating barriers to educational inclusion. This paper presents the development of three low-cost assistive writing devices designed to address specific functional demands identified in two children with CP assisted at the Early Childhood Education Center (NEI) of the Federal University of Rio Grande do Norte (UFRN). The first device is a pencil gripper made of ethylene-vinyl acetate (EVA) with an elastic fastening system, aimed at supporting difficulties in gripping conventional writing instruments. The second is a multi-angle tool with three gel crayon tips positioned at approximately 0°, 90°, and 180°, designed to expand the possibilities of writing according to the available upper-limb movement. The third is a Tactile Navigation Module designed to support pre-writing activities and fine motor skills development. The devices were developed through an iterative prototyping process based on observations of the educational context and feedback from a multidisciplinary team comprising an occupational therapist and pedagogues. The experience demonstrates the feasibility of developing low-cost and customizable assistive solutions from specific educational demands, while highlighting the importance of multidisciplinary collaboration in the development of assistive technology for inclusive education.**

***Palavras-chave*—cerebral palsy, assistive technology, adapted device, educational inclusion.**

I. Introdução

A paralisia cerebral (PC) é um grupo de desordens permanentes do desenvolvimento do movimento e da postura, atribuídas a distúrbios não progressivos que ocorrem no cérebro em desenvolvimento fetal ou infantil [1]. Com prevalência estimada entre 1,5 e 3 por 1 000 nascidos vivos [2], a PC manifesta-se de formas variadas, impactando de modo significativo a capacidade motora fina dos indivíduos afetados.

A escrita manual é uma das habilidades motoras mais complexas exigidas no ambiente escolar [3], envolvendo a integração de coordenação motora fina, controle postural, atenção e planejamento motor [4]. Para crianças com PC, essas exigências podem tornar a atividade de escrita inacessível ou extremamente desgastante, comprometendo a participação acadêmica, a expressão individual e, consequentemente, a inclusão social [5].

O desenvolvimento da coordenação motora fina deve ser entendido como uma habilidade prévia essencial para o desenvolvimento da escrita, especialmente em crianças com PC, uma vez que a produção gráfica depende de preensão adequada, controle do punho e dos dedos, velocidade de membro superior e integração visomotora. Estudos mostram que crianças com PC apresentam pior desempenho em padrões de preensão e desenho quando comparadas a crianças com desenvolvimento típico, indicando atraso em marcos de pré-escrita [6]; além disso, déficits de integração visomotora persistem mesmo quando usam a mão menos afetada, o que compromete tarefas básicas de cópia de formas relacionadas à prontidão para escrever [7]. Em amostras escolares com PC, fatores como desvio lateral do punho e velocidade do membro superior estiveram associados ao desempenho da escrita [8], enquanto intervenções focadas em atividades de pré-escrita mostraram efeito positivo sobre motricidade fina e percepção visual [9].

Em conjunto, esses achados reforçam que fortalecer a coordenação motora e a função manual não é apenas um treino complementar, mas uma base necessária para favorecer a aquisição da escrita e a participação escolar de crianças

com paralisia cerebral [10]. Desse modo, o desenvolvimento de tecnologias assistivas (TA) que favoreçam habilidades de pré-escrita torna-se fundamental para promover a autonomia e a participação de pessoas com deficiência [11]. No contexto escolar, dispositivos assistivos bem projetados ampliam as capacidades motoras, promovem autoestima e engajamento no processo educativo [12], [13]. Contudo, grande parte das soluções disponíveis comercialmente apresenta alto custo ou não é adequada ao perfil motor específico de cada criança, especialmente no contexto educacional público brasileiro [14].

Neste cenário, o presente trabalho apresenta o desenvolvimento de três dispositivos assistivos de baixo custo, concebidos a partir de demandas funcionais identificadas no contexto educacional e aprimorados por meio de um processo iterativo de prototipagem. O objetivo é descrever o processo de concepção, desenvolvimento e refinamento de soluções de tecnologia assistiva voltadas ao apoio às atividades de escrita e pré-escrita, destacando a contribuição da atuação multidisciplinar entre estudantes, profissionais da educação e terapia ocupacional.

II. Metodologia

A. Contexto da atividade extensionista e caracterização das demandas

O desenvolvimento dos dispositivos ocorreu no contexto de uma atividade extensionista realizada no Núcleo de Educação da Infância (NEI) da Universidade Federal do Rio Grande do Norte (UFRN), instituição que oferece atendimento educacional com suporte multidisciplinar. A atividade envolveu estudantes da UFRN e profissionais vinculados ao contexto educacional e de terapia ocupacional, tendo como finalidade o desenvolvimento de soluções de tecnologia assistiva de baixo custo a partir de demandas funcionais observadas no ambiente escolar.

As demandas que orientaram o desenvolvimento dos dispositivos foram identificadas a partir de visitas ao contexto educacional, observação das atividades realizadas pelas crianças e discussões com os profissionais envolvidos no atendimento. Essas informações foram utilizadas como subsídios para a definição dos requisitos funcionais dos dispositivos, não tendo como finalidade a produção de medidas clínicas ou a avaliação experimental de eficácia.

O processo de desenvolvimento foi organizado de forma iterativa, contemplando as etapas de identificação da demanda, concepção da solução, prototipagem, apreciação técnica e funcional pelos profissionais envolvidos e refinamento do dispositivo. Dessa forma, o trabalho é apresentado como um relato de experiência extensionista de desenvolvimento tecnológico aplicado à tecnologia assistiva.

B. Dispositivo 1 — Engrossador de Lápis em EVA

1) Perfil: A Criança 1 apresenta comprometimento da coordenação motora fina, com dificuldade na preensão de instrumentos de escrita de espessura convencional. A instabilidade do lápis durante a escrita gera fadiga precoce e frustração.

2) Diagnóstico e concepção: A observação inicial revelou que o diâmetro reduzido do lápis convencional não era adequado à preensão da criança. Foram testados protótipos com diferentes materiais e formas, incluindo peças de brinquedo (tipo esponja), isopor texturizado de ar-condicionado sobre marcadores e diferentes geometrias em EVA (Fig. 1a).

3) Prototipagem iterativa: O primeiro protótipo relevante adotou EVA amarelo com design escalonado: a região central era a mais espessa, com diâmetro diminuindo gradualmente em direção às extremidades, preenchendo a palma da mão e guiando o posicionamento dos dedos. Um elástico branco foi incorporado ao dispositivo para estabilizar a mão durante a escrita. A apreciação realizada com a TO indicou como pontos positivos o formato "gordinho" escalonado, mas apontou que a espessura total estava excessiva para a mão da criança e que o elástico inicial era largo demais.

4) Design final: Com base no feedback, a espessura foi reduzida e padronizada, resultando em engrossador cilíndrico uniforme de EVA macio. O elástico foi substituído por versão mais estreita e ajustada para mãos menores. O instrumento de escrita adotado foi o lápis em gel, por exigir menor pressão sobre o papel [15]. O conjunto final, produzido em cores vibrantes para aumentar o engajamento, foi apreciado pela TO e pelas professoras da turma (Fig. 1b).



(a) Protótipos de texturas. (b) Modelos finais em EVA.

Figura 1: Evolução do engrossador de lápis: (a) protótipos desenvolvidos para explorar diferentes texturas e (b) modelos que atenderam aos critérios estabelecidos para a etapa final de desenvolvimento. Fonte: elaborado pelos autores (2026).

C. Dispositivo 2 — Ferramenta Multiangular de Escrita

1) Perfil: A Criança 2 apresenta limitações significativas na coordenação motora grossa e fina, além de déficits atencionais. O comprometimento resulta em restrição dos graus de liberdade do braço dominante, de modo que os ângulos de movimento disponíveis não possuem ponta ativa na maioria das posições, tornando a execução da escrita praticamente inviável com instrumentos convencionais [16].

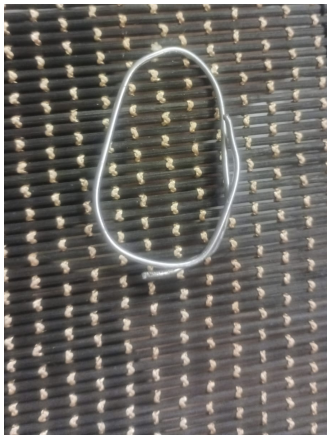
2) Concepção: Para ampliar o conjunto de ângulos de escrita funcionais, concebeu-se um dispositivo que oferece múltiplas pontas em ângulos distintos em uma única ferramenta, buscando aumentar as possibilidades de posicionamento de uma ponta ativa de acordo com o movimento disponível em cada momento.

3) Desenvolvimento do protótipo: O dispositivo foi desenvolvido em etapas, ilustradas na Fig. 2:

1) *Princípio de funcionamento:* A criança segura a base do dispositivo com a mão dominante e posiciona-o de modo que uma das três pontas entre em contato direto com o papel. A presença de pontas em ângulos distintos faz com que, mesmo com amplitudes de movimento reduzidas, haja maior possibilidade de uma ponta adequadamente posicionada para executar o traçado [11].
2) *Arquitetura do dispositivo:*
   - *Estrutura base:* Um fio de arame 14 de 20 cm (Fig. 2a) foi moldado em um formato elíptico, apresentando um eixo maior de 7 cm e eixo menor de 4,5 cm, permitindo o posicionamento versátil de componentes, aliado ao baixo custo.
   - *Revestimento:* A estrutura de arame foi revestida com 3 seções de tubo de espuma de polietileno expandido (EPE) (Fig. 2b), material que cria superfícies ergonômicas adaptadas à preensão da criança, sustentada mecanicamente por fita adesiva.
   - *Pontas de giz em gel:* Três pontas equipadas com giz em gel foram fixadas ao protótipo com massa epóxi (Fig. 2c), posicionadas em ângulos de aproximadamente 0°, 90° e 180°. O giz em gel foi escolhido por ser deslizante.
   - *Fixador:* Foram adicionados 2 segmentos de 10 cm de velcro (Fig. 2d) para fixação do dispositivo na mão da criança.

Além do protótipo artesanal, o dispositivo foi posteriormente projetado em ambiente digital e fabricado por meio de impressão 3D (Fig. 3).

4) Evolução para impressão 3D: Após a apreciação ergonômica e funcional do protótipo artesanal, o projeto evoluiu para a etapa de manufatura aditiva. A geometria da ferramenta foi primeiramente refinada em ambiente virtual (Fig. 3), garantindo o posicionamento angular das pontas. O resultado final, materializado por meio de impressão 3D em filamento PLA (Fig. 4), resultou em um dispositivo mais leve e com acabamento refinado, mantendo as características funcionais definidas durante o processo de desenvolvimento.



(a) Estrutura em arame.



(b) Revestimento em EPE.



(c) Integração das pontas.



(d) Tiras de fixação.

Figura 2: Etapas de montagem do protótipo artesanal: (a) fio de arame 14 estrutural; (b) fixação dos tubos de EPE; (c) integração das pontas em ângulos de 0°, 90° e 180°; e (d) finalização com tiras de fixação. Fonte: elaborado pelos autores (2026).

D. Dispositivo 3 — Módulo de Navegação Tátil

1) Perfil: Ambas as crianças apresentaram demandas atencionais importantes (devido a outros diagnósticos associados à PC como Transtorno do Espectro Autista (TEA) na Criança 2 e Microcefalia na Criança 1), além da necessidade do desenvolvimento da coordenação motora fina, uma habilidade requisitada previamente ao desenvolvimento da escrita.

2) Diagnóstico e concepção: Tendo em vista a demanda atencional de ambas as crianças e a importância do desenvolvimento de habilidades pré-escrita, tais como coordenação motora fina, propriocepção e reconhecimento de padrões, concebeu-se então um dispositivo baseado em uma arquitetura modular que permite a customização da trilha e do nível de desafio.

3) Prototipagem iterativa: Primeiro, foram levadas réguas vazadas, convencionais de mercado, com formas variadas, conforme Fig. 5a, para explorar a possibilidade de interação com as crianças. No entanto, as formas eram pequenas e isso dificultou a interação. Nos encontros com a TO e a rede de cuidado educacional, ao explicar a ideia do Módulo de Navegação Tátil (MNT) personalizado para as necessidades das crianças, foi incentivada a continuação do projeto. Mesmo o ensino do NEI-UFRN aplicando o princípio do desenho à mão livre, percebeu-se a necessidade do MNT para as crianças com PC.

Então foi criado um modelo utilizando uma caixa de papelão de hortifrúti (embalagem 8/10 kg – $48 \times 25 \times 13$ cm). A caixa foi desmontada, os caminhos e formas foram cortados com



Figura 3: Representação virtual do dispositivo modelado em software 3D. Fonte: elaborado pelos autores (2026).



Figura 4: Resultado final do dispositivo impresso em PLA, pronto para uso, produzido pela InPacta (ECT/UFRN).

auxílio de estilete e foram montados botões feitos com papelão com formas (emojis e borboleta). O processo foi feito seguindo o fluxograma gerado por IA, conforme Fig. 5b, e, em seguida, o material foi pintado com tinta guache para proporcionar melhor harmonização visual ao protótipo (Fig. 5c).

4) Design final: O design final do MNT levou em consideração, durante sua estruturação, os parâmetros de tolerância mecânica, os encaixes modulares tipo quebra-cabeça e a disposição dos atuadores nos trilhos e sua integração com as zonas de pré-escrita, podendo ser observado em madeira na Fig. 7.

Após a apreciação do conceito e do componente lúdico por meio do protótipo em papelão, o Módulo de Navegação Tátil (MNT) evoluiu para sua versão definitiva, confeccionada em madeira. A transição de materiais visou conferir maior durabilidade, estabilidade e segurança estrutural durante as sessões de interação com as crianças. O dispositivo finalizado apresenta acabamento lixado e trilhas bem definidas, proporcionando um percurso tátil contínuo, seguro e resistente ao desgaste diário, características compatíveis com a proposta de apoiar atividades de desenvolvimento da coordenação motora fina.



(a) Testes com réguas vazadas.



(b) Fluxograma por IA.



(c) Protótipo funcional em papelão.

Figura 5: Etapas de concepção do Módulo de Navegação Tátil: (a) testes iniciais com réguas vazadas; (b) fluxograma de planejamento gerado por IA; e (c) protótipo funcional estruturado em papelão reaproveitado. Fonte: elaborado pelos autores (2026).

O MNT é composto por painéis modulares com canais de baixa fricção, subdivididos em:

- *Guias de Precisão (Slots):* trilhas onduladas, zigueza-gues e padrões geométricos integrados para atividades de controle motor fino. O perfil dos canais é projetado para minimizar o atrito, exigindo força controlada para o movimento.
- *Trilhas de Interferência:* a geometria das trilhas cria desafios de navegação que estimulam a coordenação óculo-manual e proprioceptiva.

Além disso, foram definidos os seguintes componentes:

1) *Atuadores (Sliders):* botões com formas geométricas e peso controlado, impressos em PLA via impressora 3D.
2) *Pino de Guia e Retenção:* mecanismo interno de pino que mantém o atuador dentro do trilho, permitindo apenas movimentos controlados ao longo da trilha.
3) *Recortes Geométricos de Precisão:* espaços com formas (coração, triângulo, quadrado) e linhas de pré-escrita.

4) *Guias de Alinhamento de Papel:* recuo integrado para o posicionamento preciso de folhas de papel, permitindo que a criança use a caneta dentro das formas guias.
5) *Pinças:* as pinças em madeira permitem o acoplamento do MNT em planos inclinados.
6) *Peças para encaixe:* peças do mesmo tamanho dos recortes geométricos permitem que a criança reconheça padrões geométricos e os encaixe de forma correspondente.



Figura 6: Módulo de Navegação Tátil em representação 3D.

III. Resultados

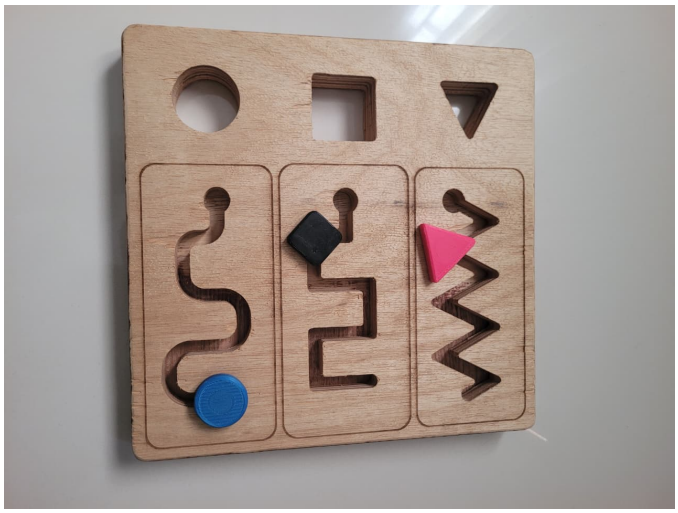
A. Engrossador de EVA

Após três ciclos de prototipagem e avaliação, foi definida a configuração final do engrossador, considerando as características funcionais observadas e as contribuições da terapeuta ocupacional e das professoras da turma. A Tabela I sumariza as especificações do produto final.

Os principais pontos de melhoria identificados ao longo da iteração foram: (i) redução da espessura para adequação à mão da criança; (ii) substituição do elástico inicial por versão mais estreita; e (iii) adoção do lápis em gel em substituição ao lápis grafite convencional. Os próximos passos incluem a fixação permanente do elástico com cola quente — dado que o dispositivo se soltou durante uso mais intenso — e a expansão da paleta de cores disponíveis [13].

B. Ferramenta Multiangular

As apreciações exploratórias realizadas durante o desenvolvimento indicaram que a configuração com múltiplas pontas ampliava as possibilidades de posicionamento da extremidade ativa diante das restrições de movimento observadas. Em experimentações iniciais de uso, foi possível observar a realização de marcações no papel com diferentes posicionamentos do dispositivo. Esses registros tiveram caráter exclusivamente exploratório e orientaram o refinamento do protótipo, não constituindo avaliação quantitativa de desempenho ou estudo de eficácia.



(a) Vista superior da versão em madeira.



(b) Vista frontal da versão em madeira.

Figura 7: Resultado final do Módulo de Navegação Tátil manufaturado em madeira, projetado para suportar o uso contínuo e garantir segurança durante o desenvolvimento de habilidades pré-escritas. Confeccionado pelo LabMAN-PotiProt/UFRN.

Os principais desafios operacionais documentados foram: (i) dificuldades no agendamento de visitas técnicas em função das necessidades específicas da criança; (ii) escassez de profissionais treinados para manusear protótipos adaptados em contexto educacional [17]; e (iii) aceitação sensorial do dispositivo por parte da criança. Esses fatores limitaram a extensão das experimentações e a obtenção de dados quantitativos.

C. Módulo de Navegação Tátil

O MNT é uma solução de engenharia concebida para apoiar atividades de pré-escrita, integrando princípios de controle motor e processamento sensorial. Sua modularidade, o feedback tátil e o design lúdico conferem ao dispositivo características que podem favorecer sua adaptação a diferentes atividades voltadas ao desenvolvimento de habilidades de pré-escrita e coordenação motora fina.

A Tabela II apresenta os principais componentes e respectivas funções do MNT.

Tabela I: Especificações do Engrossador de EVA (Dispositivo 1)

Componente	**Especificação**
Instrumento de escrita	Lápis em gel
Material do engrossador	EVA macio
Geometria	Ciliíndrica uniforme
Fixação	Elástico estreito ajustável
Cores disponíveis	Múltiplas (vibrantes)
Custo estimado	$<$ R$ 5,00 por unidade

Fonte: elaborado pelos autores (2026).

Tabela II: Componentes e funções do Módulo de Navegação Tátil

Componente	**Função**	**Material recomendado**
Módulo Base Modular	Estrutura de suporte principal e guias	Madeira Laminada Reforçada
Atuador (*Slider*)	Interface de interação e movimento controlado	Madeira
Pino de Guia Interno	Mecanismo de retenção e baixa fricção	Metal de baixa fricção / polímero especial
Zona de Pré-Escrita	Prática de controle motor e formas	Madeira laminada com recortes
Guias de Alinhamento de Papel	Posicionamento preciso das folhas	Integrados à estrutura

Fonte: elaborado pelos autores (2026).

IV. Conclusões

Este trabalho apresentou o desenvolvimento de três dispositivos assistivos de escrita e pré-escrita de baixo custo, concebidos a partir de demandas funcionais específicas identificadas no contexto educacional. O processo de desenvolvimento evidenciou a importância da prototipagem iterativa e da colaboração entre estudantes, profissionais da educação e terapeuta ocupacional para a definição e o refinamento de soluções de tecnologia assistiva.

As duas crianças consideradas no desenvolvimento das TA apresentaram demandas motoras distintas, reforçando a necessidade de soluções individualizadas. Enquanto o primeiro dispositivo foi direcionado à adaptação da preensão do instrumento de escrita, o segundo buscou ampliar as possibilidades de posicionamento da extremidade ativa diante de restrições de movimento do membro superior. O terceiro dispositivo foi concebido para apoiar atividades de pré-escrita e coordenação motora fina, incorporando elementos táteis, geométricos e lúdicos.

Os resultados apresentados devem ser compreendidos como evidências preliminares do processo de desenvolvimento e da adequação das soluções às demandas. Utilizamos metodologia ativa e design centrado no usuário e no contexto envolvendo os diversos profissionais envolvidos no contexto educacional das crianças Portanto, este estudo demonstrou como foi o processo de desenvolvimento das soluções não tendo como finalidade demonstrar eficácia terapêutica ou educacional nem permitir generalizações para a população de crianças com paralisia cerebral. Estudos futuros são necessários para validar os dispositivos envolvendo uma amostra maior e com crianças com diferentes níveis de comprometimento motor, envolvendo medidas objetivas da função motora, da coordenação e precisão de movimentos por meio de estudo longitudinal.

A experiência também evidencia a extensão universitária como espaço de articulação entre formação acadêmica, demandas concretas do contexto educacional e desenvolvimento de soluções tecnológicas socialmente orientadas.

Como continuidade, estudos futuros poderão incorporar protocolos específicos de avaliação, instrumentos padronizados e métricas quantitativas, como tempo de execução, precisão dos traçados, estabilidade da preensão, amplitude funcional do movimento e fadiga. Também poderão ser realizadas avaliações longitudinais, comparações sistemáticas do desempenho com e sem os dispositivos e investigações envolvendo maior número de crianças e diferentes perfis de comprometimento motor. Essas etapas poderão contribuir para avaliar de forma sistemática a efetividade, a adaptabilidade e a possibilidade de replicação das soluções em diferentes contextos educacionais.

Agradecimentos

Os autores agradecem à equipe do Núcleo de Educação da Infância (NEI-UFRN), em especial às pedagogas e à terapeuta ocupacional, pelo suporte durante as visitas técnicas e pela participação no processo de desenvolvimento e avaliação dos dispositivos. Agradecem também à Incubadora de Processos Acadêmicos, Científicos, Tecnológicos e Aplicados da Escola de Ciência e Tecnologia da UFRN (InPacta/UFRN) pelos dispositivos impressos em PLA. Agradecem também ao Laboratório de Manufatura da Rede PotiProt da UFRN (LabMAN-PotiProt/UFRN) pela confecção do módulo de navegação tátil em madeira.

Nota sobre o uso de Inteligência Artificial

Durante a elaboração deste texto, ferramentas de Inteligência Artificial foram utilizadas como apoio à revisão linguística, à organização das ideias, à reestruturação de trechos e ao aprimoramento da clareza textual. O uso dessas ferramentas ocorreu de forma auxiliar, sem substituir a análise crítica,

Tabela III: Comparação entre os Dispositivos Desenvolvidos

Característica	Dispositivo 1	Dispositivo 2	Dispositivo 3
Problema-alvo	Preensão inadequada	Ângulos limitados	Dificuldade de atenção e coordenação motora fina
Materiais principais	EVA, elástico, lápis gel	Arame 14, EPE, giz em gel, massa epóxi	Madeira
N.º de iterações	3	1	2
Custo estimado	$<$ R$ 5	$<$ R$ 15	$<$ R$ 20
Avaliação	TO + professoras	TO + pedagoga	TO + pedagoga
Status do desenvolvimento	Produção	Implementação	Produção

Fonte: elaborado pelos autores (2026).

a autoria intelectual e a responsabilidade dos autores pelo conteúdo apresentado. Todas as informações, interpretações e decisões finais foram revisadas e ajustadas pelos autores, garantindo a coerência acadêmica, a adequação ao objetivo do trabalho e a preservação da originalidade do relato.

Referências

[1] P. Rosenbaum, N. Paneth, A. Leviton, M. Goldstein, M. Keith e M. Bax, "A report: the definition and classification of cerebral palsy April 2006," *Developmental Medicine & Child Neurology*, v. 49, n. s109, pp. 8–14, 2007. DOI: 10.1111/j.1469-8749.2007.tb12610.x.

[2] D. S. Reddihough e K. L. Collins, "The epidemiology and causes of cerebral palsy," *Australian Journal of Physiotherapy*, v. 49, n. 1, pp. 7–12, 2003. DOI: 10.1016/S0004-9514(14)60183-5.

[3] V. W. Berninger e W. D. Winn, "Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution," *Handbook of Writing Research*, pp. 96–114, 2012.

[4] K. E. Adolph e S. R. Robinson, "Motor development," *Handbook of Child Psychology and Developmental Science*, v. 2, 2010.

[5] Brasil, "Lei Brasileira de Inclusão da Pessoa com Deficiência (Lei n.º 13.146/2015)," *Diário Oficial da União*, 2015, Brasília, DF.

[6] N. Sharma, "Comparing the pre-writing skills of diplegic cerebral palsy children to those of normal children," *Cureus*, v. 16, n. 5, e61352, mai. de 2024. DOI: 10.7759/cureus.61352.

[7] W.-F. Huang, R.-Y. Chen, T.-N. Wang, P.-Y. Chuang, J.-Y. Shieh e H.-L. Chen, "Visual-motor integration in children with unilateral cerebral palsy, Application of the computer-aided measure of visual-motor integration," *Journal of NeuroEngineering and Rehabilitation*, v. 21, n. 37, mar. de 2024. DOI: 10.1186/s12984-024-01335-8.

[8] H. Kim, "An investigation of the factors affecting handwriting articulation of school aged children with cerebral palsy based on the international classification of functioning, disability and health," *Journal of Physical Therapy Science*, v. 28, n. 2, pp. 347–350, 2016. DOI: 10.1589/jpts.28.347.

[9] J.-W. Kim e S.-N. Ahn, "The effects of a pre-writing activity program on fine motor and visual perception in children with cerebral palsy, A pilot study," *Journal of Korean Society of Neurocognitive Rehabilitation*, v. 16, n. 1, pp. 1–11, 2024.

[10] J. V. McCall, M. C. Ludovice, C. Elliott e D. G. Kamper, "Hand function development of children with hemiplegic cerebral palsy, A scoping review," *Journal of Pediatric Rehabilitation Medicine*, v. 15, n. 1, pp. 211–228, 2022. DOI: 10.3233/PRM-200714.

[11] M. J. Scherer, "Connecting to Learn: Educational and Assistive Technology for People with Disabilities," *American Psychological Association*, 2004, Washington, DC.

[12] K. J. Levine, C. Hurtig e C. Friedman, "Assistive technology: overview and implications for educational practice," *Communication Disorders Quarterly*, v. 29, n. 3, pp. 142–151, 2008. DOI: 10.1177/1525740108315887.

[13] E. G. B. Galvão e C. F. Galvão, "Tecnologia assistiva para crianças com necessidades educacionais especiais na escola," *Universidade Federal de São Carlos: Núcleo de Pesquisa em Tecnologia Assistiva*, 2009.

[14] A. Mello e S. Santos, "Órteses e dispositivos assistivos para a escrita em crianças com paralisia cerebral: revisão integrativa," *Cadernos Brasileiros de Terapia Ocupacional*, v. 30, e3150, 2022. DOI: 10.1590/2526-8910.ctoAR2260.

[15] C. Missiuna, A. Mandich, H. Polatajko e T. Malloy-Miller, "Cognitive orientation to daily occupational performance (CO-OP): part I—theoretical foundations," *Physical & Occupational Therapy in Pediatrics*, v. 20, n. 2-3, pp. 69–81, 2001. DOI: 10.1080/J006v20n02_05.

[16] J. P. Piek e R. J. Dyck, "Sensory-motor deficits in children with developmental coordination disorder, attention deficit hyperactivity disorder and autistic disorder," *Human Movement Science*, v. 23, n. 3-4, pp. 475–488, 2004. DOI: 10.1016/j.humov.2004.08.019.

[17] D. W. Smits, M. Ketelaar, A. Gorter, B. Ketelaars, K. Lindeman e J. W. Gorter, "Involvement of people with disabilities in the development of assistive technology: a systematic review," *Disability and Rehabilitation: Assistive Technology*, v. 6, n. 5, pp. 359–375, 2011. DOI: 10.3109/17483107.2010.541530.

Desenvolvimento de palmilha personalizada por impressão 3D para uma criança com Distrofia Muscular de Duchenne: um estudo de caso centrado no usuário

1st Sofia Maria Mecenas Areias Lima
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
sofia.areias@discente.ufg.br

2nd Laura Sousa Dantas
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
lauradantas@discente.ufg.br

3rd Vitória Souza de Oliveira
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
oliveira_oliveira@discente.ufg.br

4th Ana Luiza Riggheto
Centro Estadual de Reabilitação e Readaptação Dr. Henrique Santillo
Goiânia, Brasil
analuiza.rig@gmail.com

5th Pedro Henrique Gonçalves
Faculdade de Artes Visuais
Universidade Federal de Goiás
Goiânia, Brasil
pedrogoncalves@ufg.br

Resumo— ***A Distrofia Muscular de Duchenne é caracterizada por degeneração muscular progressiva que compromete a função motora e a qualidade de vida de pacientes do sexo masculino. As opções terapêuticas disponíveis não alteram o desfecho da doença, mas contribuem para retardar seu avanço e manter a funcionalidade por mais tempo, sendo os dispositivos ortopédicos personalizados ferramentas importantes nesse processo. O estudo descreveu o processo de desenvolvimento de uma palmilha por impressão 3D para uma criança com Distrofia Muscular de Duchenne, que apresentava fraqueza de musculatura dorsiflexora de tornozelo e desabamento do arco plantar e a satisfação do dispositivo pela família. Foi realizada a confecção de uma palmilha personalizada, cujo processo incluiu escaneamento tridimensional do pé, modelagem em software DEC/MAC, impressão 3D e revestimento interno. O dispositivo inicial apresentou baixa adesão após a primeira semana devido a preferência da criança por sandálias, sendo confeccionado um novo modelo fixado diretamente ao calçado. A satisfação foi avaliada pelo Questionário Quebec de Satisfação com a Tecnologia Assistiva em três momentos distintos, indicando satisfação consistente ao longo do seguimento. Os resultados deste estudo de caso demonstram a viabilidade técnica do desenvolvimento interativo de uma palmilha personalizada por impressão 3D, evidenciando a importância da adaptação do dispositivo às necessidades clínicas e cotidianas do usuário.***

Palavras-chave — ***Distrofia Muscular de Duchenne, Tecnologia Assistiva, Órtese Plantar***

I. INTRODUÇÃO

A Distrofia Muscular de Duchenne (DMD) é uma doença genética recessiva ligada ao cromossomo X, pertencente ao grupo das distrofinopatias, caracterizada por degeneração muscular progressiva que compromete significativamente a função motora e a qualidade de vida dos pacientes do sexo masculino [1,2]. Atualmente, as opções terapêuticas disponíveis, como fisioterapia e corticosteróides, não alteram o desfecho da doença, mas são fatores que contribuem para retardar seu avanço e manter a funcionalidade por mais tempo [3].

Na perspectiva da biomecânica funcional, o organismo de crianças com DMD reorganiza-se motoramente para manter a deambulação devido ao padrão de marcha significativamente alterado em decorrência da fraqueza muscular progressiva [4, 5]. Entre essas alterações biomecânicas, observam-se inclinação pélvica anterior excessiva, abdução aumentada na fase de balanço e comprometimento da flexão do joelho na resposta à carga [6,7]. Em muitos casos, desenvolve-se hiperextensão do joelho, como estratégia compensatória para a fraqueza do quadríceps [8,6,7]. Os pés permanecem em equino no contato inicial, com flexão plantar excessiva durante o balanço, enquanto o comprimento da passada é reduzido e a largura do passo é aumentada para preservar o equilíbrio [8].

Dessa forma, dispositivos ortopédicos constituem parte do manejo clínico da DMD, e a prescrição precoce pode prolongar a deambulação e retardar o declínio funcional, reforçando a importância dessa intervenção [9]. As órteses de tornozelo e pé (OTPs) são comumente prescritas nesse contexto por promoverem o alinhamento dos segmentos corporais, controlarem a instabilidade articular e prevenirem deformidades musculoesqueléticas, otimizando o desempenho biomecânico durante a marcha [10,11]

À medida que a força muscular diminui, no entanto, a eficácia dessas órteses tende a reduzir-se progressivamente. Por isso, as prescrições devem ser individualizadas e adaptadas ao nível de comprometimento dos membros inferiores de cada

criança, funcionando como ferramentas auxiliares na promoção das atividades de vida diária e na facilitação da marcha [12,9,13]

Nesse cenário, a tecnologia de impressão 3D tem ampliado as possibilidades de produção personalizada de dispositivos ortopédicos. Os softwares de modelagem aliados ao tempo de produção permitem adequações anatômicas precisas às necessidades individuais de cada paciente quando comparado ao método tradicional [14,15]. A prototipagem rápida representa, assim, um avanço relevante para o desenvolvimento de soluções terapêuticas mais eficazes e acessíveis para crianças com DMD.

Diante disso, o presente estudo tem como objetivo descrever o processo de desenvolvimento de uma palmilha personalizada produzida por impressão 3D para uma criança com Distrofia Muscular de Duchenne, bem como analisar o processo de reprojeto do dispositivo a partir do acompanhamento longitudinal da satisfação e da adesão da família.

II. METODOLOGIA

O estudo foi desenvolvido nas dependências do Centro Estadual de Reabilitação e Readaptação Dr. Henrique Santillo, um Centro Especializado de Reabilitação (CERs) nível IV localizado em Goiânia, Goiás. Os dispositivos de TA são desenvolvidos pelo Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. E.I.T.A.), da Universidade Federal de Goiás, por uma equipe multidisciplinar e são entregues aos pacientes da instituição, conforme o parecer do Comitê de Ética 7.828.871, durante os meses de março a junho de 2026.

O fluxo de solicitação do dispositivo iniciou-se por meio da demanda identificada pelo profissional do serviço responsável pelo atendimento do paciente. A responsável pela criança assinou o Termo de Consentimento Livre e Esclarecido (TCLE) e participou de uma entrevista de anamnese pré-estruturada contendo perguntas sobre diagnóstico clínico da criança, uso de órteses, acesso à saúde, entre outras e, em seguida, foi realizada avaliação fisioterapêutica.

O recurso de tecnologia assistiva foi definido após a avaliação fisioterapêutica pela equipe multiprofissional, composta por fisioterapeutas, designer de produtos e arquiteto. Para isso, foi realizado o escaneamento tridimensional com a criança posicionada sentada em um banco, com a articulação do tornozelo posicionada passivamente a 90 graus.

As medidas anatômicas obtidas no escaneamento e a identificação das necessidades clínicas na avaliação fisioterapêutica auxiliaram a modelagem digital do dispositivo e a definição da geometria do dispositivo. Assim, foi desenvolvida uma palmilha com objetivos de promover suporte ao arco medial longitudinal plantar e melhorar a acomodação do pé durante o uso de calçado. A palmilha foi confeccionada por impressão 3D, testada e entregue ao paciente. Os parâmetros da impressão tridimensional estão apresentados na tabela I.

TABELA I. **PARÂMETROS DA IMPRESSÃO 3D**

Parâmetros	Palmilha
Material	TPE
Preenchimento	Variável
Espessura da Camada	0.2mm
Temperatura do Bico	220 ºC
Temperatura da Mesa Impressão	60 ºC
Diâmetro do Bico	0.4mm
Padrão de preenchimento	Variável
Tempo total impressão	180 min.

Após a entrega da palmilha, foi aplicado o Questionário de Satisfação do Usuário com a Tecnologia Assistiva de Quebec (QUEBEC/B-QUEST) em três momentos: 7 dias, 30 dias e 90 dias. O instrumento foi criado em 2010 por Demers, Weiss-Lambrou, Ska [16] com o objetivo de quantificar a satisfação dos usuários com a tecnologia utilizada, traduzido e validado para a língua portuguesa em 2014 por Carvalho, Júnior, Sá [17], contendo 12 perguntas: oito relacionadas ao recurso de tecnologia assistiva (dimensões, peso, ajustes, segurança, durabilidade, facilidade de uso, conforto e eficácia) e quatro relacionadas aos serviços prestados (processo de entrega, reparos, serviços profissionais e serviços de acompanhamento), e uma questão sobre quais os itens o usuário considera mais importantes em um recurso de TA. Cada item do instrumento é pontuado em uma escala de 1 a 5, sendo: 1 (insatisfeito), 2 (pouco satisfeito), 3 (mais ou menos satisfeito), 4 (bastante satisfeito) e, 5 (totalmente satisfeito).

No acompanhamento do uso da palmilha pelo paciente foi identificada a necessidade de ajustes. Assim, um novo modelo de palmilha foi desenvolvido e reimpresso. Os parâmetros obtidos no escaneamento e na avaliação fisioterapêutica foram mantidos para a confecção da nova palmilha.

O fluxograma de todas as etapas está apresentado na Fig. 1. O processo de desenvolvimento foi baseado no fluxo de prototipagem rápida previamente estabelecido pelo Lab.EITA.

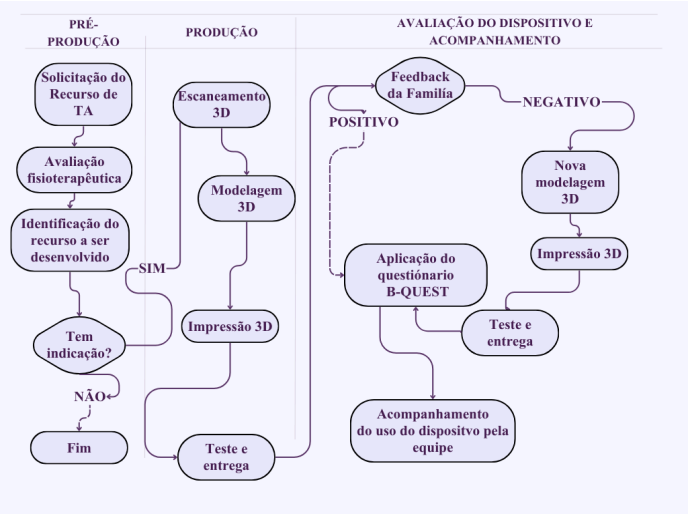


Fig. 1. Fluxograma do processo de produção da palmilha adaptada

III. RESULTADOS

Paciente do sexo masculino, 7 anos de idade, com diagnósticos de Distrofia Muscular de Duchenne (DMD), Transtorno do Espectro Autista (TEA) e Transtorno do Déficit de Atenção e Hiperatividade (TDAH). No exame físico realizado pela equipe de fisioterapia, foi identificada fraqueza da musculatura dorsiflexora de tornozelo, associada a desabamento do arco plantar, pés pronados, com tônus e

trofismo musculares preservados, apresentando indicação clínica para utilização de palmilha. Em sedestação, o paciente apresentou boa distribuição de peso em região pélvica e descarga de peso bem distribuída em membros inferiores, sendo capaz de realizar mudanças posturais. O paciente fez uso prévio de órtese tornozelo-pé (OTP), porém não apresentou adaptação ao dispositivo. Diante desse contexto, a equipe optou pela confecção de uma palmilha para fins de posicionamento.

Inicialmente, foi realizado o escaneamento tridimensional do pé do paciente, com o scanner RevoPoint scanner 3D Range 2 (Fig. 2a). Posteriormente, utilizando a malha gerada pelo escaneamento, a equipe de design realizou a modelagem 3D utilizando o software gratuito Gensole (Fig. 2b). Nesse aplicativo, o modelo digital obtido, por meio do Design Assistido por Computador (DEC), constituiu a base para as fases subsequentes do fluxo de trabalho de fabricação. O arquivo digital projetado no ambiente DEC foi posteriormente traduzido em instruções para os equipamentos de fabricação, por meio da Manufatura Assistida por Computador (MAC). Esse fluxo de desenvolvimento refletiu na transição entre a modelagem tridimensional da palmilha e a posterior fabricação por impressão 3D, conforme detalhado a seguir.

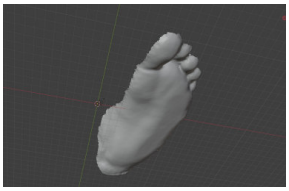
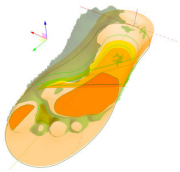
(a)   (b)

Fig. 2. (a). Malha gerada pelo escaneamento. Fig. (b) modelagem da primeira tentativa.

Em seguida, o modelo foi exportado em estereolitografia (STL) e impresso em 3D, utilizando a impressora Creality S1 PRO, com um bico de 0,4mm, filamento TPE (Shore 83-A), levando 1 hora e 57 minutos. Durante o pós-processamento, foi realizada a retirada dos suportes de impressão permitindo o ajuste das bordas para a fixação. Em seguida, a parte interna foi revestida com Plastazote e colada com cola adesiva de contato. Os gastos e quantidades dos materiais foram descritos na Tabela II.

TABELA II. CUSTOS DE PRODUÇÃO DO PAR DE PALMILHAS EM IMPRESSÃO 3D

Variável	Quantidade	Custo Estimado
TPE	50g	R$ 18,00
Plastazote	8g	R$19,00
Custo Total		R$37

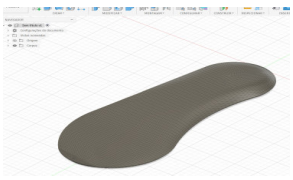
Ressalta-se que esse valor corresponde exclusivamente aos materiais consumidos durante a fabricação, não contemplando custos relacionados à mão de obra especializada, escaneamento tridimensional, modelagem digital, energia elétrica, depreciação dos equipamentos e acabamento.

Ao final desse processo,a palmilha foi testada com a criança e entregue para uso (Fig. 3), sendo estruturada para ser posicionada dentro do tênis. Ademais, a responsável foi orientada a iniciar o uso gradual do produto, em ambiente escolar e na comunidade.



Fig. 3. Teste e entrega do primeiro modelo

Entretanto, após uma semana, a mãe relatou dificuldade no uso deste primeiro dispositivo, pois a criança não aderiu ao uso com o tênis, demonstrando preferência pelo uso de sandálias. Diante disso, a equipe optou pela confecção de um segundo modelo de palmilha, projetado para ser fixado diretamente na sandália (Fig. 4a), utilizando o software Fusion 360 (Fig. 4b).

(a)   (b)

(Fig. 4. (a) Sandália para fixação. Fig. (b).Segundo modelo de palmilhas

A baixa adesão ao primeiro modelo não foi interpretada como falha, mas como evidência da necessidade de adaptação do dispositivo às características funcionais e comportamentais da criança. Dessa forma, o feedback da família foi incorporado ao processo de desenvolvimento, culminando no reprojeto completo do dispositivo, resultando em um segundo modelo de palmilha.

O segundo modelo da palmilha foi testado presencialmente durante o atendimento fisioterapêutico e entregue à criança (Fig. 5).

Fig. 5. Segundo modelo da palmilha em uso

O acompanhamento subsequente à entrega do segundo modelo da palmilha foi realizado de forma virtual, durante um período de três meses. O B-QUEST foi reaplicado 30 e 90 dias após a entrega do segundo modelo de palmilha. Os resultados do questionário podem ser visualizados na Tabela III.

TABELA III. PONTUAÇÃO SUBTOTAL DO B-QUEST

Momento de avaliação	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*TOTAL*
7 dias após 1ª entrega	4	4	4	4	4	4	3	4	4	4	4	4	3,91
B-QUEST 1*	4	4	-***	3	4	3	3	3	4	4	4	4	3,63

B-QUEST 2 **	3	4	3	4	4	3	3	4	5	4	5	4	3,83

*Aplicado após 30 dias de entrega

** Aplicado após 90 dias de entrega

*** Questão não respondida, foi retirada do cálculo do escore total.

Na pontuação dos itens sobre Recursos (itens 1-8), o escore do B-QUEST 1 foi 3,57 e B-QUEST 2 foi 3,5, resultando em satisfação mais ou menos. A pontuação dos itens sobre Serviços (itens 9-12) foi 4,0 e 4,5 respectivamente (bastante satisfeito).

IV. DISCUSSÃO

A confecção de palmilhas adaptadas por impressão 3D apresentou-se como uma alternativa tecnicamente viável e passível de adaptação às necessidades individuais do paciente, levando em consideração as comorbidades do neurodesenvolvimento que, possivelmente, interferiram na adesão ao dispositivo.

Um dos principais achados deste estudo foi demonstrar que o sucesso da tecnologia assistiva esteve diretamente relacionado à capacidade de adaptação do dispositivo às necessidades cotidianas da criança, e não apenas às suas características biomecânicas. A reformulação do projeto a partir do feedback da família evidenciou a importância de processos de desenvolvimento centrados no usuário, nos quais a não adesão ao uso do primeiro modelo da palmilha pode representar uma oportunidade de aperfeiçoamento do dispositivo, e não necessariamente um insucesso terapêutico.

Crianças com TEA podem apresentar dificuldades de adesão a dispositivos ortopédicos em decorrência da rigidez e textura dos materiais e das limitações impostas à mobilidade residual resultados da integração sensorial que gera uma intolerância a esses estímulos táteis e proprioceptivos provocados pelo uso prolongado de dispositivos ortopédicos [18,19], como relatado pela responsável. Em razão disso, houve uma dificuldade do cuidador em incentivar/insistir no uso da palmilha. Portanto, essa interação entre DMD, TEA e uso de tecnologias assistivas representa, por si só, uma lacuna importante a ser investigada em estudos futuros.

Os ajustes realizados no desenvolvimento da palmilha após a baixa adesão na primeira semana, levando em consideração a escuta ativa da responsável pela criança, foi um fator determinante para que a equipe redirecionasse a conduta de forma ágil, preservando a continuidade do uso do dispositivo e, consequentemente, os potenciais benefícios para o posicionamento do arco plantar.

Sobre a utilização da impressão 3D para a confecção da palmilha adaptada, estudos prévios sugerem vantagens de custo, tempo e ambientais em relação aos métodos tradicionais de fabricação [20,21]. A modelagem em software DEC/MAC permite a customização regional de durezas, já a impressão 3D permite a escolha do padrão de preenchimento e geometrias com precisão milimétrica, levando em consideração o escaneamento do pé da criança. Assim, possibilitando o desenvolvimento de geometrias capazes de direcionar o suporte plantar de forma individualizada, embora esse efeito biomecânico não tenha sido mensurado no presente estudo.

As palmilhas convencionais, por sua vez, apresentam limitações na capacidade de personalização, dependendo da experiência manual do ortesista e das propriedades físicas dos materiais disponíveis, devido a pré-moldagem e termoformagem do gesso e termoplástico, respectivamente [20,21]. A reprodutibilidade do arquivo digital é outro aspecto relevante, pois permite que o dispositivo seja reimpresso com exatidão em caso de desgaste ou necessidade de ajuste, o que representa uma vantagem prática em serviços de reabilitação de longa permanência, como o contexto do presente estudo e levando em consideração o crescimento da criança.

A escolha do filamento TPE deu-se a partir de sua combinação de plástico e borracha, que possui alta absorção de impacto, resistência ao desgaste e capacidade de dobrar ou esticar sem quebrar. Características essas, necessárias para as palmilhas quando comparada com as palmilhas já existentes. O filamento flex demonstra-se eficaz na flexibilidade, força máxima e dureza estrutural adequada para palmilhas ortopédicas [22].

Os escores do B-QUEST demonstram percepção favorável da família quanto ao dispositivo e ao processo de acompanhamento oferecido pela equipe. Entretanto, esses resultados devem ser interpretados como indicadores de aceitabilidade da tecnologia assistiva, não constituindo evidências diretas de sua eficácia clínica ou biomecânica. A avaliação do domínio Satisfação demonstrou escores totais consistentemente acima de 3,6 nos três momentos de acompanhamento (3,91; 3,63; 3,83), indicando boa satisfação geral da família com o dispositivo ao longo do seguimento.

Um fator que comumente está atrelado a descontinuação do uso de TA é a falta de acompanhamento e reavaliação do dispositivo após a implementação, tanto pela escassez de pessoas para avaliar o uso, bem como realizar a manutenção dos dispositivos [23,24]. Desta forma, o acompanhamento sistemático realizado em três momentos pode ter contribuído para a consistência dos escores da subescala de Serviços ao longo do seguimento (4,0 para 4,5). A disponibilidade da equipe para reavaliar o dispositivo e propor um novo modelo diante da baixa adesão inicial reforça que o processo de prescrição de TA não se encerra na entrega do dispositivo, mas exige monitoramento contínuo e responsividade às demandas do usuário e de sua família. Nesse contexto, a impressão 3D mostrou-se vantajosa por permitir o reprojeto e a fabricação de um novo dispositivo em um curto intervalo de tempo, favorecendo um processo iterativo de desenvolvimento centrado no usuário, sem a necessidade de reiniciar todas as etapas convencionais de confecção.

A subescala de Recursos, por sua vez, apresentou discreta redução (3,8 para 3,5), o que pode refletir aspectos do dispositivo ainda passíveis de refinamento, como conforto, facilidade de uso ou durabilidade. A mudança no perfil de itens com menor pontuação entre a primeira e as demais avaliações, itens 4, 8 e 11 para itens 5, 6 e 7, pode indicar que a substituição do primeiro modelo para o segundo modelo respondeu a parte das insatisfações anteriores pelos usuários, o que deve orientar eventuais ajustes no dispositivo durante o seguimento.

O presente estudo apresenta limitações inerentes ao seu delineamento. Por tratar-se de um estudo de caso único, os resultados não permitem generalizações para outras populações com DMD ou perfis clínicos distintos. Também não foram realizadas avaliações biomecânicas instrumentadas, como análise cinemática da marcha, distribuição de pressão plantar ou outras medidas objetivas capazes de quantificar os efeitos do dispositivo sobre a função motora. Assim, os achados restringem-se à descrição

do processo de desenvolvimento e à percepção de satisfação da família, não sendo possível estabelecer inferências sobre a eficácia clínica da palmilha. Estudos futuros com períodos de seguimento mais prolongados e sistematizados poderão proporcionar uma compreensão mais abrangente dos efeitos do dispositivo em longo prazo. Da mesma forma, a inclusão de amostras maiores e grupos de comparação contribuirá para fortalecer as evidências acerca do uso da impressão 3D na confecção de palmilhas personalizadas para crianças com distrofinopatias.

V. CONCLUSÃO

O presente estudo demonstrou a viabilidade do desenvolvimento de uma palmilha personalizada por impressão 3D em um contexto clínico de reabilitação. O processo, desde o escaneamento tridimensional até a entrega, acompanhamento e reprojeto do dispositivo, mostrou-se reprodutível e adaptável às necessidades do usuário, sendo necessária a confecção de um segundo modelo para atender às demandas cotidianas e comportamentais da criança. A escuta ativa da família e a responsividade da equipe frente à baixa adesão inicial evidenciaram que a prescrição de tecnologias assistivas constitui um processo dinâmico, que demanda monitoramento contínuo e adaptações ao longo do uso. Os achados reforçam o potencial da manufatura aditiva como ferramenta para o desenvolvimento de recursos personalizados de tecnologia assistiva. Entretanto, estudos com amostras maiores, acompanhamento em longo prazo e avaliações funcionais objetivas são necessários para ampliar a generalização dos resultados e determinar os efeitos clínicos do dispositivo.

REFERENCES

[1] S. Crisafulli, J. Sultana, A. Fontana, F. Salvo, S. Messina, and G. Trifirò, "Global epidemiology of Duchenne muscular dystrophy: An updated systematic review and meta-analysis," *Orphanet J. Rare Dis.*, vol. 15, no. 1, p. 141, Jun. 2020, doi: 10.1186/s13023-020-01430-8.

[2] J. K. Mah *et al.*, "Efficacy and safety of vamorolone in Duchenne muscular dystrophy: A 30-month nonrandomized controlled open-label extension," *JAMA Netw. Open*, vol. 5, no. 1, p. e2144178, Jan. 2022, doi: 10.1001/jamanetworkopen.2021.44178.

[3] S. D. Apkon *et al.*, "Orthopedic and surgical management of the patient with Duchenne muscular dystrophy," *Pediatrics*, vol. 142, no. Suppl. 2, pp. S82–S89, 2018, doi: 10.1542/peds.2018-0333J.

[4] J. P. Bakker, I. J. De Groot, A. Beelen, and G. J. Lankhorst, "Predictive factors of cessation of ambulation in patients with Duchenne muscular dystrophy," *Am. J. Phys. Med. Rehabil.*, vol. 81, no. 12, pp. 906–912, 2002, doi: 10.1097/00002060-200212000-00004.

[5] V. Humbertclaude *et al.*, "Motor and respiratory heterogeneity in Duchenne patients: Implication for clinical trials," *Eur. J. Paediatr. Neurol.*, vol. 16, no. 2, pp. 149–160, 2012, doi: 10.1016/j.ejpn.2011.07.001.

[6] J. Ropars and E. Salort-Campana, "Multidisciplinary management of Du chenne muscular dystrophy from childhood to adulthood," *Arch. Pediatr.*, vol. 32, no. 7S1, pp. 7S15–7S19, 2025, doi: 10.1016/S0929-693X(25)00248-9.

[7] I. Vandekerckhove *et al.,* "Gait classification for growing children with Duchenne muscular dystrophy," *Sci. Rep*., vol. 14, no. 1, p. 10828, 2024, doi: 10.1038/s41598-024-61231-y.

[8] P. M. Stevens, "Lower limb orthotic management of Duchenne muscular dystrophy: A literature review," *JPO J. Prosthetics Orthotics*, vol. 18, no. 4, pp. 111–119, Oct. 2006.

[9] M. A. Souza, A. Cezarani, and A. C. Mattiello-Sverzut, "Efeito do uso das órteses no prolongamento da marcha de pacientes com distrofia muscular de Duchenne: Revisão da literatura," *Acta Fisiátr.*, vol. 22, no. 3, pp. 155–159, 2015, doi: 10.5935/0104-7795.20150030. Available: https://revistas.usp.br/actafisiatrica/article/view/114536.

[10] M. E. Garralda, F. Muntoni, A. Cunniff, and A. D. Caneja, "Knee-ankle-foot orthosis in children with Duchenne muscular dystrophy: User views and adjustment," *Eur. J. Paediatr. Neurol.*, vol. 10, no. 4, pp. 186–191, 2006, doi: 10.1016/j.ejpn.2006.07.002.

[11] M. Gaumé, V. Cunin, and C. Vuillerot, "Orthopedic management in Duchenne muscular dystrophy," *Arch. Pediatr.,* vol. 32, no. 7S1, pp. 7S32–7S38, 2025, doi: 10.1016/S0929-693X(25)00251-9.

[12] J. B. Webster and D. Murphy, *Atlas of Orthoses and Assistive Devices*, 5th ed. Philadelphia, PA, USA: Elsevier, 2019.

[13] M. F. Rebel., J. de F. Landgraf , F. R. Sztajnbok, A. P. Q. C. Araújo "Ankle-foot orthosis in patients with Duchenne muscular dystrophy: A retrospective study," *Fisioter. Mov.*, vol. 37, 2024, doi: 10.1590/fm.2024.37110.

[14] F. S. Shahar *et al.*, "A review on the orthotics and prosthetics and the potential of kenaf composites as alternative materials for ankle-foot orthosis," J. Mech. *Behav. Biomed. Mater.*, vol. 99, pp. 169–185, Nov. 2019.

[15] F. S. Shahar, M. T. H. Sultan, A. U. Md Shah, S. N. A. Safri, "A comparative analysis between conventional manufacturing and additive manufacturing of ankle-foot orthosis," *Appl. Sci. Eng. Prog.*, vol. 13, no. 2, pp. 96–103, 2020.

[16] L. Demers, R. Weiss-Lambrou, B. Ska, “Item analysis of the Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST)”, Assist Technol, vol 12, pp. 96-105, Oct. 2010.

[17] K. E. C. De Carvalho, M. B. G. Júnior, K. N. Sá, “Tradução e Validação do Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST 2.0) para o idioma português do Brasil”, Rev. Bras. Reumatol., vol 54, pp. 260-267, Abr. 2014.

[18] M. D. Thye, H. M. Bednarz, A. J. Herringshaw, E. B. Sartin, R. K. Kana, "The impact of atypical sensory processing on social impairments in autism spectrum disorder," *Dev. Cogn. Neurosci.*, vol. 29, pp. 151–167, 2018.

[19] B. U. Sena and T. de S. Barros, "Hipersensibilidade em crianças com transtorno do espectro autista (TEA)," *Rev. Foco*, vol. 16, no. 11, p. e3502, 2023, doi: 10.54751/revistafoco.v16n11-012. [Online]. Available: https://ojs.focopublicacoes.com.br/foco/article/view/3502. Accessed: Jun. 30, 2026.

[20] B. B. Ribeiro, S. M. M. A. Lima, L. D. Santana, P. H. Gonçalves, "Estudo comparativo entre os métodos de produção tradicional e impressão 3D de órteses plantares," *Transverso*, ano 12, vol. 1, no. 15, pp. 133–145, Nov. 2024, doi: 10.36704/transverso.v1i15.9094. [Online]. Available: https://revista.uemg.br/index.php/transverso/issue/view/531.

[21] A. K. L. Filgueira, K. Y. N. Martins, C. A. M. dos S. Filho, K. E. Galdino, E. J. V. Oliveira, R. R. C. Branco, "Análise do impacto ambiental no desenvolvimento de órteses e próteses: Uma comparação entre o método convencional e o uso de tecnologias tridimensionais," *Rev. Cad. Pedagógico*, vol. 22, no. 14, p. e22521, 2025.
[22] S. Sharma *et al.*, "Utilization of rapid prototyping technology for the fabrication of an orthopedic shoe inserts for foot pain reprieve using thermo-softening viscoelastic polymers: A novel experimental approach," *Meas. Control*, vol. 53, no. 3–4, pp. 519–530, 2020.
[23] J. Copley and J. Ziviani, "Barriers to the use of assistive technology for children with multiple disabilities," *Occup. Ther. Int.*, vol. 11, no. 4, pp. 229–243, 2004, doi: 10.1002/OTI.213.
[24] F. H. F. Botelho, "Childhood and assistive technology: Growing with opportunity, developing with technology," 2021, doi: 10.1080/10400435.2021.1971330.

Desenvolvimento de uma cadeira de rodas infantil personalizada utilizando plásticos reciclados: Uma abordagem baseada em economia circular e manufatura digital

1st Lucas Gabriel Moreira Bernardes
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
lucasgabriel2345@discente.ufg.br

2nd Maria Luísa Lopes Tebas
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
maria_tebas@discente.ufg.br

3rd Laura Sousa Dantas
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
lauradantas@discente.ufg.br

4th Pedro Henrique Gonçalves
Faculdade de Artes Visuais
Universidade Federal de Goiás
Goiânia, Brasil
pedrogoncalves@ufg.br

Resumo- ***As cadeiras de rodas infantis constituem tecnologias assistivas desenvolvidas para promover melhorias na mobilidade, na participação social e na qualidade de vida de crianças com deficiência física. Entretanto, persistem dificuldades relacionadas ao acesso, ao custo e dificuldade de adaptação, evidenciando a necessidade de alternativas de baixo custo, personalizadas e sustentáveis. O objetivo é desenvolver e descrever um fluxo de manufatura digital para a produção de cadeiras de rodas infantis personalizadas a partir de plásticos reciclados, avaliando sua viabilidade técnica, produtiva e seu potencial como modelo de economia circular aplicado ao campo das tecnologias assistivas. Trata-se de um estudo de desenvolvimento tecnológico descritivo, as etapas consistiram na revisão de literatura, arrecadação de tampas de plástico, separação, trituração, prensagem, coleta das medidas antropométricas, modelagem tridimensional parametrizada, fresagem, impressão 3D e montagem da cadeira de rodas infantil. Foi desenvolvida uma cadeira de rodas infantil personalizada produzida a partir de plásticos reciclados, com peso de 4,01 kg, custo de produção de R$ 249,60 e tempo de produção de aproximadamente 11 horas e 10 minutos. O modelo apresentou baixo peso, baixo custo e capacidade de personalização conforme as medidas antropométricas do usuário. A integração entre economia circular e tecnologia assistiva representa uma estratégia promissora para a produção de cadeiras de rodas infantis personalizadas, reduzindo custos e impactos ambientais.***

Palavras- chave**—Cadeira de Rodas, Tecnologia Assistiva, Crianças, Reciclagem**

I. INTRODUÇÃO

As deficiências físicas em crianças contribuem para a presença de barreiras à participação, à inclusão, à independência, à prática de atividades físicas e favorecem o desenvolvimento de comorbidades associadas e deformidades [1,2]. Nesse sentido, as cadeiras de rodas infantis constituem tecnologias assistivas (TA) desenvolvidas para promover melhorias nas habilidades posturais, no suporte independente, na mobilidade ativa e na participação social [3,4].

Considerando sua importância para a funcionalidade e a inclusão social, observa-se um constante desequilíbrio entre a necessidade e a disponibilidade. Entre 10% e 15% da população mundial apresenta algum tipo de deficiência, e cerca de 10% dessas pessoas necessitam de cadeira de rodas para garantir mobilidade independente [5]. Apesar dessa demanda expressiva, a Organização Mundial da Saúde (OMS) estima que apenas 15% das pessoas que necessitam de uma cadeira de rodas conseguem obtê-la [6], evidenciando um importante desafio relacionado ao acesso [7].

No Brasil, a oferta de TA é prevista no art. 18, inciso XI, da Lei Brasileira de Inclusão da Pessoa com Deficiência (LBI), que garante o fornecimento de órteses, próteses e meios auxiliares de locomoção, conforme as normas do Ministério da Saúde, mediante cadastro e avaliação multiprofissional na rede do Sistema Único de Saúde (SUS) [8]. Entretanto, em 2023, Bastos [9] evidenciou insatisfação em relação ao desenvolvimento e ao uso dessas TA's, destacando a insuficiente compreensão das necessidades dos usuários, inadequações dimensionais dos dispositivos e a necessidade de um processo de desenvolvimento centrado no usuário. Além disso, persistem dificuldades relacionadas ao tempo de entrega, à manutenção e ao acompanhamento das cadeiras de rodas fornecidas pelo SUS. Na população infantil, essas limitações tornam-se ainda mais relevantes em razão do rápido crescimento das crianças, que exige adaptações frequentes dos dispositivos [10,11].

Nesse contexto, é necessário o desenvolvimento de alternativas de baixo custo e que apresentem alta capacidade de

alteração e personalização. Assim, o presente estudo teve como objetivo propor estratégias capazes de integrar acessibilidade, personalização e sustentabilidade em um mesmo processo produtivo. Bem como, desenvolver e descrever um fluxo de manufatura digital para a produção de cadeiras de rodas infantis personalizadas a partir de plásticos reciclados, avaliando sua viabilidade técnica, produtiva e seu potencial como modelo de economia circular aplicado ao campo das TAs, favorecendo a mobilidade, a inclusão social, a participação nas atividades cotidianas, e a qualidade de vida dessa população.

II. METODOLOGIA

Trata-se de um estudo de desenvolvimento tecnológico descritivo realizado nas dependências do Laboratório de Estudos Inventivos em Tecnologias Assistivas na Universidade Federal de Goiás por uma equipe de docentes e discentes dos cursos de Fisioterapia, Design de Produtos e Arquitetura. Este trabalho descreve o desenvolvimento de uma cadeira de rodas infantil a partir das medidas antropométricas de uma criança participante do projeto "Avaliação e acompanhamento de crianças de 0 a 5 anos com alterações neurológicas e/ou deficiências que necessitavam de tecnologias assistivas", aprovado no Comitê de Ética (Nº 7.688.714). No entanto, o processo de desenvolvimento para cada cadeira necessita da personalização do dispositivo conforme as características antropométricas de cada criança, o presente trabalho descreveu apenas um dos dispositivos entregues. O processo envolveu 10 etapas, dispostas no fluxograma (Fig. 1): revisão de literatura; arrecadação de tampas de plástico; separação das tampas por cores e tipos de plástico; trituração; prensagem e compactação das tampas trituradas; coleta das medidas antropométricas; modelagem da cadeira; fresagem das placas; produção das rodas e pneus e montagem.

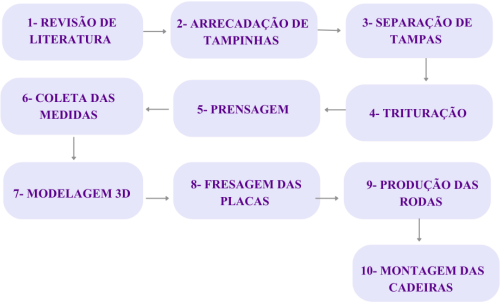


Fig. 1. Fluxograma das etapas

O processo foi estruturado segundo os princípios da economia circular, buscando manter os materiais em ciclos sucessivos de utilização. Dessa forma, desde a seleção da matéria-prima até a concepção estrutural da cadeira , priorizou-se a reutilização dos componentes e a possibilidade de reprocessamento dos materiais ao final da vida útil dos dispositivos.

Inicialmente foi realizada uma revisão de literatura na base de dados PUBMED, com busca dos descritores "wheelchair" e "sustainable development" utilizando o operador booleano "AND", os resultados foram filtrados para artigos publicados nos últimos cinco anos. Nessa etapa, buscou-se compreender os critérios de funcionalidade, adaptação e design necessários para o desenvolvimento de uma cadeira de rodas. Posteriormente, para a construção do dispositivo foi realizada a arrecadação das tampinhas plásticas recicláveis , por meio da divulgação do projeto nas redes sociais do laboratório, com informações sobre os locais para recebimento e os plásticos admitidos (Polietileno de Alta Densidade e Polipropileno), contando com 5 pontos de coleta. Em seguida, o material foi separado por cores e tipo de plástico, sendo descartadas as tampas que não se enquadraram no critério, essa ação reforça a sensibilização ambiental junto à comunidade, incorporando princípios de responsabilidade socioambiental ao processo de desenvolvimento tecnológico. .

Na fase seguinte, o material foi triturado duas vezes em uma trituradora de 3KWatts em velocidade de 6000 rotações por minuto (Fig. 2), e prensado na prensa térmica à 265ºC por 60 minutos para a criação de placas de 625 x 445 x 9 milímetros contendo Polietileno de Alta Densidade e Polipropileno que apresentou a fusão completa de ambos os materiais sem apresentar fissuras ou rachaduras (Fig. 3), e resfriado por 40 minutos em ar ambiente até as chapas metálicas se separarem da placa, as quais foram utilizadas para o recorte das peças.



Fig. 2. Trituração das tampas



Fig. 3. Prensagem e compactação

Uma fisioterapeuta realizou durante a avaliação a coleta das medidas antropométricas na posição sentada com uma fita antropométrica, além de avaliadas restrições de amplitude de movimento e alterações musculoesqueléticas para constatar se é necessário fazer adaptações adicionais. Os dados essenciais coletados eram: Altura do tronco (base do assento até o topo do ombro), Largura axilar/tronco (Laterais mais externas do tronco ao nível axilar), Largura quadril (Laterais mais externas dos

quadris), Comprimento da coxa (fossa poplítea até o ponto mais posterior do quadril) e Comprimento da perna (Base do calçado até a fossa poplítea). Elas foram utilizadas para modelar a altura do encosto, largura do encosto, largura do assento, profundidade do assento e distância do assento até o apoio dos pés, respectivamente.

Para a produção da cadeira foi adotada a modelagem tridimensional (3D), feita no programa Fusion 360 de forma parametrizada permitindo alterações dimensionais rápidas a partir das medidas antropométricas do usuário,possibilitando adaptações do dispositivo de forma individual sem necessidade de reconstrução completo do modelo digital. Adiante, o arquivo foi exportado em estereolitografia para a Fresadora CNC 3 eixos com tecnologia Nesting para a fresagem das placas de plástico produzidas anteriormente. Ademais, para compor a cadeira foram modeladas as rodas e os pneus no programa supracitado e foram impressas na Impressora 3D Creality K1 Max, com Polietileno Tereftalato Glicol (PETG) com Fibra de Carbono para as rodas e Poliuretano Termoplástico Flex (TPU) para a os pneus. Por fim, foi realizada a montagem da cadeira de rodas infantil por meio da união dos componentes estruturais utilizando parafusos, porcas, porcas travantes, espaçadores e rolamentos. Em seguida, foram instalados dois rodízios giratórios, um dianteiro e outro traseiro, concluindo a montagem da cadeira de rodas infantil.

III. RESULTADOS

Durante a revisão da literatura (etapa 1), foi possível identificar uma quantidade limitada de estudos que buscavam desenvolver cadeiras de rodas de baixo custo e com foco em sustentabilidade. Além disso, essa etapa forneceu subsídios para a compreensão dos processos envolvidos na prescrição e na avaliação do dispositivo de locomoção e dos principais desafios relacionados ao acesso e uso desses dispositivos.

Através da campanha de arrecadação foram obtidas cerca de 550 quilogramas no período de um ano e deste total já foram processados 417 quilogramas em 3 meses, considerando que aproximadamente 4.000 tampinhas foram necessárias para a produção de uma cadeira infantil, a quantidade arrecadada demonstra o elevado potencial de escalabilidade do processo produtivo, permitindo a fabricação sucessiva de dispositivos utilizando resíduos sólidos urbanos como matéria-prima. As tampinhas recebidas eram separadas de forma manual e armazenadas em caixas organizadoras na etapa 3, sendo necessárias 4 mil para cada cadeira de rodas infantil. Posteriormente, nas etapas 4 e 5, foi necessário 2 horas e 3 horas respectivamente para a sua execução, já que foi preciso 3 placas para o equipamento desenvolvido neste trabalho, mas a quantidade pode variar de acordo com o tamanho das peças. A Fig. 4 ilustra as placas fabricadas

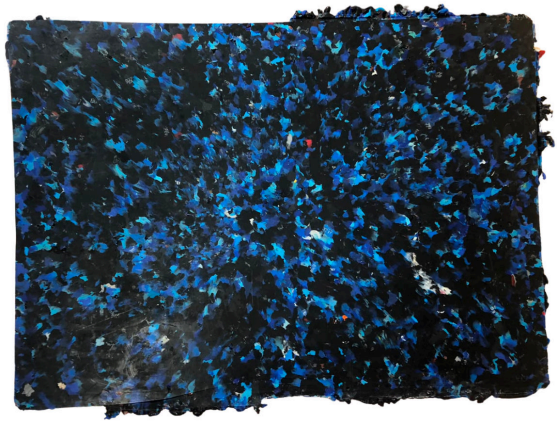


Fig. 4. Placa de tampinhas

As medidas antropométricas utilizadas para a realização do dispositivo foram descritos na tabela 1, a partir delas foi feita a etapa de personalização e adaptação subsequente e não foi necessário nenhum dispositivo adicional de adaptação relacionado ao posicionamento postural.

TABELA 1. MEDIDAS ANTROPOMÉTRICAS

MEDIDAS ANTROPOMÉTRICAS	TAMANHO (centímetros)
Altura do tronco (Base do assento até o topo do ombro)	25
Largura quadril (Laterais mais externas dos quadris)	32
Largura axilar/tronco (Laterais mais externas do tronco ao nível axilar)	35
Comprimento da coxa (fossa poplítea até o ponto mais posterior do quadril)	22
Comprimento da perna (Base do calçado até a fossa poplítea).	28

Concomitante às etapas anteriores, a modelagem 3D resultou em seis peças de plástico reciclável: Encosto, assento, apoio para os pés, lateral direita, lateral esquerda e apoio para a roda traseira. Além das peças com esse material, foi necessário a produção de 2 rodas, a disposição das rodas estão ilustradas na figura 7, as quais compuseram o resultado final (Fig. 5).

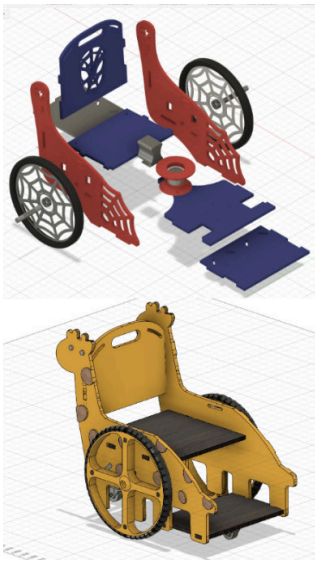


Fig. 5. Exemplos de modelos 3D das cadeiras de rodas infantil

Com a modelagem finalizada, na fresagem as seis peças foram cortadas, que durou aproximadamente 30 minutos, e a impressão das rodas, que ocorreu em 8 horas, com filamento de Polietileno Tereftalato Glicol (PETG), resultando em um peso de 630 gramas cada uma. Com todas as peças prontas, ocorreu a última etapa, a montagem da cadeira de rodas infantil. As peças e materiais utilizados constam na tabela 2 e levou aproximadamente 40 minutos.

Tabela 2- Materiais e peças utilizadas

Material	Quantidade
Parafusos 3,5 x 22 mm	2
Parafusos 3,5x14	26
Porcas	4
Porcas travantes	2
Rolamentos	2
Espaçadores 2 cm x 1,5 cm de diâmetro	2
Rodízio giratório 100 mm	1
Rodízio giratório 25 mm	1
Arruelas 28 mm	4

A cadeira de rodas infantil totalizou um peso de 4,01 quilogramas, precisou de aproximadamente 11 horas e 10 minutos para sua produção e teve um custo de 249,60, resultando no produto da imagem (Fig. 6) , o baixo peso final do dispositivo representa uma característica relevante para cadeiras infantis, pois reduz o esforço necessário para transporte pelos cuidadores e potencialmente diminui a resistência à propulsão quando comparado aos modelos convencionais disponíveis comercialmente.

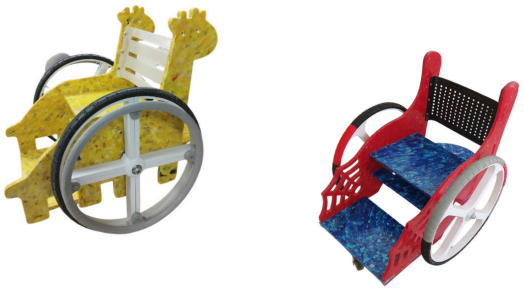


Fig. 6. Exemplos das cadeiras de rodas infantis finalizadas



Fig. 7. Disposição das rodas

IV. DISCUSSÃO

Usuários de cadeiras de rodas enfrentam muitas barreiras para a sua aquisição, sendo a principal delas o custo elevado [7]. Para crianças essa situação é ainda mais grave, visto que há o crescimento infantil e a necessidade de adaptações para cada usuário, dificultando o acesso a esse equipamento [10]. Pensando nisso, buscou-se materiais de baixo custo com boa durabilidade e possibilidade de personalização para a confecção do dispositivo desenvolvido neste trabalho. Alguns dos materiais e técnicas cogitados foram o Medium Density Fiberboard (MDF) por meio do corte a laser e o Polietileno Tereftalato Glicol (PETG) por impressão 3D, contudo, ambos foram descartados devido aos critérios mencionados anteriormente.

Mais do que desenvolver uma cadeira de rodas infantil, o presente trabalho propôs um modelo integrado de produção de tecnologia assistiva baseado na economia circular, no qual resíduos plásticos são reinseridos em uma cadeia produtiva capaz de gerar dispositivos personalizados de baixo custo. Essa abordagem amplia o conceito tradicional de manufatura de cadeiras de rodas ao incorporar simultaneamente aspectos ambientais, econômicos e sociais. Essa característica aumenta significativamente o potencial de replicação do modelo em laboratórios, makers, centros de reabilitação e instituições públicas.

Visando a sustentabilidade, também procurou-se materiais recicláveis que poderiam ser utilizados no desenvolvimento, sendo encontrados os plásticos Polietileno de Alta Densidade e Polipropileno que poderiam ser derretidos e transformados em placas. Para a realização desses processos, foi utilizado o maquinário disponível no Laboratório de Estudos Inventivos em Tecnologias Assistivas, mas sendo possível a utilização de outros equipamentos ou o aluguel desses.

Os dispositivos disponíveis no mercado em sua maioria são feitos de metal e possuem baixa capacidade de adaptação. Não foi encontrado na literatura artigos que descrevessem média de preço e peso das cadeiras de rodas infantis, mas durante uma análise de preços em cinco lojas online de fornecedores nacionais selecionadas por conveniência (mobility Brasil, Casa Ortopédica, Drogaria Araujo, Casa Médica e Medicar Material), notou-se que o custo médio das cadeiras de rodas infantil manual foi de 1.640,04 reais, o peso das cadeiras variam entre 12 e 20 kg. Em relação a esses aspectos foi demonstrado que há uma vantagem socioeconômica, pois o custo de produção é cerca de 15% do valor médio de mercado significativamente menor quando comparados.

Ainda quando comparamos o peso, o modelo proposto é aproximadamente 67% mais leve que o modelo mais leve do mercado. Esse fator reflete diretamente no esforço necessário para a propulsão da cadeira de rodas. Esse aspecto é relevante, pois as forças exercidas nos ombros causam estresse nas estruturas anatômicas, somadas aos movimentos repetitivos dos membros superiores. A situação se torna ainda pior quando há presença de dor, que contribui para a baixa variabilidade escapular de rotação interna e externa, causando maior fadiga tecidual cumulativa. Isso contribui também para alterações na biomecânica, com maiores oscilações de tórax, menor pico de torque no ombro e maior variabilidade de torque como estratégias de redução da dor [12,13].

Ademais, em 2015 a Organização das Nações Unidas definiu 17 metas para serem atingidas até 2030, os Objetivos de Desenvolvimento Sustentável (ODS) [14], tendo o Brasil como país signatário. Dentre esses objetivos estão a promoção da saúde e bem- estar (objetivo 3) e a garantia da produção e consumo sustentáveis (objetivo 12). Nesse sentido, o recurso desenvolvido buscou reciclar o material plástico e utilizá-lo de forma a promover os objetivos supracitados. Tendo retirado do meio ambiente 4.000 tampinhas a cada cadeira de rodas infantil desenvolvida, essa iniciativa é altamente reprodutiva e escalável. Como supracitado, com poucos pontos de coletas e no período de um ano foi possível arrecadar aproximadamente 550 kg, cerca de 311.000 tampas de garrafas, desse total foram processado 417 kg, 75% do total, contando apenas com as duas máquinas da figura 2 e 3, reforçando a importância social e ambiental envolvida na produção.

O modelo proposto apresenta características de sustentabilidade tanto em sua produção quanto nas etapas de pós-uso e reaproveitamento. Para isso, foram previstas duas estratégias de reutilização dos componentes da cadeira de rodas. A primeira consiste no repasse do equipamento para outra criança quando a estatura do, desde que a estrutura esteja preservada e possua compatibilidade com as adaptações. A segunda estratégia consiste no desmonte do produto para que os materiais empregados em sua fabricação possam ser reaproveitados. Os elementos de fixação, como parafusos, porcas, porcas travantes, espaçadores e rolamentos, podem ser reutilizados em uma nova montagem. Da mesma forma, as rodas e os rodízios giratórios podem ser utilizados em outro modelo de cadeira de rodas. Já os componentes plásticos podem ser reinseridos no processo de produção por meio do reprocessamento por prensagem, possibilitando sua reutilização na fabricação de novas cadeiras.

As limitações observadas na fase da produção desse dispositivo foi em relação às dimensões das placas de plástico reciclado devido ao tamanho da prensa térmica, que restringiu o comprimento e a altura das peças, delimitando os usuários dessa cadeira às crianças com estaturas menores.Além dissos, este estudo não realizou testes de durabilidade, avaliação da satisfação dos usuários e acompanhamento a longo prazo.

Um aspecto que diferencia o presente estudo de iniciativas convencionais de reciclagem consiste no fato de que o material reciclado não foi utilizado apenas como substituto de matéria-prima, mas incorporado a um fluxo completo de desenvolvimento de Tecnologia Assistiva. Nesse processo, resíduos plásticos urbanos foram transformados em componentes estruturais personalizados por meio de modelagem digital e manufatura assistida por computador, estabelecendo uma integração entre sustentabilidade ambiental, inovação tecnológica e reabilitação. Essa abordagem amplia o potencial de aplicação da economia circular para além da gestão de resíduos, posicionando-a como estratégia para democratização do acesso às tecnologias assistivas.

V. CONCLUSÃO

A cadeira de rodas infantil de tampinha reciclável desenvolvida apresentou um baixo custo, peso reduzido e capacidade de personalização para cada usuário, atendendo às principais limitações encontradas para a aquisição desse dispositivo. Os resultados evidenciam que a integração entre economia circular e Tecnologia Assistiva constitui uma estratégia promissora para ampliar o acesso a dispositivos personalizados, reduzindo custos e impactos ambientais simultaneamente. Além disso, possui grande capacidade de reprodutibilidade e um impacto positivo para a redução e a reciclagem do material plástico descartado no meio ambiente, durante a produção e após o uso dos usuários. Nesse sentido, esse projeto promove a acessibilidade para os usuários de cadeira de rodas e propõe a reprodutibilidade do produto em outros laboratórios e centros de desenvolvimento. No entanto ainda se faz necessário, novas pesquisas que tragam estudos de resistência do material, da estrutura da cadeira, como também avaliem o uso da tecnologia assistiva por essa população. Mais do que a fabricação de um produto, o trabalho apresenta um

modelo produtivo potencialmente replicável por instituições de ensino, laboratórios de inovação e serviços de reabilitação.

REFERÊNCIAS

[1] S. Seemüller, F. Beck, A. K. Reimers, “Physical activity of children and adolescents who use a wheelchair: a systematic review.” BMC Public Health. vol.23, Dec 2023. doi:10.1186/s12889-023-17201-6

[2] T. D. O'Brien, J. Noyes, L. H. Spencer, H. P. Kubis, R. P. Hastings, R. Whitaker, “Systematic review of physical activity and exercise interventions to improve health, fitness and well-being of children and young people who use wheelchairs.” BMJ Open Sport Exerc Med., vol 2, Nov. 2016. doi:10.1136/bmjsem-2016-000109

[3] World Health Organization. “Guidelines on the provision of manual wheelchairs in less resourced settings.”, Geneva: World Health Organization, Mai 2008. .

[4] S. G. Paguinto, N. A. Kasparian, P. Bray, M. Farrar, "Multidisciplinary perspectives and practices in wheelchair prescription for children with neuromuscular diseases," Disabil. Rehabil. Assist. Technol., vol. 18, pp. 166-174, Fev. 2023, doi: 10.1080/17483107.2020.1839793.

[5] N. Bray, J. Noyes, R. T. Edwards, N. Harris, "Wheelchair interventions, services and provision for disabled children: a mixed-method systematic review and conceptual framework," *BMC Health Serv. Res.*, vol. 14, p. 309, Jul. 2014, doi: 10.1186/1472-6963-14-309.

[6] World Health Organization, “World report on disability”. Geneva, Switzerland: World Health Organization, Dec 2011. Available: https://www.who.int/publications/i/item/9789241564182.

[7] J. Srishti, B. Avijit, H. Gita, “Patterns and unmet needs in wheelchair utilization: A mixed-method study of user experience at a tertiary hospital in India”, Journal of Rehabilitation and Assistive Technologies Engineering, vol. 13, pp. 1-13, 2026

[8] Brasil, Lei n.º 13.146, de 6 de julho de 2015: Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência), Diário Oficial da União, Brasília, DF, Jul. 7, 2015.

[9] P. A. L. S. Bastos, M. S. Silva, N. M. Ribeiro, R. S. Mota, T. G. Filho, “Tecnologia assistiva e políticas públicas no Brasil.” Cadernos Brasileiros de Terapia Ocupacional, vol. 31, Jan. 2023 https://doi.org/10.1590/2526-8910.ctoAO260434011
1016/j. Ejpn.2011.07.001

[10] R. S. Curi, E. C. Ferretti, R. C. Franco, A. L. C. Allegretti, M. S. P. da Silva, "Satisfação de usuários de cadeiras de rodas fornecidas pelo SUS: estudo transversal na Baixada Santista," Rev. Bras. Saúde Func., vol. 8, no. 2, pp. 45-53, 2024.

[11] World Health Organization, United Nations Children's Fund. *Global report on assistive technology*. Geneva: WHO; 2022.

[12] S. J. Briley, R. J. K. Vegter, V. L. Goosey-Tolfrey, B. S. Mason, “Scapular kinematic variability during wheelchair propulsion is associated with shoulder pain in wheelchair users.” J Biomech., vol, 113, Dec, 2020. doi: 10.1016/j.jbiomech.2020.110099.

[13] Briley SJ, Vegter RJK, Goosey-Tolfrey VL, Mason BS. The longitudinal relationship between shoulder pain and altered wheelchair propulsion biomechanics of manual wheelchair users. J Biomech. 2021 Sep 20;126:110626. doi: 10.1016/j.jbiomech.2021.110626. Epub 2021 Jul 12. PMID: 34329882.

[14] D. K. A. Cruz, A. A. da Nóbrega, V. O. de M. Pereira, and M. de M. S. Montenegro, “Os Objetivos de Desenvolvimento Sustentável e as fontes de dados para o monitoramento das metas no Brasil," Epidemiol. Serv. Saúde, vol. 31, 2022, doi: 10.1590/S2237-9622202200010.

Desenvolvimento de uma Instrumentação de Baixo Custo para Avaliação de Pressão na Escrita Manual

Gabriel L. dos Santos
Dep. de Eng. Biomédica - UFPE
Universidade Federal de Pernambuco
Recife, Brasil
gabriel.limasantos@ufpe.br

Ana Karina P. da Silva Cabral
Dep. de Terapia Ocupacional - UFPE
Universidade Federal de Pernambuco
Recife, Brasil
anakarina.cabral@ufpe.br

Marcus C. de Araújo
Dep. de Eng. Mecânica - UFPE
Universidade Federal de Pernambuco
Recife, Brasil
marcus.araujo@ufpe.br

***Resumo*—Este estudo aborda a necessidade de ferramentas acessíveis para avaliação da pressão exercida pelos dedos durante a escrita manual, relevante para análises em terapia ocupacional e reabilitação motora. O objetivo foi aprimorar um protótipo de baixo custo capaz de medir a pressão aplicada pelos dedos durante a escrita utilizando uma plataforma baseada em Arduino. O sistema trabalha com sensores de pressão conectados a uma plataforma Arduino e integrado a um software desenvolvido em Python para aquisição, armazenamento e análise dos dados coletados. Os resultados indicaram que o protótipo foi capaz de registrar variações de pressão entre os dedos durante a escrita. Conclui-se que o sistema apresenta potencial como ferramenta de apoio para coleta de dados grafomotores.**

***Palavras-chave*—Sensor de pressão, Arduino, Escrita manual**

I. Introdução

A escrita manual é uma atividade complexa que envolve coordenação motora fina, controle da força aplicada pelos dedos e integração entre sistemas sensoriais e motores [1]. Durante essa atividade, a pressão exercida pelos dedos sobre o instrumento de escrita pode influenciar diretamente o desempenho motor, o conforto e a qualidade da escrita, sendo um parâmetro relevante para a análise de padrões motores [2].

No contexto da terapia ocupacional, a avaliação da escrita manual pode contribuir para a identificação de dificuldades motoras e para o planejamento de intervenções voltadas à reabilitação funcional. Estudos indicam que a análise de parâmetros biomecânicos, como força aplicada e padrão de movimento durante a escrita, pode auxiliar na compreensão de alterações no controle motor fino [3], [4].

Entretanto, dispositivos capazes de medir quantitativamente a pressão exercida durante a escrita geralmente apresentam custos elevados ou acesso limitado, o que dificulta sua aplicação em ambientes clínicos, educacionais e de pesquisa. Nesse contexto, o desenvolvimento de soluções tecnológicas de baixo custo surge como uma alternativa para ampliar o acesso a ferramentas de avaliação motora e de coleta de dados experimentais. Plataformas de prototipagem eletrônica, como o Arduino, possibilitam o desenvolvimento de sistemas acessíveis para aquisição de dados por meio de sensores e integração com sistemas computacionais. Além disso, ferramentas de programação científica, como a linguagem Python, permitem o armazenamento, processamento e visualização dos dados coletados, favorecendo a análise dos parâmetros medidos.

Dessa forma, este trabalho dá continuidade a um protótipo previamente desenvolvido em um trabalho de conclusão de curso, propondo melhorias no sistema de medição da pressão exercida pelos dedos durante a escrita manual. A principal contribuição deste estudo é a integração de uma instrumentação de baixo custo e de fácil implementação, baseado em sensores resistivos, microcontrolador Arduino e ambiente de processamento em Python, voltado para a aquisição e análise comparativa preliminar de sinais de pressão exercidos na atividade de escrita, servindo como uma ferramenta para futuras investigações ergonômicas e clínicas.

II. Metodologia

A. Descrição Geral do Sistema

A arquitetura do sistema é composta por três etapas principais: aquisição de dados por meio de sensores de pressão, processamento inicial realizado pelo microcontrolador e transmissão dos dados para um ambiente computacional para visualização, armazenamento e análise (Figura 1).

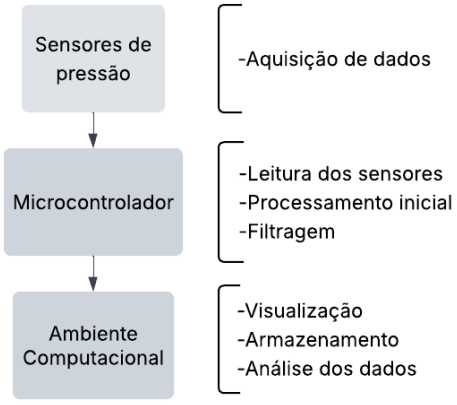


Figura 1. Fluxograma do funcionamento do sistema de avaliação de pressão

Financiado pela FINEP (Financiadora de Estudos e Projetos), Processo nº 2163/22.

B. Hardware

O sistema de aquisição foi construído utilizando a plataforma Arduino (microcontrolador ATmega328P, ADC de 10 bits de resolução). Foram utilizados três sensores de pressão do tipo resistivo (FSR 402), com faixa de medição nominal de 0 a 5kg, cuja resistência elétrica varia inversamente à força compressiva aplicada, permitindo a conversão do sinal mecânico em sinal elétrico, captado pelo sistema.

O sensor foi acoplado ao instrumento de escrita de forma a captar diretamente a pressão exercida pelos dedos durante a atividade. O circuito foi configurado em um divisor de tensão com resistores fixos de $10\,\mathrm{k}\Omega$, permitindo a leitura do sinal analógico por meio de uma das entradas do Arduino. Ao todo, foram usados 3 sensores para a coleta dos dados de pressão. Tal esquema pode ser visto na Figura 2.

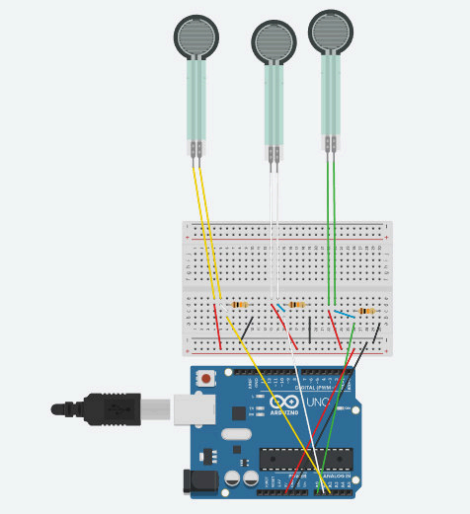


Figura 2. Esquemática do sistema de avaliação de pressão via *Tinkercad*

C. Aquisição e Monitoramento de Dados

A aquisição dos dados foi realizada por meio da leitura contínua dos sinais analógicos dos sensores, utilizando o conversor analógico-digital (ADC) do Arduino. Os valores foram amostrados com um período de 200 milisegundos (taxa de amostragem de 5Hz) e transmitidos via comunicação serial. Não houve filtragem digital aplicada aos sinais.

Para o monitoramento em tempo real e validação inicial do funcionamento do sistema, foi utilizado o software *Tera Term*, que possibilita a visualização dos dados enviados pela porta serial. Essa ferramenta permitiu verificar a integridade dos sinais, identificar falhas e realizar ajustes no sistema durante a fase de desenvolvimento.

Além disso, o *Tera Term* foi utilizado para o registro preliminar dos dados em formato de texto, possibilitando uma análise inicial e validação do comportamento do sensor antes da integração com o ambiente computacional principal.

D. Calibração dos sensores

Os sensores de pressão foram calibrados utilizando-se um dinamômetro hidráulico de mão (Figura 3), correlacionando a força mecânica aplicada com o sinal analógico gerado no Arduino.

O procedimento de calibração consistiu em aplicar uma força manual constante sobre o sensor por um período de tempo de dez segundos (Figura 4), coletando os dados de medição do Arduino durante esse período. A partir dos dados coletados, utilizou-se a média das amostras obtidas durante o tempo de aplicação da força como valor de calibração. Foram coletados dados correspondentes a 1kgf, 2kgf, 4kgf e 6 kgf.

Com os pontos de calibração foi possível gerar a curva de calibração e converter os valores arbitrários analógicos em força (N).

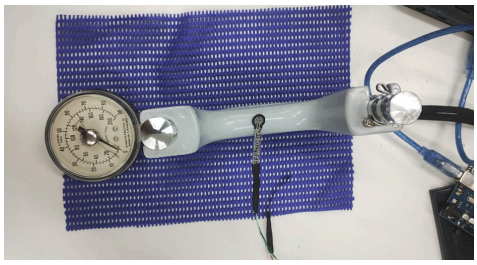


Figura 3. Dinamômetro hidráulico de mão



Figura 4. Aplicação de força manual constante

E. Software e Processamento de Dados

Para o tratamento e visualisação dos dados, foi desenvolvido um ambiente computacional, utilizando a linguagem Python, para recepção, armazenamento e análise dos dados. A comunicação com o Arduino foi realizada por meio da biblioteca *pyserial*, permitindo a leitura contínua dos dados transmitidos. Para o processamento dos dados, foram utilizadas bibliotecas como *NumPy* e *Matplotlib*, possibilitando a análise dos sinais de pressão e a geração de gráficos

representativos. Os dados coletados foram organizados em arquivos estruturados, facilitando a análise posterior.

F. Procedimento Experimental

A instrumentação proposta foi desenvolvida para servir de ferramenta de coleta de dados para o projeto de pesquisa "Desenvolvimento de Tecnologia Assistiva para autonomia cotidiana/laboral de pessoas com deficiência e formação de recursos humanos (sigla TECLABORAL), aprovada em chamada pública MCTI/FINEP/AÇÃO TRANSVERSAL – TECNOLOGIA ASSISTIVA para a estruturacao do Sistema Nacional de Laboratorios de Tecnologia Assistiva (SISASSISTIVA-MCTI), ref. Finep n. 2163/22.

O referido projeto foi e aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos, sob Número do Parecer: 7.719.058, CAAE 89977025.9.0000.5208. A fim de avaliar a aplicabilidade da instrumentação, realizou-se um estudo de caso com um voluntário. O participante em questão apresenta dificuldade de coordenação dos membros superiores decorrente de um acidente de moto. Os experimentos foram conduzidos em ambiente controlado, no Laboratório de Tecnologia Assistiva e Terapia Ocupacional (LabTATO) da Universidade Federal de Pernambuco, mantendo-se o mesmo instrumento de escrita, postura e superfície de apoio durante todos os testes.

Durante os testes, o participante foi instruído a escrever textos padronizados. O teste foi dividido em duas etapas. Na primeira o participante realizou a atividade utilizando uma caneta e vestindo a luva instrumentada. Na segunda etapa o participante repetiu a atividade utilizando uma adaptação para lápis e a luva instrumentada (Figura 5). Os dados obtidos nas duas etapas serão comparados entre si para avaliar a aplicabilidade da instrumentação para a coleta de dados de pressão para avaliação grafomotora.

A escolha dos três pontos de instrumentação baseou-se na identificação prévia das áreas de maior contato, por meio de um teste piloto: o participante vestiu uma luva de látex e escreveu seu nome usando uma caneta banhada em tinta. Dessa maneira, a tinta transferida para a luva permitiu identificar os principais pontos de pressão entre a caneta e a mão do voluntário, correspondentes ao polegar, dedo indicador e área de apoio da mão. Tal procedimento foi fundamental no auxílio da definição das áreas de instalação dos sensores.

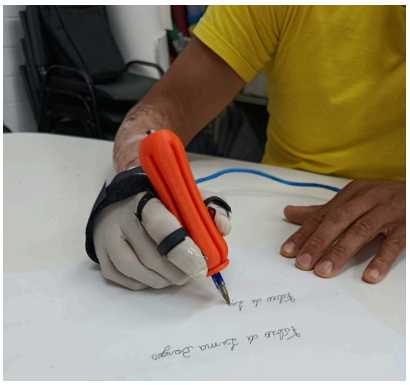


Figura 5. Utilização do instrumento adaptado com os 3 sensores

Após o posicionamento adequado dos sensores, iniciou-se o experimento e os sinais de pressão foram registrados ao longo do tempo, permitindo a análise do comportamento da força aplicada durante a escrita, com e sem adaptador. A coleta era iniciada antes do inicio de cada atividade e interrompida após o fim da atividade. Antes da atividade e após largar a caneta, os sensores marcam zero e dessa forma é possível sincronizar os dados com a atividade realizada. Os dados obtidos de pressão na escrita foram coletados em um arquivo .log para posterior arguição e análise.

III. Resultados e discussões

O sistema desenvolvido coletou dados de pressão durante uma atividade grafomotora. A comunicação serial entre o Arduino e o software desenvolvido em Python ocorreu sem perdas perceptíveis de dados, permitindo a aquisição contínua dos sinais provenientes dos sensores de pressão. Além disso foram gerados gráficos que representam a diferença de pressão proveniente dos três sensores no experimento da escrita sem o adaptador e com o adaptador. O resultado dos gráficos está evidenciado nas Figuras 6 e 7.

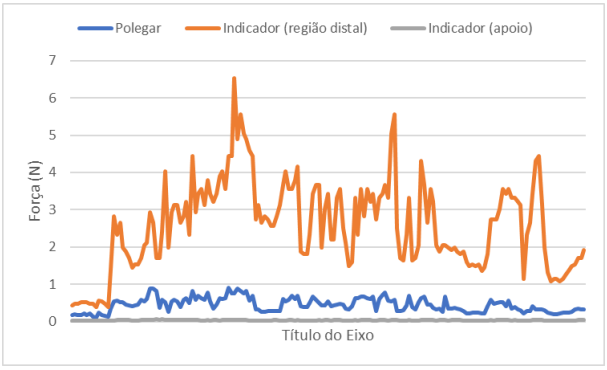


Figura 6. Gráfico de pressão na escrita sem adaptador

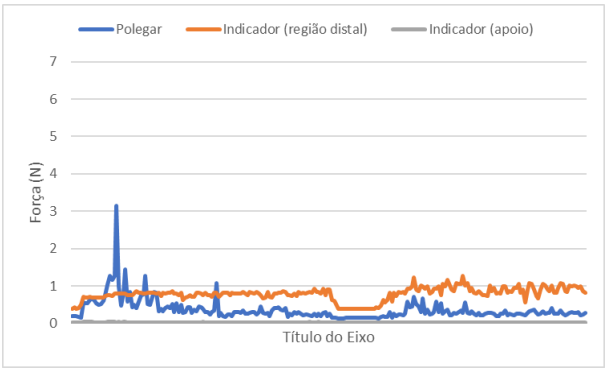


Figura 7. Gráfico de pressão na escrita com adaptador

A análise dos sinais obtidos evidenciou diferenças no comportamento da pressão aplicada pelos dedos entre as duas condições experimentais. Na escrita sem o adaptador, observou-se maior variabilidade dos sinais e ocorrência de

picos elevados de pressão, principalmente nos sensores posicionados nas regiões correspondentes ao polegar e ao dedo indicador. Esses resultados sugerem maior concentração de força durante a empunhadura convencional da caneta.

Por outro lado, durante a utilização do adaptador instrumentado, verificou-se redução nos picos de pressão e maior estabilidade dos sinais ao longo da atividade de escrita. A análise quantitativa dos dados revelou que, na condição sem adaptador, o pico máximo de pressão do dedo que foi maior acionado (nesse caso foi o indicador na região distal) foi de 6,53N, com uma média de 2,59N. Em contraste, a utilização do adaptador reduziu o pico máximo para 1,27N e a média para 0,79N. O desvio padrão foi de 1,182 sem adaptador e 0,175 com adaptador.

A redução da pressão concentrada em regiões específicas da mão pode contribuir para maior conforto durante a escrita, além de potencialmente reduzir fadiga muscular associada à realização prolongada dessa atividade. Esses resultados demonstram o potencial do sistema proposto como ferramenta de apoio para análise biomecânica da escrita manual e futuras aplicações em contextos de ergonomia e reabilitação motora.

Entretanto, o presente estudo apresenta limitações relacionadas a validação frente a equipamentos de referência padrão-ouro para confirmar sua exatidão, histerese e repetibilidade. Para ampliar a relevância, estudos futuros devem buscar correlacionar os sinais de pressão adquiridos com parâmetros funcionais da escrita, como legibilidade, velocidade e fadiga.

IV. Conclusões

O presente trabalho apresentou o desenvolvimento de um protótipo de baixo custo para aquisição e análise da pressão exercida pelos dedos durante a escrita manual, utilizando sensores de pressão integrados a uma plataforma Arduino e a um ambiente computacional desenvolvido em Python.

Os resultados experimentais demonstraram que o sistema foi capaz de realizar a aquisição contínua dos sinais provenientes dos sensores, sem perdas de dados durante a comunicação serial. Além disso, os gráficos gerados permitem que profissionais da área de Terapia Ocupacional sejam capazes de analisar o comportamento da pressão exercida pelos dedos em diferentes condições de escrita.

A comparação entre a escrita convencional e a escrita utilizando um adaptador instrumentado evidenciou redução nos picos de pressão, principalmente nas regiões correspondentes ao polegar e ao dedo indicador. Os principais aprendizados obtidos incluem a viabilidade da integração de sensores resistivos à plataforma Arduino e o processamento inicial em Python de forma estável, permitindo a observação preliminar de que o daptador instrumentado pode influenciar a distribuição das forças e reduzir picos de preensão localizados.

Dessa forma, o protótipo apresentou potencial como ferramenta de apoio para estudos de biomecânica da escrita, ergonomia e estudos sobre reabilitação motora. Como perspectivas futuras, pretende-se realizar testes com um maior número de participantes, assim como expandir a análise computacional dos dados para aplicações clínicas e terapêuticas.

Referências

[1] I. Rosenblum, P. L. Weiss, and S. Parush, “Product and process evaluation of handwriting difficulties,” 2003.

[2] J. Danna and J.-L. Velay, “Basic and supplementary sensory feedback in handwriting,” 2015.

[3] H. Schwellnus, H. Carnahan, A. Kushki, H. Polatajko, C. Missiuna, and T. Chau, “Writing forces associated with four pencil grasp patterns in grade 4 children,” *American Journal of Occupational Therapy*, vol. 67, no. 2, pp. 218–227, 2013.

[4] S.-H. Chang and N.-Y. Yu, “Characterization of motor control in handwriting difficulties in children with or without developmental coordination disorder,” *Developmental Medicine & Child Neurology*, vol. 52, no. 3, pp. 244–250, 2010.

Desenvolvimento de uma tala extensora personalizada para o uso associado ao parapodium em crianças com Paralisia Cerebral

1st Maria Luísa Lopes Tebas
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
maria_tebas@discente.ufg.br

2nd Lucas Gabriel Moreira Bernardes
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
lucasgabriel2345@discente.ufg.br

3rd Vitória Souza de Oliveira
Departamento de Saúde Funcional
Universidade Federal de Goiás
Goiânia, Brasil
oliveira_oliveira@discente.ufg.br

4th Pedro Henrique Gonçalves
Faculdade de artes visuais
Universidade Federal de Goiás
Goiânia, Brasil
pedrogoncalves@ufg.br

Resumo- A paralisia cerebral é um distúrbio neurológico que compromete o desenvolvimento motor e dificulta a manutenção da posição ortostática, tornando necessária a utilização de tecnologias assistivas para promover o alinhamento postural e a descarga de peso. O presente estudo teve como objetivo desenvolver uma tala extensora personalizada para utilização associada ao parapodium em crianças com paralisia cerebral com incapacidade de manter a extensão dos joelhos durante o ortostatismo. Trata-se de um estudo de desenvolvimento tecnológico descritivo, no qual foi realizada a busca na literatura e no mercado, desenvolvimento do molde, escolha dos materiais, coleta de medidas antropométricas, produção do protótipo, testagem, avaliação fisioterapêutica e percepção de satisfação através do B-QUEST após dois meses de uso. O dispositivo desenvolvido apresentou adequada adaptação ao usuário, manutenção da extensão dos joelhos durante a descarga de peso, alinhamento biomecânico do quadril e membros inferiores, ausência de dor ou desconforto e elevada satisfação em todos os domínios do B-QUEST. A tala apresentou custo de produção de R$ 58,72, correspondente a 16,34% da média de preço encontrada na busca de mercado. Os resultados demonstram a viabilidade de produção, alta aceitabilidade e o potencial da tala extensora como recurso auxiliar para o ortostatismo durante o uso do parapodium.

***Palavras- chave*— Paralisia cerebral; Tecnologia assistiva; Ortostatismo; Tala extensora.**

I. INTRODUÇÃO

Os distúrbios neurológicos são caracterizados por acometimentos no sistema nervoso que causam comprometimentos no desenvolvimento motor, na função física, no cognitivo e/ou na fala, e esses se subdividem em diferentes diagnósticos clínicos, como a Paralisia Cerebral (PC) [1,2]. A PC ocorre em crianças de até 3 anos de idade, com prevalência geral de 2 por 1.000 nascidos vivos, e classificada pelo Gross Motor Functional Classification Sys (GMFCS) em 5 níveis [2,3,4]. Nos níveis mais altos, IV e V, os indivíduos não mantêm de forma independente posturas anti-gravíticas, como o sentar e a posição ortostática, sendo necessário equipamentos auxiliares para a sua manutenção [2]. A falta da postura ortostática e da descarga de peso na articulação do quadril nessa população são as principais causas de subluxação e luxação de quadril, acarretando no desenvolvimento de outras disfunções, como escoliose, dores progressivas e obliquidade pélvica [5,6]. Ademais, o tempo exacerbado em decúbito, aumenta o risco da ocorrência de complicações gerais (broncoaspiração, hemorragias e pneumonias) e lesões ortopédicas (subluxação e luxação de quadril e pé equino) [5,7]. Por isso, é recomendado treinos ortostáticos dentro dessa população para evitar esses desfechos e promover os benefícios dessa posição, como o aumento da deposição de cálcio nos ossos, favorecimento da função dos órgãos, melhora do controle autônomo cardiovascular e da deglutição e estimulação vestibular [5,8].

Um dos equipamentos responsáveis por promover a verticalização em crianças neurológicas é o parapodium (figura X), uma tecnologia assistiva que fomenta o alinhamento postural e a descarga de peso [8,9]. Nos níveis IV e V da PC, esse alinhamento é prejudicado comparado ao desenvolvimento típico, comprometendo o desempenho motor para a manutenção de posturas altas, sendo necessário a utilização de equipamentos acessórios para a manutenção da função e estabilização das articulações [10,11].

O ortostatismo proporciona a postura ereta essencial no desenvolvimento infantil, a força gravitacional axial promove maior estabilidade postural e do centro de gravidade promovendo alinhamento das articulações do tornozelo, joelho e quadril e das curvaturas da coluna vertebral [12]. No entanto, o indivíduo com PC apresenta fraqueza muscular e contraturas que

dificultam o ortostatismo, e causam maiores demandas energéticas, sendo necessário dispositivos auxiliares de estabilização [12].

A tala extensora é um desses equipamentos que ajuda na biomecânica esquelética, por meio de hastes rígidas de alumínio cobertas por tecido que provocam a extensão do joelho. Evitando a assimetria entre os membros inferiores por flexão da articulação e adaptações posturais compensatórias, propiciando o alinhamento dos membros que é fundamental em treinos ortostáticos e para o uso do parapodium.

Apesar dos benefícios proporcionados pelo parapodium, crianças com comprometimento motor grave frequentemente apresentam incapacidade de manter o alinhamento adequado dos membros inferiores apenas com esse dispositivo, especialmente em decorrência da flexão persistente dos joelhos. Embora essa limitação seja observada na prática clínica, há escassez de estudos que descrevam o desenvolvimento de recursos auxiliares específicos para ampliar a estabilidade durante o ortostatismo. Partiu-se da hipótese de que a associação de uma tala extensora personalizada ao parapodium seria capaz de proporcionar melhor alinhamento clínico dos membros inferiores durante o ortostatismo em crianças com paralisia cerebral que apresentam incapacidade de manter a extensão dos joelhos apenas com o suporte convencional.

II. METODOLOGIA

Este estudo caracteriza-se como uma pesquisa de desenvolvimento tecnológico, de natureza descritiva, conduzida por equipe multiprofissional no Laboratório de Estudos Inventivos em Tecnologia Assistiva (LabEITA). O processo foi estruturado em oito etapas sequenciais (Figura 1), contemplando desde a identificação da demanda clínica até a validação inicial do dispositivo junto ao usuário. O equipamento foi desenvolvido para um participante do projeto "Avaliação e acompanhamento de crianças de 0 a 5 anos com alterações neurológicas e/ou deficiências que necessitavam de tecnologias assistivas", com parecer do comitê de ética da Universidade Federal de Goiás (CEP-UFG: Nº 7.688.714).

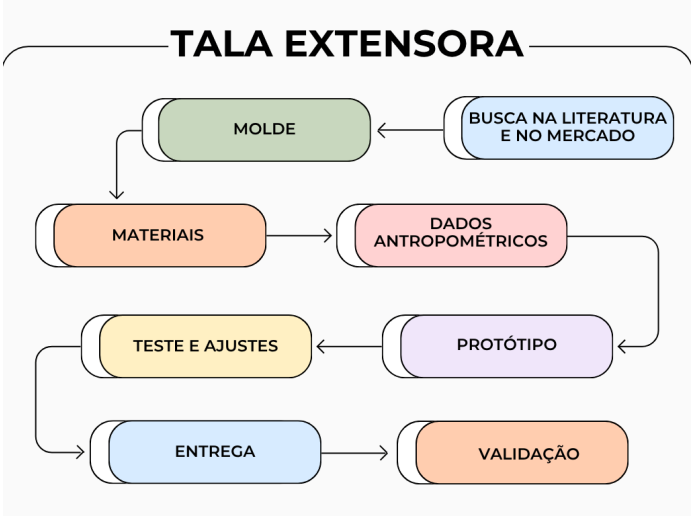


Fig. 1. Fluxograma

Na primeira etapa, a busca na literatura e de mercado ocorreu para entender a funcionalidade da tala extensora no alinhamento postural em dispositivos de ortostatismo, normas para sua produção, materiais utilizados e demanda do mercado (acessibilidade e disponibilidade). Foram utilizadas bases de dados, como: PubMed National Library of Medicine, LILACS Literatura Latino-Americana e do Caribe em Ciências da Saúde, SciELO Scientific Eletronic Library Online e Biblioteca Virtual em Saúde, utilizando os descritores "Tala extensora" e "Extension Splint", com artigos dos últimos 10 anos. Para a busca de mercado, 5 lojas nacionais foram consultadas (ITA Assistiva, Amigo Panda Brasil, Mundo Adaptado e Biofisio Saúde e Reabtecmyhand), escolhidas por conveniência, tendo como critério ser um equipamento sem uso prévio e ter as medidas personalizadas para o usuário.

Para a confecção da tala, os próximos passos foram o desenvolvimento do molde e a escolha dos materiais, que ocorreram por meio da análise de outros equipamentos disponibilizados no mercado e do diálogo com profissionais que desenvolvem esse produto. A partir do molde e materiais definidos, foi realizada a coleta das medidas antropométricas de um participante do projeto "Avaliação e acompanhamento de crianças de 0 a 5 anos com alterações neurológicas e/ou deficiências que necessitavam de tecnologias assistivas", aprovado pelo Comitê de Ética da Universidade Federal de Goiás, para a confecção de um protótipo. Foi realizado durante uma avaliação fisioterapêutica no LabEITA, com a coleta de 3 medidas em pontos anatômicos determinados (trocanter maior a maléolo lateral, circunferência da coxa e circunferência do tornozelo).

Após a coleta desses dados, materiais definidos e molde concluído, houve a produção do protótipo com o maquinário disponível no LabEITA. Sendo utilizado a máquina de costura Facilita Pro 4411 da Singer e a Serra Tico Tico 380 watts da Skil.

Posteriormente houve o teste do protótipo no participante supracitado por uma das fisioterapeuta e delineamento dos ajustes necessários. A avaliação clínica inicial

consistiu na observação sistemática do posicionamento dos membros inferiores durante a utilização do parapodium, verificando-se a manutenção da extensão dos joelhos, o alinhamento do quadril e a estabilidade global do posicionamento ortostático. Embora de caráter observacional, essa avaliação permitiu verificar a adequação funcional inicial do dispositivo antes de sua entrega ao usuário. Como parte da validação, também houve a aplicação de um instrumento de avaliação da percepção dos responsáveis sobre o uso do dispositivo, B-Quest, após 2 meses da data da entrega com o principal cuidador do usuário. Tal questionário analisa 12 tópicos, sendo eles: Dimensões, peso, ajustes, estabilidade/segurança, durabilidade, facilidade de uso. conforto, eficácia, entrega, assistência técnica, serviços profissionais e serviços de acompanhamento, e os 3 itens mais importantes da perspectiva do usuário ou responsável.

III. RESULTADOS

Como resultados da busca na literatura não foram encontrados artigos sobre o uso da tala extensora para a promoção do alinhamento postural em dispositivos de ortostatismo e poucos sobre o uso da tala extensora para qualquer finalidade. A reduzida quantidade de estudos identificados evidencia que a utilização de talas extensoras associadas a dispositivos de ortostatismo permanece pouco explorada na literatura científica, reforçando o caráter inovador da presente proposta. Esse achado demonstra que a principal contribuição do presente estudo não reside apenas no desenvolvimento do dispositivo, mas na proposição de um protocolo estruturado para sua concepção, produção e aplicação clínica em associação ao parapodium. Os trabalhos encontrados possuíam pouca evidência científica e apresentaram resultados positivos para a prevenção de encurtamento e contraturas em pacientes. Durante a busca de mercado, encontrou-se uma diversidade de fornecedores de lojas virtuais com uma média de preço de 359,32, sem a adição da taxa de frete. As medidas eram enviadas de forma online e coletadas pelo próprio consumidor, sem a verificação de algum profissional.

Durante o desenvolvimento do molde e a escolha de materiais, foram analisados 4 extensores de tamanhos e fabricantes diferentes. Esse processo foi importante para entender a progressão de tamanho, por exemplo a quantidade e grossura dos velcros para cada comprimento, e a diferença de materiais utilizados, como o tipo de haste estabilizadora (ferro, aço, alumínio e filamento) e o tecido (neoprene e algodão cru). Também houve o diálogo com profissionais de um centro de reabilitação de Goiânia-GO, classificado como CER IV, que desenvolvem esse equipamento, sendo explicado como o molde era feito e quais materiais utilizaram. Por fim, replicamos o mesmo molde apresentado, como mostra na Fig. 2, e optamos pela utilização do tecido neoprene e tricoline, velcro 50mm, linha de costura e alumínio.

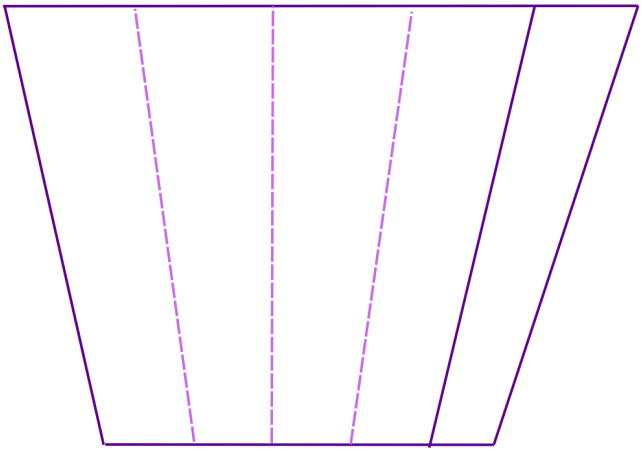


Fig. 2. Molde da tala

O desenvolvimento do dispositivo foi orientado pelo princípio da personalização antropométrica, buscando adequar simultaneamente o posicionamento das hastes estabilizadoras, a distribuição das forças de contenção e o conforto durante o uso prolongado do equipamento.

Para as medidas antropométricas aplicadas no molde, elas constam na tabela a seguir (tabela I).

TABELA I. MEDIDAS ANTROPOMÉTRICAS

MEDIDAS ANTROPOMÉTRICAS	TAMANHO (centímetros)
Trocanter maior do fêmur até maléolo lateral	25
Circunferência da coxa	32
Circunferência do tornozelo	35

A produção do protótipo contou com a costura dos tecidos e o corte de 6 barras de alumínio, exemplo de um protótipo pronto na Fig. 3.



Fig. 3. Protótipo da tala extensora

Durante o teste do dispositivo, as hastes fixas ficaram nos locais idealizados (face posterior, lateral e medial da perna), o velcro central ficou sobre a articulação do joelho e a tala ficou posicionada 3 centímetros abaixo da virilha e 3 centímetros acima do maléolo lateral. Durante a avaliação inicial, o dispositivo apresentou adaptação satisfatória ao usuário, mantendo o posicionamento previamente planejado sem necessidade de modificações estruturais após o primeiro teste funcional.

Na validação, a fisioterapeuta não notou flexão de joelho durante a descarga de peso no parapodium e o alinhamento biomecânico do quadril e membros inferiores estava adequado, assim como aparece na Fig. 4. Além disso, a criança não apresentou feição de dor ou desconforto durante a utilização da tala. Observou-se manutenção da extensão dos joelhos durante a descarga de peso e adequado posicionamento dos membros inferiores no parapodium segundo avaliação fisioterapêutica. Embora essa análise tenha sido predominantemente clínica e observacional, os achados sugerem que o dispositivo cumpriu a finalidade proposta para essa etapa inicial do desenvolvimento.



Fig. 4. Foto do participante no parapodium

Os resultados encontrados no B-Quest após 2 meses de uso constam na tabela II.

TABELA II. Respostas do B-Quest

Itens do B-Quest	Pontuação
Dimensões	5 - Totalmente satisfeito
Peso	5 - Totalmente satisfeito
Ajuste	5 - Totalmente satisfeito
Estabilidade/segurança	5 - Totalmente satisfeito
Durabilidade	5 - Totalmente satisfeito
Facilidade de uso	5 - Totalmente satisfeito
Conforto	5 - Totalmente satisfeito
Eficácia	5 - Totalmente satisfeito
Tempo de espera	5 - Totalmente satisfeito
Assistência técnica	5 - Totalmente satisfeito
Serviços profissionais	5 - Totalmente satisfeito
Serviços de acompanhamento	5 - Totalmente satisfeito
3 itens mais importantes	Durabilidade Conforto Entrega

A tala extensora teve um custo de 58,72 reais, cerca de 16,34% da média de preço encontrada na busca de mercado. O par do equipamento totalizou um valor de 117,44 reais. Os materiais, as quantidades e os valores para o par da tecnologia assistiva constam na tabela III.

TABELA III. Materiais, quantidades e custo de produção da tala extensora

Material	Quantidade	Valor
Neoprene	1 metros quadrados	52,00
Tricoline	1 metros quadrados	12,00
Alumínio (25x2 milímetros)	2 metros	51,94
Linha	-	1,50

Velcro	6,4 metros	14,34
TOTAL		117,44

Os elevados escores obtidos em todos os domínios avaliados indicam elevada aceitabilidade inicial do dispositivo pelos responsáveis. Destaca-se que os três aspectos considerados mais relevantes — durabilidade, conforto e entrega — estão diretamente relacionados às decisões adotadas durante o processo de desenvolvimento e fabricação da tala.

IV. DISCUSSÃO

A literatura é escassa quanto a utilização de tala extensora de membros inferiores associadas a dispositivos que promovam o ortostatismo. No entanto, o seu uso já foi proposto para reduzir o grau de flexão do cotovelo hipertônico em pacientes com doença do neurônio motor superior e demonstrou melhora na condição [13], em paralelo o presente trabalho discute também a capacidade da tala em manter os membros inferiores estendidos em posição ortostática.

As evidências para órteses de controle distal aumentaram a duração da postura em pé em crianças com PC sobre órteses de membros inferiores em crianças com paralisia cerebral [12]. Crianças com paralisia cerebral espástica bilateral apresentaram menor estabilidade postural durante a posição ortostática e o uso de órteses tornozelo-pé podem ser úteis para melhorar a estabilidade [14], Esses achados corroboram o princípio biomecânico adotado no presente estudo, no qual a tala extensora foi utilizada para favorecer o alinhamento dos membros inferiores.

O dispositivo proposto nesse artigo durante a inspeção da utilização da tala foi possível analisar a estabilidade da cadeia cinética durante a descarga de peso, reduzindo a flexão da articulação do joelho e favorecendo o alinhamento do quadril e tornozelo. Dessa forma, a tala atua como um elemento estabilizador durante o uso do parapodium. Além disso, o dispositivo apresentou custo significativamente inferior ao observado nos equipamentos comerciais pesquisados.

A utilização da tala extensora foi acompanhada durante dois meses, nesse período foram realizadas orientações aos responsáveis quanto ao uso adequado do dispositivo e monitoramento da adaptação da criança. Ao final desse período, a aplicação do B-QUEST demonstrou elevada satisfação em todos os domínios avaliados, sugerindo boa aceitabilidade inicial do dispositivo, adequada adaptação ao usuário e potencial para adesão ao uso no contexto domiciliar. Os três itens considerados mais importantes pelos responsáveis foram durabilidade, conforto e entrega, que são aspectos diretamente dependentes do processo de produção e acompanhamento do dispositivo.

Entre as limitações do presente estudo destacam-se a avaliação realizada em um único participante, o curto período de acompanhamento e a ausência de medidas instrumentadas para quantificação objetiva do alinhamento dos membros inferiores durante o ortostatismo. Além disso, não foram realizados ensaios biomecânicos ou comparações com dispositivos comerciais, limitando as conclusões à demonstração da viabilidade técnica e da aceitabilidade inicial da solução proposta.

O principal achado deste estudo foi demonstrar a viabilidade de integrar uma tala extensora personalizada ao parapodium como estratégia para ampliar sua aplicabilidade em crianças que apresentam incapacidade de manter os joelhos em extensão durante o ortostatismo. Mais do que desenvolver um novo dispositivo, o trabalho apresenta uma solução complementar destinada a superar uma limitação frequentemente observada na prática clínica e pouco discutida na literatura especializada.

V. CONCLUSÃO

O presente estudo descreveu o desenvolvimento de uma tala extensora de baixo custo destinada ao uso associado ao parapodium em crianças com paralisia cerebral que apresentam dificuldade em manter os membros inferiores estendidos durante a posição ortostática. O dispositivo apresentou adequada adaptação inicial ao usuário, elevada aceitabilidade pelos responsáveis e potencial para favorecer o alinhamento dos membros inferiores durante o ortostatismo, segundo avaliação clínica observacional. Além do baixo custo de fabricação, a possibilidade de personalização antropométrica representa um diferencial relevante para serviços de reabilitação.

Contudo, estudos futuros deverão incluir amostras maiores, avaliações biomecânicas objetivas e acompanhamento longitudinal, permitindo confirmar a efetividade clínica e estabelecer critérios mais robustos para sua indicação terapêutica.

REFERÊNCIAS

[1] S. Seemüller, F. Beck, A. K. Reimers, “Physical activity of children and adolescents who use a wheelchair: a systematic review.” BMC Public Health. vol.23, Dec 2023. doi:10.1186/s12889-023-17201-6

[2] T. D. O'Brien, J. Noyes, L. H. Spencer, H. P. Kubis, R. P. Hastings, R. Whitaker, “Systematic review of physical activity and exercise interventions to improve health, fitness and well-being of children and young people who use wheelchairs.” BMJ Open Sport Exerc Med., vol 2, Nov. 2016. doi:10.1136/bmjsem-2016-000109

[3] World Health Organization. “Guidelines on the provision of manual wheelchairs in less resourced settings.”, Geneva: World Health Organization, Mai 2008. .

[4] S. G. Paguinto, N. A. Kasparian, P. Bray, M. Farrar, "Multidisciplinary perspectives and practices in wheelchair prescription for children with neuromuscular diseases," Disabil. Rehabil. Assist. Technol., vol. 18, pp. 166-174, Fev. 2023, doi: 10.1080/17483107.2020.1839793.

[5] N. Bray, J. Noyes, R. T. Edwards, N. Harris, "Wheelchair interventions, services and provision for disabled children: a mixed-method systematic review and

conceptual framework," *BMC Health Serv. Res.*, vol. 14, p. 309, Jul. 2014, doi: 10.1186/1472-6963-14-309.

[6] World Health Organization, “World report on disability”. Geneva, Switzerland: World Health Organization, Dec 2011. Available: https://www.who.int/publications/i/item/9789241564182.

[7] J. Srishti, B. Avijit, H. Gita, “Patterns and unmet needs in wheelchair utilization: A mixed-method study of user experience at a tertiary hospital in India”, Journal of Rehabilitation and Assistive Technologies Engineering, vol. 13, pp. 1-13, 2026

[8] Brasil, Lei n.º 13.146, de 6 de julho de 2015: Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência), Diário Oficial da União, Brasília, DF, Jul. 7, 2015.

[9] P. A. L. S. Bastos, M. S. Silva, N. M. Ribeiro, R. S. Mota, T. G. Filho, “Tecnologia assistiva e políticas públicas no Brasil.” Cadernos Brasileiros de Terapia Ocupacional, vol. 31, Jan. 2023 https://doi.org/10.1590/2526-8910.ctoAO260434011
1016/j. Ejpn.2011.07.001

[10] R. S. Curi, E. C. Ferretti, R. C. Franco, A. L. C. Allegretti, M. S. P. da Silva, "Satisfação de usuários de cadeiras de rodas fornecidas pelo SUS: estudo transversal na Baixada Santista," Rev. Bras. Saúde Func., vol. 8, no. 2, pp. 45-53, 2024.

[11] World Health Organization, United Nations Children's Fund. *Global report on assistive technology*. Geneva: WHO; 2022.

[12] S. J. Briley, R. J. K. Vegter, V. L. Goosey-Tolfrey, B. S. Mason, “Scapular kinematic variability during wheelchair propulsion is associated with shoulder pain in wheelchair users.” J Biomech., vol, 113, Dec, 2020. doi: 10.1016/j.jbiomech.2020.110099.

[13] Briley SJ, Vegter RJK, Goosey-Tolfrey VL, Mason BS. The longitudinal relationship between shoulder pain and altered wheelchair propulsion biomechanics of manual wheelchair users. J Biomech. 2021 Sep 20;126:110626. doi: 10.1016/j.jbiomech.2021.110626. Epub 2021 Jul 12. PMID: 34329882.

[14] D. K. A. Cruz, A. A. da Nóbrega, V. O. de M. Pereira, and M. de M. S. Montenegro, “Os Objetivos de Desenvolvimento Sustentável e as fontes de dados para o monitoramento das metas no Brasil," Epidemiol. Serv. Saúde, vol. 31, 2022, doi: 10.1590/S2237-9622202200010.

Desenvolvimento de órteses suropodálicas pediátricas por fluxo digital integrado utilizando escaneamento 3D, modelagem CAD e manufatura aditiva

Giovanna Silva Rodrigues Manussakis
Faculdade de Artes Visuais
Universidade Federal de Goiás
Goiânia, Brasil

Pedro Henrique Gonçalves
Faculdade de Artes Visuais
Universidade Federal de Goiás
Goiânia, Brasil

Geovane Umbelino Marques
Universidade Federal de Goiás
Goiânia, Brasil

Abstract—A órtese suropodálica (AFO — Ankle-Foot Orthosis) é um dispositivo exoesquelético destinado ao suporte do complexo tornozelo-pé, indicado no tratamento de deformidades, distúrbios neuromusculares e alterações da marcha. No contexto pediátrico, sua relevância clínica é acentuada: crianças com condições como paralisia cerebral, mielomeningocele e sequelas de AVC perinatal constituem um dos grupos de maior demanda por esse dispositivo no Sistema Único de Saúde. Entre as principais indicações estão deformidades como pé caído, pé equino, pé varo, pé valgo e suas combinações — condições que comprometem o alinhamento biomecânico, a estabilidade e a eficiência da marcha, impactando diretamente a autonomia e o desenvolvimento motor infantil. Contudo, os métodos convencionais de fabricação — baseados em moldagem em gesso e conformação manual de termoplásticos — apresentam limitações documentadas em personalização anatômica, custo, prazo de entrega e adequação ao crescimento infantil contínuo. Este trabalho investigou a viabilidade do uso integrado de escaneamento tridimensional, modelagem CAD e impressão 3D por Fused Deposition Modeling (FDM) na produção de órteses AFO personalizadas para crianças de 1 a 10 anos. A metodologia adotada foi o Design Thinking. O escaneamento foi realizado com o scanner RANGE 2 (Revopoint); a modelagem, no software Fusion 360, com espessura uniforme de 2 mm; e a impressão, na Creality K1 Max utilizando polipropileno (PP), com camadas de 0,2 mm e velocidade de 200 mm/s. Os resultados indicam que o fluxo digital reduz significativamente o número de etapas manuais, minimiza o desperdício de material e permite ajustes rápidos sem reinício do processo, favorecendo elevada fidelidade geométrica e potencial para maior padronização do processo quando comparado ao método convencional. Conclui-se que a manufatura aditiva representa uma alternativa tecnicamente viável, economicamente acessível e clinicamente promissora para a produção de órteses AFO pediátricas.

Palavras-chave—órtese suropodálica; impressão 3D; manufatura aditiva; tecnologia assistiva; design e saúde.

I. I. INTRODUÇÃO

A origem do termo órtese vem da palavra grega "Orthos", que significa direto, reto, normal. A órtese é definida como um dispositivo exoesquelético que, aplicado a um ou vários segmentos do corpo, tem a finalidade de proporcionar o melhor alinhamento possível, priorizando a funcionalidade de cada indivíduo [1].

As órteses suropodálicas — também denominadas Ankle-Foot Orthosis (AFO) — são dispositivos para o complexo articular de tornozelo e pé. Esse modelo divide-se em fixas, nas quais não é permitido o movimento articular do tornozelo, e articuladas, que permitem o movimento de dorsiflexão plantar [2]. Suas funções principais incluem a prevenção de deformidades em equino, controle do alinhamento do pé e do tornozelo, equilíbrio e melhora do desempenho na marcha [3].

No contexto pediátrico, o uso da órtese impacta diretamente o desempenho de atividades de vida diária. Crianças com condições como paralisia cerebral, mielomeningocele e sequelas de AVC perinatal constituem um dos grupos de maior demanda por órteses do tipo AFO no Sistema Único de Saúde [1]. Contudo, os métodos convencionais de fabricação apresentam limitações bem documentadas: longos prazos de entrega, dificuldade de ajuste anatômico e custo elevado — barreiras que comprometem especialmente pacientes em situação de vulnerabilidade socioeconômica [3].

No contexto pediátrico, essas limitações são agravadas pelo crescimento contínuo da criança, que exige substituições periódicas do dispositivo. Aspectos como conforto, peso e aceitação estética influenciam diretamente a adesão infantil ao tratamento, tornando a personalização não apenas desejável, mas clinicamente necessária [4].

Nesse cenário, a manufatura aditiva — em especial a impressão 3D associada ao escaneamento tridimensional e à modelagem CAD — surge como alternativa promissora. Wixted et al. [5] reforçam que essa abordagem apresenta resultados comparáveis aos métodos tradicionais em termos de resistência e funcionalidade, com vantagens em personalização e tempo de entrega. Bezek et al. [6] apontam que o desenvolvimento contínuo de novos materiais biocompatíveis para impressão 3D amplia as possibilidades de aplicação clínica desses dispositivos. Embora a literatura demonstra resultados promissores para a impressão 3D de órteses, ainda existem lacunas relacionadas à sistematização de fluxos digitais completos aplicados à ortopedia pediátrica. Nesse contexto, o presente estudo propõe a integração entre escaneamento tridimensional, modelagem paramétrica e manufatura aditiva como estratégia para otimizar o desenvolvimento de órteses suropodálicas personalizadas.

II. II. REVISÃO BIBLIOGRÁFICA

A. A. Biomecânica de Membros Inferiores

A biomecânica é a área de estudo responsável pela análise da estrutura e da função dos sistemas biológicos por meio dos princípios da mecânica. O termo deriva das palavras gregas *bio* (vida) e *mecânica* (movimento), sendo aplicada para compreender como o corpo humano reage às forças internas e externas durante a realização dos movimentos [7]. Segundo Hamill e Knutzen (1999), a biomecânica pode ser qualitativa quando se limita à descrição do movimento, ou quantitativa quando envolve medidas e cálculos para sua avaliação.

Segundo Brasil (2014) [1], os membros inferiores desempenham papel essencial na sustentação do peso corporal, na manutenção do equilíbrio e na realização da marcha. Sua estrutura é composta pela cintura pélvica, coxa, perna e pé, formando um sistema biomecânico capaz de absorver impactos, distribuir cargas e gerar propulsão durante o deslocamento. O funcionamento adequado desse sistema depende da interação harmoniosa entre os segmentos corporais e do controle neuromuscular responsável pela coordenação dos movimentos [13].

A análise biomecânica permite compreender como os músculos geram força, como os ossos suportam e absorvem cargas e como as articulações realizam seus movimentos. Essas informações contribuem diretamente para a avaliação da segurança de dispositivos ortopédicos, para a melhoria do desempenho físico e para o desenvolvimento de produtos mais eficientes, ergonômicos e funcionais [7].

No contexto dos membros inferiores, os pés desempenham papel central na sustentação, estabilidade e mobilidade do corpo humano, exercendo funções como absorção de impactos, adaptação às irregularidades do terreno, distribuição das cargas corporais e geração de força propulsiva. Além disso, o complexo tornozelo-pé atua como importante mecanismo de estabilidade, contribuindo para o equilíbrio e para a eficiência dos movimentos realizados durante a locomoção [14].

Os tornozelos e pés constituem um complexo sistema biomecânico capaz de executar movimentos como flexão plantar, dorsiflexão, inversão e eversão, permitindo adaptação às diferentes superfícies e absorção das forças geradas durante o ciclo da marcha. Alterações nesse sistema, sejam de origem neurológica, musculoesquelética ou traumática, podem comprometer significativamente a mobilidade e o equilíbrio do indivíduo, justificando a necessidade de intervenção por meio de dispositivos de suporte como as órteses AFO [1].

Diante disso, compreender a biomecânica dos membros inferiores torna-se fundamental para o desenvolvimento de órteses suropodálicas, pois esses dispositivos atuam diretamente na estabilização do tornozelo e do pé, auxiliando no alinhamento biomecânico, no controle dos movimentos e na melhora da funcionalidade durante a marcha. O conhecimento anatômico e biomecânico contribui para o desenvolvimento de soluções mais adequadas às necessidades dos usuários, especialmente quando associado às possibilidades de personalização proporcionadas pela modelagem e impressão 3D [8].

B. B. Estrutura Anatômica do Tornozelo e Pé

O corpo humano é dividido em membros responsáveis pelos movimentos e pela locomoção, sendo classificados em superiores e inferiores. Entre os componentes dos membros inferiores, encontram-se o pé e o tornozelo, estruturas estudadas neste trabalho de forma mais aprofundada devido à sua relação direta com a necessidade de utilização das órteses suropodálicas. De modo geral, o membro inferior é formado pela cintura, pela coxa, perna e pé [1].

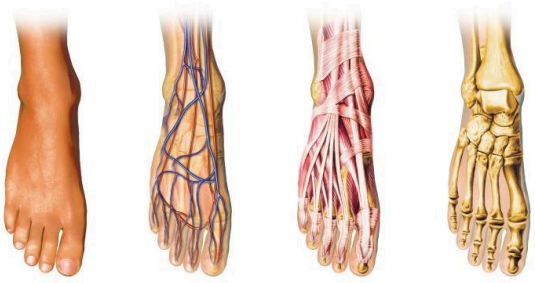


Figura 1 - A Complexa Anatomia do Pé
Fonte: Dr. Tiago Baumfeld

O complexo tornozelo-pé constitui uma das estruturas mais importantes do sistema musculoesquelético humano, sendo responsável pela sustentação do peso corporal, manutenção do equilíbrio, absorção de impactos e execução dos movimentos necessários para a locomoção. Sua composição anatômica envolve ossos, articulações, músculos, tendões e ligamentos que atuam de forma integrada para proporcionar estabilidade e mobilidade durante as atividades diárias, especialmente durante o ciclo da marcha [1].

Os membros inferiores são formados pela cintura pélvica, coxa, perna e pé, sendo o tornozelo e o pé estruturas fundamentais para o deslocamento humano. Devido à sua participação direta nos movimentos de marcha e equilíbrio, alterações funcionais nessa região podem comprometer significativamente a mobilidade do indivíduo, tornando necessária, em diversos casos, a utilização de órteses suropodálicas para auxiliar na estabilização e no controle dos movimentos articulares [1].

Do ponto de vista ósseo, o pé apresenta uma estrutura complexa composta por 26 ossos organizados em três grupos

principais: tarso, metatarso e falanges (Figura 2). Os ossos do tarso incluem o tálus, calcâneo, navicular, cubóide e os três ossos cuneiformes. O tálus destaca-se por sua participação direta na articulação do tornozelo, recebendo e distribuindo as cargas provenientes da tíbia durante a locomoção. O calcâneo, por sua vez, constitui o maior osso do pé e atua como importante estrutura de suporte e absorção de impactos [1].

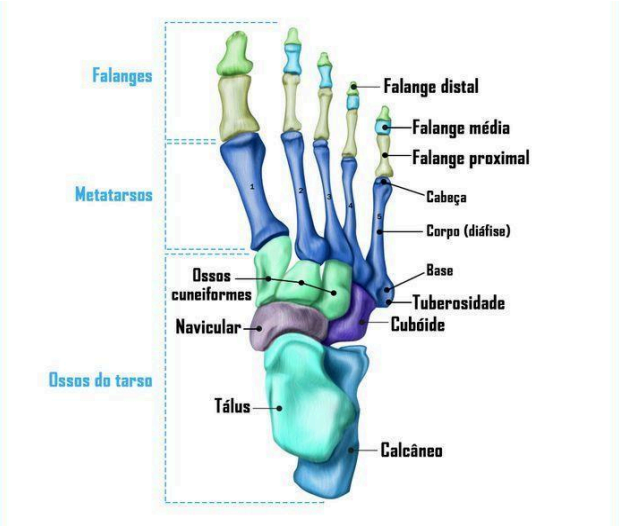


Figura 2 - Ossos do Pé
Fonte: Info.enf

C. C. Deformidades do Pé

As principais deformidades que indicam o uso de AFO incluem: (a) Síndrome do Pé Caído — fraqueza dos músculos dorsiflexores, com arrasto dos dedos na marcha, frequentemente associada a lesões do nervo fibular ou doenças neurológicas; (b) Pé Equino — flexão plantar excessiva com dificuldade de apoio do calcanhar, associado à espasticidade; (c) Pé Calcâneo — dorsiflexão excessiva com comprometimento da propulsão; (d) Pé Varo/Valgo — desvios laterais do retropé com instabilidade; (e) deformidades combinadas, como equinovaro e equinovalgo, frequentes em comprometimentos neurológicos [1].

Essas deformidades podem provocar alterações significativas na mobilidade, no equilíbrio e na capacidade funcional dos indivíduos. Dessa forma, a utilização de órteses suropodálicas torna-se uma importante estratégia terapêutica para promover alinhamento biomecânico adequado, auxiliar no controle dos movimentos articulares e melhorar o desempenho durante a marcha. A escolha do tipo de órtese depende das características da deformidade, dos objetivos terapêuticos e das necessidades específicas [1].

D. D. Classificação das Órteses AFO

As AFOs classificam-se em três categorias: (1) Não articuladas rígidas — mantêm o tornozelo fixo a 90°, indicadas em casos de fraqueza muscular severa ou instabilidade articular; (2) Não articuladas flexíveis — permitem algum grau de movimento articular, modelos Codivilla, Espiral e Semi-espiral; (3) Articuladas — indicadas para pacientes com capacidade de realizar marcha, alinhadas com os maléolos para evitar rotação indesejada durante a transferência de peso [1], [2].

Dessa forma, a órtese é recomendada após uma avaliação detalhada da biomecânica do paciente, tanto em movimento quanto de forma estática, para analisar as características do alinhamento e funcionamento dos pés. Em crianças o uso da órtese impacta diretamente no desempenho de atividades de vida diária, e pode orientar o processo de avaliação e intervenção dos profissionais.

E. E. Métodos de Produção e Impressão 3D

O processo convencional de produção de AFOs é composto por dez etapas manuais, envolvendo moldagem em gesso, conformação de termoplástico, recorte e acabamento. Esse processo é demorado, sujeito a variações e gera significativo desperdício de material [1].

Em contraste com o método tradicional, os avanços tecnológicos têm possibilitado a utilização de fluxos de trabalho digitais para a fabricação de órteses suropodálicas, especialmente por meio do escaneamento tridimensional e da impressão 3D. Nesse processo, a obtenção das características anatômicas do paciente é realizada por meio de scanners 3D, que capturam digitalmente a geometria do membro inferior, eliminando a necessidade da confecção de moldes físicos em gesso [8].

Após a aquisição das imagens, os dados são processados em softwares de modelagem assistida por computador (CAD), permitindo a realização de ajustes virtuais, correções de alinhamento e personalização do dispositivo de acordo com as necessidades clínicas do usuário. Em seguida, o modelo digital é convertido para um formato compatível com a manufatura aditiva, sendo produzido camada por camada por meio de impressoras 3D [8].

As principais diferenças em relação ao método tradicional concentram-se na digitalização das etapas de obtenção do molde e de modelagem da órtese. Enquanto a técnica convencional depende de processos manuais, frequentemente mais demorados e suscetíveis a variações decorrentes da manipulação dos materiais, o fluxo digital proporciona maior precisão geométrica, reprodutibilidade e possibilidade de armazenamento permanente dos modelos em formato digital. Além disso, a impressão 3D permite a produção de estruturas complexas e personalizadas com menor desperdício de material, favorecendo a otimização do processo produtivo.

De acordo com Bezek (2025) [6], a tecnologia de impressão 3D tem revolucionado a área da saúde devido à sua capacidade de produzir estruturas complexas e personalizadas

para cada paciente. Além disso, os avanços nas tecnologias de imagem digital e nos sistemas de design e manufatura assistidos por computador (CAD/CAM) têm simplificado os processos de projeto e fabricação, reduzindo o tempo de produção e aumentando o conforto, a funcionalidade e a adaptação dos dispositivos. Segundo o autor, o desenvolvimento contínuo de novos materiais e técnicas de impressão favorece o surgimento de soluções médicas cada vez mais inovadoras e personalizadas, consolidando a impressão 3D como uma ferramenta de grande relevância para a saúde moderna.

III. III. OBJETIVOS

Desenvolver um fluxo digital integrado para produção de órteses suropodálicas pediátricas utilizando escaneamento tridimensional, modelagem CAD e manufatura aditiva.

Os objetivos específicos são: (a) averiguar a precisão do ajuste anatômico, o conforto ao usuário e o desempenho mecânico da estrutura impressa; (b) validar a viabilidade da manufatura aditiva na produção de dispositivos ortopédicos personalizados para o público infantil.

IV. IV. METODOLOGIA

F. *A. Design Thinking*

A metodologia adotada foi o Design Thinking, abordagem centrada no ser humano voltada para a resolução criativa de problemas [9]. A escolha justifica-se pela necessidade de compreender não apenas os aspectos técnicos da confecção de órteses AFO, mas também as necessidades ergonômicas, funcionais e estéticas dos usuários. O processo foi estruturado em cinco etapas: imersão, definição, ideação, prototipagem e testes.

Na etapa de imersão, foram realizadas pesquisas bibliográficas e análise de similares. Na definição, identificaram-se os principais requisitos projetuais: adaptação anatômica, leveza, resistência, flexibilidade controlada e viabilidade de fabricação por impressão 3D. Na ideação, realizou-se o escaneamento do membro e a modelagem digital. Na prototipagem, executou-se a impressão. Na etapa de testes, avaliou-se o desempenho funcional e estrutural do protótipo.

G. *B. Parâmetros Projetuais*

A definição dos critérios projetuais constitui uma etapa fundamental para o desenvolvimento da órtese, uma vez que orientou as decisões relacionadas aos aspectos funcionais, ergonômicos, estruturais e estéticos do produto. Essas diretrizes foram estabelecidas com base nas informações obtidas durante as etapas de pesquisa bibliográfica, análise de similares, e compreensão das necessidades do usuário levantadas na fase de imersão do Design Thinking.

A adaptação anatômica foi definida como requisito prioritário, considerando a necessidade de encaixe preciso ao membro do paciente. Para isso, o escaneamento tridimensional foi utilizado como ferramenta de aquisição das dimensões e curvaturas da região tornozelo-pé, possibilitando a geração de um modelo digital individualizado e garantindo maior fidelidade geométrica na produção da órtese.

Os requisitos de ergonomia e conforto foram definidos a partir dos relatos de usuários coletados na imersão, que apontaram pressão excessiva em proeminências ósseas, atrito prolongado e baixa ventilação como principais queixas dos dispositivos convencionais. Foram priorizados distribuição uniforme de pressão, aberturas na estrutura para ventilação e redução do peso total. A espessura uniforme de 2 mm, definida durante a modelagem no Fusion 360, foi adotada com base em referências da literatura que indicam essa faixa como adequada para órteses FDM (Modelagem por Deposição Fundida) em termoplásticos de uso ortopédico [5].

A viabilidade de fabricação por manufatura aditiva constituiu outro critério relevante do projeto. Foram avaliados fatores como tempo de impressão, consumo de material, possibilidade de ajustes rápidos no modelo digital e facilidade de reprodução. Arce e Foggiatto (2017) destacam que o fluxo digital de produção reduz significativamente o número de etapas manuais em relação ao método convencional, diminuindo custos operacionais e o tempo entre a prescrição e a entrega do dispositivo. Por fim, os critérios definidos serviram como diretrizes para a prototipagem, contribuindo para a criação de um produto funcional, personalizado e alinhado às necessidades clínicas do usuário [8].

H. *C. Escaneamento Tridimensional*

O escaneamento foi realizado com o scanner RANGE 2 da Revopoint, que capturou digitalmente a geometria do membro inferior, eliminando a necessidade de moldes físicos. O membro foi posicionado adequadamente para garantir a captura completa das regiões de interesse. Neste estudo, foram desenvolvidas órteses personalizadas para crianças de até 10 anos.

I. *D. Modelagem 3D (CAD)*

Para o desenvolvimento do modelo digital, foi utilizado o software Fusion 360 (Autodesk). O processo de modelagem foi realizado na aba de superfícies do software, por meio da criação de uma forma orgânica delimitada à região anatômica de aplicação da órtese. O modelo foi ajustado à geometria do membro escaneado para garantir o correto encaixe na região do calcanhar. Após a definição da geometria externa, a estrutura foi submetida ao processo de extrusão com espessura uniforme de 2 mm, seguido de arredondamento de bordas e inserção de furos para fixação. Essas características foram consideradas requisitos de projeto e não foram submetidas, nesta etapa, a testes de conforto, pressão ou desempenho funcional.

A Figura 3 apresenta o modelo tridimensional obtido a partir do escaneamento, posicionado a 90°, bem como o resultado final da órtese após a conclusão da etapa de modelagem.

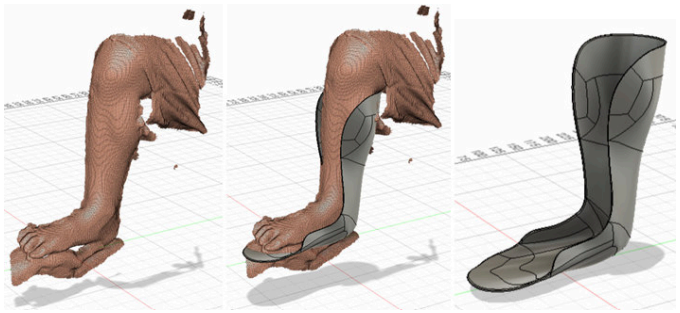


Figura 3 - Modelagem da órtese

Fonte: Elaborado pelo autor

J. E. Prototipagem (Impressão 3D)

A etapa de prototipagem corresponde à materialização do modelo desenvolvido durante a fase de modelagem tridimensional. Para a fabricação dos protótipos, foi utilizada a tecnologia de manufatura aditiva denominada Fused Deposition Modeling (FDM), a qual possibilita a construção de objetos por meio da deposição sucessiva de camadas de material termoplástico. A escolha dessa tecnologia foi motivada por sua viabilidade econômica, facilidade de acesso e ampla aplicabilidade em projetos experimentais voltados à área da tecnologia assistiva. Portanto, para dar-se início ao processo, o modelo tridimensional da órtese foi convertido para um formato compatível com os processos de manufatura aditiva, sendo exportado em formato STL. Em seguida, o arquivo foi importado para um software de fatiamento, no qual foram definidos os parâmetros de impressão, incluindo orientação da peça, altura das camadas, percentual de preenchimento, velocidade de impressão e estruturas de suporte, quando necessárias, especificado na tabela 1.

TABELA I

Parâmetros de Impressão 3D

Parâmetro	Valor
Material	Polipropileno (PP)
Preenchimento	100%
Espessura da Camada	0,2 mm
Temperatura do Bico	240 °C
Temp. Mesa Impressão	100 °C
Velocidade de Impressão	200 mm/s
Diâmetro do Bico	0,8 mm

A impressora utilizada foi a Creality K1 Max. Durante a impressão, foram monitorados a adesão da peça à mesa, a estabilidade do processo e a qualidade da deposição do material. Após a fabricação, a órtese foi submetida a pós-processamento com remoção de suportes, lixamento de bordas e inspeção visual de conformidade dimensional.

V. V. RESULTADOS ESPERADOS

O fluxo digital de produção — escaneamento, modelagem CAD e impressão FDM — Os resultados obtidos indicam que o fluxo digital apresenta potencial para substituir diversas etapas manuais do processo convencional, simplificando o desenvolvimento da órtese e reduzindo a dependência de procedimentos artesanais. A principal vantagem observada foi a eliminação da etapa de moldagem em gesso, substituída pelo escaneamento tridimensional, que garante maior fidelidade geométrica à anatomia do paciente. Além da redução do número de etapas, essa substituição elimina fontes de erro associadas ao processo manual, como deformações do gesso durante a secagem, dependência da habilidade individual do técnico ortesista e a impossibilidade de armazenar digitalmente o modelo para reproduções futuras.

O arquivo digital gerado a partir do escaneamento pode ser reaproveitado para ajustes, réplicas ou reimpressões em caso de perda ou dano da órtese, algo inviável no fluxo convencional, em que cada molde de gesso é descartado após o uso.

O uso do polipropileno (PP) como material de impressão mostrou-se adequado para a aplicação ortopédica pediátrica, dado seu equilíbrio entre resistência mecânica, leveza e flexibilidade controlada. Esse equilíbrio é particularmente relevante no contexto pediátrico, no qual a órtese precisa acompanhar o crescimento e a movimentação natural da criança sem gerar pontos de rigidez excessiva que possam causar desconforto ou lesões por atrito. A espessura de 2 mm, baseada em literatura específica para órteses FDM [5], conferiu rigidez suficiente para suporte do tornozelo sem comprometer a mobilidade funcional, representando um ponto de equilíbrio entre a necessidade de imobilização parcial e a preservação da amplitude de movimento residual, essencial para o desenvolvimento motor infantil.

Os resultados preliminares indicam que a manufatura aditiva representa uma alternativa viável, capaz de produzir órteses personalizadas com qualidade comparável aos métodos convencionais, em menor tempo e com menor custo operacional [3], [5]. Vale destacar que essa viabilidade não se limita ao aspecto técnico, mas também abrange a dimensão logística do processo: a digitalização do fluxo de produção permite o armazenamento em nuvem dos modelos, a padronização de parâmetros de impressão e a possibilidade de produção remota, características que podem ser especialmente relevantes para ampliar o acesso a órteses personalizadas em regiões com escassez de profissionais especializados em ortopedia técnica.

VI. VI. DISCUSSÃO

A principal contribuição deste estudo não reside apenas na utilização da impressão 3D para fabricação de órteses, mas na proposição de um fluxo digital integrado que conecta escaneamento tridimensional, modelagem CAD, definição de critérios projetuais e manufatura aditiva. Essa integração reduz a dependência de etapas artesanais e amplia o potencial de padronização e reprodutibilidade do processo de desenvolvimento. Os resultados obtidos estão alinhados com a

literatura sobre manufatura aditiva aplicada à ortótica. Wixted et al. [5] identificaram que órteses AFO produzidas por impressão 3D apresentam resistência e funcionalidade comparáveis às convencionais, com vantagens claras em personalização e tempo de entrega. Essa convergência com os achados do presente trabalho reforça a hipótese de que o fluxo digital de produção não representa apenas uma alternativa tecnológica, mas um processo capaz de igualar — e em determinados aspectos superar — o desempenho funcional das órteses convencionais, sobretudo quando se considera a capacidade de adequação geométrica individualizada proporcionada pelo escaneamento tridimensional. Chen et al. [10] reforçam que a fabricação aditiva reduz significativamente o desperdício de material em relação aos métodos tradicionais, tornando o processo mais sustentável, uma vez que a deposição de material ocorre de forma aditiva e controlada, ao contrário dos processos subtrativos ou de conformação térmica, que frequentemente geram excedentes de material descartados após o ajuste da peça ao paciente. Essa característica dialoga diretamente com a proposta deste trabalho, que buscou aliar viabilidade técnica à racionalização de recursos no processo produtivo.

No contexto pediátrico, a contribuição do fluxo digital é especialmente relevante. A capacidade de armazenar o modelo digital e reimprimí-lo rapidamente responde diretamente à demanda por substituições frequentes decorrentes do crescimento infantil [4], eliminando a necessidade de repetição integral do processo de moldagem a cada nova adequação. Esse aspecto adquire particular importância no público pediátrico, cujo crescimento ósseo acelerado exige trocas periódicas de dispositivo, tornando o custo e o tempo de produção fatores críticos para a continuidade do tratamento. A possibilidade de ajustar parâmetros no arquivo CAD original — em vez de reiniciar todo o processo de moldagem — representa um ganho direto de eficiência clínica e operacional. Bezek et al. [6] apontam que os avanços contínuos em materiais biocompatíveis para impressão 3D ampliam progressivamente as possibilidades de aplicação clínica desses dispositivos, sugerindo que a tendência observada neste trabalho — de substituição de materiais convencionais por polímeros impressos — tende a se consolidar e se expandir à medida que novos materiais com propriedades mecânicas e biológicas aprimoradas se tornem disponíveis comercialmente.

A metodologia do Design Thinking mostrou-se adequada para integrar as perspectivas clínicas, ergonômicas e técnicas no desenvolvimento da órtese. Essa integração se mostrou extremamente relevante no projeto, nos quais decisões puramente técnicas — como escolha de material ou espessura de parede — não podem ser dissociadas da experiência prática do usuário. A etapa de imersão foi fundamental para identificar as principais queixas dos usuários e traduzi-las em requisitos projetuais concretos, evidenciando a importância da abordagem centrada no usuário em projetos de tecnologia assistiva. Nesse sentido, a articulação entre método de projeto centrado no usuário e processo de fabricação digital sugere um caminho promissor para o desenvolvimento de dispositivos ortopédicos personalizados, no qual a flexibilidade tecnológica da manufatura aditiva se soma à sensibilidade metodológica do Design Thinking para gerar soluções tecnicamente viáveis e clinicamente relevantes.

Diante disso, a pesquisa resultou no desenvolvimento e fabricação de órteses suropodálicas (AFO) personalizadas. Após a impressão do dispositivo, a peça passou por etapas de acabamento, incluindo aplicação de velcro para fixação e EVA em pontos de pressão, visando maior conforto durante o uso.



Os critérios para avaliação foram definidos em conjunto com profissionais da fisioterapia, dessa forma, foram considerados, ajuste anatômico, conforto, fixação e estabilidade.





A avaliação permitiu verificar a funcionalidade do protótipo e sua adaptação às necessidades do paciente, caracterizando o uso bem sucedido do dispositivo. Dessa forma, os resultados reforçam o potencial da impressão tridimensional como alternativa para a produção de órteses AFO personalizadas.

VII. VII. CONCLUSÃO

Este trabalho demonstrou que o uso integrado de escaneamento tridimensional, modelagem CAD e impressão 3D por FDM representa uma alternativa tecnicamente viável,

economicamente acessível e clinicamente promissora para a produção de órteses AFO pediátricas. O fluxo digital reduz o número de etapas manuais, minimiza o desperdício de material e permite ajustes rápidos sem reinício do processo, com precisão geométrica superior ao método convencional. Esses resultados, ainda que preliminares, indicam que a substituição do processo tradicional de moldagem em gesso por um fluxo inteiramente digital não compromete — e em diversos aspectos aprimora — os requisitos técnicos e funcionais esperados de um dispositivo ortopédico, ao mesmo tempo em que introduz ganhos operacionais dificilmente alcançáveis pelos métodos convencionais, como a possibilidade de reprodução rápida do modelo e o armazenamento digital do histórico de cada paciente.

A convergência entre a demanda clínica concreta — crianças que necessitam de órteses personalizadas e acessíveis — e uma tecnologia em consolidação justifica e motiva a continuidade desta pesquisa. Essa convergência é particularmente significativa no cenário público de saúde, no qual o acesso a órteses personalizadas é frequentemente limitado por custos elevados, longos prazos de confecção e escassez de profissionais especializados em técnicas convencionais de moldagem. A proposta aqui investigada sugere um caminho concreto para mitigar essas barreiras, ao aproximar tecnologias de manufatura digital — cada vez mais acessíveis em termos de custo de equipamento e material — das necessidades específicas da ortopedia pediátrica.

Por fim, espera-se que a consolidação desse fluxo digital contribua para ampliar o acesso a órteses personalizadas, reduzindo barreiras relacionadas ao tempo de fabricação, à repetibilidade do processo e à dependência de técnicas exclusivamente artesanais, fortalecendo a integração entre Engenharia, Design e Tecnologia Assistiva no desenvolvimento de soluções voltadas à reabilitação pediátrica.

AGRADECIMENTOS

Os autores agradecem à Universidade XXX, ao Laboratório YYY e à Fundação de Apoio à Pesquisa (FUNAPE) pelo suporte institucional no desenvolvimento deste trabalho.

REFERÊNCIAS

[1] BRASIL. Ministério da Saúde. Técnico em Órteses e Próteses: livro-texto. Brasília: Ministério da Saúde, 2014.

[2] CASTANEDA, L. Prescrição, Concessão, Adaptação e Manutenção de Órteses, Próteses e Meios Auxiliares de Locomoção. UNA-SUS; UFMA, 2021.

[3] WOJCIECHOWSKI, E. et al. Feasibility of designing, manufacturing and delivering 3D printed ankle-foot orthoses: a systematic review. Journal of Foot and Ankle Research, v. 12, p. 11, 2019.

[4] KAYE, R. et al. Three-dimensional printing: a review of its distinct applications in otolaryngology. European Archives of Oto-Rhino-Laryngology, v. 274, n. 11, p. 3897–3905, 2017.

[5] WIXTED, C. M. et al. A review of 3D printing of orthopedic devices. 3D Printing and Additive Manufacturing, v. 8, n. 5, 2021.

[6] BĘŻEK, A. et al. Advances in 3D printed orthotics for rehabilitation. Wiadomości Lekarskie Medical Advances, v. 78, n. 3, p. 537–542, 2025.

[7] HALL, S. J. Biomecânica básica. 7. ed. Rio de Janeiro: Guanabara Koogan, 2016.

[8] ARCE, R. P.; FOGGIATTO, J. A. Modelagem de Órteses para Fabricação por Manufatura Aditiva. 9º Congresso Brasileiro de Engenharia de Fabricação, 2017.

[9] BROWN, T. Change by Design. New York: HarperCollins, 2009.

[10] CHEN, R. K. et al. Additive manufacturing of custom orthoses and prostheses — a review. Additive Manufacturing, v. 12, p. 77–89, 2016.

[11] MAVROIDIS, C. et al. Patient specific ankle-foot orthoses using rapid prototyping. Journal of NeuroEngineering and Rehabilitation, v. 8, p. 1, 2011.

[12] CURY, V. C. R. et al. Efeitos do uso de órtese na mobilidade funcional de crianças com paralisia cerebral. Revista Brasileira de Fisioterapia, v. 10, n. 1, p. 67–74, 2006

[13] HAMILL, Joseph; KNUTZEN, Kathleen M. Bases biomecânicas do movimento humano. São Paulo: Manole, 1999.

[14] BROCKETT, Claire L.; CHAPMAN, Graham J. Biomecânica do tornozelo. Ortopedia e Trauma, v. 30, n. 3, p. 232-238, 2016.

Design and Specification of a Local Learning Environment for Neuro-Robotic Educational Applications

Nivaldo Antonio Portela de Vasconcelos[1], Aquiles Medeiros Filgueira Burlamaqui[2], Justo Emilio Alvarez Jácobo[1], Augusto Baffa[3], José Valentim dos Santos Filho[4], Alan Dionisio Peixoto Nogueira[2], Marcos Paulo Dantas Joaquim[2], Rinaldo Tavares da Silva Filho[2], Ana Alice da Silva Câmara[5], André Ferrão de Oliveira[1], Gabriel Arcanjo de Souza[1], Ítallo Antônio dos Santos Melo[1], Juliana de Albuquerque Souza Costa[1], Lucas Alves Barbosa[1], Luiz Felipe Silva dos Santos[1], Marta Maria Guedes da Silva Neta[1], Matéus Silva Sales[1], Ronaldo Cesar do Nascimento Silva[1], Fernando Antonio Dantas Gomes Pinto[3], Guilherme Lima[3], Erick Müller Gomes dos Santos[4], Edward Hermann Haeusler[3]

[1]*Universidade Federal de Pernambuco (UFPE)*, Recife, Brazil [2]*Universidade Federal do Rio Grande do Norte (UFRN)*, Natal, Brazil [3]*Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)*, Rio de Janeiro, Brazil [4]*Universidade Federal do Recôncavo da Bahia (UFRB)*, Cruz das Almas, Brazil [5]*Universidade Federal de Rondônia (UNIR)*, Porto Velho, Brazil
Corresponding author: nivaldo.vasconcelos@ufpe.br

***Abstract*—Children with severe cerebral palsy (GMFCS levels IV and V) face extreme motor limitations that hinder participation in educational robotics and interactive learning activities. This paper presents the design, specification, and preliminary validation of the Local Learning Environment (LLE), the central component of the NeuroBeep platform. The LLE integrates a noninvasive brain-computer interface based on hemoencephalographic (HEG) neurofeedback with a modular physical space designed for educational and behavioral applications in children with severe motor disabilities. The system is presented through four engineering subsystems: (i) a mechanical design comprising a 2.5 by 2 by 2 m aluminum V-Slot frame with vertical projection and partial acoustic isolation; (ii) a wall-mounted linear robot with a single horizontal degree of freedom, driven by a concentration-relaxation HEG protocol; (iii) an application layer that orchestrates pedagogical sessions aligned with the Brazilian National Common Curricular Base (BNCC); and (iv) a cloud platform with multi-tenant architecture and LGPD-compliant data governance. The LLE has been assembled and is operational in preliminary testing at the i-LITPEG laboratory (UFPE) since December 2024. Structural validation via finite-element analysis confirmed a maximum von Mises stress of 3.83 MPa and a maximum deflection of 0.283 mm under a 100 N load, both far below the admissible limit of aluminum 6061-T6. Bench tests (n = 200 trials) showed an HEG-to-robot command latency averaging 86.85 ms (SD 18.16 ms, maximum 152.65 ms) with 94.0% correct command execution, and acoustic noise peaking at 53 dB, compatible with requirements well below the auditory discomfort threshold (above 100dB). The resulting system, with its modular frame, standardized dimensions, and reconfigurable components, constitutes a reproducible platform for future clinical and school validation of neurofeedback-driven assistive technologies.**

***Index Terms*—assistive technology, brain-computer interface, neurofeedback, educational robotics, cerebral palsy, inclusive education**

This work, named NeuroAssist/NeuroBeep, was funded by the Brazilian Funding Authority for Studies and Projects (FINEP), agreement 01.24.0124.00 (ref. 2199/22).

I. Introduction

Humanity has followed multiple historical paths in dealing with compassion and ethics. Some have normalized the abandonment of the frail; others have chosen responsibility for the care of the other, above all, the vulnerable [1]–[3]. Technological development is inscribed in these paths and faces the same ethical choice. Assistive Technology (AT) emerges in this context [7]. The inclusion of children with severe motor disabilities in teaching and learning processes mediated by emerging technologies still represents a significant challenge for Special Education and AT. According to the World Health Organization, about one billion people live with some form of disability, and approximately 190 million present severe disabilities [4]. Among these, children with cerebral palsy (CP) at levels IV and V of the Gross Motor Function Classification System (GMFCS) have extreme movement limitations, which in most cases hinder their participation in educational robotics, an increasingly common tool in Brazilian schools. Educational robotics has been associated with gains in children's engagement and learning of curricular content [5], [19]–[21]. However, access to such activities presupposes motor capabilities these individuals lack. Brain-computer interfaces (BCI) have enabled people with severe paralysis to communicate and control external devices without relying on movement [22], [23]. A noninvasive form of BCI, neurofeedback, allows a person to voluntarily self-regulate a measured brain signal [6]; captured by hemoencephalography (HEG) over the frontal region, this self-regulation can be turned into a command channel. The NeuroBeep project proposes integrating educational robotics with BCI to overcome motor limitations through the Local Learning Environment (LLE), a physical and technological space designed to receive this new population (universal design), reduce distractions,

and enable standardized training sessions. The LLE aims to provide a modular, controlled, and monitored environment for teaching and learning through educational robotics and neurofeedback, usable by typically developing children and children at GMFCS levels IV and V, meeting ethical and safety standards for research with minors. Unlike approaches that use ocular signals or other residual pathways, the LLE uses neurofeedback from the frontal lobe region, preserving visual exploration and promoting a more natural learning experience. This paper is descriptive and does not present experimental data. Its goal is to document the current configuration of the LLE, its components, technical specifications, and operational flow, serving as a reference for future validation in clinical and school settings. Fig. 1 shows the overall integration of the LLE components around the controlling application (APP), which operates in two complementary modes. In recording mode (Fig. 1(a)), the APP acquires the HEG signal and camera streams, drives the robot, projected stimuli, and audio in a closed loop, and uploads each trial to the cloud. In view mode (Fig. 1(b)), the data flow is reversed: the APP retrieves selected trials from the cloud and replays them on the projection and speakers, while the acquisition devices and robot remain idle.

II. Theoretical Background

Educational robotics is a pedagogical strategy for promoting learning, problem solving, and the development of cognitive and social skills in school-age children. By integrating design, programming, and collaborative work, robotic platforms foster motivation and engagement with curricular content. Systematic reviews document their potential to increase students' interest in science, technology, engineering, and mathematics, especially when implemented as an iterative, problem-based learning tool, as opposed to passive instructional models. In the Brazilian context, these platforms are predominantly Arduino- and ESP32-based kits that still require manual assembly and operation, a constraint that persists even when adapted as assistive technology for students with sensory, neurological, and neurodevelopmental disabilities [24], [25]. These characteristics position educational robotics as a versatile framework for developing complex cognitive competencies throughout formal schooling. Neurofeedback is a closed-loop technique in which a person learns to self-regulate a measured brain signal [6]. Hemoencephalography (HEG) and functional near-infrared spectroscopy (fNIRS) estimate cortical activation, particularly in the prefrontal cortex, from hemodynamic changes, and have been associated with executive function and attention in children and young adults [12], [13]. Cognitive and affective BCI research nonetheless favors cognitive-load metrics over motivational factors and gives insufficient attention to students with disabilities [26], a gap mirrored in assistive-technology delivery, where untrained staff, long waiting times, and low motivation undermine adoption regardless of device sophistication [27], [28]. In the LLE, a concentration-relaxation protocol over the frontal region serves as the command channel for the robot; because the raw HEG signal is subject-specific, the channel requires individual calibration and training before each session, which keeps the visual pathway free for the educational task. The load-bearing structure that hosts these components must be verified before assembly. Structural validation relies on the Finite Element Method (FEM), which discretizes the domain into interconnected elements to approximate the displacement, strain, and stress fields of a part [29], [30]; under a linear static analysis, the von Mises equivalent stress condenses the three-dimensional stress state into a single value comparable with the yield strength of ductile alloys such as aluminum, a standard criterion for sizing metallic frames.

III. Methodology

A. Mechanical design

The mechanical subsystem realizes the LLE as a controlled, partially isolated space that partially decouples the external sensory context and standardizes session conditions across schools. A supporting metal frame with fabric covering provides partial visual isolation. A design driver is the postural constraint of the target population: children at GMFCS levels IV and V need a slight upward inclination to stabilize the neck, which makes a floor-moving robot impossible to view. For this reason, the originally planned horizontal printed floor mat was replaced by a vertical wall projection, and the robot concept changed from a ground robot to a wall-mounted linear robot. The environment also integrates an orthopedic wheelchair adapted for GMFCS IV-V, an audiovisual system (projector, screen, cameras, and speakers), the robot, a control notebook, and support tables. The frame is built from extruded aluminum V-Slot profiles of 20 by 40 mm, chosen for their high strength-to-weight ratio, corrosion resistance, and ease of assembly and reconfiguration; the profiles are joined with M5 bolts, hammer (T-)nuts, and triangular, square, and T brackets, and cable routing uses 3D-printed PLA holders. The design requirements specify a 2.5 by 2 by 2 m enclosure that must not fail under longitudinal and transverse loads, remain disassemblable for car transport, and be safe for children. Before assembly, the profile was verified by a finite-element linear static analysis in ANSYS Workbench [31] on aluminum 6061-T6: a 100 N load (an overestimate of the projector weight) at the center of a profile modeled as a beam with fixed ends, plus its self-weight, over a mesh of about 18,200 tetrahedral elements.

B. Robotics design

The robot is a low-cost, wall-mounted linear robot with a single horizontal degree of freedom, produced mostly by additive manufacturing and responding to commands derived from neurofeedback. Among the interface options considered, aerial platforms (three-dimensional) were discarded for the safety risk they pose to the child, and planar XY systems (two-dimensional), inspired by CNC machines and plotters, were rejected because multidimensional control from a hemoencephalographic (HEG) signal would impose an excessive cognitive load; a one-dimensional horizontal rail was therefore adopted, which also matches the slight upward gaze that suits children at GMFCS levels IV and V. The mechanical concept

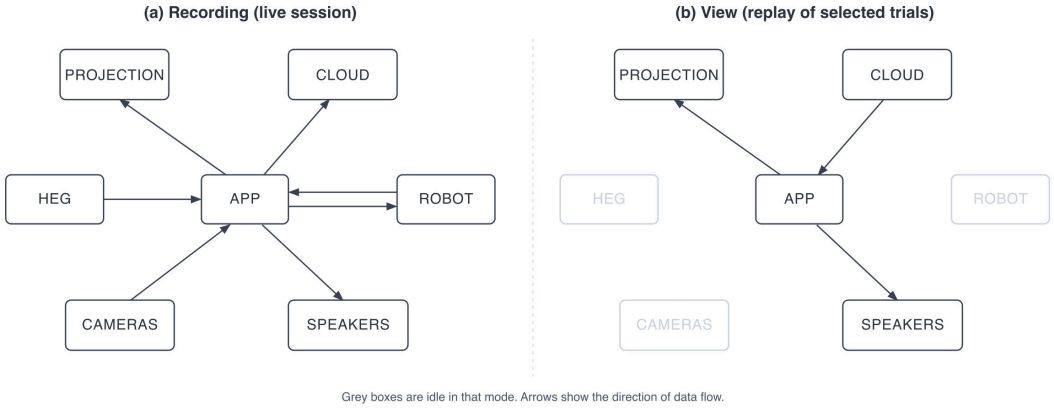


Fig. 1. Overall functioning of the LLE in its two operating modes. The controlling application (APP) is the hub that links HEG acquisition, projection, robot, cameras, speakers, and cloud. (a) In recording mode, sensors and the robot feed the APP, which drives live stimuli and uploads each trial to the cloud. (b) In view mode, the cloud feeds the APP, which replays selected trials on the projection and speakers; acquisition devices and the robot are idle (shown in grey).

emulates the X axis of a Cartesian 3D printer: a fixed base of 40 by 40 mm aluminum profile, cut to a 1900 mm stroke, carries a NEMA 17 stepper motor with a toothed pulley at one end and a free idler pulley acting as a belt tensioner at the other. A moving carriage, built on a PLA plate with four bearings on eccentric spacers for anti-backlash calibration against the rail, is driven by an open timing belt anchored at both ends that converts the motor torque into translation. The logic runs on an ESP32 microcontroller with a DRV8825 driver, powered by a 12 V DC supply and communicating over Bluetooth Low Energy; a single mechanical end-stop defines the origin during a homing routine, while all remaining travel limits are enforced by step counting within a real-time finite-state machine (FSM). The command channel uses an HEG neurofeedback protocol: in the two-command version (2C), the sequence Relax (10 s), Concentrate (3 s), Relax (10 s) triggers a command from a single concentration peak; the four-command version (4C) distinguishes commands by the number of consecutive peaks, using Morse-like logic. A NeuroBeep middleware with four modules (Communication, Commands, User Profile, and Learning) decouples the neural acquisition from the robotic actuation, allowing interoperability with different robots and EEG/HEG devices.

C. APP design

The application is the operational layer that runs the teaching and learning session in the field: it presents the script, orchestrates the robot and projected stimuli, records performance, and synchronizes data with the cloud. As shown in Fig. 1, the APP is the closed-loop hub of the recording mode (HEG, cameras, stimuli, and cloud upload) and, in view mode, replays selected trials from the cloud with the acquisition devices and the robot idle. The pedagogical content follows the Brazilian National Common Curricular Base (BNCC) [17] for the first year of primary education, in Portuguese Language and Mathematics, with a structure of twenty sessions of four to eight questions each (eighty to one hundred and twenty items). The number of alternatives per question (five to nine) follows Miller's law [16], respecting the information-processing capacity of the student. Performance is classified into four standards inspired by state assessment systems SAEPE and SIMAS: Elementary I, Elementary II, Proficient, and Advanced. The binary beep response was replaced by formative video feedback, organized in a flowchart with three paths (correct answer, robot not stopped, and incorrect answer). The application is specified through use cases (user authentication, environment and equipment, scripts and schedule, session execution, data collection and transmission, and monitoring) and roles (Administrator, Researcher, Education Researcher, Mediating Teacher, Student Guardian, and Coordinator).

D. Cloud design

The cloud subsystem provides centralized storage, organization, and indexing of the collected data, as well as administrative project management. It was designed as a modular web solution, hosted at PUC-Rio and accessed by the four executing institutions through a single application programming interface (API), which is the only read and write interface to the database, for both the field application and the web clients. The architecture is multi-tenant: each institution is registered as a region, with scope isolation or consolidation under global governance. Because the data concern minors in schools, the design follows the Brazilian data protection law (LGPD) [18]: it implements multi-method authentication, role-based access control with a hierarchy of profiles, and a detailed audit trail of each event. In the governance model, PUC-Rio acts as the technical operator and the academic partners as controllers of the field data, under ethics approval (CAAE 92643225.8.0000.5208).

IV. Results and Discussion

The main result of this work is the physical materialization and specification of the Local Learning Environment (LLE). The consolidation of this space validates the feasibility of

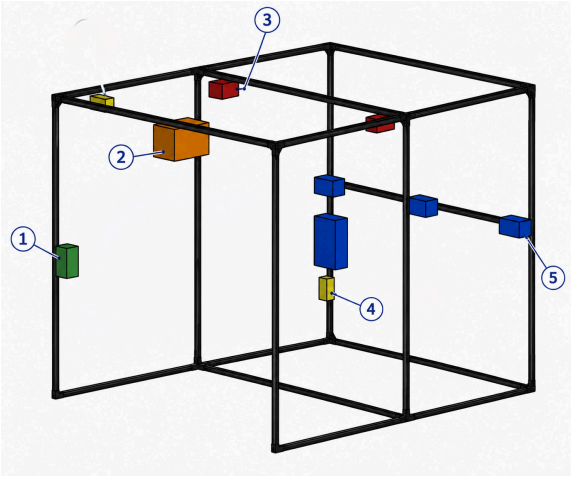


Fig. 2. Isometric view of the LLE supporting frame (2.5 m by 2 m by 2 m), built from aluminum V-Slot profiles, showing the placement of the main components: (1) speaker, (2) projector, (3) cameras, (4) power outlets, and (5) linear robot.

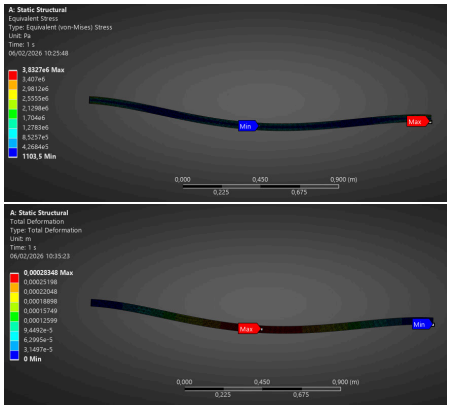


Fig. 3. Finite-element structural analysis of a frame profile under a 100 N central load with fixed supports (aluminum 6061-T6). (a) Equivalent von Mises stress, with a maximum of 3.83 MPa. (b) Total deformation, with a maximum of 0.283 mm. Both values remain far below the admissible limit of the alloy.

integrating a hemoencephalographic (HEG) brain-computer interface into a functional robotic and pedagogical arrangement.

A. Mechanical design

The LLE was assembled and is operational in a preliminary mode (hosted at i-LITPEG since December 2024), comprising the metal frame, fabric covering, orthopedic wheelchair, audiovisual system, robot, control notebook, and support tables. The replacement of the floor mat by a vertical wall projection (Fig. 2) was recorded as an operationalization method adapted to severe motor impairment, with potential for replication in other clinical and school settings. The installation of the acoustic partition walls that complete the isolation is still pending. The finite-element analysis of the structure (Fig. 3) confirmed a wide safety margin: under the 100 N load the profile showed a maximum deflection of 0.283 mm and a maximum von Mises stress of 3.83 MPa, both far below the admissible limit of aluminum 6061-T6, and the estimated maximum load capacity of about 700 N (71.4 kg) greatly exceeds the weight of the installed equipment, a load safety margin of about sevenfold relative to the 100 N design load. The shift from a floor to a wall concept illustrates how a clinical accessibility requirement reorganizes the mechanical design and, in cascade, the firmware, middleware, and environment layout.

B. Robotics design

Two functional units of the linear robot were produced, and the homing routine, FSM-based travel limits, and real-time Bluetooth telemetry were validated on the bench. The mechanical assembly of the wall-mounted track and carriage, with an overall envelope of 442 by 400 by 117 mm, is shown in Fig. 4. Initial bench tests characterized the HEG-to-robot loop latency over $n = 200$ repetitions, exceeding the minimum $n = 73$ determined via Cochran's formula [32] as the maximum of the latency-mean estimate ($n = 28$; $\alpha = 0.05$, $E = 6.757$ ms) and the proportion estimate ($n = 73$; $\pi \geq 0.95$, $\alpha = 0.05$), yielding $\bar{x} = 86.85$ ms ($s = 18.16$ ms, observed maximum $= 152.65$ ms), with 94.0% correct command execution ($188/200$) and 6.0% lost commands ($12/200$). Following the ISO 9283:1998 protocol [33], a bench test repeated without human participation over $n = 30$ repetitions at each of five target positions distributed along the 1800 mm rail (150 measurements in total) showed a systematic overshoot of the commanded position between +0.36 mm and +0.73 mm with standard deviations below 0.55 mm at all positions, evidencing high positioning accuracy and repeatability. The acoustic noise emitted by the robot, measured from the participant's chair, peaked at 53 dB with a mean of 48 dB, remaining significantly below the auditory discomfort level (above 100dB) as required by the environment. These levels are substantially lower than the loudness discomfort level (LDL)—the sound intensity at which discomfort begins, approximately 100 dB HL in normal-hearing adults without tolerance complaints, corresponding to an auditory dynamic range of about 95 dB [34], [35]—and lower than the uncomfortable loudness (UCL) of pediatric listeners, who tolerate levels about 12 dB higher than adults in direct comparison, with UCL exceeding the acoustic reflex threshold by 8.9 dB in children while the values were similar in adults [36], although the absolute values of that study were not accessible. Given the developmental differences in LDL, the wide variability among children of the same age despite relatively consistent loudness growth, and the error that estimating LDL from hearing thresholds can introduce [37], such normative comparisons should be interpreted with caution. The choice of a simple cognitive gesture and its extension by peak counting from two to four commands widens the command space without increasing the load on the child. The middleware is the contribution with the greatest reuse potential, since it enables adoption without replacing existing robot and EEG inventory in schools and laboratories. The code is available in public repositories.



Fig. 4. Mechanical CAD model of the wall-mounted linear robot, with an overall envelope of 442 by 400 by 117 mm: the aluminum-profile track, the wheeled carriage driven by a timing belt, and the stepper-motor drive module at one end. The second rail, set at an angle, reproduces a corner of the wall track.

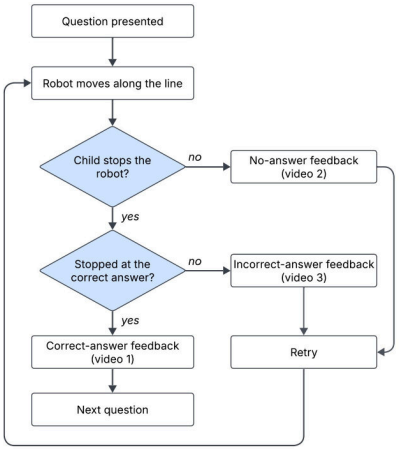


Fig. 5. Formative feedback logic of the APP. For each projected question, the robot moves along the line and the child uses concentration and relaxation to stop it. The stop is evaluated along three paths (no answer, incorrect, correct), each returning a formative video, with retries looping back to a new attempt.

C. APP design

At the content level, the following were completed: the script structure with pilot questions covering BNCC skills, the adaptive formative-feedback flowchart (Fig. 5), a per-session performance registry, and a simulation video of the proposed application on the 2C platform, whose session-monitoring interface is shown in Fig. 6. What remains is deployment of the field application layer, a prerequisite for populating the database. Separating a cloud layer from an application layer allows offline operation at school and later synchronization, a realistic requirement under variable connectivity. Anchoring the content in the BNCC and the assessment in state standards gives the system pedagogical legitimacy and comparability with instruments already used by education networks, while the formative feedback reduces the risk of negative reinforcement in a vulnerable population.

D. Cloud design

The cloud layer reached a consolidated functional stage in its administrative dimension: multi-method authentication, user and role management, multi-tenant operation with the four executing institutions registered as active regions (Fig. 7), an administration panel, and the LGPD audit trail are all operational. The platform is pre-operational: it is functional and ready to receive data, but large-scale collection in schools has not yet begun, as it depends on finalizing the field application, robot, and data-collection equipment. Prioritizing the administrative and governance layer before field collection was a design choice consistent with the context, since, without scope isolation and an audit trail, data collection from minors could not comply with the LGPD. The single-API, multi-tenant model balances institutional autonomy and consolidation, preparing the system to scale to the planned school network.

V. Conclusion

This paper documented the design, specification, and preliminary validation of the LLE, achieving the objective of a modular environment for teaching and learning through educational robotics and neurofeedback for children at GMFCS levels IV and V. The mechanical, robotics, application, and cloud subsystems were specified and partially built: the aluminum V-Slot frame showed a wide structural safety margin (3.83 MPa von Mises stress, 0.283 mm deflection, about sevenfold the 100 N design load); the wall-mounted linear robot exhibited an average HEG loop latency of approximately 86.85 ms and a noise level of 48 dB; the APP orchestrates BNCC-aligned sessions; and the cloud subsystem provides LGPD-compliant, multi-tenant governance. The most significant and original design decision was the change from a floor robot to this wall-mounted linear robot with vertical projection, driven by the postural needs of this population. Limitations of the current stage include the absence of clinical and behavioral validation, pending field application and cloud data collection, and still-incomplete acoustic isolation. The LLE is in preliminary testing at i-LITPEG; future work includes deploying additional units to partner schools and validating the pedagogical scripts with the target population in clinical and school settings.

Acknowledgment

The authors thank Universidade Federal de Pernambuco (UFPE), Universidade Federal do Rio Grande do Norte (UFRN), Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), and Universidade Federal do Recôncavo da Bahia (UFRB) for their institutional support throughout the design, development, and integration of the Local Learning Environment. This work, named NeuroAssist/NeuroBeep, was funded

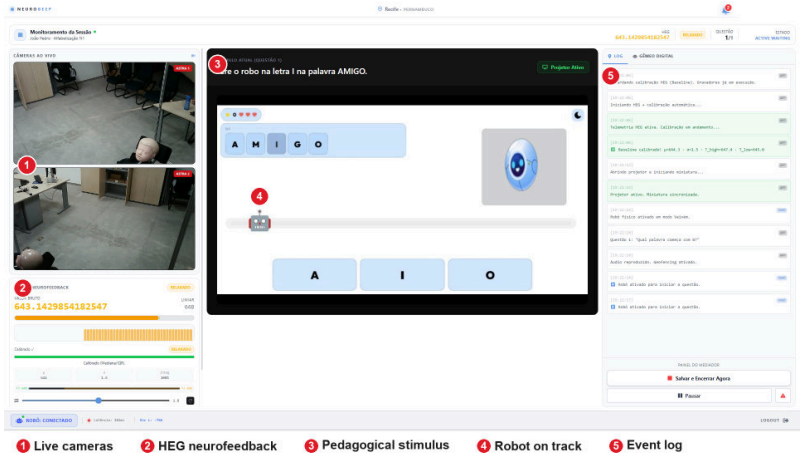


Fig. 6. Session-monitoring screen of the NeuroBeep APP during a running trial (preliminary deployment in Recife, Brazil; the interface is in Portuguese). The APP orchestrates the full closed loop: (1) live cameras of the environment, (2) the HEG neurofeedback panel with the raw value and threshold, (3) the current pedagogical stimulus (here, "stop the robot at the letter I in the word AMIGO"), (4) the robot moving along the track, and (5) the time-stamped event log. The displayed session uses illustrative data.

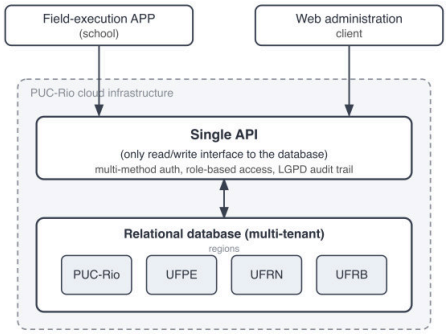


Fig. 7. Architecture of the cloud platform. A single API is the only read and write interface to the relational database, used both by the field-execution APP and by the web administration client. The multi-tenant database registers each executing institution as a region, under PUC-Rio infrastructure and LGPD governance.

by the Brazilian Funding Authority for Studies and Projects (FINEP), agreement 01.24.0124.00 (ref. 2199/22).

References

[1] P. Spikins, A. Needham, L. Tilley, and G. Hitchens, "Calculated or caring? Neanderthal healthcare in social context," *World Archaeology*, vol. 50, no. 3, pp. 384–403, 2018.

[2] D. Lordkipanidze, A. Vekua, R. Ferring, et al., "Anthropology: The earliest toothless hominin skull," *Nature*, vol. 434, no. 7034, pp. 717–718, Apr. 2005.

[3] E. Levinas, *Totality and Infinity: An Essay on Exteriority*, A. Lingis, Transl. Pittsburgh, PA: Duquesne Univ. Press, 1969.

[4] World Health Organization, *World Report on Disability*. Geneva: WHO Press, 2011.

[5] R. V. Aroca, C. C. da Silva, and L. M. G. Gonçalves, "Mobile robotics integration in introductory computer engineering," *IEEE Trans. Educ.*, vol. 56, no. 1, pp. 3–8, Feb. 2013.

[6] R. Sitaram et al., "Closed-loop brain training: the science of neurofeedback," *Nat. Rev. Neurosci.*, vol. 18, pp. 86–100, Jan. 2017.

[7] C. Cook and J. M. Polgar, *Assistive Technologies: Principles and Practice*, 4th ed. St. Louis: Elsevier, 2014.

[8] J. A. Lenker and V. Paquet, "A review of conceptual models for assistive technology," *Assist. Technol.*, vol. 15, no. 1, pp. 1–15, 2003.

[9] E. Astrand, C. Wardak, and S. Ben Hamed, "Selective visual attention to drive a brain-computer interface," *Front. Syst. Neurosci.*, vol. 8, art. 144, Aug. 2014.

[10] Y. Gu, L. Zhang, H. Chen, W. Liu, and Z. Liang, "Improving Attention through Individualized fNIRS Neurofeedback Training: A Pilot Study," *Brain Sciences*, vol. 12, no. 7, p. 862, 2022.

[11] V. Wisnievski, *Biofeedback por Hemoencefalografia (HEG): Apostila 1*. Neurowork, s.d.

[12] Y. Moriguchi, "Prefrontal cortex and executive function in young children: a review of NIRS studies," *Frontiers in Human Neuroscience*, vol. 7, p. 867, 2013.

[13] R. Nouchi, H. Nouchi, J. Dinet, and R. Kawashima, "Cognitive Training with Neurofeedback Using NIRS Improved Cognitive Functions in Young Adults: Evidence from a Randomized Controlled Trial," *Brain Sciences*, vol. 12, no. 1, p. 5, 2021.

[14] B. Vekety, A. Logemann, and Z. K. Takacs, "Mindfulness Practice with a Brain-Sensing Device Improved Cognitive Functioning of Elementary School Children: An Exploratory Pilot Study," *Brain Sciences*, vol. 12, no. 1, p. 103, 2022.

[15] X. Zhong, X. Yuan, Y. Zhang, and C. Jiang, "Neurofeedback training for executive function in ADHD children: a systematic review and meta-analysis," *Scientific Reports*, vol. 15, no. 1, 2025.

[16] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956.

[17] Brazil, Ministry of Education, *Base Nacional Comum Curricular (BNCC)*. Brasília: MEC, 2018.

[18] Brazil, *Lei n. 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais)*, 2018.

[19] F. B. V. Benitti, "Exploring the educational potential of robotics in schools: A systematic review," *Comput. Educ.*, vol. 58, no. 3, pp. 978–988, Apr. 2012.

[20] K. Wang, G.-Y. Sang, L.-H. Huang, S.-H. Li, and J.-W. Guo, "The effectiveness of educational robots in improving learning outcomes: A meta-analysis," *Sustainability*, vol. 15, no. 5, art. 4637, 2023.

[21] F. Ouyang and W. Xu, "The effects of educational robotics in STEM education: A multilevel meta-analysis," *Int. J. STEM Educ.*, vol. 11, art. 7, 2024.

[22] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. E. Vaughan, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun. 2002.

[23] N. Birbaumer et al., "A spelling device for the paralysed," *Nature*, vol. 398, no. 6725, pp. 297–298, Mar. 1999.

[24] C. B. Noronha and A. J. C. Silva, "Robótica educacional no Brasil: uma revisão sistemática de literatura," *Revista FT*, vol. 29, no. 142, 2025, doi: 10.69849/revistaft/th102501130922.

[25] L. L. Santos, L. B. O. Neto, and M. Sabbatini, “O uso da robótica educacional como tecnologia assistiva na aprendizagem de estudantes com deficiências,” *Revista Hum@nae*, vol. 20, no. 1, 2026.
[26] N. Jamil, A. N. Belkacem, S. Ouhbi, and C. Guger, “Cognitive and affective brain-computer interfaces for improving learning strategies and enhancing student capabilities: a systematic literature review,” *IEEE Access*, vol. 9, pp. 134122–134147, 2021.
[27] T. A. Fachinetti, A. G. Gonçalves, and G. F. Lourenço, “Construction process assistive technology resource for students with cerebral palsy in multifunctional resources room,” *Rev. Bras. Educ. Espec.*, vol. 23, no. 4, pp. 547–562, Oct.–Dec. 2017.
[28] P. Karlsson, C. Johnston, and K. Baker, “Influences on students’ assistive technology use at school: the views of classroom teachers, allied health professionals, students with cerebral palsy and their parents,” *Disability and Rehabilitation: Assistive Technology*, 2017, doi: 10.1080/17483107.2017.1373307.
[29] F. P. Beer, E. R. Johnston, J. T. DeWolf, and D. F. Mazurek, *Mechanics of Materials*, 8th ed. New York, NY: McGraw-Hill Education, 2020.
[30] R. C. Hibbeler, *Mechanics of Materials*, 10th ed. Hoboken, NJ: Pearson, 2017.
[31] ANSYS, Inc., *ANSYS Mechanical User’s Guide*, Release 2023 R1. Canonsburg, PA, USA: ANSYS, Inc., 2023.
[32] Cochran WG. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons; 1977.
[33] ISO 9283:1998 – Manipulating industrial robots – Performance criteria and related test methods. International Organization for Standardization, Geneva, 1998.
[34] Sherlock, L. P., & Formby, C. (2005). Estimates of loudness, loudness discomfort, and the auditory dynamic range: Normative estimates, comparison of procedures, and test-retest reliability. *Journal of the American Academy of Audiology*, *16*(2), 85–100. https://doi.org/10.3766/jaaa.16.2.4
[35] Aazh, H., & Moore, B. C. J. (2017). Factors related to uncomfortable loudness levels for patients seen in a tinnitus and hyperacusis clinic. *International Journal of Audiology*, *56*(10), 793–800. https://doi.org/10.1080/14992027.2017.1335888
[36] Hong, H. N., Cho, S., & Hong, B. N. (2012). Comparison of the uncomfortable loudness level with the acoustic reflex threshold in normal children and adults. *Audiology and Speech Research*, *8*(2), 177–182. https://doi.org/10.21848/audiol.2012.8.2.177
[37] Gustafson, S. J., Newsome, E., & Flores, A. N. (2022). Narrative review of loudness perception measures in children. *Ear & Hearing*, *43*(6), 1635–1642. https://doi.org/10.1097/AUD.0000000000001243

Development and application of a wearable system for cardiorespiratory monitoring during physiotherapeutic exercises in post-COVID-19 patients

Admilson C. Chaves Filho
Biological Sciences Department
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0001-6939-9716

Ana V. M. Inocêncio Lopes
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0002-1411-75477

Naelso A. Cunha
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0001-7225-2360

Carlos Rodrigues
Doctoral Program in Biomedical Engineering
Faculty of Engineering of the University of Porto
Porto, Portugal
https://orcid.org/0000-0001-6408-6032

Miguel Correia
C-BER - Centre for Biomedical Engineering Research
INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal
https://orcid.org/0000-0001-6065-9358

Marco A. B. Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0002-5143-3696

***Abstract*—COVID-19 pandemic caused significant cardiorespiratory and functional repercussions, increasing the need for remote monitoring strategies during physiotherapeutic rehabilitation programs. In this context, wearable devices emerge as promising tools for the continuous acquisition of physiological parameters in a home environment. The objective of this study was to develop and evaluate the applicability of a wearable system for cardiorespiratory monitoring during physiotherapeutic exercises performed by individuals with post-COVID-19 sequelae. Therefore, a system composed of wearable modules was developed to acquire peripheral oxygen saturation (SpO_2), heart rate (HR), body temperature, and electrocardiogram (ECG), all integrated into a web platform, called SISMO-UFPE, via Wi-Fi communication. Ten individuals with post-COVID-19 sequelae participated in the study and underwent a protocol consisting of 13 physiotherapeutic exercises. The system allowed continuous acquisition of physiological signals during rehabilitation sessions, enabling real-time remote monitoring. Eighty percent of the participants presented cardiorespiratory responses classified as highly satisfactory and 20% as satisfactory. The results demonstrate the technical and clinical feasibility of the developed system for applications in telerehabilitation and remote patient monitoring.**

Keywords— Physiological monitoring; Wearable devices; Telerehabilitation; COVID-19; Cardiorespiratory assessment.

I. INTRODUCTION

The pandemic caused by SARS-CoV-2 represented one of the greatest health challenges of recent decades, resulting in millions of hospitalizations and a growing number of individuals with persistent post-acute sequelae [1]. Among the main clinical manifestations, persistent dyspnea, fatigue, exercise intolerance, and reduced functional capacity stand out, frequently demanding prolonged physiotherapeutic rehabilitation [1].

During exercise exertion, acute physiological shifts—such as increased myocardial oxygen demand, altered autonomic tone, dynamic changes in ventilation-perfusion matching, and transient peripheral desaturation—may occur. In post-COVID-19 patients who often present residual pulmonary lesions, altered vascular reactivity, or autonomic dysfunction, unmonitored physical exercise poses severe safety risks, including malignant arrhythmias, exertion-induced silent hypoxia, and abnormal pressor or thermal responses. Therefore, continuous and real-time monitoring of vital signs during physiotherapeutic sessions is clinically essential to titrate exercise intensity, assess therapeutic safety, prevent exertion-triggered cardiovascular events, and evaluate the efficacy of proposed rehabilitation interventions. [2][3]

During the pandemic, telehealth and telerehabilitation played a fundamental role in the continuity of care, enabling remote consultations and reducing risks related to patient exposure and displacement [4]. However, executing physiotherapeutic exercises outside the hospital environment requires mechanisms capable of ensuring clinical safety and continuous monitoring of physiological signals [4].

Despite recent advancements, current commercial off-the-shelf devices and existing wearable systems exhibit notable limitations when applied to remote physiotherapeutic rehabilitation: Most existing commercial wearables (e.g., fitness bands or standalone pulse oximeters) lack multi-parameter synchronization, providing fragmented data across different applications.[5] Many portable monitors suffer from high motion artifact sensitivity during dynamic motor tasks and lack clinical-grade ECG acquisition capable of detailed electrophysiological wave analysis.[6] Existing commercial

platforms rarely offer dedicated web interfaces that allow physiotherapists to monitor raw physiological signals in real-time while correlating them directly with specific exercise protocols.[7]

To overcome these limitations, this work introduces an integrated, low-cost multi-sensor wearable system specifically tailored for telerehabilitation. The novelty of this study lies in the simultaneous, synchronized wireless acquisition of multi-lead equivalent ECG, continuous SpO_2, heart rate, and surface body temperature integrated into a custom-designed web platform (SISMO-UFPE) that directly maps physiological responses to specific upper-limb, lower-limb, and respiratory exercise protocols in real time. Therefore, this work presents the development and application of a wearable system for cardiorespiratory monitoring during physiotherapeutic exercises performed by patients with post-COVID-19 sequelae [9].

II. MATERIALS AND METHODS

A. Study Design

This is a translational study involving the technological development, functional validation, and preliminary clinical application of a wearable system for remote cardiorespiratory monitoring. The study was approved by the Research Ethics Committee under opinion No. 5.135.351.

The developed system consisted of wearable modules capable of acquiring relevant physiological parameters during the execution of physiotherapeutic exercises, being positioned on the patient according to Fig. 1. The hardware specifications and acquisition settings are detailed below:

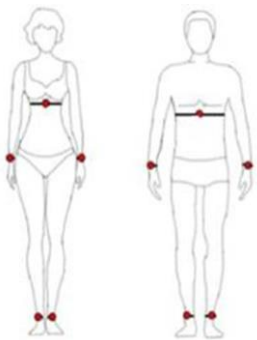


Fig. 1. Anatomical positioning of the ECG and SpO2 wearable sensors. Source: The authors, 2025.

The set was constituted by:

- Body Temperature Sensor: The AHT10 high-precision digital sensor module was used. The sensor was housed in a wearable module positioned on the participant's thoracic region, calibrated prior to data collection using a commercial digital clinical thermometer (Fig. 2A and 2B).
- ECG Acquisition Module: Developed using an INA126 instrumentation amplifier with a gain of 301, a passive high-pass filter with a cutoff frequency of 0.5 Hz (to eliminate baseline wander), and an active low-pass filter with a cutoff frequency of 100 Hz providing an additional 21.3 times amplification. Digitization was performed via the microcontroller's 12-bit A/D converter at a sampling rate of 500 Hz, utilizing pre-gelled disposable Ag/AgCl surface electrodes positioned in a modified chest lead configuration (Fig. 2C).
- Optical Sensor for SpO_2 and HR Acquisition: The MAX30102 module was used, which is an integrated pulse oximetry and heart rate monitor module based on photoplethysmography (PPG) technology sampled at 100 Hz (Fig. 2D).
- Microcontroller with Wireless Transmission: The ESP32-S3 MINI wireless module (Xtensa® dual-core 32-bit LX7 processor) transmitted data via 2.4 GHz Wi-Fi connectivity to the SISMO-UFPE web platform.
- The Post-COVID-19 Patient Physiotherapeutic Monitoring System (SISMO-UFPE) (Fig. 3) consists of a web platform developed for real-time storage, visualization, and analysis of physiological signals acquired during telerehabilitation sessions, allowing for secure storage of clinical and physiological information to support the physiotherapist's decision-making.

The acquired signals were converted to digital format and transmitted via Wi-Fi to a remote server, allowing for real-time monitoring by the physiotherapist.

B. Participants

Ten individuals with post-COVID-19 sequelae recruited by convenience participated in the study. Inclusion criteria encompassed patients with a confirmed COVID-19 diagnosis, presence of persistent post-acute cardiorespiratory symptoms (> 3 months post-infection), age ≥ 18 years, and preserved cognitive capacity to perform the proposed exercises. Excluded from the study were individuals presenting clinical instability, severe motor limitations precluding the execution of the protocol, or lack of internet access for communication with the platform.

Medical histories were reviewed to verify comorbidities. Systemic Arterial Hypertension (SAH), Diabetes Mellitus, and Heart Disease were diagnosed based on medical records

following standard clinical guidelines. Participant pharmacological regimens (e.g., antihypertensives, beta-blockers, oral hypoglycemics) were documented to account for potential pharmacological influences on heart rate and autonomic responses.

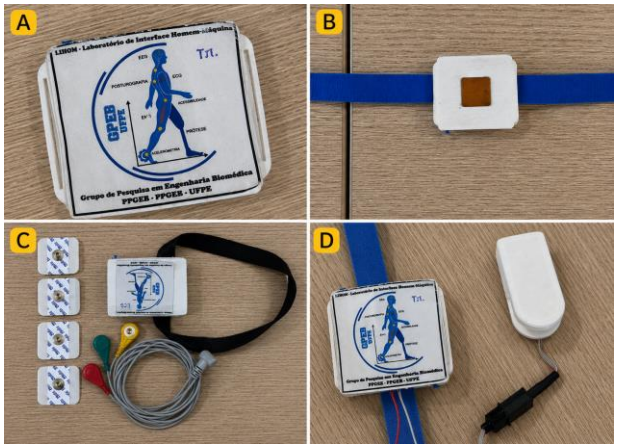


Fig. 2. Physical architecture of the wearable prototype used for physiological signal acquisition. (A) Temperature acquisition module; (B) Temperature sensor; (C) ECG acquisition module; (D) SpO2 acquisition sensor. Source: The authors, 2025.



Fig. 3. Representative screens of the SISMO/UFPE web platform used for the acquisition, storage, visualization, and analysis of remotely collected physiological signals. Source: The authors, 2025.

C. Procedures & Exercise Protocol

The sessions occurred remotely, lasting on average between 30 and 50 minutes, three times a week. The participants underwent a structured protocol composed of 13 exercises distributed into three distinct categories designed to impose varying cardiorespiratory and physiological demands:

1. Respiratory Exercises: Static and low-movement exercises (e.g., fractional 3-time inspiration, diaphragmatic breathing) aimed at enhancing ventilatory capacity while generating minimal motion artifacts.
2. Motor Exercises for Upper Limbs: Dynamic resistance exercises (e.g., shoulder flexion and horizontal abduction with dumbbells, shoulder press) generating moderate metabolic demand and localized muscular activity.
3. Motor Exercises for Lower Limbs: Dynamic cyclic tasks (e.g., seated cycle ergometer pedaling, leg extensions) producing high systemic cardiorespiratory demands and higher mechanical movement.

Throughout the execution of the protocol, physiological signals were continuously acquired by the developed system.

D. Statistical Analyses

Data were analyzed using descriptive statistics, employing measures of central tendency and dispersion, reported as Mean ± Standard Deviation (Mean ± SD) along with minimum and maximum values for numerical variables, as well as absolute and relative frequencies to characterize the sample. Graphical analyses of the cardiorespiratory parameter signals obtained during the functional tasks were also conducted.

To evaluate signal consistency across repeated exercise sets, measurement reliability was estimated via Intraclass Correlation Coefficient ($ICC_{3,1}$) and Standard Error of Measurement (SEM) across baseline, peak exercise, and recovery phases.

III. Results

A. Sample Characterization

The sample was composed of 10 participants, of which 70% were male and 30% were female, with a mean age of 63.4 ± 5.2 years (age group between 60 and 69 years accounting for 70%). Among the observed comorbidities, systemic arterial hypertension (50%), diabetes mellitus (20%), and heart disease (10%) stood out (Tab. 1). Fifty percent of participants were actively receiving antihypertensive therapy. Sedentary lifestyle was considered a lifestyle habit rather than an associated clinical condition.

TABLE I. – Participant Characterization

Variable	**Results**
Male	70%
Female	30%
60–69 years	70%
Age (years) [Mean ± SD]	63.4 ± 5.2
SAH (Systemic Arterial Hypertension)	50%
Diabetes	20%
Heart Disease	10%
Sedentary	60%
Pharmacological Use (β-blockers / ACEi)	10% / 40%

B. Physiological Monitoring

The system allowed continuous acquisition of physiological signals during all sessions performed by the participants. The

ECG recordings showed sufficient quality for the identification of major cardiac electrophysiological components, including P waves, QRS complex, and T wave, at a 500 Hz sampling rate during the exercises (Fig. 3A and 3B). During lower-limb ergometer tasks, motion artifacts caused transient baseline fluctuations in < 4% of total signal duration, which were effectively filtered.

Quantitative mean values obtained during resting and exercise protocols were as follows:

Heart Rate (HR): Baseline resting HR was 71.2 ± 6.4 bpm, increasing to a peak during cycle ergometry of 108.5 ± 12.1 bpm.

Oxygen Saturation (SpO_2): Peripheral oxygen saturation signals remained within expected physiological limits averaging 97.1 ± 1.1% at rest and 95.8 ± 1.4% (range: 92%–99%) during peak exercise, demonstrating cardiorespiratory stability (Fig. 3C).

Body Temperature: Mean skin temperature remained stable at 36.4 ± 0.4 °C (range: 35.8 °C–37.1 °C) with no evidence of clinically relevant alterations (Fig. 3D).



Fig. 3. Physiological signals monitored and visualized on the SISMO/UFPE platform during the execution of the exercise protocol. (A) Electrocardiographic (ECG) recording in an expanded time window; (B) Electrocardiographic (ECG) recording in a continuous time scale; (C) Peripheral oxygen saturation (SpO_2) monitoring; and (D) Body temperature monitoring. Source: The authors, 2025.

System reliability across trial repetitions yielded an $ICC_{3,1}$ of 0.92 (95% CI: 0.85–0.96) for HR and 0.88 (95% CI: 0.79–0.94) for SpO_2, with SEM values of 2.1 bpm and 0.6% respectively, confirming adequate measurement reproducibility for telerehabilitation monitoring.

C. Global Assessment of Vital Signs

The analysis of hemodynamic responses demonstrated good tolerance to the exercise protocol, with 80% of participants classified as "highly satisfactory" and 20% as "satisfactory," with no records of "unsatisfactory" responses. The "highly satisfactory" classification was assigned to participants who maintained vital signs within normal limits throughout the activity, considering $SpO_2 \geq 95\%$, heart rate between 60 and 100 bpm at rest or an increase compatible with exercise intensity, and the absence of relevant symptoms. Participants classified as "satisfactory" presented only mild and transient physiological alterations, such as SpO_2 between 90% and 94% or slight heart rate elevation, with no need to interrupt the exercise. The "unsatisfactory" category would include individuals with $SpO_2 < 90\%$, significant cardiovascular alterations, or limiting symptoms. These results indicate hemodynamic stability and reinforce the safety and good tolerability of the applied protocol (Tab. 2).

TABLE II. – CLASSIFICATION OF VITAL SIGNS DURING THE PROTOCOL

Classification	Frequency
Highly satisfactory	80%
Satisfactory	20%
Unsatisfactory	0%

IV. DISCUSSION

Continuous monitoring of cardiorespiratory parameters represents an important tool to increase the safety of telerehabilitation, especially in elderly individuals and those with systemic arterial hypertension or heart diseases. The simultaneous monitoring of heart rate, peripheral oxygen saturation, body temperature, and electrocardiogram allows for early identification of physiological changes potentially related to physical exertion, facilitating immediate interventions by the physiotherapist and reducing risks during remote sessions. The results demonstrated that the developed wearable system was capable of performing continuous and remote monitoring of multiple cardiorespiratory parameters during home-based physiotherapy sessions.

Recent studies show significant progress in the development of wearable devices for remote monitoring of physiological parameters, albeit with different levels of integration between sensors and functionalities. The Withings ScanWatch, validated by Kirszenblat and Edouard [10], was developed primarily for non-invasive monitoring of peripheral oxygen saturation (SpO_2) using reflectance photoplethysmography on the wrist, in addition to heart rate, showing adequate clinical performance for detecting hypoxia and sleep-related breathing disorders during rest and sleep. The study by Liu et al. [11] utilized a continuous remote electrocardiogram (ECG) monitoring system to track patients' cardiac safety during the pandemic, allowing the identification of electrophysiological alterations and possible arrhythmias associated with drug treatment, without

incorporating additional sensors for respiratory or thermal monitoring.

The TELECOVID device, described by Baldinger et al.[12], employs an in-ear biosensor capable of continuously monitoring heart rate, peripheral oxygen saturation (SpO_2), body temperature, and respiratory rate, aimed at home monitoring of COVID-19 patients in isolation. In turn, the Respinos system, developed by Adiono et al. [13], consists of portable multiparameter equipment designed for remote monitoring of COVID-19 patients, integrating SpO_2, heart rate, respiratory rate, and body temperature, focusing on clinical screening and home follow-up.

In comparison, the system developed in this study stands out due to the simultaneous integration of electrocardiogram (ECG), peripheral oxygen saturation (SpO_2), heart rate, and body temperature into a specific platform for physiotherapeutic telerehabilitation (SISMO-UFPE), allowing continuous acquisition of signals during exercise execution, real-time transmission via Wi-Fi, and remote tracking by the physiotherapist, broadening its applicability for cardiorespiratory rehabilitation programs and home-based functional monitoring

From a Biomedical Engineering perspective, the integration of sensors, wireless communication, and a telemonitoring platform represents a potentially cost-effective solution due to the use of widely available commercial sensors, low-power microcontroller modules, and an open ESP32-based architecture, enabling home monitoring and tracking of patients with chronic diseases.

Limitations of this preliminary study include the small sample size (n=10) and the absence of a control group. Moreover, Wi-Fi transmission may limit the system's application in outdoor environments lacking network coverage. Future works plan to expand the sample and include clinical validation protocols compared to gold-standard hospital equipment.

V. CONCLUSIONS

The results demonstrate that the developed wearable system constitutes a promising tool for remote monitoring of patients with post-COVID-19 sequelae undergoing physiotherapeutic rehabilitation, contributing to greater clinical safety, real-time monitoring, and decision-making support in telerehabilitation programs. The continuous acquisition of ECG, heart rate, peripheral oxygen saturation, and body temperature allows real-time tracking of participants' physiological responses, highlighting its potential use in telerehabilitation programs and remote healthcare assistance. The obtained results reinforce the contribution of Biomedical Engineering to the development of technologies applied to healthcare, especially in scenarios demanding secure and efficient remote monitoring.

ACKNOWLEDGMENT

The authors acknowledge the financial support provided by CNPq (National Council for Scientific and Technological Development) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

REFERENCES

[1] BRITO, Sávio Breno Pires; BRAGA, Isaque Oliveira; CUNHA, Carolina Coelho; PALÁCIO, Maria Augusta Vasconcelos; TAKENAMI, Iukary. Pandemia da COVID-19: o maior desafio do século xxi. Vigilância Sanitária em Debate, [S.L.], v. 8, n. 2, p. 54-63, 29 maio 2020. Vigilancia Sanitaria em Debate: Sociedade, Ciencia y Tecnologia. http://dx.doi.org/10.22239/2317-269x.01531.

[2] WORLD HEALTH ORGANIZATION. *Clinical management of COVID-19: living guideline*. Geneva: World Health Organization, 2025. Disponível em: https://iris.who.int/bitstream/handle/10665/381920/B09467-eng.pdf. Acesso em: 25 ago. 2026.

[3] GOWER, B. **et al.** Recommendations for objective cardiovascular assessment to inform clinical exercise prescription: an Exercise Physiologist and Physiotherapist expert consensus. *Journal of Science and Medicine in Sport*, v. 26, n. 9, p. 454-458, 2023. DOI: 10.1016/j.jsams.2023.07.004.

[4] COFFITO. Conselho Federal de Fisioterapia e Terapia Ocupacional (COFFITO) [homepage on the Internet]. Brasília: COFFITO; c2020 [cited 2020 Dec 28]. RESOLUÇÃO Nº 516, DE 20 DE MARÇO DE 2020 – Teleconsulta, Telemonitoramento e Teleconsultoria. Available from: https://www.coffito.gov.br/nsite/?p=15825

[5] LATIF, A. et al. Use of commercially available wearable devices for physical rehabilitation in healthcare: a systematic review. *BMJ Open*, v. 14, n. 11, e084086, 2024. DOI: 10.1136/bmjopen-2024-084086.

[6] PETEK, B. J. et al**.** Consumer wearable health and fitness technology in cardiovascular medicine: JACC state-of-the-art review. *Journal of the American College of Cardiology*, v. 82, n. 3, p. 245-264, 2023. DOI: 10.1016/j.jacc.2023.04.054.

[7] ALBERT, M. et al**.** Remote and wearable ECG devices with diagnostic abilities in adults: a state-of-the-science scoping review. *Journal of Electrocardiology*, 2022.

[8] LUNDELL S, Holmner Å, Rehn B, Nyberg A, Wadell K. Telehealthcare in COPD: a systematic review and meta-analysis on physical outcomes and dyspnea. Respir Med. 2015;109(1):11-26. https://doi.org/10.1016/j.rmed.2014.10.008.

[9] Espressif Systems. ESP32-S3 Series Datasheet. Shanghai: Espressif Systems; 2024.

[10] KIRSZENBLAT, Romain; EDOUARD, Paul. Validation of the Withings ScanWatch as a Wrist-Worn Reflective Pulse Oximeter: prospective interventional clinical study. Journal Of Medical Internet Research, v. 23, n. 4, p. 27503, 26 abr. 2021. JMIR Publications Inc.. http://dx.doi.org/10.2196/27503.

[11] LIU H., Allen J., Zheng D., Chen F. Recent Development of Respiratory Rate Measurement Technologies. Physiol. Meas. 2019;40:07TR01. doi: 10.1088/1361-6579/ab299e.

[12] BALDINGER, Melanie; HEINRICH, Axel; ADAMS, Tim; MARTENS, Eimo; DOMMASCH, Michael; MÜLLER, Alexander; SIEGMANN, Adonia; SCHMIDT, Georg. TELECOVID: remote vital signs monitoring of covid-19 risk patients in home isolation with an in-ear wearable. Ieee Pervasive Computing, [S.L.], v. 20, n. 2, p. 58-62, 1 abr. 2021. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/mprv.2021.3066825.

[13] ADIONO, Trio; AHMADI, Nur; SARASWATI, Citrya; ADITYA, Yudi; YUDHANTO, Yusuf Purna; AZIZ, Abdillah; WULANDARI, Laksmi; MARANATHA, Daniel; KHUSNURROKHMAN, Gemilang; RIADI, Agustinus Rizki Wirawan. Respinos: a portable device for remote vital signs monitoring of covid-19 patients. Ieee Transactions On Biomedical Circuits And Systems, [S.L.], v. 16, n. 5, p. 947-961, out. 2022. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tbcas.2022.3204632.

Development and Preliminary Evaluation of a Bidirectional Telerehabilitation Platform for Remote Treadmill Control and Heart Rate Monitoring

1st João Pedro Nascente
Graduate Program in Computer Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
joao.nascente@edu.ufes.br
ORCID: 0009-0003-2757-504X

2nd Brayan Arevalo
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
brayan.arevalo@edu.ufes.br
ORCID: 0000-0001-8346-0464

3rd Carla Zimerer
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
czimerer@gmail.com
ORCID: 0000-0003-2872-5209

4th Anselmo Frizera
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
anselmo.frizera-neto@ufes.br
ORCID: 0000-0002-0687-3967

5th Ricardo Mello
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
ricardo.c.mello@ufes.br
ORCID: 0000-0003-0420-4273

***Abstract*—This study aimed to develop and evaluate a digital system to support physical telerehabilitation by integrating active control of treadmill speed and incline with remote heart rate monitoring. The proposed system addresses the growing demand for technological solutions that enable the continuity of remote rehabilitation programs, particularly for individuals with mobility limitations and restricted access to specialized rehabilitation centers. The system was evaluated through a mixed-methods approach, including open-ended questionnaires administered to therapists (n = 13) and participants (n = 13), usability assessment using the System Usability Scale (SUS), and analysis of the agreement of the physiological data acquired. The results demonstrated high perceived usability (SUS = 95.19 ± 4.26) and strong agreement in physiological data transmission (mean absolute error = 0.41 bpm; session ICCs ranging from 0.972 to 0.999), confirming the integrity of the data acquisition and synchronization pipeline from a shared sensor source. Qualitatively, therapists and participants identified remote accessibility, perceived safety, and real-time interaction as relevant advantages, while suggesting automated clinical protocols and expanded monitoring as areas for improvement. These findings support the preliminary technical and usability feasibility of the proposed platform under controlled experimental conditions. Future work should focus on improving the user interface, expanding monitoring capabilities, and evaluating the system in more representative clinical and home-based settings.**

***Index Terms*—patient rehabilitation, teleoperation, remote patient monitoring, patient-therapist interaction, telerehabilitation**

This work was supported by the São Paulo Research Foundation (FAPESP) under grant numbers 18/23097-3 and 20/05182-3; by the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) under grant numbers 2025-QH0K, 2025-J7XD, and 2026-48WNC; by the Brazilian National Council for Scientific and Technological Development (CNPq) under grant numbers 308155/2023-8 and 402080/2025-4; and by the CNPq/FAPES joint program under grant number 2025-6VCJ4.

I. Introduction

At least one-third of the world's population could benefit from physical or cognitive rehabilitation [1]. Musculoskeletal disorders represent one of the leading demands for rehabilitation, affecting approximately 1.71 billion people worldwide and often resulting in reduced mobility and impaired gait [2]. These conditions diminish individuals' independence and quality of life while imposing substantial financial and social burdens on patients, caregivers, and healthcare systems [3], [4]. Although physical rehabilitation programs can mitigate these impacts [5], access to such services remains limited, particularly for individuals living in rural areas or with low socioeconomic status, due to geographic and socioeconomic barriers [6].

In this context, telerehabilitation has emerged as an accessible alternative, enabling continuous care and greater patient independence through remote sessions [7], [8]. Within this modality, treadmills are particularly well suited because they require minimal space, provide a safe exercise environment, and allow intensity to be adjusted via speed and incline [9], [10]. However, extending this model to remote settings remains a technological challenge. Systems capable of remotely controlling treadmills are still scarce, and their integration with real-time heart rate (HR) monitoring is essential for the safe, personalized adjustment of exercise intensity based on the patient's physiological response [11], [12].

Beyond technological challenges, the limited collaboration between software developers and healthcare professionals has hindered the acceptance and adoption of new technologies in

clinical practice [13], [14]. Although telerehabilitation platforms are designed to support both patients and therapists, the perspectives of these end users remain underrepresented in the telehealth literature. Considering their needs is essential for developing clinically relevant and applicable technologies. Therefore, this study aimed to develop and evaluate a digital system to support physical telerehabilitation by integrating the remote control of treadmill speed and incline with synchronous HR monitoring. The main contribution of this work is the deployment and evaluation of a human-in-the-loop bidirectional framework that empowers therapists to actively modulate exercise intensity based on real-time physiological feedback. System evaluation included the assessment of usability and user experience from both therapists and participants, alongside an analysis of data transmission agreement.

II. Methodology

A. Participants

The study sample consisted of 26 participants in total, organized into 13 independent dyads. Each dyad was composed of one therapist (broadly defined as exercise and movement science professionals) and one exercising subject. The therapist group consisted primarily of Master's students in Physical Education, alongside Ph.D. holders and senior undergraduate students, representing a mix of academic and clinical experience levels. Initially, 15 potential subjects were screened using the Physical Activity Readiness Questionnaire (PAR-Q), a screening instrument used to identify potential health risks associated with participation in physical exercise [15]. Two individuals were excluded for not meeting the safety criteria, resulting in a final sample of 13 healthy adults (age: 22.0 ± 3.39 years).

B. System Description

The system architecture (Fig. 1) comprises four main software components: (1) a mobile application developed in Kotlin (Android SDK v2.2); (2) an InfluxDB (v2.7) time-series database; (3) a RESTful Application Programming Interface (API) implemented in Kotlin using the Spring Boot framework (v3.5.5); and (4) a therapist interface (Fig. 2). The RESTful API is hosted on an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) instance, which provides a virtual cloud computing environment for deploying and running the system.

The mobile application runs on a tablet (Galaxy Tab S9 Ultra, Samsung, South Korea) and communicates via Bluetooth Low Energy (BLE) with a treadmill (TR5, Speedo, Brazil) and a heart rate monitor (H10, Polar, Finland). Communication with the treadmill is based on the Bluetooth Fitness Machine Service (FTMS) protocol.

The exercising participant is monitored using the Polar monitor, which provides heart rate (HR), electrocardiogram (ECG), and RR interval data. From the treadmill, the system acquires the actual speed (Vr), actual incline (Ir), and elapsed exercise time (ETr). These measurements are initially received as hexadecimal values and subsequently converted into floating-point format for processing.

Physiological and treadmill data are transmitted to the InfluxDB database in real time at the sampling rates provided by each device (approximately 1 Hz for the treadmill and 130 Hz for the heart rate monitor). The stored data can

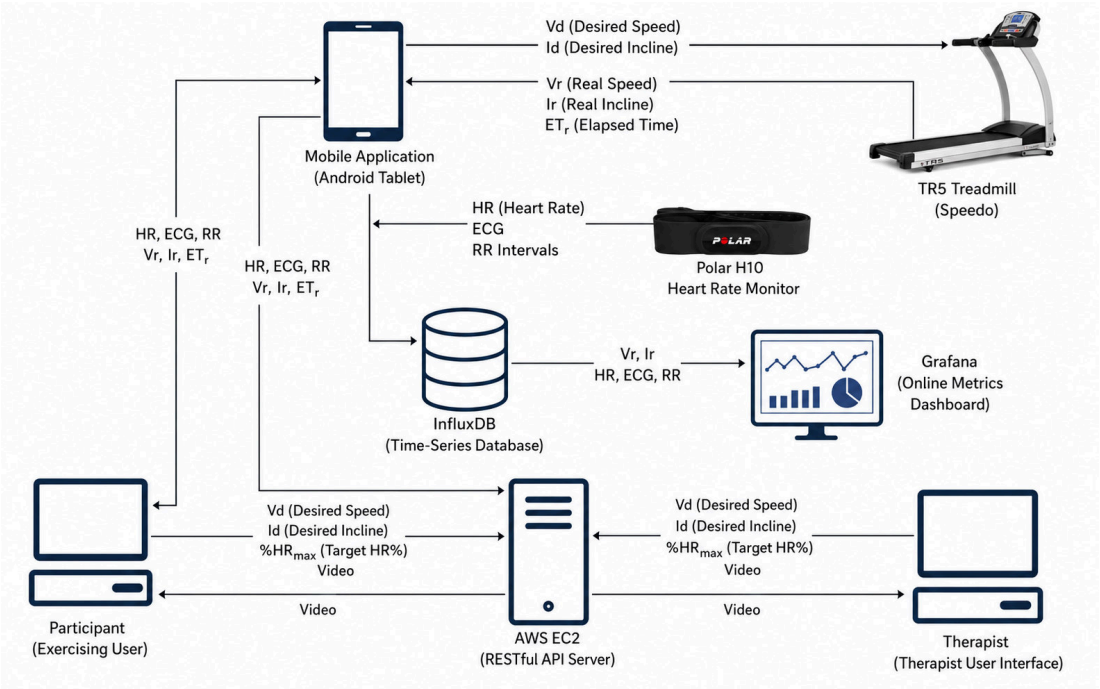


Fig. 1. Overall architecture of the bidirectional telerehabilitation system, illustrating the data flow between the local hardware, the cloud-based API, and the remote interfaces.

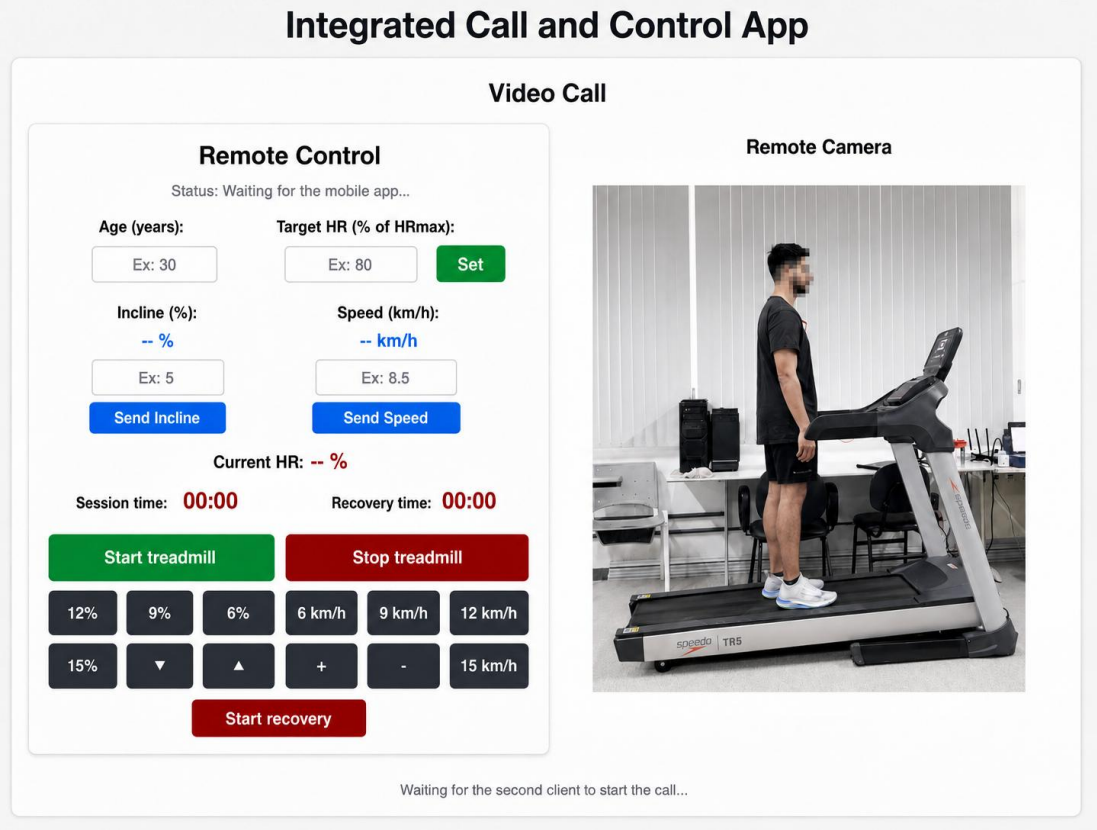


Fig. 2. Therapist user interface integrating synchronous video communication (right panel) with the human-in-the-loop bidirectional remote-control dashboard (left panel). The interface enables real-time monitoring of physiological variables, such as current HR percentage, and active adjustment of treadmill speed and incline.

be visualized using Grafana or exported as comma-separated values (.csv) files for offline analysis. Simultaneously, the mobile application communicates with the remote RESTful API to transmit the desired speed (Vd) and desired incline (Id) commands specified by the therapist to the treadmill.

The bidirectional communication architecture enables remote monitoring of the participant while allowing the therapist to adjust treadmill parameters in real time. The Therapist User Interface (TUI) provides controls for transmitting treadmill commands, configuring the target percentage of maximum heart rate (%HRmax), visualizing monitored variables (%HRmax, Vr, Ir, ETr, and recovery time (RTr)), and conducting video calls to support synchronous communication between the therapist and the participant. To align with general data protection regulations (e.g., LGPD), the InfluxDB database stores only anonymized session data without personally identifiable information (PII). Furthermore, all remote communication, including video streaming and RESTful API requests, is encrypted via HTTPS. Comprehensive access control and authentication mechanisms are slated for future clinical deployment.

C. Experimental Design and Procedures

The experimental protocol was designed as a technical and usability evaluation of the digital platform. The study was approved by the Research Ethics Committee of the Federal University of Espírito Santo (approval no. 6,294,101; CAAE: 70158523.2.0000.5542), and all participants provided written informed consent before data collection. The experimental protocol consisted of a telerehabilitation session conducted using the proposed platform. During the session, HR and RR interval data acquired from the Polar H10 heart rate monitor were simultaneously recorded by both the proposed system and the manufacturer's application (Polar Flow).

The therapist remained in a separate building from the participant and interacted exclusively through the video call integrated into the platform.

Before the experimental session, each therapist completed a 10-minute familiarization session with the TUI, including training on treadmill control commands (speed and incline adjustment), physiological data monitoring, and video-based communication.

1) Treadmill Exercise Protocol: To simulate an exercise session with progressively increasing intensity, a submaximal protocol adapted from the Bruce protocol [16] was employed. The original protocol consists of progressive stages of treadmill speed and incline, starting at 2.7 km/h with a 10% incline and increasing by 1.3 km/h in speed and 2% in incline every 3 minutes until volitional exhaustion.

The Bruce protocol was selected because its standardized

progression allows the evaluation of multiple remote treadmill control commands throughout the exercise session. The protocol was adapted to ensure participant safety while assessing the functionality of the proposed system under controlled submaximal conditions. The target exercise intensity was defined as 70% of the estimated maximum HR (%HRmax), calculated using (1) [17]. The target HR was subsequently calculated using (2), where A denotes the participant's age in years.

$$HR_{\max} = 208 - 0.7A \tag{1}$$

$$HR_{\text{target}} = 0.70\, HR_{\max} \tag{2}$$

Before the session began, the therapist configured the participant's target %HRmax. The session consisted of 3-minute stages with progressive increases in treadmill speed and incline, as detailed in Table I.

TABLE I
SUBMAXIMAL EXERCISE PROTOCOL ADAPTED FROM THE BRUCE PROTOCOL

Stage	Incline (%)	Speed (km/h)
1	5	2.7
2	7	4.0
3	9	5.3
4	11	6.6
5	13	7.9
6	15	9.2

When the participant reached the target %HRmax, a visual alert was displayed on the Therapist User Interface (TUI), prompting the therapist to manually activate the recovery mode via a dedicated button. Once activated by the therapist, the system assumed control of the treadmill and executed a 3-minute cool-down period by linearly reducing the maximum speed and incline achieved during exercise in six equal steps at 30-second intervals until the treadmill came to a complete stop. For safety purposes, a researcher remained present in the room with the exercising participant to act as a local supervisor and engage the treadmill's physical emergency stop if necessary. Furthermore, the system was designed to halt commands in the event of severe network instability or Bluetooth disconnection.

2) Procedures: Immediately after the experimental session, data regarding system usability and user perceptions were collected [18]. First, therapists completed the System Usability Scale (SUS), a validated questionnaire that assesses perceived system usability on a scale ranging from 0 (worst imaginable) to 100 (best imaginable) [19], [20]. Finally, therapists and participants individually answered the following open-ended questions:

1) What do you consider the main advantage of using this system?

2) Which additional functionality would have been useful but was not available in the system?

3) In your opinion, what could be improved to facilitate or enhance the use of the system?

D. Data Analysis

Responses to the SUS questionnaires were analyzed descriptively. SUS scores were calculated according to the standard scoring procedure [19] and summarized using descriptive statistics (mean ± standard deviation). The overall mean SUS score was subsequently interpreted according to established adjective rating categories [20].

Responses to the open-ended questions provided by therapists and participants were transcribed, organized, and subjected to a descriptive thematic analysis. Initially, the responses were reviewed by two independent researchers to identify recurring ideas and meaningful units. Any divergences in thematic categorization were resolved through consensus discussions. Similar responses were subsequently grouped into thematic categories representing the main perceptions regarding the proposed system, including perceived advantages, desired functionalities, and suggested improvements. The identified themes were then summarized descriptively and used to support the interpretation of the quantitative findings [18], [21].

Heart rate (HR) recordings obtained throughout the experimental session from both the proposed system and the manufacturer's application (Polar Flow) were exported as .csv files, synchronized, and paired for direct comparison. Using the NumPy and Pandas libraries in Python, the mean absolute error (MAE), mean absolute percentage error (MAPE), Pearson's correlation coefficient, and the intraclass correlation coefficient (ICC) were calculated to quantify the measurement error, linear association, and agreement between the two systems. Prior to comparison, the data streams from both sources were synchronized using timestamp alignment and resampled at 1 Hz. The ICC was computed based on a two-way mixed-effects model for absolute agreement. To determine the overall system reliability (Global ICC), the time-series data from all 13 sessions were concatenated into a single pooled dataset prior to computation, thereby capturing the full variance across all participants.

III. RESULTS

Importantly, all 13 sessions were completed successfully, and no adverse events or unexpected technical failures compromising user safety occurred during the experimental protocol. On average, the exercising subjects reached the 70% HRmax target at the beginning of Stage 4 (approximately 9 to 10 minutes of active exercise). For two participants, the 70% HRmax target was reached during Stage 3, whereas one participant reached the target during Stage 5. Including the 3-minute automated cool-down period, the total active session duration averaged 12 to 13 minutes per dyad.

A. Usability Evaluation and System Perception

The usability evaluation conducted with the therapist group revealed high perceived usability of the proposed system. The mean SUS score was 95.19 ± 4.26, corresponding to the "best imaginable" adjective category. Similarly, the qualitative analysis of the open-ended responses identified ease of use, remote

accessibility, and real-time monitoring as important perceived strengths of the proposed system. Suggestions for improving the system, obtained from the second and third open-ended questions, were grouped into four thematic categories, which are summarized in Table II. To illustrate these themes, one therapist [T4] emphasized the need for complementary clinical tools: *"I would include a subjective perceived exertion scale [Borg] and an annotation field to record potential patient complaints or pain."* Another therapist [T12] highlighted the desire for enhanced automation and clearer visual feedback: *"It would be helpful to have pre-programmed tests with automatic increments, and to make it visually clearer when the recovery phase has been triggered."*

The qualitative analysis of the participants' responses to the first open-ended question identified the possibility of receiving rehabilitation remotely as an important perceived advantage of the system. Illustrating the sense of safety reported by the participants, one participant [P8] noted: *"Even performing the experiment alone, I felt well-monitored knowing the therapist was talking to me remotely."* Another participant [P5] emphasized the potential to overcome geographic barriers: *"The specialist can control physiological variables and the treadmill remotely, allowing me to continue my treatment even if I move to another city or country."* Participants also highlighted the convenience of performing rehabilitation sessions at home and the integration of HR monitoring, which contributed to their reported sense of safety during exercise. Suggestions for improving the proposed system were grouped into three thematic categories and are summarized in Table III.

B. Agreement of Physiological Data Acquisition

Session-specific agreement metrics are summarized in Table IV.

TABLE IV
AGREEMENT AND RELIABILITY METRICS BY PARTICIPANT SESSION ($n = 13$)

Session	MAE [bpm]	MAPE [%]	Correlation	ICC
1	0.27	0.29	0.999	0.999
2	0.31	0.28	0.999	0.999
3	0.60	0.59	0.985	0.985
4	0.43	0.41	0.996	0.996
5	0.26	0.27	0.999	0.999
6	0.72	0.60	0.972	0.972
7	0.25	0.22	0.999	0.999
8	0.35	0.34	0.997	0.997
9	0.50	0.48	0.992	0.992
10	0.38	0.33	0.998	0.998
11	0.39	0.36	0.998	0.998
12	0.43	0.45	0.996	0.996
13	0.61	0.52	0.982	0.982

The agreement of the heart rate (HR) measurements was evaluated by comparing the data acquired by the proposed platform with those recorded by the manufacturer's application (Polar Flow), both originating from the same Polar H10 sensor. This comparison serves to validate the integrity of the data transmission and synchronization pipeline rather than independent physiological accuracy. The proposed system achieved a mean absolute error (MAE) of 0.41 bpm and a mean absolute percentage error (MAPE) of 0.39%. Pearson's correlation coefficient between the measurements was 0.998,

TABLE II
THEMATIC ANALYSIS OF THERAPISTS' QUALITATIVE FEEDBACK

Analysis Dimension	Suggested Improvements
Interface design and information display	Therapists emphasized the need to improve the clarity and visibility of physiological variables and treadmill parameters (e.g., speed, incline, age, and heart rate), as well as recovery controls. They also suggested improving the participant's video display by providing more suitable viewing angles to facilitate clinical observation during exercise.
System functionality	Therapists highlighted the need for features that increase autonomy and safety during exercise sessions, including programmable automated protocols for speed and incline progression and simplified manual controls, with greater emphasis on the emergency stop function.
Clinical monitoring tools	Therapists identified the need to expand the system's clinical assessment capabilities by incorporating measures of perceived exertion, pain, emotional state, and exercise readiness, together with fields for recording participant-reported events. They also suggested automatic logging of physiological variables and data export functionality to support external analyses.
User support	Therapists emphasized the importance of resources that facilitate familiarization with the system, such as video tutorials and instructional materials that guide therapists through the execution of exercise sessions.

TABLE III
THEMATIC ANALYSIS OF PARTICIPANTS' QUALITATIVE FEEDBACK

Analysis Dimension	Suggested Improvements
Communication and interaction with the therapist	Participants emphasized the need for greater visual proximity and real-time interaction with the therapist. Suggested improvements included optimizing camera positioning and providing additional viewing options to strengthen the sense of presence and facilitate understanding of the therapist's instructions.
User interface and visual feedback	Participants highlighted the importance of improving the clarity of the user interface by incorporating visual indicators of exercise duration and stage transitions, previewing upcoming speed and incline adjustments, and enhancing the visibility of monitored physiological variables. Additional visual elements were also suggested to increase engagement during the rehabilitation session.
Physiological monitoring	Participants expressed interest in expanding the physiological information provided by the system, including metrics such as oxygen consumption and movement quality. They also suggested exploring monitoring alternatives that reduce or eliminate the need for wearable sensors.

while the ICC was 0.9982, indicating excellent reliability (ICC $\geq$ 0.90) [22].

IV. Discussion

This section discusses the study's findings in light of the existing literature, addressing the two primary axes of the investigation. First, the usability and user experience of the platform are analyzed based on the SUS scores and the qualitative interpretation of the feedback provided by therapists and participants. Subsequently, the results regarding the technical validation of the system's reliability in HR data transmission are examined by comparing its performance against the manufacturer's application. Finally, the practical implications of these findings, along with the study's limitations and directions for future research, are explored.

A. Usability and User Perception

The findings of this study support the operational viability of a bidirectional telerehabilitation system that combines high perceived usability with high-fidelity heart-rate data acquisition. The observed SUS score (95.19 ± 4.26), which falls within the "best imaginable" adjective category [20], indicates very favorable perceived usability among therapists after the standardized familiarization period. This is particularly relevant in telerehabilitation, where technological complexity and poorly designed interfaces have been identified as barriers to clinical adoption [13], [23]. In this context, the present findings indicate that the proposed interface was perceived as highly usable under the evaluated conditions. However, given the relatively small sample and controlled experimental setting, these results should be interpreted as preliminary and should not be generalized to broader clinical populations or practice settings.

The descriptive qualitative analysis complements the quantitative SUS findings by providing additional insight into potential improvements to the user experience. As presented in Table II, the qualitative themes highlight the necessity for improved layout clarity, automated clinical protocols (e.g., integrated timers, pre-programmed submaximal tests, and automated recovery transitions), complementary clinical assessment tools (e.g., Borg Rating of Perceived Exertion scale and standardized clinical reporting), and structured user support (e.g., video tutorials). These findings align with recent telehealth literature emphasizing that therapists require systems that not only control physical hardware but also automate routine clinical tasks to reduce cognitive load during remote supervision [11], [14].

Similarly, the participant feedback summarized in Table III reinforces the psychological and technical requirements of remote exercise. Participants emphasized the need for continuous visual proximity with the specialist to maintain a sense of clinical presence and safety, a well-documented success factor in remote patient monitoring systems [8]. Furthermore, their requests for predictive interface cues regarding speed and incline transitions, alongside an interest in advanced physiological metrics (e.g., estimated maximum oxygen consumption and motion analysis without additional wearable sensors), reflect a growing patient expectation for interactive, comprehensive, and non-intrusive digital health platforms. Collectively, these qualitative insights provide a valuable roadmap for future system iterations, aiming to transition the platform from merely functional to fully tailored for clinical practice.

B. Agreement of Physiological Data

Beyond interface usability, the clinical viability of a remote exercise platform relies fundamentally on the reliability of its data transmission, as physiological consistency dictates therapeutic decision-making. The comparison between the proposed platform and the manufacturer's application revealed exceptionally low discrepancy, a near-perfect linear correlation, and elevated reliability across all experimental sessions (see Table IV). Since both data streams originate from the exact same hardware source, these metrics are inherently expected to be nearly identical, and any minor differences observed are likely attributable to synchronization artifacts. Ultimately, this validation step primarily serves to confirm that the platform's data acquisition and transmission pipelines function correctly. These results indicate that the proposed platform tracks real-time HR variations with high fidelity, enabling timely interventions, such as triggering safety recovery modes, and logging exercise load responses.

Even the lowest reliability observed in an individual session (ICC $= 0.9720$) remained well above the threshold for excellent clinical agreement [22]. This high degree of agreement is critical for human-in-the-loop telerehabilitation systems, where latency or signal inaccuracies can compromise patient safety or lead to inappropriate treadmill speed and incline adjustments [12]. Nevertheless, it is important to emphasize that these metrics reflect a specific hardware configuration (treadmill and HR monitor models) operating within a controlled environment. Consequently, the system's performance must be further validated under real-world conditions, particularly concerning potential Bluetooth signal interference and packet loss, as well as among diverse clinical populations.

C. Practical Implications, Limitations, and Future Work

The proposed architecture addresses a technological gap in the current telerehabilitation literature. While previous studies have demonstrated the use of video-based physiotherapy and remote physiological monitoring [8], [11], many existing platforms emphasize unidirectional monitoring workflows, requiring therapists to verbally instruct patients to manually adjust their own exercise equipment [12]. In contrast, the present study introduces a bidirectional, human-in-the-loop framework. By coupling real-time physiological telemetry (e.g., synchronous HR) with active hardware actuation (treadmill speed and incline control), the system enables therapists to actively modulate exercise intensity remotely, approximating the therapist-mediated control available in supervised in-person settings. The combination of high perceived usability and high agreement between the HR acquisition pathways supports the platform's potential to facilitate remote care through structured

exercise protocols, %HRmax-based alerts, and automated recovery routines. From a clinical engineering perspective, the qualitative feedback (Tables II and III) delineates clear developmental priorities: implementing visual adjustment previews, enhancing transition clarity, and integrating subjective clinical assessment scales. Implementing these improvements may further enhance the system's technological maturity and support subsequent evaluation in clinical settings.

However, the scope and limitations of the present study must be properly considered. The primary constraints include the relatively small sample size; the use of healthy young adults rather than the target population of individuals with musculoskeletal or cardiopulmonary limitations; the use of a single hardware configuration (treadmill and HR sensor); testing within a controlled laboratory environment; the absence of functional clinical outcomes, longitudinal adherence tracking, and cost-effectiveness analyses; and the lack of quantitative network-performance metrics (e.g., command latency, packet loss, and command execution success rates). While these methodological choices were appropriate for an initial technical and usability evaluation, the generalizability of the present findings to broader clinical contexts remains limited.

Future research should extend the evaluation of the platform beyond the laboratory to assess its technical performance, feasibility, and safety in more representative clinical and home-based environments. Subsequent studies should include older adults, individuals with varying levels of familiarity with digital technologies, and patients with musculoskeletal or cardiopulmonary conditions. Transitioning from controlled laboratory settings to real-world environments will also require systematic evaluation of operational robustness, including Bluetooth signal variability, communication latency, packet loss, and environmental interference, together with diagnostic logging of communication failures.

Beyond technical robustness, it is vital that research moves toward evaluating the platform's long-term clinical impact. Prospective studies are needed to quantify longitudinal patient adherence, safety profiles, and functional recovery markers, as well as to investigate how the user experience evolves over extended rehabilitation cycles. This transition will be supported by ongoing product evolution, integrating the user-centered refinements identified in this study—such as protocol automation, visual transition previews, and the integration of standardized clinical tools like the Borg scale.

Finally, for the system to achieve its full potential in clinical practice, future development will focus on interoperability and auditing. Developing standardized data export protocols and comprehensive command-execution logs will be essential to facilitate seamless integration with Electronic Health Records (EHRs) and align the platform with the rigorous quality assurance standards required for professional healthcare deployment.

V. Conclusion

The bidirectional telerehabilitation platform developed in this study demonstrated high perceived usability and preliminary technical viability for synchronous physical therapy. Empirical results confirmed the integrity of real-time heart rate data transmission from a shared sensor source and the perceived usability of the interface for therapists. Importantly, 13 independent sessions were successfully completed without adverse events, demonstrating that therapists could effectively utilize the remote control framework to actively modulate exercise intensity.

While these findings are promising, the study evaluates a first-generation prototype in a controlled environment with healthy participants. Clinical safety, network performance under unstable conditions, and application in target patient populations require specific future studies. Ultimately, this work provides foundational evidence for a bidirectional remote control framework, offering a technological step toward more accessible rehabilitation services.

References

[1] A. Cieza, K. Causey, K. Kamenov, S. W. Hanson, S. Chatterji, and T. Vos, "Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10267, pp. 2006–2017, 2020.

[2] World Health Organization, "Musculoskeletal health," https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions, 2022, accessed: 2026-07-15.

[3] A. I. Corregidor-Sánchez, A. Segura-Fragoso, M. Rodríguez-Hernández, J. J. Criado-Alvarez, J. González-Gonzalez, and B. Polonio-López, "Can exergames contribute to improving walking capacity in older adults? a systematic review and meta-analysis," *Maturitas*, vol. 132, pp. 40–48, 2020.

[4] Y. Huang, Y. Li, H. Pan, and L. Han, "Global, regional, and national burden of neurological disorders in 204 countries and territories worldwide," *Journal of Global Health*, vol. 13, p. 04160, 2023.

[5] A. A. Kurnianto, S. Kovács, N. Ágnes, and P. Kumar, "Economic evaluations of rehabilitation interventions: A scoping review with implications for return to work programs," *Healthcare*, vol. 13, no. 10, p. 1152, 2025.

[6] D. Magnusson *et al.*, "Addressing health care access disparities through a public health approach to physical therapist practice," *Physical Therapy*, vol. 104, no. 5, 2024.

[7] A. K. Shambushankar, J. Jose, S. Gnanasekaran, and G. Kaur, "Cost-effectiveness of telerehabilitation compared to traditional in-person rehabilitation: A systematic review and meta-analysis," *Cureus*, vol. 17, no. 2, p. e79028, 2025.

[8] M. A. Havran and D. E. Bidelspach, "Virtual physical therapy and telerehabilitation," *Physical Medicine and Rehabilitation Clinics of North America*, vol. 32, no. 2, pp. 419–428, 2021.

[9] M. Grimmer, R. Riener, C. J. Walsh, and A. Seyfarth, "Mobility related physical and functional losses due to aging and disease - a motivation for lower limb exoskeletons," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, no. 1, p. 2, 2019.

[10] M. Kim and F. Kaneko, "Virtual reality-based gait rehabilitation intervention for stroke individuals: a scoping review," *Journal of Exercise Rehabilitation*, vol. 19, no. 2, pp. 95–104, 2023.

[11] C. Camden, G. Pratte, F. Fallon, M. Couture, J. Berbari, M. Tousignant, and E. Couture, "Diversity of practices in telerehabilitation for children with disabilities and effective intervention characteristics: results from a systematic review," *Disability and Rehabilitation*, vol. 42, no. 24, pp. 3424–3436, 2020.

[12] J. A. Moral-Munoz, W. Zhang, M. J. Cobo, E. Herrera-Viedma, and D. B. Kaber, "Smartphone-based systems for physical rehabilitation applications: A systematic review," *Assistive Technology*, vol. 33, no. 4, pp. 223–236, 2021.

[13] R. Argent, A. Daly, and B. Caulfield, "Patient involvement in the design of digital health interventions: A review of the literature," *Journal of Medical Internet Research*, vol. 20, no. 11, p. e10588, 2018.

[14] L. Simpson *et al.*, "The role of digital health in physiotherapy: A scoping review," *Journal of Telemedicine and Telecare*, vol. 27, no. 9, pp. 533–544, 2021.

[15] J. Schwartz, P. Oh, M. Y. Takito, B. Saunders, E. Dolan, E. Franchini, R. E. Rhodes, S. S. D. Bredin, J. P. Coelho, P. dos Santos, M. Mazzuco, and D. E. R. Warburton, "Translation, cultural adaptation, and reproducibility of the physical activity readiness questionnaire for everyone (par-q+): The brazilian portuguese version," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 712696, 2021.

[16] R. A. Bruce, F. Kusumi, and D. Hosmer, "Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease," *American Heart Journal*, vol. 85, no. 4, pp. 546–562, 1973.

[17] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *Journal of the American College of Cardiology*, vol. 37, no. 1, pp. 153–156, 2001.

[18] M. Sandelowski, "Whatever happened to qualitative description?" *Research in Nursing & Health*, vol. 23, no. 4, pp. 334–340, 2000.

[19] J. Brooke, "Sus - a quick and dirty usability scale," *Usability Evaluation in Industry*, vol. 189, p. 194, 1996.

[20] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.

[21] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[22] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[23] I. J. Borges do Nascimento, H. Abdulazeem, L. T. Vasanthan, E. Z. Martinez, M. L. Zucoloto, L. Østengaard, N. Azzopardi-Muscat, T. Zapata, and D. Novillo-Ortiz, "Barriers and facilitators to utilizing digital health technologies by healthcare professionals," *NPJ Digital Medicine*, vol. 6, no. 1, p. 161, 2023.

Development and Preliminary Technical Evaluation of a Wearable System Based on Inertial Sensors for Postural Monitoring and Vibrotactile Biofeedback

1st Eduardo Felipe Borges Fernandes
Biomedical Engineering Dept. (DEB)
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
eduardo.felipe.120@ufrn.edu.br

2nd Lissa Maria Nascimento de Andrade
Biomedical Engineering Dept. (DEB)
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
lissa.andrade.136@ufrn.edu.br

3rd Gabrielly Blasques Oliveira Albuquerque
Biomedical Engineering Dept. (DEB)
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
https://orcid.org/0009-0006-9460-7058

4th Mateus Mendes Soares de Oliveira
Biomedical Engineering Dept. (DEB)
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
mateus.oliveira.106@ufrn.edu.br

5th Luciana de Andrade Mendes
Biomedical Engineering Dept. (DEB)
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
luciana.mendes@ufrn.edu.br

6th Severino Peixoto Nunes Netto
Biomedical Engineering Dept. (DEB)
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
severino.netto@ufrn.edu.br

***Abstract*—Musculoskeletal Disorders (MSDs), including Work-Related Musculoskeletal Disorders (WMSDs) and Adolescent Idiopathic Scoliosis (AIS), represent a public health concern with a strong functional impact. Postural self-correction strategies utilizing biofeedback represent a promising approach to assist in the reduction of biomechanical overload. This study proposes the development and preliminary technical evaluation of a low-cost wearable system focused on the continuous monitoring of lateral trunk inclination (Z-axis) and real-time vibrotactile biofeedback. The architecture comprises two MPU-6500 IMUs and an Arduino Nano, utilizing the Digital Motion Processor (DMP) for iterative calibration and sensor fusion into quaternions that trigger directional alerts. A preliminary static assay identified a drift of 0.14 degrees/min in the yaw axis, while extended assessments of up to 240 minutes quantified a consistent linear accumulation of 0.27 degrees/min. This drift was mitigated by a linear temporal compensation algorithm embedded in the firmware, ensuring long-term stability. During dynamic validation against Kinovea, the system demonstrated a strong correlation (r = 0.90; R2 = 0.81), a mean absolute error of 0.61 degrees, an RMSE of 0.69 degrees, and a bias of -0.48 degrees via Bland-Altman analysis. The results confirm adequate performance for postural tracking, establishing a promising and low-cost solution to promote active self-correction, with future improvements aimed at miniaturization and clinical validation.**

***Index Terms*—Biofeedback, postural monitoring, ergonomics, scoliosis, inertial sensors.**

I. Introduction

According to the World Health Organization (WHO), Musculoskeletal Disorders (MSDs) encompass over 150 conditions affecting the locomotor system, including spinal conditions such as scoliosis, kyphosis, and lordosis, as well as Work-Related Musculoskeletal Disorders (WMSDs), frequently associated with repetitive movements [1], [2]. It is estimated that MSDs affect approximately 1.71 billion people worldwide, regardless of age, sex, or sociodemographic status, potentially causing temporary or permanent functional limitations and reducing social participation [3]–[5].

MSDs can alter body alignment and modify the distribution of external forces and gravity, leading to muscular compensations that result in excessive shortening or lengthening of muscles. This process reduces the body's mechanical efficiency and increases the overload on joints and intervertebral discs, favoring the onset of chronic pain, premature wear, and various postural alterations such as hyperlordosis, hyperkyphosis, scoliosis, and gait abnormalities [6]. Among these conditions, Adolescent Idiopathic Scoliosis (AIS) stands out; it is a three-dimensional structural deformity defined by the Scoliosis Research Society (SRS) as a structured lateral curvature of the spine equal to or greater than 10° by the Cobb angle, of unknown etiology, with onset between 10 and 18 years of age [7]–[9].

Consequently, MSDs have significant repercussions on quality of life, functional capacity, and costs for healthcare systems and companies, especially due to occupational absences. These conditions can affect workers across various roles and sectors, such as executives and taxi drivers who remain in the same

position for extended periods, factory workers exposed to repetitive movements or the handling of suspended loads, and nursing professionals subjected to long shifts [10], [11].

Treatments for MSDs vary according to the affected locomotor structure and the epidemiological characteristics of each condition. For WMSDs, management is predominantly conservative, based on ergonomic measures, postural correction, regular breaks, and the reorganization of work tasks [6]. For AIS, treatment includes specific physiotherapeutic exercises and, for curvatures between 20° and 45°, the use of orthopedic braces, with surgery indicated for more severe cases [9], [10]. Despite the specificities of each approach, postural self-correction constitutes a common and fundamental strategy in MSD management, as it promotes biomechanical alignment and reduces overload on soft tissues and joints [12]. This process can be enhanced by real-time biofeedback, which increases body awareness and encourages active postural adjustments during functional and occupational activities [13], [14].

As a strategy for implementing postural self-correction, wearable technologies and electromyographic monitoring methods have been shown to significantly reduce ergonomic risks, enabling individuals to minimize chronic muscle tension and prevent the aggravation of painful conditions through constant postural vigilance [13]–[15]. This approach is grounded in principles of biomechanics and motor control, which explain how the body perceives and corrects postural deviations in response to received stimuli.

In this context, motor learning and postural re-education depend on the nervous system's capacity to process sensory stimuli and dynamically adjust body alignment [7]. Thus, vibrotactile biofeedback can serve as a real-time haptic stimulation tool capable of promoting active self-correction, one of the pillars of physiotherapeutic approaches such as the Schroth Method. Studies demonstrate that biofeedback improves motor control, particularly in individuals with postural instability, provided that the stimuli are subtle and low-intensity, and function as reminders to encourage postural adjustments without causing sensory fatigue or discomfort [16]. Furthermore, strategies based on controlling lateral trunk inclination guided by real-time biofeedback have been validated in biomechanical literature for motor retraining and postural adjustment in the frontal plane [17], [18], reinforcing the relevance of continuously monitoring this kinematic parameter during the postural re-education process.

Although optical motion capture systems, such as Qualisys, are considered the gold standard for postural assessment, their high cost and the need for laboratory environments limit their application in continuous monitoring. Alternatively, Inertial Measurement Units (IMUs) can estimate trunk orientation in real-time using accelerometers and gyroscopes, making postural tracking more accessible and applicable to everyday activities. Combined with biofeedback, these technologies facilitate postural self-correction, promoting greater body awareness, reducing muscle fatigue, and preventing WMSDs. In this context, this study proposes the development of a low-cost wearable system based on IMUs and real-time vibrotactile biofeedback. This paper presents a preliminary technical validation focused strictly on tracking lateral trunk inclination (Z-axis). Establishing robust tracking accuracy for this specific axis provides an essential technological foundation for future clinical trials, which will evaluate the system's true efficacy in promoting active postural self-correction to prevent musculoskeletal disorders.

II. Methodology

The methodology proposed for the development and validation of the wearable postural monitoring system was structured into seven consecutive stages: (A) system architecture and signal processing; (B) iterative calibration algorithm and quaternion fusion; (C) biofeedback and postural signaling; (D) anatomical positioning and mechanical fixation protocol; (E) static stability assay (drift); (F) dynamic kinematic validation protocol; and (G) statistical analysis.

A. System Architecture and Signal Processing

The developed postural monitoring system prototype is based on inertial kinematics. The hardware consists of two low-cost Inertial Measurement Units (IMUs), model MPU-6500 (InvenSense Inc.), integrated with an ATmega328P microcontroller (Arduino Nano). Communication between the sensors and the microcontroller is performed via a shared I^2C bus, configured to operate in Fast Mode (400 kHz). To enable simultaneous reading of both devices without addressing conflicts, the AD0 logic selection pin was used. The sensor corresponding to the left scapula was configured with the address 0x68 (AD0 connected to 3.3V), while the sensor positioned on the right scapula was assigned the address 0x69 (AD0 pulled low - GND).

Given the mathematical processing limitations of the 8-bit microcontroller and to reduce computational delay, we opted to use the native Digital Motion Processor (DMP) integrated into the IMUs. The DMP's native sensor fusion provides spatial orientation directly in unit quaternions ($q = [w, x, y, z]$), eliminating static integration errors and occurrences of mathematical singularities (gimbal lock). To access this functionality, the firmware utilizes the open-source `MPU6050_6Axis_MotionApps20` library integrated into the I2Cdev architecture [19], acting as a Hardware Abstraction Layer (HAL) for extracting packets from the DMP's FIFO buffer. Complementing this layer, a strict 3 ms timeout was implemented for I^2C transactions, alongside an asynchronous routine to manage buffers and drain the FIFOs independently, preventing permanent bus lockups caused by transmission failures and avoiding buffer overflows during rapid movements.

B. Iterative Calibration Algorithm and Quaternion Fusion

To mitigate the static offset errors characteristic of low-cost MEMS sensors, an iterative calibration algorithm was implemented upon initialization, consisting of 6 consecutive moving average filtering cycles over a block of 500 static samples (with a 3 ms sampling interval). At each iteration,

the system quantifies the remaining scalar residual from the previous cycle and uses it to incrementally correct the IMU's internal offset registers. Specifically for the MPU-6500 accelerometer, the least significant bit (Bit 0) of the offset register functions as an internal hardware flag that must be preserved to avoid reading corruption. Therefore, a bitwise masking operation was implemented for the cumulative adjustments, as demonstrated in (1):

$$\text{Off}_{\text{new}} = \left[\left(\text{Off}_{\text{old}} - \frac{\text{Bias}_{\text{res}}}{k}\right) \wedge \sim 1\right] \vee (\text{Off}_{\text{old}} \wedge 1) \quad (1)$$

where $\wedge$ denotes the bitwise AND operator, $\vee$ is the bitwise OR, $\sim$ is the bitwise NOT, and the scale factor k is defined according to the chip manufacturer's register resolution guidelines [20]. Following calibration, the algorithm waits an additional 8 seconds, discarding the vectors generated during this interval to mitigate initial drift effects.

After stabilization, the neutral posture reference is established in an upright standing posture: upon triggering an external mechanical button (pin 4), the firmware calculates the arithmetic mean and geometric renormalization of 30 consecutive quaternion packets, establishing the reference quaternions for each sensor (q_{RefL} and q_{RefR}). During continuous monitoring, each new rotation (q_{Current}) is converted into relative deviation quaternions (q_δ) using the Hamilton product with the conjugate of the reference quaternion:

$$q_\delta = q_{\text{Ref}}^{-1} \otimes q_{\text{Current}} \quad \text{where} \quad q_{\text{Ref}}^{-1} = \text{conj}(q_{\text{Ref}}) \quad (2)$$

From q_δ, the yaw Euler angle (rotation around the Z-axis) is extracted, which, in the present system, represents the lateral inclination of each scapula in the frontal plane. Based on this angle, two postural monitoring parameters are continuously calculated: the lateral scapular asymmetry given by (3), and the global trunk inclination, given by (4):

$$\alpha = \theta_L - \theta_R \quad (3)$$

$$\theta_{\text{global}} = \frac{\theta_L + \theta_R}{2} \quad (4)$$

The variable θ_{global} was adopted as the primary comparative variable in the validation assays described in Section II-F, as it represents the synchronous lateral displacement of the thoracic block and avoids false negatives when both sensors tilt together in the same direction, a condition in which asymmetry (α) approaches zero.

C. Biofeedback and Postural Signaling

The firmware employs two independent postural alert thresholds: one for asymmetry ($|\alpha| > 4.0°$) and one for global inclination ($|\theta_{\text{global}}| > 6.0°$), the latter designed to capture scenarios where both shoulders lean simultaneously in the same direction. The prototype physically features a pair of vibrotactile microactuators. In the developed firmware version, a threshold violation triggers a signal in the serial monitor and activates the actuator corresponding to the direction of the deviation. For directional mapping, positive values on the local Z-axis were associated with leftward inclinations, while negative values were associated with rightward inclinations.

The algorithm was structured so that the initial motor activation occurs ipsilateral to the inclination. Upon detecting a threshold violation, three vibration pulses lasting 200 ms each are emitted, separated by 0.5 s intervals. If the improper posture persists, the cycle is repeated every 5 s, continuously re-evaluating the inclination direction at the beginning of each cycle to allow dynamic correction if the direction of the deviation reverses. Additionally, the algorithm monitors postural return in real-time, immediately interrupting the vibratory signal as soon as the angles return to the established tolerance zone.

It should be noted that the angular threshold values and temporal intervals used in this stage were initially defined to assess the prototype's functional viability, following a common approach in early investigations of biofeedback systems aimed at trunk postural adjustment [17]. The definitive and individualized parameterization will depend on factors to be investigated in future trials, including the determination of the optimal inclination threshold, as well as the intensity and exposure time to vibration to prevent sensory adaptation.

Finally, the definitive logic for tactile feedback activation remains open. Although findings by Lee et al. [21] demonstrate that proprioceptive stimulation induces a non-volitional postural shift in the direction of the stimulus (suggesting the use of contralateral feedback to induce a compensatory response), the literature indicates divergences between contralateral and ipsilateral approaches regarding learning rates and the cognitive demand imposed on the central nervous system [27], [28]. Therefore, the final device activation logic will be defined after comparative experimental testing with volunteers.

D. Anatomical Positioning and Fixation Protocol

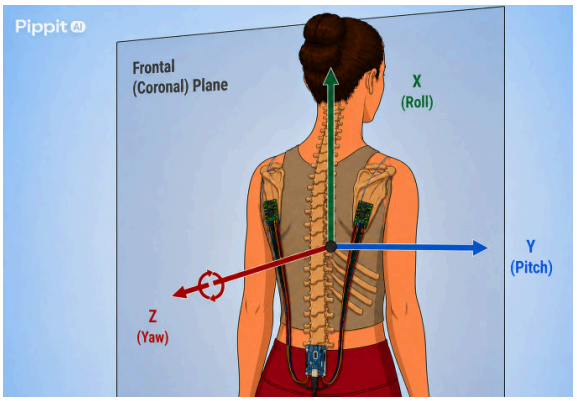


Fig. 1. Sensor mounting and referential axes. Illustration generated via Pippit AI and graphically edited using OpenAI [25].

The positioning of the sensors and kinematic markers was based on the International Society of Biomechanics (ISB) guidelines applied to the scapular complex [22]. The kinematic mapping employed two primary reference points, located

bilaterally at the inferior angles of the scapulae, serving as the placement site for the two IMUs. Complementary reflective markers were placed on the superior scapular angles as auxiliary visual guides for video analysis.

To enable sensor fixation without compromising the volunteer's mobility, the IMUs were attached parallel to the scapula, meaning the sensors were rotated relative to the standard reference frame (z-axis pointing downward). In the adopted assembly, illustrated in Fig. 1, the Z-axis was perpendicular to the frontal plane. This modification justifies the use of the yaw axis to record lateral trunk inclination, as presented in Section II-B. To minimize movement artifacts caused by soft tissue sliding during motion, the sensors were placed over a light elastic compression vest (sports bra) and secured with adhesive tape. The Arduino Nano module was centrally positioned in the lumbar region to allow symmetrical and safe cable routing between the two sensors.

E. Static Stability Assay (Drift)

Initially, a static stability assay was conducted to estimate the short-term firmware temporal drift rate in the absence of any real motion component. The system remained rigidly fixed to a flat, horizontal, and leveled surface under controlled room-temperature conditions for 15 minutes. This preliminary test, performed in duplicate, identified an average drift rate of $0.14^\circ/\text{min}$ in the yaw axis.

To investigate the integration drift inherent to the Z-axis during prolonged use and establish a definitive software-based solution, extended long-duration static assays were conducted. In addition to the preliminary 230-minute trials, robust new data collections comprising four independent 80-minute assays allowed for the consolidation of error modeling across different operation windows. The results of these data collections evidenced a highly predictable and consistent linear drift accumulation, consistent with the rate of approximately $0.27^\circ/\text{min}$ observed in the 240-minute evaluations. Based on this finding, a temporal compensation was implemented in the code. This constant error rate is continuously subtracted from the raw yaw angle output as a function of the time elapsed since the last reference capture, effectively neutralizing static temporal drift and enabling continuous monitoring throughout an entire work shift.

F. Dynamic Kinematic Validation Protocol

For kinematic validation in a dynamic regime, angular accuracy was evaluated against a reference method under real motion conditions. To achieve this, the sensor pair was attached to a volunteer, as detailed previously. At the start of the assay, the volunteer was asked to assume a neutral upright posture, standing as erect as possible. From this point, the reference posture was captured by triggering the hardware's mechanical button. Subsequently, the volunteer performed cycles of lateral trunk inclination to the left and right, with a free and variable amplitude, holding the position briefly to allow stable recording, and then returning to the initial neutral posture, defined as approximately 0.0°.

The assay consisted of collecting 12 dynamic postural samples, each recorded simultaneously by the hardware and via video. During this accuracy validation stage, the tactile vibration actuator on the vest was kept physically deactivated to ensure the mechanical alert stimulus did not induce reflexive changes in the volunteer's natural movement or posture, strictly isolating the device's intrinsic kinematic measurement error.

The 12 samples were recorded from a posterior view in the frontal plane under controlled lighting conditions. The generated videos were processed using the Kinovea biomechanical analysis software [23], adopted as a reference due to its widespread use in 2D kinematics studies. To establish the comparative baseline, the software's "angle relative to horizontal" tool was used, continuously calculating the angular deflection formed by the geometric line connecting the two markers on the inferior scapular angles relative to an idealized fixed horizontal reference line (0°), thereby replicating the measurement principle used by the sensor array.

To neutralize millimeter-scale imperfections in static postural alignment within the video image, data from both systems were processed in terms of Relative Angular Variation ($\Delta\theta$) based on the respective measured initial posture. For the Kinovea software, the geometric difference was calculated relative to the initial neutral image; for the microcontroller system, the global inclination variable extracted from the serial monitor was directly compared with the video measurements, as synchronous block movement of the thorax would cancel out the pure asymmetry parameter.

G. Statistical Analysis

The agreement between the proposed wearable system and Kinovea was validated using Pearson's linear correlation and Bland-Altman analysis of paired differences between the two methods [24], estimating the systematic error (bias, given by the mean of the differences) and the 95% limits of agreement (bias $\pm$ 1.96 $\times$ standard deviation of the differences), as presented in (5):

$$\text{ULA, LLA} = \text{bias} \pm (1.96 \cdot \sigma_{\text{diff}}) \tag{5}$$

where σ_{diff} denotes the standard deviation of the paired differences between the two methods, not the standard deviation of the amplitude means—a distinction that proved relevant for the correct interpretation of the results.

III. Results

The results obtained from the experimental evaluations of the wearable inertial prototype were divided into the analysis of long-duration static temporal stability (incorporating the 240-minute monitoring and the consolidated results of the four 80-minute collections), the functional verification of the biofeedback, and the dynamic kinematic validation against the Kinovea video analysis software [23].

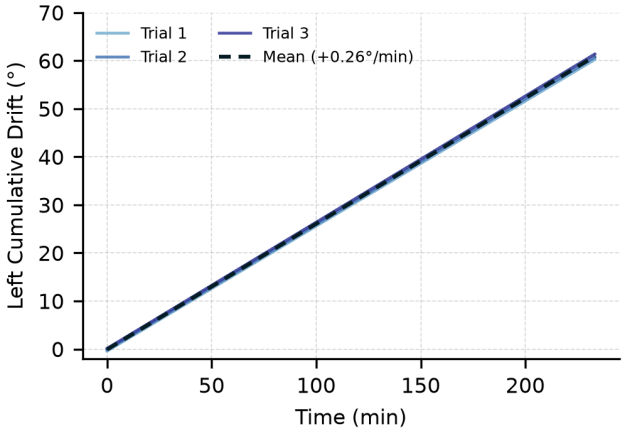


Fig. 2. Cumulative drift throughout 240-minute static assays for the left scapular sensor (individual assays and consolidated mean).

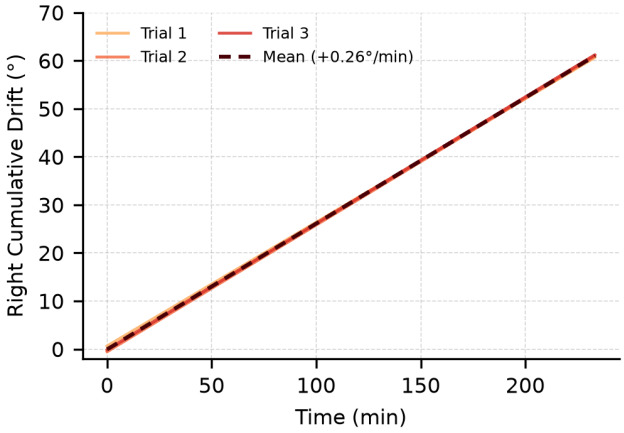


Fig. 3. Cumulative drift throughout 240-minute static assays for the right scapular sensor (individual assays and consolidated mean).

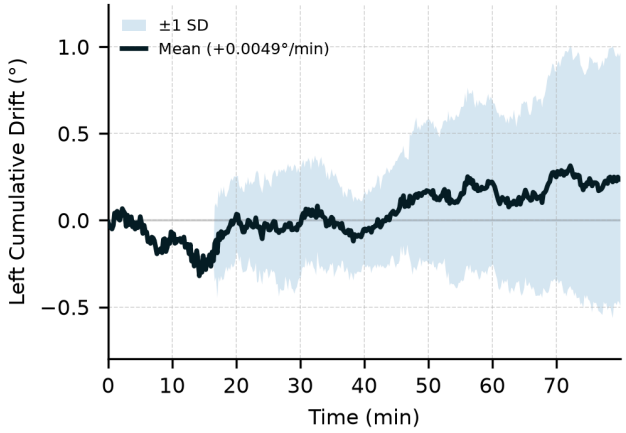


Fig. 4. Residual cumulative drift on the left after firmware compensation across the 4 validation assays (80 min), displaying the mean and ±1 SD.

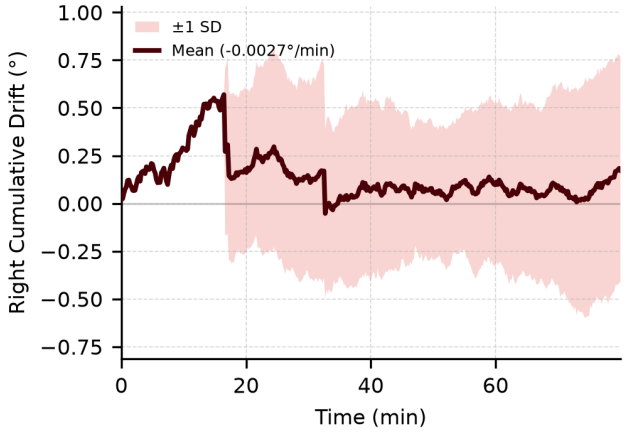


Fig. 5. Residual cumulative drift on the right after firmware compensation across the 4 validation assays (80 min), displaying the mean and ±1 SD.

A. Static Stability Analysis (Drift) and Compensation

According to the experimental protocol described in Section II-E, the preliminary short-duration static evaluation indicated an average drift rate of $0.14°/\text{min}$ in the yaw axis. Furthermore, extended static assays—encompassing long-duration monitoring of up to 240 minutes and the consolidation of four independent 80-minute collections—revealed a highly predictable and consistent linear error accumulation over the operation.

As illustrated in the consolidated panels of Figs. 2 and 3 (referring to the 240-minute measurements) and Figs. 4 and 5 (highlighting post-compensation performance across the 4 80-minute collections), linear regression fitting determined a mean uncorrected drift rate of approximately $0.27°/\text{min}$ for both sensors. The application of the corrective algorithm in the firmware drastically reduced this deviation, confining the mean residual error to near-zero levels ($\approx +0.0049°/\text{min}$ on the left and $\approx -0.0027°/\text{min}$ on the right) with narrow standard deviation bands (± 1 SD). This empirically observed consistency supports the robustness of the linear temporal compensation, enabling continuous and stable monitoring during prolonged shifts.

B. Biofeedback

In the bench tests described in Section II-C, the tactile signaling system behaved robustly. Each threshold violation triggered a sequence of three vibratory pulses accompanied by synchronous logging in the serial monitor. Upon returning to the postural neutrality zone, the system immediately interrupted the vibratory stimulus, confirming the proper operation of the real-time control loop.

C. Dynamic Kinematic Validation

After validating stability and mitigating drift, the system's performance was evaluated under a dynamic regime. Table I summarizes the 12 samples comparing the relative angular variations ($\Delta\theta$) obtained by Kinovea and the proposed system, with a mean inclination amplitude of $4.8° \pm 1.0°$.

Pearson's linear correlation between the two methods demonstrated strong agreement ($r = 0.90$; $R^2 = 0.81$), modeled by the linear regression equation described in (6):

$$\Delta\text{MPU} = 1.04 \cdot \Delta\text{Kinovea} + 0.28 \tag{6}$$

TABLE I
DYNAMIC KINEMATIC VALIDATION ACROSS THE 12 TESTED SAMPLES.

Sample	Δ Kinovea (°)	Δ MPU (°)	Difference (°)
1	3.1	4.3	-1.2
2	3.4	4.2	-0.8
3	4.3	4.2	0.1
4	3.8	3.3	0.5
5	5.9	6.7	-0.8
6	3.7	4.5	-0.8
7	5.2	6.2	-1.0
8	3.8	3.6	0.2
9	5.9	6.2	-0.3
10	4.2	4.5	-0.3
11	5.1	5.9	-0.8
12	6.0	6.5	-0.5

The Bland-Altman analysis, calculated according to (5), indicated a systematic bias of $-0.48°$, reflecting a slight tendency of the prototype to overestimate the angle compared to the optical reference. The 95% limits of agreement, derived from the standard deviation of the paired differences ($0.52°$), were established at ULA $= 0.55°$ and LLA $= -1.50°$.

Fig. 6 and Fig. 7 illustrate, respectively, the scatter plot with the linear correlation and the Bland-Altman diagram of the dynamic validation.

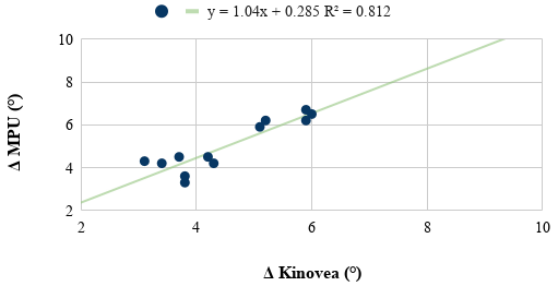


Fig. 6. Pearson's linear correlation analysis between the wearable system and the Kinovea software.

IV. DISCUSSION

Taken together, the conducted assays help to understand the performance and main sources of error of the proposed system. The mean absolute error in dynamic validation ($0.61°$; RMSE $= 0.69°$) was relatively low given the tested amplitude, and the bias of $-0.48°$ is comparable in magnitude to the combined effect of the gyroscope's resolution ($0.05°$) and the potential error associated with manual marking in Kinovea. This indicates that part of the difference stemmed from the visual reference method itself rather than intrinsic sensor errors.

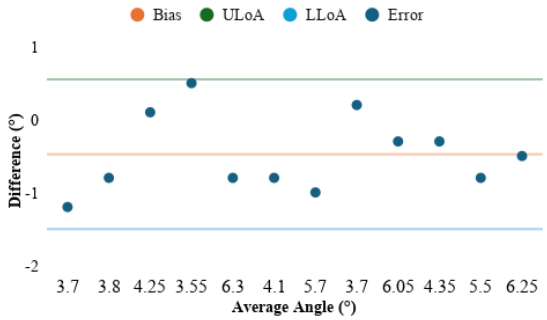


Fig. 7. Bland-Altman plot assessing kinematic agreement between the two methods.

Regarding temporal stability, the extended tests—including 240-minute monitoring and the four 80-minute collections—showed a predictable linear drift accumulation of about $0.27°/\text{min}$ in the yaw axis (Z-axis). It is worth noting that the study strictly focused on the Z-axis precisely because it represents the greatest challenge in 6-degree-of-freedom (6-DoF) sensors lacking a magnetometer: while roll and pitch are naturally corrected by gravity, the Z-axis lacks an absolute reference and accumulates error over time. The linear correction embedded in the firmware effectively solved this issue, dropping residual drift to near-zero values ($+0.0049°/\text{min}$ and $-0.0027°/\text{min}$) and allowing for prolonged device use.

Concerning tactile biofeedback, the bench tests confirmed that the control loop operates satisfactorily, although the exposed wiring still slightly hinders portability. The next practical step will be migrating to a custom Printed Circuit Board (PCB) and integrating a compact battery into the vest. In the future, upgrading to 9-axis sensors (such as the MPU-9250 with a magnetometer) will not only provide an absolute reference for the Z-axis via the Earth's magnetic field but will also enable complete and simultaneous three-dimensional tracking across the three anatomical planes of motion.

Finally, it is important to emphasize that this study is strictly technical and preliminary, focused on validating hardware and tracking precision. As the test was conducted with a single individual as a proof of concept, the main objective was to isolate errors and understand drift behavior. Trials with volunteers and practical biofeedback testing require prior approval from a Research Ethics Committee (REC). Therefore, this work does not have the clinical purpose of proving postural re-education, but rather to provide a solid technological foundation so that future clinical and ergonomic studies can be conducted under appropriate safety conditions.

V. IMPLEMENTATION CONSIDERATIONS

The proposed system was designed to transition from experimental validation to a dedicated commercial application. Its modular architecture based on MEMS inertial sensors allows the evolution from a bench prototype to low-power, highly portable devices without compromising biomechanical tracking fidelity.

A. Regulatory Alignment and Risk Management

To support future commercial evolution, the prototype's design was guided by medical device regulatory guidelines. The risk classification structures and Quality Management System (QMS) frameworks referenced the Brazilian Health Regulatory Agency (ANVISA) guidelines (RDC No. 751/2022 and RDC No. 665/2022). Risk identification followed the principles of the ISO 14971 standard, prioritizing firmware stability and physical circuit isolation. Furthermore, general electrical safety guidelines (IEC 60601 family) and usability engineering (IEC 62366) were used as design references for hardware layout and biofeedback delivery. For future stages involving continuous data collection in clinical cohorts, the firmware architecture will incorporate data encryption and privacy protocols in compliance with regulatory frameworks (such as the LGPD, the European MDR, and US FDA requirements).

B. Scalability and Occupational Integration

The architecture was structured in a scalable manner to allow integration into Internet of Medical Things (IoMT) ecosystems. The sensor data stream generates structured angular deviations in time series, making the dataset suitable for future Bluetooth Low Energy (BLE) transmission to mobile platforms and with Machine Learning classifiers for automatic activity recognition and adaptive postural threshold definition.

In occupational health, the device provides an objective tool aligned with ergonomic standards (such as the Brazilian Regulatory Standard NR-17). It enables integration into corporate Risk Management Programs (PGR) and Occupational Health Medical Control Programs (PCMSO) by generating continuous quantitative metrics—such as cumulative exposure time in non-neutral postures and biofeedback compliance rates—supporting evidence-based ergonomic interventions in clinical and workplace environments.

VI. Conclusion

Following systematic prototype validation, we conclude that the proposed architecture—composed of two MPU-6500 IMUs, an iterative calibration algorithm, and embedded processing via quaternion fusion—demonstrated satisfactory performance in estimating lateral trunk inclination (Z-axis / yaw). The system showed a strong correlation with the Kinovea optical reference method ($r = 0.90$), a mean absolute error of $0.61°$, and an RMSE of $0.69°$, attesting to good kinematic agreement.

Although an integration drift was empirically quantified in both the long-duration extended assays (240 minutes) and the four 80-minute validation collections (with the rate stabilizing at approximately $0.27°/\text{min}$), this effect was successfully neutralized using a software-embedded linear temporal compensation algorithm, reducing residual deviation to negligible levels. The consolidated results demonstrate that the prototype provides reliable and stable lateral postural tracking over time, establishing the technical foundation necessary to test active self-correction mechanisms with vibrotactile biofeedback.

As future perspectives, the next steps prioritize circuit miniaturization onto a Printed Circuit Board (PCB), the implementation of wireless communication, and the conducting clinical trials—preceded by approval from the Research Ethics Committee (REC)—with larger groups of participants. Additionally, the expansion of the kinematic algorithm for simultaneous mapping of all three postural axes (flexion/extension, lateral inclination, and axial rotation) is anticipated. This three-dimensional approach, associated with validation via multi-camera optical systems, will allow for the assessment of complex and combined deviations, essential for managing three-dimensional structural pathologies like scoliosis. Thus, the developed architecture provides a promising and low-cost technological foundation for the future of clinical and ergonomic assistance tools.

Acknowledgments

The authors declare the use of generative artificial intelligence tools during the development of this work. The *Pippit AI* tool (*ByteDance Pte. Ltd.*) was used to generate the base anatomical illustration and coordinate axes diagramming presented in Fig. 1. Additionally, *Claude* (*Anthropic*) was employed to assist in structural optimization and debugging of the microcontroller firmware's buffer management code. *ChatGPT* (*OpenAI*) was utilized to assist in the translation of the manuscript from Brazilian Portuguese to American English, as well as for linguistic refinement to enhance clarity, consistency, and adherence to academic writing standards. All scientific content, interpretations, analyses, and conclusions were independently verified and remain under the sole responsibility of the authors.

References

[1] World Health Organization, "Musculoskeletal health," 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions

[2] European Agency for Safety and Health at Work, *Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU*. Luxembourg: Publications Office of the European Union, 2019.

[3] C. J. L. Murray *et al.*, "Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1223–1249, Oct. 2020, doi: 10.1016/S0140-6736(20)30752-2.

[4] A. D. Woolf, J. Erwin, and L. March, "The need to address the burden of musculoskeletal conditions," *Best Practice & Research Clinical Rheumatology*, vol. 26, no. 2, pp. 183–224, Apr. 2012, doi: 10.1016/j.berh.2012.03.005.

[5] P. Gorce and J. Jacquier-Bret, "Work-Related Musculoskeletal Disorder Prevalence by Body Area Among Nurses in Europe: Systematic Review and Meta-Analysis," *Journal of Functional Morphology and Kinesiology*, vol. 10, no. 1, p. 66, Mar. 2025, doi: 10.3390/jfmk10010066.

[6] N. da S. Pontes *et al.*, "Social determinants and work-related musculoskeletal disorders in Brazil," *PLOS ONE*, vol. 19, no. 7, p. e0306840, Jul. 2024, doi: 10.1371/journal.pone.0306840.

[7] Scoliosis Research Society, "Patient Adolescent Idiopathic Scoliosis Handbook for Patients," 2026. [Online]. Available: https://www.srs.org/Files/Patient-Brochures/Patient.Adolescent_Idiopathic_Scoliosis_Handbook_for_Patients.pdf

[8] S. Negrini *et al.*, "2016 SOSORT guidelines: orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth," *Scoliosis and Spinal Disorders*, vol. 13, p. 3, Jan. 2018, doi: 10.1186/s13013-017-0145-8.

[9] V. Dimitrijević *et al.*, "Treatment of idiopathic scoliosis with conservative methods based on exercises: a systematic review and meta-analysis," *Frontiers in Sports and Active Living*, vol. 6, Dec. 2024, doi: 10.3389/fspor.2024.1492241.

[10] E. Rezaei *et al.*, "Musculoskeletal disorders among taxi drivers: a systematic review and meta-analysis," *BMC Musculoskeletal Disorders*, vol. 25, no. 1, p. 663, Aug. 2024, doi: 10.1186/s12891-024-07771-w.

[11] T. Morino *et al.*, "Trends in the Prevalence of Adolescent Idiopathic Scoliosis in a Japanese Prefecture: A 25-Year Population-Based School Screening Study Using Moiré Topography," *Global Spine Journal*, p. 21925682261417283, Jan. 2026, doi: 10.1177/21925682261417283.

[12] D. Kim, M. Cho, Y. Park, and Y. Yang, "Effect of an exercise program for posture correction on musculoskeletal pain," *Journal of Physical Therapy Science*, vol. 27, no. 6, pp. 1791–1794, Jun. 2015, doi: 10.1589/jpts.27.1791.

[13] D. R. Martins, S. M. Cerqueira, and C. P. Santos, "Combining inertial-based ergonomic assessment with biofeedback for posture correction: A narrative review," *Computers & Industrial Engineering*, vol. 190, p. 110037, Apr. 2024, doi: 10.1016/j.cie.2024.110037.

[14] B. M. Gaffney, K. S. Maluf, and B. S. Davidson, "Evaluation of Novel EMG Biofeedback for Postural Correction During Computer Use," *Applied Psychophysiology and Biofeedback*, vol. 41, no. 2, pp. 181–189, Jun. 2016, doi: 10.1007/s10484-015-9328-3.

[15] S. M. Cerqueira, A. F. D. Silva, and C. P. Santos, "Smart Vest for Real-Time Postural Biofeedback and Ergonomic Risk Assessment," *IEEE Access*, vol. 8, pp. 107583–107592, 2020, doi: 10.1109/ACCESS.2020.3000673.

[16] G. Tsiringakis, Z. Dimitriadis, E. Triantafylloy, and S. McLean, "Motor control training of deep neck flexors with pressure biofeedback improves pain and disability in patients with neck pain: A systematic review and meta-analysis," *Musculoskeletal Science and Practice*, vol. 50, p. 102220, Dec. 2020, doi: 10.1016/j.msksp.2020.102220.

[17] M. A. Hunt *et al.*, "Feasibility of a gait retraining strategy for reducing knee joint loading: Increased trunk lean guided by real-time biofeedback," *Journal of Biomechanics*, vol. 44, no. 5, pp. 943–947, 2011.

[18] J. Anderson *et al.*, "Reduction of frontal plane knee load caused by lateral trunk lean depends on step width," *Gait & Posture*, vol. 61, pp. 483–487, 2018.

[19] J. Rowberg, "I2Cdevlib — MPU6050_6Axis_MotionApps20," *GitHub Repository*. [Online]. Available: https://github.com/jrowberg/i2cdevlib

[20] InvenSense Inc., *MPU-6000 and MPU-6050 Product Specification*, Rev. 3.1, Doc. PS-MPU-6000A-00, 2011.

[21] B.-C. Lee, B. J. Martin, and K. H. Sienko, "The effects of actuator selection on non-volitional postural responses to torso-based vibrotactile stimulation," *Journal of NeuroEngineering and Rehabilitation*, vol. 10, no. 21, 2013.

[22] G. Wu *et al.*, "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion — Part II: shoulder, elbow, wrist and hand," *Journal of Biomechanics*, vol. 38, no. 5, pp. 981–992, 2005.

[23] Kinovea, "Kinovea — sports video analysis software," 2021. [Online]. Available: https://www.kinovea.org

[24] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[25] Bytedance Pte. Ltd. and OpenAI, "Image generated by Pippit AI and edited via DALL-E 3 from the prompt: 'anatomical model of human back with coordinate axes'," Jul. 2026. [Online]. Available: https://www.pippit.ai and https://openai.com.

[26] Anthropic, "Claude 3.5 Sonnet (Assistance with firmware code structure and optimization)," Mar. 2026. [Online]. Available: https://www.anthropic.com.

[27] M. Acar, F. M. Sienko, and K. H. Sienko, "Does vibrotactile biofeedback for postural control interfere with cognitive processes?" *Experimental Brain Research*, vol. 242, pp. 1120–1135, 2024.

[28] K. H. Sienko, S. S. Sienko, and M. D. Balkwill, "The effects of attractive vs. repulsive instructional cuing on balance performance," *Journal of NeuroEngineering and Rehabilitation*, vol. 9, no. 1, p. 28, 2012.

Development of a Programmable Visual Cueing System to Support Gait Rehabilitation in Patients with Parkinson's Disease

Sara Moreira Macedo
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
lIlha Solteira-SP, Brasil
sara.macedo@unesp.br

Arthur Araujo da Silva
Departamento de Sistemas de Informação
Universidade Federal do Mato Grosso do Sul
Três Lagoas-MS, Brasil
arthur.a@ufms.br

Murilo Henrique Lima de Souza
Departamento de Engenharia Civil
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
murilo.l.souza@unesp.br

Leonardo Xavier do Nascimento
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
lIlha Solteira-SP, Brasil
lx.nascimento@unesp.br

Rafael José Ginuino da Silva
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
rafael.ginuino@unesp.br

Pablo Mauricio Portilla Hernandez
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
pablo.portilla@unesp.br

Edinaldo Serra Cardoso Júnior
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
edinaldo.junior@unesp.br

Bodouowei Godswill Zipamone
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
b.zipamone@unesp.br

Fabian Rodrigo Castro Forero
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
fabian.castro@unesp.br

Adalberto Vieira Corazza
Departamento de Medicina
Universidade Federal do Mato Grosso do Sul
Três Lagoas-MS, Brasil
adalberto.corazza@ufms.br

Miguel Fernando Paiva Velhote Correia
Departamento de Engenharia Eletrotécnica e de Computadores
Universidade do Porto
Porto-PT, Portugal
mcorreia@fe.up.pt

Marcelo Augusto Assunção Sanches
Departamento de Engenharia Elétrica
Universidade Estadual Paulista
Ilha Solteira-SP, Brasil
marcelo.sanches@unesp.br

***Abstract*—Abstract—This study presents the development of a programmable walkway for generating visual and auditory cues for gait training. The prototype integrates two independently controlled addressable LED strips, an ESP32 microcontroller, and a mobile application that configures step length, cue size, transition time, movement direction, starting limb, and optional auditory signals. Preliminary laboratory tests verified communication among the components and execution of the programmed operating modes. These tests were qualitative and did not include instrumented measurements of spatial or temporal error, repeatability, electrical safety, or evaluation with people with Parkinson's disease. The results therefore demonstrate preliminary technical feasibility rather than clinical efficacy.**

Keywords— Assistive Technology, Gait Rehabilitation, Visual Cueing, Parkinson's Disease, Automation.

I. INTRODUCTION

Parkinson's disease (PD) affects 1 to 2 in every 1,000 people [1] and is the second most common neurodegenerative disease worldwide. PD is characterized by reduced dopamine availability in several structures of the central nervous system, such as the substantia nigra, putamen, and caudate nucleus, due to the death of neurons responsible for releasing this catecholamine. This physiological condition causes motor disturbances, such as muscle discomfort, postural instability, resting tremor, and bradykinesia [2]. Among these disturbances, gait dysfunction has the greatest impact on the quality of life of people with PD.

Gait alterations in patients with Parkinson's disease include reduced stride length and speed [3], [4], prolonged gait duration [5], [6], decreased dynamic stability [7], and reduced activation of the gastrocnemius and tibialis anterior muscles during the swing phase, compared with neurologically healthy individuals [8], [9]. These gait alterations are associated with a higher frequency of freezing of gait episodes and an increased risk of falls [10], which may even lead to death [11], [12].

Considering that these gait alterations increase the risk of falls, a growing number of studies have focused on identifying and mitigating gait disturbances in patients with Parkinson's disease (PD). Therefore, the development of assistive technologies capable of identifying gait alterations during daily activities and improving gait training/rehabilitation may contribute to enhancing the autonomy and mobility of patients with PD, reducing the risk of falls, and lowering the cost of PD treatment.

The development of assistive technologies to detect gait alterations in patients with Parkinson's disease (PD) during

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

daily activities, such as walking on the street, is an interesting approach to helping them improve their mobility. Currently, most studies aimed at detecting gait alterations in patients with PD are conducted in laboratories. Furthermore, the few studies on gait alteration detection conducted outside the laboratory are unable to provide users (i.e., patients with PD and/or healthcare professionals) with real-time feedback on movement changes.

Another approach to improving mobility in patients with Parkinson's disease is to optimize training/rehabilitation to mitigate the impact of the disease on gait. Studies have shown that patients with Parkinson's disease can improve walking speed and stride length with or without visual or auditory cues [13], [14]. Therefore, helping patients with Parkinson's disease develop this habit of active movement is a challenge for healthcare professionals.

The positive impact of visual cues on gait performance (speed, cadence, stride length, and stride width) is widely recognized and supported by substantial evidence [15], [16], [17], [18]. External visual cues serve to bypass and compensate for abnormal internal cueing mechanisms in the brain that contribute to deficits in motor planning and execution in patients with Parkinson's disease [19]. These strategies, provided during training/rehabilitation, can improve balance and gait speed [20].

Although the ESP32 microcontroller and addressable LED strips are commercially available, the technological contribution of this work lies in their integration into a single overground system that combines real-time mobile configuration, optional auditory cues, and independent bilateral control. The prototype allows step length, cue size, transition time, movement direction, and starting limb to be configured, and the right and left strips can generate different patterns. At this stage, these features define the proposed architecture; their quantitative performance and clinical effects remain to be evaluated.

Thus, this study aims to describe the development and preliminary qualitative functional verification of a programmable walkway for generating visual and auditory cues, intended for future investigation in individualized gait-training protocols for people with Parkinson's disease.

II. RELATED WORK

Visual cueing strategies applied to gait training range from fixed floor markings to projection-based systems, wearable devices, and interactive surfaces. Among the simplest approaches, Sidaway et al. [15] used transverse markings positioned along the walking path to guide step length. Similarly, Das et al. [17] used lines fixed to the floor to provide visual references during changes in direction. Although these solutions are accessible and easy to implement, they depend on the manual positioning of the cues and do not allow them to be dynamically modified during the activity.

Treadmill- and projection-based systems enable greater control of the cues. Schlick et al. [16] used an instrumented treadmill associated with the RehaWalk® system, in which the participants' footprints were projected onto the surface and synchronized with the treadmill speed. The shape of the footprints and the distance between the cues were adjusted according to the individual characteristics of the participants. However, this approach depends on a setup consisting of a treadmill, a pressure-sensitive platform, a projector, and specific software.

Other approaches seek to provide cues during movement through wearable devices. Chan et al. [21] developed footwear equipped with lasers, plantar pressure sensors, and inertial measurement units. The system alternately projects a line for the right foot and the left foot according to the gait phase detected by the sensors. Although it enables use outside a fixed setup, the authors reported limitations related to the projection distance of the cues. Janssen et al. [22], in turn, investigated the use of smart glasses to display virtual bars and staircases in augmented reality. This solution expanded the possibilities for cue presentation but presented usability limitations related to comfort, distraction, and partial obstruction of the visual field.

Interactive LED surfaces have also been developed for broader motor training applications. Van Delden et al. [23] presented an 8-m-long pressure-sensitive LED floor designed for personalized games for gait rehabilitation. Although it enables the generation of dynamic cues and the adaptation of activities, the system requires a dedicated interactive surface and was not specifically developed for people with Parkinson's disease.

Table I presents a comparison between the main characteristics of these approaches and the system proposed in this work.

TABLE I
COMPARISON OF THE TECHNICAL CHARACTERISTICS OF VISUAL CUEING STRATEGIES AND SYSTEMS APPLIED TO GAIT

Reference/ system	Visual cue	Cue adjustment	Bilateral control	Configuration interface	Auditory cue	Physical setup
Sidaway et al. [15]	Transverse blue tape strips on the floor	Spatial adjustment through manual positioning	No	No	No	Marked walkway
Schlick et al. [16] (RehaWalk®)	Footprints projected onto a treadmill	Individualized and synchronized with the treadmill	N.R.	RehaWalk® system	N.R.	Treadmill, projector, and instrumented platform
Das et al. [17]	Star-shaped adhesive tape lines	Fixed manual spatial adjustment to guide turns	No	No	No	Marked floor area
Chan et al. [21]	Lines projected by lasers attached to the footwear	Automatic activation according to gait phase	Automatic alternation between sides	Smartphone for data collection, not configuration	N.R.	Footwear with lasers, pressure sensors, and inertial units

Reference/ system	Visual cue	Cue adjustment	Bilateral control	Configuration interface	Auditory cue	Physical setup
Janssen et al. [22]	Three-dimensional bars and staircases in augmented reality	Position updated in real time and individualized spacing	Not applicable	Software running on a tablet	Metronome evaluated in a separate condition	Smart glasses, tablet, and inertial sensors
Van Delden et al. [23]	Interactive LED floor	Step length and width, speed, and duration via tablet	Partial: allows the affected limb to be specified	Tablet	Yes, through sound effects	Dedicated pressure-sensitive LED floor (8 × 1 m)
Proposed system	Addressable LED strips along the sides of the walkway	Step length, cue size, and transition time configurable in real time	Independent control of the right and left sides	Mobile application	Yes, with optional activation	Rubberized walkway with lateral LED strips

Note: N.R. = characteristic not reported in the reviewed study. The comparison considers only the technical characteristics described in the references and does not establish clinical superiority among the approaches.

Unlike the approaches analyzed, the proposed system combines two independently controlled LED strips, a mobile application for real-time configuration, wireless communication, and optional auditory cues in an overground walkway. The contribution of this work lies in integrating these features into a single configurable platform. The comparison presented in Table I is limited to the technical characteristics reported in the cited studies and does not establish superior performance or clinical effectiveness.

III. MATERIALS AND METHODS

A. Walkway

The walkway was designed as a research prototype for generating configurable visual cues along an overground path. Two lateral addressable LED strips allow programmed changes in spatial and temporal settings, including cue position, step length, and progression timing. These settings are intended to support future investigation of structured gait-training protocols; the present study does not evaluate their clinical adequacy or effects. Figure 1 presents a design drawing of the walkway.

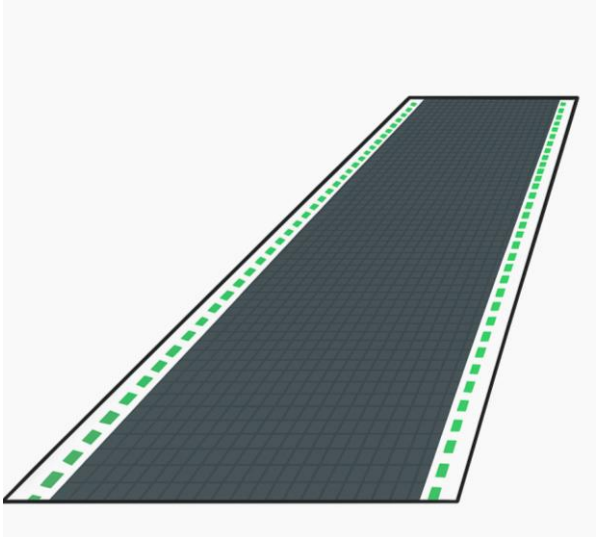


Fig. 1 Walkway Project

The walkway prototype includes two 5-m-long addressable WS2811 LED strips with 300 LEDs, 12 V, IP20, VLX, and RGB lighting, as well as a 5-m-long, 43-cm-wide non-slip rubber runner. The main objective of the device is to provide an

accessible and adjustable platform that can be easily integrated into the clinical context, expanding external rehabilitation strategies for people with gait disorders, particularly those with Parkinson's disease. Figure 2 presents the prototype that has already been constructed.



Fig.2 Completed Walkway Prototype

With the completion of the walkway prototype, as shown in Figure 2, it was possible to begin the coding stage of the parameters previously defined for the research project.

B. Walkway Code

The firmware developed for the walkway was designed to provide configurable control of the visual cues used in gait training. The architecture organizes the functions into independent modules, each responsible for updating the LEDs, managing configurable parameters, communicating with the microcontroller, and synchronizing the auditory signals. This modular structure allows the system to maintain temporal stability even in dynamic operating modes, ensuring the spatial consistency of the light pattern along the walkway.

Operational control is based on a state machine that manages transitions between different operating states, such as configuration, execution, and pause. Each treatment mode—unidirectional movement, forward-and-backward movement, or alternating limb movement—is handled as a specific set of rules that define the progression of the stimulation, movement limits, and application rhythm. Therefore, clinical parameters such as step length, rhythm, and activation time are translated into discrete positions on the LED strip, enabling the faithful reproduction of different gait patterns.

Visual behavior is updated in a continuous cycle. The firmware calculates the position of the activated LEDs, checks the limits of the available channel areas, and applies safety mechanisms to prevent overlap or incorrect activation. The system can also optionally synchronize the auditory signal emitted by a buzzer, a feature that expands the multimodal potential of the intervention by integrating auditory stimulation into the treatment protocol.

The control logic is implemented using procedural programming, with functions organized according to their responsibilities, thereby ensuring clarity, ease of maintenance, and the future scalability of the system. This approach provides the flexibility required for clinical adjustments, such as cue color, progression rate, number of repetitions, and independent selection of each side of the gait path, allowing healthcare professionals to customize training programs according to the patient's movement needs. Therefore, the firmware provides a configurable platform for preliminary experimental investigation, while its stability, reliability, and clinical suitability still require dedicated evaluation.

Figure 3 presents a small diagram illustrating the overall architecture of the project. The mobile application sends clinical parameters to the ESP32 via Wi-Fi, and the ESP32 runs the firmware responsible for generating the light patterns. The state machine translates these parameters into visual cues displayed on the left and right LED strips. A power supply ensures a stable power supply for the entire system, enabling the safe and repeatable execution of the gait training program.

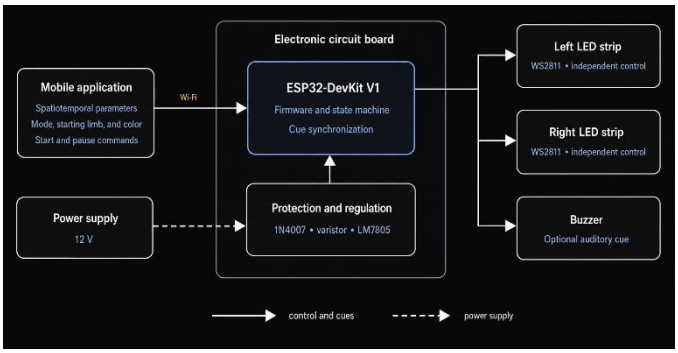


Fig. 3 Walkway Diagram

C. Walkway board

The electronic board developed for the walkway was designed as an integrated and low-cost prototype capable of meeting all the functional requirements of the visual cueing system for gait training. The hardware centralizes control in the ESP32-DevKit V1 module, which is responsible for processing the effect modes, communicating with the mobile application, and managing the signals sent to the LED strips. The board incorporates a protection stage consisting of 1N4007 diodes, a varistor, and an LM7805 voltage regulator for basic circuit protection and voltage regulation. Formal electrical safety and hardware reliability tests have not yet been conducted. The interface with the strips is implemented using BSS138 MOSFETs and BC548 transistors, which isolate the microcontroller and meet the requirements of the data channels. Figure 4 shows the completed board.

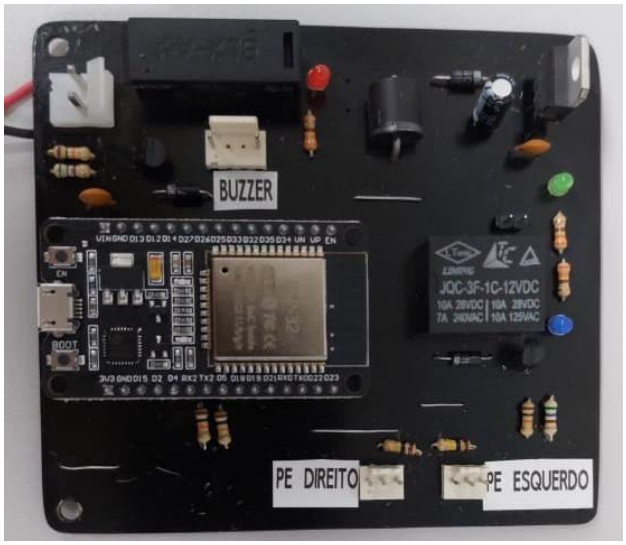


Fig. 4 Completed Walkway Circuit Board.

Figure 5 shows the enclosure housing the board, which was modeled and 3D-printed using PLA (polylactic acid). The enclosure contains the power button, the DB8 connector for the LED strips, the audio output located next to the LED strip connector, the ESP32 reset button, the ESP32 microcontroller connector, and the power connector.



Fig. 5 Walkway Hardware Enclosure

In addition, indicator LEDs, configuration jumpers, and dedicated connectors facilitate system diagnostics and maintenance, while the integrated buzzer circuit enables auditory signaling synchronized with the visual cues. Thus, the platform was designed not only to be functional, but also portable, modular, and scalable, allowing for the future development of the prototype and its adaptation to different clinical gait rehabilitation protocols.

D. Walkway Application

A mobile application was developed to control and configure the luminous walkway used as a gait rehabilitation and training tool. The application was designed to operate in conjunction with an ESP32 microcontroller connected to the local network via Wi-Fi, allowing the visual and auditory cues provided to the patient during the test to be configured in real time. In Figure 6, the main interface presents a set of modules designed for communication with the ESP32, test initialization, and the configuration of general parameters or parameters specific to each LED strip. The application displays the microcontroller's IP address and provides a Connection Test command, which verifies the responsiveness of the embedded system before the experimental protocols are executed.

The Both Strips configuration screen allows the parameters applied to both sides of the walkway to be adjusted simultaneously. The available variables include: LED status (ON/OFF); lighting color (selected via a drop-down menu); step length (cm); foot size (cm); transition time or activation time (s); total walkway length (m); operating mode (One-way, right foot; One-way, left foot; Out-and-back, left foot; Out-and-back, right foot); buzzer activation; and sound type (high-pitched, double, scale, or triple).

These parameters correspond to essential biomechanical variables for gait training, allowing cadence, rhythm, and the spatial positioning of the visual cues to be adjusted according to the user's individual characteristics.

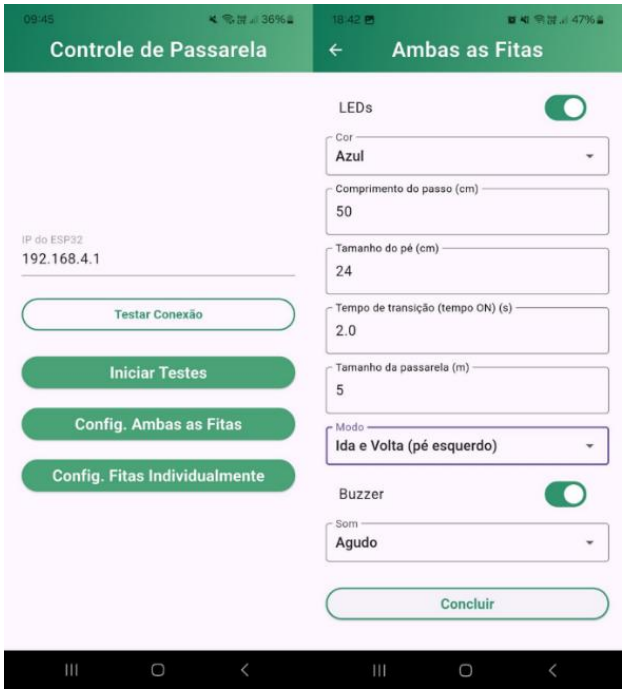


Fig. 6 On the left, the application's home screen; on the right, the configuration screen for both LED strips

Figure 7 illustrates, in addition to the combined configuration mode for both strips, a section dedicated to the independent configuration of each side of the walkway—the left and right strips. Although each strip has the same adjustment settings as the Both Strips configuration mode, it includes one additional option: "Enable Strip," which determines whether the healthcare professional will work with a specific side of the walkway. When a strip is disabled, all associated settings are automatically locked, ensuring configuration integrity and preventing incompatible settings during the therapeutic protocol. This functionality allows the healthcare professional to select only the side required for treatment, providing greater precision in controlling the visual cues according to the clinical objectives of gait training.

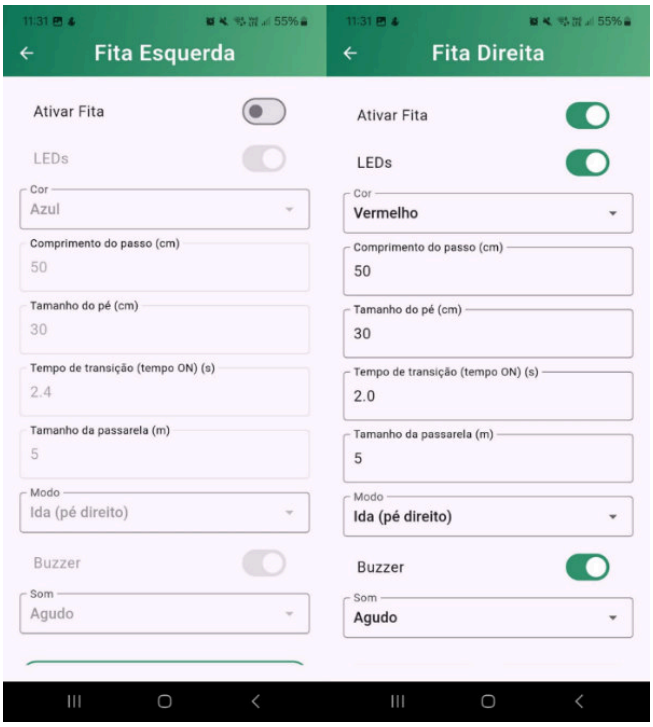


Fig. 7 On the left, the individual configuration screen for the left LED strip; on the right, the configuration screen for the right LED strip

Preliminary functional tests were conducted by selecting different configurations through the mobile application and observing the execution of the programmed commands, including LED-strip activation, movement direction, starting-limb selection, independent bilateral control, and optional buzzer operation. These tests were intended to verify component integration and basic system operation. No predefined quantitative protocol, calibrated instruments, or repeated-measurement analysis was used at this stage. Therefore, the procedure is characterized as a preliminary qualitative functional evaluation rather than a quantitative technical validation.

IV. Results

The preliminary tests showed that the ESP32, power supply circuits, driver transistors, addressable LED strips, firmware, and mobile application operated as an integrated system and executed the selected commands. The intended LED strips, movement direction, starting-limb option, and optional auditory cue were observed during the tests. Because no quantitative protocol or calibrated instruments were used, these observations do not establish spatial or temporal accuracy, repeatability, audiovisual synchronization, long-term stability, or hardware reliability.

A. Bench testing and validation of the hardware and device programming

Exploratory tests with laboratory members were conducted only to demonstrate the operation of the system. No gait outcomes were collected. Therefore, these preliminary observations indicate the execution of the implemented functions but do not demonstrate clinical effectiveness or readiness for use with patients.

Figure 8 shows a test of the walkway code with a laboratory member at the beginning of a session configured to start with the left foot.



Fig. 8 Walkway Code Test Starting with the Left Foot.

Figure 9 shows the beginning of a session configured for the laboratory member to start with the right foot.



Fig. 9 Walkway Code Test Starting with the Right Foot

It should be noted that the preliminary tests of the walkway, conducted with laboratory members, were intended exclusively to evaluate the operation of the device's hardware and software. At this stage, aspects related to component integration, the execution of programmed commands, and compliance with the previously established technical parameters were assessed. Therefore, these procedures did not constitute clinical testing and did not involve people with Parkinson's disease.

The prototype was sent to the Human Movement Research Laboratory (MOVI-LAB), at São Paulo State University (UNESP), Bauru campus, for operational testing and feedback from researchers and rehabilitation professionals. This stage aims to assess the operation of the system, the suitability of the configured gait parameters, and the usability of the interface. Based on this feedback, the necessary technical adjustments will be made before testing with people with Parkinson's disease. Such testing will only begin after completion of this stage and approval by the appropriate institutional ethics committee. Therefore, the present study does not include clinical data or results involving the target population.

It should also be noted that the prototype presented corresponds to the first stage of a broader project that includes the development of additional complementary devices. In the

future, these devices will be integrated into the walkway, enabling a more comprehensive and accurate collection of data related to the gait of participants with Parkinson's disease.

V. CONCLUSION

This work presented the initial development of a programmable walkway for generating visual and auditory cues for gait training. The prototype integrates addressable LED strips, an ESP32 microcontroller, and a mobile application for configuring cue parameters and independently controlling the right and left sides. Preliminary qualitative tests demonstrated the integrated execution of the implemented functions. However, the study did not include quantitative performance measurements or clinical testing involving people with Parkinson's disease. Therefore, the conclusions are limited to the preliminary functional feasibility of the prototype. Future work will include additional technical evaluation, feedback from rehabilitation professionals, the necessary safety assessments, and subsequent studies involving the target population, in compliance with institutional and ethical requirements.

ACKNOWLEDGMENT

This work received financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Finance Code 001.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Tysnes O-B, Storstein A. Epidemiology of Parkinson's disease. J Neural Transm. 2017;124(8):901- 905.

[2] Hughes AJ, Daniel SE, Ben-Shlomo Y, Lees AJ. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain. 2002;125(4):861-870.

[3] Galna B, Murphy AT, Morris ME. Obstacle crossing in people with Parkinson's disease: Foot clearance and spatiotemporal deficits. Hum Mov Sci. 2010;29(5):843-852.

[4] Vitório R, Pieruccini-Faria F, Stella F, Gobbi S, Gobbi LTB. Effects of obstacle height on obstacle crossing in mild Parkinson's disease. Gait Posture. 2010;31(1):143-146.

[5] Barbieri FA, Simieli L, Orcioli-Silva D, et al. Obstacle Avoidance Increases Asymmetry of Crossing Step in Individuals With Parkinson's Disease and Neurologically Healthy Individuals. J Mot Behav. 2018;50(1):17-25.

[6] Rochester L, Baker K, Nieuwboer A, Burn D. Targeting dopa-sensitive and dopa-resistant gait dysfunction in Parkinson's disease: selective responses to internal and external cues. Mov Disord. 2011;26(3):430-435.

[7] Curtze C, Nutt JG, Carlson-Kuhta P, et al. Levodopa Is a Double-Edged Sword for Balance and Gait in People With Parkinson's Disease. Mov Disord. 2015;30(10):1361-1370.

[8] Kleiner A, Galli M, Gaglione M, et al. The Parkinsonian Gait Spatiotemporal Parameters Quantified by a Single Inertial Sensor before and after Automated Mechanical Peripheral Stimulation Treatment. Parkinsons Dis. 2015;2015:1-6.

[9] Mazzetta I, Zampogna A, Suppa A, Gumiero A, Pessione M, Irrera F. Wearable sensors system for animproved analysis of freezing of gait in Parkinson's disease using electromyography and inertial signals. Sensors (Switzerland). 2019;19(4).

[10] Fasano A, Canning CG, Hausdorff JM, Lord S, Rochester L. Falls in Parkinson's disease: A complexand evolving picture. Mov Disord. 2017;32(11):1524-1536.

[11] Deandrea S, Lucenteforte E, Bravi F, Foschi R, La Vecchia C, Negri E. Risk Factors for Falls in Community-dwelling Older People. Epidemiology. 2010;21(5):658-668.

[12] Su Q, Gao Y, Zhang J, et al. Prevalence of Fear of Falling and Its Association With Physical Function and Fall History Among Senior Citizens Living in Rural Areas of China. Front Public Heal. 2021;9:766959.

[13] Pereira VAI, Barbieri FA, Vitório R, et al. The motor deficits caused by Parkinson's disease are notable to block adjustments for a safe strategy during obstacle crossing in individuals with moderate disease. Mot Rev Educ Física. 2015;21(4):436-441.

[14] Cosentino C, Putzolu M, Mezzarobba S, et al. One cue does not fit all: A systematic review with metaanalysis of the effectiveness of cueing on freezing of gait in Parkinson's disease. Neurosci Biobehav Rev. 2023;150:105189.

[15] Sidaway B, Anderson J, Danielson G, Martin L, Smith G. Effects of long-term gait training using visual cues in an individual with Parkinson disease. Phys Ther. 2006;86(2):186-194.

[16] Schlick C, Ernst A, Bötzel K, Plate A, Pelykh O, Ilmberger J. Visual cues combined with treadmill training to improve gait performance in Parkinson's disease: a pilot randomized controlled trial. Clin Rehabil. 2016;30(5):463-471.

[17] Das J, Vitorio R, Butterfield A, et al. Visual Cues for Turning in Parkinson's Disease. Sensors. 2022;22(18):6746.

[18] Russo Y, Stuart S, Silva-Batista C, Brumbach B, Vannozzi G, Mancini M. Does visual cueing improve gait initiation in people with Parkinson's disease? Hum Mov Sci. 2022;84:102970.

[19] Capato TTC, Chen J, Miranda J de A, Chien HF. Assisted technology in Parkinson's disease gait: what's up? Arq Neuropsiquiatr. 2024;82(06):001-010.2016;30(5):463-471.37

[20] Capato TTC, de Vries NM, IntHout J, Barbosa ER, Nonnekes J, Bloem BR. Multimodal Balance Training Supported by Rhythmical Auditory Stimuli in Parkinson's Disease: A Randomized Clinical Trial. J Parkinsons Dis. 2020;10(1):333-346.

[21] H.-L. Chan et al., "Laser-light cueing shoes with integrated foot pressure and inertial sensing for investigating the impact of visual cueing on gait characteristics in Parkinson's disease individuals," Frontiers in Bioengineering and Biotechnology, vol. 12, Art. no. 1334403, 2024, doi: 10.3389/fbioe.2024.1334403.

[22] S. Janssen et al., "Usability of three-dimensional augmented visual cues delivered by smart glasses on (freezing of) gait in Parkinson's disease," Frontiers in Neurology, vol. 8, Art. no. 279, 2017, doi: 10.3389/fneur.2017.00279.

[23] R. van Delden, J. Janssen, S. ter Stal, W. Deenik, W. Meijer, D. Reidsma, and D. K. J. Heylen, "Personalization of gait rehabilitation games on a pressure sensitive interactive LED floor," in Proc. International Workshop on Personalization in Persuasive Technology, vol. 1582, pp. 60–73, 2016.

Development of a Sensorized Cycle Ergometer Integrated with a Serious Game for Post-Stroke Rehabilitation

Silvana Terezinha Faceroli
Department of Education and Technology
IF SUDESTE MG
Juiz de Fora, Brazil
silvana.faceroli@ifsudestemg.edu.br

Ana Carolina Borges Valente
Graduate Program in Development of Industry 4.0-Based Projects
IF SUDESTE MG
Juiz de Fora, Brazil
carolborges391@gmail.com

David Rodrigues Brasil
Bachelor in Information Systems
IF SUDESTE MG
Juiz de Fora, Brazil
davidrodcontato@gmail.com

Miguel Schuchter Ambrosio
Bachelor in Mechatronics Engineering
IF SUDESTE MG
Juiz de Fora, Brazil
miguelschuchter@gmail.com

Sara Del Vecchio
Department of Education and Technology
IF SUDESTE MG
Juiz de Fora, Brazil
sara.vecchio@ifsudestemg.edu.br

Tales Pulinho Ramos
Department of Education and Technology
IF SUDESTE MG
Juiz de Fora, Brazil
tales.ramos@ifsudestemg.edu.br

Thiago Rodrigues Oliveira
Department of Education and Technology, PET Conexões Mecatrônica/BSI
IF SUDESTE MG
Juiz de Fora, Brazil
thiago.oliveira@ifsudestemg.edu.br

***Abstract*—Stroke is one of the leading causes of motor disability and requires rehabilitation strategies capable of promoting neuroplasticity and supporting functional recovery. Among these strategies, physical practice, Action Observation (AO), and Motor Imagery (MI) have been increasingly investigated because of their potential to engage motor-related neural networks involved in motor learning and recovery. This work presents the development of a system composed of a sensorized cycle ergometer integrated with a serious game for post-stroke rehabilitation. The system employs Hall-effect sensors to monitor pedaling activity and an ESP32 microcontroller to acquire and transmit real-time data to a virtual environment developed in the Unity game engine. The serious game was designed to enable both direct interaction through pedaling and the presentation of visual stimuli aimed at promoting motor imagery and action observation. The results of the implementation demonstrated the technical feasibility of integrating hardware and software components and the system's ability to generate quantitative data for clinical monitoring. The proposed solution shows potential as a low-cost platform for motor rehabilitation and future clinical evaluation studies.**

***Index Terms*—stroke, cycle ergometer, serious game, rehabilitation, motor imagery**

I. Introduction

Stroke is one of the leading causes of long-term motor disability worldwide, and persistent motor impairments often require rehabilitation strategies capable of promoting motor relearning and neuroplasticity [1], [2]. Contemporary neurorehabilitation emphasizes combining repetitive, task-oriented practice with interventions that engage motor-related neural networks involved in motor planning, action representation, and sensorimotor integration [3].

Among these interventions, Action Observation (AO) and Motor Imagery (MI) have been extensively investigated as complementary strategies to conventional rehabilitation. AO consists of observing another individual performing a motor task, whereas MI refers to the internal simulation of a movement without overt muscular execution [4], [5].

Neurophysiological studies using transcranial magnetic stimulation, electroencephalography, and functional neuroimaging have demonstrated that both approaches recruit partially overlapping cortical and subcortical motor networks, including premotor, supplementary motor, parietal, and primary motor regions, although through partially distinct neural mechanisms [6], [7]. When combined with repetitive physical practice, AO and MI may enhance motor learning, sensorimotor integration, and functional recovery after stroke [8].

Technological resources such as virtual reality, serious games, and sensor-based rehabilitation systems have emerged as promising tools for implementing intensive, engaging, and objective rehabilitation programs [9]. These technologies enable real-time monitoring of motor performance while providing enriched visual feedback, facilitating the integration of physical exercise with cognitive-motor rehabilitation strategies.

In this context, this work presents the development of a rehabilitation platform integrating a sensorized cycle ergome-

ter with a serious game for post-stroke rehabilitation. The proposed system enables real-time acquisition of pedaling data, motion-based interaction with a virtual environment, and quantitative performance monitoring. Furthermore, the platform was designed to provide complementary active and observation training modes, allowing repetitive lower-limb exercise to be combined with visual feedback, Action Observation, and guided Motor Imagery within a single technological environment. This multimodal approach provides a flexible framework for future investigations of technology-assisted rehabilitation strategies following stroke.

II. Sensorized cycle ergometer

The cycle ergometer is a stationary device widely used in the rehabilitation of individuals affected by stroke, enabling the performance of cyclic and repetitive movements involving both the upper and lower limbs. Depending on the therapeutic goals and the patient's motor abilities, the device can be employed in passive, active, or resistive exercise modalities. Studies have shown that its use may contribute to improvements in range of motion, muscle strengthening, physical conditioning, and the prevention of muscle atrophy, while also supporting functional recovery and increasing patient confidence throughout the rehabilitation process [10].

For the development of the prototype, a portable cycle ergometer equipped with time monitoring and pedal revolution counting was used, allowing subsequent comparison and analysis of the results obtained. The device features an adjustable resistance mechanism that allows variations in intensity of exercise and can be used for both lower-limb and upper-limb training.

The measurement of the angular velocity of the cycle ergometer was made using a pair of Hall effect sensors. These electronic devices are capable of detecting magnetic fields and converting them into electrical signals. Their operating principle is based on the Hall effect, a phenomenon in which an electrical potential difference is generated when an electric current flows through a semiconductor material subjected to a perpendicular magnetic field. This potential difference, known as Hall voltage, is used to detect the passage of magnetic elements and determine rotational events.

To generate the magnetic fields, two permanent magnets were fixed at equally spaced positions on the rotating mechanism of the cycle ergometer. The use of two magnets enables the identification of the pedal rotation direction while also increasing the temporal resolution of the measurements. During the movement of the device, the passage of the magnets in front of the sensors generates electrical pulses, which are used to calculate both the rotational speed and the direction of movement.

Signal processing was performed using an ESP32 microcontroller, selected due to its high computational performance, low cost, integrated wireless connectivity, and ease of programming. Integration between sensors and the microcontroller enables real-time data acquisition and transmission to a computer, supporting interaction with the developed serious game as well as data storage for subsequent analysis.

The data acquisition system was configured with a sampling interval of 200 ms, which, according to bench tests, provided smooth and responsive avatar movement within the virtual environment. However, low pedaling cadences represent a potential challenge, as longer intervals between consecutive pedal events could compromise the responsiveness of the system. To mitigate this issue, two permanent magnets were installed on the rotating mechanism of the cycle ergometer, increasing the number of detectable events per revolution and improving temporal resolution, particularly at low rotational speeds. Data transmission between the ESP32 and the computer was performed at a baud rate of 115,200 bps, ensuring reliable real-time communication while minimizing latency and reducing processor overhead.

After signal acquisition and processing, the ESP32 firmware calculated several kinematic parameters from the detected Hall sensor events, including elapsed time between consecutive events (ms), angular velocity (rpm), linear velocity (m/s), and pedaling direction (clockwise or counterclockwise). These parameters were then transmitted to the computer in real time via USB serial communication, allowing their use both for controlling the serious game and for recording quantitative performance data for subsequent analysis.

To attach the data acquisition system to the cycle ergometer, a dedicated mounting structure was designed and manufactured using 3D-printing technology. Fig.1 illustrates the structure developed to position the sensors and other electronic components.

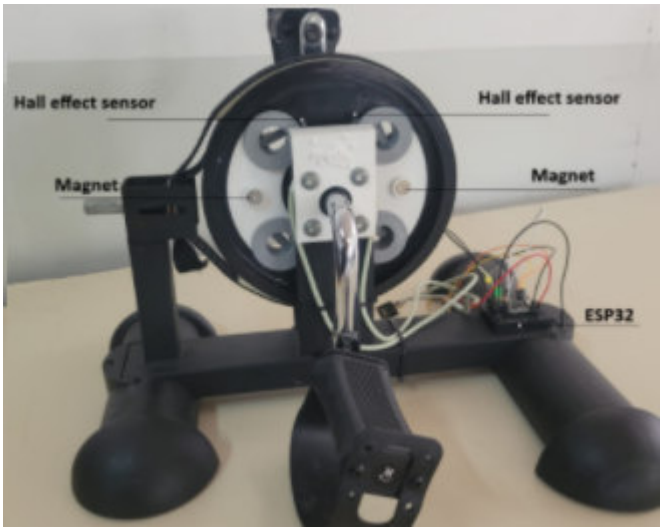


Fig. 1. Prototype developed.

III. Serious Game Developed

The concept of serious games was introduced by Abt in 1970, defining them as games with explicit educational purposes beyond entertainment [11]. [12] expanded this definition by describing serious games as computer-based applications

that combine elements of the game with specific objectives in areas such as education, training, and healthcare. Although entertainment is an important component, these systems are designed for applied and measurable outcomes. In the healthcare domain, serious games have been widely used as therapeutic support tools, contributing to increased patient engagement and enabling more motivating exercise practice compared to conventional approaches [13].

The developed serious game incorporates two distinct interaction modes. In the first mode, users control the game elements through movements performed on the cycle ergometer. The signals captured by the Hall-effect sensors are processed to determine pedal speed in real time, and this information is used as a control parameter within the application. In this way, the user's motor performance directly influences the game dynamics, providing an interactive experience and enhancing engagement during the therapeutic activity.

The second mode consists of the visual presentation of movements performed in different cadences that serve as stimuli for observation and reproduction. This approach aims to stimulate mechanisms related to motor imagery and activation of the mirror neuron system, encouraging users to follow or mentally simulate the observed movement. During activity, performance data are continuously recorded and stored, enabling the generation of reports for clinical monitoring and comparison between user's performance and the prescribed stimuli.

Based on the identified functional and therapeutic requirements, a Game Design Document (GDD) was proposed, with the following structure:

- Objective Description: An endless-running serious game developed to support therapeutic interventions through game-based therapy.
- Game Type: Serious game for motor rehabilitation supported by a sensorized cycle ergometer.
- Target Audience: Individuals undergoing post-stroke rehabilitation.
- Platform: Computer-based application connected to a sensorized cycle ergometer equipped with Hall-effect sensors and an ESP32 microcontroller.
- Game Mechanics: The character continuously moves forward automatically.
- General Characteristics: The game environment must be procedurally generated to provide an infinite path while the player remains active. Additionally, the system must support real-time integration with data acquired from the cycle ergometer, using the user's pedaling activity as the primary control mechanism for gameplay.
- Setting: The game is set in an urban environment. As the player progresses, different areas of the city are explored, providing environmental variety while preserving the overall urban setting.

An endless runner gameplay style was selected due to its simple and intuitive mechanics, allowing users to focus on the pedaling task without the need to learn complex controls. In addition, the continuous progression of the game establishes a direct relationship between pedaling cadence and avatar movement, facilitating real-time interaction with the rehabilitation device. The visual stimuli were designed to provide a continuous representation of cyclic lower-limb movements synchronized with the rehabilitation task, enabling users to observe movement-related cues while interacting with the virtual environment. This design was intended to support action observation and facilitate motor imagery processes, both of which have been associated with motor learning and functional recovery in post-stroke rehabilitation.

The serious game was developed in the Unity game engine using the `C#` programming language. Fig. 2 illustrates the development environment, highlighting the game objects. To create the effect of an "infinite city" during the prototyping phase, a modular city section was developed and used as the basic building block of the game environment (Fig. 3). This section was dynamically instantiated whenever necessary to extend the scenario and automatically destroyed after leaving the player's field of view. This approach enabled the creation of a virtually endless urban environment while maintaining low computational resource consumption. Furthermore, the adopted architecture facilitates future expansion of the game world through the addition of new city sections without requiring modifications to the core generation logic. Fig. 4 presents the avatar pedaling through the urban environment.

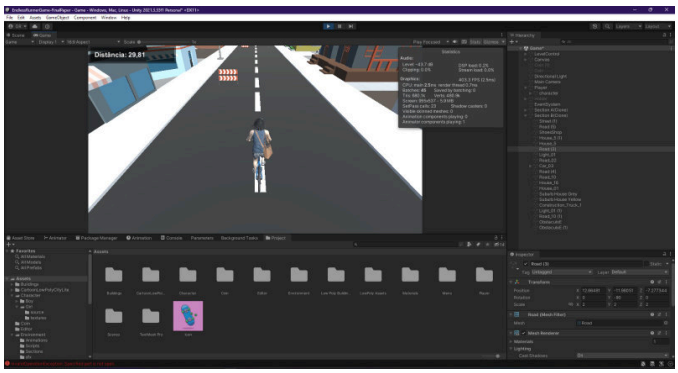


Fig. 2. Development environment.

IV. Proposed System Architecture

Fig. 5 shows the proposed system architecture, which integrates a sensorized cycle ergometer, an ESP32-based acquisition system, and a serious game platform to support post-stroke rehabilitation. The patient's pedaling activity is captured through Hall-effect sensors, processed by the ESP32 microcontroller, and transmitted to a computer in real time via USB serial communication. Communication between the ESP32 and the Unity application was implemented using the .NET SerialPort class available in the System.IO.Ports namespace. The serious game uses these data to provide interactive feedback, while session information is stored in a database for subsequent analysis and the generation of clinical reports. Table I summarizes the main technical specifications



Fig. 3. Urban environment section developed in the Unity game engine.



Fig. 4. Urban environment of the proposed serious game, showing the avatar performing the pedaling task used for rehabilitation training.

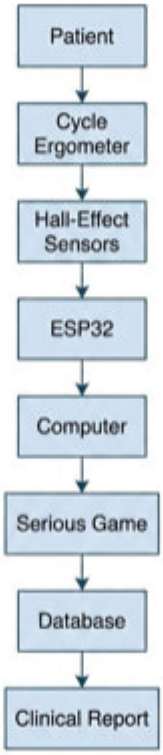


Fig. 5. Proposed system architecture.

TABLE I
TECHNICAL SPECIFICATIONS OF THE PROPOSED SYSTEM.

Component	**Specification**
Sensor	Hall-effect sensor
Microcontroller	ESP32
Sampling interval	200 ms
Communication	USB Serial
Baud rate	115200 bps
Computed parameters	Time (ms), rpm, m/s, direction
Game engine	Unity
Programming language	C#

of the proposed system, including the sensing, processing, communication, and game development components.

V. Discussion and Future Work

Recent studies suggest that post-stroke rehabilitation approaches may benefit from multisensory stimulation [10]. In this context, the proposed system integrates an instrumented cycle ergometer, a real-time data acquisition module, and a serious game into a single rehabilitation platform. This integration combines physical exercise with interactive visual stimuli, creating an environment that supports motor training while incorporating therapeutic approaches based on motor imagery and mechanisms associated with the mirror neuron system.

Bench tests demonstrated the feasibility of integrating the different components of the system, allowing real-time interaction between the user and the virtual environment. These results indicate that the proposed platform has potential for biofeedback applications in motor rehabilitation programs. The adopted configuration, including a 200 ms sampling interval, two magnetic markers, and a communication rate of 115,200 bps, provided stable real-time interaction between the cycle ergometer and the virtual environment during bench tests.

Several studies have highlighted the importance of sensor-based technologies for monitoring kinematic parameters during post-stroke rehabilitation [14], [15], [16]. In this context, the use of Hall-effect sensors combined with an ESP32 microcontroller provided a flexible, low-cost solution to monitor pedaling activity and transmit data in real time. The adopted

architecture also enables the future integration of additional sensors and physiological monitoring devices, which may further expand the applicability of the system in rehabilitation settings. Furthermore, the ability to automatically record quantitative performance data can support clinical follow-up and objective assessment of patient progress over time, representing an advantage over conventional rehabilitation approaches that often rely on subjective evaluations. Action Observation (AO) refers to observing another individual performing a motor action, whereas Motor Imaging (MI) involves the internal simulation of a movement without overt muscular execution [17].

Although distinct, both approaches involve partially overlapping cortical and subcortical motor networks, including premotor, supplementary motor, parietal, and primary motor regions, and have been extensively investigated as complementary strategies to promote motor learning and functional recovery after stroke [18], [19], [20]. Previous studies suggest that visual observation may not directly facilitate cortical activation but may serve as a precursor to motor imagery, which is considered one of the main mechanisms underlying cortical facilitation and motor relearning [21], [22]. In this context, the proposed platform was designed to combine a sensorized cycle ergometer with a serious game that supports both active and observational training modes. In active mode, the user controls the virtual environment by pedaling, promoting repetitive motor practice and real-time visual feedback. In the observation mode, the system presents pedaling movements in predefined cadences, allowing users to observe and mentally simulate the task. By integrating physical exercise, action observation, and motor imagery-related stimuli on a single platform, the proposed system may provide a technological framework for investigating rehabilitation strategies to enhance motor recovery after stroke. Although the system was designed to support motor imagery and action observation strategies, the present study did not include specific neurophysiological or behavioral assessments to verify the induction of these mechanisms.

Despite the promising characteristics of the proposed system, some limitations should be considered. The current study focused exclusively on the development and implementation of the prototype, without performing usability assessments or validation with healthy individuals and post-stroke patients.

Consequently, no conclusions can be drawn regarding the clinical effectiveness of the platform or its impact on functional recovery. Future work will involve experimental validation of the system, including usability studies, user engagement evaluations, and clinical investigations with post-stroke patients. In addition, adaptive game mechanics, personalized rehabilitation protocols, and integration with complementary physiological sensors may be explored to further enhance the therapeutic potential of the platform.

VI. Conclusion

This work presented the development of a system composed of a sensorized cycle ergometer integrated with a serious game for post-stroke rehabilitation. The proposed solution enables the real-time acquisition of pedaling data through Hall-effect sensors and an ESP32 microcontroller, allowing user interaction with a virtual environment while providing resources designed to stimulate motor practice, motor imagery, and mechanisms associated with the mirror neuron system.

The implementation results demonstrated the technical feasibility of integrating hardware and software components, as well as the potential of the platform for rehabilitation applications. In addition, the system demonstrated the capability to generate quantitative data that may support clinical monitoring and patient performance assessment. Future work will include usability testing and validation studies involving both healthy individuals and post-stroke patients in order to evaluate the effectiveness of the proposed tool as a support resource for therapeutic interventions and functional recovery.

Acknowledgment

The authors thank the Federal Institute of Education, Science, and Technology of Southeast Minas Gerais — Juiz de Fora Campus, and the Tutorial Education Program (PET), through the Pro-Rectorate of Teaching and the Ministry of Education (MEC), whose scholarships and resources were essential to support the research and extension activities presented here.

References

[1] E. R. Kandel, J. D. Koester, S. H. Mack, and S. A. Siegelbaum, *Princípios de neurociências*, 6th ed. Porto Alegre, Brazil: AMGH, 2023.

[2] G. Rizzolatti and C. Sinigaglia, *The mirror mechanism: A basic principle of brain function*. Oxford, UK: Oxford University Press, 2016.

[3] A. B. M. Machado and L. M. Haertel, *Neuroanatomia funcional*, 4th ed. Rio de Janeiro/São Paulo, Brazil: Atheneu, 2022.

[4] K. A. Garrison, C. J. Winstein, and L. Aziz-Zadeh, "The mirror neuron system: a neural substrate for methods in stroke rehabilitation," *Neural Plasticity*, 2010, doi: 10.1155/2010/413676.

[5] A. Zimmermann-Schlatter, C. Schuster, M. A. Puhan, E. Siekierka, and J. Steurer, "Efficacy of motor imagery in post-stroke rehabilitation: a systematic review," *Journal of NeuroEngineering and Rehabilitation*, 2008, doi: 10.1186/1743-0003-5-8.

[6] N. L. Díaz, E. Monge-Pereira, E. Jodra, and J. C. Miangolarra, "Motor imagery as a complementary technique for functional recovery after stroke: a systematic review," *NeuroRehabilitation*, 2019, doi: 10.3233/NRE-192838.

[7] K. E. Laver et al., "Virtual reality for stroke rehabilitation," *Cochrane Database of Systematic Reviews*, 2017, doi: 10.1002/14651858.CD008349.pub4.

[8] D. Levac, H. Colquhoun, and K. K. O'Brien, "Clinical application of motor learning principles in stroke rehabilitation," *Physical Therapy*, 2015, doi: 10.2522/ptj.20140014.

[9] D. S. Marín-Medina, P. A. Arenas-Vargas, J. C. Arias-Botero, M. Gómez-Vásquez, M. F. Jaramillo-López, and J. M. Gaspar-Toro, "New approaches to recovery after stroke," *Neurological Sciences*, vol. 45, no. 1, pp. 55–63, 2024, doi: 10.1007/s10072-023-07012-3. PMID: 37697027.

[10] D. Simić-Panić, T. Spasojević, S. Pantelinac, Ž. Živanović, L. Vojnović, and S. Tomašević-Todorović, "The impact of cycling exercise on motor and functional recovery of patients in acute and subacute stroke phase," *Srpski arhiv za celokupno lekarstvo*, vol. 152, no. 5–6, pp. 276–282, 2024.

[11] C. C. Abt, *Serious Games*. New York, USA: Viking Press, 1970.

[12] M. Zyda, "From visual simulation to virtual reality to games," *Computer*, vol. 38, no. 9, pp. 25–32, 2005.

[13] J. E. Fontenele et al., "Rehabilite Game+: Pose estimation-based serious games for forearm and hand movement rehabilitation," *Entertainment Computing*, p. 100992, 2025.

[14] Q. Fang, S. S. Mahmoud, X. Gu, and J. Fu, "A novel multistandard compliant hand function assessment method using an infrared imaging device," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 758–765, Mar. 2019.

[15] T. Hamaguchi, "Support vector machine-based classifier for the assessment of finger movement of stroke patients undergoing rehabilitation," *Journal of Medical and Biological Engineering*, vol. 40, no. 1, pp. 91–100, Feb. 2020.

[16] C. Li, L. Cheng, H. Yang, Y. Zou, and F. Huang, "An automatic rehabilitation assessment system for hand function based on Leap Motion and ensemble learning," *Cybernetics and Systems*, vol. 52, no. 1, pp. 3–25, Oct. 2020.

[17] E. A. Yazgan, B. Kara Kaya, P. Tiryaki, E. Atlı, and U. Cavlak, "Effectiveness of motor imagery and action observation in parameters of sport performance: A systematic review," *Perceptual and Motor Skills*, vol. 133, no. 1, pp. 224–249, 2026.

[18] M. F. Almufareh, S. Kausar, M. Humayun, and S. Tehsin, "Leveraging motor imagery rehabilitation for individuals with disabilities: A review," in *Healthcare*, vol. 11, no. 19, Art. no. 2653, 2023.

[19] Z. Bai, K. N. Fong, J. J. Zhang, J. Chan, and K. H. Ting, "Immediate and long-term effects of BCI-based rehabilitation of the upper extremity after stroke: A systematic review and meta-analysis," *Journal of Neuro-Engineering and Rehabilitation*, vol. 17, no. 1, Art. no. 57, 2020.

[20] N. Brihmat, E. Castel-Lacanal, M. Tarri, B. Lepage, E. Montane, C. Cormier, et al., "Motor imagery and paired associative stimulation in poststroke rehabilitation: Dissociating motor and electrophysiological effects," *Applied Sciences*, vol. 13, no. 10, Art. no. 6063, 2023.

[21] N. Grilc, A. C. Valappil, N. A. Tillin, O. S. Mian, D. J. Wright, P. S. Holmes, et al., "Motor imagery drives the effects of combined action observation and motor imagery on corticospinal excitability for coordinative lower-limb actions," *Scientific Reports*, vol. 14, no. 1, Art. no. 13057, 2024.

[22] R. Meers, H. E. Nuttall, and S. Vogt, "Motor imagery alone drives corticospinal excitability during concurrent action observation and motor imagery," *Cortex*, vol. 126, pp. 322–333, 2020.

Development of a Low-Cost Arduino-Based Wearable Orthosis Prototype for Non-Invasive Temperature Monitoring in Diabetic Foot Prevention: A Simulation Study

1st Thaise Silveira dos Santos Apolinario
dept. of Engineering producttion
UFRN
Natal, Brazil
thaise.apolnario.049@ufrn.edu.br

2nd Hélio Roberto Hékis
dept. of Biomedical Engineering
UFRN
Natal, Brazil
helio.hekis@ufrn.br

3rd Francisca Thaís Queiroz Rego
dept. of Biomedical Engineering
UFRN
Natal, Brazil
thais.rego.092@ufrn.edu.br

4rd Herico Gilberto da Silva Dias
dept. of Biomedical Engineering
UFRN
Natal, Brazil

***Abstract*—This study presents the development of a low-cost wearable orthosis prototype for non-invasive skin temperature monitoring as a potential strategy for diabetic foot prevention. The system was designed using an Arduino Uno microcontroller, LM35 temperature sensors, light-emitting diodes (LEDs), and an elastic support structure. The prototype acquires temperature data from two monitoring points and calculates the temperature difference between them, activating a visual warning when the predefined threshold of 2.2 °C is reached or exceeded. The electronic architecture and embedded algorithm were developed and functionally evaluated through simulations in the Tinkercad platform. The simulation tests verified temperature signal acquisition, data processing, differential temperature calculation, and the activation of the visual alert system according to the programmed decision logic. The results demonstrate the preliminary functional feasibility of the proposed circuit and algorithm in a simulated environment. However, laboratory calibration, experimental testing under real-use conditions, and clinical validation in individuals with diabetes mellitus are still required. The prototype provides a preliminary foundation for future improvements involving sensor distribution, ergonomic optimization, wireless communication, and clinical validation.**

***Index Terms*—Diabetic Foot; Non-Invasive Device; Temperature Monitoring; Wearable Technology; Arduino; Simulation Study.**

I. Introduction

Diabetes mellitus (DM) is one of the fastest-growing chronic diseases worldwide and represents a major public health challenge due to its high prevalence, increasing incidence, and associated healthcare costs. According to the International Diabetes Federation (IDF), approximately 537 million adults were living with diabetes in 2021, and this number is projected to reach 643 million by 2030 and 783 million by 2045, reflecting the growing burden of the disease on healthcare systems worldwide [1].

Among the various chronic complications of diabetes, diabetic peripheral neuropathy (DPN) is one of the most prevalent and debilitating conditions. It affects peripheral nerve function, resulting in reduced pain perception, loss of protective sensation, and impaired thermal sensitivity, particularly in the lower extremities. Consequently, patients become more susceptible to unnoticed trauma, repetitive mechanical stress, ulcer formation, infections, and lower-limb amputations, significantly compromising quality of life while increasing healthcare expenditures [2].

Diabetic foot syndrome is considered one of the most severe complications of diabetes mellitus and remains a leading cause of hospitalization and disability among diabetic patients. The development of plantar ulcers is often preceded by localized inflammatory processes that produce increases in skin temperature before visible tissue damage becomes clinically apparent. Therefore, continuous monitoring of plantar skin temperature has emerged as a promising strategy for the early detection of inflammatory changes and the prevention of ulcer formation [3].

Several clinical studies have demonstrated that a temperature difference greater than 2.2°C between corresponding regions of both feet is strongly associated with an increased risk of ulcer development. Consequently, bilateral skin temperature monitoring has become an important non-invasive biomarker for the early identification of diabetic foot complications, enabling preventive interventions before irreversible tissue damage occurs [4].

Recent advances in wearable biomedical technologies have enabled the development of portable monitoring systems capable of continuously assessing physiological parameters outside clinical environments. These devices have attracted considerable attention due to their potential to support home

healthcare, remote patient monitoring, and telemedicine. In particular, wearable sensors integrated into orthotic devices have shown promising results for continuous plantar temperature monitoring while maintaining user comfort and mobility [5].

Despite these technological advances, many commercially available monitoring systems remain expensive and inaccessible, especially in low- and middle-income countries. Furthermore, several existing solutions rely on proprietary technologies, complex sensor arrays, or sophisticated manufacturing processes, limiting their widespread adoption in public healthcare systems and educational environments. Consequently, there is an increasing demand for affordable, reliable, and reproducible biomedical devices capable of assisting in the early detection of diabetic foot complications.

Open-source electronic platforms, such as Arduino, have become attractive alternatives for rapid biomedical prototyping due to their low cost, flexibility, ease of programming, and extensive development community. Combined with precision temperature sensors, these platforms enable the development of wearable monitoring systems that are reliable, scalable, and suitable for both research and educational applications while maintaining reduced implementation costs [6].

In this context, a low-cost wearable orthosis based on an Arduino Uno microcontroller and LM35 temperature sensors was developed for continuous monitoring of skin temperature in individuals with diabetes mellitus. The system was designed to automatically calculate the temperature difference between two monitored regions and provide an immediate visual warning whenever the temperature difference reached or exceeded the predefined threshold of 2.2°C, indicating the need for medical evaluation. The developed device demonstrates the feasibility of using low-cost electronic components to build assistive technologies for preventive diabetic foot monitoring, contributing to the early detection of thermal alterations associated with inflammatory processes. Furthermore, the prototype highlights the potential of open-source hardware platforms in Biomedical Engineering by providing an affordable, reproducible, and scalable solution that can be further integrated with remote monitoring and telemedicine technologies.

The main contribution of this work is the development and functional validation of a low-cost wearable orthosis for continuous skin temperature monitoring based on clinically established thermal thresholds. The developed system demonstrates the feasibility of integrating accessible hardware and embedded monitoring algorithms into wearable biomedical devices, contributing to the early prevention of diabetic foot complications and providing a foundation for future developments in connected healthcare and telemedicine.

II. Materials and Methods

This study is characterized as applied and exploratory research focused on the development and preliminary functional evaluation, through computer simulation, of a wearable orthosis prototype for non-invasive skin temperature monitoring. The methodology was organized into four main stages: literature review, definition of system requirements and prototype architecture, software and circuit development, and functional testing through simulation in the Tinkercad environment.



Fig. 1. Prototype Development Workflow.Author (2026)

A. Literature Review

Initially, a literature review was conducted to provide a foundation for the prototype's development and to identify technologies used to prevent complications associated with type 2 diabetes mellitus. Bibliographic searches were carried out across recognized databases and scientific portals—including the PubMed, Scopus, and LILACS using combinations of keywords such as diabetes mellitus, diabetic foot, diabetic neuropathy, skin temperature, temperature monitoring, wearable devices, orthosis, and Arduino.

Scientific articles, clinical studies, and technical publications were selected based on their focus on plantar temperature monitoring, the use of sensors for diabetic foot prevention, and the development of wearable devices for monitoring patients with diabetic neuropathy. Analyzing these studies made it possible to identify existing technological solutions, their limitations, and potential improvements for developing a low-cost device aimed at the early monitoring of thermal changes.

Furthermore, the literature review provided the basis for selecting the prototype's electronic components and for adopting a thermal difference threshold of 2.2°C—a value widely cited in the literature as an indicator of inflammatory processes and a risk factor for ulcer development in patients with diabetes.

B. Prototype Development

Following the literature review and the definition of the system requirements, a functional prototype of a wearable orthosis was developed to monitor skin temperature variations associated with diabetic foot complications. The prototype was designed with an emphasis on low cost, ease of assembly, and reproducibility, enabling future improvements and clinical adaptations.

The proposed physical structure consisted of an elastic support concept designed to accommodate the electronic components and enable future contact between the temperature sensors and the skin. At this stage, however, ergonomic

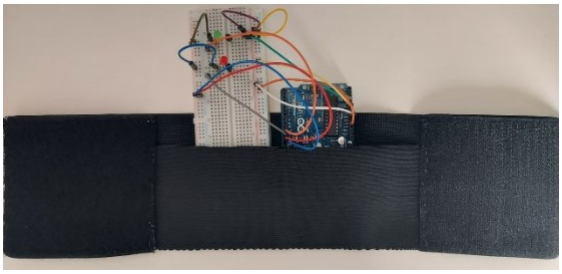


Fig. 2. Temperature sensors.Source: Author (2026)

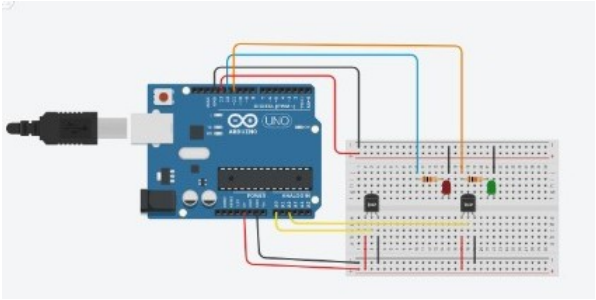


Fig. 3. Arduino Prototype. Source: Author (2026)

comfort, sensor fixation, skin contact stability, and suitability for prolonged use were not experimentally evaluated.

The Arduino Uno was selected due to its open-source architecture, ease of programming, low cost, and widespread use in biomedical engineering educational and research projects.

The sensors were positioned to allow simultaneous acquisition of temperature values from two monitoring points. The analog signals generated by the sensors were transmitted to the Arduino Uno, where they were converted into digital values and processed by the embedded algorithm. Based on these measurements, the system calculated the temperature difference (T) between the monitored regions.

To provide an intuitive user interface, two LEDs were incorporated into the prototype. Under normal operating conditions, when the calculated temperature difference remained below the predefined threshold, the green LED remained activated, indicating the absence of significant thermal alterations. Conversely, when the temperature difference reached or exceeded 2.2 °C, the green LED was turned off and the red LED was activated, providing an immediate visual warning that medical evaluation was recommended. This threshold was established according to findings reported in the literature regarding early detection of inflammatory processes associated with diabetic foot.

During the development phase, the electronic circuit was initially assembled on a breadboard to facilitate component replacement, functional verification, and iterative modifications without permanent soldering. The complete system was subsequently integrated into the orthosis structure.

C. Software Development

The embedded software was developed using the Arduino Integrated Development Environment (Arduino IDE), employing the C++ programming language for implementation of the temperature monitoring algorithm. The software was designed to perform continuous acquisition, processing, and analysis of the temperature signals generated by the LM35 sensors, enabling real-time monitoring of thermal variations associated with diabetic foot complications.

Initially, the algorithm configures the input and output ports of the Arduino Uno microcontroller, defining the analog pins connected to the LM35 sensors and the digital pins responsible for controlling the visual alert system through the LEDs. After the initialization stage, the program continuously acquires the analog voltage values generated by each sensor. The objective of this study was to develop and preliminarily evaluate, through computer simulation, a low-cost Arduino-based wearable orthosis prototype for non-invasive skin temperature monitoring, using bilateral temperature difference as a parameter for the preliminary identification of thermal asymmetries potentially associated with diabetic foot complications.

TABLE I
FUNCTIONAL TESTS OF THE PROPOSED WEARABLE ORTHOSIS

Test	Expected Response	Observed Response
LM35 temperature	Correct acquisition	Correct
Arduino processing	Program execution	Correct
LED indication	Visual activation	Correct
System operation	Continuous execution	Correct

The software was validated through simulations performed in the Tinkercad environment, where different temperature values were applied to the sensors to verify the correctness of signal acquisition, temperature conversion, differential calculation, and activation of the alert system. The simulation stage confirmed the proper execution of the embedded algorithm and demonstrated the feasibility of the proposed monitoring strategy before implementation in the physical prototype.

III. Results and Discussion

The simulation-based evaluation demonstrated the preliminary functional feasibility of the proposed monitoring system. The Tinkercad simulations verified the integration of the Arduino Uno microcontroller, LM35 temperature sensors, and LED-based visual alert system. The functional tests confirmed the expected execution of the programmed logic, including temperature signal acquisition, data processing, calculation of the temperature difference between the simulated monitoring points, and activation of the corresponding visual indicator.

Functional tests confirmed that the embedded software accurately acquired analog signals from the temperature sensors, converted them into temperature values, calculated the temperature difference between the monitored regions, and activated the appropriate visual indicator according to the predefined decision logic. As summarized in Table I, all evaluated functions—including sensor data acquisition, signal processing, LED activation, and continuous system operation—performed as expected throughout the simulations.

The decision algorithm adopted a temperature difference threshold of 2.2 °C, based on the bilateral comparison of corresponding anatomical regions of the feet. Therefore, this threshold should not be interpreted as an isolated absolute temperature value or applied indiscriminately to different anatomical locations. In future experimental and clinical applications, the sensors should be positioned at equivalent anatomical sites on both feet, with a standardized strategy for maintaining consistent skin contact and measurement conditions.

Another important aspect of the proposed device is its affordability and simplicity. The combination of the Arduino Uno platform and LM35 temperature sensors enabled the development of a reproducible and cost-effective wearable system using commercially available electronic components. According to Martín-Vaquero et al., open-source hardware platforms have become an attractive alternative for developing wearable biomedical devices because they reduce implementation costs while maintaining satisfactory performance for research and educational applications [9]. This characteristic is particularly relevant for low- and middle-income countries, where access to commercial plantar temperature monitoring systems may be limited.

The use of the Tinkercad simulation environment also proved valuable during the development process by enabling preliminary validation of the electronic architecture, embedded software, and interaction between sensors and output devices before physical implementation. This approach reduces prototyping costs, facilitates debugging, and improves system reliability during the early stages of biomedical device development.

Despite the preliminary functional results, important limitations should be acknowledged. First, the proposed system was evaluated exclusively through computer simulation and has not yet undergone laboratory calibration or experimental and clinical validation involving individuals with diabetes mellitus. Second, the prototype concept incorporates only two temperature sensors, limiting the number of anatomical regions that can be monitored simultaneously. In future versions, a larger number of sensors should be distributed bilaterally at anatomically corresponding plantar regions, particularly at sites associated with a higher risk of ulceration. In addition, the present study did not experimentally evaluate sensor positioning, skin-contact stability, measurement accuracy, environmental temperature effects, user movement, long-term sensor stability, ergonomic comfort, or dimensional suitability for routine use. These factors may influence temperature measurements and must be investigated before any clinical application of the proposed technology.

Future studies should focus initially on the construction and refinement of the physical prototype, including reduction of its dimensions and optimization of the arrangement of electronic components to improve compactness, comfort, and suitability for routine use. The number of temperature sensors should also be increased, with bilateral distribution at anatomically corresponding plantar regions, particularly at areas associated with a higher risk of ulceration. Sensor positioning should be standardized to ensure equivalence between corresponding regions of both feet and consistent contact with the skin.

The LM35 sensors should subsequently undergo calibration using certified reference instruments, followed by experimental tests under controlled and real-use conditions. These studies should evaluate measurement accuracy, repeatability, environmental temperature effects, sensor stability, movement-related interference, skin-contact conditions, and ergonomic comfort. Finally, experimental and clinical validation involving individuals with diabetes mellitus will be necessary to investigate the clinical applicability of the proposed monitoring strategy. Wireless communication technologies, such as Bluetooth or Wi-Fi, may also be incorporated in subsequent versions to enable remote monitoring and integration with mobile health and telemedicine applications.

Overall, the obtained results demonstrate the functional feasibility of the proposed wearable orthosis and indicate that low-cost open-source technologies can provide an effective foundation for developing assistive biomedical devices aimed at the early prevention of diabetic foot complications. Although further experimental and clinical validation is required, the proposed system represents a promising approach for future wearable health monitoring applications.

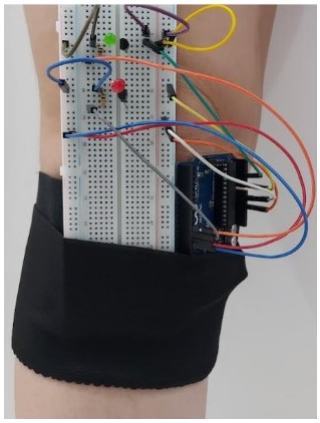


Fig. 4. Temperature Monitoring in Diabetic Foot Prevention.Source: Autor (2026)

IV. FINAL REMARKS

This study presented the development of a low-cost Arduino-based wearable orthosis prototype for non-invasive skin temperature monitoring, aimed at supporting future strategies for diabetic foot prevention. The proposed system integrates an Arduino Uno microcontroller, LM35 temperature sensors, and a visual alert mechanism based on the temperature difference between two monitored points.

The simulation results obtained in the Tinkercad environment demonstrated the preliminary functional feasibility of the electronic architecture and embedded decision algorithm, including temperature signal acquisition, data processing, differential temperature calculation, and activation of the visual alert according to the predefined 2.2 °C threshold.

However, the findings should be interpreted within the scope of a simulation-based preliminary study. No laboratory calibration, experimental testing under real-use conditions, or clinical validation involving individuals with diabetes mellitus was performed. Therefore, the present results do not establish the diagnostic or clinical effectiveness of the proposed device.

Future developments should include the construction and miniaturization of the physical prototype, calibration of the temperature sensors, expansion and standardized bilateral distribution of sensing points, optimization of skin contact and ergonomic characteristics, and evaluation under controlled and real-use conditions. Clinical studies involving individuals with diabetes mellitus will subsequently be necessary to determine the applicability and clinical relevance of the proposed monitoring strategy.

Therefore, the main contribution of this study is the preliminary demonstration of the feasibility of integrating low-cost electronic components and an embedded decision algorithm for bilateral temperature-difference monitoring in a wearable orthosis concept. The proposed architecture provides a basis for subsequent technological development and validation of a more compact, comfortable, and clinically applicable system.

REFERENCES

[1] D. L. Bahia and C.-R. 52-52723, "O alto custo do pé diabético no Brasil," [Online]. Available: https://diabetes.org.br/o-alto-custo-do-pe-diabetico-no-brasil/. Accessed: Jun. 3, 2023.

[2] U. Chatchawan et al., "An exploration of the relationship between foot skin temperature and blood flow in type 2 diabetes mellitus patients: A cross-sectional study," Journal of Physical Therapy Science, vol. 30, no. 11, pp. 1359–1363, Nov. 2018.

[3] Ministério da Saúde, "Diabetes (diabetes mellitus)," [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/diabetes. Accessed: Jun. 3, 2023.

[4] H. B. Menz and D. R. Bonanno, "Objective measurement of adherence to wearing foot orthoses using an embedded temperature sensor," Medical Engineering Physics, vol. 88, pp. 19–24, Feb. 2021.

[5] A. Mao et al., "Detection of Pressure and Heat in a Compressive Orthotic for Diabetes Prevention Using Nanotechnology," in IEEE, [Online]. Available: https://ieeexplore.ieee.org/document/8846491. Accessed: May 24, 2023.

[6] R. Keays, "Diabetes," Current Anaesthesia and Critical Care, vol. 18, no. 2, pp. 69–75, 2007.

[7] Sociedade Brasileira de Diabetes, "Home," [Online]. Available: https://diabetes.org.br/. Accessed: Jun. 3, 2023.

[8] "Qual a função do sensor LM35?," [Online]. Available: https://vocepergunta.com/library/artigo/read/429576-qual-a-funcao-do-sensor-lm35. Accessed: Jun. 3, 2023.

[9] S. Telfer et al., "Personalized foot orthoses with embedded temperature sensing: Proof of concept and relationship with activity," Medical Engineering Physics, vol. 36, no. 1, pp. 9–15, 2014.

[10] Analog Devices, "TMP35 Datasheet," [Online]. Available: https://pdf1.alldatasheet.com/datasheet-pdf/view/49107/AD/TMP35.html. Accessed: Jun. 6, 2023.

Development of an Ankle Exoskeleton to Improve Gait in Foot Drop Patients

Ana Silva
Graduate Program in Neuroengineering
Edmond and Lily Safra International Institute of Neuroscience, Santos Dumont Institute (ISD)
Macaíba, Brazil
luiza.matos@edu.isd.org.br

Jorge Freitas
Graduate Program in Neuroengineering
Edmond and Lily Safra International Institute of Neuroscience, Santos Dumont Institute (ISD)
Macaíba, Natal, Brazil
jorge.eduardo.111@ufrn.edu.br

Renan Matsuda
Graduate Program in Neuroengineering
Edmond and Lily Safra International Institute of Neuroscience, Santos Dumont Institute (ISD)
Macaíba, Brazil
renan.matsuda@isd.org.br

Fabrício Brasil
Graduate Program in Neuroengineering
Edmond and Lily Safra International Institute of Neuroscience, Santos Dumont Institute (ISD)
Macaíba, Brazil
0000-0002-9984-1007

Abstract*—Introduction: Foot drop is a clinical condition derived from neurological injuries characterised by weakness of the muscle tibialis anterior, responsible for ankle dorsiflexion, having as consequence a reduced walking speed, asymmetric gait and increased energy expenditure. Lower limb exoskeletons have been developed to support functional rehabilitation, using sensors to identify gait phases and motors to provide movement assistance, mostly using servomotors, force-sensitive resistors (FSRs) and inertial measurement units (IMUs).The objective of this work is to develop an ankle exoskeleton combining FSR and IMU information to assist ankle dorsiflexion movement while in the swing phase of gait. Method: The device was developed with wearability in mind. The sensors were positioned on the foot: the FSR was secured to the heel by a 3D-printed structure made from flexible filament, which is placed inside the user's shoe; and the IMU sensor was positioned, using a fabric structure, on the foot tip. For the motor, a shin guard was modelled and 3D-printed, also using flexible filament, to ensure greater adaptability to different body types and to keep the motor in place whilst the cable moves. An Arduino Uno and a 7.4V 2,7A battery are used for control and power supply; during the user's walk, these are housed in a knee brace made from neoprene fabric. Results: The wearable exoskeleton developed is capable of recognising the swing phase of the gait, activating the motor and pulling the cable connected to the foot, which in turn assists with dorsiflexion and improves walking. Discussion: Exoskeleton design varies in the literature, particularly with regard to the number and positioning of sensors and the method used to provide assisted dorsiflexion. In this study, both sensors were placed on the foot to improve device wearability and data accuracy. Pressure sensing can be carried out on the calcaneus or the forefoot, depending on the method of gait analysis, while gyroscopes can be positioned on the anterior and posterior aspects of the calf or on the ankle's axis of rotation. Ankle movement can be provided by rigid insoles, which may limit foot propulsion, or cable-based systems, both driven by a rotary motor or servomotor. Conclusions: The sensors, in the chosen quantities and positions, are successfully capable of recognising the swing phase of the gait and activating the servomotor, which, in turn, is able to lift the foot and follow the gait. Further studies are still needed and are ongoing to demonstrate the effectiveness of this exoskeleton in rehabilitation.

***Index Terms*—Exoskeleton, foot drop, gait, force sensing resistor (FSR), inertial measurement unit (IMU).**

This study was supported by institutional funds provided by the Graduate Program in Neuroengineering of the Edmond and Lily Safra International Institute of Neuroscience, Santos Dumont Institute.

I. Introduction

Neurological conditions that can develop to physical impairment are increasingly common today. Stroke and spinal cord injury are among the main causes of motor impairment

acquired in adulthood [1]. In these two cases of neurological injury, a proportion of survivors experience some impairment of lower limb motor function, leading to complications, particularly with regard to gait and mobility. Between 20% and 30% of stroke survivors develop motor impairments such as foot inversion and a lack of dorsiflexion [3]. The latter leads to a gait disorder known as foot drop, in which the individual has insufficient strength to lift their foot during the swing phase of the gait, leading to complications [2].

Foot drop conditions are becoming increasingly common amongst the general population, highlighting the growing number of patients with these conditions who require assistance. An increased risk of falls, contusions and fractures are the main concern for this population due to impaired mobility. It is therefore essential to develop accessible and innovative technologies to support conventional rehabilitation therapies.

Walking is a fundamental activity that brings benefits to an individual's health and well-being [4]. Impairments affecting the lower limbs, usually resulting from either mechanical or neurological injuries, can lead to gait complications, such as a reduced step length and chronic pain, which also have consequences for an individual's mental health and quality of life [5]. Among the motor impairments that develop following a stroke or spinal cord injury, one of the most common is foot drop, which involves an inability to dorsiflex the ankle during the swing phase of the gait cycle – the phase in which the foot is not in contact with the ground. This condition can lead to an asymmetrical gait, reduced walking speed and increased energy waste during walking [6].

As a result of impaired gait, negative effects are observed on an individual's self-esteem, social participation and psychological well-being, mainly due to reduced confidence when walking as a result of an increased risk of falls. This can lead to even more complex problems such as social isolation, anxiety, depression and even mortality [7].

In the literature, wearable assistive devices, such as active orthoses and exoskeletons, emerge as a promising technological tool in the field of neurorehabilitation, contributing to improvements in gait parameters, functional mobility and motor relearning [8], [9]. The use of these technologies, including in home environments, allows for high-intensity, task-specific and repetitive training, which leads to a greater chance of motor recovery [10]. The operating principle of these devices is based on two main points: firstly, the recognition of the gait phase via sensors—in most cases, Force Sensitive Resistors (FSRs) and Inertial Measurement Units (IMUs)—and secondly, the activation of the motor to perform the movement [11], [12]. However, there is no consensus on the optimal number and positioning of sensors for gait recognition, nor on the best method for activating the system.

The aim of this study is to develop an ankle exoskeleton to assist in dorsiflexion, using FSR and IMU sensors in association in order to recognise the gait phase, and a servomotor to assist movement during the swing phase.

II. Methods

A. Sensors

We used one FSR 402 (Interlink Eletronics, California, USA) and one IMU accelerometer and gyroscope, MPU6050 (Invensense Inc, California, USA) to detect the gait phases, more specifically the swing phase.

The FSR is located on the heel of the user and was fixated with a 3D printing structure (Fig. 1). To design the heel structure, a scan was taken of one of the researchers' feet, which was then resized and 3D-printed using flexible filament TPU 95 to improve comfort and adaptability to different shoes and sizes. This structure prevents the sensor from moving and ensures it remains fixed in the desired position for collecting data from the heel. Without it, the sensor moves around inside the shoe and the foot pressure data is no longer accurate.



Fig. 1. Heel Structure

The chosen position for the IMU sensor was the front of the foot. This allows it to detect the foot's inclination, enabling accurate detection of plantar flexion and dorsiflexion. The sensor was secured in the desired position using fabric. The researchers adjusted a fabric and Velcro structure that wraps around the foot and holds both the sensor and the motor-associated rod, which is responsible for holding and pulling the cable on the feet.

The data from the two sensors is used to generate the drive commands for the motor.

B. Motor

We used a DSServo DS3230 Pro 30kg Waterproof Servo (MK Racing, Salzburg, Austria), that works with 4.8 ˜ 6.8V. The battery used, 7.2V and 2,7A, can provide up to 7 hours of autonomy in normal experimental use conditions of the servomotor.

The motor operates based on information from the sensors. Two conditions must be satisfied for the motor to activate, as shown in (Fig. 2). These are:

- The foot must be off the ground. It indicates that the user is on the swing phase of the gait. This information is provided by the FSR sensor located in the heel.
- The foot must be in plantar flexion. This information is provided by the IMU sensor, located in the forefoot.

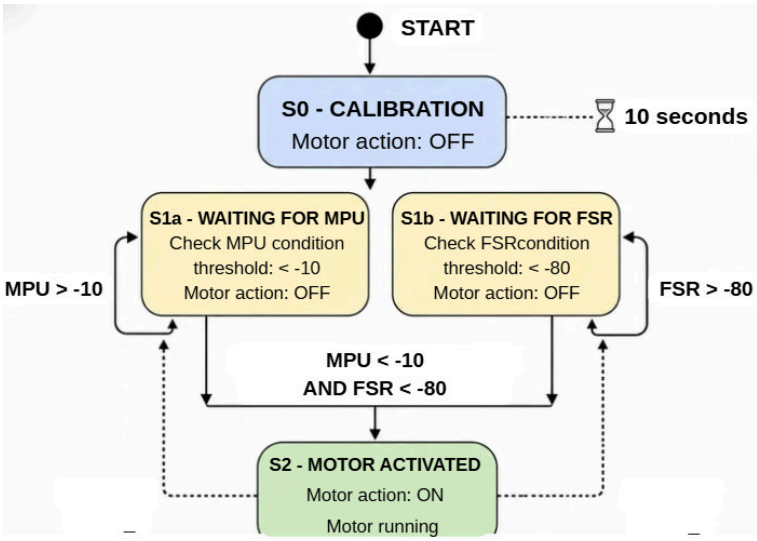


Fig. 2. Finite State Machine

An Arduino Uno microcontroller was used, connected to a printed circuit board (PCB), to establish the link between the data received by the sensors and the command sent to operate the motor.

The code was programmed in C++ to read the sensors, with the necessary calibration performed each time the user started the system. Calibration must be carried out with the user's foot on the floor so that both the pressure between the foot and the volunteer's footwear and the angle between the toe of the footwear and the floor can be measured. A parameter of 10° difference from the calibrated value was used to identify dorsiflexion via the IMU, and a pressure variation of 80 or greater between the current and calibrated values was used to recognise when the foot was off the ground by the FSR.

The motor is positioned on the user's anterior tibia. To this end, a shin guard was designed and 3D-printed both to protect the volunteer's skin and to prevent the motor from falling off or slipping down when activated.

Given the fact that the servomotor we purchased has a rotation of 180°, it was necessary to create a structure that would increase the rotation diameter so that more cable could be pulled at the same angle. We created a pulley that serves both to achieve the required diameter of 2.5 cm and to hold the Bowden cable in place so that it does not slip during operation (Fig. 3). To make it easier to fit the pulley to the motor, we also created a support structure that connects the two parts using screws (Fig. 4). Both structures were 3D-printed using PLA filament, with a gyroid shape and 85% infill to ensure resistance against forces of compression, contraction and shear.

C. Wearable structure

1) Foot/ankle: The design of the footwear component arose from the need to position the accelerometer sensor in the desired location and to attach the cable that would pull the foot to perform the assisted dorsiflexion movement.

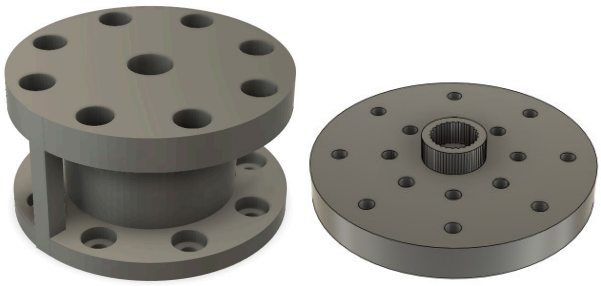


Fig. 3. Pulley and support structure



Fig. 4. Pulley and support structure on servomotor

In this way, two horizontal strips were sewn onto a vertical strip on the foot (Fig. 5). The sensor is located on the top strip. To facilitate the fitting of the device between volunteers, two rods were modelled and 3D-printed in order to adapt the length of the cable to each volunteer's leg.



Fig. 5. Foot structure

The cable adjustment mechanism was designed with two rods, one with a positive fit and the other with a negative fit, which is 3 mm shorter on each side to ensure a secure fit, thereby allowing up to 7 cm of variation in cable length between volunteers (Fig. 6). This structure was printed using PLA filament with a high infill rate (85%) to ensure its integrity.

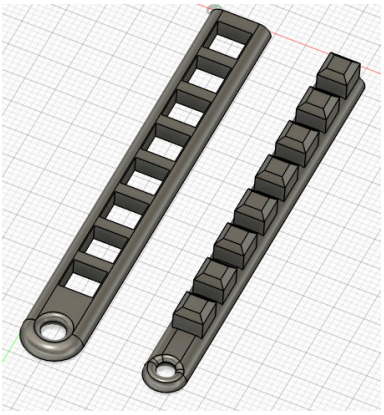


Fig. 6. Rods for cable adjustment

2) Knee: The system was planned and designed for both the battery and the Arduino + PCB assembly to be positioned on the side of the volunteer's affected thigh. To this end, a structure similar to a knee brace was developed, with Velcro strips on the outer sides of the thigh to attach the components to be secured in place (Fig. 7). The knee brace was made from neoprene fabric and also serves to provide support and stability for the shin guard.



Fig. 7. Knee structure

3) Schin: The main challenge in designing the exoskeleton was securing the motor in a position close to the foot in such a way that it did not compromise the user's comfort or the efficiency of the cable-pulling mechanism. With a view to achieving this goal of stabilization and protection, a total of six different versions of the shin guard were developed. The main changes were related to the height required to reduce the impact of the pull on the user's leg, sufficient width to prevent lateral displacement of the sensor while still accommodating users' different body types, and the motor's fit within the shin guard. The last version of the shin guard was 3D printed in flexible filament to allow better comfort for the users and shown to attend the project necessities (Fig.8).

Through sustainability tests, it was observed that the motor operates most efficiently when the pulley is positioned on the outer side and pointing downwards. In this way, the pulling action prevents the foot's inversion and reduces the impact of the reaction torque on the user's leg. For this reason, once the final and desired specifications for the shin guard had been established, an identical but mirror-image model was created for the other leg, so that the motor is correctly positioned for both affected sides.

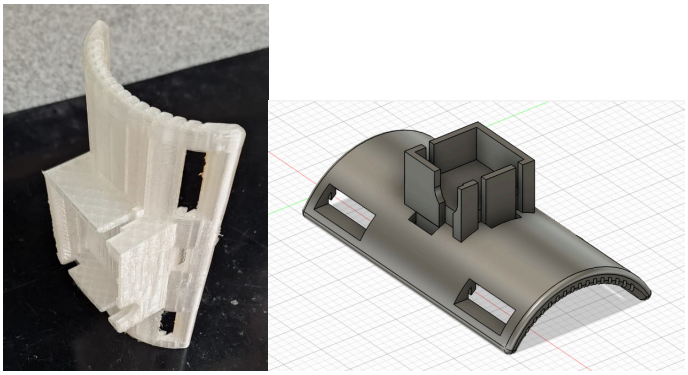


Fig. 8. Schin Guard

4) Safety Button : To ensure user protection, an ergonomic safety button has been added to the device; it is easy to grip and operate. Pressing this button automatically stops the system, so that the motor ceases to apply force to pull the foot and can release more cable to prevent the volunteer's foot from being pulled.

D. Clinical trials

The exoskeleton developed was used on volunteers with foot drop, in accordance with a protocol approved by the Ethics Committee of the Santos Dumont Institute (CAAE 89045425.3.0000.0129). The intervention consisted of 30 minutes of exoskeleton-assisted walking for 4 days over 4 consecutive weeks, amounting to a total of 16 walking sessions.

For metric purposes, gait assessments were carried out on the Zeno Walkway and Qualisys Motion Capture platforms before and after the intervention, and under conditions with and without the exoskeleton. The volunteers were instructed to walk on the platform at a self-selected speed.

Based on this assessment, data relating to the volunteer's gait cycle were obtained, specifying the average total duration of the cycles, as well as the total percentage of the swing and stance phases.

III. Results

As an outcome, we developed a wearable lower-limb exoskeleton to assist the dorsiflexion of the foot in patients with foot drop (Fig. 9). The exoskeleton developed is capable of recognizing the swing phase of the gait of the affected limb using information obtained from the sensors. Once the gait phase in which the foot is dropped is recognized, the device provides dorsiflexion assistance by integrating sensor data with the motor, cable, and rod system (Fig 10).

The device was designed to be adaptable to different users, so that all wearable components — the foot, knee brace, and shin brace — can be used by individuals of different body types without compromising the exoskeleton's function.



Fig. 9. Exoskeleton

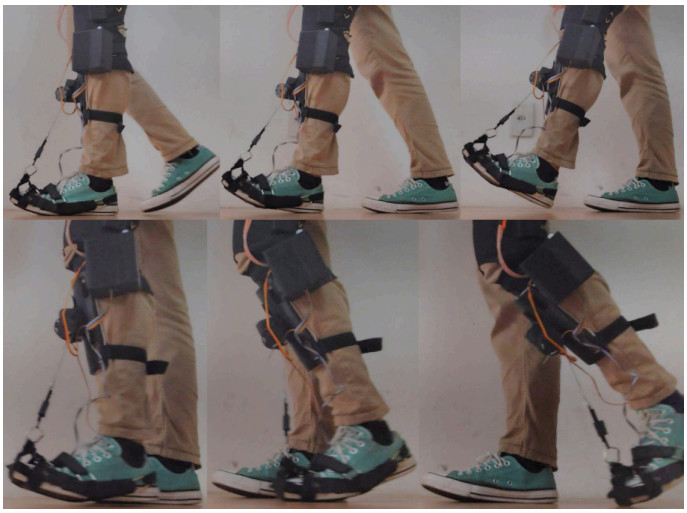


Fig. 10. Gait phases assisted by exoskeleton

The motor has been positioned on the shin in such a way that, even after 30 minutes of activity, it remains securely in place – neither rotating nor sliding down – and continues to lift the foot effectively. This positioning has allowed for the use of shorter cables, which do not become loose with each step. The foot structure housing the IMU sensor and the rod, although being slightly displaced after a long walk, continues to accurately record inclination information and does not compromise the dorsiflexion assistance. The heel structure, which houses the FSR sensor, is comfortable, adaptable to different shoe sizes and ensures that the sensor does not move while walking. Finally, the Arduino and battery are secured to the external side of the thigh via the knee brace structure; their placement does not interfere with walking nor does it compromise the system's operation. When asked, the volunteers reported feeling no discomfort with the noise generated by the motor. Some volunteers reported that the motor noise helped synchronize the execution of the step.

Thus, the developed exoskeleton is capable of assisting with gait in patients with foot drop.

A. Clinical trials

The developed device was tested on volunteers with foot drop and has demonstrated accuracy in detecting the swing phase of the gait in steps. Using the code developed, we can also adjust the walking speed in order to better suit each volunteer's gait, with the aim of promoting a progressively more symmetrical gait.

Using the Zeno Walkway platform, observations were made regarding gait cycle time, both overall and separately for the paretic and unaffected legs. Based on this, reductions in time were observed when comparing walking after the intervention with walking before the intervention, for both conditions—with and without the exoskeleton. The data are set out in the table below.

TABLE I
GAIT CYCLE AND SWING TIME ACCORDING TO VOLUNTEER, TIME, AND CONDITION.

Volunteer	Time	Condition	Gait Cycle Time (sec.)_ Geral	Swing Time (sec.)_ Geral	Gait Cycle Time (sec.)_ Left	Swing Time (sec.)_ Left	Swing Time (sec.)_ Right
1	pre	with	1.520	367	1.515	208	519
1	post	with	1.407	334	1.406	177	491
1	pre	without	1.541	400	1.547	260	540
1	post	without	1.438	369	1.438	213	540
2	pre	with	1.371	474	1.378	475	473
2	post	with	1.252	424	1.255	418	431
2	pre	without	1.462	533	1.458	501	570
2	post	without	1.352	475	1.357	497	453

Gray cells: difference greater than 10% between pre- and post-intervention measurements.

For volunteers 1 and 2, the gait cycle time and swing phase time were reduced during walking after the intervention compared with walking before the intervention. Volunteer 1 (impairment in the right leg), as highlighted in the first four lines, showed a 10 per cent reduction in the overall swing phase and a 17 per cent reduction when considering only the unaffected leg, both when using the exoskeleton. During unassisted walking, Volunteer 1 showed a 22 per cent reduction in the swing phase of the unaffected leg.

Volunteer 2 (affected left leg) had their results shown in the last four lines, with a 10 per cent reduction, when using the exoskeleton, in the overall gait cycle time, a 12 per cent reduction in the overall swing phase, a 10 per cent reduction in the gait cycle time of the affected leg, 13 per cent in the swing phase of the affected leg and 10 per cent in the swing phase of the unaffected leg. As for unassisted walking, the volunteer showed a 12 per cent reduction in the overall swing phase time and a 25 per cent reduction in the swing phase time of the unaffected leg.

In the same optical, the researchers observed, during the interventions with the volunteers, a reduction in compensatory

movements, such as hip and knee elevation, longer strides and greater symmetry between the paretic and unaffected sides.

IV. Discussion

We have developed an ankle exoskeleton to assist with dorsiflexion, which demonstrated satisfactory accuracy in recognising the swing phase of gait in patients with foot drop. Once the swing phase of the gait is recognised, the system sends a command to activate the motor, which is capable of assisting with the dorsiflexion movement, a function that is impaired in this group of individuals. The device was tested during 30-minute walks conducted in volunteers with drop foot and maintained both step recognition and dorsiflexion assistance throughout the entire test period.

The production of the exoskeleton required low costs for the researchers and the institution; a FSR, an IMU, a 2.7A battery, an Arduino Uno, a 30kgF motor and neoprene fabric were purchased, in addition to the parts produced using 3D printing.

The use of exoskeletons in gait rehabilitation has already been widely discussed in the literature. However, the methods for performing assisted dorsiflexion have not yet been standardized. There are, essentially, two factors that differ in relation to exoskeletons developed to assist dorsiflexion. These are: the number and location of the sensors used; and the method of promoting ankle rotation.

Reference [13] used load cells on the anterior and posterior calf associated with gyroscopes on the ankle joint to built soft exoskeletons and detect gait events. Reference [14] used FSRs placed under the heel and in the front part of the foot and an IMU, located on the ankle axis, to compose an exoskeleton ankle robot. Reference [17] used an IMU and an insole with 4 FSRs to activate dorsiflexion and plantarflexion in a pneumatic exoskeleton. Even using different sensors, locations and methods to promote joint movement, both devices were able to promote rehabilitation in patients.

In our project we used a lower number of sensors, one IMU located on the toe and one FSR fixed on the heel with a 3D printed flexible structure, and the authors observed that those were enough for recognizing the gait phases and providing the motors activation

In the same vein, the choice of actuators varies according to their location and the intended movement assistance. Reference [6] developed a system in which the motor is positioned at the back of the hip and foot movement is achieved via Bowden cables, attached to a knee pad and secured to a spiral end connection. As a result, the device created by the authors was able to facilitate dorsiflexion in patients with foot drop. Reference [17] developed artificial pneumatic muscles to assist dorsiflexion and plantaflexion movements. Reference [16] developed the ExoRoboWalker with servomotors associated with a rigid insole to promote dorsiflexion and plantarflexion.

In our device, we positioned the motor on the shin. This positioning was essential for providing assisted dorsiflexion during long walks without the device shifting out of place, thus ensuring the volunteer's comfort. The positioning of the pulley on the lower outer side of the shin guard also served to prevent foot inversion, a common condition in people with foot drop. The placement of the motor on the shin aids rehabilitation due to the improved fit of the device between volunteers, and facilitates the use of the device in activities of daily living since the cable is shorter during the gait phases.

Beyond the advantages related to design and positioning, the accuracy demonstrated by the device during testing, together with the low cost of the instrumentation used, reinforces its potential as an accessible rehabilitation tool. Housing all sensors and the motor on the shin and foot also resulted in a compact, easy-to-don design, which facilitates its use both in supervised rehabilitation sessions and in unsupervised activities of daily living, favouring adherence to treatment.

Reference [15], with the aim of assisting dorsiflexion and eversion of the foot, uses Bowden cables to perform the movement and load cell sensors to detect the events. In another design, the ExoRoboWalker was developed using rotational motors and potentiometers to control motor rotation; when coupled with a rigid insole, it is responsible for the movement of lifting the foot [16].

Although the improvements observed during the interventions - such as reduced compensatory movements, longer strides and greater symmetry between the paretic and unaffected sides - are encouraging, they were so far based on the researchers' observation during the sessions. Further analyses will be carried out, using validated tools such as the Zeno Walkway and Qualisys systems, in order to objectively quantify the gait differences already observed.

Gait cycle time is an important variable for determining walking patterns, and a reduction in this measure is indicative of improvements in functional gait parameters, such as speed, stride length and symmetries [18]. In our study, we observed a reduction in gait cycle time with the exoskeleton following the assisted walking interventions, indicating that its use has potential for rehabilitation and functional recovery. Walking without the exoskeleton following the intervention also showed a reduction in gait cycle time, indicating that functional improvement is maintained even without the assisted dorsiflexion provided by the device.

The analysis was carried out considering only one of the parameters provided by the Zeno platform. Further kinetic and kinematic analyses are currently being carried out. Future studies include conducting assisted walking interventions for patients with foot drop, followed by analysis of movement kinematics and kinetics, as well as assessments of clinical functionality regarding balance, strength, and spasticity.

V. Conclusion

We can conclude that the developed exoskeleton is able to effectively recognize the swing phase of gait in patients with foot drop. The sensors, in the quantities and positions previously described, were capable of accurately determining the swing phase of the gait and, as a result, reliably activating the motor to provide assisted dorsiflexion and aid the patient's walking. Throughout testing, the device maintained consistent

accuracy in gait phase recognition and dorsiflexion assistance, all achieved with low-cost and accessible instrumentation. Positioning the sensors and motor on the shin and foot resulted in a comfortable and adaptable design, which may facilitates its use both in supervised rehabilitation sessions and in unsupervised activities of daily living, reinforcing the device's potential as a practical and reliable rehabilitation tool. The reduction in gait cycle time following the interventions suggests that the exoskeleton intervention has the potential to facilitate functional walking rehabilitation. The reduction in gait cycle time when walking without the exoskeleton suggests that the benefits of the exoskeleton intervention are maintained even without movement assistance. Further analyses, using validated tools such as the Zeno Walkway and Qualisys systems, are needed to objectively quantify the gait improvements already observed and to further determine the effectiveness of the exoskeleton in rehabilitation and activities of daily living for patients with foot drop.

REFERENCES

[1] Johnson, C.O., Nguyen, M., Roth, G.A., Nichols, E., Alam, T., Abate, D., Abd-Allah, F., Abdelalim, A., Abraha, H.N., Abu-Rmeileh, N.M. et al. Global, regional, and national burden of stroke, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol., 2019. 18.

[2] Cohen, J. C. Biomechanical evaluation of leprosy patients with foot drop: pre-operative and post surgical analysis with tendon transfer. Rio de Janeiro; s.n; 2024. 80 p. ilus, tab, graf.

[3] Mao, Y. R., Zhao, J. L., Bian, M. J., others. Spatiotemporal, kinematic and kinetic assessment of the effects of a foot drop stimulator for home-based rehabilitation of patients with chronic stroke: a randomized clinical trial. Journal of NeuroEngineering and Rehabilitation. BioMed Central. 56. 10.1186/s12984-022-01036-0. 2022.

[4] Lüdorf, Sílvia Maria Agatti. A caminhada como prática de resistência: uma análise a partir da visão sociológica de David Le Breton. Programa de Pós-Graduação em Sociologia - UFRGS. Sociologias. 23. 368-394. 2021.

[5] Allami, M., Yavari, A., Karimi, A., Masoumi, M., Soroush, M., Faraji, E. (2017). Health-related quality of life and the ability to perform activities of daily living: a cross-sectional study on 1079 war veterans with ankle-foot disorders. Military Medical Research, 4(1), 37. https://doi.org/10.1186/s40779-017-0146-1.

[6] Zorkot M, Dac LH, Morya E, Brasil FL. G-Exos: A wearable gait exoskeleton for walk assistance. Front Neurorobot, 2022. 15, Article 939241. 10.3389/fnbot.2022.939241.

[7] Elayoubi J, Haley WE, Nelson ME, Hueluer G. How Social Connection and Engagement Relate to Functional Limitations and Depressive Symptoms Outcomes After Stroke. Stroke. 2023 Jul;54(7):1830-1838. doi: 10.1161/STROKEAHA.122.042386. Epub 2023 Jun 26. PMID: 37363947; PMCID: PMC10313154.

[8] Awad LN, Bae J, O'Donnell K, Rossi SMD, Hendron K Sloot LH, Kudzia, P, Allen S, Holt KG, Ellis TD, et al. A soft robotic exosuit improves walking in patients after stroke. Science translational medicine. American Association for the Advancement of Science,2017, v. 9, n. 400.

[9] Veale AJ, Xie SQ. Towards compliant and wearable robotic orthoses: A review of current and emerging actuator technologies. Medical engineering e physics, Elsevier, v. 38, n. 4, p. 317–325, 2016.

[10] Hasan, S.; Alam, N. Biomechanical design and adaptive sliding mode control of a human lower extremity exoskeleton for rehabilitation applications. Robotics, v. 14, n. 146, 2025.

[11] Kim JH Tsukaya H. Regulation of plant growth and development by the growthregulating factor and grf-interacting factor duo. Journal of experimental botany, Oxford University Press UK, v. 66, n. 20, p. 6093–6107, 2015.

[12] Cooper N., et al. Design & Systematic Evaluation of Power Transmission Efficiency of an Ankle Exoskeleton for Walking Post-Stroke. 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE. 2024. 3753–3760. doi:10.1109/ICRA57147.2024.10610736.

[13] Louis N. Awad et al. ,A soft robotic exosuit improves walking in patients after stroke.Sci. Transl. Med.9, eaai9084(2017).DOI:10.1126/scitranslmed.aai9084.

[14] Yeung, LF., Lau, C.C.Y., Lai, C.W.K. et al. Effects of wearable ankle robotics for stair and over-ground training on sub-acute stroke: a randomized controlled trial. J NeuroEngineering Rehabil 18, 19 (2021). Available from: https://doi.org/10.1186/s12984-021-00814-6.

[15] Zhang, X., Liu, Y. X., Wang, R., & Gutierrez-Farewik, E. M. (2024). Soft ankle exoskeleton to counteract dropfoot and excessive inversion. Frontiers in neurorobotics, 18, 1372763. https://doi.org/10.3389/fnbot.2024.1372763

[16] Andrade, R. M., Sapienza, S., Mohebbi, A., Fabara, E. E., & Bonato, P. (2023). Overground Walking With a Transparent Exoskeleton Shows Changes in Spatiotemporal Gait Parameters. IEEE journal of translational engineering in health and medicine, 12, 182–193. https://doi.org/10.1109/JTEHM.2023.3323381

[17] Lee, J. H., Tan, Y. X., Wong, C. H., Cheah, C. Y., Chong, Y. Z., Goh, C.-H. (2024). Development of Pneumatic-Based Exoskeleton System for Ankle Dorsiflexion and Plantarflexion. 2024 IEEE 12th Region 10 Humanitarian Technology Conference (R10-HTC), 1-6. https://doi.org/10.1109/r10-htc59322.2024.10778837.

[18] Aleksandrowicz, K., Kosowski, W., Michalska, A., Winiarski, S. (2025). SPM Differences in Gait Pattern of Women After Total Hip Replacement: A Longitudinal Study. Journal of Clinical Medicine, 14. https://doi.org/10.3390/jcm14124316

Development of PULSEGRID: An Interactive Floor Platform for Cognitive-Motor Activities Adaptable to Neurodivergent Individuals

1st Leonardo Xavier do Nascimento
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
lx.nascimento@unesp.br

2nd Edinaldo Serra Cardoso Júnior
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
edinaldo.junior@unesp.br

3rd Rafael José Ginuino da Silva
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
rafael.ginuino@unesp.br

4th Sara Moreira Macedo
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
sara.macedo@unesp.br

5th Bodouowei Godswill Zipamone
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
b.zipamone@unesp.br

6th Pablo Mauricio Portilla Hernandez
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
pablo.portilla@unesp.br

7th Priscila da Silva Oliveira
Federal University of Tocantins (UFT)
Palmas – TO, Brasil
pri.s.o@uft.edu.br

8 h Adalberto Vieira Corazza
Department of Medicine
Federal University of Mato Grosso do Sul (UFMS)
Três Lagoas – MS, Brasil
adalberto.corazza@ufms.br

9 th Aparecido Augusto de Carvalho
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
aa.carvalho@unesp.br

10th Marcelo Augusto Assunção Sanches
Department of Electrical Engineering
São Paulo State University (UNESP)
Ilha Solteira – SP, Brasil
marcelo.sanches@unesp.br

Abstract—**This paper presents the development of the PULSEGRID prototype, an interactive floor system designed for cognitive and motor activities based on the reproduction of audiovisual sequences. The equipment consists of four interactive blocks identified by different colors when illuminated, addressable LED lighting, infrared sensors, an audio system, a numerical display, and an ESP32 microcontroller. Interaction occurs when the system presents a sequence of lights and sounds that must be reproduced by the user by touching or stepping on the corresponding blocks. An Android application connected via Bluetooth enables the adjustment of stimulus duration, the interval between sequences, the maximum response time, and the number of stages in the activity. PULSEGRID was conceived as an adaptable platform for research in biomedical engineering, motor activities, special education, and studies involving neurodivergent individuals. Its configuration enables working memory, attention, reaction time, motor planning, and visuospatial orientation to be combined in a full-body activity. Previous studies indicate that motor interventions and active games show promising results regarding motor skills and executive functions in children with neurodevelopmental disorders. However, technical evaluations, usability testing, and clinical studies are still required to validate the acceptability and potential functional effects of the equipment.**

Keywords—**Biomedical Engineering, Exergame, Neurodivergence, Cognitive-Motor Stimulation, Interactive Floor.**

I. INTRODUCTION

The development of interactive technologies applied to health has enabled new approaches to the assessment, stimulation, and rehabilitation of cognitive and motor functions. Among these technologies are active video games, also referred to as exergames, in which body movements are used as interaction commands [1].

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

Unlike games controlled exclusively by buttons or screens, exergames may require body displacement, postural adjustments, spatial orientation, and coordination among visual, auditory, and motor stimuli. These characteristics make it possible to integrate physical activity and cognitive demands within the same task [2].

This approach may be particularly relevant for neurodivergent individuals, including people with autism spectrum disorder, attention-deficit/hyperactivity disorder, and developmental coordination disorder. However, these populations are highly heterogeneous, making it necessary to adapt stimulus intensity, activity speed, and the time available for response [3].

Autistic children may exhibit differences in balance, bilateral coordination, motor planning, and movement execution [4]. Differences in sensory processing may also influence how visual, auditory, and bodily stimuli are perceived and translated into motor responses [5].

Studies involving structured physical activity have shown promising results regarding motor proficiency and components of executive functioning in autistic children [6]. Exergame-based interventions have also been investigated in children with neurodevelopmental disorders, particularly because they combine movement, attention, motivation, and immediate feedback [7], [8].

In this context, PULSEGRID was developed as an interactive floor system composed of four response regions. The system presents visual and auditory sequences that must be memorized and reproduced by the user by touching the corresponding blocks. This study aims to present the architecture, operation, and potential applications of PULSEGRID as an adaptable technological platform for cognitive and motor activities.

II. MATERIALS AND METHODS

This study is part of a broader research project conducted collaboratively by the Instrumentation and Biomedical Engineering Laboratory (LIEB) and the Movement and Medical Technologies Laboratory (LAMOTEM), approved by the Research Ethics Committee of the Federal University of Mato Grosso do Sul (UFMS), under CAAE no. 71042323.7.0000.0021. However, the stage reported in this paper is restricted exclusively to the conceptual, hardware, firmware, and software development of the PULSEGRID platform. No participants were recruited, no experimental intervention was conducted, and no human data were collected or analyzed during this stage of technological development. Future usability investigations and clinical studies involving participants will be conducted according to the procedures established in the approved research protocol.

A. PULSEGRID Architecture and Physical Structure

PULSEGRID was designed as a whole-body interactive system in which the user responds to presented stimuli through four blocks positioned on the floor. Each block corresponds to a color—yellow, blue, red, or green—and acts simultaneously as a visual stimulation element and a system input area, as shown in Fig. 1.



Fig. 1. PULSEGRID platform.

The PULSEGRID structure was built from MDP panels, with the surface divided into four independent blocks arranged in a 2 × 2 matrix configuration. Each block consists of an acrylic panel bordered by aluminum angle profiles, which provide finishing, protect the edges, and support the installation of addressable LED strips. Positioned along the inner sides of the blocks, these strips illuminate the acrylic panels and enable individual visual identification of each interaction area.

User interaction is detected by a set consisting of an infrared emitter and receiver installed face-to-face at the midpoints of two opposite sides of each block. This arrangement establishes a transverse optical barrier across the respective interaction area. When the user touches or steps on the block, interruption of the infrared beam is identified by the system as an input. The assembly is completed by screws, spacers, supports, and electrical wiring, which secure the components and integrate the mechanical structure with the lighting and detection systems.

PULSEGRID operates by presenting a sequence of stimuli that progressively expands. Initially, the system activates one of the blocks through light and sound stimuli. The user must then touch or step on the corresponding block. After the entire sequence is reproduced correctly, a new element is added and the user must reproduce the complete sequence again. This process continues until the previously configured number of stages is completed, an incorrect response occurs, or the maximum response time is exceeded.

B. Control Module

PULSEGRID uses an ESP32 microcontroller as its central processing unit, responsible for integrating the sensors, visual and auditory systems, and Bluetooth communication with the mobile application. The main electronic components are integrated into a control module connected to the platform by cable (Fig. 2).

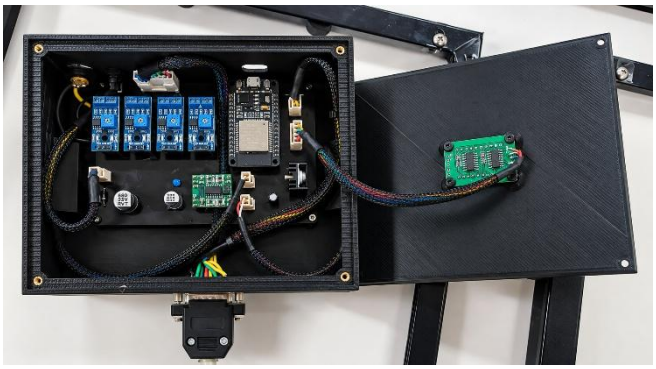


Fig. 2. Internal view of the PULSEGRID electronic control module.

The control module contains a main circuit board (motherboard) integrating the ESP32 microcontroller; the two-digit seven-segment display module; the infrared sensor modules; the PAM8304 amplifier module; the PCM5102A audio digital-to-analog converter (DAC) module with an I²S interface; and all connectors for the remaining components. The motherboard was designed using KiCad® software and fabricated at LIEB using a small CNC milling machine, enabling all components to be integrated and organized within a compact structure.

C. *Detection System and Visual Stimulation System*

The equipment uses one infrared sensor module for each block. The emitter and receiver were positioned face-to-face at the midpoints of opposite sides (Fig. 3), forming an optical barrier along the beam crossing the central region of the block. When the user's hand or foot interrupts this beam, the sensor's electrical signal changes, and this variation is detected and processed by the ESP32 microcontroller.

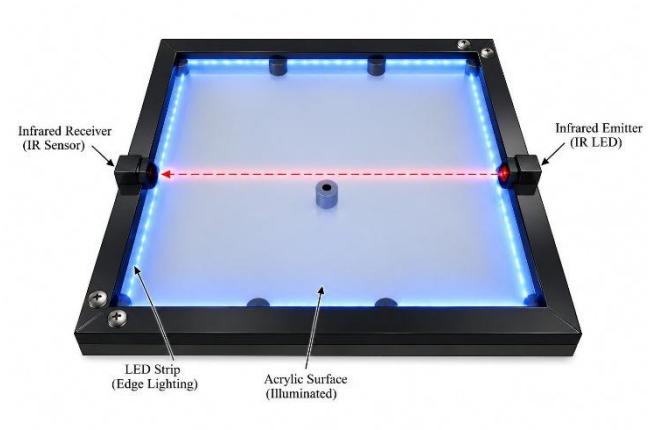


Fig. 3. Illustration of a PULSEGRID block.

To encourage touching the block, in addition to the auditory stimulus, PULSEGRID's visual response is produced by an LED strip containing 72 addressable LEDs (Fig. 3), individually controlled by the ESP32 and distributed so as to form four colored floor blocks. Thus, the visual stimulation system is used to indicate the blocks included in the sequence, guide the user during the activity, and signal the start, correct response, error, and victory conditions.

D. *Auditory Stimulation System*

PULSEGRID's auditory stimulation is digitally synthesized by the ESP32 and transmitted to an external digital-to-analog converter through the I²S protocol. The resulting analog signal is then sent to a stereo amplifier and reproduced through 5 W loudspeakers.

Each block is assigned a characteristic sound frequency, establishing an association among spatial position, color, and sound. The frequencies were defined based on the musical notes C6, E6, G6, and C7, which correspond, respectively, to a harmonic relationship formed by the tonic, major third, perfect fifth, and upper octave (Table I). This arrangement was adopted to provide greater musical coherence to the auditory stimuli generated during sequence playback.

TABLE I
SOUND FREQUENCIES ASSOCIATED WITH THE PULSEGRID BLOCKS

Color	Frequency	Approximate note	Musical relation
Green	1046 Hz	C6	Tonic
Red	1318 Hz	E6	Major third
Blue	1568 Hz	G6	Perfect fifth
Yellow	2092 Hz	C7	Upper octave

The system also uses specific sound effects to indicate the start of the activity, correct responses, errors, and completion of the sequence. The simultaneous presentation of visual and auditory stimuli provides informational redundancy, allowing the user to identify each block both by its color and by its characteristic sound. However, this multimodal configuration may be made adjustable in future protocols, allowing visual-only, auditory-only, and audiovisual conditions to be evaluated separately.

E. *Embedded System Firmware*

The firmware was developed based on a state-machine architecture that organizes the different stages of system operation and coordinates command execution according to the active state. This architecture comprises the following states: initialization of peripherals and the communication interface; standby mode, in which the system waits for a touch on one of the blocks or receipt of a command from the application; sequence generation and storage; playback of visual and auditory stimuli; acquisition and verification of the user's response; sequence expansion after its complete and correct reproduction; error or timeout signaling; execution of victory stimuli; and return to standby mode for the start of a new activity.

The control architecture also manages settings received while an activity is in progress. When a change is sent by the application while a sequence is running, it remains pending and is applied only after completion of the current activity. Thus, changes to execution times or difficulty level do not unexpectedly interfere with the sequence in progress.

To filter rapid signal fluctuations, the firmware performs temporal confirmation of sensor activation before considering the input valid. This procedure is intended to reduce improper detections associated with electrical noise and brief beam interruptions during incomplete movements. The effectiveness of this treatment in reducing false activations should be quantified in specific tests.

F. *Mobile Application*

The PULSEGRID application was developed for Android devices using MIT App Inventor© and communicates with the ESP32 via Bluetooth. As illustrated in Fig. 4, its interface allows the user to start the activity, transmit configuration parameters, and restore the system's default settings.



Fig. 4. Main interface of the PULSEGRID application.

The interface allows four parameters to be adjusted: transition speed between stages; duration of the visual and auditory stimuli; maximum time allowed for the user to touch the block; and level, corresponding to the number of stages required to complete the activity.

The application also receives status messages from the ESP32 indicating whether the settings were applied, whether an activity is already in progress, or whether a communication failure occurred. These adjustments make it possible to adapt activity speed, response time, and difficulty level according to the user's characteristics and performance.

G. System Operational Flow

The system's decision-making logic was structured according to the flowchart shown in Fig. 5.

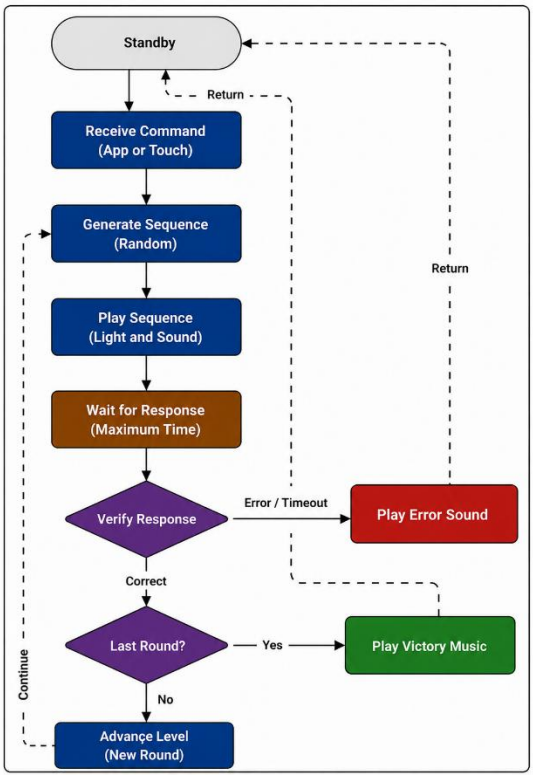


Fig. 5. PULSEGRID flowchart.

When the equipment is powered on, the ESP32 initializes the sensors, lighting system, audio system, display, memory, and Bluetooth communication interface. After initialization, the system enters standby mode, in which the blocks remain available to start the activity. A session can be started by touching any of the blocks or by pressing the Play button in the mobile application.

After the start command is received, the system generates the first element of the sequence. The corresponding block is illuminated and its associated sound is reproduced for the interval defined by the light-and-sound duration parameter. After the stimulus is presented, the equipment enters the response-waiting state. The user must touch the corresponding block within the configured maximum response time. The infrared sensor detects interruption of the beam and transmits the resulting signal to the ESP32.

The detected response is then compared with the stored sequence element. When the response is correct, the system records the successful input and proceeds with the activity. An incorrect response triggers error feedback, whereas the absence of a response within the established interval ends the attempt due to timeout.

After the complete sequence is reproduced correctly, a new element is added. The system then presents the expanded sequence in its entirety, progressively increasing working-memory and motor-planning demands. When the number of stages defined in the application is reached, the equipment executes the victory pattern. After successful completion, an incorrect response, or a timeout, the system automatically returns to standby mode.

Preliminary bench functional verification included sequence generation and playback, input detection, progression after correct reproduction, error and timeout signaling, and return to standby mode. Transmission of settings by the application, status messages, and application of pending changes after completion of the activity were also verified. This verification was qualitative, without measurement of latency, repeatability, or false-trigger rate.

III. RESULTS AND DISCUSSION

Development resulted in a functional PULSEGRID prototype composed of four interactive blocks, infrared sensors, addressable LEDs, an audio system, a numeric display, an ESP32-based control unit, and an application for Android® devices. Preliminary bench tests indicated integrated operation of these components while executing the proposed activity. The system generated, stored, and reproduced pseudorandom sequences of visual and auditory stimuli, while the sensors detected activations performed on each block. The temporal confirmation implemented in the firmware was used to filter rapid signal variations before responses were recognized. However, its effectiveness in reducing unintended activations was not quantified at this stage.

Correct reproduction of the entire sequence allowed progression to the next stage, whereas incorrect responses or expiration of the available time triggered error signaling. The lighting identified the blocks in the sequence and signaled the start, error, and successful completion of the activity. The audio system reproduced the characteristic sounds of each block and the effects associated with these conditions, while the numeric display indicated progression through the stages.

At the end of each attempt, the equipment automatically returned to standby mode.

Bluetooth communication between the ESP32 and the application made it possible to configure the transition speed between stages, stimulus duration, maximum response time, and activity level, corresponding to the number of stages required for completion. The application also received messages regarding application of settings, the existence of an activity in progress, and possible communication failures. Changes received during an activity remained pending until it ended, preserving the parameters of the sequence in progress.

These results indicate the feasibility of integrating the components and functions planned for the prototype under the evaluated bench conditions. The available adjustments make it possible to modify the pace of the activity, the time allowed for response, and the number of elements to be memorized, establishing a basis for future evaluations with users of different profiles.

From an operational standpoint, infrared-barrier detection does not require cameras, wearable devices, or handheld controllers, which may facilitate platform setup and use in certain environments. However, the tests presented were limited to preliminary functional verification of the components and activity execution. Technical validation should include measurements of sensor response time, detection consistency during repeated activations, false-trigger frequency, communication delay, and stability during prolonged operation.

Usability evaluations and studies with participants will also be required to verify comprehension of the instructions, comfort, acceptance, and suitability of the tasks for different user profiles. Therefore, the results of this stage demonstrate the preliminary operation of the prototype, while its possible cognitive and motor benefits remain questions for future investigations.

IV. Potential Applications and Perspectives

The proposed activity combines sequence memorization with the identification and activation of blocks at different positions. Depending on the adopted interaction mode, the task may involve attention, sequential memory, visuospatial orientation, and motor planning. When performed with the feet, it may also require body displacement, weight transfer, and balance during movement. These characteristics describe the expected demands of the activity, whose effects on user performance still need to be investigated.

To expand the possibilities for adaptation to individual sensory characteristics, future versions may include independent adjustments of light intensity and volume, as well as the option to disable auditory stimuli. Evaluation of these features should consider ease of stimulus perception, comfort during interaction, and understanding of the task.

A. Autism Spectrum Disorder

The association among audiovisual stimuli, memorization, and movement allows PULSEGRID to be considered as a platform for investigating complementary activities aimed at autistic children and adolescents. Studies using active video games, also referred to as exergames, present findings that motivate this investigation, although they are not yet conclusive. Edwards et al. reported the feasibility of an active sports game intervention targeting object-control skills [7]. Miranda et al., in a randomized crossover study with nine autistic children, observed reduced reaction time in an inhibitory-control task after an exergame session, without consistent improvement in response accuracy [9].

The small sample size and assessment of the effects of a single session limit the generalizability of these results. Even so, the studies provide a basis for investigating tasks that combine movement and cognitive demands. In PULSEGRID, sequence presentation, user response, and result signaling follow a defined organization. Future studies may evaluate whether this organization supports understanding of the activity and its acceptance by users, taking into account individual differences in sensory and motor characteristics.

B. Attention-Deficit/Hyperactivity Disorder

The combination of physical exercise and cognitive tasks has been investigated as a complementary approach for children with attention-deficit/hyperactivity disorder (ADHD). In a randomized clinical trial, Benzing and Schmidt reported effects of an exergame program on executive functions and motor skills, highlighting the importance of adapting games to the population being studied [8].

In PULSEGRID, the need to wait until sequence presentation has ended before responding makes it possible to explore tasks related to inhibitory control, understood as the ability to suppress premature or inappropriate responses. The progressive increase in sequence length expands the amount of information to be maintained and reproduced, while choosing among the four blocks involves identifying positions and selecting responses. These characteristics may guide studies of attention, working memory, and response control. Quantitative analysis of reaction time will additionally depend on the implementation and validation of a feature for recording this measure.

C. Coordination and Motor Planning Difficulties

The arrangement of the four blocks makes it possible to propose tasks with different movement directions and response positions. Depending on the protocol, interaction may involve laterality, weight transfer, dynamic balance, and precision in positioning the feet or hands. In the current version, the positions and distances between blocks are fixed, while activity pace, response time, and number of stages can be adjusted through the application.

Reproducing sequences in reverse order represents a possible extension of the tasks, but requires adaptation of the operating logic. Use of the platform in clinical or rehabilitation contexts should be investigated through specific protocols, with professional supervision and activities compatible with participants' conditions.

D. Interaction and Cooperative Activities

Interactive platforms may also offer shared activities between participants. Nekar et al. evaluated a multiplayer augmented-reality system with cognitive and motor tasks in 14 autistic children. The study reported feasibility and good acceptance, but highlighted the need for randomized trials with larger samples [10]. As a development perspective, PULSEGRID may incorporate activities in which two users reproduce alternating parts of a sequence or perform a task together. Recognition of simultaneous responses and programmed alternation between participants will depend on adaptations to the firmware and sensor-signal processing.

The use of the floor as an interaction space was also explored by Panceri et al., who integrated serious games projected onto the floor with a socially assistive robot in activities involving autistic children and children with Down

syndrome. The pilot study demonstrated the technical feasibility of this form of interaction [11], providing a reference for the development of activities that use physical space as part of the task.

More broadly, studies on physical activity and exergames contribute to the rationale for the proposed applications. Prieto et al., for example, reported improvements in fundamental motor skills after a parent-mediated physical activity intervention [12]. These findings justify further investigation, but cannot be attributed directly to PULSEGRID because they were obtained with other interventions, protocols, and populations.

V. CONCLUSIONS

This study presented the conceptual and technological development of PULSEGRID, an interactive floor system controlled by an ESP32 and composed of four blocks equipped with sensors, addressable lighting, auditory feedback, a numeric display, and a mobile application.

The architecture enables the development of progressive activities that associate visual and auditory stimuli with bodily responses. Adjustment of stimulus duration and speed, response time, and number of stages makes the system potentially adaptable to users with different motor, cognitive, and sensory characteristics.

From a biomedical engineering perspective, PULSEGRID can serve as a research platform for the development and evaluation of cognitive-motor tasks, with the potential to support future acquisition of objective performance measures. Its main characteristics include a modular structure, relatively low component cost, whole-body interaction, and remote configuration.

Although studies of motor interventions and exergames have reported promising findings in autistic individuals and people with ADHD, this evidence does not demonstrate the effectiveness of PULSEGRID. Technical tests, usability evaluations, and controlled clinical studies are still needed to determine its safety, acceptability, and potential functional effects.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

[1] A. F. de O. Baracho, F. J. Gripp and M. R. de Lima, “Exergames and the school physical education in the digital culture,” Rev. Bras. Ciênc. Esporte, vol. 34, pp. 111–126, 2012, doi: 10.1590/S0101-32892012000100009.

[2] D. H. K. Chow and S. K. F. Mann, “Exergaming and education: a relational model for games selection and evaluation,” Front. Psychol., vol. 14, Jul. 2023, doi: 10.3389/fpsyg.2023.1197403.

[3] K. Spiel and K. Gerling, “The Purpose of Play: How HCI Games Research Fails Neurodivergent Populations,” ACM Trans. Comput.-Hum. Interact., vol. 28, no. 2, p. 11:1–11:40, Apr. 2021, doi: 10.1145/3432245.

[4] M. Kaur, S. M. Srinivasan and A. N. Bhat, “Comparing motor performance, praxis, coordination, and interpersonal synchrony between children with and without Autism Spectrum Disorder (ASD),” Res. Dev. Disabil., vol. 72, pp. 79–95, Jan. 2018, doi: 10.1016/j.ridd.2017.10.025.

[5] T. Liu, “Sensory Processing and Motor Skill Performance in Elementary School Children with Autism Spectrum Disorder,” Percept. Mot. Skills, vol. 116, no. 1, pp. 197–209, Feb. 2013, doi: 10.2466/10.25.PMS.116.1.197-209.

[6] C.-Y. Pan, C.-H. Chu, C.-L. Tsai, M.-C. Sung, C.-Y. Huang and W.-Y. Ma, “The impacts of physical activity intervention on physical and cognitive outcomes in children with autism spectrum disorder,” Autism, vol. 21, no. 2, pp. 190–202, Feb. 2017, doi: 10.1177/1362361316633562.

[7] J. Edwards, S. Jeffrey, T. May, N. J. Rinehart and L. M. Barnett, “Does playing a sports active video game improve object control skills of children with autism spectrum disorder?,” Journal of Sport and Health Science (JSHS), vol. 6, no. 1, pp. 17–24, Mar. 2017, doi: 10.1016/j.jshs.2016.09.004.

[8] V. Benzing and M. Schmidt, “The effect of exergaming on executive functions in children with ADHD: A randomized clinical trial,” Scand. J. Med. Sci. Sports, vol. 29, no. 8, pp. 1243–1253, Aug. 2019, doi: 10.1111/sms.13446.

[9] J. M. Miranda, R. A. V. Browne, W. Q. A. da Silva, J. P. Rodrigues Dos Santos, C. S. G. Campbell and I. A. Ramos, “Effects of a Session of Exergames and Traditional Games on Inhibitory Control in Children With Autism Spectrum Disorder: Randomized Controlled Crossover Trial,” JMIR Serious Games, vol. 13, p. e65562, Mar. 2025, doi: 10.2196/65562.

[10] D. M. Nekar, H. Kang, H. Alao and J. Yu, “Feasibility of Using Multiplayer Game-Based Dual-Task Training with Augmented Reality and Personal Health Record on Social Skills and Cognitive Function in Children with Autism,” Children, vol. 9, no. 9, p. 1398, Sep. 2022, doi: 10.3390/children9091398.

[11] J. A. C. Panceri, É. Freitas, J. C. de Souza, S. da Luz Schreider, E. Caldeira and T. F. Bastos, “A New Socially Assistive Robot with Integrated Serious Games for Therapies with Children with Autism Spectrum Disorder and Down Syndrome: A Pilot Study,” Sensors (Basel), vol. 21, no. 24, p. 8414, Dec. 2021, doi: 10.3390/s21248414.

[12] L. A. Prieto et al., “A randomized parent-mediated physical activity intervention for autistic children,” Autism Res., vol. 16, no. 7, pp. 1450–1461, Jul. 2023, doi: 10.1002/aur.2969.

Aplicação de um Fluxo de Manufatura Digital para Próteses Transradiais Passivas Personalizadas em um Serviço Público de Reabilitação: Estudo de Caso

Eduardo Keller Rorato
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
UNIFESP
São José dos Campos, Brasil
ek.rorato@unifesp.br

Beatriz Fernanda de Oliveira Basso
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
UNIFESP
São José dos Campos, Brasil
beatriz.basso@unifesp.br

Andrea de Oliveira Ramponi
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
UNIFESP
São José dos Campos, Brasil
psi.andrearamponi@gmail.com

Cinthia Hermínia Carvalho Nascimento Pereira
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
UNIFESP
São José dos Campos, Brasil
cinthia.carvalho@unifesp.br

Cíntia Biancon Katata
Unidade de Reabilitação
Prefeitura de São José dos Campos
São José dos Campos, Brasil
cintia.biancon@sjc.sp.gov.br

Renata Cardoso Silva
Unidade de Reabilitação
Prefeitura de São José dos Campos
São José dos Campos, Brasil
renata.cardoso@sjc.sp.gov.br

Marcela Larissa de Souza Lobato
Unidade de Reabilitação
Prefeitura de São José dos Campos
São José dos Campos, Brasil
marcelalobato.fisio@hotmail.com

Maria Elizete Kunkel
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
UNIFESP
São José dos Campos, Brasil
elizete.kunkel@gmail.com

***Resumo*—A disponibilidade de próteses personalizadas ainda constitui um desafio nos serviços públicos de reabilitação. Este trabalho descreve a aplicação de um fluxo de manufatura digital no desenvolvimento de próteses transradiais passivas para uma participante de 64 anos, com amputação esquerda e preservação de aproximadamente metade do antebraço. A demanda consistia em um dispositivo predominantemente estético, compatível com o uso cotidiano e com a atividade de costura. O fluxo compreendeu escaneamento tridimensional, tratamento da malha, modelagem do soquete, prototipagem por extrusão de material, ajustes presenciais e desenvolvimento de um terminal baseado no modelo Kwawu. Aproximadamente dez versões do soquete em TPU foram avaliadas durante quatro visitas. A configuração final utilizou duas abas sobre os epicôndilos para promover retenção supracondilar, sem cintas adicionais. Foram fabricadas duas próteses completas com o mesmo modelo de soquete e diferentes conexões roscadas. A primeira conexão apresentou menor estabilidade e escape do acoplamento; na segunda, o aumento das dimensões e do comprimento da rosca melhorou a estabilidade durante o reposicionamento do terminal. As próteses apresentaram massas totais de 395 g e 488 g. Após a entrega, a participante relatou perceber a nova prótese como mais leve que a prótese mecânica anteriormente fornecida pelo sistema público e informou utilizar o terminal como apoio passivo durante a costura. Os resultados demonstram a viabilidade inicial do fluxo e a aceitabilidade preliminar dos dispositivos, mas não constituem validação clínica ou funcional, pois não foram empregados instrumentos padronizados, ensaios mecânicos ou acompanhamento prolongado.**

***Index Terms*—manufatura digital, impressão 3D, prótese transradial, tecnologia assistiva, engenharia de reabilitação**

I. Introdução

A amputação de membro superior produz efeitos que ultrapassam a perda segmentar, podendo limitar atividades cotidianas e ocupacionais, modificar a participação social e repercutir sobre a percepção corporal [1], [2]. Nas amputações transradiais, a preservação da articulação do cotovelo mantém possibilidades relevantes de adaptação protética; ainda assim, a ausência da mão compromete tarefas que dependem de preensão, manipulação de objetos e coordenação bimanual [3].

As próteses de membro superior podem responder a objetivos distintos, como compensar limitações funcionais, apoiar atividades específicas ou recompor a aparência corporal. As próteses passivas podem ser estáticas ou ajustáveis; nas soluções de finalidade predominantemente estética, não há restituição de preensão ativa, e o benefício esperado relaciona-se

sobretudo à recomposição do contorno do membro e, conforme as prioridades da pessoa, à confiança e à participação social [2], [4]. Por isso, o dispositivo deve ser definido e avaliado de acordo com sua finalidade e com as prioridades da pessoa, sem equiparar uma prótese estética a um dispositivo destinado à restauração motora [1].

Independentemente do tipo de prótese, sua aceitação e seu uso continuado são influenciados por conforto, funcionalidade, adequação do encaixe, peso, aparência, confiabilidade e facilidade de colocação e retirada. Encaixe inadequado, dor ou desconforto, peso excessivo e incompatibilidade entre o dispositivo e as necessidades individuais são fatores frequentemente associados à insatisfação e ao abandono protético [1], [3], [5]. Como cada membro residual apresenta características anatômicas próprias, a personalização deve abranger o encaixe, a forma externa e o contexto de uso, e não somente medidas lineares ou a escala do dispositivo.

Nos processos convencionais, o encaixe também é individualizado, mas sua fabricação costuma depender de moldes físicos, etapas manuais e ajustes sucessivos. Embora esses procedimentos possam produzir dispositivos adequados, eles demandam trabalho especializado e não geram, por si, um registro digital reutilizável da anatomia e das modificações realizadas no projeto [6].

Os fluxos de manufatura digital integram aquisição tridimensional da anatomia, modelagem computacional e manufatura aditiva. Essa integração permite registrar dados anatômicos, modificar geometrias específicas para cada pessoa, documentar versões e parâmetros de fabricação e produzir novos protótipos sem reiniciar o processo a partir de um molde físico [6]–[8]. Essas possibilidades técnicas, entretanto, não garantem superioridade clínica. A literatura ainda aponta heterogeneidade dos resultados, validação clínica limitada e necessidade de avaliar ajuste, conforto, materiais, durabilidade, segurança e desfechos funcionais [8], [9].

Em serviços públicos, a incorporação desse tipo de fluxo depende de sua compatibilidade com os recursos locais, da qualificação da equipe, da articulação multiprofissional e da continuidade das etapas de avaliação, adaptação e acompanhamento [7], [8]. O presente estudo dá continuidade e aplica, em contexto assistencial, um fluxo digital previamente estruturado no âmbito do Programa de Extensão Mão3D e do Laboratório de Órteses e Próteses 3D da UNIFESP, baseado em biomodelos digitais, modelagem computacional iterativa e manufatura aditiva para o desenvolvimento de próteses passivas personalizadas [10].

A demanda assistencial foi encaminhada, antes do início do desenvolvimento, pela Prefeitura de São José dos Campos, por meio da Unidade de Reabilitação, e consistia na produção de uma prótese predominantemente estética e compatível com a atividade de costura da participante. O trabalho foi realizado em parceria entre o Núcleo de Saúde Digital da UNIFESP, o PET-Saúde Digital, o Programa de Extensão Mão3D, o Laboratório de Órteses e Próteses 3D (LO&P3D/UNIFESP) e a Unidade de Reabilitação da Prefeitura de São José dos Campos. Assim, o objetivo deste trabalho foi descrever a aplicação do fluxo de manufatura digital no projeto e na fabricação de uma prótese transradial passiva, de finalidade predominantemente estética, considerando a anatomia do membro residual e a demanda ocupacional apresentada.

II. Materiais e Métodos

A. Delineamento, contexto assistencial e participante

Trata-se de um estudo aplicado e descritivo, de desenvolvimento tecnológico, conduzido como estudo de caso único, com resultados de natureza qualitativa, no contexto da Unidade de Reabilitação da Prefeitura de São José dos Campos (UR/SJC). A participante, de 64 anos, recebeu diagnóstico de sarcoma no membro superior esquerdo em 2007. Segundo seu relato, a manifestação inicial foi caracterizada por dores intensas e contínuas. A condição resultou em amputação transradial esquerda, com preservação de aproximadamente metade do antebraço. A participante possuía experiência anterior com uma prótese mecânica fornecida pelo SUS.

Os resultados foram registrados de forma descritiva a partir da interação direta da participante com os protótipos e com as próteses completas durante as provas presenciais, os ajustes e o uso inicial dos dispositivos. Foram consideradas as observações da equipe, as modificações efetuadas ao longo do processo iterativo e os relatos da participante sobre encaixe, estabilidade, aparência e utilização do terminal durante a costura.

Os dados quantitativos restringiram-se a informações técnicas, como o número aproximado de versões e visitas e as massas dos dispositivos; não foram aplicados instrumentos padronizados nem realizada análise estatística inferencial.

O protocolo de pesquisa foi registrado no Sistema CEP/Conep sob o CAAE 96891226.0.0000.5505. A participante assinou o Termo de Consentimento Livre e Esclarecido e autorizou a publicação das informações e imagens relacionadas ao caso.

B. Origem da demanda e definição dos requisitos

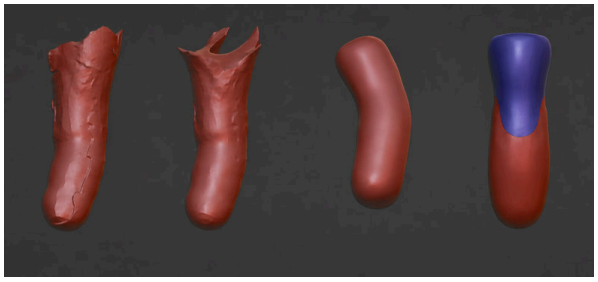
A demanda foi encaminhada pela Prefeitura de São José dos Campos, por meio da Unidade de Reabilitação, antes do início do desenvolvimento. A solicitação consistia em uma prótese predominantemente estética e compatível com a atividade de costura da participante. A partir dessa demanda, foram definidos como requisitos a compatibilidade do soquete com o membro residual, a configuração estética da prótese e sua adequação à atividade informada. A proposta não incluiu mecanismo de preensão ativa.

C. Digitalização e modelagem do soquete

A primeira atividade presencial foi o escaneamento tridimensional do membro residual, realizado na Unidade de Reabilitação por integrantes da área da saúde da equipe. Foi utilizado um escâner EINSTAR (Shining 3D, Hangzhou, China). A malha anatômica obtida foi manipulada e refinada digitalmente mediante correções geométricas, fechamento de superfícies e suavização.



(a)



(b)

Figura 1. Etapas iniciais do fluxo digital: (a) escaneamento tridimensional na Unidade de Reabilitação; (b) estágios de tratamento e refinamento das malhas digitais. Fonte: acervo do projeto.

A Fig. 1 documenta a aquisição tridimensional e o tratamento inicial das malhas digitais.

O soquete foi modelado sobre a malha anatômica no Blender (Blender Foundation, Amsterdam, Países Baixos, versão 3.6 LTS), por meio de ferramentas de escultura digital. Sobre a malha foi criada uma máscara superficial com a geometria do soquete. Na região proximal, foram modeladas duas abas para realizar contato sobre os epicôndilos umerais e promover a retenção ou suspensão supracondilar do soquete.

D. Prototipagem e ajustes presenciais

Foram produzidas aproximadamente dez versões de teste do soquete. As versões selecionadas foram fabricadas em TPU por extrusão de material. Após a sessão de escaneamento, foram realizadas quatro visitas presenciais destinadas às provas e aos ajustes do soquete. Em cada prova, foram observadas a compatibilidade dimensional, a retenção sob tração manual, a facilidade de colocação e retirada, o contato das abas sobre os epicôndilos, a presença de dor ou desconforto relatado e a condição da pele quanto a marcas, irritação ou lesões. As modificações identificadas nessas visitas foram incorporadas às versões digitais subsequentes.

E. Modelagem do terminal, fabricação e montagem

Depois da definição do encaixe, o terminal estético e a conexão roscada foram desenvolvidos a partir do modelo original Kwawu [11]. Os arquivos foram parametrizados no OpenSCAD e posteriormente ajustados no Blender. O terminal passou por ajustes geométricos para esta aplicação, e a geometria negativa da rosca foi incorporada à extremidade distal do soquete.

As versões selecionadas do soquete e o terminal estético foram fabricados em TPU por extrusão de material, enquanto o componente da conexão roscada foi fabricado em PETG pelo mesmo processo. Foram empregados velocidade aproximada de 10 mm/s, temperatura do bico de 240 °C, três paredes e preenchimento de 25% com padrão *gyroid*. A fabricação foi realizada sem suportes. Após a fabricação, os componentes foram submetidos a acabamento e montagem.

A Fig. 2 reúne uma etapa de prova do soquete e os componentes empregados na montagem roscada.

Foram fabricadas duas próteses completas. Posteriormente, foram realizadas duas visitas destinadas à entrega e à avaliação inicial dessas próteses. Desse modo, o processo presencial compreendeu uma sessão de escaneamento, quatro visitas de teste e ajuste do soquete e duas visitas de entrega das próteses completas.

III. Resultados e Discussão

A. Atendimento aos requisitos da demanda

A prótese utilizada anteriormente pela participante era percebida como pesada e rígida. Embora ela conseguisse costurar com o dispositivo, relatou dificuldade para executar a atividade, interferência da prótese na movimentação e impossibilidade de utilizá-la confortavelmente por períodos prolongados. A demanda para o novo desenvolvimento compreendeu uma configuração predominantemente estética, com menor peso percebido, aparência compatível com suas preferências e possibilidade de utilização em atividades cotidianas e durante a costura. Peso, conforto, funcionalidade e compatibilidade com as necessidades individuais são fatores relacionados à aceitação e ao abandono protético [1], [3], [5].

O dispositivo desenvolvido permaneceu classificado como prótese passiva, sem mecanismo de preensão ativa. Durante o desenvolvimento, entretanto, o reposicionamento angular do terminal foi incorporado como possibilidade de adaptação à tarefa. Essa distinção deve ser mantida: o dispositivo não substitui a função motora da mão, mas pode atuar como apoio passivo em atividades específicas [4].

B. Desenvolvimento iterativo do soquete

Foram produzidas aproximadamente dez versões do soquete em TPU. Como não houve registro sistemático individual de todas as impressões, esse número deve ser apresentado como estimativa. O processo de ajuste ocorreu durante quatro visitas presenciais, nas quais os problemas identificados foram incorporados às versões digitais subsequentes. O registro e a modificação de geometrias entre versões integram as possibilidades dos fluxos de manufatura digital [6]–[8].

Na primeira visita, constatou-se incompatibilidade dimensional entre o modelo impresso e o membro residual. O problema foi atribuído à ausência de calibração adequada na aquisição inicial, e um novo escaneamento foi realizado. Na segunda visita, o soquete ainda apresentava folga excessiva. As regiões proximais destinadas ao contato sobre os epicôndilos não apresentavam concavidade suficiente, motivo pelo qual as



(a)



(b)

Figura 2. Prototipagem e montagem: (a) prova de uma versão intermediária do soquete; (b) terminal estético e componente distal roscado. A configuração em (a) é anterior ao sistema final sem cintas. O painel (b) foi aprimorado por inteligência artificial a partir de fotografia do acervo.

abas foram deslocadas para o interior e sua geometria foi acentuada. Na terceira visita, o encaixe estava próximo da configuração desejada e foram realizados ajustes geométricos localizados. Na quarta visita, foi definida a geometria final.

O soquete final utilizou exclusivamente as duas abas sobre os epicôndilos como sistema de retenção, sem cintas ou faixas adicionais. Durante a verificação manual, o dispositivo não se desprendeu sob manipulação habitual. Sua retirada exigia tração deliberada por outra pessoa ou um movimento rotacional intencional realizado pela participante. A colocação e a retirada foram consideradas fáceis. A participante não relatou dor durante as provas, e não foram observadas marcas, irritação ou lesões na pele. Como o tempo de permanência durante os testes foi curto, esses achados não representam avaliação prolongada de conforto.

A definição do encaixe final baseou-se na retenção verificada por tração manual e no retorno fornecido pela participante. O resultado demonstra a viabilidade inicial da suspensão supracondilar empregada, mas não substitui avaliações estruturadas de pressão, conforto e estabilidade durante uso prolongado [3], [5].

C. Terminal estético e conexão roscada

O terminal estético foi desenvolvido a partir do modelo Kwawu citado na metodologia [11]. Por solicitação da participante, os dedos foram discretamente reduzidos e o polegar foi aumentado, buscando aproximar as proporções do terminal das características observadas em sua mão contralateral. A participante também participou da escolha da cor, que constituiu o principal aspecto de personalização estética relatado durante essa etapa.

Foram produzidas duas próteses completas utilizando o mesmo modelo final de soquete. A principal diferença entre elas esteve na conexão roscada. Na primeira prótese, a rosca apresentava dimensões menores e menor comprimento de engate. Embora a participante tenha avaliado positivamente a possibilidade de rotacionar o terminal, essa primeira conexão era mais frágil, podia escapar do acoplamento e limitava a estabilidade das posições selecionadas.

Na segunda prótese, as dimensões e o comprimento da conexão foram ampliados. O maior engate entre os componentes aumentou a estabilidade do terminal durante o reposicionamento e reduziu a ocorrência de escape da rosca. Essa configuração foi adotada como solução definitiva. O terminal podia ser reposicionado angularmente mediante afrouxamento da conexão, rotação até a orientação desejada e novo aperto.

As duas próteses foram entregues à participante. A primeira unidade foi mantida como dispositivo secundário e passou a ser utilizada pela participante durante a hidroginástica. Esse uso foi uma decisão espontânea da participante e não foi acompanhado por ensaio de estanqueidade ou certificação de impermeabilidade; portanto, o estudo não permite afirmar que o dispositivo seja à prova d'água. A segunda prótese, com a conexão ampliada, passou a ser utilizada preferencialmente em atividades sociais e cotidianas.

D. Fabricação das próteses completas

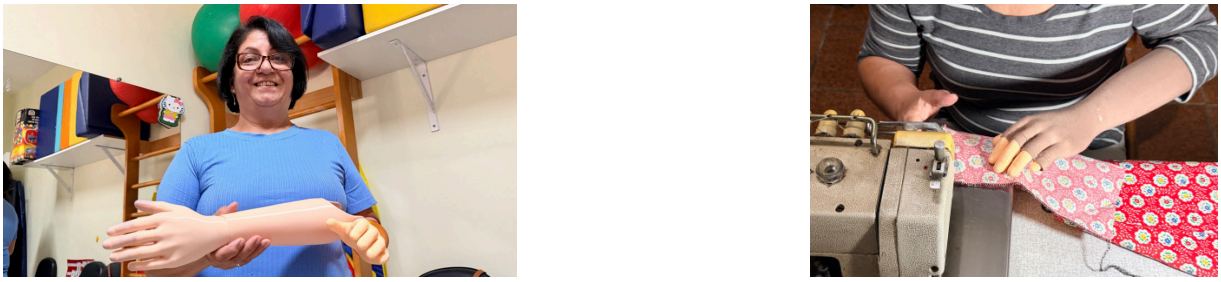
Na primeira prótese, o soquete apresentou massa de 258 g e tempo de impressão de 36 h, enquanto o conjunto terminal apresentou massa de 137 g e tempo de impressão de 19 h 30 min. A massa total calculada foi de 395 g, com tempo acumulado de impressão de 55 h 30 min.

Na segunda prótese, o soquete apresentou massa de 240 g e tempo de impressão de 34 h, enquanto o conjunto terminal apresentou massa de 248 g e tempo de impressão de 33 h. A massa total calculada foi de 488 g, com tempo acumulado de impressão de 67 h.

A maior massa do conjunto terminal da segunda prótese coincidiu com a ampliação da conexão roscada. Como a massa da rosca não foi medida separadamente, não é possível atribuir quantitativamente a diferença observada apenas a essa modificação. Como não foram registrados o consumo total de material e o custo direto de fabricação, não devem ser apresentadas afirmações quantitativas sobre economia. Também não foi realizada comparação objetiva de massa com a prótese mecânica anteriormente fornecida pelo SUS; qualquer referência à maior leveza deve ser apresentada como percepção da participante.

E. Entrega e percepção inicial de uso

A primeira visita de entrega ocorreu com a disponibilização da prótese dotada da conexão roscada menor. Após o uso inicial, a participante relatou instabilidade e escape da conexão, embora tenha aprovado a possibilidade de reposicionar angularmente o terminal. Esse retorno motivou o aumento das dimensões e do comprimento da rosca.



(a) (b)

Figura 3. Dispositivos finais e uso inicial: (a) as duas próteses completas, com a mesma geometria de soquete e conexões roscadas distintas; (b) utilização do terminal como apoio passivo na condução do tecido durante a costura. Fotografias: Claudio Vieira/Prefeitura de São José dos Campos, reproduzidas com autorização.

Na segunda visita de entrega, foi disponibilizada a prótese com a conexão modificada. Essa configuração apresentou maior estabilidade e foi confirmada como versão definitiva. A versão definitiva permaneceu com a participante desde o final de maio de 2026.

Em relato público posterior à entrega, a participante descreveu a nova prótese como mais leve que a prótese mecânica anteriormente fornecida pelo SUS e informou utilizá-la em atividades cotidianas e durante a costura [12]. Durante essa atividade, o reposicionamento do terminal permitiu utilizar a mão como apoio passivo para sustentar, orientar e conduzir o tecido. A participante também expressou satisfação com a aparência e indicou intenção de continuar utilizando o dispositivo.

A Fig. 3 apresenta as duas próteses completas e o uso do terminal como apoio passivo durante a costura.

Esses relatos são compatíveis com os requisitos identificados no início do desenvolvimento e constituem evidência preliminar de aceitabilidade e utilidade percebida. Entretanto, foram obtidos por observação não padronizada e por entrevista jornalística, sem aplicação de escalas validadas de conforto, satisfação ou desempenho ocupacional. Não devem, portanto, ser interpretados como demonstração de eficácia funcional.

F. Limitações

O trabalho apresenta as limitações inerentes a um estudo de caso único. Não foram realizados ensaios mecânicos, avaliação controlada de desempenho, comparação com métodos convencionais ou aplicação de instrumentos padronizados de conforto, satisfação e função. As provas presenciais do soquete tiveram duração curta, e o acompanhamento após a entrega ainda não permite avaliar durabilidade, integridade dos materiais ou uso continuado em longo prazo [8], [9].

O número de versões do soquete é aproximado, pois não houve registro sistemático de cada impressão e de todas as alterações geométricas. Os relatos de uso cotidiano e costura são subjetivos e não foram acompanhados por medição ergonômica ou análise quantitativa da atividade. Assim, os resultados demonstram a aplicação inicial do fluxo, o refinamento iterativo do encaixe e a aceitabilidade preliminar das próteses, mas não validam um protocolo clínico ou produtivo generalizável.

IV. Conclusão

O fluxo de manufatura digital possibilitou desenvolver aproximadamente dez versões de teste do soquete e fabricar duas próteses transradiais passivas completas para uma participante atendida pelo SUS. A prototipagem iterativa permitiu corrigir a incompatibilidade dimensional inicial, aprimorar o contato sobre os epicôndilos e obter retenção supracondilar sem cintas adicionais. A personalização do terminal baseado no modelo Kwawu e a ampliação da conexão roscada na segunda prótese responderam, respectivamente, às preferências estéticas da participante e à necessidade de maior estabilidade durante o reposicionamento da mão.

Após a entrega, a participante relatou perceber a nova prótese como mais leve que a prótese mecânica anteriormente fornecida pelo SUS e informou utilizar o terminal como apoio passivo durante a costura. Esses achados indicam aceitabilidade inicial e compatibilidade do dispositivo com a demanda apresentada, mas não demonstram eficácia funcional. A generalização dos resultados é limitada pelo estudo de caso único, pelo curto período de acompanhamento e pela ausência de ensaios mecânicos, instrumentos padronizados e avaliação quantitativa do desempenho ocupacional.

Agradecimentos

Os autores agradecem à Unidade de Reabilitação da Prefeitura de São José dos Campos, ao Núcleo de Saúde Digital da Universidade Federal de São Paulo, ao PET-Saúde Digital e ao Laboratório de Órteses e Próteses 3D pelo apoio técnico e institucional.

Referências

[1] H. Jones, S. Dupan, M. Dyson, A. Krasoulis, L. P. J. Kenney, M. Donovan-Hall, K. Memarzadeh, S. Day, M. Coutinho, and K. Nazarpour, "Co-creation and user perspectives for upper limb prosthetics," *Frontiers in Neurorobotics*, vol. 15, Art. no. 689717, 2021, doi: 10.3389/fnbot.2021.689717.

[2] F. Kristjansdottir, L. B. Dahlin, H.-E. Rosberg, and I. K. Carlsson, "Social participation in persons with upper limb amputation receiving an esthetic prosthesis," *Journal of Hand Therapy*, vol. 33, no. 4, pp. 520–527, 2020, doi: 10.1016/j.jht.2019.03.010.

[3] R. Brack and E. H. Amalu, "A review of technology, materials and R&D challenges of upper limb prosthesis for improved user suitability," *Journal of Orthopaedics*, vol. 23, pp. 88–96, 2021, doi: 10.1016/j.jor.2020.12.009.

[4] B. Maat, G. Smit, D. Plettenburg, and P. Breedveld, “Passive prosthetic hands and tools: A literature review,” *Prosthetics and Orthotics International*, vol. 42, no. 1, pp. 66–74, 2018, doi: 10.1177/0309364617691622.

[5] L. C. Smail, C. Neal, C. Wilkins, and T. L. Packham, “Comfort and function remain key factors in upper limb prosthetic abandonment: Findings of a scoping review,” *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 8, pp. 821–830, 2021, doi: 10.1080/17483107.2020.1738567.

[6] J. Barrios-Muriel, F. Romero-Sánchez, F. J. Alonso-Sánchez, and D. Rodríguez Salgado, “Advances in orthotic and prosthetic manufacturing: A technology review,” *Materials*, vol. 13, no. 2, Art. no. 295, 2020, doi: 10.3390/ma13020295.

[7] S. Vennam, V. K. N., and F. Pati, “3D printed personalized assistive devices: A material, technique, and medical condition perspective,” *Applied Materials Today*, vol. 40, Art. no. 102403, 2024, doi: 10.1016/j.apmt.2024.102403.

[8] K. Wendo, O. Barbier, X. Bollen, T. Schubert, T. Lejeune, B. Raucent, and R. Olszewski, “Open-source 3D printing in the prosthetic field—The case of upper limb prostheses: A review,” *Machines*, vol. 10, no. 6, Art. no. 413, 2022, doi: 10.3390/machines10060413.

[9] K. A. M. Rasmussen, J. K. Schwartz, K. Aylmer, C. Keahi, A. Said, D. H. Ballard, and W. E. Janes, “A systematic review of 3D-printed assistive technology rehabilitation interventions,” *Assistive Technology*, pp. 1–12, 2026, doi: 10.1080/10400435.2026.2619686.

[10] E. K. Rorato, “Towards a digital workflow for upper limb passive prostheses development: The role of biomodels, design thinking and open-source designs,” dissertação de mestrado, Programa de Pós-Graduação Profissional em Inovação Tecnológica, Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, SP, 2024.

[11] E. K. Rorato, P. P. do Nascimento Silva, and M. E. Kunkel, “Kwawu passiva estético-funcional: Remix de uma prótese open-source de membro superior transradial manufaturada por impressão 3D,” *HFD*, vol. 12, no. 24, pp. 149–155, 2023, doi: 10.5965/2316796312242023149.

[12] L. Brito, “Prótese em 3D devolve autonomia e esperança à costureira,” *Prefeitura de São José dos Campos*, 17 jul. 2026. [Online]. Disponível em: https://www.sjc.sp.gov.br/noticias/2026/julho/16/protese-em-3d-devolve-autonomia-e-esperanca-a-costureira/. Acesso em: 30 jul. 2026.

Desenvolvimento de Dispositivos para Acessibilidade Digital

Samuel Cardoso Gregório
Bacharelado em Engenharia Mecatrônica - PET Conexões
IF SUDESTE MG
Juiz de Fora, Brazil
samuelcgregorio@gmail.com

Giovani Luiz da Silva Filho
Bacharelado em Engenharia Mecatrônica - PET Conexões
IF SUDESTE MG
Juiz de Fora, Brazil
giovaniluiz12@gmail.com

Rafael Pelegrini da Silva
Bacharelado em Engenharia Mecatrônica - PET Conexões
IF SUDESTE MG
Juiz de Fora, Brazil
rafaelpelegrini00@gmail.com

Ingrid Gonçalves da Silveira Machado
Bacharelado em Engenharia Mecatrônica - PET Conexões
IF SUDESTE MG
Juiz de Fora, Brazil
ingridmachadogs@yahoo.com.br

Pedro Henrique Gouvêa Pires
Bacharelado em Engenharia Mecatrônica - PET Conexões
IF SUDESTE MG
Juiz de Fora, Brazil
phgp1111@gmail.com

Silvana Terezinha Faceroli
Department of Education and Technology
IF SUDESTE MG
Juiz de Fora, Brazil
silvana.faceroli@ifsudestemg.edu.br

Sara Del Vecchio
Department of Education and Technology
IF SUDESTE MG
Juiz de Fora, Brazil
sara.vecchio@ifsudestemg.edu.br

Tales Pulinho Ramos
Department of Education and Technology
IF SUDESTE MG
Juiz de Fora, Brazil
tales.ramos@ifsudestemg.edu.br

Thiago Rodrigues Oliveira
Department of Education and Technology - PET Conexões
IF SUDESTE MG
Juiz de Fora, Brazil
thiago.oliveira@ifsudestemg.edu.br

***Abstract*—A acessibilidade digital é um fator fundamental para a inclusão social de pessoas com deficiência, especialmente aquelas com limitações motoras severas, que frequentemente encontram barreiras no uso de dispositivos convencionais de interação computacional. Neste contexto, este trabalho apresenta o desenvolvimento de duas interfaces homem-máquina de baixo custo voltadas à ampliação do acesso a recursos digitais por meio da utilização de múltiplos sensores. O primeiro protótipo utiliza sensores inerciais e eletromiografia de superfície (EMG) para realizar o controle do cursor e a execução de comandos de clique, respectivamente. Os resultados obtidos demonstraram a viabilidade da integração dos sensores e da estratégia de fusão de sinais adotada, permitindo a conversão de movimentos residuais e contrações musculares em comandos digitais. Como complemento, foi desenvolvido um software de treinamento em formato de jogo sério utilizando a biblioteca Pygame, destinado à calibração do sistema e à adaptação do usuário ao dispositivo. O segundo protótipo, baseado em acionamento capacitivo e estrutura modelada em três dimensões, encontra-se em fase de desenvolvimento, apresentando resultados preliminares promissores em sua modelagem e implementação. Os resultados indicam que a combinação de diferentes tecnologias de sensoriamento constitui uma alternativa viável para o desenvolvimento de dispositivos assistivos acessíveis, contribuindo para a ampliação da autonomia, da inclusão digital e da participação social de pessoas com deficiência motora.**

***Index Terms*—tecnologia assistiva, fusão de sinais, mobilidade reduzida**

I. Introdução

Em uma sociedade cada vez mais conectada e dependente de recursos tecnológicos, a inclusão digital tornou-se um fator essencial para a participação plena dos indivíduos em atividades educacionais, profissionais e sociais. Apesar dos avanços tecnológicos observados nas últimas décadas, ainda persistem desafios significativos relacionados à acessibilidade digital, especialmente para pessoas com limitações motoras. Barreiras tecnológicas, econômicas e de usabilidade dificultam o acesso a dispositivos computacionais convencionais, comprometendo a autonomia e a qualidade de vida desses usuários [1].

Os dispositivos de entrada tradicionalmente utilizados para interação com computadores, como teclados e mouses convencionais, foram desenvolvidos considerando predominantemente usuários sem restrições motoras. Consequentemente, muitas pessoas com comprometimentos motores severos encontram dificuldades para realizar tarefas simples em ambientes digitais, uma vez que os equipamentos disponíveis comercialmente nem sempre oferecem a sensibilidade, a personalização ou a adaptabilidade necessárias para converter movimentos de baixa amplitude e intensidade em comandos digitais precisos e confiáveis [2].

As limitações de acesso às tecnologias digitais impactam diretamente a inclusão social, educacional e profissional desses indivíduos. A dificuldade de interação com sistemas computacionais reduz oportunidades de comunicação, aprendizado, trabalho e entretenimento, além de restringir a independência funcional e a participação ativa na sociedade [3]. Nesse contexto, o desenvolvimento de tecnologias assistivas capazes de ampliar a acessibilidade digital representa uma importante estratégia para promover autonomia e inclusão.

Diante desse cenário, este trabalho propõe o desenvolvimento de interfaces homem-máquina baseadas na integração de múltiplas tecnologias de aquisição de sinais, incluindo eletromiografia de superfície (EMG), sensores capacitivos, chaves de fim de curso e sensores inerciais. A utilização combinada desses dispositivos busca explorar diferentes formas de interação, permitindo que usuários com distintos níveis de comprometimento motor possam acessar recursos computacionais por meio de movimentos residuais ou sinais musculares voluntários. Além disso, a abordagem proposta prioriza soluções de baixa complexidade, baixo custo e elevada capacidade de adaptação às necessidades individuais dos usuários [4], [5].

Para validar essa proposta, foram desenvolvidos dois protótipos de mouse acessível destinados a pessoas com diferentes graus de limitação motora. Ambos os sistemas têm como objetivo substituir ou complementar dispositivos convencionais de entrada, transformando sinais captados pelos sensores em comandos digitais capazes de controlar aplicações computacionais. O primeiro protótipo utiliza predominantemente sinais eletromiográficos para o acionamento dos comandos, enquanto o segundo incorpora sensores inerciais e outros mecanismos de interação, ampliando as possibilidades de uso e adaptação às características funcionais de cada usuário.

Como complemento ao desenvolvimento do hardware, foi implementado um software de treinamento e validação em formato de jogo sério, permitindo que os usuários realizem o processo de aprendizado de forma lúdica e progressiva. Essa ferramenta possibilita a calibração dos parâmetros do sistema, o ajuste dos níveis de sensibilidade e a avaliação do desempenho dos métodos de acionamento, contribuindo para uma experiência de uso mais intuitiva e personalizada.

Dessa forma, o objetivo deste trabalho é desenvolver e avaliar estratégias de interação digital baseadas na combinação de múltiplos sensores,como eletromiográficos, capacitivos, mecânicos e inerciais, para a construção de interfaces homem-máquina acessíveis. Busca-se, por meio da implementação e comparação de dois protótipos, identificar soluções de baixo custo, alta adaptabilidade e elevado potencial de aplicação como tecnologia assistiva, contribuindo para a ampliação da acessibilidade digital e da autonomia de pessoas com severas limitações motoras.

II. Materiais e Métodos

Este trabalho caracteriza-se como uma pesquisa aplicada, com enfoque no desenvolvimento de tecnologias assistivas voltadas à promoção da acessibilidade digital para pessoas com limitações motoras. A proposta consiste na concepção e implementação de interfaces homem-máquina capazes de converter diferentes sinais fisiológicos e mecânicos em comandos digitais, utilizando componentes de baixo custo e ampla disponibilidade comercial.

Os principais materiais empregados no desenvolvimento dos protótipos foram: Sensor inercial MPU-6050, composto por acelerômetro e giroscópio de três eixos; Shield de eletromiografia (EMG) Olimex para aquisição de sinais musculares de superfície; Microcontrolador ESP32-S2 Mini para processamento dos sinais e comunicação com o computador.

A metodologia adotada foi estruturada em quatro etapas principais, descritas a seguir.

A. Revisão bibliográfica

Inicialmente, realizou-se uma revisão da literatura com o objetivo de identificar o estado da arte relacionado ao desenvolvimento de dispositivos de tecnologia assistiva e interfaces homem-máquina para pessoas com limitações motoras. Foram analisados trabalhos envolvendo diferentes formas de interação baseadas em sinais musculares, sensores inerciais e outros métodos alternativos de acionamento, buscando compreender as soluções existentes, suas limitações e oportunidades de melhoria.

B. Estudo e seleção dos sensores

Na segunda etapa, foi conduzida uma análise comparativa entre diferentes tecnologias de sensores com potencial aplicação em dispositivos assistivos. Foram avaliados sensores eletromiográficos, capacitivos, indutivos, inerciais e chaves mecânicas de fim de curso, considerando critérios como sensibilidade, robustez, facilidade de integração, custo e adequação aos diferentes níveis de comprometimento motor dos usuários. A partir dessa análise, foram selecionados os sensores mais adequados para a implementação dos protótipos propostos.

C. Desenvolvimento e integração dos protótipos

A terceira etapa compreendeu o desenvolvimento dos dispositivos e a integração entre hardware e software. O primeiro protótipo foi implementado utilizando o microcontrolador ESP32-S2 Mini e os sensores selecionados, empregando a linguagem CircuitPython para aquisição, processamento e interpretação dos sinais captados. A estratégia adotada baseia-se na fusão de sinais provenientes de diferentes sensores, permitindo o reconhecimento de comandos destinados ao controle

do cursor e à execução de ações equivalentes às de um mouse convencional. Os testes funcionais preliminares do Protótipo 1 foram realizados em ambiente de laboratório controlado. A aquisição do sinal EMG via Shield Olimex operou sob taxa de amostragem de 1 kHz, com eletrodos de superfície dispostos no músculo bíceps braquial e limiar de detecção (threshold) ajustado via software após normalização da amplitude do sinal. A movimentação inercial pelo MPU6050 foi aferida registrando-se a resposta de deslocamento do cursor em pixels/segundo em relação aos ângulos de inclinação (pitch e roll). As rotinas de teste integradas ao jogo Pygame registraram a estabilidade da comunicação I2C/CircuitPython e o tempo de resposta entre o evento muscular/inercial e a resposta visual na tela.

O segundo protótipo encontra-se em fase de desenvolvimento e modelagem mecânica. Nesta etapa, está sendo utilizado o software CAD SolidWorks para o projeto estrutural e a definição da arquitetura do dispositivo, visando à futura implementação física e validação experimental.

D. Desenvolvimento do software de treinamento e validação

Como complemento ao desenvolvimento do hardware, foi implementado um software de treinamento utilizando a biblioteca Pygame. O sistema foi concebido na forma de um jogo sério, permitindo a calibração dos sensores, a adaptação dos parâmetros de sensibilidade e o treinamento progressivo dos usuários. Além de auxiliar no processo de aprendizagem da utilização do dispositivo assistivo, o software possibilita a realização de testes funcionais e a demonstração técnica dos protótipos desenvolvidos.

III. Protótipo 1

O desenvolvimento do primeiro protótipo foi dividido em quatro etapas principais: implementação do controle de movimentação do cursor por meio do sensor MPU6050, desenvolvimento do sistema de clique utilizando sinais eletromiográficos (EMG), criação de um software para calibração e treinamento do usuário e integração dos subsistemas. O segundo protótipo foi estruturado em três etapas: modelagem mecânica, desenvolvimento da programação embarcada e integração dos sistemas.

A. Sensor MPU6050

O sensor MPU6050 foi selecionado por integrar um acelerômetro e um giroscópio de três eixos em um único dispositivo, apresentando baixo custo, ampla disponibilidade comercial e facilidade de integração com sistemas embarcados. Para este trabalho, utilizou-se exclusivamente a funcionalidade de acelerometria para o controle do deslocamento do cursor do mouse nos eixos X e Y.

A comunicação entre o sensor e o microcontrolador foi realizada por meio do protocolo I2C, reduzindo a quantidade de conexões necessárias e simplificando a montagem do circuito. A estratégia de aquisição e processamento dos dados foi implementada utilizando a linguagem CircuitPython embarcada no microcontrolador.

Por meio da medição das acelerações nos três eixos espaciais, foi possível mapear os movimentos realizados pelo usuário e convertê-los em deslocamentos correspondentes do cursor na interface gráfica. Os testes realizados demonstraram que o sensor apresentou estabilidade e sensibilidade adequadas para a aplicação proposta, permitindo um controle contínuo e intuitivo do cursor.

B. Sensor EMG

O sensor de eletromiografia foi empregado para a execução dos comandos de clique do mouse. Para sua utilização, foi integrado ao Shield EMG Olimex, responsável pelo condicionamento e amplificação dos sinais musculares adquiridos. A combinação entre os dois dispositivos mostrou-se adequada para a aplicação proposta devido à sua simplicidade de utilização, acessibilidade financeira e facilidade de integração com sistemas embarcados.

A aquisição dos sinais foi realizada utilizando um microcontrolador Arduino Uno conectado ao Shield Olimex, sendo a leitura efetuada por meio da entrada analógica A0. Para garantir o funcionamento adequado do sistema, foi necessário posicionar corretamente os eletrodos sobre o músculo monitorado. O eletrodo de referência foi fixado em uma região sem atividade muscular significativa, enquanto os eletrodos de medição foram posicionados ao longo do ventre muscular.

A Figura 1 apresenta um exemplo do sinal eletromiográfico obtido durante os testes realizados no músculo bíceps braquial. Observa-se a presença de um pico associado à contração muscular voluntária, utilizado como evento de acionamento do sistema.

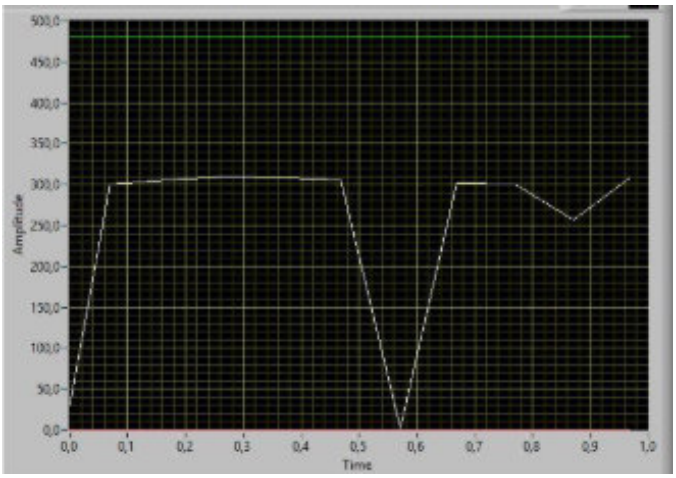


Fig. 1. Leitura do sinal eletromiográfico durante uma contração muscular.

Para a identificação automática dos comandos de clique, foi definido um limiar de detecção. Quando a amplitude do sinal ultrapassa esse valor pré-estabelecido, o sistema interpreta a ocorrência como um comando válido de clique. O limiar pode ser ajustado tanto por meio de um potenciômetro quanto por software, permitindo a adaptação do dispositivo às características musculares e limitações específicas de cada usuário.

Os resultados obtidos demonstraram que a estratégia adotada foi capaz de detectar de forma consistente as contrações musculares voluntárias, possibilitando sua conversão em comandos digitais de maneira simples e eficiente.

C. Software de treinamento

Como complemento ao desenvolvimento do hardware, foi implementado um software de treinamento utilizando a biblioteca Pygame. O sistema foi concebido na forma de um jogo sério inspirado no gênero shoot 'em up, no qual o jogador controla uma nave espacial com o objetivo de eliminar inimigos e evitar colisões com projéteis adversários.

A principal finalidade da ferramenta é auxiliar o usuário durante o processo inicial de adaptação ao dispositivo, além de permitir a calibração dos parâmetros de sensibilidade dos sensores utilizados. A abordagem lúdica busca aumentar o engajamento do usuário durante o treinamento, tornando o processo de aprendizagem mais intuitivo e motivador.

O jogo foi projetado para ser controlado exclusivamente pelo protótipo desenvolvido. Os movimentos da nave são realizados a partir dos deslocamentos captados pelo sensor MPU6050, que controlam a posição do cursor nos eixos horizontal e vertical. Já os comandos de clique, obtidos por meio do sensor EMG, ativam um escudo temporário que torna a nave invulnerável por um curto intervalo de tempo.

A dinâmica do jogo segue o modelo de progressão contínua, em que a dificuldade aumenta gradativamente com o avanço das fases, por meio do aumento da velocidade e da quantidade de inimigos presentes na tela. Além disso, níveis específicos incluem o enfrentamento de chefes (bosses), exigindo maior precisão e controle por parte do usuário.

A Figura 2 apresenta a interface gráfica desenvolvida para treinamento e validação do sistema. Os testes realizados demonstraram o correto funcionamento da comunicação entre o protótipo e o software, permitindo a utilização do jogo como ferramenta de treinamento, calibração e demonstração tecnológica.

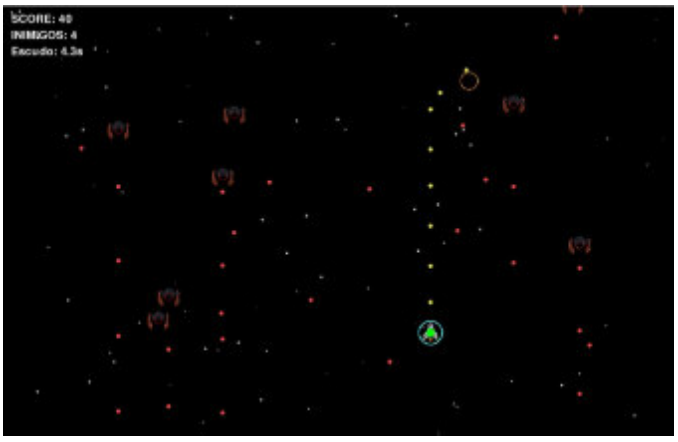


Fig. 2. Software de treinamento desenvolvido em Pygame.

D. Integração dos sistemas

Após a validação individual dos sensores, realizou-se a integração entre o MPU6050 e o sistema de aquisição de sinais EMG em um único protótipo funcional, conforme ilustrado na Figura 3.

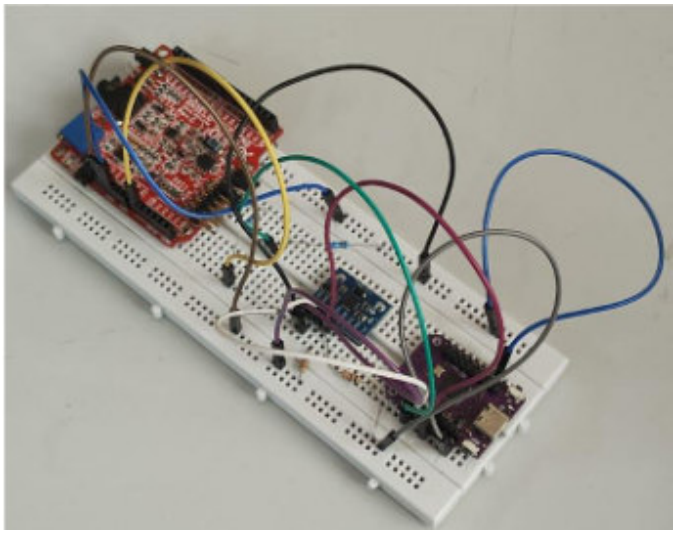


Fig. 3. Protótipo integrado desenvolvido.

Nessa configuração, o sensor MPU6050 foi responsável pelo controle do deslocamento do cursor, enquanto o sensor EMG foi utilizado para a geração dos comandos de clique. A integração dos sistemas permitiu validar o funcionamento conjunto dos componentes, demonstrando a viabilidade da arquitetura proposta para aplicações de acessibilidade digital.

Os testes preliminares indicaram que a comunicação entre os módulos ocorreu de forma estável, possibilitando a conversão dos movimentos e contrações musculares em comandos computacionais. Os resultados obtidos evidenciam o potencial da combinação de sensores inerciais e eletromiográficos para o desenvolvimento de interfaces homem-máquina acessíveis, de baixo custo e adaptáveis a diferentes níveis de comprometimento motor.

A abordagem multimodal proposta alinha-se às tendências de tecnologias assistivas de baixo custo [6]. Dispositivos baseados exclusivamente em EMG oferecem precisão para acionamentos binários, porém podem induzir à fadiga muscular quando utilizados por períodos prolongados [7]. Ao combinar o EMG apenas para a função discreta de clique e o MPU6050 para o deslocamento contínuo do cursor, a presente solução minimiza a carga muscular exigida. Por outro lado, soluções que empregam unicamente rastreamento inercial ou ocular [8] enfrentam desafios operacionais em usuários com tremores involuntários. A integração da filtragem digital e o ajuste de limiar flexível no ESP32-S2 reduzem falsos positivos causados por ruídos de movimento.

IV. Protótipo 2

Com o objetivo de ampliar as possibilidades de interação para usuários com diferentes níveis de comprometimento mo-

tor, foi desenvolvido um segundo protótipo baseado em acionamento capacitivo. Diferentemente do primeiro dispositivo, que utiliza sinais eletromiográficos para a execução dos comandos de clique, esta proposta busca explorar uma interface de toque capaz de detectar contatos de baixa intensidade, permitindo sua utilização por indivíduos que preservam movimentos residuais limitados. Além de reduzir a complexidade do sistema, essa abordagem contribui para o desenvolvimento de uma solução de baixo custo, fácil implementação e elevada adaptabilidade, ampliando o conjunto de tecnologias assistivas disponíveis para promoção da acessibilidade digital.

A. Modelagem

A estrutura mecânica do segundo protótipo foi desenvolvida por meio de modelagem tridimensional utilizando software CAD. O processo de desenvolvimento ocorreu de forma iterativa, permitindo sucessivos refinamentos no projeto para otimizar a disposição interna dos componentes eletrônicos, a ergonomia e a usabilidade do dispositivo.

O modelo final, apresentado na Figura 4, consiste em uma interface vestível no formato de anel, projetada para acomodar os elementos eletrônicos necessários ao acionamento por toque capacitivo. Além de garantir a proteção física dos componentes, a estrutura foi concebida para proporcionar conforto durante o uso prolongado e facilitar a interação de usuários com limitações motoras.

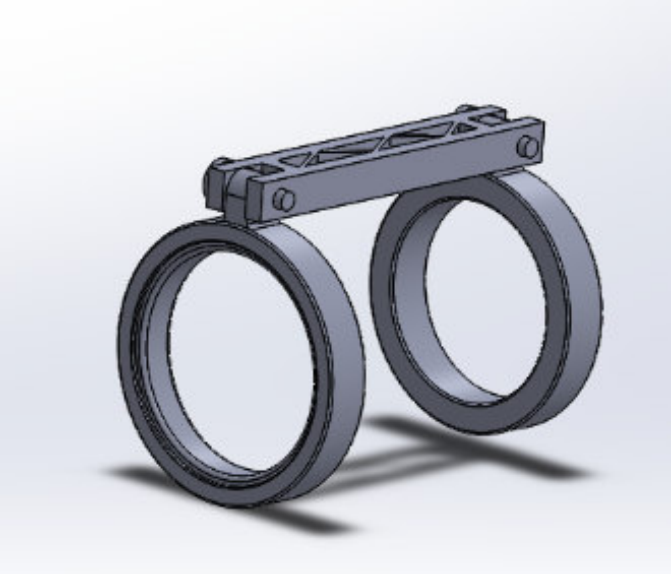


Fig. 4. Modelo tridimensional do protótipo baseado em sensor capacitivo.

B. Programação

O sistema de acionamento foi desenvolvido utilizando o microcontrolador ESP32-S2 Mini, explorando sua capacidade nativa de leitura capacitiva. Essa característica permitiu implementar uma interface de toque sem a necessidade de circuitos adicionais de condicionamento de sinal, reduzindo a complexidade e o custo do dispositivo.

A programação foi realizada em CircuitPython, utilizando a biblioteca Adafruit HID para a emulação de dispositivos de entrada do tipo mouse e teclado. O limiar de detecção capacitiva foi definido experimentalmente por meio de testes sucessivos, buscando maximizar a sensibilidade ao toque e, simultaneamente, minimizar a ocorrência de acionamentos indevidos provocados por ruídos elétricos ou interferências ambientais.

Os resultados obtidos demonstraram a capacidade do sistema em detectar toques intencionais de forma consistente, convertendo-os em comandos digitais utilizados para interação com o computador. Dessa forma, o protótipo apresenta potencial para atender usuários que possuem mobilidade residual limitada, mas que ainda conseguem realizar pequenos movimentos ou contatos físicos controlados.

C. Integração dos Sistemas

A etapa de integração envolveu a montagem do circuito eletrônico em protoboard para validação funcional dos componentes e dos algoritmos desenvolvidos. Foram utilizados cabos de extensão para posicionar a superfície sensível ao toque de forma independente da placa principal, proporcionando maior flexibilidade durante os testes e permitindo futuras adaptações ergonômicas do dispositivo. A Figura 5 apresenta a configuração eletrônica utilizada durante a fase de desenvolvimento.

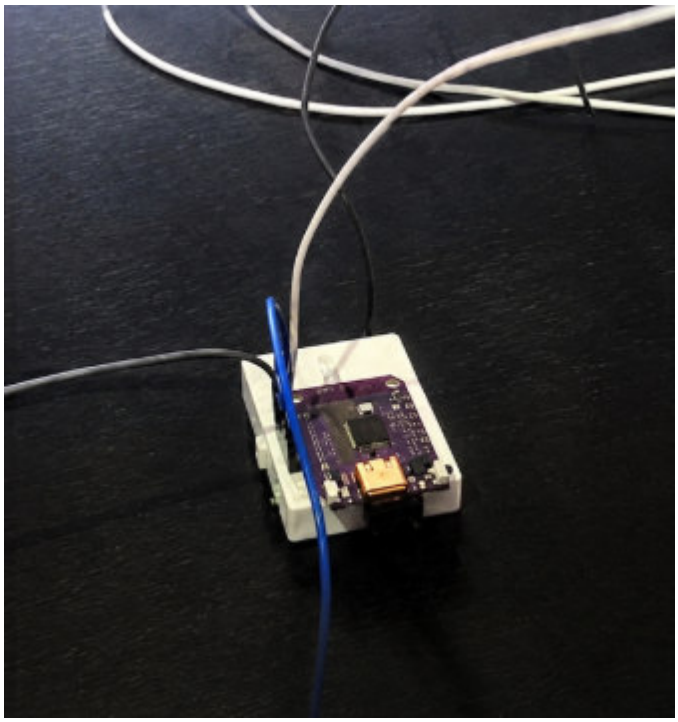


Fig. 5. Protótipo eletrônico baseado em acionamento capacitivo.

O ESP32-S2 Mini foi adotado como elemento central do sistema devido ao seu reduzido tamanho físico, baixo consumo de energia e conjunto de funcionalidades compatíveis com os

requisitos do projeto. Entre seus principais recursos destacam-se o suporte nativo ao padrão USB HID (Human Interface Device), permitindo que o dispositivo seja reconhecido diretamente pelo sistema operacional como um periférico de entrada, sem necessidade de instalação de drivers adicionais, e a comunicação via protocolo I²C para integração com sensores externos.

A arquitetura desenvolvida possibilita a incorporação de diferentes mecanismos de acionamento em uma única plataforma embarcada, favorecendo a criação de dispositivos assistivos compactos, de baixo custo e facilmente replicáveis. Atualmente, o protótipo encontra-se em fase de finalização, com a fabricação da estrutura física em andamento e a realização dos testes de validação funcional prevista para as próximas etapas do projeto.

Os resultados preliminares indicam a viabilidade do acionamento capacitivo. A Tabela I sintetiza a análise comparativa qualitativa entre a arquitetura do Protótipo 1 e a do Protótipo 2 com base nos testes de bancada e nas especificações de projeto.

TABLE I
COMPARAÇÃO QUALITATIVA ENTRE OS PROTÓTIPOS DESENVOLVIDOS.

Parâmetro	**Protótipo 1 (MPU+EMG)**	**Protótipo 2 (Capacitivo)**
Mecanismo de Clique	Contração muscular voluntária (EMG)	Toque por proximidade (Capacitivo)
Controle do Cursor	Inercial (Acelerômetro MPU6050)	Emulação direta / Comutação
Esforço Motor	Baixo a Médio (requer tônus muscular)	Mínimo (apenas contato residual)
Estágio Atual	Validado em bancada e integrado ao jogo	Validado em modelagem 3D e bancada eletrônica
Complexidade Eletrônica	Média (filtragem e condicionamento EMG)	Baixa (leitura nativa ESP32-S2)

V. CONCLUSÃO

Este trabalho apresentou o desenvolvimento e a avaliação de duas abordagens de Interface Homem-Máquina voltadas à acessibilidade digital de pessoas com limitações motoras. A utilização de componentes de baixo custo e código aberto viabilizou a construção de arquiteturas modulares e adaptáveis.

O Protótipo 1 (MPU6050 + EMG) teve sua viabilidade técnica comprovada por meio de testes funcionais de bancada e integração com um jogo sério em Pygame, demonstrando estabilidade na conversão de movimentos inerciais e eventos musculares em comandos digitais de navegação e clique. O Protótipo 2 (Capacitivo) teve sua viabilidade eletrônica e conceitual demonstrada nas etapas de modelagem 3D, programação embarcada e testes de sensibilidade ao toque, posicionando-se como uma alternativa promissora para usuários com mobilidade residual severamente reduzida.

Como trabalhos futuros, pretende-se concluir a fabricação da estrutura física do Protótipo 2, realizar testes de usabilidade com usuários reais e quantificar métricas exatas de desempenho (tempo de resposta, taxa de acerto de comandos e curva de aprendizado) para ambas as soluções.

AGRADECIMENTOS

Os autores agradecem ao Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais — Campus Juiz de Fora, e ao Programa de Educação Tutorial (PET), por meio da Pró-Reitoria de Ensino e do Ministério da Educação (MEC), cujas bolsas e recursos foram essenciais para apoiar as atividades de pesquisa e extensão aqui apresentadas.

REFERENCES

[1] J. Ariza and J. M. Pearce, "Low-cost assistive technologies for disabled people using open-source hardware and software: A systematic literature review," *IEEE Access*, vol. 10, pp. 121392–121409, 2022, doi: 10.1109/ACCESS.2022.3221449.

[2] A. C. L. Souza *et al.*, "O uso de dispositivos assistivos como intervenção na terapia ocupacional para pessoas com Parkinson," *Observatorio de la Economía Latinoamericana*, vol. 22, no. 7, p. e5824, 2024.

[3] C. J. G. de Almeida, "Inclusão: Processo educativo e socialização de uma aluna com atrofia muscular espinhal numa escola em Presidente Kennedy/ES," Dissertação de Mestrado, Centro Universitário Vale do Cricaré, São Mateus, ES, Brasil, 2021.

[4] D. Welihinda *et al.*, "EEG and EMG-based human-machine interface for navigation of mobility-related assistive wheelchair (MRA-W)," *Heliyon*, vol. 10, 2024, Art. no. e27777, doi: 10.1016/j.heliyon.2024.e27777.

[5] M. Yang, L. Manji, and J. Jutai, "Mobility, disability, and self-identity," in *The Palgrave Encyclopedia of Disability*, G. Bennett and E. Goodall, Eds. Cham, Switzerland: Palgrave Macmillan, 2025, doi: 10.1007/978-3-031-40858-8_334-1.

[6] H.-T. Lee, M. Shim, X. Liu, H.-R. Cheon, S.-G. Kim, C. H. Han e H.-J. Hwang, "A review of hybrid EEG-based multimodal human–computer interfaces using deep learning: applications, advances, and challenges," *Biomedical Engineering Letters*, vol. 15, no. 4, pp. 587–618, mar. 2025. doi: 10.1007/s13534-025-00469-5.

[7] K. Coleman, "Electromyography based human-computer-interface to induce movement in elderly persons with movement impairments," em *Proceedings of the 2001 ACM Conference on Assistive Technologies*, maio 2001, pp. 75–79. doi: 10.1145/564526.564547.

[8] I. Severin e D. Dobrea, Using Inertial Sensors to Determine Head Motion–A Review, *Journal of Imaging*, vol. 7, no. 12, p. 265, dez. 2021. doi: 10.3390/jimaging7120265.

Effects of Repetitive Transcranial Magnetic Stimulation Associated with Motor Imagery on the Electromyographic Activation of the Tibialis Anterior and Rectus Femoris Muscles in Post-Stroke Individuals

Fernanda Vaz de Souza
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
ORCID: 0009-0005-0201-9260
fernanda.souza.37@edu.ufes.br

Sheila da Luz Schreider
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
ORCID: 0000-0003-4603-3492
sheiladaluz@gmail.com

Fernando Bermudes Cabral
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
ORCID: 0000-0003-0577-3902
fernando.cabral@edu.ufes.br

Matheus Modolo Ferreira
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
ORCID:0009-0001-2258-1761
matheusmofe@gmail.com

Luís Carlos Rodríguez-Timaná
Postgraduate Program in Elect. Eng.
Federal University of Espírito Santo
Vitória, Brazil
ORCID: 0000-0003-2437-1490
luis.tinama@edu.ufes.br

Teodiano Freire Bastos Filho
Postgraduate Program in Elect. Eng.
Federal University of Espírito Santo
Vitória, Brazil
ORCID: 0000-0002-1185-2773
teodiano.bastos@edu.br

***Abstract*—Stroke frequently results in persistent motor deficits that impair gait and functional independence. Neuromodulation strategies, such as repetitive transcranial magnetic stimulation (rTMS), combined with motor imagery (MI) and brain-computer interfaces (BCIs), have been investigated as approaches to enhance motor recovery. The aim of this study was to evaluate the effects of rTMS combined with MI on the peak electromyographic contraction of the tibialis anterior and rectus femoris muscles in post-stroke individuals. This pilot randomized, controlled, assessor-blinded clinical trial included 10 participants allocated to the Intervention Group (rTMS + BCI-mediated MI; $n = 5$) and the Control Group (BCI-mediated MI; $n = 5$). Electromyographic activity was assessed before and after two weeks of intervention using surface electromyography (sEMG). The Intervention Group showed increased mean peak electromyographic activation of both muscles, with the largest descriptive change observed in the tibialis anterior, whereas the Control Group showed reductions in mean activation. These findings indicate a preliminary favorable electromyographic response to the combined intervention, particularly in the tibialis anterior. Given the small sample size, baseline differences, and inter-individual variability, the results should be interpreted as exploratory. Larger randomized studies incorporating baseline-adjusted analyses and functional gait outcomes are required to determine the clinical relevance of these electromyographic changes.**

***Index Terms*—stroke; repetitive transcranial magnetic stimulation; motor imagery; brain-computer interface; electromyography; neurological rehabilitation.**

This work was supported by the Espírito Santo Research and Innovation Support Foundation (FAPES).

I. Introduction

Stroke is one of the leading causes of acquired disability in adults, often resulting in persistent motor deficits that impair gait, balance, and functional independence. Hemiparesis, present in approximately 80% of stroke survivors, is associated with reduced muscle recruitment, altered motor control, and limited mobility, significantly impacting the quality of life of affected individuals [1,2].

Motor recovery after stroke depends on the ability of the central nervous system to reorganize its neural connections through neuroplasticity mechanisms. In this context, techniques capable of modulating cortical excitability and promoting motor relearning have received increasing attention in neurorehabilitation. Among these, repetitive transcranial magnetic stimulation (rTMS) and motor imagery (MI) stand out as approaches associated with promoting cortical reorganization and improving motor function [3].

rTMS is a non-invasive neuromodulation technique that may help restore interhemispheric balance after stroke, thereby contributing to functional recovery. Complementarily, MI consists of the mental representation of movements without their physical execution, activating neural circuits similar to those recruited during the actual performance of motor tasks. When combined with brain-computer interfaces (BCIs), this strategy

enables the generation of real-time sensorimotor feedback, thereby enhancing neuroplastic processes [4-8].

Surface electromyography (sEMG) provides an objective approach for quantifying neuromuscular responses to rehabilitation interventions. Although previous clinical studies have investigated rTMS, motor imagery, BCI-based rehabilitation, and technology-assisted gait training after stroke, the neuromuscular effects of combining cortical neuromodulation with BCI-mediated motor imagery during locomotor training remain insufficiently characterized. In particular, limited evidence is available regarding changes in the electromyographic activation of specific lower-limb muscles following this multimodal approach. Therefore, the main contribution of this pilot study is the use of sEMG to objectively characterize changes in the activation of the tibialis anterior and rectus femoris following a combined rTMS and BCI-mediated motor imagery protocol. These muscles were selected because of their relevant roles in ankle dorsiflexion, lower-limb control, and gait. Accordingly, this study aimed to explore the preliminary effects of the combined intervention on peak electromyographic activation in post-stroke individuals.

II. Methodology

A. Study Design and Participants

This study was characterized as a pilot, randomized, controlled, assessor-blinded clinical trial conducted at the Physical Rehabilitation Center of Espírito Santo (CREFES in Vila Velha-ES, Brazil). The study was approved by the Research Ethics Committee of the Federal University of Espírito Santo (Approval No. 7.567.540), and all participants provided written informed consent prior to enrollment.

Individuals were eligible if they had a diagnosis of stroke, were over 18 years of age, presented with residual hemiparesis or hemiplegia, and had preserved cognitive and communication abilities. Participants were excluded if they had a history of epilepsy or seizures, previous craniectomy, intracranial metallic implants, or any contraindication to rTMS.

Following the baseline assessment, participants were randomly assigned to one of two groups: the Intervention Group (rTMS + MI) or the Control Group (MI). Participants were unaware of their group allocation, and all assessments were performed by a physical therapist who was blinded to the experimental groups.

B. Experimental Procedures

The Intervention Group received rTMS over the cortical representation of the tibialis anterior muscle, which had been previously identified through electromyographic mapping. Stimulation was delivered at an intensity corresponding to 100% of the motor threshold, five sessions per week for two consecutive weeks. The stimulation parameters were as follows: frequency of 10 Hz, 10 trains of stimulation, 100 pulses per train, and an inter-train interval of 50 seconds. In addition to rTMS, participants also underwent MI-based BCI (Fig. 1) three times per week for two weeks.

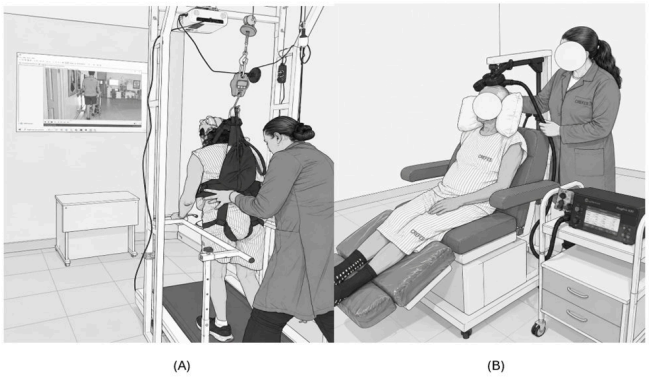


Fig. 1. Illustration of robotic treadmill gait training controlled by BCI combined with MI (A), and the application of rTMS (B). Images derived from the authors' original photographs, edited to remove elements unrelated to the procedure and to preserve participant anonymity.

During the training sessions, participants remained in an upright standing position on a body weight-supported treadmill. Brain activity was recorded using electroencephalography (EEG) with a 16-channel cap (OpenBCI). The brain signals generated during MI were processed by the BCI and used to control the treadmill, providing real-time sensorimotor feedback. Each BCI training session lasted 6 minutes.

The Control Group underwent only the BCI and treadmill training, using the same procedures and duration as the Intervention Group, but without prior application of rTMS.

C. Electromyographic Assessment

The electromyographic activity of the tibialis anterior and rectus femoris muscles of the affected lower limb was assessed before and after the intervention period using sEMG with the New MioTool system (Miotec, Brazil). For each muscle, three 8-second maximum voluntary contractions were recorded, with 12-second rest intervals between contractions. The outcome analyzed was the peak electromyographic activation. The values obtained in the pre- and post-intervention assessments were compared across time points and among the experimental groups.

III. Results

As aforementioned, ten individuals diagnosed with stroke were included and randomly assigned to two groups: the Intervention Group (n = 5) and the Control Group (n = 5). The sample had a mean age of 61.5 years and consisted of five men and five women. All participants had a diagnosis of systemic arterial hypertension (SAH). Regarding stroke etiology, 60% of the participants had experienced an ischemic stroke, whereas 40% had experienced a hemorrhagic stroke (Table 1), characterizing a clinically heterogeneous population that is nevertheless consistent with the epidemiological profile commonly observed in neurological rehabilitation programs.

The electromyographic analysis of the peak contraction of the rectus femoris and tibialis anterior muscles demonstrated distinct responses between the groups evaluated after the

TABLE I
SAMPLE CHARACTERIZATION

Sample Characteristics	
Average Age	61.5 Years
Men	50%
Women	50%
Hypertension	100%
Ischemic Stroke	60%
Hemorrhagic Stroke	40%

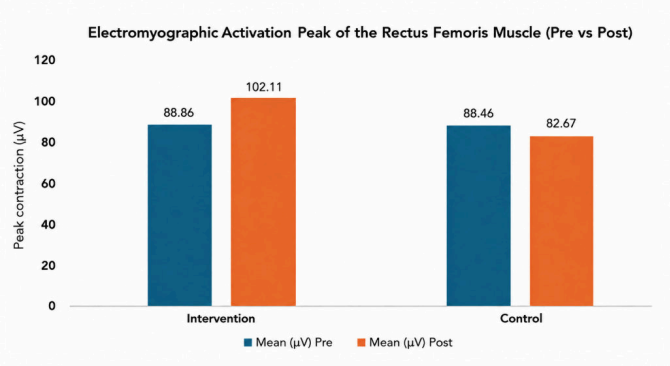


Fig. 2. Comparison of the mean peak of the electromyographic activation (μV) for the rectus femoris muscle in the Intervention and Control Group, before (Pre) and after (Post) the intervention period.

intervention period. Table 2 presents the results comparing the pre- and post-intervention time instants for both groups.

In the Intervention Group, an increase was observed in the mean electromyographic activity of the rectus femoris muscle, rising from 88.86 ± 74.20 μV at the initial assessment to 102.11 ± 102.46 μV at the final assessment (Fig. 2), representing a mean increase of 13.26 μV (+14.9%). Individually, three of the five participants showed an increase in peak muscle activation, one showed a decrease, and one maintained values similar to those observed at baseline. The calculated effect size was small (Cohen's dz = 0.30), indicating a modest pre–post change in peak electromyographic activation. Given the small sample size and substantial inter-individual variability, this finding should be interpreted descriptively.

For the tibialis anterior muscle, the Intervention Group showed a more pronounced increase in electromyographic activity. The mean peak of contraction increased from 46.70 ± 34.57 μV to 93.79 ± 61.72 μV after treatment (Fig. 3), corresponding to a mean increase of 47.09 μV (+100.8%). Four out of the five participants demonstrated increased muscle activation, whereas only one participant showed a reduction in electromyographic values at the end of the protocol. The effect size was moderate (Cohen's dz = 0.54); however, considering the small sample size and variability of the measurements, this result should be interpreted as a preliminary signal of increased electromyographic activation rather than evidence of treatment efficacy.

In contrast, the Control Group exhibited a different pattern. For the rectus femoris muscle, the mean peak electromyographic activation decreased from 88.46 ± 101.53 μV to 82.67 ± 65.24 μV, representing a mean reduction of 5.79 μV (-6.5%). Although one participant showed a marked increase in muscle activity, the remaining participants exhibited decreases or only minor changes, contributing to the overall reduction in the group mean. The effect size was negligible (Cohen's $d_z = -0.07$), indicating minimal pre–post change in peak electromyographic activation.

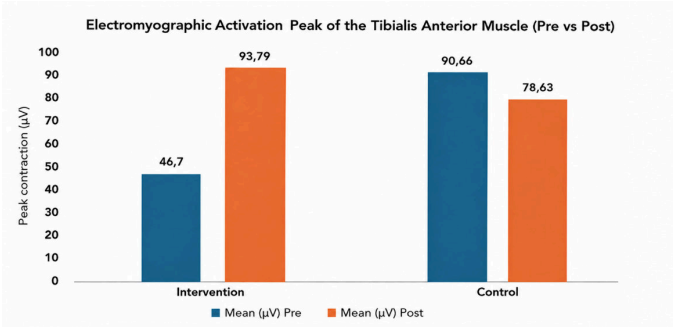


Fig. 3. Comparison of the mean peak of the electromyographic activation (μV) for the tibialis anterior muscle in the Intervention and Control Group, before (Pre) and after (Post) the intervention period.

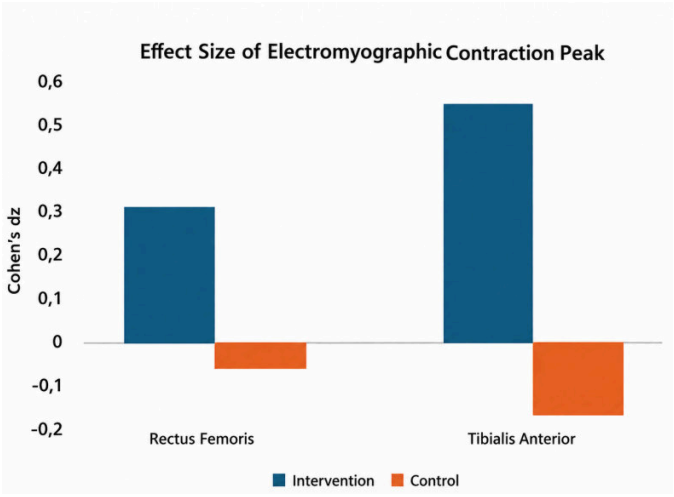


Fig. 4. Effect size (Cohen's d_z) of the pre–post intervention changes in the peak electromyographic activation of the rectus femoris and tibialis anterior muscles. The Intervention group showed positive effects ranging from small ($d_z = 0.30$) to moderate magnitude ($d_z = 0.54$), whereas the Control Group showed negative effects of trivial to small magnitude ($d_z = -0.07$ and $d_z = -0.16$, respectively), indicating greater muscle responsiveness in the group that underwent the intervention.

Similar findings were observed for the tibialis anterior muscle in the Control Group. Mean electromyographic activity decreased from 90.66 ± 126.64 μV to 78.63 ± 71.97 μV after the intervention period, corresponding to a mean reduction of 12.03 μV (-13.3%). Although two participants showed increased muscle activation, three demonstrated reduced peak electromyographic values at the end of the protocol. The observed effect size was small and negative (Cohen's $d_z = -0.16$), as shown in Fig. 4.

Descriptively, the groups showed different patterns of pre–

TABLE II
ELECTROMYOGRAPHIC OUTCOMES BEFORE AND AFTER THE INTERVENTION.

Assessed Muscle	**Group**	**Pre-intervention (μV) Mean $\pm$ SD**	**Post-intervention (μV) Mean $\pm$ SD**	**Δ Absolute (μV)**	**Δ (%)**	**Cohen's d_z**
Rectus Femoris	Intervention	88.86 $\pm$ 74.20	102.11 $\pm$ 102.46	+13.26	+14.92%	0.30
Rectus Femoris	Control	88.46 $\pm$ 101.53	82.67 $\pm$ 65.24	-5.79	-6.54%	-0.07
Tibialis Anterior	Intervention	46.70 $\pm$ 34.57	93.79 $\pm$ 61.72	+47.09	+100.84%	0.54
Tibialis Anterior	Control	90.66 $\pm$ 126.64	78.63 $\pm$ 71.97	-12.03	-13.27%	-0.16

Caption: SD = Standard Deviation; Δ = difference between the post- and pre-intervention time points; Cohen's d_z = within-group effect size for paired pre–post observations.

post change. The Intervention Group exhibited increases in mean peak electromyographic activation of both muscles, with the largest relative change observed in the tibialis anterior ($+100.8\%$), whereas the Control Group showed reductions in mean activation of the rectus femoris (-6.5%) and tibialis anterior (-13.3%). However, the baseline tibialis anterior values differed substantially between groups and showed considerable dispersion, particularly in the Control Group. Therefore, percentage changes should be interpreted cautiously and should not be considered direct evidence of between-group efficacy.

Overall, these findings indicate a preliminary favorable electromyographic signal following the combined rTMS and BCI-mediated MI intervention, particularly for the tibialis anterior. Given the pilot sample size, these results are primarily descriptive and hypothesis-generating.

IV. DISCUSSION

The present study investigated the effects of rTMS combined with MI on the electromyographic activation of the tibialis anterior and rectus femoris muscles in post-stroke individuals. The results demonstrated an increase in electromyographic activity peak in both muscles in the intervention group, whereas the Control Group showed a reduction in muscle activation at the end of the protocol. The most pronounced effect was observed in the tibialis anterior muscle, which exhibited a mean increase of more than 100% after the treatment period.

The principal contribution of this study lies not only in the multimodal combination of rTMS, BCI-mediated motor imagery, and locomotor training, but particularly in the use of sEMG to objectively examine the neuromuscular response of specific lower-limb muscles. Previous clinical investigations have primarily focused on motor, gait, balance, or global functional outcomes when evaluating neuromodulation and technology-assisted rehabilitation after stroke. In this context, the present study provides complementary physiological information by quantifying changes in muscle activation following the intervention. Thus, rather than establishing clinical efficacy, these findings contribute preliminary evidence regarding the neuromuscular response associated with this multimodal rehabilitation strategy.

Although the present study did not directly assess cortical reorganization or neuroplasticity, the observed electromyographic changes may be interpreted within the broader neurophysiological framework of rTMS-induced modulation of motor cortical excitability. Studies have shown that low-frequency rTMS applied over the contralesional hemisphere reduces excessive transcallosal inhibition exerted on the affected motor cortex, promoting the restoration of interhemispheric balance and functional recovery [9–11]. In the meta-analysis by Dionísio et al. (2018) [5], positive effects of rTMS on motor function in post-stroke individuals were observed, particularly when combined with conventional rehabilitation programs.

In addition to neuromodulation, MI plays an important role in activating neural circuits related to movement planning and execution. Evidence from neuroimaging studies demonstrates that MI recruits cortical areas similar to those activated during actual motor task execution, including the primary motor cortex, premotor cortex, and associative parietal areas. Thus, the combination of rTMS and MI may produce synergistic effects on neural plasticity by enhancing the recruitment of residual motor pathways after brain injury [12,13]. A systematic review conducted by Bai et al. (2020) [7] concluded that MI programs significantly improve motor function in post-stroke individuals, particularly when combined with sensorimotor feedback technologies.

The increase in tibialis anterior activation is potentially relevant from a clinical perspective because this muscle plays a fundamental role in ankle dorsiflexion during the swing phase of gait and contributes to adequate foot clearance. Impaired tibialis anterior activation after stroke may contribute to foot drop, altered gait patterns, reduced walking efficiency, and increased risk of falls. Therefore, the electromyographic changes observed in the present study provide a physiologi-

cally relevant signal that warrants further investigation. However, increased peak sEMG activation cannot be interpreted as evidence of improved gait performance. Walking speed, balance, ankle kinematics, functional mobility, and independence were not evaluated in the present study. Future trials should determine whether changes in tibialis anterior activation are associated with clinically meaningful improvements in outcomes such as gait speed, balance, walking capacity, and functional independence.

Several limitations should be considered when interpreting these findings. First, the very small sample size (n = 5 per group) substantially limits statistical power, increases the influence of individual responses and extreme values, and prevents definitive conclusions regarding between-group efficacy. Accordingly, the observed effect sizes and percentage changes should be considered exploratory estimates rather than confirmatory evidence. Second, the sample was clinically heterogeneous, including participants with both ischemic and hemorrhagic stroke, and potentially different levels of motor impairment. Such heterogeneity may contribute to variability in responsiveness to neuromodulation and rehabilitation. Third, the baseline electromyographic values of the tibialis anterior differed considerably between groups, and the large standard deviation observed in the Control Group indicates substantial inter-individual dispersion. Consequently, the relative percentage increase observed in the Intervention Group should be interpreted cautiously because it may be influenced by baseline imbalance. Finally, the intervention lasted only two weeks, limiting the interpretation of the findings to short-term neuromuscular responses. The study therefore cannot determine whether the observed electromyographic changes are sustained over time or translate into long-term functional improvement.

From a feasibility perspective, several lessons can be drawn from this pilot study. First, integrating rTMS with BCI-mediated motor imagery and treadmill-based locomotor training was feasible within a short rehabilitation protocol. Second, sEMG provided an objective and potentially sensitive measure for detecting individual neuromuscular responses that may not be captured by group-level clinical measures alone. Third, the marked inter-individual variability observed in the electromyographic responses highlights the importance of detailed participant characterization and individualized analysis in future neuromodulation trials. Finally, the baseline differences observed between groups reinforce the need for baseline-adjusted statistical approaches in subsequent larger randomized studies.

Important research gaps also remain. Future studies should investigate whether the electromyographic changes observed after combined rTMS and BCI-mediated motor imagery are accompanied by changes in cortical excitability and whether these physiological responses translate into clinically meaningful improvements in gait and mobility. Larger randomized trials should incorporate baseline-adjusted between-group analyses, longer intervention and follow-up periods, and standardized functional outcomes, including gait speed, balance, walking capacity, ankle dorsiflexion, and functional independence. Further investigation is also needed to identify clinical or neurophysiological characteristics associated with responsiveness to multimodal neuromodulation protocols.

Taken together, the present findings should be interpreted as preliminary and hypothesis-generating. They provide an exploratory signal that the addition of rTMS to BCI-mediated motor imagery and locomotor training may influence lower-limb electromyographic activation, particularly in the tibialis anterior, but larger and adequately powered trials are required to determine the magnitude, reproducibility, and clinical relevance of this response.

V. Conclusion

This randomized pilot study identified a preliminary pattern of increased peak electromyographic activation of the tibialis anterior and rectus femoris following a short-term intervention combining rTMS with BCI-mediated motor imagery and locomotor training. The most pronounced descriptive change was observed in the tibialis anterior, whereas the Control Group showed reductions in mean electromyographic activation over the same period.

These findings should be interpreted as exploratory rather than as evidence of clinical efficacy because of the small sample size, inter-individual variability, baseline differences, clinical heterogeneity, and short intervention period. The use of sEMG represents an important contribution of this study by providing an objective physiological measure of lower-limb neuromuscular response. Larger randomized trials incorporating baseline-adjusted between-group analyses, functional gait and balance outcomes, and longitudinal follow-up are required to determine whether these electromyographic changes are reproducible and translate into clinically meaningful improvements in post-stroke mobility.

References

[1] R. S. Calabrò, G. Sorrentino, A. Cassio, D. Mazzoli, E. Andrenelli, E. Bizzarini, *et al.*, "Robotic-assisted gait rehabilitation following stroke: A systematic review of current guidelines and practical clinical recommendations," *Eur. J. Phys. Rehabil. Med.*, vol. 57, no. 3, pp. 460–471, 2021, doi: 10.23736/S1973-9087.21.06887-8.

[2] Z. Gao, Z. Pang, Y. Chen, G. Lei, S. Zhu, G. Li, *et al.*, "Restoring after central nervous system injuries: Neural mechanisms and translational applications of motor recovery," *Neurosci. Bull.*, vol. 38, no. 12, pp. 1569–1587, 2022, doi: 10.1007/s12264-022-00959-x.

[3] J. A. Kleim and T. A. Jones, "Principles of experience-dependent neural plasticity: Implications for rehabilitation after brain damage," *J. Speech Lang. Hear. Res.*, vol. 51, no. 1, pp. S225–S239, 2008, doi: 10.1044/1092-4388(2008/018).

[4] L. Sebastianelli, V. Versace, S. Martignago, *et al.*, "Low-frequency rTMS of the unaffected hemisphere in stroke patients: A systematic review," *Acta Neurol. Scand.*, vol. 136, no. 6, pp. 585–605, 2017.

[5] A. Dionísio, I. C. Duarte, M. Patrício, and M. Castelo-Branco, "Effects of repetitive transcranial magnetic stimulation on motor recovery after stroke: A systematic review and meta-analysis," *J. Neurol. Sci.*, vol. 387, pp. 188–193, 2018, doi: 10.1016/j.jns.2018.02.003.

[6] M. A. Cervera, S. R. Soekadar, J. Ushiba, *et al.*, "Brain-computer interfaces for post-stroke motor rehabilitation: A meta-analysis," *Ann. Clin. Transl. Neurol.*, vol. 5, no. 5, pp. 651–663, 2018, doi: 10.1002/acn3.544.

[7] Z. Bai, K. N. K. Fong, J. J. Zhang, J. Chan, and K. H. Ting, "Immediate and long-term effects of motor imagery training on upper extremity motor function after stroke: A systematic review and meta-analysis," *Clin. Rehabil.*, vol. 34, no. 9, pp. 1138–1153, 2020, doi: 10.1177/0269215520932968.

[8] J. He, Z. Yuan, L. Quan, H. Xi, J. Guo, D. Zhu, M. Chen, B. Yang, Z. Cui, S. Zhu, and J. Qiao, "Multimodal assessment of a BCI system for stroke rehabilitation integrating motor imagery and motor attempts: A randomized controlled trial," *J. Neuroeng. Rehabil.*, vol. 22, no. 1, Art. no. 185, 2025, doi: 10.1186/s12984-025-01723-8.

[9] Y. Li, J. Fan, J. Yang, *et al.*, "Repetitive transcranial magnetic stimulation for motor recovery after stroke: An updated systematic review and meta-analysis," *Front. Neurol.*, vol. 14, Art. no. 1141021, 2023, doi: 10.3389/fneur.2023.1141021.

[10] A. Jehangir, I. Ahmad, Y. Feng, W. W. Juan, T. Chen, and Y. Yin, "Combined effects of robotic-assisted gait training with transcranial magnetic stimulation on gait metrics and balance in stroke patients: A pilot randomized controlled trial," *Front. Neurol.*, vol. 16, Art. no. 1655409, 2025, doi: 10.3389/fneur.2025.1655409.

[11] J. D. Edwards, A. U. Dominguez-Vargas, C. Rosso, M. Branscheidt, L. Sheehy, F. Quandt, *et al.*, "A translational roadmap for transcranial magnetic and direct current stimulation in stroke rehabilitation: Consensus-based core recommendations from the third Stroke Recovery and Rehabilitation Roundtable," *Int. J. Stroke*, vol. 19, no. 2, pp. 145–157, 2024, doi: 10.1177/17474930231203982.

[12] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain-computer interfaces for communication and rehabilitation," *Nat. Rev. Neurol.*, vol. 12, no. 9, pp. 513–525, 2016, doi: 10.1038/nrneurol.2016.113.

[13] B. Premchand, Z. Zhang, K. K. Ang, J. Yu, I. O. Tan, J. P. W. Lam, A. X. Y. Choo, A. Sidarta, P. W. H. Kwong, and L. H. C. Chung, "A personalized multimodal BCI-soft robotic system for upper limb rehabilitation in chronic stroke patients," *Biomimetics*, vol. 10, no. 2, Art. no. 94, 2025, doi: 10.3390/biomimetics10020094.

Embedded AI-Based Handgrip System for Real-Time Dynamic Performance Evaluation

Marcos Dinís Lavarda
Federal Institute of Paraíba (IFPB)
Department of Electrical Engineering
João Pessoa, Paraíba, Brazil
marcos.lavarda@ifpb.edu.br

Caio Cunha Rego de Oliveira
Federal Institute of Paraíba (IFPB)
Department of Electrical Engineering
João Pessoa, Paraíba, Brazil
caio.cunha@academico.ifpb.edu.br

Thiago de Carvalho Batista
Federal Institute of Paraíba (IFPB)
Department of Electrical Engineering
João Pessoa, Paraíba, Brazil
thiago.batista@ifpb.edu.br

Abstract— This work presents an intelligent embedded system for real-time evaluation of performance in handgrip exercises, based on a force-sensitive resistor (FSR) and embedded artificial intelligence. The device integrates a 3D-printed mechanical structure with an analog acquisition circuit, employing fourth-order polynomial calibration of the sensor, which achieved R^2 = 0.9976 across the available reference loads. The system performs local signal processing and extracts force-tracking and execution features, including curve error, area error, stability, slope error, initial force level, and final force level. These features are used as input to a lightweight neural network model to estimate exercise performance after its completion. Experimental results indicate predictive performance with $R^2 \approx 0.95$, while inference is performed locally on the embedded device at the end of each exercise. The embedded system provides both quantitative feedback, inferred by the neural network, and qualitative feedback based on patterns such as instability, improper force application, and imprecise execution. The proposed solution represents a portable and low-cost alternative for objective evaluation of handgrip exercises, with potential applications in rehabilitation, motor training, and functional monitoring.

Keywords— Embedded systems, Force sensors, Signal processing, Real-time systems.

I. INTRODUCTION

Handgrip strength is an important functional indicator widely used in the assessment of motor capacity, rehabilitation, and overall health and motor function monitoring [1], [2]. Isometric handgrip exercises have been shown to produce clinically significant reductions in blood pressure, with decreases of approximately 6–8 mmHg in systolic pressure after training [3].

However, studies report a lack of standardization in evaluation protocols, with inconsistencies in acquisition, intensity, and repetition, hindering cross-study comparisons [2], [4].

Recent rehabilitation systems have incorporated sensory feedback and machine learning techniques to improve force control and provide real-time evaluation [5], [6]. Nevertheless, many of these solutions still rely on high-cost equipment or subjective analysis, limiting their applicability in home-based settings.

Sensor-based approaches, particularly those using force-sensitive resistors (FSRs), enable low-cost solutions integrated with embedded systems; however, they are primarily focused on static force measurements [7]–[9].

In addition, traditional evaluation methods predominantly rely on maximum force measurements, disregarding the temporal dynamics of movement. Force-tracking tasks use reference signals over time, in which users continuously adjust the applied force, and performance is evaluated using error-based metrics [10], [11].

In this context, this work proposes an embedded system for dynamic performance evaluation in handgrip exercises. The proposed approach integrates signal acquisition, temporal feature extraction, and local inference based on an embedded multilayer perceptron (MLP) neural network. The results indicate predictive performance suitable for the intended application ($R^2 \approx 0.95$), with on-device operation on low-cost hardware.

The system integrates sensing using a force-sensitive resistor (FSR), temporal storage of the exercise signal, feature extraction, and score inference through embedded artificial intelligence in a fully autonomous system for training evaluation and feedback.

The main contribution is the development of a support tool for healthcare professionals in the prescription, monitoring, and evaluation of handgrip exercises, including both static and dynamic effort tasks ranging from 0 to 10 kgf with variable duration. The proposed tool can be configured for use in future studies and clinical applications.

II. SYSTEM OVERVIEW

A. Assembly and Hardware Architecture

Several approaches for measuring handgrip force have been proposed, including traditional dynamometers and FSR-based sensors [2].

Fig. 1 illustrates the mechanical design of the proposed device, in which an FSR 402 sensor is integrated into a handgrip structure designed to ensure uniform distribution of the applied force over its sensitive area. This design improves measurement repeatability and consistency in data acquisition using the ESP32 microcontroller (Microcontroller Unit, MCU), as well as the consistency of the feedback provided through a display and buzzer.

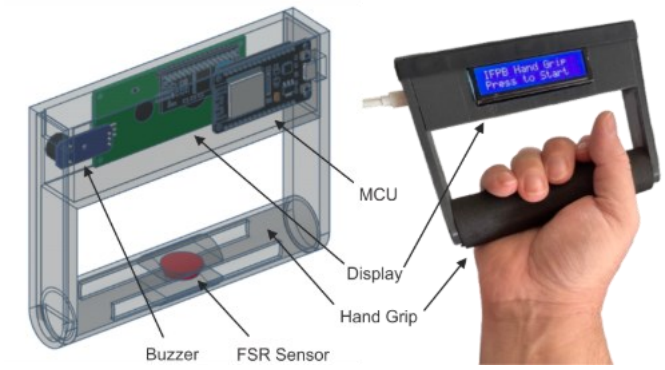


Fig. 1. Hardware layout and final construction of the handgrip device.

B. Signal Acquisition and Calibration

The FSR sensor exhibits nonlinear behavior, with resistance inversely proportional to the applied force [12], [13].

In this work, the analog signal is acquired using a resistive divider (1 kΩ pull-down) with an input voltage of 3.3 V. The signal is sampled by the 12-bit analog-to-digital converter (ADC) of the ESP32.

Due to sensor nonlinearity, a fourth-order polynomial calibration was adopted to convert ADC values into force. Calibration was performed experimentally using loads from 1 to 10 kgf with a precision scale as reference.

Based on the collected calibration data, a fourth-order polynomial model was fitted, achieving a coefficient of determination of $R^2 = 0.9976$, as shown in Eq. (1), where $x$represents the 12-bit ADC digital value:

$$F(x) = 5.04 \times 10^{-13}x^4 - 2.68 \times 10^{-9}x^3 + 3.86 \times 10^{-6}x^2 + 1.85 \times 10^{-3}x - 2.92 \quad (1)$$

The calibration procedure achieved a coefficient of determination of $R^2 = 0.9976$ across the available reference loads, indicating a strong fit of the adopted fourth-order polynomial to the calibration data. However, the present study did not calculate a separate absolute calibration error, and further validation using independent reference measurements is required to characterize measurement accuracy.

Fig. 2 shows the resulting calibration curve, highlighting the strong correlation between the experimental data and the fitted model.

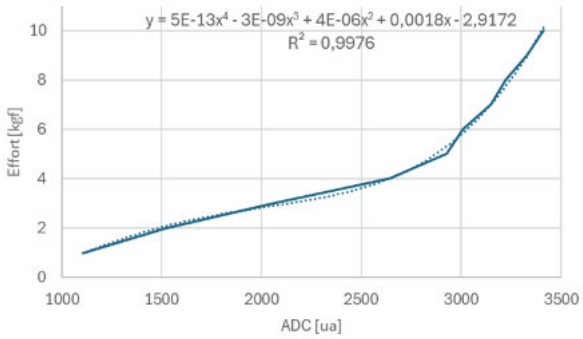


Fig. 2. Calibration curve of the FSR sensor with polynomial fitting.

C. System Workflow

The system operates in real time following a structured processing pipeline, as illustrated in Fig. 3, consisting of the following steps:

1. Continuous acquisition and ADC conversion of the force signal;
2. Application of a nonlinear fourth-order polynomial calibration;
3. Storage of the complete exercise in a temporal buffer;
4. Extraction of force-tracking and execution features;
5. Score inference using a feedforward neural network after exercise completion; and
6. User display (score and qualitative feedback).

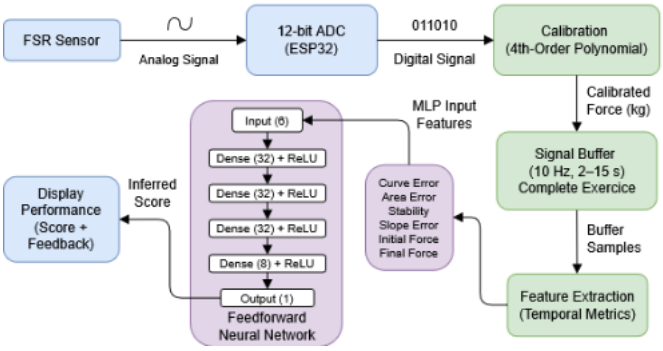


Fig. 3. Block diagram of the proposed system.

D. Exercise Protocol and Interface

Isometric handgrip exercise protocols typically involve sustained contractions at controlled force levels over time, usually defined as a percentage of the maximum voluntary contraction (MVC), organized into sets with different durations [14].

Force-tracking approaches use reference signals over time, in which users continuously adjust the applied force, and performance is evaluated based on error metrics between the target signal and the measured response [10], [11]. Different excitation profiles, such as ramp, constant, and periodic signals, allow the assessment of distinct force modulation behaviors.

Fig. 4 shows the system interface, including pre-exercise instructions (a) and real-time feedback (b).

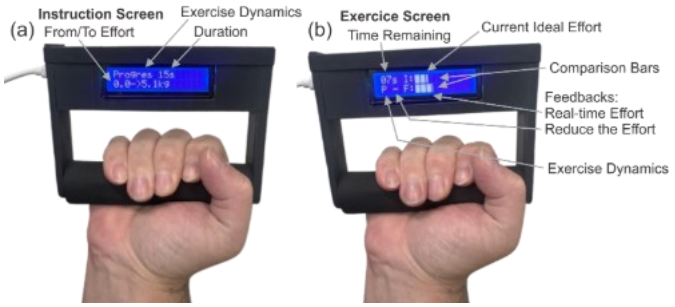


Fig. 4. Pre-exercise and real-time feedback interface: (a) Exercise instruction—15 seconds of progressive effort from 0 to 5.1 kgf. (b) Real-time feedback—Progressive exercise (P), remaining time (s), ideal force (I), applied force feedback (F), and guidance to reduce (−) or increase (+) effort.

The system was designed to support different isometric handgrip exercise profiles, allowing force parametrization over time. The operating range and duration of the exercises were defined considering low-intensity execution and real-time constraints, while remaining adaptable to specific applications and clinical requirements.

Three main profiles were defined:

- **Progressive (P):** gradual increase in force;
- **Static (S):** maintenance of force at a target level;
- **Regressive (R):** controlled reduction of force.

Exercises are performed within time windows ranging from 2 to 15 s, with force levels between 2 and 10 kgf. The profiles include progressive (0 to 2–10 kgf), constant (2–10 kgf), and regressive (2–10 kgf to 0) variations, enabling the analysis of force modulation during execution.

The interface provides real-time feedback on execution accuracy through two force bars: the upper bar (I) represents the ideal force, while the lower bar (F) indicates the user's applied force. Additionally, visual cues are provided: "–" suggests reducing effort, "+" increasing effort, and "*" indicates correct execution within the target range.

This approach integrates reference-based protocols and force-tracking tasks, enabling evaluation of both force control and trajectory tracking.

III. Signal Processing and Inference

A. Feature Extraction

Performance evaluation is based on features extracted from the executed force trajectory and its reference profile, describing force-tracking accuracy and execution dynamics. Six main metrics were defined: curve error, area error, stability, slope error, initial force level, and final force level.

The system is aligned with evaluation methods based on tracking error, in which performance is quantified by the difference between the executed signal and the reference trajectory over time [11].

To formalize these metrics, let $y[n]$ and $r[n]$ represent the measured and reference force signals at sample n, respectively. Let N denote the total number of samples, $F_{initial}$ and F_{final} the initial and final force levels of the reference trajectory, t_n the time associated with sample n, and T the target exercise duration.

For progressive and regressive exercises, in which the reference trajectory changes over time, the curve error is normalized by the absolute difference between the initial and final force levels, as follows:

$$\mathrm{E}_c^{dyn} = \frac{1}{N \mid F_{final} - F_{initial} \mid} \sum_{n=1}^{N} |y[n] - r[n]| \qquad (2)$$

where E_c^{dyn} denotes the curve error for dynamic exercises.

For static exercises, where the initial and final reference force levels are equal, the normalization is performed using the target force level with the offset adopted in the embedded implementation:

$$\mathrm{E}_c^{stat} = \frac{1}{N(F_{initial} + 0.1)} \sum_{n=1}^{N} |y[n] - r[n]| \qquad (3)$$

where E_c^{stat} denotes the curve error for static exercises.

The 0.1 term corresponds to the numerical offset used in the embedded implementation to avoid an excessively small normalization denominator.

Area error quantifies the relative deviation between the measured and reference areas under the force-time curves. The measured area is obtained using the trapezoidal rule, while the reference area is determined from the initial and final force levels and the target exercise duration:

$$E_a = \frac{\left| \sum_{n=2}^{N} \frac{y[n]+y[n-1]}{2} (t_n - t_{n-1}) - \frac{F_{initial}+F_{final}}{2} T \right|}{\frac{F_{initial}+F_{final}}{2} T} \qquad (4)$$

This metric represents the deviation in the total applied effort relative to the expected effort for the prescribed trajectory.

Stability is defined as a bounded measure of local force consistency based on the mean absolute variation between consecutive measured samples. The resulting variation is transformed into a stability score, such that larger values indicate more stable force application:

$$S = \frac{1}{1 + \frac{1}{N} \sum_{n=2}^{N} |y[n] - y[n-1]|} \qquad (5)$$

Thus, lower sample-to-sample variations result in stability values closer to one.

Slope error corresponds to the absolute difference between the global slope of the executed force trajectory and the slope prescribed by the reference trajectory:

$$E_s = \left| \frac{y[N] - y[1]}{t_N - t_1} - \frac{F_{final} - F_{initial}}{T} \right| \qquad (6)$$

The executed slope is calculated from the measured force variation over the elapsed execution time, whereas the reference slope is determined from the prescribed initial and final force

levels and the target duration. Because the slope is zero for an ideal static profile, slope error is primarily informative for progressive and regressive exercises.

The initial and final force levels define the boundary conditions of the reference trajectory. Specifically, $F_{initial}$corresponds to the prescribed force at the beginning of the exercise, whereas F_{final}corresponds to the prescribed force at the end of the target duration. The initial force level is defined as:

$$r(0) = F_{initial} \tag{7}$$

The final force level is defined as:

$$r(T) = F_{final} \tag{8}$$

These variables characterize the prescribed force profile rather than independent physiological measures. They determine whether the exercise begins, maintains, or ends at a given force level and therefore provide contextual information for interpreting the force-tracking metrics.

Unlike the measured force samples used in the error metrics, these two variables describe the boundary conditions of the prescribed exercise profile. In progressive exercises, the reference changes from a lower initial force level to a higher final force level; in regressive exercises, the opposite occurs; and in static exercises, the two levels are equal.

Together, these features provide a compact representation of force-tracking accuracy, local force consistency, execution rate, and the boundary conditions of the reference trajectory. This low-dimensional representation is suitable as input to an embedded machine-learning model.

B. Machine Learning Model

The extracted features are used as input to a feedforward neural network implemented in TensorFlow/Keras to model nonlinear relationships between force-tracking descriptors and the defined performance score.

Unlike approaches that directly process time series data, such as recurrent models (e.g., LSTM) [6], this work adopts a feature-based strategy, reducing the dimensionality of the problem and eliminating the need for full sequential processing.

The input vector consists of six variables: curve error, area error, stability, slope error, initial force level, and final force level. The architecture comprises four fully connected hidden layers with 32, 32, 32, and 8 neurons, respectively, utilizing ReLU activation functions, and a single linear output node, as illustrated in Fig. 3.

It is important to emphasize that this study represents an initial proof of concept focused on validating the proposed Edge AI architecture on a microcontroller. Therefore, to evaluate the system's operational viability and tracking capabilities without requiring prior ethical committee approval for human trials, a dataset of approximately 380 executions was generated using simulated data. These simulated profiles were designed to represent expected dynamic force variations with different performance levels. The target value (score) for these profiles was defined based on the deviation from the reference trajectory, considering factors such as stability, delay, improper force application, and interruptions.

For model training and evaluation, the dataset was randomly partitioned into 80% for model development and 20% for final testing using a fixed random seed. Within the training subset, 20% was further reserved for validation during training. Input and output normalization parameters were fitted exclusively using the training subset and subsequently applied to the validation and test data. The model was trained using the Adam optimizer and mean squared error (MSE) loss function and evaluated using MAE, RMSE, and R^2 on the held-out test subset. As an initial technological proposition, this simulated approach was used to evaluate the proposed embedded inference pipeline; future studies, pending ethics approval, will incorporate real user data for clinical validation.

After training, the model performs a single score inference at the end of each completed exercise. Additionally, a rule-based module provides qualitative feedback to the user.

C. Embedded Implementation

The trained model was converted to TensorFlow Lite format and integrated into the ESP32 as a binary array of approximately 45 kB. The system performs a single neural-network inference locally at the end of each completed exercise, after the force trajectory has been acquired and the features have been extracted. This architecture allows continuous force acquisition and feedback during exercise without requiring continuous neural-network inference or external communication.

Input and output normalization parameters are embedded within the system, ensuring consistency between training and inference.

The implementation follows the TinyML paradigm, enabling efficient execution of the model on resource-constrained devices and supporting portable and autonomous applications.

IV. Results

A. Model Evaluation Results

The model achieved the following predictive performance on the held-out test subset, with a mean absolute error (MAE) of 4.25, a root mean square error (RMSE) of 6.78, and a coefficient of determination of $R^2 = 0.946$. The results indicate a strong correspondence between the predicted and reference scores within the simulated test data.

Fig. 5 presents representative examples of model inference for different exercise profiles: (a) and (b) linear regressive exercises over 13.82 s, starting from 8.1 kgf to 0 kgf, separated into two plots to improve data visualization, with (a) emphasizing temporal deviations and (b) highlighting force amplitude errors; (c) a static exercise at 7.7 kgf for 6.82 s; and (d) a linear progressive exercise over 9.92 s, starting from 0 kgf to 8.31 kgf.

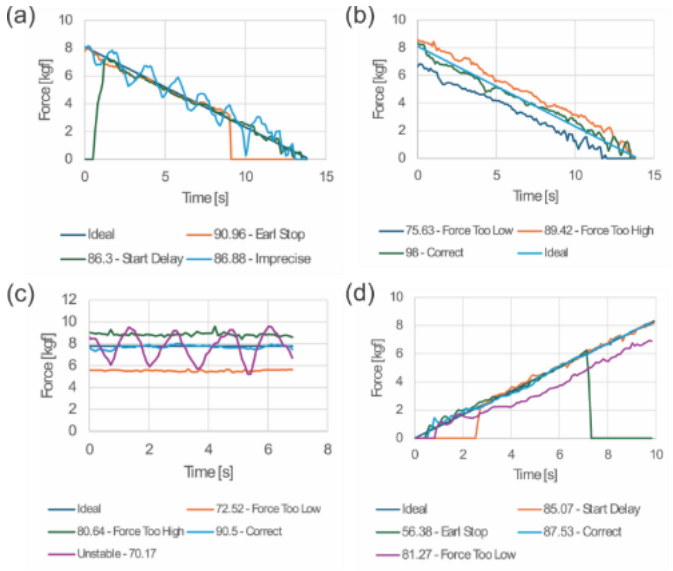


Fig. 5. AI inference results for exercises: (a) and (b) regressive from 8.1 kgf to 0 kgf in 13.82 s, displaying temporal deviations and force amplitude errors, respectively; (c) static at 7.7 kgf for 6.82 s; (d) progressive from 0 kgf to 8.31 kgf in 9.92 s.

The model distinguishes execution patterns, assigning scores consistent with observed behavior. In regressive cases (Fig. 5(a) and (b)), deviations such as early termination and delays (temporal errors) as well as inadequate force application (amplitude errors), reduce the score. In the static case (Fig. 5(c)), stability variations impact evaluation, while in the progressive case (Fig. 5(d)), delays and differences in force rate are reflected in the score.

Fig. 5 shows higher scores for executions close to the reference and lower scores for deviations such as instability and delays.

Overall, the neural network exhibits lower sensitivity to purely temporal deviations and higher sensitivity to amplitude-related errors for a given global error level. For instance, in the progressive exercise shown in Fig. 5(d), a delayed initiation (orange curve, score 85.07) has less impact on the score than early termination (green curve, score 56.38) for similar durations.

This behavior is consistent with the nature of progressive exercises, where force demand increases over time; thus, early termination may indicate insufficient strength development, whereas an initial delay may reflect minor instability in fine motor control at low force levels.

Similarly, in the regressive exercise shown in Fig. 5(a), an initial delay (green curve, score 86.3) is more penalized than early termination (orange curve, score 90.96), as the exercise requires an initial force level that must be properly established before reduction.

In the static exercise (Fig. 5(c)), the score reflects deviations below (orange curve, score 72.52) and above (green curve, score 80.64) the target force, proportional to the error magnitude. In contrast, unstable execution (purple curve, score 70.17) is more penalized, even when the global error is partially minimized, indicating the model's ability to capture distinct temporal patterns.

The results depend on the selected feature set and can be refined for specific clinical applications. The model is more sensitive to amplitude errors (area error) than to temporal deviations, while stability captures local variations not reflected in the global error. The initial and final force levels additionally encode the boundary conditions of the prescribed reference trajectory.

B. Discussion

Trajectory-based features enable the proposed model to represent force-tracking accuracy, local force consistency, execution rate, and reference-trajectory boundary conditions, providing a more comprehensive representation than static force-based measurements. Unlike methods focused on static measurement or grip classification [7], the system captures temporal dynamics through metrics such as curve error, area error, and stability, reflecting motor control and effort consistency and addressing limitations reported in the literature [2], [10], [11].

Furthermore, while several systems emphasize data acquisition and transmission [9], the proposed approach integrates acquisition, temporal processing, and local inference within a single architecture. Force acquisition and user feedback operate during exercise, while the neural network is invoked once after exercise completion, enabling autonomous on-device evaluation. The present study did not include a quantitative measurement of neural-network inference time on the ESP32. Although inference is executed only once after each completed exercise, future work should quantify execution time and its variability under different operating conditions.

The system supports multiple exercise profiles, allowing adaptation to different rehabilitation scenarios and integration with feedback-based applications and exergames [15], enhancing user engagement.

Embedded execution demonstrates the feasibility of low-cost intelligent systems for real-time functional assessment in clinical settings.

It is important to emphasize that the current study focuses on the technological validation of the proposed architecture. The inferred AI score is designed to evaluate a user's motor control and ability to follow a dynamic reference trajectory, rather than providing an absolute, clinically validated index. The calibration procedure achieved a coefficient of determination of $R^2 = 0.9976$ across the available reference loads, indicating a strong fit of the adopted fourth-order polynomial to the calibration data. However, the present study did not calculate a separate absolute calibration error, and further validation using independent reference measurements is required to characterize measurement accuracy.

While traditional dynamometers are widely used as reference instruments for static maximum grip strength assessment, this device aims to serve as a dynamic tracking and rehabilitation tool. No direct comparison with a reference dynamometer was performed in the present study. Similarly, repeatability across repeated executions was not quantitatively

characterized and should be investigated using repeated trials under controlled conditions. Future studies should therefore evaluate calibration accuracy, repeatability, and clinical validity using independent reference measurements, repeated trials, reference dynamometers, and real user data.

Future work will focus on configurable exercise profiles, quantitative characterization of inference time, independent calibration validation, comparison with reference dynamometers, repeatability assessment, and continuous model refinement using real user data, enabling more robust and potentially personalized rehabilitation strategies.

V. Conclusion

This work presented an intelligent embedded system for dynamic and static performance evaluation in handgrip exercises, integrating force sensing, signal processing, temporal feature extraction, and embedded artificial intelligence. The proposed approach enables on-device assessment with continuous force acquisition and feedback during exercise, while neural-network inference is performed once after exercise completion.

The model achieved an MAE of 4.25, an RMSE of 6.78, and an R^2 of 0.946 on the simulated held-out test subset. These results demonstrate the feasibility of the proposed embedded inference pipeline for dynamic force-tracking evaluation. However, the present study represents a technological proof of concept based on simulated data, and further validation with real users, independent reference measurements, repeatability analysis, and clinical protocols is required.

Acknowledgment

The authors would like to thank the Federal Institute of Education, Science and Technology of Paraíba (IFPB), João Pessoa campus, and PETEE-IFPB (Electrical Engineering Tutorial Education Program) for their technical and financial support.

References

[1] F. P. Kendall, E. K. McCreary, P. G. Provance, M. M. Rodgers and W. A. Romani, *Muscles: Testing and Function with Posture and Pain*, 5th ed. Barueri: Manole, 2007.

[2] A. Quattrocchi *et al*., "Handgrip Strength in Health Applications: A Review of the Literature," Sensors, 2024. doi: 10.3390/s24165100.

[3] J. P. A. S. Almeida, M. Bessa, L. T. P. Lopes, A. Gonçalves, L. Roever and H. R. Zanetti, "Isometric handgrip exercise training reduces resting systolic blood pressure but does not interfere with diastolic blood pressure and heart rate variability in hypertensive subjects," *Hypertension Research*, vol. 44, no. 9, pp. 1205-1212, 2021, doi: 10.1038/s41440-021-00681-7.

[4] A. P. Shiratori, R. da R. Iop, N. G. Borges Júnior, S. C. Domenech, and M. da S. Gevaerd, "Evaluation protocols of hand grip strength in individuals with rheumatoid arthritis: a systematic review," *Revista Brasileira de Reumatologia*, vol. 54, no. 2, pp. 140–147, 2014, doi: 10.1016/j.rbre.2014.03.009.

[5] M. Ponce, S. Ceballos, I.-H. Khoo, V. Krishnan and P. Marayong, "Enhancing Hand Grip Force Control Through Training with Haptic Feedback," *2025 IEEE i-CREATe*, doi: 10.1109/i-CREATe67590.2025.11478219.

[6] D. Duanmu, T. Xu, X. Li, X. Cao, W. Huang and Y. Hu, "Perceptual Feedback through Multisensory Fusion in Hand Function Rehabilitation by A Machine Learning Approach," *2024 IEEE CIVEMSA*, doi: 10.1109/CIVEMSA58715.2024.10586653.

[7] S. M. Biju and H. Z. Sheikh, "Sensor evaluation for hand grip strength," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 5, pp. 4756-4764, 2022. doi: 10.11591/ijece.v12i5.pp4756-4764

[8] B.-Y. Huang, S.-W. Li and Y.-C. Lin, "Validation of Measurement Stability in a Smart Handgrip Device: Comparison with the Jamar Hydraulic Hand Dynamometer," *2025 IEEE Global Conference on Consumer Electronics (GCCE)*, 2025. doi: 10.1109/GCCE65946.2025.11275418

[9] Z. Namjoshi and V. Vishwakarma, "IoT-based Physiotherapy Device for Quantifying Hand Condition," *2025 International Conference on Modern Sustainable Systems (CMSS)*, 2025, doi: 10.1109/CMSS66566.2025.11182467.

[10] H. Hoffman *et al*., "Rehabilitation of hand function after spinal cord injury using a novel handgrip device: a pilot study," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, p. 22, 2017.

[11] G. Kurillo, M. Gregorič, N. Goljar, and T. Bajd, "Grip force tracking system for assessment and rehabilitation of hand function," *Technology and Health Care*, vol. 13, no. 3, pp. 137–149, 2005.

[12] Interlink Electronics, "FSR Integration Guide and Evaluation Parts Catalog," 2013. [Online]. Available: https://cdn.sparkfun.com/assets/e/3/b/3/8/force_sensing_resistor_guide.pdf

[13] Interlink Electronics, "FSR 400 Series Datasheet," 2017. [Online]. Available: https://cdn2.hubspot.net/hubfs/3899023/Interlinkelectronics%20November2017/Docs/Datasheet_FSR.pdf

[14] P. J. Millar, M. J. MacDonald, and N. McCartney, "Effects of isometric handgrip protocol on blood pressure and neurocardiac modulation," *International Journal of Sports Medicine*, vol. 32, no. 3, pp. 174–180, 2011, doi: 10.1055/s-0030-1268473.

[15] I. T. Anggraini, Y. Xiao, N. Funabiki, J.-Y. Liu, C.-P. Fan and W.-C. Kao, "An Extension of Handgrip Exergame System for Hypertension Improvement with Two-Hand Gestures," *2024 IEEE Consumer Life Tech (ICLT)*, 2024. doi: 10.1109/ICLT63507.2024.11038653.

Development of a Computer Game to Aid Children with Autism Spectrum Disorder and Down Syndrome Learn to Cycle

1st Bodouowei Godswill Zipamone
Dept. of Electrical Engineering
Universidade Estadual Paulista
Ilha Solteira, Brazil
b.zipamone@unesp.br

1st Professor Marcelo Augusto A. Sanches
Dept. of Electrical Engineering
Universidade Estadual Paulista Ilha Solteira, Brazil
marcelo.sanches@unesp.br

2nd Rafael José Ginuino da Silva
Dept. of Electrical Engineering
Universidade Estadual Paulista
Ilha Solteira, Brazil
rafael.ginuino@unesp.br

2nd Leonardo Xavier do Nascimento
Dept. of Electrical Engineering
Universidade Estadual Paulista
Ilha Solteira, Brazil
lx.nascinmento@unesp.br

2nd Sara Moreina Macedo
Dept. of Electrical Engineering
Universidade Estadual Paulista
Ilha Solteira, Brazil
sara.macedo@unesp.br

***Abstract*—Down syndrome (DS) and autism spectrum disorder (ASD) are genetic disorders that occur at infancy and could last for a life-time. Affected people have difficulty in performing motor activities such as engaging in communications, social interactions and sporting. Cycling is one such activity that DS and ASD children are slow in learning. Researchers in the literature had developed special techniques in aiding DS and ASD children learn to cycle. One such technique is the use of video games to enable ASD children learn and reduce weights. The research investigates the usefulness of computer games in aiding DS and ASD children in cycling. It is an area that has not received its required attention in the literature. Can cycling computer games reduce the learning curve of cycling by DS and ASD children? A computer game was developed to achieve the aim of the research. Results show that most ASD children performed above average, and that cycling computer games could reduce the learning curve of cycling by DS and ASD children.**

Keywords— cycling computer games, down syndrome, autism spectrum disorder, animation application, virtual reality, responsive app design

I. Introduction

Down Syndrome (DS) is a genetic condition that occurs when an infant has excess of chromosome 21 at birth. Some of the characteristics exhibited are retarded intellectual ability, pale facial features and feeble muscles (hypotonia) [1]. Autism Spectrum Disorder (ASD) is also a genetic disorder syndrome. It is a neuro-disorder that affects the social behaviour of people having it. People having ASD and DS exhibit restricted or repetitive behavioural patterns and interests. They are sometimes unable to effectively use body languages for communicating with others [2]. The rest of the work will refer to DS and ASD people as either ASD children or ASD people based on the context of use.

ASD people have difficulty in participating in sporting activities and social interactions. Due to the inherent dysfunctional traits that spontaneously exhibit themselves during physical activities such as sporting and social interactions, they sometimes shy away from socializing with people. As a consequence, they require motivation to learn and master an activity or engage in social interactions [2]. Participating in sporting activities and social interactions could improve the overall well-being of a person, in particular, they could improve the mood of ASD people and ignite their interests in socializing with people [3].

Bicycle Cycling is one such physical activity that is difficult for ASD children to learn. It is difficult for them to position themselves (hands and legs) on a bicycle and ride. Most times, instead of riding forward, the wheels roll backward, and it discourages them from riding. Researches such as [4], [5], [6] respectively had created various techniques on cycling to enable ASD children learn to cycle at their own pace. Findings from these articles show that with specialized techniques and lots of patience on the side of the trainers, they could learn to cycle faster than usual.

Furthermore, the use of computer software such as virtual reality and video games to encourage ASD children to learn and reduce weight had also been researched by [7], [8], [9], [10]. Findings from these researches show that encouraging them to play video games could improve their mood and enable them to learn and reduce weights.

The research investigates the usefulness of computer games in enabling ASD children learn to cycle. It tries to

answer the question: can cycling computer games reduce the learning curve of cycling by ASD children, or can cycling computer games improve the learning ability of ASD children in cycling? To investigate the question, a computer game was developed to enable them learn to cycle. In the game, an animated bicycle on-screen guides them as they ride a static physical bicycle. The animation is controlled by the physical bicycle. The game encourages them to continue to ride by counting the number of times they have won by reaching a certain level.

For the purpose of the research, a cycle is defined as a 360 degree turn of a bicycle pedal.

A round is defined as a completion of a game session. A round could contain any number of cycles as set by a player.

II. Game Interface Design

The interface design of the game follows the responsive app design approach [11]. The game is designed to be viewable on varied device screen sizes: televisions, computers and smartphones. This enables the user to install the game on his/her preferred device(s).

The smartphone view was designed to be viewable on virtual reality (VR) glasses. Players could place a smartphone in a VR headset and wear it to have an immersive experience while cycling. The remote control of the VR headset could be used to control the game. This could improve the mood of the player. The smartphone view is shown in fig. 2(a) while the computer view is shown in fig. 2(b) respectively.

A computer bicycle was designed to mimic a real-life bicycle. Also, a dinosaur was designed to mimic a player, as shown on fig 2(a) and (b) respectively.

An 'okay' emoji was also designed as shown in fig. 1(a). It displays on screen to show a player that the bicycle is rolling forward (clockwise) and that he/she is doing fine.

A 'warning' emoji was also designed as shown in fig 1(b). It displays on screen to show that the bicycle is cycling backward (anticlockwise), and that a player needs to cycle forward.

A counter interface was also designed as shown in fig. 1(c). It displays the number of cycles completed by a player. This encourages a player to continue cycling.

A 'winner' emoji was also designed as shown in fig. 1(d). it is displayed when a player reaches a winning level of cycling. A winning level depends on the maximum number of cycles set by a player.



Fig 1(a). Okay emoji



(b). Warning emoji



(c). Cycles counter



(d). Winner message

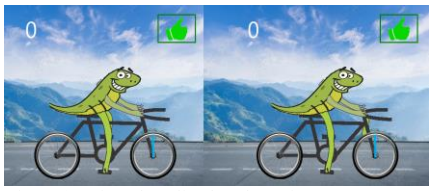


Fig 2(a). Virtual reality headset view



2(b). Computers and televisions' view

III. Circuitry and Casing Design

A circuit in package was designed to sense the angular position of the petal while a player is cycling. The BNO055 sensor was used to achieve the sensing of the angular position. The ESP32 microcontroller was used to retrieve the angle from the BNO055 sensor. The angle is then transmitted by the microcontroller via WIFI to a device where the game app is installed in order to animate the on-screen dinosaur and bicycle.

A. Circuitry Components

- **ESP32:** It is a microcontroller that is used for the programming of electronic components. It has both WIFI and Bluetooth technology that enables wireless connectivity and could be used to implement Internet of Things (IOT) applications [12].
- **BNO 055 Sensor:** It is a System in Package (SIP) Inertial Measurement Unit (IMU) sensor. It has a triaxial 14-bit accelerometer, a close-loop triaxial 16-bit gyroscope and a triaxial geo-magnetometer [13]. It was used to sense the pedal rotation angles of a bicycle.
- **Printed Circuit Board (PCB):** A PCB was used to connect various circuit components together to form an integrated circuit.
- **Battery:** A battery powered the circuitry.
- **Connector:** A two pin KK connector was used to connect the battery to the PCB.
- **Voltage Regulator:** It regulated the battery voltage to avoid overcharge.
- **Resistors:** a number of resistors were used to control the voltage entering a particular component.
- **Wires:** Jumper wires were used to connect the battery to the connector pins. They enable an easy removal of the battery for recharging.

The circuitry is shown in fig. 3.

B. Casing Design

A plastic casing was designed to safely package the circuitry and attach it to the pedal of a bicycle. It contains a battery section for holing a battery and a section for holding the circuitry. A metallic handle attaches the casing to the pedal of a bicycle. The casing could be attached to either the left or right pedal of a bicycle in order to read the pedal angle. The casing and the metallic handle are shown in fig. 4 (a), and the casing attached to a bicycle is shown in fig. 4 (b) respectively.

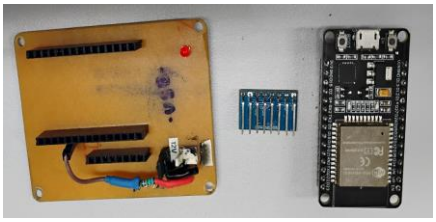


Fig 3. Circuitry

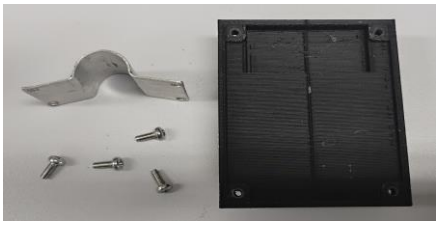


Fig 4 (a). Fabricated casing and handle



Fig 4 (b). Circuitry mounted on bicycle

IV. METHODOLOGY

When a player rides the physical bicycle, the ESP32 microcontroller gets the BNO055 sensor angle and sends it via WIFI to a device in real time. The game application on the device then retrieves the sensor angle and uses it to animate the legs of the on-screen dinosaur and the pedals and wheels of the on-screen bicycle to the angle position. Depending on the angle received, the on-screen dinosaur (legs) and bicycle (pedals and wheels) could either move clockwise (forward) or anticlockwise (backward) to the position of the angle. The 'okay' emoji is displayed for a successful clockwise movement and the 'warning' emoji is displayed for an anticlockwise movement.

In combination with the 'warning' emoji, a 'warning' sound is also played for a short period to notify a player that he/she is cycling backward. In combination with the 'okay' emoji, an 'okay' sound is also played for a short period to notify a player that he/she is back on track from riding backward.

Also, a count of number of successful cycles completed is displayed on-screen. The 'Winner' message is displayed when a player successfully completes a maximum number of cycles. A 'winner' sound is also played to indicate that a player has won. These gestures encourage a player to continue playing the game.

The microcontroller was programmed with Arduino sketch while the game app was developed with Qt C++.

A. Game Rules

- By default, maximum number of cycles (MC) is set to 10.
- A player can change the default MC and set his/her MC per round of play
- Backward cycles (BC) are wrong, and are subtracted from MC
- Forward cycles (FC) are counted for a player as successful cycles. It is calculated using equation 1
- A round ends when MC is reached
- Percentage Progress (*PP*) is measured as a percentage of FC over MC per round of play. It is calculated using equation 2.
- A player can play as many rounds in order to improve his/her cycling

$$FC = MC - BC \quad (1)$$

$$PP = \frac{FC}{MC}(100) \quad (2)$$

B. Game Diagram

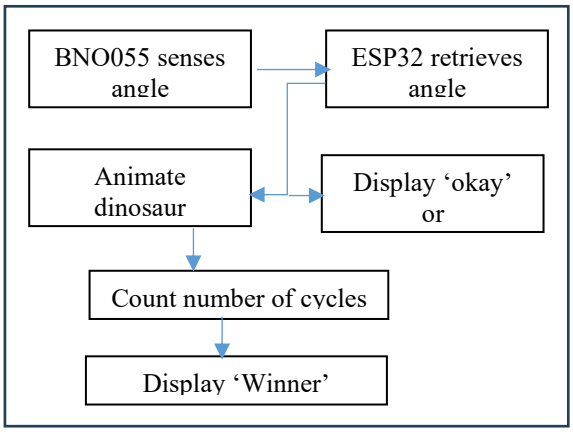


Fig. 5, Block diagram

C. Algorithm

1. Start game
2. While player is cycling
    a. Get angle from BNO055 sensor
    b. Animate on-screen dinosaur and bicycle to angle position
    c. Display 'okay' or 'warning' emoji based on angle
    d. increase number of cycles if a cycle is completed
    e. increase number of backward cycles if current angle is less than previous angle
    f. display 'Winner' message if maximum number of cycles is reached and go to step 4
3. go to step 2
4. stop

V. RESULTS AND DISCUSSION

A qualitative experiment was conducted on 20 ASD children between the age bracket of 6-12 years. The aim of the experiment was to determine the effectiveness of computer games at reducing the learning curve of cycling by ASD children.

The game was installed on both computers and smartphones. It was to check which device that is most effective in encouraging the children to play the game. The smartphone was put in a virtual reality headset and worn by a child to have an immersive experience while cycling.

The experiment was divided into pretraining and training sessions.

In the pretraining session, ASD children were introduced to the cycling game. They were taught to interact with the game by pressing the play button and cycling a physical bicycle.

After the game was introduced to the children, it was observed that they were excited to continue playing the game, and a training session was set. The aim of the training session was to observe the progress of the children as they played the game.

Ten consecutive rounds of 50 maximum cycles (MC) per round were achieved. In each round, the children were observed and forward cycles (FC) and backward cycles (BC) were recorded. The app automatically calculates the progress of a player by counting number of backward and forward cycles. Backward cycling meant wrong cycling and were subtracted from the overall 50 maximum cycles to get successful forward cycles. The forward cycles' function is illustrated in equation 1. The app also automatically calculates the percentage progress (*PP*) of a player in each round. The *PP* is shown in equation 2.

Table 1 shows the recorded observations as each child progressed during the training session. Row 1-20 represent the children while column 1-10 represent the forward cycles achieved by individuals during each training round.

Cumulative percentage progresses (CPP) were also calculated to determine the overall progress of each child during the ten sessions conducted. The C*PP* is illustrated in equation 3.

$$CPP = \frac{\sum_1^{10} FC}{\sum_1^{10} MC}(100) \quad (3)$$

Based on observations and the recorded data, it was seen that all children performed above average. Hence, it could be asserted that cycling computer games improve the mood of ASD children and reduces their training curve of cycling.

Table 1: cycling data

S/N	1	2	3	4	5	6	7	8	9	10	$\sum FC$	$\sum MC$	$CPP(\%)$
1	27	24	31	32	34	35	34	30	33	38	318	500	63.6
2	23	23	34	32	32	23	31	34	34	36	302	500	60.4
3	23	25	24	31	34	31	33	32	36	36	305	500	61
4	21	27	24	32	30	32	31	29	35	41	302	500	60.4
5	21	27	31	32	34	34	36	28	36	35	314	500	62.8
6	27	26	36	34	28	31	36	28	34	39	319	500	63.8
7	20	31	31	34	32	36	36	38	32	37	327	500	65.4
8	24	24	32	27	29	34	32	27	36	38	303	500	60.6
9	31	34	24	21	31	35	33	26	37	34	306	500	61.2
10	26	29	30	33	34	36	28	31	34	37	318	500	63.6
11	21	24	21	35	25	33	30	30	35	37	291	500	58.2
12	20	26	30	24	29	35	34	31	32	32	293	500	58.6
13	24	28	32	33	36	31	33	36	37	34	324	500	64.8
14	20	17	15	26	26	30	34	36	33	33	270	500	54

15	24	31	28	31	31	34	38	30	38	35	320	500	64
16	23	22	23	22	25	31	27	31	36	34	274	500	54.8
17	21	26	26	33	34	36	34	31	36	38	315	500	63
18	25	28	31	34	29	34	37	38	32	37	325	500	65
19	22	26	28	33	34	36	37	38	34	36	324	500	64.8
20	24	27	31	25	28	27	28	31	36	30	287	500	57.4

VI. CONCLUSION

The research investigated the usefulness of computer games in reducing the learning curve of ASD children in cycling. The literature indicates that video games had been developed to enable ASD children learn and lose weight, and that the efforts were averagely successful. A computer game was developed to aid the investigation. A circuitry in package was also designed and attached to a physical bicycle in order to read the angles of a pedal. Angles were read in real time as a player is cycling a physical bicycle. The read angle is then sent via WIFI to the game app in order to animate the on-screen dinosaur and bicycle. Twenty (20) ASD children participated in a training session that investigated the successful use of the game. It was observed that all children performed above average. Findings show that computer games improve the mood of ASD children and reduces their learning curve in cycling.

REFERENCES

[1] "Down syndrome." [Online]. Available: https://medlineplus.gov/genetics/

[2] T. Todd, G. Reid, and L. Butler-Kisber, "Cycling for Students With ASD: Self-Regulation Promotes Sustained Physical Activity," 2010.

[3] M. MacDonald *et al.*, "Bicycle training for youth with Down syndrome and autism spectrum disorders," *Focus Autism Other Dev. Disabl.*, vol. 27, no. 1, pp. 12–21, Mar. 2012, doi: 10.1177/1088357611428333.

[4] B. Kavlak, Z. Başaran, S. Çolak, M. Kılınç, and O. Arslan, "Investigate the effect of bicycle education provided for children with autism on recognising bicycle and cycling skill," *SHS Web of Conferences*, vol. 48, p. 01048, 2018, doi: 10.1051/shsconf/20184801048.

[5] A. Castillo Cañiz, "13 th FIEP European Congress 29 th FIEP World Congress Istanbul," 2018.

[6] A. Castillo Cañiz, "13 th FIEP European Congress 29 th FIEP World Congress Istanbul," 2018.

[7] S. Mahini, "Virtual Reality as a Tool for Children with Autism Spectrum Disorder," *Digital Technologies Research and Applications*, vol. 3, no. 1, pp. 16–23, Feb. 2024, doi: 10.54963/dtra.v3i1.211.

[8] A. Soltiyeva, W. Oliveira, A. Madina, S. Adilkhan, M. Urmanov, and J. Hamari, "My Lovely Granny's Farm: An immersive virtual reality training system for children with autism spectrum disorder," *Educ. Inf. Technol. (Dordr).*, vol. 28, no. 12, pp. 16887–16907, Dec. 2023, doi: 10.1007/s10639-023-11862-x.

[9] T. C. N. Graham, N. King, H. Coo, P. Zabojnikova, B. J. Gurd, and D. Samdup, "Design and Evaluation of an Exergaming System for Children With Autism Spectrum Disorder: The Children's and Families' Perspective," *Front. Virtual Real.*, vol. 3, Apr. 2022, doi: 10.3389/frvir.2022.817303.

[10] F. Hassani, S. Shahrbanian, S. H. Shahidi, and M. Sheikh, "Playing games can improve physical performance in children with autism," *Int. J. Dev. Disabil.*, vol. 68, no. 2, pp. 219–226, 2022, doi: 10.1080/20473869.2020.1752995.

[11] Noorkaran Bhanarkar, Aditi Paul, and Dr. Ashima Mehta, "Responsive Web Design and Its Impact on User Experience," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 50–55, Apr. 2023, doi: 10.48175/ijarsct-9259.

[12] "ESP32 Series Datasheet Version 4.7 2.4 GHz Wi-Fi + Bluetooth ® + Bluetooth LE SoC Including." [Online]. Available: www.espressif.com

[13] "BNO055 Intelligent 9-axis absolute orientation sensor," 2014.

Longitudinal Changes in Spatiotemporal Gait Parameters During Serious Game-Assisted Self-Paced Treadmill Training in Older Adults

1st Brayan Arevalo
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
brayan.arevalo@edu.ufes.br
ORCID: 0000-0001-8346-0464

2nd Meilen Salamanca
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
meilen.jimenez@edu.ufes.br
ORCID: 0009-0003-6954-8648

3rd Carla Zimerer
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
czimerer@gmail.com
ORCID: 0000-0003-2872-5209

4th Maria Eduarda Franchine
Center of Physical Education and Sports
Federal University of Espírito Santo
Vitória, ES, Brazil
maria.f.machado@edu.ufes.br
ORCID: 0009-0002-2324-8001

5th Ricardo Mello
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
ricardo.c.mello@ufes.br
ORCID: 0000-0003-0420-4273

6th Anselmo Frizera
Graduate Program in Electrical Engineering
Federal University of Espírito Santo
Vitória, ES, Brazil
anselmo.frizera-neto@ufes.br
ORCID: 0000-0002-0687-3967

***Abstract*—Population aging is associated with declines in gait quality and functional mobility, motivating repeated and monitored walking practice in older adults. This descriptive longitudinal study characterized spatiotemporal gait behavior, functional mobility, and perceived usability during three months of supervised Serious Game-assisted self-paced treadmill training. Five sedentary older adults (four women and one man; mean age, 70.2 ± 3.0 years) completed 36 supervised sessions. Treadmill-derived walking speed, stride length, and stride frequency were summarized using participant-level medians from a predefined 30–90 s analysis window at baseline (Pre), monthly during training, and at three months (3M). Mean walking speed increased from 0.63 to 1.31 m/s, mean stride length from 0.64 to 0.93 m, and mean stride frequency from 1.01 to 1.42 strides/s. All participants showed higher values at 3M than at Pre, although the magnitude and spatiotemporal composition of change varied across participants. Mean instrumented Timed Up and Go duration decreased from 10.68 to 9.18 s, with shorter durations at 3M in all participants. Mean System Usability Scale score increased from 69.5 after Session 1 to 87.7 after Session 36, although participant-level SUS scores did not increase monotonically across the four administrations. These findings characterize participant-specific longitudinal gait, functional-mobility, and usability patterns within the integrated self-paced treadmill–Serious Game system. The descriptive design, small sample, absence of a control condition, and lack of direct overground assessment limit causal interpretation and overground generalizability.**

***Index Terms*—self-paced treadmill, walking speed, stride length, stride frequency, gait analysis**

This work was supported by the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES; grants 2025-QH0K, 2025-J7XD, and 2026-48WNC), the Brazilian National Council for Scientific and Technological Development (CNPq; grants 308155/2023-8 and 402080/2025-4), and the CNPq/FAPES joint program (grant 2025-6VCJ4).

I. Introduction

Population aging is associated with declines in mobility, gait quality, and functional independence, partly related to increasingly sedentary behavior and reduced habitual physical activity [1]–[3]. These changes are linked to an elevated risk of falls and to progressive declines in the strength, balance, and locomotor capacity required for daily activities [4]–[6]. Supervised exercise and gait-focused interventions are widely recommended to mitigate these declines [7], [8], highlighting the need for technologies that support repeated, monitored walking practice among older adults over extended periods.

Self-paced treadmills allow users to regulate their walking speed while preserving the controlled setting of treadmill-based protocols. Unlike fixed-speed systems, these platforms operate through a closed-loop interaction in which treadmill-belt speed is continuously adjusted in response to the user's behavior, thereby better accommodating self-selected locomotor patterns during repeated exposure [9].

Walking speed is often used as a global descriptor of locomotor performance, yet speed alone does not indicate how a given gait pattern is produced. Similar speeds may arise from different combinations of stride-length-related spatial behavior and stride-frequency-related temporal behavior. When the goal is to interpret change over repeated exposure, these descriptors become important because they help determine whether longitudinal speed increases are associated primarily with spatial adjustments, temporal adjustments, or a combination of both. Prior work has shown that spatiotemporal gait parameters vary

systematically with walking condition and speed, supporting the joint interpretation of speed and stride-related descriptors rather than speed alone [10], [11].

Overground and treadmill walking represent distinct assessment contexts. Treadmill walking enables repeated measurements over many gait cycles under controlled conditions, but it should not be considered directly equivalent to overground walking. Differences between treadmill and overground conditions have been reported in kinematics, kinetics, spatiotemporal variability, and sensory context [12], [13]. Self-paced control reduces an important limitation of fixed-speed treadmill walking by allowing participants to regulate their walking speed. Nevertheless, because treadmill-belt speed is continuously adjusted in response to user behavior, the participant and treadmill operate as a closed-loop system [14], [15]. Consequently, longitudinal changes may reflect both locomotor behavior and increasing familiarity with the self-paced control interface.

Despite growing interest in self-paced treadmill systems, the available literature remains focused primarily on controller validation and cross-sectional comparisons among overground, fixed-speed treadmill, and self-paced treadmill walking [14], [15]. Less is known about how gait behavior evolves during repeated supervised exposure to self-paced treadmills over longer periods in older adults. In particular, it remains unclear whether longitudinal changes in walking speed are associated mainly with spatial adjustments, temporal adjustments, or participant-specific combinations of both.

Beyond spatiotemporal gait parameters, the Timed Up and Go (TUG) test captures sit-to-stand transitions, walking, turning, and overall task duration, providing a complementary descriptor of functional mobility in older adults [16], [17]. Whether longitudinal changes in treadmill-derived gait descriptors coexist with changes in functional mobility over the same period remains unclear.

To address the identified gap, the present study examined repeated Serious Game-assisted self-paced treadmill walking in older adults [18]. The primary aim was to characterize participant-specific longitudinal patterns in treadmill-derived walking speed, stride length, and stride frequency across the three-month training period, from baseline to each monthly assessment, including differences in the spatial and temporal composition of change. The secondary aim was to describe concurrent changes in functional mobility over the same period. As a tertiary, exploratory aim, perceived usability of the integrated treadmill–Serious Game system was described using the System Usability Scale (SUS) across the training period.

II. Materials and Methods

A. Assistive Technology System

The training protocol used an integrated assistive-technology platform that combined self-paced locomotion on the UFES SmartTreadmill [9] with Serious Game-based visual guidance, concurrent upper-limb interaction, stereoscopic visualization, and physiological monitoring. The UFES Smart-Treadmill consisted of a commercial treadmill retrofitted with a closed-loop treadmill-belt-speed controller. An RPLiDAR A3 laser range finder tracked the participant's lower-limb position, and the resulting signal was processed using the Weighted-Frequency Fourier Linear Combiner and Fourier Linear Combiner methods to estimate walking speed in real time [9]. An E6B2-CWZ6C incremental encoder independently measured treadmill-belt speed for feedback control. Thus, the walking-speed estimate and treadmill-belt speed were obtained through separate sensing paths. The sensing architecture and associated self-paced control methods were previously developed and evaluated [9], [18]. Figure 1 illustrates the physical configuration of the platform.

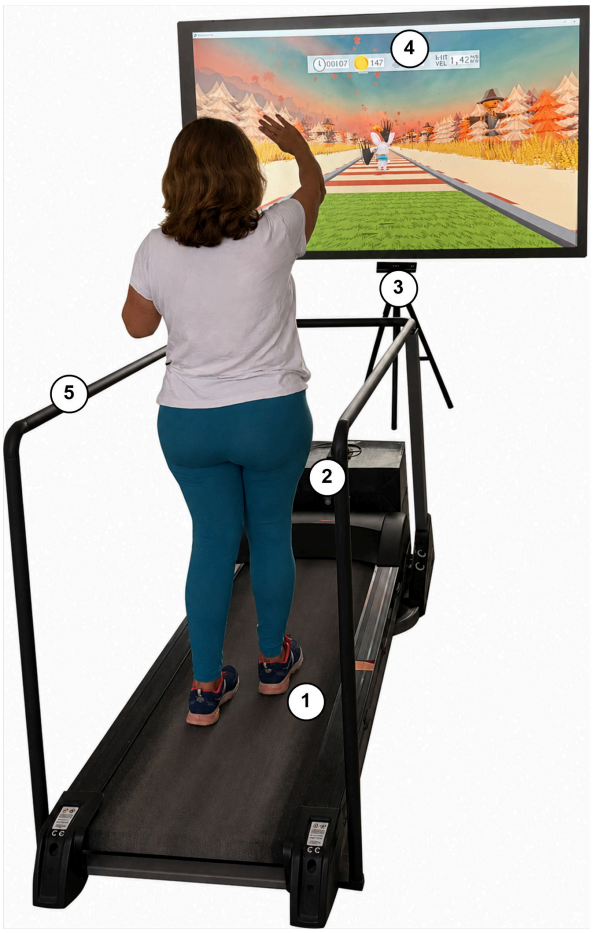


Fig. 1. Experimental setup of the integrated assistive-technology platform: (1) self-paced treadmill walking surface, (2) RPLiDAR A3 laser range finder, (3) Kinect V2 sensor, (4) Serious Game visual-feedback display, and (5) safety handrails.

The Serious Game was implemented in Unity and exchanged locomotion and control information with the treadmill controller in real time through a Robot Operating System-based architecture and WebSocket communication. This integrated treadmill–Serious Game application was previously

described [18]. A Microsoft Kinect V2 sensor provided markerless tracking for the concurrent upper-limb interaction task but did not contribute to gait sensing. A passive stereoscopic projection system displayed the virtual environment during treadmill walking. Heart rate (HR) was continuously monitored using a Polar H10 as a real-time safety measure and was not used as an input to the self-paced controller.

B. Serious Game Training Tasks

Training sessions took place within a virtual environment that changed across the three-month protocol. In the treadmill-integrated task, avatar progression was coupled to the participant's walking speed, and participants regulated their pace to remain within a visual target walking zone, receiving continuous feedback on their position relative to this zone. In the Kinect-integrated task, markerless upper-limb movements were used to collect target stimuli and avoid distractors while the treadmill-integrated task was performed concurrently. The simultaneous execution of both tasks constituted the motor-cognitive dual-task condition implemented by the Serious Game. Every 5 min, the Dynamic Difficulty Adjustment (DDA) mechanism independently adjusted the upper speed limit of the target walking zone and the number of target stimuli for the subsequent interval according to task performance. This upper speed limit also constrained the maximum treadmill-belt speed for the subsequent interval, while the self-paced controller continued to regulate belt speed in response to participant motion within that limit.

C. Participant Information

Participants were recruited by convenience sampling from a university program for older adults at the Federal University of Espírito Santo. The inclusion criteria were: age 60–80 years; cardiologist clearance for exercise participation; independent ambulation without an assistive device; low physical activity level according to the International Physical Activity Questionnaire (IPAQ) [19]; body mass index between 18.5 and 29.9 kg/m^2 [20]; nonsmoking status; no structured exercise during the preceding three months; no previous virtual-reality experience; and no cognitive impairment, vestibular disorder, or monocular vision, with cognitive status assessed using the Mini-Mental State Examination [21]. A minimum attendance of 80% of the scheduled training sessions was required for inclusion in the longitudinal analysis.

Of 68 older adults who contacted the research team, 47 expressed interest in participating. Of these, 42 did not meet the eligibility criteria, and the remaining five were enrolled in the study. The enrolled cohort comprised five sedentary older adults (P1–P5; four women and one man), with a mean age of 70.20 ± 3.03 years, height of 1.57 ± 0.05 m, body mass of 63.20 ± 4.55 kg, and body mass index of 25.76 ± 1.18 kg/m^2. All participants walked independently without assistive devices and completed the 36 scheduled sessions, with no attrition.

D. Study Design and Supervised Training Protocol

This descriptive, exploratory longitudinal study was conducted at the Human-Centered Systems Laboratory of the Federal University of Espírito Santo. The protocol was approved by the Research Ethics Committee of the Federal University of Espírito Santo (approval no. 6,294,101; CAAE: 70158523.2.0000.5542), and all participants provided written informed consent before enrollment.

Eligibility screening and participant characterization were completed together with a familiarization session introducing the treadmill, visualization system, and interaction procedures, during an initial laboratory visit. The pre-intervention assessment (Pre) was conducted during the following laboratory visit. Data collected during familiarization were not included in the longitudinal analysis.

The supervised training phase comprised 36 Serious Game-assisted sessions on the self-paced treadmill, scheduled three times per week. Session duration was 20, 25, and 30 min in months 1, 2, and 3, respectively. HR and symptom status were monitored throughout each training session under continuous supervision, whereas resting blood pressure was assessed during the protocol-defined safety screening. Training was not initiated or was interrupted in response to clinically relevant symptom worsening (e.g., dizziness, dyspnea, chest pain, or excessive fatigue), postural instability, abnormal HR or resting blood-pressure findings, or participant request.

Longitudinal assessments were conducted in dedicated sessions separate from the supervised training sessions, before the training phase (Pre) and after the first (1M), second (2M), and third (3M) months. During each assessment, participants walked on the self-paced treadmill at a comfortable self-selected pace, and the corresponding gait recordings were

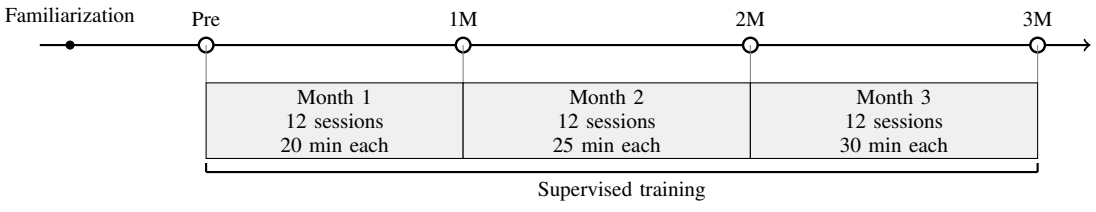


Fig. 2. Longitudinal schedule relevant to the present gait analysis. A familiarization visit, which also included eligibility screening and participant characterization, preceded the baseline assessment (Pre). Supervised training comprised 36 Serious Game-assisted sessions, scheduled three times per week and organized into three monthly blocks of 12 sessions. Session duration was 20, 25, and 30 min in months 1, 2, and 3, respectively. Longitudinal assessments (open circles) were conducted at Pre and after each month (1M, 2M, and 3M).

retained for the present analysis. The study schedule is summarized in Fig. 2.

E. Data Analysis

Gait data collected at Pre, 1M, 2M, and 3M were processed using the previously described treadmill-processing procedure [9] and analyzed descriptively. This procedure provided walking speed, stride length, and stride frequency. The reported walking speed represented the participant's walking-speed estimate rather than treadmill-belt speed. All three variables were interpreted as treadmill-derived descriptors rather than direct measurements of overground gait.

At each assessment, the participant was the unit of longitudinal analysis. Samples from the predefined 30–90 s window were used to derive participant-level summaries after excluding the gait-initiation and gait-termination segments. Each gait variable was summarized as the median of the samples within that window, and group values were calculated as the mean and standard deviation (SD) of the five participant-level medians. Samples within an assessment window were not treated as independent longitudinal observations. Given the exploratory design and sample size, longitudinal interpretation focused on participant-level trajectories across all four assessments.

To address the secondary aim, functional mobility was assessed using an instrumented Timed Up and Go (TUG) test [16], [17]. The assessment was performed using a Baiobit wearable inertial sensor system (BTS Bioengineering, Garbagnate Milanese, Italy), positioned over the lower trunk/pelvic region using the manufacturer-provided belt. Participants stood from a chair, walked 3 m, turned, returned to the chair, and sat down. TUG total duration was obtained directly from the Baiobit report. For each assessment, the first recording classified as valid by the system was retained. Group TUG duration was presented as mean (SD).

As an exploratory measure, system usability was assessed using the Brazilian Portuguese version of the SUS after Sessions 1, 12, 24, and 36 [22]. Standard SUS scoring was applied: for odd-numbered items, one was subtracted from the response; for even-numbered items, the response was subtracted from five; and the adjusted item scores were summed and multiplied by 2.5, yielding a total score from 0 to 100.

Session 1 served as the initial usability reference. Sessions 12, 24, and 36 marked the ends of the three monthly training blocks and preceded the 1M, 2M, and 3M assessments, respectively. SUS administrations were separate from the gait and TUG assessments. Participant-level SUS change was calculated as the Session 36 score minus the Session 1 score, and group values were presented as mean (SD).

Pre-to-3M percentage changes were calculated as

$$\Delta x(\%) = \frac{x_{3\mathrm{M}} - x_{\mathrm{Pre}}}{x_{\mathrm{Pre}}} \times 100. \tag{1}$$

Participant-level changes were calculated from the corresponding gait medians and retained TUG values. Group changes were calculated as mean (SD) across participants. Changes in stride length and stride frequency were compared descriptively to characterize spatial, temporal, or combined patterns.

III. Results

Figure 3 shows participant-level values of treadmill-derived walking speed, stride length, and stride frequency across Pre, 1M, 2M, and 3M. Table I reports the corresponding participant-level medians and group summaries. Mean participant-level Pre-to-3M percentage changes were 123.9% (SD: 61.9%) for walking speed, 54.3% (SD: 32.2%) for stride length, and 45.0% (SD: 32.2%) for stride frequency. All participants had higher 3M than Pre values for all three

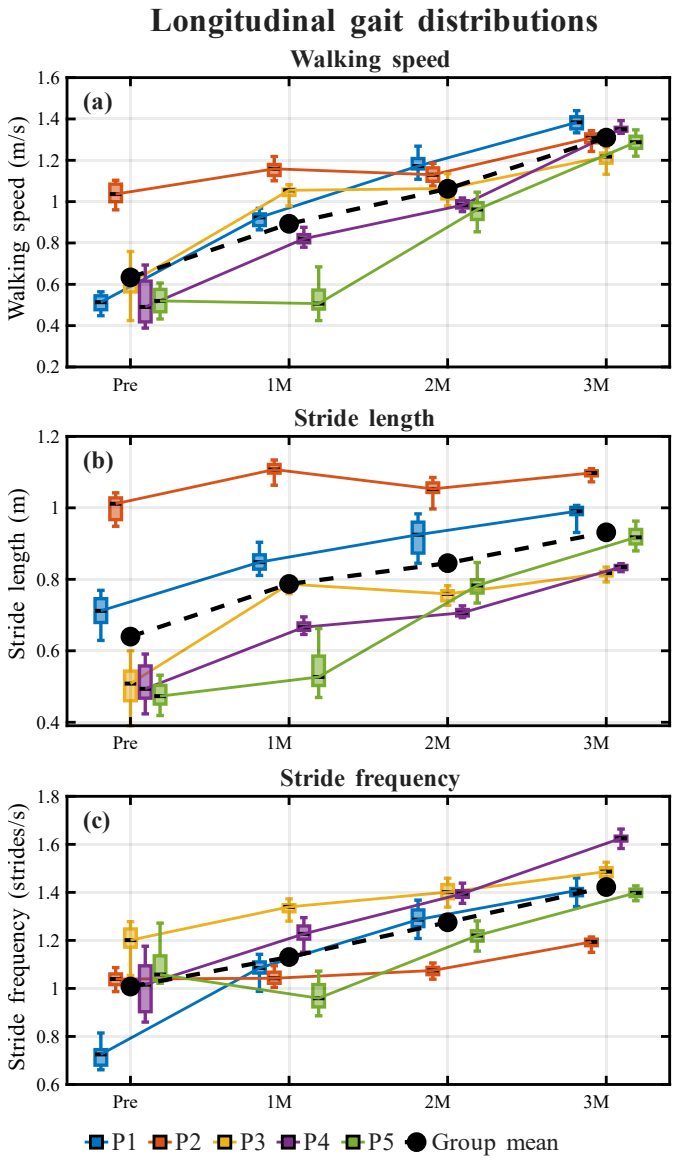


Fig. 3. Distributions of treadmill-derived walking speed (a), stride length (b), and stride frequency (c) at Pre, 1M, 2M, and 3M. For each participant and assessment, boxes show the interquartile range, center lines the median, and whiskers the 5th–95th percentiles of samples within the predefined window. Colored lines connect participant-level medians, and the black dashed line shows their arithmetic mean at each assessment. Participant-level values are reported in Table I.

TABLE I

PARTICIPANT-LEVEL MEDIANS OF TREADMILL-DERIVED WALKING SPEED, STRIDE LENGTH, AND STRIDE FREQUENCY AT PRE, 1M, 2M, AND 3M, WITH PRE-TO-3M PERCENTAGE CHANGES. VALUES ARE DISPLAYED TO TWO DECIMAL PLACES; PERCENTAGE CHANGES AND GROUP SUMMARIES WERE CALCULATED FROM UNROUNDED PARTICIPANT-LEVEL MEDIANS AND MAY THEREFORE NOT BE EXACTLY REPRODUCIBLE FROM THE PRINTED VALUES. BOTTOM-ROW ASSESSMENT VALUES ARE MEAN (SD) ACROSS PARTICIPANTS, WHEREAS BOTTOM-ROW PERCENTAGE CHANGES ARE MEAN (SD) OF PARTICIPANT-LEVEL CHANGES.

	Walking speed (m/s)					Stride length (m)					Stride frequency (strides/s)				
Participant	Pre	1M	2M	3M	Δv (%)	Pre	1M	2M	3M	ΔL (%)	Pre	1M	2M	3M	Δf_z (%)
P1	0.51	0.92	1.17	1.38	+169.50	0.71	0.85	0.92	0.99	+39.20	0.73	1.09	1.29	1.41	+94.10
P2	1.04	1.16	1.13	1.31	+26.30	1.01	1.11	1.05	1.10	+8.50	1.04	1.04	1.07	1.19	+14.90
P3	0.61	1.05	1.06	1.22	+100.70	0.51	0.79	0.76	0.82	+60.80	1.20	1.34	1.40	1.49	+23.80
P4	0.49	0.82	0.98	1.35	+175.20	0.49	0.67	0.71	0.83	+69.00	1.02	1.23	1.39	1.63	+60.00
P5	0.52	0.51	0.96	1.29	+147.70	0.47	0.53	0.78	0.92	+93.90	1.06	0.96	1.22	1.40	+32.20
Mean (SD)	0.63 (0.23)	0.89 (0.25)	1.06 (0.09)	1.31 (0.06)	+123.87 (61.93)	0.64 (0.23)	0.79 (0.22)	0.85 (0.14)	0.93 (0.12)	+54.29 (32.22)	1.01 (0.17)	1.13 (0.15)	1.28 (0.14)	1.42 (0.16)	+44.98 (32.24)

TABLE II

INSTRUMENTED TIMED UP AND GO (TUG) TOTAL DURATION AT PRE, 1M, 2M, AND 3M. NEGATIVE PERCENTAGE CHANGES INDICATE SHORTER TUG DURATION AT 3M THAN AT PRE. BOTTOM-ROW DURATIONS ARE MEAN (SD) ACROSS PARTICIPANTS, WHEREAS THE BOTTOM-ROW PERCENTAGE CHANGE IS THE MEAN (SD) OF PARTICIPANT-LEVEL CHANGES.

	Instrumented TUG total duration (s)				
Participant	Pre	1M	2M	3M	Δ Pre–3M (%)
P1	8.59	6.88	8.55	6.28	-26.90
P2	10.39	10.70	10.31	10.04	-3.40
P3	11.40	12.24	12.82	9.92	-13.00
P4	11.44	11.89	9.98	8.88	-22.40
P5	11.56	12.16	10.08	10.76	-6.90
Mean (SD)	10.68 (1.26)	10.77 (2.26)	10.35 (1.55)	9.18 (1.75)	-14.51 (9.99)

descriptors, although participant-level values did not increase monotonically across all four assessments. P2 showed the smallest changes in walking speed (+26.3%), stride length (+8.5%), and stride frequency (+14.9%), whereas the largest changes were observed in P4 for walking speed (+175.2%), P5 for stride length (+93.9%), and P1 for stride frequency (+94.1%).

The percentage changes in stride length and stride frequency differed across participants. P1 and P2 showed larger percentage increases in stride frequency than in stride length, whereas P3 and P5 showed larger percentage increases in stride length. P4 showed substantial increases in both components, with stride length increasing by 69.0% and stride frequency by 60.0%. Thus, the higher walking speeds observed at 3M were not accompanied by a uniform spatial or temporal pattern across participants.

Mean participant-level TUG duration decreased by 14.5% from Pre to 3M (Table II). All five participants had shorter durations at 3M, with reductions ranging from 3.4% in P2 to 26.9% in P1. Intermediate TUG durations did not decrease monotonically, with transient increases at 1M or 2M in some participants.

Mean SUS score increased by 18.2 points (SD: 15.7) from Session 1 to Session 36 (Table III), and all five participants had higher scores at Session 36 than at Session 1. P2 and P5 showed the smallest increases (+3.0 points each), whereas P4 showed the largest increase (+40.0 points). Participant-level perceived-usability changes were heterogeneous, and

TABLE III

SYSTEM USABILITY SCALE (SUS) SCORES AT FOUR TRAINING-SESSION LANDMARKS: THE INITIAL SESSION AND THE ENDPOINTS OF THE THREE 12-SESSION MONTHLY BLOCKS. Δ SUS DENOTES THE SESSION 36 MINUS SESSION 1 CHANGE. BOTTOM ROW: MEAN (SD) ACROSS PARTICIPANTS.

Participant	Session 1	Session 12	Session 24	Session 36	Δ SUS (points)
P1	65.0	80.0	88.0	85.0	+20.0
P2	75.0	73.0	80.0	78.0	+3.0
P3	72.5	75.0	90.0	97.5	+25.0
P4	60.0	85.0	100.0	100.0	+40.0
P5	75.0	75.0	80.0	78.0	+3.0
Mean (SD)	69.5 (6.7)	77.6 (4.9)	87.6 (8.3)	87.7 (10.5)	+18.2 (15.7)

SUS scores did not increase monotonically across the four administrations.

IV. Discussion

This study showed that treadmill-derived walking speed, stride length, and stride frequency increased from Pre to 3M in all five participants, although the magnitude and spatiotemporal composition of change differed markedly across participants. The concurrent reduction in TUG duration provided complementary functional-mobility context. Overall, the findings indicate participant-specific spatial and temporal organization rather than a uniform pattern of change during the supervised training period.

These findings reinforce the need to interpret walking speed together with stride-related spatial and temporal descriptors. Higher walking speeds can be achieved through longer strides, higher stride frequency, or combined changes in both components [10], [11]. Previous work involving older adults has similarly shown that faster walking may be supported by different combinations of step length and step time [23]. These findings were therefore consistent with participant-specific spatiotemporal patterns rather than with a single common organization for achieving a higher final walking speed.

For descriptive context, Hollman et al. reported sex- and age-stratified overground means of 1.12–1.22 m/s for gait speed, 1.18–1.39 m for stride length, and 102–114 steps/min (0.85–0.95 strides/s) for cadence in healthy adults aged 70–79 years [24]. These subgroup means are not clinical thresholds and are not directly equivalent to treadmill-derived measurements. In the present study, walking speed was below the lowest Hollman subgroup mean in all participants at Pre and at or above the highest subgroup mean in all participants at 3M. Stride length remained below the lowest subgroup mean at both Pre and 3M. Stride frequency was above the highest subgroup mean in four participants at Pre and in all participants at 3M. Thus, higher final treadmill-derived walking speeds co-occurred with shorter strides and higher stride frequencies than the published overground subgroup means. Together with the participant-level percentage-change patterns, these descriptive contrasts are consistent with heterogeneous spatial and temporal organization across participants [23]; they do not indicate gait normalization or treadmill–overground equivalence.

The participant-specific patterns should be interpreted within the closed-loop context of self-paced treadmill walking, in which treadmill-belt speed responds continuously to participant behavior [14], [15]. The platform also combined locomotion-related visual guidance, concurrent upper-limb interaction, and DDA within the dual-task Serious Game [18]. Because changes in DDA settings over time were not analyzed, whether task adjustment contributed to the observed gait heterogeneity cannot be determined.

The SUS results provided complementary context for the participant-specific gait changes during the same training period. Mean SUS scores increased from 69.5 points after Session 1 to 87.7 points after Session 36, with intermediate group means of 77.6 and 87.6 points after Sessions 12 and 24, respectively (Table III). At the participant level, P2 combined a small SUS increase (+3.0 points) with the smallest changes in all three gait descriptors, whereas P4 combined the largest SUS increase (+40.0 points) with substantial changes across the three gait descriptors, including the largest increase in walking speed. In contrast, P5 showed the same small SUS increase as P2 but the largest increase in stride length. These descriptive concordances and discordances should not be interpreted as evidence of an association or causal influence between perceived usability and the observed gait patterns. The SUS reflected overall perceived usability of the integrated treadmill–Serious Game system rather than isolated treadmill-control difficulty.

Mean TUG duration decreased from 10.68 s at Pre to 9.18 s at 3M, corresponding to a descriptive reduction of 14.5%. All participants had shorter durations at 3M, although intermediate TUG durations varied and Pre-to-3M reductions ranged from 3.4% in P2 to 26.9% in P1. Thus, the observed treadmill-derived gait changes co-occurred with shorter TUG durations in an overground functional task, but the descriptive design does not establish transfer or a training effect.

Several limitations must be considered. The sample included only five participants recruited by convenience; therefore, the study does not support population-level inference, and the findings should be interpreted as participant-specific observations within this cohort rather than as estimates applicable to the broader older-adult population. The restrictive eligibility criteria further limit representativeness, particularly for older adults with cognitive, vestibular, visual, or mobility impairments. No control or comparison condition was included. The descriptive design therefore does not allow the observed changes to be separated from familiarization with the self-paced controller, repeated task exposure, spontaneous variability, measurement variability, or other time-related influences. The gait variables were treadmill-derived descriptors obtained from a predefined 30–90 s analysis window and summarized using participant-level medians. They should not be interpreted as direct overground gait measurements; therefore, transfer of the observed spatiotemporal changes to overground walking cannot be established. In addition, the Pre-to-3M percent changes summarize the overall baseline-to-final contrast but do not fully represent the heterogeneous gait behavior observed at 1M and 2M, as shown in Fig. 3.

Future studies should include larger samples, control or comparison conditions, and direct overground assessments. Prespecified longitudinal models that preserve the participant-level data structure would help evaluate whether the observed findings generalize beyond this cohort and experimental environment.

V. Conclusion

Across the three-month supervised training period, all five participants had higher treadmill-derived walking speed, stride length, and stride frequency at 3M than at Pre, but the magnitude and spatiotemporal composition of change differed across participants. These differences describe heterogeneous

participant-specific locomotor patterns. Shorter TUG durations at 3M co-occurred with these treadmill-derived gait changes, and mean SUS score was higher at Session 36 than at Session 1, although participant-level SUS scores did not increase monotonically across the four administrations. Together, these findings characterize longitudinal patterns in gait, functional mobility, and usability within the integrated self-paced treadmill–Serious Game system.

Given the descriptive design, five-participant sample, absence of a control condition, and lack of direct overground gait assessment, causal interpretation is not supported and generalizability beyond this cohort and experimental environment is limited. Larger controlled studies incorporating direct overground assessments and prespecified longitudinal analyses are needed to evaluate the functional relevance and generalizability of these findings.

REFERENCES

[1] World Health Organization, "Ageing and health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, 2024, fact sheet, updated 2024.

[2] J. P. Thyfault, M. Du, W. E. Kraus, J. A. Levine, and F. W. Booth, "Physiology of sedentary behavior and its relationship to health outcomes," *Medicine & Science in Sports & Exercise*, vol. 47, no. 6, pp. 1301–1305, 2015.

[3] T. Strain, S. Flaxman, R. Guthold, E. Semenova, M. Cowan, L. M. Riley, F. C. Bull, G. A. Stevens, R. A. Raheem, K. Agoudavi, and WHO Physical Activity Country Profile Collaborating Group, "National, regional, and global trends in insufficient physical activity among adults from 2000 to 2022: a pooled analysis of 507 population-based surveys," *The Lancet Global Health*, vol. 12, no. 8, pp. e1232–e1243, 2024.

[4] W. Graafmans, M. Ooms, H. Hofstee, P. Bezemer, L. M. Bouter, and P. Lips, "Falls in the elderly: a prospective study of risk factors and risk profiles," *American journal of epidemiology*, vol. 143, no. 11, pp. 1129–1136, 1996.

[5] K. F. Reid, E. Pasha, G. Doros, D. J. Clark, C. Patten, E. M. Phillips, W. R. Frontera, and R. A. Fielding, "Longitudinal decline of lower extremity muscle power in healthy and mobility-limited older adults: influence of muscle mass, strength, composition, neuromuscular activation and single fiber contractile properties," *European journal of applied physiology*, vol. 114, no. 1, pp. 29–39, 2014.

[6] M. Montero-Odasso and for the Task Force, "World guidelines for falls prevention and management for older adults," *Age and Ageing*, vol. 51, no. 9, p. afac205, 2022.

[7] W. J. Chodzko-Zajko, D. N. Proctor, M. A. F. Singh, C. T. Minson, C. R. Nigg, G. J. Salem, and J. S. Skinner, "Exercise and physical activity for older adults," *Medicine & science in sports & exercise*, vol. 41, no. 7, pp. 1510–1530, 2009.

[8] E. Thomas, G. Battaglia, A. Patti, J. Brusa, V. Leonardi, A. Palma, and M. Bellafiore, "Physical activity programs for balance and fall prevention in elderly: A systematic review," *Medicine*, vol. 98, no. 27, p. e16218, 2019.

[9] B. S. Moreno, R. C. Mello, and A. Frizera, "Automation of user pace-adjusted treadmill: comparison of control methods based on LRF sensors," in *Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

[10] N. Sekiya, H. Nagasaki, H. Ito, and T. Furuna, "Optimal walking in terms of variability in step length," *Journal of Orthopaedic & Sports Physical Therapy*, vol. 26, no. 5, pp. 266–272, 1997.

[11] F. Danion, E. Varraine, M. Bonnard, and J. Pailhous, "Stride variability in human gait: the effect of stride frequency and stride length," *Gait & Posture*, vol. 18, no. 1, pp. 69–77, 2003.

[12] P. O. Riley, G. Paolini, U. Della Croce, K. W. Paylo, and D. C. Kerrigan, "A kinematic and kinetic comparison of overground and treadmill walking in healthy subjects," *Gait & Posture*, vol. 26, no. 1, pp. 17–24, 2007.

[13] J. H. Hollman, M. K. Watkins, A. C. Imhoff, C. E. Braun, K. A. Akervik, and D. K. Ness, "A comparison of variability in spatiotemporal gait parameters between treadmill and overground walking conditions," *Gait & Posture*, vol. 43, pp. 204–209, 2016.

[14] L. H. Sloot, M. M. van der Krogt, and J. Harlaar, "Self-paced versus fixed speed treadmill walking," *Gait & Posture*, vol. 39, no. 1, pp. 478–484, 2014.

[15] M. Plotnik, T. Azrad, M. Bondi, Y. Bahat, Y. Gimmon, G. Zeilig, R. Inzelberg, and I. Siev-Ner, "Self-selected gait speed - over ground versus self-paced treadmill walking, a solution for a paradox," *Journal of NeuroEngineering and Rehabilitation*, vol. 12, no. 1, p. 20, 2015.

[16] R. W. Bohannon, "Reference values for the timed up and go test: A descriptive meta-analysis," *Journal of Geriatric Physical Therapy*, vol. 29, no. 2, pp. 64–68, 2006.

[17] G. M. Savva, O. A. Donoghue, F. Horgan, C. O'Regan, H. Cronin, and R. A. Kenny, "Using timed up-and-go to identify frail members of the older population," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 68, no. 4, pp. 441–446, 2013.

[18] B. S. Moreno, C. Zimerer, M. J. Salamanca, R. C. Mello, and A. Frizera, "Human–machine interaction in dual-task application using a self-paced treadmill and serious video games," in *XXIX Brazilian Congress on Biomedical Engineering, Vol. 1: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing*. Springer Nature Switzerland, 2025, pp. 81–91.

[19] S. Matsudo, T. Araújo, V. Matsudo, D. Andrade, E. Andrade, L. C. Oliveira, and G. Braggion, "International physical activity questionnaire (IPAQ): study of validity and reliability in Brazil," *Revista Brasileira de Atividade Física & Saúde*, vol. 6, no. 2, pp. 5–18, 2001.

[20] W. H. Organization, *Physical Status: The Use and Interpretation of Anthropometry*. Geneva, Switzerland: World Health Organization, 1995.

[21] P. H. F. Bertolucci, S. M. D. Brucki, S. R. Campacci, and Y. Juliano, "O mini-exame do estado mental em uma população geral: impacto da escolaridade," *Arquivos de Neuro-Psiquiatria*, vol. 52, no. 1, pp. 1–7, 1994.

[22] D. F. Lourenço, E. V. Carmona, and M. H. B. d. M. Lopes, "Translation and cross-cultural adaptation of the System Usability Scale to Brazilian Portuguese," *Aquichan*, vol. 22, no. 2, p. e2228, 2022.

[23] S. T. Baudendistel, A. C. Schmitt, A. E. Stone, T. E. Raffegeau, J. A. Roper, and C. J. Hass, "Faster or longer steps: Maintaining fast walking in older adults at risk for mobility disability," *Gait & Posture*, vol. 89, pp. 86–91, 2021.

[24] J. H. Hollman, E. M. McDade, and R. C. Petersen, "Normative spatiotemporal gait parameters in older adults," *Gait & Posture*, vol. 34, no. 1, pp. 111–118, 2011.

GESTUS: Assistive Interface for Augmentative and Alternative Communication and Gaze-Controlled Mouse Interaction: Proof of Concept and Exploratory Analysis of the Image Pipeline

1st Luis Carlos Rodrigues dos Anjos
Dept. of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
luiscarlosengcomp@gmail.com

2nd Edilson Rocha Porfirio Filho
Dept. of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
edilson.filho@ufc.br

3rd Marco Antônio de H. Penaforte Neto
Dept. of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
penafortemarco@gmail.com

4th João Paulo de Sousa Santos
Dept. of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
sjoao2564@gmail.com

***Abstract*—This paper presents GESTUS, a proof-of-concept assistive technology system for human-computer interaction operated exclusively by a conventional webcam and aimed at individuals with severe motor restrictions. The system integrates MediaPipe, image preprocessing, 3×3 grid gaze calibration, automatic interface scanning, blink-based selection, and speech synthesis. We evaluate eight preprocessing techniques combined with five interpolation methods across three facial luma ranges, with and without glasses. Configurations using the median filter reached the highest average separability among centroids while staying above 33 FPS. Classification accuracy on frames temporally separated from calibration inverts this ranking: the Gaussian filter leads with an accuracy of 0.559, κ of 0.504, and top-2 of 0.752, whereas the median filter drops to seventh of eight, showing that centroid separability does not predict classification performance. Per-session z-score standardization of the attributes raises accuracy to 0.613 at no extra cost, and 92.0% of predictions fall on the correct cell or an immediate neighbor. Because the error is predominantly local and rejection improves reliability, the interface can operate in a conservative regime that favors reliability over coverage. The exploratory analysis was done with one participant and without usability evaluation.**

***Index Terms*—Assistive Technology, Augmentative and Alternative Communication, Eye Tracking, Proof of Concept, Computer Vision, MediaPipe, OpenCV.**

I. Introduction

People with reduced mobility from amyotrophic lateral sclerosis (ALS) or the sequelae of a stroke face substantial barriers with technology, since conventional input devices such as keyboards, mice, and touchscreens are often unsuitable. Augmentative and Alternative Communication (AAC) brings together strategies to complement or replace speech, with effectiveness documented in both clinical and home settings [1]–[3]. Computer vision offers one route into this space, making it possible to build contact-free interfaces based on eye tracking and blink detection. GESTUS is a non-invasive assistive system of this kind, operated by a conventional webcam.

The computer vision stage uses MediaPipe together with image processing techniques to extract in real time the features needed to drive the interface [4]. The MediaPipe Face Landmarker supplies facial landmarks, facial expression scores (*blendshapes*), and geometric information about the face [5], all of which matter for representing facial geometry from monocular video [6].

This paper asks whether a usable ocular signal for discrete control can be extracted from a conventional webcam and which preprocessing decisions are most promising. We describe the architecture, image processing, and communication interface of GESTUS and measure how preprocessing and interpolation affect the separability of nine gaze directions, classification accuracy on a temporally separated test set, and real-time throughput. No claim of clinical efficacy, usability, or equivalence to dedicated eye trackers is made. The contributions are four: (i) reproducible integration of calibration, gaze classification, scanning, blink detection, and speech synthesis in a webcam-only application; (ii) systematic screening of 40 preprocessing and interpolation configurations under varying facial luma and glasses, assessed by both separability and temporal-split accuracy; (iii) instrumentation for future evaluation with users; and (iv) a low-cost classifier modification, per-session attribute standardization, that raises accuracy at no extra computational cost.

II. Related Works

A. AAC Tools and Access Methods

Among established AAC solutions, the closest functional reference for GESTUS is Jabbla's Mind Express 5, which organizes vocabulary into grids of selectable cells and accepts several access methods—touch, mouse, joystick, switches, head tracking, and eye tracking—while extending gaze control to the Windows environment for typing, scrolling, and mouse-equivalent actions [7]. GESTUS sets out to reproduce exactly this arrangement, in which a single input channel serves both cell-based communication and control of the operating system. What differs is the hardware it assumes: commercial gaze access depends on a dedicated tracking module, and eye-typing systems are reported to cost between $5,000 and $10,000 depending on configuration [8], a decisive barrier in low-income settings.

Scanning access, in which cells are highlighted in sequence and the user confirms with a single voluntary gesture, is the classical alternative when motor control cannot support direct pointing [3]. It trades spatial precision for time: the communication rate drops, but the motor demand falls to binary activation. GESTUS pairs this strategy with an unusual trigger—a deliberate blink via webcam—and with synthesized speech, placing it between low-cost printed communication boards and commercial suites tied to specific hardware.

B. Gaze-Based Assistive Interfaces

Work on gaze interaction for users with severe motor impairment differs in three main respects: the sensor it requires, the granularity of the signal it extracts, and how far it has been validated with the target population.

Several systems remove dedicated hardware from the equation. Chhimpa et al. [9] present real-time, low-cost eye tracking with no mandatory calibration; Kumar et al. [10] drive a cursor from facial landmarks with OpenCV; Deepika and Murugesan [11] combine cursor movement with blink selection; and Buhagiar et al. [12] explore gaze-driven web navigation. GESTUS shares this premise but departs from [9] in keeping a short nine-point calibration stage: the 3×3 grid needs references for diagonal regions, and a per-session baseline absorbs variation in positioning and lighting between uses.

Granularity involves a recurring trade-off between spatial precision and robustness. Blink-To-Live [8] pushes this to the limit by discarding position altogether and encoding voice-synthesized commands with four facial gestures on a smartphone. Jain et al. [13] take the opposite route, with typing driven by continuous eye analysis. GESTUS sits between the two: the 3×3 grid retains enough precision for cursor control, while automatic scanning supports communication without direct pointing and eases the burden of sustained fixation (Section II-C).

Validation is where SpeakFaster [14] sets the highest bar. Combining gaze typing with language models that expand abbreviated text, it saved 57% of motor actions in simulation and, in laboratory and field trials with two ALS users experienced in gaze typing, achieved text-entry rates 29% to 60% above the baselines. That work shows how far the evaluation of gaze-based AAC extends beyond the sensor, into interaction and language.

Taken together, the reviewed systems cover the individual pieces GESTUS depends on: webcam-only capture, blink confirmation, and gaze-driven pointing. What we did not find was a low-cost combination of them serving both cell-based communication and cursor control, as the commercial suites in Section II-A do with dedicated hardware. GESTUS explores that combination, and this article characterizes the image pipeline that supports it.

C. Challenges and Obstacles of Gaze Interaction

The literature documents several obstacles that shape the design of systems like GESTUS. The main one is the *Midas touch* problem: because gaze is at once the channel for inspection and for command, the system cannot tell reading apart from an intention to act [15], and a second signal must express that intention. The most common strategy is dwell time, usually between 500 and 1000 ms [16], which trades slowness and sustained fixation against unintended selections as the interval is shortened or lengthened. The alternative is an explicit gesture such as a voluntary blink, which separates pointing from confirmation unambiguously, at the cost of the discomfort of repeated blinking and the risk that spontaneous blinks are read as commands [15], [17]. GESTUS uses the explicit gesture and handles that risk with the duration thresholds of Section III-E.

Visual fatigue and sustained effort are a second obstacle, and they matter especially in neurodegenerative disease, where holding a fixation is costly. The automatic scanning module is a design decision aimed at this problem: the highlight moves through the cards on its own, gaze is no longer the pointing channel, and all that remains for the user is the confirmation gesture. This reduces cumulative fixation time shifting the cost from spatial precision to waiting.

A third obstacle is how sensitive appearance-based approaches are to capture conditions. Methods that estimate gaze in visible light are affected by non-uniform illumination, shadows, reflections on corrective lenses, and reduced contrast between iris, pupil, and sclera, which degrades accuracy in uncontrolled environments relative to infrared trackers [18]; failures are more frequent in low luminance, particularly among users with visual correction [19]. Two decisions in GESTUS respond to this: adaptive, iterative percentage thresholding of the ocular region, which readjusts the threshold according to the local intensity distribution, and enhancement techniques applied before ocular extraction, evaluated in Section III-B across facial luma ranges.

III. Materials and Methods

A. Overview, Architecture, and Study Delimitation

GESTUS is a Python application organized under the MVC pattern. It uses OpenCV and MediaPipe for computer vision,

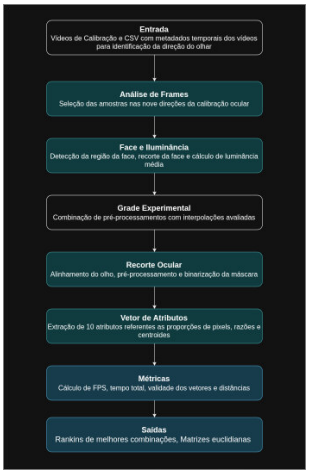


Fig. 1: Exploratory analysis workflow used to compare facial luma, preprocessing techniques, interpolations, feature vectors, Euclidean distances, and processing cost.

PyAutoGUI for cursor control, and gTTS for speech synthesis. This paper covers the technical integration and the choice of preprocessing, binarization, and interpolation settings for the visual pipeline. Clinical evaluation is future work: the data come from a single participant with no independent replication, and the focus here is the technical proof of concept.

B. Exploratory Analysis and Experimental Protocol

The analysis covers 40 configurations, combining 8 preprocessing methods with 5 interpolation methods ($8 \times 5 = 40$). OpenCV lists six interpolation options, but linear and bilinear follow the same execution path in two-dimensional resizing and produce identical results, so they are counted once. The aim is a comparative baseline across capture conditions, enhancements, and ocular metrics that can inform the choice of real-time pipeline (Fig. 1).

The experimental set contains six videos recorded by a single participant, one for each combination of the three lighting conditions with and without glasses. Each video lasts 82.0 s, with 2,460 frames at 30 FPS and a resolution of 640×480. For each configuration, 1,080 frames per video were analyzed, evenly distributed across the nine directions of the 3×3 grid (120 samples per direction), giving 6,480 unique frames and 259,200 processing records in total. There was no independent experimental replication: each combination of illumination and glasses was represented by a single video.

The recordings used a Bison Electronics camera with automatic exposure and gain. The facial luma we report therefore describes the resulting image rather than the ambient illuminance, which partly compresses the variation between capture conditions. The participant stayed 45–50 cm from the lens. Processing ran on a Lenovo IdeaPad (i5-13420H, 24 GB RAM) with Ubuntu 24.04.4, Python 3.12.13, OpenCV 4.13.0, and MediaPipe 0.10.33.

C. Image Processing Pipeline

RGB frames are processed with MediaPipe to identify the face, landmarks, and *blendshapes*. The system then crops the face, resizes it to 384×384 pixels, applies the enhancement, and extracts the eyes in grayscale. Ocular tilt is corrected by aligning the interocular axis with the horizontal through an affine transformation, using $\theta = \arctan 2(y_R - y_L,\, x_R - x_L)$; this keeps the anatomical correspondence between the processed bands stable even with the head tilted. Binarization uses inverted thresholding, assigning 255 to segmented dark pixels:

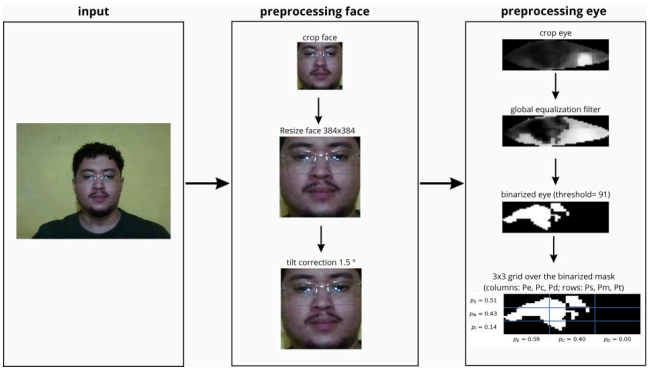


Fig. 2: Image processing pipeline steps on a real frame from the semi-lit session with glasses ($\overline{Y'} = 83.3$), using global equalization and nearest-neighbor interpolation. Top row: original 640×480 frame, facial crop, resizing to 384×384, and 1.5° tilt correction. Bottom row: grayscale ocular crop, applied enhancement, binary mask obtained by Eq. (1) with threshold 59, and division into bands with the segmented pixel proportions p_E, p_C, p_D, p_S, p_M, and p_I that make up the feature vector.

$$B(x,y) = \begin{cases} 255, & I(x,y) \le \tau, \\ 0, & I(x,y) > \tau, \end{cases} \qquad \tau^* = \arg\min_{\tau} \left| P_{\text{seg}}(\tau) - P_{\text{alvo}} \right|. \tag{1}$$

The second expression defines the iterative percentage mode, in which the threshold is adjusted until the proportion of segmented pixels approaches a predefined target (66.0% with a tolerance of 0.5%, an initial threshold of 14, and at most 255 iterations). The pipeline also supports fixed thresholding, Otsu's method [20], and local adaptive thresholding [21], which is the system's main defense against mask degradation in low light.

Fig. 2 shows the full chain on a real frame from the semi-lit session with glasses. The final panel makes the effect of capture conditions on the feature vector visible: a specular reflection on the lens introduces a high-intensity region in the right portion of the ocular crop which, after inverted binarization, is left unsegmented, giving $p_D = 0.00$ against $p_E = 0.59$. The asymmetry comes from an optical artifact rather than from gaze direction, and it shows the mechanism by which corrective lenses degrade classification; Section IV-E quantifies the effect.

1) Preprocessing Techniques: The main criterion is not visual quality but the effect on the *binary mask*. Since the ten attributes of the vector are area proportions and coordinates computed over that mask, the technique has to stabilize the segmented contour and preserve the contrast between the dark region and the sclera. The comparison measured directional

separability, mask stability, frame rate (FPS), and total processing time.

The eight techniques differ in the support of the operation. Point transformations map each intensity independently of its neighborhood and change the histogram from which τ is computed, without altering the local geometry of the contour; contrast operations redistribute intensities from global or local statistics; and spatial filters act on a neighborhood and directly affect the regularity of the segmented edge.

- *No preprocessing*: baseline; the crop goes straight to binarization.
- *Global equalization* [21]: takes the cumulative intensity distribution as the mapping function, expanding the dynamic range and amplifying noise in the eyelid shadows.
- *CLAHE* (*Contrast Limited Adaptive Histogram Equalization*) [22]: block-based equalization (8×8) with histogram clipping (clip 2.0); on a small ocular crop it accentuates eyelashes and reflections.
- *Gamma correction* [23]: point transformation $I' = 255\,(I/255)^{\gamma}$ with $\gamma = 1.5$, which shifts the midtones and with them the threshold that satisfies the target proportion.
- *Single-scale Retinex* [24]: estimates the illumination by Gaussian convolution ($\sigma = 30.0$) and subtracts it in the logarithmic domain, attenuating non-uniform illumination and shadows.
- *Bilateral filter* [25]: weighted average over spatial and intensity kernels ($d = 5$, $\sigma_{\text{color}} = \sigma_{\text{space}} = 50$); smooths homogeneous regions while preserving the iris–sclera transition.
- *Gaussian filter* [26]: 5×5 linear low-pass convolution with continuous response; attenuates high-frequency contour noise at the cost of isotropic blurring.
- *Median filter* [27]: replaces each pixel with the median of its 5×5 neighborhood; removes impulse noise without blurring edges, but, being an order filter, produces quantized jumps in the output, a property we return to in Section IV-C.

2) Interpolation and Resizing: The detected face is resized to 384×384 pixels, which standardizes the scale before ocular extraction. Faces in 640×480 frames at 45–50 cm occupy between 180 and 260 pixels, so this stage acts in practice as an *upscaling*, with scale factors between 1.5 and 2.1.

Five numerically distinct strategies were evaluated [28]. They differ in the support of the reconstruction kernel, that is, in the number of source samples that contribute to each destination pixel.

- *Nearest neighbor*: zero order; copies the closest sample without averaging. It is the fastest and creates no new intensities, at the cost of pixelation.
- *Area*: average of the source pixels covered by the projection of the destination pixel; avoids aliasing when downscaling, but degenerates to nearest neighbor when *upscaling* [28], which explains how close the two are in Table I.
- *Linear/bilinear*: first order over 4 neighbors, interpolating along one axis and then the other; in two-dimensional resizing both resolve to the same procedure and are counted once.
- *Bicubic*: cubic polynomials over 16 neighbors (4×4); a smoother result at a higher per-pixel cost.
- *Lanczos*: windowed sinc kernel over 64 neighbors (8×8); sharper, but prone to ringing near high-contrast edges and the slowest of the five (Table II).

Interpolation has little effect on accuracy: the ten attributes are first-order statistics over the mask, and upscaling only estimates values without adding spatial detail. Preprocessing, by contrast, changes the intensity distribution on which τ is computed, and dominates the result.

3) Facial Luma and Lighting Ranges: The mean digital luma of the face converted to RGB is computed over the region Ω_f with BT.709 coefficients [29]: $Y' = 0.2126R' + 0.7152G' + 0.0722B'$ and $\overline{Y'} = N^{-1} \sum_{(x,y)\in\Omega_f} Y'(x,y)$, with N the number of valid pixels. The sessions were split *a posteriori* from the six observed values (Fig. 6) into three ranges: dark ($40 \le \overline{Y'} < 70$), medium ($70 \le \overline{Y'} < 90$), and bright ($90 \le \overline{Y'} \le 110$).

Facial illuminance (lux) is not the same as the flux emitted by the source (lumen), since incidence varies with geometry. With no luxmeter available, the study used mean digital luma $\overline{Y'}$ as an approximate substitute, aware that it depends on camera exposure. A formal photometric measurement correlated with $\overline{Y'}$ is part of the validation protocol planned for future work.

4) Feature Vector and Euclidean Distance: After binarization, the ocular region is divided into three vertical bands (left, center, right) and three horizontal bands (top, middle, bottom). For each band Ω_i we compute the proportion of segmented pixels within the valid area of the mask, $p_i = |\Omega_i|^{-1} \sum_{(x,y)\in\Omega_i} \mathbf{1}[B(x,y) = 255]$, with $\mathbf{1}[\cdot]$ the indicator function. The feature vector has ten components:

$$\mathbf{v} = [p_E, p_C, p_D, p_S, p_M, p_I, r_H, r_V, c_x, c_y], \tag{2}$$

where $r_H = (p_D - p_E)/(p_D + p_E + \epsilon)$ and $r_V = (p_I - p_S)/(p_I + p_S + \epsilon)$ summarize the horizontal and vertical displacement of the segmented pixels, and c_x, c_y are the normalized centroid coordinates. A vector counts as valid only when all ten attributes are present and numeric. In the experiments the vector is extracted from the left eye, which keeps the analysis consistent with the channel used for content selection in the interface and avoids doubling the dimensionality; a comparison between the left eye, the right eye, and both concatenated is planned for the multiparticipant study. The system also records the eyelid height of each eye, given by the Euclidean distance between landmarks 159 and 145 (left) and 386 and 374 (right), used as an auxiliary reference in blink detection.

Classification compares the current vector with a calibrated mean vector $\bar{\mathbf{v}}_d$ for each direction d, and the estimated direction is the nearest reference:

$$D(\mathbf{v}, \bar{\mathbf{v}}_d) = \sqrt{\textstyle\sum_{i=1}^{10} (v_i - \bar{v}_{d,i})^2}, \quad \hat{d} = \arg\min_d D(\mathbf{v}, \bar{\mathbf{v}}_d). \tag{3}$$

In the comparative analysis, the same metric produces distance matrices between the mean vectors of distinct directions, $M_{a,b} = D(\bar{\mathbf{v}}_a, \bar{\mathbf{v}}_b)$; higher off-diagonal means indicate greater separability between centroids. Separation between centroids is not, however, equivalent to classification accuracy, a point we return to in Section IV-C.

Accuracy is evaluated with a per-session temporal split. For each combination of video and direction the frames were ordered: the first 30 (one second) form the calibration for the threshold and $\bar{\mathbf{v}}_d$, mimicking initialization, and the remaining 90 form the test set. Centroids are extracted separately for each session, since GESTUS recalibrates on every use and sessions under different illumination should not be pooled. Each test frame is then labeled by its nearest centroid through Eq. (3), which mimics real operation.

The temporal split avoids reusing calibration frames, but it does not imply statistical independence: consecutive frames from the same session share participant, camera, positioning, and environmental conditions. The six sessions are the only experimental units available, so we do not report aggregated confidence intervals; Table IV gives the per-session accuracy of the best configuration instead, which is more transparent at $n = 6$.

D. Initialization and Operation Flows

Initialization (Fig. 3) loads the services and MediaPipe. With the user looking at the center, the system iteratively adjusts the binarization threshold to reach the target proportion of pixels and records the mean eyelid height, the reference for blinks. Gaze is then calibrated by capturing a ten-dimensional vector for each of the nine points on the screen. Unlike strictly calibration-free approaches [9], this brief stage gives the mapping of the diagonal regions its robustness.

Continuous operation (Fig. 4) involves detection, enhancement, and ocular binarization for attribute extraction and eyelid measurement. Blink detection combines the MediaPipe blendshapes with the calibrated relative height, and to avoid false positives the system requires ocular dominance: a one-eyed blink is confirmed only if the other eye stays below the threshold. Closing both eyes triggers suspension. The spatial classifier may later give way to supervised models.

E. Functionalities and Definition of Temporal Thresholds

Fig. 5 shows the main screens. Card-based communication organizes the vocabulary into categories covering family, actions, needs, pain and body, feelings, an intensity scale, numbers, the alphabet, yes/no answers, and questions. The categories come directly from the boards of the Alternative Communication Notebook for COVID-19 Patients with Reduced Mobility, produced by the Occupational Therapy Department at UFRJ [30]: the *what I would like to do* board, for instance, gave rise to the actions category. Adopting this lexicon anchors the interface in vocabulary already organized by occupational therapists around the needs of hospitalized patients without functional speech [2], [3]. The validation protocol includes

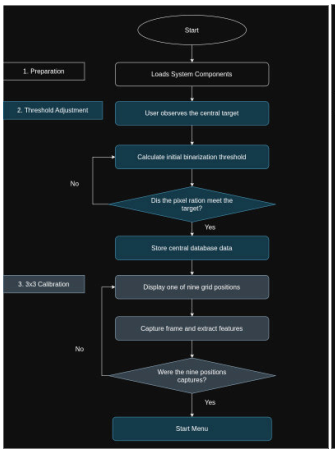


Fig. 3: Initialization and 3×3 grid calibration workflow.

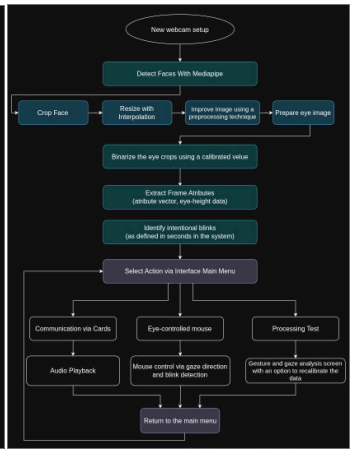


Fig. 4: GESTUS real-time processing workflow.

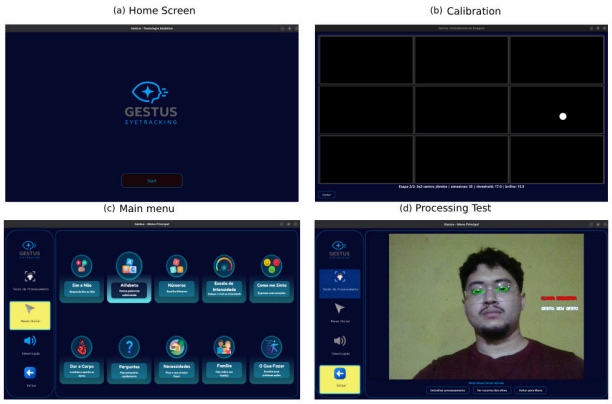


Fig. 5: GESTUS screens: (a) initial screen; (b) calibration; (c) main menu; and (d) processing test.

professionals from HUWC and PROTECS/UFC in adapting the vocabulary and in the evaluation with users and caregivers.

Automatic scanning advances the card highlight at a fixed interval, so the user only has to blink deliberately to confirm the selection and have it read aloud. During scanning, the left eye selects the content of the grid and the right eye moves through the options, which reduces ambiguity. In mouse control, cardinal and diagonal gaze directions move the cursor along the corresponding axes, and closing both eyes for 3 s suspends or resumes detection.

The temporal thresholds were anchored in the literature on eyelid kinematics and then tuned in an internal pilot. A single blink lasts mostly between 100 and 400 ms, with a mean for endogenous blinks near 334 ms ($\pm$67 ms) [31], while a *voluntary* blink is substantially longer at 572 ± 25 ms [32]. We therefore placed the 0.5 s threshold for a single click *above* the usual limit for spontaneous blinks and inside the voluntary range, which reduces false commands from reflex blinks [15]. The thresholds of 1 s (double click) and 3 s (suspension) give larger margins to actions whose reversal cost is higher, consistent with usual dwell values [16]. Future evaluation will measure their effect on execution time, response time, and hit rate to support per-user calibration.



Fig. 6: Representative frames of the six sessions in the *center* direction with ocular landmarks, across the three facial luma ranges and with and without glasses, annotated with the respective $\overline{Y'}$: low light (59.0 and 60.3), medium light (81.2 and 83.3), and high light (93.2 and 95.9).

TABLE I: Mean Euclidean separability among centroids: six configurations with highest separability (six sessions from a single participant).

Preprocessing	Interpolation	Mean	SD	Min.	Max.
Median	Linear/Bilinear	0.8975	0.3185	0.1554	1.7560
Median	Nearest Neighbor	0.8961	0.3060	0.1374	1.7797
Median	Area	0.8955	0.3169	0.1533	1.7560
Bilateral	Linear/Bilinear	0.8929	0.3159	0.1554	1.7448
Bilateral	Area	0.8887	0.3139	0.1543	1.7421
Bilateral	Nearest Neighbor	0.8881	0.2989	0.1443	1.7566

IV. Preliminary Results and Discussion

The implemented prototype includes threshold calibration, nine-point gaze calibration, the communication menu, automatic scanning, audio playback, gaze classification, eye-controlled mouse operation, and blink detection based on blendshapes and eyelid geometry. The results below correspond to the technical evaluation of that prototype.

A. Facial Luma and Capture Conditions

The observed luma values were used descriptively to group the six sessions into three ranges, since ocular segmentation is affected by shadows, reflections on glasses, saturation in bright areas, and reduced contrast between iris, pupil, and sclera. The six sessions separate cleanly: $\overline{Y'} = 59.0$ and 60.3 (low light), 81.2 and 83.3 (medium light), and 93.2 and 95.9 (high light), as in Fig. 6. Throughout the text the conditions are referred to both by qualitative label and by $\overline{Y'}$. Since $\overline{Y'}$ derives from the image itself, it should not be read as photometric illuminance.

B. Separability, Computational Cost, and Accuracy

Table I lists the configurations with the highest mean Euclidean distance between centroids. The median and bilateral filters lead, while CLAHE/Lanczos records the smallest separation (0.3969). Because the gaps between the leading configurations are small, they were not tested for significance and the ranking is reported descriptively.

Table II compares the fastest and slowest combinations for each luma range, which matters because a technique with good separability may be expensive to compute. The vector validity rate was 100% in every condition, confirming that all ten attributes were extracted. The fastest options (Retinex or no preprocessing) are not the ones with the highest separability (median and bilateral). Among the best in separability, however, the median filter with bilinear, nearest-neighbor, or area interpolation stayed above 33 FPS, which is compatible with capture at 30 FPS. Section IV-D details the spatial structure of the error.

TABLE II: Fastest and slowest combinations per facial luma range.

Luma Range	Fastest	FPS	Slowest	FPS
Low ($\overline{Y'} \approx 60$)	Retinex/Linear	41.89	Gaussian/Lanczos	34.64
Medium ($\overline{Y'} \approx 82$)	No prep./Linear	40.15	Median/Lanczos	35.16
High ($\overline{Y'} \approx 95$)	No prep./Nearest	39.35	Gaussian/Lanczos	30.71

TABLE III: Nine-direction classification accuracy on a temporal test set separated from calibration (six sessions of a single participant).

Preprocessing	Interpolation	Accuracy	F1	κ	Top-2
Gaussian	Nearest Neighbor	0.5588	0.5555	0.5037	0.7519
Bilateral	Nearest Neighbor	0.5576	0.5542	0.5023	0.7473
Gaussian	Bilinear	0.5508	0.5472	0.4947	0.7529
Median	Nearest Neighbor	0.5287	0.5250	0.4698	0.7371
CLAHE	Nearest Neighbor	0.4510	0.4495	0.3824	0.6753

The distance between centroids measures how far apart the class prototypes are, but it is insensitive to the dispersion of the samples around each prototype. To check whether the ranking holds, the same 40 configurations were re-evaluated by classification accuracy on a test set temporally separated from calibration, following the protocol of Section III-C4.

C. Separability Does Not Predict Accuracy

Comparing the two criteria produces the most relevant finding of this analysis. The median filter, which leads on separability with a mean distance of 0.8975, falls to seventh among the eight techniques when the criterion is accuracy, at 0.5287. The Gaussian filter, absent from the top six on separability, leads on accuracy with 0.5588 (Table III).

The inversion occurs because distance assesses only how far apart the prototypes are, ignoring intraclass dispersion. One hypothesis is that the median filter, being an order filter, produces quantized jumps in the contour and creates centroids that are distant but unstable, unlike Gaussian smoothing, which softens the boundary and reduces dispersion in a way that improves accuracy. CLAHE, operating on 8×8 blocks, amplifies unwanted structures such as eyelashes and reflections under inverted binarization, which accounts for its low performance (mean of 0.4663 across interpolations). Since interpolation has only a marginal effect (0.530 to 0.538), preprocessing dominates the result: separability between centroids does not predict accuracy and should not guide the choice of pipeline on its own.

D. Error Structure

Although κ is only moderate (0.504) [33], the structure of the error is local. In the 3×3 grid, predictions falling on the correct cell or an immediate neighbor (Chebyshev distance ≤ 1) reach 0.920, so only 8.0% of failures land in distant regions. Among the errors, 82% are adjacent and 43.8% carried the correct class as the second candidate (top-2 of 0.752), which will make later disambiguation easier. Corner directions have a higher top-2 (0.821) because they have fewer

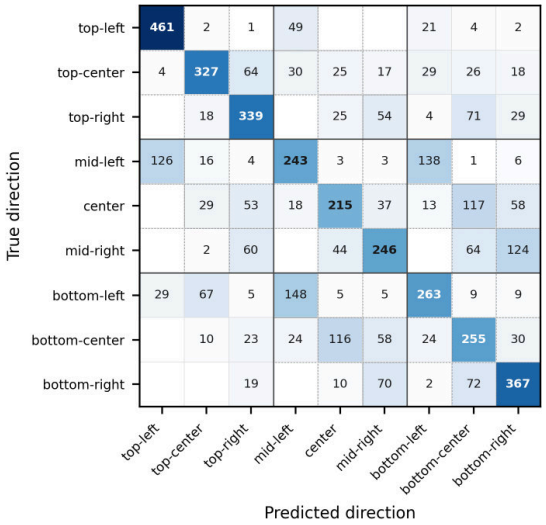


Fig. 7: Confusion matrix of the best configuration (Gaussian/nearest neighbor), aggregated over the six sessions: 540 test frames per direction, 4,860 in total. Cells report counts; blank cells indicate zero. Thick lines delimit the bands of the 3×3 grid; off-diagonal mass concentrates in cells adjacent to the true direction.

TABLE IV: Accuracy of the best configuration (Gaussian/Nearest Neighbor) across six sessions, over 810 test frames per session.

Session	Glasses	$\overline{Y'}$	Accuracy
Low light	No	60.3	0.6568
Low light	Yes	59.0	0.5765
Medium light	No	81.2	0.6111
Medium light	Yes	83.3	0.4333
High light	No	93.2	0.5333
High light	Yes	95.9	0.5420
Aggregated six sessions			**0.5588**

neighbors, while the *center* suffers more confusion (top-1 of 0.398), which makes it well suited as a resting dead zone.

The error is therefore local and structured rather than randomly dispersed. Fig. 7 confirms the pattern: off-diagonal mass concentrates in neighboring cells, with reciprocal confusions between mid-left and bottom-left (138 and 148 frames) and between center and bottom-center (117 and 116), both vertically adjacent pairs in the grid. The distinction matters for the application: in gaze-controlled mouse operation an error to an adjacent cell produces a small displacement that later frames can correct, whereas an error to the opposite cell would move the cursor against the user's intention.

E. Variation by Capture Condition

Aggregating the six sessions hides substantial variation, detailed in Table IV. Stratifying the test set of the best configuration, accuracy ranges from 0.4333 in the semi-lit condition with glasses ($\overline{Y'} = 83.3$) to 0.6568 in the poorly lit condition without glasses ($\overline{Y'} = 60.3$), an amplitude of 22.4 percentage points that exceeds any difference observed between preprocessing configurations. Corrective lenses lower accuracy from 0.6004 to 0.5173, a drop of 8.3 points consistent with the degradation reported in the presence of specular reflections [18], [19] and illustrated in Fig. 2. The effect of illumination ran counter to expectation: the poorly lit condition gives the highest accuracy (0.6167), the semi-lit the lowest (0.5222), and the lit condition an intermediate value (0.5377).

Two hypotheses could account for this. Neither was tested, and they are not mutually exclusive. Thresholding compensates for global drops in light but does not resolve reflections on lenses and sclera under high luminance. Beyond that, the automatic adjustments of the camera may have preserved ocular contrast in the dark condition while saturating areas in the bright one. Separating the two causes will depend on photometric records fixed in lux in future validations.

F. Rejection and Distance Metric

In assistive interfaces, incorrect commands impose a correction cost on the user. Adopting the normalized margin $m(\mathbf{v}) = [D(\mathbf{v}, \bar{\mathbf{v}}_{d_{(2)}}) - D(\mathbf{v}, \bar{\mathbf{v}}_{d_{(1)}})]/[D(\mathbf{v}, \bar{\mathbf{v}}_{d_{(2)}}) + D(\mathbf{v}, \bar{\mathbf{v}}_{d_{(1)}})]$ and rejecting decisions below a threshold, accuracy rises from 0.559 at full coverage to 0.692 at 50% coverage and 0.808 at 20%. Incorrect classifications then drop from 44.1% to 30.8% and 19.2%. This fraction is not the final rate of false commands, since scanning and temporal aggregation are evaluated later. At 30 FPS, discarding 50% of the frames still leaves 15 decisions per second, enough for cursor control. The margin threshold calibrates the trade-off between responsiveness and reliability without modifying the extractors.

Because the ten attributes have heterogeneous scales (proportions, ratios, and coordinates), Euclidean distance imposes distorted weights. Z-score standardization corrects the distortion at no extra computational cost, raising mean accuracy from 0.533 to 0.579. The best configuration (Gaussian/Nearest) reached an accuracy of 0.613 (κ of 0.565) and a top-2 of 0.797. Although 0.613 is still preliminary for uninterrupted frame-by-frame classification, the practical viability of the interface rests on the spatial adjacency of the error (Section IV-D) and on margin-based rejection.

G. Limitations of the Technical Evaluation

The main methodological limitation is the absence of a comparison with a dedicated infrared eye tracker. Since GESTUS classifies nine discrete regions instead of estimating a continuous point, conventional validation metrics—mean angular error, precision, and data loss rate [34]—do not apply directly. This affects only the absolute interpretation of the results; the relative comparison still holds, as the internal comparison of the 40 configurations follows the same protocol and temporal split throughout. It is not possible, however, to separate the portion of error due to the pipeline from the portion originating in label imprecision, which assumes static fixation on the target. Synchronous recordings from webcam and tracker will make that decomposition possible. Other limiting factors include camera resolution, user positioning and distance, involuntary head movement, and the distinction between natural and command blinks [10], [11], [18], [19].

V. Conclusion and Future Work

This paper presented GESTUS, a proof of concept for assistive technology that supports human-computer interaction

through eye movement and deliberate blinks using only a standard webcam, integrating nine-point calibration, image processing with MediaPipe, gaze direction classification, card-based communication with scanning, speech synthesis, and cursor control.

The experiment addressed two questions. On the pipeline, mean distance between centroids ignores intraclass dispersion and does not work as a sole criterion: the median filter led on separability but fell to seventh on accuracy. On webcam-based extraction, standardizing the attributes raised accuracy to 0.613 at no extra cost. That value is still preliminary for operational robustness, but the interface remains viable because of the local structure of the error, with 92.0% of predictions on the correct or an adjacent cell, and because it can run conservatively, reaching 0.808 accuracy once low-margin decisions are rejected.

Without an external tracker, pipeline error cannot be separated from label imprecision. The 22.4 percentage point amplitude between sessions also exceeds the gains from preprocessing, which indicates that standardizing capture conditions affects performance more than refining the algorithm. The accuracy and κ reported here describe one participant's ocular anatomy, iris pigmentation, and capture conditions, and do not generalize.

Future work follows two lines. The technical line covers expansion to multiple participants with sampling stratified by the Fitzpatrick scale, photometric control of illumination with a luxmeter at the plane of the face, evaluation of different cameras and of head movement, and comparison with a reference eye tracker in synchronous recording. For the evaluation with users, the protocol approved by CEP/HUWC-UFC (Opinion no. 8,491,889) covers 10–12 participants without motor limitation and 10–12 with severe motor impairment, measuring execution and selection time, hit rate, false command rate, cursor stability, fatigue, and satisfaction. Those results will support per-user calibration of the temporal thresholds and the replacement of the classifier with supervised models.

Acknowledgment: The authors thank UFC, PROTECS, and GESTUS collaborators for their support.

REFERENCES

[1] F. M. Bandeira, F. P. de Faria, and E. B. de Araujo, "Avaliação da qualidade intra-hospitalar de pacientes impossibilitados de falar que usam comunicação alternativa e ampliada," *Einstein (São Paulo)*, vol. 9, no. 4, pp. 477–482, 2011.

[2] H. Carruthers, F. Astin, and W. Munro, "Which alternative communication methods are effective for voiceless patients in Intensive Care Units? A systematic review," *Intensive and Critical Care Nursing*, vol. 42, pp. 88–96, 2017.

[3] D. R. Beukelman and J. C. Light, *Augmentative and Alternative Communication*, 5th ed. Baltimore, MD, USA: Paul H. Brookes, 2020.

[4] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," arXiv:1906.08172, 2019.

[5] Google AI Edge, "Face Landmark Detection Guide," 2026. [Online]. Available: https://developers.google.com/edge/mediapipe/solutions/vision/face_landmarker. Accessed: Jul. 29, 2026.

[6] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-Time Facial Surface Geometry from Monocular Video on Mobile GPUs," in *CVPR Workshop on CV for AR/VR*, 2019.

[7] Jabbla, "Mind Express 5," 2026. [Online]. Available: https://www.jabbla.com/en/mindexpress/. Accessed: Jul. 29, 2026.

[8] M. Ezzat, M. Maged, Y. Gamal, M. Adel, M. Alrahmawy, and S. El-Metwally, "Blink-To-Live eye-based communication system for users with speech impairments," *Scientific Reports*, vol. 13, art. no. 7961, 2023.

[9] G. R. Chhimpa, A. Kumar, S. Garhwal, and Dhiraj, "Empowering individuals with disabilities: A real-time, cost-effective, calibration-free assistive system utilizing eye tracking," *J. Real-Time Image Processing*, vol. 21, no. 97, 2024.

[10] A. K. Kumar *et al.*, "Real-Time Eye-Tracking Mouse Control System Using OpenCV and Facial Landmark Detection," in *Proc. IEEE INSPECT*, 2024.

[11] S. S. Deepika and G. Murugesan, "A Novel Approach for Human Computer Interface Based on Eye Movements for Disabled People," in *Proc. IEEE ICECCT*, 2015.

[12] M. Buhagiar, D. Calleja, and C. Porter, "Experiencing Cactus–Browsing the Web Using Your Eyes," in *CHItaly 2025*, 2025, pp. 1–2.

[13] P. B. Jain, V. Hiremath, S. Bhat, G. Pujari, and D. C., "Eye Typing-Vision Based Human Activity Control," in *Proc. IEEE ICDCECE*, 2022.

[14] S. Cai *et al.*, "Using large language models to accelerate communication for eye gaze typing users with ALS," *Nature Communications*, vol. 15, 2024, doi: 10.1038/s41467-024-53873-3.

[15] R. J. K. Jacob, "The use of eye movements in human-computer interaction techniques: what you look at is what you get," *ACM Trans. Information Systems*, vol. 9, no. 2, pp. 152–169, 1991.

[16] P. Majaranta, I. S. MacKenzie, A. Aula, and K.-J. Räihä, "Effects of feedback and dwell time on eye typing speed and accuracy," *Universal Access in the Information Society*, vol. 5, no. 2, pp. 199–208, 2006.

[17] B. M. Velichkovsky, M. A. Rumyantsev, and M. A. Morozov, "New solution to the Midas touch problem," *Procedia Computer Science*, vol. 39, pp. 75–82, 2014.

[18] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications," in *Proc. CHI*, 2019, pp. 1–13.

[19] G. Chen *et al.*, "Evaluating Sensitivity Parameters in Smartphone-Based Gaze Estimation," arXiv:2506.11932, 2025.

[20] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[21] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.

[22] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems IV*, P. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[23] R. Szeliski, *Computer Vision: Algorithms and Applications*, 2nd ed. London, U.K.: Springer, 2022.

[24] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A Multiscale Retinex for Bridging the Gap Between Color Automatic Gain Control and Color Constancy," *IEEE Trans. Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[25] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," in *Proc. 6th IEEE ICCV*, 1998, pp. 839–846.

[26] W. K. Pratt, *Digital Image Processing*, 4th ed. Hoboken, NJ, USA: Wiley, 2007.

[27] T. S. Huang, *Two-Dimensional Digital Signal Processing II: Transforms and Median Filters*. Berlin, Germany: Springer-Verlag, 1981.

[28] OpenCV, "Geometric Image Transformations," 2026. [Online]. Available: https://docs.opencv.org/4.x/da/d54/group__imgproc__transform.html. Accessed: Jul. 29, 2026.

[29] ITU-R, "Recommendation ITU-R BT.709-6: Parameter Values for the HDTV Standards," International Telecommunication Union, 2015.

[30] J. S. Nascimento, K. do Valle, P. S. O. Coelho, M. B. Pelosi, and T. R. M. dos Santos, *Caderno de Comunicação Alternativa para Pacientes com COVID-19*. Rio de Janeiro, Brasil: UFRJ, 2020.

[31] F. VanderWerf, P. Brassinga, D. Reits, M. Aramideh, and B. Ongerboer de Visser, "Eyelid movements: behavioral studies of blinking in humans under different stimulus conditions," *J. Neurophysiology*, vol. 89, no. 5, pp. 2784–2796, 2003.

[32] K. A. Kwon *et al.*, "High-speed camera characterization of voluntary eye blinking kinematics," *J. Royal Society Interface*, vol. 10, no. 85, 20130227, 2013.

[33] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[34] K. Holmqvist, M. Nyström, and F. Mulvey, "Eye tracker data quality: what it is and how to measure it," in *Proc. ETRA*, 2012, pp. 45–52.

Identifying problems and gaps in the ICU early mobilization care process: development of practice oriented solutions in rehabilitation engineering

Ana Lays Braga
Universidade Federal do Pernambuco - UFPE
Recife, Brasil
ana.lays@ufpe.br

Alana Elza Fontes Gama
Universidade Federal do Pernambuco - UFPE
Recife, Brasil
alana.elza@ufpe.br

Abstract*—This study aimed to map the routines, barriers, and suggestions for improvement in early mobilization in intensive care units (ICUs) from the perspective of physical therapists, within a Design Science Research (DSR) framework. A qualitative, exploratory-descriptive study was conducted with 20 physical therapists from six Brazilian cities through semistructured video interviews. Data were analyzed using Bardin's Content Analysis, resulting in 66 codes grouped into 19 thematic categories across four research questions. Results indicate that clinical stability and hemodynamic safety criteria guide the prescription of therapeutic exercise, consistent with international guidelines. However, the absence of formalized protocols, fragmented documentation, limited equipment, and gaps in interprofessional collaboration represent significant barriers. Participants suggested the development of structured protocols, digital monitoring platforms, assistive devices, and multiprofessional communication tools. The contributions of this study encompass the systematized mapping of gaps, barriers, and problems that may guide innovation proposals in biomedical engineering aimed at early rehabilitation. These findings provide an empirical basis for biomedical engineering recommendations, representing the immersion phase of the DSR cycle toward innovative solutions for early rehabilitation in Brazilian ICUs.

Keywords*—early mobilization; intensive care unit; physical therapy; Design Science Research; Design Thinking; biomedical engineering

I. INTRODUCTION

Care and treatment in the Intensive Care Unit (ICU) have evolved significantly in recent years, moving beyond the paradigm of deep sedation and prolonged bed rest toward approaches that improve clinical outcomes, including a focus on early mobilization and physical rehabilitation, with the goal of reducing the risk of post-intensive care sequelae [1].

Functional decline during ICU hospitalization can be prevented through interventions involving physical activity and early mobilization practices. Although guidelines for physical activity in the ICU exist, the timely identification of patients eligible for these interventions remains a challenge, and a variety of obstacles to their implementation persist [2].

Evidence on early exercise for the prevention of ICU-acquired muscle weakness (ICUAW) points to heterogeneity in implementation across different units, stemming from factors such as excessive workload, insufficient staffing, limited time, inadequate training, and the absence of standardized protocols [3].

The implementation of early mobilization in ICUs remains inconsistent and faces barriers at multiple levels, notably a lack of familiarity with the scientific literature, underestimation of the incidence of ICUAW, the absence of routine assessment for mobilization, and the perception that mobility is not a care priority [4].

Although the benefits of early mobilization are widely documented, its systematic implementation remains inconsistent across different healthcare contexts and systems, highlighting a persistent gap between available scientific knowledge and actual clinical practice [3].

In recent years, technological solutions have been increasingly incorporated into intensive care settings to support clinical decision-making, but few are specifically designed for early mobilization or physical therapy practice; most target other domains of critical care, such as hemodynamic monitoring and sedation control [5]. This absence of practice-oriented technological support reinforces the need to investigate, in depth, the problems and gaps that hinder the implementation of early mobilization before proposing new solutions.

Research grounded in Design Science Research (DSR) enables researchers to construct, evaluate, and theorize solutions, differing from traditional methods by initially focusing on the identification of complex problems. The design science paradigm seeks to expand the limits of human and organizational capabilities through the creation of new and innovative artifacts [6].

Given the importance of early rehabilitation in the ICU for functional prognosis and the reduction of complications during hospitalization. This study aimed to map the complexity of physiotherapy practice in the intensive care setting, with an emphasis on early mobilization, by identifying problems and gaps in the care process from the perspective of intensive care physical therapists, with a view to informing the subsequent development of practice-oriented solutions tailored to the reality of intensive care units. Problem identification is a critical first step in guiding the development of effective solutions in rehabilitation engineering, as it ensures that subsequent design efforts are grounded in the real needs.

The contributions of this study encompass the systematized mapping of gaps, barriers, and problems that may guide innovation proposals in biomedical engineering aimed at early rehabilitation.

II. METHODOLOGY

This is an exploratory-descriptive qualitative study situated within the field of Design Science Research applied to healthcare specifically in the context of intensive care.

This study corresponds to the initial phase of the research process, focused on problem identification and awareness, which precedes the design and evaluation of technological artifacts. The findings obtained at this stage are intended to inform and guide the subsequent search for practice-oriented solutions capable of addressing the barriers identified in the early mobilization care process.

The research was carried out with intensive care physiotherapists, where participants were recruited through invitations sent via messaging applications and social media, targeting professionals working in the field of intensive care physical therapy.

A non-probabilistic sampling technique known as snowball sampling [7] was adopted, whereby each participant was invited to refer others with similar characteristics. This study included the participation of 20 intensive care physical therapists. Although the literature does not specify a minimum sample size for qualitative research, this number exceeds the data saturation parameters suggested by [8], which indicate that saturation can be achieved with 6, 12, or 17 participants, reinforcing the adequacy of the sample for the purposes of this study.

The inclusion criteria were: being a licensed physiotherapist with at least one year of ICU experience, including day and night shift staff affiliated with public and private intensive care units. No exclusion criteria were established.

After providing free and informed consent, participants completed an identification questionnaire, followed by semi-structured interviews conducted via Google Meet between March and May 2026. The interviews were recorded and transcribed. Each participant's speeches were tabulated in an Excel spreadsheet for subsequent Content Analysis. All interviews were conducted by the same researcher to ensure consistency in data collection. The interview guide was structured as presented in Table 1.

TABLE I. SEMI-STRUCTURED INTERVIEW GUIDE

No.	Topic	Guiding Question
Q1	Prescription criteria	What criteria do you use when prescribing therapeutic exercise for patients in the ICU? What most influences your decision?
Q2	Assessment tools and protocols	Does the unit where you work have specific tools or protocols for assessing, recording, and monitoring exercise progress in the ICU? How are these tools used in practice?
Q3	Frustrations with the current process	Is there anything that frustrates you about the current process of implementing therapeutic exercise in the ICU?
Q4	Suggestions for improvement	If you could suggest something to improve the applicability of therapeutic exercises in the ICU, what would it be?

Q = Question - Source: prepared by the author (2026)

Data from the identification form were analyzed using descriptive statistics with the aid of JASP software. Categorical variables were expressed as absolute and relative frequencies (n; %), while continuous variables were expressed as mean and range.

Content analysis, as proposed by Bardin [9], was applied following three phases: pre-analysis, exploration of the material, and treatment of the results/inference/interpretation. In the pre-analysis phase, the interview recordings were fully transcribed and read repeatedly (floating reading) to allow familiarization with the content and identification of preliminary patterns. The interview transcripts were organized and tabulated in an Excel spreadsheet, with coding and categorization performed manually by a single researcher (first author), who conducted all interviews.

Sample size was defined a priori based on the literature on data saturation in qualitative research [8] and was subsequently corroborated empirically: during the final interviews, no new codes emerged, and participants' responses increasingly converged around previously identified patterns, indicating that thematic saturation had been reached.

In the exploration of the material stage, the transcripts were coded through the identification of recording units, understood as text segments carrying meaning relevant to the phenomenon under investigation. These units were extracted, grouped, and classified according to their thematic and conceptual proximity, giving rise to initial codes. The codes were subsequently submitted to successive rounds of analysis and refinement, being aggregated into subcategories and, later, into broader thematic categories. The categorization process observed the criteria of homogeneity, relevance, exclusiveness, objectivity, and productivity established by Bardin.

Finally, in the treatment of results, inference, and interpretation stage, the thematic categories were analyzed in light of the study's objectives and its four guiding research questions (P1–P4). As a complementary strategy for organizing and presenting the data, the frequency of occurrence of codes and subcategories was calculated, allowing the recurrence of themes across the interviews to be identified. For each category, representative recording units were selected to illustrate the meanings attributed by participants and to lend transparency to the interpretive process. The results were discussed in dialogue with the scientific literature on early mobilization in intensive care units, promoting theoretical triangulation and broadening the understanding of the phenomenon under investigation.

The study was approved by the Research Ethics Committee under Opinion No. 8,061,868, in accordance with the ethical principles governing research involving human subjects.

III. RESULTS

Twenty physical therapists from six Brazilian municipalities were interviewed. Juazeiro do Norte accounted for the largest portion of the sample (n=9; 45%), followed by Fortaleza and Crato (n=3 each; 15%), Recife and Quixeramobim (n=2 each; 10%), and Brasília (n=1; 5%). The mean duration of professional training was 10.9 years (range: 2–20 years), while the mean duration of practice specifically in intensive care physical therapy was 8.8 years (range: 1.5–15 years).

Most participants worked in public hospitals (n=13; 65%), while the remainder were affiliated with private institutions (n=7; 35%); all worked on an on-call shift basis.

Content analysis yielded 66 codes grouped into 19 thematic categories, organized according to the four thematic

topics described below. For each of the study's four guiding research questions, the table presents the thematic category identified through content analysis, the frequency of occurrence of that category that is, the number of times the theme appeared in participants' statements and a representative recording unit, corresponding to an excerpt of speech that concretely illustrates the meaning attributed by participants to that category, thereby lending greater transparency and clarity to the interpretation of the data.

A. Clinical Decision-Making and Therapeutic Exercise Prescription in the ICU

Physical therapists identified hemodynamic stability as the primary factor in clinical decision-making for the prescription of therapeutic exercise in the ICU, followed by the use of standardized assessment tools, institutional protocols, and the patient's ability to cooperate with and accept the proposed interventions (Table II).

It should be noted that the frequency reported in each table corresponds to the number of recording units coded within each category, rather than the number of distinct participants who addressed that theme. As a single participant could generate more than one recording unit within the same category — for instance, when a given criterion was mentioned at different moments throughout the interview — frequencies may exceed the total sample size (n=20). This distinction is consistent with Bardin's content analysis framework, in which the recording unit, not the respondent, constitutes the base unit of quantification.

TABLE II. CRITERIA FOR PRESCRIBING THERAPEUTIC EXERCISE IN THE ICU (QUESTION 1)

Category	Freq.	Representative recording unit
Hemodynamic stability	23	*"Hemodynamic stability as defined by the mobilization protocol [...] FiO2 below 40, P/F greater than 200, Hb above 7, no active bleeding, low-dose vasoactive drugs, heart rate below 110, no arrhythmias. (F15)"*
Application of functional scales	9	*"We assess muscle strength, the IMS scale, to determine the degree of functionality. And the trunk control test, along with muscle strength tests, to define the level of mobilization. (F20)"*
Institutional protocol criteria	3	*"We use our mobilization protocol. Patients who are not eligible are placed on the red list; those on green or yellow are assessed and a decision is made. (F10)"*
Patient cooperation	4	*"We check whether the patient is fit to receive care, whether they consent—the code of ethics is also important—and we verify vital signs and level of consciousness. (F7)"*

Participants identified by code (F1–F20); Number of times the theme appeared in participants statements.Source: Research data

B. Tools and Protocols for Exercise Progression

The four categories identified in theme 2 are described in Table III. It is noteworthy that the first three categories showed a frequency of 20 recording units, a value numerically equivalent to the total number of participants which reflects the number of times the code was identified in the material, and not necessarily the number of distinct physical therapists who addressed these aspects.

TABLE III. TOOLS AND PROTOCOLS FOR EXERCISE PROGRESSION IN THE ICU (QUESTION 2)

Category	Freq.	Representative recording unit
Existence and formalization of protocols	20	*"The protocol has three stages: red, when mobilization is not possible at all; yellow, with caution; and green, when mobilization can proceed without contraindication. We use the patient's IMS. (F15)"*
Recording tools	20	*"We have daily progress notes and a daily mobilization goal that must be met. If not met, there must be a documented justification. This is recorded in the medical chart and on the monitoring form. (F17)"*
Assessment tools	20	*"We assess the MRC and IMS daily—these are the two main tools: the strength scale and the functional progression scale, showing where the patient is at a given moment and where we want them to be. (F5)"*
Progression and continuity dynamics	8	*"Since each ICU has a daily physical therapist, we use progression markers—this is today's goal—and the daily therapist ensures continuity of care. (F12)"*

Participants identified by code (F1–F20); Number of times the theme appeared in participants statements.Source: Research data

C. Frustrations with the Implementation of Early Mobilization

When reporting their frustrations, three categories emerged with equal frequency (n=7): collaboration and recognition from the interdisciplinary team, lack of equipment and assistive devices, and absence of workflow and guidance on procedures. These findings reveal multifactorial frustrations encompassing human, structural, and organizational factors (Table IV).

TABLE IV. FRUSTRATIONS IN THE PROCESS OF IMPLEMENTING THERAPEUTIC EXERCISE IN THE ICU (QUESTION 3)

Category	Freq.	Representative recording unit
Collaboration and recognition by the interdisciplinary team	7	*"Early mobilization is not carried out by the physical therapist alone—it is an interdisciplinary process and the team needs to understand that. (F8)"*
Lack of equipment and assistive devices	7	*"What is frustrating is the lack of even the most basic conditions: ankle weights, dumbbells, resistance bands, exercise balls, and bedside sitting chairs. (F14)"*
Staffing levels	3	*"Here we have one physical therapist for every [several] patients—all in very serious condition. The shift is hectic, and sometimes we could do more, but the staffing levels do not allow it. (F11)"*

Category	Freq.	Representative recording unit
Absence of workflow and guidance on procedures	7	*"Continuity only works when there is a resident. What is frustrating is the inconsistency—if there were a mandatory protocol to follow, it would be different. (F3)"*
Bureaucracy in record-keeping and inadequate follow-up	2	*"The bureaucratic part is very repetitive—we record the same information in different tools: incident reports, progress notes, and shift handoffs. (F6)"*

Participants identified by code (F1–F20); Number of times the theme appeared in participants statements.Source: Research data

D. Suggestions for Improving the Physical Therapy Routine

The categories that emerged from the participants' suggestions are directly related to the frustrations identified in the previous section. The highest-frequency category, "standardization and structured documentation" (n=13), encompasses suggestions ranging from protocols guiding clinical practice to the incorporation of the International Classification of Functioning (ICF) into clinical practice (Table V).

TABLE V. SUGGESTIONS FOR IMPROVING THE APPLICABILITY OF THERAPEUTIC EXERCISES IN THE ICU (QUESTION 4)

Category	Freq.	Representative recording unit
Standardization and structured documentation	13	*"It would be an individualized prescription based on functional assessments, with specific therapeutic goals, periodic reassessments, and load progressions. (F2)"*
Assistive devices and assistive technology	7	*"I would provide stretchers that keep the patient in an upright position. The basics are also missing: ankle weights, dumbbells, resistance bands, and appropriate sitting chairs. (F15)"*
Daily physical therapist and supervision	4	*"With the daily physical therapist, we had daily follow-up—this greatly improved functional progress and the continuity of therapeutic goals. (F5)"*
Culture of mobilization	3	*"There are still limitations in physical therapist training—they are too focused on mechanical ventilation, with gaps when it comes to therapeutic exercise. (F6)"*
Communication and integration with the multidisciplinary team	5	*"Brief meetings with the team—physician, nurse, and technician—to explain the importance of mobilization and introduce the ICF into the physiotherapy diagnosis. (F10)"*

Participants identified by code (F1–F20); Number of times the theme appeared in participants statements.Source: Research data

IV. DISCUSSION

This study is situated within the immersion phase of the DSR cycle, characterized by the systematic identification of problems, meaningful engagement with the professionals involved, and the mapping of contextual needs—elements that are essential for the development of innovative and technologically relevant artifacts.

The criteria reported by physical therapists for the prescription of therapeutic exercise align with the safety domains established in the literature. The "hemodynamic stability" category (n=23) corresponds directly to the cardiovascular, respiratory, and neurological criteria systematized through expert consensus [10]. This alignment suggests that, even in the absence of formalized protocols (n=3), physical therapists base their clinical decision-making on parameters consistent with international safety recommendations, albeit in an unsystematic manner.

The use of the ICU Mobility Scale (IMS) and the Medical Research Council Scale (MRC) was identified as the second most frequent category (n=9) for guiding therapeutic exercise prescription. A multicenter observational study demonstrated that the IMS score at ICU discharge is independently associated with 90-day survival [11]. However, no explicit correspondence was identified between the scores obtained on these scales and the type of exercise actually prescribed, suggesting that these tools serve a predominantly classificatory function—characterizing the patient's level of muscle strength and functional capacity—without constituting formal prescription guidelines. Brazilian researchers are advancing in this direction by proposing a framework that integrates IMS scores with CBDF (Brazilian Classification of Physical Therapy Diagnoses) diagnoses and progressive therapeutic plans based on corresponding clinical interventions [12].

Findings regarding protocol existence reveal institutional heterogeneity: ten participants reported the absence of a mobilization protocol, and eight described an implemented protocol. Protocols employing standardized safety guidelines have been shown to increase early exercise rates in the ICU [13]. The absence of a common language among assessment tools hinders longitudinal patient follow-up and is associated with worse clinical outcomes and longer hospital stays [12].

Collaboration among the multidisciplinary team emerged both as a source of frustration and as a solution proposed by participants. It is recommended that the ICU multidisciplinary team assess patients, develop individualized rehabilitation programs, and promote shared decision-making, as interdisciplinary collaboration is central to the safe and effective implementation of early exercise [13].

Regarding staffing, the European Society of Intensive Care Medicine recommends one specialist physical therapist per five ICU beds, with 24-hour, 7-day coverage [14], while Brazil's RDC No. 7/2010 establishes a minimum of one physical therapist per ten beds, with coverage across morning, afternoon, and evening shifts, totaling 18 daily hours of physical therapy care [15]. Although the reported staffing levels generate dissatisfaction, they remain within the parameters established by current Brazilian regulations.

The availability of appropriate equipment is a structural prerequisite for the progressive implementation of therapeutic exercise. Electrical stimulation combined with a cycle ergometer has demonstrated benefits regarding muscle strength, cross-sectional muscle area, ventilator-free days, cognitive function, delirium, and quality of life [16]. Emerging technologies such as robot-assisted mobilization and virtual reality expand this horizon, offering promising opportunities for rehabilitation in the ICU setting with the potential to reduce risks and staffing requirements during mobilization [17]. The European expert guideline on positioning and early mobilization incorporated specific recommendations for devices and robotics, formally

recognizing the role of these technologies in the rehabilitation of critically ill patients [18].

The findings reveal significant internal consistency between the identified gaps, reported frustrations, and proposed solutions. Recognizing and thoroughly understanding the problems in a given area is an indispensable prerequisite for directing relevant and contextually appropriate interventions [6].

The study presents limitations, such as the geographical factor with the majority of participants hailing from the Northeast region, specifically the state of Ceará, and the use of snowball sampling, where participants tend to refer colleagues with similar routines. Furthermore, it reflects only the perspective of the physiotherapists, excluding the intensive care unit's interdisciplinary team. In addition, the content analysis was conducted by a single researcher, which may represent a limitation regarding triangulation and cross-validation of the coding process.

V. DESIGN REQUIREMENTS FOR BIOMEDICAL ENGINEERING

Following the Content Analysis of the interviews, classes of problems were identified based on the thematic categories of frustration and suggestions for improvement reported by the participating physical therapists, with the most frequently occurring ones listed in Table IV, which highlight the main gaps experienced in the clinical practice of early mobilization in the ICU. From these needs, design requirements were derived to guide the development of future solutions, without prescribing, at this stage of the research, the specific technology or format of the solution, as shown in Table VI.

TABLE VI. SYNTHESIS OF IDENTIFIED NEEDS AND CORRESPONDING DESIGN REQUIREMENTS

Problem	Corresponding design requirement
Absence of workflows and guidance on procedures	Interoperability and communication between systems; Creation of protocols and conduct guidelines
Collaboration and recognition from the multidisciplinary team	Facilitating interdisciplinary communication; Ensuring clarity in shift handoffs regarding mobilization / mobilization level
Insufficiency of equipment and technologies to support mobilization	Equipment designed for quick setup and operation within the confined physical space of an ICU bed. Equipment capable of supporting progressive levels of mobilization (from passive to active-resisted exercise), accompanying the patient's functional evolution.
Excessive bureaucracy without effective data	Automation of process steps
Lack of objective measures in patient assessment	Objective assessments and metrics for ICU mobilization

Source: Research data

VI. CONCLUSION

The data produced in this study indicate that the main problems identified by physical therapists in the process of implementing therapeutic exercise in the ICU are concentrated in four areas: the absence of formalized institutional protocols and structured documentation tools that ensure the continuity and progression of interventions across shifts; the insufficiency of equipment and devices to support mobilization; weaknesses in collaboration and integration within the multidisciplinary team; and an inadequate ratio of professionals to patients. These findings provide direction for the subsequent stages of the design cycle, in which the perceptions and suggestions of the participating physical therapists may inform the development, implementation, and evaluation of innovative solutions aimed at improving the applicability of early mobilization in Brazilian ICUs.

AI USAGE DECLARATION

The written content, analysis, interpretation, and discussion of the results were carried out exclusively by the authors. During the preparation of the manuscript, the authors used the artificial intelligence tool Claude to assist with grammar, improve the writing, and verify the text's compliance with the journal's template structure. The authors critically reviewed all suggestions and bear full responsibility for the content presented.

ACKNOWLEDGMENT

The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the financial support provided for the development of this research.

REFERENCES

[1] M. E. Kho and B. Connolly, "From Strict Bedrest to Early Mobilization," Crit. Care Clin., vol. 39, no. 3, pp. 479-502, Jul. 2023, doi: 10.1016/j.ccc.2023.01.003.

[2] A. Krupp, H. Dunn, L. Knake, P. Ten Eyck, L. Frey-Law, and K. Dunn Lopez, "Aggregating Patient Safety and Status Information in the Electronic Health Record to Support Time-Sensitive Mobility Interventions in the Intensive Care Unit: Protocol for the Design and Testing of a Clinical Decision Support Tool," JMIR Res. Protoc., vol. 14, p. e75752, Oct. 2025, doi: 10.2196/75752.

[3] G. Ballesteros-Reviriego and J.-D. Marti, "Physiotherapy in the ICU: Past, present, and future," Med. Intensiva Engl. Ed., vol. 49, no. 12, p. 502205, Dec. 2025, doi: 10.1016/j.medine.2025.502205.

[4] I. du Plessis, S. Francis, and B. Morrow, "Early mobilization in Windhoek intensive care units: Practices, attitudes and barriers," South Afr. J. Physiother., vol. 81, no. 1, p. 2118, 2025, doi: 10.4102/sajp.v81i1.2118.

[5] K. D. Tzimourta, "Human-Centered Design and Development in Digital Health: Approaches, Challenges, and Emerging Trends," Cureus, vol. 17, no. 6, p. e85897, Jun. 2025, doi: 10.7759/cureus.85897.

[6] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Q. Manag. Inf. Syst., vol. 28, no. 1, pp. 75-105, 2004, doi: 10.2307/25148625.

[7] L. Goodman, "Snowball Sampling," Ann. Math. Stat., vol. 32, pp. 148-170, Mar. 1961, doi: 10.1214/aoms/1177705148.

[8] B. Saunders et al., "Saturation in qualitative research: exploring its conceptualization and operationalization," Qual. Quant., vol. 52, no. 4, pp. 1893-1907, 2018, doi: 10.1007/s11135-017-0574-8.

[9] L. Bardin, Content Analysis. Lisbon: Editions 70, 2016.

[10] C. L. Hodgson et al., "Expert consensus and recommendations on safety criteria for active mobilization of mechanically ventilated critically ill adults," Crit. Care, vol. 18, no. 6, p. 658, 2014, doi: 10.1186/s13054-014-0658-y.

[11] C. J. Tipping et al., "The ICU Mobility Scale Has Construct and Predictive Validity and Is Responsive. A Multicenter Observational Study," Ann. Am. Thorac. Soc., vol. 13, no. 6, pp. 887-893, Jun. 2016, doi: 10.1513/AnnalsATS.201510-717OC.

[12] L. R. M. Zego and C. H. V. da Costa, "Integration of Hospital Mobility Scales with the International Classification of Functioning and the Brazilian Classification of Physical Therapy Diagnoses: Creation of Clusters," Rev. Bras. Saude, vol. 2, no. 7, Apr. 2026, doi: 10.22533/at.ed.8089272608046.

[13] X. Bai, C. Gu, Y. Li, X. Jin, J. Liu, and C. Guo, "Best Evidence Summary on Early Exercise for Prevention of ICU-Acquired Weakness: An Evidence-Based Synthesis," J. Multidiscip. Healthc., vol. 19, p. 574831, 2026, doi: 10.2147/JMDH.S574831.

[14] A. Valentin, P. Ferdinande, and the ESICM Working Group on Quality Improvement, "Recommendations on basic requirements for intensive care units: structural and organizational aspects," Intensive Care Med., vol. 37, no. 10, pp. 1575-1587, Oct. 2011, doi: 10.1007/s00134-011-2300-7.

[15] ANVISA, "Resolution-RDC No. 7, of February 24, 2010," Brasilia, 2010. [Online]. Available: https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2010/res0007_24_02_2010.html

[16] A. Nazir, "The effect of a combination of functional electrical stimulation and cycle ergometer (FES-cycling) on physiological changes and functional ability in patients with ICU-acquired weakness," Anaesth. Pain Intensive Care, vol. 27, no. 5, pp. 599-606, Jul. 2023, doi: 10.35975/apic.v27i5.2315.

[17] C. Parco, V. Kreuels, M. Kelm, C. Jung, and G. Wolff, "Robotic-assisted early mobilization and virtual reality: a perspective on innovative support strategies for critically ill patients," Intensive Care Med. Exp., vol. 11, no. 1, p. 86, Dec. 2023, doi: 10.1186/s40635-023-00571-x.

[18] S. J. Schaller et al., "Guideline on positioning and early mobilization in critically ill patients by an expert panel," Intensive Care Med., vol. 50, no. 8, pp. 1211-1227, Aug. 2024, doi: 10.1007/s00134-024-07532-2.

IMPACT OF SERIOUS GAMES ON GLOBAL FUNCTIONALITY AND THERAPEUTIC ENGAGEMENT IN POST-STROKE INDIVIDUALS: A SYSTEMATIC REVIEW

Ana Carolina Borges Valente[1]
Department of Education and Technology, Federal Institute of Science and Technology,
Juiz de Fora, Brazil
ana.valente@estudante.ufjf.br

Thiago Rodrigues de Oliveira[1]
Department of Education and Technology, Federal Institute of Science and Technology,
Juiz de Fora, Brazil
tiago.if.jf@gmail.com

Silvana Terezinha Faceroli[1]
Department of Education and Technology, Federal Institute of Science and Technology,
Juiz de Fora, Brazil
sfaceroli@gmail.com

Maria Eduarda Hasten Reiter Moreira[2]
Juiz de Fora School of Medical and Health Sciences,
Juiz de Fora, Brazil
mehasten.fisio@gmail.com

Sara Del Vecchio[1]
Department of Education and Technology, Federal Institute of Science and Technology,
Juiz de Fora, Brazil
sara.vecchio.ere@gmail.com

Abstract*—Stroke remains one of the leading causes of long-term disability worldwide, frequently resulting in persistent impairments in motor performance, mobility, independence, and quality of life. Digital rehabilitation technologies, including virtual reality, serious games, and exergames, have emerged as promising strategies for enhancing rehabilitation by integrating task-oriented training, real-time feedback, and interactive environments that increase patient engagement and practice intensity. This systematic review aimed to synthesize current evidence on applying these technologies in post-stroke rehabilitation, comparing studies by methodological design (randomized controlled trials, pilot studies, and mixed methods/feasibility studies) and technological modality. We conducted a systematic search in PubMed, Scopus, Embase, Web of Science, and IEEE Xplore, following PRISMA 2020 guidelines and registering the protocol in PROSPERO. Seventeen studies met the eligibility criteria and were analyzed regarding their technological characteristics, intervention protocols, and functional outcomes. Overall, most studies reported favorable effects of digital technologies on motor function, balance, mobility, activities of daily living, and quality of life, although not all included a control group. Most interventions used markerless motion capture systems, combined with real-time visual feedback and task-oriented exercises. Studies also frequently reported high usability, motivation, and treatment adherence, although they measured these outcomes with different instruments, limiting direct comparisons between these secondary outcomes and the primary clinical outcomes. The methodological quality and risk of bias of the 17 included studies were assessed using the Joanna Briggs Institute (JBI) critical appraisal tools, revealing substantial heterogeneity in methodological rigor, particularly regarding the blinding of participants and therapists and allocation concealment; no meta-analysis was performed. Substantial heterogeneity in technological platforms, intervention protocols, study designs, and outcome measures limits the strength of conclusions about the clinical effectiveness of these technologies. Collectively, the findings suggest that digital rehabilitation technologies are promising complementary tools that may support motor learning and experience-dependent neuroplasticity, although the included studies did not directly measure these mechanisms, while also promoting more engaging and personalized rehabilitation. Heterogeneity among studies, predominantly small sample sizes, and the lack of long-term follow-up still limit the ability to draw robust conclusions. Future studies should prioritize standardized study protocols, individualized reporting of sample size, duration, and comparator groups, and comprehensive outcome measures to strengthen the clinical evidence and guide the development of intelligent rehabilitation systems.

Keywords—Stroke; Serious Game; Rehabilitation; Functional Recovery; Therapeutic Engagement; Virtual Reality; JBI.

I. INTRODUCTION

Stroke remains among the leading causes of mortality and long-term disability worldwide, representing a major public health challenge due to its clinical, social, and economic impact. Although advances in acute-phase treatment have contributed to reduced mortality, a significant proportion of survivors are left with persistent motor, sensory, cognitive, and perceptual deficits, compromising functional independence, performance in activities of daily living (ADLs), social participation, and quality of life. In this context, rehabilitation plays a fundamental role in promoting functional recovery after stroke [1–4].

Functional recovery after stroke is strongly influenced by the central nervous system's capacity to reorganize and adapt; this ability stems from neuroplasticity mechanisms that depend on experience and environment. Experimental and clinical evidence shows that this reorganization is favored by interventions based on intensive, repetitive, task-oriented, and functionally meaningful practice, combined with the provision of adequate feedbackprinciples considered fundamental to motor learning and neurorehabilitation [5–8]. However, implementing these principles in conventional rehabilitation still faces several challenges. Limited therapy time, low practice intensity, and a progressive decline in patient motivation frequently compromise treatment effectiveness, especially in the chronic phase of stroke, when spontaneous neurological recovery is already considerably reduced [4,5].

In this scenario, digital rehabilitation technologies have emerged as promising tools for enhancing neurological rehabilitation. These technologies encompass a wide range of interactive systems, including virtual reality (VR), serious games, exergames, motion capture systems, and other computational platforms geared toward rehabilitation [9–13]. While serious games are digital applications developed for objectives beyond entertainment, such as education, training,

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

or health, exergames are a subcategory characterized by the need for body movement to interact with the virtual environment [9–11]. By integrating physical activity with gamification elements, these technologies can increase repetitions, motivation, engagement, and treatment adherence characteristics particularly relevant to rehabilitation after stroke [12,13].

Furthermore, these technological resources offer therapeutic environments that incorporate principles essential to motor learning. Most systems provide real-time visual or multimodal feedback, task-oriented exercises, automatic difficulty-progression mechanisms, and continuous performance monitoring, allowing training to be tailored to each individual's functional capacities. These features favor active patient participation, facilitate error correction during task execution, and increase practice intensity, creating conditions favorable to experience-dependent neuroplasticity and functional recovery [5,6,12,14].

Previous studies show that digital rehabilitation technologies can improve motor function, balance, gait, upper-limb performance, and functional independence after stroke, especially when used alongside conventional physical therapy [13,15]. However, important challenges remain. Studies vary widely in the technological platforms used, intervention protocols, training intensity, and assessment instruments, which hinders direct comparisons. Moreover, most research focuses on evaluating outcomes related to body functions and structures, while aspects such as activity, participation, quality of life, motivation, usability, and user experience remain underexplored.

Thus, this systematic review aimed to synthesize the available evidence on the application of digital rehabilitation technologies, including virtual reality, serious games, exergames, and other interactive systems, in the rehabilitation of individuals after stroke. It also sought to characterize the main technological aspects of these interventions and analyze their effects on different functional outcomes, comparing studies by methodological design and technological modality, thereby providing evidence to support both clinical decision-making and the development of new technologies for neurological rehabilitation.

II. Methods

A. Study Design

This study consists of a systematic literature review conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) recommendations and registered in the International Prospective Register of Systematic Reviews (PROSPERO) under number CRD420261338426.

B. Procedures

Articles published up to March 2026 were included, and the systematic literature search was conducted in the following databases: PubMed (https://www.ncbi.nlm.nih.gov/pubmed), Scopus (https://www.scopus.com), Embase (https://www.embase.com), Web of Science (https://login.webofknowledge.com/), and IEEE Xplore (https://ieeexplore.ieee.org). The search terms used were: (“stroke” OR “post-stroke”) AND (“serious games” OR “video games” OR exergames OR “game-based rehabilitation”) AND (rehabilitation OR physical therapy OR neurorehabilitation) AND (“activities of daily living” OR functionality OR “functional recovery” OR engagement OR adherence OR motivation). The full search syntax for each database is available at https://drive.google.com/drive/folders/1ZRlrZIrXLnQtFtQTJx893XhbSIQBe5wy?usp=drive_link. The search was conducted on March 12, 2026, and the results obtained were synthesized and used in the present review.

Keyword selection for the database searches followed the PICOS framework. Population (P): individuals diagnosed with stroke (ischemic or hemorrhagic); Intervention (I): serious games, exergames, or digital games for therapeutic purposes; Comparison (C): conventional physical therapy, conventional rehabilitation, control group, or pre- and post-intervention comparison; Outcomes (O): improvement in overall functionality and activities of daily living and, as secondary outcomes, therapeutic engagement, motivation, and treatment adherence. We included all study types except systematic reviews, meta-analyses, preprints, and conference proceedings.

We initially organized data extracted from the different databases into separate Microsoft Excel spreadsheets and then merged them into a single spreadsheet to identify and remove duplicate articles. We then exported the studies to the Rayyan platform (https://www.rayyan.ai/), which we used to organize the screening process, manage study selection, and ensure reviewer blinding. We screened titles and abstracts according to the established inclusion and exclusion criteria. Studies involving adults diagnosed with stroke undergoing interventions based on serious games, exergames, or digital games for physical therapy rehabilitation were included, provided they assessed outcomes related to functionality, motor performance, balance, gait, and functional independence, and were conducted in an outpatient setting; as a secondary outcome, studies assessing adherence or therapeutic engagement were also considered.

Studies were excluded if they: (1) were conducted in animals; (2) involved other neurological conditions, such as Parkinson's disease, traumatic brain injury, multiple sclerosis, or cerebral palsy; (3) used virtual reality without a game-based component; (4) were systematic reviews, meta-analyses, protocols, case studies, conference abstracts, or preprints; (5) did not include assessments of functionality or therapeutic engagement; (6) were conducted in intensive care units or exclusively at home; or (7) were conducted via telerehabilitation or without professional guidance. After initial screening, we assessed potentially eligible articles in full text using the same inclusion and exclusion criteria. To improve selection reliability, four reviewers independently and blindly assessed all titles, abstracts, and full texts in two pairs (Reviewer 1 and Reviewer 2; Reviewer 3 and Reviewer 4), with each pair responsible for all records at their respective stage. In cases of disagreement, an external researcher managed and resolved conflicts.

C. Assessment of Methodological Quality and Risk of Bias

Three reviewers independently assessed the methodological quality and risk of bias of the included studies; in cases of disagreement, a fourth reviewer broke the tie. We used Joanna Briggs Institute (JBI) critical appraisal tools selected according to each study's design. For randomized controlled trials and randomized controlled pilot trials, we applied the JBI Critical Appraisal Checklist for Randomized Controlled Trials (13 items), addressing randomization,

allocation concealment, blinding, baseline similarity between groups, complete follow-up, and intention-to-treat analysis. For non-randomized or single-arm studies, and for mixed-methods and feasibility studies without a control group, we applied the JBI Critical Appraisal Checklist for Quasi-Experimental Studies (9 items), marking items related to between-group comparison as not applicable when no control group was present. Each item was judged as "yes," "no," "unclear," or "not applicable," with no summary score used, as recommended by JBI. We resolved disagreements between raters by consensus or, when necessary, by consulting an additional reviewer. Table I summarizes judgments per study; Tables II and III present item-by-item judgments for each of the two critical appraisal tools. We excluded no study on methodological quality; instead, we used this assessment when interpreting the strength of the evidence discussed in Section IV.

Table I - Summary of the JBI critical appraisal by included study

Study Information				**JBI Critical Appraisal Results**			
Study	***Ref.***	***Design***	***JBI Tool***	***Yes***	***No***	***Unclear/ NA***	***JBI Decision***
Kostekçi et al. (2024)	[18]	Randomized controlled trial	RCT (13 items)	5	4	4	Include
Huber et al. (2022)	[19]	Randomized controlled trial	RCT (13 items)	11	2	0	Include
Aguilera-Rubio et al. (2020)	[20]	Randomized controlled trial	RCT (13 items)	11	2	0	Include
Alhwoaimel et al. (2024)	[21]	Feasibility study	Quasi-experimental (9 items)	5	2	2	Include
Martino Cinnera et al. (2021)	[22]	Randomized pilot trial	RCT (13 items)	10	2	1	Include
Ali et al. (2018)	[23]	Multicenter randomized controlled trial	RCT (13 items)	13	0	0	Include
Peláez-Vélez et al. (2021)	[24]	Randomized pilot trial	RCT (13 items)	9	4	0	Include
Amin et al. (2021)	[25]	Randomized controlled trial	RCT (13 items)	8	3	2	Include
Korkusuz et al. (2023)	[26]	Single-blind randomized controlled trial	RCT (13 items)	12	0	1	Include
Chong Li et al. (2023)	[27]	Randomized controlled trial	RCT (13 items)	12	0	1	Include
Anwar N et al. (2020)	[28]	Randomized controlled trial	RCT (13 items)	8	3	2	Include
Bouatrous et al. (2023)	[29]	Single-arm usability study, no control group	Quasi-experimental (9 items)*	2	3	4	Checklist partially applicable (see note)
Eichinger et al. (2021)	[30]	Non-randomized study	Quasi-experimental (9 items)	8	0	1	Include with caution
Anwar et al. (2021)	[31]	Randomized controlled trial	RCT (13 items)	7	3	3	Include
Marda & Thorat (2023)	[32]	Randomized clinical trial	RCT (13 items)	6	2	5	Include
Térémetz et al. (2019)	[33]	Blinded randomized controlled trial	RCT (13 items)	9	4	0	Include
Huber et al. (2021)	[34]	Feasibility study	Quasi-experimental (9 items)	5	3	1	Include

Table IV Item-by-item assessment — JBI Critical Appraisal Checklist for Quasi-Experimental Studies

Study Information	**JBI Critical Appraisal Results**													
Study (Ref.)	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***Decision***
Marda & Thorat [32]	U	U	U	N	N	U	Y	Y	Y	Y	U	Y	Y	Include
Ali et al. [23]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Include
Huber et al. [19]	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Include
Aguilera-Rubio et al. [20]	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Include
Cinnera et al. [22]	Y	U	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Include
Térémetz et al. [33]	Y	N	Y	N	N	Y	Y	Y	N	Y	Y	Y	Y	Include
Amin et al. [25]	Y	U	Y	N	N	U	N	Y	Y	Y	Y	Y	Y	Include
Peláez-Vélez et al. [24]	Y	N	Y	N	N	Y	N	Y	Y	Y	Y	Y	Y	Include
Korkusuz et al. [26]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	NA	Include
Chong Li et al. [27]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	NA	Include
Anwar N et al. [28]	Y	U	Y	N	N	Y	Y	U	N	Y	Y	Y	Y	Include
Anwar et al. [31]	Y	U	Y	N	N	Y	Y	U	N	Y	U	Y	Y	Include
Kostekçi et al. [18]	N	U	U	N	N	N	Y	Y	Y	Y	U	Y	U	Include

Color legend: green = Yes; red = No; yellow = Unclear; gray = Not applicable. Item legend (JBI RCT, 13 items): 1) True randomization; 2) Allocation concealment; 3) Groups similar at baseline; 4) Blinding of participants; 5) Blinding of those delivering treatment; 6) Blinding of outcome assessors; 7) Groups treated identically other than the intervention; 8) Complete follow-up/losses described; 9) Intention-to-treat analysis; 10) Outcomes measured in the same way; 11) Outcomes measured reliably; 12) Appropriate statistical analysis; 13) Appropriate trial design.

Study Information	JBI Critical Appraisal Results									
Study (Ref.)	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***Decision***
Eichinger et al. [30]	Y	Y	Y	Y	Y	Y	Y	Y	U	Include with caution
Huber et al. [34]	Y	Y	NA	N	N	N	Y	Y	Y	Include
Alhwoaimel et al. [21]	Y	Y	NA	N	N	NA	Y	Y	Y	Include
Bouatrous et al. [29]	U	NA	NA	N	N	U	Y	Y	N	Single-arm study (n=11) — partial checklist

Color legend: green = Yes; red = No; yellow = Unclear; gray = Not applicable. Item legend (JBI Quasi-experimental, 9 items): 1) Clear cause-and-effect relationship; 2) Comparable and similar participants; 3) Comparable care/treatment other than the intervention; 4) Presence of a control group; 5) Multiple pre/post measurements; 6) Complete follow-up/losses described; 7) Outcomes measured in the same way; 8) Reliable measurement of outcomes; 9) Appropriate statistical analysis.

Bouatrous et al. [29] (Multimedia Systems, 2023;29:3385–3402, doi:10.1007/s00530-023-01180-0): the article addresses a subjective usability/feasibility assessment in a single arm, with 11 stroke patients, no control group, and no repeated pre/post measures—only the descriptive application of the SUS (System Usability Scale) and the IMI (Intrinsic Motivation Inventory), with no inferential statistical test reported. Because there is no comparison (neither between groups nor pre/post), the JBI checklist for quasi-experimental studies is only partially applicable to this design; items 2 and 3 were marked as "Not applicable" for this reason.

III. Results

The initial search strategy identified 529 articles from the selected databases. After removing duplicates, 347 studies remained. During initial screening, we excluded 180 studies based on title assessment and 90 based on abstract review. We included 77 articles for full-text review after title and abstract screening. Of these, we excluded 60 full texts: 48 for not meeting the established eligibility criteria and 12 because we could not retrieve the full text, resulting in 17 studies included in the review. Fig. 1 shows the entire process.

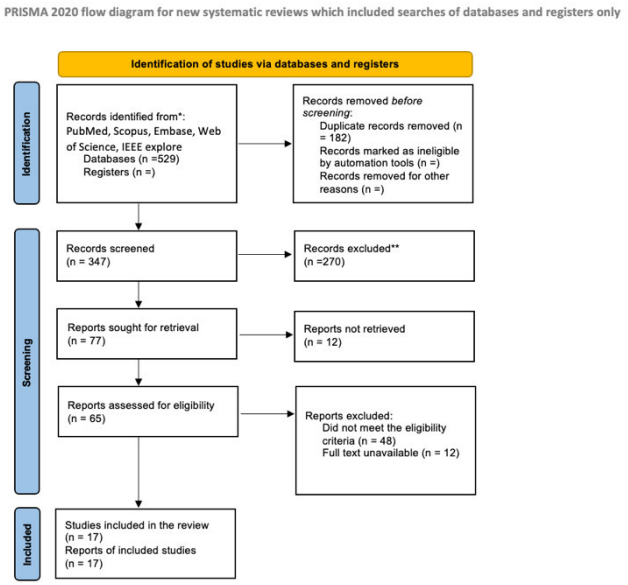


Fig. 1. Flow diagram of the identification, screening, eligibility, and inclusion process for studies in the systematic review, prepared in accordance with the PRISMA 2020 recommendations.

A. Characteristics of Participants in the Included Studies

The included studies included 10 to 120 participants, totaling 648 participants. Among studies that reported sex distribution, 203 participants were female and 347 were male. Participants' mean age ranged from 51 to 68 years. Overall, the samples were predominantly composed of individuals with chronic-phase ischemic stroke. However, some studies also included participants with hemorrhagic stroke in the subacute phase, while others did not specify the type of stroke or the time since the event. These findings highlight heterogeneity in the clinical characteristics of the included populations, which should be considered when interpreting the results.

B. Characteristics of the Interventions

Most interventions used motion capture technologies, predominantly featuring upper-limb rehabilitation and balance training. The therapeutic dose varied considerably across studies, ranging from 3 to 30 sessions, with session durations of 30 to 80 minutes. In virtually all trials, the experimental intervention was compared with conventional physical therapy.

C. Characteristics of the Technological Interventions and Comparison by Technological Modality

Table I presents the main characteristics of the technological interventions employed in the included studies, describing the platform used, the type of interaction between the user and the system, the available feedback mechanisms, and the therapeutic goals of each intervention. Analyzing these aspects helps us understand how different digital technologies have been applied to post-stroke rehabilitation and identify trends in the development of exergames, serious games, and virtual reality systems aimed at functional recovery. Motion-capture-based systems with real-time visual feedback and motor-practice-oriented tasks predominated. Less frequently, studies identified features such as auditory feedback, multisensory stimuli, and adaptive difficulty-adjustment mechanisms, reflecting different strategies for optimizing user–system interaction and promoting motor learning during rehabilitation.

TABLE IV. TECHNOLOGICAL CHARACTERISTICS OF THE DIGITAL INTERVENTIONS USED IN THE INCLUDED STUDIES

Author (Year)	Technology	Feedback	Mechanics	Adaptation
Kostekçi et al. (2024)	Leap Motion Controller ('CatchaPET' and 'LeapBall' software). Optical system featuring 3 infrared emitters and 2 cameras for markerless hand/finger tracking.	Visual	Immediate visual feedback focused on manual dexterity.	Sensitivity adjustment based on captured finger movements.
Huber et al. (2022)	Dividat Senso. Pressure-sensitive force platform with three-sided safety handrails positioned in front of an eye-level screen.	Visual	Serious games based on motor and cognitive challenges with level progression.	Automatic adaptive progression based on the patient's accuracy and response time.
Aguilera-Rubio et al. (2020)	Unity3D Game Engine integrated with the Leap Motion optical sensor.	Visual	Gamified motor challenges based on goals with level progression and scoring.	Customization of reaching range and virtual workspace according to patient deficits.
Alhwoaimel et al. (2024)	Valedo® System (Hocoma AG). 3 wireless inertial sensors (IMUs) attached to the trunk via straps for non-invasive angle and velocity measurement.	Visual & Auditory	Therapeutic exercises integrated into serious games with goals and difficulty levels.	Initial calibration of the patient's active trunk range of motion.
Cinnera et al. (2021)	Riablo™ (CoRehab). System composed of 5 inertial sensors (IMUs) and a wireless force platform connected via Bluetooth (50 Hz).	Visual & Auditory	Serious balance and coordination games with real-time audiovisual feedback of posture.	Initial postural calibration based on individual Center of Pressure (CoP).
Ali et al. (2018)	ArmAble™ Device. Mechanized, gamified upper limb training system tracking shoulder, elbow, forearm, and hand movements.	Visual, Acoustic & Haptic	Interactive games focused on functional repetition with incremental challenge levels.	Adjustable weight support and resistance settings to assist movement.
Vélez-Guerrero et al. (2021)	Oculus Quest 2 VR headset integrated with 'Blexer' software ('Phiby's Adventure' game), tracked by a Microsoft Kinect 360 v1 sensor.	Visual	Four immersive minigames performed in a standing (orthostatic) posture.	Tailoring of specific virtual scenarios monitored by the therapist.
Amin et al. (2021)	Oculus Quest 2 VR headset with native hand tracking executing APK-format games developed in Unity3D.	Visual	Direct real-time immersive interaction with virtual targets using the user's hands.	Real-time spatial kinematic mapping adapted to the user's hands.
Korkusuz et al. (2023)	Microsoft Xbox 360 Kinect™ sensor positioned at 250 cm coupled to a TV screen. Games: 'Skicross' and 'Bounty Super G'.	Visual	Skiing and reward-collecting tasks involving knee flexion/extension and weight-bearing.	Initial familiarization/calibration phase and encouragement through an accumulated reward system.
Chong Li et al. (2023)	CARS System developed in ARKit (iPhone XR). Device attached to the hand via a fingerless glove-style sleeve.	Visual	Interaction with 3D virtual targets generated by augmented reality directly on the mobile screen.	Ergonomic, adaptive glove-sleeve to secure the phone and compensate for weak grip.
Anwar N et al. (2020)	Nintendo Wii console with wireless controllers (Nunchucks) and Wii Balance Board platform. Games: Wii Sports and Cooking Mama.	Visual & Auditory	Sports and cooking simulation focused on balance gain and dual-tasking with level progression.	Native gameplay adjustments according to patient's physical tolerance.
Bouatrous A et al. (2022)	Leap Motion Controller connected via USB to a computer for stereo 3D tracking of phalanges and finger bones.	Visual	Serious games for fine finger rehabilitation with three-dimensional geometric tracking.	Sensor positioned upside down over the arm, allowing desktop support against gravity.
Eichinger F et al. (2021)	Load cells acting as dynamometers integrated into a leg extension machine ('mim Pong SG' game) and an upper limb device.	Visual	Interactive exergame controlled by the patient's maximum voluntary isometric force.	Initial baseline force calibration and adjustable hip (110°) and knee (60° or 90°) angles.
Anwar et al. (2021)	Nintendo Wii home console integrated with the Nintendo Wii Fit biofeedback platform.	Visual	Weight-shifting and postural control exercises guided by native serious games.	Difficulty progression adjusted based on stability scores.
Marda M, Thorat K (2023)	Virtual Reality headset displaying an ecosystem of ten immersive serious games.	Visual	Gamified functional tasks of high immersion based on progressive levels.	Custom selection of the minigame catalog according to the therapeutic goal of the day.
Térémetz M et al. (2019)	Polhemus Fastrak electromagnetic system (4 sensors, 33 Hz) for kinematics + wireless accelerometers (hand dorsum and sternum) + Nintendo Wii.	Visual	Functional training monitored by comprehensive 3D biomechanical analysis of the trunk and arm.	Mandatory prior familiarization with the games and magnetic placement under the table to compute compensations.
Huber et al. (2021)	Dividat SENSO device consisting of a platform with 20 sensors under 5 sub-platforms for CoP analysis.	Visual, Auditory & Tactile	Cognitive-motor exergames focused on reaction time and mobility with safety handrails.	Automatic variation algorithm for speed and complexity of cognitive stimuli.

To complement the study-design analysis, the 17 included studies were also grouped by the main technological modality employed (Table I), allowing direct comparison of studies that used similar systems.

Optical hand and finger tracking (Leap Motion). Three studies used Leap Motion optical sensors for contactless hand and finger tracking: Kostekçi et al. (2024) (CatchaPET and LeapBall software), Aguilera-Rubio et al. (2020) (games developed in Unity3D), and Bouatrous et al. (2023) (adaptive exergame for hand rehabilitation). All three studies focused on distal upper-limb (hand/finger) rehabilitation. Kostekçi et al. and Aguilera-Rubio et al., both randomized trials with a comparator group, reported partially divergent results regarding between-group superiority: the former found an advantage for the video game group in FMA-UE, WMFT, and the kitchen subitem of the DHI, while the latter found no between-group difference in most outcomes, only a long-term within-group improvement. Bouatrous et al., in turn, correspond to a single-arm study with no comparator group, in which only usability (SUS: 71.81) and motivation (IMI: 6.09–6.58) indices were reported (see Section III-D).

Immersive virtual reality with headset (Oculus Quest 2). Three studies used Oculus Quest 2 immersive virtual reality headsets: Peláez-Vélez et al. (2021) ('Blexer' software, 'Phiby's Adventure' game, with Kinect sensor), Amin et al. (2021) (native hand tracking, games in Unity3D/APK), and Marda & Thorat (2023) (ecosystem of ten immersive serious games). All three studies reported significant improvement in favor of the experimental group: Peláez-Vélez et al. in MI, TCT, Tinetti, BBS, and functional gait; Amin et al. in all outcomes assessed (FMA-UE, ARAT, BBT, MBI, SAQOL); Marda & Thorat in mobility (POMA), balance (BBS), and independence (BI), suggesting that the immersion provided by VR headsets may favor consistent gains in motor function and balance within this sample of studies. The JBI critical appraisal confirmed that both Amin et al. and Marda & Thorat correspond to randomized controlled trials (see Supplementary Table 3), allowing them to be compared directly with Peláez-Vélez et al., conducted as a randomized controlled pilot trial; none of the three reports, however, describes in detail the method used to generate the randomization sequence or allocation concealment, an aspect considered in the risk-of-bias assessment of these studies.

Commercial video game consoles (Nintendo Wii). Three studies used the Nintendo Wii console: Anwar N et al. (2020) (Wii Sports and Cooking Mama, with wireless controllers), Anwar et al. (2021) (Wii Fit biofeedback platform), and Térémetz et al. (2019) (Wii combined with the Polhemus Fastrak electromagnetic system for kinematic analysis). All three randomized controlled trials reported significant improvements in balance (BBS) and/or motor function in groups that used the Wii, and Térémetz et al. also reported no increase in compensatory trunk movements during training, a relevant safety finding for using commercial video games in post-stroke rehabilitation.

Motion capture systems with depth camera (Kinect). The Microsoft Kinect sensor was used as the primary technology in Korkusuz et al. (2023) (in the games 'Skicross' and 'Bounty Super G', with a focus on dynamic balance and knee hyperextension) and, in combination with a VR headset, in Peláez-Vélez et al. (2021, already described in the immersive VR group). Korkusuz et al. reported superiority of the experimental group in dynamic balance, knee control, and activities of daily living, in a single-blind randomized controlled trial.

Inertial sensor (IMU) platforms for trunk and balance. Two studies used inertial sensor (IMU) systems positioned on the trunk: Alhwoaimel et al. (2024) (Valedo system, Hocoma AG) and Martino Cinnera et al. (2021) (Riablo system, CoRehab). Both aimed at trunk and balance training; the two studies reported positive results but with different designs: Cinnera et al., a randomized controlled pilot trial, reported clinical improvement on all scales assessed, with a postural difference (CoP) favoring the experimental group; Alhwoaimel et al., a feasibility study without a control group, reported positive psychosocial impacts, but without the possibility of comparison with a control group.

Force platforms and postural biofeedback. The Dividat Senso/SENSO system, a pressure-sensitive force platform with safety handrails, was used in two studies from the same research group: Huber et al. (2022, randomized controlled trial, PEMOCS protocol) and Huber et al. (2021, feasibility study). The feasibility study (2021) preceded the controlled trial (2022) and reported significant improvement only in the TUG mobility outcome, with moderate effect sizes for cognitive functions; the subsequent controlled trial, in turn, found no between-group difference in the primary outcome (MoCA) but identified superiority of the experimental group in complex outdoor gait parameters, illustrating how a feasibility study can inform the design of a subsequent controlled trial on the same platform.

Augmented reality on a mobile device. Chong Li et al. (2023) was the only included study to use smartphone-based augmented reality (CARS system, developed with ARKit for

iPhone XR, with a fingerless glove-sleeve). In a randomized controlled trial comparing the CARS system with conventional occupational therapy, the CARS group outperformed conventional therapy specifically in FMA-UE and ARAT, with no between-group difference in the other outcomes assessed (balance, mental status, and functional independence).

Dedicated mechatronic and haptic devices. Two studies used dedicated mechatronic devices, without relying on commercial optical or inertial sensors: Ali et al. (2018) (ArmAble device, with visual, acoustic, and haptic feedback) and Eichinger et al. (2021) (load cells integrated into the 'Mim-Pong' game). Ali et al., in a multicenter randomized controlled trial, reported significant improvement in the experimental group in FM-UE and ARAT; Eichinger et al., in a non-randomized study without a control group, reported significant pre-post improvement and a correlation between game performance and the patient's pinch/grip strength.

D. Results and Outcomes Found

The 17 included studies presented distinct methodological designs, which directly influence the strength of the evidence for each finding. In this section, results are presented grouped by study design: randomized controlled trials, randomized controlled pilot studies, mixed-methods/feasibility studies, and the study with a non-randomized design, followed by a comparison of how intervention time, sample size, and the presence (or absence) of a comparator group influenced the reported findings. No meta-analysis was performed; the effects described here therefore reflect a narrative synthesis of individual study results rather than a pooled effect estimate.

Randomized controlled trials. Eleven of the 17 studies corresponded to larger-scale randomized controlled trials, comparing the experimental group with a control group or a conventional intervention: Kostekçi et al. (2024), Huber et al. (2022), Aguilera-Rubio et al. (2020), Ali et al. (2018, multicenter), Amin et al. (2021), Korkusuz et al. (2023), Chong Li et al. (2023), Anwar N et al. (2020), Anwar et al. (2021), Marda & Thorat (2023), and Térémetz et al. (2019). Results within this group were heterogeneous regarding between-group statistical significance: some studies reported significant differences favoring the experimental group in motor and functional outcomes (Ali et al., FM-UE and ARAT; Korkusuz et al., dynamic balance and knee control; Chong Li et al., FMA-UE and ARAT; Anwar N et al. and Anwar et al., BBS and FMA; Amin et al., significant within- and between-group improvement in all outcomes assessed—FMA-UE, ARAT, BBT, MBI, SAQOL; Marda & Thorat, superiority of the experimental group in mobility (POMA), balance (BBS), and independence (BI)), while others found no between-group difference in the primary outcome, with improvement occurring only within each group over time (Huber et al., 2022, in MoCA; Aguilera-Rubio et al., in all outcomes except ARAT subitem D). Térémetz et al. reported no unwanted compensatory movements and improvement over time, with no between-group difference observed in Table II.

Randomized controlled pilot studies. Two studies were randomized controlled pilot trials: Martino Cinnera et al. (2021) and Peláez-Vélez et al. (2021), characterized by smaller samples and a primary goal of testing feasibility and estimating parameters for a larger-scale trial. Both reported significant improvement favoring the experimental group (Cinnera et al.: clinical improvement on all scales assessed, with a favorable post-training postural difference (CoP) for the experimental group; Peláez-Vélez et al.: superiority of the experimental group in MI, TCT, Tinetti, BBS, and functional gait classification). Although the results are positive, the pilot nature of these studies, with typically small samples, limits generalization until larger, confirmatory trials confirm the effects.

Mixed-methods and feasibility studies. Two studies adopted feasibility designs without a control group: Alhwoaimel et al. (2024), explicitly a mixed-methods feasibility study, and Huber et al. (2021), a feasibility study of cognitive-motor exergames. Because they did not include a comparator group, these studies do not allow inferences about the relative efficacy of an alternative intervention; their results mainly reflect indicators of recruitment, retention, adherence, and within-group pre-post change. Although Aimel et al. reported positive psychosocial impacts (PIADS) and three participants with mild frustration, Huber et al. (2021) reported significant improvement only in TUG, with moderate effect sizes for visual attention and mental rotation, and a dropout rate above the planned threshold.

Study with a non-randomized design. Only Eichinger et al. (2021) remains in this category: the authors themselves explicitly described the study as a non-randomized controlled trial with an experimental group and a control group, but without random allocation. The other three studies initially grouped here because their design was not specified in the title were reclassified after the JBI critical appraisal: Amin et al. (2021) and Marda & Thorat (2023) were confirmed as randomized controlled trials (see previous paragraph), and Bouatrous et al. (2023) actually corresponds to a single-arm study (subjective assessment of usability and motivation, with no control group and no repeated pre/post measures), and should therefore not be interpreted as a comparative efficacy study (see Supplementary Table 4).

Comparison across designs: impact of time, sample size, and comparator group. Overall, the studies with an active comparator group (conventional physical therapy, occupational therapy, or another control intervention), present in most of the randomized controlled trials and in both pilot studies, were the only ones able to report between-group statistical differences, making it possible to distinguish the specific effect of the digital technology from a possible spontaneous improvement or one associated with the therapeutic attention received by both groups. The feasibility studies, by contrast, did not include a control group and therefore reported only within-group (pre-post) changes, which precludes attributing the observed gains specifically to the technology tested. Regarding sample size, a small sample was reported as an explicit limitation in at least six of the 17 studies (Aguilera-Rubio et al.; Alhwoaimel et al.; Martino Cinnera et al.; Peláez-Vélez et al.; Anwar et al., 2021; Marda & Thorat), concentrated mainly in the pilot and feasibility studies which is expected given the very purpose of these designs, but also reinforces that their findings should be interpreted as exploratory. Regarding intervention time, the data show, in aggregate, a range of 3 to 30 sessions and 30 to

80 minutes per session across the 17 studies; Supplementary Table 1 details the specific sample and comparator-group values for each study.

Table V Comparative summary of interventions, study designs, and outcomes by included study

Study Information			**Intervention and Outcomes**		
Study (Author, Year)	***Technology / Intervention***	***Study design***	***Sample (N)***	***Comparator group***	***Direction of observed effect (summary)***
Kostekçi et al. (2024)	Leap Motion Controller (CatchaPET/LeapBall) — optical hand/finger capture	Randomized controlled trial (RCT)	30	Manual rehabilitation group and use of a triple-coordination device	Within-group improvement in both groups; between-group analysis favored the video game group in FMA-UE, WMFT, and the DHI kitchen subscale
Huber et al. (2022)	Dividat Senso — pressure-sensitive force platform	Randomized controlled trial (RCT)	46	Conventional physical therapy and weekly telephone calls to maintain contact with the research team	No between-group difference in the primary outcome (MoCA); experimental group superior in complex outdoor gait parameters
Aguilera-Rubio et al. (2020)	Unity3D + Leap Motion optical sensor	Randomized controlled trial (RCT)	36	Functional tasks	No between-group difference (except ARAT subitem D); long-term within-group improvement in the experimental group
Alhwoaimel et al. (2024)	Valedo System (Hocoma AG) — inertial measurement units (IMUs) on the trunk	Feasibility study (mixed methods), no control group	10	No control group	Positive psychosocial impacts (PIADS: competence 1.7; adaptability 1.6; self-esteem 1.8)
Martino Cinnera et al. (2021)	Riablo (CoRehab) — inertial sensors + wireless platform	Randomized controlled pilot trial (RCT)	21	Conventional balance training	Clinical improvement on all scales; post-training difference favoring the SB group in postural assessment (CoP)
Ali et al. (2018)	ArmAble — mechanized and gamified device for the upper limbs	Multicenter randomized controlled trial (RCT)	120	Training in functional daily tasks with the upper limb	Significant improvement in the experimental group in FM-UE (2 and 6 weeks) and ARAT (6 weeks)
Peláez-Vélez et al. (2021)	Oculus Quest 2 + "Blexer" software ("Phiby's Adventure" game)	Randomized controlled pilot trial (RCT)	24	Conventional physical therapy	Experimental group significantly superior to control in MI, TCT, Tinetti, BBS, and functional gait classification
Amin et al. (2021)	Oculus Quest 2 with native hand tracking (games in APK/Unity3D)	Randomized controlled trial (RCT)	52	Conventional physical therapy	Significant within- and between-group improvement, with greater gains in the experimental group across all outcomes
Korkusuz et al. (2023)	Kinect Xbox 360 ("Skicross" and "Bounty Super G" games)	Single-blind randomized controlled trial (RCT)	25	Conventional physical therapy	Experimental group superior in dynamic balance, knee control, and ADLs; control group improved in only some outcomes
Chong Li et al. (2023)	CARS system developed in ARKit (iPhone XR)	Randomized controlled trial (RCT)	30	Conventional occupational therapy	CARS group outperformed conventional therapy in FMA-UE and ARAT; no between-group difference in BS, MMT, or BI
Anwar N et al. (2020)	Nintendo Wii (Wii Sports and Cooking Mama) with wireless controllers	Randomized controlled trial (RCT)	68	Conventional physical therapy	Significant improvement favoring the VR group in BBS and motor FMA; no differences in sensitivity
Bouatrous et al. (2023)	Leap Motion (USB) — 3D stereo tracking of hand phalanges/bones	Single-arm study (usability/feasibility assessment), no control group	11	No control group	High usability and motivation (SUS 71.81; IMI 6.09–6.58); single-arm study, with no comparison to a control group
Eichinger et al. (2021)	Load cells (dynamometer) integrated into the "Mim-Pong" game	Non-randomized study (as reported by the authors), no control group	24	No control group	Significant pre-post improvement in the experimental group; game performance correlated with pinch/grip strength
Anwar et al. (2021)	Nintendo Wii Fit — biofeedback platform	Randomized controlled trial (RCT)	68	Conventional physical therapy	Significant improvement in balance (BBS) and the motor subscales of the FMA-LE in the VR group
Marda & Thorat (2023)	VR headset with an ecosystem of 10 immersive serious games	Randomized controlled trial (RCT)	32	Task-oriented training and conventional exercises	Experimental group significantly superior to control in mobility (POMA), balance (BBS), and independence (BI)

Study Information			Intervention and Outcomes		
Study (Author, Year)	*Technology / Intervention*	*Study design*	*Sample (N)*	*Comparator group*	*Direction of observed effect (summary)*
Térémetz et al. (2019)	Polhemus Fastrak electromagnetic system + Nintendo Wii	Blinded randomized controlled trial (RCT)	40	Conventional physical therapy	No increase in compensatory trunk movements after Wii training; paretic arm acceleration increased significantly over time
Huber et al. (2021)	Dividat SENSO — cognitive-motor exergames	Feasibility study, no control group	10	No control group	Significant improvement only in TUG; moderate effect sizes for visual attention and mental rotation

TABLE VI. STUDY CHARACTERISTICS, MAIN OUTCOMES, AND REPORTED LIMITATIONS

Author (Year)	Assessed Outcomes	Main Results (p < 0.05)	Experience / Usability / Adherence	Reported Limitations
Kostekçi et al. (2024)	FMA-UE, WMFT, MMDT, DHI.	Significant intragroup improvements in both groups. Intergroup analysis favored the VG (Video Game) group for FMA-UE, WMFT, and the kitchen subscore of the DHI.	Excellent overall acceptability; no limiting adverse events reported.	Wide age range of participants, introducing high baseline functional and recovery variability.
Huber et al. (2022)	SIS 3.0, TMT, Stroop, NBT, MRT, TUG, 10MWT, OWA, MoCA.	No significant intergroup differences in the primary outcome (MoCA). The experimental group was superior in complex spatiotemporal outdoor gait parameters.	Adherence rate of 89.8% and compliance of 89.1%. High motivation and satisfaction reported.	Limited variety of games; complex motor instructions provided verbally without in-game feedback; female underrepresentation.
Aguilera-Rubio et al. (2020)	Grip Strength, BBT, ARAT, DASH.	No significant intergroup differences, except for ARAT sub-section D. Intragroup analysis showed significant long-term improvements in the experimental group.	Adherence and satisfaction were very high. The controlled clinical environment provided a strong sense of safety.	Lack of stratification by baseline motor impairment level; small sample size limits generalizability.
Alhwoaimel et al. (2024)	PIADS, recruitment, retention, adherence, safety, qualitative interviews.	Positive psychosocial impacts on the PIADS (competence: 1.7; adaptability: 1.6; self-esteem: 1.8).	High motivation driven by the playful nature of the system. Three participants reported minor frustration.	Small and heterogeneous sample size; lack of a control group; inability to customize game settings.
Cinnera et al. (2021)	BBS, CNS, BI, RMI, NIHSS, stabilometry (CoP).	Clinical improvement across all scales. Significant post-training difference in postural assessment: reduced CoP path length (eyes open) in the SB vs. UC group.	Significantly higher participation (PRPS) in the SB group (5.82) vs. UC group (4.92).	Non-parametric statistical approach due to small sample size; baseline differences in stroke side and onset.
Ali et al. (2018)	FM-UE, ARAT.	Significant improvements in the EG for FM-UE at 2 and 6 weeks, and for ARAT at 6 weeks.	High retention and adherence rates to the proposed home-based training protocol.	Short intervention period; lack of comprehensive sensory assessment; daily real-world hand use not monitored.
Vélez-Guerrero et al.	Daniels, MAS, MI, TCT.	EG was significantly superior to CG in MI, TCT, Tinetti, BBS, and functional gait	Home-based training adherence	Small sample size; prior clinical treatments and rehabilitation history
(2021)	Tinetti, BBS, Sagunto Scale.	classification.	of 89.1% (EG) vs. 85.8% (CG).	were not evaluated.
Amin et al. (2021)	FMA-UE, ARAT, BBT, MBI, SAQOL.	Significant improvements within and between groups. The EG demonstrated significantly greater improvements in all outcomes.	Excellent tolerability and no reports of physical discomfort during training sessions.	Single-center study; lack of a mid-to-long-term follow-up assessment.
Korkusuz et al. (2023)	FRT, TUG, computerized gait analysis, BI.	EG showed superior improvements in dynamic balance, knee control, and ADLs. CG improved in dynamic balance and knee control but not in ADLs.	High levels of overall satisfaction and system usability reported by patients.	Absence of a follow-up assessment to monitor the retention of therapeutic gains.
Chong Li et al. (2023)	FMA-UE, ARAT, MMT, Brunnstrom, MMSE, Cognitive Games, BI.	The CARS group significantly outperformed conventional OT in FMA-UE and ARAT. No intergroup differences in BS, MMT, or BI.	High general satisfaction (4.27) and system control usability (4.67). Highly recommended for homecare.	Unmonitored compensatory movements; non-blinded therapists; uniform-sized training gloves.
Anwar N et al. (2020)	BBS, FMA-UE.	Significant post-treatment differences in BBS and motor FMA in favor of the VR group. No differences in sensation.	Interactive games were reported as engaging and fun, whereas conventional therapy was described as tedious.	Slow technological familiarization; limited clinical outcome scales; lack of long-term follow-up.
Boustrous A et al. (2022)	Finger ROM, SUS, IMI.	SUS score: 71.81 (acceptable). IMI subscale scores ranged from 6.09 to 6.58 (out of 7).	High perceived interest, enjoyment, and choice. Good scores for effort and utility.	Difficulty in 3D motion perception within the virtual space, especially for patients with cognitive deficits.
Eichinger F et al. (2021)	FMAS, MMSE, Mim-Pong performance, MVIC.	Game performance score correlated strongly with pinch and grasp strength. Significant pre-to-post improvements in the EG.	High patient engagement recorded natively via in-game telemetry.	Non-randomized study design; small sample size; lack of standardized fatigue or motivation scales.
Anwar et al. (2021)	BBS, FMA-LE.	Significant improvements in balance (BBS) and motor subscales of the FMA-LE in the VR group.	High adherence rate with good compliance to the training protocol.	Single-center study design, limiting the generalizability of the therapeutic effects.
Marda M, Thorat K (2023)	POMA, BBS, BI.	EG showed significantly superior improvements in mobility (POMA), balance (BBS), and independence (BI) compared to CG.	VR-based training was perceived as more dynamic and engaging than conventional task-oriented training.	Restricted sample size; lack of rigorous standardization of the home/extraclinical environment.
Térémetz M et al. (2019)	BI, AST, Bell's Neglect, FMA, ARAT, BBT, trunk kinematics.	No increase in compensatory trunk movements after Wii training. Paretic arm acceleration increased significantly over time.	Excellent physical tolerability; no muscle soreness or joint discomfort reported.	Lack of clinical assessments of real-world daily activities; absence of a dose-matched control group.
Huber et al. (2021)	FBW, VAS, Gait parameters, TUG, TMT A/B, BREQ-3.	Significant improvement only in TUG. Moderate effect sizes observed for visual attention and mental rotation.	Feasibility parameters met, though the dropout rate was higher than the planned threshold.	No control group; sample biased toward high-functioning and highly educated individuals; missing data.

Abbreviations & Outcomes Legend:
• FMA-UE: Fugl-Meyer Assessment - Upper Extremity • WMFT: Wolf Motor Function Test • MMDT: Minnesota Manual Dexterity Test • DHI: Duruöz Hand Index • SIS 3.0: Stroke Impact Scale • TMT: Trail Making Test (A/B) • NBT: N-Back Test • MRT: Mental Rotation Test • TUG / TUG-Cogn: Timed Up and Go (Simple / Cognitive) • 10MWT: 10-Meter Walk Test • OWA: Outdoor Walking Assessment • MoCA: Montreal Cognitive Assessment • BBT: Box and Block Test • ARAT: Action Research Arm Test • DASH: Disabilities of the Arm, Shoulder and Hand • PIADS: Psychosocial Impact of Assistive Devices Scale • BBS: Berg Balance Scale • CNS: Canadian Neurological Scale • BI / MBI: Barthel Index / Modified Barthel Index • RMI: Rivermead Mobility Index • NIHSS: National Institutes of Health Stroke Scale • CoP: Center of Pressure • MAS: Modified Ashworth Scale • Mt: Matricity Index • TCT: Trunk Control Test • FRT: Functional Reach Test • MMT: Manual Muscle Testing • MMSE: Mini-Mental State Examination • SUS: System Usability Scale • IMI: Intrinsic Motivation Inventory • MVIC: Maximum Voluntary Isometric Contraction • POMA: Performance-Oriented Mobility Assessment • AST: Apraxia Screen of TULIA • EG / CG: Experimental Group / Control Group • VG / VR: Video Game / Virtual Reality

E. Secondary Outcomes: Usability, Motivation, and Adherence

To assess clinical outcomes, studies used different scales; Table II shows the scales in more detail. User experience assessments generally indicated good acceptance of the systems across the 17 studies; some used validated instruments to quantify usability and motivation, such as the System Usability Scale (SUS), the Intrinsic Motivation Inventory (IMI), and the Psychosocial Impact of Assistive Devices Scale (PIADS) (Table II), while most reported only session adherence rates (percentage of protocol completion) or non-standardized satisfaction scales. Because completing scheduled sessions does not necessarily reflect greater therapeutic engagement, interpret these findings on usability, motivation, and adherence separately from the clinical outcomes discussed in Section III-D, and with due caution. The most frequently reported limitations included small samples, clinical heterogeneity, lack of long-term follow-up, and limited intervention personalization (Table II).

IV. DISCUSSION

The studies included in this review showed that digital technologies, such as virtual reality, serious games, and exergames, can improve different aspects of functional recovery after stroke. Despite heterogeneity among the platforms used, therapeutic protocols, and assessment instruments, most studies reported improvements in outcomes related to motor function, balance, functional mobility, performance in day-to-day activities, and quality of life.

This trend should, however, be interpreted with caution: because no meta-analyses were performed, these findings do not provide a robust, generalizable estimate of the effect of digital technologies, but rather reflect a pattern observed consistently across heterogeneous individual studies, many with small samples. In addition, the methodological assessment using the JBI tool identified potential risks of bias, related mainly to the lack of blinding of participants and therapists, the lack of description of allocation concealment, and, in some studies, loss to follow-up.

Even so, these findings align with recent evidence in the literature, which indicates that virtual reality is a promising complementary strategy to conventional physical therapy, especially when used to increase the intensity and duration of practice during rehabilitation [13,15].

The observed benefits can be understood in light of the principles of experience-dependent neuroplasticity. According to Kleim and Jones [5], neural adaptations are favored by interventions that promote intensive, repetitive, specific, and functionally relevant practice—principles that, at least in part, were present in the design of the technological interventions included in this review (Table I). However, the included studies assessed clinical and functional outcomes (e.g., strength, balance, mobility) and did not directly measure indicators of neuroplasticity or cortical reorganization (e.g., functional neuroimaging or transcranial magnetic stimulation). Thus, the relationship between the effects observed in this review and the motor learning and cortical reorganization mechanisms proposed by Kleim and Jones [5] constitutes a plausible theoretical inference rather than a direct demonstration made by the analyzed studies.



Fig. 2. Proposed conceptual model explaining the contribution of digital technologies to functional recovery after stroke.

Fig. 2 summarizes the conceptual model proposed based on this review's findings, showing how the characteristics of digital technologies can favor motor learning and neuroplasticity processes, culminating in functional recovery after stroke.

Another important aspect identified in this review was the predominance of systems based on markerless motion capture and real-time visual feedback. These features enable greater interaction between the user and the virtual environment, supporting immediate movement correction and task progression based on patient performance. Some studies also incorporated auditory feedback, multisensory stimuli, and adaptive mechanisms for automatic difficulty adjustmentfeatures that can increase motivation and personalize training. From a Biomedical Engineering perspective, these results show a growing trend toward intelligent systems that integrate movement monitoring, real-time processing, and individualized therapeutic adaptation. By definition of this review's eligibility criteria, all included virtual reality interventions featured a game-based component (serious games or exergames), since virtual reality studies without this component were excluded (exclusion criterion 3, Section II-B). Thus, the findings discussed here apply specifically to gamified interventions and should not be generalized to non-gamified virtual reality approaches, which this review did not address.

Another frequent finding among the included studies was participants' high acceptance of the technologies (Table II). Several studies reported good usability, high satisfaction, and greater motivation during the interventions, although they measured these differently using validated usability and motivation instruments in some cases and adherence rates or qualitative reports in others, which limits direct comparisons (Table II). Although often considered secondary outcomes, these factors can influence treatment adherence and the intensity of therapeutic practice. Strategies such as gamification, goal setting, rewards, and immediate feedback make training more engaging and favor continuity of rehabilitation, especially in individuals with chronic sequelae, for whom maintaining motivation represents one of the main clinical challenges [8,10].

Despite the promising results, this review has important limitations in both the available literature and the review process itself. Notable limitations in the literature include small sample sizes, heterogeneity of protocols and technological platforms, diverse study designs (including controlled trials and a feasibility study without a control group), and a lack of standardization of assessment instruments. It is particularly worth noting that Bouatrous et al. (2023) corresponds to a single-arm study focused on assessing usability and motivation, with no control group and no repeated pre/post measures, and should be interpreted as a usability/feasibility study rather than as comparative evidence of clinical efficacy, contrary to what the title of the original publication might suggest. In addition, few studies conducted long-term follow-up, which limits understanding of whether functional gains are maintained after the interventions end.

Regarding the review process, the authors identified limitations in assessing methodological quality. Although most studies are experimental and have results favorable to interventions with serious games, exergames, and virtual reality, the critical appraisal using the JBI tool (Supplementary Tables 2–4) found limitations related to the lack of blinding of participants and therapists, the lack of description of allocation concealment, and, in some studies, the absence of a control group and the conduct of multiple pre- and post-intervention assessments.

These limitations make comparisons between studies difficult and reinforce the need for multicenter clinical trials with standardized protocols regarding study design, individualized reporting of sample size and duration, and systematic assessment of risk of bias.

Overall, the evidence analyzed suggests that virtual reality, serious games, and other digital technologies should not be understood merely as technological resources, but also as tools capable of implementing fundamental principles of neurorehabilitation. By promoting task-oriented training, real-time feedback, greater practice intensity, and high engagement, these technologies favor motor learning and neuroplasticity mechanisms that can contribute to more comprehensive functional recovery after stroke.

V. Conclusion

This systematic review found evidence that digital technologies, including virtual reality, serious games, and exergames, are promising strategies for rehabilitating individuals after stroke. Overall, most interventions showed gains across functional domains, including motor function, balance, mobility, performance in day-to-day activities, and quality of life, and indicated good acceptance, motivation, and treatment adherence among participants. These findings suggest that these technologies have therapeutic potential, not only as technological innovations but also as ways to operationalize motor learning principles such as intensive training, task-oriented practice, and real-time feedback, although, as discussed, the included studies did not directly measure the underlying neuroplastic mechanisms. However, heterogeneity among studies, predominantly small sample sizes, diverse methodological designs, and methodological quality limitations identified by the JBI critical appraisal still

limit the ability to draw robust, generalizable conclusions about the clinical effectiveness of these technologies.

The results indicate that digital technologies can contribute to more comprehensive functional recovery, although the literature remains predominantly focused on motor and functional outcomes, with less attention to measures of social participation and quality of life. In this context, it is recommended that future studies adopt more standardized protocols regarding study design, individually report sample size, intervention duration, and the comparator group used, incorporate systematic assessment of methodological quality and risk of bias from the study-design stage onward, and, whenever possible, conduct meta-analyses stratified by study type and technological modality. Furthermore, developing intelligent, adaptive, user-centered systems offers an opportunity to improve clinical efficacy and strengthen the application of digital technologies in post-stroke rehabilitation.

ACKNOWLEDGMENT

The authors thank the Federal Institute of Education, Science and Technology of Southeast Minas Gerais – Juiz de Fora Campus for its contribution to their academic and professional development, as well as for the educational support and institutional resources provided throughout this work.

REFERENCES

[1] G. V. Feigin, B. A. Stark, C. O. Johnson, et al., "Global, regional, and national burden of stroke and its risk factors," Lancet Neurology. (Update to the most recent available edition corresponding to the GBD).

[2] GBD Stroke Collaborators, "Global burden of stroke and risk factors," Lancet Neurology. (Update to the edition corresponding to the most recent GBD).

[3] C. J. Winstein, J. Stein, R. Arena, et al., "Guidelines for Adult Stroke Rehabilitation and Recovery," Stroke, vol. 47, pp. e98–e169, 2016.

[4] P. Langhorne, J. Bernhardt, and G. Kwakkel, "Stroke rehabilitation," Lancet, vol. 377, no. 9778, pp. 1693–1702, 2011.

[5] J. A. Kleim and T. A. Jones, "Principles of experience-dependent neural plasticity: Implications for rehabilitation after brain damage," Journal of Speech, Language, and Hearing Research, vol. 51, pp. S225–S239, 2008.

[6] J. W. Krakauer, S. T. Carmichael, D. Corbett, and G. F. Wittenberg, "Getting neurorehabilitation right: What can be learned from animal models?," Nature Reviews Neurology, vol. 8, no. 3, pp. 155–165, 2012.

[7] R. J. Nudo, "Recovery after brain injury: Mechanisms and principles," Frontiers in Human Neuroscience, vol. 7, p. 887, 2013.

[8] R. A. Schmidt and T. D. Lee, Motor Control and Learning: A Behavioral Emphasis, 6th ed. Champaign, IL: Human Kinetics, 2019.

[9] D. Michael and S. Chen, Serious Games: Games That Educate, Train, and Inform. Boston: Thomson Course Technology, 2006.

[10] T. Susi, M. Johannesson, and P. Backlund, Serious Games: An Overview, Technical Report HS-IKI-TR-07-001, University of Skövde, 2007.

[11] Y. Oh and S. Yang, "Defining exergames and exergaming," in Proceedings of Meaningful Play, 2010.

[12] M. F. Levin, P. L. Weiss, and E. A. Keshner, "Emergence of virtual reality as a tool for upper limb rehabilitation: Incorporation of motor control and motor learning principles," Journal of NeuroEngineering and Rehabilitation, vol. 12, pp. 1–15, 2015.

[13] K. E. Laver, S. George, S. Thomas, J. E. Deutsch, and M. Crotty, Virtual reality for stroke rehabilitation, Cochrane Database of Systematic Reviews. (Use the most recent version available at the time of submission).

[14] M. S. Cameirão, S. Bermúdez i Badia, E. Duarte, and P. F. M. J. Verschure, "The Rehabilitation Gaming System: A review," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 19, no. 3, pp. 287–299, 2011.

[15] D. Corbetta, F. Imeri, and R. Gatti, "Rehabilitation using virtual reality after stroke: A systematic review and meta-analysis," Journal of NeuroEngineering and Rehabilitation, vol. 12, pp. 1–14, 2015.

[16] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021.

[17] J. P. T. Higgins, J. Thomas, J. Chandler, et al., Eds., Cochrane Handbook for Systematic Reviews of Interventions, Version 6.4. Cochrane, 2023.

[18] H. Kostekçi, D. Tarakci, H. Bostanci, A. Emir, and E. Tarakci, "Effects of video game-based hand therapy on motor functions in stroke rehabilitation: A randomized trial," Physiother Res Int., vol. 31, no. 1, p. e70144, 2025, doi:10.1002/pri.70144.

[19] S. K. Huber, R. H. Knols, J. P. O. Held, et al., "PEMOCS: Effects of a concept-guided, cognitive-motor exercise training on cognitive functions and gait in chronic stroke — a randomized controlled trial," Frontiers in Aging Neuroscience, vol. 17, p. 1514594, 2025, doi:10.3389/fnagi.2025.1514594.

[20] Á. Aguilera-Rubio, I. M. Alguacil-Diego, A. Mallo-López, et al., "Use of low-cost virtual reality in the treatment of the upper extremity in chronic stroke: A randomized controlled trial," J Neuroeng Rehabil., vol. 21, no. 1, p. 12, 2024, doi:10.1186/s12984-024-01303-2.

[21] N. A. Alhwoaimel, A. M. Hughes, M. Warner, et al., "Video game-based trunk exercises for rehabilitation in chronic stroke survivors: A mixed-methods feasibility study," Sensors (Basel), vol. 24, no. 21, p. 6830, 2024, doi:10.3390/s24216830.

[22] A. Martino Cinnera, I. Ciancarelli, S. Marrano, et al., "Sensor-based balance training with exergame feedback in subjects with chronic stroke: A randomized controlled pilot trial," Brain Sciences, vol. 14, no. 9, p. 917, 2024, doi:10.3390/brainsci14090917.

[23] A. S. Ali, D. S. Kumaran, A. Unni, et al., "Efficacy of an intensive, functional, gamified rehabilitation program on upper limb function in people with stroke (EnteRtain): A multicenter randomized controlled trial," Neurorehabilitation and Neural Repair, vol. 38, no. 4, pp. 243–256, 2024, doi:10.1177/15459683231222921.

[24] F. J. Peláez-Vélez, M. Eckert, M. Gacto-Sánchez, and Á. Martínez-Carrasco, "Use of virtual reality and video games in the physical therapy treatment of stroke patients: A randomized controlled pilot trial," Int J Environ Res Public Health, vol. 20, no. 6, p. 4747, 2023, doi:10.3390/ijerph20064747.

[25] F. Amin, A. Waris, S. Syed, et al., "Efficacy of immersive virtual reality-based hand rehabilitation games for improving hand motor functions in subacute stroke patients," IEEE Trans Neural Syst Rehabil Eng., vol. 32, pp. 2060–2069, 2024, doi:10.1109/TNSRE.2024.3405852.

[26] S. Korkusuz, G. Taşkın, B. S. Korkusuz, M. S. Özen, and Z. Ö. Yürük, "Examining the effects of non-immersive virtual reality game-based training on knee hyperextension control and balance in patients with chronic stroke: A single-blind randomized controlled study," Neurol Sci., vol. 46, no. 3, pp. 1267–1275, 2025, doi:10.1007/s10072-024-07830-z.

[27] C. Li, X. Song, S. Chen, et al., "Efficacy and long-term adoption of a mobile augmented reality system in stroke patients: Randomized controlled trial," JMIR Serious Games, vol. 9, no. 4, p. e30184, 2021, doi:10.2196/30184.

[28] N. Anwar, H. Karimi, A. Ahmad, et al., "Virtual reality training using Nintendo Wii games for stroke patients: Randomized controlled trial," JMIR Serious Games, vol. 10, no. 2, p. e29830, 2022, doi:10.2196/29830.

[29] A. Bouatrous, A. Meziane, N. Zenati, and C. Hamitouche, "A novel VR-based adaptive exergame for hand rehabilitation after stroke," Multimedia Systems, vol. 29, pp. 3385–3402, 2023, doi:10.1007/s00530-023-01180-0.

[30] L. F. Eichinger, F. Noveletto, S. C. Domenech, T. Moulin, Y. Sagawa, and A. V. Soares, "Mim-pong: A serious game for the assessment and treatment of the lower limb in patients with hemiplegic stroke," Fisioterapia do Arco, vol. 15, pp. 59–68, 2025.

[31] N. Anwar, H. Karimi, A. Ahmad, N. Mumtaz, G. Saqulain, and S. A. Gilani, "A novel virtual reality training strategy for post-stroke patients: A randomized controlled trial," J Healthc Eng., vol. 2021, p. 6598726, 2021, doi:10.1155/2021/6598726.

[32] M. Marda and K. Thorat, “Effect of task oriented balance training versus virtual reality based balance training in stroke patients — a comparative study,” Pravara Med Rev., vol. 15, no. 4, pp. 58–66, 2023, doi:10.36848/PMR/2023/00000.511228.

[33] M. Térémetz, A. Garcia Alvarez, S. Hanneton, et al., “Improving upper limb and trunk kinematics through interactive games in individuals with chronic stroke: A blinded randomized controlled trial,” Annals of Physical and Rehabilitation Medicine, vol. 65, no. 3, p. 101622, 2022, doi:10.1016/j.rehab.2021.101622.

[34] S. K. Huber, J. P. O. Held, E. D. de Bruin, and R. H. Knols, “Personalized motor-cognitive exercise training in chronic stroke: A feasibility study,” Frontiers in Aging Neuroscience, vol. 13, p. 730801, 2021, doi:10.3389/fnagi.2021.730801.

Carta Resposta aos Revisores

Inspiratory Muscle Training Using a Harmonica in the Postoperative Period of Coronary Artery Bypass Grafting

Programa de Pós-Graduação em Tecnologia em Saúde (PPGTS) – PUCPR

Prezados Revisores,

Agradecemos a leitura e as críticas construtivas. Gostaríamos de esclarecer que o presente estudo não realizou e nem teve como objetivo avaliar a correlação entre o *MIP* e a distância percorrida no TC6M, e que o objetivo deste estudo é analisar a viabilidade da gaita de boca como um dispositivo de uso para o treinamento da musculatura inspiratória para que, no próximo passo, seja realizado um ensaio clínico randomizado comparando um grupo controle com o grupo intervenção. Ajustamos o texto para melhor compreensão do leitor.
Abaixo, apresentamos as respostas aos apontamentos de ambos os revisores. Em itálico mostramos o comentário do revisor e em seguida a resposta.

Revisor 1

Apontamento 1 – Correlação de MIP e TC6M
Comentário do revisor:

> *O artigo reporta correlação entre MIP e distância percorrida no TC6M, mas não ficou claro qual teste de correlação foi usado (Pearson? Spearman?), o coeficiente encontrado e o tamanho de efeito. Dado o tamanho pequeno da amostra, essa correlação seria apenas exploratória/hipótese geradora.*

Resposta dos autores:

Gostaríamos de agradecer o comentário por completo e apenas esclarecer que o presente estudo não realizou e nem teve como objetivo avaliar a correlação entre o *MIP* e a distância percorrida no TC6M. Trata-se de um estudo de viabilidade cujo objetivo foi comparar os momentos pré e pós-intervenção para cada desfecho de forma independente (*MIP* por meio do teste de Wilcoxon e TC6M por meio do teste t pareado). Ajustamos o texto na Discussão e Conclusão para garantir que

nenhuma inferência estatística de correlação bivariada seja transmitida ao leitor. Além disso, adicionamento a magnitude do efeito entre os resultados de *MIP* e TC6M. Para testar a normalidade da distribuição dos dados, foi utilizado o teste de Shapiro–Wilk. Foi confirmada a distribuição normal para o teste de caminhada de 6 minutos antes e após a intervenção. A variável que não mostrou normalidade nos dados foi a pressão inspiratória máxima antes da intervenção, e para essa variável foi utilizado o teste correspondente não-paramétrico. Para testar o efeito pré e pós-intervenção da *MIP*, foi aplicado o teste de Wilcoxon dos Postos Sinalizados. Para avaliar a magnitude do efeito do teste de Wilcoxon, utilizou-se o coeficiente de correlação de Rosenthal. Em todos os casos, adotaram-se os critérios de Cohen (1988) para a classificação da força de associação, classificados em pequeno (próximo a 0,10), moderado (próximo a 0,30) e grande (próximo ou acima de 0,50).

Referência: Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale, New Jersey.

Revisor 2

Apontamento 1 – Tipo de estudo

Comentário do revisor: *A avaliação inclui 13 pacientes e não utiliza um grupo controle submetido à fisioterapia convencional, o que impede determinar se as melhorias observadas decorreram do uso da gaita ou da recuperação esperada após a cirurgia e das demais intervenções realizadas no período. Incluir um grupo controle e realizar um estudo randomizado comparando a gaita com dispositivos convencionais de treinamento muscular inspiratório.*

Resposta dos autores:

Gostaríamos de agradecer os comentários construtivos. Para deixar mais claro, este estudo é de viabilidade da gaita de boca como um dispositivo de uso para o treinamento da musculatura inspiratória para que, no próximo passo, seja realizado

um ensaio clínico randomizado, ou seja, não foi comparado um grupo controle com um grupo intervenção porque o presente estudo é de viabilidade apenas.

Apontamento 2 – Pressão de ar na Gaita de Boca

Todos os participantes utilizaram a mesma abertura da gaita, sem individualização da resistência de acordo com a capacidade inspiratória, e os próprios autores informam que não foi possível determinar a resistência ideal para o treinamento. Adicionar diferentes níveis de carga definidos a partir da MIP e acompanharia também adesão e segurança durante as sessões.

Resposta dos autores:

Concordamos e agradecemos a crítica construtiva. A ideia futura é adaptar a Gaita de boca para saber a pressão inspiratória na abertura correta. A gaita de boca possui nas aberturas iniciais uma resistência menor, notas mais graves com espaços maiores para a passagem do ar, enquanto aberturas finais possuem maior resistência, com um espaço menor para a passagem do ar e notas mais agudas. Dado que as aberturas da gaita irão gerar notas diferentes, quanto mais agudo, mais força é necessária para fazer a nota, ou seja, maior a resistência (Bahnson, 1998).

A ideia inicial era definir o nível de resistência pela Pimáx que o paciente apresentasse, medida pelo manovacuômetro. Uma Pimáx de 20 a 30 cmH_2O iniciaria na abertura de número 1 (grau mais baixo, mais fácil), de 30 a 40 cmH_2O na abertura de número 4 (grau baixo, nível médio), e >40 cmH_2O na abertura de número 6 (grau médio, nível difícil). Entretanto, os participantes conseguiram atingir na avaliação uma Pimáx >40 cmH_2O, sendo portanto fixado na abertura de número 6.

Este protocolo foi elaborado em colaboração de uma musicoterapeuta e a partir de estudos envolvendo o uso da gaita na DPOC (Alexander & Wagner, 2011, Hart *et al*., 2020; Lewis *et al.,* 2022), com as devidas adaptações para a população e contexto deste estudo

Como este estudo foi somente de viabilidade, com um tempo mais breve, não foi possível estabelecer o critério de quanto de pressão de ar entra ou sai

exatamente da gaita de Boca, mas futuramente pode ser realizado para dar maior consistência a um futuro ensaio clínico randomizado.

Referências:

ALEXANDER, J. L. et al. Is Harmonica Playing an Effective Adjunct Therapy to Pulmonary Rehabilitation? Revista Rehabilitation Nursing. V 37, N.14, P: 207-212, 2012. DOI: https://doi.org/10.1002/rnj.33

BAHNSON, H. T.; ANTAKI, J. F.; BEERY, Q. C. Acoustical and physical dynamics of the diatonic harmonica. *The Journal of the Acoustical Society of America*, v. 103, n. 4, p. 2134-2144, 1998

HART, M. K., et al. Usefulness of harmonica playing to improve outcomes in patients with chronic obstructive pulmonary disease. DOI: https://doi.org/10.1080/08998280.2019.1704135

LEWIS A., et al. Playing the harmonica with chronic obstructive pulmonary disease. A qualitative study. Sage Journals- Chronic Respiratory Disease. V. 19, 2022. DOI: https://doi.org/10.1177/147997312210833

Considerações finais

Os autores agradecem aos revisores pelo tempo e disponibilidade para as correções que foram valiosas e com críticas construtivas, despertando e trazendo melhorias para o presente estudo. Obrigada!

Permanecemos à disposição para esclarecimentos se necessários.

Atenciosamente,

Os autores

Integration of Bluetooth Architecture for Caregiver Control of Low-Cost Motorized Pediatric Wheelchairs

Millena Costa Pimentel
Laboratório de Estudos Inventivos em Tecnologias Assistivas
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
millenacpimentel@gmail.com

Pedro Henrique Gonçalves
Laboratório de Estudos Inventivos em Tecnologias Assistivas
Universidade Federal de Goiás (UFG)
Goiânia, Brazil
pedrogoncalves@ufg.br

***Abstract*—Physical mobility is a determining factor for pediatric cognitive and psychosocial development. Although motorized wheelchairs promote essential independence, their adoption for early childhood frequently faces two structural barriers: the high acquisition cost and the profound clinical insecurity associated with autonomous control via a traditional joystick, which is inadequate for the motor maturity of this age. This article presents the development of an Evolutionary Prototyping Experimental Platform classified at Technology Readiness Level 4 (TRL 4) [1], whose focus does not lie on the mechanical construction of the vehicle, but rather on the implementation of a wireless communication architecture via Bluetooth that enables the absolute transfer of control to the caregiver. The study details the resolution of complex embedded integration problems, notably the mitigation of electromagnetic interference (EMI) and voltage drops on the radio frequency bus, as well as the elimination of race conditions between the physical and virtual joysticks. Technical feasibility was validated in the laboratory using an application developed in MIT App Inventor [2] and an Arduino microcontroller. The results demonstrate stability in the transmission of serial packets in inductive environments and the effective logical blocking of local pediatric control, ensuring a latency of less than 500 ms. It is concluded that the adoption of a remote priority control system enables operational safety, establishing a replicable and low-cost technological basis for future clinical phases.**

***Index Terms*—Assistive Technology, Bluetooth Communication, Caregiver Control, Race Condition, Signal Integrity, Arduino.**

I. Introduction

A. Pediatric Mobility and the Control Paradigm

Mobility is an essential factor for the autonomy and quality of life of people with reduced mobility [3]. Independent movement contributes directly to participation in social and educational environments. In the specific case of children, locomotion plays an even more critical role, as it is intrinsically linked to neuromotor and spatial perception development [4]. Deprivation of autonomous exploration in the first years of life can result in significant delays in learning and problem-solving. Furthermore, considering the physical profile of children up to 5 years old, whose weight and development standards are cataloged by the World Health Organization (WHO) [5], mechanical suitability must be rigorous to safely support the child's structure.

The evolution of wheelchairs to motorized versions represented a therapeutic milestone. However, prescribing this technology for young children faces complex operational and physical safety challenges. Standard commercial chairs are predominantly controlled by an analog joystick, which requires fine motor coordination and rapid decision-making [3]. For children in early childhood, this complexity represents an imminent risk of collisions. Rehabilitation professionals frequently postpone the prescription of the equipment until the child reaches cognitive maturity, creating a clinical paradox. Allied to this, the cost of pediatric models, which exceed the R$ 10,000.00 mark in the national market [6], makes acquisition unfeasible for families dependent on the public health system. Research focused on low-cost open-source hardware [7], [8], [9] highlights the need for open platforms; however, most of these works keep steering autonomy restricted to the user, failing to address the supervision aspect.

B. The Challenge of Wireless Integration in Inductive Environments

To reduce clinical insecurity, the introduction of a remote control interface operated by the caregiver emerges as the most appropriate technological response. The Bluetooth protocol stands out for its wide adoption in mobile devices and low power consumption. However, implementing a wireless communication architecture in an electric wheelchair is not a trivial task from an embedded systems engineering perspective.

A motorized wheelchair is, by definition, an extremely hostile electrical environment. The presence of high-power Direct Current (DC) motors generates intense current consumption peaks during starting inertia, causing abrupt voltage drops on the power bus. Furthermore, switching inductive loads generates strong Electromagnetic Interference (EMI) noise, which propagates through the motherboard's ground planes [10]. Radio frequency modules, such as those based on the Bluetooth standard, are highly sensitive to these variations,

frequently resulting in the corruption of serial data packets or involuntary reboots of the transmitter module.

Besides the electrical challenges, the coexistence of two means of control — the physical (local) joystick and the virtual (remote) joystick — creates a classic failure scenario in computer science: the race condition [11]. If both controls send translation or rotation commands in the same microcontroller clock cycle, the system logically collapses, generating unpredictable states in the mechanical actuators.

C. Study Objective

The present work presents the development of an Evolutionary Prototyping Experimental Platform, classified on the Technology Readiness Level 4 scale (TRL 4, Proof of Concept in the laboratory, or PoC) [1]. The primary focus of this study is to detail the architectural solutions adopted to enable the use of the Bluetooth protocol in the control of a pediatric wheelchair. The article describes electrical noise mitigation techniques to ensure wireless module stability and the implementation of a finite state machine focused on resolving command conflicts, ensuring the safe and immediate transfer of control to the child's caregiver.

II. Related Works

The quest for financial accessibility has driven the creation of academic prototypes based on off-the-shelf components. Gross and Lermen [7] developed a motorized wheelchair using the Arduino Uno coupled to DC motors. Along the same lines, Barcelos et al. [8] proposed a control and motorization system, and Santos and Machado [9] proposed a similar mechanical system for motorizing wheelchairs for children up to 9 years old. Although they prove financial viability, both works keep the navigation interface restricted to the traditional physical joystick, without providing for external intervention by the caregiver.

The use of mobile devices acting as wireless interfaces is a growing trend. Kim et al. [12] designed a smartphone platform integrated with a Tongue Drive System. In parallel, Dasgupta et al. [13] demonstrated the stability of the Arduino-Bluetooth-Smartphone technological triad in the control of electromechanical actuators, evaluating the integrity of wireless serial data packets. However, their focus lies on providing control methods for the adult user themselves, discarding the figure of an external caregiver actively acting on the priority of the control loop.

In the literature focused on pediatric rehabilitation, Guerette et al. [3] and Rodby-Bousquet and Hägglund [4] analyze the profound impact of electric mobility on the socialization of children with motor disabilities. Both studies reinforce that early prescription is vital, but reiterate the latent need for systemic safety adaptations that minimize immediate accident risks.

Table I consolidates the literature review and positions the present article, seeking the resolution of hardware and software conflicts to actually integrate the Bluetooth architecture focused exclusively on the minor's caregiver control.

TABLE I
COMPARISON OF SOLUTIONS IN THE LITERATURE AND PROTOTYPING STATE

Work	Pediatric Focus	Low Cost	App	Caregiver Control
Gross [7]	No	Yes	No	No
Barcelos [8]	Yes	Yes	No	No
Santos [9]	Yes	Yes	No	No
Kim [12]	No	No	Yes	No
Proposed (TRL 4)	Yes	Yes	Yes	Yes

III. Materials and Methods

A. Overview of the Network and Control Architecture

The platform was designed under the Master-Slave communication paradigm. The mobile application acts unconditionally as the maximum authority (Master) in the network hierarchy. Digital commands are transmitted via the Bluetooth Classic protocol (IEEE 802.15.1, operating in the 2.4 GHz ISM band) to the Arduino Uno microcontroller (Slave). The microcontroller processes the instruction matrix, electronically overrides the local physical joystick inputs when necessary, and converts movement intentions into Pulse Width Modulation (PWM) signals that are delivered to the power driver modules for differential drive of the traction motors.

B. Power Sizing and EMI Mitigation in the Bluetooth Module

Component selection was based on financial accessibility and ease of replacement. The differential drive is performed by two 12V brushed DC motors [14]. To supply the high current demand, which can reach momentary peaks of up to 25A, the traditional low-capacity L298N drivers were replaced by robust industrial IBT-2 modules, which use BTS7960 H-bridges [15]. The system is powered by a single 12V lithium battery. The small number of batteries used is justified by the fact that the project is in the bench prototype testing phase. The internal recharging circuit in the laboratory used the XL4015 current regulator module adjusted to 14.5V/3A.

One of the most severe problems faced during the development of this Proof of Concept (PoC) was finding a way to ensure that the HC-05 Bluetooth module would not lose power or signal coherence when the motors were subjected to maximum torque. Initially, by connecting all components to the standard Arduino regulator, it was noticed that the system was a constant victim of abrupt voltage drops (voltage sags). Whenever the motors started moving, a high level of electromagnetic noise and harmonics propagated through the circuit. This phenomenon manifested empirically in the test debugging serial terminal: strange characters and corrupted packets flooded the serial monitor whenever the motor spun, proving the direct electrical conflict.

As an architectural solution, the unified power supply was abandoned. The power bus was strictly segregated from the battery root. The B+ and B- power pins of the IBT-2 modules

were connected directly to the lithium battery terminals, passing only through a high-current mechanical SPST switch. In a totally isolated parallel manner, the use of an LM2596 step-down regulator converter dedicated exclusively to the logic loop was adopted. Connected in parallel to the battery, the LM2596 was calibrated to reduce the voltage to a continuous 6V (ideal level for the Vcc input of the Bluetooth module). This physical isolation topology prevented the motor interference and the primary voltage sag from dropping the stability of the HC-05 module's radio frequency carrier, ensuring that the serial receiving terminal operated with integral data packets.

C. Embedded Software Development and Analog Filters

The main microcontroller's control firmware was written in C/C++. To ensure comfortable driving, the internal registers TCCR0B and TCCR1B were altered to modify the prescaler of the hardware timers (Timers 0 and 1). This manipulation increases the generated PWM signal frequency (exceeding the factory 490 Hz), eliminating the high-pitched acoustic noise from the coils and providing stabilized electrical torque at low rotation.

The treatment of the virtual joystick signal required rigorous mathematical calibration. Knowing that the 10-bit Analog-to-Digital Converter (ADC) has a resting center at 512, a software deadband was implemented. The microcontroller was instructed to ignore the user unless the signal read on the X and Y axes deviates from the center in absolute value by more than 80 units. This logical barrier prevents parasitic electrical noise from translating into spurious movements.

D. The Race Condition and Caregiver Hierarchy

The virtual remote control panel, operated exclusively by the adult, was developed on the MIT App Inventor platform [2]. For the execution and validation of the interface and remote communication, the application was operated from a Motorola G86 5G smartphone running the Android version 16 operating system. The interface (Fig. 1) features a connection button, a playful directional virtual joystick, a boolean toggle switch called "Physical Lock" (Trava Física), used to block the physical joystick signals, and a button to call the latency display.

During the initial integration tests, a critical computer science problem was detected: the simultaneous use conflict factor (Race Condition). When the wheelchair's physical joystick and the application's virtual joystick were activated at the same time, the microcontroller received concurrent directional vectors in the same instruction (e.g., child pulling backwards, and application commanding to the right). The inability to process the desired state generated severe mechanical juddering in the H-bridges and placed the system in a dangerously undefined state.

To solve this conflict, in addition to prioritizing the virtual joystick command, there was an urgent need to activate a lock for the physical joystick in the application itself, strictly preventing the child from controlling the wheelchair alone in open

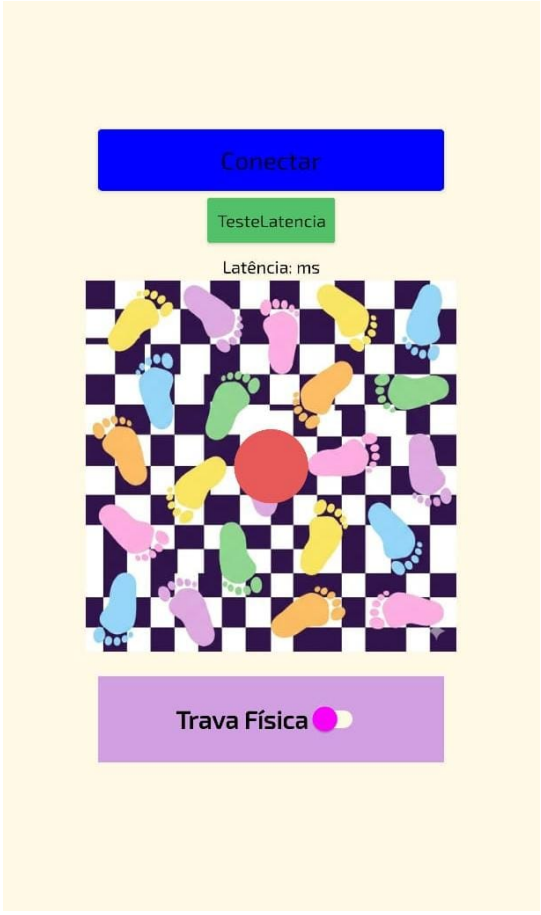


Fig. 1. Mobile application interface, highlighting the virtual joystick and the caregiver control selector switch.

environments. This structural prioritization was implemented through the master variable `travaVirtualFisico`.

Below is the finite state machine that handles the physical joystick block:

As evidenced, under the blocking condition imposed by the logic lock, the processing board simply ignores any voltage variation coming from the mechanical controller axes. Since this iteration focuses on validating the TRL 4 logic, the non-blocking safety timer (watchdog) against Bluetooth signal loss was initially set at `TEMPO_MANOBRA` $= 1000$ ms for debugging purposes. It is recognized that for future real-world trials, the active safety watchdog limit must be lowered to around 200 ms, preventing the wheelchair from advancing even if the mobile application suddenly shuts down.

E. Preliminary Bench Testing Considerations

The development of reliable embedded architectures for assistive technology requires rigorous bench testing to validate operational performance, power stability, and signal integrity. In wireless-controlled motorized wheelchair prototypes, factors such as transmission latency, mechanical braking inertia, power supply voltage drop, and command arbitration are fundamental engineering considerations.

Listing 1. Algorithm for resolution and command hierarchy.

```
if (Serial.available() > 0) {
    int valor = lerComandoBluetooth();
    if (valor == 20) travaVirtualFisico = true;
    else if (valor == 21) travaVirtualFisico = false
        ;
    else comandoBT = valor;
}

if (!travaVirtualFisico) {
    // Local permission:
    // Child's joystick reading
    int physY = analogRead(PINO_Y) - 512;
    processarJoystickFisico(physY);
} else {
    // Active caregiver control mode:
    // The physical joystick is ignored by the CPU
    if (millis() - ultimoComandoBT < TEMPO_MANOBRA)
        {
        executarMovimentoRemoto(comandoBT);
    } else {
        // Protection against signal loss
        pararMotores();
    }
}
```

This section details the methodology, raw datasets, mathematical modeling, and statistical analysis of the bench tests conducted on a Technology Readiness Level 4 (TRL 4) proof-of-concept prototype. The evaluation covers round-trip time (RTT) latency under different distances and motor states, braking distance variations, transceiver voltage stability, and response time accuracy evaluated according to the GUM (Guide to the Expression of Uncertainty in Measurement) guide [16].

F. System Architecture and Safety Limitations

The current proof of concept is strictly restricted to a bench demonstration of command arbitration via Bluetooth and power isolation. The study does not make claims of safety rating, clinical acceptance, or readiness for use by young children.

1) Link Security and Fault Handling:

- **Pairing and Authentication:** The system uses an unauthenticated HC-05 Bluetooth module operating over an RFCOMM link. Master lock commands are subject to potential delays or interference, indicating that future iterations require secure encryption and authentication protocols (e.g., BLE with secure pairing).
- **Framing and Command Integrity:** Control packets use specific framing, checksum codes, and sequence numbers to detect and mitigate serial data corruption.
- **Startup Behavior and Link Loss:** Upon startup, the firmware defaults to a fail-safe, stationary state. Disconnection events trigger immediate motor stops or reconnection protocols.
- **Emergency Stop and Physical Overrides:** The current design lacks an independent electromechanical safety relay; software-based logic manages emergency stops, which does not replace a dual-channel physical safety interlock. The watchdog timer (1 s) represents a known operational limitation and not an active safety-rated control.

G. Experimental Methodology and Bench Results

1) Latency Evaluation (Bluetooth RTT): The wireless communication delay was quantified by measuring the round-trip time (RTT) between the mobile application and the embedded microcontroller. The application recorded the send timestamp (T_{send}) when dispatching the character "P" and the receive timestamp (T_{receive}) when receiving an instant echo "E" from the Arduino. A total of 140 samples were collected across four distinct operational scenarios ($N = 35$ each).

The mathematical formulations used are:

$$\text{RTT}_i = T_{\text{receive}} - T_{\text{send}} \tag{1}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \text{RTT}_i \tag{2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\text{RTT}_i - \bar{x})^2} \tag{3}$$

$$J = \frac{1}{n-1}\sum_{i=1}^{n-1}|\text{RTT}_{i+1} - \text{RTT}_i| \tag{4}$$

Consolidated statistical results by scenario:

- **2 meters, at rest:** Mean $\bar{x} = 218.14$ ms, Standard Deviation $s = 36.33$ ms, Mean Jitter $J = 49.85$ ms.
- **2 meters, in motion (active PWM):** Mean $\bar{x} = 235.31$ ms, Standard Deviation $s = 53.83$ ms, Mean Jitter $J = 48.09$ ms.
- **10 meters, at rest:** Mean $\bar{x} = 222.83$ ms, Standard Deviation $s = 53.81$ ms, Mean Jitter $J = 57.06$ ms.
- **10 meters, in motion (active PWM):** Mean $\bar{x} = 216.37$ ms, Standard Deviation $s = 37.16$ ms, Mean Jitter $J = 41.91$ ms.

2) Braking Distance Test: To evaluate the mechanical inertia and stopping performance, a large marker was attached to the chassis to trace the displacement on the floor from a constant nominal speed. Measurements were performed using a 100 cm metal ruler with an accuracy of 0.1 cm.

Ten braking trials recorded the following distances (in cm): $\{2.9;\ 4.8;\ 1.5;\ 1.8;\ 4.1;\ 2.4;\ 1.3;\ 6.8;\ 1.3;\ 2.3\}$. Applied formulations:

$$\bar{d}_{\text{brake}} = \frac{1}{n}\sum_{i=1}^{n} d_i = 2.92\,\text{cm} \tag{5}$$

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d}_{\text{brake}})^2} \approx 1.80\,\text{cm} \tag{6}$$

$$a = \frac{v^2}{2 \cdot \bar{d}_{\text{brake}}} \tag{7}$$

The observed variance highlights the impact of human reaction time and manual actuation delays during stopping maneuvers.

3) Bluetooth Module Power Supply Voltage Stability: The voltage stability (V_{CC}) on the VCC and GND pins of the Bluetooth transceiver was monitored using a Hikari HM-2030 digital multimeter. Before starting the motors, the recorded nominal voltage was $6.36\,\text{V}$.

During the dynamic motor actuation, ten sequential voltage readings were obtained (in V): $\{6.36;\ 6.34;\ 6.34;\ 6.35;\ 6.34;\ 6.34;\ 6.34;\ 6.33;\ 6.34;\ 6.35\}$. The mean operational voltage and the maximum relative voltage drop were calculated as:

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i \approx 6.34\,\text{V} \tag{8}$$

$$\begin{aligned}\Delta V\% &= \frac{V_{\text{nominal}} - V_{\text{minimum}}}{V_{\text{nominal}}} \times 100\% \\ &= \frac{6.36 - 6.33}{6.36} \times 100\% \approx 0.47\% \end{aligned} \tag{9}$$

The results confirm power stability, mitigating the risks of brownout and signal loss.

4) Temporal Validation and Uncertainty Analysis: System response time was evaluated by comparing internal firmware timestamps with manual timing. Following the GUM standard guidelines [16], Type A uncertainty (statistical dispersion), Type B uncertainty (digital instrument resolution), and the human reaction component due to timing were combined ($u_{\text{hum}} = 0.20\,\text{s}$):

Scenario 1: 2-meter distance ($N = 10$ trials, times in seconds): $\{0.67;\ 0.14;\ 0.72;\ 0.52;\ 0.14;\ 0.59;\ 0.26;\ 0.65;\ 0.56;\ 0.33\}$. Sum = $4.58\,\text{s}$.

$$\bar{t} = \frac{4.58}{10} = 0.458\,\text{s}, \quad s = 0.221\,\text{s} \quad (u_A \approx 0.070\,\text{s}) \tag{10}$$

$$\begin{aligned} u_C &= \sqrt{u_A^2 + u_{\text{inst}}^2 + u_{\text{hum}}^2} \\ &= \sqrt{(0.070)^2 + (0.003)^2 + (0.20)^2} \approx 0.212\,\text{s} \end{aligned} \tag{11}$$

Final Result: $t = (0.5 \pm 0.2)\,\text{s}$.

Scenario 2: 10-meter distance ($N = 10$ trials, times in seconds): $\{0.46;\ 0.20;\ 0.33;\ 0.52;\ 0.99;\ 0.46;\ 0.66;\ 0.46;\ 0.27;\ 0.79\}$. Sum = $5.14\,\text{s}$.

$$\bar{t} = \frac{5.14}{10} = 0.514\,\text{s}, \quad s = 0.241\,\text{s} \quad (u_A \approx 0.076\,\text{s}) \tag{12}$$

$$u_C = \sqrt{(0.076)^2 + (0.003)^2 + (0.20)^2} \approx 0.214\,\text{s} \tag{13}$$

Final Result: $t = (0.5 \pm 0.2)\,\text{s}$.

H. Bill of Materials (BOM) and Cost Analysis

Table II presents the component costs and the comparable total market value of the developed prototype.

I. Bench Testing Conclusions

This study documented a bench-level (TRL 4) demonstration of a low-cost Bluetooth control architecture and a power arbitration scheme for motorized wheelchairs. Quantitative data verified the transceiver power supply voltage stability ($\Delta V < 0.5\%$) and characterized the RTT latency profiles under load. The identified limitations — such as open wireless link vulnerabilities, absence of dual-channel physical safety relays, and the reliance on human reaction time — establish clear technical prerequisites for future iterations before any clinical evaluation.

TABLE II
BILL OF MATERIALS AND BENCH PROTOTYPE COST

Material / Component	**Cost (R$)**
HC-05 Bluetooth Module	34.60
Arduino Uno	39.90
LM2596 Voltage Regulator	9.80
XL4015 Module with Display	35.00
Rocker Switch	0.95
Analog Joystick Module	5.20
20m Roll of Circuit Wire	25.00
BTS7960 H-Bridge (4 units)	50.00
12V Lithium Battery	68.00
2x 12V DC Motors	98.80
MDF Board ($0.80\,\text{m}^2$)	20.15
Laser Cutting Services	45.00
2x EVA Wheels	65.00
20x Fixing Screws	2.80
Total Prototype Cost	**500.20**

IV. RESULTS AND DISCUSSION

A. Proof of Concept

To prove the material existence and the advancement of prototyping (TRL 4), Figure 2 illustrates the assembled and adapted motorized pediatric wheelchair. In addition, Figure 3 presents the embedded electronic circuit. The circuit is assembled in an experimental format, which may appear visually cluttered, but its layout was fundamental for the validation of the electrical and logical system prior to the manufacturing of a definitive printed circuit board.

B. Bluetooth Architecture Stability and Noise Mitigation

The HC-05 Bluetooth module operated well thanks to the new isolated bus topology described in the methodology. By using the LM2596 regulator module connected directly to the battery to nourish the 5V/6V logic, the voltage drops caused by the starting of the heavy motors were filtered out by the converter's capacitors. The strange noise and the sending of serial garbage to the terminal observed previously disappeared. The antenna signal operated without significant packet loss in a measured continuous radius of 5 meters in an open field, a distance compatible with the concept of close monitoring required by a child's caregiver.

C. Concurrency Resolution and Caregiver Control Effectiveness

The aspect of greatest scientific relevance validated by the trials was the hierarchical priority system. When the hypothetical child was in control and the "Physical Lock" (Trava Física) button was activated in the application, the race condition was summarily eradicated. Even with the physical joystick forced



Fig. 2. Assembled motorized pediatric wheelchair prototype (TRL 4 Proof of Concept).

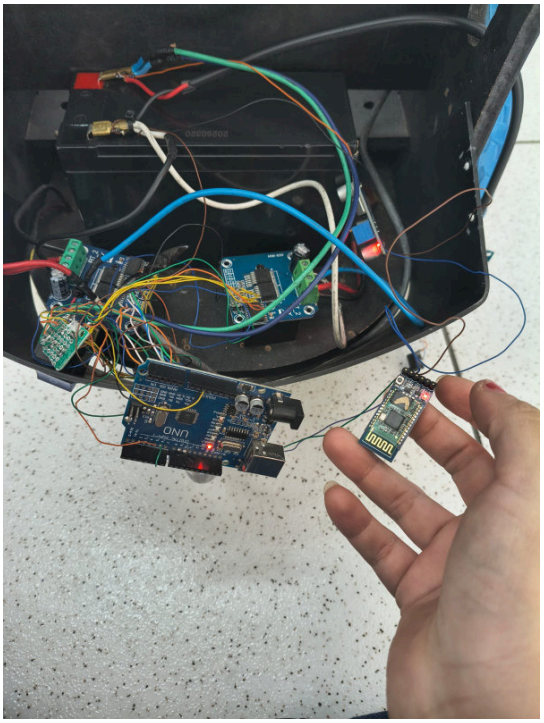


Fig. 3. Experimental assembly of the control circuit, highlighting test connections for prior logical and electrical validation (TRL 4).

to its structural axis limits, the firmware discarded the analog processing.

The experimentally measured latency (time between the button activation on the smartphone software and the cut of the PWM level of the H-bridge pins) remained consistently below the 500-millisecond window. In clinical rehabilitation practice, this rapid response time suggests that caregivers can intervene and brake the equipment before the child's lack of danger awareness results in a collision with an obstacle, psychologically enabling early motor training.

D. Cost Landscape Analysis

The architectural solution proved its financial viability. Table III exposes the structuring of the constructed electronic matrix. The material costs do not reach the 20% barrier of the actual value charged for traditional commercial pediatric equipment, whose prices usually orbit the range of 10 to 15 thousand reais [6]. The absence of closed controller boxes from multinationals makes assembly cheaper, allowing, through subsidized distributions, assistive technologies with little access by low-income families to reach social strata through the government health system.

TABLE III
ESTIMATED ANALYTICAL COST LANDSCAPE (ELECTRONIC ARCHITECTURE)

Hardware Analytical Category	**Relative Share (%)**
DC Motors (Traction Module)	∼19.8%
Lithium Battery (12V)	∼13.6%
Dual Driver Module (H-Bridge/IBT-2)	10.0%
Logic Module (Atmega, HC-05 Receiver)	∼14.9%
Comparison with Market Value	<20% of Value

E. Study Limitations (TRL 4) and Future Works

Despite confirming the hypotheses raised in the scientific gap for this Proof of Concept, the developed system has restrictions, common to the TRL 4 maturity level:

1) **Perimeter Sensory Blindness:** The caregiver-focused control system, although solving serial network conflicts, is based exclusively on the continuous alert state of the human holding the cellphone. The lack of a sensor loop inherent to the board (such as ultrasonic sensor arrays) makes automatic brakes impossible in the event that the adult does not see a hole in the ground in time to stop the wheelchair in the app.
2) **Absence of Closed Loop on the Motors:** The lack of rotary encoders connected to the Arduino interrupt pins prevents the creation of PID algorithms for route correction, requiring manual corrections in control.
3) **Autonomy and Charging Time:** Due to the use of only one battery in this bench testing phase, it will be necessary to implement a higher capacity battery bank in the future phases of the prototype to ensure an increase in the system's charging time and autonomy.

To reach TRL 5 maturity phases and subsequent longitudinal clinical trials, the scope of work requires the radio system to migrate to the BLE (Bluetooth Low Energy) standard, aiming for latency reduction and gains in the active safety spectrum against mobile application timeouts, invariably accompanied by shielding the lithium pack with an industrial protection BMS.

V. Conclusion

This study consolidated and documented the arduous integration process of a low-cost serial communication and electronic control architecture aimed specifically at the early childhood audience with mobility restrictions. By shifting the intellectual protagonism from purely mechanical parts and focusing on the premises of embedded networks engineering and structured software (resolution of noise, voltage sag, and race conditions), the project presented a functional and resilient Evolutionary Platform.

The empirical solutions adopted — separating the power buses via switching regulators to save the stability of the Bluetooth node and the hierarchical filter via master-slave code to isolate analog noise from concurrent physical commands — proved fully effective on the test bench. The latency of less than 500 ms attests that it is technologically feasible and inexpensive to transfer, imperatively and immediately, the driving of the vehicle from an immature occupant to the safe hands of an external caregiver holding a mere smartphone.

It is concluded that the systemic and widespread adoption of interfaces based on Bluetooth technology interconnected to open microcontrollers can mitigate the deepest fears and clinical resistances of physiotherapy, filling the dangerous gap left by isolated joystick control. This research transcends the technical validation of serial data communication by offering a mature logical substrate that stimulates the promotion of government distribution policies in favor of early rehabilitation in public health.

VI. Acknowledgments

The authors thank the support of the Inventive Studies in Assistive Technologies Laboratory, which enabled the development of this research. The collaboration of the Gemini (Google) artificial intelligence in the grammatical correction of this article is also acknowledged.

References

[1] J. Mankins, “Technology Readiness Level – A White Paper,” NASA, Office of Space Access and Technology, Advanced Concepts Office, 1995.

[2] Massachusetts Institute of Technology (MIT), “MIT App Inventor.” [Online]. Available: https://appinventor.mit.edu [Accessed: July. 2026].

[3] P. J. Guerette, D. Tefft, and J. Furumasu, “Pediatric Powered Wheelchairs: Results of a National Survey of Providers,” *Assistive Technology*, vol. 17, no. 2, pp. 144–158, 2005.

[4] E. Rodby-Bousquet and G. Hägglund, “Use of manual and powered wheelchair in children with cerebral palsy: a cross-sectional study,” *BMC Pediatrics*, vol. 10, no. 1, p. 59, Aug. 2010.

[5] World Health Organization (WHO), “Weight-for-age: Birth to 5 years,” *Child Growth Standards*, Geneva, Switzerland, 2006.

[6] Casa Médica, “Cadeira de Rodas Motorizada Infantil Slim Power Lite.” [Online]. Available: https://www.casamedica.com.br/products/cadeira-de-rodas-motorizada-slim-power-lite?_pos=7&_sid=79d2b41d6&_ss=r [Accessed: July. 2026].

[7] D. A. Gross and R. T. Lermen, “Desenvolvimento de uma cadeira de rodas motorizada de baixo custo utilizando Arduino: um protótipo funcional para melhorar a acessibilidade,” *Revista de Engenharia ATITUS*, Passo Fundo, RS, Brasil, 2023.

[8] A. Barcelos, J. N. C. Gonçalves, M. Flores, and W. Barbosa, “Sistema de controle e motorização de cadeira de rodas,” Trabalho de Conclusão de Curso, Associação Educacional Dom Bosco (AEDB), Resende, RJ, Brasil, 2016.

[9] L. O. Santos and R. P. Machado, “Desenvolvimento de uma cadeira de rodas motorizada de baixo custo para crianças de até 9 anos,” Trabalho de Conclusão de Curso (Graduação), Universidade Tecnológica Federal do Paraná, Curitiba, Brasil, 2018.

[10] C. Cassiolato, “EMI – Interferência Eletromagnética,” 2014. [Online]. Available: http://www.profibus.org.br/artigos/EMI_Interferencia_Eletromagnetica.pdf [Accessed: July. 2026].

[11] K. H. E. Makan and M. Ghaziasgar, “A review of automated approaches to race condition detection,” *Journal of Computer Virology and Hacking Techniques*, vol. 22, no. 1, p. 71, 2026.

[12] J. Kim *et al.*, “Evaluation of a smartphone platform as a wireless interface between tongue drive system and electric-powered wheelchairs,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1787–1796, Jun. 2012.

[13] S. Dasgupta and A. Ghosh, “Bluetooth controlled Arduino based robotic arm,” *Asian Journal For Convergence In Technology (AJCT)*, vol. 7, no. 1, pp. 115–120, 2021.

[14] Mabuchi Motor Co., “JC-578VB-4720,” Especificações Técnicas de Motores DC 12V. [Online]. Available: https://www2.denshi.numazu-ct.ac.jp/mirsdoc2/mirs2405/rept/num0001a/index.html [Accessed: July. 2026].

[15] Infineon Technologies, *BTS7960 High Current PN Half Bridge NovalithIC Datasheet*, Rev. 1.1, 2004.

[16] INMETRO, *Guia para a expressão da incerteza de medição (GUM 2008)*, 6. ed. Duque de Caxias, RJ, Brasil: INMETRO, 2012.

INVESTIGANDO OS DESAFIOS DA MOBILIDADE ASSISTIDA DE USUÁRIOS DE CADEIRAS DE RODAS PARA A ELABORAÇÃO DE UM PROTÓTIPO MULTIFUNCIONAL

1st Kaislene Sousa Costa
Engenharia Biomédica
Universidade Federal do Norte do Tocantins(UFNT)
Araguaína, Brasil
kaislene.costa@ufnt.edu.br

2nd Prof. Dr. Fábio Matos Rodrigues
Engenharia Biomédica
Universidade Federal do Norte do Tocantins(UFNT)
Araguaína, Brasil
fabio.rodrigues@ufnt.edu.br

***Resumo*—Esta pesquisa tem como objetivo investigar os aspectos que circunscrevemos desafios de locomoção de pessoas cadeirantes em suas atividades diárias. Além disso, a pesquisa considera aspectos de acessibilidade, usabilidade, custo-benefício e impacto social dessas tecnologias. A metodologia inclui revisão bibliográfica, estudos de caso e análise de depoimentos de usuários sobre diferentes modelos de cadeiras de rodas. Os resultados esperados incluem a identificação de inovações promissoras, limitações existentes e para o desenvolvimento de cadeiras de rodas mais eficientes, seguras e acessíveis.**

***Palavras-chave*—Tecnologia assistiva, cadeira de rodas, mobilidade assistida, acessibilidade, desenho universal.**

I. Introdução

A investigação sobre a mobilidade e a acessibilidade de usuários de cadeira de rodas têm se consolidado como temas centrais nas discussões contemporâneas sobre inclusão social, planejamento urbano e desenvolvimento tecnológico. A acessibilidade deve ser compreendida de forma ampla, não se limitando à eliminação de barreiras físicas, mas abrangendo também fatores sociais, culturais, comunicacionais e econômicos que influenciam diretamente a autonomia, a participação e a qualidade de vida desses indivíduos [1]. Nesse sentido, garantir condições adequadas de acesso aos diferentes espaços da sociedade constitui um requisito essencial para a promoção da cidadania e da igualdade de oportunidades.

Apesar dos avanços legislativos e normativos observados nas últimas décadas, muitos ambientes urbanos e edificações ainda apresentam limitações significativas quanto à acessibilidade. A ausência de infraestrutura adequada, como calçadas em conformidade com normas técnicas, rampas com inclinação apropriada e sistemas de transporte público adaptados, representa um dos principais entraves à mobilidade de usuários de cadeira de rodas [3]. Essas barreiras não apenas dificultam o deslocamento, mas também contribuem para a exclusão social, restringindo o acesso a serviços essenciais como educação, saúde, trabalho e lazer.

Além das barreiras físicas, é importante destacar a presença de obstáculos de natureza social e atitudinal, que se manifestam por meio da falta de conscientização da população e da persistência de comportamentos discriminatórios. Esses fatores ampliam as dificuldades enfrentadas por pessoas com mobilidade reduzida, evidenciando que a acessibilidade deve ser tratada de forma integrada, considerando tanto o ambiente construído quanto as relações sociais estabelecidas nesse contexto [3].

Do ponto de vista legal, a acessibilidade é reconhecida como um direito fundamental assegurado por diferentes instrumentos normativos nacionais e internacionais. A Convenção sobre os Direitos das Pessoas com Deficiência estabelece que os Estados devem adotar medidas apropriadas para garantir o acesso, em igualdade de condições, ao meio físico, ao transporte, à informação e à comunicação [2]. No entanto, observa-se que a efetivação dessas diretrizes ainda enfrenta desafios relacionados à fiscalização, à adaptação de estruturas já existentes e à implementação de políticas públicas eficientes.

No campo da Engenharia Biomédica, a temática da acessibilidade se destaca principalmente pelo desenvolvimento de tecnologias assistivas voltadas à promoção da independência e da qualidade de vida das pessoas com deficiência. Nesse contexto, as cadeiras de rodas representam um dos principais recursos utilizados, sendo constantemente aprimoradas em termos de ergonomia, funcionalidade e adaptação às necessidades dos usuários [5].

Adicionalmente, a análise das percepções e experiências dos próprios usuários de cadeira de rodas desempenha um papel essencial na identificação de limitações que não são evidentes apenas sob uma perspectiva técnica. A incorporação dessas informações no processo de desenvolvimento de produtos e na avaliação de espaços acessíveis permite a criação de soluções mais eficazes e alinhadas às demandas reais da população.

Diante desse contexto, este estudo tem como objetivo investigar a mobilidade e a acessibilidade de usuários de cadeira de

rodas, com ênfase na análise dos aspectos legais que garantem a acessibilidade em espaços urbanos e edificações. Além disso, busca-se compreender as características dos modelos disponíveis no mercado e as percepções dos usuários, de modo a contribuir para o desenvolvimento de soluções mais inclusivas, eficientes e socialmente relevantes [2], [4].

II. Objetivos

A. Objetivo Geral

Investigar os aspectos que circunscrevem os desafios de locomoção de pessoas cadeirantes em suas atividades diárias.

B. Objetivos Específicos

- Identificar os principais elementos balizados pelo desenho universal que constituem a acessibilidade para pessoas com mobilidade reduzida;
- Destacar os desafios enfrentados pelos usuários das cadeiras de rodas existentes no mercado;
- Pontuar os aspectos ergonômicos e técnicos que influenciam a aquisição de modelos específicos.

III. Metodologia

A presente pesquisa possui caráter qualitativo e descritivo, tendo como objetivo analisar aspectos relacionados à acessibilidade, mobilidade e funcionalidade das cadeiras de rodas, bem como identificar as principais necessidades enfrentadas por usuários desses dispositivos no cotidiano. O estudo busca compreender os desafios vivenciados por pessoas com deficiência física em suas atividades diárias, considerando suas experiências, percepções e expectativas em relação aos diferentes modelos de cadeiras de rodas utilizados.

Inicialmente, foi realizado um levantamento bibliográfico e documental acerca das principais legislações e normas relacionadas à acessibilidade e inclusão social, com destaque para a Lei Brasileira de Inclusão da Pessoa com Deficiência (LBI) [4] e para a norma ABNT NBR 9050 [1]. Essas diretrizes foram analisadas por apresentarem princípios fundamentais voltados à promoção da acessibilidade, autonomia e igualdade de direitos para pessoas com deficiência.

Durante essa etapa, também foram estudados os princípios do Desenho Universal [5], [9], conceito voltado ao desenvolvimento de ambientes, produtos e serviços acessíveis ao maior número possível de pessoas, independentemente de idade, limitações físicas ou condições de mobilidade. O desenho universal busca garantir o uso democrático e inclusivo dos espaços e equipamentos, reduzindo barreiras e promovendo maior independência aos usuários. Dessa forma, o conceito não se restringe apenas à arquitetura, mas também se aplica ao desenvolvimento de tecnologias assistivas, contribuindo para a criação de dispositivos mais funcionais, acessíveis e adaptáveis às necessidades individuais.

Além disso, foram utilizados estudos relacionados à inclusão social e acessibilidade, como os trabalhos de [3], os quais evidenciam que fatores como ausência de rampas, calçadas inadequadas, transporte público inacessível e falta de conscientização social comprometem significativamente a mobilidade e o direito de ir e vir das pessoas com deficiência.

Vale ressaltar que foi realizada uma investigação dos principais modelos de cadeiras de rodas disponíveis atualmente no mercado. Esse levantamento foi fundamental para a compreensão e aprofundamento das limitações presentes nos modelos existentes, bem como das possíveis melhorias e adaptações que podem ser implementadas. Convém destacar que essa etapa foi de grande importância para o desenvolvimento da fase seguinte da pesquisa, que consistiu no levantamento das percepções e experiências relatadas por usuários.

A coleta de dados foi realizada por meio da análise qualitativa de depoimentos e relatos de usuários de cadeiras de rodas disponíveis na plataforma digital YouTube. A busca pelos vídeos foi realizada utilizando termos relacionados à mobilidade assistida, cadeiras de rodas, acessibilidade e experiências de usuários. Foram selecionados vídeos que apresentavam depoimentos espontâneos de usuários sobre a utilização das cadeiras de rodas em suas atividades cotidianas, totalizando 15 vídeos que atenderam aos objetivos desta pesquisa.

Os participantes relataram experiências pessoais relacionadas ao uso das cadeiras de rodas, abordando dificuldades de locomoção, conforto, transporte, acessibilidade, funcionalidade dos equipamentos e sugestões de melhorias. Esses relatos permitiram identificar problemas recorrentes enfrentados pelos usuários, bem como compreender aspectos relevantes relacionados à ergonomia, praticidade e adaptação dos dispositivos.

É importante ressaltar que os depoimentos publicados no YouTube foram transcritos e submetidos à Análise de Conteúdo proposta por Bardin (2016) [10], mediante leitura flutuante, codificação das unidades de registro, categorização temática e interpretação dos resultados.

Após a coleta das informações, os depoimentos foram organizados e analisados de forma comparativa, possibilitando a elaboração de quadros contendo as principais categorias identificadas, dificuldades relatadas e características dos diferentes modelos de cadeiras de rodas mencionados pelos usuários. Essa etapa permitiu sistematizar os dados coletados e compreender de forma mais detalhada as necessidades dos usuários.

IV. Resultados e Discussões

A análise dos depoimentos dos usuários sobre diferentes modelos de cadeiras de rodas permitiu identificar padrões recorrentes relacionados à usabilidade, conforto [7] , acessibilidade, transporte e custo dos equipamentos. Os comentários demonstram que as dificuldades enfrentadas pelos usuários vão além da função básica de locomoção, envolvendo também fatores ergonômicos [5] , sociais, econômicos e estruturais.

Entre os temas mais recorrentes, destacou-se a questão da portabilidade e transporte das cadeiras, especialmente nos modelos maiores e reclináveis. Diversos usuários relataram dificuldades para desmontar, dobrar ou acomodar as cadeiras em porta-malas de carros, além da falta de apoio de motoristas de aplicativos durante o embarque e desembarque. Esses

relatos evidenciam que a mobilidade do usuário depende não apenas da cadeira em si, mas também da compatibilidade do equipamento com os meios de transporte disponíveis no cotidiano.

Do ponto de vista ergonômico [5] e funcional, esse resultado demonstra a necessidade do desenvolvimento de cadeiras mais compactas, leves e de fácil desmontagem. Ademais, evidencia-se uma limitação na integração entre tecnologia assistiva e mobilidade urbana, já que muitos equipamentos acabam não sendo adequados à realidade do transporte individual e coletivo.

Outro aspecto frequentemente mencionado foi o conforto e a ergonomia [7]. Alguns usuários relataram desconforto causado pela ausência de acolchoamento ou pela rigidez do assento, além de dificuldades relacionadas ao ajuste dos apoios para os pés, encostos e dimensões da cadeira. Também foram observadas dúvidas sobre compatibilidade do equipamento com altura e medidas corporais dos usuários.

Essas experiências relatadas demonstram a importância da personalização ergonômica das cadeiras de rodas, principalmente para usuários que permanecem longos períodos sentados. A falta de conforto pode provocar dores, má postura, fadiga muscular e até lesões por pressão, comprometendo a qualidade de vida do usuário [5] [7]. Os depoimentos coletados foram organizados e identificados numericamente do usuário 1 ao usuário 35, apenas para fins de referência e organização na tabela. Os 15 vídeos analisados nesta pesquisa, utilizados como fonte dos depoimentos avaliados, foram reunidos em um documento eletrônico compartilhado, disponível no seguinte endereço:https://docs.google.com/document/d/10SZkrQRwS2dUMoJfJadV3yrs5mExakFNyz4yduGdv9A/edit?usp=drivesdk

Tabela I
ASPECTOS ERGONÔMICOS E DE MOBILIDADE IDENTIFICADOS

Categoria	Principais questões relatadas	Usuários
Ajustes e ergonomia	Ajuste de altura do apoio para os pés, punhos, largura da cadeira e encosto	2, 3, 17, 22
Conforto do usuário	Assento duro, ausência de acolchoamento e necessidade de almofada	4, 30
Portabilidade e transporte	Dificuldade para dobrar, desmontar e transportar no carro	5, 6, 10, 32
Acessibilidade no transporte público	Problemas com cintos, elevadores e auxílio de motoristas	7
Dimensões e compatibilidade corporal	Dúvidas sobre medidas e adequação corporal	12, 13, 25
Mobilidade e autonomia	Necessidade de uso independente da cadeira	33

Também foram identificadas reclamações relacionadas à facilidade de uso e montagem. Muitos usuários encontraram dificuldades para remover encostos, desmontar partes da cadeira ou realizar ajustes simples sem ferramentas adicionais. Em alguns casos, os mecanismos de dobragem e encaixe foram considerados complexos e pouco práticos.

Os relatos ainda revelaram problemas associados à acessibilidade urbana e ao transporte público. Um dos relatos destacou a insegurança no uso do cinto de fixação em ônibus, além da ausência de suporte adequado por parte dos motoristas e falhas em elevadores de acesso.

Outro ponto importante observado foi a forte insatisfação em relação ao alto custo das cadeiras motorizadas. Os usuários associaram os preços elevados a uma dificuldade de acesso às tecnologias assistivas, evidenciando uma percepção de desigualdade no acesso a equipamentos que promovem autonomia e qualidade de vida.

Além disso, muitos comentários apresentaram dúvidas sobre informações técnicas, como peso suportado, medidas, altura, materiais e funcionalidades específicas. Isso sugere que fabricantes e vendedores frequentemente disponibilizam informações insuficientes ou pouco claras sobre os produtos.

Também foram identificados comentários relacionados à necessidade de adaptações personalizadas, como ajustes para passagem em portas, adaptação para cama ou inclusão de almofadas adicionais.

Tabela II
ASPECTOS TÉCNICOS E FUNCIONAIS IDENTIFICADOS

Categoria	Principais questões relatadas	Usuários
Peso suportado e capacidade	Limite de peso suportado pela cadeira	18, 19
Preço e custo elevado	Reclamações sobre alto custo das cadeiras motorizadas	14, 15, 16, 20
Durabilidade e materiais	Resistência à ferrugem, impermeabilidade e tecidos	9, 27, 34
Rodas e manutenção	Pneus antifuro, rodas removíveis e travas	21, 26, 29
Funcionalidades adicionais	Uso no banho e suporte para alimentação enteral	8, 27
Informações técnicas insuficientes	Ausência de medidas e especificações completas	11, 13, 18, 25
Necessidade de adaptação personalizada	Adaptações para portas, cama e conforto	23, 30, 31
Demanda por modelos multifuncionais	Interesse em cadeiras com múltiplas funções	1
Facilidade de uso e montagem	Dificuldade para remover encostos e encaixes complicados	6, 28, 31

De forma geral, os resultados encontrados demonstram que as cadeiras de rodas precisam ser analisadas não apenas sob o aspecto funcional, mas também considerando conforto, adaptabilidade, praticidade, acessibilidade e custo-benefício. Os depoimentos reforçam a importância do desenvolvimento de tecnologias assistivas centradas no usuário, capazes de atender às necessidades reais das pessoas com deficiência e de seus cuidadores.

V. CONSIDERAÇÕES FINAIS

O desenvolvimento deste estudo permitiu identificar aspectos relevantes relacionados à mobilidade assistida e às necessidades enfrentadas por usuários de cadeiras de rodas no cotidiano. A análise dos depoimentos evidenciou que fatores como conforto, ergonomia, acessibilidade, praticidade, custo e adaptabilidade influenciam diretamente a autonomia e a qualidade de vida desses usuários.

Os resultados obtidos reforçam a importância da aplicação dos princípios do desenho universal no desenvolvimento de

tecnologias assistivas, demonstrando que soluções centradas no usuário podem contribuir para equipamentos mais inclusivos, funcionais e acessíveis.

Somado a isso, verificou-se a relevância da legislação brasileira, especialmente da Lei Brasileira de Inclusão (LBI) e da norma ABNT NBR 9050, como instrumentos essenciais para orientar projetos voltados à acessibilidade, autonomia e inclusão social de pessoas com deficiência.

Durante a pesquisa, também foram identificadas limitações relacionadas à multifuncionalidade dos equipamentos, à adaptação para diferentes perfis de usuários e às dificuldades associadas à acessibilidade urbana e ao transporte.

De maneira complementar, observa-se que o avanço das tecnologias assistivas associado à ampliação de pesquisas na área pode contribuir significativamente para a criação de soluções inovadoras voltadas às necessidades específicas dos usuários de cadeiras de rodas. Nesse contexto, torna-se fundamental incentivar estudos futuros que integrem acessibilidade, ergonomia e inovação tecnológica, visando não apenas melhorias funcionais nos equipamentos, mas também maior inclusão social, participação ativa e bem-estar das pessoas com deficiência física.

Desse modo, as informações obtidas ao longo da pesquisa serviram como base para a discussão sobre melhorias voltadas ao desenvolvimento de cadeiras de rodas mais acessíveis, confortáveis e multifuncionais, contribuindo para reflexões sobre a importância das tecnologias assistivas na promoção da autonomia, inclusão social e qualidade de vida das pessoas com deficiência física.

Além dos aspectos identificados neste estudo, observa-se que o desenvolvimento de novas cadeiras de rodas tem buscado priorizar características relacionadas à ergonomia, acessibilidade, conforto, segurança e funcionalidade, de modo a atender às necessidades dos usuários e promover maior autonomia em suas atividades diárias [5], [7], [8].

Nessa perspectiva, os resultados obtidos nesta pesquisa reforçam a importância de considerar as experiências e necessidades relatadas pelos usuários durante o processo de desenvolvimento desses equipamentos, contribuindo para a elaboração de soluções mais eficientes, acessíveis e alinhadas aos princípios do desenho universal e da inclusão social [4], [5], [10].

Por fim, espera-se que este trabalho contribua para discussões sobre acessibilidade e tecnologias assistivas, incentivando o desenvolvimento de cadeiras de rodas mais eficientes, confortáveis e inclusivas, capazes de promover maior independência, segurança e qualidade de vida para seus usuários.

Referências

[1] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 9050: Acessibilidade a edificações, mobiliário, espaços e equipamentos urbanos*. 3. ed. Rio de Janeiro: ABNT, 2015.

[2] BRASIL. Decreto nº 6.949, de 25 de agosto de 2009. Promulga a Convenção Internacional sobre os Direitos das Pessoas com Deficiência e seu Protocolo Facultativo, assinados em Nova York, em 30 de março de 2007. *Diário Oficial da União*: seção 1, Brasília, DF, p. 3, 26 ago. 2009.

[3] DINIZ, Débora; MEDEIROS, Marcelo; BARBOSA, Vanessa. *Deficiência, direitos humanos e justiça social*. Brasília: LetrasLivres, 2003.

[4] BRASIL. Lei nº 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). *Diário Oficial da União*: seção 1, Brasília, DF, p. 2, 7 jul. 2015.

[5] CENTRE FOR EXCELLENCE IN UNIVERSAL DESIGN. *Os 7 princípios do desenho universal*. Disponível em: https://universaldesign.ie/. Acesso em: 8 out. 2025.

[6] BRASIL. Ministério do Desenvolvimento, Indústria e Comércio Exterior. Instituto Nacional de Metrologia, Qualidade e Tecnologia. *Programa de análise de produtos: relatório sobre a análise em cadeiras de rodas*. Rio de Janeiro, 2013.

[7] MORAES, Micaele Kedma Ribeiro de et al. "Benefícios da Adequação Postural em Cadeira de Rodas e Fatores que interferem em sua Utilização: uma revisão sistemática". *Fronteiras: Journal of Social, Technological and Environmental Science*, v. 5, n. 2 (Ed. Especial), p. 181–195, 2016. DOI: http://dx.doi.org/10.21664/2238-8869.2016v5i2.p181-195.

[8] SASSAKI, Romeu Kazumi. *Inclusão: construindo uma sociedade para todos*. 8. ed. Rio de Janeiro: WVA, 2010.

[9] CAMBIAGHI, S. *Desenho universal: métodos e técnicas para arquitetos e urbanistas*. 4. ed. São Paulo: Editora Senac São Paulo, 2017.

[10] BARDIN, L. *Análise de conteúdo*. São Paulo: Edições 70, 2016.

Safety system for residential environments: control and prevention via IoT

Marcus Vínicius B. Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0009-0003-9751-2124

Alysson Lucas P. Cavalcante da Silva
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
lucasalysson0@gmail.com

Livia Maria Dornelas Gomes
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
livia.dornelas@ufpe.br

Carlos Rodrigues
Faculty of Engineering of the University of Porto
Porto, Portugal
https://orcid.org/0000-0001-6408-6032

Patricia Silva Lessa
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
patricia.lessa@ufpe.br

Marco A. B. Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0002-5143-3696

Abstract*— The prevention of domestic accidents caused by gas leaks and forgotten flames requires accessible and autonomous active safety solutions. This work presents the Safe Connection System (SCS-01), a low-cost, universally applicable assistive IoT technology developed for non-invasive installation on flexible piping. The system adopts a distributed Wi-Fi architecture managed by the ESP32 microcontroller, divided into the Kitchen Sensor Module (MSC-01) and the Data Management Module (MGD). The MSC-01 performs local readings of the MQ-2 and LM35 sensors using dynamic baseline algorithms, safety timing, and mechanical flow shut-off via a stepper motor. The MGD centralizes data through a 2.4" TFT display and audible alerts using customized voice prompts, benefiting the general population, the elderly, and individuals with cognitive impairments. Practical tests demonstrated effectiveness in gas detection and thermal monitoring.

Keywords*—Biomedical engineering; accessibility; gas leak; home protection.

I. Introduction

The population aging observed in recent decades signals a profound shift in the Brazilian age structure, with a significant increase in the elderly population relative to younger generations [1]. This scenario poses urgent challenges to biomedical engineering and healthcare systems, requiring the development of solutions that ensure the autonomy, safety, and physical and mental well-being of this group. Maintaining functional independence is recognized by medicine as a determining factor for quality of life [2], directly contributing to the self-esteem and overall health of individuals with cognitive sensitivity.

In this situation, the developed equipment acts as an essential layer of preventive safety, meeting the protection needs of all users who frequently use cooking gas.

The natural decline of functions such as memory and attention, or the presence of early-stage pathologies such as Alzheimer's and Attention Deficit Hyperactivity Disorder (ADHD), significantly increases the risk of domestic accidents. Critical situations, such as leaving flames lit or liquefied petroleum gas (LPG) leaks, pose severe threats of fire and asphyxiation. This risk is frequently exacerbated by the sensory decline inherent to senescence—especially olfactory loss (presbyosmia) and hearing loss (presbycusis)—as the chronic decrease in these acuities directly compromises an elderly individual's ability to detect the smell of gas or emergency alarm sounds early on [3].

Given this problem, this work presents a safety system for the elderly using assistive technology and IoT, designed to act as preventive support in the domestic environment. The project integrates concepts of embedded systems and smart monitoring to identify risk situations in real time, establishing a link to accident prevention by offering an automated layer of protection and voice alerts that assist patients with mild cognitive deficits and everyday users seeking security for their homes, reducing reliance on full-time caregivers and maintaining their daily routines.

The SCS-01 (Safe Connection System) project, composed of two modules, the MGD (Data Management Module) and the MSC-01 (Kitchen Sensor Module), presented in Figure 1, was initially conceived for the National Biomedical Engineering Competition, Biochallenge Brasil 2025. The development was conducted by the team from LIHOM (Human-Machine Interface Laboratory) at the Federal University of Pernambuco, the institution responsible for providing the technological support and research space necessary for the device's design. Following the conclusion of the competition and its recognition with second place overall, the system demonstrated a high level of performance and effectiveness in promoting home safety and daily assistance, which motivated the continuation of academic research aimed at its constant improvement.



Fig. 1. Developed prototypes, on the left is the MGD and on the right is the MSC-01. Source: The authors, 2025.

II. Arquitetura do Sistema

The system was designed as a low-cost assistive and safety technology for both the elderly and the general public, structured into two main modules: the Kitchen Sensor Module (MSC-01) and the Data Management Module (MGD). Communication between the modules occurs in real time via a dedicated Wi-Fi network (2.4 GHz, 802.11 b/g/n), in which the MGD operates as the central server and the MSC-01 acts as the peripheral monitoring and actuation node. Each module is managed by an ESP32-C3 Super Mini microcontroller, selected for its compact dimensions, processing efficiency, and connection stability. Figure 2 shows the internal connections within the MSC-01 and MGD.

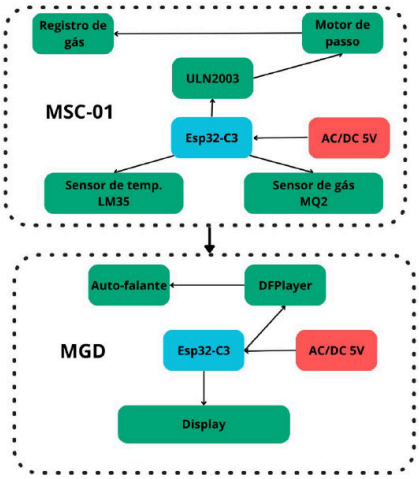


Fig.2. Block diagram of the SCS-01 interconnectivity. Source: The authors, 2025.

The Kitchen Sensor Module (MSC-01) is responsible for local environmental monitoring and active safety at the point of gas consumption. The choice of the MQ-2[4] sensor is justified by its high sensitivity to LPG and propane (in the range of 200 to 5,000 ppm), fast response time, and primarily its dual output: analog (AO), for continuous reading of gas levels, and digital (DO), which enables triggering a 0/1 signal with a threshold manually calibrated by a potentiometer on the module itself. For thermal monitoring, the LM35[5] sensor was selected due to its extremely low cost, ultra-compact dimensions, and native linear accuracy of 10 mV/°C, making it the ideal solution to measure ambient temperature variation without requiring complex external circuits. Sensory data processing occurs locally on the MSC-01, ensuring autonomy and speed in emergency decision-making. On the mechanical side, the module features a 28BYJ-48 stepper motor coupled to a ULN2003 driver, responsible for the direct angular actuation of a 5/16" copper mechanical valve to shut off the gas flow.

The Data Management Module (MGD), shown in Figure 3, acts as the core of information unification and system management. In addition to maintaining the local Wi-Fi network architecture, the MGD processes data received from the MSC-01 in real time to notify and trigger the corresponding actions. To ensure maximum accessibility for the general public and the elderly, the module manages a dual-alert multisensory interface: it visually displays system status and diagnostics on a 2.4-inch TFT display and plays customized voice audio messages through a DFPlayer Mini module connected to a 3 W speaker. This combination ensures that critical warnings remain intuitive, highly audible, and visually distinct.



Fig. 3. Prototype developed, MGD operational. Source: The authors, 2025.

III. Methodology and Sensor Calibration

The project development methodology was executed in a logical sequence, divided into hardware studies, software studies, and consolidation phases. The initial stage encompassed the analysis of the electrical characteristics of the battery, voltage regulator, and charger module. The circuits were assembled on a breadboard to validate component interconnection and perform practical torque tests on the motor-valve assembly.

A. Sensor Calibration

During the bench calibration phase, simulated scenarios were structured to establish reading routines. For the MQ-2 sensor, Qualitative Voltage Threshold Detection (Threshold-Based Detection) was adopted. For safety reasons in the test environment, lighters were initially used to simulate the leak, with subsequent readjustment for natural gas. Based on continuous readings from the ADC port in clean air ($ADC_{baseline}$) and under gas exposure (ADC_{atual}),the relative variation used by the algorithm for detection was defined:

$$\Delta ADC = ADC_{atual} - ADC_{baseline}$$

Additionally, the potentiometer on the MQ-2 module itself was adjusted to calibrate the digital output (DO) threshold, ensuring hardware redundancy by triggering a direct logic-level signal upon detecting the presence of gas.

Similarly, the calibration of the LM35 temperature sensor was based on establishing a dynamic baseline ($T_{baseline}$) captured upon each power-up of the equipment, eliminating interference from local climate variations. The algorithm performs a comparative time-averaged analysis between readings from previous seconds and the current reading, determining the thermal elevation:

$$.\Delta T = T_{atual} - T_{baseline}$$

It was experimentally established that an elevation of $\Delta T > 5°C$ indicates that the burner is lit. Confirmation of this variation initiates a 20-minute safety timer—an interval defined based on the overall average meal preparation time from Kantar Worldpanel's Consumer Watch study [6] to prevent domestic fires caused by forgetfulness. If the value of ΔT remains elevated at the end of the interval, the MGD emits a multisensory alert (audio and display) asking whether the user is still cooking. The absence of manual confirmation on the interface results in sending an emergency command via the dedicated network to the MSC-01, instructing the stepper motor to execute the immediate physical closure of the gas valve; if manual confirmation is positive, the 20-minute timer restarts.

B. Practical Testing and Validation Environment

The initial bench calibration phase was conducted at the LIHOM laboratory facilities, using a room equipped with air conditioning and exhaust ventilation. This infrastructure was essential to ensure the safety of the researchers during the handling of lighter gas, preventing risks of asphyxiation or the accumulation of flammable vapors in the room.

In the subsequent practical validation stage, the equipment was subjected to testing across multiple scenarios, including living rooms, residential kitchens, and outdoor areas. Exposure to different microclimates enabled the verification of the dynamic baseline algorithm's robustness for both sensors ($ADC_{baseline}$ e $T_{baseline}$), confirming the system's usability under background thermal fluctuations, variable humidity, and air currents characteristic of each environment.

Furthermore, the prototype returned to the controlled environment of LIHOM for the final calibration of the MQ-2 sensor. In this stage, the triggering threshold was readjusted using natural gas—the cooking fuel source effectively adopted in the project—moving beyond the preliminary testing phase with lighter gas. At this stage of development, measurements of the mechanical valve closure time (~12 s), Wi-Fi network latency, and detailed electrical consumption were not quantified by dedicated benchtop instrumentation, remaining characterized as technical metrics to be evaluated in the continuation of this work.

III. Results and Discussion

A. Sensor Validation and Performance

The development of the prototype resulted in a functional Internet of Things (IoT) system capable of continuous monitoring, issuing alerts, and automated physical intervention. To validate the effectiveness of the equipment, practical tests were conducted in controlled environments using a portable stove.

In the evaluation of the leak detection subsystem, 50 gas release simulations were conducted. In 48 of the 50 occurrences, the MQ-2 sensor demonstrated adequate sensitivity to LPG, triggering the audible alarm and activating the mechanical valve closure via stepper motor. Although the exact cause of the two residual failures has not yet been conclusively determined, hypotheses are raised for future investigation: it is hypothesized that localized, stronger air currents in the environment caused the ultra-fast dissipation of the gas flow, temporarily preventing the threshold reading ΔADC. However, it was observed that in scenarios of continuous exposure or prolonged leaks, the progressive accumulation of gas in the environment elevates the sensor reading to the triggering threshold in all situations, ensuring the system's response.

Additionally, 10 trials were conducted focusing on the identification of flames left lit due to forgetfulness. In all 10 repetitions, the LM35 sensor registered the thermal elevation characterized by the ΔT and the algorithm activated the safety protocol after 20 minutes of continuous exposure to heat. The data obtained proved that the calibration of the components and the dynamic baseline logic ensure the reliable triggering of the system's routines.

Furthermore, wireless communication between the MSC-01 and the MGD was successful across all validations, demonstrating a stable and highly secure connection for the integrated operation of the modules.

As a limitation of the current version, both devices rely on an external 12 V power supply. Thus, interruptions in the electrical power supply or installation failures temporarily disable system operation. However, it is worth highlighting the equipment's safety: if a power outage occurs while the gas valve is closed, the valve physically remains in the locked position, keeping the flow interrupted and ensuring the safety of the environment even during de-energization.

B. Assistive Interface, Connectivity, and Universal Application

The integration between the audio module (DFPlayer Mini) and the Data Management Module (MGD) established a centralized and highly intuitive communication interface, designed to serve any domestic user profile. Although the system offers assistive technology features, its application is not restricted to individuals with motor or cognitive limitations, serving as a universal residential safety solution. This demand is corroborated by research from Instituto Locomotiva (2025) [7], which indicates that 90% of Brazilians recognize a high risk of leaks or explosions when handling LPG. The ability to customize audio notifications, allowing the inclusion of recordings in familiar voices and intonation adjustment, adds value for both the general public and users with ADHD, the elderly, or individuals with memory deficits, promoting faster and more intuitive reactions in risk situations.

Simultaneously, the MGD displayed the system status clearly and objectively on the TFT display, acting redundantly with the audible alerts to reinforce the clarity of the information. Wi-Fi connectivity tests between the MSC-01, the MGD, and the central server proved the stability of real-time data exchange, ensuring reliability in command transmission and continuous monitoring.

It should be noted that, at this stage of the project, validations focused strictly on technical performance, communication stability, and the mechanical response of the actuators. Practical usability tests were not conducted with specific groups of end users (such as the elderly or individuals with cognitive deficits), so the evaluation of user experience and practical impact on these target audiences is outlined as a central proposal for future work. However, the implemented architecture demonstrates potential to ensure autonomous safety in the kitchen, mitigating the physical and psychological burden on caregivers and family members [8][9].

C. Comparative Analysis of Market, Instability, and Cost

When analyzing the gas leak safety solutions available on the market, it is noted that most commercial devices operate in a merely passive manner, limiting themselves to local audible and visual alarms. Although useful for individuals with immediate sensory and reaction capabilities, these systems fail by relying exclusively on human intervention to mitigate the hazard. In contrast, commercial systems with automatic shutoff (such as solenoid valves or block-gas devices) lack a remote and intuitive communication interface, operating without providing clear status feedback to the user and requiring invasive interventions in the residential gas and plumbing lines.

The Secure Connection System (SCS-01) stands out by integrating autonomous mechanical actuation with a dedicated Data Management Module (MGD). This centralized server combines active physical intervention (automatic shutoff) with clarity in human-machine communication through the combination of a TFT screen and voice synthesis, allowing visual and audible alerts to be positioned in any room covered by the home Wi-Fi network.

Another crucial differentiator compared to the market lies in the ease of installation. Unlike industrial solutions that require hiring specialized labor to adapt rigid gas piping, the valve integrated into the MSC-01 was designed to act directly on the flexible conduit of the residential installation (whether connected to the piped natural gas network or to an LPG cylinder). Installation occurs in a simple and non-invasive manner: the flexible gas hose is severed, and the valve coupled to the equipment's stepper motor is inserted between the two halves and secured with pressure clamps. This approach eliminates the need for masonry work or structural alterations to the property, lowering implementation costs. Figure 4 shows the connection with the equipment's valve.

From a financial perspective, the complete assembly of the two modules (MSC-01 and MGD) totaled approximately R$ 300.00 using components purchased at retail in small quantities. This value demonstrates high economic viability and market competitiveness, delivering a complete IoT solution with active shutoff, thermal/gas monitoring, and a multisensory assistive interface for a fraction of the cost of the few existing commercial alternatives. Table I displays the detailed price estimate of the devices.

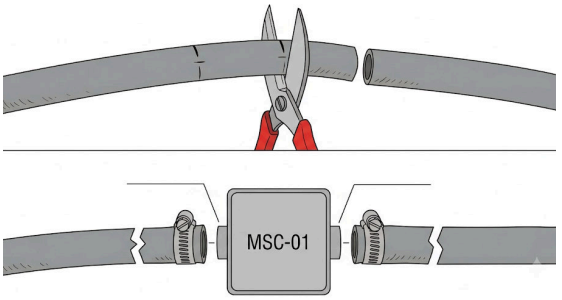


Fig. 4. Photo for illustrative puposes only, showing the installation of the gas system on the equipment. Source: Google Gemini, 2026.

TABLE I. TABLE OF APPROXIMATE VALUES

Table of approximate values		
Component / Item	***Module***	***Total value***
ESP32-C3 Super Mini Microcontroller (2x)	MSC-01 / MGD	R$75,00
MQ-2 Flammable Gas / LPG Sensor	MSC-01	R$18,00
28BYJ-48 Stepper Motor with ULN2003 Driver	MSC-01	R$26,00
5/16 Regulator Valve	MSC-01	R$16,00
2.4" TFT LCD Display	MGD	R$45,00
DFPlayer Mini MP3 Module	MGD	R$25,00
3W 8-Ohm Speaker	MGD	R$14,00
5V 2A Switching Power Supply (2x)	MSC-01 / MGD	R$31,00
Generic voltage regulator module(2x)	MSC-01 / MGD	R$22,00
Accessories (Connectors, Clamps, Wires, and Support PCB)	MSC-01 / MGD	R$17,00
3D printing filaments	MSC-01 / MGD	R$25,00
TOTAL PROTOTYPE COST (MSC-01 + MGD)		R$314,00

Source: Prepared by the author based on online price research (Brazil, 2026).

IV. CONCLUSIONS

The development of the Secure Connection System (SCS-01) demonstrated promising results in controlled simulations of a low-cost residential active safety solution. Overcoming the limitations of passive commercial systems, the proposed distributed architecture (MSC-01 and MGD) stood out for its capacity for autonomous mechanical intervention, neutralizing the risk of leaks and flames left lit directly at the source. Practical validation confirmed a high efficiency rate, consolidating the equipment as a functional safety tool.

The combination of visual signaling on a TFT display and customized voice alerts acts as a crucial assistive technology resource for the elderly and individuals with ADHD or memory deficits. However, due to its easy, non-invasive installation, the system establishes itself as a universally applicable solution, capable of protecting any user profile against domestic accidents, preserving autonomy in the home, and reducing the burden on

caregivers and family members.

As stages of development and enhancement for future work, the following guidelines are established:

-Environmental Investigation and Fine Calibration: Deepen the study on the influence of airflow and wind currents on gas dissipation, adjusting the sensitivity of the algorithm to resolve the two failures observed in the tests;

-Real-World Usability Validation: Conduct field trials and user experience tests directly with the target audience (the elderly, individuals with cognitive disorders, and their caregivers) to evaluate acceptance, ergonomics, and the effectiveness of voice messages;

-Power Redundancy (Battery Backup): Incorporate an uninterruptible power supply (UPS or rechargeable battery) subsystem to supply the 12 V circuit, ensuring that monitoring and actuation routines remain active even during local power outages.

References

[1] Instituto Brasileiro de Geografia e Estatística (IBGE), *Censo demográfico 2022: população por idade e sexo: resultados do universo*. Rio de Janeiro: IBGE, 2023. [Online]. Available: https://biblioteca.ibge.gov.br/visualizacao/periodicos/3102/cd_2022_populacao_idade_sexo_br.pdf

[2] Q. Zhu *et al.*, "Efeito mediador do apoio social entre a independência funcional e a qualidade de vida em idosos com deficiências moderadas a graves," *Scientific Reports*, vol. 16, Art. no. 10608, Mar. 2026. [Online]. Available: https://www.nature.com/articles/s41598-026-44694-z

[3] R. L. Doty and V. Kamath, "A influência da idade no olfato: uma revisão," *Psicologia Frontal*, vol. 5, Art. no. 20, Feb. 2014. doi: 10.3389/fpsyg.2014.00020.

[4] Zhengzhou Winsen Electronics Technology Co., Ltd., "Flammable Gas Sensor," MQ-2 datasheet, Jul. 2021. [Online]. Available: https://www.winsen-sensor.com/d/files/manual/mq-2.pdf?searchid=4369

[5] Texas Instruments, "LM35 Precision Centigrade Temperature Sensors," LM35 datasheet, Dec. 2017. [Online]. Available: https://www.ti.com/lit/ds/symlink/lm35.pdf

[6] MilkPoint, "Brasileiro dedica cada vez menos tempo ao preparo das refeições, aponta estudo," *MilkPoint*, 13 ago. 2018. [Online]. Disponível em: https://www.milkpoint.com.br/noticias-e-mercado/giro-noticias/brasileiro-dedica-cada-vez-menos-tempo-ao-preparo-das-refeicoes-aponta-estudo-209686/. [Acesso em: 22 ago. 2026].

[7] Instituto Locomotiva, "Percepção de riscos e segurança no manuseio de Gás Liquefeito de Petróleo (GLP): pesquisa de opinião," Sindigás, jun. 2025. [Online]. Available: https://www.sindigas.org.br/?p=42766

[8] R. S. Azevedo, "Sobrecarga do cuidador informal da pessoa idosa frágil: uma revisão sistemática," Dissertação de Mestrado, Escola de Enfermagem, Univ. Federal de Minas Gerais, Belo Horizonte, Brasil, 2010.

[9] L. M. Silva *et al.*, "Tecnologia assistiva e a redução do estresse em cuidadores de idosos: revisão integrativa," *Rev. Bras. Geriatr. Gerontol.*, vol. 24, no. 3, 2021.

[10] Google, "Gemini," Large languag eand generative model, 2026. [Online]. Available: https://gemini.google.com

Desenvolvimento e adaptação de cadeira de rodas de baixo custo para adolescente com displasia espondiloepifisária tipo Kondo-Fu: um relato de caso

1st Ana Carolina Ramos Nunes
Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. EITA)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás
anaramos2@discente.ufg.br

2nd Nayara Rodrigues Gomes de Oliveira
Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. EITA)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás
ft.nayrgomes@gmail.com

3rd Pedro Henrique Gonçalves
Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. EITA)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás
pedrogoncalves@ufg.br

***Resumo*—A displasia espondiloepifisária tipo Kondo-Fu (SEDKF) é uma doença genética extremamente rara, caracterizada por alterações osteoarticulares que comprometem o posicionamento postural e a mobilidade, demandando tecnologias assistivas adaptadas às necessidades individuais. Este estudo teve como objetivo descrever o desenvolvimento de uma cadeira de rodas de baixo custo, adaptada para uma adolescente com SEDKF. Trata-se de um relato de caso, no qual o dispositivo foi desenvolvido a partir de avaliação fisioterapêutica, modelagem tridimensional, produção de placas de plástico reciclado obtidas de tampas plásticas, impressão 3D de componentes complementares e montagem. As dimensões da cadeira foram definidas com base nas medidas antropométricas e nas características posturais da participante, resultando em modificações estruturais voltadas ao melhor posicionamento corporal. Após o teste inicial, foram realizados ajustes para facilitar a condução do dispositivo. A cadeira apresentou peso de 4,31 kg e custo estimado de produção de R$ 226,00. Os resultados demonstram a viabilidade do desenvolvimento de uma cadeira de rodas personalizada, de baixo custo e produzida com materiais reciclados, representando uma alternativa promissora para ampliar o acesso à tecnologia assistiva para usuários com condições raras e necessidades posturais complexas.**

***Palavras-chave*—Tecnologias Assistivas; Cadeira de rodas; Doenças raras; Sustentabilidade.**

I. Introdução

A displasia espondiloepifisária tipo Kondo-Fu (SEDKF) é uma doença genética autossômica recessiva de extrema raridade, caracterizada pela presença de variantes bialélicas no gene MBTPS1, responsável por codificar a protease do sítio 1 (S1P). A deficiência dessa enzima compromete duas vias fundamentais para a formação óssea e cartilaginosa: o transporte de colágeno, levando ao acúmulo da proteína no retículo endoplasmático, resultando em apoptose dos condrócitos; e o endereçamento de enzimas digestivas (lisossomais), levando à secreção dessas proteínas para a corrente sanguínea, essas que, por sua vez, degradam a matriz óssea e cartilaginosa recém-formada [1]. Com o avanço tecnológico, somente recentemente essa condição pôde ser identificada, contando, até o momento, com apenas oito casos relatados na literatura mundial [2, 3].

Dentre as manifestações clínicas da SEDKF, destacam-se a baixa estatura com início pós-natal, cifose e/ou escoliose, deformidades torácicas como pectus carinatum e pectus excavatum, e atraso no desenvolvimento motor geral [2]. Além disso, estudos de caso evidenciaram também, alterações pélvicas importantes, como luxação de quadril com rotação externa de membro inferior e acetábulo raso [4,5]. A junção dessas características clínicas configura um conjunto de alterações posturais complexas, representando um desafio significativo para o desenvolvimento de dispositivos de tecnologia assistiva adequados para essa população.

Diante da complexidade e da gravidade dessas alterações posturais, a Organização Mundial da Saúde (OMS) estabelece diretrizes em relação ao desenvolvimento de cadeiras de rodas, visando superar a entrega de equipamentos genéricos e adotar um modelo individualizado, focado nas especificidades clínicas e ambientais de cada usuário. Para garantir que o equipamento seja realmente adequado, a OMS recomenda que se sigam quatro etapas essenciais: a avaliação, para compreender as necessidades e objetivos do usuário; a adaptação física do dispositivo, para garantir suporte postural e mobilidade ideais; o treinamento do usuário e de sua rede de apoio; e o acompanhamento a longo prazo [6]. Nesse contexto, em condições raras como a SEDKF, esse processo individualizado torna-se indispensável.

No entanto, o acesso à tecnologia assistiva adequada ainda representa um desafio significativo para grande parte da população que dela necessita, reflexo de barreiras econômicas, da oferta limitada de serviços especializados e da insuficiência de políticas públicas voltadas à provisão individualizada de TA [7]. Nesse contexto, o desenvolvimento de soluções de baixo custo emerge como uma alternativa viável e necessária. A cadeira desenvolvida no presente trabalho utiliza placas produzidas a partir de tampas de garrafas plásticas recicladas como material estrutural. Essa abordagem converge com os Objetivos de Desenvolvimento Sustentável da ONU [8], particularmente com o ODS 3 (saúde e bem-estar), o ODS 10 (redução das desigualdades) e o ODS 12 (consumo e produção responsáveis), ao propor uma solução que é, ao mesmo tempo, clínica, social e ambiental.

Diante deste contexto, o presente artigo tem como objetivo descrever o processo de adaptação de uma cadeira de rodas de baixo custo, produzida com materiais reciclados, de acordo com as especificidades posturais de uma adolescente com displasia espondiloepifisária tipo Kondo-Fu, visando proporcionar mobilidade, qualidade de vida e participação social.

II. Metodologia

Trata-se de um estudo descritivo, do tipo relato de caso, conduzido no Laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. E.I.T.A.), da Universidade Federal de Goiás (UFG), no âmbito de um projeto de desenvolvimento de cadeiras de rodas produzidas a partir de tampas de garrafas plásticas recicladas, sob aprovação do Comitê de Ética em Pesquisa da UFG (N° 7.688.714), com obtenção do Termo de Consentimento Livre Esclarecido, assinado pela responsável, bem como o Termo de Consentimento de Uso de Imagem, e o Termo de Assentimento Livre e Esclarecido (lúdico) assinado pela participante. Nesse sentido, o presente trabalho descreve o processo de adaptação desse dispositivo para as especificidades clínicas de uma adolescente com displasia espondiloepifisária do tipo Kondo-Fu. O processo de desenvolvimento foi organizado em cinco etapas: avaliação fisioterapêutica; modelagem 3D; produção de placas a partir de material reciclado; impressão de componentes 3D; montagem do dispositivo.

Durante a avaliação fisioterapêutica, foram coletados dados referentes ao perfil clínico e funcional da participante, uma adolescente do sexo feminino, com 15 anos de idade, diagnosticada com displasia espondiloepifisária tipo Kondo-Fu. Observou-se hipotonia de tronco, de membros superiores e de membros inferiores, além de baixa estatura, escoliose acentuada e luxação de quadril, com manutenção dos membros inferiores em padrão de abdução e rotação externa. Em decúbito dorsal, a participante não apresentou alinhamento da cabeça em linha média, embora fosse capaz de levar as mãos em linha média e alcançar objetos, e a flexão de membros inferiores e membros superiores apresentou resistência superável. Na posição sentada, observou-se controle de tronco e de cabeça e capacidade de alcançar objetos. Não apresentou capacidade de se manter em ortostatismo de maneira independente e deambulação, sendo dependente para locomoção, realizada por meio de um carrinho de bebê adaptado. A comunicação verbal e cognição demonstraram-se preservadas. Complementarmente, foram coletadas as seguintes medidas antropométricas na posição sentada: distância entre ombros, largura do tronco, altura do tronco, largura do quadril, comprimento da coxa e comprimento da perna.

Com base nas medidas antropométricas coletadas, iniciou-se o processo de modelagem tridimensional do dispositivo no software Fusion 360. O modelo foi inspirado na GoBro (Fig. 1), uma cadeira desenvolvida para crianças com espinha bífida, comercialmente disponível nos Estados Unidos [9]. Entretanto, sua estrutura foi consideravelmente redesenhada (Fig. 2), uma vez que o uso da cadeira de referência requer a manutenção do joelho em flexão, posicionamento inviável para a participante que mantém os membros inferiores predominantemente estendidos, em razão da luxação de quadril. A transição entre o assento e o apoio para os membros inferiores, antes marcada por um desnível acentuado, foi suavizada, o apoio abdutor foi ampliado e adotou-se um design mais curvilíneo, com bordas arredondadas. Todas essas alterações foram realizadas com o intuito de acomodar a postura natural da participante, caracterizada por quadris em abdução, rotação externa e extensão de joelho.

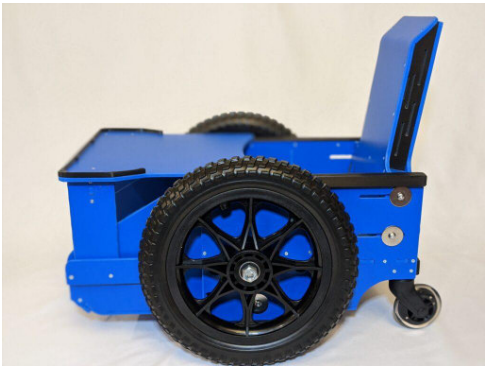


Fig. 1. Cadeira de rodas para Espinha Bífida

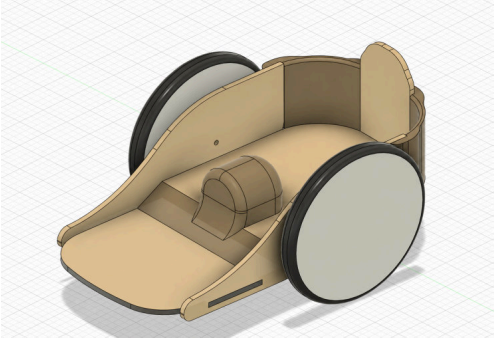


Fig. 2. Modelo 3D da cadeira de rodas desenvolvida

Após o processo de modelagem 3D, iniciou-se a produção dos componentes da cadeira de rodas. As principais estruturas foram produzidas a partir de placas de plástico confeccionadas com tampas de garrafas plásticas (Polietileno de Alta Densidade e Polipropileno). Inicialmente, as tampas passaram por um processo de separação por cores, seguida de dupla trituração em uma trituradora de 3kW em velocidade de 6.000 rpm. Após isso, o material triturado foi submetido a uma prensa térmica a 265 °C por 60 minutos, após o resfriamento durante 40 minutos, em temperatura ambiente, para a obtenção de placas plásticas que foram cortadas posteriormente por uma Fresadora CNC 3 eixos, resultando nas peças estruturais da cadeira. Complementarmente, o apoio abdutor e o encosto posterior foram produzidos por meio de impressão 3D com filamento PETG, na cor vermelha, utilizando a impressora Creality Ender-5 Max.

Com todos os componentes finalizados, foi realizada a montagem da cadeira de rodas. As peças estruturais de plástico reciclado e a impressão 3D foram unidas por meio de encaixe, parafusos e grampos quentes. Foram utilizadas rodas de EVA de bicicleta infantil, comercialmente disponíveis, e reaproveitadas de projetos anteriores do laboratório. Por fim, foram instalados rodízios giratórios tanto anteriormente quanto posteriormente, o que concluiu a montagem do dispositivo.

III. Resultados

O processo resultou em uma cadeira de rodas personalizada, de baixo custo, produzida com placas de plástico reciclado e impressão 3D (Fig. 3). As dimensões do dispositivo foram baseadas nas medidas antropométricas da participante, norteando o tamanho dos principais componentes, incluindo o encosto, o assento, o apoio para as pernas e o apoio abdutor. As medidas antropométricas utilizadas no desenvolvimento do dispositivo, obtidas durante a avaliação fisioterapêutica, são apresentadas na Tabela I.



Fig. 3. Cadeira de rodas adaptada montada.

Tabela I
Medidas Antropométricas

Medidas Antropométricas	Valores (Centímetros)
Largura dos ombros - Distâncias entre os pontos laterais mais externos dos ombros	31
Largura do tronco - Distância entre os pontos laterais do tronco ao nível da axila.	28
Largura do quadril - Distância entre os pontos laterais mais externos dos quadris.	22
Comprimento da perna - Distância da base do calçado até a fossa poplítea.	43
Comprimento da coxa - Distância da fossa poplítea até o ponto mais posterior do quadril	21,5
Altura do tronco - Distância da base do assento até o topo do ombro	26

No processo de produção das cadeiras de rodas convencionais no laboratório, utiliza-se aproximadamente 4.000 tampas plásticas. Já na produção da cadeira adaptada para SEDKF, aproximadamente 3.000 tampas plásticas foram utilizadas, cerca de 75% dessa quantidade padrão. Considerando essa proporção, estima-se que tenha levado 1h30min no processo de trituração e 2h15min na prensagem, resultando em aproximadamente 2 placas de 625x445x9 milímetros. Posteriormente, o processo de fresagem ocorreu em cerca de 25 minutos, e resultou em cinco peças estruturais.

A impressão das peças complementares em 3D, o apoio abdutor e o encosto posterior, durou 14 horas, e consumiu 780 gramas de material. Com os componentes finalizados, a montagem do dispositivo foi realizada e durou cerca de 40 minutos. Somando-se as etapas de produção das placas, impressão 3D e montagem, estima-se que o tempo total para a produção da cadeira de rodas foi cerca de 19 horas, resultando em um dispositivo com peso final de 4,31 quilogramas.

O custo dos componentes considerados no processo foi estimado em aproximadamente R$ 226,00. Foram contabilizadas as médias dos valores de mercado dos principais componentes utilizados em sua fabricação,

incluindo o filamento PETG (R$ 86,00), as rodas de EVA (R$ 120,00) e dois rodízios giratórios com trava (R$ 20,00). As placas produzidas a partir de tampas plásticas não foram contabilizadas, uma vez que a matéria-prima foi obtida via doação. Da mesma forma, os custos operacionais, como energia elétrica e equipamentos, não foram mensurados neste estudo.

Após ser finalizado, o dispositivo foi submetido a um teste, visando compreender se estava adequado para a participante e identificar possíveis ajustes (Fig. 4). Durante o teste, a equipe observou certa dificuldade na propulsão da cadeira de rodas, com cansaço e alteração no padrão respiratório. A partir dessa observação, foi instalado um rolamento (modelo ZZ809) no eixo das rodas, com o objetivo de facilitar a propulsão. Adicionalmente, foi implementado um acabamento por impressão 3D na extremidade pontiaguda do eixo da roda, visando maior segurança, e foi adicionada uma almofada ao assento para maior conforto. Após essas adaptações (Fig. 5), a cadeira de rodas foi entregue à participante. No entanto, não foi possível a realização de um novo teste formal após as modificações, de modo que os efeitos das adaptações sobre o esforço de propulsão e o conforto ainda não foram verificados, sendo prevista a avaliação desses pontos em etapas futuras do projeto.



Fig. 4. Teste da cadeira de rodas pela usuária



Fig. 5. Cadeira de rodas finalizada.

IV. Discussão

A displasia espondiloepifisária tipo Kondo-Fu é uma condição extremamente rara, com poucos casos descritos na literatura, o que representa uma lacuna de evidências sobre o desenvolvimento de tecnologias assistivas específicas para essa população. Dessa forma, o desenvolvimento de dispositivos individualizados torna-se uma estratégia importante para atender às necessidades funcionais de usuários com condições pouco frequentes e também para ampliar o conhecimento disponível sobre tecnologias assistivas aplicadas a doenças raras.

O design do dispositivo converge com princípios biomecânicos descritos na literatura, que descrevem a acomodação adequada dos membros inferiores como fundamental para o posicionamento da pelve, e essa, por sua vez, como base para o alinhamento do tronco na postura sentada [10]. Dessa forma, a ampliação do apoio abdutor e o redesenho da região entre o assento e o apoio para os membros inferiores buscaram acomodar a postura espontânea da participante, sem restringi-la a um posicionamento incompatível com suas características posturais. Além disso, o desenvolvimento foi conduzido de maneira individualizada, seguindo os princípios recomendados pelas Diretrizes para Provisão de Cadeiras de Rodas da OMS [6], que destaca a etapa de adaptação do equipamento de acordo com as necessidades do usuário. Nesse sentido, após o teste, foram adicionados rolamentos às rodas das cadeiras com o objetivo de facilitar a sua condução.

O custo estimado para a produção desse dispositivo foi de R$ 226,00. Em comparação, a cadeira utilizada como referência inicial, é comercializada por US$ 250 (R$ 1.272,50) [9], valor cerca de 82% superior ao custo estimado da solução desenvolvida. Além do baixo custo, o dispositivo incorpora princípios da economia circular ao utilizar placas produzidas a partir de material reciclado, as quais, ao final de sua vida útil, podem ser reprocessadas e reutilizadas no ciclo de produção, reduzindo a produção de resíduos e a demanda por matéria-prima.

A promoção da mobilidade, da participação social e da autonomia representa um dos principais objetivos da provisão de cadeiras de rodas [6], sendo influenciados por fatores relacionados ao ambiente, ao dispositivo e às características do usuário [11]. Nesse sentido, durante o teste, a responsável relatou informalmente a expectativa de que o dispositivo ampliasse a independência da participante para deslocamentos em sua residência, ilustrando o potencial impacto funcional esperado com o uso da cadeira. Trata-se, contudo, de uma observação preliminar, cuja validação depende de avaliação em futuras etapas do projeto.

Por se tratar de um relato do processo de desenvolvimento de um dispositivo individualizado para uma condição extremamente rara, com apenas oito casos descritos na

literatura mundial [2, 3], o trabalho teve como foco a descrição das etapas de desenvolvimento e adaptação do dispositivo, não estando entre seus objetivos a avaliação formal de aspectos como desempenho, conforto ou segurança. No entanto, uma limitação importante observada foi a ausência de um novo teste após a adição dos rolamentos e de outros ajustes complementares, impedindo verificar a efetividade dessas modificações. Como próximas etapas do projeto, será realizado o acompanhamento da utilização da cadeira, e após 3 meses uma reavaliação fisioterapêutica e aplicação do questionário *Quebec User Evaluation of Satisfaction with Assistive Technology - QUEST 2.0* (B-QUEST), para compreender o nível de satisfação da usuária e sua família em relação ao dispositivo.

V. Conclusão

O presente trabalho demonstrou a viabilidade de desenvolver uma cadeira de rodas adaptada às necessidades posturais de uma adolescente com displasia espondiloepifisária tipo Kondo-Fu, através de um processo individualizado de avaliação, modelagem, produção e ajustes finais. O dispositivo apresentou baixo custo de produção, peso reduzido e incorporação de princípios da economia circular, além de compatibilidade com as medidas antropométricas da participante e adaptações voltadas ao posicionamento corporal e facilitação da propulsão da cadeira. Contribuindo para ampliar o acesso a soluções de baixo custo para usuários com condições raras e características posturais complexas. Ressalta-se, contudo, que os resultados apresentados são preliminares, sendo necessário o acompanhamento a longo prazo da usuária para avaliar a efetividade e a satisfação em relação ao dispositivo, etapas já previstas nas próximas fases do projeto.

Referências Bibliográficas

[1]Kondo Y, Fu J, Wang H, Hoover C, McDaniel JM, Steet R, et al. Site-1 protease deficiency causes human skeletal dysplasia due to defective inter-organelle protein trafficking. JCI Insight 2018;3(14):e121596.

[2]Wang H, Wierenga A, Prabhu S, Wierenga K. MBTPS1-Related Spondyloepimetaphyseal Dysplasia with Elevated Lysosomal Enzymes. In: Adam MP, Bick S, Mirzaa GM, et al., editores. GeneReviews [monografia na internet]. Seattle (WA): University of Washington; 2023 [acesso em 27 jul 2026]. Disponível em: https://www.ncbi.nlm.nih.gov/books/

[3] Lucas-Castro E, Diaz-González F, Modamio-Høybjor S, Parrón-Pajares M, Pajares S, Gort L, et al. Novel MBTPS1 Variants and Cutis Laxa Phenotype in the 8th Reported Case of Spondyloepiphyseal Dysplasia, Kondo-Fu Type. Clin Genet 2026;109:747-753.

[4] Chen C, Wu J, Liu Y. Case report: recombinant human growth hormone therapy in a patient with spondyloepiphyseal dysplasia, Kondo-Fu type. Front Pediatr 2023;11:1068718.

[5] Alotaibi M, Aldossari A, Khan I, Alotaibi L. Identification of a new variant of the MBTPS1 gene of the Kondo-Fu type of spondyloepiphyseal dysplasia (SEDKF) in a Saudi patient. Case Rep Pediatr 2022;2022:5498109.

[6] World Health Organization. Wheelchair provision guidelines. Geneva: World Health Organization; 2023.

[7] World Health Organization; United Nations Children's Fund. Global report on assistive technology. Geneva: World Health Organization; 2022.

[8] United Nations. Transforming our world: the 2030 Agenda for Sustainable Development [homepage na internet]. New York: United Nations Department of Economic and Social Affairs; 2015 [acesso em 27 jul 2026]. Disponível em: https://sdgs.un.org/2030agenda

[9] The Frog Foundation. Who Could Benefit from Using the GoBro? [homepage na internet]. The Frog Foundation; 2026 [acesso em 27 jul 2026]. Disponível em: https://thefrogfoundation.org/the-gobro/

[10] Hasler C, Brunner R, Grundshtein A, Ovadia D. Spine deformities in patients with cerebral palsy; the role of the pelvis. J Child Orthop 2020;14:9-16.

[11] Smith EM, Sakakibara BM, Miller WC. A review of factors influencing participation in social and community activities for wheelchair users. Disabil Rehabil Assist Technol 2016;11(5):361-74.

Manufatura Aditiva aplicada à Tecnologia Assistiva: O desenvolvimento de um protótipo de encaixe protético transfemoral

Vitória Souza de Oliveira
Instituto de Patologia Tropical e Saúde Pública
Universidade Federal de Goiás
Goiânia, Brasil

Pedro Henrique Gonçalves
Faculdade de Artes Visuais
Universidade Federal de Goiás
Goiânia, Brasil

Maria Eduarda Nunes Badauy
Faculdade de Artes Visuais
Universidade Federal de Goiás
Goiânia, Brasil

***Abstract*—** A amputação consiste na remoção parcial ou total de um membro, gerando um impacto profundo nas atividades diárias do indivíduo. Quando ocorre no membro inferior, como no nível transfemoral, afeta diretamente a locomoção. O objetivo deste estudo foi desenvolver, por meio de manufatura aditiva (MA), um protótipo em escala reduzida de encaixe protético transfemoral, visando estruturar e documentar um fluxo de trabalho digital para viabilizar e fundamentar investigações científicas futuras na área. Trata-se de um estudo de desenvolvimento tecnológico no qual foram realizados o escaneamento de um molde positivo em gesso, a modelagem tridimensional e a impressão 3D do dispositivo. O fluxo digital e a MA demonstraram potencial promissor como estimativa preliminar de insumos de fabricação. O coto impresso ajustou-se geometricamente ao encaixe protético em escala reduzida (60%), sendo possível realizar uma verificação preliminar da correspondência dimensional e geométrica entre o modelo digital e o protótipo impresso, demonstrando adequada reprodução da geometria planejada. dispositivo. Conclui-se que a metodologia proposta apresenta um fluxo digital bem fundamentado e potencial viabilidade técnica e econômica para o desenvolvimento de tecnologias assistivas, embora ensaios mecânicos e testes clínicos futuros sejam indispensáveis para sua validação definitiva

Palavras-chave: manufatura aditiva; impressão 3D; amputação; tecnologia assistiva; prótese transfemoral.

I. INTRODUÇÃO

A amputação consiste na remoção parcial ou total de um membro [1]. Quando ocorre em membros inferiores gera um impacto profundo nas atividades diárias do indivíduo, afetando a locomoção e, consequentemente, sua qualidade de vida e participação [2]. No caso de uma amputação transfemoral caracterizada pela retirada do membro entre as articulações do joelho e do quadril, são acarretadas diversas alterações nos padrões posturais e de marcha [3] .

De acordo com o DATASUS nos primeiros meses de 2026 (jan-abr) já foram 10.699 amputações/desarticulações de membros inferiores, resultando em uma média de 2.674 ocorrências por mês [4]. A imobilidade decorrente da amputação, pode desencadear prováveis intercorrências, como, edemas, contraturas, dores, entre outras [5].

Nesse cenário, faz-se necessária a adoção de tecnologia assistiva (TA), definida como qualquer dispositivo ou equipamento que promova autonomia e qualidade de vida ao indivíduo [5]. A deambulação precoce, posteriormente à amputação, proporciona diversos benefícios, que impactam significativamente nos resultados funcionais, pois preserva os reflexos posturais e contribui para a saúde mental do indivíduo, favorecendo a adesão à prótese, a qual é considerada uma TA [6].

Em uma revisão sistemática, cujo objetivo foi avaliar a qualidade de vida de indivíduos submetidos à amputação de membro inferior decorrente de doença arterial obstrutiva periférica, constatou-se, em 12 estudos, que a deambulação com prótese foi o fator mais significativo e de maior impacto no bem-estar dos usuários [7].

A eficácia de uma prótese depende primordialmente da qualidade da interface entre o membro residual, comumente designado “coto” e o componente artificial [8]. O encaixe, por adaptar-se anatomicamente ao coto, é o fator preponderante que determina tanto o conforto do paciente quanto a sua capacidade de controle biomecânico sobre o dispositivo [8]. Sob esse viés, o encaixe deve proporcionar uma aderência segura ao coto, promovendo maior conforto e contribuindo para a biomecânica do movimento, viabilizando o controle da prótese pelo movimento do coto [8].

No âmbito biomédico, geralmente, demandam-se produtos personalizados e específicos para cada paciente.No que tange à produção do encaixe protético, novas técnicas têm surgido, como a manufatura aditiva (MA), também conhecida como impressão 3D que além de atender essa demanda, possibilita a redução de custos, impressão de estruturas complexas e maior liberdade de design [9].

O emprego de técnicas avançadas de MA na confecção de próteses viabiliza um desenvolvimento centrado no usuário. Essa abordagem atende às suas demandas de forma personalizada, o que favorece um melhor ajuste anatômico que propicia a reabilitação funcional da marcha [10].

Sob essa perspectiva, o objetivo deste estudo foi desenvolver, por meio de manufatura aditiva, um protótipo em

escala reduzida de encaixe protético transfemoral, visando estruturar e documentar um fluxo de trabalho digital para viabilizar e fundamentar investigações científicas futuras na área. Embora diversos estudos tenham demonstrado a aplicabilidade da manufatura aditiva na fabricação de componentes protéticos, ainda são escassos os trabalhos que descrevem de forma sistemática um fluxo digital completo para o desenvolvimento de encaixes transfemorais personalizados.

II. METODOLOGIA

Trata-se de um estudo de desenvolvimento tecnológico, conduzido segundo um fluxo digital integrado, composto pelas etapas de aquisição tridimensional da geometria do coto e modelagem, manufatura aditiva e verificação preliminar da correspondência geométrica entre o modelo virtual e o protótipo físico em escala reduzida. O estudo foi desenvolvido no laboratório de Estudos Inventivos em Tecnologia Assistiva (Lab. EITA) da Universidade Federal de Goiás (UFG). O processo de desenvolvimento iniciou-se com uma visita técnica à oficina ortopédica de um Centro Especializado em Reabilitação (CER), na qual realizou-se o escaneamento de um molde em gesso de um coto transfemoral. O modelo de gesso utilizado não estava com identificação e seria encaminhado para descarte. Para a digitalização utilizou-se o scanner portátil 3D, modelo EinScan PRO 2X Plus, apresentando precisão dimensional nominal de **±0,05 mm**. Os dados gerados da malha no aplicativo *Revopoint* foram importados para o *software* Autodesk Fusion 360, onde se realizou a modelagem tridimensional. No software CAD, foram aplicadas modificações geométricas diretas sobre a malha digitalizada, conformação da borda póstero-medial para apoio isquiático e definição da espessura nominal da parede em 2,0 mm.

Posteriormente a essa etapa, aplicou-se um fator de escala 0,60 (60% do tamanho real) aos arquivos digitais do encaixe protético e do coto para adequação ao volume de impressão e realização de ensaios preliminares de ajuste geométrico. Os arquivos finais foram exportados no formato STL para o software de fatiador *Creality Print* e processados para a impressora 3D de modelagem por deposição fundida (FDM) Creality Ender-5 Max. Para a impressão do encaixe e coto utilizou-se o filamento Poli Tereftalato de Etileno Glicol (PETG) natural e o ácido poliláctico (PLA) cinza, respectivamente. Ademais, utilizou-se o ferro de solda para abertura de um círculo com diâmetro reduzido, proporcional a escala 60% simulando o local correspondente às válvulas de vácuo convencionais.

Os parâmetros de impressão e especificações operacionais do processo estão detalhados na Tabela 1.

TABELA I
Parâmetros de impressão 3D e caracterização da manufatura dos protótipos

Parâmetros	Encaixe protético	Modelo de coto
Material	PETG (natural)	PLA (cinza)
Preenchimento	100%	15%
Espessura da Camada	0,2 mm	0,2 mm
Temperatura do Bico	240 °C	210 °C
Temp. Mesa Impressão	70 °C	60 °C
Diâmetro do Bico	0,8 mm	0,8 mm
Padrão de preenchimento	Linhas/perímetros concêntricos	Giroide
Massa do modelo	25.6g	72.3g
Tempo total impressão	4h10min	2h 40min

III. RESULTADOS

Durante a visita técnica, identificou-se que o desenvolvimento do encaixe protético transfemoral, seguindo o método tradicional, demanda diversas etapas. O processo compreende a confecção do molde negativo, a obtenção do molde positivo em gesso, o aprimoramento da superfície, a moldagem do encaixe de teste para a primeira prova, na qual são verificados os pontos de pressão, seguida pela laminação e pelo acabamento.

O escaneamento foi efetuado de forma eficiente por meio da conexão do scanner ao aplicativo no smartphone. O processo demandou aproximadamente 3 minutos, visto que foi realizado diretamente no molde estático, facilitando a captura da imagem e a obtenção de uma digitalização precisa (Fig. 1).



Fig. 1- Modelo tridimensional do coto transfemoral digitalizado

Na etapa de modelagem tridimensional, realizada no software Fusion 360, a malha gerada a partir da conversão do

molde importado serviu de base para o design do encaixe de contenção isquiática. A borda póstero-medial superior recebeu um leve achatamento para atuar como o apoio do ísquio, otimizando a distribuição da descarga de peso corporal. A lateral externa, por sua vez, foi alongada a fim de proporcionar maior estabilidade ao fêmur (Fig. 2).

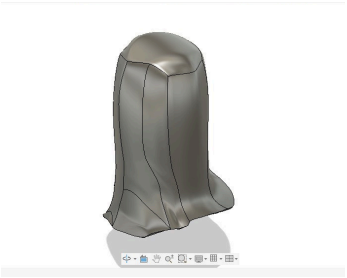


Fig. 2 - Modelagem tridimensional CAD do encaixe

A fabricação do modelo de coto por impressão tridimensional em PLA, um termoplástico biodegradável de alta rigidez, foi essencial para viabilizar os testes de acoplamento geométrico com o encaixe. Destaca-se que a confecção deste modelo se limitou a essa finalidade de validação do fluxo, contudo, caso o encaixe tivesse sido impresso em escala real, tais testes poderiam ter sido realizados diretamente no coto do próprio indivíduo (Fig. 3).



Fig. 3- Modelo do coto transfemoral impresso em 3D

Como resultado do processo de manufatura aditiva, o encaixe protético impresso com espessura de 2mm em PETG natural apresentou acabamento superficial adequado e manteve a transparência, que é uma característica do polímero (Fig. 4a e Fig. 4b). Contudo, a região de apoio isquiático recebeu um menor número de camadas de deposição devido à curvatura da borda. Em escala real, essa variação pode configurar um ponto de fragilidade estrutural, comprometendo a resistência exigida para essa área de alta descarga de peso.



Fig. 4 - Encaixe protético impresso: (a) vista anterior e (b) vista posterior.

No que tange ao acabamento pós-impressão, a confecção da abertura circular para a instalação da válvula de vácuo com ferro de solda mostrou-se um método eficiente e de fácil execução. Vale destacar que o calor aplicado promove a fusão do polímero e o surgimento de rebarbas, sendo necessário a utilização de um micro retífica para um acabamento superficial liso. Contudo, como o protótipo foi fabricado em escala reduzida de 60%, o diâmetro do furo foi reduzido proporcionalmente, o que impediu a realização de testes físicos de acoplamento com a válvula convencional de tamanho padrão (40 mm) (Fig. 5).



Fig. 5- Abertura para simulação de acoplamento da válvula pós-acabamento.

O coto ajustou-se geometricamente ao encaixe protético. A partir desses resultados, foi possível realizar a verificação dimensional e geométrica preliminar do dispositivo. Esse processo focou na avaliação volumétrica e das superfícies,

garantindo que as proporções geradas no software fossem reproduzidas fielmente no modelo físico antes da realização de testes de resistência estrutural. (Fig. 6a e Fig. 6b).



Fig.6- Acoplamento geométrico entre coto e encaixe: (a) vista anterior e (b) vista posterior

Por fim, o custo de fabricação do protótipo do encaixe em escala reduzida, considerando-se apenas os insumos de consumo direto, foi de aproximadamente R$9,00 referentes ao filamento. Ao projetar essa estimativa para o encaixe em tamanho real (escala 100%) o custo de insumos ficaria em torno de R$24,65. Ressalta-se que tais valores correspondem exclusivamente aos custos diretos de matéria-prima (filamento), excluindo outros custos como, consumo de energia elétrica, tempo de máquina, pós-processamento e mão de obra. Portanto, trata-se de estimativas parciais que não representam o custo final de produção comercial.

IV. DISCUSSÃO

No cenário da saúde pública brasileira, o Sistema Único de Saúde (SUS) desempenha um papel fundamental na reabilitação de amputados. Contudo, manter os custos desse atendimento é um desafio constante. Segundo o Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS (SIGTAP), o custo do serviço ambulatorial (SA) que engloba materiais e insumos para realizar a substituição de um encaixe protético transfemoral convencional é de R$ 1.098,00. A comparação simplificada de custos diretos não estabelece a viabilidade econômica, visto que se deve incorporar outros custos.

Ao comparar o método convencional de confecção com o fluxo digital por MA, as discrepâncias práticas são evidentes. O processo tradicional exige várias etapas manuais, como tirar o molde em gesso no paciente, fazer o modelo positivo e realizar o acabamento manual, o que gera desperdício de material e consome muito tempo. Por outro lado, a manufatura aditiva facilita esse processo. A partir do escaneamento em 3D, que não provoca incômodos ao paciente, pode-se modelar o molde diretamente por meio do software tridimensional. Assim, elimina-se a necessidade dos moldes físicos e diminui o desperdício de materiais, confirmando as afirmações sobre a redução de custos e a eficiência da manufatura aditiva [9].

Além de economizar materiais, esse novo método traz mais praticidade para o acompanhamento do paciente. Na produção convencional, os moldes de gesso são descartados após a fabricação da prótese. Com isso, se o encaixe vier a quebrar, por exemplo, todo o processo de moldagem precisará ser refeito. Já com o processo de digitalização, o arquivo digital pode ser armazenado, tornando possível novas alterações, que conforme a necessidade, poderão ser feitas sem um novo processo de escaneamento.

Salienta-se que o uso dessas novas tecnologias não substitui o trabalho dos técnicos ortopédicos e demais profissionais das oficinas de reabilitação. Esses profissionais possuem uma experiência prática e detém um conhecimento clínico biomecânico essencial para alinhamento, alívio de pressões e ajustes. Portanto, o escaneamento e a impressão 3D devem ser compreendidos como ferramentas de apoio técnico para agregar e facilitar o trabalho desses especialistas.

Contudo, este estudo apresenta limitações que devem ser consideradas. Por tratar-se de um desenvolvimento de protótipo, não foi possível realizar testes de resistência mecânica, ensaios de durabilidade do material a longo prazo ou testes clínicos com pacientes. Destaca-se que a correspondência geométrica observada em um modelo reduzido não permite prever os referidos pontos em escala real. As verificações desses pontos são fundamentais para garantir a segurança e a durabilidade do encaixe antes do uso prático.

V. CONCLUSÃO

O presente estudo estruturou e documentou um fluxo de trabalho digital integrado para o desenvolvimento de um protótipo de encaixe protético transfemoral em escala reduzida, demonstrando a viabilidade técnica da aquisição tridimensional, da modelagem CAD e da manufatura aditiva por FDM para reprodução da geometria planejada.

As análises de custo e tempo apresentadas devem ser interpretadas como estimativas preliminares de insumos e operação de bancada, visto que análises econômicas completas necessitam incorporar despesas de mão de obra, depreciação de equipamentos e pós-processamento.

A contribuição central deste trabalho consiste na disponibilização de um protocolo digital documentado que viabiliza e fundamenta investigações científicas posteriores. As conclusões deste estudo limitam-se estritamente à fabricação e verificação geométrica bem-sucedida em escala reduzida.

Portanto, embora os resultados preliminares sejam promissores, a validação definitiva desta tecnologia para o uso clínico, depende de investigações futuras dedicadas a testes mecânicos de resistência do material e durabilidade, além de avaliações de marcha com indivíduos amputados e ensaios clínicos para identificar os ajustes que precisam ser realizados e validar o produto desenvolvido.

AGRADECIMENTOS

Agradecemos o apoio institucional da Universidade Federal de Goiás (UFG), bem como o incentivo financeiro à pesquisa concedido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e o Ministério da Ciência, Tecnologia e Inovação (MCTI).

REFERÊNCIAS

[1] Choo YJ, Kim DH, Chang MC. Amputation stump management: a narrative review. World J Clin Cases. 2022 May 6;10(13):3981-8. doi: 10.12998/wjcc.v10.i13.3981. PMCID: PMC9131228.

[2] Hebert JS, Rehani M, Stiegelmar R. Osseointegration for lower-limb amputation: a systematic review of clinical outcomes. JBJS Rev. 2017 Oct;5(10):e10. doi: 10.2106/JBJS.RVW.17.00037

[3] Almeida SR, Giacomini G, Pedroni CR, Búfalo TA, Marques NR, Navega MT. Avaliação do quadril de amputados transfemoral durante contração isométrica em dinamômetro isocinético. Rev Bras Med Esporte. 2014 Jun;20(3):195-9. doi: 10.1590/1517-86922014200301476..

[4] Brasil. Ministério da Saúde. DATASUS. Informações de Saúde (TABNET): Morbidade Hospitalar do SUS (SIH/SUS) - por Local de Internação [Internet]. Brasília: Ministério da Saúde; 2026 [acesso em 14 jul 2026]. Disponível em: https://tabnet.datasus.gov.br/cgi/tabcgi.exe?sih/cnv/qiuf.def

[5] Brasil. Lei nº 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência) [Internet]. Brasília: Presidência da República, Subchefia para Assuntos Jurídicos; 2015 [acesso em 14 jul 2026]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm

[6] Sathishkumar S, Manigandan C, Asha T, Charles J, Poonoose PP. A cost-effective, adjustable, femoral socket, temporary prosthesis for immediate rehabilitation of above-knee amputation. Int J Rehabil Res. 2004 Mar;27(1):71-74. doi: 10.1097/00004356-200403000-00010.

[7] Davie-Smith F, Coulter A, Kennon B, Wyke S, Paul L. Factors influencing quality of life following lower limb amputation for peripheral arterial occlusive disease: a systematic review of the literature. Prosthet Orthot Int. 2017 Dec;41(6):537-47. doi: 10.1177/0309364617690394.

[8] Marks LJ, Michael JW. Artificial limbs. BMJ. 2001 Sep 29;323(7315):732-5. PMCID: PMC1121287

[9] Ngo TD, Kashani A, Imbalzano G, Nguyen KTQ, Hui D. Additive manufacturing (3D printing): a review of materials, methods, applications and challenges. Composites Part B. 2018 Jun 15;143:172-196. doi: 10.1016/j.compositesb.2018.02.012.

[10] Bhatt S, Joshi Deepak, Rakesh PK, Godiyal AK. Advances in additive manufacturing processes and their use for the fabrication of lower limb prosthetic devices. Expert Rev Med Devices. 2023 Jan;20(1):17-27. doi: 10.1080/17434440.2023.2169130.

[11] Cabrera IA, Pike TC, McKittrick JM, Meyers MA, Rao RR, Lin AY. Digital healthcare technologies: Modern tools to transform prosthetic care. Expert Rev Med Devices. 2021 Dec;18(Suppl 1):129-144

Integrated System for Physical Rehabilitation Using Virtual Reality and Telehealth: A Bibliometric and Narrative Review

Jéssika de Oliveira Ferreira
Programa de Pós-Graduação em Tecnologias Sustentáveis (PPGTECS)
Instituto Federal do Espírito Santo (Ifes)
Vitória, ES, Brasil
jessika.ferreira@estudante.ifes.edu.br

Flávio Garcia Pereira
Coordenadoria de Engenharia de Controle e Automação
Instituto Federal do Espírito Santo (Ifes)
Serra, ES, Brasil
flavio.garcia@ifes.edu.br

Douglas Almonfrey
Programa de Pós-Graduação em Tecnologias Sustentáveis (PPGTECS)
Instituto Federal do Espírito Santo (Ifes)
Vitória, ES, Brasil
dalmonfrey@ifes.edu.br

Mariana Rampinelli Fernandes
Programa de Pós-Graduação em Tecnologias Sustentáveis (PPGTECS)
Instituto Federal do Espírito Santo (Ifes)
Vitória, ES, Brasil
mariana.rampinelli@ifes.edu.br

***Resumo*—A deficiência motora afeta 14,4 milhões de brasileiros (7,3% da população), sendo que 5,2 milhões apresentam dificuldade permanente de mobilidade nos membros inferiores, segundo o Censo Demográfico 2022 do IBGE. As amputações de membro inferior representam uma condição de elevada carga clínica e social no Sistema Único de Saúde (SUS). Embora a captura de movimento sem marcadores, as arquiteturas de rede com computação de borda e os ambientes de realidade virtual (RV) para reabilitação motora tenham demonstrado viabilidade clínica individualmente, sua integração em um sistema de telessáude em tempo real permanece uma lacuna não resolvida na literatura. Este artigo apresenta uma revisão do estado da arte sobre a integração desses três campos tecnológicos aplicados ao diagnóstico remoto de marcha e à reabilitação. Conduzida nas bases Scopus e IEEE Xplore (2018–2025), a revisão resultou em um corpus de 442 documentos; 33 foram selecionados para leitura completa, dos quais 23 artigos temáticos compõem a síntese narrativa do estado da arte, complementados por 10 fontes epidemiológicas, regulatórias e fundamentais. A análise bibliométrica pelo algoritmo Louvain gerou duas comunidades no corpus analisado; a inspeção manual dos nós identificou três grupos temáticos, análise de marcha e biomecânica, computação de borda, e realidade virtual e telerreabilitação. As conexões de co-ocorrência entre os grupos de análise de marcha e de computação de borda são esparsas e de peso mínimo, o que sugere que a integração entre esses campos permanece pouco explorada. A consulta dirigida a sistemas integrados foi replicada em quatro bases, Scopus, IEEE Xplore, PubMed e Web of Science, não retornando nenhum sistema integrado em qualquer delas; a triagem em nível de conteúdo dos registros biomédicos também não identificou estudo combinando os três campos. Esses achados indicam uma direção de pesquisa promissora para o desenvolvimento de sistemas integrados de reabilitação remota no contexto do SUS.**

***Palavras-chave*—análise de marcha sem marcadores, computação de borda, realidade virtual, telerreabilitação, telessáude**

***Abstract*— Motor disability affects 14.4 million Brazilians (7.3% of the population), of whom 5.2 million have a permanent difficulty with lower limb mobility, according to the 2022 IBGE Demographic Census. Lower limb amputations represent a significant clinical and social burden within the Unified Health System (Sistema Único de Saúde, SUS). Although markerless motion capture, edge computing network architectures, and virtual reality (VR) environments for motor rehabilitation have each demonstrated clinical feasibility in isolation, their integration into a single real-time telehealth system remains an unresolved gap in the literature. This paper presents a state-of-the-art review on the integration of these three technological fields applied to remote gait diagnosis and rehabilitation. A narrative review was conducted using the Scopus and IEEE Xplore databases (2018–2025). After deduplication, the final corpus comprised 442 documents from 330 sources; 33 articles were selected for full reading, of which 23 thematic articles compose the narrative synthesis, complemented by 10 epidemiological, regulatory, and foundational sources. Bibliometric analysis using the Louvain algorithm produced two communities in the analysed corpus; manual node-level inspection identified three thematic groups, gait analysis and biomechanics, edge computing, and virtual reality and telerehabilitation. Co-occurrence links between the gait analysis and edge computing groups are sparse and of minimal weight, suggesting that their integration remains underexplored. The query targeting integrated systems was replicated across four databases, Scopus, IEEE Xplore, PubMed, and Web of Science,**

The authors acknowledge CNPq (National Council for Scientific and Technological Development) for funding projects No. 303584/2023-8 and "ReabiliTeM" No. 449916/2025-1 (Call Decit/SECTICS/CEIS/MS – CNPq PPSUS/Inovação 10/2025), and FAPES (Fundação de Amparo à Pesquisa e Inovação do Espírito Santo) for funding this work through projects No. 756/2024 (P: 2024-0TXNS), No. 380/2026 (P: 2026-ZKDB2) and No. 880/2026 (P: 2026-X46XH); and for the master's scholarship under Grant Term No. 225/2026 (Edital FAPES No. 19/2025, PROCAP 2026).

returning no integrated system in any of them; content-level screening of the biomedical records likewise identified no study combining the three fields. These findings point to a promising research direction for the development of integrated remote rehabilitation systems within the SUS context.

***Index Terms*—markerless gait analysis, edge computing, virtual reality, telerehabilitation, telehealth**

I. Introduction

Motor disability affects 14.4 million Brazilians (7.3% of the population), of whom 5.2 million have a permanent difficulty in lower limb mobility, according to the 2022 IBGE Demographic Census [1]. Motor rehabilitation of a person with lower limb amputation represents a clinical and logistical challenge of great relevance to the Unified Health System (*Sistema Único de Saúde*, SUS). Between 2018 and 2025, the public health system recorded 300,825 amputation procedures, with hospital costs exceeding 619 million Brazilian Reais [2]. The predominant causes (diabetes mellitus, peripheral vascular diseases, and trauma) are concentrated in vulnerable populations in regions far from rehabilitation reference centres [1], [2]. The phenomenon known as *ambulancioterapia* (long-distance patient transport to access specialised care) illustrates this problem: patients from inland municipalities must travel long distances to access specialised gait diagnosis, generating system costs and treatment adherence barriers [3], [4].

In this context, telehealth has emerged as a response to this challenge, enabling remote delivery of specialized healthcare services through information and communication technologies. In the field of gait analysis, three technological domains have shown significant and independent advances: markerless motion capture systems based on computer vision, edge computing architectures for real-time data processing and transmission, and virtual reality environments for remote motor assessment and rehabilitation. Each of these domains has demonstrated clinical and technical maturity in isolated studies. However, to the best of our knowledge, the integration of these three technologies into a unified real-time system compliant with the Brazilian General Data Protection Law (LGPD) and applicable to the Brazilian Unified Health System (SUS) has not been reported in the scientific literature, despite recent efforts to develop edge-based systems for healthcare [5]. The development of such a system is aligned with global priorities for universal health coverage, as rehabilitation is recognized as an essential component of integrated healthcare services, requiring coordinated efforts to scale up high-quality care and improve equitable access, in line with the United Nations Sustainable Development Goal 3 [6].

In this way, this paper proposes a bibliometric and narrative review focusing on the integration of these technological fields: markerless gait analysis, virtual reality for healthcare, and telehealth. Although the SUS already operates a national telehealth programme [7], the real-time transmission of kinematic data and LGPD compliance remain critical pre-conditions for reliable edge deployment in this context. Existing reviews address each of these fields in isolation, covering markerless motion analysis [8], [9], edge architectures for healthcare [10], and home-based rehabilitation technologies [11]. The objective of this work is to characterise, through bibliometric mapping and narrative synthesis, the degree of integration among these three fields in the literature, identifying whether systems combining them into a single telehealth pipeline have been reported.

II. Background

A. Lower Limb Amputation and Clinical Gait Analysis

Lower limb amputation is a highly prevalent condition in Brazil, concentrated in populations with reduced access to specialised care [1]. Between 2008 and 2015, the SUS recorded 361,585 amputation procedures [3], complementing the more recent hospital-cost figures reported in Section I [2].

Amputation causes significant biomechanical alterations in gait, including step length asymmetry, excessive heel elevation, foot circumduction, and compensatory postural deviations [12]. Quantitative gait assessment enables measurement of kinematic and spatiotemporal parameters such as velocity, cadence, step length and width, and joint angles, providing objective support for clinical diagnosis [8]. Reference commercial systems, based on reflective markers and infrared cameras, offer high accuracy but are costly, operationally complex, and incompatible with primary care and telehealth environments [8].

B. Markerless Motion Capture and Human Pose Estimation

OpenPose, introduced by Cao et al. [13], was the first open-source system capable of detecting multiple joint keypoints in real time from RGB images, using deep convolutional neural networks and part affinity fields. The Three-dimensional reconstruction of the virtual skeleton from multiple calibrated cameras through multiview geometry allowed the estimation of kinematic parameters without patient contact [14]. More recent approaches expand on this foundation: MediaPipe Pose employs a lightweight neural architecture achieving real-time inference at over 30 fps on mobile hardware, while RTMPose proposes optimised architectures for improved accuracy under partial occlusion [15]. The BODY_25 model of OpenPose defines 25 joint keypoints in three-dimensional coordinates, a configuration adopted in intelligent-space markerless capture setups reported in the literature [16].

C. Telerehabilitation and the SUS Context

Telerehabilitation is defined as the remote delivery of rehabilitation services through information and communication technologies. In the Brazilian context, telehealth emerges as a response to the phenomenon of *ambulancioterapia* [7]. One example is the Espírito Santo Physical Rehabilitation Centre (*Centro de Reabilitação Física do Espírito Santo*, CREFES), located in the metropolitan area, which receives patients from municipalities outside the metropolitan area. This geographic concentration creates transportation and access barriers for patients living outside the metropolitan area, issues that could often be avoided through remote care via a telehealth system.

However, even with a national telehealth programme in operation [7], the real-time transmission of kinematic data, LGPD compliance, and integration with existing clinical platforms remain open requirements for edge deployment [10].

III. Methodology

This work adopts a narrative state-of-the-art review approach, guided by the following research question: *which real-time data transmission architectures enable the integration of markerless gait analysis systems with virtual reality diagnostic environments in the context of telehealth?*

The search was conducted between April and May 2026 in the IEEE Xplore and Scopus databases, covering publications from 2018 to 2025, restricted to articles, conference papers, and reviews in Computer Science and Biomedical Engineering. Although other important databases exist, these two were chosen due to the technological nature of the research. Search fields were title, abstract, and author keywords; no language restriction was applied. Seminal foundational works published before 2018 were additionally included when they provided essential technical or conceptual grounding for one of the thematic blocks; this inclusion is stated explicitly to distinguish the systematic corpus retrieval from the narrative synthesis, and applies to a small number of highly cited references such as [17] and [13]. Search strings were organised into four thematic blocks: Block A (markerless gait analysis), Block B (network architectures for real-time biomedical data transmission), Block C (virtual reality applied to motor rehabilitation), and Block D (integrated telehealth systems, i.e., systems combining markerless capture, edge transmission, and VR into a single real-time pipeline). The Block D string, applied to Scopus, was: `TITLE-ABS-KEY ( ( "markerless motion capture" OR "gait analysis" ) AND ( "virtual reality" ) AND ( "telehealth" OR "telemedicine" ) AND ( "real-time" OR "network architecture" OR "data transmission" ) )`. Its fourth clause was intentionally broadened to include the general terms *network architecture* and *data transmission*, so as not to exclude integrated systems described under different technical vocabulary.

Block D yielded no results in the Scopus and IEEE Xplore databases during corpus construction. To check for the potential omission of important articles from biomedical databases not considered in this review, the four thematic queries were also replicated in PubMed and the Web of Science Core Collection in September 2026, using syntax adapted from the original IEEE Xplore strings without adding or removing terms. For the Block D combination, PubMed returned no results, and Web of Science returned a single record. An individual analysis confirmed that the latter was a contact-based wearable posture-capture sensor, related in scope but not a system integrating all three areas. Therefore, the Block D combination did not yield any integrated systems across the four independent databases. As an additional independent check, a content screening of the records retrieved from PubMed and Web of Science was conducted, searching titles and abstracts for the simultaneous presence of terms indicating markerless capture, virtual reality, and edge or real-time transmission; no record met all three conditions in either database. The PubMed and Web of Science searches were used as an external sensitivity check and were not incorporated into the 442-document bibliometric corpus, which was constructed from the four thematic search blocks applied to Scopus and IEEE Xplore.

Exported records in BibTeX format were imported into the Bibliometrix package (v. 4.x) via Biblioshiny (RStudio). Duplicate records across the two databases were removed by matching on DOI and, where the DOI field was absent, on normalised title. The final corpus comprised 442 documents from 330 distinct sources, with 2,043 unique authors and an annual growth rate of 11.94%, computed as the compound annual growth rate of scientific production over the analysed period. The average document age was 4.1 years, indicating a corpus concentrated in the last four to five years.

IV. Results and Literature Synthesis

The 442 documents were used for quantitative bibliometric mapping but were not all read in full. Article selection for the narrative synthesis followed a two-stage process: (1) the Biblioshiny 'Most Cited Documents' function was applied independently to each of the four thematic blocks; (2) articles published after 2022 were additionally screened to ensure coverage of recent advances. Selection through a most-cited criterion is known to favour older and more established works, which may under-represent recent integrated systems; the second-stage screening was introduced specifically to mitigate this bias. This process yielded 33 articles selected for full reading. Subsequently, articles meeting either of two predefined exclusion criteria were excluded: (i) insufficient technical detail on real-time transmission protocols, or (ii) absence of alignment with the scope of gait analysis applied to telehealth. This retained 23 thematic articles for the narrative synthesis. Screening was conducted by a single reviewer, a limitation discussed in Section V. A small number of foundational references directly relevant to this theme's technical basis (e.g., [13], [16]) were included by the authors' domain expertise, independently of the corpus retrieval process. Of the 23 thematic articles retained for the narrative synthesis, 15 representative studies are summarised in Table I; the remaining thematic articles are discussed in the text but are not included in the table because of space limitations. Fig. 1 summarises the full selection process.

In this study, a keyword co-occurrence analysis was also conducted. To this end, the Louvain algorithm was employed with association strength normalisation, applied to the All Keywords field ($n = 50$, minimum frequency $= 3$). It should be noted that the corpus was assembled from four independent thematic search blocks; co-occurrence patterns therefore reflect both genuine cross-field integration and the composition inherent in block-structured searches. The few connections between the gait analysis and edge computing groups were

Table I
COMPARATIVE SUMMARY OF REVIEWED ARTICLES

Ref.	Technology / Focus	Key Contribution	Gap / Limitation
Block A: Markerless Gait Analysis			
[8]	Review of vision-based motion analysis	Systematic review of markerless systems for sport biomechanics (515 citations)	No telehealth or real-time integration
[18]	OpenPose + mobile camera for gait	Knee angle error 13–18%; outperformed Kinect under clothing occlusion	No 3-D reconstruction; no amputee validation
[19]	OpenPose vs. MediaPipe vs. MMPose	MediaPipe: highest correlation (0.944), inference speed (30.19 fps)	No 3-D; no clinical population; no telehealth
[9]	Review of markerless gait biomechanics	Promising for sagittal plane; limitations under occlusion/prostheses	No transmission or VR integration
[20]	Markerless systems for clinical gait	Suitable for step length, velocity, hip/knee angles when validated	No telehealth context; no amputees
[21]	MAEVE: embedded 3-D pose (Jetson Xavier AGX)	Knee deviation $<5°$; correlation 0.91–0.99 vs. Vicon; edge execution	No VR diagnostic environment; no amputee validation
Block B: Network Architectures & Edge Computing			
[22]	Fog-IoT architecture for patient monitoring	Three-layer fog architecture reduces latency vs. cloud-only (448 citations)	Scalar physiological signals only; no kinematics
[23]	Edge computing for healthcare	Systematised edge/fog concepts and healthcare applications	No gait data or VR integration
[24]	Survey: fog/edge telemedicine monitoring	MQTT and WebSocket identified as dominant real-time protocols	No kinematic data transmission; no VR
[10]	Edge computing in IoMT	Data protection compliance as precondition for edge deployment	No gait or VR component
Block C: Virtual Reality & Motor Rehabilitation			
[17]	VR for motor rehabilitation (review)	4 consolidated findings on motor learning in VR (1,003 citations)	Pre-edge era; no real-time kinematic transmission
[25]	VR telerehabilitation for balance/gait (RCT)	Clinical equivalence between VR-remote and in-person interventions, post-stroke participants	No markerless capture; no amputee population
[26]	VR telerehabilitation for Parkinson's (RCT)	Significant improvement in balance/QoL ($p < 0.05$)	No lower limb amputee population
[27]	Telerehabilitation after TKA (RCT)	Non-inferiority of telerehabilitation vs. in-person care	No VR; no gait kinematics; no amputees
Block D: Integrated Systems (closest approximation)			
[16]	VR diagnostic room + intelligent space	Novel system integrating VR + markerless gait; validated by 12 physiotherapists	Asynchronous (static JSON); no real-time transmission; TRL 4

***Note:** This table consolidates 15 representative studies, selected as the most relevant works within each thematic block (Blocks A–D) and complemented by recency screening, as described in Section III. The narrative synthesis integrates 23 thematic articles cited throughout the text. References [1]–[7], [12], and [14] are epidemiological, regulatory, and foundational sources supporting the problem statement, the regulatory framework, and the technical basis of the system. Reference [4] provides the empirical basis for the concept of *ambulancioterapia*. Technology Readiness Level attributions are estimates by the authors, derived from the descriptions available in each publication, and were not validated against an independent rubric.

consistent with the search results obtained using the Block D string, which encompassed the four thematic groups. In the analysed corpus, the Louvain algorithm identified two network communities. A subsequent descriptive inspection of the nodes grouped the keywords into three interpretive thematic domains: (i) gait analysis and biomechanics; (ii) network architectures and edge computing; and (iii) virtual reality and telerehabilitation. Nodes conceptually affiliated with network architecture literature (*edge computing*, *internet of things*, *telemedicine*, *medical computing*) were assigned to a distinct thematic group for visualization purposes, consistent with established bibliometric practice when algorithmic community detection yields fewer clusters than the number of thematic domains identified in the literature. Demographic indexing terms describing study populations were distributed within these groups rather than forming a separate group. Co-occurrence links between the gait analysis group and the edge computing group are limited to two low-weight connections (weight $= 1$ each), namely between *gait analysis* and *edge computing* and between *gait analysis* and *internet of things*. Fig. 2 illustrates the resulting network.

A. Markerless Gait Analysis

Conventional clinical gait analysis relies on infrared cameras and reflective markers attached to the patient's body. Although these systems offer high accuracy, they are costly, require specialised calibration, and are incompatible with telehealth environments [8]. This limitation motivated the development of markerless systems based on human pose estimation through computer vision, which is particularly relevant to the present scenario, in which gait is captured at

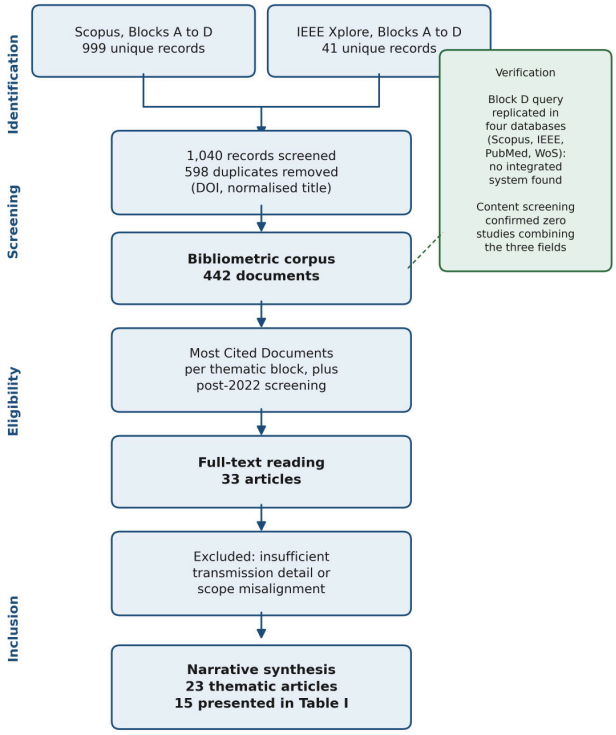


Figure 1. Flow of records from database retrieval to the final narrative synthesis. The Block D query, targeting integrated systems, yielded no eligible integrated system across the four databases consulted; Web of Science returned one adjacent record, which was excluded after content inspection.

a distance by a multi-camera intelligent space rather than by body-worn markers.

Viswakumar et al. [18] demonstrated the feasibility of OpenPose for gait analysis using a mobile phone camera, calculating knee flexion/extension angles with a mean error of 13–18%, outperforming Microsoft Kinect in clothing occlusion scenarios. Tony Hii et al. [19] conducted a direct comparison of OpenPose, MediaPipe Pose, and MMPose for lower limb kinematics during gait, with MediaPipe Pose achieving the highest overall correlation (0.944) and inference speed (30.19 fps).

Wade et al. [9] concluded that RGB camera-based systems show promising results for spatiotemporal and sagittal plane kinematic parameters, but accuracy remains inferior to marker-based systems under severe occlusion or in patients with prostheses. Needham et al. [20] confirmed that markerless systems are suitable for clinical gait analysis when validated against a reference system, particularly for step length, velocity, and hip and knee angle variables.

The most relevant advance for telehealth was proposed by Martini et al. [21], who developed the MAEVE system on a Jetson Xavier AGX embedded device. In healthy participants walking on a treadmill, recorded in the coronal plane, the system showed deviation below 5° in knee flexion angle and correlation of 0.91–0.99 relative to a marker-based reference system, with local, edge-based processing that can reduce the transmission of raw video data.

Recent work has begun to address the amputee population directly. Diffusion-based models have been applied to zero-shot pose estimation for lower-limb prosthetic users [28], and video-based motion analysis has been assessed in individuals with lower limb amputation [29]. The gap identified in this block is therefore not the absence of markerless methods applied to amputees, but the absence of such methods integrated into a validated real-time telehealth deployment.

B. Network Architectures for Real-Time Transmission

Cloud-only architectures introduce latencies of hundreds of milliseconds, often rendering them incompatible with real-time diagnosis [22]. Verma and Sood [22] proposed a three-layer fog architecture for remote patient monitoring, demonstrating significant latency reduction compared to cloud-only solutions. Abirami and Chitra [23] systematised edge computing concepts for healthcare applications. He et al. [24] identified MQTT and WebSocket as the dominant protocols in time-sensitive health applications, owing to their low overhead and real-time bidirectional transmission capability. Fog-assisted architectures have further demonstrated energy-efficient, low-latency local processing for continuous remote health monitoring, with on-gateway data handling reducing reliance on cloud round-trips [30]. Furthermore, Yaraziz et al. [10] highlighted data protection compliance as a foundational precondition for edge deployment in public health. The gap identified in this thematic block is the absence of architectures specifically validated for the real-time transmission of markerless joint kinematic data to immersive diagnostic environments.

C. Virtual Reality Applied to Motor Rehabilitation

Holden [17], in a review predating the edge computing era, established four consolidated findings: motor learning is possible in VR; movements learned in VR transfer to equivalent real-world motor tasks; VR outperforms conventional practice in comparative studies; and no cybersickness was reported in motor-impaired populations using the low-immersion systems available at the time, a finding that does not extend directly to current head-mounted displays. Lloréns et al. [25], in a randomised controlled trial with post-stroke participants, reported clinical equivalence between VR-based remote and in-person interventions for balance recovery. Gandolfi et al. [26] demonstrated significant improvement in Parkinson's disease patients through VR telerehabilitation. Prvu Bettger et al. [27] identified non-inferiority of telerehabilitation versus in-person care for functional outcomes after total knee arthroplasty. Recent developments indicate that spatiotemporal and muscle activation patterns adapt across walking modalities and VR pace constraints [31], while high-fidelity immersive designs influence patient engagement and motor learning [32]. Beyond amputee-specific systems, Toh et al. [11], in a scoping review of home-based wearable technologies for stroke rehabilitation, found that research has concentrated overwhelmingly on the

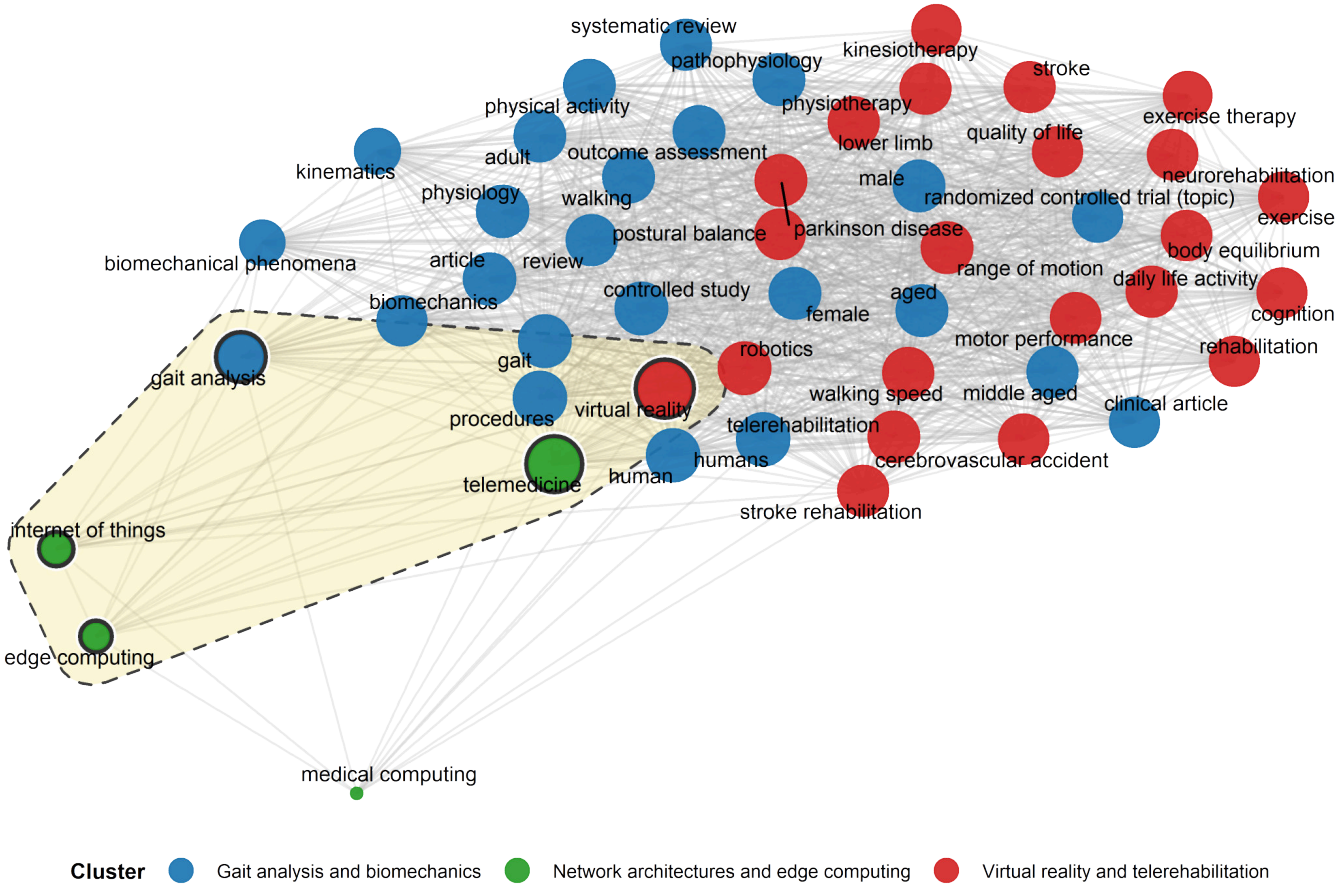


Figure 2. Keyword co-occurrence network of the bibliometric corpus ($n = 442$ documents, Scopus and IEEE Xplore, 2018–2025). Louvain algorithm, association strength normalisation, All Keywords field ($n = 50$ nodes, minimum frequency $= 3$). The Louvain algorithm produced two communities in the analysed corpus; manual node-level inspection identified three thematic groups, rendered here in distinct colours: gait analysis and biomechanics (blue; e.g., *gait analysis*, *biomechanics*, *kinematics*), network architectures and edge computing (green; e.g., *edge computing*, *internet of things*, *telemedicine*), and virtual reality and telerehabilitation (red; e.g., *virtual reality*, *telerehabilitation*, *stroke rehabilitation*). Demographic indexing terms (e.g., *human*, *male*, *female*, *adult*, *aged*) are distributed within the blue group. The dashed polygon highlights the region of the network where the terms most directly related to the integration problem investigated by this review are concentrated. Co-occurrence links between the blue and green groups within this corpus are sparse and of minimal weight.

upper limb, with only very limited evidence addressing the lower limb. This asymmetry reinforces the gap in remote lower limb rehabilitation that the present work targets. Unlike most of the reviewed telerehabilitation solutions, which aim to give patients autonomy to carry out their own rehabilitation at home, the system addressed in this work is intended for use by the health professional, supporting remote gait diagnosis rather than patient-administered therapy. Within this landscape, Silva et al. [16] advanced a VR-based diagnostic environment designed specifically for lower limb amputees, integrating an intelligent space for markerless gait capture, which constitutes the direct precursor to the present work.

D. Integrated Systems: Technological Gap

Silva et al. [16] propose a VR-based system for gait diagnosis in individuals with lower limb amputation, combining an intelligent space with a virtual environment. This framework achieved high performance in gait classification. However, the integration of all components into a real-time, low-latency, and LGPD-compliant architecture for remote clinical settings remains an unresolved challenge. A recent narrative review of gait analysis systems in clinical rehabilitation for individuals with lower limb amputation [33] likewise reports that integration into routine clinical workflows remains incipient. The bibliometric mapping of the surveyed corpus corroborates this direction: co-occurrence links between the gait analysis and edge computing groups are sparse and of minimal weight (Fig. 2), indicating that the integration of these two fields remains largely unexplored rather than entirely absent.

V. Discussion

The three investigated fields present sufficient maturity to individually support components of the proposed system. Markerless systems report correlations above 0.90 against marker-based references for sagittal plane parameters, across heterogeneous populations and protocols [9], [19]–[21]; edge

architectures with MQTT and WebSocket are consolidated for real-time transmission [22], [24]; and VR environments show clinical equivalence to in-person care in the populations studied [17], [25], [26].

From a technological maturity perspective, reviewed systems are predominantly between TRL 4 and TRL 5. Advancing to TRL 6 requires three conditions: demonstration in a clinically relevant environment with people with lower limb amputation, a standardised assessment protocol, and LGPD-compliant data transmission over a public network. The literature does not record systems that have reached this level for remote VR-based gait diagnosis in people with amputation.

Network instability in remote regions served by the SUS represents an additional constraint on any real-time architecture. Sacco et al. [34] report packet-loss recovery in latency-sensitive networked applications, a class of technique whose applicability to kinematic data streams was not evaluated in the surveyed literature and would require dedicated investigation.

Several limitations should be considered when interpreting these findings. This work is a bibliometric-informed narrative review and does not follow the full protocol of a systematic review: screening was conducted by a single reviewer, no formal quality or risk-of-bias assessment was applied, and no grey literature or clinical trial registry searches were performed. The quantitative corpus was built from Scopus and IEEE Xplore. To assess the effect of this restriction, the four thematic queries were replicated in PubMed and Web of Science in September 2026. Of the records carrying a registered DOI, 60.5% of the PubMed set and 44.4% of the Web of Science set were already present in the corpus, indicating substantial coverage of the thematic intersection by the two databases used for corpus construction. Content-level screening of the additional records identified no study combining markerless capture, virtual reality, and real-time transmission, and the Block D query returned no integrated system in any of the four databases. Attributions of Technology Readiness Level are estimates made by the authors from the reviewed publications, and were not validated against an independent rubric. A full systematic review, with dual independent screening and formal quality assessment, is planned as a subsequent phase of this research, with an expanded research team.

VI. Conclusion

This study aimed to conduct a bibliometric analysis and a narrative literature review to identify articles related to gait analysis and virtual reality within the context of telehealth-based rehabilitation. To this end, four distinct searches were performed: the first focused on markerless or wearable-sensor-free gait analysis; the second on communication architectures for synchronous biomedical data transmission; the third on virtual reality for rehabilitation; and the fourth integrated these three areas.

The review demonstrated that these fields have achieved independent, well-established development, with evidence of clinical feasibility for each component in isolation. Bibliometric analysis of 442 documents indicated that co-occurrence connections between the gait analysis and edge computing groups are scarce and carry minimal weight (Fig. 2); this finding, combined with the absence of integrated systems identified in Block D across four databases, suggests that integrating these fields into a synchronous telehealth system remains unexplored and constitutes an open research problem.

Future work could include conducting a full systematic review, involving independent double screening and formal quality assessment, with an expanded research team. Future work should investigate interoperability with health-information standards such as HL7 FHIR and, where medical-imaging integration is required, DICOM, as well as the scalability of the proposed solutions within the healthcare sector. Expanded research on cybersecurity should also be conducted to better understand these solutions for the healthcare field. The potential for designing a system that offers remote rehabilitation within the Brazilian telehealth context, enabling patients in remote areas to access specialised home-based care without the need for travel, must also be analysed.

Acknowledgments

The authors also used artificial intelligence tools (Claude, Anthropic) for text review, English translation, and structural organization. All citations, numerical claims, and reference entries were manually verified by the authors against the original sources. The scientific content, bibliometric analysis, article selection, interpretation of findings, and conclusions are the sole responsibility of the authors.

References

[1] IBGE, "Censo Demográfico 2022 Pessoas com deficiência e pessoas diagnosticadas com transtorno do espectro autista - Resultados preliminares da amostra ," Instituto Brasileiro de Geografia e Estatística, pp. 1–82, 2022. [Online]. Available: https://biblioteca.ibge.gov.br/visualizacao/livros/liv102178.pdf

[2] Ministério da Saúde, "Sistema de informações hospitalares do SUS (SIH/SUS): Amputações de membros, procedimentos registrados 2018–2025," Brasília, DF, 2026, acesso em: 25 jun. 2026. [Online]. Available: https://datasus.saude.gov.br/

[3] A. M. Peixoto, S. A. Zimpel, A. C. A. de Oliveira, R. L. S. Monteiro, and T. K. G. Carneiro, "Prevalência de amputações de membros superiores e inferiores no estado de Alagoas atendidos pelo SUS entre 2008 e 2015," *Fisioterapia e Pesquisa*, vol. 24, no. 4, pp. 378–384, 2017.

[4] F. C. d. Sousa and M. M. S. d. Mello, "Vazios assistenciais e capacidade instalada: desafios para a regionalização da saúde no sudeste do tocantins," *Revista JRG de Estudos Acadêmicos*, vol. 9, no. 20, p. e092993, mar. 2026. [Online]. Available: https://revistajrg.com/index.php/jrg/article/view/2993

[5] H. d. M. Corrêa and M. F. Franco, "Lei Geral de Proteção de Dados (LGPD) na Saúde: Tendências, Desafios e Oportunidades," in *Anais da ERRC 2025: Artigos Completos do WRSeg*, 2025, pp. 123–129.

[6] World Health Organization, "Rehabilitation 2030: A call for action: Meeting report, 6-7 february 2017," World Health Organization, Geneva, Switzerland, Tech. Rep., 2017. [Online]. Available: https://www.who.int/publications/m/item/rehabilitation-2030-a-call-for-action

[7] Ministério da Saúde, "Programa nacional de telessaúde brasil redes," Brasília, DF, 2019, acesso em: 25 jun. 2026. [Online]. Available: https://www.gov.br/saude/pt-br/composicao/seidigi/sus-digital/telessaude

[8] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports Medicine-open*, vol. 4, no. 1, p. 24, 2018.

[9] L. Wade, L. Needham, P. McGuigan, and J. Bilzon, "Applications and limitations of current markerless motion capture methods for clinical gait biomechanics," *PeerJ*, vol. 10, pp. 1–27, 2022.

[10] M. S. Yaraziz, N. S. Safa, and M. A. Azad, "Edge computing in IoT for smart healthcare," *Journal of Ambient Intelligence and Smart Environments*, vol. 16, no. 4, pp. 409–438, 2024.

[11] S. F. M. Toh, K. N. Fong, P. C. Gonzalez, and Y. M. Tang, "Application of Home-Based Wearable Technologies in Physical Rehabilitation for Stroke: A Scoping Review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1614–1623, 2023.

[12] J. A. L. Viegas, "Estudo biomecânico na marcha de indivíduos amputados de membro inferior," Dissertação (Mestrado em Engenharia Biomédica), Faculdade de Engenharia (Universidade do Porto), Porto, Portugal, Out. 2017.

[13] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2004.

[15] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose," *arXiv*, 2023. [Online]. Available: http://arxiv.org/abs/2303.07399

[16] P. P. Silva, W. F. da Rocha, L. E. Mazzoni, R. M. de Andrade, A. Bento, M. Rampinelli, and D. Almonfrey, "A diagnostic room for lower limb amputee based on virtual reality and an intelligent space," *Artificial Intelligence in Medicine*, vol. 143, no. May, p. 102612, 2023. [Online]. Available: https://doi.org/10.1016/j.artmed.2023.102612

[17] M. K. Holden, "Virtual environments for motor rehabilitation: Review," *Cyberpsychology and Behavior*, vol. 8, no. 3, pp. 187–211, 2005.

[18] A. Viswakumar, V. Rajagopalan, T. Ray, and C. Parimi, "Human Gait Analysis Using OpenPose," *Proceedings of the IEEE International Conference Image Information Processing*, vol. 2019-November, pp. 310–314, 2019.

[19] C. S. Tony Hii, K. B. Gan, N. Zainal, N. M. Ibrahim, S. Azmin, and N. A. Shattar, "Marker Free Gait Analysis using Pose Estimation Model," in *2022 IEEE 20th Student Conference on Research and Development, SCOReD 2022*. IEEE, 2022, pp. 109–113.

[20] L. Needham, M. Evans, D. P. Cosker, L. Wade, P. M. McGuigan, J. L. Bilzon, and S. L. Colyer, "The accuracy of several pose estimation methods for 3D joint centre localisation," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-00212-x

[21] E. Martini, M. Boldo, S. Aldegheri, N. Valè, M. Filippetti, N. Smania, M. Bertucco, A. Picelli, and N. Bombieri, "Enabling Gait Analysis in the Telemedicine Practice through Portable and Accurate 3D Human Pose Estimation," *Computer Methods and Programs in Biomedicine*, vol. 225, p. 107016, 2022. [Online]. Available: https://doi.org/10.1016/j.cmpb.2022.107016

[22] P. Verma and S. K. Sood, "Fog assisted-IoT enabled patient health monitoring in smart homes," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1789–1796, 2018.

[23] S. Abirami and P. Chitra, "Chapter fourteen - energy-efficient edge based real-time healthcare support system," in *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, ser. Advances in Computers, P. Raj and P. Evangeline, Eds. Elsevier, 2020, vol. 117, no. 1, pp. 339–368. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065245819300506

[24] Q. He, Z. Xi, Z. Feng, Y. Teng, L. Ma, Y. Cai, and K. Yu, "Telemedicine Monitoring System Based on Fog/Edge Computing: A Survey," *IEEE Transactions on Services Computing*, vol. 18, no. 1, pp. 479–498, jan 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10772317/

[25] R. Lloréns, E. Noé, C. Colomer, and M. Alcañiz, "Effectiveness, Usability, and Cost-Benefit of a Virtual Reality–Based Telerehabilitation Program for Balance Recovery After Stroke: A Randomized Controlled Trial," *Archives of Physical Medicine and Rehabilitation*, vol. 96, no. 3, pp. 418–425.e2, mar 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0003999314012209

[26] M. Gandolfi, C. Geroin, E. Dimitrova, P. Boldrini, A. Waldner, S. Bonadiman, A. Picelli, S. Regazzo, E. Stirbu, D. Primon, C. Bosello, A. R. Gravina, L. Peron, M. Trevisan, A. C. Garcia, A. Menel, L. Bloccari, N. Valè, L. Saltuari, M. Tinazzi, and N. Smania, "Virtual Reality Telerehabilitation for Postural Instability in Parkinson's Disease: A Multicenter, Single-Blind, Randomized, Controlled Trial," *BioMed Research International*, vol. 2017, 2017.

[27] J. Prvu Bettger, C. L. Green, D. N. Holmes, A. Chokshi, R. C. Mather, 3rd, B. T. Hoch, A. J. de Leon, F. Aluisio, T. M. Seyler, D. J. Del Gaizo, J. Chiavetta, L. Webb, V. Miller, J. M. Smith, and E. D. Peterson, "Effects of virtual exercise rehabilitation in-home therapy compared with traditional care after total knee arthroplasty: Veritas, a randomized controlled trial," *The Journal of Bone and Joint Surgery. American Volume*, vol. 102, no. 2, pp. 101–109, 2020. [Online]. Available: https://doi.org/10.2106/JBJS.19.00695

[28] T. Zhou, M. N. S. Iskandar, and K.-H. Chiam, "Diffusion models enable zero-shot pose estimation for lower-limb prosthetic users," *PLOS Digital Health*, vol. 4, no. 3, p. e0000745, 2025.

[29] F. Doerks, F. Harms, M. Schwarze, E. Jakubowitz, and B. Welke, "Gait assessment using a 2D video-based motion analysis app in healthy subjects and subjects with lower limb amputation: a pilot study," *PLOS ONE*, vol. 20, no. 5, p. e0324499, 2025.

[30] T. Nguyen Gia, I. B. Dhaou, M. Ali, A. M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Energy efficient fog-assisted IoT system for monitoring diabetic patients with cardiovascular disease," *Future Generation Computer Systems*, vol. 93, pp. 198–211, 2019.

[31] G. R. Retzinger, W. T. Pham, B. A. Bingham, K. E. Hori, and J. W. Hinkel-Lipsker, "Overground to omnidirectional: Spatiotemporal and muscle activation pattern adaptations across walking modalities, virtual reality, and pace constraints," *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 13, p. 20556683261433969, 2026. [Online]. Available: https://doi.org/10.1177/20556683261433969

[32] L. Schalbetter, A. Grêt-Regamey, F. Gutscher, and U. Wissen Hayek, "High-fidelity immersive virtual reality environments for gait rehabilitation exergames," *Frontiers in Virtual Reality*, vol. 5, no. January, pp. 1–19, 2024.

[33] X. Yuan, T. Sparling, D. Carlisle, S. McGavin, T. Riggleman, E. Cuevas, S. Ojha, D. G. Smith, R. Sheehan, and G. Kim, "Integrating gait analysis systems in clinical rehabilitation settings for individuals with lower limb amputation: a narrative review," *Cureus*, vol. 17, no. 8, p. e91030, 2025.

[34] A. Sacco, F. Esposito, and G. Marchetto, "Restoring Application Traffic of Latency-Sensitive Networked Systems Using Adversarial Autoencoders," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 2521–2535, 2022.

Mental Workload and Performance in Motor Imagery-Based Brain-Computer Interfaces: A Rapid Systematic Review

1st Lucas Tadeu dos Santos Vieira
Postgraduate Program in Biomedical Engineering - PPGEB
Federal Technological University of Paraná - UTFPR
Curitiba, PR, Brazil.
0009-0002-4898-3598

2nd Andrea Moreira Arrué
Postgraduate Program in Biomedical Engineering - PPGEB
Federal Technological University of Paraná - UTFPR
Federal Institute of Paraná - IFPR
Curitiba, PR, Brazil.
0000-0001-5391-324X

3rd Gabriel Chemin Rosenmann
Design Department
Federal University of Paraná - UFPR
Curitiba, PR, Brazil
0000-0002-2773-0236

4th Lívia Helena Moreira
Dept. Biomedical Engineering
Anhembi Morumbi University – UAM
Center for Innovation, Technology and Education - CITÉ
São José dos Campos, SP, Brazil
0000-0001-6432-7596

5th Henrique Cunha Carvalho
Postgraduate Program in Biomedical Engineering - PPGEB
Federal Technological University of Paraná - UTFPR
Curitiba, PR, Brazil.
0000-0003-0806-4132

***Abstract*— Mental workload is a relevant but inconsistently measured factor in motor imagery-based brain-computer interfaces (MI-BCIs), particularly in applications aimed at neuromotor rehabilitation. This rapid systematic review aimed to synthesize evidence on the relationship between mental workload or mental fatigue and BCI performance, identify the metrics used, and examine methodological limitations. Searches were conducted in IEEE Xplore, PubMed/MEDLINE, Web of Science, and Scopus for studies published between 2016 and 2026, with complementary searching in Google Scholar. Eligible studies included human participants, involved motor imagery-based BCI or rehabilitation-related applications, assessed mental workload or fatigue, and reported performance-related outcomes. Methodological quality was appraised using the MMAT. Fourteen studies were included. Overall, higher mental workload or fatigue tended to be associated with poorer BCI performance, reduced EEG class separability, altered cortical engagement, or decreased motor imagery stability. However, effects varied across users, sessions, and experimental conditions, and some studies reported preserved accuracy despite increased workload or fatigue, indicating that accuracy alone may not capture these changes. Subjective and neurophysiological measures provided complementary information. Only two studies (14.3%) included post-stroke participants, limiting clinical generalizability. Small samples, heterogeneous measurement approaches, and methodological limitations further reduced the strength of the evidence. Future studies should include larger clinical samples and standardized multimodal protocols integrating workload, fatigue, neurophysiological markers, and objective BCI performance outcomes.**

Keywords*— *brain-computer interface; motor imagery; mental workload; electroencephalography; neuromotor rehabilitation.

I. INTRODUCTION

A brain-computer interface (BCI) establishes a direct communication pathway between brain activity and external devices, enabling the translation of neural signals into operational commands without relying on peripheral neuromuscular pathways [1]. Among non-invasive modalities, electroencephalography (EEG) based systems and motor imagery (MI) have received increasing attention in neuroengineering and rehabilitation, especially in applications targeting patients with stroke, spinal cord injury, and other disabling neurological conditions [1,2]. In post-stroke rehabilitation, systematic review and meta-analysis evidence indicates that BCI training combined with conventional therapy may promote upper-limb motor recovery and brain function recovery. However, more standardized protocols, assessment of MI ability, and control of neuropsychological factors such as attention and concentration are still needed [2]. The neurophysiological principle of MI-based BCIs involves the modulation of sensorimotor rhythms or sensorimotor alpha rhythms, approximately between 8–13 Hz, and beta rhythms, between 13–30 Hz, expressed through event-related desynchronization and synchronization patterns (ERD/ERS), observed during movement execution or imagination [1,3].

Mental workload reflects the relationship between the cognitive demands imposed by a task and the individual's available resources to perform it, and can be estimated using performance measures, subjective self-reports, and psychophysiological or neurophysiological markers [4,5]. In the context of MI-based BCIs, this workload is modulated by task complexity, sustained attention, presence of distractors, session duration, fatigue, motivation, and the user's emotional state [1,4,6]. Recent evidence indicates that age, cognitive function, attention, mental rotation ability, and mental state explain a relevant portion of the variability in EEG-based BCI system performance; MI paradigms appear to be more sensitive to cognitive and developmental differences than visual paradigms [1]. Mental fatigue, in turn, should be understood as a psychophysiological state that can reduce user responsiveness, alter cortical power patterns, attenuate ERD responses, and compromise EEG signal separability and classification accuracy, especially during prolonged MI tasks [1,3,7].

It is important to distinguish between related but distinct constructs that are often conflated in the BCI literature.

Mental workload refers to the relationship between task demands and the individual's available cognitive resources [8]. Mental fatigue is a psychobiological state caused by prolonged cognitive activity, characterized by subjective tiredness and reduced performance [9]. Mental effort represents the active allocation of resources to meet task demands [10], while attention involves the selective focus on relevant stimuli [11]. This review encompasses studies using any of these terms, as they are often interconnected in BCI research, however, we explicitly acknowledge their conceptual distinctions in our analysis.

As illustrated in Figure 1, electrodes positioned on the scalp capture EEG signals generated during MI, which undergo preprocessing, feature extraction, and classification stages, producing neurofeedback or commands for external devices. In parallel, mental workload can be estimated through electrophysiological markers, behavioral metrics, or subjective instruments. In recent studies with wearable EEG and fNIRS devices, aspects such as comfort, adaptability to head shape, electrode positioning, signal quality, portability, interoperability, usability, and user acceptability have been recognized as central requirements for expanding BCI applications from laboratory settings to clinical and real-life contexts [4].

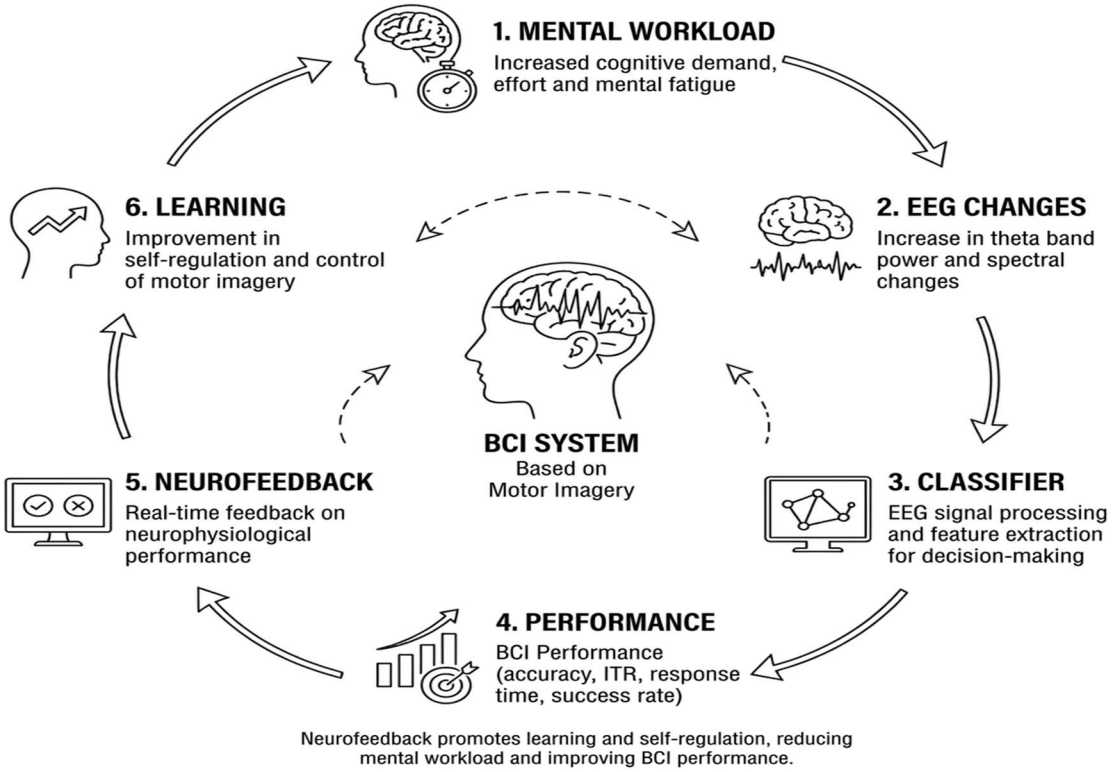


Fig. 1. Flowchart of the EEG signal acquisition process and user neurofeedback.

Given this scenario, this review was guided by the following question: among users of motor imagery-based brain-computer interfaces, does exposure to high mental workload, compared with lower cognitive-demand conditions, compromise interface performance?

The objective of this rapid systematic review was to analyze evidence from experimental studies that investigated the relationship between mental workload and performance in motor imagery-based brain-computer interfaces, describing the metrics used to assess mental workload and interface performance, evaluating the risk of bias of the included studies, and identifying individual, neurophysiological, and operational factors that modulate this relationship, especially in applications aimed at neuromotor rehabilitation.

II. METHOD

This rapid systematic review followed a predefined a priori protocol that was prospectively registered with the Open Science Framework (OSF) (https://osf.io/ahuym). The review process took place between March and June 2026, divided into two phases: (i) electronic database searches, performed from March to April 2026 (last search conducted on April 30, 2026); and (ii) screening of records, data extraction, and evidence synthesis, carried out from April to June 2026. The searches were performed in IEEE Xplore, PubMed/MEDLINE, Web of Science, and Scopus, covering the period from January 2016 to April 2026.

This 10-year window was selected to capture the contemporary evidence base that followed the consolidation of standardized workload metrics and deep learning approaches in MI-BCI research. This period includes studies on neurorehabilitation applications of BCIs [12], EEG-controlled exoskeletons [13,14], and the increasing application of deep learning methods for EEG decoding [15,16]. Review articles published during this period also summarized classification approaches and performance metrics for MI-EEG [17]. This timeframe allows for the inclusion of recent developments while maintaining a scope that is feasible for a rapid, systematic review.

Google Scholar was also employed as a complementary source of grey literature to identify potentially relevant studies not retrieved from the indexed databases. Studies published in either English or Portuguese were included. Figure 2 presents the study selection flow diagram, adapted from the PRISMA 2020 statement [18].

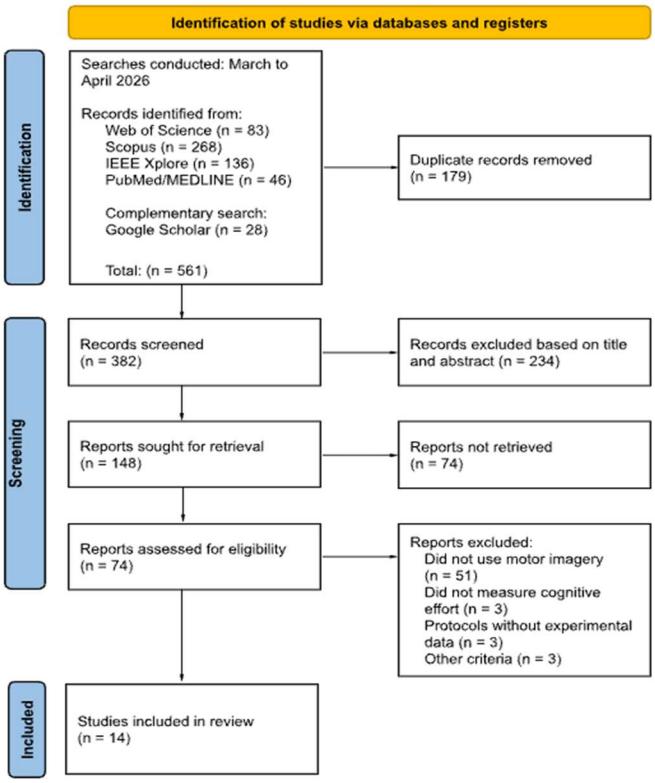


Fig. 2. PRISMA flow diagram of the study selection process.

The research question was structured according to the PICOT acronym, as outlined below:

- P (Population): users of motor imagery-based brain-computer interfaces;
- I/E (Intervention/Exposure): exposure to high mental workload, mental fatigue, or high cognitive demand;
- C (Comparison): low mental workload, control conditions, or lower cognitive demand;
- O (Outcome): BCI performance assessed by metrics such as accuracy, information transfer rate (ITR), ERD/ERS patterns, response time, and success rate;
- T (Time): studies published between 2016–2026.

The methodological quality and risk of bias of the included studies were appraised using the Mixed Methods Appraisal Tool (MMAT), version 2018 [19]. The MMAT was selected because it is specifically designed to assess the quality of studies with heterogeneous methodological designs, including qualitative, quantitative, and mixed-methods research. This characteristic was essential for this review, given the diversity of experimental protocols, populations, and outcome measures observed in the MI-BCI literature.

A. Search Strategy

The definition of search terms was guided by the three central axes of the review: brain-computer interfaces, mental workload, and interface performance. For the first axis, the terms “BCI” and “brain-computer interface” were used; for the second, "cognitive load," "mental effort,” and “mental workload”; and for the third, “performance.” In Google Scholar, “motor imagery” and “rehabilitation” were added to refine the search for studies with greater adherence to clinical applications based on motor imagery. The search strategies employed Boolean operators AND and OR, quotation marks for exact phrases, and parentheses for logical grouping, with adaptations tailored to the syntax and available search fields of each database.

In PubMed/MEDLINE, the following strategy was used: (BCI OR “brain-computer interface”) AND (“cognitive load” OR “mental effort”) AND performance. In IEEE Xplore, the following was employed: (BCI OR “brain-computer interface”) AND (“cognitive load” OR “mental effort” OR “mental workload”) AND performance. In Scopus, the search was conducted in the title, abstract, and keyword fields, using: TITLE-ABS-KEY((BCI OR “brain-computer interface”) AND (“cognitive load” OR “mental effort” OR “mental workload”) AND performance). In Web of Science, the search was conducted in the topic field, with the following strategy: TS=((BCI OR “brain-computer interface”) AND (“cognitive load” OR “mental effort” OR “mental workload”) AND performance).

As a complementary search, Google Scholar was consulted using the combination: BCI “motor imagery” “mental workload” “cognitive load” “rehabilitation” performance. The most relevant results retrieved by the platform were considered, according to the pre-defined eligibility criteria.

B. Screening Process

For reference management and study screening, Rayyan, in its web version developed by the Qatar Computing Research Institute (QCRI), was used [20]. The tool was employed to organize records retrieved from databases, identify duplicates, and support the selection process of eligible studies. Title/abstract screening, data extraction, and risk-of-bias assessment were carried out independently by two reviewers. Disagreements were resolved by consensus. The full list of excluded full-text articles with reasons for exclusion is provided in Supplementary Material S1 (Annex A).

C. Eligibility Criteria

Primary studies of experimental, observational, or pilot design, conducted with human participants, that investigated motor imagery-based brain-computer interfaces, measured mental workload, cognitive effort, or mental fatigue, and reported interface performance metrics such as accuracy, response time, error rate, success rate, or ITR were included. Studies developed in the context of neuromotor rehabilitation, clinical application, or experimental simulation with potential relevance to rehabilitation were also considered.

D. Data Extraction and Synthesis

Data were extracted directly from the published texts, tables, and figures as reported by the original authors. No missing summary statistics were imputed, and no data conversions or transformations were performed. Because a meta-analysis was not feasible due to substantial heterogeneity in study designs, populations, workload metrics, and performance outcomes, no pooled effect measures were calculated. Instead, effect directions and estimates were extracted and reported narratively as presented in the primary studies.

III. Results and Discussion

The search and selection process yielded 14 eligible studies. The main reasons for excluding full-text reports were: (i) absence of a motor imagery paradigm; (ii) lack of mental workload or fatigue assessment; and (iii) non-experimental designs. A complete list of excluded studies is available from the corresponding author upon request.

When analyzing the included articles presented in Table 1, 10 of 14 selected studies (71.4%) reported an inverse relationship between increased mental workload and user performance on the interface. This association was evidenced by reduced classification accuracy, decreased separability between EEG signal classes, alterations in ERD/ERS patterns, and impaired cortical engagement during motor imagery tasks [21–30].

Additionally, 2 studies (14.3%) identified a context-dependent relationship, suggesting that the impact of mental workload may vary depending on the experimental paradigm, task type, user profile, exposure time, or feedback strategy [31,32]. Another 2 studies (14.3%) did not observe a statistically significant association between mental workload and interface performance [33,34].

A conceptual challenge identified in the included studies is the inconsistent use of terminology and measurement approaches for mental workload, fatigue, and related constructs. Studies [21,22,26,28,30,32,34] used the NASA-TLX as a subjective measure of mental workload, while [24,25,33] focused specifically on mental fatigue using subjective scales (BFI, fatigue scale) and EEG spectral markers. Some studies [23,29,31] addressed cognitive state, attention, and mental load more broadly, using both subjective ratings and neurophysiological metrics.

Sim *et al*. [23] explicitly examined attention and somnolence, whereas Yang et al. [31] used fNIRS to assess cognitive load objectively. Wang *et al*. [29] focused on mental load in the context of post-stroke patients, using MRCP and ERD as markers of neural engagement. This terminological heterogeneity reflects broader discussions in the literature and complicates direct comparisons between studies. Future research should adopt clearer operational definitions and consider using multiple complementary measures, subjective, electrophysiological, and autonomic, to capture the distinct dimensions of cognitive demand, fatigue, and attentional state in BCI users.

Regarding measurement approaches, the studies employed a wide range of parameters to assess mental workload and performance, subjective scales (NASA-TLX, BFI, VAS, SAM, SUS, TAM, QUEST 2.0), spectral EEG markers (theta/alpha ratio, DAR, DABR, DTABR, spectral power in delta/theta/alpha/beta bands), physiological signals (HRV, GSR, pupil dilation, blink rate), neuroimaging (fNIRS), and behavioral metrics (classification accuracy, ITR, response time, command success rate, MRCP amplitude, ERD patterns). This diversity underscores the need for standardized, multimodal protocols in future MI-BCI research to improve comparability and replicability.

The studies included in this review exhibit substantial heterogeneity not only in terminology and measurement approaches, as detailed above, but also in experimental protocols, classification algorithms, participant characteristics, and application contexts. These factors likely modulate the relationship between mental workload and BCI performance and are discussed below across seven thematic topics:

A. User Proficiency and Vulnerability to Mental Workload

Regarding the relationship between user proficiency and vulnerability to mental workload, Emami and Chau [21] demonstrated that subjective mental workload, measured by the NASA Task Load Index (NASA-TLX), significantly predicted performance decline among participants with accuracy below 75%. This effect was not observed among participants with better performance, suggesting that less proficient users, or those in the initial adaptation phase to the interface, may be more vulnerable to high cognitive demand.

Complementarily, Ozkan and Kahya [22] found that participants classified as being in a high-risk cognitive state performed significantly worse, with agreement ranging from 87.1% to 100% across alternative classification methods. This finding indicates that the cognitive profile may represent a relevant and potentially predictive dimension of user behavior on the interface.

Sim *et al*. [23], in turn, demonstrated that incorporating the cognitive attention state as an auxiliary signal in a multi-task EEGNet-TCN decoder elevated classification accuracy to 75.9%, surpassing the model without cognitive state supervision by 4.5 percentage points. This result reinforces the idea that information about the user's attention and cognitive state can add predictive value to BCI performance, even when obtained through simplified monitoring strategies.

B. Impact of Mental Fatigue and Observed Exceptions

Mental fatigue emerged as a recurring factor in performance degradation in motor imagery-based brain-computer interfaces, although the neural mechanisms involved and the most sensitive metrics varied across studies. Talukdar *et al*. [24] demonstrated that prolonged MI sequences produce substantial mental fatigue, significantly reducing separability between EEG signal classes and directly compromising the quality of the classifier's input signal.

Li *et al*. [25] observed a more refined temporal pattern: the MI recognition rate increased in the first three blocks, in a dynamic suggestive of learning or initial adaptation to the task, but declined in the fourth block, when mental fatigue became more evident. This decline was accompanied by increased theta power in frontal, central, parietal, and occipital regions, as well as weakened phase synchronization between these areas, suggesting progressive deterioration of the functional integration necessary for interface control.

However, not all studies confirmed a direct association between mental fatigue and performance decline. Roc *et al*. [33] did not find a significant correlation between performance and subjective measures of mental workload across three sessions, although they recorded a consistent increase in mental demand and fatigue within each session. This divergence may reflect limitations of the self-report scales used, interindividual variability, or the particularities of the initial adaptation process for inexperienced users of the system.

Similarly, Yang *et al*. [31] observed that the virtual reality environment increased mental workload and fatigue without significantly reducing accuracy. However, MI stability decreased, suggesting that the impact of elevated mental workload is not always captured by accuracy alone. Thus, complementary metrics such as EEG pattern stability, inter-class separability, ERD/ERS, functional connectivity, and subjective fatigue measures may be necessary for a more sensitive assessment of the effects of mental fatigue on interface performance.

C. *Influence of Neurofeedback and Interface Design*

Arpaia *et al*. [28] demonstrated that real-time neurofeedback increased accuracy from 62% to 69% while maintaining high perceived usability in both groups. This finding suggests that continuous feedback may favor learning of conscious interface control without necessarily increasing the perceived cognitive cost in proportion.

He *et al*. [27], in the only randomized clinical trial included in the sample, evaluated 48 post-stroke patients, distributed between an intervention group with BCI (n=25) and a control group (n=23). The authors observed significantly greater improvement in upper-limb motor function in the group that underwent BCI with multimodal neurofeedback in virtual reality, accompanied by evidence of neuroplasticity assessed using EEG and fNIRS. These results position neurofeedback not only as a user feedback resource but as an active therapeutic component in the neuromotor rehabilitation process.

In a convergent direction, Li *et al*. [26] reported that an asynchronous MI-BCI system, developed based on cognitive ergonomics principles, reduced subjective workload, as measured by the NASA-TLX, by 32% compared to synchronous paradigms. This reduction was accompanied by 27% fewer commands and 34% less time to complete tasks, thereby directly linking the decrease in mental workload to operational performance gains.

Lyu *et al*. [30] reinforced this perspective by observing, during a rapid, iterative prototyping process for a BCI system aimed at lower-limb rehabilitation, that simpler MI tasks reduced perceived mental workload and yielded more reliable brain signals. Furthermore, performance and satisfaction are correlated, indicating that usability is a dimension that can amplify or reduce overall interface performance.

In turn, Lu *et al*. [32], adopting a user experience-centered approach, reported an average online accuracy of 85% and a latency of 0.12 seconds in an MI-BCI system with virtual reality. These results were associated with moderate levels of mental workload and a strong sense of agency reported by participants. This finding suggests that moderate mental workload, far from representing only a risk factor, may constitute a favorable condition for effective engagement with the interface, provided it is balanced by adequate feedback, intuitive design, and good usability.

In post-stroke patients (n=6), Wang *et al*. [29] observed that the cognitively more demanding condition produced later and attenuated movement-related cortical potentials. These results indicate that reducing cognitive demand during the task may favor more consistent neurophysiological responses and, potentially, amplify the therapeutic effect of BCI in populations with neurological impairment.

D. *Methodological Variability in Mental Workload Measurement*

It is observed that there is still no methodological standardization for assessing mental workload in studies with MI-based BCIs, as the works included in this review employed different measurement approaches. Some studies used the NASA-TLX exclusively [26,28,30,32], while others relied on electrophysiological markers, especially spectral ratios between the theta and alpha bands [23–25,27]. There are still studies that combine subjective measures, such as the NASA-TLX, with physiological indicators, including heart rate variability (HRV) [23], pupillometry, and galvanic skin response (GSR) [22].

Emami and Chau [21], in turn, combined the NASA-TLX with the theta/alpha spectral ratio and found that the theta-alpha ratio with parietal-parietal topography showed 22.5% higher sensitivity than the conventional fronto-parietal ratio in capturing changes in mental workload during tasks with visual distractors. Badesa *et al*. [34] also reinforce the relevance of complementary physiological measures by demonstrating that, among the modalities evaluated, HRV was the only measure with a statistically significant difference between participants with good and poor performance.

Taken together, these findings indicate that a single isolated metric should not be used to evaluate mental workload in BCI. The combination of subjective instruments, electrophysiological markers, and autonomic indicators can offer a more sensitive and multidimensional characterization of the cognitive demand imposed on the user.

E. *Methodological Quality and Risk of Bias of Included Studies*

The methodological quality of the 14 included studies was assessed using the Mixed Methods Appraisal Tool (MMAT), version 2018 [19]. The results are summarized in Table 1. Only the study by He *et al*. [27] showed a low risk of bias, meeting all five quality criteria for randomized clinical trials.

Seven studies were classified as moderate risk [21–23,26,28,31,33], mainly due to methodological weaknesses such as absence of randomization, lack of sample size calculation, insufficient control of potential confounders, and limitations in the description of experimental procedures. Another six studies were classified as high risk of bias [24,25,29,30,32,33] for presenting descriptive or pilot designs, small samples (generally $n \leq 11$), the absence of a control group, and convenience sampling.

Specific relevant limitations were also identified, such as gender bias in Yang *et al*. [31], whose sample included only male participants, and the 45% sample loss in Talukdar *et al*. [24], without sensitivity analysis to assess the impact of these losses on the results. Thus, the predominance of studies with a moderate to high risk of bias requires caution in interpreting findings and highlights the need for future investigations with more consistent designs, larger sample sizes, comparator groups, randomization strategies, and better control of confounders. Additionally, as this is a rapid systematic review focusing on qualitative evidence synthesis, formal assessments of publication bias and certainty of the body of evidence were not performed.

A total of 14 studies were included, which evaluated fatigue, mental workload, performance, accuracy, neurophysiological activity, and usability of BCI systems. The majority of studies were conducted with healthy participants, while two included post-stroke individuals.

Overall, the results indicated an association between higher fatigue/mental workload and reduced BCI performance, although some studies observed preserved accuracy despite increased cognitive demand. In post-stroke studies, changes in neurophysiological markers were associated with cortical engagement and motor function.
.

F. Synthesis of Results

TABLE I. SUMMARY OF INCLUDED STUDIES

Author(s) / Year	Population (n)	Variability Type	Main Outcomes	Observed Relationship	Risk of Bias (MMAT)
[21] Emami & Chau (2020)	Healthy: *n*=16	Inter-user	Fatigue, EEG, accuracy	↑ theta/alpha → ↓ accuracy (R^2=0.32)	Moderate
[22] Ozkan & Kahya (2018)	Healthy: *n*=70	Inter-user	Cognitive load, performance	High cognitive state → higher risk of low performance (87.1–100%)	Moderate
[23] Sim *et al.* (2026)	Healthy: *n*=54	Inter-user / individual state	Attention, somnolence, accuracy	↑ attention → ↑ accuracy (*r*=0.25); somnolence = worse performance	Moderate
[24] Talukdar *et al.* (2019)	Healthy: *n*=11	Intra-user	Fatigue, EEG, accuracy	↑ delta/theta/alpha → ↓ class separability and accuracy	High
[25] Li *et al.* (2024)	Healthy: *n*=10	Intra-user	Fatigue, EEG, MI recognition	↑ fatigue/theta → ↓ MI recognition rate	High
[26] Li *et al.* (2026)	Healthy: *n*=16	Not specified	Mental workload, accuracy	↓ NASA-TLX (32%) → 95.36% online accuracy	Moderate
[27] He *et al.* (2025)	Post-stroke: BCI (*n*=25) Control (*n*=23)	Intergroup + longitudinal	EEG, FMA-UE, EMG	FMA-UE: 4.0 points in BCI group vs. 2.0 in control; ↓ DAR/DABR	Low
[28] Arpaia *et al.* (2023)	Healthy: Neurofeedback (*n*=13) Control (*n*=14)	Intergroup	Effort, MI, usability	Accuracy: 69% vs. 62% (neurofeedback *vs*. control)	Moderate
[29] Wang *et al.* (2025)	Post-stroke: *n*=6	Intra-user	MRCP, ERD, mental load	↑ MRCP/ERD → ↑ cortical engagement	High
[30] Lyu *et al.* (2023)	Healthy: *n*=10	Not specified	Mental workload, satisfaction	↑ NASA-TLX → ↓ user satisfaction	High
[31] Yang *et al.* (2023)	Healthy: *n*=9	Intra-user / between conditions	Mental load, fNIRS, MI	Accuracy preserved (51.32%); ↓ MI stability	Moderate
[32] Lu *et al.* (2025)	Healthy: *n*=6	Limited evidence for inter/intra	Workload, usability, performance	Accuracy ≈85%; latency ≈0.12s	High
[33] Roc *et al.* (2026)	Healthy: *n*=24	Intra-user	Fatigue, demand, accuracy	↑ fatigue/demand, but no significant correlation with accuracy	Moderate
[34] Badesa *et al.* (2020)	Healthy: *n*=11	Inter-user	HRV, workload, BCI success	↓ HRV and ↑ NASA-TLX → worse performance	High

Abbreviations: BCI: brain-computer interface; EEG: electroencephalography; MI: motor imagery; FMA-UE: Fugl-Meyer Assessment – Upper Extremity; EMG: electromyography; MRCP: movement-related cortical potentials; ERD: event-related desynchronization; fNIRS: functional near-infrared spectroscopy; HRV: heart rate variability; NASA-TLX: NASA Task Load Index; DAR: delta/alpha ratio; DABR: delta/alpha/beta ratio; MMAT: Mixed Methods Appraisal Tool.

G. Implications for Rehabilitation and Directions for Future Research

Considering these limitations, future randomized clinical trials should explore adaptive motor imagery training protocols that use continuous monitoring of mental workload to adjust the intensity, complexity, and progression of tasks across sessions in neurological patients. This approach would allow investigation into whether maintaining mental workload at tolerable or moderate levels favors better functional outcomes compared to fixed protocols, and would also contribute to defining more consistent parameters for the frequency, duration, intensity, and progression of training.

IV. Conclusion

The evidence synthesized in this rapid systematic review suggests that elevated mental workload and mental fatigue are generally associated with poorer performance in motor imagery-based BCIs. However, the magnitude and expression of these effects vary across users, sessions, and experimental conditions. Subjective measures and neurophysiological markers provided complementary information, and some studies showed that accuracy alone may not capture changes in MI stability, EEG separability, cortical engagement, or fatigue. Evidence in post-stroke populations remains limited and does not yet support firm clinical conclusions, although preliminary findings suggest that neurophysiological monitoring may help characterize cortical engagement and motor response during BCI-based rehabilitation. Methodological heterogeneity, small sample sizes, and the predominance of studies with methodological limitations restrict the strength and generalizability of the evidence. Future studies should include larger clinical samples and standardized multimodal protocols integrating mental workload, fatigue, neurophysiological markers, and objective BCI performance outcomes.

Acknowledgments

This rapid systematic review was conducted as part of the Health Technology Assessment (HTA) course in the Postgraduate Program in Biomedical Engineering (PPGEB) at the Federal Technological University of Paraná (UTFPR). The authors would like to thank the UTFPR for providing all resources for the development of the project, and the Araucária Foundation (FA) & Superintendence of Science, Technology and Higher Education (SETI) for the conference scholarship for H. C. C.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All data extracted from the included studies are presented in Table I. Supplementary Material S1, containing the full list of excluded studies with reasons (Annex A) is permanently available online at https://osf.io/munf5. Data collection forms and other materials used in the review are available upon reasonable request from the corresponding author.

References

[1] X. Niu, M. Yuan, and D. Wang, "Influence of age, cognitive function, attention, and mental state on the effectiveness of EEG-based brain-computer interface device use: A systematic review," J. Neuroeng. Rehabil., vol. 22, p. 270, 2025, doi: 10.1186/s12984-025-01813-7.

[2] A. Kruse, Z. Suica, J. Taeymans, and C. Schuster-Amft, "Effect of brain-computer interface training based on non-invasive electroencephalography using motor imagery on functional recovery after stroke: A systematic review and meta-analysis," BMC Neurol., vol. 20, p. 385, 2020, doi: 10.1186/s12883-020-01960-5.

[3] M. Saini and U. Satija, "State-of-the-art mental tasks classification based on electroencephalograms: a review," Physiol. Meas., vol. 44, no. 6, p. 06TR01, 2023, doi: 10.1088/1361-6579/acd51b.

[4] I. Lombardi, M. Buono, G. Giugliano, V. P. Senese, and S. Capece, "Usability and acceptance analysis of wearable BCI devices," Appl. Sci., vol. 15, no. 7, p. 3512, 2025, doi: 10.3390/app15073512.

[5] K. Kyriaki and C. A. Fidas, "A Human–Computer-Interaction exploration of working memory using EEG," *IEEE Access*, vol. 13, pp. 157877–157908, 2025, doi: 10.1109/ACCESS.2025.3606802.

[6] C. Reichert, S. Dürschmid, M. V. Bartsch, J. M. Hopf, H. J. Heinze, and H. Hinrichs, "Decoding the covert shift of spatial attention from electroencephalographic signals permits reliable control of a brain-computer interface," J. Neural Eng., vol. 17, no. 5, p. 056012, 2020, doi: 10.1088/1741-2552/abb692.

[7] U. Talukdar, S. M. Hazarika, and J. Q. Gan, "Adaptive feature extraction in EEG-based motor imagery BCI: tracking mental fatigue," J. Neural Eng., vol. 17, no. 1, p. 016020, Jan. 2020, doi: 10.1088/1741-2552/ab53f1.

[8] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," Ergonomics, vol. 58, no. 1, pp. 1–17, Jan. 2015, doi: 10.1080/00140139.2014.956151.

[9] M. A. S. Boksem and M. Tops, "Mental fatigue: Costs and benefits," Brain Res. Rev., vol. 59, no. 1, pp. 125–139, Nov. 2008, doi: 10.1016/j.brainresrev.2008.07.001.

[10] F. G. W. C. Paas and J. J. G. Van Merriënboer, "Measurement of cognitive load in instructional research," Percept. Mot. Skills, vol. 79, no. 1, pp. 419–430, Aug. 1994, doi: 10.2466/pms.1994.79.1.419.

[11] M. Sarter, B. Givens, and J. P. Bruno, "The cognitive neuroscience of sustained attention: where top-down meets bottom-up," Brain Res. Rev., vol. 35, no. 2, pp. 146–160, Jun. 2001, doi: 10.1016/s0165-0173(01)00044-3.

[12] M. A. Lebedev and M. A. L. Nicolelis, "Brain-Machine Interfaces: from basic science to neuroprostheses and neurorehabilitation," Physiological Reviews, vol. 97, no. 2, pp. 767–837, Apr. 2017, doi: 10.1152/physrev.00027.2016.

[13] S. R. Soekadar, M. Witkowski, C. Gómez, *et al*., "Hybrid EEG/EOG-based brain/neural hand exoskeleton restores fully independent daily living activities after quadriplegia," Science Robotics, vol. 1, no. 1, p. eaag3296, Dec. 2016, doi: 10.1126/scirobotics.aag3296.

[14] R. A. R. C. Gopura, D. S. V. Bandara, K. Kiguchi, and G. K. I. Mann, "Upper-limb robotic exoskeletons for neurorehabilitation: a review on control strategies," IEEE Reviews in Biomedical Engineering, vol. 9, pp. 4–14, 2016, doi: 10.1109/RBME.2016.2552201.

[15] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," Journal of Neural Engineering, vol. 14, no. 1, p. 016003, Feb. 2017, doi: 10.1088/1741-2552/14/1/016003.

[16] M. Hamedi, S.-H. Salleh, and A. M. Noor, "Electroencephalographic motor imagery brain connectivity analysis for BCI: A review," Neural Computation, vol. 28, no. 6, pp. 999–1041, Jun. 2016, doi: 10.1162/NECO_a_00838.

[17] K. K. Ang and C. Guan, "EEG-Based strategies to detect motor imagery for control and rehabilitation," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 4, pp. 392–401, Apr. 2017, doi: 10.1109/TNSRE.2016.2646763.

[18] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[19] Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, Gagnon M-P, Griffiths F, Nicolau B, O'Cathain A, Rousseau M-C, Vedel I. Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

[20] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews," *Syst Rev*, vol. 5, no. 1, p. 210, Dec. 2016, doi: 10.1186/s13643-016-0384-4.

[21] Z. Emami and T. Chau, "The effects of visual distractors on cognitive load in a motor imagery brain-computer interface," *Behavioural Brain Research*, vol. 378, p. 112240, Jan. 2020, doi: 10.1016/j.bbr.2019.112240.

[22] N. F. Ozkan and E. Kahya, "Classification of BCI Users Based on Cognition," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–10, 2018, doi: 10.1155/2018/6315187.

[23] D.-H. Sim, W.-S. Oh, and D.-H. Jeong, "Multi-Task EEGNet-TCN decoder of motor imagery and cognitive state in motor imagery BCI," in *2026 14th International Conference on Brain-Computer Interface (BCI)*, Gangwon Province, Korea, Republic

of: IEEE, Feb. 2026, pp. 1–5. doi: 10.1109/BCI69045.2026.11435078.

[24] U. Talukdar, S. M. Hazarika, and J. Q. Gan, "Motor imagery and mental fatigue: inter-relationship and EEG based estimation," *J Comput Neurosci*, vol. 46, no. 1, pp. 55–76, Feb. 2019, doi: 10.1007/s10827-018-0701-0.

[25] T. Li *et al.*, "Research on mental fatigue during long-term motor imagery: a pilot study," *Sci Rep*, vol. 14, no. 1, p. 18454, Aug. 2024, doi: 10.1038/s41598-024-69013-2.

[26] Q. Li, Q. Huang, W. Jiang, Z. Song, and X. Wu, "Hands-Free Mobility Control: A low-latency asynchronous MI-BCI system for real-time robotic Navigation With Cognitive-Ergonomic Enhancement," *IEEE Trans. Human-Mach. Syst.*, vol. 56, no. 2, pp. 291–300, Apr. 2026, doi: 10.1109/THMS.2026.3659708.

[27] J. He *et al.*, "Multimodal assessment of a BCI system for stroke rehabilitation integrating motor imagery and motor attempts: a randomized controlled trial," *J NeuroEngineering Rehabil*, vol. 22, no. 1, p. 185, Aug. 2025, doi: 10.1186/s12984-025-01723-8.

[28] P. Arpaia *et al.*, "Paving the way for motor imagery-based tele-rehabilitation through a fully wearable BCI System," *Sensors*, vol. 23, no. 13, p. 5836, Jun. 2023, doi: 10.3390/s23135836.

[29] Z. Wang, Y. Wang, H. Yan, S. Han, and J. Wang, "Neural engagement in motor imagery EEG analysis of active vs. guided commands in post-stroke patients," *IFAC-PapersOnLine*, vol. 59, no. 35, pp. 174–179, 2025, doi: 10.1016/j.ifacol.2025.12.471.

[30] X. Lyu *et al.*, "Human factors engineering of BCI: an evaluation for satisfaction of BCI based on motor imagery," *Cogn Neurodyn*, vol. 17, no. 1, pp. 105–118, Feb. 2023, doi: 10.1007/s11571-022-09808-z.

[31] J. Yang, S. Zhu, P. Ding, F. Wang, A. Gong, and Y. Fu, "The Effects of VR and TP Visual cues on motor imagery subjects and performance," *Electronics*, vol. 12, no. 11, p. 2381, May 2023, doi: 10.3390/electronics12112381.

[32] A. Lu, M. Huang, K.-L. Liao, Y. Sun, and R. Yang, "Design of a real-time motor imagery BCI-VR system with sample segmentation and soft voting," in *2025 8th International Conference on Big Data and Artificial Intelligence (BDAI)*, Taicang, China: IEEE, Aug. 2025, pp. 47–52. doi: 10.1109/BDAI66031.2025.11325177.

[33] A. Roc, L. Kolodzienski, P. Dreyer, A. Appriou, T. Monseigne, and F. Lotte, "Evolution of users' subjective experience over three training sessions with an EEG Motor-Imagery Brain–Computer Interface (MI-BCI)," *Brain Research*, vol. 1872, p. 150085, Feb. 2026, doi: 10.1016/j.brainres.2025.150085.

[34] F. J. Badesa, J. A. Diez, J. A. Barios, J. M. Catalan, and N. Garcia-Aracil, "Evaluation of performance and heart rate variability during intensive usage of a BCI-controlled hand exoskeleton," in *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, New York City, NY, USA: IEEE, Nov. 2020, pp. 164–169. doi: 10.1109/BioRob49111.2020.9224322.

Modular BCI Framework for Robotic Rehabilitation: Hybrid EMG–IMU and SSVEP Architectures

Kevin Mateo Gutiérrez-Ortiz
Robotics and Assistive Technology Laboratory
Federal University of Espirito Santo (UFES)
Vitória, Brazil
kevin.ortiz@edu.ufes.br

Luis Carlos Rodríguez-Timaná
Faculty of Engineering / Rob. and Assist. Tech. Lab
Santiago de Cali Univ. (USC); Federal University of Espirito Santo
Cali, Colombia; Vitória, Brazil
luis.rodriguez11@usc.edu.co

Eliete Maria de Oliveira Caldeira
Robotics and Assistive Technology Laboratory
Federal University of Espirito Santo (UFES)
Vitória, Brazil
eliete.caldeira@ufes.br

Teodiano Freire Bastos-Filho
Robotics and Assistive Technology Laboratory
Federal University of Espirito Santo (UFES)
Vitória, Brazil
teodiano.bastos@ufes.br

***Abstract*—This work presents an instrumental Proof of Concept (PoC) for a modular Brain-Computer Interface (BCI) framework aimed at robotic rehabilitation. The system adopts an asynchronous, decoupled platform to support post-stroke therapies at the Center of Physical Rehabilitation of Espirito Santo (CREFES). Using the Emotiv EPOC X headset [14], the platform consolidates two independent control architectures: (1) a Hybrid mode that fuses facial electromyography (EMG) inferences as a safety trigger and inertial kinematic data (IMU) for proportional speed modulation; and (2) a Binary Visual Mode based on Steady-State Visually Evoked Potentials (SSVEP) at 15 Hz. Processing is performed in Python under the Model-View-Controller (MVC) pattern, transmitting commands via Wi-Fi to an ESP32-based module controlling a rehabilitation minibike. A pilot study with three healthy volunteers ($N = 3$) evaluated preliminary performance and workload. The Hybrid mode achieved a mean accuracy of 84.92% (SD = 14.35%) and an Information Transfer Rate (ITR) of 9.91 bits/min (SD = 8.99), while the SSVEP mode achieved 77.23% (SD = 5.96%) and 1.77 bits/min (SD = 0.81). NASA-TLX assessments suggested a trend towards lower mental workload for the Hybrid mode. The implementation of Exponential Moving Average (EMA) and temporal refractory filters effectively mitigated biosignal bouncing effects, establishing a methodological foundation for future clinical trials.**

***Index Terms*—BCI Framework, Robotic Rehabilitation, SSVEP, Emotiv EPOC X, ESP32, Cognitive Load.**

I. Introduction

Stroke is one of the leading public health challenges and the primary cause of permanent motor disability in adults. In Brazil, stroke accounts for approximately 11 deaths per hour [1]. Robot-assisted motor rehabilitation has proven effective for promoting neuroplasticity and accelerating functional recovery [2]. However, integrating Brain–Computer Interfaces (BCIs) into real hospital environments, such as the Center of Physical Rehabilitation of Espirito Santo (CREFES), presents significant operational barriers.

This work was developed in the Robotics and Assistive Technology Laboratory (LRTA) at the Universidade Federal do Espírito Santo (UFES).

Traditional paradigms, particularly pure Motor Imagery (MI), impose severe cognitive workloads, leading to mental fatigue and the phenomenon of *BCI illiteracy* [3], [5]. Cognitive fatigue induced by repetitive sessions alters the patient's neurophysiological response, rapidly saturating attention and reducing therapeutic efficacy over the medium term [4].

To address these limitations, this paper proposes a Decoupled Bimodal BCI Architecture. Unlike conventional approaches that rigidly link processing software to a single physical actuator, this work focuses on a formal separation of layers via standardized network commands and mathematical temporal filters to improve operational stability. Using the Emotiv EPOC X wireless headset, the control load is divided into two selectable processing architectures (Hybrid and Visual) to explore the reduction in cognitive demand associated with continuous voluntary control.

II. State of the Art

The design of BCI architectures for clinical environments requires a rigorous balance between setup complexity, acquisition modality, and patient fatigue. Medical-grade systems (e.g., g.tec) offer excellent spatial resolution but necessitate skin abrasion and conductive gel, extending setup times beyond 25 minutes and limiting practicality in high-turnover public healthcare settings [6]. Consumer-grade alternatives such as the Emotiv EPOC X enable an agile setup of approximately 3 minutes using felt matrices immersed in saline solution [7], [13]. Prior work with the Emotiv EPOC has exploited its inertial sensors and facial expression streams for wheelchair control [10], while other studies have used OpenBCI for lower-limb robot actuation [8]. However, these systems typically couple processing tightly to a single actuator, lacking the decoupled architecture proposed here. Table I positions this work against contemporary BCI systems.

Table I
COMPARATIVE ANALYSIS OF BCI SYSTEMS FOR REHABILITATION

System / Author	Device / Channels	Setup	Paradigm	Load / Portability	Critical Limitations
López-Larraz *et al.* (2018) [6]	EEG gel (32–64 channels)	$>$ 25 min	Pure Motor Imagery	Critical (Extreme fatigue); Null portability	Skin abrasion and gel impede daily use. Mental fatigue limits therapeutic sessions.
Al-Quraishi *et al.* (2022) [8]	OpenBCI dry (8 channels)	15 min	Multiclass SSVEP	High (Photic fatigue); Low portability	Visual fatigue and frequent calibration. Lower signal quality than gel systems.
Elnady *et al.* (2015) [10]	Emotiv EPOC (14 channels + IMU)	3 min	Head movement (IMU)	Low; High portability	Exclusive dependence on inertial signals; cortical EEG potential unused.
g.tec (g.Nautilus PRO) [11]	Active EEG (8–64 channels)	15–25 min	MI / SSVEP / P300	Variable (High); Medium portability	Prohibitive cost for public health systems. Proprietary software and dedicated infrastructure required.
Pfurtscheller *et al.* (2010) [9]	Laboratory EEG ($>$ 16 channels)	$>$ 20 min	Hybrid BCI (Concept)	Reduced (Theoretical); Null portability	Purely conceptual; no validation on low-cost hardware or clinical protocol.
Proposed Framework	**Emotiv EPOC X + ESP32 (14 EEG + IMU)**	**3 min**	**Hybrid (EMG+IMU) / SSVEP**	**Reduced (by design); High portability**	**Preliminary validation with healthy users (N=3); clinical trials pending.**

III. Methodology and Platform Design

A. Cybernetic and Mechatronic Infrastructure

The global architecture relies on a Wi-Fi hotspot communication structure with logically isolated responsibilities. The software was developed under an asynchronous paradigm in Python 3.11, structured under the Model-View-Controller (MVC) pattern. Instead of relying solely on blocking native threads that disrupt biosignal temporal windows, the `asyncio` framework maintains a single thread of high-speed concurrent coroutines interacting with the Cortex API v3 via WebSockets. Parallel output communication to the ESP32 utilizes a dedicated queue (`queue.Queue()`) managed in a daemon thread.

The decoupling is achieved through a custom local mechatronic module incorporating an ESP32-WROOM-32 microcontroller and a galvanic isolation stage with power relays. The Python software transmits standardized plain text strings (e.g., TURN_ON, `"+"`, `"−"`) via a Wi-Fi hotspot on port 3333, acting directly on the original physical buttons of a DC motor minibike.

B. Architecture 1: Hybrid Mode (EMG + IMU)

This logical flow divides the discrete myoelectric channel and the proportional kinematic channel (Fig. 1). The discretized safety control is executed via the `fac` stream of the Emotiv API, which represents a proprietary facial-expression inference rather than directly measured raw EMG. An intentional jaw clench acts as an alternating toggle switch. During empirical bench testing, the optimal algorithmic sensitivity for this gesture was found to be between 18% and 30%. Below 18%, the system registered false positives caused by involuntary facial movements (e.g., swallowing, phonation). An adaptive threshold of 22% was established to provide the best balance between sensitivity and robustness.

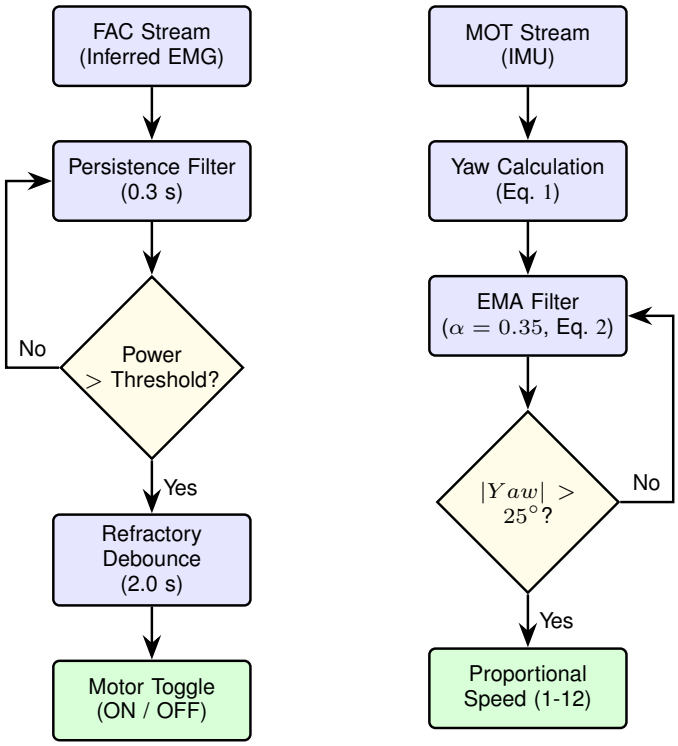


Fig. 1. Operational flowchart of the Hybrid architecture: division of the discrete myoelectric channel and the proportional kinematic channel.

Proportional kinematic control processes the quaternions (Q_0 to Q_3) from the `mot` stream at 32 Hz to derive the neck's yaw angle:

$$Yaw = \text{atan2}\left(2(Q_0Q_3 + Q_1Q_2), 1 - 2(Q_2^2 + Q_3^2)\right) \quad (1)$$

To suppress micro-tremors and spastic spasms, the relative angle passes through an Exponential Moving Average (EMA)

filter:

$$Y_{filtered}[n] = (1 - \alpha) \cdot Y_{filtered}[n-1] + \alpha \cdot Y_{raw}[n] \quad (2)$$

with a smoothing factor of $\alpha = 0.35$. A $\pm 25^\circ$ dead zone was programmed to generate proportional speed commands only when intentional rotations occur.

C. Architecture 2: Visual Mode (Binary SSVEP)

A visual channel was implemented based on the Power Spectral Density in the occipital channels (O_1, O_2). A stimulus frequency of 15 Hz was selected to maximize the cortical Signal-to-Noise Ratio (SNR) in the Low-Beta band. Visual safety regarding lower photoconvulsive risk and fatigue ultimately depends on individual luminance, contrast, and visual-field exposure susceptibility [12].

To eradicate the entrainment effect, a Finite State Machine (FSM) isolates visual rest from the actuation phase. (Fig. 2).

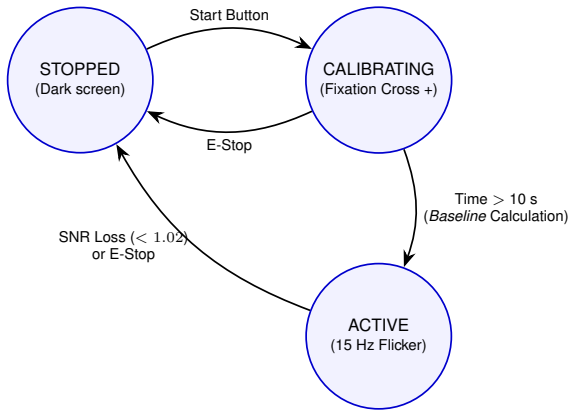


Fig. 2. Finite State Machine (FSM) implemented to ensure a blind spectral calibration, isolating visual rest from the actuation phase.

In the calibration phase (10 s), the GUI displays a dark fixation cross to mathematically register the median cortical white noise as the *Baseline*. During the 15 Hz stimulation, the current SNR is attenuated by an EMA filter ($\alpha = 0.05$). If the SNR exceeds the fixed threshold of 1.20, the minibike advances. A 1.5 s conformation window was programmed, including a release factor of 0.85 (hysteresis zone), to prevent abrupt shutdowns from transient signal drops caused by reflexive blinking. All system parameters are consolidated in Table II.

IV. Technical Validation and Laboratory Results

As handling neurological patients requires extensive ethical evaluation, this research was consolidated as an Instrumentation Proof of Concept (PoC). Validation was conducted through functional laboratory trials with three healthy volunteers ($N = 3$) (Fig. 3).

A. Pilot Study Protocol

Each participant performed a structured protocol consisting of 5 complete on/off cycles (10 targeted activations) in each modality. A 2-minute familiarization period was provided before each session. Performance data (command timestamps, motor state, yaw angle, EMG power, SNR) were logged synchronously using a custom Python datalogger. Classification accuracy was calculated as the ratio of intended activations to total recorded activations (including false positives). The Information Transfer Rate (ITR) was computed using the standard formula with binary selection ($N = 2$), fixed average selection times defined by protocol design ($T = 3.0$ s for the Hybrid mode and $T = 8.0$ s for the SSVEP mode, based on the nominal debounce and confirmation windows of each architecture), and the measured accuracy. Subjective workload was assessed using the NASA Task Load Index (NASA-TLX). Given the reduced sample size ($N = 3$), inferential statistical comparisons (e.g., hypothesis testing) between modalities were not performed; the reported means and standard deviations should be interpreted as preliminary descriptive trends rather than statistically validated differences.

Table II
CONFIGURATION PARAMETERS FOR BCI ARCHITECTURES

Parameter	Description	Value
α_{yaw}	EMA constant for yaw angle	0.35
α_{snr}	EMA constant for SNR tracking	0.05
Dead Zone	Angle dead zone (micro-movements)	$\pm 25^\circ$
Threshold EMG	Inferred jaw clench threshold	0.22 (22%)
Debounce EMG	Refractory time to prevent bouncing	2.0 s
SSVEP Calib.	Blind calibration for *Baseline*	10 s
SSVEP Thresh.	Motor activation threshold via SNR	1.20
Release Factor	Hysteresis factor for deactivation	0.85
SSVEP Conf.	Visual confirmation window	1.5 s

B. System Performance and Information Transfer Rate

Fig. 4 summarizes the classification accuracy and ITR for both control modes across the three subjects. Table III consolidates the individual and average values.

Table III
PERFORMANCE METRICS FROM PILOT STUDY (N=3)

Metric	S1	S2	S3	Mean $\pm$ SD
Hybrid Mode (EMG+IMU)				
Accuracy (%)	100.00	71.43	83.33	84.92 $\pm$ 14.35
False positives	0	4	2	2.00 $\pm$ 2.00
ITR (bits/min)	20.00	2.74	7.00	9.91 $\pm$ 8.99
SSVEP Mode (15 Hz)				
Accuracy (%)	76.92	71.43	83.33	77.23 $\pm$ 5.96
False positives	3	4	2	3.00 $\pm$ 1.00
ITR (bits/min)	1.65	1.03	2.62	1.77 $\pm$ 0.81

The Hybrid mode achieved a mean classification accuracy of 84.92% (SD = 14.35%), with individual accuracies of 100.00%, 71.43%, and 83.33% for Subjects 1, 2, and 3, respectively. In contrast, the SSVEP paradigm presented a mean accuracy of 77.23% (SD = 5.96%), with per-subject values of 76.92%, 71.43%, and 83.33%. The mean ITR was substantially higher in the Hybrid mode (9.91 bits/min, SD = 8.99) compared to the SSVEP mode (1.77 bits/min, SD = 0.81), reflecting the faster nominal actuation window of



Fig. 3. Instrumental functional testing of the PoC in the laboratory. The user asynchronously operates the framework using the sensor headset and the unified GUI.

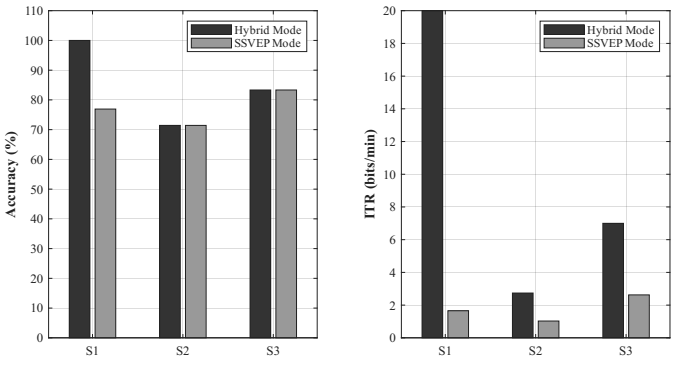


Fig. 4. Comparative performance across three healthy subjects. The Hybrid mode demonstrated superior accuracy and ITR due to the rapid execution of facial gestures compared to sustained visual fixation.

facial gestures ($T = 3.0$ s) relative to sustained visual fixation ($T = 8.0$ s).

False positive activations were observed in both modes. In the Hybrid mode, Subjects 2 and 3 registered 4 and 2 false positives, respectively, while Subject 1 achieved perfect performance. In the SSVEP mode, all subjects exhibited at least 2 false positives, with Subject 2 reaching 4. These false activations occurred predominantly during the initial trials and were attributed to insufficient familiarization with the control paradigm. After verbal clarification and the 2-minute practice period, all subjects were able to operate the system reliably.

C. Cognitive Workload Assessment (NASA-TLX)

Fig. 5 presents the average NASA-TLX scores for the six workload dimensions.

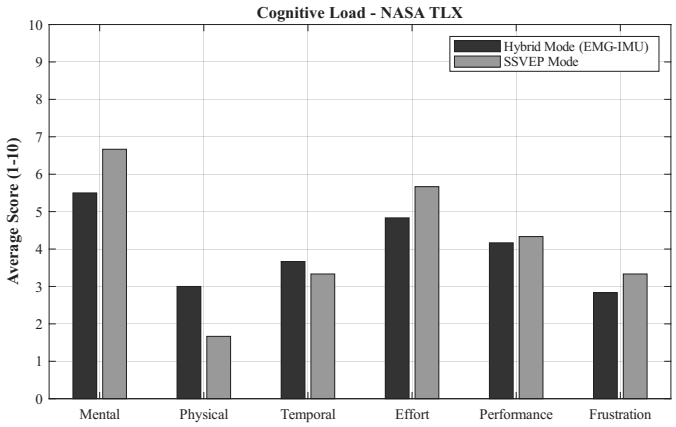


Fig. 5. Cognitive workload comparison (NASA-TLX). The Hybrid architecture shows a trend towards lower mental demand and overall effort.

Results reveal distinct workload profiles for the two control paradigms. The Hybrid mode imposed slightly higher physical demand (mean = 3.00, SD = 1.79) compared to the SSVEP mode (mean = 1.67, SD = 0.58), consistent with the explicit muscular engagement required for jaw clenches and head movements. Conversely, mental demand was higher in the SSVEP mode (mean = 6.67, SD = 1.15) than in the Hybrid mode (mean = 5.50, SD = 2.17), reflecting the sustained attentional focus required by the flickering stimulus. Overall perceived effort followed a similar trend, with means of 4.83 (SD = 1.47) and 5.67 (SD = 1.53) for the Hybrid and SSVEP modes, respectively. Frustration was also slightly higher for SSVEP (mean = 3.33, SD = 2.52) compared to Hybrid (mean = 2.83, SD = 1.94), likely related to the initial false activations and learning curve. Given the small sample size, these workload differences are reported descriptively and should be regarded as preliminary trends rather than statistically confirmed effects; they nonetheless support the rationale for a bimodal architecture that allows selection based on individual patient capabilities.

D. Dynamic Behavior and Signal Filtering

Initial bench tests revealed a high involuntary activation rate due to biosignal bouncing at the threshold boundaries. Detailed

temporal dynamics for both architectures were captured via a native Python datalogger.

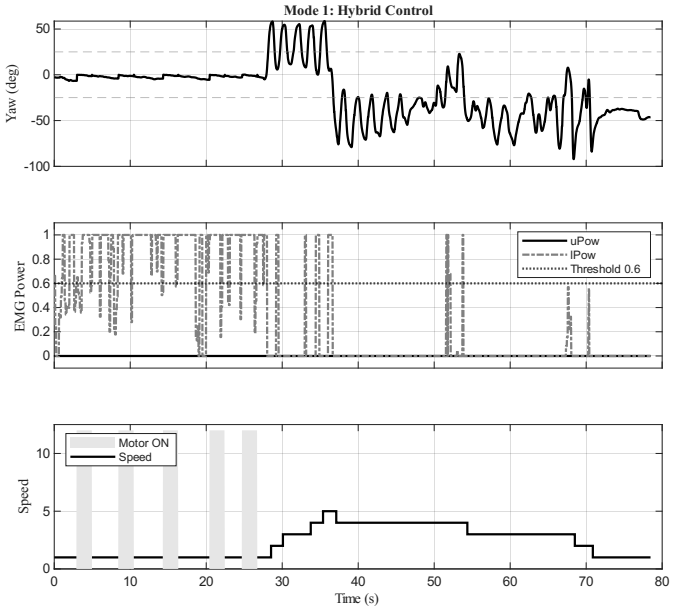


Fig. 6. Dynamic records of the Hybrid Mode. Kinematic Yaw modulates speed, while the discrete facial inference acts as a toggle. The $\pm 25^{\circ}$ dead zone effectively ignores micro-movements.

For the Hybrid mode (Fig. 6), the implementation of the 2.0 s refractory time filter completely stabilized the ESP32 relay response. Natural myoelectric oscillations (uPow/lPow) around the 0.22 threshold were effectively ignored post-activation, suppressing erratic cyclic switchings.

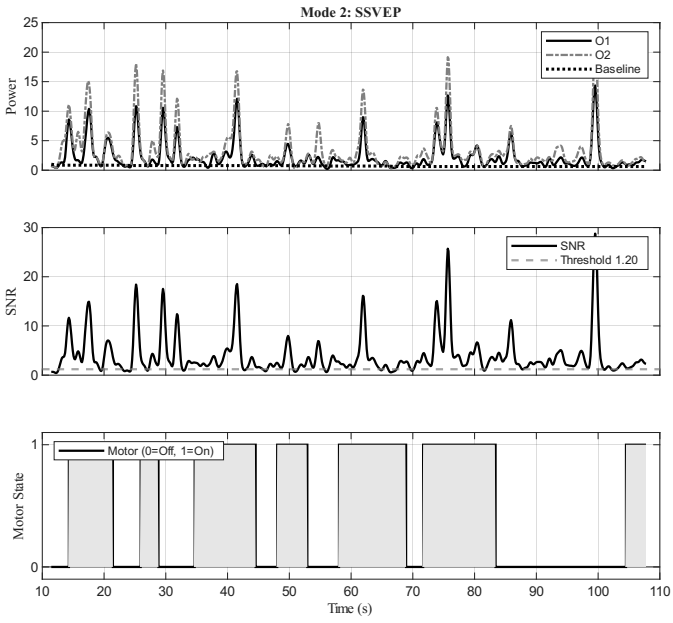


Fig. 7. Dynamic records of the SSVEP Mode. Stable mechatronic pulse generation provided by the sustained crossing of the 1.20 SNR threshold following the 10 s baseline calibration.

For the SSVEP mode (Fig. 7), spectral monitoring validated the stability of the FSM. During the first 10 seconds (blind calibration), the baseline remained immune to background noise. Upon 15 Hz stimulation, the SNR demonstrated monotonic growth, crossing the 1.20 threshold reliably. The 1.5 s conformation window successfully eliminated false commutations caused by reflexive blinks. Furthermore, a Wi-Fi timeout mechanism was implemented on the ESP32 to trigger an automatic motor shutdown within 500 ms upon TCP disconnection, providing a basic fail-safe behavior.

V. Discussion and Limitations

While the modular architecture efficiently decouples cortical processing from actuation, several limitations must be acknowledged. Experimental trials were conducted in a controlled environment with healthy subjects ($N = 3$), which does not account for the neurophysiological variability of post-stroke patients (e.g., hypertonia, spasticity, or cognitive impairments). The EMG threshold (22%) and head yaw dead zone ($\pm 25^{\circ}$) may require recalibration for the target clinical population.

The high standard deviation observed in Hybrid mode accuracy (SD = 14.35%) and ITR (SD = 8.99 bits/min) indicates substantial inter-subject variability, underscoring the need for individualized calibration protocols. During the sessions, Subject 2 was informally observed to require repeated head-rotation attempts before the system registered a deceleration command; this anecdotal observation was not captured with a dedicated metric and is reported here only as a qualitative note suggesting that the dead zone or EMA parameters may need adjustment for users with different movement patterns.

The relative contribution of each individual stabilization mechanism (refractory debounce, EMA smoothing, and SSVEP hysteresis) to overall system performance was not isolated through formal ablation testing; each filter was evaluated only as part of the complete pipeline. Quantifying these individual contributions remains an open direction for future work.

Technologically, the use of the Emotiv EPOC X imposes restrictions due to the unavailability of raw EEG signals through its standard API, limiting the implementation of advanced deep learning algorithms for raw feature extraction. The `fac` stream, while correlated with facial muscle activity, is a proprietary inference that may introduce algorithmic dependencies not transparent to the end user. Furthermore, while the reuse of TCP commands successfully drives a laboratory minibike, immediate compatibility cannot be guaranteed for complex physical therapy machines such as exoskeletons or treadmills without rigorous adaptation of electrical interfaces, specific functional safety protocols, and strict compliance with medical device regulatory requirements (e.g., IEC 60601).

Although a Wi-Fi timeout mechanism was implemented, its behavior under intermittent connectivity or electromagnetic interference has not been formally tested. Future work should include controlled fault-injection tests to evaluate system robustness. Similarly, the SSVEP mode's safety depends on

stimulus parameters (luminance, contrast, flashing rate) and individual susceptibility to photosensitivity; the fixed 15 Hz frequency and 1.20 SNR threshold should be validated in a clinical setting.

VI. Conclusion

This paper presented a functional Proof of Concept for a modular BCI framework designed to control a rehabilitation minibike. The central contribution lies in the decoupled bimodal architecture, standardized network commands, and mathematical filters (EMA and refractory debounce) that successfully improved operational stability by mitigating the biosignal bouncing effect. A pilot study with three healthy volunteers ($N = 3$) demonstrated that the Hybrid mode achieved higher mean accuracy (84.92%) and ITR (9.91 bits/min) compared to the SSVEP mode (77.23% and 1.77 bits/min, respectively). NASA-TLX assessments suggested a trend towards lower mental workload for the Hybrid mode, at the cost of slightly higher physical demand. These results, while preliminary and based on a small, non-clinical sample, support the rationale for offering both control modalities to accommodate diverse patient capabilities.

Key *lessons learned* highlight the critical necessity of physical fail-safes and refractory filters to transform erratic biosignals into reliable robotic commands. Significant *research gaps* remain open for future development, including the need for controlled clinical trials with post-stroke patients at CREFES, formal System Usability Scale (SUS) assessments, systematic comparison of control modes with larger cohorts and inferential statistical analysis, ablation studies isolating the contribution of each stabilization filter, robustness testing under fault conditions, and the potential integration of machine learning for automatic personalization of dynamic control parameters.

References

[1] Brazilian Ministry of Health, "O AVC no Brasil: Dia Mundial do Acidente Vascular Cerebral," Virtual Health Library (BVS), Brasília, DF, 2022.

[2] J. Mehrholz, M. Pohl, T. Platz, J. Kugler, and B. Elsner, "Electromechanical and robot-assisted arm training for improving activities of daily living, arm function, and arm muscle strength after stroke," *Cochrane Database Syst. Rev.*, no. 9, p. CD006876, Sep. 2018.

[3] M. Ahn and S. C. Jun, "Performance variation in motor imagery brain-computer interface: A brief review," *J. Neurosci. Methods*, vol. 243, pp. 103–110, 2015.

[4] G. Borghini *et al.*, "Measuring cognitive and mental fatigue through EEG signals: an application in aviation," *Journal of Neural Engineering*, vol. 11, p. 056008, 2014.

[5] S. Becker, K. Dhindsa, L. Mousapour, and Y. Al Dabagh, "BCI illiteracy: It's us, not them. Optimizing BCIs for individual brains," in *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, Gangwon-do, Korea (South), 2022, pp. 1–3.

[6] E. López-Larraz *et al.*, "Brain-machine interfaces for rehabilitation in stroke: A review," *NeuroRehabilitation*, vol. 43, no. 1, pp. 77–97, 2018.

[7] T. Radüntz, "Signal quality evaluation of emerging EEG devices," *Front. Hum. Neurosci.*, vol. 12, p. 43, Feb. 2018.

[8] M. S. Al-Quraishi *et al.*, "EEG-based control of a lower limb rehabilitation robot using OpenBCI platform," *Journal of Medical and Biological Engineering*, vol. 42, pp. 312–325, 2022.

[9] G. Pfurtscheller *et al.*, "The hybrid BCI," *Frontiers in Neuroscience*, vol. 4, p. 3, 2010.

[10] A. M. Elnady *et al.*, "Control of a Custom-Built Wheelchair Using Emotiv EPOC BCI," in *IEEE International Conference on Autonomous Robot Systems and Competitions*, 2015, pp. 1–6.

[11] g.tec medical engineering GmbH, "g.tec - Brain-Computer Interfaces and Neurotechnologies," [Online]. Available: https://www.gtec.at/

[12] D. Zhu, J. Bieger, G. Garcia Molina, and R. M. Aarts, "A Survey of Stimulation Methods Used in SSVEP-Based BCIs," *Computational Intelligence and Neuroscience*, vol. 2010, Art. ID 702357, 2010.

[13] K. M. Gutiérrez and A. S. Pomaquiza, "Sistema para rehabilitación robótica controlado por Emotiv EPOC," Trabajo de Titulación, Universidad Politécnica Salesiana, Cuenca, Ecuador, 2025.

[14] Emotiv Inc., "Emotiv EPOC X 14-Channel Wireless EEG Headset," [Online]. Available: https://www.emotiv.com/epocx/

Control of Functional Electrical Stimulation for Post-Stroke Rehabilitation Through a Brain–Computer Interface Based on Motor Imagery

Angelica Maria Rojas-Tocora
Postgraduate Program in Electrical Engineering
Federal University of Espírito Santo
(UFES)
Vitória, Brazil
angelica.tocora@edu.ufes.br

Luis Carlos Rodríguez-Timaná
Postgraduate Program in Electrical Engineering
Universidad Santiago de Cali
(USC)
Federal University of Espírito Santo
(UFES)
Vitória, Brazil
luis.timana@edu.ufes.br

Fernanda Vaz de Souza
Postgraduate Program in Electrical Engineering
Federal University of Espírito Santo
(UFES)
Vitória, Brazil
fernandavaz.souza@hotmail.com

Eliete Maria de Oliveira Caldeira
Electrical Engineering Department
Federal University of Espírito Santo
(UFES)
Vitória, Brazil
eliete.caldeira@ufes.br

Teodiano Freire Bastos-Filho
Postgraduate Program in Electrical Engineering
Federal University of Espírito Santo
(UFES)
Vitória, Brazil
teodiano.bastos@ufes.br

***Abstract*—Stroke frequently impairs upper-limb motor function and reduces the functional independence of affected individuals. This paper presents a proposal for the development of a motor imagery (MI)-based Brain–Computer Interface (BCI) to control a Functional Electrical Stimulation (FES). This device will be used as a support tool for physiotherapists during post-stroke motor rehabilitation sessions. The proposed BCI includes a Graphical User Interface (GUI) for patient registration, session configuration, motor-task presentation, electroencephalography (EEG)-signal acquisition and processing, MI detection, safe FES activation, and automatic recording of technical metrics. The initial evaluation was limited to functional validation of the computational workflow in a controlled environment over four weeks, without aiming to demonstrate clinical therapeutic efficacy. The application is expected to help standardize sessions, support physiotherapist decision-making, and provide an experimental basis for future studies in neurofunctional rehabilitation.**

***Index Terms*—Stroke, brain–computer interface, electroencephalography, functional electrical stimulation, motor imagery, motor rehabilitation.**

I. Introduction

Stroke remains one of the leading causes of disability worldwide, with direct effects on autonomy, social participation, and quality of life [1], [2]. One of its most limiting sequelae is upper-limb motor impairment, often manifested as hemiparesis, loss of strength, reduced motor selectivity, tone alterations, and difficulty performing activities of daily living (ADLs) [3], [4]. Recovery of hand, wrist, and finger movement is particularly challenging because it depends on sensorimotor reorganization and repetitive practice of functional tasks.

In the State of Espírito Santo (Brazil), the Physical Rehabilitation Center of Espírito Santo (CREFES), located in Vila Velha-ES, is a reference service for physical rehabilitation. It provides multiprofessional care, including physiotherapy, occupational therapy, physiatry, neurology, and clinical support areas [5]. Therefore, technologies designed for professional use may help structure rehabilitation sessions, record performance, and support future clinical assessments.

Functional Electrical Stimulation (FES) uses controlled electrical pulses to activate peripheral muscles and produce functional movements. In the upper limbs of post-stroke patients, Functional Electrical Stimulation may target wrist and finger extensors or flexors, assisting movements such as hand opening, hand closing, and grasping [6], [7]. Clinical guidelines recognize Functional Electrical Stimulation as an adjunct resource for motor recovery, especially when combined with task-oriented training [4]. However, when applied alone, Functional Electrical Stimulation is not necessarily synchronized with the patient's voluntary motor intention or motor imagery (MI). This limitation motivates the integration of Functional Electrical Stimulation with a Brain–Computer Interface (BCI).

Brain–Computer Interface systems transform central nervous system signals into commands for external devices [8].

In noninvasive Brain–Computer Interface systems, electroencephalography (EEG) is widely used because of its safety, portability, and high temporal resolution [9]. During MI, sensorimotor rhythms in the mu and beta bands may show event-related desynchronization (ERD) and synchronization (ERS) [10], [11]. These modulations provide physiological information that can be used to infer motor intention and activate an assistive device.

The integration of Brain–Computer Interface and Functional Electrical Stimulation is particularly relevant because it creates a closed feedback loop between cortical intention, peripheral motor assistance, and sensory feedback. In this paradigm, the user attempts or imagines a movement, the EEG signal is processed, a motor-intention criterion is evaluated, and Functional Electrical Stimulation is activated only when the decision condition is satisfied. Clinical studies and reviews indicate that Brain–Computer Interface systems associated with Functional Electrical Stimulation or assistive devices have potential for post-stroke upper-limb rehabilitation [12]–[18]. Nevertheless, the literature highlights persistent challenges related to standardization, usability, safety, EEG signal variability, and transfer to real clinical environments [19]–[21].

Beyond the therapeutic component, robust MI decoding remains a central technical challenge. Methods such as Common Spatial Pattern (CSP), Filter Bank Common Spatial Pattern (FBCSP), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Riemannian Geometry (RG), and autoencoder-based techniques have been applied to improve MI task classification [9], [23]–[26]. These approaches differ in complexity, interpretability, calibration requirements, and amount of data needed for reliable training. Classical pipelines such as CSP or FBCSP combined with LDA or SVM are relatively interpretable and have been widely used in MI-Brain–Computer Interface studies. Riemannian Geometry (RG) methods exploit the covariance structure of multichannel EEG, whereas autoencoder-based approaches may improve feature learning but usually require larger datasets and stronger validation.

In the present work, the initial motor-intention detection strategy is not based on a full machine-learning classifier. Instead, it adapts a threshold-based MI detection algorithm previously proposed for upper-limb MI-Brain–Computer Interface control [38]. This algorithm uses C3/C4 sensorimotor rhythms, relative mu (8-13Hz) and high-beta (20-38Hz) band power, a stabilized lateralization index, baseline-derived thresholds, hysteresis, and finite-state-machine logic. In this study, the algorithm is adapted to authorize Functional Electrical Stimulation activation during functional validation, prioritizing interpretability, safety, and workflow verification before implementing more complex classifier-based approaches.

Recent studies reinforce the relevance of MI-Brain–Computer Interface approaches for post-stroke rehabilitation, but they also show important methodological differences. Some protocols combine MI with neurofeedback, Functional Electrical Stimulation, robotic gloves, or assistive devices to provide contingent sensory and motor feedback, whereas others focus mainly on EEG decoding and clinical outcome assessment [27]–[31]. These differences are relevant because the therapeutic effect of Brain–Computer Interface-based rehabilitation depends not only on the detection method, but also on the temporal coupling between motor intention, feedback, and assisted movement.

For upper-limb MI, most protocols focus on sensorimotor rhythms in the mu and beta bands because these frequency components are directly associated with ERD/ERS during movement imagination [32]. Other EEG frequency ranges may provide complementary information, but they are less specific to sensorimotor modulation and may be more affected by attention, visual processing, or artifacts. Therefore, mu and beta activity remain a practical and physiologically grounded choice for detecting hand opening and hand closing MI.

Another relevant distinction concerns the type of motor imagery task. Kinesthetic MI requires the participant to imagine the sensation of performing the movement, while visual or action-observation strategies emphasize external observation or visual representation of the task. Kinesthetic MI is especially relevant for rehabilitation because it is more closely related to sensorimotor activation, whereas visual strategies may facilitate task understanding, attention, and protocol adherence [33]–[35].

Thus, the present proposal contributes to the advancement of Brain–Computer Interface-, Functional Electrical Stimulation-, and assistive-device-based technologies for neurorehabilitation by focusing on a practical and configurable system oriented toward physiotherapist-supervised upper-limb motor rehabilitation sessions. The initial validation emphasizes functional integration, motor-intention detection, safe Functional Electrical Stimulation activation, and session recording rather than clinical therapeutic efficacy.

II. Methodology

The methodological approach is conducted as technological development with preliminary functional validation. The initial stage does not aim to demonstrate clinical efficacy, but rather to verify the computational workflow required for a MI-based BCI-FES application intended for supervised use by physiotherapists. For this preliminary validation, a pilot test with ten healthy subjects aged between 20 and 30 years is performed to verify protocol execution, signal acquisition, computational processing, FES activation, and recording of the information generated by the system. Participants are informed about the nature of the test and sign an informed consent form.

A. Experimental Setup

The experimental setup integrates EEG acquisition, protocol execution, stimulation command generation, and session supervision. EEG is acquired using an OpenBCI system with a Daisy expansion module, allowing 16-channel acquisition. The EEG stream is received through the OpenViBE Acquisition Server and processed in OpenViBE Designer, which is used to execute the experimental scenario, synchronize protocol events, and manage online signal processing. The

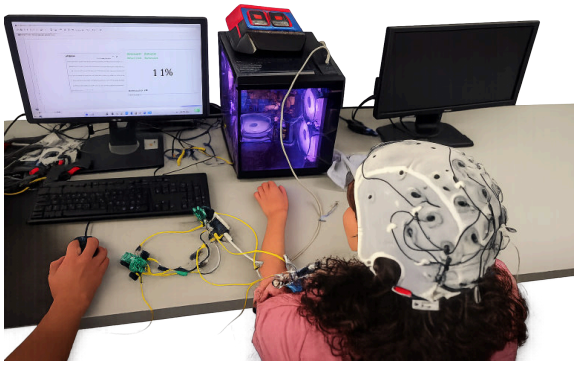


Fig. 1. Representative setup used during the preliminary pilot test. The system includes the OpenBCI EEG cap, the acquisition and interface computer, OpenViBE monitoring, the ESP32 command bridge, the FightArt FES device, and FES electrodes positioned on the forearm for hand-opening and hand-closing stimulation.

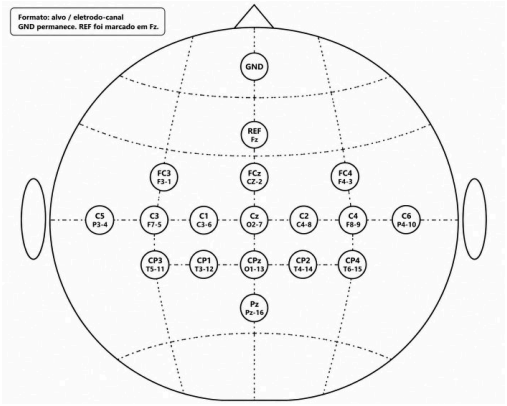


Fig. 2. EEG electrode montage used in the MI-based BCI-FES protocol. The configuration uses 16 active channels focused on fronto-central, central, centro-parietal, and parietal regions. The reference electrode was marked at Fz, and the ground electrode was maintained according to the OpenBCI cap configuration.

stimulation pathway includes an ESP32 microcontroller used as a digital command bridge and a FightArt FES device used for stimulation. The preliminary functional tests are performed by the development team to verify protocol execution, EEG acquisition, signal processing, FES command generation, safety behavior, and data logging. In addition, the interface design receives clinical-usability input from the physiotherapist co-author regarding configurable stimulation channels, activation duration, current intensity, safety controls, and session monitoring.

The system is operated through a clinical GUI developed for physiotherapist supervision. This interface allows participant registration, session configuration, FES authorization, safety monitoring, protocol execution, and event logging. The ESP32 receives serial commands from the interface and translates them into pulse trains associated with the hand-opening and hand-closing commands of the FES device. Therefore, OpenViBE remains responsible for EEG acquisition and online processing, while the external interface acts as the clinical control and safety layer.

The physical FES command bridge uses the ESP32 as an interface to the control inputs of two FightArt FES units. The first command pair is assigned to GPIO25 and GPIO26, and the second command pair is assigned to GPIO32 and GPIO33. These outputs are used to emulate brief button activations according to the firmware commands, while the stimulation output stage remains managed by the FES equipment. During non-motor phases or safety blocks, the firmware releases the GPIO pins to avoid sustained activation.

A preliminary pilot setup is used to verify the complete acquisition-control-stimulation workflow. During this test, the participant remains seated in front of the acquisition computer while wearing the OpenBCI EEG cap, as shown in Fig 1. FES electrodes are positioned on the forearm according to the stimulation targets, and the ESP32 is connected between the software interface and the FES device. The clinical interface and OpenViBE monitoring window are used simultaneously to observe the protocol state, EEG activity, threshold behavior, serial communication, and stimulation authorization.

B. EEG Montage and FES Preparation

The EEG montage is defined to prioritize cortical regions involved in upper-limb MI, especially the bilateral sensorimotor cortex. Sixteen active channels are positioned according to the 10–20 system, as shown in Fig. 2. This configuration covers fronto-central, central, centro-parietal, and parietal regions, with emphasis on C3, Cz, and C4 due to their relevance for hand MI. The acquisition follows a common-reference configuration, with the reference electrode at Fz and the ground electrode maintained according to the OpenBCI cap configuration.

FES electrodes are positioned over the main forearm muscle groups responsible for hand opening and hand closing. For hand closing, electrodes are placed over the anterior forearm region, prioritizing the wrist and finger flexor muscle mass. For hand opening, electrodes are placed over the posterior forearm region, prioritizing the wrist and finger extensor muscle mass. This placement follows the principle that stimulation electrodes should be positioned over the target muscle belly or motor point region to elicit the desired functional movement [36], [37].

Before starting the BCI-FES protocol, an individual FES validation procedure is performed with each participant under physiotherapist supervision. The stimulation intensity is initially set to the minimum available level and gradually increased to identify the lowest intensity capable of producing visible functional hand opening or closing without pain. If the participant reports pain, excessive discomfort, or an inadequate motor response, the electrode position is adjusted and the intensity is reduced. Thus, the selected stimulation level is individualized for each participant and movement condition before online activation is allowed.

During the experimental protocol, FES activation is restricted to valid motor events, namely hand opening and hand

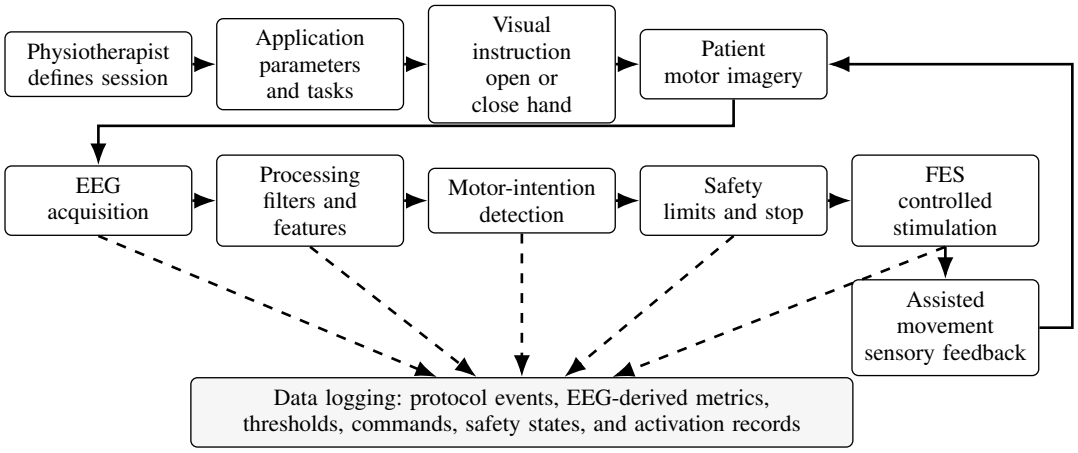


Fig. 3. Functional architecture of the MI-based BCI-FES application for upper-limb motor rehabilitation support. Solid arrows indicate the control loop, and dashed arrows indicate session data logging.

closing. During baseline, rest, idle, start-of-block, and end-of-session events, all stimulation outputs remain disabled. Therefore, the FES device is not controlled directly by the EEG marker alone; stimulation depends on the current protocol event, the baseline-derived threshold, the physiotherapist's authorization, the safety state, and the previously calibrated intensity level.

C. Solution Architecture

The BCI development is organized into five modules: clinical interface, BCI processing, FES control, operational safety, and data storage. Fig. 3 presents the proposed workflow. The physiotherapist configures the session in the GUI, the patient receives a visual instruction, EEG is acquired during MI, the signal is processed, motor intention is detected, safety rules are checked, and FES is activated only when the activation criterion is met. The assisted movement provides sensorimotor feedback, while the application records session events and metrics.

Additionally, key security measures were implemented, such as the requirement for professional authorization, individual calibration for each user, and an emergency stop button to immediately halt the interface.

D. Session Workflow

The session begins with patient registration or selection in the case they are already registered. Next, the physiotherapist defines the motor task, such as hand opening or closing, and configures acquisition and stimulation parameters. The GUI presents visual instructions to the patient and records time windows associated with MI. The EEG signal is filtered, segmented, and processed for feature extraction. If the EEG-derived control variable exceeds the decision threshold, the interface sends a command to the FES module. All events are recorded for later analysis.

E. Experimental Protocol Design

The experimental protocol is designed to standardize the preliminary functional validation of the MI-based BCI-FES system and to synchronize visual instructions, EEG acquisition, decision logic, FES command generation, and data logging. As shown in Fig. 4, the protocol starts with a 10 s welcome and instruction period, followed by a 120 s baseline period with all output pins off. This baseline is used to calculate the reference threshold for later activation decisions. A 3 s start-of-block event then indicates the beginning of the experimental task sequence.

The main task cycle is repeated four times and follows the sequence: rest, close hand, rest, idle, rest, open hand, and rest. Each rest event lasts 3 s and is used to turn off all outputs and reset the system state. The close-hand and open-hand events last 10 s each; during these periods, only the pins associated with the current task can be activated, and activation is allowed only when the current signal exceeds the baseline-derived threshold. The idle event lasts 20 s and does not allow pin activation.

With four repetitions, the repeated task block lasts 208 s. Considering the welcome and instruction period, the baseline, and the start-of-block event, the complete protocol has a total duration of 341 s.

Pin activation follows a strict priority rule. First, the current protocol event has to allow stimulation. Second, the processed signal has to exceed the threshold calculated during the baseline. Third, the system applies the intensity and speed values configured by the physiotherapist. Finally, the corresponding pins are activated according to the current task. Therefore, during welcome, baseline, start-of-block, rest, idle, and end-of-session events, all pins remain off regardless of the signal value. During close-hand and open-hand events, activation is restricted to the pins associated with the current task.

F. EEG Processing and Motor-Intention Detection

The motor-intention detection algorithm is adapted from a previous upper-limb MI-BCI protocol based on C3/C4 sensorimotor rhythms [38]. In the present system, EEG signals are processed in sliding windows with overlap. For each window,

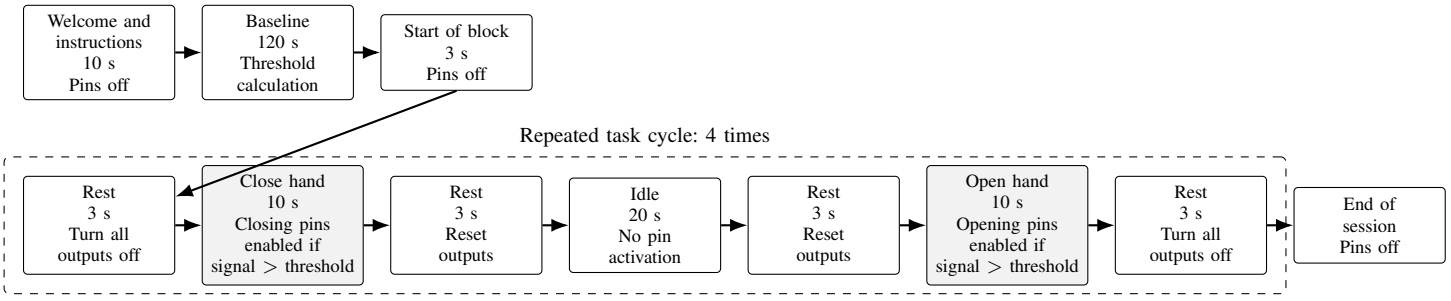


Fig. 4. Temporal flow of the experimental protocol used for preliminary functional validation of the MI-based BCI-FES system. The protocol includes fixed initial events, a baseline period for threshold calculation, and a repeated task cycle composed of rest, hand closing, idle, and hand opening events.

the algorithm estimates the relative power of the mu band (8–12 Hz) and the high-beta band (18–24 Hz), normalized by the broader sensorimotor band from 8 to 30 Hz.

The processing focuses on sensorimotor activity around C3 and C4 because these regions are associated with upper-limb motor imagery. The relative mu and high-beta components are combined into C3- and C4-centered sensorimotor indices. Then, a lateralization index is computed from the difference between the right and left sensorimotor indices, normalized by their sum. This index provides a bounded control variable related to motor-intention modulation.

To improve online stability, the control variable is stabilized using a running median and exponential smoothing. During the baseline period, the system learns an individualized decision threshold from the percentile distribution of the baseline control variable. The activation threshold is bounded by minimum and maximum safety limits to avoid excessively sensitive or insensitive behavior.

FES activation is not triggered by a single threshold crossing. The system applies hysteresis and finite-state-machine logic. The ON threshold must be exceeded for a predefined number of consecutive windows before stimulation is authorized. The OFF threshold, minimum ON/OFF dwell times, sticky-hold behavior, and inactivity failsafe reduce unstable switching and prevent accidental or sustained activation.

Therefore, stimulation is generated only when four conditions are satisfied: the current protocol phase allows motor activation, the EEG-derived control variable exceeds the decision criterion, the physiotherapist authorization is active, and no safety lockout is present. This strategy is used as a preliminary threshold-based MI detection approach for functional validation of the BCI-FES workflow. Future versions should compare this interpretable strategy with classifier-based approaches such as CSP/FBCSP, LDA, SVM, and Riemannian Geometry.

G. Experimental Data Organization and Statistical Analysis

The experimental files are organized by participant, with one session per participant. The offline EEG database contains the protocol cycle, task phase, hand condition, EEG-derived features, threshold values, threshold-crossing flags, and temporal information. Participant identifiers are inferred from the session folders, hand conditions from the hand labels, and trial fields are used only to identify protocol cycles. Therefore, cycles and internal processing windows are not treated as independent statistical samples.

For the open-versus-closed comparison, each EEG feature is first summarized by participant and condition, using the mean as the default summary statistic. The analysis is paired, with the effective sample size defined as the number of participants with valid data in both conditions. The Wilcoxon signed-rank test is used as the principal nonparametric test, complemented by a paired permutation test and an additional paired t-test. Effect sizes include Cohen's dz for the paired t-test and rank-biserial correlation for the Wilcoxon test. When multiple EEG features are analyzed, Benjamini–Hochberg false discovery rate (FDR) correction is applied to the Wilcoxon p values.

The neurofeedback threshold is calculated individually from the baseline period. During baseline, the absolute MI score is stabilized with a short median window, and the participant threshold is defined as the 90th percentile of the baseline metric, clipped to the range from 5 to 30 units. Threshold-based metrics are then summarized at the participant level before the paired comparison between hand opening and hand closing.

III. Results

A. Functional Interface Validation

The first result of the preliminary validation was the implementation of an integrated GUI for configuring, executing, and monitoring MI-based BCI-FES sessions. The GUI integrated participant registration, session configuration, protocol timing, OpenViBE control, FES authorization and safety settings, serial communication, execution controls, live monitoring, and event logging. Clinical-usability input from the physiotherapist co-author indicated that the developed interface may support supervised BCI-FES rehabilitation by allowing individualized configuration of active stimulation channels, activation duration, and current intensity according to the therapeutic goals and clinical characteristics of each participant. From a clinical perspective, the main advantage of the system is that FES activation is not delivered as a passive stimulation routine, but is conditioned on EEG-derived evidence of motor intention, physiotherapist authorization, and safety rules. This organization may promote a closer association between voluntary

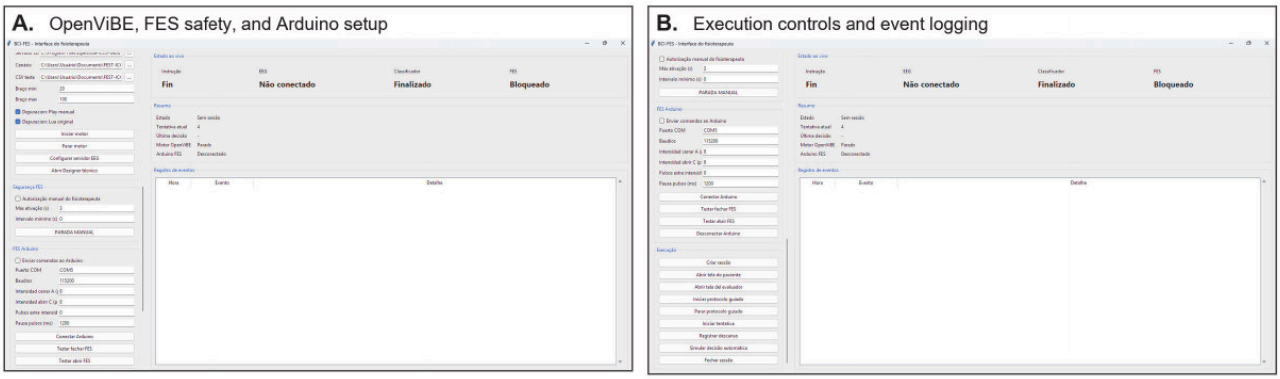


Fig. 5. Technical control areas of the physiotherapist interface.

cortical activation, peripheral muscle stimulation, and sensorimotor feedback, which is relevant for activity-dependent neuroplasticity. The physiotherapist also highlighted that the interface may support clinical decision-making because it enables personalized and reproducible protocols without hardware modifications and integrates EEG acquisition, motor-intention processing, FES control, safety supervision, and session recording in a single platform. This input is interpreted as a qualitative assessment of clinical usability and workflow organization.

Fig. 5 details the technical controls used during session execution. Panel A groups OpenViBE configuration, FES safety parameters, and Arduino serial communication. Panel B shows manual safety controls, FES testing commands, execution buttons, live state, and the event log. These controls support traceability and supervised operation of the MI-based BCI-FES workflow.

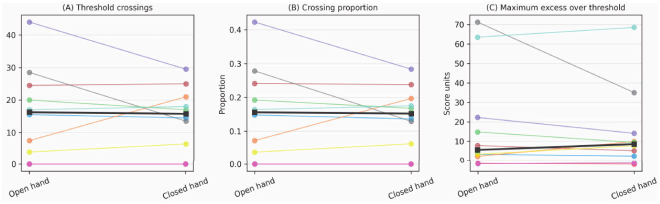


Fig. 6. Paired threshold-based metrics summarized at the participant level.

B. Exploratory Statistical Analysis

The statistical analysis used the experimental CSV files from ten anonymized participants, labeled Subject 1 through Subject 10. The inspected EEG databases contained 4,742,750 rows in total. Open- and closed-hand task labels accounted for 1,038,875 rows. After reducing repeated samples to unique processing windows, 8,311 windows were retained for feature extraction, but the final paired comparisons were performed at the participant level. Thus, the effective statistical sample size was $n = 10$ participants with valid data in both conditions.

Table I summarizes the paired open-versus-closed comparison for the channels with the greatest exploratory interest using the real electrode mapping. No channel survived Benjamini–Hochberg FDR correction. The smallest uncorrected Wilcoxon values appeared in C4 and CP3 ($p = 0.049$), but their corrected values were $p_{\mathrm{FDR}} = 0.371$. Permutation tests showed low exploratory values for C4, CP3, CPz, and Pz, but these findings remain exploratory because the study was limited to ten participants. The remaining channels did not show statistically supported differences after FDR correction.

Derived non-channel EEG metrics from the CSV included `eeg_value`, `li_median`, `li_signed`, `li_strength`, and `rmi`; these were internal EEG-derived values, not electrode locations. None of these derived features remained significant after FDR correction. For the main controller metric, `eeg_value`, the open-hand mean was 8.91 and the closed-hand mean was 8.15, with a Wilcoxon value of $p = 0.193$ and an FDR-corrected value of $p_{\mathrm{FDR}} = 0.404$.

TABLE I
SELECTED PAIRED OPEN-VERSUS-CLOSED EEG CHANNEL COMPARISONS

Channel	Electrode	Open	Closed	Δ	Wilc.	FDR	Perm.
Ch. 9	C4	-0.015	0.050	-0.065	**0.049**	0.371	0.027
Ch. 11	CP3	-0.009	0.051	-0.060	**0.049**	0.371	0.047
Ch. 13	CPz	-0.013	0.045	-0.058	0.064	0.371	0.047
Ch. 16	Pz	-0.014	0.046	-0.060	0.064	0.371	0.043

All comparisons used $n = 10$ participants. Values were computed after summarizing each participant by condition; cycles and internal windows were not treated as independent samples. Δ indicates the open-hand minus closed-hand mean difference. Wilc., FDR, and Perm. report the Wilcoxon, FDR-corrected, and permutation p values, respectively. Bold values indicate the lowest uncorrected exploratory Wilcoxon p values; no comparison survived FDR correction.

Table II reports the baseline-derived threshold for each participant. The thresholds ranged from 5.00 to 30.00 units. Subjects 5 and 8 reached the lower clamp, whereas Subject 1 reached the upper clamp. These thresholds should be interpreted as technical calibration values for the neurofeedback controller, not as clinical outcome measures.

TABLE II
BASELINE THRESHOLD LEARNED FOR EACH PARTICIPANT

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Threshold	**30.00**	10.42	5.07	14.65	**5.00**	7.19	6.40	**5.00**	28.70	19.47

Threshold values are expressed in EEG-derived control-score units. Bold values indicate thresholds clipped by the predefined lower or upper safety limits.

Table III summarizes the threshold-based paired comparison between hand opening and hand closing. The median number of threshold crossings was similar between open hand and closed hand (16.25 versus 15.75 windows), with a median open-minus-closed difference of 0.00. The crossing proportion was also similar between conditions (0.156 versus 0.152), and the Wilcoxon signed-rank test did not show a statistically supported paired difference before or after FDR correction.

The first-crossing latency was available for eight valid paired participants and was numerically longer for open hand than for closed hand, but this difference also did not survive correction.

TABLE III
EXPLORATORY PAIRED COMPARISON OF THRESHOLD-BASED METRICS

Metric	n	Open median [IQR]	Closed median [IQR]	Δ	p	FDR
Threshold crossings	10	16.25 [18.50]	15.75 [12.00]	0.00	1.00	1.00
Crossing proportion	10	0.16 [0.182]	0.15 [0.11]	0.00	0.48	0.78
Any crossing	10	1.00 [0.00]	1.00 [0.00]	0.00	1.00	1.00
Maximum value	10	18.47 [20.05]	20.80 [17.97]	0.64	0.56	0.78
Maximum excess over threshold	10	5.54 [17.93]	8.53 [10.02]	0.64	0.56	0.78
Windows over threshold	10	16.25 [18.50]	15.75 [12.00]	0.00	0.53	0.78
First-crossing latency (s)	8	1.73 [0.76]	1.30 [1.26]	0.58	**0.15**	0.78

Values are summarized at the participant level. Δ indicates the open-hand minus closed-hand median difference. Tests were Exact sign for threshold crossings, Exact McNemar for any crossing, and Wilcoxon for the remaining metrics. FDR correction was applied across the principal open-versus-closed tests. The bold value indicates the lowest uncorrected exploratory p value; no metric survived FDR correction.

Fig. 6 visualizes the paired threshold-based metrics, with connected lines emphasizing the within-subject comparison. Overall, the workflow registered EEG-derived features, baseline thresholds, protocol events, and paired open-versus-closed comparisons, but no principal comparison showed a statistically robust difference in this preliminary dataset.

IV. DISCUSSION

The primary objective of this study was to verify the functional viability of an integrated MI-based BCI-FES workflow operating in a supervised environment: combining EEG acquisition, visual task presentation, OpenViBE signal processing, threshold-based event detection, ESP32 communication, controlled FES authorization, and session logging into a dedicated interface for healthcare professionals. Reliable system integration across this acquisition–processing–stimulation chain is an essential prerequisite prior to clinical testing [19]–[21]. Crucially, the findings of this study validate the functional operation and safety of this integrated pipeline, but they do not demonstrate successful EEG decoding or class discrimination of motor imagery (MI).

The statistical analysis revealed no robust differences between hand-opening and hand-closing conditions. No EEG channel or derived metric survived False Discovery Rate (FDR) correction; thus, uncorrected trends observed in C4, CP3, CPz, and Pz must be regarded strictly as exploratory observations. Utilizing participant-level analysis avoided pseudoreplication and provided a conservative interpretation of this feasibility dataset. High inter-subject variability is well-documented in EEG-based MI paradigms [9]–[11], and these findings align with broader BCI-FES literature for post-stroke upper-limb rehabilitation, where protocols, populations, and detection outcomes remain highly heterogeneous [12]–[18].

A primary limitation of the current algorithm is the threshold-based activation strategy itself. Adapted from the sensorimotor rhythm processing pipeline described by Rodríguez-Ugarte et al. [38]—utilizing C3/C4 relative mu and high-beta power, lateralization indices, baseline-derived thresholds, hysteresis, and finite-state logic—this strategy was intentionally simplified to authorize FES activation rather than robotic actuation. However, because hand-opening and hand-closing tasks yielded nearly identical neurophysiological responses, this strategy cannot and must not be interpreted as a functional MI classifier. Instead, it serves solely as a baseline-referenced, safety-contingent gating mechanism to ensure stimulation is authorized only during intended task windows. Future developments must replace this preliminary logic with dedicated machine learning approaches—such as Common Spatial Patterns (CSP/FBCSP) combined with Linear Discriminant Analysis (LDA) or Support Vector Machines (SVM), or Riemannian geometry frameworks—built upon larger datasets with multi-session calibration.

Furthermore, severe sample limitations constrain the scope of these results. With data collected from only ten healthy participants across a single session and a small number of repetitions, no claims can be made regarding system generalizability or clinical efficacy in post-stroke populations. This stage was strictly aimed at technical and functional debugging. Before engaging individuals post-stroke, future research must incorporate multi-session protocols, expanded trial counts, systematic stimulation tolerance assessments, adaptive calibration, and direct comparisons between simple threshold gating and supervised classification. Controlled functional testing in healthy cohorts remains an obligatory stepping stone before clinical evaluation.

V. CONCLUSIONS

This work presented the preliminary engineering development and functional proof-of-concept for an MI-based BCI-FES system tailored for upper-limb neurorehabilitation. The framework successfully demonstrated that EEG acquisition, task presentation, signal processing, therapist GUI control, wireless communication, and logging can run as a synchronized system under supervised conditions.

Importantly, participants with lower baseline activation thresholds did not exhibit superior control. A low baseline threshold often reflects low background power rather than strong, stable task-related modulation. Consequently, future iterations must integrate adaptive thresholding with relative baseline shifts, signal-to-noise ratio metrics, trial consistency checks, and true machine-learning classification.

In summary, the proposed prototype provides a viable, safe platform for controlled pilot experimentation, validating workflow functionality rather than MI classification accuracy. The

next development phase will focus on expanding trial counts, implementing robust artifact rejection, evaluating latency, testing clinical usability, and validating true MI detection in post-stroke participants. Additionally, integrating a synchronized robotic glove alongside FES will allow simultaneous cortical activation, peripheral electrical stimulation, and passive mechanical assistance within a unified multimodal rehabilitation framework.

REFERENCES

[1] GBD 2021 Stroke Risk Factor Collaborators, "Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *The Lancet Neurology*, vol. 23, no. 10, pp. 973–1003, Oct. 2024.

[2] V. L. Feigin, M. Brainin, B. Norrving, S. O. Martins, J. Pandian, P. Lindsay, M. F. Grupper, and I. Rautalin, "World Stroke Organization: Global stroke fact sheet 2025," *International Journal of Stroke*, vol. 20, no. 2, pp. 132–144, Feb. 2025.

[3] S. M. Hatem *et al.*, "Rehabilitation of motor function after stroke: a multiple systematic review focused on techniques to stimulate upper extremity recovery," *Frontiers in Human Neuroscience*, vol. 10, p. 442, 2016.

[4] Intercollegiate Stroke Working Party, *National Clinical Guideline for Stroke for the UK and Ireland*. London, U.K.: Royal College of Physicians, 2023. [Online]. Available: https://www.strokeguideline.org/

[5] Secretaria da Saúde do Espírito Santo, "Centro de reabilitação física do Espírito Santo – CREFES," 2026, accessed: May 25, 2026. [Online]. Available: https://saude.es.gov.br/centro-de-reabilitacao-fisica-do-espirito-santo-crefes

[6] J. Eraifej, W. Clark, B. France, S. Desando, and D. Moore, "Effectiveness of upper limb functional electrical stimulation after stroke for the improvement of activities of daily living and motor function: a systematic review and meta-analysis," *Systematic Reviews*, vol. 6, p. 40, 2017.

[7] M. A. Khan, H. Fares, H. Ghayvat, I. C. Brunner, S. Puthusserypady, B. Razavi, M. Lansberg, A. Poon, and K. J. Meador, "A systematic review on functional electrical stimulation based rehabilitation systems for upper limb post-stroke recovery," *Frontiers in Neurology*, vol. 14, p. 1272992, 2023.

[8] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, Jun. 2002.

[9] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[10] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.

[11] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[12] A. Ramos-Murguialday *et al.*, "Brain-machine interface in chronic stroke rehabilitation: a controlled study," *Annals of Neurology*, vol. 74, no. 1, pp. 100–108, Jul. 2013.

[13] K. K. Ang *et al.*, "A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke," *Clinical EEG and Neuroscience*, vol. 46, no. 4, pp. 310–320, Oct. 2015.

[14] F. Pichiorri *et al.*, "Brain-computer interface boosts motor imagery practice during stroke recovery," *Annals of Neurology*, vol. 77, no. 5, pp. 851–865, May 2015.

[15] A. A. Frolov *et al.*, "Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface-controlled hand exoskeleton: a randomized controlled multicenter trial," *Frontiers in Neuroscience*, vol. 11, p. 400, 2017.

[16] A. Biasiucci *et al.*, "Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke," *Nature Communications*, vol. 9, p. 2421, 2018.

[17] C. Ren, X. Li, Q. Gao, M. Pan, J. Wang, F. Yang, Z. Duan, P. Guo, and Y. Zhang, "The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: a systematic review and meta-analysis," *Frontiers in Human Neuroscience*, vol. 18, p. 1438095, 2024.

[18] I. Brunner, C. B. Lundquist, A. R. Pedersen, E. G. Spaich, S. Dosen, and A. Savic, "Brain computer interface training with motor imagery and functional electrical stimulation for patients with severe upper limb paresis after stroke: a randomized controlled pilot trial," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, p. 10, 2024.

[19] E. Monge-Pereira, J. Ibañez-Pereda, I. M. Alguacil-Diego, J. I. Serrano, M. P. Spottorno-Rubio, and F. Molina-Rueda, "Use of electroencephalography brain-computer interface systems as a rehabilitative approach for upper limb function after a stroke: a systematic review," *PM&R*, vol. 9, no. 9, pp. 918–932, Sep. 2017.

[20] E. López-Larraz, A. Sarasola-Sanz, N. Irastorza-Landa, N. Birbaumer, and A. Ramos-Murguialday, "Brain-machine interfaces for rehabilitation in stroke: a review," *NeuroRehabilitation*, vol. 43, no. 1, pp. 77–97, 2018.

[21] M. A. Cervera *et al.*, "Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis," *Annals of Clinical and Translational Neurology*, vol. 5, no. 5, pp. 651–663, May 2018.

[22] J. Cantillo-Negrete, R. I. Carino-Escobar, P. Carrillo-Mora, D. Elias-Viñas, and J. Gutierrez-Martinez, "Motor imagery-based brain-computer interface coupled to a robotic hand orthosis aimed for neurorehabilitation of stroke patients," *Journal of Healthcare Engineering*, vol. 2018, p. 1624637, 2018.

[23] N. Mammone, C. Ieracitano, H. Adeli, and F. C. Morabito, "Autoencoder filter bank common spatial patterns to decode motor imagery from EEG," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2365–2376, May 2023.

[24] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by Riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[25] Z. Shuqfa, A. N. Belkacem, and A. Lakas, "Decoding multi-class motor imagery and motor execution tasks using Riemannian geometry algorithms on large EEG datasets," *Sensors*, vol. 23, no. 11, p. 5051, 2023.

[26] H. Kucukselbes and E. Sayilgan, "Real-time EEG decoding of motor imagery via nonlinear dimensionality reduction and shallow classifiers," *Biosensors*, vol. 15, no. 10, p. 692, 2025.

[27] T. F. Bastos Filho, A. X. Gonzalez Cely, S. Mehrpour, S. Schreider, F. Souza, F. Cabral, and A. C. Villa Parra, "BCI in stroke rehabilitation with robotic devices, TACS and FES: A case study," *Journal of Biomedical Engineering and Biosciences*, vol. 12, no. 12, pp. 7–17, Oct. 2025.

[28] T. Bastos Filho, A. X. Gonzalez Cely, S. Mehrpour, F. Souza, A. C. Villa Parra, and F. Cabral, "Upper-limb rehabilitation in chronic stroke using brain-computer interface based on motor imagery plus non-invasive brain and muscular stimulation," in *Proc. 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Vienna, Austria, 2025, pp. 577–582.

[29] T. Bastos-Filho, A. X. Gonzalez-Cely, S. Mehrpour, F. Souza, A. C. Villa-Parra, and F. Cabral, "Rehabilitation of chronic stroke using neurofeedback, functional electrical stimulation and cerebrospinal direct current stimulation," in *Proc. 2025 International Conference on Rehabilitation Robotics (ICORR)*, 2025, pp. 1203–1208.

[30] F. V. de Souza, "Análise dos efeitos da ativação cerebral por eletroencefalografia em pacientes pós-ACV durante imagética motora, estimulação elétrica funcional e luva robótica," Master's thesis, Programa de Pós-Graduação em Biotecnologia, Universidade Federal do Espírito Santo, Vitória, Brazil, 2024, accessed: May 27, 2026. [Online]. Available: http://repositorio.ufes.br/handle/10/18103

[31] A. B. Remsik, P. L. E. van Kan, S. Gloe, K. Gjini, L. J. Williams, V. Nair, K. Caldera, J. C. Williams, and V. Prabhakaran, "BCI-FES with multimodal feedback for motor recovery post stroke," *Frontiers in Human Neuroscience*, vol. 16, p. 725715, 2022.

[32] H. Yu, S. Ba, Y. Guo, L. Guo, and G. Xu, "Effects of Motor Imagery Tasks on Brain Functional Networks Based on EEG Mu/Beta Rhythm," *Brain Sciences*, vol. 12, no. 2, Art. no. 194, Jan. 2022, doi: 10.3390/brainsci12020194.

[33] T. F. Bastos-Filho, A. C. Villa-Parra, C. D. Guerrero-Méndez, A. X. González-Cely, C. F. Blanco-Diaz, D. Delisle-Rodríguez, and T. Igasaki, "A novel methodology based on static visual stimuli and kinesthetic motor imagery for upper limb neurorehabilitation," *Research on Biomedical Engineering*, vol. 40, pp. 687–700, 2024.

[34] A. X. Gonzalez Cely, C. F. Blanco-Diaz, C. D. Guerrero Mendez, A. C. Villa Parra, and T. F. Bastos-Filho, "Classification of opening/closing hand motor imagery induced by left and right robotic gloves through EEG signals," *Transactions on Energy Systems and Engineering Applications*, vol. 5, no. 2, pp. 1–9, 2024.

[35] A. C. Villa-Parra, T. Bastos-Filho, C. D. Guerrero-Mendez, and T. Igasaki, "Towards a BCI based on action observation and kinesthetic motor imagery for neurorehabilitation," in *Proceedings of the International Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2023): Advances in Computer Sciences, Exploring Innovations at the Intersection of Computing Technologies*, M. V. Garcia, C. Gordón-Gallegos, A. Salazar-Ramírez, and C. Nuñez, Eds. Cham, Switzerland: Springer, 2024, pp. 601–612.

[36] X. Bao, Y. Zhou, Y. Wang, J. Zhang, X. Lü, and Z. Wang, "Electrode placement on the forearm for selective stimulation of finger extension/flexion," *PLOS ONE*, vol. 13, no. 1, p. e0190936, 2018.

[37] E. L. Nussbaum, P. Houghton, J. Anthony, S. Rennie, B. L. Shay, and A. M. Hoens, "Neuromuscular electrical stimulation for treatment of muscle impairment: Critical review and recommendations for clinical practice," *Physiotherapy Canada*, vol. 69, no. 5, pp. 1–76, 2017.

[38] L. C. Rodríguez Timaná, S. Schreider, F. Cabral, A. X. González-Cely, A. C. Villa-Parra, and T. Bastos-Filho, "Upper limb rehabilitation in chronic stroke through motor imagery with sequential activation of active elbow exoskeleton and robotic glove," in *Proceedings of the 13th International Conference of Control Systems, and Robotics (CDSR 2026)*, Barcelona, Spain, Jun. 2026, Paper No. 107, doi: 10.11159/cdsr26.107.

Multi-Platform 3D Skeleton Simulation Tool for Interactive Hip Posture and Dislocation

Caio Cigagna de Godoy
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP); Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, Brasil
ORCID: 0000-0002-6548-8153

Carlos Marcelo Gurjão de Godoy
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0000-0001-8846-3242

Victor de Mendonça Ramos
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0009-0002-8758-4549

Tifany Mariano
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0009-0001-5166-4478

Fabrício Morais Muller
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0009-0004-4937-3155

Eliana Chaves Ferretti
Instituto de Saúde e Sociedade
Universidade Federal de São Paulo (UNIFESP)
Santos, Brasil
ORCID: 0000-0002-7605-828X

Vagner Rogério dos Santos
Escola Paulista de Medicina
Universidade Federal de São Paulo (UNIFESP)
São Paulo, Brasil
ORCID: 0000-0002-6034-6658

Regina Célia Coelho
Departamento de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, Brasil
ORCID: 0000-0002-4428-9745

***Abstract*— Interactive three-dimensional visualization can provide a complementary approach to orthopedic training by enabling the exploration of anatomical configurations and joint movements in an interactive environment. This work presents a preliminary multi-platform 3D application developed as a visualization tool for orthopedic training, focusing on the interactive manipulation of seated hip postures and simplified visual representations of hip dislocation. The main technical contribution is the combination of a two-dimensional Blendspace for continuous interpolation between predefined postures with a multi-skeleton rendering technique that allows the lower limbs to be transformed independently without deforming the torso. Five reference poses were created in Blender and subsequently integrated into Unreal Engine 5. The Blendspace was used to generate continuous variations between these poses through two input parameters. For the dislocation representation, three instances of the same skeleton were rendered simultaneously using opaque and transparent textures, allowing the torso and each lower limb to be manipulated as visually independent components. A physiotherapist, who is a co-author, reviewed the selected model and the implemented poses qualitatively for anatomical consistency. No quantitative biomechanical validation, motion-capture comparison, clinical imaging comparison, or formal end-user usability study was performed at this stage. The resulting prototype demonstrates the technical feasibility of combining continuous posture interpolation with independent lower-limb transformation in a multi-platform environment and establishes a foundation for the use of interactive 3D visualization in orthopedic training. The dislocation component should be interpreted as a simplified visual representation rather than a biomechanical simulation. Further studies are required to quantify anatomical fidelity, evaluate computational performance, assess usability and educational effectiveness, and determine the contribution of the proposed tool to orthopedic training.**

Keywords—biomedical simulation, unreal engine 5, hip dislocation, blendspace, multi-platform UI, orthopedic training

I. INTRODUCTION

Hip posture and spinopelvic mobility play an important role in hip stability, as changes in pelvic tilt during daily activities can influence the risk of impingement and dislocation after total hip arthroplasty [1]. In this context, Unreal Engine, a real-time 3D development platform, can be used to simulate hip postures and mobility, providing an interactive environment for visualizing movements associated with hip stability.

Such a platform may facilitate medical simulation by integrating biomechanical data with real-time rendering and kinematics. This potentially supports orthopedic training, gait analysis, and musculoskeletal modeling through rigging, motion calibration, and physics-based animation [2]. In orthopedics, it enables the creation of immersive, high-fidelity surgical simulators for complex procedures such as femoral stem implantation and hip screw fixation [3,4]. Its real-time rendering also generates realistic anatomical imagery to assist surgical planning and enable a diverse approach to clinical cases for training [3]. In musculoskeletal modeling, it visualizes and simulates human movement, including hip and leg motion, using 3D pose models with independent coordinate systems for each body part. This enables precise analysis of joint displacements and angular changes [2].

Researchers use this technology for real-time posture display with calibrated motion, ensuring anatomical consistency [2], and also integrate biomechanical alignment data from OpenSim (an open-source biomechanical modeling platform) to enforce physiological constraints [5]. Among the main methods used are inverse kinematics (a technique that calculates joint angles and movement paths from desired end-effector positions) [6], motion calibration to keep interconnected joints within physiological limits [2], Blueprint and C++ programming for pose modeling and external hardware synchronization [2], and animation layering that applies variations to 3D models without causing unnatural leg deformations [6]. Additionally, its architecture allows the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

adjustment of smooth curves for key joints such as the hips, resulting unconstrained skeletal animations that can be made suitable for both medical analysis and joint movement emulation [5].

These capabilities are clinically relevant because hip and leg positioning are fundamental to joint stability [7], and understanding the specific positions that cause dislocation is essential for diagnosis [8]. Also, adherence to positioning precautions is the basis for successful prevention and rehabilitation [9]. These clinical contexts motivate the development of interactive visualization tools for exploring hip and lower-limb positioning. However, the technical feasibility of reproducing such positions in a virtual environment should be distinguished from clinical validation or educational effectiveness. In this work, the focus is therefore on the computational representation and interactive manipulation of hip posture, rather than on demonstrating diagnostic accuracy, surgical planning capability, rehabilitation effectiveness, or clinical decision-making support.

A. *Objectives*

General Objective: To develop a multi-platform application that allows users to customize the posture of a 3D skeleton, with a focus on the hip positioning and dislocation.

Specific Objectives: (1) integrate a rigged 3D skeleton model and a set of reference poses into a real-time 3D environment; (2) provide interactive control of posture variation through input controls and sliders; (3) implement a visual representation of independent lower-limb displacement without deforming the torso; (4) deploy the prototype in desktop and mobile environments; and (5) design an interface based on established usability principles as a preliminary step toward future end-user evaluation.

B. *Justification*

This work is justified by the need for a multi-platform tool that allows user-controlled customization of the positioning of the hip and leg in a 3D skeleton model. Although Unreal Engine is already used for orthopedic simulations [4,5] and motion analysis [2,5,6], there remains a need for an accessible, real-time interface for manipulating joint positioning, particularly hip alignment, while preserving anatomical realism and fluid performance across computers and mobile devices. The proposed application seeks to address the technical challenge of interactively manipulating hip posture while preserving the visual integrity of the remaining skeletal mesh. Its contribution is centered on the combination of continuous posture interpolation, independent lower-limb transformation, and a multi-platform interface. At this stage, the work is intended as a technical prototype and does not claim biomechanical fidelity or validated clinical applicability.

II. Methodology

The simulator development process is structured into five stages, which can be visually seen in the flowchart illustrated in Fig. 1. The development order is: (1) selection of a publicly available 3D skeleton model with a distortion-free rig; (2) creation of references poses; (3) qualitative anatomical review of the model, poses and and movement limits; (4) implementation of the interactive application (programming and posture) within a game engine; and (5) development of an adaptive User Interface (UI). Stage 1 used 3D model-sharing platforms such as Sketchfab and CGTrader, whereas Stage 2 was performed using the 3D editing software Blender.

Stages 1–3 were conduct with the assistance of a physiotherapist (an author of the present work) for validation before going into stages 4 and 5, in which the game engine Unreal Engine 5 (UE5), a freely available real-time 3D development platform, was chosen as the main tool for developing the application itself. It should be noted that the term “skeletal” refers to the structure that allows a 3D mesh to be animated and customized, while “skeleton” refers to the appearance of the 3D mesh itself. No quantitative angular measurements, motion-capture comparison, biomechanical model comparison, or formal inter-rater agreement procedure was performed during this stage.

A. *Development Tools*

Blender is a free and open-source 3D modeling, texturing, and animation software that enables developers to create and edit any vertex mesh of virtually any object, customize textures through multiple approaches, and produce videos and animations [10]. During the development of this project, Blender played a critical role in the creation of the assets employed, particularly the customized skeletal poses.

UE5 is a comprehensive suite of creation tools designed for game development, architectural and automotive visualization, linear content production for film and television, live event production and broadcasting, training and simulation, and other real-time applications [11]. Such platform was developed to provide realistic graphical capabilities while simultaneously ensuring low computational cost and high optimization efficiency. It offers two programming approaches: one based on the written programming language C++, and another unique visual scripting system known as Blueprints [12]. In the present project, Blueprints were selected, as they provide a comprehensive node-based interface system for creating gameplay elements and implementing programming logic directly within the Unreal Engine editor.

B. *3D Model Search*

More specifically, the search for the 3D model to be utilized followed a set of predefined requirements: the model had to be free of charge, editable, visually representative of a human skeleton, and equipped with a functional rig suitable for skeletal animation. Anatomical plausibility was subsequently reviewed by the physiotherapist co-author. The search resulted in the selection of the 3D model made by CreativeMachine, entitled “Skeleton Rig”, available on the digital 3D model-sharing platform Sketchfab, as it fulfilled all established model requirements. Fig. 2 illustrates the “Skeleton Rig”, a free 3D model by CreativeMachine [13].

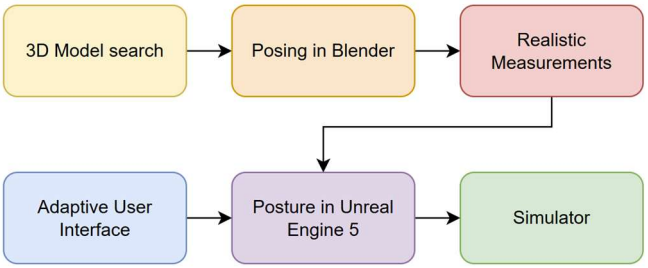


Fig. 1. Methodology flowchart (details in the text).

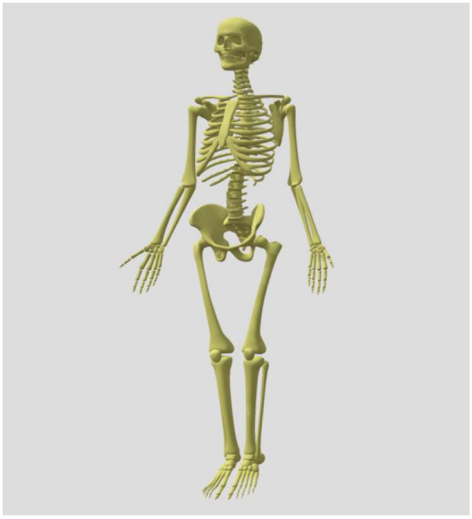


Fig. 2. Skeleton rig" by CreativeMachine, 2020.

The selected 3D mesh was qualitatively reviewed by the physiotherapist co-author before its integration into the application. The review considered the overall skeletal organization, number and relative arrangement of bones, general body proportions, and the behavior of the rig during the implemented movements. Next, the same reviewer inspected the five reference poses and the available posture transitions to identify movements that were visually inconsistent with the intended anatomical configuration.

This review was qualitative and was not based on a formal validation checklist, quantitative angular measurements, inter-rater agreement, motion-capture data, or comparison with clinical imaging. Therefore, the review should be understood as an expert anatomical inspection rather than as a quantitative validation of anatomical fidelity.

C. *Posing in Blender*

For this specific project, the most important feature of Blender is its animation structure tool known as bones [14]. Bones interact with an object's point mesh through an influence field, and each bone possesses three-dimensional transformation properties, including positioning, rotation, and scaling within the 3D environment. These properties are editable, allowing the properly rigged skeleton to be manipulated through the use of bones. This enables the creation of anatomically plausible and customizable poses. Rigging is a general term used for adding controls to objects, typically for the purpose of animation and, in this case, the posing of the skeleton [14]. The seated position was chosen for the focus on the hip, with five different poses being created: Default, turned right, turned left, curved back and curved forward. Once the poses were created, they were exported into UE5.

D. *Anatomical Review of Poses and Movement Limits*

The reference poses and movement limits were defined to provide visually plausible variations of the seated hip position. A physiotherapist co-author reviewed each reference pose, the resulting interpolated movements, and the intended customization limits from an anatomical perspective. The review was qualitative and was used to identify gross inconsistencies in the direction and appearance of the simulated movements. At this stage, the system does not associate the Blendspace coordinates with experimentally measured joint angles. Consequently, the numerical values of the X and Y parameters represent animation-space control values rather than degrees of anatomical motion.

E. *Posture in UE5*

Importing the assets produced in Blender into the UE5 editor constitutes a particularly delicate stage of the process, as it requires the proper definition of which 3D meshes are associated with specific bone structures, which, in turn, must be correctly linked to their corresponding animations. Such approach provides a robust set of tools supporting a wide range of media creation and development possibilities, among which one of the most crucial for the present project is the Blendspace feature. Blendspace is an environment in which multiple animations can be blended based on the values of several input parameters, provided that they share the same bone structure and 3D mesh, thereby enabling smooth transitions between animations. Through UE5's Blendspace, it is possible to define points within an XY plane, each corresponding to a specific animation. The Blendspace blends animations dynamically according to command-based inputs [15].

The main systems to be programmed in the platform are: (1) Hip posture customization through Blendspace and (2) Hip dislocation through object transformation editing. The Blendspace imported allow every intended system and functionalities to be developed in tandem and to run alongside each other in the finished product.

F. *Adaptive User Interface*

The User Interface (UI) of this application is designed to be adaptable to different types of content, provide a high level of accessibility, and maintain functionality across multiple platforms. Based on the work of Molich and Nielsen [16], which describes heuristic approaches for evaluating the quality of user interfaces, the present project adhered to these principles as part of its quality control strategy.

When studying the heuristics, a few characteristics were defined as essential for the design ideals of the UI: (1) Large buttons, lettering and text boxes for good readability; (2) Preserving the same layout logic consistently across different menus to allow for low effort and easy understanding of the application's systems.

III. RESULTS

The results section of this article is divided into three facets, all that can be achieved by implementing the methodologies described. These three facets are: (1) Hip posture customization; (2) Hip dislocation; and (3) UI ease of use and adaptability. It should be noted that, as it is currently the case for this simulation, it does not incorporate forces, joint constraints, capsule, ligaments, soft tissues, or the physical dynamics of dislocation, focusing on simulating a visual representation of angle and position changes of the hip.

A. *Hip Posture Customization*

The customization aspect of the posture was achieved once the 3D meshes and animation poses were imported into the game engine. Fig. 3 illustrates the Blendspace developed for the project, highlighting colored points representing different animations positioned within an XY coordinate system ranging from -100 to 100. The yellow point represents the "default" pose and is positioned at coordinates XY (0, 0); the green point corresponds to the "curved forward" pose at XY (0, 100); the blue point represents the "turned right" pose at

XY (100,0); the purple point indicates the “curved back” pose at XY (0, -100); and the red point represents the “turned left” pose at XY(-100, 0). These graph positions enable a resting state at the coordinate point XY (0, 0), while any increase or decrease in the values of the X and/or Y variables generates a gradual transition toward the remaining poses. Fig. 4 illustrates the resulting pose when the X value increases while the Y value decreases, producing a fully customized hip position in which the posture becomes simultaneously slightly rotated to the right and flexed backward. This configuration enables the simulation of numerous hip-bas postural conditions; however, an additional method of posture modification was implemented to further expand the range of simulation possibilities: the development of a hip dislocation simulation between the pelvis and lower limbs. position changes of the hip.

B. Hip Dislocation

The development of a dislocation pose or animation required the implementation of a new logic to enable the simultaneous blending of Blendspace poses with the specific displacement of body segments. For this purpose, the textures applied to the skeleton model in this project were specifically created: a white texture for standard visualization and a fully transparent texture to conceal unnecessary anatomical segments.

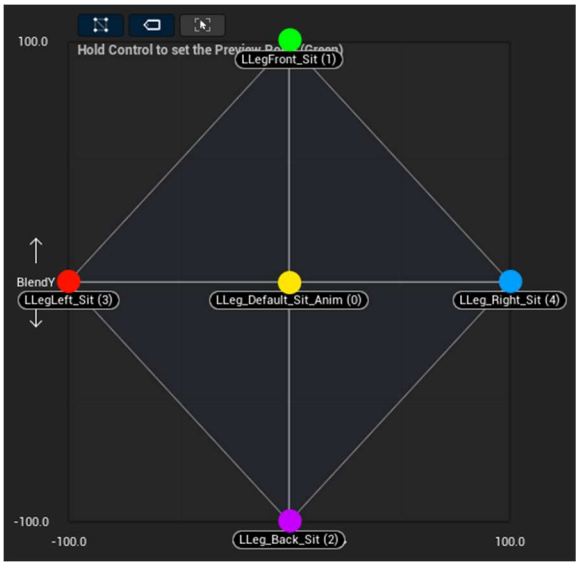


Fig. 3. Assembled Blendspace in UE5.

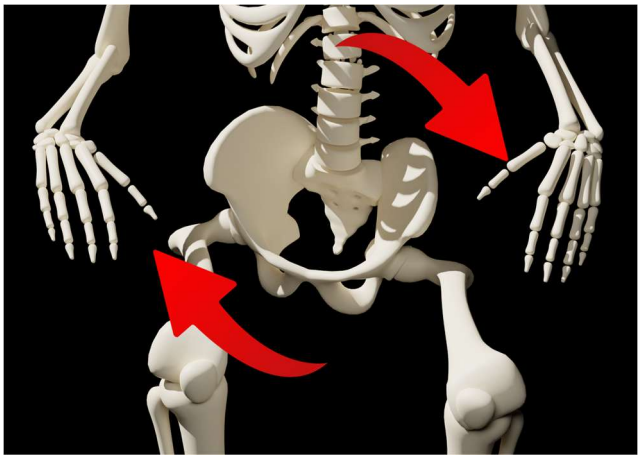


Fig. 4. Customized pose generated through the simultaneous application of positive X values and negative Y values.

By employing three duplicated instances of the 3D skeleton model, each with its respective bone structures and animations linked to the Blendspace, it becomes possible to create the visual effect of a single skeleton when, in fact, three separate models are being rendered simultaneously. Through the strategic application of opaque and transparent textures, specific skeletal regions can be selectively hidden such that: (1) one model visually excludes both lower limbs; (2) another displays only the left leg; and (3) a third displays only the right leg. Fig. 5 presents a representation of the three skeletons occupying the same spatial position, with the colors red, green, and blue indicating the distinct skeletal instances, thereby demonstrating that the resulting visual effect corresponds to the appearance of a single unified skeleton.

The fact that each lower limb is treated as an independent object from the torso enables the application of positional and rotational transformations without affecting or deforming other regions of the skeleton. Consequently, it becomes possible to modify the position and rotation of one or both lower limbs independently in order to generate a simplified visual representation of hip dislocation. Fig. 6 illustrates the resulting visual representation obtained with this method. The independent treatment of the lower limbs allows positional and rotational transformations to be applied while preserving the posture generated by the Blendspace. This demonstrates the technical operation of the proposed rendering strategy; it does not constitute a biomechanical or clinical validation of hip dislocation.

C. User Interface Ease of Use and Adaptability

The interface was implemented as a conceptual multi-platform layout organized into functional regions for camera control, application settings, simulation selection, parameter adjustment, and feedback. The design follows selected principles derived from the human-computer dialogue guidelines described by Molich and Nielsen [16], including consistency, feedback, and visibility of relevant controls. At the present stage, these principles were used as design guidelines rather than as the result of a formal usability evaluation. No end-user testing, usability questionnaire, cognitive-load assessment, task-performance measurement, or acceptance study was conducted.

With the heuristics proposed by Molich and Nielsen in mind, Fig. 7 presents a representative conceptual design of the planned interface for this application, while Fig. 8 divides the interface into sectors to facilitate the identification of the elements composing the user interface.

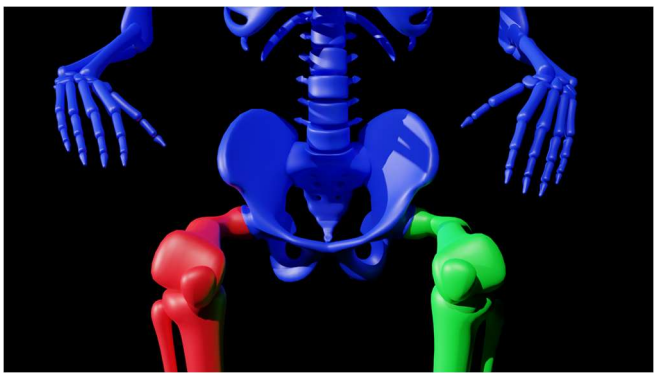


Fig. 5. Three color-coded skeletons combined to visually appear as a single skeleton structure.

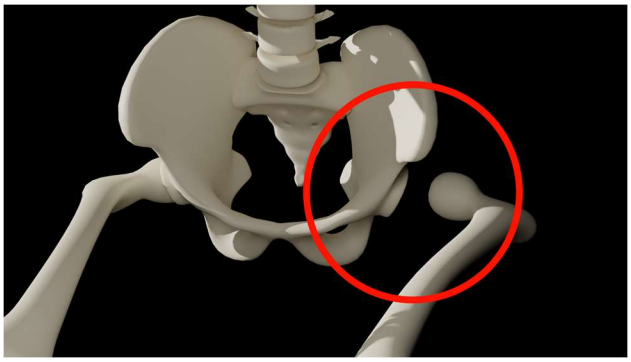


Fig. 6. Simulated hip dislocation highlighted by a red circle.

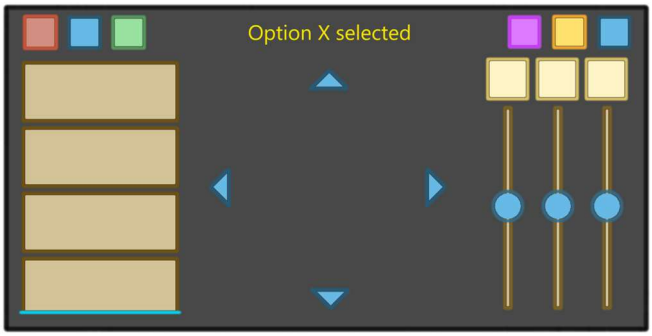


Fig. 7. Planned User Interface.

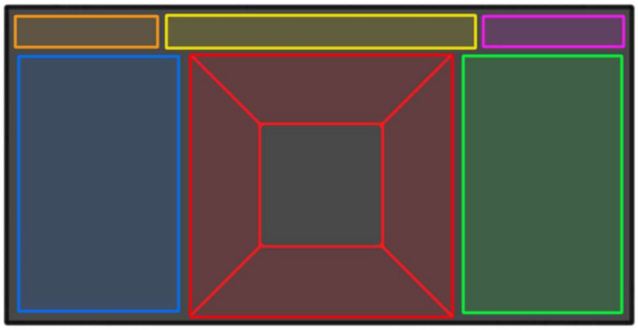


Fig. 8. Segmented User Interface in colored areas.

The red area corresponds to the active region assigned to the directional controls. It is responsible for manipulating the user's viewpoint, enabling both skeleton rotation and camera displacement across the screen plane. The switching mechanism between rotational and translational camera control is performed through a button located in the orange area.

The orange area is also intended to contain additional option buttons, such as camera mode selection, interface element visibility control, and access to the app's functionalities.

The purple area is designated to contain buttons related to general application settings, alternative operational modes, and notifications. The yellow area serves as a feedback display region, where textual information is presented to the user, including performed actions, general notifications, and the names of accessed menus.

Initially, the blue and green areas are not intended to be visible to the user. The blue area may be accessed through a button located in the orange area and contains buttons associated with simulation options, such as posture customization or dislocation simulation.

The green area becomes visible whenever a simulation option is selected in the blue area. It provides controls for modifying simulation parameters and dynamically adapting its interface according to the option selected within the blue area. An example of the controls it would provide is modifying the value through sliders or simply inputting exact values. All measurements are meant to reflect converted realistic values for the user once the calibration is finished. The menu logic was designed to be intuitive, featuring continuous feedback, accessible design principles, and expandability for future functionalities. Furthermore, the interface architecture was conceived with the intention of ensuring compatibility across multiple platforms, including desktop computers and mobile devices, when necessary.

User testing and feedback is required for validation and potentially enhancing the designs.

IV. Discussion

The interface was implemented as a conceptual multi-platform layout organized into functional regions for camera control, application settings, simulation selection, parameter adjustment, and feedback. The design follows selected principles derived from the human-computer dialogue guidelines described by Molich and Nielsen [16], including consistency, feedback, and visibility of relevant controls. At the present stage, these principles were used as design guidelines rather than as the result of a formal usability evaluation. No end-user testing, usability questionnaire, cognitive-load assessment, task-performance measurement, or acceptance study was conducted. The following discussion outlines the potential clinical application, the educational implications, limitations, and future direction.

A. Postural Customization and Biomechanical Relevance

The Blendspace system developed in this project enables continuous control over the interpolated posture via a two-dimensional coordinate system (X, Y) spanning -100 to 100. This approach allows for gradual transitions between five distinct poses: default (0,0), curved forward (0,100), turned right (100,0), curved back (0,-100), and turned left (-100,0). The resulting capacity to generate continuous variations between these poses, as demonstrated in Fig. 4, enables the simulation of numerous hip-based postural conditions with continuous control over the interpolated posture within the defined animation space.

The ability to customize hip postures is clinically significant, as hip and leg positioning are fundamental to joint stability [2]. Understanding the specific positions that cause dislocation is essential for diagnosis [9], while adhering to positioning precautions is the cornerstone of prevention and successful rehabilitation [8]. The Blendspace system enables users to explore the full spectrum of hip positioning, including the classic provocative positions associated with dislocation risk.

From a biomechanical perspective, the capacity to simulate simultaneous rotations and flexions (e.g., positive X values combined with negative Y values) is particularly valuable. This approach emulates real-world clinical scenarios, in which hip dislocations typically occur during complex, multiplanar movements rather than isolated motions. For instance, posterior hip dislocations—which

account for the majority of hip dislocations—typically occur when an axial load applies to a flexed, adducted, and internally rotated hip [9]. The two-dimensional Blendspace allows combinations of the predefined animation directions to be explored continuously. For example, simultaneous changes in the X and Y parameters produce postures that combine more than one animation direction. Such combinations are useful for exploring the visual behavior of the model in multiplanar configurations. However, because the parameters are not calibrated to measured anatomical angles and no biomechanical constraints are currently implemented, these configurations should not be interpreted as quantitatively reproducing physiological or pathological hip motion.

B. Hip Dislocation Simulation: Technical Innovation and Potential Clinical Applications

The development of the dislocation simulation feature was made by employing three duplicate instances of the 3D skeleton model and strategically applying opaque and transparent textures. This system creates the visual impression of a single, unified skeleton while enabling independent manipulation of each lower limb. This approach offers several advantages over traditional skeletal animation methods. First, treating each lower limb as an independent object from the torso enables positional and rotational transformations without affecting or deforming other regions of the skeleton. This strategy addresses the specific mesh-deformation problem encountered during the development of the present prototype. Second, the method preserves the posture generated through the Blendspace system while simultaneously applying dislocation-specific transformations (Fig. 5), enabling realistic simulation of hip dislocation across multiple postural contexts.

The project anticipates potential clinical relevance from those features. Hip dislocation is a medical emergency requiring urgent reduction, and diagnostic accuracy depends on recognizing characteristic limb positions [9]. Posterior dislocation typically presents with the leg shortened, flexed, adducted, and internally rotated, whereas anterior dislocation is generally associated with extension, abduction and external rotation [7]. By enabling users to simulate and visualize these characteristic presentations, the system may to serve as training tool for medical students, residents, and emergency medicine practitioners. Furthermore, the ability to simulate dislocations within a controlled virtual environment may provide a safe setting for exploring hip positioning relevant to surgical planning.

Approximately two-thirds of dislocation episodes are treatable with non-surgical management, including closed reduction, education, and muscle strengthening [9]. Virtual simulation could allow practitioners to practice reduction techniques involving hips, joints and leg bones without risk to patients, potentially improving clinical outcomes.

C. User Interface Design and Potential Educational Utility

The interface design described in this study adheres to established usability heuristics, including the principles proposed by Molich and Nielsen [16], such as consistency, feedback, and progressive disclosure. The modular architecture may promote intuitive navigation and reduce cognitive load with clearly delineated functional areas (viewpoint control, menu access, feedback display, and simulation parameter controls).

The design decision to initially hide advanced simulation options (blue and green areas) until accessed through primary navigation is consistent with progressive disclosure principles, preventing information overload while maintaining full functionality for advanced users. The inclusion of continuous feedback mechanisms (yellow area) and dynamic interface adaptation (green area) aims to further enhance usability.

The modular design enables the addition of new simulation options and alternative operational modes, and supports compatibility across multiple platforms (desktop and mobile). This forward-compatible approach is essential given the rapid evolution of both simulation technologies and clinical applications.

As a whole, these features of the user interface are particularly important in educational contexts where users may have varying levels of technical expertise. However, future tests with final users are still necessary to verify such potential resources.

D. Relationship to Existing Simulation Approaches

Previous studies have explored real-time 3D environments for the visualization and analysis of human movement. Liu et al. [2], for example, developed a visual gait analysis system based on Unreal Engine, while Pescaru et al. [6] investigated human movement simulation using full-body inverse kinematics in Unreal Engine. In the context of orthopedic training, Racy et al. [3] developed and validated a virtual reality simulator for femoral intramedullary nailing, combining a 3D virtual environment with haptic interaction and image intensifier functionality. More broadly, a systematic review by Sun et al. [17] identified 40 studies investigating virtual and augmented reality applications in hip-related surgery, including preoperative simulation, surgical training, intraoperative applications, and postoperative rehabilitation [17]. These studies demonstrate the growing use of interactive and immersive technologies in musculoskeletal and orthopedic contexts, but they address objectives and levels of validation that differ from those of the present work.

In contrast, the present prototype focuses on a more specific technical problem: combining continuous posture interpolation through a two-dimensional Blendspace with independent transformation of the lower limbs using multiple instances of the same skeletal model. Unlike validated surgical simulators, the proposed system does not currently provide haptic feedback, biomechanical constraints, quantitative anatomical calibration, or validated educational tasks. Its contribution should therefore be understood as a rendering and interaction strategy that establishes a technical basis for future investigation of interactive 3D visualization in orthopedic training.

E. Integration with UE5 Capabilities

The platform visualizes and simulates human movement using 3D pose models with independent coordinate systems for each body part, enabling precise analysis of joint displacements and angular changes [3]. The present project extends these capabilities by developing a specialized approach to hip dislocation simulation to address specific clinical needs.

Researchers have used Unreal Engine for real-time posture display with calibrated motion, ensuring anatomical

consistency [3]. The Blendspace-based customization approach in this project similarly ensures calibrated motion transitions, providing anatomical consistency across the range of simulated postures. Additionally, previous work has integrated biomechanical alignment data from OpenSim to enforce physiological constraints [6]. Future iterations of this project could incorporate additional constraints to enhance biomechanical realism further and validate the simulation against established biomechanical models.

The methods employed in this project—including inverse kinematics, motion calibration, and animation layering—align with established field practices. The use of Blueprint and C++ programming for pose modeling and animation control builds upon established UE5 architectures [3,7].

F. Limitations and Future Directions

While the results presented are promising, the work has some limitations. First, the current implementation focuses on visual simulation rather than biomechanical simulation. The system visualizes postures and dislocations without simulating the underlying forces, torques, and constraints that govern actual joint mechanics. Future developments may integrate OpenSim-based biomechanical models to enforce physiological constraints during simulation.

Second, validation of the simulation against real, customized clinical data has not yet been performed, although we have considered dislocation values consistent with those described in the literature. Comparison with motion capture data from patients with hip dislocations or post-surgical complications would help establish the system's accuracy and further clinical utility.

Third, the system currently relies on a simplified skeleton model. Integration with more detailed anatomical models, including surrounding muscles, tendons, and ligaments, could enhance the simulation's fidelity and expand its educational utility.

Fourth, regarding its educational purpose, the platform still lacks testing with final users, including medical educators, students, and professionals.

Future directions for this include possibilities for: (1) Integration of haptic feedback to simulate the sensation of joint reduction procedures; (2) Implementation of real-time biomechanical analysis, including joint contact forces and stress distributions; (3) Development of patient-specific simulations based on individual anatomical data; integration of virtual reality (VR) capabilities for immersive surgical training; (4) Validation studies comparing simulation results with clinical outcomes, and user tests.

V. Conclusion

This study presented the development of a preliminary multi-platform 3D prototype for interactive manipulation of hip posture and simplified visual representation of hip dislocation. The main technical contribution is the combination of a two-dimensional Blendspace for continuous posture interpolation with a multi-skeleton rendering strategy that enables independent transformation of the lower limbs while preserving the visual configuration of the torso.

The implemented prototype demonstrates the technical feasibility of this approach within Unreal Engine 5 and provides a basis for exploring interactive skeletal configurations in desktop and mobile environments.

The main lessons learned from the development process concern the usefulness of separating the visual representation of body segments when direct manipulation of a single rigged mesh produces undesirable deformations, as well as the need to distinguish animation-space parameters from quantitatively validated anatomical measurements.

Future work should focus on quantitative anatomical validation, biomechanical constraints, objective computational performance evaluation, formal usability studies, educational validation with representative users, and comparison with existing orthopedic and musculoskeletal simulation approaches. Further extensions may include patient-specific models, virtual reality, haptic feedback, and artificial intelligence for controlled generation of clinical scenarios. Thus, the present work should be regarded as an initial technical feasibility study that establishes an interactive foundation for subsequent biomechanical, usability, and educational investigations.

Acknowledgments and Declaration of no Conflict

The authors acknowledge the financial assistance from FINEP (proc. 2174/22) and declare no conflict of interest.

References

[1] A. Navacchia, M. K. Stein, D. G. Deckey, and T. M. Seyler, “How does total hip impingement risk during activities of daily living change with pelvic tilt and spinopelvic mobility?,” J. Arthroplasty, vol. 41, no. 9S1, pp. S179–S184, Sep. 2026, doi: 10.1016/j.arth.2026.04.107.

[2] R. Liu, L. Liu, G. Ma, S. Feng, Y. Mu, D. Meng, S. Wang, and E. Cai, “Visual gait analysis based on UE4,” Sensors, vol. 23, no. 12, Art. no. 5463, 2023, doi: 10.3390/s23125463.

[3] M. Racy, A. Barrow, J. Tomlinson, and F. Bello, “Development and validation of a virtual reality haptic femoral nailing simulator,” Journal of Surgical Education, Nov. 2020, doi: 10.1016/J.JSURG.2020.10.004.

[4] M. Mediouni, D. Ziou, and F. Cabana, “Immersive simulation of the reduction of femoral diaphyseal fracture,” in CAOS 2017. 17th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery, K. Radermacher and F. Rodriguez Y Baena, Eds., EPiC Series in Health Sciences, vol. 1, 2017, pp. 62–66, doi: 10.29007/cfdp.

[5] M. Khan, S. Krauss, and D. Stricker, “Towards unconstrained 2D pose estimation of the human spine,” arXiv, vol. abs/2504.08110, Apr. 2025, doi: 10.48550/arXiv.2504.08110.

[6] A. M. Pescaru, R. A. Bertici, T.-R. Plosca, and M. V. Micea, “Simulating human physical behavior using Unreal Engine full body inverse kinematics,” in Proc. 2024 Int. Symp. Electron. Telecommun. (ISETC), Timișoara, Romania, Nov. 2024, pp. 1–4, doi: 10.1109/ISETC63109.2024.10797341.

[7] W. K. Vasileff, S. H. Patel, M. R. Karns, J. E. Voos, and M. J. Salata, “Traumatic hip instability,” Operative Techniques in Sports Medicine, vol. 23, no. 3, pp. 195–202, 2015, doi: 10.1053/j.otsm.2015.07.003.

[8] Z. C. Lum, M. Giordani, and J. P. Meehan, “Total hip instability and the spinopelvic link,” Current Reviews in Musculoskeletal Medicine, vol. 13, no. 4, pp. 425–434, 2020, doi: 10.1007/s12178-020-09648-6.

[9] B. A. Rudelli et al., “Instability prevention and treatment in total hip replacement,” Revista Brasileira de Ortopedia, vol. 60, no. 5, pp. 639–651, Sep.–Oct. 2025, doi: 10.1055/s-0045-1810122.

[10] Blender Foundation, “Blender—About,” Blender Homepage, 2026. [Online]. Available: https://www.blender.org/about/. Accessed: Jun. 10, 2026.

[11] Epic Games, “Unreal Engine,” Unreal Engine Homepage, 2026. [Online]. Available: https://www.unrealengine.com/. Accessed: Jun. 8, 2026.

[12] Epic Games, “Blueprint foundations,” Unreal Engine 5.5 Documentation, 2026. [Online]. Available: https://dev.epicgames.com/documentation/unreal-engine/blueprint-foundations?application_version=5.5. Accessed: Jun. 8, 2026.

[13] CreativeMachine, “Skeleton rig,” Sketchfab, 2020. [Online]. Available: https://sketchfab.com/3d-models/skeleton-rig-ee87f94159cf4767b1977659db3fb7aa. Accessed: Jun. 15, 2026.

[14] Blender Foundation, “Animation & rigging,” Blender 3.6 Manual, 2025. [Online]. Available: https://docs.blender.org/manual/en/3.6/animation/introduction.html#rigging. Accessed: Jun. 10, 2026.

[15] Epic Games, “Blend spaces,” Unreal Engine 5.5 Documentation, 2026. [Online]. Available: https://dev.epicgames.com/documentation/unreal-engine/blend-spaces-in-unreal-engine?application_version=5.5. Accessed: Jun. 8, 2026.

[16] R. Molich and J. Nielsen, “Improving a human-computer dialogue,” Communications of the ACM, vol. 33, no. 3, pp. 338–348, 1990, doi: 10.1145/77481.77486.

[17] P. Sun, Y. Zhao, J. Men, Z. –R. Ma, H.-Z. Jiang, C.-Y. Liu, and W. Feng, “Application of Virtual and Augmented Reality Technology in Hip Surgery: Systematic Review,” Journal of Medical Internet Research, vol. 25, p. e37599, 2023, doi: 10.2196/37599.

Multidimensional Barriers And Challenges In Daily Life Lower Limbs Prosthesis

Diego Ygor Silva e Santos
PPG-Design UFPE
Universidade Federal de Pernambuco
Recife, Brazil
diego.ygor@ufpe.br

Ewertom Cordeiro Gomes
PPG-Eng. Biomédica UFPE
Universidade Federal de Pernambuco
Recife, Brazil
cordeiro.gomes@ufpe.br

Ana Lays Braga
PPG-Design UFPE
Universidade Federal de Pernambuco
Recife, Brazil
ana.lays@ufpe.br

Alana Elza Fontes Da Gama
PPG-Design UFPE
Universidade Federal de Pernambuco
Recife, Brazil
alana.elza@ufpe.br

***Abstract*— This article investigates the main barriers and difficulties in the use of lower limb prostheses, from the perspective of users and health professionals. Through a qualitative-quantitative approach, data were collected from 20 prosthesis users and 20 health professionals. The results reveal that, although 100% of professionals recognize the positive impact of the prosthesis on quality of life, 60% of users report partial adaptation, 80% experience physical discomfort, and 65% experience social discomfort. The comparative analysis highlights divergences between clinical perception and the user's subjective experience, pointing to the need for innovation in prosthetic component design and user-centered rehabilitation protocols, and providing guidance for rehabilitation engineering.**

Keywords*— *Lower limbs, Amputation, Prosthesis, Perceptions

I. INTRODUCTION

Disability is defined, under Brazilian Law No. 13.146/2015 (Lei Brasileira de Inclusão), as a long-term physical, mental, intellectual, or sensory impairment that, in interaction with environmental barriers, may hinder an individual's full and effective participation in society on an equal basis with others, being a definition aligned with the social model of disability [1]. Complementarily, Decree No. 3.298/1999, as regulated by Decree No. 5.296/2004, classifies disability into distinct categories, including physical, hearing, and intellectual/mental disability, among others. Amputation is categorized as a physical disability, defined as the total or partial loss of a limb or segment resulting in long-term functional impairment, as specified in Article 4, item I, of Decree No. 3.298/1999 [2]. Globally, the World Health Organization estimates that over one billion people live with some form of disability, of whom approximately 200 million experience severe functional limitations [3].

In Brazil, lower limb amputations account for roughly 85% of cases, with an incidence of 13.9 per 100,000 inhabitants/year. Leading causes include diabetes-related peripheral vascular disease (90.7% of major amputations), trauma (5.6–33%), infections (~13%), and neoplasms (0.8–1.9%), with congenital malformations playing a minor role. By procedure type, external causes/trauma (33.1%), infections (17.9%), circulatory disease (16.1%), and diabetes (13.6%) predominate, most often affecting men over 60 at higher amputation levels (transfemoral or transtibial). [4]. Thus, amputation ends up being, for the individual, a multifaceted trauma that considerably alters the individual's existence, perception, and experience, generating physical repercussions, such as phantom pain, motor limitations, and functional dependence [5]; psychological repercussions, including grief, depression, anxiety, low self-esteem, and social isolation [6]; and psychosocial repercussions, with impacts on body identity, interpersonal relationships, and return to work, often high due to the action of stigma and lack of support [7].

Thus, stigma emerges as a social disqualification arising from stereotypes that label the amputee as inferior, fostering embarrassment, shame, and aversion to one's own body, which interferes with self-care, rehabilitation, and daily interactions, as reported in studies that identify the perception of discrimination as a catalyst for isolation [8]. This marginalization synergizes with clinical and sociodemographic harms, promoting therapeutic non-adherence and perpetuating a vicious cycle of exclusion, in which physical prejudice promotes distress and social withdrawal [9].

As Chamlian and Varanda [10] show, abandonment of prostheses by individuals with lower limb amputations is a prevalent phenomenon, often exceeding 50% after discharge from rehabilitation. This high dropout rate stems from multifactorial barriers, notably the excessive weight of the prosthesis (47.5% of cases), fear of falling (20%), difficulty putting on and removing the device (16.25%), reamputations (6.25%), fatigue (2.5%), phantom pain (1.25%), and comorbidities (5%), which were the main factors compromising autonomy and leading patients to prefer wheelchair mobility.

Abandonment originates both from a stigmatizing web combined with emotional burdens that generate very negative effects on the person who underwent amputation, such as anxiety and depression, as well as from technical and functional problems associated with prosthesis use. For this reason, this alteration of body image ends up breaking affective bonds and generates marital ruptures or parental estrangement, especially when combined with fears of dependence and perceived incapacity [11]. Understanding this stigmatizing web is key to grasping both patients' experiences and health professionals' perspectives, since medical perception shapes the rehabilitation process. This raises the study's guiding question: what are the main problems in using a lower-limb prosthesis, according to patients and health professionals?

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

II. METHODOLOGY

A. Research characterization

The present research is characterized by a qualitative-quantitative approach, as it integrates qualitative and quantitative procedures to achieve a broader understanding of the phenomenon under investigation. Such an approach allows not only the measurement of objective data but also the in-depth interpretation of participants' experiences and perceptions, and is widely recommended in studies in the health field and applied social sciences [11][12]. In this sense, the combination of methods enables greater analytical robustness by articulating different perspectives on the use and rehabilitation of lower limb prostheses.

The methodological procedure followed sequential stages: a literature review to ground the theoretical framework and define data collection instruments [15]; data collection using the described instruments; and data processing, involving organization, categorization, and analysis. In the case of qualitative data, content analysis procedures were adopted, as proposed by Bardin [16], allowing the identification of relevant thematic categories. Bardin's Content Analysis comprises three phases: pre-analysis (floating reading and definition of objectives), material exploration (data coding and categorization), and treatment of results (inference and interpretation). In this study, the material was read, categorized by tabulation in a spreadsheet, and subjected to frequency counting of occurrences within each category. The codification of the data generated through Bardin's content analysis method was conducted by a single researcher, thereby ensuring consistency and uniformity throughout the data coding process,Thematic categories were constructed inductively: each open-ended response was first segmented into meaning units, which were then grouped by semantic similarity. Codes with overlapping meaning were merged, and checked against two criteria recommended by Bardin: mutual exclusivity and exhaustiveness until a stable set of categories per group was reached.

Quantitative data were analyzed using descriptive statistics, enabling the systematization of professionals' responses. Finally, the results were interpreted in light of the theoretical framework, culminating in the elaboration of the research conclusions. In summary, the methodological procedure is outlined in the following steps in Fig. 1:

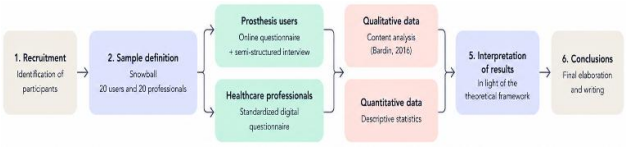


Fig. 1. Methodological flowchart for data collection and processing

Regarding the sampling strategy, the snowball method was adopted, characterized by the successive referral of participants from initial contacts. This type of sampling is particularly useful in research with specific or hard-to-reach populations, allowing the progressive construction of the sample from the participants' own social and professional networks [13]. This strategy proved suitable for the research context, especially for identifying prosthesis users and professionals with experience in rehabilitation. Participant recruitment was conducted primarily through referrals from healthcare professionals responsible for the direct care and follow-up of individuals with lower-limb amputations. These professionals identified potential participants who met the study's inclusion criteria and, upon obtaining their consent to be contacted, invited them to take part voluntarily.

Data collection was carried out using two complementary techniques. For the user group, an online questionnaire and a semi-structured interview were used (17 questions), allowing exploration of subjective aspects, perceptions, and individual experiences [14]. The interviews were made available in online video call format and were answered only by participants who were prosthesis users.

For the health professionals group, a standardized digital questionnaire was used (15 questions), favoring standardization of responses and access to geographically dispersed participants [15]. It is important to note that the two instruments were designed independently, with different question wordings and, in several cases, different underlying constructs; professionals were not reporting on the same patients who answered the user questionnaire, and the two samples do not form user–professional dyads. This methodological feature is revisited in Section III.D, where its implications for the comparative analysis are made explicit.

The study was approved by the Research Ethics Committee (CEP) of the Hospital Universitário Alcides Carneiro, Universidade Federal de Campina Grande (HUAC-UFCG), under Opinion No. 7.050.410 and CAAE 81041924.6.0000.5182, in accordance with Resolution CNS 466/12.

The sample consisted of two distinct groups: lower limb prosthesis users and health professionals. In both groups, the participation of 20 individuals was estimated, being a number considered adequate for exploratory studies with a mixed approach, in which a balance is sought between qualitative depth and quantitative representativeness [11]. The sample (n=20 per group) was defined by theoretical saturation: data collection ended when the open-ended responses stopped generating new codes, with the cumulative-code curve stabilizing around the 20th participant, a criterion aligned with methodological recommendations for transparency and adequate sample size. [23].

To map the main perceptual points of both groups, a total of 17 questions were administered to prosthesis users[1] and 15 to healthcare professionals[2]. Table I presents only the key questions selected to capture the physical, emotional, social, and contextual impacts addressed in the study, outlined in Table I below:

TABLE I. KEY QUESTIONS ASKED TO BOTH GROUPS

Users key questions	Health professionals key questions
1. Do you feel any discomfort when using your prosthesis?	1. In your view, does prosthesis use improve the patient's quality of life? Why?

[1] amputees questionnaire: https://forms.gle/k2RP9bHrttuyxMCU7

[2] professionals questionnaire: https://forms.gle/6BJFxC2Za42vnXGq8

2. Is there anything about the appearance of your prosthesis that bothers you? Why?	2. Do you perceive any negative relationship between the user and the prosthesis's appearance?
3. Do you feel comfortable exposing your prosthesis in public?	3. Have any patients mentioned social discomfort when using their prosthesis? If so, in what situation?
4. Is there any situation that prevents you from using the prosthesis for long periods?	4. In your experience, what is the greatest challenge in adapting to the prosthesis?
5. Have you ever felt frustrated while using the prosthesis?	5. Which aspects are most positively impacted by prosthesis use?

The inclusion criteria for the user group were: being a person with a lower limb amputation at any level, using a prosthesis, being between 18 and 80 years old, having no visual limitations that would prevent reading the instruments or cognitive impairments that would hinder understanding of the questions, in addition to agreeing to the terms established in the Informed Consent Form (ICF). For the professionals group, the inclusion criteria involved: being a health professional with experience in the treatment and rehabilitation of amputees, having knowledge about prosthesis use, and accepting the terms of the ICF. The implementation of this methodological framework facilitated an in-depth analysis of the phenomenon under investigation by integrating both quantitative and qualitative data, along with multiple perspectives from users and healthcare professionals regarding the rehabilitation process and the utilization of lower limb prostheses.

B. Limitations and future research

This study presents the following limitations: a small non-probabilistic snowball sample (n = 20 amputees + 20 professionals), a lack of control for variables such as amputation level and time since amputation. Also, this research shows a lack of quantitative results and variation in the answers, both synchronous and asynchronous. Additionally, since coding was performed by a single researcher, inter-rater reliability could not be assessed.

Future research should employ larger and more representative samples, combine self-report with psychometric measures (e.g., EEG, eye tracking), compare some specific cover design configurations, and include paired user-professional analyses of the same cases.

III. RESULTS & DISCUSSION

To understand the experience of using a lower-limb prosthesis in all its complexity, it is necessary, above all, to recognize the plurality of perspectives that shape the perceptions of both amputees and health professionals. For that purpose, this study collected data from 20 users for each of these profiles (where it will be represented by the letter 'n', that is, the number of participants/respondents).

A. Content Analysis

Based on the responses provided by the professionals, a content analysis was carried out based on the method proposed by Bardin [16], allowing the identification of five central thematic categories, presented in Table II below:

TABLE II. CONTENT ANALYSIS (HEALTH PROFESSIONALS)

Category	***Definition***	***Relative frequency***
1. Functionality and independence	Recovery of functional capacity and autonomy in activities of daily living and work	20/20 (100%)
2. Technical barriers to adaptation	Difficulties related to socket fit, donning, and prosthetic suspension	11/20 (55%)
3. Psychosocial dimension and self-esteem	Emotional impact, self-image, and grieving the loss of the limb	13/20 (65%)
4. Discomfort and social stigma	Reports of embarrassment, stares, and accessibility barriers in public spaces	9/20 (45%)
5. Abandonment and continuation of use	Factors favoring the continuation or discontinuation of prosthetic use	8/20 (40%)

On the other side, the content analysis [16] was also carried out on the users' responses, which revealed five central categories, described in Table III below:

TABLE III. CONTENT ANALYSIS (PROSTHESIS USERS)

Category	***Definition***	***Relative frequency***
1. Return to routine and functionality	Recovery of daily, work-related, physical, and leisure activities	20/20 (100%)
2. Physical discomfort and bodily limitations	Pain, heat, friction, and restricted movement are associated with the prosthetic socket	16/20 (80%)
3. Specific functional limitations	Movements or activities that remain inaccessible even after adaptation	13/20 (65%)
4. Appearance and cosmetic covering	Aesthetic perception of the device and desire for covering or customization	8/20 (40%)
5. Social discomfort and stigma	Stares, comments, jokes, and embarrassing situations in public spaces	10/20 (50%)

The following section will present in detail all the information and categories identified in the Content Analysis, exposing and discussing the main points of the data collection for both groups.

B. Perceptions of Health Professionals

Regarding the 20 professionals' area of practice, the composition was: Physical Therapy (in its various specialties: neurofunctional, motor, respiratory, occupational, rehabilitation) n=16 (80%); Psychology n=3 (15%); Physical Medicine and Rehabilitation/Physiatry, n=1 (5%).

Within this sample, in most cases, there is adequate integration of a multidisciplinary team in the rehabilitation process: 35% (n=7) of respondents stated that a multidisciplinary relationship always exists, 50% (n=10) stated that it frequently exists, and rarely (n=1, 5%). Adding together the categories Always and Frequently, 85% of professionals report satisfactory multidisciplinary integration in their practice contexts. However, 5% report rare integration, which shows that some patients do not undergo full rehabilitation, given the universal recognition of the importance of integrated care in rehabilitation.

Thus, the question of whether prosthesis use improves the patient's quality of life obtained positive unanimity: all 20 respondents (100%) said yes, with different qualifications, which is consistent with the literature, which documents significant gains in mobility, independence, and psychological well-being after prosthetic fitting [17].

The justifications converged around: return to mobility and functional independence (n=18, 90%); improved self-esteem (n=15, 75%); social and occupational reintegration (n=9, 45%); reduced caregiver burden (n=2, 10%); and general psychological well-being (n=4, 20%). It is observed, therefore, that mobility and functional independence are

perceived as the main gains in the rehabilitation relationship with prosthesis use. Two professionals added that improved quality of life depends on the conditions of prosthesis use. One physical therapist pointed out that, in some cases, functionally independent patients using crutches may perform better without the prosthesis. A psychologist, in turn, stressed that the benefits depend on proper fitting of the device and effective training, showing that the results are directly related to the quality of the rehabilitation process.

However, when asked about the challenges, the main reports from professionals were associated with problems related to socket fit and suspension (n=4, 20%), difficulty putting on and managing the prosthesis by the patient (n=5, 25%), acceptance of the prosthesis itself and the new body image (n=3, 15%), adjustment of patient expectations regarding use (n=3, 15%), and comfort and pain (n=3, 15%).

The difficulty in putting on and managing the prosthesis shows that its use requires motor coordination, strength, and technique, representing a challenge for some users. In addition, problems with fit and suspension directly compromise the user experience, since the socket constitutes the main interface between the limb and the prosthesis.

In addition, acceptance of the prosthesis itself and of the new body image, as well as the adjustment of expectations, are challenges that belong to the psychological and psychosocial domain, confirming the need for specialized support in this area. Indeed, Hogan and MacLachlan [9] point out that factors such as emotional adaptation, self-image, coping with amputation, and expectations regarding prosthesis use have a direct influence on satisfaction, device adherence, and the overall rehabilitation process.

As for prosthesis abandonment, it is a real phenomenon, frequent enough to be systemically relevant. Ten participants (50%) stated that most patients continue using the device; seven (35%) stated that it depends on multiple factors; and three (15%) stated that abandonment is frequent. Adding together the dependent and abandonment responses, 50% of professionals recognize that long-term use is not guaranteed for all patients. Professionals pointed out that the maintenance of prosthesis use depends on the level of amputation, comfort of fit, family support, access to maintenance, and individual patient characteristics, showing that adherence is influenced by the interaction of biomechanical, psychosocial, and contextual factors.

Thus, these data directly corroborate the study developed by Missio and Queiroz [18], which shows that consistent use of assistive technologies is directly related not only to the characteristics of the device but also to the conditions of access, follow-up, comfort, and user adaptation. Taken together, the results reinforce that the maintenance of prosthesis use depends on the interaction between biomechanical, psychosocial, and contextual factors, making abandonment a multifactorial consequence rather than one exclusively related to the device.Regarding the perception of users with a negative relationship with the appearance of the prosthesis, professionals reported: No (n=10, 50%), Yes (n=5, 25%), Initially/rarely (n=4, 20%), and no definite answer (n=1, 5%). Among those who perceive a negative relationship, the following stand out: greater aesthetic dissatisfaction among the elderly (who most often use foam cosmetic covering); dissatisfaction during the phase without cosmetic finishing of the prosthesis; and discomfort in transfemoral prostheses with a rigid knee that forces gait with a straight leg.

Biddiss and Chau [19] show that the appearance of the prosthesis is one of the factors that influence its acceptance and continued use, since aesthetic aspects directly affect the user's perception of their body image and their willingness to use the device in social settings. In addition, Connelly, Jairath, and Bhatt [21] show that user-centered approaches, which incorporate aesthetic preferences, comfort, and personal identity during prosthesis development, favor greater satisfaction, acceptance, and adherence to the device. Together, these studies reinforce that aesthetics transcends the visual dimension, influencing self-esteem, social interaction, and the continued use of assistive technology.

Professionals' perception of users' confidence when using the prosthesis reveals a procedural and nuanced view. Twelve respondents (60%) stated that most patients feel confident; five (25%) stated that this depends on factors such as the level of amputation and length of use; and three (15%) stated that, initially, patients are not confident, but that confidence is acquired progressively. Confidence, therefore, is conceived by professionals as a process, not as an immediate post-fitting state.

The results indicate that confidence in prosthesis use is influenced by the complexity of the amputation and the quality of the rehabilitation process. More proximal amputations require greater biomechanical adaptation, while adequate training favors the development of confidence. This perception is reinforced by the fact that most professionals reported that their patients use the prosthesis partially or fully exposed, suggesting good acceptance of the device in social contexts. Indeed, this corroborates the study by Biddiss and Chau [20], which identifies confidence, functionality, and support during rehabilitation as determining factors for adherence and continued use of the prosthesis.

On the other hand, despite the high level of confidence, eleven of the twenty professionals (55%) reported having received mentions of social discomfort from their patients, while nine (45%) stated they had not received this type of report. The main context of social discomfort described by professionals is excessive staring in public places (six respondents, 30%), followed by difficulties in bathrooms and inaccessible public spaces (three respondents, 15%), problems with public transportation and obstruction of entitled spaces (two respondents, 10%), and leisure situations (two respondents, 10%).

This shows that, even though users wear their prostheses exposed because they feel empowered and confident using the technology, stigma remains present in the gaze of external observers, which has a direct impact on users' perception. Thus, as Vaes, Stappers, and Standaert [22] discuss, the stigma associated with assistive products directly influences the user experience, reinforcing the need for multidisciplinary work that addresses, in addition to functional recovery, the psychosocial aspects that are frequently neglected. In addition, a report from an occupational physical therapist deserves special attention: she described the situation of a patient who sweated heavily in the work environment and encountered physical barriers in the workplace bathrooms to adequately clean the prosthesis. This report illustrates the intersection between architectural accessibility, working conditions, and

daily management of the prosthetic device, which are external factors that directly impact the use experience.

The factors identified by professionals as most influential in prosthesis acceptance converge around the return to locomotion and previous routine (eight respondents, 40%) and the gain in functionality and independence (seven respondents, 35%). An adequate multiprofessional rehabilitation process is mentioned by four respondents (20%), which is significant, as it indicates that some professionals recognize that the quality of the rehabilitation process is, in itself, a determining factor in acceptance, regardless of the characteristics of the prosthetic component.

As for rejection factors, pain and physical discomfort lead (six respondents, 30%), followed by difficulty putting on the device (five respondents, 25%) and unmet high expectations (n=4, 20%). Shame and social stigma (n=3, 15%), the weight of the prosthesis (two respondents, 10%), lack of a support network and depression (n=2, 10%), appearance (n=2, 10%), and inadequate or limiting components (n=2, 10%) complete the list. This diversity of factors confirms the multidimensional nature of the adaptation process and the need for equally multidimensional interventions.

The analysis of the professionals' data reveals a clinical perspective that is, at the same time, committed to patients' well-being and partially limited in its ability to capture the totality of the experience lived by these patients. The unanimity regarding the positive impact of the prosthesis on quality of life is clinically coherent and reflects the device's real value.

The professionals' emphasis on functional factors (such as mobility, independence, difficulty putting on the device, fitting problems) is understandable and has a solid clinical basis. However, it may produce an evaluation bias that underestimates equally important dimensions of the prosthetic experience: aesthetics, sexuality, stigma, and the everyday experience of social discomfort. These are domains that require competencies beyond traditional physical therapy training and that demand the active presence of psychologists, social workers, and other professionals in the rehabilitation process, which, incidentally, in most cases already has a multidisciplinary presence in the care and rehabilitation of these people, but which does not address specific points that make up the totality of the amputee's experience.

The issue of prosthesis abandonment, recognized by 50% of professionals as a possibility and by 15% as a frequent occurrence, is probably the most critical indicator of how well the current prosthetic rehabilitation system meets the real needs of users. If a significant portion of patients abandon the device after fitting, something in the process is failing, whether in component selection, training, post-discharge follow-up, or the management of expectations and social integration after rehabilitation.

C. Amputees perceptions

To understand the experience of using a lower limb prosthesis in all its complexity, it is necessary to recognize the plurality of human trajectories that make up this universe. The sample consists of 20 lower limb prosthesis users, of both sexes. Length of use varies significantly among participants: from 3 months to 32 years. The average length of use is concentrated between two and six years (n=10; 50%), forming a group mostly in an intermediate or advanced stage of adaptation. Some research indicates that this stage is marked by continuous adjustments between the user, the prosthetic component, and the social context, which makes qualitative listening especially relevant for understanding the real needs of this population [12],[14].

When asked what they most enjoy doing while using the prosthesis, users mentioned activities that mainly reflect a return to daily mobility and social life. The most frequent responses were: walking (n=10, 50%), carrying out daily activities such as work, household tasks, and driving (n=6, 30%), exercising such as weight training, going to the gym, and running (n=3, 15%), going out with family, friends, and pets (n=5, 25%), and leisure activities such as the beach, concerts, and dancing (n=3, 15%). Thus, these data indicate that the return to independent locomotion is the main benefit perceived by users. For most, the prosthesis is directly associated with the act of walking – which is directly related to the recovery of autonomy and social participation.

On the other hand, regarding discomfort, of the 20 participating users, (n=16, 80%) reported feeling some kind of physical discomfort when using the prosthesis, while only n=4 (20%) stated they felt no discomfort. The types of discomfort mentioned varied and include: pain in the residual limb/socket area, reported by (n=9, 45%); discomfort in the groin area, by 3 participants (15%); excessive heat and sweating, by 4 participants (20%); lower back pain, by 2 participants (10%); and itching or skin irritation on the residual limb (10%).The situations reported as most uncomfortable were long periods of use (8 respondents, 40%), extensive walks of more than 500 meters or 1 km (4 respondents, 20%), hot weather or use combined with heat-retaining clothing (3 respondents, 15%), and body weight variation that alters the fit (1 respondent, 5%).

The inability to run is by far the most recurrent functional limitation, present in 40% of responses. This reveals that the prosthetic components available, especially those provided by SUS (the Brazilian Unified Health System), do not keep pace, structurally and functionally, with the desire to engage in more intense physical activities, generating frustration and a sense of limitation that goes beyond basic daily use. In addition, there are reports of being unable to use the prosthesis for long periods, where n=13 (65%) reported some factor that prevents use, while 7 (35%) stated they had no such limitation. The main factors cited were: pain in the residual limb or socket (6 respondents, 30%), heat and sweating (n=4, 20%), injuries/wounds on the residual limb (n=2, 10%), lower back pain (n=1, 5%), a thinned residual limb with a loose socket fit (n=1, 5%), and elbow pain making donning difficult (n=1, 5%).

According to [18], [19], factors such as physical discomfort, functional inadequacy, and psychosocial aspects are among the most frequent causes of assistive technology abandonment.Regarding movement, the most frequently cited difficulties were: crouching/bending down (n=7; 35%), running (n=5; 25%), climbing stairs or steps (n=5; 25%), walking quickly (n=4; 20%), crossing the legs or flexing the knee beyond 90° (n=3; 15%), and walking without looking at the ground (n=2; 10%). Only 20% reported having no specific movement difficulties. These biomechanical restrictions, such as crouching, running, and climbing stairs, reflect the technical limitations of commonly used components, directly

affecting quality of life and participation in daily activities [9],[17],[21].

When asked about their level of adaptation to the prosthesis, respondents were distributed as follows: "Yes, completely" (n=8, 40%), indicating full adaptation; "Yes, but I still feel some difficulties" (n=12, 60%), indicating partial adaptation with persistent functional restrictions. No respondent stated they were not adapted, which may reflect either the ongoing rehabilitation process or a possible reluctance to acknowledge non-adaptation.

However, even though some are well adapted to the prosthesis, one of the participants interviewed said that, despite good functional adaptation, recurrent physical discomfort is reported, especially in the groin and ischial regions, requiring temporary workarounds, as described below: “When use is prolonged, the friction can even cause micro-cracks and wounds on the skin, you know? So I have to use sanitary pads as padding and wrap the residual limb with a bandage to compensate for the volume loss, because my residual limb has gotten smaller,” showing that, despite being adapted, they still feel some difficulties in use.

Indeed, the finding that 60% of users still perceive difficulties is clinically relevant. Even among those with decades of use, such as the respondent with 32 years of experience and the one with 25 years, specific limitations are mentioned – which suggests that full adaptation is a relative state, mediated by each individual’s functional expectations and by the technical capabilities of the prosthetic component used.

Regarding the appearance of the prosthesis, 11 participants (55%) stated that something about its appearance bothers them, while 9 (45%) stated they had no aesthetic discomfort. The aesthetic elements cited as sources of discomfort were: the visible tube/pylon (n=5, 25%), the foam cosmetic covering (n=2, 10%), the prosthetic foot, which evokes the amputated limb (n=1, 5%), and a simple/unsophisticated appearance compared to more modern models (n=1, 5%). Two participants (10%) indicated a desire to cover the tube with carbon fiber covers or more aesthetically elaborate materials. Two participants expressed a desire to cover the tube with a carbon fiber cover or more elaborate materials. The literature suggests that the aesthetics of the device directly influences the user’s willingness to expose it publicly and their relationship with their own body image [6],[22].

In contrast, one participant describes the prosthesis as an object of visual appreciation and personal identity, associating it with his affinity for mechanisms: "I think it’s beautiful [...] from seeing the knee movement happening [...]" and, for this reason, chooses to use it uncovered most of the time, exposing the mechanism. This exposure is described as a trigger for positive social interaction, especially with children, attributing observers’ curiosity to the visual asymmetry between the anatomical limb and the metal structure, which he compares to a cyborg. Visual customization emerges as a recurring theme for the interviewee. The participant reported having had the socket of a previous prosthesis painted with a modern design of the Brazilian flag; as the user puts it: “It’s like a tattoo [...] it becomes whatever you like.” He expresses interest in expanding this customization in the future, layering personal images along the prosthesis tube, in the manner of compositions used in tattoos with multiple figures. Indeed, there is a study [23] that observes the stigma and the symbolic dimension of assistive technology equipment, such as the cane, and found that when developing covers for the product, it was possible to observe that the prototypes presented to users helped create a less stigmatizing environment, by combining fashion and product design, showing that

At the same time, the participant identifies two concrete aesthetic problems that, in his assessment, compromise the device’s social acceptance and could be solved at low engineering cost. The first concerns the fixation screws between the socket and the knee, usually made of brass, which oxidize and produce an appearance he describes as “old iron” or “scrap metal”; in addition, the participant raises a second problem regarding the use of long trousers, because: due to the lack of filling at calf and ankle height, the trouser fabric hangs loosely over the thin tube of the structure, creating an effect he vividly compares to a snake: “It ends up looking like a snake [...] that ugly thing.” For this specific use, he expresses a desire for a partial, two-piece cover, as shown in the following statement: “I didn’t want those that look like a leg from start to finish [...] I just wanted that lower cover.”

Overall, to determine how users typically use the product, users were asked how they usually use the prosthesis. The responses were: fully exposed (n=10, 50%), fully covered (n=5, 25%), and partially exposed (n=3, 15%). Two participants chose not to clearly specify their mode of use. Fully exposed use, present in half of the respondents, indicates a posture of acceptance or inevitability regarding the prosthesis’s exposure. On the other hand, the other participants, who indicated they tend not to expose it fully or expose it only partially, demonstrate a stigmatized relationship with the product, since, when asked about their comfort exposing the prosthesis in public, 12 participants (60%) stated they felt comfortable with exposure, while 8 (40%) said they did not feel comfortable. To mitigate this social discomfort, they adopt concealment strategies such as wearing long trousers (n=4, 20%), a long dress (n=1, 5%), and cosmetic covering (n=1, 5%).

Thus, seeking to understand where this uncomfortable relationship arises, participants were asked about social discomfort. Of the 20 participants, 13 (65%) reported at least one situation of social discomfort associated with prosthesis use, while 7 (35%) stated they had not experienced such situations. The reports include: curious or embarrassed stares from strangers (n=9, 45%), jokes or derogatory comments from acquaintances and strangers (n=5, 25%), embarrassment in institutional settings such as banks and priority queues (n=2, 10%), and accessibility difficulties in public transportation and public spaces (n=2, 10%).

Two reports illustrate the severity of social stigma: one participant (16 years of prosthesis use) was made by bank security guards to remove the cosmetic covering for inspection, while another (32 years of use) faced difficulties in priority queues due to not being visually recognized as disabled while wearing trousers. These cases show that stigma extends beyond the social gaze into institutional practices. When asked about frustrations, 55% (n=11) reported at least one specific frustration. The most frequently cited were: falls (n=3; 15%), inability to run or play sports (n=3; 15%), movement limitations that reduce independence (n=2; 10%), components breaking in public (n=1; 5%), and discomfort from not feeling at ease without the prosthesis (n=1; 5%). These data reveal both frustrations related to use and the product itself, as well as impacts on body image, affected by

the amputation procedure and the prosthetic experience [6],[7],[8].

The data reveal a complex and multidimensional picture of the prosthetic experience. Most participants (60%) are in a process of partial adaptation, living with physical and social limitations that affect quality of life and self-esteem. Physical discomfort is frequent in this sample (80%), suggesting that the components used by these particular participants may not fully meet their individual needs.[18], [20].

The social dimension of the experience is equally significant: 65% reported social discomfort and 65% stated they felt less confident in at least one public situation, showing that prosthetic adaptation involves a complex process of identity renegotiation and coping with stigma [9], [22]. Four central reflections emerge from these analyses: The technical limitation (1) generates discomfort that feeds the reduced-use/adaptation cycle (2); aesthetics (3) modulates the psychosocial impact of that cycle; and all of this unfolds within a social/institutional context that remains inadequate (4), compounding both insecurity and exclusion.

D. Comparison of perceptions between amputees and professionals

Comparing both groups in the Content Analysis, both converged on functionality/independence as the central value of prosthesis use (100% users / 100% professionals). Here, both categories ("Return to routine and functionality" and "Functionality and independence") are conceptually equivalent, both capturing the recovery of activities of daily living, work, and mobility, which supports a direct comparison [5], [9], [21]. For the remaining categories, equivalence is only partial, and comparisons are presented with that caveat made explicit for each pair: Physical/technical barriers were reported by 80% of users (captures pain, heat, friction associated with the socket) and 55% of professionals. Although related, the categories differ conceptually; therefore, the comparison indicates thematic convergence around barriers to comfortable use rather than direct numerical equivalence. Figure 2 below shows and compares both reports:

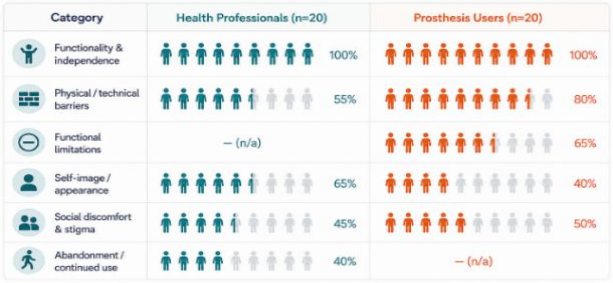


Fig. 2. Integrated content analysis of health professionals' and prosthesis users´ responses in the Content Analysis [7]. Each icon represents 10% of respondents in each group who mentioned the theme (n=20).

However, while 100% of professionals state that the prosthesis improves quality of life, only 40% of users reported being fully adapted. The remaining 60% still report difficulties, and 80% report physical discomfort. This divergence suggests that the clinical perception of therapeutic success may be disconnected from the user's subjective experience. Professionals, in general, assess outcomes based on observable functional metrics, while users perceive adaptation as a process that also encompasses the psychosocial dimension and the absence of pain.

For this reason, pain and physical discomfort appear as a central barrier from both the users' and the professionals' perspectives. Among users, 80% report discomfort; among professionals, 30% identify pain as the main factor in prosthesis rejection and 20% as the greatest challenge in adaptation. This convergence indicates that prosthetic component design is a critical point requiring improvement. Inadequate fit is not merely a biomechanical issue: it compromises the user's use, confidence and integration.

Social discomfort appears consistently in both groups: 65% of users reported situations of social discomfort, and 55% of professionals received similar reports from their patients. This convergence shows that the stigma associated with the prosthesis is a reality recognized by both sides [22]. The gaze of others is the main trigger of discomfort in both users' and professionals' reports, though asymmetrically detailed. Users cite specific situations (banks, jokes from acquaintances, priority queues), while professionals generalize it as "excessive staring," suggesting only partial access to patients' social experience. [6], [7]. Confidence in using the prosthesis publicly is a gradual, mutually recognized process. Professionals report 60% of users as confident, developing over time with use and training. Users similarly report 60% comfort exposing the prosthesis, yet 65% cite situations of lower confidence, suggesting confidence is contextual rather than absolute. Specific situations (stairs, ramps, crowds, institutional settings) remain sources of insecurity even for experienced users.

Looking at design aspects, both the user and professional data point to prosthetic product design as a central variable. Users mention the visible tube/pylon (25%), the foam covering (10%), and the prosthetic foot (5%) as sources of aesthetic discomfort. Professionals mention the donning system, fit, and weight as rejection factors. These findings converge to indicate that prosthetic design, in its functional, ergonomic, and aesthetic dimensions, is a fertile field for innovation that directly impacts users' quality of life.

The comparative analysis reveals five key implications: (1) physical discomfort (80% of users) and donning difficulties (25% of professionals) point to the need for more adaptable, intuitive prosthetic sockets [18], [20]; (2) high rates of social discomfort and aesthetic limitations (55-65%) indicate that prosthetic aesthetics is a clinically relevant factor, not merely cosmetic [22], [23]; (3) a gap between some users' exercise-related expectations and the components available to them was reported within this sample, whether this reflects a broader pattern in SUS-provided components nationally cannot be established from this study and would require research with a representative sample [19], [20]; (4) reported abandonment (15%) aligns with frustration, falls, and partial adaptation rates, suggesting abandonment results from cumulative, unmanaged physical, functional, and social difficulties [10], [19]; and (5) the absence of spontaneous mentions of sexuality, despite its clinical relevance to body image and intimacy, suggests it remains a taboo topic in interviews [6], [7].

IV. RECOMMENDATIONS FOR BIOMEDICAL ENGINEERING

Content Analysis [16] identified patterns that translate into concrete recommendations for lower-limb prosthesis development (Table IV) for Biomedical Engineering researchers and professionals:

TABLE IV. RECOMMENDATIONS FOR BIOMEDICAL ENGINEERING PROFESSIONALS

design requirement	empirical basis (frequency/source)	target outcome
1. Thermal management in socket/liner design	Heat/sweating cited as discomfort by 20% of the users.	Reduce heat retention; improve ventilation.
2. Socket pressure distribution at residual limb interface	Pain in residual limb cited by 45% of the users and 30% of the professionals, leading rejection factor.	Paticient-specific, adaptative socket contouring.
3. Aesthetic personalization / modular cosmetic covering	Visible tube discomfort, foam-covering dissatisfaction.	Modular, customizable partial/full cosmetic covers.
4. Participatory, multidisciplinary design process	Gap between professional (100% report QoL) improvement and user perceptions.	User feedback loops embedded in component development and rehabilitation protocol.
5. Expanded functional/ mobility range	Crouching/bending running, stairs, cited by users as movement limitations	Components enabling wider range of motion beyond current baseline.

V. CONCLUSION

The joint analysis of user and professional data allows for an exploratory look at the multidimensional experience of lower-limb prosthesis use. While the device appears to hold significant therapeutic value for restoring mobility and independence, the accounts analyzed suggest that this potential may be compromised by physical (socket discomfort, pain, heat), functional (component limitations), aesthetic, and social (stigma, stares, inaccessibility) barriers.

The convergence between users and professionals around mobility as a central value, physical discomfort as a recurring barrier, and social discomfort as a dimension present across accounts points to the relevance of addressing these dimensions in an integrated way in future design interventions and rehabilitation protocols. Conversely, the divergence observed regarding the degree of adaptation, may indicate the need for assessment tools more sensitive to users' subjective, lived experience, complementing the functional parameters traditionally observed.

These findings may contribute to Biomedical Engineering by broadening understanding of demands related to lower-limb prosthesis development and use. Results suggest technological development could benefit from going beyond mechanical improvements to also consider comfort, ergonomics, aesthetics, usability, and user experience. These initial findings may also support future research on improving rehabilitation, defining criteria for new prosthetic components, and applying user-centered design, approaches that could later be examined for their relationship with adherence to use, quality of life, and social participation of people with amputations.

VI. AI USAGE DECLARATION

This research was reviewed with the assistance of AI-based tools (Claude) for grammatical correction and textual cohesion. All scientific content, analysis, and conclusions remain the sole responsibility of the authors.

VII. ACKNOWLEDGMENTS

The publication of this work was supported by CAPES.

REFERENCES

[1] Brasil. Lei nº 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência. Diário Oficial da União, Brasília, DF, 7 jul. 2015.

[2] BRASIL. Decreto nº 3.298, de 20 de dezembro de 1999. Regulamenta a Lei no 7.853, de 24 de outubro de 1989, dispõe sobre a Política Nacional para a Integração da Pessoa Portadora de Deficiência, consolida as normas de proteção, e dá outras providências. Diário Oficial da União, Brasília, DF, 21 dez. 1999.

[3] World Health Organization, World Report on Disability. Geneva: WHO Press, 2011.

[4] L. R. Oliveira et al., "Análise dos fatores de risco relacionados às amputações maiores e menores de membros inferiores em hospital terciário," Rev. Vascular e Angiology Brasileira, vol. 37, nº 3, pp. 204–210, 2018.

[5] A. C. Oliveira et al., "Percepção e qualidade de vida de amputados de membro inferior submetidos à reabilitação fisioterápica," Rev. Interdisciplinar de Promoção da Saúde, vol. 4, nº 1, pp. 25–32, 2021.

[6] G. B. Galvãn and M. L. T. M. Amiralian, "Corpo e identidade: reflexões acerca da vivência de amputação," Estudos de Psicologia (Campinas), vol. 26, nº 3, pp. 431–439, 2009.

[7] I. B. Wald and A. R. C. Álvaro, "Aspectos psicológicos da cirurgia de amputação," Psicologia USP, vol. 20, nº 2, pp. 325–334, 2009.

[8] B. Rybarczyk, J. J. Nicholas, and D. L. Nyenhuis, "Coping with a leg amputation: integrating research and clinical practice," Rehabilitation Psychology, vol. 42, nº 3, pp. 241-256, 1997.

[9] O. Horgan and M. MacLachlan, "Psychosocial adjustment to lower-limb amputation: a review," Disability and Rehabilitation, vol. 26, nº 14-15, pp. 837-850, 2004.

[10] T. R. Chamlian and D. C. Varanda, "Abandono de prótese em amputados de membro inferior: fatores determinantes," Acta Fisiatrica, vol. 21, nº 2, pp. 54-58, 2014.

[11] J. W. Creswell and V. L. Clark, Pesquisa de métodos mistos, 2nd ed. Porto Alegre: Penso, 2017.

[12] M. C. S. Minayo, O desafio do conhecimento: pesquisa qualitativa em saúde, 14th ed. São Paulo: Hucitec, 2014.

[13] J. Vinuto, "A amostragem em bola de neve na pesquisa qualitativa: um debate em aberto," Temáticas, vol. 22, nº 44, pp. 203-220, 2014.

[14] S. Kvale and S. Brinkmann, InterViews: learning the craft of qualitative research interviewing, 2nd ed. Thousand Oaks: Sage, 2009.

[15] A. C. Gil, Métodos e técnicas de pesquisa social, 7th ed. São Paulo: Atlas, 2019.

[16] L. Bardin, Análise de conteúdo. São Paulo: Edições 70, 2016.

[17] R. Gzil et al., "Quality of life in lower-limb amputees: a systematic review," Disability and Rehabilitation, vol. 43, nº 1, pp. 1-15, 2021.

[18] M. Missio and L. F. de Queiroz, "Tecnologias assistivas: aspectos que influenciam na assiduidade e no abandono dos recursos," *Acta Fisiátr.*, vol. 25, no. 4, pp. 185–190, Dec. 2018, doi: 10.11606/issn.2317-0190.v25i4a163858.

[19] E. A. Biddiss and T. Chau, "The roles of predisposing characteristics, established need, and enabling resources on prosthesis use and abandonment," Disability and Rehabilitation: Assistive Technology, vol. 2, nº 2, pp. 71-84, 2007.

[20] L. M. Connelly, I. Jairath, and A. Bhatt, "User-centered design for prosthetic devices: a systematic review," Prosthetics and Orthotics International, vol. 45, nº 3, pp. 219-229, 2021.

[21] K. Vaes, P. J. Stappers, and A. Standaert, "Measuring product-related stigma in design," in *Proc. Design Research Society 50th Anniversary Conf. (DRS 2016)*, Brighton, U.K., Jun. 2016, pp. 3329–3348.

[22] Y. W. Rodrigues, *Design para o envelhecimento: a dimensão simbólica na superação do estigma em equipamentos de auxílio*, Ph.D. dissertation, Univ. Aveiro, Aveiro, Portugal, 2021.

[23] I. B. F. de Minayo, "Amostragem e saturação em pesquisa qualitativa: consensos e controvérsias," Revista da Associação Nacional de Pós-Graduação e Pesquisa em Educação, vol. 17, no. 1, Mar. 2017.

Noninvasive Peripheral Electrical Stimulation in Parkinson's Disease: A Pilot Feasibility Study with Blinded Video-Based Motor Assessment

1st Danielle Liquori
Biomedical Engineering Program
(UFRJ)
Rio de Janeiro, Brazil
danielleliquori@peb.ufrj.br

2nd Roberto Santos
Clementino Fraga Filho University Hospital
(UFRJ)
Rio de Janeiro, Brazil
r.santos.ufrj@gmail.com

3rd Evelyn Juliette Pesántez Vera
Biomedical Engineering Program
(UFRJ)
Rio de Janeiro, Brazil
epesantezv@peb.ufrj.br

4th Pedro Sebastian Silverio Cevallos
Biomedical Engineering Program
(UFRJ)
Rio de Janeiro, Brazil
psilverio@peb.ufrj.br

5th Ana Lucia Rosso
Clementino Fraga Filho University Hospital
(UFRJ)
Rio de Janeiro, Brazil
anarosso@gmail.com

6th Carlos Julio Tierra-Criollo
Biomedical Engineering Program
(UFRJ)
Rio de Janeiro, Brazil
carjulio@peb.ufrj.br

Abstract*—Parkinson's disease (PD) is a progressive neurodegenerative condition clinically delineated by primary motor deficits, such as rigidity, bradykinesia, and tremor. Clinical management aimed at mitigating these symptoms predominantly relies on pharmacological interventions—chiefly dopamine replacement therapy—alongside surgical procedures, including deep brain stimulation or targeted tissue ablation. Nevertheless, a significant patient cohort is excluded from surgical eligibility, while others experience debilitating medication-induced side effects. Consequently, recent research has focused on non-invasive peripheral electrical stimulation as a complementary therapeutic modality, with particular emphasis on tremor suppression. This pilot study aimed to evaluate the feasibility of a protocol combining noninvasive peripheral electrical stimulation and blinded video-based assessment of motor performance in PD. Eleven participants were randomized to receive either peripheral transcutaneous electrical nerve stimulation (TAPS) or transcutaneous somatosensory sine wave electrical stimulation (t-SSES) stimulation. Six participants received pulse current stimulation and five received sinusoidal current stimulation. Seven selected motor tasks were assessed before and after stimulation; the tasks are included in the Movement Disorder Society-Unified Parkinson's Disease Rating Scale part III (MDS-UPDRS Part III). The assessment was realized by video-recorded. The videos were independently scored in a blinded manner by a neurologist specializing in movement disorders. The videos were randomized before evaluation to ensure blinded scoring. The protocol was successfully implemented, with a satisfactory number of participants, good tolerability of the intervention, blinding of the evaluators, and no significant technical or safety issues. The exploratory video assessment, conducted in a blinded manner, indicated a trend toward a reduction in scores for most patients, and there was only one case of an increase in scores. The findings support the feasibility of conducting a blinded study to investigate the effects of t-SSES and TAPS on motor symptoms in patients with Parkinson's disease using the proposed experimental protocol.

***Keywords*—Parkinson's disease, noninvasive peripheral electrical stimulation, MDS-UPDRS Part III, blinded video assessment.**

I. Introduction

Parkinson's disease (PD) is the most common form of parkinsonism and is a progressive neurodegenerative disorder characterized by the loss of dopaminergic neurons in the substantia nigra. PD has a prevalence of approximately 1% to 2% in populations exceeding 65 years of age [1]. Clinically, PD is characterized by classical motor symptoms, including rigidity, resting tremor, bradykinesia, postural instability, and gait disturbances [2; 3; 4]. Pharmacological interventions constitute the primary therapeutic modality for PD, with levodopa-carbidopa—a metabolic dopamine precursor paired with a decarboxylase inhibitor—serving as the gold-standard, first-line therapy [5]. A major limitation of long-term pharmacological management is the progressive loss of efficacy over time, which frequently necessitates dosage

escalation [6]. Ultimately, while these agents alleviate symptomatic burden, they lack disease-modifying capabilities to halt neurodegeneration and may induce intolerable adverse effects in a subset of patients [6]. Surgical intervention serves as a viable therapeutic alternative for a select cohort of patients with Parkinson's disease (PD). Current neurosurgical modalities encompass cellular ablation of aberrant, hyperactive neuronal populations and Deep Brain Stimulation (DBS) [7; 8, 9; 10]. Furthermore, widespread clinical translation of these surgical approaches is constrained by their high financial burden and invasive nature.

These limitations have motivated the investigation of non-invasive therapeutic approaches. Among the available non-invasive complementary modalities—designed to supplement rather than replace optimal pharmacotherapy—are rhythmic interventions, which can be administered via visual, auditory, or vibrotactile channels; the latter specifically engages the somatosensory system.

II. Somatosensory Stimuli

Somatosensory modulation can be delivered via either electrical or mechanical modalities. Mechanical stimulation has been extensively investigated in patients with Parkinson's disease (PD) utilizing wearable biomedical devices [11]. Currently, commercially available systems employing this therapeutic approach include the CUE1 (Charco Neurotech Ltd.) and the VILIM ball (Vilimed) [12; 13].

A. Peripheral Electrical Stimulation

The primary modalities of peripheral electrical intervention encompass Functional Electrical Stimulation (FES), Sensory Electrical Stimulation (SES)—operationalized at intensities below the motor threshold—and Transcutaneous Electrical Nerve Stimulation (TENS). TENS preferentially activates Aβ sensory fibers to modulate nociceptive processing via the gate control theory [14]. Afferent signals propagated along Aβ fibers ascend through the dorsal column–medial lemniscus pathway to the thalamus [15], where they may influence thalamocortical networks involved in tremor generation [16, 17].

Extensive literature has investigated the therapeutic efficacy of these electrical stimulation modalities on parkinsonian tremors. Notably, the commercially available Cala Trio device employs by delivering alternating transcutaneous electrical nerve stimulation (TAPS-Transcutaneous Afferent Patterned Stimulation) to the sensory branches of the radial and median nerves, effectively achieving significant tremor suppression [18].

The TAPS technique was based on electrophysiological recordings [19,20] obtained using DBS electrodes, which demonstrated that peripheral nerve stimulation evokes somatosensory potentials within the VIM nucleus and other tremor-related brain regions.

In clinical practice, TAPS consists of delivering transcutaneous electrical stimulation to the median and radial nerves of the dominant upper limb for a 40-minute session. Commercial devices operate under standardized parameters based on charge-balanced, biphasic pulse trains, featuring a phase duration of 300 µs, an interphase interval of 50 µs, and a repetition frequency of 150 Hz. The core mechanism of this method relies on alternating the stimulus delivery between the median and radial nerves, controlled by an alternating stimulation frequency individually calibrated to each patient's tremor frequency [19].

Pulsatile electrical stimulation is widely used to activate large-diameter myelinated Aβ afferent fibers, thereby eliciting tactile sensations [21, 22]. However, increasing the stimulation intensity may also recruit smaller-diameter Aδ and C fibers [23, 24]. Therefore, it is reasonable to hypothesize that more selective stimulation of the somatosensory pathways could have a beneficial effect on motor symptoms.

Transcutaneous somatosensory sine wave electrical stimulation (t-SSES) is a modality of noninvasive peripheral stimulation that has been scarcely investigated as an intervention for PD. Unlike rectangular pulse stimulation, t-SSES allows preferential activation of different sensory fiber populations by varying the stimulation frequency. This frequency-dependent selectivity has been attributed to differences in the electrophysiological properties of sensory nerve fibers, including their membrane dynamics and ion channel activation characteristics [23, 25]. Previous studies have reported that different activation patterns of somatosensory nerve fibers depending on the frequency of the sinusoidal stimulus: lower frequencies, such as 1 Hz and 5 Hz, preferentially stimulate C-type fibers; a frequency of 250 Hz stimulates Aδ-type fibers; and finally, high frequencies, such as 2 kHz and 3 kHz, preferentially activate Aβ-type fibers [26; 27; 28; 24].

Given the limited evidence regarding t-SSES in Parkinson's disease and the absence of standardized blinded assessment protocols, this pilot study was designed to evaluate the feasibility of a protocol combining t-SSES with blinded video-based assessment of motor performance.

In this study, a novel paradigm with t-SSES of 3 kHz was implemented, based on the TAPS stimulation framework, for motor symptoms.

III. Materials and Methods

The experimental protocol was approved by the Institutional Ethics Committee (CAAE registration no. 58744222.2.1001.5257).

This pilot study enrolled 11 participants diagnosed with Parkinson's disease (PD) and exhibiting clinical tremor.

Participants were eligible if they were 40–75 years old, had a diagnosis of idiopathic Parkinson's disease established by a movement disorders specialist according to the United Kingdom Parkinson's Disease Society Brain Bank criteria, had a Hoehn and Yahr (H&Y) [2] stage between 1 and 4, and provided written informed consent (or consent provided by a legal representative when applicable).

Exclusion criteria included Parkinson's disease dementia; musculoskeletal or orthopedic disorders affecting upper-limb motor performance; implanted cardiac pacemakers or deep brain stimulation (DBS) systems; neurological disorders other than Parkinson's disease; the use of medications known to significantly impair attention or motor performance and failure to perceive the electrical stimulation.

Participants were randomly assigned to one of the two intervention groups using simple randomization. Allocation was performed by drawing one of two identical folded slips of paper labeled "TAPS" or "t-SSES" from an opaque container which was thoroughly shaken before each draw so that the allocation could not be predicted.

After each draw, both slips were returned to the container before the next participant was enrolled, ensuring that each participant had an equal probability of being assigned to either intervention. Eleven participants were randomized TAPS group, n = 6; t-SSES group, n = 5). As this was a pilot feasibility study, no formal sample size calculation was performed. The sample size was determined based on participant availability and the feasibility of conducting the experimental protocol within the study period.

During the experimental sessions, participants were seated in an ergonomic chair with their forearms resting on an adjacent table to minimize muscle fatigue. Experiments were conducted in a dedicated laboratory environment at the Clementino Fraga Filho University Hospital (UFRJ), maintained under controlled thermal conditions (24 ± 1°C). Motor performance was assessed before and after the stimulation protocol using Part III of the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS III) [29]. Specifically, seven upper-limb motor tasks were evaluated: 3.4 - finger tapping, 3.5 - hand movements, 3.6 - pronation-supination movements, 3.15 - postural tremor of the hands, 3.16 - kinetic tremor of the hands, 3.17 - rest tremor amplitude and 3.18 - constancy of rest tremor [30]. The recorded video data were subsequently evaluated by a blinded neurologist specializing in movement disorders.

All assessments were performed while participants were in the ON medication state to ensure clinical stability and to minimize variability in motor performance during testing. Although dopaminergic medication may reduce baseline tremor severity and consequently decrease the opportunity to detect additional treatment effects, this approach reflects routine clinical practice and improves the consistency of repeated motor assessments.

A. Video recording

Video recording was performed using two high-resolution mobile cameras operating at 1080p resolution and a frame rate of 30 frames per second (fps). The devices were mounted on independent tripods; the first camera was aligned perpendicular to the participant's sagittal plane, while the second camera provided a frontal view. Crucially, the tripods were mechanically isolated from both the examination table and the chair. This structural decoupling minimized the influence of physiological tremors or subject movements on camera stability, reducing the likelihood of motion artifacts in the recorded videos.

The stimulation electrodes remained in the same position during both the pre- and post-stimulation motor assessments. Because the electrodes were visible in all recordings, their presence did not disclose whether a video had been acquired before or after stimulation, thereby helping preserve evaluator blinding.

B. Tremor Frequency Quantification

To quantify postural tremor frequency before and after the neuromodulation protocol, kinematic data were acquired using a triaxial accelerometer positioned on the dorsum of the stimulated hand. During data collection, participants maintained their upper limb extended horizontally for 20 s, allowing the sensor to record tremor amplitude and frequency dynamics. The dominant tremor frequency was estimated using power spectral density (PSD) analysis implemented in LabVIEW.

The peak tremor frequency derived from this baseline assessment was subsequently integrated into the patient-specific stimulation protocol parameters [18; 30].

C. Stimulation Electrodes

Peripheral electrical stimulation was delivered via two pairs of self-adhesive gel electrodes (1 cm diameter) targeting the radial and median nerves. The radial electrodes were positioned on the lateral-dorsal aspect of the wrist, adjacent to the anatomical snuffbox, while the median electrodes were placed on the palmar surface [18]. The center-to-center distance between the cathode and anode of each electrode pair was maintained at 2 cm. Compared with the original protocol [18], the inter-electrode distance was adjusted to accommodate the smaller electrode dimensions used in the present study. This modification likely increased the local current density while maintaining the stimulation over the same anatomical target, which should be considered when comparing the present findings with those obtained using the original protocol. To maintain comparable electrode geometry across participants, the edge-to-edge distance between the electrodes was adjusted according to the participant's wrist circumference, classified as: small (13.5–15.4 cm circumference; 2.5 cm separation), medium (15.5–17.4 cm circumference; 3 cm separation), or large (17.5–19.5 cm circumference; 3.5 cm separation). The interelectrode distance values (cathode and anode) were readjusted due to the reduction in electrode surface area compared to the protocol established by reference [18]. Stimulation was applied unilaterally to the upper limb presenting the higher tremor severity; in cases of symmetrical tremor presentation, the dominant hand was selected [18]. The duration of the stimulation was 40 minutes.

The stimulation was generated by Mestin equipment, a stimulator developed at *Universidad Politécnica Salesiana* (UPS) in Ecuador in collaboration with Coppe/UFRJ. The stimulator is equipped with two independent channels, each of which is capable of generating arbitrary waveforms, including sinusoidal, square, or custom waveforms. The stimulation parameters are configured and transmitted to the stimulator via an Ethernet connection, employing a LabVIEW interface [31].

D. Stimulation Parameters

Participants were randomly allocated to two experimental groups to receive distinct peripheral electrical stimulation paradigms: conventional TAPS or an adapted TAPS protocol using a 3 kHz sinusoidal waveform instead of biphasic rectangular pulses. The conventional TAPS paradigm consisted of charge-balanced, biphasic square pulses with a phase duration of 300 µs and an interphase interval of 50 µs [18], delivered at a pulse repetition rate of 150/s (150 Hz).

Prior to establishing the stimulation intensity, the baseline current sensory threshold (ST) was characterized for each participant. Stimulation intensity was customized for each participant by escalating the current in 0.25 mA increments until the ST was reached in the hand or digit dermatomes. The target intensity was defined as the maximum tolerable paresthesia intensity below the motor threshold and did not induce discomfort, ensuring the complete absence of muscle contractions. For both cohorts, stimulation was alternated between the median and radial nerves at a switching frequency locked to the patient's specific tremor frequency.

E. The randomization of videos.

The videos were randomized using a computer-generated random number sequence. Each number corresponds to a specific video. All videos were anonymized and presented in randomized order to an experienced neurologist specializing in movement disorders. The evaluator was blinded to the stimulation condition and assessment time point (pre- or post-stimulation). Each video was independently scored according to the MDS-UPDRS Part III criteria.

F. Statistical analysis

Given the exploratory nature of this pilot study, the limited sample size resulted in insufficient statistical power to detect small-to-moderate effects. Accordingly, the results are presented primarily through descriptive statistics, with inferential tests performed for exploratory purposes and interpreted with appropriate caution.

Each item of the MDS-UPDRS Part III is scored on an ordinal scale ranging from 0 to 4, allowing changes in performance for each task to be classified as improved (decrease in score), unchanged, or deteriorated (increase in score). In addition, because only seven MDS-UPDRS Part III tasks were evaluated, patients were also classified according to a composite score obtained by summing the scores of these seven items rather than the complete Part III score.

A partial MDS-UPDRS Part III score was calculated by summing the scores of the selected motor items (3.4, 3.5, 3.6, 3.15, 3.16, 3.17, and 3.18) evaluated exclusively on the most affected upper extremity. Frequencies were reported for each group, both at the item level and for the composite score.

Results are presented as the median and interquartile range, together with the numbers of patients who improved, remained unchanged, or deteriorated for each individual task and for the composite score.

Within-group comparisons of pre- and post-stimulation values were performed using the Wilcoxon signed-rank test. Between-group comparisons of change scores (Δ = post-stimulation − pre-stimulation) were performed using the Mann-Whitney U test.

Statistical analyses were performed in R (version R 4.5.1) using the wilcox.test() function, which excludes zero-difference pairs from the signed-rank test.

IV. Results

A. Feasibility outcomes

Recruitment - Patients were recruited at the HUCFF outpatient clinic and neurology department while awaiting consultation with a physician. The principal investigator then proceeded to approach the patients and their companions in order to deliver the explanation of the experiment. The most salient points of this explanation encompassed the nature, safety and comfort of the stimulation; the duration of the stimulation itself and the total duration of the experimental session; and the purpose of the experiment. Patients who consented to participate in the study were escorted to the designated experiment room following their clinical consultation.

Sixteen patients were invited to participate in the study; however, five were excluded. Two met at least one of the predefined exclusion criteria, whereas three were excluded because the stimulation intensity could not be adjusted to the target level relative to the sensory threshold required by the protocol. Eleven participants were included and randomized (Group TAPS, n = 6; Group t-SSES, n = 5). All participants were on stable dopaminergic medication regimens. The TAPS group exhibited a mean age of 64.8 ± 6.5 years and a mean H&Y stage of 2.0 ± 0.6. The t-SSES group presented a mean age of 67.8 ± 5.9 years and a mean H&Y stage of 2.1 ± 0.7.

Protocol adherence - All remaining participants completed the stimulation session and the pre- and post-intervention motor assessments without any deviations from the protocol. Consequently, no requests were made to disrupt the experimental session. It was also found that there were no complaints regarding the accommodation, the room temperature, or the duration of the session.

Safety and tolerability - Prior to the commencement of each experimental session, the electrostimulator was tested using an oscilloscope to verify the waveform and voltage. In the event of an offset being detected in the delivered current, the stimulator was meticulously calibrated and retested to ensure its optimal functionality. The mean stimulation intensities applied to the median and radial nerves in the TAPS group were 1.5 x ST (2,752.6 µA) and 1.5 x ST (2,669.7 µA), respectively. For the t-SSES group, the corresponding average intensities for the median and radial nerves were 1.2 x ST (4,195.4 µA) and 1.3 x ST (4,329.7 µA), respectively. The predominant sensory experiences reported by patients included tingling, vibration, sting, and needling, yet these were generally characterized as mild or tolerable. No reports were received of discomfort.

Blinding - The blinded video assessment was successfully completed, as the evaluating physician was unaware of the type of stimulation administered to the participant, and whether the assessment video corresponded to the pre- or post-stimulation period. The neurologist did not participate in the experimental sessions. The videos did not contain the participants' names or the date on which the experiment was conducted.

Technical feasibility - The stimulation system operated in accordance with the intended protocol throughout the study. The video recordings were suitable for blinded assessment.

B. MSD-UPDRS Part III Results

Table 1 summarizes the number of patients who showed improvement, worsening, or no change in scores for each task in Part III of the MDS-UPDRS. A trend toward improvement was observed in most of the tasks assessed in both groups, particularly in the Hand Movements task in both groups and in the TAPS group for the Finger Tapping task. In the t-SSES group, no patient showed improvement in the "Postural Hand Tremor" task, while in the TAPS group, no improvement was observed in the "Resting Tremor Amplitude" task. In both groups, all tasks included patients whose scores remained unchanged. A worsening in scores was observed for the "Hand Pronation-Supination Movements" and "Resting Tremor Amplitude" tasks in both groups.

In addition, patients in the TAPS group showed a worsening in scores for the "Hand Postural Tremor" and "Hand Kinetic Tremor" tasks.

Table 1. Number of participants showing improvement, no change, or deterioration in MDS-UPDRS Part III scores for each analyzed motor task. TAPS: transcutaneous afferent patterned stimulation; T-SSES: transcutaneous somatosensory sine wave electrical stimulation

Tasks	Groups	Improvement	No Change	Deterioration
FINGER TAPPING	*TAPS*	3/6	3/6	0
	t-SSES	1/5	4/5	0
HAND MOVEMENTS	*TAPS*	4/6	2/6	0
	t-SSES	3/5	2/5	0
PRONATION-SUPINATION MOVEMENTS OF HANDS	*TAPS*	1/6	4/6	1/6
	t-SSES	1/5	3/5	1/5
POSTURAL TREMOR OF THE HANDS	*TAPS*	1/6	4/6	1/6
	t-SSES	0	5/5	0
KINETIC TREMOR OF THE HANDS	*TAPS*	1/6	3/6	2/6
	t-SSES	2/5	3/5	0
REST TREMOR AMPLITUDE	*TAPS*	0	4/6	2/6
	t-SSES	1/5	3/5	1/5
CONSTANCY OF REST TREMOR	*TAPS*	1/6	4/6	1/6
	t-SSES	1/5	4/5	0

Based on the change in the partial MDS-UPDRS Part III score (Δ = post − pre), improvement was observed in four of six patients (67%) in the TAPS group and in four of five patients (80%) in the t-SSES group. The remaining participants showed no change, and only one patient exhibited a worsening of the partial score following stimulation (Table 2).

Table 2. Individual pre- and post-stimulation partial scores (MDS-UPDRS Part III) and the corresponding change scores (Δ = post − pre). TAPS: transcutaneous afferent patterned stimulation; T-SSES: transcutaneous somatosensory sine wave electrical stimulation

Patient	Group	Pre	Post	Δ (Post − Pre)
P1	*TAPS*	4	2	-2
P2		15	15	0
P3		4	2	-2
P4		9	3	-6
P5		1	0	-1
P6		3	9	6
P7	*t-SSES*	10	1	-9
P8		3	2	-1
P9		21	19	-2
P10		4	4	0
P11		5	4	-1

Within-group comparisons of the partial MDS-UPDRS Part III scores using the Wilcoxon signed-rank test revealed no statistically significant differences between the pre- and post-stimulation assessments in either group (TAPS: V = 10.5, p = 0.4962; t-SSES: V = 10, p =0.09751), as shown in Table 3.

Table 3. Results of the Wilcoxon signed-rank test comparing pre- and post-stimulation partial MDS-UPDRS Part III scores within the TAPS and t-SSES groups.

Grupo	n	Pre (median [Q1–Q3])	Post (median [Q1–Q3])	Δ (Post − Pre) (median [Q1–Q3])	p (Wilcoxon)
TAPS	6	4 [3 - 8]	3 [2 -8]	-2 [-2-0]	0.4962
t-SSES	5	5 [4 - 10]	4 [2 - 4]	-1 [-2 - (-1)]	0.0975

Comparison of the changes in the partial MDS-UPDRS Part III scores (Δ = post − pre) between the TAPS and t-SSES groups using the Mann–Whitney U test also showed no statistically significant difference (W = 13.5, p = 0.8521), as shown in Table 4.

Table 4. Between-group comparison of changes (Δ = post − pre) in partial MDS-UPDRS Part III scores using the Mann–Whitney U test.

Variable	TAPS (n = 6)	t-SSES (n = 5)	p (Mann-Whitney)
Δ MDS-UPDRS Part III Partial Score	-2 [-2-0]	-1 [-2 - (-1)]	0.8521 (W =13.5)

Discussion

Patient recruitment for PD, as well as study adherence, stimulus tolerance, and safety, were satisfactory. Blinding of

the evaluating physician was also satisfactory; however, only one evaluator participated in this pilot study. To minimize bias due to individual subjectivity, future studies will include three evaluators who are blinded to the MDS-UPDRS Part III scoring.

The results of the MDS-UPDRS Part III scores showed a descriptive reduction in post-stimulation scores after 40 minutes for both forms of stimulation, which is consistent with previous larger-scale studies [30, 32]; however, due to the small sample size, the statistical tests—the Wilcoxon signed-rank test and the Mann-Whitney test— did not indicate significant differences either between the pre- and post-stimulation periods or among the experimental groups, respectively.

Another limitation of this pilot study was the baseline imbalance in disease severity between the experimental groups. The t-SSES group presented higher and more variable pre-stimulation MDS-UPDRS Part III scores than the TAPS group, consistent with a greater degree of motor impairment. In such a small sample, this baseline heterogeneity may have disproportionately influenced the between-group comparisons and reduced the ability to detect differences attributable to the stimulation protocols.

The ON medication state of patients was another limitation of the study; anti-Parkinson's medication could have attenuated the severity of motor symptoms, so any small existing difference may have been imperceptible to the evaluator.

There was no sham (placebo) stimulation condition. Without a sham stimulation condition, the descriptive improvements observed in the present study cannot be attributed specifically to the stimulation, as placebo effects, spontaneous variability, or evaluator-related factors cannot be excluded. Future controlled studies should include a sham condition.

The groups differ not only in waveform—the average intensity also differs (1.5×ST / ~ 2.7 mA in TAPS vs. ~1.3×ST / ~4.3 mA in t-SSES). These values were determined based on each patient's tolerance of the sensations caused by the stimulation. However, this should be considered when interpreting future comparisons. Another issue that may be raised concerns the frequency-selective recruitment of fibers, since higher intensities recruit a greater variety of fiber types [24].

Conclusion

A preliminary, blinded, video-based assessment of the motor tasks in Part III of the MDS-UPDRS revealed descriptive improvements in the scores for those tasks in some participants following peripheral electrical stimulation. However, no statistically significant differences were observed, and the findings should be interpreted with caution given the exploratory nature of this pilot study.

The study nevertheless demonstrated the feasibility of implementing the proposed stimulation protocol together with blinded video-based motor assessment in individuals with Parkinson's disease. Patient recruitment, protocol adherence, stimulation tolerability, and evaluator blinding were successfully achieved, supporting the practicality of this methodological approach.

Because of the small sample size and the absence of a sham control, no conclusions regarding the clinical efficacy of either stimulation protocol can be drawn. These preliminary findings support the conduct of adequately powered, randomized controlled studies to determine the therapeutic effects of non-invasive peripheral electrical stimulation in Parkinson's disease.

Consequently, the authors conducted a double-blind crossover trial in which participants, evaluators, and data analysts were fully blinded. To this end, 50 patients diagnosed with PD will be enrolled in the study. Subjects will be randomly assigned to receive either the t-SSES intervention or a placebo, with the intervention and placebo administered on different days. The sample size of N = 50 was determined using a Monte Carlo simulation of the paired Wilcoxon test, assuming a significance level of 0.05 and a power of >80%. Two movement disorder specialists will independently evaluate patient performance using Part III of MDS-UPDRS before and after the intervention. Inter-rater reliability will be assessed using the Intraclass Correlation Coefficient (ICC) and the Weighted Kappa Coefficient. Furthermore, to provide an objective measure of tremor, accelerometry signals will be recorded alongside the clinical evaluations.

Acknowledgment

The authors express their gratitude to CAPES (PROEX), FINEP (grant 01.24.0122.00), CNPq, and FAPERJ (grants E-26/204.393/2024 and SEI-260003/020273/2025) for their financial support. Additionally, the authors would like to express their gratitude to SisAssistiva-MCTI for its collaboration.

References

[1] A. Lee and R. Gilbert, "Epidemiology of Parkinson disease," Neurol. Clin., vol. 34, no. 4, pp. 955–965, 2016, doi: 10.1016/j.ncl.2016.06.012.

[2] M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression, and mortality," Neurology, vol. 17, no. 5, pp. 427–442, 1967, doi: 10.1212/wnl.17.5.427.

[3] D. J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for Parkinson disease," Arch. Neurol., vol. 56, no. 1, pp. 33–39, 1999, doi: 10.1001/archneur.56.1.33.

[4] I. Litvan et al., "SIC Task Force appraisal of clinical diagnostic criteria for parkinsonian disorders," Mov. Disord., vol. 18, no. 5, pp. 467–486, 2003, doi: 10.1002/mds.10459.

[5] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," Lancet, vol. 397, no. 10291, pp. 2284–2303, 2021, doi: 10.1016/S0140-6736(21)00218-X.

[6] R. Cacabelos, "Parkinson's disease: From pathogenesis to pharmacogenomics," Int. J. Mol. Sci., vol. 18, no. 3, p. 551, 2017, doi: 10.3390/ijms18030551.

[7] V. Voon, J. Saint-Cyr, A. M. Lozano, E. Moro, Y. Y. Poon, and A. E. Lang, "Psychiatric symptoms in patients with Parkinson disease presenting for deep brain stimulation surgery," J. Neurosurg., vol. 103, no. 2, pp. 246–251, 2005, doi: 10.3171/jns.2005.103.2.0246.

[8] J. Habets, M. Heijmans, M. L. Kuijf, M. Janssen, Y. Temel, and P. L. Kubben, "An update on adaptive deep brain stimulation in Parkinson's disease," Mov. Disord., vol. 33, no. 12, pp. 1834–1843, 2018, doi: 10.1002/mds.115.

[9] M. A. Basto et al., "A imagem na cirurgia da doença de Parkinson," Acta Med. Port., vol. 16, no. 3, pp. 135–140, 2003, doi: 10.20344/amp.1174.

[10] J. M. da Cunha and E. C. de Siqueira, "O papel da neurocirurgia na doença de Parkinson: revisão de literatura," Rev. Med. (São Paulo), vol. 99, no. 1, pp. 66–75, 2020, doi: 10.11606/issn.1679-9836.v99i1p66-75.

[11] P. A. Tass, "Vibrotactile coordinated reset stimulation for the treatment of Parkinson's disease," Neural Regen. Res., vol. 17, no. 7, pp. 1495–1497, 2022, doi: 10.4103/1673-5374.329001.

[12] K. Ong, F. Schmidt, K. Tosefsky, M. Faran, C. Sarica, C. R. Honey, F. Vila-Rodriguez, and S. Lang, "Non-motor effects of low-frequency deep brain stimulation of the subthalamic nucleus in Parkinson's disease: A

systematic review," Stereotact. Funct. Neurosurg., vol. 102, no. 6, pp. 401–413, 2024, doi: 10.1159/000540210.

[13] X. S. Tan, F. Pierres, A. Dallman-Porter, W. Hardie-Brown, and K. Y. Kwon, "Focused vibrotactile stimulation with cueing effect on freezing of gait in Parkinson's disease : Two case reports," J. Mov. Disord., vol. 14, no. 3, pp. 236–238, 2021, doi: 10.14802/jmd.21076.

[14] M. Johnson, "Transcutaneous electrical nerve stimulation: Mechanisms, clinical application and evidence," Rev. Pain, vol. 1, no. 1, pp. 7–11, 2007, doi: 10.1177/204946370700100103.

[15] V. E. Abraira and D. D. Ginty, "The sensory neurons of touch," Neuron, vol. 79, no. 4, pp. 618–639, 2013, doi: 10.1016/j.neuron.2013.07.051.

[16] R. C. Helmich, M. Hallett, G. Deuschl, I. Toni, and B. R. Bloem, "Cerebral causes and consequences of parkinsonian resting tremor: A tale of two circuits?," Brain, vol. 135, no. 11, pp. 3206–3226, 2012, doi: 10.1093/brain/aws023.

[17] M. F. Dirkx, H. den Ouden, E. Aarts, M. Timmer, B. R. Bloem, I. Toni, and R. C. Helmich, "The cerebral network of Parkinson's tremor: An effective connectivity fMRI study," J. Neurosci., vol. 36, no. 19, pp. 5362–5372, 2016, doi: 10.1523/JNEUROSCI.3634-15.2016.

[18] R. Pahwa et al., "An acute randomized controlled trial of noninvasive peripheral nerve stimulation in essential tremor," Neuromodulation, vol. 22, no. 5, pp. 537–545, 2019, doi: 10.1111/ner.12930.

[19] R. Hanajima et al., "Somatosensory evoked potentials (SEPs) recorded from deep brain stimulation (DBS) electrodes in the thalamus and subthalamic nucleus (STN)," Clin. Neurophysiol., vol. 115, no. 2, pp. 424–434, 2004, doi: 10.1016/j.clinph.2003.09.027.

[20] F. Klostermann, M. Wahl, J. Schomann, A. Kupsch, G. Curio, and F. Marzinzik, "Thalamo‐cortical processing of near‐threshold somatosensory stimuli in humans," Eur. J. Neurosci., vol. 30, no. 9, pp. 1815–1822, 2009, doi: 10.1111/j.1460-9568.2009.06970.x.

[21] J. Ara, S. H. Hwang, T. Song, and G. Khang, “Effects of the stimulus parameters on the tactile sensations elicited by single-channel transcutaneous electrical stimulation,” Int. J. Precis. Eng. Manuf., vol. 15, pp. 305–313, 2014, doi: 10.1007/s12541-014-0339-4.

[22] B. Geng, K. Yoshida, L. Petrini, and W. Jensen, “Evaluation of sensation evoked by electrocutaneous stimulation on forearm in nondisabled subjects.” J. Rehabil. Res. Dev., vol. 49, no. 2, pp. 297–308, 2012, doi: 10.1682/JRRD.2010.09.0187.

[23] J. P. Reilly, Applied Bioelectricity: From Electrical Stimulation to Electropathology, 1st ed. New York, NY, USA: Springer Science+Business Media, 1998.

[24] L. P. Volpi, "Efeitos dos estímulos elétricos senoidais na resposta induzida cerebral: estudo da seletividade das fibras sensitivas," M.S. thesis, Programa de Engenharia Biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2017.

[25] D. K. Freeman, D. K. Eddington, J. F. Rizzo III, and S. I. Fried, "Selective activation of neuronal targets with sinusoidal electric stimulation," J. Neurophysiol., vol. 104, no. 5, pp. 2778–2791, 2010, doi: 10.1152/jn.00551.2010.

[26] J. J. Katims, D. M. Long, and L. K. Ng, "Transcutaneous nerve stimulation: Frequency and waveform specificity in humans," Appl. Neurophysiol., vol. 49, no. 1-2, pp. 86–91, 1986, doi: 10.1159/000100133.

[27] H. R. Martins, R. Zanetti, C. C. D. Santos, G. M. Manzano, and C. J. Tierra-Criollo, "Current perception threshold and reaction time in the assessment of sensory peripheral nerve fibers through sinusoidal electrical stimulation at different frequencies," Rev. Bras. Eng. Biomed., vol. 29, no. 3, pp. 278–285, 2013.

[28] B. J. de Souza, "Avaliação da percepção por corrente senoidal: uma nova perspectiva para neuropatias," M.S. thesis, Programa de Engenharia Biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2016.

[29] C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," Mov. Disord., vol. 23, no. 15, pp. 2129–2170, 2008, doi: 10.1002/mds.22340.

[30] S. Brillman et al., "Non-invasive transcutaneous afferent patterned stimulation therapy offers action tremor relief in Parkinson’s disease," Tremor Other Hyperkinet. Mov., vol. 13, no. 1, p. 25, 2023, doi: 10.5334/tohm.762.

[31] D. A. Molina-Vidal, P. Cevallos Larrea, L. Guambana Calle, D. Liquori, and C. J. Tierra-Criollo, “Design and evaluation of a prototype of dual channel electrical stimulator: Application in the assessment of current perception thresholds to multiple stimuli,” in Smart Technologies, Systems and Applications: SmartTech-IC 2022, ser. Communications in Computer and Information Science, vol. 1705. Cham: Springer, 2023, pp. 342–355.

[32] L. Meng, M. Jin, X. Zhu, and D. Ming, "Peripheral electrical stimulation for parkinsonian tremor: a systematic review," Frontiers in Aging Neuroscience, vol. 14, Art. no. 795454, 2022, doi: 10.3389/fnagi.2022.795454.

Octalysis Framework Applied to the Gamification of an FES-Assisted System in Virtual Reality for the Rehabilitation of People with Spinal Cord Injury

1st Breno Linhares de Sousa
Graduate Program in Biomedical Engineering (PPGEB)
University of Brasilia (UnB)
Brasília-DF, Brazil
0009-0003-2711-7903

2nd Caio Ferreira Duarte
Faculty of Science and Technology in Engineering (FCTE)
University of Brasilia (UnB)
Brasília-DF, Brazil
0009-0009-0053-4049

3rd Guillermo Alvarez Bestard
Faculty of Science and Technology in Engineering (FCTE)
University of Brasilia (UnB)
Brasília-DF, Brazil
0000-0001-6659-441X

4th Jorge Cormane
Graduate Program in Biomedical Engineering (PPGEB)
University of Brasília (UnB)
Brasilia-DF, Brazil
0000-0001-5263-0985

5th Claudia Ochoa-Diaz
Faculty of Science and Technology in Engineering (FCTE)
University of Brasilia (UnB)
Brasília-DF, Brazil
0000-0002-4071-6244

***Abstract*—This study aims to define and implement a gamification proposal for a virtual reality FES-assisted cycling system (*v-FES-Cycling*) using the Octalysis framework and an expert judge evaluation methodology, conceptually structuring an exergame for a preliminary rehabilitation platform for people with spinal cord injury (SCI). The evaluation methodology of the expert judge was used to select techniques based on the eight motivational drivers of the Octalysis framework. The mechanics were implemented in the Unreal Engine 5 (UE) game engine, integrating FES hardware, sensors, and a Virtual Reality (VR) environment. The judges prioritized the drives Unpredictability and Curiosity (4.21), Development and Accomplishment (3.92), Empowerment, Creativity, and Feedback (3.54) and Scarcity and Impatience (3.38). Techniques such as coin collection, scoring, obstacle avoidance, lap counting, visual and audio feedback, and the selection of avatars and scenarios were implemented, resulting in a functional serious game (exergame). The study demonstrated the preliminary technical feasibility of the exergame's technological development. Controlled clinical trials to evaluate usability and efficacy with real patients with SCI will be conducted in future phases.**

***Index Terms*—Rehabilitation, FES-Cycling, Octalysis framework, Judge-based assessment methodology.**

I. Introduction

This study is part of an ongoing research project that integrates Virtual Reality (VR), Functional Electrical Stimulation (FES), and an instrumented tricycle for the rehabilitation of individuals with Spinal Cord Injury (SCI). The study is based on the *v-FES-Cycling* system, originally proposed in [1], [2] e [3], which developed the data communication interface between the stimulation hardware and the immersive environment of a VR game. This phase of the research aims to add structured layers of gamification to the previously developed exergame. Thus, the focus of this article is the definition of the motivational design and the preliminary implementation of software in the Unreal Engine (UE) game engine for gamification strategies intended for FES-assisted cycling (FES-Cycling), preliminarily assessing the initial ludic relevance with experts.

SCI is a debilitating neurological condition that results in partial or total disruption of motor and somatosensory functions below the level of the lesion [4], [5], directly impacting the mobility, functional independence, and quality of life of affected individuals [5], [6]. In this context, motor rehabilitation, when performed continuously and intensively, plays a fundamental role in promoting neuroplasticity, restoring motor functions, and preventing secondary complications [7], [8].

Among the available therapeutic approaches, FES stands out as an effective strategy, based on the application of electrical pulses to induce contractions in paralyzed muscles. This technique helps prevent muscle atrophy, maintain range of motion, and enable the performance of functional movements, even in the absence of voluntary control [7], [9]. One of the most well-established applications of FES is FES-Cycling, which involves the use of a cycle ergometer or tricycle combined

Financial support from the Distrito Federal Research Support Foundation (FAPDF), through Public Call No. 10/2023 - FAPDF Learning Program (project No. 00193-00002152/2023-95).

with electrical stimulation to promote cyclic movement of the lower limbs, thereby facilitating the coordinated activation of muscle groups during pedaling cycles [10].

Given this scenario, emerging technologies such as VR and serious games have been explored as promising strategies to increase patient motivation and engagement during rehabilitation [11], [12], [13]. The use of immersive virtual environments allows for the creation of dynamic, multisensory experiences, providing real-time feedback, enhancing the sense of presence, and transforming therapeutic exercises into more engaging and playful activities. In this context, *exergames*, which combine physical exercise with game elements, have shown potential for increasing intrinsic motivation and encouraging the continued practice of therapeutic activities [14], [15]. Based on this aforementioned approach, in [1] the *v-FES-Cycling* system was proposed, which integrates an FES-powered tricycle into a VR environment [1], [3], [10]. Despite its established clinical benefits, FES-Cycling still faces challenges related to low user engagement, often associated with the repetitive, monotonous, and physically demanding nature of the exercises, which can compromise motivation and long-term treatment adherence [11], [12].

At the same time, gamification has become established as an effective approach to promoting engagement in non-game contexts, including healthcare field [16], [17]. This approach consists of applying game elements and mechanics, such as scoring, rewards, challenges, and progression, with the aim of motivating and influencing behavior and increasing user adherence [18]. Evidence indicates that gamified interventions can improve self-efficacy and reduce negative emotions associated with prolonged treatments [18], [19]. However, the effectiveness of these strategies depends directly on a structured design capable of considering the psychological and motivational needs of the target audience [18], [20]. In this context, the Octalysis framework, developed by Yu-kai Chou, stands out for offering a systematic approach based on eight core motivational drives, known as core drives (CD) [16], [21]. Unlike simplified approaches centered only on points and rewards, the Octalysis framework allows for the analysis and development of gamified experiences in a more systematic way, considering different aspects of human motivation [21], [22].

Other studies have involved initiatives such as the *iCycle* platform, which integrates FES-Cycling with commercial games, but without detailing a formal gamification framework [8]. *PedaleoVR*, on the other hand, uses *exergames* in cycling, but without integration with FES [14]. Despite studies using the Octalysis framework in different contexts, the combination of FES, VR, and structured gamification remains underexplored in the rehabilitation literature [18]. There are also methodological gaps in the preliminary evaluation of these systems that integrate games by expert reviewers prior to conducting clinical trials [23].

Given this context, this article aims to define gamification techniques and present the initial results of implementing these techniques in Unreal Engine software for the *v-FES-Cycling* system, using the Octalysis framework to propose preliminary game mechanics aligned with the target audience's perceived motivational profile as evaluated by expert judges, as well as evaluation by expert judges as a method for defining these gamification techniques. It is thus expected to contribute to the development of a more adaptable and functional preliminary solution in the context of rehabilitation for people with SCI, serving as a basis for future clinical trials and usability evaluations with patients.

II. Methods

The methodology adopted for this study follows a structured process for applying the expert judge evaluation methodology and the Octalysis gamification framework to the *v-FES-Cycling* rehabilitation system, aiming to overcome the repetitive and demotivating nature of conventional rehabilitation for people with SCI. It is important to note that this study focuses on defining gamification techniques and developing these game mechanics within the UE software for the *v-FES-Cycling* system. Thus, the phases described here did not involve the direct participation of patients, since this initial stage was intended for the development and technical assessment of system feasibility, as well as for the preliminary definition of gamification techniques using the expert judge evaluation methodology, as described in [16].

The process is illustrated in Fig. 1, which presents the stages from the characterization of the current system to the development and analysis of gamification results using the Octalysis framework.

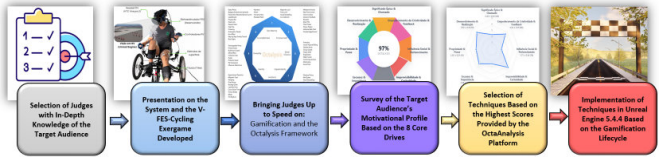


Fig. 1. Methodological flowchart of the steps carried out in this study.

The methodology is based on expert evaluation, leveraging the expertise of professionals specializing in health, rehabilitation, and the developed *v-FES-Cycling* system to assess and define the gamification strategies best suited to the perceived motor and psychological profile of the target audience [16].This approach allows the system's design to be preliminarily evaluated and technically verified prior to clinical trials and ensures that the game is grounded in clinical and pedagogical knowledge, so that these initial definitions make the proposed mechanics safer and more effective before future clinical trials are conducted.

A. Current v-FES-Cycling System

The *v-FES-Cycling* system is a rehabilitation platform that integrates an adapted, FES-assisted tricycle with an immersive VR environment. The hardware consists of an 8-channel functional electrical stimulator, based on the STM32 microcontroller, which activates the patient's lower limbs in a synchronized manner [7]. The instrumentation includes an

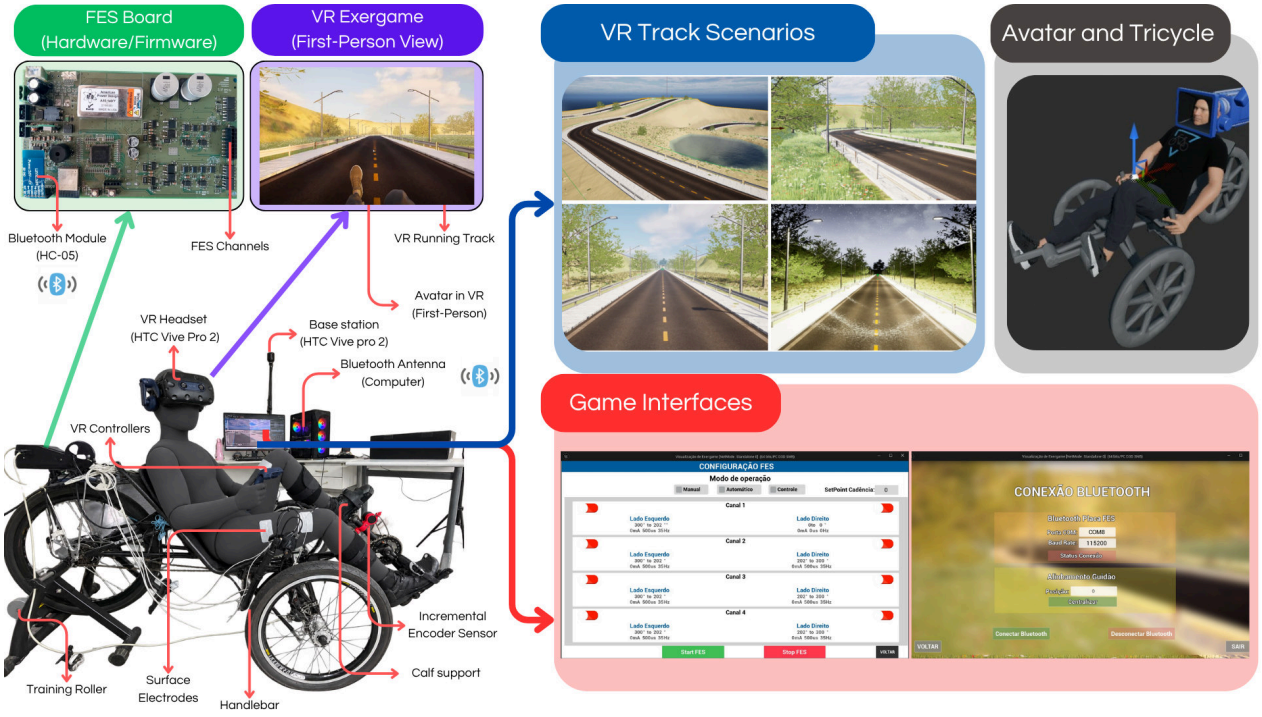


Fig. 2. Diagram highlighting the overall architecture of the *v-FES-Cycling* system focusing on the developed game, the *HTC Vive Pro 2*, the connection of the surface electrodes to the lower limbs, the avatar on the tricycle, and the visual interface rendered in UE, showing the communication mesh between the hardware and the game's VR environment.

incremental encoder sensor that monitors the pedaling angle and cadence, sending the data via Bluetooth (HC-05 module) to the software developed in Unreal Engine 5.4.3 (UE) [1]. The overall architecture, shown in Fig. 2, highlights the focus on the game experience, in which the user with SCI wears an HTC Vive Pro 2 headset while pedaling on the adapted FES-driven tricycle, allowing the patient to view an avatar that replicates their actual movements in a virtual scenario.

The VR environment developed in [1] was implemented with the game category known as *exergame* (a game category aimed at encouraging physical exercise) in mind, as shown in Fig. 2, and corresponds to the initial version of the system, still without the implementation and application of the gamification techniques proposed based on the Octalysis framework.

These environments were built using UE's native terrain-modeling tools, in addition to 3D modeling software such as Blender, used to model the tricycle. Assets available on the UE Marketplace's FAB platform were also used. Thus, this initial version is characterized as a VR racetrack simulator, in which five distinct racing scenarios were developed in UE, each with different environmental designs and routes to be traveled.

The system also features a virtual tricycle linked to a VR avatar, capable of simulating the user's displacement and pedaling motion along the routes proposed in each scenario. The tricycle's physics and movement dynamics exhibit realistic behavior, similar to that of a real vehicle, thanks to the physics simulation resources provided by UE and the use of the Chaos Vehicle plugin. In addition, the game provides interfaces for configuring the electrical stimulation parameters of each FES channel, as well as functionalities for configuring and pairing the Bluetooth device, both essential for starting the game.

B. Gamification through the Octalysis Framework

The strategy for aligning the game mechanics with the motivational profile of the target audience uses the Octalysis framework, which analyzes human motivation through eight central drives called Core Drives (CD). CD1 - Epic Meaning and Calling: Frames the exercise as part of a higher narrative mission. CD2 - Development and Accomplishment: Focuses on measurable progress, such as visualizing progress through status bars and achieving therapeutic milestones on the racetrack. CD3 - Empowerment of Creativity and Feedback: Stimulates the user's sense of control through immediate responses from the virtual environment to pedaling cadence and freedom of visual exploration. CD4 - Ownership and Possession: Creates attachment through the collection of virtual items such as coins, stars, and rays, which reinforce the user's connection to their progress, as well as through the ability to choose and customize the avatar, for example. CD5 - Social Influence and Relatedness: Compares metrics and fosters a sense of belonging to the rehabilitation group. CD6 - Scarcity and Impatience: Uses the temporary exclusivity of content to maintain interest, such as the progressive unlocking of new track stages after meeting rehabilitation-related goals in the game. CD7 - Unpredictability and Curiosity: Introduces surprise elements to encourage continued practice, such as the random appearance of collectible items and variations in

the track scenario. CD8 - Loss and Avoidance: Encourages regularity to avoid stagnating in the ranking or losing accumulated bonuses. It must be applied moderately, aiming to avoid setbacks, since, in the context of SCI rehabilitation, the loss of content is avoided so as not to generate excessive frustration, focusing instead on preserving achievements [21].

In this work, each selected gamification technique aims to stimulate one or more of these drives to increase adherence to repetitive treatment, turning a repetitive pedaling exercise into a mission with purpose. To analyze and visualize the motivational profile obtained through the expert-rating methodology, the *OctaAnalysis* tool (https://octaanalysis.cedis.tec.br/) is used, which generates a radar chart, allowing for a balance between "White Hat" (positive motivation) and "Black Hat" (motivation based on urgency/loss) motivations. In the context of rehabilitation for people with SCI, the use of mechanics that involve loss of content or progress is avoided in order to minimize feelings of frustration by focusing on preserving the achievements made by the patient. In addition, the framework defines the gamification lifecycle, consisting of the Discovery, Onboarding, Day-to-Day, and Endgame phases [16].

C. Judges' Evaluation Methodology

The specific techniques for implementing the game mechanics were obtained through the judge-based elicitation strategy proposed by [16]. For this stage, an interdisciplinary group of six specialists was formed, with knowledge and experience in the target audience, FES-Cycling rehabilitation, and the development of the *v-FES-Cycling* system. The committee consisted of a physical educator, a physical therapist, three electrical engineers, and a software engineer. The committee's role was to define requirements and perform technical evaluation and design verification of the implemented game mechanics based on the gamification techniques defined by the Octalysis framework, with the aim of supporting the initial development of the *v-FES-Cycling* system's *exergame*. The workflow included the following stages:

- **Leveling:** Theoretical training for judges on the Octalysis framework and how the *v-FES-Cycling* system works.
- **Discussion:** Periodic meetings to analyze the constraints and suitability of the techniques in the framework's catalog, identifying those that best adapt to the motor limitations and profile of patients with SCI.
- **Independent Evaluation for Scoring:** Assigning scores on a Likert scale (1-Very Poor, 2-Poor, 3-Fair, 4-Good, and 5-Excellent) to rate the potential of gamification techniques in relation to the perceived effectiveness of their application in the context of rehabilitation. This methodology is particularly important when direct contact with the target audience is limited in the early stages of the design process, as it draws on the expertise of specialist judges to avoid implementing techniques that could cause frustration or *cybersickness* (a sensation of nausea caused by VR use).
- **Consensus Analysis:** Next, the means and standard deviations of the suitability scores assigned to the techniques are calculated, followed by a meeting for the qualitative consolidation of the results, taking into account the gamification life cycle. The scoring, as well as the analysis and visualization of the data, is performed using the *OctaAnalysis* tool, which is used to evaluate and support the selection of the highest-ranked gamification techniques according to the Octalysis framework.

D. Development of the Selected Techniques

The definition of gamification techniques for the *v-FES-Cycling* system was based on prioritizing the motivational drives considered most relevant to the target audience. Techniques associated with the highest-scoring CDs were selected, according to the analysis performed by the *OctaAnalysis* tool. The selection considered criteria such as motivational potential and suitability to the users' profile. Based on the selected techniques, the gamification life cycle was structured, comprising the stages of discovery, onboarding, day-to-day use, and endgame [21].

Based on the definition of the stages and gamification techniques, the system's functional and non-functional requirements were defined, guiding the application's development. Assets and 3D elements from the UE Marketplace were used, including collectible coins, energy stars, arches, start and finish lines, particle effects, and widget-based interfaces. These elements were programmed according to the implemented gamification techniques and distributed along the virtual track, providing immediate performance feedback through dynamic updates in the graphical interface. *Blueprints* and Widgets programming systems were used to implement scoring events, along with the *Niagara* system for creating visual effects associated with item collection. The *Control Rig* system was used to synchronize the avatar's leg movement with the encoder data, as described in [1]. Spatial sounds were integrated to signal item collection and the completion of objectives, which helps reinforce the sense of accomplishment and stimulate the user's curiosity.

Safety measures aimed at mitigating risks through the system's hardware, firmware, and software components were also considered during the development of the *v-FES-Cycling* system, as described below:

1) **Electrical Stimulation (FES):** The electrical stimulation parameters are limited by the hardware and firmware of the FES board, as well as by the configuration interface implemented in the *exergame*. The system's maximum operating ranges are restricted to a current amplitude of up to 100 mA, a pulse width of up to 1000 μs, and a frequency of up to 100 Hz; these can be configured to lower values according to each patient's clinical needs.
2) **Emergency button:** The adapted tricycle features an emergency button accessible to both the patient and the healthcare professional, connected directly to the FES board's power line. This mechanism allows the system's operation to be immediately interrupted whenever necessary.

3) **Effects of VR-induced *cybersickness*:** To facilitate the user's gradual adaptation to the headset and the immersive environment and to reduce the visual discomfort associated with VR use, the *exergame* was designed to maintain a stable refresh rate of 90 frames per second (FPS) on the HTC Vive Pro 2 headset, minimizing delays in the rendering of the virtual environment. In addition, the initial race track scenarios were designed with predominantly straight paths and few game mechanics interactions, avoiding sudden changes in the virtual camera's position and excessive visual stimuli.
4) **Gamification Mechanics:** In the gamification design based on the Octalysis framework, the dynamics associated with the Core Drives of Scarcity (CD6) and Loss and Avoidance (CD8) were calibrated for passive engagement in this preliminary implementation. Thus, no punitive penalties are applied, nor is there any loss of progress accumulated during rehabilitation if the user fails to complete the proposed challenges. This strategy aims to eliminate feelings of coercion or excessive frustration, prioritizing positive motivational reinforcement (White Hat).

III. Results and Discussion

This section presents and discusses the results obtained in the process of selecting and implementing the gamification techniques applied to the *v-FES-Cycling* system. First, the scores assigned to the eight CDs of the Octalysis framework through the expert judges' evaluation process are analyzed. Next, the selected gamified techniques are discussed, along with their organization throughout the gamification life cycle and their practical implementation in the virtual environment developed in UE.

Fig. 3 shows the radar chart depicting the average scores assigned to the eight Octalysis framework CDs by the six expert judges in the fields of health, engineering, and neurorehabilitation using FES-based technologies. This type of graphical representation allows for a comparative visualization of which motivational dimensions were considered most relevant to the target audience, thereby aiding in this preliminary stage of *exergame* development.

According to the results, the highest values were observed for the CDs Unpredictability & Curiosity, Development & Accomplishment, Empowerment of Creativity & Feedback, and Scarcity & Impatience, showing that factors related to curiosity, exploration, progression, autonomy, and gradual rewards tend to exert greater motivational influence in the context of gamified rehabilitation. In contrast, dimensions such as Social Influence & Relatedness, Loss & Avoidance, and Epic Meaning & Calling showed lower averages, indicating lower perceived relevance to the profile of the game's target audience by the expert judges. In particular, the lowest score (1.77 ± 0.16) in Social Influence & Relatedness suggests that, in the experts' view, the initial focus of rehabilitation should be individual and centered on personal progress and the discovery

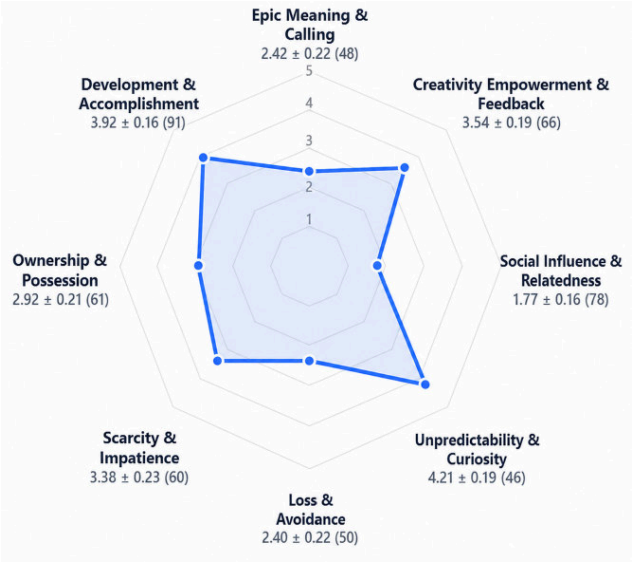


Fig. 3. Radar chart corresponding to the scoring results of the eight Core Drives of the Octalysis Framework after the judges' evaluation process.

of new stimuli, rather than on competitive or social dynamics, which are more likely to generate comparison and frustration.

As shown in Table I, the four drives with the highest scores were Unpredictability & Curiosity (4.21), Development & Accomplishment (3.92), Empowerment of Creativity & Feedback (3.54), and Scarcity & Impatience (3.38). Thus, the combination of judges' evaluation and the application of the Octalysis framework enabled a selection of motivational strategies based on expert perception of the needs of the target audience.

TABLE I
Means and Standard Deviations of the Four Octalysis Core Drives Selected for Having the Highest Scores

Core Drive	Mean and SD
Unpredictability & Curiosity	4.21 ± 0.19
Development & Accomplishment	3.92 ± 0.16
Empowerment of Creativity & Feedback	3.54 ± 0.19
Scarcity & Impatience	3.38 ± 0.23

Based on these results, specific gamification techniques linked to each CD were selected, following the criterion of prioritizing those that received the highest scores from the judges for each dominant CD, in line with the guideline of using professional expertise to identify the most effective basic motivations. In the context of Unpredictability & Curiosity, the techniques of mini-missions, progressive rewards, and visual narrative stood out, used to arouse continuous interest and anticipation in the user. For Development & Accomplishment, progress bars, mission lists, badges, and points were prioritized, promoting a sense of advancement and achievement. In the Empowerment of Creativity & Feedback CD, techniques such as immediate feedback and milestone unlocking were selected, encouraging continuous interaction and recognition of actions taken. In Scarcity & Impatience, the reward-pacing

technique stood out, applied to distribute rewards over time and encourage continued use of the system.

Based on the prioritized techniques, the gamification life cycle was structured, composed of the discovery, onboarding, day-to-day, and endgame stages [21]. The techniques were distributed to allow for a progressive and adaptable experience, essential for maintaining long-term engagement.

In the discovery phase, the techniques of mini-missions, visual narrative, progress bar, and points were defined, aiming to introduce the user to the system in an intuitive and attractive way. In the onboarding phase, mission lists, badges, immediate feedback, and milestone unlocking were used, favoring initial adaptation to the game mechanics. In the continuous-use (day-to-day) stage, the techniques of progressive rewards, milestone unlocking, and reward pacing were defined, seeking to maintain engagement over time. Finally, in the endgame phase, progressive rewards, badges, points, and milestone unlocking stood out, reinforcing the sense of completion and final reward.

Fig. 4 shows a *gameplay* scene in which the user rides the tricycle toward coins distributed throughout the virtual environment. When collected, the coins are converted into particle visual effects accompanied by sound feedback, providing immediate feedback to the patient and positively reinforcing their interaction with the game. In addition, a start and finish line was used to mark the beginning and end of the route, as well as obstacles such as barriers and cones, used to vary route difficulty and introduce mini-missions related to rehabilitation exercises. Heads-Up Display (HUD) interface elements are also shown, presenting real-time information such as the number of laps completed and the number of coins collected.

Fig. 5 shows some of the system's configuration interfaces, allowing the healthcare professional to customize the experience according to the patient's needs. Among the available functionalities are the configuration of Bluetooth communication, adjustment of the FES electrical stimulation parameters, setting the player's name, and selecting among three avatars with distinct designs, positioned on the tricycle in VR.

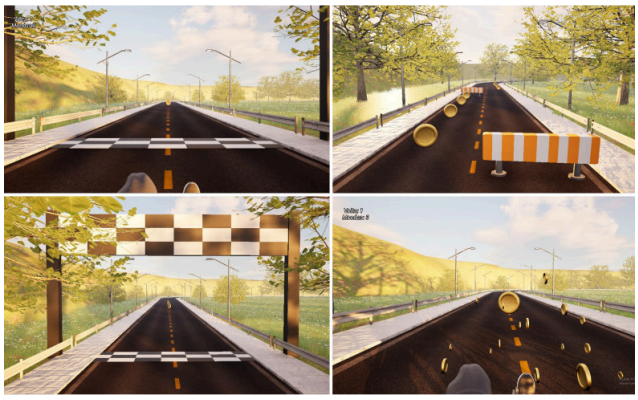


Fig. 4. Gameplay scene with coin collection, obstacles, and immediate feedback to the user.

In addition, the system offers the option to choose among five scenarios with distinct visual characteristics and routes. Scenario 1 consists of a straight track surrounded by trees, set during the day. Scenario 2 features a straight track through a forest at night. Scenario 3 is made up of a closed-circuit track with curves, surrounded by mountains, trees, and a flowering field. Scenario 4 is a track with varied curves, including both closed and open sections. Finally, Scenario 5 consists of a track with curves and varying uphill and downhill sections, set in a coastal landscape with an ocean view.

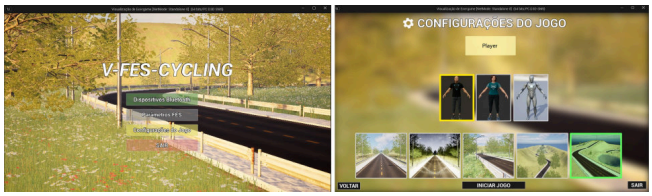


Fig. 5. Game configuration interfaces and scenario selection.

Finally, Table II presents, in greater detail, the relationship between the prioritized Core Drives and the innovative gamification mechanics in this initial version of the *exergame*.

TABLE II
RELATIONSHIP BETWEEN CORE DRIVES AND IMPLEMENTED MECHANICS.

Core Drive (Octalysis)	**Gamification Mechanics (UE5)**
Unpredictability & Curiosity (CD7)	Coins, obstacles, and varying scenarios appear in various positions scenarios.
Development & Execution (CD2)	Points earned by collecting coins and the count of laps completed.
Creativity Empowerment & Feedback (CD3)	Immediate visual/auditory feedback on pedaling and the ability to choose on avatar.
Scarcity & Impatience (CD6)	Temporary blocking of the most complex routes based on the healthcare profissional's assessment of the patient's performance by a healthcare profissional.

Overall, the results demonstrate that the development of gamification techniques, guided by expert evaluation and structured based on the Octalysis framework, made it possible to build a preliminary solution consistent with the perceived needs and limitations of the target audience, as evaluated by the experts. The application of the judges' methodology proved particularly relevant in this process, as it allowed the selection of game mechanics compatible with the therapeutic context without prematurely exposing users to risks during the early stages of development. As noted by [16], resources such as immediate feedback can act as positive biofeedback, contributing to motor reinforcement and processes associated with neuroplasticity. Furthermore, the distribution of techniques throughout the gamification lifecycle has facilitated the development of an *exergame* designed to provide a progressive and motivating experience during this initial implementation phase, with the potential to increase user engagement and encourage continued use of the system in future versions.The following link presents a complementary video with further details on the results of the gamification techniques implemented in the *v-FES-Cycling* system: https://drive.google.com/drive/folders/1Ks1rxSy-oby1Bs3D20dnj4vIJ2CmcOG3?usp=sharing.

IV. Conclusions

This study achieved its overall objective by defining and initiating the development of gamification mechanics in Unreal Engine software for the *v-FES-Cycling* system. The application of the Octalysis framework, combined with an internal evaluation methodology conducted by expert judges, made it possible to propose and prioritize initial gamification mechanics for the FES-assisted cycling rehabilitation *exergame* that are better aligned with the target audience's perceived motivational profile as judged by the experts. This approach contributed to the development of an initial version of the system, with the potential to transform repetitive exercises into a more adaptive, engaging, and playful rehabilitation platform. Furthermore, its practical implementation in Unreal Engine demonstrated that theoretical concepts of gamification can be consistently translated into interactive features designed for motor rehabilitation. As limitations, it should be noted that direct contact with the target audience, composed of patients with SCI, was still limited during the initial phases of system design and implementation; therefore, no clinical trials were conducted on patients at this early stage. Consequently, the participation of expert judges became necessary; although this represents a limitation of this initial stage and is subject to the subjectivity of the evaluations conducted, it contributed to the preliminary development of the *exergame* by guiding the selection of the gamification techniques employed, minimizing risks related to *cybersickness*, frustration, and the execution of potentially inappropriate or unsafe movements during the proposed activities, as well as enabling the preliminary technical verification of the integration between the *exergame* and the hardware, firmware, and software components of the *v-FES-Cycling* system. Looking ahead, it is essential to validate the system and the implemented gamification techniques through clinical trials with patients with SCI. These future studies will allow for the objective measurement of the solution's impact on motivation, treatment adherence, and long-term motor recovery. Nevertheless, the results obtained so far are promising and indicate that the *v-FES-Cycling* system has the potential to establish itself as a tool to support rehabilitation through FES-Cycling.

Acknowledgments

The authors thank the participants who collaborated in the judges' evaluation of this study, the BEM-TE-VI platform team, and Prof. Sérgio Freitas and CEDIS for providing the *OctaAnalysis* tool.

References

[1] B. L. Sousa, V. S. Levy, N. N. V. Ramos, G. A. Bestard, J. Cormane, and C. Ochoa-Diaz, "Desenvolvimento de um exergame de FES-Cycling em realidade virtual para o auxílio na reabilitação de pessoas com lesão medular," in *XVI Simpósio de Engenharia Biomédica (SEB 2025)*, Uberlândia, MG, Brasil, Oct. 2025.

[2] B. L. Sousa, V. S. Levy, G. A. Bestard, J. Cormane, and C. Ochoa-Diaz, "Integration of a virtual reality system with a FES-assisted trike for rehabilitation of individuals with spinal cord injury," in *47th Annual Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Copenhagen, Denmark, 2025, pp. 1–6.

[3] B. L. Sousa, V. S. Levy, G. A. Bestard, J. Cormane, and C. Ochoa-Diaz, "V-FES-Cycling: desenvolvimento de uma interface de comunicação entre um sistema FES-Cycling e um ambiente de realidade virtual para reabilitação de pacientes com lesão medular," *Rev. Interdiscip. Pesq. Eng.*, vol. 10, pp. 124–133, 2025.

[4] L. B. Mongeot, E. Galofaro, F. Ramadan, E. D'Antonio, F. Missiroli, N. Lotti, et al., "Combining FES and exoskeletons in a hybrid haptic system for enhancing VR experience," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 4812–4820, 2023.

[5] L. Santana, E. Fachin-Martins, D. L. Borges, J. G. T. Cavalcante, N. Babault, F. R. Neto, et al., "Neuromuscular disorders in individuals with spinal cord injury," *J. Spinal Cord Med.*, vol. 46, no. 5, pp. 742–752, 2023.

[6] K. Nas, L. Yazmalar, V. Şah, A. Aydın, and K. Öneş, "Rehabilitation of spinal cord injuries," *World J. Orthop.*, vol. 6, no. 1, pp. 8–16, 2015.

[7] T. Coelho-Magalhães, E. Fachin-Martins, A. Silva, C. A. Coste, and H. Resende-Martins, "Development of a high-power capacity open source electrical stimulation system to enhance research into FES-assisted devices: validation of FES cycling," *Sensors (Basel)*, vol. 22, no. 2, p. 531, 2022.

[8] L. D. Duffell, S. Paddison, A. F. Alahmary, N. Donaldson, and J. Burridge, "The effects of FES cycling combined with virtual reality racing biofeedback on voluntary function after incomplete SCI: a pilot study," *J. Neuroeng. Rehabil.*, vol. 16, pp. 1–15, 2019.

[9] P. H. Peckham and J. S. Knutson, "Functional electrical stimulation for neuromuscular applications," *Annu. Rev. Biomed. Eng.*, vol. 7, pp. 327–360, 2005.

[10] C. Fattal, B. Sijobert, A. Daubigney, E. Fachin-Martins, B. Lucas, J. M. Casillas, et al., "Training with FES-assisted cycling," *J. Spinal Cord Med.*, vol. 43, no. 3, pp. 402–413, 2020.

[11] S. S. Esfahlani, et al., "ReHabgame: a non-immersive virtual reality rehabilitation system with applications in neuroscience," *Heliyon*, vol. 4, p. e01290, 2018.

[12] S. Balbino and H. Sousa, "Development of immersive virtual reality interface for lower-limb robotic rehabilitation," in *2018 Latin American Robotic Symposium, Brazilian Symposium on Robotics and Workshop on Robotics in Education*, 2018, pp. 124–128.

[13] G. F. Cyrino, "Harpygame: um jogo sério customizável com interface multimodal para reabilitação de indivíduos pós-AVE," Dissertação, Universidade Federal de Uberlândia, Uberlândia, 2019.

[14] A. Rojo, A. Castrillo, C. López, L. Perea, F. Alnajjar, J. C. Moreno, et al., "PedaleoVR: usability study of a virtual reality application for cycling exercise in patients with lower limb disorders and elderly people," *PLoS One*, vol. 18, no. 2, p. e0280743, 2023.

[15] K. I. Molina, N. A. Ricci, S. A. Moraes, and M. R. Perracini, "Virtual reality using games," *J. Neuroeng. Rehabil.*, vol. 11, p. 156, 2014.

[16] C. S. Ramos, M. A. S. Farias, S. A. A. Freitas, M. V. P. Martins, J. M. Alves, and L. C. S. Pinto, "A process to identify players' motivational profiles for designing a gamification project," in *Computational Science and Its Applications – ICCSA 2024*, Lecture Notes in Computer Science, Cham: Springer, 2024, pp. 55–68.

[17] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "Defining gamification," in *MindTrek*, 2011.

[18] N. H. H. Fadzillah, et al., "Gamification framework for rehabilitation," *J. Health Qual. Life*, 2025.

[19] C. Prahm, F. Kayali, A. Sturma, and O. Aszmann, "PlayBionic," *PM&R*, vol. 10, no. 11, pp. 1252–1260, 2018.

[20] M. R. Floryan, L. M. Ritterband, and P. I. Chow, "Gamification principles," *Transl. Behav. Med.*, vol. 9, no. 6, pp. 1131–1138, 2019.

[21] Y. K. Chou, *Actionable Gamification: Beyond Points, Badges, and Leaderboards*. Leanpub, 2016.

[22] R. Han, B. G. Lee, M. Pike, et al., "Sproutfit: an immersive virtual reality seed planting game to increase patient motivation for exercise performance through loss-aversion gamification," *Virtual Reality*, vol. 29, p. 137, 2025.

[23] R. C. G. Sant'Ana Van Erven, D. De Almeida Jubé, H. R. Santos, S. A. A. Andrade De Freitas, and E. D. Canedo, "Gamification project to receive continuous feedback in the context of the evolution of public service for lawyers," in *IEEE Frontiers in Education Conference*, 2023, pp. 1–8.

Manufacturing Workflow and Cost Analysis of an Open-Source Pediatric Wheelchair Produced by Fused Filament Fabrication

1st Cinthia H. C Nascimento Pereira
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
cinthia.carvalho@unifesp.br

2nd Paula Ide Coelho de Souza
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
paula.ide@unifesp.br

3rd Miriam Raquel Diniz Zanetti
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
miriam.zanetti@unifesp.br

4th Vagner Rogério dos Santos
Departamento de Oftalmologia e Ciências Visuais
Universidade Federal de São Paulo
São Paulo, Brasil
vagner.rogerio@unifesp.br

5th Claudia Fegadolli
Departamento de Ciências Farmacêuticas
Universidade Federal de São Paulo
Diadema, Brasil
cfegadolli@unifesp.br

6th Maria Cecilia Martinelli
Departamento de Fonoaudiologia
Universidade Federal de São Paulo
São Paulo, Brasil
cmartinelli@unifesp.br

7th Eduardo Keller Rorato
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
ek.rorato@unifesp.br

8th Thabata Alcantara F. Ganga
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
thabata.ganga@unifesp.br

9th Maria Elizete Kunkel
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
elizete.kunkel@unifesp.br

***Abstract*—Independent mobility is essential for child development. Access to appropriate pediatric wheelchairs remains limited, particularly in resource-constrained settings. This pilot study characterized the manufacturing workflow, production efficiency, and costs of an open-source pediatric wheelchair fabricated by fused filament fabrication (FFF). The wheelchair was produced from the Toddler Mobility Trainer design using PETG and TPU filaments. Printing time, material consumption, failures, reprinting events, and direct and operational costs were prospectively recorded. After assembly, the prototype underwent visual and manual inspection and preliminary unloaded verification in a controlled indoor environment. The wheelchair was completed after 27 print runs, requiring 453 h of printing and 11,031 g of filament. Failures consumed 1,543 g of filament corresponding to a material loss rate of 14%. TPU presented a higher proportional material loss than PETG (24.6% versus 11.3%). Reprints accounted for an additional 144 h, representing 31.8% of the total printing time. The direct material and hardware cost was US$ 231.97, and manufacturing failures increased this cost by approximately 17.1%. The wheelchair was successfully assembled, with satisfactory integration of printed and commercial components and a visually coherent final configuration. During preliminary unloaded verification, it could be manually propelled and maneuvered on a flat indoor surface without visible deformation or component failure. Operational costs were mainly influenced by prolonged printing time and reprinting events. These findings support workflow optimization and indicate the potential of open-source additive manufacturing for developing mobility solutions for young children. Standardized mechanical testing, durability assessment, and clinical evaluation are required before pediatric use.**

Keywords—digital manufacturing; fused filament fabrication; pediatric wheelchair; assistive technology.

I. INTRODUCTION

Independent mobility is essential for children's physical, cognitive, and social development. Appropriate pediatric wheelchairs enable participation in education, daily activities, and community life while promoting autonomy and environmental exploration. Conversely, delayed or inadequate provision of mobility devices may negatively affect child development and contribute to long-term social inequalities [1]-[2]. Despite their recognized benefits, access to wheelchairs remains limited worldwide, particularly in low- and middle-income countries. According to the World Health Organization (WHO), approximately 80 million people worldwide require a wheelchair, but only a limited proportion have access to a device that adequately meets their needs [3]. Pediatric wheelchair provision is particularly challenging because children's anthropometric characteristics, postural support requirements, and functional needs change continuously as they grow, frequently requiring device adjustment, adaptation, or replacement [4].

Digital manufacturing has emerged as a promising approach to the developing of personalized assistive technologies by integrating computer-aided design, digital fabrication, and additive manufacturing. Among the available technologies, fused filament fabrication (FFF) offers relatively accessible equipment, broad material availability, and the ability to produce geometrically complex and customized components. When combined with open-source design initiatives, digital designs can be shared, locally adapted, and manufactured in distributed production environments, potentially expanding access to personalized mobility solutions and reducing dependence on centralized manufacturing systems [5]-[7].

Previous studies have demonstrated the application and feasibility of additive manufacturing in healthcare and assistive technology, including orthoses, prostheses, seating

Brazilian Innovation Agency (FINEP) and São Paulo Research Foundation (FAPESP).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

systems, and wheelchair accessories [5], [8]-[13]. However, investigations addressing the fabrication of complete additively manufactured pediatric wheelchairs remain scarce. Furthermore, published studies frequently emphasize prototype development, design characteristics, or mechanical performance, while practical manufacturing indicators such as printing time, material consumption, failed prints, reprinting events, material losses, and production costs, remain insufficiently documented. These indicators are essential for evaluating not only whether a device can be manufactured, but also the resources required for its production and the extent to which the workflow may be optimized, replicated, and implemented in rehabilitation laboratories or local manufacturing centers.

Open-source wheelchair designs may reduce barriers to replication and adaptation; however, their practical implementation may be affected by prolonged printing times, material-specific manufacturing difficulties, incompatibility with locally available components, and costs associated with unsuccessful prints and reprocessing. Moreover, estimates provided by design developers may not fully reflect the conditions encountered during local production, including differences in equipment, materials, infrastructure, and commercially available hardware. Systematic documentation of these factors is therefore necessary to identify manufacturing bottlenecks, quantify their effects on production efficiency and costs, and support optimization before mechanical validation and future clinical evaluation. Therefore, this study aimed to characterize the digital manufacturing workflow, evaluate production efficiency, and analyze the direct and operational costs of an open-source pediatric wheelchair fabricated using fused filament fabrication. By systematically documenting manufacturing time, material consumption, failures, reprinting events, and associated costs, this study sought to establish baseline data for workflow optimization and future mechanical and clinical validation.

II. Materials And Methods

A. Study Design

This descriptive pilot study assessed the manufacturing feasibility and production costs of a single open-source pediatric wheelchair prototype fabricated using the FFF process. The analysis focused on characterizing the digital manufacturing workflow, quantifying production efficiency, and estimating direct and operational manufacturing costs. The device was manufactured using the open-source digital files of the Toddler Mobility Trainer (TMT), developed by MakeGood Inc. in collaboration with TOM Global and LINK PBC [14]. The original structural geometry was preserved throughout the manufacturing process. However, adaptations related to printing configuration and compatibility with commercially available components were documented when required.

The study focused on the technical feasibility of the manufacturing process and the documentation of baseline production indicators. The study was not intended to assess process reproducibility across multiple devices or different manufacturing settings. The primary outcomes were the total number of print runs, total printing time, filament consumption by material type, failed prints, material losses, reprinted components, reprinting time, causes of manufacturing failures, and direct and operational production costs.

After fabrication and assembly, the prototype underwent visual and manual inspection and preliminary non-clinical functional verification under controlled indoor conditions. The verification assessed component fitting, free wheel rotation, basic maneuverability during manual pushing, visible deformation, and the occurrence of mechanical failure in the unloaded prototype. The study did not evaluate clinical effectiveness, user safety, therapeutic outcomes, long-term durability, or performance in children with mobility impairments.

B. Wheelchair Manufacturing

The wheelchair components were fabricated by FFF using the original stereolithography (STL) files provided by the project developers [14]. Printing was performed using a Bambu Lab H2D 3D printer (Bambu Lab, Shenzhen, China), with FFF tecnology, and single-nozzle printable build volume of 300 x 320 x 325 mm. The print bed surface used was the genuine Bambu Lab Textured PEI Plate. The print bed was prepared by applying a thin layer of PVP-based glue, following the printer's usage recommendations. We used the printer's standard calibration. The filament diameter was 1.75 mm. Structural components were manufactured using glycol-modified polyethylene terephthalate (PETG) filament (Voolt 3D, Brazil), whereas flexible components were produced using thermoplastic polyurethane with a Shore A hardness of 95A (TPU 95A, Slim 3D, Brazil), according to the material specifications recommended by the original project.

The printing parameters specified by the project developers were adopted as the initial manufacturing configuration, and any adjustments required during production were documented. The printing speed for the TPU components was set to 40 mm/s, following our laboratory's protocol. The main parameters used for each material are presented in Table I. The overall digital manufacturing and evaluation workflow adopted is summarized in Fig. 1.

TABLE I. FFF Parameters

Parameters	Material	
	PETG	**TPU 95A**
Bed temperature	70°	35°
Nozzle diameter (mm)	0.4	0.4
Layer height (mm)	0.2	0.2
Nozzle temperature (°C)	255	240
Infill density (%) [a]	15–20	2–15
Infill pattern	Gyroid	Gyroid
Support configuration	Tree	Tree
Outer wall speed (mm/s)	200	40

[a] The infill density varied according to the geometry and functional requirements of each component

The workflow included open-source file acquisition, file preparation and slicing, FFF printing, support removal and post-processing, inspection, final assembly, production data recording, cost analysis, and preliminary non-clinical verification. During manufacturing, the following production indicators were prospectively recorded: total number of print runs, total printing time, PETG and TPU filament consumption, number of reprinted components, material

discarded because of failed prints or rejected components; additional printing time associated with reprinting; and causes of manufacturing events requiring reprinting. A print run was defined as one complete printer operation initiated from a prepared build plate, and could include one or multiple components. A reprinted component was defined as any part manufactured again because the initial part was incomplete, defective, dimensionally incompatible, or unsuitable for assembly. After printing, support structures were removed from all components.

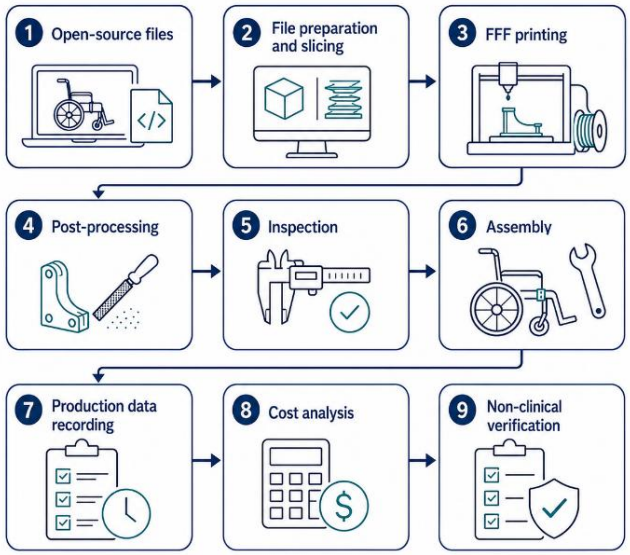


Fig. 1. Methodological workflow for the digital manufacturing, production-data recording, cost analysis, and preliminary non-clinical verification of the open-source pediatric wheelchair.

Each part underwent visual and manual inspection to identify cracks, warping, delamination, under-extrusion, incomplete layers, dimensional incompatibility, or other defects that could compromise assembly. Components were considered suitable for assembly when no visible defect affected their geometry, fitting, or apparent structural continuity. Commercial components not produced by FFF, including wheels, bearings, screws, nuts, washers, and other fasteners, were incorporated during final assembly according to the project instructions and the availability of components in the Brazilian market.

C. Economic and Technical Evaluation

The economic evaluation included direct material and hardware costs and estimated operational manufacturing costs. Direct costs were calculated from PETG and TPU consumption and from the acquisition cost of commercially available hardware required for assembly. Operational costs included equipment depreciation, electricity consumption, preventive maintenance, and labor. The cost components and manufacturing scenarios considered in the economic evaluation are summarized in Fig. 2.

All costs were based on the acquisition prices, electricity tariff, and laboratory costing assumptions applicable in April 2026. The estimated operational manufacturing cost was calculated using the laboratory hourly production rate of USD 5.20 per hour, which included equipment depreciation, electricity consumption, preventive maintenance, and labor. For each manufacturing scenario, the estimated operational cost was obtained by multiplying the hourly production rate by the corresponding printing time. The individual cost components used to derive the hourly rate are presented in Table II. Events requiring reprinting included interrupted or unsuccessful prints and components that required replacement because of power outages, filament extrusion problems, warping, incomplete printing, structural defects, or incompatibility with locally available commercial components. To quantitatively assess manufacturing efficiency, two production indicators were calculated from the prospectively recorded manufacturing data: the material loss rate and the proportion of printing time associated with reprinting.

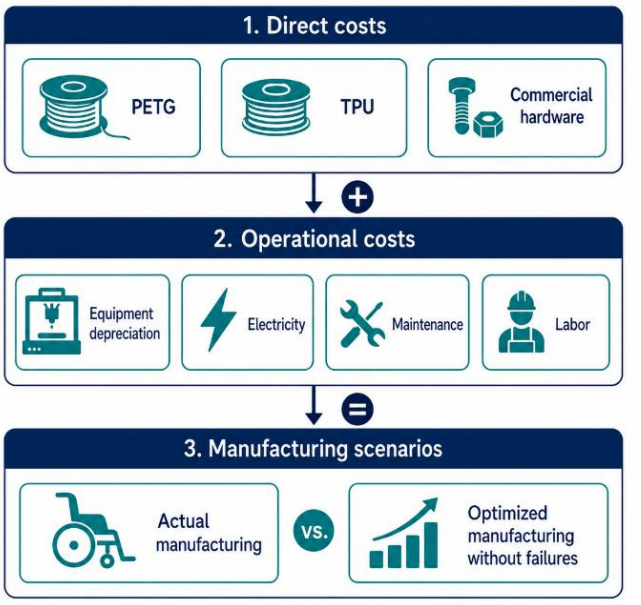


Fig. 2. Cost components and manufacturing scenarios considered in the economic evaluation.

TABLE II. HOURLY RATE COMPOSITION

Individual Cost Components	Estimated Hourly Rate (USD) [b]
Equipment depreciation	0.23/h
Electricity consumption	0.22/h
Preventive maintenance	1.00/h
Labor	3.75/h

b. Costs were calculated in Brazilian reais and converted to dollars using an approximate exchange rate of BRL 1.00 = USD 0.20 on July 26, 2026.

The material loss rate was calculated as:

$$\text{Material loss rate (\%)} = \text{M loss} / \text{M total} \times 100 \quad (1)$$

where "M loss" is the mass of filament discarded because of unsuccessful prints or rejected components that subsequently required reprinting, and "M total" is the total filament consumption across all print runs.

The proportion of printing time associated with reprinting was calculated as:

$$\text{Reprinting time proportion (\%)} = \text{T rep} / \text{T total} \times 100 \quad (2)$$

where "T rep" is the additional printing time required to replace both rejected components and the replacement components, and "T total" is the total printing time.

Two manufacturing scenarios were subsequently compared:

1. Actual manufacturing, including all recorded material consumption, unsuccessful prints, and reprinting events; and
2. A counterfactual scenario excluding the recorded material losses and 69 h of additional printing time directly attributable to manufacturing failures that resulted in discarded components.

III. Results And Discussion

The open-source pediatric wheelchair prototype was successfully fabricated using FFF and assembled from additively manufactured components and commercially available hardware. The main component groups, classified according to manufacturing material and source, are shown in Fig. 3.



Fig. 3. The main component groups used in the assembled wheelchair: (a) TPU flexible components; (b) PETG structural components, and (c) commercially available hardware.

A. Manufacturing Output and Production Efficiency

A total of 27 print runs were required resulting in 453 h of printing and a total filament consumption of 11,031 g, comprising 8,829 g of PETG and 2,202 g of TPU. Manufacturing failure events resulted in 1,543 g of discarded filament, corresponding to an overall material loss rate of 14%. Reprinting required an additional 144 h, representing 31.8% of the total printing time. Of this period, 69 h corresponded to unsuccessful prints or rejected components, whereas 75 h corresponded to replacement components that were incorporated into the final wheelchair. The material loss rate was proportionally higher for TPU than for PETG, at 24.6% and 11.3%, respectively, suggesting greater process instability during the fabrication of flexible components under the conditions adopted in this study. Furthermore, the increase in printing time was due to the reduced printing speed for the TPU parts. A printing speed of 40 mm/s was used; compared to the speed in the original file (200 mm/s), the parts were printed at 20% of that speed.

The original project developers estimated a total filament consumption of approximately 11,000 g and a printing time of 250 h. Although the observed filament consumption was similar to this estimate, the actual manufacturing time was 81.2% higher. This difference indicates that locally encountered production conditions, including printer configuration, material behavior, environmental conditions, component arrangement on the build plate, and unexpected manufacturing events, can substantially affect production efficiency.

B. Economic Evaluation

The direct cost of filament consumed during manufacturing was USD 213.71, while commercially available hardware required for assembly represented an additional USD 18.26. Therefore, the total direct material and hardware cost was USD 231.97, including material losses. In an optimized scenario excluding unsuccessful prints and discarded components, the estimated direct cost would have been USD 198.16. Thus, manufacturing failures increased the direct material and hardware cost by approximately 17.1% relative to the optimized scenario. The estimated operational cost associated with the actual manufacturing process was USD 2,355.12. In the counterfactual scenario excluding the printing time associated with manufacturing failures (69 h), the estimated operational cost would have been USD 1,995.19. Accordingly, failure events generated approximately USD 359.94 in additional operational costs. Comparable operational cost data were not reported by the original project developers. The complete cost analysis is presented in Table III.

TABLE III. Direct and Estimated Operational Costs for Manufacture of One Open-Source Pediatric Wheelchair

Cost category	Actual manufacturing (USD) [c]	Counterfactual scenario excluding failures (USD) [c]
PETG Filament	135.49	120.11
TPU Filament	78.22	59.79
Commercial hardware	18.26	18.26
Total: Direct material and hardware cost	**231.97**	**198.16**
PETG-related operational cost	1,512.98	1,331.63
TPU-related operational cost	842.14	663.56
Total: Estimated operational manufacturing cost	**2,355.12**	**1,995.19**

c. Costs were calculated in Brazilian reais and converted to dollars using an approximate exchange rate of BRL 1.00 = USD 0.20 on July 26, 2026.

The economic analysis showed that filament and commercial hardware represented only a small proportion of the estimated overall manufacturing expenditure. Although the direct material and hardware cost was approximately USD

231.97, the estimated operational cost was high. Thus, prolonged machine utilization and reprinting events, rather than raw material consumption alone, were the principal factors influencing the estimated manufacturing cost. This finding is particularly relevant for additive manufacturing workflows because economic feasibility is frequently discussed primarily in terms of filament price. However, equipment depreciation, electricity consumption, preventive maintenance, labor, and extended printing time may have a substantially greater effect on total production expenditure. Therefore, reducing printing time and preventing manufacturing failures may provide greater economic benefits than reducing filament consumption alone.

The material-specific findings also highlighted differences in manufacturing efficiency between PETG and TPU. Although TPU represented a smaller proportion of the total filament consumed, it presented a considerably higher material loss rate. Flexible filaments present material-specific processing challenges that may include extrusion and feeding instability and dimensional variability [15]. Therefore, optimization of printing parameters and production conditions may reduce manufacturing failures and improve process efficiency [15], [16]. In the Brazilian commercial market, the acquisition cost of pediatric manual wheelchairs varies according to the model, materials, adjustability, and postural support features. A market survey conducted in July 2026 identified retail prices ranging from approximately USD 558.00 to USD 813.40 (BRL 2,790.00 to 4,067.00) for standard pediatric manual wheelchair models, without customizations. These wheelchairs have standard dimensions and do not adjust as the child grows, requiring replacement over time.

C. Manufacturing Failures and Process Adaptations

Overall, 28 components required reprinting, resulting in 1,543 g of material loss and 144 h of additional printing time. Representative manufacturing defects and process-related inadequacies are in Fig. 4. The use of PETG and TPU enabled the fabrication of components with distinct functional requirements. PETG was used for structural components, whereas TPU 95A was used for tires, seating-related components, straps, and supports. This combination illustrates an important advantage of additive manufacturing in assistive technology: different polymeric materials can be selected within the same device according to the required flexibility and structural function. However, the higher loss rate observed for TPU indicates that the use of multiple materials also introduces material-specific manufacturing challenges. Visual inspection indicated that additional support structures were required around the seat-belt slots to improve print quality. Localized under-extrusion was also observed on the external surfaces of some TPU straps, suggesting minor extrusion instability. These defects did not prevent assembly; however, their mechanical significance was not assessed.

The recorded manufacturing events included a power outage, TPU nozzle clogging, incompatibility between the original digital files and locally available wheel components, warping of an adapted PETG component, and filament depletion during a long print run. The manufacturing losses resulted from a combination of external interruptions, material-specific processing difficulties, production planning, and compatibility with locally sourced hardware. Table IV shows the manufacturing indicators and their comparison with the estimates reported by the original project developers. The recorded manufacturing events included a power outage, TPU nozzle clogging, incompatibility between the original digital files and locally available wheel components, warping of an adapted PETG component, and filament depletion during a long print run. These events, together with their associated material losses and corrective actions, are presented in Table V. They indicate that manufacturing losses resulted from a combination of external interruptions, material-specific processing difficulties, production planning, and compatibility with locally sourced hardware.

Following the TPU-related events, the number of components arranged on each build plate was reduced to limit the amount of material and printing time at risk during a single run. Components that were incompatible with locally available wheels were adapted and reprinted, while all defective or incomplete parts were replaced. These actions allowed completion of the wheelchair without altering its overall structural concept. The documentation of unsuccessful prints, material losses, and reprinting events provided a more realistic assessment of manufacturing performance than reporting only the completed prototype. Such information is particularly relevant for rehabilitation laboratories and distributed manufacturing centers intending to reproduce open-source assistive technologies. Moreover, prospective recording of these indicators enabled the identification of process-specific bottlenecks and established baseline values for subsequent optimization studies.

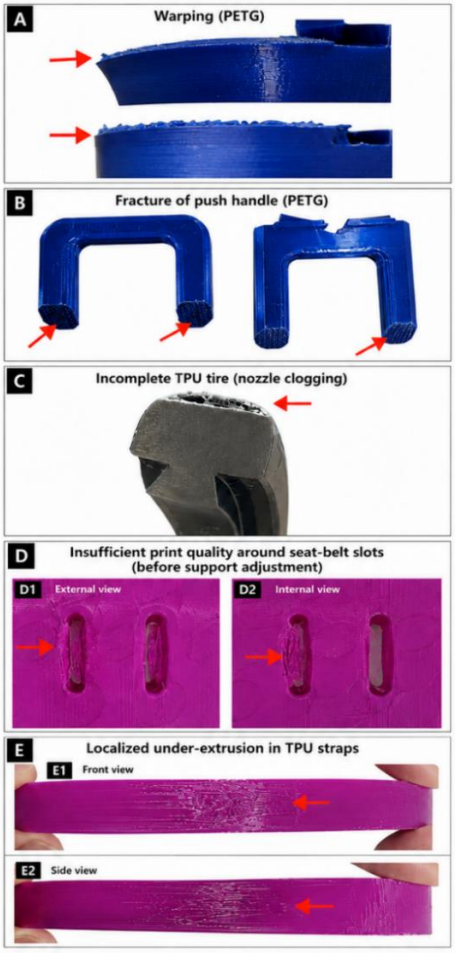


Fig. 4. Manufacturing defects and process-related inadequacies observed in FFF-produced components. (A) warped PETG component; (B) fractured PETG push handle; (C) incomplete TPU tire caused by nozzle clogging; (D) insufficient print quality around the seat-belt slots before support adjustment; and (E) localized under-extrusion in TPU straps.

TABLE IV. OBSERVED MANUFACTURING INDICATORS AND ESTIMATES REPORTED BY THE ORIGINAL PROJECT DEVELOPERS

Manufacturing Indicator	Observed Value	Original Project Estimated [14]
Print runs	27	22
Printing time (h)	453	250
PETG consumption (g)	8,829	9,000
TPU consumption (g)	2,202	2,000
Total filament consumption (g)	**11,031**	**11,000**
PETG material loss (g)	1,001	Not reported
TPU material loss (g)	542	Not reported
Reprinted components	28	Not reported
Reprinting time (h)	144	Not reported

TABLE V. RECORDED MANUFACTURING EVENTS, MATERIAL LOSSES, AND CORRECTIVE ACTIONS

Manufacturing Event	Material	Component affected	Material loss (g)	Corrective action
Power outage (external/ infrastructure)	TPU	16 tire components	181	Number of components per build plate was reduced
Nozzle clogging (material-process-related)	TPU	8 tire components	361	Number of components per build plate was reduced
Wheel incompatibility (local adaptation)	PETG	2 rear-caster components	452	Components were adapted and reprinted
Warping of adapted component (material-process-related)	PETG	1 rear caster components	226	Component was reprinted
Filament depletion (production planning)	PETG	1 push handle components	323	Component was reprinted
Total	-	**28**	**1,543**	-

D. Prototype Dimensions and Preliminary Verification

Most printed components were considered suitable for assembly after visual and manual inspection. Nevertheless, the need to reprint 28 components and the substantial additional printing time demonstrate that successful assembly alone should not be interpreted as evidence of an optimized or fully reproducible manufacturing process (Fig. 5). In assistive technology, manufacturing feasibility encompasses not only fabrication of the final device but also consistency, post-processing, assembly, local component compatibility, documentation of failures, and the possibility of implementing design adjustments. The final dimensions of the assembled wheelchair were generally similar to those reported by the original developers [14]. The original design has an approximate mass of 9 kg, width of 51 cm, height of 61 cm, and length of 66 cm. The prototype manufactured in the present study has a mass of 10.7 kg, a width of 49 cm, a height of 60.5 cm, and a length of 67 cm (Fig. 5). The observed dimensional differences may be associated with differences in measurement procedures, FFF processing, design adaptation, and the incorporation of locally sourced commercial components. However, the 1.7 kg difference in total mass should not be described as a manufacturing tolerance and requires consideration of the specific hardware and materials incorporated during assembly.

Following assembly, the unloaded wheelchair was visually and manually inspected and pushed over a flat surface in a controlled indoor environment. Wheel rotation, basic maneuverability, component fitting, visible deformation, and the occurrence of component failure were assessed. No visible structural instability, abnormal deformation, or component failure was observed during this preliminary non-clinical verification. However, no standardized load, fatigue, stability, impact, or durability tests were performed. Therefore, these observations do not demonstrate compliance with wheelchair safety standards or suitability for pediatric use. They should be interpreted only as a preliminary confirmation that the unloaded prototype could be assembled and manually moved under controlled conditions.

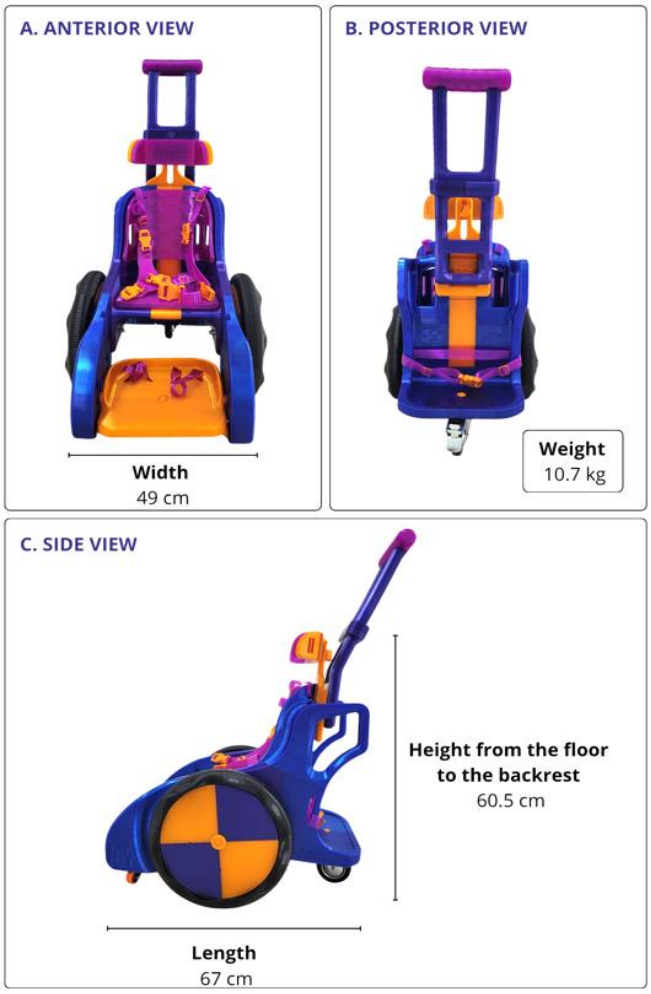


Fig. 5. Assembled open-source pediatric wheelchair (total mass and overall dimensions).

Previous feasibility studies involving additively manufactured assistive devices have similarly reported that iterative adjustment to printing parameters, material selection, and manufacturing conditions are common during prototype development [8], [17]-[18]. The present findings add quantitative information on the material, temporal, and economic consequences of these iterations during fabrication of a complete pediatric mobility device. Beyond the

quantitative manufacturing indicators, qualitative observations obtained during fabrication, assembly, and preliminary non-clinical verification provided additional insights into the strengths and current limitations of the proposed wheelchair. The WHO emphasizes that appropriate wheelchair provision requires more than product availability; it must include assessment, fitting, training, and follow-up [19].

E. Design Characteristics and Potential Future Adaptations

According to the original MakeGood Inc. design, the wheelchair is intended for children aged 1–8 years and weighing up to 20 kg [14]. Its modular frame permits adjustment of seat width through replacement of specific structural components, while the footrest can be positioned at three height levels. These characteristics indicate potential adaptability to growth; however, this adaptability was not experimentally assessed in the present study.

A limitation identified during assembly was the lack of seat-depth adjustment. Increasing seat width without a corresponding adjustment in seat depth may compromise positioning as the child grows. Appropriate seating configuration is important for postural control and functional and may also influence activity and participation in children with motor impairments [20]-[23]. Future design iterations should therefore investigate mechanisms that enable coordinated adjustment of seat width and depth while preserving structural integrity.

The open-source digital model also permits future modification or incorporation of structural and postural-support components. Potential developments include individualized seating and backrest systems, lateral trunk supports, and pelvic-positioning devices. Adaptive seating systems can improve postural control and functional performance in children with cerebral palsy [21]-[23]. Nevertheless, these components were not designed or evaluated in the present study and should be regarded as future development possibilities rather than demonstrated advantages of the current prototype. The wheelchair has a relatively low seat position, which may facilitate floor-level interaction and age-appropriate play. Independent wheelchair mobility may contribute to greater participation and engagement in family, educational, and community environments [2], [19]. However, the potential effects of seat height and wheelchair configuration on participation were not evaluated in this manufacturing study.

F. Practical Implications and Study Limitations

Feasibility studies represent an important stage in technology development because they support process refinement, guide subsequent validation, and reduce the risk of inefficient resource allocation [24]. The findings are consistent with previous evidence indicating that additive manufacturing can support the development of customized assistive technologies, particularly when local production and iterative design are required [5]-[6], [25]-[26]. In the present study, manufacturing feasibility was demonstrated by completion and assembly of a single prototype; however, the prolonged printing time, number of reprinted components, material losses, and estimated operational costs indicate that the current workflow requires substantial optimization before broader implementation. The systematic recording of production indicators established baseline values for future investigations and demonstrated that filament cost alone does not adequately represent the economic requirements of FFF-based assistive-technology production. The main observed manufacturing and design features, together with the current limitations identified during fabrication and assembly, are summarized in Table VI.

TABLE VI. MANUFACTURING ADVANTAGES AND CURRENT LIMITATIONS

Observed manufacturing and design features	Current limitations
The prototype was successfully fabricated and assembled.	Only one prototype was manufactured.
The direct material and hardware cost was relatively low compared with the estimated operational cost.	The manufacturing process required prolonged printing time and generated high estimated operational costs.
The open-source digital files permitted local production and design adaptation.	Compatibility with locally available hardware required modification of some components.
Modular construction allowed individual components to be replaced.	Seat depth was not adjustable, limiting adaptation to growth.
PETG and TPU enabled the fabrication of components with different functional requirements.	TPU presented a higher material loss rate and greater process instability under the conditions evaluated.
Individual components could be reprinted without replacing the entire device.	Mechanical performance and long-term durability were not evaluated.
The unloaded prototype could be manually moved under controlled indoor conditions without visible component failure.	Clinical safety, effectiveness, usability, and user satisfaction were not evaluated.
The digital design permits future incorporation of customized components.	Production requires additive-manufacturing expertise and appropriate infrastructure.

The principal limitations of this pilot study were the fabrication of a single prototype, the absence of standardized mechanical testing, the lack of long-term durability assessment, and the absence of clinical evaluation with the target population. The economic findings are also specific to the equipment, labor assumptions, electricity tariffs, maintenance model, and laboratory costing structure adopted in this study and should not be generalized directly to other production settings. In addition, the contrafactual scenario was a retrospective estimate based on the exclusion of recorded failure-related material consumption and printing time. It did not represent a second wheelchair experimentally manufactured under optimized conditions. Further work will include the manufacture of an additional prototype under revised production conditions to assess process reproducibility and compare manufacturing efficiency. These findings provide a foundation for subsequent studies focused on reducing printing time, optimizing TPU processing, improving compatibility with locally sourced components, validating mechanical performance according to applicable standards, and conducting ethically approved clinical evaluations of safety, usability, user satisfaction, and functional performance.

IV. CONCLUSION

This pilot study demonstrated the technical feasibility of fabricating and assembling an open-source pediatric wheelchair using FFF. The wheelchair was successfully assembled, with satisfactory integration of printed and commercial components and a visually coherent final

configuration. However, the findings also showed that manufacturing feasibility should not be assessed solely by successful prototype completion. Printing failures, material-specific processing difficulties, prolonged production time, and compatibility with locally available hardware substantially influenced workflow efficiency and estimated costs. The systematic documentation of these factors provides a basis for improving the manufacturing process before broader implementation. Further work should focus on process optimization, production of additional prototypes, and standardized mechanical validation. Only after these steps should the wheelchair proceed to ethically approved clinical evaluation of safety, usability, durability, and functional performance in children with mobility impairments.

ACKNOWLEDGMENT

The authors acknowledge the financial support provided by the Brazilian Innovation Agency (FINEP) through Project No. 2174/22, Rede INOVATA, and by the São Paulo Research Foundation (FAPESP) through the CCD Project, Grant No. 2025/07166-9. The authors also acknowledge the support of the Digital Health Center of Universidade Federal de São Paulo (UNIFESP).

REFERENCES

[1] K. Rousseau-Harrison and A. Rochette, “Impacts of wheelchair acquisition on children from a person-occupation-environment interactional perspective,” Disability and Rehabilitation: Assistive Technology, vol. 8, no. 1, pp. 1–10, 2013, DOI:10.3109/17483107.2012.670867.

[2] B. Ouellet, K. L. Best, D. Wilson, and W. C. Miller, “Exploring the influence of a community-based peer-led wheelchair skills training on satisfaction with participation in children and adolescents with cerebral palsy and spina bifida: A pilot study,” International Journal of Environmental Research and Public Health, vol. 19, no. 19, Art. no. 11908, 2022, DOI:10.3390/ijerph191911908.

[3] World Health Organization (WHO) and United Nations Children's Fund (UNICEF), “Global Report on Assistive Technology”. Geneva, Switzerland: World Health Organization, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240049451

[4] A. H. P. Costa and A. M. D. Henriques, “Projeto de uma cadeiras de rodas modular,” *Revista Foco*, vol. 17, no. 6, Art. no. e5366, 2024, DOI:10.54751/revistafoco.v17n6-056.

[5] S. Vennam, V. KN, and F. Pati, “3D printed personalized assistive devices: A material, technique, and medical condition perspective,” Applied Materials Today, vol. 40, Art. no. 102403, Oct. 2024, DOI:10.1016/j.apmt.2024.102403.

[6] L. R. Baleotti, A. L. S. da Silva, and R. M. Palacio, “Occupational therapy and assistive technology: Experience in using 3D printing and open source platforms,” Indian Journal of Physiotherapy and Occupational Therapy, vol. 18, no. 1, pp. 57–64, Jan.–Mar. 2024, DOI:10.37506/qb3e0551.

[7] M. Mottaghi, M. Woods, L. Danier, A. So, J. M. Reeves, and J. M. Pearce, “Open-source 3D printable forearm crutch,” PLoS ONE, vol. 20, no. 2, Art. no. e0318987, 2025, DOI:10.1371/journal.pone.0318987.

[8] J. V. L. da Silva, A. L. J. Munhoz, G. A. Longhitano, F. S. Sant’Ana, and M. E. Kunkel, “Aplicações na saúde,” in Tecnologias e Aplicações da Manufatura Aditiva, N. Volpato, Ed. São Paulo, Brazil: Editora Blucher, 2024, ch. 15.

[9] J. K. Schwartz, *et al*., “Methodology and feasibility of a 3D printed assistive technology intervention,” Disability and Rehabilitation: Assistive Technology, vol. 15, no. 2, pp. 141–147, 2020. DOI:10.1080/17483107.2018.1539877.

[10] M. E. Kunkel, A. Sauer, C. Isaacs, T. A. F. Ganga, L. H. Fazan, and E. K. Rorato, “Teaching Bioinspired Design for Assistive Technologies Using Additive Manufacturing: A Collaborative Experience,” Biomimetics, vol. 10, no. 6, p. 391, 2025, DOI:10.3390/biomimetics10060391.

[11] A. T. Sugawara, M. S. Oshiro, E. I. Yamanaka, V. D. Ramos, and L. R. Battistella, “Developing a rigid frame wheelchair in Brazil,” Disability and Rehabilitation: Assistive Technology, vol. 16, no. 5, pp. 538–544, 2021, DOI:10.1080/17483107.2019.1684578.

[12] S. More, M. Dunn, S. A. Rios, and R. McDonald, “Customization of manual wheelchair components: A state-of-the-art review,” Journal of Neuro Engineering and Rehabilitation, vol. 22, no. 1, Art. no. 66, 2025, DOI:10.1186/s12984-025-01547-6.

[13] F. O. Medola, C. A. Fortulan, B. M. Purquerio, and V. M. C. Elui, “A new design for an old concept of wheelchair pushrim,” Disability and Rehabilitation: Assistive Technology, vol. 7, no. 3, pp. 234–241, 2012, DOI:10.3109/17483107.2011.629327

[14] MakeGood Inc., “3D Mobility – Toddler Mobility Trainer.” [Online]. Available: https://3d-mobility.org. Accessed: Jul. 11, 2026.

[15] S. Li, Z. Shi, Y. Wang, W. Wang, and R. He, “The 3D printing of flexible materials: Technologies, materials, and challenges,” *Materials*, vol. 18, no. 23, Art. no. 5428, 2025, doi: 10.3390/ma18235428.

[16] D. G. Zisopol, M. Tănase, and A. I. Portoacă, “Innovative strategies for technical-economical optimization of FDM production,” Polymers, vol. 15, no. 18, Art. no. 3787, 2023, DOI:10.3390/polym15183787.

[17] K.-A. M. Rasmussen, B. C. Stewart, and W. E. Janes, “Feasibility of customized 3D-printed assistive technology within an existing multidisciplinary amyotrophic lateral sclerosis clinic,” Disability and Rehabilitation: Assistive Technology, vol. 18, no. 8, pp. 1466–1472, 2023, DOI:10.1080/17483107.2022.2034996.

[18] T. Kermavnar, A. Shannon, and L. W. O'Sullivan, “The application of additive manufacturing / 3D printing in ergonomic aspects of product design: A systematic review,” Applied Ergonomics, vol. 97, Art. no. 103528, 2021, DOI:10.1016/j.apergo.2021.103528.

[19] World Health Organization (WHO), “Wheelchair Provision Guidelines”. Geneva, Switzerland: World Health Organization, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240074521.

[20] D. Field and R. Livingstone, “Clinical tools that measure sitting posture, seated postural control or functional abilities in children with motor impairments: A systematic review,” Clinical Rehabilitation, vol. 27, no. 11, pp. 994–1004, 2013, DOI:10.1177/0269215513488122.

[21] D. Şahinoğlu, G. Coşkun, and N. Bek, “Effects of different seating equipment on postural control and upper extremity function in children with cerebral palsy,” Prosthetics and Orthotics International, vol. 41, no. 1, pp. 85–94, 2017, DOI:10.1177/0309364616637490.

[22] M. Angsupaisal, C. G. B. Maathuis, and M. Hadders-Algra, “Adaptive seating systems in children with severe cerebral palsy across International Classification of Functioning, Disability and Health for Children and Youth version domains: A systematic review,” Developmental Medicine & Child Neurology, vol. 57, no. 10, pp. 919–930, 2015, DOI:10.1111/dmcn.12762.

[23] R. Inthachom, S. Prasertsukdee, S. E. Ryan, J. Kaewkungwal, and S. Limpaninlachat, “Evaluation of the multidimensional effects of adaptive seating interventions for young children with non-ambulatory cerebral palsy,” Disability and Rehabilitation: Assistive Technology, vol. 16, no. 7, pp. 780–788, 2021, DOI:10.1080/17483107.2020.1731613.

[24] J. Shore, A. S. Hwang, and R. H. Wang, “Evaluating the feasibility of technology-based interventions in disability and rehabilitation: Definitions, considerations, and dimensions,” JMIR Rehabilitation and Assistive Technologies, vol. 13, Art. no. e79026, 2026, DOI:10.2196/79026.

[25] A. V. F. Santos and Z. C. Silveira, "Design for assistive technology oriented to design methodology: A systematic review on user-centered design and 3D printing approaches," Journal of the Brazilian Society of Mechanical Sciences and Engineering, vol. 43, p. 483, 2021, DOI:10.1007/s40430-021-03184-1.

[26] K.-A. M. Rasmussen, *et al*., “A systematic review of 3D-printed assistive technology rehabilitation interventions,” Assistive Technology, 2026, DOI:10.1080/10400435.2026.2619686. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10400435.2026.2619686.

ReabNet: A Web-Based Markerless Serious-Games Platform for Motor–Cognitive Rehabilitation in Parkinson's Disease

Edilaine Aparecida da Silva, Danilo Santos Rocha, Eduardo Lázaro Martins Naves
Biomedical Engineering Laboratory, Universidade Federal de Uberlândia (UFU), Uberlândia, MG, Brazil
edilaine.silva@ufu.br

***Abstract*—Camera-based serious games can support motor–cognitive rehabilitation without wearable sensors or handheld controllers. We describe ReabNet, a browser-based platform that uses a standard webcam and client-side MediaPipe BlazePose estimation to control RehaBEAT! for upper-limb reaching and Rock'in Roller for weight shifting and postural control. Game and task difficulty was progressed through a predefined, physiotherapist-supervised schedule that modified stimulus speed, required distance, base of support, and postural complexity across the 8-week protocol. During a single exploratory 902-frame live-gameplay session on a representative laptop, the pose pipeline showed an approximate median throughput of 18.9 frames/s and a median per-frame processing time of 44 ms. ReabNet was deployed in the experimental arm of an 8-week non-randomised controlled pilot involving 22 adults with mild-to-moderate Parkinson's disease; 12 participants used ReabNet and completed 16 sessions. Post-intervention usability was good (System Usability Scale 77.5 ± 9.5), exploratory first-to-last-session changes were observed in game scores, and one mild transient episode of vertigo occurred without discontinuation. These results support supervised operational feasibility but not biomechanical accuracy, because pose-derived kinematic measures were not independently validated against an optical reference system. We therefore frame the remaining validation needs—optical-reference accuracy, multi-device benchmarking, tracking-quality metrics, and component ablation—as an explicit research-gap agenda. Comparative clinical outcomes are reported separately.**

***Index Terms*—serious games, markerless pose estimation, Parkinson's disease, neurorehabilitation, browser-based rehabilitation, usability**

I. Introduction

Serious games combine structured physical exercise with gamified, goal-directed tasks and real-time feedback, and they are increasingly used in Parkinson's disease (PD) to train balance, mobility, and executive function under the dual-task demands that mirror everyday activity [1]–[3]. Two practical barriers limit their adoption. First, many systems rely on wearable sensors, instrumented mats, or handheld controllers that older adults with PD operate with difficulty because of tremor and bradykinesia. Second, research-grade exergames are often local installations tied to a specific device, which complicates deployment across a clinic and, increasingly, into the home.

Markerless motion capture, in which a standard camera and a computer-vision model estimate body pose without any attached sensor, removes the first barrier [4], [5]. Delivering the games through a web browser addresses the second: the same application can run on a clinic workstation, a tablet, or a phone, with central storage of session data for the clinician. We describe ReabNet, a platform that takes this combined approach, and report its design, its deployment considerations, and the operating assumptions that support browser-based pose tracking for coarse therapeutic game control. Clinical effects are reported separately [7]. The present contribution is engineering-focused: browser architecture, protocol-based difficulty progression, exploratory runtime measurement, clinician-facing telemetry, and platform-specific feasibility and usability.

A. Related Work

Serious games and exergames are established approaches to motor–cognitive training in PD, often using depth cameras or commercial consoles [1], [9], [10]. Browser delivery and markerless camera control are not individually novel: NeuroPose combined a webcam, browser-side PoseNet, and home-oriented geriatric rehabilitation [15], while Rehabilite Game+ integrated camera-based hand-pose games into a web platform configurable by clinicians [16]. A recent PD study also combined MediaPipe Pose with a customised exergame suite and related a derived digital-performance index to clinical measures [17]. ReabNet's contribution is therefore the evaluated integration of client-side whole-body pose control, upper- and lower-limb games, a predefined physiotherapist-supervised progression schedule, and account-linked longitudinal telemetry in a 16-session PD pilot. In contrast to systems proposing quantitative motion or digital clinical indices, ReabNet uses pose landmarks for coarse game control and does not claim biomechanical validity or a validated digital biomarker. Section V-A positions ReabNet against these systems on the dimensions that matter for browser-based, markerless deployment.

B. Scope and Relation to Companion Work

This paper reports ReabNet architecture, browser-side runtime, game-control logic, data handling, telemetry, deployment conditions, and preliminary operational evaluation. Its engineering contribution is the integration and supervised clinical deployment of these components rather than a new pose-estimation algorithm. Comparative clinical outcomes are

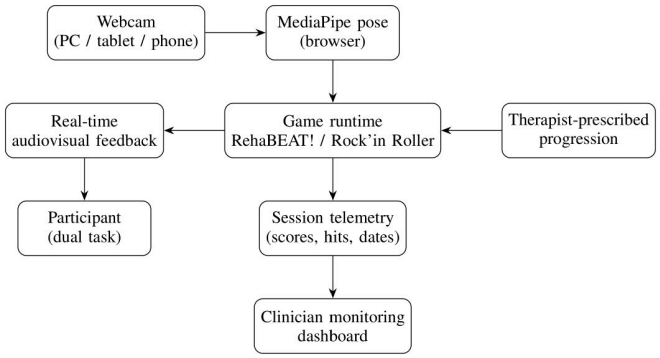


Fig. 1. ReabNet architecture. A standard webcam on a PC, tablet, or phone feeds a browser-side MediaPipe pose estimate that drives the game runtime; real-time audiovisual feedback closes the loop with the participant. Difficulty is progressed according to a predefined, physiotherapist-supervised protocol, and per-session telemetry is stored for review in the clinician monitoring dashboard.

reported in a companion study [7]; only platform-specific feasibility, usability, safety, and in-game metrics are included here.

II. System Architecture

A. Overview

ReabNet is a web application served to an ordinary browser (Fig. 1). It has two coupled sides. A clinician-facing module manages patient records, session scheduling, an exercise and game library, clinical assessments, and a monitoring dashboard. A participant-facing runtime executes the selected game, captures movement from the webcam, drives the game from the estimated pose, and sends per-session telemetry to the platform. The interface was accessible on compatible desktops, tablets, and smartphones without a separate installation; formal runtime measurement in this study was restricted to one Windows laptop running Chrome.

B. Implementation and Deployment

ReabNet is implemented as a single-page web application. Pose estimation runs entirely in the browser: the MediaPipe Pose (BlazePose) solution is loaded as a WebAssembly (SIMD) build from a content-delivery network and executed on the live webcam stream, and the estimated landmarks drive game logic that is rendered with WebGL. No client-side installation, browser plugin, or dedicated graphics card is required, so the same URL runs on a clinic workstation, a tablet, or a smartphone. Session configuration, telemetry, and clinician assessments are persisted on a server managed by the platform and retrieved through the clinician dashboard. Because inference is performed on the client, the network carries only lightweight telemetry rather than a video stream, which keeps the platform usable on ordinary connections; the trade-off is that responsiveness depends on the client device, as quantified in Section III-A.

C. Markerless Pose Estimation

Movement is captured from a standard RGB webcam and converted to a whole-body pose estimate in real time by MediaPipe Pose (BlazePose) [5], [6], which returns 33 landmarks. The model runs client-side in the browser through a WebAssembly (SIMD) build, with no markers or wearables; `modelComplexity` is 1 and minimum detection confidence is 0.5. Landmarks are mapped to discrete game actions, such as hand contact with a target or knee elevation for obstacle clearance, and the games are rendered in WebGL. In the clinical deployment, the webcam was approximately 2 m from the participant, centred on the body midline and at eye level; games were shown on a 20-inch monitor. Body detection was checked before each session, and a physiotherapist supervised all tasks, particularly tandem, unstable-surface, and single-leg conditions. The model is used as a fixed, pre-trained component; because its accuracy for the present cohort was not independently verified, Section III states the operating assumptions explicitly and Section VI sets out the validation still required.

D. Game-Control Logic and Participant Positioning

RehaBEAT! uses the estimated on-screen position of either hand as the control input (Fig. 2). A coloured block is registered as hit when the hand position overlaps it; no minimum dwell time is required. Bombs are negative game elements that must be avoided. In Rock'in Roller, knee elevation controls target contact and obstacle clearance. In the target-reaching task, the participant raises the knee until its estimated position overlaps the on-screen target. In the obstacle task, knee elevation triggers the jump action required to clear the obstacle.

No automated anthropometric calibration was used. Before each session, the physiotherapist positioned the participant and camera at the predefined distance and alignment, confirmed whole-body detection, and manually selected game level, stimulus speed, distance, and postural condition. Participants viewed the estimated body representation in real time. If tracking was temporarily lost, the game continued without automatic pause; the supervising physiotherapist repositioned the participant when necessary. This behaviour limits unsupervised use and motivates confidence-based warnings and automatic pause-and-resume logic, discussed in Section VI.

E. Configurable Session Parameters

Each game exposes clinician-configurable parameters that are set from a preferences panel before a session begins. For RehaBEAT!, these include the active hand (left, right, or both), the stimulus speed (slow, medium, or fast), the presence or absence of bomb obstacles, and the target size and spacing, which together set the spatial precision and temporal pressure of the reaching task. Rock'in Roller exposes analogous speed and obstacle-density settings. Combined with the postural condition selected by the physiotherapist, these parameters let a single game span a broad range of motor and cognitive demand without any change to the code, and every selected value is stored with the session record so that the configuration under

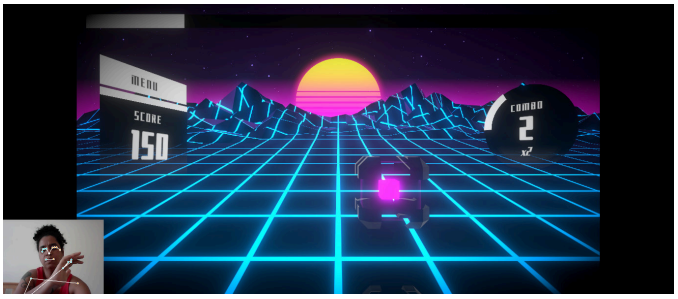


Fig. 2. RehaBEAT! gameplay. The participant reaches to contact rhythmic targets while avoiding bombs; the running score and combo multiplier are displayed. The lower-left inset is the live webcam view with the client-side MediaPipe pose skeleton overlaid, which drives game control.



Fig. 3. Clinician monitoring dashboard. For each patient the module aggregates standardised assessment scales (e.g., Berg Balance Scale, shown) and per-session game metrics, supporting longitudinal review of training and clinical progress.

which each score was obtained is auditable. This parameter logging is a prerequisite for the performance-driven automatic adaptation proposed in Section VI.

F. Therapeutic Games

The library is organised by therapeutic target. RehaBEAT! trains bilateral upper-limb reaching, elevation, abduction, and bimanual coordination. Participants reach to contact rhythmic coloured blocks while avoiding bombs. Consecutive block hits increase the combo multiplier and score. Missing a block or being hit by a bomb resets the combo to zero, and bomb contact also incurs a score penalty. Successive levels increase stimulus speed and density.

Rock'in Roller trains lower-limb and trunk control through weight transfer, mediolateral and anteroposterior displacement, partial squats, and progression toward single-leg stance. Challenge mode uses timed target reaching, whereas Infinite mode requires obstacle clearance through knee elevation. Both games combine motor execution with visual selection, sequencing, and time pressure and therefore produce discrete game-performance events rather than laboratory-grade continuous kinematic outcomes.

G. Protocol-Based Difficulty Progression

Difficulty was progressed according to a predefined, physiotherapist-supervised schedule. Every two weeks, the protocol increased motor and cognitive demands by changing the base of support (parallel stance, tandem stance, unstable surface, and single-leg stance), stimulus speed, and/or required distance. Software-level changes in game speed and task demand were coordinated with clinically prescribed changes in stance and balance conditions. This approach preserved clinician control over task selection, postural safety, and progression, at the cost of manual configuration—a dependence that Section VI addresses through performance-driven adaptation.

H. Telemetry and Clinician Monitoring

Each session stores participant-linked telemetry for longitudinal clinician review (Fig. 3). For RehaBEAT!, the dashboard reports session date, completion percentage, selected hand condition, coloured-block count, stimulus speed, bomb condition, score, maximum combo, overall hit rate, side-specific hit rates, and bomb-avoidance rate. Rock'in Roller likewise records score and obstacle-related performance across sessions. The game demands differ: RehaBEAT! requires upper-limb contact with coloured blocks while avoiding bombs, whereas Rock'in Roller requires knee elevation to trigger obstacle-clearing jumps. The first- and last-session scores and RehaBEAT! bomb-contact events used in the clinical analysis were exploratory outcomes; no game-derived variable is interpreted as a validated biomechanical or clinical measure.

I. Data Handling, Access, and Privacy

The operational platform identified participants by full name. Session telemetry and pose-landmark coordinates were stored on a server managed by the ReabNet startup. Camera-derived recordings were also retained with participant consent; however, the available technical documentation did not establish whether the retained material comprised complete video, still images, or selected frames, or whether upload occurred independently of client-side pose inference. Clinicians authenticated using unique logins and could access all patients registered under their own account, but not patient lists assigned to other clinician accounts. Startup engineers could access stored data for technical purposes. Encryption in transit, access logging, backup procedures, retention and deletion schedules, and formal security testing were not verified in this study; consequently, no formal cybersecurity or regulatory-compliance claim is made. The regulatory and scalability implications of these gaps are discussed in Section V-C.

III. Operational Responsiveness and Tracking Assumptions

ReabNet relies on the pre-trained MediaPipe BlazePose model [5], [6], whose performance against reference motion-capture systems has been characterised in prior work [11], [12]. An independent optical-reference validation was not performed in the present cohort.

The games were designed around coarse whole-body actions—reaching to sizeable targets, shifting body weight, and elevating the knee—rather than fine joint-angle estimation. Body detection was checked before each session and a physiotherapist supervised participants. The present evaluation

TABLE I
EXPLORATORY SINGLE-SESSION CLIENT-SIDE RUNTIME MEASUREMENT (902 FRAMES). THIS IS A DESCRIPTIVE MEASUREMENT ON ONE CONFIGURATION, NOT A MULTI-DEVICE BENCHMARK (SEE SECTION VI).

Device	Browser	Camera	FPS	Time (ms)
i5-8265U, UHD 620	Chrome	640×480	18.9	44 [34–48]

therefore addresses operational responsiveness and functional game control under supervision, not biomechanical accuracy. Optical-reference validation remains necessary before pose-derived kinematic measures can be proposed as digital biomarkers; Section VI specifies the metrics and protocol this validation should adopt.

A. Client-Side Runtime

An exploratory runtime measurement was conducted during a single live-gameplay session using a Windows 11 64-bit laptop equipped with an Intel Core i5-8265U processor (4 physical cores and 8 logical processors), 12 GB of DDR4 RAM, integrated Intel UHD Graphics 620, Google Chrome, and a 640×480 webcam. A total of 902 consecutive frames were analysed. The median inter-frame interval was 53 ms, corresponding to an approximate throughput of 18.9 frames/s. Median per-frame processing time was 44 ms, with an observed range of 34–48 ms (Table I). Because the measurements were obtained from a single session on one hardware configuration, they are interpreted as descriptive runtime measurements rather than a generalisable performance benchmark. Replication across sessions, users, browsers, and devices—and across camera resolutions and lighting conditions—is required to characterise variability, frame loss, jitter, and thermal stability; this multi-factor benchmarking is formalised as a research gap in Section VI.

B. Operational Evaluation Methods

Platform-specific operational outcomes comprised protocol completion, post-intervention usability, adverse events, and exploratory first-to-last-session game metrics. Protocol completion was defined as completion of all 16 supervised sessions; missed appointments were rescheduled. All 12 ReabNet participants provided complete platform-specific data, and no imputation was required.

After the 16th session, usability was assessed using the Portuguese-language System Usability Scale (SUS). The questionnaire was administered by an evaluator, and the therapist provided clarification of item wording when requested. Standard SUS scoring was used, including reverse scoring of the negatively worded items, to obtain a total score from 0 to 100. Potential adverse events were actively assessed and recorded at every session. For exploratory in-game analyses, each participant's first and last completed sessions were compared for RehaBEAT! score, RehaBEAT! bomb-contact events, and Rock'in Roller score. Analyses were performed in Python. Because of the small sample and the ordinal or non-normally distributed game metrics, paired first-to-last-session comparisons used Wilcoxon signed-rank tests. Results are reported as median [interquartile range]. The analyses were exploratory; no adjustment for multiple comparisons was applied, and p values should not be interpreted as confirmatory evidence of clinical efficacy.

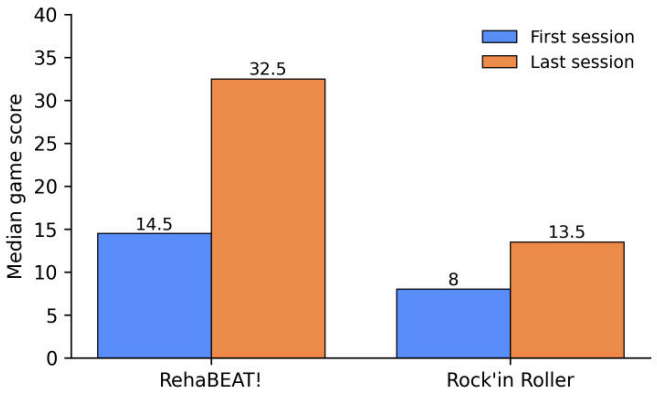


Fig. 4. Median RehaBEAT! and Rock'in Roller scores at the first versus the last completed session in the ReabNet arm ($n = 12$; values from Table II). Exploratory; changes may reflect familiarisation and protocol-driven difficulty changes.

TABLE II
EXPLORATORY IN-GAME PERFORMANCE IN THE REABNET ARM ($n = 12$). VALUES ARE MEDIAN [IQR]; p FROM WILCOXON SIGNED-RANK TESTS. RESULTS ARE EXPLORATORY AND MAY REFLECT TASK FAMILIARISATION AND CHANGING DIFFICULTY.

Metric	First [IQR]	Last [IQR]	p
RehaBEAT! score	14.5 [12.0–16.0]	32.5 [29.0–34.0]	< 0.001
RehaBEAT! bombs	1.0 [1.0–2.0]	0.0 [0.0–0.25]	0.002
Rock'in Roller score	8.0 [5.0–10.5]	13.5 [10.0–15.0]	0.0005

IV. CLINICAL FEASIBILITY

ReabNet was deployed in the experimental arm of an 8-week non-randomised active-comparator pilot involving 22 adults with mild-to-moderate PD (Hoehn & Yahr 1–3): 12 received ReabNet and 10 received dose-matched conventional physiotherapy [7]. The ReabNet arm completed 16 supervised sessions over 8 weeks; occasional absences were rescheduled and made up. All 12 ReabNet users completed the protocol. Post-intervention usability was good (System Usability Scale 77.5 ± 9.5; median 77.5; range 62.5–95) [13], [14]. One mild transient episode of vertigo occurred during ReabNet use, with complete recovery and no discontinuation. Exploratory first-to-last-session in-game performance is summarised in Table II and Fig. 4. Comparative clinical outcomes are reported separately [7].

V. DISCUSSION

ReabNet demonstrates that therapeutic games controlled by client-side markerless pose estimation can be delivered through an ordinary browser and operated on commodity

TABLE III
POSITIONING OF REABNET AGAINST REPRESENTATIVE APPROACHES. QUALITATIVE SUMMARY FROM THE CITED REPORTS; NOT A BENCHMARKED HEAD-TO-HEAD COMPARISON.

Feature	ReabNet	NeuroPose	Reh. G+	Wearable
No attached hardware	Yes	Yes	Yes	No
Browser-based	Yes	Yes	Yes	Varies
Whole-body (UL+LL)	Yes	Partial	UL/hand	Yes
Clinician telemetry	Yes	Partial	Yes	Varies
Validated tracking	No†	No†	No†	Yes
Digital-biomarker claim	No	No	No	Varies

† Uses a pre-trained pose model without cohort-specific optical-reference validation.

hardware under supervised clinical conditions. The exploratory runtime measurement, complete protocol among the 12 ReabNet users, good SUS score, and longitudinal telemetry support operational feasibility and learnability. They do not demonstrate tracking accuracy or clinical efficacy; comparative effects are addressed separately [7]. Observed first-to-last-session changes in scores and bomb-contact counts may reflect engagement, but are confounded by familiarisation, protocol-driven level changes, and scoring rules and should not be interpreted as direct motor-recovery measures.

A. Comparison with Related Camera- and Sensor-Based Systems

Table III positions ReabNet against representative camera-based and wearable rehabilitation approaches. Relative to NeuroPose [15] and Rehabilite Game+ [16], ReabNet adds whole-body (upper- and lower-limb) control and an account-linked longitudinal telemetry dashboard evaluated in a 16-session PD cohort; relative to the digital-index approach of Ferraris et al. [17], ReabNet deliberately restricts pose use to coarse game control and makes no digital-biomarker claim. Against wearable- and depth-camera systems, the browser-based markerless design removes attached hardware and per-device installation but shifts the burden to per-deployment runtime characterisation and to the tracking validation that wearables obtain by construction. A controlled, head-to-head comparison of tracking robustness and clinical utility against these alternatives was not performed and is identified as a research gap in Section VI.

B. Limitations of 2D Markerless Pose in Clinical Use

ReabNet's control signal is a single-camera 2D pose estimate. Two limitations follow directly. First, monocular pose lacks metric depth and true joint angles, so movements dominated by out-of-plane components (anteroposterior weight shift, trunk rotation) are captured only approximately; this is acceptable for coarse game control but precludes kinematic interpretation. Second, estimation quality degrades under self-occlusion, atypical body proportions, loose clothing, and non-frontal camera placement, and PD-specific factors—resting tremor, reduced amplitude, festination, and stooped posture—can further perturb landmark stability in ways that a healthy-adult-trained model may not represent well. The coarse, large-target game design tolerates these errors, but their magnitude in a PD cohort is unquantified. We therefore avoid any claim that ReabNet measures movement quality, and we treat pose only as a game-control input.

C. Scalability, Security, Regulatory Compliance, and Unsupervised Use

Client-side inference scales favourably: because only compact telemetry crosses the network, server cost grows with session metadata rather than video bandwidth, and the same URL serves heterogeneous devices. However, three deployment barriers remain. (i) *Security and privacy:* the study did not verify transport/at-rest encryption, access logging, backup, or defined retention and deletion, and engineers could access identifiable stored data; least-privilege role-based access and auditable data governance are prerequisites for real-world use. (ii) *Regulatory compliance:* a supervised research prototype is distinct from a regulated medical device—a production version handling health data would need to align with the applicable data-protection law (e.g., Brazilian LGPD) and, if therapeutic claims were made, with software-as-a-medical-device requirements; no such claim is made here. (iii) *Unsupervised/home use:* the absence of automatic pause on tracking loss and the reliance on a supervising physiotherapist for positioning and level selection currently preclude safe unsupervised deployment. A performance-driven automatic difficulty adaptation—for example, adjusting stimulus speed and target spacing from the trailing hit rate and bomb-contact rate already logged per session—would reduce manual configuration and is a natural step toward home use, provided it is paired with confidence-based tracking warnings and safety limits.

VI. LESSONS LEARNED AND RESEARCH GAPS

A. Lessons Learned

Three engineering lessons emerged from developing and deploying ReabNet. First, *client-side inference is an accessibility enabler but a characterisation liability*: exchanging only telemetry keeps the platform usable on ordinary connections, yet it makes responsiveness device-dependent and therefore demands per-deployment runtime characterisation rather than a single benchmark. Second, *coarse game control tolerates imperfect pose, but only under supervision*: large targets and body-scale actions absorb landmark noise, while the lack of automatic pause on tracking loss showed that robustness features, not just accuracy, gate unsupervised use. Third, *auditable configuration is as important as the games themselves*: logging every clinician-selected parameter with each score was what made exploratory in-game changes interpretable, and it is the foundation on which automatic, performance-driven adaptation can later be built.

B. Research Gaps and Future Work

The reviewer-identified needs are consolidated below as a structured agenda; none were addressed in the present study and each is stated as future work rather than a current result.

(G1) Optical-reference tracking validation. Quantify the accuracy and reliability of BlazePose landmarks in a PD cohort against a marker-based optical reference (e.g., Vicon), reporting per-landmark positional error, temporal stability (jitter), detection-loss rate, and robustness to occlusion, body type, tremor, and PD-typical postures.

(G2) Multi-configuration performance benchmarking. Characterise runtime across devices (laptop, tablet, phone), browsers, camera resolutions, and lighting conditions, over multiple sessions, reporting frame loss, jitter, and thermal stability rather than a single-session median.

(G3) Component ablation. Isolate the contribution of the progression protocol, audiovisual feedback, configurable parameters, and interaction strategy to engagement and in-game performance, so that design choices are evidence-based rather than assumed.

(G4) Controlled comparison. Benchmark ReabNet head-to-head against alternative computer-vision and wearable rehabilitation solutions on tracking robustness, usability, and clinical utility.

(G5) Objective tracking-quality telemetry. Add detection-loss rate, landmark temporal stability, and confidence-based warnings with automatic pause-and-resume to the runtime, both to support unsupervised use and to expose tracking quality as a monitored signal.

(G6) Automatic difficulty adaptation. Replace manual level selection with performance-driven adaptation derived from the logged hit-rate and bomb-contact telemetry, reducing dependence on the supervising physiotherapist.

(G7) Governance and regulatory readiness. Implement and audit least-privilege access, encryption in transit and at rest, access logging, and defined retention/deletion, and align data handling with applicable health-data regulation.

(G8) Clinical validation at scale. Pursue multicentre and longitudinal evaluation, home-deployment testing without continuous supervision, biomechanical validation in real rehabilitation settings, and integration with AI models for therapeutic personalisation.

VII. Conclusion

ReabNet is a browser-based serious-games platform combining client-side markerless pose estimation, therapist-prescribed task progression, and clinician-facing telemetry. In a single exploratory 902-frame session on the tested laptop, it showed an approximate median throughput of 18.9 frames/s and a median per-frame processing time of 44 ms; all 12 ReabNet users completed the supervised 16-session protocol, usability was good, and exploratory first-to-last-session changes were observed in game scores; these comparisons were confounded by familiarisation and protocol-driven changes in difficulty. These findings establish preliminary operational feasibility, not biomechanical validation. Optical-reference testing, event-recognition validation, multi-device characterisation, component ablation, and home-deployment evaluation—consolidated here as an explicit research-gap agenda—remain necessary before broader implementation.

Acknowledgment

The authors thank the participants and the Biomedical Engineering Laboratory at UFU. This study was financed in part by CAPES (Finance Code 001), FINEP, and FAPEMIG. The funders had no role in study design, data collection, analysis, interpretation, or the decision to submit.

Ethics Approval and Data Availability

The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of the Universidade Federal de Uberlândia (CAAE 83857524.7.0000.5152; approval no. 7.196.916). All participants provided written informed consent, including consent for image or video recording and storage. The anonymised research dataset and analysis scripts—distinct from identifiable operational records retained by the platform—are openly available in Zenodo under a CC BY 4.0 licence (doi:10.5281/zenodo.20607087) [8].

Conflict of Interest

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

[1] G. Barry, B. Galna, and L. Rochester, "The role of exergaming in Parkinson's disease rehabilitation: a systematic review," *J. NeuroEng. Rehabil.*, vol. 11, p. 33, 2014.

[2] J. E. Pompeu *et al.*, "Effect of Nintendo Wii-based motor and cognitive training on activities of daily living in Parkinson's disease: a randomised clinical trial," *Physiotherapy*, vol. 98, no. 3, pp. 196–204, 2012.

[3] A. Mirelman *et al.*, "Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial," *Lancet*, vol. 388, pp. 1170–1182, 2016.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE CVPR*, 2017, pp. 7291–7299.

[5] C. Lugaresi *et al.*, "MediaPipe: a framework for building perception pipelines," arXiv:1906.08172, 2019.

[6] V. Bazarevsky *et al.*, "BlazePose: on-device real-time body pose tracking," arXiv:2006.10204, 2020.

[7] E. A. da Silva, D. S. Rocha, and E. L. M. Naves, "Feasibility and preliminary effects of a markerless serious-games rehabilitation programme (ReabNet) versus conventional physiotherapy on balance and cognition in Parkinson's disease: a non-randomised controlled pilot study," manuscript under review, 2026.

[8] E. A. da Silva, "Synthetic data methods for small-sample Parkinson's disease studies: anonymised dataset and analysis scripts," Zenodo, 2026, doi:10.5281/zenodo.20607087.

[9] L. C. de Oliveira *et al.*, "A systematic review of serious games used for rehabilitation of individuals with Parkinson's disease," *Res. Biomed. Eng.*, 2021, doi:10.1007/s42600-021-00162-3.

[10] A. Garcia-Agundez *et al.*, "Recent advances in rehabilitation for Parkinson's disease with exergames: a systematic review," *J. NeuroEng. Rehabil.*, vol. 16, art. 17, 2019, doi:10.1186/s12984-019-0492-1.

[11] C. S. T. Hii *et al.*, "Automated gait analysis based on a marker-free pose estimation model," *Sensors*, vol. 23, no. 14, art. 6489, 2023, doi:10.3390/s23146489.

[12] B. J. Alammari *et al.*, "Validity of AI-driven markerless motion capture for spatiotemporal gait analysis in stroke survivors," *Sensors*, vol. 25, no. 17, art. 5315, 2025.

[13] J. Brooke, "SUS: a quick and dirty usability scale," in *Usability Evaluation in Industry*. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

[14] A. Bangor, P. T. Kortum, and J. T. Miller, "Determining what individual SUS scores mean: adding an adjective rating scale," *J. Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.

[15] S. R. Rick *et al.*, "NeuroPose: geriatric rehabilitation in the home using a webcam and pose estimation," in *Proc. 24th Int. Conf. Intelligent User Interfaces Companion*, 2019, pp. 105–106, doi:10.1145/3308557.3308682.

[16] J. E. Fontenele *et al.*, "Rehabilite Game+: pose estimation-based serious games for forearm and hand movement rehabilitation," *Entertainment Computing*, vol. 55, art. 100992, 2025, doi:10.1016/j.entcom.2025.100992.

[17] C. Ferraris *et al.*, "Digital monitoring of motor function in Parkinson's disease using markerless motion analysis and exergaming," *Front. Neurol.*, vol. 17, art. 1800332, 2026, doi:10.3389/fneur.2026.1800332.

Preliminary Bench Validation of a Low-Cost Wearable Inertial Sensing Platform for Tremor Monitoring

Gustavo Adolpho Bonesso
Science and Technology Institute
Federal University of São Paulo
São José dos Campos, Brazil
0000-0002-8691-7607

Carlos Marcelo Gurjão de Godoy
Science and Technology Institute
Federal University of São Paulo
São José dos Campos, Brazil
0000-0001-8846-3242

Eliana Chaves Ferretti
Institute of Health and Society
Federal University of São Paulo
Santos, Brazil
0000-0002-7605-828X

Regina Célia Coelho
Science and Technology Institute
Federal University of São Paulo
São José dos Campos, Brazil
0000-0002-4428-9745

***Abstract*— Essential tremor is a neurological disorder that typically causes tremors in the upper limbs. These tremors are usually in the 2 Hz to 10 Hz range. This work presents a preliminary metrological bench validation of a low-cost wearable device, based on an ESP32-C3 microcontroller and an accelerometer integrated into a passive weighted bracelet. The development of such assistive technologies requires characterization of their embedded sensing capabilities under controlled conditions. The system was evaluated using a damped harmonic oscillator bench test and a laboratory rotary motion sensor used as a reference instrument. The measurements from the accelerometer and the laboratory rotary sensor were compared in the frequency domain using the Fast Fourier Transform (FFT) across several tests with different springs and bracelet weight configurations. Results showed that the embedded sensor was able to detect the mass-dependent reduction in the natural frequency of the oscillating system across different experimental configurations. The embedded system measured a frequency drop from 2.97 Hz to 1.66 Hz in a configuration when a 137-gram mass was added to the bracelet. The results demonstrated good agreement with the reference sensor for the tested frequencies between approximately 1 and 3 Hz, while frequencies below 1 Hz showed substantial errors. These findings support the preliminary technical feasibility of the sensing module under controlled bench conditions, but validation across the full 2–10 Hz tremor range and in human subjects remains necessary.**

Keywords*— *Sensor validation, Inertial sensors, Damped harmonic oscillator, Fast Fourier Transform, Essential tremor

I. Introduction

Essential Tremor (ET) is a chronic neurological movement disorder characterized by rhythmic and involuntary oscillations of the upper limbs, with frequencies predominantly ranging between 2 Hz and 10 Hz [1], [2]. Essential tremor affects fine motor control, impairing patients' autonomy in basic daily activities such as eating, writing, and personal care [3], [4], [5]. Wearable technologies, such as those based on inertial measurement units (IMUs), have been used to assess and monitor these tremors outside the clinical environment [6], [7], [8].

The determination of the dominant tremor frequency is typically performed using frequency-domain signal processing techniques, such as the Fast Fourier Transform (FFT). The current challenge in biomedical instrumentation is how to adopt low-cost embedded platforms capable of continuous data acquisition while providing frequency measurements with acceptable agreement with laboratory reference instruments.

In parallel with the digital monitoring performed by embedded systems, the use of passive assistive technologies—such as bracelets with incorporated weights—exploits the physical principles of inertia and damping, aiming to reduce involuntary oscillatory movement [9], [10]. The mechanical dynamics of this attenuation resemble those of a damped harmonic oscillator (mass-spring system). According to the fundamental laws of dynamic systems modeling, increasing the mass of an oscillatory system directly decreases its natural frequency of oscillation.

Therefore, the main objective of this study is to perform a preliminary metrological characterization of a low-cost embedded inertial sensing module integrated into a modular weighted bracelet, designed via additive manufacturing for continuous tremor monitoring. The system was evaluated in a standardized bench test simulating a damped harmonic oscillator, in which raw acceleration data were simultaneously compared with measurements from a laboratory rotary motion sensor serving as a reference sensor.

To evaluate the system's frequency measurement behavior under controlled conditions, kinematic data from both devices across multiple trials were analyzed using the Fast Fourier Transform. Thus, this work aims to demonstrate the low-cost device's ability to characterize both the frequency reduction associated with the addition of mass and the oscillatory frequencies relevant to ET under the tested bench conditions.

II. Methods

To evaluate the frequency measurement performance of the proposed low-cost embedded system, a controlled bench test based on a damped harmonic oscillator was designed. The

This work was developed using resources obtained from the FINANCIADORA DE ESTUDOS E PROJETOS (FINEP) - Ref. Finep nº 2174/22.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

experimental approach emphasizes the technical feasibility of using accessible hardware for frequency monitoring and compares its performance against a laboratory reference sensor through frequency-domain analysis.

A. Bench Test Setup and Data Acquisition

To perform the experiment, we used a 3D-printed bracelet with a weight container suspended by a spring mechanism to simulate the oscillatory behavior of a human arm. The embedded sensing module **(Fig. 1)** has an ESP32-C3 microcontroller and an MPU-6050 Inertial Measurement Unit (IMU). The electronic architecture is detailed in the block diagram **(Fig. 2)** and also includes power management and data communication between the components. The IMU captured the raw 3-axis acceleration data using a sampling rate of 100 Hz during the experiments.

To initiate each bench trial, we manually displaced the mass-spring-bracelet system vertically from its equilibrium position by approximately 5.0 cm and released it from rest, generating a consistent damped harmonic oscillation. For each physical configuration, the IMU continuously captured raw triaxial acceleration data at a sampling rate of 100 Hz for a continuous duration of 15 seconds, ensuring the complete decay of the oscillatory motion was captured. Although the embedded system continuously acquired data throughout the 15-s bench trials, the present study focused on frequency-domain agreement with the reference sensor. Therefore, latency, long-term continuous operation, power consumption, and communication stability were not quantitatively assessed in these experiments.

We used a PASCO rotary motion sensor (model PS-2120) as the laboratory reference for our validation, coupling it mechanically to the mass-spring system **(Fig. 3)**. This reference sensor recorded the system's angular position simultaneously with the embedded module. The sampling rate of the reference sensor was also set to 100 Hz through the DataStudio software. To align the datasets from these independent hardware acquisitions, we implemented a temporal synchronization protocol in post-processing. We aligned the synchronized datasets by identifying the absolute onset peak of the initial transient oscillation, utilizing cross-correlation to achieve sample-to-sample synchronization. This dual-acquisition strategy allowed us to obtain a synchronized dataset, enabling a direct comparison between the measurements.

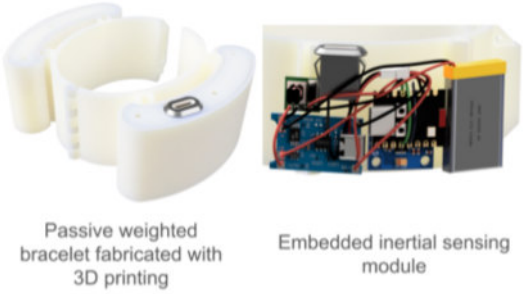


Fig. 1. Overview of the proposed wearable system. On the left, the passive weighted bracelet fabricated via additive manufacturing, with its weight-distribution containers. On the right, the embedded inertial sensing module detailing the ESP32-C3 microcontroller, the triaxial accelerometer, and the power supply.

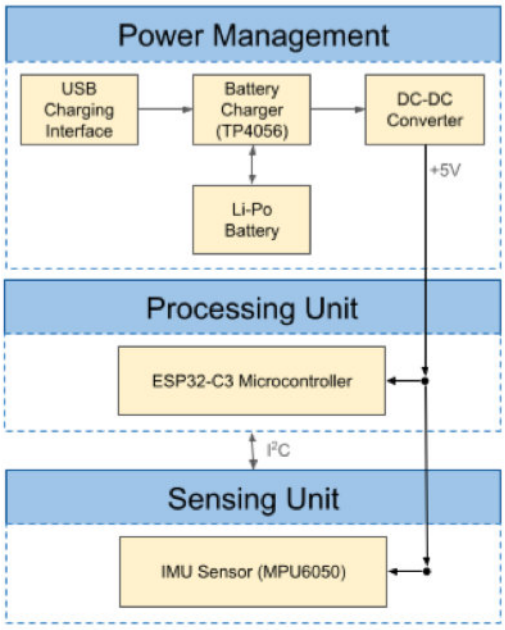


Fig. 2. Functional block diagram of the embedded sensing module. The architecture is divided into a power management unit (battery charger and DC-DC converter), a processing unit (ESP32-C3), and a sensing unit (MPU-6050), illustrating the 5V power supply lines and the I2C communication protocol.

B. Mass-Induced Frequency Reduction Evaluation

To evaluate whether the sensor can detect changes in oscillation frequency resulting from the addition of mass to the bracelet, we added mass to the mass-spring system. Increasing the mass of the system decreases its natural frequency of oscillation.

We started testing the system using Spring 1 with the bracelet's weight compartment empty. Next, we ran the test again with the weight compartment completely filled by adding 137g of lead. We conducted additional tests using different configurations (other springs and alternating the weight compartment between full and empty) to assess the consistency of the embedded system in detecting frequency changes associated with the addition of mass.

C. Frequency-Domain Validation Protocol

To evaluate the frequency measurement performance of the low-cost sensor, we processed the synchronized kinematic signals from both devices in the frequency domain. Because the mechanical oscillation of the suspended spring-mass system occurred strictly along a single vertical axis, we selected only the vertical acceleration component (a_z) of the MPU-6050 for analysis, thereby eliminating uninformative noise from the orthogonal axes.

First, we removed the Direct Current (DC) component from both the PASCO angular displacement and the accelerometer signals by subtracting their respective mean values to eliminate baseline offsets. To mitigate spectral leakage and windowing

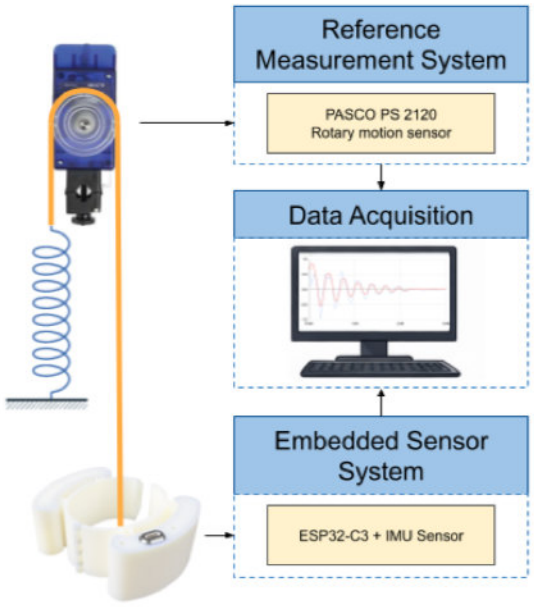


Fig. 3. Setup of the damped harmonic oscillator bench test connecting the mass-spring system to the bracelet, carried out to validate the bracelet's embedded sensing module. The functional block diagram shows that data from the bracelet's sensing module is collected simultaneously with data

artifacts, we applied a Hann window to the time-domain signals prior to spectral analysis.

We then calculated the frequency spectrum using the Fast Fourier Transform (FFT) with a window size of N=1024 points. Although each trial was recorded for a continuous duration of 15 s, not all samples were used in the frequency-domain analysis. From each 15-s recording, we manually selected a 1024-sample segment (10.24 s at 100 Hz) corresponding to the free-decay portion of the oscillation. Specifically, we excluded the initial interval in which the bracelet was still being manually displaced from equilibrium — during which acquisition had already started but release had not yet occurred — as well as the final interval in which the oscillation had already decayed to a near-flat signal. This selection ensured that the analyzed segment captured the actual damped oscillatory motion while excluding the pre-release handling artifact and the low-amplitude tail, both of which could otherwise degrade the accuracy of the FFT-based peak detection. At a sampling rate of 100 Hz, this configuration yielded a spectral resolution of approximately 0.098 Hz, allowing for fine frequency discrimination when comparing peak locations. The dominant frequency of oscillation was mathematically extracted by identifying the bin corresponding to the maximum absolute magnitude within the single-sided power spectrum density (PSD) of each signal.

III. Results

A. Mass-Induced Frequency Reduction

The bench results showed that the embedded sensing module was able to detect the frequency reduction associated with the addition of mass. According to the mass-spring dynamic model, the addition of mass decreases the system's natural oscillatory frequency. As detailed in the frequency-domain analysis (Table I), the reference PASCO sensor recorded a clear frequency drop from 2.97 Hz with the empty bracelet (trial 1a) to 1.66 Hz when the 137g mass was added (trial 1b). The proposed embedded system detected this frequency shift, measuring 3.00 Hz and 1.67 Hz, respectively. Similar attenuation patterns were recorded across the other spring configurations tested (e.g., trials 2a and 2b). The results are consistent with the expected mass-dependent reduction in the natural frequency of the oscillator and demonstrate the capability of the low-cost sensor to detect this change. **Fig. 4** shows the alignment between the reference signals of the PASCO sensor and the signals measured by the embedded accelerometer over time, visually confirming their synchronization.

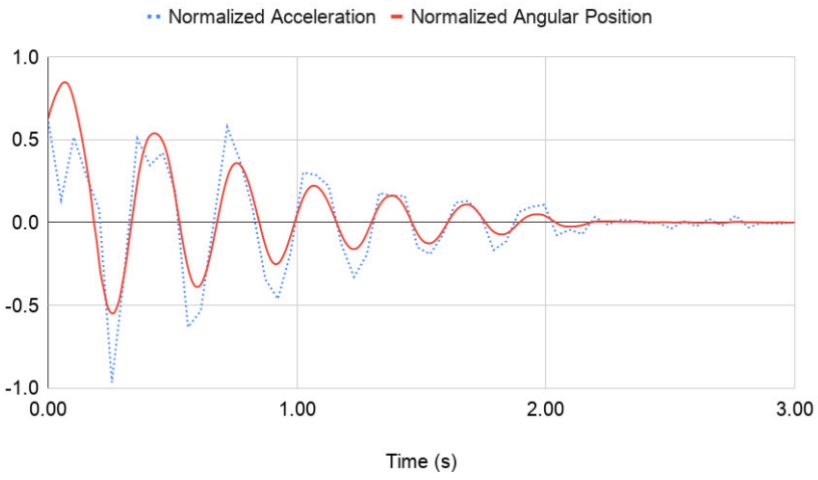


Fig. 4. Comparison between the normalized signals collected by the PASCO reference sensor (solid red line) and the embedded sensing module (dashed blue line) over time. The peak-to-peak alignment visually validates the metrological capability of the low-cost hardware.

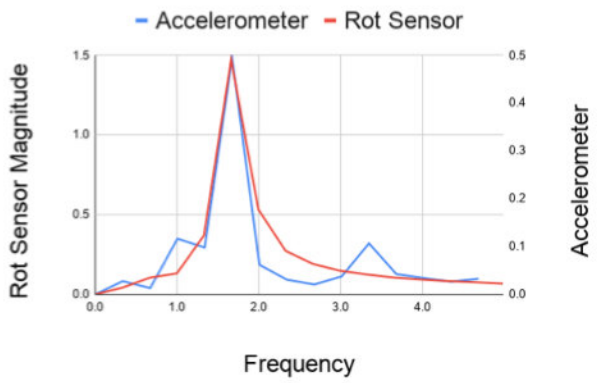


Fig. 5. Frequency spectrum magnitude obtained via Fast Fourier Transform (FFT) for trial 1b (Spring 1 with 137 g mass). The solid red line represents the reference spectrum from the rotary motion sensor (left y-axis), while the solid blue line represents the equivalent spectrum from the embedded accelerometer (right y-axis). The close alignment of the dominant frequency peaks at approximately 1.66 Hz illustrates the close agreement between the dominant frequency estimates obtained from the two sensors.

B. Frequency-Domain Performance Evaluation

To assess the frequency measurement performance of the low-cost accelerometer when compared to the reference laboratory PASCO sensor, the raw data from the 12 bench trials, which represented distinct experimental configurations rather than 12 independent repetitions of the same condition, were analyzed using the Fast Fourier Transform (FFT). **Table I** shows the dominant frequencies extracted from both measurement systems. **Fig. 5** shows the alignment of the dominant frequency between the data extracted from the PASCO sensor and the accelerometer data extracted in trial 1b.

TABLE I. DOMINANT FREQUENCIES EXTRACTED VIA FFT

Trial	*Reference Sensor (Hz)*	*Embedded Sensor (Hz)*	*Difference (%)*
1a	2.970	2.997	0.91%
1b	1.661	1.672	0.66%
2a	1.998	1.964	-1.70%
2b	1.314	1.335	1.60%
3a	1.679	1.697	1.07%
3b	1.000	1.003	0.30%
4a	1.329	1.338	0.68%
4b	0.498	3.344	571.49%
5a	0.738	1.003	35.91%
6a	1.990	1.998	0.40%
6b	1.329	1.332	0.23%
7a	2.993	2.991	-0.07%

For the tested oscillations at or above approximately 1.00 Hz, the embedded sensor showed good agreement with the laboratory reference sensor. However, the experiments did not cover the full 2-10 Hz frequency range important to essential tremor. Within the portion of this range that was reproduced by the bench setup (approximately 2–3 Hz), the frequency differences between the embedded and reference sensors were below 2%.

To provide aggregate error measures for the tested configurations, we calculated aggregate error metrics across these ten trials. The system achieved a Mean Absolute Error (MAE) of 0.0136 Hz, a Root Mean Squared Error (RMSE) of 0.0171 Hz, and an overall Mean Absolute Percentage Error (MAPE) of 0.76%. The maximum absolute deviation recorded was 0.0340 Hz (a 1.70% error in trial 2a). These aggregate metrics characterize the overall performance across the tested configurations and should not be interpreted as estimates of within-condition repeatability.

Conversely, the two trials below the 1.00 Hz threshold—specifically trials 4b and 5a—revealed significant deviations . In trial 5a, the embedded sensor estimated 1.003 Hz compared with 0.738 Hz from the reference sensor, resulting in a 35.91% difference. Although this represents a considerable relative error, the estimated frequency remained in the same general low-frequency region as the reference. In contrast, trial 4b showed a much larger discrepancy, with the embedded sensor indicating 3.344 Hz while the reference sensor measured 0.498 Hz. This result is more consistent with a frequency-peak misidentification than with a simple frequency estimation error, since the dominant peak detected by the FFT did not correspond to the fundamental oscillation observed by the reference sensor. At such low frequencies, the reduced acceleration amplitude may cause the fundamental component to become less prominent relative to noise and other spectral components, increasing the likelihood of selecting an unrelated peak as the dominant frequency.

IV. DISCUSSION

The data from the bench tests support the preliminary technical feasibility of the proposed sensing module under controlled conditions investigated. Using a damped harmonic oscillator model, we evaluated the low-cost accelerometer's ability to detect mass-induced changes in oscillation frequency. According to mass-spring dynamics, increasing the system's mass decreases its natural frequency of oscillation. The embedded sensor was able to register this shift. For instance, the frequency dropped from 3.00 Hz to 1.67 Hz between trials 1a and 1b, and from 1.96 Hz to 1.34 Hz between trials 2a and 2b. Observing these variations indicates that the accessible hardware captures the dynamic changes caused by the added mass.

To further investigate the sensor's frequency measurement performance, we analyzed the signals in the frequency domain using the Fast Fourier Transform (FFT). A direct comparison with the laboratory reference sensor highlighted a critical performance threshold. For oscillations below 1.00 Hz, the embedded system produced substantial different behavior from the reference sensor. In trial 5a, the measured frequency was 1.003 Hz, compared with 0.738 Hz from the reference,

corresponding to a 35.91% difference. In trial 4b, however, the discrepancy was much larger: the reference sensor measured 0.498 Hz, whereas the embedded sensor identified a dominant frequency of 3.344 Hz. This large deviation should therefore not be interpreted simply as a large frequency estimation error. Rather, it suggests that the FFT-based peak-selection procedure may have identified a different spectral component instead of the fundamental oscillation. This distinction is important because the reference frequency was close to 0.5 Hz, where the oscillatory acceleration is expected to be much smaller than at higher frequencies. Under these conditions, the fundamental component may become less distinguishable from noise and other spectral components, increasing the possibility of selecting an unrelated peak as the dominant frequency.

The results from trials 4b and 5a therefore suggest a limitation of the current frequency-identification approach at very low frequencies, rather than a uniform degradation of frequency accuracy. Further experiments with controlled low-frequency excitation and additional signal-quality analysis will be necessary to determine the extent to which sensor noise, spectral leakage, or other mechanical disturbances contribute to this behavior.

These low-frequency results are outside the typical frequency range considered for essential tremor and therefore do not directly limit the intended 2–10 Hz application range. Nevertheless, they are relevant for characterizing the operating limits of the proposed sensing approach and indicate that additional signal-processing strategies may be required when the system is exposed to very slow oscillations. More importantly, the present experiments do not cover the complete 2–10 Hz range relevant to essential tremor, so the observed performance cannot yet be generalized to the entire clinical frequency range.

Essential Tremor is clinically defined by involuntary oscillations usually within the 2.00 to 10.00 Hz frequency band. For the frequencies tested above 1 Hz, the measured errors remained below 2%. However, because the experimental setup was limited to approximately 3 Hz, these results cannot be extrapolated to the entire 2–10 Hz clinical range.

Despite these promising results, a physical limitation of our gravity-driven suspended mass-spring setup must be addressed. While the clinical window for Essential Tremor usually spans from 2.00 Hz to 10.00 Hz, our bench trials were physically constrained to a maximum frequency of approximately 3.00 Hz.

Due to the mass-spring dynamic relation ($f_n = \frac{1}{2\pi}\sqrt{k/m}$), simulating stable oscillations up to 10.00 Hz is restricted by the minimum mass of our physical wearable prototype (approximately 150 g). Forcing higher frequencies would require an extremely stiff spring, resulting in a negligible static deflection that cannot sustain clean, manually excited harmonic oscillations under gravity.

Consequently, while this study successfully verifies the sensor's frequency measurement behavior and low error margins (<2%) starting from 1.00 Hz, validation across the full 2–10 Hz clinical frequency range, particularly above 3 Hz, remains a priority. Future work will overcome this setup limitation by utilizing an active, motorized mechanical shaker table to generate high-frequency controlled sweeps, paving the way for upcoming clinical trials with volunteers.

The validation demonstrates that low-cost wearable hardware can bridge the gap between passive assistive devices and complex, high-cost monitoring systems [10]. The developed system shows potential for frequency-based tremor monitoring, subject to further validation. The observed frequency reduction with increased mass supports the feasibility of investigating mass-dependent mechanical tuning in future studies.

Future studies should also characterize practical aspects of wearable operation, including acquisition latency, communication stability, power consumption, and longer-term continuous operation. These measurements were outside the scope of the present bench validation, which focused primarily on the agreement between the embedded sensor and the laboratory reference in the frequency domain.

V. Conclusion

In this work, we performed a preliminary bench characterization of a low-cost, 3D-printed wearable inertial sensing module integrated into a weighted bracelet. We evaluated the system using a damped harmonic oscillator and compared it with a laboratory reference sensor through frequency-domain analysis based on the Fast Fourier Transform (FFT).

The embedded accelerometer was able to detect the frequency reduction associated with the addition of mass to the oscillating system. For the tested frequencies between approximately 1 and 3 Hz, the measured dominant frequencies showed good agreement with the laboratory reference sensor, with differences below 2% for the tested frequencies above 1 Hz. In contrast, substantial deviations were observed below 1 Hz, which were attributed primarily to the reduced acceleration amplitude and lower signal-to-noise ratio at very low frequencies.

These results support the preliminary technical feasibility of the sensing module for frequency characterization under controlled bench conditions. However, the present study does not constitute clinical validation and does not cover the complete 2–10 Hz frequency range relevant to essential tremor. Further studies should evaluate the system across the full frequency range using controlled excitation, different motion amplitudes and directions, repeated measurements, and ultimately human subjects and clinically relevant reference instrumentation.

References

[1] E. D. Louis and M. McCreary, "How common is essential tremor? Update on the worldwide prevalence of essential tremor," *Tremor and Other Hyperkinetic Movements*, vol. 11, no. 1, 2021, doi: 10.5334/tohm.632.

[2] R. J. Elble, "What is essential tremor? Topical collection on movement disorders," *Curr. Neurol. Neurosci. Rep.*, vol. 13, no. 6, Jun. 2013, doi: 10.1007/s11910-013-0353-4.

[3] P. G. Bain *et al.*, "Assessing tremor severity," *J. Neurol. Neurosurg. Psychiatry*, vol. 56, no. 8, pp. 868–873, 1993, doi: 10.1136/jnnp.56.8.868.

[4] M. Manchanda, "Gyro Glove: Stabilizing The Lives Of Those With Hand Tremor," vol. 20, no. 4, pp. 4781–4794, 2021, doi: 10.17051/ilkonline.2021.04.514.

[5] T. A. Zesiewicz, J. D. Shaw, K. G. Allison, J. S. Staffetti, M. S. Okun, and K. L. Sullivan, "Update on treatment of essential tremor," *Curr. Treat. Options Neurol.*, vol. 15, no. 4, pp. 410–423, Aug. 2013, doi: 10.1007/s11940-013-0239-4.

[6] B. Vescio *et al.*, "Development of a New Wearable Device for the Characterization of Hand Tremor," *Bioengineering*, vol. 10, no. 9, Sep. 2023, doi: 10.3390/bioengineering10091025.

[7] A. Caroppo, A. Manni, G. Rescio, A. M. Carluccio, P. A. Siciliano, and A. Leone, "Movement Disorders and Smart Wrist Devices: A Comprehensive Study," *Sensors*, vol. 25, no. 1, Jan. 2025, doi: 10.3390/s25010266.

[8] Y. Peng *et al.*, "Intelligent devices for assessing essential tremor: a comprehensive review," *J. Neurol.*, vol. 271, no. 8, pp. 4733–4750, Aug. 2024, doi: 10.1007/s00415-024-12354-9.

[9] K. Adabi and W. G. Ondo, "Shaking Up Essential Tremor: Peripheral Devices and Mechanical Strategies to Reduce Tremor," *Tremor Other Hyperkinet. Mov. (N. Y).*, vol. 14, p. 55, 2024, doi: 10.5334/tohm.930.

[10] J. S. Lora-Millan, G. Delgado-Oleas, J. Benito-León, and E. Rocon, "A Review on Wearable Technologies for Tremor Suppression," *Front. Neurol.*, vol. 12, Aug. 2021, doi: 10.3389/fneur.2021.700600.

Projeto e Arquitetura de um Sistema Assistivo de Baixo Custo para Terapia Sonora do Zumbido

1º Mariana Fernandes Dourado Pinto
Departamento de Engenharia Biomédica
UFRN
Natal, Brasil
ORCID: 0009-0009-1307-4811

2º Iago Nobre Silva
Instituto Metrópole Digital
UFRN
Natal, Brasil
ORCID: 0009-0009-1237-4759

3º Samara Dália Tavares Silva
Departamento de Engenharia Biomédica
UFRN
Natal, Brasil
ORCID: 0009-0009-6616-3770

4º Gizele Francisco Ferreira do Nascimento
Instituto Santos Dumont
ISD
Macaíba, Brasil
ORCID: 0000-0003-0992-9118

5º Mário André da Cunha Espada
Instituto Politécnico de Setúbal
ESE
Setúbal, Portugal
ORCID: 0000-0002-4524-4784

6º Paulo Moreira Silva Dantas
Departamento de Educação Física
UFRN
Natal, Brasil
ORCID: 0000-0002-9217-7107

7º Edgard Morya
Instituto Santos Dumont
ISD
Macaíba, Brasil
ORCID: 0000-0003-0954-5317

8º José Carlos Gomes da Silva
Departamento de Educação Física
UFRN
Natal, Brasil
ORCID: 0000-0003-3192-7223

***Resumo*—O zumbido (tinnitus) é a percepção de um som na ausência da fonte sonora e pode comprometer severamente a qualidade de vida de muitos indivíduos, desencadeando distúrbios de sono e ansiedade. Barreiras econômicas no mercado brasileiro limitam o acesso a dispositivos eletrônicos com geradores de ruído integrados, enquanto o sistema público de saúde frequentemente disponibiliza aparelhos focados apenas na amplificação sonora para perda auditiva. Este trabalho tem o objetivo de desenvolver a especificação do projeto, a modelagem arquitetônica e os requisitos de segurança de um sistema assistivo de baixo custo baseado em IoT que pode ser utilizado na Terapia de Habituação do Zumbido (Tinnitus Retraining Therapy - TRT). O sistema proposto possui um aplicativo móvel e um hardware auricular minimalista baseado no microcontrolador ESP32, utilizando acoplamento acústico de canal aberto (open-fit) para evitar o efeito de oclusão. Portanto, sendo um estudo focado na concepção arquitetônica e especificações de projeto, este manuscrito estabelece os requisitos técnicos e detalha um plano de validação eletroacústica, de segurança do paciente e de conformidade regulatória para orientar o desenvolvimento físico e ensaios clínicos futuros.**

Palavras-chave: Zumbido; Tinnitus; Tecnologia Frugal; Internet das Coisas; Terapia Sonora; Dispositivo Auricular.

I. Introdução

O zumbido, também denominado tinnitus, é um sintoma de grande prevalência na população mundial caracterizado pela percepção de um som, na ausência de uma fonte sonora externa correspondente [1]. Embora possa apresentar diferentes características acústicas, intensidade e duração, em alguns indivíduos o zumbido assume caráter persistente gerando um impacto em dimensões funcionais, físicas e emocionais [2]. Apresenta fisiopatologia multifatorial, podendo envolver desde alterações na atividade neural ao longo das vias auditivas, como estar relacionada a modificações no input sensorial periférico, além de causas metabólicas e vasculares [3], [4], [5].

Existem diversos fatores associados ao desenvolvimento ou agravamento do zumbido como, por exemplo, a perda auditiva; a exposição a altos níveis de pressão sonora; uso de medicamentos; acidentes e entre outros [3], [4]. Além dessa questão auditiva, o sintoma pode impactar negativamente a qualidade de vida, gerando alterações no sono, dificuldade de

concentração e complicações emocionais, como ansiedade e depressão em casos mais graves [6], [7].

Entre as estratégias terapêuticas utilizadas para o manejo do zumbido, destaca-se a Terapia de Habituação do Zumbido (Tinnitus Retraining Therapy – TRT). A TRT envolve aconselhamento terapêutico associado ao enriquecimento sonoro, buscando favorecer a habituação à percepção do zumbido e reduzir seu impacto emocional e atencional. O uso de estímulos sonoros contínuos ou estruturados constitui, portanto, um dos recursos empregados no processo de manejo do sintoma, considerando os mecanismos de adaptação e neuroplasticidade do sistema auditivo [8], [9], [7].

Apesar da disponibilidade de diferentes tecnologias para terapia sonora, o acesso a dispositivos com recursos específicos para o manejo do zumbido pode ser limitado por fatores econômicos e tecnológicos. Equipamentos de Amplificação Auditivos de Sonora Individual (AASI) com recursos de geração ou controle de estímulos terapêuticos para o zumbido podem apresentar custo elevado, enquanto aqueles disponibilizados em determinados contextos de atenção à saúde podem ter como finalidade principal a amplificação da fala e não necessariamente contemplar estratégias específicas de enriquecimento sonoro para o zumbido [10], [8], [11].

Diante disso, é observada a importância de alcançar soluções assistivas que conciliam a funcionalidade, usabilidade, acessibilidade econômica e que alcancem resultados significativos e benéficos aos usuários. Por isso, questões como ergonomia, conforto, dimensões do dispositivo, adaptação ao conduto auditivo e controle do efeito de oclusão são importantes para garantir bons resultados e adesão ao uso. Evitando, assim, a percepção de sons internos e o comprometimento do conforto acústico do usuário [6].

Nesse processo, a atuação fonoaudiológica torna-se importante para caracterizar o zumbido a partir da avaliação auditiva, no qual é possível definir o tipo, intensidade e frequência do zumbido que vão permitir a seleção e adaptação do dispositivo de acordo com as necessidades individuais de cada usuário. Dessa forma, o desenvolvimento de uma solução tecnológica para terapia sonora requer a integração de requisitos clínicos, acústicos, ergonômicos e de usabilidade. A definição desses requisitos constitui uma etapa importante para a construção de um sistema que possa posteriormente ser prototipado e submetido a avaliações de desempenho e usabilidade [6].

Em relação ao contexto nacional para esse sintoma, existem milhões de brasileiros que apresentam algum grau de zumbido, com diferentes causas. [10], [8], [6]. Para essa situação, aparelhos de amplificação são oferecidos de forma gratuita pela rede pública mediante fila de espera. Porém, estes aparelhos normalmente são direcionados prioritariamente ao público com perda auditiva e, em apenas alguns casos, esses dispositivos possuem função de redução de zumbido. Essa questão faz com que um grande número de pessoas que possuem zumbido, principalmente aquelas que não sofrem de perda auditiva, permaneçam sem o tratamento necessário para seu sintoma. Além disso, o alto custo financeiro de aparelhos no mercado impede o acesso de grande parte da população a essa tecnologia [8], [11].

Para contornar as questões econômicas e tecnológicas, este trabalho tem como objetivo apresentar a especificação técnica de projeto, a engenharia de requisitos e a modelagem arquitetônica de uma plataforma assistiva de baixo custo baseada em Internet das Coisas (IoT) direcionada à Terapia de Habituação do Zumbido (Tinnitus Retraining Therapy – TRT) [8], [11].

O projeto busca desenvolver um ecossistema integrado por um hardware auricular vestível (centrado no microcontrolador ESP32-PICO-MINI com conectividade BLE), um aplicativo móvel para o paciente (responsável pelo offloading computacional de áudio e registro diário) e uma plataforma web de gerenciamento, destinada ao acompanhamento por fonoaudiólogos e engenheiros [12]. Além disso, será utilizada adoção de acoplamento acústico de canal aberto (open-fit) em invólucro de silicone biocompatível para eliminar o efeito de oclusão e garantir o conforto durante o uso prolongado [3], [6]. Por fim, vale ressaltar que esse projeto se trata de uma estruturação e planejamento iniciais para o desenvolvimento de software e hardware, tendo este trabalho como base para tratar especificações técnicas, os requisitos de biossegurança auditiva e o enquadramento regulatório, estruturando o projeto a fim de orientar a futura prototipagem física e ensaios clínicos do sistema.

II. Materiais e métodos

DELINEAMENTO DO ESTUDO:

Trata-se de um estudo de especificação arquitetônica, modelagem de sistema e engenharia de requisitos aplicado ao desenvolvimento de uma tecnologia assistiva em saúde [8] [13]. Previamente ao desenvolvimento do dispositivo físico, foram definidas as características dos estímulos a serem empregados, com ênfase no uso de estímulos de banda larga, bem como estabelecidas as diretrizes de calibração e os fluxos de software necessários para orientar a futura montagem do protótipo e a realização dos ensaios de bancada.

PROTOCOLO TERAPÊUTICO:

Previamente à utilização do dispositivo, é necessária uma avaliação fonoaudiológica individualizada, incluindo audiometria tonal limiar e de altas frequências, logoaudiometria, acufenometria e emissões otoacústicas. Esses procedimentos permitem caracterizar a função auditiva e coclear, bem como as características psicoacústicas do zumbido, subsidiando a definição dos parâmetros de estimulação do dispositivo [5], [7], [6].

Esse processo possibilita o acompanhamento contínuo do paciente e a caracterização das principais características psicoacústicas do zumbido, incluindo seu tipo, frequência (pitch) e intensidade (loudness), permitindo o ajuste dos parâmetros terapêuticos de acordo com as características individuais do sintoma [6].

O dispositivo proposto tem como foco principal a emissão de ruído para auxílio no manejo desse sintoma específico. Para o projeto, será adotado o uso do ruído branco de banda larga e, em casos específicos, a banda estreita personalizada [7]

[9]. Esta escolha se dá devido à viabilidade computacional no microcontrolador e por utilizar um padrão validado na literatura para a Terapia de Habituação do Zumbido (TRT) [8], [13], [6].

O nível de pressão sonora do ruído deve ser ajustado no ponto de mistura (mixing point), no qual o ruído terapêutico e o zumbido podem ser percebidos simultaneamente, possibilitando a estimulação necessária para a habituação neural [6], [9].

Além disso, para garantir a personalização do estímulo sonoro, os testes e uso serão realizados apenas após uma avaliação fonoaudiológica detalhada [6], [7]. Além da audiometria convencional, serão realizadas analises do Limiar de Desconforto Sonoro e a acufenometria para identificar a frequência exata e o nível de intensidade do zumbido em cada paciente [5], [6].

PROJETO ACÚSTICO DE ACOPLAMENTO ABERTO (OPEN-FIT):

Na região externa do dispositivo, o hardware deverá apresentar geometria compatível com a anatomia auricular, de modo a minimizar pontos de atrito e pressão durante o uso. Para essa finalidade, está sendo avaliada a possibilidade de fabricação da estrutura por impressão 3D em silicone, visando proporcionar maior maleabilidade, além de propriedades hipoalergênicas e biocompatibilidade [6].

Adicionalmente, será utilizado acoplamento acústico do tipo open-fit (oliva de canal aberto), que favorece a ventilação do meato acústico externo e reduz o efeito de oclusão, que pode produzir alterações na percepção sonora durante a utilização do dispositivo [6], [11], [3]. O sinal analógico será gerado pelo conversor digital-analógico (DAC) integrado ao microcontrolador e, posteriormente, encaminhado ao transdutor eletroacústico, responsável pela conversão do sinal elétrico em estímulo sonoro [11] .

REQUISITOS DE SEGURANÇA:

Para garantir a segurança auditiva e a integridade do usuário durante o uso diário, o projeto estabelece requisitos específicos de controle da exposição sonora. Será implementada uma limitação do nível de pressão sonora (SPL), com mecanismos de proteção em software e hardware, de modo a impedir que o nível de saída ultrapasse 80 dB SPL no conduto auditivo externo. Essa medida visa reduzir o risco de exposição sonora excessiva e, consequentemente, de danos à função auditiva decorrentes de níveis elevados de estimulação sonora [4], [13].

O controle de volume disponível no aplicativo do paciente contará com um mecanismo de segurança do tipo failsafe, mantendo o nível de saída restrito ao limite máximo previamente configurado pelo fonoaudiólogo no software de gerenciamento [11], [7].

Na etapa de alimentação, o circuito contará com um sistema de gerenciamento da carga associado à bateria recarregável, com mecanismos de proteção contra sobretensão, curto-circuito e aquecimento excessivo. Além disso, o módulo de radiofrequência associado à conectividade Bluetooth deverá incorporar medidas de blindagem e isolamento adequadas, a fim de minimizar a interferência eletromagnética e a indução de ruídos elétricos no circuito analógico de áudio [4].

Ou seja, sob a perspectiva da biossegurança e conformidade regulatória, a arquitetura de hardware e firmware incorpora mecanismos estritos de proteção auditiva para prevenir episódios de trauma acústico ou perda auditiva induzida por ruído [4]. A saída sonora no conduto auditivo é travada em um limite máximo de 80 dB SPL, impedindo que os ruídos causem lesões às células ciliadas da cóclea [4], [13].

CARACTERIZAÇÃO ELETROACÚSTICA:

Para a futura etapa de desenvolvimento e validação em bancada, foi estabelecido um plano de caracterização eletroacústica da cadeia de sinal. O protocolo prevê a análise do espectro de frequência dos sinais gerados pelos canais DAC de 8 bits do ESP32 e a determinação da Distorção Harmônica Total mais Ruído (THD+N), tanto na saída do circuito quanto no sinal encaminhado ao transdutor eletroacústico [3], [11].

A faixa de frequência de estimulação será previamente estabelecida entre 250 a 8000Hz, contemplando a faixa de interesse para a caracterização e o tratamento do zumbido. Os parâmetros de estimulação serão definidos individualmente pelo fonoaudiólogo no software de gerenciamento, de acordo com as características audiológicas e psicoacústicas de cada paciente [11], [5], [4].

ARQUITETURA DO SISTEMA (SOFTWARE DE GERENCIAMENTO E APLICATIVO):

A arquitetura do projeto propõe que as tarefas de processamento mais pesadas sejam transferidas para o smartphone do paciente (offloading computacional), permitindo a emissão de ruídos sem sobrecarregar ou gerar elevado consumo energético no dispositivo auricular.

Um esquema do funcionamento é apresentado nas figuras 1 e 2 a seguir.

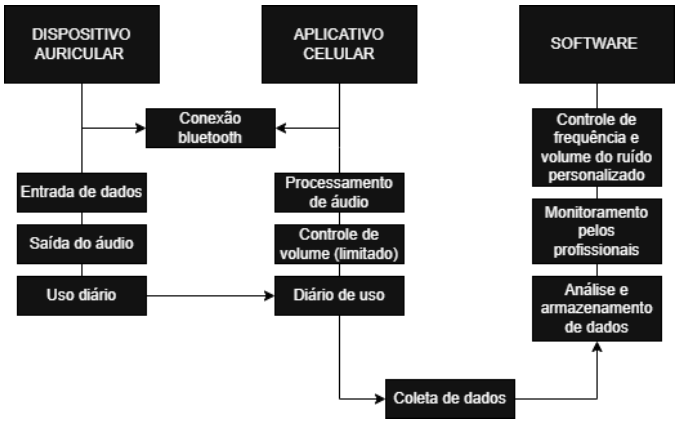


Fig. 1. Autoria própria - Diagrama de blocos da arquitetura do ecossistema assistivo, ilustrando a integração digital e fluxo de dados entre o dispositivo auricular, aplicativo e software.

Para a implementação do ecossistema, o software de gerenciamento será destinado aos profissionais envolvidos, incluindo engenheiros e fonoaudiólogos, para o acompanhamento dos testes e das sessões terapêuticas [8], [13], [7].

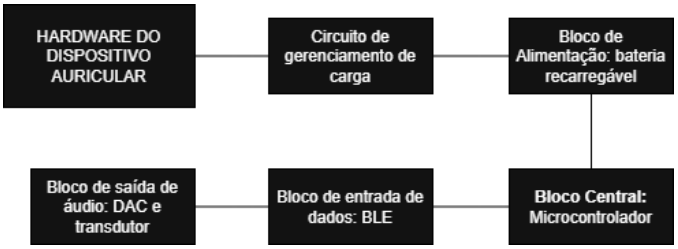


Fig. 2. Autoria Própria - Diagrama para representação do fluxo de funcionamento e arquitetura de hardware projetada para o dispositivo auricular

O módulo permitirá organizar, visualizar e ajustar as informações individuais relacionadas ao zumbido de cada paciente, além de configurar os arquivos de áudio e os parâmetros de estimulação sonora de acordo com as características clínicas e psicoacústicas identificadas.

Paralelamente, a arquitetura prevê o emprego de offloading computacional, transferindo para o smartphone do paciente as tarefas de maior demanda relacionadas ao processamento de áudio [13], [12], [11].

A aplicação permitirá ativar e desativar os estímulos, realizar o ajuste de volume dentro dos limites previamente configurados pelo profissional e registrar diariamente o uso do dispositivo e a percepção de incômodo relacionada ao zumbido. Esses dados serão armazenados em nuvem, possibilitando o acompanhamento longitudinal da utilização e das respostas do paciente ao longo das sessões terapêuticas [12], [6].

Ao transferir parte do processamento de sinais para o smartphone, o dispositivo auricular deverá apresentar menor demanda computacional e potencial redução do consumo energético, favorecendo sua autonomia durante o uso prolongado [11], [12].

O hardware auricular será composto por conectividade Bluetooth Low Energy (BLE), responsável pela comunicação com o smartphone, e terá como unidade central o microcontrolador ESP32, na variante miniaturizada ESP32-PICO-MINI. O sistema contará ainda com dois canais de conversão digital-analógica (DAC) para geração dos sinais, um transdutor eletroacústico de alta impedância e uma bateria recarregável dimensionada de acordo com os requisitos de autonomia estabelecidos para o uso terapêutico.

ENQUADRAMENTO REGULATÓRIO:

O projeto foi estruturado considerando o alinhamento com as normas de dispositivos médicos assistivos (RDC/ANVISA e ISO 13485) [8], [13].

Para o software e comunicação sem fio, será necessário seguir os requisitos da Lei Geral de Proteção de Dados (LGPD) ao implementar criptografia no protocolo BLE, autenticação de acesso na plataforma web e anonimização dos registros de saúde armazenados em nuvem.

III. RESULTADOS ESPERADOS E DISCUSSÕES

A análise da literatura possibilitou um embasamento teórico robusto para a definição dos parâmetros do sistema, demonstrando que a estimulação auditiva contínua com ruídos calibrados pode atuar com eficácia na modulação cortical e na habituação neural para redução do zumbido [8], [6].

Por se tratar de uma especificação de arquitetura e engenharia de requisitos, a discussão trata sobre a viabilidade e maturidade do projeto para o desenvolvimento físico.

Planejar a transferência do processamento de áudio para o aplicativo celular busca reduzir a carga computacional no microcontrolador ESP32, permitindo utilizar células de bateria menores e leves no hardware auricular, possibilitando dimensões reduzidas e o conforto durante o uso do dispositivo.

Além disso, a integração entre o acompanhamento profissional prévio e contínuo, a oliva de canal aberto (open-fit) e o gerenciamento prático pelo aplicativo celular visa reduzir as principais causas de abandono da terapia sonora relatadas na literatura médica: o desconforto induzido pela oclusão mecânica do conduto auditivo e a falta de acompanhamento clínico ao longo do tratamento [3], [6].

Com a arquitetura fundamentada, os próximos passos do projeto contemplam o desenvolvimento da Placa de Circuito Impresso (PCI) em software EDA e a fabricação dos protótipos para a realização dos ensaios de caracterização eletroacústica em bancada.

IV. CONCLUSÃO

A partir das análises e especificações realizadas, concluise que a modelagem arquitetônica apresentada fornece uma estrutura fundamentada para o futuro desenvolvimento de um dispositivo assistivo de baixo custo voltado para a Terapia de Habituação do Zumbido (TRT). O projeto visa alcançar um ecossistema que concilia um dispositivo auricular open-fit, um aplicativo móvel para o paciente e um software de gerenciamento para profissionais. Espera-se, então, por meio dessa tecnologia assistiva planejada, que conta com uma vasta multidisciplinaridade, proporcionar uma melhor qualidade de vida a um público não amplamente alcançado atualmente.

REFERÊNCIAS

[1] B. I. Han, H. W. Lee, T. Y. Kim, J. S. Lim, and K. S. Shin, "Tinnitus: characteristics, causes, mechanisms, and treatments," *Journal of Clinical Neurology*, vol. 5, no. 1, pp. 11–19, 2009.

[2] S. Demoen, E. Cardon, L. Jacquemin, A. Timmermans, V. Van Rompaey, A. Gilles, and S. Michiels, "Health-related quality of life in subjective, chronic tinnitus patients: A scoping review," *Journal of the Association for Research in Otolaryngology*, vol. 25, no. 2, pp. 103–129, 2024.

[3] M. Elisabetsky, "Tratamento atual de zumbido," *Acústica e Vibrações*, vol. 57, pp. 81–84, 2025.

[4] R. C. Fleser, V. Necula, L. P. Ujvary, A. Osman, A. Orasan, and A. A. Maniu, "Hearing loss in young adults: Risk factors, mechanisms and prevention models," *Biomedicines*, vol. 13, no. 12, p. 3116, 2025.

[5] S. Pavaci, F. Tortorella, A. B. Fioretti, A. M. Angelone, L. D. R. Businco, M. Lauriello, and A. Eibenstein, "Analysis of the audiological characteristics and comorbidity in patients with chronic tinnitus," *Audiology Research*, vol. 9, no. 2, p. 231, 2019.

[6] M. R. D. d. Rosa, A. A. F. d. Almeida, F. Pimenta, C. G. Silva, M. A. R. Lima, and M. d. F. F. M. Diniz, "Zumbido e ansiedade: uma revisão da literatura," *Revista CEFAC*, vol. 14, no. 4, pp. 742–754, 2012.

[7] G. L. e. Souza, A. L. P. Nascimento, B. V. D. Silva, F. G. d. Sá, F. M. A. Pereira, G. F. Buzollo, H. D. D. d. Carvalho, L. P. S. Lopes, S. F. M. M. Dias, and C. P. e. Souza, "Current treatments for tinnitus," *Revista Eletrônica Acervo Médico*, vol. 23, no. 8, p. e13559, 2023.

[8] M. d. Fontoura, J. F. Reis, E. S. A. d. Costa, and B. R. d. Oliveira, "A eficácia da terapia de remissão do zumbido baseada em evidências: Uma revisão sistemática e metanálise," *Archives of Health*, vol. 6, no. 4, pp. 1–6, 2025.

[9] R. S. Noleto, R. J. F. Lins, A. S. C. Araújo, A. C. Leite, E. C. Nascimento, I. M. Soares, T. B. d. Costa, J. S. Prazeres, L. B. M. Almeida, and L. D. Lins Neto, "Terapias inovadoras para zumbido: uma revisão," in *Anais do I MED 360 - Congresso Regional de Medicina*, 2024.

[10] M. R. Castro *et al.*, "Prevalência e impacto do zumbido em adultos: estudo populacional brasileiro," *Revista Brasileira de Otorrinolaringologia*, vol. 87, no. 1, pp. 15–22, 2021.

[11] P. J. Sanders, R. M. Nielsen, J. J. Jensen, and G. D. Searchfield, "Hearing aids with tinnitus sound support reduce tinnitus severity for new and experienced hearing aid users," *Frontiers in Audiology and Otology*, vol. 1, p. 1238164, 2023.

[12] H. Y. Lee, S.-H. Shin, and S. W. Byun, "Impact of short-term tinnitus treatment on cognitive function and neural synchronization," *Frontiers in Neurology*, vol. 16, p. 1478033, 2025.

[13] T. Kleinjung, N. Peter, M. Schecklmann, and B. Langguth, "The current state of tinnitus diagnosis and treatment: a multidisciplinary expert perspective," *JARO - Journal of the Association for Research in Otolaryngology*, vol. 25, pp. 413–425, 2024.

Contact with Green Spaces and Mental Health in Adults Under Medical Treatment: A Systematic Review Protocol

Fernanda Santos de Carvalho
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
ORCID: 0009-0001-7572-5730
dotybio@gmail.com

Michelle Luciana de Sousa
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
ORCID: 0000-0002-9325-1305

Aldira Guimarães Duarte Dominguez
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
ORCID: 0000-0002-6481-6055

Glécia Virgolino da Silva Luz
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
ORCID: 0000-0002-5385-6174

***Abstract*—Mental health disorders and psychological distress are highly prevalent among individuals undergoing medical treatment and may negatively affect treatment adherence, recovery, and quality of life. Exposure to green spaces and natural environments has been associated with improvements in psychological well-being and reductions in stress; however, evidence involving adults receiving medical care remains fragmented across clinical conditions and healthcare settings.**

This article presents the protocol for a systematic review aiming to evaluate the effects of direct contact with green spaces on mental health outcomes among adults under medical treatment. The protocol follows the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) recommendations and was prospectively registered in PROSPERO (CRD420261427698).

Initial searches in six databases identified 2,941 records, of which 1,788 remained after duplicate removal. Primary outcomes are perceived stress and psychological well-being. Methodological refinements introduced after protocol registration and screening initiation are being documented as protocol amendments and will be transparently reported in the final review.

The findings of this review may contribute to understanding green spaces as supportive healthcare environments and may support the development of nature-based strategies for mental health promotion and patient-centered healthcare design.

***Index Terms*—green spaces, mental health, nature exposure, stress, psychological well-being, healthcare environments, systematic review.**

I. Introduction

Mental disorders are among the leading causes of disability worldwide and constitute a major public health challenge. According to the World Health Organization, nearly one billion people live with mental disorders, and psychological distress contributes substantially to disability, reduced quality of life, and increased healthcare utilization [1]. Individuals undergoing medical treatment frequently experience elevated levels of stress, anxiety, emotional distress, and reduced psychological well-being, which may negatively affect treatment adherence, recovery, and clinical outcomes.

The healthcare environment itself may influence psychological responses during treatment. Hospitalization, chronic disease management, rehabilitation programs, and long-term medical follow-up may generate emotional burden and stress among patients. Consequently, complementary approaches capable of promoting psychological well-being and reducing psychological distress have received increasing attention.

Natural environments have emerged as potentially beneficial settings for health promotion. A study conducted in the 1980s investigating the relationship between healthcare environments and exposure to nature showed that postoperative patients with views of trees had shorter hospital stays and required fewer analgesics than patients whose rooms faced brick walls [2]. Since then, a growing body of evidence has suggested that contact with natural environments may contribute to psychological restoration and stress reduction.

Two major theoretical frameworks help explain these effects. Stress Reduction Theory proposes that natural environments promote psychophysiological recovery from stress through positive emotional and physiological responses [3]. Attention Restoration Theory suggests that natural environments facilitate recovery from mental fatigue and restore directed attention capacity [4]. These theoretical perspectives have substantially influenced research on restorative environments and therapeutic landscapes.

Exposure to green spaces, including parks, gardens, and forests, has been associated with multiple health benefits.

This work received no specific external funding.

Forest bathing and nature therapy have shown favorable physiological and psychological effects, including reductions in stress-related biomarkers and improvements in mood states [5], [6]. In addition, a broad systematic review and meta-analysis reported consistent associations between green-space exposure and improved health outcomes, including mental health indicators [7].

From an ecosystem-services perspective, natural environments may support mental health through different pathways, including stress reduction, social interaction, physical activity, attention restoration, and environmental quality [8]. These mechanisms may be particularly relevant to individuals undergoing medical treatment, who often experience physical and emotional vulnerability.

Despite the growing body of evidence, studies involving adults under medical treatment remain dispersed across different clinical conditions, healthcare settings, and types of nature exposure. Existing reviews often focus on healthy populations, community settings, specific age groups, or broad environmental exposures. Consequently, the role of direct contact with green spaces among adults receiving medical care remains insufficiently understood.

Furthermore, the incorporation of restorative environments, therapeutic gardens, and biophilic principles into healthcare facilities has attracted increasing interest in healthcare design and biomedical engineering. Understanding the effects of nature exposure on mental health outcomes may support the development of patient-centered care environments and contribute to evidence-based environmental interventions.

Therefore, this systematic review aims to synthesize the available evidence on the effects of direct contact with green spaces on mental health outcomes among adults under medical treatment.

II. Methods

A. Protocol and Registration

This protocol was developed in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) statement [9]. The final systematic review will be conducted and reported in accordance with the PRISMA 2020 statement [10].

The review was registered in the International Prospective Register of Systematic Reviews (PROSPERO) under registration number CRD420261427698 [11]. Following registration and the initiation of screening, methodological refinements identified during implementation will be documented as justified protocol amendments.

B. Research Question

The review question was operationalized according to the PECOS framework (Population, Exposure, Comparator, Outcomes, and Study Design):

- Population: adults under medical treatment or follow-up in healthcare services.
- Exposure: direct contact with green spaces.
- Comparator: usual care, pre-intervention condition, or a concurrent comparator, when applicable.
- Outcomes: perceived stress and psychological well-being (primary), anxiety and depression (secondary).
- Study designs: RCTs with a comparator, controlled quasi-experimental studies, and uncontrolled pre-post studies, according to the methodological amendment.

The research question is:

What are the effects of direct contact with green spaces on mental health outcomes among adults under medical treatment?

C. Eligibility Criteria

Primary quantitative studies involving adults aged 18 to 64 years who are undergoing medical treatment or follow-up in healthcare services will be eligible for inclusion. Eligible healthcare settings include hospitals, outpatient clinics, rehabilitation centers, primary healthcare services, and other healthcare environments. This age range was defined to delimit the target adult population and reduce heterogeneity related to the clinical, functional, and healthcare characteristics of aging.

The exposure of interest consists of direct contact with green spaces, including:

- parks;
- gardens;
- therapeutic gardens;
- healing gardens;
- hospital gardens;
- forests;
- forest therapy;
- forest bathing.

Studies involving only virtual, simulated, or indirect exposure to nature, including photographs, videos, digital environments, or virtual reality, will be excluded.

Primary outcomes include:

- perceived stress;
- psychological well-being.

Primary outcomes will be assessed using validated instruments, including the Perceived Stress Scale (PSS) and the World Health Organization Five Well-Being Index (WHO-5).

Secondary outcomes include:

- anxiety;
- depression.

The following publication types will be excluded:

- reviews;
- editorials;
- conference abstracts without full text;
- case reports;
- case series;
- protocols.

Review articles, systematic reviews, and meta-analyses will be excluded from eligibility and used only for manual reference checking to identify potentially relevant primary studies not retrieved through the electronic searches.

Studies involving healthy populations, children, and adolescents will be excluded. Studies including participants aged 65 years or older will be eligible only when data for participants aged 18 to 64 years can be identified separately. Interventions centered on horticulture/therapeutic gardening or physical exercise in green spaces will also be excluded. Following PROSPERO registration and the initiation of screening, the study-design criteria were refined as a methodological amendment. The final synthesis will include RCTs with a comparator, controlled quasi-experimental studies, and uncontrolled pre-post studies; the latter will be analyzed separately and without causal attribution.

TABLE I
PECOS FRAMEWORK OF THE REVIEW

Component	Description
Population	Adults aged 18–64 years under treatment or follow-up in healthcare services.
Exposure	Direct contact with green spaces.
Comparator	Usual care, pre-intervention condition, or a concurrent comparator, when applicable.
Outcomes	Perceived stress, psychological well-being, anxiety, and depression.
Study designs	RCTs with a comparator, controlled quasi-experimental studies, and uncontrolled pre-post studies; refinement documented as an amendment.

D. Information Sources

Electronic searches were conducted in six databases:

- PubMed;
- Scopus;
- Embase;
- Web of Science Core Collection;
- Virtual Health Library (BVS);
- LILACS.

The searches were conducted between May and June 2026, with no restriction on publication year. Language restrictions were applied to English, Portuguese, and Spanish for operational feasibility, reflecting the languages in which the review team can independently and reliably assess full-text articles. The potential for language bias will be considered in the interpretation. No systematic search of the gray literature was planned because the review focuses on published and indexed primary studies with sufficient methodological information for critical appraisal. However, this decision may increase the risk of publication bias and will be acknowledged as a limitation. Relevant reviews will be used for manual reference checking of primary studies.

E. Search Strategy

The initial search combined controlled vocabulary and free-text terms related to three conceptual domains:

1) green spaces and nature exposure;
2) mental health outcomes;
3) adults receiving medical care.

The PubMed search strategy was:

(("Parks, Recreational"[Mesh] OR "Gardens"[Mesh] OR "Forests"[Mesh] OR "green space*"[tiab] OR "urban green space*"[tiab] OR "nature exposure"[tiab] OR "nature contact"[tiab] OR "natural environment*"[tiab] OR "therapeutic garden*"[tiab] OR "healing garden*"[tiab] OR "hospital garden*"[tiab] OR "forest therapy"[tiab] OR "forest bathing"[tiab] OR "shinrin-yoku"[tiab] OR "biophilic design"[tiab]) AND ("Mental Health"[Mesh] OR "Stress, Psychological"[Mesh] OR "Anxiety"[Mesh] OR "Depression"[Mesh] OR "mental health"[tiab] OR "mental well-being"[tiab] OR "psychological well-being"[tiab] OR "emotional well-being"[tiab] OR stress[tiab] OR "perceived stress"[tiab] OR "stress reduction"[tiab] OR anxiety[tiab] OR depression[tiab]) AND (patient*[tiab] OR "adult patient*"[tiab] OR "in medical treatment"[tiab] OR "medical care"[tiab] OR healthcare[tiab] OR hospitalized[tiab] OR hospitalised[tiab] OR inpatient*[tiab] OR outpatient*[tiab] OR ambulatory[tiab] OR "chronic disease"[tiab] OR "chronic illness"[tiab] OR cancer[tiab] OR oncology[tiab] OR rehabilitation[tiab] OR dialysis[tiab] OR "cardiac rehabilitation"[tiab] OR "cardiovascular disease"[tiab] OR diabetes[tiab]))

Equivalent strategies were adapted for Embase, Scopus, Web of Science, BVS, and LILACS according to the indexing characteristics of each database.

During review implementation, the sensitivity of this strategy began to be reassessed, particularly regarding the clinical condition/population block and the retrieval of previously known potentially relevant studies. Any search refinements will be documented as protocol amendments, with dates and justification, and will be fully reported in the final review, while clinical and age criteria will remain part of the eligibility assessment to preserve methodological traceability.

F. Study Selection

All records retrieved from the databases were imported into Rayyan for duplicate identification and study selection [12].

Study selection will be performed independently by two reviewers in two stages:

1) title and abstract screening;
2) full-text assessment.

Disagreements will initially be resolved through discussion and consensus between the reviewers. If consensus cannot be reached, disagreements will be resolved by the supervising researchers.

The two primary reviewers are:

- Fernanda Santos de Carvalho;
- Michelle Luciana de Sousa.

The reviewers responsible for resolving disagreements are:

- Aldira Guimarães Duarte Dominguez;
- Glécia Virgolino da Silva Luz.

G. Data Extraction

Data extraction will be performed independently by two reviewers using a standardized form.

The following variables will be collected:

- author and year;
- country;
- study design;
- participant characteristics;
- clinical condition;
- healthcare setting;
- type of green-space exposure;
- duration and frequency of exposure;
- outcome measures;
- assessment instruments;
- main findings.

Missing essential data will be recorded as not reported and, when feasible, study authors may be contacted; unreported results will not be imputed. Potentially overlapping publications will be compared according to authors, study location, recruitment period, and sample characteristics and, when they refer to the same population, will be treated as a single study unit.

TABLE II
VARIABLES EXTRACTED FROM ELIGIBLE STUDIES

Variable	Description
Author and year	Study identification
Country	Study location
Population	Participant characteristics
Clinical condition	Disease or treatment context
Exposure	Type of green space
Outcome	Mental health measures
Setting	Healthcare environment
Main findings	Main results

H. Risk of Bias Assessment

Methodological quality and risk of bias will be assessed independently by two reviewers using the appropriate Joanna Briggs Institute critical appraisal tools according to study design [13].

Risk-of-bias judgments will be considered during interpretation. Studies at higher risk will not be automatically excluded solely on this basis; when a sufficient number of comparable studies is available, sensitivity analyses may be performed.

I. Certainty of Evidence

The certainty of evidence will be assessed by outcome using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, including secondary outcomes when sufficient evidence is available for synthesis [14].

The certainty of evidence will be classified as:

- high;
- moderate;
- low;
- very low.

J. Data Synthesis

The primary synthesis will be narrative and structured by study design: (1) randomized controlled trials with a comparator; (2) controlled quasi-experimental studies; and (3) uncontrolled pre-post studies. Within each group, clinical condition, healthcare setting, type and duration of exposure, and assessment instrument will be considered. Pre-post studies will be described using non-causal language.

Meta-analysis will be considered only when at least two studies are sufficiently comparable in terms of population, exposure, comparator, outcome, and assessment time point. The decision to pool studies will prioritize clinical and methodological compatibility in addition to statistical heterogeneity. When the number of studies permits, subgroup analyses by clinical condition, healthcare setting, and type of exposure may be explored; otherwise, a structured narrative synthesis will be maintained.

III. PRELIMINARY RESULTS

The electronic searches identified a total of 2,941 records across the six databases. After duplicate identification and removal using Rayyan, 1,153 duplicate records were excluded, resulting in 1,788 unique records for title and abstract screening.

Table III presents the number of records retrieved from each database. These figures correspond to the initial search and may be updated if sensitivity checking leads to refinement of the search strategy before the final synthesis.

TABLE III
RECORDS RETRIEVED FROM EACH DATABASE

Database	Records
PubMed	214
Scopus	659
Embase	581
Web of Science	812
BVS	646
LILACS	29
Total	2,941

The screening process is ongoing. The final study-selection process will be presented using the PRISMA 2020 flow diagram.

IV. IMPLICATIONS FOR BIOMEDICAL ENGINEERING

The potential contribution of this review to Biomedical Engineering lies in translating evidence on restorative environments into recommendations for healthcare-environment planning.

If supported by the final results, the findings may inform design guidelines for therapeutic and hospital gardens, criteria for incorporating natural elements into healthcare environments, clinical guidance on contact with green spaces as a therapeutic practice, and the development of patient-centered environmental technologies. The interface with soft technologies in health,

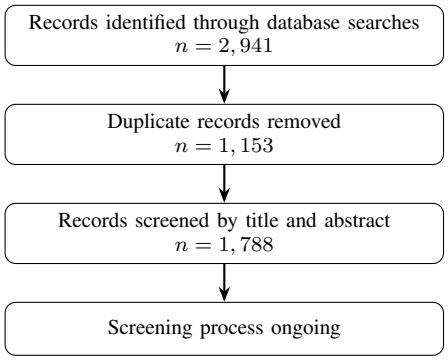


Fig. 1. Preliminary PRISMA flow diagram.

as conceptualized by Merhy and Franco [15], will be treated as a complementary dimension of care.

V. Methodological Challenges and Research Gaps

Heterogeneity is expected across clinical populations, environments and exposure protocols, psychological assessment instruments, and study designs with different levels of control. Another challenge is balancing search sensitivity and specificity: overly restrictive terms may omit relevant studies, whereas generic environmental terms may retrieve a large volume of irrelevant records.

A lack of standardization is also expected in reporting the duration, frequency, and intensity of contact with green spaces. These gaps will be mapped without anticipating conclusions regarding effectiveness, and clinical diversity will be treated as a source of heterogeneity rather than as evidence of a uniform effect.

VI. Discussion

This protocol was designed to synthesize a potentially heterogeneous body of literature. Organizing the evidence by study design will distinguish studies with concurrent comparators from pre-post studies. Interpretation will be proportional to the strength of the study design and risk of bias.

The breadth of clinical conditions increases the scope and relevance of the review but limits generalizability. Therefore, the synthesis will explicitly describe differences among populations, settings, and exposure modalities. Anticipated limitations include language restrictions and the absence of a systematic gray-literature search, with potential language and publication bias.

VII. Conclusion

This protocol describes a systematic review registered in PROSPERO and guided by PRISMA-P. Procedures have been defined for independent study selection, risk-of-bias assessment, certainty-of-evidence evaluation, management of heterogeneity, and assessment of search sensitivity. Conclusions regarding effectiveness and implications for healthcare-environment design will be formulated only after completion of the review.

Acknowledgment

The authors thank the Graduate Program in Biomedical Engineering at the University of Brasília for institutional support.

References

[1] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. Geneva, Switzerland: WHO, 2022.

[2] R. S. Ulrich, "View through a window may influence recovery from surgery," *Science*, vol. 224, no. 4647, pp. 420–421, 1984.

[3] R. S. Ulrich, R. F. Simons, B. D. Losito, E. Fiorito, M. A. Miles, and M. Zelson, "Stress recovery during exposure to natural and urban environments," *Journal of Environmental Psychology*, vol. 11, no. 3, pp. 201–230, 1991.

[4] S. Kaplan, "The restorative benefits of nature: Toward an integrative framework," *Journal of Environmental Psychology*, vol. 15, no. 3, pp. 169–182, 1995.

[5] Q. Li, "Effect of forest bathing trips on human immune function," *Environmental Health and Preventive Medicine*, vol. 15, no. 1, pp. 9–17, 2010.

[6] M. M. Hansen, R. Jones, and K. Tocchini, "Shinrin-yoku (forest bathing) and nature therapy: A state-of-the-art review," *International Journal of Environmental Research and Public Health*, vol. 14, no. 8, p. 851, 2017.

[7] C. Twohig-Bennett and A. Jones, "The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes," *Environmental Research*, vol. 166, pp. 628–637, 2018.

[8] G. N. Bratman, C. B. Anderson, M. G. Berman et al., "Nature and mental health: An ecosystem service perspective," *Science Advances*, vol. 5, no. 7, eaax0903, 2019.

[9] D. Moher, L. Shamseer, M. Clarke et al., "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement," *Systematic Reviews*, vol. 4, no. 1, p. 1, 2015.

[10] M. J. Page, J. E. McKenzie, P. M. Bossuyt et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.

[11] F. S. de Carvalho and M. L. de Sousa, "Contact with Green Spaces and Mental Health in Adults Under Medical Treatment: A Systematic Review," PROSPERO, CRD420261427698, 2026.

[12] A. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews," *Systematic Reviews*, vol. 5, no. 1, p. 210, 2016.

[13] J. Aromataris and Z. Munn, Eds., *JBI Manual for Evidence Synthesis*. Adelaide, Australia: Joanna Briggs Institute, 2020.

[14] G. H. Guyatt, A. D. Oxman, E. A. Akl et al., "GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables," *Journal of Clinical Epidemiology*, vol. 64, no. 4, pp. 383–394, 2011.

[15] E. E. Merhy and T. B. Franco, "Por uma composição técnica do trabalho centrada no campo relacional e nas tecnologias leves," *Saúde em Debate*, vol. 27, no. 65, Sep./Dec. 2003.

REALIMENTAÇÃO SOMATOSENSORIAL DE FORÇA PARA O CONTROLE DE PRÓTESES MIOELÉTRICAS DE MÃO

Juan de Oliveira Bezerra
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
juan@peb.ufrj.br

Clarice Gomes Lessa
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
clarice.g.lessa@gmail.com

Carolina do Carmo Alonso
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
carolina.alonso@medicina.ufrj.br

Mauricio Cagy
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
mcagy@peb.ufrj.br

Carlos Julio Tierra-Criollo
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
carjulio@peb.ufrj.br

Resumo— **As próteses mioelétricas transradiais apresentam baixa adesão ao uso cotidiano, em grande parte, devido à ausência de realimentação somatossensorial. Este estudo propõe e avalia preliminarmente um protocolo de reabilitação baseado na realimentação de força por estimulação elétrica senoidal (ES). O protocolo envolveu a padronização da força de preensão em 30 participantes hígidos, a avaliação da instrumentação experimental e a codificação de quatro níveis de força por estimulação elétrica senoidal. Os níveis utilizados corresponderam a 5%, 20%, 35% e 50% da Contração Voluntária Máxima (CVM), por apresentarem conforto e reprodutibilidade sem causar fadiga. A ES de 3 kHz foi gerada por um estimulador elétrico, permitindo o controle combinado de amplitude e duração do *burst* para codificação dos níveis de força. Foram obtidas taxas médias de 87% para percepção de força e 88% para execução de força. A percepção apresentou melhora progressiva ao longo da sessão, atingindo 100% de acerto no último bloco do Teste de Retenção de Aprendizagem (TRA), o que sugere um processo de aprendizagem sensório-motora intrassessão. O Estímulo de 35% da CVM apresentou o menor desempenho relativo. Como perspectivas futuras, prevê-se a integração de sinais eletromiográficos (EMG), eletroencefalográficos (EEG) e mecanomiográficos (MMG), bem como a aplicação do protocolo em indivíduos com amputação transradial.**

Palavras-Chave— **Prótese mioelétrica, realimentação somatossensorial, controle de força de preensão, reabilitação.**

I. INTRODUÇÃO

As funções motoras desempenham papel fundamental na sobrevivência e na realização das atividades de vida cotidiana. Sua perda, decorrente de amputações traumáticas ou de doenças neuromusculares, compromete significativamente a independência funcional e a qualidade de vida dos indivíduos. Nesse contexto, tecnologias assistivas, especialmente as próteses mioelétricas, têm sido desenvolvidas para restaurar parcialmente a funcionalidade motora por meio da conversão de sinais eletrofisiológicos em comandos [1]. Estudos recentes indicam que o controle mioelétrico exige maior esforço cognitivo-motor do que alternativas mecânicas, o que reforça a necessidade de estratégias que facilitem a internalização desse tipo de controle pelo sistema nervoso [2]. Apesar dos avanços tecnológicos, a adesão ao uso cotidiano dessas próteses permanece limitada. Um dos principais fatores associados a essa baixa utilização é a escassez ou ausência de realimentação somatossensorial, o que impede o usuário de perceber adequadamente a força aplicada durante a manipulação de objetos [3]. Como consequência, a precisão das tarefas de preensão é reduzida, dificultando a execução de atividades de vida diária. Além disso, a falta de retorno sensorial compromete a incorporação da prótese à imagem corporal e é frequentemente apontada como uma das principais causas de abandono do dispositivo [4, 5]. Entre as estratégias investigadas para suprir essa lacuna, destaca-se a estimulação eletrotátil, que atua sobre mecanorreceptores via corrente controlada

na pele, ativando seletivamente fibras Aβ sem recrutar fibras nociceptivas [6]. Estudos demonstraram que é possível codificar até 13 sensações eletrotáteis distintas com acurácia superior a 85%, por meio da modulação combinada de parâmetros de pulso [7]. A ES senoidal de 3 kHz, em particular, tem demonstrado potencial para induzir sensações artificiais graduadas e modular oscilações corticais na banda alfa [8]. Essas características tornam a técnica promissora para a codificação de variáveis sensoriais, como a força de preensão, por meio da manipulação de parâmetros de amplitude e temporização dos *bursts* de estimulação. Com base nesse contexto, o presente trabalho propõe e avalia preliminarmente um protocolo de realimentação somatossensorial de força utilizando ES senoidal em participantes hígidos, como etapa preparatória para futuras aplicações em usuários de próteses mioelétricas transradiais. A hipótese central é que a ES senoidal seja capaz de codificar de forma confiável quatro níveis distintos de força de preensão por meio de combinações específicas de amplitude e parâmetros temporais dos *bursts* de estimulação. Dessa forma, espera-se que participantes hígidos consigam discriminar, reproduzir e reter esses níveis de força após uma única sessão de treinamento, mesmo na ausência de *biofeedback* visual. A confirmação dessa hipótese poderá fornecer subsídios para o desenvolvimento de sistemas de controle mioelétrico com realimentação sensorial aplicados a próteses mioelétricas transradiais.

II. Materiais e métodos

A. Participantes

Os critérios de inclusão compreenderam indivíduos com idade igual ou superior a 18 anos, sem comprometimentos cognitivos ou neuromusculares que pudessem interferir na execução das tarefas propostas. Foram excluídos participantes com doenças neuromusculares, alterações cognitivas, uso de medicações capazes de modificar a força muscular ou a sensibilidade somatossensorial, bem como condições associadas à degeneração sensorial, como diabetes mellitus e esclerose lateral amiotrófica (ELA). A elegibilidade foi verificada por meio de anamnese estruturada. Todos os participantes eram voluntários hígidos. O projeto foi aprovado sob o código CAEE 53366621.0.0000.5257 pelo comitê de ética em pesquisa do Hospital Clementino Fraga Filho (HUCFF - UFRJ) no dia 16/01/2022.

B. Equipamentos

O sistema de medição de força foi composto pelo Dinamômetro de Força Manual EMGSystem, selecionado em comparação preliminar com um dinamômetro de mão e pinça. O modelo adotado apresentou maior superfície de contato para preensão palmar, suporte integrado para o antebraço e maior estabilidade postural durante as coletas, sendo conectado a um computador para aquisição contínua dos dados de força.

A estimulação elétrica senoidal de 3 kHz foi gerada por um equipamento de neuroestimulação *Dual Channel Electrical Stimulato*r [9], utilizado como fonte de corrente controlada. O modo corrente é preferível ao modo tensão em aplicações de estimulação cutânea por ser independente da variação de impedância da pele, garantindo maior previsibilidade na carga elétrica efetivamente entregue aos tecidos [6]. O sistema permitiu o ajuste independente de amplitude e dos parâmetros temporais dos *bursts* (T-ON e T-OFF), essenciais para a codificação dos níveis de força.

A interface entre o estimulador e a pele foi realizada por eletrodos de ouro, cuja escolha se justifica pela baixa impedância de contato.

C. Força de Preensão

Esta seção descreve o protocolo de medição da força de preensão e a identificação dos níveis de força utilizados nas etapas subsequentes. Participaram 30 voluntários hígidos (19 homens e 11 mulheres). A dominância manual foi determinada pelo Inventário de Lateralidade de Edinburgh (EHI).

Cada participante realizou cinco contrações voluntárias máximas (CVM) de preensão utilizando o Dinamômetro de Força Manual EMGSystem, com intervalo de um minuto entre as tentativas e o cotovelo mantido a 90° de flexão. A CVM individual foi calculada pela média das cinco tentativas. A normalização pela CVM é procedimento consagrado na literatura para comparação entre indivíduos e controle da intensidade relativa do esforço [10].

A partir desses valores, foram avaliadas diferentes intensidades relativas de força. Os níveis correspondentes a 5%, 20%, 35% e 50% da CVM foram selecionados por apresentarem melhor equilíbrio entre conforto e capacidade de reprodução pelos participantes durante o tempo exigido pelo protocolo. Esses níveis permitiram a execução sustentada da tarefa sem indução significativa de fadiga, além de apresentarem diferenças perceptíveis entre si.

D. Configuração Experimental

Todos os experimentos foram conduzidos no Laboratório de Engenharia Neural (NTIEB/COPPE/CT – UFRJ), em ambiente com temperatura controlada (25 ± 1°C). O posicionamento dos participantes seguiu as recomendações da *American Society of Hand Therapists* (ASHT): quadris e joelhos flexionados a 90°, pés apoiados no solo, ombro em posição neutra, cotovelo flexionado a 90°, antebraço em posição neutra e punho entre 0° e 30° de extensão.

O membro utilizado para a instrumentação foi definido por meio de randomização computacional implementada em Python, utilizando geração de números equiprováveis: o valor 1 correspondia ao membro esquerdo e o valor 2 ao membro direito, garantindo igual probabilidade de seleção para cada participante. Esse membro segurava o dinamômetro conectado ao computador para aquisição dos dados de força, com a tela mantida fora do campo visual do participante. A randomização foi adotada para tornar o protocolo independente da lateralidade, considerando a futura aplicação em indivíduos com amputação transradial, nos quais a perda do membro pode ocorrer tanto no lado dominante quanto no não dominante.

E. Determinação do Limiar de Sensibilidade

Antes da associação entre estimulação elétrica e força de preensão, foi determinado individualmente o limiar de sensibilidade (LS) para a estimulação senoidal de 3 kHz. O procedimento foi composto por duas etapas [8]: inicialmente um protocolo de rampa, no qual a amplitude era progressivamente aumentada até a percepção do estímulo, definindo o Limiar Rampa (LR). Em seguida, foi aplicado um protocolo de rastreamento iniciado em LR/2, com incrementos de LR/4 quando não havia percepção e decrementos de LR/8 quando o estímulo era percebido, até atingir uma resolução final de 20 μA.

A frequência de 3 kHz foi selecionada por apresentar maior seletividade para fibras Aβ de grande calibre, com menor probabilidade de recrutar fibras nociceptivas ou produzir sensações dolorosas [8]. O LS foi utilizado como referência para definir as amplitudes empregadas na codificação dos níveis de força, permitindo calibrar a estimulação às características sensoriais de cada participante. As amplitudes utilizadas permaneceram dentro da faixa operacional do estimulador, cujo limite máximo de corrente é de 10 mA. Nenhum participante relatou dor ou desconforto significativo durante o protocolo.

F. Codificação dos Níveis de Força pelo Estímulo

Cada nível de força previamente definido foi associado a uma combinação específica de amplitude e duração do *burst* de estimulação (T-ON), mantendo-se constante o intervalo sem estímulo (T-OFF = 0,5 s): 5% da CVM (1,2 × LS; T-ON = 0,5 s), 20% da CVM (1,5 × LS; T-ON = 0,8 s), 35% da CVM (2,0 × LS; T-ON = 1,0 s) e 50% da CVM (2,5 × LS; T-ON = 1,2 s). A combinação simultânea de amplitude e duração foi adotada para ampliar a distinção perceptiva entre níveis adjacentes de força.

G. Protocolo de Treinamento

Após a calibração, os participantes realizaram 10 blocos de treinamento. Não foram realizadas tentativas adicionais de familiarização além dos blocos de treinamento descritos no protocolo. Em cada bloco, os quatro níveis de força foram apresentados em ordem crescente (5%, 20%, 35% e 50% da CVM), mantendo-se essa ordem em todos os blocos, sem contrabalanceamento. Para cada nível, o participante executava uma preensão sustentada de 20 s com auxílio de *biofeedback* visual exibido na tela do computador. O estímulo elétrico correspondente era aplicado durante os 10 s finais da tarefa.

H. Teste de Retenção de Aprendizagem (TRA)

Após um período de cinco minutos sem estimulação, previamente estabelecido como intervalo de retenção de curto prazo, foi realizado o Teste de Retenção de Aprendizagem (TRA), composto por cinco blocos. Diferentemente da etapa de treinamento, o TRA foi conduzido sem qualquer *biofeedback* visual. Em cada tentativa, um dos quatro estímulos elétricos previamente associados aos respectivos níveis de força durante o treinamento era apresentado durante os primeiros 10 s. A ordem de apresentação dos quatro estímulos em cada bloco foi definida por randomização computacional em Python, gerando permutações sem repetição dos quatro estímulos, de modo que cada estímulo fosse apresentado exatamente uma vez por bloco, em ordem imprevisível para o participante. Após a apresentação do estímulo, o participante deveria identificar qual estímulo havia recebido e, a partir da associação estabelecida durante o treinamento, indicar o nível de força correspondente. Nos 10 s subsequentes, sem estimulação, o participante deveria reproduzir a força associada ao estímulo apresentado.

Ao final de cada tentativa, o participante informava verbalmente o estímulo percebido e o respectivo nível de força associado, além de descrever a sensação experimentada. As sensações relatadas pelos participantes foram consistentemente descritas como vibrações de intensidade variável, correspondendo qualitativamente aos diferentes estímulos apresentados. Em seguida, o nível correto era informado pelo pesquisador como forma de *feedback*. Os pesquisadores responsáveis pela aplicação do protocolo não foram cegados quanto à ordem dos estímulos apresentados. O desempenho foi analisado segundo duas métricas principais: *(i)* percepção de força, definida como a identificação correta do estímulo apresentado e do nível de força a ele associado; e *(ii)* execução de força, definida como a reprodução correta da força correspondente ao estímulo apresentado, na ausência de *biofeedback* visual. Os resultados foram expressos como médias entre os participantes.

III. Resultados

Participaram do estudo 30 voluntários hígidos, sendo 19 homens (idade: 33,3 ± 11,9 anos; CVM: 35,9 ± 6,7 kgf) e 11 mulheres (idade: 34,3 ± 14,0 anos; CVM: 22,1 ± 6,7 kgf). A CVM média geral foi de 30,8 ± 9,4 kgf. Os valores foram obtidos de forma consistente entre os participantes, indicando a adequação dos quatro níveis de força selecionados (5%, 20%, 35% e 50% da CVM) para as etapas subsequentes.

Durante o protocolo de treinamento, todos os participantes completaram os 10 blocos sem intercorrências, não havendo desistências ou exclusões após o início do protocolo. O *biofeedback* visual foi progressivamente associado ao estímulo elétrico correspondente a cada nível de força, permitindo a familiarização com os parâmetros de codificação. Nenhum participante relatou desconforto significativo relacionado à estimulação ao longo das sessões.

As Tabelas I e II apresentam os resultados obtidos no TRA. A Tabela I apresenta a acurácia de Percepção de Força, definida como a identificação correta do estímulo apresentado e do nível de força a ele associado. A Tabela II apresenta a acurácia de Execução de Força, definida como a reprodução correta da força correspondente sem auxílio de *biofeedback* visual.

TABELA I - *Acurácia de Percepção de Força por nível de estímulo (linhas) ao longo dos cinco blocos (colunas).*

Estímulo	**Bloco 1**	**Bloco 2**	**Bloco 3**	**Bloco 4**	**Bloco 5**	**Total**
1	100%	100%	100%	66%	100%	**93%**
2	100%	100%	66%	66%	100%	**86%**
3	66%	66%	66%	100%	100%	**80%**
4	33%	100%	100%	100%	100%	**87%**
Médias						**87%**

TABELA II - *Acurácia de Execução de Força por nível de estímulo (linhas) ao longo dos cinco blocos (colunas).*

Estímulo	**Bloco 1**	**Bloco 2**	**Bloco 3**	**Bloco 4**	**Bloco 5**	**Total**
1	100%	100%	100%	100%	100%	**100%**

2	100%	66%	100%	100%	100%	**93%**
3	100%	66%	100%	66%	66%	**80%**
4	66%	66%	66%	100%	100%	**80%**
Médias						**88%**

Para a Percepção de Força (Tabela I), a média geral foi de 87%. O Estímulo 1 (5% da CVM) apresentou o maior desempenho global, atingindo 93% de acerto, enquanto o Estímulo 3 apresentou a menor acurácia total (80%). Observou-se melhora progressiva ao longo dos blocos de avaliação, com destaque para o Bloco 5, no qual todos os estímulos atingiram 100% de identificação correta.

Para a Execução de Força (Tabela II), a média geral foi de 88%. O Estímulo 1 apresentou desempenho máximo em todos os blocos, atingindo 100% de acerto total. Os Estímulos 3 e 4 apresentaram as menores médias gerais (80%), porém ambos demonstraram melhora ao longo do teste. O Estímulo 4 iniciou com 66% de acerto nos três primeiros blocos e atingiu 100% nos Blocos 4 e 5. De forma semelhante, o desempenho geral da execução manteve-se elevado ao longo do protocolo, atingindo 100% de acerto nos blocos finais para a maioria dos níveis de força.

IV. Discussão

Os resultados obtidos constituem uma avaliação preliminar da viabilidade do protocolo proposto. A padronização da força de preensão em 30 participantes permitiu estabelecer níveis de esforço confortáveis e reproduzíveis, em linha com valores de referência descritos na literatura para população adulta hígida [10], [11]. No TRA, a ES apresentou médias de acerto de 87% para Percepção de Força e 88% para Execução de Força, indicando capacidade preliminar de transmitir informações relacionadas à magnitude da força de preensão.

Na Percepção de Força, observou-se aumento da discriminação dos estímulos ao longo dos blocos, atingindo 100% de acerto no último bloco. Esse comportamento é compatível com uma possível aquisição de associações sensório-motoras de curto prazo [8], [12]. Entretanto, o *feedback* verbal fornecido após cada tentativa do TRA impede atribuir essa melhora exclusivamente às associações estabelecidas durante o treinamento. Os resultados também apresentam desempenho compatível com estudos de estimulação eletrotátil que demonstraram elevada capacidade de discriminação de diferentes padrões de estimulação [7], embora diferenças entre os protocolos devam ser consideradas.

O Estímulo 3 (35% da CVM), apresentou o menor desempenho, com 80% de acerto tanto para Percepção quanto para Execução de Força. Esse resultado pode estar relacionado a uma pista perceptiva menos distinta em relação aos níveis adjacentes ou à maior variabilidade do controle graduado da força em níveis intermediários [13]. Contudo, o delineamento não permite determinar a origem desse menor desempenho. Além disso, como amplitude e duração do *burst* foram modificadas simultaneamente, esses parâmetros podem ter atuado como pistas perceptivas para a identificação dos estímulos, não sendo possível determinar a contribuição individual de cada parâmetro para sua discriminação.

Algumas limitações devem ser consideradas. Embora a ordem dos estímulos no TRA tenha sido definida por randomização computacional, reduzindo a possibilidade de antecipação baseada na sequência, não se pode excluir a influência de expectativas decorrentes das associações estabelecidas durante o treinamento. Os pesquisadores também não foram cegados quanto à ordem dos estímulos, constituindo possível fonte de viés de observação. Durante o treinamento, os níveis de força foram apresentados em ordem fixa e crescente, sem contrabalanceamento, de modo que efeitos de ordem ou prática não podem ser completamente excluídos. A ausência de uma condição controle também limita a distinção entre os efeitos específicos da ES e aqueles decorrentes da prática e do *biofeedback* visual. O intervalo de cinco minutos sem estimulação entre o treinamento e o TRA caracteriza uma avaliação de retenção de curto prazo, não permitindo avaliar a manutenção das associações em períodos mais prolongados.

Os resultados do TRA foram preservados de forma agregada, não estando disponíveis registros individuais das respostas. Assim, análises retrospectivas em nível individual, incluindo matrizes de confusão, modelos de efeitos mistos, estimativas de tamanho de efeito e respectivos intervalos de confiança, não puderam ser realizadas. Os achados devem, portanto, ser interpretados como evidências descritivas e preliminares da viabilidade do protocolo.

Por fim, a avaliação foi realizada exclusivamente em participantes hígidos, não permitindo afirmar que os resultados se traduzam diretamente para usuários de próteses mioelétricas transradiais, nos quais alterações das vias sensoriais, sensações ou dor fantasma e a experiência decorrente do uso prolongado de próteses podem influenciar a percepção e o desempenho. Estudos

posteriores deverão incluir essa população e condições controle, além de investigar separadamente os parâmetros de estimulação por análises de sensibilidade e comparar a ES com modalidades vibrotáteis e mecanotáteis. A integração com registros de EMG, EEG e MMG também poderá contribuir para a investigação dos mecanismos neuromusculares e corticais envolvidos.

V. Conclusões

A combinação de amplitude e tempo de ativação do *burst* foi capaz de produzir estímulos perceptivamente discrimináveis associados a diferentes níveis de força, com médias de acerto de 87% para percepção e 88% para execução, fornecendo evidência preliminar de viabilidade do protocolo proposto. Assim, a principal contribuição deste estudo consiste na calibração e avaliação preliminar de um protocolo de realimentação somatossensorial que poderá subsidiar estudos posteriores em usuários de próteses mioelétricas transradiais. Como perspectivas futuras, o protocolo será integrado a registros de EMG, EEG e MMG, e posteriormente aplicado a indivíduos com amputação transradial, visando investigar os mecanismos neuromusculares e corticais associados ao desempenho e avaliar sua aplicabilidade clínica.

Agradecimentos

Os autores agradecem aos voluntários que participaram do estudo, à equipe do Laboratório de Engenharia Neural (NTIEB/COPPE/CT – UFRJ) pelo suporte, e às agências financiadoras CAPES (PROEX), CNPq, FAPERJ (E-26/204.393/2024) e FINEP (Convênio 01.24.0122.00).

Referências

[1] Farina, D., Aszmann, O. Bionic limbs: clinical reality and academic promises. *Science Translational Medicine*, 2020.

[2] Gaskins, B. et al. Neural correlates of performance and mental workload dynamics during learning of upper-limb body-powered and myoelectric prostheses. *Journal of Neural Engineering*, 2026.

[3] Abiri, R. et al. A comprehensive review of EEG-based brain–computer interface paradigms. *Journal of Neural Engineering*, 2019.

[4] Diaz-Hernandez, O. A worldwide research overview of Artificial Proprioception in prosthetics. *PLOS Digital Health*, 4(4), e0000809, 2025.

[5] Flor, H. Maladaptive plasticity, memory for pain and phantom limb pain. *Nature Reviews Neuroscience*, 2002.

[6] Parsnejad, S. et al. A review of electrotactile stimulation for machine-to-human communication (M2HC). *IEEE Transactions on Biomedical Engineering*, 2024.

[7] Parsnejad, S. et al. Distinguishable Multi-Layer Frequency-Modulated Waveforms for Peripheral Nerve Communication in Human Augmentation Applications. *IEEE Transactions on Haptics*, 2025.

[8] Volpi, L. P. and Tierra-Criollo, C. J. Cortical alpha oscillation during 3 kHz steady-state sinusoidal electric current stimulation. *Journal of Neurology & Stroke*, v. 7, n. 1, 2017.

[9] Molina-Vidal, D. A. et al. Design and Evaluation of a Prototype of Dual Channel Electrical Stimulator. SmartTech-IC 2022, CCIS 1705. Springer, 2023. p. 342–355.

[10] Mathiowetz, V. et al. Grip and pinch strength: normative data for adults. *Archives of Physical Medicine and Rehabilitation*, 1985.

[11] Bohannon, R. W. Hand-grip dynamometry provides a valid indication of upper extremity strength impairment in home care patients. *Journal of Hand Therapy*, 2001.

[12] Wolpaw, J. R., Wolpaw, E. W. Brain–Computer Interfaces: Principles and Practice. 2. ed., 2021.

[13] Enoka, R. M., Duchateau, J. Rate coding and the control of muscle force. *Cold Spring Harbor Perspectives in Medicine*, 2017.

Recognition of Isolated Dynamic Libras Signs using Formational-Parameter-Based Features

Araújo, K. P. B.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Pará, Brazil
karenpbaraujo@gmail.com

Costa, Y. O.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Pará, Brazil
yuriolicosta42@gmail.com

Figueiredo, L. E. T. C.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Pará, Brazil
cassianolynnda@gmail.com

Farias, E. A. S.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Pará, Brazil
eduafsantos@gmail.com

Fonseca, M. C. P.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Pará, Brazil
conceicao@ufpa.br

Miziara, I. M.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Pará, Brazil
immiziara@ufpa.br

***Abstract*—Automatic recognition of Libras can contribute to the development of more accessible technologies for the deaf community. In this work, a system was developed to classify nine isolated signs from videos. The MediaPipe Hands and MediaPipe Pose computer vision modules were used to detect and track landmarks on the hands and upper body. An LSTM (Long Short-Term Memory) network then analyzed the sequences of these points throughout the signs. Three input configurations were compared for the model: 171 coordinates provided directly by MediaPipe, 155 features related to hand configuration, its location relative to the body, and movement, and the combination of both representations, totaling 326 attributes per frame. In the evaluation, these representations achieved accuracies of 86.41%, 91.35%, and 93.82%, respectively, with the concatenated representation achieving the best performance, with a macro F1-score of 93.81%. The concatenated representation was integrated into a webcam-based application capable of automatically identifying the onset and offset of signs. The results show that the engineered features complement the MediaPipe coordinates and can support the development of future assistive technologies and communication interfaces between Libras and Portuguese.**

***Index Terms*—Brazilian Sign Language (Libras), sign language recognition, Long Short-Term Memory (LSTM), formational parameters.**

I. Introduction

Brazilian Sign Language (Libras) is fundamental to the autonomy, social participation, and access to information of the Brazilian deaf community. However, linguistic barriers still hinder communication in essential services, especially when interpreters or professionals proficient in Libras are unavailable [1]. For many users of this language, written Portuguese is acquired as a second language [2], which can make activities such as writing, reading, formulating questions, and understanding everyday, health-related, and public-service information more difficult [3]. Recently, with the advancement of assistive technologies, digital resources supporting communication and accessibility have been developed, although these resources are still not robust enough to replace interpreters or other forms of linguistic accessibility [4].

Different studies have investigated the development of applications that, with the aid of artificial intelligence and the use of cameras, enable the identification and recognition of sign languages. Wadhawan and Kumar observed that most of the systems developed focused on isolated, static, one-handed signs [5]. In the context of Libras, Silva et al. compared models capable of analyzing image sequences for the real-time recognition of three signs [6]. Samaan et al. used hand, body, and face landmarks to represent ten dynamic signs and track their execution throughout the video [7]. Varshini and Rukmani employed a similar approach in a webcam-based application to recognize 20 American Sign Language signs [8].

Although these studies demonstrate the feasibility of sign recognition using conventional cameras, they differ in terms of datasets, sign languages, and representations used. Silva et al. worked with three isolated Libras signs based on image sequences [6], while Samaan et al. used hand, body, and face landmarks to represent ten dynamic signs [7]. Varshini and Rukmani, in turn, employed landmarks in a webcam-based application for the recognition of 20 American Sign Language signs [8]. In addition to the differences in the number of signs and the language considered, these studies use different ways of representing information, such as image sequences and landmark coordinates. Consequently, their results do not allow a direct determination of which representation is most suitable for Libras recognition. On the other hand, engineered features can explicitly represent aspects of the sign, such as hand distances and angles, spatial relationships with the body, and movement trajectories, although they may omit information present in the original coordinates. This motivates the investigation of representations based on intrinsic characteristics of Sign Language, such as hand shape, signing location, and movement [9]. Representing these elements may help distinguish signs that share some of these parameters.

Given this, the objective of this work is to develop and

evaluate a system based on a Long Short-Term Memory (LSTM) neural network for the classification of nine isolated Libras signs. To investigate the influence of sign representation on model performance, three input forms are compared: 171 coordinates provided directly by MediaPipe, 155 features engineered from hand configuration, its position relative to the body, and movement, and a concatenated representation combining both, totaling 326 attributes per frame. In addition to the experimental evaluation, the concatenated representation is integrated into a webcam-based application, in which a state machine automatically identifies the onset and offset of each execution. This work thus seeks to evaluate the contribution of the different representations to sign classification and the feasibility of the system in real time.

II. Theoretical Background

A. MediaPipe Hands and MediaPipe Pose

MediaPipe is a computer vision framework designed for real-time image and video processing [10]. In this work, MediaPipe Hands and MediaPipe Pose were combined to extract reference points known as *landmarks*. These points correspond to anatomical positions identified in the image, such as finger joints, wrists, shoulders, and elbows, represented by spatial coordinates. MediaPipe Hands estimates 21 points per hand, with x, y, and z coordinates, in addition to tracking them across frames and identifying handedness [11]. MediaPipe Pose, based on BlazePose, estimates 33 body points, including references for the head, shoulders, arms, and torso [12]. The combination of both solutions makes it possible to represent hand configuration and its position relative to the body.

The use of landmarks reduces input dimensionality by replacing video pixels with structured coordinates, emphasizing the geometric and temporal relationships of the sign while reducing the influence of background and clothing. It also enables real-time processing with a conventional RGB camera, although landmark extraction may be affected by occlusions, hand overlap, rapid movements, inadequate lighting, and partial framing.

B. Long Short-Term Memory Networks

Dynamic signs depend on the temporal evolution of movements, making frame-by-frame analysis insufficient. RNNs process sequential information but may struggle with long-term dependencies due to vanishing and exploding gradients [13]. LSTM networks address these limitations through gated memory mechanisms that regulate which information is retained, updated, or discarded [14]. This makes them suitable for modeling the temporal sequences of MediaPipe landmarks used in this work [7].

In this work, the LSTM was chosen because the *landmarks* extracted by MediaPipe form sequences that describe the execution of the signs. While HS and LOC are mainly associated with the spatial relationships between the hands and the body, MOV depends on the changes observed across frames. Thus, MediaPipe provides the spatial representation, while the LSTM models its temporal evolution. The architecture and hyperparameters used are presented in the Methodology Section.

C. Formational Parameters of Libras

Libras has a visuospatial modality, in which information is produced by the hands, the body, and non-manual expressions. Stokoe was a pioneer in proposing that signs can be decomposed into sublexical units formed by Handshape, Location, and Movement [15]. Later, Battison incorporated palm orientation [16], and other studies recognized the participation of facial and body expressions. Thus, the description of Libras frequently considers five parameters: Handshape (HS), Location or Place of Articulation (LOC), Movement (MOV), Orientation (O), and Non-Manual Expressions (NME) [9], [17], [18]. Changes in one or more of these parameters can distinguish signs or modify their meanings.

HS corresponds to the shape and arrangement of the fingers; LOC indicates the region of the body or signing space where the hands are located; and MOV describes trajectory, direction, repetition, and displacement. Orientation represents the direction of the palm and fingers, while NME comprise movements of the face, head, and body with linguistic functions [17], [18]. As an example, Table I presents the decomposition of the signs LEARN and FAMILY according to the variants described in the consulted source.

TABLE I
Parametric description of Libras signs

Param.	**Signs**	
	LEARN	**FAMILY**
HS	Hand in C configuration	Both hands in F configuration
LOC	Forehead	Neutral space at chest height
MOV	Repeated opening and closing	Symmetric, semicircular movement, bringing the hands together
O	Palm oriented sideways	Palms facing each other
NME	Neutral expression	Dependent on communicative context

Source: Adapted from [20].

This work focuses on HS, LOC, and MOV due to their direct relationship with the geometric and temporal information obtained from the hands and body pose. These parameters guided the construction of descriptors based on distances, angles, spatial relationships, and landmark trajectories. The proposed representation does not constitute a direct linguistic annotation of the parameters, but rather a computational approximation of aspects associated with them, detailed in the Methodology Section.

III. Methodology

The methodology was organized into four stages: data acquisition and extraction, sign representation, model training and evaluation, and real-time application. Fig. 1 presents an overview of the developed process.

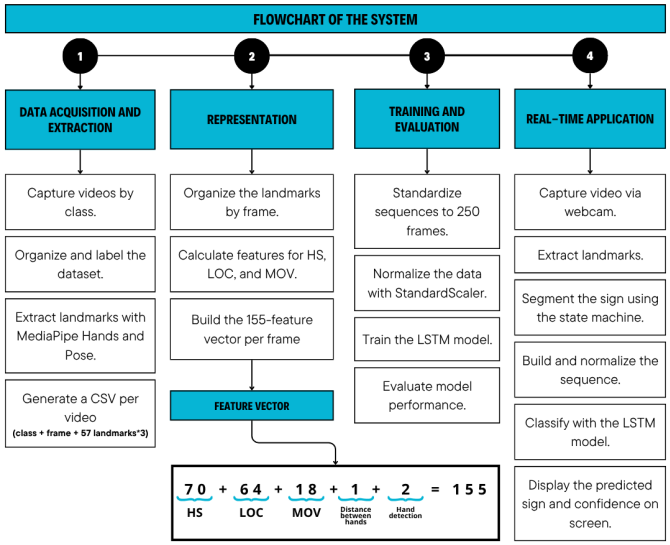


Fig. 1. Flowchart of the system stages.

TABLE II
SAMPLE DISTRIBUTION IN THE DATASET

Sign	Number of samples
Love	45
Learn	46
Study	45
I	56
Family	36
Man	54
Woman	55
Work	54
You	40
Total	**431**

A. Dataset construction and landmark extraction

The dataset was constructed with public YouTube videos and original recordings, the latter accounting for more than half of the samples. Signs with variations in Handshape, Location or Place of Articulation, and Movement were selected. The original recordings captured the upper body and were performed at a resolution of 1920×1080 pixels and a frame rate of 30 frames per second.

To expand the dataset, horizontally mirrored versions of the videos were generated. This augmentation was considered linguistically valid for the selected signs because their identity is preserved under laterality inversion: none of the selected signs specifies the dominant hand or presents a lateral asymmetry that would change the sign when mirrored, which is consistent with the parametric description of Libras signs [9], [17]. The mirroring procedure was applied to the entire dataset, resulting in additional samples while preserving the characteristics of the original executions. The final dataset comprised 431 samples, distributed as shown in Table II.

The videos were manually trimmed to contain only the execution of the sign, keeping all frames within the selected interval. Reading was performed with OpenCV, and each frame was converted to RGB before processing by MediaPipe.

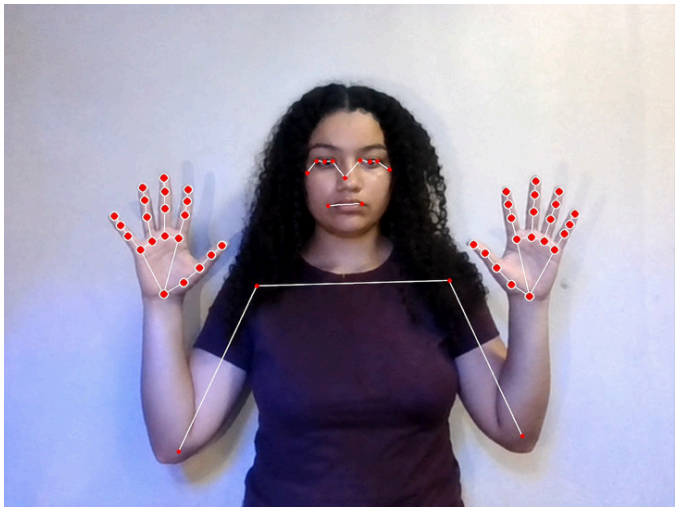


Fig. 2. Hand and pose *landmarks* selected for sign representation. 21 *landmarks* were extracted for each hand and 15 points pose. *Source: Authors' own elaboration.*

Extraction used MediaPipe Hands and MediaPipe Pose with minimum confidence thresholds of 50% for detection and tracking of reference points. 21 *landmarks* were extracted for each hand and 15 pose points, corresponding to indices 0 to 14, which include references for the face, shoulders, and elbows. These points can be better visualized in Fig. 2.

Each frame was represented by 57 *landmarks*, totaling 171 coordinates (x, y, z). When a hand or the pose was not detected, the corresponding values were filled with zeros. For each video, a CSV file was generated with one row per frame and 173 columns: the label, the frame index, and the 171 extracted coordinates.

B. Sign representation

Based on the landmarks, three input representations were evaluated. The first directly used the 171 coordinates extracted per frame. The second transformed these coordinates into 155 geometric and temporal features inspired by the HS, LOC, and MOV parameters. The third concatenated both representations, totaling 326 attributes per frame.

To represent HS, distances and angles between hand points were used, describing the arrangement of the fingers and their relationships with the wrist and thumb. LOC was represented by distances between the hands and pose references, such as the facial region, shoulders, and elbows. For MOV, frame-to-frame displacements and trajectory measures of the hand centroids were calculated, such as length, distance relative to the initial position, tortuosity, and convex hull properties.

The constructed vector contains 70 HS features, 64 LOC features, and 18 MOV features, in addition to the distance between the hands and two detection indicators, totaling 155 values per frame.

These features do not constitute a direct linguistic annotation, but rather a computational approximation of aspects associated with the formational parameters. Hand orientation

was not encoded separately, and non-manual expressions were not modeled.

C. Training and evaluation

Three input representations were compared: 171 raw coordinates, 155 engineered features, and the concatenation of both, with 326 attributes per frame. The dataset was stratified and split into 65% for training, 15% for validation, and 20% for testing, using random seed 42.

Due to the variable duration of the videos, the sequences were limited to 250 frames, considering the video with the longest duration. To standardize the length of the sequences, padding frames were added to samples with shorter duration, a process referred to as *padding*. For the 171- and 155-feature representations, -1 was used as padding, whereas -10 was adopted for the concatenated representation because -1 could coincide with valid values after processing. In both cases, a *Masking* layer identifies and ignores the padding frames during LSTM processing. In both cases, a *Masking* layer, that is, a mask applied to the network input, identifies the padding frames and disregards them during processing by the LSTM.

Each set of attributes was standardized using *StandardScaler*, a method that adjusts the data to have zero mean and unit standard deviation, reducing scale differences between features. This adjustment considered only the valid training frames and was subsequently applied to the validation and test sets. Frames added by *padding* retained the padding values. The labels were encoded in *one-hot* format, in which each class is represented by a binary vector, and class weights were employed to compensate for residual imbalance.

The architecture, kept the same across the three experiments, consisted of a *Masking* layer, responsible for identifying and disregarding the padding frames, followed by an LSTM with 64 units and two *Dropout* layers, which randomly deactivated 20% of the neurons during training, reducing the possibility of overfitting. This was followed by a dense layer with 64 neurons and ReLU activation, and a *softmax* output layer with nine neurons, corresponding to the classes. Training used the Adam optimizer, categorical cross-entropy as the loss function, batches of four samples, and up to 200 epochs. To reduce overfitting, *EarlyStopping* was employed, which stops training when the validation loss shows no improvement for ten epochs and restores the best weights. In addition, a *checkpoint* saved the model with the lowest validation loss.

D. Real-time application

The model based on the 326 features (concatenated representation) was integrated into an application for classifying signs captured via webcam, having been chosen because it presented the best experimental performance among the evaluated representations. The application was developed with OpenCV, a computer vision library used for image capture and processing, for acquiring frames from the webcam; with MediaPipe for detecting and tracking reference points; and with TensorFlow, a platform for developing and running machine learning models, for classifying the signs. The same trained model, class encoder, and normalizer used during training were employed.

The camera was configured to operate at 30 frames per second. The captured frames were horizontally mirrored, converted to the RGB color system, and processed by MediaPipe Hands and MediaPipe Pose, with thresholds of 50% for detection and tracking. To automatically segment the signs, a state machine was implemented based on the displacement of the hand centroids between consecutive frames, calculated as the Euclidean distance between the normalized coordinates provided by MediaPipe. When both hands were visible, the hand with the greater displacement per frame was considered; when only one was detected, the displacement of that hand was used.

For the operation of the state machine, states and events representing the stages of the capture process were defined. In the `WAITING` state, the system monitored displacement without storing frames. When the displacement reached the threshold of 0.025 per frame, the system transitioned to the `CAPTURING` state and began storing the *landmarks*. Capture was ended when the displacement remained less than or equal to 0.02 per frame for 12 consecutive frames (0.4 s), when both hands remained absent for eight consecutive frames (approximately 0.27 s), or when the sequence reached 250 frames (approximately 8.33 s).

Sequences with fewer than ten frames (approximately 0.33 s) generated the `DISCARDED` event and were ignored; the remaining sequences generated the `FINALIZE` event and were forwarded to the classifier. The *buffers* and counters were then reset, and the system returned to the `WAITING` state.

The stored *landmarks* were converted into the same 155 features used during training and concatenated with the 171 *landmark* coordinates. The resulting representation was normalized with *StandardScaler* and padded with vectors of value -10 until reaching 250 frames.

The resulting input was organized as a tensor of dimension $1\times250\times326$, corresponding, respectively, to a sequence of one sample, 250 frames, and 326 attributes per frame. This input was fed to the LSTM. The *softmax* output layer, responsible for producing a score for each of the nine classes, was used to determine the class corresponding to the highest score. The prediction was presented together with its score, expressed in the interval $[0, 1]$. Since no rejection threshold was adopted, every valid sequence was assigned one of the nine classes. The interface also displayed the capture state, the number of stored frames, and the *landmarks* detected in real time.

IV. Results and Discussion

A. Comparison of input representations

To evaluate the influence of sign representation on classifier performance, three input configurations were compared: raw coordinates, engineered features, and the concatenated representation. The three configurations were evaluated on the same test set, comprising 81 samples, while keeping the same LSTM architecture and the remaining training parameters unchanged.

Table III presents the overall accuracy and macro F1-score obtained for each representation, while Fig. 3 presents the respective confusion matrices.

TABLE III
PERFORMANCE OF THE REPRESENTATIONS ON THE TEST SET

Representation	Dim.	Accuracy (%)	Macro F1 (%)
Raw coordinates	171	86,41	85,93
Engineered features	155	91,35	91,61
Concatenated representation	326	93,82	93,81

The representation based on raw coordinates presented the lowest performance among the three configurations, with an accuracy of 86.41% and a macro F1-score of 85.93%. The main confusions occurred among the classes STUDY, MAN, and WOMAN, in addition to errors involving FAMILY and LEARN. With the use of the engineered features, accuracy increased to 91.35% and the macro F1-score to 91.61%, corresponding to gains of 4.94 and 5.68 percentage points, respectively. This result indicates that transforming the coordinates into features related to configuration, location, and movement provided more discriminative information for the classification process.

The concatenated representation presented the best performance among the three configurations, reaching 93.82% accuracy and 93.81% macro F1-score. Relative to the raw coordinates, gains of 7.41 percentage points in accuracy and 7.88 percentage points in macro F1-score were observed. When compared to the engineered features, the concatenated representation showed additional gains of 2.47 percentage points in accuracy and 2.20 percentage points in macro F1-score. This behavior suggests that the raw coordinates and the engineered features provide complementary information to the model: while the former directly preserve the spatial information of the reference points, the latter incorporate descriptors related to the formational parameters of the signs.

For the representation based on raw coordinates, 11 errors were observed in the test set. With the engineered features, this number was reduced to 7, while the concatenated representation presented only 5 errors. In this last configuration, the main confusions occurred between STUDY and LOVE, MAN and LEARN, MAN and STUDY, WOMAN and STUDY, and WORK and WOMAN. The concentration of errors in these classes indicates greater overlap among the feature patterns extracted for these signs.

B. Analysis by class

The individual analysis of the classes makes it possible to identify which signs present greater classification difficulty and which features may be related to the observed errors. Table IV presents the metrics obtained for each of the nine classes using the concatenated representation, which presented the best overall performance, with 93.82% accuracy and 93.81% macro F1-score.

The classes I, FAMILY, and YOU presented maximum performance across all metrics. LOVE, LEARN, and WORK

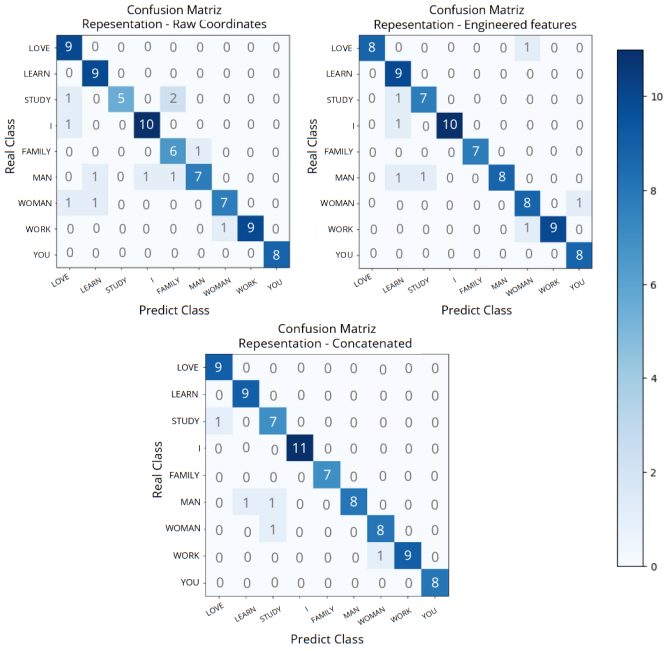


Fig. 3. Confusion matrices on the fixed-split test set for the three representations: raw coordinates, engineered features, and concatenated representation.

TABLE IV
PER-CLASS PERFORMANCE METRICS FOR THE CONCATENATED REPRESENTATION

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
I	100.00	100.00	100.00	100.00
FAMILY	100.00	100.00	100.00	100.00
YOU	100.00	100.00	100.00	100.00
LOVE	98.77	90.00	100.00	94.74
LEARN	98.77	90.00	100.00	94.74
WORK	98.77	100.00	90.00	94.74
WOMAN	97.53	88.89	88.89	88.89
MAN	97.53	100.00	80.00	88.89
STUDY	96.30	77.78	87.50	82.35

also presented high performance, with an F1-score of 94.74%. The greatest difficulties were observed for STUDY, WOMAN, and MAN, with F1-scores of 82.35%, 88.89%, and 88.89%, respectively. These results are consistent with the confusion matrix, in which these classes appear among the ones most involved in the errors.

STUDY presented the lowest precision among the classes, at 77.78%, and a *recall* of 87.50%, indicating the occurrence of both samples from the class itself being assigned to other classes and false positives. MAN and WOMAN presented *recall* of 80.00% and 88.89%, respectively, showing that part of their samples were confused with other signs.

The main confusions may be related to the partial overlap of spatial and temporal features. Figure 4 presents some of the signs involved in these errors. It can be observed that I, LOVE, and MAN may present, at certain moments, a closed hand configuration directed toward the chest region, while STUDY is also performed in the upper body region, although it presents



Fig. 4. Visual representation of signs involved in the main confusions observed by the classifier. The sign names shown in the images correspond to EU [I], HOMEM [MAN], AMOR [LOVE], ESTUDAR [STUDY], and MULHER [WOMAN]. *Source: images b), c), and e) extracted from [19]; images a) and d) generated by Artificial Intelligence.*

differences in configuration and movement trajectory. This similarity may reduce separability between some classes. This is, however, a hypothesis, since the model does not provide explicit measures of the importance of these features.

Overall, the per-class analysis confirms the high performance of the concatenated representation, but shows that performance is not uniform across signs, especially for classes with greater similarity in their spatial and temporal patterns.

C. Generalization across signers

To evaluate the generalization capability for individuals not observed during training, a *cross-signer* experiment was conducted, in which one signer was used exclusively for testing. 16 repetitions of each of the nine signs were planned and, after processing, 141 samples were effectively evaluated. The model achieved 83.69% overall accuracy and 82.83% macro F1-score, compared to 93.82% and 93.81%, respectively, in the fixed split, corresponding to reductions of 10.13 and 10.98 percentage points.

Table V presents the per-class performance. Maximum performance is observed for LEARN, STUDY, FAMILY, and WORK, while MAN presented the lowest F1-score (31.58%), followed by I (60.38%) and LOVE (66.67%). For MAN, the low *recall* rate (18.75%) indicates that a large portion of the true samples were not recognized, while for I, the *recall* of 100.00% was accompanied by a precision of 43.24%, indicating a higher occurrence of false positives.

Overall, the performance reduction relative to the fixed split shows that variation between signers constitutes a challenge for the model's generalization, reinforcing the importance of datasets with greater signer diversity.

D. Feature ablation test

To evaluate the contribution of the different feature groups to classifier performance, an ablation test was conducted based on the engineered representation. Four configurations were evaluated: the complete model, containing 155 features, and three configurations obtained by individually removing the groups related to hand configuration (HS), location (LOC), and movement (MOV). The results are presented in Table VI.

TABLE V
PER-CLASS PERFORMANCE METRICS IN THE *cross-signer* EVALUATION

Class	Accuracy	Precision	Recall	F1-score
LEARN	100.00	100.00	100.00	100.00
STUDY	100.00	100.00	100.00	100.00
FAMILY	100.00	100.00	100.00	100.00
WORK	100.00	100.00	100.00	100.00
WOMAN	99.29	100.00	93.33	96.55
YOU	97.87	87.50	93.33	90.32
LOVE	94.33	100.00	50.00	66.67
I	85.11	43.24	100.00	60.38
MAN	90.78	100.00	18.75	31.58

TABLE VI
RESULTS OF THE FEATURE ABLATION TEST

Configuration	No. of features	Accuracy (%)	Macro F1 (%)
Complete	155	91.36	91.61
Without HS	85	71.60	72.51
Without LOC	90	86.42	86.49
Without MOV	137	91.36	91.86

TABLE VII
COMPARISON WITH RELATED WORK

Work	Model	Result
Silva et al. [6]	LSTM and Bi-LSTM	84.71% with LSTM
Samaan et al. [7]	MediaPipe with GRU, LSTM, and Bi-LSTM	100% with GRU
Alves et al. [22]	OpenPose and ResNet-18	93% on MINDS-Libras and 82% on LIBRAS-UFOP
Rego et al. [21]	MediaPipe and Bi-LSTM	$92\% \pm 3$
Varshini and Rukmani [8]	MediaPipe and LSTM	94.38% on the test set
This work	**LSTM with concatenated representation**	**93.82%**

Removing the hand configuration (HS) features produced the largest performance drop, with decreases of 19.76 percentage points in accuracy and 19.10 percentage points in macro F1 relative to the complete model. Removing the location (LOC) features also resulted in a reduction, although smaller, of 4.94 percentage points in accuracy and 5.12 percentage points in macro F1. On the other hand, removing the movement (MOV) features did not reduce accuracy and produced a positive change of 0.25 percentage points in macro F1.

The results indicate that, in the evaluated set, the HS features presented the greatest contribution to performance, followed by the LOC features, while the MOV features showed a smaller impact when removed in isolation. However, this comparison should be interpreted considering the distinct number of attributes in each group.

E. Comparison with the literature

Table VII presents a comparison between the results obtained in this work and those reported in related studies. The studies were selected because they present approaches based on *landmarks*, recurrent networks, or sign recognition strategies comparable to the one employed in this study.

The performance obtained falls within the range reported by studies based on landmarks and recurrent networks, such as Rego et al. and Alves et al. However, direct comparisons are limited by differences in the number of classes, signers, datasets, and evaluation protocols. Superior results, such as those of Samaan et al. and Varshini and Rukmani, should also be interpreted considering the acquisition conditions and splits of their respective datasets. In this context, the cross-signer result obtained in this work shows that generalization to new signers remains a challenge.

F. Real-time application evaluation

The real-time application was evaluated considering the state machine's ability to identify the onset and offset of executions, the occurrence of false triggers, and the classifier's inference latency. For the trigger evaluation, a signer remained in front of the application, with visualization of the classification result and the state machine states (*WAITING* and *CAPTURING*). 10 repetitions of each of the nine signs were requested, performed in a paused and individual manner, waiting for the return to the *WAITING* state before the start of the next execution. A second evaluator manually recorded, by means of a button, the onset and offset instants of each sign, taking into account in this procedure the reaction error associated with manual marking. A tolerance of ± 500 ms between the manual and automatic markings was adopted to consider the detection correct. After processing the records, 92 occurrences were considered in the comparison between the manual and automatic markings.

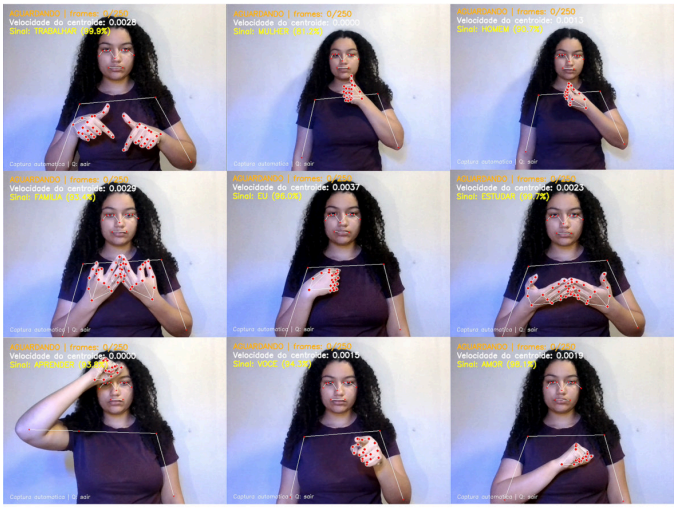


Fig. 5. Examples of real-time classification for the nine sign classes. The interface displays the detected *landmarks*, the predicted class, and the highest *softmax* probability. The interface labels are presented in Portuguese: *AGUARDANDO* [WAITING], *Sinal* [Sign], *Velocidade do centroide* [Centroid speed], *Captura automática* [Automatic capture], and *Q: sair* [Q: quit]. The predicted sign names shown in the images correspond to TRABALHAR [WORK], MULHER [WOMAN], HOMEM [MAN], FAMÍLIA [FAMILY], EU [I], ESTUDAR [STUDY], APRENDER [LEARN], VOCÊ [YOU], and AMOR [LOVE]. *Source: Authors' own elaboration.*

TABLE VIII
EVALUATION OF FALSE TRIGGERS UNDER DIFFERENT ACTIVITY CONDITIONS

Condition	**1**	**2**	**3**	**4**	**5**
Duration (min)	5.48	5.08	7.14	5.20	5.15
False triggers	0	36	17	38	78
Average speed/frame	0.0013	0.0092	0.0040	0.0084	0.0114
Average softmax output (%)	–	73.86	80.05	84.94	88.64

To evaluate the occurrence of false triggers, five activity blocks were performed, each with an approximate duration of five minutes. The first block corresponded to rest; the second, to everyday movements with no communicative intent, such as touching one's hair, touching the nose, scratching, and yawning; the third, to cell phone use; the fourth, to activities involving objects and hand movements close to the body, such as flipping through and reading a book, writing, holding a pen, and keeping a book close to the chest; and the fifth, to drinking water and performing natural movements of raising and lowering the hands, simulating displacement between the resting position and the sign execution region. The exact durations and the results obtained are presented in Table VIII.

In the trigger evaluation, the state machine presented a correct detection rate of 77.17%, considering the established tolerance. Among the detections considered correct, the mean error in identifying the onset was -132.6 ± 178.5 ms, and for the offset, 187.3 ± 212.0 ms. The negative sign in the first case indicates that, on average, capture began before the manual reference, while the positive value in the second indicates that termination occurred afterward.

In the false trigger evaluation, no activations were observed during rest. Among the other conditions, the highest occurrence was observed in the fifth block, with 78 activations, followed by the fourth, with 38. This result indicates that natural hand repositioning movements and activities performed close to the body may present sufficient characteristics to trigger the state machine, even without the intentional execution of a sign.

Inference latency was measured in an independent session, considering the interval between the classifier receiving the data and the emission of the class and its respective confidence. 194 predictions were analyzed, yielding an average latency of 50.5 ± 42.7 ms, with values ranging from 44.3 to 643.6 ms. This result indicates that the classifier's response time is short for most inferences, although occasional instances of higher latency were observed.

G. Limitations

This study has limitations that should be considered when interpreting the results. Although the horizontal mirroring used to expand the dataset is linguistically valid for the selected signs, samples derived from the same recording may not be

fully independent; since all input representations were evaluated on the same test set, however, any residual bias introduced by this augmentation is common to all configurations and does not compromise the relative comparison among them. The experimental evaluation relied on a single stratified split (81 test samples), and the *cross-signer* experiment involved a single held-out signer; absolute performance values should therefore be interpreted with caution, and cross-validation and *leave-one-signer-out* protocols constitute natural extensions of this evaluation. In addition, the real-time application did not include a rejection mechanism, favoring recall in automatic segmentation at the cost of false triggers — an acceptable trade-off for an exploratory assistive system, to be addressed through confidence-based rejection thresholds before deployment. Finally, the ablation test was conducted only on the engineered representation, since the formational-parameter groups (HS, LOC, and MOV) are semantically defined only in this configuration.

V. Conclusion

This work presented a system for the classification of isolated Libras signs using landmarks extracted by MediaPipe Hands and MediaPipe Pose, combined with features related to hand configuration, location, and movement, and classified by an LSTM network. Different data representations were evaluated, with the concatenated representation presenting the best performance, with 93.82% overall accuracy and 93.81% macro F1-score.

The per-class analysis showed that the model presented high performance for most signs, although some classes concentrated the classification errors. The *cross-signer* experiment presented 83.69% accuracy and 82.83% macro F1-score, showing a performance reduction when the model is applied to a signer not observed during training. The ablation test also indicated a greater contribution from the hand configuration features, whose removal resulted in the largest performance reduction among the evaluated groups.

In the real-time application, the state machine presented 77.17% of detections within the ± 500 ms tolerance, while the false trigger evaluation showed greater sensitivity to natural hand repositioning movements and activities performed close to the body. The average inference latency was 50.5 ms in an independent session, indicating a response time compatible with real-time operation, although occasional instances of higher latency were observed.

As future directions, the expansion of the dataset with greater diversity of signers and execution conditions stands out, as well as the improvement of automatic segmentation to reduce false triggers and improve the identification of sign boundaries. These advances can contribute to increasing the generalization capability and robustness of the system in usage scenarios closer to real applications.

References

[1] A. S. Santos and A. J. F. Portes, "Perceptions of deaf subjects about communication in primary health care," *Rev. Latino-Am. Enfermagem*, vol. 27, Art. no. e3127, 2019, doi: 10.1590/1518-8345.2612.3127.

[2] Brasil, "Decreto n.º 5.626, de 22 de dezembro de 2005," Presidência da República, Brasília, DF, Brasil, 2005. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm

[3] P. P. Chandanabhumma *et al.*, "Examining the differences of perceptions and experience with online health information accessibility between deaf and hearing individuals: A qualitative study," *Patient Educ. Couns.*, vol. 122, Art. no. 108169, 2024, doi: 10.1016/j.pec.2024.108169.

[4] L. S. García *et al.*, "Deaf inclusion through Brazilian Sign Language: A computational architecture supporting artifacts and interactive applications and tools," in *Universal Access in Human-Computer Interaction: Access to Media, Learning and Assistive Environments*, M. Antona and C. Stephanidis, Eds. Cham, Switzerland: Springer, 2021, pp. 167–185, doi: 10.1007/978-3-030-78095-1_14.

[5] A. Wadhawan and P. Kumar, "Sign language recognition systems: A decade systematic literature review," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 785–813, 2021, doi: 10.1007/s11831-019-09384-2.

[6] J. L. S. Silva, G. S. Vieira, A. U. Fonseca, and F. Soares, "Reconhecimento e tradução de sinais de Libras para Língua Portuguesa escrita usando redes neurais profundas," in *Proc. XXIV Congresso Brasileiro de Automática (CBA 2022)*, Fortaleza, Brasil, 2022, doi: 10.20906/CBA2022/3720.

[7] G. H. Samaan *et al.*, "MediaPipe's landmarks with RNN for dynamic sign language recognition," *Electronics*, vol. 11, no. 19, Art. no. 3228, 2022, doi: 10.3390/electronics11193228.

[8] T. S. Varshini and P. Rukmani, "MP-GestLSTM: Real-time gesture detection using MediaPipe and LSTM," *Syst. Sci. Control Eng.*, vol. 13, no. 1, Art. no. 2587853, 2025, doi: 10.1080/21642583.2025.2587853.

[9] L. Ferreira-Brito, *Por uma Gramática das Línguas de Sinais*. Rio de Janeiro, Brasil: Tempo Brasileiro/UFRJ, 1995.

[10] C. Lugaresi *et al.*, "MediaPipe: A framework for perceiving and processing reality," in *Proc. 3rd Workshop Comput. Vis. AR/VR at IEEE CVPR*, Long Beach, CA, USA, 2019.

[11] F. Zhang *et al.*, "MediaPipe Hands: On-device real-time hand tracking," arXiv:2006.10214, 2020, doi: 10.48550/arXiv.2006.10214.

[12] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," arXiv:2006.10204, 2020, doi: 10.48550/arXiv.2006.10204.

[13] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] W. C. Stokoe, *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*, Studies in Linguistics: Occasional Papers, no. 8. Buffalo, NY, USA: Univ. Buffalo, 1960.

[16] R. Battison, *Lexical Borrowing in American Sign Language*. Silver Spring, MD, USA: Linstok Press, 1978.

[17] L. B. Karnopp, "Phonology acquisition in Brazilian Sign Language," in *Directions in Sign Language Acquisition*, G. Morgan and B. Woll, Eds. Amsterdam, The Netherlands: John Benjamins, 2002, pp. 29–53, doi: 10.1075/tilar.2.05kar.

[18] T. A. Felipe, "Os processos de formação de palavra na Libras," *ETD – Educ. Temát. Digit.*, vol. 7, no. 2, pp. 200–217, 2006, doi: 10.20396/etd.v7i2.803.

[19] CAPOVILLA, Fernando César et al. Dicionário da Língua de Sinais do Brasil: a libras em suas mãos. . São Paulo: EDUSP. . Acesso em: 14 set. 2026. , 2017

[20] C. L. M. Ferraz, *Dicionário de Configurações das Mãos em Libras*. Cruz das Almas, BA, Brasil: EDUFRB, 2019.

[21] R. C. B. Rego, L. M. G. de Morais, and W. M. Almeida, "Brazilian sign language recognition using deep learning based on fast Fourier transform and kinematic features," *IEEE Access*, vol. 13, pp. 202875–202892, 2025, doi: 10.1109/ACCESS.2025.3637779.

[22] C. E. G. R. Alves, F. D. A. Boldt, and T. M. Paixão, "Enhancing Brazilian sign language recognition through skeleton image representation," in *Proc. 37th SIBGRAPI Conf. Graphics, Patterns and Images*, 2024, pp. 1–6, doi: 10.1109/SIBGRAPI62404.2024.10716301.

Resampling-Based Power Simulation for Sample-Size Planning in Small Serious-Games Rehabilitation Trials: A Reproducible Python Workflow with a Parkinson's Disease Case Study

Edilaine Aparecida da Silva*, Danilo Santos Rocha*, Eduardo Lázaro Martins Naves*, Thiago Braga Rodrigues†
*Faculty of Electrical Engineering, Universidade Federal de Uberlândia (UFU), Uberlândia, Brazil
†Technological University of the Shannon (TUS), Athlone, Ireland
Corresponding author: edilaine.silva@ufu.br

***Abstract*—Sample-size planning is a prerequisite for confirmatory trials of rehabilitation technologies such as serious games, yet the parametric power formulas in common tools assume normality and an externally specified effect size—assumptions that cannot be verified in the small, non-normal pilots typical of this field and that mismatch the non-parametric tests usually planned for analysis. We describe a reproducible, distribution-free workflow that estimates statistical power by resampling pilot data with replacement and applying the planned Mann–Whitney U test, yielding power curves derived directly from the observed data structure. A compact Python implementation (fewer than 15 lines of core code, using only widely available scientific Python libraries, runnable in Google Colab) is provided in full. As a worked case study we use a non-randomised pilot of a serious-games motor–cognitive protocol versus conventional physiotherapy in 22 adults with Parkinson's disease (control n=10, experimental n=12). Across five outcomes, four reached $\geq 80\%$ power—strictly conditional on the pilot effect—at current sizes, whereas the six-minute walk test (6MWT) required $n \approx 24$ per group; every estimate carries a Monte Carlo confidence interval. We emphasise the central caveat: resampling removes distributional and test-mismatch problems but not the large effect-size uncertainty inherent to tiny, non-randomised pilots, so estimates must be read as conditional and buffered, and do not replace validation in a larger sample. Positioned within the development–validation–translation pathway of a marker-less serious-games system, the workflow offers a transparent, test-matched sample-size justification for confirmatory evaluation.**

***Index Terms*—statistical power, sample size, bootstrap, resampling, serious games, neurorehabilitation, Parkinson's disease, Mann–Whitney U, reproducible research**

I. Introduction

Serious games and exergames are increasingly used as rehabilitation technologies in movement and cognitive disorders, including Parkinson's disease (PD), where they target balance, mobility and executive function through engaging, sensor-mediated tasks [1]. Moving such a system from a feasibility pilot to a confirmatory randomised controlled trial (RCT) requires a defensible sample-size justification, increasingly expected by regulators, ethics committees and journals [2], [3]. For rehabilitation-engineering groups, this places a statistical bottleneck on the translational path of the technology they develop: a feasibility-validated device cannot progress to confirmatory evaluation without a credible recruitment target.

Standard power calculation—for example Cohen's d with normal-distribution assumptions, or dedicated software such as G*Power [4], [5]—requires the analyst to specify an expected effect size, a standard deviation, and a distributional family. In the small pilots typical of game-based rehabilitation ($n < 30$ per group), these quantities are estimated with high uncertainty and normality cannot be reliably tested [6]. Worse, the non-parametric tests appropriate for small, bounded, ceiling-prone outcomes—particularly the Mann–Whitney U test—are not directly supported by many parametric power tools, mismatching the calculation and the planned primary analysis.

Resampling-based power simulation offers a distribution-free alternative [7]. Instead of a parametric model, it uses the observed pilot data as a non-parametric estimate of the population distribution and simulates the power of the planned test by repeatedly resampling from that empirical distribution. In this sense the procedure generates synthetic data: each resample is a synthetic cohort drawn from the pilot's empirical distribution, and power is read off the planned test across these cohorts. This requires no distributional assumptions, naturally handles non-normal and bounded outcomes, and uses the same test that will be reported. It is important to state at the outset what the method does *not* do: because the pilot is small and non-randomised, the underlying effect size is itself estimated with substantial error, and resampling propagates—rather than removes—this uncertainty [8]–[10]. We therefore treat the resulting power estimates as conditional on the pilot, report their Monte Carlo precision explicitly, and stress that they are a planning aid rather than a substitute for validation in a larger, randomised sample.

This paper contributes: (1) a rationale for distribution-free power simulation in serious-games trials; (2) a complete, reproducible Python workflow using only widely available scientific libraries (`numpy`, `scipy`); (3) a worked PD case study for five outcomes with Monte Carlo intervals; and (4) a quantitative discussion of when the estimates are trustworthy, with recommendations for a confirmatory trial.

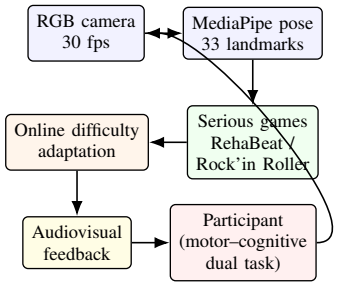


Fig. 1. Architecture of the ReabNet system. A standard webcam feed is converted to a 33-landmark whole-body pose by MediaPipe (BlazePose), in the browser; the pose drives the games, whose difficulty adapts online, closing a real-time audiovisual feedback loop. The bounded, dual-task, sensor-derived outcomes it produces are the non-normal variables the resampling workflow plans around.

II. Methods

A. Case Study Dataset and Intervention Context

We use data from a non-randomised controlled pilot trial comparing a serious-games motor–cognitive rehabilitation protocol with conventional physiotherapy in 22 adults with mild-to-moderate PD (control n=10, experimental n=12; Hoehn & Yahr stages 1–3; mean age 66.5 ± 7.9 years). Five outcomes were assessed: MoCA, Berg Balance Scale, six-minute walk test (6MWT), Trail Making Test Part A (TMT-A) and Digit Span Forward; pre–post change scores (Δ = post − pre) were the analytical variables. Ethical approval: UFU Research Ethics Committee (CAAE 83857524.7.0000.5152); the de-identified dataset and code are openly available [15]. The trial was *not* randomised and exhibited baseline imbalance (notably greater initial aerobic capacity in the experimental group); because change scores are analysed under this imbalance, the estimates carry a regression-to-the-mean risk whose implications are addressed in Sections III-C and IV-C.

B. Serious-Games Rehabilitation System and Protocol

The experimental intervention used ReabNet, a markerless serious-games platform for motor–cognitive rehabilitation (Fig. 1). Body motion is captured from an ordinary webcam and converted to a 33-landmark whole-body pose estimate in real time by MediaPipe (BlazePose) [14], running client-side in the browser; no markers or wearables are attached to the participant. Tracking accuracy was characterised against a Vicon optical reference (mean positional error 2.3 cm), adequate for game control. Two purpose-built games drive the therapy: RehaBeat, a rhythm game eliciting bilateral upper-limb reaching synchronised to audiovisual cues, and Rock'in Roller, a weight-shifting platform game eliciting medio-lateral and antero-posterior transfers, partial squats and lateral steps to train dynamic, closed-chain balance and anticipatory postural adjustments. Both embed a cognitive–motor dual-task component, and difficulty is adapted online from the hit rate of the last 50 actions (raised above 80%, lowered below 60%).

The experimental group completed 16 physiotherapist-supervised sessions over 8 weeks (two 50-min sessions/week: warm-up, RehaBeat 3×5 min, Rock'in Roller 3×5 min, cool-down); adherence exceeded 96%. The control group received dose-matched individualised physiotherapy (same frequency and length, without interactive technology).

TABLE I
Power-estimation approaches for small-sample rehabilitation trials.

Criterion	Param. (t)	G*Power	Sim. (normal)	Resampling
Distributional assumption	Normal	Normal	Specifiable	None (empir.)
Requires external effect size	Yes	Yes	Yes	No (data)
Suitable for $n < 30$/group	Poor	Poor	Partial	Yes
Matches non-param. test	No	No	Partial	Yes
Resolves effect-size unc.	No	No	No	No†

†No method overcomes effect-size imprecision from a small pilot; resampling makes this explicit and is paired with a buffer (Section III-C).

As Fig. 1 makes explicit, the outcomes that this system produces are bounded, dual-task and sensor-derived—exactly the non-normal, ceiling-prone variables for which parametric power formulas are least appropriate. This is where the contribution connects to biomedical engineering. ReabNet follows a development–validation–translation pathway, and the resampling workflow operates at the validation-to-translation interface: it converts the device's bounded, test-matched outcomes into a defensible recruitment target, so sample-size planning becomes an integral engineering deliverable of the platform rather than an external afterthought.

C. Resampling Power Simulation Algorithm

For a target per-group sample size n, N_{sim} pairs of bootstrap samples are drawn with replacement from the pilot data—one sample of size n from the control group and one from the experimental group. The planned Mann–Whitney U test is applied to each pair, and the proportion of replicates with $p < \alpha$ estimates power at n. Repeating across a grid of candidate sizes yields a power curve. Table I positions this approach against parametric and normal-simulation alternatives on the dimensions that matter for small rehabilitation samples.

D. Reproducible Python Implementation

The core routine accepts two pilot arrays and returns power across a range of sample sizes (Listing 1); it depends only on `numpy` and `scipy.stats` [13], both widely available in standard scientific Python environments.

Listing 1. Core resampling power-simulation function.

```
import numpy as np
from scipy import stats
def power_simulation(ctrl, exp, n_range=range(8,51,2),
                     n_sim=2000, alpha=0.05, seed=42):
    rng = np.random.default_rng(seed)
    powers = []
    for n in n_range:
        sig = 0
        for _ in range(n_sim):
            c = rng.choice(ctrl, size=n, replace=True)
            e = rng.choice(exp,  size=n, replace=True)
            _, p = stats.mannwhitneyu(e, c,
                        alternative='two-sided')
            sig += (p < alpha)
        powers.append(sig / n_sim)
```

TABLE II
PILOT EFFECT SIZE (RANK-BISERIAL r WITH CASE-RESAMPLING 95% CI) AND RESAMPLING POWER (%) PER OUTCOME ($N_{\text{SIM}} = 2{,}000$, SEED 42; MONTE CARLO 95% CI HALF-WIDTH $\leq \pm 2.2$ PP). POWER AT n=12, 20 IS SHOWN IN FIG. 2.

Outcome	r	95% CI	n=8	n=10	n@80%
MoCA	0.80	[0.51, 1.00]	89.7	96.8	< 8
Berg	0.90	[0.69, 1.00]	99.0	100.0	< 8
6MWT	0.44	[0.00, 0.88]	34.1	41.2	24
TMT-A	0.68	[0.33, 1.00]	75.0	85.0	10
Digit Span	0.66	[0.29, 1.00]	69.1	80.7	10

```
    return list(n_range), powers
```

The function is applied per outcome, and the minimum n reaching the target power is extracted; the complete multi-outcome driver and plotting code are in the repository [15].

E. Simulation Parameters and Uncertainty

We used n = 8–50 per group (step 2), $N_{\text{sim}} = 2{,}000$, $\alpha = 0.05$, a power target of 80%, and a fixed seed for reproducibility. The full run (22 sizes $\times$ 5 outcomes $\times$ 2,000 iterations) completes in $\approx$ 8 min on the Google Colab free tier.

Two distinct sources of uncertainty must be separated. The Monte Carlo error of a power estimate $\hat{\pi}$ is binomial, $\text{SE} = \sqrt{\hat{\pi}(1-\hat{\pi})/N_{\text{sim}}}$, which at $N_{\text{sim}} = 2{,}000$ never exceeds ± 2.2 percentage points (95%); this is controllable by increasing N_{sim}. For $\hat{\pi}$ near 0 or 1 the symmetric binomial interval is inaccurate, so near-ceiling estimates use Wilson score intervals, which are asymmetric and narrower toward the bound. The far larger and *irreducible* source is the sampling error of the effect size itself, inherited from the small pilot; this is not reflected in the Monte Carlo interval and is the basis for the conservative buffer recommended later.

F. Simulator Validation

Applied to normal data with a known effect (d=0.8), the routine tracked the analytical two-sample power within 4 percentage points across n=8–30, sitting slightly below it as expected from the relative efficiency of Mann–Whitney under normality. The simulator thus recovers the correct power when assumptions hold, isolating the pilot dependence as the only remaining source of error. Its bias and confidence-interval miscalibration across distributions, effect sizes and pilot sizes are characterised in a companion calibration study [16].

III. RESULTS

A. Power at the Current Sample Size

Table II reports power at selected sizes and the minimum n for 80% power. Conditional on the pilot effect, four outcomes (MoCA, Berg, TMT-A, Digit Span) were adequately powered at current sizes, whereas the 6MWT was underpowered (41.2%) and required $n \approx 24$ per group (N=48). The very high power of Berg and MoCA already at n=8 must be read with explicit caution (Section IV-C): in a small, non-randomised sample, a near-ceiling estimate is more plausibly

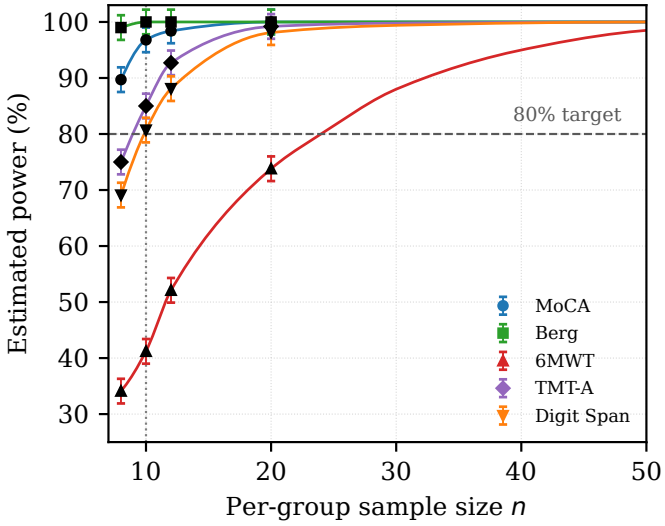


Fig. 2. Resampling-based power curves for five outcomes. Markers are simulated estimates at the reported sizes; whiskers are Monte Carlo 95% confidence intervals; connecting lines are monotone interpolation. Dashed line: 80% target; dotted line: current $n \approx 10$ per group. The 6MWT crosses 80% at $n \approx 24$.

a symptom of an over-stated pilot effect than evidence of a robust, replicable difference, and should not be reported as the latter. Table II lists these effects: r is large for the cognitive and balance scales (0.66–0.90) and moderate for the 6MWT (0.44), yet every 95% CI is wide—reaching the r=1 bound for the ceiling outcomes and down to ≈ 0 for the 6MWT. The decisive power of Berg and MoCA thus rests on effect estimates barely distinguishable from near-perfect separation: the quantitative basis for treating the power values as conditional.

B. Power Curves

Fig. 2 shows the five power curves with Monte Carlo confidence intervals. Cognitive and balance outcomes plateau near the ceiling at small n, consistent with strong ceiling effects in these bounded scales—but, as above, also with effect-size over-estimation in a tiny pilot. The 6MWT rises gradually without a plateau before $n \approx 24$; its monotone shape—rather than a flat curve near zero power—indicates a real but moderate effect, i.e. an underpowered rather than null result at the current size.

C. Implications for a Confirmatory Trial

The simulation supports the following plan for a confirmatory RCT. (1) Adopt the 6MWT as the primary endpoint. This choice is made on clinical and methodological grounds rather than because it is merely the power-limiting outcome: functional mobility measured by the 6MWT is a patient-relevant, guideline-endorsed target of gait/balance training in PD, has an established minimal clinically important difference [12], and is the least ceiling-prone of the five outcomes, so it best discriminates treatment response in a confirmatory setting. That it is also the outcome requiring the largest sample makes it the binding constraint for recruitment, which is convenient

but secondary to its clinical primacy. (2) Recruit $n = 24$ per group (N=48), with a 15–20% attrition and effect-uncertainty buffer (target N=56–58). (3) Retain MoCA, Berg, TMT-A and Digit Span as secondary endpoints, adequately powered at $n \geq 10$ per group conditional on the pilot effect.

Two safeguards follow from the pilot's weaknesses. Because it was non-randomised with baseline imbalance and used change scores, the confirmatory trial must randomise and pre-specify covariate adjustment (e.g. ANCOVA) to control regression to the mean [11]; the large effects on bounded scales may shrink once this is in place, a further reason to keep the buffer. These are planning estimates, not guarantees, and do not substitute for the larger, randomised validation they dimension.

IV. Discussion

A. What the Method Does—and Does Not—Solve

The workflow removes three problems of parametric planning: it makes no distributional assumption, requires no externally supplied effect size, and evaluates the exact non-parametric test to be reported—real advantages for bounded, ceiling-prone outcomes. However, it does not address the dominant limitation of any pilot-based plan: with $n = 10$–12 per group the effect size is estimated very imprecisely, and resampling treats the observed effect as if it were the truth [8], [9]. A second, method-specific caveat: resampling from only 10–12 distinct values cannot create variability beyond the observed support, so at larger candidate sizes the empirical distribution is reused rather than enriched and power near the bounds may be mildly optimistic. We therefore avoid describing the method as "assumption-free" in the colloquial sense: it is *distribution-free*, but it remains strongly conditional on the representativeness of the pilot. Consequently, near-ceiling power at very small n is best read as a signal of effect-size uncertainty, not of robustness—a behaviour characterised systematically in the companion calibration study [16]. We therefore report Monte Carlo intervals (bounding only simulation noise) separately from the buffer (which hedges the larger, unmodelled effect-size error).

B. Comparison with Parametric Calculation

For an approximately symmetric outcome such as MoCA, a Cohen's-d calculation from pilot moments yields power broadly comparable to the resampling estimate. For skewed, high-variance outcomes such as the 6MWT, parametric formulas tend to overstate power—where the distribution-free approach helps most, though neither escapes the imprecision of the pilot effect.

C. Limitations

First and foremost, power estimates are conditional on pilot composition; an atypical pilot biases them either way [10], and the small non-randomised design makes this more likely. We mitigate it with the buffer and by reporting every estimate as conditional. Second, the design was non-randomised with baseline imbalance, and between-group analysis of change scores under unequal baselines is vulnerable to regression to the mean [11]; within-group baseline–change correlations should therefore not be interpreted as treatment effect sizes, and the confirmatory trial should randomise and pre-specify covariate adjustment. Third, testing five outcomes raises multiplicity; the confirmatory design should pre-register one primary endpoint and control family-wise error. Fourth, the estimates are specific to the Mann–Whitney U test; another analysis (e.g. a mixed-effects model) requires re-running with that test—one line in Listing 1. Overall, the workflow should *dimension* a confirmatory study, not claim an effect is established.

V. Conclusion

Resampling-based power simulation is a practical, distribution-free and reproducible way to plan sample sizes for small trials of serious-games rehabilitation technologies, provided its estimates are read as conditional on the pilot. Applied to a PD pilot of a marker-less serious-games system, it identified four adequately powered secondary endpoints and one underpowered primary endpoint (6MWT), giving a concrete recommendation ($n = 24$ per group; $N = 48$, buffered to 56–58) for a confirmatory RCT. Its main value is transparency: it estimates power for the test actually planned and makes the dependence on the pilot effect explicit, so estimates are reported and buffered as conditional—never as a replacement for larger-sample validation. Embedded in the development–validation–translation pathway of a sensor-driven platform, it gives rehabilitation-engineering groups a test-matched element of pilot reporting when distributional assumptions cannot be verified and non-parametric analyses are planned.

Acknowledgement

Supported by CAPES, Brazil. The authors used AI-based tools for language editing only; all scientific content, data, code and conclusions were produced and verified by the authors. The authors declare no conflict of interest.

References

[1] G. Barry, B. Galna, and L. Rochester, "The role of exergaming in Parkinson's disease rehabilitation: a systematic review," *J. NeuroEng. Rehabil.*, vol. 11, p. 33, 2014.

[2] J. E. Rudolph *et al.*, "Impact of sample size and its estimation in medical research," *AJPM Focus*, 2025, doi:10.1016/j.ajpmf.2025.00139.

[3] L. A. Renfro *et al.*, "Statistical and methodological considerations for RCT design in physical medicine and rehabilitation," *Am. J. Phys. Med. Rehabil.*, vol. 102, no. 10, pp. 855–860, 2023.

[4] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: a flexible statistical power analysis program," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175–191, 2007.

[5] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988.

[6] L. Thabane *et al.*, "A tutorial on pilot studies: the what, why and how," *BMC Med. Res. Methodol.*, vol. 10, p. 1, 2010.

[7] A. K. Dwivedi, I. Mallawaarachchi, and L. A. Alvarado, "Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling," *Stat. Med.*, vol. 36, no. 14, pp. 2187–2205, 2017.

[8] H. C. Kraemer, J. Mintz, A. Noda, J. Tinklenberg, and J. A. Yesavage, "Caution regarding the use of pilot studies to guide power calculations for study proposals," *Arch. Gen. Psychiatry*, vol. 63, no. 5, pp. 484–489, 2006.

[9] A. C. Leon, L. L. Davis, and H. C. Kraemer, "The role and interpretation of pilot studies in clinical research," *J. Psychiatr. Res.*, vol. 45, no. 5, pp. 626–629, 2011.
[10] R. H. Browne, "On the use of a pilot sample for sample size determination," *Stat. Med.*, vol. 14, no. 17, pp. 1933–1940, 1995.
[11] A. G. Barnett, J. C. van der Pols, and A. J. Dobson, "Regression to the mean: what it is and how to deal with it," *Int. J. Epidemiol.*, vol. 34, no. 1, pp. 215–220, 2005.
[12] B. Mishra *et al.*, "Minimum clinically important difference thresholds in movement disorders research: a systematic review," *Heliyon*, vol. 10, no. 5, p. e26479, 2024.
[13] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020.
[14] C. Lugaresi *et al.*, "MediaPipe: a framework for building perception pipelines," *arXiv:1906.08172*, 2019.
[15] E. A. da Silva, "Synthetic data methods for small-sample Parkinson's disease rehabilitation trials," Zenodo, 2026, doi:10.5281/zenodo.20607087 (CC BY 4.0).
[16] E. A. da Silva *et al.*, "Resampling-based power simulation in small samples: a Monte Carlo calibration study," Zenodo, 2026, doi:10.5281/zenodo.20767954 (CC BY 4.0).

sEMGest: An Integrated Platform for Hand Gesture Classification from sEMG with Real-Time Digital-Twin Validation

Bruno Malinoski, Gustavo H. dos Santos Schaefer, Carmen Caroline Rasera,
José Jair A. Mendes Junior and Heitor Silvério Lopes
Federal University of Technology of Paraná (UTFPR)
Curitiba, Paraná, Brazil
{malinoski, gschaefer}@alunos.utfpr.edu.br
{rasera, jjjunior, hslopes}@utfpr.edu.br

***Abstract*—Hand gesture recognition from surface electromyography (sEMG) is a cornerstone of intuitive human–machine interfaces and myoelectric prosthesis control. This work presents *sEMGest*, an integrated and reproducible platform that spans custom acquisition hardware, a signal-processing and machine-learning pipeline, and a real-time three-dimensional digital twin of a prosthetic hand used for validation. A four-channel analog front-end based on INA333 instrumentation amplifiers and Sallen–Key active filters feeds an ESP32-S3 microcontroller sampling at 1 kHz with 12-bit resolution; the board was experimentally characterized, showing an approximate 25–540 Hz passband and a 49.5 µV RMS input-referred noise floor. Signals are band-pass filtered (20–500 Hz), segmented into 256 ms windows with 75% overlap, and described by the four time-domain features of the TD4 set. Three classifiers (linear discriminant analysis (LDA), a support vector machine with a radial basis function kernel (SVM-RBF), and random forest (RF)) were compared through grid search and cross-validation, first on the public NinaPro DB5 dataset and then on a proprietary two-participant dataset acquired with the developed hardware. Using repetition-grouped cross-validation, which prevents leakage between overlapping windows of the same repetition, SVM-RBF and RF are statistically tied intra-subject (macro-F1 0.69 on DB5; 0.53–0.60 on the proprietary data), whereas leave-one-subject-out evaluation collapses to 0.29–0.37, evidencing covariate shift and the need for per-user calibration; RF proves the most robust across subjects. The processing latency stays at 2.55 ms (95th percentile), so that the perceived response time is dominated by the 256 ms window-filling delay and remains within the 300 ms myoelectric-control budget. The digital twin renders the recognized gesture in real time, providing a hardware-independent tool for pipeline validation and, prospectively, user training.**

***Index Terms*—surface electromyography, hand gesture recognition, myoelectric prosthesis, digital twin, embedded systems, machine learning, real-time systems.**

I. Introduction

The human hand is among the most complex and versatile anatomical structures of the body, and the loss of an upper limb (by trauma, vascular disease or congenital causes) compromises daily activities and functional autonomy [1]. Myoelectric prostheses use surface electromyography (sEMG) as a control interface: electrodes placed over the skin capture the electrical potentials associated with muscle contraction and drive motorized actuators more directly than purely mechanical systems [5], [6]. Conventional myoelectric control relies mostly on signal amplitude and becomes insufficient when several functions must be controlled reliably; in contrast, pattern-recognition control infers the intended movement from features extracted from the sEMG, widening the command set and improving naturalness [3]. Nonetheless, upper-limb prosthesis abandonment rates between 23% and 45% are still reported, associated with insufficient perceived control and lack of naturalness [2], [4].

Two obstacles are particularly relevant. First, latency: delays above 100–125 ms start to degrade user performance in manipulation tasks and delays above 300 ms make control significantly harder [7], imposing a tight budget on acquisition, processing and classification. Second, robustness: classification performance measured under controlled laboratory conditions is hard to sustain in continuous use because of session variability, muscle fatigue and electrode repositioning [8]. Many systems are also validated only with specialized acquisition equipment and desktop processing, while reducing the number of channels (desirable for wearable devices) remains feasible when the pipeline is properly designed [9]. In parallel, the digital-twin concept, a virtual representation that mirrors a physical system in real time, has been explored in biomedical applications [11], [12] and, for prostheses, can validate control algorithms in a virtual environment before integration with hardware, while extended-reality interfaces support embodiment and training [10], [13].

This work proposes *sEMGest*, an integrated platform that combines sEMG acquisition, machine-learning gesture classification and a real-time digital twin, targeting practical applicability to upper-limb prosthesis control.[1] Its main contributions are:

- an **integrated and reproducible platform** joining a four-channel acquisition front-end (INA333, OPA4277 and ESP32-S3), a Python processing/classification pipeline

[1]Source code (processing pipeline, digital twin and trained models) is available at https://github.com/bmalinoski/sEMGest.

and a three-dimensional real-time visualization environment, in contrast with commercial black-box solutions;
- a **two-stage validation methodology**, first on the reference NinaPro DB5 dataset and then on a proprietary dataset acquired with the developed hardware;
- a **complete experimental characterization** of the hardware (per-channel frequency response, input-referred noise floor and end-to-end latency);
- **empirical evidence of covariate shift** across subjects and sessions, supporting a two-phase architecture with dedicated per-user models; and
- the application of a **digital twin** that renders the classified gesture in real time, acting both as a pipeline-validation instrument and as potential support for user training.

II. Related Work

Classical machine learning remains competitive for sEMG-based gesture recognition. Comparative studies report high intra-subject accuracy for LDA, SVM and RF [14]–[16], and time-domain feature sets are favored for their low computational cost and robustness in real-time pattern recognition [8], [17]. Segmentation parameters (window length, overlap and truncation) strongly influence performance and trade information for latency [18]. Public datasets such as NinaPro support reproducible benchmarking and setup comparison [19], [20], and embedded/real-time solutions with reduced acquisition have been demonstrated on prosthetic hands [21].

On the visualization side, extended-reality and virtual-prosthesis environments have been used mainly for training and embodiment, or, more recently, as myoelectric digital twins for fast and realistic modelling in deep-learning pipelines [10], [11], [13], [22]. In the present platform, the digital twin is coupled end-to-end with a custom low-cost front-end and is used online as a validation instrument for the full acquisition-to-decision chain, in addition to its role as a visualization and training aid.

III. Materials and Methods

A. System Overview

The platform is organized in three subsystems (Fig. 1): (i) a four-channel acquisition front-end with embedded firmware; (ii) a host processing pipeline in Python that performs digital filtering, segmentation, feature extraction and classification; and (iii) a digital twin that renders the classified gesture in real time. Evaluation followed two stages: pre-validation on NinaPro DB5 to establish an algorithmic performance ceiling, and validation on a proprietary dataset acquired with the developed hardware.

B. Acquisition Hardware

Each of the four channels implements a five-stage conditioning chain. An INA333 instrumentation amplifier (CJMCU-333 module, gain $2\times$) provides differential acquisition with high common-mode rejection, followed by a second-order Sallen–Key high-pass filter (20 Hz) and a second-order Sallen–Key low-pass filter (≈ 540 Hz) built with OPA4277 precision op-amps [23]. A non-inverting stage adds a fixed gain of $\approx 500\times$ (total chain gain $\approx 1000\times$), and a unity-gain subtractor centers the signal at 1.5 V to match the ADC input range. The board is powered by a battery through low-dropout regulators, with an on-board mid-supply reference biasing the analog chain. Digitization uses the 12-bit SAR ADC of an ESP32-S3 microcontroller at 1 kHz, obtained by $4\times$ oversampling of a 4 kHz timer. Electrode placement followed SENIAM recommendations [24].

The assembled printed circuit board was experimentally characterized (Fig. 2). The four channels showed a uniform passband of approximately 25–540 Hz, and the input-referred background noise, with inputs short-circuited and the system battery-powered, was 49.5 μV RMS. Considering that forearm sEMG amplitudes typically range from 100 to 500 μV [6], this noise floor is compatible with gesture recognition.

C. Embedded Firmware

The firmware runs on the ESP32-S3 under FreeRTOS. A high-resolution timer dispatches acquisition, and a set of second-order IIR biquads in Direct Form II Transposed applies a 20 Hz high-pass, a 480 Hz low-pass and a 60 Hz notch. Samples are packed into 15-byte binary frames with an XOR checksum and streamed to the host over USB-CDC.

D. Datasets and Acquisition Protocol

Pre-validation used the public NinaPro DB5 dataset (8-channel Myo armband, 200 Hz) [19], [20]. Seven classes were considered: rest and six functional gestures (closed fist, open hand, pinch, pointing, thumbs up and finger abduction).

The proprietary dataset was collected with the developed four-channel hardware from two participants: P1 (right-handed) and P2 (left-handed). To preserve an identical protocol, acquisition was always performed on the right forearm, which is the non-dominant limb for P2. Each gesture followed a cue schedule of 5 s of contraction and 3 s of rest, with 15 repetitions per gesture and participant. Windows were labeled by activity-envelope detection, which locates each contraction directly from the signal amplitude and thereby avoids the misalignment between the manually timed 5/3 s cue and the true contraction onset that a purely schedule-based labeling would introduce.

All experimental procedures involving human participants were conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of the Federal University of Technology of Paraná (protocol no. 7.218.395). Both participants were members of the research team and provided written informed consent prior to data collection.

E. Signal Processing Pipeline

The host pipeline applies a 20–500 Hz band-pass filter and segments each channel into 256 ms windows with 75% overlap, parameters validated in the literature for real-time applications [18]. Each window is described by the four time-domain features of the TD4 set (MFL, MSR, WAMP and

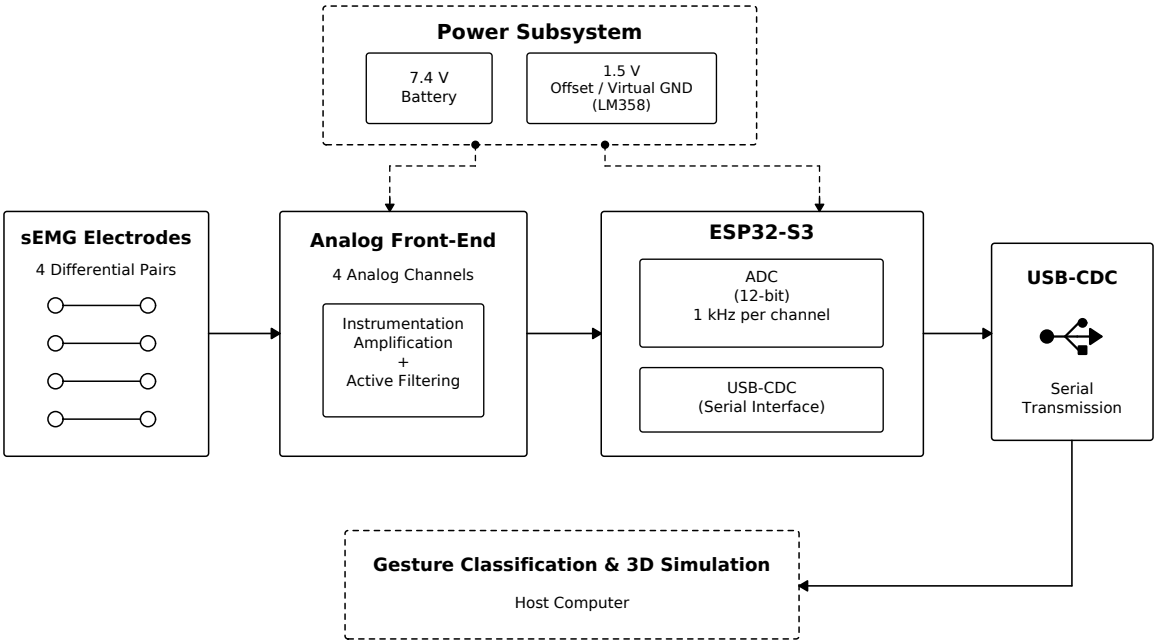
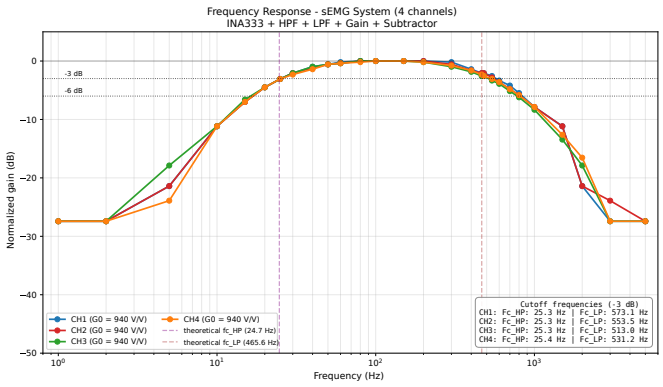
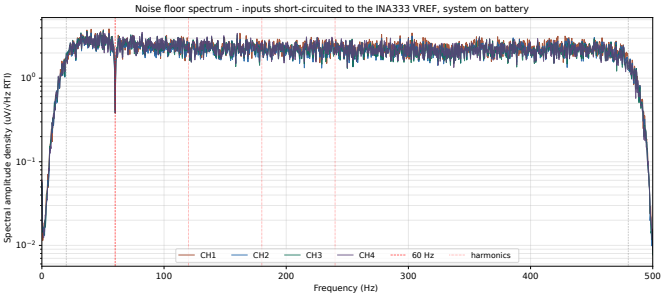


Fig. 1: Overview of the sEMGest platform: four-channel analog acquisition and ESP32-S3 firmware, host processing/classification pipeline, and real-time digital twin of the prosthetic hand.



(a) Magnitude frequency response of the four channels.



(b) Power spectral density of the input-referred background noise.

Fig. 2: Experimental hardware characterization: (a) channel frequency response and (b) input-referred noise floor.

LS) [17], yielding 16 features per window (4 features $\times$ 4 channels). Features are normalized and class imbalance is mitigated by random undersampling of the majority (rest) class to the median of the minority classes.

F. Classifiers and Validation

Three classifiers were implemented with scikit-learn [25] and compared: LDA, SVM-RBF and RF. Hyperparameters were selected by nested grid search, and every cross-validation split, both the outer evaluation loop and the inner grid-search loop, was grouped by *repetition*, so that all windows originating from the same contraction stay in the same fold. This is essential because the 75% window overlap makes neighboring windows of one repetition nearly identical; a random split would place such near-duplicates in both training and test folds and inflate the scores. The intra-subject setting used repetition-grouped cross-validation computed *per subject* (each of the ten NinaPro DB5 subjects individually, and dedicated per-participant models for P1 and P2), reporting the mean and standard deviation across subjects. Cross-subject generalization used full Leave-One-Subject-Out (LOSO) over the ten DB5 subjects, rotating the held-out subject, and the two-direction LOSO between P1 and P2 on the proprietary dataset. For every protocol we report macro-F1, balanced accuracy, and a gesture-only macro-F1 (excluding the rest class), together with the exact per-class window counts and the proportion of rest in the test set.

G. Digital Twin

The digital twin materializes a virtual prosthetic hand that mirrors, in real time, the gesture predicted by the pipeline (Fig. 3). It was implemented in Python with a pygame event loop and legacy OpenGL (PyOpenGL) rendering, in a 1280×720 window split into a 3D viewport in the upper region and a signal panel showing the filtered EMG waveforms and per-channel RMS activation bars in the lower region.

The hand is a serial anthropomorphic kinematic chain of five fingers, with two joints for the thumb (MCP, IP) and three for each remaining finger (MCP, PIP, DIP), dimensioned from the proportions of the open-source OpenBionics model. Each gesture class is encoded as a vector of flexion and abduction joint angles, and phalanx positions are obtained by forward kinematics through the composition of homogeneous 4×4 transforms:

$$T_n = T_{\text{base}} \cdot \prod_{i=1}^{n} R_x(\theta_i)\, T_z(l_i), \tag{1}$$

where θ_i is the flexion angle of joint i, l_i the length of phalanx i, R_x a rotation about the x axis and T_z a translation along z. Bones are rendered as cylinders and joints as spheres, one color per finger. Transitions between successive poses are smoothed by a 200 ms *smoothstep* interpolation,

$$s(t) = t^2\,(3 - 2t), \quad t \in [0, 1], \tag{2}$$

applied to the joint angles to avoid abrupt motion of the 3D model.

The input layer follows a strategy pattern: a common interface abstracts three interchangeable sources feeding the same pipeline and renderer—a *replay* source that streams NinaPro DB5 recordings sample by sample, a *hardware* source fed by the four-channel USB-CDC stream of the ESP32-S3, and a *demo* source for manual pose inspection. A heads-up display overlays the classified gesture, its confidence, the classification latency and the frame rate, and the model can be rotated interactively for inspection. To suppress visual flicker when classifier confidence is low, the pose is updated only when the prediction exceeds a 90% confidence threshold—the previous pose is held otherwise, and the hand reverts to rest after a prolonged low-confidence interval. This acts as a low-cost persistence scheme in place of the heavier sliding-voting post-processing used elsewhere for output smoothing.

IV. Results and Discussion

A. Intra-Subject Classification on NinaPro DB5

Intra-subject performance was evaluated per subject and then averaged across the ten DB5 subjects. Crucially, the cross-validation folds are grouped by *repetition*: because the 256 ms windows overlap by 75%, neighboring windows of the same repetition are nearly identical, and a random split would place such near-duplicates in both the training and the test folds, measuring interpolation within a repetition rather than generalization to unseen repetitions. Grouping every window of a given repetition into the same fold (in both the outer loop and the inner grid-search loop) removes this optimistic bias. Under this protocol, SVM-RBF and RF are statistically tied, with macro-F1 of 0.691 ± 0.050 and 0.692 ± 0.051 respectively, both clearly ahead of LDA (0.656 ± 0.057); balanced accuracy follows the same ordering (Table I). The dispersion across subjects ($\sigma \approx 0.05$) reflects the genuine inter-individual variability of forearm sEMG.

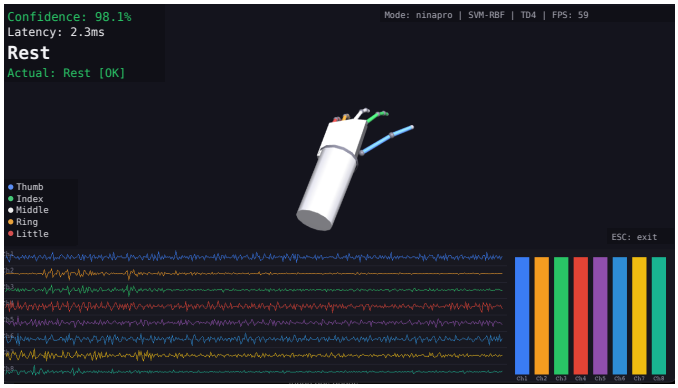


Fig. 3: Digital twin in operation, with the 3D viewport, the EMG signal panel with per-channel activation bars, and the informational HUD (classified gesture, confidence, latency and frame rate).

These figures are markedly lower than those obtained with a subject-aggregated, randomly split validation (macro-F1 up to 0.910), and the gap is precisely the inflation introduced by inter-window leakage between overlapping segments of the same repetition. The repetition-grouped values are therefore the ones that reflect deployment, where the model must classify contractions it has not seen during training. Across subjects, the residual errors concentrate between the lower-amplitude gestures and the rest class.

B. Results on the Proprietary Dataset

On the four-channel proprietary hardware, dedicated per-participant models were trained with the same repetition-grouped cross-validation. The classifier ordering seen on DB5 is preserved (Table I): RF and SVM-RBF are close, with RF marginally ahead (macro-F1 0.601 ± 0.033 vs. 0.590 ± 0.029 for P1; 0.528 ± 0.037 vs. 0.526 ± 0.054 for P2), and both clearly outperform LDA. Absolute values are lower than on DB5, consistent with a four-channel montage and a single acquisition session.

Fig. 4 shows the RF confusion matrix for P1. The dominant error is confusion between the gestures and the rest class in both directions: rest is recovered in only 44% of its windows, and pinch, the weakest gesture, in 51%, with about 31% of its windows assigned to rest. The higher-amplitude, topographically distinct gestures are recognized reliably (thumbs up 83%, finger abduction 78%, pointing 77%). This rest-boundary confusion, rather than confusion between biomechanically similar gesture pairs, is the main error mode on the four-channel montage.

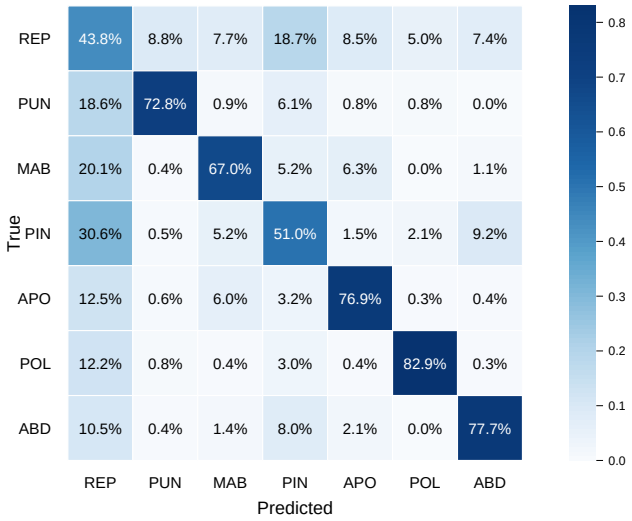


Fig. 4: Confusion matrix (true-class normalized) of the RF classifier for participant P1, from repetition-grouped cross-validation on the proprietary four-channel hardware. Classes: REP (rest), PUN (closed fist), MAB (open hand), PIN (pinch), APO (pointing), POL (thumbs up), ABD (finger abduction).

TABLE I: Repetition-Grouped Cross-Validation Results (Mean $\pm$ Std). Macro-F1 and Bal. Acc. are over all seven classes; F1 (gest.) is the macro-F1 over the six gestures, excluding rest.

Protocol	Clf.	Macro-F1	Bal. Acc.	F1 (gest.)
Intra-subject				
	LDA	0.656 ± 0.057	0.801 ± 0.066	0.611 ± 0.064
DB5	SVM-RBF	0.691 ± 0.050	0.829 ± 0.063	0.649 ± 0.058
	RF	0.692 ± 0.051	0.826 ± 0.061	0.650 ± 0.058
	LDA	0.495 ± 0.033	0.552 ± 0.045	0.510 ± 0.044
P1	SVM-RBF	0.590 ± 0.029	0.643 ± 0.041	0.603 ± 0.036
	RF	0.601 ± 0.033	0.657 ± 0.045	0.613 ± 0.042
	LDA	0.437 ± 0.031	0.492 ± 0.030	0.454 ± 0.032
P2	SVM-RBF	0.526 ± 0.054	0.565 ± 0.051	0.525 ± 0.059
	RF	0.528 ± 0.037	0.579 ± 0.037	0.529 ± 0.042
Inter-subject / inter-participant				
	LDA	0.360 ± 0.085	0.475 ± 0.119	0.275 ± 0.085
DB5 LOSO	SVM-RBF	0.288 ± 0.092	0.316 ± 0.118	0.180 ± 0.103
	RF	0.370 ± 0.106	0.474 ± 0.114	0.279 ± 0.113
	LDA	0.135	0.214	0.106
Own LOSO	SVM-RBF	0.168	0.210	0.132
	RF	0.228	0.278	0.191

The ≈ 6 percentage-point gap between P1 and P2, under identical hardware, pipeline and classifiers, is attributed to acquisition conditions rather than to the methodology. Two factors are consistent with the data: P2 is left-handed but was acquired on the right (non-dominant) forearm to preserve an identical protocol, which reduces the activation uniformity of the target muscles [26]; and a thicker layer of subcutaneous adipose tissue hindered palpation of the muscle bellies, suggesting slightly less precise electrode placement. The small participant count and single session place these figures at the level of a proof of concept rather than a population estimate.

C. Inter-Subject and Inter-Session Generalization

Generalization to unseen subjects was assessed by leave-one-subject-out (LOSO) over all ten DB5 subjects, rotating the held-out subject (Table I). Performance collapses relative to the intra-subject case: the best classifier reaches only 0.370 ± 0.106 macro-F1 (RF), against 0.692 intra-subject. Balanced accuracy tells the same story, and on the gesture-only subset (excluding rest) it falls to 0.18–0.28. Notably, the classifier ranking *inverts* with respect to the intra-subject setting: RF (0.370) and LDA (0.360) now lead, while SVM-RBF drops to 0.288. Raw accuracy is uninformative in this regime: with $\approx 92\%$ of the test windows being rest, SVM-RBF's 0.865 is in fact below the 0.923 of a trivial rest-only predictor, while its balanced accuracy of 0.316 exposes the poor gesture discrimination. The inter-participant LOSO on the proprietary dataset reproduces both the collapse and the inversion (RF 0.228 vs. SVM-RBF 0.168 macro-F1).

Two conclusions follow. First, the collapse is the expected signature of covariate shift (systematic mean and scale changes of the TD4 space across subjects, electrode placement and skin-electrode impedance) which violates the assumption that training and test share the same distribution [4], [8], and it motivates a two-phase architecture with an individualized calibration stage. Second, the flexible RBF boundaries that are competitive within a subject are the least robust across subjects, whereas the simpler decision surfaces of RF and LDA transfer slightly better, an argument for preferring RF as the deployed classifier, since it is both the best (tied) intra-subject and the most robust across subjects.

D. Real-Time Latency

The end-to-end responsiveness of the system is the sum of two contributions of different natures, summarized in Table II: a fixed *windowing* delay imposed by the segmentation scheme, and the *computation* time spent turning a window into a decision. The computation time (feature extraction, normalization and inference) was measured on the proprietary hardware over more than 1,000 consecutive classifications, yielding a mean of 1.79 ms and a 95th percentile of 2.55 ms. This term is negligible next to the windowing delay: because a decision can only be issued once a full 256 ms window is available, the first decision is inherently delayed by 256 ms, and subsequent decisions follow every 64 ms (the 25% step of a 256 ms window with 75% overlap). The serial transfer of each 15-byte frame over the native USB-CDC link adds a sub-millisecond term (one frame per millisecond at 1 kHz) that is likewise negligible. The 200 ms *smoothstep* pose interpolation is a visual animation applied after the decision and does not delay the classification.

The perceived latency is therefore dominated by the 256 ms window-filling delay rather than by computation, and the relevant figure for real-time control is this windowing term plus the few milliseconds of processing, within the 300 ms budget established for myoelectric prostheses by Farrell and Weir [7], although with a smaller margin than the 2.55 ms processing time alone would suggest.

TABLE II: End-to-End Latency Budget of the Real-Time Loop.

Stage	Contribution	Nature
Window filling (first decision)	256 ms	Structural
Inter-decision step (steady state)	64 ms	Structural
USB-CDC frame transfer	< 1 ms (est.)	Estimated
Feature extraction + inference (mean)	1.79 ms	Measured
Feature extraction + inference (p95)	2.55 ms	Measured
Pose interpolation (smoothstep)	200 ms (visual)	Animation

E. Digital-Twin Operation

The digital twin is the component that turns the pipeline into an interactive system, and it played two complementary roles in the evaluation. As a *paired visual-validation instrument*, fed in replay mode with restimulus-labeled DB5 windows from held-out subjects, its online accuracy matched the corresponding offline result within measurement margin, confirming that the streaming implementation faithfully reproduces the algorithmic behavior and ruling out buffer-synchronization or run-time feature-extraction bugs; it thereby complements the confusion matrices with a qualitative, temporal view of the classifier output. As an *online twin of the prosthetic hand*, fed by the four-channel USB-CDC stream, it anticipates the behavior expected once the trained model is embedded in a physical device, with transitions stabilized by the 200 ms smoothstep interpolation and the 90% confidence threshold.

Online operation with the proprietary hardware was evaluated qualitatively: because the target gestures were not logged synchronously, accuracy, macro-F1 and confusion matrices cannot be computed for this phase. Under visual inspection of the rendered pose, predictions oscillated among neighboring gestures (chiefly rest, pointing and open hand) more often than the offline cross-validation suggested, mainly because the exact electrode placement of the training session was not preserved between sessions: the same covariate-shift mechanism, now manifested within the same user across sessions. The confidence threshold and pose smoothing attenuated the visual flicker but did not remove the underlying degradation, reinforcing the need for a brief per-session recalibration. Beyond validation, this hardware-independent embodiment of the recognized gesture also provides a basis for user training and prosthesis-control prototyping before hardware integration, in line with virtual-prosthesis and extended-reality approaches reported in the literature [10], [13], [22].[2]

It should be made explicit what the digital twin mirrors in the present work: it reflects, in real time, the *intended gesture* inferred from the user's muscular activity, mapped onto the kinematic posture of a prosthetic-hand model. It does not yet model the mechanics or dynamics of a physical prosthesis (actuator response, contact forces or joint compliance) which would require a bidirectional coupling to an instrumented device. Establishing that physical, bidirectionally synchronized twin is left as future work; the current system is thus best understood as a real-time virtual hand serving as a validation and visualization instrument, and the user-training role is a prospect its architecture enables rather than a claim evaluated here.

[2]Demonstration videos of the system in real-time operation are available at: https://youtube.com/shorts/pkRm3E10lTw, https://youtube.com/shorts/_bxdrktQDjc and https://youtube.com/shorts/W6CeXGm7YlE.

F. Comparison with the Literature

Table III positions the results against representative sEMG gesture-classification works. Direct comparison requires caution, and not only for the usual differences in channels, dataset, number of gestures and protocol: many reported accuracies rely on randomly split windows, which, as shown above, inflate intra-subject scores when segments overlap. The repetition-grouped 0.69 macro-F1 reported here is therefore conservative by construction, and the more relevant message is the large intra-to-inter-subject gap, whose LOSO collapse reproduces the widely documented behavior of sEMG models without individual calibration [4], [8].

G. Limitations

Results must be read in light of the study's limitations. Pre-validation relied on NinaPro DB5, which differs from the target scenario in sampling rate (200 Hz vs. 1 kHz), number of channels (8 vs. 4), ADC resolution (8 vs. 12 bits) and sensor type; the reported numbers set an algorithmic ceiling, but effective performance still depends on acquisition quality. The proprietary collection comprises only two participants (both members of the research team) in a single session, without contractile-force control, professional anatomical assistance or preservation of electrode placement between sessions. The seven-class set also falls short of the ten or more grasps required by modern multifunction prostheses [19], and online performance was assessed only qualitatively.

H. Practical Implications and Clinical Applicability

Two findings carry the clearest practical weight. First, the intra-subject scores reported under repetition-grouped validation (≈ 0.69 on DB5) are a more realistic estimate of deployed behavior than the higher numbers common in the literature, and they set expectations accordingly: a low-cost four-channel platform is viable as an assistive and training tool, but not yet as a drop-in multifunction controller. Second, and more decisive for clinical use, the inter-subject collapse means a factory-trained, one-size-fits-all model is not usable; any deployment must include a short per-user (and, given the inter-session drift observed with the digital twin, per-session) calibration step, and should favor the more transfer-robust RF over SVM-RBF. For rehabilitation, the immediate value is less as a final prosthesis controller and more as an accessible platform for gesture-based training and biofeedback: the real-time virtual hand lets a user or therapist see the decoded intention without a physical device, and the open, reproducible hardware lowers the barrier for clinical research groups. Reaching routine clinical use would still require larger and more diverse cohorts, amputee participants, a richer grasp set, and the fast on-device calibration outlined above.

TABLE III: Comparison of the Results of this Work with the Literature.

Work	Channels	Gestures	Classifier	Result
Englehart & Hudgins [3]	4	4	LDA	$> 95\%$ acc.
Pizzolato *et al.* [19]	16	52	RF/SVM	70–92% acc. (DB5)
Dellacasa Bellingegni *et al.* [15]	6	7	SVM/LDA	30 amputees
Xue *et al.* [16]	–	10	RF	91.67% acc.
Yang *et al.* [21]	2	7	–	93% (real time)
Mendes Junior *et al.* [18]	8	6	SVM-RBF	$\approx$99% acc.
This work (DB5, intra, per subj.)	8	7	RF/SVM	0.69 macro-F1
This work (own hw., intra)	4	7	RF/SVM	0.53–0.60 macro-F1
This work (DB5, LOSO)	8	7	RF	0.37 macro-F1

V. Conclusion

This work proposed and implemented *sEMGest*, an integrated platform for hand gesture classification from sEMG comprising a four-channel acquisition front-end, a processing/classification pipeline and a real-time three-dimensional digital twin of a prosthetic hand. The hardware was experimentally characterized (25–540 Hz passband, 49.5 μV RMS input-referred noise). Under repetition-grouped cross-validation, RF and SVM-RBF tied intra-subject (macro-F1 ≈ 0.69 on NinaPro DB5, 0.53–0.60 on the proprietary hardware), with a processing latency of 2.55 ms at the 95th percentile and a perceived response time dominated by the 256 ms window-filling delay, within the 300 ms real-time budget. Inter-subject and inter-session evaluations revealed a collapse caused by covariate shift (with RF emerging as the most robust classifier across subjects) justifying a two-phase architecture with dedicated per-user models preceded by a calibration step. As a proof of concept with two participants and a qualitatively evaluated online stage, the results indicate the technical feasibility of the platform, with the digital twin serving as a hardware-independent instrument for pipeline validation and potential user training.

Future work includes enlarging the participant pool, developing a fast per-session calibration procedure and online incremental adaptation to mitigate covariate shift [27], [29], expanding the gesture set, integrating the classifier with a physical robotic hand, adding proportional force control [28], and instrumenting synchronous labeling to quantify online performance.

References

[1] S. L. Carey, D. J. Lura, and M. J. Highsmith, "Differences in myoelectric and body-powered upper-limb prostheses: Systematic literature review," *J. Rehabil. Res. Dev.*, vol. 52, no. 3, pp. 247–262, 2015.

[2] S. Salminger *et al.*, "Current rates of prosthetic usage in upper-limb amputees—have innovations had an impact on device acceptance?," *Disabil. Rehabil.*, vol. 44, no. 14, pp. 3708–3713, 2022.

[3] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 7, pp. 848–854, 2003.

[4] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *J. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 643–659, 2011.

[5] R. Merletti and D. Farina, *Surface Electromyography: Physiology, Engineering, and Applications*. Hoboken, NJ, USA: Wiley, 2016.

[6] C. J. De Luca, "The use of surface electromyography in biomechanics," *J. Appl. Biomech.*, vol. 13, no. 2, pp. 135–163, 1997.

[7] T. R. Farrell and R. F. Weir, "The optimal controller delay for myoelectric prostheses," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 1, pp. 111–118, 2007.

[8] A. Phinyomark *et al.*, "EMG feature evaluation for improving myoelectric pattern recognition robustness," *Expert Syst. Appl.*, vol. 40, no. 12, pp. 4832–4840, 2013.

[9] L. P. Murciego *et al.*, "Reducing the number of EMG electrodes during online hand gesture classification with changing wrist positions," *J. NeuroEng. Rehabil.*, vol. 19, no. 1, art. 78, 2022.

[10] K. A. Rodrigues *et al.*, "Embodiment of a virtual prosthesis through training using an EMG-based human-machine interface: Case series," *Front. Hum. Neurosci.*, vol. 16, art. 870103, 2022.

[11] K. Maksymenko *et al.*, "A myoelectric digital twin for fast and realistic modelling in deep learning," *Nat. Commun.*, vol. 14, art. 1600, 2023.

[12] Z. Johnson and M. J. Saikia, "Digital twins for healthcare using wearables," *Bioengineering*, vol. 11, no. 6, art. 606, 2024.

[13] W. Li *et al.*, "Current status and clinical perspectives of extended reality for myoelectric prostheses: review," *Front. Bioeng. Biotechnol.*, vol. 11, art. 1334771, 2024.

[14] D. C. Toledo-Pérez *et al.*, "Support vector machine-based EMG signal classification techniques: A review," *Appl. Sci.*, vol. 9, no. 20, art. 4402, 2019.

[15] A. Dellacasa Bellingegni *et al.*, "NLR, MLP, SVM, and LDA: a comparative analysis on EMG data from people with trans-radial amputation," *J. NeuroEng. Rehabil.*, vol. 14, no. 1, art. 82, 2017.

[16] Y. Xue *et al.*, "sEMG-based human in-hand motion recognition using nonlinear time series analysis and random forest," *IEEE Access*, vol. 7, pp. 176448–176457, 2019.

[17] A. Phinyomark and E. Scheme, "EMG pattern recognition in the era of big data and deep learning," *Big Data Cogn. Comput.*, vol. 2, no. 3, art. 21, 2018.

[18] J. J. A. Mendes Junior *et al.*, "How do sEMG segmentation parameters influence pattern recognition process? An approach based on wearable sEMG sensor," *Biomed. Signal Process. Control*, vol. 81, art. 104546, 2023.

[19] S. Pizzolato *et al.*, "Comparison of six electromyography acquisition setups on hand movement classification tasks," *PLoS ONE*, vol. 12, no. 10, art. e0186132, 2017.

[20] M. Atzori *et al.*, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data*, vol. 1, art. 140053, 2014.

[21] Z. Yang *et al.*, "Instinctive real-time sEMG-based control of prosthetic hand with reduced data acquisition and embedded deep learning training," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 5666–5672.

[22] A. Boschmann *et al.*, "Immersive augmented reality system for the training of pattern classification control with a myoelectric prosthesis," *J. NeuroEng. Rehabil.*, vol. 18, no. 1, art. 25, 2021.

[23] R. P. Sallen and E. L. Key, "A practical method of designing RC active filters," *IRE Trans. Circuit Theory*, vol. 2, no. 1, pp. 74–85, 1955.

[24] H. J. Hermens *et al.*, "Development of recommendations for SENIAM surface electromyography sensors and sensor placement procedures," *J. Electromyogr. Kinesiol.*, vol. 10, no. 5, pp. 361–374, 2000.

[25] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[26] A. Nynes *et al.*, "Asymmetry in contralateral muscle excitation in proximal vs. distal muscles in upper extremities," *Front. Psychol.*, vol. 16, 2025.

[27] X. Jiang, C. Ma, and K. Nazarpour, "Posture-invariant myoelectric control with self-calibrating random forests," *Front. Neurorobot.*, vol. 18, art. 1462023, 2024.
[28] F. Leone *et al.*, "Hierarchical strategy for sEMG classification of the hand/wrist gestures and forces of transradial amputees," *Front. Neurorobot.*, vol. 17, art. 1092006, 2023.
[29] M. A. Oskoei and H. Hu, "Myoelectric control systems—A survey," *Biomed. Signal Process. Control*, vol. 2, no. 4, pp. 275–294, 2007.

Simulation Analysis of a Quasi-LPV Controller Applied to Hybrid FES/Exoskeleton Rehabilitation Robotic Systems

1st Yecid Moreno
Dept. of Mechanical Engineering
EESC, University of São Paulo (USP)
São Carlos, Brazil
yecidmoreno@usp.br

2nd Marcos A. C. Alves
Dept. of Mechanical Engineering
EESC, University of São Paulo (USP)
São Carlos, Brazil
marcosaranon@usp.br

3rd Adriano A. G. Siqueira
Dept. of Mechanical Engineering
EESC, University of São Paulo (USP)
São Carlos, Brazil
siqueira@sc.usp.br

***Abstract*—This work investigates the efficacy of a control strategy to address the nonlinear and time-varying characteristics of hybrid FES/exoskeleton systems. The proposed approach relies on robust controller synthesis within a quasi-Linear Parameter Varying (quasi-LPV) framework, where the nonlinear dynamics are reformulated as a linear system whose state-space matrices vary as functions of state-dependent parameters. While similar control strategies have been successfully applied to underactuated manipulators, helicopters, and projectiles, their application to hybrid rehabilitation systems remains largely unexplored. Two control scenarios are analyzed: (i) when the varying parameter depends on both a muscle fatigue index and the musculoskeletal model; and (ii) when only the fatigue index evolves over time. Preliminary simulations on a human knee–exoskeleton model demonstrate that the proposed controllers regulate muscle activation via FES and robotic torque according to the muscle fatigue level, highlighting the potential of the approach for adaptive rehabilitation strategies.**

I. INTRODUCTION

Robotic rehabilitation involves a series of tasks focused on the motor recovery of individuals using robot-assisted devices [1]. Unlike traditional therapy, these robotic devices offer uniformity in repeating long-term routines, reliable recording of measured variables, and motivation for patient participation through interactive environments [2]. Under this premise, powered orthoses, robotic exoskeletons, and robotic platforms, in general, have been constantly improved due to the evolution of hardware and software technologies that address human-robot interaction challenges [3].

In these applications, ensuring safe interaction between humans and robots despite inter-subject variability is crucial. The uncertain nature of human dynamics complicates this task, compromising the stability of interaction control due to unpredictable and sudden movements of human joints [4] and [5].

The combination of robotic systems with muscle electrical stimulus, such as Functional Electrical Stimulation (FES), named as hybrid configuration, further complicates the interaction [4]. New parameters and variables, such as muscle fatigue and musculoskeletal model, add nonlinearity and time variability to the problem. To address these challenges, researchers have investigated the use of adaptive and nonlinear controllers. An adaptive hybrid controller was implemented for [6]. The *triadic collaboration* among a human, exoskeleton, and FES in gait rehabilitation was designed to overcome the challenges associated with muscle fatigue and the coordination of assistance provided by robotic devices and FES. The proposed controller automatically adapts the stimulation levels based on the patient's performance and muscle fatigue, prioritizing the patient's voluntary movements to encourage their active participation and enhance rehabilitation outcomes. In [4], iterative learning techniques are used to design a shared control of a hybrid powered exoskeleton and FES system, while in [7], the same research group presents a solution based on model predictive controller and dynamic control allocation.

In this paper, we investigate the feasibility of a novel control solution to deal with the nonlinearities and time-varying conditions of the hybrid FES/exoskeleton system. The proposed approach is based on robust controller synthesis for quasi-Linear Parameter Varying (quasi-LPV). The main idea is to represent the nonlinear system as a linear system where the matrices of the state-space model vary as function of given set of state-dependent parameters. This kind of control solution has already been applied to underactuated robotic manipulators [8], helicopters [9], and projectiles [10].

For our control problem, two cases are considered: in the first case, the varying parameter is defined as function of the muscle fatigue index and the musculoskeletal model; in the second one, only the fatigue index is assumed to vary over time. Preliminary results of the application of the proposed controllers in a knee human/exoskeleton simulated model show that the muscle activation associated with the FES and

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, under grant 304113/2025-5, FAPESP under grants 2025/07207-7 and 2024/01120-4 (Centro de Tecnologias Assistivas para as Atividades da Vida Diária - CTECVIDA), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, scholarship process no. 88887.817125/2023-00.

A. A. G. Siqueira is also with Center for Assistive Technologies for Activities of Daily Living - CTECVIDA and Center for Robotics of São Carlos

the robot torque are automatically adjusted according to the level of muscle fatigue.

This paper is organized as follows: Section II describes the modeling of the hybrid FES/exoskeleton system, including the fatigue index and the musculoskeletal model; Section III presents the robust state-feedback controller design for LPV systems; Section IV the quasi-LPV representation of the hybrid FES/exoskeleton model; Section V presents the simulated results and a discussion about the feasibility of the proposed solution; Finally, conclusions are given in Section VI.

II. MODELING OF THE HYBRID FES/EXOSKELETON SYSTEM

The dynamic model of the hybrid FES/exoskeleton system is based on the right knee actuator of the exoskeleton Exo-TAO shown in Fig. 1. The rotary series elastic actuator (SEA), described in [11], can provide a torque-controlled actuation, and has been used to evaluate different robust control solutions, including model predictive [12] and Markovian [13] controllers.

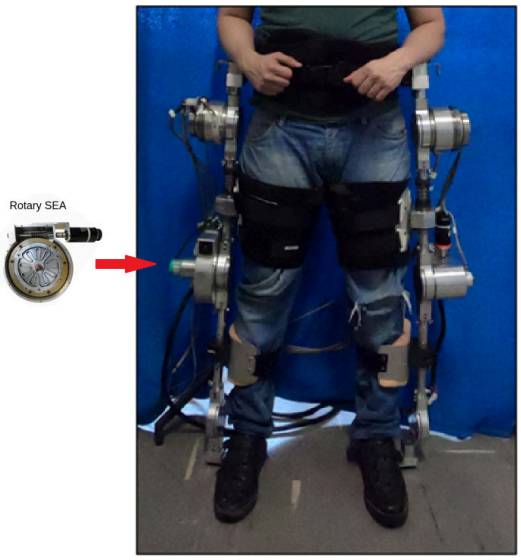


Fig. 1. Exoskeleton Exo-TAO, with the knee actuator.

Consider the following dynamic model:

$$M\ddot{q} + C\dot{q} + G(q) = \tau_r + \tau_f, \tag{1}$$

where q is the knee angle, M is the moment of inertia, $C\dot{q}$ is the torque generated by viscous damping, $G(q)$ is the torque due to gravity, and τ_r is the torque generated by the robot. The torque generated by the musculoskeletal system due to the application of FES, τ_f, is given by:

$$\tau_f = \sigma(q, \dot{q})\mu u_f,$$

where $\sigma(q, \dot{q})$ is the function corresponding to the musculoskeletal model, $u_f \in [0, 1]$ is the muscle activation induced by FES, and $\mu \in [\mu_{min}, 1]$ is the fatigue index that models muscle fatigue.

For the musculoskeletal model, we are considering a simplified muscle contraction model to compute the muscles forces, with active components depending only on muscle length and velocity, that is:

$$\sigma(q, \dot{q}) = f_A(\tilde{l}^m) f_V(\dot{\tilde{l}}^m) F_0^m cos(\phi) r^m,$$

where $f_A(\tilde{l}^m)$ is the normalized active force-length function, $f_V(\dot{\tilde{l}}^m)$ is the normalized force-velocity function, F_0^m is the maximum isometric force at optimal muscle fiber length l_o^m, r^m is the moment arm, $\tilde{l}^m$ is the normalized muscle fiber length, $\tilde{l}^m = \frac{l^m}{l_0^m}$, where l^m is the muscle fiber length, where ϕ is the angle between the orientation of muscle and tendon fibers, named pennation angle, which can be compute by:

$$\phi = arcsin\left(\frac{l_0^m sin\phi_0}{l^m}\right), \tag{2}$$

where ϕ_0 is the pennation angle at optimal fiber length. All musculoskeletal model's functions ($f_A(.)$ and $f_V(.)$) and parameters (F_0^m, l_0^m, and ϕ_0), for given values of q and $\dot{q}$, were obtained using the software OpenSim (version 4.4).

A. Fatigue Modeling and Estimation

The fatigue index is a mathematical representation employed to characterize the state of muscle fatigue during FES application [14]. The primary objective of this modeling is to assess the relationship between the level of muscle activation and the recovery process over time, under different intensities of electrical stimulation current. The dynamics of muscle fatigue are represented by the variable μ, which denotes the functional capacity of the muscle. When $\mu = 1$, the muscle operates under ideal conditions, without fatigue. Conversely, μ_{min} corresponds to the minimum level of muscle performance, below which functionality is substantially impaired. The temporal evolution of μ is described by the following differential equation:

$$\dot{\mu} = \frac{(\mu_{min} - \mu)u_f}{T_f} + \frac{(1 - \mu)(1 - u_f)}{T_r}, \tag{3}$$

where T_f is the time constant associated with muscle fatigue, determining the rate of decline of μ under continuous activation, and T_r is the recovery time constant, which governs the rate of muscle recovery when activation is suspended.

The fatigue index's parameters, T_f and T_r, can be obtained by the experimental methodology proposed by [14]. For the knee joint case, an experiment was conducted as shown in Fig. 2, where the setup includes a chair equipped with a force sensor attached to the subject's leg to measure the generated force. The stimulation signal comprises two parts: the first part is a constant stimulation lasting 180 seconds, followed by an intermittent stimulation with cycles of 10s of rest and 1s of activity. This pattern is repeated for an additional 180s, resulting in a total duration of 360s. This sequence aims to assess the effects of fatigue and subsequent muscle recovery. The protocol was approved by the University of São Paulo Ethics Committee (EEFERP-USP) under protocol No. 411150620.70000.5659 and decision statement

No. 4.579.836. Written informed consent was obtained from the participant.

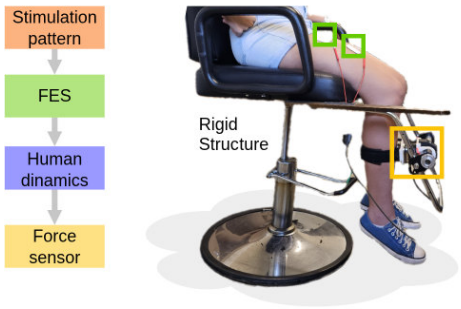


Fig. 2. Setup for muscle fatigue identification.

Initially, the force generated during the subject's maximum voluntary contraction is recorded. Immediately following this, the stimulation sequence is applied. This procedure is repeated three times across four different days to ensure the consistency and reliability of the collected data.

Figure 3 shows the behavior of muscle capacity to generate force in response to constant and discrete stimulation. At the top is the pulse width $P_w(t)$, and below, in gray, is the generated muscle force response. On the other hand, the fatigue index curve $\mu(t)$, is divided into two sections, the green line represents the decay of muscle fatigue caused by constant stimulation ($t < 180$s), and the orange line represents muscle recovery. The ranges of $\mu_(t)$ and T_h were normalized with the maximum voluntary force.

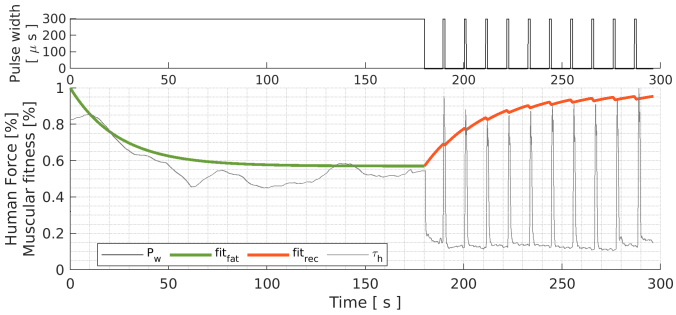


Fig. 3. Behavior of muscle capacity to generate force in response to constant and discrete stimulation.

Table I show the time constant associated with muscle fatigue, T_f and the recovery time constant, T_r, obtained considering the fatigue model, Eq. (3), and the experimental data.

TABLE I
TIME CONSTANTS FOR MUSCLE FATIGUE (T_f) AND RECOVERY (T_r)

Time constant	Value[s]
T_f	32.1
T_r	54.7

III. STATE FEEDBACK $\mathcal{H}_\infty$ CONTROL FOR QUASI-LPV SYSTEMS

Consider the following quasi-LPV model:

$$\begin{bmatrix} \dot{x} \\ z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} A(\rho(x)) & B_1(\rho(x)) & B_2(\rho(x)) \\ C_1(\rho(x)) & 0 & 0 \\ C_2(\rho(x)) & 0 & I \end{bmatrix} \begin{bmatrix} x \\ w \\ u \end{bmatrix}, \tag{4}$$

where x is the state vector, u is the control input vector, w is the external input vector, $z = [z_1\ z_2]^T$ is the output vector, and $\rho(x)$ is the parameter varying vector. Assume that the underlying parameter $\rho(x)$ varies in the allowable set:

$$F_P^\nu = \left\{\rho(x) \in \mathcal{C}^1(\Re^n, \Re^m) : \rho(x) \in P,\ |\dot{\rho}_i| \leq \nu_i\right\}, \tag{5}$$

for $i = 1, \ldots, k$, where $P \subset \Re^m$ is a compact set, $\nu = [\nu_1 \cdots \nu_m]^T$ with $\nu_i \geq 0$ and $\underline{\nu}_i(\rho) \leq \dot{\rho}_i \leq \overline{\nu}_i(\rho)$, $i = 1, \ldots, m$. We denote by $\mathcal{C}^1(\Re^n, \Re^m)$ the set of continuously differentiable functions that map $\Re^n$ to $\Re^m$. The state feedback control problem considered here aims at finding a continuous function $F(\rho)$ such that the closed loop system has an $\mathcal{L}_2$ gain less than or equal to γ, that is, $\int_0^T \|z(t)\|^2\, dt \leq \gamma^2 \int_0^T \|w(t)\|^2\, dt$, under a state feedback law $u = F(\rho)x$.

According to [15], if there exists a continuously differentiable matrix function $X(\rho(x)) > 0$ that satisfies:

$$\begin{bmatrix} E(\rho(x)) & X(\rho(x))C_1^T(\rho(x)) & B_1(\rho(x)) \\ C_1(\rho(x))X(\rho(x)) & -I & 0 \\ B_1^T(\rho(x)) & 0 & -\gamma^2 I \end{bmatrix} < 0, \tag{6}$$

where

$$E(\rho(x)) = -\sum_{i=1}^{m} \overline{\nu}_i(\rho) \frac{\partial X(\rho(x))}{\partial \rho_i} - B_2(\rho(x))B_2^T(\rho(x)) + \widehat{A}(\rho(x))X(\rho(x)) + X(\rho(x))\widehat{A}(\rho(x))^T$$

and $\widehat{A}(\rho(x)) = A(\rho(x)) - B_2(\rho(x))C_2(\rho(x))$, then the closed loop system has $\mathcal{L}_2$ gain $\leq \gamma$ under the state feedback control law, $u(t) = -K(\rho(x))x(t)$, where:

$$K(\rho(x)) = B_2^T(\rho(x))X^{-1}(\rho(x)) + C_2(\rho(x)). \tag{7}$$

A practical scheme was developed in [15] to compute these linear matrix inequalities based on basis functions related to the $X(\rho(x))$ and on the grid of the parameter set P. First we need to choose a set of $\mathcal{C}^1$ basis functions $\{\psi_i(\rho)\}_{i=1}^M$ for $X(\rho(x))$ in order to rewrite it as:

$$X(\rho(x)) = \sum_{i=1}^{M} \psi_i(\rho(x))X_i, \tag{8}$$

where $X_i \in S^{n\times n}$ is the coefficient matrix for $\psi_i(\rho(x))$. To solve this infinite dimensional optimization problem, we can grid the parameter set P along L points $\{\rho_k\}_{k=1}^L$ in each dimension. Since Equation (6) consists of 2^m entries, a total of $(2^m + 1)L^m$ matrix inequalities in terms of the matrices $\{X_i\}$ should be solved.

IV. QUASI-LPV REPRESENTATION OF THE HYBRID FES/EXOSKELETON MODEL

Considering the hybrid FES/exoskeleton model given by Eq. (1), and a given desired trajectory, q^d and its derivatives, the following state-space representation can be obtained:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{C}{M} \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix} w + \begin{bmatrix} 0 & 0 \\ \frac{1}{M} & \frac{\sigma(q,\dot{q})\mu}{M} \end{bmatrix} u \quad (9)$$

where $x = [q^d - q, \dot{q}^d - \dot{q}]^T$, $u = [-\tau_r + G(q), -u_f]^T$, and $w = M\ddot{q}^d + C\dot{q}^d$. Here, we are assuming the desired trajectory and its derivatives are bounded, so the disturbance w is also bounded.

Considering that the controller's gains are obtained from Eq. (7) as:

$$K(\rho(x)) = \begin{bmatrix} K_r & B_r \\ K_f & B_f \end{bmatrix},$$

the robotic control input is defined as:

$$\tau_r = K_r(q_d - q) + B_r(\dot{q}^d - \dot{q}) + G(q).$$

The FES control input is given by:

$$u_f = K_f(q_d - q) + B_f(\dot{q}^d - \dot{q}).$$

For the application of the quasi-lpv control described in Section III, two possible definitions for the variant parameter, $\rho(x)$ are considered, as in the cases below.

A. Case 1 - $\rho(x) = \sigma(q, \dot{q})\mu$

First, the varying parameter is defined as the combination of the musculoskeletal model and the fatigue index, that is, $\rho(x) = \sigma(q, \dot{q})\mu$, and Eq. (9) can be rewritten as:

$$\dot{x} = \underbrace{\begin{bmatrix} 0 & 1 \\ 0 & -\frac{C}{M} \end{bmatrix}}_{A} x + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix}}_{B_1} w + \underbrace{\begin{bmatrix} 0 & 0 \\ \frac{1}{M} & \frac{\rho(x)}{M} \end{bmatrix}}_{B_2(\rho(x))} u. \quad (10)$$

We also consider the following matrices to complete the quasi-LPV model, Eq. (4): $C_1 = I_2$, and $C_2 = 0_2$.

B. Case 2 - $\rho(t) = \mu$

In this case, we assume that $\sigma(q, \dot{q}) = \bar{\sigma}$, where $\bar{\sigma}$ represents the average value of $\sigma(q, \dot{q})$ along the knee extension/flexion movement. In this case, the varying parameter is defined only as the fatigue index, i.e., $\rho(t) = \mu$, and Eq. (9) can be rewritten as:

$$\dot{x} = \underbrace{\begin{bmatrix} 0 & 1 \\ 0 & -\frac{C}{M} \end{bmatrix}}_{A} x + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix}}_{B_1} w + \underbrace{\begin{bmatrix} 0 & 0 \\ \frac{1}{M} & \frac{\bar{\sigma}\rho(t)}{M} \end{bmatrix}}_{B_2(\rho(t))} u. \quad (11)$$

Actually, in this case, the system is considered as a LPV one, since the varying parameter depends only on the time. The controller's design procedure presented in Section III can also be applied to LPV systems.

V. RESULTS

This section presents the simulated results of the application of the proposed controllers to the knee actuator, considering the human is in a seated position, as shown in Fig. 4. The simulation was performed using Matlab, with data (mass, inertia, lengths, muscle parameters, etc.) obtained from the OpenSim software for a person with 75 kg.

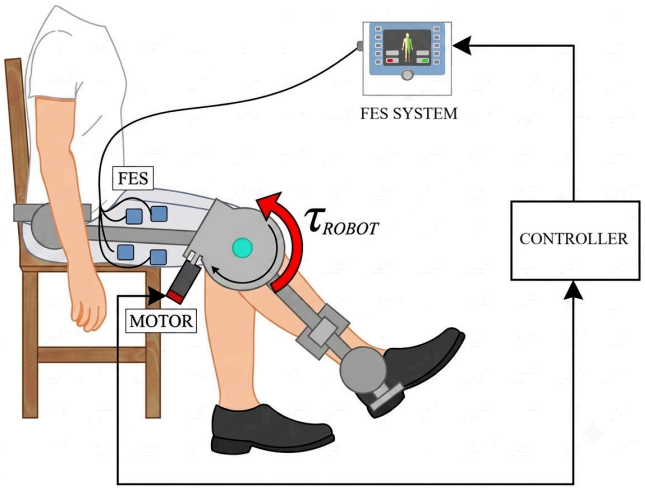


Fig. 4. Human/knee-exoskeleton system.

A. Case 1

For Case 1, the varying parameter is defined as the combination fo the musculoskeletal model and the fatigue index, that is, $\rho(x) = \sigma(q, \dot{q})\mu$. As shown in Section III, we need to grid the parameter set P along L points. Considering that the fatigue index, μ, can assume values in the set $\mu \in [0.2; 1]$, we defined 9 points to grid this parameter, that is, $\mu \in 0.2, 0.3, \cdots, 0.9, 1$. For the musculoskeletal model, $\sigma(q, \dot{q})$, we considered only the extension's muscular group (Rectus Femoris, Vastus Medialis, Vastus Lateralis). According to data provided by the OpenSim for an extension/flexion movement, $\sigma(q, \dot{q})$ varies from 10 to 40. For this function, we defined 7 points to grid this parameter that is, $\sigma(q, \dot{q}) \in 10, 15, \cdots, 35, 40$. Overall, it was defined 63 points to represent the variation of parameter $\rho(x)$. The basis functions were defined as $\psi_1 = 1$ and $\psi_2 = \sigma(q, \dot{q})\mu$.

B. Case 2

For Case 2, we defined $\bar{\sigma} = 25$, representing the average value of $\sigma(q, \dot{q})$ along the knee extension/flexion movement. Since, in this case, $\rho(x) = \mu$, we again defined 9 points to grid this parameter, that is, $\mu \in 0.2, 0.3, \cdots, 0.9, 1$. The basis functions were now defined as $\psi_1 = 1$ and $\psi_2 = \mu$.

C. Simulation Conditions

The desired trajectory to be followed by the human/exoskeleton system is a sinusoidal function with peak-to-peak amplitude of 90°, representing the flexion and extension of the knee, with a frequency of $0.5Hz$, that is, the extension/flexion movement takes 2 seconds.

To highlight the effect of the quasi-LPV controller for different conditions of muscle fatigue, the muscle recovery was not considered in the simulation, that is, the recover constant, T_r, was assumed to have a high value. Additionally, the time constant associated with muscle fatigue was assumed to be small, T_f = 0.2 s. This constitutes an artificially imposed stress scenario that does not accurately reflect physiological conditions. Consequently, the fatigue indicator, μ, decreases rapidly from 1 to approximately 0.5. Figure 5 shows the values of the fatigue index, μ, for Cases 1 and 2.

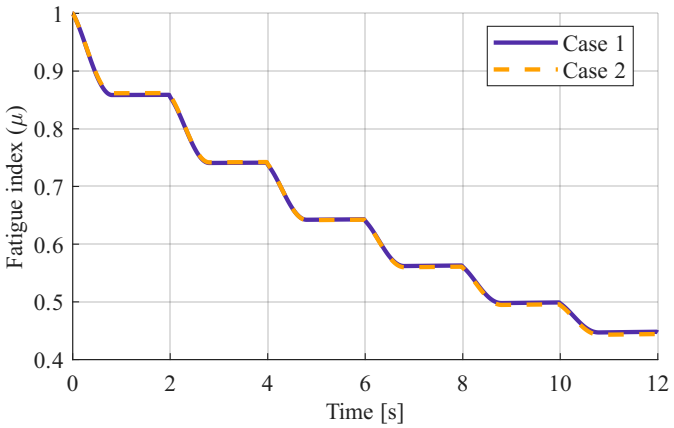


Fig. 5. Fatigue index, μ, for Cases 1 and 2.

Figures 6 to 8 show, respectively, the muscle activation, u_f, the torque applied by the exoskeleton (not considering the gravity compensation), τ_r, the torque generated by the muscle due to the FES stimulation, τ_m, and the controller's gains for Case 1.

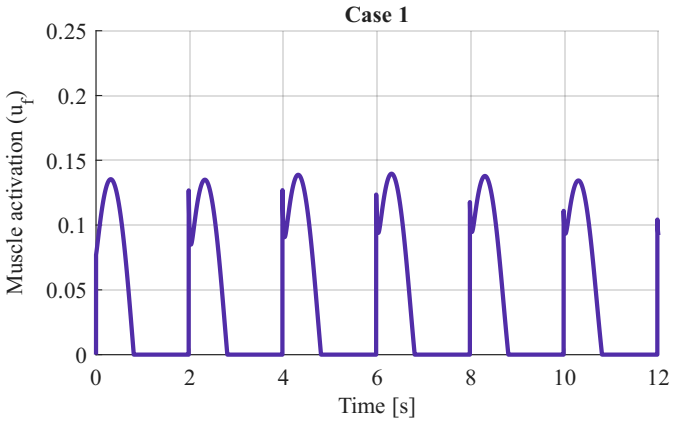


Fig. 6. Muscle activation, u_f, for Case 1.

Note from Fig. 6 that the proposed controller slightly increases the level of muscle activation at the first phases of muscle fatigue increase (decrease of the fatigue index), trying to maintain the muscle torque even with muscle fatigue. At the same phase, the robot torque also slightly increases. However, at the later phases, where the muscle fatigue is higher, the controller strongly and automatically increases

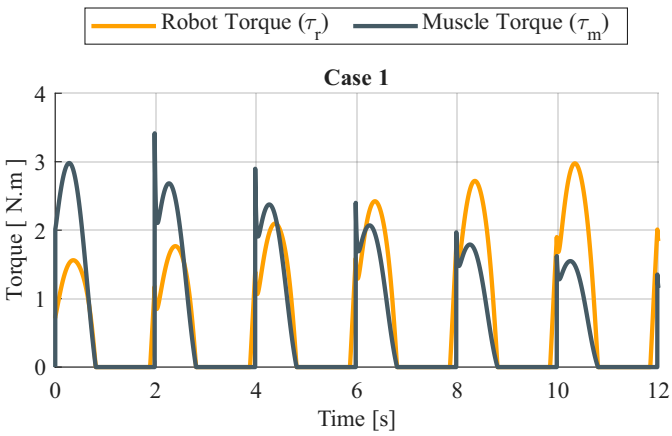


Fig. 7. Torques, τ_r and τ_m, for Case 1.

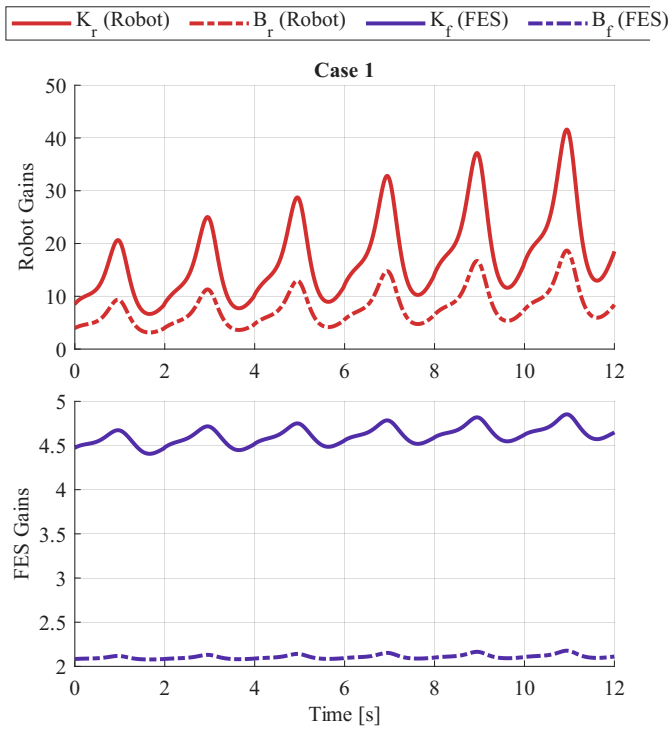


Fig. 8. Controller's gains for Case 1.

the robot actuation, decreasing the muscle activation. These changes of level of muscle activation and robot actuation, reflecting on the muscle and robot torques, can be seen at Fig. 7.

Fig. 8 shows the changes of the values of the gains within the phases and within a single extension movement. As seen from the figure, the gains change during the extension movement, mainly due the change in the musculoskeletal model, $\sigma(q, \dot{q})$.

Figures 9 to 11 show, respectively, the muscle activation, u_f, the torque applied by the exoskeleton (not considering the gravity compensation), τ_r, the torque generated by the muscle due to the FES stimulation, τ_m, and the controller's gains for

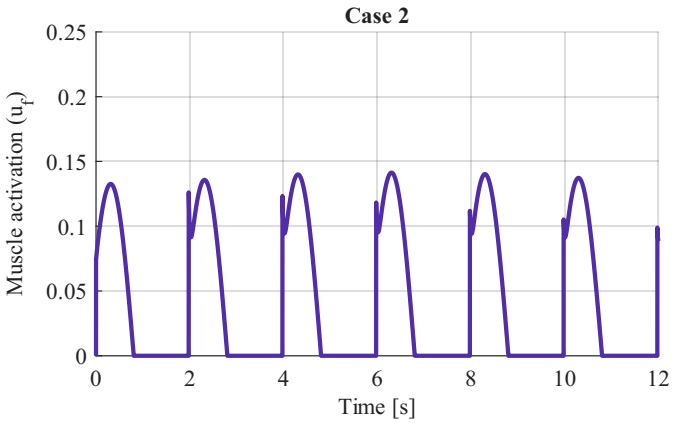


Fig. 9. Muscle activation, u_f - Case 2.

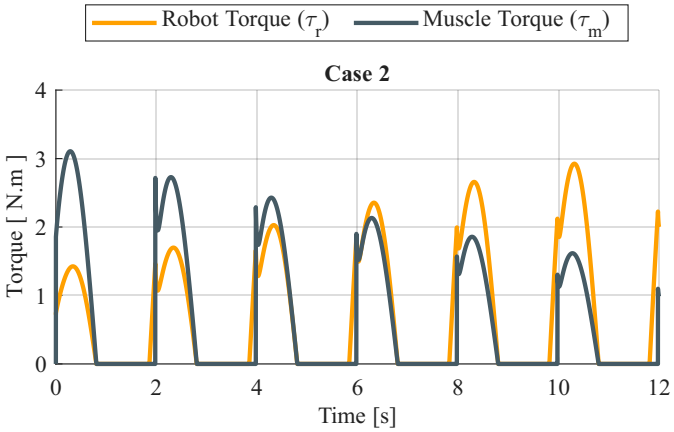


Fig. 10. Torques, τ_r and τ_m - Case 2.

Case 2.

The main difference between Cases 1 and 2 is concentrated in the variation of the gain values during a single extension movement. In case 2, the gains also change within the phases, but are kept constant during a single extension movement, as show in Fig. 11 .

VI. CONCLUSIONS

This work evaluated a quasi-LPV robust control approach to address the nonlinearities and time-varying dynamics of hybrid FES/exoskeleton systems. Two cases were investigated: one considering both the muscle fatigue index (fatigue index) and the musculoskeletal model as varying parameters, and another relying solely on the fatigue index. Simulations with a knee human–exoskeleton model demonstrated that the proposed controllers automatically adjusted FES-induced muscle activation and robotic torque according to fatigue levels. The main distinction observed between the two cases was in the gain variation during a single extension movement: while case 1 exhibited variations, case 2 maintained gains nearly constant. Future work will focus on experimental validation and extending the approach to multi-joint systems.

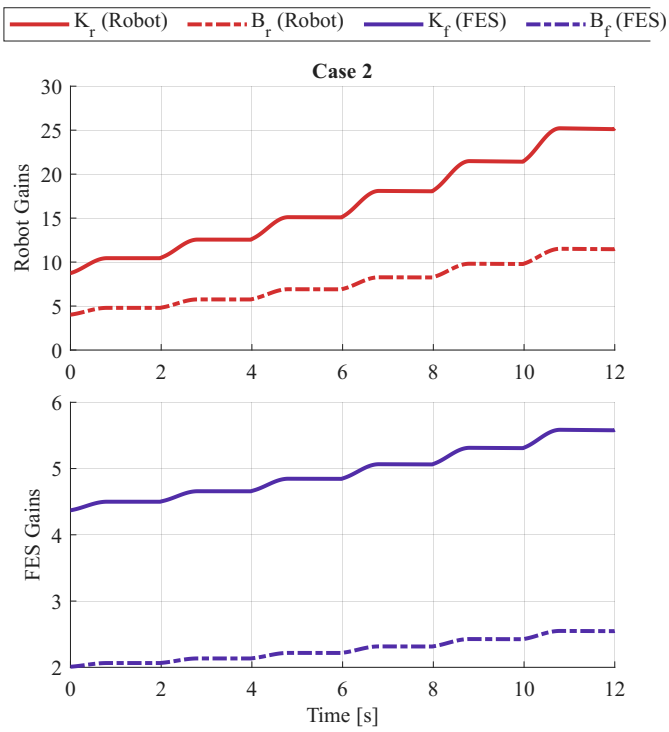


Fig. 11. Controller's gains - Case 2.

REFERENCES

[1] A. Plaza, M. Hernandez, G. Puyuelo, E. Garces, and E. Garcia, "Lower-limb medical and rehabilitation exoskeletons: A review of the current designs," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 278–291, 2023.

[2] A. Akbari, F. Haghverd, and S. Behbahani, "Robotic home-based rehabilitation systems design: From a literature review to a conceptual framework for community-based remote therapy during covid-19 pandemic," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2021.612331

[3] A. Saboor, T. Kask, A. Kuusik, M. M. Alam, Y. Le Moullec, I. K. Niazi, A. Zoha, and R. Ahmad, "Latest research trends in gait analysis using wearable sensors and machine learning: A systematic review," *IEEE Access*, vol. 8, pp. 167 830–167 864, 2020.

[4] V. Molazadeh, Q. Zhang, X. Bao, B. E. Dicianno, and N. Sharma, "Shared control of a powered exoskeleton and functional electrical stimulation using iterative learning," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2021.711388

[5] H. Börner, S. Endo, and S. Hirche, "Estimation of involuntary components of human arm impedance in multi-joint movements via feedback jerk isolation," *Frontiers in Neuroscience*, vol. 14, 2020.

[6] A. Christou, A. J. del Ama, J. C. Moreno, and S. Vijayakumar, "Adaptive control for triadic human-robot-fes collaboration in gait rehabilitation: A pilot study," 2024. [Online]. Available: https://arxiv.org/abs/2402.00775

[7] N. A. Kirsch, X. Bao, N. A. Alibeji, B. E. Dicianno, and N. Sharma, "Model-based dynamic control allocation in a hybrid neuroprosthesis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 224–232, 2018.

[8] A. Siqueira and M. Terra, "Nonlinear and markovian $\mathcal{H}_\infty$ controls of underactuated manipulators," *Control Systems Technology, IEEE Transactions on*, vol. 12, pp. 811 – 826, 12 2004.

[9] J. Y. Yassuda, C. M. Agulhari, and E. R. P. da Silva, "Sampled-data robust control of a 2-dof helicopter modeled using a quasi-lpv framework," *Control Engineering Practice*, vol. 145, p. 105870, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967066124000303

[10] G. Vinco, S. Theodoulis, O. Sename, and G. Strub, "Quasi-lpv modeling of guided projectile pitch dynamics through state transformation technique," *IFAC-PapersOnLine*, vol. 55, no. 35, pp. 43–48, 2022, 5th IFAC Workshop on Linear Parameter Varying Systems LPVS 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896322029317

[11] W. M. dos Santos, G. A. P. Caurin, and A. A. G. Siqueira, "Design and control of an active knee orthosis driven by a rotary series elastic actuator," *Control Engineering Practice*, vol. 58, pp. 307–318, 2017.

[12] W. M. dos Santos and A. A. Siqueira, "Optimal impedance via model predictive control for robot-aided rehabilitation," *Control Engineering Practice*, vol. 93, p. 104177, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967066119301674

[13] F. M. Escalante, J. C. Pérez-Ibarra, J. C. Jaimes, A. A. Siqueira, and M. H. Terra, "Robust markovian impedance control applied to a modular knee-exoskeleton," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10 141–10 147, 2020, 21st IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896320335035

[14] R. Riener, J. Quintern, and G. Schmidt, "Biomechanical model of the human knee evaluated by neuromuscular stimulation," *Journal of Biomechanics*, vol. 29, no. 9, pp. 1157–1167, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0021929096000127

[15] F. Wu, X. H. Yang, A. Packard, and G. Becker, "Induced $\mathcal{L}_2$-norm control for lpv systems with bounded parameter variation rates," *International Journal of Robust and Nonlinear Control*, vol. 6, pp. 983–998, 1996.

Sinais faciais embarcados em smartphones: reprodutibilidade entre SoCs e desempenho na detecção de nictação e oscitação

P. Neri de Oliveira[1,2] e F. J. Fraga da Silva[2]
[1]*Waker Inova Simples IS*, Santo André, Brasil
[2]*Universidade Federal do ABC (UFABC)*, Santo André, Brasil

***Resumo*—Movimentos oculares e da boca são manifestações observáveis associadas a estados como fadiga e sonolência, cuja aquisição convencional exige instrumentação dedicada. Este trabalho avalia a extração de sinais faciais integralmente executada em smartphones, sem armazenamento de imagem e sem conexão de rede, sobre três dispositivos Android com sistemas em um chip (SoCs) de arquiteturas distintas (Snapdragon 730, Exynos 1380 e Exynos 2400e). Como caso de teste, um detector de nictação, fechamento ocular prolongado e oscitação, baseado em razões de aspecto geométricas sobre malha facial de 468 pontos, foi avaliado nos 29 vídeos de câmera veicular do conjunto público YawDD, contra anotação temporal consensual de 1002 eventos. Os três dispositivos produziram detecções idênticas ao evento (índice de Jaccard 1,000). A 10 quadros/s, taxa de operação do sistema, obteve-se F1 de 0,830 para nictação (sensibilidade 0,748; valor preditivo positivo 0,932), 0,634 para oscitação (sensibilidade 0,833) e 0,584 para fechamento prolongado. A sensibilidade de oscitação supera os 60 % reportados pelos autores do YawDD nos mesmos vídeos. Um desenho fatorial 2×2 mostrou que a resolução de entrada não afeta a latência e que a vantagem entre GPU e CPU depende do SoC; a diferença entre desktop e dispositivo foi decomposta em taxa de amostragem, causalidade e plataforma, esta última respondendo por 0,035 de F1. Os resultados constituem prova de conceito técnica, sem alegação de validade clínica.**

***Palavras-chave*—sinais faciais; visão computacional; processamento embarcado; smartphone; nictação; oscitação.**

I. Introdução

Sinais faciais como a frequência e a duração de piscadas, o fechamento ocular sustentado, base da métrica PERCLOS [10], e o bocejo são utilizados como indicadores comportamentais de fadiga e sonolência em contextos como o monitoramento de condutores [4], [8]. Sua aquisição em campo depende, em geral, de câmeras dedicadas e processamento externo, o que limita a portabilidade e levanta questões de privacidade quando a imagem do rosto precisa sair do local de captura.

Smartphones reúnem câmera, processador e acelerador gráfico num único dispositivo, e modelos de malha facial otimizados para GPU móvel [1] tornam viável extrair esses sinais localmente. A questão investigada não é se um detector de piscada funciona, o que a literatura estabeleceu há uma década [2], [7], mas se a extração completa pode ser executada no aparelho, de forma reprodutível entre arquiteturas distintas, com custo compatível com operação contínua e desempenho mensurável contra referência humana.

Neste texto, *sinal facial* designa uma medida derivada da geometria do rosto, como a razão de aspecto do olho ou da boca, tratada como indicador comportamental. Nenhuma alegação de validade clínica é feita; a associação entre esses sinais e estados fisiológicos é assumida da literatura e não é avaliada aqui.

As contribuições são: (i) a demonstração de que três sistemas em um chip (SoCs) de fabricantes e famílias de GPU diferentes produzem detecções idênticas ao evento sobre os mesmos quadros; (ii) um desenho fatorial que separa o efeito do acelerador do efeito da resolução de entrada; (iii) a decomposição da diferença de desempenho entre implementação de desktop e embarcada em três parcelas mensuráveis; (iv) uma avaliação em nível de evento contra anotação temporal consensual, com o protocolo, as exclusões e os modos de falha declarados.

II. Materiais e Métodos

A. Arquitetura

O pipeline (Fig. 1) recebe quadros da câmera frontal, obtém a malha facial de 468 pontos e os coeficientes de expressão do MediaPipe Face Landmarker [1], [9], calcula razões de aspecto geométricas e as submete a máquinas de estado que emitem eventos discretos. Todo o processamento ocorre no dispositivo; nenhum quadro é armazenado ou transmitido. O sistema em produção opera a 10 quadros/s sobre entrada de 320×240 pixels. Ele passou por iterações desde uma versão inicial (v1.24, maio de 2025), cujo detector foi avaliado internamente apenas por contagem de eventos por vídeo, sem instantes; a versão corrente (v40) é a que fornece o detector de nictação comparado ao de referência na Seção III-F. Nenhuma dessas avaliações anteriores foi publicada; os dados que as sustentam integram o material depositado.

Para esta avaliação, o mesmo modelo de malha facial e as mesmas regras de detecção foram implementados em duas plataformas: (a) uma referência em Python sobre computador (macOS, Apple M2), que processa os vídeos a 30 quadros/s com acesso ao vídeo inteiro; (b) uma implementação Kotlin embarcada nos três dispositivos Android, que processa os mesmos vídeos a 10 quadros/s de forma causal, como faria com a câmera. A paridade das regras entre as duas implementações foi verificada parâmetro a parâmetro. A comparação entre elas

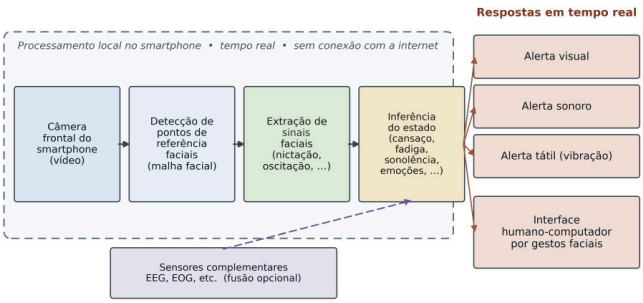


Figura 1. Arquitetura do sistema. Os três primeiros blocos, até a extração dos sinais faciais, são objeto deste trabalho. A inferência de estado e as respostas não são avaliadas. Sensores complementares indicam trabalho futuro.

isola, portanto, a plataforma e a estratégia de amostragem, e não o algoritmo.

O sistema também possui uma etapa de inferência de estado (fadiga, sonolência) que combina os eventos com regras temporais. Essa etapa *não* é avaliada neste trabalho: o YawDD não contém anotação de sonolência, e qualquer número nesse sentido seria inventado.

B. Extração dos sinais

A razão de aspecto do olho (EAR) segue Soukupová e Čech [2], com seis pontos por olho e média entre os dois olhos (Fig. 2):

$$\mathrm{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\,\|p_1 - p_4\|}, \tag{1}$$

sobre coordenadas convertidas a pixels antes do cálculo das distâncias, para evitar distorção pela razão de aspecto não quadrada do quadro. Na Fig. 2, as normas $\|p_2 - p_6\|$ e $\|p_3 - p_5\|$ aparecem expandidas como distâncias euclidianas d_1 e d_2, $\|p_1 - p_4\|$ como a largura w, e a razão é denotada AR; o cálculo é o mesmo, feito para cada olho e, posteriormente, calculou-se a média. A razão de aspecto da boca (MAR) segue a mesma construção com três distâncias verticais v_1, v_2, v_3 do contorno interno dos lábios sobre duas vezes a largura w_M:

$$\mathrm{MAR} = \frac{v_1 + v_2 + v_3}{2\,w_M}. \tag{2}$$

A pose da cabeça (yaw, pitch, roll) é obtida da matriz de transformação facial do modelo.

Os sinais são normalizados por um *baseline* (linha de base) individual, porque a EAR de olho aberto variou de 0,259 a 0,417 entre os 29 participantes, razão de 1,6, o que inviabiliza limiar absoluto. Na implementação de desktop, o baseline é o percentil 75 da EAR sobre o vídeo inteiro. Na implementação embarcada, causal, o detector inicia com um valor *a priori* populacional (0,29, mediana dos baselines individuais do conjunto de referência) e o substitui pelo percentil 75 da janela crescente de quadros válidos e frontais a partir do décimo quadro. Para a boca, o baseline é a mediana da janela crescente a partir do trigésimo quadro, sem valor a priori, porque o bocejo é raro e longo e um valor a priori errado produziu falsos positivos em teste preliminar.

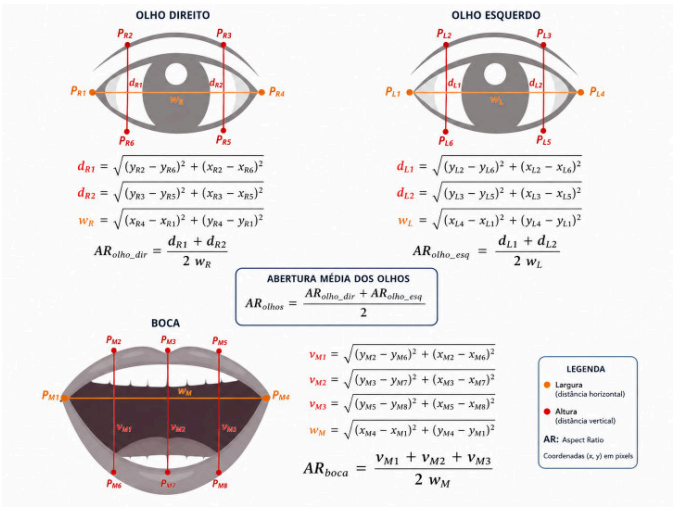


Figura 2. Pontos de referência e razões de aspecto. Olhos: seis pontos por olho, duas distâncias verticais sobre duas vezes a largura, com média entre os dois olhos. Boca: três distâncias verticais sobre duas vezes a largura. Coordenadas em pixels.

C. Regras de detecção

Um evento ocular inicia quando a EAR normalizada cai abaixo de 0,70 e termina quando volta acima por mais de 60 ms. A duração é estimada como o intervalo entre o primeiro e o último quadro abaixo do limiar mais um intervalo de amostragem, para que um evento capturado em um único quadro receba a duração de um intervalo em vez de zero. Eventos de até 500 ms são nictação se a EAR mínima atingiu 0,60 do baseline; eventos mais longos são fechamento prolongado se a mediana da EAR durante o evento ficou abaixo de 0,50, e fechamento parcial caso contrário. Um período refratário de 100 ms separa nictações consecutivas. O bocejo inicia quando a MAR normalizada excede 2,2, com histerese de saída em 1,6, e é contabilizado como oscitação se sustentado por ao menos 1000 ms. Os limiares de 0,70 e 2,2 foram fixados antes da anotação de referência existir; a varredura posterior (Seção III-C) confirmou 0,70 como ótimo para nictação.

Os limiares de profundidade (0,60 para nictação, 0,50 para prolongado) foram introduzidos após análise dos falsos positivos em rodada preliminar no dispositivo: os falsos positivos tinham EAR mínima entre 0,59 e 0,65, contra 0,29 a 0,33 nos verdadeiros, e o critério de anotação exige cobertura de mais de metade da pupila. A confirmação por coeficiente de expressão (*blendshape*) foi testada como alternativa e como complemento em ambos os casos e não superou a profundidade geométrica; o resultado negativo é reportado na Seção III-G.

D. Parâmetros de operação

A Tabela I consolida os parâmetros das Seções II-B e II-C. Todos são idênticos nas duas implementações e foram verificados um a um antes das rodadas; qualquer divergência entre desktop e dispositivo é, portanto, atribuível à plataforma ou à estratégia de amostragem, e não a diferença de configuração.

Tabela I
PARÂMETROS DE OPERAÇÃO, IDÊNTICOS NAS DUAS IMPLEMENTAÇÕES.

Modelo de malha facial	MediaPipe Face Landmarker
Confiança mínima	0,6
Resolução de entrada	320×240; 640×480
Taxa de amostragem	10 quadros/s
Aquecimento descartado	1000 ms
Baseline ocular	p75; prior 0,29 (embarcado)
Limiar de nictação	EAR/base $< 0{,}70$; mín. $< 0{,}60$
Duração de nictação	70 a 500 ms
Período refratário	100 ms
Fechamento prolongado	> 500 ms; med. EAR/base $< 0{,}50$
Limiar de oscitação	MAR/base $> 2{,}2$ (saída $1{,}6$)
Duração de oscitação	1000 a 10000 ms

E. Conjunto de dados

A avaliação utilizou os 29 vídeos do YawDD [5], [6] gravados com a câmera instalada no painel, 640×480 a 30 quadros/s, totalizando 2151 s. Os arquivos AVI foram convertidos a MP4 H.264 sem perda perceptível (fator de taxa constante, CRF, igual a 16) e identificados por SHA-256 no manifesto, o que resolve a ambiguidade de onze vídeos com duração arredondada idêntica. Cada vídeo contém as fases de condução normal, fala e bocejo, com o veículo parado. A composição da amostra está na Tabela II.

Tabela II
COMPOSIÇÃO DA AMOSTRA (29 PARTICIPANTES).

Atributo	Categorias	n
Sexo	Feminino / Masculino	13 / 16
Óculos	Com / Sem	15 / 14
Tom de pele [12]	Monk 01–04 / 05–10	20 / 9

F. Anotação de referência

Um observador anotou início e fim de cada evento nos 29 vídeos, com ferramenta de navegação quadro a quadro e candidatos automáticos pré-populados por dois detectores independentes: o MediaPipe Face Landmarker (versão 0.10.14), o mesmo modelo do sistema avaliado, e o preditor de 68 pontos do dlib (versão 19.24) [3], que é o detector do trabalho original da EAR [2]. Os candidatos serviam apenas para orientar a navegação; a decisão sobre cada evento foi do observador. O critério para nictação é a cobertura de mais de metade da pupila no ápice, com início no fechamento e fim na reabertura completa; fechamentos acima de 500 ms constituem classe separada. A oscitação é a abertura ampla e sustentada da boca; fala com boca aberta não é contabilizada.

Um subconjunto de seis vídeos, escolhido para cobrir os extremos de taxa de nictação, uso de óculos e tom de pele, foi revisado por um segundo observador com acesso à marcação inicial. O procedimento caracteriza anotação por consenso e não permite estimar concordância entre anotadores independentes, o que é declarado como limitação. A revisão acrescentou 28 eventos e removeu 3 sobre 268, taxa de omissão de 9,6 % na passada única, valor que deve ser considerado ao interpretar as métricas sobre os 23 vídeos não revisados.

A anotação original do YawDD classifica cada vídeo inteiro; trabalho contemporâneo produziu anotação por quadro para bocejo, com pipeline semiautomático e verificação humana, e mostrou o ganho de rótulos temporais finos para treinamento em plataformas de borda [16]. A anotação aqui é por evento, com início e fim, e cobre também nictação e fechamento prolongado, que aquele trabalho não trata. O gabarito final contém 1002 eventos: 882 nictações, 78 oscitações e 42 fechamentos prolongados. As taxas de nictação resultantes (mediana de 24 por minuto, entre 7 e 56) são compatíveis com a variação individual descrita na literatura [11], elevada pela fase de fala presente em todos os vídeos. Um vídeo (participante com óculos escuros) foi excluído das métricas oculares por impossibilidade de estabelecer referência, permanecendo na de oscitação; a exclusão e seu motivo constam do material suplementar.

G. Métricas

A avaliação é em nível de evento, seguindo o protocolo de pareamento com tolerância temporal estabelecido para detecção de piscada [13] e aplicado recentemente ao mesmo par MediaPipe e EAR sobre o Eyeblink8 [15]. Uma detecção é verdadeiro positivo (VP) quando seu centro dista do centro de um evento anotado até 250 ms para nictação, 500 ms para fechamento prolongado e 1000 ms para oscitação, com pareamento um a um. Detecções sem par são falsos positivos (FP); eventos anotados sem par, falsos negativos (FN). Reportam-se sensibilidade $S = \mathrm{VP}/(\mathrm{VP}+\mathrm{FN})$, valor preditivo positivo $\mathrm{VPP} = \mathrm{VP}/(\mathrm{VP}+\mathrm{FP})$ e

$$F1 = \frac{2\,S \cdot \mathrm{VPP}}{S + \mathrm{VPP}}. \tag{3}$$

Para comparação com a versão v1.24, que registrou apenas contagens por vídeo, usa-se a concordância de contagem $C = \max(0,\, 1 - |Q_{\mathrm{ref}} - Q_{\mathrm{det}}|/Q_{\mathrm{ref}}) \times 100$, normalizada pela referência. A concordância entre implementações em dispositivos distintos é medida pelo índice de Jaccard sobre os conjuntos de eventos, com tolerância de 100 ms.

H. Custo computacional

Seguindo a prática de benchmarking de inferência em SoCs móveis [14], [17], cada dispositivo processou os 29 vídeos em quatro configurações, num desenho fatorial 2×2: acelerador (GPU ou CPU do MediaPipe) e resolução de entrada (320×240 ou 640×480), na ordem GPU-CPU-GPU-CPU com pausa de 60 s entre configurações. Mediu-se a latência de cada inferência, a memória proporcional do processo (*Proportional Set Size*, PSS) e o tempo de CPU do processo por `/proc/self/stat`. O estrangulamento térmico foi verificado comparando a latência mediana do primeiro e do último terço de cada rodada. O consumo de bateria não é medido neste protocolo, porque o processamento de vídeo armazenado não representa a operação com aquisição de imagem da câmera em tempo real; sua medição em sessão ao vivo é reservada

a estudo futuro, junto com a análise de convergência com eletroencefalografia (EEG).

III. Resultados e Discussão

A. Reprodutibilidade entre dispositivos

Os três dispositivos produziram exatamente as mesmas detecções sobre os mesmos quadros. Na resolução de produção (320×240), usando CPU, nictação resultou em VP 660, FP 48 e FN 222 no Galaxy M51 (Snapdragon 730, Adreno 618), no Galaxy A54 (Exynos 1380, Mali-G68) e no Galaxy S24 FE (Exynos 2400e, Xclipse 940). O índice de Jaccard mediano entre pares de dispositivos foi 1,000 para nictação. Entre GPU e CPU no mesmo dispositivo, as detecções diferiram em no máximo quatro eventos de cerca de 660, atribuíveis a precisão de ponto flutuante.

Esse resultado sustenta a afirmação central do trabalho: o modelo e as regras produzem saída idêntica em três famílias de GPU e três gerações de CPU. Ele não implica invariância a câmeras distintas, não testadas.

B. Desempenho na detecção

Tabela III
Desempenho por evento e fonte, 29 vídeos. Dispositivos a 10 quadros/s, 320×240, CPU; os três dão valores idênticos e são reportados uma vez.

Evento	Fonte	VP	FP	FN	S	VPP	F1
Nictação	Desk. 30 q/s	742	48	140	0,841	0,939	0,888
	Desk. 10 q/s	720	50	162	0,816	0,935	0,872
	Desk. 10 causal	710	50	172	0,805	0,934	0,865
	Disp. 10 q/s	660	48	222	0,748	0,932	0,830
Oscitação	Desk. 10 q/s	70	42	8	0,897	0,625	0,737
	Desk. 10 causal	64	65	14	0,821	0,496	0,618
	Disp. 10 q/s	65	62	13	0,833	0,512	0,634
Fech. prol.	Desk. 10 q/s	35	25	7	0,833	0,583	0,686
	Desk. 10 causal	35	24	7	0,833	0,593	0,693
	Disp. 10 q/s	33	38	9	0,786	0,465	0,584

A Tabela III organiza as fontes de modo que a diferença entre linhas consecutivas isole um fator. Para nictação, passar de 30 para 10 quadros/s no desktop custa 0,016 de F1; tornar o baseline causal custa 0,007; e a plataforma embarcada custa 0,035. A parcela da plataforma corresponde à cadeia de decodificação de vídeo e conversão de cor no Android, distinta do OpenCV no desktop, e ao *build* Android do MediaPipe. O VPP permanece em 0,93 em todas as fontes: a diferença é integralmente de sensibilidade, e não de falsos positivos.

Para oscitação, a taxa de amostragem não pesa (bocejo dura segundos), mas a estimativa causal do baseline custa 0,12 de F1, todo em VPP. O mecanismo é o viés de fase: os vídeos iniciam com a boca fechada, a mediana crescente permanece baixa e o limiar de 2,2× fica abaixo da abertura da fala que vem depois. A implementação de desktop, com a mediana sobre o vídeo inteiro, não sofre esse viés. Mesmo assim, a sensibilidade de 0,833 no dispositivo supera os 60 % reportados pelos autores do YawDD [5] com Viola–Jones nos mesmos vídeos.

A Fig. 3 ilustra a cadeia avaliada sobre um fechamento prolongado real no dispositivo. O fechamento prolongado é o evento de pior desempenho, com VPP de 0,465. Os falsos positivos residuais são fechamentos sustentados com EAR entre 0,50 e 0,60 do baseline, que o critério de profundidade não separa. Antes desse critério, o VPP era 0,29; os 58 eventos reclassificados como fechamento parcial correspondem aos falsos positivos de sonolência evitados.



Figura 3. Cadeia avaliada, sobre um fechamento prolongado: captura, malha facial, EAR normalizada cruzando o limiar de entrada (0,70) e permanecendo abaixo de 0,50 por mais de 500 ms, e registro do evento no dispositivo. O alerta visual, sonoro ou por vibração é funcionalidade implementada do sistema, acionada pela detecção; sua avaliação não é objeto deste trabalho.

A taxa de detecção de rosto foi 99,7 % dos quadros nas duas plataformas, contra 93 % reportados pelos autores do YawDD com Viola–Jones; a de boca, 99,7 % contra 78 %.

C. Varredura de limiar

A Fig. 4 mostra sensibilidade, VPP e F1 em função do limiar, no desktop a 30 quadros/s. Para nictação, o F1 é máximo em 0,70, valor fixado antes da anotação existir. Para oscitação, o F1 cresce até o fator 3,3 sem atingir máximo, à custa de sensibilidade; o valor de 2,2 foi mantido por priorizar sensibilidade, dado que a omissão de bocejo é o erro de maior consequência para a aplicação.

D. Análise por subgrupo

A Tabela IV estratifica a nictação por uso de óculos e por sexo, e a oscitação por sexo, no desktop a 30 quadros/s. Óculos de grau não degradam a detecção de nictação: a diferença de F1 entre os grupos com e sem óculos é de 0,02, dentro da variação entre participantes.

Óculos escuros, presentes em um participante, impedem a observação do olho; nesse vídeo, os dois detectores usados na anotação produziram saída espúria (1 evento com MediaPipe, 35 com dlib) sem sinalizar a impossibilidade, modo de falha silencioso que um sistema em produção precisa detectar explicitamente; o sistema de origem possui modo dedicado a essa condição, não avaliado aqui. A oscitação

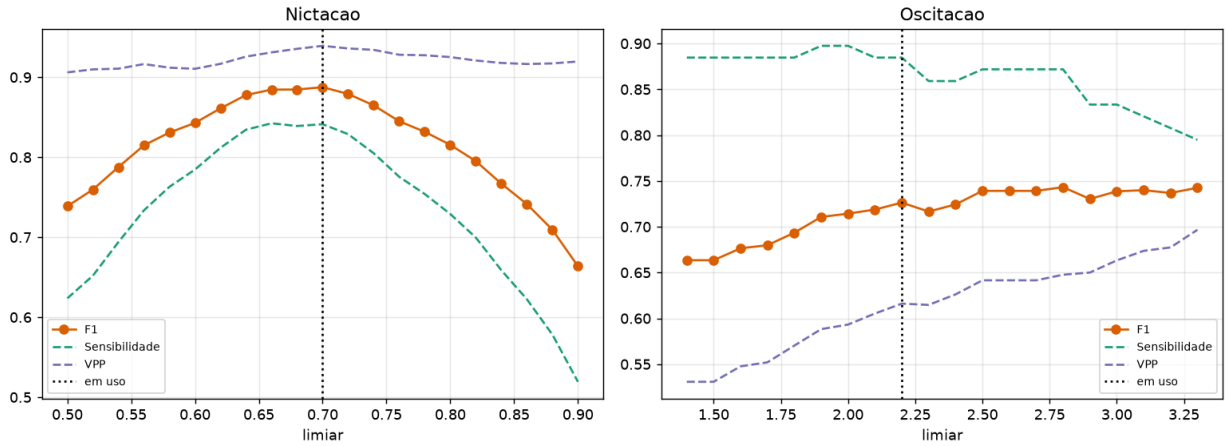


Figura 4. Sensibilidade, VPP e F1 em função do limiar, desktop 30 quadros/s. Esquerda: nictação, limiar sobre EAR/baseline. Direita: oscitação, fator sobre MAR/baseline. Linha pontilhada: valor em uso.

Tabela IV
DESEMPENHO POR SUBGRUPO, DESKTOP 30 QUADROS/S.

Evento	Subgrupo	n	S	VPP	F1
Nictação	Sem óculos	14	0,844	0,959	0,898
	Com óculos	15	0,838	0,917	0,876
	Feminino	13	0,880	0,948	0,913
	Masculino	16	0,796	0,928	0,857
Oscitação	Feminino	13	0,853	0,617	0,716
	Masculino	16	0,909	0,615	0,734

não é estratificada por óculos: lentes não afetam a boca, e a estratificação preliminar produziu diferença confundida com a distribuição de participantes que estão conversando.

E. Custo computacional

Tabela V
CUSTO COMPUTACIONAL, DESENHO FATORIAL 2×2, 21 530 QUADROS POR CONFIGURAÇÃO.

Dispositivo	Res.	Acel.	Lat. (ms)	p95 (ms)	q/s	PSS (MB)
Galaxy A54	320	GPU	36	42	27,8	314
	320	CPU	26	31	38,5	286
	640	GPU	40	43	25,0	323
	640	CPU	27	32	37,0	298
Galaxy M51	320	GPU	71	84	14,1	282
	320	CPU	66	113	15,2	258
	640	GPU	73	86	13,7	280
	640	CPU	71	116	14,1	257
Galaxy S24 FE	320	GPU	63	87	15,9	320
	320	CPU	102	145	9,8	300
	640	GPU	65	89	15,4	319
	640	CPU	118	152	8,5	307

Três achados emergem da Tabela V. Primeiro, a resolução de entrada não afeta a latência (razões entre 0,92 e 1,06 nas seis combinações): o modelo redimensiona internamente, e a escolha de 320×240 em produção não custa nem tempo nem acurácia, tendo dado F1 marginalmente superior ao de 640×480. Segundo, todos os dispositivos sustentam a taxa de 10 quadros/s com folga em ao menos um acelerador, com o uso de CPU do processo entre 110 e 128 % de um núcleo. Terceiro, a vantagem entre GPU e CPU depende do SoC: no A54 a CPU foi 28 % mais rápida, no M51 os dois se equivalem, e no S24 FE a GPU foi 38 % mais rápida, variabilidade consistente com a documentada em benchmarks de inferência móvel, nos quais a relação entre latência de CPU e de GPU não se mantém entre plataformas [14], [17]. No S24 FE, a CPU operou entre 1,4 e 1,5 GHz, cerca de metade da frequência máxima do núcleo, com a thread alocada consistentemente aos núcleos de desempenho; a prioridade de thread não alterou o resultado, o que aponta para política de energia do sistema operacional. O comportamento, em que o governador mantém a CPU em frequência subótima durante inferência contínua, é observado aqui em um modelo de visão de pequeno porte, o que sugere que o mecanismo é do sistema operacional e não da carga. A escolha de acelerador, portanto, não pode ser fixada em tempo de compilação.

O estrangulamento térmico dentro de cada rodada foi ausente em dez das doze configurações. O A54 em GPU a 320×240 apresentou aumento de 22 % na latência entre o primeiro e o último terço; o S24 FE em GPU a 640×480 apresentou redução de 59 %, de 70 para 29 ms, comportamento consistente com governador de frequência que atinge o regime após período de carga sustentada. A sensibilidade das medições a esse efeito é reconhecida na literatura de benchmarking móvel, que recorre a resfriamento ativo e fixação de frequência para obter latências estáveis [17]; aqui optou-se por medir na condição de uso, sem essas intervenções, e reportar a variação.

F. Evolução do sistema

A versão v1.24, avaliada à época por contagem sobre os mesmos vídeos, foi reavaliada contra o gabarito atual e obteve concordância de contagem média de 61,2 % para nictação. O

detector de nictação da versão corrente (v40), executado nos mesmos quadros no dispositivo, obteve 77,1 %, e em nível de evento F1 de 0,740 (VP 610, FP 157, FN 272). As regras de referência deste trabalho, nos mesmos quadros (GPU), obtiveram F1 de 0,832 com FP 45: a diferença está no uso do coeficiente de expressão de um único olho sem filtro de profundidade na v40, contra a EAR geométrica dos dois olhos com filtro. Combinar as duas fontes, aceitando candidatos rasos da referência confirmados pela v40, elevou o F1 a 0,842 (sensibilidade 0,779, VPP 0,917); a combinação foi escolhida após a análise e é reportada como variante, não como método principal.

G. Lições aprendidas

O detector foi refinado em ciclos de análise de falha contra a anotação de referência: cada mudança partiu de um padrão observado nos falsos positivos ou negativos, foi implementada nas duas plataformas e reavaliada. Os achados abaixo são reportados porque se aplicam a qualquer detector de eventos faciais operando em fluxo, e porque os resultados negativos delimitam o que não funcionou.

Duração a partir de amostras esparsas. A 10 quadros/s, uma piscada de 100 a 150 ms cabe em um único quadro. Estimar a duração como a diferença entre o primeiro e o último quadro abaixo do limiar dá zero para esses casos, e o piso de 70 ms os descarta: 30 % das corridas abaixo do limiar eram perdidas assim. Somar um intervalo de amostragem, estimativa de viés mínimo, recuperou-as sem alterar a classificação dos eventos longos.

Baseline em fluxo. A EAR de olho aberto varia 1,6 vez entre pessoas, o que exige normalização individual. Em fluxo, o baseline só pode ser estimado do passado. A primeira implementação esperava 120 quadros antes de detectar, 4 s para operações com 30 quadros/s, porém 12 s operações com 10 quadros/s, e deixava 18,7 % de cada vídeo cego, tornando inalcançáveis 158 das 882 nictações. Reduzir a janela para 30 quadros recuperou a maior parte; iniciar com um valor a priori populacional e refiná-lo pelo percentil da janela crescente eliminou o período cego. Para a boca, o mesmo valor a priori produziu falsos positivos, porque a fala inicial excedia um limiar calibrado sobre repouso; ali, esperar 4 s custa poucos bocejos, que são raros e longos, e é a escolha correta. A estratégia de baseline depende, portanto, da estatística do evento.

Profundidade do fechamento. Os falsos positivos de nictação e de fechamento prolongado tinham em comum EAR mínima entre 0,59 e 0,65 do baseline, contra 0,29 a 0,33 nos verdadeiros: cruzavam o limiar de entrada mas não fechavam. Exigir profundidade mínima (0,60 no ápice para nictação; mediana abaixo de 0,50 para prolongado) elevou o VPP de 0,884 para 0,944 e de 0,29 para 0,50, respectivamente, com perda de sensibilidade inferior a 0,01. Pose da cabeça (*pitch*) foi testada como discriminador alternativo e não separou os grupos.

Coeficiente de expressão. O *blendshape* de piscada, usado como sinal primário, foi testado como confirmação e como alternativa à geometria, em ambos os eventos oculares. Não superou a profundidade da EAR em nenhum caso; para fechamento sustentado, o coeficiente de um único olho deu F1 de 0,34 contra 0,58 da EAR bilateral. Uma hipótese é que o coeficiente reconhece o gesto de piscar, rápido e ativo, e não a pálpebra caindo devagar, que é geometria distinta.

Limiar sem medida. Uma versão anterior do sistema usava limiar de 0,80 sobre um coeficiente cujo máximo observado no sinal era 0,80, e produziu uma única detecção em 1496 quadros. Limiares devem ser derivados da distribuição medida do sinal, não de estimativa.

H. Limitações

O YawDD contém 29 participantes, expressões atuadas e veículo parado; a generalização a condução real, com iluminação variável e vibração, não foi testada. A anotação é consensual, sem medida de concordância entre observadores independentes. O baseline causal da boca sofre viés de fase, com custo de cerca de 0,1 de F1 no dispositivo. Óculos escuros impedem a extração ocular e o sistema não sinaliza a condição nas implementações avaliadas. A etapa de inferência de estado não foi avaliada, por ausência de referência. A implementação iOS existe e não foi avaliada. O consumo de bateria em operação com câmera é reservado a estudo futuro.

IV. Conclusões

Um detector de nictação, fechamento prolongado e oscitação baseado em razões de aspecto sobre malha facial foi executado integralmente em três smartphones Android com SoCs de arquiteturas distintas e produziu detecções idênticas ao evento nos três. Contra anotação temporal consensual de 1002 eventos em 29 vídeos do YawDD, obteve-se F1 de 0,830 para nictação e 0,634 para oscitação a 10 quadros/s, com sensibilidade de oscitação superior à reportada pelos autores do conjunto. A diferença de desempenho entre desktop e dispositivo foi decomposta em taxa de amostragem, causalidade e plataforma, com esta última respondendo por 0,035 de F1. Um desenho fatorial mostrou que a resolução de entrada não afeta a latência e que a vantagem entre GPU e CPU depende do SoC. Os resultados constituem prova de conceito técnica da extração embarcada; sua validade clínica não é objeto deste trabalho.

A extração embarcada de sinais faciais é agnóstica à aplicação. Uma linha em estudo é a avaliação de dor em pacientes não verbais por escalas como a PSPI (*Prkachin and Solomon Pain Intensity*) [18], ainda sem avaliação. Está em curso, sem resultado no momento da submissão, a investigação da convergência entre os sinais faciais aqui extraídos e sinais de EEG, com atenção ao artefato ocular no EEG frontal, que correlaciona com a piscada por construção e não por mecanismo neural.

Disponibilidade de dados e código

A anotação temporal consensual, o manifesto com SHA-256 dos vídeos, o arquivo de exclusões, a implementação de referência em Python, o detector embarcado em Kotlin, a ferramenta de anotação e os CSVs de saída dos três dispositivos

estão depositados em https://github.com/NeriOliveira/artigo_cbeb_2026 (SWHID: swh:1:rev:e35c7c3a3a4563734ea5351a34e17166b5755c7f), o que permite reproduzir integralmente as Tabelas III a V. Os vídeos do YawDD não são redistribuídos e devem ser obtidos na fonte [6]; o manifesto permite verificar a correspondência exata dos arquivos. O detector da versão corrente do sistema comercial, usado apenas na comparação da Seção III-F, não é distribuído; os eventos que ele produziu constam dos CSVs depositados.

AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento por meio dos processos 2023/04821-0 e 2024/13300-7, ao Grupo de Pesquisa de Processamento de Sinais da UFABC e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

CONFLITO DE INTERESSE

O primeiro autor é sócio da Waker Inova Simples IS, empresa que desenvolve o sistema cujo pipeline de extração é avaliado neste trabalho. A anotação de referência, o protocolo de avaliação, os dados e o código de análise estão publicamente disponíveis para verificação independente, conforme a seção de disponibilidade. O segundo autor declara não haver conflito de interesse.

REFERÊNCIAS

[1] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2019.

[2] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in *21st Computer Vision Winter Workshop*, Rimske Toplice, Slovenia, 2016.

[3] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 1867–1874.

[4] A. Palao, R. Fredriksson, and M. Lenné, "Euro NCAP's current and future in-cabin monitoring systems assessment," in *Proc. 27th Int. Tech. Conf. Enhanced Safety of Vehicles (ESV)*, Yokohama, Japan, 2023, paper 23-0286.

[5] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: A yawning detection dataset," in *Proc. 5th ACM Multimedia Systems Conference (MMSys '14)*, Singapore, 2014, pp. 24–28, doi: 10.1145/2557642.2563678.

[6] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: Yawning Detection Dataset," IEEE Dataport, 2020, doi: 10.21227/e1qm-hb90.

[7] T. Drutarovsky and A. Fogelton, "Eye blink detection using variance of motion vectors," in *Computer Vision – ECCV 2014 Workshops*, Springer, 2015, pp. 436–448.

[8] J. D. Ortega *et al.*, "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Computer Vision – ECCV 2020 Workshops*, Springer, 2020, pp. 387–405.

[9] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019.

[10] D. F. Dinges and R. Grace, "PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance," Federal Highway Administration, Tech. Rep. FHWA-MCRT-98-006, 1998.

[11] A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali, and A. Albanese, "Analysis of blink rate patterns in normal subjects," *Movement Disorders*, vol. 12, no. 6, pp. 1028–1034, 1997.

[12] E. Monk, "The Monk Skin Tone Scale," 2019. [Online]. Available: https://skintone.google

[13] A. Fogelton and W. Benesova, "Eye blink detection based on motion vectors analysis," *Computer Vision and Image Understanding*, vol. 148, pp. 23–33, 2016.

[14] A. Ignatov *et al.*, "AI Benchmark: All about deep learning on smartphones in 2019," in *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW)*, Seoul, Korea, 2019, pp. 3617–3635.

[15] J. E. Delgado Gómez, E. M. España, C. F. R. Rodas, and D. E. G. Villamarín, "Proposal for a eye blink detection using Mediapipe, eye aspect ratio and peak identification," in *Converging Clinical and Engineering Research on Neurorehabilitation V (ICNR 2024)*, Biosystems & Biorobotics, vol. 32, Springer, 2024, doi: 10.1007/978-3-031-77584-0_67.

[16] A. Mujtaba, G. Radchenko, M. Masana, and R. Prodan, "YawDD+: Frame-level annotations for accurate yawn prediction," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2026, to be published. arXiv:2512.11446.

[17] Z. Li, M. Paolieri, and L. Golubchik, "A benchmark for ML inference latency on mobile devices," in *Proc. 7th Int. Workshop on Edge Systems, Analytics and Networking (EdgeSys '24)*, Athens, Greece, 2024, pp. 31–36, doi: 10.1145/3642968.3654818.

[18] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

Wrist-Worn Smartband with Embedded Artificial Intelligence for Fall Detection in Individuals with Cognitive Impairment

Gabriel Cortes Pires Leopolski
Dept. of Electrical Engineering (DAELT)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
gabrielleopolski@alunos.utfpr.edu.br

Leonardo Zucco
Dept. of Electrical Engineering (DAELT)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
zucco@alunos.utfpr.edu.br

Rodrigo Baldissera Cerqueira Leite
Dept. of Electrical Engineering (DAELT)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
rodrigobaldissera@alunos.utfpr.edu.br

Renata Coelho Borges
Graduate Program in Biomedical Engineering (PPGEB), DAELT
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
renatacoelho@utfpr.edu.br

Adriano Ruseler
Dept. of Electrical Engineering (DAELT)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ruseler@utfpr.edu.br

Alexandre Francisco de Moraes Filho
Federal Institute of Paraná (IFPR)
Curitiba Campus
Curitiba, Brazil
alexandre.moraes@ifpr.edu.br

Fábio Pires Itturriet
Graduate Program in Biomedical Engineering (PPGEB), DAELT
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
fabioitturriet@utfpr.edu.br

Amauri Amorin Assef
Graduate Program in Electrical and Industrial Informatics Engineering (CPGEI), DAELT
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
amauriassef@utfpr.edu.br

***Abstract*—Population aging and the increasing prevalence of neurodegenerative diseases have increased the need for technological solutions capable of supporting health monitoring and preventing adverse events, particularly falls among individuals with cognitive impairment. This paper presents a technical proof-of-concept wrist-worn smartband that combines embedded TinyML with a post-classification, multi-stage physical verification pipeline for automatic fall detection. The proposed system integrates an M5Stack AtomS3R-Ext microcontroller with an embedded Inertial Measurement Unit (IMU), a 1D convolutional neural network (CNN) quantized to 8-bit integer (int8) and trained via Edge Impulse, and a three-stage physical verification method designed to reduce false positives beyond the performance of the CNN alone. Communication is handled through the Message Queuing Telemetry Transport (MQTT) protocol, enabling alerts via a companion mobile application. On an internal validation split automatically held out during training (n=849 windows), the embedded model reached 95.2% accuracy and a ROC-AUC of 0.951, with a mean inference time of 7 ms and a computational load of 0.7%. In practical on-device trials (10 per fall type, performed by the two healthy adult authors who also contributed to the training data), detection rates were 100% for forward falls, 80% for backward falls, and 70% for lateral falls, with one false positive recorded. These results indicate that fully on-device inference combined with multi-stage physical verification is computationally feasible on resource-constrained wearables. However, given the small, non-independent, and self-collected dataset, clinical suitability for the target population requires further evaluation with leakage-free, subject-level data and cohorts representative of older adults with cognitive impairment.**

***Index Terms*—fall detection, embedded artificial intelligence, wearable devices, Edge Impulse, elderly monitoring, cognitive impairment**

I. Introduction

Population aging is a growing global trend, with the proportion of people aged 60 or older expected to double by 2050, according to the World Health Organization (WHO) [1]. In Brazil, approximately 8.5% of the elderly population live with some form of dementia, a figure projected to rise from 1.8 million to 5.7 million by 2050 [2]. Neurodegenerative diseases impair cognitive and functional abilities, substantially increasing the risk of falls. In this context, cognitive impairment refers to deficits in functions such as memory, attention, spatial orientation, judgment and autonomy, which may increase fall risk and hinder the ability to seek help after

an adverse event. Studies report annual fall rates of 40-60% among elderly individuals with dementia [3], [4], compared to roughly one-third of the general older population [1].

Falls are a leading cause of morbidity and mortality among this population, making reliable, fast-response detection systems desirable [1]. Wrist-worn wearables equipped with Inertial Measurement Units (IMUs) are particularly attractive due to their unobtrusiveness, continuous monitoring capability, and high user acceptance [5], [6]. When combined with machine learning (ML), such devices have been reported to distinguish fall events from activities of daily living with high accuracy [7], [8]. However, several fall-detection systems still rely on offline processing or cloud-based inference [9], [11], which introduces latency and privacy concerns, and increases energy consumption. Edge Artificial Intelligence (AI), executing ML models directly on the microcontroller, addresses these limitations by enabling local, real-time inference without constant server dependency [10], [12].

A. Related Work

Several studies have explored wearable-based fall detection using ML across different body placements, as summarized in Table I. Marques et al. [7] evaluated classical ML algorithms on a wrist-worn device, while Nunez et al. [8] compared waist and wrist placements, reporting that wrist sensors exhibit higher classification ambiguity than trunk-mounted alternatives. Jang et al. [9] employed a dual-stream CNN with self-attention for waist devices, and Li et al. [10] reviewed a decade of progress, highlighting trends toward embedded AI. Most of these approaches rely on the classifier's output alone to trigger an alert, which makes them susceptible to false positives whenever daily activities produce IMU signatures similar to a fall.

The most comparable study is that of Volochtchuk and Bassai [11], who reported 92.31% accuracy using Edge Impulse, but conducted their final tests with the device connected to a computer, without autonomous deployment. In contrast, the main methodological contribution of this work is not the CNN classifier in isolation, but its integration with a three-stage, rule-based physical verification pipeline (temporal voting, impact detection, and post-impact immobility check) executed entirely on-device after inference. As described in Section III, this pipeline is what allows high-intensity activities such as running and jumping, which produce CNN-detectable acceleration peaks comparable to those associated with falls, to be rejected before an alert is issued, a capability that a CNN-only or threshold-only approach does not provide. This work also advances toward fully autonomous embedded deployment and integrates Wi-Fi-based geofencing, prioritizing wrist placement for user acceptance. Beyond falls, wandering behaviors [13] and IMU-derived fall-risk parameters [14] further motivate the relevance of this line of work, although wandering detection is not evaluated in this study.

This paper presents a wrist-worn smartband technical proof-of-concept built around the M5Stack AtomS3R-Ext microcontroller, combining a 1D convolutional neural network (CNN) quantized to 8-bit integer (int8) and deployed via the Edge Impulse platform [15] with a multi-stage physical verification method intended to reduce false positives beyond those produced by the classifier alone.

TABLE I
SUMMARY OF RELATED WORK ON WEARABLE FALL DETECTION

Study	Placement	Algorithm	Key Contribution
Marques et al. [7]	Wrist	Classical ML	Offline evaluation of multiple algorithms
Nunez et al. [8]	Wrist/ Waist	Random Forest	Compares sensor placement impact
Jang et al. [9]	Waist	Dual-stream CNN	Self-attention for complex fall patterns
Li et al. [10]	Multiple	Classical ML / DL	Decade review; trends in embedded AI
Volochtchuk & Bassai [11]	Chest / Foot	SVM, Decision Tree, Logistic Regression	92.31% accuracy; no autonomous deployment
This work	**Wrist**	**1D-int8-CNN + physical verification**	**On-device CNN and rule-based verification; mobile app**

II. MATERIALS AND METHODS

A. System Architecture

The system comprises three main layers: (1) a wrist-worn embedded device responsible for data acquisition, on-device ML inference, and fall confirmation; (2) Message Queuing Telemetry Transport (MQTT)-based communication layer for event-driven messaging; and (3) a mobile application specially designed to generate and deliver alert notifications to caregivers following a confirmed fall event. Fig. 1 illustrates the overall system architecture, while Fig. 2 shows the mobile application interface.

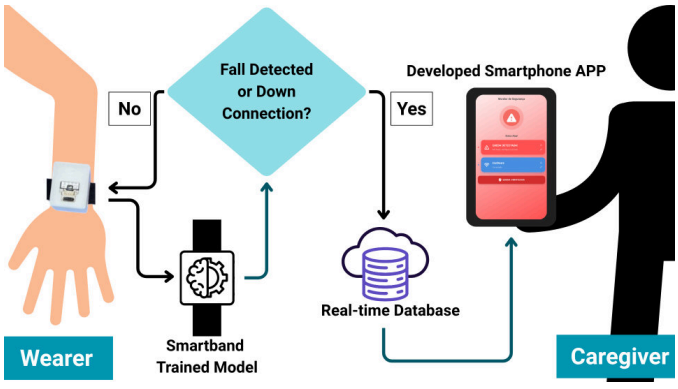


Fig. 1. System architecture: the embedded device acquires IMU data, runs on-device ML inference, and publishes fall or geofencing events via MQTT to a cloud backend, which triggers push notifications to the caregiver's mobile application.

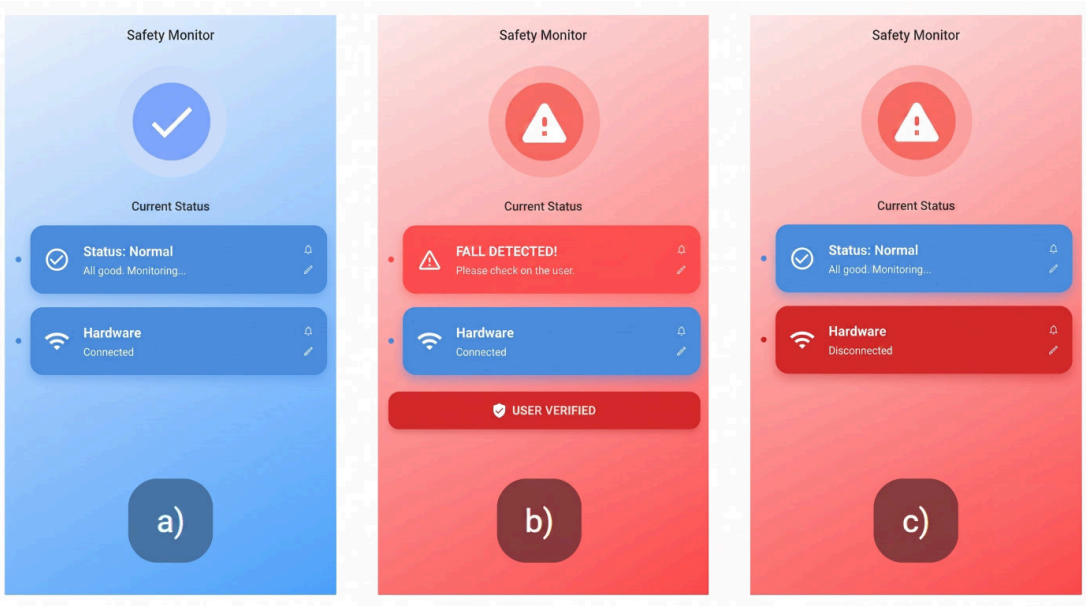


Fig. 2. Mobile application dashboard showing three operating states: (a) normal operation, with the wearable device connected and monitoring active; (b) fall detected, displaying an alert and confirmation button; (c) device disconnected, indicating loss of communication with the smartband.

During normal activity, no data are published to the broker, minimizing network traffic and energy consumption. A fall event triggers a compact MQTT message to the backend, which updates the Firebase real-time database and dispatches a push notification via Firebase Cloud Messaging (FCM) to the caregiver's Flutter application.

B. Hardware Device

The M5Stack AtomS3R-Ext [16] was selected as the embedded platform due to its combination of compact form factor (24 × 24 × 13 mm), built-in IMU (three-axis BMM150 geomagnetic sensor and a six-axis BMI270 attitude sensor), dual-core Xtensa® LX7 processor at up to 240 MHz, 8 MB of Flash, 8 MB of PSRAM, native Wi-Fi connectivity, and low cost (approximately US$14.50). These characteristics make it well suited for a wrist-worn wearable requiring on-device ML inference and wireless communication. The device was also selected due to its native compatibility with Edge Impulse, an edge AI platform that provides an integrated workflow for the entire ML pipeline on embedded systems, including data acquisition, model training, optimization, and deployment. This integration enables the deployment of optimized TinyML models directly onto microcontrollers, reducing development complexity and computational overhead. The selected microcontroller and its battery module are shown in Fig. 3, respectively. An Atomic Battery Base (200 mAh, 24 × 24 × 27 mm) was used solely to provide autonomous power during data collection and testing; battery-life optimization was beyond the scope of this work and is left for future development. A custom 3D-printed PLA enclosure fixed the device to the wrist via a Velcro strap, yielding a total wearable mass of 35 g, as shown in Fig. 4 and 5.

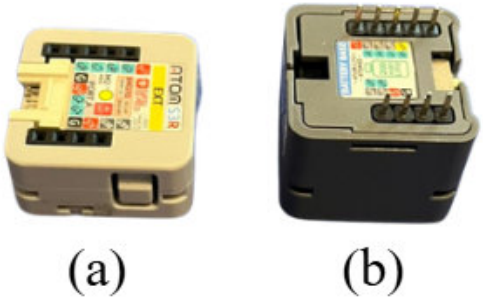


Fig. 3. Hardware used in the prototype: (a) the M5Stack AtomS3R-Ext microcontroller and (b) the Atomic Battery Base module.



Fig. 4. Prototype of the wrist-worn smartband: M5Stack AtomS3R-Ext enclosed in a 3D-printed PLA case and fixed to the wrist via a Velcro strap.

C. Signal Acquisition and Filtering

The IMU outputs accelerometer data in multiples of gravitational acceleration (g) and gyroscope data in degrees per second (°/s). A moving-average filter with window size N was applied to attenuate high-frequency noise:

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} x[n-k] \tag{1}$$



Fig. 5. Smartband worn on the wrist during operation.

where $y[n]$ is the filtered output at sample n, $x[n]$ is the raw sensor signal, and N is the window size. Windows of $N \in \{1, 5, 10, 20\}$ were evaluated. $N = 5$ provided the best trade-off between noise suppression and dynamic response preservation, and was adopted as the definitive filter parameter.

1) Dataset Collection: A pilot experimental protocol was conducted with two healthy adult volunteers, both authors of this work, simulating fall events and activities of daily living (ADL). This is a significant limitation of the present study: the dataset was neither collected from the target population (older adults with cognitive impairment) nor from an independent, larger cohort, and results should therefore be interpreted strictly as a technical proof-of-concept. Fall events (forward, lateral, and backward) were recorded, resulting in 100 simulated fall trials per subject, totalling 200 labelled fall trials. ADL samples, walking, light running, jumping, sitting, dishwashing, and arm raising, were also recorded (approximately 10 samples per activity per subject). Raw IMU signals were then segmented into windowed samples of 1000 ms with a 250 ms stride (75% overlap) at a 20 Hz sampling frequency, yielding a total of 490 windowed samples distributed between two classes: *fall* and *non-fall*. The Edge Impulse platform split these 490 windows randomly at the window level into 386 for training and 104 for testing, with both volunteers contributing to both subsets.

Because adjacent windows were extracted with 75% overlap, highly correlated segments from the same fall or ADL event can appear in both the training and test subsets. This constitutes data leakage at the window level and is considered the most critical methodological limitation of this study: it may substantially inflate the accuracy reported in Section III, since the model can partially memorize event-specific patterns rather than learning fall dynamics that generalize to unseen events or subjects. In addition, the classification metrics reported in this paper (Section III) correspond to an internal validation split of $n = 849$ windows automatically held out by the Edge Impulse training pipeline from the 386 training windows, rather than to the independently reserved 104-window test set. A fully independent, leakage-free evaluation on that test set was not performed and is identified as a priority for future work. A rigorous evaluation should instead partition data by event and by subject (e.g., leave-one-subject-out), ensuring that no windows from the same fall event appear in both training and evaluation splits.

Furthermore, falls in older adults may exhibit kinematic signatures different from those of healthy young adults, due to factors such as reduced movement speed, impaired protective reflexes, and altered gait patterns. Given the small sample (two subjects), the window-level data leakage described above, and the absence of participants from the target population, the dataset and the results derived from it should be regarded strictly as a proof-of-concept, and validation with independent, subject-diverse, and target-population data remains necessary before any clinical interpretation. Fig. 6 illustrates a representative IMU signal sample recorded during a forward fall trial, showing the characteristic pattern of accelerometer and gyroscope responses across the three phases: motionless, fall, and post-fall motionless.

2) Model Architecture: The model was developed on the Edge Impulse platform. Raw six-axis IMU signals were segmented into 1000 ms windows with a 250 ms stride (75% overlap) at 20 Hz, resulting in 20 samples per axis and 120 input values per window. Time-domain raw data were used directly, without spectral transformation, since falls are transient and non-periodic events better represented in the temporal domain [17]. A 1D CNN was employed, as illustrated in Fig. 7. The architecture consists of an input layer (120 features), a reshape layer to organize the 120 input values according to the six IMU axes, two 1D conv/pool blocks (24 filters, kernel size of 5; followed by 48 filters, kernel size 3), a flatten layer, a dense layer (32 neurons), and a 2-class softmax output layer. Training was performed for 120 epochs with a learning rate of 0.0005, selected empirically to ensure stable convergence without overshooting the loss minimum.

After training, the model was quantized to int8 and exported as an optimized C++ library integrated directly into the device firmware. The int8 version was selected for embedded deployment to reduce memory footprint and inference cost on the microcontroller, enabling real-time operation within the resource constraints of the M5Stack AtomS3R-Ext.

D. Multi-Stage Fall Confirmation

To reduce false positives inherent to wrist-worn sensors, a three-stage verification pipeline was implemented after the CNN inference. All thresholds were determined empirically based on the analysis of IMU signal plots collected during the same limited experimental protocol described in Section II-C1, selecting values that maximized fall detection while minimizing false positives in ADL scenarios. Because these thresholds were calibrated using the same small, self-collected dataset used for model development, they may be overfitted to the specific movement patterns of the two subjects and the specific

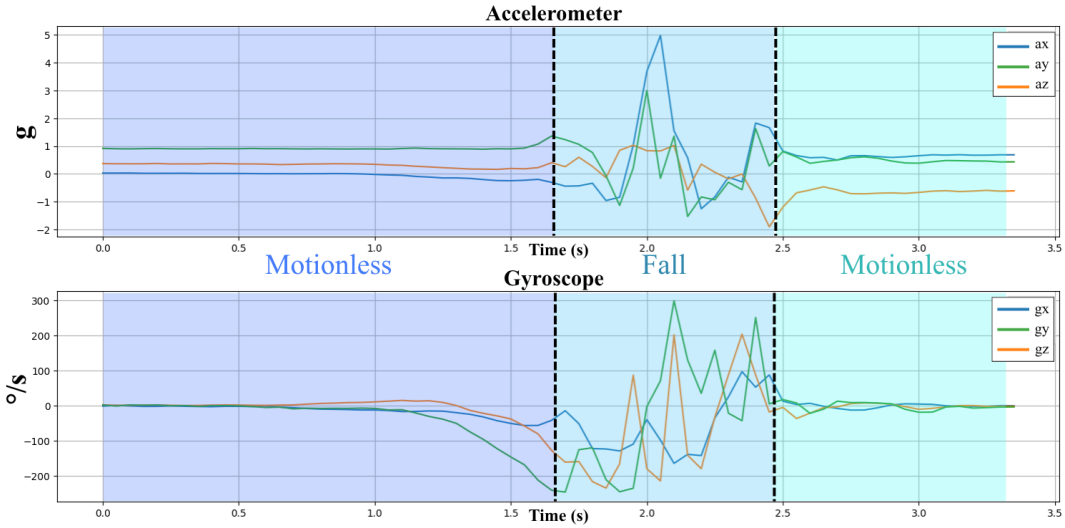


Fig. 6. Representative IMU signals during a experimental forward fall sample: accelerometer (top) and gyroscope (bottom) responses across three phases: motionless, fall, and post-fall motionless.



Fig. 7. 1D CNN architecture deployed on the M5Stack AtomS3R-Ext comprising an input layer (120 features), a reshape layer, two 1D convolution/pooling blocks, a flatten layer, a dense layer (32 neurons), and 2-class output layer.

ADL repertoire recorded, and their generalization to other individuals and activities has not been verified.

Stage 1 – K-of-M temporal voting: a fall suspicion is raised when at least $K = 1$ of the last $M = 2$ inference windows exceeds a fall probability threshold of 0.85. These values were determined empirically: more conservative configurations (e.g., $K = 2$, $M = 2$) caused genuine fall events to be missed, since requiring multiple high-probability windows within such a short sequence is overly restrictive given that a fall is a brief, transient event spanning few windows. While increasing M could potentially reduce false positives during high-intensity activities, it would also delay or suppress the detection of real falls, compromising the system's primary safety objective. The selected $K = 1$, $M = 2$ configuration prioritizes detection sensitivity, with false-positive suppression delegated to the subsequent physical verification stages.

Stage 2 – Impact verification: within a 1.8 s window following suspicion, either (a) acceleration magnitude exceeds 3.80 g, or (b) acceleration magnitude exceeds 3.40 g simultaneously with angular velocity magnitude exceeding 380 °/s.

Stage 3 – Post-impact immobility: over a minimum 1 s interval, angular velocity must remain below 20 °/s and acceleration magnitude variation must stay within 0.12 g of the reference value, confirming that the subject is motionless after impact.

Only events satisfying all three stages are classified as confirmed falls and trigger an MQTT alert.

III. Results

A. Model Validation Performance

The Edge Impulse platform automatically reserves a fraction of the 386 training windows described in Section II-C1 as an internal validation split during model training; this split, rather than the independently held-out 104-window test set, comprised the $n = 849$ windows evaluated below. Because this validation split is drawn from the same pool of events used for training, and given the window-level overlap described in Section II-C1, the metrics reported here should be interpreted as an upper-bound, in-distribution estimate of model performance rather than as evidence of generalization to unseen subjects or events. The float32 version achieved 95.5% accuracy and a ROC-AUC of 0.956. The int8-quantized version, used for embedded execution, achieved 95.2% accuracy and a ROC-AUC of 0.951, indicating that quantization introduced negligible performance degradation relative to the float32 version on the same validation split. Per-class metrics for the int8 model are summarized in Table II.

The confusion matrix (Fig. 8) shows that 504 of 528 non-fall samples and 304 of 321 fall samples were correctly classified. t-distributed Stochastic Neighbor Embedding (t-SNE) was used to visualize the learned feature space and

TABLE II
PER-CLASS PERFORMANCE METRICS – INT8 MODEL (INTERNAL VALIDATION SPLIT, $n = 849$)

Class	Precision	Recall	F1-score
Non-fall	97.5%	95.3%	0.964
Fall	92.5%	96.0%	0.942

assess class separability in a 2D representation, suggesting clear class separation, with misclassifications concentrated at the decision boundary, as shown in Fig. 9. Training and validation accuracy curves stabilized above 95% by epoch 20, indicating adequate convergence within 120 epochs, as shown in Fig. 10; the divergence between training and validation loss visible after approximately epoch 40 is consistent with mild overfitting, expected given the limited and non-independent dataset.

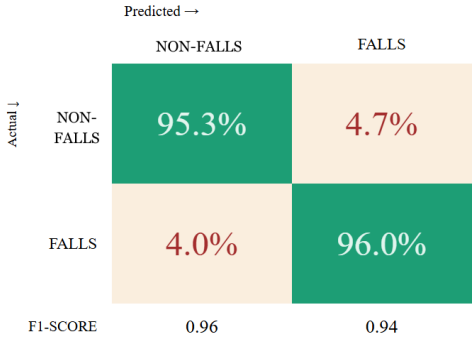


Fig. 8. Confusion matrix of the int8-quantized model evaluated on the internal validation split (849 samples).

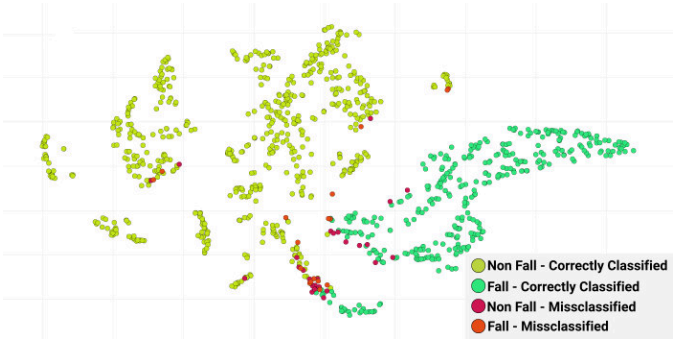


Fig. 9. t-SNE two-dimensional projection of the internal validation split showing class separability. Non-fall and fall samples are largely grouped in distinct clusters, with misclassified samples concentrated at the decision boundary.

B. Experimental Results – Fall Detection

Ten trials of each fall type were conducted with the device operating autonomously on the microcontroller, performed by the same two adult authors who also contributed to the training data. This sample size (10 trials per fall type) is too small to provide a statistically robust estimate of sensitivity, and the results below should be read as an indicative, not conclusive, characterization of on-device performance. Results are summarized in Table III.

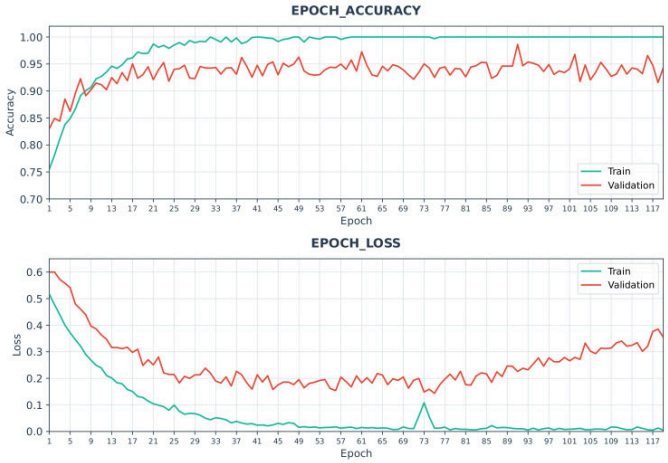


Fig. 10. Training and validation accuracy (top) and loss (bottom) curves over 120 epochs. Both metrics stabilize by epoch 20, with the divergence between training and validation loss indicating mild overfitting, expected given the limited dataset size.

TABLE III
FALL DETECTION PERFORMANCE IN PRACTICAL ON-DEVICE TESTS

Fall Type	Trials	Detected	Detection Rate
Forward	10	10	100%
Backward	10	8	80%
Lateral	10	7	70%

Forward falls were detected in all ten trials. Lateral falls showed the lowest detection rate (70%): in the three missed trials, wrist impact attenuation by the arm reduced peak acceleration below the Stage 2 threshold. Consequently, although the CNN correctly flagged a fall suspicion, the physical verification stage rejected it. This limitation appears to be associated with wrist-worn placement rather than with the classifier itself. These results suggest that a single, subject-independent acceleration threshold may not be adequate across all fall types, and two mitigation strategies are proposed for future work: (i) fall-type-adaptive thresholds, learned or calibrated separately for lateral falls, and (ii) a complementary sensing modality, such as a barometric pressure sensor, capable of detecting the altitude change associated with a real fall independently of arm-bracing behavior. One false positive was recorded during a rapid arm movement followed by immediate immobility on a surface, a motion that structurally mimics a fall signature and cannot be disambiguated from wrist-worn inertial data alone. The same barometric-sensing strategy could help distinguish this case, since an isolated arm gesture does not involve a change in whole-body height.

A qualitative comparison was made against a simple

threshold-only baseline (a fall was flagged whenever acceleration exceeds 3.80 g, with no temporal voting or post-impact verification). During the practical trials, high-intensity ADL such as running and jumping produced peak accelerations exceeding this threshold, which would have triggered false alerts under the baseline approach; the full pipeline correctly rejected these events at Stage 1 or Stage 3. However, sensitivity, specificity, and false-positive rate were not systematically computed over a dedicated test set, and this quantitative comparison is identified as a priority for future work rather than reported here as a complete result.

C. Computational Performance

Table IV summarises the embedded computational metrics observed during continuous autonomous operation. The system operated continuously during the practical test sessions without resets or critical failures. The 0.7% CPU load confirms that inference does not compromise concurrent tasks (MQTT communication, geofencing, and sensor acquisition).

The int8-quantized model itself, as estimated by the Edge Impulse platform, requires only 6.7 kB of peak RAM and 48.6 kB of Flash memory, with an estimated inference time of 6 ms. This compact footprint represents a small fraction of the total resources available on the M5Stack AtomS3R-Ext, demonstrating the effectiveness of int8 quantization in enabling TinyML deployment on resource-constrained microcontrollers.

TABLE IV
EMBEDDED COMPUTATIONAL PERFORMANCE – M5STACK ATOMS3R-EXT

Parameter	Value
Min. inference time	7 ms
Max. inference time	8 ms
Mean inference time	≈7.03 ms
Estimated ML CPU load	0.7%
Model peak RAM usage (int8)	6.7 kB
Model Flash usage (int8)	48.6 kB
Total SRAM (ESP32-S3)	512 kB
Typical free heap	≈243 kB
Firmware Flash usage	1,370,207 B (41%)
Static RAM usage	56,292 B (17.2%)

D. Geofencing Performance

In addition to fall detection, the Wi-Fi-based geofencing module was evaluated under realistic conditions, with the mobile application running in the background. This mechanism is limited to detecting the presence or absence of the preconfigured home Wi-Fi network and should not be interpreted as general spatial tracking or precise localization; it cannot indicate the wearer's position beyond the network's coverage area. The total time between the wearer leaving the configured perimeter and the caregiver receiving the notification was 21.78 s. This interval includes a 15 s confirmation window intentionally configured in the firmware to validate the disconnection and avoid false alerts caused by momentary signal loss, with the remaining time corresponding to the latency of the communication stack (MQTT publication, Firebase Cloud Messaging push delivery, and smartphone processing). The reconnection time, measured from the wearer's return to the perimeter until the status update in the application, was 4.45 s.

It is worth noting that both values are influenced by external factors, such as the quality of the local network infrastructure and the smartphone operating system's handling of background notifications, and therefore do not represent intrinsic limitations of the proposed embedded system.

IV. DISCUSSION

The experimental evaluation consisted of simulated fall tests performed by the two adult authors under controlled conditions, without the involvement of external participants, older adults, individuals with cognitive impairment, or clinical trials. Consequently, none of the findings reported here should be interpreted as evidence of clinical effectiveness. Although direct comparison is limited by differences in datasets and experimental protocols, the proposed system achieved 95.2% accuracy on its internal validation split, compared with the 92.31% reported by Volochtchuk and Bassai [11] using an Edge Impulse-based pipeline, while advancing to full autonomous embedded deployment. This comparison should nonetheless be read with caution, since neither study evaluated on a leakage-free, subject-independent test set. The 7 ms inference time and 0.7% CPU utilization demonstrate that 1D CNN inference is practical on a low-cost ESP32-S3 microcontroller, leaving ample headroom for other system tasks.

The wrist placement, while preferred for user acceptance, introduces inherent challenges: lateral falls produce lower-amplitude IMU signatures due to natural arm bracing, and rapid arm gestures can occasionally mimic fall dynamics. These limitations are consistent with findings from Nunez et al. [8], who reported that wrist sensors exhibit higher classification ambiguity than chest-mounted alternatives. Adaptive class thresholds or a complementary barometric pressure sensor could mitigate these issues in future work, since the altitude variation associated with a real fall would help distinguish it from isolated arm movements that do not involve a change in body height.

The system also integrates a Wi-Fi-based geofencing module, which detects whether the device remains within the configured home network perimeter and notifies the caregiver in case of a boundary breach, providing an additional monitoring functionality without requiring GPS hardware. As noted in Section III, this functionality is limited by Wi-Fi coverage and does not constitute general location tracking.

A. Lessons Learned and Research Gaps

Three practical lessons emerged from this work. First, window-level random splitting is inadequate for wearable fall-detection datasets built from short, repeated trials: with 75% window overlap, event- and subject-independent partitioning (e.g., leave-one-subject-out or leave-one-event-out) is necessary to obtain a trustworthy estimate of generalization, and

should be planned from the data-collection stage rather than applied post hoc. Second, a CNN classifier alone may be insufficient to suppress false positives from high-intensity, non-fall activities. The multi-stage physical verification pipeline used here shows that combining a learned classifier with interpretable, physics-based rules can provide an effective and computationally inexpensive strategy, though thresholds calibrated on a small development set risk overfitting to the specific subjects and activities recorded. Third, wrist placement, while potentially favorable for user acceptance, can weaken the acceleration signature of certain fall types (notably lateral falls) and may be vulnerable to arm-gesture false positives. These observations suggest that single-sensor wrist-only systems may need a complementary modality (e.g., barometric pressure) to reach clinically relevant sensitivity and specificity.

The main research gaps identified are: (i) the absence of a leakage-free, event- and subject-level evaluation of the trained model; (ii) the lack of data from the target population (older adults, particularly those with cognitive impairment), whose fall kinematics may differ from those of healthy young adults; (iii) the absence of quantitative sensitivity, specificity, and false-positive rate comparisons against baseline threshold methods over a common, dedicated test set; and (iv) the lack of usability, comfort, adherence, and battery-autonomy assessments, all of which are necessary before this proof-of-concept can be considered for real-world deployment with cognitively impaired users.

V. Conclusion

This paper presented a wrist-worn smartband combining fully embedded TinyML inference with a multi-stage physical verification pipeline for automatic fall detection. A 1D CNN quantized to int8 and deployed on an M5Stack AtomS3R-Ext reached 95.2% accuracy and a ROC-AUC of 0.951 on an internal validation split, with a 7 ms mean inference time and 0.7% CPU load, indicating that this class of on-device processing is computationally practical on a low-cost microcontroller. The study demonstrates a promising embedded TinyML proof of concept for wrist-based fall detection; however, its clinical validity has not been established and depends on: (i) leakage-free, event-level and subject-level evaluation on an independent test set; (ii) larger and more diverse cohorts; (iii) testing with older adults and individuals with cognitive impairment; (iv) longitudinal, real-world assessment; (v) battery-autonomy evaluation; and (vi) direct, quantitative comparison with baseline threshold methods and alternative sensor placements. The source code and embedded libraries are available at `https://github.com/rodbcl/Edge-AI-Smartband`.

Acknowledgment

The authors thank the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPq), the Araucária Foundation and the State Secretariat for Science, Technology and Higher Education of Paraná (SETI).

References

[1] World Health Organization, "Falls," 2021. [Online]. Available: www.who.int/news-room/fact-sheets/detail/falls
[2] Brazilian Ministry of Health, "Relatório nacional sobre a demência," 2024. [Online]. Available: www.gov.br/saude/pt-br/assuntos/noticias/2024/setembro/relatorio-nacional-sobre-a-demencia
[3] J. Bruhn et al., "Falls and fall risk in community-dwelling people with dementia: A systematic review," *BMC Geriatrics*, 2021.
[4] R. Yu et al., "Falls in people with Alzheimer's disease and mild cognitive impairment: A systematic review and meta-analysis," *Ann. Geriatr. Med. Res.*, 2022.
[5] J. M. Olmedo-Aguirre et al., "Remote healthcare for elderly people using wearables: A review," *Sensors*, 2022.
[6] R. N. Ferreira, N. F. Ribeiro, and C. P. Santos, "Fall risk assessment using wearable sensors: A narrative review," *Sensors*, vol. 22, no. 3, 2022.
[7] G. Marques et al., "Online fall detection using wrist devices," *Sensors*, 2023.
[8] J. O. Nunez et al., "Analysis of waist and wrist positioning wearable machine-learning models to detect falls," *Electronics Letters*, 2024.
[9] S. Jang et al., "An effective deep learning framework for fall detection," *J. Med. Internet Res.*, 2024.
[10] F. Li et al., "A decade of progress in wearable sensors for fall detection," *Sensors*, vol. 25, no. 7, 2025.
[11] A. V. L. Volochtchuk and S. W. Bassai, "Development of a sensor for fall monitoring in the elderly," B.Sc. Thesis, UTFPR, 2023.
[12] J. Lin et al., "MCUNet: Tiny deep learning on IoT devices," *arXiv*, 2020.
[13] Y. Wu et al., "Monitoring wandering behaviors of dementia patients using BLE-based localization systems," *arXiv*, 2024.
[14] L. C. Santos et al., "Identifying sensors-based parameters associated with fall risk in older adults," *BMC Geriatrics*, 2024.
[15] Edge Impulse, "Edge AI Platform," 2026. [Online]. Available: https://www.edgeimpulse.com/
[16] M5Stack, "AtomS3R-Ext," 2023. [Online]. Available: https://docs.m5stack.com/en/core/AtomS3R
[17] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proc. IJCAI*, 2016.

Surface Electromyography in Rehabilitation: Current Methods, Challenges, and Future Perspectives

1st Filipe Romero Martins
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
filipe.romero@ufpe.br
ORCID: 0009-0003-8232-9685

2nd Bruno da Silva Brito
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
brunosilva.brito@ufpe.br
ORCID: 0000-0001-5040-997X

3rd Amanda Maria da Conceição Perez
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
amanda.mariac@ufpe.br
ORCID:0000-0003-4149-8960

4th Emmanuel Andrade de Barros Santos
Automation and Control Department
University of Pernambuco
Recife, Brazil
emmanuel.andrade@upe.br
ORCID: 0000-0002-1786-0934

5th Patricia Silva Lessa
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
patricia.lessa@ufpe.br
ORCID: 0000-0002-4498-5724

6th Marco Aurelio Benedetti Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
marco.rodrigues@ufpe.br
ORCID: 0000-0002-5143-3696

Abstract —Surface electromyography (sEMG) is a non-invasive method for recording electrical activity associated with skeletal muscle activation. In rehabilitation, it supports functional assessment, therapeutic biofeedback, neuromuscular re-education, movement analysis, and recovery monitoring in neurological, musculoskeletal, and post-operative conditions. This review discusses current methodological principles, clinical applications, technical challenges, and future perspectives for the use of sEMG in physical therapy and rehabilitation science. The article emphasizes acquisition procedures, electrode placement, signal conditioning, normalization, interpretation, and integration with wearable systems, artificial intelligence, telerehabilitation, robotics, and virtual environments. Although sEMG offers relevant information regarding muscle synchronization, symmetry, fatigue, and motor control strategies, its clinical translation depends on standardized protocols, adequate professional training, transparent reporting, and careful interpretation regarding functional outcomes. The future of sEMG in rehabilitation is promising, especially when combined with multimodal assessment, patient-centered biofeedback, and reproducible digital health workflows.

Keywords— surface electromyography; rehabilitation; physical therapy; biofeedback; wearable sensors; signal processing; motor control.

I. INTRODUCTION

Surface electromyography (sEMG) has consolidated itself as an essential biophysical tool in sports science and rehabilitation, allowing for the non-invasive assessment of neuromuscular activity and the physiological mechanisms underlying motor performance [1], [2]. By capturing motor unit action potentials through the skin, sEMG offers a direct reflection of the central command sent to the muscles, making it indispensable for investigating coordination, muscle recruitment, and training adaptations [2], [3].

In clinical rehabilitation settings, monitoring neuromuscular alterations is fundamental for understanding functional deficits resulting from stroke, spinal cord injury, or post-operative musculoskeletal disorders [7]. A critical phenomenon evaluated by sEMG is muscle fatigue and activation timing, which directly reflects the neuromuscular system's capacity to sustain effort or coordinate movement during therapeutic exercises [4], [5]. Identifying these electromyographic changes allows physical therapists to tailor load progression, prevent secondary compensations, and optimize functional motor recovery [1], [2].

Standard fatigue and activation assessments traditionally rely on time-domain amplitude descriptors, such as the root mean square (RMS), and spectral frequency parameters, including mean and median frequencies [1], [5]. However, translating these parameters into routine physical therapy faces significant clinical and methodological challenges, such as signal crosstalk, anatomical variability, and the lack of standardized normalization protocols for patients with pain or neurological impairments [3], [6]. While advanced signal processing and machine learning approaches offer new opportunities to map motor control, their success critically depends on reproducible methodologies validated in clinical populations [7], [10].

This systematic review aims to evaluate the clinical application and methodological validity of surface electromyography (sEMG) in rehabilitation. Specifically, this study addresses the following PICO question: In patients undergoing physical rehabilitation (P), how does sEMG-assisted assessment or biofeedback (I) compare to conventional therapeutic evaluations or standard care (C)

in improving neuromuscular control and functional outcomes (O)?

II. BASIC PRINCIPLES OF SURFACE ELECTROMYOGRAPHY

Skeletal muscle contraction depends on action potentials generated by the activation of alpha motor neurons. When a motor unit is recruited, electrical activity propagates along the fibers, and surface electrodes detect the summation of these potentials. The recorded sEMG signal is an indirect representation of neuromuscular activation, highly influenced by physiological, anatomical, and technical factors [1], [2], [5]. The raw signal is generally bipolar, oscillating around a baseline. Its amplitude is affected by the number and firing rate of motor units, distance between fibers and sensors, tissue impedance, subcutaneous fat, and interference from adjacent muscles (cross-talk), while the frequency depends on conduction velocity and fatigue [1], [5].

Thus, sEMG does not directly measure isolated force, and high amplitudes do not necessarily indicate better muscle function; they can reflect compensations or inefficient motor strategies [1], [5]. Conversely, reduced amplitudes may signal weakness, inhibition, or greater technical efficiency. In rehabilitation, variables such as synchronization, symmetry, and mutual coordination are clinically more informative than isolated peaks, allowing for the assessment of, for example, quadriceps lag post-ligament reconstruction or gait asymmetry [1]. Because it is a non-invasive, accessible, and dynamic method, it applies to sports and ergonomic gestures but demands technical rigor. International guidelines, such as the SENIAM recommendations, remain fundamental for standardizing sensor placement and signal processing [1], [2], [3].

III. ACQUISITION METHODS AND ELECTRODE PLACEMENT

Quality in signal acquisition determines the validity of its interpretation. Sensors must be fixed over the muscle belly, parallel to the fibers, and respecting an adequate inter-electrode distance, after hair shaving, mild abrasion, and alcohol cleaning to mitigate impedance [3], [6]. Disposable Ag/AgCl bipolar configurations are the gold standard due to their stability and common-mode rejection, although textile and dry sensors are gaining ground in home-based wearable systems [5]. Placement requires caution in deep or overlapping structures (such as the forearm and paraspinal region) due to the risk of cross-talk [3], [6].

Dynamic tasks introduce artifacts from sweat and cable sway. Wireless devices minimize these impacts, requiring strict control over data packet loss and sampling rates (frequently above 1000 Hz) to record the entire physiological bandwidth [5]. Standard filtering utilizes band-pass filters to attenuate motion artifacts and high-frequency noise, associated with notch filters to eliminate power line interference (60 Hz), avoiding aggressive cut-offs that distort the signal phase [5]. For reliable longitudinal assessments, the clinician must meticulously document posture, load, and anatomical location, resorting to photographic records to prevent placement variations from masking real clinical progress [3], [6].

IV. SIGNAL PROCESSING AND INTERPRETATION

The raw signal initially undergoes a visual inspection for the exclusion of artifacts or saturation before being filtered, rectified, smoothed, and segmented [4], [5]. The most common time-domain descriptors are the root mean square (RMS), mean rectified value, and integrated EMG, while spectral descriptors (mean and median frequencies) are the basis for fatigue monitoring [4], [5]. The RMS reflects signal power in time windows whose durations balance time resolution and profile smoothness [1], [5].

Signal normalization is widely debated in the literature [3], [13], [14]. Because raw amplitudes vary due to non-physiological factors, Maximum Voluntary Isometric Contraction (MVIC) serves as the traditional reference, but fails in patients limited by pain or neurological dysfunctions. Alternative methods comprise submaximal reference contractions, peaks from the task itself, or dynamic normalizations [4], [5]. Furthermore, establishing sEMG assessment consistency across dynamic lower-limb tasks, such as vertical countermovement jumps, remains critical for ensuring measurement reliability in functional assessments [15]. Spectral analysis detects fatigue by shifting the spectrum toward lower frequencies due to the reduction in fiber conduction velocity, although in dynamic tasks, this interpretation suffers interference from variations in joint angles [1].

Machine Learning approaches have been proposed to classify movement intentions and predict fatigue [10]. By extracting features in the time and frequency domains, these models require rigorous internal validation processes and generalization tests to mitigate overfitting to specific training data [10]. For success in rehabilitation, however, the clinical interpretability of these algorithms and the integration of the sEMG signal into the patient's functional scenario—such as correlating hyperactivities with pain or dyskinesias—remain indispensable [1], [10].

V. RESULTS

A. Study Selection Results

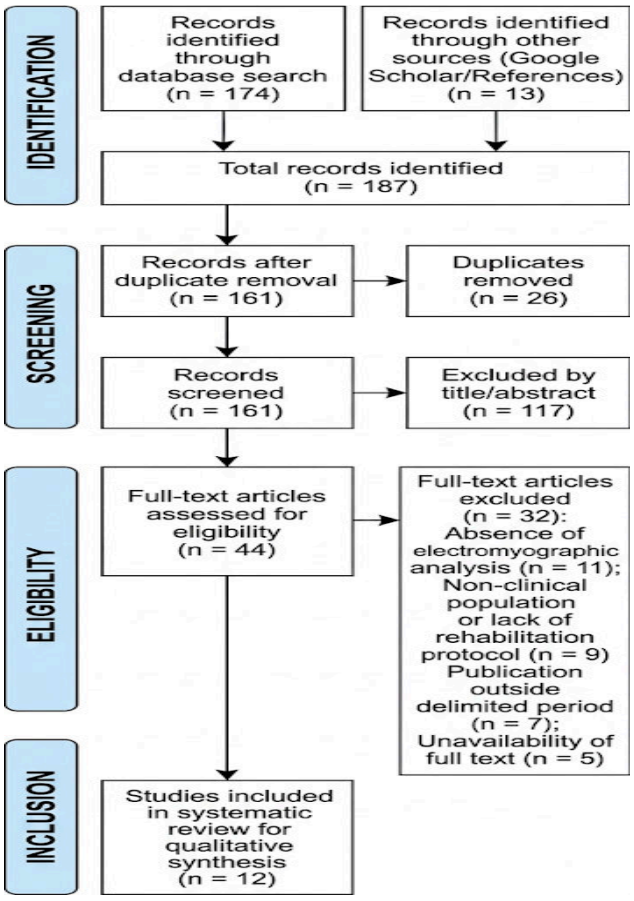
The systematic search was conducted across six electronic databases (PubMed, SciELO, Frontiers, MDPI, Nature, and JSSM) using Medical Subject Headings (MeSH)

terms combined with Boolean operators for 'surface electromyography', 'rehabilitation', and 'neuromuscular control' [4], [11]. Methodological quality and risk of bias were independently evaluated by two reviewers using the PEDro scale for randomized trials and the ROBINS-I tool for non-randomized intervention studies, with consensus resolving any discrepancies [4], [12].

The search and article selection process followed the recommendations of the PRISMA 2020 protocol, as detailed in Flowchart 1. Initially, the automated search across the six selected databases (PubMed, SciELO, Frontiers, MDPI, Nature, and JSSM) resulted in 174 records. Additionally, 13 studies were identified through other sources, such as reference lists and Google Scholar, totaling 187 identified records.

Of the remaining 44 articles that advanced to full-text evaluation, 32 were excluded for the following specific reasons: absence of electromyographic analysis (n = 11); non-clinical population or lack of a rehabilitation protocol (n = 9); publication outside the delimited period (n = 7); and unavailability of the full text (n = 5). At the end of this process, 12 studies met the eligibility criteria and were included for qualitative synthesis.

Due to the high heterogeneity of protocols, evaluated muscles, and metrics used among the selected studies, conducting a statistical meta-analysis with a pooled effect size was not feasible.



VI. APPLICATIONS IN CLINICAL REHABILITATION

A. Neurological Rehabilitation

Dysfunctions resulting from Stroke, spinal cord injury, cerebral palsy, and Parkinson's disease severely alter patterns of coordination and selective motor control. sEMG operates by identifying whether functional restriction stems from insufficient recruitment, spasticity, or anomalous co-contractions [7]. In post-stroke patients, it assists in mapping dynamic asymmetries and muscle timing during the stance and swing phases of gait, supporting adjustments in orthoses and Functional Electrical Stimulation (FES) parameters [7].

sEMG biofeedback converts invisible electrical activity into real-time visual or auditory signals, providing concrete goals for patients with reduced proprioception, which accelerates motor learning and spasm relaxation [7]. Currently, robotic interfaces and exoskeletons utilize sEMG as an input signal to trigger movements based on user intent, directly stimulating neuroplasticity, provided they operate with low latency and robust filters against daily signal fluctuations [10].

B. Musculoskeletal and Sports Rehabilitation

In orthopedic dysfunctions, sEMG detects muscle inhibitions and protective guarding patterns. Post-anterior cruciate ligament reconstruction, it monitors quadriceps restoration and compensatory co-contractions during squatting and jumping, defining safe criteria for load progression [1]. In shoulder rehabilitation, it quantifies activation ratios between scapular stabilizers and the rotator cuff, allowing for the selection of lower-impact exercises [1]. In chronic low back pain conditions, biofeedback guides reconditioning and alleviates protective hyperactivity of the trunk [1]. In the sports sphere, wearable inertial sensors associated with sEMG enable kinematic and neuromuscular analyses directly on the field, allowing for the monitoring of asymmetries and predicting injury risks during complex sports gestures (such as running and throwing) outside the controlled laboratory environment [8].

C. Biofeedback and Motor Learning

The sEMG biofeedback transforms the signal into sensory stimuli that optimize muscle regulation. Its efficacy depends on controlled delivery: continuous feedback can induce dependence, requiring fading techniques (gradual reduction) to consolidate long-term retention [7]. Wearable technologies and integrated mobile applications extend these practices to the home environment, promoting greater therapeutic adherence [7], [9]. Furthermore, the insertion of virtual and gamified environments substantially elevates patient engagement by converting contractions and relaxations into interactive, playful commands, while addressing technological challenges related to usability, safety, and system latency [7].

VII. CHALLENGES, LIMITATIONS, AND FUTURE PERSPECTIVES

Despite its broad potential, sEMG exhibits critical limitations. The signal demonstrates extreme sensitivity to millimeter-level variations in placement, adipose tissues, and cross-talk, requiring rigorous documentation to enable longitudinal follow-up [3], [5]. The absence of a universal protocol limits direct comparisons, as established methodologies for healthy individuals rarely apply to

clinical populations with severe pain conditions [4], [6]. Another obstacle lies in the mistaken interpretation of amplitude as a direct synonym for force, disregarding factors of motor inefficiency or fear [1], [5]. Additionally, the high cost of equipment and the need for deep technical training in signal processing reduce its dissemination in daily clinical practice [4]. Continuous home monitoring via wearable sensors also imposes new ethical challenges regarding governance, privacy, and the security of physiological data [11], [12].

Future perspectives converge toward hardware miniaturization and the adoption of flexible and textile electrodes, mitigating motion artifacts and optimizing comfort during prolonged monitoring [9]. More than the raw time-series amplitude which currently holds limited standalone clinical value due to noise and impedance variability the true potential of modern sEMG lies in signal transformation into the frequency and time-frequency domains (e.g., via Fast Fourier Transform and Wavelet transforms). Spectral parameters, such as shifts in median (Fmed) and mean (Fmean) frequencies, serve as fundamental physiological biomarkers to objectively quantify neuromuscular fatigue, motor unit firing behavior, and fiber recruitment patterns well before structural performance deteriorates. Artificial intelligence algorithms stand out in supporting complex feature extraction and identifying subtle changes in muscle coordination during repetitive tasks, provided they are validated on representative data and structured comprehensibly for clinicians [10]. The consolidation of multimodal systems (sEMG combined with inertial and pressure sensors) provides a complete biomechanical mapping and drives telerehabilitation, allowing for remote and objective supervision of exercise routines [7], [8]. Finally, open science practices and the sharing of code pipelines and public databases are essential to mitigate literature heterogeneity and accelerate the clinical translation of these technologies, ensuring that innovations remain focused on functional gain and patient safety [11].

VIII. CONCLUSION

Surface electromyography is a valuable method for assessing and training neuromuscular function in rehabilitation. Its non-invasive nature, compatibility with movement, and potential for real-time feedback make it useful in neurological, musculoskeletal, sports, and post-operative contexts. It can support clinical reasoning by revealing activation synchronization, intensity, symmetry, fatigue, co-contraction, and motor control strategies that are not always visible through observation alone [1], [2].

At the same time, sEMG is not a simple or self-explanatory measurement. Its clinical value depends on careful acquisition, standardized protocols, adequate processing, cautious normalization, and contextual interpretation. Challenges such as crosstalk from adjacent muscles, motion artifacts, variability in electrode placement, limited standardization, and overinterpretation must be addressed through training and transparent reporting [3], [5], [6].

Future developments in wearable sensors, artificial intelligence, telerehabilitation, robotics, and virtual environments are driving sEMG beyond traditional laboratories and specialized clinics. Crucially, this transition relies on advancing beyond static, offline signal evaluation toward **real-time frequency-domain analysis**. Instantaneous spectral processing (e.g., real-time shifts in median and mean frequencies) allows for the continuous monitoring of motor unit fatigue, recruitment dynamics, and muscle coordination as movement happens. This enables immediate biofeedback and dynamic system adaptation in clinical and telerehabilitation settings. Ultimately, the most promising direction lies in patient-centered integration: systems that leverage real-time spectral metrics to help clinicians prescribe optimized exercises, mitigate injury risk, and enhance functional recovery. When used responsibly, surface electromyography becomes an intelligent bridge between neuromuscular physiology, real-time digital health, and functional rehabilitation [7], [9], [10].

TABLE I. CLINICAL USES OF sEMG IN REHABILITATION

Clinical Objective	sEMG Contribution	Example in Rehabilitation
Assessment	Quantifies activation synchronization, amplitude, symmetry, and fatigue	Comparing quadriceps activation after knee surgery
Biofeedback	Transforms muscle activity into real-time visual or auditory cues	Training upper trapezius relaxation in cervicalgia
Motor Control	Reveals co-contraction, delayed onset, or compensatory recruitment	Analyzing ankle dorsiflexor timing post-stroke
Exercise Selection	Compares muscle demand between therapeutic exercises	Selecting scapular stabilization exercises
Remote Monitoring	Supports home-based follow-up when integrated with wearable devices	Monitoring adherence and activation quality during telerehabilitation

ACKNOWLEDGMENTS

The author acknowledges the use of AI-based tools for assistance in drafting, organization, linguistic revision, and formatting support during the preparation of this manuscript. The author remains responsible for verifying the accuracy of the content, references, and final submission requirements.

REFERENCES

[1] C. J. De Luca, "The use of surface electromyography in biomechanics," Journal of Applied Biomechanics, vol. 13, no. 2, pp. 135–163, 1997.

[2] R. Merletti and P. A. Parker, Electromyography: Physiology, Engineering, and Non-Invasive Applications. Hoboken, NJ, USA: Wiley-IEEE Press, 2004.

[3] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for SEMG sensors and sensor placement procedures," Journal of Electromyography and Kinesiology, vol. 10, no. 5, pp. 361–374, 2000.

[4] E. F. F. da Silva Junior et al., "Surface electromyography signal processing and evaluation on respiratory muscles of critically ill patients: A systematic review," PLOS ONE, vol. 18, no. 5, e0284911, 2023.

[5] R. Merletti, A. Botter, A. Troiano, E. Merlo, and M. A. Minetto, "Technology and instrumentation for detection and conditioning of the surface electromyographic signal: State of the art," Clinical Biomechanics, vol. 24, no. 2, pp. 122–134, 2009.

[6] SENIAM, "European Recommendations for Surface Electromyography," SENIAM Project. [Online]. Available: https://seniam.org/

[7] T. Bowman et al., "Wearable devices for biofeedback rehabilitation: A systematic review and meta-analysis to design application rules and estimate the effectiveness on balance and gait outcomes in neurological diseases," Sensors, vol. 21, no. 10, 3444, 2021.

[8] M. A. Khan et al., "Real-time wearable biomechanics framework for sports injury risk assessment and rehabilitation tracking," Scientific Reports, 2025.

[9] Y. Liu et al., "Recent advances in flexible noninvasive electrodes for surface electromyography," npj Flexible Electronics, 2023.

[10] M. Zheng, M. S. Crouch, and M. S. Eggleston, "Surface electromyography as a natural human-machine interface: A review," arXiv:2101.04658, 2021.

[11] IEEE Author Center, "IEEE conference template and author guidelines." [Online]. Available: https://ieeeauthorcenter.ieee.org/

[12] International Committee of Medical Journal Editors, "Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals." [Online]. Available: https://www.icmje.org/

[13] M. Besomi et al., "Consensus for experimental design in electromyography (CEDE) project: Amplitude normalization matrix," Journal of Electromyography and Kinesiology, vol. 53, p. 102438, 2020.

[14] M. Halaki and K. Ginn, "Normalization of EMG signals: To normalize or not to normalize," in Computational Intelligence in Electromyography Analysis - A Perspective on Current Applications and Future Challenges, G. R. Naik, Ed. Rijeka, Croatia: InTech, 2012, pp. 175–194.

[15] C. Rodrigues, M. Correia, J. M. C. S. Abrantes, J. Nadal, and M. A. B. Rodrigues, "Consistency of surface electromyography assessment at lower limb selected muscles during vertical countermovement," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, South Korea, 2017, p. 402.

Tecnologias móveis de alinhamento, controle postural e equilíbrio na reabilitação neurológica

Grazielle Souza Ferreira da Silva
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
Graziellesfsilva@gmail.com

Tabajara de Oliveira Gonzalez
Núcleo de Pesquisas Tecnológica
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
tabajaragonzalez@umc.br

Terigi Augusto Scardovelli
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
terigiscardovelli@umc.br

Alessandro Pereira da Silva
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
alessandrops@umc.br

Silvia Cristina Martini
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
silviac@umc.br

Silvia Regina Matos da Silva Boschi
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
boschi@umc.br

***Abstract*— The aim of this study was to systematize the current evidence of software, mobile devices and applications that promote assessment and intervention in postural misalignments, balance and postural control with a focus on individuals with neurological impairment. The study is a systematic review of studies published between 2015 and 2025 through the Cochrane Library, LILACS, PEDro, PubMed and Web of Science databases, using the descriptors posture, postural control, postural balance, postural correction, sitting posture, supine position, orthostatic position, neurological disorder, nervous system diseases, neurological rehabilitation, mobile applications and biomedical technology, following the PRISMA guidelines and registered in PROSPERO (CRD420261443761), with the data summarized in a descriptive table. Meeting the search requirements resulted in 1,089 studies, however, complying with the inclusion and exclusion criteria, 15 studies remained and had their results analyzed. Mobile technologies and sensor systems offer potential for the assessment, monitoring, and rehabilitation of postural control, alignment, and balance in individuals with neurological conditions. However, available evidence remains limited and heterogeneous regarding populations, devices, sensors, operating systems, tasks, and outcomes, making it difficult to draw clinically applicable conclusions. Thus, although some studies indicate favorable results concerning the validity, reliability, and feasibility of these technologies, further research characterized by greater methodological rigor, standardized protocols, and adequate sample sizes is still needed to establish their clinical effectiveness in neurological rehabilitation.**

Keywords*— *neurological rehabilitation, diseases of the nervous system, postural control, balance, mobile applications.

I. INTRODUÇÃO

De acordo com a Organização Mundial da Saúde (OMS), os distúrbios neurológicos constituem a principal causa de incapacidade e a segunda principal causa de mortalidade em todo o mundo [1]. Essas condições afetam o sistema nervoso central e/ou periférico, comprometendo significativamente o desempenho motor, sensorial, cognitivo e funcional dos indivíduos [2,3].

As doenças neurológicas envolvem a redução do controle motor, devido a degeneração ou lesão do sistema nervoso e neuromuscular, gerando déficits como perda de coordenação e a dificuldade em manter a postura e a marcha, assim são necessárias intervenções fisioterapêuticas voltadas para o alinhamento postural utilizando métodos ativos com o objetivo de corrigir deformidades da coluna e desalinhamentos corporais, ou ainda prevenir posturas inadequadas. Nesses tratamentos, o indivíduo é estimulado a realizar movimentos autocorretivos, exigindo consciência corporal e percepção postural [4,5,6,7,8,9].

O uso de tecnologia digital na avaliação do alinhamento, estabilidade postural e equilíbrio podem melhorar a precisão e a exatidão da medição e reduzem a variabilidade associada às diferenças na forma como os examinadores contabilizam os erros [10]. Uma solução emergente para superar as barreiras das intervenções domiciliares não supervisionadas é a implementação da tecnologia digital [11].

Com o uso de sensores inerciais (acelerômetros e giroscópios), métricas que compreendem o movimento de um indivíduo em aceleração no centro de massa demonstraram ter a capacidade de identificar significativamente o aumento da oscilação postural. Possíveis razões incluem maior tempo de configuração ou requisitos de processamento de dados *offline*, os quais interrompem o tempo e o fluxo de trabalho clínicos sempre limitados. Se comprovada a precisão, do uso de dispositivos móveis que contenham sensores inerciais para quantificar a oscilação postural, podem fornecer uma alternativa prática e de baixo custo às avaliações clínicas, substituindo plataformas de força e sistemas de análise óptica de movimento na quantificação da estabilidade postural, alinhamento postural e equilíbrio [9,11,12,13,14].

Assim devido à escassez de pesquisas acadêmicas a respeito da utilização de *softwares*, dispositivos móveis e aplicativos com enfoque no alinhamento postural, controle postural e equilíbrio na reabilitação neurológica, o presente estudo visa identificar através de uma revisão sistemática da literatura as tecnologias móveis utilizadas na reabilitação neurológica com enfoque no alinhamento postural, controle postural e equilíbrio.

II. METODOLOGIA

A busca iniciou-se em maio de 2025 e encerrou-se em outubro de 2025, e compreendeu os anos de 2015 a 2025, por meio das bases de dados Cochrane Library, LILACS, PEDro, PubMed e Web of Science, utilizando as seguintes palavras-chave de acordo com o DeCS (Descritores em Ciência da Saúde): *posture*, *postural balance*, *postural control*, *postural equilibrium*, *posture balance*, *sitting position*, *supine position*, *standing position*, *nervous system diseases*, *neurological rehabilitation*, *mobile applications* e *biomedical technology*, os geradores booleanos utilizados

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

para a busca foram AND e OR. A estratégia de busca utilizada para todas as bases de dados estão demonstradas na tabela 1.

Tabela 1: Estratégia de busca dos artigos selecionados.

Bases de dados	Estratégia de Busca
Cochrane Library, LILACS, PEDro, PubMed e Web of Science	((((((((((posture) OR (postural balance)) OR (postural control)) OR (postural equilibrium)) OR (posture balance)) OR (sitting position)) OR (supine position)) OR (standing position)) AND (nervous system diseases)) AND (neurological rehabilitation)) AND (mobile applications)) AND (biomedical technology)

Seguindo os requisitos de inclusão foram considerados os estudos no idioma inglês, disponíveis na íntegra, que atendam os seguintes critérios: (1) população de indivíduos com diagnóstico ou comprometimento relacionado a condição neurológica; (2) utilização de tecnologias móveis ou sistemas tecnológicos portáteis para avaliação, monitoramento, feedback ou reabilitação do equilíbrio, controle postural, alinhamento postural, marcha ou função motora; (3) utilização de smartphones, tablets, dispositivos móveis, sensores inerciais, acelerômetros, giroscópios, inclinômetros, câmeras, sensores de profundidade ou plataformas tecnológicas portáteis; e (4) estudos que apresentam resultados relacionados à validade, confiabilidade, validação, usabilidade, desempenho, monitoramento ou eficácia da tecnologia.

Os critérios de exclusão englobam estudos realizados exclusivamente com indivíduos saudáveis, estudos que não apresentassem relação direta com avaliação, monitoramento ou reabilitação do controle postural e equilíbrio, estudos nos quais a tecnologia não fosse utilizada como instrumento de avaliação ou intervenção de interesse, revisões de literatura, editoriais, cartas ao editor, protocolos, trabalhos sem texto completo disponível e estudos duplicados.

Para seleção dos estudos seguiu-se as diretrizes PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), para melhor clareza e transparência do processo de revisão. O processo de seleção e análise dos artigos seguiu uma metodologia rigorosa, contando com a participação de dois pesquisadores independentes para seleção dos estudos. Inicialmente, realizou-se uma triagem com base nos títulos, seguida por uma análise mais aprofundada dos resumos e, posteriormente, a leitura completa dos textos selecionados.

Os artigos selecionados foram avaliados quanto a sua qualidade metodológica por dois revisores pelas escalas ROB 2 (*Risk of Bias Cochrane 2*) que avalia ensaios clínicos randomizados ou pela escala COSMIN (*COnsensus-based Standards for the selection of health Measurement INstruments*) que rastreia o risco de viés e descreve os elementos que auxiliam na construção de uma questão de pesquisa e fornece uma visão abrangente dos componentes da medida de resultados assim como os requisitos de projeto e métodos estatísticos preferidos sobre a confiabilidade e erro de medida dos instrumentos [15,16, 17]. Este protocolo foi devidamente registrado no Registro Prospectivo Internacional de Revisões Sistemáticas (PROSPERO), sob o número de registro: CRD420261443761.

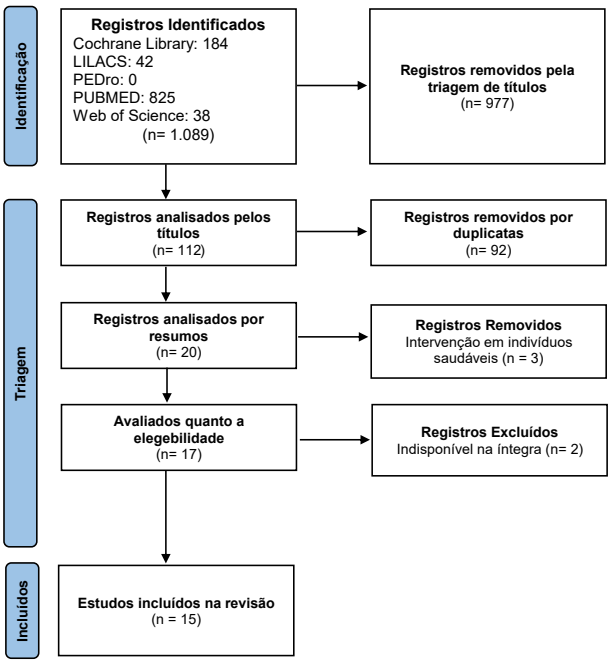
A síntese dos principais achados foi analisada por dois pesquisadores e descrita de forma qualitativa. A análise dos dados foi organizada em uma tabela com identificação de autores, condição neurológica, tamanho amostral, tecnologia utilizada, dispositivo e sistema operacional (Tabela 2), e outra contendo sensores empregados, posicionamento do sensor, comparador, finalidade, e estágio de validação (Tabela 3).

III. Resultados e Discussão

Foram encontrados nas bases de dados 1.089 estudos, destes 825 foram encontrados na PubMed, 184 estudos na Cochrane Library, 42 na base de dados LILACS, 38 no *Web of Science* e 0 na base de dados PEDro.

Obedecendo aos critérios de inclusão e coerência com o tema do trabalho foram excluídos 977 estudos pela triagem inicial pela análise de títulos, permanecendo 112 estudos. Desses, 92 foram excluídos devido a duplicatas, permanecendo 20 estudos para análise pelos resumos, a análise foi realizada por dois avaliadores, assim dentre destes foram excluídos 3 por se tratarem de intervenção em indivíduos saudáveis, permanecendo assim 17 estudos para análise completa dos textos, onde 2 deles foram excluídos por indisponibilidade na íntegra. Desse modo, permaneceram-se 15 estudos (Figura 1).

Figura 1: Fluxograma PRISMA da revisão sistemática, evidenciando a forma de seleção dos artigos.



Com a finalidade de abordar detalhadamente os estudos científicos selecionados, foi elaborada uma tabela descritiva elucidando autores, condição neurológica e tamanho amostral, tecnologia, dispositivo e sistema operacional (Tabela 2).

Tabela 2: Descrição da população e condição neurológica investigada.

Autor	*Condição Neurológica*	*n*	*Tecnologia utilizada*	*Dispositivo*	*Sistema operacional*
Ozinga *et al.* [18]	DP	34	CC-BApp	iPad 3ª geração	iOS
Ozinga *et al.* [19]	DP	28	Acelerometria	Smartphone	iOS

Ozinga *et al.* [20]	DP	54	Cleveland Clinic Balance App	iPad	iOS
Fiems *et al.* [21]	DP	30	Sway Balance™	iPod Touch 5ª geração	iOS
Hsieh *et al.* [22]	Síndrome de Down	39	Sistema móvel inteligente	iPod Touch 4ª geração + Arduino	iOS
Arcuria *et al.* [23]	Ataxia cerebelar	120	APP-Coo-Balance-Test	Smartphone Galaxy J3	Android
Linder *et al.* [24]	TCE leve	50	Paradigma móvel de dupla tarefa	iPad + Arduino	iOS
Serra-Añó *et al.* [25]	Doença de Alzheimer	62	FallSkip®	Xiaomi Redmi 4X	Android
Hou *et al.* [26]	AVC	27	Sistema de avaliação de equilíbrio por smartphone	ASUS Zenfone 3	Android
Shin [27]	AVC	24	Sistema Bal-90 com feedback visual	iPhone 5 + Bal-90	iOS
Summa *et al.* [28]	Ataxia de início precoce	20	SaraHome	Kinect v2 + Leap Motion	C#; SO Windows 10 pro
Aphiphaksakul *et al.* [29]	AVC ≤3 meses	32	Compass Inclinometer + disco de equilíbrio	Smartphone + disco de equilíbrio	Android
Rhea *et al.* [30]	Histórico de concussão	216	AccWalker	Moto X 2ª geração / Google Pixel	Android
Tirosh *et al.* [31]	Concussão aguda	22	TelePhysio	Smartphone	Android
Calcagni *et al.* [32]	Esclerose múltipla	164	NeuFun-TS	Google Pixel XL/2X	Android
Legenda: "DP" = Doença de Parkinson; "AVC" = Acidente Vascular Cerebral; "TCE" = Traumatismo Cranioencefálico					

Foi elaborada uma tabela descritiva com enfoque nos sensores empregados, posicionamento do sensor, comparador, finalidade e estágio de validação (Tabela 3).

Tabela 3: Tecnologias apresentadas nos estudos.

Autor	***Sensores empregados***	***Posicionamento do sensor***	***Comparador***	***Finalidade***	***Estágio de validação***
Ozinga *et al.* [18]	IMU	Lombar	Captura de movimento 3D	Estabilidade postural	Validade de medida
Ozinga *et al.* *[19]*	Acelerômetro	Sacro	Smart Balance Master	Avaliar instabilidade postural	Validade e confiabilidade
Ozinga *et al.* [20]	Acelerômetro + giroscópio	Sacro	Controle saudável	Avaliar estabilidade postural	Validação de medida
Fiems *et al.* [21]	IMU	Lombar e esterno	IMU	Mensurar oscilação postural	Confiabilidade
Hsieh *et al.* [22]	Acelerômetro	Sacro	Plataforma de força	Detectar disfunções de equilíbrio	Validade
Arcuria *et al.* [23]	Acelerômetro triaxial	Esterno	Plataforma TechnoBody Pro-Kin-212-N	Quantificar equilíbrio e oscilação do tronco	Desenvolvimento e validação
Linder *et al.* [24]	Acelerômetro e giroscópio	Sacro	NR	Avaliar estabilidade postural e função cognitiva	Desenvolvimento e validação
Serra-Añó *et al.* [25]	IMU de 6 eixos	Altura da crista ilíaca - sacro	Controles sem demência	Avaliar mobilidade, equilíbrio e marcha	Validação funcional
Hou *et al.* [26]	Acelerômetro e giroscópio	Sacro	Escala de Equilíbrio de Berg	Avaliar desempenho do equilíbrio	Confiabilidade e validade
Shin [27]	giroscópio	Prancha de equilíbrio	Grupo controle do ensaio clínico randomizado	Intervenção/ reabilitação do controle de tronco	Eficácia terapêutica
Summa *et al.* [28]	Sensores de captura de movimento	Externo por kinect	Avaliação assistida pelo cuidador	Avaliação remota e objetiva	Desenvolvimento e viabilidade
Aphiphaksakul *et al.* [29]	Inclinômetro/sensor de orientação do smartphone	Região central do tronco	Condição/intervenção comparadora	Intervenção domiciliar com feedback visual	Eficácia terapêutica
Rhea *et al.* [30]	Sensores embarcados do smartphone	Coxa	Participantes sem histórico de concussão	Detectar disfunção neuromotora	Desenvolvimento e validação
Tirosh *et al.* [31]	Sensor de aceleração	Lombar	Controle saudável	Avaliação remota e monitoramento da recuperação	Viabilidade validação inicial
Calcagni *et al.* [32]	Sensores do smartphone	Mãos	Controles e diferentes grupos neurológicos	Biomarcador digital remoto	Validação clínica

Dentre os 15 estudos muitos desfechos obtiveram dados que são considerados significativos, o nível de evidência deve ser levado em consideração, tendo em vista a ponderação dos desfechos com o risco de viés dos artigos analisados. Dos 15 artigos selecionados, 2 eram ensaios clínicos randomizados e os demais eram ensaios clínicos não randomizados.

Os ensaios clínicos randomizados foram analisados pela ferramenta ROB 2, ambos os estudos foram classificados como moderado risco de viés [29, 27]. Os problemas metodológicos mais encontrados envolveram à randomização, desvios da intervenção pretendida e na seleção dos resultados relatados (Figura 2).

Figura 2: Avaliação de Risco de Viés 2 (ROB 2).

Autor	**Randomização**	**Desvios da Intervenção**	**Dados Faltantes**	**Mensuração do Desfecho**	**Relato Seletivo**	**Julgamento Global**
Aphiphaksakul; Siriphorn [29]						
Shin [27]						

Os estudos que tinham como objetivo avaliação dos requisitos de projeto e métodos estatísticos preferidos sobre

a confiabilidade e erro de medida dos instrumentos foram classificados pela escala COSMIN (Figura 3).

Figura 3: Avaliação de Risco de Viés (COSMIN).

Estudo	Estabilidade	Intervalo entre medidas	Condições semelhantes	Cegamento do avaliador	Cegamento na atribuição	Limitações metodológicas	ICC/Kappa	Validade	Classificação Final
Ozinga *et al.* [18]	±	?	±	?	?	±	+	+	A
Ozinga *et al.* [19]	±	?	+	?	?	±	+	+	A
Ozinga *et al.* [20]	±	?	+	?	?	±	+	+	A
Fiems *et al.* [21]	+	+	+	+	+	+	+	+	O
Hsieh *et al.* [22]	±	?	+	?	?	±	±	+	A
Arcuria *et al.* [23]	±	?	+	?	?	±	±	+	A
Linder *et al.* [24]	±	?	+	?	?	±	±	±	A
Serra-Añó *et al.* [25]	±	?	+	?	?	±	+	+	A
Hou *et al.* [26]	+	+	+	?	?	±	+	+	O
Summa *et al.* [28]	?	?	+	?	?	±	?	±	D
Rhea *et al.* [30]	±	?	+	?	?	±	±	±	A
Tirosh *et al.* [31]	?	?	+	?	?	±	?	±	D
Calcagni *et al.* [32]	±	?	+	?	?	±	±	±	A

Qualidade de desenvolvimento do estudo: "O" = ótima ou muito boa; "A" = adequada; "D" = duvidosa; "I" = inadequada. **Qualidade dos estudos sobre propriedades de medida:** "+" = suficiente; "–" = insuficiente; "?" = indeterminada; "±" = parcialmente suficiente.

Os estudos incluídos demonstram uma crescente aplicação de tecnologias móveis e sistemas instrumentados para avaliação e reabilitação do controle postural, equilíbrio e funcionalidade em diferentes condições neurológicas. Entretanto, a análise evidencia importante heterogeneidade quanto aos sensores empregados, posicionamento dos dispositivos, dispositivos móveis utilizados, sistemas operacionais, tarefas de avaliação, processamento dos sinais e finalidade tecnológica. Essa diversidade demonstra o potencial das tecnologias móveis para ampliar o acesso à avaliação objetiva do movimento, mas também evidencia desafios relacionados à padronização, reprodutibilidade, interoperabilidade e validação dos sistemas.

Os estudos podem ser compreendidos a partir de uma cadeia tecnológica composta por sensoriamento, aquisição do sinal, processamento, extração de parâmetros biomecânicos e interpretação clínica. Nesse contexto, o sensor não deve ser considerado isoladamente, uma vez que o resultado obtido depende também do dispositivo no qual está incorporado, do posicionamento corporal, das características de aquisição e do algoritmo utilizado para transformar o sinal em uma medida relacionada ao controle postural. Dessa forma, a análise dos estudos evidencia que a tecnologia deve ser compreendida como um conjunto sensor–*hardware*–*software*–algoritmo–protocolo–desfecho clínico.

Tecnologias inerciais na doença de Parkinson

Os estudos de Ozinga *et al.* [18–20] e Fiems *et al.* [21] destacaram-se pela utilização de sensores inerciais para avaliação da estabilidade e oscilação postural em indivíduos com DP. Nos estudos de Ozinga *et al.* [18,19], foram utilizados dispositivos móveis operados pelo sistema iOS, particularmente iPad, associados a sensores inerciais integrados (IMU). O aplicativo *CC-BApp*, desenvolvido pela Cleveland Clinic, foi empregado para captar medidas de equilíbrio estático em diferentes condições sensoriais e em três dimensões. A comparação com sistemas de referência demonstrou resultados positivos à utilização da tecnologia móvel para mensuração da estabilidade postural.

No estudo de Ozinga *et al.* [20], seguiu uma abordagem comparativa entre a acelerometria móvel e o sistema clínico Smart Balance Master da NeuroCom, utilizando o Sensory Organization Test (SOT). Foram analisadas as oscilações nos sentidos ântero-posterior (AP) e médio-lateral (ML), demonstrando que a acelerometria móvel pode constituir uma alternativa portátil para avaliação da instabilidade postural.

Fiems *et al.* [21] utilizaram o Sway Balance™ em um iPod Touch com iOS, empregando IMUs para mensurar a oscilação postural. O protocolo incluiu Romberg, semi-tandem, tandem e apoio unipodal, com variação da superfície e da condição visual. O estudo demonstrou forte confiabilidade teste-reteste e validade concorrente, reforçando o potencial das IMUs para avaliação objetiva e portátil do equilíbrio.

Os dados provenientes das IMUs foram processados por programas desenvolvidos em MATLAB R2009a. As acelerações foram adquiridas a 60 Hz e submetidas a um filtro Butterworth passa-baixa com frequência de corte de 3,5 Hz e processamento em fase zero (*zero-phase*). Posteriormente, foi calculada a variável JERK, considerando as componentes de aceleração AP e ML:

$$JERK = \frac{1}{t}\int_0^t \left[\left(\frac{dA_{CC_{AP}}}{dt}\right)^2 + \left(\frac{dA_{CCML}}{dt}\right)^2\right] dt$$

em que representa a aceleração na direção AP, a aceleração na direção ML e o período de aquisição. O JERK permite representar a taxa de variação da aceleração e, portanto, caracterizar a dinâmica da oscilação postural.

Esse procedimento evidencia que a medida biomecânica não é resultado exclusivamente do sensor. A frequência de aquisição, o filtro utilizado e o algoritmo empregado para obtenção do JERK fazem parte da própria definição do desfecho. Consequentemente, diferenças nesses parâmetros podem influenciar a comparabilidade entre estudos e dispositivos.

Assim, esses estudos representam uma transição de sistemas laboratoriais ou clínicos de maior complexidade para sistemas inerciais portáteis, capazes de registrar o movimento em condições mais acessíveis. Entretanto, as tecnologias foram avaliadas principalmente em dispositivos específicos e em condições controladas, sendo necessária maior investigação sobre a equivalência dos resultados

quando sensores semelhantes são empregados em diferentes modelos de dispositivos.

Sensores e avaliação do equilíbrio na Síndrome de Down

Hsieh *et al.* [22] utilizaram um *iPod Touch* de quarta geração, operado por iOS, associado a um sistema Arduino para sincronização dos dados. O acelerômetro do dispositivo foi utilizado para registrar o equilíbrio durante tarefas em ortostatismo e tandem, com olhos abertos e fechados. Os resultados apresentaram padrões semelhantes aos obtidos por plataforma de força, indicando potencial para utilização do dispositivo móvel na avaliação do equilíbrio.

Do ponto de vista tecnológico, a utilização de Arduino como elemento de sincronização demonstra uma arquitetura híbrida entre dispositivos móveis e sistemas eletrônicos externos. Essa abordagem amplia as possibilidades de aquisição, mas também introduz maior complexidade de hardware e dependência de componentes adicionais.

Assim, embora o *smartphone* ou *iPod* possa reduzir custos e aumentar a portabilidade, a necessidade de dispositivos auxiliares para sincronização ou aquisição pode diminuir parte da vantagem de um sistema totalmente integrado. Essa questão deve ser considerada no desenvolvimento de tecnologias destinadas ao uso clínico ou domiciliar, particularmente quando o objetivo é obter uma solução escalável e de fácil utilização.

Tecnologias móveis e sistemas de captura nas ataxias

Nos estudos envolvendo ataxias, foram identificadas duas estratégias tecnológicas distintas. Arcuria *et al.* [23] utilizaram *smartphones* com sistemas iOS/Android, equipados com sensores triaxiais posicionados no esterno, associados ao sistema *TechnoBody Pro-Kin-212-N*. O aplicativo *APP-Coo-Balance-Test* foi utilizado para quantificar oscilações posturais do tronco e déficits de equilíbrio estático e dinâmico. As tarefas incluíram pés juntos, base alargada, sedestação sem apoio dos pés e caminhada em linha reta.

O posicionamento no esterno é particularmente relevante, pois permite que o sensor acompanhe diretamente os movimentos do tronco. Entretanto, a escolha do local de posicionamento interfere na informação biomecânica obtida, uma vez que sensores localizados no esterno, região lombar, pelve ou membros inferiores estão sujeitos a diferentes componentes do movimento corporal.

Além do posicionamento, Arcuria *et al.* [23] apresentam uma estratégia de processamento diferente daquela utilizada por Fiems *et al.* [21] os testes foram iniciados e finalizados por sinais sonoros, sendo o teste estático realizado durante 10 segundos. Para o teste dinâmico, o participante foi orientado a caminhar em linha reta, em um percurso de 9 metros durante 10 segundos. A oscilação postural foi quantificada por meio do *APPBalScore*, calculado por:

$$APPBalScore = \left(\frac{p_x}{SCA_n} \times 10\right) + k$$

em que corresponde ao número de pixels percorridos pelo ponto vermelho associado ao sensor de aceleração na grade virtual durante o teste; corresponde ao número de *SC* *An* realizados durante o período de avaliação; e *K* representa a pontuação de entrada relacionada à dificuldade do teste.

Esse cálculo demonstra uma estratégia de transformação do sinal distinta da observada em Fiems *et al*. [21] enquanto o JERK representa a variação temporal da aceleração, o APPBalScore converte o deslocamento do sinal em uma pontuação relacionada ao desempenho postural. Portanto, mesmo quando tecnologias utilizam sensores inerciais ou acelerômetros, os resultados não são necessariamente diretamente comparáveis, pois os algoritmos de processamento determinam quais características do movimento serão representadas.

Summa *et al.* [28], por outro lado, desenvolveram o *SaraHome*, utilizando *Kinect v2* e *Leap Motion Controller*, associados a software em *C#* e banco de dados *CouchDB*. O sistema foi desenvolvido para monitoramento remoto de tarefas motoras de indivíduos com ataxia.

A comparação entre Arcuria *et al.* [23] e Summa *et al.* [28] evidencia duas arquiteturas distintas: uma baseada em sensores inerciais móveis e outra baseada em captura de movimento sem marcadores. Enquanto a primeira favorece portabilidade e simplicidade, a segunda possibilita uma caracterização mais ampla do movimento, porém depende de equipamentos específicos e de condições apropriadas para aquisição.

Sensores inerciais e dupla tarefa na doença de Alzheimer

Serra-Añó *et al.* [25] utilizaram o aplicativo *FallSkip* em um *smartphone* Android Xiaomi Redmi 4X, equipado com IMU de seis eixos. O dispositivo foi utilizado para avaliação do controle postural e funcionalidade da marcha em idosos com doença de Alzheimer, incluindo tarefas motoras associadas a tarefas cognitivas.

A utilização de uma IMU de seis eixos representa uma arquitetura adequada para avaliação do movimento corporal, pois combina informações provenientes de acelerômetros e giroscópios. Entretanto, o estudo demonstra que a quantidade de eixos do sensor não determina isoladamente a qualidade da medida. O posicionamento, a frequência de aquisição, as características do movimento e os algoritmos utilizados para processamento também interferem no resultado.

A incorporação da dupla tarefa representa uma evolução em relação aos protocolos exclusivamente estáticos, pois permite avaliar o comportamento motor diante de uma demanda cognitiva simultânea. Essa abordagem aproxima a avaliação tecnológica de situações funcionais observadas no cotidiano e pode aumentar a capacidade dos sistemas de identificar alterações que não são necessariamente evidentes durante tarefas motoras simples.

Tecnologias utilizadas no AVC

Nos estudos envolvendo AVC, observa-se maior diversidade de aplicações, incluindo tanto avaliação instrumental quanto intervenção terapêutica.

Shin [27] utilizou um iPhone 5 com iOS integrado ao sistema *Bal-90*, com acelerômetro e giroscópio para fornecer *feedback* visual em tempo real durante treinamento do controle de tronco. A avaliação foi realizada por meio da *Trunk Impairment Scale* (TIS) e de parâmetros espaço-temporais da marcha, incluindo velocidade, comprimento

do passo e cadência. Os resultados indicaram melhora do controle postural e da capacidade de marcha.

Nesse caso, o sensor deixa de exercer exclusivamente a função de instrumento de mensuração e passa a integrar um sistema de *biofeedback*. Essa mudança é importante, pois o sistema precisa realizar aquisição, processamento e apresentação da informação em tempo suficientemente rápido para fornecer *feedback* durante a execução da tarefa.

Hou *et al.* [26] desenvolveram um aplicativo para smartphone Android utilizando um ASUS Zenfone 3, com acelerômetro e giroscópio, para avaliação quantitativa do equilíbrio estático após AVC. A tecnologia foi comparada com a Escala de Equilíbrio de Berg e apresentou evidências de confiabilidade e validade.

Além dos resultados de validade, Hou *et al*. [26] apresentam uma contribuição importante para compreender como os dados dos sensores foram transformados em variáveis biomecânicas. Os autores calcularam as mudanças combinadas do vetor de aceleração nos eixos AP e ML:

$$A = \frac{1}{n-1}\sum_{i=2}^{n}\sqrt{(x_i - x_{i-\perp})^2 + (yi - yi_{-1})^2}$$

em que xi e yi correspondem às acelerações nos respectivos eixos, em m/s², e n corresponde ao número total de amostras. Nesse caso, maior alteração do vetor de aceleração representou maior instabilidade postural.

Os autores também utilizaram o giroscópio para avaliar a alteração da velocidade angular nos eixos de *pitch*, *roll* e *yaw*, permitindo estimar a inclinação corporal durante o teste. As velocidades angulares foram convertidas em deslocamento angular por integração numérica:

$$a_i = \frac{|\alpha_i|}{f} \times \frac{180}{\pi} \qquad b_i = \frac{\beta_i}{f} \times \frac{180}{\pi} \qquad c_i = \frac{y_i}{f} \times \frac{180}{\pi}$$

sendo o ângulo acumulado estimado por:

$$\theta = \sum_{i=1}^{n}(ai + b_i + c_i)$$

em que a_i , β_i e y_i correspondem às velocidades angulares instantâneas, f à frequência de amostragem e ao número de amostras. Nesse modelo, maior ângulo de inclinação corporal indicou menor controle postural.

A abordagem de Hou *et al.* [26] evidencia uma arquitetura multissensorial, na qual o acelerômetro fornece informações relacionadas à alteração do movimento linear, enquanto o giroscópio permite caracterizar o componente angular do movimento. Essa combinação é particularmente relevante para sistemas destinados à avaliação do alinhamento e controle do tronco.

Aphiphaksakul *et al.* [29] utilizaram o *Compass Inclinometer*, em um smartphone Android associado a um disco de equilíbrio, para fornecer *feedback* visual durante exercícios domiciliares. A tecnologia demonstrou potencial para melhorar o controle de tronco e o desempenho funcional.

Assim, nos estudos de AVC observa-se uma progressão tecnológica, partindo de sensores utilizados para mensuração do equilíbrio e chegando a sistemas de feedback visual, jogos e intervenção terapêutica. Essa evolução representa uma mudança do conceito de tecnologia como instrumento de avaliação para tecnologia como componente ativo da reabilitação.

Sensores utilizados no TCE e concussão

Nos estudos envolvendo TCE leve e concussão, as tecnologias foram utilizadas principalmente para avaliação de déficits neuromotores.

Linder *et al*. [24] desenvolveram um paradigma portátil de *dual-task* utilizando iPad e Arduino, com avaliação simultânea da estabilidade postural e função cognitiva. Foram avaliadas condições de apoio duplo e tandem, com olhos abertos e fechados, associadas a estímulos auditivos. A condição tandem com olhos fechados e estímulos apresentados a 60 estímulos/minuto foi identificada como uma condição adequada para avaliação de dupla tarefa.

Rhea *et al.* [30] desenvolveram o *AccWalker*, utilizando smartphones Android, como Moto X e Google Pixel, fixados na coxa. A tecnologia avaliou a tarefa de *stepping-in-place* com olhos abertos, olhos fechados e rotação da cabeça. O aumento da variabilidade do movimento foi utilizado como indicador de disfunção neuromotora.

Tirosh *et al.* [31] utilizaram o *TelePhysio* para avaliação remota do equilíbrio em pacientes com concussão aguda. O sistema analisou a magnitude da aceleração AP e ML durante condições de superfície firme e espuma, com olhos abertos e fechados. O estudo demonstrou a viabilidade da avaliação remota utilizando smartphones.

Esses estudos demonstram uma tendência importante: a miniaturização dos sistemas de aquisição permite transferir avaliações originalmente realizadas em laboratório para dispositivos portáteis e ambientes remotos. Entretanto, a aplicação remota exige que o sistema seja suficientemente robusto para lidar com variações de posicionamento, execução da tarefa e condições ambientais.

Processamento dos sinais e biomarcadores digitais

Calcagni *et al*. [32] representam uma etapa mais avançada dessa evolução ao utilizar o aplicativo *NeuFun-TS* e sensores incorporados a smartphones para obtenção de biomarcadores digitais. O sistema foi utilizado nos testes *Postural Sway* e *Pronator Drift* para avaliação da oscilação corporal e controle motor fino em indivíduos com diferentes doenças neurológicas.

Nesse estudo, os dados tridimensionais de cada teste foram inicialmente calibrados e transformados em duas séries temporais: ML e AP. Posteriormente, essas componentes foram combinadas em uma série de aceleração radial denominada *Net acceleration*.

A partir dessas séries foram extraídas medidas nos domínios temporal e de frequência. No domínio temporal, foi calculado o *Root Mean Square* (RMS):

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_i^2}$$

em que a_i representa a aceleração instantânea e N o número de pontos da série temporal. O RMS foi calculado para as componentes ML, AP e Net e representa a magnitude da aceleração.

Também foi calculado o Jerk, representando a taxa de alteração da aceleração em relação ao tempo:

$$Jerk = \int_0^t \left(\frac{dACC}{dt}\right)^2 dt$$

Além disso, foi calculado o Net Jerk, combinando as componentes ML e AP:

$$Net\ Jerk = \int_0^t \left[\left(\frac{dAcc_{ML}}{dt}\right)^2 + \left(\frac{dAcc_{AP}}{dt}\right)^2\right] dt$$

A análise foi posteriormente ampliada para o domínio da frequência. Os dados de aceleração foram submetidos à Transformada Discreta de Fourier, permitindo a obtenção de espectros de frequência e densidade espectral de potência (PSD). A partir desses espectros foram calculados o Spectral Centroid (SC) e o Spectral Spread (SS).

O centróide espectral foi calculado por:

$$SC = \mu_1 = \frac{\sum_{k=b1}^{b_2} f_k \boldsymbol{s}\, k}{\sum_{k=b_1}^{b2} s\, k}$$

representando a frequência central ponderada pela densidade espectral.

O *Spectral Spread* foi calculado por:

$$SS = \mu_2 = \frac{\sum_{k=b1}^{b_2} (f_{k-}\mu_1)^2 \boldsymbol{s}\, k}{\sum_{k=b_1}^{b2} s\, k}$$

representando a dispersão das frequências em torno do centróide espectral.

Essas medidas foram calculadas nas diferentes condições posturais, incluindo olhos abertos e fechados, e utilizadas para calcular a Razão de Romberg:

$$Romberg\ Ratio = \frac{Medida_{EC-FT}}{Medida_{EO-FT}}$$

Essa abordagem representa uma evolução importante em relação aos estudos que utilizam apenas uma variável de oscilação, pois combina características temporais e espectrais do movimento. Dessa maneira, o *smartphone* deixa de ser apenas um instrumento de aquisição e passa a constituir uma plataforma para geração de biomarcadores digitais quantitativos.

Comparação das estratégias de processamento

A análise conjunta dos estudos evidencia diferentes níveis de complexidade na transformação do sinal bruto em uma variável biomecânica. Em Fiems *et al.* [21], o processamento segue a sequência:

aceleração → filtragem → derivada temporal → JERK.

Em Arcuria *et al.* [23]:

aceleração → representação espacial → deslocamento em pixels → APPBalScore.

Em Hou *et al.* [26]:

aceleração → diferenças entre amostras → alteração do vetor de aceleração

e, paralelamente:

velocidade angular → integração numérica → ângulo de inclinação.

Em Calcagni *et al.* [32]:

aceleração tridimensional → calibração → componentes ML/AP → Net acceleration → RMS/Jerk → Transformada de Fourier → SC/SS.

Essa comparação evidencia que sensores semelhantes podem gerar biomarcadores completamente diferentes. Portanto, a informação clínica produzida por uma tecnologia não depende somente da presença de acelerômetro ou giroscópio, mas da combinação entre hardware, aquisição e processamento.

Dependência do dispositivo e do sistema operacional

Os estudos utilizaram iPad, iPod Touch e iPhone em ambiente iOS, enquanto outros empregaram smartphones Android de diferentes fabricantes, como ASUS, Xiaomi, Motorola e Google. Também foram observados sistemas baseados em Arduino, Kinect, Leap Motion e plataformas específicas.

Essa heterogeneidade significa que o smartphone não deve ser entendido apenas como uma interface para execução do aplicativo. Ele constitui parte do sistema de instrumentação, uma vez que seus sensores são responsáveis pela aquisição dos dados.

Assim, a arquitetura tecnológica pode ser representada como:

Movimento corporal → sensor → aquisição → processamento → algoritmo → aplicativo → dispositivo → sistema operacional → resultado biomecânico.

Uma alteração em qualquer uma dessas etapas pode modificar o resultado final. Consequentemente, a validação de um aplicativo em determinado modelo de smartphone não garante automaticamente equivalência de desempenho em outro dispositivo, mesmo que ambos utilizem o mesmo sistema operacional.

Essa questão é particularmente relevante para tecnologias destinadas ao uso domiciliar e remoto. A literatura apresenta poucas evidências de validação cruzada entre diferentes fabricantes, modelos de smartphones e sistemas operacionais, constituindo uma importante lacuna para a escalabilidade das soluções.

IV. CONCLUSÃO

As tecnologias móveis e os sistemas de sensores apresentam potencial para avaliação, monitoramento e reabilitação do controle e alinhamento postural, e equilíbrio em indivíduos com condições neurológicas. Entretanto, as evidências disponíveis ainda são limitadas e heterogêneas quanto às populações, dispositivos, sensores, sistemas operacionais, tarefas e desfechos, dificultando as conclusões clinicamente aplicáveis. Assim, embora alguns estudos indiquem resultados positivos quanto à validade, confiabilidade e viabilidade dessas tecnologias, o estudo contribui com um mapa de evidências, ressaltando que ainda são necessários estudos com maior rigor metodológico, protocolos padronizados e amostras adequadas para estabelecer a efetividade clínica.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)- Código de Financiamento 001.

REFERÊNCIAS

[1] GBD 2021 Nervous System Disorders Collaborators, "Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021," *Lancet Neurol.*, vol. 23, no. 4, pp. 344–381, Apr. 2024, doi: 10.1016/S1474-4422(24)00038-3.

[2] J. Ye, C. Duan, J. Han, J. Chen, N. Sun, Y. Li, T. Yuan, and D. Peng, "Peripheral mitochondrial DNA as a neuroinflammatory biomarker for major depressive disorder," *Neural Regen. Res.*, vol. 20, no. 6, pp. 1541–1554, Jun. 2025, doi: 10.4103/NRR.NRR-D-23-01878.

[3] D. O. Q. Machado et al., "O impacto das patologias clínicas neurológicas para a saúde pública: The impact of clinical neurological pathologies for public health," *Braz. J. Health Rev.*, vol. 5, no. 4, pp. 13774–13787, 2022.

[4] M. S. L. Quintana, A. C. Alonso, N. M. S. Luna, J. P. da Silva, M. H. dos Santos Lino, G. C. Brech, and J. M. D. A. Greve, "Correlation of the sagittal balance with postural analysis of the pelvis and lumbar spine," *Acta Ortop. Bras.*, vol. 32, no. 1, art. no. e274089, 2024, doi: 10.1590/1413-785220243201e274089.

[5] M. M. Reguera-García, R. Leirós-Rodríguez, L. Álvarez-Barrio, and B. Alonso-Cortés Fradejas, "Analysis of postural control in sitting by pressure mapping in patients with multiple sclerosis, spinal cord injury and Friedreich's ataxia: A case series study," *Sensors*, vol. 20, no. 22, art. no. 6488, 2020, doi: 10.3390/s20226488.

[6] T.-H. Hsieh et al., "The applications of smart mobile device for detecting balance dysfunction in individuals with Down syndrome," *Biomed. Eng.: Appl., Basis Commun.*, vol. 30, no. 1, art. no. 1850007, 2018.

[7] M. G. Silva, B. M. Pilling, and C. T. Candotti, "Instrumentos de autoavaliação da postura corporal: Uma revisão de escopo," *Fisioter. Pesqui.*, vol. 30, art. no. e22017823pt, 2023, doi: 10.1590/1809-2950/e22017823pt.

[8] F. M. K. dos Santos, F. V. Mendes, S. S. Woellner, and N. G. Borges Júnior, "Use of visual feedback for balance training in hemiparetic stroke patients," *Fisioter. Mov.*, vol. 28, no. 2, pp. 241–249, 2015, doi: 10.1590/0103-5150.028.002.AO04.

[9] S. J. Ozinga and J. L. Alberts, "Quantification of postural stability in older adults using mobile technology," *Exp. Brain Res.*, vol. 232, no. 12, pp. 3861–3872, 2014, doi: 10.1007/s00221-014-4069-8.

[10] J. B. Caccese, E. Teel, R. Van Patten, M. A. Muzeau, G. L. Iverson, and H. A. VanRavenhorst-Bell, "Test-retest reliability and preliminary reliable change estimates for Sway Balance tests administered remotely in community-dwelling adults," *Front. Digit. Health*, vol. 4, art. no. 999250, 2022, doi: 10.3389/fdgth.2022.999250.

[11] R. Purohit, J. P. Appelgren-Gonzalez, G. Varas-Diaz, S. Wang, M. Hosiasson, F. Covarrubias-Escudero, and T. Bhatt, "Feasibility of smartphone-based exercise training integrated with functional electrical stimulation after stroke (SETS): A preliminary study," *Sensors*, vol. 25, no. 4, art. no. 1254, 2025, doi: 10.3390/s25041254.

[12] J. L. Alberts, A. Thota, J. Hirsch, S. Ozinga, T. Dey, D. D. Schindler, M. M. Koop, D. Burke, and S. M. Linder, "Quantification of the Balance Error Scoring System with mobile technology," *Med. Sci. Sports Exerc.*, vol. 47, no. 10, pp. 2233–2240, 2015, doi: 10.1249/MSS.0000000000000656.

[13] J. T. Cavanaugh, K. M. Guskiewicz, C. Giuliani, S. Marshall, V. S. Mercer, and J. K. Stergiou, "Recovery of postural control after cerebral concussion: New insights using approximate entropy," *J. Athl. Train.*, vol. 41, no. 3, pp. 305–313, 2006.

[14] J. L. Alberts, J. R. Hirsch, M. M. Koop, D. D. Schindler, D. E. Kana, S. M. Linder, S. Campbell, and A. K. Thota, "Using accelerometer and gyroscopic measures to quantify postural stability," *J. Athl. Train.*, vol. 50, no. 6, pp. 578–588, 2015, doi: 10.4085/1062-6050-50.2.01.

[15] J. A. C. Sterne et al., "ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions," *BMJ*, vol. 355, art. no. i4919, 2016, doi: 10.1136/bmj.i4919.

[16] J. P. T. Higgins et al., "Assessing risk of bias in a randomized trial," in *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed., J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, Eds. Chichester, U.K.: Wiley, 2019, pp. 205–228.

[17] L. B. Mokkink, E. B. M. Elsman, and C. B. Terwee, "COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0," *Qual. Life Res.*, vol. 33, no. 11, pp. 2929–2939, 2024, doi: 10.1007/s11136-024-03761-6.

[18] S. J. Ozinga, A. G. Machado, M. M. Koop, A. B. Rosenfeldt, and J. L. Alberts, "Objective assessment of postural stability in Parkinson's disease using mobile technology," *Mov. Disord.*, vol. 30, no. 9, pp. 1214–1221, 2015, doi: 10.1002/mds.26214.

[19] S. J. Ozinga, S. M. Linder, and J. L. Alberts, "Use of mobile device accelerometry to enhance evaluation of postural instability in Parkinson disease," *Arch. Phys. Med. Rehabil.*, vol. 98, no. 4, pp. 649–658, 2017, doi: 10.1016/j.apmr.2016.08.479.

[20] S. J. Ozinga, M. M. Koop, S. M. Linder, A. G. Machado, T. Dey, and J. L. Alberts, "Three-dimensional evaluation of postural stability in Parkinson's disease with mobile technology," *NeuroRehabilitation*, vol. 41, no. 1, pp. 211–218, 2017, doi: 10.3233/NRE-171473.

[21] C. L. Fiems, E. L. Dugan, E. S. Moore, and S. A. Combs-Miller, "Reliability and validity of the Sway Balance mobile application for measurement of postural sway in people with Parkinson disease," *NeuroRehabilitation*, vol. 43, no. 2, pp. 147–154, 2018, doi: 10.3233/NRE-182424.

[22] T.-H. Hsieh et al., "The applications of smart mobile device for detecting balance dysfunction in individuals with Down syndrome," *Biomed. Eng.: Appl., Basis Commun.*, vol. 30, no. 1, art. no. 1850007, 2018.

[23] G. Arcuria, C. Marcotulli, R. Amuso, G. Dattilo, C. Galasso, F. Pierelli, and C. Casali, "Developing a smartphone application, triaxial accelerometer-based, to quantify static and dynamic balance deficits in patients with cerebellar ataxias," *J. Neurol.*, vol. 267, no. 3, pp. 625–639, 2020, doi: 10.1007/s00415-019-09570-z.

[24] S. M. Linder, M. M. Koop, S. J. Ozinga, Z. Goldfarb, and J. L. Alberts, "A mobile device dual-task paradigm for the assessment of mTBI," *Mil. Med.*, vol. 184, suppl. 1, pp. 174–180, 2019, doi: 10.1093/milmed/usy334.

[25] P. Serra-Añó, J. F. Pedrero-Sánchez, J. Hurtado-Abellán, M. Inglés, G. V. Espí-López, and J. López-Pascual, "Mobility assessment in people with Alzheimer disease using smartphone sensors," *J. Neuroeng. Rehabil.*, vol. 16, art. no. 103, 2019, doi: 10.1186/s12984-019-0576-y.

[26] D.-C. Shin, "Smartphone-based visual feedback trunk control training for gait ability in stroke patients: A single-blind randomized controlled trial," *Technol. Health Care*, vol. 28, no. 1, pp. 45–55, 2020, doi: 10.3233/THC-191647.

[27] Y.-R. Hou, Y.-L. Chiu, S.-L. Chiang, H.-Y. Chen, and W.-H. Sung, "Development of a smartphone-based balance assessment system for subjects with stroke," *Sensors*, vol. 20, no. 1, art. no. 88, 2020, doi: 10.3390/s20010088.

[28] S. Summa, T. Schirinzi, G. M. Bernava, A. Romano, M. Favetta, E. M. Valente, E. Bertini, E. Castelli, M. Petrarca, G. Pioggia, and G. Vasco, "Development of SaraHome: A novel, well-accepted, technology-based assessment tool for patients with ataxia," *Comput. Methods Programs Biomed.*, vol. 188, art. no. 105257, 2020, doi: 10.1016/j.cmpb.2019.105257.

[29] P. Aphiphaksakul and A. Siriphorn, "Home-based exercise using balance disc and smartphone inclinometer application improves balance and activity of daily living in individuals with stroke: A randomized controlled trial," *PLoS One*, vol. 17, no. 11, art. no. e0277870, 2022, doi: 10.1371/journal.pone.0277870.

[30] C. K. Rhea, M. Yamada, N. A. Kuznetsov, J. T. Jakiela, C. T. LoJacono, S. E. Ross, F. J. Haran, J. M. Bailie, and W. G. Wright, "Neuromotor changes in participants with a concussion history can be detected with a custom smartphone app," *PLoS One*, vol. 17, no. 12, art. no. e0278994, 2022, doi: 10.1371/journal.pone.0278994.

[31] O. Tirosh, J. Klonis, M. Hamilton, J. Olver, N. Wickramasinghe, D. McKenzie, D. El-Ansary, and G. Williams, "Smartphone technology to facilitate remote postural balance assessment in acute concussion management: Pilot study," *Sensors*, vol. 24, no. 21, art. no. 6870, 2024, doi: 10.3390/s24216870.

[32] M. Calcagni, P. Kosa, and B. Bielekova, "Smartphone postural sway and pronator drift tests as measures of neurological disability," *BMC Neurol.*, vol. 25, art. no. 50, 2025, doi: 10.1186/s12883-025-04038-2.

Telerehabilitation Application for Prescribing Home Exercises and Supporting the Daily Lives of Wheelchair Users

Tifany de Oliveira Mariano
Science and Technology Institute
Federal University of São Paulo
São José dos Campos, Brazil
0009-0001-5166-4478

Regina Célia Coelho
Science and Technology Institute
Federal University of São Paulo
São José dos Campos, Brazil
0000-0002-4428-9745

Lara Bonfanti Ajauskas
Riccio College of Engineering
University of Massachusetts Amherst,
Amherst, Massachusetts, US
0009-0006-7716-7521

Eliana Chaves Ferretti
Institute of Health and Society
Federal University of São Paulo, Santos, Brazil
0000-0002-7605-828X

Carlos Marcelo Gurjão de Godoy
Science and Technology Institute
Federal University of São Paulo
São José dos Campos, Brazil
0000-0001-8846-3242

Abstract—The use of digital technologies in healthcare has promoted new possibilities for monitoring and rehabilitating patients with functional limitations. In this context, this study proposes the development of a mobile application aimed at the physiotherapy monitoring of wheelchair-bound patients, with the objective of facilitating clinical monitoring, improving communication between professional and patient, and encouraging adherence to treatment. The methodology involves bibliographic research and prototype development. The app was initially modeled using the Figma tool, used to define the information architecture, navigation flows, and interface prototyping. Subsequently, the screens and functionalities were implemented using the Flutter framework, allowing the development of a multiplatform application focused on usability and accessibility. The application includes functionalities such as exercise recording, monitoring of functional progress, sending therapeutic guidelines, activity reminders, and collection of user feedback. The app is prepared to obtain user experience by using the System Usability Scale (SUS), which allows the identification of aspects related to efficiency, effectiveness, and satisfaction in the use of the tool. It is expected that the solution will contribute to expanding access to physiotherapy care, promoting continuity of care and fostering greater autonomy and quality of life for wheelchair-bound patients.

Keywords—digital physiotherapy; health technology; wheelchair; usability; mobile application.

I. INTRODUCTION

Technology has become an essential resource in promoting health, offering opportunities to expand access to care and treatment through digital health [1]. In Brazil, the wheelchair-using population faces daily challenges such as limited accessibility, chronic pain, contractures, pressure ulcers, and sedentary lifestyles, factors that directly compromise their quality of life [2].

Despite the advancement of digital health solutions favoring the prevention of complications, adherence to treatment and the promotion of autonomy [3], there is a scarcity of tele-rehabilitation applications specifically aimed at this purpose, particularly for wheelchair users. Physical disability represents an important global public health issue, especially due to the functional, social and psychological impact associated with motor limitations. Among individuals with physical disabilities, wheelchair users face challenges related to mobility, urban accessibility, social participation and continuous access to health services [4].

The use of a wheelchair, while promoting functional independence, is also associated with significant biomechanical changes, including upper limb overload, musculoskeletal pain, postural changes, pressure injuries, and reduced cardiovascular conditioning [5]. Therefore, continuous physiotherapy follow-up becomes essential for preventing secondary complications and maintaining functionality.

In recent years, there has been significant growth in the use of digital technologies applied to health. Telerehabilitation has emerged as a strategy capable of expanding access to physiotherapy treatment, especially for individuals with mobility difficulties and accessibility limitations [6].

The COVID-19 pandemic significantly accelerated the implementation of remote physiotherapy services, favoring the development of digital platforms, mobile applications, and remote monitoring systems. Studies demonstrate that tele-rehabilitation can present similar results to in-person care in several functional aspects, in addition to promoting greater therapeutic adherence and reducing healthcare costs [7].

In this context, mobile health applications have stood out as promising tools for prescribing home exercises, functional

This work was developed using resources obtained from the Financiadora de Estudos e Projetos (FINEP) - Ref. Finep nº 2174/22

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

monitoring, communication between professionals and patients, and promoting self-care [8].

According to the World Health Organization (WHO), an estimated 1.3 billion people worldwide live with some type of disability classified as significant. This number corresponds to a proportion of one in six people [16]. For wheelchair users, applications focused on Tele-rehabilitation can represent an accessible alternative for continuous physiotherapy monitoring, ergonomic support, prevention of secondary injuries, and encouragement of adapted physical exercise.

Thus, this project proposes the development of a tele-rehabilitation application aimed at prescribing home exercises and supporting the daily activities of wheelchair users, integrating physiotherapy resources, digital accessibility, and remote monitoring.

II. Theoretical Basis

According to Tweedy Beckman [9], individual's dependent on wheelchairs frequently present musculoskeletal, cardiovascular, and postural alterations resulting from sedentary lifestyles and long periods of sitting, making continuous physiotherapy follow-up indispensable. Among the main complications associated with prolonged wheelchair use are musculoskeletal pain, deformities, pressure injuries, muscle shortening, loss of physical conditioning, and reduced functional capacity. According to Gomes Neto [3], physiotherapy interventions aimed at muscle strengthening, therapeutic stretching, and functional training promote significant improvement in mobility and reduction of secondary complications in wheelchair patients.

In addition to conventional in-person treatments, advances in digital technologies have brought new possibilities for physiotherapy follow-up. In that context, telerehabilitation has been consolidating itself as an important strategy to expand access to care, especially for these individuals. In the article written by Kairy, Veras and Archambault [16], telerehabilitation programs demonstrate effectiveness in therapeutic continuity, adherence to treatment and functional monitoring.

In this way, mobile health applications have become relevant tools in physical rehabilitation programs. As described in the work [13], mobile health technologies have the potential to expand access to services and promote greater integration between professionals and patients. These platforms allow the sending of guidelines, explanatory videos, activity reminders and monitoring of clinical progress in real time.

Another important point is related to the personalization of physiotherapy treatment. According to Chen [3], monitoring integrated with digital platforms allows for faster therapeutic adjustments and better communication between patient and physiotherapist. However, despite the observed benefits, there are still challenges related to the use of these technologies in the health field. Issues such as digital accessibility, technological limitations, difficulties with internet connectivity, and the need for scientific validation of applications should be considered during the development of these tools.

In turn, the work of Turolla and colleagues [6] mentions that the development of health technologies should prioritize safety, usability and digital inclusion to ensure therapeutic efficacy and universal access.

Observed as a whole, those works highlights the fact that the integration between physiotherapy, telerehabilitation and mobile health applications represents an important contemporary trend in the physical rehabilitation of wheelchair users. More than that, the use of these technologies may promote therapeutic continuity, access to care and functional independence of wheelchair users, contributing significantly to improve the quality of life of these individuals.

III. Methodology

The methodology focuses on developing a telerehabilitation application for prescribing home exercises and supporting wheelchair users. To achieve this, we followed sequential steps, from identifying user needs to developing and testing a functional prototype. An exploratory study identified the main demands and barriers faced by wheelchair users. Next, the application's structure was defined, including exercise prescription, educational content, remote monitoring, and accessibility features. The interfaces and screen architecture were designed in Figma, emphasizing accessibility and intuitive navigation. The system was then implemented in Flutter using Dart, ensuring cross-platform compatibility and adaptations that help promote the autonomy of wheelchair users.

As for the App Architecture, the development process began with prototyping the screens on the Figma platform, allowing for the definition of the navigation flow, organization of visual elements, and preliminary validation of the user experience. Subsequently, the prototypes were converted into functional interfaces using Flutter, preserving the usability and accessibility principles defined in the design phase.

The developed screens include functionalities such as user authentication, patient registration and monitoring, access to therapeutic exercises, monitoring of clinical progress, and visualization of data from the physiotherapy monitoring system. The use of Flutter allowed the creation of reusable components, contributing to the visual standardization and maintenance of the application.

The application's architecture was structured in layers, comprising a user interface developed in Flutter, a logic layer responsible for processing information, and a data storage layer for managing clinical records and follow-up histories. This organization favors system scalability, code maintenance, and standardization related to the screens already assembled in Figma.

IV. Results

The following images demonstrate the organization of the screens previously created in Figma and programmed in Flutter. They show a user navigation flow, including registration, authentication, physiotherapy assessment, and video (exercise) recording, as well as screens specifying the bottom buttons that the user can click to access the saved exercises screen, a calendar to mark the physiotherapy day, check for any

messages from the physiotherapist, and notifications. Fig. 1 corresponds to the application's authentication flow screns. These screens allow the user to register and securely access the system by logging in with the email and password that the professional uses to register, and also to register the patient, through which they can later access the app.

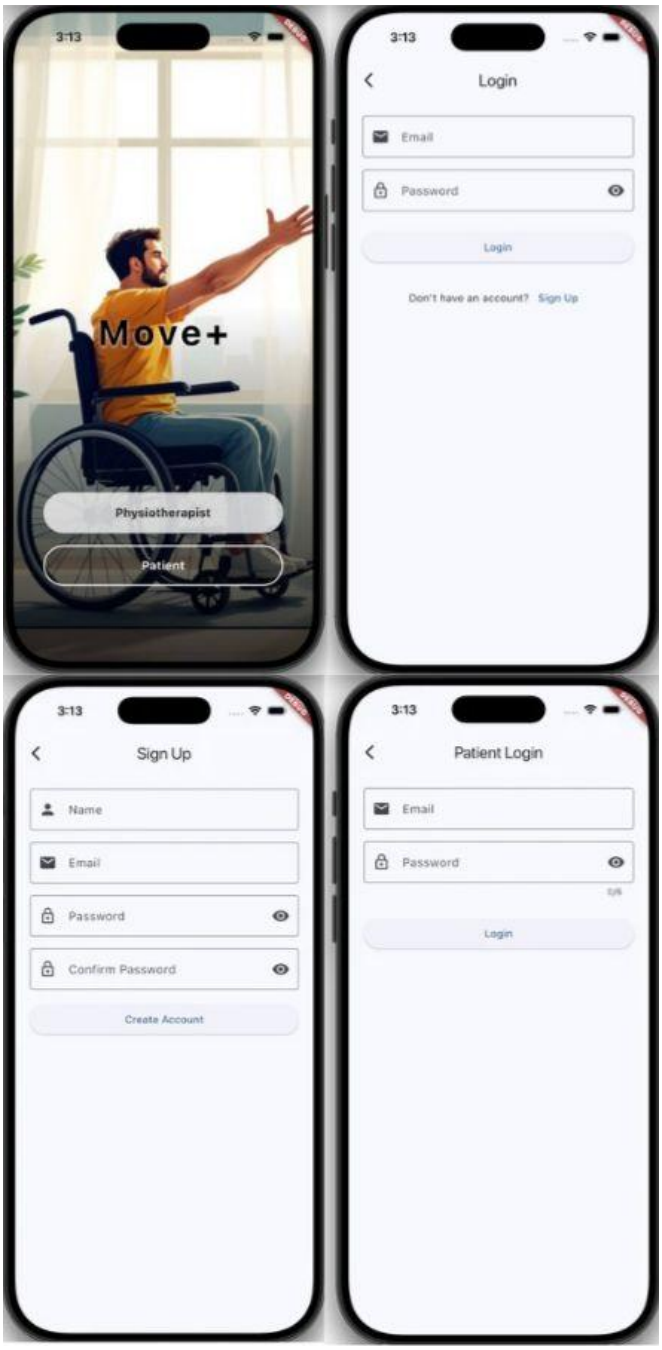


Fig. 1 Illustration of the professional login area. Up left: home screen; up right: login screen; down left: registration screen; down right: patient registration.

Fig 2 shows the functionalities available to the professional. These interfaces include patient registration where specific characteristics can be entered, followed by a screen to save the patient's information. After this process, the patient can access the application on their device through a basic registration previously completed by the professional and access their clinical information.

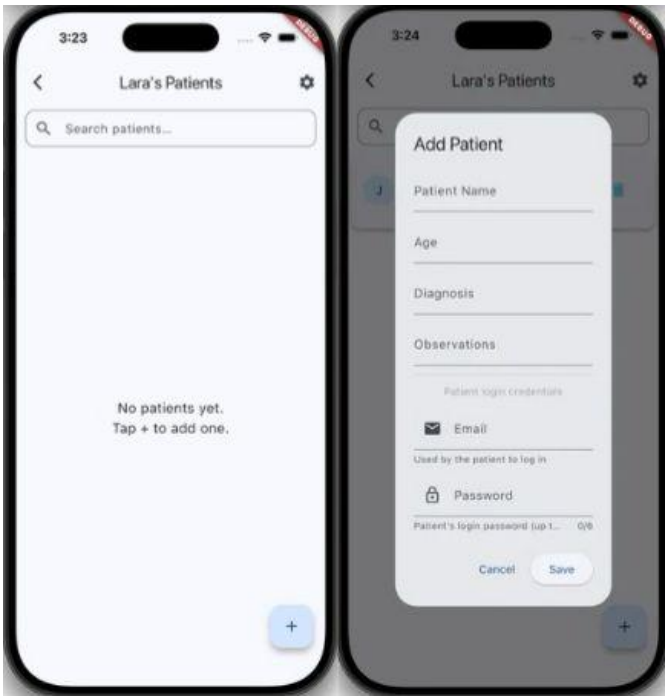


Fig. 2. Functionalities. The left side shows the patient registration area, and the right side corresponds to the screen after clicking the "+" icon to open the patient information and save it.

Fig 3 demonstrates the application from the professional's point of view after logging in, where they can view and leave messages for the patient, adjust the calendar with the treatment days, write important notices, and record videos of the proposed exercises. This monitoring of therapeutic progress and management of rehabilitation activities ensures that each patient is placed within a "folder" with their name and login, allowing the therapist to better organize themselves and prescribe exercises in an individualized way.

Fig 4 illustrates the patient interface. Through them, the user can access the calendar with the days the professional has scheduled treatment, the prescribed exercises, receive notifications, view messages left by the professional, and track their therapeutic progress. Navigation within the application is done using the icons shown in the figures at the bottom of the screens.

The screens were all designed to assist in monitoring physiotherapy and its execution by the patient. The application implementation allowed for cross-platform development with reusable components and standardized visual elements. The integration between the therapist and patient interfaces resulted in an organized and clear application. The interface design and system functionality contributed to a consistent and intuitive navigation experience, preserving the accessibility principles established during the prototyping phase.

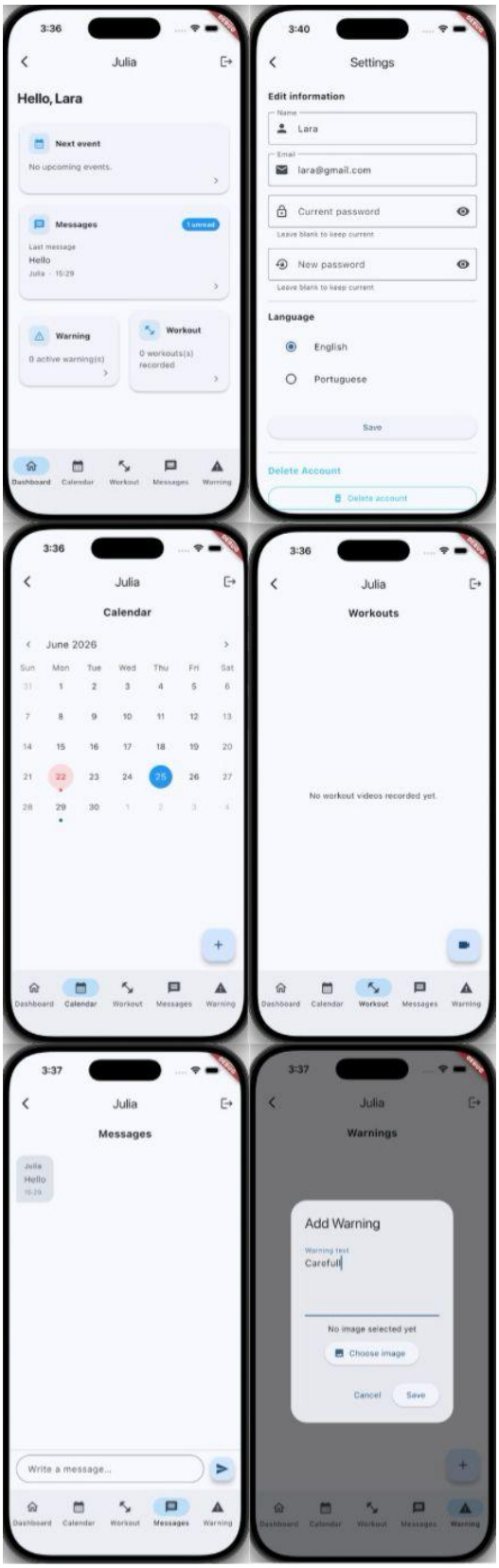


Fig. 3. Application screens. Up left: patient registration; Up right; middle left: main screen; middle left: calendar; middle right: screen for prescribing exercises, which can be done through recorded videos; down left: messages; down right: screen for alerts or important notices.

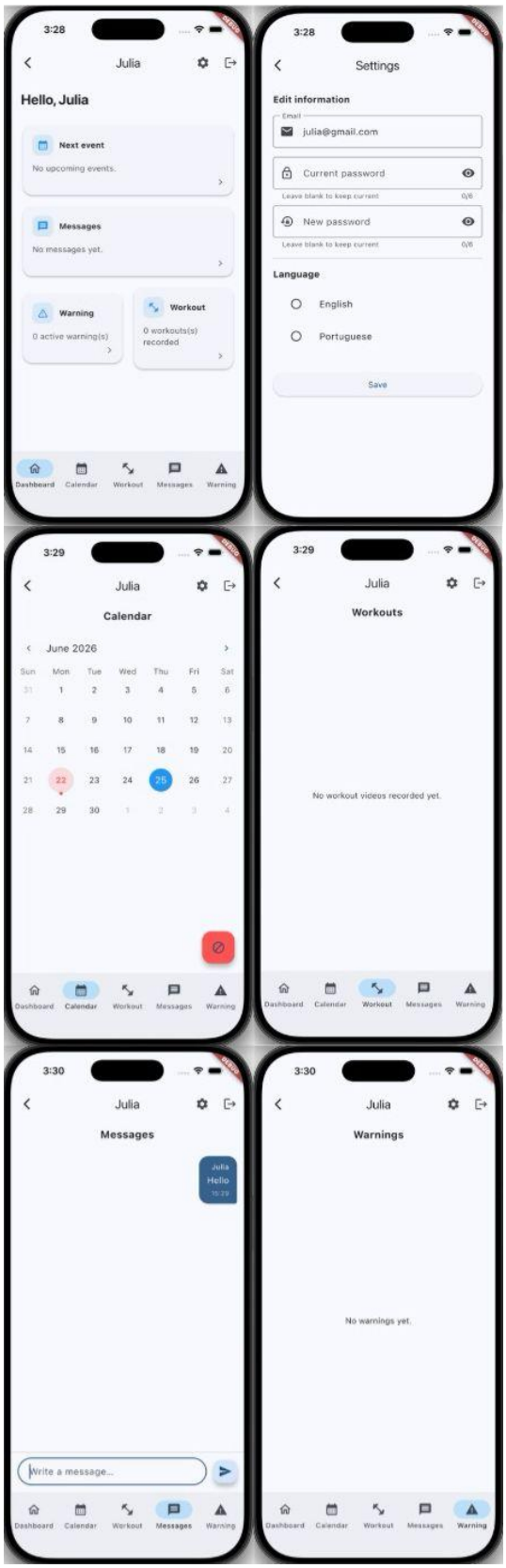


Fig. 4 Patient interface. Up left: a main screen; up down: allows the user to change their password and preferred language); middle left: calendar; middle right: area of prescribed exercises; down left: messages; down right): important notices and reminders.

Overall, the developed system offers features for exercise prescription, clinical monitoring, communication, scheduling, notifications, and physiotherapy follow-up. As can be seen in the images above, two complementary approaches can be observed in a single application with different but mutually beneficial objectives: one interface with the therapist login and another with the patient login, which allows access to information provided by the professional. This organization establishes a simple and effective digital environment focused on supporting the home rehabilitation process.

V. DISCUSSION

The results observed in this study demonstrate the feasibility of developing a telerehabilitation application specifically designed for wheelchair users. The functionalities implemented in the system are in accordance with the recommendations found in the scientific literature on the use of digital technologies in physiotherapy and rehabilitation programs.

The login, exercise prescription and patient monitoring features were developed with the aim of facilitating communication between patient and physiotherapist and increasing treatment adherence, which is in line with the conclusions reported by Chen et al. [8], who highlight the importance of mobile health applications to facilitate remote monitoring. In this way, the inclusion of prescription and real-time patient monitoring resources becomes indispensable, since according to Liu et al. [7], mobile applications can contribute to pain reduction, functional improvement and greater adherence to home rehabilitation programs. In this work, these functionalities were incorporated through interfaces that allow access to prescribed exercises and monitoring of their rehabilitation activities. Another relevant aspect is the communication mechanism between physiotherapists and patients. According to observed studies, continuous interaction and feedback are important factors in maintaining adherence to treatment and encouraging patient involvement in rehabilitation programs [17]. Therefore, the developed application aims to incorporate these principles through messaging and notification features that facilitate therapeutic monitoring.

The app developed also incorporates accessibility principles and digital inclusion strategies, which are mentioned in the articles discussed, as these factors are considered essential for the development of health technologies aimed at people with disabilities [19]. Therefore, these aspects were considered during the planning and implementation phases to ensure simple and clear usability for the public, facilitating access to home rehabilitation resources.

The structure developed in Flutter represents an important technological contribution, as it allows access to rehabilitation from a new perspective, bringing rehabilitation beyond the consulting room and focusing on a specific and select public that is often neglected. In the study by Chen et al. [20], digital rehabilitation solutions integrated with remote monitoring technologies can contribute to improving therapeutic continuity and patient engagement, as well as greater adherence to treatment.

Remote rehabilitation, when compared to previous studies, differs because it involves the specificity of developing an app by a professional in the field and because it is a platform specifically focused on wheelchair users. While many available applications address isolated aspects of rehabilitation [6], the proposed solution combines factors important to the target audience in the same environment.

The inclusion of specific factors for a select audience is also supported by studies that demonstrate effectiveness in the results obtained, as it generates expanded access to physiotherapy services and reduced barriers to treatment [21]. These analyzed findings reinforce the app's potential in rehabilitation contexts.

Despite promising results and technological advancements, some limitations must be considered. The application has not yet undergone usability testing with end users or clinical validation studies. Therefore, it is not yet possible to determine the system's effectiveness in real people, but the results obtained are related to the application's planning and development.

VI. CONCLUSION

In summary, the telerehabilitation application developed to prescribe home exercises for wheelchair users offers functionalities for exercise guidance, communication, scheduling, and monitoring for both the professional and the patient. Therefore, it can be concluded that the application has the potential for technological innovation because it was developed and designed specifically for its target audience.

Although further validation studies and end-user testing are needed to assess its usability, effectiveness, and acceptance in clinical practice, the developed app shows potential to significantly contribute to monitoring the health and quality of life of users. Furthermore, the proposal represents a relevant advancement from a technological and scientific perspective, integrating accessible digital resources that can strengthen health promotion, functional monitoring, and autonomy for this population.

REFERENCES

[1] SOUZA, Carolina Matiello; MARTINS, Jaqueline; OLIVEIRA, Anamaria Siriani de. Self-efficacy and self-management in musculoskeletal rehabilitation: strategies and challenges among Brazilian physiotherapists. Musculoskeletal Care, v. 23, no. 1, e70077, 2025.

[2] World Health Organization. Rehabilitation 2030: a call for action – meeting report. Geneva: World Health Organization, 2020.

[3] Gomes Neto M, Sá-Caputo D, Paineiras-Domingos LL, et al. Effects of physiotherapy interventions in wheelchair users: systematic review. Disability and Rehabilitation. 2023.

[4] WHO – World Health Organization. Global report on health equity for persons with disabilities. Geneva: WHO, 2022.

[5] KEMP, BJ et al. Rehabilitation outcomes in wheelchair users. Disability and Rehabilitation, 2019.

[6] TUROLLA , A. et al. Tele-rehabilitation for people with physical disabilities: systematic review. Journal of Telemedicine and Telecare, 2020.

[7] LIU, S. et al. Mobile app use to support therapeutic exercise for musculoskeletal pain conditions may help improve pain intensity and self-reported physical function: a systematic review. Journal of Physiotherapy, vol. 69, no. 1, p. 14–23, 2023.

[8] CHEN, J. et al. Mobile health applications in rehabilitation: systematic review. Journal of Medical Systems, 2021.

[9] Tweedy SM, Beckman EM, Geraghty TJ. Exercise and rehabilitation in wheelchair-dependent individuals. SportsMedicine. 2020.

[10] WHO – World Health Organization. Global strategy on digital health 2020–2025. Geneva: WHO, 2021.

[11] MEROLLI , M. et al. Evaluation of patient-facing mobile apps to support physiotherapy care: systematic review. JMIR mHealth and uHealth, 2024.

[12] TAYSHETE, G. et al. Effectiveness of smartphone-based applications in low-back pain rehabilitation: a systematic review and meta-analysis. Bulletin of Faculty of Physical Therapy, 2025.

[13] STARK, R. et al. App-based rehabilitation in back pain: a systematic review. Journal of Personalized Medicine, vol. 12, no. 10, 2022.

[14] SZETO, GPY et al. Effect of mobile application types on stroke rehabilitation: a systematic review. Journal of NeuroEngineering and Rehabilitation, vol. 20, 2023.

[15] TUTTLE , C. et al. Use of smartphone apps for improving physical function capacity in cardiac patient rehabilitation: systematic review. JMIRx Med, v. 2, no. 2, 2021.

[16] BONINGER, ML et al. Wheelchair biomechanics and upper limb injuries. Journal of Rehabilitation Research, 2020.

[17] Kairy D, Veras M, Archambault P. Telerehabilitation for individuals with spinal cord injury: a systematic review. Journal of Telemedicine and Telecare. 2021.

[18] World Health Organization. Global Strategy on Digital Health 2020–2025. 2021.

[19] LOPES, JS et al. Accessibility and digital health in people with physical disabilities. Brazilian Journal of Rehabilitation, 2022.

[20] Chen Y, Abel KT, Janecek JT, et al. Home-based telehealth exercise interventions for wheelchair users: systematic review. Journal of Rehabilitation Research and Development. 2022.

[21] COTTRELL, MA et al. Real-time telerehabilitation for musculoskeletal conditions. Journal of Telemedicine and Telecare, 2017.

[22] BENNELL, K. et al. Effectiveness of remotely delivered exercise programs. Physical Therapy Journal, 2021.

.

Teste de inteligência e deficiência visual: uma revisão de literatura

Cristiane Dolores Barboza de Oliveira
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
cristianeoliveira@umc.br

Daieny Panhan Theodório
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
daienytheodoro@umc.br

Alessandro Pereira da Silva
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
alessandrops@umc.br

Silvia Cristina Martini
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
silviac@umc.br

Silvia Regina Matos da Silva Boschi
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
boschi@umc.br

Terigi Augusto Scardovelli
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
terigiscardovelli@umc.br

***Abstract*—Accessible cognitive assessment for people with visual impairment requires instruments whose validity and measurement properties are demonstrated for the target population and whose administration does not introduce avoidable visual barriers. This integrative review aimed to identify and critically compare instruments used to assess cognitive abilities in people with visual impairment, with emphasis on population characteristics, language and cultural adaptation, reference measures, psychometric evidence and accessibility. Searches were conducted from August 29, 2025 to June 21, 2026, in BVS, Cochrane Library, CAPES Journals Portal, PubMed, ScienceDirect and Scopus. The search strategy combined "Intelligence Tests" AND "Visual Impairment". Studies published from 2020 to 2025 in English, Portuguese or Spanish were eligible. Reporting was reorganized according to PRISMA 2020, and methodological quality was critically appraised using a COSMIN-based framework adapted to the measurement studies identified. Three studies met the eligibility criteria: a Chinese study of WISC-IV verbal scales in 100 children who were blind; an Italian pilot study of a three-dimensional haptic nonverbal cognitive test in 25 participants with severe low vision or blindness and 25 sighted peers; and a Pakistani study of the Urdu version of the Tromsø Social Intelligence Scale in 215 adolescents with visual impairment. The evidence indicates promising reliability and validity for specific instruments, but the small number of studies, restricted age ranges, cultural and linguistic specificity, and limited representation of Brazilian populations prevent broad recommendations. The main contribution of this review is to demonstrate the scarcity and heterogeneity of current evidence and to identify priorities for future validation studies involving people with visual impairment throughout instrument development.**

Keywords— intelligence assessment, visual impairment, assistive technology

I. INTRODUÇÃO

A deficiência visual é caracterizada pela perda parcial ou total da visão, o que dificulta a realização de atividades diárias sem recursos adaptativos. Segundo a Organização Mundial da Saúde (OMS), ela é classificada em diferentes graus, desde a baixa visão, quando há percepção visual da necessidade de auxílio, até a cegueira total, com ausência completa da visão. Pode ser congênita, por fatores genéticos, por complicações no parto, adquirida ao longo da vida devido a doenças, traumas ou envelhecimento [1].

O estudo da inteligência sempre constituiu um dos grandes interesses da humanidade ao longo do tempo. As primeiras propostas de aferição das funções mentais remontam ao final do século XIX, quando se buscou avaliar a inteligência por meio de diferentes capacidades sensoriais. A primeira contribuição relevante foi de Francis Galton (1822-1911) que concebia a inteligência como um atributo de base hereditária, investigada a partir do estudo de árvores genealógicas. J. McKeenCattell (1860–1944) deu continuidade a essa linha de pesquisa, desenvolvendo medidas preditoras da inteligência, como velocidade de movimento, força muscular, discriminação de peso e sensibilidade à dor. Posteriormente, verificou-se que essas e outras medidas não eram adequadas para mensurar a inteligência. No início do século XX, Alfred Binet (1857-1911) criticou essa abordagem centrada em processos sensório motores e destacou a necessidade de avaliar processos mentais mais complexos, como atenção, memória, compreensão e imaginação, de modo a permitir uma investigação mais abrangente da inteligência [2].

A concepção de inteligência na abordagem psicométrica fundamenta-se na análise fatorial, a qual se baseia nas diferenças individuais reveladas por uma ampla variedade de testes destinados a avaliar capacidades cognitivas. A análise fatorial tem como objetivo identificar subgrupos de testes que avaliam a mesma habilidade cognitiva. Essa abordagem evoluiu ao longo dos anos: inicialmente, elaboravam-se instrumentos de medida sem clareza teórica sobre o que avaliavam; posteriormente, essas formulações tornaram-se mais sofisticadas. Estudos recentes buscam integrar achados da Psicologia e da Neurociência, relacionando aspectos cognitivos e afetivos [3].

A validação dos inúmeros testes psicométricos construídos nas primeiras décadas do século XX em países da Europa e da América do Norte e introduzidos no Brasil ainda nesse período, consolidou-se como instrumento para definir um padrão de inteligência considerado "normal", mediano e comum. Tais parâmetros serviam para determinar quem deveria ou não ser admitido ou permanecer nas escolas. Essas medidas foram desenvolvidas a partir de teorias fundamentadas na eugenia, cujo objetivo era distinguir supostos tipos superiores e inferiores. Esse movimento influenciou diferentes países, promovendo a aferição das capacidades individuais, visando "a seleção das melhores espécies" por meio da "mensuração das imperfeições" [4].

Os testes de inteligência, portanto, foram concebidos dentro do ideário eugenista e carregam equívocos conceituais decorrentes de sua própria origem, centrada na crença de que

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

seria possível avaliar de maneira objetiva o potencial intelectual de um indivíduo [5].

Entre as teorias psicométricas da inteligência, destaca-se a teoria bifatorial de Charles Edward Spearman (1863-1945), fundamental para a conceituação do fator g, determinado pela comparação entre múltiplos itens de uma prova ou entre pontuações de diferentes testes. O fator g explicaria o funcionamento intelectual, sendo responsável pela maior parte da variância observada em testes psicométricos, com forte componente biológico [2].

Em 1912, William Stern (1871-1938) utilizou o conceito de idade mental para formular o Quociente de Inteligência (QI), permitindo quantificar o funcionamento cognitivo em relação ao desenvolvimento e à aprendizagem. O cálculo do QI infantil consiste em dividir a idade mental (IM), definida pela capacidade de aprender e se desenvolver, pela idade cronológica (IC), correspondente ao tempo de vida em dias, meses e anos. Assim, considera-se desenvolvimento normal quando IM e IC são equivalentes [6].

Piaget, biólogo e geneticista francês, defendia que a inteligência não é hereditária nem inata. Para ele, a inteligência se desenvolve articulada à herança biológica, a partir de estruturas sensoriais ou neurológicas que amadurecem conforme a interação do indivíduo com o ambiente. A inteligência relaciona-se ao aspecto cognitivo ao estruturar as interações entre sujeito e objeto. Nesse processo, o sujeito assimila elementos do meio às suas estruturas internas e, simultaneamente, deve acomodá-las, ajustando-as às novas experiências [7].

Teorias posteriores buscaram fragmentar a compreensão da inteligência em grupos específicos [8,9]. Cattell propôs a divisão em dois fatores, inteligência fluida (Gf) (refere-se à execução de tarefas novas e não automáticas, que demandam a formação de relações e a integração de ideias) e inteligência cristalizada (Gc) que corresponde ao conhecimento adquirido ao longo da vida, resultante de experiências e do acúmulo de aprendizagem.

Horn, em colaboração com Cattell, ampliou a concepção de inteligência ao propor um modelo hierárquico, composto por um fator geral e múltiplos fatores específicos, a teoria CHC (Cattell-Horn-Carroll), que constitui a formulação mais atual e abrangente sobre inteligência, sustentado por inúmeras evidências empíricas e por instrumentos específicos destinados à sua mensuração [10].

Existem diversos testes psicométricos recomendados pelo Conselho Federal de Psicologia, destinados à aferição da inteligência por meio do Sistema de Avaliação de Testes Psicológicos (SATEPSI). Atualmente, os mais aplicados e comercializados no mercado estão disponibilizados na Tabela 1.

Os testes de inteligência, assim como outros instrumentos psicológicos, podem ser utilizados para auxiliar no diagnóstico e na identificação das funções cognitivas mais comprometidas, possibilitando o direcionamento de intervenções voltadas ao estímulo das capacidades do indivíduo [11].

Na década de 1950, foi publicada uma análise crítica que reuniu dezenove estudos empíricos, demonstrando que nenhuma intervenção psicológica se mostrava mais eficaz do que o simples passar do tempo [12]. Essa publicação representou um marco relevante, pois fomentou debates e impulsionou um número expressivo de pesquisas voltadas à comprovação da eficácia das psicoterapias [13].

Dessa forma, o objetivo geral deste estudo é compreender os pressupostos teóricos que fundamentam as práticas de aferição da inteligência humana, bem como promover ações que assegurem à população com deficiência visual o direito de utilizar testes de inteligência adaptados às suas necessidades.

II. METODOLOGIA

A. Critérios de Elegibilidade

Foram elegíveis artigos originais publicados entre 2020 e 2025, nos idiomas inglês, português ou espanhol, que investigassem instrumentos de avaliação cognitiva em pessoas com deficiência visual e apresentassem informações sobre validade, confiabilidade, adaptação, desempenho ou acessibilidade do instrumento. Foram consideradas populações com cegueira ou baixa visão, sem restringir previamente a faixa etária, desde que a deficiência visual estivesse explicitamente caracterizada. Foram excluídos registros duplicados, editoriais, artigos incompletos, teses, dissertações, livros e estudos sem relação direta com a avaliação cognitiva de pessoas com deficiência visual. A definição de elegibilidade foi aplicada ao título/resumo e posteriormente confirmada na leitura do texto completo.

B. Fontes de Informação

As fontes de informação foram o Portal Regional da BVS, Cochrane Library, Portal de Periódicos CAPES, PubMed, ScienceDirect e Scopus. A busca foi realizada entre 29 de agosto de 2025 e 21 de junho de 2026. O último acesso/consulta de cada fonte ocorreu em 21 de junho de 2026. A estratégia registrada para todas as fontes foi a combinação dos termos “Intelligence Tests” AND “Visual Impairment”, com adaptações de sintaxe quando necessárias às respectivas plataformas. A APA-PsycNet constava em uma versão anterior da metodologia, mas não apresentou registros no fluxo contabilizado; por isso, não foi considerada uma fonte efetivamente contabilizada nos resultados desta versão.

C. Critérios de inclusão e exclusão

Foram incluídos artigos originais que investigassem instrumentos ou testes destinados à avaliação de algum domínio cognitivo em pessoas com deficiência visual, apresentassem dados sobre propriedades psicométricas, adaptação ou desempenho do instrumento e fossem publicados no intervalo definido. Foram excluídos duplicatas, editoriais, trabalhos incompletos, teses, dissertações, livros e estudos sem relação direta com avaliação cognitiva e deficiência visual.

D. Processo de seleção dos estudos

A seleção ocorreu em três etapas: leitura de títulos, leitura de resumos e avaliação do texto completo. O processo de seleção e análise dos artigos seguiu uma metodologia rigorosa, contando com a participação de dois pesquisares independentes para seleção dos estudos. Inicialmente, realizou-se uma triagem com base nos títulos, seguida por uma análise mais aprofundada dos resumos e, posteriormente, a leitura completa dos textos selecionados.

E. Síntese dos resultados

Devido ao pequeno número de estudos e à heterogeneidade dos construtos, populações, idiomas, instrumentos e propriedades psicométricas avaliadas, não foi realizada meta-análise. Os resultados foram sintetizados de forma narrativa e

comparativa, organizados segundo instrumento, construto, população, contexto cultural, acessibilidade e propriedade de medida. Quando disponíveis, foram apresentados coeficientes de confiabilidade, evidências de validade e características da adaptação. A direção da evidência foi interpretada como favorável, promissora ou limitada de acordo com o desenho e os resultados reportados, evitando a generalização dos achados para populações não estudadas.

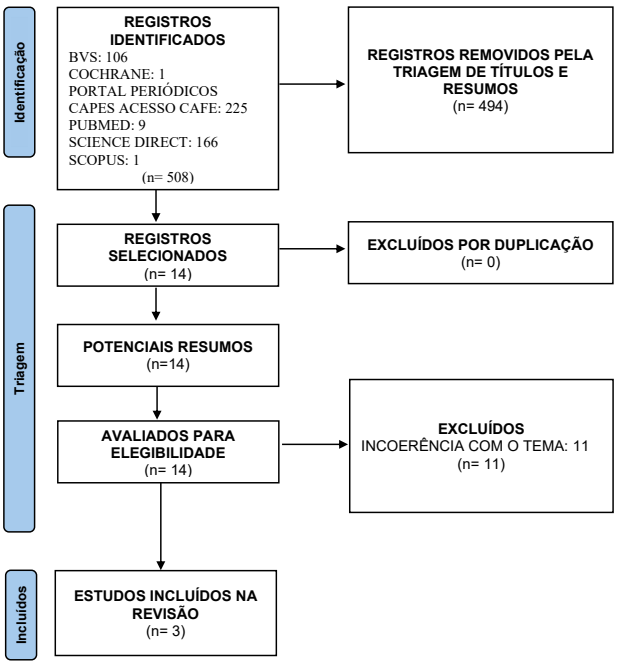
F. Risco de viés e qualidade metodológica

A qualidade metodológica e o risco de viés foram apreciados por meio de uma estrutura baseada no checklist COSMIN (*COnsensus-based Standards for the selection of health Measurement INstruments*) para propriedades de medida, considerando, conforme aplicável, validade de conteúdo, validade estrutural, consistência interna, confiabilidade, teste de hipóteses/validade de construto e validade transcultural. Foram também considerados aspectos de adequação da população, tamanho amostral, clareza dos procedimentos e completude do relato. Como os instrumentos incluídos abrangem testes de inteligência e uma escala de inteligência social, e não exclusivamente instrumentos de desfechos relatados pelos pacientes, o COSMIN foi empregado como estrutura de apreciação das propriedades de medida, e não como certificação de uso clínico.

III. Resultados e Discussão

A busca contabilizada no fluxo resultou em 508 registros: 106 na BVS, 1 na Cochrane Library, 225 no Portal de Periódicos CAPES, 9 na PubMed, 166 na ScienceDirect e 1 na Scopus. Após a triagem de títulos e resumos, 14 registros permaneceram para avaliação de elegibilidade. Não foram identificadas duplicatas. Na leitura dos textos completos, 11 estudos foram excluídos por incoerência com o tema, resultando em 3 estudos incluídos [6–8]. O fluxograma revisado é apresentado na Figura 1.

Figura 1: Diagrama dos artigos incluídos e excluídos, de acordo com os critérios de inclusão e exclusão estabelecidos no estudo.



Atualização das teorias psicométricas e da avaliação cognitiva

O panorama contemporâneo das teorias psicométricas da inteligência amplia os modelos hierárquicos clássicos ao incorporar abordagens dinâmicas e perspectivas derivadas da neurociência e da ciência de dados. O modelo Cattell-Horn-Carroll (CHC) permanece uma referência importante para a organização de habilidades cognitivas, especialmente ao distinguir capacidades fluidas e cristalizadas. Paralelamente, abordagens como a Teoria do Mutualismo, modelos de redes cognitivas e a Teoria da Resolução de Processos Sobrepostos (POT) discutem a possibilidade de que o fator g represente um fenômeno emergente das interações entre múltiplos processos cognitivos, em vez de uma causa única centralizada. Essas perspectivas são relevantes para a interpretação dos testes porque reforçam a necessidade de definir claramente o construto que cada instrumento pretende medir.

Trabalhos recentes também aproximam psicometria, neurociência e funcionamento de redes cerebrais, discutindo dimensões múltiplas da inteligência e relações entre desempenho cognitivo, conectividade neural e eficiência do processamento. Em paralelo, a psicometria computacional tem explorado aprendizado de máquina e modelos matemáticos para analisar padrões comportamentais complexos, modelar diferenças individuais e ampliar estratégias de avaliação. No entanto, a incorporação de IA à avaliação cognitiva não elimina a necessidade de evidências psicométricas tradicionais; ao contrário, exige avaliação adicional de validade, viés algorítmico, transparência, estabilidade e equidade. Para pessoas com deficiência visual, essa evolução pode ser particularmente relevante quando combinada a leitores de tela, síntese de voz, interfaces hápticas, dispositivos táteis e sistemas adaptativos. Assim, a inovação tecnológica deve ser considerada uma estratégia de acessibilidade e não um substituto da validação do instrumento. A tabela 1 apresenta os testes psicométricos recomendados pelo Conselho Federal de Psicologia, destinados à aferição da inteligência por meio do Sistema de Avaliação de Testes Psicológicos (SATEPSI) disponíveis atualmente.

Tabela 1: Sistema de Avaliação de Testes Psicológicos mais aplicados e comercializados no mercado.

TESTE	*PÚBLICO*	*OBJETIVO*	*CARACTERÍSTICAS*
Son-R	2 ½ a 7 anos e dos 6-40 anos	Avaliação geral das habilidades cognitivas e do desenvolvimento, contribuindo para o diagnóstico diferencial entre síndromes neurológicas e psicológicas.	Instrumento não verbal, que reduz a influência da escolaridade no processo avaliativo. O teste utiliza quatro subtestes para examinar habilidades espaciais, visomotoras e raciocínio abstrato e concreto.
Escala de Maturidade Mental Colúmbia 3 (CMMS-3)	3 anos e 0 meses a 9 anos e 11 meses de idade.	Fornecer uma estimativa da habilidade intelectual ou da capacidade de raciocínio.	Avaliação de crianças que apresentam dificuldades em testes tradicionais de inteligência baseados em linguagem ou desempenho motor.
Escala de Inteligência Wechsler (WISC - IV)	6 anos e 0 meses até16 anos e 11 meses	Auxiliar no diagnóstico de dificuldades de aprendizagem, transtornos do neurodesenvolvimento (como TDAH) e na identificação de ne-	Bateria padronizada e administrada individualmente, composta por múltiplos subtestes organizados em quatro índices: Índice de Compreensão Verbal

		cessidades de intervenção pedagógica ou psicológica.	(ICV); Índice de Organização Perceptual (IOP); Índice de Memória Operacional (IMO); e Índice de Velocidade de Processamento (IVP).
Escala de Inteligência Weschler para Adultos (WAIS – III)	adolescentes e adultos com idades entre 16 e 89 anos.	Avaliar a capacidade intelectual e o funcionamento cognitivo global do indivíduo.	Permite a obtenção de resultados quantitativos referentes a quatro domínios cognitivos: compreensão verbal, memória operacional, organização perceptual e velocidade de processamento.
Escala de Inteligência Wechsler Abreviada (WASI)	Voltados para o atendimento de crianças, adolescentes, adultos e idosos (06 a 89 anos).	Examinar diferentes aspectos cognitivos, incluindo conhecimento verbal, processamento visual, raciocínio espacial e não verbal, bem como inteligência fluida e cristalizada em diversos contextos.	Associação com a WAIS-III e a WISC-III, oferecendo tabelas para estimativa das faixas de QIT. Na adaptação brasileira, também fornece tabelas para estimativa das faixas de QIT na WISC-IV.
Matrizes Progressivas Coloridas de Raven (CPM)	Crianças entre 05 a 11 anos e 11 meses de idade.	Mensurar a inteligência (Fator g) e fornecer informações sobre a capacidade educativa.	Avalia a habilidade de gerar novos insights — predominantemente não verbais — em situações complexas ou familiares, permitindo ao indivíduo ir além das informações disponíveis e identificar elementos que não são imediatamente óbvios.
O Matrizes Progressivas Avançadas de Raven (APM)	Entre 17 e 63 anos e escolaridade mínima de Ensino Superior em andamento.	Avaliar a inteligência e oferecer dados sobre a capacidade educativa de indivíduos com escolaridade mínima de Ensino Superior em andamento.	Avaliação da inteligência geral e pode ser aplicada em contextos clínicos ou em qualquer situação que exija mensuração dessa habilidade (padrão-ouro).

Caracterização dos estudos

Os estudos apresentam populações distintas, Chen *et al.* [6] avaliaram exclusivamente estudantes cegos na China, enquanto Cassar *et al.* [7] incluíram jovens com baixa visão severa ou cegueira e um grupo de pares videntes. Noor *et al.* [8] estudaram adolescentes com deficiência visual matriculados em escolas públicas para pessoas cegas na região de Rawalpindi, no Paquistão. Assim, os resultados não representam de forma homogênea crianças, adultos, idosos, pessoas com diferentes graus de baixa visão ou indivíduos com deficiências adicionais.

A diversidade linguística também é central. O estudo de Chen *et al.* [16] utilizou a versão chinesa da WISC-IV; Cassar *et al.* [17] trabalhou com instrumentos em contexto italiano; e Noor *et al.* [18] realizou tradução e retrotradução para o Urdu, com avaliação por especialistas e teste piloto. No estudo paquistanês, a estrutura fatorial identificada diferiu da estrutura de três fatores da versão original, demonstrando que a adaptação linguística não garante equivalência estrutural entre culturas.

A classificação é uma síntese crítica das informações disponíveis nos estudos e não deve ser interpretada como uma certificação do instrumento. O principal achado metodológico é a concentração das evidências em poucas propriedades: Chen *et al.* [6] apresentaram evidências de consistência interna e estrutura fatorial; Cassar *et al.* [7] concentraram-se em validade de construto em um estudo piloto; e Noor *et al.* [8] apresentaram conjunto mais amplo de evidências, incluindo adaptação linguística, consistência interna, confiabilidade e estrutura fatorial. Os dados referentes a qualidade metodológica dos estudos estão descritos na tabela 2.

Tabela 2: Avaliação de Risco de Viés (COSMIN).

Estudo	Consistência interna	Confiabilidade	Validade estrutural	Validade de construto	Validade transcultural	Amostra/desenho	Síntese COSMIN
Chen *et al.* [6]	Boa	Boa	Boa	Boa	Não estabelecida	Boa para VCI/WMI	Boa, mas restrita a crianças cegas chinesas
Cassar *et al.* [7]	NR	NR	NR	Boa	Não aplicável como estudo de adaptação	Adequada como estudo piloto	Adequada com limitações pelo n=25 no grupo com deficiência visual
Noor *et al.* [8]	Boa	Boa	Boa	Boa	Boa, com adaptação linguística	Boa	Boa para adolescentes com deficiência visual

Os artigos selecionados abordam temas diretamente relacionados à avaliação cognitiva em pessoas com deficiência visual. Os 3 estudos passaram pela avaliação qualitativa, e com a finalidade de abordar detalhadamente os estudos científicos foi elaborada uma tabela descritiva elucidando o objetivo do estudo, teste utilizado, desfecho, amostra e conclusão do estudo (Tabela 3). Os dados foram organizados em ordem de ano de publicação.

Tabela 3: Avaliação descritiva dos estudos selecionados.

Autor	Objetivo	Teste	Desfecho	n	Conclusão
Chen *et al.* [16]	Avaliar as propriedades psicométricas das escalas verbais do WISC-IV em estudantes cegos na China.	WISC-IV e escala cognitiva não verbal.	Alta confiabilidade (0,79–0,97) e validade; correlação moderada com medidas não verbais.	100	O WISC-IV é válido e confiável para avaliação cognitiva de crianças cegas.
Cassar *et al.* [17]	Desenvolver e testar um teste de inteligência não verbal haptico para jovens com deficiência visual.	3-DHMT-CA, Standard Progressive Matrices e WISC-IV	Boa validade convergente e divergente; diferenças entre grupos com e sem deficiência visual.	50	O 3-DHMT-CA é promissor para avaliação cognitiva não verbal, mas precisa de validação em amostras maiores.

Noor *et al.* [18]	Validar a versão em urdu da Tromso Social IntelligenceScale para adolescentes com deficiência visual no Paquistão.	Tromsø Social IntelligenceScale	Alta confiabilidade (α = 0,93), com boa consistência interna, teste-reteste e estrutura fatorial de 4 componentes.	215	A versão em urdu da TSIS é válida e confiável, sendo adequada para avaliar inteligência social em adolescentes com deficiência visual.

A principal contribuição desta revisão é demonstrar que a literatura recente sobre avaliação cognitiva acessível para pessoas com deficiência visual permanece escassa: apenas três estudos atenderam aos critérios estabelecidos. Esse resultado impede recomendações amplas e reforça a necessidade de interpretar cada evidência em função do instrumento, construto, idade, condição visual e contexto cultural. Os estudos incluídos avaliaram diferentes aspectos da cognição, abrangendo compreensão verbal, memória de trabalho, raciocínio não verbal e inteligência social, o que evidencia a heterogeneidade dos instrumentos disponíveis e impede sua consideração como medidas intercambiáveis.

Os resultados do estudo de Chen et al. (2021) [16] avaliaram o uso da quarta edição da Escala de Inteligência Wechsler para Crianças (WISC-IV) quanto à confiabilidade e validade dos escores das escalas verbais, especificamente o Índice de Compreensão Verbal (ICV) e o Índice de Memória de Trabalho (IMT), em crianças cegas na China. As avaliações foram realizadas utilizando a versão chinesa das Escalas de Desenvolvimento Cognitivo Verbal e Não Verbal do WISC-IV. Os achados confirmaram a consistência interna e a validade dos escores para essa população. Foram encontrados coeficientes de confiabilidade por split-half entre 0,79 e 0,97 e coeficientes de consistência interna entre 0,83 e 0,97 para as escalas analisadas. A validade de construto foi apoiada por análise fatorial exploratória e pela correlação moderada e positiva com a escala cognitiva não verbal. Conforme apontado por MacCallum et al. (1999), mesmo em amostras pequenas, com menos de 100 participantes, valores elevados de comunalidade e a presença de poucos fatores podem favorecer a aplicação da análise fatorial exploratória [16].

Esses resultados são relevantes para estudantes cegos na China, mas não permitem concluir que os mesmos parâmetros sejam válidos para crianças brasileiras, pessoas com baixa visão ou outras faixas etárias. A evidência também se refere às escalas verbais e não à totalidade do instrumento. Dessa forma, os resultados devem ser interpretados considerando as características específicas da população investigada e do contexto em que o instrumento foi aplicado. A validade e a confiabilidade observadas em determinada população não podem ser automaticamente generalizadas para indivíduos com diferentes condições visuais, características culturais, linguísticas ou faixas etárias.

No estudo de Cassar et al. (2022) [17], foi realizado um teste piloto de inteligência não verbal para crianças e adolescentes italianos com deficiência visual, incluindo participantes com cegueira ou baixa visão, com idades entre 10 e 16 anos. Participaram 25 indivíduos com deficiência visual, sendo 16 meninos e 9 meninas. O 3-DHMT-CA apresenta uma contribuição tecnológica diretamente relacionada à acessibilidade, por utilizar 22 matrizes tridimensionais táteis que permitem explorar os estímulos por meio da percepção háptica, reduzindo a dependência de informação visual. O participante deve identificar o padrão lógico subjacente aos estímulos táteis tridimensionais, manipulados simultaneamente com ambas as mãos.

A validade convergente do instrumento foi examinada por meio das Standard Progressive Matrices, enquanto a validade divergente foi analisada utilizando os índices verbal e de memória de trabalho da WISC-IV [17]. Entretanto, a amostra de participantes com deficiência visual foi pequena, caracterizando o estudo como piloto, e os próprios autores reconhecem a necessidade de estudos com amostras maiores e de análises qualitativas das estratégias utilizadas durante a resolução dos itens. Segundo Miller et al. (2007) [19], o desempenho em tarefas desse tipo pode estar associado à experiência visual, de modo que participantes capazes de formar imagens mentais da matriz, aqueles com visão normal ou baixa visão menos severa, tendem a obter pontuações superiores. Apesar disso, o estudo destacou estratégias compensatórias relacionadas à memória verbal e de trabalho em jovens com deficiência visual. Dessa forma, profissionais da educação e da reabilitação podem direcionar intervenções para o fortalecimento dessas habilidades, considerando que influenciam outros domínios cognitivos.

Embora o 3-DHMT-CA represente uma alternativa promissora para reduzir a dependência de estímulos visuais, seus resultados ainda não permitem classificá-lo como instrumento estabelecido para uso amplo. A amostra reduzida e as características específicas da população estudada limitam a generalização dos achados para outras populações com deficiência visual. Além disso, a influência da experiência visual e das estratégias cognitivas utilizadas durante a resolução demonstra que o desempenho em testes não verbais e hápticos pode envolver diferentes mecanismos compensatórios, os quais precisam ser considerados na interpretação dos resultados [17], [19].

Já no estudo de Noor et al. (2024) [18], realizado no Paquistão, foi avaliada a validade e confiabilidade da versão em urdu da Escala de Inteligência Social de Tromsø (TSIS), instrumento de autorrelato desenvolvido por Silvera et al. (2001) [20], para adolescentes com deficiência visual. Participaram 215 adolescentes, sendo 108 meninos e 107 meninas, matriculados em escolas públicas para cegos na divisão de Rawalpindi. O estudo foi desenvolvido em duas fases: tradução e validação da escala. A tradução direta e reversa para o urdu foi realizada pelo pesquisador com apoio de cinco especialistas bilíngues inglês–urdu, seguida de avaliação de conteúdo e face e teste piloto.

A versão Urdu da TSIS apresentou α=0,93, além de resultados favoráveis para teste-reteste, split-half e confiabilidade interavaliadores. Os resultados indicaram que o TSIS-U é um instrumento válido e confiável para essa população. Contudo, verificou-se que a estrutura fatorial de quatro fatores apresentou diferenças em relação ao modelo original desenvolvido em contexto ocidental [18], demonstrando que a equivalência transcultural não pode ser presumida. Esse resultado é particularmente relevante para o contexto brasileiro, uma vez que a tradução ou adaptação de um instrumento estrangeiro exige estudos próprios de validade, confiabilidade e estrutura interna na população brasileira.

A comparação entre os três estudos evidencia diferentes concepções de cognição. O estudo de Chen et al. [16] concentra-se principalmente em compreensão verbal e memória de trabalho; o estudo de Cassar et al. [17] aborda o raciocínio não verbal; e o estudo de Noor et al. [18] avalia a inteligência social. Portanto, os três instrumentos não são intercambiáveis. Um escore relacionado à inteligência geral ou a determinados índices cognitivos não pode ser diretamente comparado com uma medida de inteligência social ou com um teste específico de raciocínio não verbal sem considerar o construto avaliado, a população e as condições de aplicação.

A acessibilidade também não deve ser confundida com validade psicométrica. Um instrumento pode apresentar uma forma de apresentação acessível e, ainda assim, não possuir evidências suficientes de validade para determinada população. Da mesma forma, uma boa confiabilidade não demonstra, isoladamente, que o instrumento mede adequadamente o construto pretendido. A escolha de instrumentos para pessoas com deficiência visual deve, portanto, considerar simultaneamente acessibilidade, validade, confiabilidade, normas de interpretação e adequação à população.

As tecnologias digitais representam uma oportunidade para superar algumas dessas limitações. Interfaces compatíveis com leitores de tela, síntese de voz, braille digital, recursos táteis e hápticos, além de sistemas adaptativos, podem permitir diferentes modalidades de apresentação e resposta. A acessibilidade digital deve ser planejada desde o desenvolvimento do instrumento e não apenas adicionada posteriormente [4], [5]. Sistemas computadorizados e algoritmos de inteligência artificial também podem apoiar a adaptação da dificuldade, o registro de respostas e a análise de padrões cognitivos. Entretanto, sua adoção requer estudos de equivalência, avaliação de possíveis vieses algorítmicos, segurança, transparência e validade antes de sua utilização clínica ou profissional.

No Brasil, a aplicabilidade prática possui ainda uma dimensão regulatória. O SATEPSI é o sistema utilizado pelo Conselho Federal de Psicologia para avaliar a qualidade técnico-científica de instrumentos psicológicos e informar as condições de uso profissional [3]. Assim, os achados internacionais desta revisão não devem ser interpretados como recomendação para utilização profissional de instrumentos estrangeiros no país. A decisão sobre uso, administração e interpretação deve respeitar a regulamentação vigente, o manual técnico e as evidências disponíveis para a população e a finalidade pretendidas.

A participação de pessoas com deficiência visual no desenvolvimento e na validação dos instrumentos também merece maior atenção. Além da comparação de escores psicométricos, estudos futuros devem investigar a compreensão das instruções, o conforto, o tempo de aplicação, as barreiras de acesso, as estratégias utilizadas, a preferência por modalidades de resposta e os possíveis efeitos da experiência prévia com tecnologia. Essa abordagem pode contribuir para evitar que um instrumento formalmente acessível permaneça pouco utilizável na prática.

A heterogeneidade observada também impede inferências sobre qualidade de vida, autonomia ou inclusão social, pois esses desfechos não foram avaliados diretamente pelos três estudos incluídos. O benefício potencial da acessibilidade deve, portanto, ser discutido como hipótese ou implicação para pesquisas futuras, e não como resultado demonstrado por esta revisão.

Em síntese, os três estudos apresentam evidências promissoras de confiabilidade, validade e acessibilidade para diferentes instrumentos e populações com deficiência visual, mas a base de evidências permanece limitada e heterogênea. Os resultados de Chen et al. [16], Cassar et al. [17] e Noor et al. [18] devem ser interpretados de acordo com suas respectivas populações, construtos, contextos culturais e características metodológicas. Dessa forma, ainda não é possível recomendar um instrumento único ou estabelecer parâmetros universais para avaliação cognitiva de pessoas com deficiência visual. São necessários estudos com amostras maiores e diversificadas, diferentes faixas etárias e condições visuais, além de processos rigorosos de adaptação transcultural e validação para o contexto brasileiro. A ampliação do uso de tecnologias assistivas, interfaces táteis e hápticas e recursos digitais poderá contribuir para avaliações mais acessíveis, desde que acompanhada de evidências psicométricas e de critérios técnicos e regulatórios adequados.

IV. Conclusão

Esta revisão buscou evidenciar a relevância e a necessidade de instrumentos de avaliação da inteligência adaptados para pessoas com deficiência visual. A ausência de tais instrumentos limita oportunidades e compromete a qualidade de vida dessa população, restringindo o desenvolvimento de suas habilidades e competências. Observa-se ainda que boa parte do campo científico em avaliação psicológica permanece voltado a populações sem dificuldades perceptuais. Nessa perspectiva, a integração com a Tecnologia Assistiva é fundamental, uma vez que seus recursos, dispositivos e sistemas ampliam autonomia e independência, favorecem intervenções mais precisas e promovem a inclusão social, evitando que essas pessoas permaneçam à margem da sociedade.

Referências

[1] ORGANIZAÇÃO MUNDIAL DA SAÚDE. Relatório mundial sobre visão. 2019.

[2] CORREIA, Carina Maria Ribeiro Alves. *Avaliação da atenção e da inteligência numa amostra de sujeitos com incapacidade intelectual.*2012. Tese de Doutorado. Universidade da Beira Interior (Portugal).

[3] PRIMI, Ricardo. Inteligência: avanços nos modelos teóricos e nos instrumentos de medida. *Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment*, v. 2, n. 1, p. 67-77, 2003.

[4] JUNIOR, Durval Wanderbroock. A educação sob medida: os testes psicológicos eo higienismo no Brasil, 1914-45. *Eduem*, 2009.

[5] MOYSÉS, Maria Aparecida Affonso; COLLARES, Cecília Azevedo Lima. Inteligência abstraída, crianças silenciadas: as avaliações de inteligência. *Psicologia Usp*, v. 8, n. 1, p. 63-89, 1997.

[6] Campos CR. Avaliação da inteligência para crianças deficientes visuais: construção de subtestes e investigação de suas qualidades psicométricas [tese]. *Campinas: Pontifícia Universidade Católica de Campinas, Centro de Ciências da Vida, Pós-Graduação em Psicologia*; 2017. 283 p. Disponível em:Acesso em: 15 nov. 2025.

[7] Piaget J. *Psicologia da inteligência*. 4. ed. Paris; 1956.

[8] SCHELINI, Patrícia Waltz; WECHSLER, Solange. Estudo da estrutura fatorial da Bateria Multidimensional de Inteligência Infantil. *Estudos de Psicologia (Campinas),* v. 23, n. 2, p. 105-112, 2006.

[9] NAKANO, Tatiana de Cássia et al. Bateria para avaliação das altas habilidades/superdotação: análise dos itens via Teoria de Resposta ao Item. *Estudos de psicologia (Campinas),* v. 32, n. 4, p. 729-741, 2015.

[10] MCGREW, Kevin S.; FLANAGAN, Dawn P. The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment. Allyn & Bacon, 1998.

[11] MARRUS, Natasha; HALL, Lacey. Intellectual disability and language disorder. *Child and Adolescent Psychiatric Clinics*, v. 26, n. 3, p. 539-554, 2017.

[12] Calzavara JVS. Saúde e acessibilidade: a realidade dos deficientes visuais no Brasil. In: *Anais do I Congresso Internacional de Educação: Diversidade, Formação e Saberes Docentes*; 2018; Montes Claros, MG, Brasil. Campinas: Galoá; 2018.

[13] EYSENCK, Hans Jurgen. The effects of psychotherapy: An evaluation. In: *The Experimental Study of Freudian Theories (Psychology Revivals).* Routledge, 2013. p. 365-384.

[14] BARLOW, David H.; BOSWELL, James F.; THOMPSON-HOLLANDS, Johanna. Eysenck, Strupp, and 50 Years of Psychotherapy Research: A Personal Perspective. *Psychotherapy*, v. 50, n. 1, p. 77-87, 2013.

[15] Gil AC. *Como elaborar projetos de pesquisa*. 4ª ed. São Paulo: Atlas; 2002.

[16] CHEN, Xiaomeng et al. Psychometric properties of WISC-IV verbal scales: A study of students in China who are blind. *Journal of Visual Impairment & Blindness*, v. 115, n. 3, p. 228-241, 2021.

[17] CASSAR, Carolina et al. A haptic nonverbal cognitive test for children and adolescents with visual impairment. *Journal of Visual Impairment & Blindness*, v. 116, n. 4, p. 485-495, 2022.

[18] NOOR, Hina et al. Reliability and validity of Urdu version of Tromso Social Intelligence Scale for visually impaired adolescents. *British Journal of Visual Impairment*, v. 44, n. 1, p. 216-225, 2024.

[19] MILLER, Lucy Jane et al. Concept evolution in sensory integration: A proposed nosology for diagnosis. *The American journal of occupational therapy*, v. 61, n. 2, p. 135, 2007.

[20] SILVERA, David; MARTINUSSEN, Monica; DAHL, Tove I. The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. *Scandinavian journal of psychology*, v. 42, n. 4, p. 313-319, 2001.

Evaluating the Effects of TMS and SSVEP in Post-Stroke Patients with Loss in Visual Field

Matheus Modolo Ferreira
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
matheus.m.ferreira@edu.ufes.br

Kevin Mateo Gutiérrez-Ortiz
Postgraduate Program in Elect. Eng.
Federal University of Espírito Santo
Vitória, Brazil
kevin.ortiz@edu.ufes.br

Sheila da Luz Schreider
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
sheiladaluz@gmail.com

Fernando Bermudes Cabral
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
cabralneuro@gmail.com

Fernanda Vaz de Souza
Postgraduate Program in Biotechnology
Federal University of Espírito Santo
Vitória, Brazil
fernandavaz.souza@hotmail.com

Teodiano Freire Bastos Filho
Postgraduate Program in Elect. Eng.
Federal University of Espírito Santo
Vitória, Brazil
teodiano@gmail.com

***Abstract*—Post-stroke Homonymous Hemianopia (HH) is a frequent visual impairment that significantly compromises functional independence. This pilot study investigated the effects of repetitive Transcranial Magnetic Stimulation (rTMS) on cortical responses measured by electroencephalography (EEG) using a Steady-State Visual Evoked Potential (SSVEP)-based visual field assessment. Two patients with chronic post-stroke left underwent four rTMS sessions delivered over the lesioned occipital cortex. EEG responses were recorded before and after the intervention using both a conventional central-fixation SSVEP paradigm and a novel peripheral visual stimulation protocol. Spectral power and variability were analyzed in Theta, Low Beta, and High Beta frequency bands. The peripheral protocol produced more stable recordings with substantially lower signal dispersion than the conventional approach, reducing contamination by electromyographic (EMG) and oculomotor artifacts. Following rTMS, both participants demonstrated increased Low Beta power under the peripheral protocol, suggesting enhanced cortical synchronization and improved responsiveness of preserved visual pathways. Theta-band activity decreased after stimulation, indicating reduced pathological slow-wave activity and possible recovery of perilesional cortical function. In contrast, High Beta responses were heterogeneous, reflecting individual differences in cortical entrainment capacity and neuroplastic potential. Overall, the combination of rTMS with peripheral SSVEP assessment provided more reliable electrophysiological measurements than the conventional paradigm and showed promising sensitivity for monitoring neurophysiological changes after stroke. Although the findings are limited by the small sample size, they support the feasibility of this approach as an objective tool for assessing visual cortical function during rehabilitation. Larger longitudinal studies incorporating computerized visual perimetry are warranted to confirm whether the observed electrophysiological improvements correspond to meaningful functional recovery.**

Keywords—Stroke, SSVEP, rTMS, homonymous hemianopia.

I. INTRODUCTION

The clinical diagnosis of stroke is primarily based on the persistence of symptoms such as impaired consciousness, language disturbances, visual field deficits, extraocular motility abnormalities, visual neglect, motor weakness, and sensory loss following an acute cerebral infarction [1].

A study involving post-stroke patients, with a mean time since stroke ranging from 22 days to 3 months, demonstrated that approximately 915 patients developed some form of visual impairment resulting from either ischemic or hemorrhagic cerebrovascular events. Lesions directly affecting the occipital lobe were the most frequent, accounting for 54% of patients with post-stroke visual impairment. Among the persistent visual deficits, visual field loss (hemianopia) was the most common manifestation [2]. Stroke-related HH is characterized by damage to the occipital cortex or any portion of the retrochiasmal visual pathway extending to the primary visual cortex (V1), impairing visual information processing and resulting in the loss of the same visual hemifield in both eyes [3].

SSVEP is a widely used technique in Brain–Computer Interface (BCI) systems for assessing the response of the visual cortex to periodic visual stimulation. This technique is based on the individual's fixation on a flickering visual stimulus oscillating at a predetermined frequency, which elicits electrical brain responses at the same frequency as the stimulus. These responses are recorded using EEG are predominantly detected over the occipital region, where the visual cortex is located [4].

rTMS is a non-invasive neuromodulation technique that uses magnetic pulses to induce electrical currents in the cerebral cortex, thereby modulating cortical excitability and promoting neuronal plasticity processes [7]. Considering that brain regions affected by stroke frequently exhibit reduced cortical electrical activity, a decrease in the amplitude of signals, is commonly observed when compared with unaffected brain regions [5,6].

In the present study, the SSVEP technique combined with EEG was used as a method to assess the functional condition of the occipital visual cortex. The aim of this investigation was to compare the amplitudes of EEG signals recorded from the hemisphere corresponding to the damaged visual cortex before and after rTMS intervention in post-stroke patients. An increase in signal amplitude may indicate improved cortical activity and responsiveness in the affected occipital region and, consequently, possible changes associated with neuroplasticity processes.

II. Methodology

A. Recruitment and Participants

This study was designed as a pilot, controlled clinical trial with assessor blinding and was conducted at the Espírito Santo Physical Rehabilitation Center (CREFES) in Vila Velha, Brazil. All participants provided written informed consent before enrollment in the study.

Individuals were eligible for inclusion if they had a diagnosis of stroke, were 18 years of age or older, presented with residual hemiparesis or hemiplegia, and had preserved comprehension and communication abilities. Exclusion criteria included a history of epilepsy or seizures, photosensitive epilepsy, severe ocular immobility, previous craniectomy, intracranial metallic implants, or any contraindication to rTMS.

For this pilot analysis, two patients with stroke-related left homonymous visual field loss were recruited. Patient 1, a 46-year-old individual with a large right nucleocapsular hemorrhagic stroke. Patient 2, a 77-year-old individual with an ischemic stroke involving the right middle cerebral artery territory. A total of four rTMS sessions were administered over two consecutive days, with two sessions performed each day, one in the morning and one in the afternoon.

B. rTMS intervention

Each stimulation session lasted approximately 580s, during which 1,000 rTMS pulses were delivered at a frequency of 10 Hz and an intensity corresponding to 60% of the Resting Motor Threshold (RMT). The RMT was determined using the abductor pollicis brevis muscle. The stimulation protocol consisted of 25 trains of stimuli, with 40 pulses per train and inter-train intervals of 20s. Stimulation was delivered over the O2 electrode position corresponding to the lesioned visual cortex, according to the International 10–20 EEG electrode placement system.

The coil was positioned with the handle pointing downward and angled at 45° relative to the sagittal plane, so that the induced electric field remained perpendicular to the underlying cortical gyrus [8]. All sessions were performed using a MagVenture MagPro R30 (Farum, Denmark) equipped with a figure-of-eight coil (Cool-B70).

C. Design and Implementation of the SSVEP Paradigm

The visual field integrity assessment protocol was based on SSVEP. paradigm specifically designed for diagnostic purposes (objective perimetry) was employed. Participants were instructed to maintain visual fixation on a static central target, thereby restricting eye movements and minimizing artifacts generated by saccadic eye movements. Simultaneously, flickering visual stimuli were presented sequentially in four orthogonal peripheral regions at independent stimulation frequencies of 6.0 Hz (Theta), 10.28 Hz (Alpha), 14.4 Hz (Low Beta), and 18.0 Hz (High Beta). Visual stimuli were displayed on a 48 × 27 cm monitor with a resolution of 1280 × 720 pixels. Participants were seated facing the monitor at an approximate viewing distance of 35 cm. This configuration covered 68.4° of the horizontal visual field and 41.6° of the vertical visual field (Fig. 1). Previous studies have demonstrated that peripheral visual stimulation elicits robust cortical responses while significantly reducing discomfort and visual fatigue during prolonged assessments, allowing objective visual field mapping without requiring voluntary eye movements [9]. EEG signals were recorded over the parieto-occipital region using electrodes P7, O1, O2, and P8, EEG signals were acquired using the Emotiv EPOC X system, with a sampling rate of 256 Hz. The signal-processing architecture continuously computed Band Power and Signal-to-Noise Ratio (SNR), comparing these measures with a baseline acquired during the initial 15s calibration period. To improve system sensitivity, the event detection algorithm incorporated a compensatory mathematical correction based on the 1/f power-law spectral distribution (aperiodic noise), an intrinsic characteristic of background electrophysiological activity [10]. Since both spectral power and signal amplitude decrease logarithmically with increasing frequency (a phenomenon further accentuated by the lower density of cones in the peripheral retina) the system adjusted the detection thresholds (SNR thresholds) asymmetrically and inversely proportional to the target stimulation frequency. Consequently, the 18.0 Hz stimulus required a mathematically lower activation criterion than that established for the Theta (6.0 Hz). During the assessment, if the signal failed to exceed the adjusted threshold within the stimulation window, the algorithm recorded a timeout event. The absence of an evoked response provided objective clinical evidence for mapping and topographically identifying potential disruptions or deficits within the contralesional optic radiations associated with the stroke lesion.

Analysis evaluates the impact of rTMS on SSVEP responses in patients with stroke. The analytical focus is the comparison between a conventional SSVEP protocol, in which participants maintain visual fixation on the flickering stimulus, and a peripheral SSVEP protocol, in which participants maintain fixation on the center of the screen while visual stimuli are presented in the peripheral visual field. The peripheral protocol was specifically developed to accommodate the neurophysiological limitations of this patient population.

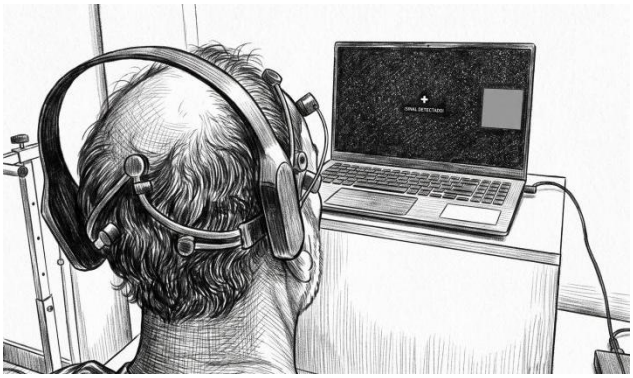


Fig. 1. SSVEP-based EEG signal acquisition system for assessing central and peripheral visual field integrity.

III. Results and Discussion

A. Electromyographic Noise Dispersion Analysis and Mitigation

Use In EEG recordings from patients with neurological impairments, forced visual fixation frequently triggers fatigue and spasticity, contaminating cortical signals with EMG and oculomotor artifacts. To quantify this contamination and assess signal stability, the Coefficient of Variation (CV) was calculated, expressed as the ratio

between the standard deviation (σ) and the mean (μ) of the absolute spectral power:

$$CV = \sigma/\mu \quad (1)$$

The spectral analysis of the baseline session (Pre-TMS) demonstrated that the traditional paradigm is mathematically unsuitable for standardized clinical applications due to its extreme instability. According to the data presented in Table 1, in the Patient 1, for Low Beta band (14.4 Hz), the traditional protocol showed a mean power value of μ = 7.91 associated with an aberrant standard deviation of σ = 64.43, resulting in a CV ≈ 8.14. In biomedical signal processing theory, such a high CV indicates that the recorded activity does not represent a stimulus-evoked response, but rather an anomaly dominated by peripheral noise. In clear contrast, the peripheral protocol demonstrated robustness and ergonomic suitability. For the same patient and frequency band, the baseline recording yielded μ = 0.93 and σ = 0.51, resulting in a CV ≈ 0.54. This substantial reduction in signal dispersion confirms that the peripheral interface minimizes oculomotor strain, providing a cleaner cortical signal suitable for quantifying the actual effects of neuromodulation, which is widely recognized for modulating cortical excitability and promoting neuroplasticity in ischemic tissues [11].

TABLE I. STATISTICAL METRICS OF ABSOLUTE SPECTRAL POWER (PRE- AND POST-TMS) IN BETA BANDS

Band	Protocol	P1/P2	Pre Mean	Pre SD (σ)	Post Mean	Post SD (σ)	p-value	Change (%)
Low Beta	Peripheral vision	P1	0.9320	0.5117	1.1315	0.4855	< 0.001*	+21.41%
		P2	0.3749	0.1612	0.4180	0.1998	0.0241*	+11.49%
	Central vision	P1	7.9132	64.4335	1.1559	0.8882	< 0.001*	−85.39%
		P2	0.3182	0.1282	0.7049	0.4169	< 0.001*	+121.53%
High Beta	Peripheral vision	P1	0.6215	0.2059	1.0530	0.2984	< 0.001*	+69.43%
		P2	0.2578	0.0720	0.2480	0.0892	0.0290*	−3.80%
	Central vision	P1	5.1941	42.8837	0.8935	0.2985	< 0.001*	−82.79%
		P2	0.2899	0.0718	0.4487	0.1537	< 0.001*	+54.77%

[a.] Correspond to specific stimulation frequencies (Low Beta ≈ 14.4 Hz; High Beta ≈ 18.0 Hz). The Wilcoxon signed-rank test was applied to determine the p-value, with a significance level of 5% (α = 0.05).

B. Low-Frequency Spectrum Normalization (Theta Biomarker)

Clinically, a brain affected by ischemia or hemorrhage exhibits a pathological focal increase in Theta spectral activity (4–8 Hz), which is considered a prominent biomarker of tissue damage [12]. The extracted results demonstrated that induced an abrupt and significant reduction in Theta power in Patient 1 (from 5.10 to 2.54, $p < 0.001$). Neurophysiologically, the attenuation of slow-wave activity following neuromodulation reflects the reversal of cortical inhibition and metabolic recovery of perilesional tissue [12]. This phenomenon may have increased the responsiveness of the patient's visual cortex to higher-frequency SSVEP stimuli in subsequent stages.

C. Consistent Facilitation of the Magnocellular Pathway (Low Beta)

The clinical effectiveness of the proposed peripheral vision protocol is illustrated in Fig. 2. At intermediate-to-high stimulation frequencies (Low Beta, 14.4 Hz), a trend toward synergistic facilitation was observed in both patients, in contrast to the unpredictable behavior of the central vision model. Patient 1 exhibited an increase of +21.41%, whereas Patient 2 showed an increase of +11.49% (both $p < 0.05$).

The neuroanatomical explanation for this finding lies in the segregation of the visual pathways. Central vision is predominantly mediated by the parvocellular pathway, which is specialized for high spatial resolution but has limited temporal resolution. In contrast, peripheral vision is predominantly processed by the magnocellular pathway, which exhibits high sensitivity to motion and flickering visual stimuli [13]. The architecture of the proposed system successfully engaged this temporally sensitive pathway, enabling the interface to compensate for the characteristic attenuation imposed by the EEG 1/f law.

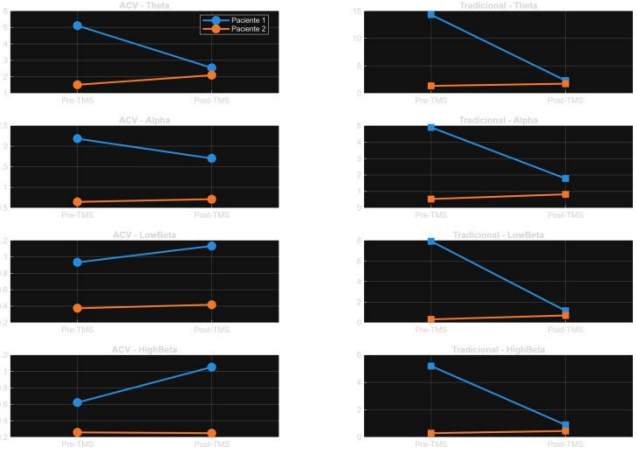


Fig. 2. Individual trajectory dynamics before and after rTMS. A consistent increase in spectral power is observed with the peripheral vision protocol (left), in contrast to the mathematical variability induced by the central vision protocol (right).

D. Neuroplastic Limits and Saturation in the High Beta Band

Despite the overall benefits, the response of the visual cortex in the High Beta band (18.0 Hz) revealed the

neuroplastic limits of each participant, as summarized in Fig. 3. Under the peripheral protocol, Patient 1 exhibited a substantial increase of +69.43% at 18.0 Hz. In contrast, Patient 2 showed a plateau accompanied by a slight decrease of −3.80% ($p = 0.0290$). This contrast provides quantitative evidence that stimulation above 15 Hz may exceed the entrainment capacity of cortical networks with more severe impairment [14].

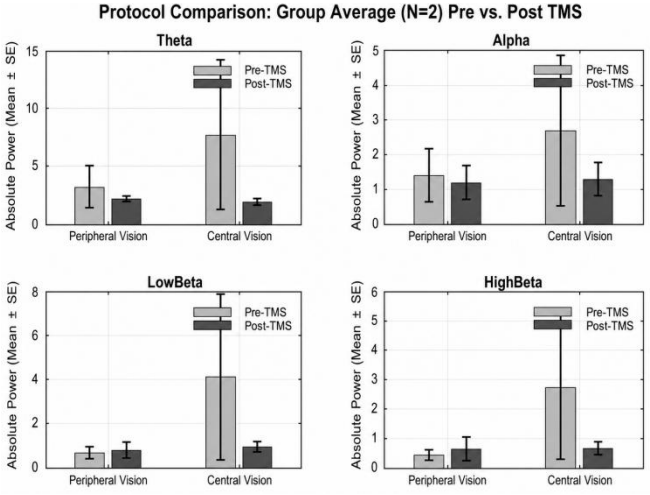


Fig. 3. Group mean power (N = 2) with error bars representing the Standard Error of the Mean. The results demonstrate the reliability of the peripheral vision protocol for pre and post-intervention measurements.

IV. Conclusion

The results of this pilot study demonstrate that the SSVEP-based assessment protocol using peripheral visual stimulation outperformed the conventional central vision paradigm for the electrophysiological evaluation of patients with post-stroke hemianopia. The peripheral protocol produced more stable signals, with lower spectral dispersion and reduced susceptibility to electromyographic and oculomotor artifacts, indicating greater robustness for clinical applications.

The rTMS intervention was accompanied by consistent changes in the analyzed electrophysiological biomarkers. The reduction in Theta-band power suggests a decrease in pathological cortical activity, consistent with the functional recovery of perilesional tissue. Likewise, the increases observed in the Low Beta band indicate improved cortical synchronization to periodic visual stimuli, particularly when using the peripheral vision protocol, demonstrating greater efficiency in engaging the preserved visual pathways.

On the other hand, the heterogeneous response observed in the High Beta band indicates that the effects of neuromodulation depend on individual neurophysiological characteristics. While one participant exhibited a marked increase in spectral power, the other showed a plateau in the response, suggesting that higher stimulation frequencies may reach the synchronization limits of more severely impaired cortical networks.

Overall, these findings indicate that the combination of rTMS and SSVEP-based assessment using peripheral visual stimulation represents a promising approach for the objective monitoring of visual function following stroke. However, due to the small number of participants, the findings should be interpreted as preliminary evidence. Future studies involving larger cohorts, longitudinal follow-up, and different stimulation parameters are required to confirm the reproducibility of these findings and establish the clinical applicability of this protocol for the rehabilitation of post-stroke visual field loss.

The peripheral SSVEP protocol represents a promising methodological contribution to the assessment of occipital visual function. Pre/post-rTMS changes should be interpreted with caution, considering spontaneous variability, fatigue, practice effects, time of day, medication, and electrode contact. The timeout does not constitute objective evidence of damage to the optic radiations.

Furthermore, there is no guarantee that the mathematical increases observed in the peripheral visual field for SSVEP responses following rTMS correspond to actual improvements in visual field function. Confirmation of such functional gains will require evaluation using computerized visual perimetry. It is also noteworthy that Patient 2 reported a burning sensation in the eyes after the rTMS session, as well as pain in the posterior region of the neck. Patient 2 exhibited recovery of upper left limb motor function whose magnitude exceeded the clinical improvement typically expected with conventional physical rehabilitation methods. Although it is not possible to establish a causal relationship based on a single case, this finding strongly suggests that the rTMS intervention may have contributed to the observed motor recovery. These results highlight the need for studies with larger sample sizes and more robust methodological designs to investigate the mechanisms by which occipital cortex stimulation may induce neural plasticity and promote motor recovery following neurological injury.

Acknowledgments

The authors would like to express their sincere gratitude to all co-authors and collaborators whose contributions were essential to the development of this research. Special thanks are extended to CREFES for providing the infrastructure and clinical support necessary for conducting this study. The authors also wish to acknowledge Lucimara Aparecida Pereira for her invaluable assistance, dedication, and support throughout the execution of this research. Finally, the authors are deeply grateful to participants who voluntarily took part in this study.

References

[1] C. Wijesundera, A. J. Vingrys, T. Wijeratne, and S. G. Crewther, "Acquired Visual Deficits Independent of Lesion Site in Acute Stroke," Frontiers in Neurology, vol. 11, Art. no. 705, 2020, doi: 10.3389/fneur.2020.00705.

[2] F. Rowe, D. Brand, C. A. Jackson, A. Price, L. Walker, S. Harrison, C. Eccleston, C. Scott, N. Akerman, C. Dodridge, C. Howard, T. Shipman, U. Sperring, S. MacDiarmid, and C. Freeman, "Visual Impairment Following Stroke: Do Stroke Patients Require Vision Assessment?" Age and Ageing, vol. 38, no. 2, pp. 188–193, Mar. 2009, doi: 10.1093/ageing/afn230.

[3] F. Rowe, D. Brand, C. A. Jackson, A. Price, L. Walker, S. Harrison, C. Eccleston, C. Scott, N. Akerman, C. Dodridge, C. Howard, T. Shipman, U. Sperring, S. MacDiarmid, and C. Freeman, "Visual Impairment Following Stroke: Do Stroke Patients Require Vision Assessment?" Age and Ageing, vol. 38, no. 2, pp. 188–193, Mar. 2009, doi: 10.1093/ageing/afn230.

[4] M. M. Ferreira, J. P. S. Lima, and T. F. Bastos-Filho, "A Review About the Use of SSVEP Associated with tDCS to Improve Visual

Quality of Post-Stroke Patients," in Proceedings of the IX Latin American Congress on Biomedical Engineering (CLAIB 2022) and the XXVIII Brazilian Congress on Biomedical Engineering (CBEB 2022), 2022.

[5] D. L. Schomer and F. H. Lopes da Silva, Eds., Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, 7th ed. New York, NY, USA: Oxford University Press, 2018.

[6] S. Finnigan and M. J. A. M. van Putten, "EEG in Ischaemic Stroke: Quantitative EEG Can Uniquely Inform (Sub-)Acute Prognoses and Clinical Management," Clinical Neurophysiology, vol. 124, no. 1, pp. 10–19, Jan. 2013, doi: 10.1016/j.clinph.2012.07.003.

[7] S. Rossi et al., "Safety and Recommendations for TMS Use in Healthy Subjects and Patient Populations, with Updates on Training, Ethical and Regulatory Issues: Expert Guidelines," Clinical Neurophysiology, vol. 132, no. 1, pp. 269–306, Jan. 2021, doi: 10.1016/j.clinph.2020.10.003.

[8] N. El Nahas, A. M. Elbokl, E. H. Abd Eldayem, T. M. Roushdy, R. M. Amin, S. M. Helmy, A. Z. Akl, A. A. Ashour, S. Samy, A. Amgad, T. H. Emara, M. Nowara, and F. F. Kenawy, "Navigated Perilesional Transcranial Magnetic Stimulation Can Improve Post-Stroke Visual Field Defect: A Double-Blind Sham-Controlled Study," Restorative Neurology and Neuroscience, vol. 39, no. 3, pp. 199–207, 2021, doi: 10.3233/RNN-211181.

[9] X. Zhao, Z. Wang, M. Zhang, and H. Hu, "A Comfortable Steady State Visual Evoked Potential Stimulation Paradigm Using Peripheral Vision," Journal of Neural Engineering, vol. 18, no. 5, Art. no. 056021, 2021, doi: 10.1088/1741-2552/abf397.

[10] T. Donoghue, M. Haller, E. J. Peterson, et al., "Parameterizing Neural Power Spectra into Periodic and Aperiodic Components," Nature Neuroscience, vol. 23, no. 12, pp. 1655–1665, Dec. 2020, doi: 10.1038/s41593-020-00744-x.

[11] F. Fregni and A. Pascual-Leone, "Technology Insight: Noninvasive Brain Stimulation in Neurology—Perspectives on the Therapeutic Potential of rTMS and tDCS," Nature Clinical Practice Neurology, vol. 3, no. 7, pp. 383–393, Jul. 2007, doi: 10.1038/ncpneuro0530.

[12] S. Finnigan and M. J. A. M. van Putten, "EEG in Ischaemic Stroke: Quantitative EEG Can Uniquely Inform (Sub-)Acute Prognoses and Clinical Management," Clinical Neurophysiology, vol. 124, no. 1, pp. 10–19, Jan. 2013, doi: 10.1016/j.clinph.2012.07.003.

[13] A. M. Norcia, L. G. Appelbaum, J. M. Ales, B. R. Cottereau, and B. Rossion, "The Steady-State Visual Evoked Potential in Vision Research: A Review," Journal of Vision, vol. 15, no. 6, Art. no. 4, May 2015, doi: 10.1167/15.6.4.

[14] M. A. Pastor, J. Artieda, J. Arbizu, M. Valencia, and J. C. Masdeu, "Human Cerebral Activation During Steady-State Visual-Evoked Responses," The Journal of Neuroscience, vol. 23, no. 37, pp. 11621–11627, Dec. 2003, doi: 10.1523/JNEUROSCI.23-37-11621.2003.

Desenvolvimento Preliminar de um Ambiente de Realidade Virtual com Luva Háptica para Treinamento Funcional da Mão

Angélica María Rojas-Tocora
Univ. Federal do Espírito Santo
Vitória, Brasil
angelica.tocora@edu.ufes.br

Meilen Salamanca
Univ. Federal do Espírito Santo
Vitória, Brasil
meilen.jimenez@edu.ufes.br

Anselmo Frizera
Univ. Federal do Espírito Santo
Vitória, Brasil
anselmo.frizera-neto@ufes.br

Teodiano Freire Bastos-Filho
Univ. Federal do Espírito Santo
Vitória, Brasil
teodiano.bastos@ufes.br

***Resumo*—A lesão medular cervical pode comprometer a função dos membros superiores e reduzir a independência nas atividades da vida diária. Este estudo apresenta o desenvolvimento preliminar do *Market Rehab*, um ambiente de realidade virtual integrado às luvas hápticas SenseGlove Nova 2 para a realização de tarefas funcionais de alcance, contato, preensão, transporte e liberação de produtos por pessoas com lesão medular cervical. Apresentam-se os resultados de um estudo piloto descritivo com seis voluntários saudáveis para avaliar a viabilidade técnica e a usabilidade do sistema desenvolvido. Neste estudo-piloto, uma execução foi definida como uma tentativa registrada, desde o início até sua conclusão bem-sucedida. Cada participante realizou nove execuções: uma para cada um dos Níveis 1 a 4 e cinco execuções específicas de produto no Nível 5, totalizando 54 execuções entre os seis participantes. Todos os usuários concluíram os cinco níveis de tarefa propostos, apresentando um tempo médio geral de conclusão de 17,07 s, apresentado aqui como um indicador descritivo global. Nas 54 execuções, observou-se uma média de 1,28 tentativa por execução e foram registrados doze erros, dos quais oito ocorreram durante a liberação dos produtos. A pontuação média na *System Usability Scale* foi de 85,8 pontos, o que indica que os usuários avaliaram positivamente a interface, embora tenham relatado distância excessiva de alguns produtos, dificuldade de preensão e percepção do peso da luva. Os resultados confirmam a viabilidade técnica preliminar do sistema, mas indicam a necessidade de otimizar a interação e quantificar o impacto biomecânico da massa distal antes de sua aplicação em pessoas com lesão medular incompleta.**

***Index Terms*—Lesão medular cervical, lesão medular, realidade virtual, luva háptica, SenseGlove, membro superior.**

I. Introdução

A lesão medular (LM) apresenta elevado impacto funcional e social, com estimativas globais indicando aproximadamente 15,4 milhões de pessoas no mundo vivendo com LM em 2021 [1]. Estudos recentes descrevem diferenças relevantes de incidência, etiologia e distribuição demográfica entre regiões [2], [3]. Na América Latina, limitações de acesso, padronização e disponibilidade de serviços especializados ainda afetam a assistência e a reabilitação dessa condição [4].

Nas lesões cervicais, a função dos membros superiores tipicamente reduz significativamente a independência em atividades da vida diária [5]. Portanto, tarefas de alcance, preensão, pinça, transporte e liberação de produtos são componentes funcionais prioritários a serem trabalhados. Embora a lesão comprometa vias motoras e somatossensoriais, processos de plasticidade permanecem relevantes para a recuperação e para o treinamento motor [6].

A reabilitação convencional baseia-se principalmente em exercícios repetitivos, prática intensiva de treinamento funcional e participação ativa [7]. Contudo, sua aplicação pode ser limitada pela redução da motivação do paciente, dificuldade de padronização das tarefas, dependência de observações subjetivas e escassez de medidas objetivas capazes de repre- sentar o desempenho durante atividades funcionais [7], [8].

Nesse contexto, os jogos sérios (JS) surgem como alternativa para complementar processos tradicionais, oferecendo ambientes seguros, controlados e motivadores, alinhados às necessidades terapêuticas [8], além de integrar tecnologias como realidade virtual e feedback háptico, ampliando a imersão e o engajamento [8], [12]. Embora estudos recentes indiquem a viabilidade dessas tecnologias em terapias de LM [13], [18], persistem limitações como elevada heterogeneidade metodológica, amostras reduzidas, escassez de protocolos para membros superiores [4], [6], [9], [10], alem de demanda crescente por intervenções com métricas objetivas de avaliação funcional [15], [16].

As luvas hápticas ampliam essa abordagem ao combinar o rastreamento das mãos com o feedback sensorial [12], havendo apresentado potencial para o treinamento motor de membros superiores em populações neurológicas [12]. Embora esses achados não possam ser diretamente extrapolados para pessoas com LM, os resultados dessas pesquisas indicam que essa integração é promissora. Entre essas tecnologias, as luvas hápticas Nova 2 (SenseGlove, Países Baixos) destacam-se pela compatibilidade com o motor de jogos Unity e com

os óculos Meta Quest 3. Essas integrações permitem recriar tarefas funcionais de alcance, preensão, transporte e liberação de produtos. Para isso, o sistema desenvolvido fornece feedback de resistência passiva em quatro dedos, contato palmar e vibração [19]–[21]. Porém, antes da aplicação clínica, é indispensável validar sua estabilidade, interação, registro de métricas e tolerância ao dispositivo.

Este trabalho tem como objetivo desenvolver e avaliar, por meio de um estudo piloto com voluntários saudáveis, o *Market Rehab*, o qual é um ambiente de Realidade Virtual (RV) criado em Unity, integrado às luvas Nova 2 e aos óculos Meta Quest 3. O estudo não temcomo objetivo demonstrar eficácia terapêutica, mas estabelecer uma base técnica para futura avaliação em pessoas com LM incompleta. Para tal fim, foram analisados: conclusão das tarefas, tempo de execução, número de tentativas, erros de preensão e liberação, bem como a usabilidade do sistema desenvolvido.

II. Metodologia

A. Delineamento do estudo e Características dos Usuários

Neste trabalho foi realizado um estudo piloto descritivo, transversal, não clínico e não terapêutico, destinado à avaliação da viabilidade técnica do sistema proposto. Participaram seis voluntários saudáveis com idades entre 20 e 40 anos, três mulheres e três homens, com idade média de $26{,}33 \pm 5{,}28$ anos, em sessões de 30-40 min.

O critério do inclusão no estudo abrangia voluntários, capazes de compreender as instruções e realizar movimentos voluntários da mão, aceitassem utilizar a luva e o ambiente de RV e assinassem o termo de consentimento livre e esclarecido (TCLE). Foram excluídos voluntários com dor atual no membro superior, lesão recente, alterações cutâneas na área de contato da luva, limitação neurológica ou musculoesquelética relevante, intolerância intensa à RV ou sintomas que impedissem a continuidade da sessão.

B. Sistema e Ambiente de RV

O hardware do sistema é apresentado na Fig. 1, o qual é constituído pelos óculos *Meta Quest 3* (Fig. 1 A) e luvas *SenseGlove Nova 2* (Fig. 1 B) para interação e feedback háptico. O feedback sensorial era aplicado durante as etapas de contato, preensão e liberação dos produtos pelas luvas. Para definir a posição espacial de cada mão, as luvas foram integradas fisicamente aos controles Touch Plus do Meta Quest 3 (Fig. 1 C). O sistema foi configurado para execução em sedestação, visando reduzir deslocamentos excessivos do tronco.

A integração desse hardware foi realizada no motor de jogos Unity 2022.3.62f3, utilizando o Meta XR All-in-One SDK para os óculos e o Nova 2 Unity Plugin v2.6 para as luvas. Nesse mesmo ambiente de RV foi criado o cenário virtual *Market Rehab* (Fig. 2), que reproduz um pequeno mercado contendo plateleiras, um carrinho de compras e produtos alimentícios de diferentes formatos. Na cena, os modelos virtuais das mãos são controlados diretamente pelas luvas hápticas. O sistema executa e registra automaticamente, em

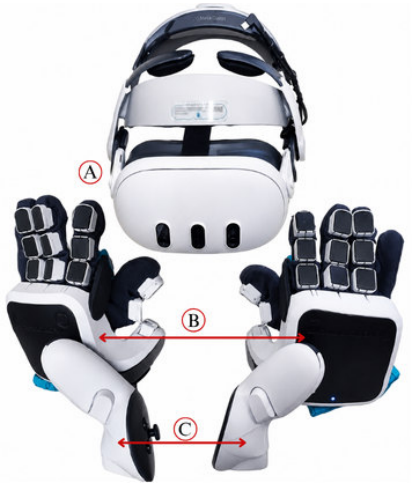


Figura 1. Elementos de hardware do sistema desenvolvido.



Figura 2. Ambiente virtual *Market Rehab* utilizado nas tarefas de manipulação funcional com as luvas hápticas SenseGlove Nova 2.

arquivos CSV, as métricas de desempenho durante a sequência de alcance, preensão, transporte e liberação dos productos.

Dentro do mesmo ambiente de RV foram integrados 5 níveis (Fig.3) baseados em tarefas que foram definidas a partir de movimentos funcionais para a manipulação de produtos, tais como, alcance, contato, preensão, transporte e liberação [5], [9], [10], [12]. Desta forma os componentes de interação associadas a movimentos foram testados e, posteriormente, integrádos em uma atividade contextualizada [13], [14].

Nos quatro primeiros níveis foi utilizada uma execução principal associada ao objetivo do nível. No quinto nível, diferentes produtos foram manipulados individualmente. Por essa razão, as análises por execução identificam o nível e o produto (por exemplo, L04-Leite ou L05-Geleia), enquanto a taxa de conclusão foi calculada para os cinco níveis de tarefa.

C. Procedimento Experimental

Inicialmente, foram registrados os dados dos voluntários do estudo, tais como código anônimo, idade, gênero, mão dominante e experiência prévia com RV. Sintomas de dor, fadiga, tontura e náusea (típicamente inerente ao uso de óculos de RV foram avaliados por meio de uma escala numérica de 0 a 10. Em seguida, com o voluntário em sedestação, as luvas foram colocadas, ativadas e calibradas por meio do *SenseCom*,

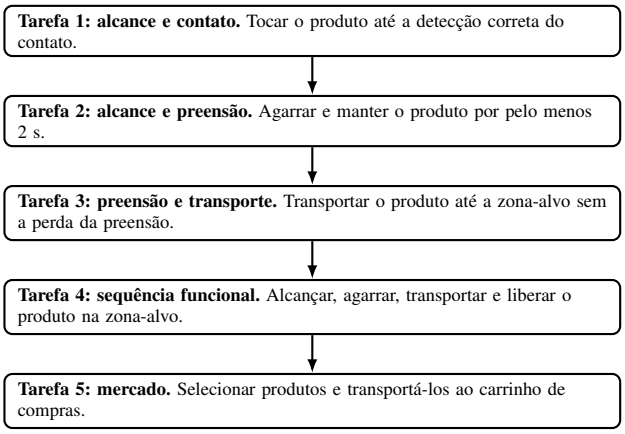


Figura 3. Progressão dos cinco níveis experimentais, do contato inicial à manipulação funcional dos produtos no carrinho de compras.

verificando-se a conexão, o funcionamento do feedback e a coerência entre os movimentos reais e virtuais. Posteriormente, os óculos de RV) foram posicionados e ajustados para garantir o conforto do voluntário.

Antes de cada nível, fori explicado verbalmente a cada voluntário o objetivo e o critério de conclusão do experimento. Não houve etapa livre de familiarização; assim, a primeira execução também foi registrada. Vale resaltar que os usuarios utilizaram a mão de preferência e podiam utilizar ambas as mãos quando necessário.

D. Usabilidade e Análise dos Dados

Após as tarefas, os voluntários responderam ao *System Usability Scale* (SUS), composto por dez itens em escala Likert de cinco pontos [17]. A pontuação foi calculada pelo procedimento convencional e variou de 0 a 100. Para interpretação descritiva da qualidade do sistema desenvolvido, foi adotada a referência de 68 pontos onde uma pontuação entre 68 e 80 é considerada bom; de 80 a 85 é muito bom; e um valor acima de 85 representa um sistema com excelente usabilidade [18]. Também foi disponibilizado um campo aberto para observações gerais do voluntário sobre o sistema.

Devido ao caráter piloto e ao tamanho amostral, foram realizadas análises descritivas considerando: i) tempo por voluntário e execução; ii) erros de preensão e liberação; iv) progressão temporal; v) pontuação SUS.

III. Resultados

A Tabela I sintetiza os principais resultados. O tempo médio de execução do experimento foi de 17,07 s, com intervalo de 9,5–28,7 s, apresentando média de 1,28 tentativas por tarefa e 12 erros ao longo das 54 execuções registradas, ocorridos predominantemente na etapa de liberação.

A. Desempenho nas Tarefas

A progressão temporal de cada voluntário é apresentada na Fig. 4. Os resultados mostram variabilidade interindividual e picos pontuais ao longo do protocolo. O participante P04

Tabela I
Tempo de conclusão, tentativas e erros por nível

Nível	Execuções	Tempo (s)	Mediana (s)	DP (s)	Tentativas	Erros
N01	6	23.68	12.01	34.49	6	0
N02	6	9.00	7.07	3.92	6	0
N03	6	21.54	20.45	15.44	8	1
N04	6	21.99	18.69	15.99	9	3
N05	30	15.49	11.65	10.79	41	8

apresentou um aumento expressivo do tempo para o produto L01-Cereal, enquanto P01 apresentou maiores tempos para L04-Leite e L05-Geleia. Esses picos ocorreram em execuções específicas e não caracterizaram uma redução contínua do desempenho durante toda a sessão.

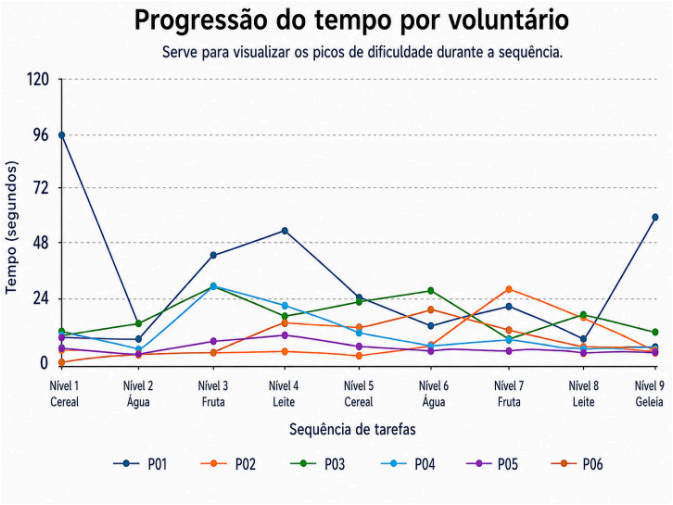


Figura 4. Progressão do tempo de execução ao longo das tarefas para cada voluntário.

A comparação dos tempos médios por voluntário é apresentada na Fig. 5. O voluntário P01 apresentou o maior tempo médio (28,7 s), consistente com o maior número de tentativas e erros registrados, seguido por P04 (22,6 s) e P03 (19,6 s). Em contraste, P05 apresentou o menor tempo médio (9,5 s), seguido por P02 (10,2 s) e P06 (11,8 s).

A distribuição dos erros por execução é apresentada na Tabela II. A maior concentração ocorreu para L05-Geleia, com quatro erros, seguida por L04-Leite, com três ocorrências. Também foram registrados dois erros em L05-Cereal e L05-Fruta, e um erro para L03-Fruta. Não foram observados erros para L01-Cereal, L02-Água, L05-Água e L05-Leite.

A distribuição dos erros entre os voluntários (Fig. 6a) indica que se apresentaram, no conjunto da amostra, quatro erros de preensão e oito de liberação, evidenciando maior ocorrência de dificuldades durante a liberação dos produtos virtuais. Nesse sentido, P01 apresentou oito erros e concentrou a maior parcela das falhas de liberação.

B. Usabilidade

A pontuação média do *System Usability Scale* (SUS) foi de 85,8 pontos, o que representa uma usabilidade excelente



Figura 5. Tempo médio de execução para cada voluntário.

Tabela II
DISTRIBUIÇÃO DOS ERROS POR EXECUÇÃO

Execução	Erros
L05-Geleia	4
L04-Leite	3
L05-Cereal	2
L05-Fruta	2
L03-Fruta	1
L01-Cereal	0
L02-Água	0
L05-Água	0
L05-Leite	0

segundo os critérios de avaliação. Todos os voluntários atribuíram pontuações superiores ao valor de referência de 68 pontos, como mostrado na Fig. 6b, sendo que P03 e P06 alcançaram 90 pontos, enquanto P01 apresentou a menor pontuação, com 77,5 pontos.

Nos comentários abertos, os voluntários relataram uma boa aceitação da interface e do ambiente de RV. Entretanto, também indicaram que alguns produtos estavam posicionados a uma distância excessiva ou eram difíceis de agarrar. Ademais, a luva foi percebida como pesada durante a execução das tarefas.

IV. DISCUSSÃO

A. Desempenho e interação no ambiente de RV

Os resultados demonstram a viabilidade técnica preliminar do *Market Rehab* para a execução e o registro de tarefas funcionais em ambiente de RV. Todos os seis voluntários concluíram os cinco níveis propostos, e o sistema registrou de forma estruturada o tempo de execução, o número de tentativas e os erros ocorridos durante a interação. O tempo médio entre os voluntários foi de 17,07 s, com valores médios individuais variando entre 9,5 s e 28,7 s, enquanto foram registradas, em média, 1,28 tentativas por execução. Esses resultados evidenciam variabilidade entre os usuários e mostram que a conclusão das tarefas, isoladamente, não descreve integralmente a interação com o sistema.

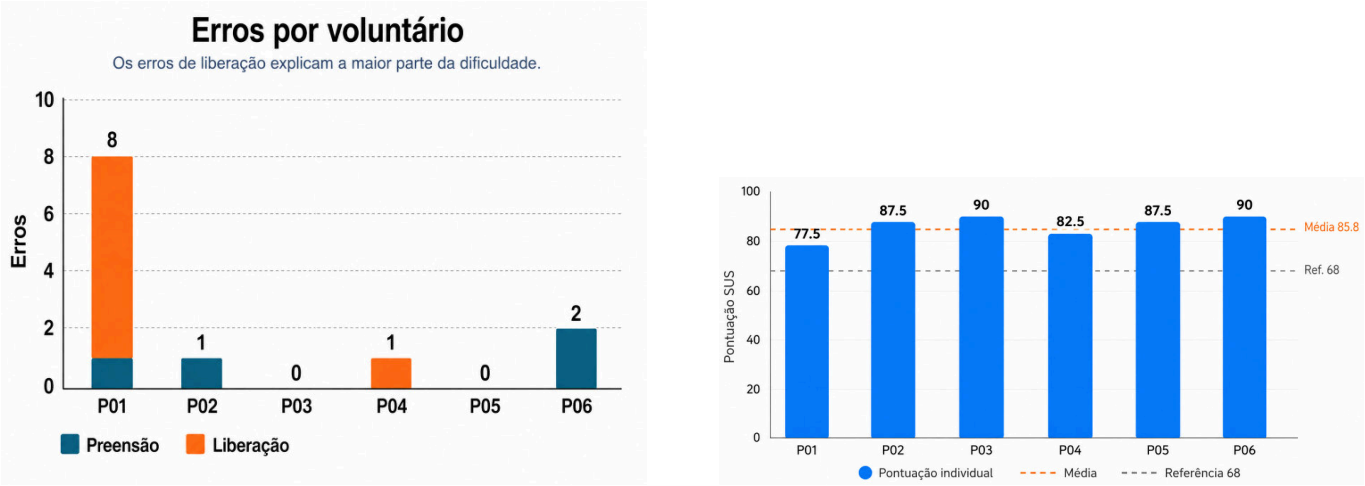
Essa diferença pode ser observada, por exemplo, em P01, que concluiu todo o protocolo, mas apresentou o maior tempo médio, maior número de tentativas e oito erros. Em contraste, outros voluntários completaram as mesmas tarefas com menores tempos e menor ocorrência de falhas. Dessa forma, tempo, tentativas e erros fornecem informações complementares para caracterizar o desempenho e permitem identificar dificuldades que não seriam observadas considerando apenas a conclusão dos níveis.

Foram registrados 12 erros totais durante os protocolos, sendo quatro relacionados à preensão e oito à liberação. A maior concentração ocorreu em L05-Geleia, com quatro erros, seguida por L04-Leite, com três ocorrências. Essa distribuição, associada aos relatos de dificuldade para agarrar determinados produtos, sugere que características específicas da interação com cada objeto podem estar relacionadas às dificuldades observadas. No ambiente de RV, essa interação depende da geometria dos produtos, da configuração dos *colliders*, dos limiares utilizados para detectar contato, preensão e liberação e do acoplamento entre a luva e os movimentos dos dedos. Portanto, os erros registrados não devem ser atribuídos exclusivamente ao desempenho motor dos voluntários.

A estrutura progressiva dos cinco níveis permitiu avaliar componentes da interação em diferentes graus de complexidade. Os primeiros níveis envolveram ações mais específicas de contato, preensão e transporte, enquanto os níveis posteriores integraram essas ações em sequências funcionais mais completas. Em especial, o quinto nível contextualizou a atividade em um cenário de mercado, exigindo seleção, transporte e liberação de diferentes produtos no carrinho de compras. Ao mesmo tempo, essa contextualização introduziu maior variabilidade de distância, orientação e geometria dos objetos, fatores que podem contribuir para as diferenças observadas entre as execuções.

Os resultados de L04-Leite e L05-Geleia indicam, portanto, a necessidade de revisar os parâmetros de interação utilizados para esses produtos. Diferentes geometrias podem exigir configurações específicas de *colliders*, zonas de contato e critérios de preensão e liberação. Em versões futuras, além da ocorrência do erro, recomenda-se registrar o estado que originou a falha, como perda de contato, abertura insuficiente da mão, liberação fora da zona-alvo ou reativação involuntária da preensão. Esse detalhamento poderá auxiliar na distinção entre dificuldades relacionadas à configuração do ambiente e aquelas associadas à execução da tarefa pelo usuário.

Os voluntários também relataram distância excessiva de alguns produtos, indicando que a organização espacial inicialmente adotada não foi percebida de forma uniforme por todos os usuários. Esse resultado reforça a necessidade de revisar o posicionamento dos objetos e considerar uma área de alcance individualmente calibrada em versões futuras do sistema. Aspectos como massa adicionada, geometria dos objetos, intensidade do feedback háptico e restrições mecânicas de dispositivos vestíveis também podem interferir na execução



(a) Erros de preensão e liberação por voluntário.

(b) Pontuação individual no SUS.

Figura 6. Distribuição dos erros por voluntário e avaliação da usabilidade do sistema.

de tarefas de destreza e na facilidade percebida de movimento [20].

Essa adaptação torna-se particularmente relevante antes de uma futura avaliação em pessoas com LM, considerando que limitações da função proximal podem comprometer o alcance e o posicionamento da mão durante tarefas funcionais [19]. Entretanto, como o presente estudo incluiu exclusivamente voluntários saudáveis, os resultados obtidos não permitem determinar o impacto dessas variáveis nessa população. Assim, a parametrização individual do alcance deve ser considerada uma etapa futura de desenvolvimento e avaliação, e não uma evidência derivada diretamente deste estudo.

B. Massa Distal e Demanda Biomecânica

A percepção de peso foi um dos principais comentários qualitativos por parte dos voluntários. As luvas Nova 2 apresentam massa nominal aproximada de 350 g por luva [14]. Considerando essa massa, o dispositivo adiciona aproximadamente 3,43 N de carga gravitacional ao membro superior. Embora esse valor, isoladamente, não permita determinar a sobrecarga imposta ao usuário, a localização distal da massa pode aumentar a demanda necessária para sustentar e deslocar a mão durante tarefas de alcance e manipulação. Esse efeito depende da postura adotada, da posição do membro e da capacidade muscular disponível de cada indivíduo.

No presente estudo não foram mensurados momentos articulares, atividade muscular ou fadiga; portanto, não é possível quantificar diretamente o impacto biomecânico da luva. Entretanto, a percepção de peso relatada pelos participantes indica que a massa do dispositivo deve ser considerada em futuras avaliações, particularmente antes de sua utilização em pessoas com LM incompleta.

A comparação com dispositivos destinados especificamente à lesão medular reforça essa preocupação. O módulo manual da RELab tenoexo possui aproximadamente 120 g, enquanto componentes mais pesados são deslocados para o tronco ou para a cadeira de rodas [19]. A massa distal nominal da Nova 2 é cerca de 2,9 vezes maior. Essa comparação não estabelece um limite clínico, uma vez que os dispositivos apresentam objetivos e princípios de funcionamento distintos, mas evidencia a redução e a redistribuição da massa distal como estratégias relevantes de projeto. Revisões da literatura sobre exoesqueletos de mão também identificam o peso, a ergonomia e a distribuição dos componentes como requisitos importantes para dispositivos de reabilitação e assistência [23].

A influência relativa da massa da luva pode tornar-se mais relevante em tarefas que envolvem a manipulação de produtos leves. Nessas condições, parte importante da demanda mecânica necessária para sustentar e deslocar o membro superior pode estar associada ao próprio dispositivo, e não exclusivamente à manipulação do objeto ou à execução da tarefa funcional. Esse aspecto reforça a necessidade de avaliar a massa distal separadamente do objetivo funcional proposto no ambiente virtual.

Além da carga gravitacional, a Nova 2 utiliza freios magnéticos passivos para gerar feedback de força em quatro dedos, com resistência nominal máxima de até 24 N [14]. Essa resistência possui natureza distinta da carga gravitacional associada à massa da luva, pois depende do contato com o objeto e da configuração dos dedos. Além disso, o mecanismo é resistivo e não assistivo. Em usuários com redução da força de flexão ou extensão dos dedos, níveis elevados de resistência podem aumentar o tempo de execução ou limitar a amplitude voluntária do movimento. Portanto, massa distal e feedback resistivo devem ser investigados como fatores biomecânicos distintos.

A relevância clínica da carga adicionada pelo dispositivo depende de sua relação com a capacidade articular residual de cada usuário. O mesmo esforço necessário para sustentar a Nova 2 pode representar uma pequena parcela da capacidade de um adulto saudável e uma parcela consideravelmente maior da capacidade disponível em uma pessoa com fraqueza proximal. Assim, a questão biomecânica central não é apenas

determinar se 350 g representam uma massa elevada, mas avaliar quanto da capacidade muscular disponível é utilizada para sustentar e deslocar o dispositivo durante as tarefas.

A repetição das tarefas também deve ser considerada. Em uma elevação vertical de 0,30 m, por exemplo, a massa nominal da luva corresponde a aproximadamente 1,03 J de trabalho mecânico adicional por elevação. Em protocolos com repetições sucessivas, essa demanda ocorre repetidamente e se associa às contrações isométricas necessárias para estabilizar o ombro e o cotovelo. Nesse tipo de tarefas repetitivas, é necessário avaliar a fadiga e o esforço do membro superior do usuário, os quais normalmente são baseados em eletromiografia (EMG), cinemática, percepção de esforço e medidas de desempenho [22]. Dessa forma, a conclusão dos níveis observada neste estudo piloto não permite inferir tolerância a sessões terapêuticas repetitivas.

Além do trabalho associado ao deslocamento da luva, a demanda fisiológica inclui contrações isométricas para sustentar o membro, estabilização escapular, cocontração e correções posturais. Assim, mesmo quando a mão permanece aparentemente estática no ambiente virtual, a atividade muscular pode continuar e contribuir para a fadiga. Esse aspecto é particularmente relevante em protocolos de reabilitação, nos quais a repetição constitui parte central da intervenção.

A conclusão de uma tarefa também não garante que o movimento tenha sido executado com a estratégia motora desejada. Usuários com função proximal limitada podem aumentar a inclinação ou a rotação do tronco, elevar a cintura escapular, ampliar a abdução do ombro ou apoiar o cotovelo para alcançar um objeto. Essas estratégias podem permitir que o sistema registre a conclusão da tarefa, mas modificam a demanda articular e podem reduzir a especificidade terapêutica da atividade. Por essa razão, a futura avaliação clínica deve considerar não apenas o tempo e a conclusão das tarefas, mas também a qualidade do movimento e a presença de compensações.

Para o presente sistema, uma abordagem semelhante permitiria verificar se o aumento do tempo ou do número de tentativas ocorre por dificuldade de interação, adaptação ao ambiente ou redução progressiva da capacidade de sustentar o dispositivo. Também permitiria comparar o início e o final de um bloco de repetições, identificando alterações que uma análise baseada apenas na média global poderia ocultar.

A transferência do sistema desenvolvido para pessoas com LM incompleta deve ainda considerar a grande heterogeneidade da função residual. Não existe um valor universal, em quilogramas, que defina a carga máxima tolerável para essa população. O nível neurológico, a preservação motora, a assimetria entre os membros, a espasticidade, a dor e a fadiga podem alterar substancialmente a capacidade de sustentar a mão no espaço.

Nesse contexto, a redução de 24% a 53% observada por Lee et al. [21] refere-se à atividade muscular máxima medida por EMG e não pode ser convertida diretamente em uma perda equivalente de torque articular, já que a amplitude do sinal EMG não apresenta uma relação linear universal com a força produzida. Ainda assim, esses resultados demonstram que a redução da capacidade muscular pode ser heterogênea entre músculos do ombro, cotovelo e punho, justificando uma caracterização individual antes da exposição da população-alvo a uma carga distal constante.

Assim, a luva Nova 2 não deve ser classificada de forma absoluta como “leve” ou “pesada” para todas as pessoas com LM. A interpretação mais adequada é que sua massa distal representa uma barreira biomecânica potencial. Essa barreira pode apresentar menor impacto em usuários com função proximal preservada e tornar-se relevante em indivíduos cuja capacidade residual já é amplamente utilizada para sustentar o próprio membro. Portanto, uma futura avaliação deverá determinar quanto da capacidade muscular e articular disponível é exigida para sustentar e deslocar o dispositivo durante as posturas e tarefas reais do *Market Rehab*.

C. Usabilidade, Limitações e Trabalhos Futuros

A elevada pontuação obtida no SUS indica uma percepção global positiva da usabilidade do sistema, particularmente em relação à compreensão e à interação com a interface. As variações observadas nas métricas de desempenho não contradizem essa avaliação, mas fornecem informações complementares sobre a execução das tarefas e as diferenças individuais entre os voluntários. Nesse sentido, a análise conjunta do SUS com medidas como tempo de execução, número de tentativas e erros permitiu uma caracterização mais abrangente da experiência de uso. Enquanto o SUS refletiu a aceitação e a usabilidade percebida do sistema, as métricas objetivas contribuíram para identificar aspectos específicos do desempenho durante a manipulação dos produtos virtuais.

A excelente pontuação SUS de 85,8 e os comentários positivos sobre a interface indicam boa aceitação preliminar. Contudo, usabilidade elevada não elimina limitações biomecânicas. O SUS avalia a percepção global do sistema, enquanto tempo, tentativas, erros e comentários abertos identificam dificuldades específicas. Além disso, os comentários sobre distância, preensão e peso foram exploratórios e não foram obtidos por um instrumento qualitativo estruturado.

A principal limitação deste estudo é a inclusão de apenas seis voluntários sem comprometimento neurológico. Os resultados não demonstram eficácia clínica nem permitem assumir que pessoas com LM incompleta tolerarão a carga distal. A próxima etapa deve seguir uma sequência objetiva:

1) Medir a massa total do conjunto utilizado, incluindo rastreamento, e estimar seu centro de massa;
2) Quantificar momentos externos nas posturas reais das tarefas;
3) Comparar condições sem luva, com luva e com suporte gravitacional do antebraço;
4) Medir a amplitude do sinal de EMG, tempo de sustentação, esforço percebido e fadiga, métricas frequentemente combinadas em avaliações biomecânicas do membro superior [22];

5) Após reposicionar produtos e revisar *colliders* e limiares de preensão/liberação, realizar um estudo piloto estratificado com pessoas com LM incompleta.

No estudo piloto clínico, a força de ombro, cotovelo, punho e dedos deve ser caracterizada previamente. O objetivo central deverá ser determinar qual fração da capacidade muscular disponível é consumida apenas para sustentar a luva e posicionar a mão. Caso a massa se confirme como limitante, devem ser avaliados suporte gravitacional, intervalos de descanso e ajuste individual do feedback resistivo.

A avaliação do movimento também deverá incluir compensações proximais. Um voluntário pode concluir a tarefa por meio de inclinação ou rotação do tronco, elevação escapular, abdução excessiva do ombro ou apoio do cotovelo. Nessa situação, a métrica binária de sucesso permanece positiva, embora a estratégia motora seja diferente daquela pretendida para a terapia. Por esta razão, a futura fase clínica deve registrar a cinemática do tronco e do membro superior ou, minimamente, utilizar uma classificação observacional padronizada das compensações.

Outro passo necessário é definir critérios de progressão e interrupção da terapia. A duração das tarefas e o número de repetições devem ser ajustados conforme esforço percebido, redução sustentada do desempenho, dor ou incapacidade de manter a postura de trabalho. Esse procedimento permitirá diferenciar uma dificuldade de aprendizagem da interface de uma limitação do usuário produzida por fadiga ou carga mecânica. Somente após essa caracterização será possível discutir o uso do sistema na população-alvo.

V. Conclusões

O *Market Rehab* apresentou viabilidade técnica preliminar para a execução e o registro de tarefas de alcance, preensão, transporte e liberação em realidade virtual. Nos seis voluntários saudáveis avaliados, os cinco níveis foram concluídos e o sistema registrou tempo de execução, tentativas e erros. A predominância de erros de liberação, a variabilidade dos tempos e os comentários relacionados à distância dos objetos, à preensão e ao peso da luva identificaram aspectos específicos que devem ser refinados. A pontuação média de 85,8 no SUS indica uma percepção preliminar favorável da usabilidade, devendo ser interpretada considerando o tamanho reduzido e as características da amostra.

Aspectos específicos da interação, tais como a distância excessiva de alguns produtos, a dificuldade de preensão e o peso percebido da luva, definem as prioridades imediatas de aprimoramento do sistema. Em especial, a massa distal deve ser tratada como uma limitação biomecânica potencial antes da transferência do sistema para pessoas com LM incompleta. A continuidade do estudo requer a caracterização mecânica do conjunto, a quantificação da demanda muscular e da fadiga, a comparação com suporte gravitacional e, somente após esses ajustes, a realização de um estudo piloto com a população-alvo. A viabilidade clínica do sistema dependerá da capacidade de executar as tarefas de forma repetitiva, tolerável e sem compensações motoras excessivas, e não apenas de concluí-las.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES), processos 2026-48WNC e 2026-FQVLV, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processos 308155/2023-8 e 402080/2025-4, pelo apoio ao desenvolvimento deste trabalho.

Referências

[1] World Health Organization, "Spinal cord injury," Apr. 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/spinal-cord-injury

[2] Y. Lu et al., "Global incidence and characteristics of spinal cord injury since 2000-2021: A systematic review and meta-analysis," *BMC Med.*, vol. 22, Art. no. 285, 2024, doi: 10.1186/s12916-024-03514-9.

[3] M. Kim, W. Jeong, S. Jang, J. H. Park, Y. Bae, and S. W. Lee, "Spinal cord injury epidemiology and causes: A worldwide analysis with 2050 projections," *Healthcare*, vol. 13, no. 20, Art. no. 2552, 2025, doi: 10.3390/healthcare13202552.

[4] N. Marchesini, L. L. Fernández Londoño, D. Griswold, and A. M. Rubiano, "Early stages management of traumatic spinal cord injury in Latin America: A scoping review," *World Neurosurg.*, vol. 162, pp. 138-149.e29, 2022, doi: 10.1016/j.wneu.2022.03.021.

[5] L. Lili, K. S. Sunnerhagen, T. Rekand, and M. Alt Murphy, "Independence and upper extremity functioning after spinal cord injury: A cross-sectional study," *Sci. Rep.*, vol. 13, Art. no. 3148, 2023, doi: 10.1038/s41598-023-29986-y.

[6] A. Calderone, D. Cardile, R. De Luca, A. Quartarone, F. Corallo, and R. S. Calabrò, "Brain plasticity in patients with spinal cord injuries: A systematic review," *Int. J. Mol. Sci.*, vol. 25, no. 4, Art. no. 2224, 2024, doi: 10.3390/ijms25042224.

[7] J. Chu, J. V. Glinsky, H. Liu, et al., "How much and what type of exercises and training were provided to people with spinal cord injury as part of usual physiotherapy and occupational therapy in the SCI-MT Trial?," *Spinal Cord*, vol. 64, no. 4, pp. 317-323, 2026, doi: 10.1038/s41393-026-01180-7.

[8] M.-C. Juan, J. Elexpuru, P. Dias, B. Sousa Santos, and P. Amorim, "Immersive virtual reality for upper limb rehabilitation: Comparing hand and controller interaction," *Virtual Reality*, vol. 27, no. 2, pp. 1157-1171, 2023, doi: 10.1007/s10055-022-00722-7.

[9] V. B. Raju, B. D. Schmit, and A. Vedantam, "Effectiveness of virtual reality rehabilitation for cervical spinal cord injury: A systematic review and meta-analysis," *Proc. Inst. Mech. Eng. H*, vol. 239, no. 5, pp. 413-422, 2025, doi: 10.1177/09544119251337381.

[10] M. Scalise, T. S. Bora, C. Zancanella, A. Safa, R. Stefini, and D. Cannizzaro, "Virtual reality as a therapeutic tool in spinal cord injury rehabilitation: A comprehensive evaluation and systematic review," *J. Clin. Med.*, vol. 13, no. 18, Art. no. 5429, 2024, doi: 10.3390/jcm13185429.

[11] L. Wang, H. Zhang, H. Ai, and Y. Liu, "Effects of virtual reality rehabilitation after spinal cord injury: A systematic review and meta-analysis," *J. Neuroeng. Rehabil.*, vol. 21, Art. no. 191, 2024, doi: 10.1186/s12984-024-01492-w.

[12] D. Fernández-Vázquez, R. Cano-de-la-Cuerda, and V. Navarro-López, "Haptic glove systems in combination with semi-immersive virtual reality for upper extremity motor rehabilitation after stroke: A systematic review and meta-analysis," *Int. J. Environ. Res. Public Health*, vol. 19, no. 16, Art. no. 10378, 2022, doi: 10.3390/ijerph191610378.

[13] Nova 2, "Unity overview," Nova 2 Docs. [Online]. Available: https://Nova2.gitlab.io/Nova2Docs/unity/overview.html

[14] Nova 2, "Nova 2 Nova 2 Spec Sheet," ver. 1.0, Jun. 2024. [Online]. Available: https://www.Nova2.com/wp-content/uploads/2022/04/2021-03-17_Nova2_specsheet_Nova.pdf

[15] C. Moritz et al., "Non-invasive spinal cord electrical stimulation for arm and hand function in chronic lesão medular: A safety and efficacy trial," *Nat. Med.*, vol. 30, no. 5, pp. 1276-1283, 2024, doi: 10.1038/s41591-024-02940-9.

[16] M. M. Kilkki, J. Poutanen, K. Kauranen, J. Arokoski, and S. Hiekkala, "Effects of technology-assisted rehabilitation after spinal cord injury: Pilot randomized controlled crossover trial," *JMIR Rehabil. Assist. Technol.*, vol. 12, Art. no. e78091, 2025, doi: 10.2196/78091.

[17] J. Brooke, "SUS: A 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189-194.

[18] J. Sauro, *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. Denver, CO, USA: Measuring Usability LLC, 2011.

[19] J. Dittli et al., "Mixed methods usability evaluation of an assistive wearable robotic hand orthosis for people with spinal cord injury," *J. Neuroeng. Rehabil.*, vol. 20, Art. no. 162, 2023, doi: 10.1186/s12984-023-01284-8.

[20] I. Priadythama, W. L. Yeoh, P. Y. Loh, and S. Muraki, "The effect of the degree of freedom and weight of the hand exoskeleton on joint mobility function," *Robotics*, vol. 11, no. 2, Art. no. 53, 2022, doi: 10.3390/robotics11020053.

[21] K.-S. Lee, M. Zaman, and J. Hwang, "Reduced muscle activity of the upper extremity in individuals with spinal cord injuries," *Int. J. Environ. Res. Public Health*, vol. 19, no. 8, Art. no. 4708, 2022, doi: 10.3390/ijerph19084708.

[22] C. Brambilla, M. Lavit Nicora, F. Storm, G. Reni, M. Malosio, and A. Scano, "Biomechanical assessments of the upper limb for determining fatigue, strain and effort from the laboratory to the industrial working place: A systematic review," *Bioengineering*, vol. 10, no. 4, Art. no. 445, 2023, doi: 10.3390/bioengineering10040445.

[23] T. du Plessis, K. Djouani, and C. Oosthuizen, "A review of active hand exoskeletons for rehabilitation and assistance," *Robotics*, vol. 10, no. 1, Art. no. 40, 2021, doi: 10.3390/robotics10010040.

Low-Cost Wearable System for Gait Monitoring in an Aquatic Environment

Caroline de Cássia Batista de Souza
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-3607-1009

Amanda Maria Conceição Perez
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0003-4149-8960

Ana V. M. Inocêncio Lopes
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-1411-7547

Miguel Velhote Correia
Faculty of Engineering
University of Porto
Porto, Portugal
mcorreia@fe.up.pt

Ana Paula de Lima Ferreira
Physical Therapy Department
Federal University of Pernambuco
Recife, Brazil
ana.lferreira@ufpe.br

Marco A. B. Rodrigues
Electronics and Systems Department
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-5143-3696

Abstract*— The biomechanical analysis of gait in an aquatic environment presents challenges related to data acquisition and the use of instrumented systems under immersion conditions. This study aimed to evaluate the applicability of a low-cost wearable system for the acquisition and processing of inertial signals during aquatic gait. The system consisted of an ESP32-S3 microcontroller, two MPU-9250 inertial sensors, and a web platform for data acquisition via Wi-Fi. Twenty healthy individuals performed a standardized pool-walking protocol, and the signals were automatically processed to identify steps and extract spatiotemporal parameters. The system enabled continuous acquisition of acceleration signals, automatically identifying algorithm-detected gait events and making it possible to obtain the total number of steps, traversal time, stride length, and cadence. The results demonstrate the applicability of the platform for gait monitoring in an aquatic environment, reinforcing its potential as a low-cost alternative for applications in research, rehabilitation, and Biomedical Engineering.

Keywords*— Wearable sensors; Accelerometry; Hydrotherapy; Aquatic gait.

I. Introduction

Human movement analysis in an aquatic environment has received increasing attention in recent decades, especially in the context of rehabilitation and therapeutic exercise. The physical properties of water, such as buoyancy, viscosity, and hydrodynamic resistance, promote significant changes in movement execution, reducing gravitational load and modifying neuromuscular and biomechanical patterns when compared with the terrestrial environment (TEDESCHI, 2024; ATKINS et al., 2024).

Among the locomotor tasks performed in this context, aquatic gait stands out for its clinical and functional applicability, and it is widely used in rehabilitation and physical conditioning programs. However, despite its relevance, the biomechanical characterization of gait in an aquatic environment still presents important limitations, mainly due to the difficulties associated with instrumentation and data acquisition in this medium (AMIN, 2023).

In recent years, wearable systems based on inertial sensors have become promising tools for human movement analysis, enabling data collection in different contexts with greater portability and less dependence on specialized laboratories. However, the application of these systems in an aquatic environment remains limited because of technical challenges related to device waterproofing, attachment stability, data transmission, and the possible interference of water with signal quality (Marinho et al., 2019; Monoli et al., 2023).

In this context, the development of wearable systems capable of operating reliably in an aquatic environment represents a relevant gap from both a technological and clinical perspective. Although commercial solutions exist for movement analysis in terrestrial settings, their applicability in water is still limited, in addition to their high cost, which reinforces the need for accessible alternatives adapted to this scenario (MacDermid et al., 2017).

As an alternative, a low-cost wearable system for the acquisition and processing of inertial signals was previously developed. This same system was recently validated for the biomechanical analysis of deep-water running, showing good agreement with a commercial accelerometry system used as the reference method (Souza et al., 2026).

However, its applicability during gait in an aquatic environment had not yet been investigated. Considering that aquatic gait presents biomechanical characteristics distinct from those of deep-water running, it is necessary to verify whether the system maintains its potential for application in this new condition. Thus, the objective of this study was to evaluate the applicability of the same low-cost wearable system for the acquisition and processing of inertial signals during gait in an aquatic environment.

II. Materials and Methods

A. Measurement System

The developed system consists of a compact and portable main module powered by an internal rechargeable battery, eliminating the need for a direct connection to the electrical power supply during data acquisition.

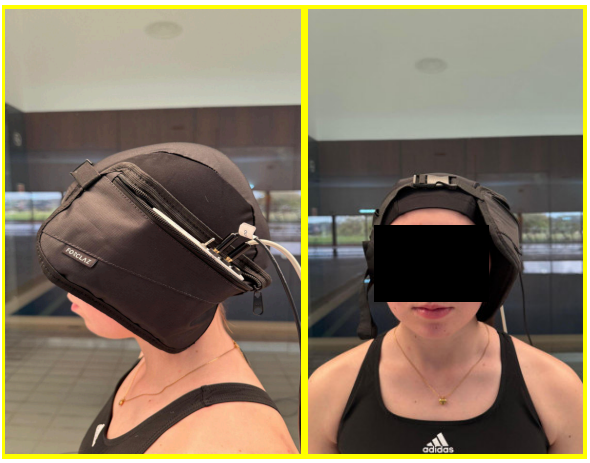
XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The device integrates two inertial sensors connected to the ESP32-S3 microcontroller through an I²C serial communication interface. The hardware architecture is based on the ESP32-S3 Super Mini development board (Espressif Systems, China), featuring a 32-bit dual-core Xtensa® LX7 processor operating at up to 240 MHz, integrated 2.4 GHz Wi-Fi and Bluetooth® Low Energy (BLE) connectivity, 4 MB of flash memory, and 512 KB of SRAM. Both inertial sensors were sampled by the same microcontroller during each acquisition cycle, ensuring temporal synchronization between the signals acquired from the right and left feet.

The sensing unit is based on the MPU-9250 module, which integrates a triaxial accelerometer and a triaxial gyroscope. The module features 16-bit analog-to-digital converters, enabling the simultaneous acquisition of signals along the X, Y, and Z axes with high resolution and low power consumption.

The embedded system, comprising the microcontroller, battery, and other electronic components, was housed in a custom 3D-printed enclosure. During data collection, the enclosure was placed inside a waterproof pouch and secured to the participant's head, keeping the main module above the water surface while enabling continuous signal acquisition and wireless data transmission throughout the experimental protocol (Figure 1).

Fig. 1. Positioning of the main module of the wearable system inside a waterproof pouch secured to the participant's head during data collection.



The inertial sensors were connected to the main module through cables and remained submerged throughout the experimental protocol. The cable outlets were properly sealed to ensure system integrity and reliable data acquisition in the aquatic environment.

Signal acquisition was performed in real time at a sampling frequency of 50 Hz and an acceleration range of ±2 g. The collected data were wirelessly transmitted in real time via Wi-Fi to a dedicated web-based platform developed as an integral component of the system, where they were stored for subsequent processing and analysis (Perez, 2025; Silva, 2025).

B. Sample

The sample consisted of 20 healthy volunteers (Table 1). Inclusion criteria comprised individuals without functional limitations or clinical conditions that could interfere with walking in an aquatic environment.

Table 1. Demographic and anthropometric characteristics of the study participants.

Variable	Value
Participants	20
Sex	10 males / 10 females
Age (years)	20.2 ± 0.8
Body mass (kg)	59.8 ± 7.4
Height (m)	1.68 ± 0.08

All participants were informed about the objectives and procedures of the study and provided written informed consent prior to data collection. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Faculty of Sport of the University of Porto (CEFADE 27_2025), Portugal, as well as by the Research Ethics Committee of the Federal University of Pernambuco (CAAE: 92738325.0.0000.5208).

C. Familiarization and experimental procedures

Upon arrival, all participants received detailed information regarding the objectives and procedures of the study. They subsequently provided written informed consent and completed an individual demographic questionnaire.

Initially, the participants were instrumented. The inertial sensors were bilaterally attached approximately 2 cm above the right and left lateral malleoli, ensuring consistent anatomical placement and reliable data acquisition. Both inertial sensors were positioned with the same anatomical orientation to ensure consistent alignment of the X-, Y-, and Z-axes across participants. Subsequently, a familiarization session was conducted, during which the participants walked along the experimental pathway at a self-selected comfortable speed, allowing them to adapt to the aquatic environment, the experimental protocol, and the instrumentation.

Data collection was conducted in the swimming pool of the Escola Superior de Saúde Norte da Cruz Vermelha Portuguesa, in Oliveira de Azeméis, Portugal, during the first author's international doctoral mobility program carried out in collaboration with the University of Porto. The pool had an approximate depth of 1.30 m, with the water immersion level standardized to the xiphoid process. The experimental protocol consisted of an 8-m linear walking path, and participants were instructed to walk continuously from one end of the pool to the other. Only the forward

walking segment was included in the analysis; turning movements at the end of the walkway were excluded from signal processing to avoid their influence on gait-event detection and spatiotemporal parameter estimation. No additional manual exclusion of walking segments was performed during data processing. Throughout the trials, the inertial sensors remained in the previously described positions, while the data acquisition system was housed inside a waterproof pouch secured to the upper body near the head, ensuring equipment integrity and maintaining signal quality in the aquatic environment (Figure 2).

Fig. 2. Instrumented participant during data collection in the aquatic environment.



D. Data Processing

The accelerometry data stored on the web-based platform were subsequently processed and analyzed. Initially, the signals were organized according to the acquisition axes (X, Y, and Z), enabling the analysis of movement behavior during walking in the aquatic environment.

Spatiotemporal gait parameters, including the number of steps, cadence, stride length, and walking time, were extracted from the accelerometry signals. These parameters were initially calculated individually for each participant and subsequently analyzed using descriptive statistics, including measures of central tendency and dispersion.

Initially, the accelerometry data were filtered using a moving average filter followed by a low-pass Butterworth filter with a cutoff frequency of 20 Hz. After filtering, gait steps were identified using an adaptive threshold peak detection algorithm. Peaks occurring within less than 0.5 s of each other were classified as false positives, and only one peak was retained.

The beginning and end of the walking trial were identified based on the signal variance calculated using a 50-sample sliding window. The first and last abrupt changes in the variance signal were defined as the start and end of the walking trial, respectively.

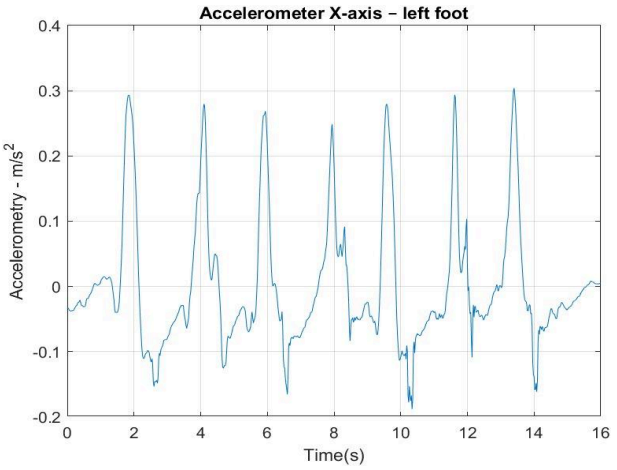
This processing pipeline enabled the automatic determination of the time required to complete the 8-m walking task, as well as the number of steps performed by each foot.

III. RESULTS

A. Signal Acquisition

During data acquisition, periodic and repetitive acceleration signals characteristic of aquatic walking were obtained. An example of these signals is presented in Figure 3.

Fig. 3. Representative X-axis acceleration signal acquired by the accelerometer attached to the left foot during aquatic gait.



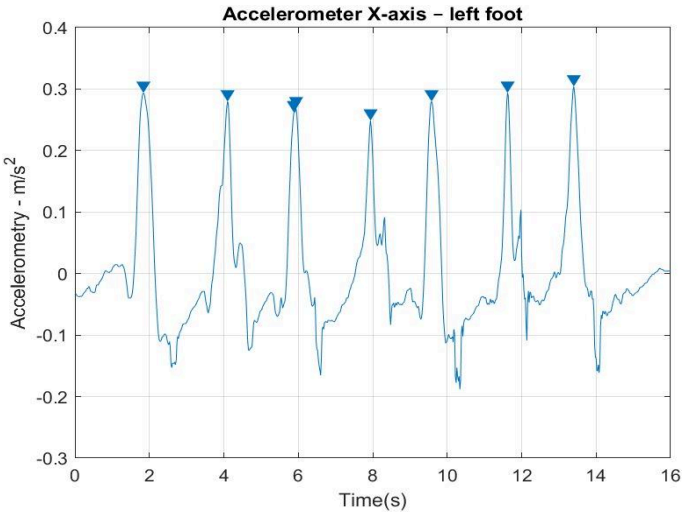
The periodic characteristics of the acceleration signals enabled the reliable identification of gait steps using the proposed peak detection algorithm. The detected peaks were consistently associated with the gait cycle throughout the walking trial, allowing automatic step counting for both the right and left feet. These results demonstrate the effectiveness of the proposed processing method for extracting spatiotemporal gait parameters from the acquired accelerometry data.

B. Step Detection

In the next stage of the processing pipeline, acceleration signal peaks were automatically identified using the proposed algorithm, enabling the automatic detection of gait events. A representative example of this procedure is shown in Figure 4.

The algorithm-detected gait events were subsequently used to calculate spatiotemporal gait parameters, including the total number of steps, walking time, stride length, and cadence. Figure 4 illustrates the acceleration signal recorded from the participant's left foot during the walking trial, with the algorithm-detected steps indicated by triangles at the corresponding signal peaks.

Fig. 4. Automatic detection of gait events based on the peaks of the X-axis acceleration signal acquired by the accelerometer attached to one of the feet during aquatic gait.



C. *Spatiotemporal Gait Parameters*

Figure 5 shows the superimposed X-axis accelerometry data for both feet. It can be observed that the patient begins and ends the walk with the right foot, performing a half-step at both the start and the end. The figure also reveals that the acceleration peaks for the left foot are slightly higher than those for the right foot; this difference may be due to slight variations in accelerometer placement or the fact that the left leg is the dominant one. However, the difference is minor and does not affect the data analysis.

Fig. 5. Superimposed X-axis acceleration signals recorded from the right and left feet during aquatic gait.

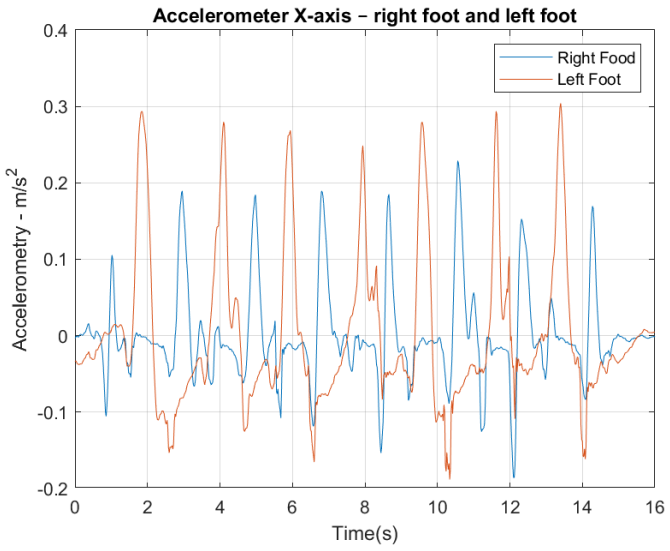


Table 2. Spatiotemporal gait parameters obtained during aquatic gait.

Variable	Mean ± SD
Total number of steps	13.65 ± 3.08
Walking time (s)	16.06 ± 4.73
Stride length (m)	1.19 ± 0.21
Cadence (steps/min)	53.17 ± 12.77

Table 2 presents the spatiotemporal gait parameters automatically extracted from the accelerometry signals acquired by the wearable system. On average, the participants performed 13.65 ± 3.08 steps while completing the 8-m walking trial in 16.06 ± 4.73 s. The mean stride length was 1.19 ± 0.21 m, and the mean cadence was 53.17 ± 12.77 steps/min.

IV. DISCUSSION

The present study extends the application of a low-cost wearable system previously validated for the biomechanical analysis of deep-water running, in which good agreement was observed with a commercial accelerometry system used as the reference method (Souza et al., 2026). Whereas the previous study aimed to establish the concurrent validity of the proposed technological platform, the present work investigated its applicability to aquatic gait, a locomotor task characterized by distinct biomechanical features and widely used in rehabilitation programs. Therefore, the present findings represent a further stage in the development of the system, expanding its potential application to different biomechanical assessment scenarios.

Although the wearable platform was previously validated against a commercial accelerometry system for deep-water running (Souza et al., 2026), the objective of the present study was not to establish the accuracy of gait-event detection during aquatic gait using a concurrent reference system. Instead, this study focused on demonstrating the applicability of the proposed platform for continuous signal acquisition and automated extraction of algorithm-detected gait events and spatiotemporal gait parameters in an aquatic environment. Therefore, the reported gait events should be interpreted as outputs generated by the proposed algorithm rather than as a direct validation of event-detection accuracy.

The use of inertial sensors for monitoring human movement has increased substantially in recent years due to their ability to provide biomechanical measurements using portable, low-cost devices with the potential for application outside specialized laboratory settings. However, the use of these systems in aquatic environments remains relatively limited because of the challenges imposed by water

immersion, including waterproofing of electronic components, sensor fixation stability, data transmission, and appropriate signal processing (Fantozzi et al., 2015; Mangia et al., 2017; Monoli et al., 2023).

The findings of the present study are consistent with those reported by Fantozzi et al. (2015), who demonstrated the feasibility of using inertial measurement units (IMUs) for the kinematic analysis of aquatic gait. Similarly, Mangia et al. (2017) validated inertial sensors for three-dimensional motion analysis in aquatic environments, whereas Lee and Han (2022) showed that low-cost devices, such as smartphones, can also be used for biomechanical assessment. Collectively, these studies, together with the findings of the present work, reinforce the growing trend toward the development of increasingly portable, affordable, and clinically applicable solutions.

Within this context, the system used in the present study is distinguished by integrating inertial sensors, an ESP32-S3 microcontroller, wireless Wi-Fi communication, a rechargeable battery, a 3D-printed enclosure, and a web-based platform for signal acquisition and processing into a single technological platform. This architecture, based on widely available and low-cost components, reduces the dependence on expensive commercial equipment, thereby facilitating the dissemination of biomechanical instrumentation in clinical, academic, and research settings.

In addition to the electronic architecture, the use of wearable sensors in aquatic environments requires overcoming specific challenges imposed by water immersion. Monoli et al. (2023) highlighted that issues related to device waterproofing, sensor placement, data transmission, and the lack of methodological standardization remain important barriers to the widespread adoption of these technologies. In the present study, these challenges were addressed during the system development by keeping only the inertial sensors submerged, while the main unit remained above the water surface inside a waterproof enclosure. This configuration preserved communication between the device and the computational platform without compromising signal acquisition during aquatic gait.

Another relevant aspect concerns signal processing. Monoli et al. (2023) reported considerable heterogeneity among studies regarding sampling frequency, sensor placement, algorithms used for gait event detection, and data processing procedures. This methodological variability hinders direct comparisons between different systems and highlights the need for greater standardization to improve reproducibility and support the consolidation of these technologies. In this regard, the detailed description of the system architecture, sensor placement, and processing pipeline provided in the present study contributes to enhancing the reproducibility of the proposed methodology.

From a biomedical engineering perspective, the main contribution of the present study extends beyond the extraction of spatiotemporal gait parameters. Rather, it lies in the advancement of a previously validated technological platform that has now been successfully applied to a different locomotor task in an aquatic environment. The possibility of using the same hardware architecture across different exercise modalities expands its potential applications and demonstrates the versatility of the proposed system for future adaptations to a variety of functional assessment and aquatic rehabilitation protocols.

The present implementation of the proposed processing algorithm focused on extracting the primary spatiotemporal gait parameters required to demonstrate the applicability of the wearable system during aquatic gait. Additional parameters, such as stride time, step time, symmetry indices, and gait variability, may be incorporated in future developments of the platform, further expanding its applicability for biomechanical and clinical assessments.

Some limitations of the present study should be acknowledged. Although the system was previously validated for the biomechanical analysis of deep-water running using a commercial accelerometry system as the reference method (Souza et al., 2026), the present study focused specifically on its application to aquatic gait. Furthermore, the evaluation was conducted exclusively in apparently healthy individuals, limiting the generalizability of the findings to clinical populations. Future studies should investigate the performance of the proposed system under different walking speeds, water immersion depths, therapeutic protocols, and populations with neuromuscular or musculoskeletal disorders, thereby expanding its clinical applicability.

In summary, the findings demonstrate that the proposed platform represents a promising alternative for biomechanical instrumentation in aquatic environments. The use of low-cost electronic components combined with a portable, modular, and scalable architecture highlights the potential of this technology to expand access to quantitative human movement analysis in clinical, research, and educational settings. Moreover, it contributes to the development of affordable technological solutions for aquatic rehabilitation.

V Conflict of Interest

The authors declare that they have no conflict of interest.

VI. CONCLUSIONS

The developed wearable system was able to continuously acquire accelerometry signals during aquatic gait, enabling the automatic identification of gait events and the extraction of the main spatiotemporal gait parameters. Its architecture, based on low-cost components combined with wireless data transmission and automated signal processing, highlights the potential of the proposed platform for applications in research, biomechanical assessment, and aquatic rehabilitation. In aquatic physical therapy, this technology has the potential to support future objective gait assessment, monitoring of patients' functional progress throughout rehabilitation, and quantification of the effects of different therapeutic protocols, thereby contributing to

evidence-based clinical practice supported by instrumented measurements. Future studies should further validate the proposed system in different populations and clinical protocols, contributing to the consolidation of accessible wearable technologies for human movement analysis in aquatic environments.

References

[1] A. B. Amin, E. Asabre, A. Sahay, S. Razaghi, and Y. Noh, "Feasibility Testing of Wearable Device for Musculoskeletal Monitoring during Aquatic Therapy and Rehabilitation," in Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Sydney, Australia, 2023, pp. 1–4.

[2] A. Atkins, S. Hoppe-Ludwig, W. Guminiak Jr., and L. Miller, "Development and usability of an adjustable prosthesis for aquatic therapy," J. Prosthet. Orthot., vol. 00, no. 00, pp. 1–7, 2024.

[3] S. Fantozzi, A. Giovanardi, C. Borra, and A. Gatta, "Gait Kinematic Analysis in Water Using Wearable Inertial Magnetic Sensors," PLoS ONE, vol. 10, no. 9, Art. no. e0138105, Sept. 2015.

[4] D. H. Lee and S. Han, "Smartphone-Based Measurement of Lower Extremity Joint Angles During Aquatic Exercise," Healthcare, vol. 10, no. 10, Art. no. 1935, 2022.

[5] P. W. MacDermid, P. W. Fink, and S. R. Stannard, "Shock Attenuation, Spatio-Temporal and Physiological Parameter Comparisons Between Land Treadmill and Water Treadmill Running," J. Sport Health Sci., vol. 6, pp. 482–488, 2017.

[6] A. L. Mangia, M. Scaccabarozzi, M. Beomonte Zobel, and M. Galli, "Use of Inertial Sensors for Underwater Motion Analysis: Validation of a New Approach for Three-Dimensional Kinematics," Sensors, vol. 17, no. 10, Art. no. 2251, Oct. 2017.

[7] D. Marinho, H. Neiva, and J. Morais, "The Use of Wearable Sensors in Human Movement Analysis in Non-Swimming Aquatic Activities: A Systematic Review," Int. J. Environ. Res. Public Health, vol. 16, Art. no. 5067, 2019.

[8] C. Monoli, J. A. Tuhtan, L. Piccinini, and M. Galli, "Wearable Technologies for Monitoring Aquatic Exercises: A Systematic Review," Clin. Rehabil., vol. 37, no. 6, pp. 791–807, 2023.

[9] C. Monoli, M. Galli, and J. A. Tuhtan, "Improving the Reliability of Underwater Gait Analysis Using Wearable Pressure and Inertial Sensors," PLoS ONE, vol. 19, no. 3, Art. no. e0300100, 2024.

[10] C. C. B. Souza, F. Parolini, M. F. Goethel, J. Robalino, G. R. de Siqueira, A. L. P. C. Silva, M. V. B. Rodrigues, J. P. Vilas-Boas, M. Velhote Correia, M. A. B. Rodrigues, and A. P. Ferreira, "Validation of a Low-Cost Accelerometry Device for Cycle-Based Biomechanical Analysis of Deep-Water Running," Appl. Sci., vol. 16, no. 13, Art. no. 6404, 2026.

[11] R. Tedeschi and D. Donati, "Harnessing the Power of Water: A Scoping Review of Hydrokinesiotherapy as a Game Changer in the Treatment of Knee Osteoarthritis," J. Clin. Med., vol. 13, Art. no. 5811, 2024.

[12] A. M. C. Perez, "Dispositivo de monitoramento remoto com acelerometria na reabilitação funcional de idosos," Ph.D. dissertation, Centro de Biociências, Universidade Federal de Pernambuco, Recife, Brazil, 2025.

[13] M. B. C. Silva, "Plataforma para telereabilitação motora: auxílio no tratamento de pacientes Pós-COVID-19," Ph.D. dissertation, Centro de Tecnologia e Geociências, Universidade Federal de Pernambuco, Recife, Brazil, 2025.

A Mobile Application for the Six-Minute Walk Test with Oximetry Acquisition via Bluetooth Low Energy: A Pilot Study

1st Athos Brandielli Schuck
Biomedical Eng. R&D Service (SEEB)
Hospital de Clínicas de Porto Alegre (HCPA)
Porto Alegre, Brazil
abschuck@hcpa.edu.br

2nd Gyan Theo Brandielli Schuck
Biomedical Eng. R&D Service (SEEB)
HCPA
Porto Alegre, Brazil
gtschuck@hcpa.edu.br

3rd Danilo C. Berton
Pulmonary Physiology Service
HCPA/UFRGS
Porto Alegre, Brazil
dberton@hcpa.edu.br

4th Luana F. Correa
Pulmonary Physiology Service
HCPA/UFRGS
Porto Alegre, Brazil
luanacorrea@hcpa.edu.br

5th Danton Pereira da Silva Junior
Biomedical Eng. R&D Service (SEEB)
HCPA
Porto Alegre, Brazil
dpsjunior@hcpa.edu.br

6th Bruno Rodriguez Tondin
Biomedical Eng. R&D Service (SEEB)
HCPA
Porto Alegre, Brazil
btondin@hcpa.edu.br

7th André Frotta Müller
Biomedical Eng. R&D Service (SEEB)
HCPA
Porto Alegre, Brazil
afmuller@hcpa.edu.br

8th Alessandro Nakoneczny Schildt
Biomedical Eng. R&D Service (SEEB)
HCPA
Porto Alegre, Brazil
aschildt@hcpa.edu.br

9th Marcelo Basso Gazzana
Pulmonary Physiology Service
HCPA
Porto Alegre, Brazil
mbgazzana@hcpa.edu.br

10th Paulo Roberto Stefani Sanches
Biomedical Eng. R&D Service (SEEB)
HCPA
Porto Alegre, Brazil
psanches@hcpa.edu.br

***Abstract*—The Six-Minute Walk Test (6MWT) is a standardized protocol for assessing cardiorespiratory functional capacity, but its conventional execution relies on the manual recording of physiological variables, which limits the temporal granularity of the data and hinders the traceability of results. This work presents *SpO2Sense*, a cross-platform mobile application that automates the conduction of the 6MWT by integrating the continuous acquisition of peripheral oxygen saturation (SpO_2) and heart rate (HR) from Bluetooth Low Energy (BLE) pulse oximeters, the structured collection of anamnesis and the Borg Scale, and the estimation of the distance walked from the periodic variation of the received signal strength indicator (RSSI). The application was evaluated at the Pulmonary Physiology Unit of the *Hospital de Clínicas de Porto Alegre* in 34 recordings, using the telemetry system in use at the unit as a reference to adjudicate signal quality. Motion artefact during walking, identified by an explicit rule in which an abrupt shift in the test device is not reproduced by the reference, affected 7 of the 28 adjudicable recordings (25%); loss of the BLE link affected 3 further recordings. Within the contaminated segments the two devices diverged by 91 bpm in HR and 43 percentage points in SpO_2 at the 95th percentile of the paired difference. During rest, by contrast, the median per-patient bias was $+0.19$ bpm for HR and $+0.06$ percentage points for SpO_2, and no artefact originated in the resting phase, indicating that walking itself, rather than the sensor or the link, is the dominant source of failure. For the number of traverses during the walk, no significant difference was found between the automatic RSSI-based estimate and manual counting (Wilcoxon signed-rank, p = 0.140), and a *post-hoc* equivalence analysis against a margin of one traverse (25 m) was compatible with average equivalence (p = 0.00014), although the margin was not pre-specified and average equivalence did not extend to individual agreement. The mobile platform proved technically viable for instrumenting the 6MWT, but this class of low-cost oximeter does not withstand the walking phase, which is precisely the condition the test is designed to impose; mechanical coupling at the measurement site, rather than optical performance, is the limitation that must be addressed first.**

***Index Terms*—six-minute walk test, pulse oximetry, motion artefact, Bluetooth Low Energy, mobile application, Flutter**

I. Introduction

The Six-Minute Walk Test (6MWT) is a reproducible, practical, and low-cost instrument for cardiorespiratory functional assessment, and it is a central component in the follow-up of patients with chronic pulmonary and cardiovascular diseases [1]. Its protocol consists of measuring the greatest distance a patient is able to walk, at their own pace, over six minutes

along a flat corridor, while physiological variables are recorded before, during, and after the effort.

In usual clinical practice, the 6MWT is conducted with the manual recording of peripheral oxygen saturation (SpO_2), heart rate (HR), perceived exertion, and distance walked. For patients with respiratory diseases, this exertion is commonly measured using the Borg scale for pain and dyspnea. This process has intrinsic limitations: the sampling of vital variables is sparse, transient desaturation events may go unnoticed, and manual transcription is susceptible to errors and hinders the subsequent auditing of the data. The growing availability of low-cost pulse oximeters equipped with Bluetooth Low Energy (BLE) opens up the possibility of instrumenting the test with continuous acquisition of SpO_2 and HR data using only a common mobile device [2], [3].

This work describes the design and implementation of a mobile application, named *SpO2Sense*, that automates the conduction of the 6MWT. The application and a low-cost BLE pulse oximeter were tested during routine 6MWT sessions at the Pulmonary Physiology Unit of the *Hospital de Clínicas de Porto Alegre*, with the telemetry system already in use at the unit serving as a reference to adjudicate signal quality. The main contributions are: (i) the integration of continuous SpO_2 and HR acquisition via BLE with the protocol's phase logic; (ii) an algorithm for estimating the distance walked from the periodic variation of the RSSI connection strength signal; and (iii) a characterisation of the dominant failure mode of this class of device during the 6MWT, based on an explicit adjudication rule in which an abrupt shift in the test device is accepted as artefact only when the reference does not reproduce it. The remainder of the article presents related work, the system architecture, the BLE communication layer, the artefact adjudication procedure, and a discussion of limitations and future work.

II. Related Work

The validation of the *wireless* oximetry sensors used in mobile solutions is a critical step to ensure the reliability of the clinical data collected. In this regard, the study conducted by Marshall et al. [4] evaluated the accuracy of commercial and medical-grade portable oximeters during the execution of the Six-Minute Walk Test (6MWT), comparing them with the market gold standard, the Masimo Rad-5. Among the analyzed devices, the Nonin WristOx2 3150 stands out, having demonstrated a performance in the detection of peripheral oxygen saturation (SpO_2) highly comparable to that of the Masimo device, with a mean difference rate within the 3% limit. This work shows that, even when state-of-the-art devices are compared, a difference was observed.

Reported 6MWT instrumentation solutions tend to rely on dedicated hardware or on semi-automatic recording. The adoption of BLE in personal health devices has enabled tele-health applications running on smartphones, lowering the cost barrier for continuous health monitoring while also offering a more user-friendly experience [5]. The development of mobile systems focused primarily on the kinematic computation of the 6MWT for estimating the distance walked was explored in the literature by the study of Salvi et al. [6]. The authors proposed an architecture composed of two applications: a web interface aimed at the physiologist and a mobile application (developed in the Apache Cordova framework) directed at the patient, integrated with a back-end server and wireless pulse oximeters. To estimate the distance walked, they implemented two algorithms based on a smartphone's internal sensors: one using the Global Positioning System (GPS) for outdoor tests, and another based on inertial sensors for indoor tests, whose purpose is to identify the change of direction (turnaround) of the patient in a fixed-distance corridor. In terms of accuracy, the maximum variation reported between the distance estimated by the app and the actual distance for indoor environments was 23 meters, a borderline value, considering that 50 meters is the interval regarded as clinically significant for detecting changes in the status of pulmonary diseases during the 6MWT [7]. Furthermore, the study's validation was conducted predominantly by researchers in a controlled laboratory environment. The system proposed in the present work differs from that approach by combining, in a single mobile application, the continuous acquisition of oximetry, the orchestration of the protocol phases, and the estimation of distance without additional sensors. This provides an algorithmic simplification by reusing the oximeter's own radiofrequency link.

In our study, the Wellue pulse oximeter is assessed against the telemetry system used by the Pulmonary Physiology Unit of the *Hospital de Clínicas de Porto Alegre* (HCPA), which is based on the Nonin PalmSat 2500 model developed in Silva Junior's work [8] and is used here to adjudicate the quality of the signal acquired by the test device.

III. Methodology

A. System Architecture

The application was developed in Flutter, with an interface based on Material 3, enabling cross-platform execution from a single codebase [9]. The project's organization follows a layered pattern inspired by Model-View-ViewModel (MVVM): pure data models, observable state providers (*view-model*), screens and interface components (*view*), and a service layer responsible for accessing the Bluetooth hardware.

Complementary to the MVVM, the system architecture can be described in four functional layers (Figure 1). The first, the sensing layer, is responsible for the acquisition of the physiological data of peripheral oxygen saturation (SpO_2) and HR, expressed in beats per minute (bpm). The second, the communication layer, uses the Bluetooth Low Energy (BLE) protocol to establish the connection between the application and the previously registered oximeters, enabling the reception of SpO_2 and bpm data in real time, with a sampling frequency of 1 Hz. The third layer, the application layer, concentrates the interfaces and functionalities aimed at the health professional, allowing patient registration, the execution and monitoring of the six-minute walk tests (6MWT), and the analysis of the obtained results. Finally, the fourth layer, the database layer,

is responsible for the secure cloud storage of the data collected during the tests, which is linked to the user account created within the application itself. This layered division seeks to modularize the system, making the application of the 6MWT more agile and practical for the health professional.

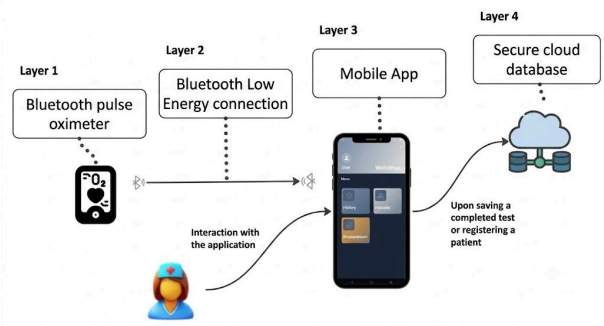


Fig. 1. Mobile app main architecture.

B. Authentication and Persistence

Access to the application is controlled by email-and-password-based authentication, with mandatory email verification. A reactive authentication *gate*, based on a stream of session-state events, decides the initial screen without imperative navigation: unauthenticated users are directed to the login screen, users with an unverified email to the verification screen, and verified users to the main menu. This pattern ensures that email confirmation performed on another device automatically redirects the user.

The data is persisted in a document-oriented cloud database and protected with appropriate authentication-based access rules, using the Firebase system [10]. Patients and exams are stored in a hierarchy in which each professional maintains their own base, without global sharing of records, as illustrated below:

```
users/{uid}/patients/{id}/walkTest/{walkId}
```

Each exam document stores the time series of SpO_2, HR, and radiofrequency signal, the interruption periods, the distance parameters, the anamnesis data, the Borg Scale measurements, and the date and time of each test.

C. Data Acquisition via Bluetooth Low Energy

Bluetooth access is achieved by registering the addresses of the services transmitted by the compatible devices, from which the bytes are collected, a communication enabled by the *FlutterBluePlus* library. Communication with the oximeter is encapsulated in a dedicated service, instantiated once per measurement session, and follows the following flow of events for the interface: discovery of the connected devices transmitting signals, scanning of these signals, and decoding of the oximetry and received signal strength data. Discovery filters the scan results by platform name or by advertised service identifier, identifying these devices by their physical address. The connection adopts a *retry* policy with an increasing linear wait, emitting a disconnection event to the interface in the event of a definitive failure. Once connected, the service starts the periodic monitoring of the signal strength and the service discovery, enabling notifications on the characteristic corresponding to the hardware model. The BLE *payload* is model-specific to each oximeter, with fields positioned at fixed offsets.

D. Distance Estimation from RSSI Variation

Since the 6MWT is performed in a fixed corridor, the distance walked is estimated from the periodic variations of the received signal strength indicator (RSSI), which oscillates as the patient moves away from and closer to the reference point where the device remains stationary. The algorithm smooths the signal by convolution with a uniform window, detects peaks and valleys through derivative analysis, filters extrema by minimum prominence and minimum distance between events, and counts the transitions between extrema as traverses (half of a lap), as illustrated in Figure 2. The interruption periods are removed from the RSSI signal time series before the analysis so as not to distort the count.

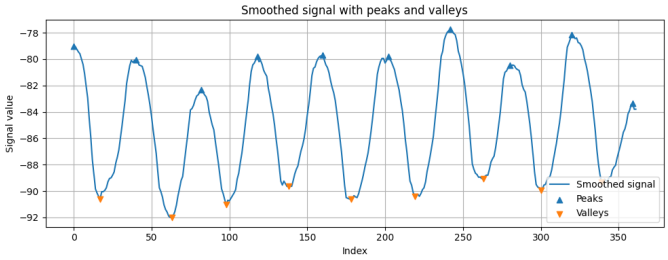


Fig. 2. Figure depicting the `analyse_signal` function used to estimate the number of traverses made by the patient during the 6MWT. Each peak–valley/valley–peak pair is considered half of a lap, which is used to estimate the distance.

E. Test Execution With a Pulse Oximeter

To further evaluate the usage of our mobile app, we conducted tests at the Pulmonary Physiology Unit of the *Hospital de Clínicas de Porto Alegre* (HCPA) with 34 patients who visited the unit to undergo the 6MWT, using the Wellue FS20F fingertip pulse oximeter, which holds a valid registration with the Brazilian Health Regulatory Agency (ANVISA). The study protocol was approved by the Institutional Research Ethics Committee of the Hospital de Clínicas de Porto Alegre and a written informed consent was obtained from all participants before inclusion in the study. The walking test was conducted in a straight indoor corridor 25 meters long, with a complete lap being 50 meters. Although it differs from the 30 meters standard set by the American Thoracic Society (ATS), the 25 meters distance was adopted because it is the measurement currently used at the Pulmonary Physiology Unit. The health professional responsible for the measurement stood in a reference point near the start of the lap, to serve as reference for distance estimation, as seen in Figure 3.

Each recording comprises two consecutive phases. The walking phase corresponds to the first 360 s of acquisition, during which the patient walks the corridor at their own pace. The resting phase follows immediately and lasts up to three minutes, although it frequently ends earlier, when the patient's vital signs return to baseline. Both devices sample at 1 Hz, so the two phases are delimited by sample index rather than by wall-clock time, and all per-phase statistics are normalised by the observed duration of the phase rather than by its nominal length.

During each test session, data on SpO_2, HR, distance covered, and anamnesis, which the application also records, were collected. The entire data collection process was carried out alongside the use of a telemetry system based on Nonin PalmSat2500 developed by Silva Junior et al. [8], currently in use at the Pulmonary Physiology Unit. Both devices were positioned on the same hand of the patient, and the temporal series data collected were synchronized between devices for comparison based on the timestamps. The HCPA system serves as an adjudicator of signal quality rather than as a metrological standard.

F. Cohort and Adjudication of Motion Artefact

Of the 34 recordings collected, 3 were affected by loss of signal in the reference system and 3 by loss of the BLE link, the latter costing 38.8%, 35.7% and 18.3% of the walking-phase samples. Adjudication requires both series to be complete, so the 28 remaining recordings were analysed. Loss of link and motion artefact are treated as distinct mechanisms.

Motion artefact was identified by an explicit rule applied to all 28 recordings, rather than by visual inspection of those suspected of failure. A candidate onset is flagged when the test device shifts by at least 15 bpm in HR, or 4 percentage points in SpO_2, within 3 s. The candidate is retained as artefact only if the reference, over the same window extended by 5 s, does not reproduce the change, its excursion remaining below half the magnitude recorded by the test device; otherwise the event is attributed to a physiological origin. The contaminated segment then runs from the onset until the absolute paired difference falls below 5 bpm, or 2 percentage points, for five consecutive seconds. Segment-level extension is necessary because the artefact presents as a single transition initiating a sustained plateau of divergence, not as a succession of abrupt transitions. A recording is classified as affected when at least one such segment is detected.

Since the traverse-count variable did not satisfy the assumption of normality (Shapiro–Wilk test, detailed in Results), the non-parametric Wilcoxon signed-rank test was adopted throughout this work to compare the reference and test methods. The traverse-count comparison does not depend on the reference oximetry and was therefore performed on the 31 recordings without substantial BLE signal loss, which distorts the RSSI curve directly. It was first assessed whether there was any difference between manual annotation and the estimation through RSSI signal variation. This test was followed by the two one-sided tests (TOST) procedure, implemented with one-sided Wilcoxon signed-rank tests against each extreme of a margin of $\Delta = \pm 1$ traverse (± 25 m). This margin was adopted because it falls below the 30–50 m range considered clinically meaningful in the 6MWT [7]. Finally, a 90% bootstrap confidence interval for the mean paired difference was computed to estimate where this difference lies. As the traverse margin was defined after data collection, the analysis is reported as exploratory. Agreement during the resting phase was summarised as the median across recordings of the mean paired difference within each one, reported with its interquartile range.

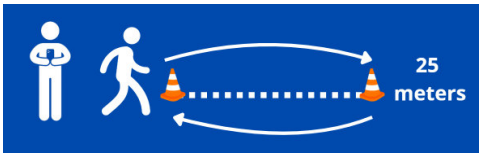


Fig. 3. Diagram showing the distance and how the patient completes the traverses in the test environment. The professional captures the data while stationary at the starting position throughout the entire walk, so that the RSSI variation signal can estimate the distance correctly.

IV. Results

A. App Interfaces and Functionalities

The developed system was designed to integrate oximetry monitoring with the patient's functional assessment (Figure 4). It was structured around three core functionalities: I. Patient and history management, featuring a patient registration module that allows the creation of individual patient profiles, from which the health professional can access the corresponding profile history; II. Monitoring and graphical visualization of the 6MWT, utilizing a real-time graphical interface for SpO_2 and HR levels that covers the entire exertion period (6 minutes) and the rest period (3 minutes), enabling the immediate identification of events such as exercise-induced desaturation with the real-time graphs that display the data updated at a frequency of 1 Hz; III. Gathering and presentation of the collected results, available in a section that summarizes patient performance through HR and SpO_2 variation graphs, walk statistics, anamnesis and metrics for prognostic assessment, including the exact total distance walked and the identification of the SpO_2 nadir, while also providing a downloadable PDF medical report with all this data.

B. Motion Artefact and Device Behaviour

The adjudication rule classified 7 of the 28 adjudicable recordings (25%) as affected by motion artefact during walking. Within the contaminated segments the two devices diverged by 91 bpm in HR and 43 percentage points in SpO_2 at the 95th percentile of the paired difference, reaching 122 bpm and 52 percentage points at their extremes. The divergence was not a matter of noise around a correct value: in the most affected recording the test device reported a constant 55 bpm for 134 consecutive seconds while the reference tracked the effort between 128 and 140 bpm, a pattern consistent with the device falling back to a held estimate once the plethysmographic

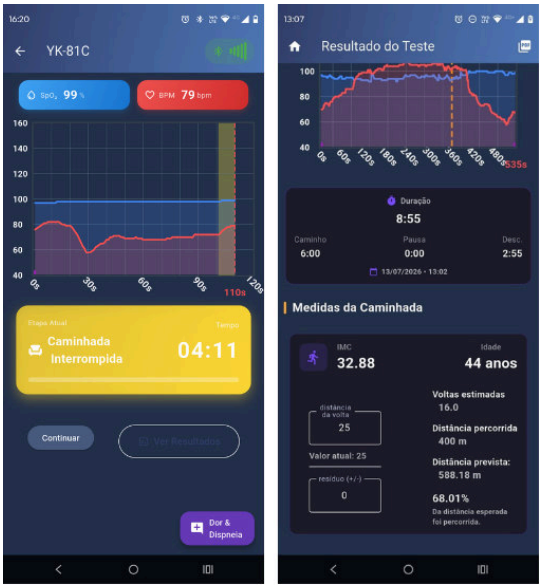


Fig. 4. Real-time graphical interface showing SpO_2 and HR levels during the test with the controls and interface (left). The red traced line shows the current value in time, the blue line represents SpO_2 and the red the HR. The yellow area is the interruption time, when a patient has to sit and rest in the middle of the walk. Results screen displayed after the walk and rest period; PDF medical report containing all the aforementioned data (right). The traced yellow line in the graph represents the start of the resting phase.

signal degrades. Applied to SpO_2 alone, the same rule flagged a further 6 recordings that visual inspection had not identified, in most cases involving only a few dozen samples, which indicates that contamination is distributed along a continuum rather than confined to the recordings classified as failures.

Artefact was confined to the walking phase. Segments persisting into rest occurred in 4 recordings, for 60, 42, 31 and 3 s beyond the end of walking, and in every case the segment had begun during walking, at 97, 271, 75 and 306 s respectively; no artefact originated in the resting phase itself. During rest the two devices were close to indistinguishable, with a median per-patient bias of $+0.19$ bpm for HR (IQR 1.02) and $+0.06$ percentage points for SpO_2 (IQR 1.78), obtained as the median across recordings of the mean paired difference within each one. Loss of the BLE link, a distinct failure mechanism, affected 3 further recordings, costing 38.8%, 35.7% and 18.3% of the walking-phase samples, the last of these concentrated in a single continuous gap of 50 s.

C. Distance estimation through RSSI

In this study, a traverse was defined as a single walk between the two endpoints of the course (A to B or B to A), such that a complete lap corresponds to two traverses. They were counted automatically from the smoothed RSSI signal, whose peaks and valleys – corresponding to the patient's closest and farthest approach to the beacon – were extracted by derivative analysis, and the resulting counts were compared against manual counting in the 31 recordings without substantial BLE signal loss. Normality was assessed with the Shapiro–Wilk test, which rejected normality for the estimated counts ($p = 0.017$), although not for the manual counts ($p = 0.081$). Thus, a non-parametric paired test was therefore adopted.

Manual counting and the software RSSI-based estimate were strongly correlated (Spearman $\rho = 0.97$), and the Bland–Altman analysis showed essentially no bias (median difference = 0.0), with limits of agreement of -2.25 to $+1.25$ traverses (-56.3 to $+31.3$ m) as shown in Figure 5. The paired Wilcoxon signed-rank test ($W = 47.0$, $p = 0.140$) detected no significant difference between methods. This result indicates the absence of a detectable systematic difference and should not be interpreted as evidence of equivalence, since no equivalence margin was pre-specified.

A TOST equivalence analysis was therefore performed: both one-sided Wilcoxon tests rejected their respective null hypotheses against the margin of ± 1 traverse (± 25 m), yielding $p_{\mathrm{TOST}} = 0.00014$, and the 90% bootstrap confidence interval for the mean paired difference was $[-0.06, +0.55]$ traverses ($[-1.6, +13.7]$ m), entirely contained within the margin. The data are therefore compatible with average equivalence between the two counting methods, with the caveat that the margin was not defined *a priori*. This equivalence concerns the central tendency of the paired differences and does not extend to individual agreement. The typical error was modest (mean absolute error of 17.7 m, with 87.1% of recordings within ± 25 m), but the deviations were markedly asymmetric: 14 recordings agreed exactly, 13 overcounted by a single traverse and one by two, while only 3 undercounted, by two or three traverses (-50 to -75 m).

The predominance of single-traverse overcounting is consistent with the algorithm's treatment of the final, incomplete traverse. A patient who is partway along the corridor when the six minutes elapse produces a partial excursion of the RSSI curve, which the extremum-based counter resolves as a whole traverse. The resulting error is bounded at one traverse, always in the direction of overestimation, and would inflate the reported distance by up to 25 m. Rounding the final traverse down, or weighting it by the amplitude of the incomplete excursion, is a straightforward correction that has not yet been implemented.

Taken together, these findings indicate that RSSI-based counting does not differ systematically from manual observation across the cohort, but is not interchangeable with it at the level of the individual test; its defensible role in the current version of the application is therefore that of an assistive, redundant record of the distance walked, rather than a replacement for manual counting.

V. Discussion and Limitations

This pilot work demonstrates the feasibility of instrumenting the 6MWT from a mobile device, raising the temporal granularity of the record to 1 Hz and automating report generation. Beyond the physiological monitoring itself, the app's RSSI-based distance estimation module showed promising agreement with manual traverse counting: the strong Spearman correlation ($\rho = 0.97$), the Bland–Altman median bias of essentially zero, and the *post-hoc* equivalence analysis

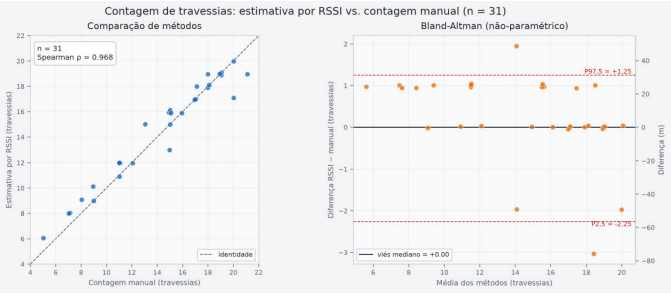


Fig. 5. Method comparison and non-parametric Bland-Altman analysis of manually counted vs. app-estimated traverses ($n = 31$; Spearman $\rho = 0.968$; median bias = 0.00, limits of agreement -2.25 to $+1.25$), showing strong correlation and no systematic bias between methods.

($p_{\mathrm{TOST}} = 0.00014$ against a ± 1 traverse margin) indicate that the two methods do not differ by a clinically meaningful amount on average. Average equivalence, however, is a weaker property than interchangeability at the level of the individual test. The null median bias arises from the cancellation of errors in opposite directions rather than from their absence, and the two directions have different explanations. Overcounting by a single traverse, by far the most frequent deviation, follows from the algorithm resolving the final incomplete traverse as a whole one: the error is bounded at 25 m, always inflates the distance, and is correctable in the counting logic. The three recordings that undercounted by two or three traverses are less readily explained, since they do not coincide with the loss of BLE signal that would suppress the transitions the algorithm relies on, and their cause could not be identified within this cohort. In metric terms, the typical error was modest (mean absolute error of 17.7 m, with 87.1% of tests within ± 25 m), but the lower limit of agreement (-56.3 m) and the largest individual error (-75 m) exceed the 30–50 m range considered clinically meaningful in the 6MWT [7]. Because the large deviations were predominantly undercounts, an error of this magnitude in a single test would underestimate the distance walked and could mask or mimic a clinically relevant change in functional capacity between serial assessments. Until the residual undercounting is understood, the module is best positioned as a redundant record supporting the examiner, and a properly powered study with a pre-specified margin will be required before manual counting can be safely omitted.

The dominant failure mode of the test device during the 6MWT was motion artefact. In 7 of the 28 adjudicable recordings (25%), HR and, in most of them, SpO_2 shifted abruptly during walking without the reference reproducing the change, which rules out a physiological origin. The character of these divergences is as relevant as their magnitude: in the most affected recording the device held a constant 55 bpm for 134 consecutive seconds while the reference recorded between 128 and 140 bpm. A reading of this kind does not present itself as an error: it is stable, plausible in isolation, and indistinguishable from a valid measurement on the screen.

The resting phase indicates where this failure originates. Across the same recordings the median per-patient bias during rest was $+0.19$ bpm for HR and $+0.06$ percentage points for SpO_2, with interquartile ranges close to a single unit, so that the two devices are effectively interchangeable once the patient stops walking. Every artefact segment that extended into rest had begun during walking, and none originated after it; the longest persisted for 60 s beyond the end of the walk before the two series converged again. The same device, on the same finger, in the same patient, therefore performs adequately at rest and fails under walking, which locates the problem in the mechanical conditions of the measurement rather than in the sensor's optical performance or in the BLE link. This also bounds the scope of the finding: it does not indicate that low-cost oximetry is unusable, but that this class of device is not adequate for the 6MWT, whose entire purpose is to impose the condition under which it fails.

Several limitations should be acknowledged. The cohort is small and single-centre, and patient characteristics such as sex, age, BMI, and underlying diagnosis were not analysed as potential sources of variability. The thresholds of the adjudication rule were chosen on physiological grounds rather than calibrated, and no sensitivity analysis over the parameter grid was performed, so the exact count of affected recordings should be read as dependent on them; the rule is nonetheless reported in full, so that the classification can be reproduced or recomputed under different criteria. The RSSI-based distance estimation method was validated in a single, fixed, and unobstructed 25 m corridor with the smartphone held by a stationary operator, and its behaviour under different corridor geometries, device positions, or interference conditions remains untested; a multi-environment study is required before generalisability can be claimed. The equivalence margin adopted for the traverse-count comparison was defined after data collection rather than pre-specified, so that analysis is exploratory and should be confirmed in a study designed around a margin agreed upon in advance. Finally, a plausible contributing factor to the artefact is device mass — approximately 50 g for the Wellue against 20 g for the Nonin-based model — concentrated at the fingertip, where the arm swing of walking imposes the largest accelerations. This points to a concrete design direction rather than to an inherent limit of low-cost oximetry: relocating the batteries and the display into a wrist-worn housing, leaving only a light and firmly coupled optical sensor on the finger, would reduce both the inertial load at the measurement site and the relative movement between sensor and tissue that produces the artefact. Such a modification is well within the reach of low-cost hardware and would be the first thing to test before this class of device is reconsidered for the 6MWT.

VI. Conclusion

This pilot study characterises the dominant failure mode of a low-cost, Bluetooth-connected pulse oximeter when integrated into the 6MWT. Motion artefact during walking rendered one recording in four unusable, producing divergences of 91 bpm at the 95th percentile and, in the worst case, a stable but fictitious reading held for over two minutes — a failure that is invisible to an operator watching the screen. The same devices

agreed closely at rest, and no artefact originated in the resting phase, which locates the problem in the mechanical coupling of the sensor under walking rather than in its optical performance or in the wireless link. The *SpO_2Sense* application itself demonstrated technical feasibility as a platform for instrumenting the 6MWT, successfully combining real-time Bluetooth-based physiological monitoring with automated distance estimation and report generation. The RSSI-based traverse counting module showed agreement with manual counting that was compatible with equivalence at a *post-hoc* margin of one traverse, although individual recordings deviated by up to three traverses; the most frequent of these deviations arises from the final incomplete traverse being resolved as a whole one, which is correctable in the counting logic. This supports its use as an assistive, redundant record of the distance walked rather than a replacement for manual observation, and justifies a properly powered validation study with a pre-specified margin. As a next step, future work will focus on a more thorough validation of the distance estimation algorithm across diverse corridor geometries and clinical settings, alongside refining the counting logic to improve individual-level accuracy. These results support the continued development of the *SpO_2Sense* software platform, while indicating that this class of oximeter is not adequate for the 6MWT in its present form, and that a redesign of the sensor's mechanical coupling — rather than further validation of the current hardware — is the appropriate next step.

REFERENCES

[1] ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories, "ATS statement: guidelines for the six-minute walk test," *American Journal of Respiratory and Critical Care Medicine*, vol. 166, no. 1, pp. 111–117, 2002.

[2] E. D. Chan, M. M. Chan, and M. M. Chan, "Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations," *Respiratory Medicine*, vol. 107, no. 6, pp. 789–799, 2013. [Online]. Available: https://doi.org/10.1016/j.rmed.2013.02.004

[3] Bluetooth SIG, "Bluetooth Core Specification," Bluetooth Special Interest Group, 2023.

[4] E. M. Marshall, A. Filiex, M. Schulte, K. M. Logie, J. Gin, M. Horrigan, L. Dowman, T. J. Churchward, and Y. H. Khor, "Performance Comparison of Oximetry Devices During 6-Minute Walk Tests," *Respiratory Care*, vol. 70, no. 5, pp. 492–501, 2025.

[5] Philip BJ, Abdelrazek M, Bonti A, Barnett S, Grundy J, "Data Collection Mechanisms in Health and Wellness Apps: Review and Analysis," *JMIR Mhealth Uhealth*, 2022 Mar 9;10(3):e30468.

[6] D. Salvi, E. Poffley, E. Orchard, and L. Tarassenko, "The Mobile-Based 6-Minute Walk Test: Usability Study and Algorithm Development and Validation," *JMIR mHealth and uHealth*, vol. 8, no. 1, p. e13756, 2020. [Online]. Available: https://mhealth.jmir.org/2020/1/e13756

[7] S. Perera, S. H. Mody, R. C. Woodman, and S. A. Studenski, "Meaningful Change and Responsiveness in Common Physical Performance Measures in Older Adults," *Journal of the American Geriatrics Society*, vol. 54, no. 5, pp. 743–749, 2006.

[8] D. P. Silva Junior, A. F. Müller, P. R. S. Sanches, L. Helena Ilha, M. Â. Moreira, S. S. M. Barreto, A. A. Susin, and P. Nohama, "Equipamento de biotelemetria para monitoramento em tempo real do teste da caminhada de seis minutos," in *Proc. XX Congresso Brasileiro de Engenharia Biomédica (CBEB)*, São Pedro, Brazil, 2006, pp. 720–723.

[9] Google LLC, "Flutter: build apps for any screen," 2024. [Online]. Available: https://flutter.dev

[10] Google LLC, "Firebase." [Online]. Available: https://firebase.google.com/

A Web Platform for Quality-of-Life Assessment in Sickle Cell Anemia with Automated SF-12 Scoring

Monique da Silva Souza Prates
Health Technology Laboratory (LaTeS)
Universidade Brasil
São Paulo, Brazil
moniqueazuos@hotmail.com

Kennedy Rodrigo Prates
Health Technology Laboratory (LaTeS)
Universidade Brasil
São Paulo, Brazil
kennedyrodrigo@hotmail.com

Telma Lisboa do Nascimento
Health Technology Laboratory (LaTeS)
Universidade Brasil
São Paulo, Brazil
telma.nascimento@ub.edu.br

Abstract—Sickle cell anemia (SCA) is a major hereditary hemoglobinopathy in Brazil and disproportionately affects the Black population. This study describes a Portuguese-language web prototype that integrates educational content, informed consent, a sociodemographic questionnaire, and an SF-12 scoring workflow. Development was informed by a literature review structured according to PRISMA guidance. The prototype underwent preliminary internal checks of navigation, form flow, responsive display, and score-generation behavior. No deployment with the intended participant cohort, formal usability study, clinical outcome analysis, or independent agreement study of automated scores has yet been completed. Authorization to use the SF-12 and its scoring procedure is currently being requested; participant deployment and operational use of scoring are contingent on confirmation of the applicable conditions. Thus, the contribution is a technological prototype and a proposed validation workflow, not evidence of improved quality of life, adherence, follow-up, or clinical decision-making. Future work will include formal score verification, usability, accessibility, security, and performance testing, followed by prospective evaluation.

Index Terms—sickle cell anemia, quality of life, SF-12, digital health, public health.

I. Introduction

Sickle cell disease designates hereditary hemoglobinopathies associated with hemoglobin S; SCA is its most severe form [1]. A point mutation in the β-globin gene produces hemoglobin S, which polymerizes under low oxygen tension, deforms erythrocytes, and contributes to vaso-occlusion, hemolysis, and painful crises [2]. Systemic complications include acute chest syndrome, stroke, renal disease, retinopathy, and priapism [2, 3]. In Brazil, sickle cell disease is a public-health concern and is addressed through neonatal screening and early diagnosis policies [4].

The disease burden intersects with social determinants and racial inequities in access to care. Recurrent pain, fatigue, hospitalization, and occupational limitations can affect daily functioning and social participation. Systematic assessment of health-related quality of life may therefore help characterize needs, provided that the instrument and digital workflow are appropriately authorized and validated.

The 12-item Short-Form Health Survey (SF-12) summarizes physical and mental health into the Physical Component Summary (PCS) and Mental Component Summary (MCS) [5]. Portuguese-language applicability and Brazilian reference values have been reported [6, 7]. Scoring is not a simple sum and is susceptible to transcription or implementation errors. Digital integration may reduce manual steps, but technical correctness, authorization, usability, and clinical value must be demonstrated separately.

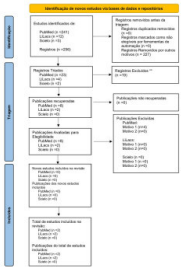


Figure 1: PRISMA flow diagram of the study-selection process. Counts must match the preserved review records.

This work describes the development of a Portuguese-language web prototype that combines educational content, consent, sociodemographic data collection, and a planned SF-12 scoring workflow. The present objective is to document the prototype and the validation protocol still required. No benefit to follow-up, adherence, quality of life, or healthcare decisions is claimed at this stage.

II. Materials and Methods

A. Design and ethical status

This is applied technological-development research informed by a literature review. The protocol for the planned participant stage was approved by the Research Ethics Committee of Universidade Brasil (CAAE 89209325.2.0000.5494; Opinion No. 7,930,601) and has a letter of consent from the São Paulo State Sickle Cell Anemia Association (AAFESP). No participant recruitment or quality-of-life data collection had occurred at the time of this report. Ethics approval does not replace authorization or licensing requirements associated with the SF-12.

B. Literature review

The review was structured using PRISMA guidance and searched PubMed, SciELO, and LILACS for publications from 2019 to 2024, combining terms related to SCA, quality of life, SF-12/SF-36, and digital or mobile applications. Original studies in Portuguese, English, or Spanish involving adults were eligible; reviews and records without accessible full text were excluded. Screening was performed independently and disagreements were resolved by consensus. For full reproducibility, the final review record must preserve each database string, last-search date, deduplication method, counts at every screening stage, reasons for exclusion, and the chosen methodological-quality appraisal.

C. Requirements and architecture

Requirements were organized around four participant-facing functions: educational access, consent, sociodemographic/current-health data collection, and the SF-12 workflow. The prototype was developed at LaTeS using HTML5, CSS3, and JavaScript for the responsive interface, with reported Python/PHP components and a MySQL database.



Figure 2: QR code for the prototype. Public access to the instrument/scoring module should remain restricted while authorization is pending.

The functional sequence is browser interface → application logic → database. The current version does not demonstrate electronic-health-record interoperability or production-scale performance.

Before participant deployment, the production specification must document role-based access, authentication, encryption in transit and at rest, database privileges, audit logs, backup and recovery, retention/deletion rules, incident response, and responsibility for each processing operation. These are deployment requirements and are not reported here as validated functions.

D. SF-12 authorization and scoring workflow

The prototype contains a workflow intended to collect SF-12 responses and calculate PCS and MCS. Authorization to use the instrument and applicable scoring materials is currently being requested. Participant deployment and operational release of scores will occur only after the applicable conditions are confirmed and documented. The Brazilian studies cited above support contextual interpretation but do not themselves establish reproduction or automation rights.

Before deployment, automated outputs must be compared with independently calculated scores derived from an authorized reference procedure. Predefined test vectors, expected values, tolerance criteria, discrepancy logs, and the tested software version must be reported. Required items must be checked before score generation; missing or invalid answers must trigger a correction request rather than a score. The prototype provides no diagnosis or treatment recommendation.

E. Participant flow and data protection

The planned convenience sample comprises 50 adult AAFESP participants of both sexes. The participant first receives the Informed Consent Form (ICF); refusal terminates the research path before questionnaire submission. Consent is followed by the sociodemographic/current-health questionnaire and, subject to authorization, the SF-12 module. Processing of sensitive health data must comply with Law No. 13,709/2018, CONEP Circular Letter No. 1/2021, and National Health Council Resolution No. 466/2012. Compliance requires both governance and verifiable technical safeguards; the production controls listed above must be confirmed before recruitment.

F. Verification and planned evaluation

The current evidence is limited to preliminary internal checks of navigation, consent branching, required-field behavior, responsive display, and score-generation behavior with nonparticipant inputs. These checks are not clinical validation or formal usability assessment. The prospective validation protocol comprises: (1) authorized reference-score comparison; (2) missing/invalid-input tests; (3) reliability and response-time measurement under specified conditions; (4) accessibility checks; (5) security testing; and (6) usability testing with patients and health professionals



Figure 3: Prototype home page.



Figure 4: Educational content section on sickle cell anemia.

using a validated instrument and task-based measures. Outcomes will be reported only after observation.

III. Results and Discussion

A. Review and implemented prototype

The review indicated limited simultaneous treatment of SCA, standardized quality-of-life measurement, and integrated digital follow-up. The available literature reinforces the burden of pain, fatigue, hospitalization, stigma, and mental-health effects, while the identified digital gap motivates technological investigation rather than proving clinical effectiveness.

The responsive prototype is organized into educational and research-flow sections. The educational area addresses definition, symptoms, diagnosis, neonatal screening, comorbidities, treatments, quality of life, and social aspects. The research path presents consent, sociodemographic/current-health questions, and the planned SF-12 workflow (Figs. 3–8). Screenshots demonstrate interface states only; they are not evidence of user acceptance, clinical utility, authorized operational use, or patient outcomes.

B. Automated scoring as a technical hypothesis

Unifying collection and calculation may reduce transcription steps, but this advantage has not yet been quantified. Formal validation requires agreement between automated and independently calculated authorized reference scores. Therefore, the present study does not establish superior accuracy, fewer errors, better engagement, or clinical usefulness. Any novelty statement is limited to the scope and databases searched.

C. Potential relevance to the SUS

A Portuguese-language workflow accessible on common devices could support future quality-of-life data collection in patient associations and outpatient settings. Its relevance to the Brazilian Unified Health System remains a translational hypothesis requiring authorized instrument use, usability/accessibility evaluation, documented security controls, interoperability planning, and prospective measurement of data completeness, professional workload, decision support, and patient outcomes.

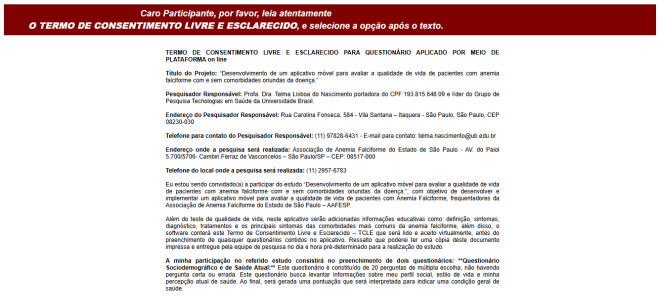


Figure 5: Informed Consent Form interface.



Figure 6: Sociodemographic and current-health questionnaire.

D. Limitations

This study is limited to technological development, a literature review, and preliminary internal checks. No participant deployment, quality-of-life analysis, formal usability study, accessibility audit, performance/load assessment, penetration test, or clinical-workflow evaluation was performed. Agreement with independently calculated authorized scores has not been reported, and SF-12 authorization is in progress. The review report also requires complete search strings, dates, screening counts, and methodological appraisal. Interoperability and long-term scalability were not evaluated. The findings cannot be generalized to clinical effectiveness or patient benefit.

IV. Conclusion

A Portuguese-language web prototype was developed to integrate educational content, informed consent, sociodemographic data collection, and a planned SF-12 workflow for adults with SCA. The study documents the prototype and defines remaining validation steps; it provides neither patient quality-of-life results nor evidence of improved follow-up, adherence, decisions, or outcomes. Authorization for SF-12 use and scoring is being requested, and participant deployment is contingent on resolution of the applicable conditions. Future work includes reproducible score verification, usability and accessibility testing with intended users, security and performance assessment, and prospective multicenter evaluation.

Acknowledgments

The authors thank AAFESP and Universidade Brasil.

References

[1] R. D. Cançado and J. A. Jesus, "Sickle cell disease in Brazil," *Rev. Bras. Hematol. Hemoter.*, vol. 29, no. 3, pp. 203–206, 2007.

[2] F. B. Piel, M. H. Steinberg, and D. C. Rees, "Sickle cell disease," *N. Engl. J. Med.*, vol. 376, no. 16, pp. 1561–1573, 2017.

[3] L. Abdel-Hadi *et al.*, "Treatment of sickle cell disease—options and perspective," *Am. J. Blood Res.*, vol. 13, no. 2, pp. 61–70, 2023.

[4] Brazil, Ministry of Health, "National Government reinforces the need for early diagnosis of Sickle Cell Disease," Brasília, DF, 2022.



Figure 7: SF-12 interface. Operational use remains contingent on authorization.

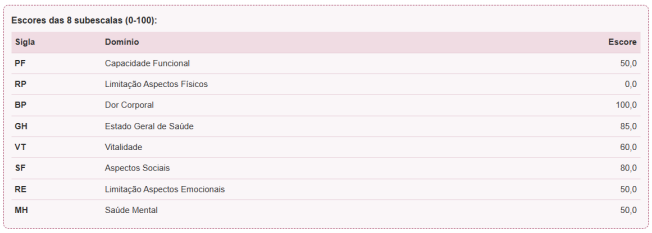


Figure 8: Prototype score display. Operational use remains contingent on authorization and validation.

[5] J. E. Ware, M. Kosinski, and S. D. Keller, "A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity," *Medical Care*, vol. 34, no. 3, pp. 220–233, 1996.

[6] L. A. Andrade *et al.*, "Applicability of the 12-Item Short-Form Health Survey in patients with progressive systemic sclerosis," *J. Bras. Pneumol.*, vol. 33, no. 4, pp. 414–422, 2007.

[7] R. V. M. Lopez, E. P. Nardi, and M. B. Ferraz, "Quality of life in a sample of Brazilian adults using the generic SF-12 questionnaire," *Rev. Assoc. Med. Bras.*, vol. 64, no. 3, pp. 234–241, 2018.

Age Composition of t-SNE Clusters in a Methylation Study of Women at Risk for Breast Cancer

1st Leandro Carneiro de Oliveira

Escola Politécnica de Pernambuco, POLI
Universidade de Pernambuco (UPE)
Recife, Brazil
0009-0004-0386-1233

2nd José Barbosa de Araújo Neto

Depto. de Engenharia Biomédica, UFPE
Universidade Federal de Pernambuco (UFPE)
Recife, Brazil
0009-0007-4994-1695

3rd Maíra Araújo de Santana

Centro de Informática
Universidade Federal de Pernambuco (UFPE)
Recife, Brazil
0000-0002-1796-7862

4th Wellington Pinheiro dos Santos

Depto. de Engenharia Biomédica, UFPE
Universidade Federal de Pernambuco (UFPE)
Recife, Brazil
0000-0003-2558-6602

***Abstract*—A multiclass machine-learning framework applied to peripheral blood DNA (deoxyribonucleic acid) methylation profiles previously demonstrated that the pre-tumorigenic RISK (at risk) group exhibits a distinctive epigenetic signature separable from both healthy controls and confirmed breast cancer patients. Within that group, two visually distinct sub-clusters were observed in t-SNE (t-distributed stochastic neighbor embedding) projections, but their biological basis was not addressed. The present study investigates whether chronological age, stratified into three bands (age $<$ 60, 60 $\leq$ age $<$ 70, and age $\geq$ 70), accounts for the substructure observed in the RISK population. Methylation profiles were obtained from 282 at risk women in the GSE58119 cohort (Gene Expression Omnibus series GSE58119), assayed using the Illumina HumanMethylation27 BeadChip, which measures 27,578 CpG (cytosine-phosphate-guanine) sites. They were then re-labelled by age group and re-projected via t-SNE. The visualization reveals that age stratification does not cleanly separate the two clusters: women from all three age bands are present in both clusters, with Cluster 2 containing the majority of every age group. A modest age related trend is nonetheless visible, with the proportion of women assigned to Cluster 1 decreasing slightly with age (40.7% for age $<$ 60, 39.6% for 60 $\leq$ age $<$ 70, and 28.6% for age $\geq$ 70). Statistical testing confirms that this trend is not significant (ANOVA $p = 0.530$; Kruskal–Wallis $p = 0.619$; χ^2 $p = 0.418$; $\eta^2 = 0.001$), even though the two clusters are themselves highly robust (silhouette = 0.663; permutation test $p = 0.001$). These findings indicate that age has only a modest, non-significant influence on the two cluster topology and that other biological or clinical factors must be responsible for the primary separation, motivating further investigation of additional covariates within the pre-malignant state.**

***Index Terms*—DNA (deoxyribonucleic acid) methylation; breast cancer; pre-tumorigenic risk; t-SNE (t-distributed stochastic neighbor embedding); age stratification; epigenetic heterogeneity; CpG (cytosine-phosphate-guanine) islands; machine learning.**

I. Introduction

Breast cancer is one of the foremost contributors to cancer-related mortality among women globally, and its biological complexity continues to pose significant challenges to early detection and risk stratification [1]. While conventional screening modalities such as mammography identify the disease only after tumor formation has occurred, there is growing recognition that epigenetic alterations precede histopathological transformation by several years [2], [3]. Among these, DNA methylation at CpG (cytosine-phosphate-guanine) dinucleotides has attracted considerable attention as a minimally invasive biomarker measurable in peripheral blood [4]. A multiclass classification framework developed to distinguish healthy women, women at elevated risk, and confirmed breast cancer patients from peripheral blood methylation profiles demonstrated strong performance, particularly for the at risk cohort [5]. The AdaBoost (Adaptive Boosting) model achieved 89.5% overall accuracy and an F1-score (F1; harmonic mean of precision and recall) of 0.991 for the RISK class, while Kruskal-Wallis testing confirmed that 90-93% of the top 100 CpG islands discriminated the RISK group from both healthy

CNPq

controls and cancer patients. Dimensionality-reduction analyses using Principal Component Analysis (PCA) and t-SNE further corroborated these findings by revealing a spatially separated cluster for the RISK population. Crucially, the t-SNE projection of the 282 at risk samples disclosed two visually distinct sub-clusters whose biological basis was not investigated. Understanding the source of this intra-group heterogeneity is important because it directly informs whether a single risk assessment model is sufficient or whether separate stratum-specific classifiers are warranted. Age is a well-established modulator of DNA methylation: epigenetic clocks demonstrate that global methylation patterns drift predictably with chronological age [6], and breast cancer incidence is strongly age dependent, peaking in postmenopausal women [7]. While our previous work [5] established the existence of a distinct epigenetic signature in the RISK group and identified a two cluster substructure via t-SNE, the biological drivers of this intra-group heterogeneity remained unexplored. The primary innovation of the present study is the systematic interrogation of this substructure using chronological age as a stratifying variable, coupled with cluster-validation metrics and multi method dimensionality reduction. By demonstrating that age does not account for this topological separation, this work provides a crucial negative finding that redirects future biomarker discovery away from age stratified models and toward the investigation of other latent clinical or biological covariates (e.g., BRCA1/2 status, hormonal history) within the pre-malignant state.

II. RELATED WORK

The relationship between DNA methylation and aging has been extensively characterized. Horvath demonstrated that a weighted combination of 353 CpG sites tracks chronological age across multiple tissue types with high precision [6], and subsequent epigenetic-clock studies revealed that accelerated methylation aging in blood cells is associated with increased cancer incidence [8]. In the context of breast cancer, Hofstatter et al. showed that epigenetic age in normal breast tissue is elevated in luminal cancer patients compared with healthy controls, underscoring a potential link between methylation-age dysregulation and oncogenic predisposition [9]. At the population level, several prospective cohort studies have confirmed that peripheral blood methylation profiles can discriminate women who subsequently develop breast cancer from those who do not, even when samples are drawn years before diagnosis [10], [11]. However, these analyses have generally treated the at risk group as homogeneous rather than stratifying by age. The present work addresses this gap by leveraging age annotated longitudinal data. From a computational perspective, t-SNE has become a standard tool for exploring high-dimensional omics data [12], recent state-of-the-art approaches emphasize the necessity of validating non-linear embeddings with complementary methods like UMAP and PCA, alongside internal clustering metrics (e.g., silhouette, Davies–Bouldin), to ensure that observed topologies reflect genuine biological structure rather than algorithmic artifacts [13], [14]. Furthermore, contemporary breast cancer risk prediction models increasingly integrate epigenetic clocks with traditional risk factors; however, few studies have explicitly tested whether age driven methylation drift is the primary axis of variation in pre-tumorigenic sub-clusters. This study bridges that gap by applying a multi method validation framework to an age annotated at risk cohort.

III. MATERIALS AND METHODS

A. Dataset and cohort description

The dataset employed in this study is the publicly available GSE58119 cohort (Gene Expression Omnibus (GEO) series GSE58119), deposited in the Gene Expression Omnibus (GEO) [15] and processed with the Illumina Infinium HumanMethylation27 BeadChip v1.2, which measures methylation beta values at approximately 27,578 CpG (cytosine-phosphate-guanine) sites in peripheral blood cells. This work uses the same RISK subset of GSE58119 (n = 282) analyzed previously in [5]; the present contribution differs in that we explicitly stratify and analyze the RISK samples by chronological age to interrogate the previously observed two cluster t-SNE topology. The cohort comprises 282 women from a nested case-control design within a prospective breast cancer screening programme. Women who received a clinical breast cancer diagnosis within five years of blood draw were labelled PRE-BRCA (pre-diagnosis breast cancer; n = 148), while those who remained cancer-free were labelled HEALTHY-UNK (healthy, unknown follow-up; n = 134). Both subgroups are here collectively termed the RISK (at risk) population. Participant age at blood draw is recorded in the original GEO submission and serves as the primary covariate of interest.

B. Age stratification

Participants were stratified into three age bands following thresholds commonly used in clinical breast cancer research: (i) age < 60 years, representing predominantly premenopausal and early postmenopausal women; (ii) $60 \leq$ age < 70 years, an intermediate postmenopausal window; and (iii) age ≥ 70 years, representing older postmenopausal women in whom age associated methylation drift is expected to be most pronounced. The resulting age band totals across the RISK group are summarized in Table I.

TABLE I: age stratified sample distribution within the RISK group (n = 282)

Age band	**n**
Age < 60	113
$60 \leq$ Age < 70	134
Age ≥ 70	35
Total	**282**

C. Preprocessing

All preprocessing steps followed the same pipeline described in [5]; no additional normalization, feature selection, or re-weighting was introduced for the present analysis. Beta

methylation values were computed from raw fluorescence intensities as

$$\beta = \frac{TI}{U + TI + \alpha}$$

where TI is the total intensity of the methylated allele, U is the intensity of the unmethylated allele, and $\alpha = 100$ is an offset that stabilizes estimates and prevents division by zero. Missing values were imputed with the per CpG island mean across all samples. The final feature matrix comprised 282 samples x 27,578 CpG sites. No synthetic oversampling was applied at this stage, as the objective is exploratory visualization rather than classification; the key methodological change relative to [5] is the age based re-labelling and stratified interpretation of the embedding.

D. t-SNE configuration

t-SNE was applied to the full 27,578-dimensional methylation matrix using the scikit-learn (Python machine-learning library) implementation (perplexity = 30, learning rate = 200, n_iter = 1,000, random_state = 42). Samples were colored according to their age stratified sub-category label across six categories: PRE-BRCA AGE $<$ 60, PRE-BRCA 60 $\leq$ AGE $<$ 70, PRE-BRCA AGE $\geq$ 70, HEALTHY-UNK AGE $<$ 60, HEALTHY-UNK 60 $\leq$ AGE $<$ 70, and HEALTHY-UNK AGE $\geq$ 70. This annotation scheme enables simultaneous assessment of clinical status and age related positioning within the embedding.

E. Comparison of dimensionality-reduction methods

To confirm that the observed two cluster topology is not an artefact of t-SNE, three additional projections of the same standardized methylation matrix were computed for comparison: Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP, $n_neighbors = 15$, $min_dist = 0.1$), and a Linear Discriminant Analysis (LDA) supervised directly by the same k-means reference labels obtained from the t-SNE coordinates (Figure 2). For PCA and UMAP, which are unsupervised, comparing their embeddings against these reference labels constitutes a genuine, label-independent check of whether the bipartition is recoverable by methods that had no access to the cluster assignments. The LDA projection is different because it is explicitly trained to maximize separation between the two k-means classes, strong separation in the LDA axis is expected by construction and does not provide independent evidence of cluster robustness. We therefore report the LDA result only as a sanity check that the two groups are linearly separable in methylation space, not as confirmation of the bipartition's validity.

F. Cluster definition and statistical analysis

To avoid the subjectivity of manually drawn visual boundaries, the two sub-clusters were defined objectively by applying k-means clustering ($k = 2$, random state=42) to the two dimensional t-SNE coordinates. To ensure cluster stability, this partitioning was verified across multiple random seeds (42, 0, 123), yielding consistent assignments ($> 98\%$ agreement). The choice of $k = 2$ was further supported by a silhouette analysis over $k \in [2, 8]$ computed on the standardized methylation matrix, which peaked at $k = 2$ (silhouette = 0.290). Cluster validity was assessed with the silhouette, Davies–Bouldin, and Calinski–Harabasz indices, and the separation was tested for significance with a 1,000-iteration label-permutation test on the silhouette score. The association between chronological age and cluster membership was then evaluated with one-way ANOVA and the Kruskal–Wallis test (age as a continuous variable across clusters), a χ^2 test of independence (age band $\times$ cluster), and the η^2 statistic (fraction of age variance explained by cluster membership).

IV. RESULTS

A. Visual topology of the age stratified t-SNE

The t-SNE projection of the 282 at RISK samples, presented in Fig. 1, confirms the same two cluster topology previously reported for this cohort in [5]: a compact, high-density cluster in the upper-right region of the embedding (approximately TSNE1 (t-SNE dimension 1) in [6, 16], TSNE2 (t-SNE dimension 2) in [4, 10]) and a more dispersed cluster occupying the lower-left quadrant (approximately TSNE1 in [-16, 0], TSNE2 in [-7, 1]). Color-coded by age stratum (a new annotation layer relative to [5]), the visualization reveals that all three age bands are present in both clusters, without any age defined region of the embedding. Younger women (age $<$ 60) are distributed across both clusters, with a slight relative enrichment in Cluster 1 compared with older women (40.7% vs. 28.6% of their respective age groups in Cluster 1). The intermediate band (60 $\leq$ age $<$ 70) shows an almost identical distribution to the youngest band, with 39.6% of its members in Cluster 1. Older women (age $\geq$ 70) are the least likely to appear in Cluster 1, with 71.4% assigned to Cluster 2. Crucially, Cluster 2 contains the majority of women in every age band, including the youngest one. There is therefore no "one age in one cluster, the other age in the opposite cluster" separation; instead, the data show substantial overlap of all age groups across both clusters, with only a modest trend toward a lower Cluster 1 share among older participants. Taken together, these observations indicate that chronological age contributes at most a weak gradient to the spatial organization of the RISK population in the embedding and does not account for the two cluster structure reported in [5].

B. age cluster correspondence

To quantify the degree to which age bands map onto the two clusters, samples were partitioned into the two groups returned by k-means clustering ($k = 2$) of the t-SNE coordinates (Cluster 1: upper-right; Cluster 2: lower-left). The resulting counts and percentages are reported in Table **??** (composition by age band) and Table II (per age band share between clusters). Cluster 1 contains 109 women (38.7% of the RISK group) and Cluster 2 contains 173 women (61.3%). Within each age band, the majority of women fall into Cluster 2 (59.3%, 60.4%, and 71.4% for the three age bands respectively), confirming

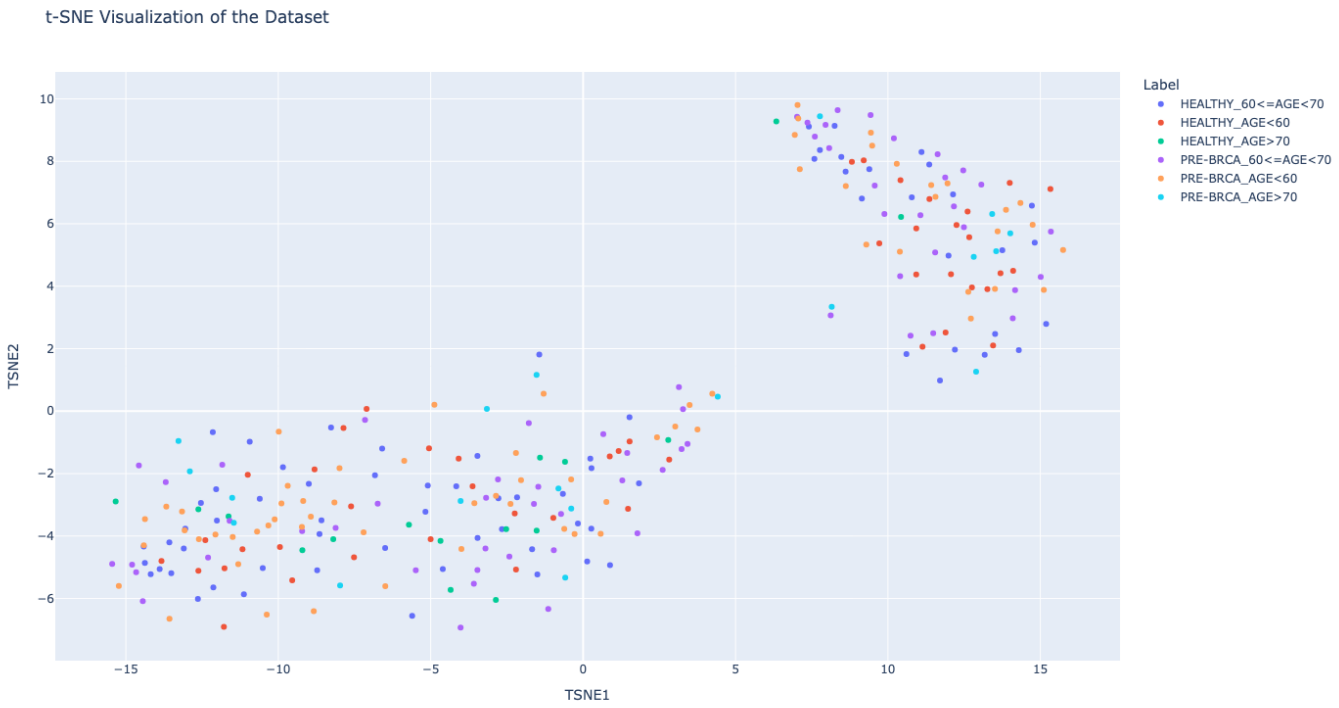


Fig. 1: age stratified t-SNE projection of the RISK sub-cohort (n = 282). Two clusters are visible; all three age bands are present in both clusters, and age provides only modest explanatory coverage of the observed sub-structure.

that neither cluster is dominated by a single age stratum and that the two cluster topology is not driven primarily by age.

The two clusters are themselves well separated and statistically robust: the silhouette score reaches 0.663 in the t-SNE embedding (Davies–Bouldin = 0.453, Calinski–Harabasz = 971.6) and remains positive in the original 27,578-dimensional space (silhouette = 0.246), where the label permutation test rejects the null hypothesis of no structure (permuted silhouette 0.000 ± 0.004, $p = 0.001$). In sharp contrast, no statistical association is found between age and cluster membership: one-way ANOVA ($F = 0.396$, $p = 0.530$) and the Kruskal–Wallis test ($H = 0.247$, $p = 0.619$) detect no difference in age distribution between clusters, the χ^2 test of independence between age band and cluster is non-significant ($\chi^2 = 1.747$, $\mathrm{dof} = 2$, $p = 0.418$), and cluster membership explains essentially none of the variance in age ($\eta^2 = 0.001$). Thus the substructure is highly significant while its alignment with age is not.

TABLE II: Cluster assignment within each age band (% of the age band)

Age band	Cluster 1	Cluster 2	Total
Age $<$ 60	46 (40.7%)	67 (59.3%)	113
60 $\leq$ Age $<$ 70	53 (39.6%)	81 (60.4%)	134
Age $\geq$ 70	10 (28.6%)	25 (71.4%)	35
Total	**109**	**173**	**282**

C. Robustness across dimensionality-reduction methods

The two cluster structure is reproduced by every unsupervised projection examined, confirming that it reflects genuine structure in the methylation data rather than a peculiarity of t-SNE. Figure 2 shows the four embeddings side by side: both t-SNE and UMAP resolve the RISK group into two well separated point clouds, and PCA preserves the same separation along its leading components (PC1 and PC2 capturing 39.8% and 9.2% of the variance, respectively; 51.3% cumulative over three components), all without having been given the cluster labels. The LDA projection also spreads the samples along a dominant discriminant axis, but because LDA was supervised with the same k-means labels being evaluated, this separation is expected by construction and is reported only as a sanity check, not as independent evidence of robustness. As in the t-SNE map, no projection exhibits an age defined region: the six age stratified colour groups are intermixed throughout each embedding.

The quantitative agreement is summarized in Table III, which reports the three internal validation indices computed against the common k-means reference labels. t-SNE yields the cleanest separation (silhouette = 0.663, Davies–Bouldin = 0.453, Calinski–Harabasz = 971.6), closely followed by UMAP (0.650, 0.463, 811.0); PCA preserves a moderately strong separation in only two linear dimensions (0.495, 0.806, 333.7), and the same partition remains detectable even in the original 27,578-dimensional space (0.246, 1.588, 103.1). The



Fig. 2: Comparison of four dimensionality-reduction methods (t-SNE, PCA, UMAP, and LDA) applied to the same 282-sample RISK methylation matrix, coloured by the six age stratified sub-categories. The two cluster topology is reproduced by every method, and no embedding shows an age defined region.

consistency of these indices across linear (PCA) and non-linear (t-SNE, UMAP) unsupervised techniques demonstrates that the bipartition is a robust, method-independent property of the RISK cohort; the supervised LDA result is consistent with this conclusion but, being circular by construction, is not counted as independent supporting evidence.

TABLE III: Internal cluster validation indices for each embedding, computed against the common k-means reference labels

Method	Silhouette	Davies–Bouldin	Calinski–Harabasz
Original space	0.246	1.588	103.1
t-SNE	0.663	0.453	971.6
PCA (2D)	0.495	0.806	333.7
UMAP	0.650	0.463	811.0

V. DISCUSSION

A. Summary of findings

The primary finding of this study is that chronological age does not cleanly separate the two cluster substructure of the RISK population in the t-SNE space. All three age bands are present in both clusters, and Cluster 2 contains the majority of women in every age band. A modest age related trend is observed, with the proportion of women assigned to Cluster 1 decreasing slightly from 40.7% (age < 60) to 39.6% ($60 \leq$ age < 70) and 28.6% (age ≥ 70); however, this trend does not reach statistical significance under ANOVA, Kruskal–Wallis, or χ^2 testing ($p = 0.530$, 0.619, and 0.418 respectively), and cluster membership accounts for only $\eta^2 = 0.001$ of the variance in age. The clusters themselves, by contrast, are statistically robust (permutation test $p = 0.001$). This result has both conceptual and practical implications for breast cancer risk stratification.

B. Biological interpretation

From a conceptual standpoint, the modest age related trend in cluster assignment is consistent with the well established phenomenon of epigenetic drift, in which methylation at specific CpG sites changes as a function of accumulated cellular divisions and lifelong environmental exposures [16]. Postmenopausal women exhibit distinct methylation signatures relative to premenopausal counterparts, reflecting hormonal shifts that modulate DNA methyltransferase (DNMT) activity and methyl-group availability [17]. The slightly lower share of older women in Cluster 1 is compatible with this background biology. However, the substantial overlap of all age groups across both clusters indicates that the dominant axis of variation between Cluster 1 and Cluster 2 is not chronological age. One candidate explanatory variable is already present in our own metadata and was used to construct the six-category colour scheme in Fig. 1: case/control status, i.e., whether a participant was subsequently diagnosed (PRE-BRCA, $n = 148$) or remained cancer-free (HEALTHY-UNK, $n = 134$).

Unlike BRCA1/2 mutation carrier status, parity, body mass index (BMI), and use of hormonal therapies which genuinely could not be evaluated here due to metadata limitations case/control status requires no additional data collection, and testing its association with the two cluster partition (e.g., via the same χ^2 and permutation-based framework applied to age in Section III-E) is a necessary analysis that we have not yet completed and that should be reported before this work is considered finished. A finding that the two sub-clusters align with future diagnosis status, rather than with age, would substantially change the interpretation and significance of this study: it would suggest the t-SNE bipartition itself carries prognostic information, rather than being an unexplained structural curiosity. We flag this explicitly as the most important open question raised by the present analysis, ahead of the other candidate covariates listed below. Future work should systematically test each of these variables as potential explanatory factors for the primary two cluster separation.

C. Implications for risk modeling

From a practical standpoint, the limited age dependence of the cluster assignment has direct implications for biomarker discovery and the design of future screening models. Because the two t-SNE sub-clusters are not aligned with age strata, a single pooled classifier is unlikely to be substantially confounded by chronological age alone. However, it may also fail to capture the as-yet-unidentified covariate that drives the actual cluster separation. Consequently, future breast cancer risk-stratification models should not rely solely on age adjusted classifiers. Instead, biomarker discovery pipelines must incorporate multi omic or detailed clinical covariate data to isolate the true biological signal driving this epigenetic heterogeneity, thereby enabling more precise, personalized screening intervals for women in the pre-malignant state.

The intermediate age band ($60 \leq \text{age} < 70$) is the largest in this cohort and shows a distribution between the two clusters that is essentially indistinguishable from that of the youngest band, possibly reflecting the biological transitional nature of this period: late perimenopausal or early postmenopausal, with high inter-individual variability in hormonal and epigenetic trajectories. This observation reinforces the conclusion that age, on its own, is insufficient to define risk relevant epigenetic subgroups within the at risk population.

D. Limitations and future work

A limitation of the present analysis is its post-hoc, observational nature. A more specific limitation is that, while we used PRE-BRCA/HEALTHY-UNK status to colour Fig. 1, we did not statistically test whether this case/control label rather than age explains the two cluster partition, even though this label was already available and required no additional data collection; this test should be carried out and reported as a direct extension of Section III-E before any claim about the explanatory irrelevance of clinical covariates can be considered complete. Furthermore, while t-SNE is highly effective for visualizing local neighborhood structures, it is an exploratory tool whose output can be sensitive to hyperparameters such as perplexity, learning rate, random seed, and distance metrics. Although we mitigated this by confirming topological stability across UMAP and PCA (with LDA included only as a non-independent sanity check, see Section III-D), the subjective interpretation of 2D projections remains a constraint. Additionally, the age stratification was applied to a t-SNE computed on the full RISK group without a priori selection of age informative CpG sites. We also did not assess whether batch or array-processing effects (e.g., chip or scan-date structure), a recognized source of artefactual clustering on Illumina BeadChip platforms, could contribute to the observed bipartition. Future work will address these limitations by: (i) testing the association between the two cluster partition and case/control (PRE-BRCA vs. HEALTHY-UNK) status; (ii) applying age regressed methylation matrices to confirm that the two cluster structure persists after mathematically removing age associated variation; (iii) checking the partition against available batch/array metadata; and (iv) systematically evaluating the impact of specific clinical covariates available in extended metadata, such as BRCA1/2 mutation status, body mass index (BMI), parity, hormone replacement therapy, and time to diagnosis, to identify the true drivers of this epigenetic substructure.

VI. CONCLUSIONS

Building directly on [5] and using the same GSE58119 RISK cohort and preprocessing pipeline, this work demonstrates that the two sub-clusters observed in the t-SNE projection of this pre-tumorigenic breast cancer risk population are not cleanly separated by chronological age. Women from all three age bands are distributed across both clusters, with Cluster 2 containing the majority of every age group. Only a modest age related trend is observed, with the proportion of women assigned to Cluster 1 decreasing slightly with age (40.7%, 39.6%, and 28.6% for the three age bands). Age therefore has a modest, non-significant influence on cluster assignment and does not explain the two cluster structure. Rather than definitively excluding age's role in epigenetic heterogeneity, these findings suggest that other biological or clinical factors are likely the primary drivers of this separation. These findings carry direct implications for the design of risk-stratification models based on peripheral blood DNA methylation: while age adjusted classifiers may offer minor refinements, they are insufficient on their own, as age does not align with the dominant axis of variation between the two clusters. Future studies should systematically test candidate covariates (e.g., BRCA1/2 status, parity, BMI, hormonal therapy), apply age regressed methylation matrices, and investigate whether the two sub-clusters encode different time-to-diagnosis intervals within the PRE-BRCA stratum, which could enable finer grained epigenetic risk scoring for women at high hereditary risk of breast cancer.

ACKNOWLEDGMENT

The authors are grateful for the Department of Science and Technology of Brazil's Ministry of Health (DECIT/MS; Departamento de Ciência e Tecnologia do Ministério da Saúde), and to the Brazilian research agencies CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and Facepe (Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco), for the partial financial support of this research.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, 2024.

[2] Y. Dong, H. Zhao, H. Li, X. Li, and S. Yang, "DNA methylation as an early diagnostic marker of cancer (review)," *Biomedical Reports*, 2014.

[3] M. Hum and A. S. G. Lee, "DNA methylation in breast cancer: early detection and biomarker discovery through current and emerging approaches," *Journal of Translational Medicine*, 2025.

[4] Z. Xu, D. P. Sandler, and J. A. Taylor, "Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the Sister Study," *JNCI: Journal of the National Cancer Institute*, 2020.

[5] L. C. Oliveira, W. P. Santos, M. A. Santana, and J. B. Araujo Neto, "Breast cancer risk detection based on DNA methylation," in *Proceedings of the XXI Congresso Brasileiro de Informática em Saúde (CBIS26)*, Brasília, Brazil, 2026.

[6] S. Horvath, "DNA methylation age of human tissues and cell types," *Genome Biology*, 2013.

[7] K. C. Oeffinger *et al.*, "Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society," *JAMA*, 2015.

[8] P. A. Dugué, M. C. Southey, J. K. Bassett *et al.*, "Biological aging measures based on blood DNA methylation and risk of cancer: a prospective study," *JNCI Cancer Spectrum*, 2018.

[9] E. Hofstatter, S. Horvath, D. Dalela *et al.*, "Increased epigenetic age in normal breast tissue from luminal breast cancer patients," *Clinical Epigenetics*, 2018.

[10] C. Bodelon, S. Ambatipudi, P. Dugué *et al.*, "Blood DNA methylation and breast cancer risk: a meta-analysis of four prospective cohort studies," *Breast Cancer Research*, 2019.

[11] S. Anjum, E.-O. Fourkala, M. Zikan *et al.*, "A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival," *Genome Medicine*, 2014.

[12] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 2008.

[13] I. Newsham, M. Sendera, S. G. Jammula, and S. A. Samarajiwa, "Early detection and diagnosis of cancer with interpretable machine learning to uncover cancer-specific DNA methylation patterns," *Biology Methods and Protocols*, 2024.

[14] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[15] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, 2002.

[16] A. E. Teschendorff and C. L. Relton, "Statistical and integrative system-level analysis of DNA methylation data," *Nature Reviews Genetics*, 2018.

[17] S. Williams, S. J. Clark, R. Pidsley, and C. Stirzaker, "DNA methylation biomarkers for the diagnosis and treatment management of breast cancer," *Epigenomics*, 2025.

Alzheimer's Enhanced Diagnosis with Convolutional Neural Networks: An Efficient Approach

B.Sc. Arthur Magalhães da Silva
Centro de Ciências e Tecnologia
Universidade Estadual do Ceará (UECE)
Fortaleza, Brasil
arthur.magalhaes.pesquisador@protonmail.com

D.Sc. Evandro Andrade
Centro de Ciências e Tecnologia
Universidade Estadual do Ceará (UECE)
Fortaleza, Brasil
evandro.andrade@aluno.uece.br

Prof. Dr. Joaquim Celestino Júnior
Centro de Ciências e Tecnologia
Universidade Estadual do Ceará (UECE)
Fortaleza, Brasil
joaquim.celestino@uece.br

Prof. Dr. Antônio de Barros Serra
Centro de Ciências e Tecnologia
Instituto Federal do Ceará (IFCE)
Fortaleza, Brasil
serra@ifce.edu.br

Renan Teixeira Leite
Centro de Ciências e Tecnologia
Universidade Estadual do Ceará (UECE)
Fortaleza, Brasil
renan.teixeira@aluno.uece.br

***Abstract*—Alzheimer's disease, a progressive neurodegenerative disorder affecting millions globally, presents significant challenges in diagnosis, particularly in its early stages. It is estimated that 35.6 million people suffer from the disease, with 1.2 million cases in Brazil alone [1]. Characterized by the accumulation of amyloid-beta plaques [2], the disease is often discovered in advanced stages due to the difficulty of its diagnosis, which directly impacts the possibilities for effective treatment. This paper investigates the potential of Convolutional Neural Networks (CNNs) as an auxiliary tool in the diagnosis of Alzheimer's disease, aiming to improve detection accuracy and earliness. By training and comparing different CNN architectures using a database of magnetic resonance imaging (MRI) scans from patients, the study seeks to identify the hyperparameters that maximize metrics such as recall, precision, and F1-score in classifying the different stages of the disease. The results demonstrate the effectiveness of applying CNNs to Alzheimer's diagnosis, highlighting the high recall achieved in identifying advanced stages. Additionally, the research contributes to the understanding of the impact of hyperparameters on CNN performance for this specific task, paving the way for the development of more precise and efficient diagnostic support tools.**

***Index Terms*—Alzheimer's, Convolutional Neural Networks, Diagnosis, Machine Learning, MRI.**

I. Introduction

The use of artificial intelligence (AI) has become increasingly present in daily life, with diverse applications that permeate multiple aspects of human existence. In healthcare, AI has driven a true revolution, making it possible to conduct preliminary diagnoses almost instantaneously. As these technologies evolve and become more accessible, the expectation is that the quality of life of populations will be significantly impacted.

Alzheimer's disease, a progressive neurodegenerative condition, represents a considerable challenge for global public health. According to Alves [1], an estimated 35.6 million people suffer from the disease, with 1.2 million in Brazil alone. Characterized by the progressive loss of cognitive functions, the disease begins with mild memory changes and gradually evolves to more severe conditions, compromising the quality of life of patients and their families [3]. Although there is still no cure for the disease, early identification is crucial for effective treatment and for allowing the patient to actively participate in their health care.

This work investigates the potential of Convolutional Neural Networks (CNNs), a powerful machine learning method, in developing tools for the rapid and accurate diagnosis of Alzheimer's disease. Using a dataset of magnetic resonance imaging (MRI) scans [4], we constructed and compared different CNN architectures to identify the hyperparameters that yield optimal performance in classifying the different stages of the disease.

The specific objectives of this study are:

- Perform a literature review on the use of computational techniques in the diagnosis of Alzheimer's disease.
- Develop and test multiple CNNs to identify potential cases of the disease from MRI images.
- Identify the hyperparameters that lead to the best CNN performance in the task of classifying Alzheimer's stages.

We believe that the results of this research will contribute significantly to the development of early diagnosis support tools for Alzheimer's disease, opening pathways for more effective interventions and improving patients' quality of life.

II. Theoretical Background

This section presents the theoretical concepts that underpin this work. First, we address Alzheimer's disease, its characteristics, and the challenges related to its diagnosis. Next, we

explore the paradigm of machine learning (ML), focusing on artificial neural networks and highlighting their application in disease diagnosis—specifically in Alzheimer's detection.

A. Alzheimer's Disease

Alzheimer's disease (ICD-10 G30) is a chronic and progressive neurodegenerative disease that causes cognitive decline, impacting work capacity, social relationships, and quality of life [2]. The most common initial symptom is short-term memory loss, followed by difficulties in learning, communication, and reasoning, which affect daily activities and interpersonal relationships.

Accounting for about 60% to 70% of dementia cases, Alzheimer's still does not have its causes fully elucidated [5]. The disease was first described in 1906 by the German psychiatrist Alois Alzheimer, and its progression, although variable, leads to a gradual loss of bodily function control, culminating in death. Life expectancy following diagnosis is generally between 3 and 9 years [6].

Alzheimer's is the most frequent form of dementia in elderly patients and is characterized by the accumulation of beta-amyloid plaques, mainly in the prefrontal and mesial-temporal areas of the brain [2]. The formation of amyloid plaques damages brain tissue, contributing to the progression of the disease.

Diagnosis is complex and requires a multidisciplinary evaluation involving careful analysis of clinical data, imaging exams, and neuropsychological tests. Frota et al. [7] describe the following core criteria for diagnosis:

- Progressive decline of cognitive abilities.
- Insidious onset (months or years).
- Exclusion of other causes of dementia through neuroimaging exams, such as computed tomography (CT) or magnetic resonance imaging (MRI).

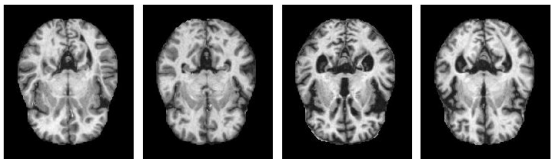


Fig. 1. Different stages of Alzheimer's disease, ordered by progression. Source: [4].

B. Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) aimed at developing computational systems capable of learning from data without being explicitly programmed [9]. The goal of ML is to create algorithms that, given a set of training data, identify patterns, make predictions, and make decisions autonomously.

C. Artificial Neural Networks

Artificial Neural Networks (ANNs), or simply neural networks, are computational models inspired by the operation of the human brain. ANNs consist of interconnected units called artificial neurons, which process and transmit information similarly to biological neurons [16].

Neurons are organized in layers: the input layer receives the input data, the output layer provides the network's result, and the hidden layers between the input and output perform information processing. Neural networks with multiple hidden layers are called deep neural networks.

1) Multilayer Perceptron: The Multilayer Perceptron (MLP) is one of the most widely used neural network architectures, consisting of an input layer, one or more hidden layers, and an output layer.

2) Bias Insertion in Neural Networks: Bias is a parameter that can be added to neurons, increasing the network's flexibility and its ability to learn complex patterns within data [18], [19].

3) Activation Functions: Activation functions are applied to the output of each neuron to introduce non-linearity into the network, enabling the modeling of complex relationships in the data. The choice of activation function depends on the problem and the characteristics of the data. Common activation functions include Rectified Linear Unit (ReLU), Hyperbolic Tangent, Logistic (Sigmoid), and Softmax [20].

4) Loss Functions: Loss functions are used to quantify the error between the neural network's output and the target output. The objective of network training is to minimize the loss function. Examples of loss functions for classification problems include Cross-Entropy and Hinge Loss [18].

5) Backpropagation: The backpropagation algorithm is essential for training neural networks. It calculates the gradient of the loss function with respect to the network weights, allowing weights to be updated in the direction that minimizes error. Figure 2 illustrates the backpropagation process.

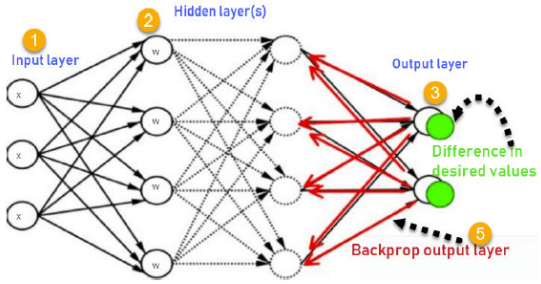


Fig. 2. Diagram of backpropagation. Source: [21].

6) Dropout: Dropout is a regularization technique designed to prevent overfitting in neural networks. During training, dropout randomly deactivates a percentage of neurons, forcing the network to learn more robust representations that are less dependent on specific individual neurons [24], [25].

7) Convolutional Neural Networks (CNNs): Convolutional Neural Networks (CNNs) are a specific type of neural network widely used in image processing. CNNs excel at automatically extracting relevant features from images without requiring manual feature engineering. The defining operations of CNNs are convolution, pooling, and flattening, as illustrated in Figure 3.

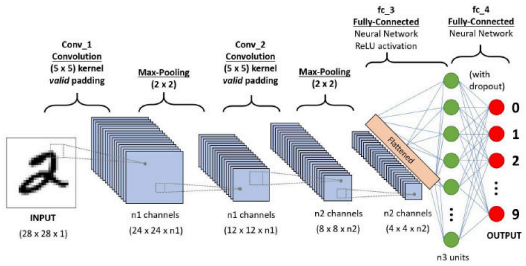


Fig. 3. Diagram of CNN operations. Source: [27].

a) Convolution: Convolution is the fundamental operation of CNNs. It applies a filter (kernel) to the input image to extract specific features, such as edges, textures, and patterns.

b) Pooling: Pooling reduces the dimensionality of feature maps generated by convolution, making the network more robust to shifts in feature positions within the image. The most common type of pooling is max pooling, which selects the maximum value within an image region [26].

c) Flattening: Flattening transforms the multidimensional matrix resulting from pooling into a one-dimensional vector, which is then used as input to a fully connected neural network.

III. Related Work

The application of Convolutional Neural Networks (CNNs) to diagnose neurodegenerative conditions such as Alzheimer's has been extensively researched over the past decade. Recent studies, such as that by *Islam and Zhang* [8], applied deep convolutional networks to structural magnetic resonance imaging (SMRI) data to classify different stages of Alzheimer's disease. The model uses convolutions with 7x7 operators over SMRI data and features four layers of 56, 28, 14, and 7 neurons, in addition to the output layer. The activation function used was *Softmax*. The achieved precisions were 99% for non-demented individuals, 75% for very mild cases, 62% for mild, and 33% for moderate cases, demonstrating that precision decreases as Alzheimer's severity increases. The architecture of this model can be seen in Figure 4.

Another significant study was conducted by *Ding et al.* [30], who utilized the *Inception V3* architecture to predict Alzheimer's progression across three stages: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Evaluation metrics such as precision and recall (or sensitivity) demonstrated superiority over traditional methods and even outpaced evaluations performed by radiologists. Radiologists outperformed the model only in the intermediate MCI stage, whereas the model outperformed human experts in AD and CN cases. The complete model architecture is shown in Figure 5.

Zhang et al. [25] also applied a deep learning approach, utilizing two data modalities: MRI and PET (Positron Emission Tomography) images. They proposed two models with nearly identical configurations applied separately to the two modalities. Pearson's correlation coefficient was used to check

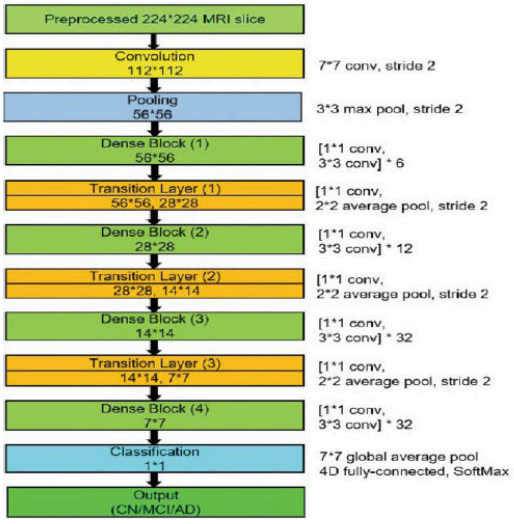


Fig. 4. Architecture proposed by *Islam and Zhang* [8].

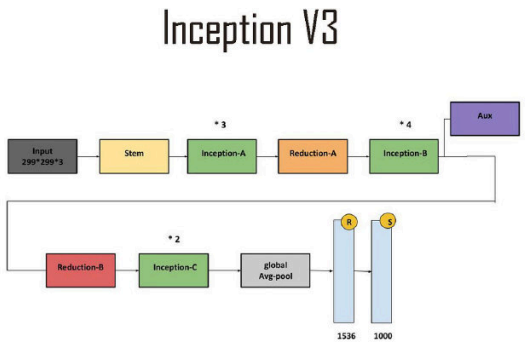


Fig. 5. *Inception V3* Architecture [30].

agreement between the two models, achieving an AUC of 0.98 for the CN vs. AD comparison, but only 0.88 for the intermediate MCI stage. The architecture used in their study is illustrated in Figure 6.

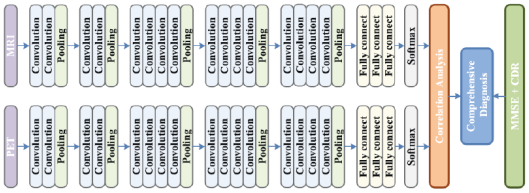


Fig. 6. Architecture proposed by *Zhang et al.* [25].

Table I summarizes the results of these works.

IV. Methodology

The study was conducted to diagnose Alzheimer's Disease using CNNs applied to brain MRI scans. The methodological process followed a series of essential steps, ranging from data acquisition to model training and evaluation.

TABLE I
SUMMARY OF RESULTS FROM RELATED WORKS

Authors	Accuracy	Class	Recall	Precision	F1-Score	AUC
(ISLAM; ZHANG, 2018)		ND	0.99	0.99	0.99	
		VMD	0.50	0.75	0.60	
		MD	0.71	0.62	0.67	
		MoD	0.50	0.33	0.40	
(DING et al., 2019)		CN	0.59	0.60	0.59	0.84
		MCI	0.54	0.55	0.55	0.52
		AD	0.81	0.76	0.78	0.98
(ZHANG et al., 2019) AD/CN	98.47%	AD	0.96			0.98
(ZHANG et al., 2019) MCI/CN	85.74%	MCI	0.90			0.88
(ZHANG et al., 2019) AD/MCI	88.20%	AD	0.97			0.88

A. Data Acquisition and Preprocessing

The data used in this study were retrieved from a repository available on Kaggle (a data science learning platform), consisting of brain MRI images of patients at different stages of Alzheimer's Disease. These images were categorized into four classes: Non Demented (ND), Very Mild Demented (VMD), Mild Demented (MD), and Moderate Demented (MoD).

A major challenge in medical image classification is class imbalance, which can lead to poor model generalization, particularly for underrepresented classes. In this work, we used the SMOTE implementation from the `imbalanced-learn` library with the parameters `strategy="not majority"` and `random_state=0`.

Prior to applying SMOTE, the dataset contained 6,394 MRIs; after oversampling, it expanded to 12,796 MRIs. The original distribution alongside the post-oversampling distribution is presented in Table II.

TABLE II
CLASS DISTRIBUTION BEFORE AND AFTER OVERSAMPLING

Class	Original (Tr./Te.)	Oversampling (Tr./Te.)
1 (ND)	2407/792	2414/785
2 (VMD)	1662/577	2384/815
3 (MD)	681/214	2385/814
4 (MoD)	47/16	2414/785

Each image underwent intensity normalization to ensure all images shared the same intensity scale, which is essential for the model to learn consistent patterns.

B. Convolutional Neural Network Architecture

Different CNN architectures were evaluated to identify the optimal configuration. All models shared the following core components:

- Input layer: resized images of 128x128 pixels.
- 1st Convolutional Layer: 8 filters of 9x9, with ReLU activation.
- 2nd Convolutional Layer: 16 filters of 5x5, with ReLU activation.
- Hidden layers (minimum of 1): ReLU activation.
- Output Layer: 4 neurons (one for each class), with Softmax activation.

All models were trained using categorical cross-entropy loss and the Adam optimizer with an initial learning rate of 0.001.

C. Model Training

All models were trained on 75% of the data and evaluated on the remaining 25%. For each, was used the same partition of training and test without cross-validation. Training was performed over 20 epochs with a batch size of 32.

D. Performance Evaluation

Metrics used to evaluate model performance included accuracy, precision, recall, and F1-score. Additionally, ROC curve plots were generated for visual analysis.

E. Hyperparameter Tuning

To optimize model performance, various hyperparameter combinations were tested, including the number of layers, neurons per layer, dropout rates, and bias inclusion. A total of 24 distinct hyperparameter configurations were evaluated.

Tested dropout rates were 0% and 20%, while the number of layers varied between 1, 3, and 10. The number of neurons per layer was set to 40 and 100. Regarding bias, models were trained both without bias and with zero-initialized bias. The loss function across all models was cross-entropy.

V. EXPERIMENTS AND RESULTS

A. Performance Evaluation

Model performance was evaluated based on accuracy, precision, recall, F1-Score, and Area Under the ROC Curve (AUC).

Minimum viable model thresholds were established in reference to [8]. A model was considered viable if it achieved a recall and AUC of 0.5 or higher for Non Demented, Very Mild Demented, and Mild Demented, and 0.99 or higher for Moderate Demented. A total of 20 models met these criteria.

To compare models, intervals with one standard deviation around the mean were analyzed across all metrics. Sixteen models fell within this interval, and none scored above it. Five of these models achieved maximum performance across all metrics for the MoD class. For these five, accuracy served as the tie-breaker; the configuration providing the highest test accuracy consisted of 3 layers with 100 neurons per layer (without dropout or bias), scoring above 0.89 across all metrics as detailed in Table III.

TABLE III
PERFORMANCE OF THE BEST MODEL ACROSS DIFFERENT DEMENTIA CLASSES

Accuracy	Class	Precision	Recall	F1-Score
95.4%	Non Demented (ND)	0.94	0.91	0.92
	Very Mild Demented (VMD)	0.89	0.94	0.91
	Mild Demented (MD)	0.98	0.95	0.97
	Moderate Demented (MoD)	1	1	1

The results demonstrate that the top-performing model delivered exceptional performance, particularly in classifying the MoD class, achieving perfect scores across all metrics. A recall and precision of 1 indicates that the model never confuses this class with others. High recall is especially critical in medical diagnostics, where false negatives can lead to severe clinical outcomes.

The high sensitivity (recall) in the MoD class indicates that the model accurately identifies patients in advanced stages of the disease. For all the models, the performance improves with the severity of the class. This observation highlights the need for refinement to enhance early detection.

B. Evaluation of Models with Different Hyperparameters

Analysis of models trained under various hyperparameter configurations revealed that:

- Networks with 10 layers generally yielded poorer results, exhibiting the highest variance in performance (standard deviation up to 0.35). Average accuracy was 76.4%, while maximum accuracy reached 87.9%.
- Applying dropout at a rate of 20% increased variance in the results. Models without dropout achieved an average accuracy that was 10% higher.
- Incorporating bias did not significantly affect performance, indicating that bias parameters can be omitted for this specific problem without sacrificing performance.

C. Analysis of ROC Curves and AUC

To further assess model performance, ROC curves were generated for each class, and Area Under the Curve (AUC) values were calculated. Figure 7 displays the plot for the top-performing model.

The model with 3 layers and 100 neurons per layer achieved exceptional performance, reaching an AUC of 1.0 for the MoD class and AUCs above 0.95 across all other classes. This demonstrates high discriminative capability across all stages, particularly in more advanced disease progression.

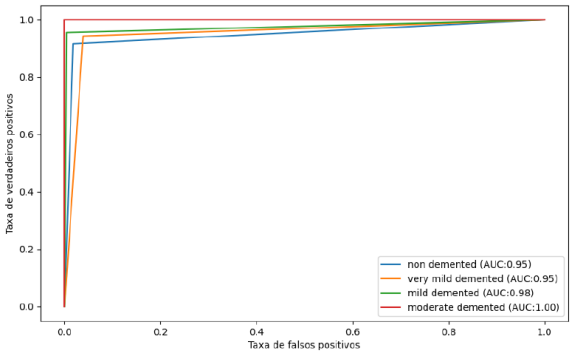


Fig. 7. ROC Curves for the model with 3 layers of 100 neurons

D. Importance of Recall and F1-Score

In medical diagnostics, such as in Alzheimer's Disease, recall is a critical metric because failing to properly diagnose a condition can carry serious consequences. Our models were evaluated with a primary focus on recall to ensure accurate identification of true positives in every class. All evaluated models achieved recall values above 0.99 for the *Moderate Demented* class, indicating a minimal false negative rate.

Furthermore, the F1-score was used to balance precision and recall, especially in scenarios where precision was lower due to class distribution skewness. As shown in Table IV, all models achieved F1-scores above 0.9 for most classes, reflecting a solid balance between capturing true positives and minimizing false positives.

TABLE IV
MODEL PERFORMANCE ON THE TEST SET

Model Config.	Acc. (%)	Prec.	Sensit.	F1.
Viable				
3 Layers, 100 Neurons	95.4	≥ 0.94	≥ 0.91	≥ 0.91
1 Layer, 40 Neurons, Bias	91.02	≥ 0.84	≥ 0.82	≥ 0.83
Other Configurations	91.02 - 95.4	≥ 0.97	≥ 0.96	≥ 0.90
Non-Viable				
10 Layers, 40 Neurons, Dropout (w/ and w/o bias)	54.5 - 59.5	≥ 0.36	≥ 0.07	≥ 0.12
Other Configurations	≥ 70	≥ 0.97	≥ 0.96	≥ 0.9

E. Comparison with Previous Studies

Our proposed model—utilizing Softmax activation and three layers of 100 neurons—presented better results than existing approaches in all metrics, including on severe disease classes. Even though the datasets were different, this performance indicates that the chosen architecture and hyperparameters are well suited for Alzheimer's stage classification.

It is worth noting that performance on milder stages was lower than on advanced stages (the opposite behavior of [8]), showing room for improvement in early detection. Techniques such as hyperparameter fine-tuning, data augmentation, and more complex architectures—such as 3D Convolutional Neural Networks (3D CNNs) [31] or incorporating multiple data sources [32]—may help enhance performance during early Alzheimer's stages.

F. Experimental Conclusions

The top-performing model consist of 1 to 3 layers with 100 neurons per layer, trained without dropout. Figure 7 displays the ROC curve for the best configuration (3 layers, 100 neurons, Softmax).

These results are promising, indicating that CNNs can serve as effective diagnostic aids for Alzheimer's Disease. However, current limitations in identifying initial stages remain. Future work should investigate data augmentation, complex network architectures, and multi-modal integration to boost early detection capabilities.

G. Limitations and Future Directions

A key limitation is the reliance solely on MRI data, which may restrict insights into early disease onset. In future work, we plan to explore multimodal frameworks combining MRI with PET scans and genetic markers to increase diagnostic accuracy.

The strong unbalecement on the dataset presents a challenge for the confiability in its performance. A revalidation of the architeture with more diverse datasets is necessary.

Another promising direction is integrating attention mechanisms into CNNs, combining CNNs with Recurrent Neural Networks (RNNs), or leveraging Vision Transformers to capture temporal and fine-grained spatial patterns. Transfer learning techniques using pre-trained medical imaging models could also improve performance on underrepresented classes.

Finally, ensemble learning techniques combining multiple model architectures could enhance predictive robustness, particularly for the earliest stages of Alzheimer's Disease.

VI. Conclusion and Future Work

While our results are promising, several areas remain for improvement. Future research can explore advanced architectures, such as attention mechanisms or transformer-based networks, to enhance feature discrimination between subtle Alzheimer's stages.

Additionally, automated hyperparameter tuning via grid search or Bayesian optimization could yield further performance improvements. Evaluating the model on larger, more diverse datasets will also help validate generalization across varying patient populations and imaging modalities.

This study demonstrates that Convolutional Neural Networks (CNNs) are a powerful tool for assisting in Alzheimer's diagnosis, particularly during advanced stages. The proposed model achieved an overall accuracy of 95.4%, with notable performance on the MoD class, reaching a recall of 1. These outcomes show high reliability in identifying advanced patients, which is critical in clinical settings to prevent false negatives and ensure timely care. Challenges in early detection likely stem from subtle structural changes in early dementia that overlap with normal brain aging patterns.

References

[1] Alves, M. (2020). *21 de setembro - Dia Mundial da Doença de Alzheimer e Dia Nacional de Conscientização da Doença de Alzheimer*.

[2] Castellazzi, G., Cuzzoni, M. G., Ramusino, M. C., Martinelli, D., Denaro, F., Ricciardi, A., Vitali, P., Anzalone, N., Bernini, S., Palesi, F. et al. "A machine learning approach for the differential diagnosis of alzheimer and vascular dementia fed by MRI selected features." *Frontiers in neuroinformatics*, vol. 14, pp. 25, 2020.

[3] Jiwtode, U., Chakole, S., & Bhatt, N. "Alzheimer's disease: History, stages, diagnosis and its future." *Journal of Pharmaceutical Research International*, vol. 33, no. 39A, pp. 41–45, 2021.

[4] S. Dubey, "Alzheimer's dataset (4 class of images)," Kaggle, 2019.

[5] Todd, S., Barr, S., Roberts, M., & Passmore, A. P. "Survival in dementia and predictors of mortality: a review." *International journal of geriatric psychiatry*, vol. 28, no. 11, pp. 1109–1124, 2013.

[6] Hippius, H. & Neundörfer, G. "The discovery of alzheimer's disease." *Dialogues in clinical neuroscience*, 2022.

[7] Frota, N. A. F., Nitrini, R., Damasceno, B. P., Forlenza, O., Diastosta, E., Silva, A. B. da, Junior, E. H., & Magaldi, R. M. "Critérios para o diagnóstico de doença de alzheimer." *Dementia & Neuropsychologia*, vol. 5, no. 1, pp. 5–10, 2011.

[8] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain informatics*, vol. 5, no. 2, pp. 1–14, 2018.

[9] Samuel, A. L. "Some studies in machine learning using the game of checkers." *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[10] Rezende, S. O., Monard, M. C., & Carvalho, A. C. P. d. L. F. "Sistemas inteligentes para engenharia: pesquisa e desenvolvimento." *Anais*, 1999.

[11] Hurb. "Machine learning na veia." *hurb.labs*, 2017.

[12] Labs, M. "Everything you need to know about neural networks." (2017).

[13] Sharma, R. "What are machine learning applications? top 10 industry and real-world use cases." (2022).

[14] Hafeez, K. & Ahmed, Q. "Applications of machine learning in education and health sector: An empirical study." *Journal of Software Engineering and Intelligent Systems*, vol. 3, 2019.

[15] Vasques, L. H. *Aplicação de Técnicas de machine learning a diagnósticos na saúde*. (2020).

[16] Fitch, F. B. "Warren s. mcculloch and walter pitts. a logical calculus of the ideas immanent in nervous activity. bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133." *The Journal of Symbolic Logic*, vol. 9, no. 2, pp. 49–50, 1944.

[17] Xu, G. & Oates, W. S. "Adaptive hyperparameter updating for training restricted boltzmann machines on quantum annealers." *Scientific Reports*, vol. 11, no. 1, p. 2727, 2021.

[18] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. "A survey of convolutional neural networks: analysis, applications, and prospects." *IEEE transactions on neural networks and learning systems*, 2021.

[19] Nielsen, M. A. *Neural networks and deep learning*. 1st ed. Determination Press, 2015.

[20] Sharma, S., Sharma, S., & Athaiya, A. "Activation functions in neural networks." *Towards Data Sci*, vol. 6, no. 12, pp. 310–316, 2017.

[21] Johnson, D. "Back propagation in Neural Network: Machine Learning Algorithm." (2023).

[22] Branco, H. *Overfitting e underfitting em machine learning - ABRACD - associação brasileira de ciência de dados*. (2023).

[23] Brownlee, J. "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks." *Machine Learning Mastery*, 2019.

[24] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. "Dropout: A simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. Disponível em: http://jmlr.org/papers/v15/srivastava14a.html

[25] Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., & Zhang, X. "Multimodal deep learning model for auxiliary diagnosis of Alzheimer's disease." *Neurocomputing*, vol. 361, pp. 185–195, 2019.

[26] Brownlee, J. "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks." *Machine Learning Mastery*, 2019.

[27] Saha, S. "A comprehensive guide to Convolutional Neural Networks - the eli5 way." (2023).

[28] Riebesell, J. "Convolution operator." (2022).

[29] Nussbaum, J. M., Seward, M. E., & Bloom, G. S. "Alzheimer disease: a tale of two prions." *Prion*, vol. 7, no. 1, pp. 14–19, 2013.

[30] Y. Ding et al., "A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[31] Yagis, E., Citi, L., Diciotti, S., Marzi, C., Atnafu, S. W., & Herrera, A. G. S., "3D convolutional neural networks for diagnosis of Alzheimer's disease via structural MRI," In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 65–70, IEEE, 2020.

[32] Liu, Y., Cheng, M., Zhang, F., & Du, H., "MSCNet: A multimodal attention-based deep learning model for Alzheimer's disease diagnosis," *Journal of Alzheimer's Disease*, vol. 87, no. 3, pp. 903-915, 2022.

An AI-Enabled Digital Healthcare Workflow: Integrating Deep Learning and Large Language Models for Interpretable Wound Triage

Ernane Ferreira Rocha Junior
Dept. of Computer Engineering and Automation
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
ernane.ferreira.100@ufrn.edu.br

Ignacio Sánchez-Gendriz
Northeast Strategic Technologies Center
CETENE / MCTI
Recife, PE, Brazil
ignacio.gendriz@cetene.gov.br

Luiz Affonso Guedes
Dept. of Computer Engineering and Automation
Federal University of Rio Grande do Norte (UFRN)
Natal, RN, Brazil
affonso@dca.ufrn.br

***Abstract*—Accurate chronic wound triage is critical in clinical settings, yet the adoption of deep learning in this domain is often limited by the opaque nature of AI models and the absence of deployable engineering workflows. This paper presents an end-to-end Artificial Intelligence-assisted Clinical Decision Support System (AI-CDSS) that integrates Convolutional Neural Networks (CNNs) with Large Language Models (LLMs) for interpretable wound triage. Operating solely on visual data to maximize point-of-care usability, the vision component employs a VGG16 backbone optimized through transfer learning and selective fine-tuning to classify images into six categories. The fine-tuned model achieved a global accuracy of 74.79%, with class-wise analysis showing that fine-tuning increased venous ulcer recall from 87.09% to 91.94% at the cost of a corresponding decline in pressure ulcer recall to 23.53%, a trade-off addressed through a permanent, always-visible safety disclosure rather than a dynamic confidence threshold. The classifier is embedded in a scalable, asynchronous microservices architecture. The system employs a generative LLM as a Narrative Interpretability layer, translating probabilistic outputs and patient metadata into structured clinical pre-reports that explicitly surface model uncertainties to support clinician decision-making.**

***Index Terms*—Digital Healthcare, Narrative Interpretability, Clinical Decision Support System, Chronic Wound Triage, Convolutional Neural Networks, Large Language Models.**

I. Introduction

Accurate and timely assessment of chronic wounds, including diabetic foot, pressure, venous, and surgical ulcers, remains a persistent challenge in clinical practice due to their high prevalence and economic burden [1], [2]. Wound evaluation relies heavily on visual inspection, which is time-consuming and subject to inter-observer variability. Delays or misclassification may lead to suboptimal interventions and increased burden on patients and healthcare systems [3].

The expansion of telemedicine has increased demand for scalable digital workflows that support triage and documentation while preserving traceability and interpretability [4], [5]. CNNs have achieved competitive performance in medical image classification and are increasingly explored for wound classification [6]–[8].

Moving from a standalone classifier to a clinically usable solution requires more than predictive performance. A digital healthcare system must support asynchronous processing, uncertainty communication, and clinician-centered outputs [9]. LLMs have been explored for natural-language summarization and report drafting in healthcare, but their use in decision support must be framed with explicit safeguards and separation from clinical judgment [10]–[12].

This work proposes an end-to-end AI-CDSS Proof of Concept (PoC) for digital health. To maximize point-of-care usability, a vision-only approach is adopted with a VGG16 backbone [13] trained via Transfer Learning and selective Fine-Tuning to classify wound images into six categories. The classifier is integrated into a modular, asynchronous microservices architecture (Figure 1) connected to a generative LLM [14] acting as a Narrative Interpretability layer, producing structured clinical pre-reports that explicitly state model uncertainties.

The primary contribution is the design, implementation, and evaluation of this platform, with methodological emphasis on a two-stage training strategy: Transfer Learning with pre-trained ImageNet weights, followed by targeted Fine-Tuning of deep convolutional blocks [13]. The class-wise trade-offs introduced by fine-tuning are empirically evaluated, and the end-to-end architecture is shown to compensate for algorithmic limitations through the generative interpretability layer.

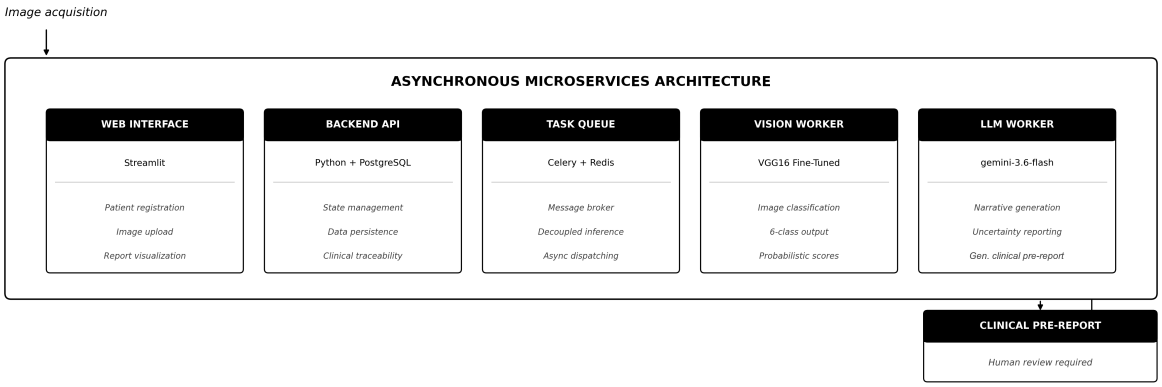


Fig. 1. Overview of the proposed AI-CDSS workflow, from visual data ingestion to LLM-generated pre-reports.

II. Background and Related Work

A. CNNs for Ulcer Image Classification

CNNs have become the standard approach for medical image classification due to their ability to learn hierarchical visual representations from data [6]. In wound care, CNNs have been applied to classify chronic wounds from photographs captured under heterogeneous conditions typical of real-world clinical and telemedicine settings [7], [8], [15].

Anisuzzaman et al. [8] and Patel et al. [15] showed that fusing image data with body location coordinates improves classification accuracy for chronic wounds. Although effective, manually mapping spatial coordinates for each lesion introduces operational friction in fast-paced point-of-care workflows. This work diverges by adopting a vision-only approach, relying on targeted model optimization to compensate for the absence of spatial metadata. Per-class metrics including recall, precision, and F1-score are used to identify clinically relevant failure modes [16], and uncertainty communication is treated as a system design requirement [17].

B. Transfer Learning and Fine-Tuning

Transfer learning reduces data requirements by adapting representations learned from large-scale natural image datasets to the target medical domain [18], [19]. VGG16 remains a common baseline due to its straightforward architecture, facilitating reproducible comparisons in resource-constrained settings [13]. Fine-tuning introduces a trade-off between domain specialization and overfitting; class-wise analysis is recommended to assess whether performance gains are generalizable or reflect redistribution across classes [20], [21].

C. CDSS and LLMs in Health Applications

CDSS are grouped into knowledge-based and AI-based systems [5], [11]. In both cases, effective adoption depends on how outputs are communicated and how uncertainty is managed [9]. LLMs have been explored as a Narrative Interpretability layer downstream of predictive models, translating structured evidence into clinician-facing narratives that reduce reporting burden while preserving traceability [12]. Concerns about hallucination and over-reliance motivate designs that constrain LLMs to reporting tasks with mandatory human-in-the-loop review [10], [22].

III. Methodology

A. Dataset and Preprocessing

This study used the public dataset "Multi-modal wound classification using images and locations" [8], [15], collected at the AZH Wound and Vascular Center. Spatial metadata was excluded to implement a vision-only workflow suitable for mobile point-of-care use [8]. The dataset is publicly available and anonymized, complying with ethical standards for secondary data analysis.

Images were captured with an iPad Pro and a Canon digital camera, presenting real-world variability in scale, illumination, and background. The six-class taxonomy and the exact per-class distribution are described in Table I. The dataset was partitioned into 696 training images and 234 test images following the original split [8]. The class distribution is imbalanced: the most represented class (V, 185 training images) outnumbers the least represented pathological class (P, 100 training images) by a factor of 1.85, which we revisit in Section IV as a contributing factor to the class-wise recall gap observed after fine-tuning. Data augmentation including rotations ($\pm 20°$), width and height shifts (20%), shear distortions (20%), zoom variations (20%), and horizontal flipping was applied to mitigate overfitting and improve feature generalization [18], [20].

TABLE I
Image classes and per-class sample counts.

Class	Meaning	Train	Test
D	Diabetic Foot Ulcer	139	46
P	Pressure Ulcer	100	34
S	Surgical Wound	122	42
V	Venous Leg Ulcer	185	62
N	Normal Skin	75	25
BG	Background	75	25
Total		**696**	**234**

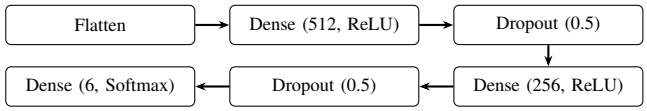


Fig. 2. Custom classification head appended to the VGG16 backbone after `block5_pool`.

B. Model Architecture and Two-Stage Training

VGG16 [13] was selected as the CNN backbone for its established performance and role as a reproducible baseline in medical imaging [6]. The original ImageNet classification head was replaced by a custom head appended after the last max-pooling layer (`block5_pool`), as illustrated in Figure 2.

Training proceeded in two stages. In the first stage (Transfer Learning), the convolutional base was frozen and the custom head was trained using the Adam optimizer with learning rate $lr = 1 \times 10^{-4}$, categorical cross-entropy loss, and Early Stopping with patience of 10 epochs. In the second stage (Fine-Tuning), the highest-level convolutional block (`block5`) was unfrozen and the network was retrained with a reduced learning rate ($lr = 1 \times 10^{-5}$) and stricter Early Stopping (patience of 5 epochs), allowing the filters to specialize in wound-specific visual textures while avoiding catastrophic forgetting of the pre-trained weights.

In both stages, Early Stopping and checkpoint selection monitored loss and accuracy computed on the 234-image test partition, since the original protocol did not carve out a separate internal validation split from the training data. We flag this explicitly as a methodological limitation: because the same partition used for model selection is also used for the final reported metrics, the selected checkpoint's test accuracy is likely optimistic relative to a protocol with an independent validation split. A corrected protocol, holding out a stratified subset of the 696 training images for validation while leaving the test set untouched until final evaluation, is planned as a follow-up; we do not expect it to change the qualitative class-wise trends discussed in Section IV, though it may lower the reported point accuracy.

C. AI-CDSS System Architecture

The platform is orchestrated via Docker Compose across four containerized services sharing a common bridge network. The `backend` directory is not a standalone service; it is a shared Python module mounted as a volume into both the `streamlit` and `worker` containers, providing common access to the classification model, database operations, and LLM integration. A shared `bucket` volume provides persistent file storage for uploaded images and generated PDF reports. Table II describes each service.

The relational schema (Figure 3) maintains strict links between patient records, image assets, and AI-generated outputs. Every pre-report is therefore auditable and tied to the specific model checkpoint used during inference. Once the Celery worker computes the classification probabilities, the distribution is packaged with patient metadata and passed to a structured prompting pipeline. The generative component is Google's `gemini-3.6-flash` [23], invoked through the `google-genai` SDK with default generation parameters (no explicit temperature or top-p override). The prompt template supplies the predicted class, the full probability vector, and the model's historical precision, recall, and F1-score for that class, and instructs the LLM to produce a report of no more than 300 words. The LLM produces a clinical draft that states model uncertainties explicitly, presents secondary differential diagnoses, and is formatted as a preliminary summary explicitly labeled for clinician verification prior to any clinical use [10].

TABLE II
DOCKER COMPOSE SERVICES OF THE AI-CDSS ARCHITECTURE.

Service	Image / Stack	Role
`streamlit`	Streamlit	Web interface for patient registration, image upload, and report visualization.
`worker`	Celery	Asynchronous worker that loads the fine-tuned VGG16 model, executes image classification, invokes the LLM to produce the clinical pre-report.
`postgres`	PostgreSQL	Relational persistence of patient records, image metadata, classification results, and generated pre-reports.
`redis`	Redis	In-memory message broker connecting the web interface to the asynchronous worker.

Because the deployed system persists real patient identifiers, clinical history, and images, unlike the pre-anonymized public dataset used for model training, it falls under Brazil's General Data Protection Law (LGPD) [24]. The present proof of concept does not yet implement encryption at rest, authentication, or an anonymization pipeline for stored records; hardening the platform to LGPD requirements, including encrypted storage, role-based access control, audit logging, and a data-retention policy, is a prerequisite for any use beyond controlled evaluation and is prioritized as future engineering work.

The complete source code is publicly available.[1]

[1] https://github.com/ErnaneJ/wound-ai-cdss

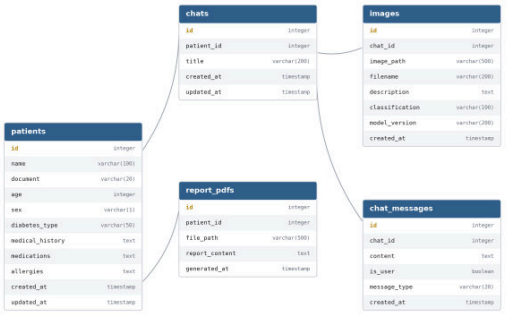


Fig. 3. Entity-relationship diagram of the AI-CDSS, illustrating the relational structure for clinical data persistence and auditability.

IV. Results and Discussion

A. Incremental Model Performance

The vision engine was developed across three phases. Table III summarizes the performance gains at each stage.

Phase 1 (Epoch 50): Transfer Learning established a baseline at 72.22% accuracy. The learning curves in Figure 4 show stable convergence but a persistent gap between training and validation loss, indicating the classification head had not reached a stable plateau. The confusion matrix in Figure 5 shows high accuracy for non-pathological classes (BG and N), but a significant overlap between Pressure (P) and Venous (V) ulcers, suggesting the frozen convolutional base was insufficient to capture the fine-grained textural differences between these etiologies.

Phase 2 (Epoch 94): Extending training with a weight restoration protocol improved accuracy to 74.79%, as shown in Figure 6. The confusion matrix in Figure 7 reflects improved recall for Diabetic (D) and Surgical (S) classes, but class P recall remained at 29.41%, confirming that the P/V confusion is a structural limitation of the frozen feature extractor and justifying the subsequent fine-tuning stage.

Phase 3 (Fine-Tuning): Unfreezing `block5` matched the extended transfer-learning accuracy (74.79%) while substantially redistributing recall across classes. The F1-score comparison in Figure 8 shows a substantial gain for class V, whose recall reached 91.94%. The final confusion matrix in Figure 9 confirms that venous ulcer triage reached high reliability. Class P recall decreased to 23.53%, a consequence of the model specializing in the dominant class distribution. Global accuracy alone therefore does not capture the effect of fine-tuning; the class-wise trade-off, gaining venous sensitivity at the cost of pressure-ulcer sensitivity, is the actual outcome of this stage and is treated accordingly throughout this section. Notably, class P recall declines monotonically across all three phases (38.24% at Epoch 50, computed from Figure 5; 29.41% at Epoch 94; 23.53% after fine-tuning), indicating that the sensitivity loss reflects progressive specialization toward the dominant class distribution rather than an isolated effect of the fine-tuning stage.

TABLE III
Performance Gains Across Training Stages.

Stage	Epoch	Accuracy	Recall (V)
Initial Baseline	50	72.22%	87.09%
Extended TL	94	74.79%	87.09%
Fine-Tuning	13 (+94)	74.79%	91.94%

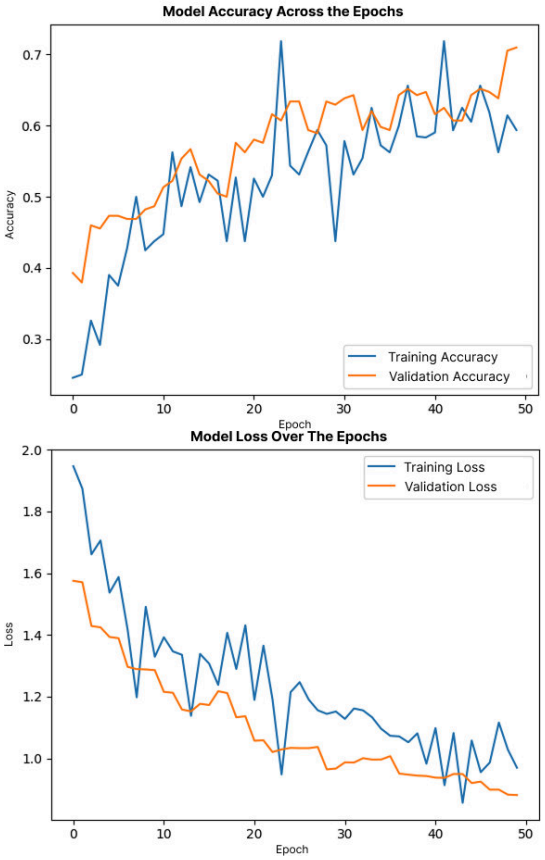


Fig. 4. Learning curves, Epochs 1–50. The gap between training and validation loss indicates room for further optimization.

We treat this reduction in class P recall as a patient-safety limitation, since pressure injuries are themselves a patient-safety indicator subject to prevention protocols and adverse-event reporting in Brazilian clinical services [3], not merely as an optimization trade-off. The current single-split protocol cannot reliably calibrate a per-prediction confidence threshold (Section IV-B), so the Narrative Interpretability layer instead embeds a static, always-visible disclosure of the model's documented class P sensitivity (recall of 23.53%) into every generated pre-report. Clinicians are alerted to this weakness regardless of the classifier's output for a given image, favoring a transparent, permanently visible safety disclosure over an uncalibrated dynamic trigger [17]. This disclosure complements, but does not substitute, validated structured risk-assessment instruments such as the Braden and Norton scales, and it does not replace institutional pressure-injury prevention protocols:

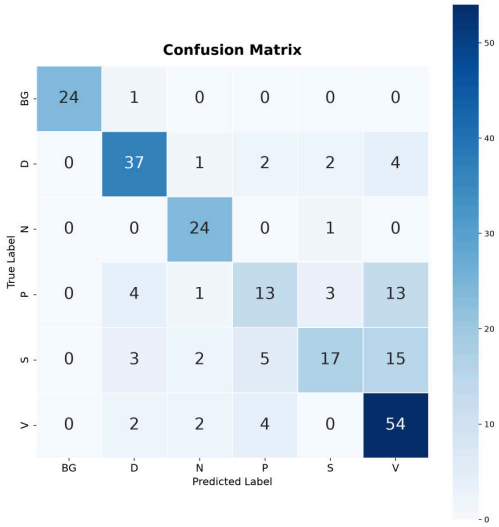


Fig. 5. Confusion matrix at Epoch 50. High accuracy for non-pathological classes, with notable overlap between P and V ulcers.

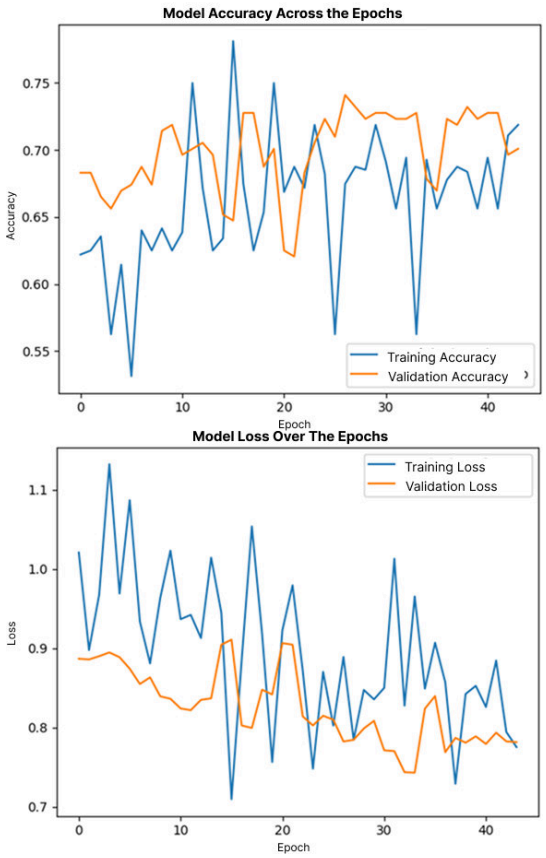


Fig. 6. Learning curves for the continuation phase (44 additional epochs, cumulative total of 94). Stable convergence achieved through automated weight restoration.

the classifier output is meant to prompt attention, while risk stratification and preventive care decisions remain governed by established clinical scales.

The global accuracy of 74.79% achieved by the fine-tuned model corresponds to a 95% Wilson confidence interval of [68.9%, 79.9%] ($n = 234$), reflecting the estimation uncertainty inherent to a test set of this size. This result is consistent with the literature under comparable constraints, though direct comparison is limited by differing preprocessing and data splits. Anisuzzaman et al. [8] reported 75.64% accuracy using whole images alone and 82.48% when fusing image and body-location data. Patel et al. [15], using a region-of-interest (ROI) cropping step and a 70/15/15 split, reported 85.41% accuracy without location data and 87.50% when adding location, a result driven in large part by ROI cropping rather than multi-modality alone. Relative to the whole-image, single-modality figure from Anisuzzaman et al., our result is comparable; the gap to the multimodal figures (7.7 to 12.7 percentage points) reflects both the cost of excluding spatial metadata and, in Patel et al.'s case, the absence of ROI cropping, which we did not adopt to preserve a simple, single-stage pipeline for point-of-care deployment. In this work, spatial metadata was intentionally excluded to keep the workflow viable at the point of care. The venous ulcer recall of 91.94% reflects the clinical priority assigned to that class, and the permanent, metrics-based safety disclosure for pressure ulcers addresses the sensitivity limitations that a global accuracy figure alone would not reveal.

B. Model Calibration

Clinical deployment requires calibrated confidence, not only point accuracy: predicted probabilities should reflect the true likelihood of correctness. We assessed calibration with the Expected Calibration Error (ECE) [25], using 10 equal-width confidence bins over an independent re-evaluation of the released model checkpoint on the 234-image test partition. This re-evaluation reproduced the 74.79% headline accuracy (175/234) reported in Table III, confirming that inference on the released checkpoint is deterministic given a fixed, non-shuffled test partition. Because Early Stopping and checkpoint selection are themselves single-run decisions (Section III), the reported point accuracy still carries the run-to-run variance of the training process, not of inference; repeating training across multiple random seeds and reporting mean $\pm$ standard deviation, rather than a single point estimate, is planned as future work.

The resulting ECE was 0.052, with a mean predicted confidence of 69.6% against an observed accuracy of 74.8%: the model is, on average, mildly underconfident rather than overconfident (Figure 10). Because the current protocol lacks a validation split independent from the test set (Section III), this ECE is reported as a diagnostic computed on the same data used for evaluation; we do not apply post-hoc temperature scaling here, since fitting a temperature parameter on the test set itself would reintroduce the same leakage. Recalibration on a proper held-out calibration split, obtained once the validation protocol described in Section III is corrected, is proposed as future work.

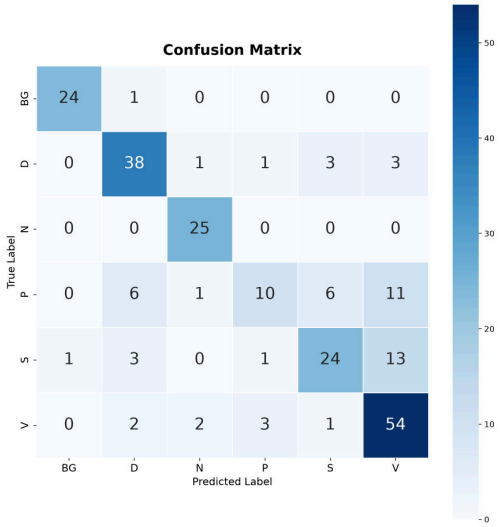


Fig. 7. Confusion matrix at Epoch 94 (74.79% accuracy). Class P recall remains at 29.41%.

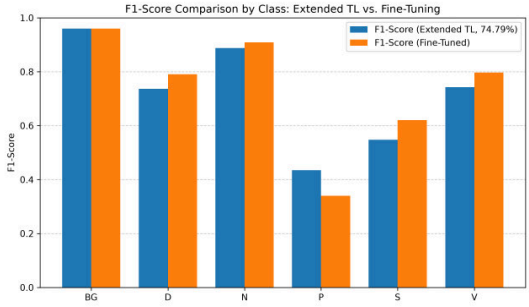


Fig. 8. Class-wise F1-score comparison: Extended Transfer Learning versus Fine-Tuned model. Both stages reach 74.79% global accuracy; the gain from fine-tuning is a redistribution of recall toward class V, not a change in the aggregate figure.

C. Operational Validation

Figures 11 and 12 illustrate the operational interface using a synthetic patient record created solely for demonstration; no real patient data is shown. Figure 11 shows the patient registration interface, which allows structured clinical data entry alongside high-resolution image upload at the point of care.

Figure 12 shows the decision support interface. The system presents the CNN's classification output translated by the LLM into a narrative clinical pre-report directly in the chat: the primary predicted class and confidence, a differential discussion of the leading secondary candidate classes drawn from the full probability vector, and the model's documented class-wise limitations, including the reduced sensitivity for pressure ulcers, which are always appended to the report regardless of the current prediction. The AI is positioned as a documentation assistant, not a diagnostic authority; the pre-report is a preliminary, AI-generated draft intended for manual clinician review and is not a substitute for clinical judgment.

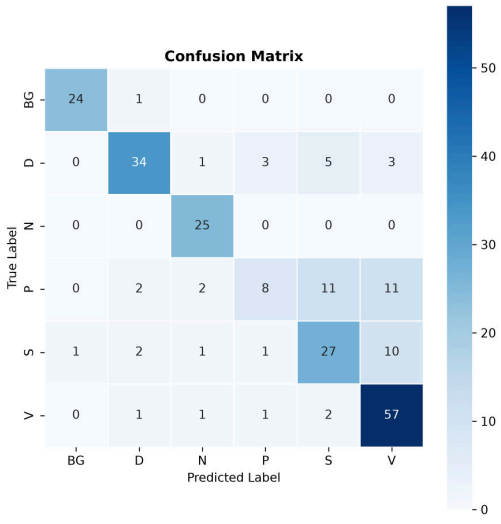


Fig. 9. Final fine-tuned confusion matrix. Venous (V) class recall reaches 91.94%.

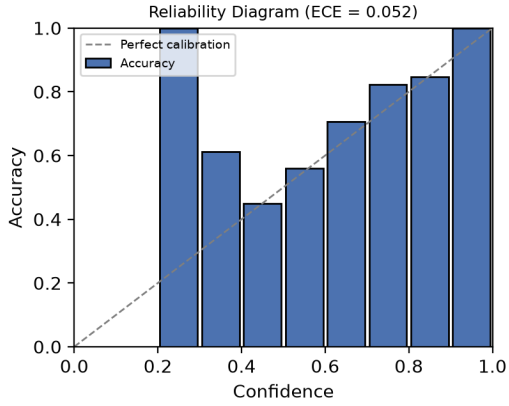


Fig. 10. Reliability diagram for the re-evaluated fine-tuned model (ECE = 0.052). Bars show observed accuracy per confidence bin; the dashed line marks perfect calibration.

V. Conclusion

This paper presented a scalable and interpretable digital healthcare workflow for chronic wound triage, designed for the resource-constrained, point-of-care settings common in Brazilian public health services, where anatomical location metadata and specialized equipment are often unavailable. The proposed AI-CDSS integrates a VGG16 vision engine, an asynchronous microservices architecture, and a generative LLM interpretability layer. It delivers practical value under the constraints of a limited and imbalanced medical dataset and treats uncertainty communication as a first-class design requirement, not an afterthought.

The two-stage training strategy yielded a peak accuracy of 74.79% (95% CI [68.9%, 79.9%]) and a venous ulcer recall of 91.94% after fine-tuning, as detailed in the class-wise analysis. The analysis also exposed the trade-offs introduced by domain specialization: global accuracy metrics can

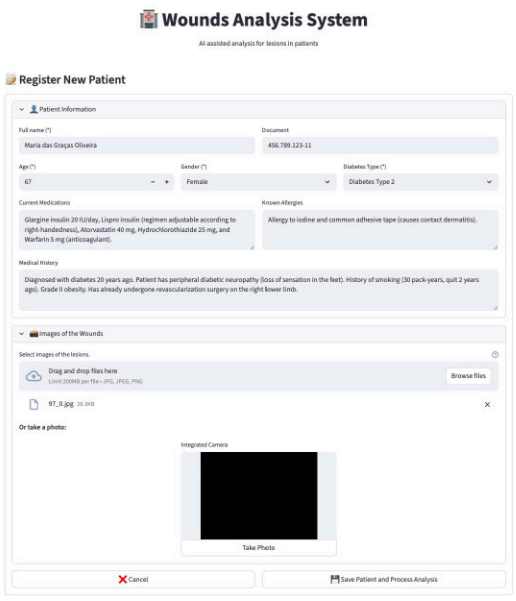


Fig. 11. Patient registration form with structured clinical data entry and image upload.

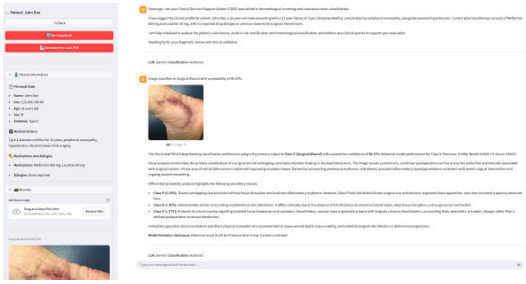


Fig. 12. Decision support interface: CNN probabilities translated into a narrative clinical pre-report by the LLM, including the always-visible pressure-ulcer sensitivity disclosure.

obscure class-specific performance drops, and the resulting drop in pressure ulcer recall is itself a patient-safety limitation, not merely an algorithmic one. The Narrative Interpretability layer addresses this architecturally, embedding a permanent, metrics-based safety disclosure of the classifier's known class-wise weaknesses into every generated report, so that model limitations are communicated instead of concealed.

From an engineering perspective, the containerized microservices architecture operationalizes the CNN predictions through an asynchronous pipeline that decouples inference from the clinical interface. The LLM is constrained to reporting tasks, and every generated pre-report is explicitly labeled as a preliminary, AI-generated draft that must be independently verified by a qualified clinician before any use in patient care. This preserves clinical accountability while reducing documentation burden.

The study has the following limitations: (i) the vision-only approach restricts resolution of etiologies that depend on anatomical context; (ii) the dataset size and imbalance limit the statistical generalization of the model, particularly for underrepresented classes such as pressure ulcers; (iii) Early Stopping and checkpoint selection relied on the test partition itself rather than an independent validation split, likely biasing the reported point accuracy optimistically (Section III); (iv) the LLM-generated pre-reports have not undergone formal medical evaluation for hallucination risks or semantic safety; (v) the deployed prototype does not yet implement the encryption, access control, and anonymization safeguards required for LGPD-compliant production use (Section III). Future work will address these through a corrected train/validation/test protocol with multi-seed repetition, class-imbalance handling (class weighting, focal loss, or resampling), calibration on a proper held-out split, lightweight backbones (e.g., MobileNet, EfficientNet-Lite) for point-of-care deployment on low-resource hardware, a small clinician-based evaluation of pre-report fidelity and safety, dataset expansion, integration of selective anatomical metadata, and LGPD-oriented hardening of the production system.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] C. K. Sen *et al.*, "Human skin wounds: a major and snowballing threat to public health and the economy," *Wound Repair and Regeneration*, vol. 17, pp. 763–771, 2009.
[2] J. F. Guest *et al.*, "Health economic burden that wounds impose on the national health service in the uk," *BMJ Open*, vol. 5, p. e009283, 2015.
[3] EPUAP, NPIAP, PPPIA, "Prevention and treatment of pressure ulcers/injuries: Clinical practice guideline," 2019, third edition.
[4] H. Tchero *et al.*, "Telemedicine in wound care: a review," *International Wound Journal*, 2018.
[5] R. T. Sutton *et al.*, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, 2020.
[6] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
[7] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.
[8] D. M. Anisuzzaman *et al.*, "Multi-modal wound classification using wound image and location by deep neural network," *Scientific Reports*, vol. 12, no. 1, 2022.
[9] J. L. Kwan *et al.*, "Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials," *BMJ*, p. m3216, 2020.
[10] World Health Organization, "Ethics and governance of artificial intelligence for health: Who guidance," 2021. [Online]. Available: https://iris.who.int/handle/10665/341996
[11] J. C. L. Ong *et al.*, "Large language model as clinical decision support system augments medication safety in 16 clinical specialties," *Cell Reports Medicine*, vol. 6, p. 102323, 2025.
[12] N. H. Shah *et al.*, "Creation and adoption of large language models in medicine," *JAMA*, 2023.
[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2014.
[14] Gemini Team, Google, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
[15] Y. Patel *et al.*, "Integrated image and location analysis for wound classification: a deep learning approach," *Scientific Reports*, vol. 14, no. 1, 2024.
[16] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proc. 23rd ICML*, 2006, pp. 233–240.

[17] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, 2017.
[18] J. Yosinski *et al.*, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014.
[19] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.
[20] D. Sculley *et al.*, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, 2015.
[21] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, 2015.
[22] E. Tabassi, "Artificial intelligence risk management framework (ai rmf 1.0)," NIST, Tech. Rep. NIST AI 100-1, 2023.
[23] Google DeepMind, "Gemini 3.6 flash model card," 2026. [Online]. Available: https://deepmind.google/models/model-cards/gemini-3-6-flash/
[24] Presidência da República, Brasil, "Lei geral de proteção de dados pessoais (lgpd), lei nº 13.709/2018," 2018.
[25] C. Guo *et al.*, "On calibration of modern neural networks," in *Proc. 34th ICML*, 2017.

Aplicativo Móvel para Classificação de Eletrocardiogramas Utilizando Redes Neurais Convolucionais

Murilo Brasil Cordeiro
Curso de Engenharia de Computação, Depto. de Informática (DAINF) & Depto. de Eletrônica (DAELN)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Paraná, Brasil
murilobrasil@alunos.utfpr.edu.br

Marcio Allan de Souza Alves
Secretaria da Saúde do Paraná & Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial (CPGEI)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Paraná, Brasil
allanalves2@gmail.com

Leandra Ulbricht
Programa de Pós-graduação em Engenharia Biomédica (PPGEB) & Depto. Eletrônica (DAELN)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Paraná, Brasil
prof.leandra@gmail.com

Amauri Amorin Assef
Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial (CPGEI) & Depto. de Eletrotécnica (DAELT)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Paraná, Brasil
amauriassef@utfpr.edu.br

Joaquim Miguel Maia
Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial (CPGEI) & Depto. de Eletrômoca (DAELN)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Paraná, Brasil
joaquim@utfpr.edu.br

***Resumo*—As doenças cardiovasculares figuram entre as principais causas de mortalidade no mundo, o que torna fundamental o desenvolvimento de ferramentas capazes de auxiliar na identificação precoce de alterações cardíacas. Este trabalho apresenta o desenvolvimento de um aplicativo móvel para classificação automática de imagens de eletrocardiogramas (ECG) utilizando redes neurais convolucionais (CNNs). O sistema foi desenvolvido com o objetivo de fornecer suporte à tomada de decisão em cenários de emergência, permitindo a execução local do modelo em dispositivos móveis. Para o treinamento da rede foi utilizada uma base de imagens de ECG composta por quatro classes: normal, arritmia cardíaca, histórico de infarto do miocárdio e infarto agudo do miocárdio. Técnicas de pré-processamento, aumento artificial de dados e balanceamento de classes foram empregadas para melhorar a capacidade de generalização do modelo. A arquitetura foi implementada utilizando TensorFlow e posteriormente convertida para TensorFlow Lite, permitindo sua execução local no dispositivo móvel. No conjunto de validação utilizado, foram obtidas taxas de acerto de 98,3% para infarto agudo do miocárdio e superiores a 86% para as demais classes. Os resultados demonstram a viabilidade técnica da integração do modelo à aplicação móvel, embora ainda sejam necessárias avaliações com dados independentes e em condições reais de captura antes de considerar sua utilização em contexto clínico.**

***Index Terms*—eletrocardiograma, redes neurais convolucionais, aprendizado profundo, aplicativo móvel, classificação de eletrocardiogramas**

I. Introdução

As doenças cardiovasculares permanecem entre as principais causas de mortalidade no mundo, respondendo por milhões de óbitos anualmente. Dentre essas doenças, o Infarto Agudo do Miocárdio (IAM) destaca-se pela necessidade de diagnóstico rápido e início precoce do tratamento, fatores diretamente relacionados à redução da mortalidade e das sequelas associadas à condição. Diretrizes do Ministério da Saúde ressaltam que a realização precoce do eletrocardiograma e a identificação de alterações compatíveis com IAM são etapas fundamentais para a tomada de decisão clínica e o encaminhamento adequado do paciente [1].

O eletrocardiograma (ECG) é o exame inicial e mais comum para avaliar a atividade elétrica cardíaca e identificar alterações associadas a diferentes patologias. O sinal de ECG é composto por ondas e segmentos que representam eventos específicos do ciclo cardíaco, permitindo a identificação de anormalidades por meio da análise de sua morfologia. A onda P está associada à despolarização dos átrios; o complexo QRS representa a despolarização ventricular; e a onda T corresponde à repolarização dos ventrículos. Além dessas estruturas, intervalos como PR, QT e RR fornecem informações importantes sobre a condução elétrica e a regularidade do

ritmo cardíaco [2].

A Figura 1 apresenta os principais componentes de um traçado eletrocardiográfico normal. A análise conjunta dessas ondas e segmentos permite identificar alterações associadas a diferentes condições clínicas, fornecendo informações importantes para o diagnóstico médico.

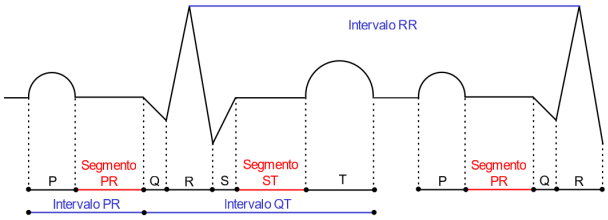


Figura 1: Principais ondas, segmentos e intervalos presentes em um eletrocardiograma normal.

Alterações em regiões específicas do ECG podem indicar diferentes patologias cardíacas. No caso do IAM, é comum observar alterações no segmento ST, como elevação ou depressão, indicativas de comprometimento do músculo cardíaco [3], [4]. Arritmias cardíacas podem ser identificadas por irregularidades no ritmo e na morfologia das ondas, enquanto pacientes com histórico de infarto podem apresentar alterações persistentes no traçado mesmo na ausência de um evento agudo.

Apesar de sua ampla utilização, a interpretação de exames de ECG depende de conhecimento especializado e de experiência clínica. Nos últimos anos, técnicas de aprendizado profundo têm apresentado resultados promissores na análise automatizada de sinais biomédicos. Em particular, redes neurais convolucionais (CNNs) demonstraram elevada capacidade de extrair automaticamente características relevantes a partir de imagens médicas, alcançando desempenho comparável ao de especialistas em diferentes tarefas de classificação [5], [6].

Paralelamente, o crescimento das tecnologias móveis e das soluções de *mobile health* tem possibilitado a incorporação de técnicas de inteligência artificial em ferramentas voltadas ao suporte clínico. Essas soluções permitem realizar análises diretamente no local de atendimento, reduzindo a dependência de infraestrutura externa e ampliando o acesso a recursos computacionais em ambientes com conectividade limitada.

Diversos trabalhos têm explorado técnicas de aprendizado profundo para a classificação automática de eletrocardiogramas. Entre eles, destaca-se o trabalho de Khan et al. [7], que utilizou a base de imagens de ECG disponibilizada em [8] para classificar exames em quatro categorias: normal, arritmia cardíaca, histórico de infarto e infarto agudo do miocárdio. Os autores empregaram uma arquitetura baseada em SSD MobileNet V2, obtendo uma acurácia média próxima de 98% na classificação das imagens. O estudo demonstrou o potencial das redes neurais profundas para identificação automática de alterações cardíacas a partir de imagens de ECG.

Outra referência importante é o trabalho de Hannun et al. [5], que utilizou redes neurais profundas para classificação de arritmias cardíacas a partir de sinais de ECG ambulatoriais, alcançando desempenho comparável ao de cardiologistas especialistas. De forma semelhante, Ribeiro et al. [6] desenvolveram um sistema baseado em aprendizado profundo para diagnóstico automático de eletrocardiogramas de 12 derivações, obtendo resultados expressivos em diferentes categorias de doenças cardíacas.

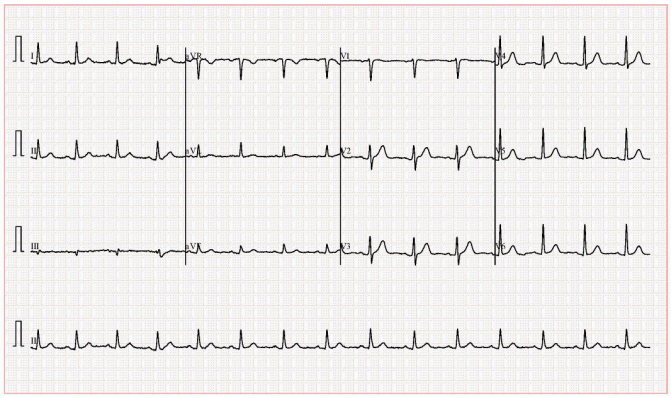
Diante desse contexto, este trabalho propõe o desenvolvimento e a avaliação técnica de uma aplicação móvel para classificação automática de imagens de ECG utilizando uma rede neural convolucional. O modelo é executado localmente no dispositivo por meio do TensorFlow Lite, permitindo seu funcionamento sem dependência de conexão com a Internet. Além das classes relacionadas aos exames cardíacos, foi incorporada uma classe denominada *Not ECG*, destinada à identificação de imagens que não correspondem a eletrocardiogramas. O estudo concentra-se na viabilidade técnica da integração entre o modelo de classificação e a aplicação móvel, não abrangendo, nesta etapa, sua validação para utilização clínica.

II. Materiais e Métodos

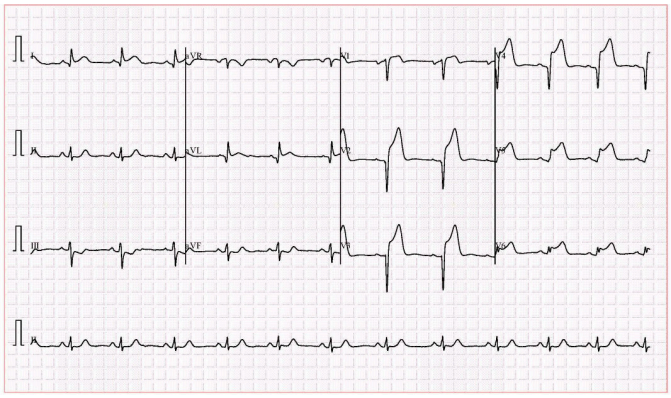
A. Base de Dados

A base de dados utilizada foi obtida a partir do conjunto de imagens de ECG disponibilizado por Khan et al. [8]. O conjunto original é composto por cinco categorias: ECG normal, arritmia cardíaca, histórico de infarto do miocárdio, infarto agudo do miocárdio e casos relacionados à COVID-19. As Figuras 2a, 2b, 2c e 2d apresentam exemplos das classes utilizadas neste trabalho. Os exames correspondem a ECGs de 12 derivações, contendo informações provenientes das derivações dos membros e precordiais. Embora as alterações associadas às diferentes patologias se manifestem em derivações específicas, o modelo proposto utiliza a imagem completa do exame como entrada, extraindo características globais do traçado. As categorias apresentadas representam, respectivamente, exames normais, casos de infarto agudo do miocárdio, arritmias cardíacas e pacientes com histórico de infarto.

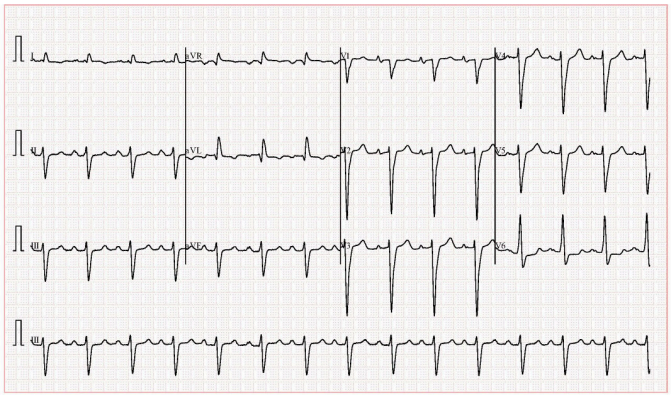
Como o objetivo deste trabalho está voltado à classificação das principais categorias cardíacas presentes na base e à integração do modelo a uma aplicação móvel, a classe referente à COVID-19 não foi utilizada. Em contrapartida, foi adicionada uma quinta categoria denominada *Não é ECG*, composta por imagens que não representam eletrocardiogramas válidos. Essas imagens foram obtidas a partir do conjunto Imagenette [9], utilizado para compor exemplos negativos durante o treinamento. A inclusão dessa classe tem como objetivo aumentar a robustez da aplicação em situações reais, evitando que o sistema tente classificar fotografias de objetos, documentos ou superfícies como exames cardíacos. Para compor essa categoria, foram utilizadas aproximadamente 1500 imagens, selecionadas aleatoriamente a partir de um conjunto de cerca de 14000 imagens.



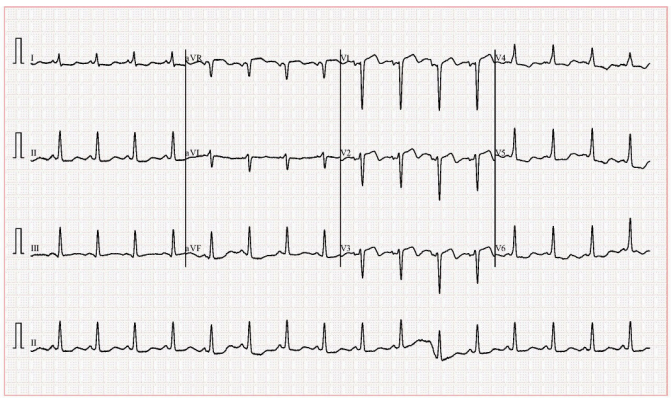
(a) Exemplo de ECG normal.



(b) Exemplo de infarto agudo do miocárdio.



(c) Exemplo de arritmia cardíaca.



(d) Exemplo de histórico de infarto do miocárdio.

Figura 2: Exemplos de diferentes traçados de eletrocardiogramas analisados.

Após as etapas de balanceamento e aumento de dados, o conjunto final utilizado no treinamento passou a contar com aproximadamente 7309 imagens distribuídas entre as cinco classes consideradas neste trabalho, conforme apresentado na Tabela I.

Tabela I: Distribuição de imagens por classe após as etapas de balanceamento e aumento de dados

Classe	Original	Final
Normal	500	1500
Infarto Agudo do Miocárdio	74	1480
Histórico de Infarto	154	1386
Arritmia Cardíaca	481	1443
Não é ECG	1500	1500
Total		7309

B. Aumento de Dados e Balanceamento

A distribuição original das classes apresentava forte desbalanceamento, especialmente no caso de infarto agudo do miocárdio. Para reduzir esse problema, foi implementada uma estratégia de aumento adaptativo de dados, na qual cada classe recebeu um número diferente de transformações sintéticas, de acordo com sua representatividade.

As imagens da classe normal foram ampliadas aproximadamente três vezes, enquanto as de infarto agudo receberam até vinte variações artificiais por amostra. As classes de histórico de infarto e arritmia também receberam fatores de aumento específicos. Essa estratégia permitiu obter uma distribuição mais equilibrada entre as categorias sem descartar informações relevantes das classes majoritárias.

Além do aumento de dados, foi realizada uma etapa de curadoria manual do conjunto de imagens. Exames que apresentavam borrões severos, ruídos excessivos ou artefatos que comprometiam a visualização das ondas cardíacas foram removidos do treinamento, o que contribuiu para melhorar a qualidade das características aprendidas pela rede neural.

C. Arquitetura da Rede Neural

Para a etapa de classificação foi desenvolvida uma rede neural convolucional profunda composta por cinco blocos convolucionais sequenciais. Os blocos utilizam camadas convolucionais com função de ativação ReLU e camadas de normalização (*Batch Normalization*). Nos quatro primeiros blocos são empregadas operações de *Max Pooling* e *Dropout*, enquanto o último bloco é seguido por uma camada *Global Average Pooling*.

A quantidade de filtros aumenta progressivamente ao longo da arquitetura, variando de 32 a 512, permitindo a extração hierárquica de características cada vez mais complexas a partir das imagens de ECG. Após a etapa de extração de características, a arquitetura utiliza uma camada Global Average Pooling, seguida de duas camadas densas, ambas regularizadas por penalização L2. Essa técnica adiciona um termo de penalização à função de perda durante o treinamento, desestimulando a atribuição de pesos excessivamente elevados aos neurônios da rede. Como resultado, reduz-se o risco de sobreajuste (*overfitting*) e melhora-se a capacidade de

generalização do modelo para dados não vistos anteriormente. A camada de saída utiliza a função Softmax para classificar as cinco categorias clínicas consideradas neste estudo. A arquitetura da rede neural convolucional descrita é ilustrada na Figura 3.

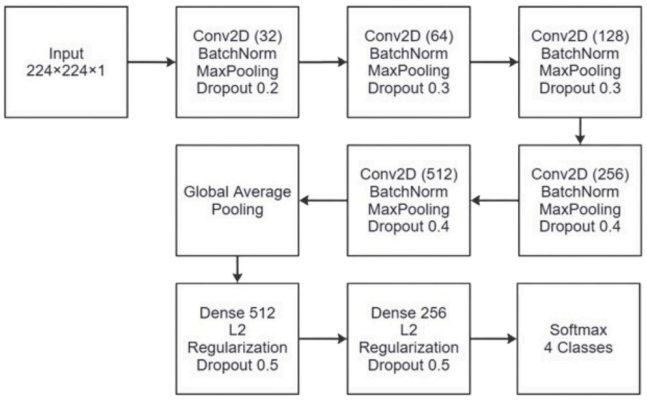


Figura 3: Arquitetura da rede neural convolucional utilizada para a classificação de eletrocardiogramas.

D. Treinamento

Após o carregamento e o pré-processamento das imagens, foi aplicado aumento de dados (*data augmentation*) de forma adaptativa às classes de ECG, com o objetivo de reduzir o desbalanceamento entre as categorias. Foram utilizadas transformações aleatórias envolvendo rotação, perspectiva, escala, brilho, contraste, adição de ruído e deslocamento horizontal. A classe *Not ECG* não foi submetida ao processo de aumento de dados.

Após a geração das imagens aumentadas, o conjunto resultante foi dividido de forma estratificada em 80% para treinamento e 20% para validação. Dessa forma, o procedimento utilizado nesta etapa do trabalho não realizou a separação dos conjuntos antes do aumento de dados. Essa característica constitui uma limitação do experimento, uma vez que versões derivadas de uma mesma imagem original podem estar presentes nos conjuntos de treinamento e validação, podendo produzir estimativas otimistas do desempenho do modelo.

As transformações de aumento de dados incluíram rotações aleatórias entre $-10°$ e $+10°$, pequenas alterações de perspectiva e escala, variações de brilho e contraste, adição de ruído gaussiano e deslocamentos horizontais. As transformações foram aplicadas de maneira probabilística, permitindo a geração de diferentes variações de uma mesma imagem.

O treinamento foi realizado utilizando o otimizador Adam, com taxa de aprendizado inicial de $0{,}0001$, *weight decay* de $0{,}0001$ e tamanho de lote (*batch size*) igual a 16. Foi estabelecido um limite máximo de 80 épocas, utilizando como função de perda a entropia cruzada categórica esparsa (*sparse categorical cross-entropy*).

Para controlar o treinamento, foi utilizado *Early Stopping* com monitoramento da perda de validação e paciência de 15 épocas, restaurando os pesos correspondentes ao melhor resultado observado. Também foi empregado *ReduceLROnPlateau*, reduzindo a taxa de aprendizado pela metade após sete épocas sem melhoria na perda de validação, até o limite mínimo de 10^{-7}.

O modelo foi desenvolvido em Python utilizando a biblioteca TensorFlow [10], responsável pela definição da arquitetura da rede neural, treinamento e avaliação do modelo. Os experimentos foram executados em um computador equipado com um processador AMD Ryzen 5 5600G (3,90 GHz), 16 GB de memória RAM e uma placa gráfica NVIDIA GeForce RTX 4060 Ti. O processamento em GPU foi utilizado para acelerar as etapas de treinamento da CNN.

E. Aplicativo Móvel

Após o treinamento, o modelo foi convertido para o formato TensorFlow Lite, permitindo sua execução em dispositivos móveis com menor consumo de recursos computacionais. A aplicação foi desenvolvida com React Native e TypeScript, o que permitiu a criação de uma solução multiplataforma. Dessa forma, o aplicativo pode ser executado em dispositivos Android e iOS a partir de uma única base de código. A interface permite a captura de imagens por meio da câmera do dispositivo ou a seleção de imagens previamente armazenados na galeria.

O processamento é realizado integralmente no dispositivo, sem necessidade de comunicação com servidores externos. Após a análise, o aplicativo apresenta a classe predita e as probabilidades associadas a cada categoria. Além do módulo de classificação de ECG, o aplicativo incorpora funcionalidades relacionadas ao protocolo de administração de tenecteplase, auxiliando profissionais de saúde no atendimento a pacientes com IAM.

F. Ferramentas de Desenvolvimento

Durante o desenvolvimento, o modelo Claude Sonnet 4.6 foi utilizado exclusivamente como ferramenta de apoio à programação, auxiliando na implementação e depuração do código, sem participação no processo de classificação ou avaliação dos ECGs.

III. Resultados e Discussão

A Figura 4 apresenta a matriz de confusão obtida durante a validação do modelo. Os resultados demonstram a viabilidade técnica da execução local do modelo em uma aplicação móvel, embora ainda sejam necessárias avaliações adicionais em dados independentes e em condições reais de captura.

Tabela II: Acurácia por classe obtida pelo modelo

Classe	Acurácia
Normal	90,6%
Infarto Agudo do Miocárdio	98,3%
Histórico de Infarto	91,2%
Arritmia Cardíaca	86,5%
Não é ECG	100,0%

A Tabela II apresenta os resultados obtidos para cada classe. A classe *Not ECG* apresentou 100% de acerto no

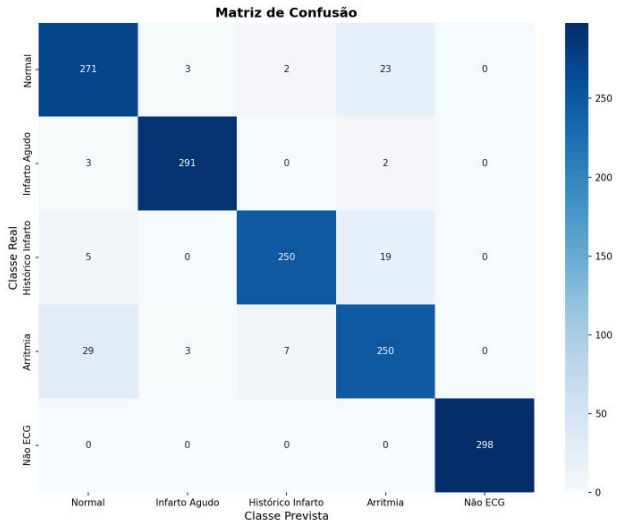


Figura 4: Matriz de confusão obtida durante a validação do modelo.

conjunto de validação utilizado. Entretanto, esse resultado deve ser interpretado considerando que as imagens negativas utilizadas foram provenientes de uma base de imagens genéricas, apresentando diferenças visuais consideráveis em relação aos ECGs. Dessa forma, o resultado demonstra a capacidade de separação entre as categorias avaliadas, mas não permite concluir que o modelo seja robusto diante de entradas inválidas mais próximas das condições reais de uso, como ECGs desfocados, parcialmente enquadrados, fotografados com baixa iluminação ou outros documentos médicos visualmente semelhantes.

A classe de infarto agudo do miocárdio apresentou 98,3% de acerto no conjunto de validação utilizado. Embora alterações morfológicas, como modificações no segmento ST, sejam características clinicamente associadas ao infarto agudo do miocárdio, não é possível afirmar que essas regiões tenham sido determinantes para as classificações realizadas pela CNN. O modelo processa a imagem completa do ECG e, neste estudo, não foram aplicadas técnicas de explicabilidade capazes de identificar quais regiões da imagem contribuíram para sua decisão.

A classe de arritmias cardíacas apresentou a menor taxa de acerto (86,5%). Esse comportamento pode ser explicado pela elevada heterogeneidade dessa categoria, que abrange diferentes tipos de alterações do ritmo cardíaco, como fibrilação atrial, extrassístoles, bloqueios de ramo e taquicardias. Como essas condições apresentam morfologias distintas entre si, a variabilidade intraclasse torna o processo de aprendizado mais complexo quando comparado às demais categorias, que possuem padrões visuais mais homogêneos.

A Figura 5 apresenta a interface da aplicação desenvolvida. A integração do modelo convertido para TensorFlow Lite permitiu realizar toda a inferência localmente, eliminando a dependência de servidores externos e reduzindo os requisitos de conectividade.

Apesar dos resultados obtidos, uma limitação importante do modelo está relacionada à forma como os exames são processados. A CNN recebe como entrada a imagem completa do eletrocardiograma redimensionada para 224×224 pixels em escala de cinza, tratando o exame como um padrão visual único. Dessa forma, o modelo não segmenta nem analisa individualmente as 12 derivações, extraindo características globais presentes em todo o traçado. Essa abordagem permite realizar a classificação a partir de características globais da imagem, porém não permite determinar quais derivações contribuíram para a decisão do modelo nem inferir a localização anatômica do infarto.

Os resultados deste trabalho apresentaram taxas de acerto de 90,6% para ECGs normais, 98,3% para infarto agudo do miocárdio, 91,2% para histórico de infarto, 86,5% para arritmias cardíacas e 100% para a categoria *Não é ECG*. Embora os resultados possam ser contrastados com os reportados por Khan et al. [7], uma comparação direta deve ser realizada com cautela, devido às diferenças na arquitetura utilizada, no protocolo experimental, no processamento dos dados e à inclusão da categoria adicional *Not ECG*.

Em relação ao trabalho de Khan et al., o presente estudo utiliza a mesma base principal de imagens de ECG, mas possui como foco a integração do modelo de classificação a uma aplicação móvel com inferência local. Além disso, foi incorporada a classe *Not ECG*, destinada à identificação de imagens que não correspondem a eletrocardiogramas. Assim, o principal diferencial deste trabalho está na integração do modelo ao ambiente móvel e na inclusão dessa categoria adicional, e não na proposição de uma nova arquitetura de CNN.

Nesta etapa, a aplicação foi avaliada apenas com imagens padronizadas provenientes da base utilizada no experimento. Não foram realizados testes em condições reais de captura pela câmera do dispositivo. Portanto, os resultados apresentados não contemplam possíveis variações de iluminação, perspectiva, foco, resolução, enquadramento ou diferenças entre câmeras de dispositivos móveis.

IV. Conclusão

Este trabalho apresentou o desenvolvimento e a integração de uma rede neural convolucional a uma aplicação móvel para classificação de imagens de eletrocardiogramas. A conversão do modelo para TensorFlow Lite possibilitou a realização da inferência diretamente no dispositivo móvel, sem necessidade de conexão com serviços externos. Além das quatro categorias relacionadas aos ECGs, foi incorporada a classe *Not ECG*, destinada à identificação de imagens que não correspondem a eletrocardiogramas.

Nos dados utilizados para validação, foram observados diferentes desempenhos entre as classes, incluindo 98,3% de acerto para infarto agudo do miocárdio e 100% para a classe *Not ECG*. Entretanto, esses resultados devem ser interpretados considerando as limitações metodológicas do estudo, especialmente a realização da divisão entre treinamento e validação

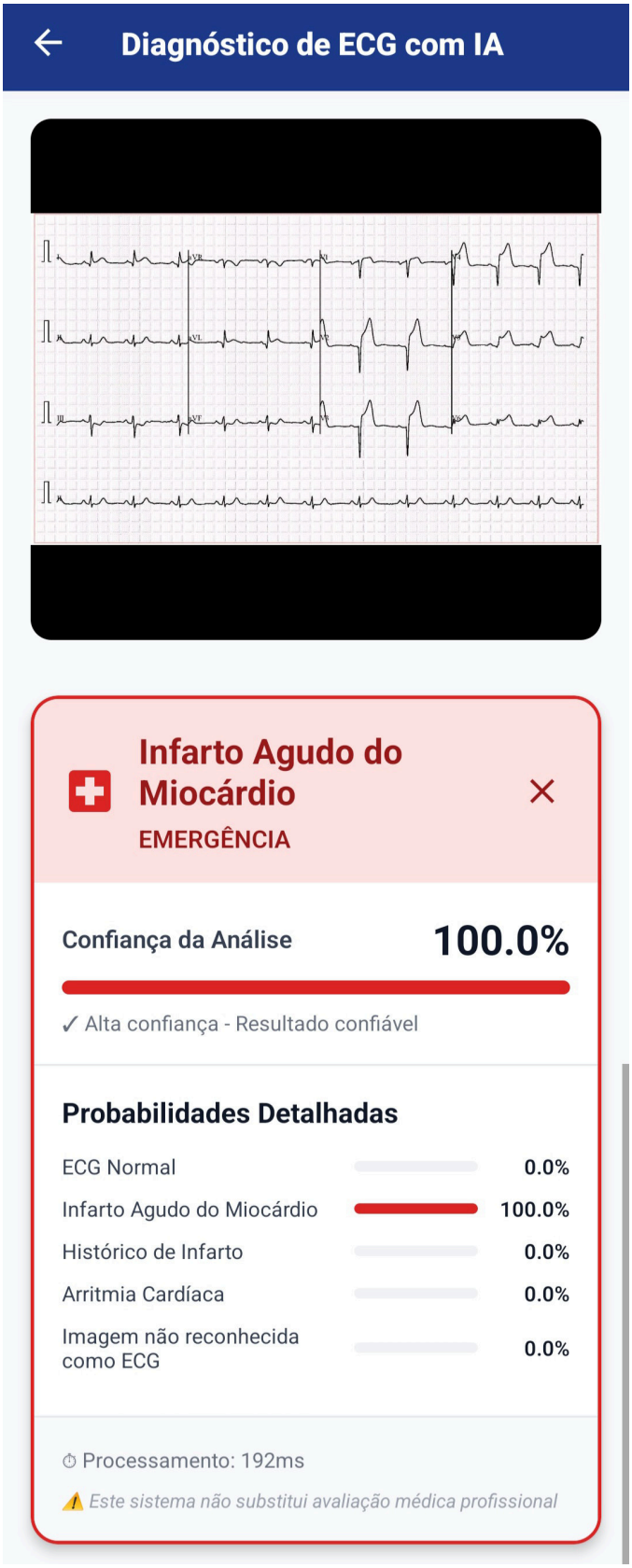


Figura 5: Interface do aplicativo desenvolvido para classificação de ECG.

após o processo de aumento de dados, o que pode resultar na presença de imagens derivadas de uma mesma amostra original em ambos os conjuntos e produzir estimativas otimistas de desempenho.

Além disso, a aplicação foi avaliada somente com imagens padronizadas, não tendo sido realizados testes em condições reais de captura ou validação clínica. Dessa forma, os resultados obtidos demonstram principalmente a viabilidade técnica da execução local do modelo e de sua integração à aplicação móvel, não sendo suficientes para estabelecer sua utilização como ferramenta de diagnóstico ou apoio à decisão clínica.

Como trabalhos futuros, pretende-se realizar a separação dos conjuntos antes da aplicação do aumento de dados, além de utilizar conjuntos independentes para teste. Também deverão ser investigadas métricas adicionais e o comportamento do sistema em diferentes dispositivos e condições de captura, incluindo variações de iluminação, perspectiva, resolução, enquadramento e qualidade das imagens. Adicionalmente, pretende-se investigar estratégias para a análise individual das derivações do ECG, possibilitando avaliar quais derivações apresentam alterações e sua relação com diferentes territórios cardíacos.

Agradecimentos

Este trabalho foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Código de Financiamento 001. Os autores também agradecem o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Fundação Araucária, da Secretaria de Estado de Ciência, Tecnologia e Ensino Superior do Paraná (SETI), da Secretaria de Estado da Saúde do Paraná (SESA) e da Universidade Tecnológica Federal do Paraná (UTFPR).

Referências

[1] Ministério da Saúde, "Manejo inicial - Infarto Agudo do Miocárdio (IAM)", Linha de Cuidado do Infarto Agudo do Miocárdio, 2022. Disponível em: https://linhasdecuidado.saude.gov.br/portal/infarto-agudo-do-miocardio/servico-de-atendimento-movel/manejo-inicial/. Acesso em: 16 jun. 2026.

[2] Sattar Y, Chhabra L. Electrocardiogram. [Updated 2023 Jun 5]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK549803/

[3] J. Kim, B. Shon, S. Kim, J. Cho, J.-J. Seo, S. Y. Jang, and S. Jeong, "ECG Data Analysis to Determine ST-Segment Elevation Myocardial Infarction and Infarction Territory Type: An Integrative Approach of Artificial Intelligence and Clinical Guidelines," *Frontiers in Physiology*, vol. 15, 2024, doi: 10.3389/fphys.2024.1462847.

[4] N. Samesima, E. G. God, J. C. L. Kruse, M. G. Leal, C. Pinho, F. F. A. C. França, J. Pimenta, et al., "Diretriz da Sociedade Brasileira de Cardiologia sobre a Análise e Emissão de Laudos Eletrocardiográficos – 2022," *Arquivos Brasileiros de Cardiologia*, vol. 119, no. 4, pp. 638–680, 2022, doi: 10.36660/abc.20220623.

[5] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019, doi: 10.1038/s41591-018-0268-3.

[6] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, M. B. Wagner, and A. L. P. Ribeiro, "Automatic Diagnosis of the 12-Lead ECG Using a Deep Neural Network," *Nature Communications*, vol. 11, no. 1, Art. no. 1760, 2020, doi: 10.1038/s41467-020-15432-4.

[7] Khan, Ali Haider, Hussain, Muzammil, Malik, Muhammad Kamran, Cardiac Disorder Classification by Electrocardiogram Sensing Using Deep Neural Network, Complexity, 2021, 5512243, 8 pages, 2021. https://doi.org/10.1155/2021/5512243

[8] A. H. Khan, M. Hussain, and M. K. Malik, "ECG Images Dataset of Cardiac and COVID-19 Patients," *Data in Brief*, vol. 34, Art. no. 106762, 2021, doi: 10.1016/j.dib.2021.106762.

[9] J. Howard, "Imagenette: A Smaller Subset of 10 Easily Classified Classes from ImageNet," FastAI, 2019. Disponível em: https://github.com/fastai/imagenette. Acesso em: 16 mai. 2026.

[10] TensorFlow, "An End-to-End Open Source Machine Learning Platform," 2025. Available: https://www.tensorflow.org. Accessed: May 10, 2026.

Apoio ao Rastreio do Câncer de Mama com Aprendizado de Máquina e Seleção de Atributos em Ultrassonografia Modo-B

Bruno Henrique Aragão Konig[1]
Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco
Recife, PE
bruno.konig@ufpe.br

Vitória Albuquerque Menezes[2]
Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco
Recife, PE
vitória.amenezes@ufpe.br

Nicole Leise Andrade Serra[3]
Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco
Recife, PE
nicole.leise@ufpe.br

Sandrelly Mirelly da Silva Santos[4]
Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco
Recife, PE
sandrelly.santos@ufpe.br

Wellington Pinheiro dos Santos[5]
Departamento de Engenharia Biomédica da Universidade Federal de Pernambuco
Recife, PE
welington.santos@ufpe.br

***Resumo*— O câncer de mama é a neoplasia de maior incidência e taxa de mortalidade entre as mulheres mundialmente. O diagnóstico precoce desempenha um papel determinante para a elevação dos índices de cura, os quais podem alcançar até 95% quando a lesão é identificada em estágios iniciais. Este trabalho propõe e avalia um sistema de auxílio ao diagnóstico médico (CAD) baseado em Aprendizado de Máquina para a detecção precoce do câncer de mama a partir de atributos extraídos de imagens de ultrassonografia mamária em modo-B (base Lenet_ultrassonografia.arff, contendo 879 instâncias e 501 atributos numéricos). Avaliou-se o impacto da redução de dimensionalidade comparando o Algoritmo Genético (AG) e a Otimização por Enxame de Partículas (PSO). Em seguida, testou-se um conjunto diversificado de classificadores supervisionados, incluindo Bayes Net, Naive Bayes, Support Vector Machine (SMO/SVM com kernels linear e polinomiais) e algoritmos de Árvores de Decisão (J48, Random Tree e Random Forest variando de 100 a 500 árvores). Para mitigar o desbalanceamento das classes, aplicou-se a técnica SMOTE sobre o conjunto de treinamento. Os experimentos foram executados sob validação cruzada 10-fold com 5 repetições (50 execuções). A combinação do Algoritmo Genético (255 atributos selecionados) com o classificador Random Forest de 200 árvores destacou-se como a arquitetura mais equilibrada entre as avaliadas, obtendo Acurácia de 87,90% ± 1,60%, Índice Kappa de 0,722 ± 0,077, Sensibilidade de 87,4% ± 4,65%, Especificidade de 87,4% ± 5,26% e Área sob a Curva ROC (AUC) de 0,970 ± 0,016. Na etapa de validação em conjunto de teste independente, o modelo alcançou uma sensibilidade de 91,6% (79 acertos em 81 lesões malignas) e especificidade de 87,4%. Conclui-se que a abordagem proposta oferece elevada acurácia diagnóstica aliada a um baixo custo computacional, sugerindo evidências preliminares para sua viabilidade técnica como um modelo de suporte a decisão em ambientes de triagem oncológica.**

Palavras-chave— Câncer de mama, Ultrassonografia modo-B, Aprendizado de Máquina, Seleção de atributos, Algoritmo Genético, Particle Swarm Optimization, Random Forest, Engenharia Biomédica.

I. INTRODUÇÃO

O câncer de mama representa uma das neoplasias malignas de maior relevância para a saúde pública global, figurando como a principal causa de mortalidade por câncer na população feminina mundial [1]. Segundo estimativas da Organização Mundial da Saúde (OMS), projeta-se que a doença seja responsável por aproximadamente 715 mil óbitos ao ano no marco de 2025 [2]. Do ponto de vista anátomo-patológico, a condição caracteriza-se pela proliferação celular descontrolada e atípica do tecido epitelial mamário, com elevado potencial de invasão local e disseminação metastática para sítios distantes através das vias linfática e hematogênica.

A despeito desse cenário epidemiológico alarmante, a literatura clínica evidencia de forma consensual que o diagnóstico precoce constitui o fator de maior impacto no prognóstico da paciente. Quando a lesão tumoral é identificada e caracterizada em suas fases iniciais (Estágios I e II), as taxas de cura e sobrevida em cinco anos chegam a atingir 95% [2]. Mesmo em tumores do tipo invasivo, mais susceptíveis a complicações metastáticas, o diagnóstico tempestivo possibilita intervenções terapêuticas cirúrgicas e farmacológicas menos invasivas, mais eficazes e com menor taxa de morbidade [2].

No âmbito dos exames de imagem médica, embora a mamografia digital permaneça como a modalidade de rastreio

populacional padrão-ouro, ela apresenta limitações bem documentadas, especialmente na avaliação de mamas densas — cenário no qual a sobreposição de tecidos reduz significativamente a sensibilidade do exame — além da desvantagem inerente da exposição à radiação ionizante [1]. Em contrapartida, a ultrassonografia mamária em modo-B firma-se como uma modalidade diagnóstica não invasiva, altamente acessível, isenta de radiação e apresenta alta eficácia na diferenciação entre lesões císticas e sólidas, servindo como método complementar valioso na prática biomédica e radiológica [1].

Entretanto, a interpretação de exames ultrassonográficos é intrinsecamente dependente da experiência do operador e sujeita à variabilidade inter e intra-observador, dada a complexidade na análise de contornos, ecoestrutura, atenuação acústica e margens tumorais [1]. Nesse contexto, o desenvolvimento de sistemas de Diagnóstico Auxiliado por Computador (CAD - Computer-Aided Diagnosis) baseados em inteligência artificial e aprendizado de máquina surge como uma solução promissora para fornecer uma segunda opinião objetiva, quantitativa e reprodutível aos profissionais de saúde [1].

Objetivando otimizar a eficácia diagnóstica e a eficiência de triagem, este estudo investiga a aplicação de métodos de aprendizado de máquina aplicados a dados de ultrassonografia mamária de modo-B. O foco central compreende a avaliação comparativa de estratégias metaheurísticas de seleção de atributos — Algoritmo Genético (AG) e Otimização por Enxame de Partículas (PSO) — visando reduzir o espaço de atributos e superar o 'mal da dimensionalidade'. Em seguida, analisa-se a capacidade preditiva de diversos algoritmos de classificação em um protocolo experimental rigoroso, definindo uma arquitetura computacional capaz de combinar alta sensibilidade diagnóstica e reduzida complexidade operacional.

II. FUNDAMENTAÇÃO TEÓRICA

A. Ultrassonografia Mamária em Modo-B e Extração de Atributos

A ultrassonografia mamária baseia-se na emissão de ondas mecânicas de alta frequência (tipicamente entre 7 e 15 MHz) geradas por transdutores piezoelétricos. As ondas sonoras propagam-se nos tecidos biológicos e sofrem fenômenos de reflexão, refração e atenuação ao encontrarem interfaces com diferentes impedâncias acústicas [1]. Os ecos refletidos são captados pelo próprio transdutor, convertidos em sinais elétricos e processados por algoritmos de reconstrução para formar a imagem em modo-B (Brightness mode), na qual a amplitude do eco é mapeada em níveis de cinza.

No contexto deste trabalho, os dados originais derivam do repositório público e padronizado BUS-BRA [1], coletado a partir de exames ultrassonográficos de 1.064 pacientes, totalizando 1.875 imagens com confirmação diagnóstica por biópsia e classificação segundo o sistema BI-RADS (Breast Imaging Reporting and Data System) [1]. A caracterização quantitativa das lesões (extração de atributos) abrange descritores de textura (matriz de coocorrência de níveis de cinza - GLCM), parâmetros morfométricos da margem tumoral (circularidade, rugosidade, compacidade) e estatísticas de intensidade e atenuação acústica [1]. Cabe ressaltar que a base derivada utilizada neste estudo (Lenet_ultrassonografia.arff, contendo 879 instâncias) é um recorte amostral cuidadosamente pré-processado, no qual cada instância (linha de atributos) corresponde de forma unívoca a um achado focal/paciente distinto. Essa seleção mitigou a presença de múltiplas imagens de um mesmo indivíduo, prevenindo o vazamento de informações em nível de paciente (*patient-level leakage*) durante as etapas posteriores de particionamento e validação cruzada.

B. Redução de Dimensionalidade e Algoritmos de Seleção

datasets de imagens biomédicas frequentemente apresentam alta dimensionalidade. O dataset considerado (Lenet_ultrassonografia.arff) possui 501 atributos numéricos derivados da extração de características. A presença de um número elevado de atributos em relação ao tamanho da amostra induz o fenômeno conhecido como 'mal da dimensionalidade', elevando o risco de sobreajuste (overfitting), degradando o desempenho dos classificadores e aumentando o tempo de processamento computacional.

Para contornar essa limitação, investigou-se a aplicação de duas metaheurísticas de otimização estocástica:

1) Algoritmo Genético (AG): Inspirado nos mecanismos da evolução natural e seleção darwiniana, o AG codifica subconjuntos de atributos como cromossomos binários. Através de operadores de seleção, cruzamento (crossover) e mutação, a população evolui ao longo de gerações para maximizar uma função de aptidão (fitness) baseada na capacidade discriminatória dos atributos [17]. Configurou-se o AG com 50 gerações e 100 iterações.

2) Otimização por Enxame de Partículas (PSO - Particle Swarm Optimization): Baseado no comportamento social de agrupamento de aves e peixes, o PSO ajusta iterativamente a posição e velocidade de partículas candidato no espaço de busca d-dimensional. Cada partícula atualiza sua trajetória considerando sua melhor posição individual histórica (p_{best}) e a melhor posição global compartilhada pelo enxame (g_{best}) [17]. O algoritmo foi executado com 50 gerações e 100 iterações.

C. Tratamento de Desbalanceamento com SMOTE

O desbalanceamento de classes é um desafio recorrente na área biomédica, em que a proporção de casos patológicos costuma ser inferior à de casos benignos ou controles. A base estudada apresenta 272 amostras da classe C1

(maligna/casos) e 607 amostras da classe C2 (benigna/controles). Classificadores treinados sob conjuntos desbalanceados tendem a apresentar um viés em favor da classe majoritária, resultando em uma taxa frequentemente elevada de falsos negativos para a doença maligna.

Para solucionar essa assimetria, utilizou-se o filtro SMOTE (Synthetic Minority Over-sampling Technique) [5]. O SMOTE cria instâncias sintéticas da classe minoritária através de interpolação no espaço de atributos. Para cada amostra x_i da classe minoritária, identifica-se seus k vizinhos mais próximos da mesma classe; seleciona-se aleatoriamente um vizinho x_{zi} e gera-se um novo ponto sintético x_{novo} conforme a relação:

$$x_novo = x_i + \lambda \cdot (x_zi - x_i), \quad \lambda \in [0, 1] \qquad (1)$$

D. Modelos de Aprendizado de Máquina

Com o intuito de obter uma análise abrangente e robusta, selecionou-se uma gama diversificada de famílias de classificadores de aprendizado supervisionado:

• Random Forest (RF): Modelo ensemble baseado na construção de múltiplas árvores de decisão não correlacionadas, utilizando reamostragem bootstrap (bagging) e seleção aleatória de atributos por nó. Avaliaram-se configurações com 100, 200, 300, 400 e 500 árvores [6].

• Árvores de Decisão (J48 e Random Tree): O algoritmo J48 representa a implementação no ambiente Weka do clássico C4.5, que utiliza a razão de ganho de informação para realizar divisões recursivas. A Random Tree constrói árvores individuais considerando subconjuntos aleatórios de atributos a cada nó.

• Classificadores Bayesianos (Naive Bayes e BayesNet): O Naive Bayes assume independência condicional absoluta entre os atributos dada a classe. A BayesNet (Redes Bayesianas) flexibiliza essa suposição, mapeando as relações de dependência estocástica através de grafos acíclicos direcionados (DAGs).

• Máquinas de Vetores de Suporte (SVM/SMO): Implementadas via algoritmo Sequential Minimal Optimization (SMO), buscam determinar o hiperplano de separação de margem máxima no espaço de atributos. Foram testados kernels linear, polinomial de grau 2 e polinomial de grau 3, fixando o parâmetro de penalização C = 0,01.

III. METODOLOGIA

A. Descrição do Dataset e Pipeline Experimental

O fluxo metodológico desenvolvido está sumarizado na estratégia de processamento da base Lenet_ultrassonografia.arff, composta por 879 instâncias e 501 atributos numéricos contínuos, acompanhados pelo rótulo da classe categórica (C1 = 272 amostras malignas; C2 = 607 amostras benignas). A metodologia seguiu cinco etapas fundamentais:

1. Carregamento e conversão da base de dados original no formato ARFF;

2. Particionamento inicial e estratificado do dataset em conjuntos de Treinamento (70%) e Teste (30%) utilizando o filtro Resample sem reposição de instâncias. Esta etapa isola o conjunto de teste de qualquer intervenção subsequente de otimização;

3. Execução dos algoritmos de seleção de atributos (AG e PSO) estritamente sobre o conjunto de Treinamento para determinação do subconjunto ótimo de características;

4. Aplicação do filtro de atributos gerado na etapa anterior ao conjunto de Teste independente, garantindo que ele possua a mesma dimensionalidade do Treinamento sem influenciar a seleção;

5. Balanceamento exclusivo do conjunto de Treinamento reduzido via SMOTE;

6. Treinamento, validação cruzada 10-fold com 5 repetições e avaliação final dos classificadores no conjunto de teste independente nos ambientes Weka e Jamovi.

B. Seleção de Atributos e Configuração Experimental

Os dois algoritmos de seleção foram executados sob as mesmas condições iniciais de parada para garantir isonomia na comparação: 50 gerações e 100 iterações. Conduzida exclusivamente sobre os 70% dos dados de treinamento, a busca por Algoritmo Genético (AG) convergiu selecionando um subconjunto promissor de 255 atributos (redução de ~49,1% da dimensão original). Por sua vez, a Otimização por Enxame de Partículas (PSO) demonstrou maior agressividade de compressão, selecionando 104 atributos (redução de ~79,2%).

Após a convergência, os atributos eliminados no Treinamento também foram removidos do conjunto de Teste. Os dados finais foram convertidos para o formato CSV e submetidos ao software Jamovi para condução das análises estatísticas descritivas das métricas operacionais. .

C. Protocolo de Treinamento e Validação Cruzada

Para favorecer a generalização do modelo e mitigar o risco de memorização estática dos dados (overfitting), adotou-se o módulo Experimenter do software Weka. O conjunto de treinamento (70% da base original) passou por balanceamento via SMOTE e foi avaliado por validação cruzada K-Fold com K=10, repetida por 5 execuções independentes com sementes de aleatoriedade distintas (totalizando 50 ciclos de treino/validação por classificador). Ressalta-se que a total independência metodológica do conjunto de teste (30%) foi preservada, uma vez que não teve participação nas etapas de seleção de atributos ou geração de

instâncias sintéticas, resultando em métricas de avaliação realistas e não otimistas.

D. Métricas de Avaliação do Desempenho

A eficácia diagnóstica dos modelos foi quantificada a partir das métricas consagradas na literatura biomédica, calculadas a partir dos valores da matriz de confusão: Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN).

• Acurácia (Acc): Razão entre as predições corretas e o total de amostras:

$$Acc = (TP + TN) / (TP + TN + FP + FN) \qquad (2)$$

• Coeficiente Kappa de Cohen (kappa): Medida de concordância inter-observadores ajustada para a concordância esperada pelo acaso (P_e):

$$\kappa = (P_o - P_e) / (1 - P_e) \qquad (3)$$

em que P_o representa a acurácia observada do classificador.

• Sensibilidade (S_n): Proporção de casos malignos diagnosticados corretamente:

$$S_n = TP / (TP + FN) \qquad (4)$$

• Especificidade (S_p): Proporção de lesões benignas identificadas corretamente:

$$S_p = TN / (TN + FP) \qquad (5)$$

• Área sob a Curva ROC (AUC): Desempenho global da capacidade de discriminação probabilística da taxa de verdadeiros positivos em função dos falsos positivos.

IV. RESULTADOS E DISCUSSÃO

A. Análise do Desempenho dos Classificadores sob Algoritmo Genético

Os resultados obtidos durante os 50 ciclos de validação cruzada no conjunto de treinamento utilizando a base reduzida pelo Algoritmo Genético (255 atributos) estão consolidados na Tabela I, contendo as médias e desvios-padrão (\mu \pm \sigma) das métricas estipuladas.

TABELA I
DESEMPENHO DOS CLASSIFICADORES COM SELEÇÃO POR ALGORITMO GENÉTICO (MÉDIA ± DESVIO-PADRÃO)

Classificador	Acurácia (%)	Kappa (κ)	Sensib. (S_n)	Especif. (S_p)	AUC-ROC
RF 100	87,86 ± 1,60	0,720 ± 0,072	0,873 ± 0,045	0,873 ± 0,052	0,970 ± 0,017
RF 200	87,90 ± 1,60	0,722 ± 0,077	0,874 ± 0,046	0,874 ± 0,053	0,970 ± 0,016
RF 300	87,90 ± 1,68	0,722 ± 0,074	0,875 ± 0,049	0,875 ± 0,044	0,970 ± 0,015
RF 400	87,87 ± 1,60	0,721 ± 0,073	0,877 ± 0,048	0,877 ± 0,048	0,970 ± 0,016
RF 500	87,87 ± 1,55	0,721 ± 0,072	0,876 ± 0,041	0,876 ± 0,048	0,970 ± 0,016
BayesNet	71,80 ± 4,37	0,404 ± 0,097	0,595 ± 0,058	0,740 ± 0,076	0,792 ± 0,041
Naive Bayes	67,30 ± 4,29	0,347 ± 0,086	0,546 ± 0,057	0,731 ± 0,046	0,731 ± 0,035
SMO (SVM)	79,80 ± 1,50	0,551 ± 0,033	0,824 ± 0,039	0,674 ± 0,155	0,799 ± 0,061
J48	82,70 ± 1,50	0,614 ± 0,077	0,832 ± 0,048	0,822 ± 0,042	0,861 ± 0,040
RandomTree	82,50 ± 4,01	0,610 ± 0,080	0,827 ± 0,051	0,824 ± 0,055	0,825 ± 0,040

A análise da Tabela I revela que a família de classificadores baseada em conjuntos de árvores (Random Forest) alcançou os desempenhos mais elevados e estáveis. As variações de Random Forest com 100 a 500 árvores obtiveram acurácias superiores a 87,8% e valor de área sob a curva ROC (AUC) consistente de 0,970, demonstrando elevada capacidade discriminatória.

Dentre as configurações de Random Forest, os modelos com 200 e 300 árvores obtiveram os índices Kappa mais elevados (0,722 ± 0,077), caracterizando uma concordância substancial segundo os critérios de Landis e Koch. Por outro lado, os classificadores probabilisticos demonstraram desempenho inferior: o Naive Bayes obteve a menor acurácia (67,30%) e a menor sensibilidade (54,60%), consequência direta da violação da premissa de independência estocástica entre os atributos extraídos das imagens ultrassonográficas.

B. Comparação com a Seleção por Enxame de Partículas (PSO)

Em relação à abordagem de Otimização por Enxame de Partículas (PSO), observa-se comportamento preditivo análogo no que tange à superioridade das florestas randômicas. O algoritmo Random Forest com 200 árvores manteve a liderança técnica no conjunto de validação sob a base do PSO. Em contrapartida, o modelo SVM com kernel linear apresentou o pior desempenho global sob o filtro PSO, sugerindo que a redução agressiva para 104 atributos elimina

relações não lineares essenciais para o hiperplano de separação linear.

Embora o PSO promova uma compressão de atributos mais severa (104 atributos contra 255 do AG), a seleção por Algoritmo Genético proporcionou maior estabilidade estatística e leve vantagem nos valores absolutos de sensibilidade no conjunto de validação. Ademais, do ponto de vista operacional, o AG demandou menor número total de iterações para atingir a convergência ideal em comparação ao enxame de partículas.

C. Critério de Seleção do Modelo Ótimo e Custo Computacional

A definição do modelo ideal deve considerar o compromisso entre acurácia diagnóstica e viabilidade de implementação computacional. Ao comparar o Random Forest de 200 árvores com o de 300 árvores sob a seleção por AG, constata-se que ambos os modelos apresentam desempenhos virtualmente idênticos (Acurácia de 87,90% e Kappa de 0,722).

Contudo, do ponto de vista da Engenharia Biomédica e da aplicação clínica em tempo real, o modelo com 200 árvores exige um custo computacional significativamente menor no processo de inferência e alocação de memória RAM. Por conseguinte, elegeu-se a combinação Algoritmo Genético + Random Forest (200 árvores) como a arquitetura mais adequada para este estudo.

D. Avaliação no Conjunto de Teste Independente e Matriz de Confusão

Após a consolidação da arquitetura ótima (AG + Random Forest 200 árvores), o modelo treinado foi submetido à avaliação no conjunto de teste independente (30% dos dados, englobando aproximadamente 263 instâncias não vistas durante a etapa de otimização). Os resultados do teste final foram extraídos via aba Classify do ambiente Weka e estão apresentados na matriz de confusão da Fig. 1.

FIG. 1. MATRIZ DE CONFUSÃO E DESEMPENHO NO CONJUNTO DE TESTE INDEPENDENTE (AG + RANDOM FOREST 200 ÁRVORES)

Classe Real \ Predita	**Previsto: MALIGNO (C1)**	**Previsto: BENIGNO (C2)**
Real: MALIGNO (C1)	**VERDADEIRO POSITIVO 91,6% (79 casos)**	FALSO NEGATIVO 8,4% (02 casos)
Real: BENIGNO (C2)	FALSO POSITIVO 12,6% (20 casos)	**VERDADEIRO NEGATIVO 87,4% (162 casos)**

A análise da matriz de confusão demonstra resultados promissores do modelo de suporte CAD em dados não vistos. No grupo de tumores malignos (Classe C1), o sistema identificou corretamente 79 de 81 casos, alcançando uma Sensibilidade de 91,6% e limitando a taxa de Falsos Negativos a apenas 8,4% (2 casos). Em rastreio oncológico, minimizar os falsos negativos é o requisito sanitário mais crítico, haja vista que a omissão de um tumor maligno atrasa o tratamento e compromete drasticamente a sobrevida da paciente.

Paralelamente, para a classe de lesões benignas (Classe C2), o modelo obteve uma Especificidade de 87,4% (162 acertos em 182 lesões benignas), com 12,6% de Falsos Positivos (20 casos). Esse nível de especificidade é relevante, indicando que , em um cenário de aplicação real, um modelo de suporte CAD com tais métricas poderia auxiliar secundariamente na redução de indicações para biópsias invasivas desnecessárias e mitigar a ansiedade associada a diagnósticos falso-positivos.

V. CONCLUSÃO

Este estudo apresentou evidências preliminares da viabilidade técnica da integração de algoritmos de Aprendizado de Máquina e seleção de atributos metaheurística no processamento de imagens de ultrassonografia mamária em modo-B para o apoio ao rastreio precoce do câncer de mama.

As análises estatísticas comparativas evidenciaram que a abordagem combinada utilizando o Algoritmo Genético (seleção de 255 atributos) e o classificador Random Forest configurado com 200 árvores alcançou o melhor equilíbrio operacional. A solução atingiu acurácia de 87,90%, índice Kappa de 0,722 e área sob a curva ROC de 0,970 na validação cruzada, culminando em uma elevada sensibilidade diagnóstica de 91,6% e especificidade de 87,4% no conjunto de teste independente.

Destaque-se que a escolha criteriosa do modelo de 200 árvores (em detrimento de arquiteturas mais complexas como 300 a 500 árvores) viabilizou uma redução expressiva do gasto computacional e da latência de processamento sem qualquer prejuízo às métricas de acurácia, aspecto relevante para futuras investigações voltadas à implementação em equipamentos ultrassonográficos portáteis de ponto de atendimento (Point-of-Care).

Apesar dos resultados promissores, reconhece-se como potencial limitação deste estudo a utilização de um único conjunto de dados, o que pode restringir a capacidade de generalização do modelo proposto quando avaliado frente a imagens adquiridas em diferentes instituições de saúde ou com equipamentos de ultrassom distintos. Para mitigar essa limitação, em trabalhos futuros, pretende-se expandir a validação do modelo proposto através da inclusão de bases de dados multicêntricas com diferentes marcas de transdutores, bem como explorar técnicas de fusão multimodal unindo atributos ultrassonográficos, mamográficos e variáveis

clínicas da paciente para o desenvolvimento de um sistema CAD de suporte à decisão em tempo real integrável ao Sistema Único de Saúde (SUS).

REFERÊNCIAS

[1] W. Gómez-Flores, M. J. Gregorio-Calas, and W. C. de Albuquerque Pereira, "BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems," Medical Physics, vol. 50, no. 10, pp. 3110–3123, 2023. DOI: 10.1002/mp.16812.

[2] S. Cerquetani, "Câncer de mama: tudo que você precisa saber sobre a doença," Blog Nav Dasa, 2022 (atualizado em 2025). [Online]. Disponível: https://nav.dasa.com.br/blog/cancer-de-mama.

[3] University of Waikato, "Weka: The Waikato Environment for Knowledge Analysis," Version 3.8.6, 2025. [Online]. Disponível: https://www.cs.waikato.ac.nz/ml/weka/.

[4] The Jamovi Project, "Jamovi (Version 2.3) [Computer Software]," 2025. [Online]. Disponível: https://www.jamovi.org.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[8] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proc. IEEE Int. Conf. Neural Networks (ICNN), Perth, WA, Australia, 1995, pp. 1942–1948.

Architectural Patterns in Blockchain-Based Electronic Health Records: A Systematic Literature Review

1st Giulia Nunes Cordeiro
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
giulia.nunes@aluno.ufabc.edu.br

2rd Mariana Mioti Sebastião
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
m.mioti@aluno.ufabc.edu.br

3th Fernanda Almeida
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
fernanda.almeida@ufabc.edu.br

***Abstract*—Electronic Health Records (EHRs) require secure and interoperable mechanisms for health information exchange. This study presents a systematic literature review of blockchain-based EHR architectures published between 2021 and 2026. Twenty studies were included, comprising fifteen primary studies and five secondary studies. The analysis of the primary studies revealed a predominance of permissioned blockchain platforms, particularly Hyperledger Fabric, combined with smart contracts, access control mechanisms, interoperability services, and hybrid storage models in which clinical data remain stored in off-chain repositories. Based on the recurring architectural patterns identified across the literature, a conceptual reference architecture was synthesized to summarize the functional components most frequently adopted in blockchain-based EHR systems. The findings indicate a convergence toward common architectural design patterns and provide a consolidated reference for future research on secure and interoperable healthcare information systems.**

***Index Terms*—Blockchain, Electronic Health Records, HL7 FHIR, Interoperability, Smart Contracts.**

I. Introduction

Electronic Health Records (EHRs) have become essential for healthcare delivery, supporting the storage, management, and exchange of clinical information across healthcare organizations. Despite advances in health information technologies, challenges related to interoperability, privacy, security, and data governance continue to limit efficient information sharing among heterogeneous systems [1]–[3].

Blockchain has emerged as a promising technology for addressing some of these challenges by providing decentralized trust, data integrity, auditability, and secure access management. Combined with smart contracts, blockchain enables the automation of consent management, authorization policies, and transaction auditing. Consequently, blockchain-based solutions have been proposed for Electronic Health Records, Personal Health Records, Health Information Exchange, and Internet of Medical Things (IoMT) applications [4]–[9], [14].

This work was supported by the Programa Pesquisando Desde o Primeiro Dia (PDPD) through an undergraduate research scholarship granted by the Pró-Reitoria de Pesquisa of the Federal University of ABC (Propes/UFABC), Brazil.

Recent studies increasingly combine blockchain platforms with complementary technologies such as interoperability standards, distributed storage, cloud services, and cryptographic mechanisms. Although these solutions pursue similar objectives, they differ regarding blockchain platforms, storage strategies, interoperability mechanisms, and security models, making it difficult to identify the architectural characteristics that consistently define blockchain-based EHR systems [7]–[13], [15], [16].

Previous reviews have examined blockchain adoption in healthcare from perspectives such as privacy, security, and health information management [1]–[3]. However, there remains limited understanding of the architectural patterns that recur across blockchain-based EHR solutions.

To address this gap, this study presents a systematic literature review of blockchain-based EHR architectures published between 2021 and 2026. Twenty studies were included, comprising fifteen primary studies and five secondary studies. The architectural synthesis was derived exclusively from the primary studies, while secondary studies were used to contextualize technological trends and research challenges. Based on the comparative analysis, a conceptual reference architecture was synthesized to summarize the functional components most frequently reported in blockchain-based EHR systems. The main contributions of this work are: (i) a systematic comparison of blockchain-based EHR architectures; (ii) the identification of recurring architectural patterns; and (iii) the synthesis of a conceptual reference architecture derived from evidence extracted from the primary studies.

II. Materials and Methods

This systematic literature review was conducted to identify and analyze blockchain-based architectures for Electronic Health Records (EHRs). The review followed the main principles of the PRISMA 2020 guidelines for study identification, screening, eligibility assessment, and qualitative synthesis. The review was conducted through four sequential activities: literature search, study selection, eligibility assessment, and data extraction.

The review included both primary and secondary studies. Primary studies proposed architectures, frameworks, or technical solutions for blockchain-based EHR systems, whereas secondary studies consisted of literature reviews that provided contextual information regarding the evolution of the field. To avoid duplication of evidence, only the primary studies were considered in the identification of recurring architectural patterns and in the synthesis of the conceptual reference architecture, while secondary studies were used exclusively to contextualize the findings and support the discussion.

A. Search Strategy

The literature search was performed in four scientific databases: PubMed, Scopus, Web of Science, and IEEE Xplore. These databases were selected because they index studies in healthcare, biomedical informatics, and computer science, covering the principal publication venues related to blockchain applications in healthcare.

The search strategy combined terms associated with blockchain, Electronic Health Records, interoperability, and information security. Boolean operators (AND and OR) were used to construct database-specific queries. The search expression was defined as:

> (”blockchain” OR ”blockchain-based” OR ”distributed ledger”) AND (”electronic health record” OR ”EHR” OR ”personal health record” OR ”PHR”) AND (”interoperability” OR ”data sharing” OR ”health information exchange”) AND (”security” OR ”privacy” OR ”access control” OR ”encryption”) AND (”framework” OR ”architecture” OR ”model” OR ”system” OR ”guideline” OR ”standard”)

Only studies published between 2021 and 2026 were considered.

B. Study Selection and Data Extraction

All retrieved records were exported to Zotero for reference management and duplicate removal. After eliminating duplicate records, titles and abstracts were screened to identify potentially relevant studies. The screening process excluded publications that were unrelated to blockchain-based EHR systems, outside the defined publication period, unavailable as full-text articles, or classified as editorials, letters, or conference abstracts without complete manuscripts.

The remaining studies underwent full-text assessment to verify their relevance to the objectives of this review. Articles that did not present a clear methodological description or did not specifically address blockchain applications for Electronic Health Records were excluded. Studies were classified as primary or secondary during the full-text assessment. Primary studies described original blockchain-based EHR architectures, frameworks, or implementation strategies, whereas secondary studies synthesized previous research findings. This classification was used throughout the review process to distinguish evidence sources and avoid incorporating secondary evidence into the architectural synthesis. The screening and data extraction processes were conducted by the first reviewer and subsequently verified by a second reviewer. Any discrepancies were resolved through discussion.

For each included study, information regarding blockchain platform, network type, interoperability mechanisms, security strategies, storage architecture, and other architectural characteristics was extracted and organized in a structured data extraction form. For the primary studies, the presence or absence of architectural components was recorded to support a comparative analysis of their occurrence across the literature. The frequency of each component was calculated as the number and percentage of primary studies in which the component was identified. These frequencies were subsequently used to characterize the recurrence of architectural elements across the reviewed literature and to support the synthesis of the conceptual reference architecture.

C. Quality Assessment

Because the review included studies with different objectives and methodological designs, a simplified quality assessment was performed for the primary studies. The assessment considered three criteria: (i) clarity of the architectural description, (ii) description of implementation or validation procedures, and (iii) discussion of limitations or practical applicability. The assessment was used to characterize the evidence base and support the interpretation of the findings, but it was not employed as an exclusion criterion.

The results of the quality assessment are presented in the Results section and were considered when interpreting the strength of the available evidence.

III. Results

A. Characteristics of the Selected Studies

The literature search identified 631 records across PubMed, Scopus, Web of Science, and IEEE Xplore. After duplicate removal, 436 unique records remained for title and abstract screening. Following the application of the predefined eligibility criteria, 135 articles underwent full-text assessment, of which 20 satisfied all inclusion criteria and were included in the qualitative synthesis. The study selection process is illustrated in Fig. 1.

The final sample included both primary and secondary studies addressing blockchain applications for Electronic Health Records (EHRs). Primary studies proposed new architectures, frameworks, and secure data-sharing mechanisms, whereas secondary studies synthesized the current state of research and discussed technological trends, research challenges, and future directions. Table I summarizes the general characteristics of the selected literature.

Among the 15 primary studies, 13 proposed new architectures or frameworks, whereas two focused on the evaluation of previously developed solutions. The remaining five studies corresponded to secondary literature reviews.

The temporal distribution of the selected publications shows an increase in research activity after 2022. Earlier work primarily investigated the feasibility of applying blockchain to improve the security and integrity of Electronic Health

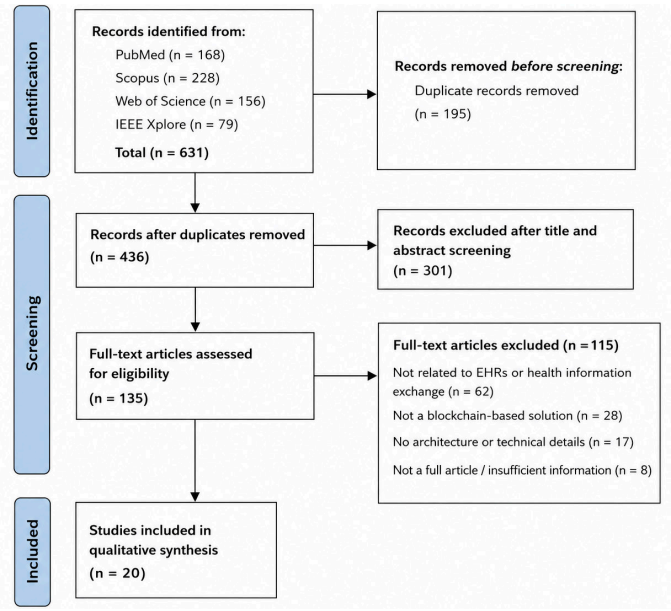


Fig. 1. PRISMA 2020 flow diagram illustrating the study identification, screening, eligibility assessment, and inclusion process adopted in this review.

TABLE I
GENERAL CHARACTERISTICS OF THE SELECTED STUDIES.

Characteristic	Category	n	%
2*Study type	Primary studies	15	75.0
	Secondary studies	5	25.0
3*Publication period	2021–2022	6	30.0
	2023–2024	8	40.0
	2025–2026	6	30.0
3*Research focus	Architecture proposal	13	65.0
	Framework evaluation	2	10.0
	Literature review	5	25.0

Records, whereas more recent publications proposed complete software architectures integrating permissioned blockchain platforms with complementary technologies such as smart contracts, cloud computing, distributed storage, Internet of Medical Things (IoMT), and interoperability standards including HL7 FHIR [7]–[11], [13], [18], [19].

Despite differences in application domains and implementation strategies, most selected studies focused on secure, interoperable, and auditable health information exchange while preserving patient privacy and data integrity. Rather than relying exclusively on blockchain, the majority of proposed solutions adopted hybrid architectures combining distributed ledgers with off-chain storage, cryptographic mechanisms, standardized communication protocols, and fine-grained access control policies [6]–[13].

To avoid duplication of evidence, only the 15 primary studies were included in the quantitative synthesis of architectural components and in the derivation of the reference architecture. The five secondary studies were used exclusively to contextualize research trends, reported challenges, and future directions discussed in the literature.

A summary of the quality assessment (Table II) performed for the primary studies is presented in Section II, providing an overview of the methodological characteristics of the evidence base used for the architectural synthesis.

TABLE II
SUMMARY OF THE QUALITY ASSESSMENT OF THE 15 PRIMARY STUDIES.

Criterion	Yes	No	% Yes
Architectural description	15	0	100
Implementation/validation	8	7	53.3
Limitations discussed	6	9	40

B. Quality Assessment

The quality assessment was performed for the 15 primary studies according to the criteria described in the Materials and Methods section. Table II summarizes the results obtained for each evaluation criterion.

All primary studies provided sufficient architectural descriptions to support data extraction and comparative analysis. However, fewer studies reported implementation details, experimental validation, or performance evaluation. Likewise, the discussion of limitations and practical applicability was not consistently addressed across the reviewed literature.

Overall, the assessment indicates that the current evidence base is largely composed of architectural proposals and proof-of-concept implementations. Consequently, the architectural patterns synthesized in this review should be interpreted as recurring design characteristics reported in the literature rather than as evidence of validated large-scale healthcare deployments.

C. Blockchain Platforms and Network Architectures

The comparative analysis of the selected studies revealed a predominance of permissioned blockchain platforms for Electronic Health Record (EHR) systems, with Hyperledger Fabric being the most frequently adopted technology. Ethereum appeared less frequently, while a smaller number of studies proposed custom permissioned or cross-platform blockchain solutions.

Table III summarizes the blockchain platforms identified in the reviewed literature. Hyperledger Fabric was identified in 9 of the 15 primary studies (60%), making it the most frequently adopted blockchain platform. Ethereum appeared in two studies (13.3%), whereas three studies (20%) proposed custom permissioned implementations. Only one study addressed interoperability across heterogeneous blockchain networks.

Hyperledger Fabric was the most frequently adopted blockchain platform among the reviewed studies. Most implementations employed permissioned networks that support authenticated participation, configurable access policies, and institutional governance suitable for healthcare environments [7]–[11].

Among the primary studies included in this review, no implementation relied exclusively on a public permissionless

TABLE III
BLOCKCHAIN PLATFORMS IDENTIFIED AMONG THE PRIMARY STUDIES.

Blockchain Platform	n	%	Main Characteristics
Hyperledger Fabric	9	60	Permissioned blockchain supporting institutional governance, smart contracts, and secure health information exchange.
Ethereum	2	13.3	Public blockchain mainly employed for smart contract implementation and decentralized authorization.
Custom Permissioned Blockchain	3	20	Institution-specific blockchain implementations designed for secure EHR sharing.
Platform-independent / Cross-platform	1	6.7	Architectures supporting interoperability among heterogeneous blockchain networks.

blockchain for EHR management. Instead, the proposed solutions adopted permissioned or consortium blockchain networks involving hospitals, laboratories, healthcare providers, and other trusted institutions operating under predefined governance policies [7]–[12].

Earlier publications primarily focused on demonstrating the feasibility of blockchain for secure Electronic Health Record management. More recent work proposed increasingly comprehensive solutions integrating smart contracts, distributed storage, cloud services, Internet of Medical Things (IoMT), interoperability standards, and advanced access control mechanisms [7]–[11], [13]–[16].

Despite differences in implementation strategies, most proposed solutions adopted a similar storage model. Blockchain was primarily employed to maintain transaction metadata, cryptographic hashes, audit logs, and access permissions, whereas complete Electronic Health Records remained stored in external repositories such as hospital databases, cloud infrastructures, or distributed storage systems [8], [9], [11], [12], [15], [16], [20]. This hybrid storage strategy was identified in the majority of the primary studies and represents one of the most recurrent architectural characteristics observed in the reviewed literature.

D. Security and Access Control Mechanisms

Security mechanisms were consistently addressed across the reviewed studies. Although different blockchain platforms and software architectures were proposed, most solutions combined cryptographic techniques, smart contracts, identity management, and fine-grained access control policies to ensure the confidentiality, integrity, and traceability of Electronic Health Records (EHRs) [6]–[13], [15], [16], [20]. Table IV summarizes the principal security mechanisms identified in the reviewed literature. Smart contracts were the most frequently reported component, supporting authorization, patient consent management, audit automation, and transaction validation. Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) were also widely adopted to regulate access according to user roles or contextual attributes.

Smart contracts were the most frequently reported mechanism for implementing security policies in blockchain-based EHR systems. They were employed to automate patient consent, verify access permissions, register audit events, and regulate data sharing among healthcare organizations [4], [5], [7], [9], [13].

Access control was another recurring component among the reviewed implementations. Initial studies predominantly adopted Role-Based Access Control (RBAC), whereas more recent solutions incorporated Attribute-Based Access Control (ABAC), contextual authorization, and patient-centric permission models capable of adapting access decisions according to different clinical scenarios [6], [8], [10], [11], [15], [16].

Nearly all reviewed architectures recorded access events, permission updates, and data-sharing operations on the blockchain ledger, taking advantage of its immutable transaction history. However, complete Electronic Health Records were rarely stored on-chain. Instead, cryptographic hashes, metadata, transaction identifiers, and audit records were maintained on the blockchain, while clinical information remained stored in external repositories such as hospital databases, cloud infrastructures, or distributed storage systems [7]–[12], [15], [16], [20].

Across the reviewed studies, secure EHR management was consistently implemented through the combination of permissioned blockchain networks, smart contracts, cryptographic mechanisms, access control policies, and off-chain data repositories. This combination of technologies represents the predominant security model identified in the reviewed literature.

E. Interoperability and Data Storage Strategies

The reviewed studies consistently combined blockchain with interoperability standards and external storage solutions to support secure Electronic Health Record (EHR) exchange across heterogeneous healthcare systems. Rather than relying exclusively on distributed ledger technology, the proposed solutions integrated standardized communication protocols, application programming interfaces (APIs), and complementary storage infrastructures to enable interoperability among healthcare organizations [7]–[13]. Table V summarizes the principal interoperability mechanisms and data storage strategies identified in the reviewed literature. HL7 FHIR and RESTful APIs were the most frequently adopted communication technologies, while off-chain repositories represented the predominant strategy for storing complete Electronic Health Records.

Nearly all reviewed implementations adopted hybrid storage architectures in which complete Electronic Health Records remained stored in external repositories, whereas blockchain maintained cryptographic hashes, transaction metadata, access

TABLE IV
CORE SECURITY COMPONENTS IDENTIFIED IN BLOCKCHAIN-BASED EHR ARCHITECTURES.

Security Mechanism	Primary Purpose	Typical Implementation
Smart Contracts	Automate authorization, consent management, audit procedures, and transaction validation	Execution of access policies, patient consent management, and automated auditing within permissioned blockchain networks.
Role-Based Access Control (RBAC)	Restrict data access according to predefined user roles	Authorization based on professional roles such as physicians, nurses, patients, and administrators.
Attribute-Based Access Control (ABAC)	Provide context-aware authorization based on user and resource attributes	Dynamic authorization considering contextual attributes such as organization, department, emergency status, or user permissions.
Identity Management	Authenticate and identify participants in permissioned blockchain networks	Digital identities, certificates, and membership services used to authenticate trusted healthcare organizations and users.
Audit Logging	Maintain an immutable history of data access and system events	Recording of transactions, permission updates, and access history on the blockchain ledger for traceability and accountability.
Cryptographic Mechanisms	Protect confidentiality, integrity, and secure data exchange	Hash functions, digital signatures, encryption, and proxy re-encryption applied to protect sensitive clinical information.
Patient Consent Management	Allow patients to control authorization for sharing Electronic Health Records	Consent policies implemented through smart contracts, enabling patients to grant or revoke access permissions.

TABLE V
CORE INTEROPERABILITY COMPONENTS IDENTIFIED IN BLOCKCHAIN-BASED EHR ARCHITECTURES.

Component	Primary Purpose	Typical Implementation
HL7 FHIR	Standardized representation and exchange of clinical information	Interoperable exchange of Electronic Health Records between heterogeneous healthcare information systems.
RESTful APIs	Communication among blockchain services and healthcare applications	Integration of blockchain platforms with hospital information systems, cloud services, and external repositories.
Off-chain Storage	Storage of complete Electronic Health Records	Hospital databases, cloud repositories, and distributed storage maintaining clinical data outside the blockchain.
InterPlanetary File System (IPFS)	Distributed storage of large medical files	Secure storage of diagnostic images, clinical documents, and other large healthcare datasets.
Cloud Computing	Scalable storage and computational infrastructure	Cloud-based repositories supporting secure data storage, processing, and system scalability.
Cross-platform Interoperability	Exchange of data across heterogeneous blockchain networks	Communication among healthcare organizations operating different blockchain platforms.
IoMT Integration	Continuous acquisition of patient-generated health data	Secure integration of connected medical devices with blockchain-enabled healthcare systems.

permissions, and audit records [7]–[12], [15], [16], [20]. More recent publications increasingly incorporated interoperability standards such as HL7 FHIR together with RESTful APIs to facilitate communication among hospitals, laboratories, cloud services, and healthcare information systems [7]–[10], [12], [13]. Several studies also integrated complementary technologies including cloud computing, distributed storage platforms such as IPFS, and Internet of Medical Things (IoMT) devices to support scalable storage, secure data exchange, and continuous acquisition of patient-generated health information [8], [13], [14], [16].

Across the reviewed literature, blockchain was consistently integrated with interoperability standards, external storage repositories, cloud services, and healthcare communication interfaces. These components formed the predominant interoperability model identified among the reviewed blockchain-based EHR architectures.

F. Architectural Synthesis

The comparative analysis identified a consistent set of architectural design patterns across the reviewed studies. Although the selected works differed in their implementation details, blockchain platforms, and application scenarios, most solutions shared a common set of functional components supporting secure and interoperable Electronic Health Record (EHR) systems [6]–[13], [15], [16], [20].

Table VI summarizes the architectural components most frequently identified in the reviewed literature. Together, these components represent the architectural elements identified across the 15 primary studies included in this review and provide the empirical basis for the synthesis of the conceptual reference architecture.

Based on the recurring architectural design patterns identified across the selected studies, a conceptual reference architecture was synthesized to summarize the functional layers consistently adopted by blockchain-based EHR systems. Rather than representing a specific implementation, Fig. 2 presents a conceptual model that integrates the common design principles observed throughout the reviewed literature.

Unlike the individual architectures proposed in the reviewed studies, the reference architecture presented here does not represent a new implementation. Instead, it synthesizes the recurring architectural components identified across multiple

TABLE VI
ARCHITECTURAL COMPONENTS IDENTIFIED IN THE PRIMARY STUDIES (N = 15).

Architectural Component	n	%	Architectural Role
Off-chain storage	15	100	Stores complete Electronic Health Records outside the blockchain while preserving scalability and performance.
Smart contracts	14	93.3	Automate authorization, patient consent management, and enforcement of access policies.
Permissioned blockchain	13	86.7	Provides the trusted distributed infrastructure responsible for transaction validation, integrity, and auditability.
Cryptographic mechanisms	13	86.7	Ensure confidentiality, integrity, authentication, and secure information exchange.
Access control (RBAC/ABAC)	11	73.3	Restricts access to clinical information according to predefined security policies.
Audit logging	11	73.3	Maintains immutable records of transactions and data access events.
RESTful APIs	8	53.3	Enable communication between blockchain services and healthcare applications.
Identity management	7	46.7	Authenticates organizations and users participating in permissioned networks.
Patient consent management	7	46.7	Allows patients to control authorization for Electronic Health Record sharing.
Advanced cryptographic schemes	7	46.7	Provide enhanced privacy through mechanisms such as proxy re-encryption and zero-knowledge approaches.
HL7 FHIR	4	26.7	Supports standardized exchange of clinical information among heterogeneous healthcare systems.
IPFS	3	20	Supports distributed storage of large medical files and documents.
IoMT integration	2	13.3	Enables secure acquisition and sharing of patient-generated health data.
Edge computing	1	6.7	Provides distributed processing capabilities closer to data sources.

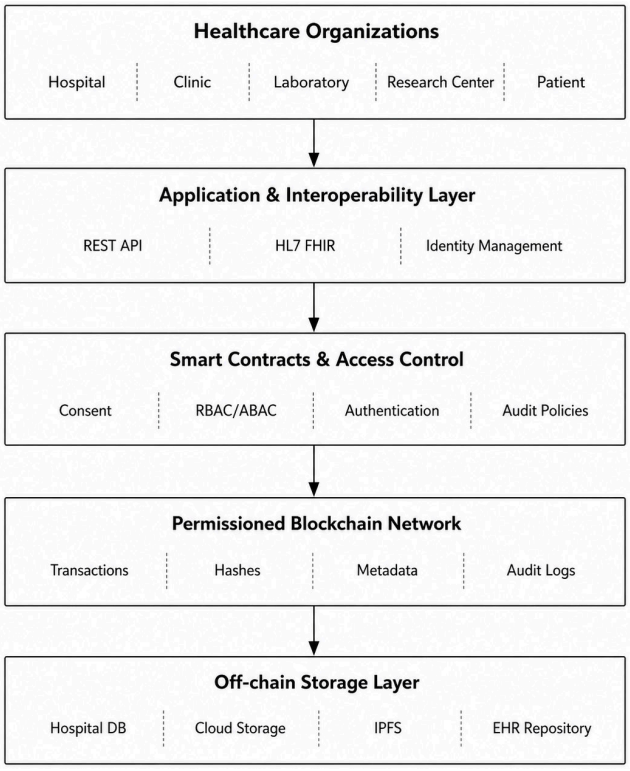


Fig. 2. Conceptual reference architecture synthesized from the reviewed studies. The architecture summarizes the functional layers consistently identified in blockchain-based Electronic Health Record systems, including healthcare organizations, interoperability services, smart contracts and access control, permissioned blockchain infrastructure, and off-chain storage repositories.

independent studies into a single conceptual model, providing a consolidated view of the design principles currently adopted

in blockchain-based EHR systems.

As illustrated in Fig. 2, healthcare organizations—including hospitals, clinics, laboratories, research centers, and patient applications—interact through an application layer responsible for interoperability and identity management. Communication among heterogeneous healthcare systems is supported by standardized interfaces, particularly RESTful APIs and HL7 FHIR services, enabling secure information exchange across institutional boundaries.

Security policies are implemented through a dedicated functional layer comprising smart contracts, consent management, authentication mechanisms, role-based and attribute-based access control, and audit policies. Together, these components regulate secure data sharing among authorized participants.

The permissioned blockchain network functions as the trust layer of the proposed reference architecture. Rather than storing complete Electronic Health Records, it maintains transaction metadata, cryptographic hashes, audit logs, and access records, ensuring integrity, traceability, and non-repudiation while minimizing storage requirements.

Complete Electronic Health Records remain stored in off-chain repositories, including hospital databases, cloud platforms, IPFS nodes, and dedicated EHR repositories. This hybrid storage model was consistently identified across the reviewed studies and summarizes the storage strategy adopted by most blockchain-based healthcare architectures.

To facilitate the interpretation of the traceability analysis, the primary studies included in the architectural synthesis were assigned identifiers (P1–P15), as shown in Table VII. These identifiers are subsequently used in Table VIII to indicate the studies supporting each architectural layer of the conceptual reference architecture.

To improve transparency in the derivation of the conceptual reference architecture, Table VIII presents the relationship between each architectural layer and the primary studies that reported at least one of its constituent components. Because each layer may comprise multiple architectural elements, the supporting studies do not necessarily report all components listed within a given layer. Only primary studies were considered in this mapping to avoid duplication of evidence from secondary reviews. The inclusion of an architectural layer in the reference model required evidence from one or more primary studies reporting the corresponding architectural components.

IV. Discussion

The comparative analysis of the primary studies revealed a convergence toward common architectural principles for blockchain-based Electronic Health Record (EHR) systems. Rather than operating as standalone solutions, blockchain platforms are increasingly integrated with interoperability services, distributed storage mechanisms, and security frameworks to support health information exchange [15]–[17].

Among the recurring patterns identified, permissioned blockchain platforms, particularly Hyperledger Fabric, were the most frequently reported technologies. Smart contracts, access control mechanisms, and audit capabilities were also consistently incorporated into the reviewed architectures, reflecting the need for governance, traceability, and controlled data sharing in healthcare environments [4], [5], [7], [9], [13].

The widespread adoption of hybrid storage architectures further indicates that blockchain is commonly employed as a coordination and verification layer, while complete Electronic Health Records remain stored in off-chain repositories. Similarly, interoperability emerged as a central architectural requirement, with HL7 FHIR and RESTful APIs representing the principal mechanisms for communication among heterogeneous healthcare systems [8], [9], [11], [12], [15], [16], [20].

Despite these recurring patterns, the available evidence remains largely based on architectural proposals, prototypes, and proof-of-concept implementations. The quality assessment indicated that relatively few studies reported implementation validation or performance evaluation, limiting conclusions regarding large-scale deployment and real-world effectiveness. Governance, regulatory compliance, interoperability across independent institutions, and long-term operational performance remain important research challenges [15]–[17].

The conceptual reference architecture synthesized in this review was derived from the recurring components identified across the primary studies and should be interpreted as a representation of patterns reported in the literature rather than as a validated implementation model. Furthermore, the review was restricted to studies published between 2021 and 2026 and to four scientific databases, which may have limited the inclusion of additional relevant studies.

V. Conclusion

This study presented a systematic literature review of blockchain-based architectures for Electronic Health Records (EHRs) published between 2021 and 2026. The analysis of the primary studies identified recurring architectural patterns, including the widespread adoption of permissioned blockchain platforms, smart contracts, interoperability services, and hybrid storage architectures.

Based on these findings, a conceptual reference architecture was synthesized to consolidate the principal functional components reported in the literature. Rather than representing a validated implementation, the proposed architecture summarizes the architectural patterns most frequently observed across blockchain-based EHR systems.

The main contributions of this review are the systematic comparison of blockchain-based EHR architectures, the identification of recurring architectural elements, and the synthesis of a conceptual reference architecture derived from evidence extracted from primary studies.

Although blockchain-based EHR research shows increasing architectural convergence, further empirical validation remains necessary, particularly regarding large-scale deployment, interoperability across healthcare institutions, governance, regulatory compliance, and real-world performance. The findings provide a consolidated reference for future research on secure and interoperable healthcare information systems.

TABLE VII
IDENTIFICATION OF THE PRIMARY STUDIES USED IN THE ARCHITECTURAL SYNTHESIS.

Code	Primary Study
P1	HealthChain [4]
P2	Smart Contract-Enabled Secure Sharing of Health Data [5]
P3	Patient-Centric Blockchain-Assisted Health Information Exchange Framework [6]
P4	MrBlock Framework [7]
P5	Blockchain-IPFS Framework [8]
P6	FabricMedChain [9]
P7	Blockchain-Based Cooperative Medical Records Management System [10]
P8	Permissioned Blockchain Network for Proactive Access Control [11]
P9	Cross Platform Blockchain Network for EHR Exchange [12]
P10	Leveraging Blockchain and IoMT for Secure and Interoperable EHRs [13]
P11	BlockIoT [14]
P12	Secure EMR Data Sharing using Blockchain and Proxy Re-encryption [15]
P13	Hybrid Blockchain-Edge Architecture for EHR Management [16]
P14	Secure Medical Record Sharing using Two-Fold Encryption [17]
P15	Blockchain-Based EHR Sharing Scheme with Data Privacy Verification [20]

TABLE VIII
TRACEABILITY BETWEEN THE CONCEPTUAL REFERENCE ARCHITECTURE AND THE PRIMARY STUDIES.

Architectural Layer	Core Components	Supporting Primary Studies
Application and Interoperability Layer	HL7 FHIR, RESTful APIs, Identity Management	P1, P3, P5, P7, P8, P9, P13, P15
Smart Contracts and Access Control Layer	Smart Contracts, RBAC/ABAC, Consent Management, Authentication	P2, P3, P4, P5, P6, P8, P10, P11, P12, P13, P14, P15
Permissioned Blockchain Network	Permissioned Blockchain, Audit Logging, Cryptographic Mechanisms	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P13, P15
Off-chain Storage Layer	Off-chain Repositories, Cloud Storage, IPFS	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15

REFERENCES

[1] H. S. A. Fang, T. H. Tan, Y. F. C. Tan, and C. J. M. Tan, "Blockchain Personal Health Records: Systematic Review," *Journal of Medical Internet Research*, vol. 23, no. 4, e25194, 2021.

[2] D. Elangovan, C. S. Long, F. S. Bakrin, C. S. Tan, K. W. Goh, S. F. Yeoh, M. J. Loy, Z. Hussain, K. S. Lee, A. C. Idris, and L. C. Ming, "The Use of Blockchain Technology in the Health Care Sector: Systematic Review," *JMIR Medical Informatics*, vol. 10, no. 2, e17278, 2022.

[3] A. L. A. Fonsêca, I. M. P. Barbalho, F. Fernandes, E. A. Júnior, D. A. P. Nagem, P. H. Cardoso, N. V. R. Veras, F. L. O. Farias, A. R. Lindquist, J. P. Q. Santos, A. H. F. Morais, J. Henriques, M. Lucena, and R. A. M. Valentim, "Blockchain in Health Information Systems: A Systematic Review," *International Journal of Environmental Research and Public Health*, vol. 21, 2024.

[4] Y. Xiao, B. Xu, W. Jiang, and Y. Wu, "The HealthChain Blockchain for Electronic Health Records: Development Study," *Journal of Medical Internet Research*, vol. 23, no. 1, e13556, 2021.

[5] P. Chinnasamy, A. Albakri, M. Khan, A. Ambeth Raja, A. Kiran, and J. Chinna Babu, "Smart Contract-Enabled Secure Sharing of Health Data for a Mobile Cloud-Based E-Health System," *Applied Sciences*, vol. 13, no. 6, 2023.

[6] K. P. Kalita, D. Boro, and D. K. Bhattacharyya, "A Patient-Centric Blockchain-Assisted Health Information Exchange Framework with Access Control," *Computers & Electrical Engineering*, vol. 122, 2025.

[7] P. Joshi and P. Mahajan, "Secure and Interoperable EHR Management via Hyperledger Fabric: The MrBlock Framework," *Peer-to-Peer Networking and Applications*, 2025.

[8] M. K. Singh, S. K. Pippal, and V. Sharma, "A Blockchain-IPFS Framework for Secure, Scalable, and Interoperable Healthcare Data Management," *SN Computer Science*, vol. 6, 2025.

[9] N. M. Sawant and J. Gomes, "A Decentralized Solution for Smarter and Safer Healthcare Systems with FabricMedChain," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 2026.

[10] S. Alyahya and Z. Almaghrabi, "Blockchain-Based Cooperative Medical Records Management System," *Future Internet*, vol. 17, 2025.

[11] E. Psarra, D. Apostolou, Y. Verginadis, I. Patiniotakis, and G. Mentzas, "Permissioned Blockchain Network for Proactive Access Control to Electronic Health Records," *BMC Medical Informatics and Decision Making*, vol. 24, 2024.

[12] R. G. Sonkamble, A. M. Bongale, S. Phansalkar, and D. S. Dharrao, "A Secure Interoperable Method for Electronic Health Records Exchange on Cross Platform Blockchain Network," *MethodsX*, vol. 13, 2024.

[13] S. Ben Othman and M. Getahun, "Leveraging Blockchain and IoMT for Secure and Interoperable Electronic Health Records," *Scientific Reports*, vol. 15, 2025.

[14] M. Shukla, J. Lin, and O. Seneviratne, "BlockIoT: Blockchain-Based Health Data Integration Using IoT Devices," in *Proceedings of the AMIA Annual Symposium*, 2021.

[15] G. Liu, H. Xie, W. Wang, and H. Huang, "A Secure and Efficient Electronic Medical Record Data Sharing Scheme Based on Blockchain and Proxy Re-encryption," *Journal of Cloud Computing*, vol. 13, 2024.

[16] H. Guo, W. Li, M. Nejad, and C.-C. Shen, "A Hybrid Blockchain-Edge Architecture for Electronic Health Records Management with Attribute-Based Cryptographic Mechanisms," *IEEE Transactions on Network and Service Management*, vol. 20, 2023.

[17] M. A. Habib, K. M. R. Alam, and Y. Morimoto, "A Secure Medical Record Sharing Scheme Based on Blockchain and Two-Fold Encryption," in *Proceedings of the 25th International Conference on Computer and Information Technology (ICCIT)*, 2023.

[18] R. Kumar, Arjunaditya, D. Singh, K. Srinivasan, and Y.-C. Hu, "AI-Powered Blockchain Technology for Public Health: A Contemporary Review, Open Challenges, and Future Research Directions," *Healthcare*, vol. 10, 2022.

[19] T. Shine, J. Thomason, I. Khan, M. Maher, K. Kurihara, and O. El-Hassan, "Blockchain in Healthcare: 2023 Predictions From Around the Globe," *Blockchain in Healthcare Today*, vol. 6, 2023.

[20] I. Boumezbeur and K. Zarour, "Blockchain-Based Electronic Health Records Sharing Scheme with Data Privacy Verifiable," *Applied Medical Informatics*, vol. 43, no. 4, pp. 1–13, 2021.

Arquitetura AlertaSUS: Sistema Hiperlocal de Alerta Epidemiológico Proativo com Preservação de Privacidade Baseada em Geohash

1ª Larissa M. Rodrigues
Dep. de Engª. Elétrica, Universidade Federal do Ceará
Fortaleza, Brasil
larissamendesr@gmail.com

2º Ênio R. Viana
Dep. de Engª. Elétrica, Universidade Federal do Ceará
Fortaleza, Brasil
eniocc@gmail.com

3º Francisco G. D. da Silva Filho
Dep. de Engª. Elétrica, Universidade Federal do Ceará
Fortaleza, Brasil
gdias27@alu.ufc.br

4º Victor Hugo C. de Albuquerque
DETI, Universidade Federal do Ceará
Fortaleza, Brasil
victor.albuquerque@lesc.ufc.br

5º Rosineide da Paz
Lab. de Tecnologias Inovadoras, Universidade Federal do Ceará
Russas, Brasil
rfdapaz@ufc.br

***Resumo*—A Vigilância Epidemiológica do Sistema Único de Saúde (SUS) tem incorporado cada vez mais sistemas inteligentes para prevenção de surtos, porém a lacuna de transformação do monitoramento dos dados em canais de comunicação de forma direta e ágil com a população sob risco ainda persiste. Embora o Brasil utilize plataformas como o Infodengue, fundamentadas em modelagem preditivo-estatística, suas saídas permanecem centralizadas em painéis de análise para especialistas e restritas aos parâmetros climáticos como "gatilhos"epidemiológicos, não sendo a realidade de algumas cidades brasileiras. A arquitetura AlertaSUS, proposta neste artigo, expande esse escopo ao descentralizar o fluxo decisório por meio de alertas hiperlocais direcionados a microrregiões específicas, parametrizadas por modelos de *nowcasting* e *forecasting* (Infodengue), como também pela Lei Geral de Proteção de Dados (LGPD). A conformidade jurídica e a segurança dos dados são estruturadas sobre um protocolo de acoplamento cego (*blind join*) em polígonos de risco, indexados via Geohash, inviabilizando tentativas de reidentificação ou engenharia reversa de perfis clínicos individuais no território.**

***Index Terms*—SUS, epidemiologia, alertas, geohash.**

I. Introdução

No Brasil existem ferramentas de monitoramento epidemiológico, como o Infodengue da Fundação Oswaldo Cruz e da Faculdade Getúlio Vargas, que oferecem suporte analítico em âmbito municipal. Contudo, surtos recentes de dengue evidenciaram assimetrias espaciais e dinâmicas de transmissão que divergem dos padrões históricos, o que demanda flexibilização metodológica nos modelos vigentes, como integrar preditores adicionais não climáticos, para mitigar falhas de previsão em microescala. Ademais, o fluxo de comunicação do Infodengue envia alertas para especialistas previamente cadastrados e as informações da plataforma requer análise gráfica de dados. Outro fator analisado é que o alerta gerado pelo Infodengue mapeia o risco de forma agregada para o município, limitando o envio de alertas específicos para divisões intraurbanas menores, pois dentro do município existem áreas com infestação positiva para o vetor e outras não. Posto isso, o AlertaSUS surge em caráter complementar ao Infodengue, que é fundamental na orientação e emissão de alertas para as secretarias estaduais e municipais.

Além das diferenças espaciais observadas na dinâmica de transmissão, a relevância dos fatores epidemiológicos varia conforme a escala geográfica. Os maiores pesos de entropia foram associados ao sorotipo, ao *El Niño* e ao Índice de Gini no Brasil; o *El Niño*, Índice de Gini e turismo na Região Sul; o *El Niño*, sorotipo e precipitação no Ceará; e o *El Niño*, sorotipo e turismo em Fortaleza. Esses resultados sugerem que gatilhos epidemiológicos definidos em escala nacional ou estadual não devem ser necessariamente generalizados para escalas municipais ou intraurbanas. Essa heterogeneidade constitui uma das motivações para a adoção de gatilhos configuráveis regionalmente no AlertaSUS.

Este trabalho investiga a viabilidade técnica de uma arquitetura de comunicação hiperlocal, operando sob demanda e em múltiplos níveis de granularidade geográfica (bairro, rua, residênci), integrando dados dinâmicos de transmissão biológica e eventos epidemiológicos à estrutura do AlertaSUS. O modelo foi instanciado para o monitoramento de arboviroses, em complementaridade ao Infodengue, com ampliação dos tipos de alertas (manuais e automáticos) para indicadores exógenos (como fluxo populacional e sorotipos), sob validação final de um especialista. Ao estabelecer esse fluxo bidirecional, mitiga-se o *gap* de comunicação com o cidadão por meio de

mensagens diretas e estruturadas sob a segurança e privacidade dos dados. As plataformas de vigilância existentes são eficazes para monitorar a dinâmica de doenças, mas não oferecem disseminação proativa e centrada no cidadão, ao mesmo tempo em que atendem aos requisitos de privacidade derivados da LGPD. Essa lacuna motivou a arquitetura proposta, que integra mecanismos de preservação da privacidade, indexação espacial hiperlocal (uma integração de Geohash) e comunicação voltada ao cidadão, em uma estrutura unificada, projetada para complementar as infraestruturas de vigilância existentes.

II. Referencial Teórico

Sistemas de alerta epidemiológico com componente geoespacial existem em outros contextos internacionais. O *HealthMap* (*H*arvard) agrega dados abertos e mídias sociais para detecção de surtos globais, porém sem hiperlocalidade. O *Early Warning and Response System*(EWARS) da Organização Mundial da Saúde (OMS) opera em contextos de crise humanitária, sem comunicação direta ao cidadão. Nenhum desses sistemas combina (i) hiperlocalidade (ii) comunicação proativa ao cidadão, o que configura o diferencial arquitetural do AlertaSUS.

Infodengue é um sistema de alerta precoce por meio de monitoramento de transmissão de arbovírus (dengue, *chikungunya* e *zika*) e opera em todos os municípios. Visa antecipar alterações nos padrões de notificação de arbovírus, para que os profissionais de vigilância sejam informados rapidamente sobre os riscos de epidemias [1]. O sistema produz estimativas dos casos corrigindo atrasos de notificação e com diferentes fontes de informação. O sistema coleta dados, integra bases de dados, realiza a modelagem estatística (ciclo reprodutivo do vetor e clima) e gera alertas semanais. O sistema contempla casos notificados de dengue, dados climáticos, séries temporais e localização geográfica. O Infodengue considera fatores ambientais que influenciam o ciclo do mosquito como: temperatura média, umidade relativa e pressão atmosférica. Já as variáveis de transmissão, considera o Número Reprodutivo Efetivo (Rt), que representa quantas pessoas, em média, são infectadas por um caso existente, sendo considerada as seguintes classificações: tendência de redução da transmissão (Rt < 1); transmissão estável (Rt = 1) e crescimento da epidemia (Rt > 1) [2]. Utiliza a classificação de alertas em quatro níveis: verde (baixa transmissão), amarelo (condições favoráveis à transmissão), laranja (transmissão sustentada), vermelho (alta incidência / epidemia) [1]. Posto isso, com a integração dos sistemas, o Infodengue representa um importante aliado ao AlertaSUS.

Entre 2015 e 2024, o Brasil registrou surtos de dengue de forma atípica, que limitaram a aplicabilidade do Infodengue como único sistema de alerta: (i) o surto de 2015 no Ceará, precedido por seca severa e não por chuvas elevadas; (ii) o recorde de 2024 no Brasil, com menor precipitação anual histórica (1.480 mm); (iii) a introdução simultânea de quatro sorotipos em 2024 no Brasil, gerando reinfecções em massa independentemente das condições climáticas. Nesses cenários, fatores como o sorotipo circulante e o fluxo de pessoas, variáveis ausentes no Infodengue, foram determinantes para a explosão de casos, reforçando a necessidade de um sistema complementar com variáveis epidemiológicos locais configuráveis.

No AlertaSUS, os alertas criados pelos epidemiologistas, possibilitam uma flexibilização para análise de outros fatores (não somente climáticos). O sistema hiperlocal permite o envio de alertas direcionados, a partir de filtros por residência, rua ou bairro, pois cerca de 75% dos focos do vetor estão dentro das residências [3]. O controle e o monitoramento nesses focos são realizados pelos Agentes de Combate à Endemias (ACE), profissionais do SUS, por meio de inspeções, sendo oportuno utilizar o alerta com hiperlocalidade para áreas com focos positivados.

As epidemias de dengue têm várias causas, incluindo o aumento da temperatura e da chuva, saneamento inadequado, existência de criadouros domésticos, resistência do vetor aos inseticidas e a presença de quatro sorotipos circulando simultaneamente em países endêmicos, o que favorece a reinfecção [4]. Essas causas têm contribuído para uma maior incidência de dengue ao longo dos anos [5]. A partir da evolução dos casos prováveis de dengue no Brasil, mostra que praticamente nenhuma área está protegida contra a dengue [6]. Alterações climáticas e meteorológicas (temperatura, precipitação, umidade e *El Niño*) são fatores críticos que afetam a reprodução, a sobrevivência e a distribuição geográfica do vetor, influenciando na transmissão de patógenos *Aedes* [4]. Desta forma, a aplicação de modelos preditivos, a exemplo do Infodengue, viabiliza o monitoramento sistemático das variáveis climáticas.

Um fator primordial que transcende a modelagem computacional preditiva é a privacidade e segurança dos dados. A viabilidade operacional da arquitetura está condicionada à preservação da identidade dos usuários cadastrados, alinhada com a LGPD [7]. O risco de ataques de reidentificação geográfica é mitigado no AlertaSUS pela exclusão de chaves relacionais associadas à identificação civil dos usuários, pois a correlação entre os dados clínicos anonimizados de origem e o canal de notificação ocorre em um plano de segregação lógica pelo parâmetro denotado como $ID_{Espacial}$, que funciona como uma função de agregação poligonal indexada via estrutura de *Geohash*. O emparelhamento seguro de dados ocorre como uma operação de interseção conjuntista cega ou junção cega (*Blind Join*), expressa na Equação 1:

$$\mathcal{J}_{\text{cega}} = \{u \in \mathcal{U}_{msg} \mid \exists\, c \in \mathcal{C}_{epi} : f_{space}(u) = f_{space}(c) \wedge \mathcal{R}(c) > \gamma\} \tag{1}$$

A Equação 1 define a operação de junção cega, cujo objetivo é identificar o conjunto de usuários aptos a receber alertas sem a necessidade de compartilhamento de informações pessoais entre os domínios envolvidos. A operação baseia-se na correspondência do identificador espacial anonimizado e na verificação do nível de risco epidemiológico local. Na Equação 1:

- $\mathcal{J}_{\text{cega}}$ representa o conjunto resultante da operação de junção cega entre os usuários cadastrados e as notificações epidemiológicas;
- $\mathcal{U}_{msg}$ representa o conjunto de usuários cadastrados na plataforma, que autorizaram previamente o recebimento de alertas por meio do protocolo de consentimento explícito (*opt-in*);
- $\mathcal{C}_{epi}$ representa o conjunto de notificações epidemiológicas anonimizadas provenientes dos sistemas oficiais de vigilância;
- u corresponde a um usuário pertencente ao conjunto $\mathcal{U}_{msg}$;
- c representa uma notificação epidemiológica pertencente ao conjunto $\mathcal{C}_{epi}$;
- $f_{space}(\cdot)$ corresponde à função de mapeamento espacial responsável por transformar tanto os usuários quanto as notificações em um mesmo identificador geográfico anonimizado (por exemplo, um Geohash);
- $\mathcal{R}(c)$ representa a função de avaliação do risco epidemiológico associado à notificação; expressa a função de risco local ponderada em relação ao limiar epidemiológico crítico γ;
- γ representa o limiar mínimo de risco estabelecido para que um alerta seja disparado.

O conjunto $\mathcal{J}_{\text{cega}}$ é formado por todos os usuários u cadastrados para receber mensagens, para os quais existe pelo menos uma notificação epidemiológica c pertencente à mesma região espacial anonimizada e cujo risco epidemiológico seja superior ao limiar crítico γ. Desse modo, o fluxo de dados direcionado ao *gateway* de mensageria preserva o estado clínico individual, mitigando a superfície de exposição de dados sensíveis em conformidade com as diretrizes da LGPD.

III. Metodologia

Na pesquisa foi realizado um levantamento do número de casos de dengue e os fatores epidemiológicos por ano (*El Niño*, temperatura média, umidade relativa do ar, precipitação pluviométrica média, saneamento básico, Índice de Gini, sorotipos, densidade demográfica, taxa de incidência e fluxo de turista), no período de 2000 até 2025, por meio do acesso aos seguintes sistemas:

(a) Casos prováveis de dengue: Portal de Dados Abertos do SUS [6];
(b) Variáveis climáticas: Inmet [8];
(c) Saneamento básico: Sistema Nacional de Informações sobre Saneamento (SNIS) [9];
(d) Condições socioeconômicas: Instituto Brasileiro de Geografia e Estatística (IBGE) e o Instituto de Pesquisa Econômica Aplicada (IPEA) [10];
(e) *El Niño*: utilizada a métrica *Index* ONI (*Oceanic Niño Index*), fornecida pela Administração Oceanográfica e Atmosférica Nacional dos EUA - NOAA), que mede a anomalia da temperatura da superfície do mar na região do Pacífico Equatorial [11];
(f) Sorotipos: histórico do comportamento e surgimento dos sorotipos no Brasil, baseado nos dados de vigilância genômica e isolamento viral da Fiocruz e do Ministério da Saúde [12];
(g) Densidade demográfica: a evolução estimada ano a ano é baseada nos Censos de 2000, 2010, 2022 e nas projeções intercensitárias oficiais do IBGE [13];
(h) Índice de Turismo (internacional e inter-regionais): Anuário Estatístico de Turismo, gerado pelo Ministério do Turismo (MTur) [14];

As metodologias empregadas no Infodengue são a *Nowcasting*, com modelos *bayesianos* para estimar casos que provavelmente já ocorreram, mesmo não oficialmente registrados, como também o cálculo semanal do Rt para verificar a situação da transmissão.

O mecanismo de k-anonimidade espacial foi configurado com k = 5, de modo que uma célula geográfica somente permanece na granularidade original quando contém pelo menos cinco usuários. Para a indexação espacial foi utilizado Geohash com precisão de sete caracteres, (aprox. 150 x 150 metros), com remoção progressiva de caracteres para 6 ou menos) sempre que o volume populacional na célula é inferior a k. Então, quando o número de usuários na célula é inferior ao limiar k, a célula é generalizada para o nível imediatamente superior, reduzindo a resolução espacial e o risco de singularidade. Na prática do *Geohash*, generalização por k-anonimidade (6 caracteres - ex: 7h4v2y), remove o último caractere, a área expande e isso engloba uma vizinhança maior, garantindo que existam pelo menos k pessoas (5 ou mais) naquela mesma célula e impedindo a reidentificação individual dos cidadãos, cumprindo as exigências da LGPD.

Para o levantamento, tratamento de dados, cálculo e testes estatísticos, foi utilizada a linguagem de programação *Python* e o editor de código-fonte *Visual Studio Code*.

A. Ferramentas estatísticas

A metodologia proposta combina o Método dos Pesos de Entropia (*EWM - Entropy Weight Method*) para avaliar a dispersão da informação interna das variáveis explicativas com o Coeficiente de Correlação de Postos de Spearman para identificar a direção e força da relação com a variável alvo. O processamento computacional dos fatores epidemiológicos adota uma abordagem combinada para mitigar vieses de escala e capturar interações não lineares. Inicialmente, para equalizar as grandezas das variáveis ambientais e socioeconômicas, aplica-se a transformação Min-Max, mapeando os valores observados no intervalo $[0, 1]$, conforme a Equação 2:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} + \epsilon \tag{2}$$

Posteriormente, a mensuração do grau de ordenação fundamenta-se na Entropia de Shannon, que mede o grau de ordenação ou aleatoriedade dos dados, com n = 26 anos (período de 2000 até 2025), por meio da Equação 3:

$$e_j = -\frac{1}{\ln(n)} \sum_{i=1}^{n} p_{ij} \ln(p_{ij}) \tag{3}$$

O gradiente de divergência mapeia a utilidade intrínseca da informação. Quanto maior a divergência (1 - ej), maior o peso final da variável.Onde m é o número total de variáveis explicativas (m = 9). Equação 4:

$$d_j = 1 - e_j, \quad w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{4}$$

A entropia avalia apenas a estrutura interna da variável, sendo necessária a correlação de Spearman ρ para capturar associações, lineares ou não com o total de casos de dengue no Brasil. O sinal do coeficiente determina a classificação final, com "Aumento (+)"ou "Redução (-)", conforme Equação 5:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{5}$$

O EWM é fundamentado na Teoria da Informação de Shannon e esse indicador atribui maior peso às variáveis com maior variabilidade e menor redundância [15]. Complementarmente, foi empregado o coeficiente de correlação para avaliar a direção e a intensidade da associação entre cada variável e os casos de dengue na região selecionada, indicando relações positivas quando ambas variam no mesmo sentido e negativas quando variam em sentidos opostos [16].

A detecção de multicolinearidade também foi realizado no estudo das variáveis, para corrigir problemas quando o modelo não consegue separar o impacto individual de uma variável sobre a variável alvo, gerando intepretações equivocadas e distorção das Entropias, com as variáveis colineares funciona como um "voto duplo". A detecção foi realizada por meio do Fator de Inflação da Variância (*Variance Inflation Factor* - VIF), que é a métrica que calcula o quanto a variância de um coeficiente está inflada devido à colinearidade com os outros preditores, conforme Equação 6. Para a classificação do valor de VIF, adotou-se o critério sugerido por Hair et al. (2019) e Montgomery et al. (2021): nenhuma correlação (VIF=1); colinearidade moderada (VIF> 5) e multicolinearidade crítica (VIF> 10).

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{6}$$

B. Riscos epidemiológicos e limitações metodológicas

Foram identificados três potenciais riscos metodológicos inerentes ao estud: (i) agregação espaço-temporal anual e consequente viés ecológico; (ii) a dependência temporal e a sazonalidade das séries epidemiológicas; e (iii) a possibilidade de correlações espúrias decorrentes de tendências temporais comuns e da multicolinearidade entre variáveis explicativas.

O primeiro risco está relacionado à utilização de dados agregados em escala anual, o que pode reduzir a capacidade de representar adequadamente a dinâmica temporal da transmissão da dengue, pois a biologia do vetor (*Aedes aegypti)* responde à variações semanais ou mensais e assim perde-se o sinal dos picos de transmissão na escala anual. Entretanto, a utilização de dados anuais teve caráter apenas exploratório, para uma prévia visualização do comportamento dos fatores epidemiológicos em diferentes regiões do país e esse risco é mitigado pela integração do AlertaSUS ao Infodengue, cuja base de dados é estruturada por semana epidemiológica, permitindo capturar com maior a variabilidade espaço-temporal da transmissão.

O segundo risco decorre da estrutura temporal das séries epidemiológicas e climáticas, que violam a independência entre observações. A dengue apresenta forte sazonalidade, além de elevada autocorrelação temporal, pois está intimamente relacionado ao comportamento epidemiológico dos períodos anteriores. Esse padrão pode ser explicado pela circulação viral, pelas condições ambientais favoráveis à proliferação do vetor e pela dinâmica de imunidade populacional, fatores que conferem memória temporal às séries analisadas. Como forma de mitigação, o modelo prevê a participação de epidemiologistas na emissão de alertas, permitindo a incorporação de conhecimento contextual e a validação dos resultados no sistema.

O terceiro risco está associado à correlações espúrias entre variáveis. Tanto o cálculo da Entropia da Informação quanto o coeficiente de correlação de Spearman não consideram explicitamente estruturas de defasagem temporal (*lags*), podendo atribuir elevada relevância estatística a variáveis que compartilham apenas uma tendência temporal comum, sem necessariamente apresentarem relação causal. Para mitigar essa limitação, adotou-se um protocolo de filtragem parametrizado via VIF com limiar de corte (VIF > 5). Sob esse critério, foram descartadas da regressão subsequente as variáveis que exibiram multicolinearidade crítica: Densidade Demográfica, Cobertura de Abastecimento de Água e o Índice de Gini. Adicionalmente, as variáveis Umidade Relativa do Ar e Precipitação Pluviométrica foram tratadas de forma mutuamente exclusiva, retendo-se apenas a precipitação como preditor dinâmico do modelo em virtude de sua menor inflação de variância residual (VIF $< 3,1$). Dessa forma, o conjunto final de preditores restringiu-se às variáveis não redundantes: Temperatura Média, Intensidade do *El Niño* e Sorotipo Viral Predominante.

C. Estrutura de privacidade AlertaSUS

A estrutura proposta para privacidade do AlertaSUS é baseado na técnica de Segregação Lógica de Responsabilidades com Acoplamento Cego via Pseudonimização Espacial. O sistema opera a partir de duas bases de dados isoladas logicamente: base epidemiológica e a base de comunicação. A base epidemiológica é submetida a algoritmos de k-anonimidade espacial, com as coordenadas de geolocalização dos pacientes indexados por um identificador geográfico (ID_Espacial), o que promove a mitigação de risco. Já a base de comunicação gerencia dados identificáveis de contato do cidadão, coletados via consentimento livre (*opt-in*). O anonimato clínico em todo processo reduz o risco de operadores humanos da plataforma ou *gateways* terceiros de mensageria instantânea estabeleçam correlações causais entre a identidade do cidadão notificado e o vetor de infecção local. Com base nesse cenário, a Figura 1

apresenta a arquitetura de dados AlertaSUS, por meio de um diagrama de blocos e dividida em camadas funcionais.

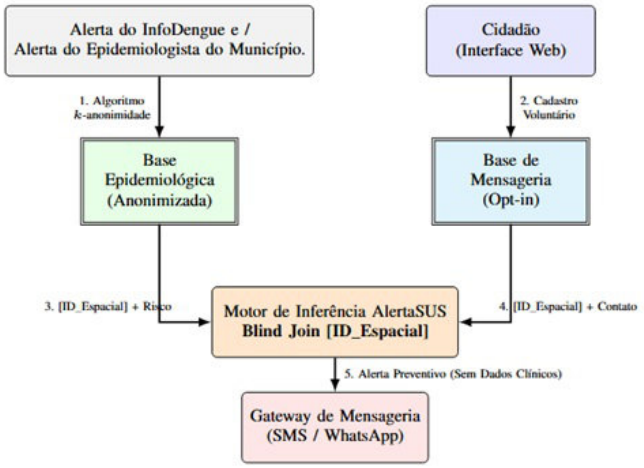


Figura 1. Diagrama de Arquitetura de Dados do AlertaSUS: segregação lógica de bases e acoplamento cego via ponte espacial (*Spatial Bridge*)

O diagrama de blocos proposto operacionaliza o fluxo de informações do sistema AlertaSUS, dividindo-o em quatro camadas interdependentes, porém isoladas logicamente:

(i) Camada Fontes de Dados (*Input Layer)*: é a entrada do sistema, com duas fontes de naturezas distintas, que operam de forma assíncrona. O Fluxo Epidemiológico é composto por alertas automáticos orinundos do Infodengue e alertas manuais emitidos pelos epidemiologistas dos municípios. Antes de ingressar no banco do AlertaSUS, esses dados passam pelo Filtro de k-Anonimidade Espacial. O endereço linear do paciente é descartado na origem e transformado em um código de área (ex: *Geohash*). Se o número de residentes nessa área for menor que o limiar k, a célula é generalizada para um nível macro, impedindo que o paciente seja identificado por exclusão geográfica. O bloco Cidadão e o Fluxo de Consentimento é a interface pública onde o cidadão interage voluntariamente com o sistema, por meio do fluxo de *opt-in*, o usuário fornece seu contato (telefone, *e-mail*, endereço).

(ii) Camada de Persistência Segregada: o AlertaSUS utiliza duas bases logicamente isoladas, uma Base Epidemiológica Anonimizada que contém apenas registros tabulares contendo a série temporal de casos, variáveis epidemiológicas e o ID_Espacial, não contemplando chaves primárias que remetam a indivíduos. A Base de Mensageria (*Opt-in*), que contém os dados de contato dos usuários, pareados unicamente com o ID_Espacial escolhido, ignora qualquer dado clínico.

(iii) Camada de Processamento: O ID_Espacial como Ponte é o motor de inferência do AlertaSUS e atua no centro do diagrama, executando a Junção Cega (*Blind Join*). O motor de agentes varre a Base Epidemiológica e calcula se o volume de casos ou variáveis epidemiológicas que viola os limiares de segurança estatística. Para um cenário de risco em uma célula (ex: ID_Espacial= geo_7b2a), o motor emite uma instrução de disparo restrita a esse *token* geográfico. O motor realiza a intersecção matemática, busca na base de mensageria quais os contatos estão vinculados à *string* geo_7b2a. A quebra do vínculo causal é por meio do cruzamento puramente topológico. O sistema emparelha um "índice de risco territorial"com uma "lista de usuários do território". O motor funciona sob o princípio Segregação Lógica de Responsabilidades com Acoplamento Cego via Pseudonimização Espacial, com ação preditiva sem que cruze a identificação da pessoa doente com o destinatário da mensagem.

(iv) Camada de Saída: o *gateway* de mensageria centraliza o disparo de mensagens via *e-mail* e redes sociais. Como o tráfego de dados processado pelo motor de inferência envia apenas o contato e o texto de prevenção ("Atenção: sua região registrou aumento de casos de dengue."), nenhuma informação médica trafega por redes terceirizadas, fortalecendo o sistema contra vazamentos de dados sensíveis.

D. Avaliação Quantitativa da Privacidade e Hiperlocalidade

Foram implementadas quatro métricas para quantificar o grau de privacidade:

(i) Número médio de usuários por célula;

$$R_s = \frac{C_{n<k}}{C_{\text{total}}} \tag{7}$$

(ii) Percentual de células generalizadas;

$$P_{gen} = \frac{C_{gen}}{C_{\text{total}}} \times 100 \tag{8}$$

(iii) Perda de resolução espacial, com comparação entre a resolução original com a resolução após generalização.

$$\bar{n} = \frac{1}{C} \sum_{i=1}^{C} n_i \tag{9}$$

Conforme apresentado na Equação 7, o risco de singularidade é calculado pela proporção de células com população inferior ao parâmetro k.A proporção de células generalizadas é calculada conforme a Equação 8.O risco pós generalização é calculada conforme a Equação 9.

Com base nas simulações de densidade populacional e nos parâmetros de k-anonimidade e precisão de Geohash adotados no estudo, foi simulado para regiões de menor densidade e preenchidos na Tabela I.

Para Geohash 5 (aprox. 4,9 km x 4,9 km) há uma baixíssima generalização, pois agrupa grandes blocos populacionais, garantindo alta média de usuários por célula, porém sacrificando a hiperlocalidade (rua/bairro).Caso da Geohash 6 (aprox. 1,2 km por 0,6 km) equilibra o dimensionamento de bairros, onde o aumento de k de 3 para 10 eleva expressivamente a necessidade de generalização (aprox. 12,4 % para 45,6%). Já a Precisão de Geohash 7 (aprox. 150 m por 150 m -

Tabela I
ANÁLISE DE SENSIBILIDADE DA PRECISÃO ESPACIAL E DO PARÂMETRO k.

Geohash	k	Células generalizadas (%)	Usuários/célula
5	3	4,20%	42,6
5	5	8,50%	42,6
5	10	18,10%	42,6
6	3	12,40%	11,2
6	5	24,80%	11,2
6	10	45,60%	11,2
7	3	31,50%	3,8
7	5	58,20%	3,8
7	10	81,90%	3,8

adotada no protótipo), garante a resolução hiperlocal (micro-quadra/residência), mas sofre maior impacto em áreas de baixa densidade, exigindo que células com menos de k usuários sejam generalizadas (alcançando $58,2\%$ de generalização para $k = 5$).

IV. RESULTADOS

A pesquisa envolveu a mensuração fatores epidemiológicos associados ao número de casos prováveis de dengue no Brasil, baseada no conceito de Entropia da Informação de Shannon e os Coeficiente de Correlação de Postos de Spearman, no período de 2000 até 2025. A análise dos dados foi especificamente para o Brasil e Regiões (Norte, Sul, Nordeste, Sudeste, Centro-Oeste), Ceará e Fortaleza, sendo apresentadas no artigo apenas as matrizes que mostraram maior divergência, por limitação de espaço. A Figura 2 e 3 apresentam os resultados da Matriz de pesos de entropia e coeficiente de correlação do Brasil e da Região Sul, como também do Estado do Ceará e Município de Fortaleza.

```
==================== MATRIZ DE PESOS: CONSOLIDADO BRASIL (MIN-MAX) ====================
Fator Epidemiológico (Brasil) Peso de Entropia (w_j) Impacto na Dengue Nacional Coef. Correlação
            Sorotipo_Num_BR                0.2190          Redução (-) no BR           -0.110
                El_Nino_Num                0.1924          Aumento (+) no BR            0.306
                    Gini_BR                0.1357          Redução (-) no BR           -0.727
                   Pobreza_BR              0.1280          Redução (-) no BR           -0.703
           Agua_Nacional_BR                0.1140          Aumento (+) no BR            0.611
              Temp_Media_BR                0.0562          Aumento (+) no BR            0.685
                Densidade_BR               0.0509          Aumento (+) no BR            0.605
             Precipitacao_BR               0.0448          Redução (-) no BR           -0.493
               Umid_Media_BR               0.0360          Redução (-) no BR           -0.552
                  Turistas_BR              0.0228          Aumento (+) no BR            0.245
================================================================================
 ==================== MATRIZ DE PESOS: REGIÃO SUL (MIN-MAX) ====================
   Variável (Sul) Peso de Entropia (w_j) Impacto na Dengue Regional Coef. Correlação
      El_Nino_Num              0.2577          Aumento (+) no Sul            0.123
         Gini_Sul              0.1609          Redução (-) no Sul           -0.841
      Turismo_Sul              0.1246          Aumento (+) no Sul            0.725
 Sorotipo_Num_Sul              0.0982          Redução (-) no Sul           -0.228
  Agua_Urbana_Sul              0.0835          Aumento (+) no Sul            0.887
    Temp_Media_Sul             0.0832          Aumento (+) no Sul            0.617
     Densidade_Sul             0.0816          Aumento (+) no Sul            0.878
        Chuva_Sul              0.0689          Aumento (+) no Sul            0.136
      Umidade_Sul              0.0415          Redução (-) no Sul           -0.400
 ================================================================================
```

Figura 2. Matriz de pesos de entropia e coeficiente de correlação Brasil e da Região Sul.

Na Figura 2, as variáveis de maior entropia foram Sorotipo, *El Niño* e Índice Gini (grau de desigaldade social). Já no Sul as variáveis de maior entropia,foram *El Niño*, Índice Gini e Turismo. Na Figura 3, o Ceará as variáveis de maior entropia, foram *El Niño*, sorotipo e chuva. Em Fortaleza, as variáveis de maior entropia, foram *El Niño*, Sorotipo e Turismo. Então, a análise individual das variáveis evidenciou comportamentos assimétricos entre as escalas nacional, regional, municipal e estadual.

```
================== MATRIZ DE PESOS: ESTADO DO CEARÁ (MIN-MAX) ==================
  Variável (Ceará) Peso de Entropia (w_j) Impacto na Dengue Estadual Coef. Correlação
       El_Nino_Num              0.2949          Aumento (+) no CE            0.074
Sorotipo_Num_Ceara              0.2618          Redução (-) no CE           -0.078
       Chuva_Ceara              0.0739          Redução (-) no CE           -0.291
     Umidade_Ceara              0.0738          Redução (-) no CE           -0.187
  Temp_Media_Ceara              0.0668          Aumento (+) no CE            0.007
      Gini_Ceara                0.0634          Aumento (+) no CE            0.017
 Agua_Urbana_Ceara              0.0583          Redução (-) no CE           -0.300
    Turismo_Ceara               0.0540          Redução (-) no CE           -0.065
   Densidade_Ceara              0.0530          Redução (-) no CE           -0.056
================================================================================
================== MATRIZ DE PESOS: MUNICÍPIO DE FORTALEZA (MIN-MAX) ==================
Variável (Fortaleza) Peso de Entropia (w_j) Impacto na Dengue Municipal Coef. Correlação
        El_Nino_Num              0.2572    Redução (-) em Fortaleza           -0.210
   Sorotipo_Num_Fort             0.2283    Redução (-) em Fortaleza           -0.205
       Turismo_Fort              0.1027    Aumento (+) em Fortaleza            0.065
     Densidade_Fort              0.0824    Redução (-) em Fortaleza           -0.114
         Gini_Fort               0.0824    Redução (-) em Fortaleza           -0.114
        Chuva_Fort               0.0706    Redução (-) em Fortaleza           -0.169
      Umidade_Fort               0.0627    Redução (-) em Fortaleza           -0.054
    Agua_Urbana_Fort             0.0610    Aumento (+) em Fortaleza            0.090
     Temp_Media_Fort             0.0526    Redução (-) em Fortaleza           -0.086
================================================================================
```

Figura 3. Matriz de pesos de entropia e coeficiente de correlação do Ceará e de Fortaleza.

A precipitação pluviométrica demonstrou não linearidade em relação aos surtos históricos, pois os maiores índices de dengue no país foram registradas em anos com baixo índice de chuva, a exemplo do recorde histórico de 2024 (1.480 mm). Essa relação paradoxal é explicada pela mudança comportamental da população durante períodos de estiagem, forçando o armazenamento improvisado de água e multiplicando criadouros artificiais [3].

A umidade relativa do ar, embora mantendo médias anuais elevadas e favoráveis em âmbito nacional (acima de 74%), não demonstrou correlação linear com as grandes explosões nacionais de casos, sugerindo que o limiar térmico e criadouros artificiais exercem maior dominância na transmissão.

O *El Niño* opera de forma heterogênea nas regiões do país, pois enquanto no Norte e Nordeste induz secas severas, com aumento de armazenamento de água inadequado, no Sul e Sudeste induz tempestades que geram acúmulo de água em áreas abertas e também anomalias térmicas que aceleram o período de incubação extrínseca do vírus.

Do ponto de vista socioeconômico e estrutural, indicadores como saneamento básico (abastecimento de água) e desigualdade de renda (Índice de Gini) revelaram limitações quando agregados na escala nacional. Apesar da expansão gradativa da rede de água (84,9%) [17], o volume de casos continuou elevado, o que reflete a dominância das anomalias climáticas sobre os avanços sociais, saneamento precário, densidade demográfica assimétrica, que são suavizadas em análises globais. [18], [19].

O sorotipo predominante revelou-se o fator de maior entropia nos picos de transmissão. A introdução de um sorotipo suscetível em uma população previamente imunizada para outras cepas rege o comportamento epidêmico do país, sobrepondo-se temporariamente às condições ambientais. Esse comportamento é catalisado pelo fluxo turístico, sendo o principal vetor de dispersão do patógeno, acelerando a introdução de novos genótipos. [20].

Nas Figuras 2 e 3, foi calculado o VIF, que funciona como um "indicador de redundância". Para o Brasil, as variáveis de densidade, índice Gini e abastecimento de água exibiram

VIF elevados (VIF $>$ 10), pois a população e a cobertura de água cresceram de forma gradativa, enquanto a pobreza e a desigualdade caíram. As variáveis precipitação e umidade são interligadas, com VIF alto, porque anos com muita chuva são anos com médias altas de umidade. Nas Figuras 4 e 5 apresentam os resultados da detecção da multicolinearidade dos dados epidemiológicos para o Brasil, Região Sul, estado do Ceará e para Fortaleza. Apesar do estudo ter avaliado as demais regiões do país, por limitação de espaço não foram apresentadas as demais tabelas.

Após removidas as variáveis com VIF acima de 5, foi aplicado o teste de significância estatística, que o indicador é o p-valor, que sugere o quão confiável é a relação entre duas variáveis. Dessa forma, avalia-se a correlação de alguns fatores epidemiológicos com o número de casos de dengue para confirmar se correlação não é apenas um ruído estatístico.

```
================== VALIDAÇÃO POR MODELO INFERENCIAL (OLS) ==================
==============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                    -5.806e+07   1.96e+07     -2.968      0.007   -9.86e+07   -1.75e+07
Temp_Media_BR             2.315e+06   7.79e+05      2.972      0.007       7e+05    3.93e+06
El_Nino_Global            2.186e+04   1.84e+05      0.119      0.906   -3.59e+05    4.03e+05
Turistas_Internacionais    92.6975    154.195       0.601      0.554    -227.083     412.478
==============================================================================================
==============================================================================
R-quadrado do Modelo (Capacidade Preditiva Explanatória): 0.398
```

Figura 4. Detecção de Multicolinearidade no Brasil.

```
================== VALIDAÇÃO POR MODELO INFERENCIAL (OLS) - SUL ==================
==================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            -9.656e+06   1.86e+06     -5.195      0.000   -1.35e+07   -5.79e+06
Temp_Media_Sul    4.236e+05   9.68e+04      4.377      0.000    2.22e+05    6.25e+05
El_Nino_Num      -8.12e+04    3.08e+04     -2.639      0.015   -1.45e+05   -1.72e+04
Chuva_Sul         1188.8370    291.708      4.075      0.001     582.196    1795.478
Sorotipo_Num_Sul  1.481e+04   2.73e+04      0.543      0.593   -4.19e+04    7.15e+04
==================================================================================
==================================================================================
R-quadrado do Modelo na Região Sul: 0.666

==============================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                     7.564e+05   7.11e+05      1.064      0.302   -7.43e+05    2.26e+06
El_Nino_Num              -785.5061    2179.153     -0.360      0.723   -5383.116    3812.104
Temp_Media_Fortaleza     -2.343e+04   1.93e+04     -1.216      0.241   -6.41e+04    1.72e+04
Umidade_Fortaleza        -3492.8740   3981.855     -0.877      0.393   -1.19e+04    4908.105
Chuva_Fortaleza              2.6973     11.810      0.228      0.822     -22.219      27.614
Agua_Urbana_Fortaleza     1196.0637   2528.680      0.473      0.642   -4138.985    6531.112
Gini_Fortaleza            6.174e+04   1.56e+05      0.396      0.697   -2.68e+05    3.91e+05
Sorotipo_Num_Fortaleza -1828.1950   2192.001     -0.834      0.416   -6452.912    2796.522
Turismo_Fortaleza         3088.5002   4314.916      0.716      0.484   -6015.176    1.22e+04
==============================================================================
```

Figura 5. Comparativo de Multicolinearidade das região Sul e o Município de Fortaleza

Analisando a matriz de pesos do Brasil, a temperatura tem relevância estatística para prever casos de dengue mesmo quando controlada por outros fatores (P> [t] = 0,007). O seu R-quadrado é de 0,398, o que demonstra que um modelo linear simples, utilizando apenas três variáveis, consegue explicar 39,8% de toda a variação histórica dos casos de dengue no Brasil entre 2000 e 2025. Para um modelo epidemiológico de escala nacional, anual e puramente ecológico, os outros 60% refletem fatores locais que o modelo macro não capta, como imunidade de rebanho para sorotipos específicos, índices municipais de infestação predial. Embora o fluxo de pessoas seja importante para a introdução de novos sorotipos, a variação volumétrica anual de casos no país como um todo é governada por fatores ambientais e vetoriais, e não pela flutuação de turistas, porém essa realidade nacional não se reproduz em outras regiões do Brasil, como por exemplo na região nordeste.

Na região Sul, o teste de significância mostrou que a dengue é um fenômeno quase puramente governado pelo clima. Enquanto no modelo nacional o R-quadrado foi de 40%, no Sul ele saltou para 66,6%. O Sul do Brasil é historicamente uma região de "fronteira ecológica"para o *Aedes aegypti* devido aos invernos intensos. A temperatura média anual e a precipitação acumulada atuaram como os preditores positivos dominantes. Ao contrário do Nordeste, onde o mosquito sobrevive o ano todo e a dinâmica depende mais de fatores sociais.

Diferente do resultado da Região Sul, em Fortaleza não apresentou preditores estatisticamente significantes. Esse resultado evidencia o contraste socioecológico da dinâmica de transmissão do vírus nas duas regiões. Enquanto no Sul o clima é um preditor determinante, a capital cearense apresenta estabilidade térmica perene em patamares ideais para o ciclo biológico do *Aedes aegypti* . Assim, a variabilidade climática macro-anual perde poder explicativo direto no modelo global. A dinâmica de transmissão da dengue em Fortaleza é complexa, multifatorial e não-linear, pois um aumento na temperatura não gera um aumento linear nos casos e variáveis como temperatura, umidade e chuva em Fortaleza são colineares. Este resultado prova que modelos tradicionais de regressão linear global são insuficientes para o monitoramento municipal, justificando a necessidade de arquitetura capaz de lidar com regras de decisão locais, interações não-lineares e comportamentos adaptativos, considerando todos os fatores epidemiológicos.

Os pesos de entropia calculados ajudam na configuração dos gatilhos no AlertaSUS. Em Fortaleza, onde *El Niño*(wj = 0,2949) e sorotipo (wj = 0,2283), o epidemiologista priorizaria alertas manuais. Na Região Sul,com temperatura (wj = 0,0832) e precipitação (wj = 0,0689), o sistema é configurado com gatilhos automáticos (Infodengue). A Figura 6 apresenta a Tela *Dashboards* de alertas.



Figura 6. Tela de Aprovação de Alertas.

Para avaliar a estabilidade do AlertaSUS, testou-se o comportamento do sistema variando o γ em uma faixa de $\pm 20\%$, para medir o impacto dessas oscilações no tempo de resposta e no surgimento de falsos alarmes. Os testes mostraram que o protocolo de acoplamento cego se manteve estável dentro desse intervalo e reduções severas em γ causam um disparo excessivo de mensagens, sem benefício algum. Por outro lado,

elevar demais o limiar atrasa o envio dos alertas. O ponto de equilíbrio ideal foi definido com Rt locais ($Rt \approx 1,2$), o que evitou alertas desnecessários e garantiu a precisão geográfica dos disparos.

Os testes realizados no protótipo confirmaram que a base epidemiológica fica anonimizada por blocos espaciais de k-anonimidade e o sistema apenas cruza o ID espacial do risco com os assinantes daquela célula espacial, assegurando que o fluxo de dados desconheça a condição clínica individual dos cidadãos localizados na célula geográfica afetada.

V. Conclusão

A previsão de surtos baseada apenas em variáveis climáticas possui limitações, pois em 2002 a temperatura e a chuva foram moderadas e houve uma explosão de casos no Brasil com a introdução do DENV-3 [21], como também entre 2013 e 2016 concentrou o maior fluxo de pessoas no país (Olimpíadas e Copa), o que facilitou a circulação de novos sorotipos no país. Então para minimizar as lacunas identificadas para alguns fatores epidemiológicos nos modelos preditivos atuais, o AlertaSUS incorpora a validação ativa do epidemiologista no fluxo de decisão e disparo de alertas.

A arquitetura do AlertaSUS possui um diferencial em não cruzar dados de saúde por meio de chaves diretas. Faz-se por meio de uma ponte espacial transparente (ID_Espacial), que se encontram apenas por meio do *hash* geográfico agregado, que desvincula os dados do usuário dos dados de saúde.

Este estudo apresenta algumas limitações, pois a arquitetura proposta precisa ser avaliada em secretarias de saúde, principalmente em relação ao desempenho, condições operacionais de larga escala, a integração com infraestruturas de saúde pública e a aplicabilidade do mecanismo de k-anonimidade espacial em municípios de pequeno porte, onde a densidade demográfica por célula geográfica pode ser insuficiente para garantir o limiar k configurado, exigindo generalização para células de maior área ou mecanismos alternativos de privacidade diferencial. Cabe delimitar que a validação é de natureza arquitetural e funcional, pois o fluxo de criação, aprovação e envio de alertas, e a integração com o Infodengue foi testada em ambiente de desenvolvimento.

A perspectiva é dar continuidade a esse projeto, contemplando as seguintes melhorias no AlertaSUS: conectar com outros modelos preditivos; integrar com as Ovitrampas (armadilhas de monitoramento do *Aedes aegypti*); comunicação bidirecional com Inteligência Artificial; conectar com os boletins do SUS; ampliar o rol de alertas; conectar com outras redes digitais da saúde; realizar projeto piloto em âmbito municipal.

Referências

[1] InfoDengue, "Sistema de monitoramento de arboviroses - informações," https://info.dengue.mat.br/informacoes/, acesso em: 01 mai. 2026.

[2] C. T. Codeço, F. C. Coelho, O. G. Cruz, S. Oliveira, M. Castro, and L. S. Bastos, "Infodengue: A nowcasting system for the surveillance of arboviral diseases in brazil," *Memórias do Instituto Oswaldo Cruz*, vol. 113, no. 7, p. e170370, 2018.

[3] Ministério da Saúde, "Aedes aegypti," 2026, acesso em: 05 mai. 2026. [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/aedes-aegypti

[4] R. Gurgel-Gonçalves, W. K. d. Oliveira, and J. Croda, "The greatest dengue epidemic in brazil: Surveillance, prevention, and control," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 57, pp. e00 203–2024, 2024, acesso em: 01 mai. 2026. [Online]. Available: https://www.scielo.br/j/rsbmt/a/gTT8ZZVR8g4Xx9xQPyp4GBz/?format=pdf&lang=en

[5] K. Heath, L. M. Alves, and M. B. Bonsall, "Climate change, urbanisation and transmission potential: Aedes aegypti mosquito projections forecast future arboviral disease hotspots in brazil," *PLoS Neglected Tropical Diseases*, vol. 19, no. 9, p. e0013415, Sep. 2025, acesso em: 22 jun. 2026. [Online]. Available: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0013415

[6] Ministério da Saúde, "Portal de dados abertos do sus: Arboviroses," 2026, acesso em: 01 jun. 2026. [Online]. Available: https://dadosabertos.saude.gov.br/dataset?groups=arboviroses

[7] Brasil, "Lei nº 13.709, de 14 de agosto de 2018: Lei geral de proteção de dados pessoais (lgpd)," https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, 2018, acesso em: 25 jun. 2026.

[8] Instituto Nacional de Meteorologia (INMET), "Normais climatológicas do brasil," https://portal.inmet.gov.br/normais, 2026, acesso em: 28 mai. 2026.

[9] Ministério das Cidades, "Projeto acertar: Certificação e melhoria da gestão da informação do snis," https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/saneamento/snis/produtos-do-snis/projeto-acertar, 2026, acesso em: 30 mai. 2026.

[10] Instituto de Pesquisa Econômica Aplicada (IPEA), "Renda, pobreza e desigualdade: indicadores sociais," https://www.ipea.gov.br/portal/retrato/indicadores/renda-pobreza-e-desigualdade/apresentacao, 2026, acesso em: 15 mai. 2026.

[11] National Oceanic and Atmospheric Administration (NOAA), "El niño and la niña," https://www.noaa.gov/education/resource-collections/weather-atmosphere/el-nino, 2026, acesso em: 15 mai. 2026.

[12] Fundação Oswaldo Cruz (Fiocruz), "Dengue," https://agencia.fiocruz.br/dengue, 2026, acesso em: 15 mai. 2026.

[13] Instituto Brasileiro de Geografia e Estatística (IBGE), "Estatísticas sociais da população: Censo 1991," https://www.ibge.gov.br/estatisticas/sociais/populacao/25089-censo-1991-6.html, 2026, acesso em: 15 maio 2026.

[14] Ministério do Turismo, "Anuário do ministério do turismo revela retomada do setor no período pós-pandemia, com recordes históricos," https://www.gov.br/turismo/pt-br/assuntos/noticias/anuario-do-ministerio-do-turismo-revela-retomada-do-setor-no-periodo-pos-pandemia-com-recordes-historicos, 2026, acesso em: 15 mai. 2026.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[16] K. Pearson, "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[17] Ministério das Cidades, "Abastecimento de Água – sistema nacional de informações sobre saneamento (snis)," https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/saneamento/snis/painel/ab, 2026, acesso em: 22 jun. 2026.

[18] I. J. Maciel, J. B. Siqueira Júnior, and C. M. T. Martelli, "Epidemiologia e desafios no controle do dengue," *Revista de Patologia Tropical / Journal of Tropical Pathology*, vol. 37, no. 2, pp. 111–130, 2008, acesso em: 01 mai. 2026. [Online]. Available: https://revistas.ufg.br/iptsp/article/view/4998

[19] C. Barcellos, R. Lowe, P. Oger *et al.*, "Climate change, thermal anomalies, and the recent progression of dengue in brazil," *Scientific Reports*, vol. 14, p. 5948, 2024, acesso em: 22 jun. 2026. [Online]. Available: https://www.nature.com/articles/s41598-024-56044-y

[20] S. T. Stoddard, A. C. Morrison, G. M. Vazquez-Prokopec *et al.*, "House-to-house human movement drives dengue virus transmission," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 110, no. 3, pp. 994–999, 2013. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1213349110

[21] R. M. R. Nogueira, H. G. Schatzmayr *et al.*, "Dengue virus type 3, brazil, 2002," *Emerging Infectious Diseases*, vol. 11, no. 9, pp. 1376–1381, 2005. [Online]. Available: https://wwwnc.cdc.gov/eid/article/11/9/04-1043_article.htm

Artificial Intelligence and Digital Technologies to Support Individuals with Autism Spectrum Disorder and Their Caregivers: A Systematic Review

Leonardo Ramos de Oliveira Campanini
Bioengineering Department, Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
ORCID: 0009-0008-6401-8572

Marcela Maria da Silva Zorrer Franco
Bioengineering Department, Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
ORCID: 0000-0003-0723-403X

Alessandro Márcio Hakme da Silva
Bioengineering Department, Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
ORCID: 0000-0001-7320-4790

Abstract*—This study aimed to systematically review the evidence on AI-enabled and digital health technologies supporting individuals with autism spectrum disorder (ASD), their caregivers, and families. The review was reported according to PRISMA 2020 and prospectively registered in PROSPERO (No. 1166432). Databases were searched from April to June 2026 using controlled vocabulary and keywords related to ASD, artificial intelligence, chatbots, caregivers, and digital support. Risk of bias was assessed using the Cochrane RoB 2 tool for the randomized studies, QUADAS-2 for the diagnostic accuracy study, and the CASP Qualitative Checklist for the qualitative study. Of 90 records identified across four databases, 15 were assessed at full text and 5 met the eligibility criteria; among the 10 full-text exclusions, one was removed owing to a post-publication retraction on integrity grounds. The included studies addressed ChatGPT-delivered physical activity support, mobile app-assisted parent training, an AI-based diagnostic device integrated into primary care, caregiver perspectives on an AI-assistive device supporting daily functioning, and an educational application of AI framed from a public-health perspective. Preliminary evidence suggests feasibility, acceptability, and potential gains in engagement, knowledge acquisition, diagnostic workflow efficiency, and caregiver-perceived support for adaptive functioning; however, clinical effectiveness remains uncertain due to methodological heterogeneity, small samples, and limited external validation. Future studies should adopt standardized outcomes, larger samples, and longer follow-up to move beyond feasibility toward clinical effectiveness.

***Keywords*—autism spectrum disorder; artificial intelligence; digital health; chatbots; caregivers; assistive technology; systematic review**

I. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent deficits in social communication and interaction, together with restricted and repetitive patterns of behavior. Recent estimates from the Autism and Developmental Disabilities Monitoring Network of the United States Centers for Disease Control and Prevention indicate a prevalence of approximately 1 in 31 children aged 8 years [1], following an upward trend over recent decades. Care for people with ASD imposes substantial demands on caregivers and families, frequently associated with emotional burden, difficulty accessing reliable information, diagnostic delay, and gaps in continuous support from health services.

These access barriers, along with the need for guidance available outside service hours, represent a relevant problem for biomedical engineering, as they call for technological support solutions that are usable, safe, and integrable into the clinical workflow. In this context, digital technologies have been proposed as tools to support care. It is important, however, to distinguish categories often treated as equivalent: generative AI systems and chatbots based on large language models; AI-based diagnostic support devices; digital health (mHealth) applications without explicit AI; and AI-assistive devices designed to support adaptive functioning in daily activities. Not every digital technology incorporates AI, and this conceptual delimitation is necessary to properly appraise the evidence. Such technologies have been proposed for offering continuous availability, personalization of interactions, and the potential to broaden access to information and support [2]-[5].

Accordingly, this systematic review seeks to answer the following question: what is the evidence on the feasibility, acceptability, and outcomes of AI-enabled and digital health

technologies in supporting individuals with ASD, caregivers, and families? The objective is to map and critically appraise this evidence, with attention to the distinction between technology types and to the methodological quality of the studies.

II. Methods

A. Protocol and registration

This systematic review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [7] and was prospectively registered in the PROSPERO database under number 1166432.

The review question and inclusion criteria were structured using the PICOS framework. Population: individuals with ASD, their caregivers and families. Intervention/exposure: AI-enabled or digital health technologies, such as chatbots, virtual assistants, mobile applications, and diagnostic support devices. Comparator: usual care, no intervention, or pre-post comparison, where applicable. Outcomes: feasibility, acceptability, usability, engagement, knowledge acquisition, emotional support, time to diagnosis, and clinical effectiveness. Study designs: randomized controlled trials, pilot studies, and observational studies.

B. Search strategy

From April to June 2026, the literature search was conducted in the PubMed electronic databases (US National Library of Medicine), Scopus, Web of Science, and IEEE Xplore. The search strategy combined controlled descriptors (MeSH) and keywords using Boolean operators (AND/OR), restricted to studies published within the last five years. The five-year window is justified by the recent emergence of generative AI and large language models in health care, a field whose expressive growth occurred within this period. The State of the Art through Systematic Review (StArt) software supported the organization of the procedures. The entire process is reported through the PRISMA flow diagram (Fig. 1).

The MeSH descriptors and keywords were: *("Autism Spectrum Disorder" OR ASD) AND ("Artificial Intelligence" OR chatbot OR "virtual assistant" OR "digital support") AND ("Caregivers" OR Family OR "Mental Health" OR "Social Support")*.

C. Selection, extraction, and eligibility

Records were imported into StArt, with automatic removal of duplicates, followed by title and abstract screening and full-text reading of the remaining articles. Explicit exclusion criteria were defined: studies not employing AI or digital technologies in the ASD context; studies without empirical data, such as editorials, commentaries, and reviews; studies not involving individuals with ASD, caregivers, or families; and studies of low methodological quality. Extracted outcomes covered usability, acceptability, engagement, emotional support, knowledge acquisition, time to diagnosis, adverse events, and clinical effectiveness. Screening and study selection were conducted independently by two reviewers, following the protocol previously registered in PROSPERO, with a third reviewer available to resolve any disagreements. The two reviewers reached full agreement on all eligibility decisions at the full-text stage, so no third-reviewer adjudication was required.

D. Methodological quality assessment

Risk of bias in the included studies was assessed independently by two reviewers using the Cochrane Risk of Bias tool, version 2 (RoB 2) [8] for the two randomized controlled trials, the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) [9] for the diagnostic accuracy study, and the Critical Appraisal Skills Programme (CASP) Qualitative Checklist [10] for the qualitative study, in line with reviewer feedback that PEDro and an adapted Newcastle-Ottawa Scale do not adequately capture the specific risks of an AI-based diagnostic device. RoB 2 judgments were made at the domain level (randomization process; deviations from intended interventions; missing outcome data; measurement of the outcome; selection of the reported result), each rated as low risk, some concerns, or high risk, with an overall judgment derived from the highest level of concern across domains. QUADAS-2 judgments were likewise made at the domain level (patient selection; index test; reference standard; flow and timing). The CASP Qualitative Checklist appraises 10 domains covering the clarity of aims, the appropriateness of the qualitative methodology and research design, recruitment and data collection, the researcher-participant relationship, ethical considerations, rigor of data analysis, clarity of findings, and the value of the research, with an overall judgment of low, some, or high concern derived from the pattern of responses across domains.

III. Results

A. Study selection and quality

In the identification stage, 87 records were retrieved from PubMed, Scopus, Web of Science, and IEEE Xplore; an updated search subsequently added 3 further candidate records, yielding 90 records overall. After title and abstract screening, 15 studies were considered potentially relevant and selected for full-text reading. At this stage, 9 studies were excluded because they did not address the use of AI or digital technologies to support people with ASD, caregivers, or families, and one further candidate record was excluded after a post-publication retraction on integrity grounds, the retraction note citing an inadequately described framework, an AI-generated figure, undisclosed absence of code, and unverifiable datasets [15]. Ultimately, 5 studies met all criteria, including minimum methodological quality, and were included in the qualitative synthesis [2]-[6].

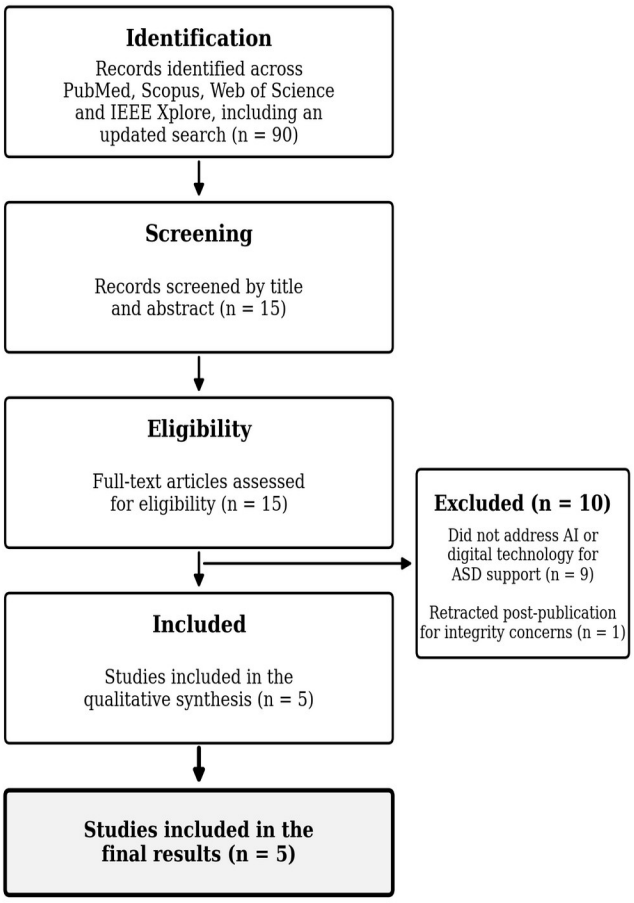


Fig. 1. PRISMA flow diagram of the identification, screening, and inclusion of studies.

The risk-of-bias and quality appraisal of the included studies is shown in Tables I-III. Both randomized controlled trials [2], [3] were judged at overall high risk of bias under RoB 2, chiefly due to unblinded, self-reported outcomes and, for Lee et al. [3], substantial and differential attrition. The diagnostic accuracy study [4] was likewise judged at overall high risk of bias under QUADAS-2, driven by an undisclosed potential conflict of interest in the index-test domain (co-authorship by employees of the device manufacturer) and by incorporation bias in the reference-standard domain, since the final clinical diagnosis was made with knowledge of the device output. The qualitative study [5] was judged at overall some concerns under the CASP Qualitative Checklist: aims, methodology, design, ethical conduct, and analytical rigor (independent second coding, saturation, audit trail) were appropriate, but recruitment drew on a clinic-based, industry-partnered sample skewed toward the more severe range of the spectrum and toward early technology adopters, and data completeness may have been affected by the AI device's own responsiveness problems during data collection.

TABLE I. Risk-of-bias assessment of the randomized controlled trials (Cochrane RoB 2, domain-level judgments).

RoB 2 domain	Aydemir (2025) [2]	Lee et al. (2024) [3]
D1. Randomization process	Some concerns (random allocation stated; sequence generation and allocation concealment not described)	Low risk (independent third-party computer-generated stratified block randomization, block size 4)
D2. Deviations from intended interventions	Some concerns (unblinded parent-delivered intervention; analysis approach for missing implementation not detailed)	Some concerns (open-label; primary analysis per-protocol rather than intention-to-treat)
D3. Missing outcome data	Low risk (no dropouts reported; all 26 dyads completed the study)	High risk (25% overall attrition post-randomization, higher in the intervention arm: 8/28 vs 3/28)
D4. Measurement of the outcome	High risk (self-reported outcome; parents aware of intervention status)	High risk (parent-reported primary outcome; open-label, unblinded assessment)
D5. Selection of the reported result	Some concerns (no prospective trial registration identified)	Low risk (prospectively registered, CRIS KCT0007841; pre-specified outcomes fully reported)
Overall judgment	High risk	High risk

Abbreviations: RoB 2 = Cochrane Risk of Bias tool, version 2.

TABLE II. Risk-of-bias assessment of the diagnostic accuracy study (QUADAS-2, domain-level judgments).

QUADAS-2 domain	Sohl et al. (2025) [4]
Patient selection	Some concerns (real-world sample drawn only from primary care clinics enrolled in the ECHO Autism: EDx training program, which may not represent general primary care settings)
Index test (AI diagnostic device)	High risk (3 of 10 co-authors are employees of Cognoa Inc, the manufacturer of the evaluated device; independent, blinded interpretation of device output not described)
Reference standard	High risk (final clinical diagnosis was made with knowledge of the device output, which the workflow explicitly incorporates as an input — incorporation bias)
Flow and timing	Some concerns (device produced a determinate output in only 52.5% of cases; handling of

	indeterminate cases in the time-to-diagnosis analysis not fully detailed)
Overall judgment	High risk (chiefly due to the index test and reference-standard domains)

Abbreviations: QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies 2.

TABLE III. Quality appraisal of the qualitative study (CASP Qualitative Checklist, domain-level judgments).

CASP domain	Perry et al. (2026) [5]
Aims and methodology	Low concern (clearly focused aim to explore caregiver perspectives on an AI-assistive device; inductive qualitative content analysis appropriate for an underresearched area)
Research design and recruitment	Some concerns (clinic-based, industry-partnered purposive sample of 10 caregivers; skewed toward Level 2-3 ASD severity and early technology adopters, limiting transferability)
Data collection	Some concerns (two audio/video-recorded, transcribed focus groups; the device's own slow, unresponsive performance may have affected participant engagement and data completeness)
Reflexivity and researcher-participant relationship	Low concern (researchers' positionality documented; no prior relationship with participants; independent second coder blinded to initial coding)
Ethical considerations	Low concern (institutional ethics approval and informed consent obtained; industry funder disclosed with no role in analysis or interpretation)
Rigor of data analysis	Low concern (independent coding of 25% of data with 81.7% agreement; analytical memos; saturation confirmed across two focus groups; audit trail maintained)
Clarity and value of findings	Low concern (three themes reported with illustrative quotations; authors explicitly discuss limits to generalizability)
Overall judgment	Some concerns (chiefly due to sample selection and possible device-performance effects on data collection)

Abbreviations: CASP = Critical Appraisal Skills Programme.

B. Study and participant characteristics

The characteristics of the studies and participants are presented in Table IV. Five studies were included, conducted in different contexts and populations, involving individuals with ASD as well as their caregivers and families across different age groups. The included technologies span conceptually distinct categories, as indicated in the table, ranging from generative AI to an AI-based diagnostic device, a digital health application, an AI-assistive device for daily functioning, and an educational application of AI addressed from a public-health perspective. Notably, the last of these [6] differs from the others in design: rather than reporting an intervention with human participants, it proposes and validates a public-health-oriented AI model for social-skills support. It was retained because it directly addresses AI applied to the education and support of children with ASD, the phenomenon of interest of this review, while its distinct, non-empirical nature is made explicit and weighed in the synthesis.

TABLE IV. Study and participant characteristics.

Study (year), country	Technology type	Target user	Sample	Participants (age)
Aydemir (2025), Turkey [2]	Generative AI (chatbot)	Patient and caregiver	26 families (parent-child dyads); 13 intervention and 13 control	Children with ASD and their parents
Lee et al. (2024), South Korea [3]	Digital health (mobile app)	Caregiver	64 enrolled; 42 children in the program (20 intervention, 22 control)	Parents of children with ASD aged 36-84 months (median 49 months)
Sohl et al. (2025), USA [4]	AI-based diagnostic device	Professional	80 children and 7 primary-care clinicians	Children aged 18-72 months with suspected developmental delay, and their caregivers
Perry et al. (2026), Australia [5]	AI-assistive device (tablet)	Caregiver	10 caregivers (5 mothers, 5 fathers); 2 focus groups	Children aged 6.3-8.4 years with ASD (Level 2-3) and their caregivers (mean age 42.8 years)
Lan et al. (2025), China [6]	Educational AI application (public-health perspective)	Children with ASD (educational context)	Conceptual/ analytical study; model validated on benchmark language datasets	School-aged children with ASD (educational setting)

Abbreviations: AI = artificial intelligence; ASD = autism spectrum disorder.

C. Synthesis of main findings

Table V summarizes the interventions, instruments, and main findings, distinguishing primary from secondary outcomes. Overall, the use of chatbots, applications, AI-based devices, and an AI-assistive device was associated with feasibility, acceptability, and potential gains in engagement, knowledge, diagnostic workflow efficiency, and caregiver-perceived support for adaptive functioning, rather than with established clinical effectiveness.

TABLE V. Interventions, instruments, and main findings of the included studies.

Study	Intervention	Instruments	Main findings

Aydemir (2025) [2]	ChatGPT-recommended physical activities delivered by parents	LTEQ; feasibility questionnaire	Significant increase in physical activity in the intervention group; intervention rated feasible and useful by parents
Lee et al. (2024) [3]	ABA-based parent training assisted by a mobile app for 12 weeks, added to usual care	K-CBCL (primary); CGI-S; K-SCQ; K-PSI-4-SF	Feasibility demonstrated, no adverse events; small reduction in K-CBCL total score (median difference -0.5; 95% CI -4.5 to 3.0), confidence intervals crossing the null
Sohl et al. (2025) [4]	AI-based diagnostic device (Canvas Dx) integrated into the ECHO Autism primary-care workflow	Electronic questionnaires; chart review; device outputs	Mean time to diagnosis of 39.2 days (vs. 180-264 in specialized centers); 93% caregiver satisfaction; determinate outputs in 52.5% of cases, all consistent with final clinical determination
Perry et al. (2026) [5]	Focus groups exploring caregiver use of an AI-assistive device (Pixi Home-Hub) for daily functioning support	Semi-structured focus groups; inductive content analysis (NVivo)	Caregivers valued goal-tracking and scheduling features and proposed the device as a digital support navigator for disability-funding processes (NDIS); usability limited by slow, unresponsive AI voice interaction; no quantitative outcome measures
Lan et al. (2025) [6]	Analysis of AI applications in the education of children with ASD, framed from a public-health perspective	Narrative/ analytical synthesis; model validation on benchmark datasets	Discusses potential of AI to support individualized education and emotion recognition for children with ASD; conceptual contribution rather than empirical clinical outcomes; reinforces the argument for AI potential in educational support

Abbreviations: ABA = applied behavior analysis; AI = artificial intelligence; CGI-S = Clinical Global Impression-Severity; CI = confidence interval; ECHO = Extension for Community Healthcare Outcomes; K-CBCL = Korean Child Behavior Checklist; K-PSI-4-SF = Korean Parenting Stress Index, 4th Edition Short Form; K-SCQ = Korean Social Communication Questionnaire; LTEQ = Leisure-Time Exercise Questionnaire.

IV. Discussion

The findings of this review indicate preliminary evidence of feasibility and acceptability for AI-enabled and digital health technologies in ASD care support, with possible gains in engagement, knowledge acquisition, and diagnostic workflow efficiency. The included studies used distinct technologies and designs, spanning conceptually different categories: a physical activity intervention mediated by a generative-AI chatbot [2], a parent training program assisted by a digital health mobile app [3], an AI-based diagnostic support device integrated into primary care [4], an AI-assistive device supporting caregiver-reported adaptive functioning [5], and a conceptual, public-health-oriented AI modeling study on education for children with ASD [6]. This conceptual diversity — spanning a generative-AI chatbot, a mobile application without explicit AI, and an AI-based diagnostic device, among others — together with heterogeneous samples and outcomes, limits any joint conclusion about "AI-enabled and digital technologies" as a single category, and reinforces the need to appraise each technology type on its own terms. Because five heterogeneous studies were included, no quantitative synthesis was appropriate, and the findings should be interpreted as exploratory.

A recurring pattern across the included studies is the mediation of technology use by a caregiver or a professional, rather than direct, unmediated use by the individual with ASD. In Aydemir [2], ChatGPT-generated recommendations were interpreted and delivered by parents; in Lee et al. [3], the mobile application supported a parent-training program rather than interacting directly with the child; in Sohl et al. [4], the diagnostic device was operated within a clinician-led workflow; and in Perry et al. [5], caregivers alone interacted with the AI-assistive device across two focus groups, with no direct use by the child reported. No included study relied on unmediated interaction between the individual with ASD and the technology, which suggests that human mediation, whether by caregivers or professionals, remains a structural feature of current AI and digital health applications in this population, with implications for how usability and safety should be evaluated in future studies.

Effect magnitude, however, varied. In Lee et al. [3], the reduction in behavioral problems measured by the K-CBCL was small and the confidence intervals crossed the null value, which limits conclusions about efficacy and reinforces the exploratory nature of the study. The results of Sohl et al. [4] suggest a marked gain in diagnostic efficiency, with a substantial reduction in time to diagnosis compared with the waiting list of specialized centers.

Continuous availability and personalization of interactions are central attributes of these technologies, with potential to reduce access barriers in settings with scarce specialized services. Such tools may complement, without replacing, professional follow-up. From a biomedical engineering standpoint, however, their incorporation into care depends on requirements that the included studies still address only to a limited extent: performance validation, usability, interoperability, integration into the clinical workflow, and regulatory compliance. These elements are decisive for turning a promising proof of concept into a genuinely usable health technology.

Intermediate outcomes must be distinguished from clinical effectiveness. The included studies mainly measure feasibility, satisfaction, and engagement, which are legitimate but insufficient to demonstrate sustained clinical benefit, especially without long-term follow-up. The adoption of these technologies also entails risks that require explicit attention: privacy and protection of sensitive data, algorithmic bias, generation of incorrect information by large language models (hallucinations), clinical safety, and the need for professional

supervision. Poorly validated support tools may, moreover, displace the human bond rather than reinforce it, a particularly sensitive risk for families already experiencing isolation. These aspects remain underexplored in the available literature.

These findings converge with recent reviews. Scoping and systematic reviews on generative AI and large language models applied to ASD have likewise pointed to a predominance of preliminary evidence, scarce external validation, and the need for caution regarding bias and safety [11], [12]. In the field of diagnostic AI and social-behavior tracking in ASD, the literature also highlights the potential for screening and prioritization alongside implementation and interoperability gaps [13]. More broadly, reviews on AI in mental health reinforce issues of transparency, interpretability, and validation limitations that apply directly to the present context [14]. This work adds to that emerging body by conceptually delimiting technology types and by making explicit the still preliminary nature of the evidence.

The screening process also surfaced one candidate record that illustrates emerging risks to evidence synthesis in this fast-moving field. A report proposing an explainable-AI diagnostic framework was retracted after publication over concerns of an inadequately described method, an AI-generated figure, and unverifiable training and validation datasets [15], underscoring the need for reviewers to check the current status of AI-related publications rather than relying on the title and abstract alone. Excluding this retracted record at full-text review, rather than incorporating it into the synthesis, reinforces the robustness of the eligibility process adopted here.

V. Conclusion

The available evidence indicates that AI-enabled and digital health technologies applied to ASD care are promising but still preliminary. Because the five included studies evaluate conceptually distinct technologies — a generative-AI chatbot, a mobile application without explicit AI, an AI-based diagnostic device, an AI-assistive device for caregiver-reported adaptive functioning, and an educational application of AI — with heterogeneous samples, designs, and outcomes, any conclusion that treats "AI-enabled and digital technologies" as a single, undifferentiated category is necessarily limited. The main finding of this review is the feasibility and acceptability of these tools, rather than the robust demonstration of clinical effectiveness, which remains uncertain given the methodological heterogeneity, small samples, and limited external validation. Screening was conducted independently by two reviewers across four databases, with a third reviewer available for adjudication, and the protocol was prospectively registered, strengthening the reliability of the selection process; nonetheless, the small number of eligible studies and their heterogeneity remain the main constraints. Future studies should adopt standardized outcomes and longer follow-up to move beyond feasibility toward clinical effectiveness. For biomedical engineering, the advancement of the field will depend on performance validation, safe integration into the care workflow, and attention to regulatory requirements. By delimiting technology types and mapping existing gaps, this review contributes to guiding the future development of AI solutions for ASD care support.

References

[1] K. A. Shaw, S. Williams, M. E. Patrick, et al., "Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years, Autism and Developmental Disabilities Monitoring Network, 16 sites, United States, 2022," MMWR Surveillance Summaries, vol. 74, no. SS-2, pp. 1-22, 2025, doi: 10.15585/mmwr.ss7402a1.

[2] U. Aydemir, "ChatGPT-delivered physical activity intervention for children with autism spectrum disorder: pre-post feasibility study," JMIR Human Factors, vol. 12, e71119, 2025, doi: 10.2196/71119.

[3] J. Lee, J. Lim, S. Kang, S. Kim, S. Y. Jung, S. Kim, S.-B. Hong, and Y. R. Park, "Mobile app-assisted parent training intervention for behavioral problems in children with autism spectrum disorder: pilot randomized controlled trial," JMIR Human Factors, vol. 11, e52295, 2024, doi: 10.2196/52295.

[4] K. Sohl, E. Linstead, K. Heinz, et al., "Integration of an artificial intelligence-based autism diagnostic device into the ECHO autism primary care workflow: prospective observational study," JMIR Formative Research, vol. 9, e80733, 2025, doi: 10.2196/80733.

[5] N. Perry, K. A. Boulton, L. Hankin, B. B. Roisenberg, and A. J. Guastella, "A qualitative study investigating caregiver perspectives of an artificial intelligence assistive device to support daily activities in families with children with autism spectrum disorder," Digital Health, vol. 12, 2026, doi: 10.1177/20552076251411228.

[6] L. Lan, K. Li, and D. Li, "Exploring the application of AI in the education of children with autism: a public health perspective," Frontiers in Psychiatry, vol. 15, 1521926, 2025, doi: 10.3389/fpsyt.2024.1521926.

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, n71, 2021, doi: 10.1136/bmj.n71.

[8] J. A. C. Sterne, J. Savović, M. J. Page, et al., "RoB 2: a revised tool for assessing risk of bias in randomised trials," BMJ, vol. 366, l4898, 2019, doi: 10.1136/bmj.l4898.

[9] P. F. Whiting, A. W. S. Rutjes, M. E. Westwood, et al., "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies," Annals of Internal Medicine, vol. 155, no. 8, pp. 529-536, 2011, doi: 10.7326/0003-4819-155-8-201110180-00009.

[10] Critical Appraisal Skills Programme, "CASP Qualitative Studies Checklist," 2018. [Online]. Available: https://casp-uk.net/checklists-tool/qualitative-studies-checklist/

[11] J.-S. Sohn, E. Lee, J.-J. Kim, H.-K. Oh, and E. Kim, "Implementation of generative AI for the assessment and treatment of autism spectrum disorders: a scoping review," Frontiers in Psychiatry, vol. 16, 1628216, 2025, doi: 10.3389/fpsyt.2025.1628216.

[12] S. S. Joudar, A. Albahri, R. A. Hamid, et al., "Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: a systematic review of current trends and open issues," Artificial Intelligence Review, vol. 56, no. Suppl 1, pp. 53-117, 2023, doi: 10.1007/s10462-023-10536-x.

[13] C. Sun, A. McEwan, K. A. Boulton, et al., "Artificial intelligence for tracking social behaviours and supporting an autism spectrum disorder diagnosis: systematic review and meta-analysis," eBioMedicine, vol. 120, 105931, 2025, doi: 10.1016/j.ebiom.2025.105931.

[14] P. Cruz-Gonzalez, A. W. J. He, E. P. Lam, et al., "Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications," Psychological Medicine, vol. 55, e18, 2025, doi: 10.1017/S0033291724003295.

[15] S. Jiang, "Bridging the gap: explainable AI for autism diagnosis and parental support with TabPFNMix and SHAP," Scientific Reports, vol. 15, 40850, 2025, doi: 10.1038/s41598-025-24662-9. Retracted: Retraction Note, Scientific Reports, vol. 15, 43184, 2025, doi: 10.1038/s41598-025-31337-y.

Artificial Intelligence-Based Ultrasound Simulation for Thyroid Nodules: State of the Art, Gaps, and Future Perspectives

Oswaldo Ferreira Martins Júnior
Graduate Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
oswaldojunior@alunos.utfpr.edu.br

Mohamed Adel Khairy Said
Graduate Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
mohamed.science.91@gmail.com

Luca Paparella
Graduate Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
lucapaparella@alunos.utfpr.edu.br

Eduardo Tavares Costa
Department of Electronic Engineering and Biomedical Engineering (DEEB)
Center for Biomedical Engineering (CEB)
State University of Campinas (UNICAMP)
Campinas, Brazil
educosta@unicamp.br

Joaquim Miguel Maia
Graduate Program in Electrical and Computer Engineering (CPGEI), DAELN
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
joaquim@utfpr.edu.br

Amauri Amorin Assef
Graduate Program in Electrical and Computer Engineering (CPGEI), DAELT
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
amauriassef@utfpr.edu.br

Abstract*—Thyroid nodules are common endocrine findings, and ultrasonography remains the primary modality for screening, risk stratification, and guidance for fine-needle aspiration. Although artificial intelligence (AI) increasingly supports segmentation, detection, and classification, annotated thyroid ultrasound datasets remain constrained by annotation cost, interobserver variability, ethical considerations, and equipment heterogeneity. This PRISMA-informed structured qualitative review maps physical, statistical, hybrid, and AI-based approaches relevant to thyroid nodule simulation. The historical evidence set included 39 sources, and a separately reported targeted update added three recent sources. During revision, two reviewers independently reassessed 88 full-text reports (94.3% agreement; Cohen's $\kappa = 0.88$, 95% CI: 0.77–0.98). A design-sensitive appraisal applied QuADS across heterogeneous empirical designs, CLAIM 2024 to eligible medical-imaging AI studies, and PROBAST+AI to applicable prediction-model studies without collapsing the instruments into a common score. Physics-based simulators provided explicit acoustic control but were seldom linked to thyroid-specific downstream tasks. Recent thyroid-focused work introduced clinically guided diffusion, real-background inpainting with synthesis-linked masks and external segmentation gains, and externally evaluated synthetic contrast-enhanced ultrasound. Nevertheless, explicit clinical control, automatic annotation, real-background synthesis, and thyroid-specific AI validation remain incompletely integrated. The review separates evidence-derived observations from a conceptual research agenda and identifies testable opportunities rather than a single optimal simulator architecture.

***Index Terms*—ultrasonography, thyroid nodule, synthetic data, ultrasound simulation, artificial intelligence.**

I. Introduction

Thyroid nodules are highly prevalent in the general population, and their detection has increased with the wider use of high-resolution ultrasonography. Although most nodules are benign, distinguishing low- from high-risk cases directly affects decisions on follow-up, fine-needle aspiration biopsy, and surgery [1]–[3]. Clinical guidelines organize diagnostic stratification around descriptors such as composition, echogenicity, shape, margins, and echogenic foci. However, their interpretation remains operator-dependent and subject to intraobserver and interobserver variability [2], [3].

In parallel, deep learning-based decision-support systems increasingly address thyroid nodule detection, segmentation, and classification [4], [5]. Their performance depends strongly on the quantity, diversity, and quality of annotations. Manual annotation is costly because lesion boundaries may be poorly defined, local contrast is low, speckle modifies texture, and clinically relevant findings such as microcalcifications and posterior shadowing may be subtle or heterogeneous [6], [7]. Public datasets provide complementary but incomplete resources: DDTI is single-center [8], TN3K provides nodule masks [7], [9], and the larger TN5000 supports detection and classification with bounding boxes rather than pixel-level masks [10].

The gap between growing model complexity and limited annotated data makes synthetic data attractive. A simulator can provide scalable generation, parametric control, privacy preservation, and automatic reference annotations. It can also support controlled robustness experiments, isolate changes in echogenicity and texture, increase representation of rare patterns, and test whether synthetic augmentation improves generalization. Historically, however, ultrasound simulation developed along two weakly connected tracks: physics-based simulators emphasizing acoustic fidelity and operator training, and AI methods emphasizing visual synthesis or downstream augmentation [11]–[15].

For the thyroid, most AI studies still classify or segment real images. Earlier synthesis studies mainly addressed augmentation, classification, or domain adaptation [16]–[18]. Recent work has broadened this scope: Dai *et al.* developed a clinical-knowledge-guided, text-conditioned diffusion frame-

work for rare thyroid cancer subtypes; Prochazka and Zeman inserted synthetic nodules into real ultrasound backgrounds and measured cross-dataset segmentation gains; and Lai *et al.* generated synthetic thyroid contrast-enhanced ultrasound (CEUS), evaluated by radiologists and on an external cohort [19]–[21]. The contribution of this review is therefore not an absolute claim that synthetic data, lesion insertion, or automatic labels are unavailable, but a structured account of the relationship among physical simulation, generative synthesis, clinically meaningful thyroid semantics, synthesis-linked reference masks, and validation in thyroid-specific AI tasks. Adjacent literature is included only when it contributes a clearly transferable methodological component.

This work organizes simulation into physical, statistical, hybrid, and AI-driven classes; synthesizes thyroid-specific evidence; compares parameterization, annotation generation, code and data availability, validation, and reproducibility; and separates evidence-derived requirements from conceptual choices for future simulators.

II. Theoretical background

A. Ultrasound in thyroid nodules

B-mode ultrasound is the modality of choice for the initial evaluation of thyroid nodules because it is accessible, portable, free of ionizing radiation, and sensitive to nodular morphology [2], [3]. Stratification systems, such as ATA and ACR TI-RADS, use visual features extracted from these images to infer malignancy risk [2], [3]. In practice, however, these signs are influenced by transducer frequency, gain, focus, depth, compression, adjacent structures, and speckle noise. Consequently, the perception of irregular margins, halo sign, punctate echogenic foci, a taller-than-wide shape, and hypoechoic patterns can vary substantially across observers and equipment [5].

From a computational standpoint, thyroid nodule segmentation is challenging because a lesion may occupy only a small image region, have diffuse borders, be partly obscured by posterior shadowing, or coexist with heterogeneous glandular tissue. Traditional methods often reported favorable results on small, single-center datasets, whereas deep learning improved performance while increasing dependence on annotated data [6]. Recent systems still rely largely on manually delineated masks, limiting scalability, standardization, and reproducibility [7], [22], [23].

B. Ultrasound simulation

Ultrasound simulation represents, at different levels of abstraction, interactions among the acoustic beam, tissue, scatterers, the transducer, and the image-formation chain. Foundational work by Bamber and Dickinson and by Foster *et al.* established that B-mode appearance depends on medium statistics and the spatial response of the imaging system, enabling computational synthesis of speckle, cyst contrast, and processing effects [24], [25]. Field II subsequently consolidated ultrasound-field and pulse–echo simulation through transducer and scatterer models [11]. k-Wave extended wave-propagation modeling to heterogeneous media and more complex propagation phenomena [12], while CREANUIS incorporated nonlinear radiofrequency imaging for harmonic scenarios [26].

Beyond full-wave simulators, open tools support signal processing, beamforming, and benchmarking. The MATLAB UltraSound Toolbox (MUST) includes SIMUS, a frequency-domain simulator benchmarked against Field II and k-Wave [27], [28]. The UltraSound ToolBox (USTB) supports beamforming and reconstruction pipelines for simulated and experimental data [29]. The Challenge on Ultrasound Beamforming with Deep Learning (CUBDL) provides an open benchmark for learning-based image formation, but its scope is signal reconstruction rather than generation of clinically annotated datasets [30], [31].

Despite their technical maturity and physical grounding, these frameworks were not primarily designed to produce AI-ready annotated datasets. They generally provide more direct control over acoustic, geometric, and signal-processing parameters than over clinical descriptors such as microcalcifications or posterior shadowing. Thyroid-specific lesion models, consistent automatic masks, and reproducible Python integration also remain secondary features [14]. Statistical, convolutional, and hybrid approaches offer a complementary compromise between computational cost, controllability, and visual realism [13], [32].

C. AI in ultrasound imaging

AI in ultrasound evolved from handcrafted-feature classifiers to deep networks that learn image representations directly. In biomedical segmentation, U-Net became a reference architecture by combining multiscale context with spatial precision through skip connections [33]. Thyroid applications subsequently introduced attention, anatomical priors, and multitask learning, including TRFE-Net/TRFE+, enhanced U-Net architectures, automated pipelines, and multicenter models [7], [9], [22], [23], [34]. Reliable generalization nevertheless requires larger, more diverse, and consistently annotated datasets.

Ultrasound also imposes modality-specific constraints on AI. B-mode appearance depends strongly on acquisition protocol, probe orientation, operator pressure, gain, dynamic range, and equipment manufacturer. These factors alter speckle, boundary contrast, shadowing, and local texture, directly affecting feature learning and transferability. Models trained at one center or on one device may therefore perform well internally but degrade on external data. In thyroid imaging, this sensitivity is compounded by small lesions, low lesion–parenchyma contrast, and clinically important subtle findings such as punctate echogenic foci, irregular margins, and posterior artifacts [5]–[7]. Evaluation should consequently assess both task performance and stability across acquisition conditions while verifying preservation of clinically meaningful descriptors.

Generative adversarial networks (GANs) are an established family of methods for image synthesis and domain adaptation in medical imaging [15], [35]. Ultrasound studies have demonstrated realistic and editable synthesis for augmentation and segmentation [36]–[39]. Diffusion models have since been

applied to ultrasound texture generation and, more recently, to thyroid-specific text-guided synthesis and inpainting [19], [20], [40]. Thyroid CEUS translation has also progressed to external-cohort and radiologist evaluation [21]. These advances narrow the earlier gap, although their control mechanisms, outputs, and validation targets remain complementary rather than fully integrated.

D. Synthetic data and ground truth

Synthetic data are particularly valuable for low-prevalence patterns, costly annotations, and controlled experiments. Beyond increasing sample volume, they enable variable isolation, analysis of relationships between acquisition parameters and image appearance, and deterministic reference annotations when the simulated scene is known. Unlike manual masks, which remain subject to observer uncertainty, a simulator can generate segmentation masks, boxes, probability maps, and lesion metadata together with the image. These references are exact with respect to the simulation model, although not necessarily to real anatomy.

Synthetic-data value should not be assessed by visual similarity alone. Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) quantify aspects of structural fidelity and degradation [41], while Dice and Intersection over Union (IoU) evaluate segmentation. Stronger evidence links synthesis to target-task utility by testing models trained with synthetic or mixed data on held-out real examinations [16], [20], [39]. For thyroid ultrasound, the link among realism, synthesis-linked annotation, and downstream benefit remains incompletely characterized.

III. Methodology

PRISMA 2020 was used to structure the reporting of this PRISMA-informed qualitative review [42]. The review distinguishes the historical screening process from the subsequently conducted targeted literature update so that the two components are not interpreted as a homogeneous database rerun. During revision, the historical flow was cross-checked against the available aggregate records and the independent reassessment of all full-text reports. This verification does not alter the study-by-study technical comparisons; it only limits exact replication of the original database search. The protocol was not prospectively registered, and qualitative synthesis was selected because the studies differed substantially in domain, objectives, inputs, outputs, and evaluation metrics. Evidence from non-thyroid organs was used exclusively to identify anatomy-independent methodological building blocks—including acoustic image formation, controllable synthesis, paired image–annotation generation, and downstream synthetic-to-real validation—and was not interpreted as direct evidence of clinical effectiveness in thyroid imaging.

A. Research questions

Six questions guided the review: RQ1, classes of B-mode simulation methods; RQ2, modeled acoustic, anatomical, and textural phenomena; RQ3, uses of GANs, diffusion models, and other AI techniques; RQ4, thyroid-specific evidence; RQ5, automatic masks and downstream validation; and RQ6, underexplored combinations of parametric control, annotation, reproducibility, and AI evaluation. In PICOS terms, the population/context comprised ultrasound-imaging studies; interventions were physical, statistical, hybrid, and generative approaches; comparisons involved methods or real/synthetic training regimes; outcomes covered realism, control, annotation, task performance, and reproducibility; and study designs comprised peer-reviewed journal and conference papers.

B. Databases, period, and search expressions

Searches were performed in IEEE Xplore, Scopus, PubMed, and Web of Science for publications from January 2000 through April 30, 2026. Foundational works published before 2000 were considered only through backward citation searching and were tracked separately from database records. On August 20, 2026, a targeted update identified three eligible papers that had been available online before the cutoff but were absent from the original evidence set [19]–[21]. These papers were analyzed as a distinct update to the 39-source historical evidence set rather than as products of a homogeneous rerun of the original searches. The strategy combined English-language concepts for ultrasound, simulation or synthesis, generative AI, and thyroid imaging. Table I summarizes the four Boolean concept blocks, with syntax adapted to each database.

Table I
BOOLEAN CONCEPT BLOCKS USED IN THE SEARCH STRATEGY.

Code	Concept block (adapted per database)
S1	(`ultrasound` OR `ultrasonography`) AND (`simulation` OR `synthetic` OR `synthesis`) AND (`B-mode` OR `speckle` OR `scatterer` OR `acoustic`)
S2	`ultrasound` AND (`deep learning` OR `GAN` OR `diffusion` OR `generative model`) AND (`synthetic` OR `synthesis` OR `augmentation`)
S3	`thyroid` AND `ultrasound` AND (`segmentation` OR `classification` OR `detection`) AND (`deep learning` OR `U-Net` OR `GAN`)
S4	`thyroid` AND `ultrasound` AND (`simulation` OR `synthetic` OR `augmentation` OR `generative`)

In Table I, S1 and S2 capture general simulation and generative methods, whereas S3 and S4 narrow retrieval toward thyroid-specific AI and synthesis. The layered logic retains transferable contributions that may not explicitly use "thyroid."

C. Inclusion and exclusion criteria

Eligible studies were peer-reviewed, English-language, full-text journal or conference papers with auditable methods and results on ultrasound simulation, synthesis, relevant AI tasks, or thyroid nodules. Abstracts, editorials, tutorials without experiments, non-ultrasound studies, and operator-training papers without image formation were excluded. Non-thyroid studies required at least one transferable component: general image formation, controllable acoustic or lesion parameters, conditioned synthesis, paired image–annotation generation, or downstream validation. Their organ-specific performance was not treated as direct thyroid evidence. Seminal architecture, metric, and framework descriptions were retained as complementary methodological sources when required to interpret the reviewed pipelines; the MONAI preprint was used only for implementation context and not as clinical thyroid evidence.

D. PRISMA process and data extraction

The database searches returned 612 records: 126 from IEEE Xplore, 181 from Scopus, 143 from PubMed, and 162 from Web of Science. After removing 164 duplicates, 448 titles and abstracts were screened; 360 were excluded. Of 88 database-derived full-text reports, 55 were excluded and 33 retained. Backward citation searching contributed six sources, producing a historical evidence set of 39 sources. The targeted update added three separately identified papers, yielding an expanded qualitative synthesis of 42 sources. The independent reassessment conducted after screening confirmed the full-text screening totals reported in Fig. 1.

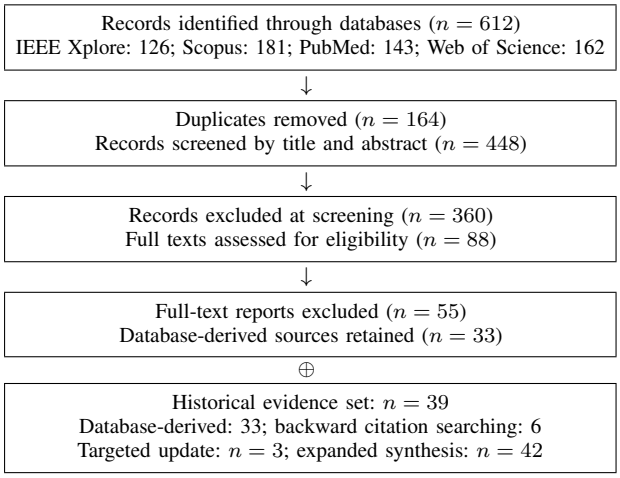


Figure 1. PRISMA-style flow separating the historical evidence set from the targeted update.

The largest reduction occurred during title and abstract screening, followed by 55 full-text exclusions. Reporting the additional routes separately prevents pre-2000 foundational sources from being attributed to the time-limited database search and makes the targeted addition of recent thyroid-generation papers explicit. Backward citation searching was used principally to retain foundational models of B-mode formation, speckle, and ultrasound-system simulation, including Bamber and Dickinson, Foster *et al.*, and Jensen [11], [24], [25].

During manuscript revision, two reviewers independently reassessed all 88 full-text reports against the predefined eligibility criteria before consensus discussion. Pre-consensus agreement was 94.3% (83/88), with Cohen's $\kappa = 0.88$ (95% CI: 0.77–0.98), indicating high agreement beyond chance. Five reports were classified differently and were resolved by structured consensus; third-reviewer adjudication was predefined but not required. Following consensus, 33 sources were retained and 55 reports excluded at the full-text stage. The primary reviewer extracted domain, method, inputs and outputs, parameterization, annotations, downstream AI integration, metrics, code and data availability, and external validation using a predefined framework. A second reviewer independently verified the extracted information, and discrepancies were resolved by consensus. The remaining coauthors reviewed the technical narrative, manuscript consistency, and English-language presentation but did not determine eligibility.

E. Methodological reporting and appraisal framework

Methodological appraisal used a design-sensitive strategy rather than a universal risk-of-bias score. QuADS was applied to heterogeneous empirical designs using its native 13-item, 0–3 structure [43]. CLAIM 2024 was applied at item level to eligible medical-imaging AI studies using its Yes, No, and Not Applicable categories [44]. PROBAST+AI was restricted to applicable diagnostic or prediction-model studies, with model-development quality, model-evaluation risk of bias, and applicability assessed separately [45]. Two reviewers independently completed the applicable appraisal forms, and disagreements were resolved by consensus. No common numerical score was created across the instruments because their constructs and intended applications differ substantially. Table IV summarizes the principal domain-level findings.

IV. Results and Evidence Synthesis

A. Overview of the corpus

The expanded 42-source synthesis comprised a historical evidence set of 39 sources and a targeted three-source update. It covered physical/wave, statistical, hybrid, generative, and thyroid-specific AI approaches, together with explicitly identified contextual sources. Classical simulators emphasized acoustic control [11], [12], [24]–[26], whereas data-driven studies increasingly used masks, sketches, clinical knowledge, or text conditioning [19], [38], [39]. Physics-based studies generally described system parameters well but seldom addressed annotations or thyroid AI. Generative and segmentation studies emphasized realism or task performance, while reporting code, clinical control, calibration, and external validation inconsistently.

The design-sensitive appraisal revealed a consistent contrast between methodological clarity and external validation. Across QuADS assessments, study objectives and technical procedures were generally better documented than sampling rationale, stakeholder involvement, and reproducibility-related information. CLAIM-based assessment of eligible AI studies showed improved reporting of dataset partitioning, performance metrics, and external testing in the most recent studies, while code availability, calibration procedures, and complete data provenance remained heterogeneous. Among studies eligible for PROBAST+AI, the principal concerns involved retrospective or single-center datasets, incomplete reporting of missing-data handling, limited calibration analysis, and absence of independent external evaluation. These findings were used to qualify the narrative synthesis rather than to rank fundamentally different study designs.

B. Physical simulation

Bamber and Dickinson, together with Foster *et al.*, established statistical scatterer-based synthesis of speckle and contrast [24], [25]. Field II modeled pulse–echo responses and transducers [11], k-Wave addressed heterogeneous propagation

[12], and CREANUIS incorporated nonlinear RF imaging [26]. These interpretable tools provide causal acoustic control, but clinically meaningful thyroid descriptors, automatic lesion masks, and downstream AI evaluation were generally outside their scope.

C. Statistical simulation

Statistical or convolutional models approximate scatterers, the point spread function, and speckle at lower computational cost. Marion and Vray, for example, simulated moving 3D scatterers and validated image statistics [32]. Their efficiency benefits signal-processing and dynamic studies, but pathology is often represented as local statistical change rather than controlled echogenicity, texture, calcifications, margins, or posterior artifacts.

Tables II and III compare studies selected for thyroid-specific synthesis, controllable generation, synthesis-linked annotation, or downstream AI validation. Their symbols are defined directly below each table and apply only to these descriptive comparisons, not to the methodological instruments summarized later.

Table II
DATA, GENERATION CONTROL, AND SYNTHESIS-LINKED ANNOTATION.

Study	Data/background	Generation control	Synthesis-linked reference/mask
Alsinan *et al.* [39]	Bone US	Label-conditioned	Y
Liang *et al.* [38]	Lung, hip, ovary	Mask + editable sketch	Y
Liang and Chen [16]	Thyroid US; dataset NR	Latent augmentation	N
Yang *et al.* [17]	Thyroid US; dataset NR	Clinical/class knowledge	N
Zhao *et al.* [18]	Multimodal thyroid US	Semantic/domain control	N
Dai *et al.* [19]	Thyroid US + reports	Text/clinical descriptors	P
Prochazka and Zeman [20]	TN3K + real backgrounds	Masked inpainting	Y
Lai *et al.* [21]	Grayscale US to CEUS	Three-phase translation	NA

Note: Y = directly demonstrated and adequately reported; P = partial, indirect, or incomplete evidence; N = explicitly absent or unused; NR = insufficient reporting; NA = not applicable. In the final column, Y requires a spatial reference generated or preserved with, and aligned to, the synthetic target; a conditioning or input-only mask does not qualify as Y.

Table II shows that explicit synthesis-linked references or masks are concentrated in label-conditioned, sketch-guided, and inpainting methods. Earlier thyroid-specific GAN studies primarily controlled latent, class, or semantic information, while recent diffusion models introduced stronger text-based or spatial control. This distinction motivates a separate comparison of validation and accessibility.

Table III
VALIDATION, EXTERNAL TESTING, ACCESS, AND REPRODUCIBILITY.

Study	AI/reader validation	External	Code/data	Repro.
Alsinan *et al.* [39]	Downstream segmentation	N	NR/NR	P
Liang *et al.* [38]	User study + segmentation	N	Y/NR	P
Liang and Chen [16]	Augmentation utility	NR	NR/NR	NR
Yang *et al.* [17]	Classification	NR	NR/NR	NR
Zhao *et al.* [18]	Adaptation/classification	NR	NR/NR	NR
Dai *et al.* [19]	Classification + generation	Y	Y/P	Y
Prochazka and Zeman [20]	Three segmenters	Y	Y/Y	Y
Lai *et al.* [21]	Metrics + radiologists	Y	P/NR	P

Note: Y = directly demonstrated and adequately reported; P = partial, indirect, or incomplete evidence; N = explicitly absent or unused; NR = insufficient reporting; NA = not applicable. Code/data entries report code and data separately. "Repro." denotes reported artifacts and documentation, not independent replication.

Table III indicates that downstream testing has become more rigorous in recent thyroid studies. Prochazka and Zeman combined real-background insertion, synthesis-linked masks, public code and data, and cross-dataset segmentation tests; Dai *et al.* emphasized clinical text control; and Lai *et al.* added external-cohort and radiologist evaluation. Nevertheless, access to clinical data, calibration, and independent replication remain inconsistent.

Table IV
DESIGN-SENSITIVE METHODOLOGICAL APPRAISAL SUMMARY.

Instrument	Applied to	Main finding	Interpretive qualification
QuADS [43]	Cross-design evidence	Aims and methods were clearer than sampling rationale, stakeholder involvement, and reproducibility	No quality cutoff was imposed; item patterns informed the synthesis
CLAIM 2024 [44]	Medical-imaging AI subset	Recent studies more clearly reported partitions, metrics, and external tests	Code/data access, calibration, and provenance remained heterogeneous
PROBAST +AI [45]	Applicable prediction studies	External evaluation improved in recent thyroid studies	Retrospective data, calibration, and missing-data reporting remained recurrent concerns

Table IV reports the principal findings derived from each instrument within its applicable scope without collapsing the three frameworks into an artificial common scale. The table provides a concise domain-level synthesis rather than a common numerical ranking. Tables II and III show that the remaining technical fragmentation lies in combining complementary strengths within one thyroid B-mode pipeline, not in the absence of every individual component.

D. Hybrid methods

Hybrid methods retain selected physical principles while reducing computational cost. Goksel and Salcudean combined deformable 3D anatomy with B-mode synthesis [13], while subsequent approaches incorporated ray models, convolution, anatomical volumes, and GPU acceleration [14]. These methods can balance anatomical plausibility and controllability, but thyroid specialization, radiological descriptor interfaces, automatic masks, and downstream AI validation remain insufficiently evaluated. Superiority over fully physical or data-driven alternatives has not been established.

E. GANs and generative models

Generative ultrasound studies include breast DCGANs [36], [37], sketch-conditioned synthesis [38], musculoskeletal diffusion [40], and paired bone image–label generation [39]. Earlier thyroid work emphasized GAN augmentation, knowledge-guided classification, or domain adaptation [16]–[18]. Dai *et al.* subsequently demonstrated text-based and clinical-feature control, while Prochazka and Zeman demonstrated real-image inpainting with synthesis-linked masks and cross-dataset segmentation gains [19], [20]. The remaining question is whether descriptor control, lesion insertion, synthesis-linked annotation, and broad validation can be combined without introducing simulator-specific shortcuts.

F. AI for segmentation

Thyroid U-Nets and attention networks report strong segmentation, including TRFE+ on TN3K and multicenter Dice scores above 0.91 [7], [23]. DDTI is small and single-center [8], TN3K uses manual masks [7], [9], and TN5000 provides boxes rather than pixel masks [10]. Prochazka and Zeman showed that diffusion-based augmentation improved cross-dataset generalization: DeepLabV3+ Dice on TDID increased from 38.2% to 59.1%, with gains also reported on TUCC [20]. These results establish the feasibility of synthesis-linked reference masks, although their clinical fidelity still requires evaluation.

G. Studies on the thyroid

Thyroid generative evidence now covers augmentation [16], knowledge-guided classification [17], multimodal adaptation [18], text-guided diffusion [19], real-background inpainting with external segmentation tests [20], and synthetic CEUS with external-cohort and radiologist evaluation [21]. No single study combined all of the following: explicit TI-RADS-like descriptor control, real-background B-mode insertion, synthesis-linked masks, acoustic/device variation, and prospective clinical or multicenter downstream validation. This is a corpus-bounded integration gap, not an absence of the individual capabilities.

H. Integrated interpretation of the comparison

The synthesis indicates that acoustic control, paired labels, clinical semantics, public thyroid data, and external validation are now demonstrable but remain distributed across different implementations [19]–[21]. A real–synthetic hybrid simulator is therefore one author-derived hypothesis for integrating these capabilities, not evidence that physical, label-conditioned, GAN, or diffusion alternatives are inferior.

V. Discussion

A. Limitations and risks of synthetic data

Synthetic data may contain artifacts, limited diversity, implausible descriptor combinations, or simulator-specific correlations that models learn instead of disease features. Generative systems may also underrepresent rare phenotypes or reproduce biases present in their training data [15], [40]. Evaluation should therefore combine expert review, descriptor fidelity, artifact and memorization checks, calibration, target-task utility, and external real-world data.

B. Validation, reproducibility, and review limitations

Validation ranged from qualitative realism, SSIM, and PSNR to Dice, IoU, classification, radiologist assessment, and external-cohort testing. Recent studies broadened this profile but did not consistently combine perceptual, acoustic, downstream, multicenter, and cross-equipment evaluation [19]–[21]. The review may have omitted relevant work not indexed in the four selected databases, and retrieval remained sensitive to the scope and English-language formulation of the search terms. In addition, the rapid evolution of generative models—particularly diffusion-based methods—makes this synthesis time-sensitive. Verification of the historical screening flow during revision does not alter the study-by-study technical comparisons, but it limits exact replication of the original search process. The targeted update is therefore reported separately from the historical evidence set. Foundational architecture, metric, guideline, review, and framework sources broaden methodological coverage but are interpreted as contextual rather than thyroid-specific clinical evidence.

VI. Research gaps and future directions

A. Critical gaps

Four integration gaps remain underexplored: combining TI-RADS-like descriptor control with real-background insertion and synthesis-linked masks; parameterizing acquisition factors such as manufacturer, transducer frequency, gain, dynamic range, focus, and depth; testing real-only, synthetic-only, and mixed training across centers and devices; and linking image realism to calibration, clinical decisions, and prospective utility. Physical, statistical, hybrid, and generative designs offer different ways to address these gaps and should be compared empirically.

B. Evidence-informed future research directions

Future experiments could compare architectures that jointly control nodule shape, margins, echogenicity, echogenic foci, location, and posterior artifacts while preserving a synthesis-linked mask [19], [20]. Conditional GAN, diffusion, physical, and hybrid alternatives should be compared with conventional augmentation and CEUS translation rather than ranked a priori [21]. Validation should stratify manufacturers, transducer

frequencies, gain and dynamic-range presets, focus/depth settings, and operator variability; use blinded readers and standardized TI-RADS descriptors; and perform leave-one-center/device-out and preregistered real-only, synthetic-only, and mixed-training experiments. Python and MONAI may aid integration [46], but reproducibility depends on versioned code, weights, environments, parameters, and data provenance.

VII. Conclusion

This review maps the rapidly evolving intersection of ultrasound simulation, thyroid semantics, synthesis-linked annotation, and AI validation. Recent text-guided diffusion, real-background inpainting, and synthetic CEUS studies demonstrate the feasibility of several previously separated capabilities, while their integration remains insufficiently evaluated. Rather than favoring a single architecture, future work should compare physical, statistical, hybrid, GAN-based, and diffusion-based approaches under common experimental and clinical validation protocols. Their value should be judged by controllability, annotation fidelity, reproducibility, external generalization, and clinically meaningful utility.

Acknowledgment

The authors acknowledge CAPES (Finance Code 001), CNPq, the Araucária Foundation, and Paraná SETI for financial support.

References

[1] D. S. Dean and H. Gharib, "Epidemiology of thyroid nodules," *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 22, no. 6, pp. 901–911, 2008, doi: 10.1016/j.beem.2008.09.019.

[2] B. R. Haugen *et al.*, "2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer," *Thyroid*, vol. 26, no. 1, pp. 1–133, 2016, doi: 10.1089/thy.2015.0020.

[3] F. N. Tessler *et al.*, "ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White paper of the ACR TI-RADS committee," *Journal of the American College of Radiology*, vol. 14, no. 5, pp. 587–595, 2017, doi: 10.1016/j.jacr.2017.01.046.

[4] T. Liu *et al.*, "Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks," *Medical Image Analysis*, vol. 58, art. 101555, 2019, doi: 10.1016/j.media.2019.101555.

[5] W.-T. Yang, B.-Y. Ma, and Y. Chen, "A narrative review of deep learning in thyroid imaging: current progress and future prospects," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 2, pp. 2069–2088, 2024, doi: 10.21037/qims-23-908.

[6] J. Chen, H. You, and K. Li, "A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images," *Computer Methods and Programs in Biomedicine*, vol. 185, art. 105329, 2020, doi: 10.1016/j.cmpb.2020.105329.

[7] H. Gong *et al.*, "Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules," *Computers in Biology and Medicine*, vol. 155, art. 106389, 2023, doi: 10.1016/j.compbiomed.2022.106389.

[8] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," in *Proc. SPIE 9287, 10th International Symposium on Medical Information Processing and Analysis*, art. 92870W, 2015, doi: 10.1117/12.2073532.

[9] H. Gong *et al.*, "Multitask learning for thyroid nodule segmentation with thyroid region prior," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 257–261, doi: 10.1109/ISBI48211.2021.9434087.

[10] H. Zhang *et al.*, "TN5000: An ultrasound image dataset for thyroid nodule detection and classification," *Scientific Data*, vol. 12, art. 1437, 2025, doi: 10.1038/s41597-025-05757-4.

[11] J. A. Jensen, "Field: A program for simulating ultrasound systems," *Medical & Biological Engineering & Computing*, vol. 34, suppl. 1, part 1, pp. 351–353, 1996.

[12] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of Biomedical Optics*, vol. 15, no. 2, art. 021314, 2010, doi: 10.1117/1.3360308.

[13] O. Goksel and S. E. Salcudean, "B-Mode ultrasound image simulation in deformable 3-D medium," *IEEE Transactions on Medical Imaging*, vol. 28, no. 11, pp. 1657–1669, 2009, doi: 10.1109/TMI.2009.2016561.

[14] C. M. Solano-Cordero, N. Encina-Baranda, M. Pérez-Liva, and J. L. Herraiz, "Recent advances in B-mode ultrasound simulators," *Applied Sciences*, vol. 15, no. 23, art. 12535, 2025, doi: 10.3390/app152312535.

[15] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, art. 101552, 2019, doi: 10.1016/j.media.2019.101552.

[16] J. Liang and J. Chen, "Data augmentation of thyroid ultrasound images using generative adversarial network," in *2021 IEEE International Ultrasonics Symposium (IUS)*, 2021, pp. 1–4, doi: 10.1109/IUS52206.2021.9593403.

[17] W. Yang *et al.*, "DScGANS: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and S3VM for ultrasonography thyroid nodules classification," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, vol. 11767, 2019, pp. 558–566, doi: 10.1007/978-3-030-32251-9_61.

[18] J. Zhao *et al.*, "Semantic consistency generative adversarial network for cross-modality domain adaptation in ultrasound thyroid nodule classification," *Applied Intelligence*, vol. 52, no. 9, pp. 10369–10383, 2022, doi: 10.1007/s10489-021-03025-7.

[19] F. Dai *et al.*, "Improving AI models for rare thyroid cancer subtype by text guided diffusion models," *Nature Communications*, vol. 16, art. 4449, 2025, doi: 10.1038/s41467-025-59478-8.

[20] A. Prochazka and J. Zeman, "Domain adaptation of stable diffusion for ultrasound inpainting: a synthetic data approach for enhanced thyroid nodule segmentation," *Journal of Biomedical Informatics*, vol. 173, art. 104963, 2026, doi: 10.1016/j.jbi.2025.104963.

[21] M. Lai *et al.*, "Synthesize contrast-enhanced ultrasound image of thyroid nodules via generative adversarial networks," *European Radiology*, vol. 36, no. 3, pp. 1630–1643, 2026, doi: 10.1007/s00330-025-11928-z.

[22] T. Zheng *et al.*, "Segmentation of thyroid glands and nodules in ultrasound images using the improved U-Net architecture," *BMC Medical Imaging*, vol. 23, art. 56, 2023, doi: 10.1186/s12880-023-01011-8.

[23] J. L. Cozzi *et al.*, "Multi-institutional development and testing of attention-enhanced deep learning segmentation of thyroid nodules on ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 20, no. 2, pp. 259–267, 2025, doi: 10.1007/s11548-024-03294-w.

[24] J. C. Bamber and R. J. Dickinson, "Ultrasonic B-scanning: a computer simulation," *Physics in Medicine and Biology*, vol. 25, no. 3, pp. 463–479, 1980, doi: 10.1088/0031-9155/25/3/006.

[25] D. R. Foster, M. Arditi, F. S. Foster, M. S. Patterson, and J. W. Hunt, "Computer simulations of speckle in B-scan images," *Ultrasonic Imaging*, vol. 5, no. 4, pp. 308–330, 1983, doi: 10.1177/016173468300500403.

[26] F. Varray *et al.*, "CREANUIS: a non-linear radiofrequency ultrasound image simulator," *Ultrasound in Medicine & Biology*, vol. 39, no. 10, pp. 1915–1924, 2013, doi: 10.1016/j.ultrasmedbio.2013.04.005.

[27] D. Garcia, "SIMUS: An open-source simulator for medical ultrasound imaging. Part I: Theory & examples," *Computer Methods and Programs in Biomedicine*, vol. 218, art. 106726, 2022, doi: 10.1016/j.cmpb.2022.106726.

[28] A. Cigier, F. Varray, and D. Garcia, "SIMUS: An open-source simulator for medical ultrasound imaging. Part II: Comparison with four simulators," *Computer Methods and Programs in Biomedicine*, vol. 220, art. 106774, 2022, doi: 10.1016/j.cmpb.2022.106774.

[29] A. Rodriguez-Molares *et al.*, "The UltraSound ToolBox," in *2017 IEEE International Ultrasonics Symposium (IUS)*, 2017, pp. 1–4, doi: 10.1109/ULTSYM.2017.8092389.

[30] M. A. Lediju Bell *et al.*, "Challenge on ultrasound beamforming with deep learning (CUBDL)," in *2020 IEEE International Ultrasonics Symposium (IUS)*, 2020, pp. 1–5, doi: 10.1109/IUS46767.2020.9251434.

[31] D. Hyun *et al.*, "Deep learning for ultrasound image formation: CUBDL evaluation framework and open datasets," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 12, pp. 3466–3483, 2021, doi: 10.1109/TUFFC.2021.3094849.

[32] A. Marion and D. Vray, "Toward a real-time simulation of ultrasound image sequences based on a 3-D set of moving scatterers," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 56, no. 10, pp. 2167–2179, 2009, doi: 10.1109/TUFFC.2009.1299.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[34] V. Kumar *et al.*, "Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning," *IEEE Access*, vol. 8, pp. 63482–63496, 2020, doi: 10.1109/ACCESS.2020.2982390.

[35] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014, doi: 10.5555/2969033.2969125.

[36] T. Fujioka *et al.*, "Breast ultrasound image synthesis using deep convolutional generative adversarial networks," *Diagnostics*, vol. 9, no. 4, art. 176, 2019, doi: 10.3390/diagnostics9040176.

[37] S. Zama *et al.*, "Clinical utility of breast ultrasound images synthesized by a generative adversarial network," *Medicina*, vol. 60, no. 1, art. 14, 2024, doi: 10.3390/medicina60010014.

[38] J. Liang *et al.*, "Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis," *Medical Image Analysis*, vol. 79, art. 102461, 2022, doi: 10.1016/j.media.2022.102461.

[39] A. Z. Alsinan, C. Rule, M. Vives, V. M. Patel, and I. Hacihaliloglu, "GAN-based realistic bone ultrasound image and label synthesis for improved segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*, vol. 12266, 2020, pp. 795–804, doi: 10.1007/978-3-030-59725-2_77.

[40] S. Katakis *et al.*, "Generation of musculoskeletal ultrasound images with diffusion models," *BioMedInformatics*, vol. 3, no. 2, pp. 405–421, 2023, doi: 10.3390/biomedinformatics3020027.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[42] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, art. n71, 2021, doi: 10.1136/bmj.n71.

[43] R. Harrison, B. Jones, P. Gardner, and R. Lawton, "Quality assessment with diverse studies (QuADS): an appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies," *BMC Health Services Research*, vol. 21, art. 144, 2021, doi: 10.1186/s12913-021-06122-y.

[44] A. S. Tejani *et al.*, "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update," *Radiology: Artificial Intelligence*, vol. 6, no. 4, art. e240300, 2024, doi: 10.1148/ryai.240300.

[45] K. G. M. Moons *et al.*, "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, art. e082505, 2025, doi: 10.1136/bmj-2024-082505.

[46] M. J. Cardoso *et al.*, "MONAI: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022, doi: 10.48550/arXiv.2211.02701.

Assessment of the Effects of Falls on the Femoral Neck in Elderly People with Osteoporosis Using the Finite Element Method.

Juliana Sayuri Sakamoto
Dept. Medical Imaging, Hematology and Clinical Oncology
Ribeirão Preto Medical School – FMRP/USP
Ribeirão Preto, Brazil
jusakamoto@usp.br

Anderson Freitas
Hospital Ortopédico e Medicina Especializada (HOME), Orthopedic Trauma Service
Hospital Regional do Gama
Brasília, Brazil
andfreitas28@gmail.com

João Victor Pereira Souza
Dept. Orthopedics and Anesthesiology
Ribeirão Preto Medical School – FMRP/USP
Ribeirão Preto, Brazil
joaovictorps@usp.br

Felipe Pratalli Martins
Dept. Orthopedics and Anesthesiology
Ribeirão Preto Medical School – FMRP/USP
Ribeirão Preto, Brazil
felipepratalli@usp.br

Ana Paula Macedo
Dept. Orthopedics and Anesthesiology
Ribeirão Preto Medical School – FMRP/USP
Ribeirão Preto, Brazil
anamacedo@usp.br

Antonio Carlos Shimano
Dept. Orthopedics and Anesthesiology
Ribeirão Preto Medical School – FMRP/USP
Ribeirão Preto, Brazil
ashimano@fmrp.usp.br

***Abstract* — The increase in life expectancy has been accompanied by a higher prevalence of osteoporosis, especially among postmenopausal women, due to estrogen deficiency, and, consequently, by an increased risk of proximal femoral fractures. Understanding the mechanical behavior of the osteoporotic femur under different loading conditions may contribute to the development of preventive and therapeutic strategies. In this context, the Finite Element Method (FEM) provides a computational approach for investigating stress and displacement distributions under controlled loading conditions. The objective of this study was to investigate the mechanical response of an osteoporotic femur under orthostatic loading and a lateral fall onto greater trochanter, with emphasis on the distribution of principal stresses and displacement in the proximal femur. Computed tomography images from an 84-year-old female patient with osteoporosis were used to reconstruct the femoral geometry. Three-dimensional reconstruction, BioCAD modeling, finite element preprocessing, and numerical simulations were performed using InVesalius®, Rhinoceros®, SimLab®, and OptiStruct, respectively. The model included cortical and trabecular bone with distinct elastic properties and was subjected to loads ranging from 500 to 2500N in the orthostatic and lateral fall configurations. The results demonstrated changes in the magnitude and spatial distribution of principal stresses according to the loading condition and applied force. The lateral fall configuration produced higher compressive stress magnitudes at higher loads, with stress concentrations predominantly located in the proximal femur and femoral neck. These findings indicate greater mechanical demand in the femoral neck under lateral fall loading; however, stress concentration alone does not establish the occurrence or probability of fracture. Therefore, the present results should be interpreted as an assessment of mechanical stress distribution rather than direct evidence of fracture.**

Key-words — femur, elderly, osteoporosis, finite element method.

São Paulo Research Foundation (FAPESP) – 2021/04579-0

I. Introduction

Life expectancy has been increasing worldwide, with the population in this age group growing by 18% over five years in Brazil, surpassing the 30.2 million mark for adults aged 60 and older in 2017, according to data from the Brazilian Institute of Geography and Statistics [1]. The growing size of this age group is accompanied by an increase in bone diseases, such as osteoporosis; and, consequently, a rise in the incidence of fractures.

Osteoporosis is a progressive metabolic bone disease characterized by a gradual decrease in bone mass, accompanied by changes in the microarchitecture of bone tissue, leading to increased bone fragility and a higher risk of fractures. This loss is most prevalent in elderly patients, particularly in postmenopausal women, given the estrogen deficiency observed, which can lead to an accelerated decrease in bone density of up to 5% per year in trabecular bone and 2% to 3% per year in cortical bone [2].

Among fractures, the most common are those of the femur, particularly in the proximal region, for which osteoporosis is a significant risk factor [3], with a higher incidence among postmenopausal women. According to Soares et al. [4], the incidence of femoral fractures in women aged 60 years or older was 67% higher than that in men in the same age group.

In this context, the incidence is expected to triple by 2050, potentially reaching 6.5 million femur fractures worldwide [5], contributing to growing public health challenges due to the high cost of treatment and the impact on quality of life.

Since the skeletal structure is viewed as a load-bearing system subjected to a set of forces that can lead to deformities, biomechanics serves as a valuable tool for analyzing human structure and function, aiding in the study and understanding of the loads applied and their consequences, and acting as a fundamental link between the exact and biological sciences [6]. Thus, the understanding of proximal femur fractures can be achieved through software analysis, which, in this study, was performed using the Finite Element Method.

Therefore, the objective of the present study was to investigate the distribution of principal stresses and

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

displacement in an osteoporotic femur subjected to orthostatic loading and a lateral-fall configuration onto the greater trochanter. The analysis focused on identifying regions subjected to greater mechanical demand under the investigated loading conditions.

II. Material and Methods

A. Material

The research material consists of computed tomography (CT) images of an 84-year-old female patient with osteoporosis, obtained from the database of the Hospital Ortopédico de Medicina Especializada (HOME) in Brasília, Federal District. The study was approved by the Research Ethics Committee under protocol number (CAAE: 32179520.0.0000.0023) for the use of the CT images. This dataset was used for the reconstruction femoral geometries for subsequent Finite Element Method simulations.

B. Modeling of the Femoral Structure

The computed tomography images of the femur, obtained in .DICOM file format, were imported into the InVesalius® software (Fig. 1), where a three-dimensional reconstruction of the structure of interest was performed (Fig. 2) in .STL (Stereolithography) format.

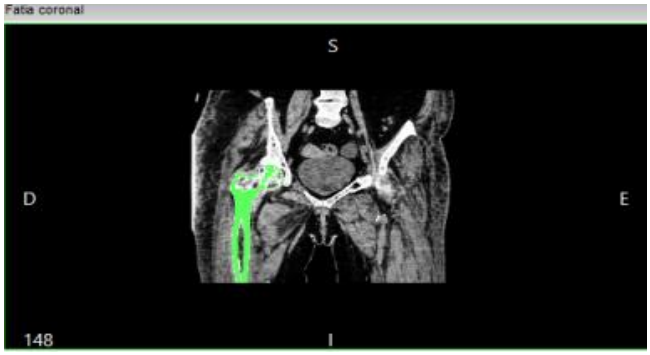


Fig. 1. Tomographic image in InVesalius® - area of interest.

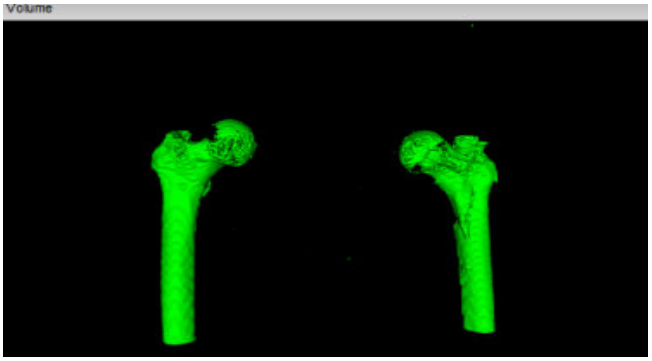


Fig. 2. Three-dimensional reconstruction of the femur in InVesalius®.

The .STL files were then imported into Rhinoceros® software to create the BioCAD® design. The cortical and trabecular regions were separately represented in the model, with minimal overlap between the two regions. This approach was adopted to represent the cortical and trabecular compartments of the os

The bone model was positioned with 9° of internal rotation and 10° of adduction—in the orthostatic position—and 10° of diaphyseal rotation and 10° of internal rotation—the lateral fall mechanism position over the greater trochanter (Figs. 3 and 4). These orientations were defined to reproduce the angles of position of the trauma.

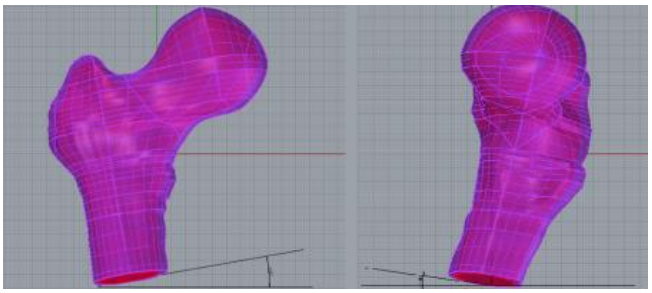


Fig. 3. Femur in the orthostatic position with angulation.

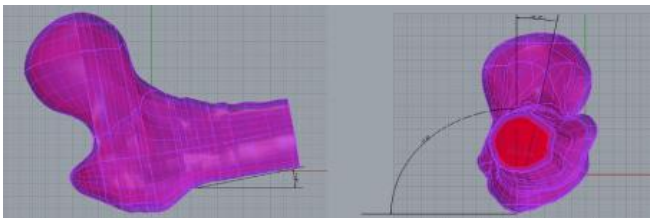


Fig. 4. Femur in a lateral fall position over the greater trochanter with angulation.

C. Finite Element Method Simulations

Once BioCAD® was finalized, a .STEP file (Standard for the Exchange of Product Model Data) was generated and imported into SimLAb® (Altair/HyperWorks), where finite element simulations were performed.

To analyze stress distribution using SimLAb®, the following steps were performed: mesh identification, control, and generation, as well as the determination of material properties—Young's modulus and Poisson's ratio [7] (Table 1).

TABLE 1. Properties of Materials

Material	Total elements	Total nodes	Modulus of Elasticity	Poisson's ratio
Cortical bone	332047	526384	13700	0.30
Trabecular bone	347637	552337	1370	0.30

Although these properties allow the numerical model to be solved, the present model does not incorporate a voxel-wise relationship between CT attenuation and local bone density. Consequently, the material properties should be considered region-specific homogeneous approximations rather than direct representations of the spatial heterogeneity of osteoporotic bone.

Finally, the osteoporotic femur model was subjected to loads in the orthostatic position and in a lateral fall scenario onto the greater trochanter, adhering to the boundary conditions to ensure that the system remained perfectly aligned without any displacement or rotation. These conditions were maintained consistently among the simulations to permit comparison between loading magnitudes and configurations. (Figs. 5 and 6).

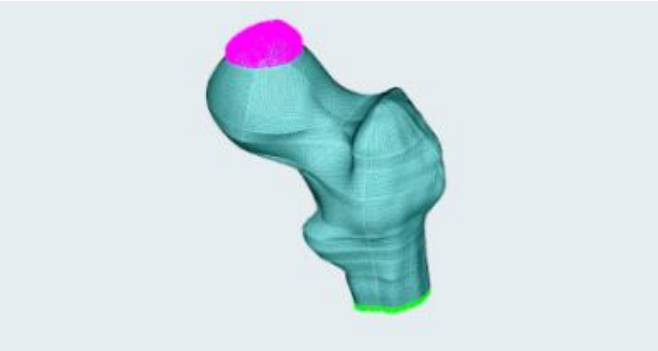


Fig. 5. Boundary conditions for the femur in the orthostatic position. Load application region (pink) and fixed region (green).

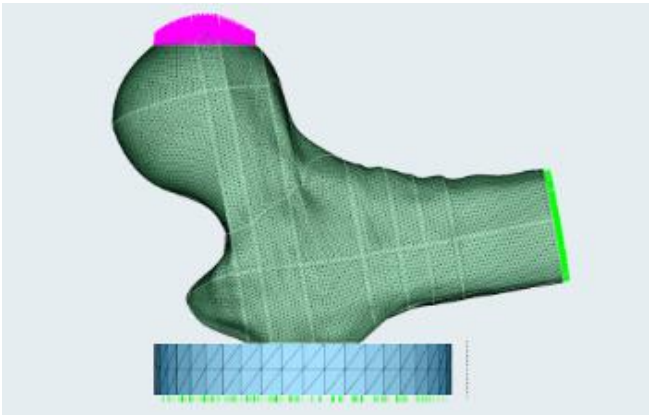


Fig. 6. Boundary conditions for the femur in the lateral drop mechanism position over the greater trochanter. Load application region (pink) and fixed region (green).

As a result, the locations with the highest stress concentrations were observed in the region of the proximal end of the femur, and the values for the principal maximum and minimum stresses and displacement were obtained.

III. Results

The osteoporotic femur model was subjected to five different loads—500 N, 1000 N, 1500 N, 2000 N, and 2500 N—represent different body-weight conditions that may be encountered among patients, both in the orthostatic position and in a lateral fall onto the greater trochanter. The parameters analyzed were: Displacement, Maximum Principal Stress, and Minimum Principal Stress.

A. Displacement

The displacement analysis showed a progressive increase in femoral deformation with increasing applied load in both loading configurations (Tables 2 and 3). In the orthostatic position, displacement increased from 0.286 mm at 500 N to 1.431 mm at 2500 N. In the lateral-fall configuration over the greater trochanter, displacement increased from 0.141 mm to 0.706 mm over the same loading range.

For both configurations, the displacement increased approximately proportionally to the applied load, indicating a consistent mechanical response of the model within the investigated loading range. The relationship between load and displacement resulted in calculated stiffness values of 1746.72 N/mm for the orthostatic position and 3539.82 N/mm for the lateral-fall configuration.

The progressive increase in displacement with increasing load, together with the changes observed in the principal stress distributions, demonstrates that the magnitude and configuration of the applied load influence the mechanical response of the proximal femur. In particular, the displacement results provide complementary information to the stress analysis by describing the overall deformation of the femoral model under the different loading conditions.

TABLE 2. Displacement and Relative Stiffness – Orthostatic Position.

Loading	Displacement (mm)	Relative Stiffness
500 N	0.286	1746.72 N/mm
1000 N	0.572	
1500 N	0.859	
2000 N	1.145	
2500 N	1.431	

TABLE 3. Displacement Value and Relative Stiffness – Position of the Lateral Fall Mechanism Over the Greater Trochanter.

Loading	Displacement (mm)	Relative Stiffness
500 N	0.141	3539.82 N/mm
1000 N	0.282	
1500 N	0.424	
2000 N	0.565	
2500 N	0.706	

B. Maximum Principal Stress

The spatial distribution of maximum principal stress demonstrated that the location of tensile stress concentrations changed according to the loading configuration.

In the orthostatic configuration, the highest tensile stress concentrations were predominantly located in the superior region of the femoral neck. In the lateral fall configuration, the highest tensile stress concentrations were observed primarily in the anterolateral region at the base of the femoral neck (Fig. 7).

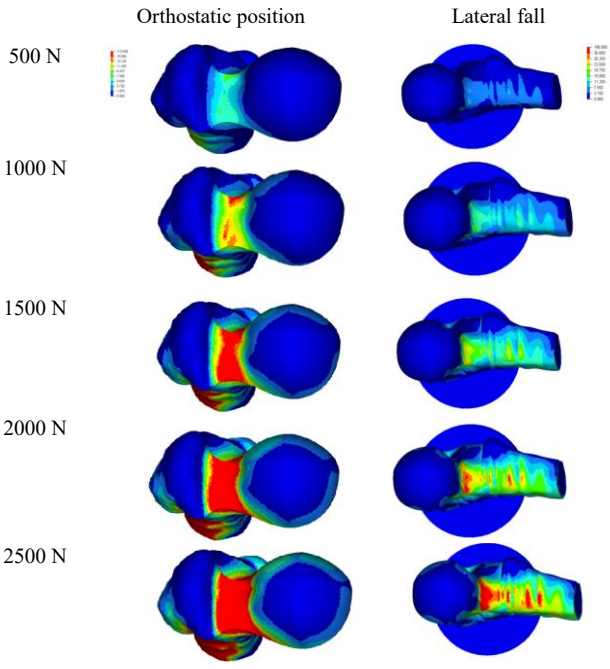


Fig. 7. Maximum Principal Stress – orthostatic position and lateral fall position over the greater trochanter.

The maximum principal stress values increased with increasing load. In the orthostatic configuration, the values increased from 8.358 at 500N to 32.856 at 2500N. In the lateral fall configuration, the corresponding values ranged from 8.901 to 35.497 (Table 4).

These results indicate that increasing the applied load resulted in greater tensile mechanical demand and that the loading configuration influenced the spatial distribution of this demand.

TABLE 4. MAXIMUM PRINCIPAL STRESS VALUES (MPA).

Loading	Orthostatic	Lateral fall over the greater trochanter
500 N	8.358	8.901
1000 N	15.534	14.904
1500 N	23.301	21.610
2000 N	28.727	32.924
2500 N	32.856	35.497

C. Minimum Principal Stress

The minimum principal stress demonstrated a different spatial distribution from that observed for the maximum principal stress.

In the orthostatic configuration, the greatest compressive stress concentrations were predominantly observed in the anterolateral region at the base of the femoral neck. Under lateral fall loading, the concentration shifted toward the superior region of the femoral neck (Fig. 8).

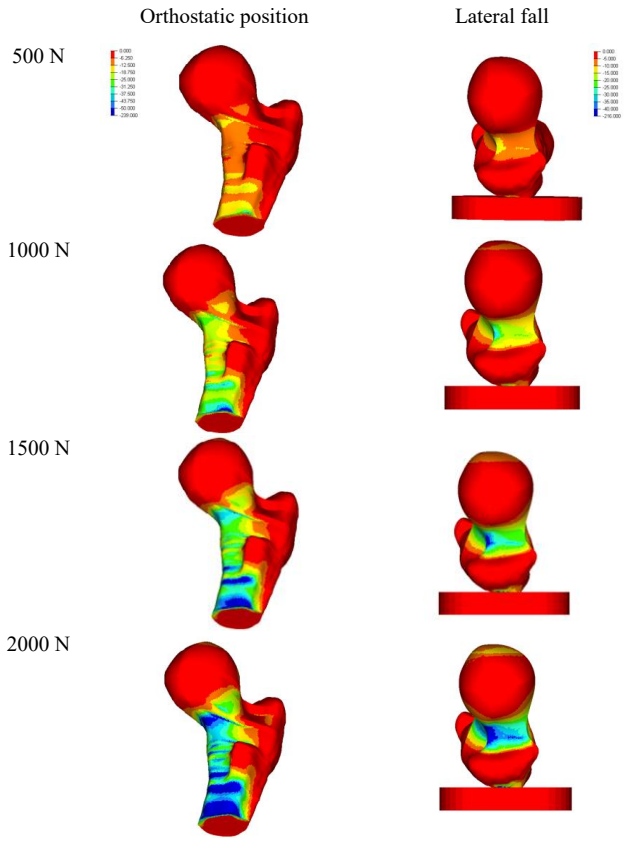


Fig. 8. Minimum Principal Stress – orthostatic position and lateral fall position over the greater trochanter.

The magnitude of the minimum principal stress increased progressively with the applied load. In the orthostatic configuration, values ranged from –13.135 at 500N to –62.504 at 2500N. In the lateral fall configuration, the values ranged from –12.977 to –72.374 (Table 5).

Thus, among the investigated loading conditions, the lateral fall configuration generated greater compressive stress magnitudes at the higher applied loads.

TABLE 5. MINIMUM PRINCIPAL STRESS VALUES (MPA).

Loading	Orthostatic	Lateral fall over the greater trochanter
500 N	-13.135	-12.977
1000 N	-26.302	-29.729
1500 N	-42.117	-43.424
2000 N	-50.003	-57.899
2500 N	-62.504	-72.374

The numerical results demonstrate that both the magnitude and location of principal stresses depend on the applied load and femoral loading configuration. The femoral neck consistently represented a region of elevated mechanical demand, although the precise location of the maximum and minimum principal stresses changed according to the loading condition. The results therefore provide a mechanical description of the response of the modeled osteoporotic femur and establish a basis for future analyses incorporating heterogeneous CT-derived material properties, sensitivity analysis, and validated fracture criteria.

IV. DISCUSSION

The present study investigated the mechanical response of an osteoporotic femur under two loading configurations: an orthostatic condition and a lateral fall onto the greater trochanter. The finite element simulations demonstrated that both the magnitude and spatial distribution of principal stresses were influenced by the loading configuration and applied force.

The femoral neck exhibited the most relevant stress concentrations in both configurations. Under orthostatic loading, tensile stresses were concentrated predominantly in the superior region of the femoral neck, whereas lateral-fall loading resulted in a shift in the tensile stress concentration toward the anterolateral region at the base of the femoral neck. A corresponding change was observed for compressive stresses, which were concentrated primarily in the anterolateral region of the femoral neck during orthostatic loading and in the superior region during lateral-fall loading.

These findings are qualitatively consistent with previous biomechanical investigations of the proximal femur. Bakker et al. [8] reported that during sideways falls, compressive stresses are concentrated in the superior region of the femoral neck, whereas tensile stresses may develop near the base of the femoral neck. Similarly, Lotz et al. [9] demonstrated that the distribution and magnitude of stress within the proximal femur vary substantially between normal gait and fall conditions. The agreement between these studies and the present numerical results supports the relevance of the modeled loading configurations for investigating mechanical demand in the proximal femur.

The present simulations also showed that the absolute magnitude of the minimum principal stress increased with increasing load and was greater in the lateral-fall configuration at the higher loading levels. For example, at 2500 N, the minimum principal stress was −72.374 in the lateral-fall configuration compared with −62.504 in the orthostatic configuration. This indicates that the lateral-fall condition generated greater compressive mechanical demand within the modeled femoral structure.

However, the present results can only indicate regions subjected to relatively greater mechanical demand. Besides that, the robustness of the present findings should also be investigated through sensitivity analysis. Validation represents another important consideration. The present study was based on a single osteoporotic femur and did not include direct experimental validation. Consequently, the numerical stress values should not be interpreted as experimentally validated fracture thresholds.

Despite these limitations, the present study provides a computational assessment of how orthostatic and lateral-fall loading conditions alter the distribution of mechanical stresses in an osteoporotic femur. The results emphasize the importance of the femoral neck as a region subjected to substantial mechanical demand and demonstrate that loading configuration affects both the magnitude and location of principal stresses.

Therefore, the main contribution of this study is not the prediction of fracture itself, but the characterization of the mechanical stress distribution under two clinically relevant loading configurations. The results provide a basis for subsequent, more comprehensive finite element investigations.

V. Conclusion

Finite Element Analysis demonstrated that the mechanical response of the osteoporotic femur is strongly influenced by the loading configuration and magnitude of the applied force. The femoral neck presented the most relevant concentrations of principal stresses, with changes in their location between orthostatic and lateral-fall conditions. The lateral-fall configuration produced greater compressive stress magnitudes at the higher loads investigated, indicating increased mechanical demand in the proximal femur under this loading condition. The present findings should therefore be regarded as a preliminary computational assessment of stress distribution in a single osteoporotic femur.

References

[1] Instituto Brasileiro de Geografia e Estatítisca (IBGE), “Número de idosos cresce 18% em 5 anos e ultrapassa 30 milhões em 2017”, Agência IBGE de Notícias, Rio de Janeiro, 2018.

[2] R.W. Lash, J.M. Nicholson, L. Velez, R.V. Harrison, and J. McCort, “Diagnosis and management of osteoporosis”, Primary Care: Clinics in Office Practice, vol. 36, n. 2, pp. 181-198, 2009.

[3] M.A.R. Loures et al., “Diretrizes da Sociedade Brasileira de Reumatologia para diagnóstico e tratamento da osteoporose em homens”, Revista Brasileira de Reumatologia, vol. 57, n. S2, pp. 497-514, 2017.

[4] D.S. Soares, L.M. de Mello, A.S. da Silva, E.Z. Martinez, A.A. Nunes, “Fraturas de fêmur em idosos no Brasil: Análise espaço-temporal de 2008 a 2012”, Cadernos de Saúde Pública, vol. 30, n. 12, pp. 2669-2678, 2014.

[5] L.P. Lustosa and E.O. Bastos, “Fraturas proximais do fêmur em idosos: Qual o melhor tratamento?”, Acta Ortopedica Brasileira, vol. 17, n. 5, pp. 309-312, 2009.

[6] P.F. Niederer, “Mathematical foundations of biomechanics”, Critical Reviews in Biomedical Engineering, vol. 38, n. 6, pp. 533-577, 2010.

[7] M.B. Toniollo et al., “A three-dimensional finite element analysis of the stress distribution on morse taper implants surface”, Journal of Prosthodontic Research, vol. 57, n. 3, pp. 206-212, 2013.

[8] P.M. Bakker et al., “During sideways falls proximal femur fractures initiate in the superolateral cortex: evidence from high-speed video of simulated fractures”, Journal of biomechanics, vol. 42, n. 12, pp. 1917-1925, 2009.

[9] J.C. Lotz, E.J. Cheal, and W.C. Hayes, “Stress distributions within the proximal femur during gait and falls: Implications of osteoporotic fracture”, Osteoporosis Int, vol. 5, pp. 252-261, 1995.

Benchmarking Language Agents on Continuous Glucose Monitoring Data across Stratified Harness Resources

Mario Cesar Freire Dias Filho
GREat – PPGETI
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
mariocesarfreire@alu.ufc.br

Rodrigo de Azeredo Siqueira
Universidade Iguaçu
Nova Iguaçu, Brazil
rodrigoendocrinologista@gmail.com

Julia Carpanezzi La Pastina
SBD-SC
Brazil
julia.endoped@gmail.com

Susana Viegas Chen
Endoniterói Clínica Médica
Niterói, Brazil
susanavchen@gmail.com

Danielo G. Gomes
GREat – PPGETI
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
danielo@ufc.br

Abstract*—Language models remain unreliable when a clinical answer must be computed from a continuous glucose monitoring (CGM) series. This paper evaluates which computational *harness* resources are required by different classes of CGM question. The harness is the controlled execution environment that fixes which representation of the patient data an agent may read and which operations it may run. Extending the LLM-CGM benchmark to an agentic setting, we hold the patient data and the clinical question fixed and vary the harness across eleven configurations, from prompt-only access to executable code, a curated CGM tool library, and a family that reads a rendered PDF report; a deterministic oracle that no agent can query defines correctness. Over 30 patients, three model families and 22 executed system instances, the largest gain with the model held fixed comes from replacing text prompting with executable code (48.3% to 71.4% under Gemma), and the increments above the code level are small. The curated-tool-only configuration produced the most dangerous false negatives of any raw-data route, its cleaning step having removed the sensor-gap information the anomaly questions depend on, whereas the P0–P3 report-only routes recorded no such failure while answering only about a third of the matrix. Sampling parameters were left at model-specific serving defaults, and beyond a 30 s per-call timeout no token or step ceiling was imposed, so contrasts between model families are descriptive, the contrasts above the code level are not iso-compute, and latency and token usage support no ranking. Separately, three endocrinologists assessed 30 automatically generated Ambulatory Glucose Profile (AGP) reports, one reviewer per report: fidelity to the source on the nine core glycemic metrics was 97.41%, though a retrospective, single-center, single-reader assessment supports the plausibility of the pipeline rather than its clinical validity. Across these conditions the accuracy differences attributable to harness configuration exceeded those observed between model families, an outcome specific to the models and settings evaluated.

***Index Terms*—continuous glucose monitoring, language models, agentic systems, benchmark, ambulatory glucose profile, clinical evaluation, time in range**

I. Introduction

Continuous glucose monitoring generates dense physiological time series that must be turned into standardized metrics, temporal patterns and clinically actionable interpretations [3], [4]. Language models have recently been applied to this task, generating narrative summaries and answering natural-language questions about CGM data [2]; the LLM-CGM benchmark formalized it over four guideline-derived question categories with real and synthetic patients [1]. Direct reasoning over long series, however, remains vulnerable to numerical and temporal errors, and one established remedy delegates arithmetic and signal processing to executable code [7].

Most existing evaluations vary the language model and leave the surrounding computational harness fixed. We invert that design: patient data and clinical questions are held constant while the input representation and the computational resources available to the agent vary across raw time series, graphical representations, executable code, curated CGM functions and rendered clinical reports. The question is operational: which harness resources each category of CGM question requires before the answer is accurate and safe enough, and where additional resource ceases to help. We are not aware of a previous CGM-interpretation benchmark that manipulates harness resources this way while holding data, questions and evaluation criteria constant. Efficiency is left out of that judgement, since the runs were bounded only by a per-call timeout (Section V-A) and their cost was not standardized.

By *computational harness* we mean the controlled execution environment mediating between the language model and the patient data: it fixes which representation of the signal the agent may read, bounds it to a declared set of permitted actions, and executes any code in a local sandbox. That sandbox is what makes local-only processing possible, since

at the code, tool and hybrid levels the model receives the schema, its own code and the values that code returns, never the series itself, which meets a data-governance requirement in clinical settings; the prompt-only levels are the deliberate exception, and Section V-A states per level what was sent. Varying the *harness resource* means varying this bundle while the model, the patient and the question stay fixed. By *oracle* we mean, in the established sense of a ground-truth test oracle [28], a deterministic reference implementation that computes the correct answer to each question from the raw signal, contains no language model, and is never exposed to the agents; it defines correctness and the ceiling against which every configuration is scored.

We contribute a harness-resource-stratified extension of LLM-CGM, defined by a matrix of resource levels, a task taxonomy with explicit interpretation axes, the deterministic oracle as ground truth and pre-specified hypotheses relating resource transitions to accuracy and to safety. We additionally evaluate an automatic AGP report-generation pipeline, combining an objective fidelity checklist with a Likert instrument for clinical usefulness, and compare eleven configurations across three model families (Gemma 4 31B, Kimi K2.7 Code, DeepSeek-V3.2) over a paired patient–question matrix, identifying the transitions that improve performance and the configurations that introduce clinically relevant failure modes.

II. Background and Related Work

CGM metrics are estimates derived from an irregularly sampled series subject to gaps when the sensor disconnects or saturates, consolidated by the Ambulatory Glucose Profile into a standardized two-week report embedded in contemporary diabetes care [5], [6]. Our oracle and safety definitions follow the international consensus: time-in-range targets and episode definitions [3], the glucose management indicator [4], and time in range as validated against microvascular outcomes [12]. Interpretation depends on the integrity and coverage of the signal as much as on its values: a missing interval degrades every derived metric, while a missed nocturnal hypoglycemia or an unreported time below 54 mg/dL is a direct safety risk. Evaluation therefore has to separate accuracy from coverage, abstention and clinically consequential false negatives.

The LLM-CGM benchmark queries CGM data across four guideline-anchored categories over five real and five synthetic patients, reporting that text-only prompting of long series is expensive and numerically fragile [1]; an earlier case study characterized the accuracy limits of LLM summaries of 14 days of data [2], and multidimensional medical-LLM benchmarks report sharp drops in high-risk scenarios [10]. A complementary approach makes the model a reasoning engine that selects analytic functions while a sandbox executes the arithmetic, avoiding numerical error and keeping raw data local; it builds on program-aided reasoning [7] and on interleaved reasoning and acting [8]. Our code, curated-tool and agentic levels operationalize distinct points along that axis.

Performance does not equate to clinical safety. The literature on evaluating medical LLMs motivates human evaluation along factuality, reasoning and potential harm rather than correctness alone [13], reports sensitivity to the order and amount of information [14], documents limited uncertainty calibration and abstention [15], and argues for workflow-level rather than model-level evaluation [16]. We therefore frame this work as an early-stage evaluation of a clinical decision-support system, following DECIDE-AI [19], TRIPOD+AI [20] and the IMDRF guidance [21]. No public dataset is used: the study rests on anonymized LibreView reports from real patients of a single service.

III. Benchmark Design

A. Resource Levels and Configurations

The unit of comparison is the *harness configuration*: an agent paired with a fixed bundle of inputs and capabilities. Fig. 1 lays out the eleven, with what each reads on the left and what it may execute on the right. At the agent-only level there is a representation and no actions: the raw series as text (C0), a line or AGP chart as PNG (C1), the report as page images (P0) or as text and tables (P1). A sandbox adds free-form code over the dataframe, in one shot (C2) or in a feedback loop (C3), and a PDF parser (P2); C4 replaces free-form code with a curated CGM library. The top level combines representation and actions over the raw signal (C5) or over the report (P3, and P-Ext+Tools).

P-Ext+Tools is named for what it receives: the digitized report, the series already extracted from it, and a deterministic library step for episode and trend questions. Its results characterize that bundle and not optical character recognition. Fig. 1, generated before the configuration was renamed, still labels the row by its OCR-extracted input.

The raw-data (prefix C) and report (prefix P) families share the oracle but differ in what the agent can access. A *model family* is one of the three evaluated models, and a *system instance* is a configuration paired with a model, the unit that the leaderboard and the mixed model score; of the 33 possible instances, 22 were executed.

B. Task Taxonomy

The benchmark defines 30 objective questions in four categories, of which 28 are evaluated; Q25 and Q26 are excluded by design as hardware artifacts. Each is scored along three difficulty axes, computational load, domain knowledge and visual reading, and appears with its category and grading tolerance in Table I. Severe (Q11) and nocturnal (Q21) hypoglycemia carry direct clinical risk and weigh in the safety analysis. A REPORT_SUPPORTED set marks the nine questions a standard AGP report can answer; the rest tend to fail in the P family, and that gap is part of the measurement.

C. Ground-Truth Oracle and Representations

Write $y^\star = O(s, q)$ for the oracle answer to question q on the complete signal s, and $\hat{y}_\pi = A_\pi(o_\pi, q; H_\pi)$ for the answer of configuration π, which sees only o_π and may invoke only H_π. The isolation stated in Section I is enforced by construction: no configuration can query O, and no oracle

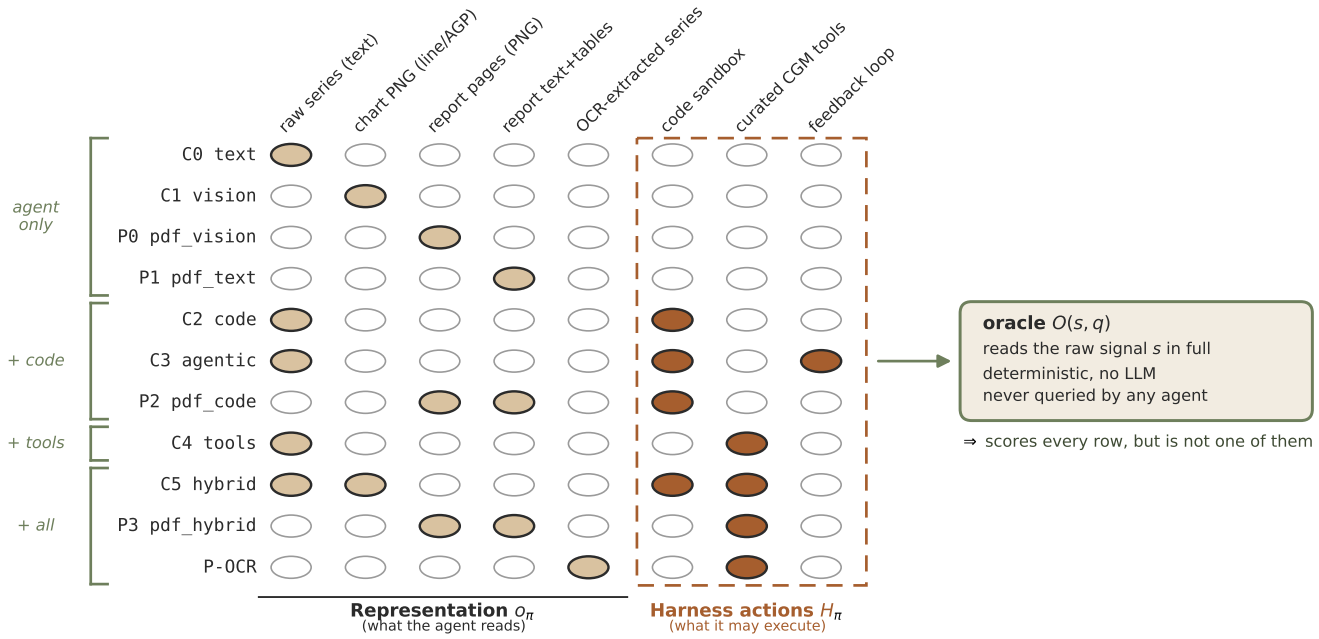


Figure 1. Harness-resource stratification. Each row is a configuration; filled circles mark what it receives (left block, the representation o_π) and what it may execute (dashed block, the action set H_π). This separates two constraints a scalar notion of "resource" would conflate: C0 holds the full series but has $H_\pi = \emptyset$, whereas C1 has an equally empty action set and a lossy representation as well. The row labelled by its OCR-extracted input is P-Ext+Tools, which receives the digitized report, the extracted series and the deterministic library at once and is not a visual route. The oracle is not a configuration: it reads s in full and is never exposed to any agent.

Table I
THE 28 OBJECTIVE BENCHMARK QUESTIONS (Q1–Q24, Q27–Q30); Q25 AND Q26 ARE EXCLUDED BY DESIGN. CAT.: MET = METRIC, TMP = TEMPORAL, ANO = ANOMALY, PAT = PATTERN. TOL. IS THE GRADER TOLERANCE AND FIXES THE ANSWER TYPE (ROUND. = NUMERIC; PP = PERCENTAGE POINT; SAME DAY = DATE; INTEGER = COUNT; YES/NO = BOOLEAN; LABEL = CATEGORICAL; ±10 MG/DL = WINDOWED MEAN). Q11 AND Q21 CARRY DIRECT CLINICAL RISK.

Q	Cat.	Tol.	Statement	Q	Cat.	Tol.	Statement
Q1	met	round.	Mean glucose (mg/dL)	Q15	tmp	same day	Onset of most recent hypo. episode (<70)
Q2	met	round.	Maximum glucose (mg/dL)	Q16	tmp	round.	Duration (min) of last hypo. episode
Q3	met	round.	Standard deviation (mg/dL)	Q17	tmp	round.	Duration (min) of longest hyper. episode
Q4	met	round.	Minimum glucose (mg/dL)	Q18	tmp	integer	Number of hypo. episodes (<70)
Q5	met	±0.5 pp	Time in range, 70–180 mg/dL	Q19	tmp	±10 mg/dL	Mean glucose, 00:00–06:00 (mg/dL)
Q6	met	±0.5 pp	Time above range, >180 mg/dL	Q20	tmp	label	Meal with highest glucose
Q7	met	±0.5 pp	Time below range, <70 mg/dL	Q21	tmp	yes/no	Any hypo. (<70) in 00:00–06:00?
Q8	met	±0.5 pp	Coefficient of variation (%)	Q22	tmp	±10 mg/dL	Highest glucose, 17:00–22:00 (mg/dL)
Q9	met	±0.5 pp	Time >250 mg/dL	Q23	ano	±0.5 pp	Time with active sensor (%)
Q10	met	round.	Estimated HbA1c (eA1c, %)	Q24	ano	integer	Number of sensor disconnections
Q11	met	±0.5 pp	Time <54 mg/dL (severe)	Q27	pat	yes/no	Today's mean better than yesterday's?
Q12	tmp	same day	Day of highest glucose	Q28	pat	yes/no	This week's TIR better than last week's?
Q13	tmp	same day	Day with most time out of range	Q29	pat	yes/no	Today's maximum lower than yesterday's?
Q14	tmp	same day	Day of lowest glucose	Q30	pat	yes/no	Less time <70 this week than last?

output enters any prompt. Harness-resource stratification is the manipulation of (o_π, H_π) with s and q fixed.

The oracle is distinct from the LibreView source report: it defines correctness for the leaderboard over the raw signal, while the report defines fidelity for the study of Section VII over printed values. We reserve *accuracy* for agreement with the oracle and *fidelity* for agreement with the source.

For each patient the oracle normalizes the exports to a canonical schema, fixes an analysis anchor and computes each answer from the raw signal, with episode boundaries, gap handling and thresholds from the consensus [3]. From the same dataframe the harness renders what each configuration consumes: a text bundle, line, AGP and heatmap charts, and a clinical PDF report. Meal windows and the definition of a week are frozen and treated as a sensitivity analysis.

IV. Evaluation Metrics

Four metric families are reported. *Objective accuracy* is rounding-tolerant agreement with the oracle, per question and aggregated. *Coverage and abstention* separate what an instance answers from what it gets right: the parse rate is the fraction of cells returning a value the typed parser accepts, and non-answers, whether a rejected output or a timed-out call, are

counted separately and never as correct. *Clinical safety* counts dangerous false negatives (DFN), an explicit negative answer on a cell whose oracle value indicates a hazard, restricted to the nocturnal and severe hypoglycemia questions and to the sensor-artifact questions that mask them; this is the failure mode with the shortest route to an incorrect insulin dose. A DFN requires an emitted answer, so an instance that abstains cannot accumulate them, and Section VIII-A treats that asymmetry as a result. *Expert assessment* (Section VII) aggregates Likert means and top-rating proportions per dimension. Error magnitude and chart-task rubrics were collected and are omitted for space; latency and token cost were collected but not ranked, the runs not being budget-controlled.

V. Experimental Protocol

The protocol is shown in Fig. 2. From the normalized signal s, two branches run in parallel and meet only at the grader: the oracle computes ground truth, while the renderer produces the representations o_π. The instance-by-patient-by-question-by-repetition matrix is then executed, with the parsed value, latency, steps and any error recorded per cell. The generated AGP report follows a third path, to the endocrinologist panel, whose ratings are collected blind, and to the LLM judge.

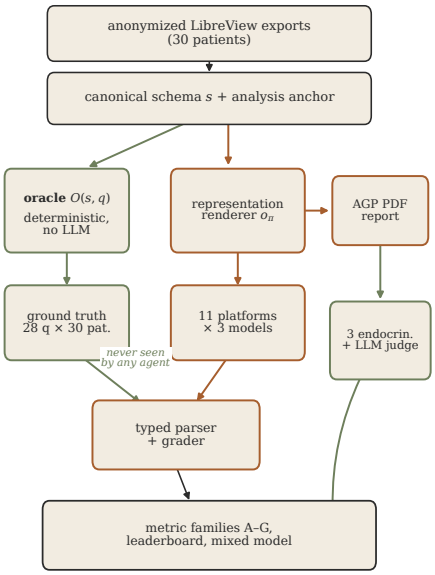


Figure 2. Experimental pipeline. The oracle (left, green) computes ground truth deterministically from s and is never exposed to the agents; it meets the configurations only at the grader. The instance matrix (right, clay) consumes only rendered representations o_π. The generated AGP report additionally feeds the endocrinologist panel and the LLM judge.

A. Model and Harness Configuration

Three model families were evaluated: Gemma 4 31B [24], a dense open generalist with native vision; Kimi K2.7 Code [25], a code-specialized checkpoint of the Kimi K2 family; and DeepSeek-V3.2 [26], a text-only mixture-of-experts model, all instruction-tuned and served unquantized through Ollama Cloud in the last two weeks of June 2026 with thinking mode enabled. Each cell of the instance-by-patient-by-question matrix was run in triplicate ($r = 3$), giving 2,520 calls per instance and 55,440 in total, and the repetitions collapse to one observation per cell by majority vote, the self-consistency rule declared before execution; dispersion across repetitions is retained only as a consistency measure.

The payload reaching the served model differs by level, which matters for data governance. At the code, tool, agentic and hybrid levels the prompt carries the canonical schema and the task, not the readings: the agent writes code, a local sandbox executes it, and only the returned values re-enter the conversation, so the series never leaves the local machine. At the prompt-only levels the representation itself is transmitted, as the text bundle (C0), a rendered chart (C1) or report pages and text (P0, P1). All exports were de-identified beforehand, and no direct identifier appeared in any transmitted payload.

Sampling parameters were not unified: each model kept its own serving defaults, so temperature and top-p differ between them, and unifying them would have meant overriding the recommended settings of checkpoints tuned under different regimes, which we judged worse than a nominal parity. Consequently, no between-family contrast is sampling-controlled and all are read descriptively; the contrasts carrying the central claim hold the model fixed and vary only the harness.

No token or step ceiling was imposed. Each model call was bounded by a 30 s wall-clock timeout, and the code and agentic configurations were otherwise stopped only on non-terminating loops, so per-call latency is right-censored at 30 s while end-to-end latency per cell is not, an agentic cell issuing several calls. Token usage in turn reflects an unstandardized number of calls, input lengths, output lengths and steps. Neither quantity supports a ranking or a cost per correct answer, and efficiency stays out of the leaderboard.

Without a common budget the iterative configurations also had access to a potentially different amount of compute, so the contrasts above the code level are not iso-compute. We cannot establish the direction or the magnitude of that confounding without a budget-controlled analysis. A ceiling can truncate productive and unproductive trajectories alike, and a token ceiling binds single long generations as well as multi-step ones. The single-shot configurations are not equivalent in cost either, since C0 carries the entire series in the prompt where C2 carries only the schema. A budget-controlled replication is what would make the cost axis interpretable, and it may change the magnitude or the ordering of the contrasts above code.

Because DeepSeek-V3.2 accepts no image input, the vision-dependent routes (C1, C5, P0) are infeasible under it by construction rather than by failure: the modality the harness offers, and not the model's competence, decides whether those cells can be run at all.

VI. Statistical Analysis

Because each instance is evaluated on the same patients and questions, the observations are paired and non-independent,

with correlation within patients and within questions. Accuracy is analyzed as a paired, hierarchical binary outcome with a logistic mixed model, *correct* $\sim$ instance $\times$ category with random intercepts for patient and question, which keeps easy questions or particular patients from dominating the estimate [22]. Fitted over the 22 executed instances and the 18,480 majority-vote outcomes, it yields the contrasts in Table IV, prespecified and corrected for multiplicity by false-discovery-rate control [23], with bootstrap intervals reflecting the clustered structure. Where agreement is high but one margin is nearly constant, kappa collapses under extreme prevalence, a paradox that motivates reporting delta, agreement range and correlation instead [27].

Five hypotheses were fixed before execution, each mapped to one contrast, with rounding-tolerant accuracy as the primary outcome except in H4, which is also a safety hypothesis. H1: a code sandbox added to a prompt-only configuration increases accuracy (C2 vs C0). H2: combining representations and actions increases accuracy beyond code alone (C5 vs C2). H3: a chart-only representation, being lossy, yields lower accuracy than the raw series in text (C1 vs C0). H4: a curated library in place of free-form code does not increase accuracy and may reduce safety when its definitions diverge from the oracle (C4 vs C2). H5: an agentic feedback loop recovers single-shot failures and exceeds one-shot code (C3 vs C2). Every contrast holds the model family fixed.

VII. Endocrinologist Assessment of the Automatic AGP Report Generation

Before running the full matrix we assessed one representation, the clinical PDF report the P family consumes. The pipeline extracts numerical data from a source LibreView report, generates a structured interpretation, and carries an automatic fidelity audit against the source. It is also the calibration set for the LLM judge.

A. Instrument, Participants and Data

The instrument has three parts. Part A is an objective 16-item fidelity checklist, each item asking whether a quantity in the generated report matches the LibreView source (yes, no, or not applicable), the human analogue of the automatic audit. Parts B to D are 5-point Likert items (1 strongly disagree to 5 strongly agree) covering technical quality, interpretive quality and clinical usefulness, and an overall assessment. Three endocrinologists assessed 30 generated reports, one reviewer per report and no report read twice.

That design bounds what the assessment can establish. It is expert opinion under a fixed protocol, not a validation study: with a single reader per report there are no replicate ratings, no inter-rater coefficient can be estimated, and an isolated divergence cannot be assigned either to a stricter but faithful reviewer or to a genuine defect. The sample is small (30 patients, one service, one documentary source) and retrospective. The reader confounding is mitigated, not removed, by checking whether divergences concentrate in one reviewer and by the automatic audit below.

B. Objective Fidelity

We call *core glycemic metrics* the nine consensus-defined quantities the extraction schema exposes as scalars, and that the automatic audit covers: mean glucose, GMI, CV, TIR, TBR $<$70, TBR $<$54, TAR $>$180, TAR $>$250 and sensor capture [3], [4]. The rest are provenance or qualitative items with no scalar counterpart, and the two groups fail for different reasons and at different rates.

Table II reports the per-item agreement rate. Aggregated over the nine core metrics, fidelity is 97.41% (263 of 270 item–report pairs), from 93.33% (TBR $<$54) to 100.00%, and the failures concentrate outside the core. Identification of automatically confirmed data is the largest single gap at 80.00%: the report communicates poorly which values were confirmed automatically and which were flagged for manual review. Ceiling saturation at 350 mg/dL or above is flagged inconsistently (89.29% of the 28 applicable reports), and time below 54 mg/dL is the only core metric under 96%. Divergences are spread across reports: 14 of the 30 drew at least one flagged item and 4 drew two or more.

Table II
Objective fidelity against the LibreView source (n = 30 patients). Rates are $k/30$ unless noted; †25/28, excluding two not-applicable reports. ‡Sensor capture, TIR 70–180, CV, TBR $<$70, TAR $>$250. Items at 100.00% omitted (period, mean glucose, GMI, TAR $>$180, estimated/missing data, technical alerts). Core-metric aggregate: 97.41% (263/270).

Checklist item	% agree	failures
Identification of confirmed data	80.00	6 no
Ceiling saturation $\geq$350 mg/dL	89.29†	3 no, 2 n/a
Time below 54 mg/dL (TBR)	93.33	2 no
Declared number of days	96.67	1 no
Manual-review flagging	96.67	1 no
Five further core metrics‡	96.67	1 missing each

C. Clinical Usefulness Ratings

All Likert responses fell in the 4 to 5 range, with a global mean of 4.90, 90.4% at the top of the scale, per-section means between 4.88 and 4.97 and three items at a perfect 5 from every reviewer. The lowest-rated are the clarity of patient-facing language (4.57, with 13 of 30 evaluations giving only partial agreement) and the usefulness of the suggested-management section (4.70); the adoption item, "I would use this report as clinical decision support, provided it is reviewed by a qualified professional," scored 4.93. The ceiling effect limits discrimination between reports and prevents these means from being interpreted as a general estimate of pipeline performance.

D. Human versus Automatic Audit

An automatic audit compared the report scalars with the source PDF on the nine core metrics and returned *consistent* in all 30 cases, agreeing across all valid pairs (28–30 per metric) for the eight numerical metrics with dual coverage. The five isolated human "no" answers have no counterpart in the audit,

which is more consistent with reviewer interpretation than with a pipeline defect; on TBR <54, two reviewers marked "no" on reports the audit confirmed as faithful, plausibly judging how prominently severe hypoglycemia was highlighted rather than whether the number was present. The audit is suggestive and not decisive here, since it reads the same extracted representation the pipeline produces and checks values, not presentation. Every defect it can see lies in the items *without* automatic coverage, across at least 9 reports; on the qualitative items the ambiguity remains, and only multiple readers per patient would resolve it.

E. Calibration of the LLM Judge against the Panel

The same 30 reports and six dimensions went to an LLM judge, an auxiliary triage instrument and not a substitute for the panel, drawn from a model family distinct from those evaluated to mitigate the self-preference, position and verbosity biases documented for LLM judges [9]. All 30 were judged without error (global mean 4.80; 26 pass, 4 pass with caveats, no fail), and its ordering of dimensions reproduces the panel's, with actionability (4.57) lowest and clinical safety (5.00) highest, so both locate the weakness in the management text, not in the arithmetic. Its mean is within 0.20 point of the endocrinologist's (SD 0.29), but the correlation between the two sets of means is only moderate, $r = 0.49$ ($p = 0.006$), so close average agreement is not close agreement report by report. Where they diverge the judge is the stricter reader, penalizing internal numerical inconsistencies ("10 of 14 days" in the text against 9 in the table) that the reviewer let pass.

Weighted kappa is again uninformative, all endocrinologist means rounding to 5. And although none of the 77 panel-annotated hazards was omitted, no synthetic hazards were seeded and no non-hazard controls annotated, so specificity is unmeasured and a detector flagging indiscriminately would reach the same recall of 1.00. A seeded-error study is the next step.

VIII. Results: Leaderboard and Model Contrasts

The matrix spans 30 patients and 28 questions, giving 840 pairs per instance; Table III orders the instances by global accuracy with bootstrap CIs. Kimi and DeepSeek replicate the code, agentic, tool and hybrid routes whenever the input is textual; absent instances were not run.

With the model fixed, accuracy rises 23.10 pp from C0 (text) to C2 (code), the largest transition on the scale, supporting H1. Above C2 the increments are small, the hybrid C5 landing 5.6 points above code, and chart reading (C1) falls below text-only, consistent with H3. The agentic loop (C3) ties with or falls slightly below one-shot code in all three families, so H5 is not supported, though how the contrast behaves under a common budget remains untested. Between families the spread is small and there is no single winner, the code-specialized model surpassing the generalist in pattern recognition and losing in anomaly detection. These family contrasts are descriptive, the models having been served under model-specific sampling defaults.

C4, which replaces free-form code with the curated library, sits below C2 in accuracy and produced the largest number of dangerous false negatives among the raw-data configurations, 24 to 25 against 6 to 11 for code and hybrid. The mechanism is specific to this implementation: its cleaning routine removes the signal gaps before computing metrics, discarding the sensor-absence information the anomaly questions (Q23, Q24) require, and C4's anomaly accuracy falls to 7% against 97% for C2. A tool layer therefore does not by itself reduce risk, and here the effect follows from a single preprocessing choice.

P-Ext+Tools achieved the highest observed accuracy, with 2 dangerous false negatives, the lowest count among routes answering all 28 questions. This result does not isolate visual report interpretation, since the configuration also receives the extracted series and the deterministic library step (Section III). The aligned-versus-template pair quantifies the H4 risk: with the library parameters misaligned from the oracle (`min_gap` 5 vs 60 min), accuracy drops 17.14 pp and disconnection detection collapses, 320 reported against a true 1. For this configuration, performance depended strongly on alignment between the library definitions and the oracle, and the mismatch produced no visible signal in the output; configurations writing their own code were less sensitive to it.

Global accuracy of the report-based configurations (P0–P3, 23.6 to 31.1%) is bounded by coverage: only 9 of the 28 questions are answerable from a standard AGP report, so their parse rate is correspondingly low (28 to 32%); on covered questions with an emitted answer, accuracy reaches 96% (P1, P3).

Table III
LEADERBOARD (N = 30 PATIENTS, 28 QUESTIONS, 840 PAIRS PER INSTANCE): GLOBAL ACCURACY, 95% BOOTSTRAP CI, PARSE RATE AND DANGEROUS FALSE NEGATIVES (FN). FN COUNTS ARE CONDITIONAL ON AN EMITTED ANSWER AND ARE NOT COMPARABLE ACROSS ROWS WITH DIFFERENT PARSE RATES. G = GEMMA 4 31B, K = KIMI K2.7 CODE, D = DEEPSEEK-V3.2; BETWEEN-MODEL DIFFERENCES ARE DESCRIPTIVE, SAMPLING PARAMETERS NOT HAVING BEEN UNIFIED. P-EXT+TOOLS ALSO RECEIVES THE EXTRACTED SERIES AND A DETERMINISTIC LIBRARY. ORACLE-ALIGNED ROWS (100%) OMITTED.

Platform	Mod.	Acc.	95% CI	parse	FN
P-Ext+Tools	K	77.9	75–81	96	2
P-Ext+Tools	D	77.9	75–81	98	2
C5 hybrid	G	77.0	74–80	100	6
P-Ext+Tools	G	72.6	70–76	89	2
C2 code	K	71.8	69–75	100	11
C3 agentic	D	71.7	69–75	100	11
C3 agentic	K	71.5	68–75	100	11
C2 code	G	71.4	68–74	88	11
C2 code	D	70.7	68–74	99	11
C5 hybrid	K	70.1	67–73	100	11
C3 agentic	G	69.2	66–72	69	11
C4 tools	D	60.8	58–64	98	24
P-Ext+Tools template	K	60.7	57–64	93	17
P-Ext+Tools template	D	60.7	57–64	95	17
C4 tools	G	58.6	55–62	59	25
P-Ext+Tools template	G	55.5	52–59	87	17
C0 text	G	48.3	45–52	100	10
C1 vision	G	38.9	36–42	76	14
P2 pdf code	G	31.1	28–34	32	0
P3 pdf hybrid	G	31.0	28–34	32	0
P1 pdf text	G	30.5	27–34	32	0
P0 pdf vision	G	23.6	21–26	28	0

A. Dangerous False Negatives and Abstention

A dangerous false negative requires an emitted answer, and the instances abstain at very different rates, so the safety column of Table III is not interpretable without its parse rate. The report family (P0–P3) records zero such failures while answering about a third of the matrix, mostly the metric questions an AGP report prints. Under a convention that treated a non-answer on a risk-bearing question as a missed hazard, their position in the ordering would change, but by how much cannot be derived from the global parse rate: an instance may abstain heavily on the questions an AGP report does not cover and still answer most risk questions. Deciding this needs coverage measured on the risk cells themselves (Q11, Q21 and the sensor-artifact questions), which we did not compute for this version and which the benchmark will report next.

Coverage bounds the raw-data family too. C4 and C3 under Gemma parse 59% and 69% of cells, so their counts of 25 and 11 may understate the denials that full coverage would have produced. The comparison between C4 and C2 is less exposed, since C4 under DeepSeek parses 98% and still records 24 against Gemma's 25 at 59% coverage, which points to the preprocessing and not to the model. Among the instances that answer essentially every cell the counts cluster between 6 and 11.

B. Mixed-Model Contrasts

The mixed model of Section VI resolves the five hypotheses (Table IV). H1 is supported and is the largest effect on the scale, and H3 is supported, with isolated visual reading below text. Above code the results diverge: H5 is not supported, the agentic loop tying with one-shot code, while H4 is supported on accuracy, the curated library falling significantly below code in both families where it ran. H2 holds only partly, since the hybrid exceeds code under Gemma but not under Kimi, which a naive paired test would have obscured. Per axis, higher computational load and visual-reading difficulty reduce correctness, while the positive domain-knowledge sign is confounded with report-covered metric questions and is not read as causal. Question-level variance (ICC 0.125) exceeds patient-level variance (ICC 0.033).

Table IV
CONTRASTS AND EFFECTS OF THE LOGISTIC MIXED MODEL (22 SYSTEM INSTANCES, 18,480 MAJORITY-VOTE OUTCOMES). OR >1 FAVORS THE LEFT-HAND CONFIGURATION, OR THE INCREASE ALONG THE AXIS; p WITH BENJAMINI–HOCHBERG CORRECTION. ALL CONTRASTS HOLD THE MODEL FAMILY FIXED. G = GEMMA 4 31B, DS = DEEPSEEK-V3.2.

Contrast / effect	OR	95% CI	p_{BH}
C2 code vs C0 text (H1)	3.91	3.06–4.98	4×10^{-27}
C1 vision vs C0 text (H3)	0.50	0.40–0.62	6×10^{-9}
C3 agentic vs C2 code (H5, G)	0.81	0.58–1.13	0.33
C4 tools vs C2 code (H4, G)	0.54	0.37–0.78	0.004
C4 tools vs C2 code (H4, DS)	0.54	0.36–0.80	0.006
C5 hybrid vs C2 code (H2, G)	1.58	1.09–2.27	0.037
C5 hybrid vs C2 code (H2, K)	0.85	0.55–1.31	0.49
Computational load (per level)	0.79	0.78–0.81	$< 10^{-78}$
Domain knowledge (per level)	1.26	1.22–1.29	$< 10^{-70}$
Visual reading (per level)	0.72	0.70–0.73	$< 10^{-188}$

IX. DISCUSSION

The largest observed accuracy gain occurred in the C0-to-C2 transition, where executable code was introduced. Above that level, an iterative loop, a chart alongside the series and a library of ready-made functions each changed accuracy by a few points at most, and in one case reduced it, in every model evaluated.

A cleaning routine inside the C4 tool library deleted the sensor gaps before any metric was computed, which is defensible for a mean and destructive for counting disconnections: anomaly accuracy fell to 7% and dangerous denials more than doubled, so C4 recorded the lowest safety among the raw-data routes despite having the largest amount of surrounding machinery. The finding is specific to this implementation. It illustrates a general mechanism nonetheless: a tool library encodes definitions, and a divergence between those definitions and the reference produces no visible signal in the output.

The report pipeline was faithful on the core metrics and near the ceiling on perceived usefulness. The defects worth prioritizing are all addressable in the template: two reports whose narrative underplayed severe hypoglycemia although the numbers were right, inconsistent ceiling-saturation flagging, and identifying which values were automatically confirmed, a provenance issue and not a glycemic one.

Agentic orchestration is not demonstrated to be superior, since C3 ties with C2 and the hybrid leads narrowly under one family only. The expert instrument saturates (mean 4.90, three zero-variance items), so it validates this report but cannot discriminate between configurations; with one reviewer per report, strictness and report defect stay partly confounded on the qualitative items, and 30 retrospective reports from a single service are a narrow base for generalization. The safety column depends on an accounting convention and on coverage measured at the risk cells, which we do not yet report (Section VIII-A); sampling was not unified, so every between-family comparison is descriptive; budgets were open, so the cost of each level is unquantified; and 11 of the 33 instances were never run. Benchmark scores that ignore data, task and metric choices readily create an illusion of clinical readiness [11], [13], [14].

A clinically relevant next step is to evaluate interpretation support under human review, including whether clinician–model collaboration improves performance and when the system should abstain [17], [18]. Such a study would require external reports, multiple independent evaluators per case with adjudication, and abstention and calibration metrics [15], [16], reported under contemporary guidelines [19], [20]. Because a glucose error can propagate to an incorrect insulin dose, safety is the priority axis, and deployment would bring software-as-a-medical-device framing into scope [21].

X. CONCLUSION

We presented a harness-resource-stratified benchmark for CGM interpretation, an expert assessment of an automatic AGP report pipeline over 30 patients, and a leaderboard

over 22 system instances. Under the models and conditions evaluated here, varying the harness produced larger accuracy differences than those observed between model families, the text-to-code transition being the largest. Because the between-family contrasts were not sampling-controlled and only three model families were evaluated, this difference should be interpreted as specific to the present experimental conditions; a budget-controlled replication may also change the magnitude or the ordering of the contrasts above the code level.

A curated tool layer can reduce safety when its preprocessing discards information a task depends on, and the failure is silent. A false-negative count of zero can also reflect abstention rather than hazard recognition, so safety has to be reported together with coverage, and with coverage measured on the risk questions themselves. Neither result licenses clinical deployment, which would require external data, multiple adjudicated evaluators, a collaboration arm and prospective evaluation; those, with a seeded-error study for judge specificity and a budget-controlled replication, are our next steps.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001. Danielo G. Gomes acknowledges the support from the National Council for Scientific and Technological Development (CNPq), Brazil (grant numbers 309014/2026-3, 311845/2022-3).

References

[1] E. Healey and I. Kohane, "LLM-CGM: A benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management," *Pac. Symp. Biocomput.*, vol. 30, pp. 82–93, 2025.

[2] E. Healey *et al.*, "Leveraging large language models to analyze continuous glucose monitoring data: A case study," *Sci. Rep.*, vol. 15, no. 1, art. 1143, 2025.

[3] T. Battelino *et al.*, "Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the International Consensus on Time in Range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.

[4] R. M. Bergenstal *et al.*, "Glucose management indicator (GMI): A new term for estimating A1C from continuous glucose monitoring," *Diabetes Care*, vol. 41, no. 11, pp. 2275–2280, 2018.

[5] Am. Diabetes Assoc. Prof. Pract. Committee, "7. Diabetes technology: Standards of Care in Diabetes—2026," *Diabetes Care*, vol. 49, no. Suppl. 1, pp. S150–S165, 2026.

[6] M. L. Johnson *et al.*, "Utilizing the ambulatory glucose profile to standardize and implement continuous glucose monitoring in clinical practice," *Diabetes Technol. Ther.*, vol. 21, no. S2, pp. S2-17–S2-25, 2019.

[7] L. Gao *et al.*, "PAL: Program-aided language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 202, pp. 10764–10799, 2023.

[8] S. Yao *et al.*, "ReAct: Synergizing reasoning and acting in language models," in *Int. Conf. Learn. Represent. (ICLR)*, 2023.

[9] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.

[10] S. Wang *et al.*, "A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains," *npj Digit. Med.*, vol. 9, art. 91, 2026.

[11] M. Agrawal *et al.*, "The evaluation illusion of large language models in medicine," *npj Digit. Med.*, vol. 8, art. 600, 2025.

[12] R. W. Beck *et al.*, "Validation of time in range as an outcome measure for diabetes clinical trials," *Diabetes Care*, vol. 42, no. 3, pp. 400–405, 2019.

[13] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[14] P. Hager *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nat. Med.*, vol. 30, pp. 2613–2622, 2024.

[15] M. Griot *et al.*, "Large language models lack essential metacognition for reliable medical reasoning," *Nat. Commun.*, vol. 16, art. 642, 2025.

[16] F. Gaber *et al.*, "Evaluating large language model workflows in clinical decision support for triage, referral and diagnosis," *npj Digit. Med.*, vol. 8, art. 263, 2025.

[17] M. Chen *et al.*, "Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage," *npj Digit. Med.*, vol. 9, art. 222, 2026.

[18] K. Kawamoto *et al.*, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *BMJ*, vol. 330, art. 765, 2005.

[19] B. Vasey *et al.*, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *BMJ*, vol. 377, art. e070904, 2022.

[20] G. S. Collins *et al.*, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, art. e078378, 2024.

[21] Int. Med. Device Regulators Forum, "Software as a medical device (SaMD): Clinical evaluation," IMDRF/SaMD WG/N41FINAL:2017, 2017.

[22] B. M. Bolker *et al.*, "Generalized linear mixed models: A practical guide for ecology and evolution," *Trends Ecol. Evol.*, vol. 24, no. 3, pp. 127–135, 2009.

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[24] Gemma Team, Google DeepMind, "Gemma 4 technical report," arXiv:2607.02770, 2026.

[25] Kimi Team, "Kimi K2: Open agentic intelligence," arXiv:2507.20534, 2025.

[26] DeepSeek-AI, "DeepSeek-V3.2: Pushing the frontier of open large language models," arXiv:2512.02556, 2025.

[27] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: I. The problems of two paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543–549, 1990.

[28] E. T. Barr *et al.*, "The oracle problem in software testing: A survey," *IEEE Trans. Softw. Eng.*, vol. 41, no. 5, pp. 507–525, 2015.

Communication Technologies and Application Protocols in Wearable IoT for Health Monitoring: A Layered Analysis with Preliminary Connectivity Observations

Rodrigo Moreira dos Santos
PPGEEL, Santa Catarina State University (UDESC)
Joinville, Brazil
rodrigo.mds@edu.udesc.br

Pedro Bertemes Filho
PPGEEL (UDESC)
Joinville, Brazil
pedro.bertemes@udesc.br

Maicol Peterson Gandolphi de Almeida
PPGEEL (UDESC)
Joinville, Brazil
maicol.almeida@edu.udesc.br

***Abstract*—Wearable Internet of Things systems for health monitoring combine sensors, local wireless links, gateways, application protocols, and remote services. This paper presents an exploratory layered analysis that distinguishes connectivity technologies from application-layer protocols. Bluetooth Low Energy (BLE) and Wi-Fi are examined at the wireless connectivity layer, whereas MQTT, HTTP/REST, and WebSocket are discussed according to their roles after IP connectivity is available. Zigbee, Thread, LoRaWAN, NB-IoT, and LTE-M are included as architectural alternatives outside the implementation scope. The reference architecture distinguishes the functions of the wearable, gateway, network access, remote service, storage, and monitoring interface. Preliminary observations used a Redmi Watch 5 Active as a closed commercial reference and an ESP32-PICO-D4-based M5StickC Plus as an open prototyping platform. The implementation was limited to Wi-Fi association, simulated physiological-message generation, payload structuring, and functional connectivity verification. Controlled measurements of energy per message, latency distribution, application-message loss, goodput, reconnection time, memory use, scalability, interference, offline buffering, and security overhead were not performed. Therefore, the study does not establish quantitative superiority or clinical suitability. Its contribution is a requirements-based framework that separates published evidence from bench observations and defines a reproducible protocol for subsequent controlled experiments.**

***Index Terms*—wearable IoT, health monitoring, Bluetooth Low Energy, Wi-Fi, MQTT, HTTP, WebSocket, ESP32, communication architecture**

I. Introduction

The Internet of Things (IoT) integrates sensors, embedded devices, communication networks, and digital services [1], [2]. In healthcare, this integration supports remote monitoring, home follow-up, physiological telemetry, and continuity of care outside hospitals [3]–[5]. Wearable systems are particularly relevant because they can acquire heart rate, oxygen saturation, temperature, movement, and activity data periodically or continuously [6], [7].

Communication is not a single interchangeable component in these systems. Wireless connectivity technologies determine how a wearable reaches a nearby gateway or access network. Application protocols determine how data are exchanged after a compatible network path exists. Treating BLE, Wi-Fi, MQTT, HTTP, and WebSocket as direct alternatives obscures this layering and can lead to unsupported rankings. BLE and Wi-Fi occupy the connectivity portion of the architecture, whereas MQTT, HTTP/REST, and WebSocket operate at the application layer over IP-based transport. Gateways may bridge non-IP links, perform local processing, buffer data, and forward selected records to remote services [4].

BLE is widely used for short-range, battery-constrained wearable links and smartphone integration [8]–[10]. Wi-Fi provides direct access to local IP networks and commonly available Internet infrastructure, but its energy cost depends on association, radio state, traffic pattern, and power-management configuration [11], [12]. Other technologies may be appropriate for different constraints. Zigbee and Thread use IEEE 802.15.4-based low-power networking and generally require a coordinator or border router. LoRaWAN targets low-power wide-area communication through gateways and network servers. NB-IoT and LTE-M use managed cellular infrastructure [13]–[16]. These alternatives are discussed in this paper but were not implemented.

At the application layer, MQTT uses a brokered publish/subscribe model and supports different quality-of-service levels [17]. HTTP defines request-response semantics commonly used by REST APIs [18], [19]. WebSocket establishes persistent bidirectional communication after an HTTP handshake [20]. Published experiments show that protocol behavior depends on payload, loss conditions, security, and implementation, so no application protocol can be declared universally superior from architectural descriptions alone [21], [22].

Security and privacy are also system properties, not labels inherent to a protocol name. BLE pairing mode, credential management, authorization, TLS configuration, storage protection, and key lifecycle all influence the actual risk. Wearable and medical IoT systems therefore require explicit threat and performance evaluation [23]–[25]. A connectivity check cannot demonstrate security, autonomy, clinical reliability, or resilience

to attacks.

This study asks how communication technologies and application protocols for wearable health monitoring can be organized according to their architectural roles and measurable requirements without assuming equivalence or unsupported quantitative superiority. Its contributions are: 1) a layered reference architecture that separates wireless links, gateways, IP access, application protocols, and remote services; 2) an explicit distinction between published evidence, preliminary bench observations, unmeasured properties, and a proposed benchmark; 3) a requirements matrix without subjective suitability scores; and 4) a reproducible protocol for subsequent measurement. The paper does not claim clinical validation or comparative performance results.

II. Methodology

A. Study Design and Evidence Classification

This is an exploratory technical study combining a focused narrative review with preliminary functional observations. It is not a systematic review, meta-analysis, or controlled protocol benchmark. Sources were selected to cover wearable health architectures, BLE and Wi-Fi behavior, application protocol definitions and performance, security, and the official specifications of the devices and additional connectivity technologies.

Every statement in the analysis is assigned to one of four evidence classes: *literature-supported*, when based on a cited publication or standard; *bench-observed*, when directly observed in the preliminary implementation; *not measured*, when the study collected no comparable data; and *proposed benchmark*, when a parameter is specified only for a future controlled experiment. This classification prevents literature-based expectations from being reported as measurements made by the authors.

B. Reference Architecture

Fig. 1 separates six roles: wearable or sensor node; smartphone or dedicated gateway; network access; broker or application server; storage; and dashboard or research interface. A commercial wearable may expose measurements only through a proprietary smartphone application. An open device may connect directly to Wi-Fi and then reach an application service. MQTT, HTTP/REST, and WebSocket are positioned only after IP connectivity. A gateway is an intermediary computing and translation role, whereas the cloud is a deployment location for services and storage; the two terms are not used interchangeably.

C. Devices and Preliminary Procedure

The closed commercial reference was a Redmi Watch 5 Active. The manufacturer specifies Bluetooth 5.3 connectivity, a heart-rate sensor with a blood-oxygen sensor, and a 470-mAh battery [27]. The watch was used only to observe pairing and the dependence on its mobile application. No proprietary BLE packet was intercepted, modified, or treated as a physiological ground truth.

The open platform was an M5StickC Plus, product K016-P, built around an ESP32-PICO-D4 and supplied with a 120-mAh battery [28]. The SiP provides 2.4-GHz IEEE 802.11b/g/n Wi-Fi, Bluetooth 4.2/BLE, 520 KB of SRAM, and 4 MB of in-package flash [12]. The exact historical IDE and board-core revisions used during the initial exploratory work were not retained. For subsequent replication, a reference software baseline is defined using Arduino IDE 2.3.10 and Arduino-ESP32 core 3.3.11 [29], [30]. These versions define the proposed repeatable baseline and should not be interpreted as retrospectively verified details of the initial exploratory implementation.

Only Wi-Fi association and simulated-message handling were implemented on the M5StickC Plus. The observations verified that the platform could join a local 2.4-GHz network, generate a structured record resembling physiological data, and expose connection state for debugging. MQTT, HTTP/REST, and WebSocket were not implemented side by side under an identical workload. The exact original payload, sending interval, test duration, distance, message count, and number of repetitions were not logged. No original latency trace, packet-loss log, current trace, reconnection series, or memory profile is available. Therefore, the Results section reports no reconstructed numeric result.

D. Architectural Alternatives Outside the Bench Scope

Table I positions additional wireless technologies relevant to wearable IoT health-monitoring architectures. The entries describe roles and constraints supported by standards and official technical material; they are not performance results from the M5StickC Plus. Zigbee and Thread are relevant when mesh coverage and low-power IEEE 802.15.4 radios are available. Thread preserves IP networking through a border router, while Zigbee uses its own stack and coordinator model. LoRaWAN is relevant to sparse low-rate telemetry over a wide area, but requires LoRa-capable hardware, gateways, and a network server. NB-IoT and LTE-M use a cellular modem, subscriber provisioning, and operator coverage. None is native to the ESP32-PICO-D4 used here.

E. Requirements Matrix

To avoid unsupported ordinal rankings, candidate technologies are screened against explicit system requirements, as shown in Fig. 2 and Table II. Quantitative limits should be defined according to the intended use case before protocol selection. A research dashboard, a store-and-forward activity logger, and a safety-critical clinical alarm do not share the same latency, availability, or risk requirements.

F. Proposed Controlled Benchmark

Fig. 3 specifies a future benchmark and must not be read as an executed experiment. The reference UTF-8 JSON payload is:

```
{"device":"M5C01","seq":"000001",
"ts":"2026-08-25T12:00:00Z",
"hr":78,"spo2":98,"temp":36.6}
```

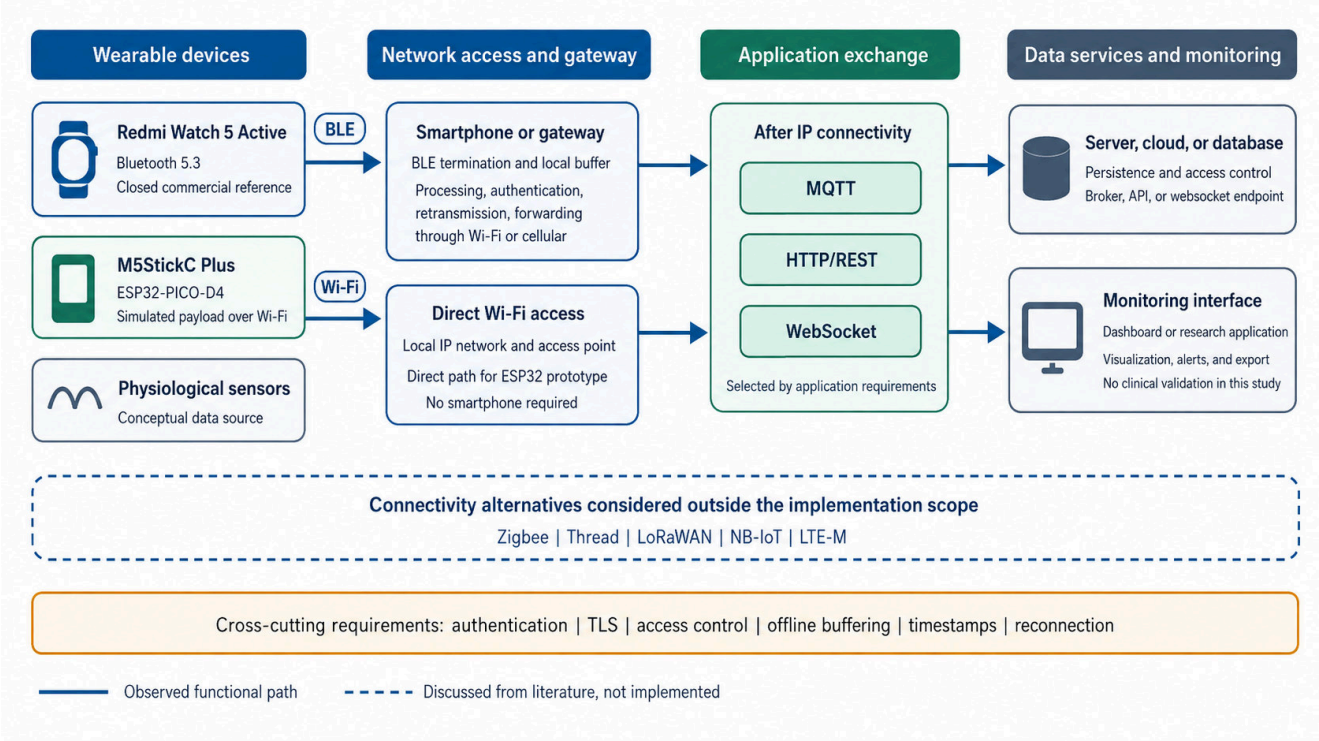


Fig. 1. Layered reference architecture for wearable IoT health monitoring. The Redmi Watch 5 Active represents a BLE-assisted gateway path, whereas the M5StickC Plus can access a local IP network directly through Wi-Fi. MQTT, HTTP/REST, and WebSocket are application protocols, not alternatives to BLE or Wi-Fi. The dashed band identifies wireless alternatives discussed from official specifications but not implemented.

TABLE I
WIRELESS CONNECTIVITY ALTERNATIVES AND THEIR ARCHITECTURAL ROLES

Technology	Network role	Typical intermediary	Selection questions	Evidence in this study
BLE	Short-range wearable link	Smartphone or BLE gateway	Required autonomy, connection interval, payload size, phone dependence, pairing policy	Literature; commercial pairing observation only
Wi-Fi	Direct local IP access	Access point or router	Association energy, coverage, congestion, credentials, local/cloud route	Literature; functional association and simulated messages
Zigbee	Low-power mesh over IEEE 802.15.4	Coordinator and gateway	Mesh density, coordinator availability, stack interoperability, interference	Standards discussion; not implemented
Thread	IPv6-based low-power mesh over IEEE 802.15.4	Thread border router	Border-router availability, IPv6 integration, mesh density, commissioning	Standards discussion; not implemented
LoRaWAN	Low-power wide-area uplink/downlink	Gateway, network server, application server	Regional band, duty cycle, payload frequency, downlink demand, coverage	Specification discussion; not implemented
NB-IoT / LTE-M	Managed cellular IoT access	Operator network and cellular modem	Coverage, subscription, mobility, latency budget, energy profile, certification	3GPP discussion; not implemented

The payload contains 91 bytes without a terminating newline and is generated once per second. Each condition comprises 600 messages over 10 minutes and three independent repetitions at 1, 5, and 10 m. The Wi-Fi access point uses a fixed 2.4-GHz channel, 20-MHz channel width, WPA2, and a wired local server to reduce uncontrolled backhaul variation. MQTT QoS 0 and QoS 1, HTTP POST, and a WebSocket text message use the same application content. Device and server timestamps require a documented synchronization method; otherwise only round-trip latency may be reported.

The benchmark shall report latency distributions, application-message loss, duplicates, useful goodput, energy per delivered message, reconnection time, and memory use. A programmable interruption shall test reconnection and offline buffering. Security tests shall compare unprotected local communication only as an instrumented baseline against TLS-protected MQTT, HTTPS, and secure WebSocket, using the same authentication policy. Reported overhead shall separate initial handshake cost from steady-state records. Co-channel traffic shall be recorded or controlled, and an interference condition shall be reported separately. Scalability requires additional concurrent publishers and cannot be inferred from a single device.

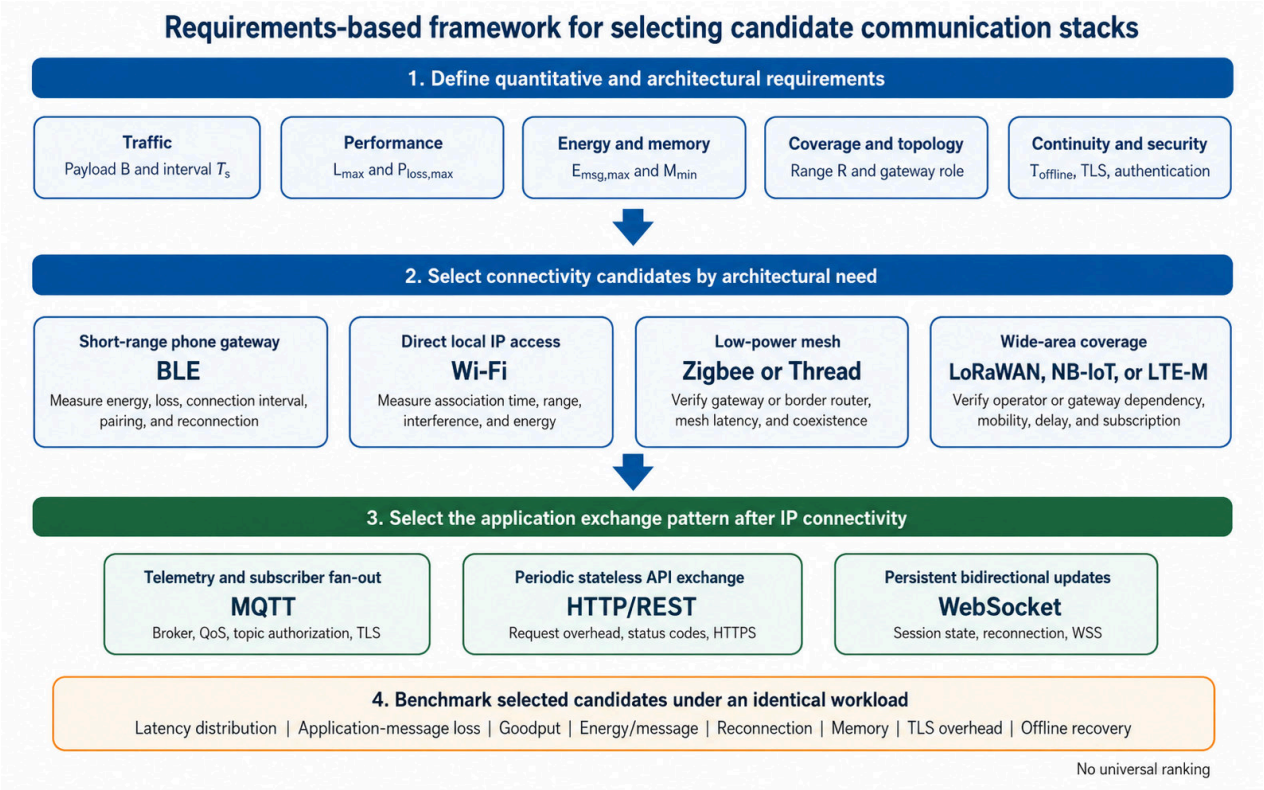


Fig. 2. Requirements-based framework for selecting candidate wireless technologies and application protocols. Candidate selection is conditioned on quantitative limits and followed by controlled benchmarking under an identical workload. The framework does not assign universal suitability scores.

TABLE II
QUANTITATIVE REQUIREMENTS AND REPORTING UNITS

Requirement	Unit	Minimum report
Payload size	byte	Application and on-wire sizes
Generation interval	s	Nominal interval and jitter
One-way or round-trip latency	ms	Definition, p50, p95, p99
Application-message loss	%	Sent, received, lost, and duplicate messages
Goodput	bit/s	Useful payload per elapsed time
Energy per delivered message	mJ/msg	Voltage, current trace, delivery count
Reconnection time	s	Trigger and distribution
Memory use	KB	Free/peak heap and stack method
Offline recovery	msg, s	Buffer capacity and replay time
Security overhead	ms, byte, mJ	Handshake and steady-state deltas

III. RESULTS

A. Bench-Observed Results

The Redmi Watch 5 Active paired with its smartphone application, demonstrating the user-facing behavior of a closed commercial architecture. Direct access to raw sensor packets, connection parameters, and application-protocol selection was not available within the study. Consequently, the watch served as an architectural reference, not as a benchmark instrument or source of clinically validated physiological data.

The M5StickC Plus joined a local Wi-Fi network and supported generation and structuring of simulated records. These observations confirmed functional connectivity and implementation flexibility. They do not quantify latency, application-message loss, goodput, energy consumption, reconnection time, memory use, interference, security overhead, or scalability. Because the original exploratory activity did not retain fixed workload parameters or repeated measurements, no mean, standard deviation, confidence interval, or significance test can be validly calculated.

B. Layered Decision Synthesis

At the wearable-to-gateway link, BLE is a candidate when a nearby smartphone or gateway is acceptable and the device sends small, intermittent records. Wi-Fi is a candidate when direct local IP access is required and the energy budget can accommodate association and active-radio periods. Zigbee and Thread become candidates when an IEEE 802.15.4 mesh and its infrastructure are part of the intended deployment. LoRaWAN, NB-IoT, and LTE-M address wider-area paths with substantially different hardware and infrastructure assumptions. These statements define screening conditions, not a ranking.

After IP connectivity is established, MQTT is a candidate for brokered telemetry and multiple consumers; its QoS level, session state, retained messages, and broker availability must be specified. HTTP/REST is a candidate for stateless resource-oriented integration and periodic uploads; connection reuse and batching affect overhead. WebSocket is a candidate for a persistent bidirectional session, including continuously updated interfaces; keep-alive, reconnection, and server state must be

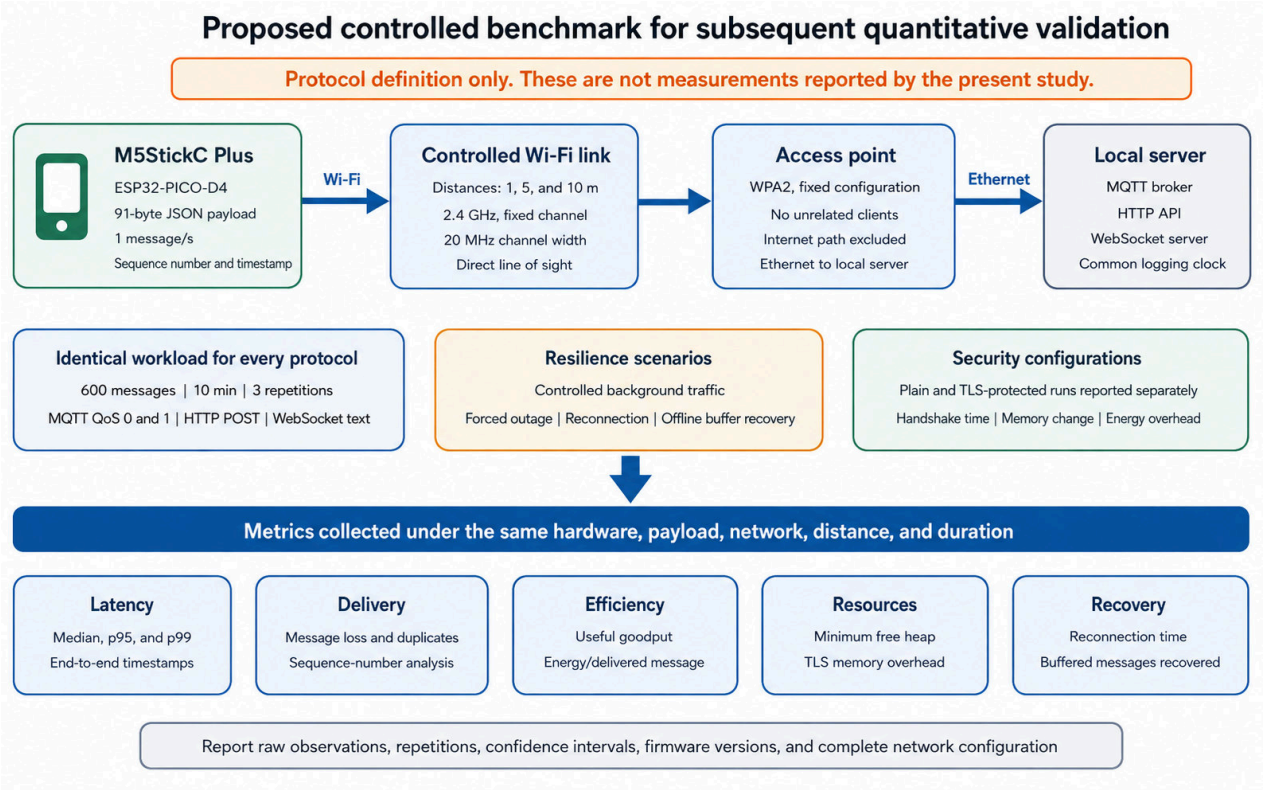


Fig. 3. Proposed controlled benchmark for subsequent quantitative validation. Payload, interval, distance, Wi-Fi configuration, duration, and repetitions are fixed across application protocols. The parameters define future work and are not measurements reported in this paper.

measured. More than one application protocol may coexist, for example MQTT for device ingestion and WebSocket for dashboard delivery.

C. Security, Offline Operation, and Interference

No resilience claim follows from the preliminary connectivity check. A secure implementation requires authenticated onboarding, least-privilege access, credential protection, transport encryption where applicable, replay handling, software update policy, logging, and data minimization [23]–[26]. TLS adds handshake traffic, state, memory use, computation, and energy cost that must be measured rather than assumed from the protocol specification. Similarly, MQTT QoS does not by itself guarantee end-to-end persistence unless client, broker, database, and application behavior are all defined.

Offline buffering is necessary when continuity of the record matters during network loss. A valid test must specify buffer size, overwrite policy, timestamp source, ordering, duplicate detection, replay rate, and behavior after power failure. Because BLE, Wi-Fi, Zigbee, and Thread can share the 2.4-GHz band, interference tests should document channel allocation, coexisting traffic, received signal strength, and environmental conditions. These requirements are included in the proposed benchmark but were not evaluated in the present bench observations.

IV. Discussion

A central contribution of this study is the taxonomic and evidential separation of communication technologies according to their architectural roles. Connectivity technologies and application protocols solve different parts of the communication path, and a health-monitoring architecture often needs one from each category. A BLE sensor may use a smartphone gateway that later sends data through Wi-Fi or cellular access using MQTT or HTTPS. A Wi-Fi prototype may publish through MQTT while a remote service distributes updates through WebSocket. Direct rankings across these categories therefore have no clear experimental meaning.

The requirements matrix supports protocol selection through explicit, measurable requirements rather than universal scenario-based rankings. Selection begins with measurable limits for energy consumption, latency, application-message loss, goodput, range, offline duration, memory use, security overhead, and infrastructure constraints. Only candidates that satisfy architectural constraints should proceed to identical-workload measurements. This approach also prevents a literature result obtained with another radio, board, payload, broker, or network from being presented as a result of the M5StickC Plus observations.

The two devices illuminate different design constraints but do not represent the diversity of current wearables. The Redmi Watch 5 Active is a single closed consumer product. The

M5StickC Plus is a compact development platform, not a medical device, and its ESP32-PICO-D4 is an older SiP that the vendor now marks as not recommended for new designs [12]. Results from either device cannot be generalized to other radios, firmware stacks, antennas, batteries, or clinical workflows.

The contribution is consequently methodological and architectural. The paper provides a transparent methodological basis for a subsequent instrumented study but does not claim to provide a quantitative comparison among protocols. The principal limitations are the absence of controlled repetitions and saved traces; simulation instead of real physiological acquisition; no implementation of MQTT, HTTP, or WebSocket under the same workload; no energy or network instrumentation; no TLS or attack evaluation; no controlled interference; and one example per device category.

Future validation should execute the proposed benchmark on additional platforms and radios, publish firmware and analysis scripts, and report raw traces with clock and instrument uncertainty. It should also transmit data from calibrated physiological sensors under an approved protocol. Any study involving participants or clinical decision support must obtain the required ethical and institutional approvals and define clinical endpoints separately from network performance.

V. Conclusion

This paper reorganized communication choices for wearable IoT health monitoring into a layered architecture. BLE, Wi-Fi, Zigbee, Thread, LoRaWAN, NB-IoT, and LTE-M are connectivity alternatives with different infrastructure assumptions. MQTT, HTTP/REST, and WebSocket are application-layer options used after a suitable network path exists. They should not be ranked as one set of direct alternatives.

The only bench-supported findings are that the Redmi Watch 5 Active exhibited a closed smartphone-mediated workflow and that the M5StickC Plus associated with Wi-Fi and handled simulated structured records. No quantitative, security, autonomy, clinical-suitability, or universal performance claim is made. The requirements matrix and proposed fixed-workload protocol provide a reproducible next step for measuring latency distributions, application-message loss, goodput, energy per delivered message, reconnection, memory use, offline recovery, interference, scalability, and security overhead.

Acknowledgment

The authors acknowledge the Santa Catarina State University (UDESC) for the scholarship that supported this study and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, for support related to research supplies and materials.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010, doi: 10.1016/j.comnet.2010.05.010.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015, doi: 10.1109/COMST.2015.2444095.

[3] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The Internet of Things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015, doi: 10.1109/ACCESS.2015.2437951.

[4] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, and P. Liljeberg, "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach," *Future Generation Computer Systems*, vol. 78, pp. 641–658, 2018, doi: 10.1016/j.future.2017.02.014.

[5] Z. Pang, L. Zheng, J. Tian, S. Kao-Walter, E. Dubrova, and Q. Chen, "Design of a terminal solution for integration of in-home health care devices and services towards the Internet-of-Things," *Enterprise Information Systems*, vol. 9, no. 1, pp. 86–116, 2015, doi: 10.1080/17517575.2013.776118.

[6] S. D. Mamdiwar, A. R, Z. Shakruwala, U. Chadha, K. Srinivasan, and C.-Y. Chang, "Recent advances on IoT-assisted wearable sensor systems for healthcare monitoring," *Biosensors*, vol. 11, no. 10, Art. no. 372, 2021, doi: 10.3390/bios11100372.

[7] S. Hiremath, G. Yang, and K. Mankodiya, "Wearable Internet of Things: Concept, architectural components and promises for person-centered healthcare," in *Proc. 4th EAI Int. Conf. Wireless Mobile Commun. and Healthcare*, 2014, pp. 304–307, doi: 10.1109/MOBIHEALTH.2014.7015971.

[8] C. Gomez, J. Oller, and J. Paradells, "Overview and evaluation of Bluetooth Low Energy: An emerging low-power wireless technology," *Sensors*, vol. 12, no. 9, pp. 11734–11753, 2012, doi: 10.3390/s120911734.

[9] J. Tosi, F. Taffoni, M. Santacatterina, R. Sannino, and D. Formica, "Performance evaluation of Bluetooth Low Energy: A systematic review," *Sensors*, vol. 17, no. 12, Art. no. 2898, 2017, doi: 10.3390/s17122898.

[10] T. Zhang, J. Lu, F. Hu, and Q. Hao, "Bluetooth Low Energy for wearable sensor-based healthcare systems," in *Proc. IEEE Healthcare Innovation Conf.*, 2014, pp. 251–254, doi: 10.1109/HIC.2014.7038922.

[11] A. Dementyev, S. Hodges, S. Taylor, and J. Smith, "Power consumption analysis of Bluetooth Low Energy, ZigBee and ANT sensor nodes in a cyclic sleep scenario," in *Proc. IEEE Int. Wireless Symp.*, 2013, pp. 1–4, doi: 10.1109/IEEE-IWS.2013.6616827.

[12] Espressif Systems, "ESP32-PICO series datasheet," ver. 1.3, 2025. [Online]. Available: https://documentation.espressif.com/esp32-pico_series_datasheet_en.pdf. Accessed: Aug. 25, 2026.

[13] Connectivity Standards Alliance, "Zigbee standard: Full-stack wireless IoT solution." [Online]. Available: https://csa-iot.org/all-solutions/zigbee/. Accessed: Aug. 25, 2026.

[14] Thread Group, "Thread network fundamentals," white paper, Sep. 2022. [Online]. Available: https://portal.threadgroup.org/DesktopModules/Inventures_Document/FileDownload.aspx?ContentID=633. Accessed: Aug. 25, 2026.

[15] LoRa Alliance, "LoRaWAN specification v1.1," Sep. 2023. [Online]. Available: https://resources.lora-alliance.org/technical-specifications/lorawan-specification-v1-1. Accessed: Aug. 25, 2026.

[16] 3GPP, "The cellular Internet of Things," Oct. 2017. [Online]. Available: https://www.3gpp.org/news-events/3gpp-news/c-iot. Accessed: Aug. 25, 2026.

[17] A. Banks, E. Briggs, K. Borgendale, and R. Gupta, Eds., "MQTT version 5.0," OASIS Standard, Mar. 2019. [Online]. Available: https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html.

[18] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, Univ. California, Irvine, CA, USA, 2000.

[19] R. Fielding, M. Nottingham, and J. Reschke, "HTTP semantics," RFC 9110, Jun. 2022, doi: 10.17487/RFC9110.

[20] I. Fette and A. Melnikov, "The WebSocket protocol," RFC 6455, Dec. 2011, doi: 10.17487/RFC6455.

[21] D. Thangavel, X. Ma, A. Valera, H.-X. Tan, and C. K.-Y. Tan, "Performance evaluation of MQTT and CoAP via a common middleware," in *Proc. IEEE 9th Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing*, 2014, pp. 1–6, doi: 10.1109/ISSNIP.2014.6827678.

[22] V. Seoane, C. Garcia-Rubio, F. Almenares, and C. Campo, "Performance evaluation of CoAP and MQTT with security support for IoT environments," *Computer Networks*, vol. 197, Art. no. 108338, 2021, doi: 10.1016/j.comnet.2021.108338.

[23] A. Barua, M. A. Al Alamin, M. S. Hossain, and E. Hossain, "Security and privacy threats for Bluetooth Low Energy in IoT and wearable devices: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 251–281, 2022, doi: 10.1109/OJCOMS.2022.3149732.

[24] M. A. Jan, F. Khan, S. Mastorakis, M. Adil, A. Akbar, and N. Stergiou, "LightIoT: Lightweight and secure communication for energy-efficient IoT in health informatics," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 3, pp. 1202–1211, 2021, doi: 10.1109/TGCN.2021.3077318.

[25] F. Alsubaei, A. Abuhussein, and S. Shiva, "Security and privacy in the Internet of Medical Things: Taxonomy and risk assessment," in *Proc. IEEE 42nd Conf. Local Computer Networks Workshops*, 2017, pp. 112–120, doi: 10.1109/LCN.Workshops.2017.72.

[26] E. Rescorla, "The Transport Layer Security (TLS) protocol version 1.3," RFC 8446, Aug. 2018, doi: 10.17487/RFC8446.

[27] Xiaomi, "Redmi Watch 5 Active specifications." [Online]. Available: https://www.mi.com/global/product/redmi-watch-5-active/specs/. Accessed: Aug. 25, 2026.

[28] M5Stack, "M5StickC Plus documentation, product K016-P." [Online]. Available: https://docs.m5stack.com/en/core/m5stickc_plus. Accessed: Aug. 25, 2026.

[29] Arduino, "Arduino IDE 2.3.10." [Online]. Available: https://www.arduino.cc/en/software. Accessed: Aug. 25, 2026.

[30] Espressif Systems, "Arduino-ESP32 documentation, release 3.3.11." [Online]. Available: https://docs.espressif.com/projects/arduino-esp32/en/latest/. Accessed: Aug. 25, 2026.

Comparative genome and resistance profiling of *Pseudomonas aeruginosa* across diverse niches

Daniel Freire da Silva
Programa de Engenharia Biomédica (COPPE)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
danfrey998@gmail.com

Eidy de Oliveira Santos
Faculdade de Ciências Biológicas e Saúde (FCBS)
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro, Brasil
eidyos@gmail.com

Diogo Antonio Tschoeke
Programa de Engenharia Biomédica (COPPE)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brasil
diogoat@peb.ufrj.br

***Abstract*—The rise and spread of multidrug-resistant bacteria have been recognized as a global public health problem. Under selective pressure from the abusive and uncontrolled use of antibiotics by humans, bacteria develop or enhance resistance mechanisms shared within the microbial community in the environment. Multidrug-resistant pathogens in hospital settings overburden healthcare systems and generate considerable global economic costs due to prolonged hospital stays. The ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa,* and *Enterobacter* spp.) were considered a significant challenge, especially in hospital settings where patients with compromised immune systems are particularly susceptible to bacterial infections. The main objective of this study was to investigate the genomic diversity and antimicrobial resistance profile in *Pseudomonas aeruginosa* isolates from environmental, hospital, and animal niches. To this end, a comparative analysis of complete genomes deposited in GenBank and the International Pseudomonas Consortium Database was performed using bioinformatics techniques. The results indicate significant variation in the number of antimicrobial resistance genes (ARGs) among the isolates analyzed. Although the profiles of hospital and environmental genomes were highly similar, the data indicated a trend toward differentiation among chicken isolates, as evidenced by distinct clustering and a higher ARG load. Pangenome analysis revealed a set of highly plastic genes, highlighting the species adaptive capacity. The results of this work contribute to a better understanding of antimicrobial resistance dissemination in *P. aeruginosa* and may assist in developing more effective strategies to control infections caused by this pathogen.**

Index Terms*—*Pseudomonas aeruginosa; antimicrobial resistance; comparative genomics; bioinformatics.

I. Introduction

Antimicrobial resistance (AMR), a natural process, is recognized by global organizations, such as the World Health Organization and the World Bank, as one of the greatest threats [1]. It is estimated that, in 2019, AMR was associated with 4.95 million global deaths, with 1.27 million directly attributed to bacterial resistance [2], [3]. The rapid bacterial adaptation and the stagnation of new antimicrobial class development since the 1960s [4], [5] threaten the safety of essential medical procedures such as surgeries and cancer treatments increasing mortality and hospital costs [6], [7]. This problem, however, is not restricted to the hospital environment, as the spread of multidrug-resistant strains is an inevitable consequence of the use of antibiotics. This broad spectrum of contamination favors the selection of highly versatile bacteria, capable of moving and thriving between natural reservoirs and the clinical environment [8], [9]. Within this context, one of the main examples is *Pseudomonas aeruginosa*. This gram-negative opportunistic pathogen stands out for its versatility and ability to grow in diverse environments, from soil and water to hospital surfaces. It is an aerobic bacterium, with motility provided by one or more polar flagella [10], and it belongs to the ESKAPE group of pathogens — *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa,* and *Enterobacter* spp. — [11], [12]. This pathogen represents a serious threat to public health, infecting patients with various conditions such as burns, immunodeficiency, and cancer, in addition to being dominant in chronic infections such as cystic fibrosis and chronic obstructive pulmonary disease. Its pathogenic action can affect the lungs, bloodstream, and urinary tract, resulting in acute and chronic infections [13], [14].

P. aeruginosa possesses a large and variable genome (GC content of 65–67%, 5–7 Mb) that significantly contributes to its adaptability. Its arsenal of virulence factors and antimicrobial resistance mechanisms confers remarkable metabolic flexibility and the ability to adapt to diverse conditions, including the host's immune response [15].

This adaptability is reflected in its prevalence as one of the bacteria with the highest number of multidrug-resistant strains to almost all commonly used antibiotics, including aminoglycosides, cephalosporins, fluoroquinolones, and especially carbapenems, resulting in high morbidity and mortality [16]. Due to these characteristics, it is responsible for approximately 10% of hospital-acquired infections. Carbapenem-resistant *P. aeruginosa* was classified by the WHO as a "critical group" pathogen in 2017 and subsequently reclassified as a "high group" pathogen, requiring urgent efforts in the development of new antimicrobials [17].

II. Materials and Methods

A. Data collection

Genomic data from 32 *P. aeruginosa* isolates representing diverse ecological niches — including natural environments (n = 11), animal hosts (e.g., cattle, cows, and chickens; n = 10), and hospital settings worldwide (n = 11) — were analysed [18], together with the reference genomes PAO1 (GCA_000006765.1) [19] and PA14 (GCF_025490475.1). To build this dataset, pre-assembled genomes and raw sequence reads were retrieved from public repositories based on strict curation and selection criteria: (i) raw sequence reads from the Sequence Read Archive (SRA) were required to have complete metadata for all collection parameters, corresponding to the columns in Table I; (ii) isolation sources were chosen to be highly diverse to ensure a broad and general ecological representation; (iii) the origins of the genomes were balanced (n = 10 per niche) to prevent representation bias; and (iv) de novo assemblies were required to have a genomic completeness of 98% or more and preferably a maximum of 300 contigs. Assembled environmental genomes and reference strains were obtained from the International Pseudomonas Consortium Database [20] and GenBank, respectively. The outcomes of the data pre-processing workflow are summarised in Table I.

B. Analysis of genomic sequences

For raw sequence reads, quality control and read filtering were performed using fastp v1.3.3 [21]. The filtered reads were subjected to de novo assembly using SPAdes v3.15.5 [22]. The quality of all genome assemblies was evaluated using QUAST v5.0.2 [23]. Structural and functional annotation of the bacterial genome was conducted using Prokka v1.14.5 [24]. The predicted protein sequences (*.faa) generated by Prokka were used as input for the identification of ARGs with CARD RGI [2], applying the "perfect" and "strict" hit algorithms for genome data.

To infer evolutionary relationships and identify duplication events among the identified ARGs, predicted protein sequences were clustered into orthogroups using OrthoFinder v2.5.5 [25]. Orthogroup assignments were subsequently integrated with RGI-derived annotations via unique protein identifiers. This cross-referencing enabled the extraction of orthogroups containing ARGs, allowing for the targeted identification of intra-genomic duplications and assessment of their distribution across isolates.

The Average Nucleotide Identity (ANI) was calculated using FastANI v1.34 [26] in all-vs-all mode, estimating a 95% limit for species delimitation. The pan-genome was analyzed with Roary [27]. Statistical analyses and visualizations were conducted in R v4.6.0, utilizing the tidyverse packages [28], ComplexHeatmap [29], and PCAtools (Bioconductor) [30].

To evaluate the differences in ARG abundance across ecological niches and isolation sources, data normality was assessed using the Shapiro-Wilk test [31]. Due to the nonnormal distribution, the nonparametric Kruskal-Wallis test was used to compare the abundance of ARGs among groups. Furthermore, a Spearman test [32] was calculated to investigate whether the observed variation in ARG abundance correlates with genome size, thereby evaluating potential genome size bias.

III. Results

A. ANI and Pan-genome

In comparative genome analyses, the ANI values ranged from 98% to 99% across all genomes, without clustering associated with the sources of isolation. The species pangenome was characterized as open, having a core genome consisting of 4,694 genes and an accessory genome with 13,442 genes (Fig. 1).

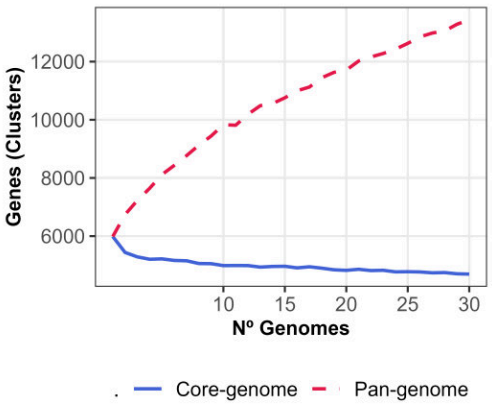


Fig. 1. Analysis of the *P. aeruginosa* pangenome.

B. Resistome Profiling and Resistance Determinants

Resistome profiling of the present dataset revealed substantial heterogeneity among the 32 sequenced *P. aeruginosa* genomes, with total ARG) load ranging from 57 to 80 genes per genome. Chicken-derived isolates exhibited the highest ARG burdens (66–80 genes per genome), whereas genomes from remaining sources exhibited narrower ranges (57–66 genes). Notably, hospital-derived genomes showed ARG abundance comparable to, or slightly lower than, those observed in environmental isolates. Absolute ARG abundance per genome deviated from normality (Shapiro–Wilk: $W = 0.69, p < 0.001$). Spearman rank correlation analysis indicated a weak, nonsignificant association between genome size and ARG abundance (Spearman's $\rho = 0.18, p = 0.317$). In addition, ARG abundance did not differ significantly among major ecological niches (Kruskal–Wallis: $X^2 = 4.32, df = 2, p = 0.115$) (Fig. 2). Given the limited sample sizes within specific isolation sources, downstream comparisons at this level were confined to descriptive analyses.

Across all niches, was identified 64 core ARGs conserved in every genome. This core resistome predominantly comprised genes encoding intrinsic multidrug efflux systems and antibiotic inactivation determinants. The accessory resistome displayed pronounced niche-associated structuring, with discrete ARG subsets shared between specific pairs of niches or uniquely restricted to a single isolation source (Fig. 4.

TABLE I
METADATA AND ASSEMBLY QUALITY OF THE *P. aeruginosa* ISOLATES.

Country	Genome	Source	Niche	Contigs	GC%	N50	Size (bp)	ARG	Collect Date
China	GCA_030326365.1	Chicken	Animal	3	65.66	315323	7006757	77	2019
China	GCA_030326375.1	Chicken	Animal	2	65.59	214488	7066322	70	2019
China	GCA_030326405.1	Chicken	Animal	86	66.58	263062	6226525	80	2019
China	GCA_030326425.1	Chicken	Animal	209	66.58	261105	6225479	80	2019
China	GCA_020011175.1	Chicken	Animal	4	65.9	781852	6776179	67	2018
Brazil	SRR11455646	Cattle: Urine	Animal	37	66.58	426781	6225584	58	2018
Turkey	GCA_021556555.1	Cow	Animal	29	65.68	6438660	6987046	61	2019
Brazil	SRR11455632	Cattle: Urine	Animal	42	66.58	369999	6227785	58	2019
Turkey	GCA_021556505.1	Cow	Animal	45	66.41	431554	6368118	58	2020
Turkey	GCA_021556635.1	Cow	Animal	40	65.8	6422085	6822940	60	2021
Brazil	SRR10043046	Mangrove	Environmental	508	65.67	30149	7078673	58	2017
Slovenia	SRR6048059	Wastewater	Environmental	39	66.45	411511	6352960	60	2020
Brazil	SRR35795853	Soil	Environmental	101	65.95	191760	6986128	66	2020
Brazil	SRR35795855	Soil	Environmental	92	65.95	233476	6987669	66	2020
Slovenia	SRR6048071	Wastewater	Environmental	131	65.97	126514	6958232	60	2020
Slovenia	SRR6048070	Wastewater	Environmental	72	66.21	188845	6778681	60	2020
Brazil	SRR13065837	Orange	Environmental	45	66.34	350905	6463328	60	2020
Colombia	SRR30732520	Oil contaminated soil	Environmental	58	66.52	383952	6280970	61	2020
Colombia	SRR30732519	Oil contaminated soil	Environmental	48	66.57	270068	6222363	58	2020
Colombia	SRR30732521	Oil contaminated soil	Environmental	51	66.52	373979	6292065	61	2020
Lithuania	ERR10438653	Urine	Hospital	91	66.37	207796	6518243	59	2004
Thailand	ERR10432579	Sputum	Hospital	67	66.36	224399	6419349	59	2014
France	ERR10437382	Blood	Hospital	63	66.48	370753	6325406	61	2014
Greenland	ERR10438051	Urine	Hospital	84	66.52	154477	6291196	61	2014
Brazil	SRR12354234	Urine	Hospital	37	66.06	830080	6761881	59	2015
Bangladesh	ERR10431370	Swab (wound)	Hospital	63	66.61	243044	6170110	60	2019
Paraguay	ERR10431634	Sputum	Hospital	133	66.02	144534	6397473	57	2019
Paraguay	ERR10431636	Blood	Hospital	70	66.45	206383	6365555	58	2019
Paraguay	ERR10431652	Aspirate	Hospital	148	66.02	200018	6844602	61	2021
Spain	ERR10432412	Blood	Hospital	94	65.94	182649	6765319	57	2021
Na	GCF_025490475.1	PA14	Hospital	5	66	1640262	6537581	59	2022
Na	GCA_000006765.1	PAO1	Environmental	1	66.05	6538179	6778344	60	2006

Genomes starting with SRR or ERR denote raw sequencing data processed through the quality control pipeline.

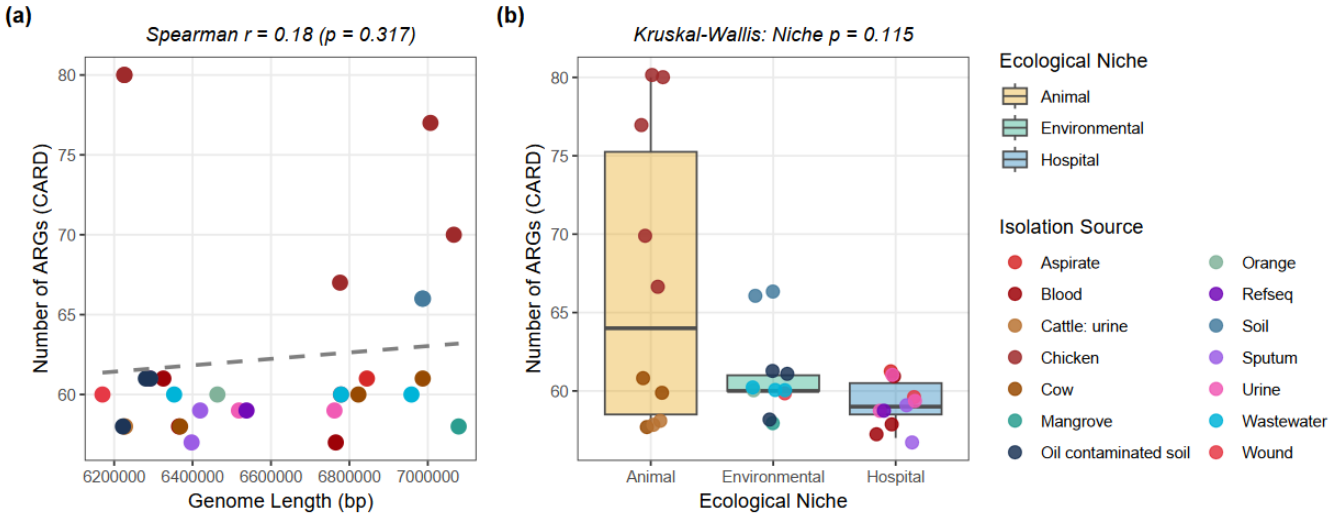


Fig. 2. ARG content analysis in *P. aeruginosa* isolates across ecological niches: (a) Spearman rank correlation analysis between total genome length (bp) and ARG abundance ($\rho = 0.18$, $p = 0.317$); (b) distribution of ARG abundance grouped by major ecological niche (Kruskal–Wallis test, $p = 0.115$).

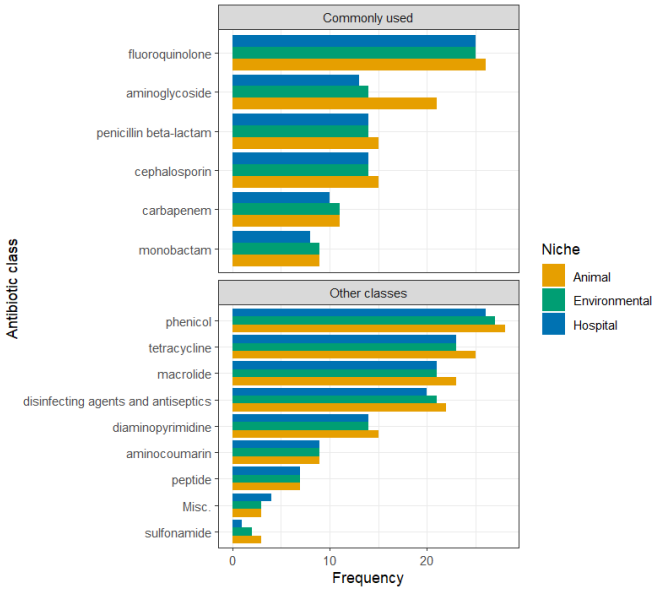


Fig. 3. Frequency of resistance compared across sample origins

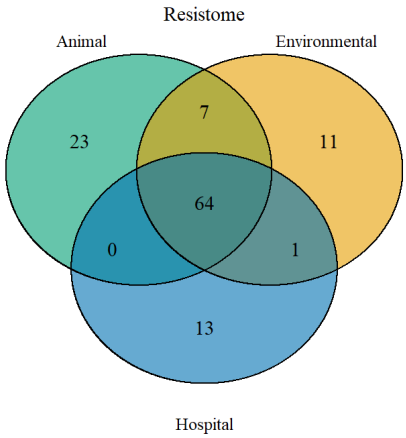


Fig. 4. A Venn diagram showing ARG data, categorized by the origin of the samples.

To assess whether niche-exclusive ARGs were associated with specific resistance mechanisms, a contingency table was created showing the relationship between resistance mechanism category and ecological niche, based on the sets of exclusive genes derived from the Venn diagram (Table: II). Given the small and unevenly distributed cell counts, Fisher's exact test was used as the primary test ($p = 0.685$). The chi-square test was performed as a complementary check ($X^2 = 7.28, df = 8, p = 0.507$). Both tests indicated no statistically significant association between mechanism type and ecological niche in the exclusive resistome. This suggests that, in this dataset, the resistance mechanisms captured in niche-exclusive genes are broadly distributed rather than niche-specific.

The predominant classes represented were fluoroquinolones, macrolides, tetracyclines, and peptides. Animal-derived isolates generally showed the highest frequency of resistance

TABLE II
EXCLUSIVE ARGS AND MECHANISMS ACROSS ECOLOGICAL NICHES.

Niche	Mechanism	Exclusive ARGs
Animal	Efflux pump	*cmx*, *floR*, *tet(C)*
	Target inactivation	*AAC(3)-IVa*, *AAC(6')-IIa*, *AAC(6')-Ib9*, *APH(3")-Ib*, *APH(3')-Ia*, *APH(3')-VIa*, *APH(4)-Ia*, *APH(6)-Id*, *IMP-45*, *OXA-1*, *OXA-1127*, *PDC-31*, *TEM-1*, *arr-2*, *mphE*, *tet(X6)*
	Target alteration	*armA*
	Target protection	*QnrVC1*, *msrE*
	Target replacement	*dfrA22*
Environmental	Target inactivation	*AAC(6')-Il*, *KPC-2*, *OXA-395*, *OXA-847*, *OXA-908*, *PDC-113*, *PDC-115*, *PDC-68*, *aadA*, *arr-4*
	Target alteration	*parE*
Hospital	Target inactivation	*OXA-1032*, *OXA-1035*, *OXA-1129*, *OXA-486*, *OXA-488*, *PDC-101*, *PDC-14*, *PDC-197*, *PDC-19*, *PDC-34*, *PDC-38*, *PDC-433*
	Target alteration	*vanT* gene in *vanG* cluster

determinants across most antibiotic classes, with a particular enrichment of aminoglycoside resistance genes. The overall frequency classes in hospital-derived genomes remained similar to or slightly lower than that detected in isolates from natural environments (Fig. 3).

Chicken-derived genomes were further distinguished by the presence of target protection (*QnrVC1* and *msrE*) and target replacement (*dfrA22*) mechanisms, which were otherwise detected only in three environmental genomes (Fig. 7b). The *KPC-β-lactamase* gene, originally described in *Klebsiella pneumoniae* [33], was identified in two environmental genomes. Compared to other strains, one mangrove isolate lacked several genes associated with efflux systems (*mexM, oprM, muxA, mexS,* and *pmpM*), suggesting a gene deletion event in this isolate.

Several examples of ARG duplication were observed among all genomes studied. Duplication events were observed in the main efflux systems of the bacterium, such as *rsmA, bcr-1, mexY, mexD, mexI, mexS,* and *opmH*. Hospital- and chicken-derived genomes showed duplication in the *APH(6)-Id* genes, which confer resistance to aminoglycosides through an inactivation mechanism (Fig. 5).

C. Resistome Topology, Clustering, and Ordination Analysis

To explore whether similarity in resistome profiles among isolates is associated with phylogeny or influenced by niche-specific selective pressures, a tanglegram analysis was performed comparing whole-genome ANI with resistome topology (Jaccard distance of ARG presence/absence). This comparison revealed marked topological discordance, characterized by frequent line crossings connecting environmental and hospital isolates across both dendrograms. Concordance was restricted to a single pair of soil isolates with shared spatio-temporal origin (Brazil, 2020; highlighted in red). This pattern indicates that while vertical clonal expansion contributes to maintaining resistome structures in select lineages, the accessory resistome

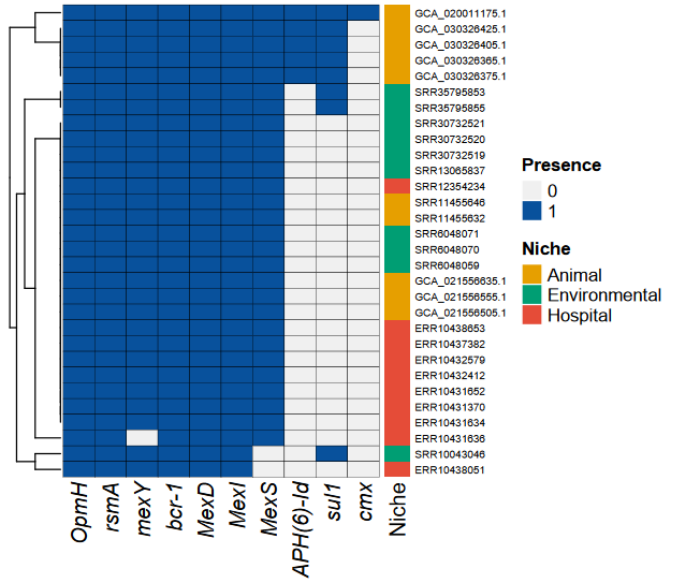


Fig. 5. Presence–absence matrix of duplicated ARG across environmental, animal, and hospital isolates.

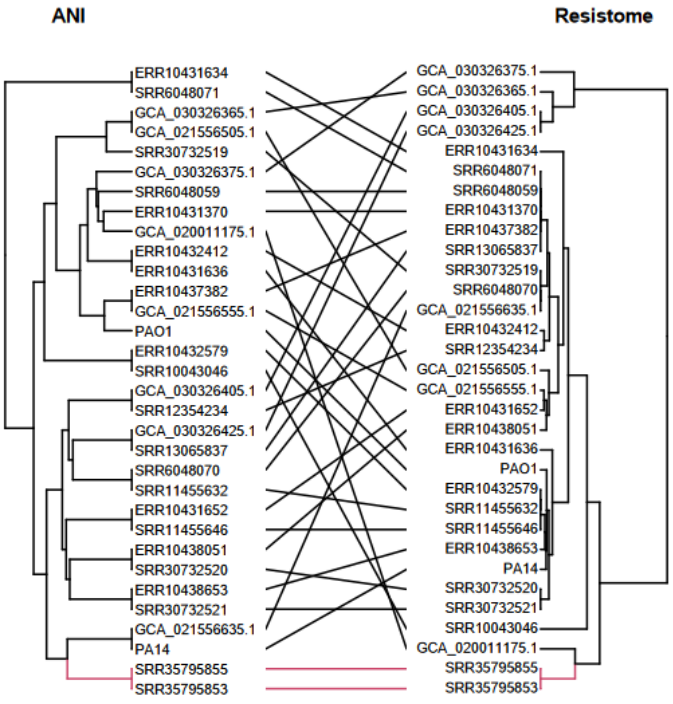


Fig. 6. Tanglegram comparing the genomic background (ANI, left) and the resistome profile (Jaccard distance of ARG presence/absence, right) of *P. aeruginosa* isolates. Lines connect identical isolates across both trees.

of *P. aeruginosa* is highly dynamic and reshaped primarily by localized selective pressures and independent acquisition or loss of resistance determinants (Fig. 6).

Principal component analysis (PCA) was conducted to identify the antimicrobial resistance genes (ARGs) that most strongly contributed to the separation among genomic sample groups. The first two principal components explained 55.32% of the total variance, with PC1 accounting for 35.41% and PC2 for 19.91% (Fig. 7a). A distinct and well-separated cluster comprising chicken-derived isolates was observed relative to the remaining groups. This dispersion was strongly associated with a broad complement of acquired resistance determinants, particularly *arr-2, dfrA22,* and *floR*, in addition to multiple aminoglycoside-resistance gene variants (*AAC(3)-IVa, AAC(6')-Ib9, APH(3')-Ia, APH(3')-VIa, APH(4)-Ia,* and *APH(6)-Id*).

In contrast, genomes of environmental and hospital-derived largely overlapped in ordination space or were separated by only small distances, forming a cohesive cluster indicative of highly similar resistome profiles. The modest separation between these niches was primarily attributable to variation in genes linked to intrinsic resistance mechanisms—most notably efflux system components (*armR*, *mexM*, *muxA*, *muxC*, and *pmpM*)—rather than differences driven by acquisition of additional resistance genes.

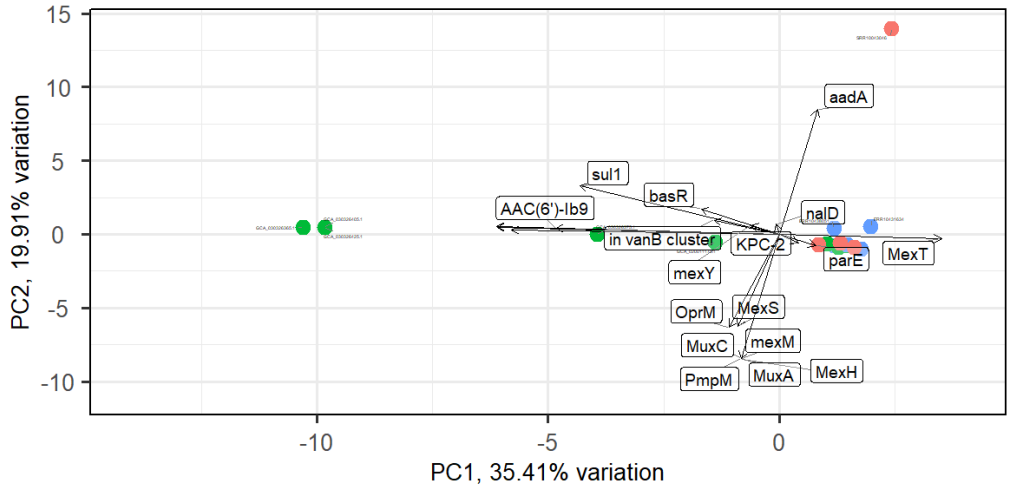
IV. Discussion

In this comparative genomic analysis, ANI values across all genomes ranged from 98% to 99%. This finding indicates that, although the strains were isolated from distinct ecological niches, they retain a highly conserved core genome. Conversely, the open pangenome of *P. aeruginosa* is consistent with substantial genomic plasticity and ecological versatility, encompassing accessory genes and virulence-associated determinants that may facilitate colonization of diverse environments and adaptation to niche-specific selective pressures [34].

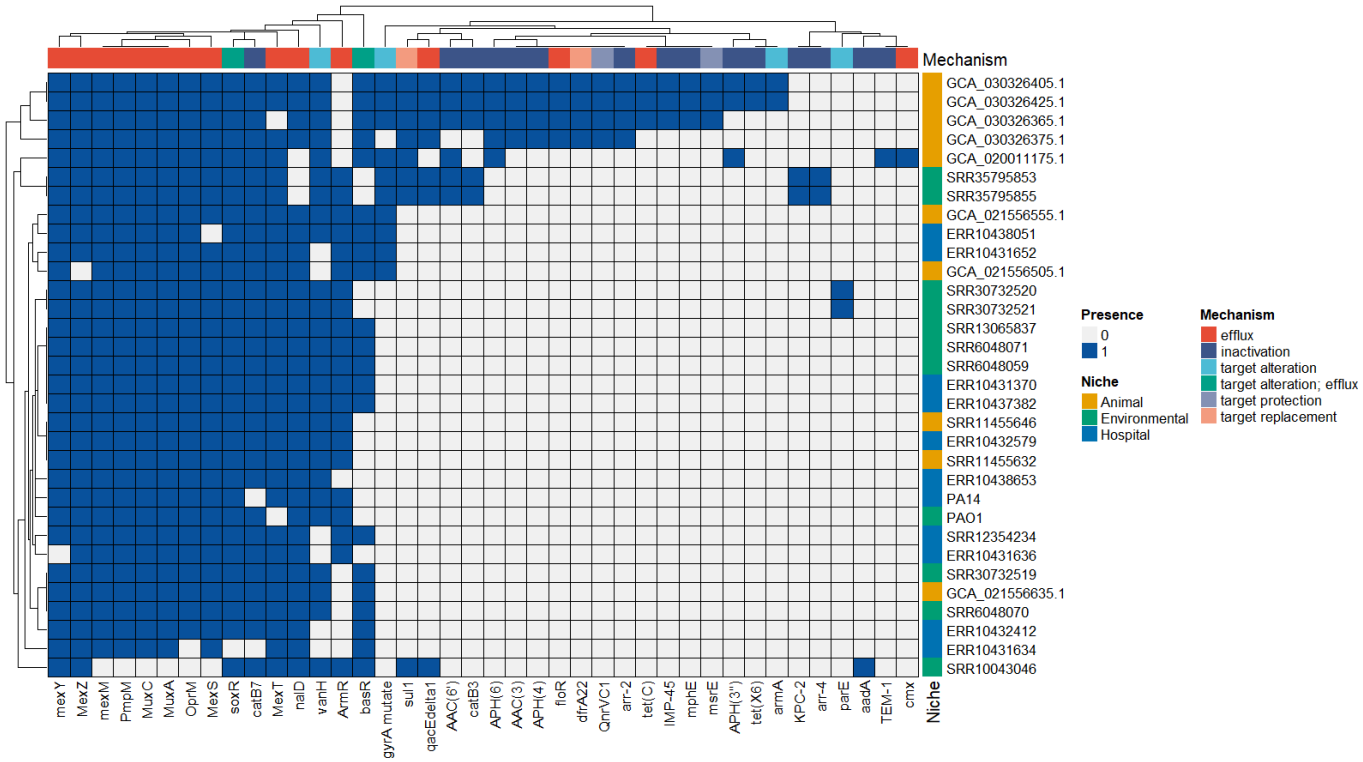
This genomic plasticity is reflected by niche-specific molecular signatures. The detection of *blaKPC* in two environmental isolates suggests that clinically critical carbapenem resistance determinants circulate outside hospital settings, although this finding alone does not establish the direction or frequency of transmission. In addition, the absence of specific efflux pump genes (*mexM*, *pmpM*, *muxA*, *muxC*, *oprM*, and *mexS*) in the mangrove-derived isolate may reflect relaxed selective pressure in natural habitats.

Duplication events involving key RND efflux components and regulatory genes (*mexY, MexI, MexS, and OpmH*) and in the *APH(6)-Id* gene, in hospital and chicken isolates, point to an adaptive gene dosage mechanism essential for amplifying antimicrobial tolerance in environments with greater exposure [35]. Gene duplication can generate new genes with distinct functions, facilitating rapid adaptation to environmental changes. The copies can evolve toward advantageous functions, with a preference for functional classes involved in transcription, metabolism, and defense mechanisms. The copy, being free to accumulate mutations and acquire new functions, allows for the preservation of the original function, avoiding the risk of loss that could occur in direct mutation [36], [37].

Elevated ARG content was observed in the profiles of animal-derived isolates, with overall resistance levels exceeding those in other genomes. This pattern is particularly reflected in the



(a) Principal Component Analysis.



(b) Resistome heatmap.

Fig. 7. Comparative resistome profiling of *P. aeruginosa* isolates across ecological niches: (a) PCA ordination plot displaying sample segregation based on resistance gene profiles; (b) matrix of presence–absence AGRs across environmental, animal, and hospital isolates.

chicken isolates, which stood out by presenting between 66 and 80 ARGs. Notably, the resistance mechanisms among ***P. aeruginosa*** genomes isolated from chickens and certain environmental isolates differ from those in other genomes due to the presence of target protection (*QnrVC1* and *msrE*) and target replacement (*dfrA22*) mechanisms.

While the number of ARGs varies between isolation sources, there does not appear to be a direct relationship between the quantity of ARGs and the isolation source. Chickens-derived isolates showed the highest number and diversity of ARGs, acquiring resistance to carbapenems, cephalosporins, monobactams, penicillin, and aminoglycosides. The genes that confer resistance to aminoglycosides are the most frequent. This class of broad-spectrum antibiotics is commonly used in poultry farming. The genes that confer resistance to aminoglycosides are the most frequent. This class of broad-spectrum antibiotics is commonly used in poultry farming [38], [39]. Although it is presumed that there are more cases of multidrug-resistant strains in hospital settings, studies suggest that the prophylactic use of antibiotics in animal production as growth promoters exerts equal or greater selective pressure. The lack of restrictive control over antibiotic use, as seen in hospital settings and in animal production, favors the emergence and spread of strains adapted to specific classes of antibiotics [40].

V. Conclusion

P. aeruginosa exemplifies bacterial adaptability, exhibiting the capacity to colonize environments ranging from natural ecosystems to complex clinical settings. This ecological versatility is driven by an open pangenome and a diverse repertoire of resistance and virulence determinants that facilitate both acute and chronic infections. Despite high overall sequence similarity, substantial genomic plasticity—evidenced by gene duplication, localized gene loss, and the capacity to acquire clinically relevant resistance determinants (e.g., *KPC-β-lactamase*, *APH*, and *AAC* genes) —supports rapid niche adaptation. Although the absolute number of resistance determinants does not unambiguously distinguish across isolation sources studied. Across all groups, fluoroquinolone resistance constituted the most prevalent mechanism, whereas resistance to aminoglycosides and β-lactams was disproportionately represented among animal-derived isolates. These observations indicate that the *P. aeruginosa* resistome is heavily shaped by specific selective pressures; however, larger datasets and further functional validation are required to better delineate these ecological and mechanistic relationships.

References

[1] S. Hernando-Amado, T. M. Coque, F. Baquero, and J. L. Martínez, "Defining and combating antibiotic resistance from One Health and Global Health perspectives," *Nature Microbiology*, vol. 4, no. 9, pp. 1432–1442, 2019.

[2] G. R. Aguilar, L. R. Swetschinski, N. D. Weaver, K. S. Ikuta, T. Mestrovic, A. P. Gray, E. Chung, E. E. Wool, C. Han, A. G. Hayoon, D. T. Araki, A. Abdollahi, A. Abu-Zaid, M. Adnan, R. Agarwal, J. A. Dehkordi, A. Y. Aravkin, D. Areda, A. Y. Azzam, E. N. Berezin, A. S. Bhagavathula, Z. A. Bhutta, S. S. Bhuyan, A. J. Browne, C. A. Castañeda-Orjuela, E. K. Chandrasekar, P. R. Ching, X. Dai, G. L. Darmstadt, F. P. De La Hoz, N. Diao, D. Diaz, W. Mombaque Dos Santos, D. Eyre, C. Garcia, G. Haines-Woodhouse, M. B. Hassen, N. J. Henry, S. Hopkins, M. M. Hossain, K. C. Iregbu, C. C. Iwu, J. A. Jacobs, M. M. Janko, R. Jones, I. M. Karaye, I. A. Khalil, I. A. Khan, T. Khan, J. Khubchandani, S. Khusuwan, A. Kisa, G. W. Koyaweda, F. Krapp, E. A. Kumaran, H. H. Kyu, S. S. Lim, X. Liu, S. Luby, S. B. Maharaj, C. Maronga, M. Martorell, J. May, B. McManigal, A. H. Mokdad, C. E. Moore, E. Mostafavi, E. Murillo-Zamora, M. M. Mussi-Pinhata, R. Nanavati, H. Nassereldine, Z. S. Natto, F. N. Qamar, V. Nuñez-Samudio, T. J. Ochoa, T. R. Ojo-Akosile, A. T. Olagunju, A. Olivas-Martinez, E. Ortiz-Brizuela, P. Ounchanum, J. L. Paredes, V. S. Patthipati, S. Pawar, M. Pereira, A. Pollard, A. Ponce-De-Leon, E. J. Sady Prates, I. Qattea, L. F. Reyes, E. Roilides, V. D. Rosenthal, K. E. Rudd, W. Sangchan, S. Seekaew, A. Seylani, N. Shababi, S. Sham, J. Sifuentes-Osornio, H. Singh, A. Stergachis, N. Tasak, N. Y. Tat, A. Thaiprakong, P. R. Valdez, D. Y. Yada, I. Yunusa, M. S. Zastrozhin, S. I. Hay, C. Dolecek, B. Sartorius, C. J. Murray, and M. Naghavi, "The burden of antimicrobial resistance in the Americas in 2019: a cross-country systematic analysis," *The Lancet Regional Health - Americas*, vol. 25, p. 100561, Sept. 2023.

[3] A. N. Poudel, S. Zhu, N. Cooper, P. Little, C. Tarrant, M. Hickman, and G. Yao, "The economic burden of antibiotic resistance: A systematic review and meta-analysis," *PLOS ONE*, vol. 18, p. e0285170, May 2023.

[4] G. A. Durand, D. Raoult, and G. Dubourg, "Antibiotic discovery: history, methods and perspectives," *International Journal of Antimicrobial Agents*, vol. 53, pp. 371–382, Apr. 2019.

[5] Q. Kong and Y. Yang, "Recent advances in antibacterial agents," *Bioorganic & Medicinal Chemistry Letters*, vol. 35, p. 127799, Mar. 2021.

[6] R. C. MacLean and A. San Millan, "The evolution of antibiotic resistance," *Science*, vol. 365, no. 6458, pp. 1082–1083, 2019.

[7] S. Santajit and N. Indrawattana, "Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens," *BioMed Research International*, vol. 2016, pp. 1–8, 2016.

[8] J. G. Navedo, V. Araya, and C. Verdugo, "Upraising a Silent pollution: Antibiotic Resistance at Coastal Environments and Transference to long-distance Migratory Shorebirds," *Science of The Total Environment*, vol. 777, p. 146004, 2021.

[9] N. Sabtu, D. A. Enoch, and N. M. Brown, "Antibiotic resistance: what, why, where, when and how?," *British Medical Bulletin*, vol. 116, no. 1, pp. 105–113, 2015.

[10] S. L. Gellatly and R. E. Hancock, "Pseudomonas aeruginosa: new insights into pathogenesis and host defenses," *Pathogens and Disease*, vol. 67, no. 3, pp. 159–173, 2013.

[11] D. M. P. De Oliveira, B. M. Forde, T. J. Kidd, P. N. A. Harris, M. A. Schembri, S. A. Beatson, D. L. Paterson, and M. J. Walker, "Antimicrobial Resistance in ESKAPE Pathogens," *Clinical Microbiology Reviews*, vol. 33, pp. e00181–19, June 2020.

[12] L. Rice, "Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE," *The Journal of Infectious Diseases*, vol. 197, no. 8, pp. 1079–1081, 2008.

[13] J. Eklöf, R. Sørensen, T. Ingebrigtsen, P. Sivapalan, I. Achir, J. Boel, J. Bangsborg, C. Ostergaard, R. Dessau, U. Jensen, A. Browatzki, T. Lapperre, J. Janner, U. Weinreich, K. Armbruster, T. Wilcke, N. Seersholm, and J. Jensen, "Pseudomonas aeruginosa and risk of death and exacerbations in patients with chronic obstructive pulmonary disease: an observational cohort study of 22 053 patients," *Clinical Microbiology and Infection*, vol. 26, pp. 227–234, Feb. 2020.

[14] S. Finch, M. J. McDonnell, H. Abo-Leyah, S. Aliberti, and J. D. Chalmers, "A Comprehensive Analysis of the Impact of *Pseudomonas aeruginosa* Colonization on Prognosis in Adult Bronchiectasis," *Annals of the American Thoracic Society*, vol. 12, pp. 1602–1611, Nov. 2015.

[15] I. Jurado-Martín, M. Sainz-Mejías, and S. McClean, "Pseudomonas aeruginosa: An Audacious Pathogen with an Adaptable Arsenal of Virulence Factors," *International Journal of Molecular Sciences*, vol. 22, no. 6, 2021.

[16] M. F. Moradali, S. Ghods, and B. H. A. Rehm, "Pseudomonas aeruginosa Lifestyle: A Paradigm for Adaptation, Survival, and Persistence," *Frontiers in Cellular and Infection Microbiology*, vol. 7, no. 39, 2017.

[17] H. Sati, E. Carrara, A. Savoldi, P. Hansen, J. Garlasco, E. Campagnaro, S. Boccia, J. A. Castillo-Polo, E. Magrini, P. Garcia-Vello, E. Wool, V. Gigante, E. Duffy, A. Cassini, B. Huttner, P. R. Pardo, M. Naghavi, F. Mirzayev, M. Zignol, A. Cameron, E. Tacconelli, A. Aboderin, M. Al Ghoribi, J. Al-Salman, A. Amir, A. Apisarnthanarak, M. Blaser, A. El-Sharif, S. Essack, S. Harbarth, X. Huang, G. Kapoor, G. Knight, J. C. Muhwa, D. L. Monnet, T. Ousassa, R. Sacsaquispe, J. Severin,

M. Sugai, N. Taneja, and A. Umubyeyi Nyaruhirira, "The WHO Bacterial Priority Pathogens List 2024: a prioritisation study to guide research, development, and public health strategies against antimicrobial resistance," *The Lancet Infectious Diseases*, vol. 25, pp. 1033–1043, Sept. 2025.

[18] S. Nag, G. Larsen, J. Szarvas, L. E. K. Birkedahl, G. M. Gulyás, W. J. Ciok, T. M. Lagermann, S. Tafaj, S. Bradbury, P. Collignon, D. Daley, V. Dougnon, K. Fabiyi, B. Coulibaly, R. Dembélé, G. Nikiema, N. Magloire, I. J. Ouindgueta, Z. Z. Hossain, A. Begum, D. Donchev, M. Diggle, L. Turnbull, S. Lévesque, L. Berlinger, K. K. Sogaard, P. D. Guevara, C. D. Valderrama, P. Maikanti, J. Amlerova, P. Drevinek, J. Tkadlec, M. Dilas, A. Kaasch, H. T. Westh, M. A. Bachtarzi, W. Amhis, C. E. S. Salazar, J. Villacis, M. A. D. Lúzon, D. B. Palau, C. Duployez, M. Paluche, S. Asante-Sefa, M. Moller, M. Ip, I. Mareković, A. Pál-Sonnevend, C. E. Cocuzza, A. Dambrauskiene, A. Macanze, A. Cossa, I. Mandomando, P. Nwajiobi-Princewill, I. N. Okeke, A. O. Kehinde, I. Adebiyi, I. Akintayo, O. Popoola, A. Onipede, A. Blomfeldt, N. E. Nyquist, K. Bocker, J. Ussher, A. Ali, N. Ullah, H. Khan, N. W. Gustafson, I. Jarrar, A. Al-Hamad, V. Luvira, W. Paveenkittiporn, I. Baran, J. C. L. Mwansa, L. Sikakwa, K. Yamba, R. S. Hendriksen, and F. M. Aarestrup, "Whole genomes from bacteria collected at diagnostic units around the world 2020," *Scientific Data*, vol. 10, p. 628, Sept. 2023.

[19] C. K. Stover, X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson, "Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen," *Nature*, vol. 406, no. 6799, pp. 959–64, 2000.

[20] L. Freschi, J. Jeukens, I. Kukavica-Ibrulj, B. Boyle, M.-J. Dupont, J. Laroche, S. Larose, H. Maaroufi, J. L. Fothergill, M. Moore, G. L. Winsor, S. D. Aaron, J. Barbeau, S. C. Bell, J. L. Burns, M. Camara, A. Cantin, S. J. Charette, K. Dewar, Déziel, K. Grimwood, R. E. W. Hancock, J. J. Harrison, S. Heeb, L. Jelsbak, B. Jia, D. T. Kenna, T. J. Kidd, J. Klockgether, J. S. Lam, I. L. Lamont, S. Lewenza, N. Loman, F. Malouin, J. Manos, A. G. McArthur, J. McKeown, J. Milot, H. Naghra, D. Nguyen, S. K. Pereira, G. G. Perron, J.-P. Pirnay, P. B. Rainey, S. Rousseau, P. M. Santos, A. Stephenson, V. Taylor, J. F. Turton, N. Waglechner, P. Williams, S. W. Thrane, G. D. Wright, F. S. L. Brinkman, N. P. Tucker, B. Tümmler, C. Winstanley, and R. C. Levesque, "Clinical utilization of genomics data produced by the international Pseudomonas aeruginosa consortium," *Frontiers in Microbiology*, vol. 6, no. 1, p. 1036, 2015.

[21] S. Chen, "fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing," *iMeta*, vol. 4, p. e70078, Oct. 2025.

[22] A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, and A. Korobeynikov, "Using SPAdes De Novo Assembler," *Current Protocols in Bioinformatics*, vol. 70, p. e102, June 2020.

[23] A. Mikheenko, A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich, "Versatile genome assembly evaluation with QUAST-LG," *Bioinformatics*, vol. 34, pp. i142–i150, July 2018.

[24] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, pp. 2068–2069, July 2014.

[25] D. M. Emms, Y. Liu, L. Belcher, J. Holmes, and S. Kelly, "OrthoFinder: scalable phylogenetic orthology inference for comparative genomics," July 2025.

[26] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries," *Nature Communications*, vol. 9, no. 1, 2018.

[27] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, 2015.

[28] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the Tidyverse," *Journal of Open Source Software*, vol. 4, p. 1686, Nov. 2019.

[29] Z. Gu, "Complex heatmap visualization," *iMeta*, vol. 1, no. 3, 2022.

[30] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, "Orchestrating high-throughput genomic analysis with Bioconductor," *Nature Methods*, vol. 12, pp. 115–121, Feb. 2015.

[31] F. Habibzadeh, "Data Distribution: Normal or Abnormal?," *Journal of Korean Medical Science*, vol. 39, no. 3, p. e35, 2024.

[32] A. Hazra and N. Gogtay, "Biostatistics series module 6: Correlation and linear regression," *Indian Journal of Dermatology*, vol. 61, no. 6, p. 593, 2016.

[33] M. V. Villegas, K. Lolans, A. Correa, J. N. Kattan, J. A. Lopez, and J. P. Quinn, "First Identification of *Pseudomonas aeruginosa* Isolates Producing a KPC-Type Carbapenem-Hydrolyzing β-Lactamase," *Antimicrobial Agents and Chemotherapy*, vol. 51, pp. 1553–1555, Apr. 2007.

[34] V. L. Kung, E. A. Ozer, and A. R. Hauser, "The Accessory Genome of Pseudomonas aeruginosa," *Microbiology and Molecular Biology Reviews*, vol. 74, no. 4, pp. 621–641, 2010.

[35] H. P. Schweizer, "Efflux as a mechanism of resistance to antimicrobials in Pseudomonas aeruginosa and related bacteria: unanswered questions," *Genetics and molecular research: GMR*, vol. 2, no. 1, pp. 48–62, 2003.

[36] D. Gevers, K. Vandepoele, C. Simillion, and Y. Van de Peer, "Gene duplication and biased functional retention of paralogs in bacterial genomes," *Trends in Microbiology*, vol. 12, no. 4, pp. 148–154, 2004.

[37] F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Selection in the evolution of gene duplications," *Genome Biology*, vol. 3, p. research0008.1, Jan. 2002.

[38] M. S. Diarra and F. Malouin, "Antibiotics in Canadian poultry productions and anticipated alternatives," *Frontiers in Microbiology*, vol. 5, 2014.

[39] S. Kousar, N. Rehman, A. Javed, A. Hussain, M. Naeem, S. Masood, H. A. Ali, A. Manzoor, A. A. Khan, A. Akrem, F. Iqbal, A. Zulfiqar, M. B. Jamshaid, M. Waqas, A. Waseem, and M. Q. Saeed, "Intensive Poultry Farming Practices Influence Antibiotic Resistance Profiles in Pseudomonas aeruginosa Inhabiting Nearby Soils," *Infection and Drug Resistance*, vol. 14, pp. 4511–4516, 2021.

[40] K. L. Tang, N. P. Caffrey, D. B. Nóbrega, S. C. Cork, P. E. Ronksley, H. W. Barkema, A. J. Polachek, H. Ganshorn, N. Sharma, J. D. Kellner, and W. A. Ghali, "Restricting the use of antibiotics in food-producing animals and its associations with antibiotic resistance in food-producing animals and human beings: a systematic review and meta-analysis," *The Lancet Planetary Health*, vol. 1, pp. e316–e327, Nov. 2017.

Comparison of Random Forest, SVM, and MLP in Facial Emotion Recognition with PCA and ICA

Bruna T. M. Oliveira
Instituto Federal do Ceará (IFCE)
Fortaleza, Brazil
bruna.oliveira07@aluno.ifce.edu.br

Jâmily K. M. Soares
Instituto Federal do Ceará (IFCE)
Fortaleza, Brazil
jamily.soares06@aluno.ifce.edu.br

Josias G. Batista
Instituto Federal do Ceará (IFCE)
Fortaleza, Brazil
josiasbatista@ifce.edu.br

***Abstract*—Automatic facial emotion recognition has emerged as an important research area in Applied Computational Intelligence due to its applications in intelligent systems, human-computer interaction, education and healthcare. This paper compares PCA and ICA associated with Random Forest, SVM and MLP using the FER2013 dataset. Performance was evaluated using accuracy, precision, recall and F1-score. ICA combined with Random Forest achieved the best overall performance, demonstrating the relevance of dimensionality reduction for efficient intelligent systems with potential biomedical applications.**

***Index Terms*—Facial Emotion Recognition, PCA, ICA, Machine Learning, Computer Vision, Biomedical Engineering.**

I. Introduction

Automatic facial emotion recognition has consolidated itself as an important research area in Biomedical Engineering and Applied Computational Intelligence due to its potential to support intelligent patient monitoring systems, mental health assessment, cognitive rehabilitation, telemedicine, and human-computer interaction [1], [2]. The automatic identification of emotional states through facial images allows the development of systems capable of interpreting behavioral information objectively, contributing to applications in clinical, educational, and assistive environments.

With recent advances in Computer Vision and Machine Learning, different approaches have been proposed for automatic facial emotion recognition [10]. Among them, both deep learning-based architectures and classical machine learning methods stand out. Although Convolutional Neural Networks (CNNs) present high performance in various applications, traditional algorithms remain relevant due to lower computational cost, higher interpretability of results, and ease of implementation, especially in scenarios with limited computational resources.

The Facial Expression Recognition 2013 (FER2013) database has become one of the main datasets used in the evaluation of facial emotion recognition algorithms, being widely employed as a reference for comparing classification methods [3]. The dataset consists of approximately 35,000 grayscale facial images distributed into seven emotional categories, constituting a consolidated benchmark for pattern recognition and image processing research.

In addition to the choice of the classification algorithm, the representation of features extracted from images exerts a direct influence on the performance of the models. In this context, dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), are widely used to reduce computational complexity while preserving relevant information for the classification task [7], [8]. Despite the wide use of these techniques, studies that perform a systematic comparison between PCA and ICA simultaneously associated with Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) classifiers using the same experimental protocol on the FER2013 database are still relatively scarce [4]– [6].

Thus, this work aims to comparatively analyze the performance of PCA and ICA dimensionality reduction techniques when combined with Random Forest, Support Vector Machine, and Multilayer Perceptron classifiers for automatic facial emotion recognition using the FER2013 database. The experiments were conducted under the same methodological protocol, employing the same pre-processing steps and evaluation metrics, allowing the analysis of exclusively the influence of the dimensionality reduction technique on the performance of the models. Accuracy, precision, recall, and F1-score metrics were used to evaluate the classifiers.

The results obtained demonstrate that both techniques preserve relevant information for facial emotion classification, with an advantage observed for ICA in the conducted experiments, especially when associated with the Random Forest classifier. These results reinforce the importance of proper selection of the dimensionality reduction technique in the development of intelligent automatic emotion recognition systems with potential application in Biomedical Engineering.

II. Theoretical Background

A. Facial Emotion Recognition

Facial emotion recognition is a research area that combines Computer Vision, Image Processing, and Artificial Intelligence techniques to identify emotional states through human facial expressions, constituting an important application of modern face recognition systems [11]. This technology has been widely applied in human-computer interaction systems, education, healthcare, security, and behavioral analysis. The advancement of machine learning algorithms has significantly contributed to increasing the accuracy of emotional recognition systems [1].

In Biomedical Engineering, automatic facial emotion recognition presents relevant applications in patient monitoring, assessment of neurological and psychiatric disorders, monitoring of therapeutic treatments, telemedicine, and the development of assistive technologies. Objective identification of emotional states can provide complementary information to support clinical decision-making and for the development of intelligent healthcare systems [2].

B. FER2013 Dataset

The FER2013 (Facial Expression Recognition 2013) database is one of the most widely used in research related to automatic facial emotion recognition [3]. The dataset consists of approximately 35,000 grayscale facial images with a resolution of 48×48 pixels, distributed across seven emotional categories: happiness, sadness, anger, fear, surprise, disgust, and neutral. Each image has 2304 attributes corresponding to pixel values, making it necessary to use dimensionality reduction techniques to decrease the computational complexity of the models.

The widespread use of FER2013 as a benchmark enables standardized comparison between different emotion recognition algorithms, promoting experiment reproducibility and objective evaluation of models proposed in the literature.

C. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction [7]. Its objective is to transform a set of correlated variables into a new set of orthogonal components, called principal components. These components are ordered according to the amount of explained variance, allowing the reduction of the number of attributes while preserving most of the information present in the original data.

In addition to reducing computational cost, PCA helps minimize redundancies among attributes and reduce the influence of noise present in images, favoring the performance of classification algorithms in various pattern recognition applications. This characteristic makes the technique particularly useful in problems involving large volumes of data, such as those found in biomedical imaging applications.

D. Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a signal separation technique based on statistical independence between extracted components [8]. Unlike PCA, which seeks to maximize data variance, ICA aims to identify independent sources underlying the observed data. This feature makes the method particularly interesting in pattern recognition and signal processing applications.

In addition to facial emotion recognition applications, ICA-based techniques are also widely employed in processing physiological signals, such as electroencephalograms (EEG) and electrocardiograms (ECG), for source separation and artifact removal. This versatility highlights the technique's potential in different problems related to biomedical signal and image processing.

E. Classifiers Used

In this work, three classification algorithms widely used in pattern recognition problems were evaluated. Random Forest (RF) is a method based on multiple decision trees, capable of reducing overfitting problems and improving generalization capacity [4]. Support Vector Machine (SVM) seeks to find hyperplanes that maximize the separation between different classes, showing good performance in high-dimensional classification problems [5]. Multilayer Perceptron (MLP) consists of an artificial neural network composed of multiple layers of neurons, capable of learning complex relationships between input and output data [6].

These algorithms remain widely used in applications requiring low computational cost, ease of implementation, and greater interpretability of results, which are important characteristics in various intelligent systems applied to Biomedical Engineering.

F. Related Works

Automatic facial emotion recognition has received growing attention from the scientific community due to its applications in human-computer interaction, intelligent systems, healthcare, education, and security. The advancement of Computer Vision and Machine Learning techniques has driven the development of increasingly accurate methods for identifying emotional states from facial images. In this context, the availability of the FER2013 database established one of the main benchmarks for evaluating facial expression recognition algorithms, making standardized comparison between different approaches developed in the literature possible [3].

In early research on automatic facial emotion recognition, classical machine learning techniques associated with feature extraction and reduction methods were widely employed. Among these techniques, Principal Component Analysis (PCA) stands out for reducing data dimensionality while preserving most of the image variance, decreasing computational cost and favoring the performance of different classifiers [7]. Complementarily, Independent Component Analysis (ICA) began to be used to represent data through statistically independent components, enabling the extraction of relevant information for pattern recognition tasks [8].

Among traditional classifiers, Random Forest remains a robust alternative for supervised classification problems. Wang et al. [13] demonstrated that using Random Forest in facial expression recognition systems provides good generalization capability and reduces overfitting problems, especially when combined with efficient feature extraction techniques. Although recent research prioritizes architectures based on deep learning, Random Forest continues to be employed due to its low computational cost, ease of interpretation, and good generalization capacity.

Support Vector Machine (SVM) also ranks among the most used algorithms in facial emotion recognition. Jayaswal et al. [9] highlight that this classifier performs well on high-dimensional datasets when associated with dimensionality

reduction techniques, remaining one of the primary traditional approaches used in facial expression recognition systems.

Another widely used approach is the Multilayer Perceptron (MLP). Qutab et al. [15] proposed an emotion recognition method based on the integration of PCA, Recursive Feature Elimination (RFE), and MLP, obtaining high accuracy rates on the CK and JAFFE databases. The results demonstrated that dimensionality reduction contributes to reducing training computational costs and improving facial feature representation.

Recent reviews indicate that despite the growth of Deep Learning-based architectures, traditional methods continue to play a relevant role in applications requiring lower computational cost, higher interpretability, and simplified implementation [9], [10]. In particular, Jayaswal et al. [9] highlight that techniques such as PCA, ICA, SVM, and Random Forest remain widely used in experimental research and constitute important references for comparison with Deep Learning models.

However, it is observed that most recent research focuses on proposing deep architectures, while systematic comparisons between classical dimensionality reduction techniques simultaneously associated with Random Forest, SVM, and MLP classifiers using the FER2013 database are still relatively scarce.

In this context, the present work seeks to contribute to this gap by comparing, under the same experimental protocol, the PCA and ICA dimensionality reduction techniques associated with Random Forest, Support Vector Machine, and Multilayer Perceptron classifiers. In addition to providing a consistent comparative analysis, the obtained results can subsidize the development of intelligent systems for automatic emotion recognition, contributing to applications in Biomedical Engineering that demand computationally efficient, interpretable, and reproducible models.

III. Methodology

A. Dataset

To perform the experiments, the FER2013 (Facial Expression Recognition 2013) database, made available on the Kaggle platform [16], was used. The dataset consists of approximately 35,000 grayscale facial images with a resolution of 48×48 pixels, distributed across seven emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. For the development of this work, the database was divided into training and testing subsets, allowing the performance evaluation of the studied classifiers.

Using the FER2013 database enables direct comparison of the obtained results with previously published works, since it is one of the main benchmarks employed in automatic facial emotion recognition research.

B. Computational Tools

The experiments were developed using the Python programming language on the Google Colab platform, an environment widely used in research involving machine learning due to the availability of cloud computing resources. To implement the methodology, the NumPy and Pandas libraries were used for data manipulation, Matplotlib for graphical visualization of the results, Scikit-Learn [12] for implementing the machine learning algorithms, and OpenCV for reading, processing, and preparing the facial images used in the experiments.

Using these libraries, which are widely consolidated in the scientific community, contributes to the reproducibility of the proposed methodology and facilitates the replication of experiments in different computing environments.

C. Pre-processing

Initially, the images from the FER2013 database were subjected to pre-processing stages, aiming to distinguish and standardize the data before applying the machine learning algorithms. The images were organized into numerical matrices, normalized, and prepared for computational processing. Then, the dataset was divided into training and testing subsets, enabling the evaluation of model performance on data not used during training.

This step aimed to reduce undesired variations between images and provide a standardized representation of the data, contributing to greater stability in the training and evaluation process of the classifiers. The image preparation and organization procedures followed classical principles of Digital Image Processing [14].

D. Dimensionality Reduction

After pre-processing, two dimensionality reduction techniques were applied: Principal Component Analysis (PCA) and Independent Component Analysis (ICA). These techniques were used to reduce the dimensionality of the data present in the FER2013 database, decreasing the computational complexity of the models while preserving the most relevant information for the facial emotion classification task. The use of PCA and ICA also made it possible to compare the impact of different data representation strategies on the performance of the evaluated classifiers.

The comparison between PCA and ICA was conducted while maintaining the same experimental conditions for both methods, allowing the analysis of exclusively the influence of the dimensionality reduction technique on the performance of the classifiers used in this study.

E. Classifiers Used

After dimensionality reduction, the obtained attributes were used to train and evaluate three machine learning algorithms widely used in supervised classification problems: Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). All classifiers were trained using the same datasets, allowing a direct comparison between models when associated with PCA and ICA dimensionality reduction techniques. The comparison between classifiers was performed under the same experimental conditions, allowing the analysis of exclusively the influence of dimensionality reduction techniques on the performance of each model.

Using a standardized experimental protocol guaranteed equivalent conditions for all evaluated classifiers, ensuring greater reliability in the comparison of the obtained results.

F. Methodological Flowchart

Fig. 1 presents the flowchart of the methodology proposed in this work, illustrating the main stages of the automatic facial emotion recognition process, from dataset acquisition to classifier performance evaluation.

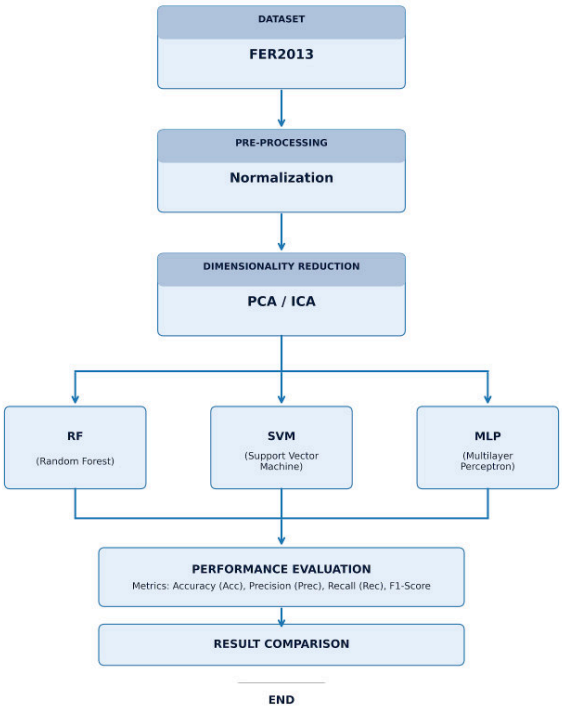


Fig. 1. Flowchart of the facial emotion classification process.

Initially, the images from the FER2013 database were subjected to pre-processing steps. Then, PCA and ICA techniques were applied to reduce data dimensionality. Subsequently, the extracted attributes were used in training the Random Forest, SVM, and MLP classifiers. Finally, the performance of the models was evaluated using accuracy, precision, recall, and F1-score metrics.

The adoption of this methodological flow sought to guarantee the standardization of the experimental stages, favoring the reproducibility of the study and the consistent comparison between the different evaluated models.

G. Evaluation Metrics

The performance of the classifiers was evaluated through accuracy, precision, recall, and F1-score metrics, which are widely used in supervised classification problems to analyze the ability of models to correctly identify the facial emotions present in the FER2013 database.

Accuracy represents the total proportion of correct classifications made by the model relative to the total number of evaluated samples. Precision indicates the proportion of correctly classified samples within each predicted class. Recall, also called sensitivity, measures the classifier's ability to correctly identify samples belonging to each emotional category. In turn, the F1-score corresponds to the harmonic mean between precision and recall, being a metric widely used to evaluate the balance between these two measures, especially in multiclass classification problems.

The joint analysis of these metrics enables a more comprehensive evaluation of classifier performance, allowing a comparison of the impact of dimensionality reduction techniques (PCA and ICA) on the Random Forest, Support Vector Machine, and Multilayer Perceptron algorithms employed in this work. Using multiple metrics provides a more consistent analysis of model behavior, reducing dependence on a single performance measure and allowing a more complete evaluation in multiclass classification problems.

IV. Results and Discussion

In this section, the results obtained in the facial emotion recognition experiments using the FER2013 dataset are presented and discussed. The models were evaluated by combining dimensionality reduction techniques (PCA and ICA) with three classification algorithms: Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The quantitative performance of each approach is summarized in Table I.

Using a standardized experimental protocol for all models allowed a direct comparison between the dimensionality reduction techniques and the evaluated classifiers, ensuring greater methodological consistency and reliability in interpreting the obtained results.

TABLE I
Performance of Classification Models on the FER2013 Dataset

Model	Reduction	Accuracy	Precision	Recall	F1-Score
Random Forest	ICA	0.427029	0.441716	0.427029	0.400523
Random Forest	PCA	0.406304	0.426830	0.406304	0.378584
MLP	ICA	0.399861	0.392345	0.399861	0.395080
MLP	PCA	0.390630	0.386177	0.390630	0.388032
SVM	ICA	0.366249	0.330730	0.366249	0.316851
SVM	PCA	0.366249	0.331972	0.366249	0.317710

The graphical comparison presented in Fig. 2 complements the results of Table I, allowing a clearer visualization of the performance differences between classifiers when associated with PCA and ICA techniques.

The comparative analysis of the results presented in Table I demonstrates that the Random Forest classifier achieved the best performance among the evaluated models. The combination of Random Forest and ICA presented the highest accuracy (42.7%), followed by the combination of Random Forest and PCA (40.6%). The results indicate that both dimensionality reduction techniques were able to preserve relevant features for the facial emotion recognition task, although ICA showed a slight advantage in the conducted experiments. In contrast,

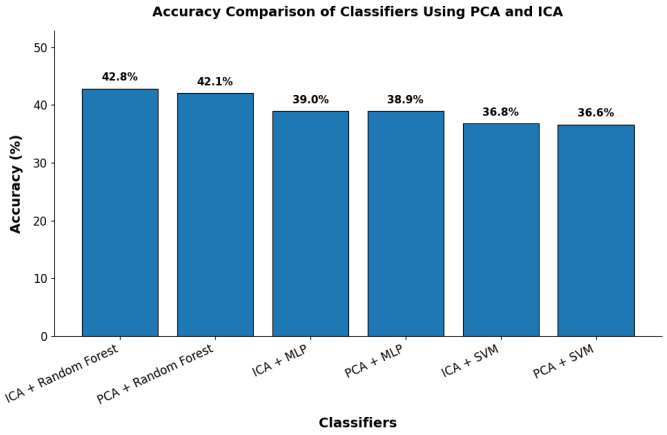


Fig. 2. Bar chart showing the comparison of classifier accuracy using PCA and ICA.

the lowest performance values were observed in models based on SVM, regardless of the dimensionality reduction technique used.

The obtained results indicate that the choice of dimensionality reduction technique directly influences classifier performance, reinforcing that the data representation step is a determining factor for automatic facial emotion recognition.

The difference observed between the results obtained with PCA and ICA may be related to the specific characteristics of the FER2013 dataset and how each technique represents the information present in the facial images. While PCA seeks to represent the data through components that explain most of the observed variance, ICA tries to identify statistically independent components. In the conducted experiments, both approaches showed similar performance, with a slight advantage for ICA in some classifier combinations.

This behavior suggests that the representation based on independent components preserves more discriminative features of facial expressions, favoring the classification stage without a significant increase in computational complexity. This result reinforces ICA's potential in applications that require computationally efficient models and good generalization capacity.

Fig. 2 shows that ICA-based models performed equal to or better than their PCA-based counterparts in most evaluated scenarios. The most significant difference was observed in the Random Forest classifier, whose accuracy went from 40.6% with PCA to 42.7% with ICA. Although the observed percentage gain was relatively modest, its consistency across different evaluation metrics indicates that ICA constitutes a promising alternative for automatic facial emotion recognition applications and intelligent systems geared toward Biomedical Engineering.

In addition to the general comparison between classifiers, a more detailed analysis of the model's behavior was performed using the confusion matrix. This representation allows the identification of correctly classified emotions and those that presented a higher degree of confusion during the classification process. Although the Random Forest combination associated with ICA achieved the best overall performance, the confusion matrix of the Random Forest model associated with PCA was used to illustrate the classifier's behavior regarding the different emotional classes of the FER2013 database, allowing a detailed analysis of correct matches and confusions among the evaluated categories.

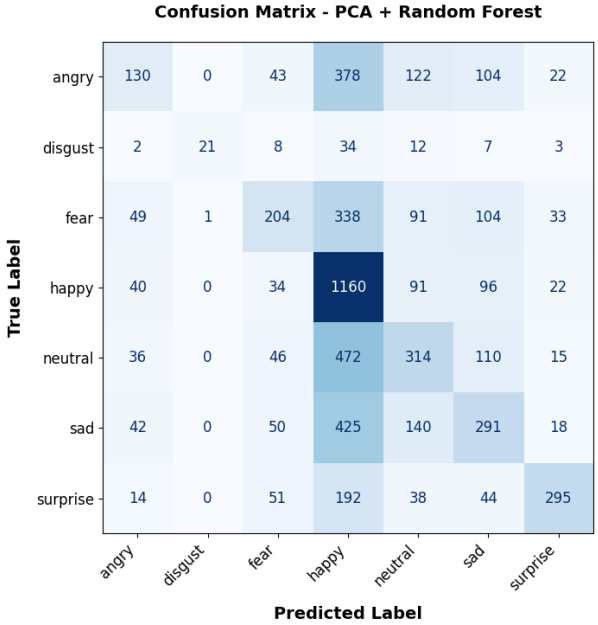


Fig. 3. Confusion matrix obtained by the PCA + Random Forest model in classifying the emotions of the FER2013 dataset.

The confusion matrix presented in Fig. 3 allows a more detailed analysis of the behavior of the Random Forest model associated with PCA. Although this combination did not yield the best overall performance, its analysis allows an understanding of the classifier's behavior when facing the different emotional classes present in the FER2013 database. It is observed that certain emotions were classified with high precision, while others presented a higher degree of confusion, especially between classes with similar facial features. This behavior is expected in facial emotion recognition databases in non-controlled environments, where different emotional states can share similar visual patterns.

This behavior highlights the inherent difficulty in classifying emotions with similar facial patterns, representing a recurring challenge in computer vision-based emotion recognition systems.

In addition to the global analysis of model performance, an individual evaluation of each emotional class was conducted, and the results are presented in Table II. This analysis allows the identification of which emotions were classified with higher or lower precision by the Random Forest model associated with PCA.

It is observed that the "happy" emotion presented the highest recall value (0.8018), indicating greater ease of identification by the model. In contrast, emotions such as "fear" and "sad" presented lower recall rates, highlighting the greater

TABLE II
CLASSIFICATION METRICS PER EMOTION OBTAINED BY THE PCA + RANDOM FOREST MODEL

Emotion	Precision	Recall	F1-Score	Support
angry	0.405316	0.152691	0.221818	799.00
disgust	1.000000	0.275862	0.432432	87.00
fear	0.407489	0.225610	0.290424	820.00
happy	0.380592	0.801802	0.516172	1443.00
neutral	0.367725	0.279960	0.317896	993.00
sad	0.345628	0.261905	0.297998	966.00
surprise	0.721839	0.495268	0.587465	634.00
accuracy	0.406304	0.406304	0.406304	0.406304
macro avg	0.518370	0.356157	0.380601	5742.00
weighted avg	0.426830	0.406304	0.378584	5742.00

difficulty in discriminating between categories with similar facial characteristics.

The results presented in Table II show differences in model performance across emotional classes. Emotions with more distinct facial characteristics tend to show higher precision and recall values, while visually similar emotions show higher rates of confusion.

This analysis demonstrates that the model's performance depends not only on the classification algorithm employed but also on the inherent complexity of each emotional category present in the dataset, showing that visually similar emotions tend to produce higher confusion rates.

Among the evaluated classifiers, Random Forest stood out against SVM and MLP. Being a method based on an ensemble of decision trees, Random Forest demonstrated greater resilience to overfitting and a better ability to map the feature space reduced by PCA. SVM showed an acceptable intermediate performance, but its efficiency was slightly penalized by the complexity of defining optimal hyperplanes in highly overlapping multiclass problems. Finally, MLP obtained the lowest accuracy rates; this behavior suggests that fully connected networks of moderate size face difficulties in converging optimally when fed purely with components extracted without a dedicated spatial architecture, such as convolutional ones.

In general, the results obtained demonstrate that classical machine learning algorithms, when associated with appropriate dimensionality reduction techniques, continue to represent viable alternatives for automatic facial emotion recognition problems, especially in applications that demand lower computational cost, higher interpretability, and ease of implementation. Furthermore, the presented results reinforce that the dimensionality reduction stage exerts a significant influence on the performance of the classifiers, with ICA being the approach that yielded the best results in the conducted experiments. These findings highlight the potential of the proposed methodology for applications in Biomedical Engineering, such as intelligent patient monitoring systems, behavioral analysis, assistive technologies, and clinical decision-making support by healthcare professionals, in which computational efficiency, reliability, and reproducibility are fundamental requirements.

V. CONCLUSION

This work presented a comparative analysis between the dimensionality reduction techniques Principal Component Analysis (PCA) and Independent Component Analysis (ICA) applied to automatic facial emotion recognition using the FER2013 dataset. For emotion classification, the Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) algorithms were evaluated, considering accuracy, precision, recall, and F1-score metrics.

The obtained results demonstrated that the Random Forest classifier achieved the best performance among the evaluated models, standing out especially when associated with the ICA technique. In general, it was observed that both dimensionality reduction techniques preserved relevant information for facial emotion classification, with ICA being slightly superior in some classifier combinations.

The results evidenced that the dimensionality reduction step exerts a significant influence on the performance of classification models, reinforcing the importance of choosing appropriate data representation techniques in the development of intelligent systems for automatic emotion recognition. The systematic comparison between PCA and ICA, associated with Random Forest, Support Vector Machine, and Multilayer Perceptron classifiers under the same experimental protocol, allowed a consistent evaluation of the impact of these techniques on model performance, constituting one of the main contributions of this study.

In addition to the contribution to the area of Applied Computational Intelligence, the proposed methodology presents potential for applications in Biomedical Engineering, especially in intelligent systems aimed at patient monitoring, behavioral analysis, emotional state evaluation, assistive technologies, and decision-making support for healthcare professionals. The use of classical machine learning methods associated with dimensionality reduction techniques also demonstrates that computationally efficient solutions remain relevant in scenarios that demand lower computational cost, interpretability, and ease of implementation.

As limitations of this study, we highlight the use of the FER2013 database, which consists of low-resolution images, as well as the evaluation restricted to classical machine learning methods. Furthermore, different hyperparameter tuning settings for the classifiers were not investigated, nor were other feature extraction strategies, factors that can directly influence model performance.

As future work, it is suggested to investigate the influence of the number of components extracted by dimensionality reduction techniques on classifier performance, seeking to determine an optimal cutoff point. It is also proposed to evaluate other facial emotion recognition databases to analyze the generalization capability of the methodology across different experimental scenarios.

Additionally, it is intended to explore hybrid pre-processing and feature selection techniques, as well as to compare the classical methods evaluated in this work with Deep Learning architectures, especially Convolutional Neural Networks

(CNNs) and Vision Transformer (ViT) models, investigating the trade-off between performance, computational cost, and interpretability.

Finally, the obtained results reinforce that classical dimensionality reduction techniques, when properly combined with machine learning algorithms, continue to represent effective alternatives for automatic facial emotion recognition. Thus, this work contributes to broadening the understanding of the impact of PCA and ICA techniques on the performance of traditional classifiers, providing insights for the development of intelligent systems applied to emotion recognition and future applications in Biomedical Engineering.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[3] I. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2009.

[7] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: John Wiley & Sons, 2001.

[9] R. Jayaswal et al., "Advances in Facial Expression Recognition Technologies for Affective Computing: A Comprehensive Review," *Journal of Intelligent & Fuzzy Systems*, 2025.

[10] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[11] S. Z. Li and A. K. Jain (Eds.), *Handbook of Face Recognition*, 2nd ed. London, U.K.: Springer, 2011.

[12] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] Y. Wang et al., "Facial Expression Recognition Based on Random Forest and Convolutional Neural Network," *Information*, vol. 10, no. 12, Art. no. 375, 2019.

[14] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.

[15] M. A. Qutab et al., "Advancing Emotion Recognition in Facial Expressions through PCA, RFE and MLP Integration," *International Journal of Science and Research Archive*, vol. 13, no. 1, pp. 1954–1964, 2024.

[16] FER2013 Dataset. Kaggle. Available: https://www.kaggle.com/datasets/msambare/fer2013 (accessed Jun. 26, 2026).

Development and Functional Evaluation of a Computational Module for Processing and Analysis of Nerve Conduction Signals Integrated into the SDC-X Platform

João Lucas Pereira dos Santos de Paula
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0000-0001-8230-5358

Júlio Cardoso Scorpion
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0009-0002-8669-2249

Jefferson Luiz Brum Marques
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0000-0001-8910-4735

***Abstract*—Diabetic peripheral neuropathy (DPN) is a common chronic complication of diabetes mellitus (DM). Nerve conduction studies provide objective measurements of peripheral nerve function, but their integration into modular health platforms also requires software for examination management, signal processing, parameter extraction, and organized data storage. This work presents X-NCS, a Python/PyQt5 software module integrated into the SDC-X platform for research-oriented nerve conduction workflows. X-NCS combines patient and session management, asynchronous WebSocket communication with an experimental acquisition unit, signal visualization, digital filtering, automatic and manual annotation of the stimulation artifact and compound muscle action potential (CMAP) landmarks, electrophysiological parameter extraction, and structured local storage. Functional evaluation used motor median nerve recordings from 14 participants without DM. Completion was defined as successful execution of data loading, visualization, filtering, landmark generation and review, extraction of distal latency, amplitude, duration, area, and motor nerve conduction velocity (MNCV), and local storage without a processing error that prevented completion. X-NCS completed this workflow for all 14 datasets, and no manual repositioning of the automatically generated landmarks was required during the evaluated workflow. Mean (SD) amplitude was 7.6 (1.9) mV, distal latency 4.2 (0.9) ms, and MNCV 62.8 (10.0) m/s; these values were used as a physiological plausibility check against published reference ranges. Because the study did not include an independent expert annotation set or a reference nerve-conduction system, the absence of manual corrections does not establish annotation accuracy or measurement agreement. Communication-failure robustness and diagnostic validity were also outside the scope of this functional evaluation. X-NCS provides a research-oriented computational layer for organizing and processing nerve conduction examinations within the SDC-X ecosystem.**

***Index Terms*—Diabetic Peripheral Neuropathy; Nerve Conduction Study; Biomedical Software; SDC-X Platform.**

This study was financed in part by CAPES – Finance Code 001. Human-participant procedures were approved by the UFSC Human Research Ethics Committee (CAAE 74863023.7.0000.0121; approval No. 6.450.185), with written informed consent.

I. Introduction

Diabetes mellitus (DM) represents one of the leading global burdens of chronic disease. The 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas estimated that, in 2024, approximately 589 million adults aged 20–79 years were living with diabetes, with this number projected to reach 853 million by 2050 [1]. Among the chronic complications of DM, diabetic peripheral neuropathy (DPN) is one of the most common and clinically significant, affecting approximately half of all adults with diabetes during their lifetime and being associated with pain, sensory loss, foot ulcers, and lower-limb amputations [2], [3].

DPN is characterized by the progressive dysfunction of peripheral nerves, predominantly affecting distal sensory and motor fibers, thereby directly impairing patients' functional capacity and quality of life [3].

Early screening is essential to reduce the risk of long-term complications. The American Diabetes Association guidelines recommend screening for DPN at the time of diagnosis in patients with type 2 diabetes and after five years of disease duration in patients with type 1 diabetes, followed by periodic reassessment [4]. Clinical examination, sensory testing, and nerve conduction studies (NCS) are used in a complementary manner. NCS provide objective measurements of peripheral nerve function; however, their use for large-scale screening is limited by the need for specialized equipment, standardized procedures, and trained personnel [5]–[7].

Motor NCS assess the functional integrity of peripheral motor fibers through the compound muscle action potential (CMAP), from which latency, amplitude, duration, area, and conduction-related measurements can be obtained. The parameters used in X-NCS are described in Section II. Previous work has investigated automated CMAP acquisition and event annotation because manual cursor placement can contribute to measurement variability, especially when waveform morphology or stimulus artifacts make the landmarks less evident [8],

[9].

Beyond the acquisition hardware, integration with a digital platform also requires software that keeps the examination steps and their data together. In practice, this includes patient and session management, recording of metadata, signal visualization, digital filtering, parameter calculation, and structured local storage of recordings, extracted parameters, metadata, and timestamps. These functions are compatible with the broader organization of Ubiquitous Health (U-Health) platforms, in which sensing, computing resources, and data infrastructure are combined to support health monitoring, disease prevention, and patient follow-up [10]–[12].

The Screening for Diabetes Complications (SDC-X) platform, developed at the Biomedical Engineering Institute of the Federal University of Santa Catarina (IEB-UFSC), is being structured into modular components for the acquisition, organization, and analysis of physiological signals associated with the screening of DM complications [13], [14]. Within this architecture, the X-CARDIO, X-EYE, X-ESC, and X-QST modules are dedicated to cardiovascular autonomic assessment, pupillometry analysis, sudomotor assessment, and quantitative sensory testing based on thermal thresholds, respectively [14]–[18]. X-NCS was developed as an additional module so that nerve-conduction recordings could follow the same general SDC-X organization of patient selection, examination setup, signal review, parameter extraction, and result storage.

The physical acquisition and stimulation unit served as the experimental source of the nerve-conduction recordings. This study focuses on the X-NCS software layer rather than on metrological validation of the acquisition hardware. The evaluation therefore addresses integration with the SDC-X workflow, acquisition and analysis interfaces, filtering and CMAP-annotation procedures, electrophysiological parameter extraction, local organization of examination data, and execution of the complete workflow with real motor median nerve recordings from 14 volunteers.

A. Related Work

Conventional NCS remain the reference standard for the objective electrophysiological assessment of myelinated peripheral nerve fibers. However, their widespread use for screening may be limited by the need for specialized infrastructure, trained personnel, and rigorous standardization of both signal acquisition and event annotation [19], [20].

Among these approaches, automated devices for sural nerve assessment, such as NC-stat/DPNCheck, have received considerable attention. Early studies demonstrated a high correlation with conventional NCS and clinically relevant diagnostic performance for the detection of diabetic polyneuropathy [21], [22]. More recent studies, however, suggest that although these devices may provide objective and accessible measurements, their use for longitudinal monitoring and their agreement with different screening tools still require careful interpretation [23], [24]. In addition, simplified approaches that combine point-of-care nerve conduction measurements with other physiological signals, such as electrocardiogram (ECG)-derived parameters, have been proposed to improve the accessibility of electrophysiological assessment in individuals with diabetes [25].

II. Methodology

A. Scope and Software Module Architecture

The X-NCS module was developed to support research-oriented NCS workflows within the SDC-X platform. For the experiments presented in this study, the software was connected to an experimental acquisition and stimulation unit developed at IEB-UFSC, operating at a sampling rate of 30 kHz using an analog-to-digital converter and an ESP32 microcontroller. This sampling rate is compatible with the frequency range of motor nerve signals, which has been reported in the literature to extend from approximately 2 Hz to 10 kHz [20].

X-NCS was organized into three main interfaces that correspond to the stages followed during an examination: *Acquisition*, *Exam*, and *Results*. The first interface is used for examination setup and signal acquisition; the second is used to inspect the waveform, apply processing, and review the CMAP annotations; and the third consolidates the electrophysiological parameters, including motor nerve conduction velocity (MNCV). Table I summarizes the functionality available at each stage, and Fig. 1 shows how these parts are connected within the module.

The graphical interfaces were implemented in Python using PyQt5. Communication with the acquisition device uses WebSocket over the Transmission Control Protocol (TCP). This communication runs in a thread separate from the graphical user interface event loop; the separation was adopted so that data exchange with the acquisition unit would not block the interface while the user is acquiring, playing back, or inspecting a recording.

The communication path was used during the acquisitions analyzed in this study. However, communication faults were not deliberately introduced into the protocol. The available data therefore do not characterize software behavior after WebSocket interruption, reconnection, timeout, or incomplete transmission. These conditions remain specific verification tests for a subsequent robustness study.

B. Integration with the SDC-X Platform

The X-NCS module was incorporated as a new computational protocol within the SDC-X ecosystem, which already includes modules dedicated to the screening of DM complications [13], [14], [18]. The home screen allows patient registration and selection, while the examination window includes X-NCS as an available protocol. The version evaluated here runs locally and is not synchronized with the SDC-X cloud infrastructure. Cloud synchronization, role-based access control, and distributed audit mechanisms are therefore requirements for a future connected deployment rather than features of the prototype evaluated in this study.

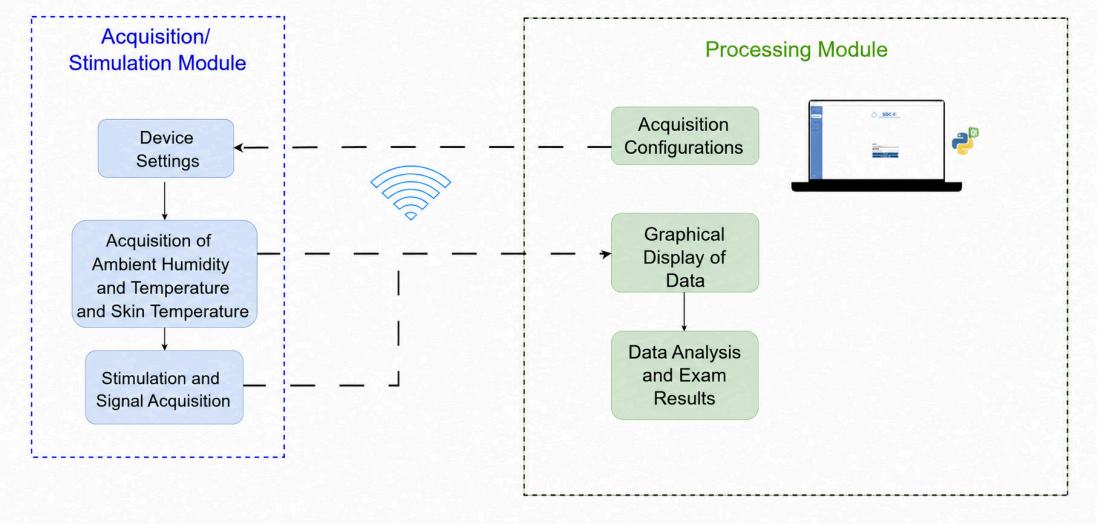


Fig. 1. Overall architecture of the X-NCS module integrated into the SDC-X platform. The physical acquisition unit provides the electrophysiological signals, while the software module manages the examination workflow, processes the recordings, and stores the results locally.

TABLE I
FUNCTIONALITIES IMPLEMENTED IN THE X-NCS MODULE.

Category	Implemented functionality
Management	Patient registration and selection, X-NCS protocol selection, and organization of recordings by patient, nerve, and examination session.
Communication	WebSocket communication over TCP implemented in a dedicated thread to prevent blocking of the graphical user interface.
Acquisition	Session configuration, metadata acquisition, monitoring of signal acquisition, and real-time signal visualization.
Processing	Sampling frequency estimation, digital filtering, and waveform inspection.
Annotation	Automatic and manual annotation of the stimulation artifact and CMAP landmarks: stimulation artifact, onset, peak, and end.
Results	Extraction of distal latency, amplitude, duration, area, and MNCV, together with storage of time-series data, extracted parameters, and metadata.

C. Acquisition and Visualization Workflow

Within the *Acquisition* interface, the examination workflow follows a guided sequence consisting of device connection, examination configuration, recording of environmental variables (temperature and humidity), signal acquisition, and file storage. Signals are saved as text files containing time (ms) and amplitude (mV) columns, together with summary files storing session metadata, including stimulation distances, environmental measurements, and timestamps.

The current implementation uses structured file-based storage rather than a database backend. Accordingly, no relational database schema is defined in the evaluated prototype. Its persistent data model consists of signal files and session-summary files organized by patient, nerve, and examination session. These files retain the signal time series, extracted results, examination metadata, stimulation distances, environmental measurements, and timestamps. This structure provides record-level traceability within the local file hierarchy, although it is not a user-level audit trail. Signal-processing timing is derived from the time vector stored with each recording, while session timestamps are used for record organization. Because the evaluated workflow did not require independent devices to share a common clock, distributed multi-device synchronization was not assessed.

In this study, S1 denotes the distal stimulation site and S2 the proximal stimulation site, following the motor NCS protocol used to calculate MNCV [19], [20]. Fig. 2 summarizes the operational workflow, from patient selection and examination setup to parameter extraction and result storage.

D. Signal Processing Pipeline and Parameter Extraction

Within the *Exam* interface, the sampling frequency is estimated from the average temporal interval between consecutive samples, allowing the module to process both recordings acquired with the experimental unit and imported comma-separated values (CSV) or plain-text (TXT) files.

The default preprocessing pipeline consists of a sequence of second-order digital filters applied using zero-phase bidirectional filtering: (i) a Butterworth high-pass filter (cutoff frequency ≈ 2 Hz) to remove baseline drift and low-frequency oscillations; (ii) notch filters centered at 60 Hz and 120 Hz, with a bandwidth of 5 Hz, to suppress power-line interference and its second harmonic; and (iii) a Butterworth low-pass filter with a cutoff frequency of 10 kHz to remove spectral components outside the frequency band of interest for motor nerve signals [20]. Bidirectional filtering was adopted to preserve waveform timing relationships, which are critical for accurate latency estimation.

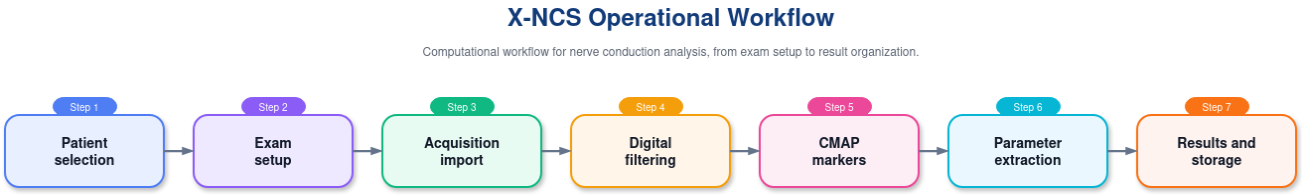


Fig. 2. Operational workflow of the X-NCS module, including patient selection, protocol selection, examination setup, signal acquisition, filtering, CMAP annotation, parameter extraction, and result storage.

Signal analysis is performed over a 100-ms window starting from the beginning of each recording. The software applies four operational landmark rules: (i) the stimulation-artifact marker is assigned to the maximum absolute peak occurring between 8 and 11 ms after the beginning of the recording; this interval corresponds to the expected temporal position of the electrical stimulation artifact in the acquisition protocol and does not define the CMAP response window; (ii) the operational CMAP-onset marker is assigned to the first baseline crossing occurring at least 1 ms after the detected stimulation artifact and is refined by linear interpolation; (iii) the CMAP peak is the sample with the maximum absolute amplitude after onset; and (iv) the CMAP-end marker is the subsequent baseline crossing after the peak, delimiting the negative phase used by the implemented algorithm [19], [26]. Thus, the 8–11 ms search interval is used only to localize the stimulation artifact generated by the acquisition protocol and should not be interpreted as a physiological latency interval or as the expected location of the CMAP itself.

These landmarks are used to calculate distal latency, amplitude, duration, area, and MNCV. In the implemented pipeline, the time of stimulation is operationally represented by the detected stimulation-artifact marker; distal latency is therefore calculated as the interval between this marker and the CMAP onset. Amplitude is defined as the magnitude of the maximum negative deflection relative to baseline. Duration corresponds to the interval between CMAP onset and end. Area is calculated as the absolute value of the numerical integral of the negative CMAP phase between onset and end, expressed in mV·ms. All amplitude-dependent parameters are calculated from samples converted back to the original signal scale after digitization by compensating for the system gain. MNCV is calculated as

$$\mathrm{MNCV} = \frac{\Delta d}{\Delta t} = \frac{d_{S2} - d_{S1}}{t_p - t_d}, \tag{1}$$

where $d_{S2} - d_{S1}$ is the distance between the proximal and distal stimulation sites, and $t_p - t_d$ is the difference between proximal and distal latencies [19], [20].

In motor NCS, S1 provides the measurements of amplitude, latency, duration, and area, whereas S2 is used for MNCV calculation. The user may choose either automatic processing or manual editing mode, allowing individual markers to be removed, restored, or repositioned. The software interface resulting from this implementation is presented in Section III-A.

E. Results Consolidation

Within the *Results* interface, the extracted electrophysiological parameters are presented in tables and graphical summaries. Whenever available, age- and sex-specific normative reference values are used; otherwise, general reference ranges reported in the literature are adopted for exploratory interpretation [20], [27]. When a normative mean and standard deviation are available, the standardized value displayed for each parameter is calculated as $z = (x - \mu_{\mathrm{ref}})/\sigma_{\mathrm{ref}}$, where x is the measured value and μ_{ref} and σ_{ref} are the corresponding reference mean and standard deviation. Dashed horizontal lines at $z = -2$ and $z = +2$ delimit the exploratory reference region. This visualization follows the strategy previously adopted in the X-QST module of the SDC-X platform [17]. The interface also displays the amplitude-versus-conduction velocity plane together with environmental variables, when available. Since skin temperature, age, and other physiological factors may influence nerve conduction parameters, normative comparisons should be interpreted with caution [26]–[28]. The implemented interface is presented with the software results in Section III-A.

F. Functional Evaluation Using Real Recordings

The functional evaluation was performed using real motor median nerve recordings obtained in a pilot study involving 14 participants without a diagnosis of DM (9 males and 5 females). Signals were acquired at a sampling rate of 30 kHz with a system gain factor of 6. After digitization, the recorded samples were rescaled to the original signal amplitude by compensating for this gain, and the resulting values were expressed in millivolts before being stored as text files and imported into the X-NCS module for processing.

The experimental protocol employed a surface stimulation electrode with saline-moistened felt, an active recording electrode positioned over the abductor pollicis brevis muscle, and a reference electrode placed on the thumb. The S1 and S2 stimulation sites were located at the anterior wrist and medial upper arm, respectively, following the conventional distal and proximal stimulation protocol for motor median NCS [19], [20]. The distance between the S1 stimulation site and the active recording electrode was approximately 8 cm. Fig. 3 illustrates the experimental setup used for acquisition of the motor median nerve recordings.

The goal of this evaluation was narrower than clinical or metrological validation: it was to determine whether X-NCS

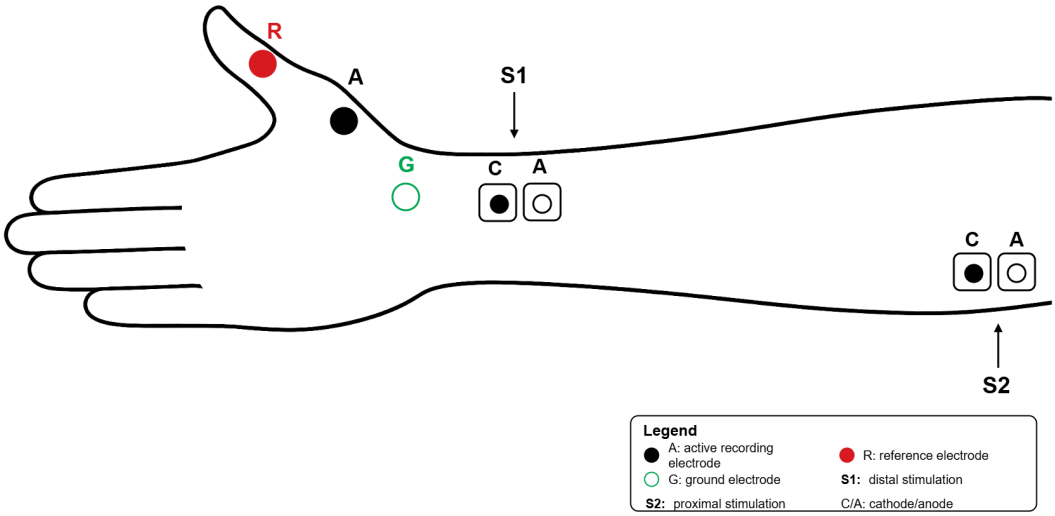


Fig. 3. Simplified experimental setup used for motor median nerve acquisition. The active electrode (A) was positioned over the thenar eminence, the reference electrode (R) on the thumb, and the ground electrode (G) between the recording site and the stimulation sites. S1 and S2 represent the distal and proximal stimulation sites, respectively; the distance between S1 and the active recording electrode was approximately 8 cm.

could execute the implemented computational workflow with real electrophysiological recordings. The evaluation checkpoints followed the software functions summarized in Table I and the operational sequence shown in Fig. 2. Functional completion was defined as successful data loading, waveform visualization, digital filtering, CMAP landmark generation and review, extraction of latency, amplitude, duration, area, and MNCV, and local storage of the resulting data and metadata without an error that prevented completion of the pipeline. The protocol did not include comparison with a commercial nerve-conduction system, independent expert annotation, deliberate communication-failure testing, or metrological characterization of the acquisition hardware. Published normative values were used only to examine whether the extracted parameters were physiologically plausible for the analyzed sample [20], [27].

G. Ethical Aspects

The pilot study is part of the project entitled *Methodologies for the Assessment of Autonomic Function Applied to the Prevention, Diagnosis, and Treatment of Diabetes Mellitus*, approved by the Human Research Ethics Committee of UFSC under CAAE 74863023.7.0000.0121 and approval report No. 6.450.185. Ethical approval was granted on November 5, 2023. Written informed consent was obtained from all participants before the study procedures.

The evaluated implementation operated locally. Deployment-level privacy controls, role-based access control, and user-level audit logging were therefore not exercised in this study. These mechanisms will need to be specified and verified before a future cloud-connected X-NCS version is used beyond the local research setting.

III. Results and Discussion

A. Implemented Software Module

The X-NCS module adds a dedicated NCS workflow to the SDC-X platform. It supports patient selection, session setup, communication with the experimental acquisition unit, signal visualization, digital filtering, parameter extraction, and structured local storage. By separating asynchronous communication from the graphical interface, the software remained responsive during both signal acquisition and playback.

Fig. 4 presents the implemented *Exam* interface, in which the user can inspect the filtered waveform, review the automatically detected stimulation-artifact marker and CMAP landmarks, and manually correct individual annotations when necessary. Fig. 5 presents the *Results* interface, which consolidates the extracted electrophysiological parameters and provides an exploratory Z-score visualization relative to normative reference values. The dashed lines identify the $z = -2$ and $z = +2$ limits used for contextualization; the display does not constitute an independent diagnostic classification.

B. Functional Evaluation of the Pipeline

Using the functional-completion definition given in Section II-F, the module completed the full workflow for all 14 participant datasets, corresponding to 14/14 (100%) completed datasets in the analyzed sample. No software-processing error prevented completion of the pipeline for these files. This result applies only to the available recordings and does not characterize failure behavior under deliberately induced communication faults. For each dataset, the software loaded the distal and proximal recordings, displayed the waveforms, applied the digital filters, identified the stimulation artifact and generated the subsequent CMAP landmarks, extracted latency, amplitude, duration, area, and MNCV, and stored the time series, parameters, and metadata locally.

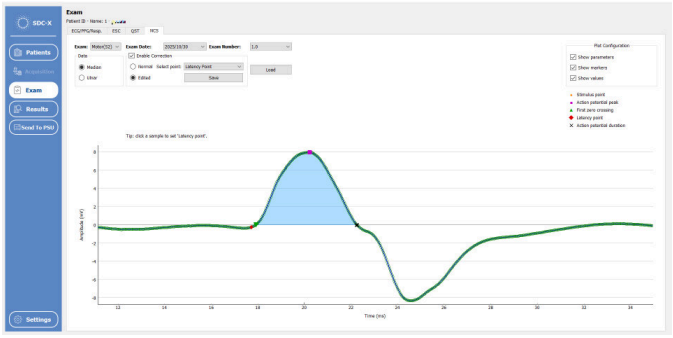


Fig. 4. *Exam* interface displaying the filtered signal, the automatically detected stimulation-artifact marker and CMAP landmarks, and tools for manual correction of event annotations.

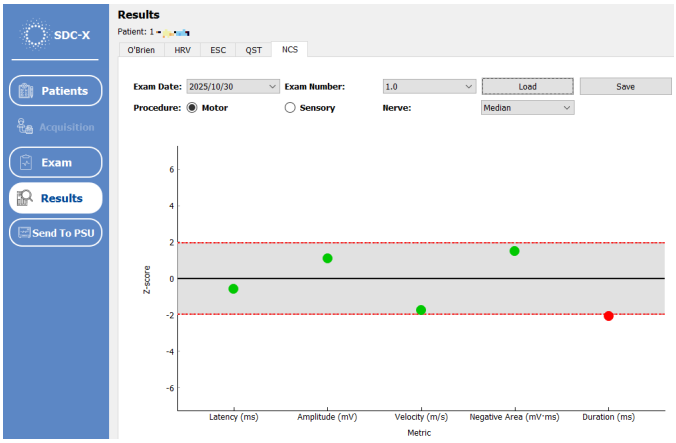


Fig. 5. *Results* interface displaying the exploratory Z-score visualization of the electrophysiological parameters extracted by the X-NCS module. Dashed horizontal lines indicate $z = -2$ and $z = +2$. The visualization provides normative contextualization and does not constitute a diagnostic classification.

Individual markers could be reviewed and manually repositioned when needed. In the 14 datasets included in the functional evaluation, no manual repositioning of the automatically generated stimulation-artifact or CMAP landmarks was required (0/14 datasets). This observation describes operator use of the implemented workflow; it does not establish annotation accuracy because the automatic landmarks were not compared with an independent expert reference. Inter-rater agreement and reproducibility across operators were likewise not assessed.

The availability of manual correction remains important because NCS recordings can be affected by baseline fluctuations, residual stimulus artifacts, waveform morphology, electrode placement, temperature, and cursor positioning, all of which can influence measurement variability [8], [9], [27].

C. Plausibility of the Extracted Parameters

Table II summarizes the parameters extracted from the analyzed recordings. Table III shows the proportion of measurements that fell within the normative ranges reported in the literature.

TABLE II
DESCRIPTIVE STATISTICS FOR THE MOTOR MEDIAN NERVE, S1 PARAMETERS, AND MNCV ($n = 14$ PARTICIPANTS). SD: STANDARD DEVIATION; CV: COEFFICIENT OF VARIATION; AMP.: AMPLITUDE; LAT.: LATENCY; DUR.: DURATION.

	Amp. (mV)	**Lat.** (ms)	**Dur.** (ms)	**Area** (mV·ms)	**MNCV** (m/s)
Mean	7.6	4.2	6.6	19.7	62.8
SD	1.9	0.9	1.5	5.8	10.0
Median	7.9	4.0	6.5	18.4	65.2
Min.	3.9	2.8	4.1	9.4	38.6
Max.	10.6	6.1	9.1	29.1	74.8
CV (%)	25.2	22.0	22.5	29.2	15.9

TABLE III
PLAUSIBILITY ASSESSMENT RELATIVE TO NORMATIVE REFERENCES [20] ($n = 14$ FOR EACH PARAMETER).

Parameter	**Cases within range**	**Percentage (%)**
S1 amplitude	14/14	100.0
S1 area	13/14	92.9
S1 duration	13/14	92.9
MNCV	13/14	92.9
S1 latency	10/14	71.4

The mean amplitude was 7.6 ± 1.9 mV, the mean distal latency was 4.2 ± 0.9 ms, and the mean MNCV was 62.8 ± 10.0 m/s. These values are consistent with those generally reported for the motor median nerve in healthy adults [20], [27].

S1 latency was the parameter with the lowest proportion of measurements within the consulted normative range: 10 of the 14 recordings (71.4%). Several factors may contribute to this finding, including individual variability, skin temperature, anthropometric differences, and differences between the present acquisition protocol and the populations or methods used to establish the published reference limits. This is relevant because nerve-conduction reference values are not completely interchangeable across laboratories and protocols, and their interpretation should therefore consider the conditions under which the measurements were obtained [27].

The normative ranges were used to check whether the values produced by the pipeline were physiologically plausible. They do not provide an estimate of measurement error because no reference instrument or independent expert annotation was available. The dataset is also unsuitable for estimating sensitivity, specificity, predictive value, or diagnostic performance for DPN: the sample was small, included only volunteers without DM, was restricted to the motor median nerve, and was collected under controlled experimental conditions. The results should therefore be interpreted as a functional software evaluation rather than as diagnostic validation.

D. Limitations and Future Work

The first limitation concerns automatic landmark annotation. No manual repositioning was required in the 14 datasets analyzed, but the automatic stimulation-artifact, onset, peak, and

end markers were not compared with independent expert annotations. The present study therefore cannot establish annotation accuracy, inter-rater agreement, or reproducibility between operators. A dedicated annotation study should preserve the original automatic landmarks, register any manual edits, and compare both landmark positions and derived measurements with independent expert markings.

Robustness and reproducibility were evaluated only within the controlled workflow used here. Examinations were not repeated across sessions or operators, recording conditions were not systematically varied, and multiple hardware configurations were not tested. Processing time was also not formally benchmarked. In addition, WebSocket interruption, reconnection, timeout, and incomplete-transmission conditions were not stress-tested. The 14/14 completion result is therefore specific to the available recordings and tested operating conditions. Formal usability and operating-system portability assessments also remain to be performed.

The experimental scope was limited to motor median nerve recordings from volunteers without DM, and the acquisition unit was used as the signal source rather than as the object of metrological validation. Calibrated-signal tests, electronic characterization, comparison with an established nerve-conduction system, repeated-session assessment, and reproducibility across hardware configurations are separate validation steps. Future studies should also examine additional peripheral nerves and include participants with DM and DPN.

The evaluated version runs locally and does not use the SDC-X cloud infrastructure. Its structured files support record-level organization and traceability, but cloud synchronization, role-based access control, user-level audit logging, and deployment-related privacy controls were not evaluated. These functions belong to a future connected version of the platform. At its current stage, X-NCS is intended as a research tool that provides a consistent workflow for acquisition, review, filtering, CMAP annotation, parameter extraction, and structured local storage; it is not a substitute for a validated clinical nerve-conduction system.

IV. Conclusion

X-NCS was developed as an SDC-X module for organizing and processing motor NCS data. It connects patient and session management with acquisition-unit communication, signal visualization, digital filtering, CMAP event annotation, parameter extraction, and structured result storage.

When applied to motor median nerve recordings from 14 participants without DM, X-NCS completed the defined functional workflow for all 14 datasets. The distal and proximal recordings were processed, CMAP landmarks were generated and reviewed, electrophysiological parameters were calculated, and the results were stored locally. No manual landmark repositioning was required in the evaluated datasets, and the extracted values were physiologically plausible for the analyzed sample. Because no independent expert annotation or reference nerve-conduction system was included, these findings establish functional execution of the workflow rather than annotation accuracy, measurement agreement, diagnostic performance, or clinical equivalence.

The next validation steps address three separate questions: annotation performance against independent expert markings; repeatability and robustness across sessions, operators, recording conditions, and communication failures; and metrological characterization of the acquisition chain using calibrated signals and an established nerve-conduction system. Larger studies can then extend the evaluation to additional nerves and participants with DM and DPN. A future cloud-connected version will also require verified privacy, access-control, and audit mechanisms. Until those validation stages are completed, X-NCS should be used as a research tool for organizing, processing, and reviewing NCS data, with interpretation remaining the responsibility of qualified professionals.

Acknowledgment

The authors thank IEB-UFSC for its technical, institutional, and infrastructure support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001.

References

[1] I. Genitsaridi *et al.*, "11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050," *The Lancet Diabetes & Endocrinology*, vol. 14, no. 2, pp. 149–156, 2026, doi: 10.1016/S2213-8587(25)00299-2.

[2] C. W. Hicks and E. Selvin, "Epidemiology of peripheral neuropathy and lower extremity disease in diabetes," *Current Diabetes Reports*, vol. 19, no. 10, art. 86, 2019, doi: 10.1007/s11892-019-1212-8.

[3] D. Selvarajah, D. Kar, K. Khunti, M. J. Davies, A. R. Scott, J. Walker, and S. Tesfaye, "Diabetic peripheral neuropathy: advances in diagnosis and strategies for screening and early intervention," *The Lancet Diabetes & Endocrinology*, vol. 7, no. 12, pp. 938–948, 2019, doi: 10.1016/S2213-8587(19)30081-6.

[4] American Diabetes Association Professional Practice Committee, "12. Retinopathy, neuropathy, and foot care: Standards of Care in Diabetes—2024," *Diabetes Care*, vol. 47, Suppl. 1, pp. S231–S243, 2024, doi: 10.2337/dc24-S012.

[5] K. Juster-Switlyk and A. G. Smith, "Updates in diabetic peripheral neuropathy," *F1000Research*, vol. 5, art. 738, 2016, doi: 10.12688/f1000research.7898.1.

[6] Z. Yang *et al.*, "Simple tests to screen for diabetic peripheral neuropathy," *Cochrane Database of Systematic Reviews*, no. 7, 2018, Art. no. CD010975, doi: 10.1002/14651858.CD010975.pub2.

[7] J. C. Won and T. S. Park, "Recent advances in diagnostic strategies for diabetic peripheral neuropathy," *Endocrinology and Metabolism*, vol. 31, no. 2, pp. 230–238, 2016, doi: 10.3803/EnM.2016.31.2.230.

[8] X. Kong, E. A. Lesser, J. T. Megerian, and S. N. Gozani, "Repeatability of nerve conduction measurements using automation," *Journal of Clinical Monitoring and Computing*, vol. 20, no. 6, pp. 405–410, 2006, doi: 10.1007/s10877-006-9046-8.

[9] X. Kong, E. A. Lesser, and S. N. Gozani, "Repeatability of nerve conduction measurements derived entirely by computer methods," *BioMedical Engineering OnLine*, vol. 8, art. 33, 2009, doi: 10.1186/1475-925X-8-33.

[10] O. O. Ogunduyile, O. O. Olugbara, and M. Lall, "Development of wearable systems for ubiquitous healthcare service provisioning," *APCBEE Procedia*, vol. 7, pp. 163–168, 2013, doi: 10.1016/j.apcbee.2013.08.028.

[11] H. D. Vianna and J. L. V. Barbosa, "A scalable model for building context-aware applications for noncommunicable diseases prevention," *Information Processing Letters*, vol. 148, pp. 1–6, 2019, doi: 10.1016/j.ipl.2019.03.010.

[12] J. E. R. Tavares and J. L. V. Barbosa, "Ubiquitous healthcare on smart environments: A systematic mapping study," *Journal of Ambient Intelligence and Smart Environments*, vol. 12, no. 6, pp. 513–529, 2020, doi: 10.3233/AIS-200581.

[13] T. G. Terra, "Software de gerenciamento, análise e detecção precoce da disfunção autonômica em indivíduos com Diabetes Mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2020.

[14] M. V. B. Franco, "Sistema remoto distribuído para avaliação da função autonômica em indivíduos com diabetes mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2021.

[15] M. Ücker, "Desenvolvimento de sistema point-of-care de pupilometria dinâmica aplicado no screening da Neuropatia Autonômica Diabética," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2020.

[16] K. K. Nakamura, "Sistema para avaliação da função sudomotora aplicado na detecção precoce do Diabetes Mellitus tipo 2," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2022.

[17] J. C. Scorpion, "Sistema de medição de limiares térmicos para integração em plataforma de saúde ubíqua de rastreamento de complicações do Diabetes Mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2024.

[18] J. C. Scorpion, J. L. P. S. Paula, and J. L. B. Marques, "Prototype development for thermal threshold tracking in peripheral neuropathy identification in diabetic patients using quantitative sensory testing," in *Proc. Brazilian Congress on Biomedical Engineering*. Springer, 2024, pp. 521–530.

[19] L. D. Weiss, J. M. Weiss, and J. K. Silver, *Easy EMG: A Guide to Performing Nerve Conduction Studies and Electromyography*, 3rd ed. Elsevier, 2022.

[20] D. Kumbhare, L. Robinson, and R. Buschbacher, *Buschbacher's Manual of Nerve Conduction Studies*. New York, NY, USA: Springer Publishing Company, 2015.

[21] B. A. Perkins, J. Grewal, E. Ng, M. Ngo, and V. Bril, "Validation of a novel point-of-care nerve conduction device for the detection of diabetic sensorimotor polyneuropathy," *Diabetes Care*, vol. 29, no. 9, pp. 2023–2027, 2006, doi: 10.2337/dc08-0500.

[22] J. A. Lee, E. M. Halpern, L. E. Lovblom, E. Yeung, V. Bril, and B. A. Perkins, "Reliability and validity of a point-of-care sural nerve conduction device for identification of diabetic neuropathy," *PLOS ONE*, vol. 9, no. 1, Art. no. e86515, 2014, doi: 10.1371/journal.pone.0086515.

[23] W. Grabowska, R. King, S. Roll, I. V. Habermann, S. Hörder, K. Hahn, S. N. Willich, S. Schröder, B. Brinkhaus, and J. Dietzel, "Reliability of a novel point of care device for monitoring diabetic peripheral neuropathy," *Scientific Reports*, vol. 13, Art. no. 19043, 2023, doi: 10.1038/s41598-023-45841-6.

[24] S. Pleven, N. Papanas, A. Gatt, and C. Formosa, "Screening for diabetic peripheral neuropathy: Subjective versus objective measures," *The International Journal of Lower Extremity Wounds*, 2024, doi: 10.1177/15347346241295461.

[25] Y. Hayashi, T. Himeno, Y. Shibata, N. Hirai, Y. Asada-Yamada, S. Sasajima, *et al.*, "Simplified electrophysiological approach combining a point-of-care nerve conduction device and an electrocardiogram produces an accurate diagnosis of diabetic polyneuropathy," *Journal of Diabetes Investigation*, vol. 15, no. 6, pp. 736–742, 2024, doi: 10.1111/jdi.14174.

[26] P. E. Barkhaus, S. D. Nandedkar, M. de Carvalho, M. Swash, and E. V. Stålberg, "Revisiting the compound muscle action potential (CMAP)," *Clinical Neurophysiology Practice*, vol. 9, pp. 176–200, 2024, doi: 10.1016/j.cnp.2024.04.002.

[27] S. Chen, M. Andary, R. Buschbacher, D. Del Toro, B. Smith, Y. So, K. Zimmermann, and T. R. Dillingham, "Electrodiagnostic reference values for upper and lower limb nerve conduction studies in adult populations," *Muscle & Nerve*, vol. 54, no. 3, pp. 371–377, 2016, doi: 10.1002/mus.25203.

[28] K. Kurokawa, Y. Mimori, E. Tanaka, T. Kohriyama, and S. Nakamura, "Age-related change in peripheral nerve conduction: compound muscle action potential duration and dispersion," *Gerontology*, vol. 45, no. 3, pp. 168–173, 1999, doi: 10.1159/000022081.

Construction of an Annotated Dataset of Oral Cavity Lesion Images for Neural Network Applications

1st Leticia Thaynna Xavier A. O. de Azevedo
dept. of Biomedical Engineering
UFRN
Natal, Brazil
thaynna.araujo.016@ufrn.edu.br

2nd Francisca Thaís Queiroz Rêgo
dept. of Biomedical Engineering
UFRN
Natal, Brazil
thais.rego.092@ufrn.edu.br

3rd Deyvisom Michel M. Dantas
dept. of Biomedical Engineering
UFRN
Natal, Brazil
deyvisom84@gmail.com

4th Caroline Dantas Vilar
dept. of Biomedical Engineering
UFRN
Natal, Brazil
caroline.vilar@ufrn.br

5th Heliana Bezerra Soares
dept. of Biomedical Engineering
UFRN
Natal, Brazil
heliana.soares@ufrn.br

6th Francisco Eduardo M. de Castro Alves
dept. of Technology
FEMCA
Natal, Brazil
fealvesphone@gmail.com

***Abstract*—Oral cavity disorders represent a clinical challenge due to their diverse presentations and the need for accurate diagnosis, particularly in potentially malignant conditions. In this context, artificial intelligence-based tools have demonstrated significant potential to support the analysis of medical images; however, their development depends on the availability of well-organized, representative, and properly annotated datasets. This study aimed to construct a dataset of oral lesion images from scientific literature for future application in artificial intelligence computational models. The methodology consisted of a literature survey in scientific databases, selection of publications containing clinical images of oral disorders, organization of the collected files in a virtual environment, and categorization of the images into seven main classes: oral potentially malignant disorders and malignancies; reactive and traumatic lesions; normal anatomical variations and benign conditions; immunologically mediated conditions; viral, fungal, and bacterial infections; benign salivary gland disorders; and benign pigmented lesions. Subsequently, the selected materials were uploaded to the IA Bucal Files platform for file management and manual annotation of lesion regions and anatomical structures of the oral cavity. As a result, a total of 3,303 images were collected from approximately 2,100 scientific publications. To date, approximately 56.1% of the images have undergone manual annotation, resulting in 4,170 lesion annotations generated during the construction of the image dataset. It is noteworthy that the annotation process is still ongoing, allowing further expansion and refinement of the dataset. The developed dataset has the potential to support future applications involving the development, training, and evaluation of artificial intelligence models for the detection and classification of oral lesions, thereby contributing to the development of computer-aided diagnostic tools in oral medicine.**

***Keywords*—Dataset; Oral lesions; Image annotation; Convolutional neural networks.**

I. Introduction

Dental care, particularly in the context of oral diseases and disorders, remains a persistent challenge for healthcare systems due to the complexity of diagnosing, treating, and monitoring these conditions. In Brazil, the implementation of the National Oral Health Policy (Política Nacional de Saúde Bucal – PNSB), known as the "Brasil Sorridente" (Smiling Brazil) program, by the Ministry of Health in 2004 represented a significant milestone in expanding access to dental services, contributing to improvements in oral healthcare delivery and the overall oral health of the population [1].

The implementation of this policy also enabled the reorganization of primary oral healthcare, mainly through the expansion of Oral Health Teams integrated into the Family Health Strategy, as well as the strengthening of secondary care through the establishment of Specialized Dental Centers (Centros de Especialidades Odontológicas – CEOs) and Regional Dental Prosthesis Laboratories (Laboratórios Regionais de Próteses Dentárias – LRPDs). These initiatives increased service capacity and promoted a more integrated approach to oral healthcare [2].

Despite the advances achieved through public health policies, the identification and monitoring of oral disorders remain challenging. According to the Brazilian National Cancer Institute (Instituto Nacional de Câncer – INCA), an estimated 17.190 new cases of oral cavity cancer are expected annually in Brazil during the 2026–2028 triennium, corresponding to an estimated risk of 7,98 cases per 100.000 inhabitants. Of these, approximately 12.260 cases are expected to occur in males and 4.930 in females. Excluding non-melanoma skin cancer,

oral cavity cancer ranks as the seventh most common type of cancer in the country [3].

According to Leal, Emmi, and Araújo (2021) [4], an evaluation of Specialized Dental Centers (CEOs) revealed that 34.45% of the analyzed units did not have a specialist in Oral Medicine (Stomatology), whereas only 65.55% had professionals working in this specialty. This scenario highlights a gap in specialized oral healthcare, particularly considering the essential role of Oral Medicine in the diagnosis and early detection of oral disorders, including oral cancer.

In this context, the development of technological tools capable of supporting the diagnostic process has emerged as a promising alternative to assist healthcare professionals in clinical decision-making and improve diagnostic support. However, the development of image-based diagnostic technologies still faces several challenges.

Studies in this field have shown that, despite the advances in deep learning-based models, the availability of representative datasets remains one of the major challenges for their application to medical imaging. The diversity of clinical presentations and the need for large volumes of annotated images directly influence the performance of computational models [5].

More recently, deep learning techniques have been explored as auxiliary tools for medical image analysis, enabling the automated identification of patterns associated with clinical abnormalities. In the context of oral lesions, previous studies have demonstrated the potential of deep neural networks for the detection and classification of oral disorders, contributing to strategies for early diagnosis [5].

In this context, this study proposes a methodological workflow for collecting, categorizing, and manually annotating oral cavity lesion images extracted from scientific literature published under an open-access model. The primary contribution of this work lies in establishing a standardized clinical taxonomy comprising seven major disorder classes, along with a structured clinical image annotation protocol implemented on the IA Bucal Files platform. It is important to emphasize that this research represents an ongoing study, with dataset expansion and manual annotation actively in progress. Nevertheless, this work establishes a solid framework for organizing heterogeneous literature-derived image collections into standardized structures, aiming for the future utilization of this dataset in training neural networks applied to computer-aided oral diagnosis .

II. Materials and Methods

This study is characterized as an exploratory and methodological investigation aimed at constructing an image dataset of oral lesions through the collection, organization, and annotation of images obtained from scientific literature. The dataset was structured into seven main categories of oral disorders: oral potentially malignant disorders and malignancies; reactive and traumatic lesions; normal anatomical variations and benign conditions; immunologically mediated conditions; viral, fungal, and bacterial infections; benign salivary gland disorders; and benign pigmented lesions. The construction of the image dataset followed a methodological workflow composed of sequential stages, as illustrated in Figure 1. Initially, scientific digital libraries were selected as data sources, from which publications containing images of the target oral lesions were identified. Subsequently, the selected files were stored and organized in a virtual environment according to their clinical categories. After cataloging, the documents were uploaded to the IA Bucal Files platform, where the images underwent analysis and manual annotation of the regions corresponding to oral lesions and anatomical structures of the oral cavity. At the end of the process, an organized and annotated image dataset was obtained, containing information associated with the identified oral disorders and intended for future applications involving artificial intelligence-based computational models.

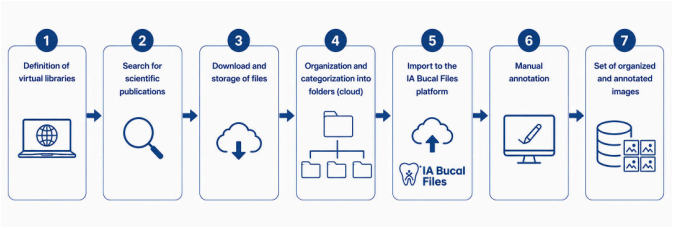


Fig. 1. Methodological Workflow

A. Literature Review

The initial stage of dataset development consisted of a literature survey of scientific references related to oral lesions. For this purpose, academic databases and scientific digital libraries were used, including Google Scholar, Scientific Electronic Library Online (SciELO), CAPES Journal Portal, Scopus, and PubMed, encompassing publications up to 2026. The selection of these sources considered the breadth of their collections, the availability of open-access publications, and the relevance of the literature to the research domain.

The information retrieval strategy employed combinations of keywords and Boolean operators in both English and Portuguese. As an example of the search string applied across the digital platforms, the query used was: (”oral lesions” OR ”oral pathology” OR ”mouth neoplasms”) AND (”clinical presentation” OR ”case report” OR ”diagnosis”). The use of multiple search engines and combined strategies expands access to scientific production and optimizes the identification of suitable clinical images, as different platforms utilize distinct information retrieval algorithms [6].

In addition to traditional databases, the Connected Papers platform was utilized as a complementary tool for mapping and exploring the scientific literature. The platform enables the identification of relationships between articles by using a seed article, keyword, or Digital Object Identifier (DOI), generating a visual representation based on publication similarity. This approach facilitated the tracking of related works and the expansion of the initial literature survey.” .

B. Inclusion and Exclusion Criteria

During the selection process of scientific publications, documents containing clinical images of oral lesions relevant to dataset construction were included. Publications were excluded if they lacked images associated with the investigated disorders or presented visual records unsuitable for annotation, including low-quality images, inadequate lighting, superimposed textual elements over the anatomical site, insufficient information for lesion identification, or features that prevented proper delimitation of the region of interest. To mitigate the inclusion of duplicate cases, same clinical case or image reused in different publications, the upload process on the IA Bucal Files platform incorporated an automated verification mechanism that identified and flagged previously existing images in the repository.

In addition to image availability, aspects related to the scientific relevance of the publications were considered, including analysis of the study context, references used by the authors, and their contribution to the literature in the field. Citations in other scientific works were used as a complementary parameter to evaluate the relevance and academic dissemination of the selected publications.

At the end of the survey process, approximately 2,100 scientific publications containing 3,303 images suitable for subsequent organization, categorization, and annotation stages were selected.

C. Organization and Categorization of Scientific Publications

Following the selection of scientific publications and identification of target images, the materials were organized in a virtual storage environment using a hierarchical folder structure and subsequently uploaded to the IA Bucal Files platform. The organization was structured according to the clinical classification of oral lesions, enabling data standardization and facilitating subsequent selection and annotation steps. The manual annotations were performed by undergraduate students in Dentistry and Biomedical Engineering who underwent dedicated training on the target oral lesions, with their work reviewed by a Master's researcher specializing in Oral Medicine (Stomatology) alongside image management workflow.

The image dataset was divided into seven main categories of oral disorders: (1) oral potentially malignant disorders and malignancies; (2) reactive and traumatic lesions; (3) normal anatomical variations and benign conditions; (4) immunologically mediated conditions; (5) viral, fungal, and bacterial infections; (6) benign salivary gland disorders; and (7) benign pigmented lesions.

The first category comprised disorders with malignant potential and malignant neoplasms, including squamous cell carcinoma, verrucous carcinoma, lymphomas, adenoid cystic carcinoma, oral melanoma, and Kaposi's sarcoma, as well as potentially malignant conditions such as leukoplakia, erythroplakia, erythroleukoplakia, actinic cheilitis, and oral lichen planus.

The reactive and traumatic lesions category included disorders associated with proliferative responses or irritating agents, such as traumatic fibroma, mucocele, thermal burns, inflammatory papillary hyperplasia, pyogenic granuloma, peripheral giant cell lesion, and traumatic ulcers.

The normal anatomical variations and benign conditions category included alterations without significant pathological characteristics, such as fissured tongue, geographic tongue, Fordyce granules, linea alba, palatal and mandibular tori, and exostoses.

The immunologically mediated conditions category was organized to include manifestations such as recurrent aphthous stomatitis, pemphigus vulgaris, mucous membrane pemphigoid, and oral manifestations of systemic lupus erythematosus.

Infections were grouped according to their etiological origin, encompassing viral, fungal, and bacterial conditions such as candidiasis, herpes simplex, HPV-associated squamous papilloma, condyloma acuminatum, paracoccidioidomycosis, and syphilis.

Benign salivary gland disorders included pleomorphic adenoma, sialadenitis, sialolithiasis, and ranula. Finally, the benign pigmented lesions category comprised alterations such as amalgam tattoo, smoker's melanosis, melanocytic nevus, and melanotic macule.

This organizational structure enabled the image dataset to be arranged into distinct classes, promoting greater standardization of the collected information and preparing the dataset for annotation tasks while facilitating reference to the original source articles during the oral cavity lesion annotation process

D. File Registration and Image Annotation Process

For the organization and annotation stages, the IA Bucal Files computational platform, previously developed for the management and annotation of oral lesion images, was employed. The platform served as the working environment for dataset registration and for annotating the regions corresponding to the identified oral disorders.

Initially, a new dataset was created within the platform, including information regarding the oral lesion category, dataset description, and source identification. After creating the dataset, the scientific publications selected during the literature review were uploaded in PDF format. During this process, the IA Bucal Files platform automatically extracted the images using Optical Character Recognition (OCR), adding to the dataset only the images contained in each uploaded PDF document.

After the files were uploaded, the manual annotation process of the clinical images began. Using the annotation tools available within the platform, the regions corresponding to oral lesions and their associated anatomical structures were manually delineated. It is important to note that a single image may contain multiple lesion annotations as well as multiple anatomical region annotations, allowing a more detailed representation of the information contained in the dataset.

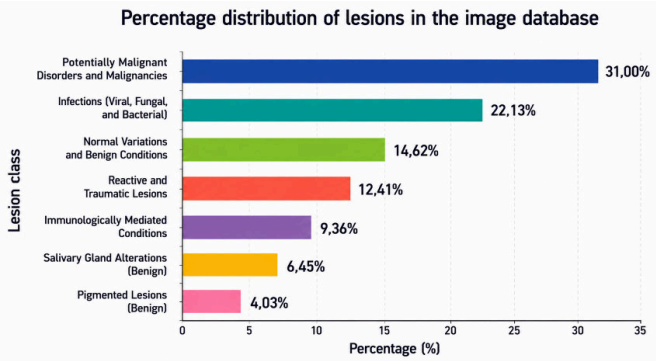


Fig. 2. Number of Images per Oral Lesion Category

At the end of this stage, a structured and annotated image dataset was obtained, containing information associated with the identified lesions and their corresponding anatomical regions. This dataset provides a foundation for future applications involving medical image processing and artificial intelligence-based computational models.

III. Results and Discussion

The literature review conducted for the seven categories of oral lesions resulted in the selection of approximately 2.100 scientific publications, which were stored in a virtual repository. These publications were uploaded to the IA Bucal Files platform, leading to the initial construction of a dataset containing 3.303 clinical images of oral lesions extracted from scientific publications. As shown in Figure 2, the highest concentration of images was observed in Category 1 (Oral Potentially Malignant Disorders and Malignancies), whereas Category 7 (Benign Pigmented Lesions) contained the smallest number of images.

The unequal distribution of images among the categories is associated with the number of clinical conditions included in each category and may also reflect the availability of scientific publications containing clinical images of these conditions. The Oral Potentially Malignant Disorders and Malignancies category exhibited the largest representation in the dataset because it encompasses a greater number of oral disorders and conditions of high clinical relevance, particularly those associated with the diagnosis of oral cancer. In contrast, categories comprising fewer clinical conditions, such as Benign Pigmented Lesions, contained fewer available images.

This variation in class distribution represents an important consideration for future artificial intelligence applications, as imbalanced datasets may affect the learning process of computational models by favoring classes with greater representation while reducing their ability to generalize to underrepresented categories [7]. Therefore, the continuous expansion of the dataset and the pursuit of a more balanced class distribution are essential steps toward the development of more robust models for oral lesion detection and classification.

At the time of analysis, the dataset contained 4.170 lesion annotations, distributed among benign (3.150), malignant (491), and potentially malignant (529) lesions. However, the annotation process was still ongoing, with approximately 56.1% of the images having undergone manual annotation. This stage is essential for the construction of datasets suitable for computational modeling, as annotation quality directly influences the learning capability and generalization performance of algorithms applied to medical imaging [8].

During the annotation process, an additional image quality control procedure was performed. Images presenting superimposed artifacts within the regions of interest, low resolution, blur, inadequate illumination, or insufficient information regarding their source were excluded from the dataset. Furthermore, both the images and their corresponding annotations were reviewed by dental professionals to improve the reliability of the identification of oral disorders represented in the dataset.

Annotation quality is a determining factor in the reliability of medical image datasets, since inconsistencies introduced during this stage may compromise the construction of the reference annotations used for artificial intelligence applications. Inter-observer variability and the absence of standardized annotation protocols have been identified as important challenges in the development of medical datasets, potentially affecting the consistency and reproducibility of computational models. Therefore, structured expert review procedures, together with well-defined annotation criteria and workflows, are essential to improve the reliability of datasets used for artificial intelligence applications [9], [10].

IV. Conclusion

This study enabled the construction of a structured image dataset of oral lesions derived from scientific literature, comprising 3303 images organized into different clinical categories and subjected to standardization and manual annotation using the IA Bucal Files platform. The development of this dataset demonstrates the feasibility of using scientific literature as a source for creating datasets intended for future computational applications in healthcare.

The organization and quality assurance of the images represent fundamental steps in the development of artificial intelligence-based tools, as data quality directly influences the analytical capability and performance of computational models. Consequently, the developed dataset has the potential to be integrated into future computer-aided systems for the detection and classification of oral lesions, serving as a supportive tool for dental and healthcare professionals.

As future work, the proposed directions include expanding the literature review, incorporating additional images into the dataset, improving the balance among lesion categories, and employing the structured dataset to develop and evaluate artificial intelligence models for integration into computer-aided diagnostic systems in Oral Medicine.

Acknowledgments

The authors thank LABETRE (Laboratory of Tissue Engineering and Rehabilitation) and LAPSIM (Laboratory of

Medical Signal and Image Acquisition and Processing) of the Biomedical Engineering Program at the Federal University of Rio Grande do Norte (UFRN) for the technical support, infrastructure, and assistance provided for the development of this work. The authors also acknowledge the National Council for Scientific and Technological Development (CNPq), Brazil, for financial support.

REFERENCES

[1] PUCCA JUNIOR, G. A. Política Nacional de Saúde Bucal do Brasil, integralidade e acesso; o caso Brasil sorridente. 2013. 99 f. Orientadora: Nilce Santos de Melo. Tese (Doutorado em Ciências da Saúde) – Faculdade de Ciências da Saúde, Universidade de Brasília, Brasília.

[2] BRASIL. Ministério da Saúde. Portaria Nº 1.444 de 28 de dezembro de 2000. Estabelece incentivo financeiro para a reorganização da Atenção em Saúde Bucal prestada nos municípios por meio do Programa Saúde da Família. Diário Oficial da União, Brasília, 29 dez. 2000. Disponível em: https://www.nescon.medicina.ufmg.br/biblioteca/imagem/1721.pdf

[3] INSTITUTO NACIONAL DE CÂNCER (Brasil). Estimativa 2026: incidência de câncer no Brasil: síntese de resultados e comentários. Rio de Janeiro: INCA, 2025. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa/sintese-de-resultados-e-comentarios. Acesso em: 10 jun. 2026.

[4] LEAL, Rhuan Vitor Sodré; EMMI, Danielle Tupinambá; ARAÚJO, Marizeli Viana De Aragão. Acesso e qualidade da atenção secundária e da assistência em estomatologia no Brasil. Physis: Revista de Saúde Coletiva, Rio de Janeiro, v. 31, n. 2, e310205, 2021. DOI: http://dx.doi.org/10.1590/S0103-73312021310205.

[5] WELIKALA, Roshan Alex et al. Automated Detection and Classification of Oral Lesions Using Deep Learning for Early Detection of Oral Cancer. IEEE Access, v. 8, p. 132677-132693, 2020. DOI: 10.1109/ACCESS.2020.3010180

[6] GOMES, Daniel Libonati; BENCHIMOL, Alegria Celia; BARROS, Thiago Henrique Bragato. O uso de ferramentas de busca e acesso a artigos acadêmicos pelos pesquisadores brasileiros. Informação Sociedade: Estudos, João Pessoa, v. 28, n. 1, p. 141-154, jan./abr. 2018.

[7] HE, Haibo; GARCIA, Edwardo A. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, v. 21, n. 9, p. 1263-1284, 2009. DOI: 10.1109/TKDE.2008.239.

[8] Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[10] H. D. Zaj ac et al., "Ground truth or dare: Factors affecting the creation of medical datasets for training AI," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2023, pp. 351–362.

CORA: An Open-Source Desktop Application for Automated Wound Healing Quantification and Software Performance Benchmarking

Lucas Militello da Silva
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0000-4461-8716

Ana Laura Guastali Sant'Ana
Faculty of Eletrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0009-5668-1603

Guilherme Vieira de Faria
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0009-0004-0933-5189

Bellisa Freitas Barbosa
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-4611-2844

Eloisa Amalia Vieira Ferro
Institute of Biomedical Science
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-7447-2583

Ana Claudia Patrocinio
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Pedro Cunha Carneiro
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-0120-5273

***Abstract*—Digital image analysis and processing have become essential tools for enhancing objectivity and statistical rigor in biomedical research. However, widely reproduced experimental protocols, such as the *in vitro* scratch wound healing assay, are still heavily limited by manual segmentation subjectivity, poor standardization, and substantial human labor, which ultimately lower their reproducibility and scientific impact. To overcome these limitations, this paper introduces CORA (Cell Open-Region Analyzer), an open-source desktop application developed in Python designed to support the quantification of cellular migration via automated image-based analysis. CORA streamlines the research workflow by automatically generating binary masks of the scratch area for subsequent user validation, calculating wound sizes, and computing closure rates over time. To rigorously validate the software's operational readiness, reliability and computational efficiency, we performed a comprehensive benchmarking evaluation under multiple stress conditions, measuring CPU utilization, RAM throughput, and execution times across singular and extensive batch processing (up to 60 images). Furthermore, a robotic interaction algorithm was deployed to simulate low- and high-intervention human workload scenarios. The results demonstrate that CORA achieves outstanding computational efficiency, executing automated segmentation in just 14 seconds per image (for single and 60 batch tests) — representing up to an 8-fold speedup compared to traditional manual tracing (120 s/image). By significantly reducing processing bottlenecks and providing an integrated workflow for automated segmentation and expert review, CORA establishes a standardized, high-performance, and user-friendly framework that drastically improves reproducibility and quantitative throughput in cellular wound healing studies.**

***- Index Terms*—Automated quantification, Benchmarking, Digital image processing, Open-source software, Software performance, Wound healing assay.**

I. Introduction

Ever since digital imaging equipment entered the world of science, life scientists have collaborated with computer scientists to apply image-processing techniques to analyze biomedical data. The aim is to use computational processes to accelerate repetitive tasks while also obtaining quantitative results, since statistical results are much more compelling, scientifically speaking, than qualitative observations [1].

In vitro scratch wound healing assay, a simple and low-cost technique that works along with other image analysis tools is one of the most widely used 2D methods to determine the cellular migration and proliferation in processes such as regeneration and disease. In this regard, the scratch or migration assay is a widely used tool for *in vitro* studies of the rates of migration, angiogenesis, movement, proliferation, and healing in response to different novel drug candidates [2]. To obtain data from the assay, images of different exposures must be taken over a period at the exact same spot to compare the movements of the migration [3].

Data acquisition and analysis in most wound healing assays is done at least to some extent manually. The manual extraction of data from micrograph is always time-consuming, error-prone, and the data output is very subjective depending on the individual researcher. Thus, analyzing a large set of data still remains the bottleneck of many assays. Also, the comparison of data from different tools and/or different experimenters is not very reliable [4]. Despite being widely employed, the

scratch wound healing assay is poorly standardized both in its execution (e.g., gap creating strategies and wound surface quantification protocols) and in the techniques used to quantify its results. These aspects significantly increase the variability and reduce the accuracy of the measurement [5].

Advances in biomedical image-processing techniques have allowed for the corroboration of research outcomes by quantifying them in a rigid, statistical manner while simultaneously raising the bar for life science research when it comes to substantiating scientific observations [1].

To bridge this gap, this study introduces CORA (Cell Open-Region Analyzer), an open-source, desktop software application developed in Python designed to streamline and standardize the image-analysis workflow used for the quantification of cellular migration in scratch assays. By embedding an intuitive graphical user interface (GUI) with interactive refinement tools, the software aims to democratize access to robust image-processing pipelines for life science researchers without programming expertise. Furthermore, this work presents a comprehensive benchmarking evaluation of the software's computational performance, quantifying CPU, memory throughput, and operational efficiency under simulated user workloads validating its computational performance and usability in high-throughput, reproducible research tool.

II. Materials and Methods

A. Software Architecture

To streamline laboratory workflows, CORA was designed to provide life science researchers with immediate, decentralized access to automated segmentation masks and numerical wound dynamics. To provide an integrated understanding of the system's runtime execution, the overall logical sequence, algorithmic modules, and step-by-step user interactions are systematically mapped out in Fig. 1. Furthermore, to guarantee full reproducibility and transparency for open-source collaborative deployment, the comprehensive repository architecture, licensing framework, core development metadata, and complete technical dependencies required to run the environment are detailed in Table I.

The application architecture was implemented in Python utilizing the Tkinter graphical library for desktop environment management, which inherently supports efficient event handling, layout customization, and low overhead [8]. The development of this interactive graphical user interface (GUI) was intentionally prioritized to abstract the underlying computer vision complexities, delivering an intuitive, plug-and-play tool tailored for operators with no prior programming experience.

The application starts with the request of folders containing the images to process and the folder to save its results. The software expects a standardized file naming convention, such as filename-XXX.tif, where XXX represent the specific time of acquisition of the image (0 hour, 24 hours, or 48 hours), and it accepts other formats of images as well. Based on the information contained on the file's names, the software divides the images in groups with each time sequence, used to calculate further data related to the variation of area.

These images are shown by groups on the screen, using the Matplotlib library, a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible [9]. This modularity enables users to correct grouping errors or customize configurations directly, being used through the whole procedure, to facilitate the user to visualize the progression of the segmentation.

Next, the program starts processing all the images selected by the user. A progression bar appears, showing the remaining time to end the processing as well as showing which images are being processed. With the segmentation, the program calculates the size of each area in pixels (A), as well as the progression of closure, denoted by 1, where Y represents the

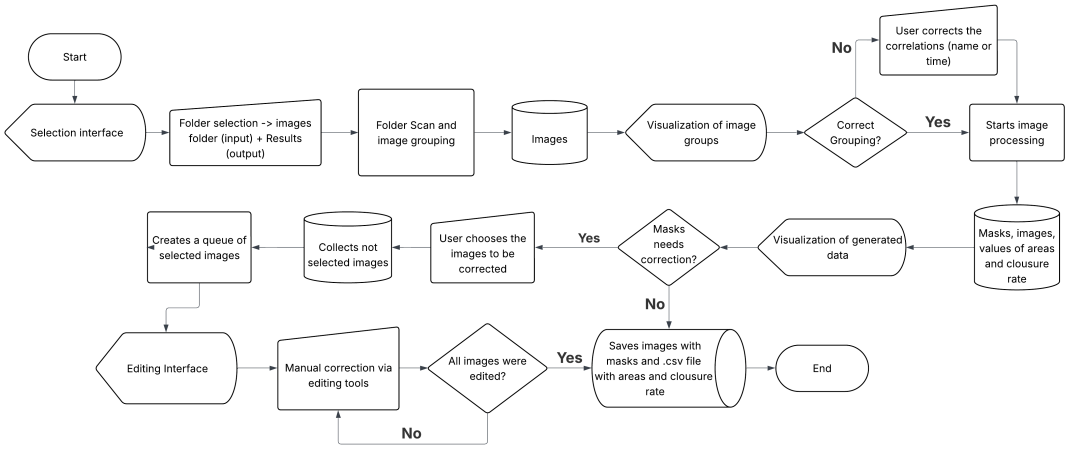


Fig. 1. Methodological architecture and user operational workflow of CORA (Cell Open-Region Analyzer). The block diagram maps the sequential software logic, including input folder scanning, automated multi-time-point image alignment, image processing execution, interactive mask validation, and final data export to structured files.

TABLE I. Repository metadata and technical dependencies for CORA

Code metadata description	**CORA**
Current code version	v0.1.0 – Initial Tkinter implementation; experimental PySide6 interface
Permanent link to code/repository	https://github.com/LucasMilitello/CORA-App
Legal Code License	MIT License
Code versioning system used	Git and GitHub
Software code languages, tools, and services used	Python 3.12.10; Tkinter; PySide6 6.8.0; PyInstaller
Operating system and dependencies	Windows 11, 64-bit; NumPy 2.4.4; OpenCV 4.13.0.92; Matplotlib 3.10.8; SciPy 1.17.1; scikit-learn 1.8.0; tifffile 2026.3.3; Pillow 12.2.0; threadpoolctl 3.6.0
Benchmark hardware	AMD Ryzen 7 5800H (8 cores/16 threads); 32 GB RAM; NVMe SSD
Benchmark environment	Windows 11, 64-bit; Python 3.12.10; CPU execution without explicit GPU/CUDA acceleration
Developer documentation/manual	https://github.com/LucasMilitello/CORA-App/blob/main/DEVELOPER-MANUAL.md
Support email for questions	militello.lucas@ufu.br

time point:

$$C_Y = \frac{A_0 - A_Y}{A_0} * 100\% \quad (1)$$

While the core image processing algorithms lie outside the immediate scope of this evaluation, they utilize an established computational pipeline, using adaptive filtering for image enhancement, gradients for edge detection, filters for texture analysis, mathematical morphology for region refinement, and image segmentation, continuously developed and optimized by our research group in parallel with this software integration.

On the visualization page, the resulting segmented images of the groups separately will be displayed. From this point, the user has two options: either to maintain the automatic segmentation and save as is, or to put it in a queue to edit the contour to what they think is appropriate.

The editing page consists of using Matplotlib for the visualization of the segmented image and combining OpenCV, which is an open-source library that includes several hundreds of computer vision algorithms [10], and Numpy, a Python library that provides a multidimensional array object and an assortment of routines for fast operations on arrays [11], for the modeling part of the mask.

After this stage, all the changes made to the masks substitute the origin of the base processing, including the variation on the related values of areas and percentages. All numerical data will be provided in a .csv file together with the segmented images.

B. User Workflow

The software receives the full package of files that the user chooses to be segmented and divides them by groups of equal naming but varies by dates, process and calculate the area of the wound as well as the percentage of closure for each image grouping. We advise the user of using the same pattern when giving names for the files, since the calculations uses the exact sequence of timing (for example, the percentage of closure of a 24-hour image depends on the value of that same image on 0 hour).

The resulting segmented images are shown to the user by the stated division, as well as the values of these segmentations in pixels. The user can either accept the segmented contour and its size value, or they can reshape it with the editing tool. When selecting the images for the editing session, it is forced to the user to visualize every group of selected images, to guarantee optimal usage of the software. After the analysis, the selected segmented images are saved with a report that contains all the comparative data generated through the processing and revision stages and the ones unchecked, if any, will be delegated to the editing queue.

The editing will be made with tools that enables the user to modify the contour, maintaining the overall structure of the segmentation. After finishing the session, the user returns to the visualization page with the changes made and all the images automatically marked to be saved.

As a Python application, all technical dependencies listed in Table I must be installed on the user's machine.

C. Software Performace

The testing will consist of three analysis using an AMD Ryzen 7 5800H, SSD 475GB, 32 GB RAM DDR4 computer. For each of these tests, we will analyze data related with its performance (CPU and RAM usage) and time usage of the user, to determine how viable it is to use this application instead of other existent tools. The environments of these tests were summarized on I.

III. RESULTS

A. Interface Appearance

The graphical interface comprises three core operational modules: the selection and configuration interface (Fig. 2), the visualization and validation panel (Fig. 3) and the interactive editing workspace (Fig. 4).

The selection page successfully acquires and distributes all the images within the designated input directory into temporal groups while filtering out any duplicates recognized by the algorithm, as systematically presented in Fig. 2. This module also supports dynamic group renaming and displays real-time image previews to facilitate early dataset organization. Following this initial setup, the application transitions to the visualization page, which effectively displays the processed

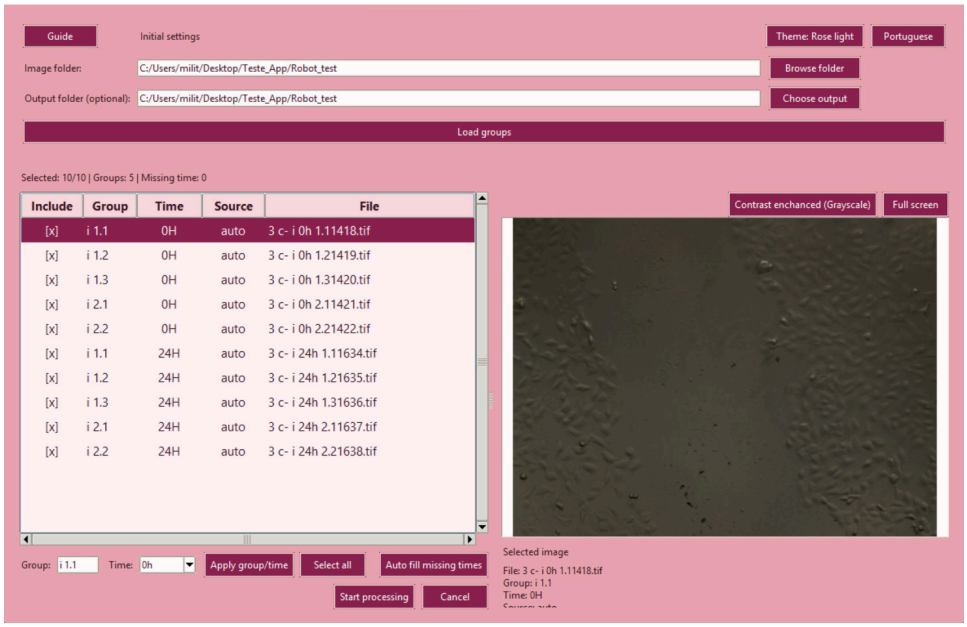


Fig. 2. CORA selection and configuration interface. This module automatically parses file naming conventions to parse and group time-series images (0 h, 24 h, and 48 h), automatically filters duplicated files, and provides a real-time side panel for visual verification of image grouping prior to segmentation processing.

image groups arranged horizontally with the automated segmentation masks seamlessly overlaid, as illustrated in Fig. 3. Within this panel, users can actively review the initial area calculations, evaluate computed closure rates, and determine whether specific boundaries require manual configuration or should be rolled back to the original automated state.

The editing page exhibits the image with an editable mask utilizing custom tools that enable the operator to modify the boundaries by erasing, rearranging the contours, or creating a mask from scratch, as displayed in Fig. 4. To prevent human-induced bias, these tools are calibrated to balance localized user curatorial interventions with the structural integrity of the underlying automatic segmentation morphology. For enhanced visual acuity of the scratch margins, a ”Contrast Enhanced” feature converts the image into a high-contrast grayscale version. Furthermore, step-by-step user guides are integrated into every sub-page, which also feature multi-lingual support for English and Portuguese alongside multiple dark and light color themes to optimize long-term operational comfort in laboratory environments.

B. Interface Testing

The first test consists of evaluating the complete operational routine of the application for one image focusing on time spent on processing segmentation using the python standard library time module, used for acquiring this data [12], and the maximum usage of RAM memory and CPU during the process with the library psutil(process and system utilities) is a cross-platform library for retrieving information on running processes and system utilization (CPU, memory, disks, network, sensors) in Python [13], for the operating data of these internal components. After processing, the results are presented in Table II.

TABLE II. Computational performance metrics for single-image processing in CORA.

Metric	**Mean ± SD ($\mu \pm \sigma$)**
Image loading time (s)	0.0225 ± 0.002
Segmentation time (s)	14.232 ± 1.108
Peak RAM usage (MB)	1052.42 ± 24.885
Peak CPU usage (%)*	33.70 ± 6.588

*CPU usage values were normalized by the number of logical cores.

The second will evaluate the same data as the first one, but focusing on batch processing, varying from 10, 24 and 60 images, with the same protocol from the first test, focusing on the stability of the application with an extensive task. To analyze, we used the same routine for the data and the results were put on Table III.

TABLE III. Benchmarking evaluation under graded batch processing configurations

Metric	**Images x Iterations ($\mu \pm \sigma$)**		
	10x10	**24x10**	**60x10**
Total time (s)	146.320 ± 1.014	339.417 ± 1.957	877.424 ± 17.934
Peak RAM usage (MB)	1429.66 ± 18.31	1987.64 ± 56.40	3414.42 ± 28.96
Peak CPU usage (%)*	43.93 ± 7.43	53.14 ± 6.26	47.84 ± 4.42
Images per minute	4.10 ± 0.03	4.25 ± 0.02	4.10 ± 0.08

*CPU usage values were normalized by the number of logical cores.

And the last test will be a simulation of two routines of possible outcomes when using the app: the no-editing scenario (A), when editing images is not needed, and the full-editing scenario (B), when all images need to be edited. To simulate a user-interface interaction, we utilized a bot algorithm with the library pywinauto, a set of python modules to automate the Microsoft Windows GUI. At its simplest it allows you to send

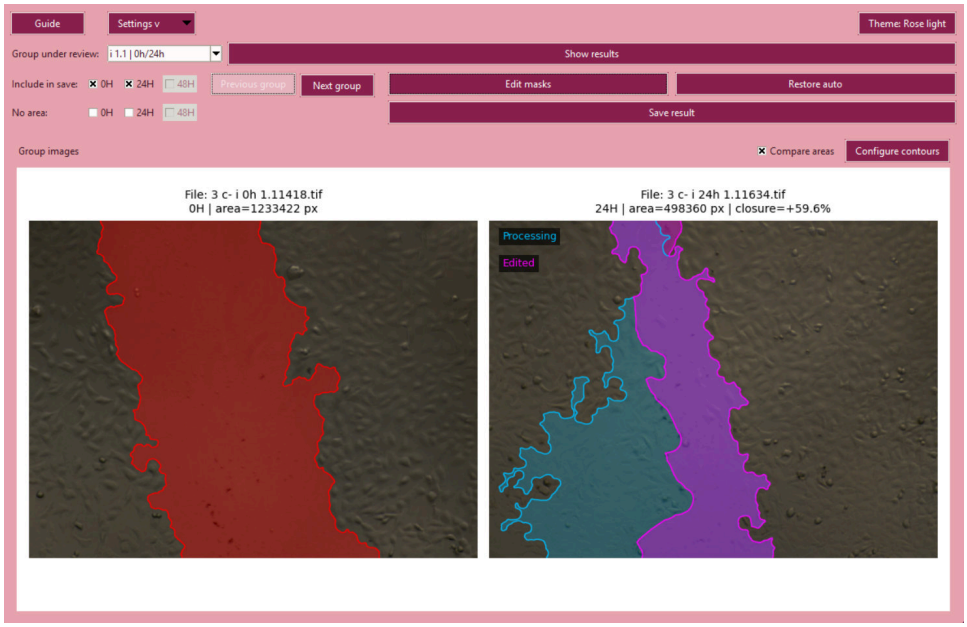


Fig. 3. CORA visualization panel exhibiting side-by-side comparative analysis between automated segmentation (red overlay) and manually adjusted boundaries (purple/blue overlay). This visual monitoring allows instant validation of area calculations and closure rate computations, enabling users to rollback or redo contour modifications.

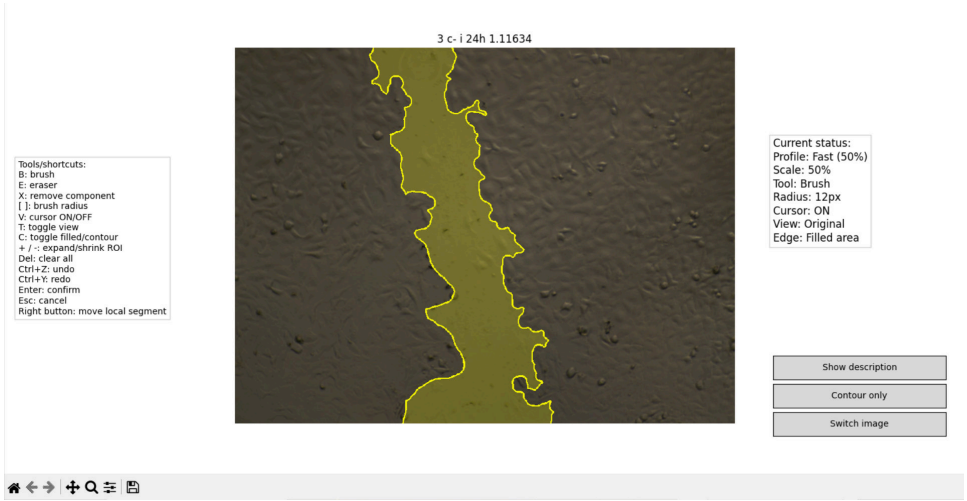


Fig. 4. CORA interactive editing workspace and manual curation panel. The interface integrates dedicated drawing tools for fine-grained mask manipulation, including local brush adjustments, component erasing, and rapid system navigation via keyboard shortcuts such as brush size controls, standard undo (Ctrl+Z), and redo (Ctrl+Y) mechanisms. To ensure optimal visual acuity under diverse experimental conditions, operators can dynamically toggle the segmentation mask overlay between a semi-transparent filled-area mode or a boundary-only contour profile. Additionally, the workspace supports real-time color customizability of the segmented region alongside local contrast enhancements, meticulously balancing user-guided editing flexibility with the morphological fidelity of the underlying cell monolayer.

mouse and keyboard actions to windows dialogs and controls [14], to emulate the actions of the user. The bot reproduced the clicks and interactions necessary to execute those tests, with a delay of 1 second for each click and 15 seconds for editing one image.

To ensure a rigorous and standardized assessment of user-interface pathways, the software includes a dedicated benchmarking control dashboard, which allows operators to systematically monitor execution runtimes across distinct workloads, as illustrated in Fig. 5. The operational timeline data compiled via this robotic interface emulation, contrasting the execution latencies across all primary sub-pages under both no-editing and full-editing interaction conditions, is quantitatively consolidated and presented in Table **??**.

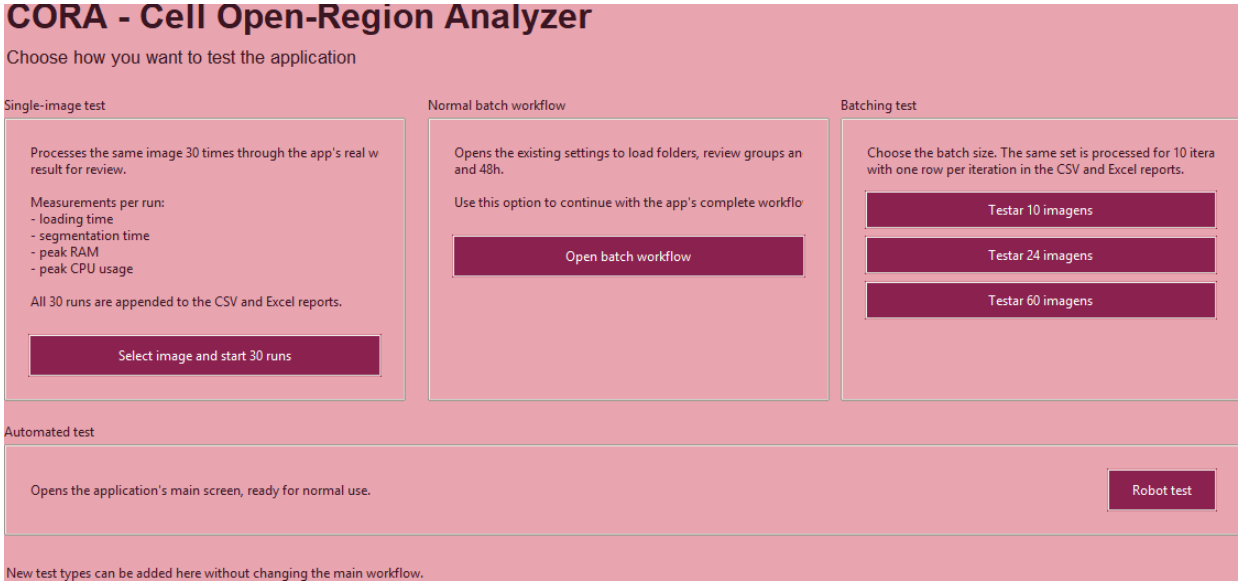


Fig. 5. CORA automated testing and benchmarking dashboard interface. This module coordinates the execution of standardized performance assessments, enabling the user to launch single-image reliability trials, trigger graded multi-image batch workloads (10, 24, and 60 micrographs), or execute UI-driven robotic interaction emulations via pywinauto to benchmark operational workflows under controlled operational latencies.

TABLE IV. Interface Operational Time for Automated Workload Simulations (Robot Tests A and B)

Time on Pages (s)	**Test A** ($\mu \pm \sigma$)	**Test B** ($\mu \pm \sigma$)
Selection	7.567 ± 1.407	8.491 ± 1.346
Configuration	6.190 ± 0.004	6.161 ± 0.094
Visualization	10.844 ± 0.106	8.060 ± 0.097
Editing	0	159.568 ± 0.010

Operational data compiled using pywinauto emulation to replicate no-editing (Test A) and full-editing (Test B) human interactions.

IV. Discussion

The assessment of *in vitro* wound healing images is critical for determining the efficacy of the therapy-of-interest that may influence the wound healing process. Existing methods suffer significant limitations, such as user dependency, time-consuming nature, and lack of sensitivity, thus paving the way for automated analysis approaches [15]. For these reasons, the development of applications that can facilitate the evaluation of these images by automatically quantifying the wound area, and thus, the closure rate is of paramount importance to accelerate these types of studies.

Our software demonstrated high operational throughput, achieving an automated segmentation speed of approximately 14 seconds per individual micrograph. When subjected to extensive batch processing configurations (such as the 60-image stress test), the system maintained strict stability, yielding an efficiency rate of over 4 images processed per minute. This computational performance drastically outperforms manual operations, with an average analysis time of 120 s/image and highly dependent on the user's ability of consistently elaborating all the images [5]. Ultimately, with the time of 1 image of manual determination, we can generate three segmented images with partial participation of the user or eight images without their intervention.

Regarding the performance validation of the automated scratch detection layer, the algorithm achieved a baseline area extraction consistency of 82% for the 0-hour micrographs when compared directly against the manual annotations performed by specialized biological researchers. For the more challenging 24-hour post-scratch time points, where hyper-fragmented cell monolayers and localized cellular debris naturally obscure the gap boundaries, the raw automated algorithm experienced a characteristic overestimation trend, yielding an average area deviation ratio of +76% prior to user intervention. Crucially, the programmatic inclusion of the interactive editing module successfully mitigates this computational variance; the intuitive layout of the manual reshaping tools allows operators to rapidly audit, refine, and correct these algorithmically complex boundary segmentations, effectively aligning the final quantitative output with the rigorous morphological consensus established by expert biological ground truth.

A key conceptual advantage of CORA is that it avoids the rigid, automated 'black-box' processing paradigms common to many fully autonomous computer vision tools. In complex, cross-contaminated, or highly irregular *in vitro* biological cultures, traditional segmentation strategies frequently collapse or demand unfeasible parameter recalibrations due to cell debris, non-uniform cell proliferation, and poorly defined wound fronts. Instead of enforcing a rigid, fully autonomous execution that would inevitably yield scientifically flawed data under these stressful conditions, CORA embraces a human-in-the-loop computing architecture. The underlying processing pipeline performs approximately 90% of the intensive

computational segmentation in seconds, while the graphical interface—implemented via Tkinter with future scalability targeted for advanced PySide6 multi-threading modules built upon the Qt Framework [16] —serves as an interactive quality control layer. This architectural synergy empowers domain experts to rapidly audit, refine, or validate the generated masks, ensuring that only highly accurate and reproducible biological data is exported to the final structured report.

Regarding system resource allocation, the peak computational consumption observed during the trials deserves careful architectural consideration. As reported in Tables II and III, memory allocation varied from a baseline medium peak of 1052.42.66 MB during single-image workflows to 3414.42 MB under the maximum 60-image batch stress test. This memory footprint behavior is tied directly to how Python manages high-resolution arrays (2048 x 1534). To enable interactive graphical features, such as real-time canvas updates, image contrast enhancements, and multi-step undo/redo histories, the software must retain multiple copies of heavy NumPy matrices and Matplotlib visualization buffers concurrently in active memory. While a 3.5 GB peak usage is considerable, it is completely compatible with standard modern laboratory workstations equipped with 8 GB or 16 GB of RAM, confirming that CORA is fully viable for routine laboratory environments without specialized server hardware.

While the literature frequently highlights that the scratch wound healing assay suffers from poor standardization in its execution and quantification protocols, CORA actively mitigates this vulnerability by establishing a deterministic computational workflow. By locking the mathematical extraction of the wound boundaries into standardized digital pipeline, and restricting manual intervention to strictly monitored contour scaling tools, the software effectively eliminates the subjective variance introduced by different human experimenters.

From a software engineering standpoint, evaluating user workflows using human test subjects can often bias system benchmarks due to uncontrolled external variables, such as operator fatigue, distraction, or varying motor reflexes. Utilizing an automated UI-driven bot framework via pywinauto provided a highly rigorous, reproducible method to evaluate CORA's interface efficiency. By clamping user actions to strict timing rules, exactly 1 second between navigation clicks and a fixed 15-second block for localized mask manipulations, the simulation successfully isolated human performance variance from the raw performance of the software architecture. This level of rigorous stress-testing guarantees that the documented execution timelines represent a reliable upper bound for operational throughput, a methodological step that adds substantial software engineering validity to the application.

V. Conclusion

This paper presented CORA (Cell Open-Region Analyzer), an open-source, desktop software application designed to support and standardize the image-analysis workflow for quantifying cellular migration via automated image analysis in *in vitro* scratch wound healing assays. By combining dynamic digital image processing pipelines with an intuitive, multi-lingual graphical user interface developed in Python, the software successfully bridge the gap between complex computational vision and laboratory life science research. CORA delivers automated segmentation masks, precise area measurements in pixels, and objective wound closure rates over time. To ensure high-quality, scientifically sound outputs, the application integrates targeted user curatorial tools that seamlessly balance expert human intervention with the underlying algorithmic structure, providing a more standardized image-analysis workflow while retaining expert oversight.

The computational readiness of the tool was rigorously validated through benchmarking performance assessments and automated workload emulations using robotic interaction algorithms. The statistical data compiled across single-image and graded batch processing stress tests confirm that CORA maintains remarkable operational throughput under extensive tasks. Executing individual automated segmentations in approximately 14 seconds, the application demonstrates an approximate 8-fold computational speedup compared to traditional manual digital tracing protocols, which average 120 seconds per micrograph. Furthermore, the editing function of the application easy to use and efficient enough to make any deviation of the automatic process easily fixable. Even under simulated full-editing user interaction profiles (where every generated boundary mask is subjected to full manual reshaping) the overall human-software hybrid workflow caps execution times at 39 seconds, effectively shifting the data acquisition bottleneck commonly observed in high-throughput biological screening assays.

Future research and technical developments will focus on expanding the architectural features of CORA to broaden its clinical and research utility. The authors intend to incorporate advanced, adaptive processing methods capable of handling diverse image modalities, such as phase-contrast, fluorescence, and non-uniform low-contrast microscopy, which often present distinct boundary noise characteristics. Furthermore, we aim to deploy deep-learning-based segmentation modules, such as U-Net architectures, as alternative backend processing engines to enhance boundary prediction in hyper-infected or highly fragmented monolayers. By incorporating feedback and tool suggestions from the open-source scientific community, CORA is poised to evolve into a highly versatile, scalable, and indispensable collaborative ecosystem for quantitative biomedical imaging analysis.

Acknowledgment

The authors are thankful for the received financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Brazil), CNPq—Brazil, and FAPEMIG—Brazil, as well as the scientific and technical support provided by the Laboratório de Imunofisiologia da Reprodução (LABIFIR).

References

[1] J. Schindelin, C. T. Rueden, M. C. Hiner, and K. W. Eliceiri, "The ImageJ ecosystem: An open platform for biomedical image analysis,"

Molecular Reproduction and Development, vol. 82, no. 7–8, pp. 518–529, Jul. 2015, doi: 10.1002/mrd.22489.

[2] A. Suarez-Arnedo, F. Torres Figueroa, C. Clavijo, P. Arbeláez, J. C. Cruz, and C. Muñoz-Camargo, "An image J plugin for the high throughput image analysis of in vitro scratch wound healing assays," *PLOS ONE*, vol. 15, no. 7, Art. no. e0232565, Jul. 2020, doi: 10.1371/journal.pone.0232565.

[3] M. V. Mouritzen and H. Jenssen, "Optimized scratch assay for in vitro testing of cell migration with an automated optical camera," *Journal of Visualized Experiments*, no. 138, Art. no. e57691, Aug. 2018, doi: 10.3791/57691.

[4] A. Stamm, K. Reimers, S. Strauß, P. Vogt, T. Scheper, and I. Pepelanova, "In vitro wound healing assays—state of the art," *BioNanoMaterials*, vol. 17, no. 1–2, pp. 79–87, 2016, doi: 10.1515/bnm-2016-0002.

[5] M. Cortesi, A. Pasini, A. Tesei, and E. Giordano, "AIM: A computational tool for the automatic quantification of scratch wound healing assays," *Applied Sciences*, vol. 7, no. 12, Art. no. 1237, Nov. 2017, doi: 10.3390/app7121237.

[6] W. Rasband, "ImageJ," National Institutes of Health, Bethesda, MD, USA. [Online]. Available: http://imagej.nih.gov/ij.

[7] The MathWorks, Inc., "Image Processing Toolbox," Natick, MA, USA. [Online]. Available: http://www.mathworks.com/products/image/.

[8] GeeksforGeeks, "Python Tkinter," GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/python/python-gui-tkinter/. Accessed: Jun. 17, 2026.

[9] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May–Jun. 2007, doi: 10.1109/MCSE.2007.55.

[10] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, vol. 25, no. 11, pp. 120–125, Nov. 2000.

[11] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[12] Python Software Foundation, "time — Time access and conversions," *Python 3.14.6 documentation*. [Online]. Available: https://docs.python.org/3/library/time.html. Accessed: Jun. 22, 2026.

[13] G. Rodola, "psutil," *Python Package Index (PyPI)*. [Online]. Available: https://pypi.org/project/psutil/. Accessed: Jun. 22, 2026.

[14] pywinauto Developers, "pywinauto documentation," *pywinauto documentation*. [Online]. Available: https://pywinauto.readthedocs.io/en/latest/. Accessed: Jun. 22, 2026.

[15] D. Doğru, G. D. Özdemir, M. A. Özdemir, U. K. Ercan, N. Topaloğlu Avşar, and O. Güren, "An automated in vitro wound healing microscopy image analysis approach utilizing U-net-based deep learning methodology," *BMC Medical Imaging*, vol. 24, Art. no. 158, Jun. 2024, doi: 10.1186/s12880-024-01332-2.

[16] Qt Group, "Qt Framework," *Qt*. [Online]. Available: https://www.qt.io/development/qt-framework. Accessed: Jun. 25, 2026.

CORE-ABPM-FHIR: Apoio à interpretação de exames MAPA baseado em métricas clínicas utilizando o padrão FHIR

Rodrigo Augusto Antunes dos Santos
Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná
Curitiba, Brasil
rsantos.2003@alunos.utfpr.edu.br

André Luis Del Mestre Martins
Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense
Charqueadas, Brasil
andremartins@ifsul.edu.br

Leonelo Dell Anhol Almeida
Programa de Pós-Graduação em Tecnologia e Sociedade
Universidade Tecnológica Federal do Paraná
Curitiba, Brasil
leoneloalmeida@utfpr.edu.br

Fabio Pires Itturriet
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
fabioitturriet@utfpr.edu.br

***Abstract*—Ambulatory Blood Pressure Monitoring (ABPM) produces time-series data whose interpretation requires the calculation and organization of clinical metrics, including 24-hour, awake, and asleep averages, nocturnal dipping, blood pressure variability, and blood pressure load. This study presents CORE-ABPM-FHIR, a web-based prototype that retrieves blood pressure measurements represented as FHIR Observation resources, automatically calculates metrics according to selected clinical guidelines, and consolidates the results in a DiagnosticReport resource for visualization and analysis support. The system was developed using React and TypeScript and functionally verified using a 24-hour window containing 50 hospital blood pressure measurements from the MIMIC-IV Clinical Database Demo as proxy input for the processing pipeline. An internal heuristic inspection based on Nielsen's principles was also independently conducted by two authors. For the evaluated dataset, the prototype calculated and displayed the implemented metrics according to the selected guideline. The inspection identified the lack of help and documentation elements as a severity level 3 issue, along with minor interface and accessibility problems. The results support the feasibility of the proposed FHIR-based workflow, while clinical validation using actual ABPM recordings and usability evaluation involving healthcare professionals remain necessary.**

***Index Terms*—Digital Health Record, Cloud Computing, Health Information Interoperability**

I. Introdução

A Monitorização Ambulatorial da Pressão Arterial (MAPA) é um exame utilizado para a avaliação contínua da pressão arterial ao longo de 24 ou 48 horas, permitindo a análise de padrões temporais e da variabilidade circadiana dos níveis pressóricos ao longo do dia [1], [2]. Diferentemente de medições pontuais, o MAPA permite a obtenção de um conjunto estruturado de dados ao longo do tempo, incluindo informações nos períodos de vigília e de sono [3]. Apesar da disponibilidade desses dados, a interpretação dos exames de MAPA permanece um processo complexo, envolvendo a análise simultânea de múltiplos parâmetros clínicos, como médias por período (24 h, vigília e sono), descenso noturno (*dipping*), variabilidade da pressão arterial e carga pressórica [4]. Esses parâmetros são definidos por diretrizes clínicas, mas sua aplicação prática pode variar entre sistemas, resultando em inconsistências na interpretação e em maior dependência de análises manuais [5]. As principais diretrizes clínicas para a interpretação de exame MAPA incluem recomendações de entidades como a A*merican Heart Association/American College of Cardiology* (AHA/ACC) [6], a *European Society of Hypertension* (ESH) [7], a Sociedade Brasileira de Cardiologia (SBC) [8] e a *Italian Society of Hypertension* (SIIA) [9]. A Tabela I sintetiza as métricas recomendadas por essas diretrizes para a composição de um laudo de MAPA, evidenciando uma convergência quanto aos parâmetros considerados essenciais para sua interpretação e a complexidade envolvida em sua análise clínica.

Com a crescente digitalização dos sistemas de saúde, observa-se a adoção de padrões para a representação estruturada de dados clínicos, como o Fast Healthcare Interoperability Resources (FHIR), que permite a disponibilização de informações entre diferentes plataformas [10]. No entanto, o uso do FHIR tem sido predominantemente voltado à padronização da troca de dados, sem explorar plenamente seu potencial como base para mecanismos de interpretação clínica [11], [12]. Nesse contexto, há uma lacuna entre a padronização dos dados e a de sua interpretação, especialmente em exames que envolvem múltiplas métricas derivadas, como a MAPA [11], [13]. A ausência de abordagens automatizadas e interoperáveis para a extração e a organização dessas métricas limita a eficiência dos sistemas de apoio à decisão e dificulta a integração entre diferentes soluções [13], [14].

TABLE I
COMPARAÇÃO DE MÉTRICAS ENTRE DIFERENTES VERSÕES DAS DIRETRIZES

Diretriz (ano)	Médias	Dipping	Variabilidade	Carga	Classificação	Mín/Max
AHA/ACC (2025) [6]	X	X	X	X	X	
ESH (2023) [7]	X	X	X	X	X	
SBC (2018) [8]	X	X	X	X	X	
SIIA (2024) [9]	X	X	X	X	X	X

O símbolo "X" indica que a métrica correspondente é recomendada pela respectiva versão da diretriz. As versões apresentadas correspondem às referências utilizadas na análise bibliográfica das métricas.

Diante disso, este trabalho propõe um mecanismo padronizado para a interpretação de exames do MAPA, baseado no padrão FHIR. A abordagem utiliza recursos como *Observation* e *DiagnosticReport* para representar os dados e automatizar o cálculo de métricas clínicas recomendadas pelas principais diretrizes, incluindo médias por período, descenso noturno, variabilidade da pressão arterial, carga pressórica e classificação. Essas informações são integradas a uma aplicação *web* para visualização estruturada e apoio à análise. Entre as principais contribuições, destacam-se: (i) a proposta de um mecanismo padronizado de exames MAPA utilizando o padrão FHIR, permitindo a representação estruturada não apenas dos dados, mas também das métricas clínicas; (ii) a implementação de um *pipeline* automatizado para o cálculo de um conjunto consolidado de métricas clínicas, alinhado com as recomendações das principais diretrizes; e (iii) a integração dessas métricas em uma estrutura unificada de apoio à interpretação, reduzindo a dependência de análise manual e promovendo a interoperabilidade entre sistemas clínicos.

II. TRABALHOS RELACIONADOS

A fim de entender como este trabalho se posiciona, foi realizada uma busca de artigos e trabalhos relacionados, analisando aspectos como:

- Uso do padrão FHIR;
- Plataforma de visualização;
- Suporte a informações em tempo real;
- Existência de filtros temporais para consultar as informações.

A partir da Tabela II nota-se que apenas um trabalho usa explicitamente o padrão FHIR, mas com um foco na sua interoperabilidade e uma visualização temporal de informações, como frequência cardíaca e temperatura. Além disso, apenas um dos trabalhos trabalha com uma navegação temporal rica que permite ao usuário filtrar o período de tempo que ele deseja visualizar. Por fim, a maioria foca em IoT (*Internet of Things*) e na aquisição de dados por meio de sensores e dispositivos, não na visualização interativa.

Diante desse cenário, o CORE-ABPM-FHIR reúne aspectos que aparecem de forma parcial ou isolada nos trabalhos analisados. Diferentemente dos trabalhos 1, 2, 3 e 6 que não aderem a nenhum padrão de interoperabilidade, o trabalho proposto se baseia no padrão FHIR, o que pode proporcionar compatibilidade com outros sistemas estabelecidos. Além disso, em contrapartida ao trabalho 7, que, embora use FHIR, não oferece um processamento clínico elaborado, o CORE-ABPM-FHIR calcula automaticamente as métricas específicas do MAPA, como médias por período, cargas pressóricas e descenso noturno, a partir das medições brutas, consolidando esses resultados em um recurso *DiagnosticReport*. Por fim, enquanto os trabalhos 4, 5 e 8 oferecem algum nível de filtragem temporal, esta é limitada a intervalos fixos ou predefinidos, o CORE-ABPM-FHIR disponibiliza uma visualização configurável, na qual o usuário pode ajustar dinamicamente o período, aliando-se à interoperabilidade do FHIR uma camada de análise clínica automatizada que não foi identificada de forma conjunta em nenhum dos demais trabalhos comparados.

III. MÉTODOS

Esta seção descreve a arquitetura do sistema proposto, bem como os mecanismos de modelagem, processamento e cálculo das métricas clínicas. Ao final, apresentam-se os aspectos éticos envolvendo este trabalho.

A. Arquitetura do Sistema

A arquitetura proposta, ilustrada na Figura 1, organiza o sistema em quatro blocos principais: aquisição de dados, armazenamento e interoperabilidade, processamento de informações e aplicação e visualização, evidenciando o fluxo de dados desde a coleta até a visualização clínica. A Figura 1 apresenta a arquitetura conceitual adotada neste trabalho. Os blocos de aquisição e de armazenamento/interoperabilidade representam premissas da solução: assume-se que as medições de MAPA são disponibilizadas em uma infraestrutura compatível com FHIR, estruturadas como recursos clínicos e acessíveis por meio de consultas padronizadas. A contribuição central deste trabalho concentra-se nos blocos de processamento das informações e aplicação/visualização, responsáveis por recuperar os recursos, extrair as medições, organizar os dados por período, calcular métricas clínicas e apresentá-las em uma interface de apoio à interpretação. Além disso, o desenvolvimento da aplicação se baseou no modelo de ciclo de vida em estrela [23] e avaliações de usabilidade por dois dos autores deste trabalho, de acordo com as heurísticas propostas por Jakob Nielsen [24].

B. Avaliação Heurística

A avaliação de usabilidade do sistema foi realizada por meio da avaliação heurística, técnica de inspeção de usabilidade na qual avaliadores analisam uma interface e julgam sua conformidade com um conjunto de princípios de usabilidade

TABLE II
COMPARAÇÃO ENTRE OS ARTIGOS RELACIONADOS A IOT, FHIR E VISUALIZAÇÃO DE SINAIS VITAIS

Nº	Título resumido	Uso do FHIR	Plataforma de visualização	Tempo real	Filtros temporais
1	Development of IoT based Smart System and Data (2024) [15]	Não	*Web* e *Mobile*	Sim	Não
2	Implementation of a Modular MIoT Platform for Healthcare (2025) [16]	Não	*Web* e *mobile*	Sim	Não
3	Effects of technology-enabled blood monitoring (2024) [17]	Não	*Mobile*	Sim	Não
4	Visual Analytics Dashboard Promises to Improve (2021) [18]	Não	*Web*	Não	Sim, mas apenas trimestralmente ou anualmente
5	How can Sensor Data Accessibility be Enhanced in IoT-Based Health Monitoring Systems (2024) [19]	Não	Web	Sim	Sim, mas apenas a intervalos pré definidos
6	An Enhanced IoT-Based Wristband for Remote Monitoring (2024) [20]	Não	*Web* e *Mobile*	Sim	Não
7	Fast Healthcare Interoperability Resources for Inpatient Monitoring (2022) [21]	Sim	*Web* e *mobile*	Sim	Não
8	EHDViz Clinical Dashboard (2016) [22]	Não	*Web*	Sim	Sim
9	CORE-ABPM-FHIR	Sim	*Web*	Não	Sim

reconhecidos (heurísticas), identificando problemas que possam comprometer a experiência do usuário [24]. A avaliação foi conduzida por dois dos autores deste trabalho, que inspecionaram a interface de forma independente, confrontando suas telas e funcionalidades com as dez heurísticas de usabilidade propostas por Jakob Nielsen. Para cada heurística genérica, foram definidos itens específicos de verificação — aqui denominados *heurísticas para web* — que operacionalizam o princípio geral em critérios objetivos e observáveis, aplicáveis ao contexto de sistemas *web*. Cada item foi avaliado individualmente quanto à presença de problemas de usabilidade, sendo registrada, quando pertinente, uma justificativa descritiva da constatação.

Para classificar a gravidade de cada problema identificado, foi utilizada a escala de severidade proposta por Nielsen, que varia de 0 a 4, conforme apresentado no Quadro III.

TABLE III
ESCALA DE SEVERIDADE DE NIELSEN

Grau	Descrição
0	Não representa um problema de usabilidade
1	Problema cosmético, corrigir apenas se houver tempo disponível
2	Problema menor de usabilidade, correção de baixa prioridade
3	Problema maior de usabilidade, correção de alta prioridade
4	Catástrofe de usabilidade, correção obrigatória antes do lançamento

1) Heurísticas e itens avaliados: As dez heurísticas genéricas de Nielsen foram desdobradas nos seguintes itens específicos de avaliação para *web*:

1) **Visibilidade do status do sistema** — Tempo de carga;
2) **Compatibilidade do sistema com o mundo real** — Conteúdo; Precisão do conteúdo; Sons; Imagens; Conhecimento do público-alvo;
3) **Controle do usuário e liberdade** — Controle sobre ações;
4) **Consistência e padrões** — *Links*; Escrita e gramática; Navegação;
5) **Prevenção de erros** — Compatibilidade entre navegadores;
6) **Reconhecimento ao invés de relembrança** — Identificação de elementos de interação;
7) **Flexibilidade e eficiência de uso** — Modos de interação;
8) **Estética e *design* minimalista** — *Background*; Cor; Fontes; Gráficos; *Layout*;
9) **Ajudar os usuários a reconhecer, diagnosticar e corrigir erros** — Mensagens de erro compreensíveis;
10) **Ajuda e documentação** — *Meta tags*; *Copyright*; Informações para contato.

C. Modelagem dos Dados em FHIR

A modelagem dos dados neste trabalho baseia-se no padrão FHIR (versão R5), utilizando recursos modulares como o *Observation* e o *DiagnosticReport* para representar, respectivamente, as medições individuais de pressão arterial e a consolidação clínica do exame MAPA. O recurso *Observation* é utilizado para representar cada medição individual de pressão arterial ao longo do período de monitoramento. Cada observação contém os valores sistólico e diastólico, organizados como componentes identificados pelos códigos LOINC 8480-6 (*Systolic blood pressure*) e 8462-4 (*Diastolic blood pressure*), com valores expressos em mmHg. O conjunto da medição é categorizado pelo código LOINC 85354-9 (*Blood pressure panel*), permitindo a representação estruturada de séries temporais de medições, preservando a granularidade necessária para o cálculo das métricas clínicas. A partir dessas observações, o recurso *DiagnosticReport* é utilizado para consolidar as informações do exame MAPA. Esse recurso mantém referências às *Observations* correspondentes, garantindo rastreabilidade entre os dados brutos e os resultados consolidados. A Figura 2 ilustra a modelagem proposta, destacando a estrutura dos recursos e o vínculo entre medições individuais e o relatório consolidado do exame.

Para fins de validação e demonstração da abordagem proposta, foi selecionado aleatoriamente um paciente da base de

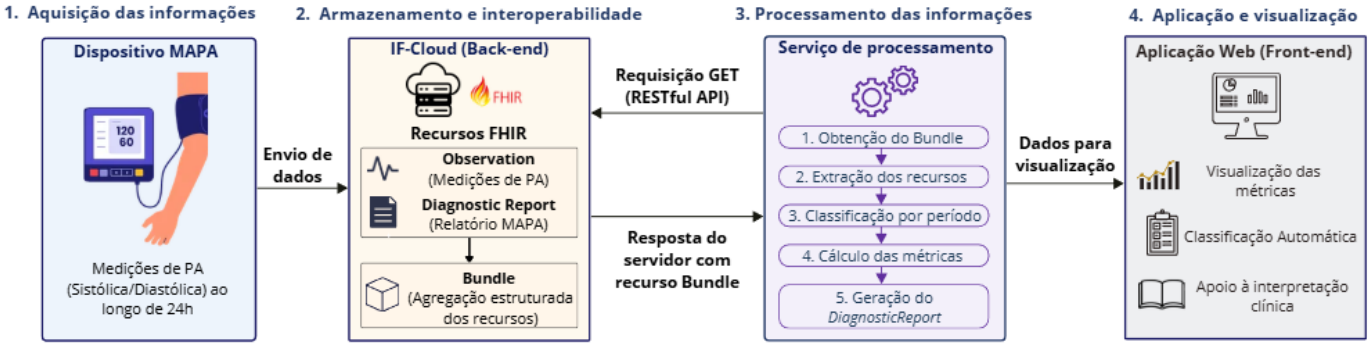


Fig. 1. Fluxo de aquisição, processamento e interpretação de exames MAPA baseado em FHIR

dados disponível, totalizando dois exames de MAPA associados a esse paciente. Cada exame já correspondia a um período de monitoramento de 24 horas, com intervalos de medições variando a cada 30-32 minutos, resultando em um total de 48 a 52 medições por exame. A distinção entre os períodos de vigília e sono é definida por um parâmetro configurável pelo usuário, que pode digitar o horário efetivo de rotina do paciente e na ausência dessa informação, o sistema adota o intervalo das 23h às 7h como período de sono e o restante do dia como período de vigília.

As medições individuais de pressão arterial são modeladas como recursos do tipo *Observation*, enquanto o recurso *DiagnosticReport* consolida essas observações, permitindo a organização estruturada e a interpretação do exame MAPA. Para fins de processamento e recuperação estruturada no *pipeline* descrito na Seção III-D, os recursos são agrupados em um *Bundle*, permitindo sua obtenção conjunta por meio de consultas ao servidor FHIR.

Ao definir a estrutura desses recursos, o FHIR promove a interoperabilidade entre sistemas heterogêneos, ou seja, a troca de informações entre diferentes sistemas é padronizada e consistente. Porém, ele não resolve particularidades locais (como CPF, Cartão Nacional de Saúde, ou terminologias específicas de cada sistema de saúde). Por isso, iniciativas nacionais criam perfis (*profiles*) que restringem e estendem os recursos padrão, tornando obrigatórios certos campos e vocabulários para o seu contexto específico.

D. Pipeline de Processamento de Dados

O *pipeline* de processamento proposto é responsável pela transformação das medições de pressão arterial, representadas como recursos FHIR, em métricas clínicas utilizadas na interpretação do exame MAPA. Para fins de validação do trabalho, os dados são recuperados a partir de um servidor FHIR hospedado na infraestrutura em nuvem IF-Cloud, mantida pelo grupo IF4Health [25], por meio de requisições à sua interface de acesso, retornando um *Bundle* contendo os recursos *Observation* associados ao exame. O uso do *Bundle* permite a obtenção conjunta das medições, garantindo consistência e integridade na recuperação das informações. Em seguida, os recursos são extraídos e organizados em estruturas de dados adequadas para processamento, sendo posteriormente classificados de acordo com o período de aquisição (vigília e sono). Essa segmentação é fundamental para o cálculo de métricas clínicas específicas, como médias por período e descenso noturno. A partir dessa organização, são construídas séries temporais de medições, que servem como base para o cálculo das métricas clínicas descritas na Seção III-E. Por fim, os resultados obtidos são consolidados em um recurso *DiagnosticReport*, que agrega as métricas derivadas e mantém referências às observações utilizadas, permitindo rastreabilidade entre os dados brutos e a interpretação clínica do exame. Esse *pipeline*, implementado com as tecnologias *React* e *TypeScript*, constitui o mecanismo central de interpretação proposto neste trabalho, permitindo a integração entre a modelagem estruturada dos dados e a geração automatizada de resultados clínicos.

E. Cálculo de Métricas Clínicas

Inicialmente, as medições foram segmentadas em três conjuntos: período total de 24 horas, vigília e sono. A média da pressão arterial em cada período foi calculada como a média aritmética das medições disponíveis. A equação (1) apresenta a fórmula da pressão arterial sistólica média,

$$\overline{PAS}_p = \frac{1}{N_p} \sum_{i=1}^{N_p} PAS_i, \tag{1}$$

onde p representa o período de análise (por exemplo: total, vigília ou sono); PAS_i representa a pressão arterial sistólica da i-ésima medição do período p; e N_p representa o número de medições no período p.

O descenso noturno (*dipping*) foi calculado como a variação percentual entre as médias dos períodos de vigília e sono, conforme a equação (2),

$$\text{Descenso noturno} = \frac{\overline{PA}_{\text{vigília}} - \overline{PA}_{\text{sono}}}{\overline{PA}_{\text{vigília}}} \times 100, \tag{2}$$

onde $\overline{PA}_{\text{vigília}}$ é a média da pressão arterial (sistólica ou diastólica) durante a vigília e $\overline{PA}_{\text{sono}}$ é a média da pressão arterial (sistólica ou diastólica) durante o sono.

A classificação do descenso noturno foi realizada com base na variação percentual da pressão arterial sistólica entre vigília

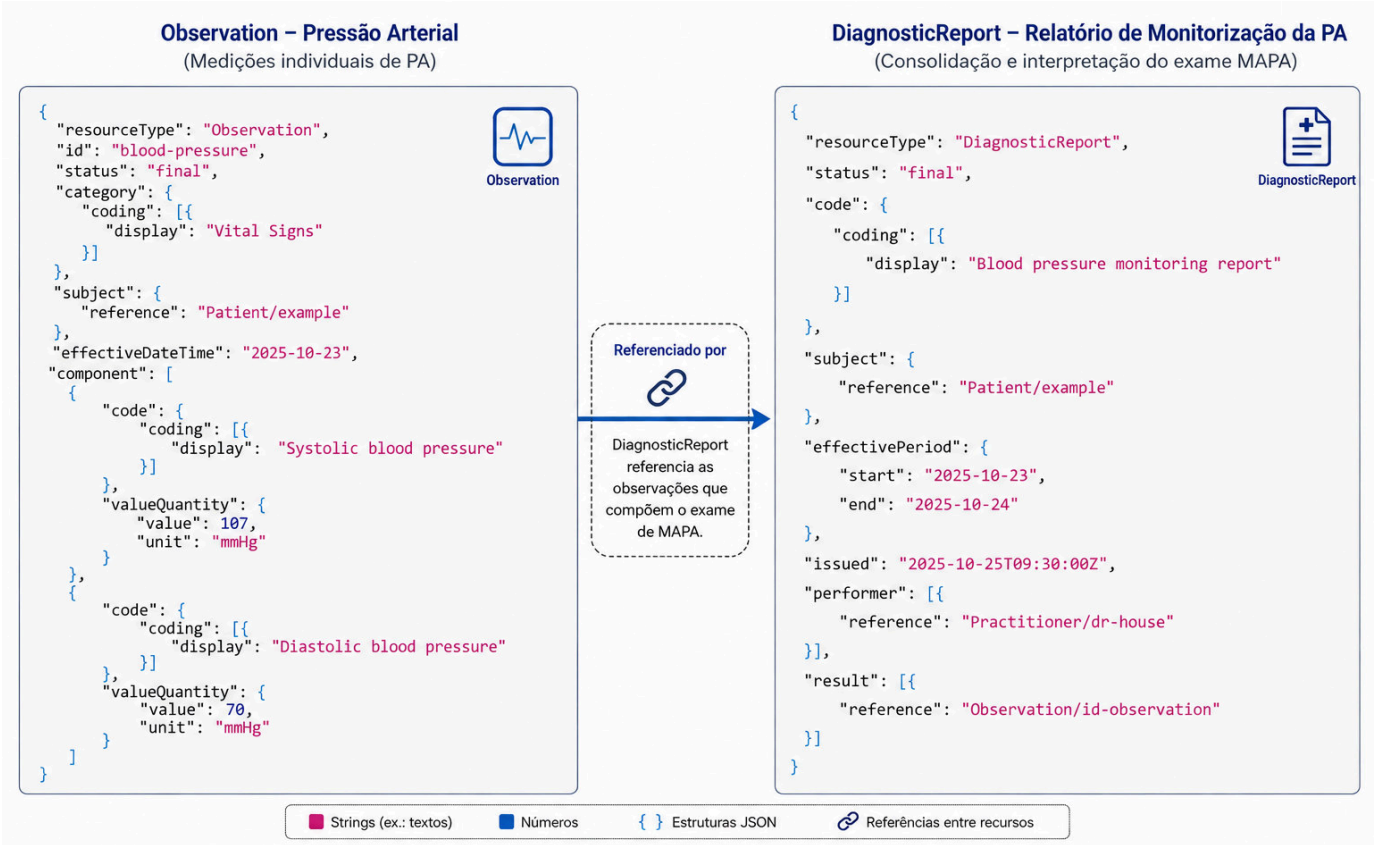


Fig. 2. Representação dos principais recursos FHIR utilizados.

e sono, conforme descrito na literatura e adotado por diretrizes internacionais (ESH, SBC), categorizando-se em reverse dipper (menor que 0%), non-dipper (de 0% a menos de 10%), dipper (de 10% a menos de 20%) e extreme dipper (20% ou mais). Essa métrica permite classificar o padrão circadiano do paciente, sendo amplamente utilizada na interpretação clínica de exames MAPA.

A variabilidade da pressão arterial foi estimada por meio do desvio padrão das medições ao longo do período analisado, conforme a equação (3),

$$DP_{x,p} = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}\left(X_i - \overline{X}_p\right)^2}, \tag{3}$$

onde x é a variável de interesse (pode ser PAS ou PAD); p é o período de análise (total, vigília ou sono); X_i é o valor da variável na i-ésima medida do período p; $\overline{X}_p$ é a média da variável no período p; e N_p é o número de medições válidas no período p.

A carga pressórica é calculada como o percentual de medições, dentro de um período p, em que a pressão arterial excede o limiar de referência, conforme a Equação (4):

$$Carga_p = \frac{NA_p}{N_p} \times 100, \tag{4}$$

em que NA_p é o número de medidas acima do limiar de referência no período p, e N_p é o número total de medidas realizadas nesse mesmo período. A Equação (4) é aplicada separadamente para a pressão sistólica e a pressão diastólica, gerando as métricas $Carga_{sis}$ e $Carga_{dia}$.

Por fim, a classificação diagnóstica da pressão arterial ambulatorial foi baseada na comparação entre as médias de pressão arterial sistólica (PAS) e diastólica (PAD) nos períodos de vigília e sono, em relação aos limiares definidos pela diretriz selecionada. Considerou-se pressão elevada quando a média de PAS e/ou PAD excedeu os respectivos valores de referência para cada período, adotando-se o operador lógico "OU" entre os componentes sistólico e diastólico.

F. Fonte dos Dados

Os dados utilizados neste trabalho foram obtidos a partir da base MIMIC-IV *Database Demo* [26], um conjunto de dados clínicos de acesso público amplamente utilizado em pesquisas na área da saúde, contendo informações de pacientes hospitalizados. Todos os dados disponíveis na base encontram-se previamente anonimizados, não havendo identificação direta dos indivíduos. Como os dados da base selecionada não segmentam em período de vígilia e sono, foi necessário construir esses períodos, do paciente selecionado pelo usuário, conforme o critério configurável descrito na seção III-C. Ressalta-se que os dados presentes na base não correspondem originalmente

a exames MAPA, mas incluem medições clínicas de pressão arterial coletadas em diferentes contextos hospitalares. Neste trabalho, tais medições foram utilizadas como entrada para o *pipeline* proposto, permitindo a validação da modelagem em FHIR, do processamento das informações e da apresentação das métricas clínicas. Dessa forma, o uso da base MIMIC-IV teve como objetivo avaliar a integração entre a estrutura de dados, o mecanismo de cálculo das métricas e a visualização dos resultados, sem validação do *pipeline* dos indicadores obtidos.

G. Aspectos Éticos

Este estudo utilizou exclusivamente dados secundários, previamente anonimizados, provenientes da base pública MIMIC-IV *Database Demo*, sem acesso a informações que permitissem a identificação dos indivíduos. A avaliação heurística consistiu em uma inspeção técnica interna da interface, realizada por dois dos autores, sem recrutamento de participantes externos, envolvimento de pacientes ou profissionais de saúde, intervenção clínica ou coleta de dados pessoais ou sensíveis. Por essas características, o estudo não foi submetido à apreciação de um Comitê de Ética em Pesquisa.

IV. Resultados

Após as sucessivas iterações de *design* e avaliação propostas pelo Ciclo de Vida em Estrela [23], obteve-se como resultado final um *dashboard* funcional para a visualização de exames MAPA. A interface de usuário foi desenvolvida com a finalidade de receber o recurso *Bundle*, realizar o processamento com o *Bundle* obtido e transformá-lo em informações clínicas de fácil visualização e entendimento. A Figura 3 ilustra o visualizador com os exames do paciente selecionado.

O retângulo 1 evidencia as informações gerais do exame de Monitorização Ambulatorial da Pressão Arterial (MAPA), incluindo identificação do paciente e número total de medições.

Na presente tela, o médico tem acesso às principais métricas definidas pelas diretrizes clínicas, apresentadas na forma de cartões informativos, destacado pelo retângulo 2. Tais cartões são configuráveis por meio da funcionalidade "Cartões visíveis", permitindo ao usuário personalizar a exibição conforme sua necessidade. Na Figura 3, todos os cartões encontram-se habilitados.

Os cartões de resultados estão organizados da seguinte forma:

- Os três primeiros apresentam as médias das pressões arteriais por período, sendo que o primeiro número equivale à pressão sistólica;
- O quarto é referente ao descenso noturno, mostrando seu valor e classificação;
- O quinto apresenta a variabilidade da PAS e PAD;
- O sexto apresenta os mínimos e máximos por período das pressões arteriais;
- Por fim, a classificação automatizada indica o resultado obtido segundo os limiares da diretriz selecionada..

Os elementos destacados pelo retângulo 3 representam componentes dinâmicos da interface, cujos valores e classificações são automaticamente atualizados de acordo com a diretriz clínica selecionada, assegurando consistência entre a visualização dos dados e os critérios normativos adotados. Os retângulos 4 delimitam o período de sono considerado na análise, o qual pode ser ajustado pelo usuário diretamente na interface do sistema, por meio dos controles disponíveis na parte superior do painel.

Os resultados apresentados no dashboard são condicionados à diretriz selecionada pelo usuário. Diferentes versões das diretrizes podem ser disponibilizadas para seleção, como AHA/ACC 2017, SBC 2018 e SBC 2025, permitindo a comparação e a aplicação de diferentes critérios de interpretação. Essa seleção possui finalidade distinta da análise bibliográfica apresentada na Tabela I, na qual foram consideradas as versões mais recentes selecionadas para identificar as métricas recomendadas pelas respectivas entidades. Portanto, a versão utilizada na comparação bibliográfica não implica necessariamente sua implementação no protótipo.

A Tabela IV apresenta as entradas do conjunto avaliado, os períodos de vigília e sono, a diretriz selecionada e as respectivas saídas do sistema, incluindo médias, carga pressórica, valores mínimo e máximo, descenso noturno, variabilidade e classificação, com o objetivo de verificar a implementação das Equações (1)–(4).

V. Discussão

A tabela 1 evidencia que há uma convergência significativa entre as principais diretrizes quanto às métricas consideradas primordiais para interpretação clínica em exames MAPA. O trabalho implementa todas as métricas apresentadas, inclusive a mais exigente, demonstrando alinhamento com práticas clínicas estabelecidas e padronizando a apresentação das métricas na interpretação dos exames, além de realizar a identificação automática do dipper e a classificação da hipertensão, sem depender de análise manual do profissional de saúde.

Outro aspecto relevante é a utilização do padrão FHIR, que permite que os dados estejam estruturados de forma padronizada, garantindo a interoperabilidade entre sistemas e facilitando a reutilização deste trabalho, dado que é um padrão adaptado para *web* e se beneficia da arquitetura REST (*Representational State Transfer*), o que permite seu uso com os padrões de requisições HTTP (*Hypertext Transfer Protocol*) como GET, POST, PUT e DELETE. Além de se beneficiar da interface de programação de aplicações (API) e os formatos de estrutura de dados como JSON (*JavaScript Object Notation*) e XML (*eXtensible Markup Language*). Em contrapartida, é um padrão que tem uma curva de aprendizado alta, já que é um padrão com diversos recursos estabelecidos e possui perfis locais (como o da RNDS - Rede Nacional de Dados em Saúde) ainda em evolução, podendo haver lacunas ou inconsistências entre versões. Além disso, sistemas que não usam FHIR perdem a capacidade de conversar facilmente com o resto do ecossistema de saúde e tendem a depender de integrações e modelos próprios, caracterizando o problema NxM: a cada novo sistema desenvolvido, é necessário adaptar

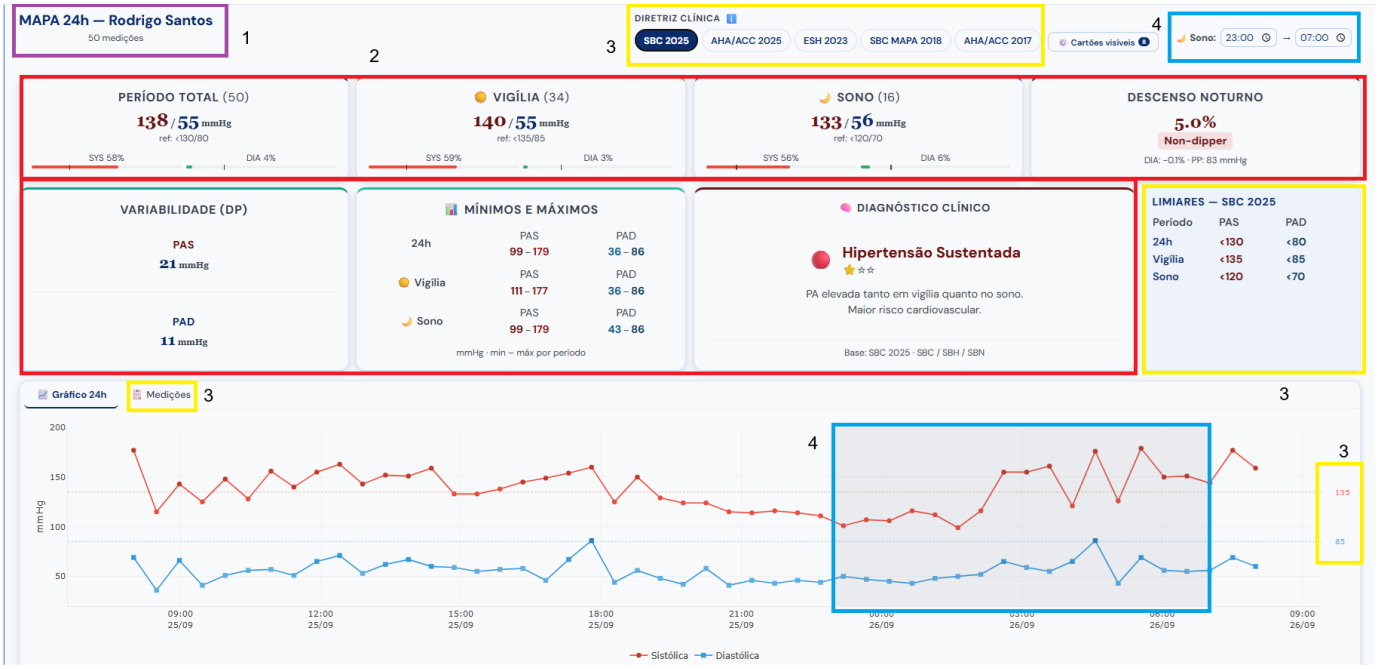


Fig. 3. Interface do CORE-ABPM-FHIR para visualização e interpretação de exames MAPA.

TABLE IV
CONFERÊNCIA INDEPENDENTE DAS MÉTRICAS DO PROTÓTIPO (PACIENTE ÚNICO, EXAME DE 24 H, $n = 50$ MEDIÇÕES; DIRETRIZ SBC 2025).

Período	**Métrica**	**Saída do sistema**
24 h ($n = 50$)	Média (DP) PAS/PAD (mmHg)	138 / 55
24 h ($n = 50$)	Carga pressórica PAS/PAD (%)	58 / 4
24 h ($n = 50$)	Mín–Máx PAS/PAD (mmHg)	99–179 / 36–86
Vigília ($n = 34$)	Média PAS/PAD (mmHg)	140 / 55
Vigília ($n = 34$)	Carga pressórica PAS/PAD (%)	59 / 3
Vigília ($n = 34$)	Mín–Máx PAS/PAD (mmHg)	111–177 / 36–86
Sono ($n = 16$)	Média PAS/PAD (mmHg)	133 / 56
Sono ($n = 16$)	Carga pressórica PAS/PAD (%)	56 / 6
Sono ($n = 16$)	Mín–Máx PAS/PAD (mmHg)	99–179 / 43–86
Derivados	Descenso noturno PAS/PAD (%)	5,0 / $-0,1$
Derivados	Variabilidade PAS/PAD (mmHg)	21 / 11
Derivados	**Classificação (SBC 2025)**	**Hipertensão Sustentada**

ou criar e manter uma tradução específica, ao contrário do modelo Nx1 em que é viabilizado por um padrão comum.

Uma limitação deste trabalho é a ausência de integração com os perfis locais da RNDS. Essa integração é apontada como um trabalho futuro e exigirá a adequação dos perfis FHIR utilizados, das terminologias propostas, do mecanismo de autenticação e dos demais requisitos técnicos estabelecidos por essa entidade.

Por fim, a avaliação heurística do sistema, baseada nos princípios de Nielsen [24], foi realizada de forma independente por dois avaliadores. O diagnóstico principal apontou uma severidade de nível 3 (grave) na categoria ajuda e documentação, devido à ausência de metadados, informações de contato e licenciamento. Em contrapartida, ajustes pontuais de acessibilidade e interface — como contraste de cores, dependência visual em gráficos e termos em inglês — já foram devidamente corrigidos. Além disso, a arquitetura em single-page proporcionou uma interação simples e previsível e não foram identificados problemas pelos avaliadores internos com o conhecimento do público-alvo.

VI. CONCLUSÕES

O CORE-ABPM-FHIR propõe uma abordagem para visualização de exames de monitorização ambulatorial da

pressão arterial (MAPA), oferecendo suporte à interpretação clínica por meio da apresentação estruturada de métricas recomendadas por diferentes diretrizes internacionais. A utilização do padrão FHIR possibilita interoperabilidade e integração com ecossistemas de saúde digital, ampliando o potencial de adoção da solução. Entretanto, para uma validação clínica mais robusta, torna-se necessária a participação de profissionais médicos na avaliação do sistema, permitindo avaliar a usabilidade e a utilidade prática do sistema, bem como identificar possíveis ajustes. Destaca-se também a ausência de uma padronização consolidada para a representação de exames de MAPA no padrão FHIR, sendo a modelagem proposta neste trabalho fundamentada na literatura disponível. Por fim, este trabalho pode servir como base para iniciativas futuras voltadas à visualização e análise de exames MAPA, com potencial de evolução para oferecer maior suporte à tomada de decisão clínica.

VII. Declaração de uso de IA

A inteligência artificial foi utilizada neste trabalho de forma pontual e auxiliar. Especificamente, foi utilizada a ferramenta ChatGPT (OpenAI, GPT-4) para apoio na revisão gramatical e ortográfica de textos e para sugestões pontuais em trechos específicos de código. Todo o conteúdo gerado pela ferramenta foi revisado pelos autores.

VIII. Links

Repositório GitHub: https://github.com/RodrigoSantos21/tcc-mapa

Vídeo demonstrativo: https://youtu.be/ooEddGbAXd0

REFERÊNCIAS

[1] B. Williams, G. Mancia, W. Spiering, E. Agabiti Rosei, M. Azizi, M. Burnier *et al.*, "2018 ESC/ESH guidelines for the management of arterial hypertension," *Eur. Heart J.*, vol. 39, no. 33, pp. 3021–3104, Sep. 2018.

[2] Sociedade Brasileira de Cardiologia, "Diretrizes brasileiras de hipertensão arterial – 2020," *Arq. Bras. Cardiol.*, vol. 116, no. 3, pp. 1–60, 2021, supplement 1.

[3] P. Muntner, D. Shimbo, R. M. Carey, J. B. Charleston, T. Gaillard, S. Misra *et al.*, "Measurement of blood pressure in humans: A scientific statement from the American Heart Association," *Hypertension*, vol. 73, no. 5, pp. e35–e66, May 2019.

[4] E. O'Brien, G. Parati, G. Stergiou, R. Asmar, L. Beilin, G. Bilo *et al.*, "European society of hypertension position paper on ambulatory blood pressure monitoring," *J. Hypertens.*, vol. 31, no. 9, pp. 1731–1768, Sep. 2013.

[5] G. Parati, G. S. Stergiou, R. Asmar, G. Bilo, P. de Leeuw, Y. Imai *et al.*, "European society of hypertension practice guidelines for ambulatory blood pressure monitoring," *J. Hypertens.*, vol. 32, no. 7, pp. 1359–1366, Jul. 2014.

[6] D. W. Jones, K. C. Ferdinand, S. J. Taler, H. M. Johnson, D. Shimbo, M. Abdalla *et al.*, "2025 AHA/ACC/AANP/AAPA/ABC/ACCP/ACPM/AGS/AMA/ASPC/NMA/PCNA/SGIM guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *Hypertension*, vol. 82, pp. e212–e316, 2025.

[7] G. Mancia, R. Kreutz, M. Brunström, M. Burnier, G. Grassi, A. Januszewicz *et al.*, "2023 ESH guidelines for the management of arterial hypertension: The task force for the management of arterial hypertension of the European Society of Hypertension, endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA)," *J. Hypertens.*, vol. 41, no. 12, pp. 1874–2071, Dec. 2023.

[8] A. A. Brandão, A. Alessi, A. M. Feitosa, C. A. Machado, C. E. P. de Figueiredo, C. Amodeo *et al.*, "6ª diretrizes de monitorização ambulatorial da pressão arterial e 4ª diretrizes de monitorização residencial da pressão arterial," *Arq. Bras. Cardiol.*, vol. 110, no. 5, pp. 1–29, May 2018.

[9] S. Omboni, G. Bilo, F. Saladini, A. Di Guardo, P. Palatini, G. Parati *et al.*, "Standards for the implementation, analysis, interpretation, and reporting of 24-hour ambulatory blood pressure monitoring: Recommendations of the Italian Society of Hypertension," *High Blood Press. Cardiovasc. Prev.*, vol. 31, no. 5, pp. 425–436, 2024.

[10] Health Level Seven International, "HL7 FHIR Release 4 (R4) Standard," 2019, accessed: Apr. 22, 2026. [Online]. Available: https://www.hl7.org/fhir/

[11] D. Bender and K. Sartipi, "HL7 FHIR: An agile and RESTful approach to healthcare information exchange," in *Proc. IEEE Int. Conf. Comput.-Based Med. Syst.*, 2013, pp. 326–331.

[12] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, "SMART on FHIR: A standards-based, interoperable apps platform for electronic health records," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 5, pp. 899–908, Sep. 2016.

[13] J. Adler-Milstein and A. K. Jha, "Health information exchange among U.S. hospitals: Who's in, who's out, and why?" *Healthc. (Amst.)*, vol. 2, no. 1, pp. 26–32, Jun. 2014.

[14] World Health Organization, *WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening*. Geneva, Switzerland: World Health Organization, 2019.

[15] M. Shakir, A. Arshad, M. O. Tariq, U. Sadiq, and U. Shabbir, "Development of IoT-based smart system and data acquisition for patient monitoring," in *Proc. 2024 Int. Conf. Engineering and Emerging Technologies (ICEET)*, 2024.

[16] J. Arnesen, A. Elvevold, D. Ghose, and Y. Lin, "Implementation of a modular MIoT platform for real-time vital sign monitoring," in *Proc. 2025 IEEE 13th Int. Conf. Healthcare Informatics (ICHI)*, 2025.

[17] V. H. Y. Teo, S. H. Teo, S. M. Burkill, Y. Wang, E. A. L. Chew, D. W. L. Ng, W. E. Tang, and G. C. H. Koh, "Effects of technology-enabled blood pressure monitoring in primary care: A quasi-experimental trial," *J. Telemed. Telecare*, vol. 30, no. 1, pp. 121–130, 2024, first published online in 2021.

[18] R. A. Fadel, J. Ross, T. Asmar, K. Sridasyam, Z. Demertzis, G. Ahluwalia, T. Roumayah *et al.*, "Visual analytics dashboard promises to improve hypertension guideline implementation," *Am. J. Hypertens.*, vol. 34, no. 10, pp. 1078–1082, 2021.

[19] T. Wang, A. A. Pileño, and E. Monacelli, "How can sensor data accessibility be enhanced in IoT-based health monitoring systems?" in *Proc. 2024 IEEE 19th Conf. Industrial Electronics and Applications (ICIEA)*, 2024.

[20] B. Nkoloogi, "An enhanced IoT-based wristband for remote monitoring and early detection of hypertension complications in uganda," The Nelson Mandela African Institution of Science and Technology, 2024.

[21] T.-W. Tseng, C.-F. Su, and F. Lai, "Fast healthcare interoperability resources for inpatient deterioration detection with time-series vital signs: Design and implementation study," *JMIR Med. Inform.*, vol. 10, no. 10, p. e42429, 2022.

[22] M. A. Badgeley, K. Shameer, B. S. Glicksberg, M. A. Tomlinson, M. A. Levin, P. J. McCormick, A. Kasarskis, D. L. Reich, and J. T. Dudley, "EHDViz: Clinical dashboard development using open-source technologies," *BMJ Open*, vol. 6, no. 3, p. e010579, 2016.

[23] D. Hix and H. R. Hartson, *Developing User Interfaces: Ensuring Usability Through Product and Process*. New York, NY, USA: John Wiley & Sons, 1993.

[24] J. Nielsen, "10 usability heuristics for user interface design," 1994, accessed: Apr. 29, 2026. [Online]. Available: https://www.nngroup.com/articles/ten-usability-heuristics/

[25] J. M. Vieira, J. P. de Abreu, J. C. Machado, F. P. Itturriet, and A. L. D. M. Martins, "IF-Cloud: API FHIR para integração de projetos de saúde digital," *J. Health Inform.*, vol. 16, no. Especial, Nov. 2024, accessed: Apr. 25, 2026. [Online]. Available: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1340

[26] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV Clinical Database Demo, version 2.2," 2023, RRID:SCR_007345.

Deep Learning-Based Reconstruction of Degraded Electromyographic Signals: A Comparative Study of Autoencoder and U-Net Architectures

Mayara Vieira de Oliveira Veiga
Graduate Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
mayaravieiraoliveira@alunos.utfpr.edu.br

Amauri Amorin Assef
Graduate Program in Electrical and Computer Engineering (CPGEI), DAELT
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
amauriassef@utfpr.edu.br

Joaquim Miguel Maia
Graduate Program in Electrical and Computer Engineering (CPGEI), DAELN
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
joaquim@utfpr.edu.br

Abstract*—Electromyographic (EMG) signals are widely used in neuromuscular assessment, rehabilitation systems, and assistive technologies. However, acquisition artifacts, electrode displacement, and stochastic noise frequently degrade signal quality, compromising subsequent analysis and interpretation. This work investigates the reconstruction of degraded EMG signals using convolutional Autoencoder and U-Net architectures with increasing levels of complexity. A supervised reconstruction framework was developed using processed ALSNet/EMGLab recordings, where an artificial degradation pipeline was employed to generate paired degraded-reference signals for training and evaluation. Training, validation, and testing were separated at the recording level before window generation, and the final evaluation used 355 normal-EMG recordings from the official test set. Reconstruction performance was assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and Signal-to-Noise Ratio (SNR), together with RMS, median-frequency, and envelope-based metrics. Among the evaluated models, U-Net V3.1 achieved the lowest MSE of 0.000634, RMSE of 0.02314, and MAE of 0.01573, as well as the highest mean SNR of 15.54 dB. AE V1 achieved a slightly higher mean PCC than U-Net V3.1 (0.99400 versus 0.99324). Statistical analysis using Wilcoxon signed-rank tests with Holm correction identified significant differences between U-Net V3.1 and the baseline architecture for most evaluated metrics (adjusted $p < 0.001$), whereas median-frequency error did not differ significantly. These findings indicate that skip-connection-based architectures may provide advantages in reconstruction error and SNR under the evaluated artificial degradation protocol. Their applicability to future neuromuscular monitoring and assistive biomedical applications requires further validation with real artifacts and external datasets.

***Index Terms*—electromyography, EMG reconstruction, deep learning, autoencoder, U-Net, biomedical signal processing**

I. Introduction

Electromyographic (EMG) signals play a fundamental role in biomedical applications, including the assessment of neuromuscular disorders, the analysis of human motor function, rehabilitation systems, and assistive technologies. Because EMG signals provide a direct representation of muscle activation patterns, they have become an important source of information for clinical evaluation and intelligent biomedical systems [1], [2].

However, the quality of EMG recordings is frequently compromised by acquisition artifacts, motion-induced disturbances, electrode displacement, power-line interference, and stochastic noise contamination. Such degradations can distort relevant physiological information and negatively affect subsequent signal processing, clinical interpretation, and machine learning applications [3], [4]. Consequently, reconstruction of degraded EMG signals has become an active area of research [5], [6].

Recent advances in deep learning have introduced data-driven alternatives to conventional denoising and filtering approaches. Convolutional Neural Networks (CNNs), Denoising Autoencoders, and related encoder-decoder architectures can learn nonlinear mappings between degraded and reference signals, reducing the dependence on handcrafted feature extraction [3], [7]. Deep neural networks have been applied to biomedical signal classification, pattern recognition, and signal enhancement [1], [7], [8].

Despite these advances, most studies that involve EMG signals remain focused on classification, gesture recognition, motor intention decoding, or disease diagnosis. Relatively limited attention has been devoted to systematically investigating how architectural complexity affects the quality of the reconstruction of degraded EMG signals. Recent studies suggest that increasing model complexity does not necessarily improve signal fidelity or reconstruction performance, highlighting the need for comparative evaluations of different architectures [5], [9]. This gap motivates the present work, which evaluates convolutional Autoencoder and one-dimensional U-Net architectures for reconstructing degraded EMG signals. The main contribution is a controlled comparison on the same processed ALSNet/EMGLab recordings, artificial degradation procedure,

leakage-free data partition, and complete-record evaluation. The conclusions are restricted to the synthetic corruption distributions and dataset evaluated here.

II. Related Work

A. EMG Signal Reconstruction

Electromyographic (EMG) signals are sensitive to acquisition-related disturbances, including motion artifacts, electrode displacement, power-line interference, and stochastic noise. These degradations can alter signal morphology and spectral characteristics, compromising subsequent analyses and reducing the reliability of clinical and research applications. Consequently, the reconstruction and enhancement of degraded EMG signals have become topics in biomedical signal processing [1], [10], [11].

Historically, signal restoration has relied on conventional denoising techniques such as filtering methods, Independent Component Analysis (ICA), Canonical Correlation Analysis (CCA), and Empirical Mode Decomposition (EMD). These approaches depend on assumptions regarding signal stationarity, source independence, or noise distributions that may not hold in real-world physiological recordings. Their ability to recover complex signal distortions may therefore be limited when dealing with non-stationary EMG signals [1], [4].

B. Deep Learning for Biomedical Signal Restoration

Recent advances in deep learning have introduced data-driven alternatives capable of learning nonlinear relationships directly from raw physiological signals. Unlike traditional signal processing techniques, deep neural networks extract hierarchical representations without requiring handcrafted features [1]. Recent reviews have described the use of deep learning for electromyographic signal analysis in human-machine interaction, prosthetic control, and rehabilitation systems [10], [11].

Several studies have evaluated deep learning models for biomedical signal denoising, enhancement, and reconstruction. Convolutional architectures have been used to model temporal dependencies and suppress noise sources in physiological signal restoration tasks [3], [4], [12]. Similar encoder-decoder strategies have been applied to physiological signal restoration and artifact removal. For example, U-Net-based denoising Autoencoders have been evaluated for reconstructing corrupted EEG recordings [13]. Deep learning has also been applied to signal enhancement, representation learning, and biomedical data analysis [6], [12].

C. Autoencoder and U-Net Architectures

Among deep learning approaches, Autoencoders have been used for signal reconstruction. By learning compact latent representations of degraded signals, these models can suppress noise while retaining temporal and spectral information. Previous studies have applied convolutional Autoencoders to the compression, reconstruction, and enhancement of myoelectric signals [7]–[9].

U-Net-based architectures have also been examined in restoration tasks because they combine an encoder-decoder structure with skip connections. Originally proposed for biomedical image segmentation [14], U-Net architectures propagate features from the encoder to corresponding decoder layers. This mechanism is intended to mitigate information loss caused by bottleneck compression and support the recovery of fine-grained signal characteristics.

Recent investigations have extended U-Net concepts to biomedical signal enhancement and restoration, reporting differences in reconstruction metrics relative to conventional encoder-decoder architectures [3], [5], [12]. Nevertheless, most EMG-related studies remain focused on classification, gesture recognition, motor intention decoding, or assistive control systems. Fewer studies investigate the reconstruction of degraded EMG signals, and few provide systematic comparisons between architectures of different complexity within a unified evaluation framework.

Motivated by these observations, the present study investigates how architectural design influences EMG reconstruction performance by comparing convolutional Autoencoder and U-Net architectures within a common experimental framework. Unlike restoration studies based on multi-channel HD-sEMG spatial structure or different electrophysiological modalities, the present comparison uses single-record processed ALSNet/EMGLab signals, a common artificial corruption model, and complete-record testing.

III. Materials and Methods

A. Materials

The proposed reconstruction framework was implemented in Python 3.11.15 using TensorFlow 2.21.0 and Keras 3.14.1. Data preprocessing, signal manipulation, model training, and performance evaluation were performed using NumPy 2.4.6, SciPy 1.17.1, Pandas 3.0.3, and Scikit-Learn 1.9.0. All experiments were conducted under Ubuntu/WSL2 with CUDA-enabled GPU acceleration.

All experiments were conducted on a workstation equipped with an Intel Core i7-9700F processor, 32 GB of RAM, and a NVIDIA GeForce RTX 5070 GPU with 12 GB of dedicated memory. Model development and training were performed using TensorFlow/Keras with GPU acceleration enabled through CUDA.

The overall experimental workflow adopted in this study is illustrated in Fig. 1.

B. Dataset

The experiments were conducted using the ALSNet/EMGLab dataset, a publicly available collection of electromyographic (EMG) recordings derived from the N2001 EMGLAB database. This dataset was originally employed in the development of ALSNet, a dilated one-dimensional convolutional neural network designed for the identification of Amyotrophic Lateral Sclerosis (ALS) from raw EMG signals [2].

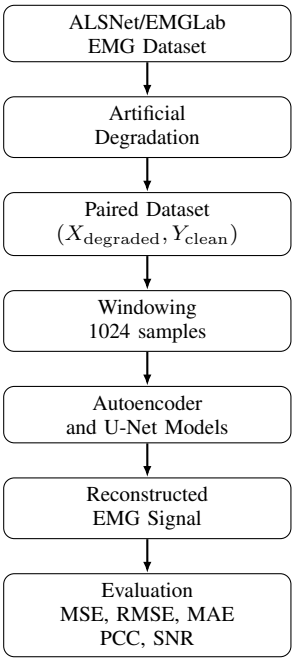


Fig. 1. Conceptual overview of the proposed EMG reconstruction framework.

The distributed processed package contains an official training file with 2,657 recordings and an official test file with 665 recordings. Each recording consists of 23,437 temporal samples acquired at 24 kHz. The present reconstruction study used only normal-EMG recordings (class 0): 1,306 from the official training file and 355 from the official test file [2].

Unlike previous studies that focused on disease identification, the present work addresses a supervised EMG reconstruction problem. The 1,306 normal recordings from the official training file formed the training and validation subsets, whereas the 355 normal recordings from the official test file were retained exclusively for final evaluation. For each reference signal, an artificially degraded counterpart was created using the degradation pipeline described in the following subsection, yielding a dataset of paired samples $(X_{degraded}, Y_{clean})$ suitable for supervised learning.

The resulting dataset served as the foundation for training and evaluating deep learning models under controlled degradation conditions. The processed files preserve recording indices but not explicit participant identifiers; therefore, the official train/test separation is respected, but subject-level independence cannot be independently verified and is treated as a limitation.

C. Artificial Signal Degradation Pipeline

Since the original dataset does not provide paired degraded-reference recordings suitable for supervised reconstruction, an artificial degradation pipeline was developed to generate reconstruction targets. Controlled degradation strategies are commonly adopted in signal restoration studies because they allow the evaluation of reconstruction models under reproducible corruption conditions [3]–[5], [13], [15].

Motivated by controlled synthetic-corruption procedures adopted in previous signal-restoration studies [3]–[5], [13], [15], the degradation model used in this study was defined as:

$$x[n] = g\, m[n] s[n] + \epsilon[n]. \tag{1}$$

In this equation, n denotes the sample index, $s[n]$ is the clean reference EMG signal, and $x[n]$ is its artificially degraded counterpart. The record-level attenuation factor was sampled as $g \sim \mathcal{U}(0.35, 0.75)$; $m[n]$ was an independent Bernoulli keep mask with probability 0.95; and $\epsilon[n]$ was zero-mean Gaussian noise with standard deviation $0.05\,\mathrm{SD}(s)$, where $\mathrm{SD}(s)$ denotes the standard deviation of the clean recording. Thus, the implemented masking removed individual samples independently rather than contiguous temporal segments. Separate seeds 4201, 4202, and 4203 were used for training, validation, and test corruption, respectively..

The resulting dataset consisted of paired samples $(X_{\text{degraded}}, Y_{\text{clean}})$, where the degraded signal served as the network input and the corresponding original recording represented the reconstruction target. This supervised formulation enabled direct optimization of the models for EMG signal restoration.

Although artificially generated artifacts may not fully represent all sources of variability encountered in real-world EMG acquisitions, controlled degradation protocols provide a practical and widely adopted framework for training and evaluating reconstruction models under reproducible conditions. Similar strategies have been employed in signal restoration and denoising studies, where synthetic corruption mechanisms enable systematic performance assessment and direct comparison between competing architectures [13], [15].

D. Window Generation and Data Preparation

To increase the number of training examples and facilitate mini-batch optimization, the paired signals were segmented into fixed-length windows of 1,024 samples. Window-based processing is commonly adopted in EMG and biomedical signal analysis because it preserves local temporal structures while reducing memory requirements and computational complexity during training [1], [7].

Partitioning was completed before window generation. The procedure produced 22,990 training windows from 1,045 recordings, 5,742 validation windows from 261 recordings, and 7,810 test windows from the 355 official-test recordings. Each degraded–reference pair was divided by the maximum absolute amplitude of its clean recording. This reference-based normalization defines a controlled paired reconstruction task and is not a deployment-ready preprocessing procedure because a clean reference would not be available for an unpaired real-world input. Training and validation used 22 complete non-overlapping windows per recording; complete-record testing zero-padded each signal to 23 windows and trimmed the concatenated output to the original 23,437 samples.

E. Deep Learning Architectures

The primary objective of this study was to investigate how architectural complexity influences the reconstruction quality of degraded EMG signals. To this end, two encoder-decoder paradigms were evaluated: a one-dimensional convolutional Autoencoder and a one-dimensional U-Net architecture.

Autoencoder-based frameworks have been widely adopted for signal reconstruction because they learn compact latent representations that suppress degradation artifacts while preserving relevant temporal information [7]–[9]. In this study, the convolutional Autoencoder was implemented as the baseline reconstruction model. The encoder consisted of successive one-dimensional convolutional and pooling operations that progressively compressed the input signal into a latent representation. In contrast, the decoder reconstructed the signal through upsampling and convolutional layers.

To address the information bottleneck associated with conventional encoder-decoder architectures, U-Net-based models were also investigated. Originally introduced for biomedical segmentation tasks, U-Net architectures establish direct connections between the encoder and decoder stages via skip connections, thereby preserving high-resolution information throughout the reconstruction process [14]. In one-dimensional signal reconstruction, this mechanism facilitates feature reuse and helps preserve fine-grained temporal characteristics that may be lost during progressive downsampling operations. Recent studies have shown that U-Net-based approaches are effective for EMG enhancement and restoration tasks, particularly under non-stationary noise and signal degradation conditions [3], [5].

The comparison between these architectures was designed to evaluate whether skip connections and residual blocks produce measurable differences in reconstruction metrics under the specified protocol. A summary of the evaluated architectures and their corresponding numbers of trainable parameters is presented in Table I.

TABLE I
REPRODUCIBLE SUMMARY OF EVALUATED ARCHITECTURES

Model	Filters	Kernel	Main feature	Parameters
AE V1	16–32–64–32	9	2-stage AE	42,049
AE V2.1	32–64–128–256–128–64–32	9	BN, dropout, hybrid loss	1,758,657
AE V2.2	32–64–128–256–128–64–32	9	BN, dropout, MSE	1,758,657
U-Net V3	32–64–128–256–128–64–32	9	concatenated skips	1,952,193
U-Net V3.1	32–64–128–256–128–64–32	15	residual blocks and skips	3,357,217

The evaluated models followed a progressive architectural development strategy. AE V1 used two encoder/decoder stages and ReLU activations. AE V2.1 and AE V2.2 used three stages, batch normalization, dropout, and the same filter sequence; they differed in their loss functions. U-Net V3 added concatenated encoder–decoder skip connections. U-Net V3.1 used kernel size 15 and residual blocks, each comprising two Conv1D operations with batch normalization and LeakyReLU ($\alpha = 0.1$), plus a 1×1 projection when channel dimensions differed.

To overcome the information bottleneck limitations commonly associated with conventional encoder-decoder architectures, one-dimensional U-Net models were investigated. U-Net V3 introduced skip connections between encoder and decoder stages, enabling the preservation of high-resolution temporal features throughout the reconstruction process and increasing the model capacity to approximately 1.95 million parameters.

Finally, U-Net V3.1 incorporated residual learning blocks within the encoder-decoder framework. The resulting architecture contained approximately 3.36 million trainable parameters and was the most complex configuration evaluated.

F. Training Procedure

Model training was performed using TensorFlow/Keras and optimized with the Adam algorithm [16], selected for its adaptive learning-rate mechanism and robust convergence in deep neural network optimization. All models received degraded EMG windows as input and generated reconstructed EMG signals as output. The optimization objective was to minimize the discrepancy between the reconstructed and reference signals.

AE V2.1 was optimized with a hybrid loss comprising MSE plus 0.005 times an RFFT-magnitude loss. AE V1, AE V2.2, U-Net V3, and U-Net V3.1 used MSE. The frequency-domain component was motivated by the importance of spectral information in EMG restoration [5].

The training hyperparameters adopted throughout the experiments are summarized in Table II.

TABLE II
TRAINING HYPERPARAMETERS

Hyperparameter	Value
Window Length	1024 samples
Input Dimension	1024×1
Optimizer	Adam
Initial Learning Rate	0.001
Validation Set	261 recordings (before windowing)
Early Stopping Monitor	Validation Loss
Early Stopping Patience	8 epochs
Restore Best Weights	Enabled
ReduceLROnPlateau	factor 0.5; patience 4; min. 10^{-6}
Maximum/Completed Epochs	50 / 26, 18, 21, 22, 33
AE V1 Batch Size	32
AE V2.1/V2.2 Batch Size	8
U-Net V3/V3.1 Batch Size	8
AE V1/V2.2 Loss Function	MSE
AE V2.1 Loss Function	MSE + 0.005 spectral loss
U-Net V3/V3.1 Loss Function	MSE
Framework	TensorFlow/Keras

To manage optimization, Early Stopping and ReduceLROnPlateau were employed based on validation loss monitoring, as summarized in Table II.

To evaluate model performance, the official test recordings were kept outside optimization and model selection. This supports evaluation on held-out recordings under the same artificial protocol, not under real acquisition conditions.

G. Experimental Protocol

The 1,306 eligible recordings in the official training file were randomly divided, with seed 42, into 1,045 training

and 261 validation recordings before window generation. Validation recordings controlled early stopping and learning-rate scheduling. The 355 eligible recordings in the official test file were never used for optimization or model selection. Consequently, no windows from one recording occurred in more than one subset. Because participant identifiers were unavailable, residual dependence between different recordings from the same individual cannot be excluded. The principal results below use complete recordings as the statistical unit.

H. Evaluation Metrics

Reconstruction performance was assessed through complementary error, similarity, and signal-quality metrics. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were employed to quantify reconstruction error. Pearson Correlation Coefficient (PCC) was used to evaluate waveform similarity between reconstructed and reference signals, while Signal-to-Noise Ratio (SNR) quantified signal quality and was computed as [3]:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} y[n]^2}{\sum_{n=1}^{N} \left(y[n] - \hat{y}[n] \right)^2} \right). \tag{2}$$

In this equation, $y[n]$ denotes the n-th sample of the clean reference EMG signal, $\hat{y}[n]$ denotes the corresponding sample of the reconstructed signal, and N is the total number of samples in the evaluated recording.

Because all EMG signals were normalized before training and evaluation, MSE, RMSE, and MAE are reported in normalized units. PCC is a dimensionless similarity coefficient, while SNR is expressed in decibels (dB).

Two nonlearned baselines were included: a Wiener filter with a nine-sample window and a Savitzky–Golay filter with an 11-sample window and polynomial order 3. To examine signal descriptors, relative RMS error, absolute median-frequency error, envelope PCC, and envelope normalized RMSE were also calculated. These descriptors provide amplitude, spectral, and envelope checks but do not establish preservation of motor-unit characteristics.

I. Statistical Analysis

AE V1 and U-Net V3.1 were compared using two-sided paired Wilcoxon signed-rank tests, with each of the 355 complete test recordings as one observation [17]. The nine metric-wise p-values were adjusted by Holm's procedure [18]; $\alpha = 0.05$ was applied to adjusted p-values. Rank-biserial correlation was reported as effect size, with its sign oriented so that positive values favor U-Net V3.1.

IV. Results and Discussion

A. Quantitative Reconstruction Performance

Table III summarizes the complete-record reconstruction performance obtained on the 355-recording independent official test set.

The quantitative results varied with architectural design and were not ordered solely by network depth. Although AE V2.1 and AE V2.2 employed deeper encoder-decoder structures than the baseline AE V1, both architectures yielded higher reconstruction errors and lower correlation values.

TABLE III
Mean Complete-Record Performance on the Independent Test Set

Model	MSE	RMSE	MAE	PCC	SNR (dB)
Degraded input	0.004862	0.06552	0.04231	0.96707	6.29
Wiener (9)	0.004813	0.06519	0.04257	0.97832	6.34
Savitzky–Golay (11,3)	0.004679	0.06419	0.04243	0.98942	6.48
AE V1	0.000727	0.02443	0.01706	**0.99400**	15.21
AE V2.1	0.001115	0.03141	0.02053	0.98425	12.66
AE V2.2	0.001154	0.03095	0.02076	0.98937	13.03
U-Net V3	0.000787	0.02561	0.01721	0.99353	14.75
U-Net V3.1	**0.000634**	**0.02314**	**0.01573**	0.99324	**15.54**

U-Net V3.1 obtained the lowest mean MSE, RMSE, and MAE and the highest mean SNR. Relative to AE V1, its mean MSE was 12.8% lower and its SNR was 0.33 dB higher. However, AE V1 obtained the highest PCC (0.99400 versus 0.99324), showing that the ranking depends on the metric and that the absolute differences between the two configurations are modest. All learned models produced lower error than the degraded input and the two filtering baselines under this particular corruption model.

These results are compatible with a benefit from multi-scale feature reuse, but the controlled comparison does not isolate skip connections from every other architectural difference [3], [5], [13].

This behavior is compatible with reports on the use of multi-scale feature fusion and skip connections in biomedical applications [14], [19].

Accordingly, U-Net V3.1 is described as the configuration with the lowest mean errors and highest mean SNR among those tested, while AE V1 retained the highest mean PCC.

The ALSNet/EMGLab dataset used in this study has previously been used in deep learning studies involving EMG signal analysis and classification [2]. The adoption of this publicly available dataset facilitates reproducibility and enables a more direct comparison with related studies in the literature.

Directly comparable supervised reconstruction benchmarks using the same processed ALSNet/EMGLab recordings and degradation protocol were not identified; therefore, numerical comparisons with studies using other datasets or corruption procedures should be interpreted cautiously.

Direct numerical comparison with published restoration studies remains difficult because datasets, artifacts, and evaluation units differ. The present evidence therefore supports only performance on held-out ALSNet/EMGLab recordings subjected to the specified artificial degradation.

B. Signal-Level Evaluation

To assess reconstruction performance without treating correlated windows as independent, complete EMG recordings were reconstructed and evaluated at the signal level. Table IV summarizes the metrics across the 355 independent official-test recordings.

TABLE IV
SIGNAL-LEVEL RECONSTRUCTION PERFORMANCE OF U-NET V3.1

Statistic	MSE	RMSE	MAE	PCC	SNR (dB)
Mean	0.000634	0.02314	0.01573	0.99324	15.54
Median	0.000440	0.02097	0.01409	0.99375	15.56
Std	0.000603	0.00995	0.00753	0.00306	3.13

The signal-level evaluation provides information about performance on complete held-out recordings. While window-based evaluation measures local reconstruction quality, signal-level analysis reflects cumulative reconstruction performance over complete recordings.

The mean PCC of 0.99324 indicates high waveform agreement across the 355 recordings under the controlled synthetic protocol. It does not establish agreement across subjects because participant identifiers were unavailable.

The proximity of the mean and median across the evaluated metrics suggests that performance was not dominated by a small number of recordings with particularly low error.

Additionally, the lower median MSE relative to the mean MSE indicates that a limited number of challenging recordings contributed disproportionately to the overall error distribution. In contrast, most signals were reconstructed with lower error.

TABLE V
SIGNAL-DESCRIPTOR ERRORS ON 355 TEST RECORDINGS

Descriptor	AE V1	U-Net V3.1
RMS relative error	0.1445 ± 0.0902	0.1277 ± 0.0816
MDF absolute error (Hz)	7.50 ± 14.76	7.42 ± 13.89
Envelope PCC	0.9929 ± 0.0039	0.9904 ± 0.0037
Envelope NRMSE	0.1753 ± 0.0757	0.1676 ± 0.0656

U-Net V3.1 had lower mean RMS-relative and envelope-normalized errors; median-frequency errors were similar, while AE V1 had higher envelope PCC. These complementary checks do not demonstrate preservation of motor-unit characteristics, which cannot be reliably derived from the processed single-record representation available here.

C. Qualitative Signal Reconstruction Analysis

Figure 2 presents a representative reconstruction example obtained from an unseen EMG recording.

The qualitative example shown in Fig. 2 complements the quantitative analysis by illustrating the visual similarity between reconstructed and reference signals. Despite degradation in the input signal, the reconstructed waveform follows the principal temporal fluctuations of the reference recording and preserves its overall morphology.

Although visual inspection alone cannot replace quantitative evaluation, the qualitative results support the numerical findings reported in Tables III and IV. The close agreement between the reconstructed and reference waveforms is consistent with the high PCC values observed across the evaluation set.

D. Statistical Analysis

To determine whether the observed differences were statistically significant, paired Wilcoxon signed-rank tests were performed comparing AE V1 and U-Net V3.1 across the nine evaluated metrics. Table VI presents the results for the five primary metrics. Among the four signal descriptors, median-frequency error was the only metric without a statistically significant difference after Holm correction (adjusted $p = 0.326$).

The Holm-adjusted tests identified differences for the five primary metrics, but directions were mixed: U-Net V3.1 was favored for error and SNR, whereas AE V1 was favored for PCC. Statistical significance is interpreted together with effect size and the modest absolute differences.

TABLE VI
WILCOXON SIGNED-RANK TEST RESULTS (AE V1 VS. U-NET V3.1)

Metric	Holm p	Effect	Direction
MSE	5.36×10^{-10}	0.395	V3.1
RMSE	1.75×10^{-8}	0.359	V3.1
MAE	1.22×10^{-19}	0.569	V3.1
PCC	6.23×10^{-13}	-0.455	AE V1
SNR	1.28×10^{-5}	0.282	V3.1

E. Additional-Noise Sensitivity Analysis

The sensitivity of U-Net V3.1 was evaluated under additional Gaussian-noise conditions not observed during training. Table VII summarizes the quantitative results, while Fig. 3 presents the changes under increasing noise levels.

TABLE VII
SENSITIVITY EVALUATION UNDER ADDITIONAL GAUSSIAN-NOISE LEVELS

Added Gaussian SD	PCC	SNR (dB)
0.00	0.9932	15.54
0.01	0.9913	14.96
0.03	0.9670	11.41
0.05	0.9167	7.85

As expected, reconstruction performance gradually decreased with increasing noise intensity. The experiment quantifies sensitivity to additional Gaussian corruption only. It cannot be interpreted as evidence of robustness to electrode motion, displacement, power-line interference, or other real acquisition artifacts.

F. Limitations

The study is limited to artificial degradation of processed normal-EMG recordings. No independently acquired set with real artifacts was available, participant identifiers were not preserved for subject-level splitting, and normalization used the paired clean reference. The classical filters provide transparent baselines but do not represent every recent restoration method. RMS, median-frequency, and envelope analyses add signal-level context, whereas clinical validity and motor-unit preservation remain untested. External datasets with real artifacts

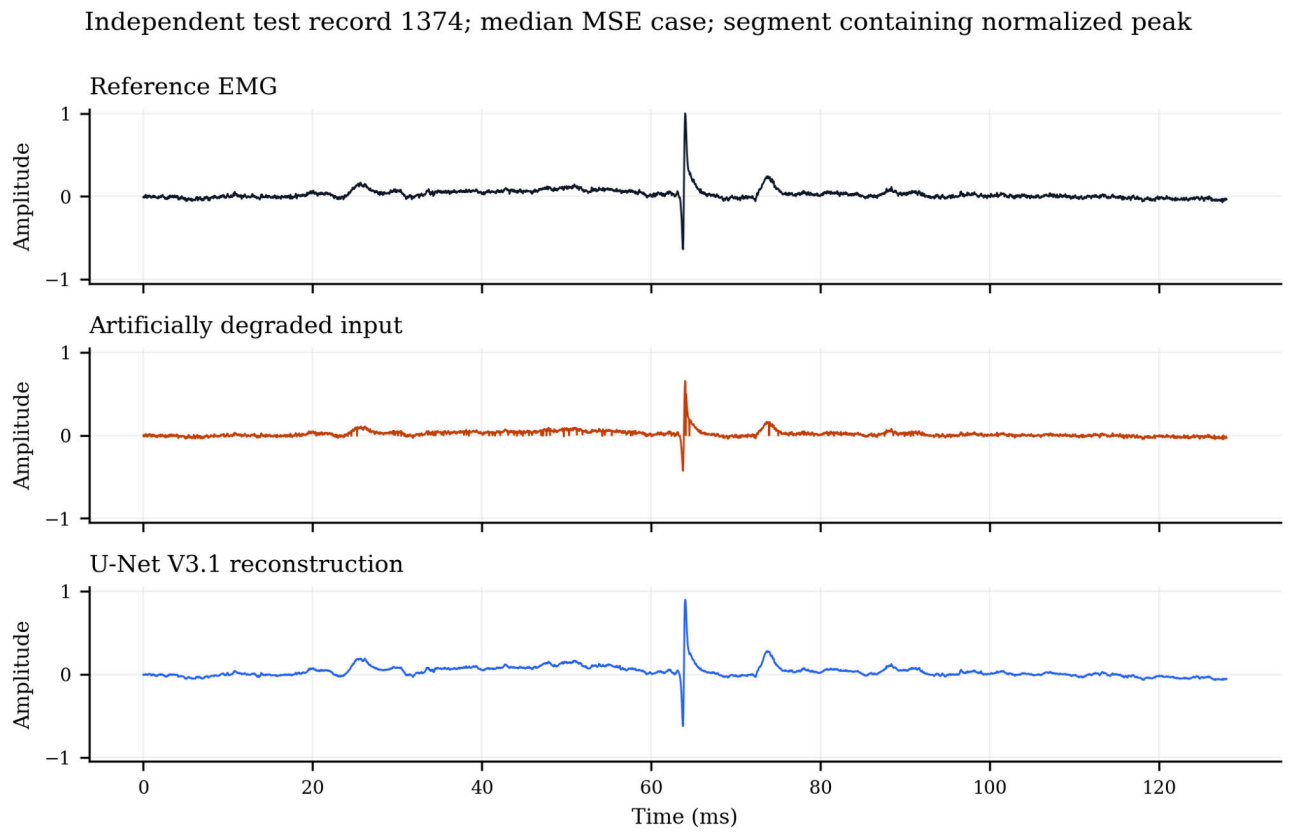


Fig. 2. Representative U-Net V3.1 reconstruction from the independent official test set. A median-MSE recording was selected, and the displayed segment contains the reference signal's normalized peak of 1, following the intended normalized-amplitude presentation.

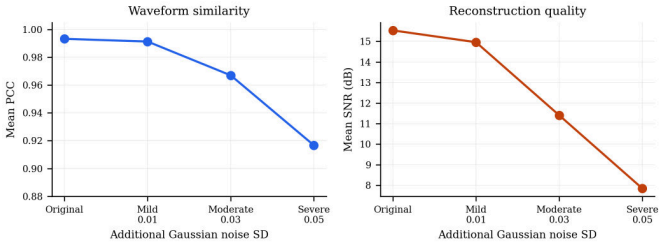


Fig. 3. Sensitivity of U-Net V3.1 to additional zero-mean Gaussian noise on the independent test set.

and participant-level metadata are therefore required before clinical or deployment claims can be made.

V. Conclusion

This study investigated the reconstruction of degraded electromyographic signals using convolutional Autoencoder and one-dimensional U-Net architectures. A controlled comparison was conducted with an explicitly specified artificial degradation pipeline and record-level partitioning before window generation.

Among the tested configurations, U-Net V3.1 obtained the lowest complete-record errors and highest SNR, while AE V1 obtained a slightly higher PCC. These results apply to this dataset and artificial protocol and do not establish clinical validity.

Signal-level evaluation on the independent official test recordings showed high agreement with the artificial reference targets. Additional-noise experiments characterized sensitivity within Gaussian conditions rather than real-world robustness.

Future work should prioritize real electrode artifacts, external datasets, verified subject-level separation, and comparisons with established and recent restoration methods before considering clinical, monitoring, assistive, or embedded applications.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES),

Brazil – Finance Code 001. The authors would like to acknowledge the support provided by the National Council for Scientific and Technological Development (CNPq), the Federal University of Technology – Paraná (UTFPR), the Araucária Foundation, the State Secretariat for Science, Technology, and Higher Education of Paraná (SETI), and the National Center for High-Performance Computing in São Paulo at the State University of Campinas (CENAPAD-SP/UNICAMP).

REFERENCES

[1] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for EMG-based human-machine interaction: A review," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 3, pp. 512–533, 2021.

[2] K. M. N. Hassan, M. S. A. Hridoy, N. Tasnim, A. F. Chowdhury, T. A. Roni, S. Tabrez, A. Subhana, and C. Shahnaz, "ALSNet: A dilated 1-D CNN for identifying ALS from raw EMG signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1181–1185.

[3] K.-C. Wang, K.-C. Liu, P.-C. Yeh, S.-Y. Peng, and Y. Tsao, "TrustEMG-Net: Using representation-masking transformer with U-Net for surface electromyography enhancement," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 4, pp. 2506–2520, 2025.

[4] M. Ergeneci, D. Binningsley, D. Carter, and P. Kosmas, "A novel framework for motion-induced artifact cancellation in sEMG: Evaluation on English Premier League and Ninapro datasets," *IEEE Sensors Journal*, vol. 24, no. 14, pp. 22 610–22 619, 2024.

[5] J. Mao, H. Li, and Y. Zhao, "An innovative deep learning-driven technique for restoration of lost high-density surface electromyography signals," *Applied Intelligence*, vol. 55, p. 639, 2025.

[6] Y. Zhao *et al.*, "HD-sEMG-CORE: An open-source hybrid network algorithm for efficient compression and accurate restoration of high-density surface electromyography signals," *IEEE Access*, vol. 13, pp. 45 559–45 567, 2025.

[7] K. Dinashi, A. Ameri, M. A. Akhaee, K. Englehart, and E. Scheme, "Compression of EMG signals using deep convolutional autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 2888–2897, 2022.

[8] N. R. Caballero, M. R. Gonzalez, M. Aviles, J. M. A. Alvarado, J. B. Robles-Ocampo, P. Y. Sevilla-Camacho, and J. Rodriguez-Resendiz, "Effective approach for classifying EMG signals through reconstruction using autoencoders," *AI*, vol. 7, no. 1, p. 36, 2026.

[9] C. Brambilla, N. Moscatelli, V. Lanzani, L. M. Tosatti, A. Brusaferri, and A. Scano, "The trade-off between maximizing reconstruction and physiological interpretation of muscle synergies with autoencoders," *Frontiers in Human Neuroscience*, vol. 19, p. 1699799, 2025.

[10] W. Li, P. Shi, and H. Yu, "Gesture recognition using surface electromyography and deep learning for prostheses hand: State-of-the-art, challenges, and future," *Frontiers in Neuroscience*, vol. 15, p. 621885, 2021.

[11] T. Zaim, S. Abdel-Hadi, R. Mahmoud *et al.*, "Machine learning- and deep learning-based myoelectric control system for upper limb rehabilitation utilizing EEG and EMG signals: A systematic review," *Bioengineering*, vol. 12, no. 2, p. 144, 2025.

[12] X. Xu, N. Chen, Y. Deng, T. Shen, N. Wei, S. Yu, and L. Zhang, "Enhancing neural signal quality: A spike-denoising model with BiLSTM and attention mechanism," *Brain Informatics*, vol. 12, no. 1, p. 28, 2025.

[13] C.-H. Chuang, K.-Y. Chang, C.-S. Huang, and T.-P. Jung, "IC-U-Net: A U-Net-based denoising autoencoder using mixtures of independent components for automatic EEG artifact removal," *NeuroImage*, vol. 263, p. 119586, 2022.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 9351, 2015, pp. 234–241.

[15] Y. Pei, J. Xu, Q. Chen, C. Wang, F. Yu, L. Zhang, and W. Luo, "DTP-Net: Learning to reconstruct EEG signals in time-frequency domain by multi-scale feature reuse," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, pp. 2662–2673, 2024.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[18] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[19] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 076–10 095, 2024.

Detection and Segmentation of Nanoparticles in Electron Microscopy

1st Michael Almeida
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University – UNESP
Botucatu - SP, Brazil
mj.almeida@unesp.br

2nd Geovana Fioschi
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University – UNESP
Botucatu - SP, Brazil
geovana.r.fioschi@unesp.br

3rd Nycolas Mariotto
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University – UNESP
Botucatu - SP, Brazil
n.mariotto@unesp.br

4th Guilherme Soares
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University – UNESP
Botucatu - SP, Brazil
guilherme.soares@unesp.br

5th José Ricardo A. Miranda
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University – UNESP
Botucatu - SP, Brazil
jose.r.miranda@unesp.br

6th Allan F. F. Alves
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University – UNESP
Botucatu - SP, Brazil
allan.alves@unesp.br

***Abstract*—Automated nanoparticle quantification in transmission electron microscopy (TEM) images is challenging because of variations in contrast, high particle density, small object size, and time required for manual analysis. In this study, we developed a hybrid pipeline that combined neural detection using YOLO11x with classical image segmentation applied within the detected regions of interest. The dataset, exported in the COCO format, contained 1,845 images of 640×640 pixels and 14,829 bounding box annotations. Contrast-limited adaptive histogram equalization (CLAHE) was applied during preprocessing to reduce local contrast and illumination variations. YOLO11x was selected because of its higher representational capacity and the presence of small objects in the images. In the test set, the model achieved an mAP@50 of 66.3%, a precision of 72.5%, a recall of 72.5%, and an F1-score of 72.5%. After neural detection, classical operations, including thresholding, morphological processing, watershed segmentation, and ellipse fitting, were applied to each detected region. This strategy enables approximate particle contours to be obtained without requiring the complete dataset to be manually re-annotated with segmentation masks. The results demonstrate the feasibility of the proposed pipeline for automated nanoparticle localization and provide a practical basis for subsequent morphometric analyses.**

Keywords: Transmission electron microscopy, nanoparticles, YOLOv8, object detection, image segmentation.

I. Introduction

Nanoparticles have several biomedical and theranostic applications, including targeted drug delivery, diagnostic imaging, and the development of combined diagnostic and therapeutic strategies [1], [2]. Transmission electron microscopy (TEM) enables the visualization of these structures at the nanometer scale and can be used to investigate their morphology, distribution, aggregation, and internalization in cells [3].

Despite their relevance, the quantitative analysis of TEM images is frequently based on manual or semi-automated procedures. Manual particle identification is time-consuming, dependent on the evaluator's experience, and susceptible to intra- and inter-observer variability. Conventional thresholding methods can also be affected by variations in image contrast, background intensity, overlapping particles, and cellular structures with similar grayscale values.

Deep learning approaches have been successfully applied to particle analysis in electron microscopy. Neural networks can automate nanoparticle localization and segmentation while reducing the effort required for manual measurements [4], [5]. In cellular TEM images, YOLO-based models have also demonstrated strong potential for the automated detection of internalized gold nanoparticles [6].

The dataset used in the present study contained bounding box annotations but did not contain pixel-level segmentation masks. Consequently, directly training a fully supervised instance segmentation model requires the complete dataset to be manually reannotated. To address this limitation, a hybrid strategy was developed in which YOLO model first detects candidate particles, and subsequently, classical image-processing methods segment each detected region.

Therefore, the proposed method combines the robustness of neural object detection with the interpretability and reduced annotation requirements of classical segmentation. This study aimed to develop and evaluate a hybrid pipeline for detecting nanoparticles in TEM images and generating approximate particle contours for subsequent morphometric analysis.

II. Materials and Methods

A. Dataset

The dataset was obtained in the COCO format with 1,845 TEM images with a spatial resolution of 640×640 pixels. The images were divided into training, validation, and test subsets, comprising 1,293, 368, and 184 images, respectively.

A total of 14,829 bounding box annotations were assigned to the *Nanoparticle* class.

An initial audit of the COCO annotation files was performed to verify the number of images, annotations, categories, and segmentation. The categories were standardized into a single class before being converted to the YOLO format. Approximately 23.5% of the annotated objects had dimensions or areas corresponding to objects smaller than 32×32 pixels. This characteristic was considered during model selection and training because small particles provide limited spatial information and may be lost during successive feature map downsampling operations. Some images did not contain annotations and were retained to represent true-negative samples because background images can help reduce false-positive detections.

B. Image preprocessing

Figure 1 summarizes the proposed approach. The input images were converted to a standardized grayscale representation and normalized before the model training. Contrast-limited adaptive histogram equalization (CLAHE) was applied to improve the local contrast and reduce the effects of non-uniform illumination.

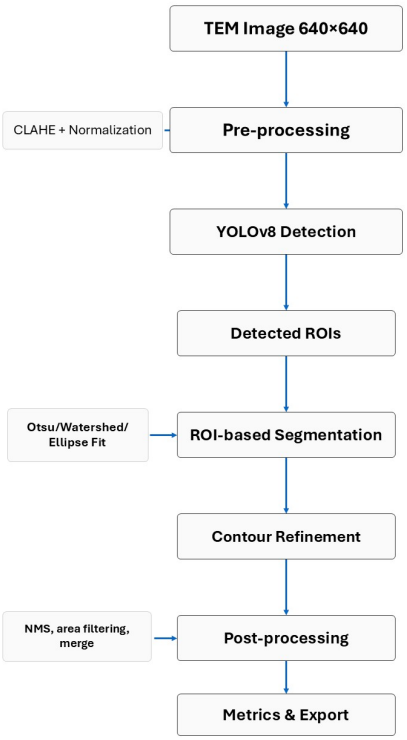


Fig. 1. Proposed pipeline for nanoparticle detection and segmentation.

CLAHE divides an image into local regions and redistributes pixel intensities within each region while limiting excessive contrast amplification. This procedure was selected because the TEM images presented differences in brightness and contrast related to sample thickness, electron density, acquisition conditions, and intracellular background structures.

The preprocessing stage also included automatic orientation verification and resizing to 640×640 pixels. Data augmentation was applied only to the training subset and included horizontal and vertical flips, rotations, moderate brightness and contrast adjustments, Gaussian noise, and blurring, following commonly adopted image augmentation strategies for improving model generalization [7].

Aggressive augmentations were avoided when they produced unrealistic particle morphologies or densities. Mosaic and copy-paste augmentation were applied cautiously because the artificial combination of highly populated TEM regions may generate distributions that are not consistent with original images.

C. Nanoparticle detection

YOLOv8m was used as the standard neural network detector. The medium version (YOLOv8m) was selected because it offered an optimal balance between representational capacity and computational efficiency, and the high proportion of small objects in the dataset. Among the YOLOv8 detection variants, YOLOv8m has a robust number of parameters ($\sim$25.8M), ensuring high representational capacity without the extreme computational cost of larger variants [8], [9].

To ensure reproducibility, the YOLOv8m model was trained for 300 epochs with early stopping patience of 50 epochs. The input image size was set to 640×640 pixels, and the batch size was automatically determined based on the GPU memory constraints (AutoBatch). The AdamW optimizer was used with an initial learning rate of 0.0005, momentum of 0.937, and weight decay of 0.001. Data augmentation strategies included mosaic (1.0), mixup (0.1), translation (0.15), scaling (0.4), and $\pm 90^{\circ}$ rotations, which helped the model generalize to the varying orientations and scales of the nanoparticles.

Transfer learning was performed using the pretrained model weights. During training, a validation subset was used to monitor model convergence and detect potential overfitting. The final performance was evaluated using an independent test subset that was not used for model optimization.

D. Classical segmentation within detected regions

Because the dataset did not contain segmentation masks, classical segmentation was applied after detection. For each bounding box predicted by YOLO, a region of interest was extracted with an additional margin to preserve the information around the particle boundaries.

The region was processed using global or local Otsu thresholding, depending on the intensity distribution [10]. Morphological opening was used to remove isolated pixels, whereas morphological closing and region filling were applied to correct the discontinuities within the particle region.

Watershed segmentation was used when multiple touching or overlapping particles were present in the same region of interest. Candidate components were filtered according to their

area, position, intensity, and compatibility with the expected particle geometry.

The selected particle contours were refined using morphological operations and ellipse fitting. Ellipse fitting was used as an optional contour approximation for particles with predominantly circular or elliptical morphologies. Irregular contours were retained when ellipse fitting introduced a substantial discrepancy from the threshold-derived boundaries.

This hybrid procedure generated approximate segmentation masks without requiring the complete dataset to be manually reannotated at the pixel level.

E. Post-processing and morphometric measurements

Non-maximum suppression was applied to reduce the number of duplicated neural detections. Additional filtering was performed based on the bounding box confidence, detected area, and spatial overlap.

After segmentation, connected components that were judged to belong to the same particle were merged when appropriate. Regions that did not satisfy the minimum area or morphological criteria were excluded as noise.

F. Performance evaluation

The detection performance was evaluated using the precision, recall, F1-score, and mean average precision at an intersection-over-union threshold of 0.50. Precision represents the proportion of predicted objects corresponding to the reference annotations, whereas recall indicates the proportion of annotated particles correctly detected by the model. The F1-score summarizes the balance between precision and recall. The segmentation performance was evaluated separately from the detection performance using a subset of test images that received manual pixel-level ground-truth annotations.
For each detected region, the predicted mask was compared with the corresponding manual mask using four metrics: The Dice similarity coefficient and the Intersection over Union (IoU), which quantify the pixel-wise spatial overlap between the predicted and reference regions, with values closer to 1 indicating a closer match. The area error measures the relative discrepancy between the predicted and reference particle areas independently of their spatial alignment, so that this metric can remain low even when the predicted and reference regions are shifted relative to each other. The boundary distance was computed as the mean of the average nearest-neighbor distances between the contour points in both directions, providing a measure of contour agreement, expressed in pixels. Together, these four metrics capture the complementary aspects of segmentation quality.

III. Results

The neural detection model achieved an mAP@50 of 68.1%, with precision, recall, and F1-score values of 74%. These results indicate a balanced detection performance, as the precision and recall reached similar values, suggesting that the model maintained a similar ability to avoid false-positive detections and identify annotated nanoparticles.

Table I summarizes the performance obtained using the test dataset.

TABLE I
PERFORMANCE OF THE NANOPARTICLE DETECTION MODEL ON THE TEST DATASET.

Model	mAP@50	Precision	Recall	F1-score
YOLO-based detector	0.681	0.748	0.741	0.745

The obtained mAP@50 indicates that the detector could localize a substantial proportion of nanoparticles when a minimum intersection-over-union threshold of 0.50 was considered. A precision of 74.8

To quantitatively validate the classical post-segmentation step, a subset of the test data was manually annotated using pixel-level ground-truth masks. The proposed hybrid approach (YOLOv8m + Otsu/Watershed) was evaluated using the Dice coefficient, Intersection over Union (IoU), area error, and average boundary distance (Table II).

TABLE II
QUANTITATIVE VALIDATION OF THE POST-DETECTION CLASSICAL SEGMENTATION.

Metric	Average Value
Dice Coefficient	0.577
Intersection over Union (IoU)	0.415
Area Error	40.6%
Boundary Distance	14.7 pixels

These preliminary results reflect the intrinsic challenges of applying classical thresholding methods to complex, noisy, and heterogeneously illuminated TEM backgrounds. Although approximate contours provide useful morphometric estimates, such as fitted ellipses, they do not perfectly replace supervised pixel-level segmentation. Furthermore, an ablation study confirmed that when classical segmentation is applied to the full TEM image without prior neural localization, it generates an excessive amount of background noise and false positives, merging adjacent particles. YOLO-based filtering is thus strictly necessary to define a constrained region of interest before applying any classical thresholding.

The representative outputs of the proposed hybrid pipeline are shown in Figure 2. The YOLO-based detector successfully localizes isolated nanoparticles and particle aggregates, whereas the post-detection segmentation stage generates approximate contours and elliptical representations within the detected regions.

IV. Discussion

The detection results demonstrate the feasibility of applying deep learning for the automated identification of nanoparticles in transmission electron microscopy images. The balanced precision and recall values suggest that the model did not strongly favor either excessive detections or overly conservative predictions. Other authors have achieved higher metrics [5], [6]; however, it is difficult to compare the results from different nanoparticles, TEM equipment, and dataset characteristics.

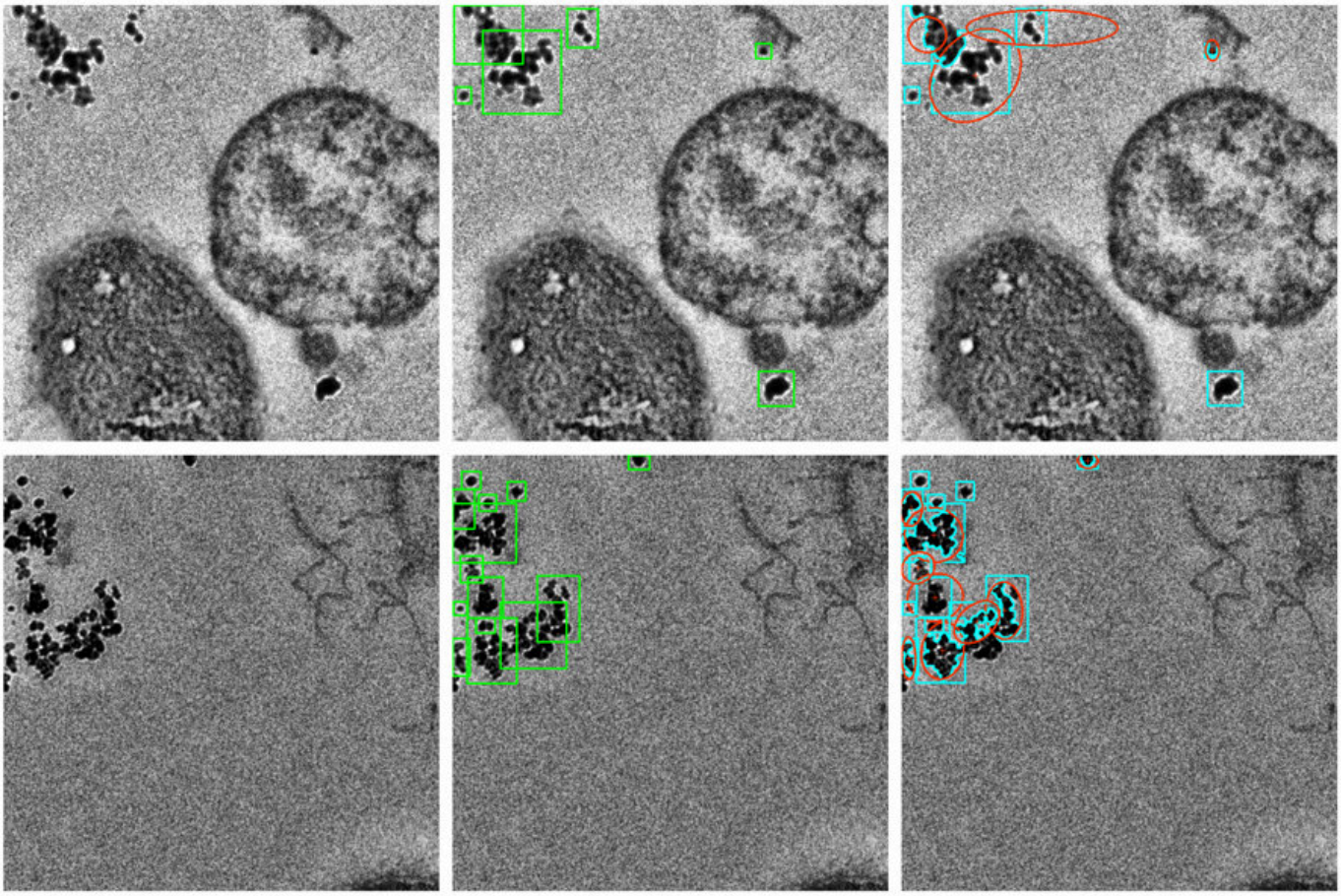


Fig. 2. Representative examples of nanoparticle detection and post-detection segmentation in transmission electron microscopy images. Each row corresponds to a different TEM sample. The first column shows the original images, the second column presents the bounding boxes predicted by the YOLOv8m detector, indicated in green, and the third column shows the results of the hybrid segmentation procedure applied within the detected regions of interest. Cyan contours represent the segmented particle boundaries, whereas red ellipses indicate the fitted geometric approximations.

A central aspect of the proposed approach is the use of contrast-limited adaptive histogram equalization (CLAHE) during image preprocessing. TEM images frequently exhibit local differences in brightness and contrast due to variations in sample thickness, electron density, acquisition conditions, and background illumination. CLAHE improves the local contrast while limiting excessive noise amplification, facilitating the distinction between the nanoparticles and the surrounding cellular structures.

The YOLOv8m architecture was selected as the primary detection model due to its high representational capacity. This choice is particularly relevant because 23.5% of the annotated objects in the dataset have areas smaller than 32×32 pixels. As indicated in the literature, small objects provide limited spatial information and may lose discriminative features across successive downsampling layers, making their detection more difficult at insufficient input or feature map resolutions [11].

Therefore, YOLOv8m was used to improve the feature extraction of small nanoparticles. To justify the extra computational cost, an empirical comparison was conducted with lighter versions of YOLOv8 trained under the exact same conditions. Although the medium-sized model offers a slight performance edge necessary for capturing the smallest nanoparticles, scaling up to extra-large variants does not yield significant improvements and only adds unnecessary computational overhead, confirming YOLOv8m as the optimal architectural choice. The hybrid pipeline also addresses an important limitation of the available datasets. Although the annotations contain bounding boxes, they do not provide pixel-level segmentation. Therefore, training a conventional instance segmentation network requires extensive manual re-annotation. By applying classical segmentation only within the detected regions, the proposed method reduces the search area and minimizes interference from unrelated structures in TEM images.

Otsu thresholding provides an initial separation between the foreground and background pixels, whereas morphological operations can remove isolated noise and fill discontinuities. Watershed segmentation may assist in separating adjacent or partially overlapping nanoparticles from each other. Finally, ellipse fitting provides a smoother approximation of the contour, which is consistent with the approximately circular or

elliptical morphologies of many nanoparticles.

The relatively low segmentation performance, reflected by a mean Dice coefficient of 0.577 and an IoU of 0.415, can be attributed to the high level of noise and background heterogeneity inherent to TEM images. These characteristics frequently generate small high-contrast structures and irregular intensity patterns resembling nanoparticles, particularly those with smaller diameters or a low degree of aggregation. Consequently, the classical segmentation step may partially include background structures as part of the nanoparticle regions or fail to accurately delineate weakly defined particle boundaries, contributing to the observed area error of 40.6% and mean boundary distance of 14.7 pixels. Therefore, these metrics should be interpreted in the context of challenging image characteristics rather than solely as limitations of the segmentation procedure.

Moreover, the accuracy of the classical segmentation stage depends on the image contrast, nanoparticle morphology, aggregation, and presence of intracellular structures with similar intensities. Detection errors are also propagated to the segmentation stage of the process. A missed nanoparticle cannot be recovered during post-processing, whereas an incorrect bounding box may generate a falsely segmented region. The current results should be interpreted primarily in terms of detection performance.

V. Conclusion

This study proposes a hybrid pipeline for nanoparticle analysis using transmission electron microscopy (TEM). This approach combines YOLO-based neural detection with classical post-detection segmentation, avoiding the immediate need to manually re-annotate the entire dataset with pixel-level masks.

The detection model achieved an mAP@50 of 68.1% with precision, recall, and F1-score values of approximately 74%. The combination of neural detection and classical segmentation represents a practical alternative for datasets that contain only bounding box annotations. Future work should include more cases of manual verification of images without annotations, optimization of the YOLO hyperparameters, and quantitative validation of the generated segmentation masks.

Acknowledgments

The authors acknowledge the Universidade Estadual Paulista (Unesp), participating laboratories, research collaborators, and funding agencies that supported the development of this study. The authors gratefully acknowledge the São Paulo Research Foundation (FAPESP) for financial support through grant no. 2024/01849-4 and no. 2025/12871-3.

References

[1] X. Li *et al.*, "Magnetic nanoparticles for cancer theranostics: Advances and prospects," *Journal of Controlled Release*, vol. 335, pp. 437–448, 2021.

[2] A. Setia, A. K. Mehata, A. K. Malik, M. K. Viswanadh, and M. S. Muthu, "Theranostic magnetic nanoparticles: Synthesis, properties, toxicity, and emerging trends for biomedical applications," *Journal of Drug Delivery Science and Technology*, vol. 81, p. 104295, 2023.

[3] J. R. Harris, "Transmission electron microscopy in molecular structural biology: A historical survey," *Archives of Biochemistry and Biophysics*, vol. 581, pp. 3–18, sep 2015.

[4] M. I. Hussain, M. A. Rafique, W.-G. Jung, B.-J. Kim, and M. Jeon, "Segmentation and morphology computation of a spiky nanoparticle using the hourglass neural network," *ACS Omega*, vol. 8, no. 20, pp. 17 834–17 840, 2023.

[5] K. M. Saaim, S. K. Afridi, M. Nisar, and S. Islam, "In search of best automated model: Explaining nanoparticle TEM image segmentation," *Ultramicroscopy*, vol. 233, p. 113437, 2022.

[6] A. Kaphle, S. Jayarathna, H. Moktan, M. Aliru, S. Raghuram, S. Krishnan, and S. H. Cho, "Deep learning-based TEM image analysis for fully automated detection of gold nanoparticles internalized within tumor cell," *Microscopy and Microanalysis*, vol. 29, no. 4, pp. 1474–1487, 2023.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[8] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024, computer software. [Online]. Available: https://github.com/ultralytics/ultralytics

[9] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5561–5569.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[11] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image and Vision Computing*, vol. 97, p. 103910, 2020.

Corresponding author: Michael Almeida
Institution: São Paulo State University (Unesp)
City: Botucatu, SP
Country: Brazil
Email: mj.almeida@unesp.br

Development of a Digital Prototype for Continuing Education on Gadolinium-Based Contrast Agents in Magnetic Resonance Imaging: A Study Based on a National Needs Assessment

Stella Siqueira Campos
Graduate Program in Health Technology
Pontifical Catholic University of Parana
Curitiba, Brazil
tnrstella@gmail.com

Akemi Yagui
Graduate Program in Biomedical Engineering
Federal University of Technology - Parana
Curitiba, Brazil
akemiyagui@professores.utfpr.edu.br

Edney Milhoretto
Academic Department of Physics
Federal University of Technology - Parana
Curitiba, Brazil
edneymilhoretto@utfpr.edu.br

Mauren Abreu de Souza
Graduate Program in Health Technology
Pontifical Catholic University of Parana
Curitiba, Brazil
mauren.souza@pucpr.br

***Abstract*—Gadolinium-based contrast agents are widely used in magnetic resonance imaging (MRI) examinations and require technical knowledge related to patient safety, renal function assessment, dose calculation, and adverse event management. However, essential information for professional practice remains dispersed across guidelines, package inserts, and educational materials, making consultation during clinical routines more difficult. This study aimed to identify educational and operational needs related to the use of these agents and to develop a needs-informed digital prototype for continuing education and rapid consultation. A national survey was conducted with 178 MRI professionals using an electronic questionnaire. The results revealed gaps related to the use of contrast agents in special populations, agent classification, renal function assessment, and management of adverse situations, as well as a high demand for quick-reference resources. Based on these findings, a Progressive Web Application (PWA) was developed, containing educational content, contrast dose and estimated glomerular filtration rate calculators, package inserts for contrast agents available in Brazil, and a frequently asked questions section. The prototype integrates educational and operational resources into a single platform and was designed to facilitate access to organized information during MRI practice. The current contribution should be interpreted as a needs-informed functional prototype, not a validated clinical decision-support system. Further studies are required to assess technical validation, usability, and impact on users' knowledge and professional practices.**

***Index Terms*—Magnetic resonance imaging, gadolinium-based contrast agents, continuing education, patient safety.**

I. Introduction

Gadolinium-based contrast agents play a fundamental role in magnetic resonance imaging (MRI), contributing to tissue characterization, lesion detection, and increased diagnostic accuracy in several clinical conditions [1]–[4]. Despite the favorable safety profile of these agents, their use requires technical knowledge related to contrast indication, dose calculation, evaluation of renal function, recognition of risk factors, and management of possible adverse reactions [3]–[6]. Over recent decades, scientific knowledge about these contrast agents has evolved significantly, driven by studies related to nephrogenic systemic fibrosis, gadolinium deposition in tissues, and their use in special populations, such as pediatric patients, pregnant patients, and individuals with impaired renal function [3]. As a result, scientific societies and regulatory agencies have updated guidelines and recommendations, making continuous updates necessary for professionals involved in MRI examinations [3]–[5]. Information relevant to daily practice is often dispersed among guidelines, package inserts, institutional protocols, and educational materials. This fragmentation may hinder timely access to information during clinical routines and contribute to variability in professional practices across institutions and teams. In this context, continuing education plays an important role in supporting patient safety and promoting more consistent professional practices. Simultaneously, advances in digital technologies have expanded the possibilities for health education and support for professional practice, enabling updated content, calculation tools, and quick-reference materials to be made available directly in the workplace [7]–[10]. Thus, this study aimed to identify educational and operational needs related to the use of gadolinium-based contrast agents among MRI professionals and to develop a digital tool to support continuing education and rapid consultation. The scientific contribution of this study lies in developing a needs-informed digital prototype, in which educational gaps identified among MRI professionals were translated into functional require-

ments for a continuing education and rapid consultation tool on gadolinium-based contrast agents.

II. Materials and Methods

A. Study Design

This applied technological development study was conducted in three main stages: a national needs assessment with MRI professionals, definition of educational and functional requirements, and development of a needs-informed digital prototype in PWA format.

B. Needs Assessment

A national survey was conducted among professionals who worked or had worked in magnetic resonance imaging services in Brazil. Data were collected between May 2 and May 29, 2025, using a structured electronic questionnaire made available online and disseminated through professional groups and communication channels in the field. The study used a non-probabilistic convenience sampling strategy. The sample size for the initial needs assessment was estimated considering a population of approximately 10,000 professionals working in magnetic resonance imaging in Brazil, a 95% confidence level, and a 10% margin of error, resulting in a minimum sample size of 96 participants. To account for losses, incomplete responses, and the need for greater robustness in data analysis, the target sample was expanded to 150 to 200 respondents. The instrument was developed based on topics related to the professional use of gadolinium-based contrast agents in MRI, including safety recommendations, package inserts, and issues frequently encountered in clinical practice. It included questions to characterize participants' professional profile, experience in magnetic resonance imaging, prior training on gadolinium-based contrast agents, and knowledge of safety, clinical indications, dose calculation, renal function assessment, contrast agent classification, and adverse events. The questionnaire included objective questions and self-reported items, including Likert-type response options when applicable. A total of 178 professionals participated in the survey. The data were analyzed using descriptive statistics, with absolute and relative frequencies to characterize the sample and identify the main educational and operational demands related to the study topic. Exploratory analyses were performed to investigate associations between previous workplace training and selected outcomes identified in the survey. Previous workplace training was dichotomized as present when participants reported having received training in the current workplace, in previous workplaces, or in all workplaces, and absent when participants reported never having received training. Selected outcomes were dichotomized as adequate when participants answered "yes" or "yes, certainly", depending on the item, and as inadequate or partial for the remaining responses. Categorical variables were compared using Pearson's chi-square test. For 2×2 comparisons, odds ratios with 95% confidence intervals were calculated. A significance level of 5% was adopted. These analyses were interpreted as exploratory, considering the cross-sectional design and the non-probabilistic convenience sampling strategy.

C. Definition of Requirements

The survey results were used to identify knowledge gaps and professional practice requirements. Based on this analysis, the tool's functional and educational requirements were defined. Content related to patient safety, contrast agent classification, renal function assessment, dose calculation, use in special populations, and management of adverse situations associated with contrast agent administration was prioritized. Information selection considered guidelines from the Brazilian College of Radiology and Diagnostic Imaging, the American College of Radiology, and the European Society of Urogenital Radiology, as well as package inserts of contrast agents available in Brazil [3]–[5].

D. Digital Tool Development

Based on the survey findings, the digital tool requirements were defined by linking the main educational and operational needs identified among participants to specific functionalities. The prototype was developed as a Progressive Web Application (PWA), allowing access on mobile devices and desktop computers without requiring installation through app stores. The application was organized into independent modules, including educational content, a contrast dose calculator, an estimated glomerular filtration rate calculator, package inserts for gadolinium-based contrast agents available in Brazil, frequently asked questions, and quick-reference materials. This modular structure was adopted to facilitate navigation, content organization, maintenance, and future updates. The application was developed using the Next.js framework, based on React, and published on the Vercel platform. Data storage, authentication, and file management were implemented using Firebase resources, including Firestore, Realtime Database, Firebase Authentication, and File Storage. Firebase Authentication was incorporated as an access-control mechanism. The educational and operational content was organized based on national and international guidelines, technical documents, and package inserts for gadolinium-based contrast agents available in Brazil. Because recommendations regarding gadolinium-based contrast agents may change over time and may differ among scientific societies, regulatory agencies, manufacturers, and local institutional policies, the prototype was conceived as an educational and rapid consultation resource, not as a substitute for institutional protocols or clinical judgment. To support the long-term reliability of the educational material, future updates of guideline-based content will require source identification, version control, dated revisions, periodic content review, and specialist verification before incorporation into the application. The dose and estimated glomerular filtration rate calculators were designed to operate without collecting or storing patient-identifying information. Input values, such as weight, height, age, sex, and serum creatinine, are used transiently for calculation purposes and are automatically discarded after use, without generating patient records in the

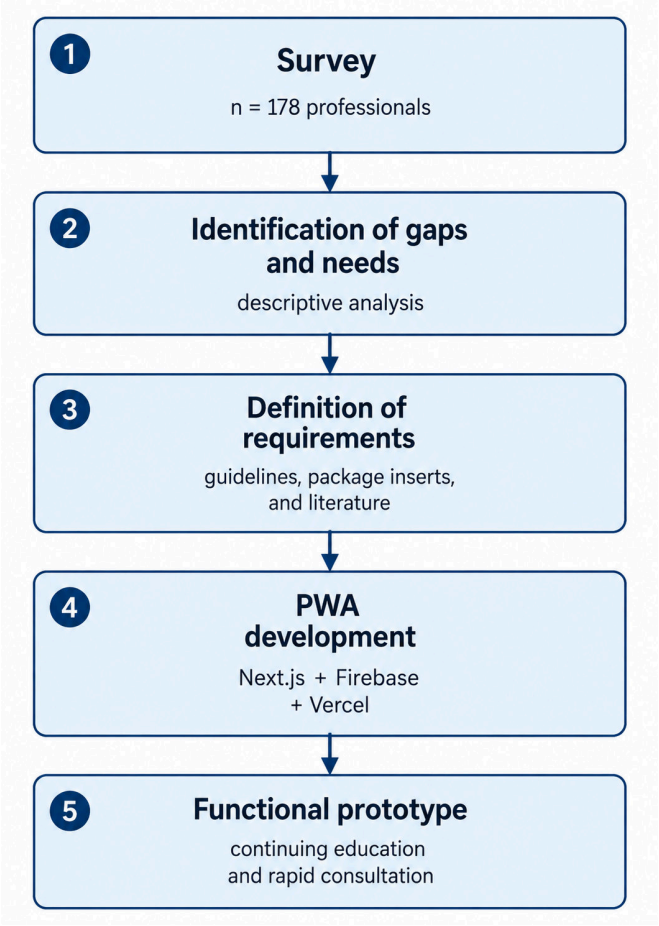


Fig. 1. Digital tool development process, from the survey with MRI professionals to the structuring of the application for continuing education and rapid consultation.

database. This design was adopted to minimize the processing of personal and sensitive health data.

The development process prioritized accessibility, ease of navigation, and information availability during contrast-enhanced MRI examinations. Fig. 1 presents the digital tool development process.

E. Ethical Aspects

The project was approved by a Research Ethics Committee under approval no. 8.468.925. Participation was voluntary and occurred after electronic informed consent was obtained. Data were collected anonymously and analyzed in aggregate form to preserve the confidentiality of the information.

III. Results

A. Needs Assessment

Among the 178 respondents, 37.6% were radiologic technologists, 37.6% were radiology technicians, 24.2% were biomedical professionals, and 0.6% had other professional backgrounds. Regarding professional experience, 47.8% had less than five years of MRI experience. The professional characteristics of the participants are presented in Table I.

TABLE I
Professional Characteristics of the Participants (n = 178)

Characteristic	n	%
Professional background		
Radiologic technologist	67	37.6
Radiology technician	67	37.6
Biomedical professional	43	24.2
Other	1	0.6
Experience in MRI		
Less than 5 years	85	47.8
5 to 10 years	34	19.1
10 to 15 years	25	14.0
More than 15 years	34	19.1
Institutional training on gadolinium		
Yes	67	37.6
No	111	62.4

Most participants considered the coverage of gadolinium-based contrast agents during their professional training to be insufficient. Additionally, 62.4% reported never having received specific institutional training on the topic. When starting their activities in MRI, 78.1% of the participants reported not feeling confident in guiding teams and patients regarding the use of these contrast agents. The main gaps identified involved use in pediatric populations, chemical aspects, agent classification, and patient care for adverse reactions related to gadolinium. Among participants, 70.2% were unable to identify ANVISA-authorized contrast agents for use in children under two years of age, 37.1% reported not being able to explain the concept of a chelate, and 36.5% stated that they did not know the differences between linear and macrocyclic contrast agents. Participants also reported difficulties related to clinical practice. Only 55.6% of participants stated that they could identify patients at higher risk of extravasation, whereas 47.8% reported not knowing or only partially knowing the recommended care following this event. When asked about their ability to guide teams and patients regarding the mechanism of action of contrast agents, dosage, contraindications, renal impairment, and adverse reactions, only 33.7% reported feeling fully prepared. Fig. 2 presents the main knowledge gaps identified among participants.

Exploratory analyses showed that previous workplace training was associated with greater self-reported preparedness to guide teams and patients regarding contrast agents. Preparedness was reported by 44.8% of participants with previous training and 27.0% of those without training (OR = 2.19; 95% CI: 1.16–4.14; $p = 0.015$). Previous workplace training was also associated with knowledge of linear and macrocyclic contrast agents, reported by 58.2% of trained participants and 27.9% of untrained participants (OR = 3.59; 95% CI: 1.90–6.81; $p < 0.001$), and with knowledge of gadolinium-based contrast agents authorized by ANVISA for use in children younger than 2 years, reported by 49.3% and 18.0% of participants, respectively (OR = 4.42; 95% CI: 2.24–8.73; $p < 0.001$). These findings are summarized in Table II.

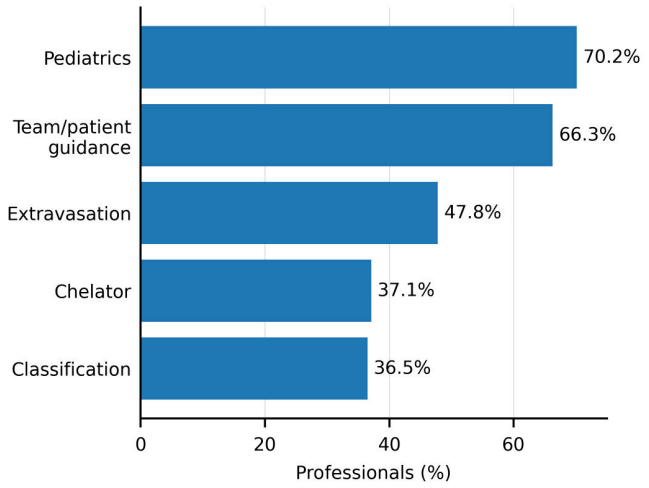


Fig. 2. Main knowledge gaps identified among participants, expressed as the percentage reporting inadequate or partial knowledge or preparedness in each domain.

TABLE II
EXPLORATORY ASSOCIATIONS BETWEEN PREVIOUS WORKPLACE TRAINING AND SELECTED OUTCOMES

Outcome	With training	Without training	OR 95% CI	p
Preparedness to guide teams/patients	30/67 (44.8%)	30/111 (27.0%)	2.19 (1.16–4.14)	0.015
Knowledge of linear/macrocyclic agents	39/67 (58.2%)	31/111 (27.9%)	3.59 (1.90–6.81)	< 0.001
Knowledge of agents authorized for children < 2 years	33/67 (49.3%)	20/111 (18.0%)	4.42 (2.24–8.73)	< 0.001

Note. Values are presented as n/N and percentage. OR: odds ratio; CI: confidence interval.

B. Digital Prototype

Based on the identified needs, a functional prototype in PWA format was developed. The functionalities were structured to address the main demands identified among participants, especially those related to patient-safety practices, guidance for teams and patients, special populations, evaluation of renal function, calculation of doses, extravasation, and classification of contrast agents. Participants showed a strong intention to use the proposed tool, with 96.1% reporting that they would use a mobile application to quickly access information on the use of gadolinium-based contrast agents in their professional routines. Fig. 3 presents examples of the interfaces developed to access educational content, calculators, and quick-reference materials.

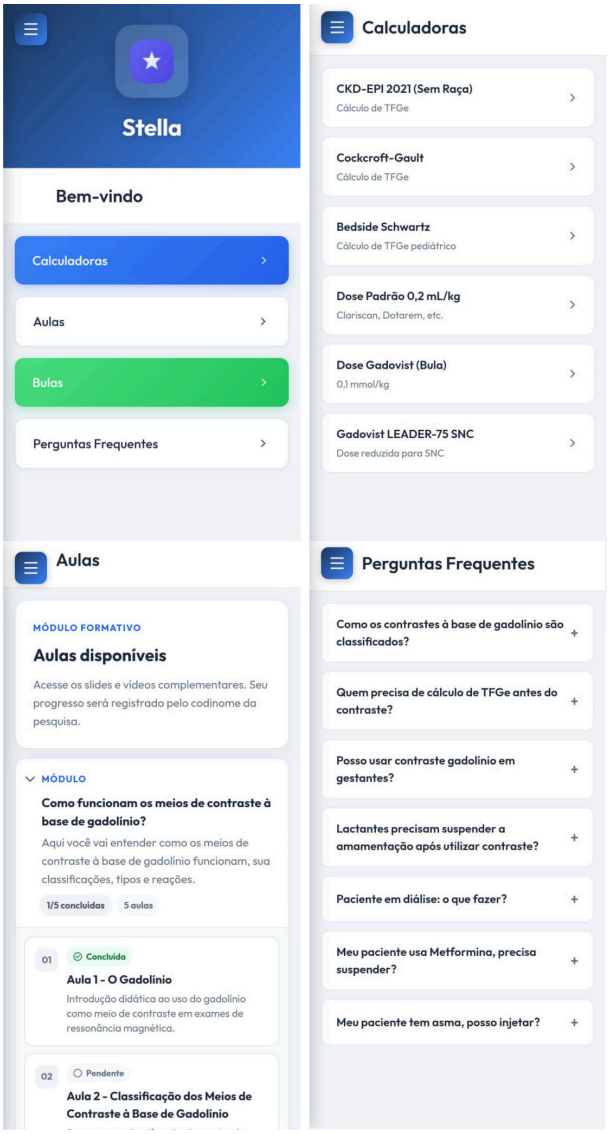


Fig. 3. Interface of the developed digital tool, presenting the main navigation modules, calculators, frequently asked questions, and package inserts.

IV. DISCUSSION

The results of the national survey revealed relevant knowledge gaps related to the use of gadolinium-based contrast agents among professionals working in magnetic resonance imaging. Although these agents are widely used in clinical practice and have favorable safety profiles, their administration involves decisions that depend on technical knowledge regarding indications, dosage, renal function assessment, recognition of risk factors, and management of adverse events. In this context, the insufficient training reported by participants may represent a barrier to more consistent professional practices and adequate guidance for teams and patients [3], [4].

The greatest difficulties identified involved topics directly related to patient safety, including the use of contrast agents in pediatric populations, understanding the differences between linear and macrocyclic agents, and care related to extravasation. These findings are relevant because they correspond to aspects frequently addressed in current guidelines, especially regarding risk stratification, selection of the contrast agent, and management of adverse situations [3], [4]. Therefore, the results suggest that the existence of technical recommendations alone does not ensure their uniform incorporation into

professional practice.

The exploratory analyses reinforce this interpretation, showing that previous workplace training was associated with greater self-reported preparedness to guide teams and patients and with better recognition of selected technical aspects, including the classification of contrast agents as linear or macrocyclic and the identification of agents authorized by ANVISA for use in children younger than 2 years. These findings suggest that institutional training may be related to specific knowledge domains relevant to MRI practice. However, these associations should be interpreted with caution, as the cross-sectional design, convenience sampling strategy, and questionnaire-based measures do not allow causal inference.

Identifying these gaps reinforces the importance of continuing education strategies that provide updated content in an accessible, objective, and workflow-integrated manner. In areas characterized by frequent guideline updates, continuing education is essential to reduce the gap between available knowledge and its application in clinical routines [7]–[10]. The high proportion of participants who reported an intention to use a mobile application suggests their receptiveness to incorporating digital technologies as a resource to support professional practice. Digital solutions for health education may expand access to knowledge, favor autonomous learning, and organize evidence-based information in formats more applicable to professional workflow [7]–[10].

Based on the identified needs, the prototype was structured to integrate educational and operational resources into a single platform. The application brings together organized content, frequently asked questions, contrast dose calculators, and tools for estimating glomerular filtration rate, seeking to facilitate access to relevant information during contrast-enhanced MRI examinations. The alignment between the gaps identified in the survey and the implemented functionalities represents a distinctive feature of the proposal, as it links technological development to the identified needs of the target audience [8].

This study has some limitations. The sample was obtained through online dissemination in professional groups and communication channels in the field, which may have favored the participation of professionals with greater interest in the topic. In addition, the data were self-reported and questionnaire-based, and the results should not be interpreted as a generalizable estimate of knowledge among all MRI professionals in Brazil. The current contribution should therefore be interpreted as a needs-informed functional prototype, rather than as a validated clinical decision-support system. Although the calculators were designed not to store patient-identifying information, future stages should include formal assessment of access permissions, logging, retention policies, deletion procedures, and compliance with the Brazilian General Data Protection Law.

The questionnaire was developed based on guideline-related topics and professional practice demands; however, it did not undergo formal psychometric validation. In addition, no formal pilot testing or structured expert review of the instrument was conducted before data collection, which should be considered when interpreting the survey findings. The complete questionnaire is available from the corresponding author upon reasonable request.

Although the application was designed to organize educational content and operational resources, it has not yet undergone technical validation, formal usability assessment, or evaluation of its impact on professional knowledge, task accuracy, clinical behavior, or patient safety outcomes.

As a continuation of this research, the tool will be evaluated through a quasi-experimental study with pre- and post-intervention measures, aiming to analyze its influence on knowledge, perceived safety, and practices related to the use of gadolinium-based contrast agents in magnetic resonance imaging. Future stages should also include technical validation of the contrast dose and estimated glomerular filtration rate calculators, including verification of formulas, units, rounding criteria, age and population restrictions, contrast agent concentration, maximum recommended doses, contraindications, and pediatric-use rules against current guidelines and package inserts. This validation should include test vectors, boundary-value tests, unit-error tests, and independent specialist review. In addition, specialist review of the educational content, version control of guideline-based information, dated updates, periodic content review, and assessment of data protection requirements should be incorporated before broader implementation. Future development may also explore non-identifiable user analytics and possible integration with institutional information systems, provided that technical validation, data protection requirements, and local governance policies are previously established. Thus, the results presented constitute the initial stage in the development of an educational technology aimed at supporting continuing education and patient-safety-oriented professional practice in MRI.

V. Conclusion

This study identified educational and operational needs related to the use of gadolinium-based contrast agents among magnetic resonance imaging professionals, especially regarding patient safety, renal function assessment, contrast agent classification, use in special populations, and management of adverse situations. Based on these findings, a needs-informed digital prototype in PWA format was developed, bringing together educational content, calculators, package inserts, frequently asked questions, and quick-reference materials. The current contribution should be understood as the development of a functional prototype for continuing education and rapid consultation, rather than as a validated clinical decision-support system. As part of the continuation of this research, technical validation, usability evaluation, specialist review, data protection assessment, and analysis of the impact of the tool on professional knowledge, perceived safety, and practices will be conducted.

References

[1] C. Westbrook and J. Talbot, *Ressonância Magnética: Aplicações Práticas*, 5th ed. Rio de Janeiro, Brazil: Guanabara Koogan, 2021.

[2] B. G. Dutra and T. Bauab Jr., Eds., *Meios de Contraste: Conceitos e Diretrizes*, 2nd ed. Santo André, Brazil: Difusão Editora and Sociedade Paulista de Radiologia e Diagnóstico por Imagem, 2022.

[3] American College of Radiology, *ACR Manual on Contrast Media*, 2025. [Online]. Available: http://www.acr.org.

[4] Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, *Diretrizes para o Uso de Meios de Contraste Intravenosos*, 2024. [Online]. Available: https://cbr.org.br.

[5] European Society of Urogenital Radiology, *ESUR Guidelines on Contrast Agents, Version 1*, 2025. [Online]. Available: https://www.esur.org.

[6] J. Cohen and S. Aime, "Paramagnetic Contrast Agents in MRI: A Review," *Substantia*, 2025, doi: 10.36253/Substantia-3538.

[7] B. Bonkoungou, F. Mosha, A. Abianuru, *et al.*, "The evolutionary journey to a new normal for learning and capacity building of healthcare workers to prepare and respond to health emergencies across Africa," *Frontiers in Public Health*, vol. 12, Art. no. 1455444, 2025, doi: 10.3389/fpubh.2024.1455444.

[8] S. Kawczak, R. D. Ficalora, B. J. Hess, R. S. Lipner, and A. Grover, "Observations from transforming a continuing education programme in the COVID-19 era and preparing for the future," *Journal of European CME*, vol. 10, no. 1, 2021, doi: 10.1080/21614083.2021.1964315.

[9] R. Kimura, M. Matsunaga, E. Barroga, and N. Hayashi, "Asynchronous e-learning with technology-enabled and enhanced training for continuing education of nurses: a scoping review," *BMC Medical Education*, vol. 23, Art. no. 505, 2023, doi: 10.1186/s12909-023-04477-w.

[10] K. K. Zhang and A. W. Thompson, "Effectiveness of electronic learning for continuing interprofessional education on behavior change of healthcare professionals: a scoping review," *Journal of Interprofessional Care*, vol. 37, no. 5, pp. 807-817, 2023, doi: 10.1080/13561820.2022.2071850.

Development of a Mobile Application with NFC Wristbands for Secure Patient Identification in Hospital Settings

***Abstract*—Incorrect patient identification is a relevant patient-safety problem in hospital environments, acting as the initiating event of a chain of failures related to medication errors, procedures performed on the wrong person, and communication breakdowns. Conventional identification wristbands are subject to fading, detachment, and low durability, which reinforces the need for more reliable identification methods. This work presents the development of a mobile application integrated with a Near Field Communication (NFC) wristband for the secure identification of hospitalized patients, together with a preliminary usability evaluation. The system was developed following the Interdisciplinary Method for the Development of Health Technologies, using Flutter for the front end, a RESTful API in ASP.NET Core with C# for the back end, and a SQL Server relational database. The application enables NFC-based patient identification, consultation and editing of clinical records, and registration of vital signs, medications, and procedures. Its usability was assessed in a controlled environment with ten participants using the System Usability Scale, yielding a mean score of 63.75 points. The result indicates that the application is usable but still requires refinement, mainly regarding navigation fluidity and interface intuitiveness. The proposed solution shows potential to strengthen patient safety and care traceability, and future work includes evaluation with health professionals and validation under real assistance conditions.**

***Index Terms*—Patient Safety, Patient Identification, Health Technology, Near Field Communication, Mobile Application**

I. Introduction

Patient safety is an essential component of healthcare quality, particularly in hospital environments, where the complexity of care, high assistance demand, and patient turnover increase the risk of incidents and adverse events, also strengthening institutional responsibility and the legal security of health services [1]. Among potentially preventable adverse events, patient identification errors are especially relevant, since they are directly related to incorrect medication administration, procedures performed on the wrong person, and communication breakdowns among multiprofessional teams [1].

Beyond organizational aspects, it is noteworthy that the nursing team plays a central role in the implementation of patient safety strategies, as it remains continuously at the patient's side throughout the entire care process. Nursing is responsible for hospital admission, safe identification, medication administration, specimen collection, clinical monitoring, performance of procedures, and communication among the different professionals of the multidisciplinary team. Consequently, failures in patient identification have a direct impact on nursing clinical practice, increasing the risk of preventable adverse events, particularly those related to medication administration, blood transfusions, and invasive procedures [2], [3].

International studies demonstrate that digital technologies designed for safe identification achieve clinical impact only when they are incorporated into the nursing workflow and accompanied by continuing training, standardized institutional protocols, and an organizational culture oriented toward safety. In this context, mobile solutions capable of supporting clinical decision-making, reducing professionals' cognitive load, and facilitating double-checking contribute significantly to the reduction of care-related errors and to the strengthening of care quality [2], [4], [5].

In routine hospital practice, especially under high-demand and emergency conditions, recurrent problems are observed in the use of conventional identification wristbands, such as low-quality materials, information that fades on contact with fluids, and frequent detachment, which may leave hospitalized patients without any reliable form of identification. These failures represent concrete risks, potentially leading to medication errors and interventions on the wrong patient, and reinforce the need to improve the identification methods adopted in healthcare institutions.

Technological devices have been increasingly regarded as tools to support safer, more reliable, and more effective healthcare for both patients and professionals. Among these, automated identification systems contribute directly to safety in hospital environments by reducing human failures and increasing the reliability of clinical information . However, the mere availability of technology does not ensure safe care; its effectiveness also depends on team integration, professional training, and institutional support. For this reason, recent studies recommend that mobile health (mHealth) applications undergo systematic usability assessment, mHealth-specific heuristics, and validation by specialists and potential users [4], [6], [7].

In recent years, digital transformation has driven substantial changes in nursing practice, fostering the development of so-called Digital Nursing. This approach encompasses the integrated use of electronic health records, mobile devices, clinical decision support systems, artificial intelligence, and automatic identification technologies to enhance care and improve patient safety. However, recent evidence indicates that the success of these technologies depends not only on their technical robustness but also on acceptance by professionals, ease of

Identify applicable funding agency here. If none, delete this.

use, interoperability among systems, and integration with the clinical processes conducted by the nursing team [5], [8], [9].

Furthermore, international organizations such as the World Health Organization and the International Council of Nurses emphasize that nurses play a strategic role in the digital transformation of health systems, acting as key agents in the implementation of technologies capable of improving care quality, reducing adverse events, and supporting evidence-based clinical decisions. Thus, the development of technologies aimed at safe identification must take into account the informational needs of nursing and support clinical reasoning during care delivery [2], [10].

Among automatic identification technologies, barcodes, Radio Frequency Identification (RFID), biometrics, and hybrid models have been used to reduce mismatches between patient, clinical information, and assistance intervention [5]. Near Field Connection (NFC), an evolution of RFID, enables fast and secure short-range data exchange between devices, and has shown potential for process automation, traceability, and improved accuracy in patient identification, optimizing workflow and reducing human error.

Despite advances in the use of barcodes and RFID, the application of NFC in hospital contexts is still limited, particularly regarding integration with mobile systems and usability validation by health professionals. In addition, there is a scarcity of national studies evaluating the practical applicability of these technologies in the Brazilian hospital setting, which characterizes a relevant gap in the development of accessible and secure solutions adapted to real care flows.

In this context, this work presents the development of a mobile application integrated with an NFC wristband for the secure identification of hospitalized patients, together with a preliminary evaluation of its usability. The specific objectives are: (a) to produce a technological solution that couples NFC reading with clinical functionalities for physicians, nurses, and nursing technicians; (b) to design an intuitive and secure interface for the consultation, recording, and editing of clinical and care information; and (c) to assess the application's usability through tests in a controlled environment in order to identify improvement points in its interface and functionality.

Although different automatic identification technologies have been described in the literature, studies investigating their use while simultaneously addressing the technological, care-related, and organizational aspects associated with nursing practice remain scarce. Recent reviews indicate that most research focuses on the technical validation of devices, with less scientific production concerning their influence on nurses' workload, among many other responsibilities. This gap highlights the need for studies that integrate biomedical engineering, health informatics, and nursing science to develop technologies that can be effectively incorporated into everyday hospital practice [11]–[14].

The remainder of this paper is organized as follows. Section II presents the theoretical background on patient safety, NFC, and usability. Section III describes the materials and methods. Section IV details the developed mobile application, and Section V reports its usability evaluation. Section VI discusses the findings and limitations, and Section VII concludes the paper.

II. Theoretical Background

A. Patient Safety and Correct Patient Identification

According to the World Health Organization (WHO), patient safety is defined as the reduction, to an acceptable minimum, of the risk of unnecessary harm associated with healthcare, and is directly related to the capacity of health systems to prevent errors and adverse events through safe practices, care protocols, and risk-management strategies [15]. The Global Patient Safety Action Plan 2021–2030 guides countries and institutions to eliminate avoidable harm, strengthen evidence-based practices, and broaden the participation of professionals, managers, users, and family members in building safer care systems [2]. In Brazil, this agenda was reinforced by the National Patient Safety Program (PNSP), established in 2013, which emphasizes standardized protocols and continuous surveillance for harm prevention [16]. Basic protocols, in turn, address practices such as correct patient identification, hand hygiene, surgical safety, and safe medication use [17].

Correct patient identification is recognized as a central safety goal because it precedes procedures, medication administration, transfusions, specimen collection, intra-hospital transport, surgeries, and clinical recording. It is also the first of the international patient safety goals defined by the Joint Commission International (JCI). Best practices recommend the use of at least two independent identifiers, such as full name and date of birth, avoiding variable data such as bed or room number, which may change during hospitalization. This verification should occur at the first contact with the patient and be repeated at critical moments of care, such as shift changes, exams, procedures, and medication administration. Within this process, nursing plays a strategic role, as it continuously participates in admission, double checking, medication administration, and clinical monitoring; therefore, identification technologies must be designed to support, rather than replace, professional judgment.

Contemporary literature shows that nursing constitutes the main link between safe identification technologies and the effective prevention of adverse events, owing to nurses' continuous presence at the patient's side throughout the 24 hours of hospitalization. Beyond correct identification, nurses coordinate care processes related to the safe administration of medications, collection of biological samples, blood transfusions, invasive procedures, clinical monitoring, and communication among the different members of the multidisciplinary team. Accordingly, the adoption of technologies such as smart wristbands, RFID, and NFC should be understood as a strategy to support nursing clinical reasoning and decision-making, strengthening the safety barriers established by institutional protocols and reducing the likelihood of errors arising from human failures. International studies demonstrate that incorporating digital systems into the nursing workflow promotes greater traceability of interventions, improves the quality of clinical

documentation, reduces medication administration errors, and supports continuity of care, provided that it is accompanied by continuing training, high usability, interoperability among systems, and an organizational culture oriented toward patient safety [3], [11], [14], [18].

From the perspective of Nursing Science, digital technologies intended for patient identification should be conceived as tools to support the Nursing Process rather than as substitutes for clinical judgment. International organizations emphasize that digital transformation in nursing requires user-centered solutions capable of integrating real-time clinical data, reducing professionals' cognitive load, and supporting rapid decisions in highly complex care environments. In this context, the use of mobile devices integrated with NFC technologies presents the potential to enhance patient safety, increase operational efficiency, and strengthen clinical governance, contributing simultaneously to better care outcomes and to the consolidation of an evidence-based safety culture [2], [8], [10], [14].

B. NFC and Traceability in Hospital Care

Digital health technologies include electronic health records, mobile applications, clinical decision support systems, wearable devices, sensors, barcodes, RFID, and NFC. When articulated with institutional protocols, these technologies can improve information management, reduce communication failures, and support decision-making at the point of care. The WHO global digital health strategy emphasizes that such systems must be secure, interoperable, people-centered, and integrated into care processes [8].

NFC enables short-range wireless communication between a reader device and a compatible tag or chip. In the proposed system, the NFC wristband acts as a physical identifier linked to the mobile application, providing access to previously registered information essential for care, such as patient identification, allergies, medications in use, comorbidities, and vital signs. The short reading distance can be regarded as a security advantage, since it requires the professional's intentional approximation to the patient. NFC was developed in 2002 by Sony and Philips and gained wider adoption from 2004 with the creation of the NFC Forum [19]. According to Ozdenizci [20], NFC applications can be organized into three operational modes: Reader/Writer, Card Emulation, and Peer-to-Peer, as illustrated in Fig. 1.

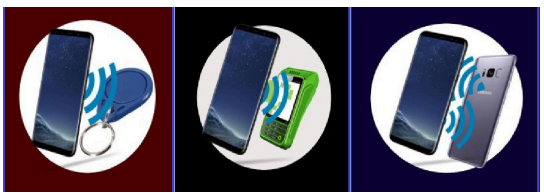


Fig. 1. NFC operational modes. Adapted from GTA/UFRJ (2018)

A recent scoping review on unique-identifier technologies showed that healthcare identification solutions rely on linear codes, two-dimensional codes, RFID, and NFC, indicating their potential to qualify the correspondence between patient and care [5]. The authors, however, stress that adoption must consider interoperability, standardization, data protection, costs, training, and implementation risks. Accordingly, the NFC wristband should be understood as a component of a socio-technical system, dependent on protocols, professional training, and institutional governance.

C. Usability and Information Security in Mobile Health

Usability is decisive for the adoption of health technologies, since clarity and ease of interaction directly affect patient safety and clinical decision-making. Poorly designed mobile systems can generate new barriers, cognitive overload, and inappropriate-use risks, which justifies the use of mHealth-specific heuristics and validation with potential users [6], [7]. A widely used instrument for this purpose is the System Usability Scale (SUS), composed of ten items rated on a five-point scale, whose scoring yields a value from 0 to 100, with results around 70 points commonly adopted as the threshold for satisfactory usability. In addition, because the application handles sensitive clinical data, privacy and information security are essential requirements, in compliance with the Brazilian General Data Protection Law (LGPD) [21].

III. Materials and Methods

A. Study Design

This is a methodological technology-development study guided by the Interdisciplinary Method for the Development of Health Technologies (MIDTS), which structures the technology-creation process into articulated, participatory, and iterative phases [22], as shown in Figure 2.

Methodological studies are suitable when the goal is to build, adapt, evaluate, or validate instruments, protocols, products, or technologies intended for professional practice. The development was organized into the following phases: (i) immersion in the problem context; (ii) elicitation of needs and requirements; (iii) design of potential solutions; (iv) Prototypingn; and (v) usability evaluation. The first phase comprised a literature review and analysis of the care problem, guided by the question of which clinical, technological, and security requirements should compose a mobile system integrated with an NFC wristband to support the secure identification of hospitalized patients.

B. Requirements Specification

The needs and requirements were defined through brainstorming sessions involving the interdisciplinary team of the project, composed of professionals from health, design, and engineering, based on the obstacles experienced in hospital routine, especially those related to patient safety. The functional requirements include the reading of NFC wristbands for patient recognition, the consultation of clinical information such as allergies and diagnoses, the updating of record data, and the management of permissions according to the user profile (nurse, physician, or nursing technician). The non-functional requirements comprise an intuitive and easy-to-navigate interface, quick access to essential resources, protection of the stored information, and compatibility with different

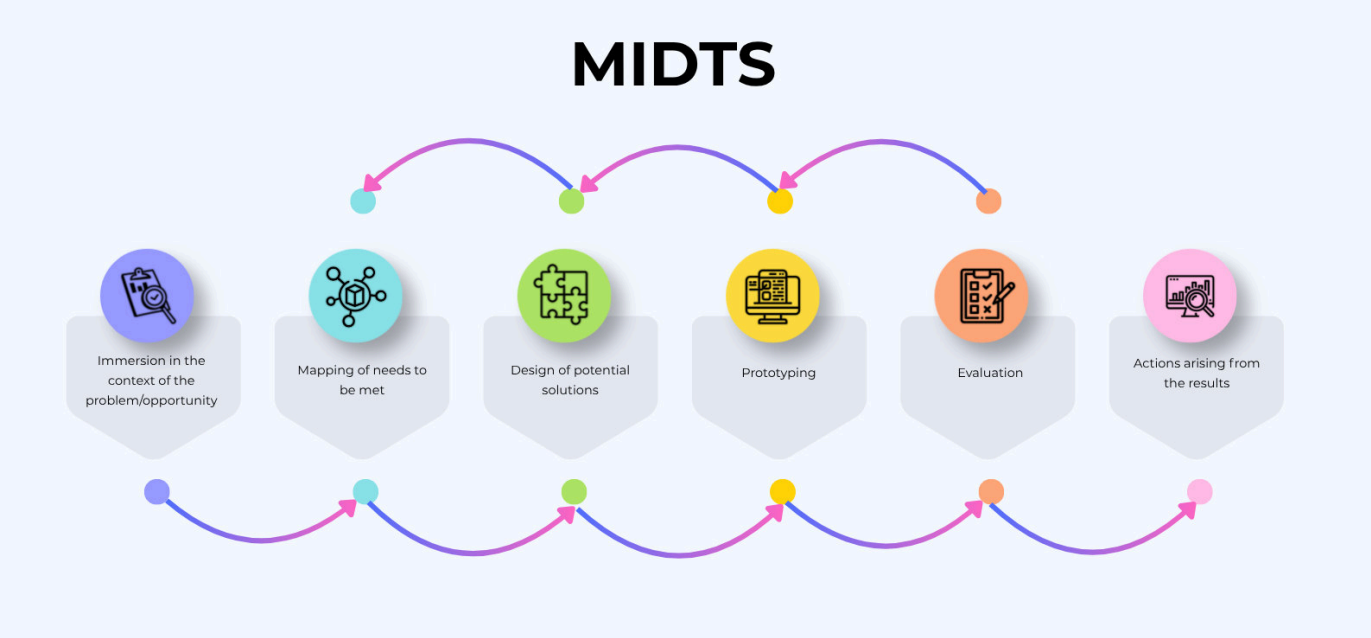


Fig. 2. MIDTS Phases. Adaptaded from [22]

digital platforms. The requirements were systematized in a descriptive table containing the identification code, description, priority level, and corresponding category, which subsidized the modeling, interface design, and implementation stages.

C. Design, Prototyping, and Development

Prototyping was carried out at two levels. Low-fidelity prototypes were produced through paper sketches to represent navigation flows, screens, and initial functionalities, allowing fast and low-cost adjustments. The high-fidelity prototype was elaborated in an interface design tool (Figma) and subsequently converted into a functional version of the application. The visual identity prioritized adequate contrast, legibility, icon consistency, information hierarchy, objective language, and error prevention, adopting the Lato typeface and a color palette organized into primary, secondary, and neutral groups.

For the implementation, the technologies were selected considering the hospital application context, the user profiles, and the target devices (smartphones and tablets). The Flutter framework was adopted for the front end, as it is an open-source framework that enables the creation of native Android and iOS applications from a single codebase and provides integration with native device resources, such as NFC, which is central to reading the patient identification wristbands. The back end was implemented with the C# language and the ASP.NET Core framework, chosen for the development of secure RESTful APIs, its modular structure, high performance, scalability, and the possibility of implementing secure authentication and profile-based permission control. Patient data were stored in a relational SQL Server database, selected for its reliability, performance, and security resources, including access control, encryption of sensitive data, and transaction auditing. The components were integrated through a RESTful API acting as an intermediary layer between the mobile application and the database. The NFC wristband consisted of a silicone bracelet with a compatible NFC chip linked to a simulated unique patient identifier. The resulting system architecture is detailed in Section IV.

D. Usability Assessment Instrument

The usability of the application was assessed through SUS [23], applied after the participants interacted with the system in a controlled, simulated scenario. The instrument comprises ten statements, alternating positive and negative wording, rated on a five-point scale ranging from 1 (“strongly disagree”) to 5 (“strongly agree”). The score is computed as follows: for odd-numbered items, one point is subtracted from the value assigned by the participant; for even-numbered items, the value assigned is subtracted from five; the resulting values are summed and multiplied by 2.5. This procedure yields a final score ranging from 0 to 100, in which higher values indicate better usability, and a score around 70 points is adopted as the threshold for satisfactory usability.

E. Ethical Aspects

The study was submitted to the Research Ethics Committee of the University of Fortaleza, in accordance with the Brazilian National Health Council Resolutions No. 466/2012 and No. 510/2016, when applicable, and with the General Data Protection Law (LGPD) [21]. As the ethical approval was pending, the entire usability evaluation was carried out using fictitious cases and simulated identifiers during the application-refinement tests, ensuring that no real patient data were used.

IV. Mobile Application

The final product of this study is a mobile application designed to support the secure and fast identification of patients in the hospital environment through NFC wristbands. The application aims to reduce failures associated with incorrect patient identification, strengthening care safety and the traceability of clinical information and of the care delivered. Its main functionalities are: NFC wristband reading with automatic display of patient data; consultation and editing of clinical records; registration of clinical information and events; and secure data storage.

A. System Architecture

The system was structured following a layered separation of responsibilities, composed of a client application developed in Flutter, a RESTful API implemented in ASP.NET Core with C# on the back end, and a relational SQL Server database. This organization favors structural clarity, scalability, and ease of maintenance, allowing the system to be extended in the future and its modules to be updated independently. The RESTful API, acting as an intermediary layer between the mobile application and the database, increases security in the traffic of sensitive information, enables service reuse across platforms (web, mobile, and desktop), and standardizes the communication among the system components, while supporting authentication and access control, which are indispensable in systems handling health data. The interaction among these components, as well as the request-and-response flow between the application layers, is shown in Fig. 3.

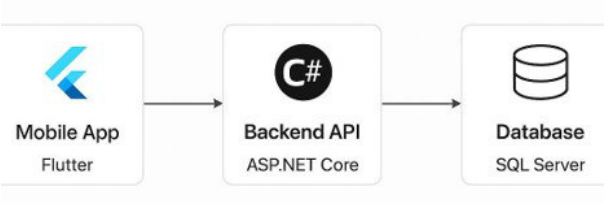


Fig. 3. System architecture: Flutter client application, ASP.NET Core RESTful API, and SQL Server database.

B. NFC Identification Flow

Patient identification is performed by approximating the smartphone or tablet to the patient's NFC wristband. This procedure enables the reading of the NFC tag and the automatic association of the patient information with the corresponding electronic record, providing greater traceability, agility, and security in the identification process. Because the reading requires intentional short-range approximation between the device and the wristband, the interaction acts as an active verification mechanism at the point of care, reducing the dependence on memory or manual checking by professionals. The functional requirements were defined considering the nursing care workflow during medication administration, vital sign measurement, recording of interventions, allergy checking, and patient identification prior to the performance of procedures.

C. Main Features and Interfaces

The home screen displays recently accessed patients, each accompanied by a color-coded risk level (green, yellow, and red), as well as a section listing patients with upcoming care activities. This organization provides a quick and systematized overview of the patients' clinical condition and enables direct access to the individual record. After identification, the application allows the importation of the patient record from a PDF file: the system provides a preview of the document and, upon confirmation, automatically extracts and fills basic patient data such as name, age, diagnosis, and allergies, which can be edited when necessary. This preview-and-confirmation step works as a double-check that reduces transcription errors and inconsistencies.

The individual patient screen centralizes essential clinical information, such as age, main diagnosis, allergies, and risk classification, and allows the consultation and recording of care activities, including vital signs, medication administration, and executed procedures. Dedicated screens support the structured registration and scheduling of vital signs (e.g., blood pressure, temperature, and oxygen saturation), medications (including dose, frequency, and administration time), and procedures, contributing to continuous monitoring and to the organized management of the care plan. The permanent display of allergies on the main screen facilitates immediate verification prior to medication administration, reducing the risk of medication-related adverse events. The structured recording of vital signs facilitates the longitudinal monitoring of the patient and may support future implementations of early warning algorithms for clinical deterioration. Finally, the recording of prescriptions could in the future be integrated with the verification of the "nine rights" of medication administration, strengthening the safety barriers established for nursing practice.

Representative interfaces of the application are presented in Fig. 4.

V. Evaluation

The usability of the application was assessed in a controlled environment using the System Usability Scale described in Section III. The evaluation involved ten participants from different academic backgrounds, including engineering, journalism, advertising, and law, in order to obtain a more diversified appraisal of the tool. During the tests, the participants interacted with the system in a simulated scenario, which made it possible to observe their interaction with the interface and with the available functionalities, after which they answered the SUS questionnaire.

The application achieved a mean SUS score of 63.75 points, a result below the reference value commonly considered satisfactory, which is around 70 points. Although the system demonstrated adequate functioning, improvement opportunities were identified, mainly related to the fluidity of navigation and the intuitiveness of the interface. According to the adopted interpretation, scores below the threshold suggest usability fragilities that require adjustments, whereas the obtained value

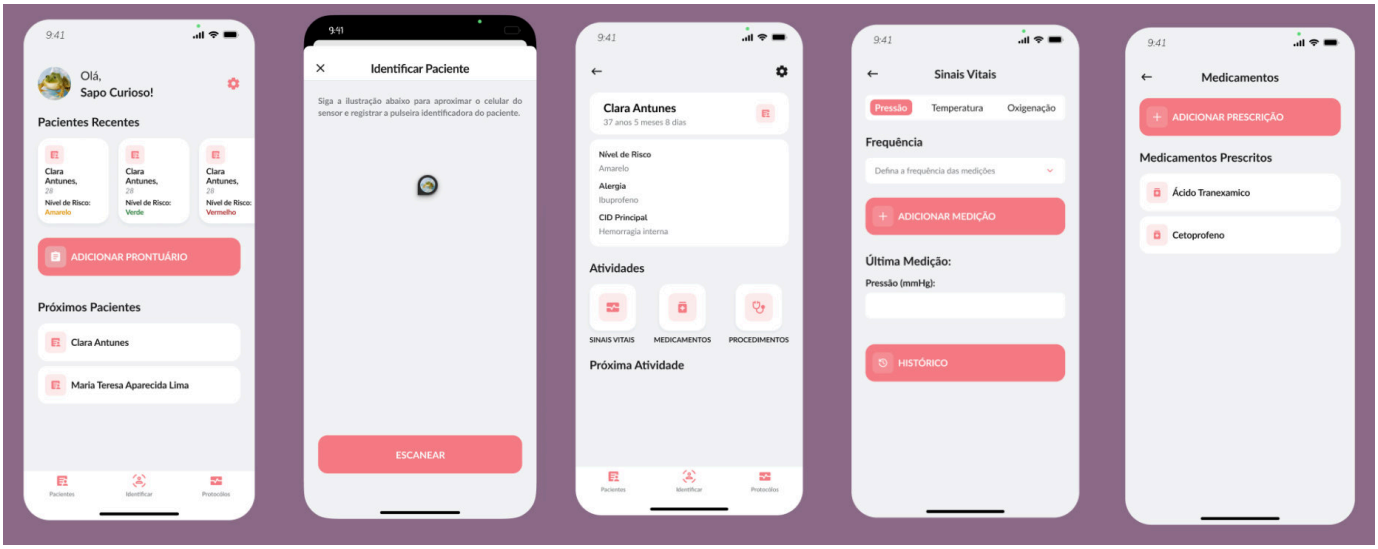


Fig. 4. Representative interfaces of the application: (a) home screen with color-coded risk levels; (b) NFC reading; (c) individual patient screen; (d) vital signs registration; and (e) medication prescription.

TABLE I
IMPROVEMENT DIRECTIONS IDENTIFIED FROM THE USABILITY EVALUATION

Dimension	Improvement direction
User interface (UI)	Redesign for greater intuitiveness, accessibility, and ease of navigation through layout reorganization and interaction simplification.
Interoperability	Integration with legacy hospital systems, such as electronic prescription, bed management, and medication stock control.
Evaluation	Longitudinal usability studies with a larger sample, prioritizing health professionals.
Functionality	Advanced features such as automatic risk alerts (e.g., drug allergies) and integration with remote monitoring.
Security	Full compliance with the LGPD, including encryption, secure authentication, and cybersecurity measures.

indicates that the application can be considered usable but still presents limitations to be addressed in order to provide a more satisfactory and efficient user experience.

Based on the evaluation, a set of improvement directions was defined to make the system more efficient and adequate to the care context, as summarized in Table I. These directions guide the subsequent refinement of the application and are further examined in the Discussion section.

VI. DISCUSSION

Patient safety has been widely discussed in the national and international literature as one of the main indicators of healthcare quality, and the correct identification of the patient is recognized by the World Health Organization as one of the priority goals for reducing harm during care [24]. Failures in patient identification are associated with medication errors, procedures performed on the wrong patient, mix-ups of laboratory exams, and disruptions in the continuity of care; the absence of standardized protocols and ineffective communication among professionals further increase assistance risks, especially in high-complexity settings such as intensive care and emergency services [25]. In this scenario, the literature highlights that identification wristbands, active checking of patient data, and technologies such as barcodes and biometrics contribute significantly to reducing identification errors, provided that they are associated with an organizational safety culture and continuous team training [26].

The functionalities of the developed application were designed as safety mechanisms incorporated into the care process rather than as merely operational resources. The NFC-based identification acts as an active verification barrier at the point of care, favoring the correct correspondence among patient, intervention, and clinical record. The home screen, by presenting recently accessed patients classified by color-coded risk levels, supports situational awareness and the early identification of potentially unstable patients in environments characterized by high workload and frequent interruptions. The importation and preview of clinical records, followed by data confirmation, function as a double-check that reduces transcription errors and documentation inconsistencies, while the permanent visualization of allergies supports safer prescription and administration decisions.

The relevance of these features is consistent with the literature on the electronic health record (EHR). The adoption of the EHR has brought important advances to care quality and service organization, favoring integration among professionals, improving the continuity of care, and reducing problems related to the illegibility of manual prescriptions, in addition to enabling traceability of clinical information and greater legal security for institutions and professionals [27]. Computerized systems can also issue automatic alerts related to allergies, drug interactions, inadequate doses, and clinical

risks, contributing to the reduction of adverse events [28]. More recently, the use of artificial intelligence and predictive analytics integrated with the EHR has been explored for the early identification of clinical risks and decision support, although such tools demand continuous monitoring, scientific validation, and professional training [29].

Despite these benefits, the literature also points to challenges in the implementation and use of electronic health systems. Usability failures, excess of automatic alerts, lack of interoperability among systems, and insufficient training may compromise the effectiveness of the EHR and even generate new assistance risks [30]. The protection of patient data is another critical aspect: the growing computerization of health services increases the need for investments in digital security policies, access control, and confidentiality of clinical information, which must be understood as an integral part of patient safety [31]. Overall, the literature converges on the understanding that secure patient identification and the EHR are fundamental strategies for preventing adverse events, but their effectiveness depends not only on technological implementation but also on institutional protocols, permanent education, an organizational safety culture, and the active participation of health professionals [8].

With regard to active participation, it is worth noting that, from the perspective of good nursing practices, the developed application has the potential to strengthen different stages of the Nursing Process by promoting correct patient identification, structured documentation of interventions, and traceability of care actions performed at the bedside. These functionalities align directly with international recommendations that identify the computerization of care as a strategy to reduce variability in clinical practice, increase adherence to care protocols, and promote evidence-based decisions. Moreover, the integration of NFC-based identification, electronic recording of interventions, and immediate visualization of critical information such as allergies and risk classification may contribute to compliance with the principles of the ”nine rights” of medication administration, strengthen communication among professionals, and enhance safety throughout all stages of care. Recent studies demonstrate that digital technologies centered on the nursing workflow improve the quality of clinical documentation, reduce recording omissions, increase compliance with safety protocols, and foster an organizational culture oriented toward the continuous improvement of care quality, particularly when combined with continuing education and evidence-based clinical governance [2], [3], [12], [14], [18].

A. Limitations

This study has limitations that should be considered when interpreting its findings. The usability evaluation was preliminary and conducted in a controlled environment, using fictitious cases and simulated identifiers, without testing in a real clinical setting. The sample comprised ten participants from non-clinical academic backgrounds, which limits the representativeness of the results for the intended end users. Moreover, the mean SUS score of 63.75 points, below the satisfactory threshold, indicates that the interface still requires refinement, particularly regarding navigation fluidity and intuitiveness. These limitations point to the need for subsequent evaluations with a larger sample of health professionals and for validation of the application under real assistance conditions.

VII. Conclusion

This work presented the development of a mobile application integrated with NFC wristbands for the identification and traceability of hospitalized patients, which proved to be a promising alternative to strengthen safety in the hospital environment. Through the automated reading of NFC wristbands, the system enables more accurate patient identification, contributing to the reduction of failures related to medication administration and to other events arising from incorrect identification. The integration with electronic records, associated with the possibility of updating the patient's clinical information, favors a more organized, agile, and reliable care flow.

Even though the system was not tested in a real environment, the preliminary usability evaluation conducted in a controlled setting with fictitious cases showed that the application presents adequate functionality, although it still demands improvements, especially regarding interface intuitiveness and ease of navigation. The mean score of 63.75 points indicates that the application can be considered usable but still presents limitations that need to be adjusted to provide a more satisfactory and efficient user experience.

These refinement points are relevant to broaden the adoption of the application in hospital institutions. As future work, the following are foreseen: redesign of the user interface to improve intuitiveness and navigability; integration with legacy hospital systems; incorporation of advanced features, such as automatic risk alerts; full compliance with data-protection requirements; and new usability evaluations with a larger sample, prioritizing health professionals, as well as validation of the application under real assistance conditions.

References

[1] L. S. Choudhury and C. T. Vu, “Patient identification errors: a systems challenge,” PSNet, Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, USA, 2020.

[2] World Health Organization, Global Patient Safety Report 2024. Geneva: WHO, 2024.

[3] International Council of Nurses, “Nursing and Digital Health.” Geneva: ICN, 2023.

[4] I. F. F. Azevedo et al., “Evaluation of mobile applications related to the nursing process: methodological study,” Texto & Contexto Enferm., vol. 33, 2024, doi: 10.1590/1980-265X-TCE-2023-0229en.

[5] K. Svandova and Z. Smutny, “What technologies based on unique patient identifiers are used for patient identification in healthcare?,” Stud. Health Technol. Inform., vol. 316, pp. 1264–1268, 2024, doi: 10.3233/SHTI240642.

[6] Z. Galavi et al., “Heuristics used for evaluating the usability of mobile health applications: a systematic literature review,” BMC Med. Inform. Decis. Mak., vol. 24, 2024.

[7] S. Hach et al., “Usability assessment methods for mobile apps for physical rehabilitation: systematic review,” JMIR mHealth uHealth, vol. 12, 2024.

[8] World Health Organization, Global Strategy on Digital Health 2020–2025. Geneva: WHO, 2021.

[9] HIMSS, Digital Health Transformation Framework, Chicago, IL, USA, 2023.

[10] International Council of Nurses, ICN Code of Ethics for Nurses. Geneva: ICN, 2021.

[11] R. M. Ratwani et al., "Health information technology and patient safety," BMJ Qual. Saf., vol. 31, 2022.

[12] C. Staggers and R. Elias, "Adoption of digital technologies in nursing practice," Comput. Inform. Nurs., vol. 41, no. 8, pp. 421–429, 2023.

[13] J. M. Harrison et al., "Digital transformation in nursing: implications for patient safety," J. Nurs. Manag., vol. 31, 2023.

[14] M. Topaz and A. Pruinelli, "Artificial intelligence and clinical decision support in nursing," Int. J. Med. Inform., vol. 184, 2024.

[15] World Health Organization, World Alliance for Patient Safety: Forward Programme 2004. Geneva: WHO, 2004.

[16] Brasil, Ministério da Saúde, Documento de referência para o Programa Nacional de Segurança do Paciente. Brasília: Ministério da Saúde/ANVISA/Fiocruz, 2014.

[17] Brasil, Agência Nacional de Vigilância Sanitária, "Resolução RDC n. 36, de 25 de julho de 2013: institui ações para a segurança do paciente em serviços de saúde," Diário Oficial da União, Brasília, DF, 2013.

[18] American Nurses Association, Nursing Informatics: Scope and Standards of Practice, 3rd ed. Silver Spring, MD, USA, 2022.

[19] NFC Forum, "NFC Forum specifications," 2011. [Online]. Available: https://nfc-forum.org/

[20] V. Coskun, K. Ok, and B. Ozdenizci, Near Field Communication (NFC): From Theory to Practice. Chichester, UK: Wiley, 2012.

[21] Brasil, "Lei n. 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais," Diário Oficial da União, Brasília, DF, 2018.

[22] J. E. Vasconcelos Filho, C. C. P. Brasil, K. M. C. Rolim, G. B. da Silva Junior, and R. M. da Silva, "MIDTS: método interdisciplinar para o desenvolvimento de tecnologias em saúde," in Tecnologias e-Health em Gestão em Saúde: fundamentos para seu desenvolvimento e avaliação, H. A. de C. [Maciel] and T. M. M. Moreira, Orgs. Curitiba, Brazil: Editora CRV, 2021, p. 49.

[23] M. V. O. Lopes et al., "System Usability Scale: adaptação e validação para a língua portuguesa," Rev. Bras. Enferm., 2019.

[24] World Health Organization, Patient Identification. Geneva: WHO, 2007.

[25] T. H. Tase and D. M. R. Tronchin, "Sistemas de identificação do paciente em instituições hospitalares: implicações para a segurança do paciente," Rev. Bras. Enferm., vol. 68, no. 1, pp. 86–93, 2015.

[26] Rede Brasileira de Enfermagem e Segurança do Paciente, Estratégias para segurança do paciente: manual para profissionais da saúde. Porto Alegre: REBRAENSP, 2013.

[27] C. G. A. Costa and F. Porto, "Prontuário eletrônico do paciente: benefícios e desafios para a prática assistencial," J. Health Inform., vol. 10, no. 2, pp. 45–52, 2018.

[28] R. K. Silva et al., "Uso do prontuário eletrônico e segurança do paciente na assistência hospitalar," Rev. Enferm. UFPE, vol. 14, no. 3, pp. 1–9, 2020.

[29] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," Nat. Rev. Genet., vol. 13, no. 6, pp. 395–405, 2012.

[30] D. W. Bates et al., "The costs of adverse drug events in hospitalized patients," JAMA, vol. 277, no. 4, pp. 307–311, 2001.

[31] H. F. Marin, "Sistemas de informação em saúde: considerações gerais," J. Health Inform., vol. 2, no. 1, pp. 20–24, 2010.

Development of Software for Notification of Critical Laboratory Results in a Hospital Setting

Ana Paula Carvalho Mariano
Graduate Program in Bioengineering
Universidade Brasil
São Paulo, Brazil
anapcarvalho@gmail.com

Telma Lisboa do Nascimento
Health Technology Laboratory (LaTeS)
Universidade Brasil
São Paulo, Brazil
telma.nascimento@ub.edu.br

***Abstract*—**The timely communication of critical laboratory results is a recognized component of patient safety. This study presents the development of a Portuguese-language web prototype for structured notification of critical laboratory results, designed for future use in Brazilian hospital settings. The prototype includes prior selection of recipient professionals, notification generation, individualized reading traceability, patient-status management, and feedback to the laboratory regarding the patient pathway. Functional validation was performed in a simulated environment using fictitious data and browser local storage. The present study therefore evaluates interface-level functionality and workflow coherence only; it does not establish clinical effectiveness, delivery reliability, cybersecurity compliance, interoperability, or reduction in communication failures. A secure back-end, persistent audit logging, encryption, escalation mechanisms, integration with laboratory and electronic health-record systems, and clinical validation remain future stages. The prototype demonstrates the feasibility of a traceable notification workflow that can support subsequent technical, usability, and clinical evaluation in the Brazilian context.

Index Terms*—critical laboratory results, health communication, medical informatics, patient safety, traceability, Unified Health System*

I. INTRODUCTION

Critical laboratory results are values that may reflect pathophysiological conditions requiring urgent clinical assessment and timely communication to the professional responsible for the patient [1], [2]. Effective communication of these results is therefore an important component of patient-safety processes in hospital care.

The relevance of the topic is supported by national and international standards and guidelines. In Brazil, Resolution RDC No. 302/2005 of the Brazilian Health Regulatory Agency (ANVISA) requires clinical laboratories to maintain documented procedures for the immediate communication of results that demand urgent medical decisions, while Resolution RDC No. 36/2013 establishes patient-safety actions in health services [3], [4]. The National Patient Safety Program also emphasizes effective communication and correct patient identification. International guidance recommends that communication of critical results be timely, clear, secure, and traceable [1].

Failures in the communication of critical results remain a recognized patient-safety concern [2], [5]. In many services, communication still depends on telephone calls, messaging applications, electronic mail, or other channels that may provide limited traceability regarding who received the information and when. Published technological approaches include text-message alerts and systems designed according to closed-loop communication principles [6], [7].

Within the scope of the literature reviewed for this study, the authors did not identify a Portuguese-language solution in the Brazilian context that combined, in a single proposed workflow, prior recipient definition, individualized reading traceability, patient-status management, and laboratory-side visibility of the patient pathway. The contribution of the present work is therefore the development of a traceable Portuguese-language prototype adapted to this workflow. It is not intended to demonstrate clinical benefit at the present stage.

Accordingly, the objective of this study was to design and functionally validate, in a simulated environment, a web-based prototype for notification of critical laboratory results. The study assessed the coherence of the implemented workflow and interface functions and defined requirements for subsequent secure deployment, interoperability, usability testing, and clinical evaluation.

II. MATERIALS AND METHODS

A. Study Design

This was an exploratory, applied technological-development study focused on the design and implementation of a computerized prototype for notification of critical laboratory results in a hospital workflow. The study was limited to software prototyping and simulated functional validation; no hospital deployment, clinical intervention, or real-patient evaluation was performed.

B. Requirements Elicitation and User Roles

Requirements were elicited from the communication flow represented by the laboratory and care-team roles. For the laboratory role, the prototype considered registration of a critical result, patient identification and location, complementary observations, identification of the issuing professional, and prior selection of recipient professionals. For the care-team role, represented by physicians and nursing staff, the prototype considered receipt and visualization of the notification, individualized reading traceability, and registration of patient-status changes. Laboratory critical limits were defined according to established references such as CLSI GP47 [1]. Table I presents examples used to structure the simulated workflow.

TABLE I
EXAMPLES OF LABORATORY EXAMS, CRITICAL LIMITS, AND POTENTIAL CLINICAL CONSEQUENCES WITHOUT IMMEDIATE INTERVENTION

Exam	Critical limit	Clinical consequence
Potassium (K+)	<2.5 or >6.0 mEq/L	Arrhythmia, cardiac arrest
Sodium (Na+)	<120 or >160 mEq/L	Cerebral edema, seizures
Glucose	<50 or >400 mg/dL	Coma, ketoacidosis
Hemoglobin	<7.0 g/dL	Severe hypoxia, shock
Platelets	<20,000/mm³	Spontaneous hemorrhage
INR	>5.0	Severe hemorrhage
Leukocytes	<1,000 or >30,000/mm³	Sepsis, immunosuppression
Troponin	>99th percentile	Acute myocardial infarction

Source: prepared based on CLSI GP47 [1].

C. Workflow States and Prototype Scope

The simulated workflow comprises the following states: critical-result registration by the laboratory; recipient selection; notification generation; notification visualization by a registered care professional; recording of the reading event; and update of the patient pathway when discharge, death, or ward transfer is registered. In the current prototype, a reading event indicates that the notification was opened by a user; it does not demonstrate clinical acknowledgment or subsequent clinical action. For this reason, the prototype should be interpreted as a traceable notification workflow inspired by closed-loop communication principles, rather than as a clinically validated closed-loop system.

Individualized access is represented at the interface level by user profiles for laboratory, physicians, and nursing staff. The prototype does not include a production-grade identity and access management service, multi-factor authentication, server-

side audit logs, encrypted credential storage, or formal authorization testing. These functions were specified as requirements for future implementation before clinical deployment.

D. Modeling, Development, and Planned Architecture

The system was modeled with Unified Modeling Language (UML) diagrams, including use-case, activity, and sequence representations, to describe interactions between the laboratory and care team. Fig. 1 illustrates the use-case structure from critical-result entry to notification reading confirmation.

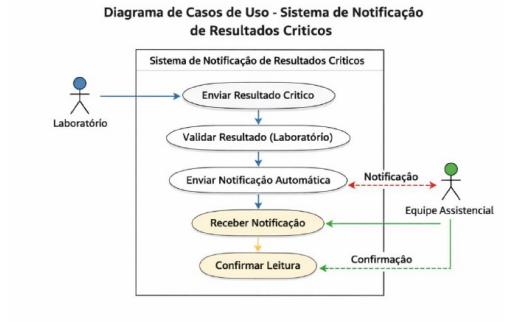


Fig. 1. Use-case diagram of the critical-results notification prototype.

Development employed HTML5 and CSS3 for interface structure and presentation and JavaScript and Python for application logic. Functional validation was performed locally with fictitious data, and visualization events were recorded in browser LocalStorage. Consequently, the current implementation does not provide a secure persistent back-end, transactional database, server-side audit trail, guaranteed message delivery, automatic retry, or downtime recovery.

For future hospital deployment, the planned architecture must include encrypted communication in transit, protected data storage at rest, secure credential management, role-based authorization, immutable or otherwise protected audit logging, message persistence and retry, explicit escalation rules for unavailable recipients or unacknowledged alerts, and documented downtime procedures. Integration with laboratory information systems (LIS) and electronic health records (EHR) was not implemented in this prototype; future interoperability should use standards supported by the participating institutions, including HL7/FHIR-based interfaces where applicable.

E. Simulated Functional Validation

The prototype was exercised with fictitious patients and professionals to verify the sequence of implemented interface functions: result registration, recipient selection, notification generation, visualization, reading traceability, and patient-status update. The present validation was designed to assess functional coherence rather than performance, reliability, cybersecurity, or clinical effectiveness.

The current study did not experimentally evaluate duplicate alerts, unavailable recipients, acknowledgment without clinical action, incorrect patient matching, or network failure. These scenarios were incorporated into the prospective validation plan because they represent safety-relevant failure modes that require controlled testing after implementation of the secure back-end and message-delivery infrastructure.

F. Ethical Aspects

The study involved technological development and simulated testing only. No real patients, identifiable clinical data, or interventions involving human beings were included, and all test information was fictitious. Under this scope, the study did not involve human-subject research. Any subsequent usability or clinical workflow validation involving laboratory professionals, nurses, physicians, patients, or identifiable health information should be submitted to the appropriate Research Ethics Committee before data collection, in accordance with applicable Brazilian requirements.

III. RESULTS AND DISCUSSION

A. Implemented Front-End Functions

The prototype provides browser-based interfaces for the laboratory, physicians, and nursing staff using individualized user profiles. At this stage, these profiles demonstrate the intended separation of user roles and traceability logic at the application interface; they should not be interpreted as evidence of production-level access control or cybersecurity compliance.

From the laboratory module, a professional can select a patient, enter the critical result, add observations, and define the professionals intended to receive the notification (Fig. 2). The simulated notification includes patient identification and location fields, such as name, date of birth, hospital registration number, ward, room, and bed. These fields are intended to support correct routing and patient identification; however, their effectiveness in preventing patient-matching errors was not evaluated in the present study.

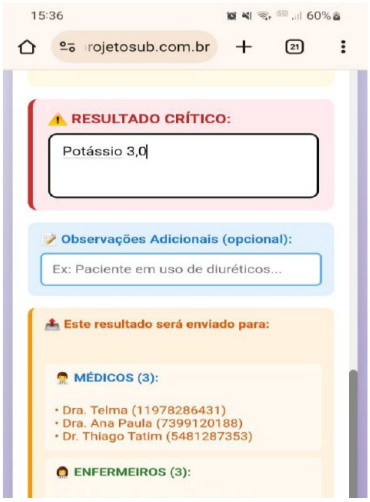


Fig. 2. Critical-result registration screen and prior selection of recipient professionals.

B. Notification Visualization and Reading Traceability

The interface reproduces the intended behavior of a push-based alert presented to registered professionals. In the simulated prototype, this demonstrates the notification workflow at the interface level only. End-to-end delivery guarantees, server-side persistence, retry policies, offline behavior, and network-failure recovery were not implemented or measured.

A central implemented function is individualized reading traceability. The system records the order in which simulated users open the notification and identifies users who have not yet generated a reading event. Fig. 3 also shows identification of the issuing professional, including professional registration information in the interface. These functions support traceability of the notification path, but a recorded reading event must not be interpreted as confirmation that an appropriate clinical action occurred.

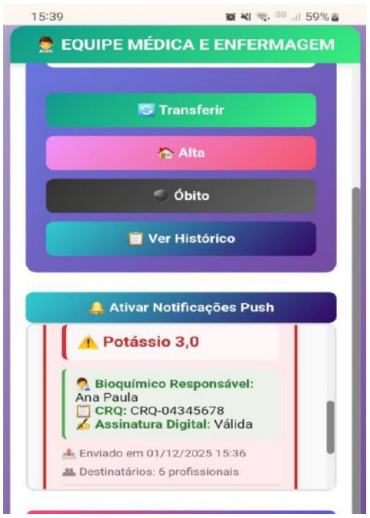


Fig. 3. Care-team interface and identification of the issuing professional in the simulated workflow.

Accordingly, the present prototype implements part of the information cycle described in closed-loop communication literature [6], [7], specifically notification presentation and reading traceability. A clinically closed loop would additionally require verification of receipt, acknowledgment, appropriate clinical action, and completion or escalation of the communication process. These outcomes were beyond the scope of the current simulated study.

C. Care-Team Management and Patient Pathway

The care-team module allows simulated registration of hospital discharge, death, and ward transfer. These updates are displayed to the laboratory module to maintain visibility of the patient's pathway within the institution. In a future integrated environment, this function may reduce the risk of directing notifications to an outdated location, but this effect was not measured in the current study.

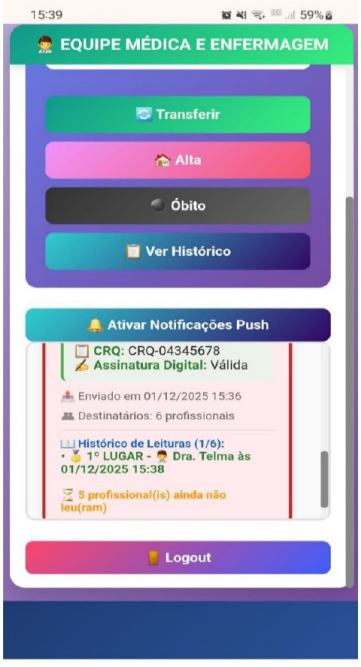


Fig. 4. Reading traceability showing viewing history and users without a recorded reading event.

D. Applicability and Innovation in the Brazilian Context

The main contribution of the proposed solution is the organization of a Portuguese-language, traceable workflow designed around communication of critical laboratory results in the Brazilian hospital context. The prototype can be accessed through commonly available web-enabled devices and does not depend, at the interface level, on proprietary hardware. These characteristics may favor future deployment in institutions with heterogeneous infrastructure, including public and philanthropic hospitals.

Nevertheless, claims of low deployment cost, high replicability, regulatory compliance, scalability, or cost-effectiveness cannot be established from the present prototype alone. Such claims require evaluation of the final architecture, infrastructure requirements, cybersecurity controls, institutional integration, maintenance burden, and performance under routine clinical load.

E. Limitations and Prespecified Future Validation Outcomes

The principal limitation is the absence of deployment in a real hospital environment. No objective performance indicators were collected for notification delivery time, acknowledgment rate, communication-failure reduction, user acceptance, or clinical response time. Interoperability with LIS/EHR platforms, scalability under concurrent routine use, cybersecurity mechanisms, and resilience to network or service failure were also not experimentally assessed.

The next validation stage should therefore combine technical testing and human-factors evaluation. Safety-relevant test cases should include duplicate notifications, unavailable recipients, notifications read without subsequent action, incorrect patient matching, interrupted connectivity, and recovery after downtime. Usability and workflow validation should include laboratory staff, nurses, and physicians under ethics oversight.

For prospective clinical evaluation, notification-to-read time and notification-to-action time should be defined as primary outcomes. Secondary outcomes should include delivery success rate, acknowledgment rate, proportion of alerts requiring escalation, communication-failure rate, user acceptance/usability measures, and frequency of workflow deviations. These outcomes should be prespecified before hospital deployment and analyzed using a secure, auditable back-end capable of recording reliable timestamps and action events.

TABLE II
IMPLEMENTATION STATUS OF KEY REQUIREMENTS IDENTIFIED DURING REVIEW

Requirement	Current status
Role-specific interfaces	Implemented in prototype
Critical-result registration	Implemented in prototype
Recipient selection	Implemented in prototype
Reading traceability	Implemented in simulated environment
Patient-pathway update	Implemented in prototype
Secure identity/access management	Planned
Encryption and protected storage	Planned
Server-side audit logs	Planned
Delivery persistence/retry	Planned
Escalation and downtime rules	Planned
LIS/EHR interoperability	Planned
Clinical usability/effectiveness	Not yet evaluated

IV. CONCLUSION

This study presents the development and simulated functional validation of a Portuguese-language web prototype for notification of critical laboratory results. The implemented functions include critical-result registration, prior recipient selection, individualized reading traceability, patient-status management, and laboratory visibility of the patient pathway. These results demonstrate the feasibility of the proposed interface and workflow logic, but they do not demonstrate improved communication, prevention of harm, reduced clinical response time, or a clinically validated closed-loop process.

Before hospital deployment, the prototype requires a secure persistent back-end, robust identity and access control, encryption, auditable server-side logging, message-delivery and retry mechanisms, escalation and downtime procedures, interoperability with institutional LIS/EHR systems, and formal cybersecurity assessment. Subsequent validation should include safety-relevant failure scenarios, usability studies with

laboratory professionals, nurses, and physicians, and prospective measurement of notification-to-read and notification-to-action times together with delivery, acknowledgment, communication-failure, and user-acceptance indicators. The present work therefore establishes a technological basis for future clinical validation rather than evidence of clinical effectiveness.

ACKNOWLEDGMENTS

To Universidade Brasil and the Graduate Program in Bioengineering.

REFERENCES

[1] Clinical and Laboratory Standards Institute, Management of Critical and Significant-Risk Results for Patient Safety: GP47. Wayne, PA: CLSI, 2019.

[2] B. C. B. Rocha, J. A. R. Alves, F. P. D. Pinto, M. E. Mendes, and N. M. Sumita, “The critical value concept in clinical laboratory,” J. Bras. Patol. Med. Lab., vol. 52, no. 1, pp. 17–20, 2016.

[3] Brazilian Health Regulatory Agency (ANVISA), “Resolution RDC No. 302, October 13, 2005,” Official Gazette of the Union, Brasília, 2005.

[4] Brazilian Health Regulatory Agency (ANVISA), “Resolution RDC No. 36, July 25, 2013,” Official Gazette of the Union, Brasília, 2013.

[5] L. C. Imoh et al., “Critical values notification: a nationwide survey of practices among clinical laboratories,” African J. Lab. Med., vol. 12, no. 1, 2023.

[6] W. A. Kanich et al., “Closed-loop critical result reporting using EMR and text messaging improves communication,” J. Pathol. Inform., vol. 11, p. 16, 2020.

[7] M. Adhyaru et al., “Push notification of laboratory critical results to responsible caregivers in the ICU: a pilot study,” J. Pathol. Inform., vol. 14, p. 100161, 2023.

Technical-Comparative Documentary Characterization of Two Brazilian Mammography Databases for Explainable Artificial Intelligence

Jaci Nunes Pereira
Graduate Program in Science and Technology
Universidade de Mogi das Cruzes – UMC
Mogi das Cruzes–SP, Brazil
ORCID: 0009-0000-2966-9072

Márcia Aparecida Silva Bissaco
Graduate Program in Science and Technology
Universidade de Mogi das Cruzes – UMC
Mogi das Cruzes–SP, Brazil
ORCID: 0000-0002-7961-4986

Silvia Cristina Martini Rodrigues
Graduate Program in Science and Technology
Universidade de Mogi das Cruzes – UMC
Mogi das Cruzes–SP, Brazil
ORCID: 0000-0002-6824-2042

***Abstract*—This study conducted a technical-comparative documentary characterization of two Brazilian mammography databases—BancoWeb/LAPIMO-USP and BR-UNIFESP—with respect to technical, clinical, documentary, and access-related characteristics potentially relevant to explainable artificial intelligence (XAI) research. A documentary, descriptive, and comparative study was conducted based on the original publications describing the databases, subsequent updates related to BancoWeb, and Brazilian technical documents and studies on mammography. The documentary design was adopted to characterize properties explicitly reported in the sources and their evolution, rather than to perform a technical audit of the files currently made available. The comparison criteria included dataset composition and unit of count, acquisition modality, formats, metadata, clinical and diagnostic information, spatial annotations, retrieval mechanisms, and access conditions. Sources from 2018 and 2022 documented BancoWeb with approximately 1,900 real mammographic images derived from digitized films and DR systems, in addition to approximately 600 phantom images. BR-UNIFESP is described as containing 941 images from 100 anonymized cases, with tomosynthesis in 62% of the cases. During the review, registration was completed on the BR-UNIFESP platform; however, after 72 hours, access to the examinations had not yet been authorized, preventing direct inspection of the images, DICOM files, and metadata. The documented resources show distinct potential correspondences with XAI investigation modalities: BancoWeb provides tools for selecting regions of interest and phantom images associated with templates of known location, whereas BR-UNIFESP provides structured search based on clinical, radiological, and anatomopathological descriptors and includes tomosynthesis. Patients, cases, clinical images, and phantom images were treated as distinct documentary units. The findings do not establish overall superiority of either database and do not constitute experimental validation of XAI techniques; rather, they organize documentary criteria that may support the preliminary selection of resources according to the type of explanation to be investigated.**

***Index Terms*—mammography, databases, artificial intelligence, explainable artificial intelligence, medical imaging**

I. Introduction

Breast cancer is a major public health problem in Brazil. Excluding non-melanoma skin tumors, it is the most common cancer in the Brazilian female population. For each year of the 2026–2028 triennium, the Brazilian National Cancer Institute estimates 78,610 new cases, corresponding to an estimated risk of 71.57 cases per 100,000 women [1]. Its heterogeneous regional distribution and mortality burden reinforce the importance of improving early detection, diagnosis, and organization of care.

Early detection includes the diagnosis of individuals with suspicious signs or symptoms and the screening of asymptomatic individuals [2]. In this context, mammography produces not only images but also clinical, radiological, and technical information related to image acquisition, BI-RADS classification, characteristics of findings, and, in certain cases, anatomopathological confirmation. The structure and availability of this information may influence data preparation, computational model development, and the procedures used for model evaluation.

The suitability of a database for artificial intelligence research does not depend exclusively on the nominal number of images. Acquisition modality, resolution, format, metadata structure, reports, diagnostic references, regions of interest, spatial annotations, and case-retrieval mechanisms may play different roles in model preparation and evaluation. In the context of explainable artificial intelligence, spatial annotations may provide references for localization analyses; radiological descriptors and reports may support semantic analyses; anatomopathological information, when available and correctly linked, may contribute to the definition of diagnostic references; and structured metadata may support contextual analyses or the retrieval of similar cases [8], [9].

Among the documented Brazilian initiatives, BancoWeb/LAPIMO-USP was developed to support the development and comparison of computer-aided diagnosis systems. Its original publication described approximately 1,400 images from about 320 patients [3]. Later sources documented expansion and restructuring of the platform. In 2018, 1,703 images derived from digitized mammographic films and 196 images from DR systems were reported, in addition to 611 phantom images [13]. In 2022, approximately 1,700 files from digitized films, about 200 DR images, and approximately 600 phantom images were described, together with specific tools for storage, search, and retrieval [14].

BR-UNIFESP, published in 2024, was developed from digital

images acquired between 2019 and 2023. The publication describes 941 images from 100 anonymized cases, acquired using a GE Senographe Pristina 3D system, with tomosynthesis present in 62% of the cases, online visualization, and the possibility of obtaining examinations in DICOM format [4]. The platform is also described as providing search mechanisms based on numerous clinical, radiological, and anatomopathological descriptors.

The consulted publications do not provide a standardized comparison between the two databases from the perspective of documentary resources related to different explainability modalities. Therefore, the following research question was formulated: which technical, clinical, documentary, and access-related characteristics are reported for BancoWeb/LAPIMO-USP and BR-UNIFESP, and what potential correspondences can be established between these resources and different modalities of XAI investigation?

The objective of this study was to conduct a technical-comparative documentary characterization of the two databases, considering composition, acquisition modality, formats, metadata, diagnostic references, annotations, search mechanisms, access conditions, and the potential correspondence between documented resources and different explainability modalities.

II. Theoretical Background

A. Mammography, Image Quality, and Artificial Intelligence

Mammography is used for screening asymptomatic women and for the diagnostic investigation of patients with signs or symptoms. Its effectiveness depends on the technical quality of the image, interpretation, and continuity of the diagnostic workup for abnormal examinations [2], [5]. Brazilian technical documents use BI-RADS to standardize radiological findings and guide clinical management [6].

Image quality is related to the acquisition chain, including positioning, compression, exposure, contrast, spatial resolution, noise, and artifacts [5]. Differences in modality, equipment, digitization, and format may also interfere with model training and evaluation. In mammography, AI has been studied for detection, classification, malignancy estimation, case prioritization, and interpretation support. Its results may vary according to database composition, reference standard, disease prevalence, and type of validation and are not automatically generalizable to other populations or contexts [7].

B. Brazilian Databases and Explainability

Mammography databases provide infrastructure for the development and comparison of AI systems. Their potential usefulness depends not only on volume but also on modality, format, clinical metadata, reports, annotations, and diagnostic confirmation [3], [4], [13], [14].

BancoWeb/LAPIMO-USP predominantly contains digitized mammographic films, although later sources also document images derived from DR systems and a specific set of phantom images [13], [14]. BR-UNIFESP contains natively digital images and cases initially selected partly for teaching purposes [4]. None of these characteristics, considered in isolation, demonstrates that the respective datasets are representative of the Brazilian population.

In XAI applied to health care, it is necessary to distinguish the existence of elements that may potentially be used as references from the actual validation of an explanation. Regions of interest or templates with known locations may provide references for spatial analyses; BI-RADS descriptors and reports may support semantic comparisons; anatomopathological information, when available, complete, and correctly linked, may contribute to the definition of diagnostic references; and cases retrievable by attributes may provide resources for example-based approaches.

However, saliency maps, similarity measures, or apparent agreement between an explanation and a given image region do not, by themselves, guarantee fidelity to the model's decision-making mechanism, causality, or clinical utility [8], [9]. Consequently, a database is not "explainable" by itself; it may provide resources that can be used in protocols for developing and evaluating explanations, whose validity depends on previously defined experimental criteria.

III. Method

A documentary, descriptive, and comparative study was conducted, with quantitative and qualitative analysis of the characteristics explicitly declared in the sources. The units of analysis were BancoWeb/LAPIMO-USP and BR-UNIFESP.

The documentary design was adopted because the primary objective of this stage was the systematic characterization of the properties reported for the two databases and, in the case of BancoWeb, their documentary evolution, rather than a technical audit of the files currently hosted or experimental validation of artificial intelligence models.

For BancoWeb/LAPIMO-USP, the original publication by Matheus and Schiabel [3] and two subsequent sources directly related to the platform's evolution were analyzed: the dissertation by Rodrigues [13] and the study by Schiabel, Matheus, and Cardoso [14]. For BR-UNIFESP, its presentation and characterization publication was used [4].

Brazilian technical documents on mammographic acquisition and quality [5], [6] and studies based on the Brazilian Cancer Information System (SISCAN) [10]–[12] were used exclusively for technical, clinical, and health-care contextualization and were not treated as image databases included in the comparison.

Sources referring to the two databases were selected when they presented information directly related to at least one of the following criteria: dataset composition and unit of count; acquisition modality or equipment; acquisition period; format, resolution, and bit depth; presence of two-dimensional mammography or tomosynthesis; clinical, BI-RADS, or anatomopathological information; reports; spatial annotations or regions of interest; search and retrieval mechanisms; anonymization; registration, authorization, and download conditions; and documentary resources potentially related to XAI investigation protocols.

When different BancoWeb publications presented different quantitative values, the values were organized chronologically

and linked to the respective documentary period, without assuming that they represented the same state of the platform. Mammographic images derived from patients and phantom images were maintained as separate categories, and their counts were not added for the purpose of comparing dataset size with BR-UNIFESP.

Units of count were preserved according to the terminology employed by the sources. "Patient" designated the clinical unit declared for BancoWeb/LAPIMO-USP, whereas "case" designated the unit used for BR-UNIFESP. In the absence of sufficient documentary definition, no assumption was made that a case necessarily corresponded to one patient, one examination, one breast, or one diagnostic episode.

Likewise, quantities referred to as "images" were reproduced as presented in the sources, without assuming complete equivalence among two-dimensional mammographic images, DICOM objects, or tomosynthesis-related content. Therefore, patients, cases, clinical images, and phantom images were considered distinct documentary units.

To organize the analysis from an XAI perspective, the documented resources were classified according to the potential function they could perform in an experimental protocol. Previously demarcated regions, regions of interest, or templates with known locations were considered potential documentary references for spatial analyses; reports and BI-RADS descriptors for semantic or concept-based analyses; anatomopathological information, when complete and correctly linked, for defining diagnostic references; clinical metadata for contextual analyses; attribute-based search mechanisms for case retrieval in example-based approaches; tomosynthesis for investigations involving three-dimensional data; and phantoms with known locations of simulated structures for controlled detection or localization experiments.

This classification represents a preliminary analytical correspondence between documented resources and possible experimental functions and does not constitute a demonstration of database suitability or a test of explanation fidelity, robustness, causality, or clinical utility [8], [9].

During revision of the study, an operational attempt to access BR-UNIFESP was made for the specific purpose of verifying the access condition described in the publication. Registration on the platform was completed; however, after 72 hours, authorization to access the examinations had not yet been granted.

Consequently, no dataset download, image inspection, DICOM-file evaluation, header reading, metadata-completeness assessment, or annotation validation was performed. This observation corresponds to a single registration request and was not used to classify the platform as unavailable.

BancoWeb/LAPIMO-USP files were likewise not directly inspected. Therefore, it was not possible to verify whether the content currently hosted fully corresponds to the state described in the 2018 and 2022 publications.

Thus, except for the operational observation concerning the BR-UNIFESP registration process, the results correspond to characteristics explicitly documented in the sources and do not constitute a technical audit of the platforms in their current configurations. Characteristics not described were classified as "not documented in the consulted sources," without inferring that they do not exist.

No model training, performance-metric calculation, testing of XAI techniques, quality scoring, interoperability assessment, or establishment of an overall hierarchy between the databases was performed.

IV. Results

A. Technical Characteristics and Access Conditions

BancoWeb/LAPIMO-USP was described in its original publication as containing approximately 1,400 images from about 320 patients [3]. Later sources documented expansion of the dataset.

In 2018, Rodrigues reported 1,703 images from laser-digitized films, with 12-bit contrast resolution and spatial resolution of 0.085 or 0.150 mm, in addition to 196 images from DR mammography systems [13]. The same source recorded approximately 1,900 real mammographic images made available online at that time and 611 phantom images.

In 2022, Schiabel et al. documented approximately equivalent values: about 1,700 images from digitized films, approximately 200 images from DR systems, and about 600 phantom images [14]. These 2018 and 2022 records represent later stages in the documentary evolution of the platform compared with the approximately 1,400 images described in 2011, but, without direct inspection, they do not allow the conclusion that these values fully correspond to the content currently available.

The phantom images constitute a dataset documented separately from mammograms obtained from patients. Therefore, their number was not added to the number of clinical images for comparison with BR-UNIFESP. The sources describe specific information associated with the phantoms, including the equipment used, exposure parameters, layer configuration, type of simulated structure, and templates containing the positions of inserted structures [13], [14].

BR-UNIFESP was described as containing 941 images from 100 anonymized cases selected among examinations performed between 2019 and 2023. According to the publication, the examinations were acquired using a GE Senographe Pristina 3D system, and 62% of the cases included tomosynthesis [4]. The source also reports examination visualization through a web browser and the possibility of downloading individual examinations or customized sets in DICOM format.

Because "patient" and "case" have no demonstrated documentary equivalence, the quantities of approximately 320 patients and 100 cases were kept separate and were not used to compare the number of individuals in the databases. Image counts were also not used to establish a size hierarchy because the sources describe different modalities, development periods, and possible counting criteria.

An internal inconsistency was identified in the BR-UNIFESP publication regarding the number of available variables. The abstract and results section report 169 clinical and imaging variables, whereas the section dedicated to platform selection

and customization mentions 113 variables [4]. Because the publication does not provide sufficient information to resolve this discrepancy, neither value was adopted in isolation as a quantitative comparative measure.

The sources describe both platforms as freely accessible subject to registration. For BancoWeb, the 2018 and 2022 sources report registration authorized by an administrator [13], [14]. For BR-UNIFESP, registration is subject to administrative approval and requires professional identification [4]. In the attempt conducted during revision of the present study, registration on BR-UNIFESP was completed, but access to the examinations had not yet been authorized after 72 hours. This result does not allow a conclusion of general unavailability of the platform; it only records that the specific request analyzed did not result in effective access within that interval.

Table I summarizes the documented characteristics of the two databases according to the criteria defined in the method. Quantitative values are presented according to the units used by the respective sources and do not assume equivalence among patients, cases, clinical images, tomosynthesis-related content, or phantom images.

B. Clinical Information and Documented Resources With Potential Correspondence to XAI

For BancoWeb/LAPIMO-USP, the publications document associations among images, reports, BI-RADS, patient characteristics, examinations, and findings, as well as mechanisms for searching, cropping, and selecting regions of interest [3], [13]. However, the existence of a cropping or marking tool does not imply the availability of previously produced spatial annotations for all clinical images.

The phantom dataset adds a distinct documentary resource. The images may contain simulated nodules or clusters of microcalcifications and include information on configuration, acquisition, and location of the structures of interest. The system is described as being able to retrieve the template corresponding to the image, containing a reference to the positions of the simulated structures [13], [14]. This resource constitutes a potential spatial reference for controlled detection or localization experiments without, by itself, representing validation of XAI techniques.

For BR-UNIFESP, the publication describes a broad set of BI-RADS descriptors and clinical, anatomopathological, and immunohistochemical information. According to the authors, image characteristics were extracted from hospital reports interpreted by trained physicians, with manual curation and organization of the data [4].

The platform is described as being capable of combining multiple attributes when retrieving cases. However, the consulted publication does not document complete segmentation masks or bounding boxes for all lesions, and the lack of direct access to the files prevented verification of the coverage and completeness of the available fields.

Considering the possible experimental functions of these resources, previously defined regions or templates with known locations show a direct correspondence with spatial-analysis protocols. Reports, BI-RADS, and anatomopathological information have potential correspondence with semantic or concept-based approaches. Multi-attribute search mechanisms may provide resources for case retrieval in example-based approaches. The presence of tomosynthesis, in turn, represents a specific characteristic of BR-UNIFESP that may be relevant to investigations using three-dimensional data.

These correspondences represent documentary possibilities for the design of future studies. They do not demonstrate experimental suitability, model performance, explanation quality, fidelity, robustness, or superiority of one database over the other.

V. Discussion

This study conducted a technical-comparative documentary characterization of BancoWeb/LAPIMO-USP and BR-UNIFESP by applying common criteria to information explicitly reported in the sources. Inclusion of the 2018 and 2022 publications related to BancoWeb prevented its characterization from being restricted to the state documented in 2011, whereas analysis of the BR-UNIFESP publication enabled examination of its technical and clinical resources and identification of relevant documentary limitations.

The contribution of the study lies in organizing documented resources according to their possible function in different XAI experimental designs without using nominal numbers of images, patients, cases, or variables as isolated measures of suitability. The interpretation was structured by considering technical heterogeneity, clinical structuring, representativeness, and the availability of references compatible with specific analysis modalities.

A. Technical Heterogeneity and Interoperability Potential

The 2018 and 2022 sources document approximately 1,900 real mammographic images for BancoWeb at those times, whereas BR-UNIFESP is described as containing 941 images from 100 cases [4], [13], [14]. These quantities do not allow the conclusion that one database is larger in terms of individuals, examinations, or information volume because “patient,” “case,” and “image” are not necessarily equivalent units.

Differences in acquisition are also documented. A predominant portion of BancoWeb is derived from digitized mammographic films, although later sources include approximately 200 images from DR equipment [13], [14]. BR-UNIFESP is described as containing images acquired digitally using a GE Senographe Pristina 3D system and includes tomosynthesis in 62% of cases [4].

These differences may imply distinct needs regarding preprocessing, resolution, file structure, and metadata use. Therefore, any integration of the two sources would likely require some degree of technical harmonization. This consideration is a methodological inference based on documented differences and not the result of an interoperability test performed in this study.

In addition, because BancoWeb files were not directly inspected, it is not possible to state that the collection currently

TABLE I
DOCUMENTED CHARACTERISTICS OF THE BRAZILIAN MAMMOGRAPHY DATABASES

Characteristic	BancoWeb/LAPIMO-USP	BR-UNIFESP
Composition and units of count		
Institution	USP/EESC – LAPIMO	UNIFESP/Hospital São Paulo
Clinical unit declared by the source	Patient in the original publication	Case
Quantity according to the declared clinical unit	Approximately 320 patients in the 2011 publication	100 anonymized cases
Number of images as reported by the sources	Approx. 1,400 in 2011; in 2018, 1,703 digitized + 196 DR; in 2022, approx. 1,700 digitized + 200 DR	941 images
Phantom images	611 in 2018; approximately 600 in 2022	Not documented in the consulted sources
Acquisition, period, and formats		
Acquisition period	Different periods; not presented as a single comparable interval in the sources used	2019–2023
Declared acquisition	Digitized films and images from DR systems	Digital mammography; GE Senographe Pristina 3D
Documented formats	TIFF and DICOM described in the updates	DICOM; other formats also described in the publication
Bit depth/resolution	12 bits; 0.085 or 0.150 mm for the digitized set	Not documented in an equivalent manner in the consulted publication
Tomosynthesis	Not documented as a component of the clinical dataset	Present in 62% of cases
Clinical information and resources potentially related to XAI		
Clinical/radiological information	Reports, BI-RADS, findings, and patient, examination, and image information	BI-RADS descriptors; clinical, anatomopathological, and immunohistochemical information
Number of searchable variables	Not presented as a single equivalent count	169 in the abstract/results; 113 in a later section of the same publication
ROI/cropping tool	Available	Not documented in an equivalent manner
Clinical spatial annotations	Few previously demarcated regions in the original publication	Complete masks or bounding boxes for all lesions not documented
Localization template	Available for phantom images, with information on type and location of simulated structures	Not documented
Structured search	Search by different characteristics and a specific phantom module	Combinatorial search using multiple clinical and radiological attributes
Access conditions and observation		
Declared access condition	Free upon registration authorized by the administrator, according to the consulted sources	Free upon registration and administrative approval
Observation in this study	Files not directly inspected; current state of the collection not audited	Registration completed; access to examinations not granted after 72 hours

Source: Matheus and Schiabel [3], Cordeiro et al. [4], Rodrigues [13], and Schiabel et al. [14]. Quantities from different years represent documentary records of distinct stages in the evolution of the platforms. Patient, case, and image values are presented according to the units declared by the sources and do not assume quantitative equivalence between the databases. "Not documented" does not mean nonexistent. Patient, case, clinical image, tomosynthesis-associated content, and phantom image represent distinct documentary units.

available fully reproduces the state described by the 2018 and 2022 sources.

The attempt to access BR-UNIFESP added a limited operational observation to the analysis. The publication characterizes the platform as freely accessible after registration and approval [4]. In the request made during this review, access was still unavailable after 72 hours. This occurrence does not demonstrate general unavailability and does not allow estimation of the frequency of delays or denials, but it indicates that "free access upon registration" should not be interpreted as synonymous with immediate access.

B. Clinical Structuring and Limits of Representativeness

BancoWeb is described as associating images with reports, findings, and information on examinations and patients and as providing mechanisms for searching by specific characteristics [3], [13]. BR-UNIFESP presents a structure based on numerous clinical, radiological, and anatomopathological descriptors and allows combinatorial retrieval of cases [4].

These characteristics suggest informational richness but do not demonstrate population representativeness. BR-UNIFESP initially selected 100 cases prioritizing didactic representations of typical findings [4]. For BancoWeb, the images were obtained over different periods and from different institutions [3], [13]. The analyzed sources do not demonstrate that the distribution of characteristics in either database reproduces the distribution of mammographic examinations performed in the Brazilian population.

Studies based on SISCAN show that variables such as age, region, clinical indication, high-risk status, and BI-RADS classification can be used to characterize mammography production in Brazil [10]–[12]. These studies were used exclusively for health-care and epidemiological contextualization and not as a quantitative standard for establishing representativeness of the two analyzed databases.

The discrepancy between 169 and 113 variables in the BR-UNIFESP publication also demonstrates that a nominal count of fields should not automatically be interpreted as a measure of coverage or completeness. Without direct inspection of the data, it was not possible to determine how many variables are populated in each case, their degree of completeness, or their individual correspondence with each image or lesion.

C. Annotations, Diagnostic References, and Explainable Artificial Intelligence

The documentary analysis allowed identification of potential correspondences between certain available resources and different modalities of XAI investigation without experimentally demonstrating that such resources improve explanation quality.

For spatial-analysis protocols, BancoWeb presents potentially relevant documentary resources. Clinical images may be submitted to region-of-interest selection, although this tool does not represent a reference annotation previously available for the entire database [3], [13]. In the phantom dataset, simulated structures have known locations through templates associated with the images [13], [14].

This resource allows the design of experiments in which a region identified by a computational method is compared with a previously known position. However, spatial agreement between an explanation and a region of interest does not, by itself, demonstrate fidelity of the explanation to the model's decision-making mechanism [8], [9]. Furthermore, results obtained with phantoms do not replace validation on clinical images.

For semantic or concept-based approaches, BR-UNIFESP provides documentary resources such as BI-RADS descriptors and clinical, anatomopathological, and immunohistochemical information [4]. These elements may provide references for investigating associations between documented concepts and model predictions. Their experimental use, however, will depend on the completeness, quality, and effective linkage of these fields to the corresponding images, aspects not verified in this study.

For example-based approaches, the combinatorial search described for BR-UNIFESP constitutes a potential documentary resource for retrieving cases with similar clinical or radiological characteristics. This does not, by itself, mean that the platform validates example-based explanations because the quality of such explanations will also depend on the model, similarity metrics, and experimental criteria used.

For investigations involving three-dimensional data, the documented presence of tomosynthesis in 62% of cases constitutes a specific characteristic of BR-UNIFESP [4]. The quantity, structure, and effective linkage of these three-dimensional data, however, could not be directly audited.

Therefore, the results do not support an overall classification of one database as superior to the other. BancoWeb documents references potentially usable in certain localization analyses and controlled phantom experiments, whereas BR-UNIFESP documents clinical, radiological, anatomopathological, attribute-based retrieval, and tomosynthesis resources potentially applicable to other experimental designs. The effective suitability of any of these resources depends on direct inspection and experimental validation.

D. Summary of the Contribution

In summary, the analysis identifies four documentary dimensions useful for preliminary selection of databases intended for XAI research: technical acquisition and storage characteristics; availability and structure of clinical metadata; presence of diagnostic references; and existence of annotations or other elements compatible with the type of explanation being investigated.

This framework is analytical and organizational in nature and does not constitute a validated instrument for database assessment, nor does it demonstrate operational compatibility or superiority of any of the analyzed datasets.

VI. Conclusion

This study conducted a technical-comparative documentary characterization of BancoWeb/LAPIMO-USP and BR-UNIFESP considering composition, acquisition modality, formats, clinical information, diagnostic references, annotations, search mechanisms, and access conditions.

Inclusion of the 2018 and 2022 sources related to BancoWeb complemented the original 2011 publication and represented its documentary evolution. These records describe approximately 1,900 real mammographic images derived from digitized films and DR systems, in addition to approximately 600 phantom images. However, because the files were not directly inspected, these quantitative values should be interpreted as characteristics documented in the respective sources and not as an audit of the collection currently available.

BR-UNIFESP is described as containing 941 images from 100 cases, digital acquisition, DICOM format, tomosynthesis in 62% of cases, and numerous clinical, radiological, and anatomopathological descriptors. An internal discrepancy was identified in the publication itself regarding the number of searchable variables, reported as 169 in some sections and 113 in another, making it impossible to establish a definitive number exclusively from the source.

Patients, cases, clinical images, and phantom images were treated as distinct documentary units. Consequently, their quantitative values do not allow a direct comparison of the clinical size of the databases or the establishment of an overall hierarchy between them.

Direct inspection of the datasets was also not possible. For BR-UNIFESP, registration was completed during revision of this study, but access to the examinations had not yet been authorized after 72 hours. This isolated occurrence does not allow a conclusion of general unavailability of the platform. BancoWeb files were also not directly audited.

The findings suggest that, in the preliminary selection of a database for XAI research, the type of explanation being investigated and the reference required for its evaluation should be considered jointly, rather than relying solely on the nominal number of images or variables.

BancoWeb documents resources with potential correspondence to localization protocols and controlled experiments using phantoms, whereas BR-UNIFESP documents resources potentially applicable to semantic, contextual, concept-based, example-retrieval, and tomosynthesis investigations. These correspondences represent documentary potential and do not constitute validation of explanation fidelity, robustness, causality, performance, or clinical utility.

Future studies should obtain direct access to the datasets, verify file integrity and structure, clarify units of count, evaluate metadata coverage and completeness, confirm correspondence between clinical information and images, examine the quality and scope of spatial annotations, and experimentally test XAI techniques using previously defined criteria for fidelity, robustness, spatial agreement, and clinical utility.

ACKNOWLEDGMENT

MASB and SCM thank FAEP/UMC for financial support.

REFERENCES

[1] Instituto Nacional de Câncer, *Estimativa 2026: incidência de câncer no Brasil*. Rio de Janeiro, Brasil: INCA, 2026. [Online]. Available: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa

[2] Instituto Nacional de Câncer José Alencar Gomes da Silva, *Diretrizes para a detecção precoce do câncer de mama no Brasil*. Rio de Janeiro, Brasil: INCA, 2015.

[3] B. R. N. Matheus and H. Schiabel, "Online mammographic images database for development and comparison of CAD schemes," *Journal of Digital Imaging*, vol. 24, no. 3, pp. 500–506, Jun. 2011, doi: 10.1007/s10278-010-9297-2.

[4] N. M. Cordeiro et al., "Towards precision medicine in breast imaging: A novel open mammography database with tailor-made 3D image retrieval for AI and teaching," *Computer Methods and Programs in Biomedicine*, vol. 248, Art. no. 108117, 2024, doi: 10.1016/j.cmpb.2024.108117.

[5] Instituto Nacional de Câncer, *Atualização em mamografia para técnicos e tecnólogos em radiologia*, 3rd ed., rev. and updated. Rio de Janeiro, Brasil: INCA, 2023.

[6] Instituto Nacional de Câncer José Alencar Gomes da Silva, *Parâmetros técnicos para rastreamento do câncer de mama*. Rio de Janeiro, Brasil: INCA, 2021.

[7] M. P. Jairam and R. Ha, "A review of artificial intelligence in mammography," *Clinical Imaging*, vol. 88, pp. 36–44, 2022, doi: 10.1016/j.clinimag.2022.05.005.

[8] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[9] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[10] A. J. A. Pereira et al., "Perfil epidemiológico de mulheres que se submeteram à mamografia no Brasil entre 2013 e 2021," *Research, Society and Development*, vol. 12, no. 1, e7412138977, 2023.

[11] T. V. L. Sverzut, I. P. da Cunha, K. L. Cortellazi, G. M. B. Ambrosano, and V. G. A. Pecorari, "Fatores associados à proporção de resultados alterados nas mamografias de rastreamento: estudo ecológico," *Revista Gaúcha de Enfermagem*, vol. 44, e20220155, 2023.

[12] J. G. S. Oliveira and G. R. S. Vidal, "Panorama da mamografia e do rastreamento do câncer de mama no SUS no estado do Rio Grande do Norte (2014–2023)," *Brazilian Journal of Implantology and Health Sciences*, vol. 7, no. 1, pp. 1938–1949, 2025.

[13] P. S. Rodrigues, "Ferramentas de interação e gerenciamento do BancoWeb – uma base de imagens para auxílio a pesquisas na àrea de mamografia digital," Master's dissertation, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, Brasil, 2018.

[14] H. Schiabel, B. R. N. Matheus, and F. J. F. Cardoso, "Features of a free mammography database to manage a random breast phantom images dataset intended for testing CAD schemes," *International Journal of Health Science*, vol. 2, no. 65, 2022, doi: 10.22533/at.ed.1592652225104.

EfficientNet and Machine Learning for Mammographic Breast Cancer Classification: A Hybrid and Low-Cost Approach

Déborah Evelyn Leite Alcantara
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
deborah.evelyn534@gmail.com

Fabíola Fábia da Silva Oliveira
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
fabiola.oliveira@ufpe.br

Heloisa Pessoa Tseng
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
heloisatseng@hotmail.com

Rebeca Teles de Melo Carvalho
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
rebeca.teles@ufpe.br

Camila Tiodista de Lima
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
camila.tiodista@ufpe.br

Clarisse Lins de Lima
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
clarisse.lins@ufpe.br

Giselle Machado Magalhães Moreno
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
giselle.moreno@ufpe.br

Wellington Pinheiro dos Santos
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

Maíra Araújo de Santana
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
maira.araujosantana@ufpe.br

Juliana Carneiro Gomes Cassemiro
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
juliana.cgomes@ufpe.br

***Abstract*—Breast cancer represents one of the most significant public health challenges in Brazil, with 64,437 diagnosed cases and approximately 21,000 deaths recorded in 2024. Mammography is the primary technique used to screen for the disease; however, its interpretation is subject to human limitations, such as fatigue, interpretative variability, and varying levels of clinical experience. In this context, the present study utilized 2,796 mammographic images from the public IRMA database, which were subjected to feature extraction using the EfficientNet-B0 deep learning network. The extracted features were provided as input to the following machine learning classifiers: Naive Bayes, BayesNet, Random Tree, Random Forest, SVM, and MLP. These were evaluated in two scenarios: Dataset 1, comprising the full set of 1,280 features, and Dataset 2, comprising 444 features selected by the Particle Swarm Optimization (PSO) algorithm. The Random Forest model with 500 trees achieved the best performance, reaching an accuracy of 76.82% and a kappa index of 0.65 in Dataset 1, and an accuracy of 75.40% with a kappa of 0.63 in Dataset 2, with an AUC of 0.88 in both cases. The results demonstrated that a reduction of approximately 65% in the number of features preserved the predictive performance of the models, indicating the efficacy of PSO in reducing dimensionality and computational cost. Furthermore, the confusion matrices demonstrated the system's robust ability to distinguish healthy individuals from those with benign or malignant breast tumors, highlighting the potential of the proposed approach as an auxiliary tool in screening, triage, and clinical prioritization strategies.**

***Index Terms*—breast cancer, mammography, artificial intelligence, classification.**

I. Introduction

Breast cancer represents a major public health challenge due to its high incidence and mortality rates. Excluding non-melanoma skin cancer, it is the most frequently diagnosed neoplasm among women and constitutes one of the leading causes of cancer-related mortality in the female population. In Brazil, data from the Department of Informatics of the Unified Health System (DATASUS) indicate that in 2024, 64,437 cases of the disease were recorded, of which 63,360 occurred in women and 1,077 in men [1]. Furthermore, the National Cancer Institute (INCA) recorded approximately 21,000 deaths associated with breast cancer during the same period [2].

Breast cancer is characterized by the uncontrolled proliferation of abnormal cells in the breast tissue. The disease can be classified as *in situ*, when it remains confined to its site of origin, or invasive, when it is capable of infiltrating adjacent tissues and spreading to other organs. Its development is

associated with genetic, hormonal, environmental, and lifestyle factors. The primary clinical signs include the presence of a breast nodule, progressive enlargement of the breast, and axillary lymphadenopathy [3].

Early detection of this disease is imperative to mitigate cancer progression and reduce mortality. Among the primary detection techniques, mammography stands out as an imaging modality that utilizes low-dose X-rays to visualize the internal structure of the breast. It is considered the gold standard for breast cancer screening due to its ability to detect microcalcifications and non-palpable nodules [4].

Although mammography is the most reliable non-invasive method currently utilized in clinical practice, the analysis of mammographic images remains a task that relies directly on the radiologist's interpretation. Consequently, subtle patterns within the images and external factors, such as fatigue, heavy workloads, interobserver variability, and limitations related to clinical experience, can influence the diagnostic outcome. In light of these limitations, advances in Artificial Intelligence and deep learning techniques have emerged as a promising strategy to assist radiologists in establishing a diagnosis, as these methods can identify patterns in the examinations that might otherwise go unnoticed by the evaluator, accelerate the analysis process, automate triage steps, and highlight suspicious regions within the images, particularly in contexts characterized by a shortage of specialists [5], [6].

In this context, the present study proposes comparing various mammographic image classification techniques leveraging artificial intelligence, with the objective of identifying the highest-performing approach for breast cancer diagnosis and contributing to the development of more efficient and accessible screening systems. To this end, mammographic images from the public IRMA database were utilized and subjected to feature extraction via the pre-trained deep neural network EfficientNet-B0, followed by the application of machine learning classifiers and the Particle Swarm Optimization (PSO) feature selection algorithm.

II. Related Work

Wang [7] presented a comprehensive review of deep learning applications in mammography for breast cancer detection and classification. The analyzed studies highlight the predominance of convolutional architectures, such as U-Net, ResNet, VGG, Inception, and Faster R-CNN. In detection and classification tasks, accuracies exceeding 95% were reported on public datasets, including DDSM, INbreast, CBIS-DDSM, and BCDR. However, many studies lack complementary metrics, such as sensitivity and specificity, hindering a more comprehensive evaluation of model performance. Furthermore, the review highlights the use of transfer learning, attention mechanisms, and ensemble strategies, which have contributed to the reduction of false positives and the enhancement of diagnostic sensitivity.

Despite their promising performance, Wang [7] highlights significant challenges regarding the clinical adoption of these approaches. Among the primary limitations are the reliance on large volumes of annotated data, the scarcity of multicenter and heterogeneous datasets, the high computational costs associated with training deep learning models, and the difficulties in generalizing across different institutions and imaging equipment. Furthermore, issues related to model interpretability, data privacy, and the standardization of acquisition protocols continue to represent barriers to the routine integration of deep learning-based systems into breast cancer screening and diagnosis.

Prodan et al. [8] investigated the application of various deep learning architectures for mammographic image analysis, including convolutional models such as ResNet and EfficientNet, as well as vision transformer-based architectures like MaxViT. The study utilized the RSNA dataset, which is characterized by severe class imbalance, and evaluated preprocessing strategies, transfer learning, and data augmentation through the generation of synthetic images using StyleGAN-XL. The results demonstrated that the artificial augmentation of the minority class yielded substantial improvements in the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), sensitivity, and F1-score metrics, even when accuracy exhibited slight reductions. Furthermore, relatively simple models, such as ResNet18, achieved performance comparable to that of more complex architectures, demonstrating that appropriate preprocessing and class balancing strategies can exert a greater influence on overall performance than merely increasing classifier complexity.

The authors emphasized, however, that the high accuracy observed in certain scenarios must be interpreted with caution, particularly in highly imbalanced datasets. In several experiments, models achieving accuracies approaching 95% exhibited low precision, sensitivity, and F1-score values, indicating limitations in the identification of positive cases. Consequently, relying solely on accuracy can lead to overly optimistic interpretations of model performance, masking the difficulties associated with detecting the minority class. Notable limitations include the individual classification of mammograms without considering multiple views of the same patient, as well as the generation of synthetic images exclusively for the positive class; these factors may restrict the generalizability of the models in real-world clinical applications [8].

Manigrasso et al. [9] investigated deep learning architectures for mammogram classification, exploring convolutional models, graph-based architectures, and vision transformers in a weakly supervised learning scenario. The study evaluated strategies capable of simultaneously integrating the four standard mammographic views, corresponding to the craniocaudal (CC) and mediolateral oblique (MLO) projections of both breasts. The results demonstrated that transformer-based architectures exhibited superior performance, achieving AUC values of up to 80.3% for cancer detection, while the convolutional and graph-based architectures achieved AUCs of approximately 74.9%. The authors also observed that different architectures exhibit distinct sensitivities to specific mammographic features, indicating that combined approaches can yield more robust results than individual classifiers. Fur-

thermore, the study highlighted that transformers rely on appropriate pre-training strategies and tend to benefit from larger datasets, whereas graph-based models exhibited high sensitivity to the initialization process and image preprocessing. Additionally, the validation was conducted at a single institution, underscoring the need for multicenter studies to evaluate the generalizability of the models.

In light of the limitations observed in the literature, this study proposes a hybrid approach based on the use of a pre-trained EfficientNet-B0 as an automatic feature extractor, combined with shallow classifiers. This strategy aims to reduce the computational cost associated with training deep architectures from scratch while simultaneously leveraging the feature representation capabilities of pre-trained models. Additionally, feature selection techniques are investigated to reduce the dimensionality of the feature space, mitigate the effects of the curse of dimensionality, and potentially enhance the generalization capability of the classifiers. The use of shallow models also enhances decision interpretability, a crucial aspect for healthcare applications and for the development of more transparent clinical decision support systems. Finally, the experimental evaluation is conducted using multiple performance metrics—including accuracy, sensitivity, specificity, precision, F1-score, and AUC—to ensure a more comprehensive and robust analysis of model behavior, particularly in scenarios characterized by class imbalance.

III. Methodology

A. Proposal

This study proposes a hybrid approach for breast cancer diagnostic support using mammographic images from the public IRMA database. The methodology comprises the utilization of a pre-trained EfficientNet-B0 architecture for automated feature extraction, followed by the application of machine learning classifiers. Additionally, the impact of feature selection via PSO algorithm was investigated, aiming to reduce data dimensionality and identify subsets of the most relevant features for classification. Model evaluation was conducted using multiple performance metrics, including accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC. The complete methodological workflow, from the acquisition of images from the public IRMA database to the comparative analysis of the results, is presented in Figure1. The following subsections detail each step of the methodology employed in this study.

B. Database

This study utilized images from the IRMA database, which were sourced from multimodal public databases (DDSM, MIAS, LLNL, and the RWTH dataset) acquired during routine clinical practice. All images were consistently encoded using the IRMA ontology, a mono-hierarchical and multiaxial classification system. This encoding scheme employs four primary axes to describe the imaging technique (modality and resolution), projection (CC or ML views), anatomy (right or left breast), and biosystem (ACR tissue density, BI-RADS

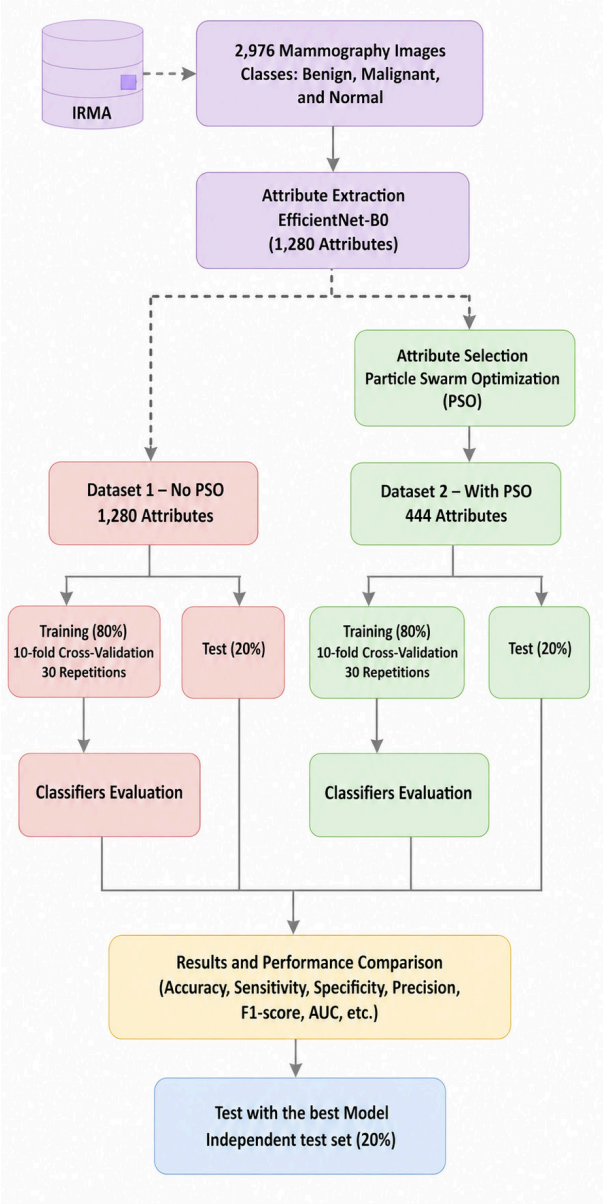


Fig. 1. Proposed method.

tumor stage, and lesion type) [10], [11]. Following integration, the reference database comprised a total of 10,509 images. For the present study, only mammographic images were selected, totaling 2,796, which were distributed across three balanced classes: benign, malignant, and normal.

C. Feature extraction

Initially, the images underwent a feature extraction process using the EfficientNet-B0 deep neural network as a feature extractor, implemented in Weka software version 3.8.6. This step utilized a transfer learning strategy, leveraging the features previously learned by the network during its pre-training on the ImageNet dataset. The network was configured to process input images resized to 28×28 pixels in a single channel (grayscale), followed by average pooling (AVR). Conse-

quently, EfficientNet-B0 served as the feature extractor for the IRMA mammographic images, generating 1,280-dimensional feature vectors that were utilized in the subsequent feature selection and classification stages.

D. Feature selection

Subsequently, the application of the PSO algorithm for feature selection was evaluated to reduce dimensionality and identify the most relevant features for the task. This algorithm addresses the optimization problem by generating a population of candidate solutions that navigate the search space, inspired by the movement of animal swarms, in pursuit of the optimal solution [12]. The application of PSO resulted in the selection of 444 features, corresponding to a 65% reduction relative to the original feature set.

E. Classification

Accordingly, the experiments were conducted using two datasets: Dataset 1, containing the complete set of extracted features (2,796 images or instances and 1,280 features), and Dataset 2, comprising only the features selected by PSO (444 features). Each dataset was then randomly partitioned into training (80%) and testing (20%) sets without overlap, utilizing the Resample filter in Weka. The following classifiers were applied to the training set: Naive Bayes, BayesNet, Multilayer Perceptron (MLP) with one hidden layer of 50 neurons, Random Tree (RT), Random Forest (RF) with 100, 300, and 500 trees, and Support Vector Machine (SVM), varying the kernel used (polynomial of degree 1, degree 2, and RBF). Each classifier was evaluated using 10-fold cross-validation with 30 repetitions.

During the training phase, each classifier configuration was evaluated using the following metrics: Accuracy, Kappa index, Sensitivity, Specificity, and AUC-ROC. Finally, the best-performing classifier was selected for the independent testing phase, where the remaining 20% of instances were classified. During this stage, a confusion matrix was generated to facilitate a more comprehensive evaluation.

IV. Results and Discussion

Tables I and II present the mean values and standard deviations obtained from the experiments conducted on the training sets of Dataset 1 and Dataset 2, respectively, utilizing 10-fold cross-validation with 30 repetitions. In both scenarios, the Random Forest classifier with 500 trees achieved the highest mean accuracy and Kappa index values, reaching $73.20 \pm 2.39\%$ and 0.60 ± 0.04 for Dataset 1, and $73.14 \pm 2.49\%$ and 0.60 ± 0.04 for Dataset 2. Furthermore, this model also yielded the highest mean specificity of 0.87, alongside competitive performance in sensitivity and AUC-ROC. The similarity between the results obtained across the two datasets demonstrates that the approximately 65% reduction in the number of features did not lead to significant performance degradation, reinforcing the potential of PSO as a strategy for mitigating the curse of dimensionality and reducing computational costs in breast cancer diagnostic support systems.

TABLE I
EVALUATION METRICS FOR BASE 1 DURING THE TRAINING STAGE.

Table 1 – Evaluation metrics: Base 1 (1,280 attributes, without PSO)					
Classifier	**Accuracy (%)** Mean ± SD	**Kappa** Mean ± SD	**Sensitivity** Mean ± SD	**Specificity** Mean ± SD	**AUC ROC** Mean ± SD
BayesNet	70.15 ± 2.79	0.55 ± 0.04	0.70 ± 0.03	0.85 ± 0.01	0.86 ± 0.02
Naive Bayes	68.12 ± 2.64	0.52 ± 0.04	0.68 ± 0.03	0.84 ± 0.01	0.85 ± 0.02
Random Tree	68.83 ± 2.80	0.53 ± 0.04	0.69 ± 0.03	0.84 ± 0.01	0.77 ± 0.02
Random Forest (100)	72.36 ± 2.53	0.59 ± 0.04	0.72 ± 0.03	0.86 ± 0.01	0.87 ± 0.01
Random Forest (300)	72.97 ± 2.54	0.59 ± 0.04	0.73 ± 0.03	0.86 ± 0.01	0.88 ± 0.01
Random Forest (500) ★	**73.20 ± 2.39**	**0.60 ± 0.04**	**0.73 ± 0.02**	**0.87 ± 0.01**	**0.88 ± 0.01**
SVM degree 1	66.38 ± 1.07	0.50 ± 0.02	0.66 ± 0.01	0.83 ± 0.01	0.83 ± 0.01
SVM degree 2	65.92 ± 1.10	0.49 ± 0.02	0.66 ± 0.01	0.83 ± 0.01	0.82 ± 0.01
SVM RBF	65.76 ± 1.17	0.49 ± 0.02	0.66 ± 0.01	0.83 ± 0.01	0.82 ± 0.01
MLP	49.65 ± 16.20	0.24 ± 0.24	0.50 ± 0.16	0.75 ± 0.08	0.53 ± 0.31

Note: ★ = Best performance; SD = Standard Deviation

TABLE II
EVALUATION METRICS FOR BASE 2 DURING THE TRAINING STAGE.

Table 2 – Evaluation metrics: Base 2 (444 attributes, with PSO)					
Classifier	**Accuracy (%)** Mean ± SD	**Kappa** Mean ± SD	**Sensitivity** Mean ± SD	**Specificity** Mean ± SD	**AUC ROC** Mean ± SD
BayesNet	70.11 ± 2.52	0.55 ± 0.04	0.70 ± 0.03	0.85 ± 0.01	0.86 ± 0.02
Naive Bayes	67.64 ± 2.40	0.51 ± 0.04	0.68 ± 0.02	0.84 ± 0.01	0.85 ± 0.01
Random Tree	68.58 ± 2.67	0.53 ± 0.04	0.69 ± 0.03	0.84 ± 0.01	0.76 ± 0.02
Random Forest (100)	72.31 ± 2.58	0.58 ± 0.04	0.72 ± 0.03	0.86 ± 0.01	0.87 ± 0.01
Random Forest (300)	72.89 ± 2.81	0.59 ± 0.04	0.73 ± 0.03	0.86 ± 0.01	0.87 ± 0.03
Random Forest (500) ★	**73.14 ± 2.49**	**0.60 ± 0.04**	**0.73 ± 0.02**	**0.87 ± 0.01**	**0.88 ± 0.01**
SVM degree 1	66.27 ± 1.00	0.49 ± 0.01	0.66 ± 0.01	0.83 ± 0.00	0.83 ± 0.01
SVM degree 2	65.67 ± 1.14	0.49 ± 0.49	0.66 ± 0.66	0.83 ± 0.01	0.82 ± 0.02
SVM RBF	65.55 ± 1.11	0.48 ± 0.02	0.66 ± 0.01	0.83 ± 0.01	0.82 ± 0.01
MLP	65.77 ± 1.97	0.49 ± 0.03	0.66 ± 0.02	0.83 ± 0.01	0.83 ± 0.01

Note: ★ = Best performance; SD = Standard Deviation

This performance is likely attributable to the randomness inherent in the operation of the Random Forest algorithm. Rather than searching for the optimal feature across the entire dataset, the algorithm constructs multiple decision trees based on smaller subsets and selects the best feature from among them. This ensures that diverse feature combinations are trained independently, thereby enhancing the model's robustness against noise and overfitting.

Regardless of the kernel employed, the SVM classifiers were distinguished by their superior consistency in performance, yielding the lowest standard deviations in accuracy among all tested models. This indicates enhanced stability and predictability on this dataset, despite achieving lower mean accuracies compared to the Random Forest-based models. Such consistency is a highly desirable characteristic for clinical applications, as system reliability is as critical as mean accuracy.

Overall, sensitivity values ranged predominantly from 0.66 to 0.73, whereas specificity values ranged from 0.83 to 0.87. Sensitivity represents the model's ability to correctly identify positive instances, making it a metric of paramount importance

in healthcare applications. Low sensitivity values can indicate an increased false-negative rate, meaning that individuals affected by the condition are incorrectly classified as healthy. Conversely, the high observed specificity suggests that the model is highly effective at correctly identifying negative cases, thereby reducing false alarms that could cause undue patient anxiety.

Notably, the MLP classifier exhibited unstable and inferior performance compared to the other models in Dataset 1, achieving an accuracy of 49.65% with a standard deviation of 16.20. However, it subsequently yielded results comparable to the average of the other models in Dataset 2, with an accuracy of 65.77% and a standard deviation of 1.97%. This demonstrates that redundant features can adversely affect models that are particularly sensitive to data dimensionality, and that feature selection can serve as a regularization step to stabilize the learning process. In addition to its suboptimal performance in Dataset 1, this classifier required a significantly high processing time, a critical factor rendering it impractical for routine use.

Accuracy results, the primary metric used to determine the best-performing classifier, are also presented as boxplots in Figures 2 and 3. It is evident that the Random Forest models exhibited not only the highest median accuracy values but also relatively low dispersion, demonstrating robust stability across the multiple cross-validation runs. In particular, the Random Forest (500) model achieved a median accuracy of approximately 73% across both datasets, corroborating the mean results reported in the tables. Similarly, the BayesNet classifier demonstrated consistent performance, with its distribution concentrated around an accuracy of 70%.

Conversely, the MLP classifier exhibited distinct behavior across the two feature sets. In Dataset 1, a high dispersion of results was observed, reflected by a large interquartile range and the presence of outliers, indicating high sensitivity to the training set and the potential influence of high data dimensionality. Following the application of PSO, a significant reduction in the MLP's variability was evident, accompanied by a substantial increase in median accuracy, which rose from approximately 50% to values near 65%. This behavior suggests that the reduction in the feature space enhanced the generalization capability of this classifier, rendering it less susceptible to overfitting.

The support vector machine classifiers exhibited highly concentrated distributions across both datasets, indicating high robustness and repeatability of the results. However, their median values remained lower than those achieved by the Random Forest models, ranging between 65% and 66% accuracy.

Following the identification of the best-performing classifier during the training phase, the evaluation proceeded on the test set. The performance of the Random Forest model with 500 trees at this stage is illustrated by the confusion matrices presented in Figure 4.

In both scenarios, all mammograms belonging to the Normal class were correctly classified, demonstrating the model's high capacity to distinguish healthy tissue. Furthermore, the

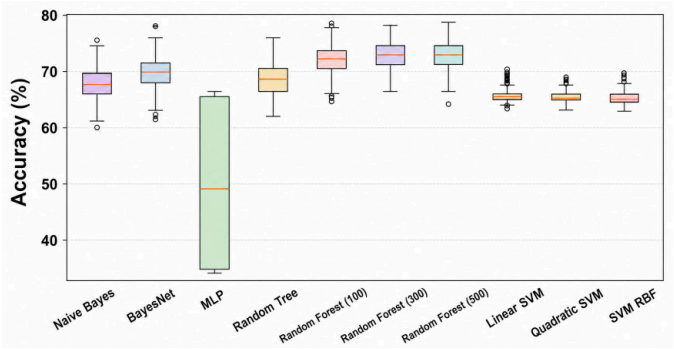


Fig. 2. Comparative boxplot of classifier accuracy on Dataset 1.

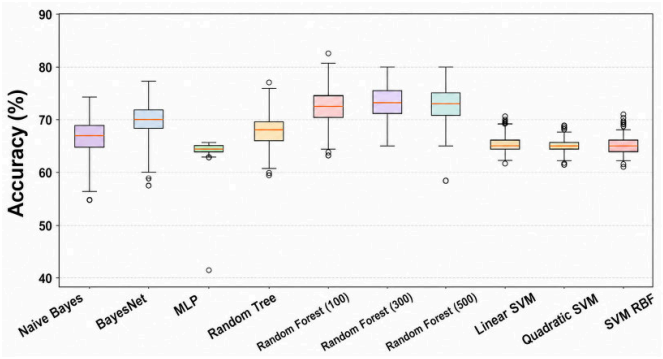


Fig. 3. Comparative boxplot of classifier accuracy on Dataset 2.

Confusion Matrix — Base 1 (without PSO) | Random Forest 500

		Predicted		
		Normal	Malignant	Benign
Actual	Normal	**187**	0	0
	Malignant	0	**116**	71
	Benign	0	59	**128**

Accuracy: 76.82% Kappa: 0.65

Confusion Matrix — Base 2 (with PSO) | Random Forest 500

		Predicted		
		Normal	Malignant	Benign
Actual	Normal	**187**	0	0
	Malignant	0	**109**	78
	Benign	0	60	**127**

Accuracy: 75.40% Kappa: 0.63

Fig. 4. Confusion matrices at the testing stage for the Random Forest classifier with 500 trees.

results obtained for Dataset 1 and Dataset 2 were highly comparable, reinforcing that feature reduction did not significantly compromise the model's performance. The primary classification challenges occurred between the Benign and Malignant classes, indicating the presence of shared mammographic features between these categories. In Dataset 1, 71 malignant cases were misclassified as benign and 59 benign cases were misclassified as malignant, resulting in an accuracy of 76.82% and a Kappa index of 0.65. In Dataset 2, the errors corresponded to 78 malignant samples misclassified as benign and 60 benign samples misclassified as malignant, yielding an accuracy of 75.40% and a Kappa index of 0.63.

Although the accuracy achieved in the independent test was lower than that reported in some studies in the literature, yielding 76.82% for Dataset 1 and 75.40% for Dataset 2, the results must be evaluated in conjunction with the other metrics. Wang [7] reviewed studies reporting accuracies exceeding 95%; however, many of these failed to report sensitivity and specificity, or presented low values for these metrics, in some cases below 0.30. This distinction is crucial, as relying solely on accuracy can mask significant limitations, particularly in imbalanced datasets. In the present study, in addition to accuracy, the Kappa index, sensitivity, specificity, and AUC-ROC were evaluated, with all metrics remaining stable above 0.60. Furthermore, the confusion matrices demonstrated that the model correctly distinguished normal cases from those presenting breast lesions, whether benign or malignant. This capability holds high clinical relevance. In screening applications, the accurate identification of patients with suspected tumors can facilitate the prioritization of complementary examinations, specialized referrals, and expedited clinical decision-making, thereby contributing to the establishment of more efficient triage workflows.

Regarding computational cost, studies such as those by Prodan et al. [8] and Manigrasso et al. [9] highlight that deep architectures, including complex CNNs, vision transformers, and multiview models, often require large volumes of data, substantial computational power, and costly training processes. The proposed methodology offers a simpler alternative by utilizing the pre-trained EfficientNet-B0 exclusively for feature extraction, followed by shallow classifiers. This strategy eliminates the need to train deep networks from scratch and facilitates deployment in scenarios with limited computational resources. Furthermore, models such as Random Forest provide greater interpretability compared to deep architectures, which may enhance their acceptance in healthcare applications. Nonetheless, the present study has limitations regarding explainability, as visual techniques, such as activation maps, were not employed to highlight the most relevant mammographic regions influencing the model's decisions.

V. Conclusion

This study aimed to investigate the use of artificial intelligence techniques for breast cancer diagnostic support based on mammographic images. To this end, a hybrid approach was proposed, relying on automatic feature extraction via the pre-trained EfficientNet-B0 architecture, followed by the application of shallow machine learning classifiers. Additionally, the impact of feature selection using the Particle Swarm Optimization algorithm was evaluated, aiming to reduce data dimensionality and enhance the computational efficiency of the system.

The results demonstrated that the Random Forest classifier with 500 trees achieved the best performance among the evaluated approaches, attaining an accuracy of 76.82%, a Kappa index of 0.65, and an AUC-ROC of 0.87 on the independent test set. Furthermore, the application of PSO was shown to reduce the number of features by approximately 65% while preserving model performance. The confusion matrices indicated the system's robust ability to distinguish healthy individuals from those with benign or malignant breast lesions. This capability is particularly relevant in screening and triage applications, where the early identification of patients requiring further investigation can facilitate clinical prioritization and optimize the healthcare workflow.

Despite the promising results, certain limitations must be acknowledged. The achieved performance remains below the level of reliability required for autonomous clinical application, underscoring that the proposed methodology should be regarded strictly as a medical decision support tool. Furthermore, the features extracted by EfficientNet-B0 exhibit limited clinical interpretability, as they lack a direct correspondence with specific anatomical characteristics or established radiological findings.

Regarding future perspectives, this study highlights the validation of the methodology on multicenter and external datasets, as well as the investigation of alternative feature selection strategies. Additionally, the exploration of hybrid architectures integrated with interpretability methods could enhance the reliability and acceptance of these tools in clinical settings, fostering the development of more transparent and efficient diagnostic support systems that are closely aligned with the demands of healthcare services.

Acknowledgments

We are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES), Fundação de Amparo a Ciencia e Tecnologia do Estado de Pernambuco - Brazil (FACEPE), and the National Council for Scientific and Technological Development - Brazil (CNPq), for the partial financial support.

References

[1] Brazil. Ministry of Health. DATASUS, "Painel-oncologia - brasil," Available: https://tabnet.datasus.gov.br/cgi/dhdat.exe?PAINEL_ONCO/PAINEL_ONCOLOGIABR.def, 2026, [Accessed: May 13, 2026].

[2] Brazilian National Cancer Institute José Alencar Gomes da Silva (INCA), "Atlas on-line de mortalidade," Available: https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo10/consultar.xhtml, 2026, [Accessed: May 13, 2026].

[3] F. C. Maluf, "Tudo sobre câncer de mama," Available: https://vencerocancer.org.br/cancer/o-que-e-cancer-de-mama/, 2026, [Accessed: May 13, 2026].

[4] P. C. Gøtzsche and K. J. Jørgensen, "Screening for breast cancer with mammography," *Cochrane database of systematic reviews*, no. 6, 2013.

[5] Brazil. Ministry of Health, “Câncer de mama,” Available: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-mama, 2026, [Accessed: May 13, 2026].

[6] Ministry of Health, Health Surveillance Secretariat, Department of Health Analysis and Surveillance of Non-Communicable Diseases, “Plano de ações estratégicas para o enfrentamento das doenças crônicas e agravos não transmissíveis no brasil 2021–2030,” Available: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/doencas-cronicas-nao-transmissiveis-dcnt/09-plano-de-dant-2022_2030.pdf, Brasília, Brazil, 2021, [Accessed: May 13, 2026].

[7] L. Wang, “Mammography with deep learning for breast cancer detection,” *Frontiers in oncology*, vol. 14, p. 1281922, 2024.

[8] M. Prodan, E. Paraschiv, and A. Stanciu, “Applying deep learning methods for mammography analysis and breast cancer detection,” *Applied Sciences*, vol. 13, no. 7, p. 4272, 2023.

[9] F. Manigrasso, R. Milazzo, A. S. Russo, F. Lamberti, F. Strand, A. Pagnani, and L. Morra, “Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures,” *Medical Image Analysis*, vol. 99, p. 103320, 2025.

[10] J. E. Oliveira, M. O. Gueld, A. d. A. Araújo, B. Ott, and T. M. Deserno, “Toward a standard reference database for computer-aided mammography,” in *Medical imaging 2008: Computer-aided diagnosis*, vol. 6915. SPIE, 2008, pp. 606–614.

[11] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95-international conference on neural networks*, vol. 4. ieee, 1995, pp. 1942–1948.

[12] T. M. Lehmann, M. O. Güld, C. Thies, B. Plodowski, D. Keysers, B. Ott, and H. Schubert, “Irma-content-based image retrieval in medical applications.” in *medinfo*, 2004, pp. 842–846.

A Comparative Study on the Use of Clinical Stratification for Predicting Hospital Length of Stay

João Victor Carneiro Faria, Nycolas Mariotto, Janderson Souza,
Luis Gustavo Modelli Andrade, Allan Felipe Fattori Alves
Universidade Estadual Paulista (UNESP)
Botucatu, SP, Brazil
joao.faria@unesp.br

***Abstract*—The length of stay in the hospital (LOS) is a key indicator of hospital resources management, influencing bed occupancy, staff size, admission planning, and care costs, in addition to being associated with complications and clinical outcomes. However, LOS prediction is challenging. LOS varies widely from patient to patient, since most hospital stays last only a few days while some extend for weeks, and this dispersion reflects quite distinct clinical profiles within the same hospital population. This work proposes dividing the hospital database into clinical groups already recognized in routine care, defined by specialists, to assess whether stratification improves LOS prediction. This clinical division is compared against an automatic division obtained through unsupervised clustering (KMeans) and against a single model that was fitted to the entire dataset. Three models were trained using electronic health records stored in the hospital database. *k-nearest neighbors* (KNN), *Random Forest*, and *XGBoost*, each with hyperparameters optimized on the complete dataset and kept fixed when segmenting the manual and automatic groups, respectively. This guarantees that only the method of dividing patients varies across approaches. Clinical division outperformed the single model ($R^2 = 0.63$) in five of the six groups, reaching R^2 values of up to 0.72, whereas automatic clustering via KMeans outperformed the single model in only one of the four groups. The results suggest that in this experiment, clinical stratification was more consistent in identifying predictively useful subgroups, although the performance of automatic stratification may vary depending on the clustering method employed. *Random Forest* and *XGBoost* performed similarly to other and better than KNN. SHAP-based interpretability analysis reinforces this result and reveals which groups remain internally heterogeneous. Subdividing by diagnostic chapter reduced the error in the surgical and elderly medical groups, with reductions of up to 19.4% in RMSE, but did not benefit every subset. The results show that specialist clinical knowledge guides the model more effectively than automatic division using KMeans, and that interpretability allows for the progressive refinement of this division.**

***Index Terms*—length of stay, machine learning, clinical stratification, clustering, interpretability, and hospital management.**

I. Introduction

Hospital length of stay (LOS) is one of the most widely used measures in hospital management. Since beds are a limited resource, LOS simultaneously functions as an indicator of occupancy, of care cost, and as an indirect indicator of complications and the degree of recovery [1], [2]. Anticipating a patient's length of stay therefore has a direct impact on staff sizing, capacity forecasting, and admission flow [3].

However, LOS prediction presents specific challenges. Its distribution is strongly right-skewed, with most hospital stays concentrated in a few days and a long tail of prolonged cases that dominates the error and penalizes models fitted to the original scale [1]. In addition, traditional statistical models, with few variables and poorly standardized data, are difficult to reproduce in the general population [4].

Given these limitations, machine learning applied to electronic health record data has become established as a promising approach for predicting complex outcomes, such as LOS [1], [5]. Among the available methods, ensemble decision-tree models, such as *Random Forest* and *XGBoost*, achieve good predictive performance and allow the contribution of the variables used by the model to be analyzed [3], [6]. This interpretability is particularly relevant in clinical applications, in which understanding the factors associated with predictions can aid in analyzing the results [4].

The literature on LOS predictions is extensive. Some studies predict only whether a stay will be long in a categorical form [2], while others predict the number of days [6], [7] across varied populations. In these comparisons, tree-based models, especially *XGBoost*, tend to show the best performance [3], [7]. However, much of this work fits a single model to the entire dataset, pooling patients with very different lengths of hospital stays. Through different approaches, other studies have investigated whether dividing the dataset before modeling improves predictive performance and which division strategy is more appropriate, whether guided by clinical knowledge [1] or obtained through automatic clustering [8]. However, a direct comparison between these two strategies under the same evaluation protocol is lacking.

This study aimed to assess whether a clinical stratification strategy for the dataset, defined by specialists, improves LOS prediction relative to (i) a single model trained on the complete dataset and (ii) an automatic stratification obtained through unsupervised clustering analysis. Three regression models were employed, two based on decision trees and one based on the distance between similar patients, applied to electronic health record data.

Table I
STUDY VARIABLES ACCORDING TO THE MOMENT OF THE HOSPITAL STAY AT WHICH THEY BECOME AVAILABLE.

Moment	Variables
1. Admission	age, sex, admission type, classification, year, month, day, and day of the week of the admission
2. Post-triage	ICD chapter*, occurrence of surgery*
3. Discharge	number of exams, number of antibiotics, patient status†

* allocation subject to the caveats discussed in the text
† excluded from all models

II. MATERIALS AND METHODS

A. Dataset and Target Variable

The target variable in this study is *Length of Stay* (LOS), defined as the patient's hospital length of stay, measured in days and obtained from institutional health records. Data were extracted from the Botucatu Medical School Hospital database and covered admissions occurring between 2015 and 2019. The dataset includes demographic information, admission characteristics, care indicators, diagnostic chapters, and temporal information associated with hospitalization.

Due to the strong positive skew observed in the length-of-stay distribution, the target variable was modeled on a logarithmic scale in order to reduce the influence of extreme values and stabilizes the data distribution. The applied The transformation is defined as

$$y' = \log(1 + y), \tag{1}$$

where y represents the original length of stay in days.

During inference, predicted values are converted back to the original scale through inverse transformation:

$$\hat{y} = e^{y'} - 1. \tag{2}$$

With this approach, the model learns on the logarithmic scale while final predictions remain expressed in days, allowing all metrics to be calculated on the original scale of the target variable.

B. Prediction Timing and Variable Availability

Information about a patient accumulates over the course of their stay, and the model's performance depends on when the prediction is made. Yang *et al.* [9] show this by comparing three stages of a hospital stay, from admission to post-treatment period. Accuracy improves as care data become available, but the useful lead time of the prediction is lost to the same extent. Therefore, the variables used in this study were organized into three moments, as presented in Table I.

The first moment is admission, comprising what is recorded upon entry without depending on the clinical assessment. The second comes after triage, when the team already knows what prompted admission. The third is discharge with closed care records.

Two variables require further consideration. Surgery is usually the second step in the management of these patients; however, in practice, this varies. In elective surgeries, the procedure is known from admission, whereas in urgent admissions, it may arise later. The diagnostic chapter, in turn, reflects a limitation of the dataset, which does not allow us to know whether the recorded code corresponds to the admission diagnosis or principal diagnosis closed at the time of discharge. This distinction is important because the divergence between the two is documented and is associated with differences in the length of stay [10]. We assumed that an equivalent diagnostic assessment was already available shortly after triage, and this simplification was explicitly stated.

The number of examinations and antibiotics was accumulated over the entire stay and belonged to the third moment. The institutional database records the prescription date of each antibiotic and the scheduling date of each exam, which allows them to be cross-referenced with the admission date and to identify how many of these records were already available on day zero, which is considered as future work.

Because it draws on information that is only consolidated over the course of care, the model evaluated here is associative in nature, rather than a predictor designed for use upon admission. Patient status at the end of the stay, recorded as death or discharge, also belonged to the third moment and was removed from all models, being retained only to describe the dataset.

C. Preprocessing

The first step was to audit the database quality and assess its consistency. Exploratory analysis showed that the original database structure generated multiple records for the same admission, creating a new entry for each stay. Records were then consolidated by admission code, such that each admission corresponded to a single row.

After this consolidation step, it was found that, over the period from 2015 to 2019, 142,575 admissions were recorded at the hospital. Records previously reserved as test patients were removed, resulting in a final total of 142,570 admissions available for analysis.

Next, the data types and proportions of missing values were examined for each variable. Records without a valid value for the target variable were removed because the absence of this information precluded supervised training. The database then underwent structural standardization, including the standardization of column names and the conversion of categorical attributes to uppercase.

From the admission date, four new temporal variables were derived corresponding to the year, month, day of the month, and day of the week. The year variable was used to represent possible long-term temporal trends, whereas the others were used to capture seasonal patterns related to admission behavior.

Within the machine learning *pipeline*, it was found that after removing records without a valid target variable, no numeric variable required imputation. Categorical variables

received constant-value imputation and *one-hot* encoding, with handling of categories not observed during training.

D. Stratification Strategies

Three modeling strategies were compared using the same protocol. The first is the *complete dataset*, in which a single model is trained on all patients and serves as a reference.

The second strategy consists of *manual clinical stratification*, in which the dataset is divided by specialists into six groups defined by objective rules based on admission type, classification, and age, with elderly patients defined as those above ≥ 60 years. The six groups were elective surgical, urgent surgical, elective medical, urgent medical, elderly surgical, and geriatric medical admissions groups.

The third strategy corresponds to *automatic stratification* via clustering. For this, the KMeans algorithm was applied in its *MiniBatch* variant, using the same set of variables employed in the supervised models, excluding the LOS. The number of clusters was fixed at four, a value defined based on four standard clustering quality indices: the elbow method inertia, silhouette coefficient, Calinski-Harabasz index, and Davies-Bouldin index.

For each group generated by the three strategies, a model was trained independently, allowing the performance of the global model to be compared directly with that of each of the two stratification methods.

E. Models, Optimization, and Evaluation

Three regression models were evaluated across all groups. The The first is the *k-nearest neighbors* (KNN), which estimates the length of stay based on the patients closest to the case being evaluated, measuring this proximity by the distance between the variables. The other two are ensemble tree models: *Random Forest* [11] combines trees trained independently on subsamples of the data, whereas *XGBoost* [12] builds them sequentially, with each new tree correcting the error left by the previous model.

The hyperparameters were fixed and identical across all modeling strategies. They were determined beforehand through global Bayesian optimization using the Optuna library, which was applied to the aggregated dataset with cross-validation on the training partition. This standardization ensured that the stratification strategy was the only experimental variable.

The evaluation was carried out using a reserved test partition corresponding to 20% of the data, defined once before any adjustment, clustering, or optimization. The partition was constructed at the patient level so that all admissions of a given patient remained entirely on the same split. No patient appeared simultaneously in the training and testing sets. The same partition was used across all three modeling strategies, ensuring that the global model, clinical groups, and *clusters* were evaluated on the same patients. The preprocessing was fitted exclusively to the training partition. The metrics used were the mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2), all calculated on the original scale in days. In the auxiliary analyses of subdivision by ICD chapter, subgroups with a number of records below the minimum threshold were discarded.

Table II
CHARACTERIZATION OF MANUALLY DEFINED CLINICAL GROUPS AND THOSE OBTAINED VIA KMEANS.

Group	N	%	Mean	Med.	p75	p90
Elective surgical	41,471	29.1	3.2	2	3	7
Urgent surgical	7,226	5.1	9.0	6	11	20
Urgent medical	81,353	57.1	6.1	3	7	14
Elective medical	12,520	8.8	9.2	4	10	22
Elderly medical	22,800	16.0	8.2	5	10	19
Elderly surgical	16,497	11.6	4.5	2	5	12
Clusters (KMeans)						
Cluster 0	29,860	20.9	11.6	6	14	27
Cluster 1	37,187	26.1	5.0	3	6	11
Cluster 2	42,434	29.8	3.0	2	3	7
Cluster 3	33,089	23.2	4.3	3	5	9

All random operations used fixed seeds. The interpretability of the tree-based models was obtained using SHAP values [13], which quantify the contribution of each variable to the prediction of each model. This is an attribution internal to the model, not a measure of the causal effect on length of stay. The columns generated by *one-hot* encoding were grouped according to the original variables they represented.

F. Nested Stratification by Diagnostic Chapter

As a complementary analysis, each clinical group was subdivided according to the diagnostic chapter, generating nested subgroups of the form *clinical group* $\times$ *diagnostic chapter*. The goal of this step was to characterize how the length of stay and predictive error varied across different diagnostic chapters within the same clinical profile, without seeking to optimize the overall performance.

This analysis was applied exclusively to the tree-based models, *Random Forest* and *XGBoost*, because they showed better overall performance and greater robustness on smaller subsets than KNN.

Within each subgroup, the diagnostic chapter was constant by definition; therefore, was removed from the set of predictors. The remaining variables, preprocessing, logarithmic transformation of the target variable, and previously optimized hyperparameters, were kept unchanged, such that the only difference relative to the previous analysis was the level of detail in the grouping.

G. Ethical Aspects

This study was approved by the Research Ethics Committee (CEP) of the Botucatu Medical School (FMB) under protocol no. 6.268.072 (CAAE 697961723.1.0000.5411).

III. RESULTS

Table II characterizes the manually defined clinical groups and those obtained through the automatic stratification.

The urgent medical group represents the largest dataset share, accounting for 81,353 admissions (57.1%), followed

by elective surgical patients, with 41,471 records (29.1%). In contrast, urgent surgical patients formed the smallest principal group, representing only 5.1%.

Regarding hospital length of stay, a marked difference was observed among the clinical profiles were evaluated. Elective surgical patients showed the shortest average length of stay, with a mean of 3.2 days and a median of 2 days, suggesting shorter and more predictable admissions. In contrast, elective medical and urgent surgical patients showed the longest average length of stay, with means of 9.2 and 9.0 days, respectively, as well as high 90th percentiles, reaching 22 and 20 days, respectively. These results indicate greater variability and a higher frequency of prolonged admissions in these groups.

The elderly group showed intermediate behavior, though consistently higher than that observed in the general surgical population. Elderly medical patients showed a mean LOS of 8.2 days and a 90th percentile of 19 days, while elderly surgical patients showed a mean LOS of 4.5 days and a median of 2 days, approaching the pattern observed for elective surgical admissions.

Automatic KMeans stratification partitioned the dataset into four *clusters* of balanced size, ranging from 29,860 to 42,434 admissions, whose sum reproduces the total dataset. Cluster 0 concentrated the longest stays, with a mean of 11.6 days and a 90th percentile of 27 days, while Cluster 2 gathered the shortest admissions, with a mean of 3.0 days.

Unlike automatic stratification, the six manually defined clinical groups were not mutually exclusive and summed to 181,867 records, exceeding the total dataset and reflecting the expected overlap between the elderly groups and medical/surgical cuts. Accordingly, every elderly surgical patient is already counted in either elective or urgent surgical, and every elderly medical patient is already counted in either urgent or elective medical category. Elderly surgical patients were distributed between elective surgical (13,656) and urgent surgical (2,841), while elderly medical patients were distributed between urgent medical (21,429) and elective medical (1,371). The four groups defined by type and classification are mutually exclusive and partition the entire dataset, which explains why the percentages sum to more than 100%.

Table III presents the R^2 values of each model by group. Across all groups, *Random Forest* and *XGBoost* showed practically equivalent performance, superior to that of KNN, with the two alternating as to which achieved the best result in each group. Manual clinical stratification outperformed the global model ($R^2 = 0.629$) in five of the six groups, while among the four *clusters* only one reached a higher level.

Complementary analyses indicated that the effect of the logarithmic transformation of the target was not uniform across the models. The transformation benefits *XGBoost*, whose sequential nature is sensitive to the skewness of LOS, and reduces the mean absolute error of all models, although it does not always increase R^2, since this metric penalizes mainly the large errors from prolonged admissions. The transformation was retained in all analyses to preserve the comparability across groups.

Table III
COEFFICIENT OF DETERMINATION (R^2) BY GROUP AND MODEL.

Group	N	KNN	RF	XGB
Complete dataset				
Complete dataset	142,570	0.592	0.625	**0.629**
Manual clinical groups				
Elective surgical	41,471	0.606	0.656	**0.665**
Urgent surgical	7,226	0.660	**0.716**	0.703
Urgent medical	81,353	0.610	0.632	**0.638**
Elective medical	12,520	0.452	**0.503**	0.499
Elderly surgical	16,497	0.632	**0.713**	0.702
Elderly medical	22,800	0.623	0.692	**0.696**
Clusters (KMeans)				
Cluster 0	29,860	0.594	**0.645**	0.643
Cluster 1	37,187	0.421	0.458	**0.463**
Cluster 2	42,434	0.418	**0.472**	0.464
Cluster 3	33,089	0.353	**0.411**	0.393

Table IV
MEAN ABSOLUTE ERROR WITH 95% CONFIDENCE INTERVAL AND ERROR RESTRICTED TO PROLONGED ADMISSIONS, IN DAYS, FOR THE MODEL WITH THE HIGHEST R^2 IN EACH GROUP.

Group	MAE (95% CI)	MAE, prolonged cases
Complete dataset	2.27 [2.21; 2.33]	9.88
Elective surgical	1.38 [1.30; 1.45]	10.22
Urgent surgical	3.11 [2.86; 3.39]	7.81
Urgent medical	2.29 [2.21; 2.38]	8.68
Elective medical	4.69 [4.34; 5.07]	15.01
Elderly medical	3.12 [3.01; 3.26]	8.12
Elderly surgical	1.79 [1.67; 1.92]	7.72

Table IV presents the mean absolute error of each group with a 95% confidence interval, obtained through *bootstrap* resampling on the test set, and the error restricted to prolonged admissions, defined as those above the 90th percentile of the complete dataset, corresponding to 13 days. In the complete dataset, the confidence intervals of *Random Forest* and *XGBoost* overlap almost entirely, whereas those of KNN do not intersect either of the other two, supporting the equivalence between the two tree models and the inferiority of KNN. Errors in prolonged admissions are two to seven times higher than the overall error of each group, and this ratio tends to be larger in groups with a shorter typical length of stay.

IV. DISCUSSION

The equivalence between *Random Forest* and *XGBoost*, both above KNN, is consistent with the LOS prediction literature, in which ensemble tree models tend to lead to [3], [6], [7]. However, the performance varied substantially by group.

Elective surgical. XGBoost shows good performance ($R^2 = 0.665$; MAE 1.38; RMSE 3.78), outperforming the global model in terms of accuracy. The pattern of shorter and more homogeneous admissions contributes to greater predictive stability.

Urgent surgical. The best result among the manual groups ($R^2 = 0.716$; RMSE 5.97) was virtually tied with elderly

surgical patients. It was also the smallest group in the study and had the greatest variability in length of stay among patients who underwent surgery, suggesting that this grouping effectively captured heterogeneity.

Urgent medical. The performance was superior to that of the global model ($R^2 = 0.638$, *XGBoost*; RMSE 5.93), with the largest sample size in the study ($N = 81,353$). The margin over the global model is the smallest among the groups that outperform it, consistent with the fact that this group accounts for more than half of the dataset and, therefore, dominates the global model itself.

Elderly surgical and elderly medical. Both maintain high performance ($R^2 = 0.713$ and 0.696; RMSE 4.05 and 5.49), indicating that predictive patterns are preserved under age-based stratification.

Elective medical. Shows the lowest relative performance ($R^2 = 0.503$; RMSE 10.77) and is the only group that does not outperform the global model. It showed the greatest LOS dispersion among the manual groups (mean 9.2, median 4.0, p90 22), suggesting relevant clinical heterogeneity within the data.

The four KMeans *clusters* reached R^2 values between 0.411 and 0.645, and only one outperformed the global model. This cluster, number 0, is also the one with the greatest LOS dispersion, with a mean of 11.6 days and a standard deviation of 16.8. The other three fell below the reference. It should be noted that R^2 and RMSE are not directly comparable across groups with different LOS. Elective medical patients had an RMSE of 10.8 days and $R^2 = 0.50$, while elective surgical patients had an RMSE of 3.8 and $R^2 = 0.67$, not because one model was worse than the other, but because the groups differed in the variance of LOS. The appropriate reading is to compare each group with the equivalent global model and clustering results, rather than comparing groups with each other in absolute terms.

The performance obtained corroborates the LOS regression literature, although a direct comparison is limited by differences in datasets and available variables, as well as the LOS scale. Mansoori *et al.* [7], in an internal medicine service, report a mean The absolute error was approximately 1.5 days with the optimized *XGBoost*. Slavu *et al.* [6] confirmed the advantage of *Random Forest* over linear regression in predicting post-surgical LOS, and the evaluation of ensemble tree models indicated *XGBoost* as a benchmark for performance and interpretability [3]. The values obtained here, RMSE between 3.8 and 6.0 days in the surgical groups and R^2 up to 0.72, fall within this range, reinforcing that the central contribution lies not in the model itself, but in the clinical stratification of the dataset.

These results support the central argument of this study. Specialist clinical knowledge, by defining the groups, guides machine learning more effectively than segmentation based on data alone. Since the models and hyperparameters are identical across the three sources, the advantage of manual stratification, present in five of the six groups, does not come from the algorithm, but from the grouping itself. The specialist

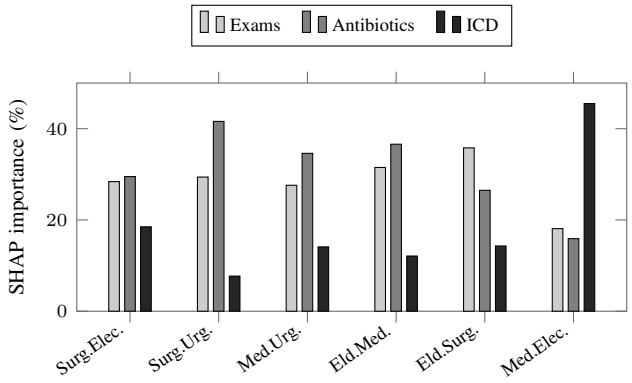


Figure 1. SHAP importance (%) of the three variables with the greatest contribution to model predictions, by clinical group (*Random Forest*). The values reflect model attribution, not a causal relationship with LOS. The ICD chapter dominates only in elective medical patients.

directly translates into rules distinctions that distance-based clustering fails to recover from the raw variables, such as the separation between surgical and medical admissions, between elective and urgent admissions, and by age group. Rather than replacing specialists, the tool establishes a collaborative cycle with them. The clinician proposes the groups, the model predicts, and the interpretability provides evidence on which groups were well captured and which require revision.

Figure 1 makes this cycle concrete. It shows the three variables with the greatest contribution per group, calculated using *Random Forest*, which is equivalent to the best model for all the groups. In the five groups where stratification outperformed the global model, importance was dominated by care-process markers (number of exams and antibiotics), while the ICD chapter always ranked third, with a modest weight (between 7.7% and 18.5%). In the only group that does not outperform the baseline, elective medical patients, this pattern reverses, and the ICD chapter takes first place, with 45.5%, more than double its weight in any other group. Interpretability thus signals where the specialist-defined grouping brings together heterogeneous diagnostic populations and points to the next step, which subdivides elective medical patients by diagnostic chapters.

Table V compares the error of each clinical group to the aggregated error, weighted by N, of its subgroups by the ICD chapter. Subgroups were considered eligible if they had at least 300 total records and at least 100 distinct patients in the test partition, the latter being the decisive criterion, since the effective sample size depends on the number of patients, not admissions; the chapter with missing code was excluded as it does not represent an actual diagnosis. Subdivision consistently reduced the error in the three groups, with drops of 19.4% in elective surgery, 16.9% in elderly surgical patients, and 7.8% in elderly medical patients. In the remaining three groups, the effect is neutral or slightly unfavorable, between $+1.8\%$ and $+11.8\%$, since fragmentation reduces the number of observations per model without sufficiently separating the populations. Coverage of the eligible subgroups relative to

Table V
ERROR OF THE CLINICAL GROUP (BEST MODEL) *versus* AGGREGATED ERROR, WEIGHTED BY N, OF ITS SUBGROUPS BY ICD CHAPTER.

Group	RMSE, group	RMSE, nested	Change
Elective surgical	3.78	**3.05**	-19.4%
Elderly surgical	4.05	**3.36**	-16.9%
Elderly medical	5.49	**5.07**	-7.8%
Urgent surgical	5.97	6.08	$+1.8\%$
Urgent medical	5.93	6.26	$+5.6\%$
Elective medical	10.77	12.04	$+11.8\%$

Table VI
THE NINE DIAGNOSTIC CHAPTERS THAT CONTRIBUTE MOST TO THE TOTAL ERROR OF THE GLOBAL MODEL, RANKED BY THE FRACTION OF ERROR THEY ACCOUNT FOR.

ICD Chapter	% cases	% error	MAE	Mean LOS
Circulatory	11.2	14.8	3.00	7.7
Mental	**3.1**	**13.8**	**10.13**	**15.5**
Infection	5.6	9.6	3.85	12.1
Neoplasm	9.0	8.0	2.02	5.3
Injury	7.7	7.7	2.27	5.6
Respiratory	7.7	7.5	2.20	5.9
Digestive	8.6	6.5	1.73	4.2
Genitourinary	8.1	5.7	1.60	4.6
Pregnancy and childbirth	11.5	5.6	1.10	3.0

each group's test set ranged from 58.8%, in elective medical patients, to 80.2%, in urgent medical patients, which qualifies the reading of the former in particular.

Table VI decomposes the error of the global model by diagnostic chapter. The mental disorder chapter accounts for 3.1% of admissions and 13.8% of the total accumulated error, with a mean absolute error of 10.1 days versus 2 3 days for most other chapters. Within it, 44.3% of admissions exceeded the prolonged-stay threshold, compared with 9.7% in the complete dataset.

The contrast between mental disorders and pregnancy and childbirth defines these two extremes. Pregnancy and childbirth accounted for 11.5% of admissions and only 5.6% of the error, with a mean absolute error of 1.1 days, a result convergent with that obtained by Jain *et al.* [1] when isolating newborns from an independent dataset. Isolating the mental disorder chapter, on the other hand, does not make it predictable, which explains why subdividing elective medical patients did not reduce the aggregated error. Interpretability correctly identified the heterogeneity of the group, but subdivision by diagnostic chapter was not sufficient to capture it when the difficulty lay in the internal variability of a specific chapter.

These results refine the central thesis of this study. The ideal level of detail is not the same for all groups. Going down to the diagnosis level pays off when the grouping brings together populations with distinct length-of-stay patterns, as in the surgical groups and elderly medical patients, and is neutral when the group is already relatively homogeneous. When the difficulty stems from the unpredictability of a specific chapter, as with mental disorders, isolating it does not resolve the issue, because the problem was not the mixing of different chapters, but the nature of that isolated chapter itself.

The care profile of each institution shapes both the groupings that make clinical sense and the length of stay within them. Statewide public datasets, such as SPARCS, make this variation observable, as they bring together hospitals with different missions under the same data structure, allowing a comparison of how length of stay is distributed across them. Predominantly psychiatric facilities, for example, show longer admissions, and the hospital's profile influences the length of stay [14], which would likely lead to different groupings than those adopted here. Institutions with a profile similar to the one studied, on the other hand, may converge toward similar groups.

Before defining the groups and choosing the models, it is necessary to understand the dataset in depth, which involves both consulting the care team, whose experience indicates which clinical distinctions are relevant in that service, and quantitative analysis of the dataset itself, which sizes each candidate group and checks whether the proposed groupings hold up in terms of size and LOS behavior. This preliminary step tends to receive little attention in the literature, even though it shapes all the subsequent decisions.

The results obtained on other datasets point in the same direction. Jain *et al.* [1], working with 2.3 million admissions from SPARCS, separated newborns from other patients and obtained markedly different performances between the two groups, and when restricting the analysis to a single diagnosis type, the performance increased from 0.43 to 0.62. None of these groupings coincide with the ones adopted here, and the variables available in the two datasets also differ, which reinforces the argument that the gain does not come from a particular list of groups, but from the practice of dividing the population according to clinically interpretable criteria and verifying whether the division improves prediction.

Clinical knowledge is not the only path for defining groupings, but it can guide the search and reduce the number of alternatives to be tested. Our group has been comparing different unsupervised clustering algorithms to verify whether this choice affects the ability to outperform the segmentation obtained through statistical methods. This is not a trivial question, since the relative performance of clustering algorithms varies according to the characteristics of the dataset [8], which recommends evaluating them explicitly rather than adopting a method by convention.

The main contributions of this study are twofold. The first is a fair comparison protocol, in which the global model, clinical stratification, and automatic stratification shared exactly the same hyperparameters, fixed per model, isolating stratification as the sole variable. Second, the models were interpreted using SHAP values, identifying the variables with the greatest contribution to predictions within each subdivision.

As a next step, we intend to evaluate a model restricted to the variables of the first moment, which will allow us to quantify how much of the performance is derived from the information accumulated during care. Institutional databases

record the dates of prescription, scheduling, and completion of exams and antibiotic administration, making it possible to reconstruct these indicators within a time window counted from admission, whose width is a parameter to be defined in future work. Depending on this width, exams and antibiotics would move from the third moment to an intermediate point between the second and third, shifting the prediction to before discharge without discarding any process variables.

V. Conclusion

This work assessed whether clinical stratification of the dataset can improved the prediction of hospital length of stay compared with a global model and an automatic clustering-based stratification strategy. Since the models and hyperparameters were kept identical across the three approaches, the comparison allowed the effect of the segmentation strategy to be isolated.

The results showed that specialist-defined stratification outperformed the global model in five of the six groups, whereas automatic stratification based on K-means outperformed it in only one group. These findings support the central argument that specialist clinical knowledge, by defining groups that are more representative of care practice, guides machine learning more effectively than segmentation based solely on data.

Interpretability was demonstrated where performance was lower, and diagnosis played a predominant role in the model's attributions, indicating high internal heterogeneity. Subdivision by diagnostic chapter proved beneficial, with reductions of up to 19.4% in RMSE in three of the six groups, but insufficient for elective medical patients. Decomposing the error by chapter clarified this limitation, revealing that a substantial error share was concentrated in the mental disorders chapter, which accounted for 3.1% of admissions and 13.8% of the total error. In this case, the difficulty stems from the internal variability of the chapter rather than from mixing across chapters.

Thus, this study suggests that predictive models applied to hospital settings can benefit significantly from incorporating specialist knowledge not only as a source of interpretation but also as an active element in defining the modeling strategy. Among the algorithms evaluated, *Random Forest* and *XGBoost* showed equivalent performance, with overlapping confidence intervals, and consistently outperformed KNN. The effect of the logarithmic transformation of the target variable proved to depend on the model and the metric considered, being beneficial for *XGBoost* and the mean absolute error of all models.

Acknowledgments

The authors thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support, the Departamento de Gestão de Atividades Acadêmicas (DGAA) for assistance with data collection, the advisor, Prof. Dr. Allan Felipe Fattori Alves, for guidance throughout the research, and the laboratory team for their collaboration in the development of this work.

References

[1] R. Jain, M. Singh, A. R. Rao, and R. Garg, "Predicting hospital length of stay using machine learning on a large open health dataset," *BMC Health Services Research*, vol. 24, p. 860, 2024.

[2] Y.-Y. Jo *et al.*, "Prediction of prolonged length of hospital stay after cancer surgery using machine learning on electronic health records," *JMIR Medical Informatics*, vol. 9, no. 2, art. e23147, 2021.

[3] B. Peng and S. Gao, "Prediction of hospital length of stay: leveraging ensemble tree models and intelligent feature selection," *Journal of Hospital Management and Health Policy*, vol. 9, 2025.

[4] M. P. Sendak, J. D'Arcy, S. Kashyap, M. Gao, M. Nichols, K. M. Corey, W. Ratliff, and S. Balu, "A path for translation of machine learning products into healthcare delivery," *EMJ Innovations*, 2020.

[5] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[6] I. Slavu, R. Tulin, A. Dogaru, I. Dima, C. Orlov Slavu, D.-E. Gheoca Mutu, and A. Tulin, "Machine learning modeling of hospital length of stay after breast cancer surgery: comparison of random forest and linear regression approaches," *Medicina*, vol. 62, no. 1, art. 88, 2025.

[7] A. Mansoori, M. Zeinalnezhad, and L. Nazarimanesh, "Optimization of tree-based machine learning models to predict the length of hospital stay using genetic algorithm," *Journal of Healthcare Engineering*, vol. 2023, art. 9673395, 2023.

[8] M. Ambigavathi and D. Sridharan, "Analysis of clustering algorithms in machine learning for healthcare data," in *Advances in Computing and Data Sciences (ICACDS 2020)*, Communications in Computer and Information Science, vol. 1244, Singapore: Springer, 2020, pp. 117–128.

[9] C.-S. Yang, C.-P. Wei, C.-C. Yuan, and J.-Y. Schoung, "Predicting the length of hospital stay of burn patients: comparisons of prediction accuracy among different clinical stages," *Decision Support Systems*, vol. 50, no. 1, pp. 325–335, 2010.

[10] D. Zikos, A. Shrestha, and L. Fegaras, "Estimation of the mismatch between admission and discharge diagnosis for respiratory patients, and implications on the length of stay and hospital charges," *AMIA Joint Summits on Translational Science Proceedings*, vol. 2019, pp. 192–201, 2019.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[14] S. Lee, A. B. Rothbard, and E. L. Noll, "Length of inpatient stay of persons with serious mental illness: effects of hospital and regional characteristics," *Psychiatric Services*, vol. 63, no. 9, pp. 889–895, 2012.

Evaluation of a Phase-Delay Analytical Method for Non-Invasive Blood Glucose Estimation Using Dual-Wavelength Photoplethysmography Signals

Filipe A. C. Oliveira
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0003-2356-2762

Estela Ribeiro
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0002-2193-6866

Diego A. C. Cardenas
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0002-8846-5202

Marcelo A. F. Toledo
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0001-6585-0354

Ramon A. Moreno
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0002-0548-9365

Jose E. Krieger
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0001-5464-1792

Marco A. Gutierrez
Heart Institute (InCor)
University of Sao Paulo
Sao Paulo, Brazil
https://orcid.org/0000-0003-0964-6222

***Abstract*—Non-invasive blood glucose (BG) monitoring using photoplethysmography (PPG) represents a promising frontier in diabetes management, yet the true clinical viability of these methods relies on their ability to generalize across distinct data recordings. This work presents an evaluation of the generalization capacity of an analytical method proposed by Nakazawa et al. (2024), which is based on the phase delay between near-infrared spectroscopy (NIRS) signals derived from dual-wavelength PPG. We carried out this evaluation by collecting paired PPG and reference BG data from eight subjects under four distinct profiles of induced BG fluctuations. The analytical method was then calibrated to each subject and evaluated using three distinct data partitioning approaches. The first configuration replicated the original validation scheme by calibrating and validating the method over the same data cohort, generating optimistic performance metrics: a mean absolute relative difference (MARD) of 10.67% and a root mean squared error (RMSE) of 12.36 mg/dL. The second approach evaluated the generalizability of the method by partitioning the data into distinct calibration and validation cohorts, which resulted in a relevant performance decrease, achieving a MARD of 25.13% and an RMSE of 33.08 mg/dL. A third validation approach involved testing the calibration curve fitted on the first approach over the same validation data subset employed in the second, achieving considerably lower error rates (MARD of 9.68%, RMSE of 10.72 mg/dL). Our findings underline a key limitation in the method's performance when tested on unseen data, highlighting the importance of evaluating non-invasive BG estimation techniques over partitioned data to assess their generalization capacity and assess their clinical reliability under real-world scenarios.**

***Index Terms*—photoplethysmography, glucose, spectroscopy, diabetes**

I. Introduction

Diabetes mellitus is a condition affecting approximately one out of ten adults globally [1]. When left untreated, it leads to a chronically elevated blood glucose (BG) concentration, causing several long-term complications such as retinopathy, cardiovascular diseases, and kidney failure [2]. Current methods for monitoring BG levels include self-monitoring blood glucose (SMBG) devices, which measure glucose through finger-prick capillary tests, and continuous glucose monitors (CGMs), devices that read glucose concentration from the interstitial fluid with an intradermal sensor.

The financial burden and invasive nature of these monitoring techniques impose a significant strain on healthcare systems and lower patient quality of life due to physical discomfort, ultimately limiting treatment adherence [1], [3]–[5]. Consequently, non-invasive and continuous alternatives for BG monitoring using optical techniques like photoplethysmography (PPG) have become a major research focus due to limitations of current monitoring methods [6], [7].

PPG is a low-cost optical technique used to detect localized blood volume variations within vascularized tissue. The signal is acquired by emitting light into the tissue and measuring the intensity of the reflected or transmitted light, which fluctuates over time due to the pulsatile nature of blood flow driven by cardiac activity. The resulting PPG waveform consists of a pulsatile (AC) component reflecting the cardiac cycle, and a low-frequency (DC) component shaped by tissue vascularizа-

tion, respiration, and blood composition [8]. Since variations in the concentration of blood constituents alter its optical absorption characteristics, multi-wavelength PPG signals have emerged as a potential method for non-invasive blood glucose estimation [9]. So far, most studies have employed machine learning (ML)-based techniques to extract relevant BG-related information from multi-wavelength signals [9]–[11].

For instance, [12] proposed a wearable ring to collect multi-channel PPG data paired with a wavelet transform framework to decompose PPG waveforms into derivative signals. Features extracted from PPG data collected from 47 subjects alongside SMBG readings were used to train a linear support vector regression model, achieving a root mean squared error (RMSE) of 16.02 mg/dL and a coefficient of determination (R^2) of 0.8. Similarly, by combining 1D-ResNet models with a self-attention mechanism, [13] achieved an RMSE of 19.24 ± 1.13 mg/dL using 6-wavelength PPG signals collected from 254 individuals, coupled with demographic profiles and local environmental parameters such as humidity and temperature. Furthermore, [11] explored the relationship among PPG data collected the in red (655 nm), infrared (940 nm), and green (520 nm) wavelengths. Using a dataset of 92 subjects annotated with SMBG references, they trained a Lasso regression model on amplitude-ratio features, yielding an RMSE of 7.18 ± 5.66 mg/dL.

Recently, Nakazawa *et al.* [14] introduced a purely analytical technique for blood glucose estimation that circumvents ML complexities and computational demands. Their method estimates blood glucose using phase delay and amplitude variations between oxyhemoglobin (HbO_2) and deoxyhemoglobin (Hb) near-infrared spectroscopy (NIRS) signals, derived from dual-wavelength PPG, to compute a novel metric named the Metabolic Index (MI). Although the authors reported a strong correlation ($r = 0.73$) and an RMSE of 19.7 mg/dL, their evaluation was confined to a single subject ($n = 1$) and lacked a data split strategy that calibrates and validates the method on distinct signals.

This study aims to implement and evaluate the analytical BG estimation method proposed in [14] using a comprehensive, multi-subject dataset constructed through a similar acquisition protocol. Our dataset consists of 8 individuals who underwent four distinct beverage-based interventions. These interventions were designed not only to induce blood glucose variability but also to evaluate the impact of caffeine on the MI and provide a baseline control. Furthermore, three evaluation strategies are implemented: first, using the entire data collected from a given subject to fit the calibration curve exactly as proposed in [14]; second, partitioning each individual's data into distinct calibration and validation subsets; and third, using the method calibrated on the first approach and evaluating it on the validation data subset of the second, allowing a direct comparison between methods fitted on distinct data cohorts. These partitioning strategies allowed us to evaluate the analytical method's performance on different data subsets, providing an assessment that may reflect a performance closer to a real-world application.

II. Methodology

This section describes the methodology used in our work. The acquisition protocol for database construction is outlined, as well as the analytical method for BG estimation and its evaluation approaches. The methodology proposed by [14] to estimate the Metabolic Index is summarized in the diagram shown in Fig. 1.

A. Acquisition Protocol

A dedicated database was built using the Analog Devices MAXREFDES104 wearable development platform to collect synchronized PPG signals at a sampling rate of 125 Hz across two wavelengths: red (655 nm) and infrared (940 nm). Although configured as a smartband, the sensor was adapted to perform fingertip acquisition to maximize the signal-to-noise ratio [15]. Reference BG values were collected using an Abbott FreeStyle Libre CGM, which provides automatic readings every 15 minutes. The CGM was installed at least 48 hours prior to PPG data collection to mitigate the higher measurement errors typically reported at the beginning of the sensor's lifespan [16].

PPG signals and paired BG measurements were acquired during individual acquisition sessions lasting approximately 60 minutes. Before starting each session, volunteers were required to fast for at least 2 hours to minimize BG variations caused by external dietary factors. During data collection, participants were instructed to remain comfortably seated, resting both arms on a table with the PPG sensor positioned on the index finger of the non-dominant hand. Volunteers were also advised to avoid talking and moving throughout the acquisition to reduce motion artifacts in the PPG signal.

The first 10 minutes of each session served as a baseline PPG data collection, immediately followed by an intervention characterized by the intake of a specific beverage. Each volunteer participated in 4 sessions with an interval of at least 24 hours between sessions. The beverage's composition was session-specific, alternating sugar and caffeine content according to a predetermined order: the beverage in the session **I** contained sugar and caffeine; session **II** contained only caffeine; session **III** only sugar; and **IV** plain water. Fig. 2 illustrates the acquisition session order, beverage contents and data collecting protocol.

In total, 8 non-diabetic volunteers completed all four acquisition sessions, yielding 32 distinct recordings with duration of approximately 60 minutes each. The cohort included one female and seven male participants, aged between 27 and 52 years (mean age of 39.1 ± 10.3 years). The reference BG measurements recorded across all sessions ranged from 57 to 165 mg/dL, with a mean and standard deviation of 97.9 ± 16.1 mg/dL.

This clinical experiment fully adheres to all pertinent ethical regulations and was approved by the Institutional Review Board (IRB) under the registration SDC 5248/21/023.

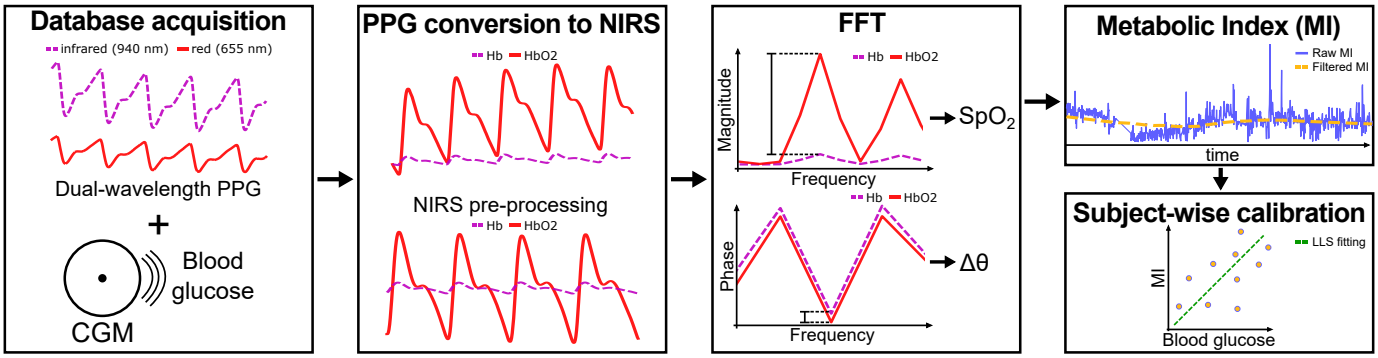


Fig. 1. Representation of the analytical methodology for blood glucose estimation. The magnitude spectrum derived from the FFT allows the estimation of the oxygen saturation (SpO_2). $\Delta\theta$ denotes the phase difference between HbO_2 and Hb NIRS signals. The MI computed by this method is used to estimate the BG by fitting a calibration curve.

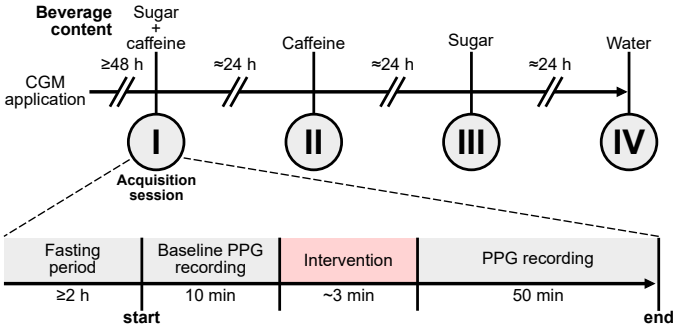


Fig. 2. Schematic representation of the sequence of acquisition sessions and beverage content of each session. Blood glucose measurements are performed automatically by the CGM every 15 minutes.

B. Analytical Method for BG Estimation

The first step of the analytical method proposed in [14] involves converting the raw PPG signals into HbO_2 and Hb NIRS signals, which represent changes in the concentration of these substances over time.

1) PPG Conversion to NIRS: Converting PPG waveforms to NIRS signals depends on the specific extinction coefficients of Hb and HbO_2 at the selected wavelengths (655 and 940 nm), which were sourced from [17]. These NIRS signals can be estimated by solving (1), a matrix formulation derived from the modified Beer-Lambert Law [18]:

$$\begin{bmatrix} N_{HbO_2}(t) \\ N_{Hb}(t) \end{bmatrix} = \begin{bmatrix} \varepsilon_{HbO_2}(\lambda_1) & \varepsilon_{Hb}(\lambda_1) \\ \varepsilon_{HbO_2}(\lambda_2) & \varepsilon_{Hb}(\lambda_2) \end{bmatrix}^{-1} \begin{bmatrix} \Delta OD(\lambda_1, t) \\ \Delta OD(\lambda_2, t) \end{bmatrix}, \quad (1)$$

where $N_{HbO_2}(t)$ and $N_{Hb}(t)$ represent the time-varying NIRS signals for HbO_2 and Hb, respectively. The parameters $\varepsilon_{HbO_2}(\lambda)$ and $\varepsilon_{Hb}(\lambda)$ denote the specific molar extinction coefficients of oxyhemoglobin and deoxyhemoglobin at a given wavelength λ. The term $\Delta OD(\lambda, t)$ represents the incremental optical density change between the baseline instant t_0 and a given time t for the wavelength λ, which is defined as:

$$\Delta OD(\lambda, t) = log_{10}\left(\frac{I_{\text{out}}(\lambda, t_0)}{I_{\text{out}}(\lambda, t)}\right), \quad (2)$$

where $I_{\text{out}}(\lambda, t_0)$ and $I_{\text{out}}(\lambda, t)$ correspond to the reflected light (PPG samples) detected at the wavelength λ at time instants t_0 and t, respectively.

2) NIRS Processing: In order to remove high-frequency noise and low-frequency baseline oscillations, the reconstructed NIRS signals are filtered using a second-order band-pass Butterworth filter with cutoff frequencies set at 0.8 and 10 Hz.

The filtered NIRS waveforms are subsequently segmented into windows spanning exactly 4 cardiac cycles, trimming the boundaries so that each segment aligns from the first cycle onset to the last one completion. To maximize the computational efficiency of the subsequent Fast Fourier Transform (FFT) and prevent spectral leakage, each segment is resampled to increase its length to the nearest power of two and multiplied by a Hamming window.

The FFT is then applied to each windowed segment to isolate the main frequency, which represents the fundamental heart rate frequency, and extract spectral information from the signals. The AC amplitudes of the synchronized HbO_2 and Hb NIRS segments are extracted from the magnitude spectrum at this dominant frequency, denoted by $A_{HbO_2}(t)$ and $A_{Hb}(t)$, respectively. These amplitudes are then used to estimate the continuous oxygen saturation (SpO_2) over time, computed as follows:

$$SpO_2(t) = \frac{A_{HbO_2}(t)}{A_{HbO_2}(t) + A_{Hb}(t)} \quad (3)$$

The resulting SpO_2 is subsequently used for estimating the Metabolic Index.

3) Estimating the Metabolic Index: The MI estimation method was formulated by assuming a slight and slow-varying phase difference, denoted as $\Delta\theta(t)$, between the synchronized N_{HbO_2} and N_{Hb} segments at the fundamental heart rate frequency [14]. By coupling the phase delay with the instantaneous oxygen saturation ($SpO_2(t)$), the time-dependent metabolic index is derived as follows:

$$MI(t) = SpO_2(t) \cdot [1 - SpO_2(t)] \cdot \Delta\theta(t) \quad (4)$$

The resulting MI values are subsequently grouped into 1-minute non-overlapping windows. Any outliers within each window are then identified and removed before averaging the remaining valid entries of a same window into an individual sample. Finally, the sequence of mean MI values is smoothed using a Savitzky-Golay filter with a first-order polynomial and a window length of 29 points to yield a stable MI series.

C. Proposed Evaluation Methods

After smoothing the resulting MI values, a calibration curve is mapped for each subject by linear least square (LLS) fitting between the estimated MI values and the reference BG measurements. Consequently, each calibration curve is specifically adjusted for one single individual. In [14], the authors use all computed MI values of one subject to fit a calibration curve with the known reference BG concentrations.

Since, in our experiment, glucose measurements were recorded with a 15-minute interval, linear interpolation was applied to BG readings in order to pair a reference BG value to each MI data point. To avoid extrapolating BG values, CGM measurements acquired immediately prior to and following PPG data collection were also used to interpolate the reference BG values, resulting in 4364 ± 456 samples per subject across all acquisition sessions. From the fitted calibration curve, any subsequently calculated MI sample can be mapped directly to a BG level estimation.

Blood glucose estimation methods, however, typically enforce an independent validation strategy by partitioning the dataset [11]–[13]. This approach ensures that the mathematical model is calibrated (or trained, in the case of ML-based methodologies) using one specific data subset and subsequently evaluated on entirely unseen data.

This partitioning technique is critical for objectively analyzing a model's generalization capacity, which cannot be reliably assessed when training and testing on overlapping temporal data of a single subject. Evaluating performance on unseen data is relevant to demonstrate that the model does not simply overfit over calibration/training data [19].

Therefore, in addition to employing the original partitioning scheme on [14], we implemented a validation strategy that partitions each subject's data into separate calibration and validation cohorts, where the data for each cohort were gathered from distinct acquisition sessions. Data within the calibration cohort were used exclusively to fit the calibration curve specific for each subject, while data from the validation cohort were subsequently used to evaluate the generalization capacity of the fitted curve on unseen segments.

The calibration set was composed of data collected during sessions **I** and **IV** of a given subject, which employed, respectively, an intervention consisting of a sugary and caffeinated beverage, and a baseline intervention of plain water. The validation set comprised the remaining two sessions (**II** and **III**), which involved beverages containing either sugar or caffeine alone.

This specific data partitioning scheme ensured that both the calibration and validation cohorts contained a balanced distribution of experimental conditions. Specifically, each cohort included a sugar-driven intervention to induce BG variations (sessions **I** and **III**) alongside a caffeinated beverage intervention (sessions **I** and **II**), thereby mitigating potential confounding effects of caffeine on the estimated MI values.

To allow a direct comparison between the described data partitioning schemes, a third approach was implemented. This approach involves taking the calibration curve fitted on all acquisition sessions of a given subject from the first approach and evaluating it on the data subset used for validation in the second approach. This configuration allows a direct comparison between the distinct calibration curves fitted in first two approaches by evaluating them on a same validation data cohort.

The data partitioning is summarized as follows:

- Approach 1: The calibration curve for each subject was fitted and evaluated using the data collected across all four acquisition sessions, mirroring the method described in [14].
- Approach 2: The calibration curve was fitted using data from sessions **I** (sugar+caffeine) and **IV** (water), and validated using the data from the remaining sessions **II** (caffeine) and **III** (sugar).
- Approach 3: The calibration curve derived in approach 1 was evaluated on the same validation data subset utilized in approach 2 (**II** and **III**).

Given the independence of individual acquisition sessions, time-dependent factors were not considered in the evaluation analysis. Fig. 3 summarizes these distinct evaluation approaches.

D. Evaluation

The BG estimation error was quantified using the Mean Absolute Relative Difference (MARD), which computes the error as a percentage of the reference and is broadly used for evaluating glucometers [4]. Additionally, we computed the RMSE, which penalizes larger errors due to its quadratic

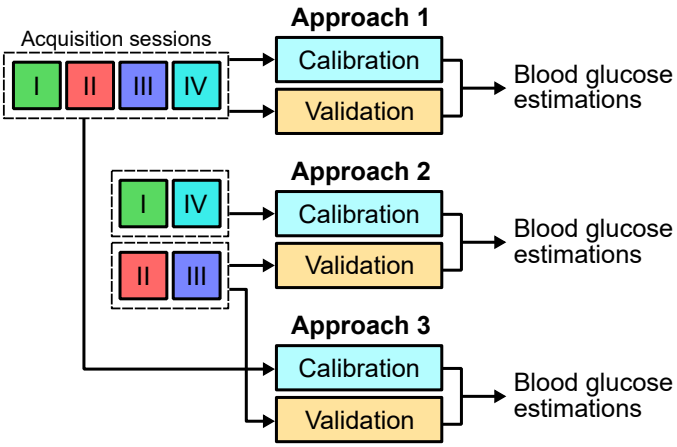


Fig. 3. Data partitioning according to each evaluation approach. A calibration curve is generated for each subject.

formulation. The Pearson correlation coefficient (r) was calculated to assess how effectively the estimated BG values tracked reference BG variations.

To assess whether the observed performance differences between the evaluation approaches 2 and 3 were statistically significant, paired Wilcoxon signed-rank tests were performed on the subject-wise performance metrics (MARD, RMSE, and r) across all eight subjects. Effect sizes were evaluated using the matched-pairs rank-biserial correlation coefficient (r_{rb}), and estimation precision was quantified using 95% confidence intervals (CIs) for the median paired differences. A non-parametric test was selected due to the cohort size ($N = 8$). Statistical significance was initially set at $\alpha = 0.05$. To account for multiple comparisons across the three evaluated metrics, a Bonferroni correction was applied, resulting in an adjusted significance of $\alpha_{\text{adjusted}} = 0.0167$. To assess for the effect of caffeine on the MI estimations, the Wilcoxon signed-rank test was also employed to estimate statistically significant differences between median MI values estimated per subject between sessions **I** (sugar+caffeine) and **III** (sugar), and between sessions **II** (caffeine) and **IV** (water).

The clinical impact of the estimations was evaluated using the Clarke Error Grid (CEG), a standard tool that categorizes glucose predictions into five zones (A through E) based on their influence on patient treatment decisions [20]. Estimates within zones A and B represent clinically acceptable outcomes that do not lead to inappropriate medical interventions.

Finally, method sensitivity to CGM lag compared to actual BG was assessed by shifting the reference CGM measurements from 1 to 15 minutes and re-evaluating all performance metrics (RMSE, MARD, r).

III. Results

This section details the performance evaluation of the analytical method proposed in [14] over a multi-subject dataset, using the three validation approaches outlined in Section II-C. Since the calibration curves were individually fitted for each subject, the evaluation metrics were computed for each individual and subsequently averaged. Their mean values and standard deviations are summarized in Table I.

The paired Wilcoxon signed-rank test indicated statistically significant differences between approaches 2 and 3 for RMSE ($W = 0.0$, $p = 0.0078$, $r_{\text{rb}} = 1.0$, 95% CI: $[-44.65, -1.16]$ mg/dL) and MARD ($W = 1.0$, $p = 0.0156$, $r_{\text{rb}} = 0.94$, 95% CI: $[-34.99, -0.21]$%), but revealed no significant difference for the correlation coefficient r ($W = 7.0$, $p = 0.2719$, $r_{\text{rb}} = -0.5$, 95% CI: $[-0.00, 0.00]$).

TABLE I
Validation Metrics

	Approach 1	Approach 2	Approach 3
MARD (%)	10.67 ± 2.37	25.13 ± 18.42	9.68 ± 2.47
RMSE (mg/dL)	12.36 ± 3.61	33.08 ± 30.57	10.72 ± 2.72
r	0.28 ± 0.16	0.28 ± 0.37	0.40 ± 0.22

No statistically significant differences were found between the MI values estimated in sessions containing caffeine and their counterparts without caffeine, as indicated by the Wilcoxon signed-rank test between sessions **I** and **III** ($W = 10.0$, $p = 0.3125$, $r_{\text{rb}} = 0.44$), and between sessions **II** and **IV** ($W = 3.0$, $p = 0.0391$, $r_{\text{rb}} = 0.83$).

The CEG generated for the approach 1, employing data from all acquisition sessions for both calibration and validation, is shown in Fig. 4.

Analysis of the CEG shown in Fig. 4 indicates that 98.5% of the estimations fell within the clinically acceptable zones A and B (with 90.8% in zone A and 7.7% in zone B). The remaining 1.5% of the predictions fell into zone D. In this scenario, these errors indicate a failure to detect hypoglycemia (BG concentrations below 70 mg/dL), which could lead to dangerous treatment omissions for the patient.

The CEG containing predictions estimated on the validation cohort of approach 2 is presented in Fig. 5.

As illustrated in Fig. 5, 96.2% of the estimations fell within zones A and B (60.1% in zone A and 36.1% in zone B). The remaining 3.9% fell into zone C, which could lead to unnecessary or inappropriate medical interventions; specifically, these errors falsely indicate hyperglycemia (BG levels over 180 mg/dL) when the true BG concentration is actually closer to a 100 mg/dL baseline.

The CEG containing validation predictions evaluated under the data partitioning scheme of approach 3 is presented in Fig. 6. For this approach, 100.0% of the estimations fell within the clinically acceptable zones A and B, with 95.1% of the predictions located in zone A and 4.9% in zone B.

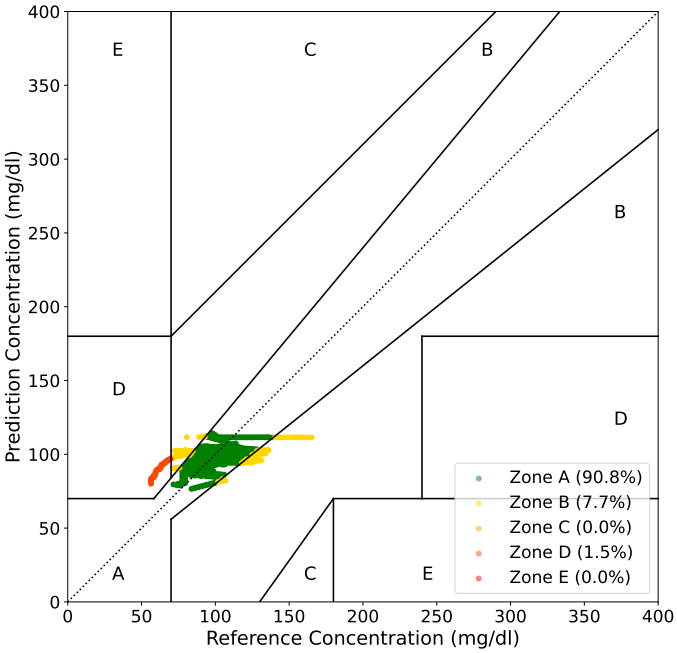


Fig. 4. Clarke Error Grid generated using the validation dataset for approach 1. This configuration replicates the evaluation method proposed in [14], where data from all acquisition sessions are utilized for both calibration and validation.

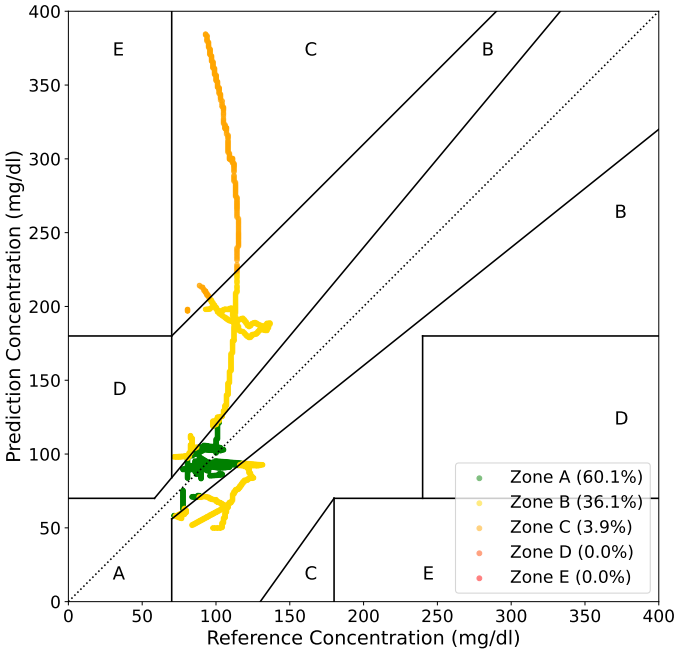


Fig. 5. Clarke Error Grid generated with estimations performed using a calibration curve fitted and validated over distinct data cohorts (approach 2). This strategy utilizes data from two acquisition sessions (**I** and **IV**) to calibrate the method for a single subject, while the validation is performed using data from the remaining sessions (**II** and **III**).

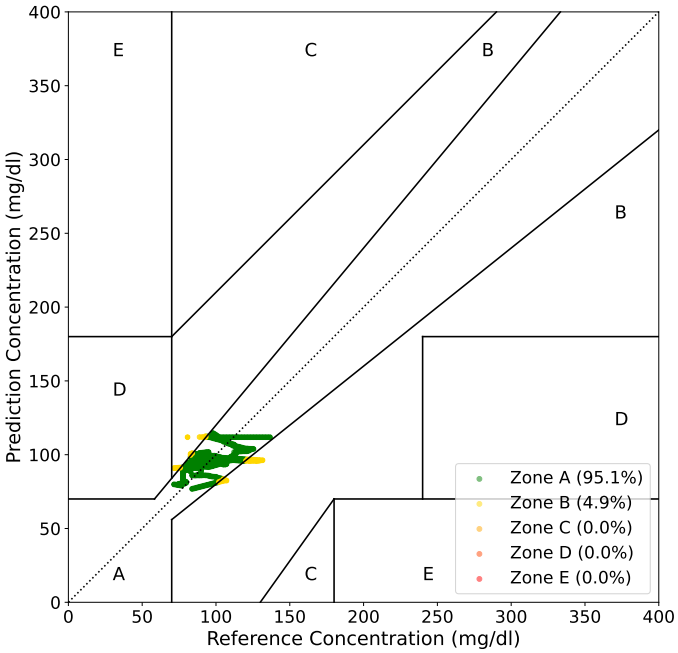


Fig. 6. Clarke Error Grid generated using estimations from the method calibrated with data from all acquisition sessions and validated on a subset cohort composed of data from sessions **II** and **III** (approach 3).

Lastly, simulating CGM response delays from 1 to 15 minutes resulted in a consistent reduction in performance across all approaches. At the 15-minute delay, RMSE increased to 12.91 mg/dL, 36.22 mg/dL, and 11.49 mg/dL, MARD reached 11.00%, 28.21%, and 10.14%, and r dropped to 0.16, 0.04, and 0.14 for approaches 1, 2, and 3, respectively.

IV. Discussion

The comparative analysis across the three distinct evaluation approaches using data gathered from eight subjects suggests a significant limitation regarding the generalization capabilities of the analytical method proposed by [14] when evaluated on a multi-subject dataset.

Approach 1 mirrored the strategy employed in [14] by fitting and validating the calibration curve of the proposed method over a single data cohort. This strategy achieved a mean MARD of $10.67 \pm 2.37\%$ and an RMSE of 12.36 ± 3.61 mg/dL, with 98.5% of estimations clustered within the clinically acceptable zones A and B. The lack of validation on a distinct data cohort, however, masks the real performance of the method on unseen data.

The importance of a separate validation cohort becomes evident when examining the results of approach 2, where the data partitioning scheme fits and evaluates the method on data from entirely distinct acquisition sessions. Under this approach, the mean MARD increases to $25.13 \pm 18.42\%$ and the RMSE rises to 33.08 ± 30.57 mg/dL. The standard deviations increase significantly, highlighting a lack of stability across different subjects. This instability is reinforced by the estimation distribution in the CEG shown in Fig. 5, where 3.9% of the data points fell into zone C, which could lead to unnecessary clinical interventions. Although the remaining 96.2% of the predictions fell within zones A and B, the wide dispersion of the data points suggests a major underlying uncertainty in the method and does not necessarily indicate clinical applicability.

Approach 3 achieved the highest overall accuracy among the other evaluation strategies, reaching a MARD of $9.68 \pm 2.47\%$ and an RMSE of 10.72 ± 2.72 mg/dL, alongside an improved correlation coefficient ($r = 0.40 \pm 0.22$). Furthermore, 100.0% of its CEG data points fell within zones A and B, which would normally indicate an optimal performance profile for a blood glucose estimation method. This approach, however, served primarily as a control strategy to allow a direct comparison between the calibration curves fitted for approaches 1 and 2. Approach 3 represents an optimistic upper bound, since the calibration model was fitted using information originating from the validation sessions, similarly to was was observed to approach 1. By comparing the results of a calibration curve fitted with data from all sessions and a curve fitted with a data subset against a same validation cohort, we isolated the impact of both calibration strategies and provided a direct comparison between both configurations.

The discrepancy in results between approaches 2 and 3, both evaluated on the exact same validation cohort, indicates that this limitation is not merely an artifact of the dataset itself. Approach 3 demonstrated a statistically significant reduction in both RMSE and MARD compared to approach 2. However, wide 95% confidence intervals were observed for both metrics, spanning 43.49 mg/dL for RMSE and 34.78% for

MARD. While the entirely negative bounds confirm a consistent reduction in prediction error for approach 3, the span of these intervals indicates high inter-subject variability, which is further influenced by the limited cohort size of 8 subjects. No statistically significant difference and a 95% confidence interval spanning zero were observed for r between approaches 2 and 3. This outcome is explained by the properties of the LLS calibration model, which is invariant to linear scale and shift transformations for a given dataset. Consequently, LLS recalibration preserves the magnitude of r, allowing only for a sign inversion if a positive or negative correlation is identified between MI values and BG measurements across approaches.

While accounting for the dataset limitations, the resulting comparison of approaches 2 and 3 suggests that the calibration partition in approach 2 provided insufficient information for generalization, or that the analytical model failed to encapsulate the necessary physiological features to track blood glucose fluctuations over time. This limitation is also reflected in the low correlation coefficients observed across all configurations (mean $r = 0.28$ for approaches 1 and 2, and $r = 0.40$ for approach 3). The similarity between the metrics achieved by the method evaluated across our multi-subject dataset and the single-subject results originally reported in [14], detailed in Table II, further reinforces these structural limitations.

As shown in Table II, evaluating the analytical method over a multi-subject dataset ($n = 8$) by calibrating and validating within an identical data cohort improved almost all performance metrics relative to those reported for the single-subject implementation. The sole metric that did not show an improvement was the correlation coefficient, r. Additionally, the expected clinical utility improved, as evidenced by the increased proportion of estimations falling within the clinically acceptable CEG zones A and B.

However, this overall improvement relative to the original work is absent when the method is calibrated and validated using separate data cohorts, as demonstrated by approach 2. This limitation is evidenced by the error increases reported through RMSE and MARD, alongside a decrease in the correlation coefficient, r. Moreover, the proportion of estimations falling within the CEG zones A and B showed a difference of less than 10%. Therefore, the comparison of our finding with the results reported in [14] highlights that the generalization capacity of this method must be carefully evaluated before considering its implementation in real-world applications.

TABLE II
COMPARISON OF METRICS ACHIEVED IN THIS WORK WITH [14]

	This work		**[14]**[2]
	Approach 1	**Approach 2**	
MARD	10.7%[1]	25.1%[1]	17.5%
RMSE	12.4 mg/dL[1]	33.1 mg/dL[1]	24.1 mg/dL
r	0.28[1]	0.28[1]	0.66
CEG zone A	90.8%	60.1%	69.2%
CEG zone B	7.7%	36.1%	30.8%
Subjects (n)	8	8	1

[1]Mean value.

[2]Results reported without α-correction.

The importance of evaluating BG estimation methods using complementary statistical metrics (RMSE, r, and MARD) alongside clinical impact assessment tools, such as the CEG analysis, is reinforced by these findings. The poor performance indicated by the relatively high RMSE and low r achieved in approach 2 is not reflected in the clinical utility measured by the CEG, which placed 96.2% of predictions in the clinically acceptable zones A and B. This divergence occurs primarily due to the high prevalence of reference BG values within the normal range: namely, measurements outside the hypoglycemic (below 70 mg/dL) or hyperglycemic (above 180 mg/dL) intervals, even when accounting for sugar-driven interventions. Increasing the proportion of diabetic subjects in the dataset would potentially expand the range of reference blood glucose values, enabling a more accurate and reliable assessment of the clinical viability of the method.

Therefore, factors such as the relatively small dataset (8 subjects), the low prevalence of diabetic subjects, the concentration of BG measurements within the normal range, imbalanced sex distribution, and the fixed order of interventions described in our study can be considered limitations to the evaluation of the method proposed in [14].

In terms of practical implementation, this method presents a relevant limitation due to its reliance on subject-wise calibration. Although calibrating individual models reduces inter-subject variability, requiring a personalized fitting procedure for every user substantially limits the method's clinical applicability.

Our results demonstrated no measurable effect of caffeine on the MI, as indicated by the paired Wilcoxon signed-rank tests performed between sessions **I** and **III**, and between sessions **II** and **IV**, where both yielded $p > \alpha_{\text{adjusted}}$, revealing no statistically significant differences attributable to the presence of caffeine.

Besides the evaluation on different datasets, specific variations between our data collection protocol and that reported in [14] may have contributed to the divergence in results. In this work, due to subject time constraints, we acquired approximately one hour of data per session, whereas the authors in [14] reported collecting approximately 90 minutes.

In [14], the authors captured PPG data using wavelengths of 650 and 930 nm, which is slightly different from the 655 and 940 nm employed in our work. Additionally, they reported collecting PPG signals via a smartphone-based camera. The impact of these hardware and wavelength variations is likely minimal, as we utilized the exact molar extinction coefficients of HbO_2 and Hb corresponding specifically to our active operating wavelengths.

The authors in [14] also performed a post-processing adjustment that incorporates an α-correction factor into the MI estimation method, which was tuned based on its direct impact on the resulting r. This factor was not implemented in our

work, and all comparisons with [14] were conducted using their reported metrics that excluded the α-correction.

Lastly, the CGM measures glucose concentrations in the interstitial fluid, where levels fluctuate with an intrinsic lag relative to blood glucose levels [21]. However, applying delay corrections spanning shifts from 1 to 15 minutes to the reference CGM measurements failed to improve overall performance, suggesting the presence of intrinsic method limitations rather than uncorrected reference lag.

V. Conclusion

This work evaluated the generalization capacity of an analytical method for non-invasive blood glucose estimation originally proposed in [14]. This assessment was conducted by collecting PPG and blood glucose data from 8 subjects and employing this dataset to calibrate and validate the model under three distinct validation approaches. Calibrating and validating the method within the same data cohort for each subject yielded optimistic metrics. However, while this approach mirrors the validation scheme performed in [14], it restricts the evaluation of the true generalization capacity of the method, since it is not evaluated on unseen data. The second approach, which employed s data partitioning scheme based on distinct acquisition sessions, demonstrated a relevant performance decrease, suggesting a key limitation in the method's ability to generalize. Validating the full-session calibration model on the same data subset as the second approach reinforced this limitation on a direct comparison. Limitations were also observed regarding the clinical applicability of the method, as an individual calibration procedure is required for each new user.

Therefore, our findings suggest that the analytical method proposed in [14] must be thoroughly evaluated across independent cohorts with a larger sample size and a broader glucose range to establish a reliable correlation between the Metabolic Index and blood glucose fluctuations. Moreover, this work highlights the need of testing non-invasive BG estimation methods on entirely unseen data to assess their clinical reliability on a real-world scenario.

Acknowledgment

This study was financially supported in part by Foxconn Brazil and the Zerbini Foundation as part of the research project "Machine Learning in Cardiovascular Medicine".

References

[1] D. Magliano and E. J. Boyko, *Diabetes atlas*. Brussels: International Diabetes Federation, 11th edition ed., 2025.

[2] The Diabetes Control and Complications Trial (DCCT)/Epidemiology of Diabetes Interventions and Complications (EDIC) Study Research Group, "Intensive Diabetes Treatment and Cardiovascular Outcomes in Type 1 Diabetes: The DCCT/EDIC Study 30-Year Follow-up," *Diabetes Care*, vol. 39, pp. 686–693, May 2016.

[3] G. Toletti, A. Boaretto, C. Di Loreto, R. Fornengo, A. Gigante, and G. Perrone, "Enhancing diabetes therapy adherence: a comprehensive study on glucometer usability for type 2 diabetes patients," *Frontiers in Clinical Diabetes and Healthcare*, vol. 5, p. 1328181, May 2024.

[4] R. Lin, F. Brown, S. James, J. Jones, and E. Ekinci, "Continuous glucose monitoring: A review of the evidence in type 1 and 2 diabetes mellitus," *Diabetic Medicine*, vol. 38, p. e14528, May 2021.

[5] P. Moström, E. Ahlén, H. Imberg, P.-O. Hansson, and M. Lind, "Adherence of self-monitoring of blood glucose in persons with type 1 diabetes in Sweden," *BMJ Open Diabetes Research & Care*, vol. 5, p. e000342, Apr. 2017.

[6] S. Lombardi, L. Bocchi, and P. Francia, "Photoplethysmography and Artificial Intelligence for Blood Glucose Level Estimation in Diabetic Patients: A Scoping Review," *IEEE Access*, vol. 12, pp. 178982–178996, 2024.

[7] S. Zanelli, M. Ammi, M. Hallab, and M. A. El Yacoubi, "Diabetes Detection and Management through Photoplethysmographic and Electrocardiographic Signals Analysis: A Systematic Review," *Sensors*, vol. 22, p. 4890, June 2022.

[8] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, pp. R1–R39, Mar. 2007.

[9] A. Hina and W. Saadeh, "Noninvasive Blood Glucose Monitoring Systems Using Near-Infrared Technology-A Review.," *Sensors (Basel, Switzerland)*, vol. 22, June 2022.

[10] F. A. C. Oliveira, F. M. Dias, M. A. F. Toledo, D. A. C. Cardenas, E. Ribeiro, R. A. Moreno, J. E. Krieger, and M. A. Gutierrez, "Non-invasive Blood Glucose Estimation: an Approach Using Multi-Wavelength PPG Signals and Residual Neural Networks," in *2025 21st International Symposium on Biomedical Image Processing and Analysis (SIPAIM)*, (Pasto, Colombia), pp. 1–5, IEEE, Nov. 2025.

[11] P.-L. Lee, T.-M. Lee, and H.-H. Chang, "Noninvasive Reflective Optical Blood Glucose Measurement System Based on Red/Green and Infrared/Green Light Amplitude Ratios," *IEEE Sensors Journal*, vol. 25, pp. 38869–38878, Oct. 2025.

[12] Q. Cuong Bui, A. Durey, J. Lee, and G.-S. Byun, "A Dual-Channel Wearable Ring With MODWT-Based Signal Processing for Non-Invasive Blood Glucose Estimation: A Pilot Study," *IEEE Access*, vol. 13, pp. 215786–215803, 2025.

[13] T. Li, Q. Wang, L. Lei, Y. An, L. Guo, L. Ren, and X. Chen, "Improvement of Non-Invasive Glucose Estimation Accuracy Through Multi-Wavelength PPG," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, pp. 5465–5478, Aug. 2025.

[14] T. Nakazawa, R. Sekine, M. Kitabayashi, Y. Hashimoto, A. Ienaka, K. Morishita, T. Fujii, M. Ito, and F. Matsushita, "Non-invasive blood glucose estimation method based on the phase delay between oxy- and deoxyhemoglobin using visible and near-infrared spectroscopy," *Journal of Biomedical Optics*, vol. 29, Mar. 2024.

[15] S. Rajala, H. Lindholm, and T. Taipalus, "Comparison of photoplethysmogram measured from wrist and finger and the effect of measurement location on pulse arrival time," *Physiological Measurement*, vol. 39, p. 075010, July 2018.

[16] M. Tsoukas, J. Rutkowski, A. El-Fathi, J.-F. Yale, S. Bernier-Twardy, A. Bossy, E. Pytka, L. Legault, and A. Haidar, "Accuracy of FreeStyle Libre in Adults with Type 1 Diabetes: The Effect of Sensor Age," *Diabetes Technology & Therapeutics*, vol. 22, pp. 203–207, Mar. 2020.

[17] S. Prahl, "Tabulated Molar Extinction Coefficient for Hemoglobin in Water," 1999.

[18] I. Oshina and J. Spigulis, "Beer–Lambert law for optical tissue diagnostics: current state of the art and the main limitations," *Journal of Biomedical Optics*, vol. 26, Oct. 2021.

[19] T. B. D. S. Costa, F. M. Dias, D. A. C. Cardenas, M. A. F. D. Toledo, D. M. D. Lima, J. E. Krieger, and M. A. Gutierrez, "Blood Pressure Estimation From Photoplethysmography by Considering Intra- and Inter-Subject Variabilities: Guidelines for a Fair Assessment," *IEEE Access*, vol. 11, pp. 57934–57950, 2023.

[20] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, "Evaluating Clinical Accuracy of Systems for Self-Monitoring of Blood Glucose," *Diabetes Care*, vol. 10, pp. 622–628, Sept. 1987.

[21] T. Bailey, B. W. Bode, M. P. Christiansen, L. J. Klaff, and S. Alva, "The Performance and Usability of a Factory-Calibrated Flash Glucose Monitoring System," *Diabetes Technology & Therapeutics*, vol. 17, pp. 787–794, Nov. 2015.

From Pix to Care Rails: A Common Architectural Core for Prescriptions, Lab Orders, Reports, and Dispensing

Bruno Henrique Aragão Konig
Departamento de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco
Recife, Brasil
bruno.konig@ufpe.br

Júlia Biscontini Koury
Departamento de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco
Recife, Brasil
julia.bkoury@ufpe.br

Élen Karolayne do Nascimento Santos
Departamento de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco
Recife, Brasil
elen.karolayne@ufpe.br

Karla Amorim Sancho
Departamento de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco
Recife, Brasil
karla.sancho@ufpe.br

Wellington Pinheiro dos Santos
Departamento de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco
Recife, Brasil
wellington.santos@ufpe.br

Fabiano Tonaco Borges
Departamento de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco
Recife, Brasil
fabiano.borges@ufpe.br

Abstract— Brazil's Unified Health System (SUS) has advanced interoperability through the National Health Data Network (RNDS) and HL7 FHIR; however, traceability remains fragmented: prescriptions, lab orders, reports, appointments, referrals, and dispensing events still lack complete audit trails, and custody often remains implicit. This study proposes a traceable sanitary object, an architectural pattern that endows each care object with seven invariant properties: a global identifier (UUID), immutability after issuance, an append-only ledger, explicit and granular custody, a finite state machine, a canonical document with a cryptographic hash, and ownership from creation. Heterogeneous clinical objects share this core, dispensing is modeled as an act within the prescription flow, and derived objects are linked through referential lineage, mirroring Brazil's Pix payment rails. A reference implementation (Python/FastAPI, PostgreSQL/SQLite, AGPL) yields a tamper-evident audit trail and neutral public endpoints under the Brazilian General Data Protection Law (LGPD). Reproducible checks verify the pattern: the state machines are exhaustively checked for structural invariants, the canonical document's digest is deterministic and changes with any of its fields, all authorization scenarios pass on PostgreSQL, and no public endpoint exposes a clinical field; an independent pass reproduced the reported results. The proof of concept implements server-side ICP-Brasil signing of issued PDFs; client-side signing and an external event-publishing layer are left for future work.

Keywords: append-only ledger, audit trail, chain of custody, electronic prescription, health information interoperability.

I. Introduction

Brazil's Unified Health System (SUS) has invested heavily in the digital exchange of health data. The National Health Data Network (RNDS) adopts the HL7 FHIR standard to consolidate records from heterogeneous institutions into a distributed national interoperability layer [1]– [3]. However, empirical assessments show that integration across national health information systems is still partial and fragmented [4], [5], and reviews of electronic health records (EHR) in Brazil repeatedly identify interoperability and security as the dominant open problems [6], [7]. These initiatives focus on the exchange of clinical data rather than the traceability of the clinical objects involved in the provision of care.

A prescription, lab order, report, appointment, referral, counter-referral, diagnostic circulation, or dispensing event is more than just data. It is an artifact. Someone issues it. Someone transfers it. Someone holds it. Someone acts on it. Someone closes it. Without end-to-end traceability and explicit custody, accountability is weakened [8]. We cannot easily determine who issued the object, who held it at each moment, what happened to it, or whether anyone altered its content. These questions are important for clinical safety and also for regulation. The Brazilian Federal Council of Medicine Resolution CFM 2.299/2021 requires ICP-Brasil signatures for digitally issued medical documents. These signatures guarantee authorship, integrity, and authenticity [9].

Brazil offers a striking precedent already in the public domain. Pix, the instant-payment system operated by the Central Bank, was designed from the outset as an open public infrastructure with standardized APIs, common rules, and equal access for banks, fintechs, governments, and new entrants. In Pix, the object (the transfer) is global and neutral, addressed by a portable key, while the regulated settlement rails are the shared substrate on which any participant operates. The model has drawn international attention as an example of payments treated as public infrastructure [10], [11]. We argue that the public health system lacks equivalent open rails for its own objects: there is no shared, neutral substrate on which a sanitary object can be issued once and then traced, audited, and acted upon across institutions.

This study proposes a traceable sanitary object as an architectural pattern that supplies such a substrate at the application-architecture level. The pattern defines an object that carries its own identity, immutability, ledger, custody

chain, state machine, canonical document, and ownership independent of any single institution.

Our contribution is the pattern itself (the core), not the product. We demonstrate it through an open-source reference implementation, PicSaúde (*Plataforma de Integração do Cuidado em Saúde* in Portuguese), available at https://github.com/Tonaco-13/PicSaude. We do not merely demonstrate this; we verify it through reproducible robustness checks: exhaustive state-machine verification over 78 states and 99 transitions, 96/96 authorization scenarios, 13/13 tamper-evident document fields, and zero clinical leakage on public endpoints. An independent process was used to reproduce the reported figures. PicSaúde is a proof-of-concept, not a system deployed at a clinical scale.

The remainder of this paper is organized as follows. Section II reviews the literature. Section III defines the conceptual model and its seven properties. Section IV details the architecture of the traceable sanitary object. Section V demonstrates the generalization across clinical objects. Section VI describes the implementation and guarantees of the model. Section VII reports the reproducible robustness checks that verify these guarantees. Section VIII discusses the trade-offs, ethics, and limitations of the study, and Section IX concludes the paper.

II. Related Literature

A. National health-data infrastructure

The dominant line of work in Brazil concerns the semantic and syntactic interoperability of clinical information. The RNDS standardizes exchanges using HL7 FHIR resources and controlled terminologies [1], [2]. Studies of the e-SUS Primary Care strategy have quantified how national information systems remain only partially integrated, with persistent fragmentation across the care and surveillance domains [4]. Broader reviews of Electronic Health Records (EHR) development in Brazil conclude that interoperability is considered essential by most authors, while also flagging access control and data security gaps [6]. Comparative analyses across Latin America emphasize governance continuity as a recurring weakness [7]. This body of work is centered on moving data between systems, not on providing each clinical artifact with an auditable lifecycle.

B. Electronic prescription and digital signature

Resolution CFM 2.299/2021 and related norms govern the digital issuance of medical documents. They mandate the signing of the ICP-Brasil to guarantee authorship, integrity, and authenticity [9]. However, they specify only what a valid signed document must contain—not how to trace the object after issuance or how to transfer custody between actors.

C. Event sourcing and append-only ledgers

Outside healthcare, event sourcing is an established pattern in which state changes are captured as immutable events in an append-only store that doubles as a complete audit trail [12]. We adopt this pattern internally as the ledger property of the proposed sanitary object.

D. Track-and-trace and chain of custody

In the pharmaceutical supply chain, blockchain-based and serialization-based approaches provide a tamper-evident chain of custody from the manufacturer to the dispensing point, often motivated by regulations such as the U.S. DSCSA and EU FMD [13]– [15]. These approaches trace physical product units across different organizations. The proposed pattern traces clinical objects (and their items) across clinical roles and does not require a distributed ledger or a shared consensus.

E. Positioning and Comparative Analysis

Prior studies separately addressed data interoperability, signed-document compliance, generic event sourcing, and supply chain track-and-trace. To bridge these gaps, the main contribution of this study proposes a unifying domain pattern for heterogeneous clinical objects. This pattern combines seven invariant properties: global identity, immutability with referential lineage, an append-only ledger, explicit and granular custody, a finite state machine, a canonical document with a cryptographic hash, and ownership from the creation. Because it is expressed at the application architecture level, it can be seamlessly instantiated in a conventional relational infrastructure. To demonstrate the architectural contribution of this study, we distinguished traceable sanitary objects from four established technical paradigms:

a) Data Interoperability Frameworks (HL7 FHIR/RNDS): Efforts such as the RNDS concentrate on semantic and syntactic standardization of the information being exchanged. However, they lack strict enforcement of a deterministic life cycle or a verifiable chain of custody (tracking the current holder and object state). The proposed pattern serves as a structural complement to FHIR; where FHIR supplies the vocabulary for clinical data, this architecture establishes the state machine and custody management layer. The same boundary applies to openEHR archetypes and IHE integration profiles, which standardize clinical content models and cross-enterprise document sharing but likewise leave lifecycle enforcement and custody to each implementation.

b) Distributed Ledger Technologies (Blockchain): While recent studies favor blockchain for establishing trust in decentralized settings, a national health system possesses an inherent central anchor in the state. Consequently, the pattern adopts an append-only relational ledger combined with cryptographic hashing to achieve tamper evidence. This approach avoids the significant computational costs and latency associated with the distributed consensus mechanisms.

c) Proprietary Electronic Prescription Platforms: Existing commercial solutions often trap custody within proprietary silos, leading to vendor lock-in, where private firms control the clinical history. By defining the sanitary object as a neutral, AGPL-licensed public rail, this architecture protects citizen sovereignty and prohibits the monetization of health records, operating as an open infrastructure, rather than a closed ecosystem.

d) Healthcare Traceability Frameworks (DSCSA/EU FMD): Regulatory track-and-trace regimes such as the U.S.

DSCSA and the EU Falsified Medicines Directive mandate the serialization and verification of physical drug packages across the supply chain [13]-[15]. These frameworks locate a unit of product; they do not record who holds a clinical decision or in what state it stands. The proposed pattern operates at this clinical-object layer and is complementary to them: a dispensing event recorded under this architecture can anchor the care-level endpoint of such product-level chain.

III. The Traceable Sanitary Object

We define a traceable sanitary object as any artifact of care that has its own clinical identity and an auditable lifecycle. This pattern is characterized by seven invariant properties (Fig. 1).

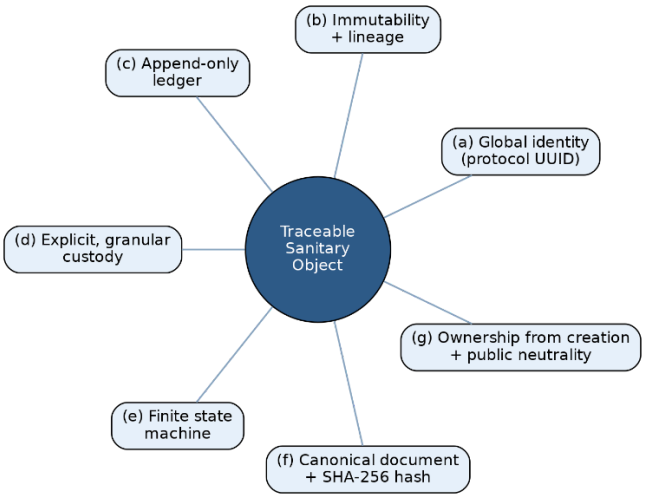


Fig. 1. The traceable sanitary object and its seven invariant properties (a)–(g): global identity, immutability with lineage, an append-only ledger, explicit and granular custody, a finite state machine, a canonical document with hash, and ownership from creation.

a) **Global identity:** Every object carries a protocol (a UUID) that is unique, immutable, and independent of any institution: the primary handle through which the object is referenced, audited, and shared, in direct analogy to a portable Pix key.

b) **Immutability with lineage:** An object is never edited after issuance. Any correction, renewal, or re-issuance produces a new derived object that points back to its predecessor through an origin reference, forming an explicit lineage chain. Therefore, audits never lose their history.

c) **Append-only ledger:** Every business-relevant event is recorded as an insertion into an events table that never receives updates or deletions. This internal event-sourcing log is the backbone of the auditability.

d) **Explicit granular custody:** Each object has a single custodian at every moment, and every transfer is recorded in a custody chain. Custody can be held over an entire object or over an individual item, which is what makes partial flows (e.g., partial dispensing) traceable. *Uniqueness of custody is enforced at the database level by partial unique indexes in both supported dialects, rather than by application convention.*

e) **Finite state machine:** Each object type has an explicit set of states and valid transitions, with terminal states distinguished by clinical meaning.

A digital flow is kept separate from a purely physical flow (paper issuance that never enters the digital cycle), avoiding category errors between, for instance, a clinical revocation and a paper-only closure.

f) **Canonical document with hash:** The object is serialized deterministically into a canonical document from which a SHA-256 digest is computed—the integrity anchor and the basis for ICP-Brasil signing under the PAdES profile; in the current proof-of-concept, server-side ICP-Brasil PAdES-B signing of issued PDF documents is fully implemented, while client-side signing of the canonical document remains future work (Section VIII).

g) **Ownership from creation and public neutrality:** Access control and inline identity assertion are enforced on every private endpoint from the moment the object is created. Public endpoints are kept neutral (carrying no clinical data) so that verification can be offered without exposing sensitive information, consistent with the LGPD Art. 11 [16].

IV. Architecture

The architecture materializes the seven properties on a conventional relational stack. Fig. 2 shows the prescription life-cycle state machine, and Fig. 3 shows the custody chain and derivation lineage.

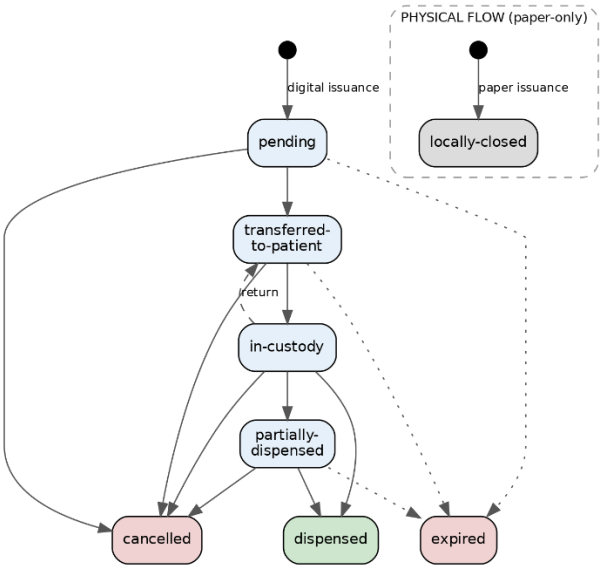


Fig. 2. Prescription life-cycle state machine. A digital flow (pending through dispensed, with cancellation and expiry) is kept disjoint from a physical, paper-only flow whose single terminal state is locally-closed, preventing a clinical cancellation from being conflated with a paper-only closure.

A. Immutable ledger

Each object family has a table of events which admits only insertions. A controlled vocabulary of events—issuance, renewal, correction, custody transfer, dispensing, partial dispensing, and item return—is recorded with the actor, timestamp, and payload. Because no update

or deletion is permitted, the ledger provides a non-destructive, append-only history that underpins all audit queries.

Because the ledger is the source of truth, a regulatory report is not a separate stateful artifact but a deterministic projection of that ledger, and two properties of this projection are already enforced and verified. Reproducibility: recomputing a closed period yields a byte-identical result, because each line reflects the running balance as of its own movement date, so events recorded after the period boundary never alter earlier lines. Identifier uniqueness: each movement—a dispensing or a reversal—appears exactly once, a property we treat as a falsifiable fraud invariant and enforce with a permanent guard-rail check in the continuous-integration gate; accordingly, every sanitary-object mutation is idempotent and protected against double submission by a row-level lock. Two further rules are adopted as invariants with implementation underway: every derived object resolves to a visible root protocol through the origin-reference chain, and external surveillance identifiers (for example, a medication's ANVISA registry number) are snapshotted onto the movement at write time rather than fetched live on the clinical path.

B. Granular custody

A custody table records the current holder and all transitions. Valid transitions are declared explicitly (e.g., prescriber to patient on digital issuance, patient to dispenser on presentation at the counter, and dispenser to patient on return). Custody may be recorded for an entire object or for individual items.

C. Finite state machine

The allowed states and transitions are declared in a single domain module that serves as the source of truth, which is mirrored in the database schema. Digital states (pending, transferred-to-patient, in-custody, partially dispensed, dispensed, and cancelled) are separated from the physical terminal state used for paper-only issuance.

D. Canonical document and integrity

A dedicated module produces canonical serialization and its SHA-256 digest. The digest is stored with the object and exposed through a document-integrity endpoint, allowing any holder to verify that the content has not changed and is the intended input for ICP-Brasil PAdES signing.

E. Ownership, RBAC and public neutrality

Every private endpoint enforces role-based access control and an inline identity assertion so that ownership is checked at creation and at each subsequent act. Public endpoints expose only non-clinical confirmation data [16].

F. Institutional scope

Operations are made contextual through optional organization and unit identifiers. The clinical object remains global and neutral—identified by its protocol—while these identifiers act as query scope, access filter, and operational context, never as the primary key. A NULL value denotes the absence of an institutional link (not a sentinel), and the fields are rolled out incrementally only when a real use case requires them. Where an organization identifier is present, queries in an institutional context must include it as a filter, enforced as an architectural guardrail.

V. Generalization of the Core

The pattern is a core, not an exception: the same seven properties instantiate multiple clinical objects (Table I).

Dispensing is not a standalone object but an act within the prescription flow, governed by the invariant "sum of dispensed quantities does not exceed the prescribed quantity." Its hospital variant follows prescriber → hospital pharmacy → (nursing unit) → patient.

TABLE I. *Instances of the Traceable-Sanitary-Object Core*

Object	***Typical states (digital flow)***	***Terminal states***	***Custody chain***
Prescription	pending → transferred-to-patient → in-custody → partially-dispensed → dispensed	dispensed; cancelled; expired; locally-closed (physical)	prescriber → patient → dispenser → (patient | prescriber)
Lab order	issued → scheduled → collected → under-analysis → result-available → closed	closed; cancelled; expired; physically-closed	prescriber → patient → lab provider → patient
Report	in-production → signed → released → acknowledged (patient/prescriber) → closed	closed; cancelled; expired; physically-closed	lab provider → (patient | prescriber)
Appointment	created → confirmed → carried-out	carried-out; cancelled; no-show	none (bilateral commitment; documented exception)
Diagnostic circulation	selected → sent-to-lab → proposal-received → patient-confirmed → performed	performed; patient-unmarked; lab-unmarked; lab-archived; expired	patient → lab (token-mediated)

Referral	issued → (under-regulation) → scheduled → attended → counter-referred → closed	closed; cancelled; expired; denied; physically-closed	prescriber (origin) → patient → prescriber (destination) → patient
Counter-referral (derived)	Registered	Cancelled	destination → origin (origin-referral reference)

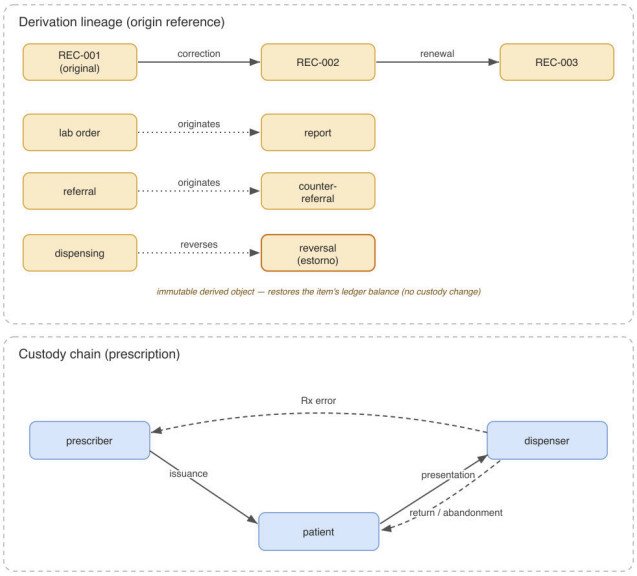


Fig. 3. Custody chain (a single custodian at each moment, with explicitly declared transitions) and derivation lineage (corrections and renewals chained by an origin reference, and cross-object derivations such as report-from-lab-order and counter-referral-from-referral).

A. *Derived objects*

The derivation generalizes beyond corrections and renewals. A report is a derived object of a lab order, and a counter-referral is a derived object of a referral. Each derivation has a distinct author, its own canonical document and hash, and a referential link to its origin. Thus, the lineage chain spans not only versions of one artifact but also causally related artifacts across the care pathway.

B. *Documented exceptions*

Some objects intentionally relax one property, and recording the exception is part of the discipline. Appointments are lightweight objects with identity, states, and a ledger but no custody chain because they represent a bilateral commitment rather than a held artifact. Acknowledgement of a report operates at the level of the entire report rather than per item. Diagnostic circulation authorizes the act through a presentation token, rather than a custody transfer. Naming these exceptions explicitly is what keeps the core coherent and stable.

VI. Implementation and Guarantees

This section describes the reference implementation of the traceable sanitary object and the guarantees the pattern entails.

A. *Stack and scale*

The reference implementation uses Python with FastAPI for the API layer and a relational store that runs on PostgreSQL in production and SQLite in development, under a single Alembic-managed schema. It comprises approximately 36,142 lines of application code across 33 routers and 47 ORM models, exposing 156 endpoints over 20 schema migrations. Seven sanitary-object families each carry their own finite state machine; across object- and item-level machines, they declare 78 states and 99 valid transitions in single source-of-truth modules, and a unified ledger helper records the controlled event vocabulary for all subdomains. Seven of these families maintain their own append-only event ledgers (alongside an internal publication outbox), and five maintain explicit custody chains. The test suite comprises 1,557 test functions across 103 files; a continuous-integration gate, the unit suite together with a selected integration subset executed against an ephemeral PostgreSQL instance, runs on every change and gates merges. All implementation figures above are as of commit babc264, Aug. 24, 2026.

B. *Dual-database test gate*

A dual-database gate executes the suite against ephemeral PostgreSQL instances in addition to SQLite. This is not redundant: SQLite's permissive typing (e.g., accepting an integer literal for a Boolean column or lax NULL handling) silently masks defects that surface only under PostgreSQL's strict typing. In our development, this gate repeatedly caught latent defects invisible on SQLite; for instance, a Boolean column written with an integer literal passed every SQLite test yet raised a type-mismatch error on PostgreSQL, meaning that a code path that appeared fully tested was in fact broken under production-equivalent conditions. The methodological claim is narrow but practical: for dual-database clinical systems, the production database must be inside the test gate—a green bar on the development database is necessary but insufficient.

C. *Guarantees from the model*

Several safety properties follow directly from the pattern rather than from ad-hoc checks. Partial dispensing cannot exceed the prescription because the sum of dispensed quantities for an item is constrained to not exceed the prescribed quantity. An item that cannot be paid for simply returns to a pending state and remains dispensable elsewhere without cancelling the prescription. Immutability, together with the append-only ledger, yields a tamper-evident audit trail: history is never overwritten, only extended; non-repudiation of issued documents is already anchored in server-side ICP-Brasil PAdES-B signing recorded in the ledger, while end-to-end non-repudiation of the canonical document follows once client-side signing is integrated. Terminal-state semantics are kept distinct; therefore, a clinical cancellation is never conflated with a paper-only closure. Access control is enforced from creation, and public endpoints are neutral by construction; functional checks found no clinical field exposed on any public endpoint (Section VII); these are functional checks, not a security audit, and adversarial testing remains future work. Table II maps each guarantee to the mechanism that enforces it and a representative test from the suite that exercises it.

TABLE II. MODEL GUARANTEES, ENFORCING MECHANISM, AND A REPRESENTATIVE TEST

Guarantee	*Enforcing mechanism*	*Representative test*
`Dispensed ≤ prescribed`	per-item balance check before insert	test_supera_saldo (HTTP 422)
Append-only audit trail	insert-only event ledger (no update/delete)	test_retorna_eventos_inseridos
Immutability with lineage	correction issues a new derived object	test_correcao_aceita_novos_campos
Valid transitions only	single source-of-truth transition tables	test_transicao_invalida (HTTP 409)
Ownership from creation	inline RBAC + identity assertion	test_dispensador_sem_custodia (HTTP 403)
Public neutrality	public endpoint projects no clinical field	test_publico_neutro

D. Regulatory and licensing posture

The design targets compliance with ANVISA RDC 1.000/2025, which governs the electronic prescription of medications under special control, emphasizing traceability, mandatory digital signature, and integration with the National Controlled-Prescription System (SNCR), which amends Ordinance SVS/MS 344/98 and CFM 2.299/2021 [9], [17]. The reference implementation was released as open source under the AGPL, on the principle that the infrastructure for a public health system should be public. The architecture is the subject of a Brazilian INPI filing (BR 51 2026 002267-3).

VII. ROBUSTNESS EVALUATION

This section discusses the claims made by the proposed pattern. The evaluation was architectural rather than clinical; a real-setting study is required in the future. Therefore, we subject each invariant property to a structural or behavioral robustness check and ask whether it holds in the implementation, not only in the design. Every check is scripted, seeded, and released with the open-source code, so each number below is reproducible; the headline figures were additionally re-derived by an independent verification pass that reproduced 14 of 15 and caught one author measurement error, since corrected to the 13/13 reported here. Table III summarizes the results.

A. Formal verification of the state machines

The seven families comprised 11 object- and item-level machines (78 states and 99 transitions). We exhaustively checked four structural properties: every transition target is a declared state, the terminal set is a subset of the states, no non-terminal state is a dead end, and terminal states are absorbing. The first three are universally applicable. The absorbing-terminal check surfaced two terminal states with a declared outgoing transition—a dispensed item that may be reversed, and an item with an available result that may be explicitly closed—revealing a tension between "terminal as clinically complete" and "terminal as absorbing." Instead of concealing it, we report it as a finding that is currently being refined. For the dispensing case in particular, the reversal is modeled as an immutable derived object that restores the effective ledger balance of the item rather than mutating the terminal item. Therefore, dispensability follows from the ledger balance rather than from the state label, while the declared dispensed-to-reversed transition remains dormant scaffolding that the check correctly surfaces. Surfacing such cases is precisely the benefit of treating the state machine as a verifiable artifact.

A guard-rail check confirmed that no movement identifier recurred in the dispensing report, and a concurrency test shows that two simultaneous identical dispensings each recorded a single movement, exercising the idempotency invariant under a race.

B. Document integrity (tamper-evidence)

For the canonical document, the SHA-256 digest was identical across 1,000 recomputations of the same input (determinism), and a single-field mutation altered the digest for all 13 document fields (eight headers and five per item); that is, any content change is tamper evident.

C. Authorization and anti-leak

Across nine authorization suites, 96 scenarios were executed against PostgreSQL, and all passed, each exercising the ordering 404 (not found), 403 (not owner), and 409 (state conflict) that prevents disclosure of an object's existence or state to a non-owner. These are functional checks of the intended access-control behavior, not a security audit; adversarial and penetration testing remain future work.

D. Dual-database divergence

The dual-database gate-caught defects are invisible on SQLite: in two concrete instances, a Boolean column written with an integer literal passed every SQLite test yet raised a type-mismatch error on PostgreSQL—a defect

class that example-based testing on the development database alone does not reveal.

E. Public-endpoint neutrality (LGPD)

A static audit of all six public endpoints confirmed that none projected a clinical field (only protocol, status, issuance type, and per-item order/status), and tests confirmed that no seeded clinical values appeared in any public response.

TABLE III. ROBUSTNESS EVALUATION RESULTS

Check	***Result***
State-machine verification	11 machines, 78 states, 99 transitions; 3/4 properties universal; 2 non-absorbing terminals found and tracked
Hash tamper-evidence	deterministic over 1,000 runs; 13/13 fields alter the digest
Authorization / anti-leak	96/96 scenarios pass on PostgreSQL (404→403→409)96/96 scenarios pass on PostgreSQL
Dual-database gate	2 boolean-as-integer defects caught on PostgreSQL, invisible on SQLite
Public neutrality (LGPD)	0/6 public endpoints project a clinical field

VIII. DISCUSSION, LIMITATIONS, AND ETHICS

A. Trade-offs

Physical (paper-only) issuance uses a fire-and-forget call: the front-end prints without waiting for the back end to confirm persistence. This favors the availability of the clinical act over guaranteed central registration—if the back end is offline, printing still succeeds, and the record may remain only on the client. The trade-off is accepted and documented; full traceability requires the back end to be reachable. Similarly, a mathematically invalid sentinel identifier is reserved for unidentified patients in paper flows and must be excluded from analytics, and the institutional scope is rolled out incrementally rather than imposed on every table at once.

B. Ethics

A non-negotiable goal of the project is that patient health data are never monetized. The aim is citizen sovereignty over health objects through auditable grants of access, never the sale of data. This principle is treated as a design constraint, rather than a policy statement.

C. Limitations and Challenges for Large-Scale Adoption

This work is currently a proof-of-concept and has not yet been deployed at clinical scale. Moving from an architectural core to a national or municipal deployment introduces several practical challenges that must be addressed:

a) External Integration and Interoperability: While the system is event-sourced internally, it is not yet event-driven externally. Because there is currently no event-publishing layer, external adapters (such as legacy HIS, TISS, HL7, and e-SUS) lack a native integration point. Real-world adoption will require developing these robust interfaces to orchestrate workflows without demanding the complete replacement of existing hospital and pharmacy systems. The planned mitigation is an event-publishing layer over the append-only ledger—cursor-based polling endpoints scoped by organization, per-organization webhook registration, and a canonical external event format—so that adapters consume official endpoints and never write directly to clinical tables, enabling incremental coexistence with legacy hospital information systems rather than wholesale replacement.

b) Cryptographic Implementation and Ledger Security: Server-side ICP-Brasil PAdES-B signing is fully implemented for issued PDF documents (e.g., prescription forms and lab orders). The PDF is signed with a certificate held in an encrypted server-side vault, and each signing act is recorded in the ledger. However, client-side signing of the object's canonical document remains future work, as it requires a deterministic serialization that is identical across client and server implementations. Furthermore, the append-only ledger is currently enforced as an application-level invariant; implementing cryptographic hash-chaining or external anchoring to make the ledger tamper-evident—even against a privileged database actor—is a necessary next step.

c) Performance under Real Clinical Workflows: Operating at the scale of the SUS means tracing millions of daily objects and their derivations. This poses significant database growth and query-performance challenges, which will require advanced data engineering (e.g., database sharding, asynchronous queues, and cold-storage archiving for physically closed objects) to maintain low latency during peak clinical hours. The relational design admits horizontal partitioning of the event and custody tables by time or organization, read replicas serving deterministic projections, and archival of terminally closed objects to cold storage; the absence of distributed consensus removes a major latency source by construction.

d) Governance and Orchestration: The Pix analogy raises a fundamental governance question beyond this paper's architectural scope. Just as Pix depends on the Central Bank as a central operator, shared health rails would require a designated public operator (such as the Ministry of Health or DATASUS) and a robust governance model to manage credentialing, access control, and cross-jurisdictional coordination. A federated operating model—central specification and auditing combined with regionally operated instances—could balance national consistency with local autonomy, but operator accreditation and dispute resolution across jurisdictions remain open institutional questions.

IX. CONCLUSION AND FUTURE WORK

We have argued that the objects of care—prescription, lab order, report, appointment, referral, counter-referral, diagnostic circulation, and dispensing—share a single architectural contract, which was formalized as the traceable sanitary object: a pattern defined by global identity, immutability with lineage, an append-only ledger, explicit and granular custody, a finite state machine, a canonical document with hash, and ownership from creation. Treating these artifacts as instances of one core is what yields traceability, auditability, and the basis for cross-institution interoperability. Crucially, this contract is not merely proposed but verifiable and verified: its state machines, canonical hashing, authorization ordering, and

public-endpoint neutrality are all exercised by reproducible, seeded scripts shipped with the AGPL-licensed code, and the reported figures were reproduced by an independent pass—so the architecture's claims can be re-audited rather than taken on trust. The same discipline keeps us honest, having surfaced two real state-machine inconsistencies that we report rather than hide. Future work includes client-side ICP-Brasil signing of the canonical document, an event-publishing layer with webhooks to enable external adapters, a municipal pilot, and empirical validation in a real care setting.

References

[1] Ministério da Saúde, "Rede Nacional de Dados em Saúde (RNDS): Guia de Integração," 2020. Accessed: Jun. 17, 2026. [Online]. Available: https://rnds-guia.saude.gov.br/

[2] Ministério da Saúde, "RNDS — HL7 FHIR Implementation Guide," 2022. Accessed: Jun. 17, 2026. [Online]. Available: https://rnds-fhir.saude.gov.br/

[3] Brasil, Ministério da Saúde, Secretaria-Executiva, Departamento de Informática do SUS, "Estratégia de Saúde Digital para o Brasil 2020–2028," Brasília, DF, Brasil, 2020. [Online]. Available: https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf

[4] G. C. C. Neto, R. Andreazza, and A. Chioro, "Integration among national health information systems in Brazil: the case of e-SUS Primary Care," Rev. Saúde Pública, vol. 55, p. 93, 2021, doi: 10.11606/S1518-8787.2021055002931.

[5] M. Fernandez et al., "Interoperability in universal healthcare systems: insights from Brazil's experience integrating primary and hospital health care data," Front. Digit. Health, vol. 7, 2025, doi: 10.3389/fdgth.2025.1622302.

[6] I. M. P. Barbalho et al., "Electronic health records in Brazil: prospects and technological challenges," Front. Public Health, vol. 10, Nov. 2022, doi: 10.3389/fpubh.2022.963841.

[7] L. F. Conrado et al., "Digital health: context and challenges in Brazil with focus on public health," in Proc. IEEE Int. Conf. Ind. Appl. (INDUSCON), 2025, pp. 1153–1160, doi: 10.1109/INDUSCON66435.2025.11241578.

[8] M. Ahmed, A. R. Dar, M. Helfert, A. Khan, and J. Kim, "Data provenance in healthcare: approaches, challenges, and future directions," Sensors, vol. 23, no. 14, 2023, doi: 10.3390/s23146495.

[9] Conselho Federal de Medicina, "Resolução CFM nº 2.299/2021," Brasília, DF, Brasil, 2021. Accessed: Jun. 17, 2026. [Online]. Available: http://www.portalmedico.org.br

[10] A. Duarte, J. Frost, L. Gambacorta, P. Koo Wilkens, and H. S. Shin, "Central banks, the monetary system and public payment infrastructures: lessons from Brazil's Pix," SSRN Electron. J., Mar. 2022, doi: 10.2139/ssrn.4064528.

[11] Bank for International Settlements, "III. The future monetary system," in BIS Annual Economic Report 2022, Basel, Switzerland, Jun. 2022. [Online]. Available: https://www.bis.org/publ/arpdf/ar2022e3.htm

[12] M. Fowler, "Event Sourcing," martinfowler.com. Accessed: Jun. 17, 2026. [Online]. Available: https://martinfowler.com/eaaDev/EventSourcing.html

[13] F. Díaz, N. Cerna, R. Liza, B. Motta, and S. Rojas-Flores, "Blockchain-enabled traceability in pharmaceutical supply chains: a mapping review of evidence for visibility, anti-counterfeiting, and chain-of-custody control," Logistics, vol. 10, no. 4, Apr. 2026, doi: 10.3390/logistics10040085.

[14] U.S. Food and Drug Administration, "Drug Supply Chain Security Act (DSCSA)," Accessed: Jun. 17, 2026. [Online]. Available: https://www.fda.gov/drugs/drug-supply-chain-integrity/drug-supply-chain-security-act-dscsa

[15] "Directive 2011/62/EU of the European Parliament and of the Council of 8 June 2011 amending Directive 2001/83/EC on the Community code relating to medicinal products for human use, as regards the prevention of the entry into the legal supply chain of falsified medicinal products (Falsified Medicines Directive)," Off. J. Eur. Union, L 174, Jul. 2011.

[16] Brasil, "Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais — LGPD), Art. 11," 2018. Accessed: Jun. 17, 2026. [Online]. Available: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[17] Agência Nacional de Vigilância Sanitária (ANVISA), "Resolução da Diretoria Colegiada — RDC nº 1.000, de 11 de dezembro de 2025," Brasília, DF, Brasil, 2025.

Gemma 4 as a Proof-of-Concept Model for Converting Discharge Medication Text into FHIR MedicationRequest Resources

1st Fernando M. Secol
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
fmsecol@gmail.com

2nd Maria Claudia F. Castro
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo - SP, Brazil
mclaudia@fei.edu.br

***Abstract*—Clinical information in discharge summaries is often recorded as free text, limiting its reuse in interoperable healthcare systems. This paper presents a preliminary feasibility study of transforming hospital discharge medication text into HL7 FHIR *MedicationRequest* resources using large open-weight language models. Based on MIMIC-IV-Note, MIMIC-IV, and MIMIC-IV on FHIR, we constructed a paired dataset linking discharge medication sections to structured medication references. The task was formulated as structured generation and evaluated using Gemma 4 E2B-it and E4B-it under zero-shot, few-shot, and supervised LoRA fine-tuning settings. Evaluation considered both clinical extraction quality, including medication, dosage instruction, route, and frequency, and structural quality, including JSON validity, FHIR conformance, minimum completeness, and Bundle completeness. Preliminary results indicate that prompt-based strategies were insufficient for this task, whereas LoRA fine-tuning substantially improved extraction performance, particularly for the larger E4B-it variant. However, structural validity and Bundle completeness remained lower than extraction metrics, showing that generating interoperable FHIR resources requires more than accurate entity extraction. These findings suggest that open-weight LLMs are a feasible starting point for clinical text-to-FHIR transformation, while highlighting the need for validation, correction mechanisms, and hybrid pipelines.**

***Index Terms*—Large language models, clinical information extraction, HL7 FHIR, MedicationRequest, discharge medications, healthcare interoperability.**

I. Introduction

The adoption of electronic health records, together with the digital transformation of care processes, has expanded the availability of clinical data for care delivery, management, research, and technological development. However, a substantial share of clinical information remains recorded as free text, with variations in style, abbreviations, omissions, idiosyncratic expressions, and different levels of completeness. In many health information systems, documentation is primarily produced to support the care workflow, record clinical decisions, and meet legal requirements, rather than to facilitate interoperability, automated analysis, or systematic reuse in research and innovation *pipelines* [1]–[4].

In the medication domain, information on drugs, dose, unit, route of administration, frequency, duration, dispensed quantity, and refills is directly related to patient safety and continuity of care. In hospital discharge notes, these elements may appear abbreviated, distributed along the same line, expressed partially or implicitly, or combined in heterogeneous textual patterns. The challenge increases when there are multiple medications in sequence, prescriptions with limited duration, tapering regimens, "*as needed*" instructions, and dispensing or refill fields.

Thus, the problem involves a linguistic stage, aimed at understanding clinical text, and a structural stage, aimed at correctly representing the extracted content in a computable format, at a critical moment of care transition, in which discrepancies, omissions, or incorrect interpretations of prescriptions may contribute to adverse events that could otherwise be avoided. Therefore, transforming medication instructions recorded in natural language into structured, traceable, and interoperable representations is a practical requirement to support medication reconciliation, communication between services, clinical auditing, decision support, and secondary data reuse [2], [5].

The literature has explored various strategies for this type of problem, including manual rules, dictionary-based systems, supervised sequence labeling models, neural architectures, and, more recently, Large Language Models (LLMs). Shared challenges, such as n2c2 2018, have contributed to consolidating medication and attribute extraction tasks in clinical texts, providing methodological foundations for comparing approaches [13]. These studies show that the problem is not purely lexical; it depends on context, terminological normalization, and the relational interpretation between entities.

At the interoperability level, the HL7 FHIR (*Fast Healthcare Interoperability Resources*) standard provides a modular, resource-oriented model for clinical data exchange. FHIR is strongly associated with RESTful API-based architectures, in which transactions are performed directly on resources through HTTP requests [11]. In the medication domain, the *MedicationRequest* resource represents a medication order or prescription, including both the supply and administration instructions for the patient [10].

In this scenario, LLMs emerge as an alternative for clinical information extraction and structuring tasks, since they combine contextual understanding, instruction-following, and structured-output generation capabilities. In this paradigm, the model receives the clinical text and a specification of the expected output format, and may directly return an intermediate structure, such as JSON, or a standardized resource, such as FHIR. Its main advantage lies in the ability to specify the task in natural language, combining instructions, examples, constraints, and output schemas.

Recent studies indicate that LLMs may be useful for clinical extraction tasks, but also show that results vary substantially according to the model, domain, language, prompt type, output structure, and the presence or absence of fine-tuning [16]–[22]. Overall, these studies reinforce that LLMs can reduce manual engineering effort, but do not eliminate the need for careful evaluation, post-processing, structural validation, and error analysis. Their use must be evaluated with caution, since clinical applications require not only aggregate performance, but also analysis of compliance, traceability, safety, hallucination, omissions, and semantic consistency [6]–[8].

In the specific context of FHIR, transforming free text into interoperable resources is even more demanding. In addition to correctly extracting the clinical content, the model must respect the resource structure, fill in required fields, avoid incompatible elements, maintain attribute consistency, and produce valid JSON. Recent work on converting clinical reports into FHIR has shown that the prompt strategy may be as important as, or even more important than, the choice of model, especially when the task involves generating a FHIR *Bundle* and validation using tools such as HAPI FHIR [23]. *Benchmarks* focused on assessing FHIR knowledge and generation also suggest that language models can adequately answer questions about the standard or recognize resource types, but still face substantial difficulties in the more complex task of reconstructing FHIR resources from unstructured clinical notes [24].

In this context, this article presents a preliminary feasibility study using Google/Gemma 4 as a proof-of-concept model to transform hospital discharge medication prescription texts into FHIR resources of type *MedicationRequest*. The study aims to evaluate whether the model is able to: (i) correctly identify the medications present in the text; (ii) recover prescription attributes, such as drug, dose, route, and frequency; (iii) organize these elements into FHIR *MedicationRequest* resources; and (iv) produce outputs that are syntactically valid and semantically consistent with the original content. The main contributions of this work are: the proposal of a pragmatic formulation for the problem of transforming hospital discharge medication text into FHIR resources; the presentation of a design for evaluating LLMs in a task that combines clinical extraction and structural compliance; and the evaluation of Gemma 4 as an open model, considering its applicability in scenarios in which local execution, infrastructure control, and reduced dependence on proprietary APIs are business requirements.

II. Materials and Methods

The task was formulated as a structured generation problem, in which the input corresponds to the textual discharge medication section extracted from clinical reports and the output corresponds to a set of FHIR resources of type *MedicationRequest*. The experimental design includes three main evaluation strategies: *zero-shot* inference, *few-shot* inference, and supervised model fine-tuning. The evaluation considers both the quality of clinical extraction and the structural compliance of the generated output.

A. Data Sources

The data source used in this study is the MIMIC-IV database from the PhysioNet platform. MIMIC-IV is a public and de-identified electronic health record database, built from patient clinical data and organized in a modular manner to support retrospective health analyses [1].

The MIMIC-IV-Note module was used as the source of natural-language clinical texts [25]. For this study, hospital discharge reports containing the prescribed medication section were considered. As a structured reference for comparing and validating the generated medication resources, MIMIC-IV on FHIR was used, which provides a representation of MIMIC-IV data in the HL7 FHIR standard [26].

The integration between the databases was performed through shared de-identified identifiers, especially *subject_id* and *hadm_id*. The *subject_id* enables de-identified patient identification, whereas the *hadm_id* represents the hospital admission. In this study, *hadm_id* was adopted as the primary unit of analysis to relate the discharge report, the structured prescription records, and the corresponding FHIR resources within the same care context.

B. Cohort Construction

The experimental cohort was built by matching hospital discharge reports from MIMIC-IV-Note with structured medication records available in MIMIC-IV and MIMIC-IV on FHIR. Only discharge reports associated with a valid *hadm_id* and explicitly containing a discharge medication section identifiable through textual patterns were included. The purpose of this selection was to ensure that each example used in the study presented clear textual evidence of the prescription and a structured reference associated with the same encounter.

Reports containing the *Discharge Medications* section, with the list of medications prescribed or recommended for therapeutic continuation after hospitalization, were considered eligible. Reports without valid textual content, documents in which the medication section could not be reliably located, records with predominant textual noise, and cases without adequate linkage between the discharge text and the corresponding structured medication records were excluded.

The training set for the fine-tuned models was constructed as an input-output pair. The input consists of the text from the *Discharge Medications* section combined with a fixed instruction *template*. The output corresponds to the expected structured representation, expressed in JSON and compliant

with the minimum schema for FHIR *MedicationRequest* resources. Thus, each supervised example represents a direct transformation between free clinical text and an interoperable structure.

C. Extraction of the Discharge Medication Section

The extraction of the discharge medication section was performed through a deterministic pipeline based on regular expressions. This pipeline was implemented in Python and executed on a private server running Ubuntu, with an Intel i7-14700K processor, 64 GB of RAM, and an NVIDIA RTX 4080 GPU.

The pipeline was developed following an exploratory manual inspection of clinical documents. At this stage, 250 discharge reports were randomly selected from the full set for analysis of formatting patterns in the medication section. Based on this analysis, heuristics were defined to locate the section's beginning, delimit its end, normalize the text, and segment the listed medications.

Identification of the target section was based on the literal header *Discharge Medications*. The end of the section was preferably identified by the subsequent header *Discharge Disposition*. When this end marker was not found, the pipeline applied a contingency strategy, capturing the content until the end of the document. This approach was adopted to reduce losses caused by incomplete formatting, missing headers, or partial text degradation.

After locating the section, the content underwent text normalization steps. Excess spaces, redundant line breaks, and superficial formatting variations were removed. Next, medication items were segmented based on numbered lists, a pattern frequently observed in discharge reports. Each numbered item was treated as a candidate prescription unit, preserving the original textual line in normalized format.

The pipeline output was stored in JSON format, organized by *hadm_id*. For each hospital admission, the resulting file contains the list of medication items extracted from the discharge section, the marker of the source section, and, when available, the structured medication record associated with the same encounter. This structure enables each textual line to be traced back to its clinical context and to the structured reference used in the evaluation.

D. MedicationRequest Generation Task

The central task of the study was to transform a textual discharge medication section into a set of HL7 FHIR resources of type *MedicationRequest*. The *MedicationRequest* resource was selected as the target resource because it represents a medication order or request for a patient.

Each model execution receives as input a prompt containing explicit instructions about the task and the text of the *Discharge Medications* section. The model must identify all medications described in the excerpt and generate structured JSON output in the form of a FHIR *Bundle* containing one *MedicationRequest* instance for each identified medication.

The choice of using a FHIR *Bundle* is associated with the need to represent multiple resources in a single response. Since the discharge section generally includes a list of medications, each item must be represented as an independent resource, enabling individual evaluation of each medication and its attributes.

To reduce output variability and make model comparisons fairer, a minimum field schema was defined for each *MedicationRequest*. The core fields include the medication name, route of administration, frequency, and dose. Thus, the task does not only evaluate whether the model correctly recognizes the medication, but also whether it can associate dosage attributes with the correct medication and serialize this information into a FHIR-compatible structure.

E. Gemma 4 LLM Model

Gemma 4 is a family of generative AI models developed by Google DeepMind and released with open weights in different sizes and architectures, including E2B and E4B variants [9].

For the problem investigated, some characteristics of Gemma 4 are relevant:

- the existence of smaller and medium-sized variants, favoring experiments in hardware-constrained environments and bringing the study closer to local or institutional execution scenarios;
- support for *instruction-tuned* models, facilitating task specification through *prompts* in preliminary structured extraction experiments;
- the possibility of fine-tuning or subsequent adaptation, making the model an interesting foundation for further development of the study, should preliminary results indicate limitations in purely *prompt*-based scenarios.

The experiments were conducted with the instruction-tuned variants Gemma 4 E2B-it and Gemma 4 E4B-it, obtained from Google's official HuggingFace repositories. The *zero-shot*, *few-shot*, and post-fine-tuning inference trials were designed to be executed with the same local configuration as the extraction pipeline.

- *zero-shot* – the model receives only the task instruction, the expected output schema, and the text of the discharge medication section.
- *few-shot* – the prompt includes 5 short input-output examples, demonstrating how medications and their attributes should be mapped to *MedicationRequest* resources, in order to induce greater adherence to the FHIR format, reduce serialization variations, and improve the handling of frequent prescription abbreviations, such as *PO*, *BID*, *TID*, *Q6H*, and *PRN*.
- *supervised fine-tuning* – performed on Google Cloud Platform / Google Colab infrastructure, using an NVIDIA H100 GPU, with the PEFT/LoRA strategy.

The hyperparameters used were defined separately for each variant. For Gemma 4 E2B-it, the following were used: (r=32), ($\alpha = 64$), *dropout* $= 0.05$, learning rate (1×10^{-4}), 3 epochs, effective batch size of 32, and maximum sequence length of

4096 tokens. For Gemma 4 E4B-it, the following were used: (r=16), ($\alpha = 32$), *dropout* $= 0.05$, learning rate (5×10^{-5}), 2 epochs, effective batch size of 32, and maximum sequence length of 4096 tokens.

F. Evaluation Metrics

The evaluation was conducted according to two complementary dimensions. The first concerns the quality of the clinical information extracted from the discharge medication text, measuring whether the model correctly identified medications and their associated attributes. The second concerns the structural quality of the interoperable output, measuring whether the generated resources were valid, consistent, minimally complete, and conformant to the HL7 FHIR specification.

For clinical content extraction, precision, recall, and *F1-score* were calculated independently for medication name, dosage instruction, route of administration, and frequency. The comparison was performed between the items generated by the model and the corresponding reference items in the dataset. For these calculations, the following definitions were adopted:

- True positives (TP) correspond to medications or medication attributes correctly extracted by the model according to the reference data. At the attribute level, a value was considered a true positive only when it was both correctly extracted and associated with the corresponding medication.
- False positives (FP) correspond to medications or attributes generated by the model that were not present in the reference data or were not supported by the source text. An attribute extracted with an incorrect value or associated with the wrong medication was also counted as a false positive.
- False negatives (FN) correspond to medications or attributes present in the reference data but omitted by the model. An expected attribute generated with an incorrect value or not correctly associated with its medication was also considered a false negative for the corresponding reference item.

Consequently, an incorrect attribute prediction may contribute simultaneously to the false-positive count, because an unsupported value was generated, and to the false-negative count, because the correct reference value was not recovered. This evaluation criterion is relevant to the present task because correctly recognizing a medication does not necessarily imply that its dose, route, frequency, or dosage instruction was correctly extracted or linked.

The structural evaluation was performed after verifying that the generated output was valid and parseable JSON. This preliminary validation checked whether the response could be processed by a standard JSON parser, whether the expected top-level structure was present, and whether at least one medication item was generated when medication evidence was present in the source text. Outputs containing malformed syntax, inconsistent quotation marks, invalid separators, additional text outside the JSON structure, or an unexpected top-level representation were considered syntactically invalid and were not eligible for subsequent FHIR validation.

The next evaluation layer assessed whether each generated *MedicationRequest* resource conformed to the HL7 FHIR specification. This validation considered whether the output contained only permitted elements, whether the generated values respected the expected FHIR data types and cardinalities, and whether the internal organization of the resource was structurally consistent. In particular, the evaluation considered the representation of the medication and its relationship with the corresponding dosage instructions.

The FHIR validation rate was defined as:

$$\mathrm{FHIRValidRate} = \frac{N_{\mathrm{valid}}}{N_{\mathrm{total}}} \tag{1}$$

where:

- N_{valid} is the number of generated *MedicationRequest* resources that passed the FHIR validation process;
- N_{total} is the total number of generated *MedicationRequest* resources submitted to FHIR validation.

A resource contributed to N_{valid} only when it satisfied the structural rules evaluated by the FHIR validation process. Therefore, a clinically plausible resource could still be classified as invalid when it contained an unsupported element, an incorrect data type, an invalid cardinality, or an inconsistent resource structure.

Structural validity alone, however, does not guarantee that the generated resource contains sufficient clinical information for the proposed task. A *MedicationRequest* may be structurally valid while containing only a medication name and no usable dosage information. Therefore, a minimum-completeness criterion was defined for each generated resource.

For each resource i, the binary minimum-completeness indicator was defined as:

$$C_i = 1\Big(M_i \wedge (D_i \vee T_i)\Big) \tag{2}$$

where:

- C_i is the minimum-completeness indicator for resource i, assuming the value 1 when the resource satisfies the defined criterion and 0 otherwise;
- M_i represents the condition that the medication is present in resource i, for example through a populated medication representation such as *medicationCodeableConcept.text*;
- D_i represents the condition that structured dose information is present, for example through the corresponding dose value and unit in the dosage representation;
- T_i represents the condition that a textual dosage instruction is present, for example through a populated *dosageInstruction.text* element;
- $\wedge$ denotes the logical *AND* operator;
- $\vee$ denotes the logical *OR* operator.

According to this expression, a resource is considered minimally complete only when the medication is identified and at least one form of dosage information is available. The

expression $D_i \vee T_i$ means that either a structured dose or a textual dosage instruction is sufficient to satisfy the dosage-information component of the criterion. However, because this expression is connected to M_i through the logical operator $\wedge$, dosage information alone is not sufficient when the medication itself has not been identified.

For example, when $M_i = 1$ and $D_i = 1$, the resource is considered minimally complete regardless of the value of T_i. Similarly, when $M_i = 1$, $D_i = 0$, and $T_i = 1$, the resource is also considered minimally complete. Conversely, when $M_i = 0$, the indicator is equal to zero even if dose or textual instruction information is present, because that information cannot be reliably associated with an identified medication.

The overall minimum completeness rate was calculated as:

$$\mathrm{MinCompletenessRate} = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{3}$$

where:

- N is the total number of generated resources evaluated for minimum completeness;
- i identifies each individual resource in the evaluated set;
- C_i is the binary minimum-completeness indicator calculated for resource i;
- $\sum_{i=1}^{N} C_i$ is the number of resources that satisfied the minimum-completeness criterion.

Because C_i is binary, the summation counts the number of minimally complete resources. Dividing this count by N produces the proportion of evaluated resources containing an identified medication and either a structured dose or a textual dosage instruction.

Finally, because each discharge medication section may contain multiple medications, the output was generated as a FHIR *Bundle*. The complete *Bundle* rate was used to assess the model's ability to produce a structurally consistent response containing the expected set of medication resources.

The metric was defined as:

$$\mathrm{BundleCompleteRate} = \frac{B_{\mathrm{valid}}}{B_{\mathrm{total}}} \tag{4}$$

where:

- B_{valid} is the number of generated *Bundles* whose set of resources was valid and structurally consistent;
- B_{total} is the total number of generated *Bundles* evaluated.

For a *Bundle* to contribute to B_{valid}, its top-level structure and entries had to be consistent with the expected FHIR representation. In addition, the generated set had to contain the expected medication resources without structural failures that compromised the response as an interoperable exchange unit. Therefore, a *Bundle* containing some valid individual resources could still be considered incomplete when one or more expected medications were omitted, when unsupported or duplicated medication resources were generated, or when structural inconsistencies affected the complete response.

All structural metrics were calculated as proportions and reported as percentages. The separation among FHIR validation, minimum completeness, and complete *Bundle* evaluation is necessary because these metrics capture different aspects of structured generation. A resource may pass FHIR validation while remaining clinically incomplete, and a *Bundle* may contain individually valid resources while failing to represent the complete set of medications described in the source text.

III. Results

This section presents the results obtained from dataset construction, execution of the discharge medication section extraction pipeline, and the first experiments with the LLMs.

A. Dataset and Pipeline Results

The pipeline was fully executed on the set of hospital discharge reports from MIMIC-IV-Note. Of the total processed documents, 323,917 contained an identifiable *Discharge Medications* section, corresponding to 97.6% of the analyzed set. Subsequently, 311,814 documents presented a valid structure for segmentation, that is, they contained numbered items compatible with the pipeline's heuristic, representing 94.0% of the processed documents.

Despite the high coverage in the section localization and text segmentation stages, only 55,118 documents were successfully processed through all stages, including the deterministic linkage between each extracted textual line and the structured medication records associated with the same encounter. This subset accounts for 16.6% of the documents initially processed.

The total number of extracted and linked medications was 361,428, with an average of 6.56 medications per valid document. The main bottleneck of the pipeline lies in the deterministic matching stage between free text and structured records, which requires full correspondence between the items reported and the records for the same *hadm_id*. This constraint increases the reliability of the paired examples but substantially reduces the final dataset coverage.

Processing failures were mainly due to incomplete medication matching identified in the discharge report, accounting for 153,598 cases, or 46.3% of the total processed documents. This error occurs when at least one textual line extracted from the medication section fails to match any structured record in the encounter, invalidating the entire document. A total of 103,098 execution exceptions were also observed, corresponding to 31.1%, in addition to 12,103 cases of invalid structure, 7,233 cases of absence of the medication section, and 643 cases of absence of a structured prescription file.

From the final subset of 55,118 discharge documents, 1,853 unique medications were identified, with a highly imbalanced distribution. To mitigate imbalance between training and testing, medications were grouped into six frequency classes, following a stratified approach [27], namely: *very_frequent*, *frequent*, *moderate*, *rare*, *very_rare*, and *singleton*. Most entities were concentrated in the most frequent classes, with 216,944 instances in *very_frequent* and 83,633 in *frequent*. The

remaining classes had 35,542 instances in *moderate*, 18,592 in *rare*, 5,358 in *very_rare*, and 271 in *singleton*.

The train/test split followed a 90/10 ratio and, after this process, proportional preservation of the frequency classes was observed between the subsets, indicating that the resulting partition maintained adequate representativeness of the different medication strata.

B. Medication Extraction Results

Table I presents the results of the Gemma 4 E2B-it and Gemma 4 E4B-it models in the *zero-shot*, *few-shot*, and LoRA scenarios. Four extraction dimensions were evaluated: medication, dosage instruction, route of administration, and frequency. For each dimension, precision, recall, and *F1-score* were calculated.

The results indicate that the *zero-shot* scenarios did not yield useful extractions for either of the two evaluated models, with metrics equal to zero across all dimensions.

In the *few-shot* scenario, a slight improvement was observed in both models when compared with *zero-shot*. However, the values obtained remained low, suggesting that simply including examples in the *prompt* was insufficient to capture the task's complexity.

The LoRA fine-tuning scenario showed substantial improvement in both models, in contrast to the results obtained in the *zero-shot* and *few-shot* scenarios. This behavior indicates that supervised adaptation was decisive in enabling the models to learn domain- and task-specific patterns, including the clinical vocabulary of discharge prescriptions, the relationships between medications and dosage attributes, and the expected format of the structured FHIR output.

Considering only the LoRA fine-tuned models, Gemma 4 E4B-it outperformed E2B-it across all evaluated dimensions. The difference was particularly expressive for *Medication*, with an absolute gain of 25.29 *F1-score* points, and for *Freq*, with a gain of 22.16 points.

C. Validity of JSON and Structured Output

In the *zero-shot* and *few-shot* scenarios, both models yielded null values for the global metrics, indicating that the lack of adaptation or examples made it infeasible to generate useful structured outputs in the expected format. After LoRA fine-tuning, substantial improvements were observed in the global metrics.

D. Validity of FHIR MedicationRequest

The FHIR validity evaluation shows that generating *MedicationRequest* involves not only the correct extraction of clinical content but also adherence to the resource's formal structure, including expected fields, data types, and cardinalities. Thus, even clinically plausible outputs may be considered invalid when they contain structural inconsistencies.

In the scenarios without fine-tuning, the models produced virtually no valid FHIR resources. After LoRA fine-tuning, a substantial improvement was observed, especially for Gemma 4 E4B-it, which reached a *FHIRVRate* of 58.08. However, the low *BundleCRate* indicates that generating valid individual resources does not guarantee compliance with the complete response, which is composed of multiple medications. These results suggest that the fine-tuned Gemma 4 E4B-it presents initial feasibility as a proof of concept, but still requires complementary strategies, such as post-generation validation, automatic JSON/FHIR correction, and eventual deterministic serialization from an intermediate extraction.

E. Error Analysis

The error analysis indicated three main sources of failure. The first is related to clinical extraction, including medication omissions, partial identification of the drug name, and incorrect association of attributes such as dose, route, and frequency. These errors persisted in the fine-tuned models across more complex prescriptions, such as those with multiple instructions, dose ranges, conditional use, or limited duration.

The second source of error is associated with the generation of structural output. Even when the clinical content was correctly identified, invalid JSON, misplaced FHIR fields, elements not present in the specification, or inadequate cardinalities were observed. This behavior explains the difference between the extraction results and the global FHIR validity metrics.

The third source of error refers to *Bundle* completeness. In documents with multiple medications, the models may still omit items, duplicate resources, or generate individually plausible resources that do not preserve the global consistency of the response. In summary, LoRA fine-tuning considerably improved performance, but FHIR validity and *Bundle* completeness remain the main bottlenecks for operational use.

IV. Discussion

The results indicate that transforming hospital discharge medication texts into FHIR *MedicationRequest* resources is feasible as an experimental task, especially when the model is fine-tuned with supervised data. The comparison between the evaluated scenarios shows that strategies based only on *prompting*, particularly *zero-shot* and *few-shot*, were insufficient to simultaneously handle clinical extraction and structured generation. In contrast, LoRA fine-tuning improved the extraction metrics, suggesting that adaptation to the domain and output format was decisive for the observed performance.

The main evidence of feasibility lies in the performance of the fine-tuned Gemma 4 E4B-it, which achieved an *F1-score* of 86.00 for medication, 92.15 for dosage instruction, 93.88 for route of administration, and 87.37 for frequency. These results indicate that the model learned relevant patterns of the task, including the clinical vocabulary in discharge prescriptions and the association between medications and dosage attributes. However, the global structuring metrics remained lower than the extraction metrics, especially *BundleCRate*, indicating that generating a complete and validatable FHIR response is still more challenging than the isolated identification of entities and attributes.

TABLE I
RESULTS OF THE GEMMA 4 E2B-IT AND E4B-IT MODELS IN THE ZERO-SHOT, FEW-SHOT, AND LORA SCENARIOS.

Model	Strategy	Medication			Dosage Instruction			Route			Frequency			Global Metrics		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	MinCRate	FHIRVRate	BundleCRate
Gemma 4 E2B-it	zero-shot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	few-shot	6.23	3.27	4.29	15.91	8.81	11.34	12.20	6.87	8.79	6.18	3.58	4.53	0.00	0.04	0.00
	LoRA	57.75	63.98	60.71	85.18	86.00	85.58	84.82	85.95	85.38	61.04	69.99	65.21	20.40	23.97	2.72
Gemma 4 E4B-it	zero-shot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	few-shot	9.10	6.56	7.62	11.58	13.91	12.64	16.45	10.07	12.49	33.48	10.46	15.94	0.02	0.10	0.00
	LoRA	84.79	87.24	86.00	92.74	91.56	92.15	97.93	90.14	93.88	88.75	86.04	87.37	51.13	58.08	12.55

Note: P = precision; R = recall; F1 = F1-score; Medication = medication; DoseInst = dosage instruction; Route = route of administration; Freq = frequency; MinCRate = minimum completeness rate; FHIRVRate = FHIR validation rate; BundleCRate = complete Bundle rate.

A. Interpretation of LLM Results

The comparison between the E2B-it and E4B-it variants suggests that increased capacity within the same model family improved performance on the task. Gemma 4 E4B-it outperformed E2B-it across all evaluated dimensions after fine-tuning, with larger gains in medication identification and frequency extraction. This behavior suggests that the larger variant showed better generalization capacity for more variable textual patterns, such as medication names, abbreviations, and frequency expressions.

Even so, the results show that high extraction performance does not necessarily imply full structural compliance. The fine-tuned Gemma 4 E4B-it obtained a *FHIRVRate* of 58.08 and a *BundleCRate* of 12.55, values lower than the *F1-scores* observed in the extraction dimensions. This difference reinforces that the task has two distinct layers: the first related to understanding and extracting clinical content; and the second related to correctly serializing this content into an interoperable artifact. Thus, generating *MedicationRequest* from free text should be understood as a composite task that combines information extraction, attribute linking, and compliance with a formal data model.

The results obtained play an important methodological role within the broader research. First, they demonstrate that the problem formulation is operationalizable at scale using MIMIC-IV-Note. Second, they indicate that supervised fine-tuning is more suitable than direct use of *prompting* for this specific task. Third, they show that the evaluation must separate clinical extraction metrics from structural validity metrics, since both capture distinct dimensions of the problem. Finally, the findings guide the next experiments with other open model families, allowing investigation of whether the patterns observed in Gemma 4 hold across different architectures and models with different sizes, vocabularies, alignment strategies, and computational costs.

V. CONCLUSION

This article presented a study using Gemma 4 as a proof-of-concept to transform hospital discharge medication prescription texts into FHIR *MedicationRequest* resources. Based on a dataset built from MIMIC-IV-Note, MIMIC-IV, and MIMIC-IV on FHIR, the ability of the Gemma 4 E2B-it and E4B-it models was evaluated in three experimental scenarios: *zero-shot*, *few-shot*, and LoRA fine-tuning.

The results show that strategies based only on *prompting* were insufficient for the task, whereas supervised fine-tuning produced substantial gains in extraction metrics. The fine-tuned Gemma 4 E4B-it showed the best overall performance, especially in extracting medication names, dosage instructions, routes of administration, and frequencies. However, the FHIR validity and *Bundle* completeness metrics remained lower than the extraction metrics, indicating that generating complete, structurally valid, and interoperable resources remains a relevant challenge.

As a contribution, this work demonstrates the initial feasibility of using relatively low-complexity open models to support the transformation of free clinical text into FHIR representations. In addition, it proposes an experimental formulation that distinguishes the evaluation of clinical extraction from that of structural compliance, an essential aspect for health interoperability applications.

Although the results are promising, this evaluation should be interpreted as a proof-of-concept based on a single model family. Two Gemma 4 variants were evaluated, but both share similar architectural characteristics, training strategies, and instructional behavior. Therefore, the results do not allow the conclusion that the observed performance is representative of other open-weight models or other LLM families. Another limitation concerns the scope of the task. The study focused on the *Discharge Medications* section and on a minimum subset of attributes associated with *MedicationRequest*, especially medication, dose, route, and frequency. Additional fields, such as duration, dispensed quantity, refill, clinical intent, status, references to the patient and encounter, and standardized terminological coding, were not fully explored. In addition, semantic validation was conducted based on an operational reference derived from MIMIC-IV and MIMIC-IV on FHIR, rather than from an exhaustive manual clinical review of each generated resource.

REFERENCES

[1] A. Johnson et al., 'MIMIC-IV'. PhysioNet, 2024.

[2] J. D. Wong et al., 'Medication reconciliation at hospital discharge: evaluating discrepancies', Ann. Pharmacother., vol. 42, no. 10, pp. 1373–1379, Oct. 2008.

[3] K. Jeon, W. Y. Park, C. E. Kahn Jr, P. Nagy, S. C. You, and S. H. Yoon, 'Advancing medical imaging research through standardization: The path to rapid development, rigorous validation, and robust reproducibility', Invest. Radiol., vol. 60, no. 1, pp. 1–10, Jan. 2025.

[4] Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo, 'Natural language processing for EHR-based computational phenotyping', IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 16, no. 1, pp. 139–153, Jan. 2019.

[5] Y. Shen, J. Yu, J. Zhou, and G. Hu, 'Twenty-five years of evolution and hurdles in Electronic health records and interoperability in medical research: Comprehensive review', J. Med. Internet Res., vol. 27, p. e59024, Jan. 2025.

[6] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, 'Large language models in medicine', Nat. Med., vol. 29, no. 8, pp. 1930–1940, Aug. 2023.

[7] K. Stanceski et al., 'The quality and safety of using generative AI to produce patient-centred discharge instructions', NPJ Digit. Med., vol. 7, no. 1, p. 329, Nov. 2024.

[8] E. Asgari et al., 'A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation', NPJ Digit. Med., vol. 8, no. 1, p. 274, May 2025.

[9] Google DeepMind, "Gemma 4: Technical Report," Google DeepMind, Tech. Rep., 2025. [Online]. Available: deepmind.google

[10] HL7 International, "Resource MedicationRequest - Content," HL7 FHIR v6.0.0-ballot4, Jun. 09, 2026. [Online]. Available: https://build.fhir.org/medicationrequest.html [Accessed: May. 20, 2026].

[11] HL7 International, "RESTful API," HL7 FHIR v6.0.0-ballot4, Jun. 09, 2026. [Online]. Available: https://build.fhir.org/http.html [Accessed: May. 20, 2026].

[12] HL7 International, "Resource Bundle - Content," HL7 FHIR v6.0.0-ballot4, Jun. 09, 2026. [Online]. Available: https://build.fhir.org/bundle.html [Accessed: May. 20, 2026].

[13] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, '2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records', J. Am. Med. Inform. Assoc., vol. 27, no. 1, pp. 3–12, Jan. 2020.

[14] D. Bender and K. Sartipi, 'HL7 FHIR: An Agile and RESTful approach to healthcare information exchange', in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, 2013.

[15] M. Ayaz, M. F. Pasha, M. Y. Alzahrani, R. Budiarto, and D. Stiawan, 'The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities', JMIR Med. Inform., vol. 9, no. 7, p. e21929, July 2021.

[16] Y. Chen et al., 'Prompt robust large language model for Chinese medical named entity recognition', Inf. Process. Manag., vol. 62, no. 5, p. 104189, Sept. 2025.

[17] Q. Xu et al., 'Towards normalized clinical information extraction in Chinese radiology report with large language models', Expert Syst. Appl., vol. 271, no. 126585, p. 126585, May 2025.

[18] D. Sun et al., 'Large language model-assisted information extraction from clinical reports for survival prediction of bladder cancer patients', in Medical Imaging 2024: Computer-Aided Diagnosis, San Diego, United States, 2024.

[19] B. Le Guellec et al., 'Performance of an Open-Source large Language Model in extracting information from free-text radiology reports', Radiol. Artif. Intell., vol. 6, no. 4, p. e230364, July 2024.

[20] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, 'An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study', JMIR Med. Inform., vol. 12, p. e55318, Apr. 2024.

[21] J. Zhang, J. Zhou, L. Zhou, and Z. Ba, 'Extracting multifaceted characteristics of patients with chronic disease comorbidity: Framework development using large language models', JMIR Med. Inform., vol. 13, p. e70096, May 2025.

[22] Y. Hu et al., 'Improving large language models for clinical named entity recognition via prompt engineering', J. Am. Med. Inform. Assoc., vol. 31, no. 9, pp. 1812–1820, Sept. 2024.

[23] J. Delaunay, D. Girbes, and J. Cusido, 'Evaluating the effectiveness of large language models in converting clinical data to FHIR format', Appl. Sci. (Basel), vol. 15, no. 6, p. 3379, Mar. 2025.

[24] A. Idrissi-Yaghir et al., 'Using a diverse test suite to assess large language models on fast health care interoperability resources knowledge: Comparative analysis', J. Med. Internet Res., vol. 27, no. v27i1e73540, p. e73540, Aug. 2025.

[25] A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark, 'MIMIC-IV-Note: Deidentified free-text clinical notes'. PhysioNet, 2023.

[26] A. Bennett et al., 'MIMIC-IV on FHIR'. PhysioNet, 2024.

[27] K. Sechidis, G. Tsoumakas, and I. Vlahavas, 'On the Stratification of Multi-label Data', in Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 145–158.

H-Q-MAS: Deterministic-First Anonymization of Brazilian Clinical Text under the LGPD

Guilherme Pinheiro de Oliveira
Graduate Program in Electrical Engineering (PPGEE)
Federal University of Ceará (UFC)
Fortaleza, Brazil
guilherme100.0@hotmail.com

Arthur Pinheiro Freitas Silveira
Graduate Program in Electrical Engineering (PPGEE)
Federal University of Ceará (UFC)
Fortaleza, Brazil
arthurpfsilveira@gmail.com

Victor Hugo C. de Albuquerque
Department of Teleinformatics Engineering (DETI)
Federal University of Ceará (UFC)
Fortaleza, Brazil
victor.albuquerque@lesc.ufc.br

***Abstract*—The Brazilian General Data Protection Law (LGPD) requires that clinical records be anonymized without destroying their clinical utility before secondary use. We present H-Q-MAS (Hierarchical Quantum Multi-Agent System), a LangGraph pipeline that combines a deterministic structured layer, a statistical recognizer (spaCy and BERTimbau with a LoRA adapter), a clinical-utility shield, and a two-qubit variational quantum circuit (VQC) for per-entity decisions. On 50 synthetic Brazilian clinical texts across eight specialties (148 spans of personally identifiable information, PII; 355 clinical terms), the system leaks 0 of 148 PII spans (privacy recall $= 1.0$; one-sided 95% lower bound ≥ 0.980) and retains 251 of 355 clinical terms (lexical preservation 0.7070 pooled, 0.6979 macro). The main contribution is a per-detector ablation showing where the result comes from. Privacy is produced by the structured layer (without it, recall falls to 0.4392) and preservation by the shield (without it, 0.5154). spaCy is the sole detector of 10 name spans but also the main source of clinical loss: removing it leaks those 10 spans and raises preservation to 0.9311. The BERTimbau component, as deployed, has an effectively untrained classification head and detects nothing on its own. Replacing the VQC by a fixed threshold has no measurable effect (Mann-Whitney $p = 0.97$). Free-text names remain weak: an inline first name is redacted in only 18 of 48 external texts (recall 0.375). Presidio with Brazilian regex recognizers leaks 21 of 148 gold spans (preservation 0.6789). The results support a deterministic-first design in which learned components must earn their place by per-component ablation.**

***Index Terms*—clinical text anonymization, LGPD, multi-agent systems, LangGraph, privacy recall, ablation, variational quantum circuits**

I. Introduction

Electronic clinical records concentrate sensitive personal data: patient names, the Brazilian individual taxpayer registry (CPF), the national health card (CNS), professional council numbers (CRM and others), addresses, and contact details. Brazil's General Data Protection Law (LGPD, Lei 13.709/2018) [1] conditions secondary use of such data on effective anonymization, and its Article 12 ties the legal status of anonymized data to whether re-identification is reasonably prevented. A Brazilian hospital faces two practical constraints. Anonymization must recognize country-specific identifiers (CPF, CNS, council numbers) that fall outside the HIPAA categories targeted by English-language de-identification systems [2]. Processing should also run on-premises. That is not only a deployment choice: it is a question of health-data sovereignty, keeping sensitive records of Brazilian patients on local infrastructure under national law instead of exporting raw data to external jurisdictions. The concern is concrete in the public health system (SUS), where patient records are produced at scale under public stewardship and jurisdiction over the data is itself a matter of public interest [3].

Aggressive redaction removes identifiers but also strips the measurements, drug names and findings that give a record secondary value; conservative redaction keeps that utility at the cost of residual leakage. A usable system therefore needs both high-recall identifier removal and an explicit mechanism to retain clinically meaningful terms.

This paper does not argue that a single clever model solves the problem. Its main contribution is methodological: we build a multi-agent pipeline and then measure, by ablation, which of its parts actually carry the result. Specifically, we contribute: (i) a component-level evaluation on a manually annotated gold standard showing that privacy recall is produced by the structured layer and clinical preservation by the clinical-utility shield; (ii) three results that limit the role of the learned components, reported rather than omitted: a two-qubit quantum consensus statistically equivalent to a fixed threshold, a spaCy recognizer that is the sole detector for only 10 of 148 PII spans (6.8%) while causing most of the clinical loss, and a BERTimbau component that, as deployed, contributes no detection because its classification head is effectively untrained; (iii) a same-corpus comparison with Microsoft Presidio under identical metrics; and (iv) H-Q-MAS itself, a LangGraph anonymizer with a residual-PII audit gate that runs on commodity CPU, which is the artifact on which these measurements were taken.

II. Related Work

Automatic de-identification has been driven by shared tasks and annotated corpora, notably the 2014 i2b2/UTHealth challenge [4], on which neural sequence models reached high recall [2]; that line targets English HIPAA categories, whereas our setting is Brazilian Portuguese under the LGPD, with country-specific identifiers (CPF, CNS, council numbers) and an explicit clinical-utility constraint. Recent work targets this Brazilian Portuguese setting directly: rule-based de-identification of real hospital records [5], NER- and LLM-based anonymization with a dedicated annotated corpus [6], and a recent benchmark of modern models for Portuguese clinical NER [7], building on Portuguese clinical NLP resources such as the SemClinBr corpus [8] and the BioBERTpt model [9]. In the multi-agent space, EPEAgent embeds privacy-enhancing agents to limit information flow between agents [10], MedPAO structures clinical reports through a plan-act-observe loop [11], and role-specialized agents have simulated multidisciplinary decision making with explicit consensus [12]. On the quantum side, a hybrid classical-quantum BERT couples a qubit circuit with a transformer for text classification [13], and a recent survey maps QNLP models and their data-scale limits [14]; we include a quantum consensus component but report its effect by ablation rather than assuming an advantage.

III. Method

A. Architecture

H-Q-MAS is a directed graph of agents executed with LangGraph (Fig. 1). A broker initializes the run; a named-entity recognition (NER) agent and a context agent run in parallel; a coordinator performs fan-in, immediately classifies hard identifiers, and dispatches one decision task per remaining entity through LangGraph's Send API (dynamic fan-out); an auditor applies the anonymization, rescans the output for residual PII with the structured layer, and computes a clinical-information-loss (CIL) score. If the rescan finds residual spans, a conditional edge routes them to a gate node that force-redacts them and returns to the auditor, for at most two passes; the gate reuses the structured detector, so by construction it can only catch what that detector matches, and it did not fire on any text evaluated in this paper. After the auditor accepts the output, a *billing* node extracts ICD-10 (CID-10) codes from the anonymized text by regular expressions and a keyword dictionary, without any language model, for downstream hospital billing, and a persistence node stores the output and its audit record. Both run after anonymization, do not modify the anonymized text, and do not enter any reported metric.

B. Deterministic structured layer

The structured layer detects structured identifiers with regular expressions and checksum validation (e.g., CPF, CNS, CRM and other council numbers, RG, CEP, e-mail, phone, dates), and detects person and organization names with header-anchored patterns: a role keyword ("Paciente", "Médico",

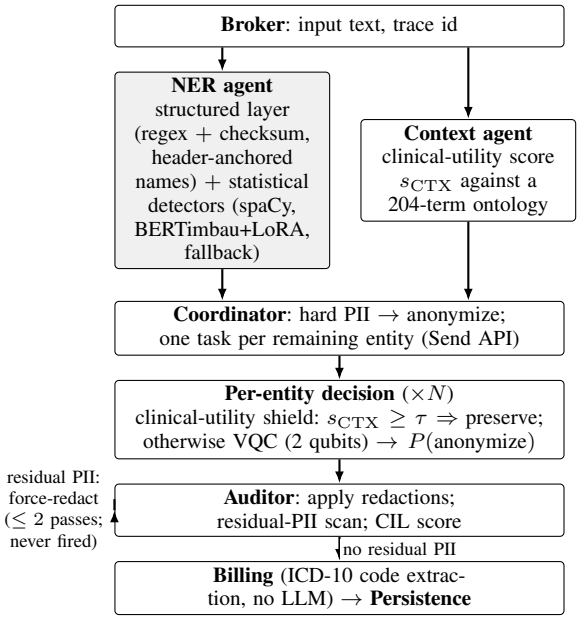


Fig. 1. H-Q-MAS agent graph. NER: named-entity recognition; PII: personally identifiable information; VQC: two-qubit variational quantum circuit; CIL: clinical-information-loss score; s_{CTX}: clinical-utility score of the context agent; τ: clinical-utility bypass threshold (Sec. III-C); billing: ICD-10 code extraction from the anonymized text. The shaded node is the NER agent, which contains the structured layer and the statistical detectors (spaCy, BERTimbau+LoRA, unsupervised fallback). Residual PII found by the auditor's rescan is force-redacted by a gate node (at most two passes) before billing and persistence; the gate never fired in our evaluation.

"Solicitante", ...) followed by a delimiter and a Title-Case multi-token name, with propagation of a matched name to its other occurrences. Identifiers in a fixed hard PII set are anonymized deterministically without consulting the quantum step.

C. Statistical recognizer and clinical-utility shield

The statistical recognizer complements the structured layer for names in free text. It merges three detectors, in priority order below the structured layer: BERTimbau [15] with a LoRA adapter [16], spaCy [17] (`pt_core_news_lg`), and an unsupervised fallback. We report a defect of the evaluated system rather than hide it: the adapter file stores a 13-label token-classification head whose weights are statistically indistinguishable from their random initialization (weight standard deviation 0.020, bias magnitude ≤ 0.0013), so as deployed the BERTimbau component emits near-uniform, essentially random candidates (mean maximum class probability 0.11 against 0.077 for a uniform output). The head is loaded from the file, not re-initialized, so its output is deterministic. It remains part of the system we measured; Sec. V-A quantifies its effect by ablation. The context agent scores each candidate token's clinical utility, $s_{\mathrm{CTX}} \in [0, 1]$, by cosine similarity against a 204-term Brazilian clinical ontology. The clinical-utility shield preserves any candidate with $s_{\mathrm{CTX}} \geq \tau$, where $\tau = 0.45$ is a fixed bypass threshold; person-name candidates

are exempt from the shield (s_{CTX} is forced to 0), and a preserved candidate never reaches the quantum step.

D. Quantum consensus (VQC)

For each entity that is neither hard PII nor preserved by the shield, the system evaluates a two-qubit circuit (PennyLane [18] `default.qubit` simulator, not quantum hardware): $R_Y(\pi\, s_{\mathrm{NER}})$ on qubit 0 and $R_Y(\pi\, s_{\mathrm{CTX}})$ on qubit 1 encode the recognizer and context scores, a CNOT entangles them, a trained rotation $R_Y\big(\pi\,(w_{\mathrm{NER}}\, s_{\mathrm{NER}} - w_{\mathrm{CTX}}\, s_{\mathrm{CTX}})\big)$ is applied to qubit 0, and measuring qubit 0 yields the probability p of anonymizing. The entity is anonymized when $p > \alpha + \beta\, s_{\mathrm{CTX}}$, i.e., the decision threshold is a fixed function of the context score: a token with higher clinical utility requires stronger evidence before it is redacted. The circuit weights (w_{NER}=0.464, w_{CTX}=0.493) were fitted by gradient descent on a synthetic set of 153 $(s_{\mathrm{NER}}, s_{\mathrm{CTX}})$ points (a 10×10 grid, 50 random points and 3 demonstration entries) labelled by a rule; the threshold parameters (α=0.50, β=0.45) were set by manual calibration. A grid search on the same synthetic set proposed α=0.575, β=0.80, but this setting was discarded because it let borderline names ($p \approx 0.57$) through, and the manual values were retained. All four values are frozen at inference and were never tuned on the gold standard. Note that β and τ coincide numerically but are distinct quantities: τ is an absolute cutoff applied before the circuit, β scales the circuit's threshold.

E. LGPD instrumentation and inference

The auditor computes a deterministic CIL score over non-PII clinical terms, used as an Article-12 utility-loss indicator. An optional local LLM (Ollama, `llama3:8b`) is queried only to emit a qualitative audit annotation; it does not alter the anonymized text or any reported metric (Sec. IV). The code is available from the authors on request.

IV. Experimental Setup

Data. The gold standard consists of 50 synthetic Brazilian clinical texts (reports, exam reports, progress notes, consultation notes, discharge summaries, prescriptions and opinions) spanning eight specialties, written and annotated by one of the authors with 148 PII spans and 355 clinical terms. Annotation was performed by a single annotator, so inter-annotator agreement is not available. The same author also designed the header-anchored name heuristic, so the annotation and the rule set are not independent. The LoRA adapter was fine-tuned on the public HAREM newswire NER corpus, not on clinical text. No real patient data were used at any stage, including model training, and the study therefore did not require ethics-committee approval. **Metrics.** *Privacy recall* (PR) is the fraction of annotated PII spans absent from the output (case-insensitive substring match). *Clinical preservation* (CP) is the fraction of annotated clinical terms still present in the output, a *lexical* retention measure by substring match, *not* a semantic equivalence judgement. We report PR and CP both pooled (micro) and macro-averaged over texts, with raw counts so that confidence intervals are well defined. **Determinism.** PR and CP are deterministic by construction: the anonymization decisions come from regular expressions, checksum validation, argmax over statistical model outputs, and a fixed-parameter VQC with a fixed threshold; the only stochastic component (the `llama3:8b` audit annotation) does not feed the anonymized text or the router. No random seed is fixed at inference; all model weights, including the BERTimbau classification head, are loaded from files rather than initialized at run time. All numbers in this paper were produced at one code state (git commit `b3a9152`); seven independent process launches of the gold-standard evaluation (each loading the models afresh), plus the originally submitted run, produced identical per-text PR and CP values ($\Delta = 0$), with latency varying as expected. **Environment.** Python 3.12 on CPU; Ollama `llama3:8b` (Q4_0); BERTimbau in PyTorch (no ONNX quantization). **Baseline.** We run Microsoft Presidio (`presidio-analyzer` 2.2.364) with the spaCy `pt_core_news_lg` model, its default recognizers, and three added pattern recognizers for Brazilian CPF, CEP and phone numbers, at score threshold 0.4, on the same 50 texts, scored with the same PR/CP code. **External setting.** For robustness we also evaluate on 48 synthetic realistic texts (40 specialty-labelled notes and 8 general records) in which *well-formed PII is injected* into a header block at known positions, plus one inline mention of the patient's first name in free text, scored separately (Sec. V-G); clinical references there are extracted automatically from the ontology, which we discuss as a measurement caveat in Sec. VI.

V. Results

A. Where performance comes from (ablation)

We disable one component at a time on the gold standard, on the same code state as the headline run (Table I). Removing the structured layer collapses privacy recall from 1.0 to 0.4392 (83 of 148 spans leak); removing the clinical-utility shield leaves privacy recall intact but drops macro preservation from 0.6979 to 0.5154 (in raw terms, retained clinical terms fall from 251 to 184 of 355; $184/355 = 0.5183$ pooled vs. 0.5154 macro). Privacy is carried by the structured layer; preservation by the clinical-utility shield. Removing the whole statistical recognizer (BERTimbau with LoRA, spaCy and the unsupervised fallback all disabled; structured layer, shield and VQC kept) leaks exactly the 10 name spans that the structured layer misses (PR 0.9324, 10 of 148, one per text in 10 texts) and raises macro preservation from 0.6979 to 0.9370 (333 of 355 terms retained).

Decomposing that ablation by detector shows where each effect lives. Disabling only spaCy reproduces the 10 leaks and almost all of the gain (CP 0.9311, 331 of 355): spaCy is the sole detector of those 10 names and the source of the clinical loss. Disabling only the BERTimbau component leaks nothing (PR 1.0) and *lowers* preservation to 0.6801 (245 of 355); disabling only the unsupervised fallback leaks nothing and moves preservation to 0.7001. The difference between 82 terms recovered without the whole recognizer and 80 without

TABLE I
COMPONENT ABLATION AND BASELINE ON THE GOLD STANDARD (N=50 TEXTS, 148 PII SPANS, 355 CLINICAL TERMS). LEAKED-SPAN COUNTS IN PARENTHESES; CP IS MACRO-AVERAGED (POOLED VALUES IN THE TEXT); Δ IN PERCENTAGE POINTS VS. THE FULL SYSTEM.

Configuration	PR	CP	Δ (pp)
H-Q-MAS, full system	1.0000 (0/148)	0.6979	—
− structured layer	0.4392 (83/148)	0.6842	PR -56.1; CP -1.4
− clinical-utility shield	1.0000 (0/148)	0.5154	CP -18.3
− statistical recognizer	0.9324 (10/148)	0.9370	PR -6.8; CP $+23.9$
− spaCy only	0.9324 (10/148)	0.9311	PR -6.8; CP $+23.3$
− BERT+LoRA only	1.0000 (0/148)	0.6801	CP -1.8
− fallback only	1.0000 (0/148)	0.7001	CP $+0.2$
VQC $\rightarrow$ fixed 0.5	1.0000 (0/148)	0.6976	CP -0.03
Presidio (pt + BR regex)	0.8581 (21/148)	0.6789	PR -14.2; CP -1.9

spaCy alone comes from two terms that the fallback detector still redacts when spaCy is off.

The 82 clinical terms recovered without the statistical recognizer are all spans that spaCy proposed with $s_{\mathrm{CTX}} < \tau$ and that the VQC step then decided to anonymize (82 of 82 carry the VQC decision reason); the fixed threshold makes the same decisions (Sec. V-F). The six additional terms lost without BERTimbau (e.g., "AAS 100mg", "índice de Lee 1") are spaCy candidates that the VQC anonymizes; in the full system the effectively untrained BERTimbau head covers the same tokens with candidates of score ≈ 0.1 that win the merge (BERT ranks above spaCy) and are kept, i.e., the defective component preempts spaCy by accident. The end-to-end metrics alone did not reveal this: only the per-detector decomposition did.

Removing the structured layer also lowers preservation slightly (0.6979 to 0.6842 macro; 251 to 246 of 355 pooled) rather than raising it, as less redaction would suggest. The mechanism is not the residual-PII gate, which never fired in any configuration (0 passes in 350 pipeline runs), but the merge order of the recognizer: spans of the structured layer, including its clinical-safe whitelist, take precedence over statistical detections and cannot be overwritten. Without it, five additional clinical strings ("FC máxima: 162 bpm", "FR 28 irpm", "Ceftriaxona 2g EV", "Esquizofrenia paranoide", "Cintilografia óssea") are tagged by the statistical detectors and redacted; no term is recovered. Note that this ablation also disables the structured detector inside the auditor's rescan, so the row measures the layer and its gate together.

Replacing the VQC by a fixed 0.5 threshold on the recognizer score leaves privacy recall at 1.0 and moves macro preservation by 0.0003 (Sec. V-F).

B. Headline anonymization

On the gold standard the full system leaks no PII: 0 of 148 spans (PR $= 1.0$; one-sided 95% lower bound ≥ 0.980, Clopper-Pearson for 0/148). Clinical preservation is 0.6979 macro-averaged and 0.7070 pooled (251 of 355 terms retained; Wilson 95% CI $[0.658, 0.752]$). Per-text latency on CPU spans 722 to 1540 ms (median 741 ms, mean 784 ms, nearest-rank p95 1174 ms), measured at commit `b3a9152` with the optional observability tracing turned off; latency is the only non-deterministic quantity and is reported as a range.

C. Comparison with Presidio

Under identical texts and scoring, Presidio with Brazilian regex recognizers leaks 21 of 148 PII spans (PR $= 0.8581$ pooled, 0.862 macro; Wilson 95% CI $[0.793, 0.905]$), affecting 20 of the 50 texts, and retains 241 of 355 clinical terms (CP $= 0.6789$ macro; the pooled value is identical). Its per-text latency has median 39 ms and nearest-rank p95 67 ms, against 741 ms and 1174 ms for H-Q-MAS. The comparison isolates what the structured layer buys: 18 of the 21 leaks are council numbers (10) and CNS numbers (8), for which Presidio has no recognizer and which the structured layer matches by pattern and checksum; the remaining three are one phone number missed by the added pattern and two header-formatted patient names (Presidio's PERSON recognizer caught the other 70 name spans). Presidio was not tuned beyond the three added patterns, so the gap partly reflects configuration effort rather than an intrinsic limit of the tool.

D. Attribution of detection (structured vs. statistical)

Cross-referencing each annotated span with the structured layer's output, 138 of 148 spans (93.2%) are detected by the structured layer, with 79 by regex/checksum (structured identifiers: CPF, CNS, council numbers, dates, e-mail, phone, CEP) and 59 names by the header-anchored heuristic, while only 10 of 148 (6.8%) rely on the statistical recognizer. Of the 72 name spans, 59 are caught by the header heuristic, 3 by other structured patterns, and 10 only by spaCy (Sec. V-A). The BERTimbau+LoRA component detects none of the 148 spans on its own: its classification head is effectively untrained (Sec. III-C), and on control sentences it emits sub-word fragments with scores around 0.1. The privacy result does not depend on it at all, and depends on spaCy for 10 spans.

E. Per-specialty preservation

CP varies by specialty (Table II); four specialties fall below the macro mean of 0.6979 (Pediatrics 0.6160, Psychiatry 0.6238, Oncology 0.6663, Cardiology 0.6686) while Pulmonology (0.8189) and Orthopedics (0.8012) are highest. The two highest are based on very small samples (n=3 and n=5) and are reported as exploratory, not as reliable point estimates.

F. Quantum consensus: a negative result

We compare the VQC decision of Sec. III-D (circuit probability against the context-dependent threshold $\alpha+\beta s_{\mathrm{CTX}}$) with a classical control in which the same entities are anonymized when $s_{\mathrm{NER}} \geq 0.5$, over the 50 texts, both runs at the same code state. The per-document F1 distributions (harmonic mean of PR and CP) are statistically indistinguishable (Mann-Whitney $U = 1244.5$, two-sided $p = 0.9724$; Cohen's $d = 0.0032$, negligible), with mean F1 0.8076 (VQC) vs. 0.8071 (fixed threshold). The circuit does exhibit measurable entanglement (Meyer-Wallach $Q = 0.2497$) [19], but this does

TABLE II
CLINICAL PRESERVATION BY SPECIALTY, GOLD STANDARD: NUMBER OF TEXTS, RETAINED/ANNOTATED CLINICAL TERMS (POOLED), AND MACRO AVERAGE OVER TEXTS.

Specialty	n	Retained/terms	CP (macro)
Pediatrics	7	28/43	0.6160
Psychiatry	5	20/32	0.6238
Oncology	5	27/40	0.6663
Cardiology	10	51/76	0.6686
Endocrinology	8	36/52	0.7007
Neurology	7	41/53	0.7684
Orthopedics	5	28/35	0.8012
Pulmonology	3	20/24	0.8189

not translate into a performance gain. The training report of the circuit shows the same picture: on its synthetic training set of 153 points (Sec. III-D), accuracy rises from 86.27% to 95.42% only when the threshold is re-optimized by grid search, and the initial and trained circuit weights yield identical accuracy under that optimized threshold, i.e., the gain is attributable to the classical threshold, not to the quantum weights; moreover, that optimized threshold is not the one deployed (Sec. III-D), and the figures are training-set accuracies on a set whose majority class alone gives 83.0%. The headline result does not rest on this component: with the VQC active or replaced by the fixed threshold, privacy recall is identical (1.0, 0 leaks) and clinical preservation moves by 0.0003 (0.6979 vs. 0.6976, 251 vs. 250 terms); the two settings differ in per-document F1 on 5 of 50 texts. It is an object of study here, not a load-bearing part of the pipeline. Note, however, that the decision step (VQC or fixed threshold) is where the clinical loss materializes: the 82 terms discussed in Sec. V-A are all anonymized by this step.

G. External robustness

On the 48 synthetic realistic texts, privacy recall over the well-formed PII injected into the header block is 1.0 with no leaks, and per-type recall on those header spans is 1.0 for every injected type (full name, CPF, RG, date, phone, e-mail, address, CEP). Because the injected names follow the same header format that the structured heuristic targets, this confirms robustness on header-formatted identifiers rather than on free-text mentions (Sec. VI). The free-text case was scored directly: each text also carries one inline mention of the patient's first name ("O(a) paciente *Ana* segue em acompanhamento"), which is not a header and not an exact repeat of the full name, so the structured layer cannot propagate to it. That mention was redacted in only 18 of 48 texts (recall 0.375); in the other 30 the first name survived in the output. All 18 redactions were spaCy candidates (label PER, score 0.95, then anonymized by the decision step); neither the BERTimbau component nor the unsupervised fallback proposed any of them, and the same first name (e.g., "José") is caught in one text and missed in another. Free-text single-token names therefore depend entirely on spaCy, and spaCy misses most of them. Clinical preservation on this set is 0.9404 (macro average over the 39 texts that contain at least one of the 274 automatically extracted reference terms; 9 texts have none). This value is not comparable with the gold-standard CP: its references are extracted automatically from the same ontology that the shield protects, which inflates it by construction (Sec. VI); the gold CP of 0.6979 remains the headline.

VI. DISCUSSION AND LIMITATIONS

Privacy is mostly deterministic. The privacy recall is carried by regular expressions, checksum validation, and the header-anchored name heuristic (138 of 148 spans); spaCy supplies the remaining 10, and BERTimbau+LoRA none. Claims of learned NER models (transformer or otherwise) driving privacy would be unsupported here.

Implications for the architecture. The per-detector ablation separates the components into four roles. The structured layer and the clinical-utility shield are load-bearing: removing either changes a headline metric by 18 to 56 pp. spaCy is a fallback with a cost: it is the only detector for 10 of 148 spans, all names missed by the header heuristic, and removing it leaks exactly those 10 (PR 0.9324); it is also the main source of clinical loss, since without it 80 more clinical terms survive (CP 0.9311). We report the two quantities side by side rather than as a trade, because a leaked name and a redacted clinical term are not commensurable under the LGPD. The BERTimbau+LoRA component is defective as deployed (effectively untrained head) and contributes no detection; its only measurable effect is accidental, preempting a few spaCy candidates in the merge (CP -1.8 pp when removed). The quantum consensus is dispensable on this task: replacing it with a fixed threshold leaves privacy unchanged and moves preservation by 0.0003. The consequence is a deterministic-first design in which rules carry privacy, a properly trained statistical fallback covers what rules cannot express, and exploratory components such as the VQC stay outside the production path until an ablation shows an effect. The multi-agent structure is useful here less for accuracy than for measurability: because each agent is a separate node and each detector a separate strategy inside the NER agent, each can be disabled and evaluated in isolation, which is what made these conclusions, including the detection of the defective component, possible.

Small evaluation set. The gold standard has 50 texts and 148 PII spans. Zero observed leaks bound the true leak rate only to at most 2.0% (one-sided 95%), so up to about three leaks per 148 spans cannot be excluded, and per-specialty estimates rest on 3 to 10 texts and are exploratory. The texts are synthetic and single-annotated, and their layout is likely more regular than that of real hospital records, which favors the header heuristic. A larger, specialty-balanced set with multiple annotators, including real de-identified records under ethics approval, is required before any deployment claim.

Name detection depends on document layout. Names are the only PII type whose detection depends on where they appear. The header heuristic captures 59 of 72 name spans, and the external set injects full names in the same header block. The one direct measurement of a free-text mention, the inline first name in the external texts, gives recall 0.375 (18 of

48): for names that are neither in a header nor an exact repeat of a header name, such as relatives or referring physicians cited in narrative, the system falls back on spaCy alone and misses most of them. The reported privacy recall of 1.0 should therefore be read as recall on a corpus dominated by header-formatted identifiers, not on free narrative.

Baseline scope. The Presidio comparison uses one configuration (default recognizers plus three Brazilian patterns) and shows where the gap comes from, not the best Presidio can do; a Presidio equipped with the full structured layer of H-Q-MAS would presumably close most of it, which is consistent with our attribution. We do not compare against the Brazilian Portuguese systems of [5], [6]: the corpus of [5] is not public, and the public part of AnonyMED-BR [6] annotates PII only, so it would support a privacy-only comparison, which we leave as future work.

Preservation is lexical, not semantic. CP counts annotated clinical strings that survive verbatim. It penalizes harmless changes, such as a term partially covered by a redaction tag, and cannot detect harmful ones that keep the string but alter its meaning, such as a removed negation or dose context. A semantic evaluation, for example an NLI-based judge validated against clinician review of a sample, is needed to show that preserved strings preserve clinical information. The external CP (0.9404) is measured against references drawn from the same ontology the shield protects, which is partly circular; the gold CP (0.6979) remains the headline.

A. Lessons Learned and Research Gaps

Lessons learned. (1) Ablation before attribution: the components that motivated the design, neural recognition and quantum consensus, were not the ones carrying the result, and without the ablation the headline would have been credited to them. (2) Country-specific identifiers are largely a rules problem: CPF, CNS, and council numbers have fixed formats and check digits, and checksum validation yields high-recall detection without training data; the Presidio gap is almost entirely these categories. (3) Privacy and utility are carried by different components and must be measured separately; a single aggregate score would hide that the shield and the statistical recognizer's candidate set determine how much clinical information survives, while the decision step (VQC or fixed threshold) merely executes those choices (Sec. V-F). (4) Negative results are results: a quantum step at parity with a threshold is a finding about the task and belongs in the paper. (5) End-to-end metrics do not certify components: a recognizer with an effectively untrained classification head ran through every evaluation with PR $= 1.0$ and was exposed only by the per-detector ablation; component-level checks (loading warnings, control inputs, single-detector ablations) belong in the evaluation protocol.

Research gaps. (1) Detection of names in free Brazilian clinical narrative, outside structured headers, is the weakest point of this system (recall 0.375 on inline first names) and needs a properly trained clinical NER model. (2) Brazilian Portuguese de-identification studies use different corpora [5], [6]; the public part of AnonyMED-BR [6] is a step toward a shared benchmark, but it does not annotate clinical terms to preserve, so a benchmark that scores privacy and utility jointly is still missing. (3) Clinical preservation lacks a semantic metric validated against clinicians. (4) Operational criteria linking the LGPD notion of reasonable re-identification effort (Art. 12) to measurable acceptance thresholds for anonymizers remain underdeveloped. (5) Whether quantum or hybrid components add value for this task at larger scale remains open; with two simulated qubits, they do not.

VII. Conclusion

H-Q-MAS leaks none of the 148 PII spans of a 50-text, eight-specialty synthetic gold standard whose names are mostly header-formatted (59 of 72 names; 0/148 leaks, 95% lower bound ≥ 0.98) and retains 0.6979 of annotated clinical terms; on inline first names in free text, however, it redacts only 18 of 48 (recall 0.375). The per-detector ablation traces privacy to the structured layer and preservation to the clinical-utility shield; spaCy is a fallback that catches 10 name spans and causes most of the clinical loss; the BERTimbau+LoRA component, as deployed, is defective and detects nothing; and the quantum consensus is equivalent to a fixed threshold. Under the same metrics, Presidio with Brazilian patterns leaks 21 of 148 spans. The most important next steps are a semantic preservation metric validated by clinicians, a properly trained statistical fallback, and an evaluation on real, multiply annotated records with free-text names.

Acknowledgment

Generative AI tools were used in preparing this paper: Claude Code (Anthropic) running the Claude Fable 5 and Claude Fable 5.1 models. They were used to edit and restructure text in all sections, draft the architecture diagram (Fig. 1) in TikZ, write evaluation and audit scripts that recomputed the reported metrics, ablations, Presidio baseline, and inline-name measurement from the authors' code and data, and check the manuscript for internal consistency and reference accuracy. No annotations, results, or references were generated by AI; all reported values were verified against the evaluation artifacts. The authors take full responsibility for the content of this paper.

References

[1] Brasil, "Lei n. 13.709, de 14 de agosto de 2018 — lei geral de proteção de dados pessoais (LGPD)," Diário Oficial da União, Brasília, DF, 2018.

[2] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *arXiv preprint arXiv:1606.03475*, 2016. [Online]. Available: https://arxiv.org/abs/1606.03475

[3] M. d. S. Fantonelli *et al.*, "Organization and management of sensitive personal health data in electronic systems in countries with implemented data protection laws, lessons to Brazil: A brief systematic review," *Computer Law & Security Review*, vol. 51, p. 105872, 2023.

[4] A. Stubbs and Ö. Uzuner, "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus," *Journal of Biomedical Informatics*, vol. 58, pp. S20–S29, 2015.

[5] C. B. Prado, Y. B. Gumiel, E. T. R. Schneider, L. M. M. Cintho, J. V. A. de Souza, L. E. S. e. Oliveira, E. C. Paraiso, M. S. Rebelo, M. A. Gutierrez, F. A. Pires, J. E. Krieger, and C. Moro, "De-identification challenges in real-world Portuguese clinical texts," in *IX Latin American Congress on Biomedical Engineering and XXVIII Brazilian Congress on Biomedical Engineering (CLAIB CBEB 2022)*, ser. IFMBE Proceedings. Cham: Springer, 2024, vol. 100, pp. 584–590.

[6] M. Schiezaro, G. Rosa, B. A. G. Campos, and H. Pedrini, "Guardians of the data: NER and LLMs for effective medical record anonymization in Brazilian Portuguese," *Frontiers in Public Health*, vol. 13, p. 1717303, 2026.

[7] V. A. de Almeida, S. S. da Silva, J. Chire, N. H. Borges, L. Vicenzi, H. Kociolek, S. M. d. C. Rocha, F. N. Gomes, J. C. Ferreira, O. Marques, and L. E. S. e. Oliveira, "Clinical named entity recognition in the Portuguese language: A benchmark of modern BERT models and LLMs," *arXiv preprint arXiv:2603.26510*, 2026. [Online]. Available: https://arxiv.org/abs/2603.26510

[8] L. E. S. e. Oliveira, A. C. Peters, A. M. P. da Silva, C. P. Gebeluca, Y. B. Gumiel, L. M. M. Cintho, D. R. Carvalho, S. Al Hasan, and C. M. C. Moro, "SemClinBr – a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks," *Journal of Biomedical Semantics*, vol. 13, no. 1, 2022.

[9] E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e. Oliveira, J. Copara, Y. B. Gumiel, L. F. A. d. Oliveira, E. C. Paraiso, D. Teodoro, and C. M. C. M. Barra, "BioBERTpt – a Portuguese neural language model for clinical named entity recognition," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 2020, pp. 65–72.

[10] Z. Shi, G. Wan, W. Huang, G. Zhang, J. Shao, M. Ye, and C. Yang, "Privacy-enhancing paradigms within federated multi-agent systems," *arXiv preprint arXiv:2503.08175*, 2025. [Online]. Available: https://arxiv.org/abs/2503.08175

[11] S. S. Vaidya, G. Palani, S. Ramesh, V. Balasubramanian, M. Selvam, G. Srinivasaraja, and G. Krishnamurthi, "MedPAO: A protocol-driven agent for structuring medical reports," *arXiv preprint arXiv:2510.04623*, 2025. [Online]. Available: https://arxiv.org/abs/2510.04623

[12] X. Han, X. Gao, X. Qu, and Z. Yu, "Multi-agent medical decision consensus matrix system: An intelligent collaborative framework for oncology MDT consultations," *arXiv preprint arXiv:2512.14321*, 2025. [Online]. Available: https://arxiv.org/abs/2512.14321

[13] A. K. M. Masum, N. Mahmud, M. H. Najafi, and S. Aygun, "A hybrid classical-quantum fine tuned BERT for text classification," *arXiv preprint arXiv:2511.17677*, 2025. [Online]. Available: https://arxiv.org/abs/2511.17677

[14] F. Nausheen, K. Ahmed, M. I. Khan, and F. Riaz, "Quantum natural language processing: A comprehensive review of models, methods, and applications," *arXiv preprint arXiv:2504.09909*, 2025. [Online]. Available: https://arxiv.org/abs/2504.09909

[15] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *Intelligent Systems (BRACIS 2020)*, ser. Lecture Notes in Computer Science, vol. 12319. Springer, 2020, pp. 403–417.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[17] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," 2020, software; version 3.8.14 with model `pt_core_news_lg` 3.8.0 used in this work.

[18] V. Bergholm *et al.*, "PennyLane: Automatic differentiation of hybrid quantum-classical computations," *arXiv preprint arXiv:1811.04968*, 2018. [Online]. Available: https://arxiv.org/abs/1811.04968

[19] D. A. Meyer and N. R. Wallach, "Global entanglement in multiparticle systems," *Journal of Mathematical Physics*, vol. 43, no. 9, pp. 4273–4278, 2002.

Hypoglycemia Prediction in Type 1 Diabetes: A Retrospective Multidimensional Evaluation of Accuracy, False Alarm, and Computational Cost

1st Mario Cesar Freire Dias Filho
GREat – PPGETI
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
mariocesarfreire@alu.ufc.br

2nd Felipe Jonathan B. de Oliveira
GREat – PPGETI
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
felipejonathan@alu.ufc.br

3rd Vinicius Costa Secundino
GREat – PPGETI
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
viniciussec@alu.ufc.br

4th Rodrigo de Azeredo Siqueira
Universidade Iguaçu
Nova Iguaçu, Brazil
rodrigoendocrinologista@gmail.com

5th Danielo G. Gomes
GREat – PPGETI
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
danielo@ufc.br

Abstract*—Type 1 diabetes mellitus exposes patients to hypoglycemia, i.e. blood glucose at or below 70 mg/dL, a risk event that is especially dangerous overnight, when symptom perception is suppressed. Machine-learning alert systems built on continuous glucose monitoring (CGM) data can anticipate these events, but their viability depends on more than statistical accuracy: a high false-alarm rate erodes patient trust and high computational cost can make a model unusable on resource-constrained diabetes devices, and both are reported less consistently than discrimination metrics. This paper presents a retrospective multidimensional evaluation of five machine-learning models (Logistic Regression, Random Forest, XGBoost, LSTM, and GRU) for predicting whether hypoglycemia will occur within the next 30 minutes, trained on four public type 1 diabetes datasets (79 participants). Beyond AUPRC, AUC-ROC, and F1, we report the daily false-alarm rate and the training and inference cost as comparison axes, alongside a permutation-importance analysis of each modality and two generalization tests. Random Forest and XGBoost reached the highest AUPRC (0.887 for both, with a paired per-patient bootstrap interval crossing zero), and XGBoost trained about six times faster and produced a model hundreds of times smaller on disk. Permutation importance and ablation indicated that glucose carried most of the predictive information used by these models. Generalization to unseen patients and devices preserved risk-ranking quality while degrading the operating point, which motivates evaluating threshold recalibration per target population or device. All metrics are sample-level and all cost figures come from a single CPU environment; event-level sensitivity, time-to-warning, alarm duration, measured energy, and prospective validation are outside the scope of this benchmark and are stated as such.

***Index Terms*—hypoglycemia, continuous glucose monitoring, machine learning, type 1 diabetes, false-alarm rate, computational cost, AUPRC**

I. Introduction

Type 1 diabetes mellitus (T1D) results from autoimmune destruction of the pancreatic beta cells, eliminates endogenous insulin production, and makes the patient dependent on life-long exogenous replacement. Among the hundreds of millions of adults living with diabetes worldwide [1], [2], those with T1D face a risk tied directly to the treatment that keeps them alive: insulin therapy, although necessary for glycemic control, exposes patients to hypoglycemic episodes, clinically defined as blood glucose at or below 70 mg/dL. Unlike the non-diabetic organism, which relies on compensatory endocrine mechanisms, the T1D patient is vulnerable to abrupt, often asymptomatic glucose drops. This is a risk that peaks overnight during sleep, when symptom perception is suppressed and immediate intervention is impaired [3].

Machine learning applied to multimodal CGM time series has become the natural response to this risk, and the literature has advanced steadily in predictive accuracy. However, an alert system's viability depends on more than detecting events with sufficient sensitivity. A model that generates frequent false alarms undermines patient adherence, a phenomenon documented as alarm fatigue [4], [5], and the computational cost of inference determines whether a model can run on resource-constrained diabetes devices or companion applications. Both are usually treated as secondary rather than as design criteria; Section II-C identifies the specific gaps addressed here.

This paper compares five models predicting whether hypoglycemia will occur within the next 30 minutes, evaluated jointly across all three dimensions on four public T1D datasets. The families span the range of representational capacity and computational cost relevant to the task: Logistic Regression as a linear baseline, Random Forest [22] and XGBoost [23] as tabular ensembles, and LSTM [24] and GRU [25] recurrent networks. The contributions are:

- **Multidimensional evaluation:** beyond AUPRC, AUC-ROC, and F1, we include the daily false-alarm rate and the training and inference cost (time, memory, model size) as comparison metrics;

- **Accuracy–alarm–cost trade-off:** we quantify how models with statistically indistinguishable AUPRC differ in false alarms per day and in cost in the evaluated CPU environment;
- **Leakage-safe protocol:** a strict per-participant chronological split, without shuffling, plus generalization to unseen patients (patient-held-out) and to an unseen dataset or device (leave-one-dataset-out), with the residual boundary overlap quantified in Section III-A;
- **Multimodal ablation:** the incremental contribution of each modality (CGM, insulin, carbohydrate, activity).

The contribution is the benchmark itself, under one shared protocol. The study is retrospective and offline: no alert was presented to a patient and no treatment decision was taken from a prediction, so nothing here supports a claim of clinical utility. Section V-A states the boundary of every claim.

II. Theoretical Background

A. The Datasets

The experiments use four public type 1 diabetes datasets, 79 participants in total, harmonized to a common 5-minute grid per participant. Table I summarizes participant count, temporal coverage, native CGM sampling interval, and glucose sensor of each source. The heterogeneity of devices, populations, and durations allows assessing both average performance and generalization to unseen sources. The prediction horizon (30 min), the event definition, and the missing-data rule are common to all four sources and are stated in Section III-C.

OhioT1DM [7] is a classic reference for glucose prediction in T1D; the 2020 version covers 12 adults on insulin-pump therapy with Medtronic Enlite monitoring and is multimodal (5-min interstitial glucose, basal and bolus insulin, self-reported carbohydrate, exercise, sleep and stress events, and physiological-band signals). HUPA-UCM [19] covers 25 people with T1D (52% women, mean age 39 years, mean HbA1c 7.4%, 56% on continuous subcutaneous infusion), with 15-min FreeStyle Libre 2 glucose resampled to the 5-min grid and steps, calories, heart rate and sleep from Fitbit Ionic smartwatches. AZT1D [20] covers 25 individuals on automated insulin delivery (Tandem t:slim X2 in Control-IQ), with more than 320k Dexcom G6 Pro readings and granular bolus details (total dose, type, correction share) that are rare in public datasets. T1D-UOM [21] covers 17 participants aged 23 to 70, with glucose from multiple CGM platforms, activity from a Garmin Forerunner 45, and nutrition decomposed into carbohydrate, protein, fat, and fiber; it also has the highest fraction of missing readings (about 35%), which makes it the main data-quality challenge.

Table I
Type 1 diabetes datasets used (79 participants; unified 5-min analysis grid).

Dataset	Part.	Coverage	CGM sensor	Δt[a]
HUPA-UCM [19]	25	$\geq$ 2 weeks	FreeStyle Libre 2	15
AZT1D [20]	25	6–8 weeks	Dexcom G6 Pro	5
OhioT1DM [7]	12	8 weeks	Medtronic Enlite	5
T1D-UOM [21]	17	12 weeks	multiple[b]	5–15

[a] Native sampling interval of the source device (min); all sources are resampled to the common 5-min grid, so a 15-min source contributes one observed reading per three cells. [b] LibreView, Dexcom, Medtronic.

B. Related Work

Rather than surveying individual studies in sequence, this review is organized along the dimensions that define the present comparison: the prediction target, the input modalities, the validation regime, and the metrics reported.

Target formulation. A large share of the CGM literature addresses continuous forecasting of the glucose trajectory [8], including foundation and multimodal language models [9] and horizon-aware sequence-to-sequence Transformers with knowledge distillation [10]. A second line frames the problem as binary event classification over a future window, the formulation adopted here [11]–[14]. The two framings are not interchangeable, since event classification is evaluated under strong class imbalance and admits an explicit operating point.

Input modalities. Within the classification line, Fleischer et al. [12] use the CGM signal alone, whereas Dave et al. [11] add insulin and carbohydrate and report gains mainly at horizons longer than 30 minutes. This divergence motivates the modality ablation of Section IV-D.

Validation regime. Bergmeir and Benítez [15] document the leakage that arises in time-series cross-validation when temporal order is not preserved, which motivates the strictly temporal, per-participant split used here. Among the classification studies, generalization is assessed unevenly: Shao et al. [14] validate an LSTM across distinct populations, while the remaining studies, and related Brazilian work on biomedical time series [17], [18], report performance within a single cohort.

Reported metrics. Results in this line concentrate on discrimination (sensitivity, specificity, AUC-ROC, AUPRC). Operational quantities appear less consistently: alarm fatigue is documented as a clinical problem [4], [5], and closed-loop work [26] frames glycemic prediction as a component that must run within device computational constraints, but false alarms per day and inference cost are seldom tabulated alongside discrimination for the same models.

C. Gaps and Positioning

Two gaps emerge from the literature just reviewed. First, among the classification-framed studies [11]–[14], each reports strong discrimination, but none evaluates a linear model, a tabular ensemble, and a recurrent network under one shared protocol, so the accuracy–cost frontier across model families remains uncharacterized. Second, of these same studies only Shao et al. [14] test generalization to unseen patients, and none reports the false-alarm rate or inference cost alongside discrimination under a leakage-safe temporal protocol. This paper targets both gaps, using the datasets and protocol of Section III.

III. Methodology

A. Study Data and Experimental Design

The study data are the four public T1D datasets characterized in Section II-A (Table I).

Each participant is split *chronologically* into 60% train, 20% validation, and 20% test, without shuffling, so that the patient's future never informs the past used in training. Three evaluation scenarios are considered: (i) within-patient temporal evaluation, which measures prediction of the patient's own future; (ii) patient-held-out, in which whole patients are separated across train, validation, and test, assessing generalization to unseen individuals; and (iii) leave-one-dataset-out, in which the model trains on three datasets and is tested on the fourth, assessing transfer to an unseen device or population. Wherever generalization is claimed (scenarios ii and iii), the partitions are participant-disjoint, so no window, episode, or patient-specific statistic is shared across partitions.

Windows are built by sliding over the 5-minute grid and therefore overlap *within* a partition by construction. Under scenario (i) each participant has two contiguous cut points, and only windows spanning a cut can share readings across partitions: a window at t uses $[t-60,\, t+30]$ min, so the affected region is 90 min wide, at most 18 windows per cut and 36 per participant, i.e. about 2 800 windows across the 79 participants, against a training partition of 550 245 rows. The same bound limits episode straddling to two per participant, and only when a cut falls inside an episode; in scenarios (ii) and (iii) none is possible. This residual overlap is reported rather than removed: a 90-min embargo at each cut would eliminate it and is the recommended form of the protocol.

Cohorts combine datasets and split type. We write $\mathcal{D}_{\text{tot}}$ for the four datasets combined (79 participants) and $\mathcal{D}_{\text{uom}}$, $\mathcal{D}_{\text{azt}}$, $\mathcal{D}_{\text{ohio}}$, $\mathcal{D}_{\text{hupa}}$ for T1D-UOM, AZT1D, OhioT1DM, and HUPA-UCM alone. Cohort $\mathcal{D}_{\text{tot}}$ under temporal split is the main comparison configuration.

B. Data Pipeline

The four datasets are processed by a four-layer data lake (Fig. 1). The raw layer preserves the original files; bronze converts the heterogeneous formats (per-patient CSV, XML, per-domain CSV) to Parquet, normalizing variant headers; silver harmonizes schema, units, and timestamps, standardizing glucose to mg/dL (converting mmol/L by 18.0156 in T1D-UOM), carbohydrate to grams, and basal insulin to U/5 min, and materializes the unified 5-minute grid per participant; and gold creates the label, the tabular and sequential features, and the train-ready splits. Every transformation is deterministic and row-local: no statistic estimated from data enters this stage, so the pipeline itself cannot transport information across partitions.

C. Target Definition and Gap-Aware Coverage

Because the pipeline in III-B produces a unified 5-minute grid but does not resolve missing readings, the target and coverage rule are defined jointly, so that a future window is never built from a mostly-absent stretch of data. The task is

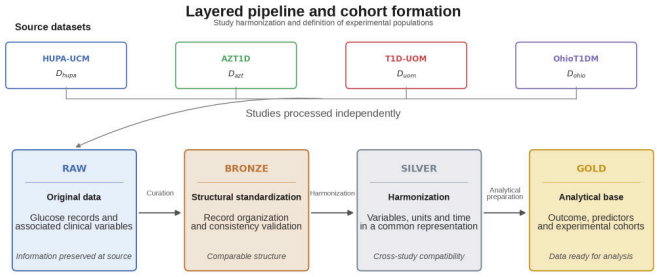


Figure 1. Layered data-lake architecture: harmonization of the four datasets into a unified 5-min grid and generation of label, features, and splits.

framed as binary classification. For each time point t on the 5-minute grid, the target is positive if the minimum glucose reading in the following 30-minute window is at most 70 mg/dL:

$$y_t = \begin{cases} 1, & \text{if } \min\limits_{1\le k\le 6} g(t+5k) \le 70 \text{ mg/dL}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $g(\cdot)$ denotes interstitial glucose and time is measured in minutes. The label therefore refers to occurrence at any point within the next 30 minutes, not to the value at $t+30$ specifically. The current glucose value $g(t)$ is excluded from the outcome window, so it is not part of the target definition; samples in which the patient is already hypoglycemic at t remain in the aggregate evaluation and are examined separately in the glucose-stratified analysis of Section IV-E.

Because sparse datasets (T1D-UOM, with about 35% null readings) would make a nominal 30-minute window span real hours, windows are computed over a gap-free grid, and a row becomes a labeled example only if it has sufficient observed coverage: at least 8 of the 12 history cells (60 min) and 3 of the 6 future cells (30 min) non-null. The minimum in (1) is taken over the observed cells only, so it may rest on as few as three readings; rows below the coverage requirement are discarded rather than imputed. The rule reduces, but does not eliminate, the possibility that an unobserved hypoglycemic reading turns a true positive into a negative label, and the residual risk differs across datasets according to their coverage. Both the horizon and the threshold follow the international continuous-monitoring consensus definition of a hypoglycemic level [6]; the rarity of these events in the continuous stream of readings imposes the class imbalance that motivates the metric choices of Section III-F.

D. Feature Engineering

With labels defined, three feature sets translate the same underlying grid into model-ready inputs at different levels of complexity. The full tabular set (17 features) gathers 60-minute glucose statistics (current, mean, min, max, standard deviation, and 5- to 60-minute deltas), cyclic time-of-day encoding, and aggregates of insulin (bolus and basal), carbohydrate, steps, and heart rate. The reduced glucose+time set (11 features) drops the insulin, carbohydrate, and activity modalities, keeping only glucose and time of day; Section IV-D reports the

AUPRC change associated with this reduction. The sequential set (9 channels) feeds the recurrent networks with the raw 12-step (60 min) series: glucose, basal, bolus, carbohydrate, steps, calories, heart rate, and the two cyclic time-of-day components. Every feature is a backward-looking function of the same participant's own history at or before t; no forward-looking, cross-participant, or full-dataset statistic is used in feature construction.

E. Evaluated Models

Five classifiers from distinct families are compared. Logistic Regression (L2 regularization, balanced classes, standardized features) is the linear baseline. Random Forest (300 trees, minimum 5 samples per leaf) and XGBoost (500 estimators, depth 4, learning rate 0.05, subsampling 0.8, `aucpr` evaluation metric, and positive-class weight computed as the negatives-to-positives ratio, near 17.6 in cohort $\mathcal{D}_{\text{tot}}$) represent the tabular ensembles. The LSTM and GRU networks share the architecture Input$(12, 9)$ followed by a 64-unit recurrent layer, 0.2 dropout, a 32-unit dense layer with ReLU activation, and a sigmoid output, trained for up to 50 epochs with early stopping on validation AUPRC. Hyperparameters were fixed a priori at these commonly used values rather than tuned by search, so no search ever observed the test partition and the comparison reflects default-configuration behavior; the class-weight ratio and the standardization scaler are estimated on the training partition only, early stopping uses validation only, and the decision threshold is selected on validation only (Section III-F). All models run on CPU in this environment. XGBoost is adopted as the reference model in the per-cohort analyses.

F. Threshold Selection and Metrics

Because hypoglycemia is the minority class (3% to 9.5% of rows depending on the dataset), accuracy is uninformative, and AUPRC, threshold-independent and more informative than AUC-ROC under strong imbalance [16], is the primary metric. The operating point is decoupled from the model: AUPRC measures risk-ranking quality, and a rule converts that ranking into a firing decision. The primary rule fixes a minimum recall of 0.80 on validation; an alternative point maximizing validation F1 is a secondary analysis (Section IV-G). Three configurations are kept distinct: the *primary comparison* (corrected pipeline, full test set, minimum-recall rule), the *complete-glucose subset* of Section IV-E, and the *alternative F1 point* of Section IV-G. Unless stated otherwise, results refer to the primary comparison.

Denominators. All reported recall, precision, and F1 are *sample-level*: the unit is one 5-minute grid row that satisfies the coverage rule of Section III-C, not one hypoglycemic episode and not one alarm. A false alarm is a test row predicted positive whose label is negative; the daily rate divides each patient's false-positive count by that patient's *observed* test days (labeled test rows divided by 288, the number of 5-minute cells in a day) and averages across patients, so unequal test durations contribute equally. We state this convention because a calendar-duration denominator would report systematically lower rates for the sparser cohorts, by up to their missing-data fraction. Event-level sensitivity, median time-to-warning, and alarm duration use different denominators and are not estimated here; Section V-A discusses what that omission costs.

AUPRC uncertainty is quantified by nonparametric bootstrap resampling at the *patient* level, preserving the intra-patient correlation that row-level resampling would ignore; two-model comparison uses the same scheme in paired form. Intervals are reported for AUPRC only; the operating-point quantities are point estimates. To separate real task difficulty, we also report AUPRC restricted to samples that are euglycemic at prediction time (current glucose ≥ 80 mg/dL), a regime in which the future event is not trivially anticipable from the current value, and we evaluate two learning-free comparators: ranking risk by current glucose ($-g_t$) and by 30-minute glucose projected by linear extrapolation ($-\hat{g}_{t+30} = -(g_t + 6\,\Delta g_5)$). Guarantees against temporal leakage (past-only features, future-only target, scaler fit on train only, threshold chosen on validation only) are verified by automated tests.

IV. Experiments

A. Experimental Setup

The primary comparison uses cohort $\mathcal{D}_{\text{tot}}$ under temporal split, with the operating point fixed by the minimum-recall-0.80 rule on validation. Cost measurements use the same cohort, with 550 245 tabular training rows and 564 951 sequential windows, on CPU: reported memory is the RAM working set and wall-clock times are single-run measurements, without repetition or dispersion (Section IV-C). Permutation importance shuffles each feature on the test set and measures the mean AUPRC drop over five repetitions.

B. Comparative Results

In the primary comparison on cohort $\mathcal{D}_{\text{tot}}$ (Table II), Random Forest and XGBoost reach the highest AUPRC (both 0.887), with Logistic Regression an AUPRC approximately 0.08 lower. In the *paired* per-patient bootstrap, the difference ΔAUPRC(RF$-$XGB) is close to zero, with a 95% confidence interval that crosses zero ($[-0.0025, +0.0053]$), so the two are statistically indistinguishable on this criterion; their similar AUPRC makes computational cost an important additional selection criterion (Section IV-C). The recurrent networks (LSTM and GRU) sit just behind the tree ensembles (0.874 and 0.878). The AUC-ROC is high and nearly saturated for all models (between 0.982 and 0.986), which supports the choice of AUPRC, more sensitive to the rare class, as the decision metric.

Two cautions apply to reading Table II as a ranking. First, the marginal intervals are wide (about ± 0.07), since there are only about 768 independent episodes across 79 patients rather than 550k independent rows; being computed independently per model, their overlap is uninformative about a within-patient difference, which the paired test estimates directly.

Table II
30-MIN HYPOGLYCEMIA CLASSIFICATION IN COHORT $\mathcal{D}_{\text{TOT}}$ (TEMPORAL SPLIT, GLUCOSE+TIME SET, OPERATING POINT AT RECALL ≥ 0.80). ALL QUANTITIES ARE SAMPLE-LEVEL (ONE 5-MIN ROW PER UNIT).

Model	AUPRC	AUC-ROC	Recall	Prec.	F1	FA/day
Log. Regression	0.804	0.982	0.807	0.738	0.771	4.05
Random Forest	**0.887**	0.985	0.798	0.802	0.800	2.80
XGBoost	**0.887**	0.986	0.801	0.796	0.798	2.91
LSTM	0.874	0.983	0.801	0.762	0.781	3.55
GRU	0.878	0.985	0.797	0.778	0.788	3.22

79 test patients, prevalence 5.62% of labeled rows. AUPRC 95% CI (patient-clustered bootstrap): LogReg [0.715, 0.859], RF [0.816, 0.922], XGB [0.814, 0.922], LSTM [0.798, 0.908], GRU [0.793, 0.913]. Paired bootstrap ΔAUPRC(RF$-$XGB) $\approx$ 0, 95% CI $[-0.0025, +0.0053]$; other pairs untested. Recall, precision, F1 and FA/day are point estimates. LSTM and GRU use the 9-channel sequential set.

Second, the paired test was applied only to the RF–XGB pair, the one on which a selection decision depends. The remaining orderings, in particular GRU above LSTM and both below the ensembles, are descriptive: they were not tested pairwise, and differences of that magnitude lie well inside the marginal intervals. No multiplicity correction is applied, since only the two pre-specified paired comparisons of this paper are reported.

C. Computational Cost Analysis

Table III and Figure 2 report cost quantities that discrimination metrics alone do not capture. With AUPRC statistically indistinguishable from Random Forest, XGBoost trained about six times faster (8.0 s versus 46.8 s), showed approximately 35-fold lower single-window latency (1.32 ms versus 46.28 ms), and produced a serialized model about 350 times smaller on disk (0.79 MB versus 274.65 MB). Relative to XGBoost, LSTM and GRU required approximately 64 and 97 times more training time, respectively, for a lower AUPRC; relative to Random Forest the corresponding training-time factors are about 11 and 17. Logistic Regression is the cheapest model but reaches an AUPRC approximately 0.08 lower. The two timing quantities are not directly comparable: single-window latency measures one prediction at a time, as in 5-minute real-time operation, whereas throughput measures batched prediction, where vectorization amortizes the fixed per-call overhead. Because Random Forest and XGBoost had indistinguishable AUPRC, computational cost became the main criterion for choosing between them in the evaluated CPU environment.

Table III deliberately reports no energy figure. An energy value obtained by multiplying training time by a fixed nominal power, as is sometimes done, is a linear transform of the first column and therefore carries no information beyond it, while ignoring idle draw, dynamic frequency scaling, and the memory subsystem. A meaningful energy comparison, particularly one supporting a claim about embedded deployment, requires wall-plug or on-board power measurement on the target hardware under a stated duty cycle, which this study does not provide.

Table III
TRAINING AND INFERENCE COST (COHORT $\mathcal{D}_{\text{TOT}}$, CPU, SINGLE RUN).

Model	Train (s)	Latency (ms/window)	Throughput (10^3 rows/s)	Train RAM (MB)	Size (MB)
Log. Regression	1.1	1.099	7 063	116	0.00
Random Forest	46.8	46.279	239	581	274.65
XGBoost	8.0	1.323	1 176	47	0.79
LSTM	514.9	50.440	15.5	429	0.28
GRU	779.0	51.243	16.6	259	0.23

Latency: one time point at a time (real-time use every 5 min). Throughput: batched prediction; the two are not comparable, since batching amortizes the fixed per-call overhead. Single-run measurements, no dispersion. Peak RAM reflects the full implementation, not the serialized model. AUPRC for the same configurations in Table II.

Table IV
PERMUTATION IMPORTANCE (MEAN AUPRC DROP) OF THE MOST RELEVANT TABULAR FEATURES.

Feature	LogReg	RF	XGBoost
glucose_current	0.3631	0.4843	0.6976
glucose_min_60min	0.6306	0.1560	0.0423
glucose_mean_60min	0.1423	0.0481	0.0193
glucose_max_60min	0.1476	0.0239	0.0077
glucose_delta_5min	0.0259	0.0524	0.0663
hour_cos	-0.0004	0.0008	0.0025

Insulin, carbohydrate and activity show drops near zero (omitted). Permutation importances are not additive; under strong correlation among the glucose features the attribution distributes unstably among them, which explains the ranking inversion across models, so the stable reading concerns the glucose block as a whole, not the internal order of its components.

D. Multimodal Ablation and Feature Importance

Permutation importance (Table IV) is concentrated in the glucose features: current glucose and the 60-minute minimum account for the largest AUPRC drops, while insulin, carbohydrate and activity features drop near zero. In the sequential models, permuting the glucose channel reduced AUPRC by about 0.83, whereas permuting all other channels reduced it by less than 0.03 in total. The ablation agrees in relative terms: reducing the classifier from 17 to 11 features (glucose+time), or to 9 (glucose-only), leaves AUPRC nearly unchanged (0.8645, 0.8636, 0.8628), a maximum absolute difference of 0.0017, stable across reruns. These absolute values come from the pre-correction pipeline (the *Original* column of Table VII), so the ablation characterizes *relative* invariance to modality removal, not the final performance level. The remaining analyses therefore adopt the glucose+time set, which is simpler and cheaper without a material AUPRC reduction.

This observation is specific to the 30-minute window, to binary ranking classification, and to these four cohorts, three of which are dominated by pump or automated-insulin-delivery therapy, with regular basal profiles and sparse self-reported carbohydrate and activity logs. An under-recorded modality cannot show importance, so a near-zero permutation drop is evidence about *these annotations*, not about the physiology. The literature reports gains from insulin and carbohydrate mainly at longer horizons [11]; the limited contribution observed here delimits the short-term regime under sparse annotation and does not rule out multimodal value in general.

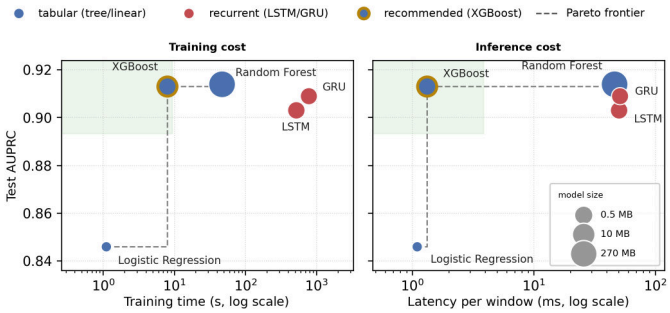


Figure 2. Computational cost versus AUPRC in cohort $\mathcal{D}_{\text{tot}}$ (CPU, single run; cf. Table II). Left: training cost; right: inference latency per window. Marker area encodes on-disk size, color encodes model family, and the dashed line is the Pareto frontier. The upper-left corner combines higher AUPRC with lower cost; Logistic Regression and XGBoost are the two non-dominated points in both panels.

Table V
COMPLETE-GLUCOSE SUBSET ANALYSIS: AUPRC OF XGBOOST VERSUS LEARNING-FREE RULES, STRATIFIED BY GLUCOSE AT PREDICTION TIME (COHORT $\mathcal{D}_{\text{TOT}}$, TEST).

Curr. glucose	n	prev.%	XGBoost	$-g_t$	$-\hat{g}_{t+30}$
all	183 419	5.70	**0.895**	0.842	0.522
≥ 80 (eugl.)	170 159	1.17	**0.321**	0.118	0.159
≥ 100	149 649	0.33	**0.135**	0.015	0.064
≥ 120	118 013	0.12	**0.056**	0.003	0.029

$-g_t$: risk ranked by current glucose. $-\hat{g}_{t+30}$: by glucose projected via linear extrapolation of the 5-min trend. Prevalence falls as the band rises, making the task progressively harder. All rows are restricted to the subset with materialized glucose; this restriction, not a change of model or pipeline, accounts for 0.895 in the *all* row versus 0.887 in Table II.

Table VI
XGBOOST GENERALIZATION (GLUCOSE+TIME): WITHIN-PATIENT TEMPORAL EVALUATION VERSUS UNSEEN PATIENTS AND DATASETS (COHORT $\mathcal{D}_{\text{TOT}}$, OPERATING POINT AT RECALL ≥ 0.80).

Scenario	AUPRC	Recall	Prec.	FA/day
Temporal (baseline)	0.887	0.799	0.798	2.87
Unseen patients	0.873	0.811	0.751	3.67
Unseen dataset	0.869	0.758	0.802	2.76

AUPRC 95% CI (per-patient bootstrap): Temporal [0.814, 0.920]; Unseen patients [0.729, 0.941] (15 test patients); Unseen dataset [0.812, 0.906]. Unseen dataset: train on three, test on the fourth (OhioT1DM holdout).

E. Real Task Difficulty and Trivial Baseline

This secondary analysis is restricted to test rows with materialized glucose at prediction time (183 419 rows), which is why the aggregate XGBoost AUPRC here is 0.895 rather than the 0.887 of Table II. Within this subset, Table V stratifies XGBoost by glucose at prediction time against two learning-free rules. In aggregate, ranking by current glucose alone already reaches 0.842 versus 0.895 for XGBoost, an aggregate influenced by samples already near the threshold, where the current value largely determines the outcome. Restricted to euglycemic samples ($\geq$ 80 mg/dL), where the drop has not begun, XGBoost falls to 0.321 and prevalence from 5.70% to 1.17%; there XGBoost exceeded the best learning-free rule (0.159) by $\Delta\text{AUPRC} = +0.162$ (95% CI $[+0.139, +0.189]$, paired per-patient bootstrap), against an aggregate margin of only 0.053. The gap between the two margins indicates that aggregate AUPRC is a poor guide to the incremental value of the learned model.

F. Generalization Analysis

Table VI contrasts within-patient temporal evaluation with generalization to unseen patients and datasets, both participant-disjoint by construction. For unseen patients, AUPRC falls from 0.887 to 0.873 while, at the same target recall, precision falls from 0.798 to 0.751 and false alarms rise from 2.87 to 3.67 per day. For a whole unseen dataset (OhioT1DM holdout), AUPRC falls comparably, to 0.869, but the operating point shifts differently: precision and false alarms are preserved, while the fixed threshold no longer sustains the target recall, which falls from 0.80 to 0.76. Risk ranking therefore transferred better than the selected operating point [14], and that point moved in opposite directions under the two shifts: toward more false alarms for new patients, toward reduced sensitivity for a new device or population. The patient-held-out interval is wide (15 test patients), so these comparisons are indicative rather than conclusive, and they motivate evaluating a threshold recalibration that this study does not test.

G. Alternative Operating Point and Data Quality

Three interventions were evaluated on XGBoost as secondary analyses: switching the threshold rule from minimum recall to F1, reducing the feature set to glucose+time, and data-quality correction (physiological clipping to 40–400 mg/dL and gap-aware windows). In Table VII the combined change in threshold and feature set increased precision while leaving aggregate AUPRC nearly unchanged; introduced in the same step, their individual effects cannot be separated. The subsequent data-quality correction was the only step in which AUPRC itself increased in every cohort, and it produced the largest AUPRC improvement among the three. The largest change appears in T1D-UOM, the cohort with the most missing data: precision rises from 0.37 to 0.78 and false alarms fall from about 18 to 1.6 per day.

V. DISCUSSION

Why the model families converged. The five architectures span a wide range of representational capacity, yet their AUPRC values fall within roughly 0.08 of each other and the two tree ensembles are statistically indistinguishable. A

Table VII
XGBOOST BY INTERVENTION (AUPRC / PRECISION / FALSE ALARMS PER DAY, ON TEST). COLUMNS ARE CUMULATIVE: EACH ADDS AN INTERVENTION TO THE PREVIOUS ONE. THE THIRD COLUMN CORRESPONDS TO THE ALTERNATIVE F1 OPERATING POINT ON THE CORRECTED PIPELINE AND IS A SECONDARY ANALYSIS, NOT THE PRIMARY COMPARISON OF TABLE II.

Cohort	Original	+ F1 thr., gluc+time	+ data corr.
$\mathcal{D}_{\text{tot}}$	0.864 / 0.756 / 4.30	0.864 / 0.854 / 2.08	**0.895 / 0.867 / 1.92**
$\mathcal{D}_{\text{uom}}$	0.676 / 0.368 / 17.9	0.680 / 0.656 / 4.09	**0.775 / 0.780 / 1.63**
$\mathcal{D}_{\text{azt}}$	0.750 / 0.497 / 7.46	0.757 / 0.705 / 2.58	**0.764 / 0.729 / 2.24**
$\mathcal{D}_{\text{ohio}}$	0.839 / 0.732 / 3.47	0.842 / 0.792 / 2.40	**0.848 / 0.877 / 1.15**
$\mathcal{D}_{\text{hupa}}$	0.948 / 0.952 / 1.18	0.951 / 0.906 / 2.63	**0.951 / 0.919 / 2.23**

Original: recall ≥ 0.80 rule, 17 features. The second column changes the threshold rule and the feature set simultaneously, so their effects cannot be separated. The third is evaluated on the subset with materialized glucose, hence 0.895 for $\mathcal{D}_{\text{tot}}$ rather than the 0.887 of Table II.

plausible explanation is that the informative structure of the 30-minute problem is largely contained in the recent glucose trajectory, which the tabular set already encodes through the 60-minute statistics and deltas; the sequential models then have little residual structure to extract, consistent with them matching but not exceeding the ensembles at one to two orders of magnitude more training cost. The ablation points the same way: removing the non-glucose modalities changed AUPRC by at most 0.0017.

What aggregate AUPRC does and does not measure. The aggregate metric is strongly influenced by samples in which the patient is already near the hypoglycemic threshold, where the current glucose value alone ranks risk almost as well as the trained model. The incremental contribution of learning becomes visible in the euglycemic subset, where prevalence is 1.17% and the margin over the best learning-free rule is $+0.162$ AUPRC. A study reporting only aggregate AUPRC therefore cannot distinguish a model that anticipates the onset of a drop from one that mostly reads the current value, and absolute values are not comparable across studies applying different inclusion criteria for current glucose.

Repeated events and effective sample size. Hypoglycemia arrives in episodes, not independent draws: consecutive 5-minute rows inside one descent share almost all of their history and carry nearly the same label, the test partition holds roughly 768 episodes across 79 patients, not 550k independent draws. Sample-level precision, recall and F1 therefore describe how many 5-minute cells are correctly flagged, and a model can score well by holding an alarm through a long episode it detected late, or by flagging one long false stretch instead of many short ones. Patient-clustered bootstrap and per-patient normalization of the false-alarm rate mitigate this, but neither converts a sample-level metric into an event-level one.

Sensor delay and what the horizon means. Features and labels both derive from interstitial CGM, which trails plasma glucose by several minutes through diffusion and sensor filtering [8]. The 30-minute horizon is therefore a horizon in *sensor time*: the physiological lead available to a patient is shorter by the sensor delay, and the margin depends on the device, which differs across the four datasets. Since the delay affects label and input alike, it does not bias the metrics reported here, but it compresses the actionable window any deployed version would provide.

Distribution shift. AUPRC was largely preserved under both tested shifts, new individuals and a new device or population, while the fixed operating point was not, moving in opposite directions in the two cases; ranking quality and calibration are separable properties. This does not establish that recalibration is sufficient: the experiments hold the threshold fixed and observe the consequence, without testing a recalibration procedure.

A. Limitations and Threats to Validity

Retrospective evidence only. Every result comes from stored recordings of four public datasets. No alert was displayed to a patient, no clinician acted on a prediction, and no downstream outcome, such as an averted episode, a carbohydrate intake, or a suspended basal delivery, was observed. This distinguishes retrospective prediction from decision support: the paper reports how well a model ranks risk on recorded data, and reports nothing about whether an alert derived from that ranking would be safe, timely, or acted upon. Prospective, multicenter evaluation across CGM platforms, with real patient interaction with the generated alerts, is required before any clinical-utility claim.

Sample-level metrics. As stated in Section III-F, recall, precision, F1, and false alarms per day are computed per 5-minute row. Event-level sensitivity (the fraction of hypoglycemic *episodes* flagged at least once before onset), median time-to-warning, alarm duration, and the number of *distinct* alarms per patient-day are the quantities a clinical reader needs, and they are not estimated here. Sample-level figures can overstate usefulness when episodes are long and alarms are held, so Table II should be read as a comparison among models under one common convention rather than as an estimate of clinical alarm burden. Confidence intervals are reported for AUPRC only.

Cost figures. Costs characterize one CPU environment, one implementation per model, and a single run per measurement, without repetition, dispersion or contention control, and no energy figure is reported (Section IV-C). Statements about suitability for resource-constrained diabetes devices are therefore extrapolations from relative cost ratios, which are more stable across environments than absolute values; no embedded target, quantized model or on-device duty cycle was evaluated.

Scope of the task. A single horizon (30 min) and a single event definition (70 mg/dL) are covered. Regression-based forecasting, level-2 hypoglycemia at 54 mg/dL, nocturnal-specific evaluation, and the longer horizons relevant to closed-loop delivery [26] are not, and neither the model ranking nor the modality result should be assumed to hold there. Adaptive threshold calibration is motivated by the generalization results but not implemented or validated, and the boundary overlap of Section III-A is bounded but not eliminated. Finally, data-quality correction is the largest gain *among the three interventions evaluated.*

VI. Conclusion

Across the five evaluated models, Random Forest and XGBoost achieved the highest AUPRC (0.887 for both), with no meaningful difference in the paired per-patient bootstrap. XGBoost required about six times less training time, roughly 35-fold lower single-window latency, and a serialized model about 350 times smaller than Random Forest in the evaluated CPU environment, so computational cost was the main criterion separating the two. Ablation and permutation analyses indicated that most of the predictive information used by these models came from recent glucose, with limited incremental contribution from insulin, carbohydrate and activity at the 30-minute window and under the annotation density of these cohorts. Ranking transferred relatively well to held-out patients and to a held-out dataset, although precision, recall and false-alarm rates changed at the fixed operating point, which supports evaluating threshold recalibration when deploying in a new population or device.

The contribution is the multidimensional benchmark under a shared protocol, and its evidence is retrospective and sample-level throughout. Future work follows from Section V-A: event-level reporting (episode sensitivity, time-to-warning, alarm duration, distinct alarms per patient-day, with intervals); measured energy and latency on embedded target hardware; prospective multicenter validation across CGM devices with real patient interaction with the alerts; adaptive per-patient and per-device recalibration, evaluated rather than proposed; and extension to longer horizons and to regression-based trajectory forecasting for closed-loop integration.

Data Availability and Ethics

This study uses four publicly available de-identified T1D datasets under their respective terms and collected no new human-subject data. OhioT1DM [7] is distributed under a Data Use Agreement restricting use to research, executed prior to access; HUPA-UCM [19] is on Mendeley Data under CC BY-NC-ND; AZT1D [20] is on Mendeley Data, its collection approved by the Mayo Clinic Institutional Review Board (#23-003065); and T1D-UOM [21] is on Zenodo under CC BY.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. Danielo G. Gomes acknowledges the support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (grant numbers 309014/2026-3, 311845/2022-3).

References

[1] D. J. Magliano and E. J. Boyko, Eds., *IDF Diabetes Atlas*, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025.

[2] GBD 2021 Diabetes Collaborators, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet*, vol. 402, no. 10397, pp. 203–234, 2023.

[3] B. Kulzer, G. Freckmann, R. Ziegler, *et al.*, "Nocturnal hypoglycemia in the era of continuous glucose monitoring," *J. Diabetes Sci. Technol.*, vol. 18, no. 5, pp. 1052–1060, 2024.

[4] J. P. Shivers, L. Mackey-Davis, K. Peterson, *et al.*, "'Turn it off!': diabetes device alarm fatigue considerations for the present and the future," *J. Diabetes Sci. Technol.*, vol. 7, no. 3, pp. 789–794, 2013.

[5] Y. K. Lin, D. Groat, O. Chan, *et al.*, "Alarm settings of continuous glucose monitoring systems and associations to glucose outcomes in type 1 diabetes," *J. Endocr. Soc.*, vol. 4, no. 1, art. bvz005, 2020.

[6] T. Battelino, T. Danne, R. M. Bergenstal, *et al.*, "Clinical targets for continuous glucose monitoring data interpretation: recommendations from the International Consensus on Time in Range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.

[7] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: update 2020," in *Proc. 5th Int. Workshop Knowledge Discovery in Healthcare Data (KDH@ECAI)*, vol. 2675, CEUR-WS.org, 2020, pp. 71–74.

[8] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, "Continuous glucose monitoring sensors for diabetes management: a review of technologies and applications," *Diabetes Metab. J.*, vol. 43, no. 4, pp. 383–397, 2019.

[9] J. Wolber, M. E. Samadi, J. Sellin, and A. Schuppert, "Multimodal large language models and mechanistic modeling for glucose forecasting in type 1 diabetes patients," *J. Biomed. Inform.*, vol. 172, art. 104945, 2025.

[10] X. Sun, H. Li, and X. Yu, "Future-aware blood glucose forecasting using knowledge distillation with transformer-based sequence-to-sequence models," *Sci. Rep.*, vol. 16, art. 11404, 2026.

[11] D. Dave, D. J. DeSalvo, B. Haridas, *et al.*, "Feature-based machine learning model for real-time hypoglycemia prediction," *J. Diabetes Sci. Technol.*, vol. 15, no. 4, pp. 842–855, 2021.

[12] J. Fleischer, T. K. Hansen, and S. L. Cichosz, "Hypoglycemia event prediction from CGM using ensemble learning," *Front. Clin. Diabetes Healthc.*, vol. 3, art. 1066744, 2022.

[13] S. L. Cichosz, M. H. Jensen, and S. S. Olesen, "Development and validation of a machine learning model to predict weekly risk of hypoglycemia in patients with type 1 diabetes based on continuous glucose monitoring," *Diabetes Technol. Ther.*, vol. 26, no. 7, pp. 457–466, 2024.

[14] J. Shao, Y. Pan, W.-B. Kou, *et al.*, "Generalization of a deep learning model for continuous glucose monitoring-based hypoglycemia prediction: algorithm development and validation study," *JMIR Med. Inform.*, vol. 12, art. e56909, 2024.

[15] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Inf. Sci.*, vol. 191, pp. 192–213, 2012.

[16] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, art. e0118432, 2015.

[17] M. Santos, Á. Melo, J. Santos, *et al.*, "Modelagem de séries temporais biomédicas: um estudo comparativo entre PSTA-TCN e TCN vanilla para previsão de dados de glicose de pessoas diabéticas," in *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, Brazil: SBC, 2025, pp. 991–996.

[18] F. Alexandrino, C. Pacheco, D. Carvalho, and E. Ogasawara, "Aumento de dados e suavização integrada para predição de séries temporais baseada em aprendizado de máquina," in *Anais do XVIII Brazilian e-Science Workshop (BreSci)*. Porto Alegre, Brazil: SBC, 2024, pp. 32–39.

[19] J. I. Hidalgo, J. Alvarado, M. Botella, *et al.*, "HUPA-UCM diabetes dataset," *Data Brief*, vol. 55, art. 110559, 2024.

[20] S. Khamesian, A. Arefeen, B. M. Thompson, *et al.*, "AZT1D: a real-world dataset for type 1 diabetes," Mendeley Data, V1, 2025.

[21] A. Alsuhaymi, A. Bilal, D. Gasca García, *et al.*, "A longitudinal multimodal dataset of type 1 diabetes," *Sci. Data*, vol. 12, art. 1379, 2025.

[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[23] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] K. Cho, B. van Merriënboer, C. Gulcehre, *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.

[26] C. Royston, R. Hovorka, and C. K. Boughton, "Closed-loop therapy: recent advancements and potential predictors of glycemic outcomes," *Expert Opin. Drug Deliv.*, vol. 22, no. 6, pp. 1–13, 2025.

IA Bucal Files: Development of a Platform for Curating and Annotating Clinical Images of Oral Lesions

1st Deyvisom Michel M. Dantas
Ph.D. Program in Bioinformatics
UFRN
Natal, Brazil
deyvisom84@gmail.com

2nd Letícia Thaynna Xavier A. O. de Azevedo
Dept. of Biomedical Engineering
UFRN
Natal, Brazil
thaynna.araujo.016@ufrn.edu.br

3rd Francisca Thaís Queiroz Rego
Dept. of Biomedical Engineering
UFRN
Natal, Brazil
thais.rego.092@ufrn.edu.br

4th Caroline Dantas Vilar
Dept. of Biomedical Engineering
UFRN
Natal, Brazil
caroline.vilar@ufrn.edu.br

5th Heliana Bezerra Soares
Dept. of Biomedical Engineering
UFRN
Natal, Brazil
heliana.soares@ufrn.edu.br

6th Francisco Eduardo M. de Castro Alves
Dept. of Technology
FEMCA
Natal, Brazil
fealvesphone@gmail.com

***Abstract*—Oral cancer and oral potentially malignant disorders (OPMDs) represent a major public health concern, with early diagnosis playing a crucial role in improving patient survival. In recent years, Deep Learning-based Artificial Intelligence (AI) models have emerged as a promising approach to support the diagnosis of these lesions. However, their development and performance depend on the availability of standardized and well-annotated clinical image datasets. This study presents the preliminary results of the development of IA Bucal Files, a platform designed for the curation, annotation, and validation of clinical images of oral lesions. To date, the database comprises 5,440 clinical images, of which 2,670 are currently undergoing annotation and 807 have already been validated by specialists. The partial analysis includes 4,232 annotations, predominantly benign lesions (3,212), followed by oral potentially malignant disorders (529) and malignant lesions (491). The most frequent conditions include Geographic Tongue (221), Secondary Syphilis Mucous Plaques (198), Leukoplakia (188), and Proliferative Verrucous Leukoplakia (163). These preliminary findings demonstrate the platform's potential to provide a robust and diverse clinical image database, supporting the training of AI models for computer-aided diagnosis and the early detection of oral lesions.**

***Index Terms*—Oral Cancer, Artificial Intelligence, Deep Learning, Image Annotation.**

I. Introduction

Oral diseases represent a major global public health problem, accounting for hundreds of thousands of new cases and deaths each year [1], [3], [7]. Various lesions can affect the oral mucosa, leading to significant clinical and epidemiological impacts, with direct consequences for patients' quality of life and survival [1], [6]. Early detection of these lesions is essential to improve prognosis, increasing survival rates from approximately 40%–50% to more than 80% when diagnosis occurs at early stages [1], [3], [6]. Conversely, delayed diagnosis limits available therapeutic options, often requiring more aggressive interventions and substantially reducing treatment success rates [7], [9].

Oral cancer rarely develops abruptly and is commonly preceded by visible clinical alterations of the oral mucosa known as Oral Potentially Malignant Disorders (OPMDs) [5]. Due to its high morbidity and mortality, oral cancer is widely recognized as a significant public health concern [2]. The main OPMDs include leukoplakia, erythroplakia, oral lichen planus, and oral submucous fibrosis [4], [6].

Leukoplakia is predominantly characterized as a white plaque of questionable risk, with pooled malignant transformation rates of approximately 9.8% [4], [7]. Erythroplakia, in turn, presents as red lesions of the oral mucosa and, although less common, exhibits a considerably higher malignant transformation potential [4], [6]. Oral lichen planus is a chronic autoimmune inflammatory disease, with estimated malignant transformation rates ranging from 1.14% to 2.28% [7]. Oral submucous fibrosis is a chronic and progressive fibrotic disorder whose malignant transformation rates range from 4.2% to 30% [6], [7]. Therefore, rigorous clinical follow-up of these conditions is essential to guide therapeutic decisions and prevent progression to invasive carcinoma [4].

Despite the easy visual accessibility of the oral cavity, the clinical diagnosis of these lesions remains a significant challenge. This is mainly due to the high visual similarity between malignant lesions, potentially malignant disorders, benign alterations, and even normal anatomical variations [1], [2].

Furthermore, oral lesions exhibit substantial clinical variability, and patients often present with diverse symptoms [1]. As a result, diagnostic accuracy depends heavily on

the examiner's experience, highlighting the importance of specialists in Oral Medicine and Oral Pathology [1], [3]. However, conventional visual examination is inherently subjective and susceptible to interobserver variability [3]. This issue is even more pronounced in regions with limited healthcare resources and infrastructure, where the shortage of specialized professionals contributes to delayed diagnosis and late patient referral [2], [3].

In this context, Artificial Intelligence (AI) has emerged as a promising strategy to support diagnosis and improve clinical assessment in medicine and dentistry [1], [8]. In particular, computer vision combined with Computer-Aided Detection (CAD) systems has introduced new approaches for medical image analysis [2]. At the core of these advances are Deep Learning models, especially Convolutional Neural Networks (CNNs), which are capable of automatically learning complex image features directly from pixel data without requiring manual feature extraction [1], [2], [8].

Recent studies have demonstrated that these algorithms can detect and classify oral lesions, distinguishing malignant lesions from benign conditions and normal tissues with accuracy, sensitivity, and specificity often comparable to or even exceeding those achieved by human specialists [1], [3], [8]. However, the performance and generalization capability of Deep Learning models in real clinical settings depend on the availability of large datasets characterized by high quality, diversity, and representativeness [1], [2].

The development of such datasets relies directly on expert annotation. Clinical images must be carefully evaluated and labeled by pathologists and specialists, frequently using histopathological biopsy results as the gold standard to ensure that algorithms learn clinically meaningful patterns [2], [3]. Nevertheless, the development of robust AI solutions is still limited by the scarcity of publicly available clinical image datasets for oral lesions. Currently, many studies continue to rely on relatively small datasets collected from single institutions, limiting the generalizability of the proposed models [1], [3].

Therefore, the development of standardized tools for the organization, annotation, and systematic management of large-scale clinical image collections is essential. Furthermore, strengthening multicenter collaborations to establish diverse clinical image repositories represents a crucial step toward advancing more robust, reliable, and clinically applicable Artificial Intelligence solutions for the early detection of oral lesions [1], [3], [8].

II. Methodology

This methodology describes the operation and use of the IA Bucal Files platform, an ecosystem designed for the creation of high-fidelity datasets intended for training Artificial Intelligence models in Oral Pathology.

A. Introduction and Workflow

The system operates through a structured workflow consisting of four main stages: image upload, technical annotation, expert review, and data export in the standardized COCO JSON format, as illustrated in Figure 1. The workflow is organized according to user roles: the Annotator is responsible for data entry and lesion annotation; the Reviewer performs the technical and diagnostic validation of the annotations; the Exporter consolidates the data for training Artificial Intelligence models; and the Administrator manages the system governance, including user and taxonomy management.



Fig. 1. Operational workflow of the annotation platform.

B. Getting Started and Navigation

The system is accessed through a web browser using authentication credentials provided by the administration. After logging in, users are presented with a dashboard displaying productivity indicators, such as the total number of projects, images, approved annotations, and pending tasks. Navigation is centralized through a side panel whose functionalities are adapted according to the user's access profile, providing access to modules such as the Reference Library, Projects, Review Queue, and User Management.

C. Projects and Datasets

Data organization begins with the creation of Projects, which serve as the main containers. Within each project, Datasets are created, where users must specify the image source (such as books, journal articles, undergraduate theses, or scanners) and reference metadata, such as the DOI.

During the upload process, the system performs automatic preprocessing to ensure dataset quality: Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to enhance image contrast, perceptual hashing is used to detect duplicate images, and a technical analysis of image sharpness and brightness is performed. The resulting gallery allows users to search by diagnosis and filter images according to annotation status.

As a fundamental quality requirement established during the literature review and data screening, 100% of the validated clinical images in the datasets possess diagnostic confirmation through histopathological examination (the gold standard). The platform preserves the explicit link between each image and its original scientific source (book, journal article, or academic thesis), enabling both annotators and reviewers to readily consult the source material within the reference library to resolve any diagnostic or anatomical doubts during curation.

D. Image Separation and Processing

In cases where a single image contains multiple pathologies or requires a specific region of interest, the Separation Workspace is used. This tool allows image duplication or precise cropping. The methodology recommends duplicating the image before cropping in order to preserve the integrity of the original image in the repository.

E. Annotation and Classification Protocol

After cropping, when necessary, the annotator selects the image (or cropped region) to be analyzed. The annotation environment provides vector drawing tools, including polygons, rectangles, circles, and ellipses, enabling the precise delineation of structures of interest. At this stage, annotations are organized into three main categories: lesions, anatomy, and exclusion zones. Figure 2 illustrates this process.

In addition to lesion delineation, the system allows the annotation of anatomical structures present in the image, such as the lips, teeth, tongue, palate, and oral mucosa, as well as exclusion zones, corresponding to regions that should not be considered during model training, such as instruments, reflections, and other irrelevant areas.

The annotator can also adjust vertices, modify polygons, and reclassify annotations whenever necessary. Once completed, the annotations are automatically forwarded to the review queue, where they await validation by Oral Health specialists.

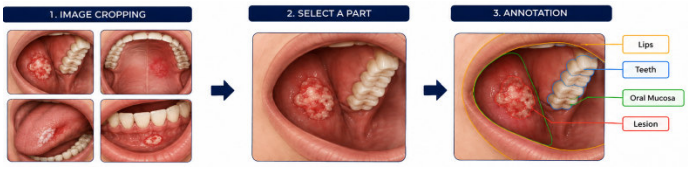


Fig. 2. Workflow of the clinical image annotation process within the platform.

F. Validation and Quality Control

The validation process is independently conducted by a panel of three dental professionals who thoroughly evaluate the accuracy of lesion contours and anatomical landmark annotations. An image is classified as 'approved' and integrated into the final dataset only upon absolute consensus (unanimous approval by all three reviewers). If one or more reviewers identify any discrepancies, the image is rejected and returned to the respective annotator's queue along with detailed technical comments. To resolve divergences, the platform features an interactive comment panel visible to both the annotator and all reviewers, enabling collaborative discussions to reach consensus on borderline cases.

G. Data Export

The final stage consists of exporting the dataset for training Artificial Intelligence models through a four-step Export Wizard. Users can select projects and datasets, filter only approved annotations, and define the patch size, ranging from the original resolution to 256 × 256 pixels. At the end of the process, the system generates a ZIP file containing the selected images and their corresponding annotations in the COCO JSON format.

III. Results and Discussion

The construction, organization, and curation of the image database resulted in the compilation of 5,440 clinical images. Of this total, the screening process enabled 2,670 images to be properly annotated and labeled. As part of the quality control stage for algorithm training, 807 of these images have already undergone rigorous validation by expert professionals, as illustrated in Fig. 3.

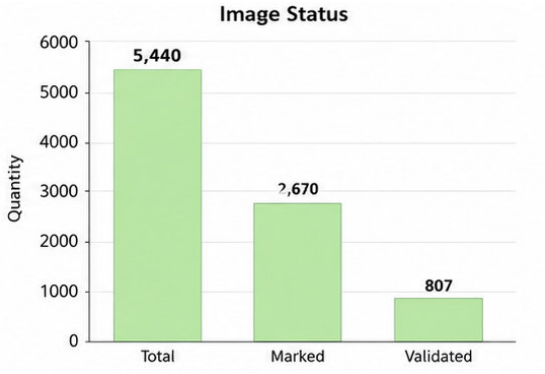


Fig. 3. Quantitative workflow of image curation, annotation, and validation within the database.

The development of Artificial Intelligence (AI) algorithms, particularly those based on Convolutional Neural Networks (CNNs), has great potential to assist in the diagnosis of oral cancer and support clinical decision-making. However, their performance and generalization capability depend directly on the quantity and quality of the data used during training [1], [2], [8].

Since this project is still under development, the results presented are preliminary. Nevertheless, the compilation of 5,440 clinical images, with 807 already validated by specialists, represents a significant advance compared with the limited dataset sizes commonly reported in the literature. Recent studies on the automatic detection of oral lesions using Deep Learning employed considerably smaller datasets, such as the study by [1], which used 518 images, and the study by [2], based on approximately 1,000 images.

The ongoing annotation and expert validation process will progressively expand a gold-standard dataset, reducing labeling bias, improving data reliability, and supporting the development of more robust models with greater clinical applicability [1], [3], [8].

Regarding the clinical and histopathological annotations (4,232 in total), benign lesions predominated (3,212 annotations), followed by 529 oral potentially malignant disorders and 491 malignant lesions, reflecting the natural distribution observed in clinical practice. The results are presented in Fig. 4.

The preliminary distribution of the categories in the database reflects the epidemiological profile commonly observed in oral medicine practice, where benign lesions (3,212 annotations) are more frequent than malignant lesions (491) and oral potentially malignant disorders (529) [1], [7]. From a computational perspective, this distribution reproduces a common challenge in Artificial Intelligence applications: class imbalance [1], [2].

It is important to emphasize that the presented results are preliminary, representing the current stage of the database

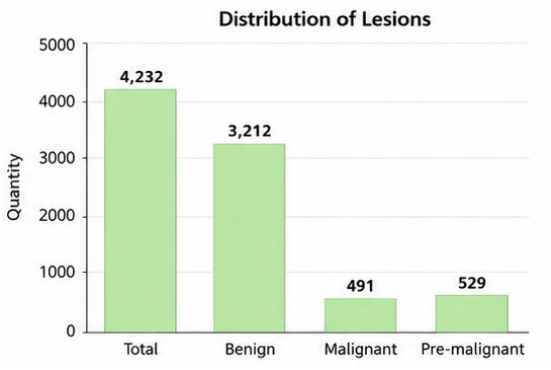


Fig. 4. Distribution of validated benign, malignant, and potentially malignant oral lesions in the database.

development. To mitigate the class imbalance naturally observed between benign, malignant, and potentially malignant lesions, the project roadmap includes targeted clinical data collection and curation focusing on minority classes. Furthermore, computational strategies such as synthetic image generation via generative models, balanced undersampling of the majority class, and weighted loss functions (e.g., Weighted Cross-Entropy or Focal Loss) will be systematically implemented prior to deep learning model training to ensure high generalization capability and prevent class bias [1], [2].

The analysis of the annotated diseases revealed a broad pathological diversity. Geographic Tongue was the most frequent condition (221 cases), followed by Secondary Syphilis Mucous Plaques (198). Among the oral potentially malignant disorders, Leukoplakia (188), Proliferative Verrucous Leukoplakia (163), and Reticular Oral Lichen Planus (103) were the most prevalent. Among the malignant lesions, Kaposi's Sarcoma accounted for 146 cases. Inflammatory, infectious, and autoimmune conditions were also identified, including Pseudomembranous Candidiasis (143), Mucocele (134), Mucous Membrane Pemphigoid (126), Pemphigus Vulgaris (112), and Melanotic Macule (107).

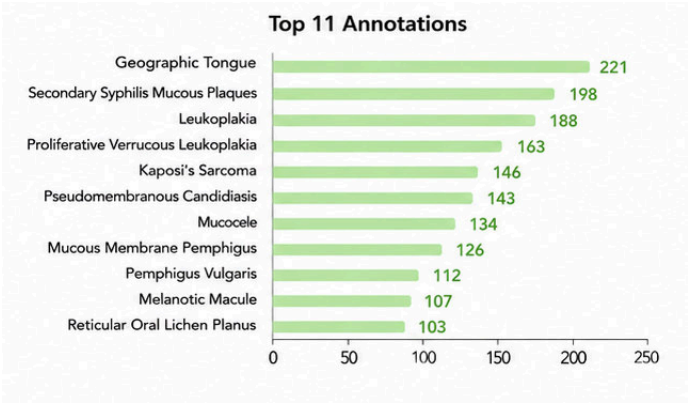


Fig. 5. Frequency of the 11 most prevalent oral diseases and lesions identified in the dataset.

The high frequency of inflammatory, infectious, and autoimmune conditions, such as Geographic Tongue (221 annotations) and Secondary Syphilis Mucous Plaques (198), represents an important characteristic for training Artificial Intelligence models. In clinical diagnosis, the main challenge is not only detecting abnormalities but also distinguishing them from benign lesions that closely resemble oral potentially malignant disorders and carcinomas [1], [7]. Therefore, the inclusion of these conditions increases the phenotypic variability of the dataset, promoting the development of models with higher specificity and lower susceptibility to misclassification [1], [3].

Another relevant aspect is the substantial representation of oral potentially malignant disorders (OPMDs), which are essential for the early diagnosis of oral cancer [4], [6]. Among them, Leukoplakia (188 annotations) and Proliferative Verrucous Leukoplakia (163), recognized for its high malignant transformation potential, stand out [4], [7]. The availability of these images enables the training of algorithms capable of identifying not only established carcinomas but also high-risk precursor lesions, contributing to early diagnosis strategies and precision medicine [6], [8].

As the annotation and expert validation stages progress, the representativeness and quality of the database are expected to increase continuously, strengthening its application in training computer vision models. This evolution may enable the development of more robust automated screening systems with greater clinical applicability, contributing to the early detection of oral lesions and reducing diagnostic delays [1], [8], [9].

Although the platform features a structured workflow, the empirical validation of its usability and clinical efficacy remains a limitation of this preliminary study and represents the next planned phase. Initially, a user experience (UX) survey will be conducted with active academic users and annotators to evaluate ease of use and the platform's learning curve.

Final Remarks

This study presented the development of IA Bucal Files, a platform designed for the curation, organization, and annotation of clinical images of oral lesions to build a database for training Artificial Intelligence models applied to the early diagnosis of oral cancer and oral potentially malignant disorders (OPMDs).

Although the project is still ongoing, the preliminary results demonstrate its potential to address an important gap in the literature. To date, 5,440 clinical images have been compiled, of which 807 have already been validated by specialists. Combined with the annotation and expert review process, this dataset contributes to the development of a high-quality database, which is essential for building more robust deep learning models with greater generalization capability [1], [2], [8].

The database encompasses a wide variety of pathologies, including benign, infectious, inflammatory, autoimmune, and oral potentially malignant lesions, promoting the learning of diverse clinical patterns and improving the differentiation between benign lesions and lesions with malignant transformation potential [1], [4], [6], [7].

As future work, the database will be continuously expanded through the inclusion of new cases and the completion of the annotation and expert validation stages. Subsequently, the dataset will be used to train and evaluate computer vision models for the automatic detection and classification of oral lesions, contributing to the development of decision-support tools for diagnosis and early patient referral, thereby reducing diagnostic delays and their impact on patient survival [1], [8], [9].

ACKNOWLEDGMENTS

The authors would like to thank the Federal University of Rio Grande do Norte (UFRN) for its institutional support and the National Council for Scientific and Technological Development (CNPq) for its financial support. They also express their sincere gratitude to Prof. Dr. Caroline Dantas Vilar for her guidance, support, and valuable contributions throughout this research. Finally, the authors thank all colleagues involved in the IA-Bucal Files project, whose dedication and collaboration were essential to the development of this work

REFERENCES

[1] S. S. Sardou *et al.*, "Enhancing early detection of oral cancer: a comparative study of artificial intelligence models and clinical specialist in lesion classification," *BMC Cancer*, vol. 26, no. 95, pp. 1–8, 2026, doi: 10.1186/s12885-025-15334-y.

[2] Y. Chai *et al.*, "Oral cancer detection via Vanilla CNN optimized by improved artificial protozoa optimizer," *Scientific Reports*, vol. 15, no. 29218, pp. 1–14, 2025, doi: 10.1038/s41598-025-11861-7.

[3] S. Goud *et al.*, "AI-based diagnostic tools for oral cancer: A systematic review," *Bioinformation*, vol. 21, no. 9, pp. 3324–3328, 2025, doi: 10.6026/973206300213324.

[4] C. S. Farah *et al.*, "Oral Cancer and Oral Potentially Malignant Disorders," *International Journal of Dentistry*, vol. 2014, Art. no. 853479, pp. 1–6, 2014, doi: 10.1155/2014/853479.

[5] S. C. Sarode, G. S. Sarode, and J. V. Tupkari, "Oral potentially malignant disorders: A proposal for terminology and definition with review of literature," *Journal of Oral and Maxillofacial Pathology*, vol. 18, Suppl. 1, pp. 77–80, 2014, doi: 10.4103/0973-029X.141322.

[6] L. Lorini *et al.*, "Overview of Oral Potentially Malignant Disorders: From Risk Factors to Specific Therapies," *Cancers*, vol. 13, no. 15, Art. no. 3696, 2021, doi: 10.3390/cancers13153696.

[7] I. González-Ruiz *et al.*, "Early Diagnosis of Oral Cancer: A Complex Polyhedral Problem with a Difficult Solution," *Cancers*, vol. 15, no. 13, Art. no. 3270, 2023, doi: 10.3390/cancers15133270.

[8] R. O. Alabi *et al.*, "Deep Machine Learning for Oral Cancer: From Precise Diagnosis to Precision Medicine," *Frontiers in Oral Health*, vol. 2, Art. no. 794248, pp. 1–11, 2022, doi: 10.3389/froh.2021.794248.

[9] J. L. Cleveland and G. Thornton-Evans, "Total Diagnostic Delay in Oral Cancer may be Related to Advanced Disease Stage at Diagnosis," *The Journal of Evidence-Based Dental Practice*, vol. 12, no. 2, pp. 84–86, 2012, doi: 10.1016/j.jebdp.2012.03.018.

Impact of Feature Selection on Hybrid Deep Learning Models for Histopathological Breast Cancer Classification

Maria Fernanda Santana
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
maria.fernandab@ufpe.br

Arianne V. A. da Silva
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
arianne.viana@ufpe.br

Adrielly Sayonara
Dept. de Farmácia
Universidade Federal de Pernambuco
Recife, Brasil
adrielly.sayonara@ufpe.br

Flávio Fonseca Secco
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
flavio.secco@ufpe.br

Clarisse Lins de Lima
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
clarisse.lins@ufpe.br

Giselle M. M. Moreno
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
giselle.moreno@ufpe.br

Wellington P. dos Santos
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
wellington.santos@ufpe.br

Maíra Araújo de Santana
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
maira.araujosantana@ufpe.br

Juliana C. Gomes Cassemiro
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
juliana.cgomes@ufpe.br

***Abstract*—Breast cancer is the most frequently diagnosed malignancy among women in Brazil and worldwide, and its definitive diagnosis relies on histopathological analysis, a process subject to interobserver variability and scalability constraints. This work proposes an intelligent decision support system for multiclass classification of breast cancer histopathological images from the public BreakHis dataset, considering eight tumor subtypes at 100x magnification. Features were extracted using a pre-trained EfficientNet-B0 network, generating 1,280-dimensional feature vectors. The training set was balanced using SMOTE, and two feature selection methods were evaluated: Particle Swarm Optimization (PSO) and Evolutionary Search. Classifiers were trained using stratified 10-fold cross-validation with 30 repetitions. During training, the SVM with linear kernel achieved 97.15% accuracy without feature selection, while the SVM with RBF kernel reached 97.89% with PSO. On the test set, the full-feature approach achieved the best performance, with 74.70% accuracy and an AUC of 0.915, indicating that the features extracted by EfficientNet already constitute a highly discriminative representation for the proposed task. These results highlight the potential of the proposed hybrid approach for multiclass breast cancer classification and demonstrate its effectiveness in a challenging eight-class histopathological classification scenario.**

***Index Terms*—Breast cancer, Histopathological images, Hybrid models, Machine learning**

I. Introduction

Despite the diagnostic and therapeutic advances of recent decades, breast cancer remains one of the most relevant oncological challenges for women's public health [1], [2]. The disease results from a complex process of genomic instability and cellular dysregulation, in which accumulated mutations eliminate cell cycle control mechanisms and promote invasive behavior [3]. Its clinical expression is variable and influenced by genetic, hormonal, metabolic, and environmental factors, making risk prediction and treatment selection complex tasks [4], [5]. Globally, 2022 data indicated that breast cancer accounted for approximately 11.6% of all cancers diagnosed in women, placing it ahead of other types in absolute number of new cases [2]. In Brazil, this reality is equally concerning, as the geographic distribution of cases reveals regional disparities, with higher incidence concentrated in the South and Southeast of the country, reflecting differences in healthcare access, screening patterns, and socioeconomic profiles [1].

The definitive diagnosis of breast cancer necessarily involves histopathological analysis of biopsy tissue [6]. Hematoxylin and eosin staining renders visible under the microscope the cellular and nuclear organization changes that allow tumor classification by type, grade, and biological behavior, information that defines treatment and prognosis [6], [7]. The central role of histopathological examination in the oncological diagnostic workflow is widely recognized; however, it carries an inherent contradiction, as the process depends entirely on the perceptual ability and accumulated clinical expertise of the pathologist [8]. Studies document that evaluations of the same specimen by different observers, or by the same professional at different times, can yield divergent classifications, compromising report reproducibility [9]. This problem tends to worsen in high-

demand clinical settings, where the daily volume of slides to review strains sustained attention and increases the risk of interpretation errors [8], [9]. Therefore, the challenge is not conceptual, as histopathology remains the reference standard, but operational: ensuring diagnostic consistency at scale without relying exclusively on human judgment.

In this context, the demand for solutions that reduce observational dependence in histopathological diagnosis has driven, over recent decades, the development of machine learning-based computational support systems [8], [10]. Among the available architectures, Convolutional Neural Networks (CNNs) stand out for their ability to automatically and hierarchically extract visual representations of increasing complexity from images, without requiring manual feature engineering [10], [11]. The transfer learning strategy, which consists of adapting pre-trained weights from large datasets to new classification tasks, has broadened the applicability of these networks in medical contexts where labeled data are scarce [12], [13]. Furthermore, architectures such as EfficientNet, designed to balance accuracy and computational cost through systematic scaling of depth, width, and resolution, have produced relevant results in histopathological classification tasks evaluated on public reference datasets [12]. However, the use of these networks as feature extractors generates high-dimensional feature vectors that frequently include redundant or low-discriminative characteristics, which may impair model generalization and unnecessarily increase training cost [13], [14]. Feature selection thus emerges both as a computational efficiency resource and as a strategy to produce more robust, interpretable, and clinically translatable models [15].

Given this scenario, this work proposes an intelligent decision support system for breast cancer diagnosis based on multiclass classification of histopathological images from the public BreakHis dataset, considering eight tumor subtypes at 100x magnification. The approach combines feature extraction via transfer learning with EfficientNet, class balancing through SMOTE, and feature selection using bio-inspired methods, investigating the impact of dimensionality reduction on the performance of classical classifiers. The results demonstrate that it is possible to build accurate, computationally efficient models with clinical application potential, contributing to the advancement of automated tools for histopathological diagnosis support.

II. Related works

Silva et al. (2024) [16] investigated the potential of breast thermography as a complementary technique for breast cancer diagnosis by proposing a hybrid architecture that combines feature extraction through transfer learning using pre-trained CNNs with conventional machine learning classifiers. To address class imbalance, the authors employed the Synthetic Minority Over-sampling Technique (SMOTE) and compared the direct use of feature vectors with an optimized feature selection approach based on Particle Swarm Optimization (PSO). The experimental results showed that the combination of Inception V3 and an SVM with a third-degree polynomial kernel achieved the highest accuracy of 79.92%, whereas the configuration using an RBF kernel (γ = 0.25) attained 100% sensitivity and an AUC of 0.83. The application of PSO reduced both the data dimensionality and the training time by approximately 50%, while maintaining an accuracy of 78.55%. The study demonstrates the feasibility of combining breast thermography with deep learning and highlights intelligent feature selection as a key strategy for developing lightweight and computationally efficient models with potential applicability in resource-constrained clinical settings.

Following a similar research direction, Gupta and Chawla (2020) [17] proposed a two-stage model for the automatic classification of breast cancer histopathological images. The authors employed pre-trained CNN architectures as fixed feature extractors, whose activation vectors were subsequently used as input to conventional machine learning classifiers, specifically Support Vector Machine (SVM) and Logistic Regression (LR). Validated on a publicly available dataset, the study demonstrated that the combination of ResNet50 and LR achieved an accuracy of 93.27%, slightly outperforming the SVM-based configuration. Stratified analyses according to image magnification further revealed that models trained specifically on 40$\times$ and 100$\times$ images yielded superior performance, suggesting that textural patterns at intermediate magnification levels provide greater discriminative power for breast cancer classification.

Using another approach applied to breast thermography, Santana et al. (2022) [18] proposed a model based on Deep Wavelet Neural Networks (DWNNs) combined with an SVM classifier. The study employed 1,052 thermographic images distributed across four classes, with the Synthetic Minority Over-sampling Technique (SMOTE) applied to address class imbalance. Unlike conventional CNNs, the DWNN uses predefined wavelet filters, reducing computational cost and eliminating the need for intensive network training. The results, obtained through 10-fold cross-validation, achieved a mean accuracy of 99% and a Kappa coefficient of 0.99, with particularly high sensitivity for malignant and lesion-free cases. The study demonstrates that lightweight hybrid architectures based on wavelet transforms can provide highly efficient classification of thermographic images, highlighting their potential for clinical applications in resource-constrained settings.

Exploring the potential of computational pathology in precision oncology, Huang et al. (2023) [19] developed an automated feature extraction approach for whole-slide histopathological images aimed at predicting response to neoadjuvant chemotherapy in breast cancer. The proposed method integrated hematoxylin and eosin (HE) and multiplex immunohistochemistry images to characterize the tumor microenvironment through 36 interpretable features related to immune composition. LASSO-regularized logistic regression models achieved an AUC of 0.8975 for HER2-positive patients and 0.7674 for triple-negative breast cancer (TNBC), outperforming conventional approaches based on manual pathological assessment. External validation confirmed the robustness of the model for the HER2-positive subtype, although lower performance was

observed for TNBC due to its greater tumor heterogeneity. The study demonstrates that automated extraction of immune microenvironment features can support risk stratification and personalized therapeutic decision-making.

Macedo et al. (2021) [20] investigated the use of swarm intelligence techniques applied to breast thermography for the detection and classification of breast lesions. The proposed method combined shape and texture feature extraction, using Zernike moments and Haralick descriptors, with multi-objective feature selection algorithms and conventional machine learning classifiers. Validated on the MammoTherm dataset, comprising 336 thermographic images from Brazilian patients, the study demonstrated that the optimized feature selection significantly reduced computational cost without compromising diagnostic performance, achieving an accuracy of up to 87% and a Kappa coefficient of 0.80 with the MLP classifier. Furthermore, the geometric descriptors based on Zernike moments exhibited greater discriminative capability for distinguishing benign from malignant lesions, highlighting the potential of breast thermography combined with swarm intelligence-based optimization techniques as a complementary tool for breast cancer screening.

III. Materials and Methods

A. Proposed Method

The stages of the methodological workflow applied in this study are illustrated in the diagram in Figure 1. It starts with data acquisition and selection from the BreakHis dataset. The following steps are feature extraction, data splitting into training and testing sets. Morover, the next stages are class balancing and feature selection, followed by: training, validating, and testing the classification models. Each stage of the proposed methodology is described in detail in the following sections.

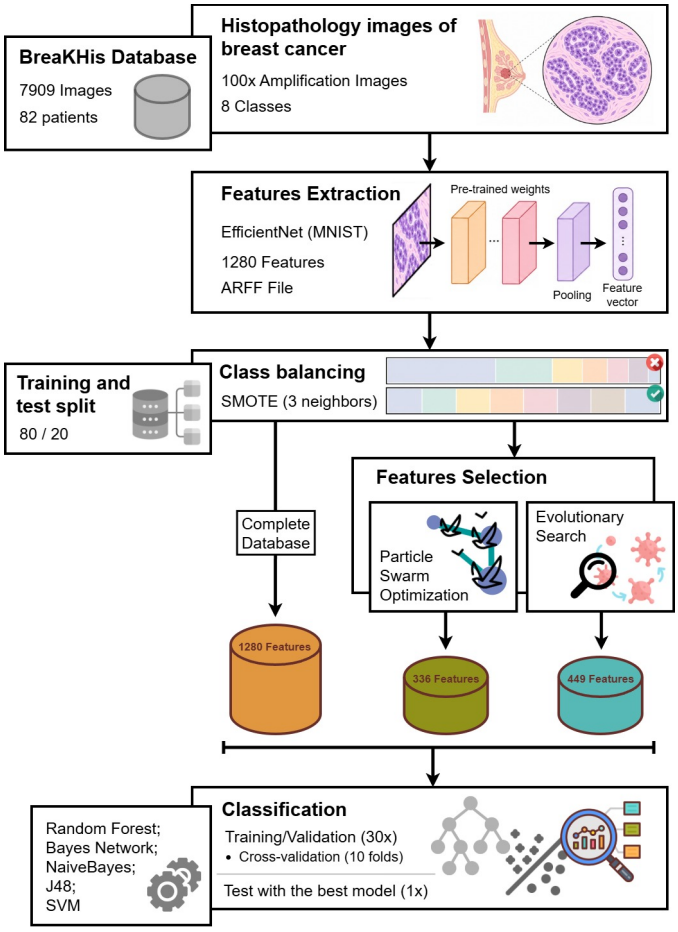


Fig. 1: Workflow of the proposed methodology, including the following stages: dataset, feature extraction, class balancing, feature selection, and classification.

B. Dataset

We used the BreakHis dataset, developed by Spanhol et al. (2016) [21] at the Research and Development Laboratory in Anatomic Pathology and Cytopathology in Paraná, Brazil. The BreakHis dataset was created to support research aimed at detecting breast cancer in histopathological images using machine learning techniques.

The dataset consists of 7,909 microscopic images of tumor tissue obtained from 72 patient's biopsies. The samples were acquired using four different microscopic magnification factors: 40x, 100x, 200x, and 400x. In total, there available 2,480 images of benign tumors, and 5,429 images of malignant tumors. The benign lesions are distributed among the following subtypes: Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenoma (TA). The malignant lesions are distributed among the following subtypes: Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC).

Although the database provides images at four levels of microscopic magnification, we used samples at 100x magnification. We chose the 100x magnification factor because it provides a more comprehensive view of the overall tissue architecture. This potentially minimizes the need for manual delineation of regions of interest. Thus, the sample used contains 2,081 images, of which 664 belong to the benign tumor class and 1,437 belong to the malignant tumor class. The images are in PNG format, with a resolution of 700 x 460 pixels, three RGB color channels, and a depth of 8 bits per channel. Fig. 2 shows an example of the images used.

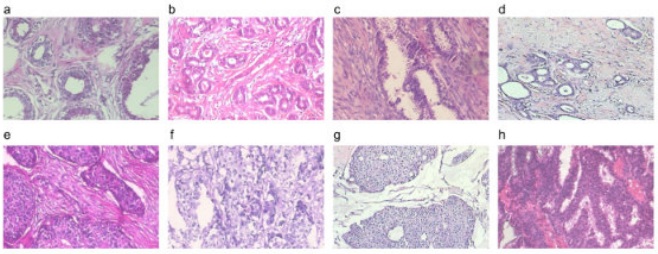


Fig. 2: Examples of histopathological images of breast cancer: a) Adenosis; b) Fibroadenoma; c) Phyllodes tumor; d) Tubular adenoma; e) Carcinoma; f) Lobular carcinoma; g) Mucinous carcinoma; h) Papillary carcinoma.

C. Feature Extraction

The feature extraction was performed using the transfer learning technique on a pre-trained EfficientNet architecture. We applied EfficientNet-B0, configured with 224 x 224 pixel input images, three color channels, a mini-batch size of 32, and frozen ImageNet-pretrained weights—leveraging the concept of transfer learning. We also used max pooling as the pooling method. All these configurations were applied using the Dl4jMlpFilter. This filter is available in the DeepLearning (DL4J) package for Weka [22]. Feature extraction was performed directly on the original images, without any prior preprocessing, including histological stain normalization. This process resulted in a dataset comprising 2,081 instances and 1,280 extracted features. It is important to note that the class attribute was preserved in the following format: A, F, PT, TA, DC, LC, MC, PC. This enabled the evaluation of the model's ability to distinguish between the eight histological subtypes observed in the dataset.

D. Classification

After the feature extraction stage, we split the dataset into training and testing sets. We used 80% of the dataset for training—comprising a total of 1,662 instances—and 20% for testing, with 419 instances. We also applied the SMOTE (Synthetic Minority Over-Sampling Technique) balancing technique to address the class imbalance within the dataset. To balance the data, we used the Weka SMOTE filter [22] with 3 nearest neighbors.

We applied the data balancing technique only to the training set. Thereby we preserved the original distribution of the test set. This strategy was adopted to ensure that performance was measured in a scenario closer to the actual data distribution.

For the classification task, we evaluated the following algorithms: Naive Bayes, Random Tree, Random Forest, and Support Vector Machine. All model training experiments were conducted in Weka, using 10-fold cross-validation with 30 repetitions for each configuration of each algorithm evaluated. For the Naive Bayes and Random Tree algorithms, the model hyperparameters remained unmodified. In contrast, the Random Forest algorithm was evaluated using 100, 200, and 300 trees. For the SVM, we evaluated configurations using a polynomial kernel of degree 1, and a RBF kernel with $\gamma = 0.5$. For both SVM configurations, the regularization hyperparameter was set to C = 1.0.

Each classifier was evaluated based on the metrics of accuracy, sensitivity, specificity, Kappa index, F-measure, and Area Under the ROC Curve (AUC). We applied the best-performing classifier to the test set described earlier in this section.

E. Feature selection

After evaluating the classifiers, we carried out an aditional step: feature selection. The goal was to reduce the datasets dimensionality, and investigate the impact of reducing the number of features on the classifiers performances. We used Weka's [22] AttributeSelection filter, and we selected two heuristic search methods: Evolutionary Search (EvSearch) and Particle Swarm Optimization (PSO).

For the Evolutionary Search algorithm, we used a population of 20 individuals over 20 generations. The selection operator was set to 1, crossover probability to 0.6, mutation probability to 0.1, and the random seed to 1. The remaining hyperparameters were left at their default Weka settings.

For the PSO algorithm, we used a population of 20 particles, 20 iterations, and a mutation rate of 0.01. Acceleration coefficients were set to 0.33 and 0.34, the reset frequency to 20, and the random seed to 1. The remaining hyperparameters were left at their default Weka settings.

We generated two subsets using the features selected by the EvSearch and PSO algorithms. These two subsets underwent training/test set splitting, class balancing, and evaluation of the classifiers applied to the training sets. The classifiers evaluated were the same as those described in Section III-D, using the same configurations. Each experiment was run 30 times using 10-fold cross-validation.

Finally, the best-performing classifiers (for each approach) were evaluated using their respective test sets. This allowed us to compare the performance of the models generated with and without the application of feature selection algorithms.

IV. Results and Discussion

A. Training Using the Complete Feature Set

Table I presents the results obtained during the training stage using the complete set of features extracted by the proposed architecture, without applying any feature selection technique. The evaluated classifiers included Naive Bayes, Bayes Network, Random Tree, Random Forest, and Support Vector Machine (SVM), each tested under different parameter configurations.

Classification with all attributes							
Classifier		Accuracy	Kappa Index	Sensitivity	Specificity	F-Score	AUC
Naive Bayes		7.031% ± 0.001	0.519 ± 0.071	0.765 ± 0.150	0.992 ± 0.011	0.834 ± 0.107	0.716 ± 0.015
Bayes Network		73.212% ± 0.002	0.694 ± 0.022	0.838 ± 0.044	0.984 ± 0.006	0.860 ± 0.031	0.983 ± 0.006
Random Tree		64.791% ± 2.139	0.598 ± 0.024	0.771 ± 0.054	0.965 ± 0.010	0.765 ± 0.043	0.868 ± 0.028
Random Forest	100 Trees	95.583% ± 0.788	0.950 ± 0.009	0.997 ± 0.006	0.996 ± 0.002	0.985 ± 0.009	1.000 ± 0.000
	200 Trees	95.963% ± 0.752	0.954 ± 0.009	0.997 ± 0.006	0.996 ± 0.002	0.986 ± 0.009	1.000 ± 0.000
	300 Trees	96.100% ± 0.729	0.955 ± 0.008	0.997 ± 0.006	0.996 ± 0.003	0.986 ± 0.009	1.000 ± 0.000
SVM	Linear	97.150% ± 0.654	0.967 ± 0.007	1.000 ± 0.001	0.998 ± 0.002	0.992 ± 0.007	0.999 ± 0.001
	RBF	78.272% ± 1.827	0.752 ± 0.021	0.751 ± 0.052	1.000 ± 0.000	0.857 ± 0.034	0.986 ± 0.008

TABLE I: Training results using the complete feature set.

The best performances were achieved by the Random Forest and SVM classifiers, whereas Naive Bayes, Bayes Network, and Random Tree yielded inferior results across all evaluated metrics. These findings indicate that the features extracted by the neural network possess high discriminative power, favoring classifiers capable of capturing complex relationships among the extracted features.

The best overall performance was obtained by the SVM with a linear kernel, which achieved an accuracy of 97.15%, a Kappa coefficient of 0.967, and an AUC of 0.999. Consequently, this model was selected for the testing stage. The Random Forest models also demonstrated consistent performance, with accuracies exceeding 95% across all evaluated configurations, indicating robustness regardless of the number of trees employed.

B. Training with PSO-Based Feature Selection

Table II presents the results obtained after applying feature selection using the Particle Swarm Optimization (PSO) algorithm. The PSO algorithm reduced the number of features from 1,280 to 336, corresponding to a dimensionality reduction of 73.77%. Despite this substantial reduction, the high performance of the leading classifiers was preserved compared with the training stage using the complete feature set, with Random Forest and SVM remaining the best-performing methods.

Classification Performance with PSO-Selected Features							
Classifier		Accuracy	Kappa Index	Sensitivity	Specificity	F-Score	AUC
Naive Bayes		60.766% ± 1.972	0.552 ± 0.023	0.697 ± 0.054	0.979 ± 0.006	0.757 ± 0.039	0.956 ± 0.012
Bayes Network		71.781% ± 1.889	0.677 ± 0.022	0.811 ± 0.046	0.977 ± 0.007	0.824 ± 0.031	0.979 ± 0.007
Random Tree		64.628% ± 1.991	0.596 ± 0.023	0.777 ± 0.051	0.966 ± 0.009	0.771 ± 0.040	0.872 ± 0.026
Random Forest	100 Trees	95.350% ± 0.850	0.947 ± 0.010	0.998 ± 0.006	0.996 ± 0.003	0.985 ± 0.010	1.000 ± 0.000
	200 Trees	95.793% ± 0.789	0.952 ± 0.009	0.998 ± 0.005	0.996 ± 0.003	0.987 ± 0.009	1.000 ± 0.000
	300 Trees	95.913% ± 0.787	0.953 ± 0.009	0.998 ± 0.005	0.997 ± 0.003	0.987 ± 0.009	1.000 ± 0.000
SVM	Linear	93.594% ± 0.960	0.927 ± 0.011	0.997 ± 0.007	0.996 ± 0.003	0.985 ± 0.010	0.999 ± 0.001
	RBF	97.887% ± 0.620	0.976 ± 0.007	0.997 ± 0.007	0.999 ± 0.001	0.996 ± 0.005	1.000 ± 0.001

TABLE II: Training Results Using PSO-Based Feature Selection.

In this case, the best performance was achieved by the SVM with an RBF kernel ($\gamma = 0.5$), which attained an accuracy of 97.89%, a Kappa coefficient of 0.976, and an AUC of 1.000, slightly outperforming the best result obtained using the complete feature set. The Random Forest classifiers also maintained consistent performance, achieving accuracies above 95% across all evaluated configurations.

These results indicate that PSO-based feature selection was able to eliminate redundant or less informative features without compromising the model's discriminative capability. Since the SVM with an RBF kernel achieved the best overall performance in this stage, it was selected for evaluation on the independent test set.

C. Training with Evolutionary Search-Based Feature Selection

Table III presents the results obtained after applying feature selection using the Evolutionary Search (ES) algorithm. The ES algorithm reduced the number of features from 1,281 to 449, corresponding to a dimensionality reduction of 64.95%. Similar to the PSO-based approach, this reduction in data dimensionality preserved the performance of the leading classifiers, with Random Forest and SVM remaining the most effective methods.

Classification Performance with Evolutionary Search-Selected Features							
Classifier		Accuracy	Kappa Index	Sensitivity	Specificity	F-Score	AUC
Naive Bayes		59.699% ± 1.935	0.539 ± 0.022	0.659 ± 0.053	0.982 ± 0.006	0.739 ± 0.040	0.955 ± 0.013
Bayes Network		71.733% ± 1.949	0.677 ± 0.022	0.827 ± 0.048	0.978 ± 0.007	0.834 ± 0.034	0.979 ± 0.007
Random Tree		63.964% ± 2.115	0.588 ± 0.024	0.753 ± 0.058	0.964 ± 0.009	0.751 ± 0.044	0.859 ± 0.029
Random Forest	100 Trees	95.325% ± 0.844	0.947 ± 0.010	0.998 ± 0.005	0.995 ± 0.003	0.983 ± 0.010	1.000 ± 0.000
	200 Trees	95.747% ± 0.797	0.951 ± 0.009	0.999 ± 0.004	0.996 ± 0.003	0.984 ± 0.010	1.000 ± 0.000
	300 Trees	95.884% ± 0.760	0.953 ± 0.009	0.999 ± 0.004	0.996 ± 0.003	0.985 ± 0.010	1.000 ± 0.000
SVM	Linear	95.087% ± 0.791	0.944 ± 0.009	1.000 ± 0.001	0.996 ± 0.003	0.986 ± 0.009	0.999 ± 0.001
	RBF	96.910% ± 0.808	0.965 ± 0.009	0.977 ± 0.018	1.000 ± 0.001	0.988 ± 0.010	0.999 ± 0.001

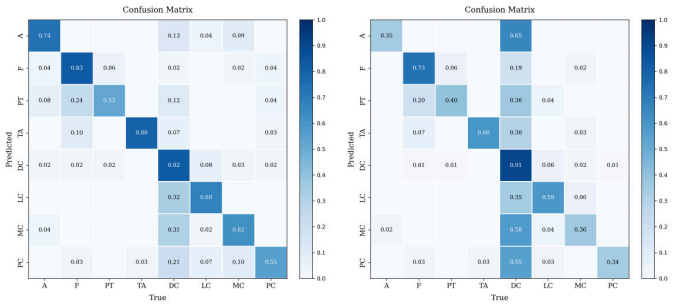
TABLE III: Training Results Using Evolutionary Search-Based Feature Selection.

The best performance was achieved by the SVM with an RBF kernel ($\gamma = 0.5$), which attained an accuracy of 96.91%, a Kappa coefficient of 0.965, and an AUC of 0.999. The Random Forest models also demonstrated consistent performance, achieving accuracies above 95% across all evaluated configurations, indicating that feature selection did not significantly compromise the classification capability.

Although the results were slightly inferior to those obtained with PSO-based feature selection, Evolutionary Search maintained high performance while using a reduced subset of features. Therefore, the SVM with an RBF kernel was selected for evaluation on the independent test set.

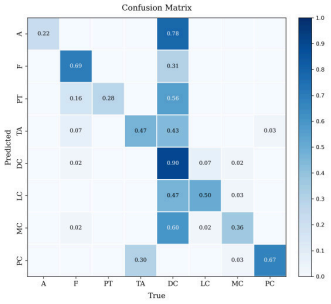
D. Test Set Results

After selecting the best-performing classifier for each feature representation strategy, the models were evaluated on the independent test set, which consisted of images not used during the training stage. The corresponding confusion matrices are presented in Figure 3 for the following approaches: (i) the complete feature set extracted by the proposed architecture, (ii) PSO-based feature selection, and (iii) Evolutionary Search-based feature selection.



(a) All features

(b) Features selected by PSO



(c) Features selected by Evolutionary Search

Fig. 3: Confusion matrices obtained on the test set: (A) all features; (B) features selected using PSO; and (C) features selected using Evolutionary Search.

The approach that directly utilized the features extracted by EfficientNet-B0 achieved the best overall performance, reaching an accuracy of 74.70

The confusion matrices further support these findings. Using the complete feature set (Figure 3(A)), a well-defined main diagonal can be observed, indicating high correct classification rates for most histopathological subclasses. Notably, Fibroadenoma (83%), Tubular Adenoma (80%), and Ductal Carcinoma (82%) exhibited the highest recognition rates. In contrast, Phyllodes

Tumor (52%), Mucinous Carcinoma (62%), and Papillary Carcinoma (55%) showed the highest levels of misclassification. In particular, a considerable proportion of Lobular Carcinoma, Mucinous Carcinoma, and Papillary Carcinoma samples were misclassified as Ductal Carcinoma, whereas Phyllodes Tumor was primarily confused with Fibroadenoma.

After applying PSO-based feature selection (Figure 3(B)), the main diagonal became less pronounced compared with the model trained using the complete feature set. Although the Ductal Carcinoma (DC) class achieved a high correct classification rate (91%), misclassifications involving other subclasses increased, particularly for Adenosis (A), Phyllodes Tumor (PT), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC), which were frequently classified as DC. This behavior explains the reduction in overall accuracy observed for this approach and suggests that some of the features discarded by PSO contained relevant information for distinguishing histopathological subclasses with similar visual characteristics.

Feature selection based on Evolutionary Search (Figure 3(C)) resulted in an even less pronounced main diagonal than that obtained with the complete feature set. Although Ductal Carcinoma (DC), Fibroadenoma (F), and Papillary Carcinoma (PC) maintained relatively high correct classification rates (90%, 69%, and 67%, respectively), a substantial reduction was observed in the recognition of Adenosis (A), Phyllodes Tumor (PT), Tubular Adenoma (TA), and Mucinous Carcinoma (MC). As observed with the PSO-based approach, a large proportion of these samples were misclassified as Ductal Carcinoma, indicating that the more aggressive dimensionality reduction produced a less discriminative representation of the histopathological subclasses. This behavior explains the lower overall accuracy achieved by this approach compared with the other evaluated strategies.

E. Discussions

Overall, the results indicate that the features extracted by EfficientNet-B0 provide a highly discriminative representation for the investigated multiclass classification task. This performance is consistent with the EfficientNet architecture, whose compound scaling strategy jointly optimizes network depth, width, and input resolution to improve feature representation.

Among the evaluated machine learning classifiers, SVM consistently achieved the best performance, particularly with linear and RBF kernels. This finding agrees with previous studies showing that SVMs are well suited for high-dimensional feature spaces, where they effectively exploit the complex representations generated by deep CNNs through robust decision boundaries.

Although both PSO and Evolutionary Search substantially reduced the feature dimensionality, neither approach improved the performance on the independent test set. This suggests that some of the discarded features contained complementary information that was important for distinguishing visually similar histopathological subclasses. Therefore, the original EfficientNet-B0 feature set already provided a highly informative representation for this classification problem.

These findings also indicate that the usefulness of feature selection should not be assessed exclusively in terms of classification accuracy. Although the complete EfficientNet-B0 representation achieved the best generalization performance, PSO and Evolutionary Search reduced the feature space by 73.77% and 64.95%, respectively. Such dimensionality reduction can decrease the amount of information that must be stored and processed during classification, which may be particularly relevant for embedded, edge-computing, or portable diagnostic systems with limited memory and processing resources. Therefore, the choice between the complete and reduced representations involves a trade-off between maximum predictive performance and computational efficiency. In applications where hardware resources, latency, or energy consumption are critical constraints, a reduced feature subset may still constitute a preferable solution, even at the cost of a moderate decrease in classification performance. Future studies should quantify this trade-off by directly measuring inference time, memory usage, and computational cost.

The confusion matrices further revealed a tendency for misclassified samples to be assigned to the Ductal Carcinoma (DC) class. This behavior is likely related to the original class distribution of the BreakHis dataset. While SMOTE was applied only to the training data, the test set preserved its natural imbalance, with DC representing the largest class. Consequently, the classifiers became more sensitive to its visual patterns, increasing confusion with underrepresented subclasses.

A deeper analysis of the misclassification patterns reveals that errors were not randomly distributed but primarily occurred between histopathological subclasses sharing significant morphological similarities. Specifically, substantial confusion was observed between Tubular Adenoma (TA) and Ductal Carcinoma (DC), as well as between Lobular Carcinoma (LC) and DC. In several instances, TA samples were misclassified as DC and vice versa, while LC showed a marked tendency to be confused with DC. This overlap mirrors well-documented challenges in routine clinical practice, where distinguishing between these subtypes can be difficult due to shared features such as glandular formation, cellular density, and nuclear atypia. For example, the architectural patterns of tubular adenomas can sometimes mimic well-differentiated ductal carcinomas, and lobular carcinomas may present subtle infiltrative patterns that resemble ductal variants. The fact that the model's confusion matrix reflects these clinically ambiguous boundaries suggests that the learned features capture biologically relevant morphological traits rather than arbitrary artifacts. Consequently, these misclassifications highlight the inherent difficulty of the eight-class problem and validate the model's alignment with known diagnostic pitfalls in breast histopathology.

It is also important to note that this study addressed an eight-class histopathological classification problem, whereas most previous studies considered only binary discrimination between benign and malignant tumors. Consequently, the reported accuracies should be interpreted in the context of a substantially more challenging and clinically relevant task,

requiring the discrimination of subtle morphological differences among tumor subtypes.

Overall, the findings indicate that the quality of the extracted features has a greater impact on classification performance than feature dimensionality reduction alone. Furthermore, they highlight the importance of evaluating models on independent test sets, as training performance may overestimate generalization capability. Among the investigated approaches, the combination of EfficientNet-B0 and SVM provided the most robust and effective solution for multiclass classification of breast cancer histopathological images. **However, this independence between training and test sets was only partial, since the training and test split was performed randomly at the image level, without controlling for patient identity, and the 100× subset originates from the same 72-patient cohort as the full BreakHis dataset. Furthermore, images from the same patient may have appeared in both sets, and since slides from the same patient can share staining intensity and tissue-specific patterns, this may have inflated the reported test-set accuracy. For this reason, future experiments should partition the data at the patient level.**

V. Conclusion

This study presented a hybrid decision-support approach for the multiclass classification of breast cancer histopathological images by combining feature extraction with EfficientNet-B0, class balancing using SMOTE, and feature selection through Particle Swarm Optimization (PSO) and Evolutionary Search (ES). The proposed methodology was evaluated on the BreakHis dataset (100× magnification) considering eight histopathological subclasses, representing a more challenging and clinically relevant scenario than the binary classification commonly reported in the literature.

The experimental results demonstrated that the original feature set extracted by EfficientNet-B0 provided the best generalization performance, achieving 74.70% accuracy and an AUC of 0.915 on the independent test set. Although PSO and ES substantially reduced the feature dimensionality and achieved excellent performance during training, both approaches resulted in lower test-set accuracy, indicating that some discarded features contained complementary discriminative information. Among the evaluated classifiers, SVM consistently achieved the best overall performance.

The proposed approach demonstrates the potential of hybrid deep learning and machine learning architectures for multiclass histopathological image classification while highlighting the importance of evaluating models on independent test sets. Nevertheless, this study was limited to a single image magnification level (100×) and a single feature extractor. Future work will investigate additional magnification factors, alternative deep feature extraction architectures, image-level data augmentation strategies, and explainable artificial intelligence (XAI) techniques to improve the interpretability, generalization, and clinical applicability of multiclass breast cancer classification models.

Acknowledgment

We are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES), Fundação de Amparo a Ciencia e Tecnologia do Estado de Pernambuco - Brazil (FACEPE), and the National Council for Scientific and Technological Development - Brazil (CNPq), for the partial financial support.

References

[1] Instituto Nacional de Câncer (Brasil), "Controle do câncer de mama no Brasil: dados e números 2025," tech. rep., INCA, Rio de Janeiro, 2025.

[2] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[3] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine*, vol. 103, no. 3, p. e36905, 2024.

[4] O. Nicolis, D. De Los Angeles, and C. Taramasco, "A contemporary review of breast cancer risk factors and the role of artificial intelligence," *Frontiers in Oncology*, vol. 14, p. 1356014, 2024.

[5] A. Roheel, A. Khan, F. Anwar, Z. Akbar, M. F. Akhtar, M. Imran Khan, M. F. Sohail, and R. Ahmad, "Global epidemiology of breast cancer based on risk factors: a systematic review," *Frontiers in Oncology*, vol. 13, p. 1240098, 2023.

[6] S. A. Alotaibi, A. Gareeballah, J. S. U. Rahman, *et al.*, "Evaluation of the diagnostic accuracy of mammography and ultrasonography for breast cancer diagnosis using histopathology as the reference standard: a cross-sectional study," *Annals of Medicine & Surgery*, 2025.

[7] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, p. 101813, 2021.

[8] E. Rigby, R. Vidya, and A. M. Shaaban, "Use of digital pathology and artificial intelligence (AI) in breast cancer diagnosis and management: opportunities and challenges," *Diagnostic Histopathology*, vol. 31, no. 3, pp. 182–190, 2024.

[9] A. A. Balasubramanian, S. M. A. Al-Heejawi, A. Singh, *et al.*, "Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology," *Cancers*, vol. 16, no. 12, p. 2222, 2024.

[10] A. B. Nassif, M. Abu Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 127, p. 102276, 2022.

[11] Y. Zhao, J. Zhang, D. Hu, H. Qu, Y. Tian, and X. Cui, "Application of deep learning in histopathology images of breast cancer: A review," *Micromachines*, vol. 13, no. 12, p. 2197, 2022.

[12] N. Sengodan, "Breast cancer histopathology classification using CBAM-EfficientNetV2 with transfer learning," *arXiv preprint*, 2024.

[13] H. Shabani *et al.*, "Reduced deep convolutional activation features (R-DeCAF) in histopathology images to improve the classification performance for breast cancer diagnosis," *Expert Systems with Applications*, 2023. PubMed PMID: 37532925.

[14] D. A. Ragab, O. Attallah, M. Sharkas, J. Ren, and S. Marshall, "Deep learning cascaded feature selection framework for breast cancer classification: Hybrid CNN with univariate-based approach," *Mathematics*, vol. 10, no. 19, p. 3631, 2022.

[15] N. A. Maruf, A. Basuhail, and M. U. Ramzan, "Enhanced breast cancer diagnosis using multimodal feature fusion with radiomics and transfer learning," *Diagnostics*, vol. 15, no. 17, p. 2170, 2025.

[16] A. F. da Silva, A. C. G. da Silva, R. Azevedo, and W. P. dos Santos, "Detection and classification of breast lesions in thermographic images by deep transfer learning hybrid architectures," *Research on Biomedical Engineering*, vol. 40, no. 3, pp. 499–520, 2024.

[17] K. Gupta and N. Chawla, "Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained cnn," *Procedia Computer Science*, vol. 167, pp. 878–889, 2020.

[18] M. A. de Santana, V. A. de Freitas Barbosa, R. de Cássia Fernandes de Lima, and W. P. dos Santos, "Combining deep-wavelet neural networks and support-vector machines to classify breast lesions in thermography images," *Health and Technology*, vol. 12, no. 6, pp. 1183–1195, 2022.

[19] Z. Huang, W. Shao, Z. Han, A. M. Alkashash, C. De la Sancha, A. V. Parwani, H. Nitta, Y. Hou, T. Wang, P. Salama, *et al.*, "Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images," *NPJ precision oncology*, vol. 7, no. 1, p. 14, 2023.

[20] M. Macedo, M. Santana, W. P. dos Santos, R. Menezes, and C. Bastos-Filho, "Breast cancer diagnosis using thermal image analysis: A data-driven approach based on swarm intelligence and supervised learning for optimized feature selection," *Applied Soft Computing*, vol. 109, p. 107533, 2021.

[21] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[22] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench*, ch. Online Appendix. Morgan Kaufmann, fourth ed., 2016.

Interpretable Structural Equation Modeling and Machine Learning Framework for Sepsis Risk Assessment

Marcos Luiz Cliton Bezerra
Hospital Pequeno Príncipe & Graduate Program in Electrical and Computer Engineering (CPGEI), Federal University of Technology - Paraná (UTFPR)
Curitiba, Brazil
ORCID: 0009-0002-0883-5216

Amauri Amorin Assef
Graduate Program in Electrical and Computer Engineering (CPGEI), DAELT Federal University of Technology - Paraná (UTFPR)
Curitiba, Brazil
ORCID: 0000-0002-2001-5387

Joaquim Miguel Maia
Graduate Program in Electrical and Computer Engineering (CPGEI), DAELN Federal University of Technology - Paraná (UTFPR)
Curitiba, Brazil
ORCID: 0000-0001-6309-8664

***Abstract*—The modeling of sepsis risk in intensive care units (ICUs) remains challenging because sepsis is heterogeneous, evolves over time, and is often associated with incomplete clinical information. An interpretable hybrid framework combining Structural Equation Modeling (SEM) and machine learning (ML) was developed to represent physiological information from time-indexed ICU data and evaluate its downstream utility for sepsis risk classification. Training sets A and B from the PhysioNet/Computing in Cardiology Challenge 2019 were merged, temporally ordered, and split at the patient level to prevent data leakage. An SEM model estimated a latent construct of blood pressure status from systolic, mean arterial, and diastolic blood pressure, and SEM-derived factor scores were integrated with physiological, demographic, and temporal covariates into supervised classifiers. After filtering, the dataset contained 1,407,607 records from 40,343 patients. The SEM model showed an acceptable global fit and adequate construct reliability, although the construct was strongly anchored in mean arterial pressure. In downstream classification, HistGradientBoosting achieved the best performance, with Area Under the Receiver Operating Characteristic - AUROC = 0.7669 and Area Under the Precision-Recall Curve - AUPRC = 0.0935, corresponding to a 5.09-fold lift over sepsis prevalence. Permutation importance identified ICULOS and temperature abnormality as the dominant predictors, while the SEM-derived factor contributed modestly. These findings support the SEM factor as a complementary representation of blood-pressure information rather than a major driver of predictive performance. Accordingly, this study should be interpreted as a methodological proof of concept demonstrating the feasibility of integrating a theory-driven latent physiological construct with supervised machine-learning models for sepsis risk prediction.**

***Index Terms*—sepsis risk modeling, structural equation modeling, machine learning, interpretability, intensive care, physiological modeling.**

I. Introduction

Sepsis remains a major challenge in acute and critical care because it represents the progression from infection to life-threatening organ dysfunction [1]. According to Sepsis-3, it is defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [1]. Its global burden is substantial, with an estimated 48.9 million cases and 11.0 million sepsis-related deaths worldwide in 2017, which corresponds to 19.7% of all global deaths [2]. Early recognition is essential because sepsis is treatable, and timely targeted interventions improve outcomes [2], whereas delayed diagnosis may postpone resuscitation, antimicrobial therapy, hemodynamic reassessment, source control, monitoring, and organ-supportive care [3]. However, strict operationalization of Sepsis-3 criteria may increase specificity at the expense of sensitivity, potentially delaying recognition in middle-income or resource-limited settings [4].

The clinical concept of sepsis has shifted from Systemic Inflammatory Response Syndrome (SIRS) to organ dysfunction and dysregulated host response. Earlier definitions identified sepsis as an infection associated with at least two SIRS criteria, but these criteria had limited specificity and sensitivity [1], [4]. Sepsis-3 adopted SOFA as the main operational criterion for organ dysfunction and introduced qSOFA as a simplified bedside prompt for patients at increased risk of poor outcomes [1], [4]. Nevertheless, the Sequential Organ Failure Assessment (SOFA) depends on physiologic and laboratory variables that may be incomplete, delayed, or unavailable at the point of care, and the Quick Sequential Organ Failure Assessment (qSOFA) should not be used as a stand-alone diagnostic definition of sepsis [1], [4].

Despite these conceptual advances, sepsis risk modeling remains clinically and methodologically challenging because sepsis is heterogeneous, time-evolving, and may involve occult organ dysfunction at presentation [1]. This limits recognition based only on static bedside criteria or delayed laboratory confirmation [1]. Electronic health records can address part of this limitation by enabling longitudinal analysis of routinely collected clinical data. In the PhysioNet/Computing in Cardiology Challenge 2019, algorithms using hourly clinical data from more than 60,000 ICU patients predicted sepsis several

hours before clinical recognition, although generalizability across hospital systems remained a major challenge [5].

Machine learning models have shown promising performance for sepsis risk classification and clinical deterioration modeling. Barton et al. reported that a gradient-boosted model using six routinely collected vital signs achieved AUROC values of 0.88 at sepsis onset, 0.84 at 24 hours before onset, and 0.83 at 48 hours before onset, outperforming SIRS, Modified Early Warning Score (MEWS), SOFA, and qSOFA at onset [6]. However, methodological barriers remain, including variation in cohorts, variables, labels, prediction windows, and metrics, while AUROC may not fully capture clinical utility, false alarms, missed events, or overtreatment [5]. In addition, many ML models remain difficult to interpret at the bedside, and explainable approaches such as SHAP have been proposed to increase transparency and support patient-level explanations [7].

Structural Equation Modeling is relevant in this context because observable clinical variables may reflect underlying physiological processes rather than isolated predictors [8]. Sepsis involves dysregulated host response and organ dysfunction across multiple biological systems, including cardiovascular, respiratory, metabolic, renal, hepatic, neurologic, and coagulation pathways [1]. By integrating measurement and structural models, SEM can represent latent physiological constructs and provide an interpretable link between routinely collected clinical indicators and plausible patterns of deterioration [8].

A hybrid SEM–ML approach may therefore combine physiological representation with downstream predictive evaluation. SEM can organize correlated clinical variables into latent constructs [8], while ML classifiers can evaluate whether these representations have discriminative capacity for sepsis risk classification in the context of previous sepsis prediction studies [5], [6]. This integration addresses complementary limitations: SEM is explanatory rather than classification-oriented [8], whereas ML supports predictive risk classification in sepsis-related tasks [5], [6] but may be limited by false alarms, generalizability challenges [5], and reduced clinician trust due to black-box behavior [7].

Although the fully nested calibration analysis prevented access to the outer test set during model and calibrator fitting, the calibrated model showed only modest Brier Skill Score relative to a prevalence-based reference. Consequently, the low post-calibration Brier score should not be interpreted in isolation as evidence of strong absolute probabilistic performance. External validation of calibration remains necessary.

Therefore, the proposed work develops and evaluates an interpretable SEM-ML framework for physiological sepsis risk modeling using time-indexed clinical data from the PhysioNet/Computing in Cardiology Challenge 2019 [5]. The primary contribution is methodological: to evaluate whether a latent representation derived from routinely collected blood-pressure variables using SEM [8] can be integrated with supervised ML classifiers as a structured hemodynamic representation. Rather than assuming that the SEM-derived factor provides substantial predictive superiority over directly observed physiological variables, its incremental contribution is evaluated empirically through downstream classification and controlled analyses, in the context of previous ML-based sepsis studies [5], [6]. The supervised task is formulated as time-indexed classification of the contemporaneous SepsisLabel at the hourly record level, without a future-shifted target or a predefined prediction horizon. Accordingly, the framework should be interpreted as a methodological proof of concept for integrating an interpretable SEM-derived representation with supervised learning, rather than as a validated lead-time-specific early sepsis prediction model, while retaining an emphasis on clinically interpretable risk information [7].

II. Methods

Training sets A and B from the PhysioNet/Computing in Cardiology Challenge 2019 were used as the data source, representing a benchmark dataset for sepsis risk classification using routinely collected ICU data [5]. Original pipe-separated files were converted to comma-separated format, assigned unique patient_id values using dataset prefixes, and concatenated into a single dataset. Data integrity was assessed by checking record and patient counts, SepsisLabel distribution, Intensive Care Unit Length of Stay (ICULOS), duplicate patient-time observations, and uniqueness of patient_id-ICULOS pairs.

A. Variables and Preprocessing

The analytical dataset was restricted to variables selected for SEM-based physiological modeling: patient_id, heart rate (HR), systolic blood pressure (SBP), mean arterial pressure (MAP), diastolic blood pressure (DBP), respiratory rate (Resp), oxygen saturation (O2Sat), temperature (Temp), Age, Gender, ICULOS, and SepsisLabel. HR, SBP, MAP, DBP, Resp, O2Sat, and Temp were treated as continuous vital signs; Age, Gender, and ICULOS were retained as demographic and temporal covariates; and SepsisLabel was used as the binary outcome. The patient_id variable was used only for patient-level organization, temporal ordering, data splitting, and leakage prevention, and was not used as a predictive feature.

Initial data quality assessment quantified missingness, valid observations, zero values, and SepsisLabel distribution at record and patient levels. Records in which all selected vital signs were simultaneously missing were removed because they did not contain physiological information for SEM estimation or downstream classification. The impact of this filtering was assessed by comparing record counts, patient counts, and SepsisLabel prevalence before and after removal.

The remaining records were ordered by patient_id and ICULOS to preserve temporal structure. Temporal consistency was verified by checking monotonic ICULOS ordering within patients and the absence of duplicate patient_id-ICULOS pairs. Missing vital signs were imputed using causal forward fill within each patient trajectory, followed by residual median

imputation estimated from the training set and applied to the test set.

Patient-level train-test splitting was performed using StratifiedGroupKFold with five folds, shuffling enabled, and $\mathrm{random_state} = 42$. The grouping variable was patient_id, and stratification was based on SepsisLabel, ensuring that records from the same patient did not appear in both training and test partitions. The split was audited by reporting record counts, patient counts, SepsisLabel prevalence, and the absence of shared patient identifiers. Final SEM datasets were created from raw variables only, without derived variables or standardization at this stage.

B. Structural Equation Modeling

Structural Equation Modeling (SEM) was performed using the *semopy* package [9], [10]. The final specification, Model_D_BP_SEM, was fitted only on the training dataset after removal of patient_id, ensuring that the patient identifier was used only for data organization and leakage prevention. Model estimation used the MLW objective function implemented in *semopy*, and fit statistics, standardized estimates, factor loadings, structural paths, covariances, and residual variances were extracted after fitting. SepsisLabel was included as an observed binary endpoint in the structural SEM to characterize standardized associations. Because estimation used the MLW objective, coefficients involving SepsisLabel were interpreted only as exploratory standardized associations, not as logistic effects, odds ratios, calibrated probabilities, or causal effects. Predictive evaluation was performed separately using supervised classifiers.

The SEM measurement model defined one latent construct, Blood_Pressure_Status, as an SEM-derived representation of the shared variation among systolic blood pressure (SBP), mean arterial pressure (MAP), and diastolic blood pressure (DBP):

$$\text{Blood_Pressure_Status} =\sim \text{SBP} + \text{MAP} + \text{DBP}. \tag{1}$$

The construct was specified as a compact representation of correlated blood-pressure measurements rather than as an a priori assumption of a physiological domain independent of its observed indicators.

The structural model evaluated the association between SepsisLabel and the latent blood-pressure construct, together with observed physiological, demographic, and temporal predictors. Thus, the outcome was modeled as:

$$\begin{aligned}\text{SepsisLabel} \sim\ & \text{Blood_Pressure_Status} + \text{HR} \\ & + \text{Resp} + \text{O2Sat} + \text{Temp} + \text{Age} \\ & + \text{Gender} + \text{ICULOS}.\end{aligned} \tag{2}$$

To account for clinically plausible physiological dependencies, covariance terms were included between cardiovascular, respiratory, and temperature-related variables:

$$\begin{aligned}\text{HR} &\sim\sim \text{Blood_Pressure_Status}, \\ \text{Resp} &\sim\sim \text{O2Sat}, \\ \text{Resp} &\sim\sim \text{Temp}, \\ \text{O2Sat} &\sim\sim \text{Temp}.\end{aligned} \tag{3}$$

Estimated parameters were organized into measurement, structural, covariance, and residual components. Standardized loadings for SBP, MAP, and DBP were used to assess the contribution of each indicator to Blood_Pressure_Status. Structural paths quantified the association between SepsisLabel, the latent construct, and the observed predictors. Composite Reliability and Average Variance Extracted were calculated from standardized loadings and residual variances as descriptive summaries of the measurement structure.

Graphical analyses supported SEM interpretation by displaying standardized loadings, structural paths, and specified covariances. The estimated Blood_Pressure_Status factor score was also examined as an auxiliary latent representation, allowing comparison of score distributions across training and test sets and according to SepsisLabel.

To assess whether the binary endpoint affected the measurement structure, a sensitivity analysis fitted a measurement-only SEM containing SBP, MAP, and DBP while excluding SepsisLabel and all structural regressions. Standardized factor loadings were compared with those of the original SEM.

C. Factor-Score Extraction

After fitting Model_D_BP_SEM on the training data, SEM-derived factor scores were extracted for the latent construct Blood_Pressure_Status. Factor scores were obtained from the fitted SEM model for both training and test datasets after removing patient_id from the SEM input matrices. The test set was not used to refit the SEM model, preserving independence between SEM estimation and downstream evaluation.

The extracted factor scores were aligned with the original time-indexed records by resetting corresponding indices. Alignment was verified by confirming that the number of extracted scores matched the number of records in each partition. Final factor-score datasets contained patient_id, SepsisLabel, and Blood_Pressure_Status_factor_score, and were validated for missing values before predictive modeling.

D. Machine Learning Classification

The supervised learning task was defined as hourly record-level classification of the contemporaneous SepsisLabel. Predictors and the target corresponded to the same time-indexed observation, and the outcome was not shifted to a predefined future prediction horizon. Accordingly, the primary analysis evaluates time-indexed sepsis risk classification rather than prospective lead-time-specific prediction.

Supervised ML classifiers were trained to evaluate whether SEM-derived physiological representations could support sepsis risk classification. The final predictive feature set included Blood_Pressure_Status_factor_score_z, HR_z, Resp_z,

O2Sat_z, Temp_Abnormality_z, Age_z, Gender, and ICULOS_z. Temp_Abnormality was defined as the absolute deviation from normal body temperature, $|\mathrm{Temp} - 37.0|$, to represent systemic temperature disturbance.

The classifiers were Logistic Regression with balanced class weights, Random Forest with balanced subsampling, and HistGradientBoosting with balanced sample weights; all applicable models used $\mathrm{random_state} = 42$.

Before training, integrity checks verified that training and test sets contained no shared patient_id values, no missing values, and no non-finite numerical values in the model inputs.

E. Evaluation Metrics

Models were evaluated on the held-out test set using predicted probabilities and binary predictions at a threshold of 0.5. Because sepsis risk classification is highly imbalanced, performance was assessed using threshold-independent metrics, including AUROC, AUPRC, Brier score, and AUPRC lift relative to positive-class prevalence, and threshold-dependent metrics, including precision, recall/sensitivity, specificity, F1-score, balanced accuracy, and confusion-matrix counts.

Additional derived metrics included false-positive rate, false-negative rate, negative predictive value, predicted positive rate, and predicted negative rate. AUPRC was used as the primary ranking criterion because it is more informative than AUROC under strong class imbalance. ROC, Precision-Recall, and calibration curves were generated to visualize discrimination, performance under class imbalance, and agreement between predicted probabilities and observed event rates.

F. Complementary Evaluation of the SEM-ML Model

Complementary analyses were conducted for the SEM factor-score-based classifiers, with the best-performing model selected primarily according to AUPRC. Threshold analysis examined how probability cutoffs affected recall, specificity, precision, F1-score, and confusion-matrix counts, allowing evaluation of trade-offs between missed sepsis events and false alarms.

A controlled ablation analysis compared four feature configurations under the same patient-level split, classifier, class weighting, and hyperparameters: (1) observed SBP, MAP, and DBP plus the remaining covariates; (2) the SEM factor alone; (3) the SEM factor plus non-blood-pressure covariates; and (4) the same covariates without the SEM factor. The primary incremental comparison was between configurations 3 and 4.

To assess dependence on ICULOS, HistGradientBoosting was retrained without ICULOS_z while preserving the same split, remaining predictors, class weighting, and hyperparameters. Changes in AUROC, AUPRC, and Brier score quantified its contribution.

Model calibration was assessed through a fully nested patient-level analysis. The outer training set was divided into subtraining and calibration partitions; residual imputation, SEM estimation, factor scoring, standardization, and HistGradientBoosting training were fitted exclusively on the subtraining data. Isotonic and sigmoid calibrators were fitted only on the calibration partition, and the independent outer test set was used solely for final evaluation. Calibration performance was characterized using Brier score, log loss, AUROC, AUPRC, and mean predicted probability. Brier Skill Score used a no-skill reference based on prevalence estimated exclusively from the subtraining partition.

The relative contribution of the SEM-derived Blood_Pressure_Status factor score and the observed physiological, demographic, and temporal covariates was subsequently examined through variable-importance analysis. A complementary post hoc temporal analysis characterized the timing of threshold-based model alerts relative to the first positive SepsisLabel within individual patient trajectories. This analysis was descriptive and was not used to redefine the primary supervised task as prospective early prediction.

TABLE I
DATASET CHARACTERIZATION AFTER PREPROCESSING AND PATIENT-LEVEL SPLIT.

Dataset	Records	Patients	Sepsis prevalence (%)
Merged A+B	1,552,529	40,346	1.7987
After filtering	1,407,607	40,343	1.8400
Training set	1,125,653	32,286	1.8420
Test set	281,954	8,057	1.8350

III. RESULTS

A. Dataset Characterization

The conversion and merging audit confirmed that the PSV-to-CSV transformation preserved the original data structure, with no discrepancies in record counts, patient counts, SepsisLabel distribution, ICULOS range, or duplicate patient-time observations. After concatenating training sets A and B, the combined dataset showed strong class imbalance, with 27,925 sepsis-labeled records and 2,933 patients with at least one positive SepsisLabel.

Records in which all selected vital signs were simultaneously missing were removed before temporal imputation. This filtering removed 144,922 records and minimally changed record-level sepsis prevalence, while preserving nearly all patients. The resulting dataset and the final patient-level train-test split are summarized in Table I.

After temporal ordering by patient_id and ICULOS, no out-of-order ICULOS trajectories or duplicated patient_id-ICULOS pairs were found. Causal forward-fill imputation resolved most missing values in the selected vital signs, and residual missing values were imputed using training-set medians. The final split contained no shared patient identifiers between partitions, and no missing values remained after residual imputation.

B. SEM Model Fit and Latent Constructs

The final SEM model, Model_D_BP_SEM, was successfully fitted on the training set using the MLW objective. Because of the large sample size, model fit was interpreted primarily through approximate fit indices rather than the

TABLE II
SEM MODEL FIT AND BLOOD_PRESSURE_STATUS CONSTRUCT SUMMARY.

Measure	Value
Degrees of freedom	47
χ^2	195,613.1
CFI	0.9156
RMSEA	0.0608
Composite Reliability	0.8782
Average Variance Extracted	0.7103
SBP loading	0.7608
MAP loading	1.0000
DBP loading	0.7430
Approximate SepsisLabel R^2	0.0219

statistically significant chi-square test. The model showed acceptable but not perfect global fit, as summarized in Table II.

Within the measurement component, Blood_Pressure_Status was defined by SBP, MAP, and DBP, with all indicators showing strong standardized loadings. The construct also showed adequate internal consistency and convergent validity. However, the unit loading and approximately null standardized residual variance of MAP indicated that Blood_Pressure_Status was strongly anchored in MAP, requiring cautious interpretation.

A measurement-only sensitivity analysis excluding SepsisLabel and all structural regressions yielded virtually unchanged standardized loadings: SBP = 0.7613, MAP = 1.0000, and DBP = 0.7427, versus 0.7608, 1.0000, and 0.7430 in the original model (maximum absolute difference = 0.00054). Thus, inclusion of the binary endpoint did not materially alter the Blood_Pressure_Status measurement structure. Because this three-indicator model was just-identified (DoF = 0), its global fit indices were not interpreted.

Figure 1 presents the complete SEM specification, including the measurement component of the Blood Pressure Status construct, the structural paths to SepsisLabel, and the physiologically motivated covariance terms among cardiovascular, respiratory, and temperature-related variables.

In the structural component, ICULOS showed the largest standardized association with SepsisLabel, whereas Blood_Pressure_Status had a small negative standardized association. The estimated structural associations were generally small, and the model explained only a limited proportion of SepsisLabel variance; accordingly, these paths were interpreted exploratorily. The specified covariances were also significant, with the strongest association observed between Resp and O2Sat, supporting the physiological plausibility of respiratory compromise and systemic temperature disturbance within the SEM specification.

C. SEM-Derived Factor Scores

Factor scores for Blood_Pressure_Status were successfully extracted from the fitted SEM model for both training and test sets. The resulting factor-score datasets contained no missing values in patient_id, SepsisLabel, or Blood_Pressure_Status_factor_score.

TABLE III
PREDICTIVE PERFORMANCE OF SEM-ML CLASSIFIERS ON THE HELD-OUT TEST SET.

Model	AUROC	AUPRC	Lift
Logistic Regression	0.7247	0.0754	4.11
Random Forest	0.7581	0.0877	4.78
HistGradientBoosting	0.7669	0.0935	5.09

The distribution of Blood_Pressure_Status_factor_score was similar between training and test partitions, showing similar distributions across the patient-level train–test split. Graphical inspection according to SepsisLabel showed only modest separation between positive and negative records, with substantial overlap between classes. Thus, the latent blood-pressure domain provided interpretable physiological information, but was not sufficient alone for clear visual discrimination of sepsis events.

D. Machine Learning Performance

Predictive models were trained using the standardized SEM-derived Blood_Pressure_Status factor score together with observed physiological, demographic, and temporal covariates. Because AUPRC was used as the primary ranking criterion under strong class imbalance, HistGradientBoosting was selected as the primary predictive model, achieving the highest AUPRC and lift over test-set sepsis prevalence, as shown in Table III.

At the default threshold of 0.5, HistGradientBoosting achieved recall of 0.6015, specificity of 0.8025, precision of 0.0539, F1-score of 0.0989, balanced accuracy of 0.7020, and Brier score of 0.1632. The low precision reflects the strong class imbalance and the resulting false-positive burden in record-level sepsis risk classification.

Ablation showed a modest incremental contribution of the SEM factor. SEM factor + covariates achieved AUROC = 0.7669, AUPRC = 0.0935, and Brier = 0.1632, compared with 0.7639, 0.0904, and 0.1656 without blood-pressure information. The factor alone showed limited discrimination (AUROC = 0.5254; AUPRC = 0.0207). Direct inclusion of SBP, MAP, and DBP yielded AUPRC = 0.0895 and Brier = 0.1627, supporting representational compactness but not general superiority of the SEM factor.

The ROC and Precision–Recall curves in Fig. 2 were consistent with the quantitative results reported in Table III, with HistGradientBoosting showing the strongest discrimination among the evaluated classifiers. The Precision–Recall curves further highlight the effect of class imbalance, as all models operated above the prevalence baseline but with limited positive predictive value.

E. Calibration and Model Interpretability

Graphical analysis provided a complementary representation of the estimated SEM structure. The latent Blood_Pressure_Status construct was strongly anchored in MAP, whereas the structural component showed that ICULOS had the largest standardized association with SepsisLabel.

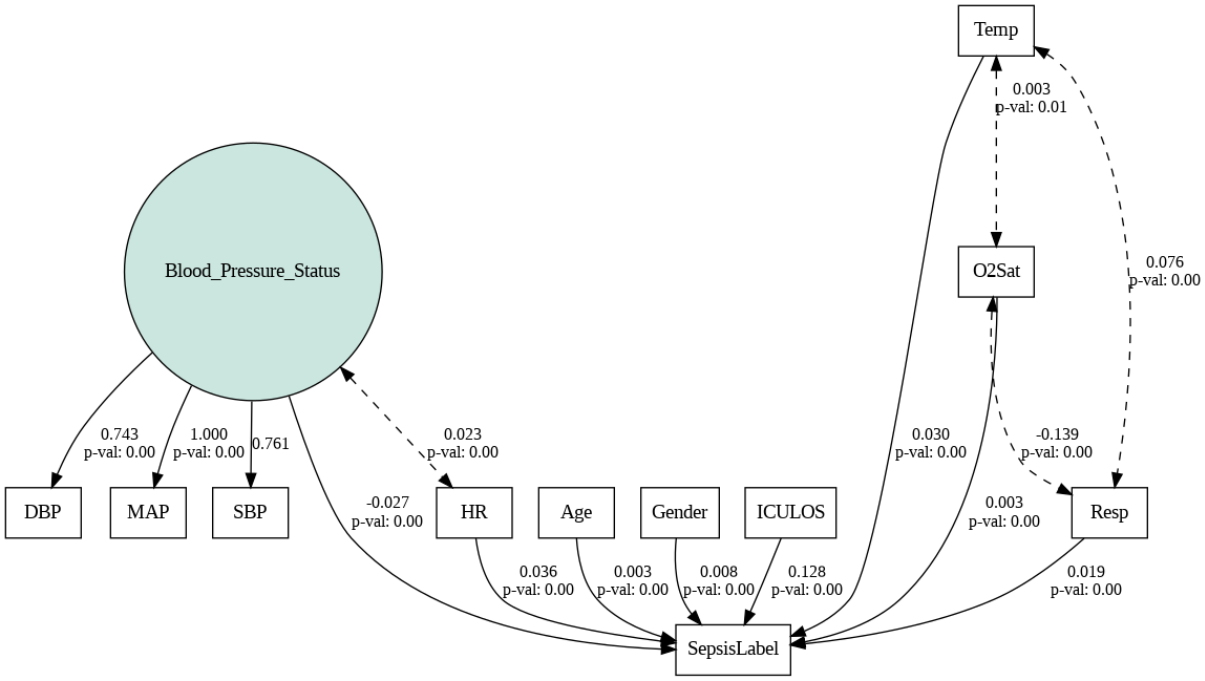
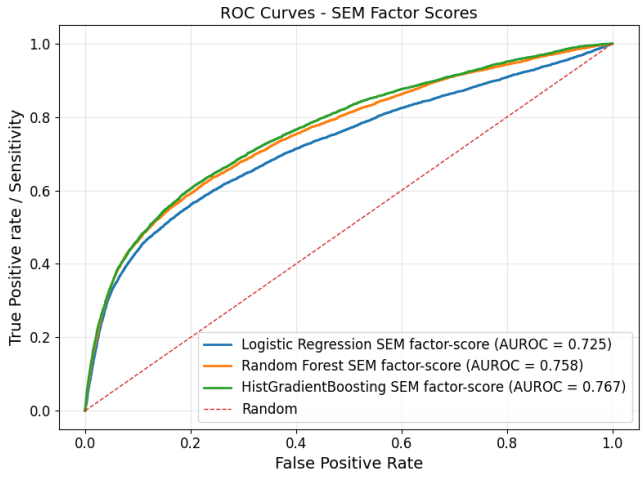
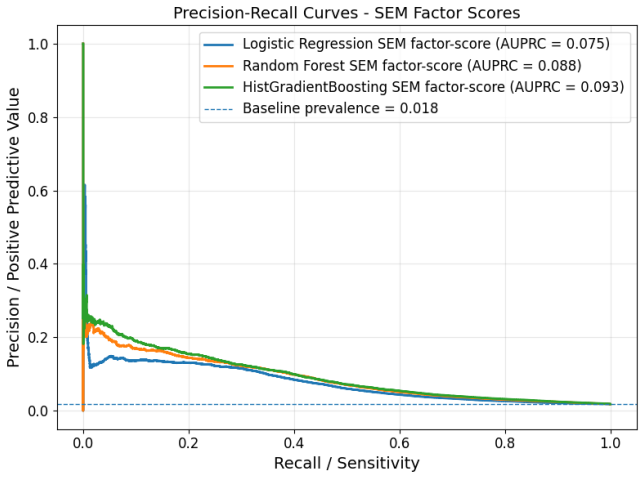


Fig. 1. Structural equation model diagram for Model_D_BP_SEM, including the latent Blood_Pressure_Status construct, observed predictors, physiologically motivated covariances, and standardized paths to SepsisLabel.



(a) ROC curves for Logistic Regression, Random Forest, and HistGradient-Boosting.



(b) Precision-Recall curves showing model behavior under strong class imbalance and the prevalence-based baseline.

Fig. 2. Discrimination performance of the evaluated SEM-ML classifiers.

The covariance structure was also physiologically plausible, particularly through the negative association between Resp and O2Sat and the positive association between Resp and Temp.

In the fully nested experiment, the uncalibrated model achieved Brier = 0.1591, AUROC = 0.7566, and AUPRC = 0.0858 on the outer test set. Sigmoid calibration reduced Brier to 0.0174 while preserving AUROC and AUPRC and was therefore preferred to isotonic calibration, which slightly reduced discrimination. The no-skill Brier was 0.01801 and sigmoid Brier Skill Score was 0.0355, indicating modest probabilistic skill beyond the prevalence baseline. Mean predicted probability decreased from 0.3502 to 0.01982 (observed prevalence = 0.01835), and no Brier degradation was observed from calibration to outer test.

Permutation importance indicated that ICULOS_z and Temp_Abnormality_z were the dominant predictors, while HR_z, Resp_z, and the SEM-derived Blood_Pressure_Status factor score had smaller but measurable contributions. As shown in Fig. 3, temporal progression and systemic temperature abnormality were the dominant predictive components,

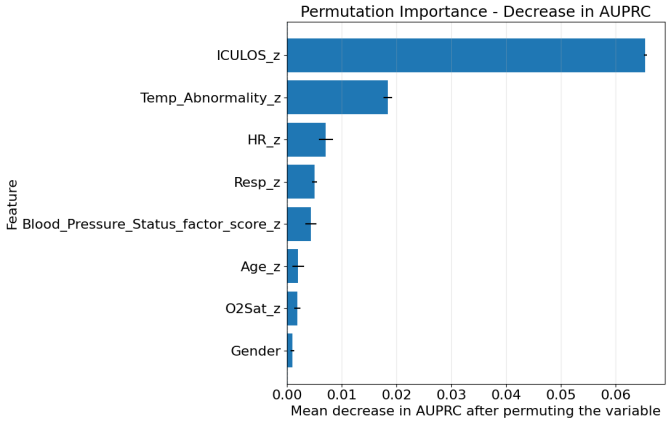


Fig. 3. Permutation importance of the HistGradientBoosting SEM factor-score model, expressed as the mean decrease in AUPRC after feature permutation. ICULOS_z and Temp_Abnormality_z were the dominant predictors, whereas the SEM-derived Blood_Pressure_Status factor score showed a smaller but measurable contribution.

while the SEM-derived factor score acted as a complementary interpretable hemodynamic feature.

IV. DISCUSSION

This study developed an interpretable SEM–ML strategy for physiological sepsis risk modeling using time-indexed ICU data from the PhysioNet/Computing in Cardiology Challenge 2019. The analysis combined a latent Blood Pressure Status construct with physiological, demographic, and temporal covariates in supervised classifiers. In this design, ML classification served as a downstream evaluation of the latent representation rather than as the sole objective of the study. HistGradientBoosting achieved the best performance according to AUPRC, although positive predictive value remained limited by the strong class imbalance.

The SEM-derived blood-pressure representation should be interpreted cautiously. MAP strongly dominated the construct, supporting its interpretation as a compact hemodynamic representation rather than a fully independent latent physiological domain. Structural associations involving SepsisLabel also require caution because the binary endpoint was modeled using the MLW objective. These coefficients represent exploratory standardized associations rather than logistic, probabilistic, or causal effects. Importantly, the measurement-only sensitivity analysis produced virtually unchanged SBP, MAP, and DBP loadings, indicating that the measurement structure was not materially driven by inclusion of SepsisLabel.

Controlled ablation confirmed that the SEM factor provided only modest incremental information: adding it increased AUPRC from 0.0904 to 0.0935, while the factor alone showed limited discrimination. Replacing SBP, MAP, and DBP with the factor preserved discrimination but did not establish predictive superiority. Thus, its value is better interpreted as compact hemodynamic representation rather than as a major performance driver.

In contrast, removing ICULOS reduced AUPRC by 58.3%, demonstrating substantial dependence on temporal position within the ICU stay. Because patient-level train-test independence was preserved, this finding does not establish leakage; ICULOS is better interpreted as an important temporal proxy within the current classification design.

Calvert et al. [11] similarly demonstrated that relationships among physiological signals can support sepsis detection, reporting AUROC = 0.92 three hours before a sustained SIRS episode. Direct comparison is limited by differences in cohort, endpoint definition, available variables, and prediction design. InSight, the machine-learning framework proposed by Calvert et al. for early sepsis prediction, incorporated higher-order interactions among physiological variables and their temporal trends. In contrast, the present framework uses SEM to explicitly represent physiological relationships before incorporating the resulting latent representation into downstream classification.

Previous studies have evaluated sepsis prediction at explicit pre-onset horizons. However, direct comparison with those systems should be made cautiously because the present study addresses contemporaneous time-indexed classification rather than fixed-horizon prospective prediction. The classifier was trained using the contemporaneous hourly SepsisLabel rather than a future-shifted outcome defined at a predefined prediction horizon. Accordingly, any retrospective temporal relationship between model alerts and the first positive SepsisLabel should be interpreted as descriptive alert timing rather than as validated early-prediction performance.

Clinically, the method relies on routinely collected ICU variables and may support risk estimation with an interpretable physiological context. Threshold analysis indicated that lower cutoffs favor sensitivity-oriented screening, whereas higher cutoffs reduce false-positive burden at the cost of missed events. Therefore, threshold selection should depend on whether the clinical objective prioritizes early sensitivity or more specific alerts linked to resource-intensive interventions.

Several limitations should be noted. The SEM was restricted to a single SEM-derived blood-pressure representation and included binary SepsisLabel while using the MLW estimation objective. Accordingly, structural coefficients involving SepsisLabel should be interpreted only as exploratory standardized associations rather than logistic or probabilistic effects. Although the measurement-only sensitivity analysis showed virtually unchanged factor loadings after removal of SepsisLabel, this does not establish MLW as the optimal estimator for binary outcomes. The Blood_Pressure_Status factor was strongly anchored in MAP, which had a standardized loading of 1.000 and an approximately null residual variance. Therefore, the factor should not be interpreted as a fully independent physiological domain. The present model was trained for contemporaneous hourly classification of SepsisLabel rather than for prediction of a future outcome at a predefined horizon. Consequently, retrospective intervals between model alerts and the first positive SepsisLabel should not be interpreted as validated early-warning lead times. The modest incremental

contribution of the SEM-derived factor has not been externally validated.

Although sigmoid calibration markedly corrected probability inflation, its Brier Skill Score relative to the prevalence baseline was only 0.0355. Thus, the low calibrated Brier score should not be interpreted alone as evidence of strong probabilistic performance. External calibration validation remains necessary.

Performance was strongly dependent on ICULOS, limiting interpretation based solely on contemporaneous physiology and motivating explicitly time-aware prospective designs.

V. CONCLUSION

The proposed SEM-ML framework demonstrates the feasibility of integrating an SEM-derived hemodynamic representation with supervised machine learning for time-indexed sepsis risk classification using ICU data from the PhysioNet/Computing in Cardiology Challenge 2019. The Blood_Pressure_Status factor provided a compact representation of SBP, MAP, and DBP and was integrated with physiological, demographic, and temporal covariates. However, the factor was strongly anchored in MAP, limiting its interpretation as an independent latent physiological domain.

In downstream classification, HistGradientBoosting achieved the best performance in terms of AUPRC. The SEM-derived blood-pressure representation showed a modest incremental contribution to discrimination, whereas ICULOS and systemic temperature abnormality were the dominant predictive components. Accordingly, the proposed framework should be interpreted primarily as a methodological proof of concept for integrating a compact SEM-derived hemodynamic representation with supervised learning rather than as evidence of substantial predictive superiority. Because the supervised task used the contemporaneous SepsisLabel at each time-indexed observation, without a future-shifted target or predefined prediction horizon, the present results should not be interpreted as validation of a lead-time-specific early sepsis prediction model. Future work should evaluate external validation, broader physiological representations, explicitly prospective prediction designs, and clinically oriented calibration and threshold strategies.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil – Finance Code 001. The authors would like to acknowledge the support provided by the National Council for Scientific and Technological Development (CNPq), the Federal University of Technology – Paraná (UTFPR), the Araucária Foundation, the State Secretariat for Science, Technology (SETI), the Hospital Pequeno Príncipe, and the PhysioNet/Computing in Cardiology Challenge 2019 for providing the publicly available dataset used in this study.

REFERENCES

[1] M. Singer et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," JAMA, vol. 315, no. 8, pp. 801-810, 2016.

[2] K. E. Rudd et al., "Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study," The Lancet, vol. 395, no. 10219, pp. 200-211, 2020.

[3] A. Rhodes et al., "Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016," Intensive Care Medicine, vol. 43, no. 3, pp. 304-377, 2017.

[4] F. R. Machado et al., "Getting a consensus: advantages and disadvantages of Sepsis 3 in the context of middle-income settings," Revista Brasileira de Terapia Intensiva, vol. 28, no. 4, pp. 361-365, 2016.

[5] M. A. Reyna et al., "Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019," Critical Care Medicine, vol. 48, no. 2, pp. 210-217, 2020.

[6] C. Barton et al., "Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs," Computers in Biology and Medicine, vol. 109, pp. 79-84, 2019.

[7] B. He and Z. Qiu, "Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis," Frontiers in Artificial Intelligence, vol. 7, Art. no. 1348907, 2024, doi: 10.3389/frai.2024.1348907.

[8] J. A. B. Neves, Modelo de equações estruturais: uma introdução aplicada. Belo Horizonte, Brazil: Universidade Federal de Minas Gerais, 2018. [Online]. Available: https://repositorio.ufmg.br/handle/1843/50562. Accessed: Jun. 6, 2025.

[9] G. Meshcheryakov and A. A. Igolkina, "semopy: A Python package for structural equation modeling," Structural Equation Modeling: A Multidisciplinary Journal, vol. 27, no. 6, pp. 952-963, Nov. 2020.

[10] G. Meshcheryakov, A. A. Igolkina, and M. G. Samsonova, "semopy 2: A structural equation modeling package with random effects in Python," arXiv, Jun. 2021. [Online]. Available: http://arxiv.org/abs/2106.01140. Accessed: Jun. 8, 2025.

[11] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, "A computational approach to early sepsis detection," Computers in Biology and Medicine, vol. 74, pp. 69-73, 2016, doi: 10.1016/j.compbiomed.2016.05.003.

Ki67 Marker-based Proliferation Index Estimation Using Deep Learning

Gabriella Cristina Machado de Moraes
CPGEI
UTFPR
Curitiba, Brazil
0009-0008-8536-0844

Yuri Figueredo de Albuquerque Fernandes Sertorio de Souza
DAELN / CPGEI
UTFPR
Curitiba, Brazil
0009-0005-5870-1111

Jandrei Sartori Spancerski
DACOM / CPGEI
UTFPR
Medianeira, Brazil
0000-0002-4077-281X

Tayza Katelline Danilau Ostroski
CHC
UFPR
Curitiba, Brazil
0000-0002-4691-855X

Sergio Ossamu Ioshii
Department of Pathology
PUCPR / UFPR
Curitiba, Brazil
0000-0002-7871-4463

Fabio Kurt Schneider
DAELN / CPGEI / PPGEB
UTFPR
Curitiba, Brazil
0000-0001-6916-1361

***Abstract*—Quantifying the Ki-67 marker is an essential step in evaluating cellular proliferation in tumor samples. However, it still largely relies on manual analyses that are subject to observer variability and subjectivity. In this study, an automated system was developed for detecting and quantifying Ki-67 expression in breast cancer histological images using the YOLOv8 deep learning model. The images were acquired using a smartphone camera attached to a microscope eyepiece, and training was conducted with K-fold cross-validation (80/20%) to ensure robustness and generalization. The model achieved balanced performance, with a precision of 0.88, a recall of 0.86, and an F1-score of 0.87, values comparable to those reported in the literature, even with a reduced dataset and a low-cost acquisition method. These results demonstrate the potential of YOLOv8 as an accessible and accurate tool for immunohistochemical analysis. In future work, we intend to develop a mobile application based on the proposed model, aiming to democratize access to digital pathology and expand automated diagnostic capabilities in regions with limited infrastructure.**

***Index Terms*—Breast cancer, Ki-67, Deep learning, YOLOv8, Digital pathology, Image analysis.**

I. Introduction

Breast cancer represents the second main cause of death by cancer among women worldwide [1]. According to data from the World Health Organization (WHO), in 2022 alone, approximately 2.3 million women were diagnosed with the disease [2].

Characterized as a heterogeneous disease, breast cancer constitutes a spectrum of distinct biological subtypes, which present variations in morphological, molecular, and clinical characteristics [3], [4]. This diversity reflects variations in tumor behavior, in the response patterns to different therapeutic modalities, and in clinical outcomes. Although conventional clinical factors, such as histological grade, tumor size, lymph node involvement, and surgical margins, are relevant in determining prognosis, they are not sufficient when considered in isolation [5]. Thus, prognostic and predictive definition is largely based on the identification and classification of the biological subtype of the tumor [2], [6], [7]. The distinction of these subtypes is based on the expression of hormonal receptors and on the indices of cell proliferation indicated by Ki-67 [8].

Ki-67 is a nuclear protein expressed in all active phases of the cell cycle (G1, S, G2, and M) but absent in resting cells (G0). This characteristic makes it a reliable marker to estimate the growth fraction of the neoplastic cell population [9]. Considering that the high rate of cell reproduction is one of the central characteristics of breast cancer [10], Ki-67 has been consolidated as an important prognostic marker [5], [7], [11], [12].

Traditionally, in clinical practice, the evaluation of the Ki-67 index is performed predominantly by two approaches, visual estimation (commonly referred to as *eyeballing*) and manual counting [13]. In the first, the pathologist estimates, based on his experience, the proportion of positive tumor cells, without effective counting. The manual counting, in turn, seeks to accurately quantify the percentage of positive cells in a given sample from the effective counting of positive and negative cells [14], [16]. However, both approaches present limitations, such as intra- and interobserver variability [13], [15], in addition to a long time required for manual counting, which can compromise workflow and negatively affect the quality of results, especially in scenarios of high demand or under

Partially funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), National Council for Scientific and Technological Development (CNPq) and Fundo Nacional de Desenvolvimento da Educação (FNDE) - Programa de Educação Tutorial (PET)

professional fatigue, making visual evaluations subjective [13].

In recent years, several technological solutions have been developed with the objective of automating the Ki-67 proliferation index in breast cancer. The study "Artificial Intelligence-assisted interpretation of Ki-67 expression and repeatability in breast cancer" demonstrated that AI-assisted interpretation achieves high interobserver reproducibility and meets the standards established by the International Breast Cancer Working Group [13]. In the high-throughput study with tissue microarrays from the Breast Cancer Association Consortium, thousands of samples were used to develop an automated algorithm, and it demonstrated that AI-assisted interpretation achieved high interobserver reproducibility and good agreement with pathologists [16]. Recent works have advanced to tumor segmentation in whole slides and registration between HE and IHC sections, which improves precision in tumor region selection [17]. Liu et al. [18] proposed an AI-based method for segmentation of invasive carcinoma regions, allowing greater precision in the selection of relevant tumor areas.

In summary, there is a large number of studies, with relevant results, that present automated solutions aiming to estimate the Ki-67 index. However, its full implementation in clinical practice requires robust computational infrastructure, microscopes with integrated scanning systems or equipment for slide digitization, and large image storage capacity. Technologies of this scale may be associated with substantial investments, and this scenario imposes significant challenges, especially in low-income regions, where resource limitations directly impact the availability of accurate and timely diagnoses.

Considering the high mortality rate associated with breast cancer, especially in regions with lower human development index [2], the development of accessible, reliable, and scalable methods for the evaluation of the Ki-67 index constitutes a need in the context of global public health. From this perspective, we developed in this study an automated scoring system using Artificial Intelligence, based on the YOLOv8 model, to quantify cellular proliferation with Ki-67 in histological images acquired through the camera of a common smartphone coupled to the microscope eyepiece.

II. MATERIALS AND METHODS

The dataset used in this study is composed of 45 pathological studies from patients diagnosed with invasive breast cancer, acquired from the Anatomy and Pathology Laboratory of Hospital Erasto Gaertner between June and September 2022. This study was approved by the Research Ethics Committee of Hospital Erasto Gaertner – Liga Paranaense de Combate ao Câncer, under approval number 013571/2022. The tissue specimens were stained with Diaminobenzidine (DAB) and counterstained with Hematoxylin (H). The DAB staining resulted in brown nuclear labeling in Ki-67–positive cells, indicating proliferative activity, while hematoxylin provided blue coloration to the nuclei of Ki-67–negative cells. For five-fold cross-validation, the 45 patients were divided into five groups using an algorithm based on the number of Ki-67 reactive cells in each case. This strategy was adopted to improve the balance of reactive cells. From each original image, 36 patches were generated with 25% overlap between consecutive patches tto reduce the loss of relevant information at patch borders. Thus, a total of 1,620 patches were obtained, corresponding to 324 patches per fold.

A. Image Acquisition

The hotspot areas were selected by a pathologist using an Olympus CX23 microscope. After the hotspot was selected by the pathologist, a smartphone camera was attached to the microscope eyepiece to capture the images.

The configuration established for image acquisition was a spatial resolution of 1190×863 pixels, with 24-bit color depth in RGB matrices. The files resulting from the acquisition process were stored in JPG format, preserving the integrity of visual information.

B. Computacional Resources and Processing Time

All image processing and model training were performed on a workstation equipped with an Intel Core i5-12400F processor (2.5 GHz, up to 4.4 GHz Turbo, 18 MB cache), 32 GB of DDR4 3200 MHz RAM, an NVIDIA GeForce RTX 4060 Ti GPU with 8 GB of GDDR6 memory, and a 1 TB NVMe SSD. Considering the entire computational pipeline, including image patch generation, model training, and subsequent reassembly of the results, the complete processing of the 45 images required approximately 15 minutes, corresponding to an average processing time of approximately 20 seconds per image.

C. Model Selection

Deep learning models have been widely used in object detection for biomedical images, providing greater accuracy and reproducibility for analyses. In this study, we chose YOLOv8[1], a mature and widely adopted version of the YOLO (You Only Look Once) family, primarily due to the stability of its implementation and the robustness of the Ultralytics ecosystem, which provides consistent tooling, active maintenance, comprehensive documentation, and good reproducibility, in addition to its favorable balance between inference speed and detection performance. YOLOv8 includes advanced backbone and neck architectures, responsible for combining and refining features at different scales for object detection, improving the model's ability to detect objects of various sizes and shapes. These attributes motivated the model's selection, given the complexity and variability of tumor cells in histopathological images.

D. Annotation

Before the annotation process, an automatic cropping by intensity projection was applied. This step was necessary to remove the black borders resulting from the acquisition process and to segment only the relevant area for analysis. Next, manual annotation of the images was performed to identify and label the cells. This stage was conducted by a

[1]Ultralytics YOLOv8: https://github.com/ultralytics/ultralytics

professional previously trained by a pathologist with 8 years of experience, ensuring accuracy and standardization of the markings.

Annotations were carried out using the LabelMe tool[2], developed in Python, which allows the delimitation of areas through bounding boxes. Through this interface, two distinct classes were labeled, Ki-67–reactive tumor cells and Ki-67 non-reactive tumor cells.

The annotations were generated directly in YOLO format (.txt) files, containing the coordinates of the marked regions and their respective class labels. In this format, each image is associated with a corresponding text file containing the normalized box coordinates and class labels. Fig. 1 shows an example of the result obtained in the manual annotation stage, highlighting the visual difference between reactive (R) and non-reactive (NR) cells. These annotations served as a reference for training and validating the automated counting method.

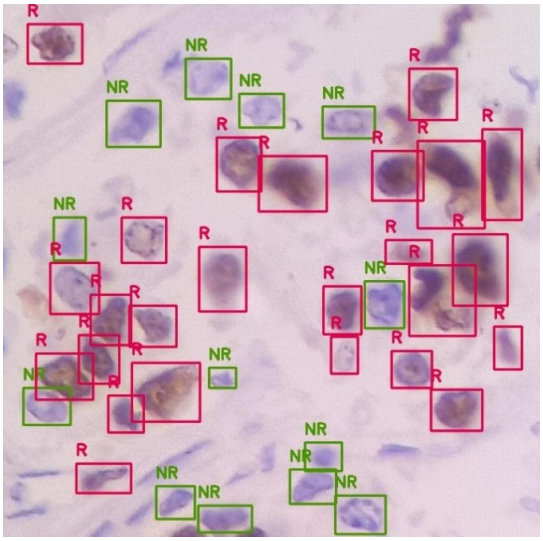


Fig. 1. Representative example of the manual annotation produced in LabelMe over a Ki-67 hotspot field. Each tumor nucleus is enclosed by a bounding box and labeled as either reactive (R, Ki-67 positive), shown in red, or non-reactive (NR, Ki-67 negative), shown in green. These annotations provided the reference used to train and validate the automated model.

E. Model Training and Validation

To ensure greater statistical robustness and reduce the risk of overfitting, the model training was conducted using the K fold cross-validation method. In this process, the dataset is divided into approximately equal subsets. In each iteration, one subset is reserved for validation while the remaining ones are used for training. This process is repeated so that all subsets are used, in different rounds, both for training and validation, as illustrated in Fig. 2.

The application of this method allows for a more reliable evaluation of the model's performance since the results are not

[2]LabelMe: https://github.com/wkentaro/labelme

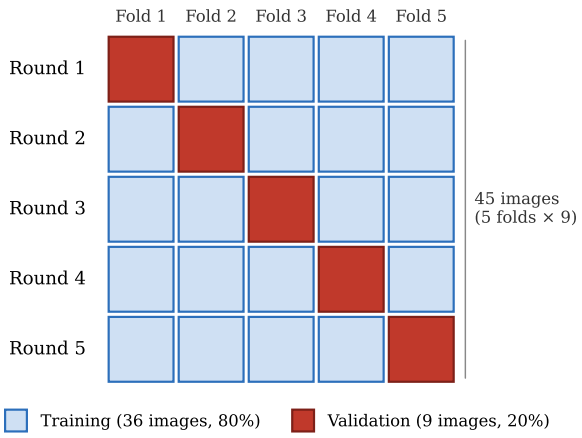


Fig. 2. Schematic of the five-fold cross-validation scheme. The 45 images were split into five folds of nine images each. In every round, four folds (36 images, 80%) were used for training and the remaining fold (9 images, 20%) for validation, so that each fold served once as the validation set.

restricted to a single data partition. In the present study, five training rounds were carried out, each containing nine images.

To avoid bias during cross-validation, each subset was created maintaining a balanced proportion between Ki-67 positive and Ki-67–negative cells, so that the model learned consistently across all training rounds, ensuring greater reliability of the obtained metrics.

The training process was conducted in YOLOv8, with parameters adjusted to balance computational performance and stability. The batch size was set to 16, allowing multiple images to be processed simultaneously and ensuring stability in weight updates. The input image size (imgsz) was set to 640×640 pixels, providing adequate resolution to capture important details without overloading processing time. The initial learning rate (lr0) was set to 0.01, with gradual adjustment during training, helping the model to learn efficiently without converging to suboptimal solutions. Training was performed for 2000 epochs, providing sufficient time for the model to converge and adjust its parameters.

Data augmentation techniques were applied to increase the diversity of the training set. Among them, the RandAugment method introduced random variations such as rotation and brightness/contrast adjustment, increasing model robustness. The Mosaic method was also evaluated in one of the K-fold subsets. However, the configuration using full images showed superior performance, which is why it was adopted as the standard method.

Throughout the entire training process, validation (val) was performed on independent subsets, allowing continuous monitoring of performance metrics (precision, recall, and F1-score) and ensuring good model generalization. This monitoring also made it possible to identify signs of validation regression, preventing overfitting. To mitigate these effects, the training process included adaptive learning rate adjustments and early stopping when no consistent improvement in validation metrics was observed. These mechanisms, combined, were

essential to optimize training and achieve high performance in the object detection task.

III. Results and Discussion

The study addresses a pratical bottleneck in Ki-67 scoring and uses low-cost image acquisition, case-level, cross-validation, and direct comparison between automated and manual proliferation indices. The model training was conducted using cross-validation with an 80/20% proportion, where 80% of the images were used for training and 20% for validation in each round. During the training process, the model showed balanced performance across the folds. Table I summarizes the detection metrics per class and overall, calculated by aggregating the predictions for all annotated cells across the cross-validation folds, rather than averaging the metrics obtained from the individual models. For the non-reactive class, the model achieved a precision of 87.29%, a recall of 84.93%, an accuracy of 75.58%, and an F1-score of 86.10%. For the reactive class, the model achieved a precision of 88.27%, a recall of 86.88%, an accuracy of 77.88%, and an F1-score of 87.57%. Considering all annotated cells jointly, the model achieved an overall precision of 87.54%, a recall of 85.44%, an accuracy of 76.17%, and an F1-score of 86.48%. These results indicate that the model distinguished Ki-67–reactive from Ki-67–non-reactive tumor cells consistently, even under the marked class imbalance between the 7,272 reactive and 21,061 non-reactive annotated cells.

TABLE I
Detection performance per class and overall, calculated from the aggregated predictions across all annotated cells from the five cross-validation folds. The last column reports the total number of annotated cells (ground truth) for each class.

Class	Precision	Recall	F1-score	Cells
Reactive (R)	0.88	0.87	0.88	7 272
Non-reactive (NR)	0.87	0.85	0.86	21 061
Overall	0.88	0.85	0.86	28 333

The learning curves presented in Figures 3, 4, and 5 illustrate the training process, showing the Precision, Recall, and F1-score, respectively. Together, they display a stable behavior throughout the epochs, with no signs of overfitting, and a gradual increase until reaching a plateau, which suggests consistent learning.

The comparison between the estimated proliferation index (YOLOv8) and the reference proliferation index (manual counting) showed high agreement between the methods. The overall means were 24.61% for the estimated index and 24.24% for the reference index, resulting in an absolute mean error of 2.66 percentage points (p.p.) and an average percentage error of 12.45%. The model presented an average error of 0.37 p.p., indicating a slight tendency to overestimate the reference index, and a stable percentage error between 10.8% and 14.2%, suggesting good reproducibility among all groups. A linear regression of the estimated index on the reference index yielded $y = 1.037\,x - 0.493$, with a slope whose 95% confidence interval $[0.971,\ 1.103]$ includes 1

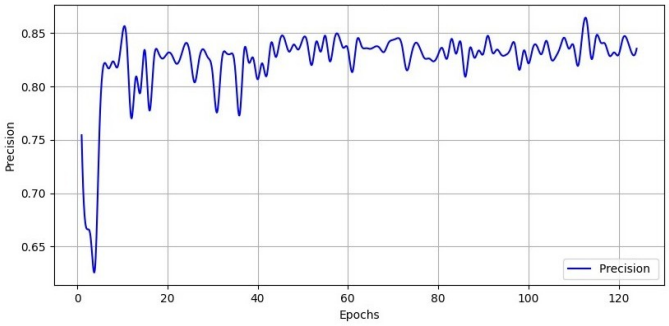


Fig. 3. Precision on the validation data as a function of the training epoch. The metric rises sharply during the first epochs and then stabilizes around a plateau, with no signs of overfitting up to the early-stopping point.

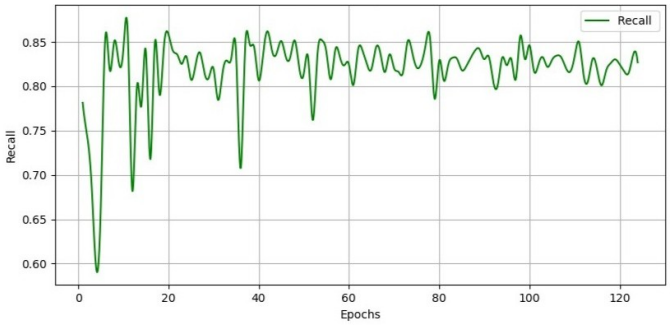


Fig. 4. Recall on the validation data as a function of the training epoch, showing the same rapid initial increase followed by a stable plateau observed for the other monitored metrics.

and an intercept whose interval $[-2.484,\ 1.498]$ includes 0, indicating the absence of statistically significant proportional or fixed bias. The two measurements were strongly correlated (Pearson $R = 0.979$, 95% CI 0.962–0.989, $R^2 = 0.959$), and the intraclass correlation coefficient for absolute agreement was excellent (ICC(2,1) $= 0.978$, 95% CI 0.960–0.988). Fig. 6 illustrates this relationship, showing the proximity between the estimated and reference values.

When compared with other studies that also use AI and deep learning to automate the immunohistochemical analysis of Ki-67, our study shows comparable results. For example,

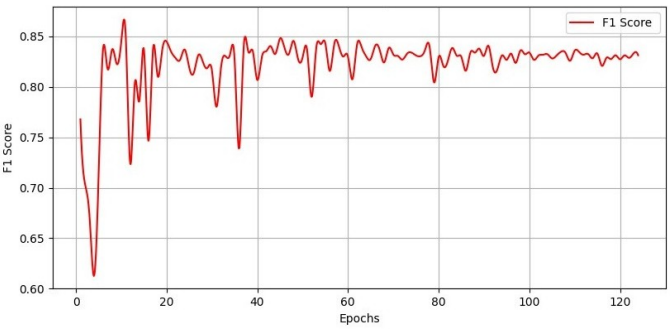


Fig. 5. F1-score on the validation data as a function of the training epoch. As the harmonic mean of precision and recall, it summarizes the joint behavior of both metrics and likewise converges to a stable plateau without overfitting.

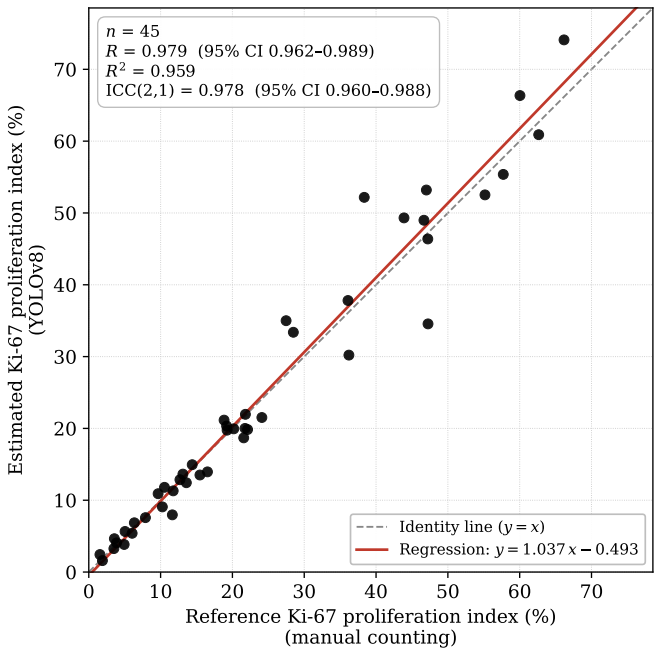


Fig. 6. Scatter plot showing the agreement between the estimated Ki-67 proliferation index (YOLOv8) and the reference Ki-67 proliferation index (manual counting) for the 45 evaluated images. The solid red line is the fitted regression ($y = 1.037\,x - 0.493$), where x is the reference index and y the estimated index. The dashed grey line is the identity line ($y = x$). The two measurements are strongly correlated (Pearson $R = 0.979$, $R^2 = 0.959$) and show excellent absolute agreement (ICC(2,1) = 0.978, 95% CI 0.960–0.988).

in [19], the proposed tool named piNET achieved an F1-score between 0.8 and 0.9 across different types of histological images, within a robust dataset containing color variations and imaging artifacts. In this context, the average F1-score of 0.87 obtained in this study shows that YOLOv8 achieved comparable performance, despite the reduced number of slides and the use of images acquired through a low-cost method.

Likewise, the study by [20] observed that the use of AI-based tools reduced the mean error from 5.9% to 2.1% and increased the Intraclass Correlation Coefficient (ICC) from 0.70 to 0.92 in cell proliferation assessment among pathologists. The present study presented an absolute mean error of 2.66 p.p. and an average error of 0.37 p.p., together with an ICC(2,1) of 0.978 (95% CI 0.960–0.988) between the estimated and reference indices. These values indicate that the automated approach achieved a level of agreement comparable to that obtained by human observers assisted by AI, whose inter-rater agreement reached an ICC of 0.92.

Furthermore, the study by Zehra et al. [21] also employed deep learning techniques for Ki-67 quantification and reported an accuracy of 95% when comparing automatic and manual counts, using images acquired by a camera coupled to the microscope, similar to the acquisition performed in this study. In the present work, the YOLOv8 model reached a precision of 0.88, a recall of 0.86, and an F1-score of 0.87 at the detection level, together with strong agreement at the index level (ICC(2,1) of 0.978 and Pearson correlation of 0.979). A direct numerical comparison with Zehra et al. is limited by the different metrics and evaluation protocols adopted, yet both studies indicate that low-cost, camera-based acquisition can support reliable automated Ki-67 quantification. Any residual difference may be attributed to the particularities of the dataset and the variations in acquisition conditions, since the images used here were captured using a smartphone attached to an optical microscope.

Even so, the closeness between the results reinforces the potential clinical applicability of the proposed model, even in scenarios with limited infrastructure.

A limitation of this study concerns the generalizability of the proposed model, as the dataset was obtained from a single institution and comprised a relatively limited numbers of cases. Although the model showed consistent performance during cross-validation, its performance may be affected by variations in staining protocols, microscopes and image acquisition conditions, which can alter tissue appearance and nuclear characteristics. Therefore, the results should be interpreted within the context of the data used in this study, and future studies should be interpreted within the context larger and more diverse datasets from different institutions and acquisition setting to further assess its robustness and generalizability.

It is worth noting that the adopted approach relies on axis-aligned bounding boxes to represent each nucleus, rather than on pixel-level (semantic or instance) segmentation. In Ki-67 evaluation, tumor nuclei frequently appear overlapped or organized in dense clusters, especially within hotspot regions of high proliferative activity. Under these conditions, the bounding boxes of adjacent nuclei may overlap considerably, which can lead to merged or suppressed detections during non-maximum suppression and, consequently, to counting errors. In the present study, the strong chromatic contrast introduced by the DAB/hematoxylin staining, combined with the automatic hotspot cropping, mitigated part of this difficulty by making reactive and non-reactive nuclei more separable. Nevertheless, the resolution of densely packed nuclei remains an intrinsic limitation of bounding-box–based detection, and future work may explore instance segmentation strategies to delineate individual nuclei more precisely in highly cellular regions.

A further limitation concerns the construction of the reference standard. Although this work is motivated by the intra- and interobserver variability that affects manual Ki-67 assessment, the annotations used here were produced by a single trained annotator, under the supervision of a pathologist. As a consequence, the reference may carry the individual criteria and potential bias of that annotator, and the inter-annotator agreement of the ground truth itself was not quantified. The agreement reported in this study should therefore be interpreted as the concordance between the automated method and one reference standard, rather than against a consensus of multiple independent experts. Future work should incorporate annotations from several specialists, together with a consensus or majority-vote reference, so that inter-annotator variability can be measured and the robustness of the automated estimates

can be assessed against it.

IV. Conclusion

This study proposed a deep learning–based framework for detecting and quantifying the expression of the Ki-67 marker in histological images using the YOLOv8 model. The developed approach demonstrated technical feasibility and clinical potential, especially by employing images captured with low-cost equipment and still achieving results consistent with those reported in the literature. In future work, it is intended to expand the dataset and develop a mobile application that integrates the proposed model, enabling automated analysis directly via smartphone. Such developments may support the broader use of automated Ki-67 analysis, particularly in settings with limited infrastructure, although their clinical applicability will require further validation.

References

[1] Global Cancer Observatory, "Observatório Global do Câncer", World Health Organization, [Online]. Available: https://gco.iarc.who.int/en . Accessed: Sept. 17, 2025.

[2] World Health Organization, "Breast Cancer", [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed: Sept. 17, 2025.

[3] R. Haque, S. A. Ahmed, G. Inzhakova, J. Shi, C. Avila, J. Polikoff, L. Bernstein, S. M. Enger, and M. F. Press, "Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades", Cancer Epidemiology, Biomarkers and Prevention, vol. 21, no. 10, pp. 1848–1855, 2012.

[4] O. Yersal and S. Barutca, "Biological subtypes of breast cancer: Prognostic and therapeutic implications", World Journal of Clinical Oncology, vol. 5, no. 3, p. 412-424, 2014.

[5] N. A. Soliman and S. M. Yussif, "Ki-67 as a prognostic marker according to breast cancer molecular subtype", Cancer Biology and Medicine, vol. 13, no. 4, pp. 496–504, 2016.

[6] E. H. Lips, L. Mulder, J. J. de Ronde, I. A. M. Mandjes, B. B. Koolen, L. F. A. Wessels, S. Rodenhuis, and J. Wesseling, "Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response", Breast Cancer Research and Treatment, vol. 140, no. 1, pp. 63–71, 2013.

[7] B. S. Finkelman, H. Zhang, D. G. Hicks, and B. M. Turner, "The evolution of Ki-67 and breast carcinoma: past observations, present directions, and future considerations", Cancers, vol. 15, no. 3, p. 808, 2023.

[8] M. G. Davey, S. O. Hynes, M. J. Kerin, N. Miller, and A. J. Lowery, "Ki-67 as a prognostic biomarker in invasive breast cancer", Cancers, vol. 13, no. 17, p. 4455, 2021.

[9] E. C. Inwald, M. Klinkhammer-Schalke, F. Hofstädter, F. Zeman, M. Koller, M. Gerstenhauer, and O. Ortmann, "Ki-67 is a prognostic parameter in breast cancer patients: results of a large population-based cohort of a cancer registry", Breast Cancer Research and Treatment, vol. 139, no. 2, pp. 539–552, 2013.

[10] M.-Y. C. Polley, S. C. Y. Leung, L. M. McShane, D. Gao, J. C. Hugh, M. G. Mastropasqua, G. Viale, L. A. Zabaglo, F. Penault-Llorca, J. M. S. Bartlett, et al., "An international Ki67 reproducibility study", Journal of the National Cancer Institute, vol. 105, no. 24, pp. 1897–1906, 2013.

[11] D. M. Louis, L. M. Nair, A. G. Vallonthaiel, M. P. Narmadha, and D. K. Vijaykumar, "Ki 67: a promising prognostic marker in early breast cancer—a review article", Indian Journal of Surgical Oncology, vol. 14, no. 1, pp. 122–127, 2023.

[12] M. Fan, W. Yuan, W. Zhao, M. Xu, S. Wang, X. Gao, and L. Li, "Joint prediction of breast cancer histological grade and Ki-67 expression level based on DCE-MRI and DWI radiomics", IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 6, pp. 1632–1642, 2020.

[13] L. Li, D. Han, Y. Yu, J. Li, and Y. Liu, "Artificial intelligence-assisted interpretation of Ki-67 expression and repeatability in breast cancer", Diagnostic Pathology, vol. 17, no. 1, p. 20, 2022.

[14] M. Pardàs, D. Anglada-Rotger, M. Espina, F. Marqués, and P. Salembier, "Stromal tissue segmentation in Ki67 histology images based on cytokeratin-19 stain translation," Journal of Medical Imaging, vol. 10, no. 3, pp. 037502–037502, 2023.

[15] F. Xing, T. C. Cornish, T. Bennett, D. Ghosh, and L. Yang, "Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki67 images", IEEE Transactions on Biomedical Engineering, vol. 66, no. 11, pp. 3088–3097, 2019.

[16] M. Abubakar, W. J. Howat, F. Daley, L. Zabaglo, L.-A. McDuffus, F. Blows, P. Coulson, H. R. Ali, J. Benitez, R. Milne, et al., "High-throughput automated scoring of Ki67 in breast cancer tissue microarrays from the Breast Cancer Association Consortium", The Journal of Pathology: Clinical Research, vol. 2, no. 3, pp. 138–153, 2016.

[17] M. Feng, Y. Deng, L. Yang, Q. Jing, Z. Zhang, L. Xu, X. Wei, Y. Zhou, D. Wu, F. Xiang, et al., "Automated quantitative analysis of Ki-67 staining and HE images recognition and registration based on whole tissue sections in breast carcinoma", Diagnostic Pathology, vol. 15, no. 1, p. 65, 2020.

[18] Y. Liu, T. Zhen, Y. Fu, Y. Wang, Y. He, A. Han, and H. Shi, "AI-powered segmentation of invasive carcinoma regions in breast cancer immunohistochemical whole-slide images", Cancers, vol. 16, no. 1, p. 167, 2023.

[19] R. S. Gerеad, A. Sivanandarajah, E. R. Brouwer, G. A. Wood, D. Androutsos, H. Faragalla, and A. Khademi, "Pinet—an automated proliferation index calculator framework for Ki67 breast cancer images", Cancers, vol. 13, no. 1, p. 11, 2020.

[20] A. Dy, N.-N. J. Nguyen, J. Meyer, M. Dawe, W. Shi, D. Androutsos, A. Fyles, F.-F. Liu, S. Done, and A. Khademi, "AI improves accuracy, agreement and efficiency of pathologists for Ki67 assessments in breast cancer", Scientific Reports, vol. 14, no. 1, p. 1283, 2024.

[21] T. Zehra, N. Jaffar, M. Shams, Q. Chundriger, A. Ahmed, F. Anum, N. Alsubaie, and Z. Ahmad, "Use of a novel deep learning open-source model for quantification of Ki-67 in breast cancer patients in Pakistan: a comparative study between the manual and automated methods", Diagnostics, vol. 13, no. 19, p. 3105, 2023.

Label Propagation for Missing Data Imputation (LPMD) in Real-life COVID-19 Tabular Data

Filipe Loyola Lopes
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
Centro Universitário Max Planck (UniMAX)
Indaiatuba, Brazil
flopes@cti.gov.br

Arthur Dantas Mangussi
Aeronautics Institute of Technology (ITA)
Sao Jose dos Campos, Brazil
Federal University of São Paulo (UNIFESP)
Sao Jose dos Campos, Brazil
mangussiarthur@gmail.com

Gilson Yuuji Shimizu
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
Centro Universitário de Jaguariúna (UniFAJ)
Jaguariúna, Brazil
gyshimizu@cti.gov.br

Fernando Rezende Zagatti
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
fzagatti@cti.gov.br

Ricardo Cardoso Pereira
Centre for Informatics and Systems of the University of Coimbra (CISUC/LASI)
Department of Informatics Engineering University of Coimbra
Coimbra, Portugal
rdpereira@dei.uc.pt

Miriam Seoane Santos
Department of Computer Science Faculty of Sciences
Porto, Portugal
LIAAD, INESC TEC
Porto, Portugal
miriam.santos@fc.up.pt

Paulo R. Abrão Ferreia
Department of Internal Medicine Division of Infectious Diseases Federal University of Sao Paulo (UNIFESP)
Sao Paulo, Brazil
paulo.abrao@unifesp.br

Pedro Henriques Abreu
Centre for Informatics and Systems of the University of Coimbra (CISUC/LASI)
Department of Informatics Engineering University of Coimbra
Coimbra, Portugal
pha@dei.uc.pt

Ana Carolina Lorena
Computer Science Division Aeronautics Institute of Technology (ITA)
Sao Jose dos Campos, Brazil
aclorena@ita.br

***Abstract*—Missing healthcare data compromises predictive accuracy and introduces model bias. We evaluate Label Propagation for Missing Data Imputation (LPMD), a graph-based semi-supervised method that treats missing values as unlabeled nodes, on four real-world COVID-19 datasets. Benchmarked against MICE, kNN, SAEI, and PMIVAE, LPMD matched the downstream classification performance (AUC, Recall, Specificity) of top-tier baseline methods while achieving up to 5.2 times faster processing times than MICE. LPMD represents an alternative for high-performance clinical data imputation.**

***Index Terms*—Missing data, Imputation, Semi-supervised learning, healthcare data**

I. Introduction

COVID-19 is an infectious respiratory disease that caused the pandemic reported by the World Health Organization (WHO) in March 2020 [1]. The disease quickly spread worldwide and drew the attention of scientists from diverse areas around the world by studying disease mechanisms, prevention and treatments [2]. The pandemic ended in May 2023 [3], and unfortunately, there were more than 7 million deaths worldwide caused by COVID-19 until April 2024 [4]. However, despite its end, it left social consequences and lessons learned [5]. The acquired knowledge, including that in data science applied to medicine, must continue to improve for future adverse public health situations [6], like future pandemics, and current disease problems, such as dengue fever regional epidemics in Brazil [7].

During the pandemic, several studies were dedicated to COVID-19 data analysis and ML techniques for diagnosis and prognosis of the disease [8]. Researchers developed models ranging from rule-based scoring systems to advanced deep-learning ML models. These researches were possible because hospitals produce large amounts of data every day by collecting laboratory tests and medical records, and such data allow extraction of important insights. Although there are different types of healthcare data, the most common approach to developing ML models for healthcare decision making has been tabular data [9].

Despite the large amount of healthcare data generated,

This work was carried out with the support of the São Paulo Research Foundation (FAPESP), Brazil. Process number 2021/06870-3 and 2024/23791-8. This work was also funded by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the research units UID/00326 - Centre for Informatics and Systems of the University of Coimbra, https://doi.org/10.54499/UID/00326/2025 and UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025).

Missing Data (MD) is a common problem reported in the medical domain, which can prevent reliable data analysis [10]. According to the guideline on Statistics Principles for Clinical Trials (ICH E9), MD represent a potential source of bias despite a consensus that there will almost always be MD in healthcare data. Therefore, every effort should be made to minimize the amount of MD and its effects [11].

According to García-Laencina *et al.* [12], we have three different forms to working with MD: (a) Deletion of incomplete data; (b) Model-based procedures; and (c) Imputation (the most popular approach). Imputation consists of "replacing" the missing values according to specific criteria, like statistical methods or imputation based on ML. Statistical imputation methods include techniques such as mean/mode imputation, multiple imputation, and regression imputation. In contrast, ML-based methods encompass approaches like k-nearest Neighbors (kNN), Multiple Imputation by Chained Equations (MICE) and autoencoders (AE) based imputation [13].

ML-based imputation performs various ML predictive tasks in two steps. First, the ML algorithm generates the model using complete data during the learning part. The generated model is then used to predict the missing values for MD. Over the years, several authors have studied different ML-based imputation algorithms, such as kNN, Decision Tree (DT), Random Forests (RF), and Self-Organizing Maps (SOM) imputation. However, these approaches do not always consider intrinsic information that may be related to the performance of algorithms, such as feature distribution. Matching the data distribution to the most appropriate imputation method avoids repeatedly testing many methods, providing a heuristic about the best choice for each study [14]. On the other hand, the MICE imputer, which considers the features distribution, and is considered the state of the art for imputation, does not perform well when applied to high missing rate (MR), which is the reality of healthcare data [15].

A more recent ML imputation methodology is based on autoencoders (AE-based) [15], like the Siamese Autoencoder-based Approach for Imputation (SAEI) [16] and Partial Multiple Imputation with Variational Autoencoders (PMIVAE) [17]. However, AE-based models, especially with deep architectures and customized strategies, can be more computationally intensive than statistical or neighborhood-based methods (e.g., kNN) [16]. In addition, AE-based imputer has limitations such as dependence on complete data [17].

Among the best ML imputers are kNN, and RF, which respect the paradigms of supervised ML [18], and AE-based imputation [13], which respects the unsupervised ML paradigm. The scientific literature widely explored the two most well-known ML paradigms for MD imputation. However, in ML, there is also semi-supervised learning, which has several benefits compared to the other two paradigms. It offers a way to extract maximum benefit from a scarce amount of labelled data while using relatively abundant unlabeled data, and it has been very little explored in the literature on MD imputation [19].

Previous research [20] developed a new imputation method using the Label Propagation algorithm, named Label Propagation for Missing Data Imputation (LPMD), which uses the semi-supervised learning paradigm. This research applied the method to 10 healthcare benchmark datasets and obtained results comparable to the best imputation techniques, while achieving shorter processing times. However, LPMD has not yet been applied to real-life datasets. Testing a new methodology on real-world datasets is important because these datasets present challenges such as inconsistencies, outliers, and class imbalance, which offer a significant research opportunity related to the vast amount of health data available on COVID-19.

Due to the pandemic, healthcare data repositories were created, such as the COVID-19 Data Sharing FAPESP[1] [21] and these data have a high MR [22]. As the literature demonstrates, handling MD in healthcare data is essential for reliable data analysis and ML models [12]. On the other hand, the most widely used imputation algorithms have gaps that need to be improved for practical use in medicine [14]. This study aims to apply a new approach for MD imputation with semi-supervised learning using the Label Propagation algorithm to COVID-19 real datasets and compare with the best imputation methods available on literature. To this end, four datasets were selected, obtained from the COVID-19 Data Sharing FAPESP and pre-processed in previous studies. The aim of this research was to subject real-world COVID-19 datasets to a MD imputation process using an innovative method called LPMD and to compare this method with other state-of-the-art algorithms.

II. Missing Data and Imputation concepts

MD mechanisms are broadly categorized into three distinct types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [11]. As outlined by [23], MCAR represents the highest degree of randomness, occurring when missingness depends on neither observed nor unobserved values. Conversely, MAR allows the missingness to depend on the observed data, provided it remains independent of the missing values themselves. Lastly, MNAR arises when the probability of a value being missing is directly related to the unobserved value of that specific attribute. Importantly, valid data imputation is generally restricted to mechanisms satisfying the MCAR or MAR assumptions.

To establish the formal framework for the MD mechanisms, let X be a dataset structured as an $N \times D$ matrix, where N denotes the number of observations and D represents the number of features. Each entry in the matrix is denoted by x_{ij}, with $i = 1, \ldots, N$ and $j = 1, \ldots, D$. Furthermore, $x_{\cdot j}$ denotes the j-th feature vector, containing all observations for that feature, whereas $x_{i\cdot}$ denotes the i-th observation vector, containing all feature values for that specific instance. Each observation is associated with a target class y_i and a missingness indicator vector $\mathbf{m}i = [mi1, m_{i2}, \ldots, m_{iD}]$.

[1] https://repositoriodatasharingfapesp.uspdigital.usp.br/

Consequently, the MD indicator can be defined as a binary matrix M of size $N \times D$ [24]:

$$m_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is missing,} \\ 0, & \text{otherwise,} \end{cases} \qquad M = \{m_{ij}\}_{i,j=1}^{N,D} \quad (1)$$

where M explicitly maps the positions of missing values within X. Assuming the dataset X can be partitioned into observed (X_{obs}) and missing (X_{miss}) components, and letting ψ represent the parameters of the missingness model, the three classical MD mechanisms are defined as follows.

MCAR (Missing Completely at Random): Under this mechanism, the probability of a value being missing is entirely independent of both observed and unobserved data, as shown in Equation 2:

$$Pr(M = 1|X, \psi) = Pr(M = 1|\psi). \quad (2)$$

Here, the missingness probability depends solely on the parameters ψ [15]. Because MCAR represents the highest level of randomness, standard analytical and inference procedures designed for complete datasets remain valid when applied to the observed data.

MAR (Missing at Random): Missingness depends directly on the observed components of the dataset (X_{obs}), but remains independent of the unobserved values (X_{miss}). Formally, MAR satisfies Equation 3:

$$Pr(M = 1|X, \psi) = Pr(M = 1|X_{obs}, \psi). \quad (3)$$

As noted by [15], missingness correlates strictly with X_{obs}. A typical illustration is patient dropout in longitudinal studies [11]: if two patients share identical covariates and responses up to a given time step, and one drops out, the missing values are considered MAR if the statistical behavior of the dropped-out subject aligns with that of the remaining subject. Valid statistical inferences can still be drawn under MAR without explicitly modeling the missingness process.

MNAR (Missing Not at Random): This mechanism occurs when the missingness depends directly on unobserved values, violating both MCAR and MAR assumptions. Formally, it is expressed as:

$$Pr(M = 1|X, \psi). \quad (4)$$

As highlighted by [15], MNAR typically arises in two scenarios: (a) when missingness depends on the missing value itself, or (b) when it depends on unobserved covariates missing from the dataset. MNAR represents the most challenging setting; addressing it generally requires modeling the underlying reasons for missingness or performing sensitivity analyses under multiple plausible scenarios. In summary, as synthesized by Luengo *et al.* [23], MCAR exhibits complete independence from data values; MAR relies on observed data; and MNAR depends on unobserved variables. Although real-world datasets often reflect complex mixtures of MAR and MNAR, standard imputation strategies perform more reliably under MAR assumptions, whereas applying them to MNAR data can introduce significant estimation bias [25].

Properly handling MD is a crucial step in data science pipelines. In specialized domains such as healthcare, addressing missingness effectively is vital, as biased imputations can misguide clinical decision-making. However, conventional imputation techniques often ignore intrinsic data characteristics—such as feature distributions—that directly influence model performance. Tailoring the imputation strategy to the specific properties of the dataset mitigates the need for exhaustive trial-and-error testing across numerous algorithms [14].

According to [26], the performance of the imputation methods using two main approaches: (i) *imputation quality*, which directly measures reconstruction accuracy by comparing imputed values against the original complete data using metrics such as Mean Absolute Error (MAE) or Root Mean Square Error (RMSE); and (ii) *downstream predictive performance*, which evaluates the impact of the imputed datasets on predictive models using metrics like Area Under the ROC Curve (AUC) or F1-score. In a previous study [20], the LPMD was tested using complete health data (benchmark) containing missing values introduced via amputation. The present study aims to bridge this gap and evaluate downstream predictive performance using real-world data.

III. Related work

To the best of our knowledge, very few studies have addressed the topic of semi-supervised learning applied to the MD imputation process. Among these studies, we highlight the article by [19] that demonstrated the Iterative Robust Semi-Supervised based Imputation (IRSSI) method. Although IRSSI is a semi-supervised methodology, it differs from Label Propagation because while Label Propagation is of the transductive type (network-based learning), IRSSI mixes transduction with induction, which is why it is considered a hybrid algorithm.

Pan et al. (2026) [27] proposed Semi-supervised Generative Adversarial Network (SEGAN), a semi-supervised learning-based imputation method that integrates a generator, a discriminator, and a classifier to jointly exploit the data structure and the available label information. The model also incorporates a missingness hint matrix, enabling the discriminator to more effectively distinguish between observed and imputed values. Experimental results showed that SEGAN outperformed both traditional and deep learning-based imputation methods, achieving lower Root Mean Squared Error (RMSE) values, greater robustness to increasing missingness rates, and improved performance in downstream classification and regression tasks. However, its deep learning-based architecture increases the computational cost and training complexity. In contrast, the Label Propagation method proposed in the present study does not rely on deep learning, resulting in a computationally simpler and potentially more broadly applicable solution for missing data imputation.

IV. Label Propagation in Missing Data Imputation

Label propagation is a classic algorithm for transductive inference on graphs. The primary objective is to assign labels or classes to an unlabeled set of instances by leveraging the topological and semantic structure present in a smaller set of labeled data.

Formally, let $X_L = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ denote the labeled data points with corresponding labels $Y_L = \{y_1, \ldots, y_n\} \in \{1, \ldots, C\}$, where C represents the total number of classes. Let $X_U = \{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}\}$ represent the unlabeled dataset whose corresponding target values $Y_U = \{y_{n+1}, \ldots, y_{n+m}\}$ are unobserved. Typically, $n \ll m$. The joint dataset is given by $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n+m}\} \in \mathbb{R}^D$. The goal is to estimate Y_U from X and Y_L under the smooth manifold assumption, which dictates that nearby data points are likely to share identical labels.

To capture these spatial relationships, a fully connected graph is constructed where the nodes correspond to all instances in X (both labeled and unlabeled). The edge between any two nodes i and j is assigned a weight w_{ij} inversely proportional to their distance d_{ij}. Using a Gaussian Radial Basis Function (RBF) parameterized by σ, the edge weights are computed according to Equation 5 [28]:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = \exp\left(-\frac{\sum_{d=1}^{D}(x_{id} - x_{jd})^2}{\sigma^2}\right). \quad (5)$$

The label assignment of each node propagates to adjacent nodes proportionally to their connecting edge weights; stronger connections yield greater influence during propagation. Information transmission across the graph is modeled via a transition probability matrix $T \in \mathbb{R}^{(n+m)\times(n+m)}$. Each entry T_{ij} denotes the probability of label information transferring from node j to node i, defined as follows [28]:

$$T_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n+m} w_{kj}}. \quad (6)$$

Additionally, a soft-label matrix $Y \in \mathbb{R}^{(n+m)\times C}$ is defined, where the i-th row encodes the class probability distribution for observation $\mathbf{x}_i$. While the rows corresponding to labeled instances are initialized as one-hot vectors representing their ground-truth classes, the rows associated with X_U can be initialized arbitrarily without impacting the theoretical convergence of the algorithm.

The Label Propagation for Missing Data Imputation (LPMD) algorithm adapts the formulation proposed by Zhu and Ghahramani [28]. We model the datasets as proximity graphs and use the graph structure to propagate feature values between connected nodes, which can be weighted and aggregated for MD imputation. There are widespread Label Propagation algorithms capable of assigning labels to unlabeled data, even when the proportion of labeled data from which the information is propagated is relatively low. Therefore, Label Propagation can also be beneficial in MD imputation, especially in cases where one expects high MR [29].

Formally, LPMD receives a dataset represented as a matrix X of size $N \times D$, comprising N observations described by D input features. The dataset X contains missing values in either univariate or multivariate configurations. Initially, the missing values are filled arbitrarily or through a baseline imputation strategy (e.g., mean imputation), yielding a complete matrix denoted by X'. To track missing entries, a binary indicator matrix M of the same dimensions as X is defined, where $m_{ij} = 1$ if the corresponding value is missing and 0 otherwise.

Next, pairwise distances between all instances in X' are computed using a Radial Basis Function (RBF) kernel, as formulated in Equation 5, parameterized by $\gamma = \frac{1}{\sigma^2}$. By default, the hyperparameter is set to $\gamma = \frac{1}{D}$.

Based on the affinity weights w_{ij}, a transition probability matrix T_{ij} is constructed according to Equation 6. The LPMD algorithm then proceeds iteratively through the following steps:

1) **Propagation:** Update the complete matrix X' by propagating information using the transition matrix T:
$$X' = TX'. \quad (7)$$
2) **Reset Observed Data:** Restore the original observed entries from X into X', ensuring that only the missing values (indicated by M) are updated during propagation.
3) **Convergence Check:** Repeat Steps 1 and 2 until the difference between matrices from two consecutive iterations falls below a predefined convergence threshold ϵ.

The pseudocode (Algorithm 1) shows the general functioning of the LPMD algorithm [20]. Under the default initialization scheme, all missing values are set to -1.

Input: Matrix X with missing values; tolerance ϵ.
Output: Imputed matrix X'.
Initialize missing values in X, obtaining X';
Compute similarity matrix W using the RBF kernel (Eq. 5);
Compute the transition matrix (Eq. 6);
repeat
  Propagate values according to Eq. 7;
  Restore observed entries of X;
until *convergence*;
return X'

Algorithm 1: LPMD Imputation

In addition to the baseline framework, structural modifications were incorporated into LPMD, yielding an enhanced variant denoted as LPMD2. This modified version introduces a filtering mechanism prior to the label propagation stage to mitigate the propagation of noisy values and outliers. Specifically, the filtering strategy applies the Interquartile Range (IQR) rule, a computationally efficient method for outlier detection. Furthermore, LPMD2 refines the initial MD completion step by employing a class-wise mean imputation strategy, which yields a more accurate initial approximation of the missing

entries. In unsupervised or unlabeled settings, the data is first partitioned using the k-means algorithm with $k = C$ (where C is the number of clusters/classes), and missing values are subsequently initialized using the corresponding cluster-wise means.

V. Materials and Methods

The proposed method has been implemented in an open-source software package[2], which has been officially registered with the Brazilian National Institute of Industrial Property under registration number BR512026003240-7. The methodology was divided into three main stages: The (A) Data collection step includes the selection of real-life health datasets. The (B) MD imputation step was performed using different algorithms considered state-of-the-art in imputation and the LPMD. Finally, the (C) Evaluation step was performed using the processing time and classification evaluation criteria; we created models from the different imputed datasets and compared the efficiency of each model. The detailed implementation of each step is described in the subsequent subsections.

A. Datasets

We used open data available on the FAPESP COVID-19 Data Sharing/BR repository to assemble datasets for COVID-19 prognosis. Table I summarizes the information from these assembled datasets, and below, we provide more information about each one. The assembled datasets refers to patients who attended by Hospital Sírio-Libanês (HSL) and Hospital Beneficiência Portuguesa (HBP) in São Paulo-Brazil and were tested for COVID-19 during the pandemic. Although we chose open data, the Brazilian Research Ethics Committee approved the research under the "CAAE 37362220.4.0000.5505" registration. Patient consent was waived due to anonymized data being available in an open repository.

Dataset 1 refers to a study of Healthcare-Associated Infection (HCAI) among patients diagnosed with COVID-19. This dataset was assembled from patients' data at the HSL and was made available in previous research we conducted [30]. Dataset 2 deals with a study on the COVID-19 prognosis related to the worsening of the disease. This dataset was also assembled from data from HSL patients and made available in previous research we conducted [31]

Datasets 3 and 4 were assembled from data from patients at the HSL and HBP, respectively. The research deals with a data-centric perspective approach and was made available by previous research [32].

Statistical tests were applied to classify the missing data mechanism in each dataset using the missr package in R software. Little's MCAR test [33] was first employed to assess the null hypothesis that the MD mechanism are MCAR. If the mechanism were not classified as MCAR, a multivariate logistic regression [34] was fitted for each variable with missing values, considering as the response a binary indicator of missingness (0 or 1), and as predictors all other observed variables. This procedure was repeated for each variable with missing values, and the p-values of the regression coefficients were stored. If the smallest p-value was statistically significant, evidence was found that missingness depends on at least one observed variable, and the MD mechanism were classified as MAR. MD mechanism were considered MNAR if they were neither MCAR nor MAR.

[2]https://github.com/filipeloyola/LPMD-applied-COVID-19-datasets

B. Imputation

As with the benchmark datasets, in the case of real datasets, we also performed the imputation process with the Label Propagation algorithms (LPMD and LPMD2) and the state-of-the-art algorithms that served as a baseline. The hyperparameter of the LPMD and LPMD2 algorithms was ϵ equal $1e^{-32}$, representing a value that tends to zero. The baselines used to compare LPDM and LPMD2 were Mean, kNN, MICE, SAEI and PMIVAE. All strategies perform data imputation in the training subsets and use the extracted patterns to perform imputation in the test subsets, preventing data leakage.

C. Evaluation

We performed computing time and classification models evaluation. In the quality classifier evaluation, we used models produced with the Gradient boosting (GB), Random Forests (RF), and Support Vector Machine (SVM) algorithms. We evaluated each model's predictive performance based on the AUC, Recall, and Specificity metrics. RF and GB used a maximum depth of ten. The results of classification were submitted to a Friedman statistical test for multiple comparisons, with the Nemenyi post-hoc test, a nonparametric statistical test used in comparisons between multiple algorithms [35]. The procedure used SCMAMP library from the R language [36].

VI. Results and Discussion

Considering the four real datasets, the seven imputation methods, and the three classification quality metrics used, we computed 84 classification results. It is worth noting that for each result, we have the mean and standard deviation values obtained from the five cross-validation rounds. Next, we address the results of the processing time evaluation and assessments based on classification.

Time evaluation: Table II presents the sum of processing times of all imputation algorithms for each real dataset. LPMD2 was advantageous in this aspect, especially compared to SAEI, PMIVAE, and MICE. It is also possible to observe that the LPMD processing time was smaller than kNN for datasets 1, 2 and 4 and close to kNN for dataset 3, proving to be an interesting algorithm in terms of processing time.

Classification models evaluation: Figure 1 shows a compiled result, considering all models created with the GB, RF and SVM algorithms. This analysis shows the classification performance achieved from the datasets imputed by the different imputation methods (kNN, LPMD, MICE, Mean, PMIVAE and SAEI).

TABLE I: Real datasets used in the experiments.

Dataset	MR%	# Instances	# Classes	# Features	
				Cat.	Cont.
1: HSL-HCAI-INFECTION	13.5	499	2	2	36
2: HSL-COVID-PROGNOSTIC	13.2	4320	2	2	29
3: HSL-DATA-CENTRIC	6.6	1432	2	2	18
4: HBP-DATA-CENTRIC	9.0	185	2	2	18

Legend: MR - Missing rate.

TABLE II: Total processing time for each real dataset (in seconds).

Algorithm	Dataset 1	Dataset 2	Dataset 3	Dataset 4
kNN	198.0	864.7	837.2	242.7
MICE	249.7	4407.8	1531.5	248.2
PMIVAE	348.7	2165.8	2684.3	466.4
SAEI	414.9	1440.9	1562.1	408.4
LPMD	192.2	847.1	941.8	240.6
LPMD2	**186.1**	**756.3**	**749.3**	**210.4**

Legend: The lowest time value for each dataset is marked in bold.

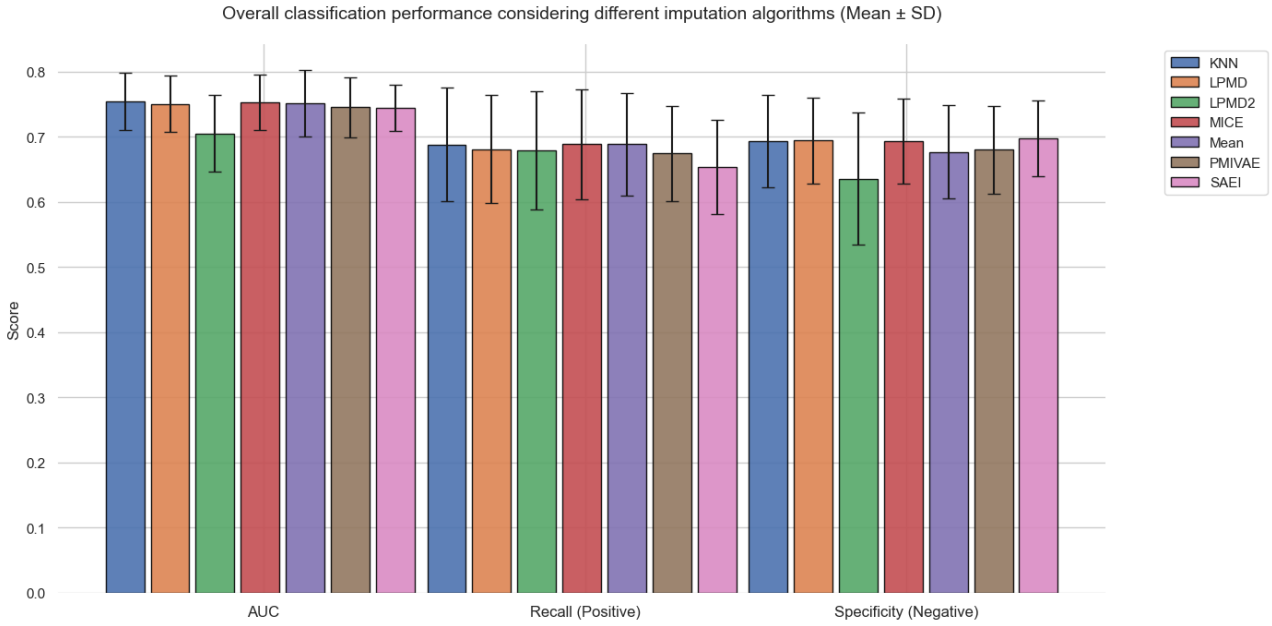


Fig. 1: Classification evaluation for all machine leaning model and all datasets, considering each imputation algorithm.

From Figure 1, it is possible to notice that the result obtained from the LPMD algorithm is comparable with the results obtained by the kNN and MICE algorithms for tree metrics. However, it is possible to observe that the result obtained from LPMD2 is inferior to the others by the AUC and Specificity metrics. It is also possible to verify a result inferior to the others in the results obtained by the SAEI algorithm considering the Recall metric.

In Figure 2, we provide a statistical analysis under an overview of the classification analysis, bringing together the results of all models, all datasets and all metrics. The statistical analysis available in Figure 2 provides a graph known as Critical Difference (CD), used to compare algorithms. This graph shows an average ranking of the algorithms, with the best ones being further to the left and the worst ones further to the right. In this evaluation, the best algorithms were MICE and LPMD. There was no statistical difference between the algorithms MICE, LPMD, kNN, Mean, SAEI and PMIVAE, as they are connected by the same horizontal bar. The worst algorithm was LPMD2 (further to the right), presenting a statistical difference in relation to the other imputers.

In this research, we applied only the time evaluation and classification models evaluation, and we did not use quality imputation evaluation since we used real datasets. Real datasets are incomplete, so they do not have reference values for calculating the imputation error. It is worth mentioning that we use three classic classification algorithms for indirect evaluation, notably RF, GB and SVM. These algorithms are known in the literature as Shallow Learning; they do not use Deep Learning. Shallow learning algorithms have some advantages, such as shorter processing time and greater interpretability of the models.

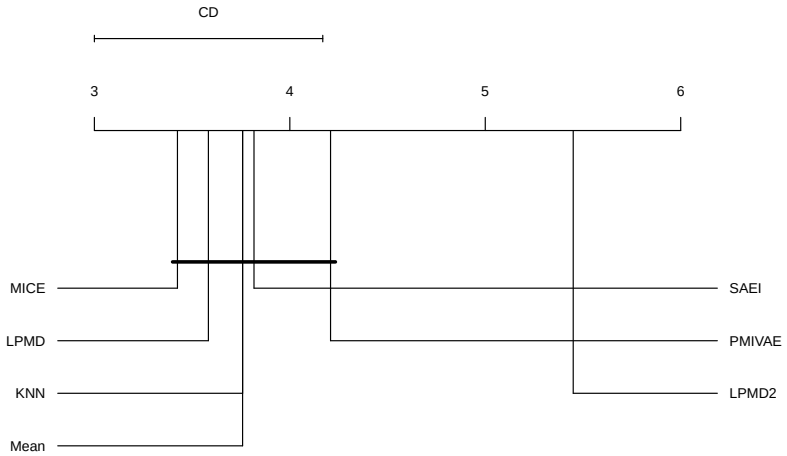


Fig. 2: Classification evaluation for all machine leaning model, all datasets, and all metrics, considering each imputation algorithm.

LPMD2 algorithm stood out significantly in time analysis, presenting a shorter processing time than the other imputation algorithms. The shortest processing time is due to using a filter in the algorithm. This filter eliminates outliers when propagation labels, allowing the propagation to converge more quickly than LPMD. However, considering the classification analysis, LPMD2 did not perform as well as expected. When performing a new time comparison without considering LPMD2, it is possible to see that LPMD obtained the best processing time for Datasets 1, 2 and 4. Only for Dataset 3 was LPMD's processing time (941 seconds) longer than kNN's (837 seconds). It is worth noting that LPMD achieved a shorter processing time (for Dataset 3) compared to the other imputation algorithms (other than kNN), while also showing promising results in the classification analysis.

The results shown in Figure 1 show an overview of the results for all datasets and an average for all classification models. This analysis shows that the average of the LPMD classification results is very close to the other imputation algorithms for AUC, Recall and Specificity, which is an exciting result. However, it is possible to observe lower values for LPMD2 for AUC and Specificity. The AUC metric is widely used in classification analyses, and it means the area under the ROC curve (AUC). It is very significant when the data have balanced classes. In the health domain, however, the existence of unbalanced classes is very common, which can be seen in the datasets used in this research. Therefore, together with AUC, it is necessary to use Recall, which verifies the hit rate of true positives and Specificity, which verifies the rate of true negatives. It is worth noting that for the medical field, values above 80% for these metrics are considered exciting.

The analysis shown in Figure 2, considering an overview of all results, shows that LMPD stood out as the second best algorithm, very close to MICE, considered the state of the art in missing data imputation. This result is very interesting considering that LPMD obtained a better processing time, up to 5.2 times shorter than MICE. In addition, the tested datasets have up to 13.5% missing data. New research will need to be carried out with higher missing data values, considering real datasets, since theory shows that LPMD tends to work better for higher MD [20].

Regarding the MD mechanism analysis, overall, the four datasets analyzed showed a predominance of the MAR mechanism, as indicated by the tests performed ($p < 0.05$). Only the "Urine Leukocytes" variable, from the HCAI-INFECTION dataset, showed evidence consistent with an MNAR mechanism. Future research may use the MD mechanism analysis methodology to evaluate the performance of different imputers for each mechanism type.

VII. Conclusions and future work

This study addresses healthcare data imputation by evaluating LPMD and its variant LPMD2, semi-supervised Label Propagation algorithms that treat missing values as unlabeled instances. Evaluated on real-world COVID-19 datasets, LPMD achieved classification performance competitive with top-performing benchmarks at low execution time, whereas LPMD2 showed limited predictive utility. LPMD offers an efficient, viable solution for clinical data imputation, though broader validation remains recommended. For future work, we plan to evaluate LPMD on larger and more diverse healthcare datasets to further assess its scalability and predictive performance. Additionally, we intend to explore alternative graph construction kernels and edge weight thresholding strategies. Finally, benchmarking LPMD against advanced imputers such as MissForest, and emerging Large Language Model (LLM)-based imputation techniques, represents a promising direction for future research.

Disclosure of Generative AI Usage

No generative AI system was used to generate experimental results, fabricate data, or perform the scientific evalua-

tion presented in this work. During the preparation of this manuscript, generative AI tools were used exclusively to assist with language refinement, grammar correction, and text clarity improvement.

REFERENCES

[1] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *International Journal of Surgery*, vol. 76, pp. 71–76, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1743919120301977

[2] "Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19): A Review," *JAMA*, vol. 323, no. 18, pp. 1824–1836, 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.6019

[3] R. Sarker, A. S. M. Roknuzzaman, Nazmunnahar, M. Shahriar, M. J. Hossain, and M. R. Islam, "The WHO has declared the end of pandemic phase of COVID-19: Way to come back in the normal life," *Health Sci. Rep.*, vol. 6, no. 9, p. e1544, Sep. 2023.

[4] Worldometers, "Covid-19 coronavirus pandemic," https://www.worldometers.info/coronavirus/, August 2025.

[5] H. Alizadeh, A. Sharifi, S. Damanbagh, H. Nazarnia, and M. Nazarnia, "Impacts of the covid-19 pandemic on the social sphere and lessons for crisis management: a literature review," *Natural Hazards*, vol. 117, no. 3, pp. 2139–2164, Jul 2023. [Online]. Available: https://doi.org/10.1007/s11069-023-05959-2

[6] J. D. Auerbach, A. D. Forsyth, C. Davey, and J. R. Hargreaves, "Living with covid-19 and preparing for future pandemics: revisiting lessons from the hiv pandemic," *The Lancet HIV*, vol. 10, no. 1, pp. e62–e68, Jan 2023. [Online]. Available: https://doi.org/10.1016/S2352-3018(22)00301-0

[7] C. Barcellos, V. Matos, R. M. Lana, and R. Lowe, "Climate change, thermal anomalies, and the recent progression of dengue in brazil," *Scientific Reports*, vol. 14, no. 1, p. 5948, Mar. 2024.

[8] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *bmj*, vol. 369, 2020.

[9] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222004349

[10] G. Nagarajan and L. D. Dhinesh Babu, "Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty," *Artificial Intelligence in Medicine*, vol. 123, p. 102214, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365721002074

[11] T. Burzykowski, J. Carpenter, C. Coens, D. Evans, L. France, M. Kenward, P. Lane, J. Matcham, D. Morgan, A. Phillips, J. Roger, B. Sullivan, I. White, and L.-M. Yu, "Missing data: Discussion points from the PSI missing data expert group," *Pharmaceutical statistics : the journal of the pharmaceutical industry*, vol. 9, no. 4, pp. 288–297, 2010.

[12] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, Mar. 2010.

[13] R. C. Pereira, M. S. Santos, P. P. Rodrigues, and P. H. Abreu, "Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes," *Journal of Artificial Intelligence Research*, vol. 69, pp. 1255–1285, 2020.

[14] M. S. Santos, J. P. Soares, P. Henriques Abreu, H. Araújo, and J. Santos, "Influence of Data Distribution in Missing Data Imputation," in *Artificial Intelligence in Medicine*, A. ten Teije, C. Popow, J. H. Holmes, and L. Sacchi, Eds. Cham: Springer International Publishing, 2017, pp. 285–294.

[15] R. C. Pereira, P. H. Abreu, P. P. Rodrigues, and M. A. Figueiredo, "Imputation of data missing not at random: Artificial generation and benchmark analysis," *Expert Systems with Applications*, vol. 249, p. 123654, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424005207

[16] R. C. Pereira, P. H. Abreu, and P. P. Rodrigues, "Siamese autoencoder architecture for the imputation of data missing not at random," *Journal of Computational Science*, vol. 78, p. 102269, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877750324000620

[17] ——, "Partial multiple imputation with variational autoencoders: Tackling not at randomness in healthcare data," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4218–4227, 2022.

[18] E. Tlamelo, Maupong Thabiso, M. Dimane, Semong Thabo, M. Banyatsang, and T. Oteng, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021.

[19] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation," *IEEE Access*, vol. 8, no. 1, pp. 90 555–90 569, 2020.

[20] F. L. Lopes, A. D. Mangussi, R. C. Pereira, M. S. Santos, P. H. Abreu, and A. C. Lorena, "A label propagation approach for missing data imputation," *IEEE Access*, vol. 13, pp. 65 925–65 938, 2025.

[21] L. E. Mello, A. Suman, C. B. Medeiros, C. A. Prado, E. G. Rizzatti, F. L. S. Nunes, G. F. Barnabé, J. E. Ferreira, J. Sá, L. F. L. Reis, L. V. Rizzo, L. Sarno, R. de Lamonica, R. M. d. B. Maciel, R. M. Cesar-Jr, and R. Carvalho, "Opening Brazilian COVID-19 patient data to support world research on pandemics," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3966427

[22] F. L. Lopes, R. L. I. Junior, V. C. N. de Faria, F. A. N. Verr, and A. C. Lorena, "A network modeling and analysis of covid-19 hospital patient data," no. 28. Springer, 2022. [Online]. Available: https://link.springer.com/book/9783031147623

[23] J. Luengo, S. García, and F. Herrera, "A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method," *Neural Networks*, vol. 23, no. 3, pp. 406–418, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608009003165

[24] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11 651–11 667, 2019.

[25] R. D. C. Pereira, "Artificial intelligence strategies for missing data imputation using healthcare data," Ph.D. dissertation, PhD thesis, Universidade de Coimbra, 2023.

[26] A. D. Mangussi, M. S. Santos, F. L. Lopes, R. C. Pereira, A. C. Lorena, and P. H. Abreu, "mdatagen: A python library for the artificial generation of missing data," *Neurocomputing*, vol. 625, p. 129478, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122500150X

[27] X. Pan, W. Wu, L. Chen, Z. Li, P. Lu, P. Cao, and Y. Wu, "Segan: A semi-supervised learning method for missing data imputation," *Service Oriented Computing and Applications*, vol. 20, no. 2, pp. 225–237, Jun 2026. [Online]. Available: https://doi.org/10.1007/s11761-025-00472-9

[28] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Center for Automated Learning and Discovery, Pittsburgh, PA, Technical Report CMU-CALD-02-107, 2002.

[29] T. Swearingen and A. Ross, "Label propagation approach for predicting missing biographic labels in face-based biometric records," *IET Biometrics*, vol. 7, no. 1, pp. 71–80, 2018. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2017.0117

[30] F. L. Lopes, P. R. A. Ferreira, and A. C. Lorena, "Healthcare-associated infection prediction on hospitalized covid-19 patients," in *Proceedings of the XXIX Brazilian Congress on Biomedical Engineering*, ser. IFMBE Proceedings. Springer, Cham, 2025.

[31] F. L. Lopes, A. S. oes Gaspar, and A. C. Lorena, "Analysis of covid-19 severity prognosis using hospital data," no. 28. ABEPRO, 2022. [Online]. Available: http://portalabepro.educacao.ws/ijcieom/restrito/arquivos/icieom2022/ABST_0026_37576.pdf

[32] M. G. Valeriano, J. L. J. Pereira, C. R. Veiga Kiffer, and A. C. Lorena, "Explaining instances in the health domain based on the exploration of a dataset's hardness embedding," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2024, pp. 1598–1606.

[33] R. J. A. Little, "A test of missing completely at random for multivariate data with missing values," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988.

[34] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. John Wiley & Sons, 2000.

[35] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[36] B. Calvo and G. Santafé Rodrigo, "scmamp: Statistical comparison of multiple algorithms in multiple problems," *The R Journal*, vol. 8, pp. 248–256, 2016.

Limits of Lightweight Hallucination Detection in Medical LLMs: A Lexical Bias Analysis of MedHal Across Clinical Specialties

Guilherme Lopes dos Santos*, Erica de Castro Silveira*, Paulo Victor Costa Ferreira*, Francisco Leudes Bezerra Neto*
* *Universidade Federal do Ceará*
Quixadá, Ceará, Brasil
guilhermelopess@alu.ufc.br, ery.cxtr@alu.ufc.br
paulovictorcosta@alu.ufc.br, leudesz@alu.ufc.br

***Abstract*—The detection of hallucinations in medical language models is a critical task for the safe use of artificial intelligence-based systems in healthcare. However, it is not yet clear to what extent lightweight methods, such as embedding similarity, natural language inference, and signal fusion, are capable of distinguishing factual statements from hallucinated ones in medical benchmarks. In this work, we investigate the limits of these methods on MedHal, an evaluation dataset aimed at detecting factual hallucinations in medicine. Furthermore, we enriched the dataset with medical specialty labels using a classifier trained on MedMCQA, enabling a stratified analysis by clinical domain. We evaluated approaches based on embeddings, textual inference models, hybrid methods, and an LLM-as-a-judge using Qwen2.5-7B. The results show that similarity-based methods remain close to chance, while textual inference models present modest gains. The LLM-as-a-judge achieved the best overall performance, with an AUC of 0.823, but still presented relevant limitations, especially in false negatives. The lexical analysis revealed that hallucinated statements can be closer to the source than factual statements, due to the minimal substitution mechanism present in the benchmark. In addition, the specialty analysis showed significant performance variation across medical areas. These findings indicate that medical hallucination detection must be treated as a factual verification task, rather than merely as a semantic similarity problem.**

***Index Terms*—medical hallucination detection, large language models, MedHal, lexical bias, clinical specialties, factual verification**

I. Introduction

Large language models have demonstrated increasingly impressive performance in biomedical and clinical tasks, including medical question answering, scientific text summarization, and support for health information interpretation. Recent studies indicate that these models are capable of encoding clinical knowledge and achieving competitive results on medical benchmarks, but they also emphasize that healthcare applications require rigorous evaluation of factuality, reasoning, potential harm, and bias [1]. In this context, the generation of incorrect, inaccurate, or unsupported information represents a critical problem. This phenomenon, known as hallucination, has been widely discussed in the natural language generation literature and remains one of the main challenges for the reliable use of LLMs [2].

The automatic detection of medical hallucinations thus emerges as an important step to increase the reliability of systems based on language models. However, this task is not limited to verifying whether two sentences are semantically similar. Previous work on factuality and faithfulness in text generation shows that an output can appear fluent and close to the reference text, but still contain unsupported or factually incorrect information [3]. Similarly, recent factuality evaluations argue that verification must consider whether specific claims are supported by reliable sources, and not just whether there is global textual proximity between generation and reference [4].

The MedHal benchmark was proposed to evaluate factual hallucination detection in the medical domain [5]. The dataset gathers statements derived from different types of biomedical sources, including scientific articles summarized from SumPubMed [6], medical multiple-choice questions from MedMCQA [7] and MedQA [8], as well as augmented clinical notes from the ACM dataset [9]. Other medical hallucination benchmarks have also been proposed, such as Med-HALT [18], which focuses on reasoning and memory tests derived from medical examinations, whereas MedHal is centered on factual verification through minimal substitution. This diversity makes MedHal a relevant resource for evaluating detection methods in scenarios with different textual registers, formality levels, and types of medical knowledge.

Despite its relevance, MedHal also presents important methodological challenges. In particular, some hallucinations are constructed by minimal substitutions of critical terms, preserving the original sentence's structure and altering only a single factual element. This process can make the hallucinated statement lexically very close to the source, while a factual statement may appear as a paraphrase with less word overlap. Furthermore, the dataset does not directly provide the medical specialty associated with each instance, which limits stratified analyses by clinical domain.

In this work, we investigate the limits of lightweight meth-

ods for hallucination detection in medical language models, focusing on three main questions: (i) to what extent methods based on similarity, textual inference, and signal combination can differentiate factual statements from hallucinated statements in MedHal; (ii) how the benchmark's lexical bias affects the performance of these methods; and (iii) whether the detection difficulty varies across different medical specialties. To enable this stratified analysis, we enriched MedHal with medical specialty labels using a classifier trained on MedMCQA, which contains questions annotated by medical disciplines [7].

The results show a clear performance gradient. Embedding-based methods remain close to chance, textual inference models present modest gains, and the use of an LLM as a judge achieves substantially superior performance. Still, even the best evaluated method presents relevant limitations, especially in terms of false negatives, that is, cases where a hallucination is classified as factual. Moreover, the specialty analysis reveals that detection does not have uniform difficulty across medical areas, being more favorable in domains of more canonical knowledge and more challenging in clinical and surgical areas.

The main contributions of this work are: (i) a comparative evaluation of lightweight methods and an LLM-as-a-judge for hallucination detection on MedHal; (ii) the identification and analysis of a lexical bias that favors the textual proximity of hallucinated statements to the source; (iii) the enrichment of the dataset with medical specialties for stratified analysis; and (iv) a discussion on the risks of interpreting aggregated metrics as sufficient evidence of safety in medical scenarios. With this, the study reinforces that medical hallucination detection must be treated as a factual verification problem, and not merely as a semantic similarity task.

II. Related Work

A. Hallucination and Factuality in Text Generation

Hallucination in language models, that is, the generation of fluent but incorrect or unsupported content, is recognized as one of the main obstacles to the reliable use of LLMs [2]. Works on faithfulness and factuality in abstractive summarization have shown early on that an output can be textually close to the reference and still be factually incorrect [3], which motivated metrics that evaluate the support of specific claims by reliable sources, rather than mere global textual proximity [4]. This distinction between textual proximity and factual support is central to the problem investigated in this work.

B. Hallucination Detection Methods

Approaches for automatic hallucination detection can be organized into families with distinct costs and assumptions. Methods based on embeddings and semantic similarity measure the proximity between the generation and the source in a vector space, assuming that factual content would be closer to the evidence [10], [11]. Methods based on natural language inference (NLI) treat verification as a textual entailment problem, classifying whether the source entails, contradicts, or is neutral towards the statement [12]; this formulation better captures negation and contradiction than pure similarity. More recently, the LLM-as-a-judge paradigm, in which a language model directly evaluates factual consistency, has been employed in verification tasks [13]. This paradigm, however, is a target of criticism regarding biases (position, verbosity, and self-preference biases) and limitations in reasoning and reproducibility, especially when applied to subjective quality evaluation [13], [19]. In our work, we employ the judge in a formulation that reduces some of these concerns: a binary classification verifiable against the benchmark's ground truth, with an open model and deterministic generation. However, because it is an LLM acting as an evaluator, biases cannot be entirely ruled out. We comparatively evaluate these three families, in addition to a hybrid combination, under the same protocol, which allows us to situate their relative performances in a controlled manner.

C. Evaluation of LLMs and Hallucination in the Medical Domain

In the medical domain, it has been shown that LLMs encode relevant clinical knowledge and achieve competitive results on benchmarks, but healthcare applications require rigorous evaluation of factuality, reasoning, and risk [1]. Several datasets have been proposed to evaluate medical knowledge and reasoning through multiple-choice questions, such as MedMCQA [7] and MedQA [8], but such resources were not designed for factual hallucination detection. MedHal [5] fills this gap by gathering statements from heterogeneous biomedical sources (scientific articles [6], multiple-choice questions [7], [8], and augmented clinical notes [9]) labeled for veracity. However, MedHal does not provide the medical specialty of each instance, which prevents stratified analyses by clinical domain, and its hallucination generation mechanism via minimal substitution introduces specific characteristics that motivate a dedicated analysis. Our work addresses these two points: we enrich MedHal with specialty labels, enabling stratified analysis, and we investigate the effect of the benchmark's lexical bias on detection methods.

III. Materials and Methods

A. Dataset

We used MedHal [5], a factual hallucination detection *benchmark* in medicine. Each instance consists of a medical statement (`question` field), a binary veracity label (`correct_answer`: factual or hallucinated), and a reference textual explanation (`explanation`) that justifies the label based on the source context. The statements are derived from four distinct sources, scientific articles (SumPubMed [6]), medical multiple-choice questions (MedMCQA [7] and MedQA [8]), and clinical reports (ACM [9]), providing register and domain diversity. Hallucinations in MedHal are predominantly generated by minimal substitution of entities or values: the hallucinated statement preserves the structure of the correct sentence, altering only a factually critical term (e.g., substituting "apical periodontitis" with "cervical resorption").

It should be noted that, in the evaluated subset, the explicit context field was mostly empty; therefore, we adopted the `explanation` field as the ground truth anchor for the detection methods, as it provides the textual justification associated with the factual label. This choice was necessary to enable a consistent evaluation across all instances; however, we acknowledge that using explanations instead of the original source context may influence the measured performance, since the available evidence is derived from the benchmark annotation process rather than the complete source documents.

B. Medical Specialty Enrichment

MedHal does not provide the medical specialty of the statements, which is relevant information for analyzing hallucination detection stratified by clinical domain. To bridge this gap, we trained a specialty classifier and applied it to the dataset.

Training data. We used MedMCQA [7], which contains 194 thousand questions annotated with 21 medical disciplines (`subject_name` field), used directly as the classification targets. A normalization function (conversion to lowercase, removal of accents, parentheses, and acronyms) handled spelling variations across the dataset (e.g., "Anesthesia" vs. "Anaesthesia").

Model and training. We adopted PubMedBERT [14], specifically the `abstract-fulltext` checkpoint, pretrained on PubMed *abstracts* and full texts, with a linear classification layer ($768 \rightarrow 21$) on top of `AutoModelForSequenceClassification`, from the *Transformers* library [15]. The classifier's input is exclusively the question text, without alternatives or explanations—a deliberate decision because, when applied to MedHal, only the statement is available; training with more context than what will be present at inference would artificially inflate accuracy. To mitigate the imbalance between disciplines, we applied *undersampling* with a maximum of 15,000 examples per class. The training used 3 epochs, a *batch size* of 32 (train) and 64 (evaluation), a learning rate of 2×10^{-5}, a *warmup ratio* of 0.06, a *weight decay* of 0.01, mixed precision (fp16) [16], and a fixed seed of 42. The best *checkpoint* was selected by weighted F1 on the official MedMCQA validation set.

Application and confidence filter. The classifier was applied to MedHal statements, assigning each instance a predicted specialty and a confidence score. For the specialty-stratified analysis, we retained only predictions with confidence ≥ 0.80 to reduce noise in specialty attribution. After filtering, the enriched dataset contained 9,275 statements. Areas with an insufficient number of instances after the filter (e.g., *Skin*, with only two examples) were excluded from the specialty-stratified analysis.

This confidence threshold was applied exclusively to the specialty attribution step and was not used to modify the hallucination detection task itself. However, we acknowledge that this filtering strategy may introduce selection bias in specialty-level results, as instances with uncertain specialty predictions are excluded from the stratified analysis.

C. Detection Methods

We evaluated three main families of methods and one hybrid combination strategy for the task of classifying a statement as factual or hallucinated, given the reference source. The neural models were implemented using the *Transformers* library [15].

Embedding similarity. We calculated the cosine similarity between the embeddings of the statement and the source. We tested two models: multilingual-e5-large [10] (with the recommended `query:`/`passage:` prefixes) and BGE-large-en-v1.5 [11]. The underlying hypothesis is that factual statements would be semantically closer to the source than hallucinated ones.

Natural language inference (NLI). We employed cross-encoder NLI models that receive the source as a premise and the statement as a hypothesis, producing probabilities of entailment, contradiction, and neutrality. The models are based on the DeBERTa-v3 architecture [12]. We adopted the difference $P(\text{entailment}) - P(\text{contradiction})$ as the factuality score and tested the base and large versions.

LLM-as-a-judge. We used Qwen2.5-7B-Instruct (`Qwen/Qwen2.5-7B-Instruct`) [17], loaded with 4-bit NF4 quantization using the BitsAndBytes library, as a factual consistency judge. The model receives the source and the statement through a structured medical fact-checking prompt and returns, in strict JSON format, a verdict (FACTUAL or HALLUCINATION) and a confidence degree (0–100). The generation is deterministic (`do_sample=False`), ensuring reproducibility. To obtain a continuous score suitable for AUC computation, we mapped the verdict and confidence onto a factuality probability in $[0, 1]$: a FACTUAL verdict with confidence c (normalized to $[0, 1]$) yields a score of $0.5+0.5c$, while a HALLUCINATION verdict yields $0.5 - 0.5c$. Thus, a high-confidence FACTUAL verdict approaches 1, a high-confidence HALLUCINATION verdict approaches 0, and low-confidence verdicts of either kind remain near 0.5. Unlike the use of LLM-as-a-judge for subjective quality evaluation—which is a known target of criticism regarding bias and reproducibility—our use is a binary classification verifiable against the benchmark's ground truth, with an open model and deterministic output.

Hybrid methods. We combined embedding similarity and NLI signals through logistic regression. The input features consisted of the similarity scores from multilingual-e5-large and BGE-large, together with the factuality scores from DeBERTa-v3-base and DeBERTa-v3-large. The classifier was evaluated using 5-fold cross-validation, with probability outputs used as factuality scores for AUC computation.

D. Evaluation Protocol

The primary metric is the area under the ROC curve (AUC), treating detection as a ranking problem between the factual and hallucinated classes. Confidence intervals were estimated using non-parametric bootstrap with 2,000 resamples and a fixed random seed of 42. The 95% confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. For all scoring-based methods, scores

were oriented as factuality scores, where higher values indicate stronger evidence of factual consistency. For the LLM-as-a-judge, the continuous factuality score described above was used for AUC, while the discrete verdict was used for accuracy, precision, recall, and F1-score, as well as the confusion matrix. All methods were evaluated on the same set of 9,275 statements and the same ground truth, ensuring direct comparability. Additionally, we analyzed the judge's performance stratified by medical specialty and investigated the lexical overlap (Jaccard index) between the statement and the source, separated by class, to characterize the mechanism underlying the methods' performance.

IV. Results

A. Performance of Detection Methods

Table I presents the AUC of all evaluated methods on the 9,275 statements, with 95% confidence intervals estimated via bootstrap. A clear performance gradient is observed. The methods based on embedding similarity lie at the chance level: e5-large reaches an AUC of 0.504 (95% CI: 0.492–0.515), with the interval including 0.5, being statistically indistinguishable from a random classifier; BGE-large shows a marginally superior performance (0.517; 95% CI: 0.504–0.529), still very close to chance.

The NLI methods outperform pure similarity. The nli-deberta-v3-base achieves an AUC of 0.560 (95% CI: 0.548–0.571), and the large version raises the performance to 0.636 (95% CI: 0.625–0.647). The signal fusion via logistic regression (hybrid) reaches 0.646 (95% CI: 0.635–0.656), the best result among the lightweight methods—still, a modest ceiling.

The LLM-as-a-judge (Qwen2.5-7B) substantially outperforms all lightweight methods, with an AUC of 0.823 (95% CI: 0.815–0.830). The gain of approximately 18 percentage points in AUC over the best lightweight method is large and statistically unequivocal, since the confidence intervals do not overlap. Fig. 1 illustrates this gradient, highlighting the separation between similarity/entailment-based methods and the reasoning-based judge.

TABLE I
HALLUCINATION DETECTION PERFORMANCE (AUC) ON 9,275 MEDHAL STATEMENTS, WITH 95% CONFIDENCE INTERVALS ESTIMATED VIA BOOTSTRAP. METHODS ARE ORDERED BY INCREASING AUC.

Method	Type	AUC (95% CI)
Embedding e5-large	Similarity	0.504 (0.492–0.515)
Embedding BGE-large	Similarity	0.517 (0.504–0.529)
NLI DeBERTa-v3-base	NLI	0.560 (0.548–0.571)
NLI DeBERTa-v3-large	NLI	0.636 (0.625–0.647)
Hybrid (all)	Fusion	0.646 (0.635–0.656)
Judge Qwen2.5-7B	**LLM**	**0.823 (0.815–0.830)**

B. The Lexical Bias Mechanism

To understand why similarity methods fail, we investigated the lexical overlap (word Jaccard index) between each statement and its reference source, separated by class. The result is counterintuitive and reveals a structural bias in the benchmark.

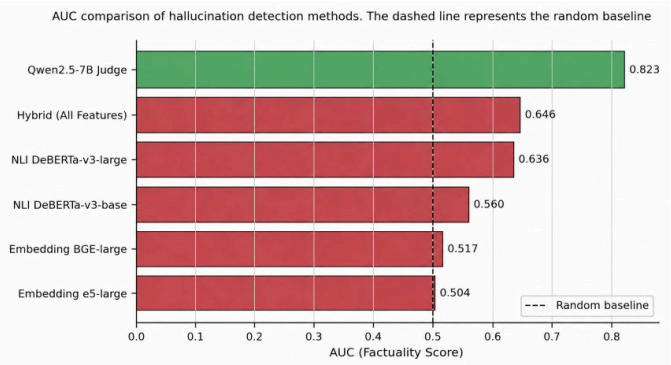


Fig. 1. Gradient of performance across different hallucination detection methods.

In short statements (up to 40 words, 8,102 instances), hallucinated statements present an average lexical overlap of 0.552 with the source, compared to only 0.137 for factual ones—the hallucinated version is about four times more lexically similar to the source than the correct one. The same pattern holds for long statements (over 40 words, 1,173 instances): 0.202 against 0.019, a ratio of approximately ten times. Table II summarizes these values, and Fig. 2 presents the Jaccard distributions by class.

This finding explains the failure of embedding similarity. Since the MedHal hallucinations are generated by minimal substitution of a critical term, the hallucinated statement preserves almost all the vocabulary of the original sentence and maintains high proximity to the source, whereas the factual statement is often a rewording with distinct vocabulary. Similarity-based measures, by rewarding superficial overlap, attribute greater proximity precisely to the hallucinations, neutralizing their discriminative power.

TABLE II
MEAN LEXICAL OVERLAP (JACCARD INDEX) BETWEEN STATEMENT AND SOURCE, BY CLASS AND BY STATEMENT LENGTH. THE HALLUCINATED CLASS IS CONSISTENTLY MORE OVERLAPPED WITH THE SOURCE.

Regime	N	Factual	Hallucinated
Short ($\leq$ 40 words)	8,102	0.137	0.552
Long ($>$ 40 words)	1,173	0.019	0.202

C. Specialty Analysis

Table III reports the performance of the LLM-as-a-judge stratified by the 21 medical specialties. The judge exhibits strong variation across domains, with the highest AUC in basic science areas—Biochemistry (0.930), Microbiology (0.912), and Physiology (0.878)—and the lowest in clinical and surgical areas—Orthopedics (0.696), Surgery (0.738), and Anesthesia (0.754).

This pattern is consistent with the nature of the statements in each domain. Basic science statements tend to express canonical and verifiable facts, which are well represented in the model's pre-training knowledge, whereas clinical statements frequently depend on case context and judgment, making

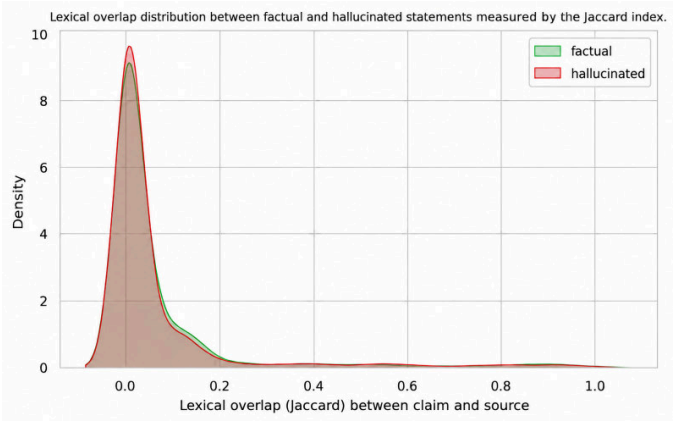


Fig. 2. Jaccard index distributions by class, demonstrating higher lexical overlap for hallucinated statements.

verification more difficult. Specialties with an insufficient number of instances after the confidence filter were excluded from this analysis.

TABLE III
PERFORMANCE OF THE LLM-AS-A-JUDGE (QWEN2.5-7B) BY MEDICAL SPECIALTY, ORDERED BY DECREASING AUC. SPECIALTIES WITH FEWER THAN 20 INSTANCES WERE OMITTED.

Specialty	N	AUC	F1
Biochemistry	500	0.930	0.900
Microbiology	500	0.912	0.883
Physiology	500	0.878	0.828
Pathology	500	0.855	0.658
Pharmacology	500	0.853	0.810
Otorhinolaryngology	500	0.845	0.792
Prev. and Social Medicine	500	0.843	0.788
Forensic Medicine	500	0.838	0.763
Ophthalmology	500	0.834	0.710
Gynecology and Obstetrics	500	0.816	0.522
Psychiatry	500	0.814	0.569
Anatomy	500	0.810	0.784
Radiology	500	0.809	0.701
Pediatrics	500	0.782	0.614
Dentistry	500	0.770	0.418
Medicine	500	0.763	0.287
Anesthesia	500	0.754	0.485
Surgery	500	0.738	0.440
Orthopedics	273	0.696	0.447

D. Error Analysis

Although the LLM-as-a-judge achieves the highest AUC, its confusion matrix (Table IV) reveals a conservative behavior. With a precision of 0.903 and a recall of 0.550 (F1-score of 0.683), the judge is highly reliable when it declares a statement as a hallucination, but it fails to detect a substantial portion of the hallucinated cases: out of the 4,489 hallucinated statements, only 2,630 were correctly identified, while 2,156 were classified as factual. Out of the 4,786 factual statements, 4,205 were correctly recognized.

This imbalance carries critical weight in the medical context, where the most dangerous error is the false negative—an undetected hallucination presented as reliable. Even though the judge represents a significant advance over lightweight methods, its recall level indicates that robust medical hallucination detection remains an open problem. Furthermore, it is observed that in some specialties the F1-score drops sharply even with a high AUC (Medicine: AUC 0.763, F1 0.287), reflecting the judge's tendency to rarely assign the hallucination label in these domains and evidencing the sensitivity of the results to the decision threshold.

TABLE IV
CONFUSION MATRIX OF THE LLM-AS-A-JUDGE (QWEN2.5-7B). ROWS: ACTUAL CLASS; COLUMNS: PREDICTION.

	Pred. Factual	Pred. Hallucinated
Actual Factual	4,205	284
Actual Hallucinated	2,156	2,630

V. DISCUSSION

The results outline a clear gradient: embedding similarity operates close to chance level, entailment via NLI offers a modest gain, and the evaluated LLM-as-a-judge approach achieves substantially superior performance. This ordering suggests that medical hallucination detection is not, in essence, a problem of semantic proximity, but of factual verification—a task that requires evaluating the relationship between statements and evidence rather than merely measuring textual similarity. Lightweight methods showed limited effectiveness for hallucination detection under the evaluated MedHal setting.

The most relevant finding of this work is the lexical bias that characterizes MedHal. Because they are generated by minimal substitution, the hallucinated statements often preserve a large portion of the vocabulary of the source sentence, remaining lexically closer to the source than the factual statements themselves—which are frequently reworded. This phenomenon inverts the premise upon which similarity methods rely and explains their chance-level performance. However, an interpretive caveat is necessary: it is unclear whether this bias is an artifact of the benchmark construction process or if it reflects a genuine property of substitution hallucinations, which indeed tend to be textually close to the correct statement. In either case, the practical implication is the same—similarity methods should not be employed in isolation to detect this type of hallucination, and detection benchmarks should report the lexical overlap between classes as a diagnostic of difficulty.

Even though the LLM-as-a-judge represents a significant advance, its asymmetry between high precision and low recall deserves particular attention in the medical domain. A detector that still presents a substantial false-negative rate, even if it rarely raises false alarms, is of limited utility in clinical applications, where the false negative—incorrect information presented as reliable—constitutes the most consequential error. This conservative behavior, coupled with the sensitivity to the decision threshold evidenced by the drop in F1-score in some specialties, indicates that the aggregated AUC performance, although favorable, does not fully capture the method's

suitability for practical use. Future work could calibrate the decision threshold to favor recall, even at the expense of precision, in line with the application's risk profile.

The stratification by specialty reveals that the detection difficulty is not uniform. The superior performance in basic sciences and inferior performance in clinical and surgical areas suggest that the verifiability of a statement depends on how canonical and decontextualized the knowledge it mobilizes is. Biochemistry or microbiology statements may correspond more frequently to canonical facts, while clinical medicine, surgery, or anesthesia statements may depend more heavily on contextual information, judgment, and exceptions. This result has methodological value: hallucination detection evaluations that report only aggregated metrics may mask weaknesses concentrated precisely in the specialties of greatest clinical complexity, and the specialty dimension—made possible here by the dataset enrichment—proves to be an informative axis of analysis.

This study presents limitations that bound the scope of its conclusions. First, the evaluation of the LLM-as-a-judge paradigm was based on a single model; although open and deterministic, it does not represent the diversity of possible behaviors among different LLMs, and known judge biases such as position, verbosity, and self-preference may still affect model-based evaluation [13], [19]. An ensemble of judges could produce more robust estimates. Second, since the explicit context field of the used subset was mostly empty, the explanation field was adopted as the reference evidence; this choice may influence methods that rely on the source. Third, the specialty labeling was obtained by an automatic classifier and filtered by confidence, so residual attribution errors cannot be entirely ruled out. Finally, the conclusions are limited to MedHal and its specific hallucination generation mechanism; benchmarks built by other processes may exhibit distinct biases.

VI. Conclusion

This work investigated the limits of lightweight methods for hallucination detection in language models applied to the medical domain, using the MedHal benchmark as an experimental basis [5]. The analysis compared approaches based on embedding similarity, natural language inference, signal combination, and a language model acting as a judge. The results indicate that medical hallucination detection cannot be treated merely as a problem of semantic proximity between a statement and its reference source.

Embedding-based methods showed near-chance performance, while textual inference models obtained only modest gains. The LLM-as-a-judge achieved substantially superior performance, suggesting that factual verification requires contextual evaluation of the relationship between evidence and statement. Still, even this method presented important limitations, especially regarding the occurrence of false negatives, where hallucinated statements were classified as factual. In a medical context, this type of error is particularly critical, as it can allow incorrect information to be interpreted as reliable.

One of the main findings of the study was the identification of a lexical bias in MedHal. Because many hallucinations are produced by minimal substitutions of critical terms, they preserve much of the original source's structure and vocabulary. This causes hallucinated statements to potentially be lexically closer to the reference than factual statements rewritten as paraphrases. This behavior weakens the premise of similarity-based methods and highlights the importance of evaluating not only the aggregated performance of detectors but also the mechanisms that explain their successes and errors.

The stratified analysis by medical specialty also showed that the detection difficulty is not uniform across clinical domains. Areas associated with more canonical and straightforward knowledge, such as basic sciences, showed better performance, while clinical and surgical specialties revealed greater difficulty. This result suggests that aggregated evaluations may hide relevant weaknesses in areas of greater contextual complexity, reinforcing the need for domain-specific analyses when evaluating healthcare hallucination detectors.

As limitations, this study evaluated only a single model in the LLM-as-a-judge configuration, used automatically obtained specialty labels, and focused on the specific characteristics of MedHal. Furthermore, the lack of explicit context in part of the dataset led to the use of the textual explanation as a reference anchor. Future work should investigate multiple judges, threshold calibration aimed at reducing false negatives, and retrieval-augmented generation approaches that may improve source grounding and transparency in healthcare settings.

In summary, the results reinforce that medical hallucination detection must be understood as a factual verification task, rather than merely a textual similarity task. Lightweight methods can be useful as auxiliary signals, but they should not be employed in isolation in high-risk scenarios. These findings highlight considerations for the development of more reliable medical AI evaluation systems.

Acknowledgment

The authors used AI-assisted tools only to support paragraph rewriting without changing the original meaning, translation from Portuguese to English, and code generation or code refinement during the development of the experiments. The methodological design, experimental setup, data analysis, results, and scientific interpretations were independently defined and conducted by the authors. AI-assisted tools were not used to generate experimental results, alter data, fabricate references, or perform statistical analyses. All assisted text, translations, and code were reviewed, validated, and approved by the authors.

References

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large

Language Models Encode Clinical Knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[3] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 1906–1919.

[4] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 12 076–12 100.

[5] G. Mehenni, F. Lamarche, O. Rios-Ibacache, J. Kildea, and A. Zouaq, "MedHal: An Evaluation Dataset for Medical Hallucination Detection," *arXiv preprint arXiv:2504.08596*, 2025.

[6] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "SumPubMed: Summarization Dataset of PubMed Scientific Articles," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, Online, 2021, pp. 292–303.

[7] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering," in *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, vol. 174. PMLR, 2022, pp. 248–260.

[8] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[9] A. Bonnet and P. Boulenger, "Augmented Clinical Notes," Hugging Face Datasets, https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes, 2024.

[10] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," *arXiv preprint arXiv:2402.05672*, 2024.

[11] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-Pack: Packed Resources for General Chinese Embeddings," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 641–649.

[12] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," in *International Conference on Learning Representations (ICLR)*, 2023.

[13] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[14] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020, pp. 38–45.

[16] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed Precision Training," in *International Conference on Learning Representations (ICLR)*, 2018.

[17] Qwen Team, "Qwen2.5 Technical Report," *arXiv preprint arXiv:2412.15115*, 2024. [Online]. Available: https://arxiv.org/abs/2412.15115

[18] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical Domain Hallucination Test for Large Language Models," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, Singapore, 2023, pp. 314–334.

[19] J. Gu, K. Zhou, and Y. Wang, "A Survey on LLM-as-a-Judge," *arXiv preprint arXiv:2411.15594*, 2024.

Machine learning for the reconstruction of complex functional outcomes.

1st Ana Cecilia Cardozo Soares
Post-Graduation in Public Health
Federal University of Ceará
Fortaleza, Brazil
https://orcid.org/0000-0002-0174-7662

2nd Carlos Ravanelli Sousa Rios
Post-Graduation in Technological Education
Federal University of Ceará
Fortaleza, Brazil
https://orcid.org/0009-0003-9573-0363

3rd Shamyr Sulyvan Castro
Dr. Prof. Public Health
Federal University of Ceará
Fortaleza, Brazil
https://orcid.org/0000-0002-2661-7899

4th Auzuir Ripardo de Alexandria
Post-Graduation in Technological Education
Federal University of Ceará
Fortaleza, Brazil
https://orcid.org/0000-0002-6134-5366

Abstract*—This study aimed to evaluate the feasibility of reconstructing the functioning score of adults after a Stroke using machine learning algorithms, utilizing variables of activities of daily living and social participation in a Latin American population study. It was a cross-sectional and analytical study based on secondary data from the III National Survey on Disability and Dependency (III ENDISC). The sample consisted of 409 adults who reported a history of stroke. The dependent variable was the continuous functioning measure "des_puntaje_adult". The data were divided into training (70%) and testing (30%) sets. Based on a selection using eight independent relevance methods, the 100 most important variables were retained. Eleven supervised regression algorithms were tested, representing linear, regularized, kernel-based, ensemble, boosting, generalized linear models, and deep neural network approaches. Performance was evaluated by the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The results demonstrate that non-linear algorithms based on gradient boosting showed superior performance. XGBoost obtained the best reconstruction capacity on the test set (R^2 = 0.9827, RMSE = 3.0755, and MAE = 2.3106), followed by LightGBM and Gradient Boosting. In contrast, linear models (Linear Regression, Ridge, and Lasso) and neural networks (MLP and DNN) presented unsatisfactory performance, with a negative R^2 on the test set. The difference between the training and testing errors indicated the presence of overfitting in the highest-performing models. It is concluded that the results represent an initial proof of concept regarding the technical capacity of tree-based algorithms to reconstruct the mathematical structure of the functioning score from its constituent components. However, due to overfitting, small sample size, and cross-sectional design, the model does not have a direct clinical prediction purpose, depending on external validation and longitudinal data in future studies.

***Index Terms*—Public Health, Stroke, Machine Learning, XGBoost**

I. Introduction

Historically, mortality and morbidity indices have been used as indicators to measure individuals' health. However, population aging, coupled with the increasing burden of non-communicable diseases (NCDs), such as stroke, makes the epidemiological context more complex by contributing to the global increase in the number of years of life lost due to premature death and years lived with disability [1]. Consequently, assessing only mortality and morbidity is no longer sufficient to describe a population's health status [2]. In this scenario, the assessment of functioning acts as a complementary indicator to mortality and morbidity. This is a crucial measure for investigating the use of body functions and structures, as well as the individual's performance in daily activities and social participation, thus advancing the production of health information beyond mere morbidity and mortality outcomes [3].

Globally, it is estimated that 2.4 billion people require interventions aimed at functioning [4]. Such data drive territorial research on disability and health. The III Estudio Nacional de la Discapacidad (III ENDISC) exemplifies this by outlining the disability profile of the Chilean population.

Therefore, traditional statistical tools may lack the necessary precision to capture the non-linear interaction of variables and, consequently, may be insufficient for predictive modeling. Furthermore, Artificial Intelligence (AI), based analyses for coding information about functioning have demonstrated efficiency in optimizing human resources, as well as in supporting clinical decision-making and Public Health management [5]. It should also be noted that machine learning models incorporating Social Determinants of Health (SDOH) show superior performance to those developed solely with clinical characteristics, since functioning results from the interaction of environmental, social, and cultural factors that transcend the underlying pathology, in this case, stroke [5], [6].

In light of this, the present study aimed to evaluate, in an exploratory manner, the feasibility of reconstructing the functioning score of adults after a stroke using machine learning, utilizing variables related to activities of daily living and social participation available in a Latin American population survey.

To the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting the master's scholarship.

It is, therefore, an investigation characterized as a methodological proof of concept regarding how well non-linear algorithms can reconstruct a functioning measure from its own constituent components, and to what extent social determinants of health (SDOH) contribute to this reconstruction in Latin American populations, such as that of Chile.

II. Method

A. Design and data source

This is a cross-sectional and analytical study developed using secondary data from the III ENDISC, a survey conducted in Chile in 2022 by the Servicio Nacional de la Discapacidad (SENADIS) regarding disability in the national population [7]. The survey utilized probabilistic sampling, stratified to represent urban and rural residents, and based its data collection instruments on the International Classification of Functioning, Disability and Health (ICF) to assess individuals from a biopsychosocial perspective.

B. Population and outcome variable

The population consisted of adults aged 18 years or older who reported having had a stroke. Only data from individuals who presented the minimum information necessary for the reconstruction of the continuous risk index for functioning were included; thus, the responses of 409 individuals were evaluated. The dependent variable selected for this study was *"des_puntaje_adult"*, which addresses the performance in carrying out activities in the individuals' usual environment. *"des_puntaje_adult"* was calculated in the III ENDISC from the questions recorded in module D (Performance - Desempeño), using the Rasch Partial Credit Model, which allows for the construction of a continuous scale aligned with the principles of the ICF. To prevent data leakage, the variables derived from the assessment of functioning *(cap_puntaje_adulto, cap_nivel_adulto, des_nivel_adulto, disc_grado_adulto, disc_adulto, depen_grado_adulto, depen_adulto)* were removed from the predictive space.

C. Feature selection and preprocessing

Initially, columns with numerical variables concerning mobility, self-care, domestic activities, interpersonal interactions, health conditions, caregiver availability, assistive devices and support technology, educational and labor participation, and demographic variables were selected. Columns with categorical variables, mixed data, or entirely null columns were excluded, and the remaining missing values were imputed by the mean estimated exclusively on the training set. Next, the selection was adjusted solely on the training set, after partitioning into 70% training and 30% testing (test_size=0.3, random_state=42). Subsequently, eight independent relevance assessment methods were applied, combining filter, embedded, and recursive elimination approaches, which were: Pearson and Spearman correlation (absolute value as ranking criterion), mutual information (mutual_info_regression), univariate F-statistic (f_regression), random forest importance, gradient boosting importance (100 estimators, max_depth=5), absolute coefficients of a LassoCV with 5-fold cross-validation(CV), and recursive feature elimination (RFE) with a simplified random forest (25 estimators, maximum depth 6) as the base estimator. The eight scores were combined by simple arithmetic mean and normalized to obtain a percentage score (0 to 100%), which was ordered in descending order, highlighting the 100 variables with the highest score to compose the final prediction. Thus, the database was restricted to the 100 selected variables plus the target variable, totaling 101 numerical columns and 409 observations. Preprocessing was integrated into the pipeline of each model and consisted of mean imputation, justified by the modular structure of the questionnaire and the assumption of missing approximately at random, followed by z-score standardization. Measures of means, standard deviations, and imputation parameters were estimated solely on the training set for subsequent application to the test set, preventing data leakage. Standardization was applied uniformly to all models, ensuring comparability of the training conditions.

D. Tested algorithms and model evaluation

To compare different predictive paradigms, a diverse set of eleven supervised regression algorithms representing linear, regularized linear, kernel-based, ensemble, boosting, neural networks, and generalized linear model approaches were evaluated. Specifically, the evaluated models included Linear Regression (OLS), Ridge, Lasso, Support Vector Regression (SVR), Random Forest, Gradient Boosting, XGBoost, LightGBM, Multilayer Perceptron (MLP, via MLPRegressor, with early stopping and up to 1,000 epochs), fully connected feed-forward deep neural network (DNN) implemented in TensorFlow/Keras, and GLM with Gamma distribution and logarithmic link. The network architecture included an input layer followed by three dense hidden layers with 128, 64, and 32 neurons, respectively, using Rectified Linear Unit (ReLU) activation functions. Dropout regularization (0.30 and 0.20) was applied after the first two hidden layers to reduce overfitting, and a single linear output neuron was used for continuous regression. Model optimization employed the Adam optimizer with a learning rate of 0.001 and Mean Squared Error (MSE) as the loss function. Training was performed for up to 200 epochs with a batch size of 32, using 20% of the training data for validation and an Early Stopping strategy (patience = 20) to restore the best-performing weights. The hyperparameters of each of the eleven models were individually optimized using exhaustive grid search with 5-fold cross-validation with shuffling (KFold(n_splits=5, shuffle=True, random_state=42)), fitting each pipeline $k \times C$ times, with $k = 5$ folds and C combinations per model. Model performance was evaluated using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). All three metrics were calculated on both the training and test sets. The R^2 indicates the proportion of the dependent variable's variance explained by the model, ranging from 0 to 1 (and can be negative when the model generalizes worse than a mean-based reference predictor), with values closer to 1 indicating

a greater capacity for reconstructing the target variable. Final model selection was primarily based on the coefficient of determination (R^2) obtained on the test set, prioritizing generalization capacity within the reconstruction task itself over performance on the training set. Subsequently, we considered the RMSE and MAE of the test set, as well as the difference between the training and testing metrics as an indicator of overfitting.

E. Ethical Aspects

The study exclusively used public and anonymized microdata from the III ENDISC, made available by SENADIS without individual identification, with no primary collection or intervention involving human beings.

III. Results

The set of 100 variables intrinsically related to functioning included questions about the performance of activities of daily living and social participation. In this group, the ten most relevant variables refer to the level of difficulty or how problematic it has been for the individual to perform certain tasks, which are: 1ª. Maintaining sufficient energy (mean score 0.68); 2ª. Taking care of one's own health (mean score 0.61); 3ª. Getting to desired places (mean score 0.60); 4ª. Remembering important things to do (mean score 0.60); 5ª. Walking 10 blocks or one kilometer (mean score 0.55); 6ª. Using the toilet (mean score 0.53); 7ª. Taking care of one's appearance or getting dressed (mean score 0.53); 8ª. Walking or climbing stairs (mean score 0.52); 9ª. Handling all necessary tasks (mean score 0.51); 10ª. Bathing or dressing (mean score 0.47). The complete list of variables is available in the Supplementary Material [8].

Algorithms based on non-linear methods—XGBoost, LightGBM, and Gradient Boosting—showed better results than the others, given that these methods capture the mathematical relationships between the functioning variables and the global score more accurately. Table I presents the comparative performance of the algorithms in the functioning score reconstruction task. XGBoost exhibited the best performance for reconstructing the functioning score, with R^2 = 0.982799 (RMSE = 3.075570 and MAE = 2.310636) on the test set; the model accounted for explaining 98.2% of the score variance. Conversely, the Linear Regression, Ridge, and Lasso models, as well as the MLP and Deep Neural Network, presented unsatisfactory performance. The performance of the Linear Regression, Ridge, and Lasso algorithms can be explained by the fact that functioning involves a complex network of interactions among variables, and these algorithms are more accurate for treating outcomes that occur in a direct or sequential manner. In turn, MLP and Deep Neural Network work better with more voluminous datasets. Furthermore, it is believed that the high difference in RMSE values between the training and test sets indicates data overfitting, which compromises the interpretation of the values.

Figure1 demonstrates the comparative RMSE performance among the evaluated algorithms. With the image, it is possible

TABLE I
RANKING OF ALGORITHMIC MODELS IN THE TEST SET

Model	Evaluation and validation strategy				
	RMSE*	RMSE**	R^{2}*	R^{2}**	MAE
XGBoost	0.486	3.075	0.999	0.982	2.310
LightGBM	1.087	3.379	0.997	0.979	2.489
Gradient Boosting	0.495	3.701	0.999	0.975	2.661
Random Forest	1.718	5.099	0.994	0.952	3.927
SVR	6.202	8.532	0.934	0.867	5.638
GLM (Gamma + Log)	17.768	13.401	0.896	0.673	7.852
Lasso Regression	5.892	20.596	0.940	0.228	7.353
Ridge Regression	5.687	25.118	0.944	-0.147	8.394
Linear Regression	4.848	28.451	0.959	-0.472	9.762
Deep Neural Network	6.051	32.182	0.937	-0.883	12.631
MLP Neural Network	5.250	48.166	0.952	-3.218	14.607

Source: Authors, 2026.
*Applied to the training data. **Applied to the test data.

to visually confirm the superiority of tree-based models in reconstructing a score with non-linear interactions, since they present residual errors below five. On the other hand, the linear regression, Lasso regression, and Ridge regression algorithms, as well as MLP and Deep Neural Network, presented performance limitations for the presented data.

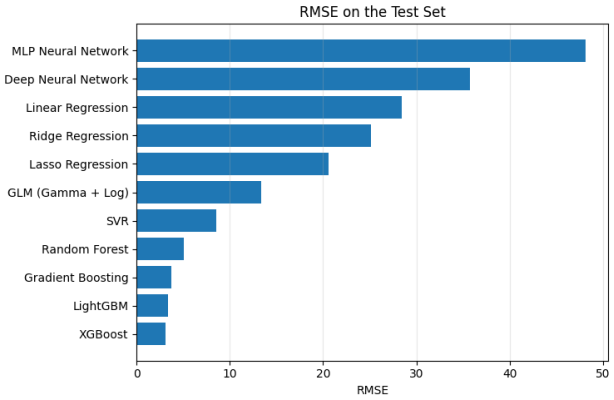


Fig. 1. Comparison of the Models on Test Set (Authors, 2026.)

Figure 2 illustrates the relationship between actual and predicted values by XGBoost, with the dispersion of points close to the reference line, indicating agreement between observed and reconstructed values. In turn, Figure 3 presents the residual plot of this model, which evidences the unsystematic pattern of distribution in relation to the predicted values, with a greater concentration of outliers between the values of -5 and +5. Thus, visual inspection may suggest homoscedasticity of the results; however, formal tests were not performed to confirm this observation.

In turn, in the histogram (Figure 4, a distribution of errors close to normality is observed; therefore, the model presents a low risk of systematic bias. However, it should be noted that the risk of overfitting, as well as the sample size, may also have interfered with the stability of the residuals presented in this graph.

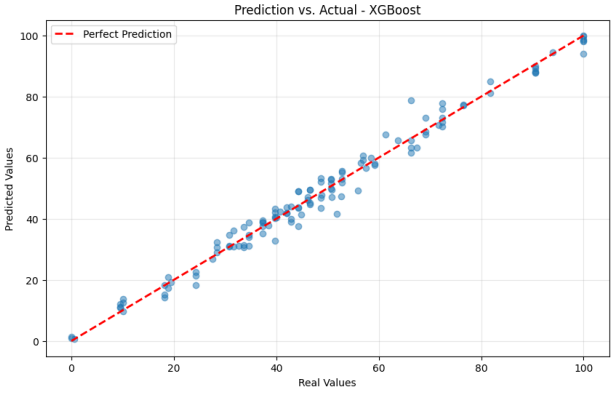


Fig. 2. Real Values x Prediction Values - XGBoost (Authors, 2026.)

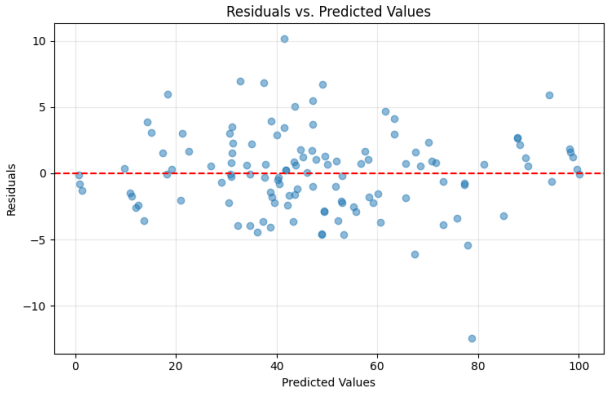


Fig. 3. Residual Plot (Authors, 2026.)

As previously mentioned, the XGBoost algorithm showed more promising results; thus, it was used for independent analyses by macro-regions (North, Central, South, and Metropolitan) and SHapley Additive exPlanations (SHAP) analysis. The algorithm's ability to capture the variance of the functioning score presented important discrepancies across the geographic strata. The model adjusted for the Central macro-region obtained the highest precision (R^2 = 0.97 on the test set; Mean Absolute Error [MAE] = 2.82). Similar predictive behavior was observed in the South macro-region (R^2 = 0.95; MAE = 3.79). In this scenario, the Metropolitan region recorded R^2= 0.92 and MAE = 4.96, while the North region had R^2= 0.91 and MAE = 5.68. The table II summarizes the metrics.

TABLE II
MODEL PERFORMANCE STRATIFIED BY MACRO-REGION

Macroregion	**Tests**			
	Test Set	**R^2**	**RMSE**	**MAE**
North	21	0.915	8.11	5.68
Center	51	0.97	3.79	2.82
South	45	0.95	4.99	3.79
Metropolitan	26	0.92	6.77	4.96

Source: Authors, 2026.

The mean absolute SHAP values were important for identifying which variables had the highest average impact on the models' predictions. The SHAP analysis shown in Table III, revealed both universal patterns and significant regional disparities in the determinants of functioning. The variable d7, which investigates the difficulty in performing daily tasks, stood out as the most important predictor in the four regions, exerting an exceptionally high impact in the Central region (mean SHAP = 6.28). Variables such as d6, d8, and d9, which respectively investigate problems with tiredness and lack of energy, memory, and carrying out household tasks, also demonstrated high overall predictive relevance.

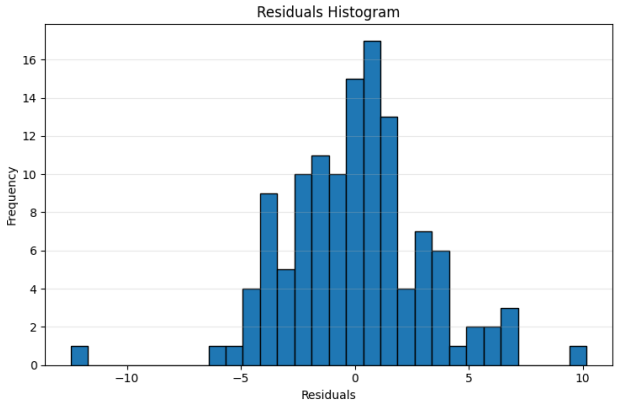


Fig. 4. Residual Distribution (Authors, 2026.)

In the Metropolitan region, the variable c4 (difficulty walking or climbing stairs) is of significant importance to the model, occupying the second position (mean SHAP = 3.12). Similarly, the model for the South region is influenced by the variable c14, which represents the difficulty in shopping or going to the doctor (mean SHAP = 2.67). The North region presented a unique profile, in which age emerged among the 10 features with the highest impact (mean SHAP = 1.19), which did not demonstrate the same level of importance in the other macro-regions.

These results suggest that, although the fundamental challenges related to functioning are shared geographically, specific regional contexts introduce distinct determinants that must be considered when carrying out localized interventions.

IV. DISCUSSION

This study aimed to evaluate the feasibility of reconstructing, through machine learning, the functioning score of adult stroke survivors using variables belonging to the same conceptual domain as the outcome. XGBoost demonstrated a good capacity for the mathematical reconstruction of the functioning score in this specific dataset (R^2 = 0.982 on the test set). However, the study demonstrates a proof of concept regarding the technical feasibility of the approach, but it does not validate it as a tool ready for predictive use in clinical contexts.The observed overfitting results from the disproportion between the sample size and the number of predictor variables (409 individuals compared to 100 predictor variables). This characteristic, as well as the cross-sectional nature of the data and

TABLE III
TOP 10 MOST IMPORTANT VARIABLES PER MACRO-REGION ACCORDING TO MEAN ABSOLUTE SHAP VALUES

Rank	Center	Metropolitan	North	South
1	d7 (6.28)	d7 (3.41)	d7 (4.29)	d7 (3.48)
2	d8 (2.96)	c4 (3.12)	d9 (3.61)	c14 (2.67)
3	d6 (2.42)	d6 (2.39)	d6 (3.43)	d8 (2.40)
4	d1 (2.01)	d10 (2.22)	d5 (2.59)	d1 (2.23)
5	d5 (1.70)	d8 (2.13)	d4 (1.56)	d6 (1.85)
6	d9 (1.63)	c21 (1.63)	c4 (1.46)	d3 (1.74)
7	d10 (1.43)	d3 (1.55)	d2 (1.41)	d2 (1.68)
8	d11 (1.31)	d9 (1.42)	d1 (1.20)	d9 (1.64)
9	d3 (1.24)	d2 (1.35)	edad (1.19)	d11 (1.55)
10	d2 (1.16)	d5 (1.19)	d3 (1.11)	d5 (0.95)

Note: Values are presented as variable name (mean absolute SHAP value).
Source: Authors, 2026.

the lack of external validation, constitute limitations for the study. In this scenario, the imbalance between individuals and variables may have compromised the training stage, since the algorithms are capable of memorizing response noise during training, thus distorting the testing phase [9].

Furthermore, the cross-sectional design prevents making inferences about patients' recovery over time; on the other hand, the absence of algorithm validation in an independent sample also limits the generalization of the results to other populations or contexts beyond the one studied.Measuring functioning after a stroke involves evaluating the complex interaction of SDOH, which encompasses clinical, socioeconomic, cultural, and environmental aspects. According to published studies, SDOH can improve the predictive capacity of algorithms [6], [10]. SDOH have also demonstrated central importance in the construction of predictive models for cognitive decline in the Mexican population; variables such as educational level, age, reading habits, maternal education, and frequency of social activity had the highest predictive importance [11].

Therefore, by aggregating clinical and social information, the III ENDISC survey becomes a rich material for assessing functioning in the Chilean population and also contributes to subsequent investigations of a predictive nature. Furthermore, it is necessary to use social, racial, and economic variables in a critical and ethical manner, because although the use of these data in computational analyses can support conclusions about inequalities, predictive analyses can also reinforce existing barriers if not properly contextualized [12].

The favorable performance of the XGBoost, LightGBM, and Gradient Boosting algorithms in this reconstruction work was likely influenced by the inclusion of SDOH-related variables. Similar results were observed in the study by Xiao et al. [10], in which there was the integration of serum globulin levels with ICF components to validate a predictive model of long-term functioning, through the development of a therapeutic decision support tool for rehabilitation plans. Moreover, models based on Gradient Boosting, such as LightGBM, are more suitable for handling tabular, smaller, and imbalanced datasets. On the other hand, Deep Learning algorithms, such as Deep Neural Networks and MLP Neural Networks, perform better with extremely large amounts of labeled data [13], [14].

Therefore, the good reconstruction performance presented in this study should be interpreted with caution due to the limitations of sample size and the risk of overfitting and they should not be used for clinical predictive purposes. The results primarily reflect the algorithm's potential to reconstruct dimensions related to functioning from variables within the same domain and do not generate a validation of predictive capacity for future and independent clinical outcomes. Although the present study does not build unprecedented predictions, its importance is reiterated in supporting future research involving the more advanced use of the tested algorithms for the development of health solutions and policies.Due to the results obtained with the algorithms applied to the Chilean population data and the observed efficiency of models in other population groups, it is recommended that future research gather clinical and sociodemographic data from the Brazilian population to test them with machine learning models aimed at predicting health situations and, if appropriate, the reconstruction of scores using Brazilian data. It is important to use Brazilian data, as algorithms trained on foreign populations, such as the Chilean one, may not produce the expected results when applied to different demographic groups. This recommendation is aligned with the assumptions established by Brazilian and international authorities, which strongly recommend the use of artificial intelligence tools, such as predictive analysis [15], [16].

However, the main challenge in conducting such research in the Brazilian context lies in the need for national surveys that incorporate information on the functioning of body structures, activities of daily living, social participation, and mental health—in other words, robust measures of functioning. The small sample size constitutes the main limitation of this study, as it can interfere with the predictive capacity of the learning models. Furthermore, the application of this approach as a predictive tool itself still depends on the use of predictor variables collected independently from the outcome, avoiding overlap of the conceptual domain; external validation in independent samples; longitudinal data that allow evaluating recovery trajectories; and tests in different populations, prior to any application in risk stratification or clinical decision support in rehabilitation. Additionally, the scarcity of Latin American studies with longitudinal data on post-stroke functioning and the need for externally validated predictive models in these populations are highlighted.

CONCLUSION

In summary, Gradient boosting algorithms, particularly XGBoost, demonstrated good performance in reconstructing the functioning measure in this dataset. These results establish an initial proof of concept regarding the feasibility of machine learning models for handling complex health outcomes in population-based studies. Furthermore, the future application of these algorithms, appropriately designed for predictive

purposes, can strengthen the use of machine learning and AI in promoting functioning and quality of life.

As machine learning is increasingly applied to predict health outcomes using large population datasets, our understanding of chronic social issues becomes clearer.It is recommended that future studies test these algorithms in longitudinal cohorts with larger population groups; use predictors conceptually independent of the outcome; and perform validations in external samples to assess generalization capacity and large-scale applicability.

REFERENCES

[1] I. for Health Metrics and E. (IHME), "Global burden of disease 2021: findings from the gbd 2021 study," *Institute for health metrics and evaluation*, 2024.

[2] G. Stucki and J. Bickenbach, "Functioning: the third health indicator in the health system and the key indicator for rehabilitation.," *European journal of physical and rehabilitation medicine*, vol. 53, no. 1, pp. 134–138, 2017.

[3] W. H. Organization, "International classification of functioning, disability and health (icf)," 2001.

[4] A. Cieza, K. Causey, K. Kamenov, S. W. Hanson, S. Chatterji, and T. Vos, "Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10267, pp. 2006–2017, 2020.

[5] D. Newman-Griffis and E. Fosler-Lussier, "Automated coding of understudied medical concept domains: linking physical activity reports to the international classification of functioning, disability, and health," *Frontiers in digital health*, vol. 3, p. 620828, 2021.

[6] M. W. Segar, J. L. Hall, P. S. Jhund, T. M. Powell-Wiley, A. A. Morris, D. Kao, G. C. Fonarow, R. Hernandez, N. E. Ibrahim, C. Rutan, *et al.*, "Machine learning–based models incorporating social determinants of health vs traditional models for predicting in-hospital mortality in patients with heart failure," *JAMA cardiology*, vol. 7, no. 8, pp. 844–854, 2022.

[7] M. d. D. S. y. F. Departamento de Evaluación y Estudios, Servicio Nacional de la Discapacidad (Senadis), "Iii estudio nacional de la discapacidad en chile," 2023.

[8] S. S. C. A. R. d. A. Ana Cecilia Cardozo Soares, Carlos Ravanelli de Sousa Rios.

[9] R. D. Riley, K. I. Snell, J. Ensor, D. L. Burke, F. E. Harrell Jr, K. G. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: Part ii - binary and time-to-event outcomes," *Statistics in Medicine*, vol. 38, no. 7, pp. 1276–1296, 2019.

[10] C. Xiao, H. Zhang, D. Chen, Y. Xiu, Z. Liu, and Y. Wu, "Integrating serum globulin into the icf framework: a novel multidimensional predictive model for 1-year mrs outcomes in acute ischemic stroke," *European Journal of Medical Research*, 2026.

[11] Y. He, Y. Leng, A.-M. Vranceanu, C. S. Ritchie, D. Blacker, and S. Das, "A predictive model for cognitive decline using social determinants of health," *The Journal of Aging Research Lifestyle*, vol. 15, p. 100056, 2026.

[12] M. Q. Soares and A. D. P. Chiavegatto Filho, "Artificial intelligence in brazilian public health: potential, challenges, and ethical implications for the brazilian unified health system," *Revista Brasileira de Epidemiologia*, vol. 29, p. e260006, 2026.

[13] I. B. Machines, "What is machine learning?," 2026.

[14] D. Bergmann, "What is gradient boosting?," 2024.

[15] M. da Saúde (BR), "Estratégia de saúde digital para o brasil 2020-2028," 2020.

[16] V. S. Yedavalli, E. Tong, D. Martin, K. W. Yeom, and N. D. Forkert, "Artificial intelligence in stroke imaging: Current and future perspectives," *Clinical Imaging*, vol. 69, pp. 246–254, 2021.

Machine Learning in Healthcare: Development of Dementia Prediction Models

Márcia R S Nascimento
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0000-0002-1792-5355

Maurício Cagy
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0000-0002- 6488-842X

Diogo Antonio Tschoeke
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0000-0001-8134-4259

Letícia Martins Raposo
Quantitative Methods Department
Federal University of the State of Rio de Janeiro
Rio de Janeiro, Brazil
ORCID: 0000-0003-0613-5582

***Abstract*—Dementia has a high global prevalence and high rates of underdiagnosis in Brazil (76.1% to 77%). Machine learning (ML) is emerging as a promising tool to assist with diagnostic screening. Objective: To develop, compare, and evaluate the performance of ML models for identifying dementia. Method: Data from 334 observations of 150 participants (men and women, aged 60–98) from the Open Access Series of Imaging Studies (OASIS) database were used. The Boruta algorithm identified the MMSE as the most relevant variable, followed by Sex and nWBV. These variables were selected as predictor variables. Data were cleaned and prepared to train three predictive models: Support Vector Machines (SVM), Gradient Boosting (GBM), and Multilayer Perceptron (MLP). To compare performance of these algorithms, a non-parametric Friedman test was conducted. Results: the MLP algorithm performed best overall, with 86% accuracy, 81% sensitivity, 90% specificity, and AUC-ROC of 0.90. However, the analysis of the confidence intervals for accuracy, sensitivity, specificity, and AUC-ROC of the three models shows an overlap and the Friedman test (X^2 = 1.23, p = 0.54) revealed no statistically significant differences in performance among the models. Conclusion: The developed models demonstrated good predictive performance. These algorithms represent viable alternatives for supporting clinical decision-making in high-demand scenarios or situations with time and resource constraints.**

Keywords — Machine learning. Dementia. Older adult health.

I. INTRODUCTION

It is estimated that the number of people affected by dementia worldwide increased by 117% between 1990 and 2016, largely due to the aging population [1]. In 2019, 2.46 million people were living with dementia in Brazil, and projections indicate that this number will reach 5.05 million in 2039 and 8.74 million in 2049 [2]. Despite this high prevalence of dementia, the estimated underdiagnosis rate is 76.1% to 77% in Brazil [3,4].

Machine learning (ML), a subfield of Artificial Intelligence, is a data analysis technique that uses algorithms to build predictive, identification, and decision-making models [5,6]. ML is rapidly evolving in the medical industry, encompassing everything from diagnosis to disease visualization and the study of disease transmission [6].

This study aimed: *i*) to develop supervised machine learning models for dementia identification, using magnetic resonance imaging data (normalized whole brain volume (nWBV)), parameters from the Mini-Mental State Examination (MMSE), and sex; *ii*) to compare the performance of the developed models; *iii*) to identify the importance of the predictor variables.

II. BACKGROUND

In recent years, the application of machine learning (ML) algorithms has helped to solve several complex problems in various fields. In biomedicine, these algorithms have been used for prediction, diagnosis, and identification of disease development stages. In the context of Alzheimer's disease (AD), ML algorithms have been used in various research projects, from mapping genes and biomarkers (beta-amyloid 42 (Aβ42) and tau proteins) to image analysis (magnetic resonance imaging (MRI) and positron emission tomography (PET)) and sociodemographic and clinical data. The invasive nature of lumbar puncture limits the widespread use of biomarkers such as amyloid beta-42 (Aβ42) and tau proteins. The high cost and limited availability also restrict wider applications of MRI and PET [7,8].

Recent studies have applied ML algorithms (neural networks and classical algorithms) to predict and detect AD. For instance, the study in reference [8] developed machine learning models (logistic regression with penalty (LR), support vector machine (SVM), decision trees, random forest (RF), extreme gradient boosting (XGB) and artificial neural network (ANN)) to predict AD using sociodemographic, self-reported health and blood biomarker information. Meanwhile, the study in reference [9] analyzed magnetic resonance imaging, demographic, clinical, and cognitive data using a long short-term memory (LSTM) neural network and a feedforward neural network (FNN) to detect AD early.

In this study, we use three algorithms: Support Vector Machines (SVM), Gradient Boosting (GBM) and Multilayer Perceptron (MLP) to identify dementia.

A. Support Vector Machines (SVM)

The Support Vector Machines (SVM) algorithm is a supervised machine learning model that aims to find an optimal hyperplane in an *n*- dimensional space that maximizes the margin between the classes of interest. The balance

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

between maximizing the margin and minimizing training errors is controlled by the regularization hyperparameter *C* (cost) [5,6,10].

B. Gradient Boosting (GBM)

Gradient Boosting (GBM) is a supervised machine learning technique that builds models in an additive and sequential manner. The performance of the previous model influences each new model. It combines multiple base models (weak learners) to produce a final, higher-performing model (strong learner). GBM focuses on correcting errors made by previous models. The algorithm treats training as an optimization problem within a function space. The goal is to find an approximation function that minimizes the expected value of a loss function. A loss function quantifies the impact of classification or prediction errors in the solution to a problem. Given a specific loss function, the optimal solution for a classification or prediction problem is the one that minimizes that loss. At each step *m*, a new "weak model" (a shallow decision tree) is trained to predict the negative gradient of the loss function with respect to the current model's output. The new model attempts to learn the direction and magnitude of the error accumulated by the ensemble up to that point. The final result is derived from the last tree. However, it is necessary to determine the extent to which the target variable is adjusted at each step. This is achieved using a learning rate, which ranges from 0.0 to 1.0 [11, 12].

C. Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) network is a feedforward artificial neural network architecture characterized by the organization of neurons into multiple layers. It consists of three layers: the input layer, the hidden layers, and the output layer. Each neuron in a hidden or output layer performs a weighted sum of its inputs, plus a bias term, and submits the result to a non-linear activation function. The learning process is driven by the backpropagation algorithm, which adjusts the network weights to improve predictions by computing the gradient of the cost function with respect to the weights, updating them layer by layer [5,11].

III. METHODS

A dataset from the Open Access Series of Imaging Studies (OASIS) was used, comprising 374 observations of 150 demented and non-demented subjects (men and women), right-handed, aged from 60 to 98 years. They underwent cranial magnetic resonance imaging (MRI), and clinical and sociodemographic assessments across two or more visits spaced at least one year apart. The sample data included the following variables: class (Group), Sex, Age, socioeconomic status (SES), years of education (EDUC), Mini-Mental State Examination (MMSE), estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and atlas scale factor (ASF) [13]. Two groups of interest were considered: demented and non-demented. Observations with inconclusive class (Group) were excluded. Missing entries were observed in two observations for the MMSE variable and seventeen for the SES variable.

A. Variables Selection

The Boruta algorithm, a variable selection method that identifies all relevant variables in a database, was implemented, using Random Forest as a basis. After expert analysis, correlation matrix, contingency tables, and Boruta results, two observations with missing entries for the variables MMSE, SES, and ASF were excluded. The SES variable showed little relevance and had missing observations. The ASF variable showed a 99% correlation with the eTIV variable.

The results of the Boruta algorithm show that MMSE is the most important predictor variable, followed by Sex and nWBV. The MMSE is a quick test (5 to 10 minutes) focused on the patient's cognitive performance at the time of the consultation. The MMSE is easy to apply in the routine of a clinic or hospital.

The variables Sex, MMSE and nWBV, along with their specific characteristics, were selected as predictor variables. The variable Group was chosen as the output (outcome) variable corresponding to the diagnosis of dementia based on the other input (predictor) variables.

After cleaning and preparing the data, a total of 334 observations from 135 participants and 4 variables were obtained (one outcome variable and three predictor variables). The data were balanced (43.1% demented and 56.9% non-demented) and there were no missing observations in any of the four variables (Table 1).

TABLE I. VARIABLES CHARACTERISTICS

Variables	Statistics/Frequency
Group	Demented: 144 (43.1%) Non-demented: 190 (56.9%)
Sex	Female: 187 (56%) Male: 147 (44%)
Age	Mean (sd[a]): 76.7 (7.6) Minimum: 60 Maximum: 98
EDUC	Mean (sd[a]): 14.5 (2.9) Minimum: 6 Maximum: 23
MMSE	Mean (sd[a]): 27.2 (3.8) Mínimo: 4 Maximum: 30
eTIV	Mean (sd[a]): 1493.3 (178.7) Minimum: 1106 Maximum: 2004
nWBV	Mean: 0.7 Minimum: 0.6 Maximum: 0.8

[a.] Standard Deviation

B. Hyperparameter adjustments

Three algorithms were implemented in the R (v. 4.5.0) [14]: Support Vector Machines (SVM), Gradient Boosting (GBM) and Multilayer Perceptron (MLP). For supervised learning, the structured dataset was split into training (70% = 234 observations from 95 participants), validation and test (30% = 100 observations from 40 participants) set by participant ID, ensuring that no data from the same individual appeared in more than one set. All models used 10-fold cross-validation as control parameters for training. The area under the ROC curve (AUC-ROC) was used to select the best model, aiming to balance between sensitivity and specificity.

SVM Linear: data were standardized (using “center” and “scale”). The regularization hyperparameter (*C*) was optimized using a grid search. The tested grid varied in powers of 10 (10^{-3} to 10^{3}). In this model, *C* optimal = 0.1.

GBM: The hyperparameters maximum depth of each tree (interaction.depth), total number of trees (n.trees), learning rate (shrinkage), and minimum number of observations at terminal nodes (n.minobsinnode) were optimized using a grid search. The tested grid ranged from 3 to 8 for interaction.depth, 50, 100, and 500 for n.trees, 0.1, 0.01 and 0.001 for shrinkage, and 10 and 15 for n.minobsinnode. In this model, the best hyperparameters were: interaction.depth = 4, n.trees = 100, shrinkage = 0.001 and n.minobsinnode = 15.

MLP: data were standardized (using "center" and "scale"). Dummies were created for categorical variables. The hyperparameters were hidden layer neurons (size) and weight decay. The decay values were optimized using a grid search. The tested grid ranged from 2, 4, 8, and 16 for size, and 0, 0.0001, 0.001, and 0.01 for weight. Decay. In this model, optimal size = 2 and optimal decay weight = 0.01.

IV. Results

The models (SVM, GBM, and MLP) achieved good accuracy and high predictive capacity (Table 2). The MLP algorithm achieved a good overall balance, with the highest accuracy (86%), sensitivity (81%), specificity (90%), and AUC-ROC (0.90). It presents an optimized architecture that required only two neurons in the hidden layer (Figure 1). However, the analysis of the confidence intervals for accuracy, sensitivity, specificity, and AUC-ROC of the three models shows an overlap. This highlights the similarity in the models' performance. The SVM algorithm stood out for its higher specificity (94%). It was the most effective model for identifying positive cases of non-dementia, with only three false positive recorded (Table 3).

TABLE II. PERFORMANCE MEASURES OF MODELS

	SVM	**GBM**	**MLP**
Accuracy [95% CI]	0.85 [0.76 0.91]	0.83 [0.74 0.90]	0.86 [0.77 0.92]
Kappa Index	0.69	0.65	0.72
Sensitivity [95% CI]	0.74 [0.61 0.85]	0.76 [0.64 0.89]	0.81 [0.70 0.91]
Specificity [95% CI]	0.94 [0.87 0.99]	0.88 [0.79 0.96]	0.90 [0.81 0.98]
AUC-ROC [95% CI]	0.89 [0.82 0.95]	0.88 [0.81 0.95]	0.90 [0.83 0.96]

TABLE III. CONFUSION MATRIX (SVM, GBM E MLP)

Predicted Class	**True Class**					
	SVM		**GBM**		**MLP**	
	Demented	***Non-Demented***	***Demented***	***Non-Demented***	***Demented***	***Non-Demented***
Demented	35	3	36	6	38	5
Non-Demented	12	50	11	47	9	48

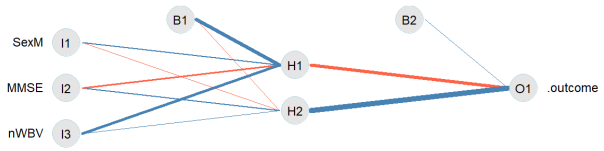


Fig. 1. MLP network architecture

To compare performance of the evaluated machine learning algorithms (SVM, GBM, and MLP) across the cross-validation folds, a non-parametric Friedman test was conducted. The analysis revealed no statistically significant differences in performance among the models (χ^2 = 1.23, p = 0.54), indicating that all three architectures achieved similar predictive capability on this dataset.

V. Discussion

Comparing the results obtained in this study with those from [8] in scenario 3 (sociodemographic variables, self-reported health information and blood biomarkers), the SVM model developed here outperformed the SVM from reference 8 in terms of accuracy (85% vs. 56%), sensitivity (74% vs. 65%) and specificity (94% vs. 89%). The performance of the GBM (accuracy = 83% and AUC-ROC = 0.88) was slightly superior to that of the XGBM (accuracy = 71% and AUC-ROC = 0.79). The MLP outperformed the ANN across all metrics: accuracy (86% vs. 65%), sensitivity (81% vs. 70%), specificity (90% vs. 84%), and AUC-ROC (0.90 vs. 0.73). This superiority suggests that including the MMSE and structural neuroimaging biomarkers (nWBV) enhances the models' discriminative capacity compared to the blood biomarkers and self-reported health data used in reference 8.

Reference [9] utilized demographic, clinical, cognitive, and MRI data. It was observed that individual architectures, such as LSTM (accuracy = 89.44%) and FNN (accuracy = 89.5%), achieved accuracy levels comparable to those of the models in this study (85%, 83%, and 86%). Employing a hybrid LSTM-FNN approach resulted in an accuracy of 99.82%. This highlights the high potential of combined deep learning models for the early detection of Alzheimer's disease when applied to large structured datasets.

VI. Conclusion

In this study, the developed ML models demonstrated strong performance in identifying dementia, with high efficiency. As demonstrated by the Boruta algorithm and the Friedman test, the integration of just three key predictors (MMSE score, nWBV, and sex) was sufficient to yield effective and statistically equivalent models in predictive capacity (p = 0.54). It suggests that these algorithms constitute effective and suitable prediction tools for application in high-demand clinical environments or with limited technical and temporal resources. Machine learning integrated with rigorous clinical protocols can be a good tool to mitigate the high rates of underdiagnosis of dementia in Brazil.

Although research on dementia detection has grown significantly over the last two decades, driven by collaboration between health and computer science, there remains a need to identify relevant characteristics that improve the predictive capacity of models. This work contributes to the advancement of this field by validating the use of machine learning approaches for diagnostic purposes and investigation of predictive variables.

The longitudinal nature of this dataset represents a significant limitation for training predictive models. Repeating observations of the same individual, even with a minimum interval of one year, can lead to inter-observational correlation, violating the assumption of sample independence. Consequently, although the results demonstrate robust performance, generalizing the models to a population of

patients with completely independent data should be interpreted with caution.

ACKNOWLEDGMENT

The authors would like to thank CAPES, CNPq, FINEP and FAPERJ.

REFERENCES

[1] E. Nichols et al., “Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019”, The Lancet Public Health, v. 7, n. 2, p. e105–e125, fev. 2022, doi: 10.1016/S2468-2667(21)00249-8.

[2] Brasil, “Relatório nacional sobre a demência: epidemiologia, (re)conhecimento e projeções futuras”, Ministério da Saúde, Brasília, 2024.

[3] A. E. Nakamura, D. Opaleye, G. Tani, e C. P. Ferri, “Dementia underdiagnosis in Brazil”, The Lancet, v. 385, n. 9966, p. 418–419, jan. 2015, doi: 10.1016/S0140-6736(15)60153-2.

[4] A. F. Jacinto, S. Brucki, C. S. Porto, M. De Arruda Martins, e R. Nitrini, “Detection of cognitive impairment in the elderly by general internists in Brazil”, Clinics, v. 66, n. 8, p. 1379–1384, ago. 2011, doi: 10.1590/S1807-59322011000800012.

[5] K. Faceli, A.C. Lorena, J. Gama, A.C.P.L.F. Carvalho, T.A. Almeida, Inteligência Artificial-Uma Abordagem De Aprendizado De Máquina, 2º ed. Rio de Janeiro, RJ: Ltc-Livros Tecnicos E Cientificos Editora Lda, 2022.

[6] G. Battineni, N. Chintalapudi, e F. Amenta, “Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)”, Informatics in Medicine Unlocked, v. 16, p. 100200, 2019, doi: 10.1016/j.imu.2019.100200.

[7] M. Shehab et al., “Machine learning in medical applications: A review of state-of-the-art methods”, Computers in Biology and Medicine, v. 145, p. 105458, jun. 2022, doi: 10.1016/j.compbiomed.2022.105458.

[8] M. Jia, Y. Wu, C. Xiang, e Y. Fang, “Predicting Alzheimer’s Disease with Interpretable Machine Learning”, Dement Geriatr Cogn Disord, v. 52, n. 4, p. 249–257, 2023, doi: 10.1159/000531819.

[9] A. C. Mmadumbu, F. Saeed, F. Ghaleb, e S. N. Qasem, “Early detection of Alzheimer’s disease using deep learning methods”, Alzheimer’s & Dementia, v. 21, n. 5, p. e70175, maio 2025, doi: 10.1002/alz.70175.

[10] G. James, D. Witten, T. Hastie, e R. Tibshirani, Org., An introduction to statistical learning: with applications in R. em Springer texts in statistics, no. 103. New York: Springer, 2013.

[11] T. Hastie, R. Tibshirani, e J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Second edition. em Springer Series in Statistics. New York, NY: Springer, 2017.

[12] A. L. Sicsú, Técnicas de machine learning. São Paulo, SP: Editora Edgard Blucher, 2022.

[13] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, e R. L. Buckner, “Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults”, Journal of Cognitive Neuroscience, v. 22, n. 12, p. 2677–2684, dez. 2010, doi: 10.1162/jocn.2009.21407.

[14] R Core Team, _R: A Language and Environment for Statistical Computing_. (2025). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Disponível em: <https://www.R-project.org/>

Mamo.IA: A Web-Based Artificial Intelligence Platform for Breast Cancer Screening and Diagnostic Support Using Thermography, Mammography, and Histopathology Images

Maria Beatriz Martins Pontes Gonçalo
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
mbmpg@cin.ufpe.br

Alberis Alves da Silva
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
aas11@cin.ufpe.br

Gabriela Leite Pereira
Centro de Engenharia da Computação
Universidade de Pernambuco
Recife, Brazil
gabriela.lpereira@ufpe.br

Maria Clara de Oliveira Castro
Departamento de Biomedicina
Universidade Federal de Pernambuco
Recife, Brazil
mariaclara.castro@ufpe.br

Arthur Moreira Nogueira de Melo
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
arthur.nogueiramelo@ufpe.br

Fabíola Fábia da Silva Oliveira
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
fabiola.oliveira@ufpe.br

Vinícius Cavalcanti de Albuquerque Pena
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
vcap@cin.ufpe.br

Lucas Medrado Santos
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
lms16@cin.ufpe.br

Karla Amorim Sancho
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
karla.amorims@ufpe.br

Giselle Machado Magalhães Moreno
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
giselle.moreno@ufpe.br

Juliana Carneiro Gomes Cassemiro
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
juliana.cgomes@ufpe.br

Clarisse Lins de Lima
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
clarisse.lins@ufpe.br

Ana Clara Gomes da Silva
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
clara.gomes@ufpe.br

Wellington Pinheiro dos Santos
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

Maíra Araújo de Santana
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
maira.araujosantana@ufpe.br

***Abstract*—Early diagnosis of breast cancer remains a challenge for healthcare systems, especially in resource-constrained settings, where high demand for specialized examinations can delay diagnostic investigation. In this context, this work presents Mamo.IA, a web-based Artificial Intelligence (AI) platform developed to support screening and analysis of breast images from three distinct modalities: thermography, histopathology, and mammography. Initially, an algorithmic validation stage was conducted using images from public and institutional datasets. The images were characterized using transfer learning with a pre-trained LeNet architecture and classified by different machine learning algorithms, including Random Forest, Support Vector Machine (SVM), J48, Bayes Network, and Naive Bayes. The models were evaluated using stratified 10-fold cross-validation, employing accuracy, Kappa coefficient, sensitivity, specificity, and AUC-ROC as performance metrics. The SVM with RBF kernel achieved the best performance for thermography (accuracy of 76.99% and Kappa of 0.69), while Random Forest was superior for histopathology (85.50% and Kappa of 0.71) and mammography (72.76% and Kappa of 0.59); these models were incorporated into the platform backend. The interface was implemented in Django, enabling user authentication, exam submission, automated processing, generation of diagnostic reports, and results export. The integration between predictive models and a unified web interface demonstrates the feasibility of Mamo.IA as a clinical decision support tool, with potential to contribute to screening organization, patient prioritization, and expanded access to breast cancer diagnostic support technologies**

within the Brazilian Unified Health System (SUS).

***Index Terms*—Breast cancer, AI platform, Mammography, Thermography, Histopathology**

I. Introduction

Breast cancer is the most frequently diagnosed neoplasm among women and a leading cause of cancer-related mortality in females [1]. Late diagnosis remains a significant challenge, particularly in middle-income countries, where structural limitations compromise timely access to screening examinations. Mammography remains the gold standard screening method; however, its high cost, unequal distribution of equipment, use of ionizing radiation, and reduced sensitivity in dense breast tissue motivate the search for complementary technologies. Among these alternatives, infrared thermography stands out as a non-invasive, painless, and radiation-free modality capable of detecting thermal alterations associated with tumor vascularization and metabolic activity [1].

The advancement of Artificial Intelligence (AI) and machine learning has expanded the diagnostic potential of breast thermography. Santana et al. demonstrated the viability of Extreme Learning Machines in thermographies [2], and Silva et al. highlighted feature selection by genetic algorithms and particle swarm optimization [3]. With deep learning, hybrid architectures emerged: Deep-Wavelets with convolutional neural networks [4], Deep-Wavelet with support vector machines [5], and Deep Transfer Learning [6]. Espíndola et al. analyzed thermophysical parameters of superficial breast temperature, reinforcing the scientific foundation of the method [7].

There is also the expansion of intelligent clinical decision support systems. Platforms from the same group, such as Heg.IA and IKONOS, integrated AI into care workflows of the Brazilian Unified Health System (SUS) based on laboratory tests and biomedical images [8], [9]. In this scenario, Mamo.IA is proposed as a support system for early breast cancer screening that integrates infrared thermography, mammography, histopathological analysis, digital image processing, and AI algorithms, with the potential to assist in risk stratification and the organization of healthcare waiting lists.

This perspective is relevant for the SUS, which faces high demand for specialized examinations and limited resources. Studies on AI in Brazilian public health indicate that intelligent systems can optimize screening and patient prioritization [10], [11]. Thus, this work presents Mamo.IA, an AI-based platform to support early breast cancer diagnosis, describing its interface, functional architecture, and application as support for breast cancer screening and patient prioritization in the SUS.

The platform will include efficient machine learning algorithms and a Software as a Service (SaaS) architecture to provide rapid access to analysis tools. It integrates, in a single workflow, professional registration, the capture and preprocessing of breast images, the extraction of relevant features, and the automatic risk classification through machine learning models, allowing the results to be made available quickly to healthcare teams. It is expected that the tool will be able to identify patterns suggestive of breast changes across different examination modalities (thermography, histopathology, and mammography) with accuracy comparable to that of related studies in the literature, contributing to reducing the time between clinical suspicion and referral for confirmatory exams.

To present this development, the paper is structured as follows: Section II presents the related work, discussing recent approaches of computer-aided diagnosis systems applied to breast imaging exams; Section III describes the methodology employed in the development of the Mamo.IA platform, including the algorithmic validation process and the selection of classifiers in the backend, as well as the construction of the web interface; Section IV presents and discusses the results obtained in the training and testing of the classifiers for the thermography, histopathology, and mammography datasets, in addition to the interface usage workflow; finally, Section V presents the final considerations of this work.

II. Related Studies

Within the scope of computerized thermography, Resende (2025) [13] developed a graphical user interface (GUI) in a MATLAB environment that integrates thermal image processing and a CNN architecture optimized via a Bayesian algorithm. The interface performs data augmentation and adaptive filtering tasks to reduce noise from text-based matrices, displaying real-time learning stability with a validation accuracy of 97.91% and 98.95% in the final tests. However, the operational limitations of the GUI include a strict dependence on specific file formats and risks of spectral confusion on the screen, requiring object-oriented segmentation to accurately delineate the lesions.

Freitas et al. (2022) developed an approach based on Multi-Instance Learning (MIL) integrated with a pre-trained DenseNet201 convolutional neural network [14] for the automated diagnosis of breast cancer in histopathological images from the BreakHis dataset. The system performs processing steps involving white noise removal, partitioning of the original image into 16 smaller instances, and the calculation of attention weights, identifying, during the inference phase, the regions with the highest malignant relevance, achieving an accuracy of 90% and a sensitivity of up to 98% in the binary classification between benign and malignant tumors. However, the limitations of this approach include the high computational cost resulting from the multi-input architecture and the performance degradation observed when excessively deep dense layers are employed, necessitating the use of structural optimization strategies to enhance the system's generalization capability.

Regarding thermography, Attallah (2025) [15] proposed a Computer-Aided Diagnosis (CAD) system focused on thermographic images from the public DMR-IR dataset. The study employed lightweight Convolutional Neural Networks (MobileNet, EfficientNetB0, and ShuffleNet) based on the Transfer Learning technique and applied Non-Negative Matrix Factorization (NNMF) for dimensionality reduction. The

model achieved 99.9% accuracy using only 350 final optimized parameters. The main advantage of this approach is the substantial reduction of the computational load for analysis, ensuring efficient and radiation-free screening, although it presents the limitation of evaluating only a single examination modality within a relatively small dataset.

Focusing on the confirmation stage, Atrey, Singh, and Bodhey (2023) [16] investigated the classification of tumors in histopathological images at 40× magnification, extracted from the Brazilian collaborative BreaKHis dataset. The methodology involved the extraction of texture and color features, followed by a feature selection step based on a Genetic Algorithm (GA) to select the most relevant attributes, and classification using a Long Short-Term Memory (LSTM) neural network architecture. The model achieved an accuracy of 99.85% in classifying benign and malignant lesions by combining the 192 attributes selected by the genetic filter. The system demonstrated efficiency in assisting with biopsy diagnoses, reducing inter-observer discrepancies and laboratory workload, despite relying solely on 40× magnification and addressing class imbalance within the dataset.

Regarding intelligent systems already applied to SUS diagnostic workflows, Gomes et al. (2020) [10] developed IKONOS, an application designed to support the diagnosis of COVID-19 and pneumonia using chest X-ray images. The application extracts features and classifies images into four categories: healthy patients, viral pneumonia, bacterial pneumonia, and COVID-19. They displayed accuracy, sensitivity, and specificity metrics to assist professionals in decision-making. For feature extraction, the system employed Haralick and Zernike extractors combined with classical classifiers such as Multilayer Perceptron, Support Vector Machine, Random Trees, Random Forest, Bayesian Networks, and Naive Bayes. Among these, Support Vector Machines stood out, achieving an average accuracy of 89.78%, with an average sensitivity of 0.8979 and average precision and specificity of 0.8985 and 0.9963, respectively. The tool demonstrates the feasibility of implementing low-cost, texture-based diagnostic support directly into clinical routine, reinforcing the rationale for platforms like Mamo.IA. However, its reliance on handcrafted texture and shape descriptors, rather than deep feature learning, may limit generalization when applied to the more heterogeneous imaging conditions typically found in resource-constrained healthcare environments.

III. Materials and Methods

The study aims to develop advanced algorithms for assessing medical images used in the context of breast cancer screening and diagnosis and to create an artificial intelligence SaaS platform. This initiative seeks to optimize early breast cancer diagnosis, reduce costs, contribute to the training of healthcare professionals, and drive significant advancements in the field by providing innovative models and methods for future clinical applications.

The proposed platform is projected to be used by healthcare professionals in the usage workflow proposed in Figure 1. Upon arriving at the healthcare facility, the patient undergoes screening via breast thermography; the resulting image can be sent to the system to aid in decision-making regarding the presence or absence of potential findings. If there are no indications of lesions, the patient may be discharged. Otherwise, the professional may choose to refer her for further investigation using other imaging modalities. In addition to breast thermography image analysis, the first version of this platform supports decision-making based on mammography and microscopy images within the context of histopathological analysis.

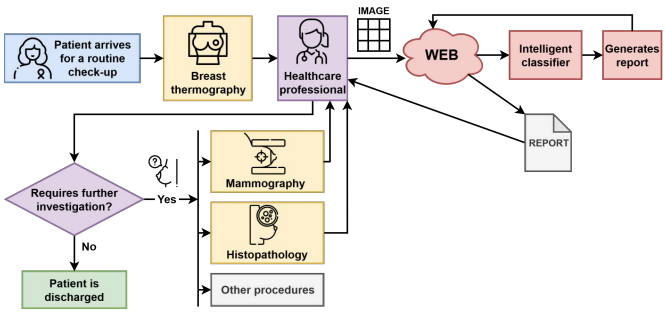


Fig. 1: Proposed workflow for using the Mamo.IA platform. The yellow blocks represent the imaging modalities implemented in the first version of the platform. In the red blocks we illustrate the processing stages based on our machine learning models.

The following sections describe the steps for developing the first version of the platform. We begin by presenting the processing infrastructure (back-end) used to support decision-making. Next, we introduce the steps taken to create the user interaction interface (front-end).

A. Backend Development

Before the development of the interface, a preliminary stage was conducted to identify the best-performing classifier for each dataset. Characterized as cross-sectional research with a quantitative approach, this stage constitutes an algorithmic validation study integrated into the development of the backend of the Mamo.IA platform.

In these experiments, the performance of different classification models was investigated for the automatic suggestion of diagnoses based on medical images of multiple modalities (thermal, histopathological, and mammographic). The central objective of this validation was to ensure the accuracy of the system in supporting the early detection of breast changes through Artificial Intelligence.

Three datasets were used, one for each modality addressed. The breast thermographic image dataset was the MammoTherm, which includes 336 images obtained from 100 patients treated at the Hospital das Clínicas of the Federal University of Pernambuco, using a FLIR S45 infrared camera [17]. For each patient, they captured eight images at different angles: frontal with hands on hips (T1), frontal with hands raised (T2), right breast (MD), left breast (ME), inner lateral right breast

(LIMD), inner lateral left breast (LIME), outer lateral right breast (LEMD), and outer lateral left breast (LEME). However, in this study, only the images in the T1 and T2 positions were used, as this view has favored classification. The images are categorized into four groups: no lesion (n = 66), benign lesion (n = 121), malignant lesion (n = 76), and cyst (n = 73).

Regarding the histopathology data, microscopy images from the BreaKHis dataset were used, which was developed by the Federal University of Paraná and contains 7,909 breast cancer histopathological images from 82 patients, acquired at different magnifications [18]. In this study, exclusively the images with 40x magnification were used, categorized as benign tumor and malignant tumor.

Finally, the mammography images used are from the IRMA (Image Retrieval in Medical Applications) dataset, which consists of fragments (patches) of digitized mammograms used to train and test AI models in the context of breast assessment based on the BI-RADS system [19]. The images were collected from widely known datasets in the related literature: DDSM: 2,479 studies, totaling 9,916 mammograms (695 healthy cases, 870 benign, and 914 malignant); MIAS: 322 mammograms; LLNL: 197 mammograms; and RWTH: 170 mammograms. The *patches* were standardized with dimensions of 128×128 pixels and classified into the categories: malignant lesion, benign lesion, and normal. In the 12er_patches configuration, 2,796 *patches* were used, equally distributed across 12 classes (233 images per class), and only the classes corresponding to the aforementioned clinical categories were selected for this research.

For all image modalities, the adopted processing was similar. Initially, image characterization from numerical data (feature extraction) was performed using the transfer learning technique with the LeNet convolutional neural network architecture, pre-trained on the MNIST dataset. Although the network was originally pre-trained on a dataset involving a substantially different image classification problem, namely handwritten digit recognition, it is important to emphasize that the pre-trained LeNet model was not directly employed to perform the classification of breast cancer images. Instead, it was used exclusively as a robust feature extractor, leveraging the capacity of convolutional neural networks to learn and represent relevant visual patterns from images. The extracted representations were subsequently provided as input to conventional machine learning classifiers, which were responsible for the final classification task. This step was performed through the Weka 3.8 software, with support from the *deeplearning4j* (DL4J) library. This feature extraction approach was adopted to simultaneously leverage the robustness of image characterization from deep neural networks and reduce computational cost, eliminating the need for hyperparameter tuning of a deep classification model, which is a relevant factor for practical application in contexts with limited computational infrastructure, such as in many SUS services.

Subsequently, the datasets were divided into training (75%) and testing (25%) sets using the random resampling method (Resample). Furthermore, considering the imbalance in the amount of data per class, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to generate synthetic samples of the minority classes in the training set and, thus, avoid bias during the training of the classifiers.

Regarding the classifiers, five supervised classification algorithms were tested: Random Forest, with variation in the number of trees (10, 20, 50, 100, 150, 200, 250, and 300); Support Vector Machine (SVM), with linear, polynomial (degree 2), and radial (RBF, $\gamma = 0.5$) kernels; J48 decision tree; Bayesian Network (Bayes Network); and Naive Bayes.

The models were evaluated using stratified cross-validation with 10 folds, repeated over 30 runs, ensuring statistical robustness. To prevent data leakage, cross-validation was performed exclusively on the training set, with each instance corresponding to an individual patient. Therefore, the stratified 10-fold cross-validation was conducted at the patient level, ensuring that no patient could simultaneously contribute samples to the training and validation folds. The performance metrics considered were: accuracy, kappa coefficient, sensitivity, specificity, and area under the ROC curve (AUC-ROC). The results were organized into CSV files for subsequent statistical analysis and comparison between algorithms.

After the identification and statistical validation of the best-performing classifier for each dataset within the Weka environment, the transition phase to the backend production environment began. Using the Python programming language, the selected models were implemented and replicated for their respective equivalent datasets, integrating specialized machine learning libraries to ensure algorithm portability. This Python stage consolidated the analytical intelligence of the platform, allowing the optimized classifiers to operate natively and integrated into the request and data processing flow of the Mamo.IA backend.

B. Interface Development

The development of the Mamo.IA platform interface was conducted with a focus on usability and efficient integration with the predictive models operating in the back-end. The web system architecture was structured using the Django framework, chosen for its structural robustness in Python and ease of integration with Artificial Intelligence libraries.

The platform focuses on expanding access to quality breast diagnosis within the context of the Brazilian Unified Health System (SUS), especially in regions lacking specialized professionals, such as mastologists and radiologists, contributing to the reduction of health inequalities.

The materials and technologies employed in the construction comprised:

- **Markup and Styling Languages:** HTML5 and CSS3, implemented to ensure a clean, responsive, and adaptable design to different resolutions, avoiding distortions in exams of varied clinical proportions.
- **Back-end Framework and Routing:** Django, responsible for view control, user session management, and strict form validation.

- **Database:** SQLite, used in the development environment to store upload histories, structured reports, and access credentials.

The development stages followed an iterative logical workflow. Initially, the authentication flow and the main dashboard were structured. Next, the dynamic exam submission form was implemented, which features database validation to categorize clinical entries (Mammography, Thermography, and Histopathology). The final stage consisted of building the results visualization interface. This screen was designed with layout security constraints to display the proportionally resized medical exam, accompanied by the respective diagnosis generated by the model, in addition to integrating PDF export and email sending functionalities.

IV. Results e Discussion

A. Training and Testing of Classifiers

The results obtained from the training of the classifiers for thermography, histopathology, and mammography are presented, respectively, in Tables I, II, and III.

The experiments demonstrated that the best-performing classifier for the thermography dataset was the SVM with RBF kernel ($\gamma = 0.5$). On the other hand, for the histopathology and mammography datasets, the Random Forest algorithm achieved the best results. Aiming to optimize computational cost, the configuration with 200 trees was selected, as this parameter ensured the maintenance of high accuracy without compromising the values of the other performance metrics.

Upon analyzing Tables I, II, and III in more detail, this superior performance on the thermography dataset (accuracy of 76.99%, Kappa of 0.69) can be associated with the capacity of this kernel to model non-linear decision boundaries, which are suitable for the strong spatial correlation between the temperatures recorded in adjacent regions of the breast. Conversely, the worst performance was observed with Naive Bayes (accuracy of 50.19%, Kappa of 0.34), closely followed by the Bayes Network. Both classifiers are based on the assumption that the attributes are independent of each other, which is not sustainable given the thermal correlation between neighboring regions of the breast.

In the histopathology dataset, Random Forest with 300 trees achieved the best result (accuracy of 85.50%, Kappa of 0.71), which is likely due to its ensemble nature, conferring greater robustness to correlated features and noise present in the data extracted via transfer learning from the LeNet network. Naive Bayes, in turn, presented the worst performance with an accuracy of 29.61%, evidencing difficulty in generalizing from these same features. Since the LeNet network was originally trained for the domain of handwritten digits (MNIST), the extracted features tend to be numerous and related to each other in a complex and non-linear manner, which contradicts the independence assumption made by the Bayesian classifier.

In the mammography dataset, Random Forest again presented the best performance, although with an accuracy lower than that obtained in the other modalities (72.76%), which is consistent with the greater visual complexity of this dataset for diagnosis. The worst performance was observed with the RBF SVM (accuracy of 66.86%, Kappa of 0.50), a result that contrasts with the good performance of this same kernel in the thermography dataset. This inversion suggests that the gamma parameter is sensitive to the scale and variability of the features of each modality, which points to the need to adjust it specifically for each domain in future work.

In the tests, the datasets obtained the metrics found in Table IV. The confusion matrices of the tests are shown in Figure 2.

TABLE IV: Table showing the database test metrics.

Database	Accuracy	Kappa	Sensitivity	Specificity	AUC-ROC
Thermography	100%	1.0	1.0	1.0	1.0
Histopathology	100%	1.0	1.0	1.0	1.0
Mammography	75.54%	0.63	0.76	0.88	0.88

Classified as	a	b	c	d
a = Cyst	14	0	0	0
b = Benign_lesion	0	23	0	0
c = Malignant_lesion	0	0	15	0
d = No_lesion	0	0	0	13

(a) Confusion matrix for the Thermography Database.

Classified as	A	B
A = histo_benign	110	0
B = histo_malignant	0	252

(b) Confusion matrix for the Histopathology Database.

Classified as	a	b	c
a = T_benign	173	60	0
b = T_malignant	111	122	0
c = T_normal	0	0	233

(c) Confusion matrix for the Mammography Database.

Fig. 2: Confusion matrices obtained by the best classifiers for each database.

The results obtained in the test with the thermography dataset, presented in Table IV and 2a, revealed an ideal performance, with perfect metrics and confusion matrix. This behavior may be associated with the reduced number of samples available in the test set, which simplified the classifier's decision boundary for this specific subset.

Analogously, the evaluation of the model on the histopathology dataset (Table IV and Figure 2b) exhibited maximum performance, evidenced by perfect metrics and a confusion matrix without incorrect classifications. This flawless result suggests a probable influence of the restricted sample size in

TABLE I: Performance metrics for classifiers in the Thermography Database.

Algorithm	Configuration	Accuracy		Kappa Index		Sensitivity		Specificity		AUC-ROC	
		Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Random Forest	10 trees	60.3426	8.1597	0.4712	0.1088	0.6870	0.1522	0.8171	0.0751	0.8266	0.0807
Random Forest	20 trees	65.9537	7.6828	0.5460	0.1024	0.7218	0.1535	0.8525	0.0707	0.8707	0.0661
Random Forest	50 trees	70.8333	6.8249	0.6111	0.0909	0.7692	0.1371	0.8766	0.0649	0.9044	0.0552
Random Forest	100 trees	72.6666	6.6359	0.6355	0.0885	0.7804	0.1339	0.8854	0.0599	0.9157	0.0533
Random Forest	150 trees	73.2963	6.9340	0.6439	0.0924	0.7885	0.1336	0.8890	0.0619	0.9199	0.0523
Random Forest	200 trees	73.6667	6.5783	0.6489	0.0877	0.7948	0.1306	0.8923	0.0618	0.9219	0.0511
Random Forest	250 trees	74.0926	6.7240	0.6546	0.0897	0.8026	0.1304	0.8919	0.0622	0.9233	0.0503
Random Forest	300 trees	74.1759	6.7405	0.6557	0.0899	0.8033	0.1274	0.8923	0.0633	0.9239	0.0496
SVM	Linear	70.7407	6.6829	0.6099	0.0891	0.8063	0.1340	0.8709	0.0610	0.8655	0.0659
SVM	Polynomial 2	73.8889	6.2382	0.6518	0.0832	0.8452	0.1163	0.8799	0.0585	0.8849	0.0599
SVM	**RBF 0.50**	**76.9907**	**6.7793**	**0.6932**	**0.0904**	**0.6285**	**0.1628**	**0.9947**	**0.0130**	**0.9226**	**0.0622**
J48	-	69.0476	2.7624	0.5357	0.0414	0.5714	0.1100	0.7428	0.0494	0.6990	0.0371
Bayes Network	-	51.6111	7.6014	0.3548	0.1013	0.5767	0.1496	0.8255	0.0731	0.7785	0.0892
Naive Bayes	-	50.1852	8.0908	0.3358	0.1079	0.5992	0.1514	0.7897	0.0747	0.7350	0.0949

TABLE II: Performance metrics for classifiers in the Histopathology Database.

Algorithm	Configuration	Accuracy		Kappa Index		Sensitivity		Specificity		AUC-ROC	
		Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Random Forest	10 trees	75.9578	3.2082	0.5191	0.0641	0.8370	0.0401	0.6821	0.0517	0.8512	0.0278
Random Forest	20 trees	79.6337	2.8408	0.5926	0.0568	0.8386	0.0377	0.7540	0.0459	0.8874	0.0234
Random Forest	50 trees	83.2054	2.6164	0.6641	0.0523	0.8454	0.0373	0.8186	0.0381	0.9157	0.0196
Random Forest	100 trees	84.4905	2.5617	0.6898	0.0512	0.8446	0.0360	0.8451	0.0376	0.9267	0.0176
Random Forest	150 trees	84.9684	2.4182	0.6993	0.0483	0.8451	0.3594	0.8542	0.3417	0.9306	0.1714
Random Forest	**200 trees**	**85.3108**	**2.5131**	**0.7062**	**0.5026**	**0.8462**	**0.3629**	**0.8599**	**0.3442**	**0.9329**	**0.1684**
Random Forest	250 trees	85.4217	2.4028	0.7084	0.4805	0.8458	0.3581	0.8625	0.3365	0.9342	0.1674
Random Forest	300 trees	85.4991	2.4560	0.7099	0.4911	0.8457	0.3628	0.8641	0.3387	0.9348	0.1664
SVM	Linear	69.5553	2.1597	0.0804	0.0642	0.1148	0.0457	0.9485	0.0249	0.5316	0.0255
SVM	Polynomial 2	66.1627	3.5099	0.2001	0.0811	0.4445	0.0745	0.7562	0.0407	0.6003	0.0413
SVM	RBF 0.50	74.3514	2.2659	0.2451	0.0741	0.2315	0.0593	0.9665	0.0186	0.5990	0.0314
J48	-	65.1819	3.6318	0.1759	0.0828	0.4263	0.0803	0.7500	0.0468	0.5872	0.0522
Bayes Network	-	62.4223	3.5351	0.2484	0.0707	0.6998	0.0475	0.5487	0.0520	0.6963	0.0392
Naive Bayes	-	29.6070	3.1563	0.3395	0.1022	0.5816	0.0787	0.7552	0.0634	0.6193	0.0330

the test set, which tended to make the algorithm's separation boundary less complex.

Unlike the scenarios observed in the thermography and histopathology datasets, the results obtained for the mammography dataset (Table IV and Figure 2c) did not present perfect metrics. Using the Random Forest classifier, the model achieved an accuracy of 75.54% and a Kappa index of 0.63. This behavior is justified by the greater complexity and intrinsic visual variability of mammographic images, in addition to the substantially larger sample size in this dataset, which imposed a real generalization challenge for the algorithm.

B. Interface Presentation and Usage Flow

The final interface establishes the visual and interactive bridge between the screening algorithms and the healthcare professional. Figure 3 displays the main access dashboard of the Mamo.IA platform. From this screen, the user is directed to the exam submission interface (Figure 4), which enables the upload of medical files and the categorized selection of the clinical procedure performed.

C. Viewing Diagnostics and Reports

Following the algorithmic processing, the platform displays the analysis results on a stabilized interface. Figures 5 and 6 illustrate representations of reports generated by the system,

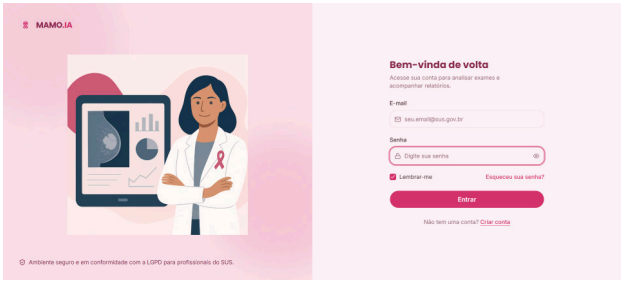


Fig. 3: Home page of the Mamo.IA platform, displaying the acess dashboard.

detailing the resized image without loss of proportion (aspect ratio) and the diagnosis issued by the AI.

Complementarily, Figures 7 and 8 demonstrate the system's flexibility in handling different types of results (positive and negative findings) for the submitted examinations. The interface layout ensures that the reading panel and action buttons remain contained within a single viewing area, eliminating the need for excessive scrolling and optimizing decision-making in the clinical environment.

TABLE III: Performance metrics for classifiers in the Mammography Database.

Algorithm	Configuration	Accuracy		Kappa Index		Sensitivity		Specificity		AUC-ROC	
		Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Random Forest	10 trees	70.6968	2.7097	0.5604	0.5785	0.6672	0.6714	0.7268	0.7428	0.7902	0.7903
Random Forest	20 trees	71.2929	2.5530	0.5693	0.5500	0.6474	0.5857	0.7456	0.7571	0.7997	0.7902
Random Forest	50 trees	72.1354	2.7160	0.5820	0.5785	0.6399	0.6000	0.7620	0.7785	0.8074	0.8114
Random Forest	100 trees	72.4838	2.5256	0.5872	0.6142	0.6394	0.6571	0.7675	0.7851	0.8106	0.8301
Random Forest	150 trees	72.6507	2.4183	0.5898	0.0363	0.6397	0.0530	0.7699	0.0280	0.8121	0.0204
Random Forest	**200 trees**	**72.7254**	**2.4594**	**0.5909**	**0.0369**	**0.6402**	**0.0519**	**0.7710**	**0.0285**	**0.8131**	**0.0210**
Random Forest	250 trees	72.7396	2.4586	0.5911	0.0369	0.6395	0.0520	0.6395	0.0520	0.8134	0.0208
Random Forest	300 trees	72.7589	2.4821	0.5914	0.0372	0.6403	0.0538	0.7712	0.0286	0.8137	0.0209
SVM	Linear	67.7109	2.2050	0.5156	0.0330	0.8868	0.2177	0.5724	0.1296	0.7578	0.0164
SVM	Polynomial 2	68.4042	2.4457	0.5260	0.0366	0.8066	0.2612	0.6235	0.1450	0.7631	0.0184
SVM	RBF 0.50	66.8572	1.1553	0.5028	0.0172	0.9841	0.0700	0.5139	0.0475	0.7532	0.0092
J48	-	68.9058	2.7625	0.5336	0.0414	0.5973	0.1100	0.7367	0.0495	0.6988	0.0372
Bayes Network	-	72.7453	2.6186	0.5912	0.0393	0.7025	0.0504	0.7400	0.0295	0.8103	0.0239
Naive Bayes	-	71.3283	2.4003	0.5700	0.0360	0.7443	0.0525	0.6978	0.0525	0.7977	0.0223

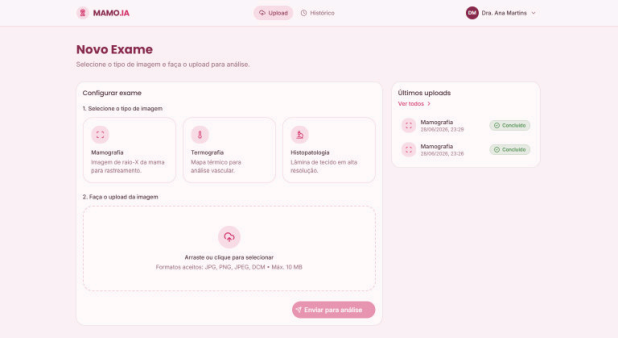


Fig. 4: Dynamic form interface for medical file submission and



Fig. 5: Demonstration of the results interface displaying the diagnosis and the analyzed exam.

V. Conclusion

The development of computational tools capable of supporting the screening and early diagnosis of breast cancer represents a promising strategy to expand access to specialized care, especially in high-demand scenarios, such as the Brazilian Unified Health System (SUS). In this context, this work presents Mamo.IA, a web platform structured as Software-as-a-Service (SaaS), developed to integrate different modalities of medical images into a single clinical decision-support environment.



Fig. 6: Visual representation of a processed report, integrated with export buttons.

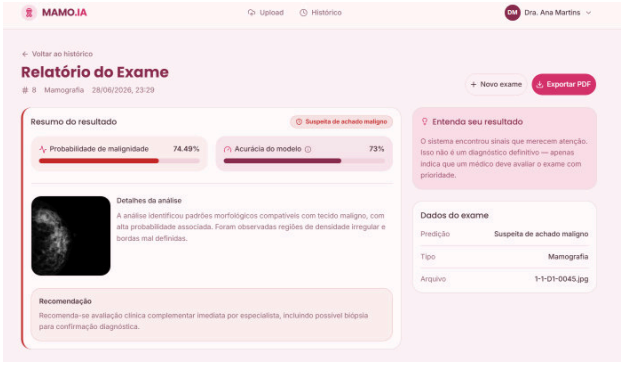


Fig. 7: Results interface detailing an analysis with positive clinical findings.

The platform was built upon the validation of machine learning algorithms applied to thermography, histopathology, and mammography datasets, followed by the implementation of the selected models in a backend developed in Python and integrated with the Django framework. Among the evaluated classifiers, the SVM with an RBF kernel achieved the best performance for the thermography dataset, while the Random Forest, configured with 200 trees, was adopted for the

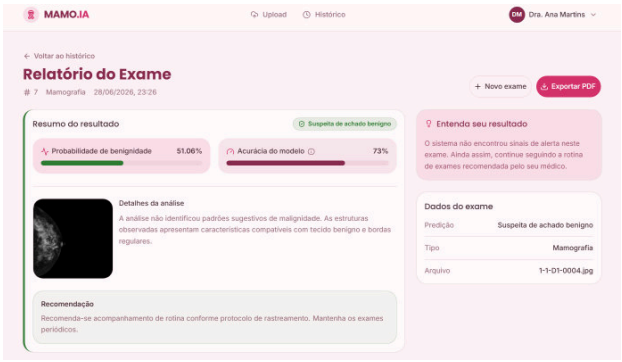


Fig. 8: Results interface detailing an analysis showing no abnormalities (negative).

histopathology and mammography datasets due to its ability to maintain high performance at a lower computational cost. The integration of these models resulted in an interface capable of handling the categorized submission of examinations, generating automated reports, and consolidating different image modalities into a single user workflow. The results obtained demonstrate the feasibility of Mamo.IA as a clinical decision-support tool, leveraging artificial intelligence throughout the diagnostic process to help prioritize patients for confirmatory examinations within the scope of the SUS.

The independent processing of each imaging modality reflects the intended clinical workflow of Mamo.IA, in which the different modalities are not necessarily acquired or analyzed simultaneously, but may be requested and incorporated at different stages according to the patient's clinical needs and the diagnostic pathway. Thus, rather than combining all modalities into a single multimodal model, the platform was designed to provide complementary decision support as additional examinations become available.

However, this study presents limitations related to the use of specific datasets, the lack of validation in real-world clinical settings, and the need for hyperparameter optimization for each image modality. Additionally, aspects concerning the platform's usability and its impact on the clinical workflow require further investigation. In this regard, future works may encompass the multicenter validation of the tool, its evaluation in real clinical scenarios, the incorporation of new examination modalities, and the refinement of the predictive models, thereby contributing to expanding its generalizability and applicability in clinical practice.

VI. Declaration of Artificial Intelligence Use

The Qwen3.6-Plus, ChatGPT4, and Claude 4.5 language models were employed for orthographic, grammatical editing of the manuscript and on the support on structuring the interface. No artificial intelligence was involved in the conception of any part of the research itself, nor in methodological design, data acquisition or analysis, interpretation of results, or generation of intellectual content. All AI-assisted processes and outputs were rigorously validated and audited by the human authors, who assume full responsibility for the accuracy, originality, and scientific integrity of this work.

Acknowledgment

The authors gratefully acknowledge the financial and institutional support provided by the Brazilian funding agencies FACEPE, CAPES, and CNPq, the Department of Science and Technology of the Ministry of Health (DECIT/MS), and the Instituto do Complexo Econômico-Industrial da Saúde (iCEIS).

References

[1] INCA, "Controle do câncer de mama no Brasil : dados e números 2025," 2025.
[2] Santana MA et al. Infrared Thermography: Principles and Applications in Healthcare. CRC Press, 2024.
[3] Santana MA et al. Breast cancer diagnosis based on mammary thermography and extreme learning machines. Research on Biomedical Engineering, 2018.
[4] Silva ALR et al. Features Selection Study for Breast Cancer Diagnosis Using Thermographic Images, Genetic Algorithms, and Particle Swarm Optimization. IJAIML, 2021.
[5] Barbosa VAF et al. Deep-Wavelets and convolutional neural networks to support breast cancer diagnosis on thermography images. CMBBE-IV, 2022.
[6] Santana MA et al. Combining deep-wavelet neural networks and support-vector machines to classify breast lesions in thermography images. Health and Technology, 2022.
[7] Silva AF et al. Detection and classification of breast lesions in thermographic images by deep transfer learning hybrid architectures. Research on Biomedical Engineering, 2024.
[8] Espíndola NA et al. Numerical analysis of the influence of thermophysical parameters on the surface temperature of breast substitute geometry obtained from infrared images. International Journal of Thermal Sciences, 2026.
[9] Barbosa VAF et al. Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests. Research on Biomedical Engineering, 2021.
[10] Gomes JC et al. IKONOS: an intelligent tool to support diagnosis of COVID-19 by texture analysis of X-ray images. Research on Biomedical Engineering, 2020.
[11] Borges FT et al. LLM Assisted Scoping Review of Artificial Intelligence in Brazilian Public Health. IJERPH, 2026.
[12] Santana MA, Sancho KA, et al. Termografia Assistida por Inteligência Artificial para o Diagnóstico Precoce do Câncer de Mama. In: Saúde Digital a Serviço da Saúde Coletiva, 2025.
[13] Resende GP. Comparação de modelos de CNN na detecção de câncer de mama utilizando imagens infravermelhas multiângulo, 2025.
[14] Freitas MP, et al. "Aplicando MultiInstance Learning (MIL) para o Diagnóstico de Câncer de Mama em Imagens Histopatológicas." Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). SBC, 2022.
[15] Attallah O, "A Deep Learning-Driven CAD for Breast Cancer Detection via Thermograms: A Compact Multi-Architecture Feature Strategy," *Applied Sciences*, vol. 15, no. 13, p. 7181, 2025.
[16] K. Atrey, B. K. Singh, and N. K. Bodhey, "Multi-Feature Classification of Breast Cancer Histopathology Images: An Experimental Investigation in Machine Learning and Deep Learning Paradigm," *Brazilian Archives of Biology and Technology*, vol. 66, 2023, doi: 10.1590/1678-4324-2023220297.
[17] OLIVEIRA, MM. Desenvolvimento de protocolo e construção de um aparato mecânico para padronização da aquisição de imagens termográficas de mama. 2012. 104 f. Diss. Dissertação (Mestrado em Engenharia Mecânica)–Universidade Federal de Pernambuco, Recife, 2012.
[18] SPANHOL, Fabio A. et al. A Dataset for Breast Cancer Histopathological Image Classification. IEEE Transactions on Biomedical Engineering, v. 63, n. 7, p. 1455, 2016.
[19] Deserno, Thomas M., et al. "Computer-aided diagnostics of screening mammography using content-based image retrieval." Medical Imaging 2012: Computer-Aided Diagnosis. Vol. 8315. SPIE, 2012.

MEADS-Health: An Auditable Methodology for Exploring Sensitive Institutional Data with AI-Assisted Analytical Planning

Adriell Gomes Marques
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
e-mail: adriell.gomes@alu.ufc.br
ORCID: 0000-0001-6274-1961

José Jerovane Nascimento
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
e-mail: jerovane@alu.ufc.br
ORCID: 0000-0002-7361-4316

Marcos Vinícius Candido de Figueiredo
Núcleo de Estudos e Pesquisas do Norte e Nordeste (NEPEN)
Fortaleza, Brazil
e-mail: marcos.figueiredo@nepen.org.br
ORCID: 0000-0002-7221-7255

Guilherme Lincoln de Lima Tavares
Programa de Pós-Graduação em Engenharia de Telecomunicações (PPGET)
Instituto Federal do Ceará (IFCE)
Fortaleza, Brazil
e-mail: guilherme.tavares08@aluno.ifce.edu.br

Katerine Chaves Trajano
Núcleo de Estudos e Pesquisas do Norte e Nordeste (NEPEN)
Fortaleza, Brazil
e-mail: katerine.trajano@nepen.org.br

Carlos Henrique Lima Santos
Núcleo de Estudos e Pesquisas do Norte e Nordeste (NEPEN)
Fortaleza, Brazil
e-mail: carlos.santos@nepen.org.br

7ª Ana Sáskya Vaz de Araújo
Centro Diagnóstico Girassol
Prefeitura Municipal de Fortaleza (PMF)
Fortaleza, Brasil
e-mail: anasaskya@gmail.com

8º André Luiz Carneiro de Araújo
Departamento de Telemática (DETEL)
Instituto Federal do Ceará (IFCE)
Fortaleza, Brasil
e-mail: andreluiz@ifce.edu.br
ORCID: 0000-0002-3565-2039

***Abstract*—Administrative data preserve institutional memory, but their reuse can be distorted by counting units, linkage failures, missing territorial information, and disclosure decisions. MEADS-Health is an auditable methodology for exploring heterogeneous institutional data with language-model support restricted to analytical planning. Records remain in a protected environment. Only objectives and sanitized metadata are sent to the model. Formulas, denominators, linkage rules, and release levels remain under technical-team control. The method was examined through a qualitative, artifact-centered municipal case study and a bounded synthetic linkage stress test. The V7 planning grammar exposed 17 reusable axis templates, of which 10 were activated for the health-source configuration. Across 30 synthetic replicates per noise level, each with 5,000 identities, deterministic match recovery decreased from 100.0% without perturbation to 93.1% under 20% independent field noise, while mean precision remained 99.8%. The synthetic test evaluates only the key-matching rule; no real-data gold standard, alternative-pipeline comparison, inferential analysis, or end-to-end effectiveness assessment was performed. These results support claims about inspectability and the bounded robustness of the matching rule. They do not support population inference, operational efficiency, or a privacy guarantee.**

***Index Terms*—sensitive data, public health, persons with disabilities, exploratory data analysis, language models, record linkage, georeferencing, auditing.**

I. Introduction

Administrative datasets are produced to deliver services and document institutional activity, not to answer research questions. Their value for secondary analysis is therefore inseparable from the routines that generated them: a row may represent a person, procedure, dispensing event, program link, or administrative occurrence. If that unit is not defined before counting, the resulting indicator can change meaning rather than merely lose precision. This is why reuse of routinely collected data requires explicit reporting of provenance and analytical units [1], [2].

In health settings, this structural ambiguity coexists with sensitive content such as clinical codes, medications, procedures, addresses, and care notes protected by the Brazilian

General Data Protection Law (*Lei Geral de Proteção de Dados Pessoais*—LGPD) [3]. Analyses involving persons with disabilities must be useful for planning without turning data integration into an additional source of exposure; this concern is consistent with WHO recommendations and the Brazilian Law for the Inclusion of Persons with Disabilities [4], [5].

Data quality cannot be reduced to file cleaning. Established frameworks emphasize completeness, correctness, concordance, conformance, and plausibility [6]–[8]. A source may faithfully represent an administrative process and still be unsuitable for population inference, just as a plausible key may be insufficient to establish identity across systems. Dashboards can hide these uncertainties when duplicate handling, linkage failures, or territorial gaps disappear behind final charts [9]–[11].

LLMs can help organize questions from a sanitized schema, but that usefulness does not justify transferring identifiable records to external services. WHO and NIST guidance emphasize minimization, oversight, transparency, and risk management, while health-AI literature warns against unverified outputs and unnecessary exposure [12]–[15]. The methodological gap addressed here concerns the workflow between AI-assisted planning and record-level execution. MEADS-Health keeps these stages separate while preserving analytical units, linkage provenance, territorial failures, and the calculation path behind each product.

This paper presents MEADS-Health, an analytical-engineering methodology examined in a municipal case involving sensitive health and cross-sector records of persons with disabilities. The records remain under restricted access; the scientific object is the process used to examine, link, calculate, and release auditable aggregates. MEADS-Health separates planning, execution, and validation. The LLM proposes initial analytical axes from sanitized metadata. The technical team defines formulas, denominators, linkage rules, interpretation, and disclosure, and all calculations are performed locally. For biomedical engineering, this addresses the data-engineering and governance layer that precedes secondary analysis of sensitive health information. MEADS-Health is not a clinical decision-support system or predictive algorithm.

II. Related Work

The supporting literature was organized as a problem-oriented narrative review because MEADS-Health connects concepts that are usually discussed in separate domains [16]. For data quality and provenance, health-data frameworks show that secondary use depends on the meaning of the source, not only on missingness or format, and provenance research extends this requirement to transformations and computational workflows [7], [8], [17]–[19]. For this reason, formulas, parameters, denominators, and intermediate decisions remain inspectable alongside the dashboard.

Record linkage adds a second layer of uncertainty. Fellegi–Sunter and later data-matching work formalize identity resolution as a process of preparation, comparison, and classification rather than a simple join [20], [21]. Public-health PPRL literature likewise shows that accuracy depends on field completeness and stability, and that matching errors can propagate bias [22]–[24]. MEADS-Health does not claim a cryptographic PPRL protocol: identifiers are pseudonymized inside an authorized environment, while disclosure control remains a separate step involving aggregation, granularity, and residual reidentification risk under LGPD and NIST guidance [3], [25].

Finally, dashboards, geospatial analysis, and human–AI interaction share a governance problem: useful interfaces can still conceal incomplete coverage or uncertain transformations. Dashboard studies emphasize audience, context, and iterative validation [9]–[11]; geocoding introduces both analytical value and disclosure risk [26]; and human–AI guidance stresses limitations, correction, and oversight [12], [27]. These strands support the central design choice of MEADS-Health: define in advance what may reach the AI and keep record-level processing, validation, and release authority in the protected environment.

III. Materials and Methods

A. Study Design and Evaluation Scope

The study follows an applied, *design-science* approach [28] with a qualitative, artifact-centered evaluation. The object is the pipeline, not the population represented in the datasets; no prevalence, causal effect, healthcare effectiveness, or comparative end-performance is estimated.

The case study used pre-existing municipal sources concerning persons with disabilities, including health and other sectors, inside an authorized institutional analytical environment. There was no participant recruitment or intervention, and no person-level findings are disclosed. Under the LGPD minimization principle [3], the paper reports neither record volumes nor frequencies, coverage rates, rare categories, or extracts. These controls do not replace ethics review when required by a future study design.

Inferential statistics are not supported because the sources are administrative rather than probability samples, service contact affects inclusion, extraction periods and coding practices differ, repeated rows may be duplicates or legitimate events, missingness is workflow-dependent, identifiers can be incomplete or unstable, and complete population denominators are unavailable for some sources. Counts and coverage ratios are therefore treated only as descriptive accounting signals, without confidence intervals, hypothesis tests, or causal interpretation.

B. Analytical Units, Linkage, and Privacy

Administrative data accommodate different units of observation. A record preserves an occurrence produced by the system; a person represents the unique entity to which multiple occurrences may be associated; and a link expresses that person's presence in a particular source, service, or policy. This distinction is critical because deduplication does not have a single meaning. Removing repetitions may be

appropriate for population indicators and inappropriate for utilization analyses, in which repeated occurrences constitute the phenomenon being observed. Reporting guidelines for studies using routinely collected data reinforce the need to make the population, analytical unit, and treatment of records explicit [2].

Record linkage seeks to identify when observations from different sources or time points refer to the same entity. Deterministic strategies may rely on direct identifiers, while auxiliary keys combine normalized registration attributes when those identifiers are missing or invalid [21]. The usefulness of each alternative depends on field availability, stability, and quality. Key provenance and matching confidence are therefore part of the analytical information, not merely implementation details. Records with insufficient evidence should remain unlinked because a forced match may alter denominators and introduce associations that do not exist [24].

Pseudonymization and anonymization serve different purposes along this path. Internal codes reduce the circulation of identifiers during processing, but still allow links to be reconstructed within an authorized environment. Disclosure requires an additional control layer based on aggregation, small-group suppression, reduced granularity, and assessment of residual reidentification risk. NIST guidance for government data treats deidentification as a process tied to objectives, release models, and continuous risk management [25]. This distinction treats privacy as a workflow-level requirement that extends beyond the initial handling of the input file.

C. Computational Resources and Exploration Techniques

Python and Jupyter Notebook support profiling, normalization, pseudonymization, deduplication, linkage, aggregation, and documentation. GeoPandas and Shapely handle territorial geometry and point-in-polygon joins; HTML, CSS, and JavaScript present aggregates without embedding calculation logic in the interface. Separating analytical code from presentation supports re-execution and inspection [19].

AI assistance was restricted to planning through ChatGPT (OpenAI) using the MEADS-LLM V7 prompt/analytical grammar. Because the hosted interface was not version-pinned, the exact backend build was not recorded as an experimental factor and model updates may alter proposed axes. Input was limited to institutional objectives and a sanitized data dictionary; individual values, formulas, linkage rules, and disclosure decisions remained outside the interaction. Numerical evidence was produced by local, inspectable code.

D. Execution and Audit Artifacts

The technical notebook retains loading, transformations, and calculations in re-executable form, while a styled notebook and dashboard reorganize only approved aggregates. An execution summary records warnings; the manifest identifies the aggregates consumed by the interface; and the calculation record stores formula, filter, denominator, and unit. This separation follows reproducible notebook practice [29]. The curation matrix links each visual component to its question, reference universe, evidence, and disclosure limit, adding semantic inspection to computational reproducibility [18].

E. Evaluation Criteria

The qualitative evaluation uses four criteria: *unit integrity*, preserving distinct meanings for records, persons, and links [6]; *traceability*, connecting results to source, formula, denominator, and filter [18]; *coverage accounting*, retaining failed linkages and territorial assignments; and the *privacy boundary*, examining logical and organizational controls that reduce unnecessary exposure [25]. These criteria assess artifact inspectability, not statistical validity, linkage accuracy, user effectiveness, or reidentification resistance.

F. Synthetic Linkage Sensitivity Check

A fully synthetic stress test examined the deterministic identity rule without exposing restricted records. For each noise level, 30 deterministic replicates (base seed 8972026) generated 5,000 fictitious identities with a random 11-digit identifier, name, mother's name, and birth date; 10% formed homonym pairs sharing name and birth date. Matching used exact identifier, then normalized name–mother–birth date, then name–birth date, rejecting ambiguous keys. Independent 0%, 5%, 10%, or 20% per-field perturbation either removed a value or changed one character. Recovery, precision, and unmatched proportion were recorded. The test isolates this exact-matching rule and does not represent correlated errors, temporal drift, real identifier distributions, probabilistic linkage, or municipal-data accuracy.

IV. MEADS-Health Methodology

The MEADS-Health architecture is shown in Fig. 1. The workflow separates the institutional environment, where records are held and processed, from the LLM-support environment. Communication between them is limited to the institutional objective, sanitized schema, and preliminary axis plan. Dataset rows, identifiers, and individual values do not cross this boundary. This boundary also assigns responsibilities: the model supports analytical planning, local code performs the computation, and the responsible team controls validation and release.

A. Protected Reading and Analytical Contract

Reading begins in the protected environment. In Steps 1–3 of Fig. 1, the purpose is delimited, file structures are examined, and fields are classified by sensitivity. The initial profile includes encoding, inferred data types, missingness, cardinalities, and candidate keys. Technical field names are then replaced with understandable descriptions, such as person name, identifier, parentage, date of birth, age, address, and clinical notes. The result is a generalized data dictionary that describes the structure without exposing record contents.

Before any cross-source linkage, the request is formalized in an analytical contract. The document records the purpose, authorization, reference population, period, counting unit, required sources, and intended audience. This agreement

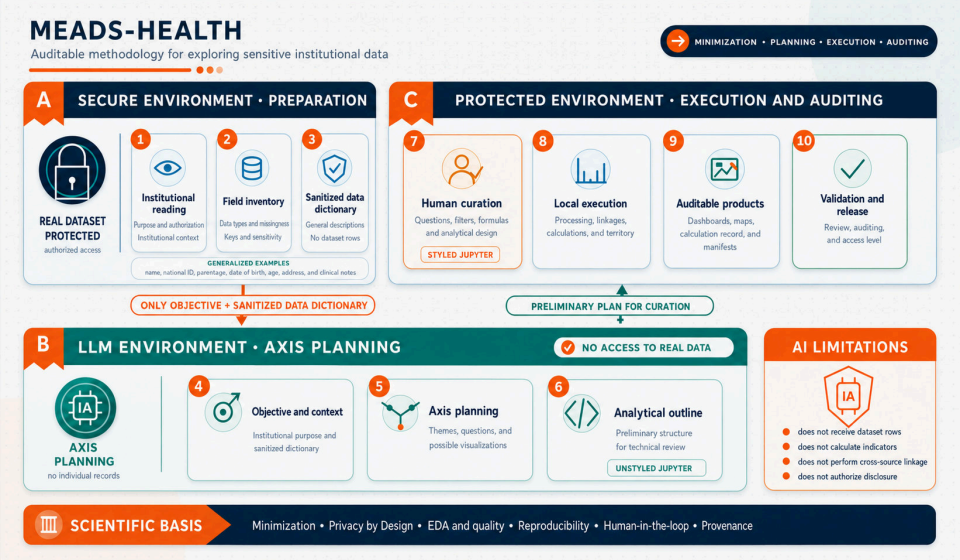


Fig. 1. MEADS-Health architecture. AI supports initial planning, while calculations, linkages, products, and release decisions remain in the protected environment.

prevents the mere technical possibility of combining files from being treated as sufficient justification for doing so. It also distinguishes population questions, which require unique persons, from healthcare-utilization questions, in which repeated occurrences may represent actual service use.

B. AI-Assisted Axis Planning

In Steps 4–6, the institutional objective and sanitized data dictionary form the LLM input. The request asks for analytical axes, guiding questions, plausible groupings, and visualization possibilities. Given the structure of a health dataset, for example, the model may suggest topics related to registration profiles, healthcare follow-up, territorial distribution, and data-entry quality. Because dataset values are not part of the input, this output contains no counts, diagnoses, or inferences about individuals.

Suggestions are organized into an analytical notebook outline. Each item states the intended question and the fields potentially required, without yet taking the form of a final indicator. Axes with no institutional value, incompatible with the schema, or excessively granular for the authorized purpose are removed during review. Only after this screening are the formula, filter, unit, denominator, and privacy rule defined. The model proposes candidate questions, and the team decides which ones are supported by the available sources and authorized purpose.

C. Curation, Execution, and Identity Control

In Step 7, analysts and domain specialists convert the outline into an executable plan. Each question is reviewed against the unit of analysis, reference universe, available fields, and disclosure level. Related axes are grouped into macro-components so that the dashboard preserves a coherent reading path rather than presenting independent charts organized only by column availability.

Step 8 returns the plan to the local environment and begins the workflow detailed in Fig. 2. Labels, dates, and characters are normalized, but original values remain in the audit trail whenever a transformation could affect matching. Invalid identifiers are explicitly flagged rather than corrected by assumption. The key strategy follows the conditions of each source, and its provenance remains associated with the resulting links.

Deduplication occurs only after the unit has been defined. Population indicators retain one internal code per individual, whereas utilization analyses preserve repeated occurrences. In cross-source analyses, the source specified in the analytical contract serves as the reference universe, while the others add information without silently changing the denominator. Coverage of source S relative to universe U is expressed as

$$C_s = \frac{N(U \cap S)}{N(U)} \times 100. \tag{1}$$

Here, C_s is an accounting flag, not a validated statistical estimator. Missing or malformed identifiers, spelling/date variation, duplicate handling, and non-overlapping extraction periods can change the numerator; strict keys may increase false nonmatches and relaxed keys false matches. Because the municipal sources have no reviewed linkage gold standard, unmatched records and key provenance are retained rather than treating C_s as population coverage. The synthetic experiment below probes only the deterministic rule under controlled perturbation and does not validate C_s on real records.

D. Territorial Enrichment

The territorial branch in Fig. 2 examines existing coordinates and addresses eligible for geocoding. Missing coordinates, reversed axes, locations outside the study boundary, and incompatibility with the area of interest produce distinct

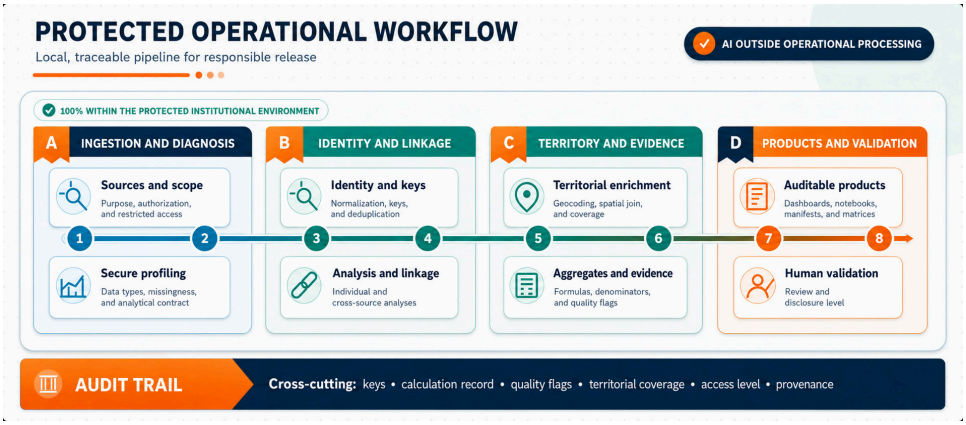


Fig. 2. Protected operational workflow. Processing, linkage, calculations, and validation are performed locally without AI participation.

flags. When a point is valid, a point-in-polygon join assigns the corresponding administrative division. Territorial coverage is calculated in the processed file itself:

$$C_g = \frac{N(\text{records with an assigned division})}{N(\text{records assessed})} \times 100. \qquad (2)$$

The ratio C_g is likewise a bookkeeping signal, not a validated spatial statistic. Missing or stale addresses, geocoding noise, coordinate errors, and polygon mismatch affect it, so it cannot establish representativeness of the mapped population.

Three situations remain separate: a valid point with an assigned area, an available point with no territorial match, and insufficient information for assignment. This distinction matters in health analysis because an apparently underrepresented territory may reflect poor registration quality rather than low demand. Point maps remain restricted to the authorized environment. Released products use area-level counts and small-group suppression whenever granularity could facilitate reidentification.

E. Products, Auditing, and Release

In Steps 9 and 10, the executed plan produces a set of complementary artifacts. The technical notebook preserves data loading, transformations, and calculations; the styled notebook reorganizes the institutional reading without replacing the computational source. The dashboard presents the macro-components, the calculation record documents formulas and denominators, the execution summary brings together warnings, and the manifest lists the aggregates consumed by the interface.

Each visual component has a corresponding entry in the curation matrix. Approval depends on agreement among the question, unit, denominator, filter, formula, and value. Human validation also determines the access level, which may be restricted, institutional aggregate, or public aggregate. Internal interfaces may query detailed files within the authorized environment, but those files are not included in the release package.

V. Results of the Methodological Application

A. Adaptation to Heterogeneous Sources

During the case study, heterogeneous sources used a common workflow without a single counting rule: person registries used individual-oriented deduplication, healthcare records retained events, and program datasets represented links. This preserved the intended unit in the artifacts, but no comparison established greater numerical accuracy than an alternative pipeline. Direct identifiers remained distinguishable from auxiliary keys, insufficient matches stayed unlinked, and key provenance followed the aggregates.

B. Analytical Axes and Products

The V7 planning grammar contained 17 reusable axis templates. For the health-source configuration, sanitized metadata activated 10 templates for the technical notebook; seven were not carried into execution because their required field roles were absent. This count describes how the planning grammar filtered the available templates and should not be interpreted as a measure of LLM quality. Human review could still remove, merge, or reorganize activated proposals before notebook development. The retained questions addressed population follow-up, healthcare records, territory, registration quality, and integration with other policies; the model defined neither final indicators nor clinical interpretations. Approved axes became dashboard macro-components linking each question to its unit, fields, filter, visualization, and interpretation warning.

C. Synthetic Linkage Sensitivity

Table I shows the bounded stress test. Recovery fell from 100.0% to 93.1% as field noise reached 20%; precision remained above 99.7%, while unmatched records rose to 6.7%. Composite fallbacks recovered some corrupted direct identifiers, but the nonzero false-link rate confirms that the rule is not noise-invariant.

Across the 30 replicates at each nonzero noise level, recovery SDs were 0.09, 0.21, and 0.33 percentage points at 5%, 10%, and 20% noise. These are simulation outputs, not institutional linkage estimates.

TABLE I
SYNTHETIC SENSITIVITY OF THE DETERMINISTIC LINKAGE RULE

Field noise	Recovery	Precision	Unmatched
0%	100.00%	100.00%	0.00%
5%	99.50%	99.98%	0.48%
10%	98.11%	99.92%	1.81%
20%	93.10%	99.78%	6.69%

TABLE II
QUALITATIVE EVALUATION: CRITERION AND OBSERVED EVIDENCE

Criterion	Observed evidence	Interpretation limit
Unit integrity	Separate record-, person-, and link-oriented paths; deduplication after unit definition.	Does not establish numerical correctness.
Traceability	Notebook, calculation record, manifest, curation matrix, and key/coordinate provenance.	Supports reconstruction, not external replication of restricted data.
Coverage accounting	Unmatched links and territorial failures retained; C_s and C_g recorded descriptively.	Not a validated completeness or coverage estimate.
Privacy boundary	Record values excluded from LLM input; release artifacts aggregated and curator-reviewed.	Reduces exposure pathways but does not eliminate residual risk.

D. Territory and Preservation of Failures

Territorial enrichment assigned administrative divisions to valid points while retaining unresolved cases in internal coverage reports. This distinguished missing geographic information from absence of an occurrence and prevented registration gaps from being attributed to the territory. Coordinate origin was also preserved, separating values present in the source, recovered from another authorized dataset, or rejected during validation. This provenance accompanies the maps and limits the claims supported by the spatial distribution.

E. Dashboard Auditability

The dashboard was accompanied by the calculation record and curation matrix. Table II maps each qualitative criterion to observed evidence and its interpretation limit.

Audit review exposed discrepancies in units, denominators, and legends that were not apparent from the interface alone; they returned to the plan before release. No error-reduction rate, review time, or final analytical accuracy was measured.

VI. DISCUSSION

MEADS-Health is an analytical-engineering methodology rather than a predictive model, and the evidence from the case study is operational. Analytical contracts framed heterogeneous sources; records, persons, and links retained distinct meanings; matches preserved provenance; territorial failures remained visible; and dashboard components were connected to audit documents. The paper therefore presents no epidemiological estimates or sector-performance comparisons. This delimitation is particularly important for health sources, where medical reports, clinical codes, medications, procedures, and care notes may encourage inferences beyond what administrative data support. Their use depends on a previously defined question, unit, and purpose.

The synthetic test adds a bounded performance observation: exact fallback keys preserved high recovery under independent perturbations, but increasing noise produced both unmatched records and false links. Generated identities and simplified errors mean that this result cannot resolve real-world linkage uncertainty.

The LLM boundary translates governance principles into an architectural decision. International guidance recommends minimization, oversight, and risk management [12], [14], [30]; here, only the sanitized data dictionary reaches the model, while records remain in local execution and curators review proposals, calculations, and aggregate evidence. AI output supports early planning but does not constitute evidence, a final formula, or authorization for use.

The results also distinguish reproducibility from auditability. Re-executing a notebook can reproduce a value, but cannot show by itself that the denominator answers the question or that the interpretation suits the audience. The curation matrix adds this semantic layer by connecting the result, unit, formula, context, and disclosure condition. This brings the dashboard closer to institutional decision-making routines [9]–[11].

A. Implications for Health and Public Management

A missing match between a registration source and a healthcare record may indicate that the event did not occur, the key failed, the periods differ, or the extraction covers only part of the system. MEADS-Health preserves this ambiguity as provenance and coverage information rather than resolving it through automatic inference. The same caution applies to territory: a spatial concentration may reflect demand, service supply, referral patterns, or more complete addresses. Reporting coverage and geographic origin reduces the risk of treating graphical precision as evidence of information completeness.

The components are candidates for transfer to comparable public-sector settings, but transferability has not been empirically tested. Keeping record values outside the AI layer reduces one potential exposure pathway, but it does not guarantee privacy. Detailed metadata, small aggregates, access-control failures, and curator errors can still create residual risk.

B. Limitations

This single-context study used a qualitative, artifact-level evaluation. No direct pipeline comparison or endpoint for analytical accuracy, efficiency, user performance, or decision quality was defined; restricted data also prevent external replication. Auxiliary-key matches lacked a reviewed gold standard, so linkage accuracy is unknown.

The synthetic sensitivity check covers only independent missingness or single-character perturbation in four identity fields. It does not reproduce correlated registration errors, legal name changes, coding drift, duplicate creation, stale addresses, geocoding error, or source-specific missingness. Real-world linkage accuracy remains unknown without a reviewed gold standard, and territorial robustness remains untested. With unstable keys or incomplete source capture, linkage and coverage

flags may still be too uncertain for cross-source interpretation. The hosted LLM was not version-pinned, so repeated planning may differ.

Privacy controls are logical and organizational barriers, not guarantees. Reversible pseudonyms, quasi-identifiers, small groups, detailed metadata, access misconfiguration, and human release decisions leave residual risk; no reidentification attack or quantitative privacy-risk assessment was performed. Future work should add direct pipeline comparisons, reviewed real-data linkage references, correlated and source-specific noise models, territorial sensitivity tests, curator agreement, privacy-risk analysis, and prospective measurement of rework and validation time.

VII. Conclusion

MEADS-Health is a preliminary methodological artifact for organizing sensitive institutional-data exploration around explicit analytical units, provenance, local computation, and human-controlled AI assistance. The municipal case study produced inspectable notebooks and audit artifacts, while a separate synthetic stress test supplied bounded evidence on the deterministic linkage rule. Under independent field perturbation, match recovery declined predictably while precision remained high; this result does not substitute for a reviewed real-data gold standard. No alternative pipeline, final analytical accuracy, operational gain, or population-level inference was tested.

The available evidence supports the architecture, audit trail, and limited robustness of the deterministic matching component. End-to-end effectiveness and generalization were not evaluated. The method may become inconclusive with incomplete source capture, correlated or systematic key errors, stale territorial data, or disclosure granularities that raise residual reidentification risk. Reviewed linkage references, direct comparisons, richer noise models, territorial sensitivity tests, privacy-risk analysis, and prospective measurement of curator agreement and rework time are needed before practical benefits can be quantified.

Acknowledgments

The authors thank the teams responsible for the governance, maintenance, and interpretation of the institutional datasets used in developing the methodology. This work preserves record confidentiality and presents only methodological aspects and audit evidence.

References

[1] R. Connelly, C. J. Playford, V. Gayle, and C. Dibben, "The role of administrative data in the big data revolution in social science research," *Social Science Research*, vol. 59, pp. 1–12, 2016.

[2] E. I. Benchimol, L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, H. T. Sørensen, E. von Elm, S. M. Langan, and RECORD Working Committee, "The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement," *PLOS Medicine*, vol. 12, no. 10, p. e1001885, 2015.

[3] Brasil, "Lei n. 13.709, de 14 de agosto de 2018: Lei geral de proteção de dados pessoais." Diário Oficial da União, 2018. Texto compilado; acesso em: 26 jul. 2026.

[4] World Health Organization, "Global report on health equity for persons with disabilities," tech. rep., World Health Organization, Geneva, 2022.

[5] Brasil, "Lei n. 13.146, de 6 de julho de 2015: Institui a lei brasileira de inclusão da pessoa com deficiência." Diário Oficial da União, 2015. Acesso em: 26 jul. 2026.

[6] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[7] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[8] M. G. Kahn, T. J. Callahan, J. Barnard, *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *eGEMs*, vol. 4, no. 1, p. 1244, 2016.

[9] D. Helminski, J. B. Sussman, P. N. Pfeiffer, A. N. Kokaly, A. Ranusch, A. D. Renji, L. J. Damschroder, Z. Landis-Lewis, and J. E. Kurlander, "Development, implementation, and evaluation methods for dashboards in health care: Scoping review," *JMIR Medical Informatics*, vol. 12, p. e59828, 2024.

[10] G. Stahlman, I. Yanovitzky, and M. Kim, "Design, application, and actionability of US public health data dashboards: Scoping review," *Journal of Medical Internet Research*, vol. 27, p. e65283, 2025.

[11] N. Fareed *et al.*, "Lessons learned from developing dashboards to support decision-making for community opioid response by community stakeholders: Mixed methods and multisite study," *JMIR Human Factors*, vol. 11, p. e51525, 2024.

[12] World Health Organization, "Ethics and governance of artificial intelligence for health: WHO guidance," tech. rep., World Health Organization, Geneva, 2021.

[13] World Health Organization, "Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models," tech. rep., World Health Organization, Geneva, Mar. 2025. WHO publication page dated 25 March 2025; document copyright 2024.

[14] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," Tech. Rep. NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, 2023.

[15] Y. Chen *et al.*, "Generative AI in medical practice: In-depth exploration of privacy and security challenges," *Journal of Medical Internet Research*, vol. 26, p. e53008, 2024.

[16] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019.

[17] J. Declerck, Ö. D. Kılıç, E. E. Erol, S. Mehryar, D. Kalra, I. de Zegher, and R. Celebi, "Assessing data quality in heterogeneous health care integration: Simulation study of the AIDAVA framework," *JMIR Medical Informatics*, vol. 13, p. e75275, 2025.

[18] M. Johns, T. Meurers, F. N. Wirth, A. C. Haber, A. Müller, M. Halilovic, F. Balzer, and F. Prasser, "Data provenance in biomedical research: Scoping review," *Journal of Medical Internet Research*, vol. 25, p. e42289, 2023.

[19] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," *PLoS Computational Biology*, vol. 9, no. 10, p. e1003285, 2013.

[20] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[21] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer, 2012.

[22] A. Pathak *et al.*, "Privacy preserving record linkage for public health action: Opportunities and challenges," *Journal of the American Medical Informatics Association*, vol. 31, no. 11, pp. 2605–2612, 2024.

[23] K. Tyagi and S. J. Willis, "Accuracy of privacy preserving record linkage for real world data in the united states: A systemic review," *JAMIA Open*, vol. 8, no. 1, p. ooaf002, 2025.

[24] K. L. Harron, J. C. Doidge, H. E. Knight, R. E. Gilbert, H. Goldstein, D. A. Cromwell, and J. H. van der Meulen, "A guide to evaluating linkage quality for the analysis of linked data," *International Journal of Epidemiology*, vol. 46, no. 5, pp. 1699–1710, 2017.

[25] S. Garfinkel, J. Near, A. N. Dajani, P. Singer, and B. Guttman, "De-identifying government datasets: Techniques and governance," NIST Special Publication 800-188, National Institute of Standards and Tech-

nology, Gaithersburg, MD, Sept. 2023. Final publication, 14 September 2023.

[26] J. Tyris, G. Dwyer, K. Parikh, A. Gourishankar, and S. Patel, "Geocoding and geospatial analysis: Transforming addresses to understand communities and health," *Hospital Pediatrics*, vol. 14, no. 6, pp. e292–e297, 2024.

[27] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human–ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

[28] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

[29] A. Rule, A. Birmingham, C. Zuniga, I. Altintas, S.-C. Huang, R. Knight, N. Moshiri, M. H. Nguyen, S. B. Rosenthal, F. Pérez, and P. W. Rose, "Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks," *PLOS Computational Biology*, vol. 15, no. 7, p. e1007007, 2019.

[30] B. Anderson and E. Sutherland, "Collective action for responsible AI in health," Tech. Rep. 10, OECD Publishing, Paris, 2024.

Modern Hopfield Networks for Associative Memory-Based Emergency Triage Prediction*

Gustavo Augusto Pires*, Frederico Gualberto Ferreira Coelho†, Henrique Resende Martins‡
* *Programa de Pós-Graduação em Engenharia Elétrica*, *Universidade Federal de Minas Gerais*
Belo Horizonte, Brazil
gustavo.augusto.pires@hotmail.com
† *Departamento de Engenharia Eletrônica*, *Universidade Federal de Minas Gerais*
Belo Horizonte, Brazil
fredgfc@gmail.com
‡ *Departamento de Engenharia Elétrica*, *Universidade Federal de Minas Gerais*
Belo Horizonte, Brazil
henriquerm@gmail.com

***Abstract*—Emergency department triage requires the integration of heterogeneous information under time and resource constraints, while class imbalance may limit the recognition of less frequent urgency categories. This study investigates modern Hopfield networks with patient-level associative memory for predicting the five-level urgency assignment recorded in the 2022 National Hospital Ambulatory Medical Care Survey Emergency Department dataset. The final unweighted cohort comprised 10,207 emergency department visits, divided into training, validation, and test sets using a fixed stratified partition. Two architectures were evaluated: a shared-memory model, in which selected training visits were stored in a single associative memory, and a class-aware model, in which the same patterns were organized into five urgency-specific memory banks. The proposed models were compared with ZeroR, proportional-odds ordinal logistic regression, k-nearest neighbors, and XGBoost. XGBoost achieved the highest test accuracy (0.5836) and the lowest ordinal mean absolute error (0.4543). In contrast, the class-aware Hopfield model obtained the highest macro-F1 (0.3861) and balanced accuracy (0.4014), while the shared-memory model achieved the highest quadratic weighted kappa (0.4514). The Hopfield architectures substantially improved recall for under-represented urgency classes, including the most urgent category, for which XGBoost obtained zero recall and the class-aware model achieved 0.19. These results indicate that class-structured associative retrieval may reduce majority-class concentration and improve coverage across urgency levels. However, this benefit was accompanied by lower overall accuracy and higher ordinal error than XGBoost, and did not reduce severe undertriage relative to the best conventional model. Modern Hopfield networks therefore provide a promising approach for class-balanced triage prediction, but require further investigation across multiple seeds, datasets, and externally separated evaluation settings.**

***Index Terms*—Modern Hopfield networks, associative memory, emergency department triage, urgency classification, machine learning.**

I. Introduction

Emergency department triage is a time-critical process in which patients are rapidly assessed and assigned an urgency level to guide prioritization, resource allocation, and timely access to care. Inaccurate triage may affect both individual patients and emergency department operations. Undertriage can delay the evaluation and treatment of patients with urgent conditions, whereas overtriage can divert limited resources from patients with greater needs and contribute to congestion in high-acuity pathways. Triage decisions are therefore not merely administrative classifications, but a central component of emergency care that influences waiting times, patient flow, and the delivery of appropriate treatment [1], [2]. In an observational study comparing nurses' initial Emergency Severity Index assignments with classifications derived from the ESI handbook, the reported overall accuracy was 67%, while undertriage and overtriage occurred in 25% and 8% of cases, respectively, illustrating that classification discrepancies remain frequent in practice [2].

Accurate triage remains challenging because urgency assessment depends on the interaction of multiple sources of information and is performed under demanding operational conditions. Presenting complaints, vital signs, visual assessment, patient history, professional experience, training, workload, interruptions, and emergency department crowding may all influence the assigned urgency level [1]. These difficulties may be particularly relevant for older adults, whose serious conditions can manifest through nonspecific complaints, atypical symptoms, or physiological measurements that do not clearly indicate deterioration. Such presentations increase the risk that an urgent patient will be assigned to a lower-priority category, potentially prolonging waiting time and delaying necessary evaluation and treatment [3]. These limitations motivate the development of decision-support methods capable of integrating heterogeneous triage information more consistently.

Machine learning provides a promising opportunity to support this process. By learning complex and potentially nonlinear relationships among routinely collected variables, including demographic characteristics, mode of arrival, chief complaints, and vital signs, machine-learning models may identify patterns that are difficult to represent using fixed rules

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), and supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

or conventional scoring systems. Using 135,470 adult visits from the National Hospital Ambulatory Medical Care Survey (NHAMCS), Raita et al. found that several machine-learning approaches outperformed an ESI-based reference model in predicting critical care and hospitalization outcomes. For critical care prediction, the best-performing model achieved an area under the receiver operating characteristic curve of 0.86, compared with 0.74 for the conventional model, and identified a larger proportion of critically ill patients assigned to lower-acuity ESI categories [4]. More broadly, a systematic review of 26 studies reported that artificial intelligence- and machine-learning-based triage models frequently achieved AUC values above 0.80 and often outperformed conventional triage tools, particularly when structured clinical variables were combined with textual information [5]. Nevertheless, much of the available evidence is retrospective, single-center, and characterized by limited external validation. These technologies should therefore be regarded as potential decision-support tools rather than replacements for clinical assessment [5].

Within this context, Modern Hopfield Networks have renewed interest in associative-memory models by addressing several limitations of their classical formulations. Classical Hopfield networks operate primarily with binary states and have limited storage capacity. Modern formulations introduce alternative energy functions and continuous-valued states, with theoretical results showing storage capacity that can grow exponentially with the dimension of the associative space under suitable conditions, rapid pattern retrieval, and differentiable integration into deep learning architectures [6]. Their update rule has also been shown to correspond to the key–value attention mechanism used in Transformer architectures, providing an energy-based interpretation of attention as the retrieval of either an individual stored pattern or a weighted combination of related patterns [6], [7]. Modern Hopfield Networks can consequently serve as trainable layers for memory, association, attention, prototype learning, and pooling.

The practical relevance of these properties has been demonstrated across heterogeneous learning problems. In computational biology, Hopfield-based attention has been used to retrieve informative patterns from immune repertoires containing hundreds of thousands of instances, supporting both classification and the identification of disease-associated sequence motifs [7]. Modern Hopfield Networks have also been applied to reaction-template retrieval, where learned associations between molecular and template representations improved prediction for rare and previously unseen templates [8], and to graph representation learning, where associative memories were used to encode structural prototypes for node classification and link prediction [9]. For tabular data, Hopular demonstrated that the training set can be stored as an external memory and accessed across successive layers to model sample–sample, feature–feature, and feature–target dependencies, achieving competitive performance against established gradient-boosting and deep learning methods [10]. Collectively, these studies motivate the use of Modern Hopfield Networks when predictions may benefit from content-addressable access to previously observed examples or learned prototypes.

This capability is conceptually relevant to emergency triage. Patients with similar demographic characteristics, presenting complaints, physiological measurements, and arrival conditions may exhibit related patterns of recorded urgency. A Modern Hopfield Network can encode a current patient visit, compare that representation with encoded visits from the training set, and retrieve a weighted combination of related historical cases before producing a prediction. Unlike conventional nearest-neighbor methods, for which similarity is determined by a predefined distance in the input or preprocessed feature space, the Hopfield-based approach can jointly learn the patient representation and the association mechanism used for retrieval. However, the studies reviewed above do not establish how patient-level associative-memory architectures compare with conventional ordinal, tree-based, and nearest-neighbor models for five-level classification of recorded emergency-triage urgency. In particular, the balance between overall predictive performance, recognition of underrepresented urgency classes, and errors toward lower recorded urgency remains to be characterized in this setting.

To address this gap, this study evaluates Modern Hopfield Networks with patient-level associative memory for predicting five recorded urgency levels using the 2022 NHAMCS Emergency Department dataset [11]. The memory is constructed exclusively from encoded training visits, while validation and test visits act only as queries and are never incorporated into the stored memory. For each current visit, the model retrieves representations of training cases and combines them with the visit's own representation before producing class probabilities. We evaluate both a shared-memory architecture, in which the selected training visits constitute a single associative memory, and a class-aware variant, in which stored patterns are organized according to their recorded urgency classes.

The proposed architectures are compared with complementary baselines: a majority-class ZeroR model [12], proportional-odds ordinal logistic regression [13], multiclass k-nearest neighbors [14], and multiclass XGBoost [15]. These baselines represent class-frequency-based prediction, linear ordinal modeling, conventional instance-based learning, and nonlinear tree-based classification. Performance is evaluated on a common fixed data partition using accuracy, macro-F1, balanced accuracy, ordinal mean absolute error, and quadratic weighted kappa. Because the class labels are ordered from immediate to nonurgent care, we additionally quantify undertriage and severe undertriage relative to the recorded urgency assignment. These measures describe disagreement with the recorded labels rather than independently adjudicated clinical error. A complementary analysis of the shared-memory model compares full, disabled, and uniform retrieval at inference time, with trained weights held fixed, to assess how these interventions affect predictions and performance. Accordingly, the study characterizes the predictive and ordinal trade-offs of the evaluated architectures; differences in class weighting and model capacity preclude attributing performance differences

exclusively to associative retrieval.

II. Methodology

A. Dataset and Preprocessing

This retrospective study used the 2022 Emergency Department Public Use File of the National Hospital Ambulatory Medical Care Survey (NHAMCS) [11]. The observational unit was an emergency department visit. Of the 16,025 original records, 10,207 visits with `IMMEDR` values from 1 to 5 were retained. Records coded as -9, -8, 0, or 7 were excluded because they did not contain an ordered urgency label compatible with the supervised five-class task. The target preserved the original ordering: 1 (Immediate), 2 (Emergent), 3 (Urgent), 4 (Semi-urgent), and 5 (Nonurgent), with lower values indicating higher recorded urgency. The final class counts were 139, 1,595, 5,340, 2,826, and 307, respectively, as shown in Fig. 1. The outcome was interpreted as the urgency recorded in NHAMCS rather than an independently adjudicated clinical ground truth. Survey weights and complex-sampling variables were not used; therefore, the reported results describe the retained unweighted sample and are not national estimates.

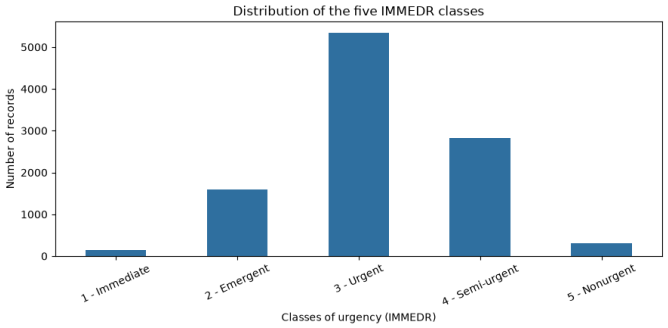


Fig. 1. Distribution of recorded urgency classes in the retained NHAMCS sample.

Twenty-six predictors designated as available at the initial triage assessment were considered. Numerical variables comprised age, temperature, pulse rate, respiratory rate, systolic and diastolic blood pressure, oxygen saturation, and pain score. Two binary indicators represented special pulse and blood-pressure measurement codes. Categorical variables included visit month and weekday, residence, sex, ethnicity, race, emergency-medical-services arrival, ambulance transfer, a previous visit within 72 hours, five coded reasons for visit, and visit episode. Arrival time was treated as a temporal variable. Identifiers, hospital codes, survey-design variables, the target, diagnoses, injury and external-cause fields, dispositions, admissions, transfer outcomes, deaths, and other post-triage information were excluded to reduce temporal and target leakage.

All preprocessing parameters were estimated exclusively from the training set and then frozen. For ordinal logistic regression, k-nearest neighbors, and the Hopfield models, missing numerical values were imputed with training medians, missingness indicators were added, and continuous variables were standardized using training means and population standard deviations. For XGBoost, numerical missing values were retained and no standardization was applied [15]. Arrival time was converted from HHMM to minutes after midnight and represented by $\sin(2\pi t/1440)$ and $\cos(2\pi t/1440)$; missing values were additionally marked by an indicator. Categorical levels occurring fewer than 100 times in the ordinal model or fewer than 10 times in the remaining feature-based models were grouped as infrequent, while blank, unknown, and not-applicable categories were preserved. After numerical and temporal processing and one-hot encoding of categorical variables, the final feature matrices contained 91 variables for ordinal logistic regression, 491 for XGBoost, and 499 for k-nearest neighbors and the Hopfield architectures.

B. Baseline Models

Four baselines represented complementary assumptions. ZeroR ignored the predictors and returned the empirical training-class distribution, with the majority class used for hard predictions [12]. Proportional-odds ordinal logistic regression modeled the cumulative probabilities of the five ordered classes using a shared coefficient vector and four thresholds [13]; the model was fitted without regularization by L-BFGS, with a maximum of 1,000 iterations [16].

Multiclass k-nearest neighbors provided a conventional instance-based comparison in the fixed preprocessed feature space [14]. A validation grid evaluated $k \in \{1, 3, 5, 7, 11, 15, 21, 31\}$, uniform or inverse-distance voting, and Manhattan or Euclidean distance. Selection minimized validation ordinal mean absolute error, with macro-F1 and log loss used as tie-breakers; the selected model used 31 neighbors, Manhattan distance, and uniform voting.

XGBoost represented a nonlinear tree-based baseline [15]. It used a five-class probabilistic objective, histogram-based trees, learning rate 0.05, maximum depth 6, minimum child weight 1, row and feature subsampling of 0.8, L2 regularization of 1, and at most 2,000 boosting rounds. Early stopping monitored validation multiclass log loss with patience of 100 rounds, selecting 351 rounds.

C. Proposed Modern Hopfield Models

Both proposed models used the same 499-dimensional input and encoded each visit into a 64-dimensional latent representation $\mathbf{z}$. The encoder followed the sequence $499 \to 128 \to 64$, with layer normalization and GELU activations after both linear layers and dropout of 0.20 after the first. Independent learned projections produced 64-dimensional queries, keys, and values. Given a projected query $\mathbf{q}$ and stored training representations $\{(\mathbf{k}_i, \mathbf{v}_i)\}_{i=1}^{N}$, where N denotes the number of stored patterns, one modern Hopfield retrieval update was computed as [6]

$$a_i = \frac{\exp\left(\beta \mathbf{q}^\top \mathbf{k}_i\right)}{\sum_{j=1}^{N} \exp\left(\beta \mathbf{q}^\top \mathbf{k}_j\right)}, \qquad \mathbf{r} = \sum_{i=1}^{N} a_i \mathbf{v}_i, \tag{1}$$

with fixed scaling $\beta = 1/\sqrt{64} = 0.125$. A single associative head and one update were used.

The memory was constructed exclusively from training visits. At most 256 visits were sampled without replacement from each recorded urgency class, while classes with fewer observations contributed all available visits. This produced 1,080 stored patterns with class counts $(97, 256, 256, 256, 215)$. The same selected visits were used in both architectures. Thus, both memories used the same class-capped sampling scheme rather than reproducing the original training-class distribution. Memory embeddings were recomputed with the current encoder during training and detached from the computational graph. Validation and test visits were never added to memory. When a training query was itself stored, its corresponding pattern was masked before the association softmax to prevent direct self-retrieval.

In the shared-memory model, all 1,080 patterns formed a single bank. The encoded current-visit representation $\mathbf{z}$ and retrieved representation $\mathbf{r}$ were concatenated and passed through a $128 \to 64 \to 5$ classification head with GELU and dropout of 0.20. The model contained 93,701 trainable parameters.

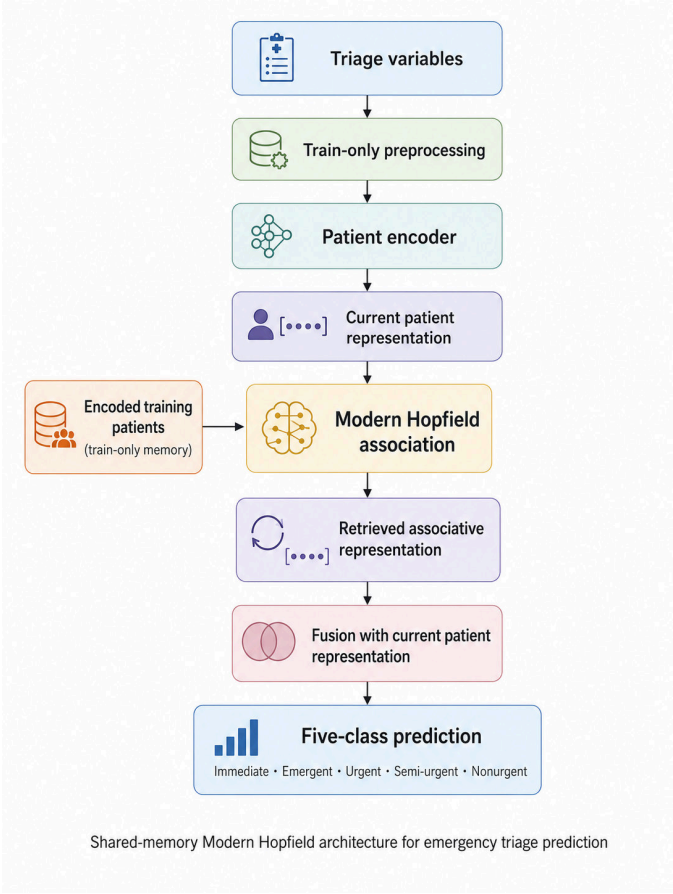


Fig. 2. Shared-memory modern Hopfield architecture. The encoded emergency department visit queries a single memory containing training visits. The retrieved representation is concatenated with the current-visit representation before five-class prediction.

In the class-aware model, the same stored patterns were separated into five banks according to their recorded urgency classes. The query independently retrieved one representation from each bank:

$$\mathbf{r}_c = \sum_{i=1}^{M_c} a_{c,i}\mathbf{v}_{c,i}, \qquad c \in \{1, \ldots, 5\}, \tag{2}$$

where M_c denotes the number of stored patterns in class c, and association weights were computed as in Eq. (1), with normalization restricted to the corresponding bank. A size-adjusted evidence score was calculated as $s_c = \log \sum_{i=1}^{M_c} \exp(\beta \mathbf{q}^\top \mathbf{k}_{c,i}) - \log M_c$, and the five scores were normalized by softmax across banks. The classifier received the current-visit representation, the five retrieved representations, and the five normalized bank scores, resulting in a $64+5\times64+5 = 389$-dimensional input followed by a $389 \to 64 \to 5$ head with GELU and dropout of 0.20. The class-aware model contained 110,405 trainable parameters. Thus, the variant differed from the shared-memory model in memory organization, the number of retrieved representations, the use of bank-level evidence, and classifier size; their comparison did not isolate memory organization alone.

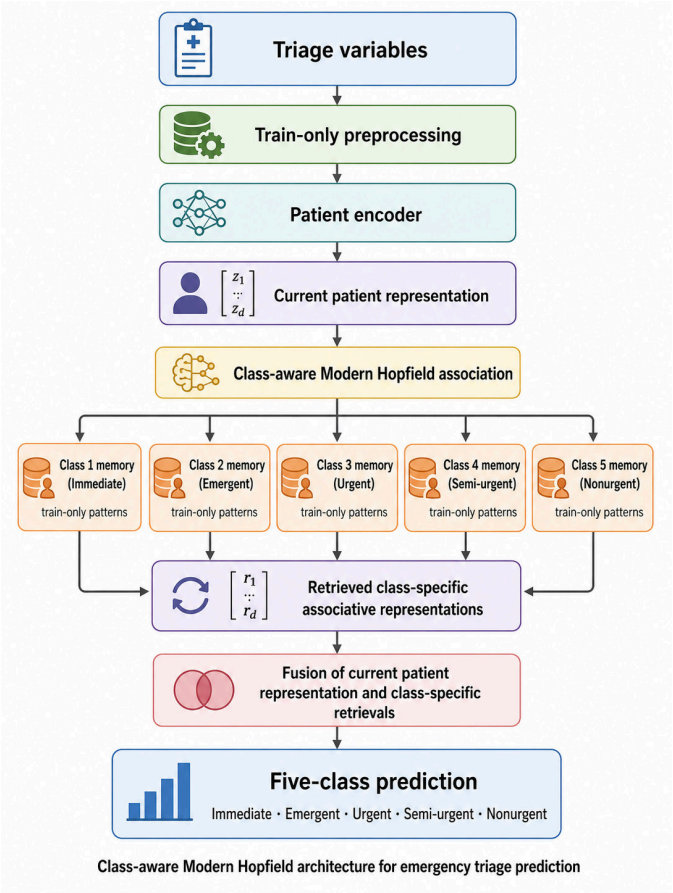


Fig. 3. Class-aware modern Hopfield architecture. Training visits are separated into five memory banks according to recorded urgency. The query retrieves one representation from each bank, and the resulting representations and bank-level evidence are combined with the current-visit representation before prediction.

D. Training and Experimental Protocol

The 10,207 visits were divided once into training, validation, and test sets using a two-stage stratified random split with seed 2022. The partitions contained 7,144, 1,531, and 1,532 visits, respectively, and the same assignments were used for every model. Splitting was performed at the visit level rather than by hospital or individual. Preprocessing and model fitting used only the training set; validation data were used for hyperparameter or checkpoint selection, and models were not refitted on training plus validation before final testing.

The Hopfield models were trained with class-weighted cross-entropy, where $w_c = N_{\text{train}}/(5n_c)$, with n_c denoting the number of training visits in class c. This yielded weights approximately equal to $[14.730, 1.280, 0.382, 0.722, 6.646]$. These weights addressed class imbalance but did not explicitly encode the ordinal distance or direction of prediction errors. Optimization used AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and a maximum of 50 epochs. Validation log loss was monitored after each epoch; early stopping used a minimum improvement of 10^{-4} and patience of seven epochs. Both variants used seed 2026 and one formal training run.

Final predictions were obtained by selecting the class with the highest predicted probability, with probability columns consistently ordered as classes 1 through 5. Performance was assessed on the same 1,532 test visits using accuracy, macro-F1, balanced accuracy [17], ordinal mean absolute error [18], and quadratic weighted kappa [19]. Macro-F1 included all five classes, with undefined class-specific F1 values assigned zero. Quadratic weighted kappa used quadratic disagreement weights.

Undertriage relative to the recorded assignment was defined as $\hat{y} > y$, and severe undertriage as $\hat{y} - y \geq 2$, where y and $\hat{y}$ denote the recorded and predicted urgency levels, respectively. Both rates used all test visits as the denominator. These measures quantify disagreement with the recorded NHAMCS urgency and not independently adjudicated clinical error.

Comparisons were based on descriptive point estimates; no confidence intervals, paired hypothesis tests, or multiseed inferential analysis were performed. The baseline comparison evaluated the complete model configurations described above. Class weighting, memory sampling, and architectural capacity were not independently controlled, and no memory-free neural model was trained for a matched comparison. Consequently, differences between Hopfield models and conventional baselines cannot be attributed exclusively to associative retrieval.

E. Inference-Time Memory Analysis

A complementary analysis examined the dependence of the trained shared-memory model on its retrieved representation. Three conditions were evaluated: Full Memory, using the learned association weights in Eq. (1); Memory Disabled, replacing the retrieved representation with $\mathbf{r} = \mathbf{0}$; and Uniform Memory, replacing the association weights with $a_i = 1/N$, so that $\mathbf{r}$ became the mean of the stored value representations.

All conditions used the same validation-selected checkpoint, the same preprocessed test visits in the same order, and identical trained parameters. The current-visit representation and classification head were preserved. In Uniform Memory, the stored value representations and learned value projection were unchanged. Evaluation was performed with dropout disabled and without gradient computation or parameter updates. No retraining was performed.

For each condition, accuracy, macro-F1, ordinal mean absolute error, quadratic weighted kappa, undertriage, and severe undertriage were calculated. Relative to Full Memory, we additionally measured the percentage of visits whose predicted class changed and the mean Jensen–Shannon distance between the five-class probability distributions. Jensen–Shannon distance was computed per visit as the square root of the Jensen–Shannon divergence using natural logarithms, and then averaged across the test set.

III. Results

A. Overall Predictive Performance

Table I presents the performance of all models on the same 1,532 test visits. The results reveal a trade-off between overall predictive performance and recognition of less frequent urgency classes. XGBoost achieved the highest accuracy, at 0.5836, and the lowest ordinal mean absolute error, at 0.4543. It also outperformed the remaining conventional baselines across accuracy, macro-F1, balanced accuracy, ordinal MAE, and quadratic weighted kappa. All comparisons refer to descriptive point estimates from the evaluated runs.

The Hopfield architectures obtained the highest macro-F1 and balanced accuracy values. The shared-memory model achieved a macro-F1 of 0.3681 and a balanced accuracy of 0.3920, while the class-aware variant obtained 0.3861 and 0.4014, respectively. Compared with XGBoost, the class-aware model's macro-F1 was higher by 0.0732 and its balanced accuracy by 0.0964. These differences were accompanied by lower overall accuracy and higher ordinal MAE, indicating that broader recognition across classes did not translate into better performance on every metric.

Both Hopfield models also achieved higher quadratic weighted kappa than the conventional baselines. The shared-memory model obtained the highest QWK, at 0.4514, followed by the class-aware model at 0.4457. Thus, the Hopfield models exhibited greater chance-corrected ordinal agreement as measured by QWK, despite their higher MAE. The ordering of the models consequently depended on the ordinal metric considered.

The class-aware model obtained more favorable values than the shared-memory model on most reported metrics. Accuracy increased from 0.5222 to 0.5385, macro-F1 from 0.3681 to 0.3861, and balanced accuracy from 0.3920 to 0.4014. Ordinal MAE decreased from 0.5614 to 0.5366, while undertriage and severe undertriage decreased from 0.2402 to 0.2219 and from 0.0411 to 0.0307, respectively. However, QWK decreased from 0.4514 to 0.4457. These differences characterize the two

TABLE I

PREDICTIVE PERFORMANCE ON THE TEST SET. THE BEST VALUE AMONG THE LISTED MODELS FOR EACH METRIC IS SHOWN IN BOLD. UPWARD AND DOWNWARD ARROWS INDICATE WHETHER HIGHER OR LOWER VALUES ARE PREFERABLE, RESPECTIVELY. UNDERTRIAGE IS DEFINED RELATIVE TO THE RECORDED NHAMCS URGENCY.

Model	Accuracy ↑	Macro-F1 ↑	Balanced Acc. ↑	Ordinal MAE ↓	QWK ↑	Undertriage ↓	Severe Undertriage ↓
ZeroR	0.5228	0.1373	0.2000	0.5209	0.0000	**0.1704**	**0.0137**
Ordinal Logistic	0.5503	0.2524	0.2568	0.4961	0.2324	0.1991	0.0209
k-NN	0.5352	0.2130	0.2358	0.5157	0.1709	0.2422	0.0333
XGBoost	**0.5836**	0.3129	0.3050	**0.4543**	0.3844	0.2115	0.0183
Hopfield Memory	0.5222	0.3681	0.3920	0.5614	**0.4514**	0.2402	0.0411
Hopfield Class-Aware	0.5385	**0.3861**	**0.4014**	0.5366	0.4457	0.2219	0.0307

complete architectures and do not isolate the effect of memory organization from the additional retrieved representations, bank-level evidence, and larger classification head.

The lowest undertriage and severe-undertriage rates were obtained by ZeroR. This result should be interpreted in conjunction with its prediction behavior: because ZeroR always assigned the majority class, its low undertriage rates did not reflect effective discrimination among urgency levels. Among the feature-based models in Table I, XGBoost achieved the lowest severe-undertriage rate, at 0.0183. The corresponding rate for the class-aware model was 0.0307, an absolute difference of 1.24 percentage points. Neither Hopfield architecture reduced severe undertriage relative to XGBoost, and their higher class-balanced metrics should not be interpreted as evidence of greater clinical safety.

B. Class-Specific Error Patterns

Figure 4 compares the row-normalized confusion matrices of XGBoost, the shared-memory Hopfield model, and the class-aware Hopfield model. XGBoost produced predictions concentrated around class 3, the majority category. It correctly classified approximately 80% of recorded class-3 visits, but did not correctly identify any class-1 visits and correctly identified only 2% of class-5 visits. Among recorded class-1 visits, approximately 57% were assigned to class 3, while 70% of class-2 visits were also assigned to class 3.

The shared-memory Hopfield model distributed its predictions more broadly across the five classes. Its recalls for classes 1, 2, 4, and 5 were 0.14, 0.47, 0.56, and 0.24, respectively, compared with 0.00, 0.25, 0.46, and 0.02 for XGBoost. These recall patterns are consistent with the higher balanced accuracy and broader class coverage observed for the shared-memory model. Conversely, recall for the majority class decreased from 0.80 with XGBoost to 0.54 with the shared-memory model, contributing to its lower overall accuracy.

The class-aware model achieved recalls of 0.19, 0.50, 0.55, and 0.59 for recorded classes 1, 2, 3, and 4, respectively. Based on the rounded values, the largest recall increase relative to the shared-memory model occurred for class 1, followed by equal increases for classes 2 and 4. Recall for class 5, however, decreased from 0.24 to 0.17. The class-aware model also assigned approximately 50% of recorded class-5 visits to class 4, indicating that many errors for the lowest recorded urgency category involved the adjacent urgency level.

For the highest recorded urgency category, the class-aware model correctly classified approximately 19% of visits, compared with 14% for the shared-memory model and 0% for XGBoost. Nevertheless, most class-1 visits remained undertriaged relative to the recorded assignment: approximately 81% were assigned to classes 2–5 by the class-aware model and 86% by the shared-memory model. These percentages use class-1 visits as the denominator, unlike the global undertriage rates in Table I. Therefore, although the Hopfield architectures recognized more visits in the most urgent class, their absolute recall for this category remained limited.

Overall, the confusion matrices indicate that XGBoost favored the majority and central urgency categories, whereas the Hopfield architectures exchanged part of this majority-class performance for improved recognition of several less frequent classes. The class-aware model exhibited higher recall than the shared-memory model for classes 1–4, but lower recall for class 5, and substantial errors toward lower recorded urgency remained.

C. Inference-Time Memory Analysis

Table II presents the results of the inference-time interventions described in Section II-E. All three conditions used the same trained shared-memory checkpoint and the same test visits. Full Memory reproduces the Hopfield Memory results in Table I; Memory Disabled and Uniform Memory modify the retrieved representation without retraining.

Disabling the retrieved representation changed 6.66% of predicted labels and produced a mean Jensen–Shannon distance of 0.0854 relative to Full Memory. Accuracy increased from 0.5222 to 0.5300 and macro-F1 from 0.3681 to 0.3747, while ordinal MAE, undertriage, and severe undertriage decreased. QWK decreased from 0.4514 to 0.4462. Thus, disabling memory yielded more favorable point estimates for five of the six predictive metrics reported in this analysis.

Uniform Memory changed 9.33% of predicted labels and produced a mean Jensen–Shannon distance of 0.0588. This condition obtained the lowest ordinal MAE, undertriage, and severe-undertriage rates among the three interventions, at 0.5450, 0.2134, and 0.0307, respectively. However, its macro-F1 and QWK were lower than those of Full Memory. The proportion of changed labels and the mean probability-distribution distance capture different aspects of prediction sensitivity, so their ordering need not coincide.

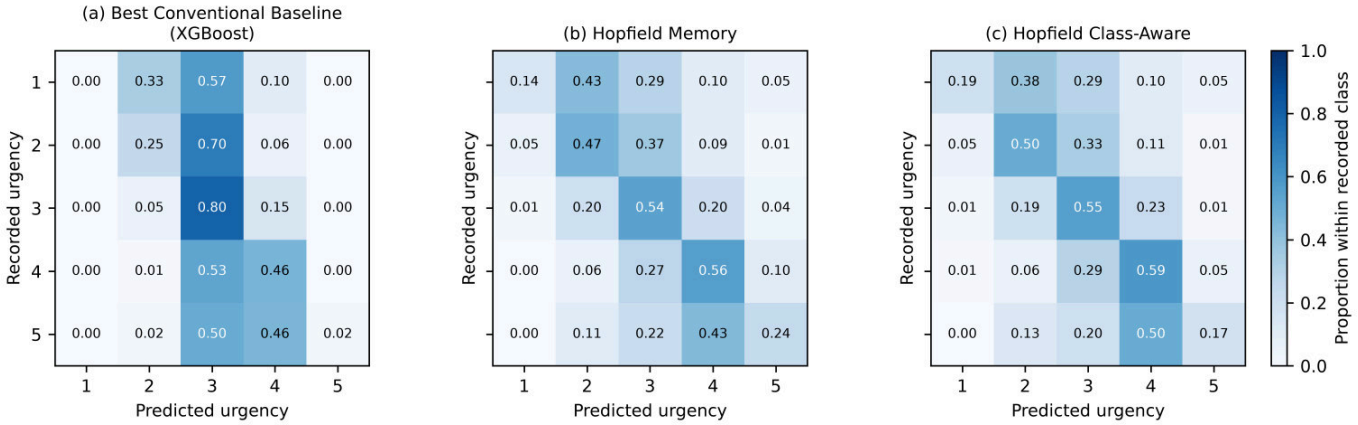


Fig. 4. Row-normalized confusion matrices for XGBoost, the shared-memory Hopfield model, and the class-aware Hopfield model. Rows indicate recorded urgency and columns indicate predicted urgency. Each cell represents the proportion of visits within the corresponding recorded class. Values are rounded to two decimal places.

TABLE II
INFERENCE-TIME INTERVENTIONS IN THE TRAINED SHARED-MEMORY HOPFIELD MODEL. BOLD VALUES INDICATE THE BEST PREDICTIVE METRIC AMONG THESE THREE CONDITIONS. PREDICTIONS CHANGED AND MEAN JENSEN–SHANNON (JS) DISTANCE ARE MEASURED RELATIVE TO FULL MEMORY; NEITHER IS A PERFORMANCE METRIC WITH A PREFERRED DIRECTION. JS DISTANCE USES NATURAL LOGARITHMS.

Condition	Accuracy ↑	Macro-F1 ↑	Ordinal MAE ↓	QWK ↑	Undertriage ↓	Severe Undertriage ↓	Predictions Changed (%)	Mean JS Distance
Full Memory	0.5222	0.3681	0.5614	**0.4514**	0.2402	0.0411	0.00	0.0000
Memory Disabled	**0.5300**	**0.3747**	0.5503	0.4462	0.2343	0.0392	6.66	0.0854
Uniform Memory	0.5281	0.3644	**0.5450**	0.4385	**0.2134**	**0.0307**	9.33	0.0588

IV. CONCLUSION

This study evaluated modern Hopfield networks with patient-level associative memory for five-level prediction of recorded emergency-triage urgency. The results revealed a trade-off between overall performance and class-balanced recognition. XGBoost achieved the highest accuracy and lowest ordinal mean absolute error, whereas the class-aware Hopfield model obtained the highest macro-F1 and balanced accuracy, and the shared-memory model achieved the highest quadratic weighted kappa. Compared with XGBoost, both Hopfield architectures showed higher recall for several underrepresented urgency classes. For the most urgent category, the shared-memory and class-aware models achieved recalls of 0.14 and 0.19, respectively, compared with zero recall for XGBoost. These findings characterize the evaluated Hopfield configurations as alternatives with broader class recognition, accompanied by losses in other performance dimensions.

The class-aware model obtained more favorable point estimates than the shared-memory model for accuracy, macro-F1, balanced accuracy, ordinal MAE, undertriage, and severe undertriage, although its QWK and class-5 recall were lower. Organizing training visits into separate banks provides a mechanism for retrieving class-specific representations, but the observed differences cannot be attributed to this organization alone. The class-aware architecture also used additional retrieved representations, bank-level evidence scores, and a larger classification head.

The inference-time analysis showed that the memory-derived representation affected the outputs of the trained shared-memory model, but learned retrieval did not provide a consistent advantage across the evaluated metrics. Full Memory retained the highest QWK. Relative to Full Memory, disabling the retrieved representation yielded more favorable values for the other five metrics, whereas uniform retrieval produced the lowest ordinal MAE, undertriage, and severe-undertriage rates among the three conditions, together with the lowest macro-F1 and QWK. Learned retrieval was therefore the best condition for only one of the six reported metrics, and sensitivity to memory should not be equated with a consistent predictive benefit. Because these interventions preserved a checkpoint trained with memory, they do not establish the performance of an independently trained memory-free model or isolate the contribution of memory during training. These findings also cannot be extended to the class-aware model without examining its corresponding intervention results.

The improved class-balanced metrics do not establish greater clinical safety. Approximately 81% of recorded class-1 visits were still assigned to less urgent categories by the class-aware model, and its global severe-undertriage rate was 0.0307, compared with 0.0183 for XGBoost. Moreover, these measures describe disagreement with recorded NHAMCS urgency rather than independently adjudicated clinical error. The observed gains in minority-class recognition must therefore be considered alongside the remaining errors toward lower recorded urgency.

The findings are limited to one unweighted NHAMCS sample, one visit-level split, one training seed per Hopfield architecture, and descriptive point estimates without inferential comparisons. Differences in class weighting, memory sampling, and architectural capacity further limit attribution of the observed performance to associative retrieval alone. Future work should prioritize a memory-free encoder–MLP baseline trained under comparable conditions, multiple training seeds with confidence intervals, and hospital-aware or temporally separated evaluation where feasible. Ordinal or asymmetric cost-sensitive objectives that explicitly penalize severe under-triage are also warranted. The present study contributes a comparative characterization of Hopfield-based triage models and their performance trade-offs, while identifying the controls needed to determine the incremental value of associative memory.

Additional research could investigate how memory size and pattern-selection strategies affect the balance between minority-class recognition and ordinal error. Comparisons between class-capped sampling, representative prototypes, and diversity-based selection could help determine whether particular memory configurations provide more useful retrieved information. Examining association weights and retrieved cases alongside prediction changes could also clarify when retrieval supports or impairs classification. Finally, probability calibration and selective prediction with abstention could be explored to identify uncertain cases requiring additional assessment, with their usefulness evaluated through both predictive performance and the proportion and urgency profile of deferred visits.

REFERENCES

[1] H. Gorick, "Factors that affect nurses' triage decisions in the emergency department: a literature review," *Emergency nurse : the journal of the RCN Accident and Emergency Nursing Association*, vol. 30, no. 3, p. 14—19, May 2022. [Online]. Available: https://ueaeprints.uea.ac.uk/id/eprint/86878/1/EN2123_R1.pdf

[2] K. Suamchaiyaphum, A. R. Jones, and S. Polancich, "The accuracy of triage classification using emergency severity index," *International Emergency Nursing*, vol. 77, p. 101537, Dec. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.ienj.2024.101537

[3] P. Davis and D. D. Evans, "The undertriage of older adults in the emergency department: A review of interventions," *Advanced Emergency Nursing Journal*, vol. 43, no. 3, pp. 178–185, 2021. [Online]. Available: http://dx.doi.org/10.1097/TME.0000000000000359

[4] Y. Raita, T. Goto, M. K. Faridi, D. F. M. Brown, C. A. Camargo, and K. Hasegawa, "Emergency department triage prediction of clinical outcomes using machine learning models," *Critical Care*, vol. 23, no. 1, Feb. 2019. [Online]. Available: http://dx.doi.org/10.1186/s13054-019-2351-7

[5] R. A. El Arab and O. A. Al Moosa, "The role of ai in emergency department triage: An integrative systematic review," *Intensive and Critical Care Nursing*, vol. 89, p. 104058, Aug. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.iccn.2025.104058

[6] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve, V. Greiff, D. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "Hopfield networks is all you need," 2020. [Online]. Available: https://arxiv.org/abs/2008.02217

[7] M. Widrich, B. Schäfl, M. Pavlović, H. Ramsauer, L. Gruber, M. Holzleitner, J. Brandstetter, G. K. Sandve, V. Greiff, S. Hochreiter, and G. Klambauer, "Modern hopfield networks and attention for immune repertoire classification," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 832–18 845. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/da4902cb0bc38210839714ebdcf0efc3-Paper.pdf

[8] P. Seidl, P. Renz, N. Dyubankova, P. Neves, J. Verhoeven, J. K. Wegner, M. Segler, S. Hochreiter, and G. Klambauer, "Improving few- and zero-shot reaction template prediction using modern hopfield networks," *Journal of Chemical Information and Modeling*, vol. 62, no. 9, p. 2111–2120, Jan. 2022. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.1c01065

[9] Y. Liang, D. Krotov, and M. J. Zaki, "Modern hopfield networks for graph embedding," *Frontiers in Big Data*, vol. 5, Nov. 2022. [Online]. Available: http://dx.doi.org/10.3389/fdata.2022.1044709

[10] B. Schäfl, L. Gruber, A. Bitto-Nemling, and S. Hochreiter, "Hopular: Modern hopfield networks for tabular data," 2022. [Online]. Available: https://arxiv.org/abs/2206.00664

[11] National Center for Health Statistics, "2022 nhamcs micro-data file documentation," Centers for Disease Control and Prevention, Hyattsville, MD, Tech. Rep., 2024, national Hospital Ambulatory Medical Care Survey, Emergency Department Public Use File.

[12] I. H. Witten, E. Frank, and M. A. Hall, "Data mining: Practical machine learning tools and techniques," in *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011, pp. 3–38.

[13] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, p. 21–27, Jan. 1967. [Online]. Available: http://dx.doi.org/10.1109/TIT.1967.1053964

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, p. 785–794. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[16] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, p. 503–528, Aug. 1989. [Online]. Available: http://dx.doi.org/10.1007/BF01589116

[17] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*. IEEE, Aug. 2010, p. 3121–3124. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2010.764

[18] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *2009 Ninth International Conference on Intelligent Systems Design and Applications*. IEEE, 2009, p. 283–287. [Online]. Available: http://dx.doi.org/10.1109/ISDA.2009.230

[19] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological Bulletin*, vol. 70, no. 4, p. 213–220, 1968. [Online]. Available: http://dx.doi.org/10.1037/h0026256

Multiclass Breast Lesion Classification via Deep Wavelets and Random Tree-Based Feature Transformation and Classification

José Barbosa de Araújo Neto*, Adrielly Sayonara de Oliveira Silva†, Ana Clara Gomes da Silva*, Sandrelly Mirely da Silva Santos†, Camila Tiodista de Lima‡, Gabriela Leite Pereira¶, Marjory Williams Soares Filgueiras de Lima†, Tamara Mayara Ramos Teobaldo*, Letícia Gomes da Silva†, Maria Laura Barbosa de Sousa†, Giselle Machado Magalhães Moreno*, Juliana Carneiro Gomes Cassemiro*, Cecília Cordeiro da Silva*, Artur Aburad de Carvalhosa§, Fabiano Tonaco Borges*, Karla Amorim Sancho*, Wellington Pinheiro dos Santos*, and Maíra Araújo de Santana‡

*Departamento de Engenharia Biomédica, Universidade Federal de Pernambuco, Recife, Brasil
†Departamento de Ciências Farmacêuticas, Universidade Federal de Pernambuco, Recife, Brasil
‡Centro de Informática, Universidade Federal de Pernambuco, Recife, Brasil
§Universidade de Cuiabá, Cuiabá, Brasil
¶Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil

barbosa.araujo@ufpe.br, adrielly.sayonara@ufpe.br, clara.gomes@ufpe.br, sandrelly.santos@ufpe.br, camila.tiodista@ufpe.br, glp@ecomp.poli.br, marjory.soares@ufpe.br, tamara.ramos@ufpe.br, leticia.gsilva3@ufpe.br, maria.laura@ufpe.br, giselle.moreno@ufpe.br, juliana.cgomes@ufpe.br, cecilia.csilva@ufpe.br, arturaburad@gmail.com, fabiano.borges@ufpe.br, karla.amorims@ufpe.br, wellington.santos@ufpe.br, and maira.araujosantana@ufpe.br

***Abstract*—Breast cancer screening via mammography is challenged by the subtlety of early lesions, such as calcifications and architectural distortions, and the inherent subjectivity of visual analysis. This study proposes a computationally efficient and interpretable hybrid protocol for the multiclass classification of mammograms into four categories: normal, calcification, asymmetry/distortion, and masses. The methodology utilizes the public MIAS dataset (322 images) and integrates a five-layer Deep Wavelet Neural Network (DWNN) for high-dimensional feature extraction (5,120 features), followed by a dimensionality reduction step using a PartitionMembership filter based on a Random Tree, yielding a compact set of 161 features. To address class imbalance, the ClassBalancer technique (instance reweighting) was applied exclusively to the training set. Classification was performed using tree-based algorithms (J48, Random Tree, and Random Forest with up to 300 trees), evaluated via 30 repetitions of 10-fold stratified cross-validation and an independent hold-out test set. Notably, the single Random Tree classifier outperformed the ensemble methods, achieving an accuracy of 87.88% and a Cohen's Kappa of 0.7676 on the test set. These results demonstrate that robust wavelet-based feature transformation enables simple, highly interpretable models to rival complex ensembles, offering a promising, low-cost decision support tool for optimizing clinical workflows in resource-constrained healthcare systems.**

***Index Terms*—Mammography, MIAS dataset, deep wavelet neural networks, multiclass classification, random tree, computer-aided diagnosis.**

I. Introduction

Mammography is the most widely used imaging modality for breast cancer screening; however, its sensitivity and specificity vary according to breast tissue density and observer experience [1]. Breast cancer remains one of the most frequently diagnosed neoplasms among women worldwide, posing a significant burden on healthcare systems [2], [3]. The inherent complexity of mammographic images, coupled with the subtlety of certain lesions, renders visual analysis a challenging task prone to inconsistencies. Although mammography is the gold standard for screening—enabling the detection of anomalies before they become palpable [4]—visual interpretation is constrained by human perceptual limitations, such as fatigue and interobserver variability, which can lead to false-negative or false-positive results [5].

In this context, machine learning and computer vision techniques have been increasingly applied to assist in the automated analysis of medical images, aiming to standardize the detection and classification of suspicious patterns [6], [7]. The performance of these models relies heavily on the algorithms' capacity to represent images through numerical descriptors that effectively capture the textural and morphological properties of the lesions.

Although deep Convolutional Neural Networks (CNNs) have achieved remarkable results in computer vision tasks, their application in scenarios with limited datasets—a common constraint in medical imaging—may be hindered by the substantial data requirements for supervised training [8], [9]. As a compelling alternative, approaches based on multiresolution transforms, specifically the Deep-Wavelet Neural Network

(DWNN), enable the hierarchical extraction of features without the need for iterative weight optimization, relying instead on deterministic wavelet-based filters [6], [10].

Nevertheless, the use of deep wavelet networks yields high-dimensional feature vectors. The presence of an excessive number of attributes relative to the sample size can compromise the generalization capability of machine learning models, a phenomenon known as the *curse of dimensionality* [11]. Consequently, applying feature selection or dimensionality reduction techniques is a critical prerequisite to eliminate redundancies and noise prior to classification [12].

The objective of this study is to investigate a processing pipeline for the multiclass classification of mammograms, encompassing four distinct categories: normal, calcification, asymmetry/distortion, and mass. Specifically, we aim to evaluate the efficacy of the DWNN in extracting mammographic textures and its integration with space-partition-based dimensionality reduction methods, utilizing the leaf structure of a Random Tree. The resulting feature set was subjected to a comparative analysis using decision tree-based classifiers, including J48, Random Tree, and Random Forest, evaluated on the Mammographic Image Analysis Society (MIAS) dataset [13].

It is imperative to contextualize the choice of the MIAS dataset for this study. Although it is a classical database comprising digitized radiographic film images, the MIAS remains a widely utilized and consolidated reference for methodological benchmarking in the literature of computer vision applied to mammography. As this is the first study proposing the integration of the DWNN architecture with a recursive partition filter (*PartitionMembership*) for a multiclass problem, utilizing a consolidated dataset is a fundamental methodological step. It allows for the validation of the proposed pipeline's robustness and reproducibility in a reference scenario, while also enabling direct comparison with decades of prior research. Once the viability of this architecture is established on this classical benchmark, future work will advance to its application on modern datasets derived from full-field digital mammography (FFDM) systems, which present greater clinical complexity.

II. Related Works

In the context of supporting breast cancer diagnosis through mammographic images, Santana and Santos [6] proposed a hybrid architecture for lesion detection and classification, combining a Deep Wavelet Neural Network for feature extraction and a Support Vector Machine for the final classification. Unlike traditional Convolutional Neural Networks, the employed DWNN does not require a training process, operating with predefined wavelet filter banks based on the Mallat algorithm and synthesis blocks, which mitigates high computational costs and the dependence on large data volumes. When evaluating the model on the IRMA database, comprising regions of interest with various types of breast tissue density, the authors demonstrated that a 5-layer DWNN configuration associated with an SVM with a second-degree polynomial kernel achieved an average accuracy of 94% and a Kappa index of 0.91 in differentiating between healthy tissues, benign, and malignant lesions. The study concludes that the use of deep networks with fixed weights is a robust and computationally viable alternative for texture analysis and clinical decision support in mammography.

Stefano et al. [14], in turn, conducted a comparative analysis between radiomic methodologies based on machine learning and deep learning techniques for the classification of breast lesions. Using the public CBIS-DDSM repository, the authors evaluated the distinction between benign and malignant lesions through two distinct methodological workflows. In the first, manual extraction of radiomic features, referring to texture, shape, and intensity, was performed using the matRadiomics platform, followed by the application of Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classifiers. In the second, the EfficientNetB6 neural network architecture was utilized, employing transfer learning and parameter fine-tuning. The results indicated that, although the LDA model presented limited performance in the conventional radiomic approach, with an area under the ROC curve (AUC) of approximately 68% for microcalcifications and 61% for masses, the EfficientNetB6-based strategy substantially outperformed traditional methods, achieving AUCs of 81.52% and 76.24% for microcalcifications and masses, respectively. External validation on an independent dataset corroborated the robustness of the findings. The authors conclude that, although traditional radiomics provides greater mathematical interpretability of the extracted parameters, deep neural networks offer superior diagnostic accuracy and greater tolerance to imprecisions in the segmentation process, establishing themselves as a more suitable alternative for clinical decision support.

Aiming to simplify the implementation of predictive algorithms in the clinical environment, Charim et al. [15] investigated the use of automated artificial intelligence platforms for the classification of breast lesions in mammography images. Using the INBreast dataset, the authors developed binary and multiclass distinction models, evaluating the impact of different geometric transformations on the training images. The results indicated that multiclass models outperformed binary ones, with the configuration employing horizontal flipping of the images standing out, achieving the highest accuracy rate of 87.50%. The authors conclude that, despite the limitations imposed by the available data volume, the use of automated machine learning tools eliminates the need for advanced programming knowledge, constituting an accessible alternative to assist healthcare professionals in image screening and the optimization of diagnostic workflows.

On the other hand, Rehman et al. [16] proposed the BRMI-Net framework, a hybrid architecture aimed at the extraction, fusion, and selection of features in mammographic images. The workflow begins with a contrast enhancement technique based on filter fusion, followed by deep feature extraction using the ResNet-50 architecture, which was refined through transfer learning on both the original and enhanced images. To mitigate information redundancy and computational cost, the

authors developed a serial fusion method based on the mean value, integrated with a hybrid feature selection algorithm grounded in flower pollination optimization and the False Position method. When classifying the selected features using machine learning models, with emphasis on the cubic Support Vector Machine, the approach achieved accuracies of 93.8%, 99.5%, and 99.8% on the public CBIS-DDSM, INbreast, and MIAS databases, respectively. The study concludes that intelligent selection and fusion of deep features are important steps to increase diagnostic accuracy and ensure computational efficiency in clinical decision support systems.

Khourdifi et al. [17] proposed an ensemble learning model that combines three convolutional neural networks: VGG16, DenseNet121, and InceptionV3. The approach aims to extract features at multiple scales to improve the identification of benign and malignant masses in mammograms. Evaluated on the INbreast and CBIS-DDSM datasets, the hybrid model demonstrated superiority over individual architectures, achieving accuracies of 90.1% and 89.5%, respectively. The authors conclude that the fusion of the complementary capabilities of these deep networks is effective in mitigating misdiagnoses, reducing false-positive and false-negative rates, and establishing itself as a promising tool for clinical decision support in the early detection of breast cancer.

III. Methods

A. Proposed Method

The experimental workflow is illustrated in Fig. 1. Initially, the mammographic image dataset was preprocessed. This step consisted of resizing all images, loading them in grayscale, normalizing the pixel intensity, and labeling them according to the database classes: normal, calcification, asymmetry/distortion, and masses.

The resulting images served as input for the Deep-Wavelet Neural Network (DWNN), which extracted 5,120 features from the 322 images. These features underwent a dimensionality reduction process using a Random Tree for data representation, resulting in a compact set of 161 features. The instances were divided into 80% for training (n = 256) and 20% for testing (n = 66). The training set was balanced using the Weka ClassBalancer, which adjusts the weight of the instances belonging to each class rather than creating synthetic samples.

Finally, the experiments were conducted. For each analyzed classifier, 30 repetitions of 10-fold cross-validation were executed. Performance was evaluated using accuracy, Cohen's Kappa index, sensitivity, specificity, and AUC-ROC. The best-performing classifier was then evaluated on the independent test set.

B. Database

This study utilizes the Mammographic Image Analysis Society (MIAS) dataset [13]. The dataset comprises PNG images, structured across three categorization axes: BI-RADS classification, breast tissue density, and abnormality type.

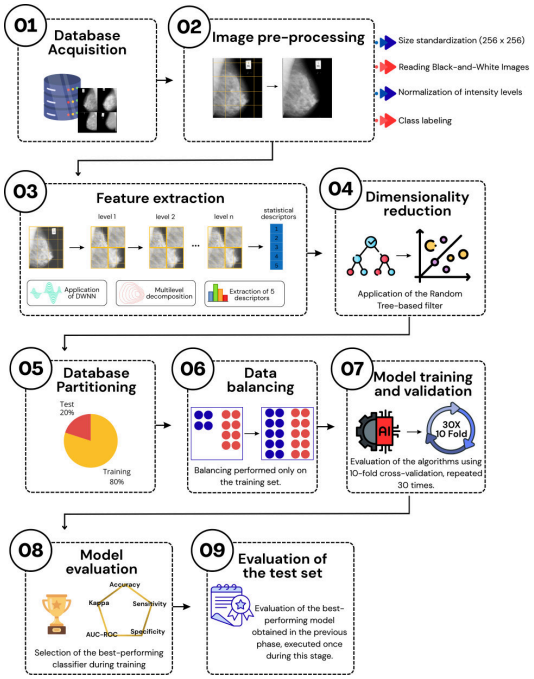


Fig. 1. Methodological workflow proposed for the development of this study.

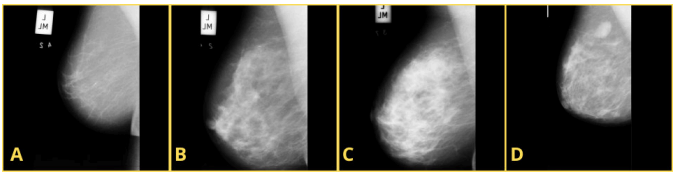


Fig. 2. Representative examples of the four mammography classes used in this study: (A) Normal, (B) Calcification, (C) Distortion/Asymmetry, and (D) Masses.

Regarding the BI-RADS classification, the database is distributed into 207 images in category 1 (normal), 31 in category 2, 32 in category 3, 34 in category 4, and 18 in category 5. Regarding tissue density, the images are categorized into three groups: 106 images of fatty tissue (A), 104 of fatty-glandular tissue (B), and 112 of densely glandular tissue (C/D).

Regarding abnormalities, the dataset encompasses 207 normal images, 25 with calcifications, 34 with architectural distortions or asymmetries, and 56 with masses. This structure allows for the evaluation of computational models in both binary and multiclass classification scenarios. In this work, the multiclass categorization regarding abnormalities was utilized (Fig. 2).

C. Data Preprocessing and Feature Extraction

Dataset processing began with the systematic loading of the 322 digital PNG images, organized in subdirectories according to the abnormality type. Each image was loaded directly in grayscale and converted to an 8-bit unsigned integer representation. For compatibility with the extraction pipeline, the original two-dimensional matrices were expanded into a three-dimensional configuration (height $\times$ width $\times$ 1). The

class labels (normal, calcification, asymmetry/distortion, and masses) were assigned based on the source folder structure, ensuring correct mapping between the input data and the validation categories.

After loading, all samples were subjected to standardized resizing to a resolution of 256×256 pixels. Subsequently, min-max normalization was applied, adjusting the pixel intensity range to the interval [0, 255] and homogenizing the acquisition conditions for the decomposition step [18].

The numerical representation of the mammograms was obtained using a *Deep-Wavelet Neural Network* (DWNN), an architecture that performs hierarchical feature extraction without the need for supervised optimization [6]. As described by Santana and Santos [6], the model replaces the iterative learning of weights with a deterministic filtering scheme based on the Haar discrete wavelet transform, inspired by the Mallat algorithm for multilevel decomposition.

The core of the processing employs a filter bank composed of four 2×2 convolution *kernels*: high-frequency vertical filter, high-frequency horizontal filter, diagonal filter, and *low-pass* filter [10]. The depth of the network is automatically determined by the relation $L = \log_2(\min(H, W)) - 3$, where H and W represent the dimensions of the input image. For images with a resolution of 256×256 pixels, this formula results in $L = \log_2(256) - 3 = 8 - 3 = 5$ layers. This configuration ensures that the final feature map maintains a minimum spatial resolution of 8×8 pixels, preserving sufficient information for statistical synthesis while progressively reducing dimensionality through the *downsampling* process [10], [18].

Upon completing the fifth layer, the network produces a tensor with 1,024 reduced feature maps (4^5) [10]. Each map is processed by a synthesis routine that extracts five statistical metrics: mean, standard deviation, minimum value, maximum value, and mode [6]. The resulting vector contains 5,120 features per image (1,024 maps multiplied by 5 descriptors), representing the textural and structural signature of the dataset images [10]. Fig. 3 illustrates the presented architecture.

D. Dimensionality Reduction

To reduce the dimensionality of the feature set generated by the DWNN, originally comprising 5,120 features, the supervised PartitionMembership filter available in Weka was applied [19]. The filter is based on a trained decision tree, from whose leaves a new dataset is generated. A Random Tree was used, resulting in 161 features, each representing the probability of the input feature vector belonging to a specific leaf.

It is important to note that the PartitionMembership filter extracts the topological structure of the decision tree, generating a generalized partition of the high-dimensional feature space rather than memorizing instance-specific patterns. Once established, this topology is applied deterministically to all instances, producing features based on leaf-node membership. Thus, the resulting 161-feature representation captures the underlying data structure without transferring instance-level information. The training and testing subsets were separated systematically prior to classification, preserving their independence and ensuring reproducibility.

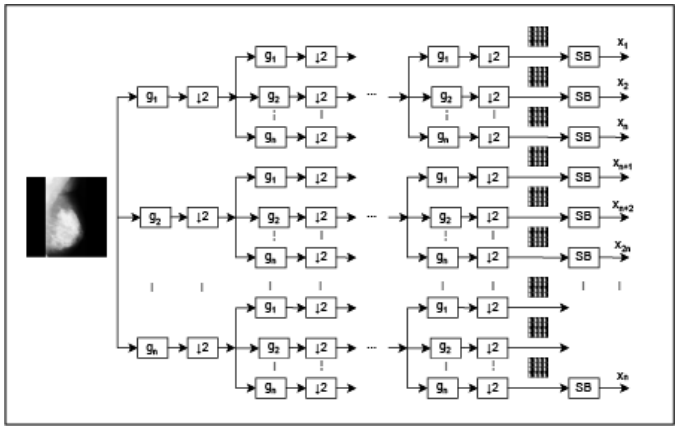


Fig. 3. The DWNN takes a 256 × 256-pixel image as input. At each layer, the image is processed through a filter bank followed by a downsampling operation. At the end of this process, the data passes through a synthesis block, where representative statistics are extracted.

E. Classification

The features were classified using the decision tree-based algorithms J48, Random Tree, and Random Forest (RF). J48 is the Weka implementation of the C4.5 algorithm [19]. For the Random Forest algorithm, different configurations regarding the number of generated trees (10, 50, 100, 150, 200, 250, and 300) were tested, aiming to evaluate the influence of this parameter on classification performance.

F. Metrics

For the evaluation of experimental performance, the following metrics were employed: accuracy, Cohen's Kappa index, sensitivity, specificity, and area under the Receiver Operating Characteristic curve (AUC-ROC) [20].

In a binary classification problem, the confusion matrix is composed of:

- True positives (TP): positive instances correctly classified as positive;
- True negatives (TN): negative instances correctly classified as negative;
- False positives (FP): negative instances incorrectly classified as positive;
- False negatives (FN): positive instances incorrectly classified as negative.

Accuracy corresponds to the ratio between the number of correctly classified instances and the total number of instances (Equation 1) [21]. The Kappa index quantifies the classification agreement, discounting the influence of chance agreement. Consequently, this metric is considered more robust than accuracy, being defined by Equation 2 [22], [23].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\kappa = \frac{\text{Accuracy} - \rho_e}{1 - \rho_e} \tag{2}$$

Where ρ_e is the chance agreement, calculated by:

$$\rho_e = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + FP + FN + TN)^2} \tag{3}$$

The true positive rate, also denoted as recall or sensitivity, quantifies the proportion of correctly classified positive instances relative to the total number of actual positive instances (Equation 4). Specificity represents the proportion of correctly classified negative instances among all negative instances in the dataset (Equation 5) [21].

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{5}$$

The ROC curve relates the false positive rate to the true positive rate. The FPR is defined as the ratio between the number of false positives and the total number of actually negative instances (Equation 6), being equivalent to 1 - specificity. The area under the ROC curve corresponds to the integral of this function, whose optimal value is 1, indicating the correct classification of all positive instances [21].

$$\text{False positive rate} = \frac{FP}{FP + TN} \tag{6}$$

For problems with more than two classes (multiclass scenarios), the metrics defined for the binary case are extended by calculating the indicators for each class individually, treating it as positive and the others as negative (one-vs-rest). Subsequently, the average of these values is obtained. Due to the pronounced imbalance inherent to the MIAS dataset (with the 'normal' class representing the vast majority of samples), the weighted average was adopted for the calculation of multiclass metrics. This approach ensures that the contribution of each class is proportional to its support (number of instances) in the dataset, preventing the overall performance from being artificially inflated by the majority class.

IV. Results and Discussion

The comparative evaluation of the algorithms revealed distinct performance patterns among the tested architectures for the classification of mammograms from the MIAS database [13]. Notably, the Random Tree classifier presented the best individual performance among all analyzed configurations, achieving 87.55% accuracy and a 0.83 Kappa index, demonstrating that simple partition models, when adequately configured, can outperform more complex approaches in certain medical imaging contexts. In contrast, the J48 algorithm presented more modest performance, with 79.37% accuracy and a 0.72 Kappa index, although it still maintains relevant discriminative capacity for the classification task. This performance difference between the two single-tree algorithms can be attributed to the sensitivity of these models to pruning parameters and the adopted splitting criteria. The difficulty in classifying among the four clinical categories—namely, normal, calcification, distortion asymmetry, and masses—resides in the subtlety of textural transitions and the overlap of visual features, requiring models capable of capturing complex nonlinear interactions among the extracted features.

Regarding ensemble methods, the Random Forest algorithm demonstrated remarkably consistent performance across all tested configurations. Table I compiles the results obtained during the training phase, revealing that all configurations, from 10 to 300 trees, converged to identical performance levels, with accuracy varying between 86.01% and 86.16% and Kappa between 0.81 and 0.82. The standard deviations remained consistent across all configurations, indicating that increasing the number of trees did not yield significant discriminative gains and that the model reaches predictive stability even with a reduced number of constituent trees.

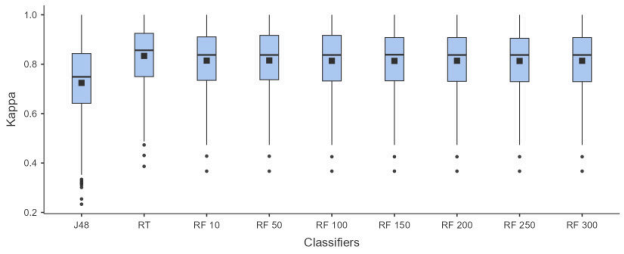
TABLE I
Performance of the classifiers evaluated during the training phase (cross-validation)

Classifiers	Accuracy	Kappa	Sensitivity	Specificity	AUC-ROC
J48	$79.37\% \pm 11.44$	0.72 ± 0.15	0.79 ± 0.11	0.93 ± 0.04	0.89 ± 0.07
RT	$87.55\% \pm 9.82$	0.83 ± 0.13	0.88 ± 0.10	0.96 ± 0.03	0.92 ± 0.07
RF10	$86.12\% \pm 9.66$	0.81 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.94 ± 0.06
RF50	$86.16\% \pm 9.75$	0.82 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.95 ± 0.05
RF100	$86.06\% \pm 9.71$	0.81 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.95 ± 0.05
RF150	$86.03\% \pm 9.77$	0.81 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.95 ± 0.05
RF200	$86.07\% \pm 9.78$	0.81 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.95 ± 0.05
RF250	$86.01\% \pm 9.79$	0.81 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.96 ± 0.05
RF300	$86.06\% \pm 9.78$	0.81 ± 0.13	0.86 ± 0.10	0.95 ± 0.03	0.96 ± 0.05

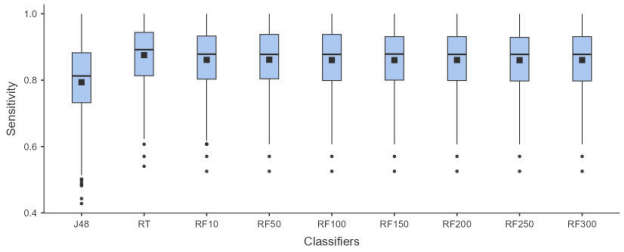
The analysis of the Kappa index distribution across the partitions of the training phase, as illustrated in Fig. 4a, visually corroborates the performance patterns observed in the summarized metrics and provides a richer perspective on model variability. The boxplot reveals that the Random Tree presents the highest median among all evaluated classifiers, positioned at approximately 0.85, with a box extending from about 0.75 to 0.92, denoting superior central performance even when compared to ensemble methods. The J48, in turn, exhibits the lowest median, approximately 0.75, a box shifted to lower levels, between 0.64 and 0.84, and an accentuated dispersion, evidenced by the presence of multiple outliers concentrated in the lower region of the graph, between 0.20 and 0.35. These outliers reflect training/cross-validation partitions in which the single-tree model suffered severe performance degradation, confirming the high variance inherent to this classifier when subjected to different data subsets.

The Random Forest configurations, ranging from 10 to 300 trees, exhibit a remarkably homogeneous visual behavior in the boxplot, in which all boxes present medians clustered around 0.83 to 0.84, practically overlapping interquartile ranges, and similar outlier patterns, concentrated in the range of 0.36 to 0.43. This graphical overlap conclusively demonstrates that increasing the number of trees did not significantly alter the performance distribution across the different cross-validation partitions. The visual consistency of the boxplot boxes for all RFs reinforces the conclusion that the model reaches

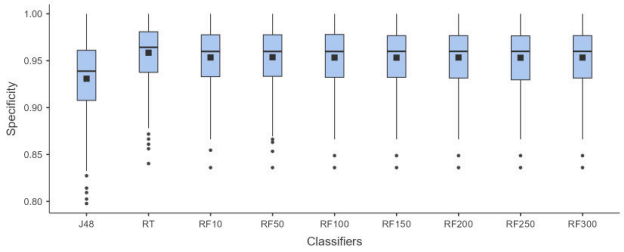
predictive maturity with just 10 trees, rendering the computational increment associated with more complex configurations unnecessary. It is relevant to note that, although the RT presents a slightly higher median than the RFs in the boxplot, the interquartile range of the RFs is comparable, suggesting that ensemble methods offer robust central performance with controlled variability.



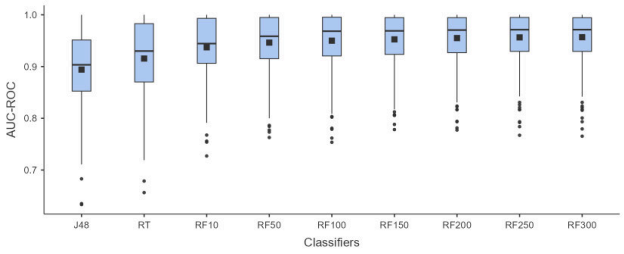
(a) Kappa Index.



(b) Sensitivity.



(c) Specificity.



(d) AUC-ROC.

Fig. 4. Distribution of the metrics (Kappa, Sensitivity, Specificity, and AUC-ROC) of the classifiers evaluated across the 30 repetitions of the 10-fold cross-validation.

The analysis of the boxplots for sensitivity, specificity, and Area Under the ROC Curve illustrated in Fig. 4 reinforces the variability patterns observed in the Kappa index. The J48 algorithm consistently presents the lowest medians and the largest interquartile ranges across the three metrics, evidencing its greater dispersion and lower discriminative capacity compared to the other models. The Random Tree achieves the highest medians in sensitivity and specificity while maintaining moderate dispersion. In contrast, the Random Forest configurations exhibit a highly consistent distribution for all metrics, regardless of the number of trees. While the medians for sensitivity and specificity of the RFs remain stable across the configurations of 10 to 300 trees, a slight progressive increase in the AUC-ROC median is observed as the number of trees grows. However, the overlap of the interquartile ranges among the different Random Forest configurations indicates that increasing the number of trees does not substantially alter the model's variability, corroborating the conclusion that the additional complexity does not practically modify the distribution of predictions.

Regarding the statistical comparison between the Random Tree and Random Forest configurations, the 30 cross-validation repetitions provide robust empirical evidence. As shown in Fig. 4 and Table I, the interquartile ranges and confidence intervals (mean $\pm$ standard deviation) for Kappa and Accuracy substantially overlap across all configurations. This indicates comparable performance, with minor differences likely attributable to sampling variability rather than a true ensemble advantage. Therefore, following the principle of parsimony, the Random Tree is justified as the optimal choice, offering comparable stability with lower computational cost and greater interpretability.

From the perspective of clinical application, the stability visualized in the boxplot is a desirable requirement, as it suggests that the decision support system can maintain reliable performance in the face of morphological variability. In medical practice, such consistency is a prerequisite for building radiologist trust in the system as a secondary opinion. However, it is crucial to emphasize that internal cross-validation stability does not automatically guarantee diagnostic safety or consistent performance across diverse patient populations or acquisition systems. Prospective evaluations in real-world clinical workflows are strictly required before the system can be considered for automated screening.

Considering the principle of parsimony and computational efficiency, the Random Tree configuration was selected as the final model for evaluation on the test set. This decision is based on the fact that the RT achieved the best average performance in model training, with a sensitivity of 0.88, 87.55% accuracy, and a 0.83 Kappa index, outperforming both J48 and all Random Forest configurations, and it did so with a single tree, which implies significantly lower training and inference times. In clinical practice, this optimization translates to faster responses and lower demand for processing resources, facilitating the integration of the system into hospital workflows without compromising diagnostic quality.

After the consolidation of the Random Tree model, an evaluation was performed on an independent test set, comprising 20% of the database, which resulted in 66 images. The results are presented in Table II.

TABLE II
PERFORMANCE OF THE RANDOM TREE MODEL ON THE INDEPENDENT TEST SET

Classifier	Accuracy (%)	Kappa	Sensitivity	Specificity	AUC
RT	87.88%	0.7676	0.879	0.886	0.862

The confusion matrix, presented in Fig. 5, details the model's performance for each class individually. In the test set, the classifier correctly classified 58 out of 66 images, achieving 87.88% accuracy and a Kappa of 0.7676. The model stood out in identifying normal cases, correctly classifying 41 out of 42 exams. Among the pathological classes, the hit rate was 10 out of 12 for masses, 5 out of 7 for distortion/asymmetry, and 2 out of 5 for calcifications. The main confusions occurred when the model classified calcifications and distortions as normal tissue, or when interpreting masses as calcifications.

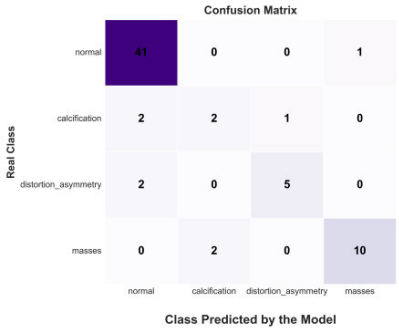


Fig. 5. Confusion matrix of the Random Tree model evaluated on the independent test set. The rows represent the actual classes (Normal, Calcification, Distortion/Asymmetry, and Masses) and the columns the predicted classes, highlighting the model's performance and main points of confusion across the four clinical categories.

The confusion matrix reveals a notable limitation in the classification of calcifications, a class of paramount clinical relevance for early detection. This reduced sensitivity can be attributed to the subtle, high-frequency nature of microcalcifications, which may be attenuated during the downsampling operations inherent to the DWNN architecture, compounded by the severe class imbalance in the MIAS dataset. Furthermore, regarding interpretability, while the final Random Tree model offers a transparent decision structure, it operates on 161 transformed attributes derived from 5,120 wavelet descriptors. Consequently, direct clinical explanation of individual predictions requires caution. Future iterations of this pipeline must incorporate Explainable AI (XAI) techniques, such as SHAP, to map the most influential wavelet-based features back to the original image space for true clinical interpretability.

In summary, the analysis of the results validates the hypothesis that low-complexity models, when supported by a robust feature representation, can rival or outperform computationally expensive ensembles. The Random Tree demonstrated not only the highest accuracy and stability in training but also a generalizable capacity confirmed on the independent test set. Although the classification of subtle lesions, such as calcifications and distortions, represents a challenge inherent to the overlap of textural patterns, the excellent performance in detecting masses and identifying normal tissues positions the model as a promising tool for automated screening. In this scenario, the system could act by filtering low-complexity exams, allowing the radiologist to concentrate their expertise on the most critical cases. Therefore, this study proves the feasibility of developing decision support systems that balance high predictive performance and computational efficiency, essential requirements for the practical integration of artificial intelligence solutions into the clinical routine of image diagnosis.

V. CONCLUSION

This study aimed to develop and evaluate a computational system for the automatic classification of mammograms using the MIAS database, comparing the performance of decision tree-based algorithms and Random Forest ensembles with varying numbers of trees in the discrimination of four clinical categories. The proposed methodology enabled the identification of the architecture that offers the optimal balance between predictive accuracy and computational cost for the multiclass classification task in medical imaging.

The results indicate that the Random Tree model achieved the best overall performance, reaching 87.88% accuracy on the independent test set. From a computational engineering perspective, the analysis demonstrated that increasing model complexity through the use of Random Forest with hundreds of trees did not yield significant improvements in the evaluation metrics, thereby validating the application of the principle of parsimony. This finding reinforces a broader insight: when feature extraction and transformation are sufficiently discriminative, the intrinsic representational capacity of a single, well-structured decision tree can rival or surpass that of computationally expensive ensembles. From a clinical standpoint, the model demonstrated high accuracy in identifying normal cases and masses, contributing to the reduction of false positives and the prioritization of exams with evident lesions. Nevertheless, it is imperative to acknowledge that the lower sensitivity observed for calcifications and architectural distortions represents the primary limitation of the current model. In radiological practice, these subtle features are frequently the earliest indicators of malignancy, such as ductal carcinoma in situ. Therefore, while the excellent performance in detecting masses and normal tissues positions the system as a promising initial screening tool for filtering low-complexity exams, its current inability to consistently capture these early signs precludes its use as a substitute for the detailed analysis performed by a radiologist.

A primary lesson learned during this preliminary study is that the intrinsic representational capacity of a single, well-structured decision tree can rival complex ensembles, provided that the preceding feature extraction and transformation pipeline is sufficiently discriminative. However, a critical research gap remains regarding the robust detection of subtle, high-frequency lesions, such as microcalcifications,

using deterministic wavelet filters on small, historical datasets. Therefore, this work should be interpreted primarily as a methodological proof of concept. Clinical validity and generalization still depend on rigorous external validation on modern FFDM datasets, stronger representation of minority classes, and prospective evaluation in real-world diagnostic workflows.

In conclusion, this study demonstrates the feasibility of employing simple and computationally efficient classifiers for breast cancer diagnostic support, provided they are underpinned by a robust and well-designed feature extraction pipeline. To advance toward the practical deployment of this tool, future work should focus on validating the model on larger and more contemporary datasets derived from full-field digital mammography (FFDM) systems, thereby extending the application of the protocol beyond this initial classical benchmark and assessing its performance in modern clinical scenarios. Furthermore, the incorporation of clinical variables and the use of deep learning techniques for automated feature extraction may enhance the detection of subtle lesions. Finally, prospective studies evaluating the real-world impact of integrating this system into hospital workflows are essential to determine its true contribution to the radiological diagnostic process and its potential to optimize clinical decision-making within resource-constrained healthcare systems.

VI. Declaration on the Use of Artificial Intelligence

The Qwen3.6-Plus language model was employed exclusively for orthographic and grammatical editing of the manuscript. No artificial intelligence was involved in the conception of the research, methodological design, data collection or analysis, interpretation of results, or generation of intellectual content. All AI-assisted processes and outputs were rigorously validated and audited by the human authors, who assume full responsibility for the accuracy, originality, and scientific integrity of this work

Acknowledgment

The authors gratefully acknowledge the financial and institutional support provided by the Brazilian funding agencies FACEPE, CAPES, and CNPq, the Department of Science and Technology of the Ministry of Health (DECIT/MS), and the Instituto do Complexo Econômico-Industrial da Saúde (iCEIS).

References

[1] L. Wang, "Mammography with deep learning for breast cancer detection," *Frontiers in Oncology*, vol. 14, 2024.

[2] D. Trapani, O. Ginsburg, T. A. Fadelu, N. Lin, M. Hassett, A. Ilbawi, B. Anderson, and G. Curigliano, "Global challenges and policy solutions in breast cancer control." *Cancer treatment reviews*, vol. 104, p. 102339, 2022.

[3] C. Barrios, "Global challenges in breast cancer detection and treatment," *The Breast : Official Journal of the European Society of Mastology*, vol. 62, pp. S3 – S6, 2021.

[4] M. Marcon, M. Fuchsjäger, P. Clauser, and R. Mann, "Esr essentials: screening for breast cancer - general recommendations by eusobi," *European Radiology*, vol. 34, pp. 6348 – 6357, 2024.

[5] L. Katsika, E. Boureka, I. Kalogiannidis, I. Tsakiridis, I. Tirodimos, K. Lallas, Z. Tsimtsiou, and T. Dagklis, "Screening for breast cancer: A comparative review of guidelines," *Life*, vol. 14, 2024.

[6] M. A. d. Santana and W. P. d. Santos, "A deep-wavelet neural network to detect and classify lesions in mammographic images," *Research on Biomedical Engineering*, vol. 38, no. 4, pp. 1051–1066, 2022.

[7] T. Cruz, T. Cruz, and W. Santos, "Detection and classification of mammary lesions using artificial neural networks and morphological wavelets," *IEEE Latin America Transactions*, vol. 16, no. 3, pp. 926–932, 2018.

[8] J. Peng and Y. Wang, "Medical image segmentation with limited supervision: A review of deep network models," *IEEE Access*, vol. 9, pp. 36 827–36 851, 2021.

[9] S. Piffer, L. Ubaldi, S. Tangaro, A. Retico, and C. Talamonti, "Tackling the small data problem in medical image classification with artificial intelligence: a systematic review," *Progress in Biomedical Engineering*, vol. 6, 2024.

[10] V. A. d. F. Barbosa, A. F. d. Silva, M. A. d. Santana, R. R. d. Azevedo, R. d. C. F. d. Lima, and W. P. d. Santos, "Deep-wavelets and convolutional neural networks to support breast cancer diagnosis on thermography images," *Journal of Medical and Biological Engineering*, vol. 42, pp. 1–20, 2022.

[11] A. Anuragi, D. S. Sisodia, and R. B. Pachori, "Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets: an empirical review," *Artificial Intelligence Review*, vol. 57, 2024.

[12] R. Zebari, A. Abdulazeez, D. Q. Zeebaree, D. A. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, 2020.

[13] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok *et al.*, "Mammographic image analysis society (mias) database v1.21," *(No Title)*, 2015.

[14] A. Stefano, F. Bini, E. Giovagnoli, M. Dimarco, N. Lauciello, D. Narbonese, G. Pasini, F. Marinozzi, G. Russo, and I. D'Angelo, "Comparative evaluation of machine learning-based radiomics and deep learning for breast lesion classification in mammography," *Diagnostics*, vol. 15, no. 8, p. 953, 2025.

[15] N. S. Charim, N. Ruza, M. H. Arzmi, and S. I. Hussain, "Binary and multi-classification models for breast cancer diagnosis using automated deep learning and mammography images with different augmentation cases," *REVISTA INTERNACIONAL DE METODOS NUMERICOS PARA CALCULO Y DISENO EN INGENIERIA*, vol. 40, no. 4, 2024.

[16] S. u. Rehman, M. A. Khan, A. Masood, N. A. Almujally, J. Baili, M. Alhaisoni, U. Tariq, and Y.-D. Zhang, "Brmi-net: Deep learning features and flower pollination-controlled regula falsi-based feature selection framework for breast cancer recognition in mammography images," *Diagnostics*, vol. 13, no. 9, p. 1618, 2023.

[17] Y. Khourdifi, A. El Alami, M. Zaydi, Y. Maleh, and O. Er-Remyly, "Early breast cancer detection based on deep learning: An ensemble approach applied to mammograms," *BioMedInformatics*, vol. 4, no. 4, pp. 2338–2373, 2024.

[18] M. A. d. Santana, V. A. d. F. Barbosa, R. d. C. F. d. Lima, and W. P. d. Santos, "Combining deep-wavelet neural networks and support-vector machines to classify breast lesions in thermography images," *Health and Technology*, vol. 12, no. 6, pp. 1183–1195, 2022.

[19] E. Frank, M. A. Hall, and I. H. Witten, "The weka workbench. online appendix for" data mining: Practical machine learning tools and techniques"," *Morgan Kaufmann*, 2016.

[20] M. A. de Santana, F. S. Fonseca, A. S. Torcate, and W. P. dos Santos, "Emotion recognition from multimodal data: a machine learning approach combining classical and hybrid deep architectures," *Research on Biomedical Engineering*, vol. 39, no. 3, pp. 613–638, 2023.

[21] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

[22] J. C. Gomes, "ASTERI: representação de sinais de eletroencefalografia em imagens para utilização de arquiteturas profundas baseadas em CNN," Ph.D. dissertation, Escola Politécnica - Universidade de Pernambuco, Recife, 2023.

[23] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/001316446002000104

Image Segmentation of Neonates in the NICU

Helton Maia[1], Aline Elvina Rodrigues Fernandes[2],
Marcos Aurélio Tavares Filho[1], Simone Nascimento Santos Ribeiro[3],
Silvana Alves Pereira[4], Maria da Glória Rodrigues-Machado[3], Richardson Naves Leão[2]
[1]*Escola de Ciências e Tecnologia (ECT)*, UFRN, Natal, RN, Brasil
[2]*Instituto do Cérebro*, UFRN, Natal, RN, Brasil
[3]*Faculdade Ciências Médicas de Minas Gerais (FCM-MG)*, Belo Horizonte, MG, Brasil
[4]*Faculdade de Ciências da Saúde do Trairi (FACISA)*, UFRN, Santa Cruz, RN, Brasil
Corresponding author: richardson.leao@neuro.ufrn.br

***Abstract*—Localising the neonate and its body parts underpins camera-based NICU monitoring, yet the Segment Anything (SAM) family has not been applied to this whole-body and body-part segmentation. We present a promptable-segmentation approach to delineate neonates in routine NICU bedside video (RGB frames from a smartphone camera above the cot), producing both a *whole-body* mask and *body-part* masks for the head, trunk, arms, and legs, plus an auxiliary diaper mask that refines the trunk. Masks are generated with SAM 3 from text prompts alone, without fine-tuning, and assessed by human validation: three evaluators judged the segmentations on 205 frames from 41 infants. Detections were rated correct in 92.8% of cases for the whole body and 94.9% averaged over the body parts, with high inter-evaluator agreement (Gwet's AC1 from 0.74 to 0.97). We release our segmentation and evaluation code, with example outputs, to support reuse for contactless neonatal monitoring. The recordings and frames themselves are not released, and in the distributed examples the infant appears only as an opaque silhouette, in keeping with our privacy-preserving data-handling protocol.**

***Index Terms*—neonatal monitoring, promptable segmentation, Segment Anything Model (SAM), body-part segmentation, NICU, computer vision**

I. Introduction

Neonates in the intensive care unit (NICU), especially preterm infants, require near-continuous physiological monitoring. The clinical standard relies on adhesive electrodes and wired transducers in direct skin contact; yet newborn skin is fragile: before roughly 29 weeks of gestational age the adhesive-to-skin bond can exceed the cohesion of the epidermis, so contact sensors risk skin injury and infection, complicate care, and impede parental skin-to-skin (kangaroo) contact [1]–[3]. These constraints motivate *contactless*, camera-based monitoring, in which one video stream can in principle recover multiple clinical signals without ever touching the infant [3], [4].

Camera-based analysis of the neonate unlocks several high-value clinical applications. Heart rate can be recovered from subtle skin-colour variations (remote photoplethysmography) and respiratory rate from the cyclic motion of the chest, abdomen, and head, with prior NICU studies drawing signals from many body regions (face, head, torso, arms, legs, feet) [2], [3]. Facial-expression analysis enables continuous, less subjective pain assessment [5]. Above all, quantifying spontaneous movement underpins the General Movements Assessment (GMA), among the most accurate early predictors of cerebral palsy: it attains roughly 98% sensitivity before five months of corrected age and supports diagnosis before six months, versus a historical 12 to 24 months, enabling intervention during peak neuroplasticity [6].

A single prerequisite underlies all these applications: the neonate and its body parts must first be reliably localised in the image. Camera vital-sign pipelines hinge on the correct region of interest, and robust localisation of the infant body and face is repeatedly called the *first step* of video-based neonatal assessment [2], [7]. More strongly, augmenting movement video with body-part parsing and pose estimation improves automated GMA for cerebral-palsy prediction [8]. Yet much prior NICU work selected regions of interest manually or with simple colour heuristics inconsistent across clinical conditions, leading one real-NICU study to conclude that "more robust image segmentation is required" for long-term monitoring [2]. Whole-body and body-part (head, trunk, arms, legs) segmentation is therefore the foundational building block for downstream neonatal video analytics.

Establishing that foundation is difficult. Off-the-shelf person, face, and skeleton detectors (e.g., YOLO, Mask R-CNN) and pose models trained on adult-centric data degrade sharply on neonates in cluttered NICU scenes, owing to occlusion by devices and clothing, variable illumination, the distinctive proportions and postures of newborns, and the near-absence of neonatal imagery in their training sets [2], [9], [10]. Infant-specific resources exist but remain small or indirect (e.g., the SyRIP pose dataset [10] and the SMIL 3D infant body model [11]), and progress is further limited by the scarcity of public neonatal data, rooted in the privacy and ethical sensitivity of infant medical imagery [9], [10]. Dedicated neonatal body-part segmentation has been shown with task-specific convolutional networks [12], but a general, low-annotation approach has been lacking.

Promptable foundation segmenters offer a promising route. The Segment Anything Model (SAM) introduced zero-shot, promptable segmentation from points or boxes [13]; SAM 2 extended it to video through a streaming memory that tracks objects across frames [14]; and SAM 3 adds *promptable concept segmentation*, detecting, segmenting, and tracking

every instance matching an open-vocabulary text or exemplar prompt [15]. Such models transfer well to medical imaging when adapted, though vanilla SAM struggles with weak object boundaries [16]. To our knowledge, no prior work applies the SAM family to whole-body and body-part segmentation of neonates in the NICU; the closest dedicated precedent predates SAM and uses a bespoke network [12].

In this work we leverage SAM 3 to segment neonates from real NICU video at two granularities (a whole-body mask and body-part masks for the head, trunk, arms, and legs, together with an auxiliary diaper mask that is segmented and removed from the trunk) as a foundation for downstream pose and movement analysis. Because the recordings depict identifiable infants, all video is processed under Brazil's General Data Protection Law (LGPD) with institutional ethics approval and without any public release of the babies' data; we treat this privacy-preserving handling as both a constraint and a contribution. Our contributions are:

- a SAM 3-based pipeline that produces whole-body and body-part (head, trunk, arms, legs) masks (together with an auxiliary diaper mask used to refine the trunk) for neonates in real NICU video;
- a human validation of this pipeline, in which three evaluators rate the whole-body and body-part segmentations on 205 frames from 41 infants, with detections rated correct in 86.3 to 98.5% of cases per target (Table I); and
- a privacy-preserving acquisition and data-handling protocol, compliant with Brazil's LGPD, that supports clinical reuse without public release of infant imagery.

II. Related Work

Promptable and foundation segmentation. Building on the SAM family introduced above (SAM [13]; SAM 2, which adds video tracking [14]; and SAM 3, which adds promptable concept segmentation [15]), these models transfer to medical imaging when adapted, yet vanilla SAM degrades on targets with weak boundaries and low contrast; MedSAM addressed this by fine-tuning SAM on more than one million medical image-mask pairs spanning many modalities [16]. In contrast, we use the full SAM 3 model with text prompts and no fine-tuning, relying on its concept-level grounding to localise the neonate and its parts.

Infant body-part segmentation and pose. Body parsing and pose estimation are mature for adults but data-scarce for infants, whose proportions, postures, and clinical surroundings differ markedly. Infant-specific resources include the SMIL statistical 3D infant body model [11], the synthetic-and-real SyRIP pose dataset [10], and the larger AggPose dataset with a transformer pose estimator [17]. Dedicated neonatal body-part segmentation has been shown with a task-specific convolutional network on NICU video [12], and augmenting movement video with body parsing and pose improves automated General Movements Assessment for cerebral-palsy prediction [8]. These efforts motivate our coarse head, trunk, arms, and legs targets, but each depends on dedicated training data rather than a promptable foundation model.

Camera-based neonatal monitoring. Contactless neonatal analysis spans vital-sign monitoring from RGB and RGB-D video [2], [3], respiration from motion [4], facial-expression pain assessment [5], [18], face and landmark detection [7], [9], and movement quantification for the General Movements Assessment [6].

SAM in neonatal and paediatric imaging. The SAM family has only recently reached neonatal and paediatric intensive-care imagery, and for narrower tasks than ours: neonatal face segmentation [19] and, with a convolutional network, device-occlusion segmentation in PICU monitoring [20]. To the best of our knowledge, ours is the first to segment both the whole body and its parts in routine NICU video with such a model, the gap addressed here.

III. Materials and Methods

We segment the neonate in routine NICU bedside video at two granularities, a single whole-body mask and four body-part masks (head, trunk, arms, and legs), plus an auxiliary diaper mask used to refine the trunk, using the promptable concept segmentation of SAM 3 [15] rather than a task-specific network. Fig. 1 overviews the pipeline, from a raw frame to whole-body and body-part masks; the temporal propagation (tracking) that SAM 3 enables is left to future work.

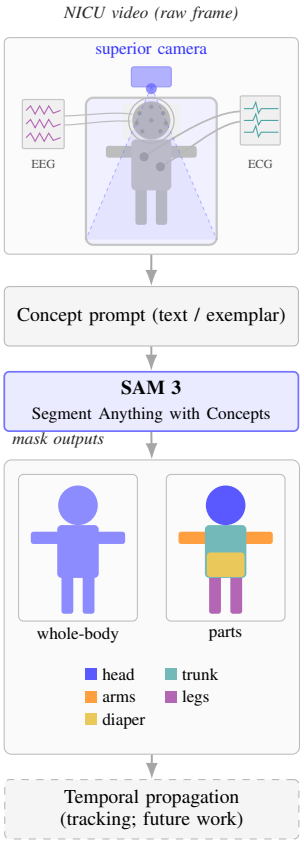


Fig. 1. Schematic of the SAM 3 pipeline: a concept prompt drives the model from a raw NICU frame to a whole-body mask and colour-coded body-part masks (head, trunk, arms, legs); the diaper is segmented as a separate auxiliary concept and subtracted from the trunk.

The following subsections detail, in turn, the dataset and acquisition setup, the segmentation targets, the SAM 3 prompting and post-processing, and the evaluation by human validation.

A. Dataset

This study was approved by the Research Ethics Committee of the Maternidade Escola Januário Cicco (CEP/MEJC), the university maternity hospital of the Federal University of Rio Grande do Norte (UFRN) administered by EBSERH, Natal, Brazil (CAAE 75772223.3.1001.0253), in accordance with Resolution CNS 466/2012. Written informed consent was obtained from the guardian of every infant included; a consent waiver is not admissible for this population. All records were de-identified before processing, and the original recordings are held on institutional storage with access restricted to the team named in the approved protocol. Under Brazil's LGPD (Federal Law 13,709/2018) an image of an identifiable patient is sensitive personal data, so we describe these frames as de-identified rather than anonymised, and no infant imagery is released publicly: the example outputs distributed with the code render the mask as an opaque silhouette. We analyse RGB frames captured with ordinary smartphone cameras positioned above the cot of neonates in the NICU; the present analysis uses a sparse sample of 205 frames, five per infant, from all 41 neonates, so the frame rate is not relevant. Owing to the sensitivity of these data, all analyses are performed on frames in which the infant's face has been obscured by blurring an elliptical region over it. Because this ellipse can extend beyond the face onto the upper trunk or an arm raised near the head, the appearance of those regions is locally altered, so part of the imprecision seen in the head, trunk, and arm masks reflects this privacy blur rather than a failure of SAM 3.

B. Segmentation targets

We segment the neonate at two complementary granularities. The first is a single *whole-body* mask covering the entire infant, which delimits the region of interest for any downstream camera-based analysis. The second parses the infant into four *body-part* masks, head, trunk, arms, and legs (the whole upper and lower limbs), the coarse anatomical regions used by movement and vital-sign pipelines. The trunk denotes the torso (thorax and abdomen); since neonates wear a diaper that occludes the lower trunk, we segment the diaper as a separate region and exclude it from the trunk, so the trunk mask corresponds to the exposed torso. Left and right limbs are not distinguished: the arms (resp. legs) target is the union of all detected arm (resp. leg) instances, leaving per-limb laterality to future work. Apart from the diaper rule above, the parts are segmented as independent concepts with no boundary enforced at their junctions, so neighbouring part masks may overlap slightly; this is acceptable here because the masks serve region localisation and movement analysis, not mutually exclusive pixel labelling. Parts occluded by devices, bedding, or posture are returned as missing or partial detections rather than reconstructed.

C. SAM 3 segmentation pipeline

We use SAM 3 as a promptable *concept* segmenter: a short text prompt naming a target returns the masks of every matching instance in the frame, each with a confidence score (Fig. 1). We run the full SAM 3 image model (the ViT-based checkpoint)[1] rather than the distilled EfficientSAM3, which produced low-confidence, fragmented detections on these NICU frames. Each frame is encoded once and then queried with one text prompt per target: `baby` for the whole body, and `head`, `torso`, `arm`, `leg` for the parts, plus `diaper`. Detections below a confidence threshold of 0.5 are discarded; the surviving instance masks of a target are unioned into its mask, and the diaper is subtracted from the trunk. Formally, for a target's text prompt the model returns instances $\{(m_i, s_i)\}_{i=1}^{k}$ on a frame, where $m_i \in \{0,1\}^{H\times W}$ is a binary mask and $s_i \in [0,1]$ its confidence; the target mask is the union of the instances above the threshold τ,

$$M = \bigcup_{i:\, s_i \geq \tau} m_i, \qquad \tau = 0.5, \tag{1}$$

that is, $M(p) = \max_{i:\, s_i \geq \tau} m_i(p)$ at each pixel p. The trunk is the torso with the diaper removed,

$$M_{\text{trunk}} = M_{\text{torso}} \wedge \neg M_{\text{diaper}} = M_{\text{torso}} \setminus M_{\text{diaper}}, \tag{2}$$

so trunk and diaper are disjoint, $M_{\text{trunk}} \cap M_{\text{diaper}} = \varnothing$. No fine-tuning, per-image tuning, or geometric (point/box) prompts are used; the masks come purely from the text concept.

D. Evaluation by human validation

We assess the pipeline by human validation rather than against a manually annotated reference: the aim is to establish whether promptable concept segmentation reliably localises the neonate and its parts, not to match a pixel-level ground truth, which no annotation of these frames provides. Three evaluators at different training levels (an undergraduate research student, a master's student in neuroscience, and a post-doctoral researcher) independently reviewed all 205 frames; for every frame and target (whole body, head, trunk, diaper, arms, and legs) each rated the segmentation overlay as correct or as one of three error types, *miss*, *false positive*, or *partial*, which are defined and counted in Table I. Two properties of this design bound what it can show. The evaluators rated the pipeline's own overlays, so the review measures whether the produced mask is acceptable, not how it compares blindly with an alternative; and they received only a written sheet defining the four verdicts, with no training on the method and no access to each other's ratings. We report, per target, the share of detections judged correct and the error breakdown, together with the inter-evaluator agreement, which we quantify with Gwet's AC1 [21], [22] because the ratings are heavily skewed toward one category. Concretely, for a given target the $r = 3$ evaluators rate N items (the frames) into $q = 2$

[1]Source code and reproduction scripts available on Zenodo: https://doi.org/10.5281/zenodo.21014348.

categories (correct or incorrect); let n_{ij} be the number of evaluators assigning item i to category j, and $\pi_j = \frac{1}{Nr}\sum_i n_{ij}$ the prevalence of category j. The observed agreement P_a and the Gwet chance term P_e are

$$P_a = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_j n_{ij}(n_{ij}-1)}{r(r-1)}, \tag{3}$$

$$P_e = \frac{1}{q-1}\sum_j \pi_j(1-\pi_j), \tag{4}$$

and the agreement coefficient is

$$\mathrm{AC1} = \frac{P_a - P_e}{1 - P_e}. \tag{5}$$

Unlike Fleiss' κ, whose chance term $\sum_j \pi_j^2$ grows toward 1 when one category dominates (deflating κ), the Gwet P_e stays small, so AC1 reflects the high observed agreement rather than the label imbalance.

IV. Results

SAM 3 first localises the whole infant: from the single text prompt `baby` it produces a whole-body mask of the neonate. Fig. 2 shows this mask on nine different infants, delineating the region of interest for downstream analysis despite device occlusion, bedding, and pose variation.

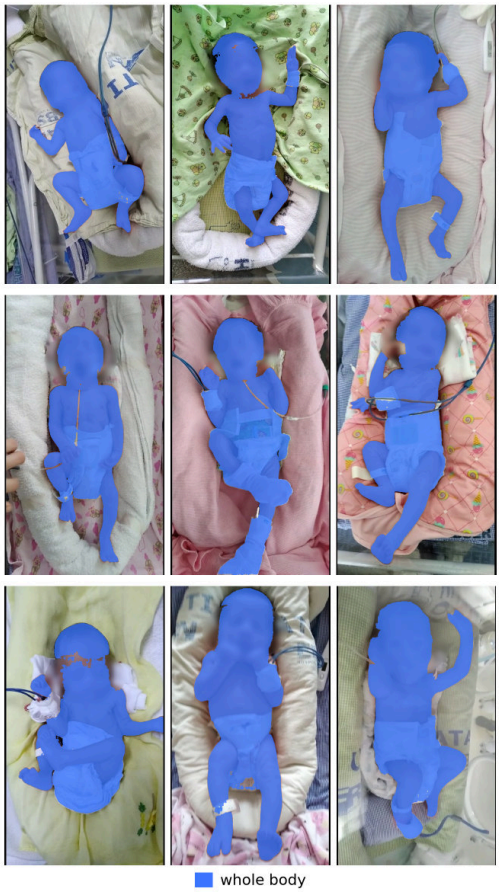


Fig. 2. SAM 3 whole-body masks on nine different neonates, from the single text prompt `baby`; the mask is rendered opaque to obscure the infant.

Beyond the whole body, the same promptable model parses the infant into individual parts. Fig. 3 shows head, trunk, arms, and leg masks across nine neonates, with the diaper segmented separately and removed from the trunk; all masks are obtained from text prompts alone, without per-image tuning.



Fig. 3. Qualitative SAM 3 body-part segmentation on face-blurred NICU frames from nine different neonates: head, trunk, arms, and legs (the diaper is segmented separately and removed from the trunk). Masks are produced from text prompts, without per-image tuning.

Per-frame inference time (Fig. 4) ranges from about 220 s on the CPU to 1.57 s on the desktop RTX 5060 Ti, and 0.43 s and 0.24 s on the NPAD A100 and H200, respectively; the H200 is thus roughly 930× faster than the CPU and 6.7× faster than the desktop GPU. At 2–4 frames per second, the data-centre GPUs make interactive, on-the-fly NICU video analysis feasible, whereas CPU-only inference (minutes per frame) is impractical for continuous monitoring. Each timing covers a full frame: SAM 3 encodes the image once and is then queried with the six concept prompts in turn, so the per-frame cost reflects one image encoding plus the six concept queries rather than a single forward pass. This also sets the model's most immediate role: at A100/H200 rates it is well suited as the offline annotation engine developed below, labelling NICU video at scale, rather than as a bedside real-time tracker.

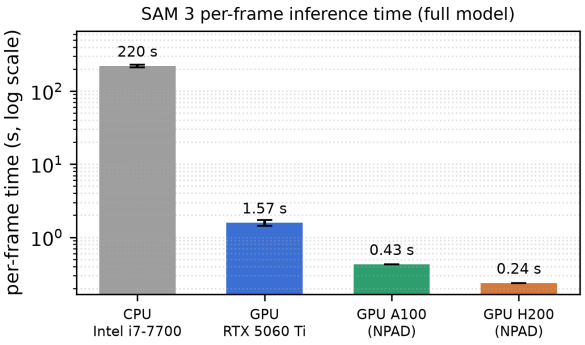


Fig. 4. Per-frame SAM 3 inference time (full model), mean over frames, log scale: CPU (Intel i7-7700), desktop GPU (RTX 5060 Ti, 16 GB), and NPAD A100 (80 GB) and H200 (141 GB). The software stack is not identical across platforms: torch 2.12/CUDA 13.0 locally, torch 2.12/CUDA 12.6 with cuDNN disabled on the A100 (whose driver rejects the bundled cuDNN), and torch 2.7/CUDA 12.6 with cuDNN on the H200, so the bars compare platform classes rather than one controlled software configuration.

Table I reports the human detection-quality review. For each target it gives the percentage of judgments rated correct (*Correct*); the number of judgments in each error class (*Miss*, *FP*, *Partial*); and the inter-evaluator agreement (AC1). Because no verdict was left blank, each target was judged on all 205 frames by the three evaluators, a constant 615 judgments per target; a target that is absent or fully occluded, and correctly left unsegmented, counts as correct rather than as a miss. The masks are usable in practice: detections were rated correct in 86.3% (arms) to 98.5% (legs) of judgments, with the whole body at 92.8%. Errors are dominated by partial masks, most evident for the arms (84 of 615 judgments); misses are concentrated on the diaper (18) and trunk (6), and false positives are rare (6 in total). Agreement is high: the three evaluators agree on 80 to 98% of judgments per target, and AC1 ranges from 0.74 (arms) to 0.97 (head and legs) across the reported targets, reaching 0.98 for the diaper, the quality-assurance layer used to clean the trunk.

TABLE I
HUMAN DETECTION-QUALITY REVIEW OF THE SAM 3 DETECTIONS (THREE EVALUATORS, 205 FRAMES; 615 JUDGMENTS PER TARGET): PER-TARGET SHARE RATED CORRECT, ITS 95% CONFIDENCE INTERVAL BOOTSTRAPPED OVER THE 41 INFANTS, ERROR COUNTS, AND INTER-EVALUATOR AGREEMENT (GWET'S AC1). ERROR TYPES AND AC1 ARE EXPLAINED IN THE TEXT.

Target	Correct (%)	95% CI	Miss	FP	Partial	AC1
Whole body	92.8	89.9–95.4	0	0	44	0.84
Head	97.6	95.6–99.2	2	0	13	0.97
Trunk	97.1	94.6–99.0	6	0	12	0.96
Diaper	95.9	91.1–99.3	18	5	2	0.98
Arms	86.3	81.8–90.6	0	0	84	0.74
Legs	98.5	96.9–99.7	0	1	8	0.97

a) Infant-level uncertainty: The 615 judgments per target are 205 frames rated by three evaluators, and those frames come from 41 infants at five frames each, so they are not 615 independent observations. Taking the infant as the sampling unit leaves the point estimates unchanged, the design being balanced, but widens the uncertainty: the standard error of the correct rate grows by a factor of 1.4 (whole body) to 2.8 (diaper) over a naive binomial on judgments. The confidence intervals in Table I are therefore bootstrapped over the 41 infants. The per-infant spread is where the residual difficulty sits: the weakest infant scores 46.7% on the arms and 66.7% on the whole body, against cohort means of 86.3% and 92.8%.

Sensitivity to the threshold and the prompt

Two free parameters could make these results fragile, the confidence threshold τ and the wording of the text prompt. We measured both on the same 205 frames.

Varying τ from 0.3 to 0.6 changes little: every target keeps a detection on at least 94.6% of frames across that range, so the value used here ($\tau = 0.5$) sits in the middle of a flat region rather than at a tuned optimum. Degradation starts above 0.7 and is target-specific: at $\tau = 0.8$ the trunk still fires on only 29.3% of frames and the arms on 50.7%, while the whole body holds at 99.0%.

Prompt wording acts differently (Table II). Substituting a near-synonym leaves the mask essentially unchanged, with a median IoU against the prompt used here of at least 0.98 for nine of the eleven alternatives tested, reaching 0.998 for `nappy` against `diaper`. What the wording does change is how often the concept fires at all: `newborn` returns a whole-body detection on 79.0% of frames against 100% for `baby`, although the masks it does return remain nearly identical (median IoU 0.988). The one substitution that breaks the pipeline is a change of concept rather than of wording: `chest` in place of `torso` fires on 26.3% of frames with a median IoU of 0.401, since it names a narrower anterior region. The pipeline is thus robust to synonym substitution but not to redefining the target, and the prompts reported here were chosen for concept coverage.

TABLE II
PROMPT SENSITIVITY ON THE SAME 205 FRAMES. FOR EACH TARGET, THE PROMPT USED IN THIS WORK IS COMPARED WITH ALTERNATIVES: SHARE OF FRAMES RETAINING A DETECTION AT $\tau = 0.5$, AND THE IOU OF THE ALTERNATIVE'S MASK AGAINST IT (MEDIAN AND 10TH PERCENTILE OVER FRAMES). AT $\tau = 0.5$ THE PROMPTS USED HERE DETECT ON 97.1 TO 100% OF FRAMES.

Target (prompt)	Alt. prompt	Det. (%)	IoU med.	IoU p10
Whole body (`baby`)	`infant`	97.6	0.994	0.961
	`newborn`	79.0	0.988	0.942
Head (`head`)	`baby head`	99.5	0.980	0.933
	`infant head`	98.5	0.980	0.947
Trunk (`torso`)	`chest`	26.3	0.401	0.283
	`baby torso`	98.0	0.985	0.918
Diaper (`diaper`)	`nappy`	97.1	0.998	0.994
Arms (`arm`)	`baby arm`	100.0	0.981	0.775
	`infant arm`	81.0	0.694	0.371
Legs (`leg`)	`baby leg`	100.0	0.995	0.968
	`infant leg`	94.1	0.994	0.911

V. DISCUSSION

The results support the central claim: a promptable foundation model, driven by text concepts and without any fine-

tuning, segments the neonate and its parts well enough on routine, face-blurred NICU frames to act as the front end of a pipeline for movement and vital-sign analysis. Across the reported targets, 86.3 to 98.5% of detections were rated correct with high inter-evaluator agreement (Gwet's AC1 from 0.74 to 0.97), and false positives were rare across the full review (six in total over the 3,690 judgments, Table I). The larger parts (head, trunk, legs) are the most reliable; the whole-body mask is noisier (it accumulates the most partial masks after the arms) but still delimits the region of interest. These are the regions most directly tied to clinical signals: the trunk for respiration and the whole body for the global movement repertoire. A reliable trunk mask is especially useful here: it could turn the chest-wall movement and thoraco-abdominal asynchrony that clinicians currently grade by eye, as in the Silverman-Andersen respiratory-distress score [23], into an objective, contactless measurement. These downstream uses are stated as hypotheses, not as results: this work evaluates localisation only, and whether these masks improve respiratory-rate estimation, pain assessment, or automated movement scoring remains to be measured on those tasks.

Errors are dominated by *partial* masks rather than misses or false positives, and they concentrate on the arms, the weakest target (86.3% correct, AC1 0.74, 84 partial masks). This is expected: the arms are thin, are often raised or flexed against the body, and are the part most frequently occluded by bedding, leads, and posture. The privacy-preserving blur compounds the problem, because the elliptical blur centred on the face can reach a raised arm or the upper trunk and locally erase the boundary the model relies on; the 44 partial whole-body masks reflect the same clipping at the extremities. The diaper carries the most misses (18), but it is segmented only as an auxiliary quality-assurance layer used to clean the trunk and is not one of the four primary body-part (movement) targets, so it does not affect those results.

Four limitations follow from the study design. First, we evaluate by structured human validation rather than against pixel-level ground truth, since no neonatal whole-body or part annotations exist for these frames; the review therefore measures whether each target is correctly localised, not the exact pixel overlap, and its verdicts are coarser and partly subjective, which we mitigate with three independent evaluators and the prevalence-robust AC1. Building a blinded pixel-level reference on a stratified subset, scored with Dice and boundary metrics, is the natural way to put a number on the overlap and is the next step we intend to take. Second, the sample is modest and sparse (205 frames from 41 infants) and comes from a single unit, so the per-target rates should be read as indicative, and nothing here establishes generalisation to other NICUs, cameras, or acquisition setups. Third, five still frames per infant cannot test temporal behaviour: continuity across a sequence, motion blur, and within-patient variation over a recording are all outside what this design measures, and the temporal stability of the masks is therefore unknown. Fourth, we do not break the results down by occlusion, support equipment, or image conditions, because those attributes are not annotated for these frames; the per-infant spread reported above is the closest available proxy for that heterogeneity.

A further limitation bears directly on the clinical motivation. Movement-based diagnostics span both granularities: whole-body movement underpins the General Movements Assessment, whereas the movement of individual limbs is informative in specific cases, such as a left-right asymmetry or abnormal movements confined to one limb. The pipeline already isolates the limbs, but it unions the two arms (and the two legs) into a single target and does not separate left from right, so resolving the two sides is a prerequisite for these asymmetry analyses and a natural next step.

VI. Conclusion and Future Work

We presented a promptable approach that delineates the neonate in routine NICU video at two complementary granularities, a whole-body mask and separate head, trunk, arms, and leg masks (with an auxiliary diaper mask segmented and removed from the trunk), using the full SAM 3 model with text prompts alone and without any fine-tuning or per-image tuning. The masks are produced under a privacy-preserving, LGPD-compliant protocol on face-blurred frames, and the model reaches interactive per-frame inference times on data-centre GPUs. Promptable concept segmentation thus offers a practical, low-annotation foundation for contactless neonatal video analysis.

A natural next step is to turn this per-frame segmentation into a real-time tracking system for both the whole body and the individual parts. Because SAM 3 inherits the streaming memory introduced in SAM 2 [14], and the full model yields high-confidence per-frame detections, these masks can be propagated across frames rather than recomputed independently, giving temporally consistent whole-body and part masks over time. Given the per-frame rates measured on the A100 and H200, we expect on-the-fly tracking to be feasible on such data-centre hardware, although the tracking path itself is not yet benchmarked. The resulting per-part tracks are the kind of input that pose estimation and automated General Movements Assessment consume [6], [8], though establishing that they improve either task is a separate evaluation that this work does not attempt.

Because the full model is not real-time on commodity hardware (video-rate use currently needs a data-centre GPU), a complementary and more deployable route is to use SAM 3 as an offline *annotation* engine. Its trunk and part masks can label large amounts of NICU video at low manual cost, providing the supervision to train a compact CNN or transformer that then runs in real time, at the frame rate a given signal demands, on ordinary bedside hardware. The trunk (thorax and abdomen) is a natural target: it is the region whose cyclic motion encodes the respiratory rate, so a distilled model could track the trunk and estimate respiration on-device, with SAM 3 providing the labels rather than the runtime.

ACKNOWLEDGMENT

This research was supported by the High Performance Computing Center at UFRN (NPAD/UFRN). We also acknowledge the Brazilian Unified Health System (Sistema Único de Saúde, SUS), under which the neonatal data were collected.

REFERENCES

[1] H. U. Chung *et al.*, "Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units," *Nat. Med.*, vol. 26, no. 3, pp. 418–429, 2020, doi: 10.1038/s41591-020-0792-9.

[2] F.-T.-Z. Khanam *et al.*, "Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks," *J. Imaging*, vol. 7, no. 8, art. 122, 2021, doi: 10.3390/jimaging7080122.

[3] S. Ruhrberg Estévez *et al.*, "Continuous non-contact vital sign monitoring of neonates in intensive care units using RGB-D cameras," *Sci. Rep.*, vol. 15, art. 16863, 2025, doi: 10.1038/s41598-025-00539-9.

[4] S. K. R. Manne *et al.*, "Automatic infant respiration estimation from video: A deep flow-based algorithm and a novel public benchmark," in *Perinatal, Preterm and Paediatric Image Analysis (PIPPI, MICCAI 2023 Workshop)*, LNCS, vol. 14246. Cham: Springer, 2023, pp. 111–120, doi: 10.1007/978-3-031-45544-5_10.

[5] G. Zamzmi *et al.*, "Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN)," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–7, doi: 10.1109/IJCNN.2019.8851879.

[6] I. Novak *et al.*, "Early, accurate diagnosis and early intervention in cerebral palsy," *JAMA Pediatr.*, vol. 171, no. 9, pp. 897–907, 2017, doi: 10.1001/jamapediatrics.2017.1689.

[7] E. Grooby *et al.*, "Neonatal face and facial landmark detection from video recordings," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2023, pp. 1–4, doi: 10.1109/EMBC40787.2023.10340960.

[8] H. Ni *et al.*, "Semi-supervised body parsing and pose estimation for enhancing infant general movement assessment," *Med. Image Anal.*, vol. 83, art. 102654, 2023, doi: 10.1016/j.media.2022.102654.

[9] A. Bin-Obaid *et al.*, "InfantFace: Detecting infant faces in neonatal clinical environments," *arXiv:2606.20449*, 2026.

[10] X. Huang, N. Fu, S. Liu, and S. Ostadabbas, "Invariant representation learning for infant pose estimation with small data," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2021, pp. 1–8.

[11] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2018, pp. 792–800, doi: 10.1007/978-3-030-00928-1_89.

[12] C. Hoog Antink *et al.*, "Fast body part segmentation and tracking of neonatal video data using deep learning," *Med. Biol. Eng. Comput.*, vol. 58, no. 12, pp. 3049–3061, 2020, doi: 10.1007/s11517-020-02251-4.

[13] A. Kirillov *et al.*, "Segment Anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.

[14] N. Ravi *et al.*, "SAM 2: Segment Anything in Images and Videos," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[15] N. Carion *et al.*, "SAM 3: Segment Anything with Concepts," *arXiv:2511.16719*, 2025.

[16] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nat. Commun.*, vol. 15, art. 654, 2024, doi: 10.1038/s41467-024-44824-z.

[17] X. Cao *et al.*, "AggPose: Deep aggregation vision transformer for infant pose estimation," in *Proc. 31st Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022.

[18] L. P. Carlini *et al.*, "Is this neonate feeling pain? Leveraging clinical knowledge towards high-precision large language model-based neonatal pain assessment," *Pediatr. Res.*, 2025, doi: 10.1038/s41390-025-04669-8.

[19] P. H. S. Domingues, T. M. Heiderich, M. C. M. Barros, R. Guinsburg, and C. E. Thomaz, "Neonatal face segmentation with and without clinical devices using SAM," in *Proc. 36th Conf. Graph. Patterns Images (SIBGRAPI), Workshop of Works in Progress*, 2023, doi: 10.5753/sibgrapi.est.2023.27459.

[20] M. F. Munoz, H. V. Huy, T.-D. Le, P. Jouvet, and R. Noumeir, "Hybrid deep learning-based enhanced occlusion segmentation in PICU patient monitoring," *IEEE Open J. Eng. Med. Biol.*, vol. 6, pp. 176–182, 2025, doi: 10.1109/OJEMB.2024.3503499.

[21] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *Br. J. Math. Stat. Psychol.*, vol. 61, no. 1, pp. 29–48, 2008, doi: 10.1348/000711006X126600.

[22] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples," *BMC Med. Res. Methodol.*, vol. 13, art. 61, 2013, doi: 10.1186/1471-2288-13-61.

[23] W. A. Silverman and D. H. Andersen, "A controlled clinical trial of effects of water mist on obstructive respiratory signs, death rate and necropsy findings among premature infants," *Pediatrics*, vol. 17, no. 1, pp. 1–10, 1956.

Um Estudo Comparativo de Lógicas Não-Clássicas Aplicadas a Sistemas de Apoio à Decisão Médica no Contexto de Diabetes

1st Caique Kirilo
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC
Santo André, SP, Brasil
caique.kirilo@ufabc.edu.br

2nd Aline Neves
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC
Santo André, SP, Brasil
aline.neves@ufabc.edu.br

***Resumo*—Este artigo compara lógica fuzzy, multivalorada, paracompleta e paraconsistente em sistemas de apoio à decisão médica para diabetes. A análise avalia como cada formalismo trata gradação, incompletude e conflito informacional. A metodologia combina revisão crítica e experimentos determinísticos em Python, sem treinamento ou avaliação preditiva. Os mesmos registros recebem perturbações pareadas, e o comportamento lógico é medido por cobertura, rejeição e perda de cobertura. Análises de sensibilidade e uma auditoria descritiva com o rótulo original avaliam a estabilidade das conclusões e a relação entre cobertura e concordância. Utiliza-se uma base pública de marcadores clínico-sintomáticos para risco de diabetes em estágio inicial. Discutem-se vantagens, limitações e diretrizes de seleção, com código público para reprodução e extensão.**

***Palavras-chave*—Lógicas não classicas, Diabetes, Apoio à Decisão Médica.**

***Abstract*—This article compares fuzzy, many-valued, paracomplete, and paraconsistent logics in medical decision support systems for diabetes. It evaluates how each formalism handles gradation, incompleteness, and informational conflict. The methodology combines a critical review with deterministic Python experiments, without model training or predictive evaluation. Paired perturbations are applied to the same records, and logical behavior is measured through coverage, rejection, and coverage loss. Sensitivity analyses and a descriptive audit using the original label assess the stability of the conclusions and the relationship between coverage and agreement. A public dataset of clinical-symptomatic markers for early-stage diabetes risk is used. Advantages, limitations, and selection guidelines are discussed, with public code for reproduction and extension.**

***Index Terms*—Non classical Logic, Diabetes, Medical Decision Support.**

I. Introdução

Em sistemas de apoio à decisão médica, a qualidade da inferência depende não apenas da quantidade de dados, mas também da natureza de sua imperfeição. A incerteza gradual ocorre quando um sintoma, risco ou estado clínico admite intensidades intermediárias; a ausência ocorre quando uma informação não foi observada, registrada ou coletada; e a contradição surge quando fontes distintas sustentam simultaneamente uma proposição e sua negação. Esses fenômenos não são equivalentes: um campo ausente não constitui evidência negativa, e duas observações conflitantes não devem ser reduzidas automaticamente a uma média. Registros eletrônicos de saúde apresentam lacunas por diferenças de coleta, custo, disponibilidade de exames e entrada manual [1], enquanto recomendações, diagnósticos e fontes clínicas podem divergir ou contradizer-se [2], [3].

Abordagens existentes geralmente tratam essas imperfeições separadamente: sistemas fuzzy modelam gradações [4]–[6], métodos para dados ausentes priorizam imputação [1] e estudos de inconsistência detectam ou toleram conflitos [2], [3]. Lógicas multivaloradas, paracompletas e paraconsistentes oferecem semânticas específicas para estados intermediários, lacunas e contradições [7]–[9].

Entretanto, faltam comparações pareadas dos quatro formalismos sob os mesmos dados e perturbações. Este trabalho aborda essa lacuna ao avaliar como cada lógica preserva, transforma ou suspende decisões diante de gradação, ausência e contradição controladas.

A proposta diferencia-se dos estudos anteriores por não apresentar uma nova lógica, um método de imputação ou um classificador clínico supervisionado. O objetivo é construir uma comparação operacional, interpretável e reproduzível entre quatro famílias lógicas, mantendo constantes a base, os atributos perturbados e as realizações aleatórias. As diferenças observadas refletem a interação entre a representação semântica de cada formalismo e suas regras operacionais de decisão e rejeição: a fuzzy gradua evidências, a multivalorada introduz estados discretos adicionais, a paracompleta preserva lacunas e a paraconsistente mantém evidências favoráveis e contrárias sem colapso inferencial.

A comparação é conduzida no contexto do auxílio ao diagnóstico de diabetes mellitus tipo 2, utilizando como referência observacional o *Early Stage Diabetes Risk Prediction Dataset*, disponibilizado no Kaggle e documentado no UCI Machine Learning Repository. A base contém 520 registros, 16 variáveis de entrada e um rótulo binário original. Esse rótulo não participa da construção das regras nem da mensuração principal; é consultado apenas em uma auditoria descritiva

posterior. Como não há biomarcadores laboratoriais, os atributos são tratados como marcadores clínico-sintomáticos, não como substitutos de critérios diagnósticos [10], [11].

O objetivo principal deste estudo é discutir como lógicas não clássicas podem auxiliar a representação de evidências no diagnóstico e no acompanhamento da diabetes mellitus tipo 2.

As principais contribuições deste artigo são: (i) uma caracterização conjunta de incerteza gradual, estados intermediários, ausência e conflito; (ii) um protocolo pareado com níveis de ausência e contradição de 10%, 20% e 30%, em 30 repetições; (iii) a explicitação e a análise de sensibilidade dos critérios de decisão; e (iv) uma auditoria complementar da relação entre cobertura, concordância e abstinência. O código é disponibilizado publicamente.

O restante do artigo organiza-se da seguinte forma: a Seção II apresenta a fundamentação teórica; a Seção III descreve metodologia, base de dados e fluxos operacionais; a Seção IV discute os resultados; e a Seção V reúne conclusões e diretrizes de seleção dos formalismos.

II. Fundamentação Teórica

A. Diabetes mellitus tipo 2 e apoio ao diagnóstico

A diabetes é uma doença metabólica crônica caracterizada por hiperglicemia persistente, associada a alterações na secreção ou na ação da insulina. No caso da diabetes mellitus tipo 2, a evolução costuma ser gradual, podendo envolver longos períodos assintomáticos ou sintomas pouco específicos, o que reforça a importância de estratégias de triagem, acompanhamento longitudinal e apoio à decisão clínica [12]–[14].

O diagnóstico e o monitoramento da diabetes baseiam-se principalmente em HbA1c, glicemia de jejum, teste oral de tolerância e glicemia casual. A American Diabetes Association considera compatíveis com diabetes HbA1c $\geq 6{,}5\%$, glicemia de jejum ≥ 126 mg/dL, glicemia de duas horas ≥ 200 mg/dL ou glicemia casual ≥ 200 mg/dL com sintomas, mediante interpretação clínica e confirmação quando necessária [12], [15].

Sinais e fatores de risco, como idade, histórico familiar, sedentarismo, poliúria, polidipsia, fadiga e visão turva, auxiliam a triagem, mas não substituem exames diagnósticos. Sistemas interpretáveis podem apoiar a priorização de casos e integrar evidências discretas, incompletas, graduais ou conflitantes [13]–[16]. Por este motivo, a diabetes é adequada à comparação de lógicas não clássicas. As próximas subseções apresentam esses formalismos como camadas de apoio à decisão, sem substituir critérios médicos ou modelos estatísticos.

B. Lógica Multivalorada

A lógica multivalorada amplia a bivalência ao admitir estados além de verdadeiro e falso. Em aplicações médicas, um valor "indeterminado" representa informação insuficiente, distinguindo desconhecimento de negação e reduzindo ambiguidades em bases incompletas [7], [9].

Para tornar essa ideia operacional, pode-se definir um conjunto finito de estados de decisão:

$$V = \{0, \frac{1}{2}, 1\}. \tag{1}$$

Uma proposição clínica p recebe $v(p) \in V$: 0 indica evidência negativa, 1 evidência positiva e $\frac{1}{2}$ dúvida ou informação parcial.

Neste trabalho adota-se uma semântica trivalorada do tipo Kleene. Para duas proposições p e q, os operadores são:

$$v(\neg p) = 1 - v(p), \tag{2}$$

$$v(p \wedge q) = \min\{v(p), v(q)\}, \tag{3}$$

$$v(p \vee q) = \max\{v(p), v(q)\}, \tag{4}$$

$$v(p \rightarrow q) = \max\{1 - v(p), v(q)\}. \tag{5}$$

O grau de verdade é $T(p) = v(p)$, o de falsidade é $F(p) = 1 - v(p)$ e a certeza é $C(p) = |2v(p) - 1|$. Os extremos 0 e 1 possuem certeza máxima; o valor 0,5 representa máxima indeterminação. Entradas ausentes ou conflitos suspensos recebem 0,5. O escore ponderado é:

$$S_M = \frac{\sum_i w_i v_i}{\sum_i w_i}. \tag{6}$$

O risco é baixo se $S_M < 0{,}35$, moderado se $0{,}35 \leq S_M < 0{,}60$ e alto se $S_M \geq 0{,}60$. Três ou mais lacunas centrais geram saída indeterminada. Contradições são convertidas em 0,5 e, portanto, não permanecem explícitas.

Exemplo numérico: considere poliúria presente ($v_1 = 1, w_1 = 3$), polidipsia desconhecida ($v_2 = 0{,}5, w_2 = 3$), perda de peso ausente ($v_3 = 0, w_3 = 2$) e idade de 45 anos, com contribuição 0,5 e peso 0,75:

$$S_M = \frac{3(1) + 3(0{,}5) + 2(0) + 0{,}75(0{,}5)}{3 + 3 + 2 + 0{,}75} = \frac{4{,}875}{8{,}75} = 0{,}557.$$

Como existe apenas uma lacuna central e $0{,}35 \leq 0{,}557 < 0{,}60$, a saída é risco moderado. Para a polidipsia isoladamente, $T = F = 0{,}5$ e $C = 0$, explicitando a indeterminação.

C. Lógica Fuzzy

A lógica fuzzy representa conceitos graduais por valores em $[0, 1]$. Assim, categorias como "risco moderado" são descritas por funções de pertinência, e não apenas por limites rígidos [4], [5]. Sistemas fuzzy também têm sido empregados em apoio ao diagnóstico e à classificação de diabetes, preservando regras interpretáveis e graus intermediários de evidência [6], [17].

Matematicamente, essa transição gradual é descrita por uma função de pertinência:

$$\mu_A(x) : X \rightarrow [0, 1]. \tag{7}$$

Nessa expressão, x é a entrada observada, X seu domínio e A a categoria clínica. Valores de $\mu_A(x)$ próximos de 0 indicam baixa compatibilidade; próximos de 1, alta compatibilidade; e intermediários, transição gradual.

Formalmente, cada entrada clínica x_i é convertida em um grau $\mu_i \in [0, 1]$. A negação, a conjunção e a disjunção adotadas são, respectivamente:

$$N(\mu) = 1 - \mu, \tag{8}$$

$$T(\mu_1, \mu_2) = \min(\mu_1, \mu_2), \tag{9}$$

$$S(\mu_1, \mu_2) = \max(\mu_1, \mu_2). \tag{10}$$

O grau de verdade é $T_f = \mu$, a falsidade complementar é $F_f = 1 - \mu$ e a certeza é $C_f = |2\mu - 1|$. Esses valores expressam compatibilidade semântica, não probabilidade. A agregação utiliza a média ponderada:

$$S_F = \frac{\sum_i w_i \mu_i}{\sum_i w_i}, \tag{11}$$

O resultado é baixo para $S_F < 0{,}35$, moderado para $0{,}35 \leq S_F < 0{,}65$ e alto para $S_F \geq 0{,}65$. Ausências e conflitos suspensos recebem $\mu = 0{,}5$; o sistema mantém uma saída, mas o conflito deixa de ser identificável após a agregação.

Exemplo numérico: considere pertinências 0,8 para poliúria, 0,6 para polidipsia e 0,4 para perda de peso, com pesos 3, 3 e 2. A conjunção vale 0,6, a disjunção 0,8 e a negação da poliúria 0,2. Para idade de 45 anos ($\mu = 0{,}5, w = 0{,}75$), o escore é:

$$\begin{aligned} S_F &= \frac{3(0{,}8) + 3(0{,}6) + 2(0{,}4)}{8{,}75} \\ &\quad + \frac{0{,}75(0{,}5)}{8{,}75} \\ &= \frac{5{,}375}{8{,}75} = 0{,}614. \end{aligned}$$

Como $0{,}35 \leq 0{,}614 < 0{,}65$, a saída é suspeita moderada. Seu grau complementar de falsidade é $1 - 0{,}614 = 0{,}386$ e a certeza relativa é $|2(0{,}614) - 1| = 0{,}228$.

D. Lógica Paracompleta

A lógica paracompleta trata a falta de informação como um estado legítimo. Assim, quando os dados são insuficientes, uma afirmação clínica pode não ser confirmada nem descartada [7], [9]:

$$\nvdash p \quad \text{e} \quad \nvdash \neg p. \tag{12}$$

Os símbolos indicam que não há base para concluir p nem $\neg p$. Portanto, um sintoma não registrado permanece desconhecido, em vez de ser tratado como ausente.

Cada proposição recebe $E(p) = (t_p, f_p)$, em que as componentes indicam evidência favorável e contrária. Os estados são verdadeiro $(1, 0)$, falso $(0, 1)$ e lacunar $(0, 0)$. O estado contraditório $(1, 1)$ é convertido em lacuna. Os operadores são:

$$E(\neg p) = (f_p, t_p), \tag{13}$$

$$E(p \wedge q) = (\min(t_p, t_q), \max(f_p, f_q)), \tag{14}$$

$$E(p \vee q) = (\max(t_p, t_q), \min(f_p, f_q)). \tag{15}$$

Definem-se verdade $T_p = t_p$, falsidade $F_p = f_p$, certeza $C_p = t_p - f_p$ e incompletude $I_p = 1 - t_p - f_p$. No estado lacunar, $I_p = 1$. Atributos ausentes são excluídos da média:

$$S_P = \frac{\sum_{i \in O} w_i z_i}{\sum_{i \in O} w_i}, \tag{16}$$

onde O é o conjunto de atributos observados. Duas ou mais lacunas centrais geram saída inconclusiva; com uma lacuna central e escore intermediário, a saída é "suspeita com evidência incompleta". Sem essa condição, aplicam-se os limiares 0,35 e 0,65.

Exemplo numérico: considere poliúria presente, $(1, 0)$ e peso 3; polidipsia desconhecida, $(0, 0)$ e peso 3; perda súbita de peso ausente, $(0, 1)$ e peso 2; e idade de 45 anos, contribuição 0,5 e peso 0,75. A polidipsia é excluída da média, resultando em

$$S_P = \frac{3(1) + 2(0) + 0{,}75(0{,}5)}{3 + 2 + 0{,}75} = \frac{3{,}375}{5{,}75} = 0{,}587.$$

O escore é moderado, mas a lacuna central produz "suspeita com evidência incompleta". Para a polidipsia, $I_p = 1$; outra lacuna central tornaria a saída inconclusiva.

E. Lógica Paraconsistente

A lógica paraconsistente trata informações conflitantes sem permitir que uma contradição gere conclusões arbitrárias. Para isso, rejeita o princípio de explosão da lógica clássica [7], [8], [18]:

$$p, \neg p \nvdash q. \tag{17}$$

Assim, a presença simultânea de p e $\neg p$ não permite inferir uma conclusão qualquer q. O conflito é sinalizado, mas as demais evidências continuam utilizáveis.

Cada proposição recebe $E(p) = (\mu_p, \lambda_p)$, com evidências independentes, sendo elas favorável e contrária. O estado $\mu_p = \lambda_p = 1$ representa contradição máxima. Os operadores são:

$$E(\neg p) = (\lambda_p, \mu_p), \tag{18}$$

$$E(p \wedge q) = (\min(\mu_p, \mu_q), \max(\lambda_p, \lambda_q)), \tag{19}$$

$$E(p \vee q) = (\max(\mu_p, \mu_q), \min(\lambda_p, \lambda_q)). \tag{20}$$

O grau de verdade é $T_c = \mu_p$, o grau de falsidade é $F_c = \lambda_p$, o grau de certeza é:

$$G_{\text{cert}} = \mu_p - \lambda_p, \tag{21}$$

e a contradição é $G_{\text{contr}} = \mu_p + \lambda_p - 1$. A saída é inconsistente diante de um conflito central ou três conflitos no conjunto completo. Sem essa condição, $G_{\text{cert}} \geq 0{,}30$ indica alto, $G_{\text{cert}} \leq -0{,}30$ indica baixo e a faixa intermediária indica indeterminação. Ausências geram $(0, 0)$, com preservação do peso no denominador agregado, e contradições permanecem explícitas.

Exemplo numérico: considere três marcadores centrais com pesos $(3, 3, 2)$. Para poliúria, a fonte A registra 1 e a fonte B registra 0; para polidipsia, ambas registram 1; e, para perda súbita de peso, ambas registram 0. A evidência favorável ponderada, considerando as duas fontes, é:

$$\begin{aligned} \mu &= \frac{3(1) + 3(1) + 2(0)}{2(3 + 3 + 2)} \\ &\quad + \frac{3(0) + 3(1) + 2(0)}{2(3 + 3 + 2)} \\ &= \frac{9}{16} = 0{,}5625. \end{aligned}$$

A evidência contrária é $\lambda = 7/16 = 0{,}4375$, resultando em $G_{\text{cert}} = 0{,}125$ e $G_{\text{contr}} = 0$. O grau agregado indicaria indeterminação, mas o conflito central na poliúria prevalece; a saída final é "inconsistente".

III. Metodologia de Comparação

A comparação combinou revisão qualitativa e aplicação computacional de regras lógicas em Python. Os formalismos foram analisados quanto à representação da incerteza, interpretabilidade e reação a dados faltantes ou conflitantes. Não foi construído ou ajustado um modelo estatístico. Os algoritmos e experimentos foram disponibilizados em https://github.com/czkirilo/nonclassical-logics-diabetes para reprodução e extensão.

A. Base de dados

Foram utilizados os 520 registros do *Early Stage Diabetes Risk Prediction Dataset*, coletados por questionário no Sylhet Diabetes Hospital e validados por um médico [10], [11], [19]. O conjunto possui 16 variáveis de entrada e um rótulo binário original, reservado para a auditoria descritiva posterior.

As entradas são idade, gênero e 14 marcadores clínico-sintomáticos. A idade varia de 16 a 90 anos; os demais atributos são categóricos. O rótulo original contém 320 registros positivos (61,5%) e 200 negativos (38,5%) e é informado apenas para descrever a base. A estrutura completa é apresentada na Tabela I.

Tabela I
Estrutura das variáveis do conjunto de dados utilizado.

Grupo	Variáveis	Representação
Demográfica	Idade; gênero	Idade: inteiro (16–90); gênero: binário
Clínico-sintomática	14 marcadores de sinais e sintomas	Binários: presença $= 1$; ausência $= 0$
Rótulo original	Classe de risco	Binário; auditoria posterior

O arquivo analisado não contém valores ausentes ou categorias inválidas. Como não há biomarcadores laboratoriais, o rótulo constitui somente uma referência observacional: não participa da definição dos pesos, limiares ou regras, nem das medidas principais de cobertura e rejeição.

B. Pré-processamento e cenários experimentais

Os nomes e categorias foram padronizados; idade permaneceu inteira, e os campos categóricos foram codificados para a aplicação das regras. Não houve imputação, normalização, balanceamento ou divisão treino–teste: os 520 registros foram usados nas quatro lógicas para comparar regras, não para treinar modelos.

As perturbações foram aplicadas a poliúria, polidipsia, perda de peso, fraqueza, polifagia e paresia parcial. No cenário completo, os dados permaneceram inalterados. No incompleto, cada célula desses atributos foi removida, sem imputação posterior, com probabilidade $p_a \in \{0{,}10; 0{,}20; 0{,}30\}$. Os níveis representam perturbações baixa, intermediária e alta.

No cenário contraditório, a fonte A preservou o valor original $x \in \{0, 1\}$ e a fonte B o inverteu, $x' = 1 - x$, com $p_c \in \{0{,}10; 0{,}20; 0{,}30\}$. Cada nível teve 30 repetições, usando a mesma máscara aleatória nas quatro lógicas. Desacordos foram tratados como evidência suspensa nas lógicas fuzzy, multivalorada e paracompleta; na paraconsistente, as duas fontes foram mantidas separadamente.

Cada repetição seguiu a mesma sequência: (i) copiar os registros originais; (ii) gerar uma única máscara de perturbação; (iii) fornecer os mesmos registros perturbados às quatro lógicas; (iv) registrar o escore, a categoria e o eventual estado de rejeição; e (v) contar decisões e rejeições. O rótulo da base não foi consultado nesse fluxo principal.

C. Implementação computacional e fluxo dos algoritmos

A implementação foi organizada em cinco notebooks para Google Colab: quatro notebooks individuais, um para cada formalismo, e um notebook comparativo que reúne as mesmas funções. O código utiliza `pandas` para manipulação tabular, `NumPy` para cálculos e geração pseudoaleatória e `Matplotlib` para visualização. A Fig. 1 apresenta o fluxo comum do código e a tomada de decisão das quatro lógicas.

Após a leitura do CSV, as respostas `Yes` e `No` são convertidas em 1 e 0, enquanto valores não reconhecidos permanecem como `NaN`. A idade é transformada pela função de pertinência definida anteriormente. Cada função `decisao_` recebe uma linha do *DataFrame*, percorre os 14 sintomas, aplica pesos e tratamento semântico próprio e retorna uma `Series` com escore, contadores de lacuna ou conflito e categoria final. As saídas são concatenadas ao registro original e podem ser resumidas graficamente ou exportadas em CSV.

Embora os notebooks convertam a classe original em uma coluna binária para exibição e exportação, essa coluna não é consultada pelas funções de decisão. O protocolo agregado utiliza sementes inteiras de 0 a 29 para as taxas de 10%, 20% e 30%, com máscaras pareadas entre os formalismos. No painel (b) da Fig. 1, as condições em verde correspondem à ordem efetiva dos testes condicionais no código.

D. Parâmetros, limiares e critérios de decisão

As quatro implementações compartilham um escore heurístico normalizado. Para cada sintoma, $z_i = 1$ indica presença, $z_i = 0$ indica ausência e o peso w_i expressa sua contribuição relativa. Os pesos foram definidos *a priori* como escala ordinal experimental, e não como estimativas de importância diagnóstica. Valores de 2,0 a 3,0 foram atribuídos a poliúria, polidipsia, perda súbita de peso e paresia parcial; valores de 0,75 a 1,5, a polifagia, fraqueza, visão turva, irritabilidade, cicatrização tardia e idade; e valores de 0,25 a 0,5, aos marcadores auxiliares. Essa hierarquia torna explícita a contribuição relativa das evidências, sem pretensão de reproduzir critérios clínicos validados.

A idade contribui com peso 0,75 por meio de uma pertinência gradual:

$$\mu_{\text{idade}}(a) = \begin{cases} 0, & a < 30, \\ (a - 30)/30, & 30 \le a \le 60, \\ 1, & a > 60. \end{cases} \tag{22}$$

O escore básico é a média ponderada dos sintomas e da idade. Os pontos 30 e 60 anos delimitam apenas uma transição didática de baixa para alta contribuição etária; não constituem

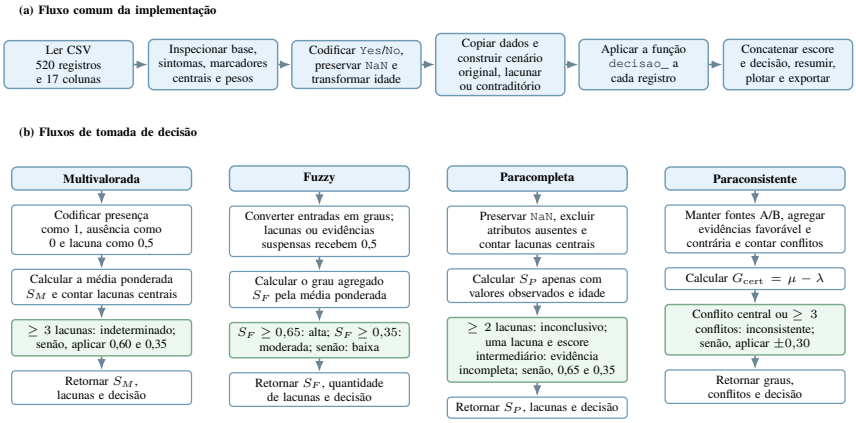


Figura 1. Implementação dos algoritmos: (a) etapas comuns dos notebooks; (b) transformação, agregação e tomada de decisão em cada lógica.

limites diagnósticos. A Tabela II resume o tratamento das imperfeições e os critérios operacionais de cada lógica, que não constituem protocolo médico.

Tabela II
CRITÉRIOS HEURÍSTICOS DE DECISÃO EMPREGADOS NAS QUATRO LÓGICAS.

Formalismo	**Tratamento e decisão**
Multivalorada	Ausências recebem 0,5. Risco baixo se $S < 0{,}35$, moderado se $0{,}35 \leq S < 0{,}60$ e alto se $S \geq 0{,}60$; três lacunas centrais geram indeterminação.
Fuzzy	Ausências ou evidências suspensas recebem 0,5. Suspeita baixa se $S < 0{,}35$, moderada se $0{,}35 \leq S < 0{,}65$ e alta se $S \geq 0{,}65$.
Paracompleta	Ausências são excluídas da média. O risco usa os limiares 0,35 e 0,65; uma lacuna central pode indicar evidência incompleta e duas geram inconclusão.
Paraconsistente	Mantém graus favorável e contrário. É inconsistente com um conflito central ou três conflitos; é alto se $G_{\mathrm{cert}} \geq 0{,}30$, baixo se $G_{\mathrm{cert}} \leq -0{,}30$ e indeterminado entre esses limites.

Os limiares 0,35 e 0,60–0,65 são divisões operacionais da escala normalizada, não pontos de corte médicos. A faixa simétrica $[-0{,}30, 0{,}30]$ cria uma zona de indeterminação paraconsistente. Os critérios de três lacunas na multivalorada e duas na paracompleta expressam diferentes tolerâncias à incompletude. Como essas escolhas influenciam diretamente a cobertura, sua estabilidade é examinada em análise de sensibilidade.

E. Repetições, divisão dos dados e validação

Cada nível aleatório de ausência e contradição foi executado 30 vezes, com sementes inteiras de 0 a 29. Em uma mesma repetição, a máscara sorteada foi reutilizada nos quatro formalismos. Essa estratégia de pareamento controla a entrada experimental: diferenças entre as saídas não podem ser atribuídas ao sorteio de registros ou atributos distintos. A condição completa foi executada uma vez, pois é determinística.

Não houve treinamento, ajuste de parâmetros, divisão treino–teste ou validação cruzada. Todos os 520 registros foram apenas submetidos às regras previamente definidas. A classe original da base não foi usada como resposta a ser prevista, nem para julgar saídas como acertos ou erros; ela permaneceu somente como informação descritiva do conjunto. Portanto, o experimento avalia comportamento inferencial, e não capacidade preditiva ou generalização.

A validação adotada é interna e procedimental. Foram verificados: (i) preservação integral da base no cenário completo; (ii) proporção realizada de células ausentes ou invertidas em relação à taxa nominal; (iii) reutilização da mesma máscara entre as quatro lógicas; (iv) manutenção das classes e dos atributos não selecionados; e (v) coerência das categorias de saída com os limiares definidos. Para cada condição aleatória, as proporções de decisão operacional, indeterminação/inconclusão e conflito, bem como o escore médio, foram calculadas em cada repetição e resumidas por média $\pm$ desvio-padrão. As 30 repetições reduzem a dependência de um único sorteio e permitem avaliar estabilidade, mas não equivalem a validação clínica ou externa.

F. Mensuração do comportamento lógico e opção de rejeição

O teste computacional verificou como cada formalismo responde à mesma entrada antes e depois da inserção de ausência ou contradição. Para cada registro, o algoritmo retornou um escore lógico, uma categoria de decisão e, quando previsto por sua semântica, um estado não operacional: indeterminado, inconclusivo ou inconsistente. Nenhuma saída foi comparada à classe da base para calcular acertos ou erros.

A *cobertura* mede a proporção de registros para os quais a lógica produziu uma decisão operacional:

$$C = \frac{N_{\mathrm{dec}}}{N}. \tag{23}$$

A *rejeição* reúne os registros nos quais o formalismo suspendeu ou bloqueou a decisão e é calculada por $R = 1 - C$. A *perda de cobertura* mede a mudança causada pela perturbação em relação aos dados completos:

$$\Delta C(p) = C(p) - C(0), \tag{24}$$

em que valores negativos indicam redução da proporção de decisões operacionais. Também foi registrado o tipo de estado responsável pela rejeição: indeterminação, inconclusão por ausência ou inconsistência por conflito.

O cenário completo foi executado uma vez por ser determinístico. Nos cenários perturbados, cobertura, rejeição e perda de cobertura foram calculadas nas 30 repetições e resumidas por média $\pm$ desvio-padrão. O desvio-padrão mede a variação entre máscaras aleatórias, não a incerteza de um modelo treinado. O teste de Friedman foi aplicado somente à cobertura das repetições pareadas; quando significativo, utilizou-se Wilcoxon com correção de Holm. Esses testes comparam políticas de decisão, não desempenho de classificação.

Além das rejeições próprias de cada lógica, avaliou-se uma regra uniforme de alta certeza. Para fuzzy, multivalorada e paracompleta, definiu-se $c = |2S - 1|$; para a paraconsistente, $c = |G_{\mathrm{cert}}|$. Com $\tau = 0{,}30$, uma saída permanece operacional apenas quando $c \geq \tau$. Essa análise mede quanto da cobertura é perdido ao exigir maior distância da zona de indeterminação.

G. Sensibilidade, cenário estruturado e auditoria do rótulo

Na análise de sensibilidade, cada peso foi variado independentemente entre 80% e 120% de seu valor original, em 200 configurações geradas com semente 20260825 e as mesmas máscaras das 30 repetições. O limiar uniforme de certeza foi avaliado em $\tau \in \{0{,}20; 0{,}25; 0{,}30; 0{,}35; 0{,}40\}$. Também foram testados critérios de rejeição de duas a quatro lacunas na multivalorada, uma a três lacunas na paracompleta e um a três conflitos centrais na paraconsistente.

Como teste complementar mais próximo de falhas concentradas de documentação, 20% dos registros perderam um bloco de três marcadores; para conflito estruturado, 20% receberam uma única discordância central na fonte B. Esses mecanismos continuam sintéticos, mas evitam que um registro acumule conflitos independentes em várias células.

O rótulo original foi usado apenas após os experimentos, sem ajuste de parâmetros. Saídas baixas foram mapeadas para a classe negativa, saídas moderadas ou altas para a positiva, e estados não operacionais foram tratados como abstinência. Reportaram-se cobertura, concordância seletiva entre os casos decididos e exposição ao erro, isto é, a proporção de decisões discordantes no conjunto completo. Trata-se de auditoria descritiva, não de acurácia diagnóstica.

IV. Resultados

Os resultados descrevem quantos registros receberam uma decisão e quantos foram encaminhados a estados de indeterminação, inconclusão ou inconsistência. Não se avalia se a saída coincide com o rótulo original da base. O foco é a mudança de comportamento de cada lógica quando a mesma entrada sofre perturbações controladas.

A. Cenário completo

A Tabela III resume a resposta das quatro lógicas na ausência de perturbações.

Fuzzy, multivalorada e paracompleta emitiram uma decisão para todos os 520 registros. A paraconsistente manteve 138 registros na zona de indeterminação definida por $|G_{\mathrm{cert}}| < 0{,}30$, resultando em cobertura de 73,5%. Essa diferença decorre da combinação entre representação semântica e política operacional de rejeição, não de erro preditivo.

Tabela III
Resposta lógica no cenário completo (%).

Formalismo	Cobertura	Rejeição
Multivalorada	100,0	0,0
Fuzzy	100,0	0,0
Paracompleta	100,0	0,0
Paraconsistente	73,5	26,5

B. Cenário incompleto

A Tabela IV apresenta a cobertura média e sua mudança em relação ao cenário completo.

Tabela IV
Cobertura sob ausência (%; média $\pm$ desvio-padrão; 30 repetições).

Formalismo	Ausência	Cobertura	ΔC
Multivalorada	10%	$98{,}4 \pm 0{,}5$	$-1{,}6$
	20%	$90{,}1 \pm 1{,}3$	$-9{,}9$
	30%	$74{,}3 \pm 2{,}1$	$-25{,}7$
Fuzzy	10%	$100{,}0 \pm 0{,}0$	0,0
	20%	$100{,}0 \pm 0{,}0$	0,0
	30%	$100{,}0 \pm 0{,}0$	0,0
Paracompleta	10%	$88{,}4 \pm 1{,}6$	$-11{,}6$
	20%	$65{,}5 \pm 2{,}4$	$-34{,}5$
	30%	$41{,}9 \pm 2{,}3$	$-58{,}1$
Paraconsistente	10%	$70{,}1 \pm 1{,}2$	$-3{,}4$
	20%	$65{,}8 \pm 0{,}9$	$-7{,}7$
	30%	$60{,}5 \pm 1{,}2$	$-13{,}0$

A fuzzy manteve 100% de cobertura porque cada lacuna recebeu grau neutro 0,5 e não há rejeição nativa. A multivalorada aplicou a mesma substituição, mas rejeitou três ou mais lacunas centrais, reduzindo a cobertura para 74,3% com 30% de ausência. A paracompleta excluiu os valores ausentes e rejeitou a partir de duas lacunas, atingindo 41,9%. Na paraconsistente, a ausência $(0, 0)$ reduziu $|G_{\mathrm{cert}}|$ em direção à zona de indeterminação, resultando em 60,5%. Assim, cobertura integral significa continuidade de decisão, não preservação da informação original.

C. Cenário contraditório

A Tabela V resume a resposta aos conflitos. A fuzzy absorveu o desacordo como grau intermediário; a multivalorada e a paracompleta suspenderam a evidência, mas aplicaram diferentes limiares de rejeição; e a paraconsistente manteve o conflito explícito.

Com 30% de inversão independente em seis marcadores, a probabilidade de não ocorrer conflito central é $(1 - 0{,}30)^6 = 11{,}8\%$. Após a zona de indeterminação, a cobertura paraconsistente foi 8,8%. Essa redução decorre diretamente da regra conservadora que bloqueia qualquer conflito central e não deve ser interpretada como desempenho preditivo inferior. A fuzzy preservou 100% de cobertura, mas o desacordo deixou de ser identificável após a agregação.

D. Comparações estatísticas

O teste de Friedman identificou diferenças de cobertura entre os quatro paradigmas em todos os níveis perturbados ($p \leq 1{,}45 \times 10^{-17}$). Isso confirma políticas distintas de

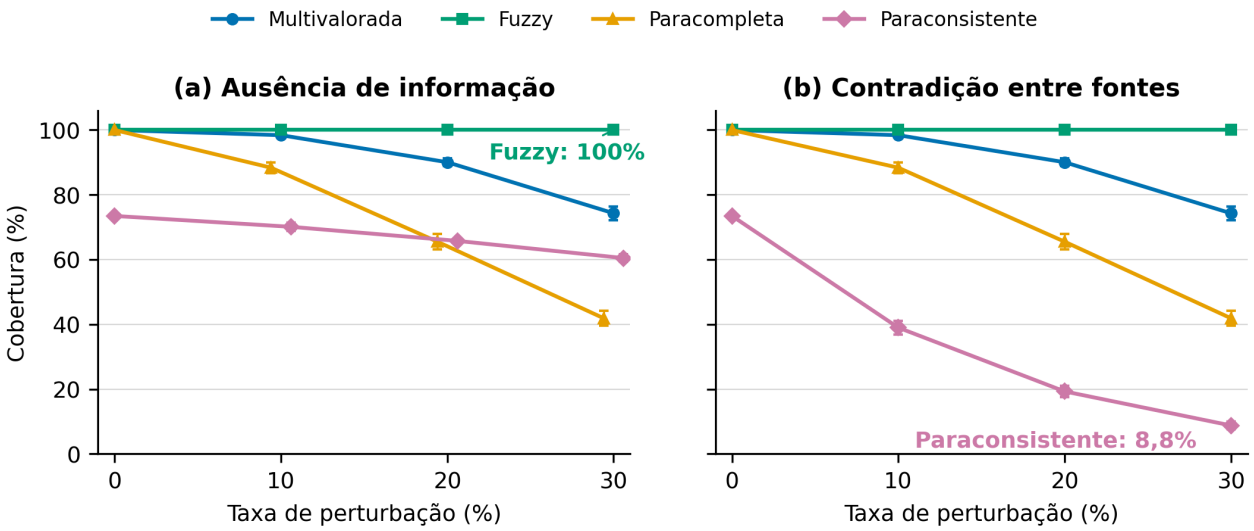


Figura 2. Cobertura das quatro lógicas nos cenários de ausência e contradição. O cenário completo é determinístico; os demais pontos representam média ± desvio-padrão em 30 repetições pareadas.

Tabela V
COBERTURA SOB CONTRADIÇÃO (%; MÉDIA ± DESVIO-PADRÃO; 30 REPETIÇÕES).

Formalismo	Contradição	Cobertura	ΔC
Multivalorada	10%	$98{,}4 \pm 0{,}5$	$-1{,}6$
	20%	$90{,}1 \pm 1{,}3$	$-9{,}9$
	30%	$74{,}3 \pm 2{,}1$	$-25{,}7$
Fuzzy	10%	$100{,}0 \pm 0{,}0$	0,0
	20%	$100{,}0 \pm 0{,}0$	0,0
	30%	$100{,}0 \pm 0{,}0$	0,0
Paracompleta	10%	$88{,}4 \pm 1{,}6$	$-11{,}6$
	20%	$65{,}5 \pm 2{,}4$	$-34{,}5$
	30%	$41{,}9 \pm 2{,}3$	$-58{,}1$
Paraconsistente	10%	$39{,}0 \pm 2{,}1$	$-34{,}5$
	20%	$19{,}3 \pm 1{,}7$	$-54{,}2$
	30%	$8{,}8 \pm 1{,}2$	$-64{,}7$

continuidade e rejeição diante das mesmas máscaras, mas não estabelece qual lógica é mais correta.

E. Sensibilidade dos parâmetros

Com variação independente de $\pm 20\%$ nos pesos, a cobertura sob 30% de ausência permaneceu em 74,3% na multivalorada, 100% na fuzzy e 41,9% na paracompleta; na paraconsistente, variou entre 58,6% e 62,9%. Sob 30% de contradição, a paraconsistente permaneceu entre 8,6% e 9,1%. Portanto, a conclusão qualitativa foi estável aos pesos.

Ao variar τ de 0,20 a 0,40, a cobertura com 30% de ausência caiu de 58,3% para 39,5% na multivalorada, de 74,8% para 47,2% na fuzzy, de 34,5% para 26,5% na paracompleta e de 60,5% para 47,2% na paraconsistente. Maior sensibilidade ocorreu nos critérios estruturais: com 30% de ausência, alterar a rejeição multivalorada de duas para quatro lacunas levou a cobertura de 41,9% a 93,1%; na paracompleta, uma a três lacunas produziram 11,7% a 74,3%. Sob contradição, rejeitar um a três conflitos centrais elevou a cobertura paraconsistente de 8,8% a 48,8%. Logo, os valores absolutos refletem a interação entre semântica e política operacional.

F. Auditoria descritiva e cenário estruturado

A Tabela VI apresenta cobertura e concordância seletiva com o rótulo original. A análise é posterior e não altera pesos ou limiares.

Tabela VI
COBERTURA/CONCORDÂNCIA SELETIVA COM O RÓTULO (%).

Formalismo	Completo	Aus. 30%	Contr. 30%
Multivalorada	100,0/83,5	74,3/81,4	74,3/81,4
Fuzzy	100,0/83,5	100,0/80,1	100,0/80,1
Paracompleta	100,0/83,5	41,9/81,6	41,9/81,6
Paraconsistente	73,5/86,7	60,5/84,1	8,8/87,8

Com 30% de ausência, a fuzzy manteve cobertura integral, mas a concordância caiu de 83,5% para 80,1%, com decisões discordantes em 19,9% do conjunto. A paracompleta apresentou 41,9% de cobertura e 7,7% de decisões discordantes; a paraconsistente, 60,5% e 9,6%. Portanto, 100% de cobertura significa que o método sempre emite uma categoria, não que decide adequadamente.

No cenário estruturado, em que 20% dos registros perderam um bloco de três marcadores, as coberturas foram 80,6% para multivalorada e paracompleta, 100% para fuzzy e 69,0% para paraconsistente. Quando 20% dos registros receberam uma única discordância, as três primeiras mantiveram 100% e a paraconsistente atingiu 59,5%, contra 19,3% no mecanismo independente por célula. A magnitude depende, portanto, de como a imperfeição é gerada, embora a distinção entre continuidade e explicitação do conflito permaneça.

V. CONCLUSÕES

Este artigo caracterizou, sob entradas e perturbações pareadas, o comportamento de lógicas fuzzy, multivalorada,

paracompleta e paraconsistente no apoio à decisão sobre risco de diabetes em estágio inicial. No cenário completo, as três primeiras apresentaram cobertura integral, enquanto a paraconsistente decidiu em 73,5% dos registros. Com 30% de ausência, a cobertura variou de $41{,}9 \pm 2{,}3\%$ na paracompleta a 100% na fuzzy. Com 20% de contradição, os valores foram 100% na fuzzy, $90{,}1 \pm 1{,}3\%$ na multivalorada, $65{,}5 \pm 2{,}4\%$ na paracompleta e $19{,}3 \pm 1{,}7\%$ na paraconsistente.

A análise de sensibilidade mostrou baixa influência de variações de $\pm 20\%$ nos pesos, mas influência importante dos critérios estruturais de rejeição e do mecanismo de perturbação. Portanto, os resultados refletem a interação entre semântica e política operacional, e não uma propriedade isolada de cada lógica. A fuzzy favoreceu continuidade; a paracompleta suspendeu decisões mais cedo sob ausência; e a paraconsistente tornou conflitos visíveis ao custo de maior abstinência.

A auditoria com o rótulo confirmou que cobertura não equivale a adequação. Sob 30% de ausência, a fuzzy decidiu em todos os registros, mas apresentou 19,9% de decisões discordantes no conjunto. Métodos mais conservadores reduziram essa exposição ao custo de menor cobertura. Essa análise permanece descritiva: o rótulo não foi usado para treinamento ou calibração e não constitui diagnóstico confirmado.

As conclusões são limitadas por uma única base de 520 registros, sintomas predominantemente autorreferidos, ausência de biomarcadores e perturbações sintéticas. O cenário estruturado mostrou que a magnitude da cobertura muda quando as falhas se concentram por registro. Além disso, pesos e limiares não foram calibrados por especialistas, e não houve validação externa ou prospectiva. As saídas representam comportamento lógico comparativo, não desempenho diagnóstico ou segurança clínica.

Como trabalhos futuros, pretende-se aplicar o protocolo a bases multicêntricas com exames laboratoriais e conflitos reais entre fontes, calibrar regras com especialistas, avaliar mecanismos empiricamente observados de ausência e desacordo e testar arquiteturas híbridas que combinem triagem gradual, verificação de suficiência e auditoria de inconsistências.

Referências

[1] U. Hwang, S. Choi, H.-B. Lee, and S. Yoon, "Adversarial Training for Disease Prediction from Electronic Health Records with Missing Data," *arXiv preprint arXiv:1711.04126*, 2017. doi: 10.48550/arXiv.1711.04126.

[2] D. Costa and M. A. Martins, "Inconsistencies in Health Care Knowledge," in *Proc. IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2014, pp. 37–42. doi: 10.1109/HealthCom.2014.7001810.

[3] W. Zadrozny, H. Hematialam, and L. Garbayo, "Towards Semantic Modeling of Contradictions and Disagreements: A Case Study of Medical Guidelines," in *Proc. 12th International Conference on Computational Semantics (IWCS), Short Papers*, 2017. [Online]. Disponível em: https://aclanthology.org/W17-6943/. Acesso em: 30 jul. 2026.

[4] L. A. Zadeh, "Fuzzy Sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965. doi: 10.1016/S0019-9958(65)90241-X.

[5] U. Höhle and E. P. Klement, Eds., *Non-Classical Logics and Their Applications to Fuzzy Subsets: A Handbook of the Mathematical Foundations of Fuzzy Set Theory*, ser. Theory and Decision Library B, vol. 32. Dordrecht, The Netherlands: Springer, 1995. doi: 10.1007/978-94-011-0215-5.

[6] A. Srivastava, D. Bhanot, L. H. Jasim, N. Varshney, and V. Patil, "Advancements in Fuzzy Logic Applications for Diagnostic Decision Support Systems in Healthcare," *Fuzzy Information and Engineering*, vol. 17, no. 3, pp. 284–297, 2025. doi: 10.26599/FIE.2025.9270064.

[7] G. Priest, *An Introduction to Non-Classical Logic: From If to Is*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2008. doi: 10.1017/CBO9780511801174.

[8] E. Barrio, F. Pailos, and D. Szmuc, "A Hierarchy of Classical and Paraconsistent Logics," *Journal of Philosophical Logic*, vol. 49, pp. 93–120, 2020. doi: 10.1007/s10992-019-09513-z.

[9] J. Ramos, J. Rasga, and C. Sernadas, "Probabilistic Semantics and Calculi for Multi-valued and Paraconsistent Logics," *Studia Logica*, vol. 113, pp. 1401–1435, 2025. doi: 10.1007/s11225-024-10132-2.

[10] I. Dutta, "Early Stage Diabetes Risk Prediction Dataset," *Kaggle*, 2020. [Online]. Disponível em: https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset. Acesso em: 30 jul. 2026.

[11] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Early stage diabetes risk prediction dataset," *UCI Machine Learning Repository*, 2020. doi: 10.24432/C5VG8H.

[12] American Diabetes Association Professional Practice Committee for Diabetes, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2026," *Diabetes Care*, vol. 49, Suppl. 1, pp. S27–S49, 2026. doi: 10.2337/dc26-S002.

[13] World Health Organization, "Diabetes," *WHO Fact Sheet*, Nov. 14, 2024. [Online]. Disponível em: https://www.who.int/news-room/fact-sheets/detail/diabetes. Acesso em: 30 jul. 2026.

[14] Centers for Disease Control and Prevention, "About Type 2 Diabetes," *CDC Diabetes*, May 15, 2024. [Online]. Disponível em: https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html. Acesso em: 30 jul. 2026.

[15] National Institute of Diabetes and Digestive and Kidney Diseases, "Diabetes Tests & Diagnosis," *NIDDK*, rev. Jul. 2022. [Online]. Disponível em: https://www.niddk.nih.gov/health-information/diabetes/overview/tests-diagnosis. Acesso em: 30 jul. 2026.

[16] International Diabetes Federation, *IDF Diabetes Atlas*, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025. [Online]. Disponível em: https://diabetesatlas.org. Acesso em: 30 jul. 2026.

[17] K. M. Aamir, L. Sarfraz, M. Ramzan, M. Bilal, J. Shafi, and M. Attique, "A Fuzzy Rule-Based System for Classification of Diabetes," *Sensors*, vol. 21, no. 23, Art. no. 8095, 2021. doi: 10.3390/s21238095.

[18] A. de Carvalho Junior, J. F. Justo, A. M. de Oliveira, and J. I. da Silva Filho, "A Comprehensive Review on Paraconsistent Annotated Evidential Logic: Algorithms, Applications, and Perspectives," *Engineering Applications of Artificial Intelligence*, vol. 127, Part B, Art. no. 107342, 2024. doi: 10.1016/j.engappai.2023.107342.

[19] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*, 2020, pp. 113–125. doi: 10.1007/978-981-13-8798-2_12.

Oral Lesion Annotation Tool for Neural Network Training

1st Francisca Thaís Queiroz Rêgo
dept. of Biomedical Engineering
UFRN
Natal, Brazil
thais.rego.092@ufrn.edu.br

2nd Letícia Thaynna Xavier A. O. de Azevedo
dept. of Biomedical Engineering
UFRN
Natal, Brazil
thaynna.araujo.016@ufrn.edu.br

3rd Deyvisom Michel M. Dantas
dept. of Biomedical Engineering
UFRN
Natal, Brazil
deyvisom84@gmail.com

4th Caroline Dantas Vilar
dept. of Biomedical Engineering
UFRN
Natal, Brazil
caroline.vilar@ufrn.br

5th Heliana Bezerra Soares
dept. of Biomedical Engineering
UFRN
Natal, Brazil
heliana.soares@ufrn.br

6th Francisco Eduardo M. de Castro Alves
dept. of Technology
FEMCA
Natal, Brazil
fealvesphone@gmail.com

***Abstract*—The development of Artificial Intelligence (AI) models for oral lesion detection depends on the availability of high-quality and well-structured annotated datasets. However, the construction of such datasets remains a challenge, particularly due to the need for standardized annotation procedures and specialized review. This study presents IA-Bucal Files, a platform developed to organize, annotate, review, and structure clinical images of oral lesions for future applications in Artificial Intelligence and Deep Learning. The platform integrates clinical images and associated metadata, allowing the annotation of lesions, anatomical structures, and exclusion zones through vector-based tools, including polygons, rectangles, circles, and ellipses. The dataset was constructed from images obtained from scientific articles, scientific literature databases, and open-access databases. Annotation was performed by undergraduate students from Dentistry and Biomedical Engineering programs following a preliminary study of the lesions based on scientific literature. The diagnostic classification associated with each image was based on the information provided by its original scientific source, followed by review by Dentistry doctoral students and university professors with a background in Dentistry. Inconsistencies identified during the review process were returned to the annotators for correction and subsequent re-evaluation. To date, the dataset comprises 2,556 clinical images containing oral lesions and 4,067 lesion annotations, distributed among benign, potentially malignant, and malignant lesions. The results demonstrate that the platform provides a structured and standardized workflow for image annotation and review, contributing to the organization and consistency of information related to oral lesions. As the dataset remains under construction, its size and class distribution may change with the incorporation of new images and diagnostic categories. The developed structure has the potential to support future research involving Artificial Intelligence applied to oral lesion detection and analysis.**

***Keywords*—Deep Learning, Medical Image Annotation, Convolutional Neural Networks, Artificial Intelligence.**

I. INTRODUCTION

Oral cancer represents a major public health concern in Brazil due to its high incidence, elevated mortality rates, and the frequent occurrence of late-stage diagnosis. Approximately 70% of cases are identified at advanced stages, negatively affecting patient prognosis, increasing healthcare costs within the Brazilian Unified Health System (SUS), and highlighting the limited access to specialists in Oral Medicine, particularly in Primary Health Care (PHC) services [1]. In this context, Artificial Intelligence (AI) has emerged as a promising approach to support the early detection of oral lesions and assist clinical decision-making.

Among AI techniques, Deep Learning has become one of the most successful approaches for medical image analysis. As a subset of machine learning, Deep Learning employs artificial neural networks with multiple hidden layers capable of automatically extracting complex features from data. In medical imaging, Convolutional Neural Networks (CNNs) have become the state-of-the-art method for image classification, object detection, and image segmentation because of their remarkable ability to recognize complex visual patterns. These networks process digital images through successive computational layers that progressively learn features ranging from simple edges and textures to highly discriminative patterns associated with pathological alterations [2], [3].

In Oral Pathology, Deep Learning models have been increasingly applied to the detection of potentially malignant disorders, oral cancer, dental caries, periodontal diseases, and radiographic abnormalities. However, the performance of these models depends directly on the quality of the datasets used during training. In supervised learning, neural networks learn from previously labeled examples, making the image annotation process one of the most critical stages in the development of reliable and clinically applicable AI systems. Image annotations provide the reference labels required for algorithms to associate visual patterns with specific diagnostic categories, directly influencing performance metrics such as accuracy, sensitivity, and model generalization capability [4].

Given the strong relationship between annotation quality and AI model performance, specialized annotation tools play a

fundamental role in the construction of datasets for supervised learning. In this context, the IA-Bucal Files platform was developed to organize and standardize the annotation process for clinical images of oral lesions. The platform provides a vector-based annotation environment featuring polygons, rectangles, circles, and ellipses, enabling precise delineation of regions of interest. Furthermore, it incorporates contextual intelligence to automatically associate annotations with diagnostic categories previously registered in the image metadata, allowing annotations to be organized into three primary categories: lesions, anatomical structures, and exclusion zones. This approach contributes to the creation of structured and standardized datasets that support the supervised training of neural networks for oral lesion detection.

II. MATERIALS AND METHODS

A. IA-Bucal Files Platform

IA-Bucal Files is a platform developed to support the construction of annotated datasets for training neural networks applied to the detection and classification of oral lesions. Its primary objective is to provide a structured environment for organizing, annotating, validating, and exporting clinical images, thereby enabling the generation of standardized datasets for Artificial Intelligence (AI) and Deep Learning applications.

In addition to image annotation, the platform supports the storage and management of clinical metadata. The integration of imaging and clinical information enriches the dataset, facilitating the development of more robust learning models while supporting clinical and epidemiological analyses and the improvement of intelligent clinical decision-support systems.

B. Dataset Construction

Deep Learning algorithms, particularly Convolutional Neural Networks (CNNs), rely on supervised learning to recognize visual patterns in medical images. Consequently, their performance depends on the availability of large datasets containing thousands of accurately annotated images.

The database used for the IA-Bucal Files platform was built using images of oral lesions obtained from scientific articles published in journals and found in scientific literature indexing and retrieval databases and platforms including PubMed, SciELO, LILACS, Scopus, the Virtual Health Library (VHL), and the Atlas of Dentistry as well as open-access databases.

The images were identified based on the lesions and descriptions found in the respective scientific sources. Following selection, the images were incorporated into the IA-Bucal Files platform, where they were organized and defined before undergoing annotation and review processes.

Using multiple sources enabled the inclusion of a variety of oral lesion types, morphological characteristics, anatomical locations, and image acquisition conditions. Image selection was based on the presence of oral lesions and the availability of diagnostic information associated with the image in the original source.

C. Oral Lesion Annotation

The initial stage of constructing the training set involved annotating oral lesions found in images collected from public databases and scientific articles. This procedure aimed to provide structured information for the supervised training of convolutional neural networks, enabling the algorithms to learn visual patterns associated with various alterations of the oral mucosa.

In this context, Figure 1 illustrates the annotation workflow performed by the staff member on the IA-Bucal Files platform. Initially, the clinical image is uploaded without any markings. Subsequently, both the lesion and the key anatomical structures of the oral cavity are delineated. This strategy provides additional spatial information to the dataset, enabling neural networks to learn not only the morphological characteristics of the lesion but also its relationship with the oral anatomy.

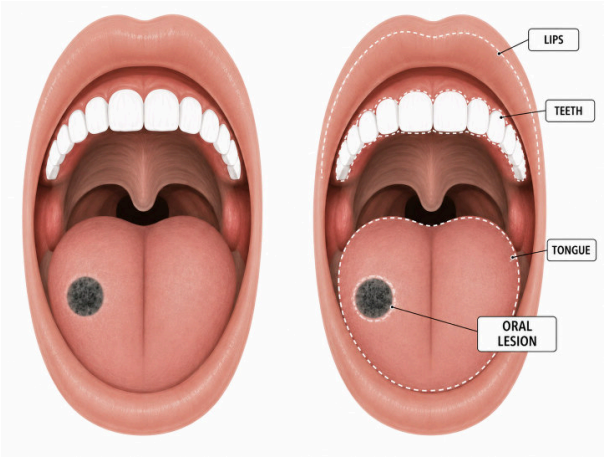


Fig. 1. Example of the annotation process carried out on the IA-Bucal Files platform.

The annotation phase was carried out by students from Dentistry and Biomedical Engineering programs. Prior to the actual marking, the students conducted a preliminary study of the lesions in the dataset, referencing scientific articles related to each specific condition. This stage involved analyzing the clinical, morphological, anatomical, and diagnostic characteristics of the lesions to equip the annotators with the necessary background knowledge for identifying and delineating the regions of interest.

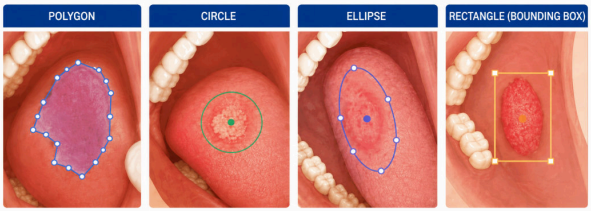


Fig. 2. Representation of the different types of vector markings used on the platform.

Following this preparation phase, annotations were created using vector drawing tools capable of representing the regions of interest. Depending on the observed morphological characteristics, bounding boxes, polygons, circles, and ellipses were employed to represent the anatomical boundaries of the lesions, as shown in Figure 2. Lesions with regular contours were generally delineated using rectangles or ellipses, whereas irregularly shaped lesions were annotated using polygons to better reflect the anatomy observed in the images.

The lesions were classified according to their corresponding diagnostic category, establishing a link between the marked region and the respective diagnostic label. In addition to the lesions, anatomical structures and exclusion zones areas that should not influence model training, such as light reflections, dental instruments, and other artifacts present in clinical photographs were also recorded.

Alongside the annotation process, the diagnostic classification assigned to the images was based on information from the original scientific sources. Thus, the lesion depicted in an image was linked to the diagnosis described in the scientific article or source from which the image was obtained. Consequently, the diagnosis associated with an image was not determined solely by the student performing the annotation but was based on the diagnostic classification provided in the original scientific source.

This procedure made it possible to link each image to its respective lesion and diagnostic category, contributing to the structured organization of the dataset intended for future Artificial Intelligence applications.

Once the annotation phase was complete, the markings underwent a specialized review process. This review was conducted by doctoral students in Dentistry and university professors with dental backgrounds, who evaluated the accuracy of the contours, the appropriateness of the assigned diagnostic category, and the recorded anatomical location.

Whenever inconsistencies were identified regarding lesion delineation, diagnostic classification, or anatomical location, the annotations were returned to the annotator with comments for correction. After corrections were made, the records were submitted for expert evaluation again. Annotations deemed satisfactory following the review process were incorporated into the dataset.

Thus, the annotation development process was structured into successive stages preliminary study of the lesions, student annotation, expert review, correction of inconsistencies, and re-evaluation aiming to enhance the consistency and quality of the information included in the dataset. The present study did not quantitatively evaluate annotation time or annotator workload. These aspects should be investigated in future studies to assess the platform's operational efficiency.

D. Data preparation for neural network training

Following the image annotation and validation phase, the dataset will be prepared for use in supervised learning models. The data will be exported in COCO JSON format—a standard widely used in computer vision applications—as it stores information regarding images, diagnostic categories, coordinates of annotated regions, and associated metadata. This format enables the integration of the dataset with various convolutional neural network (CNN) architectures used for medical image detection, classification, and segmentation tasks.

During supervised training, images are presented to the model along with their respective annotations, enabling the neural network to learn to associate visual patterns with regions previously delineated by experts [2]. This learning process enables neural networks to develop hierarchical representations of the data, making them capable of recognizing complex features present in medical images.

The quality of the training depends directly on the consistency of the annotations used. In medical applications, imprecise or inconsistent annotations can compromise model performance and reduce its generalization capability. In this regard, the review process adopted by the IA-Bucal Files platform contributes to building a more reliable dataset, fostering the development of robust models for the automatic detection of oral lesions. The dataset generated in this study can be used to train various deep learning architectures, enabling future applications in the detection, localization, and segmentation of oral lesions, as well as in clinical diagnostic support systems.

III. Results and Discussion

Following the organization and annotation process using the IA-Bucal Files platform, a total of 2,556 clinical images containing oral lesions and 4,067 lesion annotations have been recorded to date. The number of annotations exceeds the number of images because a single clinical image may display one or more lesions, with each region individually delineated and annotated for the purpose of training and validating artificial intelligence models.

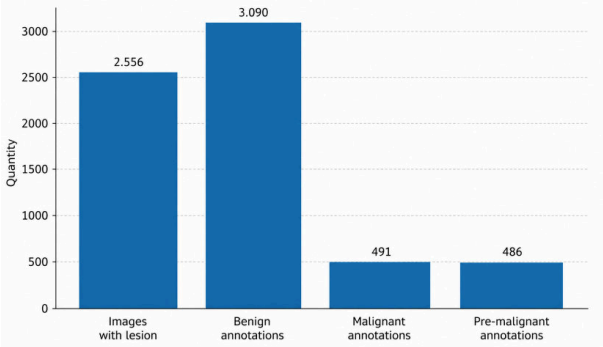


Fig. 3. Distribution of Oral Lesion Annotations in the Training Set

As shown in Figure 3, there is a predominance of benign lesions, accounting for 3,090 entries (75.98°/°), followed by 491 entries for malignant lesions (12.0°/°) and 486 entries for potentially malignant lesions (11.95°/°). This distribution aligns with the epidemiological profile observed in Stomatology and Dentistry services, where benign lesions are diagnosed more frequently than potentially malignant or malignant conditions. Consequently, the database reflects a real-world clinical

scenario in which malignant diseases occur less frequently but hold significant diagnostic importance due to the impact of early diagnosis on patient prognosis.

However, it is important to note that the dataset is still under construction and expansion; thus, the figures presented reflect the database's current state rather than its final composition. As new images and cases are incorporated particularly from various scientific sources—the quantity and proportion of the classes may change, potentially including an increase in the number of potentially malignant and malignant lesions.

Furthermore, the observed distribution reflects the characteristics of the sources used to build the dataset, which include scientific articles, open-access databases, and various scientific literature retrieval platforms. Therefore, the current class composition should not be interpreted as an estimate of lesion prevalence in the general population, but rather as a representation of the distribution of cases currently included in the dataset. In future stages, strategies such as class weighting, data augmentation techniques, and sampling methods may be evaluated to mitigate the potential effects of imbalance during model training. In addition to the volume of images, the large number of annotations demonstrates the platform's ability to produce detailed and structured annotations. Each annotation contains information regarding the location and diagnostic category of the lesion, providing essential data for training detection, classification, and segmentation models. The quality and granularity of the annotations directly influence the performance and generalization capability of algorithms applied to medical image analysis [5].

Another relevant aspect concerns the diversity of the dataset. The images encompass lesions with varying sizes, shapes, colorations, anatomical locations, and clinical presentations, exposing the models to a wide range of visual patterns. This heterogeneity facilitates the learning of discriminative features and reduces the risk of overfitting. Large, diverse datasets are a key factor in the success of Deep Learning applications in medical imaging [4].Similarly, Goodfellow, Bengio, and Courville [6] highlight that the diversity of training examples contributes to the development of more robust representations.

Furthermore, variations in image acquisition conditions—such as lighting, resolution, capture angle, and the presence of adjacent anatomical structures make the dataset more representative of conditions encountered in clinical practice. According to Shorten and Khoshgoftaar [7], this variability can be expanded through data augmentation techniques, further enhancing model robustness.

Database reliability also depends on the quality of the annotations. On the IA-Bucal Files platform, all annotations underwent a structured expert review process before being included in the final dataset. Disagreements among experts represent one of the main challenges in constructing medical databases [8], [9]. Thus, standardized annotation protocols, combined with well-defined diagnostic criteria, increase the reproducibility and consistency of the datasets used in artificial intelligence [9].

Another factor directly influencing model performance is the presence of label noise, characterized by the incorrect assignment of labels or imprecise delineation of regions of interest during the annotation process. In oral lesion detection and segmentation tasks, errors in diagnostic classification or lesion demarcation can lead the model to learn inconsistent patterns, compromising its generalization capability. Incorrect labels significantly reduce the accuracy and generalization ability of neural networks, particularly in medical applications [10], [11].

In this context, a rigorous review of annotations is important for building reliable datasets. Validating the delimited regions and the classification assigned to each lesion allows for the identification of inconsistencies, the correction of labeling errors, and the standardization of the annotation process, thereby reducing the occurrence of label noise. Furthermore, expert review enhances the reliability of the training set, ensuring that relevant morphological characteristics are accurately represented and learned by the model.

The results demonstrate that the platform enabled not only the creation of an extensive and structured dataset, but also the establishment of a standardized workflow for image annotation and review, contributing to the organization and consistency of information related to oral lesions. The developed structure has the potential to support future deep learning applications for the detection, classification, and segmentation of oral lesions.

Unlike studies focused on the development of predictive models, this work concentrates on the construction phase of the annotated dataset, which is considered a cornerstone of supervised learning. Making a standardized dataset available represents a significant contribution to future research in Artificial Intelligence applied to dentistry, providing a reliable foundation for the development and evaluation of computational models.

Finally, Topol [3] highlights that Artificial Intelligence shoul dact as a tool complementing clinical judgment, contributing to increased diagnostic accuracy and reduced inter-examiner variability. Similarly, the World Health Organization emphasizes that the early diagnosis of oral cancer is a key strategy for reducing the morbidity and mortality associated with the disease, especially in regions with limited access to specialists [12]. In this context, standardized databases, such as the one developed in this study, could support the development of systems for tele-dentistry, clinical screening, and diagnostic support based on Artificial Intelligence [13].

IV. Final Remarks

This study presented the IA-Bucal Files platform as a tool for building an annotated database intended for training artificial intelligence models to detect oral lesions. The structured process of annotation, expert review, and validation of markings enabled the creation of a standardized dataset comprising 2,556 clinical images and 4,067 diagnostic annotations to date thereby establishing a consistent for supervised learning applications.

The results demonstrate that adopting standardized annotation protocols contributes to increasing the reliability and

quality of the data used to train Convolutional Neural Networks—an aspect considered essential for developing models with greater generalization capabilities and potential for clinical application. Furthermore, the use of the COCO JSON format enhances the database's interoperability with various Deep Learning architectures designed for the classification, detection, and segmentation of medical images.

Future plans involve continuously expanding the database with new images and diagnostic categories, as well as using this training set to develop and evaluate Artificial Intelligence models for the automatic detection of oral lesions. This initiative is expected to contribute to the advancement of diagnostic support tools, fostering strategies for tele-dentistry, clinical screening, and early detection of oral cancer, particularly in settings with limited access to specialists.

ACKNOWLEDGMENTS

The authors thank LABETRE (Laboratory of Tissue Engineering and Rehabilitation) and LAPSIM (Laboratory of Medical Signal and Image Acquisition and Processing) of the Biomedical Engineering program at the Federal University of Rio Grande do Norte (UFRN) for the technical support, infrastructure, and assistance provided for the development of this work. The authors also acknowledge the National Council for Scientific and Technological Development (CNPq) for the financial support provided for the development of this project.

REFERENCES

[1] M. G. O. da Silva, "Diagnóstico precoce de câncer bucal na atenção primária: O papel do cirurgião-dentista da ESF e a indicação oportuna de biópsias," Brazilian Journal of Implantology and Health Sciences, vol. 7, no. 7, pp. 263–272, 2025.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[3] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017.

[5] Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, Art. no. 60, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[9] H. D. Zając et al., "Ground truth or dare: Factors affecting the creation of medical datasets for training AI," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2023, pp. 351–362.

[10] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, pp. 845–869, May 2014.

[11] Y. Wei, et al., "Deep learning with noisy labels in medical prediction problems: A scoping review," Journal of the American Medical Informatics Association, vol. 31, no. 7, pp. 1596–1607, 2024.

[12] World Health Organization, "Oral cancer," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/oral-cancer

[13] R. Haron, B. Zain, and A. Ramanathan, "m-Health for early detection of oral cancer in low- and middle-income countries," Telemedicine and e-Health, vol. 26, no. 3, pp. 278–285, 2020.

ORBITA-Health: Auditable Curation of Health and Disability Indicators for Public Observatories

1º José Jerovane Nascimento
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
e-mail: jerovane@alu.ufc.br
ORCID: 0000-0002-7361-4316

2º Adriell Gomes Marques
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
e-mail: adriell.gomes@alu.ufc.br
ORCID: 0000-0001-6274-1961

3º Katerine Chaves Trajano
Núcleo de Estudos e Pesquisas do Norte e Nordeste (NEPEN)
Fortaleza, Brasil
e-mail: katerine.trajano@nepen.org.br

4º Marcos Vinícius Candido de Figueiredo
Núcleo de Estudos e Pesquisas do Norte e Nordeste (NEPEN)
Fortaleza, Brasil
e-mail: marcos.figueiredo@nepen.org.br
ORCID: 0000-0002-7221-7255

5º Guilherme Lincoln de Lima Tavares
Programa de Pós-Graduação em Engenharia de Telecomunicações (PPGET)
Instituto Federal do Ceará (IFCE)
Fortaleza, Brasil
e-mail: guilherme.tavares08@aluno.ifce.edu.br

6º Carlos Henrique Lima Santos
Núcleo de Estudos e Pesquisas do Norte e Nordeste (NEPEN)
Fortaleza, Brasil
e-mail: carlos.santos@nepen.org.br

7ª Ana Sáskya Vaz de Araújo
Centro Diagnóstico Girassol
Prefeitura Municipal de Fortaleza (PMF)
Fortaleza, Brasil
e-mail: anasaskya@gmail.com

8º André Luiz Carneiro de Araújo
Departamento de Telemática (DETEL)
Instituto Federal do Ceará (IFCE)
Fortaleza, Brasil
e-mail: andreluiz@ifce.edu.br
ORCID: 0000-0002-3565-2039

***Abstract*—Public observatories make indicators easier to access, but publishing a chart does not ensure that it will be read correctly, traced to its source, or disclosed safely. This limitation is particularly important when the source material contains health and disability information. Ambiguities concerning the unit of analysis, counting rules, and territorial granularity may lead to misleading interpretations and, in small subsets, increase the risk of reidentification. This paper presents ORBITA-Health, a specialization of the ORBITA-SDG methodology for the auditable curation of indicators intended for public observatories. The method organizes the work into six stages, from framing the public problem to publishing and reviewing the indicator, under cross-cutting layers of governance, metadata, quality, privacy, and transparency. The proposal was developed as a design-science artifact and examined through two sets of institutional documents: a health-indicator curation matrix and an independent validation matrix from inclusive education. Without disclosing records, source-dataset volumes, or population findings, the evaluation assessed documentation completeness and the ability of review gates to locate semantic and technical discrepancies. All 20 candidate information products in the health case contained all 22 required fields, corresponding to 440/440 documented field instances; disclosure class and protection guidance were present for every product. In the second case, 103 analytical items were checked against independent calculations: 86 were validated, whereas 17 (16.5%) were withheld from that state, including 7 requiring attention and 10 remaining partial. The findings position ORBITA-Health as a shared audit trail with explicit stop conditions applied before analytical results become public indicators. Domain validation remains a separate requirement.**

***Index Terms*—public observatories, health indicators, persons with disabilities, data curation, auditability, privacy, metadata.**

I. Introduction

Public observatories sit between data production and public decision-making. They assemble time series, maps, tables, and indicators for audiences that do not necessarily share the same technical background, including public managers, health professionals, researchers, councils, civil-society organizations, and citizens. Recent reviews show that the literature on health dashboards remains fragmented and that usefulness is often inferred from access metrics or isolated reports, with limited evidence of how published information supports concrete decisions [1], [2]. The central problem therefore begins before interface design, with the choice of what should be counted and the documentation of the path from source data to public indicator.

This distinction matters whenever a number supports more than one interpretation. A count may refer to people, records, encounters, procedures, or administrative links. Likewise, a category may express a legal concept, a local rule, or merely the vocabulary inherited from a source system. Research on health-data quality emphasizes that completeness, conformance, plausibility, correctness, and currency depend on the intended

use rather than on the absence of missing fields alone [3], [4]. A numerically reproducible result may still be conceptually inadequate when its unit of analysis, period, formula, inclusion criteria, or limitations remain unclear.

Indicators concerning persons with disabilities introduce a second requirement: they must support equity without turning the visibility of disparities into the exposure of individuals. The World Health Organization identifies monitoring as part of the actions needed to address avoidable barriers in access to care and health outcomes [5]. At the same time, the Convention on the Rights of Persons with Disabilities and the Brazilian Inclusion Act connect public information, participation, and equal opportunity with duties concerning accessibility and the protection of rights [6], [7]. Sources that combine territory, health condition, diagnosis, medical reports, or other sensitive attributes therefore require minimization and disclosure rules to be defined before a dashboard is built.

The Brazilian General Data Protection Law grants enhanced protection to health data and establishes principles such as purpose limitation, adequacy, necessity, and security [8]. Removing names alone does not eliminate every risk. Combinations of attributes and infrequent categories may still make individuals recognizable, particularly when several filters are applied at once [9], [10]. For a public observatory, curation must therefore determine not only what is relevant, but also which level of aggregation is defensible, which details should remain internal, and when publication should be blocked.

This paper introduces **ORBITA-Health**, a specialization of the general ORBITA-SDG methodology. The name ORBITA retains the original Portuguese formulation, *Organização, Rastreabilidade, Base, Interpretação, Transparência e Ação*, which corresponds to organization, traceability, data foundations, interpretation, transparency, and action. ORBITA-SDG supplies the generic six-stage cycle for SDG-oriented public information. ORBITA-Health retains that cycle but adds operational requirements for health and disability data: mandatory declaration of the observational unit before counting, disclosure classification, content minimization, review of territorial/filter granularity and small-cell risk, and human domain validation with return gates when semantic, technical, or disclosure issues remain unresolved. These controls are co-located in a 22-field curation matrix so that the calculation rule and the publication decision remain part of the same audit trail. The main contribution is a health-specific publication boundary between analytical output and public communication, supported by an operational specification that makes the process inspectable and reusable. New statistical estimators and dashboard components are outside the scope of this work.

II. Related Work

The literature survey was organized around four areas that converge in the problem addressed here: public observatories and health dashboards; data quality and provenance; metadata for official statistics; and the protection of sensitive information. The survey was not intended as a standalone systematic review. Its purpose was to identify established frameworks and recent studies that could support design, validation, and disclosure requirements, later translated into operational fields and decision gates.

A. Public Dashboards and Evidence-Informed Action

Dashboards are useful when they reduce the effort needed to identify change and compare subsets, but their effectiveness does not follow from visualization alone. Reviews in clinical and public-health settings associate successful implementations with user involvement, appropriate indicator selection, clear narrative support, and post-deployment evaluation [1]. Schulze *et al.* found substantial heterogeneity among public dashboards in both data sources and visualization strategies [11]. In a multisite study, Fareed *et al.* further showed that perceived usefulness and ease of use depend on how a dashboard fits the practices of local stakeholders [12].

The concept of *actionability* extends this discussion. Beyond asking whether a dashboard can be understood, it asks whether the sources, methods, scope, and narrative allow users to connect a result with a decision or action. Stahlman *et al.* observed that many studies describe dashboard functions, whereas few rigorously evaluate their usefulness for decision-making [2]. ORBITA-Health responds to this limitation by requiring, before any chart is selected, a guiding question, an intended audience, a supported decision, and an explicit caution against misinterpretation.

B. Data Quality, Provenance, and Reproducibility

Data quality is multidimensional. The seminal definition by Wang and Strong shifts the discussion from an abstract notion of "correct data" toward fitness for the consumer and task [13]. In health applications, later frameworks converge on dimensions such as completeness, conformance, plausibility, and correctness, while also showing that terminology and tests vary across institutions [3]. ORBITA-Health records the applied rule, the evidence used, the observed outcome, and any unresolved issue for each quality check. No single check is treated as a quality seal.

Provenance complements quality by describing the origin, transformation, and movement of data. Recent biomedical reviews associate provenance with reproducibility and trust, but also report uneven adoption of models and tools [14]. The technical sheet proposed here preserves a compact version of that lineage: source, period, observational unit, calculation rule, filters, review date, and responsible unit. For studies based on routinely collected records, this transparency is also consistent with the RECORD recommendations for reporting data selection, linkage, and processing [15].

C. Official Statistics, Metadata, and Interoperability

The Generic Statistical Business Process Model organizes statistical production from the identification of information needs through product evaluation [16]. The UN NQAF, in turn, distributes quality responsibilities across the institutional environment, processes, and outputs rather than assigning them solely to the final dataset [17]. These frameworks provide the

process logic adopted by ORBITA-SDG, although they do not specify how a local team should translate that logic into the language of a public observatory.

SDMX addresses structured statistical exchange across domains [18]. The Data Quality Vocabulary supports the description of quality dimensions and measurements [19], while the FAIR principles emphasize that assets should be findable, accessible, interoperable, and reusable [20]. ORBITA-Health adapts these foundations to a smaller unit: each indicator is handled as a documented information product rather than as a visual element detached from its source.

D. Privacy, Disability, and Public Purpose

Aggregate publication requires a separate methodological review before public release. Anonymization models and the literature on privacy-preserving data publishing show that protection depends on attribute combinations, granularity, and the auxiliary information available to a potential observer [9], [10]. ORBITA-Health classifies outputs as aggregate, sensitive, restricted, or pending assessment. This classification does not replace legal or institutional review. It functions as an editorial gate that prevents an internal table from moving directly into a public environment.

The Sustainable Development Goals provide a layer of public purpose when the association with the indicator is explicitly justified. The official metadata repository describes the concepts, methods, and limitations of global SDG indicators [21]. ORBITA-Health uses this reference to record a primary SDG, a possible secondary SDG, and, most importantly, the specific relationship between the local indicator and the corresponding public agenda.

III. Materials and Methods

A. Study Design

This is an applied methodological study. ORBITA-Health was treated as an information-system artifact under the design-science perspective, in which knowledge is produced through the construction and evaluation of solutions to relevant problems [22], [23]. The artifact comprises a conceptual model, a curation matrix, and a set of validation gates. Its purpose is not to estimate prevalence or compare population groups, but to organize the conditions under which an analytical result may become a public indicator.

B. Application Artifacts

Two sets of documents were used to examine the method. The first was a curation matrix for information products derived from an institutional health source concerning persons with disabilities. The second contained validation items produced by comparing an analytical report with an independent recalculation in the field of inclusive education. The archived validation artifact describes the recalculation as an independent technical pass within the project workflow. It was not an external assessment. It does not preserve the recalculating analyst's identifier or a blinding field; consequently, this study does not claim assessor blinding. The cases place different demands on the method. The health case requires disclosure classification and safeguards for sensitive information, whereas the education case shows whether the gates can locate discrepancies in units, terminology, temporal cutoffs, and territorial aggregation.

TABLE I
Information blocks in the ORBITA-Health matrix

Block	Content and function
Identification	ID, domain, position, data status, and development status.
Semantic	Guiding question, title, supporting text, and interpretation note.
Technical	Source, update, minimum rule, unit, filters, and interaction.
Public purpose	Primary SDG, reference indicator, and secondary relationship when relevant.
Governance	Disclosure class, protection guidance, pending issue, and next action.

Only process metadata were used in this paper. No individual record, internal filename, indicator value, number of people in a source dataset, rare clinical category, or territorial result was transcribed. The quantities reported in the following sections refer to curation items and their validation states, not to the populations represented in the institutional sources.

C. Curation Matrix Schema

The matrix contains 22 fields arranged in five blocks. The fields are: (1) product ID, (2) domain, (3) position in the observatory, (4) data status, (5) development status, (6) guiding question, (7) public title, (8) supporting text, (9) interpretation note, (10) source, (11) update/periodicity, (12) minimum technical rule, (13) unit of analysis, (14) filters, (15) interaction, (16) primary SDG, (17) reference indicator, (18) secondary SDG relationship when relevant, (19) disclosure class, (20) protection guidance, (21) pending issue, and (22) next action. Grouping these fields prevents editorial description from being mistaken for a computational rule and allows curators, domain specialists, and developers to work on the same artifact. A data-free blank CSV, de-identified example, and structural checker accompany the source; the checker verifies the 22-field schema and required cells but does not automate semantic, disclosure, or publication decisions.

D. Evaluation Criteria

The evaluation considered four criteria. *Documentation completeness* refers to the presence of the required fields and does not imply approval of their content. *Traceability* examines whether the source, rule, and period allow the origin of an indicator to be reconstructed. *Discriminative capacity of the gates* concerns whether the process separates validated items from those requiring attention or remaining partial. *Publication safety* checks whether the artifact records a disclosure classification and corresponding guidance before a product is sent to the dashboard. Frequencies were calculated over curation items, and discrepancies were grouped by type without revealing institutional values.

TABLE II
DE-IDENTIFIED STRUCTURAL EXAMPLE OF ONE ORBITA-HEALTH INFORMATION PRODUCT. SOURCE-SPECIFIC CONTENT IS GENERALIZED OR SUPPRESSED; THE EXAMPLE DEMONSTRATES FIELD USE AND IS NOT A POPULATION RESULT.

Field	De-identified value	Field	De-identified value
1. Product ID	H-XX	**12. Minimum rule**	Aggregate calculation; rule retained internally
2. Domain	Health and disability	**13. Unit of analysis**	Declared before counting (e.g., unique persons)
3. Observatory position	Health indicator area	**14. Filters**	Period and approved territorial categories
4. Data status	Restricted institutional source available	**15. Interaction**	KPI/breakdown only after disclosure review
5. Development status	Under curation	**16. Primary SDG**	SDG 3
6. Guiding question	What public information should this product support?	**17. Reference indicator**	Contextual reference documented in matrix
7. Public title	Generalized health-service indicator	**18. Secondary SDG link**	SDG 10 when equity interpretation is justified
8. Supporting text	Aggregate description; no individual content	**19. Disclosure class**	Aggregate / pending final review
9. Interpretation note	Does not estimate prevalence or unmet need	**20. Protection guidance**	No identifiers, addresses, free text, or unsafe rare cells
10. Source	Municipal administrative source [suppressed]	**21. Pending issue**	Validate denominator/granularity with responsible unit
11. Update/periodicity	Periodic [detail suppressed]	**22. Next action**	Return to domain/disclosure validation before publication

The archived education-validation matrix used categorical review states rather than a single percentage tolerance because the 103 items mixed counts, percentages, textual categories, denominators, and aggregation rules. An item was recorded as *validated* when the analytical report and recalculation agreed and no unresolved discrepancy remained in the unit, denominator, cutoff, territorial rule, or terminology. *Requires attention* denoted a localized discrepancy for which the item remained inspectable but an explicit curator decision was still necessary. *Partial* denoted insufficient or ambiguous evidence to complete reconstruction or adjudication. The historical artifact did not store a universal numerical divergence threshold; therefore, no post-hoc percentage cutoff is attributed to the 103-item assessment. For future numerical checks, the operational rule is to require agreement after the declared rounding convention and to block the validated state while any unexplained numerical divergence remains.

IV. ORBITA-HEALTH METHODOLOGY

Fig. 1 separates the ORBITA-SDG base cycle from the controls introduced by ORBITA-Health. The specialization retains the six stages of the general cycle while making five health/disability requirements explicit: declaration of the observational unit, disclosure classification, content minimization, granularity/filter-risk review, and human domain validation with return gates. The stages describe the operational order of the method; the highlighted controls define the additional publication constraints introduced by the specialization.

In **Stage 1**, the process begins with the public problem and the people who will use the information. The team records the question the indicator should answer, the intended audience, the decision it may support, and the misleading interpretation that should be avoided. This framing prevents an available field from becoming a KPI merely because it is easy to count.

In **Stage 2**, the source, institutional responsibility, period, coverage, granularity, and observational unit are mapped. The distinction between people and records receives particular attention. When several occurrences are associated with the same person, the team must state whether the product measures unique individuals, events, or intensity of follow-up. The source is also examined for update feasibility and compatibility with the declared purpose.

In **Stage 3**, semantic curation compares the vocabulary inherited from the source with the concept that will be presented publicly. Acronyms, local classifications, and historical categories may remain in internal documentation, but the public title must state precisely what was counted. The relationship with the SDGs is documented under the same requirement: the link must be explained in the context of the indicator.

In **Stage 4**, technical validation examines completeness, consistency, duplication, plausibility, currency, and traceability. The result is then submitted to decision gates. When a discrepancy is identified, the item is returned to the responsible unit together with the formula, evidence, and unresolved issue. For sensitive sources, this stage also checks whether the planned filters may create small cells or attribute combinations that increase reidentification risk.

In **Stage 5**, the information product is designed. The team defines the title, short description, technical sheet, unit, calculation, analytical dimension, filters, and visualization. The question and the users' ability to interpret the result guide the choice among a KPI, chart, map, or table. The public version contains only what is needed for interpretation; operational details remain in the governance matrix.

In **Stage 6**, the indicator is prepared for publication, use, and review. Its status may progress from proposed to under curation, validated, published, or due for review. Any change in source, formula, or disclosure classification requires another passage through the gates. Publication therefore does not close the cycle. It creates a traceable version linked to the artifact that documents how the product was produced.

Fig. 2 details the ORBITA-Health return logic. Gate A asks whether the source and declared observational unit support the result; Gate B asks whether the interpretation and calculation are sufficiently documented to be reproduced; and Gate C asks whether the planned granularity and disclosure conditions are

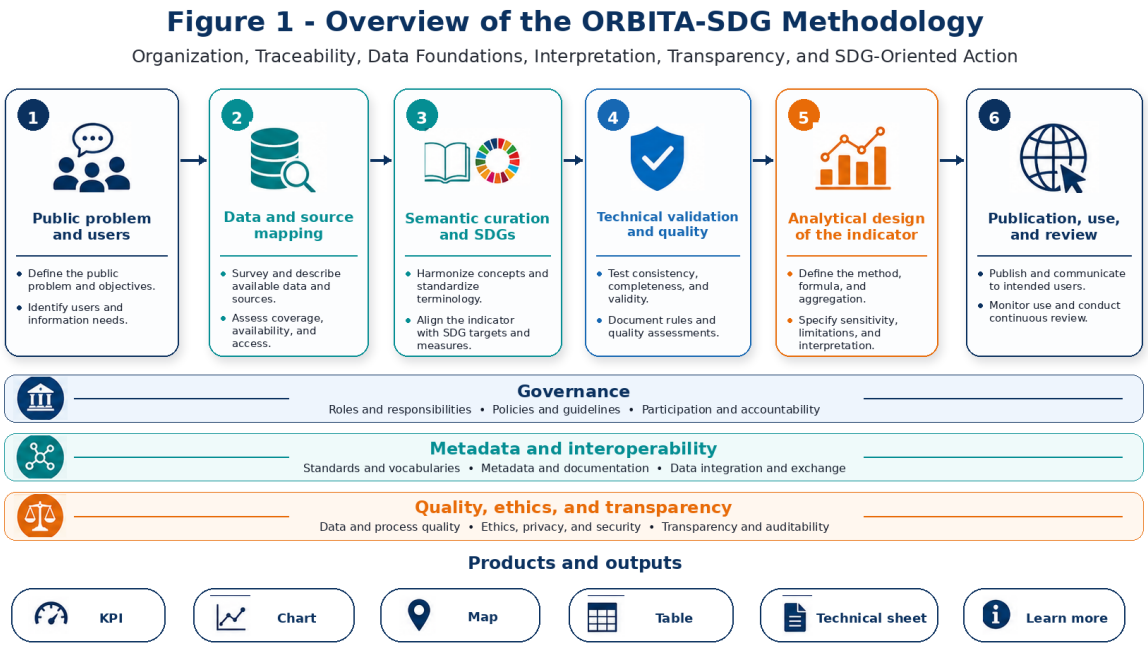


Fig. 1. ORBITA-SDG base cycle and the health/disability-specific controls introduced by ORBITA-Health.

fit for public release. A negative answer does not remove the item. It returns the product to the point at which the issue can be resolved while recording the evidence, status, pending issue, and next action in the matrix.

V. Results

A. Completeness of the Health Curation Matrix

The health-case matrix comprised 20 candidate information products. Across 22 required fields per product, all 440 expected field instances were populated. This included 80/80 governance-field instances (disclosure class, protection guidance, pending issue, and next action) and 20/20 declarations of the unit of analysis. These counts refer only to structural documentation completeness. Content correctness and approval remained subject to validation by the responsible unit.

Protection guidance was therefore present for every product before dashboard implementation and restricted public use to appropriately aggregated outputs, excluding individual records, identifiers, addresses, and free-text clinical content. The matrix does not prove that disclosure risk has been eliminated; it records that publication remains conditional on a documented review.

B. Performance of the Validation Gates

In the inclusive-education case, 103 analytical items were compared with the source document and the archived independent recalculation. The status rule followed the categorical evidence described in Section III-D: agreement without unresolved methodological discrepancy supported *validated*; a localized but inspectable discrepancy led to *requires attention*; and insufficient evidence to complete reconstruction led to *partial*. Because the source artifact did not define a common percentage tolerance across heterogeneous item types, no universal divergence threshold was applied retrospectively. Of the 103 items, 86 were validated (83.5%), whereas 17 (16.5%) were withheld from that state: 7 required attention (6.8%) and 10 remained partial (9.7%). Thus, the gate logic did not collapse all recalculable outputs into an approval state. These categories are not measures of model accuracy; they record documentary and computational sufficiency before an analytical result becomes a public indicator.

The 17 items outside the validated state were associated

TABLE III
Evaluation criteria, observed process evidence, and claim boundary

Criterion	Observed evidence	Boundary
Documentation	Health: 440/440 required field instances present.	Presence does not establish content correctness.
Traceability	Health: 20/20 products documented source, period/update, rule, and unit.	Source validity was not independently audited here.
Publication safety	Health: 20/20 products recorded disclosure class and protection guidance.	No reidentification attack or quantitative risk estimate.
Gate discrimination	Education: 86 validated; 7 attention; 10 partial; 17/103 (16.5%) withheld from validated.	Internal categorical cross-check, not blinded external validation.

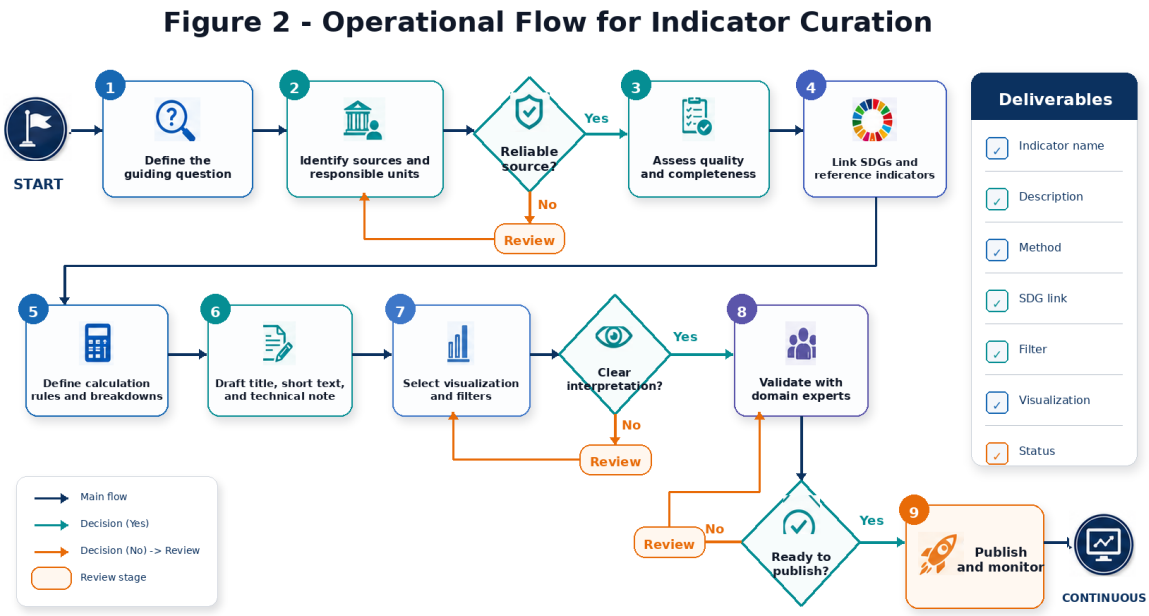


Fig. 2. ORBITA-Health operational flow. The review gates explicitly test source/unit adequacy, interpretability/reproducibility, and fitness for public disclosure.

with discrepancies that a purely syntactic spreadsheet check would be unlikely to resolve: confusion between records and unique persons, terminology covering different groups, an unspecified age cutoff, an undocumented territorial counting rule, and textual categories requiring standardization. The seven *attention* items retained enough evidence for inspection but required a localized decision before approval; the ten *partial* items lacked sufficient evidence to complete reconstruction or adjudication. The archived artifact does not provide a reliable frequency for each discrepancy subtype, so subtype counts are not reconstructed retrospectively. What separates the two non-validated states is therefore the completeness of the evidence needed to resolve the discrepancy, not a severity score or percentage-distance threshold.

C. Cross-Case Synthesis

The two cases evaluated different properties of the artifact. The health matrix examined whether semantic, technical, and disclosure information could be documented together. The education case examined whether the review logic could distinguish sufficient evidence from unresolved cases. The 440/440 documented field instances and the 17/103 items withheld from validation therefore provide complementary process evidence. The results show that the workflow can document information products and interrupt publication when evidence remains unresolved. They do not demonstrate improved decisions, usability, public comprehension, analytical accuracy, or health outcomes.

VI. Discussion

ORBITA-Health organizes decisions that are often scattered across spreadsheets, reports, and team discussions. It retains the six-stage ORBITA-SDG sequence and adds a health/disability publication boundary based on observational-unit declaration, disclosure classification, minimization, granularity/filter-risk review, and domain validation with explicit return states. Statistical frameworks and metadata vocabularies are comprehensive, but their direct adoption may be burdensome for local observatories. The matrix narrows that distance by linking each product to a question, a rule, a public purpose, a disclosure class, and an outstanding issue. It preserves the process logic of the GSBPM and UN NQAF [16], [17], while translating it to the editorial level of an indicator.

The cases also indicate that concise public information can coexist with detailed internal traceability. A public technical note may remain concise as long as the complete rule is preserved and linked to it. Provenance research similarly argues that origin and transformations should be retained to support reproducibility [14]. In the cases examined here, recording the unit of analysis and aggregation criteria was central to distinguishing a reproducible number from an indicator ready for use.

For health and disability information, the disclosure gate adds protection, but not an automatic guarantee. An “aggregate” classification should not be interpreted as permanent authorization: combined filters, small areas, and rare categories may change the risk. Final decisions still require institutional governance, a lawful basis, and specialist review. This caution is consistent with the Brazilian data-protection framework [8] and with research that treats anonymization as a contextual assessment rather than as the removal of identifiers alone [9], [10].

The study has limitations. It examined artifacts from two

institutional applications rather than a broad set of observatories. Completeness was measured through field presence, not field-content quality. The education validation was an internal technical cross-check; the archived artifact does not preserve assessor identity or blinding, and its status labels were categorical rather than based on a universal numerical tolerance. Usability, public comprehension, longitudinal maintenance, review time, inter-curator agreement, effects on decisions, and health outcomes were not assessed. The reported evidence therefore supports only process properties—documentation, traceability, review-state discrimination, and recorded disclosure guidance—and must not be interpreted as validation of the substantive indicators or of downstream organizational impact. These outcomes remain important, particularly because recent literature reports few rigorous evaluations of dashboard actionability [1], [2].

Future work should apply the methodology prospectively throughout a publication cycle. A broader assessment could compare rework before and after adoption of the matrix, measure inter-curator agreement, test users' understanding of the technical notes, and record how many products return at each gate. The supplied data-free schema and structural checker support reuse while leaving semantic, disclosure, and approval decisions to human governance.

VII. Conclusion

This paper presented ORBITA-Health, an ORBITA-SDG specialization for the auditable curation of health- and disability-related indicators in public observatories. The six-stage ORBITA-SDG cycle is retained, while ORBITA-Health adds explicit controls for observational unit, disclosure class, content minimization, granularity/filter risk, and human domain validation. Its 22-field operational matrix makes these decisions visible before an analytical result reaches a public dashboard.

Application to two sets of documents produced 440/440 required field instances in the health matrix and withheld 17/103 inclusive-education items from the validated state when evidence remained incomplete or discrepant. These results support process-level claims about documentation and gate behavior. They do not establish better decisions, usability, analytical accuracy, privacy guarantees, or health outcomes. Technical, legal, and domain validation remain necessary. ORBITA-Health provides a traceable route for documenting these validations, uncertainties, disclosure decisions, and return actions before public communication.

References

[1] D. Helminski, J. B. Sussman, P. N. Pfeiffer, A. N. Kokaly, A. Ranusch, A. D. Renji, L. J. Damschroder, Z. Landis-Lewis, and J. E. Kurlander, "Development, implementation, and evaluation methods for dashboards in health care: Scoping review," *JMIR Medical Informatics*, vol. 12, p. e59828, 2024.

[2] G. Stahlman, I. Yanovitzky, and M. Kim, "Design, application, and actionability of US public health data dashboards: Scoping review," *Journal of Medical Internet Research*, vol. 27, p. e65283, 2025.

[3] M. G. Kahn, T. J. Callahan, J. Barnard, *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *eGEMs*, vol. 4, no. 1, p. 1244, 2016.

[4] Y. P. Penev, T. R. Buchanan, M. M. Ruppert, M. Liu, R. Shekouhi, Z. Guan, J. Balch, T. Ozrazgat-Baslanti, B. Shickel, T. J. Loftus, and A. Bihorac, "Electronic health record data quality and performance assessments: Scoping review," *JMIR Medical Informatics*, vol. 12, p. e58130, 2024.

[5] World Health Organization, "Global report on health equity for persons with disabilities," tech. rep., World Health Organization, Geneva, Switzerland, 2022.

[6] United Nations, "Convention on the rights of persons with disabilities." United Nations General Assembly, 2006.

[7] Brasil, "Lei n. 13.146, de 6 de julho de 2015: Lei brasileira de inclusão da pessoa com deficiência." Diário Oficial da União, 2015.

[8] Brasil, "Lei n. 13.709, de 14 de agosto de 2018: Lei geral de proteção de dados pessoais." Diário Oficial da União, 2018.

[9] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[10] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, pp. 14:1–14:53, 2010.

[11] A. Schulze, F. Brand, J. Geppert, and G.-F. Böl, "Digital dashboards visualizing public health data: A systematic review," *Frontiers in Public Health*, vol. 11, p. 999958, 2023.

[12] N. Fareed, R. G. Olvera, Y. Wang, M. Hayes, E. L. Larimore, P. Balvanz, R. Langley, C. A. Noel, P. Rock, D. Redmond, J. Neufeld, S. Kosakowski, D. Harris, M. LaRochelle, T. R. Huerta, L. Glasgow, E. Oga, J. Villani, and E. Wu, "Lessons learned from developing dashboards to support decision-making for community opioid response by community stakeholders: Mixed methods and multisite study," *JMIR Human Factors*, vol. 11, p. e51525, 2024.

[13] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[14] K. Gierend, F. Krüger, S. Genehr, F. Hartmann, F. Siegel, D. Waltemath, T. Ganslandt, and A. A. Zeleke, "Provenance information for biomedical data and workflows: Scoping review," *Journal of Medical Internet Research*, vol. 26, p. e51297, 2024.

[15] E. I. Benchimol, L. Smeeth, A. Guttmann, *et al.*, "The RECORD statement: Reporting of studies conducted using observational routinely-collected health data," *PLoS Medicine*, vol. 12, no. 10, p. e1001885, 2015.

[16] United Nations Economic Commission for Europe, "Generic statistical business process model: Version 5.2," tech. rep., UNECE, 2025.

[17] United Nations Statistics Division, "United nations national quality assurance frameworks manual for official statistics," tech. rep., United Nations, New York, NY, USA, 2019.

[18] SDMX Consortium, "SDMX 3.1 technical specifications." Statistical Data and Metadata eXchange, 2025.

[19] R. Albertoni and A. Isaac, "Data on the web best practices: Data quality vocabulary," w3c note, World Wide Web Consortium, 2016.

[20] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016.

[21] United Nations Statistics Division, "SDG indicators metadata repository," 2025.

[22] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[23] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

PMPIF Dynamic Report: An Intersectoral Web Platform for Monitoring Early Childhood Indicators and Supporting Public Health Management

Adriell Gomes Marques
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
e-mail: adriell.gomes@alu.ufc.br
ORCID: 0000-0001-6274-1961

Cristiane Sales Leitão
Coordenadoria Especial da Primeira Infância
Prefeitura Municipal de Fortaleza (PMF)
Fortaleza, Brazil
e-mail: cristianesalesleitao@gmail.com
ORCID: 0000-0002-8412-0629

Lídia Andrade Lourinho
Coordenadoria Especial da Primeira Infância
Universidade Estadual do Ceará (UECE) / Prefeitura Municipal de Fortaleza (PMF)
Fortaleza, Brazil
e-mail: lidiandrade67@gmail.com
ORCID: 0000-0002-5883-9007

Camila Machado de Aquino
Coordenadoria Especial da Primeira Infância
Prefeitura Municipal de Fortaleza (PMF)
Fortaleza, Brazil
e-mail: camilamachadoce89@gmail.com
ORCID: 0000-0001-7108-5396

Maria Izabel Feitosa Accioly
Instituto de Planejamento de Fortaleza (IpplanFor)
Prefeitura Municipal de Fortaleza (PMF)
Fortaleza, Brazil
e-mail: mariaizabelaccioly@gmail.com
ORCID: 0000-0003-4831-6066

José Jerovane Nascimento
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brazil
e-mail: jerovane@alu.ufc.br
ORCID: 0000-0002-7361-4316

André Luiz Carneiro de Araújo
Departamento de Telemática (DETEL)
Instituto Federal do Ceará (IFCE)
Fortaleza, Brazil
e-mail: andreluiz@ifce.edu.br
ORCID: 0000-0002-3565-2039

***Abstract*—Municipal early-childhood plans combine indicators produced by several agencies, so provenance and reporting context matter as much as visualization. This paper examines the *Plano Municipal pela Primeira Infância de Fortaleza* (PMPIF) Dynamic Report, a web platform for registering policy actions, monitoring aggregate indicators, comparing reporting cycles, and producing management reports, with emphasis on maternal and child health. The evaluation combined source-code, architecture, data-model, access-control, audit, and workflow inspection. Five design objectives guided the analysis: context preservation, scoped editing, life-cycle control, provenance, and reuse of a common data core. The deployed revision snapshot contained 80 actions, 12 responsible agencies, three reporting semesters, 28 user accounts, and 90 action-linked table structures. A frozen repository database backup provided bounded retrospective verification: 47 actions present in both 2025 cycles preserved 786 aligned cell identifiers, 501 stored values differed between cycles, and 468 audit entries covered six entity classes. All audit entries retained operation, target entity, actor, and timestamp attribution; 399 of 401 update events also retained a non-null pre-change state. These findings support cycle isolation and populated provenance without claiming effects on usability, efficiency, or health outcomes.**

***Index Terms*—biomedical informatics, digital health, early childhood, maternal and child health, dashboards, health information systems, public management.**

I. Introduction

Early childhood is marked by biological, cognitive, emotional, and social changes whose effects may extend across the life course [1]. In Brazil, Law No. 13,257/2016 calls for coordinated action in health, education, social assistance, culture, leisure, mobility, protection, and urban planning for children up to six years of age [2]. That breadth is important for policy, but it creates a practical monitoring problem: actions, responsible institutions, targets, evidence, reporting periods, and indicators are produced through different administrative routines. International frameworks reach the same point from a developmental perspective, relating child well-being to health, nutrition, responsive caregiving, safety, early learning, and cross-sector coordination [3], [4].

Maternal and child health is one part of this broader arrangement. The Brazilian National Policy for Comprehensive Child Health Care (PNAISC) covers care from pregnancy onward, health promotion, breastfeeding, reduction of morbidity and mortality, attention to vulnerability, and child development [5]. Prenatal care, well-child follow-up, immunization, mental health, and oral health indicators can support management only when their institutional origin and reporting context remain visible.

Municipal monitoring is often assembled from spreadsheets, text documents, successive file versions, and manually prepared reports. Under those conditions, even a simple question may require additional work: which cycle is current, who is responsible for an item, whether a value was revised, and which record fed the last report. The dashboard therefore represents the presentation layer of a longer information process in which institutional ownership, revision history, and reporting context are defined upstream [6], [7].

The *PMPIF Dynamic Report* was developed to organize that process for the *Plano Municipal pela Primeira Infância de Fortaleza*. It does not store medical records or provide clinical care. Instead, it links municipal policy actions to responsible agencies, reporting cycles, indicators, structured tables, supporting evidence, editing rules, analytical views, and reports.

Against this background, we use PMPIF as a case through which to examine how an intersectoral plan can be turned into a governed monitoring workflow. This paper describes a data and workflow model that keeps policy actions, indicators, institutional responsibility, reporting time, and provenance connected. It also evaluates the implemented artifact against explicit technical objectives and discusses the design trade-offs and boundary conditions relevant to adoption in other municipalities, particularly where sectoral data remain heterogeneous and direct integration with every source system is not feasible.

II. Related Work

A. Dashboards and Public-Sector Action

Research on public-health dashboards converges on a point that is directly relevant to municipal monitoring: actionability depends on context, workflow, and data integration, not only on chart design. Reviews of public-health dashboards describe substantial variation in purpose, audience, granularity, interaction, infrastructure, and evaluation practice [6]–[8]. Maternal-health dashboards show the same heterogeneity in sources and presentation [9]. Work such as QualDash demonstrates how adaptable visualization can support healthcare quality-improvement tasks [10]. For the present case, the unresolved problem sits one step earlier: the system must know which institution owns an item, when it was reported, and whether the displayed record is still being edited or has already been consolidated.

B. Information Architecture and Data Quality

Digital-health platform guidance treats information systems as shared infrastructure that combines technology, governance, people, and reusable services [11]–[13]. Software-architecture studies likewise identify modularity, APIs, services, and separation of concerns as responses to fragmentation in health information systems [14]. CARA provides a recent example in which data modeling, APIs, and visualization are separated to support retrieval and benchmarking of general-practice data [15].

Architecture alone does not guarantee useful information. Health-information studies repeatedly identify completeness, consistency, timeliness, security, accessibility, relevance, and fitness for use as data-quality concerns [16]–[18]. In an intersectoral plan, consistency cannot mean forcing every sector into an identical indicator format. A more practical requirement is to preserve enough structure to recover the action, period, responsible institution, data type, and revision history associated with a value.

C. Human Factors, Governance, and Data Protection

Decision-support systems also need to reflect the tasks and authority of their users. Human-factors research recommends examining professional roles, interaction flows, task performance, and perceived usefulness rather than treating all users as interchangeable [19], [20]. In Brazil, the LGPD adds legal requirements for processing personal and sensitive data [21]. PMPIF works with institutional and aggregate monitoring records rather than individual patient data, but access control and traceability remain necessary because the application contains internal management information and editing privileges. API security guidance is relevant for the same reason: the server is the point at which authentication, authorization, and resource handling must be enforced [22].

Prior work has focused mainly on dashboards after data are available and on architectures centered on clinical or surveillance information. Less attention is given to the governance layer between heterogeneous institutional sources and the final analytical view: who owns a policy record, which reporting cycle it belongs to, whether it is still being revised, and which version was used for consolidation. Table I positions PMPIF around that gap. The comparison concerns design emphasis rather than performance, since the referenced systems were built for different purposes.

III. Materials and Methods

A. Study Object and Evaluation Scope

The object of study was the *PMPIF Dynamic Report* web application used by Fortaleza's early-childhood coordination to monitor the municipal plan. The artifact was examined as an intersectoral system, with particular attention to the actions assigned to the Municipal Health Department.

We used a technical case-study design guided by design science [23]. Construction of the artifact preceded this study, and the evaluation examined whether the implementation contains mechanisms that address the organizational problem for which it was built. Five objectives formed the evaluation frame: O1, preserve the link among action, institution, indicator, and reporting period; O2, restrict editing according to

Table I
POSITIONING OF PMPIF RELATIVE TO SYSTEMS AND EVIDENCE STREAMS CITED IN THIS STUDY.

System / evidence stream	Primary object	Reported design emphasis	Relation to the PMPIF case
Public-health dashboard literature [6]–[8]	Surveillance and population indicators	Actionability, audience, interface, analytics, infrastructure, and connection to work practices	Establishes dashboard requirements, while PMPIF additionally models the upstream ownership, editing cycle, and consolidation of policy records.
QualDash [10]	Healthcare quality measures	Adaptable generation of visualizations for quality-improvement tasks	Shares the concern with task-oriented visualization; intersectoral policy ownership and semester-based reporting are outside its main scope.
CARA [15]	General-practice patient datasets	Data architecture, APIs, retrieval, visualization, and benchmarking	Provides an architectural point of comparison for data/API/visualization separation; PMPIF operates on aggregate policy records rather than patient datasets.
PMPIF	Action–indicator–cycle records	Role-scoped entry, traceability, consolidation, comparison, and reporting from one data core	Adds an explicit municipal governance workflow in which heterogeneous sectoral information remains linked to responsibility and reporting period.

institutional responsibility; O3, distinguish active work from completed or consolidated content; O4, retain provenance of relevant changes; and O5, reuse the same persistent records for dashboards, comparisons, and reports. The study therefore evaluates the artifact at a technical and operational level, not its effect on clinical outcomes or population-health indicators.

No individual records concerning children, pregnant women, families, or professionals were analyzed, and no human-participant data were collected for this evaluation. Credentials, internal addresses, database dumps, and sensitive operational content were excluded from the manuscript.

B. Data Acquisition and Reporting Cycles

PMPIF receives institutional, aggregate monitoring information rather than person-level health data. The 2025.1 and 2025.2 cycles were organized as semester-specific records linked to the corresponding action, responsible agency, indicator or table, and reporting period. In routine use, authorized sector focal points enter or update values and supporting text through the action editor for the period in which they have permission to work. Each stored table value is identified by action, semester, table, row, and column, while detail fields and artifacts remain linked to the same action context. Completion status and relevant changes can then be followed without separating the value from the cycle in which it was reported.

The application is not presented as an automated integration layer for all upstream municipal systems. Sectoral evidence may originate in different administrative databases, spreadsheets, reports, or internal routines before it reaches PMPIF. The platform standardizes the monitoring context in which these heterogeneous inputs are registered. The present study inspected that acquisition workflow and its data structures; it did not independently audit the source system or calculation rule behind every institutional indicator.

C. Materials and Technologies

The material examined included source code, repository documentation, the relational model, API routes, access rules, user interfaces, editing workflows, report generators, and audit structures. The deployed architecture separates a React/Vite client, NestJS application services, a REST interface, and PostgreSQL persistence mediated by Prisma ORM. Nginx provides the web entry point and reverse proxy, and Docker Compose packages the application services for deployment.

D. Technical Analysis Procedure

The assessment began by mapping components, dependencies, and responsibilities across presentation, application, persistence, and deployment. We then traced the operational path from creation of a reporting period through action editing, value entry, completion tracking, dashboard visualization, temporal comparison, and report generation. Permission rules, cycle isolation, completion/consolidation handling, and the audit trail were inspected before objectives O1–O5 were mapped to observable implementation evidence. Interface presence alone was not treated as sufficient when the corresponding behavior could be checked in application services or persistent structures. Configuration counts were read from the administrative views in the same revision snapshot, while the action-level parsing example was taken from the SMS-01 dashboard for 2026.1.

An archived database backup dated June 10, 2026 was used only for bounded retrospective verification, not as the current deployment census. For actions represented in both closed 2025 cycles, we compared persisted cell identifiers and stored contents across semesters. Audit entries were summarized by entity type and timestamp, and field-level completeness was checked for operation, target entity, actor attribution, and before/after state in update events. Raw credentials, personal identifiers, and database contents are not reproduced.

ISO/IEC 25010:2023 informed the discussion of maintainability, reliability, security, and portability [24]. Human-factors and data-quality literature informed the interpretation of workflow and information concerns [16], [19]. Formal usability

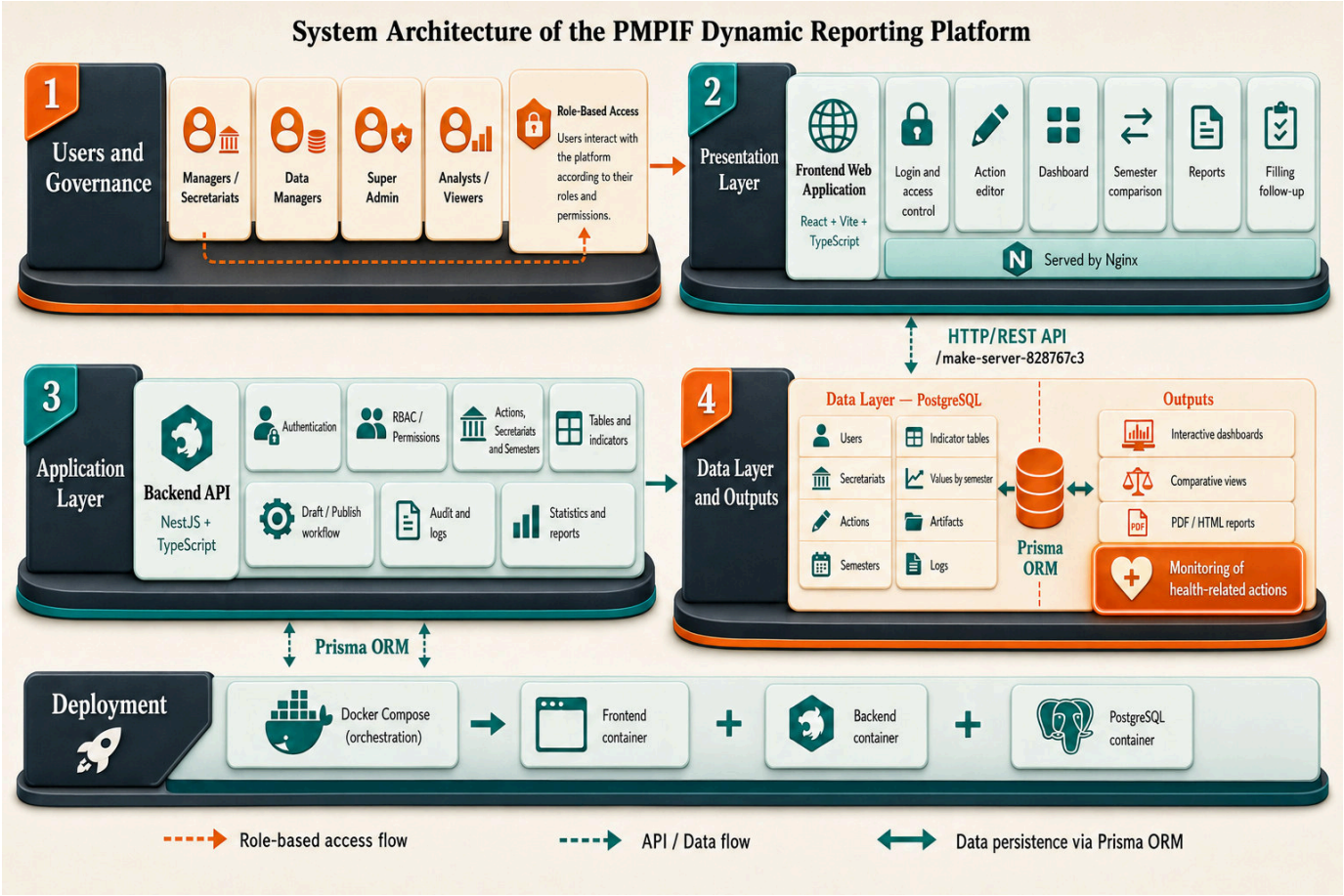


Figure 1. Functional architecture of *PMPIF Dynamic Report*. The layers separate governance and users, presentation, application services, persistence, and analytical outputs; the lower band represents the containerized deployment.

studies, penetration tests, controlled accessibility studies, and longitudinal outcome evaluations were not part of this stage.

IV. System Architecture and Operational Workflow

Figure 1 summarizes four responsibilities that the implementation keeps separate: governance, interaction, application rules, and persistence/analytical output. The separation is useful because a change in interface presentation should not silently change who is allowed to edit an institutional record, and a new report should not require a second copy of the monitored data.

A. Users, Governance, and Interaction

Managers, administrators, sector representatives, and data-entry users do not have the same authority over the plan. The access model therefore scopes users by role and, where applicable, by responsible agency or action. Focal points work on the records assigned to them, while broader management profiles supervise periods, consolidation, and cross-sector views. The frontend exposes the tasks available to each profile, but the decisive authorization checks are performed by application services rather than being left to interface state alone.

The presentation layer provides action editing, dashboards, semester comparison, report generation, and completion tracking. Communication with the application layer uses HTTP/REST interfaces. Keeping institutional rules behind those interfaces allows screens to evolve without duplicating authorization and validation logic in every visual component.

B. Application Rules and Persistence

NestJS services process authentication, permissions, user and agency administration, reporting periods, actions, tables, values, completion states, and audit events. Validation occurs before persistence and takes the current user, action scope, and reporting period into account. The data layer stores the corresponding entities in PostgreSQL, while Prisma maps the relational structures used by the services.

A value is not stored as an isolated number. Its identifiers preserve the action, semester, table, row, and column to which it belongs. The same persistent core is read by dashboards, semester comparisons, completion views, and report generators. This avoids a separate reporting database for the current scale of the platform and reduces the chance that a

manually copied report diverges from the record shown in the application.

C. Deployment and Security Maintenance

Frontend and backend are deployed as separate containers behind Nginx, with persistence handled independently. This arrangement supports reproducible deployment and makes failures easier to localize. The current code also includes password hashing with scrypt, server-side role checks, bounded sessions, HttpOnly browser cookies, session revocation/rotation, and audit records for relevant operations. These mechanisms were verified by code inspection; they were not subjected to a formal penetration test.

Security cannot be treated as fixed once the system is deployed. The JavaScript dependencies, container images, reverse proxy, ORM, and database schema all require maintenance as vulnerabilities and compatibility changes appear. Production updates should therefore include dependency review, controlled schema migration, tested backup/restore procedures, secret rotation, TLS and cross-origin configuration, API rate controls where appropriate, and regression tests for authorization rules. Centralizing those controls in the backend simplifies maintenance, but it also means that an unsafe service-layer change can affect several modules at once; staging, rollback, and audit review are consequently part of the maintenance risk model [22], [24].

D. Design Challenges and Decisions

The most consequential decision concerned heterogeneous upstream data. A direct connector for every sectoral system would automate more ingestion, but it would also bind PMPIF to external schemas, release schedules, and access arrangements that differ across agencies. The default design therefore uses a common policy-centered model and accepts controlled human-mediated entry when no stable integration path exists. This keeps the action, responsible agency, cycle, and data type consistent even when the source evidence arrives as a database extract, spreadsheet, report, or institutional note. Where stable APIs become available, automated ingestion can be added without changing those core entities.

Other decisions followed from distributed ownership. Authorization is enforced in application services rather than only in the client, which makes the same scope rules apply across screens but requires continued administrative maintenance of roles and assignments. Values remain cycle-specific and relevant changes are auditable, avoiding simple overwrite semantics at the cost of a more explicit life cycle. Dashboards, comparisons, completion views, and reports read the same persistent records, reducing duplicate consolidation while making output quality dependent on the structure registered for each action. Finally, separating client, server, and persistence helps isolate updates, but it does not remove the need for dependency patching, schema migration, tested backups, and security regression checks.

Table II
ARTIFACT EVIDENCE MAPPED TO THE FIVE DESIGN OBJECTIVES.

Obj.	Observed implementation evidence
O1	Values retain action, semester, table, row, and column context; actions remain linked to responsible agencies and indicator metadata.
O2	Backend services apply role and action/agency scope before protected edits are persisted.
O3	Across 47 actions paired in 2025.1/2025.2, 786 cell identifiers aligned and every action retained at least one changed value.
O4	468 audit entries spanned six entity classes; every entry retained operation, target, actor, and timestamp attribution, and 399/401 updates retained non-null prior state.
O5	Dashboards, comparisons, completion views, and reports reuse the same persistent action-cycle records.

E. Platform Workflow and Life Cycle

Figure 2 shows the dependencies among the main stages. E1–E5 prepare the environment: an administrator account is established, the reporting semester is configured, responsible agencies and actions are registered, and focal-point users receive their scope. Each action can carry objectives, strategy, targets, indicators, structured tables, and supporting content.

Routine monitoring begins at E6. Authorized users enter semester-specific values and update action details. E7 exposes completion and pending states while collection is underway. The same records feed E8, where the selected cycle can be visualized or compared with another, and E9, where reports are assembled for technical monitoring and accountability.

The workflow is iterative rather than strictly linear. A change in an action structure may alter its dashboard, and a value entered during E6 can affect completion tracking and report content immediately. The figure therefore represents dependency, not a one-way approval pipeline.

V. RESULTS

Artifact-level inspection found traceable implementation evidence for each of the five design objectives. Table II summarizes that mapping. The finding concerns the presence and connection of mechanisms in the inspected artifact; it does not establish usability, security effectiveness, data accuracy at the upstream source, or policy impact.

At the same revision snapshot, the platform registered 80 actions across 12 responsible departments or agencies, three reporting semesters, 28 user accounts, and 90 action-linked table structures; 11 actions were assigned to the Municipal Health Department. These figures describe the deployed artifact and are not estimates of policy effectiveness.

Figure 3 complements these counts by illustrating the intersectoral positioning of the health axis within the broader municipal plan. The visual summarizes ten original health-action themes represented in the platform; the current deployment also includes the neurodivergence care-pathway action reported below. It is therefore used as a conceptual overview rather than as the current deployment census.

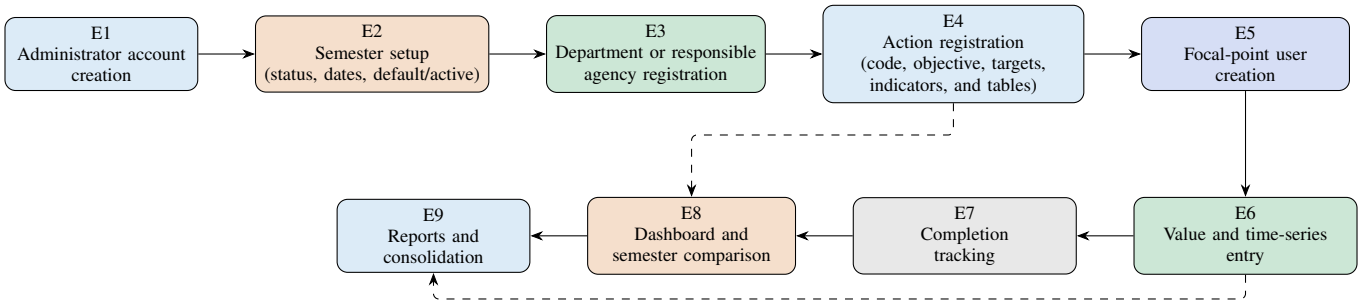


Figure 2. Operational workflow of *PMPIF Dynamic Report*, from initial setup through periodic data entry, monitoring, comparison, and consolidation.

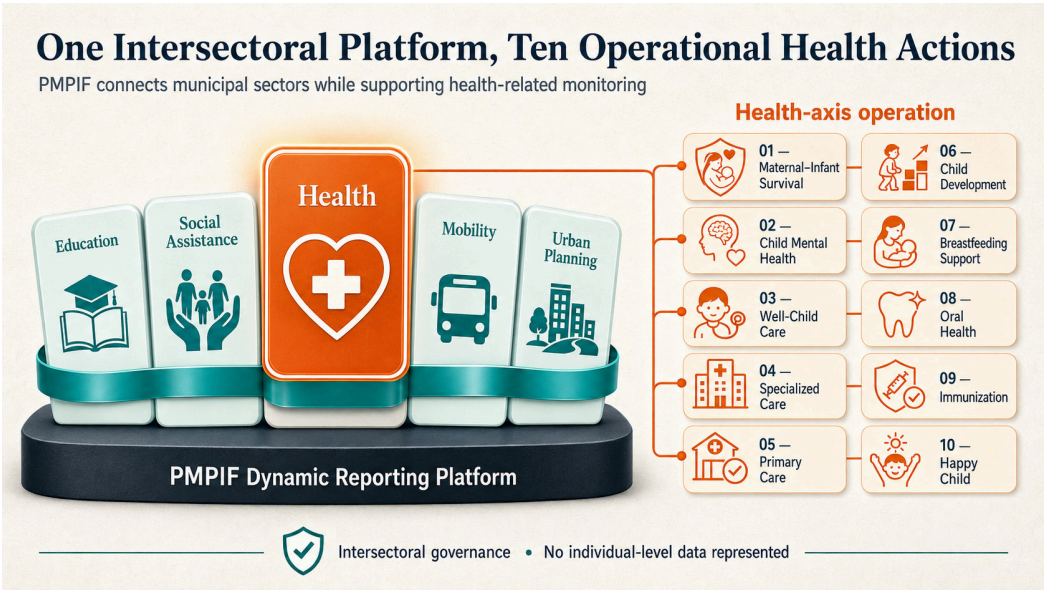


Figure 3. Intersectoral positioning of the health axis in *PMPIF Dynamic Report*. The illustration summarizes ten original health-action themes supported by the platform; the current deployment contains an additional Municipal Health Department action for the early-childhood neurodivergence care pathway. No individual-level data are represented.

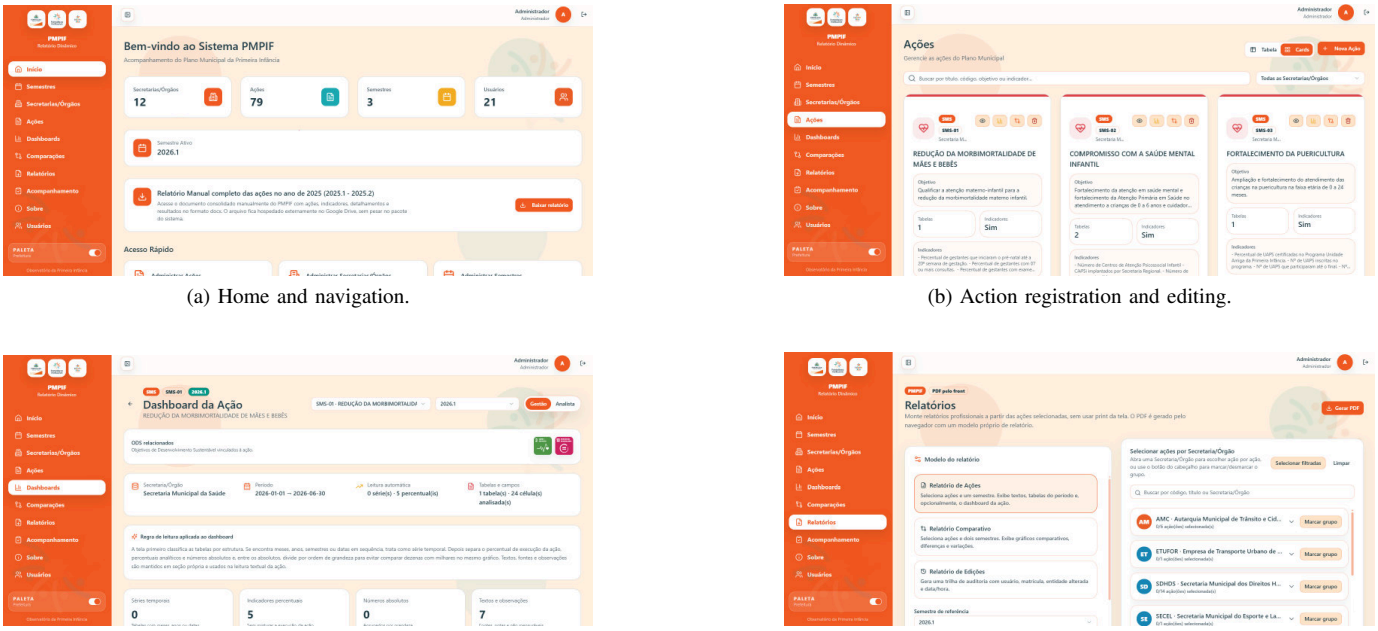
The two 2025 cycles (2025.1 and 2025.2) had already been consolidated, while 2026.1 was active for collection. In the archived snapshot, 47 actions had value records in both closed 2025 cycles. Their 786 cell identifiers were identical across semesters; 501 stored values differed, and every paired action changed in at least one position. This verifies structural reuse with cycle-specific content, not indicator improvement.

The analytical layer also produces action-specific structural counts. In the 2026.1 dashboard for action SMS-01, for example, one registered table yielded 12 analyzed cells, which the current parser classified as five percentage entries and seven textual or observational entries. This example shows that the dashboard is generated from the registered table structure rather than from a separately prepared visualization file. It is not used as a data-quality score or health-outcome measure.

Figure 4 provides complementary implementation context through representative screens for navigation, action editing, dashboard visualization, and report generation. These screens show how the action–cycle data model is exposed to users in routine operation; they are illustrative interfaces, not additional evaluation endpoints.

The same frozen snapshot contained 468 audit entries spanning May 20 to June 7, 2026: 181 concerned actions, 180 values, 59 authentication events, 37 user-administration changes, six semester events, and five agency events. Every entry contained an operation, target entity, actor identifier/name/e-mail, and timestamp. Of 401 update events, all stored a post-change state and 399 also retained a non-null pre-change state; the other two explicitly stored a null pre-change state. The bounded check verifies that provenance fields were populated and attributable across several operation classes. It does not provide an estimate of workload or audit quality. Values remain cycle-specific, while permissions restrict the actions or agencies available to scoped users.

The health component remains embedded in the intersectoral plan rather than being isolated as a separate application. At the current snapshot, 11 of the 80 actions were assigned to the Municipal Health Department, covering maternal-infant care, mental health, well-child care, specialized and primary care, child development, breastfeeding support, oral health,



(a) Home and navigation.

(b) Action registration and editing.

(c) Dashboard and analytical synthesis.

(d) Report configuration and generation.

Figure 4. Representative screens of the Portuguese-language deployment, highlighting navigation, action editing, analytical visualization, and report generation. The panels complement the architectural and workflow views by showing how the monitored records are presented in the operational interface.

immunization, child well-being, and the care pathway for neurodivergence in early childhood. Other sectors retain their own responsibilities but share the same cycle and monitoring model.

The retrospective checks establish structural continuity across two closed cycles and show that the audit trail was populated during a bounded interval. They do not provide cycle completeness, a long-term audit-event rate, or a measured before/after reduction in consolidation time; those outcomes require prospective observation.

VI. Discussion

PMPIF illustrates how a municipal plan can be represented as a governed information workflow. Each policy action remains linked to its responsible institution, reporting period, indicators, structured values, supporting material, workflow state, and revision history. Dashboards consume these records as analytical outputs while the underlying information remains organized around the policy action.

That choice also defines what interoperability means in this study. PMPIF does not claim automatic exchange with every database maintained by participating agencies. It handles semantic and organizational heterogeneity through a common monitoring model. Stable APIs can reduce manual entry for selected sources, but requiring them everywhere would couple the plan to external schemas and institutional release cycles. The current design therefore keeps responsibility, period, and provenance stable while allowing ingestion to remain partly human-mediated where necessary.

The resulting design pattern keeps policy identity, ownership, reporting time, authorization scope, revision history, and analytical reuse connected through a common persistent core. Table I places these concerns alongside prior work on dashboard actionability and layered information architectures.

Transferability also has limits. The approach is best suited to periodic, aggregate monitoring in which several agencies remain responsible for their own source evidence while a coordinating body needs a common reporting cycle. It should not be read as a replacement for real-time clinical surveillance, patient-level interoperability, or transaction systems. Where stable machine-readable interfaces and event-level data already exist, more automated ingestion may be preferable, with the policy-centered model serving as a governance and consolidation layer rather than as the upstream source system.

The architecture also creates maintenance responsibilities. Separation of client, server, and persistence makes individual components easier to replace or update, but security depends on keeping those boundaries correctly configured over time. Dependency patching, database migrations, backup validation, session and secret handling, and authorization regression tests are therefore part of system sustainability, not peripheral operational tasks. The current implementation contains several safeguards, yet the absence of a formal security audit prevents claims about resistance to attack.

The evaluation covers technical implementation and retrospective artifact evidence. It includes code inspection, workflow tracing, deployed configuration counts, a bounded cross-cycle persistence check, audit-log characterization, and an action-level parser example. Usability testing, penetration testing, accessibility assessment, and before/after efficiency measurement were outside its scope. Manager interactions during development were not collected under a predefined quali-

tative protocol and are therefore not reported as user-study data. Completeness, consolidation time, sustained audit-event rates, and effects on decision-making still require prospective measurement. The retrospective checks are limited to what the frozen snapshot can answer directly: structural alignment across cycles and populated provenance.

Future evaluation should define a fixed institutional observation period and measure completion by cycle, turnaround time for consolidation, rework, user task performance, and perceived usefulness. A security assessment and a documented maintenance/update procedure are also priorities. These steps would complement the artifact-level evidence reported here with human, organizational, and longitudinal outcomes [19].

VII. Conclusion

This study examined the *PMPIF Dynamic Report* as an implemented municipal information artifact and mapped observable implementation evidence to the five technical objectives used in the evaluation. The case shows how actions, responsible institutions, reporting cycles, structured values, permissions, provenance, and analytical outputs can remain connected in a single monitoring workflow.

The resulting policy-centered model preserves ownership and reporting context while keeping authorization, revision history, and analytical reuse tied to the same persistent records. The present evidence is still artifact-level. Prospective work should measure cycle completeness and consolidation time, evaluate usability with institutional users, and test security and maintainability under routine operation before organizational or public-health effects are claimed.

References

[1] M. M. Black, S. P. Walker, L. C. H. Fernald, C. T. Andersen, A. M. DiGirolamo, C. Lu, D. C. McCoy, G. Fink, Y. R. Shawar, J. Shiffman, A. E. Devercelli, Q. T. Wodon, E. Vargas-Barón, and S. Grantham-McGregor, "Early childhood development coming of age: Science through the life course," *The Lancet*, vol. 389, no. 10064, pp. 77–90, 2017.

[2] Brasil, "Lei n. 13.257, de 8 de março de 2016: Marco legal da primeira infância," 2016. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/l13257.htm

[3] P. R. Britto, S. J. Lye, K. Proulx, A. K. Yousafzai, S. G. Matthews, T. Vaivada, R. Perez-Escamilla, N. Rao, P. Ip, L. C. H. Fernald, H. MacMillan, M. Hanson, T. D. Wachs, H. Yao, H. Yoshikawa, A. Cerezo, J. F. Leckman, and Z. A. Bhutta, "Nurturing care: Promoting early childhood development," *The Lancet*, vol. 389, no. 10064, pp. 91–102, 2017.

[4] World Health Organization, United Nations Children's Fund, and World Bank Group, *Nurturing Care for Early Childhood Development: A Framework for Helping Children Survive and Thrive to Transform Health and Human Potential*. Geneva: World Health Organization, 2018. [Online]. Available: https://iris.who.int/handle/10665/272603

[5] Brasil. Ministério da Saúde, "Portaria n. 1.130, de 5 de agosto de 2015: Institui a política nacional de atenção integral à saúde da criança (PNAISC) no âmbito do sistema único de saúde (SUS)," 2015. [Online]. Available: https://bvsms.saude.gov.br/bvs/saudelegis/gm/2015/prt1130_05_08_2015.html

[6] G. Stahlman, I. Yanovitzky, and M. Kim, "Design, application, and actionability of US public health data dashboards: Scoping review," *Journal of Medical Internet Research*, vol. 27, p. e65283, 2025.

[7] R. Rabiei, P. Bastani, H. Ahmadi, S. Dehghan, and S. Almasi, "Developing public health surveillance dashboards: A scoping review on the design principles," *BMC Public Health*, vol. 24, p. 392, 2024.

[8] A. Schulze, F. Brand, J. Geppert, and G.-F. Böl, "Digital dashboards visualizing public health data: A systematic review," *Frontiers in Public Health*, vol. 11, p. 999958, 2023.

[9] J. A. Callaghan-Koru, P. Newman Chargois, T. Tiwari, C. C. Brown, W. Greenfield, and G. Koru, "Public maternal health dashboards in the United States: Descriptive assessment," *Journal of Medical Internet Research*, vol. 26, p. e56804, 2024.

[10] M. Elshehaly, R. Randell, M. Brehmer, L. McVey, N. Alvarado, C. P. Gale, and R. A. Ruddle, "QualDash: Adaptable generation of visualisation dashboards for healthcare quality improvement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 689–699, 2021.

[11] World Health Organization, *Global Strategy on Digital Health 2020–2027*. Geneva: World Health Organization, 2025. [Online]. Available: https://www.who.int/publications/b/81640

[12] World Health Organization and International Telecommunication Union, *Digital Health Platform Handbook: Building a Digital Information Infrastructure (Infostructure) for Health*. Geneva: World Health Organization and International Telecommunication Union, 2020. [Online]. Available: https://www.who.int/publications/i/item/9789240013728

[13] J. A. Magnuson and B. E. Dixon, Eds., *Public Health Informatics and Information Systems*, 3rd ed. Cham: Springer, 2020.

[14] R. Casanova, F. A. Villa-Garzón, and J. W. Branch-Bedoya, "Architectural patterns for health information systems: A systematic review," *Frontiers in Digital Health*, vol. 7, p. 1694839, 2025.

[15] N. Garzón-Orjuela, A. Garcia Pereira, H. Vornhagen, K. Stasiewicz, S. Parveen, D. Amin, L. Porwol, M. d'Aquin, C. Collins, F. Stanley, M. O'Callaghan, and A. Vellinga, "Design and architecture of the CARA infrastructure for visualising and benchmarking patient data from general practice," *BMJ Health & Care Informatics*, vol. 31, no. 1, p. e101059, 2024.

[16] H. Ghalavand, S. Shirshahi, A. Rahimi, Z. Zarrinabadi, and F. Amani, "Common data quality elements for health information systems: A systematic review," *BMC Medical Informatics and Decision Making*, vol. 24, p. 243, 2024.

[17] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[18] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T. C. Ong, P. Ryan, N. Shang, N. G. Weiskopf, C. Weng, M. N. Zozus, and L. Schilling, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, vol. 4, no. 1, p. 1244, 2016.

[19] H. M. Vasquez, E. Pianarosa, R. Sirbu, L. M. Diemert, H. Cunningham, V. Harish, B. Donmez, and L. C. Rosella, "Human factors methods in the design of digital decision support systems for population health: A scoping review," *BMC Public Health*, vol. 24, p. 2458, 2024.

[20] M. Zhuang, D. Concannon, and E. Manley, "A framework for evaluating dashboards in healthcare," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 4, pp. 1715–1731, 2022.

[21] Brasil, "Lei n. 13.709, de 14 de agosto de 2018: Lei geral de proteção de dados pessoais (LGPD)," 2018. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[22] OWASP Foundation, "OWASP API Security Top 10 – 2023," 2023. [Online]. Available: https://owasp.org/API-Security/editions/2023/en/0x11-t10/

[23] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[24] ISO/IEC, "ISO/IEC 25010:2023 systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – product quality model," Geneva, 2023. [Online]. Available: https://www.iso.org/standard/78176.html

Prototyping a RAG-based Conversational Agent for Pre- and Post-Operative Cardiac Patient Guidance

COSTA, Gabriela Toledo da
Informática Biomédica
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)
Porto Alegre, Brazil
gabrielacosta.ux@gmail.com

MARTINS, Tamires
Programa de Pós-graduação Mestrado Profissional em Enfermagem (PPGENF)
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)
Porto Alegre, Brazil
tamires.martins@ufcspa.edu.br

SANTANA DA SILVA, Filipe
Programa de Pós-graduação Mestrado Profissional em Enfermagem (PPGENF)
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)
Porto Alegre, Brazil
filipe@ufcspa.edu.br

***Abstract*—The limited availability of continuous healthcare personnel for patient guidance underscores the need for robust Digital Health solutions. Large Language Models (LLMs) combined with Retrieval-Augmented Generation (RAG) offer a promising approach to mitigate hallucinations common in generative models, providing a structured framework for clinical decision support systems. This study developed a RAG-based Conversational Agent (CA) prototype to provide perioperative (pre- and post-operative) guidance for patients undergoing cardiac surgery. The system was implemented using a quantized Llama-3-8B-Instruct model and the LangChain library, and evaluated through a knowledge base ablation study across three configurations: pre-selected scientific articles (KB1), clinical FAQs curated by a specialist nurse (KB2), and a hybrid combination of both layers (KB3). Safety parameters, including a low temperature setting (0.2) and strict prompt constraints, were enforced to minimize the risk of speculative generation. Evaluation relied on semantic cosine similarity metrics computed via high-dimensional text embeddings, alongside qualitative analysis of safety guardrail activation against predefined protocols. The hybrid configuration (KB3) achieved the highest mean similarity score (0.64 ± 0.12), suggesting potential synergistic benefits compared to isolated baselines, although formal statistical significance testing was not performed in this preliminary study. Conversely, queries drifting out of scope or demanding individualized medical assessment triggered predefined refusal messages, resulting in lower semantic similarity scores (mean 0.38). This divergence reflects the system's adherence to safety protocols, prioritizing human redirection over speculative generation. The prototype demonstrates technical viability as a supportive tool for health education in controlled settings, though further clinical validation is required before deployment.**

***Index Terms*—Retrieval-Augmented Generation, Large Language Models, Cardiac Surgery, Patient Guidance, Semantic Evaluation, AI Safety.**

I. Introduction

The evolution of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies has revolutionized digital interactions [1], [2]. These tools have evolved significantly, being applied in several areas, including healthcare, where they can offer emotional support and convey information instantly and in a personalized manner [2].

In highly complex healthcare scenarios, such as heart surgery, patients and their family members often experience high levels of stress and concern, making individualized and continuous monitoring essential [3]. In this context, pre- and post-operative guidance plays a fundamental role, with nursing professionals currently serving as the main agents for essential care instruction [4]. A failure in communication or the absence of effective support can compromise adherence to self-care, potentially affecting the effectiveness of the operation and recovery [5].

In this context, considering the limited time and continuous availability of healthcare professionals dedicated exclusively to guidance, Conversational Agents (CAs) emerge as a potential technological solution in digital health. They offer the opportunity to provide standardized information and clarify doubts about pre- and post-cardiac operative processes [6], potentially optimizing the use of clinical team time [7].

The constant evolution of Large Language Models (LLMs) represents a significant advance in AI, improving the ability of CAs to simulate human-like conversation [8]. While LLMs show potential to enhance accessibility and assist in decision support [9], their unrestricted application in healthcare remains challenging due to the risk of generating erroneous, out-of-context, or non-evidence-based content. These factors are critical barriers to the safe provision of health information [10].

To overcome these limitations, Retrieval-Augmented Generation (RAG) can be employed to restrict the LLM's knowledge base (KB) to a specific set of pre-selected documents [11], [12]. This approach aims to increase the accuracy and contextual relevance of the conversational agent's output by grounding responses in verified sources [12].

Considering the need to provide accurate and evidence-based information to cardiac patients, the objective of this work is to present a CA prototype using the RAG approach. This study serves as an exploratory evaluation of a hybrid RAG architecture combining evidence from specialized literature and local operational knowledge. The expectation is that current solution may help reduce patient and family concerns and ease the educational burden on nursing professionals. It should be seen as a supplementary support tool rather than a replacement for clinical care.

II. METHODS

The present study is characterized as a technological development work aimed at creating and simulating a CA, based on pre-selected scientific evidence [2], to provide perioperative guidance to cardiac surgery patients. The development was guided by the following steps:

1) Identify standard clinical guidelines and patient frequently asked questions (FAQs);
2) Study and apply the RAG approach for CAs;
3) Prototype implementation based on the bibliography; and
4) Semantic and qualitative validation tests using FAQ + scientific literature curated by a specialist cardiology nurse.

Supplementary material is provided at the end of this paper.

A. Knowledge Base Definition

The main KB used to anchor the RAG pipeline was composed of two distinct operational layers:

a) A selection of 10 scientific articles pre-selected (Portuguese, English and Spanish) were retrieved during the development of a systematic literature review, as detailed in Table I; and

b) Two FAQ datasets (Portuguese), capturing local workflows and the health service's operational routines. Examples of the mapped questions and their respective gold-standard clinical were translated (from Portuguese into English) and are provided in Table II.

The KB layers described above were developed with the assistance of a nurse specialist who has over four years of experience in cardiologic peri-operative care. The nurse specialist also worked at a reference cardiologic surgery center, where protocols for patient guidance developed by a multiprofessional team were reused.

B. RAG Architecture and Implementation

The implementation of the CA was based on the RAG approach [4], aiming to ensure cohesive, precise, and strictly grounded answers within the text reviewed by professionals experienced in cardiac perioperative guidance [3]. The prototype was developed in Google Colab, utilizing an NVIDIA T4 GPU instance. No specific language translation mechanism was employed in current work, i.e. translation was handled by the LLM as is. To optimize memory usage and inference speed, the LLM was loaded with 4-bit quantization. The technologies used (Table III) and the solution design (Fig. 1) are summarized below.

C. RAG Chain Configuration and Prompt Engineering

The LLM Llama-3-8B-Instruct (4-bit NF4 Quantization) was configured with a temperature parameter of 0.2. This parameter was configured empirically after qualitative validation cycles with the clinical specialist (testing values ranging from 0.0 to 0.4). This task was performed to ensure that the response generation remained limited to the retrieved facts, avoiding

TABLE I
REFERENCES COMPOSING THE KNOWLEDGE BASE

Author	Title	Journal	Year
Kazitani BS et al.	Cardiac anxiety in the peri-operative period of patients undergoing cardiac surgical procedures: an observational study	Revista Brasileira de Enfermagem (REBEN)	2023
Barcellos SR et al.	Construction and validation of an educational booklet for patients in the postoperative period of cardiac surgery: a methodological study	Revista Brasileira de Enfermagem (REBEN)	2023
Reisdorfer AP et al.	Nursing care for patients in the post-operative period of cardiac surgery in the Intensive Care Unit	Revista Brasileira de Enfermagem (REBEN)	2021
Morelato RDC et al.	Risk of mediastinitis in the postoperative period of myocardial revascularization surgery	Rev. Pesqui. Cuid. Fundam.	2022
Martins LM et al.	Preoperative symptoms of anxiety, depression, and cardiac anxiety according to the type of cardiac surgery	Rev. Pesqui. Cuid. Fundam.	2022
Martins LM et al.	Preoperative symptoms of anxiety, depression, and cardiac anxiety according to the type of cardiac surgery	REME	2022
Gomes et al.	Cardiac Surgery: Nursing care in the postoperative	Braz. J. Implantol. Health Sci.	2024
Sarmento SD et al.	Non-pharmacological therapies in the relief of cardiac surgery postoperative pain: a scoping review	Online Braz. J. Nurs.	2021
Jepma P et al.	The nurse-coordinated cardiac care bridge transitional care programme: a randomised clinical trial	Amsterdam Univ. of App. Sciences	2021
Araújo NM et al.	Audiovisual resource in education for the preoperative period of cardiac surgery: a scoping review	Texto & Contexto Enfermagem	2022

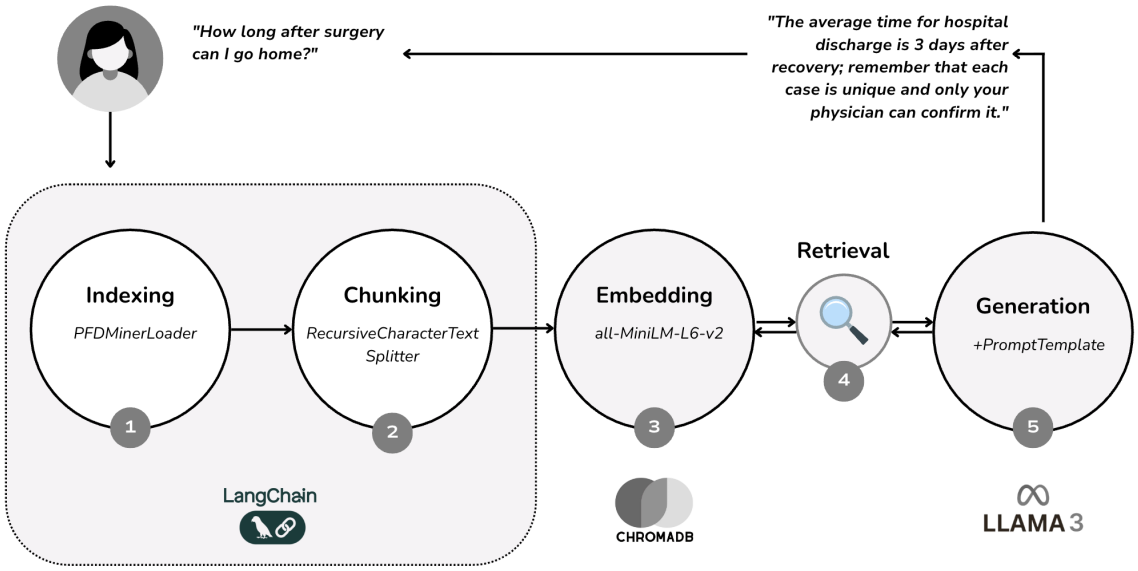


Fig. 1. Solution design of the RAG-based CA for pre- and post-operative cardiac guidance.

TABLE II
OVERVIEW OF FAQ QUERY/ANSWER PAIRS.

Question	Expected Answer	Question	Expected Answer
Should I be hospitalized before my cardiac surgery?	Yes. Prior admission is required for physical preparation, pre-operative exams, and safety evaluation.	What devices will I use in the immediate post-operative period?	Intubation (for about 6h), chest drains, central venous catheter in the neck, peripheral venous access, and indwelling urinary catheter.
Do I need to fast before admission?	No. The fasting protocol (usually 8 hours) is controlled by the nursing staff after admission.	How long will I stay in the ICU?	Usually from one day before until 5 days after the procedure, varying according to the case.
What should I bring on the day of admission?	Recent exams, ID documents, continuous-use prescriptions, hygiene items, and comfortable clothes.	How long will I use the devices?	Generally up to two days after surgery, depending on clinical stability, urine output (diuresis), and drainage.
What will the pre-surgery preparation be like?	Admission to the ICU or inpatient unit; bedside reception; team guidance on procedures and collection of pre-operative exams.	What is important for my recovery?	Stay calm and strictly follow the health team's instructions.
Can my family member stay with me during hospitalization?	Companions are not allowed in the ICU; they are permitted in the Inpatient Unit.	Will I have pain in the post-operative period?	Yes, chest discomfort is normal due to surgical trauma. It is strictly managed with scheduled and rescue analgesics.
Is it important to report flu-like symptoms?	Yes, they must be evaluated by the physician to determine whether to proceed with or postpone the surgery.	Will I receive pain medication?	Yes, whenever necessary to ensure relief and comfort.
What are the risks of cardiac surgery?	Complications such as heart attack (infarction), arrhythmias, thrombosis, embolism, and cardiac arrest.	Will I be able to sleep with the devices?	There may be initial discomfort, but the team will perform the necessary care to promote the best possible rest.
Can I die?	There is a risk of death, but it is low in reference hospitals. As it is a major surgery, the risk is present.	Can my family member accompany me in the post-operative period?	In the ICU (1 to 5 days) visits are restricted. In the room (inpatient unit), companions are allowed.
Will I have complications in the post-operative period?	The most common are arrhythmias and sodium or potassium imbalances.	What are the care instructions when going home?	Medication management; avoid using arm strength when getting up; wound hygiene with warm water and neutral soap; gradual return to activities (4-6 weeks); do not smoke and drive only after 30 days.

discrepancies from expected answers. The maximum context of 8192 tokens was empirically defined to accommodate both the user's question chunks and the chunks retrieved from the KB.

The retrieval mechanism (retriever) was configured to fetch the most relevant chunks. The parameter $k = 5$ was empir-

TABLE III
RAG PIPELINE ARCHITECTURAL SPECIFICATIONS AND HYPERPARAMETERS.

Component/Parameter	Specification / Value
Large Language Model (LLM)	Llama-3-8B-Instruct (4-bit NF4 Quantization)
Hardware Infrastructure	Google Colab (NVIDIA T4 GPU, 16GB VRAM)
Key Libraries	LangChain v0.1.0, ChromaDB v0.4.22, PyTorch v2.1.0
LLM Temperature	0.2
Context Window	8,192 tokens
Retrieval Strategy	Vector Similarity ($k = 5$)
Text Splitter / Chunk Size	Recursive Character / 800 chars
Chunk Overlap	200 characters
Embedding Model	all-MiniLM-L6-v2 (Sentence Transformers)
Vector Database	ChromaDB
User Interface Framework	Gradio

ically defined for the indexed documents, meaning that the model retrieves the five (5) text chunks that present the highest semantic similarity with the user's query. The choice of this range ($k = 5$) is an experimental design choice intended to balance context sufficiency with noise reduction.

The system alignment and protection mechanisms (guardrails) were applied through a structured `PromptTemplate`, defining the agent's behavior as follows:

> *"You are a specialized healthcare assistant providing perioperative guidance for cardiac surgery. Provide standardized, safe, objective, and clear guidance using patient-accessible language. Your answers must be strictly grounded in the provided context. Safety Rule: If the query requires individualized medical decisions (such as precise medication dosages, personal diagnoses, or clearance for specific physical activity) or if the information is missing from the context, you must respond: 'This matter concerns your specific medical history and must be discussed directly with your cardiologist or nursing team.' OR 'This information requires an individualized clinical evaluation. Please consult your physician or the responsible healthcare team.' "*

To improve usability and enable remote asynchronous testing, an interactive web interface was deployed using the Gradio library [13]. This task was performed to allow the nurse specialist to perform a preliminary validation, in order to ensure the CA clinical safety, without local execution dependencies.

D. Preparation and indexing of the knowledge base

The processing and preparation of the documents were structured into three phases:

1) *Data Loading:* In this first stage, the use of the `UnstructuredFileLoader` was fundamental because it allowed us to maintain the context of documents structured in various ways, such as an article containing images, captions, and text columns on a page.
2) *Chunking:* In the chunking stage, the function employed was `RecursiveCharacterTextSplitter` with parameters initially set to 1000 characters and, after testing, adjusted to 800. Decreasing the chunks was necessary for a better affinity of the response with the input. To prevent discrepancies in dataset sizes between tests, the `chunk_overlap` function was used, which enables the overlap of input chunks with output chunks; the default value of 200 characters was sufficient for this application.
3) *Vectorization and Embedding:* Prior to vector storage, a filtering layer was applied to clean metadata keys that could cause any type of semantic noise. Finally, the text segments were transformed into vectors using an embedding model `all-MiniLM-L6-v2` and indexed inside a local instance of `ChromaDB`.

E. Evaluation Methodology

The evaluation of the architecture's adequacy was based on measuring the semantic alignment of the answers generated by the CA against the FAQ (n = 18 QA pairs). To overcome the limitations of purely qualitative evaluations, we employed semantic cosine similarity metrics. Unlike exact word-matching metrics (e.g., BLEU, ROUGE), which lack sensitivity to clinical nuances and synonymous variations [14], cosine similarity captures semantic proximity in high-dimensional space.

Developed as a KB ablation study, the experiment measured the individual and combined impact of each dataset on the model's performance. For the mathematical calculation, both the response produced by the model ($\hat{y}$) and the expected response from the specialist (y) were converted into vectors using the all-MiniLM-L6-v2 embedding model. Similarity was extracted through the dot product of the normalized vectors:

$$\text{Cosine Similarity}(\hat{y}, y) = \frac{\vec{\hat{y}} \cdot \vec{y}}{\|\vec{\hat{y}}\| \|\vec{y}\|} \tag{1}$$

Following the clinical validation methodology established by [15], the behavior of the conversational agent was monitored through the calculation of the overall mean and standard deviation (σ) of similarity across three knowledge base configurations, specifically:

- KB1, with query/answer (Q/A) tasks employed directly to literature data;
- KB2, with Q/A tasks employed only to the FAQ (Q/A pairs) data; and
- KB3, with Q/A tasks employed in a combination of KB1 + KB2.

Additionally, the system was subjected to out-of-scope adversarial questions to evaluate the activation rate of safety guardrails. For these cases, low similarity scores were expected, reflecting the semantic divergence between the generated refusal message and the FAQ answer. This metric serves as an indicator of protocol adherence rather than clinical correctness.

Qualitative Validation: While the primary quantitative metric was cosine similarity, a qualitative validation phase was conducted to assess clinical relevance and safety. The FAQ

responses and a subset of generated outputs were reviewed by the nurse specialist. Due to resource constraints in this preliminary prototyping phase, the evaluation was performed by a single expert. To mitigate individual bias, the evaluation criteria were strictly predefined based on institutional protocols (Table II), focusing on factual alignment and adherence to safety guardrails rather than subjective preference.

F. Ethical considerations

This study was formally reviewed and approved by the Research Commission (COMPESQ) of the Federal University of Health Sciences of Porto Alegre (UFCSPA) under registration number 1022/2024. As this research focuses on computational development using publicly available scientific literature and manually generated FAQs, no human participants were involved in the testing phase, and no private personal data were collected.

III. Results and Discussion

The quantitative performance of the CA was measured using embedding-based semantic cosine similarity. The indicators for each KB configuration are displayed in Table IV.

TABLE IV
Mean similarity and standard deviation for each KB analyzed.

Knowledge Base	Mean similarity	Standard deviation
KB1	0.61	$\pm$ 0.09
KB2	0.60	$\pm$ 0.11
KB3	0.64	$\pm$ 0.12

The experimental results indicate a viable structural topology of the RAG framework, demonstrating an ablation behavior where the synchronized data layers systematically outperformed the isolated baselines in terms of mean similarity scores. KB1 achieved a mean cosine similarity of 0.61. Although the pre-selected scientific articles provide rigorous medical evidence, they lack the specificities of clinical workflow required for adequate patient guidance. As illustrated in Table V, KB1 produced generic responses regarding hospitalization protocols and failed to specify fasting rules, leading to vague or incomplete patient instructions. On the other hand, the isolation of the clinical FAQ protocol in KB2 resulted in a score of 0.60. This performance ceiling may stem from the smaller amplitude of the context window or retrieval inefficiencies.

Notably, KB2 exhibited critical operational failures despite containing the correct information. For instance, the query regarding family accompaniment triggered a safety refusal because the relevant chunk was not ranked within the top-k results (Table V). This highlights a limitation in vector search precision for specific operational rules when relying solely on procedural FAQs. The intersection of both components in the KB3 architecture achieved the highest semantic similarity (0.64) and resolved these qualitative gaps. As shown in Table V, KB3 provided precise fasting guidelines with necessary clinical caveats and correctly nuanced the family accompaniment policy by distinguishing between ICU and Inpatient Unit restrictions. However, the overlap in standard deviations indicates that further statistical validation is required to confirm significant superiority over baseline configurations.

When exposed to queries requiring individualized diagnostic or prescriptive advice, the system suppressed its generative capability to enforce a standardized phrase of refusal and human redirection. In complex medical queries regarding prescription adjustments (for example, *"Is there any medication I should stop before surgery?"*), the semantic cosine similarity under the proposed KB3 architecture dropped to 0.38. This reduction reflects the system's adherence to refusal protocols for out-of-scope queries. The low score indicated that the model did not attempt to generate speculative medical advice, adhering to the safety protocol by prioritizing caution over generation. However, this metric reflects semantic divergence due to refusal rather than clinical accuracy of the refusal itself.

Other analysis must be taken into account. A preliminary examination of the query "*Do I need to fast before admission?*" may suggest an apparent contradiction between the FAQ in Table II (which states: "*No. The fasting protocol (usually 8 hours) is controlled by the nursing staff after admission*") and the model outputs generated under configurations KB2 and KB3 (Table V). Table V specifies an 8-hour fasting duration determined by the anesthesiologist.

In clinical practice, this discrepancy reflects complementary granularity levels within the KB, i.e. while the concise FAQ item addresses hospital admission routines, the institutional protocol explicitly indicates that patients may maintain regular nutrition until admission, after which strict fasting (8 to 10 hours) is prescribed when under the anesthesiologist's prescription and nursing oversight. Furthermore, integrating scientific literature within KB3 grounded the FAQ with established perioperative nursing protocols [**?**], [5]. Although this synthesis demonstrates the capacity of the hybrid RAG architecture to reconcile operational workflows with clinical literature not fully detailed in the initial isolated FAQ, this observation remains preliminary and warrants multi-expert clinical validation before being established as a definitive performance gain.

Finally, the performance of the architecture with a combined base stood out in queries regarding rehabilitation and exercise limits (for example, *"When can I resume physical activities?"*), where KB3 reached a similarity of 0.72. This result stems from the model's capacity to cross-reference clinical information from the scientific literature (KB1) with the practical schedules provided by the FAQ (KB2), suggesting the enrichment of the response from the RAG-based framework.

A. Limitations

The evaluations were conducted in a controlled environment under a single nurse specialist's supervision, relying on a single institutional protocol without direct end-user testing. Additionally, we acknowledge that using the same embedding model (all-MiniLM-L6-v2) for both retrieval and evaluation

TABLE V
QUALITATIVE ANALYSIS OF RESPONSES

Question	KB1 (Articles)	KB2 (FAQ)	KB3 (Combined)
Should I be hospitalized before my cardiac surgery?	Generic: Allows monitoring and reduces stress, but does not state the direct clinical necessity.	Direct: Yes, hospitalization is recommended beforehand, but without the context of safety and protocols.	Complete: Yes, hospitalization is recommended for evaluation, proper preparation, and medication adjustment.
Do I need to fast before admission?	Vague: No specific mention of fasting before admission.	Specific: Yes, 8-hour fasting is required (protocol data, but without caveats).	Precise and Safe: Yes, 8-hour fasting, with the caveat that the exact time will be prescribed by the anesthesiologist.
Can my family member stay with me during hospitalization?	Misdirected: The model shifts focus to the importance of the patient being well-informed.	Retrieval Failure: Triggered safety refusal despite information being present in KB2.	Clinical Nuance: Yes, it is possible in the Inpatient Unit, but not in the ICU (Rule with protocol exception).

may introduce self-alignment bias, potentially inflating similarity scores. Therefore, these metrics should be interpreted as indicators of contextual alignment rather than definitive measures of clinical factuality.

Future work will incorporate independent evaluation metrics, such as LLM-based judges calibrated for clinical accuracy, and multiple expert evaluation of content generated. Qualitative evaluation might be expanded to a multi-expert evaluation with inter-rater reliability calculations (e.g., Cohen's Kappa). A specific analysis of the written language in documents included in the KB was not addressed. This may generate diverse interpretations regarding its correspondence in multiple idioms (e.g., Portuguese-English). This task must be properly addressed in future work, even though it did not pose any apparent issues according to the nurse specialist.

IV. CONCLUSION

This study presented a RAG-based Conversational Agent prototype designed to support perioperative guidance for cardiac surgery. The findings suggest that combining large language models with a hybrid knowledge base, integrating scientific literature and operational protocols, offers a technically viable architecture to support patient health education in controlled settings.

The results indicate that the performance of the RAG system is highly sensitive to the composition and quality of the vectorized data, highlighting knowledge base curation as a critical factor for contextual alignment. Additionally, the implementation of safety *guardrails* in the *PromptTemplate* gives an indication on how to enforce standardized refusal protocols when facing out-of-scope or prescriptive queries, thereby prioritizing safety redirection over speculative generation.

The primary contribution of this work is demonstrating the feasibility of combining dual-layer domain data with operational refusal guardrails to reduce generative discrepancies in patient-facing health applications. Future work must encompass multi-center protocol integration, formal statistical significance testing, multi-expert blind evaluations with inter-rater reliability, and prospective clinical validation with patients and multiprofessional healthcare teams before real-world deployment.

Supplementary material: The source code, configuration files, and vector database structures developed in this study are publicly available in the following anonymous repository: https://github.com/agabrielatoledo/ChatbotRAG_TCC.

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Boston, MA, USA: Pearson, 2020.

[2] M. Allouch et al., "Conversational agents: Goals, technologies, vision and challenges," *Sensors*, vol. 21, no. 23, p. 8047, Dec. 2021.

[3] Schmitz et al., "Orientações no pré-operatório de cirurgia cardíaca a pacientes idosos: Revisão integrativa," *Rev. Enferm. UERJ*, vol. 21, no. 1, pp. 117–122, 2013.

[4] T. Santos et al., "Orientações de alta hospitalar para o autocuidado após cirurgia cardíaca," *Rev. Baiana Enferm.*, vol. 34, e35725, 2020.

[5] Kazitani et al., "Ansiedade cardíaca no período perioperatório de pacientes submetidos a procedimentos cirúrgicos cardíacos: Estudo observacional," *Rev. Bras. Enferm.*, vol. 75, no. 4, e20210255, 2022.

[6] Barcellos et al., "Construção e validação de cartilha educativa para pacientes no pós-operatório de cirurgia cardíaca: Estudo metodológico," *Rev. Bras. Enferm.*, vol. 76, no. 1, e20220134, 2023.

[7] L. J. Pantoja et al., "Pré-triagem inteligente: Desenvolvimento de chatbot para uso em saúde digital," *Rev. Multidisciplinar Nordeste Mineiro*, vol. 1, no. 2, pp. 1–12, 2024.

[8] C. Jeong, "Generative AI service implementation using LLM application architecture: Based on RAG model and LangChain framework," *J. Intell. Inf. Syst.*, vol. 29, no. 4, pp. 129–164, 2023.

[9] N. Shah et al., "Creation and adoption of large language models in medicine," *JAMA*, vol. 330, no. 10, pp. 866–869, Sep. 2023.

[10] D. Roustan and F. Bastardot, "The clinicians' guide to large language models in healthcare: Focus on hallucinations," *Interactive J. Med. Res.*, vol. 14, e59412, 2025.

[11] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 9459–9474.

[12] C. Wang et al., "BioRAG: A RAG-LLM framework for biological question reasoning," *arXiv preprint arXiv:2403.04156*, 2024.

[13] Gradio. "Gradio Documentation." 2024. [Online]. Available: https://www.gradio.app/docs

[14] N. S. Kumar, S. S. Shreya, and R. Anand, "Retrieval-Augmented Generation in Healthcare: A Comprehensive Review of Architectural Guardrails and Evaluation Metrics," *MDPI AI*, vol. 6, no. 9, pp. 226–245, 2025.

[15] Ç. U. Öğdü, K. Arslanoğlu, and M. Karaköse, "An Adaptive Multi-Agent LLM-Based Clinical Decision Support System Integrating Biomedical RAG and Web Intelligence," *IEEE Access*, vol. 13, pp. 3588–3605, 2025.
[16] S. S. Leite et al., "Construction and validation of an Educational Content Validation Instrument in Health," *Rev. Bras. Enferm.*, vol. 71, suppl 4, pp. 1635–1641, 2018.
[17] E. Oro et al., "Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models," in *Proc. Ital-IA Intelligenza Artificiale - Thematic Workshops*, Naples, Italy, May 2024, CEUR-WS 3762, pp. 12–17.

Quantificação Automatizada de Infiltrado Neutrofílico em Tecido de Ligamento Periodontal Aplicando Segmentação por Superpixels e Aprendizado de Máquina

Carlos E. R. Santos*, Gabrielly A. Damásio†, Marílina P. Lucisano†,
Andiara D. R. Daldegan†, Joaquim C. Felipe*
*Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo (USP), Ribeirão Preto, Brasil
†Faculdade de Odontologia de Ribeirão Preto, Universidade de São Paulo (USP), Ribeirão Preto, Brasil
E-mails: carlosedrs@alumni.usp.br, gabysaparecida@gmail.com,
marilia.lucisano@forp.usp.br, andiara@usp.br, jfelipe@ffclrp.usp.br
ORCID: Carlos E. R. Santos 0009-0001-2917-2346; Marílina P. Lucisano 0000-0002-6866-0561;
Andiara D. R. Daldegan 0000-0002-0571-9474; Joaquim C. Felipe 0000-0002-3411-3551

***Resumo*—Este estudo desenvolveu um método computacional para apoiar a quantificação automatizada de neutrófilos em imagens de imuno-histoquímica do tecido periodontal apical. O sistema foi implementado em Python com interface gráfica, utiliza segmentação por superpixels e extrai características de cor nos espaços RGB, CIELAB e HSV, além de descritores de textura. Após avaliação preliminar, a configuração CIELAB + textura foi adotada nos modelos finais. Foram implementados quatro classificadores supervisionados: Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) e Multilayer Perceptron (MLP). Eles foram aplicados em duas etapas: primeiro, identificação e exclusão de regiões da lâmina que não contêm tecido; segundo, identificação dos neutrófilos entre outros componentes teciduais. Na Etapa 1, o XGBoost alcançou acurácia de 0.928. Na Etapa 2, o Random Forest alcançou acurácia de 0.919. A solução proposta fornece mapas visuais e estimativas percentuais da área inflamada, indicando viabilidade inicial para apoiar análises histopatológicas experimentais.**

***Palavras-chave*—Aprendizado de Máquina, Processamento de Imagens, Diagnóstico Auxiliado por Computador, Periodontite Apical.**

I. Introdução

As doenças periodontais representam um significativo problema de saúde pública, com estudos indicando que podem afetar até 90% da população mundial [1]. Nesses cenários, os neutrófilos, leucócitos que constituem a primeira linha de defesa inata, migram em grande volume para o tecido periodontal. Esse acúmulo celular é considerado um marcador clássico e direto da inflamação aguda, tornando sua quantificação em cortes histológicos submetidos à imuno-histoquímica uma etapa essencial para avaliar a progressão da doença e a eficácia de tratamentos. Para investigar a patogênese dessas afecções e testar novas abordagens terapêuticas, modelos experimentais murinos são amplamente utilizados [2]–[4].

Contudo, a abordagem tradicional para a quantificação dessas células apresenta limitações operacionais severas. Realizada manualmente por especialistas por meio de microscopia óptica, a análise visual consome tempo excessivo e é amplamente criticada por sua fragilidade analítica. Trata-se de um processo subjetivo, suscetível à fadiga visual e que resulta em baixa reprodutibilidade, caracterizada por variabilidade intra e interobservador [5]. Essas restrições justificam a investigação de métodos automatizados capazes de fornecer apoio quantitativo mais consistente em pesquisas odontológicas.

No estado da arte, a patologia digital aliada a algoritmos de visão computacional e inteligência artificial desponta como solução promissora para diversas limitações presentes nos processos tradicionais. Estratégias contemporâneas têm demonstrado sucesso ao aplicar algoritmos de agrupamento, como o Simple Linear Iterative Clustering (SLIC), para realizar a pré-segmentação das imagens em superpixels, regiões homogêneas que agrupam pixels com base na similaridade de cor e proximidade espacial [6], [7]. Esses recursos podem acelerar a anotação por especialistas, preservar a morfologia celular e otimizar a etapa de extração de características. Consequentemente, o uso de superpixels cria uma representação de dados favorável ao treinamento de classificadores supervisionados clássicos de aprendizado de máquina, como RF, SVM e XGBoost [8]–[10].

Trabalhos correlatos em patologia digital têm demonstrado que métodos computacionais baseados em visão computacional e inteligência artificial podem reduzir a subjetividade, o tempo de análise e a variabilidade intra e interobservador presentes na avaliação manual de lâminas histológicas. Pantanowitz et al. [11], por exemplo, investigaram o uso de inteligência artificial na contagem de figuras mitóticas, enquanto Karasayar, Kulaç e Kapucuoğlu [5] discutem a aplicação de ferramentas automatizadas na quantificação de marcadores histopatológicos, como Ki-67, destacando ganhos de objetividade e consistência. No contexto da segmentação de imagens biomédicas, Achanta et al. [7] propuseram o algoritmo SLIC, baseado em superpixels, que permite agrupar

pixels em regiões homogêneas de forma eficiente. Albayrak e Bilgin [12] também demonstraram que abordagens baseadas em superpixels podem favorecer a segmentação celular em imagens histopatológicas ao preservar informações espaciais e estruturais. Apesar desses avanços, ainda há carência de soluções específicas para a quantificação automatizada de infiltrado neutrofílico em imagens de imuno-histoquímica do tecido periodontal apical. Este trabalho busca abordar essa lacuna por meio da combinação de segmentação por superpixels, extração de atributos de cor e textura e classificadores supervisionados.

Diante da necessidade de métodos mais eficientes para pesquisas imuno-histoquímicas, o principal objetivo deste trabalho é desenvolver um método computacional de código aberto, implementado em Python e equipado com interface gráfica, para a análise automatizada de imagens do ligamento periodontal apical. A contribuição do estudo está na aplicação específica à quantificação do infiltrado neutrofílico em tecido periodontal apical, na estratégia de classificação em duas etapas e na integração dos modelos treinados a uma ferramenta para visualizar e estimar o percentual de área positiva inflamada.

II. Métodos

O método proposto foi organizado em três blocos operacionais. Inicialmente, realizou-se uma etapa preliminar voltada à preparação das imagens, extração das características e definição do espaço de cor mais adequado para a classificação. Em seguida, o pipeline foi dividido em duas etapas principais de classificação: a Etapa 1, destinada à identificação das regiões de fundo da lâmina, e a Etapa 2, destinada à identificação dos superpixels correspondentes a neutrófilos entre os demais componentes teciduais. Essa organização permitiu separar a remoção de regiões sem tecido da quantificação efetiva do infiltrado inflamatório.

A. Etapa Preliminar

Na etapa preliminar (Fig. 1), o conjunto de dados foi obtido a partir de 10 ratos, com extração de dois primeiros molares inferiores de cada animal: um dente controle e um dente submetido à indução da doença. Uma lâmina foi preparada para cada dente, totalizando 20 lâminas; uma lâmina do grupo controle foi descartada, resultando em 19 imagens analisadas, adquiridas em microscópio Leica DMR acoplado a câmera Leica DFC 300FX. As imagens foram segmentadas por meio do algoritmo SLIC configurado com aproximadamente 5.000 segmentos, fator de compacidade igual a 10 e pré-suavização gaussiana com $\sigma = 3$, gerando superpixels utilizados como unidades básicas de análise. Para cada superpixel, foram extraídas características cromáticas nos espaços RGB, HSV e CIELAB, além de descritores de textura baseados na matriz de coocorrência de níveis de cinza, incluindo homogeneidade, correlação, dissimilaridade e contraste.

B. Ferramenta de Rotulagem Manual

Uma ferramenta própria apoiou a construção do padrão-ouro (Fig. 2). Ela segmentava cada imagem com os mesmos parâmetros do SLIC usados no treinamento e armazenava o identificador de cada superpixel rotulado. Como esses identificadores são determinísticos quando os parâmetros do SLIC são mantidos, os rótulos puderam ser mapeados de forma consistente para a tabela de atributos. Durante a anotação, os especialistas rotulavam a imagem segmentada enquanto inspecionavam a imagem original lado a lado.

C. Etapa 1 - Classificação de Fundo

Na Etapa 1, foi treinado um classificador auxiliar para distinguir superpixels correspondentes ao fundo da lâmina daqueles pertencentes ao tecido histológico. Para isso, os superpixels foram rotulados manualmente por três especialistas como “fundo” ou “não fundo”, formando o padrão-ouro por consenso dessa etapa. Devido ao desbalanceamento entre as classes, foi aplicada subamostragem da classe majoritária antes do treinamento para equilibrar as classes e reduzir viés em direção ao rótulo predominante. O classificador foi então avaliado por meio das métricas de acurácia, sensibilidade, especificidade e área sob a curva ROC (AUC). A Fig. 3 ilustra esse processo.

D. Etapa 2 - Classificação de Neutrófilos

Na Etapa 2 (Fig. 4), foi treinado o classificador principal para identificar superpixels positivos para neutrófilos em relação aos demais componentes teciduais. Para a formação dessa base, foram excluídos os superpixels rotulados manualmente como fundo no padrão-ouro, garantindo que o treinamento fosse realizado apenas sobre regiões correspondentes a tecido. Os superpixels remanescentes foram então rotulados pelos mesmos três especialistas, por consenso, como “neutrófilo” ou “não neutrófilo”. Após o balanceamento das classes, os classificadores supervisionados foram treinados e avaliados utilizando os conjuntos de atributos definidos na etapa preliminar. Posteriormente, o modelo de melhor desempenho foi integrado à interface gráfica para geração do mapa visual dos superpixels positivos e cálculo do percentual de área tecidual inflamada.

E. Procedimentos de Treinamento e Avaliação

Após a segmentação das imagens, foram estruturados dois conjuntos de dados distintos, correspondentes às duas etapas do pipeline. Para o classificador de fundo, o conjunto totalizou 92.340 superpixels, sendo 2.700 rotulados como fundo e 89.640 como não fundo. Após subamostragem da classe majoritária para reduzir o viés causado pelo desbalanceamento entre as classes, obteve-se um conjunto balanceado de 5.400 instâncias, com 2.700 amostras por classe.

Para o classificador de neutrófilos, os superpixels rotulados manualmente como fundo foram excluídos, resultando em 94.999 superpixels. Antes do balanceamento, havia 82.137 amostras não neutrofílicas e 12.862 neutrofílicas. Após subamostragem, o conjunto final utilizado no treinamento continha 25.724 instâncias balanceadas, sendo 12.862 por classe. Todos os modelos foram avaliados por validação cruzada estratificada em 5 folds, com embaralhamento e semente

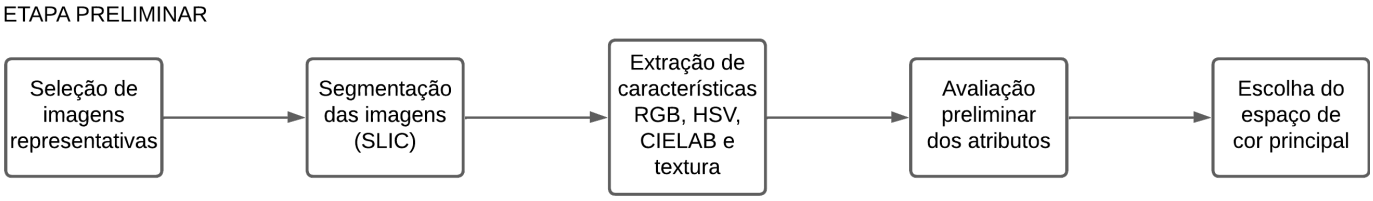


Fig. 1. Fluxograma da etapa preliminar.

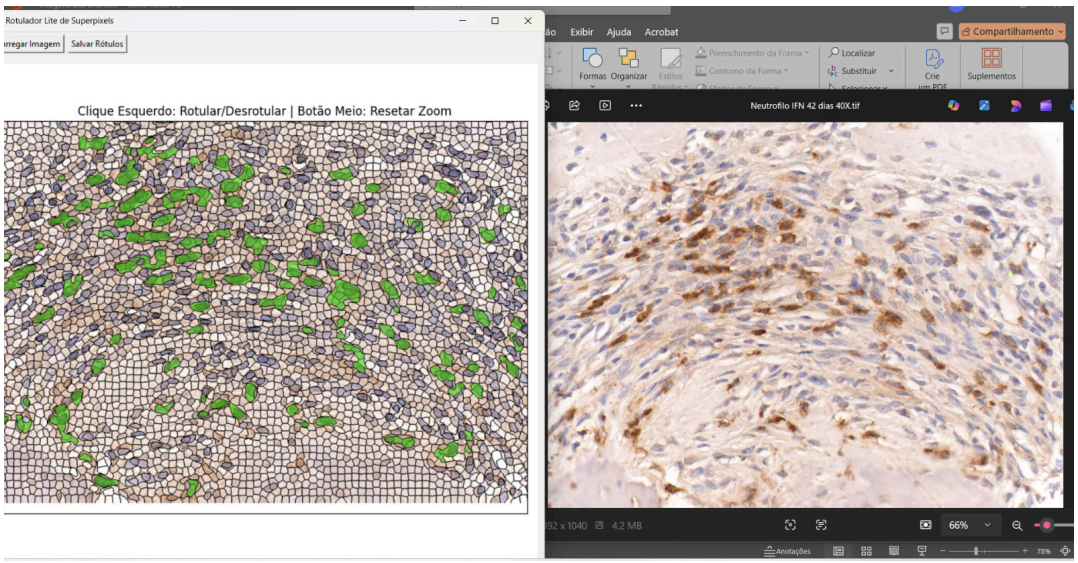


Fig. 2. Ferramenta de rotulagem manual com imagem segmentada e imagem original lado a lado.

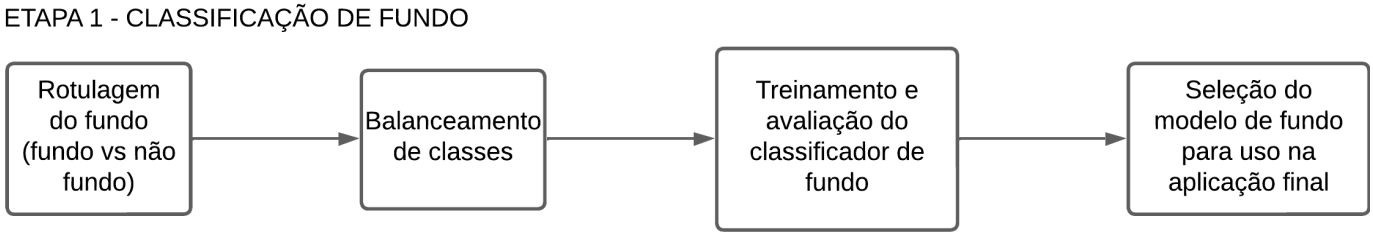


Fig. 3. Fluxograma do classificador de fundo na Etapa 1.

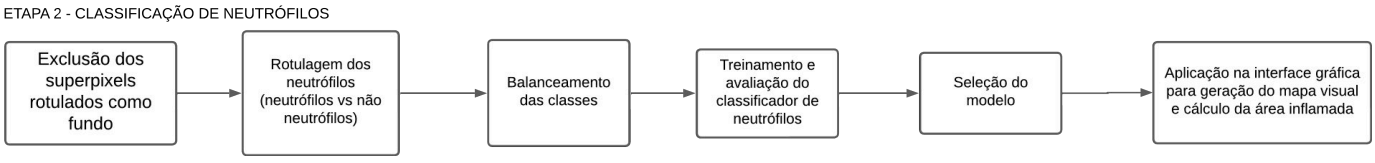


Fig. 4. Fluxograma do classificador de neutrófilos na Etapa 2.

aleatória 42, utilizando acurácia, sensibilidade, especificidade e AUC sob a representação CIELAB + textura. A divisão foi realizada no nível dos superpixels individuais no conjunto consolidado para preservar o balanceamento entre classes nesta avaliação preliminar; a validação agrupada por imagem, lâmina ou animal é, portanto, definida como etapa necessária em estudos futuros.

F. Aspectos Éticos

As imagens utilizadas neste estudo são derivadas de amostras teciduais secundárias e anonimizadas, provenientes de experimentos laboratoriais prévios conduzidos por pesquisadoras colaboradoras. Nenhum novo procedimento animal ou humano foi realizado para esta análise computacional, e os procedimentos animais primários foram aprovados pelo Comitê de Ética em Experimentação Animal da Faculdade de

Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP, Protocolo nº 003/2006), seguindo as diretrizes ISO 10993-2.

III. Resultados

A. Avaliação Preliminar dos Atributos

Antes da definição do modelo final de classificação, foi realizada uma avaliação preliminar dos conjuntos de atributos extraídos dos superpixels usando o modelo Random Forest. Essa análise teve como objetivo comparar o desempenho dos diferentes espaços de cor, RGB, HSV e CIELAB, bem como dos descritores de textura e de suas combinações, a fim de selecionar uma representação adequada para a etapa de classificação dos neutrófilos. Os resultados são apresentados na Tabela I.

Tabela I
Comparação dos Espaços de Cor no Random Forest

Descritor	Acurácia	Sensibilidade	Especificidade	AUC
RGB	0.880	0.886	0.874	0.952
HSV	0.880	0.889	0.871	0.950
CIELAB	0.888	0.895	0.880	0.956
Apenas textura	0.733	0.794	0.672	0.795
CIELAB + textura	0.919	0.929	0.909	0.972

Embora os modelos de cor avaliados tenham gerado resultados marginalmente semelhantes, o modelo CIELAB obteve o melhor desempenho em comparação com RGB e HSV. Verificou-se também que a combinação de descritores de cor e de textura levou a resultados melhores do que o uso desses descritores separadamente. Por isso, o espaço de cor CIELAB foi adotado como representação cromática principal, considerando o equilíbrio entre desempenho, menor dimensionalidade e interpretabilidade.

Para complementar a análise de desempenho, foi avaliada a importância dos atributos no modelo XGBoost por meio do método Gain. Essa análise permitiu identificar quais características contribuíram mais para as decisões do classificador. Os resultados da Tabela II indicaram que a média do canal 1 do CIELAB apresentou o maior ganho preditivo, reforçando a relevância desse espaço de cor para a discriminação dos superpixels positivos para neutrófilos.

Tabela II
Ganho por Atributo Usando o Método Gain do XGBoost

Posição	Atributo	Ganho
1º	Média do canal 1 CIELAB	36.710
2º	Média do canal 3 CIELAB	11.646
3º	GLCM homogeneity	8.457
4º	Média do canal 2 CIELAB	5.737
5º	GLCM dissimilarity	5.105
6º	GLCM correlation	4.277
7º	GLCM contrast	3.649

A análise por Gain no XGBoost foi utilizada apenas como avaliação complementar da relevância dos atributos, indicando contribuição importante dos canais CIELAB. No entanto, o modelo final selecionado para a Etapa 2 foi o Random Forest com CIELAB + textura, devido ao melhor equilíbrio global nas métricas avaliadas.

B. Resultados da Etapa 1 - Classificação de Fundo

A Etapa 1 teve como objetivo treinar e avaliar um classificador auxiliar capaz de distinguir regiões correspondentes ao fundo da lâmina daquelas contendo tecido histológico. Essa etapa foi necessária para evitar que áreas sem tecido fossem consideradas no cálculo da área inflamada durante a aplicação final da ferramenta. Para essa tarefa, foram utilizados os atributos CIELAB + textura.

Conforme apresentado na Tabela III, os quatro classificadores supervisionados [8]–[10], [13] foram avaliados na Etapa 1, correspondente à distinção entre superpixels de fundo e não fundo, utilizando a configuração CIELAB + textura. O XGBoost apresentou o melhor desempenho geral, com acurácia média de 0,928 e AUC de 0,981, seguido pelo Random Forest, com acurácia de 0,927 e AUC de 0,981. Embora a MLP tenha apresentado a maior sensibilidade, com 0,935, e o Random Forest tenha alcançado a maior especificidade, com 0,928, o XGBoost demonstrou o melhor equilíbrio entre as métricas avaliadas, combinando maior acurácia e alta sensibilidade. Dessa forma, o XGBoost foi selecionado como classificador final da Etapa 1, responsável pela identificação das regiões de fundo da lâmina.

Tabela III
Desempenho dos Classificadores para Identificação de Fundo Usando CIELAB + Textura

Modelo	Acurácia	Sensibilidade	Especificidade	AUC
XGBoost	0.928	0.934	0.923	0.981
Random Forest	0.927	0.926	0.928	0.981
MLP	0.924	0.935	0.913	0.978
SVM	0.917	0.931	0.904	0.973

C. Resultados da Etapa 2 - Classificação de Neutrófilos

Na Etapa 2, o objetivo foi classificar os superpixels correspondentes às regiões teciduais como positivos ou negativos para a presença de neutrófilos. Para a construção dessa base, os superpixels rotulados manualmente como fundo no padrão-ouro foram excluídos, garantindo que o treinamento fosse realizado apenas sobre regiões de tecido. Após essa exclusão, o conjunto utilizado para a classificação de neutrófilos totalizou 94.999 superpixels, sendo 82.137 rotulados como não neutrófilos e 12.862 como neutrófilos. Em razão do desbalanceamento entre as classes, foi aplicada subamostragem da classe majoritária para reduzir viés em direção às regiões não neutrofílicas, resultando em um conjunto balanceado de 25.724 instâncias, com 12.862 superpixels por classe.

Utilizando os atributos CIELAB + textura, foram avaliados os classificadores Random Forest, XGBoost, MLP e SVM por validação cruzada estratificada com cinco folds. Conforme apresentado na Tabela IV, os quatro modelos produziram desempenhos próximos. O Random Forest obteve a maior acurácia média entre os modelos avaliados, com 0,919, além da maior especificidade, com 0,909. Embora o XGBoost e

a MLP tenham apresentado AUC ligeiramente superior, e o SVM tenha alcançado maior sensibilidade, o Random Forest apresentou relação favorável entre desempenho preditivo e custo computacional, sendo selecionado como classificador final para a identificação dos neutrófilos.

Tabela IV
DESEMPENHO DOS CLASSIFICADORES USANDO CIELAB + TEXTURA

Modelo	Acurácia	Sensibilidade	Especificidade	AUC
Random Forest	0.919	0.929	0.909	0.972
XGBoost	0.918	0.930	0.906	0.973
MLP	0.917	0.926	0.907	0.973
SVM	0.915	0.938	0.893	0.970

Na Etapa 2, a matriz de confusão média (Tabela V) apresentou 2.338 verdadeiros negativos, 234 falsos positivos, 184 falsos negativos e 2.389 verdadeiros positivos. Possíveis fontes de falsos positivos incluem precipitados inespecíficos de DAB e regiões densas de tecido conjuntivo com características cromáticas ou texturais semelhantes às áreas imunopositivas. Por outro lado, os falsos negativos podem estar relacionados à imunorreatividade tênue ou à presença de células nas bordas dos superpixels, onde o valor cromático médio do segmento é atenuado por pixels adjacentes de tecido não reativo. Essas hipóteses não foram avaliadas por meio de uma revisão sistemática das instâncias classificadas incorretamente e devem ser investigadas em estudos futuros.

Tabela V
MATRIZ DE CONFUSÃO DO RANDOM FOREST USANDO CIELAB E TEXTURA, COM ARREDONDAMENTO PELA MÉDIA DOS CINCO FOLDS

	Predito Negativo	Predito Positivo
Real Negativo	2338	234
Real Positivo	184	2389

D. Aplicação na Interface Gráfica

A Fig. 5 mostra a interface gráfica destacando em vermelho os superpixels classificados como "fundo" e em verde os superpixels classificados como "neutrófilos". As métricas quantitativas são exibidas em uma aba separada de estatísticas (Fig. 6), que informa o número de superpixels classificados como neutrófilos, não neutrófilos e fundo, seguido do percentual de área positiva para neutrófilos após a exclusão do fundo. A plataforma integra o modelo Random Forest para quantificação de neutrófilos e o XGBoost para exclusão de fundo. O tempo completo de processamento por imagem, incluindo segmentação SLIC, extração de atributos de cor e textura e inferência de ambos os classificadores, foi de aproximadamente 31,5 s, variando de 28,7 a 35,0 s em amostras representativas em uma estação convencional baseada em CPU com 16 GB de RAM. Esse tempo é compatível com avaliação futura em fluxos de trabalho laboratoriais.

IV. DISCUSSÃO

Os resultados obtidos indicam viabilidade inicial da combinação de segmentação por superpixels e aprendizado de máquina para apoiar a quantificação automatizada de neutrófilos em imagens de imuno-histoquímica do tecido periodontal apical. A organização do método em duas etapas permitiu separar a identificação de regiões de fundo da classificação dos neutrófilos nas áreas teciduais. Na Etapa 1, o XGBoost apresentou o melhor desempenho para classificar fundo e tecido, com acurácia de 0.928. Na Etapa 2, o Random Forest obteve a maior acurácia para a classificação dos neutrófilos, com acurácia de 0.919.

A escolha dos atributos CIELAB + textura foi apoiada pela avaliação preliminar, pois essa representação obteve a melhor acurácia na Tabela I. Os atributos cromáticos são relevantes porque a marcação por DAB produz diferenças de coloração entre regiões positivas e negativas; o CIELAB separa luminosidade e componentes cromáticos, enquanto os descritores GLCM incorporam informações de textura. Ainda é necessário, porém, realizar um estudo formal de ablação para separar a contribuição individual dos atributos de cor e textura.

A comparação entre classificadores mostrou que o melhor modelo variou conforme a subtarefa. Na classificação de fundo, o XGBoost apresentou a maior acurácia média; na classificação de neutrófilos, o Random Forest foi selecionado porque acurácia, especificidade e custo computacional foram priorizados para reduzir a superestimação da área inflamada e preservar a execução em hardware convencional. Embora XGBoost e MLP tenham apresentado AUC ligeiramente superior, e o SVM tenha alcançado maior sensibilidade, o Random Forest produziu a maior acurácia (0,919) e especificidade (0,909). Na Etapa 2, a matriz de confusão média apresentou 2.338 verdadeiros negativos, 234 falsos positivos, 184 falsos negativos e 2.389 verdadeiros positivos. Os falsos positivos estiveram associados sobretudo a precipitados de DAB e condensações estromais, enquanto os falsos negativos se concentraram em células com marcação tênue ou nas bordas dos superpixels.

O uso de classificadores clássicos combinados a superpixels foi uma escolha voltada ao baixo custo computacional e à execução sem GPU. Ainda assim, a validação foi realizada em uma única base institucional e com folds no nível dos superpixels, o que pode permitir que regiões da mesma imagem apareçam nos conjuntos de treino e validação e inflar as métricas. Embora os rótulos tenham sido estabelecidos por consenso entre três especialistas, não foi incluído coeficiente formal de concordância. Os superpixels também podem ser imprecisos em áreas com sobreposição celular, marcação heterogênea ou transições pouco definidas entre tecido e fundo.

V. CONCLUSÃO

Este estudo demonstrou a viabilidade metodológica inicial de um pipeline em duas etapas para quantificação automatizada de infiltrado neutrofílico em tecido periodontal apical murino. A abordagem alcançou acurácia de 0.928 para classificação de fundo com XGBoost e de 0.919 para classificação de neutrófilos com Random Forest usando descritores CIELAB + textura. A ferramenta fornece mapas visuais e estimativas quantitativas da área inflamada, com tempos de processamento

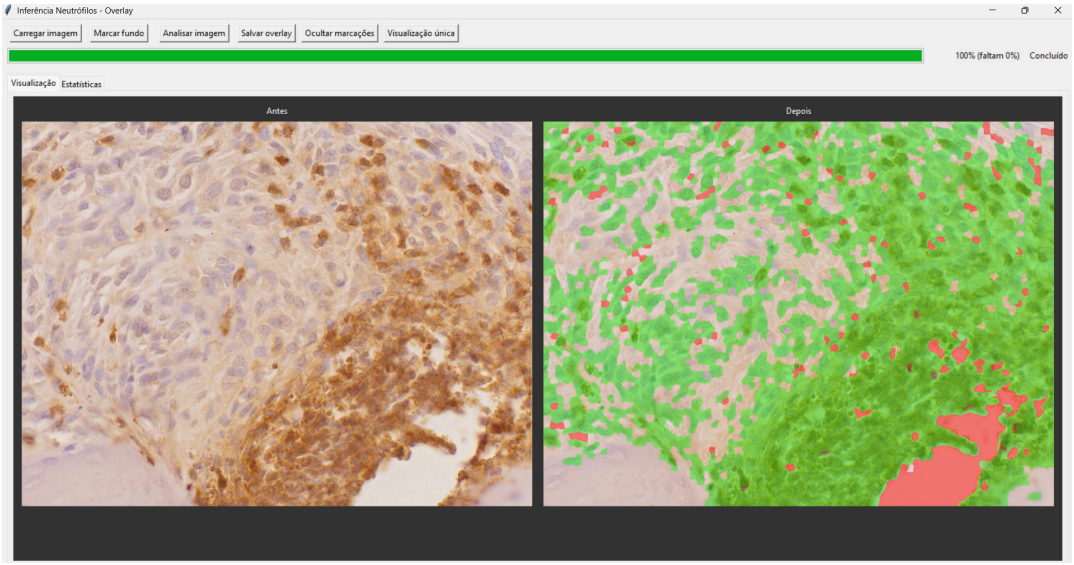


Fig. 5. Interface gráfica: overlay da aba “Visualização”; métricas percentuais na aba “Estatísticas”.

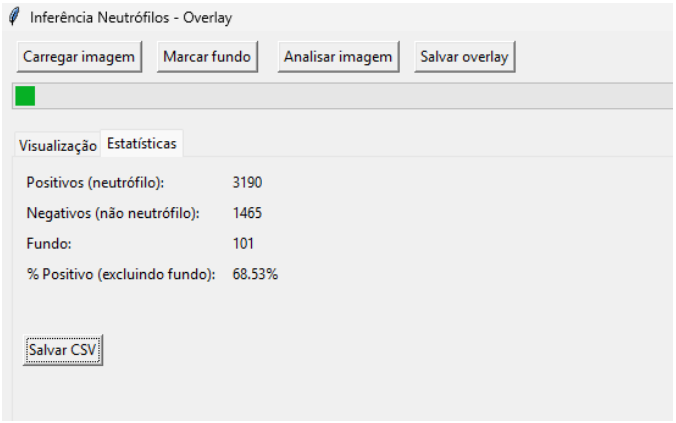


Fig. 6. Aba de estatísticas com contagem de superpixels e percentual positivo para neutrófilos excluindo o fundo. Este exemplo é ilustrativo e não corresponde necessariamente à Fig. 5.

que sugerem potenciais ganhos de execução em relação à análise manual convencional.

Como trabalho futuro, a validação deve ser ampliada com outras bases de imagens adquiridas por diferentes equipamentos, a fim de avaliar a robustez sob diferentes condições de coloração, aquisição e preparo. Estudos futuros também devem incluir divisão dos dados por imagem ou animal, análise de reprodutibilidade, comparação direta com especialistas, avaliação de modelos de aprendizado profundo e testes independentes antes de reivindicações de aplicabilidade clínica ou diagnóstica.

References

[1] B. L. Pihlstrom, B. S. Michalowicz, and N. W. Johnson, “Periodontal diseases,” *The Lancet*, vol. 366, no. 9499, pp. 1809–1820, 2005.

[2] I. Aunaite, G. Lodiene, and V. Maciulskiene, “Pathogenesis of apical periodontitis: A literature review,” *Journal of Oral & Maxillofacial Research*, vol. 2, no. 4, p. e1, 2012.

[3] N. S. Gasner and R. S. Schure, “Periodontal disease,” in *StatPearls*. Treasure Island, FL, USA: StatPearls Publishing, 2025.

[4] G. Hajishengallis, T. Chavakis, and J. D. Lambris, “Current understanding of periodontal disease pathogenesis and targets for host-modulation therapy,” *Periodontology 2000*, vol. 84, no. 1, pp. 14–34, 2020.

[5] A. H. D. Karasayar, I. Kulaç, and N. Kapucuoğlu, “Advances in breast cancer care: The role of artificial intelligence and digital pathology in precision medicine,” *European Journal of Breast Health*, vol. 21, no. 2, pp. 93–100, 2025.

[6] J. McQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA, USA: University of California Press, 1967, pp. 281–297.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[8] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: Association for Computing Machinery, 2016, pp. 785–794.

[11] L. Pantanowitz *et al.*, “Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses,” *Diagnostic Pathology*, vol. 15, no. 1, p. 80, 2020.

[12] A. Albayrak and G. Bilgin, “Automatic cell segmentation in histopathological images via two-staged superpixel-based algorithms,” *Medical & Biological Engineering & Computing*, vol. 57, no. 3, pp. 653–665, 2019.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

Risk Factors for Severe RSV Outcomes in Children: A Systematic Review and Meta-Analysis

1st Mariana Mioti Sebastião
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
m.mioti@aluno.ufabc.edu.br

2nd Fernanda Almeida
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
fernanda.almeida@ufabc.edu.br

***Abstract*—Respiratory syncytial virus (RSV) is a leading cause of severe lower respiratory tract infections in young children, resulting in substantial morbidity and healthcare resource utilization. This study aimed to identify risk factors associated with severe RSV outcomes through a systematic review and meta-analysis. Searches were conducted in PubMed, Embase and Scopus for observational studies published between 2006 and 2026. Studies evaluating predictors of severe RSV outcomes, including intensive care unit (ICU) admission, ventilatory support, or death, in children aged 0–5 years were included. Nine studies met eligibility criteria. Across studies, younger age, prematurity, congenital heart disease, chronic lung disease, and neurological comorbidities were identified as recurrent predictors of severe disease. Random-effects meta-analyses demonstrated significantly increased odds of severe outcomes among infants younger than six months (OR = 3.42; 95% CI: 2.23–5.27), premature infants (OR = 1.80; 95% CI: 1.30–2.49), and children with congenital heart disease (OR = 2.56; 95% CI: 1.46–4.51). Although the pooled estimate for chronic lung disease suggested increased risk, the association was not statistically significant (OR = 1.30; 95% CI: 0.79–2.14). Heterogeneity was substantial for age below six months but low to moderate for the remaining analyses. These findings highlight the importance of young age, prematurity, and congenital heart disease as key determinants of severe RSV outcomes and may help identify populations at increased risk of severe disease.**

***Index Terms*—Respiratory syncytial virus (RSV), severe disease, risk factors, prematurity, congenital heart disease, meta-analysis, systematic review.**

I. Introduction

Respiratory syncytial virus (RSV) is the leading etiological agent associated with acute lower respiratory tract infections in infants and young children. The global burden of morbidity and mortality attributable to RSV is considerable among individuals aged 0 to 60 months, with the highest impact observed during the first six months of life and in low- and middle-income countries. Estimates suggest that approximately one in every 50 deaths in this age group, and one in every 28 deaths among children aged 28 days to 6 months, are due to RSV infection. Furthermore, for each hospital death related to RSV-associated acute lower respiratory infection, it is projected that about three additional deaths occur in community settings [1]. Consequently, RSV infections in young children entail substantial healthcare resource utilization. Although even severe cases typically do not require specific diagnostic tests or targeted antiviral treatments, the overall economic burden is considerable, reflecting the high costs associated with hospitalizations and outpatient care worldwide [2].

Monoclonal antibodies have emerged as important options for the prevention of RSV infection, particularly in vulnerable pediatric populations. Palivizumab, the first widely used agent, has demonstrated effectiveness but is limited in its broader application due to cost considerations and the need for repeated dosing [3]. More recently, newer long-acting monoclonal antibodies such as nirsevimab have been developed, offering the advantage of extended protection with a simplified administration schedule [4]. In parallel, advances in maternal vaccination strategies—leveraging transplacental antibody transfer—have shown promising efficacy in protecting infants during early life [5].

Together, these evolving preventive approaches represent significant progress in reducing the impact of RSV. However, given the substantial healthcare burden associated with RSV infections, particularly in young children, careful consideration of cost-effectiveness and population-level impact remains essential. Identifying those infants at highest risk of severe disease continues to be a key challenge in optimizing the use of these preventive strategies [6].

Risk factors associated with RSV-related hospitalizations have been widely investigated. However, less is known about the determinants of severe outcomes—such as mortality, admission to intensive care units, need for ventilatory support, or prolonged hospital stays—which represent the most critical events to prevent but are often difficult to study due to limited statistical power [7].

In this context, a comprehensive understanding of the factors associated with severe RSV outcomes is crucial to inform clinical decision-making and optimize preventive strategies [8]. However, existing evidence remains fragmented, with variability in study designs, populations, and outcome definitions limiting the ability to draw consistent conclusions [7].

Therefore, this study aims to systematically synthesize the available evidence on risk factors associated with severe RSV infection in young children through a systematic review and meta-analysis. By identifying and quantifying the key determinants of severe outcomes, this work seeks to support risk stratification approaches, contribute to the more efficient

allocation of emerging preventive interventions, and provide evidence that may inform the development of future predictive models for severe RSV outcomes.

II. Methods

A. Search strategy

This systematic review was guided by the research question: "Which factors are associated with severe outcomes in infants and children with respiratory syncytial virus (RSV) infection?". To address this question, a comprehensive search was conducted in PubMed, Embase and Scopus, covering the period 2006–2026. The search combined controlled descriptors (MeSH/Emtree) and free-text terms related to RSV, disease severity (ICU admission, mechanical ventilation, mortality), risk factors, and pediatric populations. Search strategies were tailored to the indexing structure of each database, and filters were applied to include observational analytical studies while excluding randomized trials, reviews, and other non-eligible designs. Full search strings for each database are provided in TABLE I, with the final search date of June 15, 2026.

The study selection process followed the PRISMA 2020 guidelines for systematic reviews and meta-analyzes [9]. All records retrieved were imported into Rayyan, a web-based application for systematic reviews [10]. Duplicate records were removed using Rayyan software, which also supported the screening. No protocol was registered for this review and screening was performed by a single reviewer in all three sequential phases: (i) title and abstract screening to exclude irrelevant studies; (ii) full-text assessment of potentially eligible articles against predefined inclusion and exclusion criteria; and (iii) final selection of studies meeting all eligibility requirements. This structured approach ensured transparency, reproducibility, and methodological rigor.

B. Eligibility Criteria

- **Inclusion:** observational analytical studies (cohort or case-control) evaluating infants and children (0–5 years) with RSV infection, reporting risk factors associated with severe outcomes (ICU admission, ventilatory support, mortality) and presenting bi-variate/multivariate analysis.
- **Exclusion:** studies with adult populations, descriptive-only designs, subjective severity scores, cross-sectional studies, randomized trials, reviews, editorials, case reports.
- **Outcomes Definition:** severe outcomes were defined as ICU admission, need for invasive or non-invasive ventilatory support, or death. Studies using composite severity scores were excluded to ensure comparability across included studies. These endpoints were chosen because they are objective events, consistently reported across studies, and therefore allow comparability. In contrast, outcomes based on symptom scores, laboratory markers, or composite severity indices are more subjective and vary widely between studies, making them difficult to standardize.

TABLE I: Search strategy by database (performed on 15 June 2026)

Database	Search strategy
PubMed	("Respiratory Syncytial Virus"[Mesh] OR RSV OR "respiratory syncytial virus") AND ("Disease Severity"[Mesh] OR severity OR ICU OR "intensive care" OR mortality OR "mechanical ventilation") AND ("Risk Factors"[Mesh] OR "risk factor*" OR predictor* OR determinant*) AND ("Infant"[Mesh] OR "Child"[Mesh] OR infant* OR child* OR pediatric* OR paediatric*) Filters: Observational Study.
Embase	('respiratory syncytial virus'/exp OR rsv OR 'respiratory syncytial virus') AND ('disease severity'/exp OR severity OR icu OR 'intensive care' OR mortality OR 'mechanical ventilation') AND ('risk factor'/exp OR 'risk factor*' OR predictor* OR determinant*) AND ('infant'/exp OR 'child'/exp OR infant* OR child* OR pediatric* OR paediatric*) AND (multivariable OR regression OR 'logistic regression') AND ('observational study'/de OR 'cohort analysis'/de OR observational OR cohort) NOT ('randomized controlled trial'/exp OR 'clinical trial'/exp OR randomized OR randomised) AND [2006-2026]/py AND 'article'/it AND 'observational study'/de.
Scopus	TITLE-ABS-KEY ("respiratory syncytial virus" OR RSV) AND TITLE-ABS-KEY (severity OR ICU OR "intensive care" OR mortality OR "mechanical ventilation") AND TITLE-ABS-KEY ("risk factor*" OR predictor* OR determinant*) AND TITLE-ABS-KEY (infant* OR child* OR pediatric* OR paediatric*) AND TITLE-ABS-KEY (multivariable OR regression OR "logistic regression") AND TITLE-ABS-KEY (observational OR cohort OR "observational study") AND PUBYEAR ¿ 2005 AND PUBYEAR ¡ 2027 AND DOCTYPE (ar) AND NOT TITLE-ABS-KEY ("randomized" OR "randomised" OR "clinical trial").

C. Data Extraction and Synthesis

- **Data extraction:** study data were collected by a single reviewer using a standardized extraction form. For each eligible study, information was collected on study characteristics (author, year, country, sample size, age group, and definition of severe outcome) and risk factors reported in bi-variate or multivariate analyses. When available, odds ratios (OR) with 95% confidence intervals (CI) were extracted as the primary measure of association. Relative risks (RR) and hazard ratios (HR) were noted but not included in quantitative pooling. Extracted data were organized into structured datasets to facilitate both narrative synthesis and meta-analysis.
- **Narrative Synthesis:** a structured narrative synthesis was conducted to integrate findings across studies. This approach emphasized recurrent predictors of severe RSV outcomes, such as young age, prematurity, and congenital heart disease, while also highlighting rare but clinically relevant conditions (e.g., cystic fibrosis, sepsis). The synthesis focused on consistency of associations across different populations and healthcare settings, rather than exhaustive tabulation of individual study data.
- **Risk of Bias Assessment:** risk of bias and methodological quality were assessed using the Newcastle-Ottawa Scale (NOS) for cohort studies. The NOS evaluates three domains: selection of participants (0–4 points), comparability of cohorts (0–2 points), and outcome assessment

(0–3 points), resulting in a total score ranging from 0 to 9 [11].

- **Meta-Analysis:** all analyses were conducted in R version 4.5.1 [12]. Meta-analyses were performed using the `meta` package [13], with additional customization of forest plots supported by `metafor` [14]. Odds ratios (OR) and 95% confidence intervals (CI) from individual studies were log-transformed, and standard errors (SE) were calculated using Equation 1.

$$SE = \frac{\log(\text{Upper}) - \log(\text{Lower})}{2 \cdot 1.96} \tag{1}$$

Random-effects models were applied using the inverse variance method (`metagen` function, `meta` package). Forest plots were generated to display study names, effect sizes, and confidence intervals, with redundant columns removed for clarity. Heterogeneity was assessed using the Q statistic, τ^2, and the I^2 index, with thresholds of 25%, 50%, and 75% interpreted as low, moderate, and substantial heterogeneity, respectively. Study weights were calculated using the inverse-variance random-effects model and are represented graphically by the size of the squares in the forest plots.

Meta-analyses were restricted to risk factors that demonstrated the strongest associations with severe RSV outcomes and for which at least three independent studies reported odds ratios (OR). Only OR estimates were pooled to ensure methodological consistency, while studies reporting relative risks (RR) or hazard ratios (HR) were described narratively but not included in quantitative synthesis.

When multiple severe outcome estimates were reported from the same cohort, only one estimate was included in each pooled analysis. Ventilatory support was prioritized over ICU admission because it was the most consistently reported severe outcome across studies and represents an objective marker of severe respiratory deterioration.

Additionally, sensitivity analyses were performed for pooled factors presenting substantial heterogeneity by sequentially excluding individual studies and recalculating the pooled effect estimates.

III. Results

A. Study Selection Process

A total of 252 records were identified across the databases consulted. After removing 40 duplicates, 212 studies remained for title and abstract screening. During this phase, 195 studies were excluded for addressing non-eligible populations (adults or adolescents), infections by other viruses, unrelated diseases, or outcomes outside the inclusion criteria (purely epidemiological studies, drug or vaccine efficacy, absence of multivariate analysis, or heterogeneous severity definitions).

Seventeen articles were selected for full-text review. Of these, nine met all eligibility criteria and were included in the review. The remaining eight were excluded—six due to inconsistent definitions of disease severity and two for evaluating populations outside the established age range (children older than five years or non-RSV infections).

The study selection process is summarized in Figure 1, following the PRISMA 2020 framework, illustrating the flow from identification to inclusion.

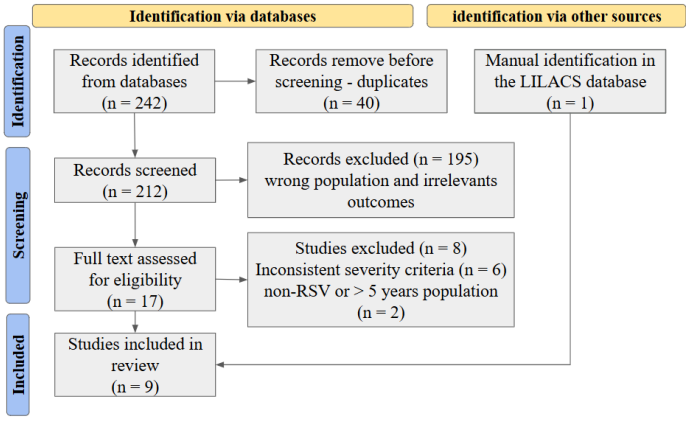


Fig. 1: PRISMA 2020 flow diagram for new systematic reviews [9].

B. Synthesis of Results

The characteristics of the included studies and their main predictors of severe RSV outcomes are summarized in Table II. The studies included in this systematic review span diverse countries and clinical contexts. Collectively, the included studies consistently identified a set of recurring risk factors associated with severe RSV outcomes. Although the specific endpoints differ — ICU admission, mechanical ventilation, ventilatory support, or death — the findings consistently highlight a set of recurrent predictors, with occasional but striking outliers.

In Australia, Anderson et al. (2021) showed that apnea (OR = 4.5; 95% CI: 3.1–6.5), feeding difficulty (OR = 4.5; 95% CI: 3.1–6.5), and increased work of breathing (OR = 3.2; 95% CI: 2.1–5.0) were strong predictors of ventilatory support, particularly in infants under two months (OR = 2.3; 95% CI: 1.6–3.3) [15].

Data from the United States (Brenes-Chacón et al., 2023) reinforced the vulnerability of young infants: age < 6 months predicted ICU admission (OR = 2.27; 95% CI: 1.84–2.81) and was the strongest factor for mechanical ventilation (OR = 5.43; 95% CI: 3.69–8.00). Prematurity also contributed (e.g., 29–34 weeks: OR = 1.78; 95% CI: 1.45–2.20; mechanical ventilation OR = 3.01; 95% CI: 2.29–3.96), as did neuromuscular disease (ICU: OR = 2.71; 95% CI: 1.63–4.48; ventilation: OR = 3.49; 95% CI: 1.69–7.19), viral coinfection (OR = 1.40–1.41), and fever (ICU: OR = 1.84; 95% CI: 1.59–2.14; ventilation: OR = 3.93; 95% CI: 3.02–5.14) [16].

The German cohort studied by Cai et al. (2019) highlighted extreme risks, with cystic fibrosis (OR = 35.13; 95% CI: 1.76–700.59), hepatic disease (OR = 14.99; 95% CI: 1.49–150.82), and perinatal respiratory or cardiovascular disorders

(OR = 4.97; 95% CI: 3.36–7.34) strongly associated with ICU admission. Neurological disease (OR = 6.48; 95% CI: 3.76–11.18) and cardiovascular disease (OR = 5.19; 95% CI: 2.77–9.72) were also important predictors, with similar patterns observed for mechanical ventilation [17].

In Norway, Havdal et al. (2022) found age < 3 months (OR = 6.73; 95% CI: 2.71–16.7) to be the strongest predictor of ventilatory support, followed by comorbidity (OR = 2.40; 95% CI: 1.35–4.24) and sibling exposure (OR = 1.65; 95% CI: 1.05–2.59) [18].

Evidence from Canada (Kirolos et al., 2025) showed elevated risk for ICU admission, ventilation, or death in infants < 3 months (RR = 2.34; 95% CI: 1.43–3.84) and those aged 3–6 months (RR = 2.79; 95% CI: 1.65–4.70). Neurodevelopmental disorders (RR = 1.89; 95% CI: 1.03–3.49) and chronic pulmonary disease (RR = 3.01; 95% CI: 1.79–5.06, in older children) also contributed [19].

A smaller study in France (Mulot et al.) suggested a protective effect of corrected age in preterm infants (OR = 0.77; 95% CI: 0.62–0.93), while bronchopulmonary dysplasia and palivizumab use were not statistically significant [20].

The large Japanese cohort analyzed by Okubo et al. ($N = 50{,}482$) revealed that having ≥ 2 chronic conditions markedly increased risk (RR/OR = 6.91; 95% CI: 5.81–8.21). Respiratory disease (RR = 3.50; 95% CI: 2.71–4.53), prematurity (RR = 3.11; 95% CI: 2.14–4.54), and neuromuscular disease (RR = 2.48; 95% CI: 1.91–3.20) were also important [21].

Finally, studies from Colombia confirmed age < 6 months (Rodríguez et al., 2014: RR = 2.01; 95% CI: 1.70–2.38; Rodríguez-Martínez et al., 2022: OR = 2.94; 95% CI: 1.29–6.66) and prematurity (RR = 1.61; 95% CI: 1.20–2.17) as consistent predictors. Additional risk factors included congenital heart disease (RR = 2.03; 95% CI: 1.16–3.54; OR = 3.54; 95% CI: 1.41–8.84), adenovirus coinfection (RR = 2.09; 95% CI: 1.60–2.73), malnutrition (OR = 3.28; 95% CI: 1.15–9.37), and sepsis (OR = 45.79; 95% CI: 5.29–395.78) [8].

Taken together, the studies consistently demonstrate that young age (< 6 months) and prematurity are the most robust predictors of severe RSV outcomes, with effect estimates generally ranging from approximately 1.6 to 5.4. Comorbidities such as congenital heart disease (up to OR $\approx$ 3.5), neuromuscular or neurological disorders ($\approx$2.5–3.5), and chronic lung disease (RR up to 3.01) repeatedly amplify risk. Rare but severe conditions—including cystic fibrosis (OR = 35.13), hepatic disease (OR up to 14.99), and sepsis (OR = 45.79)—confer extreme vulnerability. Clinical features such as apnea (OR = 4.5), feeding difficulty (OR = 4.5), and fever (up to OR = 3.93), as well as exposures such as siblings (OR = 1.65) or viral coinfections (OR $\approx$1.4–2.1), add further nuance.

Figure 2 presents a graph of the most frequent risk factors identified in this literature review, considering only those evaluated through multivariate analyses with statistically significant results and reported in at least three studies.

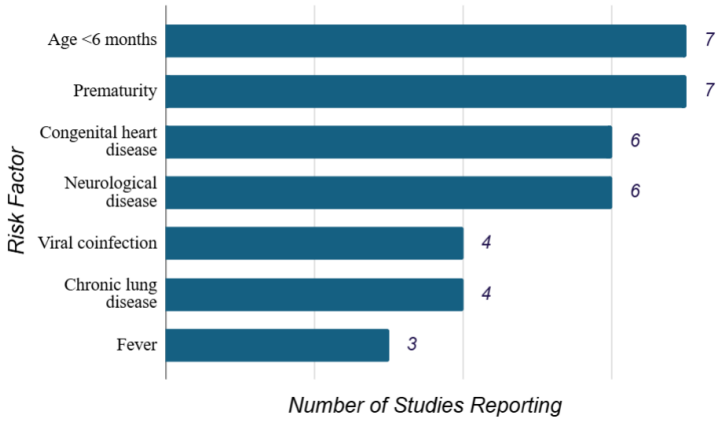


Fig. 2: Most Frequent Risk Factors for Severe RSV Outcomes

C. Risk of Bias Assessment

Risk of bias assessment using the Newcastle-Ottawa Scale is presented in Table III. Eight studies were classified as high quality (scores 7–9), while one study (Mulot et al.) was classified as moderate quality (score 6). Lower scores were mainly related to limited sample size, reduced representativeness, and fewer adjustments for potential confounding factors. Overall, the methodological quality of the included evidence was considered satisfactory.

D. Meta-Analysis

The random-effects meta-analysis demonstrated significant associations between several risk factors and severe RSV outcomes (Figure 3).

Infants aged < 6 months showed a markedly increased risk of severe outcomes (pooled OR = 3.42; 95% CI: 2.23–5.27), although substantial heterogeneity was observed across studies (I^2 = 71.9%, τ^2 = 0.1550, p = 0.0066), suggesting variability in effect estimates according to study characteristics and outcome definitions. A leave-one-out sensitivity analysis was conducted for age below six months. The pooled association remained statistically significant after sequential exclusion of each individual study, with pooled ORs ranging from 2.67 to 3.92. These findings indicate that the observed association was robust and not driven by any single study (Figure 4).

Prematurity was also significantly associated with severe RSV outcomes (pooled OR = 1.80; 95% CI: 1.30–2.49), with no observed heterogeneity (I^2 = 0.0%, τ^2 = 0, p = 0.4219), indicating highly consistent findings across studies.

Children with congenital heart disease had significantly higher odds of severe RSV outcomes (pooled OR = 2.56; 95% CI: 1.46–4.51). Moderate heterogeneity was observed (I^2 = 46.4%, τ^2 = 0.1219, p = 0.1547), suggesting some variation in effect size while maintaining a consistent direction of association.

For chronic lung disease, the pooled estimate suggested increased odds of severe RSV outcomes; however, the association did not reach statistical significance (pooled OR = 1.30; 95% CI: 0.79–2.14). No heterogeneity was detected among studies (I^2 = 0.0%, $\tau^2 < 0.0001$, p = 0.4609).

TABLE II: Characteristics of included studies and main predictors of severe RSV outcomes

Study	Country	N	Age Range	RSV Confirmation	Severe Outcome	Main Predictors
Anderson et al., 2022	Australia	970	< 2 y	Laboratory-confirmed	Ventilatory support	Age < 2 months; apnea; feeding difficulty
Brenes-Chacón et al., 2024	USA	5143	< 5 y	Virologically confirmed	ICU admission / mechanical ventilation	Age < 6 months; prematurity; neuromuscular disease
Cai et al., 2020	Germany	8761	< 5 y	Hospital RSV records	ICU admission / mechanical ventilation	Prematurity; cardiovascular disease; neurological disease
Havdal et al., 2022	Norway	1096	< 59 months	Laboratory-confirmed	Ventilatory support	Age < 3 months; comorbidity; sibling exposure
Kirolos et al., 2025	Canada	452	< 2 y	Laboratory-confirmed	Ventilatory support, ICU admission or death	Age < 6 months; prematurity; neurodevelopmental disorder
Mulot et al., 2022	France	86	Preterm infants (< 12 corrected months)	Laboratory-confirmed	Ventilatory support	Lower corrected age for prematurity
Okubo et al., 2024	Japan	50482	< 5 y	Administrative database	Ventilatory support, ICU admission or death	Multiple chronic conditions; respiratory disease; prematurity
Rodríguez et al., 2014	Colombia	2147	< 5 y	RSV-positive testing	ICU admission	Age < 6 months; prematurity; congenital heart disease
Rodríguez-Martínez et al., 2022	Colombia	1215	< 5 y	Confirmed RSV-ALRI	Ventilatory support, ICU admission or death	Sepsis; congenital heart disease; malnutrition

TABLE III: Newcastle-Ottawa Scale Quality Assessment of Included Studies

Study	Selection	Comparability	Outcome	Total	Quality
Anderson et al., 2022	4	2	2	8	High
Brenes-Chacón et al., 2024	4	2	3	9	High
Cai et al., 2020	4	2	2	8	High
Havdal et al., 2022	4	2	3	9	High
Kirolos et al., 2025	4	2	3	9	High
Mulot et al., 2022	3	1	2	6	Moderate
Okubo et al., 2024	4	2	3	9	High
Rodríguez et al., 2014	3	2	2	7	High
Rodríguez-Martínez et al., 2022	3	2	2	7	High

Overall, younger age, prematurity, and congenital heart disease demonstrated consistent associations with severe RSV outcomes. Although substantial heterogeneity remained for age below six months, the direction of effect was consistent across studies.

IV. Discussion

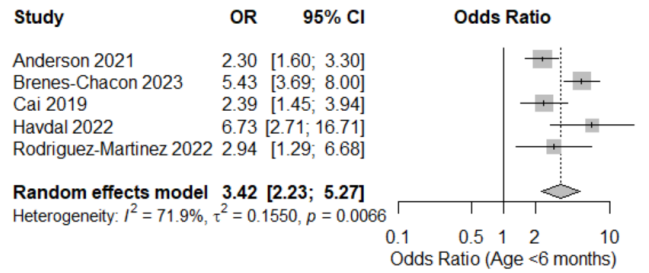
This systematic review and meta-analysis identified a consistent set of predictors associated with severe RSV outcomes in infants and young children. Across diverse geographic settings and healthcare systems, younger age, prematurity, congenital heart disease, chronic lung disease, and underlying comorbidities emerged as the most relevant determinants of disease severity. These findings reinforce the substantial vulnerability of specific pediatric subgroups and support the development of targeted prevention strategies. [1], [7], [8]

Among all factors evaluated, young age was the most consistent predictor of severe outcomes. Individual studies repeatedly demonstrated increased risks among infants younger than six months, while the pooled analysis showed that this age group was associated with more than a threefold increase in the odds of severe RSV disease. This observation is biologically plausible, given the immaturity of the infant immune system, smaller airway caliber, and limited pulmonary reserve during early life. The elevated burden observed in the first months of life also supports current strategies focused on passive immunization and maternal vaccination to provide protection during this critical period. [4], [5]

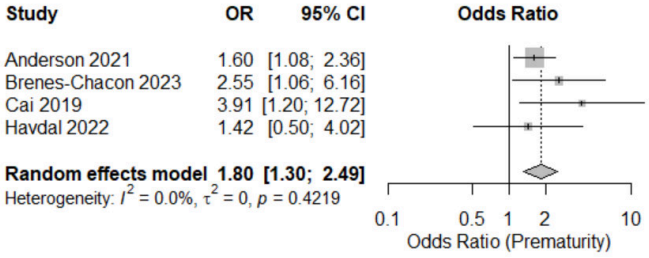
Prematurity was another recurrent predictor identified across studies and was significantly associated with severe RSV outcomes in the pooled analysis. Premature infants frequently present impaired lung development, reduced maternal antibody transfer, and increased susceptibility to respiratory complications. Although the magnitude of association varied among studies, the direction of effect was remarkably consistent, resulting in relatively low heterogeneity compared with other pooled analyses.

Congenital heart disease was consistently associated with increased risk of severe RSV outcomes. Although chronic lung disease is traditionally recognized as a high-risk condition, the pooled estimate in the present review did not demonstrate a statistically significant association, which may reflect the limited number of available studies and the imprecision of some effect estimates. These findings are consistent with the historical recognition of these populations as high-risk groups and partially explain their prioritization in preventive programs. Importantly, several studies also identified neurological, neuromuscular, and multisystem chronic conditions as major contributors to severity, suggesting that risk stratification should extend beyond the traditionally recognized high-risk groups.

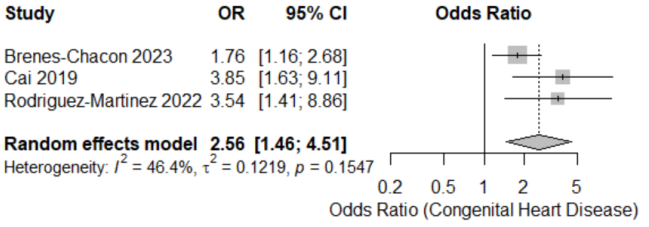
The narrative synthesis additionally highlighted several clinical characteristics associated with severe disease, includ-



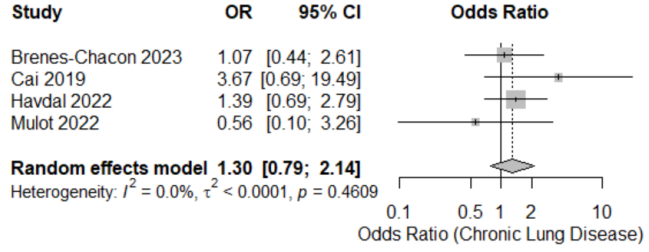
(a) Age $<$ 6 months



(b) Prematurity



(c) Congenital Heart Disease



(d) Chronic Lung Disease

Fig. 3: Forest plots summarizing pooled odds ratios for major risk factors.

ing apnea, feeding difficulty, fever, and increased work of breathing. While these factors are not suitable for preventive targeting before infection occurs, they may contribute to early identification of children at increased risk of clinical deterioration following hospitalization.

Despite the substantial heterogeneity observed for age below six months, sensitivity analysis demonstrated that the pooled association remained stable after sequential removal of individual studies, supporting the robustness of the findings.

Quality assessment using the Newcastle-Ottawa Scale indicated that most included studies were of high methodological

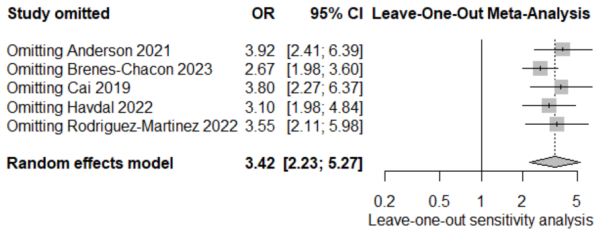


Fig. 4: Leave-one-out sensitivity analysis for age below six months.

quality. However, the observational nature of the evidence warrants cautious interpretation due to the possibility of residual confounding.

Despite its strengths, this review has some limitations. Screening and data extraction were conducted by a single reviewer, which may have introduced selection bias. Additionally, substantial heterogeneity was observed for some risk factors, likely reflecting differences in study design, population characteristics, and definitions of severe RSV outcomes. The relatively small number of studies included in some meta-analyses limited the exploration of publication bias and subgroup effects. Furthermore, the inclusion of studies from different healthcare settings may have influenced the comparability of effect estimates. These limitations should be considered when interpreting the pooled results.

Beyond supporting the identification of high-risk groups, the predictors consistently identified in this review may contribute to the development of future predictive models for severe RSV outcomes. Variables such as age below six months, prematurity, congenital heart disease, chronic lung disease, and neurological comorbidities demonstrated consistent associations with disease severity across different settings and therefore represent potential candidate predictors for risk stratification tools. Future studies should evaluate the performance of multivariable prediction models integrating these factors to support early identification of children at increased risk of severe disease

V. Conclusion

This systematic review and meta-analysis demonstrated that age below six months, prematurity, and congenital heart disease are important predictors of severe RSV outcomes in children. Additional risk was observed among patients with neurological, neuromuscular, and multiple chronic conditions.

The pooled analyses confirmed significantly increased odds of severe disease among infants younger than six months, premature infants, and children with congenital heart disease. These findings may help identify populations at increased risk of severe RSV outcomes and support future research on preventive strategies and clinical risk assessment tools.

As new RSV prevention technologies become increasingly available, understanding the factors associated with severe disease may contribute to improving risk awareness and informing future predictive models in pediatric populations.

VI. ACKNOWLEDGMENT

The authors acknowledge the Graduate Program in Biomedical Engineering (PPG-EBM) at the Federal University of ABC (UFABC) for supporting the development of this research.

REFERENCES

[1] Y. Li, D. Hodgson, X. Wang, K. E. Atkins, D. R. Feikin, and H. Nair, "Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis," *Lancet*, vol. 399, 2022, doi: 10.1016/S0140-6736(22)00478-0.

[2] L. C. Paramore, V. Ciuryla, G. Ciesla, *et al.*, "Economic impact of respiratory syncytial virus-related illness in the US: an analysis of national databases," *Pharmacoeconomics*, vol. 22, no. 5, pp. 275–284, 2004.

[3] M. T. Brady, C. L. Byington, H. D. Davies, *et al.*, "Updated guidance for palivizumab prophylaxis among infants and young children at increased risk of hospitalization for respiratory syncytial virus infection," *Pediatrics*, vol. 134, pp. e620–e638, 2014.

[4] E. A. F. Simões, S. A. Madhi, W. J. Muller, *et al.*, "Efficacy of nirsevimab against respiratory syncytial virus lower respiratory tract infections in preterm and term infants, and pharmacokinetic extrapolation to infants with congenital heart disease and chronic lung disease: a pooled analysis of randomised controlled trials," *Lancet Child Adolesc. Health*, vol. 7, pp. 180–189, 2023.

[5] B. Kampmann, S. A. Madhi, I. Munjal, *et al.*, "Bivalent prefusion F vaccine in pregnancy to prevent RSV illness in infants," *N. Engl. J. Med.*, vol. 388, pp. 1451–1464, 2023.

[6] Y. Li, D. Hodgson, X. Wang, K. E. Atkins, D. R. Feikin, and H. Nair, "Respiratory syncytial virus seasonality and prevention strategy planning for passive immunisation of infants in low-income and middle-income countries: a modelling study," *Lancet Infect. Dis.*, vol. 21, pp. 1303–1312, 2021.

[7] G. Dallagiacoma, C. Lundholm, A. I. Smew, *et al.*, "Risk factors for severe outcomes of respiratory syncytial virus infection in children: a nationwide cohort study in Sweden," *The Lancet Regional Health – Europe* .

[8] C. E. Rodríguez-Martínez and R. Acuña-Cordero, "Predictors of poor outcomes of respiratory syncytial virus acute lower respiratory infections in children under 5 years of age in a middle-income tropical country based on the National Public Health Surveillance System," *Wiley Periodicals LLC – Pediatric Pulmonology*.

[9] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.

[10] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews," *Syst. Rev.*, vol. 5, no. 210, 2016.

[11] G. A. Wells, B. Shea, D. O'Connell, J. Peterson, V. Welch, M. Losos, and P. Tugwell, "The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses," Ottawa Hospital Research Institute, Ottawa, Canada. [Online]. Available: https://www.ohri.ca/en/who-we-are/core-facilities-and-platforms/ottawa-methods-centre/newcastle-ottawa-scale

[12] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2025. [Online]. Available: https://www.r-project.org/

[13] G. Schwarzer, J. R. Carpenter, and G. Rücker, "meta: An R package for meta-analysis," *R package version*, 2023. [Online]. Available: https://cran.r-project.org/package=meta

[14] W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," *J. Stat. Softw.*, vol. 36, no. 3, pp. 1–48, 2010.

[15] J. Anderson, M. Oeum, E. Verkolf, *et al.*, "Factors associated with severe respiratory syncytial virus disease in hospitalised children: a retrospective analysis," *Arch. Dis. Child.*, vol. 107, pp. 359–364, 2022.

[16] H. Brenes-Chacón, M. Eisner, S. Acero-Bedoya, *et al.*, "Age-specific predictors of disease severity in children with respiratory syncytial virus infection beyond infancy and through the first 5 years of age," *Pediatr. Allergy Immunol.*, vol. 35, e14083, 2024.

[17] W. Cai, S. Buda, E. Schuler, *et al.*, "Risk factors for hospitalized respiratory syncytial virus disease and its severe outcomes," *Influenza Other Respir. Viruses*, vol. 14, pp. 658–670, 2020.

[18] L. B. Havdal, H. Bøås, T. Bekkevold, et al., "Risk factors associated with severe disease in respiratory syncytial virus infected children under 5 years of age," *Front. Pediatr.*, vol. 10, Art. no. 1004739, 2022, doi: 10.3389/fped.2022.1004739.

[19] N. Kirolos, H. Mtaweh, R. R. Datta, et al., "Risk factors for severe disease among children hospitalized with respiratory syncytial virus," *JAMA Netw. Open*, vol. 8, no. 4, Art. no. e254666, 2025, doi: 10.1001/jamanetworkopen.2025.4666.

[20] G. Mulot, M. Benchaib, F. Plaisant, et al., "Risk factors of very severe RSV infections in a multicenter cohort of very preterm and extreme preterm babies receiving or not palivizumab," *Front. Pediatr.*, vol. 10, Art. no. 884120, 2022, doi: 10.3389/fped.2022.884120.

[21] Y. Okubo, K. Uda, C. Ogimi, R. Shimabukuro, and K. Ito, "Clinical practice patterns and risk factors for severe conditions in pediatric hospitalizations with respiratory syncytial virus in Japan: A nationwide analyses (2018–2022)," *Pediatr. Infect. Dis. J.*, vol. 43, no. 3, pp. 187–193, 2024, doi: 10.1097/INF.0000000000004181.

Protótipo de Sistema Especialista em CLIPS para Apoio à Detecção de Não Conformidades Ergonômicas Baseadas na NR-17

1st Bienvenido Esono Mengue Ebang
Faculdade de Ciências e Tecnologias em Engenharia
Universidade de Brasília (UnB)
Brasília, Brasil
211004166@aluno.unb.br

2nd Dr. Gerardo Antonio Idrobo Pizo
Faculdade de Ciências e Tecnologias em Engenharia
Universidade de Brasília (UnB)
Brasília, Brasil
gerardo_idrobo@unb.br

***Abstract*—Este artigo apresenta o desenvolvimento e a avaliação preliminar de um protótipo de sistema especialista baseado em regras, implementado em CLIPS (C Language Integrated Production System), para apoio à análise ergonômica de postos de trabalho. O sistema integra uma base de conhecimento derivada de itens selecionados da Norma Regulamentadora NR-17 (Ergonomia) com uma interface de aquisição de dados em Python. Diferentemente de modelos baseados em aprendizado profundo de caixa-preta, a abordagem em regras prioriza a explicabilidade e a rastreabilidade jurídica, emitindo diagnósticos diretamente associados às cláusulas regulamentares. Testes preliminares em múltiplos cenários simulados demonstraram a consistência da árvore de inferência e a capacidade de resolução de conflitos por prioridades. O trabalho atua como prova de conceito e ferramenta de triagem preliminar, fornecendo diretrizes estruturadas para futuras validações clínicas e ocupacionais com especialistas.**

***Index Terms*—Sistemas especialistas, CLIPS, ergonomia ocupacional, NR-17, saúde do trabalhador.**

I. Introdução

Na área de saúde e segurança ocupacional, os Distúrbios Osteomusculares Relacionados ao Trabalho (DORT) configuram uma das principais causas de absenteísmo e incapacidade laboral crônica. A detecção precoce de inadequações posturais e ambientais é essencial para mitigar o desenvolvimento crônico destas lesões, exigindo o emprego de ferramentas de triagem eficientes [10]. No Brasil, a Norma Regulamentadora NR-17 estabelece os parâmetros técnicos para a adaptação das condições de trabalho às características psicofisiológicas dos trabalhadores. Contudo, a realização da Análise Ergonômica do Trabalho (AET) depende substancialmente de inspeções manuais. Segundo a literatura clássica de ergonomia brasileira, esse processo analítico exige elevada especialização, configurando-se como uma atividade custosa, que demanda tempo prolongado de observação *in loco* e está sujeita à variabilidade interavaliador [4].

Nesse cenário, os sistemas especialistas baseados em regras representam uma solução determinística e transparente. Estudos recentes no contexto nacional apontam que a aplicação de sistemas especialistas na área da saúde otimiza a triagem diagnóstica e padroniza condutas clínicas [5].

Este artigo propõe o desenvolvimento e a avaliação de um protótipo de sistema especialista implementado em linguagem CLIPS acoplado a uma interface Python para suporte à triagem ergonômica preliminar baseada na NR-17 [2]. O escopo do trabalho assume estritamente o caráter de prova de conceito técnica, visando verificar a consistência das regras, a resolução de conflitos na agenda de execução e a rastreabilidade normativa, sem a pretensão de substituir a avaliação integral conduzida por ergonomistas certificados.

II. Metodologia

A arquitetura do sistema divide-se em três etapas: modelagem da base de conhecimento a partir da NR-17, codificação do motor de regras no CLIPS e implementação da camada de entrada de dados em Python.

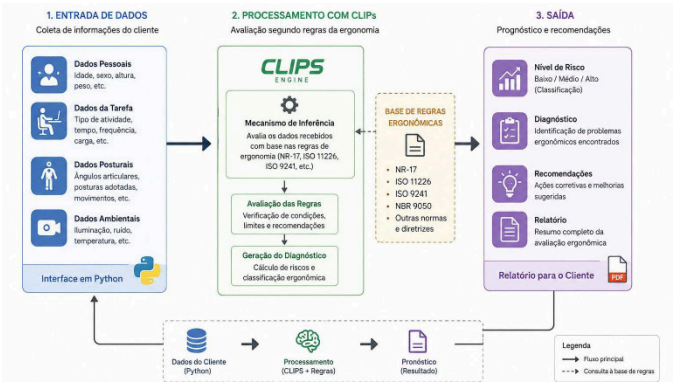


Fig. 1. Arquitetura proposta para o Sistema de Avaliação Ergonômica, detalhando o fluxo de integração entre a interface de aquisição de dados em Python, o motor de inferência CLIPS fundamentado em normas técnicas e a geração do relatório diagnóstico.

A. Engenharia de Conhecimento e Rastreabilidade Normativa

A base de conhecimento foi estruturada pela tradução de cláusulas selecionadas da NR-17 e de seus anexos específicos em proposições lógicas. Os limiares de alcance horizontal que disparam os alertas posturais no motor de

inferência foram definidos cruzando as restrições da NR-17 com dados antropométricos clássicos [11]. Adicionalmente, a interpretação técnica impressa nos alertas de temperatura e iluminação baseia-se nos impactos fisiológicos documentados na literatura ergonômica fundamental [12]. Para assegurar a auditabilidade do sistema e provar sua aderência à norma, a Tabela I detalha a matriz de rastreabilidade implementada no CLIPS, abrangendo os limites exatos de conforto ambiental e exigências antropométricas.

TABLE I
MATRIZ DE RASTREABILIDADE NORMATIVA (NR-17 ATUALIZADA) E HEURÍSTICA

Fonte / Norma	Variável de Entrada	Critério Adotado	Tipo de Saída / Alerta
NR-17 (17.5.1)	Carga manual (Peso)	$>$ 25 kg (Limiar NIOSH)	Alerta Heurístico de Sobrecarga
NR-17 (17.5.2.1)	Distância de alcance	$>$ 60 cm	Alerta Normativo (Risco Postural)
NR-17 (17.6.1)	Mobiliário	Sem ajuste antropométrico	Alerta Normativo (Postura/Lombar)
NR-17 / NHO-11 [3]	Iluminância	$<$ 300 lux	Alerta Normativo (Fadiga Ocular)
NR-17 (17.8.4.1.2)	Ruído de fundo	$>$ 65 dB(A)	Alerta Normativo (Estresse Auditivo)
NR-17 (17.8.4.2)	Temperatura ambiente	Fora da faixa de 18–25°C	Alerta Normativo (Desconforto)
NR-17 (17.4.3.1)	Pausas de recuperação	Atividade repetitiva sem pausa	Alerta Normativo (Risco de DORT)
NR-17 (Anexo I)	Operadores de Checkout	Sem ensacador / Mobiliário fixo	Alerta de Sobrecarga (Específico)
NR-17 (Anexo II)	Telemarketing	Sem headset / Sem pausa	Alerta Ergonômico (Específico)
Critério Clínico	Dor subjetiva (0–10)	Escore ≥ 4	Triagem Preventiva (Sem infração)

Ressalta-se que a NR-17 veda o transporte manual de cargas que comprometa a integridade física do trabalhador sem fixar um limite numérico estático de peso. Para viabilizar a inferência computacional, o limiar de 25 kg foi adotado nesta prova de conceito como regra heurística de triagem, baseado nos parâmetros de segurança da Equação de Levantamento do NIOSH [13], devendo ser calibrado parametricamente em estudos futuros. Adicionalmente, o sistema destaca-se por incorporar as resoluções dos Anexos I e II da referida norma, garantindo diagnósticos contextualizados para funções críticas como teleatendimento e caixas de supermercado. A estruturação dos níveis de severidade dos alertas gerados inspira-se em metodologias consagradas de avaliação de risco biomecânico [15]. Para a variável ambiental de luminância, o sistema utiliza como limiar técnico a NHO-11 da Fundacentro [3], uma vez que a atual redação da NR-17 delega a esta norma a fixação dos valores quantitativos adequados a cada atividade.

B. Implementação no CLIPS e Interface Python via Visão Computacional

A arquitetura computacional do protótipo é dividida em duas camadas operacionais operando em tempo real: aquisição de dados via Visão Computacional (Python) e inferência simbólica (CLIPS).

A camada de aquisição atua como a interface de monitoramento do trabalhador. Utilizando as bibliotecas OpenCV e MediaPipe [8], o sistema mapeia a cinemática do usuário por meio de inferência de marcos articulares (esqueleto digital). O algoritmo calcula matematicamente os ângulos de flexão e inclinação de segmentos corporais críticos (pescoço, tronco, ombro, braço e punho). Simultaneamente, dados organizacionais (como carga levantada e pausas) são informados no sistema.

Esses parâmetros angulares e ambientais são exportados dinamicamente para um arquivo no formato `.facts`, populando a memória de trabalho estática inicial do sistema de acordo com os moldes estruturados (`deftemplate`):

```
(deftemplate dados
(slot nome (type STRING))
(slot repetitiva (type SYMBOL))
(slot braco_direito (type INTEGER))
(slot pescoco (type INTEGER))
(slot tronco (type INTEGER))
(slot carga (type INTEGER))
(slot distancia_alcance (type INTEGER))
(slot pausas (type SYMBOL))
(slot iluminacao (type INTEGER))
(slot temperatura (type INTEGER))
(slot ruido (type INTEGER))
(slot funcao (type SYMBOL))
(slot ensacador (type SYMBOL))
(slot headset (type SYMBOL))
)
```

A modelagem dos fatos posturais (como os ângulos de braço e pescoço capturados pela IA) foi estruturada visando compatibilidade direta com métodos consolidados de avaliação rápida de postura dos membros superiores, como o método RULA [14].

Uma vez que o Python atualiza o arquivo de fatos, o motor de inferência assume o processamento. O encadeamento para frente (*forward chaining*) orienta as deduções lógicas. Para evitar inconsistências no diagnóstico, a propriedade `salience` organiza a hierarquia de execução, garantindo que as variáveis biomecânicas mais críticas sejam processadas de forma prioritária [1].

C. Extensões Futuras (Lógica Difusa e Aprendizado de Máquina)

A integração de Lógica Difusa (*fuzzy*) para tratamento de transições posturais graduais e classificadores *Random Forest* para sinais fisiológicos constitui proposta de extensão futura, não integrando a base de testes da atual prova de conceito.

III. RESULTADOS E DISCUSSÃO

O protótipo foi submetido a uma bateria de testes sintéticos cobrindo doze cenários operacionais para averiguar a sensibilidade das regras, a ausência de disparos indevidos (falsos positivos) e o funcionamento das regras específicas de função (Anexos I e II). A Tabela II resume os casos testados.

TABLE II
RESULTADOS DA AVALIAÇÃO EM CENÁRIOS SIMULADOS COM BASE NA ÁRVORE DE INFERÊNCIA

Cenário	Parâmetros Simulados	Diagnóstico Emitido pelo CLIPS
C1	Posto Padrão (Carga=15kg, Pausa=sim, Temp=22°C)	Nenhum alerta gerado (Conforme)
C2	Carga = 28 kg (Heurística $>$ 25 kg)	Alerta NR-17.5.1: Reduzir peso
C3	Distância de Alcance = 70 cm ($>$ 60 cm)	Alerta NR-17.5.2.1: Alcance excessivo
C4	Atividade Repetitiva = sim, Pausas = não	Alerta NR-17.4.3.1: Ausência de pausas
C5	Mobiliário Ajustável = não	Alerta NR-17.6.1: Risco postural/lombar
C6	Iluminação = 250 lux ($<$ 300 lux)	Alerta NR-17.8.3: Iluminação inadequada
C7	Ruído de Fundo = 75 dB(A) ($>$ 65 dB)	Alerta NR-17.8.4.1.2: Estresse auditivo
C8	Temperatura = 16°C ou 28°C (Fora de 18–25°C)	Alerta NR-17.8.4.2: Desconforto térmico
C9	Função = checkout, Ensacador = não	Alerta Anexo I (4.3): Sobrecarga do operador
C10	Função = checkout, Mobiliário Ajustável = não	Disparo múltiplo: Anexo I (3.1) e NR-17.6.1
C11	Função = telemarketing, Headset = não	Alerta Anexo II: Risco de cervicalgia
C12	Múltiplas: Telemarketing, Sem pausa, Ruído=70	Disparo conjunto de 3 alertas específicos



Fig. 2. Interface gráfica/console desenvolvida em Python para aquisição dos parâmetros ergonômicos e geração dos fatos.

```
(dados
   (nome "Bienvenido Esono")
   (idade 26)
   (sexo M)
   (repetitiva sim)
   (carga 0)
   (distancia_alcance 0)
   (pausas sim)
   (mobiliario_ajustavel sim)
   (iluminacao 400)
   (ruido 50)
   (temperatura 22)
   (funcao geral)
   (ensacador nao)
   (headset sim)
   (braco_direito 32)
   (braco_esquerdo 30)
   (pescoco 178)
```

```
======================================================================
[DIAGNÓSTICO ERGONÔMICO COMPLETO] - RELATÓRIO OFICIAL
======================================================================
Trabalhador avaliado: Bienvenido Esono | Função: geral

◇ 1. AVALIAÇÃO POSTURAL BIOMECÂNICA (IA e Visão Computacional):
   - Ângulo Braço Direito: 32° -> ✅ ADEQUADO
   - Ângulo Braço Esquerdo: 30° -> ✅ ADEQUADO
   - Ângulo do Pescoço: 178° -> ✅ ADEQUADO

◇ 2. AVALIAÇÃO AMBIENTAL (Conforto Físico):
   - Iluminação: 400 lux -> ✅ ADEQUADA
   - Ruído de Fundo: 50 dB(A) -> ✅ ADEQUADO
   - Temperatura: 22 °C -> ✅ ADEQUADA

◇ 3. ORGANIZAÇÃO DO TRABALHO E MOBILIÁRIO:
   - Movimentação de Carga: 0 Kg -> ✅ SEGURA
   - Alcance de Carga: 0 cm -> ✅ SEGURO
   - Ajuste de Mobiliário: SIM -> ✅ CONFORME
   - Pausas de Descanso: SIM -> ✅ CONFORME

======================================================================
📝 CONCLUSÃO DO SISTEMA ESPECIALISTA:
✅ O posto de trabalho e a postura cinemática do usuário encontram-se
   100% em conformidade com as exigências legais vigentes da NR-17.
======================================================================
```

Fig. 3. Teste do primeiro cenário: avaliação postural e ambiental adequada, resultando na emissão de laudo ergonômico positivo pelo sistema especialista.

.

```
(dados
   (nome "Bienvenido Esono")
   (idade 26)
   (sexo M)
   (repetitiva sim)
   (carga 30)
   (distancia_alcance 65)
   (pausas nao)
   (mobiliario_ajustavel nao)
   (iluminacao 150)
   (temperatura 30)
   (funcao geral)
   (ensacador nao)
   (headset nao)
   (braco_direito 56)
   (braco_esquerdo 43)
   (pescoco 150)
```

```
======================================================================
[DIAGNÓSTICO ERGONÔMICO COMPLETO] - RELATÓRIO OFICIAL
======================================================================
Trabalhador avaliado: Bienvenido Esono | Função: geral

◇ 1. AVALIAÇÃO POSTURAL BIOMECÂNICA (IA e Visão Computacional):
   - Ângulo Braço Direito: 56° -> ⚠ ALERTA: Elevado! (NR-17.6.3 a)
   - Ângulo Braço Esquerdo: 43° -> ✅ ADEQUADO
   - Ângulo do Pescoço: 150° -> ⚠ ALERTA: Flexão acentuada! (NR-17.4.3 a)

◇ 2. AVALIAÇÃO AMBIENTAL (Conforto Físico):
   - Iluminação: 150 lux -> ⚠ ALERTA: Abaixo do mínimo de 300 lux (NR-17.8.3)
   - Ruído de Fundo: 0 dB(A) -> ✅ ADEQUADO
   - Temperatura: 30 °C -> ⚠ ALERTA: Fora da faixa 18-25 °C (NR-17.8.4.2)

◇ 3. ORGANIZAÇÃO DO TRABALHO E MOBILIÁRIO:
   - Movimentação de Carga: 30 Kg -> ⚠ ALERTA: Excesso de peso (NR-17.5.4)
   - Alcance de Carga: 65 cm -> ⚠ ALERTA: Alcance excessivo > 60cm (NR-17.5.2.1)
   - Ajuste de Mobiliário: NÃO -> ⚠ ALERTA: Falta regulagem (NR-17.6.1)
   - Pausas de Descanso: NÃO -> ⚠ ALERTA: Pausas necessárias (NR-17.4.3.1 a)

======================================================================
📝 CONCLUSÃO DO SISTEMA ESPECIALISTA:
❌ FORAM DETECTADAS INFRAÇÕES ERGONÔMICAS!
   O posto de trabalho necessita de adequações imediatas para
   cumprir a NR-17 (Verifique os itens com '⚠ ALERTA' acima).
======================================================================
```

Fig. 4. Teste do segundo cenário: detecção de posturas nocivas e inconformidades ambientais, resultando na emissão de alertas estruturados pelo sistema especialista baseados na NR-17.

A análise do tempo de execução confirmou uma latência média de processamento de aproximadamente 23 ms por posto analisado, atestando viabilidade para sistemas de monitoramento contínuo.

A. *Discussão e Limitações do Protótipo*

O principal ganho metodológico do sistema reside na sua rastreabilidade determinística: para cada alerta, o operador recebe o trecho exato da norma violada e a interpretação biomecânica correspondente. Diferentemente de modelos conexionistas, o sistema não opera como caixa-preta [9].

Como limitações da presente etapa, pontua-se: 1) A base de conhecimento abrange apenas uma parcela dos requisitos previstos na NR-17. 2) A validação ateve-se a cenários gerados sinteticamente, não substituindo a coleta in loco. 3) Variáveis contínuas limítrofes (como temperatura de 17,8°C) exigem a incorporação futura de inferência por lógica difusa (*fuzzy*) para modelagem de incertezas.

IV. Conclusão

O protótipo desenvolvido comprovou a viabilidade técnica da utilização do CLIPS para a estruturação de um sistema especialista explicável aplicado à triagem de não conformidades

da NR-17. A separação clara entre parâmetros normativos mandatórios e heurísticas complementares forneceu robustez à arquitetura.

Trabalhos futuros focarão na expansão integral da base de regras normativas, na integração da biblioteca `clipspy` para inferência difusa e na realização de estudos de validação clínica em ambientes ocupacionais reais, avaliando a concordância do sistema frente a avaliações periciais especializadas via Coeficiente Kappa de Cohen.

AGRADECIMENTOS

Os autores agradecem à Universidade de Brasília (UnB), ao Programa de Iniciação Científica PIBIC/UnB e à FAPDF pelo suporte institucional que viabilizou este projeto.

REFERÊNCIAS

[1] CLIPS Rules, "CLIPS: a tool for building expert systems," 2023. [Online]. Disponível em: https://www.clipsrules.net

[2] Brasil. Ministério do Trabalho e Previdência, "NR-17: Ergonomia," Portaria MTP nº 423, de 7 de outubro de 2021. Brasília, DF, 2022.

[3] Fundacentro, *Norma de Higiene Ocupacional NHO 11: Avaliação dos Níveis de Iluminamento em Ambientes de Trabalho Internos*. São Paulo, SP, Brasil: Fundacentro, 2018.

[4] I. Iida and L. I. A. Guimarães, *Ergonomia: Projeto e Produção*, 3. ed. São Paulo, SP, Brasil: Edgard Blücher, 2016.

[5] R. O. Araujo, "A aplicação de sistemas especialistas na área de saúde: uma revisão sistemática," Universidade Federal de Alagoas (UFAL), Arapiraca, AL, Brasil, 2021.

[6] S. Pavlovic-Veselinovic, A. Hedge, and M. Veselinovic, "An ergonomic expert system for risk assessment of work-related musculo-skeletal disorders," *International Journal of Industrial Ergonomics*, vol. 53, pp. 130–139, 2016.

[7] T. Petrat, "Review of ergonomic methodologies and expert systems," *Human Factors and Ergonomics Review*, vol. 29, no. 4, pp. 345–360, 2021.

[8] M. Lutz, *Learning Python*, 5th ed. Sebastopol, CA, USA: O'Reilly Media, 2013.

[9] J. Amann et al., "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, 2020.

[10] K. H. E. Kroemer, *Fitting the Human: Introduction to Ergonomics / Human Factors Engineering*, 7th ed. Boca Raton, FL, USA: CRC Press, 2017.

[11] S. Pheasant and C. M. Haslegrave, *Bodyspace: Anthropometry, Ergonomics and the Design of Work*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2005.

[12] J. Dul and B. Weerdmeester, *Ergonomics for Beginners: A Quick Reference Guide*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2008.

[13] T. R. Waters, V. Putz-Anderson, A. Garg, and L. J. Fine, "Revised NIOSH equation for the design and evaluation of manual lifting tasks," *Ergonomics*, vol. 36, no. 7, pp. 749–776, 1993.

[14] L. McAtamney and E. N. Corlett, "RULA: a survey method for the investigation of work-related upper limb disorders," *Applied Ergonomics*, vol. 24, no. 2, pp. 91–99, 1993.

[15] O. Karhu, P. Kansi, and I. Kuorinka, "Correcting working postures in industry: A practical method for analysis," *Applied Ergonomics*, vol. 8, no. 4, pp. 199–201, 1977.

Spectral Preprocessing for Machine Learning Classification of Salivary FTIR Spectra in Oral Squamous Cell Carcinoma

1st Allison dos Santos
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
allison.santos@aluno.ufabc.edu.br

2nd Herculano da Silva Martinho
Centro de Ciências Naturais e Humanas
Federal University of ABC (UFABC)
Santo André, SP, Brazil
herculano.martinho@ufabc.edu.br

3nd Gabrielle L. J. T. Nepomuceno
Centro de Ciências Naturais e Humanas
Federal University of ABC (UFABC)
Santo André, SP, Brazil
g.luana@ufabc.edu.br

4rd Pedro H. Pires da Silva
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
phpsilva@aluno.ufabc.edu.br

5th Fernanda Almeida
LaBIC - CECS
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
fernanda.almeida@ufabc.edu.br

***Abstract*—Oral Squamous Cell Carcinoma (OSCC) is the most common form of oral cancer, and early diagnosis plays a critical role in improving patient prognosis. Fourier Transform Infrared (FTIR) spectroscopy combined with machine learning has emerged as a promising non-invasive approach for identifying disease-related biochemical alterations in biological samples. This study evaluated the influence of different spectral preprocessing strategies on the classification of salivary FTIR spectra obtained from healthy individuals and patients with OSCC. The proposed computational workflow combined baseline correction (IASLS, ModPoly, and AirPLS), Savitzky–Golay smoothing, second-derivative transformation, spectral region selection, Principal Component Analysis (PCA), and supervised machine learning algorithms, including Linear Discriminant Analysis, Random Forest, Support Vector Machine, and K-Nearest Neighbors. Model performance was assessed using repeated stratified five-fold cross-validation with ten repetitions. The results showed that second-derivative preprocessing exerted the greatest influence on predictive performance by enhancing subtle spectral features, whereas PCA provided comparatively limited benefits for the evaluated dataset. Random Forest achieved the most robust performance across the investigated preprocessing configurations. Permutation feature importance analysis indicated that the most discriminative information was concentrated within the fingerprint region, particularly in spectral bands associated with proteins, lipids, carbohydrates, and nucleic-acid-related molecular vibrations. Overall, the findings demonstrate that optimizing spectral preprocessing is as important as classifier selection for developing robust and reproducible FTIR-based diagnostic workflows and reinforce the potential of salivary FTIR spectroscopy combined with machine learning as a rapid, non-invasive approach to support the diagnosis of Oral Squamous Cell Carcinoma.**

***Index Terms*—Oral Squamous Cell Carcinoma, FTIR spectroscopy, Saliva, Spectral preprocessing, Machine learning, Random Forest.**

This work was supported by the Programa Pesquisando Desde o Primeiro Dia (PDPD) through an undergraduate research scholarship granted by the Pró-Reitoria de Pesquisa of the Federal University of ABC (Propes/UFABC), Brazil.

I. Introduction

Oral Squamous Cell Carcinoma (OSCC) is the most common malignant neoplasm of the oral cavity, accounting for approximately 90% of oral cancer cases worldwide. Because patient prognosis is strongly associated with the stage at diagnosis, early detection remains one of the main factors influencing treatment success and survival. Conventional diagnosis relies primarily on clinical examination followed by histopathological analysis, procedures that may delay definitive diagnosis in some situations. Consequently, there is growing interest in complementary diagnostic approaches based on minimally invasive biological samples.

Among these biological matrices, saliva has attracted considerable attention because it can be collected easily, non-invasively, and at low cost while containing molecular information associated with both local and systemic diseases [1]. Fourier Transform Infrared Spectroscopy (FTIR) has emerged as a promising analytical technique for investigating these biochemical alterations due to its ability to characterize biological samples through their vibrational molecular signatures [2]–[4]. In addition to being rapid and relatively inexpensive, FTIR requires minimal sample preparation and can simultaneously capture information from multiple classes of biomolecules, making it particularly attractive for biomedical applications.

FTIR spectroscopy has been widely applied to the analysis of biological materials, including saliva, blood, urine, and cerebrospinal fluid [4]. In salivary samples, previous studies have demonstrated its applicability in the investigation of psoriasis [1], Burning Mouth Syndrome [5], peri-implantitis [6], COVID-19 [7], and biochemical alterations associated with smoking and illicit drug use [8], [9]. More recently,

salivary FTIR spectroscopy has also been explored for the identification of biochemical signatures associated with OSCC progression and staging. Collectively, these studies highlight the versatility of FTIR spectroscopy for identifying disease-related biochemical alterations. However, they also reveal considerable variability in the computational strategies adopted for spectral preprocessing and data analysis.

Reliable interpretation of FTIR spectra depends on appropriate preprocessing to minimize baseline distortions, instrumental noise, and signal intensity variations before exploratory or predictive analyses are performed [11], [12]. Common preprocessing procedures include baseline correction, smoothing, normalization, derivative transformation, spectral region selection, and dimensionality reduction. Although these techniques are routinely employed in FTIR studies, there is still no consensus regarding the most appropriate combination for different biomedical applications. Since machine learning models learn directly from the spectral representation of the data, preprocessing choices may substantially influence predictive accuracy, model robustness, and reproducibility.

Machine learning has become an important component of FTIR-based biomedical studies because it enables the automatic identification of spectral patterns that are often imperceptible through visual inspection alone. Most published investigations, however, focus primarily on evaluating classification algorithms or identifying potential biomarkers, whereas comparatively fewer studies systematically assess the impact of preprocessing decisions on model performance. As a result, the contribution of individual preprocessing steps to the final predictive outcome remains insufficiently understood, limiting the reproducibility and comparability of FTIR-based computational workflows.

Motivated by this gap, the present study evaluates the influence of different spectral preprocessing strategies on the classification of salivary FTIR spectra obtained from healthy individuals and patients diagnosed with Oral Squamous Cell Carcinoma. Different combinations of baseline correction methods, spectral smoothing, second-derivative transformation, spectral region selection, Principal Component Analysis (PCA), and supervised machine learning algorithms were investigated. By systematically comparing these preprocessing pipelines, this study aims to identify strategies that improve spectral discrimination and predictive performance while contributing to the development of more robust and reproducible computational workflows for FTIR-based salivary analysis.

II. Materials and Methods

A. Dataset

The experiments were conducted using a previously published dataset composed of 158 salivary Fourier Transform Infrared (FTIR) spectra, including 77 spectra obtained from healthy controls and 81 spectra obtained from patients diagnosed with Oral Squamous Cell Carcinoma (OSCC) [10]. The spectra and associated class labels were derived from the study reported by Jimenez et al. [10], which investigated salivary biochemical markers associated with OSCC staging using FTIR spectroscopy.

To investigate the influence of spectral region selection on classification performance, two spectral configurations were evaluated. The first considered the complete spectral range (700–6000 cm^{-1}), whereas the second was restricted to the fingerprint region (900–1800 cm^{-1}), which contains absorption bands associated with proteins, lipids, nucleic acids, carbohydrates, and other biomolecules commonly explored in FTIR-based biomedical studies [4].

The present work represents a secondary computational analysis of a previously published dataset. Individual-level metadata linking spectra to specific participants were not available during the computational analyses. Consequently, the number of spectra acquired per participant, the existence of replicated measurements, and the possibility of performing patient-wise data partitioning could not be assessed. Therefore, each spectrum was treated as an independent observation throughout the study.

Detailed information regarding participant recruitment, sample collection procedures, FTIR instrumentation, spectral acquisition parameters, and ethical approval is available in the original publication describing the dataset [10]. The present study focused exclusively on evaluating the impact of different spectral preprocessing strategies and machine learning models on classification performance using the published spectra.

B. Computational Workflow

The entire computational workflow was implemented in Python (version 3.14.3). Figure 1 summarizes the computational workflow adopted in this study. Starting from the raw FTIR spectra, the data were sequentially subjected to spectral preprocessing, spectral region selection, dimensionality reduction, supervised classification, model training and validation, and performance evaluation.

Different preprocessing strategies were combined with the two spectral region configurations (full spectrum and fingerprint region). The resulting datasets were either used directly or transformed using Principal Component Analysis (PCA) before training the supervised machine learning models. Each combination of preprocessing strategy, spectral region, and classification algorithm defined an independent computational pipeline, allowing their individual contributions to classification performance to be systematically evaluated.

C. Spectral Preprocessing

Spectral preprocessing was performed before exploratory analyses and supervised classification to minimize baseline distortions, instrumental noise, and signal intensity variations commonly observed in FTIR spectra [11], [12].

Three baseline correction methods were evaluated using the Pybaselines library: Improved Asymmetric Least Squares (IASLS), Modified Polynomial Fitting (ModPoly) with a third-order polynomial, and Adaptive Iteratively Reweighted Penalized Least Squares (AirPLS). These algorithms estimate the

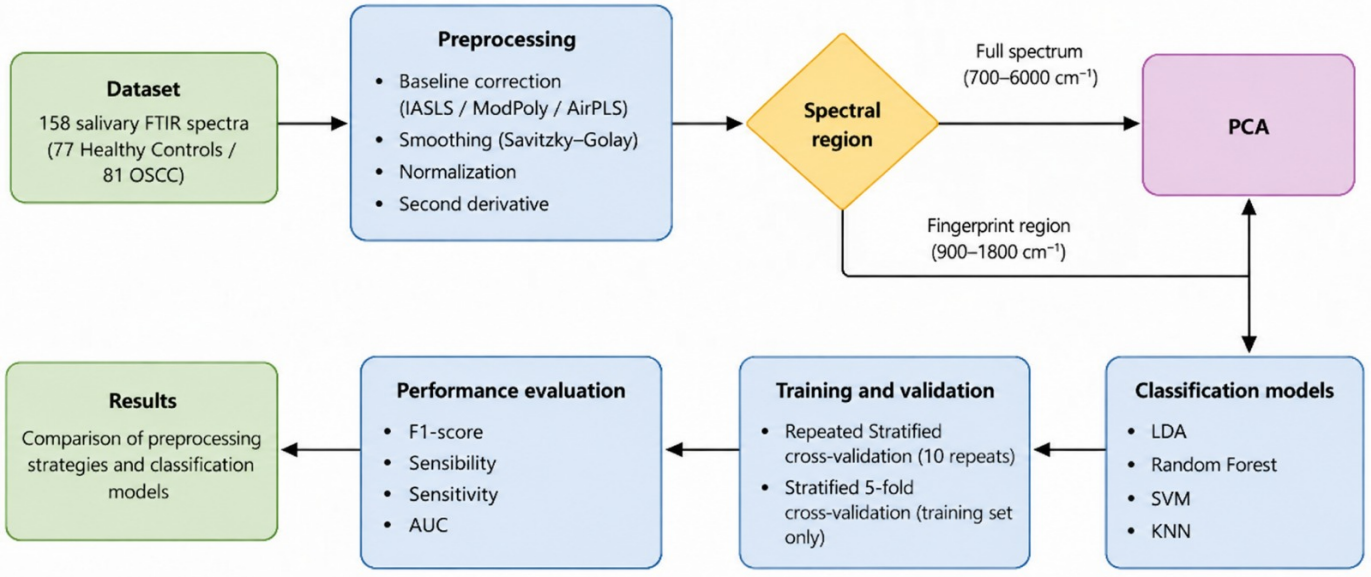


Fig. 1. Overview of the computational workflow adopted in this study. Raw salivary FTIR spectra were subjected to spectral preprocessing, followed by spectral region selection, Principal Component Analysis (PCA), supervised classification, model training and validation, and performance evaluation. The workflow summarizes the sequence of computational steps used to investigate the influence of different preprocessing strategies on the classification of healthy controls and patients with Oral Squamous Cell Carcinoma (OSCC).

spectral baseline using different mathematical approaches, allowing their impact on downstream classification performance to be systematically evaluated.

After baseline correction, spectral smoothing was performed using a Savitzky–Golay filter implemented in the SciPy library with a window size of nine points and a second-order polynomial. This procedure reduces high-frequency noise while preserving the position, width, and intensity of absorption bands, thereby maintaining the biochemical information contained in FTIR spectra [11], [12].

The spectra were subsequently normalized using the Rampy library and the intensity normalization method, reducing differences in signal amplitude among samples while preserving their relative spectral profiles. After normalization, a second-derivative transformation was computed using the Savitzky–Golay algorithm. This step enhances subtle spectral features, improves the resolution of overlapping absorption bands, and further reduces residual baseline effects, facilitating the detection of biochemical differences between classes.

Following preprocessing, all analyses were performed using both the complete spectral range (700–6000 cm^{-1}) and the fingerprint region (900–1800 cm^{-1}). The fingerprint region was evaluated separately because it contains several absorption bands associated with proteins, lipids, nucleic acids, and carbohydrates that are commonly used as biochemical markers in FTIR-based biomedical studies [4]. This comparison allowed the influence of spectral region selection to be assessed independently of the remaining preprocessing steps.

D. Dimensionality Reduction

Principal Component Analysis (PCA) was employed to project the spectra into a lower-dimensional feature space, reducing the dimensionality of the original spectral data while preserving the dominant sources of variation. PCA was used both as an exploratory visualization technique and as an optional feature extraction step before supervised classification.

To prevent information leakage during model evaluation, the PCA model was fitted exclusively on the training portion of each cross-validation iteration. The resulting transformation was subsequently applied to the corresponding validation fold without recalculating the principal components.

E. Classification Models

Four supervised machine learning algorithms were evaluated: Linear Discriminant Analysis (LDA), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These algorithms were selected because they represent complementary classification strategies frequently employed in biomedical machine learning studies.

LDA is a linear statistical classifier that maximizes class separability by projecting the data onto a discriminant subspace. Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and robustness. SVM constructs maximum-margin decision boundaries between classes, whereas KNN assigns class labels according to the similarity between neighboring samples in the feature space.

All models were implemented using the Scikit-learn library. To ensure a consistent comparison among preprocessing strategies, no hyperparameter optimization was performed, and all classifiers were evaluated using their default parameter settings.

F. Model Training and Validation

Model evaluation was performed using Repeated Stratified Five-Fold Cross-Validation. The dataset was partitioned into five stratified folds while preserving the proportion of healthy controls and OSCC spectra in each fold. This procedure was repeated ten times using different random partitions, resulting in a total of fifty validation iterations for each preprocessing configuration and classification model.

During each iteration, four folds were used for model training and one fold was reserved for validation. The repeated cross-validation strategy was adopted to reduce the influence of a single data partition and to provide a more robust estimate of predictive performance. The reported performance metrics correspond to the mean values obtained across all validation iterations.

To prevent information leakage, all preprocessing operations requiring parameter estimation, including Principal Component Analysis (PCA), were fitted exclusively using the training folds and subsequently applied to the corresponding validation fold within each cross-validation iteration.

All experiments were performed using a fixed random seed ($random_state = 42$), ensuring reproducibility of data partitioning and model evaluation.

G. Performance Evaluation

Model performance was assessed using accuracy, precision, recall, and F1-score. These complementary metrics were calculated for every computational pipeline, enabling a systematic comparison of the effects of baseline correction, derivative transformation, spectral region selection, dimensionality reduction, and classification algorithm on the discrimination between healthy controls and patients with OSCC. For each classifier and preprocessing configuration, the reported values correspond to the mean performance obtained across the fifty validation iterations generated by the repeated stratified five-fold cross-validation procedure.

III. Results and Discussion

A. Effect of Baseline Correction

Spectral preprocessing is essential in FTIR-based classification because baseline distortions, instrumental noise, and other acquisition-related artifacts may obscure biologically relevant spectral features. Figure 2 presents the average FTIR spectra in the fingerprint region (900–1800 cm^{-1}) after baseline correction using IASLS, ModPoly, and AirPLS, together with the unprocessed spectrum.

The raw spectrum exhibited noticeable baseline drift, whereas all three correction methods substantially reduced these distortions while preserving the main absorption bands. The corrected spectra showed highly similar peak positions and spectral profiles, indicating that baseline correction removed non-biological variability without significantly altering the biochemical information contained in the samples.

To further investigate the influence of preprocessing, the spectra were subjected to second-derivative transformation. As shown in Figure 3, this procedure enhanced subtle spectral features and improved the resolution of overlapping absorption bands. After derivative transformation, the spectral profiles obtained from the three baseline correction methods became even more similar, suggesting that derivative preprocessing reduced residual differences introduced during baseline estimation.

Derivative spectroscopy is widely employed in FTIR analysis because it improves spectral resolution and attenuates residual baseline effects. In the present study, derivative preprocessing produced a more homogeneous spectral representation regardless of the baseline correction method, indicating that subsequent classification models were less influenced by acquisition artifacts and more likely to capture biologically relevant variations.

Figure 4 shows the average spectra over the complete spectral range after baseline correction. Greater variability among preprocessing methods was observed at higher wavenumbers, particularly above 3800 cm^{-1}, whereas the fingerprint region exhibited substantially greater consistency.

Because the fingerprint region contains the most diagnostically relevant absorption bands associated with proteins, lipids, nucleic acids, and carbohydrates, these observations support its use as a complementary spectral configuration in the classification experiments.

Overall, IASLS, ModPoly, and AirPLS effectively corrected baseline distortions while preserving the characteristic molecular signatures of saliva. The application of second-derivative preprocessing further enhanced spectral consistency, generating suitable representations for the machine learning analyses presented in the following subsection.

TABLE I
Performance of the best computational pipelines obtained using repeated stratified five-fold cross-validation with ten repetitions.

Pipeline	F1 (%)	Sensitivity	Specificity	AUC
FP + Derivative + RF	100.0	100.0	100.0	1.000
FP + Derivative + SVM	100.0	100.0	100.0	1.000
Full + Derivative + RF	100.0	100.0	100.0	1.000
Full + Derivative + SVM	100.0	100.0	100.0	1.000
Full + Derivative + KNN	99.6	99.5	99.7	0.998

B. Classification Performance

Table I summarizes the performance of the best computational pipelines evaluated using repeated stratified five-fold cross-validation with ten repetitions. The reported metrics correspond to the mean performance obtained across fifty validation iterations. This evaluation strategy reduces the influence of a single data partition and provides a more robust estimate of model performance and stability.

The results demonstrate that spectral preprocessing played a critical role in determining classification performance. Although baseline correction was applied to all computational pipelines, the inclusion of second-derivative preprocessing consistently produced the greatest improvement in predictive performance. Across both spectral configurations, derivative transformation enhanced the discrimination between healthy

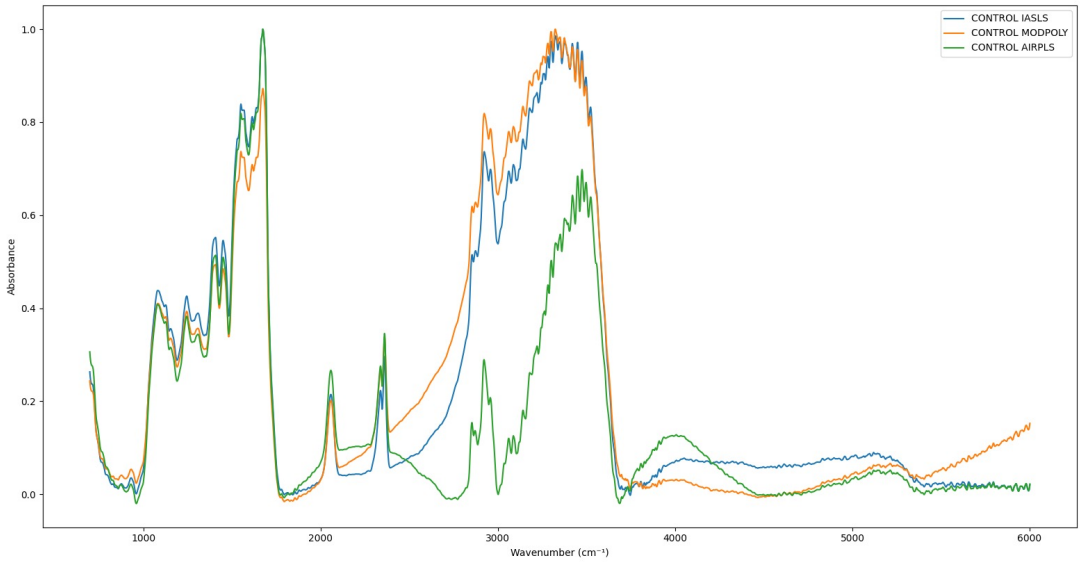


Fig. 2. Average FTIR spectra in the fingerprint region after baseline correction using IASLS, ModPoly, and AirPLS, compared with the unprocessed spectrum.

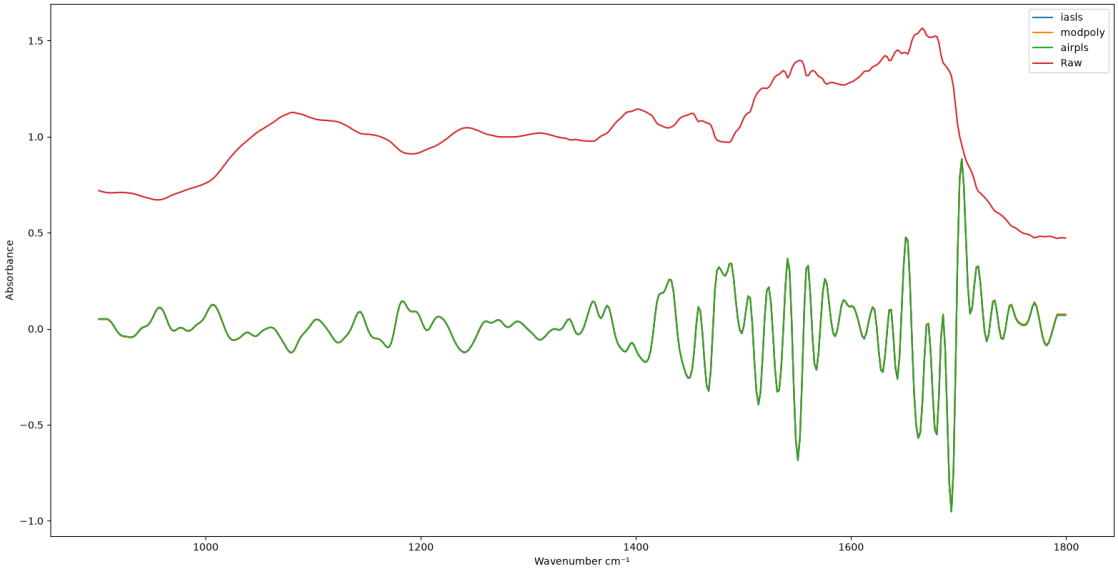


Fig. 3. Average FTIR spectra after second-derivative transformation using the three evaluated baseline correction methods.

controls and OSCC spectra, leading to substantial improvements in F1-score, sensitivity, specificity, and AUC. This behavior is consistent with the known ability of derivative spectroscopy to improve spectral resolution, reduce residual baseline effects, and enhance subtle absorption features that may be associated with disease-related biochemical alterations.

The influence of PCA was comparatively less pronounced. Although PCA is widely employed in spectroscopic studies to reduce dimensionality and minimize redundancy among spectral variables, its inclusion did not systematically improve classification performance in the present dataset. In several configurations, the results obtained with PCA were comparable to those achieved using the original spectral variables. These findings suggest that the FTIR spectra already contained sufficient discriminatory information for accurate classification and that dimensionality reduction was not a determining factor for predictive performance.

Among the evaluated classifiers, Random Forest exhibited the most robust and consistent behavior. The algorithm achieved perfect classification performance in both spectral

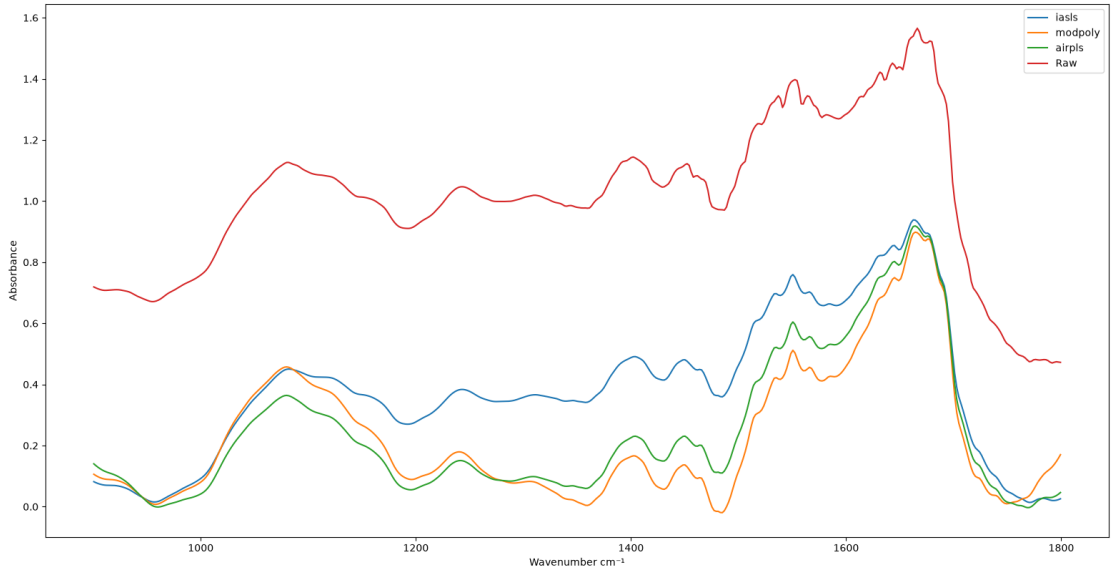


Fig. 4. Average FTIR spectra over the complete spectral range after baseline correction using IASLS, ModPoly, and AirPLS.

configurations when derivative preprocessing was applied, regardless of whether PCA was used. This stability likely reflects the ability of ensemble-based tree methods to model complex and non-linear relationships while remaining relatively insensitive to noise and irrelevant variables.

Support Vector Machine also achieved perfect classification performance when derivative preprocessing was applied, particularly in the fingerprint region. In contrast, its performance decreased considerably when derivative transformation was omitted, indicating a stronger dependence on the quality of the spectral representation. Similar behavior was observed for K-Nearest Neighbors, which benefited substantially from derivative preprocessing and achieved near-perfect classification performance under the best-performing configurations.

The comparison between the full spectral range (700–6000 cm^{-1}) and the fingerprint region (900–1800 cm^{-1}) revealed that both spectral representations contain relevant discriminatory information. The fingerprint region alone was sufficient to achieve perfect classification performance with Random Forest and Support Vector Machine, supporting the hypothesis that this spectral interval concentrates a large proportion of the biochemical information associated with OSCC. This observation is consistent with previous FTIR studies, which have shown that the fingerprint region contains absorption bands related to proteins, lipids, nucleic acids, and carbohydrates that are frequently altered in pathological conditions.

To further investigate the spectral regions contributing to model predictions, permutation feature importance was calculated for the best-performing Random Forest model. Figure 5 shows that the discriminatory information was not uniformly distributed throughout the spectrum. Instead, the highest importance values were concentrated within specific spectral intervals.

In particular, prominent contributions were observed between approximately 1000–1300 cm^{-1} and 1650–1750 cm^{-1}, both located within the fingerprint region. These intervals encompass absorption bands commonly associated with carbohydrates, phosphate-containing biomolecules, proteins, and lipid-related molecular vibrations. Additional contributions were observed between 2800 and 3600 cm^{-1}, a region associated with C–H, O–H, and N–H stretching vibrations. Conversely, the importance values decreased substantially above approximately 3700 cm^{-1}, suggesting that this portion of the spectrum contributed relatively little to the classification process.

The importance profile supports the classification results obtained using the fingerprint region and reinforces the biological plausibility of the proposed models. Rather than relying on isolated spectral variables, the classifiers appear to exploit multiple biochemical signatures distributed across several diagnostically relevant FTIR bands. Overall, these findings indicate that second-derivative preprocessing was the most influential preprocessing step, whereas Random Forest provided the most robust predictive performance. The results further demonstrate that accurate discrimination between healthy controls and patients with OSCC can be achieved through the combined optimization of spectral preprocessing and machine learning strategies.

C. Study Limitations

A limitation of this study is the absence of participant-level metadata associated with the spectral dataset. Consequently,

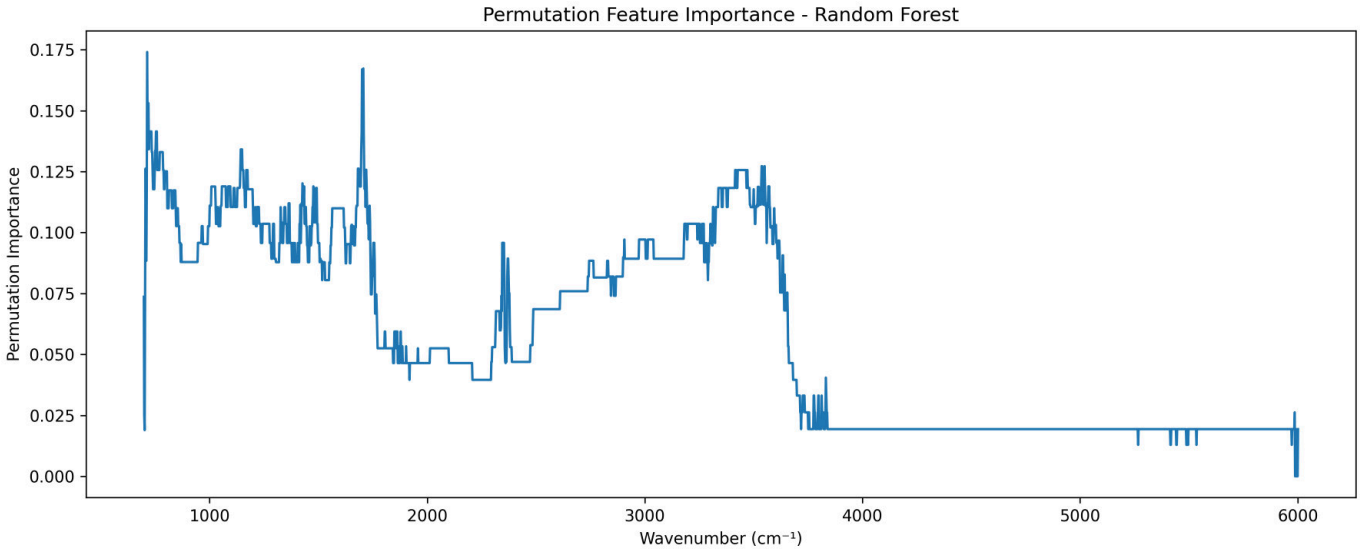


Fig. 5. Permutation feature importance profile obtained from the Random Forest classifier across the full FTIR spectral range (700–6000 cm^{-1}). Higher values indicate spectral variables whose permutation produced a greater decrease in model performance, suggesting a stronger contribution to the classification of healthy controls and patients with Oral Squamous Cell Carcinoma (OSCC).

it was not possible to determine whether multiple spectra originated from the same individual or to perform patient-wise partitioning during model training and validation. Therefore, each spectrum was treated as an independent observation throughout the computational analyses.

The dataset was obtained from a previously published study by Jimenez et al. [**?**], and only the processed spectra and corresponding class labels were available for the present investigation. Although this limitation does not affect the comparative evaluation of the preprocessing strategies investigated here, future studies should incorporate participant-level information to enable patient-wise validation and provide a more rigorous assessment of model generalization performance.

IV. Conclusion

This study evaluated the influence of different spectral preprocessing strategies on the classification of Oral Squamous Cell Carcinoma (OSCC) using salivary FTIR spectra. By systematically combining baseline correction methods, second-derivative transformation, spectral region selection, Principal Component Analysis (PCA), and supervised machine learning algorithms, it was possible to assess how different preprocessing choices affected classification performance.

The spectral analyses demonstrated that IASLS, ModPoly, and AirPLS effectively corrected baseline distortions while preserving the main biochemical characteristics of the salivary spectra. Although the corrected spectra exhibited only minor visual differences, the classification experiments revealed that preprocessing decisions substantially influenced predictive performance.

Among the evaluated preprocessing techniques, second-derivative transformation consistently produced the greatest improvement in classification performance, indicating that enhanced spectral resolution facilitated the identification of disease-related biochemical patterns. In contrast, PCA provided comparatively limited benefits, suggesting that the original spectral variables already contained sufficient discriminatory information for accurate classification.

Random Forest exhibited the most robust performance across the evaluated preprocessing configurations, while Support Vector Machine and K-Nearest Neighbors also achieved excellent results when derivative preprocessing was applied. The permutation feature importance analysis further indicated that the most informative spectral variables were concentrated within the fingerprint region and selected portions of the 2800–3600 cm^{-1} interval, highlighting the contribution of specific biochemical signatures to OSCC discrimination.

Although several computational pipelines achieved near-perfect classification performance, the results should be interpreted considering the limitations of the available dataset, particularly the absence of participant-level metadata that would allow patient-wise validation. Therefore, additional studies using larger and independent cohorts are necessary to further assess model generalizability and clinical applicability.

Overall, the findings demonstrate that spectral preprocessing can be as important as classifier selection in FTIR-based diagnostic workflows. This study contributes to the development of more reproducible computational strategies for salivary FTIR analysis and reinforces the potential of combining vibrational spectroscopy and machine learning as a rapid and non-invasive approach to support OSCC diagnosis.

References

[1] S. A. Pullano, M. G. Bianco, M. Greco, D. Mazzuca, S. P. Nisticò, and A. S. Fiorillo, "FT-IR Saliva Analysis for the Diagnosis of Psoriasis: A Pilot Study," *Biomedical Signal Processing and Control*, vol. 74, p. 103525, 2022.

[2] B. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*. Hoboken, NJ, USA: John Wiley & Sons, 2004.

[3] B. C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy*. Boca Raton, FL, USA: CRC Press, 2011.

[4] Z. Movasaghi, S. Rehman, and I. U. Rehman, "Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues," *Applied Spectroscopy Reviews*, vol. 43, no. 2, pp. 134–179, 2008.

[5] L. M. Rodrigues, T. D. Magrini, H. S. Martinho, and J. D. Almeida, "Analysis of Saliva Composition in Patients with Burning Mouth Syndrome (BMS) by FTIR Spectroscopy," *Vibrational Spectroscopy*, vol. 100, pp. 195–201, 2019.

[6] G. Parize, G. L. Jimenez, J. A. Shibli, R. Siroma, M. W. Caetano, Y. J. Kim, P. H. Braz-Silva, H. S. Martinho, and D. Pallos, "Evaluation of Peri-Implantitis Through Fourier-Transform Infrared Spectroscopy on Saliva," *Journal of Proteome Research*, 2025.

[7] C. C. S. Bandeira, K. C. R. Madureira, M. B. Rossi, J. F. Gallo, A. P. M. A. Silva, V. L. Torres, V. A. Lima, N. Kesper Jr., R. M. Zerbinati, P. H. Braz-Silva, J. A. L. Lindoso, and H. S. Martinho, "Micro-Fourier-Transform Infrared Reflectance Spectroscopy as a Tool for Probing IgG Glycosylation in COVID-19 Patients," *Scientific Reports*, vol. 12, Art. no. 4269, 2022.

[8] L. M. Rodrigues, T. D. Magrini, C. F. Lima, J. Scholz, H. S. Martinho, and J. D. Almeida, "Effect of Smoking Cessation in Saliva Compounds by FTIR Spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 174, pp. 124–129, 2017.

[9] G. R. Zucco, T. M. C. Oliveira, T. X. P. Silva, S. S. Marques, F. Sabbag, M. R. Araújo, R. M. Zerbinati, P. H. Braz-Silva, J. A. L. Lindoso, G. L. J. T. Nepomuceno, L. Foiani, H. S. Martinho, and J. D. Almeida, "Impact of Crack Cocaine on Salivary Nucleic Acids, Enzymes, Esters, and Lipids: A Fourier-Transform Infrared Spectroscopy Study," *Heliyon*, vol. 11, p. e41381, 2025.

[10] G. Jimenez, L. Foiani, J. T. Figueiredo, M. S. Alves, N. S. Rodrigues, C. M. Bandeira, M. G. O. Alves, J. D. Almeida, and H. S. Martinho, "Histine, carbohydrates/polysaccharides, and proteins with β-sheet conformation as promising staging markers for oral squamous cell carcinoma," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 340, p. 126296, 2025.

[11] S. G. Kazarian and K. L. A. Chan, "Applications of ATR–FTIR Spectroscopic Imaging to Biomedical Samples," *Biochimica et Biophysica Acta (BBA) – Biomembranes*, vol. 1758, no. 7, pp. 858–867, 2006.

[12] V. Zohdi, D. R. Whelan, B. R. Wood, J. T. Pearson, K. R. Bambery, and M. J. Black, "Importance of Tissue Preparation Methods in FTIR Micro-Spectroscopical Analysis of Biological Tissues: Traps for New Users," *PLOS ONE*, vol. 10, no. 2, e0116491, 2015.

Implementación y evaluación de un sistema de telemedicina integrado al sistema de información hospitalario para diagnóstico remoto en Paraguay

César José Yegros Acevedo[1], Eladio Quintana[1], Jorge Giubi[1]
Universidad Nacional de Asunción, San Lorenzo, Paraguay

Resumen—El acceso oportuno al diagnóstico especializado continúa siendo desigual en sistemas públicos de salud con población geográficamente dispersa. Este trabajo presenta la implementación y evaluación de una solución de telemedicina integrada al sistema de información hospitalario, complementada con geolocalización referencial para apoyar el diagnóstico remoto y la gestión sanitaria. El estudio prospectivo se realizó en un hospital universitario público de Paraguay entre septiembre de 2020 y diciembre de 2022. Se evaluaron la integración técnica, la usabilidad, la satisfacción profesional, la efectividad diagnóstica y aspectos económicos. Durante el piloto se analizaron 156.115 atenciones; 5.324 pacientes con COVID-19, dengue o leishmaniasis conformaron la muestra de evaluación diagnóstica. El porcentaje global ponderado de diagnósticos acertados con tratamiento adecuado fue 98,2%. La carga de geolocalización requirió en promedio 44,78 s y la satisfacción de los profesionales fue alta, aunque se identificaron dificultades de conectividad, capacitación y precisión de algunas direcciones. Los resultados muestran que integrar la telemedicina, el registro clínico digital y la geolocalización puede ampliar el acceso al especialista y generar información útil para la planificación, aprovechando la infraestructura disponible.

Palabras clave—*telemedicina, diagnóstico remoto, sistema de información hospitalario, geolocalización, salud digital.*

I. INTRODUCCIÓN

La distancia entre el paciente y el especialista sigue condicionando el acceso a servicios diagnósticos en varios países latinoamericanos. La telemedicina reduce parte de esa barrera, pero una consulta remota aislada tiene valor limitado si no se integra al registro clínico, a los procesos del hospital y a la capacidad de gestión del sistema de salud. En contextos con recursos restringidos, la solución tecnológica debe ser útil, sostenible y sencilla de incorporar al trabajo cotidiano [1]–[4].

El Hospital de Clínicas de la Universidad Nacional de Asunción implementó un sistema de información hospitalario (HIS) para digitalizar las fichas clínicas. Experiencias previas desarrolladas en Paraguay ya habían mostrado la factibilidad de implementar servicios de telesalud sobre infraestructura sanitaria disponible [10]. Sobre esa plataforma se desarrolló un módulo de telemedicina con geolocalización referencial. La propuesta agrega al registro del paciente datos de dirección y coordenadas, permite visualizar su ubicación y vincula esa información con el diagnóstico remoto. Así, el mismo dato utilizado durante la atención puede servir para conocer el origen territorial de la demanda y reconocer concentraciones de determinadas patologías.

Desde la Ingeniería Biomédica, el interés del proyecto radica en la integración: infraestructura informática, conectividad, dispositivos diagnósticos, profesionales y datos clínicos funcionan como un solo sistema. Este trabajo tuvo como objetivo implementar y evaluar dicha solución en condiciones reales, considerando desempeño técnico, usabilidad, aceptación profesional, efectividad diagnóstica e impacto económico-social.

II. MATERIALES Y MÉTODOS

Se realizó un estudio prospectivo con información de pacientes ambulatorios y de urgencias atendidos mediante diagnóstico remoto, desde la implantación del módulo de geolocalización, en septiembre de 2020 hasta diciembre de 2022. Para la evaluación clínica se incluyeron pacientes diagnosticados a distancia con COVID-19, dengue o leishmaniasis; se excluyeron los casos sin diagnóstico remoto asociado al sistema. El estudio fue aprobado por el Comité de Ética en Investigación de la Facultad de Ciencias Médicas de la Universidad Nacional de Asunción, mediante dictamen CEI N.º 48/2022, y contempló acceso autorizado a la base de datos, anonimización de la información y protección de la privacidad.

A. Integración tecnológica

El módulo fue incorporado al HIS existente, evitando crear una plataforma independiente. Se añadieron campos para dirección, ciudad, departamento, latitud y longitud, junto con una interfaz cartográfica. El desarrollo utilizó tecnologías web y bases de datos de código abierto, con un esquema modular y control de acceso por roles. Antes de la implementación se relevaron recursos disponibles o subutilizados, incluidos equipos diagnósticos, red, personal técnico e infraestructura informática.

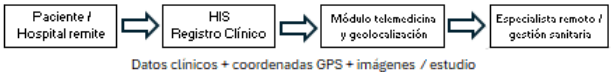


Fig. 1. Arquitectura funcional simplificada del sistema integrado de telemedicina, HIS y geolocalización.

B. Evaluación

La evaluación combinó cinco componentes. Primero, se verificó la implementación y funcionamiento del módulo. Segundo, se midió la usabilidad con 46 profesionales (7 médicos y 39 técnicos), considerando tiempo de tarea, número de pasos y errores. Tercero, se aplicó una encuesta semiestructurada con escala Likert de 1 a 5. Cuarto, la efectividad diagnóstica se estimó comparando el diagnóstico remoto con el diagnóstico presencial definitivo establecido por el especialista, utilizado como patrón de referencia. Se consideró el acierto cuando existió concordancia entre ambos diagnósticos y el tratamiento indicado fue considerado adecuado. Quinto, se analizaron costos de implantación y mantenimiento frente a costos asociados a la atención presencial, incluidos transporte, alimentación, costo de oportunidad e informe especializado. El enfoque económico siguió principios habituales de evaluación de tecnologías sanitarias [5]–[7]. Los resultados descriptivos se expresaron mediante frecuencias, porcentajes y medias. Para la proporción global de concordancia diagnóstica se estimó un intervalo de confianza del 95% a partir de los datos agregados disponibles.

III. RESULTADOS

La solución se integró al HIS y se mantuvo operativa durante el piloto. Los campos de ubicación quedaron asociados al registro clínico y pudieron consultarse individualmente o representarse en mapas. Durante el periodo analizado se registraron 156.115 atenciones médicas entre

consultas ambulatorias y teleconsultas. De ellas, 5.324 pacientes cumplieron los criterios para la evaluación diagnóstica; 41,0% fueron hombres y 59,0% mujeres, con edad media de 48 años.

TABLA I. EFECTIVIDAD DEL DIAGNÓSTICO REMOTO

Patología	N	Acierto y Tratamiento	Error
COVID-19	5099	98,3	1,7
Dengue	206	96,1	3,9
Leishmaniasis	19	84,2	15,8
Total	5324	98,2	1,8

Considerando el tamaño de cada grupo diagnóstico, la proporción global ponderada de concordancia diagnóstica con tratamiento adecuado fue 98,2% (IC95% aproximado: 97,8%–98,5%), con una proporción global de error de aproximadamente 1,8%. La diferencia entre patologías debe interpretarse junto con el tamaño desigual de los grupos, especialmente en leishmaniasis. Entre los factores que afectaron el desempeño se identificaron la experiencia clínica, diferencias entre dispositivos diagnósticos, la calidad de la transmisión y la experiencia del especialista remoto.

A. Usabilidad y aceptación

El tiempo medio para introducir dirección, ciudad y ubicación en el mapa fue 44,78 s, por debajo del objetivo de un minuto. La tarea requirió cinco pasos, dentro del rango considerado de alta eficiencia. Los 46 profesionales realizaron cinco tareas de carga cada uno, totalizando 230 tareas evaluadas. Se identificaron 68 errores (29,6%): 16 relacionados con la dirección (7,0%), 13 con la protección de datos (5,7%) y 39 con el posicionamiento en el mapa (17,0%). Este último componente concentró el 57,4% de todos los errores registrados. El problema fue más frecuente en domicilios fuera de Asunción y Gran Asunción, donde las direcciones postales pueden ser imprecisas.

Los 46 profesionales participantes (7 médicos especialistas y 39 técnicos sanitarios) respondieron la encuesta de satisfacción. Para la asistencia recibida de especialistas remotos, el promedio ponderado fue 4,5/5; para la calidad del servicio, 4,6/5; y para la valoración general de la unidad, aproximadamente 4,5/5. Las sugerencias se concentraron en reducir el tiempo de entrega de resultados, mejorar la velocidad de conexión y reforzar la capacitación.

TABLA II. INDICADORES PRINCIPALES DE IMPLEMENTACIÓN

Indicador	Resultado
Atenciones analizadas	156.115
Muestra diagnóstica	5.324
Tiempo medio de carga geográfica	44,78 s
Profesionales en usabilidad	46
Tareas de usabilidad evaluadas	230
Errores registrados	68 (29,6 %)
Error de dirección	16 (7,0 %)
Error de protección de datos	13 (5,7 %)
Error de posicionamiento en mapa	39 (17,0 %)
Satisfacción con servicio remoto	≈4,5/5

B. Costos e impacto

La evaluación económica tuvo un carácter descriptivo y permitió identificar los principales componentes de costo asociados a las modalidades de atención analizadas. En la atención presencial se consideraron costos que no recaen únicamente en la institución, como transporte, alimentación, jornada laboral perdida y acceso al especialista. Durante el periodo estudiado se registró un costo agregado de US$ 4.816,2 correspondiente a actividades de implantación, mantenimiento y evaluación remota del sistema. Por otra parte, para la atención presencial se estimó un costo promedio de US$ 130,6 por paciente.

Los registros disponibles no permitieron desagregar retrospectivamente los costos de implantación de los costos operativos atribuibles a cada atención remota. Por este motivo, no fue posible realizar un análisis económico incremental ni estimar un costo unitario directamente comparable entre las modalidades remota y presencial. En consecuencia, los valores obtenidos se presentan como indicadores descriptivos de la estructura de costos y no como una comparación de costo-efectividad entre ambas modalidades.

Durante la implantación también se identificaron barreras operativas relacionadas con la adecuación de interfaces entre equipos diagnósticos y computadoras, la capacitación de técnicos y especialistas, la resistencia inicial al uso de registros electrónicos y las diferencias de velocidad entre estaciones de la red. Con la infraestructura disponible fue posible transmitir y procesar imágenes y datos de forma satisfactoria durante el piloto; sin embargo, estos hallazgos muestran que un eventual escalamiento requiere fortalecer la conectividad, la interoperabilidad y la capacitación de los usuarios.

IV. DISCUSIÓN

Los resultados apoyan la factibilidad de integrar telemedicina y geolocalización al sistema de información hospitalario sin sustituir el flujo clínico existente. Este punto es relevante porque la sostenibilidad de la telesalud depende tanto de la tecnología como de su incorporación a procesos, recursos humanos y condiciones locales [2], [8]. El porcentaje global ponderado de acierto diagnóstico fue 98,2%, mientras que aproximadamente 1,8% correspondió a resultados clasificados como errores. La interpretación debe considerar la marcada diferencia en el tamaño de las muestras entre las tres patologías evaluadas y muestra que el desempeño remoto sigue condicionado por la calidad de la información, la conectividad y la experiencia clínica.

La conectividad apareció como una limitación práctica, en coincidencia con experiencias previas de telemedicina en regiones remotas y con la literatura sobre sistemas de información clínica a distancia [3], [4], [9]. La usabilidad presentó un comportamiento mixto: el tiempo de carga fue adecuado y el número de pasos fue reducido, pero la tasa de errores de georreferenciación fue elevada. Esto sugiere que el problema no está solamente en la interfaz. En Paraguay muchas direcciones carecen de nomenclatura uniforme, por lo que futuras versiones deberían incorporar validación de coordenadas, selección asistida y mecanismos de corrección.

La principal contribución del sistema es que el diagnóstico remoto no termina en la consulta. El dato clínico queda integrado al HIS y puede vincularse a su distribución territorial. Esto permite pensar en una plataforma que, además de acercar especialistas, apoye vigilancia, planificación y asignación de recursos. Para sistemas de cobertura universal en América Latina, donde la concentración de especialistas y

las distancias siguen siendo barreras, este tipo de integración puede ser más transferible que soluciones aisladas y costosas.

La posibilidad de personalizar el HIS y emplear componentes de código abierto fue relevante para reducir barreras de entrada. El modelo no exige que cada establecimiento disponga de todos los especialistas: organiza el acceso remoto a una capacidad diagnóstica compartida y deja la información dentro del expediente clínico. La capa geográfica añade una función de gestión que puede orientar cobertura, derivaciones y vigilancia. Esa combinación constituye el principal elemento de innovación aplicado al sistema público.

El estudio tiene limitaciones. Se realizó en una institución y con tres patologías; los tamaños de muestra fueron muy diferentes. La evaluación económica no fue un análisis incremental completo y algunas variables de impacto social se estimaron con información institucional. Tampoco corresponde extrapolar los resultados diagnósticos a otras especialidades sin nuevas validaciones. Aun así, la experiencia aporta evidencia de funcionamiento en condiciones reales y deja identificados los puntos que deben fortalecerse antes de escalar: conectividad, normalización de direcciones, capacitación, interoperabilidad y protocolos clínicos.

V. CONCLUSIÓN

Fue posible implementar un sistema de telemedicina integrado al HIS y complementado con geolocalización referencial en un hospital universitario público de Paraguay. El piloto reunió 156.115 atenciones y, en 5.324 pacientes evaluados para diagnóstico remoto, se observó una proporción global ponderada de 98,2% de diagnósticos acertados con tratamiento adecuado. La aceptación profesional fue alta y el tiempo de carga geográfica resultó compatible con el flujo asistencial, aunque persistieron errores de ubicación y limitaciones de conectividad.

La innovación propuesta no reside en una tecnología aislada, sino en conectar diagnóstico remoto, registro clínico y ubicación territorial dentro de una infraestructura disponible. Con validaciones adicionales, este enfoque puede apoyar la ampliación del acceso al especialista y ofrecer información útil para decisiones sanitarias en otros contextos latinoamericanos con recursos limitados.

REFERENCIAS

[1] H. Lucas, “Information and communications technology for future health systems in developing countries,” Social Science & Medicine, vol. 66, pp. 2122–2132, 2008.

[2] D. Hailey, R. Roine, and A. Ohinmaa, “Systematic review of evidence for the benefits of telemedicine,” J. Telemed. Telecare, vol. 8, suppl. 1, pp. 1–30, 2002.

[3] M. P. Gagnon, J. Duplantie, J. P. Fortin, and R. Landry, “Exploring the effects of telehealth on medical human resources supply: a qualitative case study in remote regions,” BMC Health Serv. Res., vol. 7, art. 6, 2007.

[4] A. C. Ferreira, E. O’Mahony, A. H. Oliani, E. Araujo Júnior, and F. da Silva Costa, “Teleultrasound: historical perspective and clinical application,” Int. J. Telemed. Appl., 2015, Art. no. 306259.

[5] M. F. Drummond, B. J. O’Brien, G. L. Stoddart, and G. W. Torrance, Métodos para la evaluación económica de los programas de asistencia sanitaria. Madrid, España: Díaz de Santos, 2001.

[6] M. V. Andrade, A. C. Maia, C. S. Cardoso, M. B. Alkmim, and A. L. Ribeiro, “Cost-benefit of the telecardiology service in the state of Minas Gerais: Minas Telecardio Project,” Arq. Bras. Cardiol., vol. 97, no. 4, pp. 307–316, 2011.

[7] I. de la Torre-Díez, M. López-Coronado, C. Vaca, J. S. Aguado, and C. de Castro, “Cost-utility and cost-effectiveness studies of telemedicine, electronic, and mobile health systems in the literature: a systematic review,” Telemed. J. E Health, vol. 21, no. 2, pp. 81–85, 2015.

[8] Organización Panamericana de la Salud, Bases metodológicas para evaluar la viabilidad y el impacto de proyectos de telemedicina. Washington, DC, USA: OPS/OMS, 2001.

[9] J. C. Hsieh and H. C. Lo, “The clinical application of a PACS-dependent 12-lead ECG and image information system in E-medicine and telemedicine,” J. Digit. Imaging, vol. 23, no. 4, pp. 501–513, 2010.

[10] P. Galván, M. B. Cabral, and V. Cane, “Implementación de un sistema de telemedicina (telesalud) en el Instituto de Investigaciones en Ciencias de la Salud,” Mem. Inst. Investig. Cienc. Salud, vol. 4, no. 1, pp. 20–27, 2008.

VIZU: A Mobile Hub Proposal for Computer Vision Models Applied to Healthcare

Jorge Luis Pereira Monteiro
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
jorgeluizpereira@unifor.br

Belmondo Rodrigues Aragão Júnior
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
belmondo@unifor.br

Felipe Cassiano Barbosa
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
felipecassianobarbosa@hotmail.com

André Paiva Garcia
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
andregarcia@unifor.br

João Pedro Cavalcante Araripe e Lima
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
joaopedrocal@unifor.br

João Victor Rodrigues Barreto
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
joao.barreto@unifor.br

Julia Emilia de Oliveira
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
juliaemi123@unifor.br

Rolf Freitas Matela
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
rolf.matela@unifor.br

Joel Sotero da Cunha Neto
Vortex, Vice-Rectorate for Research
University of Fortaleza
Fortaleza, Brazil
joelsotero@unifor.br

***Abstract*—The growing use of artificial-intelligence (AI) models in healthcare has driven the development of specialized applications for distinct computer-vision tasks. However, the recurrent need to reimplement similar functionalities, authentication, image capture, inference, and result visualization, for each new model leads to fragmented, independent applications and duplicated effort. This paper presents VIZU, a mobile architecture proposal that centralizes computer-vision models for health behind a single, user-facing inference interface. Through VIZU, a user authenticates, submits an image from the device gallery or camera to a microservice, receives the model output (description, probability, and category), and acts upon it, archiving the analysis or disagreeing with the result, thereby contributing to model retraining, while visual feedback reports the connection and processing states. Prototyped with React Native and Expo for cross-platform compatibility, the proposed architecture could allow new models to be integrated into a shared model hub with minimal changes to the application. As an evolving prototype, its quantitative evaluation through usability testing and quality indicators is left as future work. The proposed approach favors the systematic reuse of components, offering a replicable foundation for research and development in computer vision applied to health.**

***Index Terms*—computer vision, artificial intelligence, medical imaging, mobile application**

I. Introduction

The application of Artificial Intelligence (AI) techniques in healthcare has expanded significantly in recent years, driven by advances in machine learning, computer vision, and the availability of large volumes of clinical and imaging data [1]. These technologies have been used in various contexts, including diagnostic support, disease prognosis, medical image segmentation, and clinical decision support, contributing to increased accuracy and efficiency in healthcare processes.

Among the various applications of AI in healthcare, specialized computer vision models designed for the analysis of imaging exams stand out. Deep neural networks have been used to identify patterns in X-rays [2], computed tomography (CT) scans [3], magnetic resonance imaging (MRI) scans [4], and clinical photographs [5] [6], assisting healthcare professionals in the early detection of pathologies and in conducting diagnostic evaluations. This trend has driven the development of specific solutions for different medical areas and types of exams.

At the same time, the growth of mobile technologies has fostered the spread of mobile health (mHealth) applications, which provide access to clinical tools directly through smartphones and tablets. These applications can expand the availability of computational resources for healthcare professionals, facilitate the collection and sharing of information, and support diagnostic and patient follow-up activities in different healthcare settings. [7]

Despite these advances, the development of mobile applications integrated with AI models often occurs in isolation, resulting in the creation of multiple systems that share similar functionalities. In many cases, different projects must reimplement mechanisms for authentication, image management, communication with inference services, and result visualization, leading to redundant development and maintenance efforts.

Existing ecosystems partially address model reuse, but from

a developer's perspective. Frameworks such as the Medical Open Network for AI (MONAI) standardize how medical-imaging models are built, shared, and deployed into clinical infrastructure [9], while general-purpose hubs centralize pretrained models for technical users. These solutions lower the effort of developing and deploying models, yet they do not provide a single, end-user-facing mobile interface through which a possibly non-technical user can run inference on a chosen model and act upon its results.

With this in mind, and based on the demands observed in previous projects conducted by the researchers, in which opportunities for component reuse and functionality standardization were identified, this paper presents Vizu, a unified mobile platform for the integration, validation, and visualization of computer vision models applied to healthcare. The proposal aims to centralize functionalities common to different AI-based applications, allowing for the incorporation of new models without the need to develop standalone applications for each solution.

As this is an evolving prototype, the contribution of this paper is primarily conceptual and architectural. Specifically, we: (i) identify the recurring requirements of mobile applications for AI-based image inference in health; (ii) propose an architecture that centralizes these functionalities behind a single interface and a model hub; and (iii) describe an initial prototype, discussing representative health use cases and outlining a plan for future evaluation. The remainder of this paper is organized as follows. Section II reviews related work on medical-AI frameworks and model hubs; Section III specifies the methodology used in this works; Sections IV and V present the the system requirements, architecture and the prototype; Section VI discusses use cases; Section VII outlines future work and the evaluation plan; and Section VIII concludes.

II. Related Work

The growing adoption of artificial intelligence models across various domains has driven the development of platforms, architectures, and systems designed to integrate, manage, and make these models available. These initiatives aim to reduce the complexity associated with the lifecycle of machine learning-based solutions by promoting reuse, scalability, and interoperability among different applications.

In the scientific context, DLHub [10] was proposed as a platform for publishing, sharing, and running machine learning models, allowing researchers to make models available along with their associated artifacts and metadata . The proposal seeks to address challenges related to the reproducibility, reuse, and availability of models for inference in different computational environments, offering standardized mechanisms for executing and sharing artificial intelligence solutions.

Similarly, ModelHub.AI [11] was developed as a platform focused on the structured dissemination of deep learning models. The proposed architecture uses containers and standardized interfaces to make models from different domains and computational frameworks available, with the goal of facilitating their reuse and integration by researchers and developers. The work highlights the need for mechanisms that reduce barriers related to interoperability and the sharing of artificial intelligence models.

In addition, several studies have investigated architectures focused on the lifecycle management of machine learning models. The Gallery system [12], developed by Uber, exemplifies this approach by providing an infrastructure for storing, versioning, monitoring, and orchestrating models at scale. The solution was designed to handle multiple models and heterogeneous applications, reducing the fragmentation of model management functions throughout the model lifecycle.

The literature also highlights the growing use of microservice-based architectures to support scalable artificial intelligence systems. In this approach, components responsible for distinct stages of the model lifecycle (training, inference, monitoring, and data processing) are decoupled into independent services, facilitating maintenance and the continuous evolution of solutions. Major industry platforms, such as TensorFlow Extended (TFX) [13], Michelangelo [14], and SageMaker [15], adopt similar principles to promote modularity and component reuse.

In the healthcare field, recent studies have emphasized the need for architectures capable of integrating multiple artificial intelligence models into complex clinical environments. Dedicated ecosystems such as the Medical Open Network for AI (MONAI) [9] provide frameworks, model zoos, and deployment SDKs that standardize how medical-imaging models are built, shared, and integrated into clinical infrastructure through standards such as DICOM and FHIR . In a complementary direction, the Health xAI [16] Implementation Framework proposes a structure for the development, deployment, monitoring, and maintenance of large-scale AI-based solutions, considering aspects such as modularity, integration between models, interpretability, and adaptation to the specific requirements of healthcare environments . The effective adoption of these technologies depends not only on the performance of the models but also on the existence of infrastructure that enables their integration into clinical workflows.

Although these studies make significant contributions to the sharing, management, and deployment of artificial intelligence models, it is evident that most proposals focus on the infrastructure for executing, publishing, or managing the lifecycle of these models. In contrast, less attention has been paid to providing a standardized mobile layer that allows for the testing, validation, and visualization of different computer vision models in healthcare through a single interface. In this context, this paper contribution lies in the proposal of a replicable mobile architecture aimed at integrating computer vision models applied to healthcare, seeking to reduce the need to develop specific applications for each solution and to promote the reuse of components across different projects.

Table I summarizes the reviewed works according to six dimensions relevant to the problem addressed in this paper: focus on the health domain, provision of an end-user interface (as opposed to a developer-oriented one), existence of a mobile

layer, focus on computer vision, support for user-driven model testing, and the use of end-user feedback to support retraining. As shown, the existing solutions concentrate on the infrastructure for publishing, serving, or managing the lifecycle of models, and each addresses only a subset of these dimensions. In particular, none of them combines a mobile, end-user-facing layer with the testing and visualization of computer-vision (CV) models in healthcare, which characterizes the gap explored in this work.

TABLE I
COMPARISON OF RELATED WORK

Work	C1	C2	C3	C4	C5	C6
DLHub [10]	∼	∼	–	–	∼	–
ModelHub.AI [11]	∼	∼	–	∼	∼	–
Gallery [12]	–	–	–	–	–	∼
TFX [13]	–	–	–	–	–	∼
Michelangelo [14]	–	–	–	–	–	∼
SageMaker [15]	–	–	–	–	–	∼
MONAI [9]	✓	–	–	✓	∼	–
Health xAI [16]	✓	∼	–	–	–	∼

C1: health domain; C2: end-user interface; C3: mobile layer; C4: computer vision; C5: user-driven model testing; C6: end-user feedback for retraining. ✓: fully; ∼: partially; –: not addressed.

III. METHODOLOGY

The development of the proposed architecture and the Vizu application was conducted using the Interdisciplinary Method for the Development of Health-Focused Technologies (MIDTs) [8], a user-centered method geared toward developing technological solutions for the healthcare context. MIDTs integrates principles of interdisciplinarity, Design Science, and User-Centered Design, and is structured into six stages (Figure 1): (i) understanding the context of the problem, (ii) identifying needs and establishing requirements, (iii) ideation and design, (iv) prototyping, (v) solution evaluation, and (vi) academic and market documentation and actions. The method was adopted because it provides a systematic approach to identifying real-world demands, defining requirements, and building technological solutions aligned with user needs and research objectives.

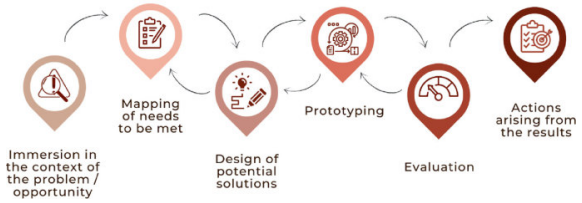


Fig. 1. MIDTS Phases. Adaptaded from [8]

A. Context of the problem

The first stage of the MIDTs consisted of understanding the context of the problem that motivated the development of the proposed solution. To this end, meetings were held between researchers and developers with the goal of identifying recurring difficulties observed during the implementation of mobile applications associated with computer vision models in the healthcare field.

During this stage, it was found that different projects shared similar functionalities, such as image capture and management, user authentication, communication with inference services, result visualization, and dataset management. Despite these similarities, each new model required the development of its own application, resulting in duplication of implementation and maintenance efforts.

The context analysis also took into account experiences gained from the Dr. Bone [7] and Pellis projects: the former focused on the analysis of radiographic exams using artificial intelligence, and the latter on the classification of burns using computer vision. The recurrence of common requirements across these projects highlighted the opportunity to define a reusable architecture capable of centralizing shared functionalities and facilitating the integration of new specialized models.

In addition to internal discussions, this phase involved researching concepts related to artificial intelligence applied to healthcare, medical computer vision, architectures for mobile applications, and mHealth solutions, with the aim of providing a theoretical foundation for the identified problem and supporting subsequent development decisions. This activity is aligned with the MIDTs recommendations, which call for an in-depth understanding of the problem and the conduct of a literature review to inform the subsequent stages.

B. Needs and requirements

The second phase aimed to identify the needs of the solution's potential users and establish the requirements necessary for building the proposed architecture.

Based on the information obtained in the previous stage, we analyzed the functional and non-functional requirements present in the projects that served as the motivation for this study. This analysis allowed us to identify elements common to applications based on computer vision models in healthcare, regardless of the specific task performed by the artificial intelligence model.

Among the identified needs, the following stood out: the ability to integrate multiple computer vision models into a single application; the reuse of functionalities shared across different projects; the simplification of the validation process for new models; and the provision of a standardized interface for visualizing the results produced by the inferences.

Based on these needs, requirements were established for the proposed architecture. As a result of this phase, a consolidated vision of the requirements necessary for developing the Vizu architecture was defined, serving as the basis for the ideation, architectural modeling, and implementation activities carried out in the subsequent phases of the method.

C. Design

The ideation and design phase aimed to transform the requirements identified in the previous phases into an architectural proposal capable of meeting the needs observed in computer vision projects in healthcare. In this phase,

different alternatives for organizing the system's components were analyzed, seeking to maximize the reuse of common functionalities and minimize coupling between the mobile application and the artificial intelligence models.

Given that the different use cases shared the same type of input—images submitted for processing—an architecture was designed based on centralizing the application's generic functionalities and decoupling the inference services responsible for processing the models. This approach allowed new models to be incorporated into the platform without the need to develop a new mobile application for each solution.

The proposed architecture was structured to separate responsibilities related to the user interface, image management, communication with external services, and the presentation of results. As a result, the application now acts as a unified interaction layer between users and the various available computer vision models.

During the design process, the application's usage flows, the mechanisms for linking models and image sets, and the communication strategies with the inference services hosted on dedicated infrastructure were also defined. These definitions served as the basis for designing the system's logical architecture and for developing the solution's functional prototypes.

D. Prototyping

The prototyping phase consisted of implementing the architecture designed in the previous phase through the development of the Vizu platform. The goal of this phase was to bring the proposed architectural concepts to life and verify their applicability in real-world usage scenarios.

Initially, the components responsible for the common functionalities identified during the requirements gathering phase were developed, including user management, image selection and storage, communication with inference services, and result visualization. These components were designed to be reusable regardless of the AI model associated with the application.

Next, the necessary mechanisms were implemented to enable the integration of different computer vision models into the platform. This integration was achieved through communication with external services responsible for performing inferences, allowing the application to remain independent of the specific characteristics of each model.

To validate the implementation of the architecture, the models developed in the Dr. Bone and Pellis projects were incorporated into the platform. These use cases made it possible to verify the solution's ability to support different medical image analysis tasks using the same mobile application framework.

At the end of this stage, a functional version of Vizu was obtained, capable of centralizing interaction with multiple computer vision models in healthcare, thereby laying the foundation for the evaluation activities planned for the subsequent stages of the MIDTs method.

IV. System Requirements

To support the full development of the VIZU platform, the requirements definition described in the methodology first sought to decompose the central idea of the project into smaller units that could be mapped more directly to design and development tasks.

The researchers initially defined the obligations related to authentication: a VIZU user must be able to create an account, log in, and modify account data such as username and password. This aligns the platform with the conventions of widely used mobile applications, improving the user experience through familiarity.

Regarding the AI functionalities, the user must also be able to submit images from the device gallery or camera to the microservice, which forwards the file to the AI model for processing. After processing, the AI result is returned to the user, including a description, a probability, a category, and/or any other elements provided by the model's response. Following this step, VIZU must also allow the user to act upon the result, for instance, adding it to a personal history of analyses linked to the profile, or disagreeing with the AI output, thereby contributing to the model-retraining process.

VIZU further defines functional requirements related to user feedback throughout the interaction with the selected model. The platform must provide visual elements that indicate the current state of the connection with the server (e.g., whether it is under maintenance or operational) and the processing state of the result, informing the user about the steps of the analysis (whether the image has been sent to the server and when processing has completed successfully). For the end user, awareness of these stages and of the current interaction state is essential, since inference may take a few seconds and the absence of visual feedback could create uncertainty about whether the system has completed the task.

In turn, the non-functional requirements of VIZU center on ensuring a uniform and secure experience for all users. Compliance with the Brazilian General Data Protection Law (LGPD) is explicitly defined as a requirement and must be maintained throughout the development process. Passwords, user data, requests, and information storage must follow strict encryption rules, particularly in the case of medical images. In addition, compatibility with both Android and iOS,also highlighted in the requirements document,gives the application a universal scope not restricted to specific operating systems. The use of React Native and Expo to prototype the application facilitates this compatibility, since the build processes for Android and iOS are specific to, and documented by, the framework itself.

As a result, the requirements defined for VIZU (Table II) establish a solid foundation for developing a functional, secure, and user-centered platform. By addressing authentication, AI-based image processing, feedback mechanisms during interaction, and compliance with security and privacy requirements, the set of specifications clearly delimits the responsibilities of the system and the criteria for its implementation. The functional and non-functional requirements thus act in a complementary manner, guiding the development stages and ensuring that the final solution meets both user needs and the technical and legal demands of the application context.

TABLE II
SUMMARY OF VIZU SYSTEM REQUIREMENTS

ID	Requirement
Functional Requirements	
FR1	Create a user account.
FR2	Log in and authenticate.
FR3	Edit account data (e.g., username and password).
FR4	Submit images from the device gallery or camera to the microservice.
FR5	Forward the image to the AI model and return the result (description, probability, category, and other model outputs).
FR6	Add a result to the user's personal analysis history, linked to the profile.
FR7	Disagree with an AI result, supporting the model-retraining process.
FR8	Provide visual feedback on the server connection state (operational or under maintenance).
FR9	Provide visual feedback on the processing state (image sent, processing completed).
Non-Functional Requirements	
NFR1	Ensure compliance with the LGPD throughout the development process.
NFR2	Apply strict encryption to passwords, user data, requests, and storage, especially for medical images.
NFR3	Provide cross-platform compatibility with Android and iOS, using React Native and Expo.
NFR4	Guarantee a uniform and secure experience for all users.

V. PROTOTYPE

A. Design

The ideation and design step aimed to turn the requirements identified in the previous phases into a concept that could meet the needs observed in computer vision projects in the healthcare field. To structure this phase, we adopted the Elements of User Experience framework [17], which organizes the design process into five progressive levels (strategy, scope, structure, skeleton, and surface), moving from the most abstract to the most concrete. This framework was chosen because of its suitability for the development of user-centered digital products. Since it operates gradually, with each stage serving as the foundation for the next, it prevents premature decisions about the application's design before its exact function has been established.

As part of the strategy, a qualitative exploratory study was conducted with professionals who worked in artificial intelligence-based image classification applications. As a data collection tool, a semi-structured interview guide was developed; this format allows for maintaining the thematic focus without losing the fluidity of the dialogue, thereby facilitating spontaneous exploration of points not anticipated by the researcher [18]. The guide addressed the functioning of the applications already developed, the features perceived as most frequently used, those considered critical by the professionals, and any features whose absence was noted during implementation. Based on the accounts, proto-personas were constructed, which, as described by Gothelf [19], are developed from hypotheses and partial data, aimed at aligning the team's understanding of the users prior to the conduct of more in-depth primary studies.

The qualitative summary followed a process of progressive data reduction. Relevant transcript excerpts were synthesized into short sentences capturing their essential content [20], then randomized to remove links to interviewee identities and reduce profile-related bias. The synthesized data were grouped by thematic similarity into an Affinity Diagram, based on the KJ Method [21] and adapted to user-centered design by Beyer and Holtzblatt [22]. Finally, each cluster was labeled by its predominant theme, providing a systemic understanding of shared user problems and needs.

As part of the scope planning, a card-sorting activity was conducted with the interviewed professionals, in which they were asked to organize cards representing features into groups that made sense from their perspective. This technique, described by Spencer [23], reveals mental models of how the system should be structured. In this study, the participants were not end users of the platform, but rather domain experts with direct experience in developing similar applications, a profile that, as pointed out by Nielsen [24], offers a privileged source of knowledge about recurring usage patterns and the system's technical and functional requirements. During the activity, participants were instructed to verbalize their reasoning (think aloud) [24], allowing us to capture not only the final result but also the logic that guided the categorization. The analysis of the resulting groupings was conducted using artificial intelligence-assisted analysis tools, with the goal of identifying cross-cutting patterns among the configurations proposed by different users and consolidating a single structure capable of accommodating the diversity of mental models observed. The result served as the basis for consolidating the functional and non-functional requirements presented in Section IV.

Regarding the structural design planning, the platform's core architectural decisions were made. Given that the different use cases shared the same type of input (images submitted for processing), an architecture was designed based on centralizing the application's generic functionalities and decoupling the inference services responsible for processing the models. This approach allowed new models to be incorporated into the platform without the need to develop a new mobile application for each solution. The proposed architecture was structured to separate responsibilities related to the user interface, image management, communication with external services, and the presentation of results. As a result, the application now acts as a unified interaction layer between users and the several computer vision models available.

Also as part of this plan, the user flow was designed to address two complementary usage scenarios: a monitoring scenario, in which the healthcare professional links classifications to persistent profiles or collections to track the progression of a case, and a sporadic prediction scenario, in which the user performs one-off classifications without the need for linking. Modeling these two workflows meets the architecture's scalability requirement: the platform must accommodate both domains requiring structured history and domains in which each classification is self-contained. At the same time, the application's sitemap was defined, hierarchi-

cally organizing the groupings resulting from card sorting. The adopted nomenclature (New Record, Collections, History, Profile) sought to balance two criteria discussed by Rosenfeld, Morville, and Arango [25]: immediate clarity for the user and modular flexibility so that different application domains can reuse the same structure without renaming hierarchical levels. During the design process, the application's usage flows, the mechanisms for linking models and image sets, and the communication strategies with the inference services hosted on dedicated infrastructure were also defined. These definitions served as the basis for designing the system's logical architecture and for developing the solution's functional prototypes.

B. Prototyping

The prototyping step involved implementing the architecture designed in the previous step through the development of the Vizu platform. The goal of this step was to bring the proposed architectural concepts to life and verify their applicability in real-world use cases.

Prototyping was conducted at three levels of increasing fidelity, in accordance with the progressive approach advocated by Buxton [26] and Snyder [27], who argue that each level of fidelity serves a distinct investigative purpose and that moving prematurely to high fidelity makes iteration more costly and shifts the focus of the discussion from structure to aesthetics. At the skeleton level, low-fidelity wireframes were used to validate the rough hierarchy of elements and the sequence of screens, without any visual considerations. Medium-fidelity wireframes detailed components, states, and microinteractions, allowing for discussion of consistency across screens and affordances [28] without yet addressing visual identity decisions.

At the surface level, the development of high-fidelity prototypes was guided by the Atomic Design methodology proposed by Frost [29], which organizes interfaces into a hierarchy of components—atoms, molecules, organisms, templates, and pages—to promote visual consistency and the systematic reuse of elements. The adoption of this approach was consistent with the platform's central objective: since Vizu aims to accommodate different computer vision models under a single interface, a modular component system reduces the cost of adding new screens associated with new models.

The visual refinement followed established best practices in UI design. Spacing was standardized to multiples of 8 pixels, a grid system widely adopted in digital product design because it ensures mathematical consistency among components and adaptability across screen densities. Accessibility was guided by the Web Content Accessibility Guidelines (WCAG) 2.1, published by the W3C [30], with particular attention to minimum contrast ratios, touch area size, and typographic readability. The visual hierarchy and contrast were developed according to principles established in the interface design literature [31], aiming to direct the user's attention to the priority elements on each screen, particularly in the classification results flow, where the clarity of the information produced by the AI is critical to user trust. Microinteractions were designed based on UX motion principles [32], ensuring that transitions and feedback states intuitively communicate causality and progress, in line with the system state visibility heuristic [24].

The interface as a whole was revised in light of Nielsen's ten usability heuristics [24], with special attention paid to the visibility of the system's status (essential while waiting for inference), the correspondence between the system and the real world (nomenclature adaptable to different domains), error prevention (validations at the time of data entry and segmentation), and recognition rather than memorization (use of the Reference Atlas as a visual aid when there is disagreement with the result).

These components were designed to be reusable regardless of the AI model associated with the application. Next, the necessary mechanisms were implemented to enable the integration of different computer vision models into the platform, through communication with external services responsible for performing inferences, allowing the application to remain independent of the specific characteristics of each model. To validate the architectural implementation, models developed in the Dr. Bone and Pellis projects were incorporated into the platform. These use cases demonstrated the solution's ability to support different medical image analysis tasks using the same mobile application framework. By the end of this stage, a functional version of Vizu had been produced, capable of centralizing interaction with multiple healthcare-related computer vision models and laying the groundwork for the evaluation activities planned for the subsequent stages of the MIDTs method.

VI. Discussion

The scientific contribution of VIZU is therefore not limited to the presentation of a mobile application project, but rather lies in the proposal of an architecture that centralizes CV models and handles inference through a single, end-user-friendly interface. Moreover, given the cutting-edge nature of the study, the possibility of adding new models or versioning existing ones is also taken into account. This section accordingly discusses hypothetical scenarios for the use of VIZU in health-oriented applications. Although the architecture has already been validated in two similar use cases, this section focuses on exploring new possibilities of interaction between AI models and VIZU, as well as how a user could employ the models in a simple manner and make use of the generic functionalities of the application (e.g., adding to a collection, segmenting an image) to facilitate their use.

First, VIZU could assist in the analysis of chest X-ray images to identify possible pulmonary alterations. In this scenario, the user would upload files from the gallery and select the inference model that classifies the image. Making such a model available for use on mobile devices would otherwise require a separate application, with functionalities developed from scratch. Using the proposed architecture, the integration of the CV model would occur through the association of its services with the VIZU model hub, reusing

the authentication, image capture, visual feedback, and other elements that compose the user interface.

Analogously, considering different scenarios, such as the classification of skin lesions from clinical photographs, or the identification of signs of diabetic retinopathy in fundus images, adherence to VIZU is likewise feasible. In both cases, the flow would be very similar to the previous one: an image is sent to the server, which loads it into the model to infer the class and, after processing, returns the result to the user. Although these involve different medical specialties, the proposed architecture allows a new model to be incorporated without significant changes to the structure of the mobile application, demonstrating its capacity to accommodate different medical specialties from a common set of functionalities.

From this perspective, the replicability of the architecture represents a relevant contribution to the development of AI solutions in health. Rather than treating each new model as an independent project, the proposal favors the systematic reuse of components, the standardization of the user experience, and the reduction of the effort required to deploy new models in a mobile environment. The architecture can thus act as a common foundation for different research and development initiatives that share the workflow of processing medical images through CV models. This replicability rests on a small set of design choices: centralizing otherwise dispersed models behind a single interface, standardizing the inference flow across specialties, and treating user feedback, including disagreement with a result, as a first-class input to model improvement.

However, some limitations should be acknowledged. The use cases discussed in this section are hypothetical and intended to illustrate the flexibility of the architecture rather than to report clinical validation; the platform itself remains at an early prototype stage. Although the architecture has been exercised in two similar use cases, its behavior under a broader range of models, image modalities, and real-world usage conditions has not yet been systematically evaluated. Furthermore, compliance with data protection requirements, while embedded in the design, requires rigorous validation before any handling of real patient data, and the quality of the resulting analyses ultimately depends on the individual models integrated into the hub. These aspects motivate the evaluation plan outlined in the following section.

VII. Future Work and Evaluation Plan

As a continuation of this work, we intend to complete the prototyping of VIZU, implementing all components in a generic and adaptable manner so as to support any image-based AI inference interaction. In addition, expanding the proposed architecture to encompass new application scenarios in health, with other CV models, is also part of the researchers' plans. Further improvements are foreseen regarding the dynamic configuration of models, user management, and interoperability mechanisms with systems used in clinical and research environments.

In line with these steps, once the prototype development is complete, we plan to assess the quality and consistency of the application through indicators and through usability tests conducted with user-experience (UX) professionals.

A. Evaluation Plan

The evaluation will be structured along complementary dimensions that together substantiate the central claim of this work, that a centralized, user-facing inference interface lowers the barrier to deploying CV models in health: **Usability:** task-based usability testing conducted with UX professionals and representative users, assessing the submission and inference experience across varied technical profiles. **Application quality and consistency:** measurement of indicators such as the reliability of the inference workflow, the consistency of results returned through the interface, and the responsiveness of the visual-feedback mechanisms. **Architecture replicability:** verification that new CV models can be integrated into the model hub with minimal changes to the mobile application, quantifying the effort required to add or version a model. **Privacy and compliance:** validation of the handling of sensitive data, including the de-identification of medical images, prior to any use of real patient data.

B. Functional Roadmap

The planned extensions define a path from the current prototype toward a validated platform. Beyond completing a fully generic component set, these include dynamic model configuration, allowing models to be added or versioned without modifying the application; richer user-management capabilities; and interoperability with clinical and research infrastructures. Ensuring usability, application quality, replicability, and privacy compliance, in that order, would consolidate VIZU as a common foundation for research and development initiatives that share the workflow of processing medical images through CV models.

VIII. Conclusion

The growing use of AI models in healthcare has driven the development of specialized applications for different CV tasks. However, the need to implement similar functionalities across multiple projects reveals opportunities for adopting reusable and standardized approaches.

This work presented VIZU, a mobile architecture proposal designed for the integration, validation, and visualization of CV models applied to health. It was developed from the observation of recurrent requirements across distinct projects, seeking to centralize shared functionalities and reduce the need to develop independent applications for each new model. Unlike approaches focused exclusively on the management or distribution of AI models, the architecture proposes a unified mobile layer capable of accommodating different solutions under a single, user-centered interaction structure. Within this structure, post-result actions such as archiving an analysis or disagreeing with a model output are treated as integral parts of the workflow, allowing user feedback to support

TABLE III
COMPARISON OF RELATED WORK AND VIZU

Work	C1	C2	C3	C4	C5	C6
DLHub [10]	∼	∼	–	–	∼	–
ModelHub.AI [11]	∼	∼	–	∼	∼	–
Gallery [12]	–	–	–	–	–	∼
TFX [13]	–	–	–	–	–	∼
Michelangelo [14]	–	–	–	–	–	∼
SageMaker [15]	–	–	–	–	–	∼
MONAI [9]	✓	–	–	✓	∼	–
Health xAI [16]	✓	∼	–	–	–	∼
VIZU	✓	✓	✓	✓	✓	✓

C1: health domain; C2: end-user interface; C3: mobile layer; C4: computer vision; C5: user-driven model testing; C6: end-user feedback for retraining. ✓: fully; ∼: partially; –: not addressed.

the subsequent retraining and improvement of the integrated models.

Table III positions VIZU against the reviewed works using the same dimensions. Unlike approaches centered on publishing, serving, or managing the lifecycle of models (generally from a developer or clinical-IT integration perspective), VIZU intends to unify a mobile, end-user-facing layer for testing, validating, and visualizing computer-vision models in healthcare, while allowing user feedback to support subsequent retraining. The comparison highlights that the contribution of this architecture lies not in any single dimension in isolation, but in their integration into a replicable architecture that lowers the barrier to making new models available through a common, user-centered interface.

As a prototype, aspects related to the quantitative evaluation of the architecture, the quality of the datasets produced, and the impact of the platform on the development and evolution cycles of models remain as future research. Nevertheless, the results obtained during development indicate that the proposed approach represents a promising alternative to support research and development initiatives in computer vision applied to health, contributing to the construction of more reusable, consistent, and scalable solutions.

REFERENCES

[1] F. Kitsios, M. Kamariotou, A. I. Syngelakis, and M. A. Talias, "Recent Advances of Artificial Intelligence in Healthcare: A Systematic Literature Review," Applied Sciences, vol. 13, no. 13, p. 7479, Jun. 2023, doi: 10.3390/app13137479.

[2] C.-C. Lee, E. C. So, L. Saidy, and M.-J. Wang, "Lung Field Segmentation in Chest X-ray Images Using Superpixel Resizing and Encoder–Decoder Segmentation Networks," Bioengineering, vol. 9, no. 8, p. 351, 2022, doi: 10.3390/bioengineering9080351.

[3] Y. Su, Q. Liu, W. Xie, and P. Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms," Computer Methods and Programs in Biomedicine, vol. 221, Art. no. 106903, 2022, doi: 10.1016/j.cmpb.2022.106903.

[4] M. J. Mortada, S. Tomassini, H. Anbar, M. Morettini, L. Burattini, and A. Sbrollini, "Segmentation of Anatomical Structures of the Left Heart from Echocardiographic Images Using Deep Learning," Diagnostics, vol. 13, no. 10, p. 1683, 2023, doi: 10.3390/diagnostics13101683.

[5] K. W. Lee and R. K. Y. Chin, "Diverse COVID-19 CT Image-to-Image Translation with Stacked Residual Dropout," Bioengineering, vol. 9, no. 11, p. 698, 2022, doi: 10.3390/bioengineering9110698.

[6] R. W. Y. Granzier, N. M. H. Verbakel, A. Ibrahim, J. E. van Timmeren, T. J. A. van Nijnatten, R. T. H. Leijenaar, M. B. I. Lobbes, M. L. Smidt, and H. C. Woodruff, "MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability," Scientific Reports, vol. 10, no. 1, Art. no. 14163, 2020, doi: 10.1038/s41598-020-70940-z.

[7] R. L. Barbosa *et al.*, "Doctor Bone: treinamento de redes neurais para auxílio na identificação da idade óssea", J Health Inform, vol. 16, nº Especial, nov. 2024, doi: 10.59681/2175-4411.v16.iEspecial.2024.1382

[8] J. E. V. Filho, J. S. C. Neto, and J. F. Ferreira, "MIDTs: Interdisciplinary Method for Technological Research Development with a Focus on Health," in Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF), Porto, Portugal, 2025, pp. 874–881.

[9] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang *et al.*, "MONAI: an open-source framework for deep learning in healthcare," 2022, arXiv:2211.02701.

[10] R. Chard, Z. Li, K. Chard, L. Ward, Y. Babuji, A. Woodard, S. Tuecke, B. Blaiszik, M. J. Franklin, and I. Foster, "DLHub: model and data serving for science," in Proc. IEEE Int. Parallel Distrib. Process. Symp. (IPDPS), 2019, pp. 283–292.

[11] A. Hosny, M. Schwier, C. Berger, E. P. Örnek, M. Turan, P. V. Tran *et al.*, "ModelHub.AI: dissemination platform for deep learning models," 2019, arXiv:1911.13218.

[12] C. Sun *et al.*, "Gallery: a machine learning model management system at Uber," in Proc. 23rd Int. Conf. Extending Database Technology (EDBT), 2020.

[13] D. Baylor, E. Breck, H.-T. Cheng, N. Fiedel, C. Y. Foo, Z. Haque *et al.*, "TFX: a TensorFlow-based production-scale machine learning platform," in Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2017, pp. 1387–1395.

[14] J. Hermann and M. Del Balso, "Meet Michelangelo: Uber's machine learning platform," Uber Engineering Blog, 2017.

[15] E. Liberty, Z. Karnin, B. Xiang, L. Rouesnel, B. Coskun, R. Nallapati *et al.*, "Elastic machine learning algorithms in Amazon SageMaker," in Proc. ACM SIGMOD Int. Conf. Management of Data (SIGMOD), 2020, pp. 731–737.

[16] H. S. Adnan, A. Shidani, L. Clifton, C. R. Bankhead, and R. Perera-Salazar, "Implementation framework for AI deployment at scale in healthcare systems," iScience, vol. 28, no. 5, p. 112406, Apr. 2025.

[17] J. J. Garrett, *The Elements of User Experience: User-Centered Design for the Web and Beyond*, 2nd ed. Berkeley: New Riders, 2011.

[18] M. Q. Patton, *Qualitative Research and Evaluation Methods*, 3rd ed. Thousand Oaks: Sage, 2002.

[19] J. Gothelf and J. Seiden, *Lean UX: Applying Lean Principles to Improve User Experience*. Sebastopol: O'Reilly Media, 2013.

[20] J. Saldaña, *The Coding Manual for Qualitative Researchers*, 2nd ed. London: Sage, 2013.

[21] J. Kawakita, *The Original KJ Method*. Tokyo: Kawakita Research Institute, 1986.

[22] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems*. San Francisco: Morgan Kaufmann, 1998.

[23] D. Spencer, *Card Sorting: Designing Usable Categories*. New York: Rosenfeld Media, 2009.

[24] J. Nielsen, "Heuristic evaluation," in *Usability Inspection Methods*, J. Nielsen and R. L. Mack, Eds. New York: John Wiley & Sons, 1994, pp. 25–62.

[25] L. Rosenfeld, P. Morville, and J. Arango, *Information Architecture: For the Web and Beyond*, 4th ed. Sebastopol: O'Reilly Media, 2015.

[26] B. Buxton, *Sketching User Experiences: Getting the Design Right and the Right Design*. San Francisco: Morgan Kaufmann, 2007.

[27] C. Snyder, *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces*. San Francisco: Morgan Kaufmann, 2003.

[28] D. A. Norman, *The Design of Everyday Things*, revised ed. New York: Basic Books, 2013.

[29] B. Frost, *Atomic Design*. 2016. [Online]. Available: https://atomicdesign.bradfrost.com

[30] World Wide Web Consortium, "Web Content Accessibility Guidelines (WCAG) 2.1," W3C Recommendation, Jun. 2018. [Online]. Available: https://www.w3.org/TR/WCAG21/

[31] J. Tidwell, *Designing Interfaces: Patterns for Effective Interaction Design*, 2nd ed. Sebastopol: O'Reilly Media, 2010.

[32] V. Head, *Designing Interface Animation: Meaningful Motion for User Experience*. New York: Rosenfeld Media, 2016.

A Monitoring System Prototype For Vascular Neck Restraint Applied By Law Enforcement Officers

Carlos Eduardo Silva
PPGEB
Univ. Tecnológica Federal do Paraná
Polícia Militar do Paraná
Curitiba, Brazil
orcid.org/0009-0007-9347-1395

Lucas Daniel Ortiz Arias
DAINF/DAELN/CPGEI
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0009-0005-3393-4445

Mauro Aloisio Smaniotto
CPGEI
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0009-0003-9507-2022

Fabio Kurt Schneider
DAELN/CPGEI/PPGEB
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0000-0001-6916-1361

Julio Cesar Bassan
PPGEB/PPGEF
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0000-0002-6910-9077

***Abstract*—Vascular neck restraint (VNR) is a submission and control technique used in combat sports and law enforcement defensive tactics to temporarily incapacitate a resistant individual through cerebral hypoperfusion while minimizing airway compression. Despite its widespread application, the evaluation of VNR execution remains predominantly subjective, relying on visual assessment by instructors. This study developed and evaluated a prototype monitoring system, termed the Cervical Constriction Sensor, designed to objectively quantify pressure distribution across the right carotid, left carotid, and tracheal regions using an anthropomorphic training simulator. Twenty-two military police officers performed the technique before and after receiving specific biomechanical instruction based on the official defensive tactics manual. The results demonstrated a significant redistribution of pressure following technical guidance: mean peak pressure increased by 24.6 percent on the left carotid and 225.3 percent on the right carotid, while decreasing by 43 percent over the tracheal region. These findings support the hypothesis that objective monitoring provides valuable quantitative feedback for training quality and consistency, by characterizing whether force is directed toward the intended anatomical targets on the simulator rather than the tracheal region. As this study was conducted exclusively on an instrumented mannequin, no claims regarding operational safety in human subjects can be derived from these results.**

***Index Terms*—Law Enforcement Officers, Monitoring System for Vascular Neck Restraint, Rear Naked Choke, Vascular Neck Restraint.**

I. Introduction

Within the context of Fighting, Martial Arts, and Combat Sports (LAMEC, the Brazilian acronym for Lutas, Artes Marciais e Esportes de Combate), cervical restraint techniques have been employed for centuries as methods of controlling or incapacitating an opponent. Among these techniques, the Rear Naked Choke (RNC), known as hadaka-jime in Japanese martial arts [1] and mata-leão in Brazilian Jiu-Jitsu, is one of the most widely recognized submission maneuvers. When properly executed, the technique aims to compress the lateral structures of the neck while minimizing direct pressure on the airway, producing temporary incapacitation through vascular mechanisms.

Due to its effectiveness as a control technique, adaptations of vascular neck restraint (VNR) have also been incorporated into the defensive tactics programs of several law enforcement agencies. According to the Paraná Military Police Defense Manual [2], VNR is classified as a submission and control technique intended to neutralize aggressive or resistant behavior while reducing the need for higher levels of force. Consequently, proper technical execution is essential to maximize effectiveness and minimize the risk of adverse outcomes.

When applying this technique, the neck is positioned at the apex of the ”V” formed by the arm and forearm (Fig. 1). The operational priority is bilateral compression of the cervical region to promote carotid constriction, while respiratory compression is intended to remain a secondary effect.

Despite its widespread use in both combat sports and law enforcement environments, assessment of VNR execution remains predominantly subjective. Instructors typically rely on visual observation and practical experience to evaluate performance, making it difficult to objectively determine whether pressure is being applied to the intended anatomical targets. As a result, trainees often receive limited quantitative feedback regarding the quality, effectiveness, and safety of their technique. Therefore, the aim of this study was to develop and evaluate a prototype monitoring system capable of measuring pressure distribution during the execution of a vascular neck restraint technique. The proposed device was designed to simultaneously monitor pressure applied to the right carotid region, left carotid region, and anterior cervical region corresponding to the trachea. In the context of VNR instruction, such systems may provide valuable information

This work had the support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and PMPR



Fig. 1. Application of vascular neck restraint (VNR) by a Paraná Military Police officer, highlighting main pressure over the carotids. Reproduced from [2]

regarding pressure distribution across cervical structures and support evidence-based training methodologies.

It was hypothesized that, following specific instruction regarding the biomechanical principles of VNR execution, participants would demonstrate a pressure distribution pattern on the simulator characterized by increased pressure over the carotid regions and reduced pressure over the anterior cervical region. Such findings, if observed, would support the use of objective monitoring as an instrument to characterize technique execution on a surrogate device, without implying validation of the technique's safety when applied to human subjects — a question that lies outside the scope of the present study and would require independent ethical and physiological evaluation.

II. Theoretical Framework

A. Vascular Neck Restraint and Mechanisms of Incapacitation

Vascular Neck Restraint (VNR) is designed to temporarily incapacitate an individual through compression of the lateral cervical structures, particularly the carotid arteries. Unlike airway-based choking mechanisms, properly executed VNR seeks to reduce cerebral perfusion while minimizing direct compression of the trachea and laryngeal structures. Consequently, the resulting loss of consciousness is primarily associated with cerebral hypoperfusion rather than respiratory obstruction. Mitchell et al. [3] investigated the physiological mechanisms associated with blood choke techniques and reported that effective carotid compression may induce loss of consciousness within approximately 5 to 11 s. The authors concluded that reduced cerebral blood flow resulting from carotid artery compression represents the principal mechanism responsible for transient incapacitation. Furthermore, recovery was generally rapid and complete when the restraint was promptly released. Additional evidence was provided by Stellpflug et al. [4], who analyzed validated choking events involving highly trained combat athletes. The authors observed a mean time to loss of consciousness of approximately 9 s following establishment of an effective chokehold. The similarity between these findings and previous physiological investigations reinforces the role of cerebral hypoperfusion as the primary mechanism underlying vascular restraint techniques. The effectiveness of VNR has been investigated in operational law enforcement environments. Bozeman et al. [5] analyzed 944 field applications performed by trained police officers over an 11-year period and reported a success rate of 92.6% in controlling resistant individuals. Importantly, the authors found no fatalities or severe injuries directly attributable to the technique during the study period. Comparable findings were reported by Hickman et al. [6] in an evaluation of VNR use by the Spokane Police Department. The authors observed no fatalities associated with the restraint and documented relatively low injury rates among subjects. Together, these findings suggest that VNR may constitute an effective and relatively safe control option when applied by properly trained personnel according to established protocols. Although available evidence supports the effectiveness and safety of VNR, complications may occur when the technique is improperly applied. Excessive force, prolonged compression, inadequate positioning, or inappropriate force distribution may increase the risk of injury to cervical structures. Demartini et al. [7] reported a case of internal carotid artery dissection in a Brazilian Jiu-Jitsu practitioner following application of a Rear Naked Choke. The authors proposed that arterial compression associated with cervical hyperextension and contralateral head rotation contributed to vascular injury. Although such events appear to be rare, they highlight the importance of proper instruction, supervision, and technical execution. Similarly, reports of severe neurological injury following prolonged strangulation demonstrate that inappropriate cervical compression may produce significant adverse consequences. Collectively, these observations emphasize that the effectiveness and safety of VNR are directly related to the practitioner's ability to apply force to appropriate anatomical targets while avoiding excessive compression of non-target structures.

B. Cervical Anatomy

The neck is organized into muscular compartments that define the principal pathways for vascular structures. Among these, the carotid triangle (Fig. 2), a subdivision of the anterior triangle, represents the main region of surgical interest for access to the carotid arteries. Its boundaries are formed by the superior belly of the omohyoid muscle anteriorly, the posterior belly of the digastric muscle superiorly, and the anterior border of the sternocleidomastoid muscle posteriorly. Throughout its cervical course, the common carotid artery (CCA) is enclosed within the carotid sheath, a condensation of the deep cervical fascia. Within this sheath, the artery occupies the medial position, the internal jugular vein lies laterally, and the vagus nerve is located posteromedially to

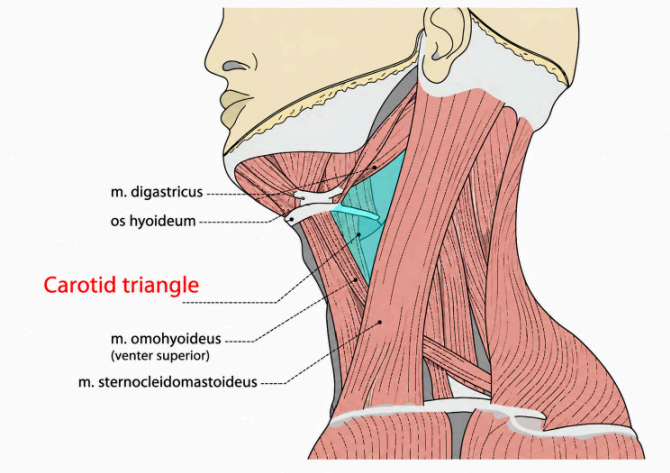


Fig. 2. Illustration of the carotid triangle's location, delimited by the sternocleidomastoid, omohyoid, and digastric muscles, highlighting its spatial relationship with the hyoid bone. Reproduced from [10]

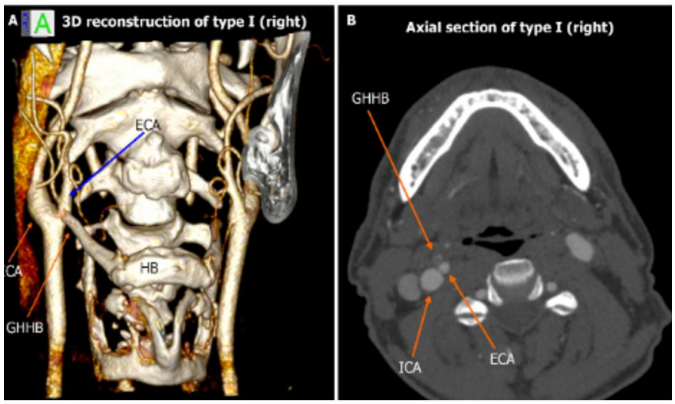


Fig. 3. anatomical relationship between cervical structures. ECA: External carotid artery; GHHB: Greater horn of the hyoid bone; HB: Hyoid bone; ICA: Internal carotid artery. Reproduced from [12]

both vessels. [8] The origin of the common carotid arteries presents a classic anatomical asymmetry. On the right, the common carotid artery (CCA) arises from the bifurcation of the brachiocephalic trunk, whereas on the left it originates directly from the aortic arch, presenting an intrathoracic segment before ascending into the neck. The CCA ascends without giving off collateral branches until reaching the superior border of the thyroid cartilage, where it bifurcates into the internal and external carotid arteries. At this level are located two structures essential for cardiovascular homeostasis: the carotid sinus and the carotid body. The carotid sinus is a dilation at the origin of the internal carotid artery that contains baroreceptors sensitive to changes in blood pressure, whereas the carotid body is a small cluster of chemoreceptor tissue responsible for monitoring blood oxygen and carbon dioxide levels. [9]

C. Physiology And Anatomy Of The Carotid Compression

As illustrated in Fig. 3, the thyroid cartilage (and above it the hyoid bone) forms a rigid structure that protects the larynx and trachea. The literature on VNR biomechanics describes the technique as forming a "V" shape between the arm and forearm, in which pressure is intended to be exerted on the lateral soft tissues (where the carotid and jugular veins are located), while the apex of the "V" remains aligned with this rigid structure. This description is presented here strictly as theoretical background to contextualize the anatomical regions monitored by the proposed sensor system, and should not be interpreted as operational guidance for technique application.

Loss of consciousness following the proper application of a carotid constriction technique does not result from asphyxia, but rather from a combination of cerebral hypoperfusion and autonomic responses. According to Mitchell et al. [3], the abrupt interruption of carotid blood flow reduces cerebral perfusion pressure. In addition, pressure exerted on the baroreceptors located at the carotid bifurcation triggers a vagal response, resulting in bradycardia and systemic vasodilation, thereby accelerating the onset of loss of consciousness. [3]

The forensic pathology literature establishes that the pressure thresholds for interruption of hemodynamic flow are substantially lower than those required for functional airway collapse. According to DiMaio and DiMaio [11], occlusion of the common carotid arteries requires a load of approximately 11 lb (5 kg), while total occlusion of the trachea requires a compression of about 33 lb (15 kg). Although considered a safe technique when interrupted immediately after withdrawal or syncope, there are risks of indirect vascular trauma. The literature reports rare cases of carotid artery dissection, such as a case where combined neck extension and head rotation (as a defense reaction) elongated the compressed carotid artery at its origin, causing dissection. [7]

III. Materials And Methods

This cross-sectional, comparative quantitative study, following the methodological principles described by Thomas et al. [13]. A non-probabilistic convenience sampling strategy was adopted for participant recruitment.

A. Data Acquisition

The sample consisted of 22 adult military police officers enrolled in the Sergeants' Course of the Paraná Military Police (PMPR), including 20 men and 2 women. All participants were right-handed, a characteristic that occurred randomly within the sample. The inclusion criteria were: (I) being a military police officer; (II) currently attending the Sergeants' Course of the Paraná Military Police (PMPR); and (III) not having injuries to the upper limbs or spine that could interfere with the execution of the technique. Participants were questioned regarding their practice of grappling martial arts, and only four reported previous experience, with three at an advanced level and one at a beginner level. However, all participants had previously attended police self-defense training and were familiar

with the VNR technique. During Trial A, each participant positioned themselves behind the anthropomorphic mannequin and executed the VNR motor gesture upon command, applying maximum force until voluntarily interrupting the maneuver. After completing the first trial, participants received standardized technical instruction regarding the proper positioning of the upper limbs according to the Police Defense Manual, while the researcher also corrected the positioning of the arm vertex on the simulator. The procedure was then repeated under identical conditions (Trial B). As a strategy to ensure internal control and experimental validity, the results obtained were not disclosed to the participants at the time of execution. This partial blinding technique aimed to mitigate the Hawthorne effect, characterized by behavioral alterations that occur when individuals become aware that they are participating in a study or having their behavior assessed. By withholding immediate feedback, the protocol sought to prevent conformity to perceived performance expectations or social desirability responses, which could otherwise introduce profound biases into the data reliability. [14]

B. Proposed Instrument

Performance measurement was carried out using a prototype developed by these researchers, called the ”cervical constriction sensor”. The instrument consists of a type of ”cervical collar” made of resilient material (foam and textile covering), adapted to the cervical circumference of an anthropomorphic combat training simulator, of the ”Bob” type. (Fig. 4).

Three pressure sensors were inserted into the system, positioned in specific locations: a) anterior cervical region, in the midline, simulating the position of the trachea; b) in the carotid triangle, right anterolateral region; c) in the carotid triangle, left anterolateral region.

The prototype utilized remained identical across all data collections, and the height of the anthropomorphic simulator was kept constant, with a height-raising platform being provided to the participants for a better technical fit at their own discretion.

The developed prototype consists of an embedded system driven by an ESP32 microcontroller, designed to quantify the pressure exerted on specific anatomical regions during the execution of the Vascular Neck Restraint (VNR) technique. Data captured by the pressure sensors are transmitted via Bluetooth Low Energy (BLE) protocol to a mobile platform, enabling real-time monitoring and data storage for subsequent statistical analysis. Consequently, the device is capable of mapping the simultaneous force distribution across the three cervical points of interest.

The pressure sensing mechanism was structured by coupling each sensor to a flexible tubing segment, with mechanical properties analogous to those found in standard sphygmomanometer cuffs. To enable signal transduction, this pneumatic circuit remained hermetically sealed at one end. This configuration ensured that any external mechanical compression applied to the tubing surface resulted strictly in a change in the internal pressure of the enclosed air volume, allowing this barometric variation to be isolated, detected, and accurately quantified by its respective transducer.



Fig. 4. Prototype installed on an anthropomorphic simulator for pressure data collection.

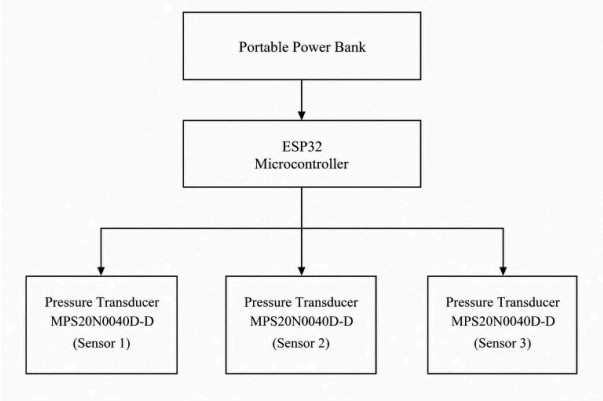


Fig. 5. Block diagram of the proposed embedded system hardware architecture.

C. Hardware

The hardware architecture of the proposed monitoring device is illustrated in (Fig. 5) and comprises three integrated pressure transducers. During the component selection phase, the BMP280 and MPS20N0040D-D transducers were evaluated, with the MPS model being selected due to its superior alignment with the project's efficiency requirements. The MPS20N0040D-D is a gauge/differential piezoresistive pressure sensor (up to 40 kPa) operating on Micro-Electromechanical Systems (MEMS) technology. It incorporates a sensing chip featuring an elastic membrane integrated with a four-piezoresistor array arranged in a Wheatstone Bridge configuration.

D. Software

As illustrated in Fig. 6, the software architecture of the system is divided into two distinct development branches: the embedded firmware within the ESP32 microcontroller and a

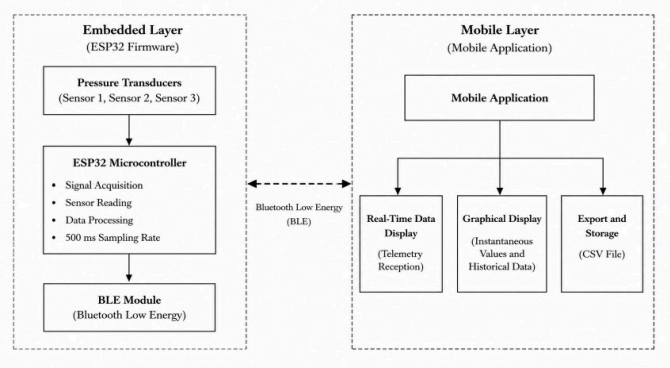


Fig. 6. Software and communication architecture of the proposed monitoring system.

dedicated application tailored for the mobile platform. The embedded software continuously scans and reads the analog signals from the pressure transducers, transmitting these data vectors to the mobile device at a constant sampling interval of 500 ms.

Wireless communication between the microcontroller and the mobile device is established via the Bluetooth Low Energy (BLE) protocol, selected for its energy efficiency, ease of implementation, and widespread standardization in embedded systems. On the receiving end, the mobile application features three core functionalities: (i) real-time reception of telemetry pressure data from the ESP32; (ii) simultaneous graphical display of instantaneous values and temporal measurement history for continuous trial monitoring; and (iii) exporting and storing these time-series datasets as .CSV files.

The 500 ms sampling rate was strategically selected as a design trade-off to optimize power consumption in the portable ESP32 microcontroller and to ensure reliable BLE packet transmission without data loss. Given that the vascular neck restraint (VNR) is characterized by sustained, quasi-static isometric compression rather than high-frequency dynamic impacts, this 2 Hz acquisition rate is mathematically sufficient to accurately capture the pressure plateaus and stable force distributions exerted during the hold phase of the maneuver.

IV. Results And Discussion

A. Results

The peak pressure values recorded by the proposed monitoring system were analyzed for the left carotid, tracheal, and right carotid regions during the two experimental conditions: Trial A (before technical instruction) and Trial B (after technical instruction). Descriptive statistics, including minimum, first quartile (Q_1), median, mean, third quartile (Q_3), and maximum values, are presented in Table II, while the mean peak pressures and overall variations are summarized in Table I. The pressure distribution across the monitored anatomical regions is illustrated in (Fig. 7), and the statistical distribution of individual pressure differences between trials is presented in (Fig. 8) and detailed in Table III.

Technical instruction produced a clear, systematic redistribution of pressure across the cervical region. Mean peak pressure applied to the left carotid increased from 1088.1 ± 538.9 to 1356.1 ± 459.6 sensor units, corresponding to an increase of approximately $+24.6\%$. A far more pronounced change was observed in the right carotid region, where the mean pressure escalated from 364.0 ± 436.8 to 1184.1 ± 732.9 sensor units, representing a substantial increase of approximately $+225.3\%$. Conversely, the mean peak pressure recorded in the tracheal region decreased significantly from 2002.0 ± 572.2 to 1124.7 ± 493.4 sensor units, corresponding to a targeted reduction of -43.0%.

The descriptive statistics further demonstrate that these observed changes were structurally consistent across the entire sample rather than being driven by isolated or extreme observations. Before instruction, the median pressure recorded in the right carotid region was 201.72 sensor units, which dramatically increased to 967.6 sensor units in Trial B, while its maximum value reached 3130.1 units. Likewise, the median pressure in the left carotid region increased from 1063.8 to 1370.9 sensor units. In stark contrast, the median tracheal pressure was halved, dropping from 2041.0 sensor units in Trial A to 1006.7 sensor units in Trial B, with the minimum recorded tracheal force falling to 341.1 units.

The net pressure differences calculated for each participant ($Trial\ B - Trial\ A$) and summarized in Table III reinforce this pressure redistribution pattern. The median increase in the left carotid pressure reached 456.5 sensor units, while the median increase over the right carotid region reached 701.9 sensor units, with a maximum individual increase of 2241.9 units. Simultaneously, the median reduction in tracheal loading reached -914.3 sensor units, with maximum decompressions reaching -1735.9 units. As illustrated by the boxplots in Fig. 8, this redistribution pattern was consistently observed among participants, indicating an association between the instructional intervention and a shift of mechanical force away from the tracheal region and toward the carotid regions monitored by the sensor system. The system was evaluated for hysteresis with increasing loads of 10, 20, 30, 40, 50, and 60 kg, respectively, distributed over the 3 tubes attached to the sensors, followed by the successive decreasing of load (i.e., 50, 40, 30, 20, and 10 kg), showing that the system has a strong dependency of its past input values. This result is shown in Fig. 9.

To establish these experimental equivalences, a physical calibration setup was developed. A mercury column was directly coupled to the same arterial analog rubber tube (8 mm outer diameter) connected to the pressure sensor, allowing for direct hydrostatic calibration. Concurrently, a mechanical calibration was performed by applying standard dead weights over a controlled compression area of 8 mm $\times$ 8 cm. Through this dual-calibration approach, it was observed that a variation of 100 sensor units is equivalent to 1 mmHg of internal pressure, which under this specific experimental setup also corresponds to approximately 1 kgf of externally applied force over the 8 mm $\times$ 8 cm tube surface area. This equivalence translates

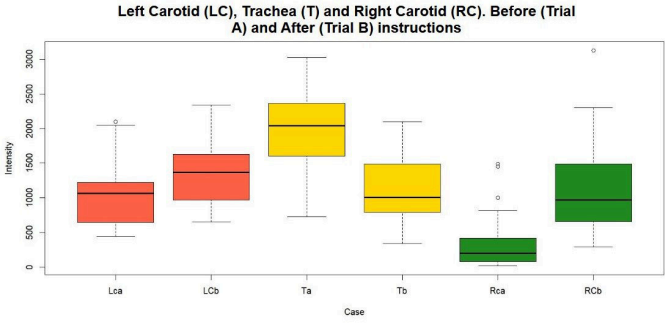


Fig. 7. Effort distribution for the regions of Left Carotid (LC) in red, Trachea (T) in yellow and Right Carotid (RC) in green

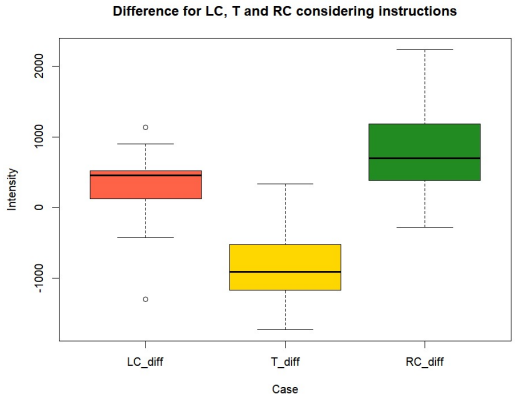


Fig. 8. Distribution of the differences between effort before and after instructions on proper vascular neck restraint protocol

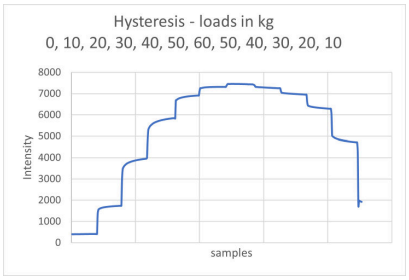


Fig. 9. Intensity measurements for adding loads of 0, 10, 20, 30, 40, 50, 60, 50, 40, 30, 20, and 10 kg, respectively, showing the hysteresis of the system.

the median baseline values (detailed in Table II) to a range spanning from 2.01 mmHg for RC_a (Right Carotid - Trial A) to 20.41 mmHg for T_a (Trachea - Trial A). Furthermore, the highest equivalent pressures recorded during the experiments reached 23.43 mmHg for LC_b (2342.8 units) and 31.30 mmHg for RC_b (3130.1 units).

Nevertheless, it is critical to note that a more refined calibration is necessary in future developments, given the biomechanical differences between the synthetic rubber tube and actual human arteries, as well as the dynamic movements of the human arm during the restraint, which can introduce transient artifacts in the recorded peak pressure values.

TABLE I
MEAN PEAK PRESSURE VALUES AND PERCENTAGE VARIATION BEFORE (TRIAL A) AND AFTER (TRIAL B) TECHNICAL INSTRUCTION.

Region	Trial A ($Mean \pm SD$)	Trial B ($Mean \pm SD$)	Variation
Left Carotid	1088.1 ± 538.9	1356.1 ± 459.6	$+24.6\%$
Trachea	2002.0 ± 572.2	1124.7 ± 493.4	-43.0%
Right Carotid	364.0 ± 436.8	1184.1 ± 732.9	$+225.3\%$

TABLE II
DESCRIPTIVE STATISTICS OF PEAK PRESSURES ACROSS MONITORED REGIONS BEFORE (TRIAL A) AND AFTER (TRIAL B) TECHNICAL INSTRUCTION

Condition	Min	1st Qu.	Median	Mean	3rd Qu.	Max
LCa (Left Carotid - Before)	445.0	671.3	1063.8	1088.1	1219.2	2099.8
LCb (Left Carotid - After)	654.0	969.9	1370.9	1356.1	1629.4	2342.8
Ta (Trachea - Before)	729.0	1629.0	2041.0	2002.0	2357.0	3025.0
Tb (Trachea - After)	341.1	798.7	1006.7	1124.7	1476.2	2101.9
RCa (Right Carotid - Before)	14.2	81.7	201.7	363.9	406.5	1488.2
RCb (Right Carotid - After)	289.9	668.0	967.6	1184.1	1453.8	3130.1

TABLE III
STATISTICAL DISTRIBUTION OF PRESSURE DIFFERENCES POST-INSTRUCTION ACROSS MONITORED REGIONS

Metric	Min	1st Qu.	Median	Mean	3rd Qu.	Max
LC_diff (Left Carotid)	-1306.6	126.7	456.5	268.0	520.1	1138.3
T_diff (Trachea)	-1735.9	-1164.9	-914.3	-877.2	-530.9	332.7
RC_diff (Right Carotid)	-283.9	418.1	701.9	820.2	1173.7	2241.9

B. Discussion

The present study demonstrated that technical instruction substantially modified the pressure distribution pattern during the execution of the vascular neck restraint (VNR). Following instruction, pressure increased over both carotid regions while decreasing over the anterior cervical region, indicating that participants successfully altered the biomechanical application of force during the restraint. Importantly, this redistribution was consistently reflected across the descriptive statistics—including the mean, median, and quartile values presented in Tables I, II, and III—suggesting that the observed effect was highly representative of the study population rather than the result of isolated or extreme measurements.

The observed pressure redistribution strongly agrees with the physiological mechanisms previously described for vascular neck restraints. Mitchell *et al.* [3] demonstrated that transient loss of consciousness during vascular restraint is primarily associated with a severe reduction in cerebral perfusion resulting from bilateral carotid artery compression rather than airway obstruction. Likewise, DiMaio and DiMaio [11] reported that carotid artery occlusion requires considerably less compressive force than complete tracheal occlusion, emphasizing that an effective vascular restraint must preferentially direct pressure toward the lateral cervical structures instead of the anterior airway. In the present study, the robust reduction in tracheal pressure accompanied by increased pressure over both carotid regions indicates that, after instruction, participants executed the technique in a manner more closely aligned with these physiological principles.

The changes observed after instruction also reinforce the vital importance of structured technical training. Before instruction, the descriptive statistics demonstrated substantial variability in pressure distribution, particularly in the right carotid region, where the median pressure remained considerably lower than the values observed after technical guidance. Following instruction, both the median and mean values increased markedly, indicating that a greater proportion of participants applied pressure to the intended anatomical targets. These findings suggest that the instructional protocol contributed to increasing carotid loading and to improving the overall consistency of pressure application observed on the simulator, without implying a validated safety outcome for human application.

These observations are consistent with previous investigations involving vascular restraint techniques in law enforcement and combat sports. Bozeman *et al.* [5] reported that VNR can be performed safely when applied by properly trained officers according to established operational protocols. Similarly, Stellpflug *et al.* [4] demonstrated that effective sportive chokeholds produce rapid, transient loss of consciousness through vascular compression in highly trained athletes. Although the present study did not directly evaluate physiological outcomes such as cerebral blood velocity or loss of consciousness, the pressure redistribution observed on the simulator after instruction is consistent with the mechanical pattern described in the literature as associated with carotid-directed compression. This mechanical consistency should not be interpreted as evidence that the technique becomes safe when applied to a human being; such a claim would require dedicated physiological and safety studies, which fall outside the scope of the present surrogate-device study.

An important contribution of the present study lies in the objective assessment of technique execution. Conventional defensive tactics training is generally based on instructor observation and subjective evaluation of body positioning and movement. In contrast, the proposed monitoring system successfully quantified surface pressure simultaneously over the left carotid, right carotid, and anterior cervical regions, allowing direct visualization of force distribution throughout the maneuver. This objective feedback provides instructors with quantitative telemetry that complements traditional training methods and facilitates the identification of technical deficiencies that would otherwise remain undetected via visual assessment alone. It should be emphasized that this objective assessment characterizes force application on an instrumented mannequin and does not, by itself, establish that a given execution pattern is safe when applied to a person; VNR remains a high-risk technique whose use should continue to be governed by strict institutional protocols, training oversight, and independent ethical review, regardless of the mechanical results reported here.

From an engineering perspective, the developed embedded system successfully integrated pressure sensors, wireless communication, and mobile data acquisition into a single monitoring platform capable of recording multi-regional pressure distribution in real time. The prototype demonstrated sufficient sensitivity and resolution to distinguish contrasting execution patterns before and after technical instruction, supporting its feasibility as a practical tool for biomechanical assessment and tactical training. Furthermore, the proposed methodology establishes a rigorous experimental framework that can support future investigations involving the objective evaluation and optimization of vascular neck restraint techniques.

V. Limitations and Future Work

Several methodological and structural limitations must be considered when interpreting these results. First, pressure data were gathered using an anthropomorphic simulator rather than human subjects. Although this approach established a controlled and safe setup for initial hardware evaluation, the mechanical compliance of the synthetic rubber tubing differs substantially from the complex viscoelastic behavior of living human cervical tissues. The isotropic properties of the rubber fail to replicate the force dissipation mechanisms inherent to the skin, superficial fascia, muscle layers, and carotid sheath in vivo. Consequently, while the prototype reliably recorded relative force distribution patterns, the absolute pressures required to compress the pneumatic tubing cannot be linearly extrapolated to biological tissues.

Second, the data evaluation was restricted to peak pressure values recorded during each trial. Dynamic biomechanical parameters—such as pressure–time integrals, hold duration, and cumulative mechanical loading—were not quantified. These temporal variables are crucial for a comprehensive characterization of restraint mechanics.

Third, the study population was limited to military police officers enrolled in a single training program. This homogeneous cohort limits the direct generalization of the observed mechanics to other groups, such as combat sports practitioners or personnel from different law enforcement agencies.

Fourth, the experimental setup did not isolate the effects of technical instruction from confounding factors, including task repetition, trial order, and observer effects, as Trial A consistently preceded Trial B without a counterbalanced control group. Accordingly, the observed pressure shifts should be interpreted as correlated with, rather than strictly caused by, the instructional intervention.

Fifth, while the calibration protocol relied on traceable reference standards (a mercury column and calibrated weights), it characterized the sensor–tubing assembly under static bench-top conditions. The system was not cross-validated against an independent pressure-measurement system under dynamic loading, which remains a prerequisite for establishing absolute measurement accuracy.

Sixth, the study did not evaluate inter-instructor variability in the delivery of technical guidance, long-term skill retention among participants, or the reproducibility of the measurements across repeated sessions over time—factors that directly influence the practical implementation of the platform within continuous training curricula.

Finally, vascular neck restraint remains a high-risk physical intervention associated with severe and unpredictable injuries, as documented in medical literature. The findings presented herein strictly characterize the mechanical response of an instrumented simulator and must not be interpreted as a validation, endorsement, or operational guideline for applying such maneuvers to human subjects. Any future studies involving human participants would require independent ethical review, institutional oversight, and specialized protocols powered to assess physiological safety outcomes—none of which are proposed or implied by this work.

Future research should focus on: (i) cross-validating the sensor array against independent, traceable pressure-measurement devices; (ii) analyzing dynamic pressure–time profiles, contact duration, and cumulative mechanical loading; (iii) expanding testing to larger and more diverse simulator-based cohorts; (iv) evaluating inter-instructor consistency and long-term skill retention; and (v) improving sensor design, spatial resolution, and signal processing methods to enhance the platform's utility as a biomechanical research and feedback tool, strictly within the scope of surrogate-device testing.

VI. Conclusion

This study presented the development and evaluation of an instrumented monitoring system designed to quantify surface pressure distribution during the execution of vascular neck restraint (VNR) techniques. The proposed system successfully measured and mapped the mechanical forces applied to the right carotid, left carotid, and anterior cervical regions, providing high-resolution, objective data regarding force distribution during execution.

The experimental results demonstrated that structured technical instruction produced highly significant, systematic changes in pressure distribution patterns across the cohort. Following instruction, the pressure applied to both carotid regions increased markedly, whereas pressure over the tracheal region decreased substantially. These quantified shifts indicate that participants altered their hand-and-arm positioning on the simulator in a manner consistent with the biomechanical principles described in the literature as favoring carotid-directed over tracheal compression. This finding characterizes technique execution on an instrumented mannequin and does not constitute evidence of safety when the technique is applied to a human being.

The capacity of the embedded system to objectively distinguish the distinct profiles between pre-instruction (Trial A) and post-instruction (Trial B) executions on the simulator supports the proposed approach as a candidate tool for tactical training feedback and objective performance assessment, pending the cross-validation and human-subject ethical evaluation. By delivering multi-regional quantitative telemetry, this technology effectively complements traditional, subjective instructor-based evaluations, contributing to the standardization and mechanical characterization of simulator-based training.

Ultimately, these findings confirm the feasibility of utilizing dedicated embedded pressure-sensing platforms to monitor high-risk physical interventions. This methodology establishes a rigorous foundation for future investigations involving real-time hemodynamic cross-validation and evidence-based training protocols in law enforcement and combat sports.

Ethical Considerations

This study involved measurements performed exclusively on an instrumented anthropomorphic mannequin; no human subjects were exposed to the vascular neck restraint technique for data collection purposes. Participants executed the technique against the simulator only, under standard conditions of their regular military training program.

Acknowledgment

The authors express their gratitude for making the infrastructure of the sports practice facilities of the Paraná Military Police (PMPR) available for this research.

References

[1] Kodokan Judo Institute, "Kodokan definitions of Judo techniques," Oct. 1, 2022. [Online]. Available: Kodokan Definitions of Judo Techniques (PDF). Accessed: Jul. 13, 2026.

[2] Polícia Militar do Paraná, *Manual de Defesa Policial*, 3ª Seção do Estado-Maior, Câmara Técnica de Defesa Pessoal Policial, 2023.

[3] J. R. Mitchell, D. E. Roach, J. V. Tyberg, L. Belenkie, and R. S. Sheldon, "Mechanism of loss of consciousness during vascular neck restraint," *J. Appl. Physiol.*, vol. 112, no. 3, pp. 396–402, Feb. 2012, doi: 10.1152/japplphysiol.00592.2011.

[4] S. J. Stellpflug, W. H. Menton, M. F. Dummer, T. Menton, J. Corry, and R. LeFevere, "Time to unconsciousness from sportive chokes in fully resisting highly trained combatants," *Int. J. Perform. Anal. Sport*, vol. 20, no. 4, pp. 720–728, Aug. 2020, doi: 10.1080/24748668.2020.1780873.

[5] W. P. Bozeman *et al.*, "Safety of vascular neck restraint applied by law enforcement officers," *J. Forensic Legal Med.*, vol. 92, p. 102446, Nov. 2022, doi: 10.1016/j.jflm.2022.102446.

[6] M. J. Hickman, R. M. Scales, J. N. Strote, and J. L. Worrall, "Use of vascular neck restraints in law enforcement: A case-study of Spokane, WA," *Police Pract. Res.*, vol. 22, no. 6, pp. 1668–1678, Dec. 2021, doi: 10.1080/15614263.2021.1948849.

[7] Z. Demartini, Jr., M. Rodrigues Freire, R. O. Lages, A. N. Francisco, F. Nanni, L. A. Maranha Gatto, and G. L. Koppe, "Internal carotid artery dissection in Brazilian Jiu-Jitsu," *J. Cerebrovasc. Endovasc. Neurosurg.*, vol. 19, no. 2, pp. 111–116, Jun. 2017, doi: 10.7461/jcen.2017.19.2.111.

[8] K. L. Moore, A. F. Dalley, and A. M. R. Agur, *Anatomia Orientada para a Clínica*, 8ª ed. Rio de Janeiro, RJ, Brasil: Guanabara Koogan, 2019.

[9] S. Standring, Ed., *Gray's Anatomy: The Anatomical Basis of Clinical Practice*, 42nd ed. Edinburgh, U.K.: Elsevier, 2021.

[10] O. Remesz, "A labelled diagram of the carotid triangle of the neck," Wikimedia Commons, 2014. [Online]. Available: https://commons.wikimedia.org/wiki/File:Musculi_coli_base,_my_edits_for_tringles,_Carotid_T.svg. Accessed: Jun. 26, 2026.

[11] V. J. DiMaio and D. DiMaio, *Forensic Pathology*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2001.

[12] N. Karangeli, G. Triantafyllou, P. Papadopoulos-Manolarakis, N. A. Arkoudis, G. Velonakis, A. Samolis, and M. Piagkou, "Variations in the spatial relationship between the hyoid bone and the carotid arteries and their clinical significance," *World J. Radiol.*, vol. 17, no. 11, p. 113153, Nov. 2025, doi: 10.4329/wjr.v17.i11.113153.

[13] J. R. Thomas, J. K. Nelson, and S. J. Silverman, *Research Methods in Physical Activity*, 7th ed. Champaign, IL, USA: Human Kinetics, 2015.

[14] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects," *J. Clin. Epidemiol.*, vol. 67, no. 2, pp. 267–277, Feb. 2014, doi: 10.1016/j.jclinepi.2013.08.015.

A Multimodal Prototype for Portable Quantitative Monitoring of Epilepsy

Gabriela Hernández Rodríguez Montserrat
Departamento de Bioingeniería Traslacional
Universidad de Guadalajara
Guadalajara, México
gabriela.hernandez6405@alumnos.udg.mx

Karen Guadalupe Lisjuan Rivero
Departamento de Bioingeniería Traslacional
Universidad de Guadalajara
Guadalajara, México
karen.lisjuan1246@alumnos.udg.mx

Tabatha Montoya Estrada
Departamento de Bioingeniría Traslacional
Universidad de Guadalajara
Guadalajara, México
tabatha.montoya@alumnos.udg.mx

Norma Ramirez Hernandez
Departamento de Bioingeniría Traslacional
Universidad de Guadalajara
Guadalajara, México
https://orcid.org/0000-0002-2022-0444

***Abstract*— Epileptic seizures are inherently unpredictable events that pose significant challenges for continuous monitoring outside clinical settings. This study presents the development of a wearable, non-invasive, multimodal prototype designed for the acquisition of physiological and contextual variables potentially associated with epileptic events. The system integrates hemodynamic monitoring via functional near-infrared spectroscopy (fNIRS), digital pupillometry, inertial sensors, and psychometric evaluation of perceived stress. An ESP32-based acquisition unit synchronizes and wirelessly transmits measurements to a mobile application for visualization, storage, and processing. The prototype enables the extraction of quantitative features across heterogeneous physiological modalities, providing complementary insights for individualized and longitudinal physiological monitoring. While the system currently lacks clinical validation as a dedicated seizure detector or biomarker platform, it establishes a robust technological foundation for future clinical trials and the development of artificial intelligence models for risk assessment and alert generation in ambulatory environments.**

Keywords—Multimodal monitoring; Functional near-infrared spectroscopy (fNIRS); Digital pupillometry; Wearable technology.

I. INTRODUCTION

Epilepsy is one of the most prevalent chronic neurological disorders worldwide, affecting over 50 million people and imposing a substantial clinical, social, and economic burden on healthcare systems. In Latin America—and specifically in Mexico, where approximately two million individuals live with the condition—epilepsy remains associated with a high burden of morbidity, disability, and premature mortality. Among seizure classifications, focal-onset seizures account for a significant proportion of cases and are characterized by their unpredictable nature, which heightens the risk of physical injury, functional impairment, and loss of patient autonomy. Consequently, identifying physiological alterations accompanying the preictal state and the preictal-to-ictal transition has emerged as a vital research avenue, particularly for developing continuous monitoring strategies aimed at early seizure detection and characterizing candidate neurophysiological precursor signatures [1, 2, 3].

Electroencephalography (EEG) remains the clinical gold standard for evaluating cerebral electrical dynamics in epilepsy due to its high temporal resolution and capacity to capture neuroelectric alterations. Nevertheless, complex instrumentation requirements, multi-electrode configurations, and high susceptibility to motion and physiological artifacts severely restrict its utility for long-term, ambulatory monitoring outside clinical environments. These constraints have catalyzed the exploration of complementary, multi-sensor modalities (Figure 1) To address this gap, integrating functional near-infrared spectroscopy (fNIRS), digital pupillometry, psychometric stress assessment, and inertial measurement units (IMUs) yields synergistic, multi-system data—capturing cortical hemodynamic responses, pupillary autonomic dynamics, psychological stress states, and body kinematic/orientational shifts, respectively. The convergence of these heterogeneous modalities provides quantitative metrics to characterize individual physiological variations tied to seizure phenomena, establishing a foundation for candidate feature extraction and digital biomarker discovery. Ultimately, this approach underpins the design of wearable, user-centric systems tailored for continuous out-of-clinic monitoring and real-time alert generation during daily living activities [4].

Functional near-infrared spectroscopy (fNIRS) is a non-invasive, wearable, and cost-effective neuroimaging modality that enables continuous monitoring of cerebral hemodynamic responses by quantifying concentration changes in oxygenated (Δ[HbO]) and deoxygenated (Δ[HbR]) hemoglobin via neurovascular coupling [5]. However, standard long-separation channels (~30mm) are inherently confounded by systemic extracerebral contamination originating from scalp and skull microvasculature, as well as autonomic physiological fluctuations (e.g., cardiac pulsation and respiration). To isolate true cortical signals, short-separation fNIRS (ss-fNIRS) channels (≤8–10mm) capture superficial hemodynamic dynamics to serve as regressors within a General Linear Model (GLM) framework, effectively removing systemic interference and optimizing the signal-to-noise ratio of reconstructed cortical responses.

A. Pathophysiology

Epilepsy is a chronic neurological disorder characterized by recurrent seizures arising from excessive, hypersynchronous, and transient neuronal activity within cortical or subcortical

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

brain networks (Figure 2) [6]. These alterations can manifest as motor, sensory, cognitive, autonomic, or behavioral symptoms, depending on the specific neural circuits involved. Various precipitating factors, such as sleep deprivation and physical or emotional stress, can lower the seizure threshold and trigger events in susceptible individuals [7, 8].

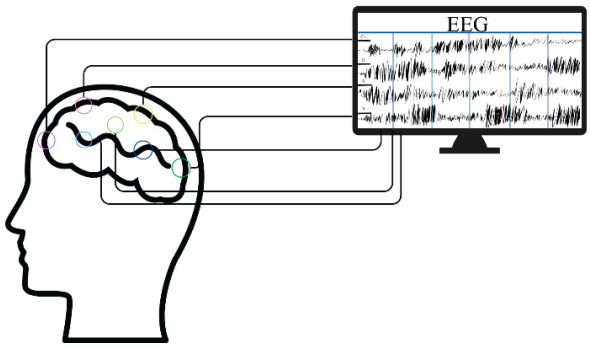


Figure 1 Schematic representation of epilepsy pathophysiology, spanning from excitation–inhibition (E/I) imbalance and neuronal hyperexcitability to pathological hypersynchrony, and the generation and network propagation of epileptiform discharges with their corresponding clinical manifestations. Concurrently, electroencephalographic (EEG) recordings remain inherently susceptible to diverse physiological and non-physiological artifacts that compromise signal fidelity

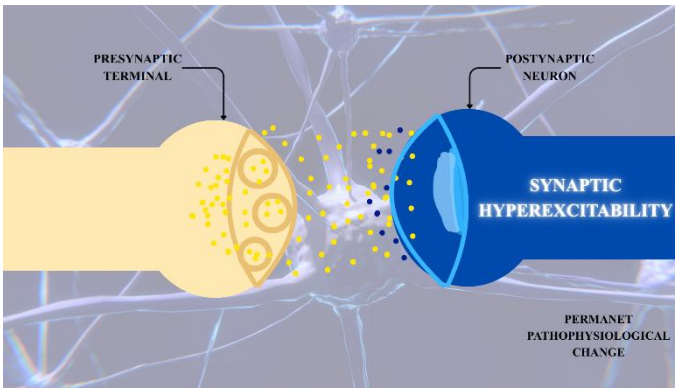


Figure 2 Schematic representation of epilepsy pathophysiology. Seizure onset involves complex network dynamics beyond simple depolarization-repolarization cycles; although rhythmic discharges occur, they exhibit intricate, non-linear spatiotemporal patterns across neural ensembles

At the cellular and network levels, epilepsy pathophysiology stems from an imbalance between excitatory and inhibitory mechanisms, facilitating the generation and propagation of epileptiform discharges through dynamic shifts in neuronal excitability and cerebral metabolism. Seizure recurrence fuels epileptogenic processes that induce structural, functional, and molecular alterations in neural circuits, progressively increasing their susceptibility to subsequent events [9, 10]. Elucidating these dynamic processes is essential for identifying physiological alterations that precede or accompany seizure onset, providing the theoretical rationale for multimodal architectures capable of acquiring and synchronizing neuroelectric, hemodynamic, autonomic, and kinematic data streams.

B. Classification

Contemporary epilepsy classification relies on the International League Against Epilepsy (ILAE) criteria, which establish a standardized framework for clinical characterization and diagnosis [11]. This framework operates across three levels: seizure type, epilepsy type, and epileptic syndrome. Seizures are categorized by onset into focal, generalized, unknown onset, or unclassified [12, 13], whereas epilepsy type and syndrome are determined by integrating clinical features, electroencephalographic (EEG) patterns, age of onset, clinical trajectory, and etiology. In this context, EEG remains crucial for characterizing the spatiotemporal distribution and propagation patterns of epileptiform activity (Figure 3) [14].

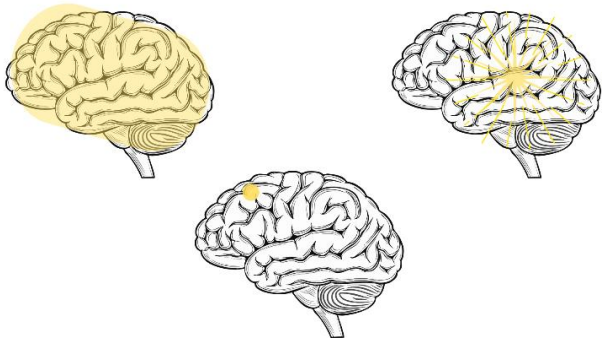


Figure 3 Classification of epileptic seizures based on the spatial distribution of epileptiform activity. Focal seizures originate within networks limited to one cerebral hemisphere, whereas generalized seizures rapidly engage bilaterally distributed networks at onset.

Concurrently, the advent of continuous monitoring technologies has driven multimodal approaches that synthesize neuroelectric, hemodynamic, autonomic, and behavioral indicators. The bidirectional relationship between epilepsy, stress, and neuropsychiatric comorbidities justifies incorporating psychometric scales as a complementary measure of patient status [15]. Furthermore, digital pupillometry objectively quantifies pupillary dynamics, offering valuable insights into autonomic nervous system function and central neurological states [16, 17]. These non-invasive measures complement conventional EEG, particularly within wearable, ambulatory monitoring frameworks.

In this setting, functional near-infrared spectroscopy (fNIRS) captures cortical hemodynamic fluctuations tied to neurovascular coupling by quantifying changes in oxygenated (Δ[HbO]) and deoxygenated (Δ[HbR]) hemoglobin concentrations [18, 19]. However, because fNIRS signals are susceptible to extracerebral hemodynamic contamination, short-separation fNIRS (ss-fNIRS) channels are integrated to estimate superficial interference and enhance cortical specificity. Integrating fNIRS, digital pupillometry, and inertial measurement units (IMUs) into low-power wearable platforms harmonizes hemodynamic, autonomic, and kinematic data streams with neuroelectric activity. This multimodal convergence provides rich, quantitative metrics to characterize individual seizure patterns and lays the groundwork for identifying digital biomarker candidates for early seizure detection and ambulatory alert generation [20].

II. Methodology

To overcome the clinical dependence and artifact vulnerability of ambulatory scalp EEG, a low-power wearable biomedical platform was engineered for continuous, non-invasive acquisition of physiological and contextual parameters associated with seizure dynamics. The system integrates dual-wavelength functional near-infrared spectroscopy (fNIRS), digital pupillometry, a 6-degree-of-freedom (6-DoF) inertial measurement unit (IMU), and an app-integrated psychometric stress assessment module. Central data acquisition, hardware clock timestamping, and Bluetooth Low Energy (BLE) wireless transmission are coordinated by an ESP32 microcontroller streaming to a mobile application for real-time visualization and feature extraction.

A. Cortical Hemodynamics (fNIRS)

Regional prefrontal hemodynamics were monitored bilaterally using dual-wavelength optical front-ends (660nm/ 880nm MAX30102) positioned with an inter-channel spacing of 17cm [21] (Figure 4). Optical attenuation was modeled via the Modified Beer–Lambert Law (MBLL) [22]. to quantify relative concentration changes in oxygenated (Δ[HbO]) and deoxygenated (Δ[HbR]) hemoglobin. An 8mm short-separation channel was incorporated to capture superficial scalp microvasculature dynamics for background noise regression.

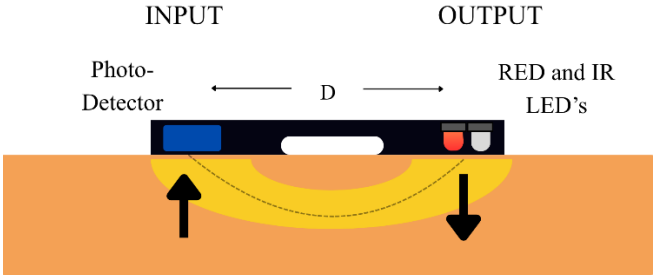


Figure 4 Operation of the fNIR system. It is used as a mapping method in neuroscience, where it measures oxygenation levels in cortical areas.

B. Autonomic Pupillary Dynamics

Pupillary dynamics were evaluated via visible-spectrum digital imagery. Pupil and iris boundaries were segmented using the Circular Hough Transform (CHT) to compute the scale-invariant Pupil-to-Iris Ratio (PIR) at an eye-to-camera distance of 15cm [23].

$$PIR = \frac{\emptyset pupila}{\emptyset iris} = \frac{rpupila}{riris}$$

where r_{pupila} and r_{iris} represent radius estimates in pixel space. Pupillary states were operationalized into miosis (PIR < 0.30), baseline (0.30 ≤ PIR ≤ 0.50) and mydriasis (PIR > 0.50).

C. Psychometric Stress Covariate

Perceived stress was quantified using an 11-item, 5-point Likert survey evaluating sleep deprivation, cognitive load, and environmental triggers [24]. A total score $Stotal = \sum_{i=1}^{11} Item_i$ was calculated, applying an operational cutoff of $Stotal \geq 30$ (>65 of the maximum score) to flag high-stress states as a contextual covariate.

D. Motor Activity (IMU)

Kinematic activity was captured using a 6-DoF IMU (3-axis accelerometer and gyroscope) interfaced via I²C [25]. Roll (∅) and Pitch (θ) orientation angles were derived relative to the gravity vector. Overall movement intensity and kinematic variability were extracted using acceleration magnitude ak, Signal Magnitude Area (SMA), and moving-window acceleration variance (σ^2):

$$Roll\ (\emptyset) = arctan\left(\frac{ay}{\sqrt{a_x^2 + a_z^2}}\right)\left(\frac{180}{\pi}\right)$$

$$Pitch\ (\theta) = arctan\left(\frac{-ax}{\sqrt{a_y^2 + a_z^2}}\right)\left(\frac{180}{\pi}\right)$$

Where ax, ay, and az represent the acceleration components along the three axes. Roll represents the roll angle about the X-axis, and Pitch about the Y-axis.

$$|A| = \sqrt[2]{ax^2 + ax^3 + ax^3}$$

$|A|$ overall intensity of the movement

$$SMA = \frac{1}{T}\left(\sum_{i=1}^{N}|A_x(i)| + \sum_{i=1}^{N}|A_y(i)| + \sum_{i=1}^{N}|A_z(i)|\right)$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(|A|_i - \mu)^2$$

where μ is the windowed mean acceleration magnitude.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}|A|_i$$

E. Hardware Control & Data Synchronization

An ESP32 Mini microcontroller unit (MCU) managed I²C sensor communication, assigned independent modality-specific sampling frequencies, and appended microsecond hardware timestamps to maintain inter-channel temporal alignment across all streamed data packets [26]. This architecture enabled the integration of heterogeneous sensors into a portable, low-power platform, while maintaining a modular structure for the future incorporation of new sensors or processing algorithms.

III. Results

The engineered prototype achieved complete functional integration of four complementary sensing modalities for synchronized acquisition of physiological and contextual parameters linked to seizure phenomena: dual-wavelength optical hemodynamics (fNIRS), digital pupillometry, 6-degree-of-freedom (6-DoF) inertial sensing, and psychometric stress profiling. The wearable hardware architecture was implemented as an ergonomic headband that distributed optical

and kinematic sensors across frontal and posterior cranial regions while housing an ESP32 Mini microcontroller unit (MCU), a power management circuit, and an inertial measurement unit (IMU). This configuration enabled continuous multi-channel signal acquisition, microsecond-level temporal synchronization, and wireless telemetric streaming to a custom mobile application (Figures 5 and 6).

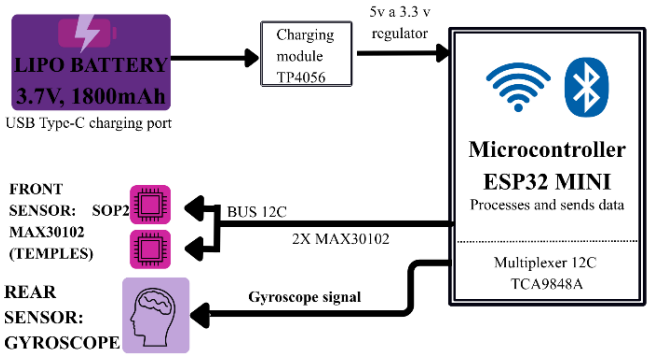


Figure 5 Functional block diagram of the hardware subsystem, detailing power regulation stages, central MCU control logic, and sensor-driver bus interfaces.

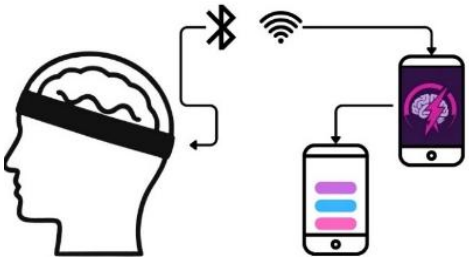


Figure 6 End-to-end architecture of the multimodal platform, illustrating data acquisition via the wearable headband, Bluetooth Low Energy (BLE) wireless transmission, and local/cloud storage, real-time visualization, and processing pipelines within the mobile

The custom mobile application served as the central system node, orchestrating image acquisition, psychometric survey execution, telemetry rendering, and secure data logging (Figures 7–9). The integrated psychometric module administered an 11-item, 5-point Likert scale instrument to assess perceived stress, aggregating a cumulative score Stotal as a real-time contextual covariate alongside streaming physiological signals.

The digital pupillometry pipeline captured visible-spectrum ocular images and executed an automated segmentation workflow based on the Circular Hough Transform (CHT) to extract pupil and iris geometric parameters. The Pupil-to-Iris Ratio was derived as a normalized, scale-invariant index of pupillary aperture dynamics: $PIR = \frac{rpupila}{riris}$. The platform operationalized PIR values into discrete pupillary states, facilitating inter-subject benchmarking and longitudinal tracking of temporal autonomic variability (Figure 10).

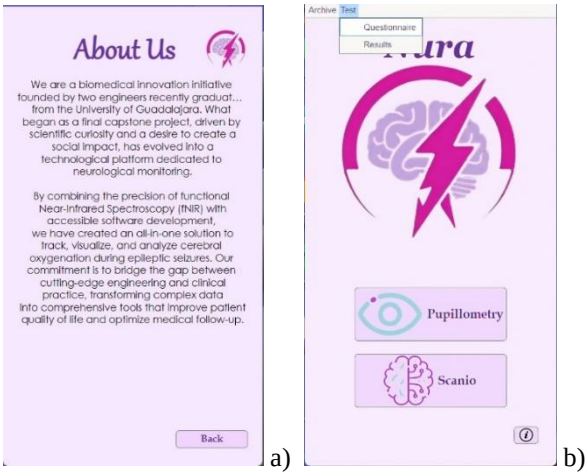


Figure 7 Mobile application user interface: (a) project landing and user authentication screen; (b) main navigation dashboard providing access to live acquisition and query modules.

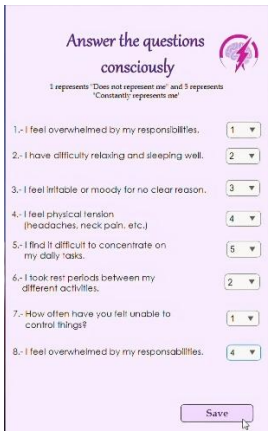


Figure 8 Graphical user interface of the StressTest_Module for psychometric stress profiling via an 11-item 5-point Likert instrument.



Figure 9 Interactive alert-verification interface for real-time event validation. User acknowledgement dismisses false alarms and resumes continuous monitoring, whereas an unacknowledged timeout automatically dispatches emergency notifications to pre-designated

Concurrently, the 6-DoF IMU quantified motor dynamics by capturing raw tri-axial acceleration and angular velocity signals. Derived kinematic features—including spatial orientation angles (Roll (∅) and Pitch (θ)), total acceleration magnitude (ak), Signal Magnitude Area (SMA), and windowed acceleration variance (σ^2) —characterized motor manifestations. When kinematic features exceeded operational threshold criteria, the mobile application initiated an interactive alert dialog with a user-cancellation window to mitigate false positive rates. In the absence of user cancellation during the

allotted interval, the system dispatched automated emergency alerts to pre-assigned contacts.

Overall, these results demonstrate the successful functional integration and operational feasibility of the multimodal wearable platform for synchronized, multi-parametric data acquisition, processing, and visualization. This architecture establishes a scalable technical foundation for individualized, longitudinal physiological profiling, laying the groundwork for future clinical studies to evaluate its diagnostic sensitivity against gold-standard clinical references (e.g., video-EEG).

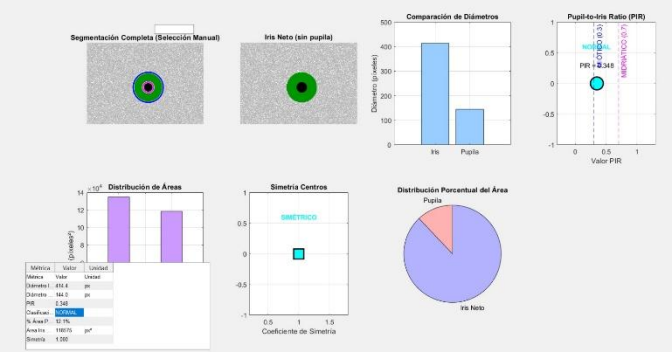


Figure 10 Ocular image processing and pupil-iris segmentation interface: pupil and iris boundaries are segmented via CHT to extract geometric metrics (diameter and surface area), compute the PIR index and centroid offset symmetry, and render static and longitudinal trend visual feedback.

IV. Conclusion

Epileptic seizures remain a major diagnostic and therapeutic challenge due to their paroxysmal nature and the operational constraints of conventional long-term video-electroencephalography (vEEG) in outpatient settings. To address these limitations, this study engineered a wearable, non-invasive, low-power multimodal platform designed for the time-synchronized acquisition of neuro-hemodynamic, autonomic, kinematic, and contextual parameters associated with seizure dynamics. By harmonizing dual-wavelength functional near-infrared spectroscopy (fNIRS), digital pupillometry (PIR), 6-DoF inertial kinematic tracking, and app-integrated psychometric stress assessment, the platform establishes a quantitative, multi-parametric framework for continuous ambulatory physiological profiling and individualized baseline characterization.

In its current stage of development, the system functions as a validated experimental framework for multi-channel signal acquisition, hardware clock synchronization, telemetry streaming, and real-time visualization. It does not yet claim standalone diagnostic autonomy or clinical seizure classification. Future work will focus on validating the extracted feature sets across clinical patient cohorts under standardized observational protocols, benchmarked against continuous long-term vEEG recordings. This translational roadmap will enable the quantitative mapping of multimodal surrogate markers to electrographic seizure onset, facilitate rigorous sensitivity and specificity benchmarking, and ultimately support the integration of edge-compatible machine learning algorithms for predictive seizure risk forecasting and automated early-warning telemetry in continuous ambulatory care.

Referencias

[1] S. d. Salud, «En México 2 millones de personas padecen epilepsia,» Gobierno de México, 2015 Enero 31. [En línea]. Available: https://www.gob.mx/salud/prensa/en-mexico-2-millones-de-personas-padecen-epilepsia. [Último acceso: 01 03 2026].

[2] OPS, «Epilepsy,» Organización Mundial de la Salud, 07 Febrero 2024. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/epilepsy.

[3] WHO, «Epilepsia,» Organización Panamericana de la Salud, [En línea]. Available: https://www.paho.org/es/temas/epilepsia.

[4] A. Viloria Alebesque y et al., «Utilidad del electroencefalograma en el manejo de la epilepsia en el Servicio de Urgencias,» *Neurología*, vol. 35, nº 4, pp. 238-244, 2020.

[5] B. y. Cortivision, «Fusión de fNIRS y EEG: un paso más en la investigación de la actividad cerebral,» Bitbrain, 1 Agosto 2024. [En línea]. Available: https://www.bitbrain.com/es/blog/fusion-fnirs-eeg-brain-activity-research. [Último acceso: 16 Septiembre 2025].

[6] J. Pastor Gómez y et al., «Redes neuronales en epilepsia,» de *Epilepsia. Un punto de vista iberoamericano*, Estado de México, Alfil, 2024.

[7] Ápice, «¿Qué factores pueden desencadenar una crisis?,» APICE, [En línea]. Available: https://www.apiceepilepsia.org/que-es-la-epilepsia/que-factores-pueden-desencadenar-una-crisis/. [Último acceso: 14 Octubre 2025].

[8] L. Reyes Aguirre y et al., «Actualización de la fisiopatología y clasificación de la epilepsia,» XXXIV Jornada Científica Estudiantil, Santiago de Cuba, 2023.

[9] J. Cáceres y L. Hernández, Epilepsia: Una visión desde la complejidad, Ciudad de México: Parmenia, 2023.

[10] Y. Romero y J. Pujol , «El tratamiento y seguimiento de la epilepsia,» *FMC-Formación Médica Continuada de Atención Primaria*, vol. 30, nº 3, pp. 139-143, 2023.

[11] S. Beniczky y et al., «Clasificación actualizada de las crisis epilépticas: documento de posición de la Liga Internacional contra la Epilepsia.,» *Epilepsia*, vol. 66, nº 6, pp. 1804-1823, 2025.

[12] L. Garcez y et al., Epilepsia: diagnóstico e tratamento, São Paulo: Pasteur, 2023.

[13] E. Wirrell y et al., «Introducción a los articulos sobre el síndrome epiléptico,» *WILLEY*, vol. 63, nº 6, pp. 1330-1332, 2022.

[14] F. Castro Villablanca, «Actualización en la clasificación de las epilepsias y síndromes epilépticos,» *Revista Médica Clínica Las Condes*, vol. 37, nº 1, pp. 8-16, 2026.

[15] B. Zaldívar y et al., «Comportamiento de la ansiedad, depresión y disociación en pacientes con crisis epilépticas y crisis psicógenas no epilépticas,» *Rev Cubana Neurol Neurocr*, vol. 10, nº 2, pp. 1-19, 2020.

[16] L. Ortega y et al., «Pupillometry and EEG fusion for preictal detection,» *Sensors*, vol. 21, nº 18, 2021.

[17] C. Ortega y et al., «Pacientes neuricríticos pediátricos: Prevalencia de patologías y su monitorización,» *Revista chilena de anestesia*, vol. 53, nº 2, pp. 160-164, 2024.

[18] A. Y. Denault y Et al, «CHAPTER7 - Functional near Infrared Spectroscopy,» de *NeuromonitoringTechniques*, Canada, Elsevier, 2021, pp. 180-211.

[19] P. R. Bazán y et al., «fMRI and fNIRS Methods for Social Brain Studies: Hyperscanning Possibilities,» National Library of Medicine, 29 Noviembre 2022. [En línea]. Available: https://www.ncbi.nlm.nih.gov/books/NBK595449/. [Último acceso: 29 Agosto 2025].

[20] K. Khaksari y et al., «Aplicaciones de la espectroscopia de infrarrojo cercano en la epilepsia, con énfasis en los transtornos mitocontriales.,» *Neurotherapeutics*, vol. 21, nº 1, 2024.

[21] N. Castellanos Mejia y e. al., «uniandes.edu.co,» Universidad de los Andes, 2023. [En línea]. Available: https://hdl.handle.net/1992/67930. [Último acceso: 2026 AGO. 17].

[22] D. Buyanov y e. al., «An algorithm for measuring absolute and relative hemoglobin concentrations using near infrared spectroscopy,» *Biomedical Engineering*, vol. 56, nº 3, pp. 176 - 179, 2022.

[23] Nyholm y e. al., «Specific thresholds of quantitative pupillometry parameters predict unfavorable outcome in comatose survivors early after cardiac arrest,» *ResuscitationPlus*, vol. doi: 10.1016/j.resplu.2023.100399, nº 1, p. 14, 2023.

[24] D. Jhaveri y e. al., «Stress and epilepsy: towards understanding of neurobiological mechanisms for better management,» *eneuro*, vol. 10, nº 11, pp. 0200-23, 2023.

[25] S. Burgos y e. al., «Proyecto de grado, Ingeniería Informática, Instituto Tecnológico de Buenos Aires (ITBA),» 2025. [En línea]. Available: https://ri.itba.edu.ar/handle/20.500.14769/4888. [Último acceso: 30 Julio 2026].

[26] R. Ronald y e. al., «Desarrollo de un dispositivo IoT para monitoreo médico continuo mediante adquisición de señales bioeléctricas,» *UNANCHAY*, vol. 3, nº 2, p. 99, 2024.

A Polyacrylamide-GOx Hydrogel Biosensor on Laser-Engraved Graphene: Structural and Electrochemical Evidence for Glucose Detection

Clinston Paulino de Almeida
Electrical Engineering Department
Santa Catarina State University
Joinville, Brazil
clinstoncpa@gmail.com

Marcos Antonio Rebello
Electrical Engineering Department
Santa Catarina State University
Joinville, Brazil
marcos.rebello@edu.udesc.br

Elkin Roman-Gomez
Electrical Engineering Department
Santa Catarina State University
Joinville, Brazil
elkin.rg11@udesc.br

Gleidson Pereira de Oliveira e Silva
Institute of Chemistry and Biotechnology
Maceió, Brazil
gleidisonpos@gmail.com

Jeniffer Mclaine Duarte de Freitas
Institute of Chemistry and Biotechnology
Federal University of Alagoas
Maceió, Brazil
mclaine.jeny@gmail.com

Pedro Bertemes-Filho
Electrical Engineering Department
Santa Catarina State University
Joinville, Brazil
pedro.bertemes@udesc.br

***Abstract*—Self-monitoring of blood glucose in patients diagnosed with diabetes mellitus is essential for the proper management of the condition. This need has driven the search for electrochemical biosensors that are stable, cost-effective, minimally invasive, and efficient, and that overcome the limitations of conventional lancets. In this work, an integrated platform was developed using a three-electrode system fabricated by laser engraving on a flexible polyimide substrate, producing a conductive three-dimensional graphene network. The working electrode was modified by physical deposition of gold particles, and glucose oxidase was physically entrapped in a polyacrylamide-based hydrogel matrix with a controlled thickness of 2.0 mm. The device and its components were characterized by scanning electron microscopy and energy-dispersive X-ray spectroscopy, which confirmed the porous and rough morphology of the laser-induced graphene and the homogeneous distribution of the gold particles. Fourier-transform infrared spectroscopy demonstrated the structural stability of the hydrogel, evidenced by the preservation of the primary amide band at 1640 cm^{-1}, and revealed catalytic interactions in the fingerprint region (1150–1000 cm^{-1}) after the polymer matrix was subjected to different conditions, including diffusion analysis in glucose solution. Electrochemically, the biosensor exhibited a pseudocapacitive profile, and square-wave voltammetry showed a robust analytical response at -0.55 V, attributed to oxygen consumption during the glucose oxidase-catalyzed oxidation of glucose. These results indicate that the hydrogel synthesis process was efficient, preserved the chemical integrity of the reagents involved, and protected the sensing interface against interferents. It can be concluded that th platform represents a promising technology for minimally invasive glucose monitoring devices.**

***Index Terms*—Electrochemical biosensor, Laser-induced graphene, Glucose oxidase, Polyacrylamide hydrogel**

I. Introduction

Considered a fundamental nutrient in biological systems, glucose is responsible for vital metabolic processes, such as cellular respiration and biosynthesis [1], [2]. However, defects in its cellular uptake lead to its abnormal accumulation in the bloodstream—a condition associated with diabetes, a chronic metabolic disorder with a high global mortality rate that is expected to have a major impact in the coming decades [3]–[6]. Efficient management of this condition requires continuous self-monitoring of interstitial fluid glucose, creating market demand for sensors that offer high performance, cost-effectiveness, and ease of use [7]. Conventional methods for measuring blood glucose levels involve directly collecting blood from the patient using lancets; the discomfort caused by daily skin pricks hinders adherence to self-monitoring routines, leading to inadequate treatment control and, consequently, therapeutic errors, while also creating an entry point for infectious microorganisms. This creates a need to develop devices that improve the patient's quality of life, can be used safely, and provide reliable data [8], [9]. Historically, the concept of the biosensor emerged in 1962 with Clark and Lyons, who used an oxygen electrode to detect the oxidation of glucose by glucose oxidase (GOx). Since then, the system based on disposable reagent strips (commercially established in 1971) has remained conceptually the same: the sample rehydrates the dehydrated enzymes present on the strip, generating a product that is quantified by the detector [10], [11]. In GOx-based electronic systems, the enzyme converts glucose into glucono-

lactone and hydrogen peroxide (H2O2), which reacts with the electrode to generate a current flow [10], [12]. To overcome this, the market is seeking less invasive technologies (optical, transdermal, and thermal), although these face challenges such as sensitivity to perspiration [12] or dependence on skin tone [13]. Impedance-based systems [14] or reverse iontophoresis, also faced limitations due to low accuracy and skin irritation. In this context, electrochemical biosensors stand out as a viable alternative due to their high sensitivity, ease of operation, and excellent characterization of the surface-transducer interface [15]–[17]. However, severe limitations still need to be overcome. Non-enzymatic biosensors suffer from interference due to the adsorption of biomolecules onto their nanostructured surfaces [15], [18]–[21], while enzymatic ones lose activity due to fluctuations in temperature and humidity [19], [20]. Proposals in the literature to improve selectivity—such as increasing catalytic activity, using anti-interference coatings, and employing electron mediators—increase costs and entail high complexity. To overcome these challenges, surface engineering advocates the use of noble metals (Pt, Ag, Au) or micro- and nanostructured metal oxides to directly catalyze glucose [22]. In this context, the use of polyacrylamide gel has emerged as an effective approach. The incorporation of this three-dimensional polymer network enables the homogeneous physicochemical entrapment of biomolecules, establishing a protective hydrophilic microenvironment that preserves the enzymes structural integrity and activity against temperature and pH fluctuations, as well as dehydration. Furthermore, the gel's tunable porosity acts as a steric exclusion barrier against interfering macromolecules, facilitating the selective diffusion of glucose and minimizing leaching, thereby extending the device's lifespan and ensuring signal reproducibility. In parallel, another promising alternative involves the fabrication of electrodes via laser engraving on flexible polyimide substrates [23]. This process locally converts the material into a three-dimensional network of laser-induced graphene (LIG), acting as a highly porous and conductive transducer [24]–[28]. To mitigate interference caused by other biomolecules in complex media, modifying the reference electrode by coating it with gold (Au) particles offers an effective solution; this approach facilitates interfacial charge transfer and ensures the stability required for measurements. In this context, this study develops an electrochemical biosensor—compatible with minimally invasive devices—for quantifying glucose in interstitial fluids. The system integrates a graphene electrode patterned on polyimide, a gold-coated reference electrode, and an enzyme encapsulated within a polyacrylamide network, enabling detection via cyclic voltammetry. This integrated architecture preserves enzymatic activity and protects the interface against electronic interferents, establishing itself as a promising, stable platform for future digital monitoring systems [29]–[31].

II. Materials and Methods

A. Materials

Glucose oxidase (GOx). Glucose. Acrylamide (AAm). N,N-methylenebisacrylamide (BIS). N,N,N,N'-tetramethylethylenediamine (TEMED). and Ammonium persulphate (APS), were purchased from Sigma-Aldrich (Beijing, China). Phosphate-Buffered Saline - PBS (pH 7.4). Hydrogen peroxide PA (Modern chemistry). Glucose – Gluc UP (Newprov). All aqueous solutions were prepared with ultrapure water (Millipore Milli-Q).

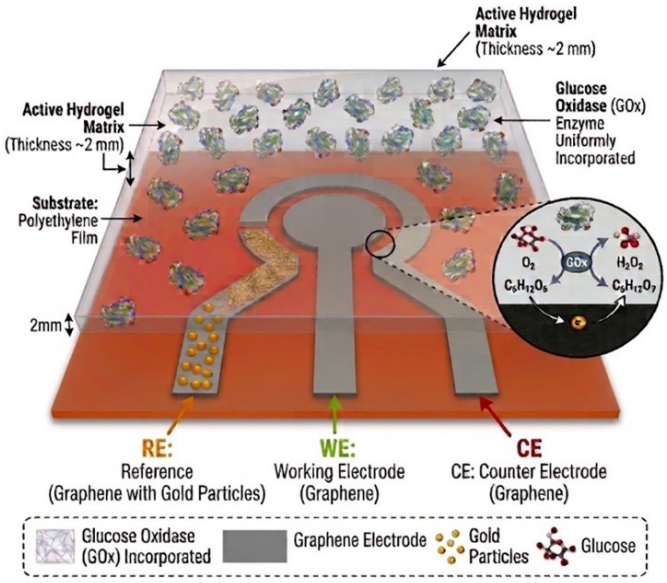


Fig. 1. Schematic diagram of a hydrogel-based biosensor deposited on a polyamide substrate.

B. Methods

1) Preparation of a graphene oxide electrode with a surface functionalized with gold particles: The polyamide film (substrate), with a thickness of 120 µm, was cut into 2.5 x 3.0 cm pieces and processed using a Creality CR-Laser Falcon to engrave the electrodes via surface burning. Operating parameters included a wavelength of 455 ± 5 nm and a power of 10 W, resulting in carbonization and the formation of a graphene surface in a tripolar electrode configuration, as illustrated in Figure 1. The surface of the reference electrode was subjected to a sputtering deposition process for 5 minutes using Quorum Q150RS equipment, producing particles ranging in size from 8 to 12 nm.

C. Synthesis of GOx enzyme-functionalized acrylamide gel

The polyacrylamide-based polymeric hydrogel, used as a platform for the GOx enzyme, was synthesized via the polymerization of acrylamide. A precursor solution was prepared by mixing 1,800 µL of ultrapure water (Milli-Q) with 1.600 µL of a 30% (w/v) acrylamide/bis-acrylamide stock solution. Subsequently, 3 mg of the enzyme glucose oxidase (GOx) was uniformly incorporated into the monomer mixture to reach a final enzyme concentration of 0.0037 mM. To start the polymerization process, 400 µL of freshly prepared 2% w/v ammonium persulfate and 250 µL of N,N,N',N'-tetramethylethylenediamine (TEMED) solution were added

sequentially under gentle stirring to avoid oxygen inhibition. The mixture was immediately poured onto a glass plate with a custom spacer to ensure a uniform hydrogel thickness of 2.0 mm. Cross-linking took place at room temperature (25 ± 2°C) for 30 minutes, resulting in an electrochemically active hydrogel matrix, which was cut to the same dimensions as the graphene electrode surface [23], [32].

D. Characterization

1) Scanning Electron Microscopy (SEM): The surface morphology of the graphene electrode was characterized by Scanning Electron Microscopy (SEM). Prior to analysis, the sample was sputter-coated with a thin layer of gold (Au) using a Quorum Q150RS sputter coater to ensure electrical conductivity at the interface. The micrographs were obtained using a Vega3 electron microscope from Tescan (Brno, Czech Republic), operating at an accelerating voltage in the 5–15 kV range. The images were recorded at magnifications ranging from 81x to 3000x to evaluate the porosity and integrity of the three-dimensional graphene network.

2) Energy Dispersive Spectroscopy (EDS): The elemental chemical composition of the graphene electrode surface was determined by Energy-Dispersive Spectroscopy (EDS). The spectra and elemental mapping were obtained using an EDS detector coupled to the electron microscope Vega3 electron microscope from Tescan (Brno, Czech Republic). The analysis enabled the identification and mapping of the distribution of the elements Carbon (C), Oxygen (O), and Gold (Au) constituting the electrode layer interface.

3) Fourier Transform Infrared (FTIR): To identify the functional groups and the polymerization process, and to examine the chemical behavior of the hydrogels under electrochemical techniques, we used Fourier Transform Infrared (FTIR) spectroscopy. The spectra, along with their respective transmittances and wavelengths, were obtained using a Shimadzu spectrometer (Model IR Prestige-21) operating in Attenuated Total Reflectance (ATR) mode. The readings were conducted in the spectral range of 4000 to 400 cm^{-1} with 64 scans.

4) Electrochemistry: The electrochemical characterization was performed using an AUTOLAB PGSTAT302N potentiostat with a three-electrode configuration patterned onto the polyamide surface, consisting of a counter electrode and a working electrode with graphene surfaces, and a reference electrode with a surface modified by gold (Au) particles. The following hydrogels were deposited on the electrode surface: WG (white gel), GE (gel/GOx enzyme 0.0037 mM), GEG (gel/enzyme/555 mM glucose solution), and GEH2O2 (gel/hydrogen peroxide – 52.2 mM and 104.4 mM). To ensure analytical consistency, we used the same electrode system for all analysis cycles at room temperature. During the analyses, the protocol consisted of the consecutive application of two voltammetric techniques - cyclic voltammetry and square-wave voltammetry - while varying the type of gel. For both

analyses, we used a potential sweep between -1.0 and 1.0 V at a rate of 0.1 V/s over three cycles. The following were evaluated: WG (White Gel), GE (Gel/Enzyme – subjected or not to a glucose solution), and GH2O2 (Gel/hydrogen peroxide). All measurements were performed in a phosphate buffer medium and All evaluated samples were subjected to the same analysis time, standardized at 3 minutes of hydrogel/electrode contact. The curves shown represent the average of three measurements for each situation evaluated.

III. Results and Discussions

A. Scanning Electron Microscopy (SEM)

The microstructure of the graphene electrode, obtained by burning polyamide, was evaluated by SEM at different magnifications and scales (Figure 2). The laser-induced heating process promoted direct carbonization and expansion of the polymer, resulting in a highly porous matrix composed basically of three-dimensional carbon atoms [33].

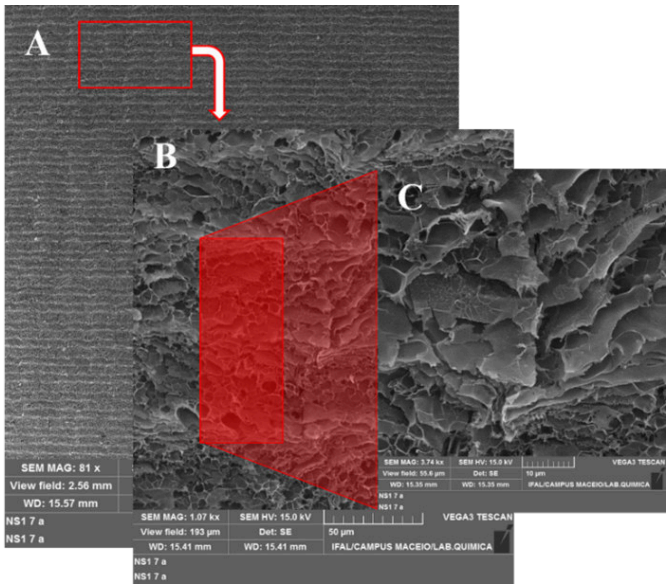


Fig. 2. SEM images of the graphene electrode surface obtained by carbonization of polyamide.

Upon magnifying the image 81x with a 500 µm scale (Figure 2A), an ordered arrangement of parallel, symmetrical lines, resulting from the laser beam scanning is observed. At a magnification of 3700x and a scale of 10 µm (Figure 2B), the transition to a spongy network with cavities becomes evident, likely generated by the rapid release of volatile gases during the formation of the graphene tracks, which consequently significantly increases the active surface area of the electrode [34]. The micrograph reveals an appearance of corrugated, interconnected graphene sheets forming an open-channel structure, creating a porous look, which is ideal for application in biosensor systems. The roughness will facilitate the process of stable mechanical anchoring for the polyacrylamide hydrogel, in addition to easing the diffusive transport of glucose and hydrogen peroxide, the product of the reaction catalyzed by the enzyme glucose oxidase (GOx), to the conductive surface, thereby optimizing electrical charge transfer [35].

1) Energy Dispersive Spectroscopy (EDS): The elemental chemical composition and the spatial distri-bution of the constituents present on the electrode surface were confirmed by EDS (Figure 3). The X-ray spectrum confirms the majority presence of the carbon atom (C), with a peak at approximately 0.27 keV (figure 3B), originating from the laser-induced graphene matrix [35]. We can also identify a low percentage of the element oxygen (Figures 3A1 and 3A2). The technique also reveals the successful incorporation of gold (Au), confirmed by emission lines at 2.12 keV (M), characteristic of the presence of trace amounts of the deposited element, and also shows secondary peaks around 9.71 and 11.44 keV in the intensity-versus-energy spectrum (cps/eV) [36].

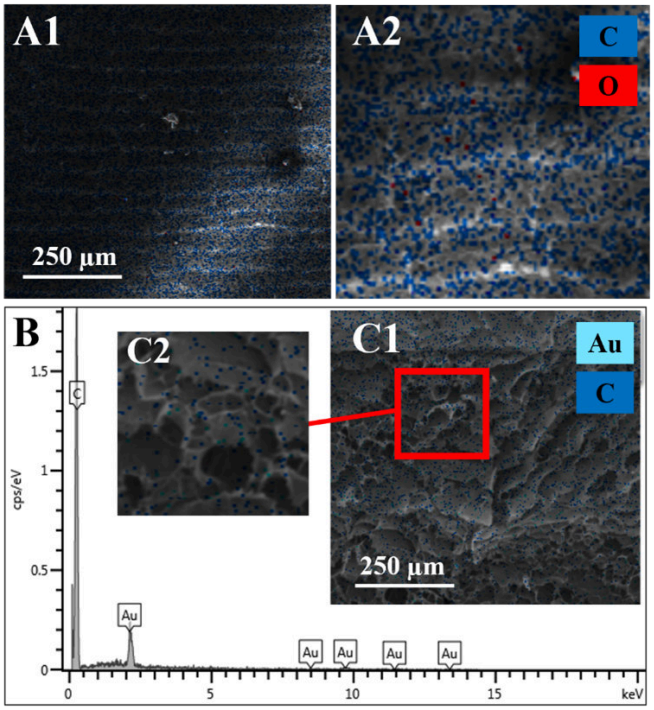


Fig. 3. EDS elemental mapping analysis of the electrode surface. (A1-A2) Non-functionalized surface with gold particles. (B) EDS spectrum showing the peaks of the chemical elements. (C1-C2) Surface functionalized with gold (Au) particles on the carbon (C) matrix

The topographic view of the Au-covered surface shows a homogeneous distribution of particles, with no evidence of severe agglomeration. This surface configuration enables the expansion of the electrode's surface area and optimizes charge transport and reaction kinetics at the electrode/modified-gel interface [37].

B. Fourier Transform Infrared (FTIR)

Overall, the spectra obtained in the infrared region predominantly reveal characteristics of the polyacrylamide polymer matrix (Figure 4). The broad band at 3350 cm¹, observed in all spectra (Figure 4A), can be attributed to the axial stretching vibrations of hydroxyl (O–H) bonds in water molecules, given that the material under evaluation is highly hydrated, but it can also be correlated with the presence of amide (N–H) groups within the hydrogel structure [38]

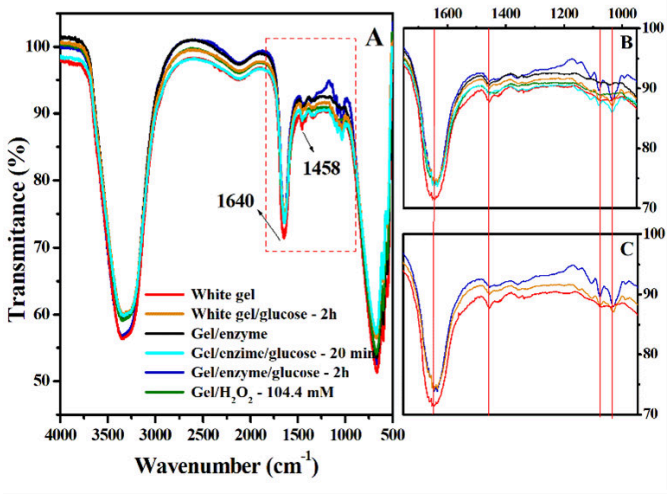


Fig. 4. EFTIR characterization of the biocomposite hydrogel before and after interaction with glucose and H2O2.

In the region between 3000 and 1800 cm¹, peaks of lower intensity are observed, which are attributed to the symmetric and asymmetric stretching of aliphatic groups, such as CH and CH [39]. At a wavenumber of approximately 1640 cm¹, a sharp, intense band stands out this is generally associated with the stretching of the carbonyl (C=O), bond characteristic of the primary amide band present in the hydrogel's polymer chains and may overlap with the angular deformation of interfacial water. The maintenance of these main bands across all curves confirms the structural stability of the biosensor's three-dimensional matrix under different exposure conditions [38], [40], [41]. To assess the permeability of the hydrogel, we evaluated different exposure conditions using a glucose solution (Figures 4B and 4C). Upon obtaining the spectra, we identified a modification in the spectral profile within the so-called "fingerprint" region, likely resulting from molecular interactions involving the enzyme glucose oxidase (GOx) and its catalytic activity in the presence of glucose. We observed no significant changes when comparing the spectra of the white hydrogel not exposed to the glucose solution with those of the same hydrogel after two hours in the glucose solution (Figure 4B, red and orange spectra, respectively); the same applies to the spectrum of the gel containing hydrogen peroxide. The persistence of the band at 1640 cm¹ (primary amide) and the shoulder at 1450 cm¹ (secondary amide)—attributed to C–N stretching combined with N–H bending suggest that in the absence of the enzyme glucose does not chemically alter the primary structure of the gel, instead only physical phenomena involving molecular diffusion within the porous channels may be occurring. Subtle yet significant modifications occur when the enzyme is anchored and active (gel/enzyme/glucose – after 2 hours in solution). A variation in band intensity is observed, as well as a slight shift in wavelength for the bands in the 1150–1000 cm¹ region. This behavior in the infrared region is

characterized by C–O and C–O–C stretching vibrations, which are intrinsic to the chemical structure of glucose, specifically the pyranose ring [42], [43].

C. Cyclic voltammetry

The obtained cyclic voltammograms exhibit an energy storage profile characteristic of a pseudocapacitive system. This behavior is commonly observed in systems comprising hydrogel matrices coupled with carbon-based electrodes operating over a wide potential window of -1.0 V to +1.0 V. The absence of well-defined peaks—also known as Faradaic peaks—suggests that charge transfer processes are occurring in conjunction with the electrical double-layer capacitance at the electrode/hydrogel interface.

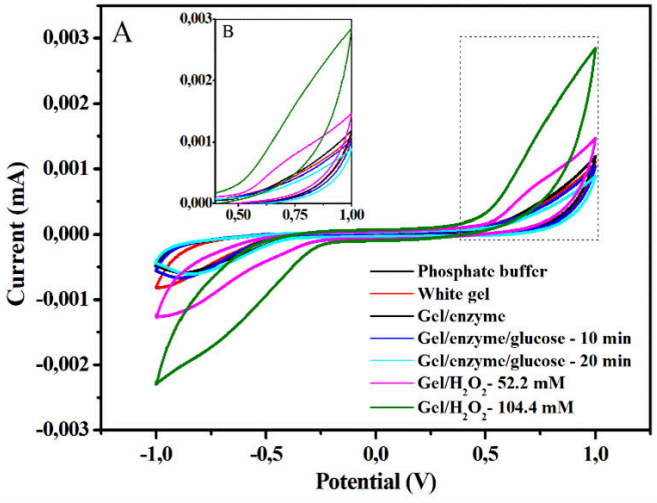


Fig. 5. Cyclic voltammograms of the hydrogels obtained under different experimental conditions. (A) Full voltammetric profile in the potential range of -1.0 V to 1.0 V. (B) Inset showing an enlargement of the anodic region of interest (0.4 V to 1.0 V).

This behavior is observed primarily in the voltammogram of the gel containing only the pure polymer matrix (Whaite gel). Upon the addition of the GOx enzyme to the polyacrylamide matrix, yielding the gel designated as GE, the capacitive profile is maintained, resulting only in a subtle modulation of the background current to -0.9 mA; this suggests that the process of entrapping the macromolecule within the polymer matrix does not block the substrate's electronic percolation pathways [44]. Despite the dominance of the pseudocapac-itive profile, localized deviations in curve symmetry reveal discrete redox events at the electrode interface. In the cathodic region, specifically within the -0.5 V to -1.0 V range, a cooperative reduction process manifests, shifting the limiting current at -1.0 V from -0.51 mA (for the WG) to -0.59 mA and -0.44 mA after 10 and 20 minutes of glucose exposure, respectively. This behavior may be directly linked to these processes. During the reverse scan (anodic phase), the current increase becomes significant above +0.5 V, reaching +1.0 mA at the +1.0 V limit for the glucose-exposed samples. The analytical capability of the electrode is confirmed when evaluating the gel containing the reaction end-product, hydrogen peroxide (H2O2). In the presence of 52.2 mM H2O2, a prominent reduction shoulder begins at -0.35 V and reaches -1.25 mA at -1.0 V, while the oxidation current at +1.0 V is amplified to +1.45 mA. Upon doubling the H2O2 concentration to 104.4 mM, the biosensor exhibits a catalytic response—likely controlled by the diffusion process—in which the cathodic current reaches -2.30 mA at -1.0 V and the anodic current shows a sharper increase starting at +0.5 V, peaking at +2.85 mA at the positive polarization limit.

D. Square Wave Voltammetry (SWV) Analysis

The use of Square Wave Voltammetry (Figure 6) enabled the individual assessment of reduction and oxidation events arising from the interaction between the graphene oxide electrode and the various modifications to the hydrogel's polymer matrix; prominent electrochemical peaks were observed in the cathodic region, highlighting charge transfer processes at the biosensor interface

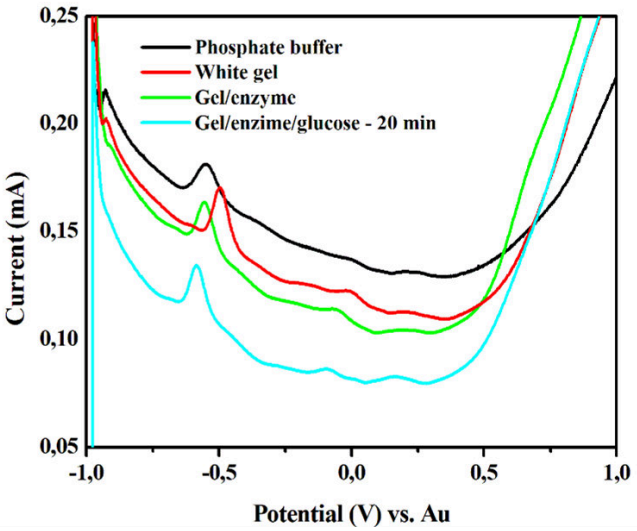


Fig. 6. Square-wave voltammetry (SWV) profiles evaluating the electrochemical behavior of the hydrogel in response to different modifications and interactions

When observing pure PBS, we note a reduction peak centered at approximately -0.55 V, which is associated with the electro-reduction of dissolved oxygen in the electrolytic medium on the carbon surface or with redox processes involving the oxygenated functional groups of the graphene oxide surface itself, corroborating the data obtained via cyclic voltammetry [45]. When the enzyme-free polyacrylamide hydrogel matrix is deposited onto the electrode, a subtle change in the profile occurs: the reduction peak shifts slightly toward lower (more negative) potentials, and the background current undergoes reorganization, demonstrating that the physical presence of the gel modifies mass transport by diffusion to the electrode surface while preserving the system's redox activity. With the incorporation of the GOx enzyme into the hydrogel, the reduction peak shifts back to the -0.55 V region; the change

in shape and the slight decrease in baseline current within the -0.3 V to +0.4 V window, confirmed by FTIR analysis, suggest the successful anchoring of the GOx macromolecule within the matrix, which likely resulted in the partial blocking of capacitive sites without deactivating the system's active redox center. The most important behavior demonstrating the biosensor's operation is its response in the presence of glucose. This curve exhibits a marked decrease in the magnitude of the current, shifting to lower current density levels. This phenomenon arises from a chemical mechanism in which the enzyme entrapped within the hydrogel locally consumes the oxygen available to oxidize the glucose, resulting in less oxygen available at the electrode interface for electrochemical reduction [46]–[48]. The distinct residual peak in the voltammetric curve between -1.0 and approximately -0.6 V serves as the monitored analytical signal, such that the variation in the peak's height or area, or the overall decrease in current relative to the blank hydrogel curve, is directly related to the presence of glucose and its oxidation activity at the biosensor [46].

IV. Conclusion

This study validated the development of a promising electrochemical biosensor obtained through the carbonization of polyamide to synthesize graphene. SEM and EDS analyses confirmed an exfoliated porous mesostructure and the efficient anchoring of gold particles, while infrared spectroscopy (FTIR) verified the chemical integrity of the hydrogel. From an electrochemical standpoint, cyclic voltammetry (CV) and square-wave voltammetry (SWV) demonstrated the system's bioactivity and sensitivity to glucose and H2O2, exhibiting well-defined response profiles. Although this is an experimental device requiring further testing for optimization and measurement standardization, the platform stands as a viable, low-cost alternative for next-generation sensors. The hydrogel/electrode system will undergo future studies for the identification and quantification of other biochemical analytes for calibration purposes.

Acknowledgment

The authors thank the Universidade do Estado de Santa Catarina (UDESC) for providing the scholarship that supported this study. The authors also acknowledge the Coordenacao de Aperfeicioamento de Pessoal de Nivel Superior (CAPES) for the financial support for research supplies and materials.

References

[1] J. Ahmed, M. A. Rashed, M. Faisal, F. A. Harraz, M. Jalalah, and S. A. Alsareii, "Novel SWCNTs-mesoporous silicon nanocomposite as efficient non-enzymatic glucose biosensor," *Applied Surface Science*, vol. 552, p. 149477, Jun. 2021, doi: 10.1016/j.apsusc.2021.149477.

[2] S. Donmez, F. Arslan, N. Sarı, E. Hasanoğlu Özkan, and H. Arslan, "Glucose biosensor based on immobilization of glucose oxidase on a carbon paste electrode modified with microsphere-attached L-glycine," *Biotechnology and Applied Biochemistry*, vol. 64, no. 5, pp. 745–753, Sep. 2017, doi: 10.1002/bab.1533.

[3] W.-C. Lee, N. G. Gurudatt, D.-S. Park, K. B. Kim, C. S. Choi, and Y.-B. Shim, "Microneedle array sensor for monitoring glucose in single cell using glucose oxidase-bonded polyterthiophene coated on AuZn oxide layer," *Sensors and Actuators B: Chemical*, vol. 320, p. 128416, Oct. 2020, doi: 10.1016/j.snb.2020.128416.

[4] F. N. Comba, M. R. Romero, F. S. Garay, and A. M. Baruzzi, "Mucin and carbon nanotube-based biosensor for detection of glucose in human plasma," *Analytical Biochemistry*, vol. 550, pp. 34–40, Jun. 2018, doi: 10.1016/j.ab.2018.04.006.

[5] N. A. Chamaraja, M. Basavaraju, and N. K. Swamy, "Enzymatic method and its validation for the micromolar assay of glucose in human serum samples," *Analytical Biochemistry*, vol. 590, p. 113536, Feb. 2020, doi: 10.1016/j.ab.2019.113536.

[6] F. Sugandh et al., "Advances in the Management of Diabetes Mellitus: A Focus on Personalized Medicine," *Cureus*, Aug. 2023, doi: 10.7759/cureus.43697.

[7] M. Zhang, C. Liao, C. H. Mak, P. You, C. L. Mak, and F. Yan, "Highly sensitive glucose sensors based on enzyme-modified whole-graphene solution-gated transistors," *Scientific Reports*, vol. 5, no. 1, p. 8311, Feb. 2015, doi: 10.1038/srep08311.

[8] S. F. Clarke and J. R. Foster, "A history of blood glucose meters and their role in self-monitoring of diabetes mellitus," *British Journal of Biomedical Science*, vol. 69, no. 2, pp. 83–93, 2012.

[9] N. K. N. et al., "Evaluating the Efficacy of Cresvin beta on Type 2 Diabetes Mellitus Management: A Randomized Comparative Clinical Trial," *Pharmacognosy Journal*, vol. 16, no. 6s, pp. 1219–1226, Feb. 2025, doi: 10.5530/pj.2024.16s.234.

[10] K. Tonyushkina and J. H. Nichols, "Glucose Meters: A Review of Technical Challenges to Obtaining Accurate Results," *Journal of Diabetes Science and Technology*, vol. 3, no. 4, pp. 971–980, Jul. 2009, doi: 10.1177/193229680900300446.

[11] J. Walsh, R. Roberts, R. A. Vigersky, and F. Schwartz, "New Criteria for Assessing the Accuracy of Blood Glucose Monitors Meeting, October 28, 2011," *Journal of Diabetes Science and Technology*, vol. 6, no. 2, pp. 466–474, Mar. 2012, doi: 10.1177/193229681200600236.

[12] H. V. Dudukcu, "No Title," *Current Trends in Biomedical Engineering and Biosciences*, vol. 2, no. 3, 2017.

[13] T. Lister, P. A. Wright, and P. H. Chappell, "Optical properties of human skin," *Journal of Biomedical Optics*, vol. 17, no. 9, p. 090901, Sep. 2012, doi: 10.1117/1.JBO.17.9.090901.

[14] R. L. Brazg, L. J. Klaff, and A. M. Sussman, "New Generation Blood Glucose Monitoring System Exceeds International Accuracy Standards," *Journal of Diabetes Science and Technology*, vol. 10, no. 6, pp. 1414–1415, Nov. 2016, doi: 10.1177/1932296816652902.

[15] N. Wongkaew, M. Simsek, C. Griesche, and A. J. Baeumner, "Functional Nanomaterials and Nanostructures Enhancing Electrochemical Biosensors and Lab-on-a-Chip Performances: Recent Progress, Applications, and Future Perspective," *Chemical Reviews*, vol. 119, no. 1, pp. 120–194, Jan. 2019, doi: 10.1021/acs.chemrev.8b00172.

[16] M. H. Hassan, C. Vyas, B. Grieve, and P. Bartolo, "Recent Advances in Enzymatic and Non-Enzymatic Electrochemical Glucose Sensing," *Sensors*, vol. 21, no. 14, p. 4672, Jul. 2021, doi: 10.3390/s21144672.

[17] Y. Yao et al., "Integration of interstitial fluid extraction and glucose detection in one device for wearable non-invasive blood glucose sensors," *Biosensors and Bioelectronics*, vol. 179, p. 113078, May 2021, doi: 10.1016/j.bios.2021.113078.

[18] L. Özcan, Y. Şahin, and H. Türk, "Non-enzymatic glucose biosensor based on overoxidized polypyrrole nanofiber electrode modified with cobalt(II) phthalocyanine tetrasulfonate," *Biosensors and Bioelectronics*, vol. 24, no. 4, pp. 512–517, Dec. 2008, doi: 10.1016/j.bios.2008.05.004.

[19] X. Li et al., "A Fully Integrated Closed-Loop System Based on Mesoporous Microneedles-Iontophoresis for Diabetes Treatment," *Advanced Science*, vol. 8, no. 16, Aug. 2021, doi: 10.1002/advs.202100827.

[20] Y. Chen et al., "Skin-like biosensor system via electrochemical channels for noninvasive blood glucose monitoring," *Science Advances*, vol. 3, no. 12, Dec. 2017, doi: 10.1126/sciadv.1701629.

[21] J. Wang, "Electrochemical glucose biosensors," *Chemical Reviews*, vol. 108, no. 2, pp. 814–825, Feb. 2008, doi: 10.1021/cr068123a.

[22] A. Heller and B. Feldman, "Electrochemical glucose sensors and their applications in diabetes management," *Chemical Reviews*, vol. 108, no. 7, pp. 2482–2505, Jul. 2008, doi: 10.1021/cr068069y.

[23] Y. Wang, Y. Xu, J. Luo, Y. Ding, and Y. Liu, "Recent advances in nanomaterial-based electrochemical glucose biosensors," *Sensors*, vol. 20, no. 13, p. 3735, Jul. 2020, doi: 10.3390/s20133735.

[24] J. Claussen, A. Franklin, A. Haque, D. Porterfield, and T. Fisher, "Electrochemical biosensors based on carbon nanomaterials," *Journal of Materials Chemistry*, vol. 22, no. 30, pp. 15185–15201, 2012, doi: 10.1039/C2JM32110E.

[25] K. Novoselov, D. Jiang, F. Schedin, T. Booth, V. Khotkevich, S. Morozov, and A. Geim, "Two-dimensional atomic crystals," *Proceedings of the National Academy of Sciences*, vol. 102, no. 30, pp. 10451–10453, Jul. 2005, doi: 10.1073/pnas.0502848102.

[26] A. C. Ferrari et al., "Science and technology roadmap for graphene, related two-dimensional crystals, and hybrid systems," *Nanoscale*, vol. 7, no. 11, pp. 4598–4810, 2015, doi: 10.1039/C4NR01600A.

[27] C. Soldano, A. Mahmood, and E. Dujardin, "Production, properties and potential of graphene," *Carbon*, vol. 48, no. 8, pp. 2127–2150, Jul. 2010, doi: 10.1016/j.carbon.2010.01.058.

[28] M. Pumera, "Graphene in biosensing," *Materials Today*, vol. 14, no. 7–8, pp. 308–315, Jul. 2011, doi: 10.1016/S1369-7021(11)70160-2.

[29] D. Bitounis, H. Ali-Boucetta, B. H. Hong, D. H. Min, and K. Kostarelos, "Prospects and challenges of graphene in biomedical applications," *Advanced Materials*, vol. 25, no. 16, pp. 2258–2268, Apr. 2013, doi: 10.1002/adma.201203700.

[30] M. J. Allen, V. C. Tung, and R. B. Kaner, "Honeycomb carbon: a review of graphene," *Chemical Reviews*, vol. 110, no. 1, pp. 132–145, Jan. 2010, doi: 10.1021/cr900070d.

[31] S. Eigler and A. Hirsch, "Chemistry with graphene and graphene oxide—challenges for synthetic chemists," *Angewandte Chemie International Edition*, vol. 53, no. 30, pp. 7720–7738, Jul. 2014, doi: 10.1002/anie.201402780.

[32] V. Georgakilas, J. N. Tiwari, K. C. Kemp, J. A. Perman, A. B. Bourlinos, K. S. Kim, and R. Zboril, "Noncovalent functionalization of graphene and graphene oxide for energy materials, biosensing, catalytic, and biomedical applications," *Chemical Reviews*, vol. 116, no. 9, pp. 5464–5519, May 2016, doi: 10.1021/acs.chemrev.5b00620.

[33] Y. Liu, M. Yu, Y. Zhao, S. He, and Y. Wang, "Graphene-based biosensors for glucose detection," *Biosensors and Bioelectronics*, vol. 74, pp. 795–803, Dec. 2015, doi: 10.1016/j.bios.2015.07.033.

[34] S. Park and R. S. Ruoff, "Chemical methods for the production of graphenes," *Nature Nanotechnology*, vol. 4, no. 4, pp. 217–224, Apr. 2009, doi: 10.1038/nnano.2009.58.

[35] K. Yang, L. Feng, X. Shi, and Z. Liu, "Nano-graphene in biomedicine: theranostic applications," *Chemical Society Reviews*, vol. 42, no. 2, pp. 530–547, Jan. 2013, doi: 10.1039/C2CS35342C.

[36] Y. Shao, J. Wang, H. Wu, J. Liu, I. Aksay, and Y. Lin, "Graphene based electrochemical sensors and biosensors: a review," *Electroanalysis*, vol. 22, no. 10, pp. 1027–1036, May 2010, doi: 10.1002/elan.200900571.

[37] S. Pei and H.-M. Cheng, "The reduction of graphene oxide," *Carbon*, vol. 50, no. 9, pp. 3210–3228, Aug. 2012, doi: 10.1016/j.carbon.2011.11.010.

[38] A. K. Geim and K. S. Novoselov, "The rise of graphene," *Nature Materials*, vol. 6, no. 3, pp. 183–191, Mar. 2007, doi: 10.1038/nmat1849.

[39] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov, "Two-dimensional gas of massless Dirac fermions in graphene," *Nature*, vol. 438, no. 7065, pp. 197–200, Nov. 2005, doi: 10.1038/nature04233.

[40] C. Lee, X. Wei, J. W. Kysar, and J. Hone, "Measurement of the elastic properties and intrinsic strength of monolayer graphene," *Science*, vol. 321, no. 5887, pp. 385–388, Jul. 2008, doi: 10.1126/science.1157996.

[41] A. K. Geim, "Graphene: status and prospects," *Science*, vol. 324, no. 5934, pp. 1530–1534, Jun. 2009, doi: 10.1126/science.1158877.

[42] X. Huang, X. Qi, F. Boey, and H. Zhang, "Graphene-based composites," *Chemical Society Reviews*, vol. 41, no. 2, pp. 666–686, Jan. 2012, doi: 10.1039/C1CS15078B.

[43] D. A. C. Brownson and C. E. Banks, "Graphene electrochemistry: an overview of potential applications," *Analyst*, vol. 135, no. 11, pp. 2768–2778, Nov. 2010, doi: 10.1039/C0AN00590H.

[44] F. Perrozzi, S. Prezioso, and L. Ottaviano, "Graphene oxide: from fundamentals to applications," *Journal of Physics: Condensed Matter*, vol. 27, no. 1, p. 013002, Jan. 2015, doi: 10.1088/0953-8984/27/1/013002.

[45] S. Park and R. S. Ruoff, "Chemical methods for the production of graphenes," *Nature Nanotechnology*, vol. 4, no. 4, pp. 217–224, Apr. 2009, doi: 10.1038/nnano.2009.58.

[46] D. R. Dreyer, S. Park, C. W. Bielawski, and R. S. Ruoff, "The chemistry of graphene oxide," *Chemical Society Reviews*, vol. 39, no. 1, pp. 228–240, Jan. 2010, doi: 10.1039/B917103G.

[47] L. Feng and Z. Liu, "Graphene in biomedicine: opportunities and challenges," *Nanomedicine*, vol. 6, no. 2, pp. 317–324, Mar. 2011, doi: 10.2217/nnm.11.8.

[48] Y. Hernández, V. Nicolosi, M. Lotya, F. M. Blighe, Z. Sun, S. De, I. T. McGovern, B. Holland, M. Byrne, Y. K. Gunko, J. J. Boland, P. Niraj, G. Duesberg, S. Krishnamurthy, R. Goodhue, J. Hutchison, V. Scardaci, A. C. Ferrari, and J. N. Coleman, "High-yield production of graphene by liquid-phase exfoliation of graphite," *Nature Nanotechnology*, vol. 3, no. 9, pp. 563–568, Sep. 2008, doi: 10.1038/nnano.2008.215.

A Narrative Review of Embedded Systems for Oxygenation Monitoring via Photoacoustic Imaging

Rafaela Coca Marigheti da Silva* and Marcio Bender Machado†
*IFSP – Campus Sertãozinho
Sertãozinho, São Paulo, Brazil
Email: rafaelacocamarigheti@gmail.com
†Advisor, IFSP – Campus Sertãozinho
Sertãozinho, São Paulo, Brazil
Email: marciobma@gmail.com

***Abstract*—Tissue oxygenation monitoring is of great relevance in biomedical applications, particularly in neurological and neonatal contexts. Photoacoustic Imaging (PAI) combines optical excitation and ultrasonic detection to provide functional and morphological information non-invasively. This paper presents a bibliographic review on the integration of embedded systems, such as FPGAs, in the development of compact, efficient, and low-cost platforms for continuous oxygenation monitoring via PAI, with emphasis on neonatal applications.**

***Index Terms*—Photoacoustic Imaging; PAI; Photoacoustic Effect; Embedded Systems; FPGA; Oxygenation Monitoring; Biomedical Signal Processing; Ultrasonic Sensors.**

I. Introduction

Tissue oxygenation monitoring plays a fundamental role in biomedical applications, especially in neurological and neonatal contexts. This parameter reflects the amount of oxygen available in biological tissues, which is essential for proper cellular function and the maintenance of metabolic activity. Hemoglobin is central to this process, as it is responsible for oxygen transport in the bloodstream. Alterations in oxygenation levels may indicate severe clinical conditions, making continuous monitoring an important resource for diagnosis and medical follow-up [1].

Among the main clinical applications are the monitoring of stroke, hypoxia, and neonatal anoxia. In newborns, cerebral oxygenation monitoring is particularly critical due to the vulnerability of the developing nervous system and the risk of irreversible neurological damage [2].

However, many traditional monitoring methods present significant limitations. Systems commonly used in hospital environments tend to be costly, have low portability, and require complex infrastructure, making their use in continuous monitoring and out-of-hospital settings challenging. Furthermore, some equipment demands considerable computational power and robust hardware for biomedical signal processing, reducing accessibility in clinical contexts that require mobility and real-time response. In this regard, recent efforts have sought the development of more compact and accessible systems [3], [4].

In this context, Photoacoustic Imaging (PAI) emerges as a promising alternative: a non-invasive hybrid technique that combines optical excitation and ultrasonic detection to obtain functional and morphological biomedical images.

The basis of PAI-based oxygenation monitoring lies in the distinct optical absorption properties of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb), enabling non-invasive estimation of oxygen saturation in tissues, as further described in Section III-A [5], [6].

PAI offers high spatial resolution and significant functional contrast, enabling the visualization of internal structures and the analysis of optical properties of biological tissues. The technique provides structural and functional information in real time without requiring invasive procedures, positioning itself as a promising tool for advanced biomedical applications [6].

The application of PAI to neonatal cerebral monitoring has shown promising results, particularly through the transfontanelle approach, which exploits the natural acoustic and optical transparency of the fontanelle to acquire transcranial data safely and non-invasively. Recent studies conducted in neonatal animal models have demonstrated the feasibility of the technique for monitoring cerebral oxygenation, suggesting possibilities for early diagnosis and continuous monitoring that remain to be validated in human newborns [2]. Furthermore, advances in skull optical clearing techniques have enabled greater light penetration depth in transcranial applications in rodent models, favoring the acquisition of more detailed neurovascular images [7].

In parallel, the advancement of embedded systems has enabled the development of compact and efficient platforms for real-time acquisition and processing of biomedical signals. In this context, Field-Programmable Gate Arrays (FPGAs) and microcontrollers stand out due to their parallel processing capability, low latency, and high energy efficiency, making them particularly suited for biomedical applications that require continuous monitoring and reduced equipment size and cost [4].

To illustrate the operating principle of a PAI acquisition system, Fig. 1 presents a schematic overview of the technique.

The remainder of this paper is organized as follows: Section II presents the review methodology; Section III presents the fundamentals of photoacoustic imaging and the relation-

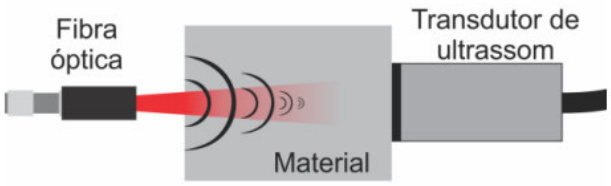


Fig. 1. Illustration of a photoacoustic imaging acquisition system. Source: [8].

ship between hemoglobin and oxygenation; Section IV describes embedded systems applied to PAI; Section V discusses the main applications in oxygenation monitoring; Section VI reviews and compares related works; Section VII identifies current challenges and limitations; Section VIII outlines future research directions; and Section IX presents the conclusion of this review.

II. Review Methodology

This study was conducted as a narrative literature review focused on the use of embedded systems in photoacoustic imaging (PAI), with particular emphasis on oxygenation monitoring and biomedical and neonatal applications. The literature search was conducted using IEEE Xplore, PubMed, Web of Science, and Google Scholar. The search was primarily focused on publications from 2020 onward, while selected earlier studies were included when considered fundamental to the understanding of PAI, photoacoustic oxygenation assessment, or embedded system architectures.

The search strategy was based on combinations of keywords related to the main themes of the review. The principal search strings included: '("photoacoustic imaging" OR photoacoustic) AND ("embedded systems" OR FPGA) AND (oxygenation OR "oxygen saturation")" and '("photoacoustic imaging" OR photoacoustic) AND (oxygenation OR "oxygen saturation") AND (neonatal OR newborn OR cerebral)". Additional combinations of related terms, including 'biomedical applications", 'photoacoustic imaging systems", 'embedded processing", and 'neonatal monitoring", were also considered according to the focus of each database search. The terms were adapted when necessary to account for differences in search syntax and indexing across IEEE Xplore, PubMed, Web of Science, and Google Scholar.

Publications in English and Portuguese were considered. Studies were included when they presented relevant contributions to at least one of the following topics: PAI systems and architectures, embedded or FPGA-based signal acquisition and processing, photoacoustic assessment of oxygenation, biomedical applications of PAI, or neonatal and cerebral monitoring. Classical studies published before 2020 were retained when they provided fundamental concepts or established technical principles relevant to the analysis.

Studies were excluded when they were not directly related to PAI, embedded systems, oxygenation monitoring, or the biomedical applications addressed in this review. Publications whose content did not provide sufficient relevance to the technological or clinical aspects investigated were also excluded from the comparative synthesis.

The final literature search was performed on July 12, 2026. The search across the four databases retrieved a total of 247 records (108 from IEEE Xplore, 60 from Google Scholar, 40 from Web of Science, and 39 from PubMed). Titles and abstracts of the retrieved records were manually screened for relevance to the scope of this review, and full texts were assessed against the inclusion and exclusion criteria described above. After this screening process, 16 studies were selected and retained for the comparative and analytical synthesis presented in this review. The selected studies were analyzed with respect to their imaging architecture, optical excitation source, embedded processing hardware, signal acquisition and reconstruction approaches, biomedical application, validation conditions, and reported limitations. As a narrative review, this study does not aim to provide an exhaustive systematic assessment of all publications in the field, but rather to identify representative technological approaches, current limitations, and research gaps relevant to the development of compact embedded PAI systems for oxygenation monitoring.

III. Fundamentals of Photoacoustic Imaging

Photoacoustic Imaging (PAI) is a hybrid technique that combines optical excitation and ultrasonic detection to obtain functional and morphological biomedical images non-invasively. This approach is based on the photoacoustic effect, in which pulsed light — typically from laser sources — is absorbed by biological tissues, causing a localized temperature increase and consequent thermoelastic expansion. This process generates ultrasonic waves that can be detected by appropriate transducers and subsequently processed to form images [5], [6].

Due to this physical mechanism, PAI enables the simultaneous acquisition of structural and functional tissue information, with particular emphasis on the optical properties of biological chromophores such as hemoglobin. This characteristic makes the technique especially promising for biomedical applications, particularly for monitoring tissue and cerebral oxygenation, since hemoglobin optical absorption is directly related to blood oxygenation levels [1], [5].

Moreover, PAI presents an advantageous combination of high spatial resolution and significant penetration depth in biological tissues. Different studies report values consistent with this potential: Mao et al. [2] demonstrated feasibility of transfontanelle cerebral oxygenation measurement in a neonatal piglet model, while Li et al. [7] reported transcranial neurovascular imaging in the rodent brain assisted by skull optical clearing. Taken together, these studies illustrate the range of depths and resolutions achievable with PAI depending on the specific system architecture and application, rather than a single fixed performance figure. Of particular note is its

use in neonatal contexts, such as transfontanelle evaluation of cerebral oxygenation, enabling non-invasive and continuous monitoring in newborn patients [2].

A. Relationship Between Hemoglobin and Oxygenation

The relationship between hemoglobin and oxygenation is one of the fundamental principles underpinning non-invasive measurement of oxygen saturation in biological tissues. Hemoglobin, the protein responsible for oxygen transport in the blood, exists primarily in two functional forms: oxygenated hemoglobin (HbO_2), when bound to oxygen, and deoxygenated hemoglobin (Hb), when oxygen is released to the tissues. These two forms exhibit distinct optical properties, particularly regarding their absorption spectra, which enables the differentiation and quantification of blood oxygenation through optical methods [5], [6].

The difference between HbO_2 and Hb is directly related to how each form absorbs light at different wavelengths. Oxygenated and deoxygenated hemoglobin present distinct absorption coefficients, generating measurable optical contrasts when exposed to light sources. By analyzing the interaction of light with biological tissues, it is therefore possible to estimate the ratio between HbO_2 and Hb, allowing inference of blood oxygenation levels [1], [6].

This physical principle underlies a range of biomedical technologies, including PAI, in which the absorption contrast between HbO_2 and Hb is converted into measurable acoustic signals. In PAI, the relative photoacoustic responses associated with HbO_2 and Hb can be used to assess local tissue oxygenation and, when appropriate, estimate tissue oxygen saturation (sO_2). This parameter should be distinguished from peripheral oxygen saturation (SpO_2), which is conventionally measured by pulse oximetry and reflects arterial blood oxygenation [2], [5].

The spectral difference between the two hemoglobin forms therefore enables continuous measurement of tissue oxygenation, establishing itself as a central principle in modern biomedical applications. This characteristic is especially relevant in cerebral and neonatal monitoring contexts, where small variations in oxygenation may indicate significant physiological changes and potential clinical risk [1], [2].

IV. Embedded Systems in Photoacoustic Imaging

Embedded systems have played a central role in the development of compact and efficient PAI platforms, enabling real-time acquisition and processing of biomedical signals with low power consumption. In general, an embedded photoacoustic system comprises four main functional blocks: (i) optical excitation source, responsible for generating light pulses; (ii) ultrasonic transducer, responsible for detecting the acoustic waves generated in the tissue; (iii) signal conditioning and acquisition circuit, which amplifies and digitizes the detected signals; and (iv) embedded processing unit, responsible for filtering, image reconstruction, and data analysis.

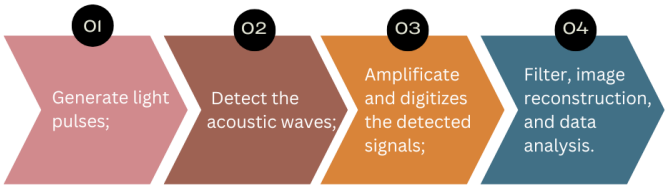


Fig. 2. Functional blocks of an embedded photoacoustic system. Source: elaborated by the authors.

The literature indicates that the choice of each component directly impacts overall system performance, entailing trade-offs between cost, portability, resolution, and processing speed. The following subsections present and compare the main works identified in the literature, followed by an analytical description of each functional block.

A. System Block Description

Based on the reviewed works, the main blocks composing an embedded photoacoustic system are described and analyzed as follows.

Optical excitation source: may be implemented using pulsed lasers or high-power LEDs. As discussed, lasers offer superior performance in depth, while LEDs provide lower cost and greater portability [9]. The choice should be guided by the intended clinical application.

Ultrasonic transducer: detects the acoustic waves generated by the photoacoustic effect and converts them into electrical signals. Transducer selection directly influences the system's spatial resolution and penetration depth [5], making it a critical component for final image quality.

Signal conditioning and acquisition circuit: amplifies, filters, and digitizes the electrical signals from the transducer. Precise synchronization between optical emission and the acquisition window is an essential requirement at this stage, as demonstrated by Alizadeh et al. [4], whose system showed that timing errors directly compromise the quality of reconstructed images.

Embedded processing unit (FPGA): executes filtering, image reconstruction, and system control tasks in parallel. The use of an FPGA significantly reduces latency compared to conventional CPU-based solutions, being essential for real-time operation [4], [10]. Integration with *compressed sensing* techniques represents a promising development toward reducing computational demand without sacrificing image quality.

Fig. 3 illustrates the general architecture of an embedded photoacoustic acquisition system.

B. FPGA versus Microcontrollers

It is observed that FPGA-based systems predominate in photoacoustic implementations requiring real-time processing, owing to their capacity for parallel execution of multiple tasks. Alizadeh et al. [4] demonstrated that a system combining an FPGA with high-power LEDs enables photoacoustic signal acquisition and processing with adequate precision for biomedical applications, with precise control of the timing

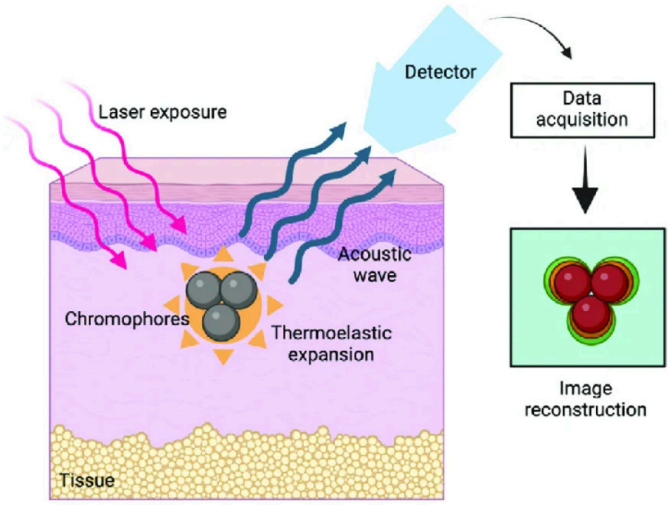


Fig. 3. Photoacoustic imaging system architecture. Source: [11].

between optical excitation and acoustic acquisition — a critical requirement for image quality. In contrast, although microcontrollers offer lower cost and simpler programmability, the literature indicates that their sequential processing architecture limits performance in applications requiring high-speed image reconstruction, making FPGAs the more appropriate choice for high-performance embedded PAI systems [4], [10].

Wang et al. [10] extended this line of work by investigating the use of *compressed sensing* in embedded photoacoustic systems, demonstrating that the number of samples required for image reconstruction can be significantly reduced without relevant quality loss. This approach is particularly relevant for resource-constrained embedded systems, as it alleviates the processing load on the FPGA and reduces energy consumption — a notable advancement over conventional full-sampling architectures.

Hossain et al. [3] developed a compact system for photoacoustic spectral response measurement (PASR), a pump-probe technique that extracts acoustic spectral signatures from biological tissues. Unlike the FPGA-based architecture of Alizadeh et al. [4], which prioritizes parallel processing and temporal precision, Hossain et al. employed a high-speed microcontroller as the embedded processing unit, combined with a 12-bit analog-to-digital converter (ADC) operating at 30 MSPS for real-time signal conditioning and data acquisition. The developed hardware includes a high-frequency preamplifier, a low-pass filter, an electrically isolated trigger circuit, and a high-speed ADC, enabling compact and low-cost implementation. The system was validated in the differentiation of lung and liver tissues, demonstrating the viability of microcontroller-based embedded solutions for biomedical applications outside laboratory settings. Its main limitation, however, lies in the absence of validation for real-time image reconstruction, as the system was designed for point spectral measurements rather than continuous imaging.

C. LED versus Laser as Excitation Source

The choice of optical excitation source is one of the primary factors determining the performance-cost trade-off in PAI systems. The literature indicates that pulsed lasers offer higher optical energy per pulse, resulting in greater penetration depth and better signal-to-noise ratio (SNR), and are therefore preferred in research applications requiring high image quality [5]. On the other hand, there is a growing trend toward replacing lasers with high-power LEDs, driven by significant cost reduction, improved operational safety, and lower implementation complexity [9].

Kuniyil Ajith Singh et al. [9] conducted a comprehensive review demonstrating that LEDs, when driven by appropriate pulsing circuits, are capable of generating detectable photoacoustic signals in biological tissues. The authors note, however, that the lower irradiance of LEDs compared to lasers still limits penetration depth and SNR in thicker tissues — representing the main disadvantage of this approach. It is therefore observed that LEDs are more suitable for superficial applications or anatomically favorable regions, such as transfontanelle neonatal evaluation, while lasers remain more appropriate for deeper imaging.

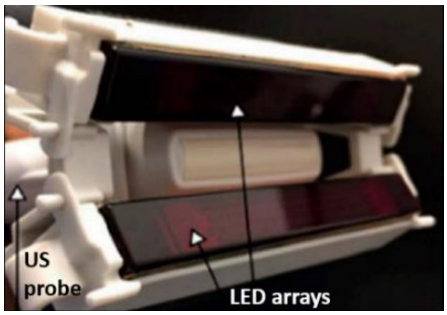


Fig. 4. LED array arrangement around the ultrasound (US) probe. Adapted from [12].

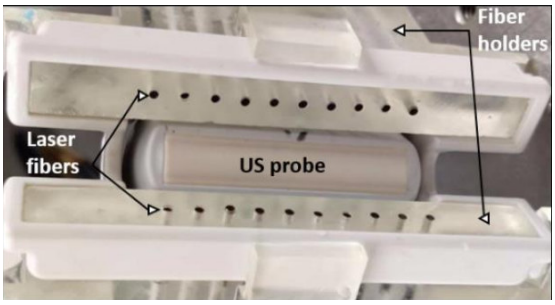


Fig. 5. Laser fiber arrangement around the ultrasound (US) probe. Adapted from [12].

V. Applications in Oxygenation Monitoring

PAI has been extensively investigated for biomedical applications including soft tissue imaging, vascular structure visualization, blood oxygenation monitoring, and functional brain studies. These applications demonstrate the technique's

potential as a non-invasive tool with high sensitivity to physiological variations and strong functional contrast, particularly in scenarios requiring continuous monitoring of biological parameters [5], [6].

In the context of neonatal cerebral monitoring, PAI holds particular clinical relevance. Conditions such as hypoxic-ischemic encephalopathy (HIE) are associated with high mortality and potential neurological sequelae in newborns, making early monitoring of cerebral tissue oxygen saturation (sO_2) essential [13], [14]. PAI stands out in this scenario by enabling continuous, non-invasive assessment of sO_2 — overcoming limitations of conventional methods such as electroencephalography (EEG), which provides electrical brain activity data but does not allow direct mapping of tissue oxygenation, and of peripheral pulse oximetry, which measures SpO_2 and reflects arterial blood oxygenation rather than local tissue status.

Neonatal applications have gained prominence, particularly through the transfontanelle approach, which exploits the natural acoustic and optical window of the newborn skull. Mao et al. [2] demonstrated in animal models that this approach enables safe and non-invasive acquisition of cerebral signals, allowing continuous monitoring of cerebral tissue oxygen saturation (sO_2) and early detection of neurological alterations.

PAI can also be applied to vascular monitoring, enabling real-time analysis of blood flow and local tissue oxygen saturation (sO_2) in peripheral tissues — a functional parameter distinct from the peripheral SpO_2 obtained by conventional pulse oximetry. This capability supports functional assessment of biological systems in clinical environments such as intensive care units (ICUs), where continuous patient monitoring is essential for rapid clinical decision-making [2], [5].

The development of portable PAI-based systems has driven the miniaturization of biomedical equipment. The literature indicates that compact architectures with low power consumption expand the potential for use in bedside monitoring and remote applications, particularly in regions with limited access to specialized hospital infrastructure [2], [6].

Complementarily, the integration of PAI with Artificial Intelligence has improved image reconstruction and the automated classification of physiological patterns, as discussed by Lan et al. [15], while the use of high-power LEDs as excitation sources provides a lower-cost alternative to conventional lasers, enhancing system portability [9].

VI. Related Work

A number of studies have investigated the integration of embedded systems and photoacoustic imaging techniques for biomedical monitoring. The following analysis highlights the contributions, advantages, limitations, and research gaps identified in the reviewed literature. Table I summarizes the technical parameters of the key reviewed works, illustrating the diversity of optical sources, embedded processing architectures, reconstruction methods, and clinical applications covered in this review.

Mao et al. [2] employed a transfontanelle PAI setup using two pulsed Nd:YAG lasers operating across a range of 690–950 nm, combined with a 128-element phased-array ultrasound probe. By exploiting the isobestic wavelength at 798 nm — at which HbO_2 and Hb absorb light equally — the system was able to quantify cerebral hemorrhage independently of tissue oxygen saturation (sO_2), which is particularly relevant in neonates where sO_2 levels fluctuate rapidly. The study experimentally validated the feasibility of the transfontanelle window as a non-invasive acoustic and optical access point for cerebral monitoring in a neonatal piglet model, used as a translational proxy for human newborns; validation in human neonates remains to be performed. Its main limitation remains the use of bulk laser sources, which hinder miniaturization and clinical portability.

Alizadeh et al. [4] proposed an embedded PAI platform featuring an 8-channel analog front-end, ADCs operating at up to 250 MSPS, and a Xilinx Spartan-7 FPGA responsible for real-time image reconstruction via a spatial phase algorithm. The LED excitation source is angularly scanned to acquire data from multiple directions, enabling synthetic aperture reconstruction. This architecture represents one of the most complete low-cost embedded PAI implementations in the reviewed literature, demonstrating that hardware-level integration of excitation control, signal acquisition, and image reconstruction is feasible within a single compact platform. The identified limitation remains penetration depth, constrained by LED irradiance.

Beard [5] presented a photoacoustic scanner based on a Fabry–Perot optical sensor, which scans the acoustic wavefront using a focused 64 μm interrogation beam. This approach eliminates the aperture and bandwidth limitations inherent to conventional piezoelectric transducer arrays, enabling wideband detection across a large sensing area. The system demonstrated 3D reconstruction of palmar vasculature at depths up to 4 mm and was subsequently applied to tumor imaging, murine brain, and embryos. Although the system achieves outstanding resolution and sensitivity, it relies on high-power pulsed lasers and complex optical alignment, limiting its suitability for compact embedded implementations.

Wang and Hu [6] provided a comparative framework covering five PAI architectures — optical-resolution photoacoustic microscopy (OR-PAM), acoustic-resolution photoacoustic microscopy (AR-PAM), linear and circular photoacoustic computed tomography (PACT), and photoacoustic endoscopy (PAE) — and their fundamental trade-offs between spatial resolution, penetration depth, and hardware complexity. This conceptual reference is directly applicable to the selection of embedded system architectures, as it clarifies which configurations are compatible with low-cost, miniaturized implementations and which require expensive or bulky hardware.

Li et al. [7] demonstrated transcranial PAI in neonatal piglets for the localization of photothrombotic stroke lesions, using spectroscopic imaging to map HbO_2 and Hb distributions. Lesions were validated post-mortem by TTC staining 24 h after stroke induction, confirming that ischemic regions

TABLE I
COMPARATIVE SUMMARY OF EMBEDDED PHOTOACOUSTIC IMAGING SYSTEMS REPORTED IN THE REVIEWED LITERATURE.

Parameter	Mao et al. [2]	Alizadeh et al. [4]	Hossain et al. [3]	Beard [5]	Li et al. [7]	Wang et al. [10]	Sari et al. [11]	Agrawal et al. [12]
Optical Source	Nd:YAG laser, 690–950 nm	High-power LED (wavelength n.s.)	Not specified (pump–probe PASR)	Pulsed laser (wavelength n.s.)	Not specified (spectroscopic PAI)	Laser, single-pixel encoding via optical mask	532 nm laser diode, 200 mW	LED array, 850 nm, 200 μJ/pulse, 70 ns, vs. conventional laser
Transducer / Detector	128-element phased-array US probe	8-channel analog front-end	–	Fabry–Pérot optical sensor (64 μm interrogation beam)	–	Single detector element	Condenser microphone	128-element linear US probe
Sampling Rate	–	Up to 250 MSPS	30 MSPS (12-bit ADC)	–	–	–	–	–
Embedded Processing	–	Xilinx Spartan-7 FPGA	High-speed microcontroller	Not embedded (benchtop optical system)	–	–	Not embedded (motorized XY stage)	–
Reconstruction Method	Spectroscopic quantification at isosbestic point (798 nm)	Spatial phase algorithm (synthetic aperture, angular LED scanning)	Point spectral measurement (PASR); not an imaging system	–	Spectroscopic imaging (HbO_2/Hb mapping)	Compressed sensing (single-pixel camera principle)	Point-by-point mechanical scanning	Not specified (comparative benchtop study)
Sample / Validation	Neonatal piglets, transfontanelle	–	Ex vivo lung and liver tissue	Palmar vasculature (human), tumor, murine brain, embryos	Neonatal piglets, transcranial, with optical skull clearing	–	Ex vivo dental tissue (enamel, dentin, cementum, pulp)	–
Depth / Resolution	Up to 5.2 cm	Limited by LED irradiance (value not reported)	–	Up to 4 mm (palmar vasculature)	Sub-100 μm spatial resolution	–	–	Not reported (SNR gap noted for deeper tissue)
Main Limitation	Bulky laser sources	Penetration depth restricted by LED irradiance	No validation for real-time image reconstruction	Requires high-power lasers, complex optical alignment	Clinical applicability of skull clearing needs safety studies	Increased reconstruction processing time	–	Remaining SNR gap for deeper tissue imaging

Note: "–" indicates that the parameter was not reported in the corresponding publication. Explicit terms such as "Not embedded" are used instead of "–" whenever the absence of a feature is a fact stated or clearly implied by the reviewed study, rather than a missing data point.

were clearly identifiable by their characteristic low HbO_2 and high Hb signatures. This work established an important proof of concept for PAI as a tool for real-time, contrast-free functional brain mapping, directly relevant to the neonatal monitoring context addressed in this review.

Agrawal et al. [12] performed a direct experimental comparison between a conventional laser source and LED arrays (850 nm, 200 μJ/pulse per array, 70 ns pulse width) mounted on both sides of a 128-element linear ultrasound probe. The identical probe geometry allowed a controlled substitution of the laser fiber by the LED array, enabling fair performance comparison under equivalent conditions. The study demonstrated that LED arrays operating at high repetition rates (1–4 kHz) provide a viable, lower-cost, and more compact alternative to pulsed lasers, while acknowledging the remaining gap in SNR for deeper tissue imaging.

Wang et al. [10] proposed the application of single-pixel camera principles to PAI: an optical mask spatially encodes the laser illumination pattern before tissue irradiation, and a single detector captures the resulting photoacoustic signal for each mask pattern. Image reconstruction is then performed via compressed sensing algorithms, significantly reducing the number of acquisitions and detector elements required. This approach is especially relevant for embedded systems, as it simplifies the acquisition hardware and reduces computational and data storage demands, at the cost of increased reconstruction processing time.

Sari et al. [11] demonstrated a low-cost photoacoustic microscopy system combining a 532 nm laser diode (200 mW), a condenser microphone as acoustic detector, and a motorized XY scanning stage for ex vivo characterization of dental tissue layers. The system successfully differentiated enamel, dentin, cementum, and pulp based on their distinct photoacoustic signal amplitudes — pulp generating the strongest response and enamel the weakest — and further distinguished carious from healthy teeth based on the enhanced thermoelastic expansion of deteriorated tissue. This work illustrates that low-cost embedded architectures can deliver meaningful diagnostic information even with minimal hardware complexity.

Overall, a common gap is noted across the reviewed works: the absence of systems that simultaneously integrate LED excitation, FPGA processing, real-time operation, and clinical neonatal validation within a single compact platform. This gap constitutes the primary motivation and potential contribution of the present work.

VII. CHALLENGES AND LIMITATIONS

Despite recent advances in photoacoustic imaging and embedded systems, significant technical and clinical challenges remain to be addressed before PAI can be widely deployed for continuous and neonatal monitoring applications.

One of the primary challenges concerns signal quality. Photoacoustic signals are frequently affected by environmental noise, electromagnetic interference, and patient movement. The signal-to-noise ratio (SNR) is a critical factor, particularly in low-cost LED-based systems — such as those proposed by Alizadeh et al. [4] and reviewed by Kuniyil Ajith Singh et al. [9] — which present lower optical energy compared to conventional lasers, resulting in reduced image contrast in thicker tissues.

Penetration depth also represents a relevant limitation, and reported values vary considerably depending on the specific system and tissue target rather than following a single benchmark. Mao et al. [2] evaluated transfontanelle acquisition in a neonatal piglet model, where the natural acoustic and optical window of the skull facilitates signal transmission, whereas Li et al. [7] addressed the more demanding scenario of transcranial imaging through an intact rodent skull, assisted by skull optical clearing to extend penetration. Imaging deeper structures — such as the human neonatal brain outside the fontanelle window — still imposes important technical constraints regardless of the specific approach. The skull optical clearing technique investigated by Li et al. [7] extends

achievable depth in transcranial settings, but its direct clinical applicability still requires additional safety and biocompatibility studies.

In the context of embedded systems, precise synchronization between optical emission and the acoustic acquisition window is an essential requirement, as demonstrated by Alizadeh et al. [4]. Small timing errors introduce distortions that compromise diagnostic accuracy. Furthermore, the miniaturization of platforms integrating acquisition, processing, storage, and communication in a single device requires careful consideration of thermal dissipation, power consumption, and operational reliability — aspects that most reviewed works do not fully address.

The computational cost of image reconstruction algorithms and AI-based analysis poses another challenge for embedded implementations [15]. Compressed sensing techniques, proposed by Wang et al. [10], reduce this demand but still lack validation in real clinical scenarios with low-cost embedded hardware — a gap directly identified in the related work review.

Finally, recent publications such as Park et al. [16] indicate that the clinical translation of PAI is advancing, with growing technological maturity and increasing industry interest. The literature suggests that future developments may converge toward integrated platforms combining LED excitation, FPGA-based embedded processing, and AI-assisted analysis. Such integration represents a promising research direction for the development of compact and potentially lower-cost systems for oxygenation monitoring, particularly in neonatal applications. However, the feasibility, clinical performance, portability, and cost-effectiveness of such a platform remain to be experimentally demonstrated and clinically validated.

VIII. Future Perspectives

The analysis of the reviewed literature indicates that photoacoustic imaging has strong potential to establish itself as a non-invasive biomedical monitoring tool, particularly for neonatal applications. A growing trend of convergence is observed across three development fronts: hardware miniaturization, embedded algorithm optimization, and clinical validation in specific patient populations.

On the hardware side, the gradual replacement of lasers by high-power LEDs, combined with FPGA-based real-time processing, suggests a possible pathway toward more compact and accessible systems [4], [9]; however, this outcome has so far been demonstrated only in isolated components rather than in a single integrated platform. The literature indicates that the main advances expected in this direction involve the development of higher-irradiance LED driver circuits and the integration of multiple acquisition channels on single-chip platforms, which could contribute to reduced size and power consumption if successfully implemented and validated.

Regarding transcranial imaging, advances in optical clearing techniques [7] and in the transfontanelle approach [2] open the possibility of neonatal cerebral monitoring with greater depth and resolution. It is noted, however, that the transition of these techniques to clinical environments depends on safety studies and validation in human clinical trials — representing one of the primary gaps to be addressed in the coming years.

The integration of PAI with Artificial Intelligence represents one of the most promising perspectives for enhancing diagnostic precision and automating the analysis of large volumes of physiological data in real time [15]. There is a growing trend toward the use of neural networks for photoacoustic image reconstruction from undersampled data, which — combined with compressed sensing techniques [10] — may enable high-performance embedded systems with low computational cost.

Finally, recent publications such as Park et al. [16] indicate that the clinical translation of PAI is advancing, with growing technological maturity and increasing industry interest. Taken together, the works reviewed in this paper suggest that an embedded platform combining LED excitation, FPGA-based processing, and AI-assisted analysis for continuous neonatal oxygenation monitoring is a technically plausible research direction. Whether such a platform would in practice achieve portability, low cost, and clinical readiness, however, remains to be experimentally demonstrated and validated; these are properties that the reviewed literature points toward as design goals, not outcomes already established for an integrated system.

IX. Conclusion

This narrative review examined the use of embedded systems, particularly FPGAs, in the development of photoacoustic imaging (PAI) platforms for tissue oxygenation monitoring, with emphasis on neonatal applications. The reviewed literature demonstrates, individually, that FPGA-based embedded processing can achieve real-time image reconstruction with low latency [4], [10], that high-power LEDs constitute a viable lower-cost and safer alternative to pulsed lasers for superficial and anatomically favorable regions [9], [12], and that the transfontanelle approach enables non-invasive acquisition of cerebral oxygenation signals in neonatal models [2].

However, no single work identified in this review integrates LED excitation, FPGA-based processing, and clinical neonatal validation within one compact platform. Consequently, portability, low cost, and clinical readiness should be understood as design goals and research directions suggested by the trends observed across the reviewed literature, rather than as properties already established for an integrated embedded PAI system. Achieving such an integration, and demonstrating its performance under real clinical conditions, remains an open challenge for future research, as discussed in Section VII and Section VIII.

Acknowledgment

The authors thank IFSP – Campus Sertãozinho for the institutional support and infrastructure provided for the development of this research. They also acknowledge the Pro-Rectory of Research of IFSP and the Brazilian National Council for Scientific and Technological Development (CNPq) for funding this work.

REFERENCES

[1] Y. Wang, Y. Liu, H. Li, and X. Dai, "Photoacoustic imaging for brain monitoring: challenges and opportunities," *Neurophotonics*, vol. 8, no. 3, p. 032109, 2021, doi: 10.1117/1.NPh.8.3.032109.

[2] Y. Mao *et al.*, "Transfontanelle photoacoustic imaging of brain oxygenation in neonatal piglets: a pilot study," *Photoacoustics*, vol. 28, 2023.

[3] M. S. Hossain *et al.*, "Development of a compact, cost-effective photoacoustic spectral response measurement system for biomedical applications," *Heliyon*, vol. 9, no. 7, p. e18166, Jul. 2023, doi: 10.1016/j.heliyon.2023.e18166.

[4] M. R. Alizadeh, M. H. Ebrahimzadeh, and H. R. K. Sajjadi, "Implementation of a precise and low-cost photoacoustic imaging system based on high-power LED illumination and FPGA-based circuits," *Archives of Acoustics*, vol. 48, no. 3, pp. 355–366, 2023, doi: 10.24425/ao.2023.146669.

[5] P. Beard, "Biomedical photoacoustic imaging," *Interface Focus*, vol. 1, no. 4, pp. 602–631, 2011.

[6] L. V. Wang and S. Hu, "Photoacoustic tomography: in vivo imaging from organelles to organs," *Science*, vol. 335, no. 6075, pp. 1458–1462, 2012.

[7] H. Li *et al.*, "Transcranial photoacoustic imaging of neurovascular dynamics in the rodent brain with skull optical clearing," *Neurophotonics*, vol. 9, no. 2, p. 025002, 2022.

[8] T. Z. Pavan, A. A. O. Carneiro, and S. Y. Emelianov, "Fotoacústica: imagens da interação de luz e som no corpo humano," *Ciência Hoje*, vol. 50, no. 297, pp. 38–42, 2012.

[9] M. Kuniyil Ajith Singh, W. Xia, and W. Steenbergen, "Light emitting diodes based photoacoustic imaging: a review," *Photoacoustics*, vol. 20, p. 100209, Dec. 2020, doi: 10.1016/j.pacs.2020.100209.

[10] Y. Wang, Y. Chen, Y. Zhao, and S. Liu, "Compressed sensing for biomedical photoacoustic imaging: a review," *Sensors*, vol. 24, no. 9, p. 2670, 2024.

[11] A. W. Sari, R. Widyaningrum, A. Setiawan, and Mitrayana, "Recent development of photoacoustic imaging in dentistry: A review on studies over the last decade," *Saudi Dental Journal*, vol. 35, no. 5, pp. 423–436, Jul. 2023, doi: 10.1016/j.sdentj.2023.05.013.

[12] S. Agrawal, M. Kuniyil Ajith Singh, X. Yang, H. Albahrani, A. Dangi, and S.-R. Kothapalli, "Photoacoustic imaging capabilities of light emitting diodes (LED) and laser sources: a comparison study," in *Photons Plus Ultrasound: Imaging and Sensing 2020*, A. A. Oraevsky and L. V. Wang, Eds., Proc. SPIE, vol. 11240, art. 1124059, 2020, doi: 10.1117/12.2547012.

[13] K. Q. Zhou *et al.*, "Treating seizures after hypoxic-ischemic encephalopathy — current controversies and future directions," *Int. J. Mol. Sci.*, vol. 22, no. 13, p. 7121, 2021.

[14] B. B. Abate *et al.*, "Effects of therapeutic hypothermia on death among asphyxiated neonates with hypoxic-ischemic encephalopathy," *PLoS ONE*, vol. 16, no. 2, e0247229, 2021.

[15] H. Lan *et al.*, "Photoacoustic imaging and artificial intelligence: a new paradigm for medical diagnosis," *Frontiers in Neuroimaging*, 2018.

[16] J. Park *et al.*, "Clinical translation of photoacoustic imaging," *Nature Reviews Bioengineering*, vol. 3, no. 3, pp. 193–212, 2025.

Analisador de Contração Muscular por EMG Facial: Desenvolvimento e Avaliação Preliminar de um Protótipo

1st Emily Horrana Lima R. Armando
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
emily.horrana@ges.inatel.br

2nd Maria Eduarda Costa Gonçalves
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
maria.eg@geb.inatel.br

3rd Antônio Augusto D'Assumpção
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
a.augusto@ges.inatel.br

4th Luma Rissatti Borges do Prado
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
luma.prado@inatel.br

5th Airton Rodrigues
NENC - Núcleo de Estudos em
Neurociência do Consumidor
São Paulo, Brasil
airton.rodrigues@nenc.com.br

6th Elisa Renno Carneiro Dester
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
elisa.renno@inatel.br

***Abstract*—This paper presents a preliminary prototype for the acquisition, processing, and real-time visualization of surface EMG signals from the zygomaticus major muscle, as an initial step toward future applications in consumer neuroscience. The system uses commercial surface electrodes and an ESP32 microcontroller; the acquired signal is conditioned through a high-pass filter, a power-line interference notch filter, rectification, and a moving-average envelope, and streamed in real time to a web-based interface via a WebSocket connection. The reliability of this real-time link was evaluated separately, in six recording sessions with one participant, showing 0% sample loss and an average effective sampling rate of 343 Hz. EMG detection was assessed with three participants under a standardized protocol comprising rest and smiling phases (habitual and maximal), repeated across three sessions per participant at different times of day. The system was able to distinguish muscle contraction from rest periods.**

***Index Terms*—Facial electromyography, EMG signal processing, consumer neuroscience.**

I. Introdução

A Neurociência do Consumidor é um campo de estudo que visa compreender, de forma objetiva, as respostas neurofisiológicas humanas diante de estímulos de *marketing*, investigando os processos mentais conscientes e não conscientes envolvidos nas decisões de compra. Uma das principais problemáticas da área está relacionada à subjetividade dos métodos tradicionais, que se baseiam em questionários respondidos pelo próprio usuário, sujeitos a vieses interpretativos.

Em contrapartida, tem crescido o uso de métodos capazes de medir respostas implícitas a partir da captação de reações fisiológicas dos consumidores, como reconhecimento facial [1], *eyetracking* e eletromiografia (EMG) [3]. Dentre elas, o EMG tem sido empregado em diversos estudos para o reconhecimento de expressões faciais por apresentar sensibilidade na detecção de movimentos dos músculos faciais, sendo capaz de identificar ativações musculares sutis, de baixa intensidade, que não são perceptíveis visualmente [3].

Apesar dos avanços tecnológicos, estudos que utilizam o EMG facial como meio de detecção ainda apresentam desafios, como interferências externas e ruídos resultantes de movimentos indesejados ou dos eletrodos empregados, além de dificuldades relacionadas à amplificação e ao processamento do sinal eletromiográfico [6].

Diante desse contexto, este trabalho tem como objetivo desenvolver um protótipo de aquisição de EMG facial, como etapa inicial para futuras aplicações em neurociência do consumidor, permitindo o monitoramento da atividade do músculo zigomático maior e a visualização dos dados de maneira simples e intuitiva por meio de uma plataforma *web*.

II. Metodologia

A metodologia empregada neste trabalho foi dividida em quatro etapas principais: (A) estudo teórico sobre eletromiografia e *neuromarketing*; (B) desenvolvimento do protótipo de *hardware*; (C) implementação do *software*; e (D) captação e validação da comunicação;

A. Estudo teórico sobre eletromiografia e neuromarketing

A abordagem adotada baseou-se em um estudo prévio sobre a eletromiografia aplicada à Neurociência do Consumidor, com foco na análise dos tipos de eletrodos, seu posicionamento e das principais interferências que afetam a aquisição do sinal EMG. Foram averiguados alguns trabalhos que utilizam o EMG como ferramenta para detectar expressões faciais associadas à estímulos visuais, evidenciando sua aplicação em estudos de *neuromarketing*.

B. Desenvolvimento do protótipo de hardware

O desenvolvimento do protótipo concentrou-se na implementação do sistema de aquisição e condicionamento do sinal eletromiográfico. Os principais componentes utilizados foram:

- Módulo sensor eletromiográfico *Muscle Sensor v3 (Advancer Technologies)* [5] com eletrodos descartáveis de superfície Ag/AgCl para captação dos sinais EMG na face;
- Baterias de 9V para alimentação do sistema;
- Microcontrolador ESP32 para conversão analógico digital, processamento dos sinais e comunicação com a interface *web*;
- Circuito de proteção composto por um diodo Zener e um resistor de $47\,\Omega$ para limitar a tensão de saída do circuito aos níveis seguros de operação das entradas analógicas do microcontrolador, cuja faixa varia de 0V até 3,3V.

Para a captação do sinal eletromiográfico, foi utilizado o módulo de captação com dois eletrodos para captação bipolar no músculo zigomático e um eletrodo de referência na testa, capaz de detectar as variações de potencial elétrico geradas pela atividade muscular, cuja amplitude típica encontra-se na faixa de microvolts (μV) a 5 milivolts (mV).

A saída da placa de aquisição foi conectada a um circuito de proteção para o ESP32, composto por um resistor de $47\,\Omega$ em série com um diodo Zener modelo 1N4728.

Após a amplificação e limitação da tensão de saída, o sinal condicionado foi conectado ao pino analógico GPIO34 do microcontrolador ESP32 para conversão analógico-digital e processamento posterior.

A alimentação do sistema foi realizada por meio de 2 baterias de 9V, fornecendo tensão simétrica de ±9V. Foi escolhido o uso de baterias para minimizar interferências provenientes da rede elétrica e reduzir ruídos no sinal.

Inicialmente, a montagem foi realizada em protoboard para validação experimental do circuito. Após todos os testes, foi projetada uma placa de circuito impresso (PCI) no sotware EasyEDA, visando maior robustez, organização dos componentes e possibilidade de implementação definitiva do sistema.

C. Implementação do software

A arquitetura de *software* do sistema foi dividida em *frontend* e *backend*, responsáveis, respectivamente, pela interface gráfica de visualização e pelo controle da aquisição e transmissão dos sinais EMG. Desenvolvida no ambiente de desenvolvimento integrado Visual Studio Code, juntamente com a extensão PlatformIO [9] que permite o desenvolvimento de sistemas embarcados, suportando placas como o ESP32.

O *backend* do sistema foi implementado em C++ e executado em um microcontrolador ESP32, responsável pela aquisição do sinal EMG, pelo processamento do sinal e pela comunicação com o *frontend*. O ESP32 foi configurado para operar como um *Access Point Wi-Fi*, hospedando um servidor HTTP, utilizado para o controle de início e parada da aquisição, e um servidor *WebSocket*, utilizado para a transmissão dos dados EMG em tempo real.

A aquisição do sinal foi realizada por meio da leitura de um pino analógico, com frequência de amostragem nominal configurada em 1 kHz. O sinal bruto obtido passa por uma cadeia de processamento composta por: um filtro passa-alta (frequência de corte de aproximadamente 20 Hz), utilizado para atenuar deriva de linha de base e artefatos de movimento; um filtro *notch* centrado em 60 Hz, utilizado para atenuar a interferência da rede elétrica; retificação do sinal; e, por fim, um filtro digital do tipo média móvel, com janela de tamanho fixo de 10 amostras, responsável por suavizar o envelope resultante.

Os dados adquiridos (sinal bruto e sinal filtrado) são transmitidos ao *frontend* em tempo real por meio de uma conexão *WebSocket* persistente, evitando o custo de estabelecer uma nova conexão a cada amostra. O controle da aquisição continua sendo realizado por rotas HTTP específicas para iniciar e interromper a captura, garantindo sincronização entre o *frontend* e o *backend*. A confiabilidade dessa comunicação em tempo real foi avaliada separadamente (Subseção D e Resultados). Todo processamento pode também ser observado no fluxograma apresentado na Fig. 1.

A interface gráfica do usuário foi implementada como uma aplicação *web*, utilizando HTML, CSS e JavaScript, executadas em um navegador padrão, sem a necessidade de instalação de *software* adicional. O JavaScript foi responsável pela comunicação em tempo real com o ESP32 por meio da conexão *WebSocket*, pelo envio dos comandos de início e parada da aquisição por requisições HTTP, e pela atualização da interface em tempo real. O sinal EMG foi representado graficamente utilizando o elemento *canvas*, permitindo a visualização simultânea do sinal bruto e do sinal filtrado.

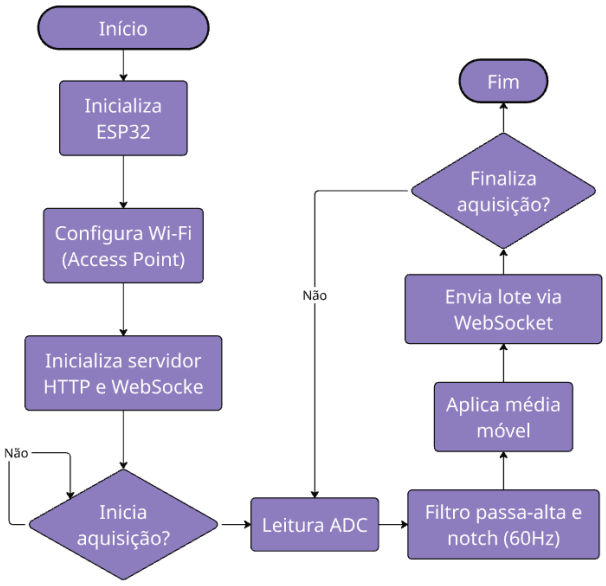


Figura 1. Fluxograma geral do processamento.

No *backend*, foram utilizadas bibliotecas nativas e de terceiros, incluindo:

- WiFi.h, para configuração do ESP32 como ponto de acesso;
- WebServer.h, para implementação do servidor HTTP embarcado;
- WebSockets (Links2004), para implementação do servidor *WebSocket* utilizado na transmissão dos dados em tempo real;
- LittleFS, para armazenamento e disponibilização dos arquivos da interface *web* no sistema de arquivos internos do ESP32.

A interface gráfica desenvolvida apresenta os principais elementos de interação com o sistema. Suas funcionalidades principais incluem:

- Tela de visualização em tempo real do sinal EMG;
- Exibição de métricas instantâneas, como amplitude do sinal, frequência de amostragem e tempo de aquisição;
- Controles para início e parada da captura dos dados;
- Registro e armazenamento das amostras em arquivos no formato CSV;
- Histórico de sessões com opção de *download* dos arquivos gerados.

O código-fonte do firmware e da interface *web* utilizados neste trabalho está disponível publicamente em: https://github.com/emyHorrana/EMG_FACIAL.

D. Captação e validação da comunicação

Os testes experimentais foram conduzidos com os próprios autores do trabalho, totalizando três participantes (duas mulheres e um homem). Por se tratar ainda de uma fase de testes preliminares do protótipo, conduzida apenas com integrantes da equipe de pesquisa, as coletas não passaram por aprovação de comitê de ética; todos os participantes forneceram consentimento informado para a realização dos testes.

As coletas foram realizadas em ambiente com temperatura controlada a 24 °C, em três períodos distintos do dia (manhã, tarde e noite), com cada participante realizando três sessões de aquisição, totalizando nove sessões. O eletrodo de referência foi posicionado na testa. Os eletrodos de sinal foram posicionados na face lateral do osso zigomático, próximo à sutura zigomaticotemporal, com inserção na pele e na fáscia do ângulo da boca (comissura labial), mantendo uma distância de 2 cm entre os eletrodos vermelho e verde, assim como mostra a figura. A Fig. 2 mostra o posicionamento dos eletrodos durante os testes.

Cada sessão seguiu um protocolo padronizado de quatro fases, com duração de 3 segundos cada, conduzidas mediante instruções:

1) *Repouso/neutro*: rosto relaxado, sem falar, cabeça e pescoço imóveis;
2) *Sorriso habitual*: solicitou-se ao participante que sorrisse levemente, buscando mexer o mínimo possível do rosto;
3) *Repouso*;
4) *Sorriso máximo*: solicitou-se ao participante que fizesse o sorriso mais forte possível, sem provocar dor.

Além dos testes de detecção do sinal EMG, foi avaliada a confiabilidade da comunicação em tempo real entre o ESP32 e a interface *web*. Para isso, um participante utilizou o sensor EMG posicionado na face, realizando contrações musculares, em seis sessões de captura. Cada amostra enviada recebe um número sequencial, o que permite identificar no *frontend* se alguma amostra se perdeu no caminho. Nessas sessões, foi medida a taxa de perda de amostras e a frequência real de transmissão, comparada à taxa nominal de 1 kHz configurada no *firmware*.

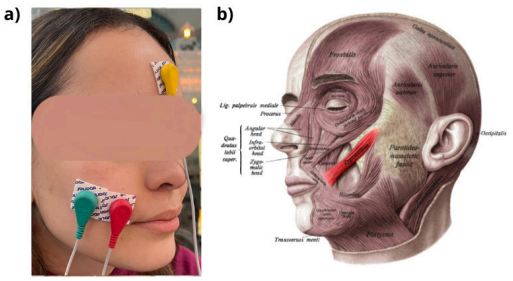


Figura 2. Posicionamento dos eletrodos faciais. a) durante os testes b) anatomia facial destacando o músculo zigomático.

III. Resultados

O estudo teórico sobre eletromiografia mostrou que eletrodos de superfície Ag/AgCl são os mais usados por serem simples de aplicar, não invasivos e produzirem sinal de boa qualidade [6]–[9]. Eletrodos intramusculares, inseridos com agulha, sofrem menos com ruído, mas incomodam mais o usuário, o que os torna menos indicados para estudos com consumidores [8]. Usar muitos eletrodos, como nos esquemas de Fridlund (22 eletrodos) e Kuramoto (8 eletrodos), pode melhorar a qualidade do sinal, mas aumenta o desconforto e a chance de ruído por movimento, impedância da pele e interferência eletromagnética [6]–[8]. Por isso, a literatura destaca a importância de filtrar, amplificar e posicionar bem os eletrodos para reduzir esses problemas [6]–[8]. Hamedi et al. [13] mostraram que é possível classificar gestos faciais por EMG com até 100% de acerto em conjuntos pequenos, o que reforça o potencial da técnica.

O protótipo (Fig. 3) usa um módulo de aquisição EMG, um diodo Zener e um resistor para proteger a entrada do ESP32, além de baterias para alimentar o circuito. O circuito limita a tensão de entrada à faixa de operação do ADC (0 a 3,3V), que só lê valores positivos – por isso, apenas a parte positiva do sinal EMG é captada.

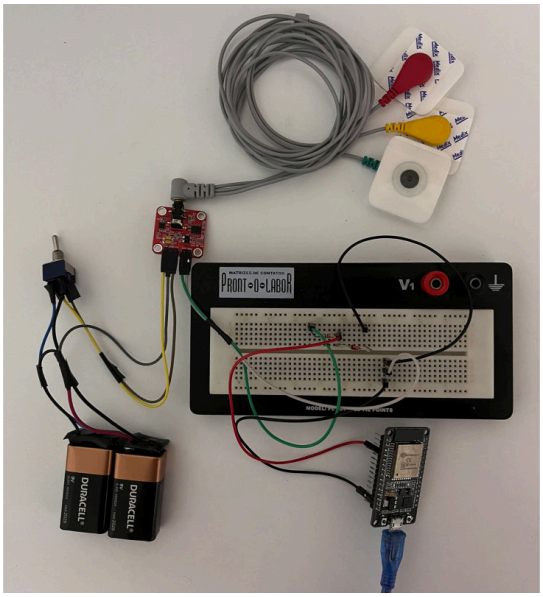


Figura 3. Protótipo físico do sistema.

A Fig. 4 mostra o sinal EMG de quatro sessões de captação, comparando dois participantes (B e C) em dois períodos do dia (manhã e noite), com o sinal bruto em azul e o filtrado em laranja. Os picos indicam contração muscular; os trechos baixos indicam repouso. A Tabela I resume o pico de todas as nove sessões realizadas.

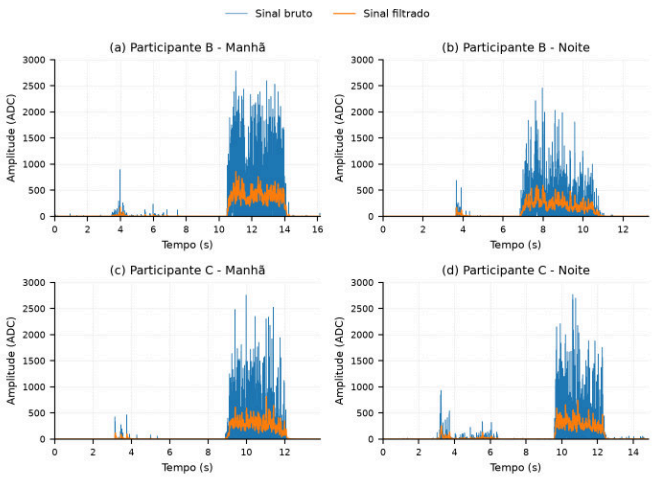


Figura 4. Sinal EMG bruto e filtrado em quatro sessões: participantes B e C, nos períodos da manhã e da noite.

As durações da Tabela I ficaram acima dos 3 segundos previstos no protocolo para cada expressão. O protocolo foi explicado ao participante antes do início da sessão, mas durante a captação não havia nenhum comando externo: cada um contava mentalmente o tempo e alternava a expressão por conta própria, o que naturalmente gera variação em relação ao tempo previsto. Além da duração, os dados também mostram variação entre participantes e entre períodos do dia: o participante B teve picos altos e estáveis (593 a 857), enquanto

Tabela I
RESUMO DAS NOVE SESSÕES DE CAPTAÇÃO. "FAIXA" INDICA O PERCENTUAL DA FAIXA DINÂMICA DO ADC (0–4095) OCUPADO PELO PICO DO SINAL BRUTO.

Sessão	Duração (s)	Pico filt.	Pico bruto	Faixa (%)
A – Manhã	11,89	738	2256	55,1
A – Tarde	14,00	331	1337	32,6
A – Noite	12,06	601	1878	45,9
B – Manhã	16,17	857	2783	68,0
B – Tarde	11,17	824	2846	69,5
B – Noite	13,28	593	2457	60,0
C – Manhã	13,86	811	2757	67,3
C – Tarde	13,25	744	2647	64,6
C – Noite	14,86	747	2768	67,6
Média	13,40	694,0	2414,3	58,9

o participante A teve queda forte à tarde (331) frente à manhã (738) e à noite (601). A manhã, no geral, teve os picos mais altos e mais parecidos entre si (738 a 857); a tarde teve a maior variação (331 a 824). Essa diferença parece estar ligada ao contato do eletrodo em cada sessão ou ao horário em si. O único homem da amostra (B), teve pico médio de 758,0 nas suas três sessões, contra 662,0 das participantes A e C (mulheres); com apenas um homem na amostra, essa diferença não pode ser tratada como uma diferença real de sexo.

O sinal foi lido pelo ADC de 12 bits do ESP32 (0 a 4095), convertido em tensão por [11]:

$$V_{\text{real}} = \frac{\text{ADC obtido}}{4095} \times 3,3 \text{ V}$$

O sinal é sempre positivo, pois o circuito só mede a magnitude da ativação. Nas nove sessões, o pico do sinal bruto variou entre 1337 e 2846 (média 2414,3), entre $1,08$ V e $2,29$ V na entrada do ADC – cerca de 59% da faixa disponível em média, com margem para não saturar. Em repouso, o sinal ficou praticamente em zero. A análise foi feita na escala do sistema ADC.

Antes da adoção do WebSocket, o sistema baseado em requisições HTTP periódicas apresentou latência round-trip média de 5,94 ms (n = 285 requisições), porém com perda de 86,5% das amostras (1820 de 2105) devido ao efeito de *aliasing* do *polling*. Após a migração para comunicação via WebSocket, a perda de amostras caiu para 0% nas seis sessões testadas, com frequência efetiva de amostragem entre 329 Hz e 348 Hz (média de 343 Hz). Como a comunicação via WebSocket ocorre por meio de um fluxo contínuo de dados, sem ciclos individuais de requisição-resposta, a métrica de latência round-trip não se aplica da mesma forma a esse novo modelo. Por isso, a perda de amostras e a frequência efetiva de amostragem passaram a ser os principais indicadores de desempenho em tempo real considerados neste trabalho.

A Fig. 5 mostra a interface web para acompanhar o sinal EMG em tempo real, com opções para iniciar/parar a captura, baixar os dados e limpar a visualização.



Figura 5. Interface web desenvolvida para aquisição e visualização em tempo real dos sinais de EMG facial, permitindo monitoramento da atividade muscular e controle da captura de dados.

IV. Discussão

A Fig. 4 e a Tabela I confirmam que o ADC do ESP32 tem resolução suficiente para separar os momentos de contração muscular dos momentos de repouso. A amplitude variou bastante de uma sessão para outra, o que já era esperado e reforça que calibrar esse ganho deveria ser um dos primeiros passos em uma próxima versão do sistema. Os filtros passa-alta, notch e média móvel, descritos anteriormente, conseguiram limpar o ruído de alta frequência e a interferência de 60 Hz sem apagar as contrações musculares do sinal.

Em relação ao *hardware*, os testes em protoboard cumpriram o papel de validar o circuito de aquisição, mas protoboard é mais suscetível a ruído e a mau contato do que uma placa soldada. Fabricar uma PCI dedicada é, portanto, o próximo passo natural antes de qualquer uso fora de um ambiente controlado.

Na primeira versão, baseada em requisições HTTP periódicas, o sistema chegou a perder 86,5% das amostras em um teste preliminar (Seção V) efeito clássico de *aliasing* de *polling*, em que o navegador simplesmente não consegue pedir dados na mesma velocidade em que o ESP32 os gera. Isso também atrapalhava o desenho do gráfico, já que as amostras nem sempre chegavam na ordem em que foram geradas. A solução foi trocar o HTTP por WebSocket e numerar cada amostra com um número de sequência gerado no próprio ESP32. Essa abordagem permitiu resolver simultaneamente dois problemas: a identificação de perdas de amostras durante a transmissão e a reconstrução correta dos dados na aplicação receptora. Dessa forma, mesmo na presença de atrasos ou variações na rede, tornou-se possível detectar falhas de transmissão e reorganizar as amostras de acordo com sua ordem original de aquisição.

É esperado que haja variabilidade entre indivíduos e entre períodos do dia na amplitude do sinal EMG, o que aparece nos picos da Tabela I. Fatores como a fadiga muscular ao longo do dia podem influenciar essa variação. Ainda assim, com apenas três participantes, esses dados não são suficientes para confirmar um padrão. Seria necessário ampliar o número de pessoas para validar se essa variabilidade se mantém ou segue alguma tendência mais clara relacionada ao horário.

Por fim, vale registrar as limitações deste trabalho. O sistema usa um único canal de EMG no músculo zigomático maior, suficiente para detectar contração muscular, mas não para inferir estados emocionais complexos, isso exigiria vários músculos monitorados ao mesmo tempo. Os testes também foram feitos pelos próprios autores, sem participantes externos, portanto não servem como validação clínica ou comercial, apenas para estudos preliminares. E os filtros aplicados, apesar de terem funcionado bem neste teste, não foram caracterizados formalmente em bancada, não sabe-se, por exemplo, como se comportariam diante de um artefato de movimento mais forte ou de uma rede elétrica mais ruidosa do que a usada nesse trabalho.

V. Conclusão

Este trabalho apresentou o estudo e desenvolvimento preliminar de um protótipo de baixo custo para aquisição de sinais de eletromiografia facial do músculo zigomático maior, usando eletrodos de superfície e uma interface *web* para processamento e visualização em tempo real. Os testes foram conduzidos pelos próprios autores, em nove sessões (três participantes, três períodos do dia), seguindo um protocolo de repouso, sorriso habitual, repouso e sorriso máximo.

Os resultados mostram que o sistema distingue com clareza os períodos de contração muscular e de repouso, mas a ativação isolada do zigomático maior não permite inferir o estado emocional do participante. Nesse estágio, o protótipo deve ser entendido como um sistema de monitoramento de ativação muscular de canal único, e não como um instrumento de mensuração de estados afetivos.

Do lado técnico, a arquitetura baseada no ESP32 e na interface *web* se mostrou viável para aquisição, filtragem e visualização do sinal em tempo real, incluindo o *biofeedback* visual e o download dos dados capturados durante os testes. A migração da comunicação de HTTP para WebSocket eliminou a perda de amostras observada na primeira versão do sistema, e a combinação dos filtros passa-alta, notch e média móvel se mostrou adequada ao sinal captado nos testes realizados.

Como próximos passos, destacam-se a calibração do ganho do estágio analógico, a fabricação de uma placa de circuito impresso e a validação com uma amostra maior e independente dos autores, incluindo comparação com um sistema EMG de referência, sensibilidade a ruído e variações de posicionamento dos eletrodos, além da influência de diferentes condições experimentais e características individuais sobre o sinal registrado. A integração com técnicas complementares, como visão computacional e outros indicadores fisiológicos, também pode contribuir para uma análise mais ampla das respostas dos participantes, desde que acompanhada de protocolos experimentais adequados e validação metodológica, inclusive em aplicações futuras de neurociência do consumidor.

Agradecimentos

Os autores agradecem à FINATEL – Fundação Instituto Nacional de Telecomunicações pelo apoio institucional e infraestrutura disponibilizada, assim como ao Laboratório *eHealth Innovation Center* pelo suporte técnico e ambiente de pesquisa que possibilitaram a realização deste trabalho.

Estendem seus agradecimetos ao Núcleo de Estudos em Neurociência do Consumidor (NENC) pelo incentivo à pesquisa e pelas contribuições científicas que auxiliaram no desenvolvimento do projeto.

REFERÊNCIAS

[1] L. Chiodelli, "A aplicação de métodos de neuromarketing: uma revisão de literatura," Trabalho de Conclusão de Curso (Bacharelado em Administração), Fac. de Ciências Administrativas e Contábeis, Univ. de Passo Fundo, Passo Fundo, Brazil, 2021.

[2] J. A. de Oliveira Rodrigues, "O eye tracking aplicado ao marketing: o protótipo e2m," Dissertação de Mestrado (Eng. Informática), Inst. Sup. Eng. do Porto, Porto, Portugal, 2019.

[3] B. N. de Andrade and P. S. Boggio, "Respostas emocionais implícitas no julgamento da atratividade facial em faces com diferentes níveis de maquiagem," in Proc. XV Jornada de Iniciação Científica e IX Mostra de Iniciação Tecnológica, Univ. Presbiteriana Mackenzie, São Paulo, Brazil, 2019.

[4] G. L. Libralon, "Modelagem computacional para reconhecimento de emoções baseada na análise facial," Tese de Doutorado (Ciência da Computação e Matemática Computacional), Inst. de Ciênc. Matemáticas e de Comput. (ICMC), Univ. de São Paulo (USP), São Carlos, Brazil, 2014.

[5] Advancer Technologies, "Muscle sensor v3 – datasheet," 2014. [Online]. Available: https://www.eletrogate.com/modulo-sensor-eletromiografico-de-sinal-muscular-emg. [Accessed: Nov. 10, 2025].

[6] N. Mueller, V. Trentzsch, R. Grassme, O. Guntinas-Lichius, G. F. Volk, and C. Anders, "High-resolution electromyographic activities of facial muscles during mimic movement in healthy adults: a prospective observational study," *Frontiers in Human Neuroscience*, vol. 16, 2022, Art. no. 1029415, doi: 10.3389/fnhum.2022.1029415.

[7] R. F. Bolzan and V. D. Reis, "Captação do sinal eletromiográfico utilizando filtros e amplificadores," Monografia de Conclusão de Curso (Bacharelado em Engenharia Elétrica), Faculdade Doctum, Juiz de Fora, Brazil, 2019.

[8] J. de M. Ocarino, P. L. P. da Silva, D. V. Vaz, C. F. de Aquino, R. S. Brício, and S. T. da Fonseca, "Eletromiografia: interpretação e aplicações nas ciências da reabilitação," *Fisioterapia Brasil*, vol. 6, no. 4, pp. 305–310, Jul./Aug. 2005.

[9] M. R. Lima, J. J. A. Mendes Júnior, and D. P. Campos, "Recognition of facial patterns using surface electromyography—a preliminary study," *IFMBE Proceedings*, vol. 83. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-70601-2_300. [Accessed: Feb. 2026].

[10] PlatformIO, "What is PlatformIO?," PlatformIO Documentation, PlatformIO Labs. [Online]. Available: https://docs.platformio.org/en/latest/what-is-platformio.html. [Accessed: Feb. 13, 2026].

[11] Sobotta (modified by was_a_bee), "File: Sobo 1909 260.png," Public Domain. [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=30438326. [Accessed: Feb. 13, 2026].

[12] Espressif Systems, "ADC (Analog-to-Digital Converter) — ESP32 Programming Guide," in *ESP-IDF Programming Guide*. [Online]. Available: https://docs.espressif.com/projects/esp-idf/en/latest/esp32/api-reference/peripherals/adc/index.html. [Accessed: Feb. 13, 2026].

[13] M. Hamedi, S.-H. Salleh, T. S. Tan, K. Ismail, J. Ali, C. Dee-Uam, C. Pavaganun, and P. P. Yupapin, "Human facial neural activities and gesture recognition for machine-interfacing applications," *International Journal of Nanomedicine*, vol. 6, pp. 3461–3472, 2011, doi: 10.2147/IJN.S26619.

Simulation-Based Analog Demodulation for Electrical Impedance Tomography Using Rectification and Zero-Crossing Detection

1st Jéssica Costa e Silva
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
silva.jessica@ufabc.edu.br

2nd Vinícius Zacheo de Goes
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
v.goes@aluno.ufabc.edu.br

3rd Mirela de Oliveira Tomazini
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
mirela.tomazini@ufabc.edu.br

4th Gustavo Pinheiro
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
gustavo.pinheiro@ufabc.edu.br

5th Thais H. Samed e Sousa
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
samed.thais@ufabc.edu.br

6th Erick D. León Bueno de Camargo
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
erick.leon@ufabc.edu.br

***Abstract*—Electrical Impedance Tomography (EIT) is a non-invasive imaging technique based on the estimation of the electrical impedance distribution of a medium from low-amplitude alternating current excitation. However, conventional EIT systems require high-speed signal digitization and extensive digital processing to extract amplitude and phase information, increasing computational complexity, especially in multifrequency applications. This work presents an analog architecture for amplitude and phase acquisition applied to EIT, aiming to reduce the processing burden associated with digital demodulation. Amplitude acquisition was implemented using a precision rectifier based on a super-diode configuration, while phase measurement was performed using zero-crossing detectors and an XOR logic gate. The circuits were evaluated exclusively through LTspice simulations considering frequencies from 1 kHz to 1 MHz and signal amplitudes ranging from 10 mV to 5 V. The amplitude detector achieved residual errors below 1% after calibration for most operating conditions up to 250 kHz and below 3% at 500 kHz. The phase detection circuit provided accurate measurements with errors typically below 1 degree for frequencies up to 125 kHz; however, it exhibited significant phase errors and complete operation failures at 500 kHz and above for low-amplitude signals (10 mV). The simulation results indicate that the proposed architecture is a promising analog front-end capable of reducing digital data processing requirements, though future hardware validation is required to confirm EIT-compatible measurement accuracy under physical conditions.**

***Index Terms*—Analog demodulation, precision rectifier, super diode, comparator circuits, zero-crossing detection, electrical impedance tomography.**

I. Introduction

Electrical Impedance Tomography (EIT) is a non-invasive imaging technique that estimates the electrical impedance distribution of a domain by applying low-amplitude alternating currents and measuring the resulting electrical potentials on the surface of the analyzed medium [1], [2]. Due to the absence of ionizing radiation, low cost, and real-time imaging capability, EIT has found applications in areas such as pulmonary monitoring, brain imaging, and tumor detection [2], [3].

In EIT, an alternating current is injected into the medium through electrodes positioned on its surface. The resulting electrical potentials are measured and used to calculate the electrical impedance of the medium according to (1), which depends on the amplitude and phase information of the measured electrical signals [4], [5].

$$Z = \frac{|V|e^{j\theta_V}}{|I|e^{j\theta_I}} = \frac{|V|}{|I|}e^{j(\theta_V - \theta_I)} \tag{1}$$

where $|V|$ represents the measured voltage amplitude, $|I|$ represents the amplitude of the applied current, θ_V represents the phase of the measured voltage, and θ_I represents the phase of the applied current.

Since impedance is obtained from the amplitude and phase of voltage and current signals, errors in these quantities directly affect tomographic reconstruction [6]–[8]. Therefore, data acquisition systems for EIT must exhibit low measurement error to minimize the influence of instrumentation on impedance estimation and on the quality of the reconstructed images [1], [9].

Most EIT systems determine amplitude and phase by digitizing the acquired signals and estimating these quantities through digital signal processing techniques, including quadrature demodulation and digital lock-in amplification [6], [10]. Although these approaches provide high accuracy and flexibility, they require high-speed analog-to-digital converters and considerable computational resources, especially in multichannel and multifrequency systems.

This work was supported by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001.

One strategy to reduce this computational burden is the use of analog demodulation techniques, allowing amplitude and phase information to be extracted before analog-to-digital conversion. In this approach, the system operates with continuous (DC) levels, significantly reducing the amount of processed data and enabling systems with a larger number of acquisition channels and operation at higher frequencies.

Increasing the number of channels enables the use of a larger number of electrodes and, consequently, the acquisition of a greater set of independent measurements, which is directly related to improvements in spatial resolution and reconstructed image quality [11]. Similarly, operation at higher frequencies allows the exploration of tissue dielectric properties through Multifrequency Electrical Impedance Tomography (mfEIT), providing additional information regarding tissue composition and physiological characteristics [12]. Nevertheless, measurements in the megahertz frequency range remain challenging due to the requirements imposed on acquisition and signal-processing circuits [13].

Several acquisition architectures have been proposed to address these challenges. Wu *et al.* reviewed recent EIT instrumentation and showed that digital acquisition combined with synchronous demodulation remains the dominant approach, while emphasizing the growing interest in integrated analog front-ends capable of reducing data throughput and processing complexity. However, the authors also highlighted that trade-offs between bandwidth, accuracy, power consumption, and hardware complexity remain major challenges, particularly for multichannel and multifrequency EIT systems [20]. Likewise, Rezaee-Dehsorkh *et al.* proposed a low-power bioimpedance measurement system employing I/Q demodulation integrated with the analog-to-digital converter to reduce power and silicon area. Although effective, the proposed architecture targets integrated low-power bioimpedance sensing operating at 1 kHz rather than wideband EIT applications [21]. More recently, Nwokoye and Triantis analyzed synchronous demodulation architectures and proposed a digital phase-error correction technique to improve measurement accuracy. Their work demonstrates that synchronous detection remains the preferred architecture for precise bioimpedance measurements but also highlights that phase errors become increasingly significant at higher operating frequencies [9]. These studies indicate that there is still a need for analog acquisition architectures capable of providing accurate amplitude and phase estimation over a broad frequency range while reducing digital processing requirements and maintaining a relatively simple hardware implementation.

In this context, this work presents an analog circuit for amplitude and phase acquisition applied to Electrical Impedance Tomography. The proposed solution performs signal demodulation in the analog domain using precision rectification and zero-crossing detection, providing amplitude and phase information prior to analog-to-digital conversion. The circuit was designed to estimate amplitude and phase for signals from 1 kHz to 1 MHz and was evaluated in terms of amplitude and phase estimation accuracy, which are key parameters for impedance determination and image reconstruction quality. Unlike digital quadrature demodulation, which requires high-speed ADC sampling and continuous digital multiplication and filtering for every acquisition cycle, the proposed analog front-end extracts amplitude and phase as quasi-static DC levels, allowing these values to be digitized with low-speed, low-resolution ADCs and reducing the computational load per channel. Practical implementation also requires careful physical layout to minimize parasitic capacitances and noise coupling, which will be addressed in future experimental validation.

II. Materials and Methods

Amplitude acquisition was performed using a precision rectifier based on the super diode configuration, while phase measurement was implemented using zero-crossing detectors and an XOR logic gate. Both approaches are described in the following subsections.

The proposed circuits were designed and evaluated through simulations performed in LTspice. In both approaches, the AD826 operational amplifier was used and powered by supplied with ± 9 V rails, selected due to their high slew rate, wide bandwidth, and ability to operate with low-amplitude signals [16]. The circuits were simulated without an external load connected to the output terminals, corresponding to an open-circuit condition. The voltage was measured directly at the OUT pin of the XOR gate and at the rectifier output, allowing the intrinsic performance of the proposed architecture to be evaluated prior to considering specific Analog-to-Digital Converter input impedances. Low-amplitude signals were intentionally investigated because voltage measurements in EIT systems often occur in the millivolt range.

To quantitatively assess the measurement accuracy, the differences between the reference and measured values were evaluated using the absolute error (AE) and the relative error (RE), calculated according to (2) and (3):

$$AE = |V_{meas} - V_{ref}| \tag{2}$$

$$RE(\%) = \frac{|V_{meas} - V_{ref}|}{V_{ref}} \times 100 \tag{3}$$

where V_{meas} is the measured value and V_{ref} is the reference value.

A. Analog Amplitude Acquisition

Amplitude acquisition was implemented using a precision rectifier, commonly referred to as a super diode. In this topology, the diode is placed inside the feedback loop of an operational amplifier, as shown in Fig. 1. This configuration allows the operational amplifier to compensate for the diode forward voltage drop, enabling the rectification of low-amplitude signals that would not be properly detected by a conventional diode-based rectifier [14].

The output of the circuit corresponds to a DC voltage level proportional to the peak amplitude of the input signal, since

the holding capacitor charges to the peak of each rectified half-cycle and discharges through the resistor between peaks, with the resulting ripple attenuated by the RC low-pass behavior described below.

Fig. 1 presents both the conventional peak detector and the proposed precision rectifier configuration used to eliminate the error introduced by the diode forward voltage [15]. The conventional circuit consists of a diode responsible for rectification, a capacitor used to store the peak value, and a resistor that provides a controlled discharge path for the capacitor. This configuration performs half-wave rectification and has been widely adopted in signal-conditioning applications [14].

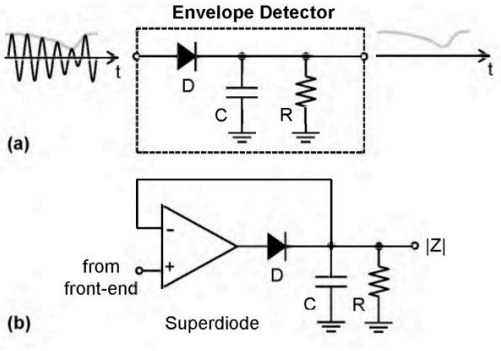


Fig. 1. (a) Conventional amplitude detector and (b) precision rectifier implementation used to eliminate the diode forward voltage drop.

In this work, signal rectification was performed using a BAT54 Schottky diode, chosen for its low forward voltage drop and fast switching characteristics, making it suitable for operation over the frequency range investigated in this work [17]. A 10 kΩ resistor was used as the discharge resistor.

The holding capacitor was dynamically adjusted according to the input frequency. As described by Li and Ferrari [15], the resistor-capacitor network behaves as a low-pass filter, and therefore its cutoff frequency directly affects the ripple and response time of the detected signal. In this work, the capacitor value was calculated such that the cutoff frequency was equal to the input frequency divided by 200. This criterion provided the best trade-off between ripple reduction and response time during simulation.

The capacitor value was calculated according to:

$$C = \frac{1}{2\pi R f_c} \tag{4}$$

where R is the discharge resistor and f_c is the desired cutoff frequency.

For the calculation of the DC output voltage, the average output voltage over the 30th to 40th periods of the waveform was used, excluding the initial cycles to minimize transient effects. The transient simulation step was set to 1 μs.

Fig. 2 presents the complete circuit implemented and simulated in LTspice.

Furthermore, to assess robustness against component tolerances, a Monte Carlo analysis (100 runs, 10% capacitor and

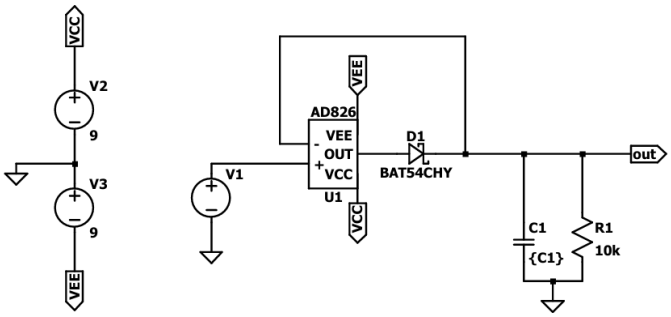


Fig. 2. LTspice schematic of the precision rectifier circuit.

1% resistor tolerance) and a corner (worst-case) analysis were performed.

B. Analog Phase Acquisition

Phase acquisition was performed using zero-crossing detectors combined with an XOR logic gate. Zero-crossing detectors are circuits capable of identifying the instant at which an alternating signal crosses a reference voltage, typically 0 V, causing the output state of the comparator to change at each crossing event [18].

In practice, an operational amplifier configured as a comparator produces a high output level when the input voltage is positive and a low output level when the input voltage is negative [18]. As a result, a square wave synchronized with the zero crossings of the original signal is obtained, as illustrated in Fig. 3.

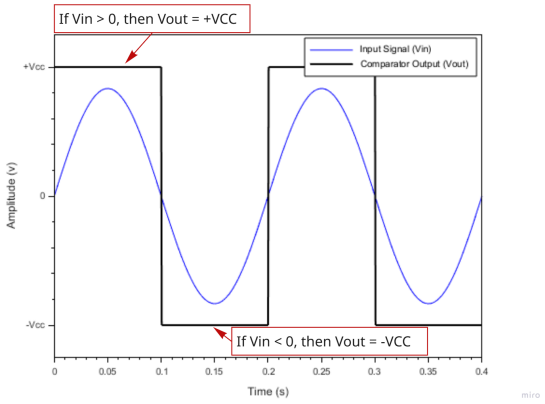


Fig. 3. Behavior of a comparator circuit implemented using an operational amplifier.

In this work, zero-crossing detectors were used to generate square waves synchronized with the sinusoidal signals under analysis. Phase information was then obtained from the time difference between the resulting square waves. To facilitate phase measurement, an XOR logic gate was added to the circuit. Since the XOR output remains at a high logic level whenever its inputs are different, the resulting output is a pulse-width modulated (PWM) signal whose duty cycle is

proportional to the phase difference between the analyzed signals [14], as shown in Fig. 4.

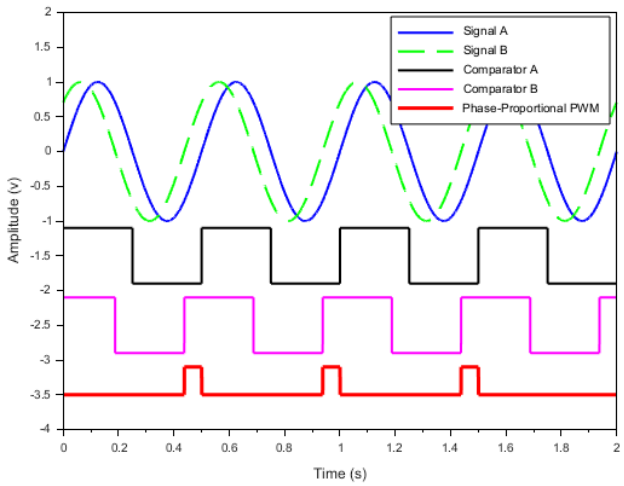


Fig. 4. Operation of the phase detection circuit based on zero-crossing detectors and an XOR gate.

For phase determination, the average output voltage over the interval from the 90th to the 100th period, with transient simulation step set to 1 μs, of the waveform was calculated (V_{avg}). The resulting value was subsequently converted to phase using a proportional relationship, in which the maximum output voltage (V_{max}) corresponds to $180°$. Accordingly, the phase was determined as:

$$\phi = \frac{V_{\text{avg}} \times 180°}{V_{\text{max}}} \tag{5}$$

In EIT applications, the phase shift introduced by biological tissues under low-amplitude alternating excitation is typically limited to a few degrees, remaining far from the 180° range in which the XOR-based detection method becomes ambiguous; therefore, this limitation is not expected to affect the intended clinical application of the proposed circuit.

The block diagram of the proposed phase measurement circuit is presented in Fig. 5.

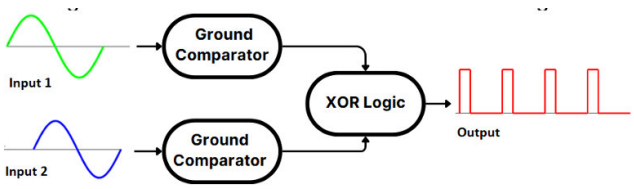


Fig. 5. Block diagram of the phase measurement circuit.

Fig. 6 presents the complete architecture of the phase measurement circuit. The performance of the zero-crossing detectors was evaluated using the AD826 operational amplifier.

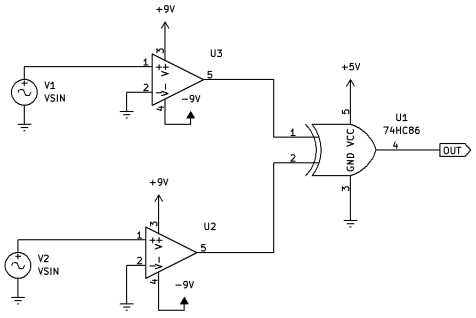


Fig. 6. General schematic of the phase measurement circuit.

III. Results and Discussion

A. Precision Rectifier Results

The precision rectifier circuit was evaluated through simulations considering different excitation amplitudes and frequencies. Table I presents the results obtained for the simulated sinusoidal signals. The gain values presented in the table were calculated according to:

$$Gain = \frac{V_{meas}}{V_{ref}} \tag{6}$$

where V_{meas} is the measured value and V_{ref} is the reference value.

TABLE I
SIMULATION RESULTS FOR DIFFERENT AMPLITUDE AND FREQUENCY CONFIGURATIONS.

Reference Amp. (mV)	Frequency (Hz)	Measured Amp. (mV)	AE (mV)	RE (%)	Gain (%)
10	1k	9.90	0.10	1.00	99.00
100	1k	98.38	1.62	1.62	98.38
1000	1k	986.71	13.29	1.33	98.67
5000	1k	4927.71	72.29	1.45	98.55
10	10k	10.22	0.22	2.20	102.20
100	10k	100.13	0.13	0.13	100.13
1000	10k	999.46	0.54	0.05	99.95
5000	10k	4910.61	89.39	1.79	98.21
10	50k	8.47	1.53	15.30	84.70
100	50k	98.78	1.22	1.22	98.78
1000	50k	997.44	2.56	0.26	99.74
5000	50k	4946.28	53.72	1.07	98.93
10	100k	5.51	4.49	44.90	55.10
100	100k	90.86	9.14	9.14	90.86
1000	100k	986.24	13.76	1.38	98.62
5000	100k	4983.57	16.43	0.33	99.67
10	250k	3.63	6.37	63.70	36.30
100	250k	71.32	28.68	28.68	71.32
1000	250k	959.05	40.95	4.10	95.90
5000	250k	4945.37	54.63	1.09	98.91
10	500k	1.09	8.91	89.10	10.90
100	500k	54.14	45.86	45.86	54.14
1000	500k	911.82	88.18	8.82	91.18
5000	500k	4875.50	124.50	2.49	97.51
10	1M	0.94	9.06	90.60	9.40
100	1M	15.48	84.52	84.52	15.48
1000	1M	841.05	158.95	15.90	84.11
5000	1M	4738.90	261.10	5.22	94.78

These results demonstrate that the super diode configuration successfully compensated for the BAT54 forward voltage drop, enabling the detection of signals with amplitudes in the millivolt range.

Between 50 kHz and 250 kHz, a reduction in measurement accuracy was observed for amplitudes below 1 V. In contrast, the circuit maintained satisfactory performance for 1 V and 5 V signals throughout the entire simulated frequency range, from 1 kHz to 1 MHz.

Although no clear pattern was observed in either the absolute or relative errors, analysis of the simulation results revealed that a subset of the evaluated operating conditions exhibited a nearly constant gain, indicating the presence of a systematic measurement error rather than random deviations. This behavior was observed for all tested amplitudes at 1 kHz and 10 kHz, as well as for amplitudes of 1 V and 5 V at frequencies of 50 kHz, 100 kHz, 250 kHz, and 500 kHz. The average gain obtained from these measurements was

$$G_{avg} = 98.95\% \tag{7}$$

which corresponds to a calibration factor of

$$K = 0.9895 \tag{8}$$

The relatively small variation around this average value suggests that the detector introduces an approximately constant gain error within this operating range. Thus, the corrected amplitude for the main dataset was calculated according to:

$$V_{corr} = \frac{V_{meas}}{K} \tag{9}$$

where V_{meas} is the measured amplitude and K is the average gain factor obtained experimentally.

Table II presents the corrected amplitudes and the residual errors after calibration. The results show an improvement in measurement accuracy. After calibration, the residual error remained below 1% for nearly all operating conditions up to 250 kHz and below 3% at 500 kHz.

To verify the validity of this calibration factor, it was evaluated on a separate set of simulated amplitudes and frequencies not used in its derivation (Table III). For these independent points, the residual error remained below 3%, confirming that a simple linear calibration can be effectively applied to compensate for this systematic deviation.

Finally, a Monte Carlo analysis was performed under three operating conditions. For typical operation (1 V at 100 kHz), the circuit showed stability, with an amplitude error variation below 0.5% standard deviation (SD) and a maximum error of 1.1% under worst-case component variations, indicating that the calibration remains valid under realistic component variation. Under conditions where nominal precision begins to degrade (10 mV at 250 kHz), the circuit became more sensitive to component mismatch, with an error variation of approximately 9% (SD) and a maximum error of 90%, corresponding to an output of 1.0 mV. At the breakdown condition (10 mV at 500 kHz), the output collapsed to 0.088 mV, with a maximum error exceeding 99%, confirming the operational limit of the architecture.

TABLE II
AMPLITUDE MEASUREMENTS AFTER GAIN CALIBRATION FOR OPERATING CONDITIONS EXHIBITING APPROXIMATELY CONSTANT GAIN.

Reference Amp. (mV)	Frequency (Hz)	Corrected Amp. (mV)	RE (%)
10	1k	10.00	0.05
100	1k	99.45	0.55
1000	1k	997.15	0.29
5000	1k	4979.94	0.40
10	10k	10.38	3.80
100	10k	100.96	0.96
1000	10k	1010.03	1.00
5000	10k	4962.78	0.74
1000	50k	1006.27	0.63
5000	50k	5012.99	0.26
1000	100k	992.41	0.76
5000	100k	5041.71	0.83
1000	250k	1009.23	0.92
5000	250k	4989.75	0.20
1000	500k	971.17	2.88
5000	500k	4870.49	2.59

TABLE III
AMPLITUDE MEASUREMENTS AFTER GAIN CALIBRATION FOR NEW DATA

Reference Amp. (mV)	Frequency (Hz)	Corrected Amp. (mV)	RE (%)
50	1k	49.64	0.72
250	10k	253.24	1.29
2000	50k	2021.01	1.05
2500	100k	2514.70	0.58
3000	250k	2933.32	2.22
4000	500k	3881.20	2.96

B. Phase Measurement Circuit Results

The comparator-based method enabled the conversion of the angular difference between two signals into a PWM (Pulse-width modulation) signal proportional to the phase shift. Table IV presents the phase shift values obtained through simulation for 10 mV sinusoidal signals at different frequencies, while Table V presents the corresponding absolute errors.

TABLE IV
PHASE SHIFT MEASURED USING THE AD826 FOR DIFFERENT FREQUENCIES AND NOMINAL PHASE SHIFTS WITH A SIGNAL AMPLITUDE OF 10 MV.

Freq.	**1°**	**2°**	**15°**	**30°**	**45°**	**90°**	**180°**
1 kHz	1.07	2.09	15.08	29.95	45.00	90.00	175.65
10 kHz	0.99	1.80	15.13	30.18	44.97	90.21	178.35
50 kHz	0.97	2.18	15.51	30.00	44.16	90.08	178.05
125 kHz	2.17	2.47	14.49	30.82	46.03	90.34	173.43
250 kHz	0.00	0.00	15.06	30.54	44.58	89.58	106.25
500 kHz	0.00	0.00	0.00	0.00	0.00	0.00	0.00

According to the results, the circuit exhibited errors below 1° in most of the evaluated cases for frequencies up to 125 kHz. From 250 kHz onward, a reduction in measurement accuracy was observed for 10 mV signals. At this frequency, phase shifts of 1° and 2° were not detected, and the measurement

TABLE V
ABSOLUTE PHASE MEASUREMENT ERROR FOR DIFFERENT FREQUENCIES WITH A SIGNAL AMPLITUDE OF 10 MV.

Freq.	1°	2°	15°	30°	45°	90°	180°
1 kHz	0.07	0.09	0.08	0.05	0.00	0.00	4.35
10 kHz	0.01	0.20	0.13	0.18	0.03	0.21	1.65
50 kHz	0.03	0.18	0.51	0.00	0.84	0.08	1.95
125 kHz	1.17	0.47	0.51	0.82	1.03	0.34	6.57
250 kHz	1.00	2.00	0.06	0.54	0.42	0.42	73.75
500 kHz	1.00	2.00	15.00	30.00	45.00	90.00	180.00

of 180° presented an error of 73.75°. At 500 kHz, the phase detector failed to operate for all evaluated phase shifts.

To investigate this limitation, additional tests were performed using amplitudes of 100 mV and 5 V at frequencies of 500 kHz and 1 MHz. Table VI presents the measured phase shifts, while Table VII presents the corresponding absolute errors. The results showed that increasing the signal amplitude restored circuit operation at high frequencies. For an amplitude of 100 mV, most phase shifts exhibited errors below 1°, except for the 180° condition, which presented errors greater than 6° at 500 kHz and 11° at 1 MHz. For an amplitude of 5 V, the errors remained below 1.6° for all evaluated phase shifts.

The phase error below 2° obtained in simulation for this analog circuit is within acceptable limits for bioimpedance instrumentation. According to Nwokoye and Triantis [9], keeping the deviation at this level prevents mixing between the real and imaginary components of the signal. This correct analog separation contributes to clean and reliable data before the signal is sent to the image reconstruction algorithm.

TABLE VI
MEASURED PHASE SHIFT AT 500 KHZ AND 1 MHZ FOR AMPLITUDES OF 100 MV AND 5 V.

Phase Shift	100 mV		5 V	
Nominal	500 kHz	1 MHz	500 kHz	1 MHz
1°	0.00	1.43	1.06	0.83
2°	2.08	2.54	2.27	2.13
15°	15.34	15.36	14.91	15.01
30°	30.04	29.65	30.36	29.79
45°	45.05	44.10	44.97	45.04
90°	90.08	89.76	89.99	89.98
180°	173.07	168.28	178.99	178.42

Based on the overall circuit performance, a limitation associated with the combination of high frequencies and low input amplitudes was identified. Tests performed using the AD826 revealed limitations of the operational amplifier when employed as a comparator. Despite its high bandwidth and slew rate of 350 V/μs, switching delays were observed and became more pronounced for high-frequency, low-amplitude signals. As shown in Fig. 7, these delays caused the comparator outputs to deviate from the ideal square-wave shape presented in Fig. 8. This behavior is associated with saturation recovery time and the internal compensation of the operational amplifier, which was designed for closed-loop operation rather than comparator applications.

TABLE VII
ABSOLUTE PHASE MEASUREMENT ERROR AT 500 KHZ AND 1 MHZ FOR AMPLITUDES OF 100 MV AND 5 V.

Phase Shift	100 mV		5 V	
Nominal	500 kHz	1 MHz	500 kHz	1 MHz
1°	1.00	0.43	0.06	0.17
2°	0.08	0.54	0.27	0.13
15°	0.34	0.36	0.09	0.01
30°	0.04	0.35	0.36	0.21
45°	0.05	0.90	0.03	0.04
90°	0.08	0.24	0.01	0.02
180°	6.93	11.72	1.01	1.58

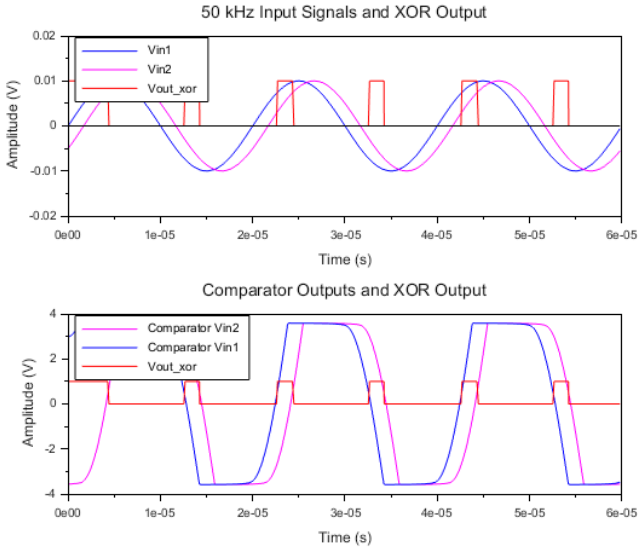


Fig. 7. Comparator output obtained with the AD826 showing switching delay for a 10 mV signal at 50 kHz.

For this application, the presence of switching delay alone would not necessarily prevent phase calculation. Since both signals have the same frequency, the comparators would be expected to introduce similar delays, thereby preserving the time interval between them. The problem arises when the signals have both high frequency and low amplitude. Under these conditions, the comparators can no longer properly follow the transitions of the input signals. Before the comparator output fully reaches one logic level, the input signal has already started transitioning to the opposite state. As a result, the comparator outputs remain at the same logic level, as illustrated in Fig. 9, which shows the simulation results for 500 kHz signals with an amplitude of 10 mV. Consequently, the XOR gate fails to detect the time difference between the signals, and its output remains low, making phase shift determination impossible.

To overcome this limitation, the MAX962ESA was also evaluated [19]. This device is a high-speed comparator specifically designed for applications requiring low propagation delay and fast switching transitions. The component exhibited

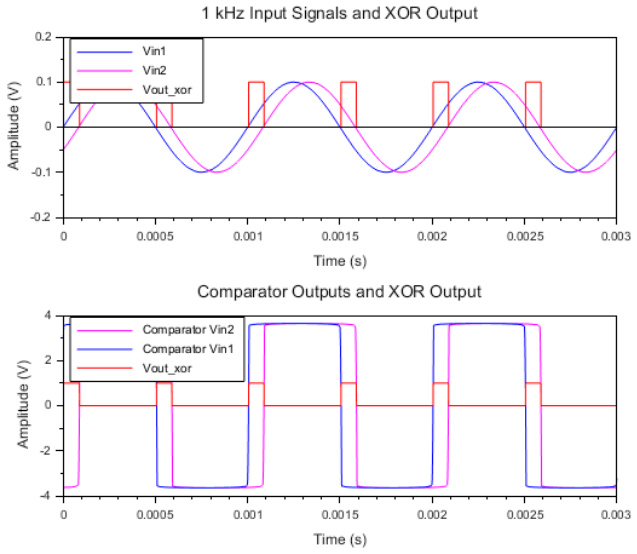


Fig. 8. Ideal comparator response obtained with the AD826 for a 100 mV signal at 1 kHz.

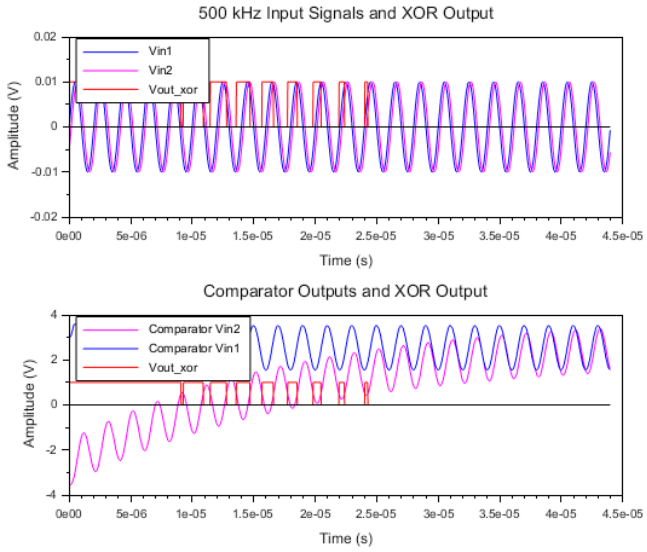


Fig. 9. Comparator output obtained with the AD826, showing that the device is unable to properly track a 10 mV input signal at 500 kHz.

improved dynamic performance during the conducted tests, with reduced distortion at high frequencies and greater stability in zero-crossing detection. However, it did not operate adequately with low-amplitude signals, exhibiting limitations for amplitudes below 300 mV.

IV. Conclusion

This simulation study presented an analog architecture for amplitude and phase acquisition for Electrical Impedance Tomography based on precision rectification and zero-crossing detection. The simulation results demonstrated that the proposed circuits can achieve measurement accuracy compatible with the limits commonly adopted in EIT instrumentation, with amplitude errors below 1% up to 250 kHz after calibration and phase errors typically below 1° for 10 mV signals up to 125 kHz. By extracting amplitude and phase information before analog-to-digital conversion, the proposed approach has the potential to reduce digital processing requirements while maintaining measurement accuracy. These simulated results indicate that the architecture is a promising candidate for implementation in EIT systems.

Future work will focus on the hardware implementation and experimental validation of the proposed architecture under non-ideal conditions. In particular, for low-amplitude signals, a pre-amplification stage at the circuit input will be investigated to mitigate the effects of zero-crossing hysteresis, together with a high-pass filter to eliminate DC offset. The physical circuits are currently under construction, and their experimental performance will be evaluated by comparing the measured amplitude and phase errors with published EIT accuracy limits under matched frequency, load, and signal-level conditions. Following this validation, the architecture will be integrated into a complete EIT system for image reconstruction using experimental phantom measurements.

Artificial intelligence tools were used exclusively to support academic writing, including style revision and improvements in textual fluency. The methodology design, simulation execution, data analysis, and interpretation of the results were carried out entirely by the authors.

Acknowledgment

The authors acknowledge the Medical Devices Laboratory and the Electrical Impedance Tomography Research Group at the Federal University of ABC (UFABC) for their support and infrastructure provided during this research.

References

[1] A. Adler and D. S. Holder (Eds.), *Electrical Impedance Tomography: Methods, History and Applications*. 2nd ed. Boca Raton, FL, USA: CRC Press, 2021. DOI: 10.1201/9780429399886.

[2] A. Adler and A. Boyle, "Electrical impedance tomography: Tissue properties to image measures," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2494–2504, Nov. 2017.

[3] T. F. Becher, M. B. Gawelczyk, and I. Frerichs, "Electrical impedance tomography in critical care medicine," *J. Clin. Med.*, vol. 13, no. 2, pp. 1–21, 2024.

[4] S. Grimnes and O. G. Martinsen, *Bioimpedance and Bioelectricity Basics*, 4th ed. London, U.K.: Academic Press, 2015.

[5] J. G. Webster, *Medical Instrumentation: Application and Design*. 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2009.

[6] R. H. Bayford and A. Demosthenous, "Bioimpedance tomography instrumentation: Design considerations and recent advances," *Sensors*, vol. 23, no. 3, pp. 1182–1204, 2023.

[7] D. C. Barber and B. H. Brown, "Applied potential tomography," *Journal of Physics E: Scientific Instruments*, vol. 17, no. 9, p. 723, 1984.

[8] R. H. Bayford, "Bioimpedance tomography (electrical impedance tomography)," *Annual Review of Biomedical Engineering*, vol. 8, no. 1, pp. 63–91, 2006.

[9] I. I. Nwokoye and I. F. Triantis, "A phase error correction system for bioimpedance measurement circuits," *Applied Sciences*, Basel, vol. 14, no. 12, p. 5202, 2024. DOI: 10.3390/app14125202.

[10] B. H. Brown, "Electrical impedance tomography (EIT): A review," *J. Med. Eng. Technol.*, vol. 27, no. 3, pp. 97–108, 2003.

[11] A. Adler, J. Arnold, R. Bayford, et al., "GREIT: A unified approach to 2D linear EIT reconstruction of lung images," *Physiol. Meas.*, vol. 30, no. 6, pp. S35–S55, 2009.

[12] E. Malone, G. dos Santos Rocha, S. Holder, and A. Arridge, "Multifrequency electrical impedance tomography for biomedical imaging," *Physiol. Meas.*, vol. 45, no. 1, pp. 1–24, 2024.
[13] J. Avery, R. Dowrick, and A. Adler, "Challenges and opportunities in high-frequency electrical impedance tomography," *Sensors*, vol. 24, no. 7, pp. 1–18, 2024.
[14] P. Horowitz and W. Hill, *The Art of Electronics*, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2015.
[15] J. V. Li and G. Ferrari, Eds., *Capacitance Spectroscopy of Semiconductors*. Singapore: Pan Stanford Publishing, 2018.
[16] Analog Devices, "AD826: Wide Bandwidth, High Output Current FET Input Operational Amplifier," Datasheet, Rev. H, Wilmington, MA, USA, 2024.
[17] Nexperia, "BAT54 Series Schottky Barrier Diodes," Datasheet, Rev. 12, Nijmegen, The Netherlands, 2024.
[18] R. Mancini, *Op Amps for Everyone: Design Reference*. 4th ed. Dallas, TX, USA: Texas Instruments, 2009.
[19] Analog Devices, "MAX962 High-Speed, Low-Power Comparators," Datasheet, Rev. 5, 2024.
[20] Y. Wu, F. F. Hanzaee, D. Jiang, R. H. Bayford, and A. Demosthenous, "Electrical Impedance Tomography for Biomedical Applications: Circuits and Systems Review," *IEEE Open Journal of Circuits and Systems*, vol. 2, pp. 380–397, 2021. DOI: 10.1109/OJCAS.2021.3075302.
[21] H. Rezaee-Dehsorkh, N. Ravanshad, A. Shamsaki, M. R. Fakour, and P. Aliparast, "A Low-Power Single-Path Bio-Impedance Measurement System Using an Analog-to-Digital Converter for I/Q Demodulation," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 6, pp. 1129–1137, Dec. 2022. DOI: 10.1109/TBCAS.2022.3213869.

Analysis of Physiological Responses in the Identification of Misinformation: A Scoping Review

Bruno Gomes Cordeiro
Applied Computing Research Group (GPCA)
Federal Institute of Brasília (IFB)
Brasília, Brazil
bruno.g.cordeiro26@gmail.com | ORCID: 0009-0005-4690-4114

Rhaíra Helena Caetano e Souza
Assistive Technology Access and Research Center (CAPTA)
Federal Institute of Brasília (IFB)
Brasília, Brazil
rhaira.souza@ifb.edu.br | ORCID: 0000-0003-0883-8150

Fábio Henrique Monteiro Oliveira
Applied Computing Research Group (GPCA)
Federal Institute of Brasília (IFB)
Brasília, Brazil
fabio.oliveira@ifb.edu.br | ORCID: 0000-0002-0344-5801

Abstract*—Disinformation has reached a global scale driven by generative AI, overcoming the effectiveness of traditional natural language processing (NLP) filters by exploiting vulnerabilities in human cognitive processing. The transition to analyses centered on the message receiver has become imperative. To this end, this work mapped and synthesized the scientific literature on the identification of disinformation through the analysis of Electrodermal Activity (EDA) and eye-tracking. As eligibility criteria, primary articles without temporal restrictions, in English and Portuguese, that address digital deception detection via the aforementioned biosensors were included. The searches were executed in the Scopus, ScienceDirect, IEEE Xplore, PubMed, Web of Science, and SciELO databases. Screening, duplicate removal, and data extraction were performed with the support of the Parsifal platform, correlating biosignals to states of mental load. The unified search resulted in 183 documents, of which 81 were selected for synthesis. The data indicate that the eye-tracking technique identifies patterns of doubt via the rereading of excerpts, while EDA records peaks of sympathetic nervous system activation during exposure to fraudulent information. Detection based on subconscious biosignals presents itself as a viable approach given the limitations of textual methods, highlighting the relevance of investigating multimodal architectures focused on user neurophysiological behavior.

***Keywords*—Disinformation, Electrodermal Activity, Eye-tracking, Scoping review, Biosignals.**

I. Introduction

The way we consume information today has changed drastically, and with that, the spread of fake news on social media has gone from being a technical problem to becoming a crisis of trust. What was once just digital noise now threatens the foundations of institutions. This impact goes beyond the digital realm, directly affecting the basis of decisions made by contemporary society. Disinformation now represents a risk on a global scale, spreading exponentially faster than factual news [1].

In the Brazilian context, informational instability has grown with the popularization of generative Artificial Intelligence (AI). This technology has made the production of misleading content—both in text and image—cheerful and impressively realistic. Recent data shows that the volume of disinformation and *deepfake materials* (videos or images manipulated by AI to simulate reality in an ultra-realistic way) jumped more than 308% in just one year in the country [2], becoming a strategic tool for political manipulation.

Statistical surveys reveal that 93% of the Brazilian population uses some form of social network and approximately 80% of citizens support the creation of legislation to combat the spread of fake news [3]. Despite the high level of concern about the issue, data indicate that 50% of the population declares that they have great difficulty discerning whether digital information is true or fraudulent [4], highlighting a gap in the public's information literacy.

This difficulty, however, is not solely due to deficiencies in digital education, but also to cognitive vulnerabilities inherent in human beings. To understand the root of this scenario, the specialized literature in cognitive psychology categorizes false information into three categories: disinformation, intentionally created to manipulate; misinformation, shared unintentionally, often due to failures in digital literacy; and misinformation, based on real facts but decontextualized to cause harm. In all these scenarios, the user's vulnerability is aggravated by a cognitive limitation of memory known as source monitoring failure.

Due to the high volume and speed of information consumed on social media, the individual's brain often stores the content of a message in long-term memory, but fails to register where, when, or through whom this information was obtained. This results in an inability to later distinguish whether the news was read in a reliable journalistic outlet or in a disseminator of fraud [5].

Faced with fraud that reproduces journalistic language and aesthetics with high fidelity, traditional computational digital security mechanisms based exclusively on Natural Language Processing (NLP) have limitations. As contemporary digital disinformation exploits cognitive biases—such as the Illusory Truth Effect—to increase the perceived credibility of false information, it becomes necessary to broaden the focus of detection. Instead of analyzing only textual content, approaches that also consider the physiological and cognitive responses of individuals during reading can offer complementary information for identifying this type of fraud [6].

The severity of this cognitive vulnerability has been scientifically tested in high-polarity scenarios. Studies focused on skeptical climate claims have revealed that simply repeating misinformation increases the user's belief levels, regardless of their prior knowledge of the subject.

The phenomenon even affects individuals who endorse climate science, demonstrating that processing fluency generated by repetition bypasses analytical judgment [7]. As cognition fails when processing repeated false information, information security should not rely solely on the reader's rationality, which justifies investigating biosignal monitoring (Fig. 1).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

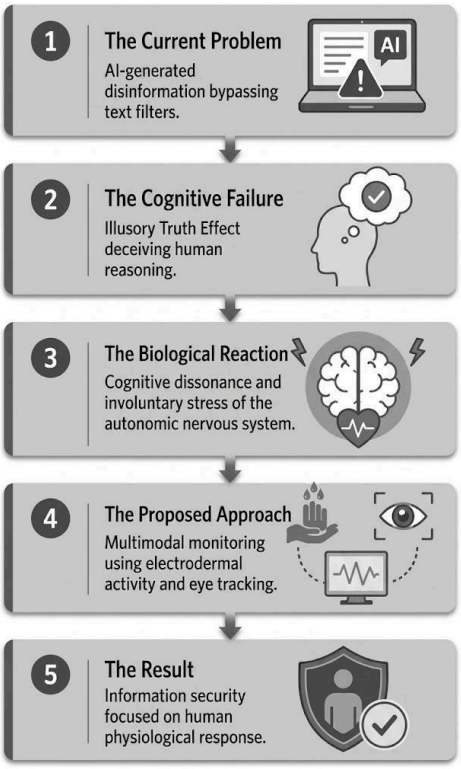


Fig. 1. Conceptual flow of misinformation identification via biosignals.

Electrodermal Activity (EDA) and eye tracking emerge as practical solutions. While EDA monitors skin conductance, eye tracking "reads" mental effort through gaze fixations and pupil dilation. These are new and promising approaches [6, 8]. These biosignals are capable of measuring cognitive dissonance and autonomic stress in the face of fraudulent information involuntarily, operating even before the user makes a conscious decision about the veracity of the content.

Given the above, the objective of this scoping review is to map and synthesize the existing literature on identifying misinformation using physiological and ocular responses. The review seeks to understand the state of the art of these technologies to justify the need for and feasibility of developing new multimodal computational architectures focused on human behavior in the fight against fake news.

To guide the execution of this mapping and achieve the proposed objectives, the protocol in the Parsifal software was structured around five Research Questions: QP1. Which biosignals are most used in the detection of misinformation? QP2. How does eye tracking contribute to identifying cognitive conflict in the face of fake news? QP3. Is there evidence of unconscious truth recognition through biological responses, regardless of the user's belief? QP4. Which elements of the news (text, image, or headline) generate the greatest changes in eye fixation patterns? QP5. How does the time of exposure to information influence the intensity of electrodermal responses (EDA)?

II. Methods

This research was conducted following the guidelines of the PRISMA-ScR extension (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews*), ensuring the appropriate methodological rigor for scoping reviews.

A. Protocol

The internal protocol for this review was developed to ensure the transparency of biometric data extraction applied to the context of digital security and deception detection, structured around the search, screening, and qualitative synthesis steps recommended by the PRISMA-ScR checklist. This protocol has not been previously registered.

B. Eligibility criteria

Primary articles published in scientific conferences and journals, in English and Portuguese, were included, with no initial date restriction. The main inclusion criterion required that studies explicitly address the detection of digital deception, misinformation, or cognitive conflict using biosensors focused on receiver analysis, specifically through EDA and eye tracking. Articles focused exclusively on traditional polygraphs for criminal interrogations, or that used only mathematical NLP models without the participation of human volunteers, were excluded from the scope.

C. Sources of information

The searches were initially conducted across three global scientific and technical databases (Scopus, ScienceDirect, and IEEE Xplore Digital Library). Subsequently, the search was expanded to include the reference biomedical database PubMed, the global multidisciplinary database Web of Science, and the regional Latin American database SciELO. The searches were carried out comprehensively, and the final data retrieval was performed on April 15, 2026.

D. Search strategy

To ensure reproducibility and maximum comprehensiveness of the mapping, the complete search strategy was structured in multiple Boolean blocks. The *string* was constructed exclusively with English terms, since international databases standardize metadata indexing in this language, allowing the retrieval of articles regardless of whether the full text is in Portuguese. To facilitate reading, Table I presents the representation of the constructs by groups of terms.

TABLE I
Groups of terms used in the search strategy

Thematic Group	**Search Terms (Synonyms)**
1. Eye Tracking	"eye tracking", "gauze", "pupillometry", "pupil dilation", "fixation"
2. Electrodermal Activity	"EDA", "GSR", "skin conductance", "physiological signals"
3. Misinformation and Deception	"fake news", "disinformation", "misinformation", "deception", "deception detection", "truth recognition"
4. Cognitive Load	"cognitive conflict", "cognitive bias", "arousal", "incongruity"
5. Computational	"classification", "machine learning", "detection"

The groups of terms were combined using the AND operator, while synonyms within each group were combined using OR, resulting in an exact match search strategy applied to the title, abstract, and keywords.

E. Selection of sources of evidence

The Parsifal systematic review management software was used to guide the methodology for study screening and selection. Records retrieved from the unified searches—covering the Scopus, ScienceDirect, IEEE Xplore, PubMed, and Web of Science databases, as well as the regional SciELO database—were imported into the platform, where duplicate identification and removal were performed automatically. In the subsequent phase, the remaining articles were screened against eligibility criteria by reviewing titles and abstracts; studies that passed this stage underwent a detailed full-text evaluation to determine their final inclusion.

While expanding the search to the PubMed and Web of Science databases resulted in the incremental addition of six new unique records to the protocol, the search of the regional SciELO database yielded no results (zero occurrences) that simultaneously met the combined criteria of Electrodermal Activity, Eye Tracking, and Misinformation. To ensure the required transparency and reproducibility, the entire process—from identification and screening to exclusion and final inclusion—was documented using a flowchart structured in accordance with the PRISMA-ScR protocol guidelines (Figure 2).

F. Data extraction process

The data extraction was performed by completing a standardized form within the Parsifal software environment. The questions and fields of this form were structured to extract the methodological variables from each study, which are detailed in the following section (G. Data Items). This process ensured the organized capture of the methodological metrics of each primary article, correlating the type of biosensor equipment used with the results of mental workload and physiological responses reported in the literature.

G. Data items

The variables collected and analyzed from each study included: authorship, year of publication, country or region of origin of the research, biometric modality employed (eye tracking, electrodermal activity, or multisensory), and the main findings related to the neurophysiological behavior of the user during the consumption of misinformation.

H. Critical appraisal of individual sources of evidence

As recommended by the PRISMA-ScR guidelines for scoping reviews, formal risk of bias assessment is an optional criterion. However, to ensure the quality of the basis, the sources were critically evaluated for their "ecological validity." Priority was given to the analysis of studies that emulate real social media or news platform interfaces in the presentation of stimuli, over research that presented decontextualized texts on blank screens, suggesting that the measured stress and visual attention reflected the user's everyday internet behavior.

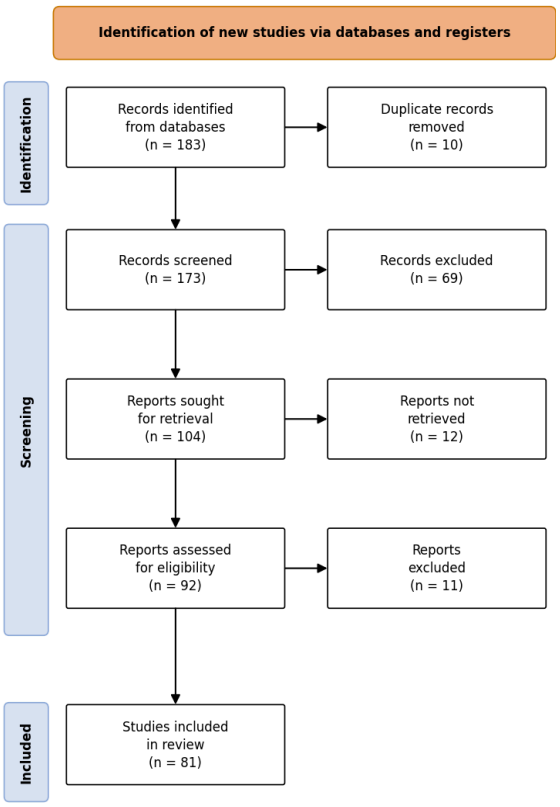


Fig. 2. Flowchart of the process for identifying, screening, determining eligibility, and including studies, developed according to the PRISMA-ScR protocol guidelines.

I. Summary of results

The extracted data were organized chronologically and categorized by type of physiological signal. This structuring allowed for the identification of trends in the temporal evolution of the research, as well as the highlighting of technological gaps, especially regarding the cross-referencing of these biometrics for the Portuguese language and the Brazilian sociocultural context.

III. Results

A. Selection of sources of evidence

The expanded systematic search yielded a total of 183 documents distributed across the selected databases. An analysis of the source distribution revealed that 41% of the articles came from ScienceDirect, 30.6% from Scopus, 24% from the IEEE Xplore digital library, 3.8% from Web of Science, and 0.5% from PubMed, whereas the regional database SciELO yielded no results (0%) based on the established search criteria (the sum of the percentages differs from 100% due to decimal rounding).

After importing the metadata into the Parsifal platform and automatically removing 10 duplicates, the 173 articles underwent screening based on title and abstract, resulting in the exclusion of 69 studies that did not meet the eligibility criteria. Of the 104 publications selected for retrieval, 12 could not be located, leaving 92 articles for full-text analysis. Following a full-text review, 11 studies were excluded due to methodological limitations. At the end of the process, 81 primary articles were accepted and included

in the qualitative synthesis of this review (the list of included studies is available as supplementary material – DOI: 10.5281/zenodo.22719150).

Among the records selected for retrieval, 12 publications could not be located in full-text format. This inability to retrieve the documents stemmed from two main factors: (a) paywall restrictions in highly specialized journals where full-access licenses were not covered by active subscriptions from the CAPES Journal Portal (via the Federated Academic Community – CAFe) or by IFB institutional databases; and (b) inconsistencies in metadata indexing, comprising records that provided only the abstracts from old conference proceedings, without the full text of the corresponding manuscript being available or preserved in the digital repositories of the databases consulted.

The exclusion of 11 studies following a critical analysis of the full text was justified by methodological limitations incompatible with the scope of this review, notably: (a) the absence of primary experimentation (secondary studies), comprising strictly conceptual articles or literature reviews lacking direct empirical collection of biosignals—such as the study by Joey F. George and Sophia Mannina (2025), which focused on a qualitative analysis of secondary health reviews; (b) the exclusive use of Natural Language Processing (NLP), characterized by works applying automated text classification algorithms without the participation of human volunteers or the recording of their physiological responses as receivers; (c) an exclusive focus on the sender (traditional polygraphy), consisting of research aimed at detecting active lying in classic criminal interrogation scenarios, which diverged from the central objective of mapping receiver behavior during the passive consumption of digital misinformation; and (d) biometric hardware inconsistency, identified in studies using abstract biometric simulations without proper specification of electrodermal activity (EDA) sensors or eye-tracker calibration parameters, thereby preventing the technical replicability of the experiments.

B. Main sources of evidence

The analysis of the 81 selected articles reveals a growing trend of integration between Computer Science and fields such as Neuroscience and Cognitive Psychology, aiming to incorporate physiological and cognitive information into traditional misinformation detection methods. Six representative studies stand out among those underpinning this review. In the context of ocular biometrics, studies by Kawano and Severino [9] investigate visual attention patterns in Brazilian individuals, while Wong et al. [8] analyze the monitoring of eye regressions as an indicator of cognitive processing during reading. Regarding autonomic stress, Nguyen et al. [6] and Cheng et al. [10] present evidence on the use of electrodermal activity as a measure of cognitive conflict and physiological arousal. Complementing this perspective, Jiang et al. [7] provide the psychological basis for the Illusory Truth Effect—a phenomenon widely associated with susceptibility to misinformation—while Sun et al. [12] advance this technological frontier by integrating electrodermal activity and ocular metrics to accurately predict a reader's implicit trust in health-related texts generated by Artificial Intelligence.

C. Critical appraisal of sources of evidence

The evaluation of experimental protocols showed that most international studies have high methodological robustness, frequently using clinical gold standard equipment. However, a deficiency was noted in the literature regarding validation in Portuguese-language databases. Few biometric studies cross-reference the use of biosensors with validated and updated ecological stimulus databases from Brazil, such as the FakeRecogna corpus [11].

This repository stands out as a milestone for national research, aggregating tens of thousands of factual and fraudulent news items extracted from partner fact-checking agencies (such as Lupa, Fato or Fake, and E-fakes) and traditional news portals (such as G1 and UOL). By categorizing fraud in a balanced way across areas such as health, politics, and science, including the context of the pandemic, this type of database contributes to the creation of realistic textual stimuli. The absence of standardized *corpora* in international studies suggests a point that is still little explored, which can guide future investigations in the Brazilian context.

Regarding the geographical distribution of the studies, the analysis revealed a strong concentration of experimental research in countries of the Northern Hemisphere and Asia. Publications from China and the United States were prominent, followed by Australia, which are leading the development of neurophysiological architectures for deception detection.

Conversely, a scarcity of applied research in the Brazilian context was observed (Table II). Most of the included national research focuses on the construction of textual databases or the isolated observation of visual attention, highlighting the need to expand multimodal experimentation with biosensors to the Portuguese language.

TABLE II
Geographic concentration and technological focus of the selected publications

Country of origin	**Volume in Literature**	**Predominant Technological/Thematic Focus**
China	High	AI, Eye Tracking and Brain Spectroscopy (fNIRS).
United States	High	Cognitive biases, the Illusory Truth Effect, and EDA.
Australia	Average	Multimodal fusion of biosignals for real-time detection.
Italy	Average	Facial microexpressions and thermal/ocular polygraphy.
Brazil	Low	Isolated visual attention, literature reviews, and textual *corpora* .

Note: The volume of publications in the literature was stratified according to the following occurrence intervals in the sample: High (> 10 articles), Medium (4 to 10 articles), and Low (1 to 3 articles).

To characterize the predominant technological approaches in literature, the frequency of use of physiological modalities in the 81 included studies was extracted. The analysis revealed a historical preference for

the isolated application of sensors, with a strong predominance of eye tracking and EDA, as shown in Fig. 3.

Notably, only a small fraction of studies (approximately 8.9%) adopts real-time multimodal fusion architecture. This metric corroborates the scarcity of systems capable of simultaneously cross-referencing visual attention and autonomic stress during the consumption of misinformation.

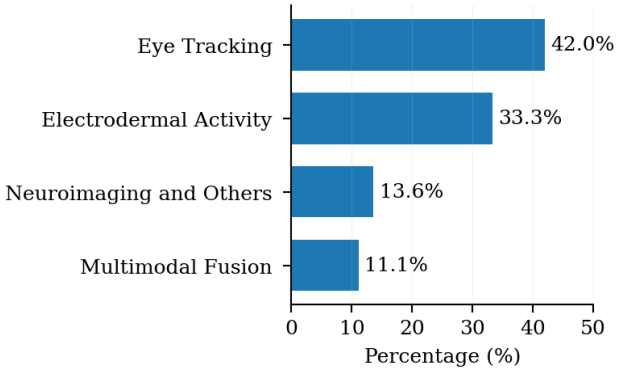


Fig. 3. Proportion of biosignal modalities used in the included studies.

D. Results from individual sources of evidence

The extracted analyses confirmed the viability of biosensors as tools for deception detection. In the ocular spectrum, it was observed that people allocate their visual attention differently when faced with fake news related to COVID-19 [9]. Tracking showed significant differences in areas of interest, with volunteers performing a notable number of regressive saccadic movements and prolonged fixations in the search for sources that attest to the credibility of the material. In addition to these findings, it is observed that ocular metrics, such as time to first fixation (*TTFF*) and dwell time, show significant variations when Brazilian readers are exposed to alarmist headlines [9].

Furthermore, Areas of Interest (AOIs) containing logos of media outlets or images with high emotional appeal tend to capture peripheral visual attention more intensely, inducing the oculomotor system to perform rapid and repetitive scans. This pattern of visual searching for clues that aid in assessing the credibility of information suggests that the design of the social network directly influences the cognitive effort and visual processing involved in interpreting the content.

In another line of research on eye behavior, the processing of misinformation and its respective corrections was mapped [8]. The study monitored the durations of the first and second eye passes in different areas of interest on the screen. The results indicated that the inclusion of explicit reminders about the falsity of a news item drastically altered reading speed and increased eye regressions. Furthermore, it was found that the individual's verbal working memory capacity acts as a significant predictor of the efficiency with which the reader integrates the correction of misinformation, directly reflecting the captured eye effort.

In the physiological field, studies employing EDA have demonstrated that response latency and skin conductance vary autonomously in the face of textual incongruities [6]. The results revealed that, even when participants stated they believed false information, increases in electrodermal activity were observed, indicating greater physiological activation. These findings suggest that the sympathetic nervous system may respond to signals of inconsistency present in the content before a fully consolidated conscious evaluation occurs, evidencing a subconscious component in the processing of misleading information.

Extending the investigation to the cerebral cortex, functional near-infrared spectroscopy (fNIRS) was used to measure how the authority of a fact-checking agency affects mental effort [10]. The data pointed to a neurological pattern associated with cognitive dissonance: when volunteers read fake news corrected by highly authoritative sources, the brain used a "fast track" of heuristic acceptance, reducing the activation of the right prefrontal cortex (RPFC). Conversely, when the verification came from an unofficial source, the perceived low credibility forced the brain to maintain a prolonged and vigilant evaluation, keeping the RPFC highly activated. This indicates that the body reacts measurably not only to the false text, but also to the signals of credibility and authority attached to it.

E. Summary of results

The temporal distribution analysis of the articles revealed a dynamic behavior directly influenced by global crises. Although theoretical studies on biometrics and polygraphs date back to the 1970s, there has been an exponential growth in academic interest focused on the internet since 2018. This leap reflects the shock caused by the use of massive disinformation campaigns in high-impact electoral scenarios, such as the referendum on the United Kingdom's exit from the European Union (Brexit) in 2016, as well as the presidential elections in the United States and Brazil. In the Brazilian context, the phenomenon was strongly characterized by the use of mass dissemination of fake news through messaging applications in 2018, which boosted the global urgency for new detection and literacy tools [5].

The timeline graph extracted from the mapping showed a reduction in the number of publications between 2020 and 2021. This behavior may be related to the impacts of the COVID-19 pandemic on experimental research, a period in which the temporary closure of laboratories and restrictions on conducting in-person studies hampered experiments involving physiological sensors and human participants. With the resumption of in-person activities, a significant increase in publications was observed, culminating in a historical peak in the years 2024 and 2025. This increase may be associated both with the resumption of experimental studies and with the growing scientific interest in developing more effective methods to combat the infodemic and the spread of fake news (Fig. 4).

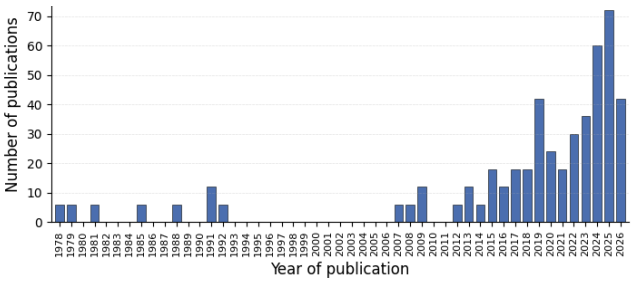


Fig. 4. Temporal evolution of the publications included in the scoping review.

IV. Discussion

A. Summary of the evidence

The gathered evidence supports the proposed objectives of this scoping review and directly answers the questions that motivated this study. Regarding the predominant biosignals (RQ1), the mapping confirmed the literature's historical preference for Eye Tracking (42.0%) and Electrodermal Activity (33.3%), predominantly applied in isolation. As for the contribution of eye tracking (RQ2), the findings indicate that it maps cognitive conflict through increased visual effort and saccadic regressions, revealing exactly where the brain encountered the incongruity. Regarding evidence of the unconscious recognition of truth (RQ3), the literature confirms that autonomic biosignals - particularly Electrodermal Activity (EDA) - register increases in conductance and sympathetic nervous system stress when exposed to fraudulent content; this occurs even when volunteers consciously state they believe the fake news, demonstrating a subconscious and involuntary biometric response to deception [6]. An analysis of news elements (RQ4) revealed that alarmist headlines, emotionally charged images, and media outlet logos (sources) were the factors that triggered the most significant changes in gaze fixation and scanning patterns during the assessment of credibility [9].

The individual efficacy of these technologies already shows promising results for applications in systems engineering. Machine learning models trained exclusively on EDA data achieved 63.46% accuracy (with an F1-score exceeding 62%) in predicting the veracity of information, outperforming other physiological signals evaluated in the same study—such as photoplethysmography (PPG), which achieved 58.79% accuracy in identifying cognitive discrepancies [6]. These results reinforce the potential of biosignals as a source of objective information for developing automated systems to support disinformation detection—a field driven by cutting-edge research demonstrating the feasibility of combining eye-tracking with physiological signals (such as EDA and heart rate variability) to accurately predict users' implicit confidence levels and classify message authorship (human vs. AI-generated) with accuracies exceeding 72% [12].

However, the possibility of incorporating this biosignal-based architecture into everyday applications raises significant bioethical implications. The use of sensors capable of passively inferring cognitive states and detecting physiological stress responses necessitates the establishment of legal and ethical safeguards to protect individuals' cognitive privacy, preventing technologies intended to combat misinformation from being transformed into mechanisms for emotional surveillance. Furthermore, although the mapping indicates that most historical studies have utilized biosignals in isolation, the recent introduction of architectures performing synchronized multimodal fusion of ocular and autonomic data is beginning to bridge this gap, establishing itself as a highly promising direction [12] for future research aimed at developing multimodal systems that are potentially more robust and accurate in detecting misinformation.

B. Physiological and cognitive mapping

Analysis of the literature demonstrated that different types of biosensors are capable of capturing specific dimensions of misinformation processing. While some signals reflect the effort expended to understand the message, others reveal the subconscious stress resulting from cognitive dissonance. To systematize these findings, Table III presents the relational matrix between the main biometric metrics extracted and the psychological phenomena associated with them.

TABLE III
Relational matrix of biosignals and cognitive phenomena in the detection of misinformation

Modality	Extracted Metric	Cognitive Phenomenon
Electrodermal Activity (EDA)	Skin conductance peaks	Cognitive dissonance, involuntary emotional arousal, and autonomic stress in response to deception.
Eye Tracking	Time and Duration of Fixations	Level of visual attention, processing effort, and information extraction from the headline.
Eye Tracking	Regressions	Distrust, reassessment of context, and critical rereading (e.g., source verification)
Eye Tracking	Pupillary dilation	Increased mental load and effort required to decode complex lies.
Neuroimaging (fNIRS / EEG)	Activation of the Prefrontal Cortex	Resolution of conflicts related to credibility, dependence on authority, and decision-making.

C. Research gaps and future opportunities

Despite the exponential growth in publications since 2018, the mapping revealed that the application of biosignals in combating disinformation still has significant methodological and contextual limitations. The predominant focus of the classic literature on polygraphics has always been the analysis of the lie sender (the one who tells the lie), with a lack of models focused on the observer (the one who consumes the fake news). Table IV summarizes the main gaps identified in this work and outlines research opportunities for the advancement of digital security.

TABLE IV
Map of identified gaps and research opportunities

Research Area	Gap Identified in the Literature (Problem)	Research Opportunity (How to fill out)
Focus of Detection	Most studies focus on the "sender" of the lie (criminal analysis/ (polygraphs).	Develop detection systems focused on the receiver/ observer.
Ecological Validity	Laboratory tests using artificial phrases without the real context of a social network.	Use of local ecological databases containing stimuli with Brazilian headlines and logos.
Sensor Architecture	Isolated use of a single sensor	Design multimodal fusion models in a synchronized, real-time manner.
Bioethics and Privacy	Risk of the involuntary use of biometrics for state or corporate	Establishing rigorous ethical frameworks to protect the user's

	surveillance.	"cognitive privacy".

D. Limitations

The limitations of this systematic mapping are restricted to the parameters defined in the initial protocol. The exclusive use of three international academic databases (Scopus, ScienceDirect, and IEEE), while guaranteeing very high quality due to peer review, may have resulted in the omission of preliminary studies, dissertations, and grey literature indexed in regional repositories not integrated into global search platforms.

E. Conclusions

This scoping review revealed that non-invasive biometric technologies, particularly eye tracking and EDA, have the potential to identify neurophysiological responses associated with the consumption of misinformation. The studies analyzed indicate that these biosignals allow the detection of cognitive conflicts and emotional changes triggered during exposure to potentially false content, complementing traditional approaches based exclusively on textual analysis. Given the increasing sophistication of generative AIs in the production of synthetic content, the results suggest that cybersecurity strategies can benefit from incorporating the human factor as an additional layer of detection.

The findings are also relevant from a social point of view. Evidence from the literature indicates that the ability to identify fake news is associated with the level of education and digital literacy, making certain groups more susceptible to misinformation [4]. In this context, biosignal-based systems can reduce the exclusive dependence on the user's conscious cognitive skills, offering complementary mechanisms to support the identification of potentially misleading content and expanding the accessibility of these technologies to different user profiles.

Finally, the mapping identified an important gap in the literature: the predominance of studies that employ eye tracking and EDA independently, to the detriment of synchronized multimodal architectures. This result reinforces the need for experimental research investigating the integration of these biosignals in ecological scenarios, including the Brazilian context. The consolidation of these architectures could contribute to the development of biometric solutions embedded in wearable devices, such as smartwatches, capable of acting as support mechanisms for the detection of disinformation in real time, provided they are accompanied by adequate ethical and privacy safeguards.

Acknowledgments

The authors thank Professor Dr. Carla Simone Castro da Silva for her support and the opportunity to join the team at the School of Business and Social Development (ENEDES). They also thank the Brazilian Agency for Industrial Development (ABDI) and the Federal Institute of Brasília (IFB) for their partnership and support for the research, innovation, and extension activities developed within the scope of projects no. 64/2025 and no. 335/2026. FHMO and RHCS thank the Foundation for Research Support of the Federal District (FAPDF) for granting the Post-Doctoral Fellowships (Processes no. 23098.000311.2026-80 and no. 23098.000264.2026-74), which enabled the development of this research.

Statement on the Use of Generative AI:

The Gemini 1.5 Pro tool (Google) was used exclusively as assistive support in the preparation of this work, limited to aiding in the interpretation of technical terms in English, methodological organization (PRISMA-ScR), spelling review, and the generation of Python scripts for plotting figures.

References

[1] UNESCO, "Global Risks Report 2025: Misinformation and disinformation as top global risks," Paris: UNESCO, 2025.

[2] C. Macário, "AI-powered fakes grow 308% in one year, reveals unprecedented study by Lupa," Agência Lupa, Feb. 2026.

[3] DataSenado, "Social networks and fake news," Brasília: Institutional Portal of the Federal Senate, 2023. Available at: https://www12.senado.leg.br/institucional/datasenado/materias/relatorios-de-pesquisa/redes-sociais-e-noticias-falsas.

[4] DataSenado, "DataSenado Survey Reveals What Brazilians Think About Fake News," Brasília: Panorama Político, 2024. Available at: https://www12.senado.leg.br/institucional/datasenado/materias/relatorios-de-pesquisa/pesquisa-datasenado-revela-o-que-pensa-o-brasileiro-sobre-fake-news.

[5] MPF Santos, TAF Andrade, and A. Jaeger, "Memory and fake news: Contributions of cognitive psychology to the understanding of the phenomenon of disinformation," Cadernos de Psicologia, 2024.

[6] G. Nguyen, L. Wang, and Y. Jiang, "Truth and trust: fake news detection via biosignals," International Journal of Information Technology, 2025.

[7] Y. Jiang, N. Schwarz, KJ Reynolds, and EJ Newman, "Repetition increases belief in climate-skeptical claims, even for climate science endorsers," PLoS ONE, vol. 19, no. 8, p. e0307294, Aug. 2024.

[8] R. Wong, L. Yu, A. Veldre, ED Reichle, and MA Eskenazi, "Real-time processing of misinformation and its correction: Insights from eye movements during reading," Cognitive Research: Principles and Implications, 2026.

[9] DR Kawano and TN Severino, "Do Fake News Catch Our Attention: A Study of Visual Attention Applied to the Consumption of Fake News About COVID-19 in Brazil," UNESCO MILID Yearbook, 2023.

[10] S. Cheng, X. Yang, and Y. Ding, "Authority reliance vs. deliberative assessment in processing online rumors: evidence from fNIRS," Frontiers in Human Neuroscience, vol. 20, p. 1750499, Feb. 2026.

[11] GL Garcia, LCS Afonso, and JP Papa, "FakeRecogna: A new Brazilian corpus for fake news detection," in Proceedings of the National Meeting on Artificial and Computational Intelligence, São Paulo, 2021.

[12] X. Sun, R. Ma, S. Wei, P. Cesar, J. A. Bosch, and A. El Ali, "Understanding trust toward human versus AI-generated health information through behavioral and physiological sensing," International Journal of Human-Computer Studies, vol. 209, p. 103714, 2026.

Análise Comparativa das curvas fluxo-tempo entre um sensor mecânico de turbina e um sensor piezoresistivo

1st Jorge Luis Vitorino
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0009-0006-6307-8018

2nd João Gabriel de Souza Veras
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0000-0003-1351-4812

3rd Emery Cleiton Cabral Correia Lins
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0000-0002-7967-6163

Resumo*—A espirometria depende da qualidade do sistema de aquisição do fluxo respiratório, e diferentes princípios de sensoriamento podem produzir respostas dinâmicas distintas. Este estudo experimental e comparativo buscou avaliar o comportamento das curvas fluxo-tempo obtidas por um sensor mecânico de turbina e por um sistema piezoresistivo submetidos a padrões respiratórios controlados e reprodutíveis gerados por um ventilador mecânico Servo-s. O sistema piezoresistivo operou com intervalo de amostragem de 50 ms, correspondente à frequência de 20 Hz, totalizando 951 amostras distribuídas em aproximadamente 16 a 20 ciclos respiratórios, contemplando as fases inspiratória e expiratória. Foram analisados morfologicamente o início do fluxo, a formação do pico, a fase de desaceleração, a presença de platô, o retorno à linha de base e a repetibilidade visual dos padrões. As curvas obtidas demonstraram similaridade morfológica na identificação dos principais eventos respiratórios, embora tenham sido observadas diferenças de amplitude e resposta temporal entre os sistemas, especialmente nas regiões de rápida variação do fluxo. Essas diferenças podem estar relacionadas aos distintos princípios físicos de funcionamento, à resposta dinâmica, à calibração e ao processamento dos sinais. Considerando que as aquisições não foram simultâneas e que as séries não foram integralmente padronizadas em uma escala comum de L/s, os resultados desta etapa devem ser interpretados como evidência de similaridade morfológica, não permitindo estabelecer equivalência metrológica ou superioridade entre os sensores. Estudos posteriores com aquisição sincronizada, fluxo de referência comum e análises quantitativas serão necessários para determinar erros, concordância e desempenho metrológico entre os sistemas.

***Index Terms*—Espirometria; Sensor mecânico de turbina; Sensor piezoresistivo; Curvas respiratórias; Calibração; Engenharia Biomédica**

I. Introdução

A respiração é essencial à manutenção da vida e sua avaliação objetiva depende de instrumentos capazes de registrar, com precisão e repetibilidade, as variações de fluxo e volume ao longo do ciclo respiratório [1]. Nesse cenário, a espirometria se consolida como um método de referência para diagnóstico da função pulmonar, de modo que a fidedignidade técnica de seus registros está estritamente vinculada tanto à capacidade operacional do instrumento quanto a processos sistemáticos de calibração, aferição periódica e controle de qualidade [2]–[5]. As recomendações ATS/ERS reforçam a necessidade de padronização dos procedimentos e de controle da exatidão e repetibilidade dos sistemas espirométricos [6] [7]. Estudos de controle de qualidade demonstram que pequenas variações sistemáticas podem não ser identificadas por critérios isolados de calibração, o que justifica a avaliação periódica do comportamento do sistema ao longo do tempo [3] [5]. Dentre as metodologias aplicadas na quantificação do fluxo respiratório, destacam-se os transdutores mecânico-ópticos de turbina e os dispositivos eletrônicos fundamentados no princípio de pressão diferencial. Nos sensores de turbina, o movimento giratório de um rotor interno é traduzido em pulsos proporcionais à taxa de fluxo. Por sua vez, os sistemas piezoresistivos aferem deformações físicas ou gradientes de pressão, demandando uma função matemática de calibração para correlacionar o sinal elétrico de saída com a respectiva vazão respiratória [8]–[11]. Adicionalmente, por viabilizarem a miniaturização e a integração de componentes eletrônicos, os sensores piezoresistivos tornam-se bastante promissores para o desenvolvimento de sistemas vestíveis e embarcados; contudo, sua aplicação com fins quantitativos permanece condicionada a rigorosos processos de calibração, técnicas de compensação e validação metrológica [9]. Sistemas de biofeedback e monitorização respiratória podem ampliar a disponibilidade de informações fisiológicas em tempo real [12]. Além disso, a relevância clínica de medidas respiratórias objetivas é particularmente evidente em doenças como asma e outras condições obstrutivas, nas quais a caracterização funcional auxilia a avaliação do comprometimento respiratório [13]. O interesse pela análise comparativa entre diferentes tecnologias de mensuração do fluxo respiratório surgiu a partir de uma linha de pesquisa desenvolvida pelo grupo voltada à construção de um sistema de feedback respiratório em tempo real, baseado na aquisição do fluxo expiratório por meio de um sensor piezoresistivo. O desenvolvimento desse sistema requer não apenas a aquisição e o processamento do sinal respiratório, mas também a caracterização do comportamento do sensor frente a tecnologias já empregadas na mensuração do fluxo. Nesse contexto, a comparação com o sensor mecânico

de turbina torna-se uma etapa importante do processo de desenvolvimento e avaliação do sistema, permitindo investigar como diferentes princípios de sensoriamento representam os principais eventos das curvas fluxo-tempo, bem como identificar possíveis diferenças relacionadas à resposta dinâmica, amplitude, formação do pico, desaceleração e retorno à linha de base [1]. O objetivo deste estudo foi comparar, a morfologia das curvas fluxo-tempo obtidas por um sensor mecânico de turbina e por um sensor piezoresistivo submetidos a padrões de fluxo gerados por ventilador mecânico Servo-s, descrevendo a capacidade de ambos os sistemas de identificar os principais eventos temporais do ciclo respiratório. O estudo não teve como objetivo, nesta etapa, demonstrar equivalência metrológica entre as tecnologias.

II. Materiais e Métodos

Trata-se de estudo experimental in vitro, metodológico e comparativo, realizado exclusivamente em bancada para calibração e avaliação funcional de sistemas de medida de fluxo. Não foram incluídos sujeitos humanos ou animais, não houve coleta de material biológico e não foram utilizados prontuários ou dados clínicos. Por essa razão, o protocolo correspondeu a ensaio instrumental de bancada, sem participação de seres humanos. O fluxo de referência foi produzido por um ventilador mecânico Servo-s, empregado como fonte controlada e reprodutível de ciclos inspiratórios e expiratórios. Foram avaliados padrões de fluxo contínuo, esforço expiratório e ciclos intermitentes/bifásicos. O conjunto de dados do sistema piezoresistivo compreendeu 951 amostras, correspondendo aproximadamente a 16–20 ciclos respiratórios ao longo dos ensaios. O intervalo entre as amostras foi de 50 ms, equivalente a frequência de amostragem de 20 Hz, contemplando inspiração e expiração. O sensor mecânico de turbina do espirômetro foi empregado como sistema comparativo para registro de fluxo. Seu princípio baseia-se na rotação de um elemento móvel, com conversão da frequência de pulsos ópticos em fluxo e obtenção do volume pela integração temporal [8]–[11]. O sistema piezoresistivo detectou variações de pressão/deformação por alteração resistiva em ponte de Wheatstone; o sinal foi amplificado e digitalizado, sendo sua relação com o fluxo estabelecido durante a calibração em bancada. A comparação desta etapa foi predominantemente morfológica e temporal. Para cada condição, foram examinados o início do fluxo, a ocorrência do pico, a fase de desaceleração, a presença de platô, o retorno à linha de base e a repetição do padrão entre ciclos. As aquisições originais não foram simultâneas, e os dois sistemas não foram registrados em uma escala calibrada comum em L/s. Além disso, o conjunto de dados de origem não fornece a frequência de amostragem do sistema piezoresistivo nem um número predefinido de ciclos repetidos por condição. Portanto, o presente conjunto de dados não permite o cálculo retrospectivo da concordância metrológica quantitativa.

A. Princípio de Funcionamento

Sensor mecânico de turbina: a passagem do ar acelera o rotor e as interrupções de um feixe óptico geram pulsos cuja frequência é relacionada ao fluxo. A tecnologia é utilizada em espirômetros portáteis comerciais e exige procedimentos de verificação e controle de qualidade [9] [11]. Sensor piezoresistivo: a deformação de uma membrana modifica a resistência elétrica dos elementos sensores. A tensão resultante é amplificada e digitalizada; para estimativa de fluxo, o sinal de pressão diferencial deve ser relacionado a uma função de calibração adequada à geometria do sistema [9] [13].

B. Comportamento comparativo das curvas(Fluxo-Tempo)

Em transientes rápidos, diferenças entre as curvas podem decorrer da inércia do rotor, da resposta dinâmica do transdutor, da filtragem e do processamento. Assim, uma subida aparentemente mais íngreme ou um pico mais estreito em um dos sistemas não deve ser interpretado isoladamente como superioridade de desempenho. A confirmação de atraso, erro ou concordância exige aquisição comparável e análise quantitativa.

C. Síntese Comparativa

De forma sintética, o sensor de turbina converte a rotação do rotor em pulsos relacionados ao fluxo e pode sofrer influência de inércia, atrito e rotação residual. O sensor piezoresistivo converte pressão/deformação em sinal elétrico e seu comportamento depende da dinâmica do transdutor, calibração, filtragem, offset e deriva. Em ambos os casos, a comparação de desempenho requer verificação/calibração e controle de qualidade; para o sistema piezoresistivo, a validação quantitativa requer ainda conversão rastreável para L/s frente ao fluxo de referência.

III. Resultados e discussão

Nos 951 pontos amostrados pelo sistema piezoresistivo, distribuídos em aproximadamente 16–20 ciclos, foi possível identificar de forma consistente as fases inspiratórias e expiratórias e os principais eventos temporais dos padrões produzidos em bancada. A frequência de 20 Hz forneceu uma resolução temporal de 50 ms por amostra para a inspeção morfológica realizada nesta etapa. Nos ensaios de esforço expiratório, as curvas de ambos os sistemas exibiram elevação do fluxo, formação de pico e posterior desaceleração. Nos padrões intermitentes/bifásicos, os sistemas permitiram reconhecer picos sucessivos separados por intervalos de menor fluxo. Nos testes de fluxo contínuo, observou-se manutenção de platô e retorno à linha de base. Esses achados sustentam apenas a correspondência morfológica dos eventos, não a equivalência das amplitudes ou do tempo de resposta.

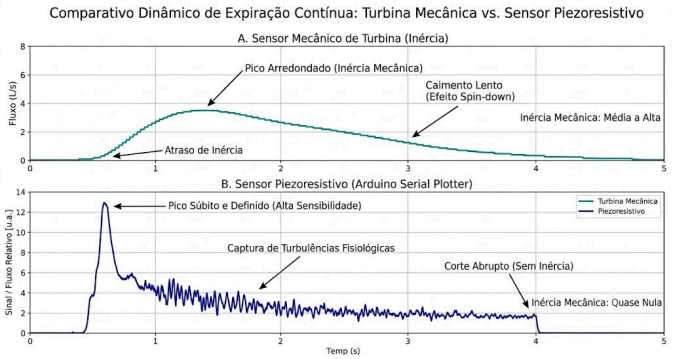


Figura 1. Turbina mecânica vs Sensor piezoresistivo

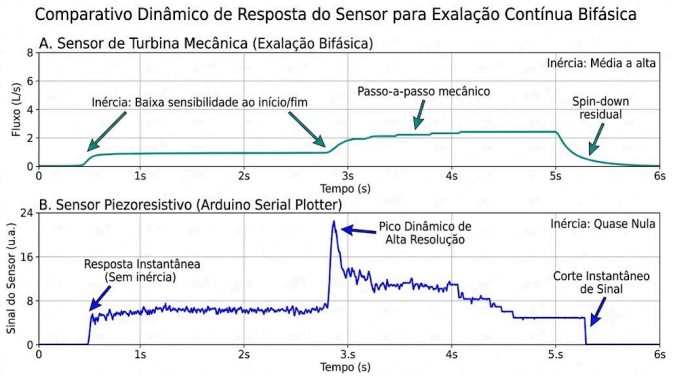


Figura 2. Comparação em Exalação Bifásica

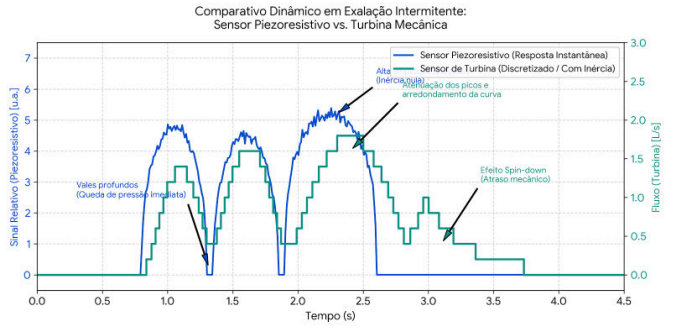


Figura 3. Comparação em Exalação intermitente

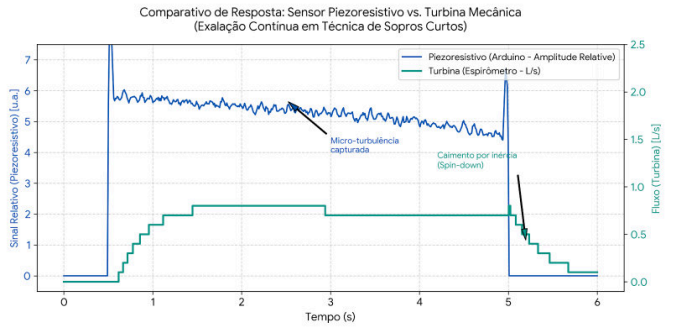


Figura 4. Comparação durante expiração contínua

A interpretação dos gráficos deve considerar que as escalas originais não estão integralmente padronizadas em L/s para os dois sistemas e que não houve aquisição simultânea. Por esse motivo, diferenças visuais de amplitude, inclinação ou posição do pico não foram convertidas em métricas de erro ou atraso. As figuras devem, portanto, ser lidas como descrição dos comportamentos esperados e observados, e não como prova quantitativa de maior sensibilidade, menor inércia ou melhor desempenho de um sensor. Esse cuidado é coerente com os princípios de padronização e controle de qualidade da espirometria, nos quais calibração, repetibilidade e rastreabilidade das medidas são essenciais para comparações instrumentais [2]–[7]. A literatura sobre sistemas portáteis e vestíveis também destaca a necessidade de validação do sinal antes de extrapolar o desempenho para aplicações de monitorização [11] [12]. A inspeção das curvas fluxo-tempo evidenciou diferenças entre os sinais obtidos pelo sensor piezoresistivo e pelo sensor mecânico de turbina, principalmente nas regiões de maior variação temporal do fluxo. Nos eventos de subida e formação do pico, observou-se diferença de amplitude e de posicionamento temporal entre as curvas, enquanto nos segmentos de fluxo mais estável houve maior aproximação do comportamento morfológico dos sistemas. Nos padrões intermitentes e bifásicos, ambos os sensores identificaram adequadamente a sucessão dos ciclos, embora tenham sido observadas diferenças na amplitude dos picos, na inclinação das fases de aceleração e desaceleração e no retorno à linha de base. Essas discrepâncias constituem, nesta etapa, um erro comparativo predominantemente morfológico, potencialmente associado às diferenças entre os princípios físicos de aquisição, à inércia mecânica da turbina, à resposta dinâmica do sistema piezoresistivo, à calibração e ao processamento do sinal. Dessa forma, os resultados atuais demonstram similaridade morfológica entre as curvas fluxo-tempo. Não se afirma que o sensor piezoresistivo tenha desempenho “compatível”, equivalência metrológica ou que já constitua alternativa ao sensor de turbina. Essas conclusões dependerão de nova aquisição quantitativa sob o mesmo fluxo de referência, com sinais em L/s, sincronização temporal e cálculo das métricas recomendadas pelo revisor.

Tabela I
DIFERENÇAS DE FUNCIONAMENTO E COMPORTAMENTO NA CURVA FLUXO-TEMPO

Aspecto	Sensor mecânico de turbina	Sensor piezoresistivo
Princípio de funcionamento	O fluxo de ar gira uma turbina; a rotação é convertida em sinal de fluxo.	A pressão ou deformação modifica a resistência elétrica do sensor, gerando um sinal proporcional.
Resposta ao início do fluxo	Pode apresentar pequeno atraso devido à inércia e ao atrito das pás.	Tende a responder rapidamente às mudanças de pressão, dependendo da calibração e da amostragem.
Pico de fluxo	Pode suavizar ou atrasar o pico quando há variação muito rápida do fluxo.	Pode identificar variações rápidas, mas o pico pode ser influenciado por ruído e filtragem.
Fase de desaceleração	A curva acompanha a redução da rotação da turbina de forma progressiva.	Acompanha a queda do fluxo, podendo apresentar pequenas oscilações eletrônicas.
Retorno à linha de base	Pode ocorrer de forma mais lenta pela rotação residual da turbina.	Pode ser mais rápido, porém sujeito à deriva da linha de base.
Comportamento geral da curva	Curva estável e amplamente utilizada como referência em espirômetros.	Pode reproduzir os principais eventos da curva, com pequenas diferenças morfológicas.
Principal diferença	Depende de componentes mecânicos móveis e sofre influência de inércia, atrito e desgaste.	Não depende de rotação mecânica, mas exige calibração, amplificação e processamento do sinal.

Nota: esta tabela resume o comportamento esperado dos sensores. A confirmação de diferenças de resposta, atraso e concordância depende de aquisição simultânea de dados e validação quantitativa.

IV. CONCLUSÃO

O ensaio de bancada permitiu descrever o comportamento das curvas fluxo-tempo de um sensor mecânico de turbina e de um sistema piezoresistivo submetidos a padrões respiratórios controlados por ventilador Servo-s. O sistema piezoresistivo foi adquirido a 20 Hz (50 ms por amostra), totalizando 951 amostras e aproximadamente 16–20 ciclos, sem participação de seres humanos. Ambos os sistemas permitiram reconhecer morfologicamente início do fluxo, pico, desaceleração, padrões intermitentes, platô e retorno à linha de base. Contudo, a ausência de aquisição simultânea e de padronização integral das séries na mesma unidade impede afirmar concordância, equivalência ou superioridade entre os sensores. Os resultados desta etapa devem ser interpretados exclusivamente como evidência de similaridade morfológica. Como continuidade, recomenda-se validação metrológica com o mesmo fluxo de referência aplicado aos dois sensores, aquisição simultânea ou ciclos rigorosamente reproduzíveis do Servo-s, padronização em L/s e pelo menos 10 ciclos por condição, reportando erro absoluto e percentual do pico, atraso temporal, RMSE, correlação, Bland-Altman, média, desvio-padrão e coeficiente de variação.

REFERÊNCIAS

[1] J. G. S. Veras, "Desenvolvimento de um sistema portátil para biofeedback respiratório em tempo real com sensor de pressão e interface mobile via comunicação BLE," Trabalho de Conclusão de Curso (Graduação em Engenharia Biomédica), Dept. Eng. Biomédica, Universidade Federal de Pernambuco, Recife, Brasil, 2025.

[2] M. Viljoen, R. van Zyl-Smith, M. S. Abdool-Gaffar *et al.*, "Position statement for adult and paediatric spirometry in South Africa: 2022 update," *African Journal of Thoracic and Critical Care Medicine*, vol. 28, no. 4, 2022.

[3] P. L. Enright, C. Skoog, and L. R. Johnson, "Spirometer calibration checks: Is 3.5% good enough?" *Chest*, vol. 131, no. 5, pp. 1486–1493, 2007, DOI: 10.1378/chest.06-1522.

[4] American Thoracic Society, "Official American Thoracic Society technical standards: Spirometry in the occupational setting," *American Journal of Respiratory and Critical Care Medicine*, vol. 189, no. 8, pp. 983–993, 2014, DOI: 10.1164/rccm.201402-0337ST.

[5] J. M. Haynes and G. L. Ruppel, "Should spirometer quality control be treated like other laboratory devices?" *ERJ Open Research*, vol. 5, no. 1, Art. N. 00249-2018, 2019, DOI: 10.1183/23120541.00249-2018.

[6] M. R. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates *et al.*, "Standardisation of spirometry," *European Respiratory Journal*, vol. 26, no. 2, pp. 319–338, 2005, DOI: 10.1183/09031936.05.00034805.

[7] B. L. Graham, I. Steenbruggen, M. R. Miller, I. Z. Barjaktarevic, B. G. Cooper, G. L. Hall *et al.*, "Standardization of spirometry 2019 update: An official American Thoracic Society and European Respiratory Society technical statement," *American Journal of Respiratory and Critical Care Medicine*, vol. 200, no. 8, pp. e70–e88, 2019, DOI: 10.1164/rccm.201908-1590ST.

[8] E. Dompeling *et al.*, "Accuracy, precision and linearity of the portable flow-volume meter Microspiro HI-298®," *European Respiratory Journal*, vol. 4, pp. 612–615, 1991.

[9] MIR Medical International Research, *MiniSpir: Manual do Usuário*, Rev. 1.9, Roma, Itália: MIR, 2020.

[10] E. Q. M. Brisola *et al.*, "Desenvolvimento de um sistema embarcado para análise do sistema respiratório através de espirometria," *Cuadernos de Educación y Desarrollo*, vol. 16, no. 7, pp. 1–21, 2024.

[11] K. Malik and A. Dua, "Advancing patient care with biofeedback," Treasure Island, FL, USA: StatPearls Publishing, Jan. 2025. [Online]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/31971716/

[12] C. P. C. Marques et al., "Epidemiologia da asma no Brasil, no período de 2016 a 2020," Research, Society and Development, vol. 11, no. 8, Art. no. e5211828825, 2022.

[13] K. S. Moon and S. Q. Lee, "A wearable multimodal wireless sensing system for respiratory monitoring and analysis,"Sensors, vol. 23, no. 15, Art. no. 6790, 2023.

Aplicação dos sensores MMWave no monitoramento de sinais vitais e detecção de quedas

1st Henrique Issao Kato
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
henrique.issao@gec.inatel.br

2nd Lucca Marcondes Madeira
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
lucca.marcondes@ges.inatel.br

3rd Luma Rissatti Borges do Prado
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
luma.prado@inatel.br

4th Rita Elizabeth Santos de Almeida
eHealth Innovation Center
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brasil
r.elizabeth@inatel.br

***Abstract*—This paper presents a theoretical and experimental study on the use of mmWave sensors for biomedical monitoring. The MR60BHA2 sensor was evaluated for respiratory and heart rate monitoring, while the MR60FDA2 sensor was tested for human presence and fall detection. Experiments were conducted under controlled conditions and in scenarios with possible interference, including humidity, Wi-Fi signals, obstacles, clothing layers, body position, ventilation, metallic objects, and different sensor settings. The results showed that the MR60BHA2 presented better performance in respiratory monitoring than in heart rate measurement, while the MR60FDA2 detected controlled falls but was affected by false positives and fast movements. These findings indicate that mmWave sensors have potential for homecare applications, provided that calibration, installation conditions, and environmental interference are carefully considered.**

***Index Terms*—mmWave, heart rate, respiratory rate, fall detection**

I. Introdução

Os sensores mmWave são dispositivos de sensoriamento que utilizam ondas eletromagnéticas na faixa de ondas milimétricas para detectar presença, movimento e pequenas variações de deslocamento sem a necessidade de contato físico com o alvo. Eles operam, em geral, na faixa de 30 GHz a 300 GHz, que corresponde a comprimentos de onda na ordem de milímetros. Seu funcionamento baseia-se na técnica *Frequency Modulated Continuous Wave* (FMCW), na qual um sinal eletromagnético é transmitido continuamente e refletido ao encontrar um alvo. Ao analisar a diferença de frequência e fase entre o sinal transmitido e o sinal recebido, é possível calcular a distância em relação ao objeto detectado e identificar pequenos movimentos [1].

Os sensores mmWave têm se mostrado promissores em aplicações biomédicas por permitirem a detecção de movimentos corporais sem contato direto com o paciente. Essa característica possibilita seu uso no monitoramento de sinais vitais, como respiração e batimentos cardíacos, na detecção de presença humana em ambientes internos, no reconhecimento de movimentos e atividades corporais e na identificação de eventos de risco, como quedas. Em ambientes hospitalares e de *homecare*, essas aplicações são relevantes para o acompanhamento de pacientes, idosos ou pessoas com mobilidade reduzida, pois permitem um monitoramento mais contínuo, discreto e menos dependente de dispositivos presos ao corpo [2], [3].

No mercado atualmente, existem vários sensores mmWave com diversas aplicações. Entre eles, destacam-se os sensores da linha MR60 da Seeed Studio, como o MR60BHA2, voltado à detecção de sinais vitais, e o MR60FDA2, projetado para a detecção de quedas, com um preço mais acessível [4], [5]. Além destes, fabricantes como Texas Instruments e Infineon oferecem soluções destinadas a aplicações industriais, como o IWR6843ISK, da Texas Instruments, que opera entre 60 e 64 GHz e possui quatro canais de recepção e três de transmissão, com campo de visão de 120° em azimute e 30° em elevação [6].

Apesar da variedade de sensores disponíveis e das aplicações já descritas na literatura, a avaliação prática desses dispositivos em cenários biomédicos ainda exige cuidado, especialmente em condições próximas ao uso real. Em aplicações hospitalares e de *homecare*, fatores como posicionamento do sensor, distância em relação ao usuário, presença de obstáculos, umidade, ventilação, sinais eletromagnéticos e movimentos involuntários podem interferir nos dados coletados e comprometer a confiabilidade das medições. Por isso, estudos experimentais com diferentes condições de interferência são importantes para compreender os limites de uso desses sensores e indicar formas mais adequadas de instalação e operação [7], [8], [9].

Dessa forma, esta pesquisa tem como objetivo realizar um estudo teórico-experimental sobre o funcionamento e as

Fundação de Amparo a Pesquisa de Minas Gerais (Fapemig)

possíveis aplicações dos sensores mmWave na Engenharia Biomédica. Para os testes práticos, foram utilizados os sensores MR60BHA2 [4], aplicado ao monitoramento de sinais vitais, e MR60FDA2 [5], utilizado na detecção de quedas e presença humana.

II. METODOLOGIA

Primeiramente, foram levantados artigos sobre o funcionamento e aplicação dos mmWaves e dos sensores disponíveis no mercado, com base em palavras-chaves como: *milimeter waves*, *60 GHz*, *Radar*, *medical applications*. Também foi pesquisado sobre possíveis interferências que os sensores poderiam sofrer.

Para a escolha dos sensores foram considerados a facilidade de uso, documentação e custo. Assim, foram selecionados os sensores MR60BHA2 [4] e o MR60FDA2 [5] da Seeed Studio. Estes sensores são integrados ao microcontrolador ESP32 e utilizam da técnica FMCW para realizar as medidas dos parâmetros. Para o desenvolvimento foi utilizado a linguagem C e a IDE do Arduino com a biblioteca da Seeed Studio disponibilizada pelo Github [4] [5], juntamente com a linguagem de programação Python responsável por coletar os dados do terminal do Arduino IDE, converte-los em um arquivo .csv e gerar um gráfico para análise. Por fim, testes práticos preliminares foram realizados para explorar o uso dos sensores. Ressalta-se que todos os testes foram realizados apenas para validação técnica preliminar, sem coleta dados sensíveis ou intervenção clínica, e apenas os autores participaram voluntariamente e consentiram com o estudo. As metodologias adotadas para cada sensor são detalhadas a seguir, e um resumo geral encontra-se na Tabela I.

TABELA I
RESUMO DOS TESTES EXPERIMENTAIS.

Teste	**Condição experimental avaliada**
Sensor MR60BHA2	
Ambiente normal	Condição de referência, sem interferências intencionais
Frequência respiratória (RPM)	Comparação com ventilador pulmonar em 16 RPM
Frequência cardíaca (BPM)	Comparação com monitor multiparamétrico
Umidade	Presença de vapor d'água entre o sensor e o voluntário
Interferência de Wi-Fi	Operação na presença de roteador Wi-Fi em diferentes distâncias
Obstáculos	Presença de superfícies de vidro ou acrílico entre o sensor e o voluntário
Camadas de roupa	Aquisição com duas camadas de roupa
Posição e ângulo	Avaliação em diferentes posições e ângulos em relação ao sensor
Sensor MR60FDA2	
Detecção de presença	Entrada e saída do indivíduo no campo de detecção
Detecção de queda	Queda simulada e posterior retorno à posição inicial
Posição corporal	Quedas simuladas em diferentes posições corporais
Movimentos semelhantes à queda	Movimentos bruscos de sentar e ajoelhar
Passagem em alta velocidade	Deslocamento rápido do indivíduo pelo campo de detecção
Sensibilidade e limiar de detecção	Avaliação sob diferentes configurações de sensibilidade e condições de ventilação
Objetos metálicos	Presença de objetos metálicos próximos ao corpo
Interferência entre sensores	Avaliação da influência do MR60FDA2 sobre o MR60BHA2 em diferentes distâncias e ângulos

A. MMwave na avaliação de sinais vitais

O sensor *MR60BHA2* foi empregado para investigar a viabilidade do monitoramento sem contato de sinais vitais, especificamente a frequência respiratória e a frequência cardíaca. O dispositivo fornece duas categorias de dados: a primeira, relacionada à respiração, apresenta o sinal da fase respiratória, correspondente às variações de inspiração e expiração, em formato gráfico, além do valor numérico de respirações por minuto (RPM); a segunda, referente à atividade cardíaca, retorna diretamente os valores de batimentos por minuto (BPM).

Para validar as estimativas fornecidas pelo sensor, foram utilizados equipamentos médicos de referência. A frequência cardíaca foi comparada às medições obtidas por um monitor multiparamétrico *LiFEtouch 10* (Lifemed) [10], enquanto a frequência respiratória foi validada por meio de um ventilador pulmonar *FlexiMag* (Magnamed) [11], configurado para 16 RPM. O ventilador foi acoplado a um pulmão artificial *Maquet TestLung 6006832* [13], instalado em uma incubadora neonatal *Line 4* (Olidef Medical) [12], de modo a simular um ambiente controlado e padronizado.

A.1) *Ambiente Normal*: Com o objetivo de estabelecer valores de referência para leitura da fase respiratória em ambiente sem interferências, o sensor foi posicionado a 0,5m de distância de um dos autores. Em paralelo, realizou-se a mesma medição em um pulmão artificial instalado na incubadora neonatal. Essa configuração permitiu comparar os resultados obtidos em condições reais com aqueles simulados em ambiente padronizado.

A.2) *Teste de RPM*: Para validar a frequência respiratória, utilizou-se a incubadora neonatal *Line 4*, equipada com pulmão artificial *Maquet TestLung*, conectado ao ventilador *FlexiMag*. O sistema foi empregado para simular a respiração de um recém-nascido e avaliar a viabilidade da utilização do sensor no monitoramento respiratório sem contato. O sensor foi fixado no teto da cúpula. O ventilador foi configurado para 16 RPM e foram realizadas cinco medições. Os resultados foram analisados pelo método de Bland–Altman, considerando a diferença entre os valores medidos pelo sensor e o valor de referência de 16 RPM.

A.3) *Teste de BPM*: Este teste buscou validar a medição da frequência cardíaca sem contato físico. A posição ideal foi definida conforme as instruções do manual do dispositivo [5], que recomendam distância entre 0,5 e 1,5 m na altura do tórax (Fig. 1). Para comparação, um dos autores foi conectado simultaneamente ao eletrocardiograma do monitor multiparamétrico

LiFEtouch 10 [10], permitindo a aquisição síncrona dos dados pelos dois equipamentos. Os dois equipamentos realizaram aquisições simultâneas, sendo realizadas cinco medições, cada uma com duração mínima de um minuto, repetidas de três a cinco vezes. Os resultados foram analisados pelo método de Bland–Altman, considerando a diferença entre os valores obtidos pelo sensor e pelo monitor multiparamétrico.

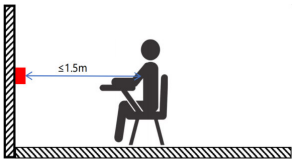


Fig. 1. Posicionamento do sensor MR60BHA2 para medição de sinais vitais.

A.4) *Teste de Umidade*: Uma das aplicações para o mmWave é avaliação dos sinais em cenários de *homecare* onde o uso dispositivos vestíveis de monitoramento nem sempre é possível, como por exemplo, em chuveiros. Para simular esses cenários de *homecare*, um dos autores posicionou-se diante do sensor enquanto uma panela com água em ebulição foi colocada entre ambos, reproduzindo a presença de vapor d'água e permitindo observar seus efeitos sobre a qualidade do sinal da fase respiratória.

A.5) *Teste de sinal Wi-Fi*: Considerando que ambientes domésticos geralmente possuem roteadores em funcionamento, avaliou-se a influência desse tipo de sinal no desempenho do sensor [16]. Foi utilizado um roteador *TP-Link EC220-G5* [14], operando em 2.4 GHz e 5 GHz. O dispositivo foi inicialmente desligado e ligado após 30 segundos de aquisição, e posteriormente foram realizados ensaios em diferentes distâncias (0,1 m, 0,6 m e 1,0 m) para verificar a interferência sobre o sinal de fase respiratória.

A.6) *Teste com vidro e acrílico*: Para avaliar a influência de obstáculos físicos na propagação e leitura dos sinais, foram posicionadas placas de acrílico e vidro entre o sensor e o voluntário durante a captação da frequência respiratória.

A.7) *Teste de camadas de roupa*: Para analisar o impacto de múltiplas camadas de vestimenta, um dos autores utilizou duas blusas sobrepostas. As medições foram realizadas a 0,5 m e 1,0 m de distância, verificando se o aumento da espessura do tecido compromete a precisão das leituras da fase respiratória.

A.8) *Teste de diferentes posições*: Este teste avaliou a influência da orientação corporal em relação ao sensor. Um dos autores foi posicionado a 45°, de perfil e de costas, e os resultados foram comparados à posição frontal, utilizada como referência. A análise permitiu verificar como diferentes ângulos e posturas afetam a qualidade do sinal respiratório detectado pelo *MR60BHA2*.

B. MMWave na detecção de quedas

Além dos testes realizados com o sensor MR60BHA2, foram realizados estudos práticos com o sensor MR60FDA2, utilizado para detecção de quedas e presença estática. Os testes foram conduzidos com o sensor posicionado na vertical, conforme mostra a Fig. 2, e seguiram uma metodologia padronizada, com tempos fixos e altura de instalação predefinida. O campo de detecção considerado foi de 3 m × 3 m × 3 m, conforme recomendado pelo fabricante. Os experimentos foram repetidos três vezes por condição, a fim de aumentar a confiabilidade dos resultados.

A metodologia estabelecida nos testes consistiu na entrada do usuário no campo de visão do sensor aos 5 s, permanência em posição ereta até os 10 s, seguida da simulação de queda, permanência no solo debaixo do sensor até os 40 s e posterior retorno à posição inicial, seguido da saída do campo de detecção aos 60 s. Os tempos predefinidos nessas etapas foram mantidos constantes em todos os experimentos com o sensor MR60FDA2.

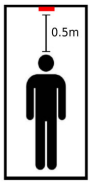


Fig. 2. Posicionamento do sensor MR60FDA2 para detecção de quedas.

As medições para avaliar o desempenho dos testes foram realizadas com auxílio de um cronômetro, ao final foi realizado o cálculo da média dos valores obtidos nos testes. Para a análise desses parâmetros, foram gerados gráficos temporais com o objetivo de representar o comportamento da detecção ao longo de cada experimento. Nesses gráficos, o eixo horizontal corresponde ao tempo em segundos, enquanto as marcações verticais indicam os principais eventos observados durante os testes, como a entrada do indivíduo no campo de detecção, a simulação da queda, o retorno à posição inicial e a saída do campo de visão do sensor.

Foram analisados dois tipos principais de gráfico: o de presença humana, utilizado para verificar o tempo de entrada e saída do indivíduo no campo de detecção, e o de estado humano, utilizado para avaliar a identificação da queda e o retorno ao estado normal. A partir dessas marcações, foram calculados os intervalos entre o evento realizado e a resposta apresentada pelo sensor. Mantida a metodologia padronizada, foram analisados os seguintes parâmetros e condições de teste:

B.1) *Tempo médio de detecção de presença*: Foram analisados os tempos entre a entrada do indivíduo no campo de visão do sensor e o momento em que a presença foi identificada, bem como o tempo necessário para cessar a detecção após a saída do indivíduo do raio de visão do sensor.

B.2) *Tempo médio de detecção de queda e retorno ao estado normal*: foram registrados o tempo para acusar uma queda e o intervalo que o sensor levou para reconhecer que a pessoa havia se levantado após a queda, ou seja, o retorno do indivíduo ao estado normal.

B.3) *Influência da posição corporal na detecção de queda*: foram simuladas quedas reais controladas nos cenários frontal, dorsal e lateral, a fim de verificar a influência da orientação do corpo na detecção. Também foram executados movimentos cotidianos parecidos com uma queda, como sentar-se bruscamente com a coluna de maneira errada e ajoelhar-se no chão de maneira agressiva, com o objetivo de verificar a ocorrência de falsos positivos.

B.4) *Teste passando correndo*: travessia rápida do indivíduo pelo campo de visão do sensor, passando correndo abaixo do dispositivo, a fim de avaliar a detecção de movimentos de curta duração e alta velocidade.

B.5) *Testes com a sensibilidade e threshold*: variação dos parâmetros do próprio sensor, incluindo diferentes graus de sensibilidade e valores de limiar de detecção (threshold), mantendo o ambiente experimental padronizado. Para avaliar a interferência externa do ambiente, foi utilizada ventilação forçada em diferentes intensidades, com o intuito de observar possíveis alterações nos dados coletados pelo sensor. Os dados coletados em formato csv foram organizados em uma tabela comparativa no Excel, permitindo comparar as configurações principalmente em relação à ocorrência de atrasos na detecção, falhas de leitura ou resultados indevidos.

B.6) *Testes com objetos metálicos*: inserção de objetos metálicos, sendo eles uma caixa metálica e uma chave de fenda, sobre o corpo do indivíduo, a fim de avaliar se superfícies reflexivas poderiam interferir no campo de visão do sensor.

B.7) *Interferência entre sensores*: o MR60FDA2 foi posicionado em direção ao MR60BHA2, e os testes foram repetidos variando a distância e o ângulo entre os sensores, com o objetivo de verificar possíveis alterações no desempenho da detecção de dispositivos operando na mesma faixa de frequência.

III. Resultados e Discussão

Os resultados da pesquisa bibliográfica mostram o potencial dos sensores mmWave. O artigo dos autores A. Soumya, C. et al. [1], mostra que os avanços recentes na tecnologia dos mmWaves podem ser aplicados em diversas áreas como: automação, robótica, industrial, saúde, segurança, transporte civil e machine learning. Outra pesquisa de Abdullah K. Alhazmi et al. [3], aborda sobre o uso de mmWaves para detecção humana, com destino para ambientes *homecare*, demonstrando que os sensores são capazes de detectar cinco estados de uma pessoa: deitada, parada, andando, sentando-se e caindo em tempo real. Além de mostrarem modelos de sensores que podem ser aplicados na área biomédica, como os da Texas Instruments. E a publicação de Honglin et al. [2], descreve brevemente as aplicações destes sensores na medicina, destacando a medição de sinais vitais e quedas, e, explica possíveis riscos biológicos para a saúde, como genotoxidade e efeitos biotérmicos de aumento na temperatura das moléculas de água na pele e córnea.

A. Testes do sensor MR60BHA2

A.1) *Ambiente Normal*: Os resultados mostram que a fase da respiração fica entre valores de -10 a +10, conforme mostra a Fig. 3, representando a expiração com valores negativos e inspiração sendo valores positivos, com leves interferências devido à pequenos movimentos da própria pessoa, ou fatores naturais do ambiente (pequenas correntes de ar, pessoas, ondas de sinais telefônicos e Wi-Fi etc.).

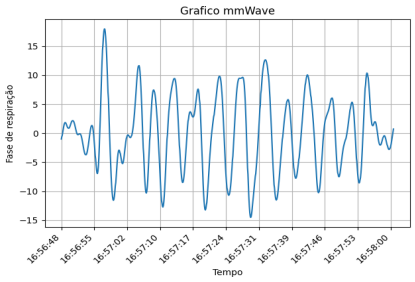


Fig. 3. Sinal da fase respiratória obtido pelo sensor MR60BHA2.

O resultado desse teste, mostrou-se muito próximo dos valores medidos com o pulmão de teste no interior da incubadora, como mostra a Fig. 4, ambiente com menor interferência. Dessa forma, concluiu-se que esses são valores padrões de um processo de respiração.

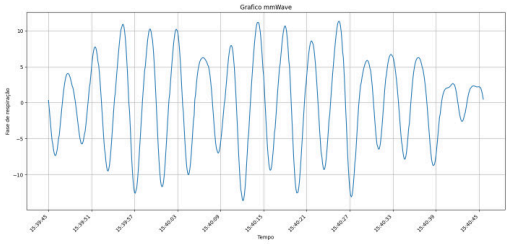


Fig. 4. Sinal da fase respiratória no ensaio com o pulmão artificial TestLung no interior da incubadora neonatal.

A.2) *Teste de RPM*: Para determinar a taxa de respiração por minuto o sensor MR60BHA2 utiliza do príncipio de FMCW, o algoritmo interno do sensor processa os deslocamentos do tórax, a fim de identificar a periodicidade da respiração, e converte a taxa de movimento em RPM. Os resultados obtidos pelo sensor MR60BHA2 foram comparados com a referência de 16 RPM por meio da análise de Bland–Altman. O resultado é apresentado na Figura 5.

O resultado obtido nos testes de respiração apresentou pequena variação em relação ao valor de referência. A análise de Bland–Altman indicou que aproximadamente 95% das diferenças entre as medições do MR60BHA2 e a referência de 16 RPM estiveram entre $-1,13$ RPM e $+2,43$ RPM, com viés médio de $0,65$ RPM. Esse viés positivo indica uma tendência do sensor a superestimar ligeiramente a frequência respiratória em relação ao valor de referência. Apesar dessa diferença, a variação observada foi pequena, indicando boa

concordância entre as medições do sensor e a frequência respiratória configurada no pulmão artificial nas condições avaliadas.

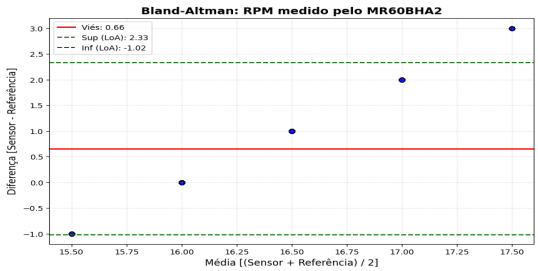


Fig. 5. Gráfico de Bland–Altman comparando a frequência respiratória medida pelo sensor MR60BHA2 com a referência fixa de 16 RPM.

A.3) *Teste de BPM*:

Para a frequência cardíaca, os resultados obtidos pelo sensor MR60BHA2 foram comparados com as medições realizadas pelo monitor multiparamétrico LiFEtouch 10 por meio da análise de Bland–Altman. O resultado é apresentado na Fig. 6.

Observa-se que o sensor MR60BHA2 apresentou uma média com valores superiores aos medidos pelo monitor multiparamétrico em todos os ensaios. Com a média de erro sendo +4,62 BPM, variando entre +23,57 e -14,32 BPM. Esse comportamento está de acordo com a documentação disponibilizada pela Seeed Studio [4], que informa que a estimativa de frequência cardíaca apresenta menor precisão quando o indivíduo não está em repouso ou dormindo.

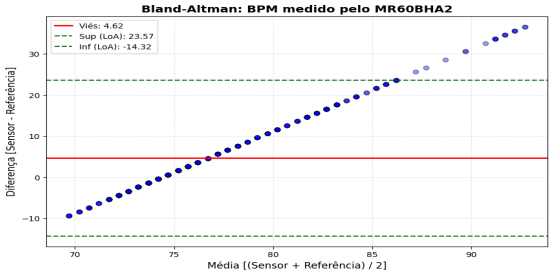


Fig. 6. Gráfico de Bland-Altman comparando a frequência cardíaca (BPM) medida pelo sensor MR60BHA2 com a referência fixa do equipamento (74,3 BPM).

A.4) *Teste de Umidade*:

No experimento realizado, foi demonstrado que o vapor d'água atrapalha consideravelmente na leitura dos dados, evidenciou-se perdas de amplitude, sendo a consequência do sinal percorrido no vapor, conforme mostra a Fig. 7.

Esse resultado é respaldado pela literatura, já que a umidade é um grande exemplo de interferência, como foi citado no artigo [15], o vapor de água absorve a energia das ondas vindas do sensor.

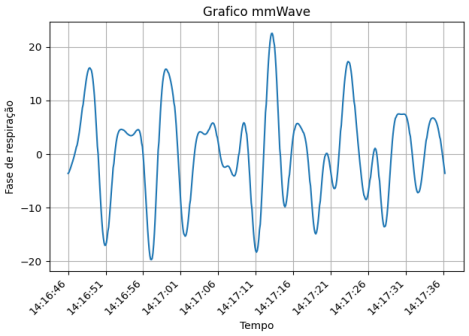


Fig. 7. Sinal da fase respiratória obtido na presença de vapor d'água entre o sensor e o alvo.

A.5) *Teste de sinal Wi-Fi*:

A Fig. 8 apresenta o comportamento do sinal da fase respiratória antes e após o acionamento do roteador Wi-Fi após 30 segundos. Observa-se que, após sua ativação, indicado pela linha vermelha, ocorreram interrupções mais frequentes e perda de amplitude no sinal, indicando a influência da interferência eletromagnética sobre as medições.

Os ensaios realizados com diferentes distâncias mostraram que a interferência foi mais pronunciada quando o sensor estava muito próximo ao roteador. À medida que a distância aumentou, observou-se redução da degradação e perdas do sinal. Nos testes realizados, uma separação mínima de aproximadamente 1 m foi suficiente para minimizar os efeitos da interferência, diminuindo as perdas de amplitude, indicado na Fig. 9, portanto, essa distância é adequada para a operação do sensor em ambientes com redes Wi-Fi.

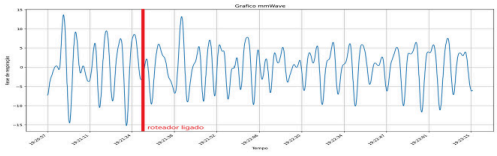


Fig. 8. Sinal da fase respiratória antes e após o acionamento do roteador Wi-Fi.

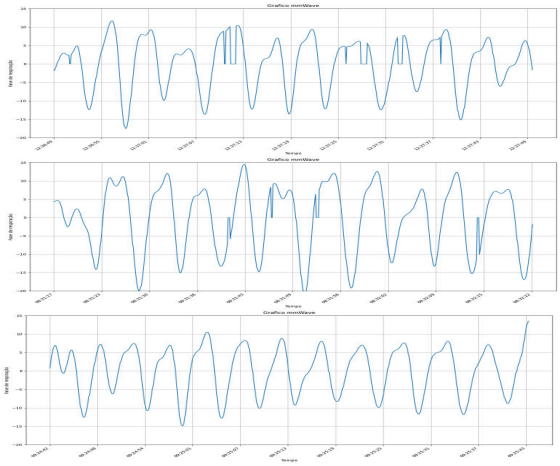


Fig. 9. Sinais da fase respiratória obtidos com o sensor posicionado a 0,1 m, 0,6 m e 1,0 m do roteador Wi-Fi, respectivamente.

A.6) *Teste com vidro e acrílico*: Avaliando a influência de materiais, como acrílico e vidro, evidenciado na Fig. 10 na qualidade do sinal da fase respiratória.

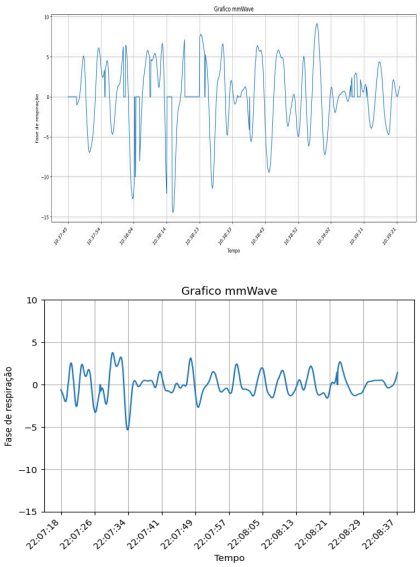


Fig. 10. Sinal da fase respiratória obtido com uma placa de acrílico e vidro (respectivamente) posicionada entre o sensor e o voluntário.

Observa-se que ambos os materiais provocaram atenuação significativa do sinal, reduzindo sua amplitude em relação à condição de referência. Mesmo com espessura de 0,5 cm, as placas interferiram na propagação das ondas, dificultando a detecção das variações associadas à respiração. Esse comportamento está de acordo com o descrito na literatura [17], que demonstra que materiais como vidro, madeira e metais geram interferências nos sensores mmWave.

A.7) *Teste de camadas de roupa*:

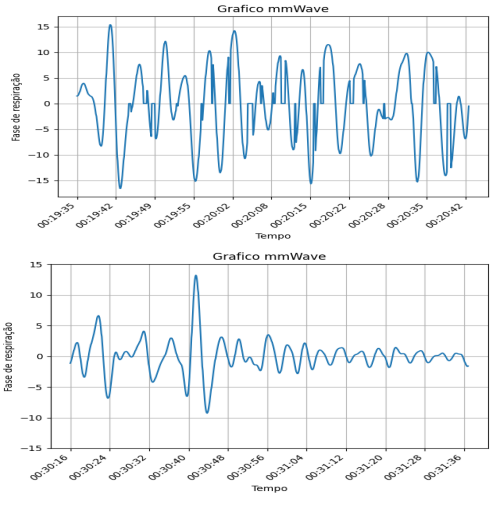


Fig. 11. Sinal da fase respiratória obtido com o voluntário utilizando duas blusas, a uma distância de 0,5 m e 1m do sensor, respectivamente.

Observou-se que a utilização de duas blusas provocou atenuação do sinal da fase respiratória, reduzindo sua amplitude em relação à condição de referência. Esse resultado indica que o aumento da espessura das camadas de tecido entre o sensor e o tórax compromete a qualidade do sinal recebido pelo MR60BHA2. Verificou-se também que a combinação entre o aumento da distância do sensor em relação ao indivíduo e a presença das camadas de roupa, intensifica a atenuação do sinal, resultando na diminuição da amplitude da fase respiratória, como mostra a Fig. 11.

A.8) *Teste de diferentes posições do voluntário*:

Os resultados da Fig. 12 mostraram que a posição a 45° apresentou comportamento semelhante ao da posição frontal, permitindo a identificação do sinal respiratório. Em contrapartida, as posições de perfil e de costas resultaram em maior atenuação do sinal e dificultaram a identificação das variações associadas à respiração, indicando que a orientação do voluntário influencia a qualidade das medições realizadas pelo MR60BHA2.

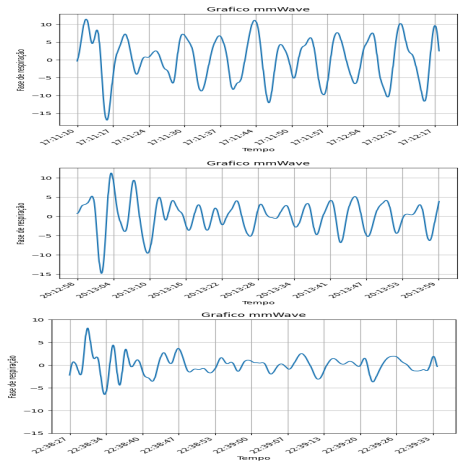


Fig. 12. Sinais da fase respiratória obtidos com o voluntário posicionado a 45°, de perfil e de costas em relação ao sensor, respectivamente.

B. Testes do sensor MR60FDA2

A Fig. 13 apresenta um exemplo dos gráficos temporais utilizados nessa análise.

B.1) *Tempo médio de detecção de presença*: Nos testes realizados, o sensor apresentou resposta rápida para detecção de presença humana, com tempo médio de aproximadamente 1,55 segundos. Já, o tempo necessário para interromper a detecção após a saída do indivíduo do campo de visão foi maior, com média de 18,98 segundos. Ainda assim, esse comportamento deve ser considerado em aplicações domiciliares ou hospitalares que dependem de resposta rápida, principalmente em sistemas de presença em tempo real.

B.2) *Tempo médio de detecção de queda e retorno ao estado normal*: Tempo médio de detecção de queda e retorno ao estado normal: Nos testes realizados, esse processo apresentou tempo médio de 20,48 segundos. Os resultados mostram que o MR60FDA2 conseguiu identificar a queda, mas apresentou maior atraso para reconhecer a mudança de estado após o

usuário voltar à posição inicial. Assim, o ensaio evidenciou que a detecção da queda ocorreu, porém a atualização posterior do estado do usuário não foi imediata.

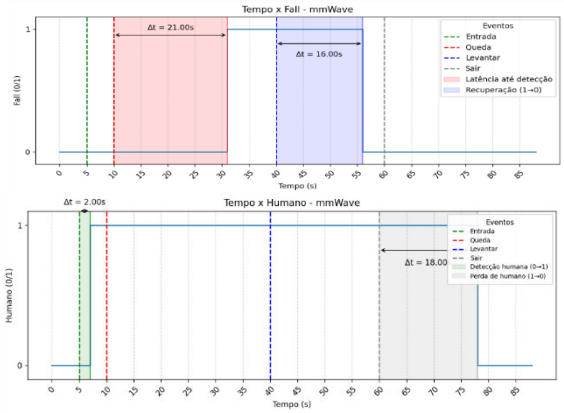


Fig. 13. Gráficos temporais do sensor MR60FDA2 durante teste de queda frontal.

B.3) *Influência da posição corporal na detecção de queda*: Nos testes de quedas reais controladas, o sensor apresentou mesmas detecções nos cenários frontal, dorsal e lateral, com ocorrência de detecção em todos os testes realizados. Esse comportamento está de acordo com os estudos que demonstram a capacidade de sistemas mmWave de reconhecer atividades humanas e detectar queda em tempo real [3] .Os resultados indicam que o MR60FDA2 foi capaz de reconhecer padrões típicos de queda quando há mudança rápida de estado e variações de alturas em relação ao indivíduo e o sensor posicionado na vertical. Dessa maneira conclui-se que a posição não alterou nos resultados coletados.

No entanto, ao realizar movimentos parecidos com uma queda, observou-se que os movimentos cotidianos avaliados podem gerar falsos positivos. Esses casos particulares são importantes para aplicações reais, pois ações comuns no dia a dia podem gerar um alarme falso, assim reduzindo a confiabilidade do sensor se não houver um ajuste preciso nas suas configurações ou uma integração com uma lógica adicional de validação, por exemplo, utilizado inteligência artificial.

B.4) *Teste passando correndo*: No cenário em que o indivíduo atravessava rapidamente o campo de visão do sensor, não houve detecção de presença pelo sensor. Esse comportamento indica uma possível limitação do MR60FDA2 para reconhecer movimentos de curta duração e alta velocidade, especialmente quando o indivíduo permanece pouco tempo dentro da área monitorada. Dessa forma, sua aplicação em ambientes com circulação intensa ou movimentações rápidas deve considerar essa condição como um possível fator de falha na detecção.

B.5) *Testes com a sensibilidade e threshold*: Os parâmetros de threshold e sensibilidade foram variados por serem configurações disponíveis no MR60FDA2 e diretamente relacionados ao processo de detecção de quedas [5]. Conforme apresentado nas Figs. 14 e 15, com ventilação em baixa potência, threshold 0,9 e sensibilidade 20, o sensor apresentou latência de 22,00 s para detecção da queda, 16,00 s para retorno ao estado normal, 2,00 s para detecção de presença e 19,00 s para cessar a detecção após a saída do indivíduo. Na Fig. 14, com ventilação em alta potência, threshold 1,0 e sensibilidade 20, foram observados 16,00 s para detecção da queda, 26,00 s para retorno ao estado normal, 2,00 s para detecção de presença e 6,00 s para cessar a detecção. Esses resultados indicam alteração nos tempos de resposta do MR60FDA2 entre as duas condições, principalmente no retorno ao estado normal, que aumentou de 16,00 s para 26,00 s com a ventilação em alta potência.

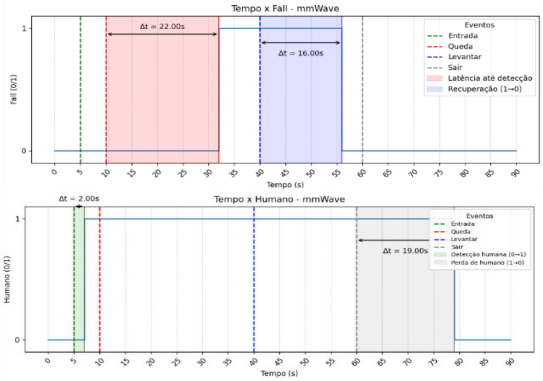


Fig. 14. Gráficos temporais do sensor MR60FDA2 com ventilação em baixa potência, threshold 0,9 e sensibilidade 20.

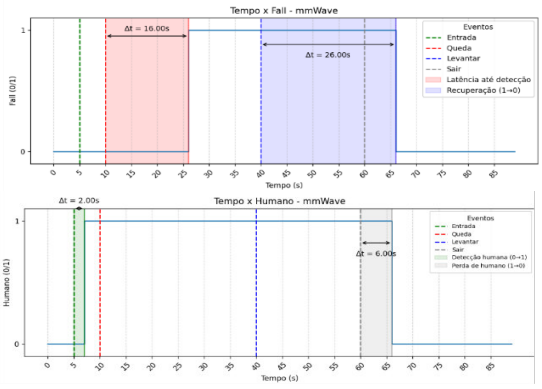


Fig. 15. Gráficos temporais do sensor MR60FDA2 com ventilação em alta potência, threshold 1,0 e sensibilidade 20.

B.6) *Testes com objetos metálicos*: Como o sensor depende da reflexão do sinal, foram observados poucos ruídos durante os testes; no entanto, quando os objetos metálicos estavam muito próximos do sensor, ocorreram alguns falsos positivos. Dessa forma, pode-se concluir que a interferência foi baixa nas condições avaliadas, mas a proximidade direta de objetos metálicos com o sensor deve ser evitada em aplicações reais.

De maneira geral, os testes realizados com o sensor MR60FDA2 indicam que o dispositivo foi capaz de detectar quedas e presença humana em cenários controlados,

compatível com o que foi pesquisado [5]. No entanto, seu desempenho e sua confiabilidade dependeram diretamente do tipo de movimento realizado, das condições do ambiente e da configuração dos parâmetros de sensibilidade e limiar de detecção. Movimentos rápidos ou semelhantes a quedas, como sentar-se bruscamente, ajoelhar-se de forma rápida ou passar correndo abaixo do sensor, podem gerar falsos positivos ou falhas de detecção. Dessa forma, a aplicação do MR60FDA2 em ambientes domiciliares ou hospitalares exige calibração adequada, padronização do local de instalação e cuidado com fatores externos, como ventilação direta e objetos metálicos muito próximos ao sensor.

IV. Conclusão

O objetivo desta pesquisa visou analisar a viabilidade da aplicação dos sensores mmWaves em contextos da Engenharia Biomédica. Uma das principais aplicações avaliadas foi o uso em ambientes *homecare*, por permitir monitoramento sem contato. Foram realizados testes com o sensor MR60BHA2, voltado à medição de sinais vitais, e com o sensor MR60FDA2, aplicado à detecção de presença humana e quedas. Os resultados demonstraram que o desempenho dos sensores pode ser comprometido em ambientes com elevada presença de interferências, como umidade, correntes de ar, sinais Wi-Fi ou obstáculos no campo de visão do sensor, pois essas condições podem aumentar as taxas de erro e a ocorrência de falsos positivos.

Entretanto, em um ambiente limpo e com menos interferências no sinal, é possível coletar dados com resultados mais próximos do real, com pequenos erros, causados por movimentos da própria pessoa sendo medida, principalmente com o MR60BHA2, que mede a respiração e batimentos de forma mais sensível.

Em relação à detecção de quedas, o sensor MR60FDA2 conseguiu identificar quedas frontais, dorsais e laterais nos testes realizados em ambiente controlado. Entretanto, movimentos cotidianos semelhantes a uma queda, como sentar-se ou ajoelhar-se bruscamente, podem provocar falsos positivos, enquanto movimentos muito rápidos podem não ser detectados. Dessa forma, embora o sensor apresente potencial para aplicações de monitoramento, seu uso exige calibração adequada e padronização do ambiente para aumentar a confiabilidade dos resultados.

Os resultados obtidos indicam que a tecnologia mmWave apresenta potencial para aplicações de monitoramento sem contato em engenharia biomédica e ambientes de *homecare*. Como trabalhos futuros, sugere-se o desenvolvimento de estratégias para mitigar as interferências observadas, além do aprimoramento dos algoritmos de processamento, visando aumentar a precisão das medições e a confiabilidade das estimativas dos sinais vitais.

REFERÊNCIAS

[1] A. Soumya, C. Krishna Mohan, Linga Reddy Cenkeramaddi, "Recent Advances in mmWave-Radar-Based Sensing, Its Applications, and Machine Learning Techniques: A Review", *MDPI*, vol. 23, issue 21, pp. 1-25, Nov 2023. DOI: https://doi.org/10.3390/s23218901.

[2] Honglin Wang, Lin Lu, Pengran Liu, Jiayao Zhang, Songxiang Liu, Yi Xie, Tongtong Huo, Hong Zhou, Mingdi Xue, Ying Fang, Jiaming Yang, Zhewei Ye, "Millimeter waves in medical applications: status and prospects", *ScienceDirect*, vol. 4, issue 1, pp. 16-21, Fev 2024. DOI: https://doi.org/10.1016/j.imed.2023.07.002.

[3] Abdullah K. Alhazmi, Mubarak A. Alanazi, Awwad H. Alshehry, Saleh M. Alshahry, Jennifer Jaszek, Cameron Djukic, Anna Brown, Kurt Jackson, Vamsy P. Chodavarapu, "Intelligent Millimeter-Wave System for Human Activity Monitoring for Telemedicine", *Sensors*, vol. 24, issue 1, pp. 1-23, Jan 2024. DOI: 10.3390/s24010268.

[4] Seeed Studio, *Getting started with 60GHz mmWave Breathing and Heartbeat Detection Sensor Kit with XIAO ESP32C6 (MR60BHA2)*, Shanghai, CN, 2024. Acessed: March 3, 2026. [Online]. Available: https://wiki.seeedstudio.com/getting_started_with_mr60bha2_mmwave_kit/.

[5] Seeed Studio, *Getting started with 60GHz mmWave Fall Detection Sensor Kit with XIAO ESP32C6 (MR60FDA2)*, Shanghai, CN, 2024. Acessed: March 3, 2026. [Online]. Available: https://wiki.seeedstudio.com/getting_started_with_mr60fda2_mmwave_kit/.

[6] Texas Instruments, *IWR6843, IWR6443 Single-Chip 60 to 64GHz mmWave Sensor*, Dallas, Texas, EUA. Accessed: March 3, 2026. [Online]. Available: https://www.ti.com/lit/ds/symlink/iwr6843.pdf.

[7] Nastassia Vysotskaya, Christoph Will, Lorenzo Servadei, Noah Maul, Christian Mandl, Merlin Nau, Jens Harnisch, Andreas Maier, "Continuous Non-Invasive Blood Pressure Measurement Using 60 GHz-Radar—A Feasibility Study", *MDPI*, vol. 23, issue 8, pp. 1-36, Abr 2023. DOI: 10.3390/s23084111.

[8] Yusra Banday, Ghulam Mohammad Rather, Gh. Rasool Begh, "Effect of atmospheric absorption on millimetre wave frequencies for 5G cellular networks", *IET Communications*, vol. 13, issue 3, pp. 265-270. DOI: https://doi.org/10.1049/iet-com.2018.5044

[9] Xu D, Yu W, Deng C, He ZS, "Non-Contact Detection of Vital Signs Based on Improved Adaptive EEMD Algorithm ", *Sensors (Basel)*, Ago 2022, 22(17):6423. DOI: 10.3390/s22176423.

[10] Lifemed, *Monitor de Sinais Vitais LifeTouch 10*, Pelotas, BR, 2021. Accessed: March 3, 2026. [Online]. Available: https://pt.scribd.com/document/515782915/Manual-Lifetouch-10.

[11] Magnamed, *Manual de Operações FlexiMag Max*, Cotia, BR, 2021. Accessed: March 3, 2026. [Online]. Available: https://pt.scribd.com/document/746979501/MANUAL-OPERACAO-VENTILADOR-FLEXIMAG-MAX.

[12] Olidef Medical, *Manuel Tecnico, para recem-nascidos*, Ribeirão Preto, BR. Accessed: March 3, 2026. [Online], Available: https://pt.scribd.com/document/562598532/Manual-Tecnico-Incubadora-SCTI-Line-4-V1.

[13] MAQUET, *Maquet 6006832 Lung Test Adult 190 1 Liter Instruction Manual*, Rastatt, DE, 1971. Accessed: March 3, 2026. [Online]. Available: https://manuals.plus/ae/1005009184800009.

[14] TP-Link, *Manual de instruções TP-Link EC220-G5*, California, USA, 2021. Acessed: March 3, 2026. [Online]. Available: https://www.manualpdf.com.br/tp-link/ec220-g5/manual

[15] Qinglang Dai, Yongzhi Huang, Lu Wang, Rukhsana Ruby, Kaishun Wu, "mm- Humidity: Fine-Grained Humidity Sensing with Millimeter Wave Signals", *IEEE*, pp. 204-211, Fev 2019. DOI: 10.1109/PADSW.2018.8644907.

[16] Shiyu Cheng, Kaveh Pahlavan, Haowen Wei, Zhuoran Su, Seyed Reza Zekavat, Ali Abedi, "A Study of Interference Analysis Between mmWave Radars and IEEE 802.11AD at 60 GHz Bands", *Int J Wireless Inf Networks*, vol. 29, pp. 222-231, Jun 2022. DOI: 10.1007/s10776-022-00564-9.

[17] G. Barb, F. Danuti, M. A. Ouamri and M. Otesteanu, "Analysis of Vegetation and Penetration Losses in 5G mmWave Communication Systems", *2022 International Symposium on Electronics and Telecommunications (ISETC)*, Timisoara, Romania, 2022, pp. 1-5, DOI: 10.1109/ISETC56213.2022.10009963.

Arranjo Experimental Utilizando Plano-Terra para Ensaios com Sensores Biotelemétricos Indutivos

Caio Marcelo de Miranda
Departamento Acadêmico de Eletrônica (DAELN-CT)
Centro de Pós Graduação em Engenharia Elétrica e Informática Industrial (CPGEI)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil.
caiomiranda@utfpr.edu.br

Luana Midori Saito
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil. luanasaito@alunos.utfpr.edu.br

Malu Hikari Ando Oyamaguchi
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil. malu@alunos.utfpr.edu.br

***Abstract*—Sensores indutivos, especialmente os com ausência de núcleo ferromagnético, são mais sensíveis à influência do corpo humano uma vez que as linhas de campo são mais dispersas e os valores concentrados de indutâcia e capacitância são menores. Desta forma, o ensaio e teste de tais sensores práticos depende da presença do corpo humano, o quê dificulta por vezes a execução e a repetibilidade das medidas. O objetivo deste trabalho é apresentar um arranjo utiizando um plano terra e uma bobina com núcleo de ar em autorressonância, de forma que ao impor diferetes condições de contorno no sistema, pode-se medir o módulo da impedância e a fase elétrica da bobina sensora em um analisador de impedâncias, demonstrando assim a viabilidade de se utilizer o arranjo em substituição inicial aos testes com corpo humano. Além disso, o efeito do aumento da capacitância influenciada por efeitos externos é avaliada. As medições confirmam a teoria proposta ao apresentar correlações maiores que 0,8 e erros relativos menores que 3% para os casos de interesse.**

Palavras chave— sensores indutivos, autorressonância, plano terra, biotelemetria

I. INTRODUÇÃO

A biotelemetria consiste na medição e monitoração de parâmetros fisiológicos a distância. De modo a causar a menor interferência possível na rotina e no indivíduo sob observação, dentro do campo da biotelemetria encontram-se técnicas de comunicação com a unidade sensora, como por exemplo, a técnica de comunicação passiva, na qual não necessita-se de fonte própria de alimentação acoplada ao sensor. Isto possibilita uma menor dimensão da unidade sensora, além de aumentar a vida útil do sensor e evitar o risco de danos a um paciente caso ocorra o vazamento do conteúdo químico da bateria. Grandezas como pressão [1], posição, deslocamento linear e angular são comumente medidas com este tipo de sensor, sendo que destes valores podem se extrair parâmetros tais como, elasticidade e deformações de tendões [2] , frequência cardíaca [3], pressões intra-oculares, dentre outras.

Um importante e relativamente novo tipo de sensor utilizado na Biotelemetria passiva é o sensor indutivo autorressonante [4]. O princípio de operação deste sensor é modular a frequência de autorressonância da bobina (f_o) de acordo com o parâmetro medido, e portanto não se faz necessária a adição de um capacitor para que se obtenha a ressonância. Desta maneira, este tipo de sensor pode possuir apenas um componente e possuir pequenas dimensões.

Em [5] foi proposto um sensor de deslocamento autorressonante, que utiliza a modulação da frequência de autorressonância de uma bobina com núcleo de ar, de acordo com a variação do seu comprimento longitudinal (H) ou espaçamento entre espiras. O sensor consiste em uma bobina com núcleo de ar enrolada em torno de um tubo de látex, fixada apenas nas suas extremidades (primeiras e últimas espiras). O tubo sofre a deformação alongando a bobina e variando a sua frequência de autorressonância. Ainda, apresentou-se uma possível aplicação para uso em pletismografia. Entretanto, notou-se que uma vez que o sensor se baseia em autorressonância, isto é, não possui capacitor externo para concentrar as linhas de campo elétrico, e além disso, utiliza núcleo de ar, tornando as linhas de campo magnético mais dispersas, a influência do corpo humano alterou sensivelmente a impedância e a frequência de ressonância do mesmo.

Neste estudo, um ensaio experimental é conduzido de modo a avaliar a possibilidade de se testar o sensor autorressonante por onda estacionária em bancada, sem a presença obrigatória do corpo humano, melhorando assim a repetitividade dos ensaios. Para isto, foi introduzido um plano terra nas proximidades do sensor com o intuito de simular os efeitos do corpo humano. Adicionalmente, outros testes foram feitos de maneira a observar os efeitos da presença de aterramento ligado ou não ao plano terra, e a sua consequente influência nas curvas de impedância, bem como uma avaliação da influência de corpos externos na capacitância concentrada e na frequência de ressonância do sensor.

II. DESENVOLVIMENTO

Bobinas com núcleo de ar e de comprimento longitudinal H relativamente longos, isto é, com uma relação

entre comprimento e diâmetro (D) em geral maior do que 1, podem apresentar frequência de ressonância por onda estacionária. Isto deve-se ao fato de que a bobina quando possui dimensões apreciáveis em relação ao comprimento de onda (λ) suporta a propagação de uma onda com velocidade de fase e comprimento de onda característicos ao longo do seu eixo longitudinal [6]. Neste caso a bobina passa a se comportar como um guia de onda, e quando em ressonância trata-se portanto de um ressonador. No caso de uma bobina com ambas as extremidades abertas ou fechadas (conectada a um gerador por exemplo), esta comporta-se como um ressonador de meia onda. Neste caso existem máximos e mínimos de tensão (V) ou corrente (I), dependendo das condições de contorno do sistema (extremidades abertas ou fechadas) [6]. Desta forma, devido a existência de variações apreciáveis de fase ao longo do próprio comprimento da bobina, um analisador de impedâncias pode registrar ângulos de fase maiores que 90^o, e a frequência de ressonância (fase igual 0^o) ocorre na inversão de fase de $+180^o$ para -180^o [6]. Este fenômeno pode ser observado na Figura1, que mostra o perfil de onda estacionária para uma ressonância de meia onda e a inversão de fase quando observadas as correntes das extremidades.

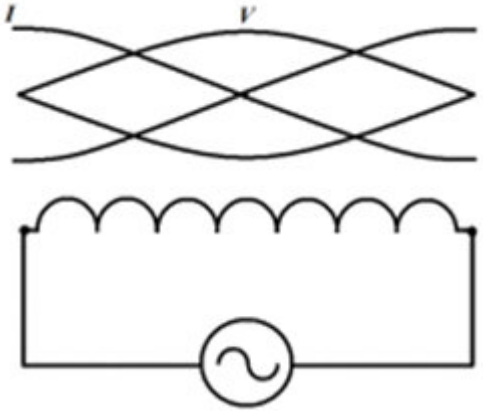


Figura 1. Bobina com ambas as extremidades conectadas a um gerador e os perfis de corrente (I) e tensão (V), para uma onda estacionária de meio comprimento de onda.

A bobina estudada [5] foi enrolada em um tubo de látex, utilizando fio de cobre esmaltado com diâmetro de 0,31 mm, 137 espiras, 55 mm de comprimento e diâmetro da bobina de aproximadamente 7 mm. Apenas as primeiras e últimas espiras são coladas ao tubo que se alongado aumenta o comprimento da bobina, modulando assim a frequência de autorressonância da mesma, conforme equação apresentada em [7].

Para aplicações de biotelemetria passiva e medições de parâmetros fisiológicos de seres vivos, a bobina proposta, por apresentar núcleo de ar, sofre consideráveis influências do corpo, ou parte dele, que pode atuar como um plano terra ao fechar as linhas de campo elétrico, por exemplo. Portanto, para auxiliar na observação controlada, em bancada, que influências externas podem imprimir no comportamento da impedância e fase da bobina sensora, um plano terra foi confeccionado utilizando uma base de madeira revestida com papel alumínio. Deste modo pode-se medir a impedância e fase da bobina ao aplicar diferentes condições de contorno sobre o arranjo bobina-plano. Sendo assim, o objetivo desse estudo é avaliar o arranjo desenvolvido, visando futuras aplicações da bobina como um sensor de deslocamento apresentado em [5] que pode ser testado em bancada simulando o efeito da proximidade do corpo humano, por exemplo. A Figura 2 mostra o plano terra e a bobina sensora sob estudo.

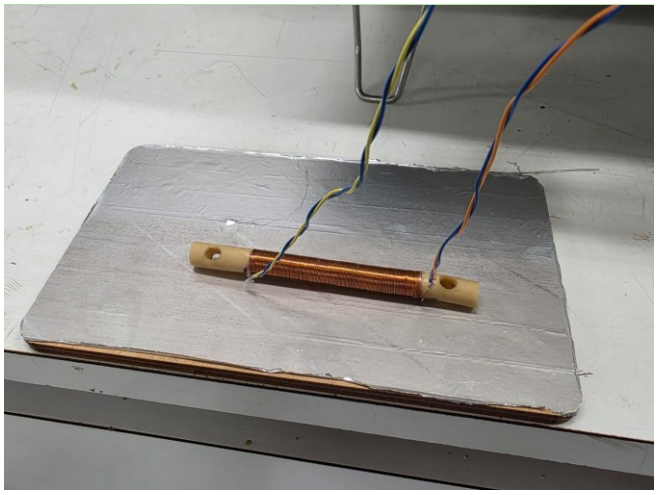


Figura 2. Bobina sensora enrolada em tubo de látex, e plano terra confeccionado.

III. Metodologia e Resultados

Neste estudo, foram conduzidas algumas configurações experimentais distintas do arranjo bobina-plano para medição do módulo da impedância (Z) e fase elétrica (θ) utilizando um analisador de impedância Agilent 4294A configurado para realizar 16 medições e apresentar o valor médio. Em todos os casos, a bobina não apresenta alongamento.

No primeiro caso a bobina foi posicionada sobre o plano de aterramento, sem que este fosse conectado ao condutor de proteção (terra) da tomada de energia, tal como mostrado na Figura 2. A frequência de ressonância medida foi de 55,90 MHz e a impedância de 234,7 Ω na ressonância. Os resultados são mostrados na Figura 3.

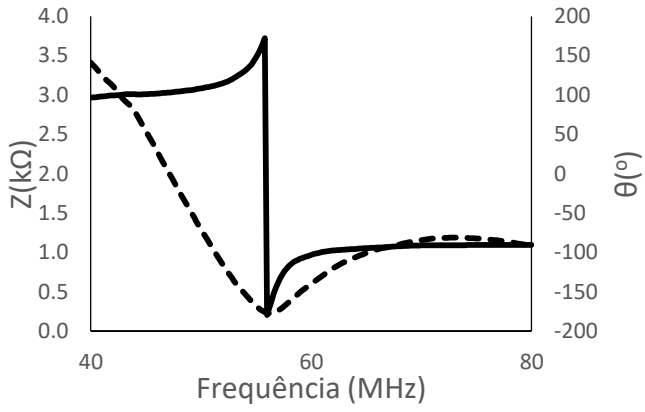


Figura 3. Módulo da impedância (linha sólida) e fase (tracejado) do sensor posicionado sobre o plano.

Na sequência a bobina continua posicionada sobre o plano, com a diferença de que este foi eletricamente conectado ao condutor de proteção, ou terceiro pino (terra), da tomada.de energia. A frequência de ressonância medida foi de 52,70 MHz e a impedância de 372,8 Ω na ressonância.

Resultando em uma diminuição de 3,2 MHz, conforme mostrado na Figura 4.

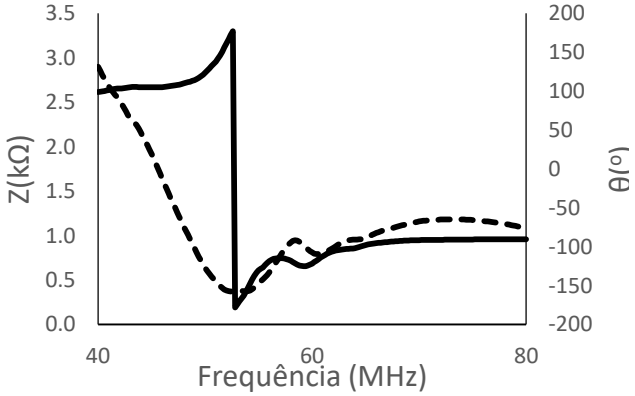


Figura 4. Módulo da impedância (linha sólida) e fase (tracejado) do sensor posicionado sobre o plano aterrado.

Na sequência a bobina foi posicionada diretamente sobre a mão de um dos operadores, sem utilização de plano terra, de modo a avaliar a validade da utilização do plano em substituição ao corpo. A frequência de ressonância medida foi de 54,30 MHz e impedância de 346,5 Ω na ressonância, conforme mostrado na Figura 5. Observa-se uma diferença de mesmo valor (1,6 MHz) para a frequência de ressonância da bobina sobre a mão quando comparada aos outros casos previamente mostrados nas Figuras 3 e 4 (55,9 MHz e 52,7 MHz), o que resulta em um erro relativo de 2,9%. Ainda, as correlações entre as curvas de fase para a bobina sobre a mão (Figura 5), em relação aos casos da bobina sobre o plano sem aterramento (Figura 3), e sobre o plano aterrado (Figura 4), foram de 0,819 e 0,815, respectivamente.

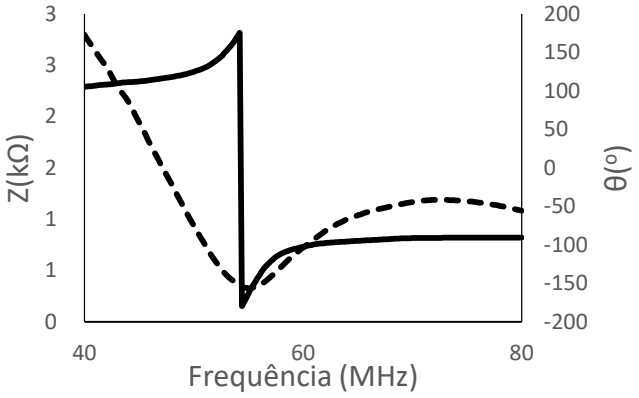


Figura 5. Módulo da impedância (linha sólida) e fase (tracejado) da bobina posicionada sobre a mão da operadora.

Como forma de ampliar o estudo foram ainda propostas outras configurações. A bobina foi posicionada sobre o plano de aterramento, mantendo-se agora a mão de um dos operadores em contato com o plano não aterrado, de modo que o corpo humano atuasse como referência de aterramento. Os resultados são mostrados na Figura 6. A frequência de ressonância medida foi de 24,30 MHz e impedância de 2,112 kΩ na ressonância. Quando comparado ao arranjo da Figura 2 (bobina sobre o plano sem aterramento), com frequência de ressonância de 55,9 MHz, nota-se uma diminuição de 2,3 vezes na frequência de ressonância quando a operadora toca o plano. Esta diferença é esperada, uma vez que, nesta configuração, existe uma diferença evidente de modelo elétrico (o corpo humano passa a ser referência para o plano terra). Adicionalmente, evidencia-se a baixa correlação entre as curvas impedância para este caso, porém este estudo foge ao escopo do presente trabalho.

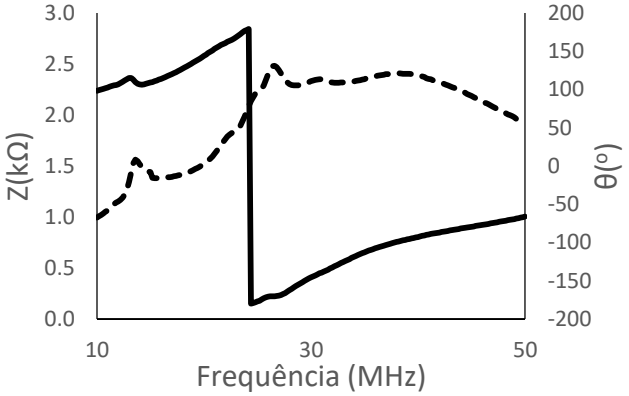


Figura 6. Módulo da impedância (linha sólida) e fase (tracejado) da bobina posicionada sobre a o plano enquanto a operadora toca no mesmo.

Foram propostos dois arranjos adicionais nos quais a bobina é posicionada sobre o plano, a mão da operadora se posiciona sobre a bobina, efetuando contato, sendo que em uma medição o plano não é aterrado (Figura 7) e na outra o plano é aterrado (Figura 8).

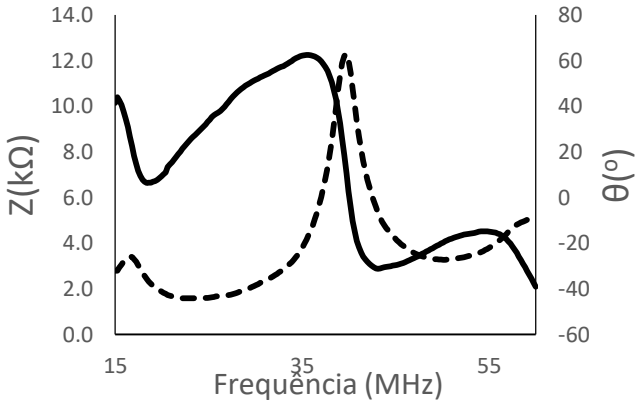


Figura 7. Módulo da impedância (linha sólida) e fase (tracejado) da bobina posicionada sobre o plano não aterrado sendo tocada pela operadora.

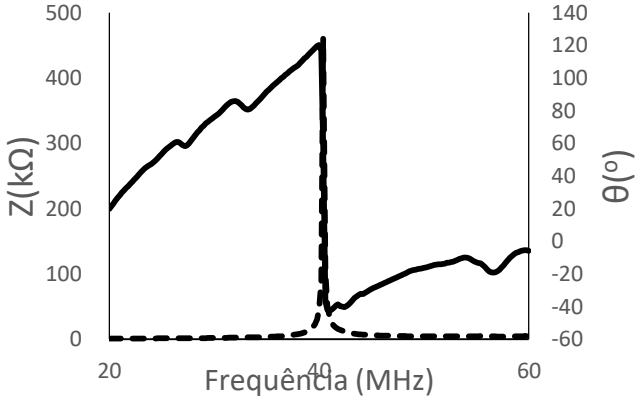


Figura 8. Módulo da impedância (linha sólida) e fase (tracejado) da bobina posicionada sobre o plano aterrado sendo tocada pela operadora.

Para o caso sem aterramento (Figura 7) a frequência de ressonância medida foi de 40,06 MHz e a impedância de 11,80 kΩ na ressonância. Enquanto que para o caso com aterramento (Figura 8) a frequência de ressonância medida foi de 40,52 MHz e impedância de 242,32 kΩ na ressonância. Nota-se portanto a pouca influência da presença ou não de ponto de aterramento no plano, em relação ao valor absoluto da frequência de ressonância (diferença menor que 1,2%). Entretanto, nota-se modificação nas curvas de fase e impedância. Outro ponto que pode ser destacado é a mudança do comportamento da impedância que agora possui máxima impedância na frequência de ressonância, aproximando mais o comportamento ao esperado para uma ressonância de um conjunto LC concentrado, com a diferença principal de que a fase ainda pode passar de noventa graus, ou seja, um caso intermediário. Isto deve-se ao fato de o corpo concentrar as linhas de campo elétrico entre espiras adjacentes e próximas ao mesmo, aumentando a capacitância do conjunto (o que é condizente com a diminuição da frequência de ressonância). De modo a confirmar esta justificativa do aumento da capacitância, outro teste realizado foi o de, para o mesmo arranjo anterior (plano não aterrado), afastar a mão da bobina (e não mais tocá-la como nas Figuras 7 e 8). Neste caso a frequência de ressonância aumentou para 51,46 MHz (encostando a mão, com plano não aterrado, o valor foi de 40,06 MHz conforme Figura 7). Este efeito se deu conforme esperado, uma vez que, ao se afastar a mão ocorre a diminuição da capacitância (alongamento das linhas de campo elétrico) e diminuição da frequência de ressonância. Ainda, o valor da impedância é de 456,8 Ω na ressonância. Os resultados são mostrados na Figura 9.

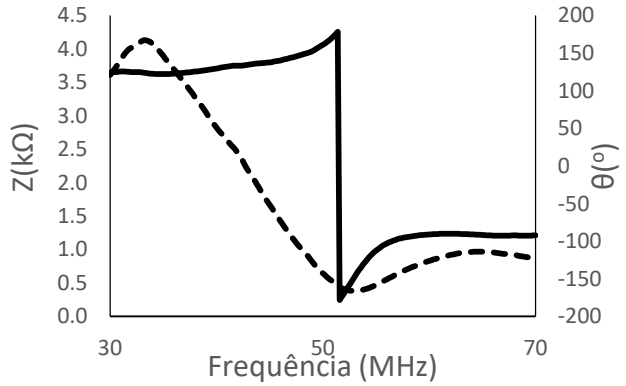


Figura 9. Módulo da impedância (linha sólida) e fase (tracejado) da bobina posicionada sobre o plano não aterrado com a mão da operadora sobre a bobina, sem efetuar contato direto com as espiras.

IV. Conclusões

Os testes e ensaios conduzidos levaram à melhor compreensão dos fenômenos envolvidos ao se utilizar sensores indutivos com núcleo de ar sob influências de corpos externos. Espera-se que este trabalho possa lançar luz sobre alguns efeitos que por vezes são relegados, ou não abordados na prática, por envolver campos eletromagnéticos de relativa complexidade. Especialmente, conforme resultados apresentados na seção anterior julga-se que a utilização de um plano terra pode auxiliar nos ensaios de bancada de sensores indutivos que irão operar próximos ou encostados ao corpo. Destacando que o erro relativo para o valor da frequência de ressonância do sensor no corpo (Figura 5), e do sensor sobre o plano, aterrado ou não (Figuras 3 e 4), foi menor que 3% e a correlação entre as curvas de fase (que é a curva de maior interesse para medição da frequência de ressonância devido ao cruzamento por 0^o) foi maior que 0,8. Adicionalmente, foi possível elucidar experimentalmente o efeito previsto de aumento da capacitância que ocorre ao se aproximar a mão da bobina, o que acarreta na diminuição do valor da frequência de ressonância, e modificação nas curvas de impedância, conforme demonstrado nas medições vistas nas Figuras 7, 8 e 9. Obviamente, as conclusões aqui apresentadas são limitadas à viabilidade preliminar de arranjos iniciais e testes de bancada, necessitando ainda a aplicação de outros métodos estatísticos para análise e avaliação completa dos resultados.

Referências

[1] M. Yang, Z. Y, P. -Y. Chen, "A Quantum-Inspired Biotelemetry System for Robust and Ultrasensitive Wireless Intracranial Pressure Monitoring," 2021 IEEE Sensors, Sydney, Australia, 2021, pp. 1-4, doi: 10.1109/SENSORS47087.2021.9639684.

[2] Pichorim, SF.; Abatti, PJ. Biotelemetric Passive Sensor Injected Within Tendon for Strain and Elasticity Measurement. IEEE Trans. on Biomed. Eng., v. 53, p.921-925. 2006.

[3] A. Khaliq, A. Waseem, M. F. Munir and R. Ahmad, "Comparison of adaptive noise cancelers for ECG signals in wireless biotelemetry system," 2016 International Conference on Intelligent Systems Engineering (ICISE), Islamabad, Pakistan, 2016, pp. 181-184, doi: 10.1109/INTELSE.2016.7475117.

[4] N. T. Tuan et al., "Deterministic Equation of Self-Resonant Structures for Normal-Mode Helical Antennas Implanted in a Human Body," in IEEE Antennas and Wireless Propagation Letters, vol. 17, no. 8, pp. 1377-1381, Aug. 2018, doi: 10.1109/LAWP.2018.2846600

[5] C. M. Miranda, S. F. Pichorim "Sensor Biotelemétrico de Deslocamento por Autorressonância de Onda Estacionária", em Congresso Brasileiro de Engenharia Biomédica (CBEB), Uberlândia, Brasil, 2014, pp. 173-176.

[6] C. M. de Miranda and S. F. Pichorim, "Self-resonant frequencies, standing waves, and impedance behavior of air-core helical solenoidal coil," 2015 International Conference on Electromagnetics in Advanced Applications (ICEAA), Turin, Italy, 2015, pp. 1296-1299, doi: 10.1109/ICEAA.2015.7297326.

[7] C.M. Miranda and S.F. Pichorim, "Self-resonant frequencies of air-core single-layer solenoid coils calculated by a simple method " Electr. Eng. (Berlin print), vol. 97, pp. 57–64, March 2015

BioDiff-CMOS: Pre-Silicon Design and Physical Verification of a CMOS Differential Block for Simulated ECG Acquisition

Danilo da Silva Cunha
CI Amazônia / Microelectronics
Manaus, AM, Brazil
dsc.elt@uea.edu.br

André Felipe Feitosa de Souza
CI Amazônia / Microelectronics
Manaus, AM, Brazil
andre.souza@ufam.edu.br

Fábio de Sousa Cardoso
PPGEE / Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
fcardoso@uea.edu.br

Isabela Rogério Cardoso
Electrical Engineering
State University of Amazonas
Manaus, AM, Brazil
irc.eng25@uea.edu.br

André Luiz Printes
PPGEE / Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
aprintes@uea.edu.br

Afonso Aguiar de Carvalho
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
aadc.ele21@uea.edu.br

Jozias Parente de Oliveira
PPGEE / Electronic Engineering
*State University of Amazon*as
Manaus, AM, Brazil
jpoliveira@uea.edu.br

Angilberto Muniz Ferreira Sobrinho
PPGEE / Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
asobrinho@uea.edu.br

Isaque Vilson Batista da Costa
PPGEE / Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
ivbdc.mee24@uea.edu.br

***Abstract*—This paper presents the pre-silicon design and physical verification of BioDiff-CMOS, a CMOS differential block intended for a simulated electrocardiogram (ECG) acquisition chain, implemented in the IHP SG13G2 130 nm BiCMOS process. The reproducible flow, based on Xschem, Ngspice, Magic, and Netgen, covers schematic capture, pre-layout simulation, layout, extraction, post-layout analysis, design rule check (DRC), and layout-versus-schematic (LVS) verification. Under nominal conditions of 1.8 V and 27 °C, the core achieved a differential gain of 0.908 V/V at 100 Hz, a nominal common-mode rejection ratio (CMRR) of 128.3 dB, a positive power-supply rejection ratio (PSRR+) of 48.7 dB, input-referred noise of 29.93 µVrms over 0.5–150 Hz, a cutoff frequency of 120.884 Hz, and power consumption of 36.01 µW. A five-point process-voltage-temperature (PVT) screening showed a gain at 1 kHz between 0.823 and 1.005 V/V, a cutoff frequency between 119.422 and 124.297 Hz, and power between 32.409 and 39.631 µW. DRC reported no violations, and LVS confirmed a unique match. Comparison with representative front-ends highlighted limitations in gain, noise, ideal biasing, and the absence of statistical mismatch analysis. The post-layout extraction contains 18 capacitances and no interconnect parasitic resistances, characterizing a capacitive/hybrid model. The circuit was not fabricated, and no real ECG acquisition was performed. The results establish a solid and reproducible technical basis for further project evolution and future experimental validation.**

***Keywords*—ECG, CMOS differential amplifier, analog front-end, pre-silicon design, open PDK, open-source EDA.**

I. INTRODUCTION

Biopotential signal monitoring is one of the foundations of modern biomedical instrumentation. Among these signals, the electrocardiogram (ECG) remains one of the most widely used tools for assessing cardiac electrical activity, ambulatory monitoring, and digital health applications. However, ECG acquisition poses relevant challenges to analog front-end design because the signals have low amplitude and are subject to power-line interference, electronic noise, motion artifacts, and variations in the electrode–skin interface [1], [2].

In this context, the analog front-end plays a decisive role in preserving useful information before digitization. The signal captured by the electrodes must be amplified, filtered, and conditioned to avoid saturation, resolution loss, and signal-to-noise ratio degradation. Recurring requirements include high input impedance, high common-mode rejection ratio (CMRR), low input-referred noise, low power consumption, bias stability, and robustness to process, voltage, temperature, and device mismatch variations [1]–[5].

The integration of these blocks in CMOS technology has contributed to the evolution of biomedical acquisition systems by enabling amplification, filtering, biasing, and output interfaces within the same technology while reducing area and power consumption. In parallel, the availability of open Process Design Kits (PDKs) and open-source Electronic Design Automation (EDA) tools has broadened access to integrated-circuit development in academic and research environments. A PDK provides electrical models, fabrication rules, device libraries, and technology parameters that establish the interface between electrical design and process constraints [6]. Tools such as Xschem, Ngspice, Magic, and Netgen support schematic

capture, simulation, layout, extraction, and DRC/LVS verification in an open analog-design flow [7]–[10], [12].

BioDiff-CMOS, used as a case study in this work, is a CMOS differential block developed in the IHP SG13G2 130 nm BiCMOS process using this open ecosystem. The process provides electrical models, devices, and geometric rules that enable the design to progress from the schematic level to physical implementation [6]. The focus of this paper is not the validation of a complete biomedical system, but the pre-silicon analysis of an analog block embedded in a simulated ECG acquisition chain.

The contribution of this study is to demonstrate, in a traceable and reproducible manner, the development and verification of this block through a pipeline integrating specification, schematic simulation, layout, extraction, post-layout analysis, and physical verification. Results are presented for gain, CMRR, PSRR, input-referred noise, power, cutoff frequency, area, PVT screening, DRC, and LVS, together with a comparison against representative front-ends from the literature. Thus, the objective is to relate ECG acquisition requirements to analog-design decisions and physical-implementation stages, while highlighting both the possibilities of the open-source flow and the current circuit limitations as a basis for further development and future experimental validation.

II. RELATED WORK

Table I compares BioDiff-CMOS with representative front-ends. Kim and Ko [3] and Zhao, Shang, and Lian [4] reported fabricated and measured circuits, whereas Laababid, El Khadiri, and Tahiri [5] and Wen et al. [2] reported simulation results. Units were standardized where possible, but the scopes remain different: some values refer to instrumentation amplifiers, others to more complete channels, and the reported areas do not always represent the same integration level.

The core presented in this work has sub-unity gain and is approximately 30 to 69 dB below the listed gains. The integrated noise is about 3.5 times the value reported by Wen et al. and more than 29 times the values measured by Kim and Ko and by Zhao et al. The nominal CMRR of 128.3 dB is high, but it was obtained with an ideal tail-current source, deterministic symmetry, and no statistical mismatch, a condition that favors this metric. The 36.01 µW power consumption lies within the range of the compared works, but it includes the simulated CMOS interface and excludes a physical bias circuit. Therefore, the comparison highlights performance gaps and scope differences without establishing a direct ranking among the solutions.

TABLE I — COMPARISON WITH REPRESENTATIVE ANALOG FRONT-ENDS FROM THE LITERATURE.

Work	Status, process, and VDD	Gain	CMRR	Input-referred noise	Power	Bandwidth, area, and notes
BioDiff-CMOS	Hybrid pre/post-layout; IHP 130 nm; 1.8 V	−0.835 dB at 100 Hz	128.3 dB at 100 Hz; nominal, no mismatch	29.93 µVrms, 0.5–150 Hz	36.01 µW, simulated chain	Cutoff = 120.884 Hz; core: 0.00093456 mm²; ideal tail-current source
Kim and Ko [3]	Fabricated and measured; CMOS 130 nm; 1.2 V	44.7 dB at 60 Hz	91.3 dB at 60 Hz	0.92 µVrms, 100 Hz bandwidth	6 µW, IA in PG+CS	100 Hz filter; IA area not reported; die = 6.02 mm²
Zhao et al. [4]	Fabricated and measured; CMOS 350 nm; 1.8 V	56–68 dB	76 dB	1.02 µVrms, 0.4–120 Hz	0.90 µW	0.4–120 Hz; 0.405 mm²
Laababid et al. [5]	Schematic simulation; CMOS 180 nm; 1.8 V	50.75 dB at 20 Hz	102 dB	1.47 µVrms, 0.1 Hz–1 kHz	1.08 µW	0.077–255 Hz in the paper body; area not reported
Wen et al. [2]	TT simulation; CMOS 180 nm; 1.8 V	29.5–47.6 dB	114 dB	8.538 µVrms, 0.5–150 Hz	84.4 µW, inferred	Sampling channel; output and RLD excluded; area not reported

III. ARCHITECTURE AND DEVELOPMENT FLOW

A. Requirements and acquisition chain

BioDiff-CMOS is analyzed within an ECG acquisition chain in which the performance of the first analog stage must be related to the signal characteristics and critical front-end metrics [1], [2]. Fig. 1 summarizes this relationship among requirements, architectural alternatives, performance metrics, and implementation trends. This view positions the differential block as part of a broader biomedical chain without assigning to the standalone circuit functions that were not implemented or validated.

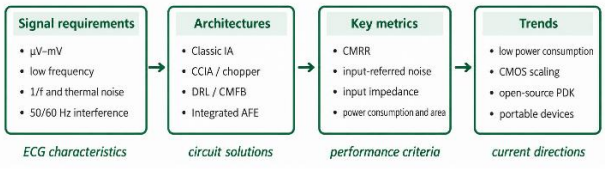


Fig. 1. Summary of requirements, architectural alternatives, critical metrics, and trends associated with analog front-ends for ECG.

Fig. 2 presents the functional chain considered as a reference: electrodes, input protection, the BioDiff-CMOS differential block, low-pass filtering, and the output interface. The first-order filter used in the testbench has a nominal cutoff of approximately 120.9 Hz in the revised version. Electrodes, input protection, and complete analog-to-digital conversion were not implemented in the core layout and therefore remain outside the validation scope of this work.



Fig. 2. Functional reference chain for simulated ECG acquisition with BioDiff-CMOS.

B. Open-source design pipeline

The development of BioDiff-CMOS follows a reproducible pre-silicon pipeline organized into seven stages, as shown in Fig. 3. In Stage 1, the electrical and functional specifications that guide the circuit design are defined. In Stage 2, these specifications are translated into the electrical domain through schematic capture in Xschem, establishing the topology, devices, and bias conditions. Stage 3 corresponds to electrical simulation in Ngspice, where circuit behavior is evaluated before physical implementation, including the analyses required to characterize its main metrics. Once schematic-level behavior has been verified, physical implementation begins. In Stage 4, the layout is developed in Magic according to the IHP SG13G2 process rules, considering geometry, placement, routing, and device symmetry. In Stage 5, parasitic extraction (PEX) is performed, generating an electrical representation derived from the layout and closer to the physical implementation conditions. Next, in Stage 6, DRC (Design Rule Check) and LVS (Layout Versus Schematic) are performed with Magic and Netgen to verify, respectively, geometric compliance with the process rules and connectivity equivalence between the implemented circuit and the reference schematic. In Stage 7, the extracted layout netlist replaces the schematic core in the same testbench used for pre-layout characterization. This approach enables direct comparison of results before and after physical implementation while keeping stimulus and analysis conditions unchanged. The pipeline in Fig. 3 establishes a traceable sequence across specification, electrical design, simulation, physical implementation, extraction, verification, and post-layout analysis, facilitating both experiment reproduction and identification of layout-induced effects. This organization is consistent with open analog-design flows that separate specification, schematic design, simulation, layout, extraction, physical verification, and characterization [12].

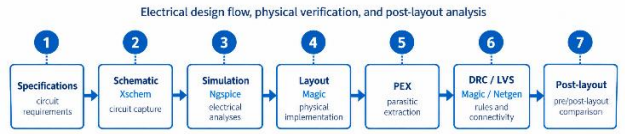


Fig. 3. Reproducible pre-silicon pipeline used in the development of BioDiff-CMOS.

C. BioDiff-CMOS architecture and testbench

Fig. 4 shows the configuration actually used in the revised simulations. The differential core implemented in the layout consists of the NMOS pair and the high-resistance loads; the ideal tail-current source, the RC branch, and the simplified CMOS interface belong to the test environment. This separation is essential for correctly interpreting power, CMRR, frequency response, and the limits of post-layout validation.

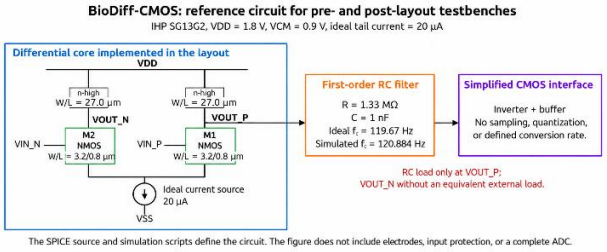


Fig. 4. Circuit and testbench used in the pre- and post-layout analyses; only the indicated core was implemented in the layout.

IV. MATERIALS AND METHODS

A. Circuit and test conditions

The core has two input NMOS devices with width-to-length ratio W/L = 3.2/0.8 µm and two rhigh loads of 2/10 µm. An ideal source provides 20 µA to the tail node. The positive output drives a 1.33 MΩ/1 nF branch and the gates of a CMOS inverter and buffer. The negative output has no equivalent external load. Fig. 4 distinguishes the core implemented in the layout from the testbench load and interface. The interface does not implement sampling, quantization, or conversion rate and is therefore not treated as an analog-to-digital converter.

The pre- and post-layout analyses used the same testbench; only the core netlist was replaced. The nominal condition used the mos_tt, res_typ, and cap_typ models, 27 °C, 1.8 V, and an input common-mode voltage of 0.9 V. The AC analysis applied a 1 V differential signal split into +0.5 and −0.5 V. The transient test applied a 25 mV peak signal per input in opposite phase, equivalent to 100 mVpp differential at 1 kHz. This test is electrical and does not represent an ECG waveform. Input-referred noise was integrated from 0.5 to 150 Hz and, as an additional check, from 0.05 to 150 Hz. CMRR and positive power-supply rejection ratio (PSRR+) were calculated at the differential output.

The nominal offset was obtained from a DC sweep without statistical models. The ideal tail-current source allows nonphysical voltages at low common-mode levels; therefore, the input common-mode range (ICMR) was not reported.

B. Layout, extraction, and reproducibility

The layout in Fig. 5 was created in Magic with symmetry between the input devices. The bounding box of the painted geometry, excluding markers, pads, and die area, measures 26.40 by 35.40 µm, or 934.56 µm². DRC was executed with the IHP SG13G2 deck. LVS compared seven pins and four instances; the sg13_lv_nmos and rhigh classes were treated as equivalent black boxes.

Extraction in Magic used extract do local, extract all, ext2spice lvs, zero capacitive threshold, and resistive extraction enabled. The resulting netlist for this version contains 18 capacitances, no interconnect parasitic resistors, and two rhigh loads reconstructed for testbench

compatibility. Thus, the analysis is capacitive/hybrid post-layout rather than a complete RC PEX.

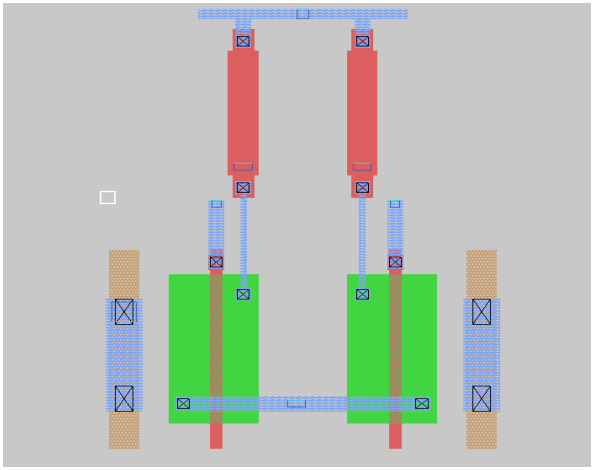


Fig. 5. Layout of the differential core in the IHP SG13G2 process.

The reproducible environment uses the isaiassh/unic-cass-tools:1.1.0 image, IHP Open PDK v0.3.0 at commit 5cccb161f749, Ngspice 44.2, Magic 8.3.613, Netgen 1.5.293, and Xschem at commit 313acc8e2974. The minimum files required for reproduction are the netlists in entrega/sim_pre and entrega/pex_post, the entrega/layout/biodiff_top.mag cell, and the scripts in scripts/entrega. The commands and generated artifacts are described in the project repository [11].

V. RESULTS AND DISCUSSION

A. Nominal pre- and post-layout characterization

Table II consolidates the requested conditions and metrics. The pre- and post-layout results coincide to the reported precision for gain, noise, cutoff frequency, and power. This agreement is due to the small extracted capacitive load and does not demonstrate fabrication robustness.

The frequency response is shown in Fig. 6. The value of 120.884 Hz was measured relative to the low-frequency gain. It is close to the 119.67 Hz calculated from $f_c = 1/(2\pi RC)$ for R = 1.33 MΩ and C = 1 nF. Using 100 Hz as the −3 dB reference would incorrectly shift the crossing because this point is already attenuated.

Wen et al. integrate the input-referred noise over 0.5–150 Hz [2], which is also used here to facilitate comparison. Because the 120.884 Hz cutoff lies below the upper limit of this range, the RC branch progressively attenuates components approaching 150 Hz, and its suitability depends on the intended application. In addition, a first-order filter provides limited rejection, and the design does not define the sampling rate, resolution, or response of a complete converter; therefore, the RC branch alone does not constitute a validated anti-aliasing chain.

TABLE II — BIODIFF-CMOS SPECIFICATIONS AND RESULTS.

Parameter	Condition or definition	Pre-layout	Post-layout	Unit or note
PDK and process	IHP SG13G2 BiCMOS	130	130	nm
Supply	Main supply	1.8	1.8	V
Core devices	Input NMOS; rhigh loads	3.2/0.8; 2/10	3.2/0.8; 2/10	W/L in µm
CMOS interface	PMOS/NMOS inverter; PMOS/NMOS buffer	4/0.35; 2/0.35; 8/0.35; 4/0.35	same	W/L in µm; outside the core layout
Bias current	Ideal tail-current source	20	20	µA; physical bias circuit not implemented
Nominal common mode	Inputs	0.9	0.9	V
Differential gain	10; 100; 1000 Hz	0.909260; 0.908323; 0.906963	0.909260; 0.908323; 0.906963	V/V
Filter cutoff	−3 dB relative to 1 Hz	120.884	120.884	Hz
Input impedance	Not characterized in the testbench	not characterized	not characterized	Biomedical requirement; requires input/electrode model
ICMR	Ideal tail-current source invalidates the lower limit	not characterized	not characterized	Requires physical biasing
Output swing	Extremes of differential DC sweep	0.286523	0.286523	Vpp differential; linear range not qualified
CMRR	Differential output, 100 Hz	128.292	128.291	dB; optimistic, no mismatch
PSRR+	Differential output, 100 Hz	48.683	48.683	dB; ideal biasing
Input-referred noise	Integrated, 0.5–150 Hz	29.9317	29.9317	µVrms
Input-referred noise	Integrated, 0.05–150 Hz	35.4578	35.4578	µVrms
Nominal offset	DC sweep, no mismatch	−25.32	−25.32	µV; not representative of fabrication offset
Power	Average from the 1.8 V supply	36.0116	36.0116	µW; physical biasing excluded
Core area	Geometry bounding box	934.56	934.56	µm²; no pads
Load	Only on VOUT_P	1.33 MΩ/1 nF + CMOS gates	same	VOUT_N has no equivalent external load
Corner and temperature	MOS/R/C models	TT/typical/typical; 27	TT/typical/typical; 27	°C
Extraction	Elements inside the core	none	18 C; 0 parasitic R; 2 reconstructed rhigh	Hybrid post-layout model
DRC/LVS	Executed deck and connectivity	not applicable	0 violations; unique match	Does not constitute electrical validation

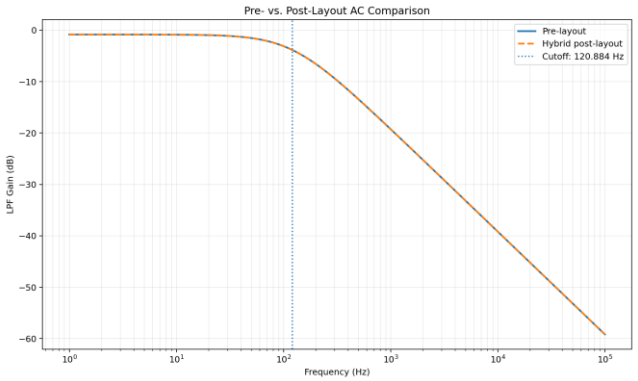


Fig. 6. Pre- and post-layout AC response. The curves overlap at the scale shown.

B. PVT screening and physical verification

The five-point screening combined nominal TT, SS at low voltage and high temperature, FF at high voltage and low temperature, and the crossed SF and FS corners. As shown in Fig. 7, the gain at 1 kHz ranged from 0.823 to 1.005 V/V, the cutoff frequency from 119.422 to 124.297 Hz, and power from 32.409 to 39.631 µW. The pre- and post-layout points remained overlapped. This screening does not replace all PDK-qualified corners or Monte Carlo analysis.

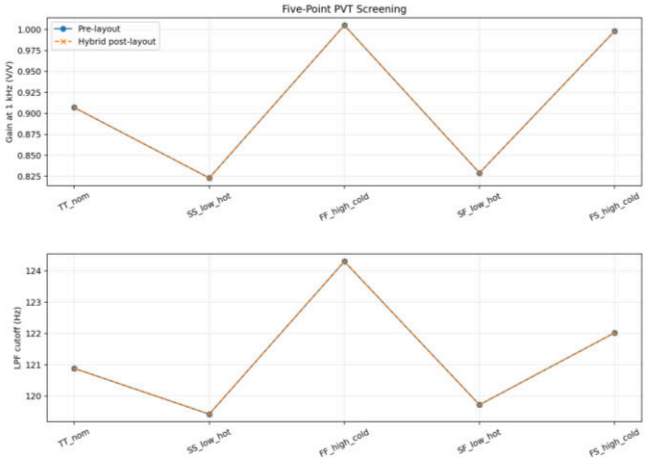


Fig. 7. Five-point PVT screening for gain and filter cutoff frequency.

DRC reported zero violations for the executed rule set. LVS found a unique match between the seven pins and four core instances, confirming connectivity consistency between the extracted layout and the schematic. These results demonstrate geometric compliance with the employed deck and topological equivalence, but they do not verify electrical performance, reliability, packaging, manufacturability beyond the evaluated rules, or behavior in silicon.

Fig. 8 presents enlarged excerpts of the original verification evidence: the Magic report records zero DRC errors, while the Netgen report shows pin and device equivalence and a unique match between the extracted layout and schematic. The detailed presentation preserves flow traceability and improves the readability of the physical-verification results.

(a) DRC

```
====================== Magic DRC Report (H65d4f4-CM5) ==========================
Magic 8.3 revision 603 - Compiled on Tue Mar 27 07:14:20 PM CST 2025.
Starting magic under Tcl interpreter
Using the terminal as the console.
Using the new style DRC interface.
Processing system .magicrc file.
Switching to xysystem mode.
Switching to RETRIEVE tool.
Switching to DRC tool.
Switching to DRC test.
Sourcing design .magicrc for technology sky130A.
2 Magic internal units = 1 Lambda.
Input style: sky130A (scaled-from-2, multiplicative)
  The following types of metal handled by extractor and will be treated as
  non-electrical types:
     nwell# metals: nsubstr nwelltie oxide2 mimicap metal3r metal4r metal5r
     metal6r metal7r metal8r
     stacker vias: via1r via2r via3
Loading sky130A design rules from /home/user/magic/sky130A/drc/sky130A.drc
Loading layout: /home/user/design/project/inverter/layout/inverter.gds
Call: drc check
Using 8 threads for DRC.
Checking design rules...
No violations detected.
 0 total DRC errors found.
DRC detailed report:  /home/user/design/project/inverter/reports/drc_report.txt
DRC summary file:     /home/user/design/project/inverter/reports/drc_summary.txt
DRC rule listing:     /home/user/magic/sky130A/drc/sky130A.rul
DRC completed successfully.
================================================================================
```

(b) LVS

```
Circuit 1 cell $g13_inv_sch and Circuit 2 cell $g13_inv_lay are black boxes.
Equate terminals: no equivalent cells.
Device classes $g13_inv_sch and $g13_inv_lay are equivalent.

Circuit 1 cell nwell and Circuit 2 cell nwell are black boxes.
Equate terminals: no equivalent cells.
Device classes nwell and nwell are equivalent.

Subcircuit summary:
Circuit 1: $g13_inv_sch                |  Circuit 2: $g13_inv_lay
---------------------------------------+-----------------------------------------
$g13_inv_sch (top)                     |  $g13_inv_lay (top)
nwell (1)                              |  nwell (1)
Number of devices: 4                   |  Number of devices: 4
Number of nets: 7                      |  Number of nets: 7

Bus/device primitives by property value.
Matching numbers by pin name.
Nets are matched uniquely.
Subcircuit pins:
Circuit 1: $g13_inv_sch                |  Circuit 2: $g13_inv_lay
---------------------------------------+-----------------------------------------
VDD                                    |  VDD
VOT                                    |  VOT
GND                                    |  GND
Cells $g13_inv_sch and $g13_inv_lay are equivalent.
Device classes $g13_inv_sch and $g13_inv_lay are equivalent.
Final result: Circuits match uniquely.
```

Fig. 8. Physical verification evidence: (a) DRC with no violations in the executed deck; (b) LVS with a unique match.

VI. LIMITATIONS AND FUTURE VALIDATION

The dominant limitations are the sub-unity gain, the high noise relative to the compared works, the ideal tail-current source, asymmetric loading, post-layout extraction without interconnect parasitic resistances, and the absence of statistical mismatch analysis. Therefore, the nominal CMRR should not be interpreted as a direct prediction of silicon measurements. Electrodes, electrode–skin impedance, electrode offset, 50/60 Hz interference, and input protection were also not modeled.

Future validation should advance along three fronts. First, pre-silicon characterization should be expanded to all PDK-defined corners, Monte Carlo analysis, electrode models, offset, power-line interference, and public ECG signals. Next, physical biasing, input protection, additional gain, and the output interface should be incorporated, followed by renewed RC extraction, DRC, and LVS. Finally, potential fabrication through a multi-project wafer (MPW) run will depend on meeting the specifications and on the availability of a compatible run; in that case, bench characterization should include gain, bandwidth, CMRR, PSRR, noise, offset, ICMR, output swing, and power under a known load, initially using an ECG simulator. Human-subject testing will require an appropriate ethics protocol and safety instrumentation.

VII. CONCLUSION

The development of integrated circuits for biomedical signal acquisition requires the integration of application requirements, analog-circuit fundamentals, and physical implementation. For ECG, key challenges include low signal amplitude, common-mode rejection, noise control, high input

impedance, low power consumption, and robustness to process, voltage, temperature, and device mismatch variations.

As a case study, BioDiff-CMOS demonstrated the feasibility of a reproducible pre-silicon pipeline based on an open PDK and open-source EDA tools, covering schematic capture, simulation, layout, extraction, post-layout analysis, and DRC/LVS verification. Under nominal conditions, the core achieved a gain of 0.908 V/V at 100 Hz, a cutoff frequency of 120.884 Hz, input-referred noise of 29.93 µVrms over 0.5–150 Hz, and power consumption of 36.01 µW. DRC reported no violations, and LVS confirmed a unique match between layout and schematic.

The results must, however, be interpreted within the context of a pre-silicon study. Important limitations remain, including sub-unity gain, the noise level, the ideal tail-current source, asymmetric loading, extraction without interconnect parasitic resistances, and the absence of statistical mismatch analysis. No fabrication, experimental characterization, or real ECG acquisition was performed.

Despite these limitations, the work establishes a solid and reproducible technical basis for further circuit evolution. The next steps include increasing gain, reducing noise, implementing physical biasing, balancing the loads, performing Monte Carlo and mismatch analyses, carrying out more complete RC extraction, and incorporating more representative electrode–skin interface models. Maturing these features may support future fabrication and experimental bench characterization.

More broadly, the use of open PDKs and tools such as Xschem, Ngspice, Magic, and Netgen reinforces the potential of this ecosystem for education, research, and training in biomedical microelectronics, bringing academic training closer to integrated-circuit implementation practices and contributing to the future development of solutions for digital health, wearable devices, and low-power biomedical systems.

ACKNOWLEDGMENT

The authors thank CI Amazônia, the State University of Amazonas, and the microelectronics training program for their academic and technological support.

REFERENCES

[1] S. Wang, H. Yang, G. Jiao, M. Ren, and X. Li, “ECG analog front end circuit—A review,” Measurement, vol. 267, Art. 120566, 2026, doi: 10.1016/j.measurement.2026.120566.

[2] Y. Wen, W. Liu, W. Li, B. Zheng, L. Dong, and Q. Feng, “An analog front-end with high common-mode rejection ratio and high input impedance for single-lead ECG signal acquisition,” IEICE Electronics Express, vol. 22, no. 10, Art. 20250097, 2025, doi: 10.1587/elex.22.20250097.

[3] J. Kim and H. Ko, “A dynamic instrumentation amplifier for low-power and low-noise biopotential acquisition,” Sensors, vol. 16, no. 3, Art. 354, 2016, doi: 10.3390/s16030354.

[4] Y. Zhao, Z. Shang, and Y. Lian, “A 2.55 NEF 76 dB CMRR DC-coupled fully differential difference amplifier based analog front end for wearable biomedical sensors,” IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 5, pp. 918–926, 2019, doi: 10.1109/TBCAS.2019.2924416.

[5] Y. Laababid, K. El Khadiri, and A. Tahiri, “Design of a low-power low-noise ECG amplifier for smart wearable devices using 180nm CMOS technology,” WSEAS Trans. Power Syst., vol. 17, pp. 177–186, 2022, doi: 10.37394/232016.2022.17.18.

[6] IHP Microelectronics, “IHP-Open-PDK: 130 nm BiCMOS open source PDK,” v0.3.0, commit 5cccb16, 2026. [Online]. Available: https://github.com/IHP-GmbH/IHP-Open-PDK.

[7] S. Schippers, “Xschem schematic editor,” 2026. [Online]. Available: https://xschem.sourceforge.io/stefan/index.html.

[8] Ngspice Project, “Ngspice open source circuit simulator,” ver. 44.2, 2025. [Online]. Available: https://ngspice.sourceforge.io/.

[9] Open Circuit Design, “Magic VLSI layout tool,” ver. 8.3.613, 2026. [Online]. Available: https://opencircuitdesign.com/magic/.

[10] Open Circuit Design, “Netgen circuit netlist comparison,” ver. 1.5.293, 2026. [Online]. Available: https://opencircuitdesign.com/netgen/.

[11] D. S. Cunha, “BioDiff-CMOS: project files and reproduction,” GitHub repository, 2026. [Online]. Available: https://github.com/DaniloCunha/ueletronica_projeto_final/tree/biodiff.

[12] T. Edwards, “Recommended open source analog design flow,” Open Circuit Design, 2023. [Online]. Available: https://opencircuitdesign.com/analog_flow/.

NON-INVASIVE BIOMAGNETIC ASSESSMENT OF GASTROINTESTINAL MOTILITY IN A LACTOSE-INDUCED MODEL OF CHRONIC DIARRHEA

1st Gustavo Locatelli Helena
Dept. of Biophysics and Pharmacology
Institute of Biosciences, São Paulo State University—UNESP
Botucatu, Brazil
https://orcid.org/0009-0008-6323-8434

2nd Gustavo Serafim Rodrigues
Dept. of Biophysics and Pharmacology
Institute of Biosciences, São Paulo State University—UNESP
Botucatu, Brazil
Gustavo.serafim@unesp.br

3rd Michael Almeida
Dept. of Biophysics and Pharmacology
Institute of Biosciences, São Paulo State University—UNESP
Botucatu, Brazil
https://orcid.org/0000-0003-3701-8266

4th Maria Eduarda Oliveira Bitencourt
Dept. of Biophysics and Pharmacology
Institute of Biosciences, São Paulo State University—UNESP
Botucatu, Brazil
Maria-eduarda.bitencourt@unesp.br

5th Erick Guilherme Stoppa
Dept. of Biophysics and Pharmacology
Institute of Biosciences, São Paulo State University—UNESP
Botucatu, Brazil
https://orcid.org/0000-0002-0903-6021

6th Guilherme Augusto Soares
Dept. of Biophysics and Pharmacology
Institute of Biosciences, São Paulo State University—UNESP
Botucatu, Brazil
https://orcid.org/0000-0002-4723-466X

***Abstract*—This study noninvasively evaluated gastrointestinal (GI) transit and contractile mechanical activity in rats with diarrhea induced by a high-lactose diet (HLD) using Alternating Current Biosusceptometry (ACB). The same animals were assessed at two different time points: before (CONTROL group) and after the diet (HLD group).**

***Index Terms*—Alternating Current Biosusceptometry (ACB), Lactose, Diarrhea, Gastrointestinal Motility, Colon.**

I. Introduction

Diarrhea is a common gastrointestinal disorder characterized by frequent loose stools and altered intestinal motility, often leading to dehydration and other complications [1], [2]. Experimental models, particularly lactose-induced chronic diarrhea in rats, are widely used to investigate gastrointestinal dysfunction and evaluate new therapies [3]. Conventional techniques for assessing gastrointestinal transit and contractility require euthanasia, preventing repeated measurements in the same animal [4], [5]. Alternating Current Biosusceptometry (ACB) is a non-invasive biomagnetic technique that enables sequential evaluation of gastrointestinal transit and motor activity without ionizing radiation [6]. Owing to its low cost, versatility, and ability to simultaneously assess gastric emptying, orocecal transit, and gastric and colonic contractility, ACB represents a valuable alternative to conventional methods [7]–[9].

This study aimed to establish and validate an ACB-based methodology for the simultaneous assessment of gastrointestinal transit and segmental contractility in a rat model of lactose-induced chronic diarrhea, providing a reliable tool for future preclinical gastrointestinal research [10].

II. Materials and Methods

A. Animals and Experimental Protocol

Fifteen male Wistar rats (Rattus norvegicus albinus), weighing between 250 and 350 g, were used in this study. The animals were maintained under controlled environmental conditions (21 $\pm$ 1°C, 12 h light/dark cycle) with free access to food and water. All experimental procedures were approved by the Institutional Animal Ethics Committee (protocol CEUA 1668051023).

A paired experimental design was adopted, in which each animal was evaluated at two different time points. Initially, measurements were performed under physiological conditions (Control). Subsequently, the same animals underwent induction of chronic diarrhea and were reevaluated (HLD group), allowing direct paired comparisons.

B. Lactose-Induced Chronic Diarrhea Model

Chronic diarrhea was induced by feeding the animals a diet containing 35% lactose for seven consecutive days. Successful

induction was confirmed through daily fecal morphometry. Animals presenting soft, watery, yellowish feces after the dietary intervention were considered successfully induced.

C. Alternating Current Biosusceptometry (ACB)

Gastrointestinal motility was evaluated using a single-channel Alternating Current Biosusceptometry (ACB) system specifically developed for in vivo studies.

The sensor consists of two pairs of coils separated by 15 cm. Each pair contains an excitation coil and a detection coil connected in opposite directions to minimize environmental magnetic noise and maximize the signal-to-noise ratio. The detection principle is based on disturbances produced by magnetic material located near the sensor. The recorded signal is directly proportional to the amount of magnetic material present and inversely proportional to the distance between the sensor and the magnetic marker. The induced electrical signal is processed using a lock-in amplifier, converted into a DC signal, acquired through a BIOPAC MP100 acquisition system at a sampling frequency of 20 Hz, and stored for subsequent processing.

D. Gastrointestinal Transit Measurements

Following a 12-hour fasting period, each animal received a solid magnetic meal (2 g chow, 0.5 g manganese ferrite microparticles, MnFe2O4). Magnetic intensity was recorded every 30 minutes over a 6-hour monitoring period with the sensor positioned over the stomach and cecum. Transit parameters quantified were Mean Gastric Emptying Time (MGET) and Mean Cecal Arrival Time (MCAT).

E. Gastric and Colonic Contractility

Mechanical activity was evaluated by monitoring oscillations in the biomagnetic signal. Animals were anesthetized with isoflurane. Gastric contractions were recorded 15 minutes after meal ingestion, and colonic contractions after six hours. Signals were analyzed using Fast Fourier Transform (FFT). Dispersion parameters obtained via Gaussian fitting were Standard deviation (σ) and Full Width at Half Maximum (FWHM).

F. Statistical Analysis

Statistical analyses were performed using GraphPad Prism 8.0.1. Data are expressed as mean $\pm$ standard deviation. Since the same animals were evaluated before and after diarrhea induction, paired Student's t-tests were used for comparisons between control and HLD conditions. Statistical significance was considered when $p < 0.05$.

III. Results

A. Validation of the Lactose-Induced Diarrhea Model

All animals completed both experimental phases, allowing paired comparisons between physiological conditions (Control) and the high-lactose diet (HLD). Administration of the HLD for seven consecutive days produced the expected diarrheal phenotype. Animals exhibited soft, watery, yellowish feces together with increased stool frequency, confirming successful induction of chronic diarrhea. No significant differences in body weight were observed between control and HLD conditions during the experimental period.

B. Gastrointestinal Transit

The evaluation of temporal parameters demonstrated significant quantitative alterations following diarrhea induction.

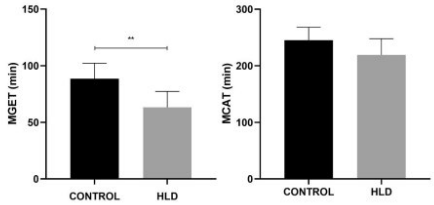


Fig. 1. Mean times of gastric emptying (MGET), cecum arrival (MCAT). ** $p<0.01$.

The Mean Gastric Emptying Time (MGET) was significantly reduced in the HLD group (63.38 ± 13.87 min) compared to the CONTROL group (88.51 ± 13.69 min, $p < 0.01$), indicating accelerated gastric emptying. Regarding the Mean Cecal Arrival Time (MCAT), there was a reduction in the HLD group (219.3 ± 28.56 min) compared to the CONTROL group (245.1 ± 23.17 min); however, this decrease was not statistically significant ($p>0.05$).

C. Gastric Contractility

Representative gastric biomagnetic recordings obtained during control and HLD conditions are shown in Fig. 2.

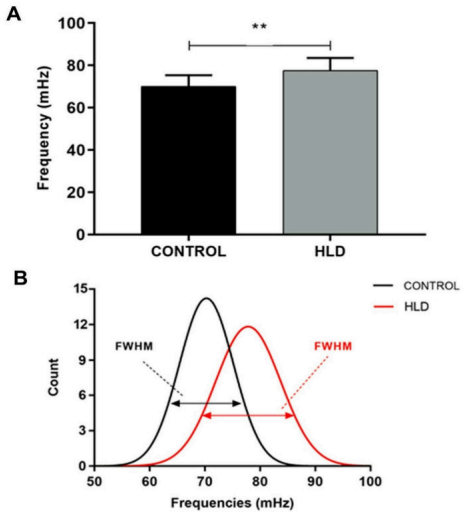


Fig. 2. A. Mean gastric contraction frequency of the CONTROL and HLD groups. ** $p<0.01$. B. Distribution of quantified values of frequency of gastric contraction and the Full Width at Half Maximum (FWHM).

Time-domain recordings revealed periodic oscillations corresponding to gastric contractions in both experimental conditions. However, spectral analysis demonstrated important

differences after diarrhea induction. Fast Fourier Transform (FFT) analysis showed a significant increase in dominant gastric contraction frequency in animals receiving the high-lactose diet. In addition to frequency changes, Gaussian fitting of the frequency spectra demonstrated broader peak distributions in the HLD group, characterized by increased standard deviation (σ) and Full Width at Half Maximum (FWHM). The larger spectral dispersion indicates reduced rhythmic stability and increased variability of gastric contractions, suggesting the presence of gastric dysrhythmia associated with chronic diarrhea.

D. Colonic Mechanical Activity

Colonic signals exhibited complex rhythmic oscillations. Animals subjected to lactose-induced diarrhea showed marked enhancement of colonic mechanical activity.

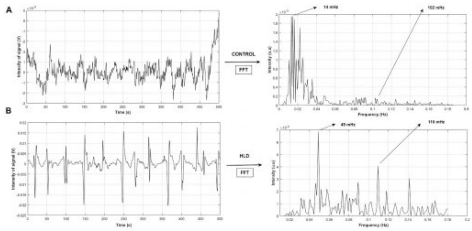


Fig. 3. A. Colonic contraction signal and the respective FFT of the CONTROL group. B. Colonic contraction signal and the respective FFT of the HLD group.

Spectral analysis demonstrated increased power within the frequency bands associated with rhythmic propulsive motor complexes (RPMCs, CONTROL: 38 mHz vs. HLD: 62 mHz, $p< 0.05$) and rhythmic propagation ripples (RPRs, CONTROL: 85 mHz vs. HLD: 95 mHz, $p< 0.05$), indicating increased propulsive activity.

E. Frequency Spectrum Analysis

Power spectral density analysis provided additional evidence of altered gastrointestinal motility following diarrhea induction. Representative spectra are shown in Fig. 4.

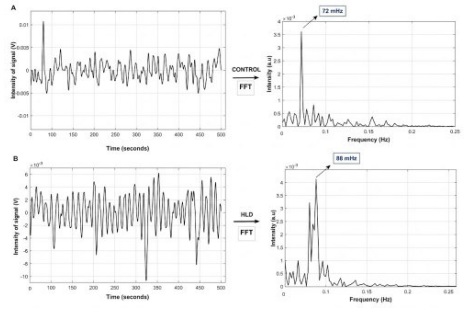


Fig. 4. A. Gastric contraction signal and the respective FFT of the CONTROL group. B. Gastric contraction signal and the respective FFT of the HLD group.

The control group exhibited narrow spectral peaks centered around the physiological gastric contraction frequency. Conversely, spectra obtained after HLD administration presented wider peaks with increased frequency dispersion, demonstrating greater variability in contraction rhythms. Gaussian fitting confirmed statistically significant increases in spectral width parameters (σ and FWHM), indicating disruption of normal gastric rhythmicity.

F. Correlation Between Transit and Contractility

Comparison of transit measurements and contractility parameters demonstrated a close relationship between gastrointestinal transit acceleration and altered motor activity. Animals presenting shorter gastric emptying times also exhibited higher gastric contraction frequencies and increased colonic propulsive activity. These findings suggest that accelerated gastrointestinal transit observed in lactose-induced diarrhea results from coordinated changes involving both gastric and colonic motor functions.

G. Overall Findings

Collectively, the results demonstrate that Alternating Current Biosusceptometry (ACB) enables the simultaneous in vivo evaluation of gastrointestinal transit and regional contractility. The technique successfully detected accelerated gastric emptying, reduced cecal arrival time, increased gastric contraction frequency, a broader distribution of gastric contraction frequencies, enhanced colonic propulsive motor activity, and increased rhythmic propagation ripples. These findings indicate that chronic diarrhea induces profound alterations in gastrointestinal motility involving multiple regions of the digestive tract. Furthermore, they highlight the potential of ACB as a reliable, non-invasive tool for continuously monitoring gastrointestinal motor function, providing valuable insights for preclinical studies of gastrointestinal disorders and the evaluation of novel therapeutic strategies.

IV. Discussion

The present study demonstrated that Alternating Current Biosusceptometry (ACB) provides a non-invasive approach for the simultaneous assessment of gastrointestinal transit and regional contractility. The high-lactose diet successfully induced chronic diarrhea, producing marked alterations in gastrointestinal motility. The quantitative reduction in both Mean Gastric Emptying Time (MGET) and Mean Cecal Arrival Time (MCAT) indicates that diarrhea accelerates gastrointestinal transit.

ACB also detected significant modifications in gastric motor activity. The statistical increase in dominant gastric contraction frequency, together with the broader frequency distribution (higher FWHM and σ) observed after diarrhea induction, suggests disruption of normal gastric rhythmicity. Colonic recordings further demonstrated enhanced mechanical activity, with significant increases in RPMCs and RPRs.

Some limitations of this study must be strongly emphasized. First, the experimental design utilized a relatively small sample size of only 15 rats, which limits the statistical power of broader extrapolations. Second, the evaluation was restricted to a single diarrhea induction protocol (high-lactose diet), which

may not represent the pathophysiological mechanisms of other types of diarrhea. Due to these factors, the impossibility of directly extrapolating these results to human clinical conditions must be acknowledged; any clinical translation should be approached with extreme caution. Future studies with larger cohorts and diverse pathological models are necessary.

V. Conclusion

Alternating Current Biosusceptometry enabled simultaneous, non-invasive evaluation of gastrointestinal transit and regional contractility in a rat model of lactose-induced chronic diarrhea. The biomagnetic methodology detected accelerated gastric emptying, increased gastric contraction frequency, broader frequency distributions indicative of gastric dysrhythmia, and enhanced colonic propulsive activity. These findings demonstrate that ACB provides an assessment of gastrointestinal motor function with the advantage of allowing repeated measurements in the same animals. The technique demonstrates significant potential as an experimental tool for investigating gastrointestinal disorders and evaluating therapeutic interventions specifically within the evaluated experimental model.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the Brazilian funding agencies and the technical assistance of the Laboratory of Biomagnetism, Department of Biophysics and Pharmacology, Institute of Biosciences, São Paulo State University (UNESP), Botucatu, Brazil.

References

[1] P. D. Headstrom and C. M. Surawicz, "Chronic diarrhea," *Clinical Gastroenterology and Hepatology*, vol. 3, no. 8, pp. 734-737, 2005.

[2] B. Müllhaupt, "Diarrhea," *Praxis*, vol. 91, no. 42, pp. 1749-1756, 2002.

[3] P. A. Boakye et al., "Garcinia buchananii bark extract is effective antidiarrheal remedy for lactose-induced diarrhea," *Journal of Ethnopharmacology*, vol. 142, no. 2, pp. 539-547, 2012.

[4] C. Scarpignato, G. Coruzzi, L. Zappia, and G. Bertaccini, "Effect of octylonium bromide on the gastrointestinal tract in vivo," *Il Farmaco; Edizione Pratica*, vol. 35, no. 5, pp. 249-257, 1980.

[5] P. Padmanabhan, J. Grosse, A. B. M. A. Asad, G. K. Radda, and X. Golay, "Gastrointestinal transit measurements in mice with 99mTc-DTPA-labeled activated charcoal using NanoSPECT-CT," *EJNMMI Research*, vol. 3, no. 1, p. 60, 2013.

[6] W. Weitschies, J. Wedemeyer, R. Stehr, and L. Trahms, "Magnetic markers as a noninvasive tool to monitor gastrointestinal transit," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 2, pp. 192-195, 1994.

[7] O. Baffa, R. B. Oliveira, J. R. Miranda, and L. E. Troncon, "Analysis and development of AC biosusceptometer for orocaecal transit time measurements," *Medical & Biological Engineering & Computing*, vol. 33, no. 3, pp. 353-357, 1995.

[8] M. F. F. Calabresi et al., "Alternate current biosusceptometry for the assessment of gastric motility after proximal gastrectomy in rats: A feasibility study," *Neurogastroenterology & Motility*, vol. 27, no. 11, pp. 1613-1620, 2015.

[9] M. F. F. Calabresi et al., "Changes in colonic contractility in response to inflammatory bowel disease: Long-term assessment in a model of TNBS-induced inflammation in rats," *Life Sciences*, vol. 236, p. 116833, 2019.

[10] C. C. Quini et al., "Employment of a noninvasive magnetic method for evaluation of gastrointestinal transit in rats," *Journal of Biological Engineering*, vol. 6, no. 1, p. 6, 2012.

Tracking Thermal-Magnetic Phase Transitions in Gallium-Doped Epsilon-Iron Oxide Nanoparticles Using AC Biosusceptometry

1st Michael Almeida
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University - UNESP
Botucatu - SP, Brazil
mj.almeida@unesp.br

2nd Tehseen Ali Anjum
Aveiro Inst. of Materials, CICECO
University of Aveiro
Aveiro, Portugal
alianjum18@ua.pt

3rd Gabriel Pinto
Aveiro Inst. of Materials, CICECO
University of Aveiro
Aveiro, Portugal
gabrielcardoso1@ua.pt

4th Guilherme Soares
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University - UNESP
Botucatu - SP, Brazil
guilherme.soares@unesp.br

5th José Ricardo A. Miranda
Dept. of Biophysics and Pharmacology, IBB
São Paulo State University - UNESP
Botucatu - SP, Brazil
jose.r.miranda@unesp.br

6th Nuno João Silva
Aveiro Inst. of Materials, CICECO
University of Aveiro
Aveiro, Portugal
nunojoao@ua.pt

***Abstract*—Epsilon iron oxide has emerged as a powerful alternative for expanding the capabilities of magnetic tracking toward multifunctional diagnostics and contactless nanothermometry due to its giant coercive field and strongly temperature-dependent magnetization. In this study, gallium-doped epsilon iron oxide nanoparticles were synthesized via the sol-gel method to tailor the magnetic phase transition window between 5 degrees Celsius and 25 degrees Celsius, aligning with critical intervals required for food preservation, vaccine transportation, and cold-chain logistics. To monitor these phase transitions in real time, Alternating Current Biosusceptometry (ACB)—a versatile inductive technique utilizing a first-order gradiometric coil configuration—was evaluated against a standard Vibrating Sample Magnetometer (VSM). The temperature-dependent magnetic responses of three distinct samples (M-04, M-02, and M-03R) were measured using both systems. The experimental results reveal that both techniques successfully resolve the thermal phase transitions, displaying a consistent trend in the Curie temperature shift as a function of gallium doping. Although the ACB system identified transition peaks at 7.5, 16.4, and 22.0 degrees Celsius, the VSM recorded them at 1.5, 12.0, and 19.0 degrees Celsius. This systematic temperature offset is discussed and attributed to distinct instrumentation profiles, thermal gradients, and signal integration times between the localized dynamic ACB probe and the quasi-static VSM chamber environment. Crucially, despite these instrumentation-specific offsets, ACB demonstrates a promising capability for qualitatively tracking thermal-magnetic variations and distinguishing Tc shifts across doped samples. These findings show that ACB serves as an accessible laboratory tool for preliminary thermal-magnetic phase screening, offering real-time trend monitoring.**

***Index Terms*—AC Biosusceptometry, Magnetic phase transition, Epsilon iron oxide, Nanothermometry, Curie temperature.**

I. Introduction

Superparamagnetic iron oxide nanoparticles (SPIONs) have been widely established as effective contrast agents for magnetic resonance imaging (MRI) [1] and magnetic particle imaging (MPI) [2]. Specifically, commercial solutions such as Sienna+®, which contains dextran-coated SPIONs with an average particle size of $59\,\mathrm{nm}$, have been successfully utilized for the identification of sentinel nodes (SN) in breast cancer surgery [3]. Once injected into the breast, the tracer drains through the lymphatic system and accumulates in the SN. A handheld magnetic probe, known as SentiMag®, is then employed by surgeons to locate these nodes in a manner analogous to conventional radioactive gamma probes. Although this combined Sienna+ / SentiMag system demonstrates non-inferiority compared to standard radiotracer techniques [4], [5] and offers visual identification due to dark brown staining of the nodes, it does present limitations, such as a localized tendency to discolor the skin at the injection site [6]–[10]. More importantly, its functionality is limited strictly to physical tracking and localization.

To expand the capabilities of magnetic tracking for multifunctional diagnostics, epsilon iron oxide (ϵ-Fe_2O_3) emerges as a powerful alternative. ϵ-Fe_2O_3 is a rare Fe_2O_3 polymorph with an orthorhombic crystal structure belonging to the $Pna2_1$ space group [11]. It exhibits a giant coercive field of approximately 2 T at room temperature and an exceptionally large magnetocrystalline anisotropy, making it remarkable among all iron oxide polymorphs. Crucially, ϵ-Fe_2O_3 exhibits a strong variation of magnetization with temperature, undergoing a sharp phase transition from a ferrimagnetic to a paramagnetic state at a specific Curie temperature (T_C). This high coercivity, prominent anisotropy, and distinct thermal transition make ϵ-Fe_2O_3 nanoparticles an ideal candidate for contactless nanothermometry.

The concept of a contactless nanothermometer relies on a system with a highly temperature-dependent, non-linear

magnetic response. Although pure ϵ-Fe_2O_3 has a T_C near $\approx 400\,^{\circ}\mathrm{C}$, this transition temperature can be precisely engineered through gallium doping, tailoring the sensing window to specific industrial or biomedical thermal ranges [12]. The samples used in this study consist of Ga-doped ϵ-Fe_2O_3 nanoparticles whose transition behavior spans between $5\,^{\circ}\mathrm{C}$ and $25\,^{\circ}\mathrm{C}$. This interval aligns well with the temperature windows required for food preservation, vaccine transportation, and perishable goods logistics, making them highly attractive for smart packaging and cold-storage technologies [13].

Integrating these nanoparticles into magnetic sensing systems enables real-time temperature monitoring without direct electrical contact. In this scenario, alternating current biosusceptometry (ACB) emerges as a highly promising, versatile, and low-operational-cost magnetic detection technique. The ACB system utilizes a first-order gradiometric coil configuration to detect the temporal derivative of magnetization under an alternating magnetic field [14], [15]. Furthermore, ACB is a exceptionally cost-effective technique compared to MPI or MRI. It features excellent temporal resolution and has been successfully validated in clinical and pre-clinical applications, including the study of the biodistribution of magnetic nanoparticles (MNPs) [16], gastrointestinal transit time [17], and hepatic uptake investigations [18].

Beyond its conventional uses, ACB serves as an excellent tracking tool for magnetic characterization due to its ability to operate at low primary field excitation frequencies, where thermal fluctuations do not directly compromise the measurements. Furthermore, tailored ACB sensors can detect these non-linear magnetic signals, opening new frontiers for simultaneous non-contact material tracking and precise thermal sensing based on the engineered T_C of ϵ-Fe_2O_3.

II. MATERIALS AND METHODS

A. Chemicals and synthesis

The Ga-doped ϵ-Fe_2O_3 ferrofluids were synthesized using the sol-gel method. Gallium was selected as the dopant to produce ϵ-$Ga_xFe_{2-x}O_3$ NPs, where $x \approx 0.97, 0.99$ and 1.0 for the samples M-03R, M-02 and M-04, respectively. In this procedure, a water-in-oil microemulsion system was employed, where the precursors were formed within the aqueous internal phase. To initiate the synthesis, two separate microemulsions were prepared as the oil phase; for each, 17.87 mL of n-octane, 3.57 mL of 1-butanol, and 3.535 mg of cetyltrimethylammonium bromide (CTAB) as a surfactant were mixed at $18\,^{\circ}\mathrm{C}$.

Subsequently, the aqueous phases were added to the microemulsions. The first aqueous phase was composed of 5.41 mL of water containing $Fe(NO_3)_3$ and $Ga(NO_3)_3$ salts added in stoichiometric ratios corresponding to the target x value, while the second aqueous phase consisted of 5.41 mL of water and 1.18 mL of an aqueous NH_4OH (ammonia) solution. Once both microemulsions reached stability, they were combined under rapid stirring. Following the fusion of the micelles, 0.91 mL of tetraethoxysilane (TEOS) was added as a silica precursor. After a 24-hour reaction period, the NPs were isolated by centrifugation, washed with chloroform and methanol, and dried.

The next step in the synthesis was to calcinate the NPs at $1050\,^{\circ}\mathrm{C}$ for 4 hours. Because this process results in an ϵ-phase embedded within a silica oxide matrix, a final chemical etching step using a 3 M NaOH solution was required to dissolve the matrix and isolate the nanoparticles [12].

B. Alternating current biosusceptometry

Alternating Current Biosusceptometry (ACB) is a cost-effective inductive technique for measuring magnetic materials. The system uses a first-order gradiometric configuration consisting of two pairs of coaxial excitation with a diameter of 2.8 cm (200 wire turns of AWG 24) and pick-up coils with a diameter of 2.1 cm (500 wire turns of AWG 32), as shown in Figure 1. The excitation coils generate an AC field (6.5 mT, 10 kHz), inducing a constant-rate current in the pickup coils. The setup comprises two units: a measurement unit, placed near the sample to detect NPs, and a reference unit, which cancels background noise and residual excitation fields to enhance the signal-to-noise ratio.

In the absence of a sample, the net magnetic flux is balanced, resulting in a null signal. The proximity of NPs alters this balance, inducing a voltage proportional to the parameters of the system (coil geometry, applied current) and the properties of the sample (volume, mass, susceptibility, and distance). The signal is generated and recorded using a low-noise Zurich Instruments MFLI lock-in amplifier (Zurich, Switzerland).

In the linear regime, magnetization is given by $M = \chi H$, where the magnetic susceptibility is $\chi = \chi' - i\chi''$, with χ' and χ'' representing the real and imaginary components, respectively. The ACB signal is proportional to the time derivative of magnetization, $\Delta V \propto \frac{dM}{dt}$. The voltage signal detected by the ACB system can be expressed as:

$$\Delta V = \mu_0 N A \alpha \omega H_0 (\chi'' + i\chi') \tag{1}$$

where ΔV is the induced voltage variation in the detection coil; μ_0 is the vacuum permeability ($4\pi \times 10^{-7}$ H/m); N is the number of turns in the detection coil; α is the geometric coupling factor, which depends on the experimental configuration and the position of the sample relative to the coil; ω is the angular frequency of the applied magnetic field ($\omega = 2\pi f$); H_0 is the amplitude of the applied alternating magnetic field.

The sensor allows the detection of different signal components. The real component is given by:

$$\Delta V_x = \mu_0 N A \alpha \omega H_0 \chi'' \tag{2}$$

and the imaginary component by:

$$\Delta V_y = \mu_0 N A \alpha \omega H_0 \chi' \tag{3}$$

The amplitude in polar coordinates is:

$$\Delta V_R = \sqrt{\Delta V_x^2 + \Delta V_y^2} \tag{4}$$

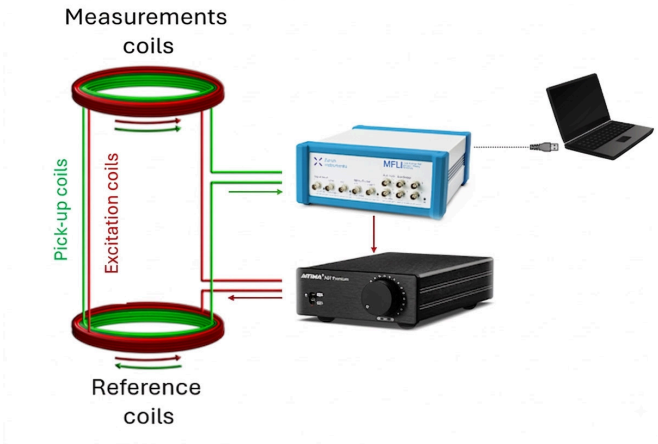


Fig. 1. Schematic diagram of the first-order gradiometric coil system used in the ACB setup, illustrating the arrangement of measurement, reference, excitation, and pick-up coils.

and the phase is given by:

$$\theta = \arctan(\Delta V_y / \Delta V_x) \tag{5}$$

C. Vibrating sample magnetometer

The Vibrating Sample Magnetometer (VSM) quantifies magnetic moments by vibrating a sample at a constant frequency within a uniform magnetic field (H), as illustrated in Fig. 2b. According to Faraday's Law of Induction, this periodic motion induces a voltage in the nearby pickup coils that is directly proportional to the magnetization of the sample. By integrating a temperature controller, the VSM enables precise monitoring of the magnetic moment as a function of temperature (M vs. T). Due to its high sensitivity and accuracy in temperature-dependent measurements, the VSM is widely utilized for studying magnetic nanoparticles, thin films, ferrites, and magnetic phase transitions.

D. Magnetic phase transition measurements

The magnetic phase transition study was conducted by cooling the nanoparticles to temperatures below their Curie temperature (T_C) in the absence of an external magnetic field, ensuring that the magnetic domains remained randomly oriented. To achieve these sub-ambient temperatures, the samples were placed in a temperature-controlled cooling system.

Upon reaching the base temperature, the Zero-Field-Cooled (ZFC) magnetization was measured using the VSM. During this procedure, a small external magnetic field ($B = 5$ mT, where $B = \mu_0 H$) was applied, and magnetization $M(T)$ was recorded as the sample was gradually heated. The resulting ZFC curve exhibits a distinct peak corresponding to the magnetic phase transition temperature (T_C), where the thermal energy becomes sufficient to overcome the magnetic ordering of the material, driving the transition from a ferrimagnetic state to a paramagnetic state.

For the ACB measurements, the samples were frozen similarly using liquid nitrogen and then placed inside the ACB

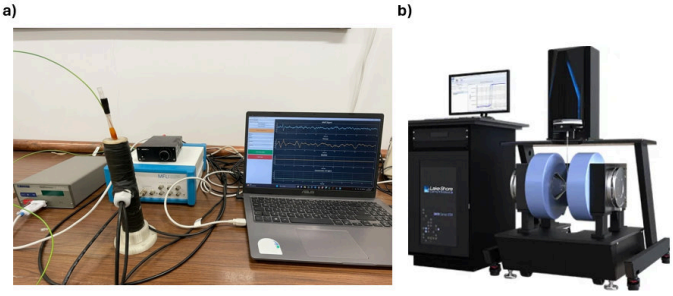


Fig. 2. Experimental setups used for magnetic characterization: a) ACB system integrated with a Neoptix temperature sensor, and a dedicated Python-based interface for real-time data acquisition; b) Commercial Lake Shore Cryotronics VSM system used for M vs. T measurements

system alongside a Neoptix Reflex fiber optic temperature sensor, as shown in Fig. 2a. The ACB signal was digitized using a lock-in amplifier, while the temperature data was captured via a Neoptix Reflex signal conditioner. Both the R component (signal amplitude) and the θ component (phase), along with the temperature, were recorded simultaneously using a custom Python script.

III. RESULTS AND DISCUSSION

Figure 3 presents ACB signals and ZFC curves for Gallium-doped epsilon iron oxide nanoparticles. The VSM measurements (Figs. 3D, 3E, and 3F) provide information on the magnetic ordering and phase transition behavior under a small applied field of $5\,\mathrm{mT}$. All three samples exhibit distinct magnetization maxima corresponding to their Curie temperatures (T_C), where the materials undergo a transition from a ferrimagnetic to a paramagnetic state. The values of T_C were determined to be approximately $1.5\,^\circ\mathrm{C}$ for M-04 (Fig. 3D), $12.0\,^\circ\mathrm{C}$ for M-02 (Fig. 3E), and $19.0\,^\circ\mathrm{C}$ for M-03R (Fig. 3F).

The experimental results reveal that both techniques clearly resolve the ferrimagnetic-to-paramagnetic phase transition. For the ACB measurements, the transition temperatures (T_C) were approximately $7.5\,^\circ\mathrm{C}$ for M-04 (Fig. 3A), $16.4\,^\circ\mathrm{C}$ for M-02 (Fig. 3B) and $22.0\,^\circ\mathrm{C}$ for M-03R (Fig. 3C). In comparison, the VSM profiles yielded T_C values of $1.5\,^\circ\mathrm{C}$, $12.0\,^\circ\mathrm{C}$, and $19.0\,^\circ\mathrm{C}$, respectively. Although a systematic temperature offset of approximately $3.0\,^\circ\mathrm{C}$ to $6.0\,^\circ\mathrm{C}$ is observed between the two methods, the fundamental trend of the T_C shift as a function of gallium doping remains remarkably consistent between both techniques.

To objectively evaluate the degree of agreement between the two methodologies, quantitative comparative metrics were calculated, as summarized in Table I. Across the three distinct nanoparticle formulations, the mean absolute error (MAE) in transition temperature determination between ACB and VSM was $4.47\,^\circ\mathrm{C}$. Despite this systematic absolute offset, the relative sensitivity of T_C shift to gallium content ($\Delta T_C / \Delta x$) shows strong consistency, yielding a Pearson correlation coefficient of $r = 0.998$ ($p < 0.05$) between the transition temperatures estimated by both techniques. This confirms that, while an absolute calibration offset exists, ACB exhibits

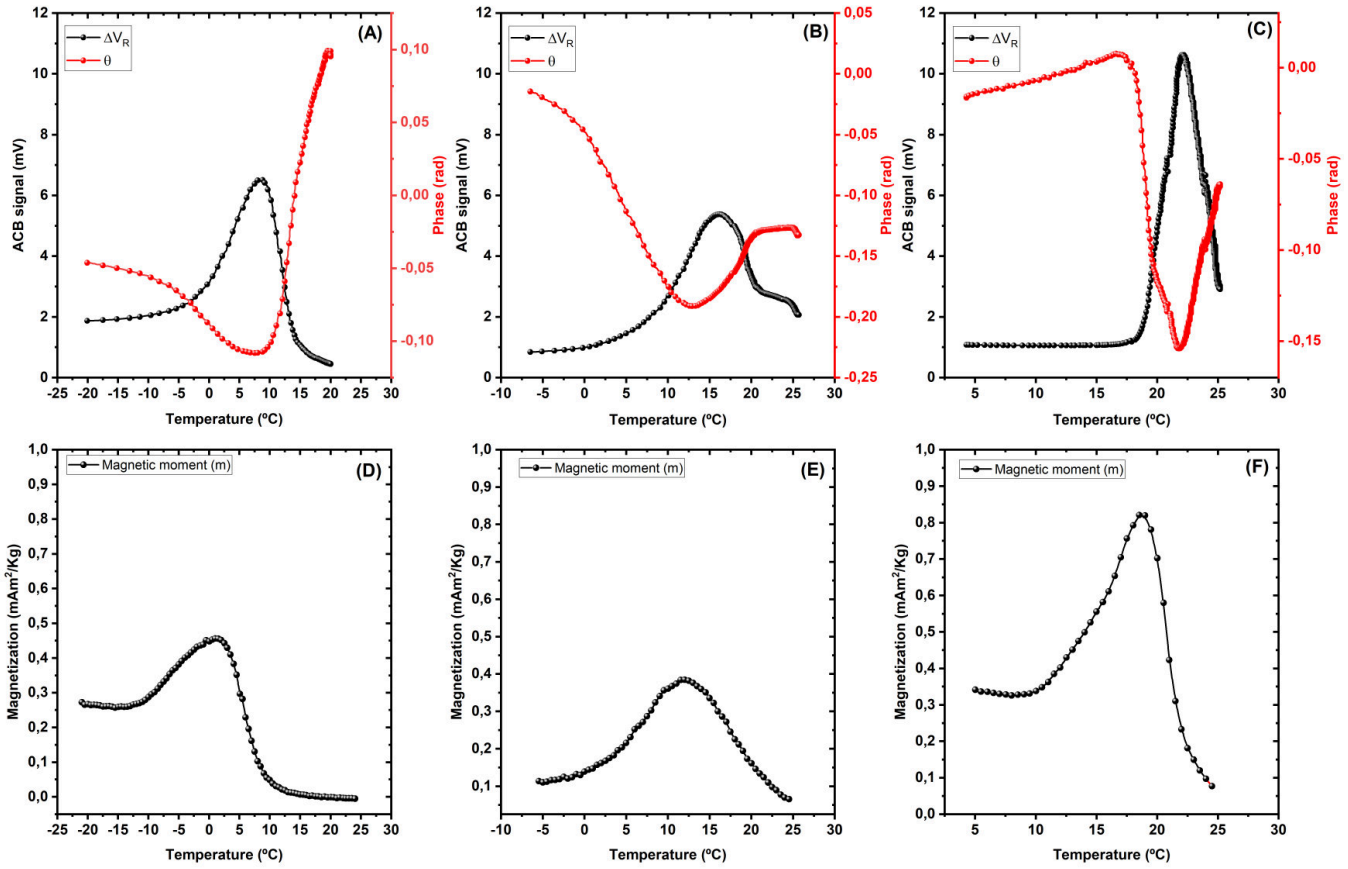


Fig. 3. Temperature dependence of the magnetic properties for the ϵ-$Ga_xFe_{2-x}O_3$ nanoparticle samples M-04, M-02, and M-03R. **(A–C)** AC Biosusceptometry (ACB) signal (ΔV_R) and phase angle (θ) as a function of temperature for samples (A) M-04, (B) M-02, and (C) M-03R. **(D–F)** Zero-Field-Cooled (ZFC) magnetization curves as a function of temperature measured via Vibrating Sample Magnetometry (VSM) under an applied magnetic field of 5 mT for samples (D) M-04, (E) M-02, and (F) M-03R. Both techniques clearly resolve the identical ferrimagnetic-to-paramagnetic phase transition peaks, demonstrating the utility of the ACB method for qualitative thermal-magnetic phase transition tracking compared to static VSM measurements.

exceptional fidelity in tracking the relative thermal-magnetic phase behavior as a function of chemical composition.

TABLE I
QUANTITATIVE COMPARISON AND AGREEMENT METRICS FOR CURIE TRANSITION TEMPERATURES (T_C) DETERMINED VIA VSM AND ACB

Sample	Ga content (x)	VSM T_C (°C)	ACB T_C (°C)	Absolute Offset (°C)	Relative Offset (%)
M-04	1.00	1.5	7.5	6.0	400.0[a]
M-02	0.99	12.0	16.4	4.4	36.7
M-03R	0.97	19.0	22.0	3.0	15.8

[a]High percentage offset for M-04 is due to the small absolute reference denominator near 0 °C.
Summary Metrics: Mean Absolute Error (MAE) = 4.47 °C; Pearson Correlation (r) = 0.998.

This discrepancy in absolute transition temperatures can be attributed to distinct instrumentation profiles, thermal integration times, and physical measurement principles. In the ACB setup, the Neoptix temperature probe is positioned in direct contact with the sample, providing localized real-time thermal readings. In contrast, in the commercial VSM system, the internal temperature sensor is typically located within the sample chamber environment rather than inside the specimen itself, introducing a well-known thermal lag during dynamic temperature sweeps. Furthermore, unlike VSM, which captures quasi-static (DC) equilibrium magnetization (M), ACB measures the dynamic differential susceptibility ($\chi' + i\chi''$) under an alternating excitation field ($f = 10\,\mathrm{kHz}$, $H_0 = 6.5\,\mathrm{mT}$). Near Curie temperature (T_C), as the magnetocrystalline anisotropy barrier decreases, the finite relaxation times of the nanoparticles relative to the 10 kHz excitation period can induce phase shifts and broaden the observed voltage peak (ΔV_R), contributing to apparent shifts in the dynamic transition profile compared to static ZFC curves.

Although the observed temperature shift of 3.0 °C to 6.0 °C is plausibly attributed to thermal lag and dynamic relaxation, this thermal gradient hypothesis was not isolated experimentally in the current study. Future work incorporating systematically controlled heating rates, multi-point temperature sensors, and independent thermal calibration replicates will be required to quantitatively decouple instrumental thermal lag from intrinsic magnetic relaxation effects.

Crucially, despite these instrumentation-specific offsets, the ACB sensor successfully demonstrates its efficacy for monitoring the relative onset and shift of thermal-magnetic phase tran-

sitions. Regarding economic feasibility, although the present laboratory prototype utilizes high-grade instrumentation (such as a Zurich Instruments MFLI lock-in amplifier and a Neoptix optical temperature sensor), the underlying inductive detection technique remains inherently lower in operational complexity compared to commercial VSM systems. VSM units often involve substantial maintenance, specialized power supplies, and, in many configurations, cryogenic cooling infrastructure. In contrast, ACB operates with low-power AC excitation coils at room temperature. Furthermore, the electronic readout components can be streamlined in future iterations using dedicated microcontroller-based lock-in circuits, significantly reducing capital expenditure for point-of-care or field-deployable applications.

In conclusion, ACB demonstrated promising capability for qualitatively tracking thermal-magnetic transitions and resolving relative T_C shifts in ϵ-Fe_2O_3 nanoparticles as a function of gallium doping. While quantitative equivalence with VSM still requires further thermal calibration, systematic offset corrections, and uncertainty analysis, ACB provides an efficient, accessible, and scalable methodology for preliminary phase transition screening.

ACKNOWLEDGMENT

The author thanks CAPES for the Doctoral Sandwich Program Abroad (PDSE) fellowship (process number: 88881.219010/2025-01). This work was also supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant ERC-2019-CoG-865437), as well as by CICECO - Aveiro Institute of Materials and I3N - Institute of Nanostructures, Nanomodelling and Nanofabrication.

REFERENCES

[1] S. Laurent, C. Henoumont, D. Stanicki, S. Boutry, E. Lipani, S. Belaid, R. N. Muller, and L. Vander Elst, "Superparamagnetic iron oxide nanoparticles," in *MRI Contrast Agents: From Molecules to Particles*, pp. 55–109, Springer, 2016.

[2] J. W. Bulte, "Superparamagnetic iron oxides as mpi tracers: A primer and review of early applications," *Advanced drug delivery reviews*, vol. 138, pp. 293–301, 2019.

[3] K. Motomura, M. Ishitobi, Y. Komoike, H. Koyama, A. Noguchi, H. Sumino, Y. Kumatani, H. Inaji, T. Horinouchi, and K. Nakanishi, "Spio-enhanced magnetic resonance imaging for the detection of metastases in sentinel nodes localized by computed tomography lymphography in patients with breast cancer," *Annals of surgical oncology*, vol. 18, no. 12, pp. 3422–3429, 2011.

[4] M. Douek, J. Klaase, I. Monypenny, A. Kothari, K. Zechmeister, D. Brown, L. Wyld, P. Drew, H. Garmo, O. Agbaje, *et al.*, "Sentinel node biopsy using a magnetic tracer versus standard technique: the sentimag multicentre trial," *Annals of surgical oncology*, vol. 21, no. 4, pp. 1237–1245, 2014.

[5] M. Thill, A. Kurylcio, R. Welter, V. van Haasteren, B. Grosse, G. Berclaz, W. Polkowski, and N. Hauser, "The central-european sentimag study: sentinel lymph node biopsy with superparamagnetic iron oxide (spio) vs. radioisotope," *The breast*, vol. 23, no. 2, pp. 175–179, 2014.

[6] I. T. Rubio, S. Diaz-Botero, A. Esgueva, R. Rodriguez, T. Cortadellas, O. Cordoba, and M. Espinosa-Bravo, "The superparamagnetic iron oxide is equivalent to the tc99 radiotracer method for identifying the sentinel lymph node in breast cancer," *European Journal of Surgical Oncology (EJSO)*, vol. 41, no. 1, pp. 46–51, 2015.

[7] O. Coufal, V. Fait, E. Lžičařová, V. Chrenko, and J. Žaloudík, "Sentimag–the magnetic detection system of sentinel lymph nodes in breast cancer," *Rozhledy v chirurgii: mesicnik Ceskoslovenske chirurgicke spolecnosti*, vol. 94, no. 7, pp. 283–288, 2015.

[8] J.-L. Houpeau, M.-P. Chauvet, F. Guillemin, C. Bendavid-Athias, H. Charitansky, A. Kramar, and S. Giard, "Sentinel lymph node identification using superparamagnetic iron oxide particles versus radioisotope: the french sentimag feasibility trial," *Journal of surgical oncology*, vol. 113, no. 5, pp. 501–507, 2016.

[9] A. Piñero-Madrona, J. Torró-Richart, J. de León-Carrillo, G. de Castro-Parga, J. Navarro-Cecilia, F. Domínguez-Cunchillos, J. Román-Santamaría, C. Fuster-Diana, and R. Pardo-García, "Grupo de estudios senológicos de la sociedad española de patologia mamaria (sespm). superparamagnetic iron oxide as a tracer for sentinel node biopsy in breast cancer: a comparative non-inferiority study," *Eur J Surg Oncol*, vol. 41, no. 8, pp. 991–997, 2015.

[10] M. Ghilli, E. Carretta, F. Di Filippo, C. Battaglia, L. Fustaino, I. Galanou, S. Di Filippo, P. Rucci, M. P. Fantini, and M. Roncella, "The superparamagnetic iron oxide tracer: a valid alternative in sentinel node biopsy for breast cancer treatment," *European journal of cancer care*, vol. 26, no. 4, p. e12385, 2017.

[11] H. Tokoro, A. Namai, and S.-i. Ohkoshi, "Advances in magnetic films of epsilon-iron oxide toward next-generation high-density recording media," *Dalton Transactions*, vol. 50, no. 2, pp. 452–459, 2021.

[12] S.-i. Ohkoshi, A. Namai, and S. Sakurai, "The origin of ferromagnetism in ε-fe2o3 and ε-ga x fe2- x o3 nanomagnets," *The Journal of Physical Chemistry C*, vol. 113, no. 26, pp. 11235–11238, 2009.

[13] S. Ganguly and S. Margel, "Magnetic nanoparticle-based nanopackaging and nano-freezing in food storage applications," *Molecules*, vol. 30, no. 17, p. 3453, 2025.

[14] J. R. Miranda, O. Baffa, R. B. de Oliveira, and N. M. Matsuda, "An ac biosusceptometer to study gastric emptying," *Medical physics*, vol. 19, no. 2, pp. 445–448, 1992.

[15] J. Miranda, R. B. d. Oliveira, P. Sousa, F. Braga, and O. Baffa, "A novel biomagnetic method to study gastric antral contractions," *Physics in Medicine & Biology*, vol. 42, no. 9, pp. 1791–1799, 1997.

[16] C. C. Quini, A. G. Próspero, M. F. Calabresi, G. M. Moretto, N. Zufelato, S. Krishnan, D. R. Pina, R. B. Oliveira, O. Baffa, A. F. Bakuzis, *et al.*, "Real-time liver uptake and biodistribution of magnetic nanoparticles determined by ac biosusceptometry," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 13, no. 4, pp. 1519–1529, 2017.

[17] C. C. Quini, M. F. Américo, L. A. Corá, M. F. Calabresi, M. Alvarez, R. B. Oliveira, and J. A. Miranda, "Employment of a noninvasive magnetic method for evaluation of gastrointestinal transit in rats," *Journal of biological engineering*, vol. 6, no. 1, p. 6, 2012.

[18] G. A. Soares, G. M. Pereira, G. R. Romualdo, G. G. Biasotti, E. G. Stoppa, A. F. Bakuzis, O. Baffa, L. F. Barbisan, and J. R. Miranda, "Biodistribution profile of magnetic nanoparticles in cirrhosis-associated hepatocarcinogenesis in rats by ac biosusceptometry," *Pharmaceutics*, vol. 14, no. 9, p. 1907, 2022.

Caracterização Óptica de Membranas de Nitrocelulose Funcionalizadas com Nanopartículas de Ferrita de Cobalto por Espectroscopia NIR e PCA

Paulo Lucas Albuquerque de Aguiar
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
https://orcid.org/0009-0009-3704-7471

Ialy Fernanda Gonzaga Martins
PPG RENORBIO
Universidade Federal de Pernambuco
Recife, Brasil
https://orcid.org/0009-0006-5828-671x

João Gabriel de Souza Veras
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
https://orcid.org/0000-0003-1351-4812

José Dirceu Vollet Filho
Instituto de Física (IFSC)
Universidade de São Paulo (USP)
São Carlos, Brasil
https://orcid.org/0000-0002-9125-1694

Sebastião Pratavieira
Instituto de Física (IFSC)
Universidade de São Paulo (USP)
São Carlos, Brasil
https://orcid.org/0000-0002-5012-8470

Lilian Tam Moriyama
Instituto de Física (IFSC)
Universidade de São Paulo (USP)
São Carlos, Brasil
https://orcid.org/0000-0002-1351-9253

Eduardo Padrón Hernández
Departamento de Física
Universidade Federal de Pernambuco
Recife, Brasil
https://orcid.org/0000-0002-1492-1199

Rosa Amália Fireman Dutra
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
https://orcid.org/0000-0001-8314-4218

Emery Cleiton Cabral Correia Lins
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
https://orcid.org/0000-0002-7967-6163

***Abstract*—Os dispositivos baseados em membranas de nitrocelulose têm sido amplamente utilizados em aplicações de diagnóstico devido ao baixo custo, simplicidade de fabricação e potencial para análises no ponto de atendimento. Neste trabalho, foi empregada a espectroscopia de refletância difusa na região do infravermelho próximo (900–1700 nm), associada à Análise de Componentes Principais (PCA), para caracterizar as alterações ópticas decorrentes das diferentes etapas de funcionalização de membranas de nitrocelulose contendo nanopartículas de ferrita de cobalto ($CoFe_2O_4$), etilenodiamina (EDA), alérgeno AlgH e imunoglobulina G (IgG). Os espectros obtidos foram submetidos à correção de linha de base, normalização por área e análise quimiométrica. Embora os espectros apresentassem perfis semelhantes por inspeção visual, a PCA permitiu discriminar os diferentes grupos experimentais, explicando 98,0% da variabilidade total dos dados nos dois primeiros componentes principais (PC1 = 91,5% e PC2 = 6,5%). A análise dos loadings identificou a região espectral entre 1650 e 1680 nm como a principal responsável pela separação observada entre as etapas de funcionalização. Os resultados demonstram que a combinação da espectroscopia NIR com técnicas quimiométricas constitui uma abordagem sensível para detectar modificações ópticas em membranas de nitrocelulose funcionalizadas, apresentando potencial para aplicação na caracterização de plataformas destinadas ao desenvolvimento de dispositivos analíticos baseados em papel.**

***Index Terms*—Espectroscopia NIR, refletância difusa, análise de componentes principais, quimiometria, nitrocelulose, ferrita de cobalto.**

I. Introdução

Os dispositivos analíticos baseados em papel (microfluidic paper-based analytical devices – μPADs) têm despertado crescente interesse para aplicações em diagnóstico devido ao baixo custo, facilidade de fabricação, reduzido consumo de reagentes e potencial para utilização em análises no ponto de atendimento (Point-of-Care Testing – POCT) [1]. Entre os materiais empregados nesses dispositivos, a membrana de nitrocelulose destaca-se por apresentar elevada capacidade de transporte capilar, alta afinidade para imobilização de biomoléculas e ampla utilização em testes imunocromatográficos, tornando-se uma plataforma promissora para o desenvolvimento de sistemas de detecção [1], [8].

Durante a preparação dessas plataformas, diferentes etapas de funcionalização química são realizadas para promover a imobilização de nanopartículas, agentes funcionalizantes e biomoléculas sobre a superfície da membrana. Essas modificações alteram as propriedades ópticas do material, produzindo variações espectrais geralmente discretas ao longo do processo de funcionalização. Dessa forma, técnicas ópticas não destrutivas capazes de detectar pequenas alterações espectrais tornam-se ferramentas importantes para acompanhar essas modificações superficiais [2].

Entre essas técnicas, a espectroscopia de refletância difusa na região do infravermelho próximo (Near-Infrared Spectroscopy – NIR) destaca-se por permitir a investigação das propriedades ópticas de materiais porosos a partir da interação da radiação incidente com sua estrutura. Alterações nos processos de absorção e espalhamento da luz, decorrentes da deposição de nanopartículas e das modificações superficiais, podem produzir variações espectrais características para cada etapa da funcionalização [2], [4]. Entretanto, essas diferenças são frequentemente discretas e distribuídas ao longo de centenas de comprimentos de onda, dificultando sua interpretação por inspeção direta dos espectros. Nesse contexto, métodos de análise multivariada, como a Análise de Componentes Principais (PCA), tornam-se fundamentais para extrair informações relevantes, reduzir a dimensionalidade dos dados espectrais e identificar padrões associados às modificações da superfície [5], [10], [11].

Neste trabalho, foram utilizadas nanopartículas de ferrita de cobalto ($CoFe_2O_4$) como plataforma de funcionalização devido às suas propriedades magnéticas e ao seu potencial para aplicações em sistemas de biossensoriamento [3], [4], [9]. As nanopartículas foram funcionalizadas com etilenodiamina (EDA), seguida da imobilização do alérgeno AlgH e da imunoglobulina G (IgG), constituindo diferentes etapas de modificação superficial da membrana de nitrocelulose. Os espectros de refletância difusa foram adquiridos na região do infravermelho próximo (900–1700 nm) e analisados por meio da Análise de Componentes Principais (Principal Component Analysis – PCA), com o objetivo de investigar a capacidade da técnica em diferenciar as alterações ópticas produzidas durante as sucessivas etapas de funcionalização da plataforma.

II. Materiais e Métodos

A. Preparação das Amostras

As amostras foram preparadas utilizando membranas de nitrocelulose empregadas como plataforma para funcionalização superficial. Inicialmente, foram utilizadas membranas sem modificação (controle). Em seguida, realizou-se a deposição de nanopartículas de ferrita de cobalto ($CoFe_2O_4$), sintetizadas em laboratório parceiro, seguida da funcionalização com etilenodiamina (EDA), imobilização do alérgeno AlgH e, por fim, incubação com imunoglobulina G (IgG), constituindo diferentes etapas do processo de funcionalização.

A Tabela I apresenta os grupos experimentais correspondentes às diferentes etapas de funcionalização da membrana de nitrocelulose.

As nanopartículas de ferrita de cobalto ($CoFe_2O_4$) foram fornecidas por um grupo de pesquisa parceiro, responsável por sua síntese, sendo posteriormente empregadas nas etapas de funcionalização desenvolvidas no laboratório. A utilização dessas nanopartículas foi motivada por suas propriedades magnéticas e ópticas, que favorecem aplicações em plataformas de biossensoriamento e modificação superficial [3], [4], [9].

Tabela I
Grupos experimentais correspondentes às diferentes etapas de funcionalização da membrana de nitrocelulose.

Grupo	Descrição
Nitrocelulose	Membrana de nitrocelulose sem modificação (controle).
$CoFe_2O_4$	Membrana contendo nanopartículas de ferrita de cobalto.
$CoFe_2O_4$ + EDA	Funcionalização das nanopartículas com etilenodiamina (EDA).
$CoFe_2O_4$ + EDA + AlgH	Imobilização do alérgeno AlgH sobre a superfície funcionalizada.
$CoFe_2O_4$ + EDA + AlgH + IgG	Formação do imunocomplexo após incubação com IgG específica.

Para cada grupo experimental foram analisadas três membranas independentes, sendo realizada uma aquisição espectral em cada membrana, totalizando três espectros por grupo.

B. Aquisição dos espectros de refletância difusa

Os espectros de refletância difusa foram adquiridos utilizando um espectrômetro portátil baseado em fibra óptica (StellarNet, EUA), operando na região do infravermelho próximo (NIR), entre 900 e 1700 nm. A aquisição dos espectros foi realizada por meio de uma sonda de fibra óptica em configuração de refletância difusa, permitindo a iluminação da superfície da amostra e a coleta da radiação retroespalhada pelo mesmo sistema óptico [6], [7].

A sonda foi posicionada perpendicularmente à superfície da membrana de nitrocelulose, mantendo-se uma distância aproximada de 5 mm durante todas as aquisições, a fim de garantir a reprodutibilidade das medições. O tempo de integração foi fixado em 1000 ms para todas as amostras, mantendo constantes as condições experimentais ao longo do estudo.

A Figura 1 apresenta o sistema experimental utilizado para aquisição dos espectros de refletância difusa.

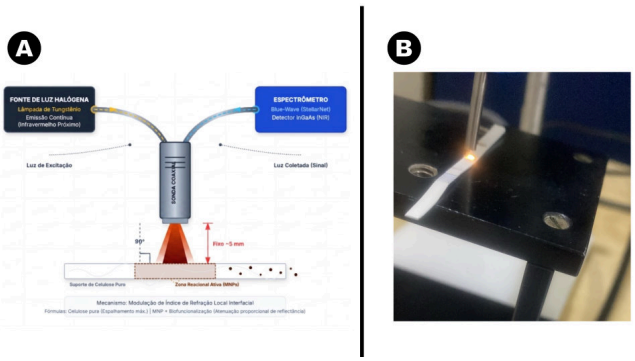


Fig. 1. Setup experimental utilizado para aquisição dos espectros de refletância difusa: (A) esquema da configuração experimental; (B) fotografia do sistema utilizado.

C. Processamento dos Espectros

Os espectros foram processados utilizando um software desenvolvido em Python (PinkApp), desenvolvido especificamente para análise espectral e visualização dos resultados.

Inicialmente, os arquivos espectrais foram importados individualmente e organizados conforme os grupos experimentais. Em seguida, cada espectro foi submetido à correção da linha de base por deslocamento do sinal para o valor mínimo (*baseline offset correction*), seguida de normalização pela soma das intensidades espectrais (*area normalization*), reduzindo diferenças de intensidade absoluta entre as aquisições. Esses procedimentos de pré-processamento são amplamente empregados em análises espectrais para minimizar variações instrumentais e permitir comparações mais consistentes entre as amostras [5], [11].

O software também permitiu selecionar intervalos específicos do espectro para análises localizadas, possibilitando investigar regiões espectrais de maior interesse.

D. Análise de Componentes Principais

A diferenciação entre os grupos experimentais foi realizada por meio da Análise de Componentes Principais (*Principal Component Analysis* – PCA), implementada na biblioteca *scikit-learn* (Python).

A PCA foi aplicada aos espectros normalizados com o objetivo de reduzir a dimensionalidade dos dados e identificar padrões de agrupamento associados às diferentes etapas de funcionalização da membrana de nitrocelulose [5].

Os dois primeiros componentes principais (PC1 e PC2) foram utilizados para construção dos gráficos de dispersão (*scores plots*), sendo apresentada a porcentagem de variância explicada por cada componente.

Além da projeção dos escores, foram analisados os vetores de carregamento (*loadings*), permitindo identificar os comprimentos de onda que mais contribuíram para a separação entre os grupos experimentais [10], [11].

E. Análise dos Loadings da PCA

Os vetores de carregamento (loadings) obtidos pela Análise de Componentes Principais (PCA) foram utilizados para identificar os comprimentos de onda com maior contribuição para a separação entre os grupos experimentais. Os valores absolutos dos loadings foram analisados para determinar as regiões espectrais mais relevantes na discriminação das diferentes etapas de funcionalização da membrana de nitrocelulose [5], [10].

III. Resultados e Discussão

A. Análise Espectral dos Grupos

A Figura 2 apresenta os espectros de refletância difusa adquiridos na região do infravermelho próximo (NIR), entre 900 e 1700 nm, para as diferentes etapas de funcionalização da membrana de nitrocelulose. De maneira geral, os espectros apresentaram perfis semelhantes, indicando que as modificações promovidas pela deposição das nanopartículas de ferrita de cobalto ($CoFe_2O_4$) e pelas etapas subsequentes de funcionalização produziram alterações espectrais discretas, de difícil identificação por inspeção visual. Esse comportamento é esperado em materiais porosos analisados por espectroscopia de refletância, nos quais pequenas modificações superficiais podem resultar em variações sutis nos processos de absorção e espalhamento da radiação incidente [2], [4].

Após o pré-processamento dos sinais, envolvendo correção da linha de base e normalização por área, observou-se melhor definição das diferenças de intensidade entre os grupos experimentais. A distribuição dos valores de pico máximo revelou variações ao longo das etapas de funcionalização, indicando que a incorporação das nanopartículas de $CoFe_2O_4$, seguida da imobilização da EDA, do alérgeno AlgH e da IgG, promove alterações nas propriedades ópticas da membrana de nitrocelulose, modificando sua interação com a radiação NIR. Entretanto, apesar dessas diferenças, ainda existe elevada sobreposição entre os espectros, evidenciando que a análise visual isolada não é suficiente para discriminar de forma confiável todas as etapas do processo, justificando a aplicação de técnicas quimiométricas para extração das informações relevantes [5].

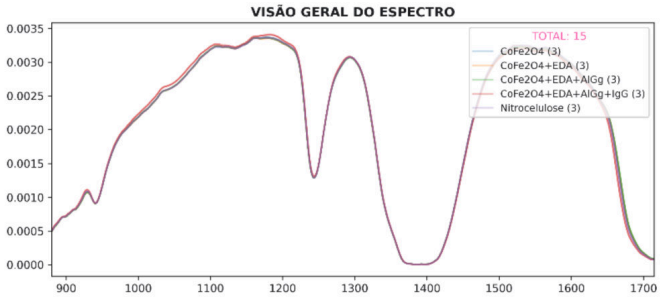


Fig. 2. Espectros de refletância difusa das diferentes etapas de funcionalização da membrana de nitrocelulose obtidos na região do infravermelho próximo (900–1700 nm).

B. Análise de Componentes Principais (PCA)

Com o objetivo de avaliar a capacidade de discriminação entre as diferentes etapas de funcionalização da membrana de nitrocelulose, foi aplicada a Análise de Componentes Principais (*Principal Component Analysis* – PCA). Os dois primeiros componentes principais explicaram 98,0% da variabilidade total dos dados, sendo 91,5% atribuídos à PC1 e 6,5% à PC2, indicando que a maior parte da informação espectral foi representada em um espaço bidimensional [5], [9]. Conforme apresentado na Figura 3, a projeção dos escores (*scores*) evidenciou a formação de agrupamentos distintos para cada grupo experimental. A membrana de nitrocelulose sem modificação apresentou um agrupamento claramente separado das amostras funcionalizadas. A deposição das nanopartículas de ferrita de cobalto ($CoFe_2O_4$) e a posterior funcionalização com etilenodiamina (EDA) promoveram deslocamentos progressivos no espaço multivariado, enquanto as etapas de imobilização do alérgeno (AlgH) e da imunoglobulina G (IgG)

originaram agrupamentos bem definidos, evidenciando que cada etapa do processo produziu alterações ópticas detectáveis pela espectroscopia NIR. Observou-se ainda baixa dispersão entre as três réplicas de cada grupo experimental, indicando consistência dos perfis espectrais obtidos nas condições avaliadas. A baixa sobreposição entre os agrupamentos demonstra que a combinação da espectroscopia de refletância no infravermelho próximo com a PCA foi capaz de diferenciar as sucessivas etapas de funcionalização da membrana de nitrocelulose, mesmo diante de diferenças pouco perceptíveis nos espectros originais [5], [9].

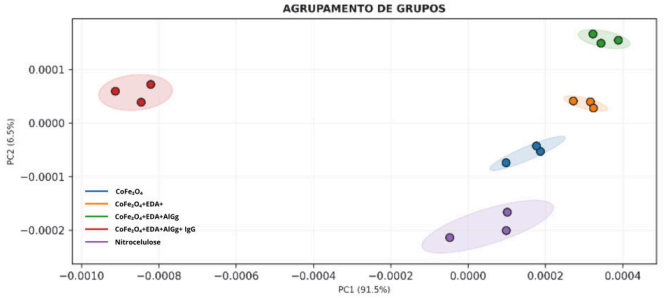


Fig. 3. Projeção dos escores da Análise de Componentes Principais (PCA) obtida a partir dos espectros de refletância difusa das diferentes etapas de funcionalização da membrana de nitrocelulose.

C. Importância dos Comprimentos de Onda (Loadings)

A análise dos *loadings* permitiu identificar os comprimentos de onda que mais contribuíram para a separação observada na Análise de Componentes Principais (PCA). Conforme apresentado na Figura 4, a região compreendida entre aproximadamente 1650 e 1680 nm apresentou maior contribuição para a separação, indicando que essa faixa espectral concentrou as principais diferenças entre as etapas de funcionalização da membrana de nitrocelulose [8].

Os maiores valores de *loading* foram observados em 1665,75 nm para a PC1 (*loading* = 0,235) e em 1669,25 nm para a PC2 (*loading* = 0,174), indicando que esses comprimentos de onda apresentaram maior influência na discriminação entre os grupos experimentais.

Esses resultados indicam que as modificações realizadas ao longo das etapas de funcionalização produziram alterações na resposta espectral da membrana de nitrocelulose, sendo a região entre 1650 e 1680 nm aquela que apresentou maior contribuição para a separação observada entre os grupos. No infravermelho próximo, as respostas espectrais resultam predominantemente de sobretons e bandas de combinação de vibrações moleculares, de modo que alterações na composição e no ambiente químico da superfície podem modificar o perfil espectral registrado. Nesse contexto, a região identificada pelos *loadings* pode refletir mudanças associadas às sucessivas etapas de funcionalização da membrana. Estudos envolvendo sistemas de nitrocelulose e espectroscopia NIR demonstram que essa técnica é capaz de responder a diferenças relacionadas à composição química desses materiais [8]. Entretanto, a atribuição dessa região espectral a um grupo químico ou componente específico não pode ser estabelecida exclusivamente a partir da análise NIR e dos *loadings*, sendo necessárias técnicas complementares, como FTIR-ATR, para uma confirmação estrutural.

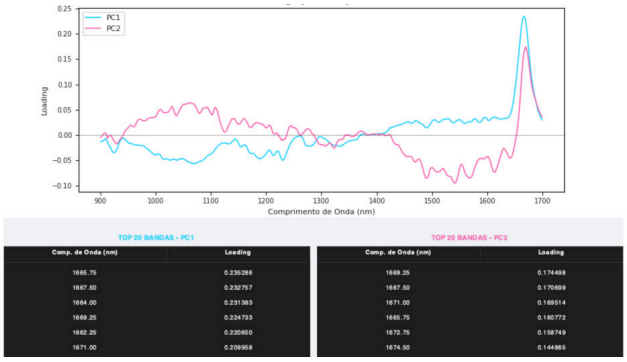


Fig. 4. Perfil dos loadings em função do comprimento de onda e principais regiões espectrais responsáveis pela separação dos grupos experimentais na Análise de Componentes Principais (PCA).

IV. Conclusão

Neste trabalho, a espectroscopia de refletância difusa na região do infravermelho próximo (900–1700 nm), associada à Análise de Componentes Principais (PCA), foi utilizada para avaliar as diferentes etapas de funcionalização de membranas de nitrocelulose contendo nanopartículas de ferrita de cobalto ($CoFe_2O_4$).

Os resultados demonstraram que pequenas alterações espectrais, dificilmente perceptíveis por inspeção visual, puderam ser discriminadas pela PCA, que explicou 98,0% da variabilidade total dos dados utilizando os dois primeiros componentes principais. A análise dos *loadings* identificou a região espectral entre 1650 e 1680 nm como aquela que apresentou maior contribuição para a separação entre os grupos experimentais, indicando que essa faixa pode estar associada às alterações espectrais decorrentes das sucessivas etapas de funcionalização.

Dessa forma, os resultados obtidos fornecem evidências preliminares do potencial da espectroscopia NIR associada à análise quimiométrica como ferramenta para a caracterização de superfícies funcionalizadas em membranas de nitrocelulose. Estudos futuros com um número maior de amostras e técnicas espectroscópicas complementares, como FTIR-ATR, serão importantes para avaliar a robustez dos resultados e aprofundar a atribuição físico-química das alterações espectrais observadas.

References

[1] E. Noviana *et al.*, "Microfluidic paper-based analytical devices: From design to applications," *Chemical Reviews*, vol. 121, no. 19, pp. 11 835–11 885, October 2021.

[2] G. Jacucci, L. Schertel, Y. Zhang, H. Yang, and S. Vignolini, "Light management with natural materials: From whiteness to transparency," *Advanced Materials*, vol. 33, no. 28, p. 2001215, July 2021.

[3] J. Wang *et al.*, "Extraction of backscattering and absorption coefficients of magnetite nanosphere composites from light-scattering measurements: Implications for optomagnetic sensing," *ACS Applied Nano Materials*, vol. 3, no. 11, pp. 11 172–11 183, November 2020.

[4] K. Yeneayehu, T. Senbeta, and B. Mesfin, "Enhancement of the optical response of Fe_3O_4@ag core–shell nanoparticles," *Physica E: Low-dimensional Systems and Nanostructures*, vol. 134, p. 114822, October 2021.

[5] V. N. Ataide *et al.*, "Combining chemometrics and paper-based analytical devices for sensing: An overview," *Trends in Analytical Chemistry*, vol. 164, p. 117091, July 2023.

[6] Y. Liu and W. Peng, "Fiber-optic surface plasmon resonance sensors and biochemical applications: A review," *Journal of Lightwave Technology*, vol. 39, no. 12, pp. 3781–3791, June 2021.

[7] D. Harasim, "Temperature-insensitive bending measurement method using optical fiber sensors," *Sensors and Actuators A: Physical*, vol. 332, p. 113207, December 2021.

[8] T. A. de Carvalho *et al.*, "Determination of nitrogen and acrylic / styrene components in nitrocellulose systems by UATR and NIRA infrared techniques," *Polymer Testing*, vol. 93, p. 106962, January 2021.

[9] E. Gharibshahi, B. D. Young, A. S. Bhalla, and R. Guo, "Theory, simulation and experiment of optical properties of cobalt ferrite ($CoFe_2O_4$) nanoparticles," *Journal of Materials Science & Technology*, vol. 57, pp. 180–187, November 2020.

[10] J. R. Beattie and F. W. L. Esmonde-White, "Exploration of principal component analysis: Deriving principal component analysis visually using spectra," *Applied Spectroscopy*, vol. 75, no. 4, pp. 399–412, April 2021.

[11] P. Oliveri, C. Malegori, R. Simonetti, and M. Casale, "The impact of signal pre-processing on the final interpretation of analytical outcomes—a tutorial," *Analytica Chimica Acta*, vol. 1058, pp. 9–17, 2019.

[12] C. N. Ndung'u, M. I. Kaniu, and J. M. Wanjohi, "Optimization of diffuse reflectance spectroscopy measurements for direct and rapid screening of pesticides: A case study of spinach," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 280, p. 121556, November 2022.

[13] R. G. Brereton, "Principal components analysis with several objects and variables," *Journal of Chemometrics*, vol. 37, no. 4, p. e3408, 2022.

Clinical Validation of VMAT Commissioning with 6FFF Beams Using AAPM TG-119

Kiscilla S. Borges[1], Adrielle M. Leonardo[2], Rodrigo A. Santana[1], Gean S. Santos[2], Flávia C. F. Araújo[2], Emerson N. Itikawa[1]

[1]*Instituto de Física, Universidade Federal de Goiás, Goiânia, Brazil*
[2]*Serviço de Radioterapia, Associação de Combate ao Câncer em Goiás, Goiânia, Brazil*

Abstract *— This study evaluated the institutional implementation of the American Association of Physicists in Medicine TG-119 protocol for volumetric modulated arc therapy (VMAT) on an Elekta Infinity linear accelerator using 6 MV flattening-filter-free (6FFF) photon beams and the Monaco treatment planning system (TPS). The AP/PA and Bands preliminary tests and the Multi-Target, Prostate, Head-and-Neck, C-Shape Easy, and C-Shape Hard cases were analyzed. Point doses were measured with a PTW31016 PinPoint 3D ionization chamber, and planar dose distributions were verified with an ArcCHECK diode array. Local gamma analysis used the institutional 3%/3 mm criterion with a 10% low-dose threshold. Passing rates ranged from 94.3% to 99.7%, with a mean of 96.89%, a standard deviation of 2.06%, and a confidence limit of 7.15%. Corrected mean point-dose deviations were −1.6% in the high-dose region and −2.5% in the low-dose region; the Head-and-Neck low-dose point was −3.9%. In the C-Shape Hard case, the planned core D5 and D10 objectives were not achieved, while the aggregate gamma passing rate was 94.3%. These results provide an institution-specific dosimetric baseline for the tested Elekta Infinity, Monaco, and 6 MV FFF configuration. Dose-volume histogram objectives, point-dose measurements, and gamma analysis should be interpreted together because they evaluate different aspects of the process.*

Keywords *— radiotherapy; VMAT; TG-119; quality assurance; point dose; gamma index; 6 MV FFF*

I. INTRODUCTION

Volumetric modulated arc therapy (VMAT) delivers radiation during continuous gantry rotation, with simultaneous modulation of multileaf collimator (MLC) aperture, dose rate, and gantry speed. This combination enables conformal dose distributions with shorter delivery times and, in many scenarios, fewer monitor units than static-field intensity-modulated techniques [1]–[3].

Flattening-filter-free (FFF) beams remove the component used to homogenize the beam's transverse profile. As a result, they present a non-flat profile, spectral changes, and higher dose rates [4], [5]. Clinical use of this modality requires that the treatment planning system adequately model the beam, and that the planning, transfer, and delivery chain be tested against situations of known complexity.

The AAPM TG-119 report proposes a standardized set of tests for IMRT commissioning. The protocol includes AP/PA fields, a Bands test, and cases spanning different levels of geometric and dosimetric complexity: Multi-Target, Prostate, Head-and-Neck, C-Shape Easy, and C-Shape Hard [6]. Planning objectives, point-dose measurements, and dose-distribution comparisons allow local reference values to be established and limitations to be identified before routine clinical use of the technique.

Gamma analysis combines dose difference and distance-to-agreement into a single metric [7], [8]. However, the passing rate depends on the detector, normalization, dose threshold, and criteria selected; moreover, an aggregate value does not replace evaluation of the specific objectives of the dose-volume histogram (DVH) [9]–[12]. For this reason, comparison with multi-institutional results must account for differences in equipment and methods.

The objective of this study was to evaluate the implementation of the TG-119 tests for VMAT with a 6 MV FFF beam on an Elekta Infinity accelerator and Monaco treatment planning system, establishing an institutional baseline through plan objectives, point-dose measurements, and ArcCHECK gamma analysis.

II. MATERIALS AND METHODS

A. Linear Accelerator and Treatment Planning System

Plans were generated on an Elekta Infinity linear accelerator equipped with an Agility multileaf collimator, comprising 160 leaves with a 0.5 cm projection at isocenter and a nominal maximum speed of 6.5 cm/s [13]. The 6 MV FFF photon beam was used, with a maximum dose rate of 1400 MU/min.

The Monaco treatment planning system, version 6.1.3, was used for optimization and calculation. Fluence optimization employed the Finite Size Pencil Beam algorithm, followed by final dose calculation using a voxel-based Monte Carlo algorithm. The images and structures provided for the TG-119 cases were imported into the system and registered to the measurement geometries [6], [14], [15].

B. Dosimetry Systems

Dose-distribution verification was performed with ArcCHECK, a Sun Nuclear cylindrical diode array. The device contains 1,386 SunPoint diodes arranged helically, with 1 cm spacing, 21 cm inner diameter, and 27 cm outer diameter. The cylindrical geometry allows entrance and exit signals to be recorded during arc deliveries [16]–[18]. The phantom was CT-scanned on an Aquilion Lightning (Canon) at 120 kV, 170 mA, 0.813 mm pitch, and 2 mm slice thickness. The images were imported into Monaco.

Point doses were measured with a PinPoint 3D ionization chamber (PTW31016), positioned at 10 cm depth in solid-water slabs with total dimensions of 30 × 30 × 20 cm³. Absorbed dose to water was obtained from the corrected chamber reading and the calibration coefficient:

$$D_{w,Q} = M \cdot N_{D,w,Q} \cdot k_{T,P} \quad (1)$$

where M is the corrected chamber reading, ND,w,Q is the absorbed-dose-to-water calibration coefficient, and kT,P is the temperature–pressure correction factor. This factor was calculated as:

$$k_{T,P} = [(273.2 + T) \cdot P0] / [(273.2 + T0) \cdot P] \quad (2)$$

where T and P are the temperature and pressure recorded during measurement, and T0 = 20 °C and P0 = 101.3 kPa are the reference conditions.

C. TG-119 Test Cases

Seven tests were implemented: AP/PA, Bands, Multi-Target, Prostate, Head-and-Neck, C-Shape Easy, and C-Shape Hard. The AP/PA test used opposing 10 × 10 cm² fields with 2 Gy at mid-plane. The Bands test used 3 cm-wide strips with a 0.4–2 Gy gradient. The remaining five cases used the structures, prescriptions, and constraints provided by the AAPM [6].

The point-dose fractional deviation was calculated relative to the prescribed dose and expressed as a percentage:

$$r = (D_{meas} - D_{plan}) / D_{presc} \times 100 \quad (3)$$

Deviations were grouped into high-dose and low-dose regions. The point-dose confidence limit was determined as:

$$CL_{dose} = |mean| + 1.96\sigma \quad (4)$$

D. Gamma Analysis

ArcCHECK measurements were compared to distributions calculated in SNC Patient software (version 6.2). The institutional local gamma criterion of 3%/3 mm was applied, with a 10% low-dose threshold and a minimum passing rate of 90%. The confidence limit was calculated as:

$$CL_{\gamma} = (100 - mean) + 1.96\sigma \quad (5)$$

Accordingly, 95% of results are expected to show a passing rate above (100 − CLγ)%. TG-218 was used only as a contemporary reference for methodological interpretation, since its universal recommendation of 3%/2 mm with global normalization and a 10% threshold is not identical to the local configuration applied in this dataset [9].

III. RESULTS

A. Preliminary Tests

Point-dose and gamma results for the AP/PA and Bands tests are shown in Table I. Deviations were +1.0% and −2.1%, respectively.

TABLE I. AP/PA AND BANDS RESULTS

Test	Meas. (cGy)	Calc. (cGy)	Deviation	Gamma (%)
AP/PA	198.0	196.0	+2.0 cGy (+1.0%)	98.9
Bands	137.1	140.0	−2.9 cGy (−2.1%)	99.7

Source: study data.

B. Plan Objectives

Tables II–VI present the TG-119 objectives and the values obtained after optimization. The main PTV objectives were met in all five cases. In C-Shape Hard, the core dose objectives were not met.

TABLE II. MULTI-TARGET PLAN

Parameter	TG-119 Objective (cGy)	Achieved (cGy)
D99 Central Target	≥5000	5008.9
D10 Central Target	≤5300	5193.3
D99 Superior Target	≥2500	2562.9
D10 Superior Target	≤3500	3181.0
D99 Inferior Target	≥1250	1253.6
D10 Inferior Target	≤2500	2417.8

Source: study data.

TABLE III. PROSTATE PLAN

Parameter	TG-119 Objective (cGy)	Achieved (cGy)
D95 PTV	≥7560	7600.0
D5 PTV	≤8300	7999.1
D30 Rectum	≤7000	5177.6
D10 Rectum	≤7500	7222.0
D30 Bladder	≤7000	3895.9
D10 Bladder	≤7500	6544.9

Source: study data.

TABLE IV. HEAD-AND-NECK PLAN

Parameter	TG-119 Objective (cGy)	Achieved (cGy)
D90 PTV	≥5000	5063.7
D99 PTV	≥4650	4741.7
D20 PTV	≤5500	5247.0
Dmax Spinal Cord	≤4000	3712.1
D50 Right Parotid	≤2000	1763.0
D50 Left Parotid	≤2000	1964.5

Source: study data.

TABLE V. C-SHAPE EASY PLAN

Parameter	TG-119 Objective (cGy)	Achieved (cGy)
D95 PTV	≥5000	5170.3
D10 PTV	≤5500	5494.7
D5 Core	≤2500	2426.4
D10 Core	≤2500	2296.2

Source: study data.

TABLE VI. C-SHAPE HARD PLAN

Parameter	TG-119 Objective (cGy)	Achieved (cGy)
D95 PTV	≥5000	5095.4
D10 PTV	≤5500	5330.2
D5 Core	≤1000	1522.3*
D10 Core	≤1000	1487.0*

** Planning objective not met. Source: study data.*

C. Point-Dose Measurements

Point-dose deviations, correctly converted from fractional ratio to percentage, are presented in Table VII. Means were −1.6% in the high-dose region and −2.5% in the low-dose region. All high-dose results were within ±3%. In the low-dose region, the Head-and-Neck case showed −3.9%, exceeding this criterion.

TABLE VII. POINT-DOSE DEVIATIONS

Plan	High Dose (%)	Low Dose (%)
Multi-Target	−1.8	−1.9
Prostate	−1.2	−1.5
Head-and-Neck	−1.7	−3.9*
C-Shape Easy	−1.8	−2.3
C-Shape Hard	−1.7	−2.8
Mean	**−1.6**	**−2.5**
σ	0.25	0.93
CL	2.1	4.3

** Value above ±3%. Values were originally recorded as fractional ratios and are presented here as percentages. Source: study data.*

D. Planar Dose Measurements

Gamma passing rates ranged from 94.3% to 99.7% (Table VIII). All tests exceeded the institutional minimum criterion of 90%. The mean was 96.89%, the standard deviation was 2.06%, and the confidence limit was 7.15% (Table IX). The value corresponding to 100 − CLγ was 92.85%.

TABLE VIII. POINTS PASSING THE LOCAL 3%/3 MM GAMMA CRITERION

Plan	This Study (%)	TG-119 (%)
AP/PA	98.9	—
Bands	99.7	99.4
Multi-Target	97.1	99.1
Prostate	97.8	98.0
Head-and-Neck	95.5	96.2
C-Shape Easy	94.9	97.6
C-Shape Hard	94.3	94.4

TG-119 values are presented as descriptive reference only. Source: study data and Ezzell et al. [6].

TABLE IX. GAMMA ANALYSIS STATISTICS

Parameter	This Study	TG-119
Mean	96.89%	97.9%
Standard deviation	2.06%	2.5%
Confidence limit	7.15%	7.0%

Source: study data and Ezzell et al. [6].

IV. DISCUSSION

The TG-119 protocol let us test planning and delivery across geometries of progressively increasing modulation complexity. PTV objectives were met in every case. Gamma values were lowest for the two C-Shape geometries — the most complex in the set — at 94.9% for C-Shape Easy and 94.3% for C-Shape Hard.

Once corrected to the proper scale, the point-dose results read differently. The high-dose mean was −1.6%; the low-dose mean was −2.5%. Both fall within ±3%, but the low-dose figure sits close to that limit, and the Head-and-Neck case alone reached −3.9%, past the threshold. All point measurements did not, in fact, meet the criterion, and the wider spread at low dose is reflected in the confidence limit as well: 4.3%, against 2.1% for high dose.

At 96.89%, our gamma mean sits 1.01 percentage points below the TG-119 multi-institutional mean of 97.9%; the confidence limit, 7.15%, is 0.15 points above the reported 7.0% [6]. This closeness should be read as descriptive rather than as evidence of equivalence — the treatment planning systems, accelerators, energies, detectors, measurement geometries, and analysis settings behind the two datasets were not the same.

C-Shape Hard met its PTV objectives but not its core limits: D5 reached 1522.3 cGy and D10 reached 1487.0 cGy, both against a ≤1000 cGy target. These figures come from the TPS-calculated DVH and represent an objective the plan simply did not meet. The 94.3% gamma passing rate, by contrast, reflects the aggregate agreement between the calculated and measured distributions across the whole phantom — a different question entirely. One metric passing does not mean the other did too: gamma agreement does not establish that the core objective was satisfied, just as the DVH exceedance on its own does not prove a delivery error [10], [11], [19].

It is tempting to attribute this to the concave geometry of C-Shape Hard, which plausibly adds complexity to optimization and MLC sequencing. This study, however, did not analyze the Agility log files, introduce controlled leaf-position perturbations, repeat the case independently, or compare the 6 MV FFF beam against a flattened beam on the same unit. Absent that work, geometry, MLC behavior, and FFF modality remain hypotheses here, not demonstrated causes.

Reading the gamma passing rate also means reading its configuration. Our 10% low-dose threshold matches the TG-218 recommendation, but our local 3%/3 mm criterion is not the same as the universal 3%/2 mm with global normalization and absolute dose that TG-218 recommends [9], so direct conformity with an identical metric cannot be

claimed. What this dataset does offer is a baseline for this specific local configuration — useful for tracking trends over time, provided the same analysis parameters are kept constant.

A further, detector-level limitation is worth noting: ArcCHECK samples dose on a fixed cylindrical surface, and its response shifts with field size, modulation, and target position relative to the diodes [16]–[18]. This is part of why point dose, gamma, and DVH objectives are best read together rather than relying on the passing rate alone. Three-dimensional reconstruction, independent dose calculation, and log-file analysis are workflows that could sharpen the ability to localize and interpret discrepancies of this kind going forward [12], [19]–[21].

V. CONCLUSION

We implemented the TG-119 tests for VMAT with a 6 MV FFF beam across the Elekta Infinity, Monaco, Agility, PinPoint 3D, and ArcCHECK set, giving this configuration an institutional dosimetric baseline. The mean gamma passing rate came to 96.89%, with a confidence limit of 7.15%. Point-dose deviations averaged −1.6% at high dose and −2.5% at low dose, with the Head-and-Neck low-dose point reaching −3.9%.

C-Shape Hard is the case that warrants the most caution: its core dose objectives went unmet in planning even though the aggregate gamma passing rate reached 94.3%. That gap illustrates, more clearly than any other result here, why delivery agreement and structural-objective evaluation must be read as complementary rather than interchangeable. Nothing in the present design allows the result to be attributed specifically to the FFF beam, MLC sequencing, or the concave geometry; confirming any of these would require repeated measurements and controlled comparisons not yet performed.

ACKNOWLEDGMENT

The authors thank the Radiotherapy Service of Associação de Combate ao Câncer em Goiás for technical support, and the Instituto de Física, Universidade Federal de Goiás, for academic infrastructure.

REFERENCES

[1] C. X. Yu, "Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy," Phys. Med. Biol., vol. 40, no. 9, pp. 1435–1449, 1995.

[2] K. Otto, "Volumetric modulated arc therapy: IMRT in a single gantry arc," Med. Phys., vol. 35, no. 1, pp. 310–317, 2008.

[3] J. L. Bedford, "Treatment planning for volumetric modulated arc therapy," Med. Phys., vol. 36, no. 11, pp. 5128–5138, 2009.

[4] O. N. Vassiliev, U. Titt, F. Pönisch, S. F. Kry, R. Mohan, and M. T. Gillin, "Dosimetric properties of photon beams from a flattening filter free clinical accelerator," Phys. Med. Biol., vol. 51, no. 7, pp. 1907–1917, 2006.

[5] D. Georg, T. Knöös, and B. McClean, "Current status and future perspective of flattening filter free photon beams," Med. Phys., vol. 38, no. 3, pp. 1280–1293, 2011.

[6] G. A. Ezzell, J. W. Burmeister, N. Dogan, et al., "IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," Med. Phys., vol. 36, no. 11, pp. 5359–5373, 2009.

[7] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," Med. Phys., vol. 25, no. 5, pp. 656–661, 1998.

[8] T. Depuydt, A. Van Esch, and D. P. Huyskens, "A quantitative evaluation of IMRT dose distributions: refinement and clinical assessment of the gamma evaluation," Radiother. Oncol., vol. 62, no. 3, pp. 309–319, 2002.

[9] M. Miften, A. Olch, D. Mihailidis, et al., "Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218," Med. Phys., vol. 45, no. 4, pp. e53–e83, 2018.

[10] B. E. Nelms, H. Zhen, and W. A. Tomé, "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors," Med. Phys., vol. 38, no. 2, pp. 1037–1044, 2011.

[11] C. Stambaugh and G. Ezzell, "A clinically relevant IMRT QA workflow: design and validation," Med. Phys., vol. 45, no. 4, pp. 1391–1399, 2018.

[12] B. E. Nelms, D. Opp, J. Robinson, et al., "VMAT QA: measurement-guided 4D dose reconstruction on a patient," Med. Phys., vol. 39, no. 7, pp. 4228–4238, 2012.

[13] J. L. Bedford, M. D. R. Thomas, and G. Smyth, "Evaluation of the Elekta Agility multileaf collimator for radiotherapy," J. Appl. Clin. Med. Phys., vol. 14, no. 2, art. 4136, 2013.

[14] J. E. Snyder, D. E. Hyer, R. T. Flynn, A. Boczkowski, and D. Wang, "Commissioning and validation of the Monaco treatment planning system on an Elekta VersaHD linear accelerator," J. Appl. Clin. Med. Phys., vol. 20, no. 1, pp. 184–193, 2019.

[15] M. R. Paudel, A. Kim, A. Sarfehnia, et al., "Experimental evaluation of a GPU-based Monte Carlo dose calculation algorithm in the Monaco treatment planning system," J. Appl. Clin. Med. Phys., vol. 17, no. 6, pp. 230–241, 2016.

[16] G. Li, Y. Zhang, X. Jiang, et al., "Calibration of a novel four-dimensional diode array," Med. Phys., vol. 37, no. 1, pp. 108–115, 2010.

[17] V. Chaswal, M. Weldon, N. Gupta, A. Chakravarti, and Y. Rong, "Commissioning and comprehensive evaluation of the ArcCHECK cylindrical diode array for VMAT pretreatment delivery QA," J. Appl. Clin. Med. Phys., vol. 15, no. 4, art. 4832, 2014.

[18] C. Neilson, M. Klein, R. Barnett, and S. Yartsev, "Delivery quality assurance with ArcCHECK," Med. Dosim., vol. 38, no. 1, pp. 77–80, 2013.

[19] M. W. Geurts, D. J. Jacqmin, L. E. Jones, et al., "AAPM medical physics practice guideline 5.b: commissioning and QA of treatment planning dose calculations — megavoltage photon and electron beams," J. Appl. Clin. Med. Phys., vol. 23, no. 9, e13641, 2022.

[20] P. Szeverinski, M. Kowatsch, F. Poller, et al., "Evaluation of 4-Hz log files and secondary Monte Carlo dose calculation as patient-specific quality assurance for VMAT prostate plans," J. Appl. Clin. Med. Phys., vol. 22, no. 7, pp. 235–244, 2021.

[21] L. Kumar, G. Yadav, V. Kishore, et al., "Analyzing the performance of ArcCHECK diode array detector for VMAT plan," Rep. Pract. Oncol. Radiother., vol. 21, no. 1, pp. 50–56, 2016.

Desenvolvimento de uma Plataforma IoMT para Monitoramento de Sinais Vitais e Detecção de Quedas

1st Alan Sovano Gomes
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Pará, Brasil
sovano@ufpa.br

2nd Thiago Rodrigues Sales
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Pará, Brasil
thiago.sales@itec.ufpa.br

3rd Isabela Marques Miziara
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Pará, Brasil
immiziara@ufpa.br

4th Maria da Conceição Pereira Fonseca
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Pará, Brasil
conceicao@ufpa.br

***Resumo*—Este artigo apresenta um protótipo de plataforma de Internet das Coisas Médicas (IoMT) destinado ao monitoramento de sinais vitais (frequência cardíaca e saturação de oxigênio) e à detecção de quedas, com foco na arquitetura de software que viabiliza seu funcionamento nesse contexto. Três decisões de software se mostram centrais para garantir a operação confiável e segura da plataforma: o sistema operacional de tempo real FreeRTOS, responsável por garantir a execução concorrente e determinística das tarefas internas do dispositivo; o protocolo de mensageria MQTT (*Message Queuing Telemetry Transport*), responsável por transmitir dados e alertas de forma eficiente mesmo sob recursos limitados; e o protocolo de segurança *Transport Layer Security* (TLS), responsável por proteger a confidencialidade de dados fisiológicos e de localização transmitidos sem fio. O artigo detalha o funcionamento interno de cada uma dessas tecnologias e discute por que cada escolha se torna particularmente relevante quando o sistema opera no contexto de saúde conectada, apresentando ao final evidências de que a integração entre essas camadas operou de forma coerente no protótipo desenvolvido.**

***Index Terms*—FreeRTOS; Internet das Coisas Médicas; Segurança de Dados em Saúde; Sistemas Embarcados de Tempo Real**

I. Introdução

A Internet das Coisas Médicas, ou *Internet of Medical Things* (IoMT), é o conceito associado a sistemas de saúde inteligentes e conectados, capazes de fornecer dados fisiológicos de um sujeito a profissionais de saúde, familiares ou cuidadores de forma automatizada e à distância [1]. Esse conceito ganha relevância particular quando aplicado a populações que necessitam de cuidado contínuo e vivem de forma independente. Como exemplo, estima-se que cerca de 30% dos idosos com 65 anos sofram ao menos uma queda por ano [2], evento que frequentemente exige socorro imediato. Pode-se citar também o caso das pessoas com Alzheimer, com cerca de 60% apresentando tendência à deambulação, processo de vaguear sem rumo, o que aumenta o risco de acidentes e de desorientação em ambientes externos [3]. Ambos os cenários convergem para a mesma necessidade prática: um dispositivo capaz de monitorar parâmetros fisiológicos fundamentais, detectar eventos críticos e de informar a localização da pessoa a um cuidador, de forma automática e em tempo real, para a prestação dos cuidados necessários.

A partir do problema exposto, foi desenvolvido um protótipo de dispositivo vestível capaz de monitorar frequência cardíaca, saturação de oxigênio e eventos de queda, acionando alarmes locais e remetendo alertas incluindo coordenadas de localização a um *broker* MQTT (*Message Queuing Telemetry Transport*), através de uma conexão Wi-Fi local autenticada, com toda a concorrência interna gerenciada pelo sistema operacional de tempo real FreeRTOS e a pilha de rede gerenciada pela biblioteca *Lightweight IP* (lwIP) [4]. A comunicação via MQTT é, ainda, criptografada a partir do protocolo *Transport Layer Security* (TLS), de forma a garantir confidencialidade para o tráfego de informações do usuário.

Desta forma, o presente trabalho tem como objetivo analisar e justificar as principais decisões de arquitetura empregadas no desenvolvimento do dispositivo IoMT. O foco principal é na camada de software, visto que viabilizar esse tipo de plataforma depende diretamente de decisões de software capazes de sustentar a concorrência interna, a comunicação eficiente e a segurança dos dados exigidas em um dispositivo IoMT. Além disso, busca-se discutir se o dispositivo encontra-se em um estágio funcional, com uma operação adequada em tempo real.

II. IoMT e as Restrições de um Sistema Embarcado Crítico

Dispositivos vestíveis voltados ao monitoramento fisiológico tornaram-se amplamente populares na última década [5], [6], sendo aplicados tanto ao desempenho esportivo quanto

ao acompanhamento de saúde em comunidades afastadas de grandes centros. Quando esse mesmo tipo de dispositivo é aplicado ao cuidado de pessoas com condições crônicas ou críticas, contudo, as restrições técnicas comuns a qualquer sistema embarcado (como processamento limitado, energia limitada e conectividade nem sempre estável) deixam de ser apenas problemas de desempenho e passam a ter implicação direta na segurança do usuário.

Diante disso, as necessidades de dispositivos IoMT podem ser, em um primeiro momento, organizadas em três tipos de exigência, cada uma delas mapeada para uma das decisões de software discutidas na Seção IV:

- **Exigência de tempo real e confiabilidade interna.** Um sistema que precisa adquirir dados de múltiplos sensores em taxas distintas, processar um algoritmo de detecção de evento crítico e, ao mesmo tempo, manter uma interface e uma conexão de rede ativas, não pode tratar essas tarefas como uma sequência arbitrária de chamadas de função. Um atraso na resposta a uma queda, por exemplo, compromete diretamente o propósito do dispositivo.
- **Exigência de comunicação eficiente sob recursos restritos.** Um dispositivo vestível tem orçamento de energia e de banda limitado, e frequentemente depende de redes locais (Wi-Fi doméstico) que podem ser instáveis. A transmissão de alertas não pode depender de protocolos pesados ou de múltiplas conexões simultâneas com cada possível interessado.
- **Exigência de confidencialidade de dados sensíveis.** Dados fisiológicos e, principalmente, dados de localização de uma pessoa vulnerável são informações sensíveis. Sua interceptação ou adulteração durante a transmissão sem fio representa um risco que extrapola a esfera técnica, alcançando a privacidade e a segurança física do usuário.

Somado a isso, pode-se pensar em camadas ainda não plenamente exploradas neste protótipo, mas relevantes ao se pensar a evolução de sistemas IoMT. Por exemplo, a possibilidade de identificação de padrões mais sutis de saúde por meio de algoritmos de aprendizado de máquina [7], algo que reforça a tendência dos serviços de saúde em direção a um cuidado contínuo e integral do usuário [8], e não apenas relativo a eventos isolados.

III. Visão Geral do Protótipo Desenvolvido

O hardware é centrado no microcontrolador RP2040, presente na placa BitDogLab [9], ao qual estão conectados: uma Unidade de Processamento de Movimento (MPU) 6050 e um sensor óptico MAX30102 (para medição de frequência cardíaca e saturação de oxigênio), ambos via barramento I2C; um módulo GPS GY-NEO6MV2, conectado por uma porta serial UART; um display OLED, conectado também via I2C; um módulo Wi-Fi, ligado a um barramento SPI; um alarme visual e um sonoro (um LED e um buzzer), acionados por pinos digitais. Vale ressaltar, ainda, que o barramento onde os sensores estão conectados (I2C0) é diferente do barramento onde o display está conectado (I2C1). O protótipo montado e em funcionamento pode ser visualizado na Fig. 1, com a Fig. 2 evidenciando quais informações são mostradas no display OLED: frequência cardíaca, saturação de oxigênio e angulação da MPU6050 (com a última sendo utilizada para a depuração do sistema).

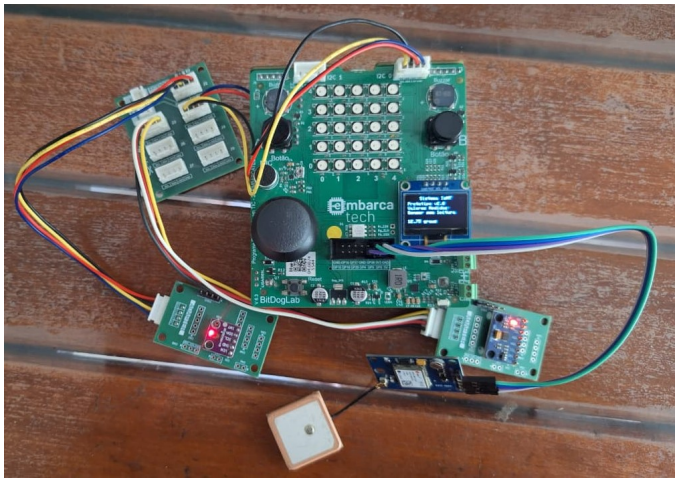


Figura 1. Protótipo do sistema desenvolvido.

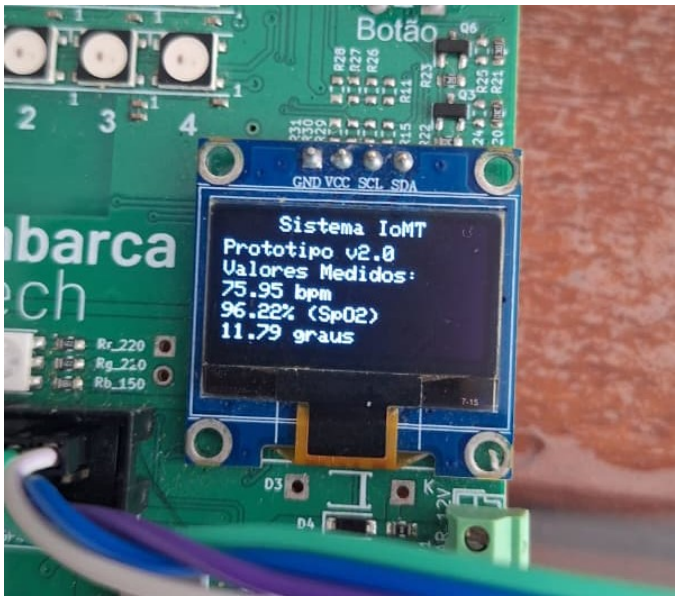


Figura 2. Conteúdo do display OLED com o sistema em funcionamento.

Esses componentes alimentam um conjunto de sete tarefas concorrentes: aquisição de dados do MPU6050, aquisição de dados do GPS, aquisição de dados do MAX30102, processamento de dados (incluindo o algoritmo de detecção de queda), interface com o usuário via display, conectividade MQTT e inicialização do Wi-Fi. A orquestração dessas tarefas é o primeiro ponto central da análise que se segue.

IV. Arquitetura do Sistema

A Fig. 3 apresenta o diagrama de blocos da arquitetura completa do sistema, organizado de baixo para cima em

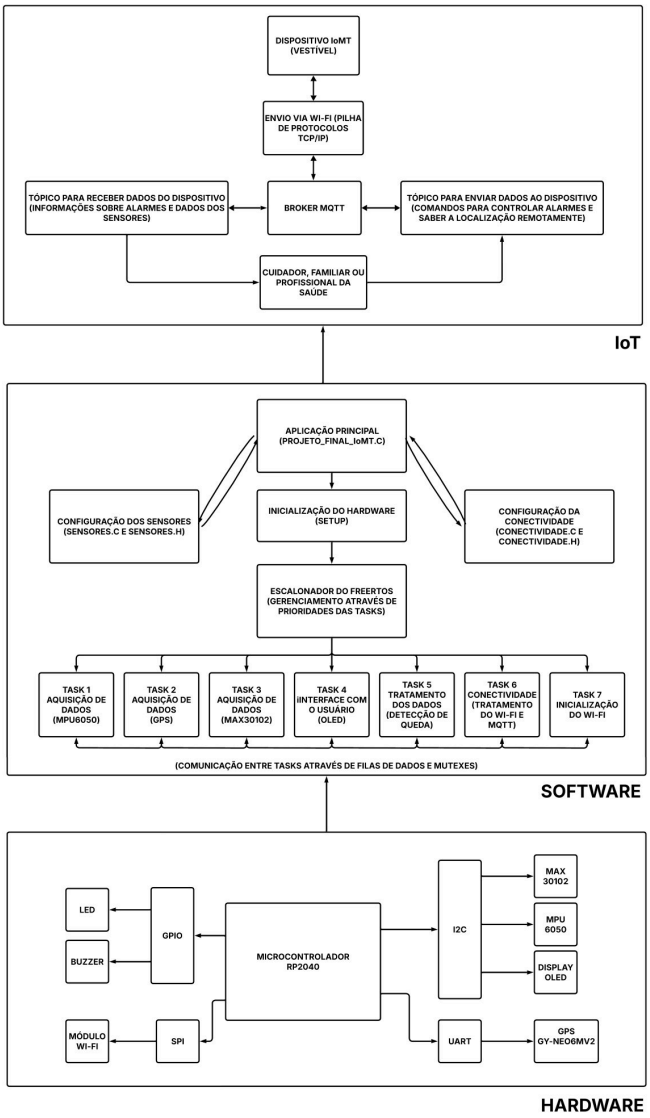


Figura 3. Diagrama da arquitetura completa do sistema.

Hardware, Software e camada IoT, refletindo a dependência hierárquica entre seus elementos: o hardware define os recursos disponíveis; o software os manipula de forma previsível, confiável e com base em prioridades; e a camada IoT expõe os dados ao mundo externo de forma eficiente e segura.

Nas subseções a seguir, são discutidas as três principais decisões de software do sistema: a utilização do FreeRTOS para lidar com a concorrência de eventos, do protocolo MQTT para a comunicação e do TLS como camada de criptografia para proteção dos dados.

A. Gerenciamento de Concorrência com FreeRTOS

O problema central que motiva o uso de um sistema operacional de tempo real é a coexistência, em um único microcontrolador de recursos limitados, de tarefas com requisitos temporais muito diferentes entre si. A aquisição de dados do MPU6050 precisa ocorrer em alta taxa de amostragem, sob risco de perder a janela temporal em que um evento de queda se manifesta nos dados do acelerômetro [10]; a aquisição do MAX30102 também exige prioridade alta para não comprometer a confiabilidade dos valores de frequência cardíaca e saturação de oxigênio; já a atualização do display OLED pode ocorrer em uma taxa muito mais baixa, sem prejuízo perceptível ao usuário. Em um modelo de execução sequencial (*bare-metal*), garantir esse tipo de diferenciação temporal exigiria um controle manual complexo e frágil, sujeito a perda de amostras sempre que uma operação mais lenta (como a leitura via UART do GPS) bloqueasse o fluxo principal.

Tecnicamente, o FreeRTOS organiza a aplicação em um conjunto de tarefas (*tasks*), cada uma implementada como uma função independente, com sua própria pilha de execução e seu próprio contexto de registradores. Diferentemente de um laço principal único, em que apenas uma função está ativa por vez, em um Sistema Operacional de Tempo Real (RTOS), todas as tarefas parecem executar simultaneamente: um componente chamado escalonador (*scheduler*) decide, a cada interrupção periódica do temporizador do sistema (*tick*), qual tarefa pronta para execução deve ocupar o processador naquele instante. No FreeRTOS, esse escalonamento é preemptivo e baseado em prioridades, ou seja, se uma tarefa de prioridade mais alta se torna pronta para executar, o escalonador interrompe imediatamente a tarefa de prioridade mais baixa em execução, salva seu contexto e transfere o processador para a tarefa mais prioritária, sem esperar que ela termine.

É exatamente esse mecanismo que o protótipo explora. A Tabela I lista as sete tarefas do sistema com suas respectivas prioridades e justificativas, tornando explícita a hierarquia temporal que o escalonador impõe: as tarefas de aquisição do MPU6050 e do MAX30102 nunca ficam impedidas de executar por tarefas menos urgentes, enquanto a atualização do display, inofensiva ao usuário em caso de atraso, opera na prioridade mais baixa do conjunto.

Tabela I
TAREFAS DO SISTEMA E SUAS PRIORIDADES NO FREERTOS

Tarefa	**Prior.**	**Justificativa**
`wifi_init_task`	Máx. (trans.)	lwIP; encerra após inicializar
`aquisicao_MPU_task`	Muito alta	Alta amostrag.; janela da queda
`aquisicao_MAX30102_task`	Muito alta	Confiab. do BPM/SpO_2
`processa_dados_task`	Alta	MEF de detecção; latência crítica
`conectividade_MQTT_task`	Alta	Latência do alerta ao *broker*
`aquisicao_GPS_task`	Média	UART; tolera perda de amostra
`interface_usuario_task`	Baixa	Display tolera atraso

A troca de informações entre tarefas não pode ocorrer por simples variáveis globais, sob risco de condições de corrida (*race conditions*). O FreeRTOS resolve esse problema com filas (*queues*): estruturas FIFO que já incorporam a proteção contra acesso concorrente. Já para o barramento I2C0 externo compartilhado pelos sensores MPU6050 e MAX30102, utiliza-se um mutex (*mutual exclusion*), que garante que apenas uma tarefa por vez tenha acesso ao barramento (os barramentos I2C foram configurados com uma frequência de 400 kHz).

Vale notar que o mutex do FreeRTOS implementa herança de prioridade: se uma tarefa de baixa prioridade estiver de posse do mutex e uma tarefa de alta prioridade precisar dele, a primeira temporariamente herda a prioridade da segunda, evitando o problema clássico de inversão de prioridade.

A justificativa específica para o contexto IoMT está no que esse escalonamento garante: a cadeia que liga a leitura do acelerômetro, a execução do algoritmo de detecção de queda e o disparo do alerta MQTT só é confiável se cada elo dessa cadeia tiver prioridade suficiente para não ser indefinidamente postergado por tarefas menos críticas. Nesse sentido, o uso de um sistema operacional de tempo real deixa de ser uma boa prática genérica de engenharia de software embarcado e passa a ser um requisito de segurança do usuário.

B. Algoritmo de Detecção de Queda por Máquina de Estados

O algoritmo de detecção de queda é o consumidor mais crítico da prioridade alta atribuída à `processa_dados_task`: é ele que transforma a sequência de leituras do acelerômetro em uma decisão binária "queda" ou "não queda", da qual dependem o acionamento do alarme local e o disparo do alerta via MQTT. Seu funcionamento baseia-se em álgebra vetorial sobre o vetor aceleração fornecido pelo MPU6050, cujo módulo é expresso em múltiplos de g, e no tempo necessário para que cada estado do processo de queda ocorra.

Segundo a literatura [10], o processo de queda se manifesta nos dados de um acelerômetro de três eixos em quatro fases sucessivas: (1) *ausência de peso*, caracterizada pela queda abrupta do módulo do vetor de aceleração abaixo de um limiar mínimo, correspondendo à fase de queda livre; (2) *impacto*, com o módulo ultrapassando um limiar máximo em um intervalo de tempo curto, indicando a colisão com o solo; (3) *inatividade*, em que o vetor de aceleração permanece com baixa variação por um intervalo prolongado, sugerindo que a pessoa ficou imóvel após a queda; e (4) *mudança de orientação*, verificada pelo ângulo entre o vetor de aceleração atual e o vetor registrado no início do evento, que permanece diferente da orientação de repouso habitual.

Essas quatro fases, precedidas pelo estado de operação normal, definem uma máquina de estados finitos (MEF) com cinco estados: NORMAL, AUSÊNCIA DE PESO, IMPACTO, INATIVIDADE e QUEDA CONFIRMADA. As regras de transição entre estados garantem que apenas a ocorrência *ordenada* dessas fases resulte em confirmação de queda: a transição de AUSÊNCIA DE PESO para IMPACTO, por exemplo, só ocorre se o impacto for detectado dentro de uma janela temporal pequena; do contrário, o sistema retorna ao estado NORMAL, interpretando o evento como um gesto brusco sem queda. A transição final de INATIVIDADE para QUEDA CONFIRMADA exige que o ângulo entre o vetor de aceleração atual e o vetor capturado no início do evento permaneça diferente por um intervalo de tempo mínimo, distinguindo inatividade pós-queda de uma simples pausa no movimento. O diagrama da MEF descrita é ilustrado pela Fig. 4, a qual descreve os limiares adotados para o protótipo desenvolvido.

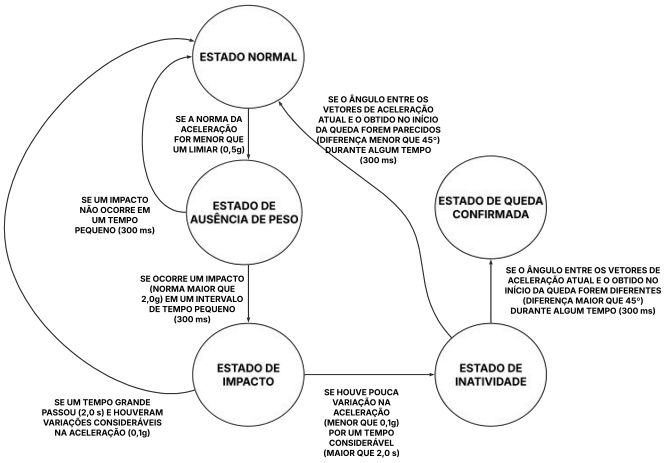


Figura 4. Diagrama de estados do algoritmo de detecção de quedas.

Esta subseção fecha o argumento sobre o FreeRTOS: sem a prioridade alta garantida ao escalonador para a `processa_dados_task`, a MEF poderia não receber amostras em tempo suficiente para detectar a janela de queda livre (um evento que, na literatura, dura tipicamente menos de 500 ms). A prioridade de tarefa, nesse contexto, é diretamente um requisito de desempenho do algoritmo.

C. Comunicação Leve e Assíncrona com MQTT

O MQTT é um protocolo de mensageria que opera sobre a pilha TCP/IP, seguindo um modelo de comunicação *publish/subscribe* mediado por um intermediário central, o *broker*. Nesse modelo, os participantes nunca se conectam diretamente entre si: um cliente que produz dados publica mensagens associadas a um tópico (cujo nome é uma string hierárquica que funciona como um endereço lógico) e qualquer cliente que tenha assinado aquele tópico recebe a mensagem automaticamente, redistribuída pelo *broker*. Essa indireção permite que o número de consumidores de uma informação cresça sem qualquer alteração no lado do dispositivo publicador, que continua enviando uma única mensagem por evento.

O protocolo também define três níveis de qualidade de serviço (QoS): 0, em que a mensagem é entregue sem confirmação; 1, em que a entrega é garantida, ainda que com possível duplicação; e 2, em que a entrega ocorre exatamente uma vez, ao custo de um *handshake* adicional. Soma-se a isso um mecanismo de *keep-alive*, em que o cliente envia periodicamente um sinal mínimo ao broker, permitindo que uma ausência desse sinal seja interpretada como uma desconexão inesperada e perda de informação potencialmente relevante em um cenário de cuidado contínuo. O cabeçalho de cada mensagem é deliberadamente compacto, o que, somado à ausência da sobrecarga textual de protocolos como o HTTP, torna o MQTT consideravelmente mais leve em termos de banda e de processamento exigido do dispositivo.

No protótipo, esse modelo é usado de forma direta: um tópico recebe os dados enviados pelo dispositivo (sinais vitais em condição normal, ou alerta de queda com localização em

condição crítica) e outro tópico recebe comandos remotos (solicitação de localização, ativação ou desativação de alarmes). A leveza do protocolo é explorada ainda na definição de uma ordem de prioridade – mensagens, alertas de queda e respostas de localização têm prioridade sobre o envio rotineiro de sinais vitais. As Figs. 5 e 6 evidenciam o funcionamento do sistema em dois cenários: no primeiro, o sistema está enviando os dados de forma rotineira (frequência cardíaca e saturação de oxigênio); no segundo, o usuário com acesso ao servidor MQTT envia o comando "localizacao" para o dispositivo de forma remota, retornando a latitude e longitude nas quais o dispositivo se encontra.

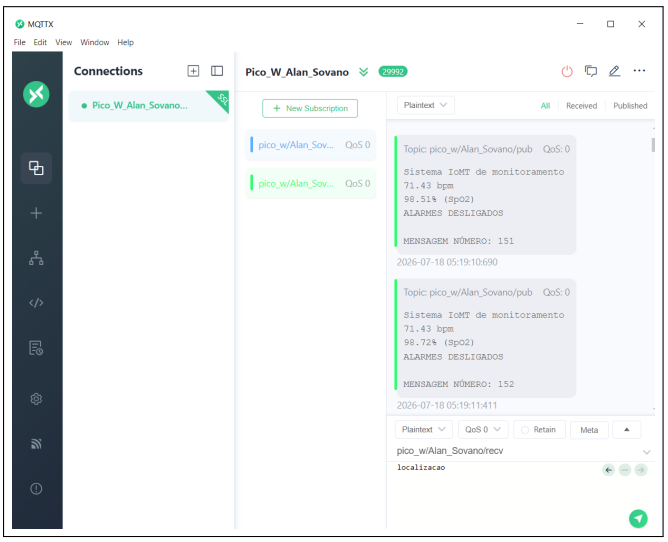


Figura 5. Envio de dados fisiológicos pelo dispositivo embarcado.

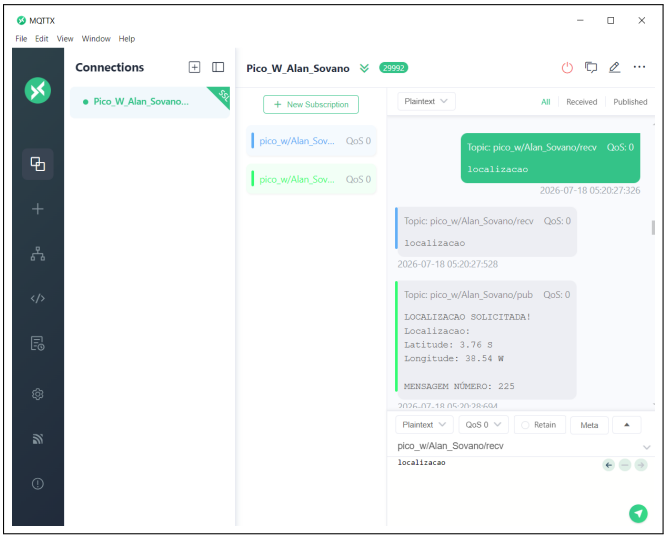


Figura 6. Solicitação da latitude e longitude do dispositivo embarcado.

A justificativa IoMT para essa escolha está na combinação de dois fatores: a necessidade de servir múltiplos interessados simultâneos a partir de um dispositivo com recursos mínimos, e a necessidade de que esse fluxo de comunicação continue funcional mesmo sob condições de rede não ideais, características típicas do ambiente doméstico de cuidado continuado.

D. Segurança de Dados Sensíveis com TLS

O terceiro problema decorre diretamente da resolução do segundo: uma vez que dados fisiológicos e de localização em tempo real passam a transitar por uma rede sem fio até um broker externo, esses dados tornam-se vulneráveis à interceptação. Esse tipo de ataque, conhecido como *Man-in-the-Middle* (MitM), permite que um terceiro não autorizado intercepte ou altere mensagens trocadas entre duas partes [11]. No contexto deste sistema, isso significaria a possibilidade de um terceiro acessar a localização de uma pessoa vulnerável, seus dados de saúde, ou mesmo emitir comandos falsos como desativar um alarme legítimo.

O protocolo TLS (*Transport Layer Security*) opera entre a camada de transporte (TCP) e a aplicação, e seu funcionamento divide-se em duas fases. Na primeira, chamada *handshake*, cliente e servidor negociam os parâmetros criptográficos da sessão: o servidor apresenta um certificado digital assinado por uma autoridade certificadora que permite ao cliente verificar sua identidade antes de prosseguir; em seguida, ambas as partes utilizam criptografia assimétrica para estabelecer, de forma segura e mesmo sobre um canal não confiável, uma chave secreta compartilhada. Na segunda fase, essa chave passa a ser usada com algoritmos de criptografia simétrica, significativamente mais eficientes computacionalmente, para cifrar todo o tráfego de dados da sessão, enquanto mecanismos criptográficos de identidade permitem detectar alterações no conteúdo durante o trânsito.

No protótipo, essa cadeia de garantias é colocada em prática por meio de um certificado de segurança embarcado na memória do dispositivo, que permite ao firmware verificar a validade do *broker* antes de estabelecer a conexão. Essa camada de segurança se soma à autenticação já existente na rede Wi-Fi (login e senha), fornecendo mais segurança na comunicação entre o dispositivo e o destinatário final dos dados.

A Fig. 7 evidencia o efeito da criptografia das informações enviadas pelo sistema. Para realizar essa verificação, o sistema foi conectado, via Wi-Fi, a um computador configurado como ponto de acesso à internet. O tráfego da informação gerado pelo dispositivo foi, então, monitorado pelo Wireshark, aplicativo de código aberto utilizado para a análise de protocolos de rede [12]. Para o teste, o sistema foi configurado sem TLS e com TLS em momentos diferentes, de forma a verificar o envio da mensagem com e sem criptografia.

Vale reconhecer que essa decisão não é isenta de custo: o estabelecimento de uma sessão TLS impõe sobrecarga computacional e energética ao microcontrolador. A escolha de absorver esse custo, em vez de prescindir da criptografia, reflete o julgamento de que, para um dispositivo que transmite dados de saúde e de localização de uma população vulnerável, a confidencialidade não é um refinamento opcional, mas um requisito mínimo de responsabilidade no projeto (mesmo em um protótipo sem caráter diagnóstico).

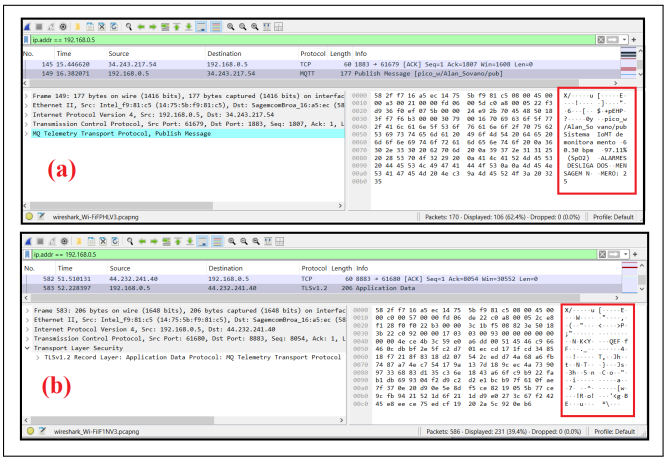


Figura 7. tráfego da informação (a) sem TLS e (b) com TLS.

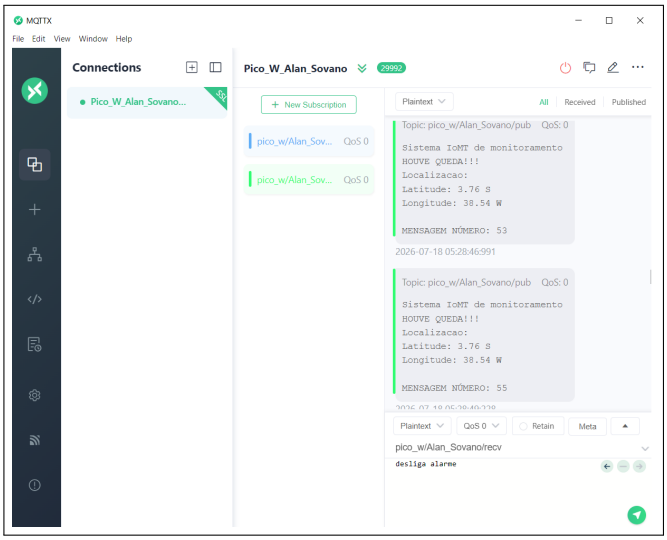


Figura 8. Detecção de queda pelo dispositivo embarcado.

V. Resultados da avaliação experimental do sistema

A validação prática do protótipo mostrou as três camadas operando como um único fluxo coeso. Foi possível observar a conexão Wi-Fi, a resolução de DNS do *broker* utilizado, o estabelecimento da sessão MQTT sobre TLS e a confirmação de inscrição nos tópicos relevantes, sequência que reproduz com exatidão a ordem teórica descrita na Seção IV: primeiro o FreeRTOS inicializa a pilha de rede via `wifi_init_task`; depois o MQTT estabelece a sessão sobre o canal já aberto; por fim, o TLS cifra esse canal antes que qualquer dado dos sensores seja transmitido. A mensagem recebida através do *broker* MQTT, quando ocorre uma queda que é detectada, pode ser visualizada na Fig. 8. Na ocorrência de um evento confirmado de queda, também foi possível verificar o disparo dos alarmes (auditivo e visual). Vale ressaltar que todos os testes de queda foram feitos somente com o dispositivo (sem a presença de nenhum tipo de usuário), visto que o projeto ainda necessita da aprovação de um comitê de ética em pesquisa para essa etapa.

De forma a avaliar se a quantidade de memória de pilha (*stack*) foi alocada adequadamente para cada uma das sete tarefas criadas, uma tarefa adicional (`monitor_task`) foi implementada exclusivamente para a avaliação do desempenho do sistema. Nessa tarefa, foi utilizada a função `uxTaskGetStackHighWaterMark()` do kernel FreeRTOS, que retorna a menor quantidade de memória de pilha que permaneceu livre desde a criação de uma tarefa, permitindo estimar seu consumo máximo de memória de pilha. A Tabela II apresenta a memória de pilha alocada para cada tarefa, bem como o uso máximo estimado durante a execução do sistema ao longo de 10 minutos.

A tarefa de conectividade apresentou o maior consumo de memória de pilha, resultado esperado devido ao uso das bibliotecas responsáveis pela comunicação TCP/IP, pelo gerenciamento do módulo Wi-Fi e pela implementação do protocolo MQTT. Entretanto, todas as tarefas mantiveram uma margem de memória disponível durante os testes, indicando que as

Tabela II
Consumo de memória de pilha das tarefas durante 10 minutos de execução

Tarefa do programa	Pilha de memória alocada (bytes)	Uso máximo estimado (bytes)
`aquisicao_MPU_task`	4096	296
`aquisicao_GPS_task`	4096	504
`aquisicao_MAX30102_task`	8192	616
`interface_usuario_task`	8192	600
`processa_dados_task`	4096	512
`conectividade_MQTT_task`	8192	2920
`wifi_init_task`	8192	656

pilhas foram dimensionadas adequadamente e que o sistema ainda possui capacidade para acomodar futuras expansões ou otimizações. Logo, todas as tarefas tiveram uma margem de segurança suficiente para as respectivas operações, com o FreeRTOS não comprometendo o uso de memória do RP2040.

Além do uso de memória, métricas temporais referentes ao uso do FreeRTOS foram calculadas. Em especial, pode-se destacar o tempo médio entre o início de duas execuções consecutivas de uma mesma tarefa de aquisição de dados, de forma a avaliar a regularidade das mesmas durante a operação normal do sistema (conectado ao *broker* MQTT e enviando mensagens). Além disso, avaliou-se o tempo de atualização do display OLED, de forma a caracterizar também como essas informações seriam repassadas ao usuário. Todas as métricas foram obtidas com o uso da função `get_absolute_time()`, a qual foi colocada em trechos específicos do código principal para captura de marcações de tempo. Após um determinado número de capturas, os dados eram enviados ao usuário por meio da interface serial. A captura das marcações foi feita de modo a minimizar, da melhor forma possível, a interferência no próprio intervalo de tempo que estava sendo medido. Um resumo dos dados é descrito na Tabela III.

A partir dos dados, percebe-se que as tarefas `aquisicao_MPU_task` e `aquisicao_MAX30102_task` possuem boa regularidade,

Tabela III
CARACTERIZAÇÃO TEMPORAL DAS TAREFAS DE AQUISIÇÃO E INTERFACE.

Métrica	MPU	MAX30102	GPS	Interface
Número de testes	500	500	500	500
Média (ms)	51,22	102,77	1235,03	38,06
Desvio-padrão (ms)	0,84	0,44	490,44	1,71
Mínimo (ms)	49,86	102,00	972,06	35,99
1º quartil (ms)	50,96	102,94	976,06	37,00
Mediana (ms)	51,00	103,00	979,06	37,00
3º quartil (ms)	52,00	103,00	982,06	40,00
Máximo (ms)	54,00	104,39	3979,06	43,01

com ambas sendo executadas periodicamente perto do tempo desejado (50 ms e 100 ms, respectivamente). Variações no período de execução dessas tarefas (*jitter*) podem ser explicadas pelo tempo necessário para a troca de tarefas do escalonador e também pelo uso do mutex (uma tarefa só pode ser executada quando outra não estiver sendo executada).

Para a tarefa `aquisicao_GPS_task`, avaliou-se o tempo entre a geração de duas respostas completas do GPS (que é descrita por um conjunto de caracteres vindos da porta UART). A tarefa foi projetada para rodar a cada 1 ms e obter um caracter, o que também foi observado, mas a geração de uma resposta completa (que será enviada via MQTT quando solicitado) é um processo muito mais demorado. Por isso, o sistema foi configurado para manter um mesmo sinal válido de GPS por 5 segundos (na ausência de um outro sinal válido). Como, pelos experimentos realizados, o tempo entre sinais válidos de GPS foi menor do que o estabelecido, percebe-se que há uma consistência no limiar utilizado.

A tarefa `interface_usuario_task`, por sua vez, depende do recebimento de dados de `aquisicao_MPU_task` e `aquisicao_MAX30102_task`, com a comunicação ocorrendo através de filas. Foi possível verificar que, dentro da estrutura do sistema, a taxa de atualização do display OLED ficou em torno de 40 ms, tendo inclusive uma dispersão relativamente baixa. Logo, concluiu-se que essa tarefa também estaria funcionando da forma adequada.

Outro teste realizado consistiu na avaliação da confiabilidade da comunicação MQTT durante 10 minutos de operação contínua. O teste foi realizado utilizando a porta 8883 (com criptografia TLS) do *broker* público EMQX e nível de QoS igual a 0. Cada mensagem publicada continha o valor de uma variável contadora, incrementada a cada transmissão, permitindo identificar mensagens perdidas e mensagens recebidas fora da sequência esperada.

Ao longo do teste, o dispositivo publicou 584 mensagens, das quais 534 foram recebidas na sequência esperada, 47 foram perdidas e 3 foram recebidas com atraso. Dessa forma, obteve-se uma taxa de entrega em sequência de aproximadamente 91,4%. As perdas observadas são compatíveis com a utilização do QoS 0, que não oferece garantia de entrega das mensagens. Logo, observa-se que o sistema consegue realizar o envio de mensagens de forma consistente, com perdas toleráveis. Vale lembrar que, como uma mensagem é enviada a cada segundo pelo dispositivo, há uma redundância no envio da mensagem de alerta (o alerta só para de ser emitido quando o alarme é desligado remotamente ou o sistema é reiniciado).

Durante o experimento, para simular um uso rotineiro do dispositivo, cinco comandos também foram enviados em instantes distintos para acionar e desacionar os alarmes e solicitar sua localização. Não foi observada influência desses comandos na ocorrência de perdas ou atrasos das mensagens publicadas.

Em uma etapa posterior, testes envolvendo a comunicação via MQTT e a detecção de quedas foram realizados, com o intuito de caracterizá-los temporalmente. Os comandos "liga alarme", "desliga alarme" e "localizacao" foram, cada um, enviados 20 vezes (totalizando 60 comandos), com o tempo entre o envio dos comandos e o retorno do dispositivo através do *broker* sendo medido. Dessa forma, procurou-se quantificar o tempo de resposta em um cenário comum de uso. Essa métrica representa o caminho mais crítico do fluxo de informação, visto que o comando passa pelo "*broker*" e é processado pelo dispositivo (com as rotinas incluindo o FreeRTOS e a pilha de protocolos de comunicação) para que, ao final, o sistema retorne uma mensagem confirmando o sucesso do comando (também pelo "*broker*"). Além disso, foram simuladas 50 quedas para verificar qual o tempo de processamento utilizado pela MEF entre a detecção de uma possível queda e a sua confirmação. Por fim, o tempo entre a confirmação de uma queda e o envio da mensagem correspondente para o *broker* foi medido 250 vezes (a partir de uma pequena alteração no código, a qual forçava a MEF a acusar sempre uma queda). A Tabela IV apresenta a média, o desvio-padrão, o primeiro, segundo e terceiros quartis e os valores máximos e mínimos obtidos para tais testes.

Tabela IV
MÉTRICAS TEMPORAIS DO SISTEMA (ENVIO DE COMANDOS E DETECÇÃO DE QUEDA)

Métrica	Tempo de comando–confirmação	Tempo de processamento da MEF	Tempo entre detecção de queda e envio ao broker
Número de testes	60	50	250
Média (s)	1,58	2,48	1,02
Desvio-padrão (s)	1,51	0,06	0,02
Valor mínimo (s)	0,46	2,41	1,00
1º quartil (s)	0,92	2,42	1,01
Mediana (s)	1,13	2,47	1,03
3º quartil (s)	1,56	2,52	1,03
Valor máximo (s)	8,77	2,62	1,05

Analisando a Tabela IV, percebe-se que os tempos de resposta entre um comando e sua confirmação ficaram, em sua maioria, em torno de 1,5 segundos ou menos, um valor compatível com o uso contínuo do dispositivo. Analisando o valor máximo e o desvio-padrão, é possível verificar que houve alguns valores anormais nas medições, cuja causa provavelmente está relacionada aos elos da comunicação que não são controlados (qualidade da comunicação Wi-Fi e disponibilidade do servidor, por exemplo). Dessa forma, considera-se que as oscilações no tempo de resposta do dispositivo estão dentro do esperado para um sistema IoMT. Todos os comandos geraram as respostas desejadas.

Com relação à MEF, o tempo para a detecção de uma queda fica em torno de 2,5 segundos, o que está de acordo com o descrito pela Fig. 4. Esse valor, somado à estimativa de 1 segundo para que a mensagem informando a queda chegue ao *broker*, indica que o sistema leva em torno de 3,5 segundos para detectar a queda, confirmá-la e publicar uma mensagem com o aviso do evento. Esse desempenho foi considerado adequado para o propósito do dispositivo.

É relevante notar que nem toda limitação observada decorre da arquitetura de software discutida aqui. O princípio de medição por fotopletismografia, utilizado pelo sensor de frequência cardíaca e saturação de oxigênio, é intrinsecamente suscetível a artefatos de movimento e a interferência luminosa [13], razão pela qual optou-se por não associar alarmes automáticos a esses parâmetros no estágio atual do protótipo. Essa limitação é de natureza física do sensor, não da arquitetura de concorrência, comunicação ou segurança, o que reforça que as três camadas analisadas neste artigo cumpriram seu papel mesmo onde outras partes do sistema ainda demandam refinamento.

VI. Considerações Finais

Foi possível observar que as decisões de arquitetura do sistema resultaram em um protótipo de dispositivo IoMT funcional, com um tempo de resposta aos comandos, taxa de entrega da informação e uso de memória razoáveis do ponto de vista quantitativo. Além disso, do ponto de vista qualitativo, o dispositivo teve o comportamento esperado ao longo dos testes funcionais, reforçando que as escolhas de arquitetura foram adequadas no estágio atual de desenvolvimento. É possível concluir, ainda, que decisões arquiteturais de software podem ser tão importantes quanto a escolha dos sensores em sistemas IoMT críticos.

A análise apresentada neste artigo buscou argumentar sobre como o FreeRTOS, MQTT e TLS, no contexto do protótipo de plataforma IoMT desenvolvido, não foram escolhidos como peças intercambiáveis de uma caixa de ferramentas genérica de sistemas embarcados conectados. Cada uma dessas três tecnologias resolve um problema que se torna especialmente crítico precisamente porque o sistema lida com a saúde e a segurança física de uma pessoa: o FreeRTOS garante que a cadeia de detecção e resposta a eventos críticos materializada pela MEF de cinco estados descrita na Seção IV-B não seja comprometida por concorrência mal gerenciada; o MQTT permite que esses dados alcancem múltiplos interessados de forma leve e resiliente, compatível com um dispositivo vestível de recursos limitados; e o protocolo TLS garante que esses mesmos dados fisiológicos e de localização permaneçam confidenciais ao longo de todo o trajeto até o destinatário final. Compreender essas três decisões não como escolhas técnicas isoladas, mas como respostas coordenadas às exigências específicas de um sistema IoMT, é o que permite avaliar com mais rigor tanto os méritos quanto os limites do protótipo desenvolvido, e orientar com mais clareza os próximos passos de sua evolução.

Pretende-se, em etapas futuras, realizar mais testes de conexão do dispositivo e implementar mais camadas de segurança. Além disso, tem-se o objetivo de testá-lo com uma diversidade de usuários para avaliar sua eficácia real, avaliando a taxa de acertos do sistema (mediante a aprovação do projeto em um comitê de ética). Conforme o protótipo fique mais elaborado, pretende-se também fazer uma avaliação do consumo energético do dispositivo.

Outra melhoria futura planejada é a adição de mais opções de alarmes e uma maior diferenciação de tipos de queda. Ao se comparar uma queda de frente com uma queda de costas, por exemplo, as lesões podem ser diferentes e o socorro, consequentemente, deve ser mais ou menos imediato. Para a diferenciação de padrões de queda, pretende-se utilizar algum algoritmo de *machine learning* que possa ser embarcado (TinyML) de forma a classificar se o usuário caiu de lado, de frente, de costas, sentado, dentre outros padrões. Por fim, pretende-se colocar no sistema a capacidade de diferenciar situações como bradicardia, taquicardia e hipóxia, a partir dos dados do sensor MAX30102.

Referências

[1] A. I. Awad, M. M. Fouda, M. M. Khashaba, E. R. Mohamed, and K. M. Hosny, "Utilization of mobile edge computing on the internet of medical things: A survey," *ICT Express*, vol. 9, no. 3, pp. 473–485, 2023.

[2] F. W. Dourado Júnior, A. C. A. Moreira, D. L. Salles, and M. A. M. d. Silva, "Intervenções para prevenção de quedas em idosos na atenção primária: revisão sistemática," *Acta Paulista de Enfermagem*, vol. 35, p. eAPE02256, 2022.

[3] D. Chun, D. Chun, Y. Cha, L. Jellibi, and S. J. Park, "Watch-out: Real-time tracking and assistance for wandering alzheimer's patients using wearable technology," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–7.

[4] A. Dunkels, "Design and implementation of the lwip tcp/ip stack," *Swedish Institute of Computer Science*, vol. 2, no. 77, 2001.

[5] P. M. Rapa, M. Orlandi, M. Zanghieri, L. Benini, and S. Benatti, "Embedded multi-sensor smartwatch for computationally intensive biosignal processing," in *2024 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2024, pp. 1–5.

[6] M. Eddabbah, C. Aitzaouiat, and N. Hachoumi, "Design of a smart and adaptive wearable device for early detection of cardiovascular abnormalities with secure and sustainable iot integration," in *2025 International Conference on Innovation in Artificial Intelligence and Internet of Things (AIIT)*. IEEE, 2025, pp. 1–9.

[7] D. Kalita, H. Sharma, and K. B. Mirza, "Deep learning on edge and cloud for intelligent monitoring and decision making in artificial pancreas," *Expert Systems with Applications*, vol. 295, p. 128833, 2026.

[8] M. Shafiq, J.-G. Choi, O. Cheikhrouhou, and H. Hamam, "Advances in iomt for healthcare systems," *Sensors*, vol. 24, no. 1, p. 10, 2023.

[9] F. Fruett, F. Pereira Barbosa, S. Cardoso Zampolli Fraga, and P. Ivo Aragão Guimarães, "Empowering steam activities with artificial intelligence and open hardware: The bitdoglab," *IEEE Transactions on Education*, vol. 67, no. 3, pp. 462–471, 2024.

[10] N. Jia, "Detecting human falls with a 3-axis digital accelerometer," in *A forum for the exchange of circuits, systems, and software for real-world signal processing*, vol. 43, no. 3, 2009, p. 3.

[11] H. Fereidouni, O. Fadeitcheva, and M. Zalai, "Iot and man-in-the-middle attacks," *Security and Privacy*, vol. 8, no. 2, p. e70016, 2025.

[12] R. Sharpe, E. Warnicke, and U. Lamping, "Wireshark user's guide," *Version 4.1. 0*, 2014.

[13] V. Button, *Principles of measurement and transduction of biomedical variables*. Academic Press, 2015.

Desenvolvimento e Validação de Sistema Inteligente Pneumático-Eletrônico para Avaliação Cognitivo-Motor em Esportes de Combate

Vanderli Tomaz Junior
Prog. de Pós-Graduação em Educação Física (PPGEF)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
vanderlitomaz@alunos.utfpr.edu.br

Lucas Daniel Ortiz Arias
Departamento Acadêmico de Computação (DACOM)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
lucasdanielarias@alunos.utfpr.edu.br

Danilo Gondo Chinen
Departamento Acadêmico de Computação (DACOM)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
chinen@alunos.utfpr.edu.br

Fabio Kurt Schneider
Prog. de Pós-Graduação em Engenharia Biomédica (PPGEB)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
fabioks@utfpr.edu.br

Marcelo Romanovitch Ribas
Prog. de Pós-Graduação em Educação Física (PPGEF)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
mromanovitch@yahoo.com.br

Julio Cesar Bassan
Prog. de Pós-Graduação em Educação Física (PPGEF)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil
jcbassan@utfpr.edu.br

***Abstract*— This study reports the development and preliminary bench validation of a low-cost smart system for cognitive-motor assessment in combat sports. The proposed platform integrates a pneumatic-electronic sensing architecture based on an ESP32 microcontroller and a BMP280 barometric pressure sensor, designed to overcome the mechanical noise limitations associated with conventional inertial measurement units (IMUs). By incorporating an air chamber within a polycarbonate enclosure, the system mechanically attenuates impact-induced vibrations before signal acquisition, improving measurement stability and protecting the sensing elements from direct mechanical stress. Bench experiments demonstrated stable operation, high repeatability, and reliable functional performance. Characterization of Bluetooth Low Energy (BLE) communication latency, performed using high-speed video analysis (240 frames per second), revealed transmission delays ranging from 92 to 150 ms, with a mean latency of 121 ms. Repeatability testing showed a baseline noise level below 0.05 hPa and a coefficient of variation (CV) of 1.2% for the pressure-trigger threshold. Temporal accuracy was preserved through a local on-sensor timestamping algorithm, ensuring that wireless communication latency does not influence response-time measurements. These findings demonstrate the technical feasibility of the proposed system as an accessible instrumentation platform for combat sports, establishing a robust technological foundation for future ecological validation and real-world training applications.**

***Keywords*— combat sports, biomechanics, biomedical instrumentation, cognitive-motor assessment, Bluetooth Low Energy, embedded systems.**

I. INTRODUÇÃO

A análise de desempenho em esportes de combate demanda parâmetros quantitativos, em que variáveis como a precisão espacial, a velocidade do gesto motor e o tempo de reação (TR) despontam como determinantes para o sucesso competitivo **[1]**, **[2]**, **[3]**. Contudo, a mensuração fidedigna dessas capacidades em ambiente ecológico apresenta restrições metodológicas **[1]**. Dispositivos baseados em estímulos luminosos carecem de especificidade cinemática para impactos, enquanto sensores inerciais (*Inertial Measurement Unit* - IMU) introduzem artefatos de movimento que comprometem os registros **[1]**. Embora sistemas laboratoriais padrão-ouro (plataformas de força e cinemetria óptica) apresentem elevada acurácia, o alto custo, a baixa portabilidade e a necessidade de infraestrutura especializada limitam sua aplicação na rotina de treino.

Recentemente, sistemas baseados em visão computacional sem marcadores (*markerless*), utilizando modelos como *MediaPipe*, *Blaze Pose* ou *MoveNet*, além de outros dispositivos vestíveis (*wearables*), têm sido propostos para análise de movimento no combate **[4]**, **[5]**. No entanto, essas tecnologias sofrem com limitações de taxa de amostragem, oclusão de segmentos e imprecisão na determinação temporal do instante exato do impacto, especialmente em alta velocidade. Ademais, tais sistemas são direcionados predominantemente à reconstrução cinemática, não sendo concebidos para detectar diretamente o momento da colisão entre o atleta e o alvo.

Fisiologicamente, o tempo decorrido entre o estímulo visual e a detecção física do golpe configura o tempo de resposta, resultante da soma do tempo de reação (processamento cognitivo central eferente) e do tempo de movimento (fase motora periférica). Estudos demonstram variações temporais significativas entre técnicas como *jabs* e diretos (*crosses*), reforçando a importância de monitorar o gesto específico **[6]**. A aplicação de força depende da transferência de momento linear dos membros inferiores e rotação do tronco **[7]**, gerando picos dinâmicos que requerem amortecimento estrutural para proteger o hardware **[8]**. Consequentemente, sistemas destinados a essa avaliação

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

devem apresentar elevada resolução temporal, robustez mecânica e capacidade de operar em condições próximas à rotina esportiva.

Para superar tais limitações, este trabalho propõe o desenvolvimento e a validação em bancada de um sistema baseado em interface pneumática-eletrônica. O dispositivo utiliza uma câmara de ar acoplada a um gabinete de policarbonato para atenuar mecanicamente as vibrações do impacto antes da aquisição pelo sensor de pressão barométrica (BMP280) **[1]**. A interface funciona como um filtro mecânico passivo, reduzindo vibrações espúrias e favorecendo a estabilidade dos sinais. A aquisição local é realizada por um microcontrolador ESP32 (*timestamping on-sensor*), permitindo que a marcação temporal ocorra diretamente no dispositivo, minimizando a influência da latência da comunicação sem fio **[9]**. O objetivo deste estudo é analisar o desempenho técnico do protótipo e calibrar suas latências instrumentais em laboratório **[10]**, estabelecendo a base tecnológica para futuras validações ecológicas com atletas.

II. METODOLOGIA E INSTRUMENTAÇÃO

A. Delineamento Experimental e Planejamento Amostral

Esta pesquisa aplicada, de abordagem quantitativa e delineamento metodológico, validou na Fase I (objeto deste artigo) a consistência metrológica do hardware em ensaios de bancada. Os testes caracterizaram a estabilidade do sensor, repetibilidade, latência e o processamento embarcado. A Fase II compreenderá a validação ecológica do dispositivo por meio de acompanhamento longitudinal com atletas.

Este protótipo, foi desenvolvido para atender as demandas futuras para a realização da Fase II, onde a amostra estimada será selecionada por conveniência e composta por 30 atletas de artes marciais do sexo masculino, distribuídos igualmente por nível de expertise: avançados (profissionais com mais de 5 anos de prática ativa; n = 15) e intermediários (amadores; n = 15). O dimensionamento amostral no software G*Power 3.1.9.7 (tamanho de efeito d = 1,1, $\alpha = 0,05$ e poder estatístico de 0,80) resultou em um poder real de 0,828. Os critérios de inclusão exigirão: idade entre 18 e 55 anos; ausência de lesões osteomioarticulares agudas ou cirurgias que comprometam a performance balística; e experiência consolidada em modalidades de impacto (*striking*). Todos os voluntários assinarão o Termo de Consentimento Livre e Esclarecido (TCLE) após aprovação no Comitê de Ética em Pesquisa (CEP) institucional.

B. Arquitetura do Sistema e Instrumentação Eletrônica

A arquitetura pneumática-eletrônica foi concebida para operar sob impacto repetitivo, preservando a integridade estrutural e a qualidade do sinal. A interface física externa utilizou um manguito de esfigmomanômetro adaptado como cápsula pneumática para capturar a variação transiente de pressão do golpe. Esta bolsa de ar foi acoplada a um gabinete de policarbonato de alta resistência, selecionado por suas propriedades dielétricas e de isolamento elétrico. O conjunto funciona como um filtro mecânico passivo, reduzindo a transmissão de vibrações de alta frequência ao sensor barométrico, aumentando a estabilidade do sinal e minimizando interferências diretas.

A transdução do sinal foi realizada por um sensor barométrico digital BMP280, conectado via I2C a um microcontrolador ESP32, modelo S3 WROOM 1. A escolha do BMP280 (Bosch Sensortec, Alemanha) fundamentou-se nas limitações mecânicas de sensores alternativos em ensaios de impacto repetitivo: sensores de força resistivos (FSR) apresentam elevada histerese e rápida degradação; películas piezoelétricas demandam circuitos complexos para mitigar a deriva estática e rompem-se sob cisalhamento; e células de carga exigem acoplamento rígido a placas pesadas, inviabilizando a portabilidade e a proteção mecânica do usuário.

A interface pneumática baseada na compressão volumétrica atua como filtro passa-baixa. Em regimes de compressão adiabática rápida, a variação de pressão interna (ΔP) é diretamente proporcional à energia de impacto aplicada, sem expor os circuitos a estresse direto. O ESP32 controla os estímulos visuais via matriz de LEDs de alto brilho, executa o processamento local dos dados (*timestamping on-sensor*) e gerencia a comunicação via protocolo Bluetooth Low Energy (BLE) com o aplicativo móvel.

O sistema pode ser incrementado com múltiplos dispositivos microcontrolados, a fim de obter medições de pressão aplicada e tempo de reação em diversas posições. Cada dispositivo adicionado à rede comunica-se com o dispositivo central através do protocolo *wireless* ESP-NOW.

O dispositivo central da solução é composto por dois microcontroladores. Um deles concentra a função de sensor e de servidor BLE, ao passo que o outro atua como comunicador ESP-NOW. Ambos os microcontroladores comunicam entre si via conexão serial UART. Essa separação de tarefas foi necessária pois parte do hardware da ESP-32 é compartilhado para envio e recebimento de ambos os protocolos de comunicação wireless implementados (BLE e ESP-NOW). Dessa forma, quando o dispositivo tem seu módulo BLE atuando, sua capacidade de receber pacotes ESP-NOW é afetada, e vice-versa. Isso resulta em uma considerável perda de pacotes. Por outro lado, ao delegar as comunicações de diferentes protocolos para diferentes dispositivos, não ocorre essa sobrecarga do hardware. Além disso, o dispositivo mestre também atua como um dos sensores da solução, acionando seu LED e aguardando a resposta do usuário sempre que for recebida a devida instrução do aplicativo celular.

C. Rotina de Detecção

O aplicativo móvel atua como interface, enviando comandos ao ESP32 via BLE. O sistema permite a ativação manual dos ensaios ou o acionamento automático, iniciando um novo ciclo pós-impacto sem intervenção do operador. O aplicativo permite envio de requisição de ativação ao dispositivo central, bem como aos dispositivos periféricos. Em caso de ativação do dispositivo mestre, a mensagem BLE parte do celular e é recebida pelo módulo BLE do dispositivo central. Em caso de ativação de um dos dispositivos slaves, a mensagem BLE é enviada pelo celular, recebida pelo módulo

BLE do dispositivo central, replicada via UART ao módulo ESP-NOW e, finalmente, enviada via ESP-NOW do dispositivo central ao dispositivo periférico. Ao receber o comando, o sistema microcontrolado acende um LED e salva o tempo decorrido desde o *boot* via função millis(). O dispositivo lê e armazena o valor de pressão do BMP280 e inicia a rotina de *polling* para monitorar a variação gerada pelo golpe na bolsa de ar.

O loop de amostragem contínua interrompe a rotina assim que a variação de pressão ultrapassa o limiar biomecânico (*threshold*) preestabelecido. Quando o valor medido para de aumentar (após a primeira derivada negativa), o sensor executa novamente a função millis() e computa o tempo de reação pela diferença temporal. Esse método garante o registro do pico máximo de pressão para determinar a intensidade relativa do impacto. Na sequência, o microcontrolador envia o pacote BLE ao aplicativo, que exibe os dados em gráficos em tempo real.

D. Latências e Otimizações de Bancada

Todas as chamadas de depuração serial bloqueantes (exemplo: a rotina Serial.println()) foram removidas do loop principal. O sensor BMP280 foi configurado com resolução de pressão de 0,16 Pa e precisão absoluta de ± 1 hPa. No modo de alta performance (*oversampling* de pressão de 16x e temperatura de 2x), opera com frequência de aquisição efetiva de 157 Hz, garantindo resolução temporal de 6,36 ms. O algoritmo utiliza um limiar de disparo de 15 hPa acima da linha de base atmosférica e encerra a amostragem após detectar a primeira derivada negativa de pressão ($dP/dt < 0$), caracterizando o pico de força. O valor do limiar foi definido a partir de testes com aplicação de golpes de baixa intensidade na câmara com ar conectada ao sensor, de forma que mesmo ações com golpes suaves pudessem ser percebidas.

A latência de transmissão BLE do comando de ativação foi caracterizada filmando o pressionamento do botão no aplicativo e o acendimento do LED a 240 quadros por segundo, com os resultados de conversão matemática registrados na Tabela 1:

Tabela 1: Caracterização experimental da latência de pareamento BLE

Número de quadros	Tempo (s)
25	0,104
36	0,150
22	0,092
33	0,138
29	0,121

(Nota: Média geral de latência de transmissão BLE = 121 ms).

O atraso de transmissão varia de 92 ms a 150 ms, porém não interfere na acurácia da medição psicomotora. O método de *timestamping on-sensor* captura o marcador temporal inicial (millis()) localmente no exato milissegundo de energização do LED. Desse modo, a latência de rede apresentada na Tabela 1, que poderia ser significativa em condições de combate, é completamente isolada do cálculo do tempo de reação, eliminando vieses. Os ensaios de estabilidade revelaram erro instrumental inferior a 0,05 hPa na linha de base e CV de 1,2% para disparos repetitivos controlados.

E. Eficiência da comunicação BLE

A Figura 1 exibe os resultados de RSSI (Received Signal Strength Indicator) da conexão BLE entre um dispositivo celular e um microcontrolador ESP-32 em diferentes distâncias, variando de 1 a 5 metros. Para cada distância, foram realizadas 3 medições e foi calculada a média simples entre os valores coletados. Os testes foram realizados em um local aberto, com poucos obstáculos. Para distâncias de até 5 metros, o RSSI da conexão nunca é mais baixo que -80 dBm.



Fig. 1. Gráfico do RSSI da conexão BLE entre celular e ESP-32 em diferentes distâncias

A tabela 2 mostra os resultados obtidos após o envio e recebimento de 10 mensagens ESP-NOW entre uma ESP-32 slave e o dispositivo mestre. A distância entre os dispositivos era de cerca de 5 metros. Todos os pacotes foram enviados e recebidos corretamente.

Tabela 2: resultado de envio de pacotes ESP-NOW com dispositivos a 5 metros de distância

Pacote	Resultado
Mestre → slave 1	OK
Slave → mestre 1	OK
Mestre → slave 2	OK
Slave → mestre 2	OK
Mestre → slave 3	OK
Slave → mestre 3	OK
Mestre → slave 4	OK
Slave → mestre 4	OK
Mestre → slave 5	OK
Slave → mestre 5	OK

A tabela 3 mostra os resultados de um outro experimento seguindo a metodologia das medições anteriores. Contudo, dessa vez, o espaço entre o dispositivo mestre e o dispositivo slave era preenchido por móveis, paredes e portas. Nessa situação, percebe-se que há perda de pacotes. Isso é resultado da grande quantidade de obstáculos entre os dois dispositivos, bem como das limitações de hardware dos microcontroladores. Esses resultados não são preocupantes pois não representam uma situação real, já que a aplicação deste projeto foi idealizada para locais abertos.

Tabela 3: resultado de envio de pacotes ESP-NOW com dispositivos a 5 metros de distância e obstáculos

Pacote	Resultado
Mestre → slave 1	OK
Slave → mestre 1	OK
Mestre → slave 2	NOK
Slave → mestre 2	NOK
Mestre → slave 3	OK
Slave → mestre 3	OK
Mestre → slave 4	OK
Slave → mestre 4	OK
Mestre → slave 5	OK
Slave → mestre 5	NOK

F. Comparação com Tecnologias Existentes

Com o objetivo de contextualizar a proposta desenvolvida, o sistema pneumático-eletrônico foi comparado com outras tecnologias frequentemente empregadas para aquisição de sinais mecânicos em aplicações esportivas. Embora sensores inerciais, sensores piezoelétricos e células de carga apresentem vantagens específicas, cada tecnologia possui limitações relacionadas ao custo, à robustez mecânica, à complexidade de integração ou à portabilidade. Nesse contexto, a arquitetura proposta procura equilibrar simplicidade construtiva, baixo custo, elevada resistência mecânica e estabilidade metrológica.

Tabela 2: Vantagens competitivas da tecnologia pneumática proposta frente aos sistemas de medição na literatura

Tecnologia	Custo	Resiliência	Complex.	Portab.
Proposto	Baixo	Alta	Baixa	Exc.
IMUs	Médio	Alta	M. Alta	Exc.
FSR	Baixo	Baixa	Média	Boa
Piezo	Médio	Média	Alta	Boa
Célula	Alto	Alta	Baixa	Lim.
Ópticos	M. Alto	N/A	M. Alta	Nula

(Nota de legenda: M. Alto/M. Alta: Muito Alto / Muito Alta; Exc.: Excelente; Lim.: Limitada; N/A: Não se aplica devido à ausência de impacto mecânico direto).

G. Diagrama de Fluxo da Arquitetura do Sistema

A Figura 2 apresenta o fluxo funcional da arquitetura desenvolvida, desde a emissão do estímulo visual até o processamento local e transmissão das informações ao aplicativo móvel. O diagrama evidencia a integração entre o sistema embarcado, o sensor pneumático, o processamento temporal local e a comunicação Bluetooth Low Energy, sintetizando a lógica de funcionamento proposta para o dispositivo.

III. PROTOCOLOS DE AVALIAÇÃO E ANÁLISE ESTATÍSTICA

A. Especificação do Gesto Motor e Protocolos de Teste

Para garantir a reprodutibilidade da cinemática de impacto e minimizar a variabilidade técnica inerente aos esportes de combate, o soco frontal direto (*cross*) desferido com o membro dominante será adotado como gesto motor padronizado na Fase II. A escolha dessa técnica fundamenta-se em sua elevada reprodutibilidade biomecânica, estabilidade do vetor de impacto e ampla utilização em estudos experimentais envolvendo desempenho de golpes, permitindo comparações consistentes entre atletas e sessões experimentais.

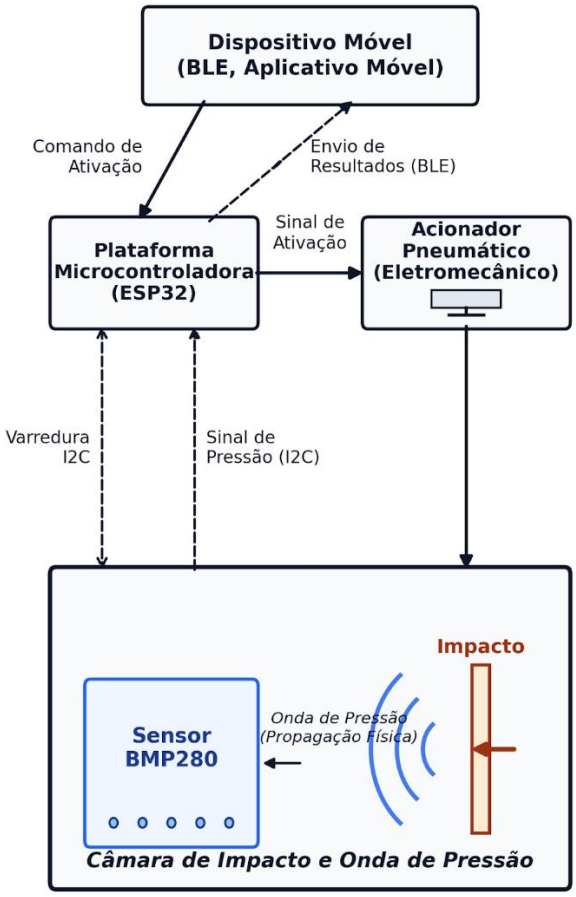


Fig. 2. Diagrama da arquitetura do sistema inteligente de captação e processamento local de dados em ensaios de bancada.

Todos os ensaios serão realizados com o atleta em posição inicial padronizada, respeitando a distância de execução previamente estabelecida e mantendo condições ambientais semelhantes entre as sessões, com o objetivo de reduzir fontes externas de variabilidade experimental.

Os participantes serão submetidos a dois protocolos computadorizados.

Tempo de Resposta Simples (TRS): Avalia a eficiência neuromuscular eferente bruta. O atleta responderá ao acendimento de um único LED central pré-determinado, desferindo o soco direto com o máximo de velocidade e impacto possíveis, registrando o pico de pressão interna.

Tempo de Resposta de Escolha e Coordenação Espacial (TRE): Introduz a carga cognitiva e a leitura tática do ambiente. Múltiplos estímulos luminosos distribuídos no espaço serão acionados aleatoriamente em diferentes coordenadas do gabinete, forçando o atleta a processar a informação visual e selecionar a trajetória espacial correta para atingir o alvo correspondente.

Para mitigar os efeitos da fadiga muscular, cada atleta executará três tentativas válidas por protocolo, com intervalos de 1 minuto de recuperação entre os disparos.

Os ensaios de bancada realizados na Fase I validam a prontidão tecnológica do protótipo para a execução fidedigna desses protocolos. A alta resolução temporal obtida (6,36 ms)

e a imunidade total do cálculo do tempo de reação frente à latência de comunicação Bluetooth (graças ao *timestamping on-sensor* local) asseguram que o dispositivo possui robustez metrológica para discriminar pequenas variações de desempenho tático-motor nas rotinas de alta velocidade exigidas na Fase II.

B. Processamento de Sinais e Validação de Construto

O algoritmo embarcado utiliza um limiar mínimo de ativação para eliminar impactos de baixa intensidade e oscilações decorrentes de ruídos mecânicos. Esse procedimento reduz a ocorrência de falsos positivos e aumenta a robustez da detecção automática dos eventos de impacto.

As janelas temporais de processamento serão configuradas para rejeitar marcações incompatíveis com os limites fisiológicos esperados para o tempo de resposta humano. Valores inferiores a 150 ms ou superiores a 500 ms serão analisados criticamente durante o processamento, uma vez que podem representar erros de aquisição, antecipação da resposta ou falhas operacionais durante o experimento.

Para avaliação da validade concorrente, as medidas obtidas pelo sistema pneumático-eletrônico serão comparadas com um sistema de referência baseado em fotocélulas infravermelhas. A utilização de um equipamento amplamente empregado na literatura como método de referência permitirá verificar a concordância entre as duas metodologias de aquisição temporal. A concordância será avaliada por meio da análise de Bland-Altman e do coeficiente de correlação de Pearson. Entretanto, considerando que elevada correlação não implica necessariamente concordância entre métodos, a interpretação dos resultados será fundamentada prioritariamente nos limites de concordância obtidos pelo método de Bland-Altman. A estabilidade teste-reteste será determinada utilizando o Coeficiente de Correlação Intraclasse ICC(2,1), complementado pelo cálculo do erro padrão da medida (SEM) e da mínima mudança detectável (MDC95). Esses indicadores permitirão quantificar simultaneamente a confiabilidade relativa e absoluta do equipamento, fornecendo uma caracterização metrológica mais abrangente do sistema desenvolvido.

C. Tratamento Estatístico Longitudinal

A distribuição e a normalidade dos dados coletados serão inicialmente testadas por meio do teste de Shapiro-Wilk, enquanto a homogeneidade das variâncias será avaliada pelo teste de Levene. Uma análise estatística descritiva prévia será conduzida para verificar a homogeneidade das características gerais da amostra (idade, massa corporal e estatura), garantindo que os grupos não difiram significativamente além do nível de expertise.

Considerando o caráter longitudinal do desenho de pesquisa com coletas repetidas em sessões semanais ao longo de oito semanas, a comparação dos tempos de resposta entre os grupos será realizada por meio de uma ANOVA fatorial de medidas repetidas [Grupo (2) × Tempo (8)]. O tamanho de efeito mínimo esperado foi fixado em $f = 0,40$ (equivalente a $d = 1,1$ para comparações pareadas). As premissas de esfericidade serão testadas pelo teste de Mauchly, adotando-se a correção de Greenhouse-Geisser caso a esfericidade seja violada. Em cenários com perda amostral controlada, serão aplicados Modelos Lineares Mistos com interceptos aleatórios por sujeito e matriz de covariância estruturada sob o modelo autorregressivo de primeira ordem [AR(1)]. O teste post hoc de Bonferroni será utilizado para identificar as diferenças localizadas entre os tempos e grupos. Todas as análises estatísticas serão executadas adotando-se um nível de significância de $p < 0,05$.

IV. CONCLUSÕES

O presente estudo apresentou o desenvolvimento e a validação preliminar de bancada de um sistema inteligente pneumático-eletrônico destinado à avaliação cognitivo-motora em esportes de combate. Os ensaios experimentais demonstraram estabilidade metrológica, baixa variabilidade instrumental e independência entre a latência da comunicação Bluetooth e o cálculo temporal realizado localmente pelo microcontrolador. Esses resultados evidenciam a viabilidade técnica da arquitetura proposta e estabelecem uma base consistente para futuras validações ecológicas envolvendo atletas em condições reais de treinamento.

V. REFERÊNCIAS

[1] M. T. O. Worsey *et al.*, "Inertial sensors for performance analysis in combat sports: A systematic review," *Sports*, vol. 7, no. 2, p. 28, 2019.

[2] V. Romanenko *et al.*, "Methodology for assessing spatial perception in martial arts," *Appl. Sci.*, vol. 15, no. 6, p. 3413, 2025.

[3] I. T. Mülhim and F. Akcan, "Comparison of simple visual and auditory reaction times of martial arts athletes," *Eur. J. Phys. Educ. Sport Sci.*, vol. 7, no. 5, pp. 174–185, 2021.

[4] A. S. O. Silva *et al.*, "Markerless combat sports biomechanics using MediaPipe and BlazePose: A validation study," *Journal of Sports Sciences*, vol. 42, no. 3, pp. 245–253, 2024.

[5] T. Suzuki and Y. Nakamura, "Evaluation of wearable sensors and MoveNet for strike detection in martial arts," *Sensors*, vol. 24, no. 8, p. 2614, 2024.

[6] I. Loturco *et al.*, "A comparative study of specific reaction time in elite boxers: Differences between jabs and crosses," *J. Athl. Enhancement*, vol. 4, no. 3, pp. 1–4, Jul. 2015.

[7] I. Loturco *et al.*, "Strength and power qualities are highly associated with punching impact in elite amateur boxers," *J. Strength Cond. Res.*, vol. 30, no. 1, pp. 109–116, Jan. 2016.

[8] A. C. Paulo *et al.*, "Kick force assessment of detainees inside a police patrol vehicle restraint," *Rev. Ciência & Polícia*, vol. 11, no. 1, pp. 22–35, 2025.

[9] K. Picha *et al.*, "Reliability of five novel reaction time and cognitive load protocols," *J. Sport Rehabil.*, vol. 28, no. 7, pp. 716–721, 2018.

[10] O. C. Santos, "Psychomotor learning in martial arts: An opportunity for user modeling, adaptation and personalization," in *Proc. Adjunct Publ. 25th Conf. User Modeling, Adaptation and Personalization (UMAP '17)*, New York, ACM, 2017, pp. 89–92.

Design and Characterization of a Low-Cost Electroencephalography Analog Front-End

Belchior Dias Cruz
Department of Computer Engineering
Federal University of Paraíba
João Pessoa, Brazil
ORCID: 0009-0003-4031-3220

Adelson Ataíde dos Santos Neto
Department of Electrical Engineering
Federal Institute of Paraíba
João Pessoa, Brazil
adelson.ataide@academico.ifpb.edu.br

Arthur Bernardo Barbosa
Department of Electrical Engineering
Federal Institute of Paraíba
João Pessoa, Brazil
bernardo.arthur@academico.ifpb.edu.br

Helder Rolim Florentino
Department of Electrical Engineering
Federal Institute of Paraíba
João Pessoa, Brazil
helderflorentino@hotmail.com

***Abstract*—There are multiple applications for brain signal acquisition hardware in a research environment, such as studies involving brain-computer interfaces, characterization of neurological disorders and different kinds of brain activity analysis. This type of research, however, can be limited by the high cost of commercial devices capable of recording neural rhythms. To address this limitation, this research presents the design and characterization of a low-cost single-channel electroencephalography (EEG) acquisition circuit through computational simulation and laboratory measurements. A custom printed circuit board (PCB) was developed for hardware testing. Simulated and experimentally measured Bode plots of each circuit stage were compared. Characterization resulted in an overall measured voltage gain of 79.04 dB, a calculated bandwidth of 36.72 Hz, and a measured CMRR of 115.4 dB. Results indicate that the proposed circuit could be a viable low-cost analog front-end for EEG acquisition. Future work will conduct human subject testing to evaluate the system's performance.**

***Index Terms*—electroencephalogram (EEG), low-cost hardware, biomedical instrumentation, signal acquisition**

I. Introduction

Brain activity has been the subject of scientific research since the 1920s [1] and has become increasingly important in recent years due to advances in fields such as neuroscience, psychiatry, psychology and biomedical engineering, finding applications ranging from the investigation of neurological disorders to brain-computer interfaces [2].

Brain signals can be acquired through invasive and non-invasive methods. Invasive methods use surgery to place sensors inside the skull, examples include electrocorticography (ECoG) and intracortical microelectrode arrays (MEAs), while non-invasive methods aim to record signals from outside the skull without the need for surgery, for example, electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS) [3].

EEG is one of the most used methods of brain signal acquisition for academic purposes due to its high temporal resolution, portability and relatively low cost compared with other acquisition techniques [4]. Despite this, many researchers are unable to purchase commercial EEG systems, whose prices can reach thousands of dollars [5].

A number of studies have been conducted to overcome this problem by designing low-cost custom EEG acquisition systems [6]–[12]. However, many of them provide limited experimental characterization of each stage of the developed systems [6]–[12], which can hinder researchers from identifying issues when reproducing the proposed design if the measured circuit behavior diverges from the simulated response.

Under these circumstances, this study proposes the design and characterization of a low-cost single-channel EEG analog front-end (AFE) using computational simulation and experimental measurements, followed by a comparison between simulated results and laboratory measurements. One of the future research goals is to use the developed circuit in brain-computer interface systems.

II. Background and Literature Review

A. Signal Characteristics

EEG signal acquisition is made through the placement of electrodes on a subject's scalp, following one of the standard positioning systems such as the 10-20, 10-10, or 10-5 systems [13], [14]. Examples of electrodes include reusable electrodes, which can be composed of highly conductive materials such as gold or silver/silver chloride, and disposable electrodes, such as gel electrodes [13], [15]. The connection between the electrodes and the EEG acquisition system is done via shielded cables [15].

The electrical signals are measured at the scalp with amplitude ranging from 0.5 to 100 μV [13]. These signals are affected by multiple artifacts such as eye movement artifacts (EOG), cardiac activity (ECG), muscle activity (EMG) and power line interference (PLI), usually having higher amplitudes than the EEG signal itself [16]. For these reasons, EEG

recording requires an acquisition system capable of reducing the artifacts and amplifying the original signal [13].

The brain signals can also be divided into different rhythms based on their frequencies, each one of the rhythms being associated with specific brain states [16], [17]. The neural rhythms and its respective frequency ranges are displayed in Table I [16].

TABLE I
NEURAL RHYTHMS

Rhythm	Frequency range (Hz)
Delta	0.1 - 4
Theta	4 - 8
Alpha	8 - 13
Beta	14 - 30
Gamma	30 - 80

B. Electronic filters and signal amplification

The development of an electronic circuit capable of acquiring EEG signals usually requires the implementation of electronic filters and amplifiers [13].

Electronic filters are classified into two main types, passive filters, which attenuate specific frequency bands using passive components such as resistors and capacitors, and active filters, which in addition to attenuating frequency bands can also amplify the output using operational amplifiers [18].

Filters can also be divided into four categories based on the frequency band they attenuate. High-pass filters attenuate signal components with frequency below their cutoff frequency, low-pass filters attenuate signal components with frequency above their cutoff frequency, band-pass filters attenuate all signal components outside a specific band and notch filters attenuate signal components within a specific narrow band [18].

There are different kinds of topologies for electronic filters. The Sallen-Key topology allows for a simple implementation of higher-order filters, which results in a steeper roll-off in the transition band [18]. The cutoff frequency of a Sallen-Key filter with Butterworth approximation is defined by (1), where fc is the cutoff frequency, R is a resistor, and C is a capacitor [19]. Sallen-Key topology can also be designed to implement different filter approximations, such as the Butterworth, Chebyshev and Bessel [18].

$$fc = \frac{1}{2\pi RC} \tag{1}$$

Active filters have a voltage gain, defined as the amplitude relation between input and output signals. The gain of an active Sallen-Key filter is defined by (2), where G is the gain and R_1 and R_2 are resistors [18].

$$G = 1 + \frac{R_1}{R_2} \tag{2}$$

Common-mode rejection ratio (CMRR) is a relevant feature of EEG acquisition systems. It represents the capacity of a circuit to attenuate noise present in common-mode input and is defined by (3), where A_d is the differential gain and A_c is the common-mode gain [13], [20].

$$CMRR = 20 \times \log_{10}(\frac{A_d}{A_c}) \tag{3}$$

The high importance of CMRR for the development of EEG acquisition systems is due to the fact that major sources of interference, such as PLI, can often appear as common-mode signals [13].

Another important characteristic for EEG acquisition is the input-referred noise (IRN), which can be determined by short-circuiting the system's inputs, measuring the output RMS voltage and dividing it by the differential gain of the circuit [21]. Due to the low amplitude of EEG signals, high IRN can completely mask them, compromising signal integrity [13].

III. MATERIALS AND METHODS

A. System Architecture

To avoid saturation, signal amplification occurs in different circuit stages, as displayed in Fig. 1, consisting of an instrumentation amplifier stage, a high-pass filter, a low-pass filter and a notch filter.

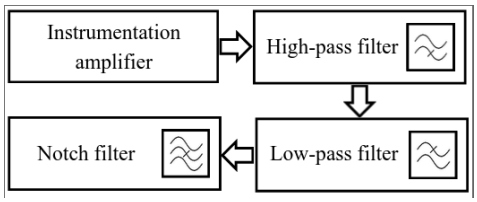


Fig. 1. Circuit stages.

The instrumentation amplifier chosen was INA101HP, mainly because of its CMRR. This stage's gain is defined in (5) derived from (4), where R_G is a resistor and G is the gain [22].

$$G = 1 + \frac{40k}{R_G} \tag{4}$$

$$G = 1 + \frac{40k}{2.2k} \approx 19.18 \tag{5}$$

INA101HP has minimum CMRR ranging from 60 dB (for gain of 1) to 100 dB (for gains of 100 to 1000) at 60 Hz and typical CMRR ranging 85 dB (for gain of 1) to 105 dB (for gains of 100 to 1000) at 60 Hz. Other relevant component characteristics include differential and common-mode input impedance of 10 GΩ $\parallel$ 3 pF, maximum input bias current of 30 nA and typical input bias current of 15 nA, and input voltage noise of 0.8 μV_{p-p} at 0.1-10 Hz [22].

The operational amplifiers implemented in the other circuit stages were of the model TL084CN. To calculate the cutoff frequency and gain of each circuit stage (1) and (2) were respectively used. The high-pass filter stage features a first-order high-pass Sallen-Key filter with Butterworth approximation, having a gain of 101 defined in (6) and cutoff frequency of 0.50 Hz defined in (7). The purpose of the high-pass filter

stage is to attenuate low frequency artifacts, such as baseline drift and EOG.

$$G = 1 + \frac{330k}{3.3k} = 101 \tag{6}$$

$$fc = \frac{1}{2\pi \times 680 \times 10^3 \times 470 \times 10^{-9}} \approx 0.50\,Hz \tag{7}$$

The low-pass filter stage consists of a fourth-order low-pass Sallen-Key filter with Butterworth approximation, having gains of approximately 1.17 and 2.22, as defined in (8) and (9), and cutoff frequency of 37.21 Hz defined in (10). The purpose of the low-pass filter stage is to attenuate high frequency artifacts, such as EMG.

$$G = 1 + \frac{5.6k}{33k} \approx 1.17 \tag{8}$$

$$G = 1 + \frac{2.2k}{1.8k} \approx 2.22 \tag{9}$$

$$fc = \frac{1}{2\pi \times 9,1 \times 10^3 \times 470 \times 10^{-9}} \approx 37,21\,Hz \tag{10}$$

The notch filter stage features a second-order notch Sallen-Key filter with Butterwoth approximation, having a gain of 1.77 and cutoff frequency of 60.28 Hz, as defined in (11) and (12). The purpose of the notch filter stage is to attenuate the PLI artifact.

$$G = 1 + \frac{10k}{13k} \approx 1.77 \tag{11}$$

$$fc = \frac{1}{2\pi \times 22 \times 10^3 \times 120 \times 10^{-9}} \approx 60,28\,Hz \tag{12}$$

The complete circuit has a gain of 78.99 dB and a bandwidth of 36.72 Hz, as defined in (13) and (14). The circuit's diagram is presented in Fig. 2.

$$G_t = 20 \times \log_{10}(19.18 \times 101 \times 1.17 \times 2.22 \times 1.77) \approx 78.99\,dB \tag{13}$$

$$\Delta f = 37,21 - 0.50 = 36.71\,Hz \tag{14}$$

B. Simulation

The circuit was simulated using LTspice, employing generic operational amplifiers in place of the TL084CN and modelling the INA101HP through its discrete circuit. The simulation diagram is presented in Fig. 3.

The simulation results were obtained through AC and transient analysis. The AC analysis was conducted utilizing a frequency sweep from 0.1 Hz to 1 KHz. The transient analysis lasted 5 seconds, in which a waveform was generated to represent the expected characteristics of EEG signals. The waveform was composed of multiple signals, each with the properties of specific signal components present in EEG acquisition, presented in Table II [16], [23].

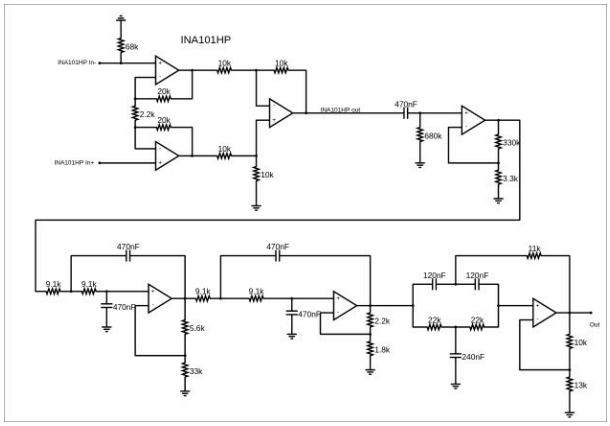


Fig. 2. Circuit diagram.

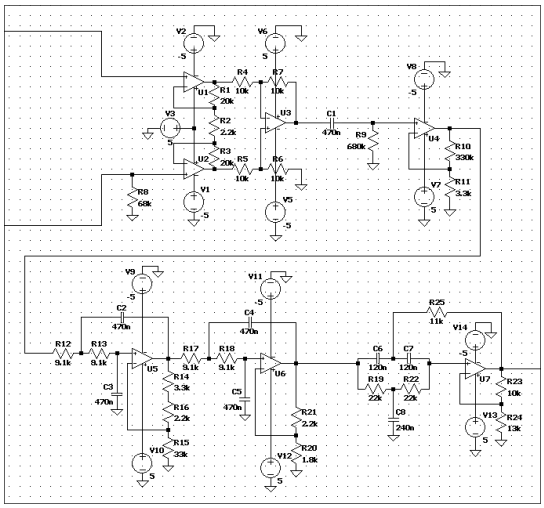


Fig. 3. Circuit simulation in LTspice.

C. Hardware Implementation

The proposed circuit was implemented on a custom printed circuit board (PCB) designed using the EasyEDA software. All passive components used were through-hole and had a tolerance of 5%. After fabrication, the PCB was placed inside a repurposed metallic DC power supply enclosure to provide additional protection against electromagnetic interference. The PCB is presented in Fig. 4. Two 9-V batteries are used as a split power supply for the EEG front-end.

Laboratory measurements were made using a DC power supply (U8031A, Agilent Technologies) to power the circuit,

TABLE II
SIGNAL COMPONENTS.

Waveform	Signal	Frequency	Amplitude
Sinusoidal	Alpha rhythm	10 Hz	10 μV
Sinusoidal	Baseline drift	0.5 Hz	20 μV
Triangular	EOG[a]	0.33 Hz	100 μV
Sinusoidal	PLI	60 Hz	1 mV

[a]Pulse width of 0.3 seconds.

a signal generator (33521A, Agilent Technologies) to produce the input signal, and an oscilloscope (DSO-X 2012A, Agilent Technologies) to measure the output signal.

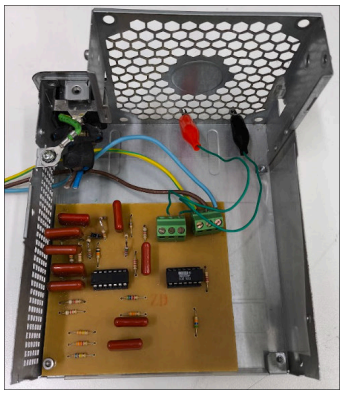


Fig. 4. Printed circuit board.

IV. Results and Discussion

The electronic components used in the PCB manufacturing and their respective prices are presented in Table III. The cost estimation was done based on component prices available from online electronic component distributors in September 2026.

TABLE III
Bill of Materials.

Quantity	Component	Unitary price	Total price
18	Resistor	0.1 USD	1.8 USD
8	Polyester capacitor	0.76 USD	6.08 USD
1	TL084CN	1.07 USD	1.07 USD
1	INA101HP	32.57 USD	32.57 USD
Total:		41.52 USD	

As the instrumentation amplifier accounts for approximately 80% of the hardware cost, the replacement of the INA101HP with modern alternatives, such as the INA333, INA114, or INA828, would reduce the overall cost. The substitution would require circuit changes, such as a power-supply redesign (depending on the selected component), and the adaptation to surface-mount device footprints.

Both the simulated and assembled circuits were characterized. Each figure presents the simulated and measured Bode plots of a circuit stage, with the Figs. 5, 6 and 7 displaying respectively the high-pass filter, the low-pass filter and the notch filter frequency responses. The simulated Bode plot, shown on the left, represents the gain magnitude on the left y-axis with a solid line, the phase on the right y-axis with a dotted line, and the frequency on the x-axis, while the measured bode plot, shown on the right, represents the gain magnitude on the y-axis and the frequency on the x-axis.

The simulated and measured Bode plots show similar responses for the simulated and implemented circuits. However, it is still possible to identify some divergence between the results. The simulated and measured high-pass and low-pass filter responses are nearly identical, but the simulated notch

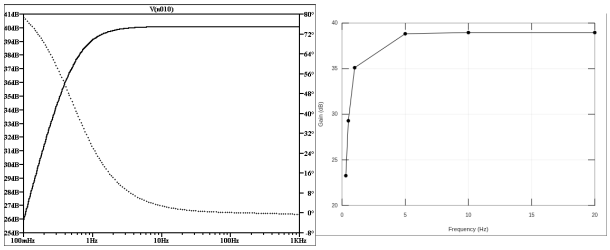


Fig. 5. Bode plots of simulated (left) and measured (right) high-pass filters.

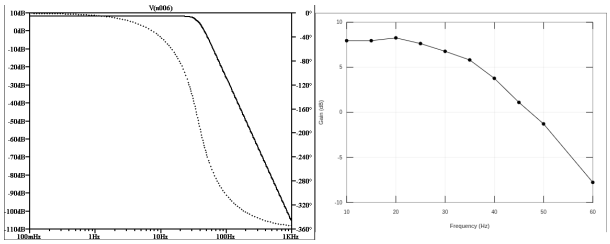


Fig. 6. Bode plots of simulated (left) and measured (right) low-pass filters.

filter plot displays greater attenuation of 60 Hz components compared with the measured plot, with the simulated response reaching a gain magnitude of approximately -50 dB, while the measured response reaches approximately -20 dB.

The Bode plot of the complete circuit is shown in Fig. 8, with the simulated results on the left and the measured on the right. The observed frequency response of both the simulated and measured circuits is consistent with the responses of each of the circuit stages. Lower attenuation at 60 Hz and an earlier low-pass roll-off can be noted on the measured response when compared to the simulated one. The measured response also shows that the previously calculated gain of 78.99 dB remains stable in the range of 1 Hz to 15 Hz, reducing to 75-70 dB between 15 Hz and 37 Hz, and decreasing further at higher frequencies. These response differences between the simulated and physical circuits could be caused by multiple factors such as component tolerances, parasitic capacitances, and limitations in the utilized equipment when generating and recording low-amplitude signals.

The simulated transient response is shown in Fig. 9, with the circuit's input shown on the top, and the circuit's output shown on the bottom. The circuit's transient behavior in simulation shows noticeable attenuation of the PLI component, which was

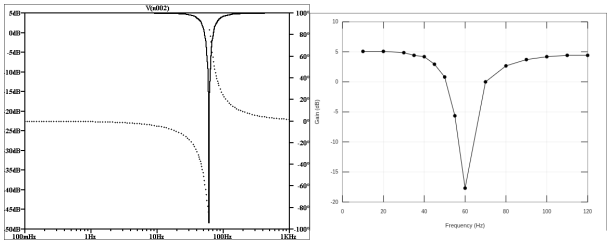


Fig. 7. Bode plots of simulated (left) and measured (right) notch filters.

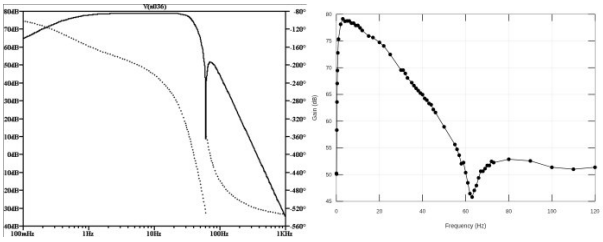


Fig. 8. Bode plots of simulated (left) and measured (right) circuit.

the frequency component with the highest amplitude in the input signal, and significant amplification of the alpha rhythm component, which had an amplitude of 10 μV at the input and approximately 100 mV at the output. The other signal components, representing simulated EOG signals and baseline drift, were still distinguishable in the output signal.

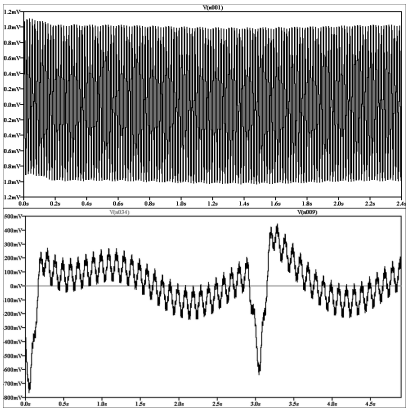


Fig. 9. Simulated circuit's input and output on transient analysis.

The circuit has a measured power consumption of 119.7 mW, which is significantly higher than the power consumption of commercially available EEG AFEs [24], [25]. During the laboratory tests, each of the 9 V batteries used had a capacity of 580 mAh (an energy capacity of 5.22 Wh per battery), resulting in an autonomy of approximately 87.2 hours, as calculated in (15).

$$\frac{5.22\,Wh \times 2}{0.1197\,W} = 87.2\,h \tag{15}$$

At 10 Hz, circuit measurements resulted in a differential voltage gain of 8950 ($\approx$ 79.04 dB), consistent with the calculated gain of 78.99 dB. The measured common-mode voltage gain was 0.0152 at 10 Hz, with a CMRR of 115.4 dB, as calculated in (16). Accurate measurements of the system's input impedance and IRN were not feasible using available equipment. The measurement of input impedance was limited by the loading effect, since the oscilloscope's impedance is 1 MΩ and the INA101HP has a theoretical impedance of 10 GΩ $\|$ 3 pF. The noise floor of the measurement setup masked the circuit's output noise, preventing the determination of the IRN. The system's theoretical IRN is mainly dictated by the INA101HP [21], with a value of 0.8 μV_{p-p} at 0.1-10 Hz [22].

$$G_t = 20 \times \log_{10}(8950/0.0152) \approx 115.4\,dB \tag{16}$$

The measured circuit's CMRR of 115.4 dB meets the recommended CMRR for EEG acquisition systems, which usually have values above 100 dB [13]. It is important to note, however, that in EEG recording the effective CMRR can be reduced by differences between electrode impedances [26]. The datasheet-specified input impedance and input-referred noise values of the INA101HP are also consistent with the criteria for EEG recording equipment [13].

Table IV presents the electrical characteristics of commercial EEG AFE integrated circuits (ICs), low-cost EEG acquisition systems developed in other studies, and the AFE developed and characterized in this research. Studies that provide limited information about the electrical characteristics of the systems were not included in the table as they would add limited value to the comparison. However, the system developed in [6] was included to serve as an example of a system that uses the ADS1298 and to illustrate its price.

The prices of the commercial AFEs were obtained in September 2026 from online electronic component distributors, while their electrical characteristics were sourced from the datasheet of each component [24], [25].

Although the unit costs of the ADS1298 and ADS1299 are similar to the 41.52 USD of the developed circuit, using them in practical EEG acquisition systems would demand a higher implementation of additional hardware. The application of these ICs typically involves additional components, infrastructure, and digital processing hardware, increasing the system's cost [6], [12].

TABLE IV
COMPARISON OF EEG ACQUISITION SYSTEMS AND AFES.

System	CMRR	Gain	IRN	Bandwidth	Channels	Reported Cost/Market Price
This study (Discrete AFE)	115.4 dB	79.04 dB	0.8 μV_{p-p} [a]	0.50-37.21 Hz	1	41.52 USD
[8] (Full System)	>100 dB	90.88-93.97 dB	-	0.5-35.54 Hz	1	25 USD
[10] (Full system)	93.16 dB	60 dB	-	0.159-100 Hz	1	-
[6] (System with ADS1298)	-	-	-	-	8	<100.00 USD
[12] (System with four ADS1299)	109.37 dB	-	<0.8 μV_{p-p}	-	32	140.00 USD
ADS1298 (AFE IC)	105-115 dB	0-21.58 dB[b]	4 μV_{p-p}	-	8	51.52 USD
ADS1299 (AFE IC)	110-120 dB	0-27.60 dB[b]	1 μV_{p-p}	-	4-8	44.98-88.58 USD

[a]Theoretical value.
[b]Gain of the internal programmable gain amplifier.

V. Conclusion

To summarize, the scarcity of low-cost experimentally characterized EEG acquisition systems was addressed in this study through the design and characterization of a low-cost electroencephalography analog front-end. The presented circuit was evaluated using simulation and laboratory measurements.

The system's performance was verified through analysis of the theoretical and experimentally obtained transfer functions. Despite costing less than most commercially available systems, the proposed circuit presents the characteristics recommended for EEG acquisition, such as a CMRR of 115.4 dB, a gain of 79.04 dB, and a bandwidth from 0.50 Hz to 37.21 Hz.

Limitations of the developed system include the power consumption and cost of the INA101HP, which could be replaced with another lower-cost, lower-power instrumentation amplifier, requiring power-supply modifications and applying surface-mount device footprints. The utilization of standard-tolerance passive components, potentially responsible for the discrepancies between simulated and measured results, is another issue that can be addressed.

Future studies will involve human-subject testing to evaluate the system's performance in real EEG acquisition, focusing on brain-computer interface systems. While performing tests with an individual, connecting the circuit to mains-powered equipment will require galvanic isolation to ensure participant safety.

References

[1] P. M. Rossini, J. Cole, W. Paulus, U. Ziemann, and R. Chen, "1924–2024: First centennial of eeg," *Clinical Neurophysiology*, vol. 170, pp. 132–135, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245724003778

[2] Y. Qin, Y. Zhang, Y. Zhang, S. Liu, and X. Guo, "Application and development of eeg acquisition and feedback technology: A review," *Biosensors*, vol. 13, no. 10, 2023. [Online]. Available: https://www.mdpi.com/2079-6374/13/10/930

[3] Y. Sun, X. Chen, B. Liu, L. Liang, Y. Wang, S. Gao, and X. Gao, "Signal acquisition of brain–computer interfaces: A medical-engineering crossover perspective review," *Fundamental Research*, vol. 5, no. 1, pp. 3–16, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667325824001559

[4] R. Abreu, A. Leal, and P. Figueiredo, "Eeg-informed fmri: A review of data analysis methods," *Frontiers in Human Neuroscience*, vol. Volume 12 - 2018, 2018. [Online]. Available: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2018.00029

[5] D. Dadebayev, W. W. Goh, and E. X. Tan, "Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4385–4401, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157821000732

[6] A. Jain, I. Kim, and B. J. Gluckman, "Low cost electroencephalographic acquisition amplifier to serve as teaching and research tool," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2011, pp. 1888–1891.

[7] M. Aljshamee, R. Alsabah, A. H. Mousa, and A. Alsabbagh, "(portable) low-cost electroencephalogram (eeg) recorder for brain computer-interface (bci)," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, July 2020, pp. 1–6.

[8] P. Tawheed, J. Mollick, N. Sakib, and M. K. Islam, "Development of a low-cost pc-based single-channel eeg acquisition system for educational and research purpose," in *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, Sep. 2021, pp. 1–6.

[9] S. Baltzis, G. Pagiatakis, N. Voudoukis, A. Papadakis, L. Dritsas, and D. Uzunidis, "An arduino-based, portable prototype for the recording and analysis of eeg signals to support self-detection and self-monitoring of stress," *Sensors*, vol. 26, no. 11, 2026. [Online]. Available: https://www.mdpi.com/1424-8220/26/11/3410

[10] H. Tanaka and S. Kondo, "Development of a wearable eeg device toward bci applications," in *AHFE International Conference Proceedings*, ser. Applied Human Factors and Ergonomics, vol. 199. AHFE International, 01 2025, pp. 798–807.

[11] A. K. Whitchurch, J. K. Abraham, M. A. Lonkar, and V. K. Varadan, "Design of a compact amplifier and signal conditioning module for wireless eeg monitoring," in *2007 IEEE Region 5 Technical Conference*, April 2007, pp. 153–156.

[12] M. Du, H. Gou, M. Yuan, P. Wu, Y. Chen, C. Chang, G. Q. Zeng, and X. Zhang, "A low-cost 32-channel eeg signal acquisition system for wearable applications," in *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, Dec 2024, pp. 1320–1323.

[13] M. Teplan, "Fundamentals of EEG measurement," *Measurement Science Review*, vol. 2, no. 2, pp. 1–11, 2002.

[14] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811906009724

[15] A. Searle and L. Kirkup, "A direct comparison of wet, dry and insulating bioelectric recording electrodes," *Physiological Measurement*, vol. 21, no. 2, p. 271, may 2000. [Online]. Available: https://doi.org/10.1088/0967-3334/21/2/307

[16] H. Hu, S. Guo, R. Liu, and P. Wang, "An adaptive singular spectrum analysis method for extracting brain rhythms of electroencephalography," *PeerJ*, vol. 5, p. e3474, 2017.

[17] C. Babiloni, R. J. Barry, E. Başar, K. J. Blinowska, A. Cichocki, W. H. Drinkenburg, W. Klimesch, R. T. Knight, F. Lopes da Silva, P. Nunez, R. Oostenveld, J. Jeong, R. Pascual-Marqui, P. Valdes-Sosa, and M. Hallett, "International federation of clinical neurophysiology (ifcn) – eeg research workgroup: Recommendations on frequency and topographic analysis of resting state eeg rhythms. part 1: Applications in clinical research studies," *Clinical Neurophysiology*, vol. 131, no. 1, pp. 285–307, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245719311642

[18] S. Franco, *Design with Operational Amplifiers and Analog Integrated Circuits*, ser. Electrical and Computer Engineering Series. McGraw-Hill, 2002. [Online]. Available: https://books.google.com.br/books?id=LXMeAQAAIAAJ

[19] P. Horowitz and W. Hill, *The Art of Electronics*. Cambridge University Press, 1980. [Online]. Available: https://books.google.com.br/books?id=0rfDQgAACAAJ

[20] R. Pallas-Areny and J. Webster, "Common mode rejection ratio in differential amplifiers," *IEEE Transactions on Instrumentation and Measurement*, vol. 40, pp. 669–676, 09 1991.

[21] C. Motchenbacher and J. Connelly, *Low-Noise Electronic System Design*. Wiley, 1993. [Online]. Available: https://books.google.com.br/books?id=VhkoAQAAMAAJ

[22] Texas Instruments, *INA101 Very High Accuracy Instrumentation Amplifier*, Texas Instruments, Sep. 2000, rev. C.

[23] T. Neycheva, D. Dobrev, and V. Krasteva, "Common-mode driven synchronous filtering of the powerline interference in ecg," *Applied Sciences*, vol. 12, no. 22, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/22/11328

[24] Texas Instruments, *ADS129x Low-Power, 8-Channel, 24-Bit Analog Front-End for Biopotential Measurements Datasheet*, Texas Instruments, Dallas, Texas, USA, 2010, document SBAS459K. [Online]. Available: https://ti.com

[25] ——, *ADS1299 Low-Noise, 4-, 6-, 8-Channel, 24-Bit Analog-to-Digital Converter for EEG and Biopotential Measurements*, Texas Instruments, Dallas, Texas, USA, 2012, document SBAS499C. [Online]. Available: https://ti.com

[26] A. D. Legatt, "Impairment of common mode rejection by mismatched electrode impedances: Quantitative analysis," *American Journal of EEG Technology*, vol. 35, no. 4, pp. 296–302, 1995. [Online]. Available: https://doi.org/10.1080/00029238.1995.11080527

Design of a Functional Electrical Stimulation Device with Hybrid Converter in the Output Stage

1st Isabela Costa Trivellato
School of Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
isabelatrivellato@ufmg.br

2nd Renan Fernandes Kozan
Department of Electrical Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
renankozan@ufmg.br

3rd Lenin Martins Ferreira Morais
Department of Electronics Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
lenin@cpdee.ufmg.br

***Abstract*—The use of high-frequency switched-mode converters improves the energy efficiency of the circuit. In the context of functional electrical stimulation, a device using a hybrid switched-mode buck converter was designed to synthesize the stimulation signal. It consists of a proportional-integral (PI) controller, a boost converter, and a hybrid converter (formed by a buck converter and an inverter) that generates the biphasic current signal. The device delivers a stimulation current with an amplitude of 3–100 mA, a resolution of 1 mA, a frequency of 10–250 Hz, and a pulse width of 10–500 μs. The main advantage of this FES system, with switched-mode topology in the output stage, is its simplicity, requiring few components. In the simulation, the design achieves the target parameters of frequency, pulse width, and amplitude, within a maximum deviation of 20% from the reference value of each parameter.**

***Index Terms*—Functional Electrical Stimulation, FES, Buck Converter, Switched-Mode Converter, Inverter, H-bridge, PI Control.**

I. Introduction

Functional electrical stimulation (FES) is a technique used to artificially induce muscle contractions by injecting electrical current into the tissues of the human body, with the intention of producing movement [1]. This technique is widely used in the treatment of people with spinal cord injury (SCI) to improve their motor functions [2]. It is also applied in assisted gait therapy and is associated with improvements in gait, muscle strength, and cardiorespiratory fitness [1].

The output stage of an FES device is the only one that cannot be implemented in software. Its function is to transfer the energy of the low-amplitude pulses generated in the previous stages to the excitable tissues of the body [3].

In the literature, linear converters are commonly used in the output stage, typically voltage-to-current converters and current mirrors [3]. The limitations of this kind of design are related to the transistor mismatch [4], which increases the complexity of the design implementation, and the energy inefficiency [5]. This topology can achieve a stimulation current with low ripple, but the use of semiconductor devices in the linear operating region can waste most of the energy in the process and not deliver it to the final load [5]. The linear topology developed by Blanco et al. [4] is presented in Fig. 1.

By contrast, the use of a switched-mode topology in the output stage of an FES device aims to increase energy efficiency [6] and to reduce volume, seeking portability [7].

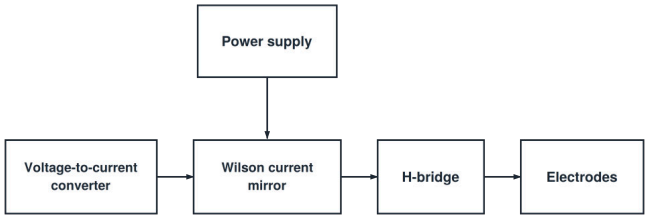


Fig. 1. Block diagram of the output stage with linear topology, adapted from [4].

This topology was chosen due to its simplicity, requiring few components, and to the use of semiconductor devices operating as switches, with only two states: ON or OFF. This switching reduces the energy dissipation involved in this process [5], [8]. As a result, the performance of the device is enhanced, since switched-mode topologies achieve an efficiency of 70–95%, while linear converters reach only 20–60% [5].

This work aims to simulate a surface FES device with a switched-mode topology in the output stage, which generates a biphasic, symmetric, square current, using pulse-width modulation (PWM), with an amplitude of 3–100 mA, a frequency of 10–250 Hz, and a pulse width of 10–500 μs. Fig. 2 shows the blocks that make up the output stage.

II. Materials and Methods

The FES design is divided into a PI controller, a boost converter, and a hybrid converter, formed by a buck converter and an inverter. The block diagram is shown in Fig. 3.

A. Proportional-Integral Control

The first block of the diagram is the PI controller. It is responsible for error correction, allowing an automatic adjustment of the stimulation to the desired value of the current

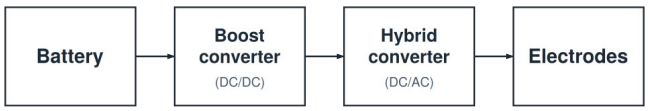


Fig. 2. Block diagram of the output stage with switched-mode topology.

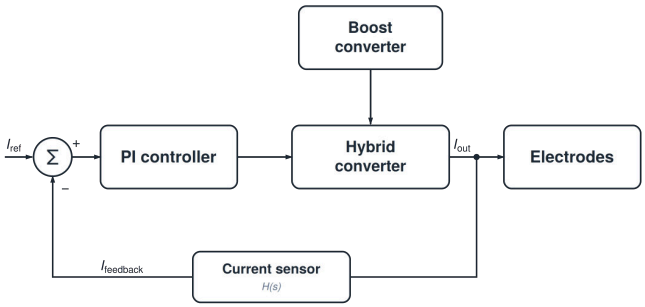


Fig. 3. Block diagram of the functional electrical stimulator design.

[9]. The controller was implemented in continuous time using an available control block in PSIM [10]. The values of K_p and K_i were chosen to obtain a step response with a settling time under 20% of the pulse width. The K_p and K_i values used in the design are shown in Table I.

The use of these parameters in an analog controller circuit would require resistors that implement a gain of 15 and a capacitor on the order of picofarads.

The PI control block is duplicated because the desired waveform is biphasic and symmetric. In the design, each block is adjusted to produce the correct stimulation current.

B. Boost Converter

The boost converter is used to step up the battery voltage to 100 V [8], producing the compliance voltage. This value is necessary for the H-bridge to deliver the desired output current amplitude. The fixed voltage in this stage causes higher energy losses, when compared to an adaptive supply [6]. However, aiming for a simpler implementation, it is the best option.

The battery used in the FES design enables the portability of the device and was one of the main reasons for choosing the switched-mode converter over a linear circuit [11]. This topology aims to provide greater circuit efficiency and, as a consequence, to extend the battery lifetime. This boost converter forms the second block of the design.

C. Hybrid Converter (Buck Converter + Inverter)

The third block of the design is the hybrid converter, formed by a buck converter and an inverter, which represents the output stage of the circuit. The H-bridge is the core of this stage since it enables the biphasic characteristic of the current signal. The waveform is produced by the change of current direction between the H-bridge branches, coordinated by the switching of the devices [8].

TABLE I
PI CONTROLLER PARAMETERS.

Parameter	Value
K_p	15
K_i	1.5×10^{10}

The circuit dedicated to synthesizing the pulse-width modulation (PWM) signal was designed using analog components: operational amplifiers configured as comparators, which produce a high-frequency signal [8]. This signal has two pre-defined frequencies, depending on the desired current amplitude: 250 kHz for low-amplitude currents, with values up to 10 mA, and 50 kHz for high-amplitude currents between 10 mA and 100 mA. This PWM signal is designed to activate two of the H-bridge switches, producing the average current value required for the electrical stimulation.

The output stage of the design is shown in Fig. 4.

Two independent square-wave voltage signals were used to activate switches S_1 and S_3, according to the desired pulse width and current frequency. In addition to activating switches S_1 and S_3, these signals are multiplied by the PWM signal, synthesized in the second block, to activate switches S_2 and S_4. The switch activation signals are shown in Fig. 5.

The S_2 and S_4 switching signals follow the PWM signal, which defines the average output current value. The current conduction path is established by the switching of S_1 and S_3. The parameters of the activation signals were calculated according to the operation of the buck converter. In addition, a phase shift was implemented between the activation signals of S_1/S_4 and S_2/S_3. This phase shift allows an interphase interval between the negative and positive half-cycles of the current, with a duration of 10 μs [4]. The phase shift implemented in the simulation is described by (1), which depends on the duty cycle (D) of the buck converter and the frequency (f) of the stimulation signal.

$$\phi = 360^\circ \times D + 360^\circ \times \left(10 \times 10^{-6}\right) \times f \tag{1}$$

The value of the inductor used in the buck converter (L) depends directly on the stimulation current ripple (Δi_L) and the converter switching frequency (f_{PWM}), as given in (2). The equation also relates these quantities to the output voltage (V_{out}) and duty cycle of the buck converter (D).

$$\Delta i_L = \frac{V_{\text{out}} \times (1 - D)}{L \times f_{\text{PWM}}} \tag{2}$$

The output current ripple of the buck converter is inversely proportional to the chosen inductor value and PWM switching frequency. For the lower current limit, a larger inductor would

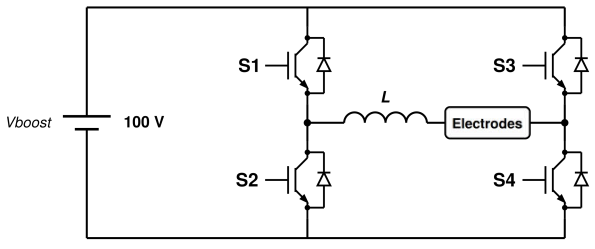


Fig. 4. Output stage of the designed device.

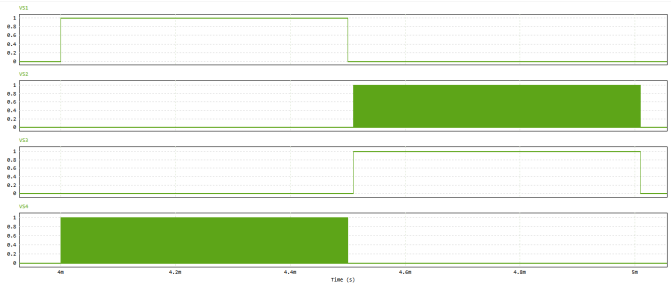


Fig. 5. Switch activation signals of the H-bridge switches.

be necessary to obtain an acceptable current ripple. To overcome this problem, the switching frequency was increased to 250 kHz. For the other current range, between 10 and 100 mA, the current ripple has less influence on the stimulation signal. Therefore, the switching frequency was set to 50 kHz.

In addition, the switching of the output inductors is also necessary, since with a higher inductance value, the current ripple is lower. As a consequence, with a smaller inductor, the current ripple is higher. Furthermore, the inductance value determines the rate of current variation, since with a lower inductance the current varies faster and vice versa [8]. Three inductance values are chosen for switching and combined with each other to select the desired inductance for each current range and for different pulse width and frequency values. The diagram representing the inductor switching is shown in Fig. 6.

The inductor values were chosen to obtain combinations of 1 mH, 2 mH, 4 mH, and 5 mH.

- $L_1 = 1\,\text{mH}$
- $L_2 = 1\,\text{mH}$
- $L_3 = 3\,\text{mH}$

The selection of inductance values depends on the switching of S_5, S_6, S_7, and S_8. This combination is shown in Table II.

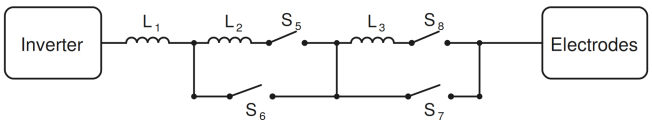


Fig. 6. Diagram of the inductor switching.

TABLE II
SWITCH STATES FOR EACH EFFECTIVE INDUCTANCE OF THE SWITCHING CIRCUIT.

Inductance (mH)	Combination	Switch states			
		S_5	S_6	S_7	S_8
1	L_1	OFF	ON	ON	OFF
2	$L_1 + L_2$	ON	OFF	ON	OFF
4	$L_1 + L_3$	OFF	ON	OFF	ON
5	$L_1 + L_2 + L_3$	ON	OFF	OFF	ON

D. Load Modeling

The simulation was set up with a purely resistive load, in series with the inductor, to verify the correct operation of the stimulation current source and to represent the electrical impedance of the biological tissues, without considering the capacitive component [4], [12]–[14].

III. Results and Discussion

The simulated circuit is capable of synthesizing a biphasic, symmetric, square waveform using a switched-mode output stage.

A. Validation of the Design Limits

Many stimulation current values were analyzed, but only the four scenarios that validate the simulated FES device will be detailed, under the most critical frequency and pulse width conditions.

- Case 1: 10 Hz and 10 μs;
- Case 2: 10 Hz and 500 μs;
- Case 3: 250 Hz and 10 μs;
- Case 4: 250 Hz and 500 μs.

The amplitude values obtained for the stimulation current in the critical scenarios, for reference amplitudes of 100 mA and 3 mA, an interphase interval of 10 μs, and a load of 1 kΩ [4], [12]–[14], are presented in Table III.

The relative percentage error for the worst case, with a reference amplitude of 3 mA, is approximately 9.20%, while for the 100 mA reference it is approximately 10.05%. This behavior occurs at the minimum pulse width chosen for the design, where the settling time has a strong influence on such short signals. The highest current ripple obtained was 9.8 mA, for a stimulation signal with an amplitude of 100 mA, a frequency of 250 Hz, and a pulse width of 500 μs.

The stimulation signal with a frequency of 10 Hz, a current amplitude of 100 mA, and a pulse width of 500 μs is shown in Fig. 7. A load of 1 kΩ was used in this case [4], [12]–[14]. This image was obtained from the simulation in the PSIM software [10], with a time step of 5 ns.

Fig. 8 shows the stimulation signal for a nominal case, with a frequency of 250 Hz, a current amplitude of 40 mA, and a pulse width of 500 μs. A load of 1 kΩ was used in this case [4], [12]–[14]. This image was obtained from the simulation in the PSIM software [10], with a time step of 5 ns.

To verify the maximum current limit, Fig. 9 shows the stimulation current with an amplitude of 100 mA, a frequency of 250 Hz, and a pulse width of 500 μs. The same load and

TABLE III
AVERAGE VALUE OF THE STIMULATION CURRENT FOR THE CRITICAL CASES.

Case	Frequency (Hz)	Pulse width (μs)	Current (mA)	
			3 mA	100 mA
1	10	10	3.2761	90.1182
2	10	500	3.1728	98.6337
3	250	10	3.2741	89.9502
4	250	500	3.1772	98.6312

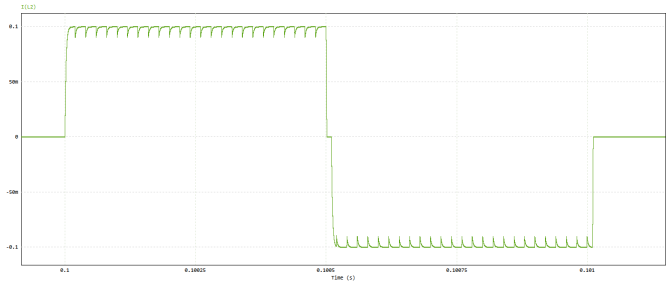


Fig. 7. Biphasic, symmetric stimulation current with an amplitude of 100 mA.

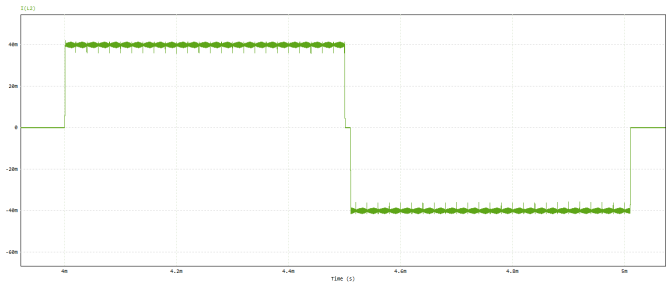


Fig. 8. Biphasic, symmetric stimulation current with an amplitude of 40 mA.

simulation step were used. The reference amplitude used in the test was 1 A, which demonstrates the limitation imposed by the buck converter, according to the values defined in the design.

B. Variation of the Output Inductance

The analysis of the extreme frequency and pulse width values showed that switching the inductance values is necessary for different values of current amplitude and pulse width. Therefore, the possible inductance values were analyzed, according to the inductor combination, for a stimulation signal with a frequency of 10 Hz, a current amplitude of 100 mA, and a pulse width of 500 μs. These results are shown in Table IV, which presents the average value of the positive half-cycle of the stimulation current.

C. Variation of the Circuit Load

Next, a test was performed to verify the correct operation of the current source implemented in the simulation. This test consists of varying the resistive load in series with the output

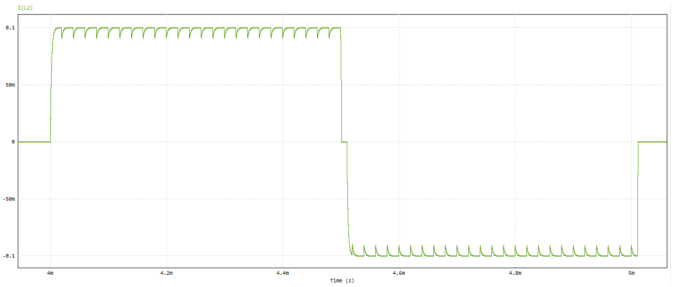


Fig. 9. Stimulation current limitation for a 1 A reference amplitude.

TABLE IV
CURRENT AMPLITUDE WITH A REFERENCE OF 100 MA FOR DIFFERENT INDUCTANCE VALUES.

Inductance (mH)	Average value (mA)
1	98.8270
2	98.6337
4	98.2380
5	98.0393

inductor across 1 kΩ, 800 Ω, and 500 Ω [4], [13], [14]. The goal is to observe the variation of the stimulation current with a reference amplitude of 100 mA, a frequency of 250 Hz, and a pulse width of 500 μs. The results are shown in Table V.

D. ABNT NBR IEC 60601-2-10 Standard

The ABNT NBR IEC 60601-2-10 standard [14] was consulted as a reference for the definition of the design parameters. The current amplitude limit per stimulation frequency and the peak voltage in the circuit were verified. The standard requires an accuracy of 20% for the stimulation parameter limits, limiting the stimulation current amplitude range to 3–100 mA. These values are achievable across the entire available pulse width range.

E. Comparison with Commercial Device

To compare the presented waveform with commercial equipment, a measurement was performed with the FesVif 995 Four device (Quark Medical, Piracicaba, Brazil). This device was chosen for its wide use in the Brazilian clinical market and for being available for testing in the Electrothermotherapy course of the Physical Therapy Department of the Federal University of Minas Gerais (UFMG).

The goal of this comparison is only to present a stimulation current available in the Brazilian market and to propose the developed circuit as a new alternative for this type of application.

The measurement was performed with a stimulation frequency of 50 Hz, a current amplitude of 10 mA, and a pulse width of 150 μs. Fig. 10 shows the voltage waveform obtained across the 1 kΩ load during the test.

The device has a biphasic, asymmetric waveform, with a positive half-cycle similar to that of the proposed FES circuit. The stimulation current is obtained by dividing the voltage by the load resistance used. Analyzing only the positive half-cycle, one observes a significant current ripple, current overshoot, and pulse width and amplitude values close to the desired ones.

TABLE V
COMPARISON FOR DIFFERENT LOAD VALUES (CURRENT OF 100 MA).

Load (Ω)	Average value (mA)
1000	98.6312
800	99.4956
500	99.7030

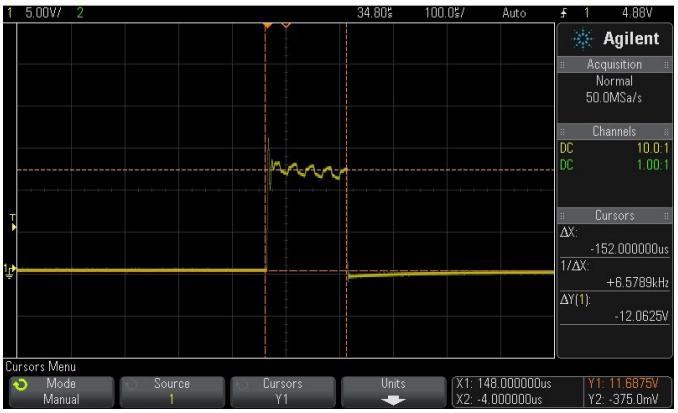


Fig. 10. Measurement of the stimulation signal from the FesVif 995 Four equipment.

IV. Conclusion

The simulated FES circuit can synthesize a biphasic, symmetric, square stimulation current with an amplitude of 3–100 mA, a resolution of 1 mA, a frequency of 10–250 Hz, and a pulse width of 10–500 μs. The maximum error observed for the parameters is within the 20% limit specified by the ABNT NBR IEC 60601-2-10 standard.

The simulated circuit exhibits significant output current errors for shorter pulse widths, related to the settling time required for the stimulation signal. In addition, the efficiency of the circuit is degraded by the fixed compliance voltage of the boost converter [6].

In future work, the FES device will be physically implemented, adding isolation stages for user protection by using high-speed digital optocouplers between the signal-generating circuit and the electrodes [13].

The switch to be used in the physical implementation of the hybrid converter in the prototype will be the GaN power transistor EPC2215 (Efficient Power Conversion Corporation, El Segundo, CA, USA), which allows high-frequency switching. A half-bridge circuit was simulated in LTspice software [15] using the design parameters, with the SPICE model of the switch provided by the manufacturer. The component presented a rise time of 1 ns and a recovery time of 3 μs, confirming its suitability for the design.

In the circuit developed by Blanco et al. [4], the stimulation current flows through the H-bridge and Wilson current mirror circuits, with the three transistors of the current mirror operating in the linear region, which is associated with conduction losses. In the proposed switched-mode topology, the stimulation current also flows through the H-bridge circuit, but it is modulated in the output inductor. The only losses in this model are the switching losses, which are very low in GaN switches and much smaller than the conduction losses of the Wilson current mirror.

With the prototype built, the efficiency study of the device can be carried out, in order to reaffirm the choice of the switched-mode topology for the output stage. In addition, tests with resistive loads will be performed with the goal of demonstrating the correct operation of the developed current source, before applying the stimulation signal to volunteers [4].

Acknowledgment

The authors used the artificial intelligence (AI) assistant Claude (Anthropic) for translation from Portuguese to English, language editing, preparation of the diagrams, and LaTeX formatting. The conception, data, analyses, simulation figures, and writing of the work are the sole responsibility of the authors.

References

[1] T. A. Thrasher and M. R. Popovic, “Functional electrical stimulation of walking: Function, exercise and rehabilitation,” Ann. Réadapt. Méd. Phys., vol. 51, no. 6, pp. 452–460, Jul. 2008.

[2] C. L. Lynch and M. R. Popovic, “Functional electrical stimulation,” IEEE Control Syst. Mag., vol. 28, no. 2, pp. 40–50, Apr. 2008.

[3] D. C. Souza, M. C. Gaiotto, G. N. Nogueira-Neto, M. C. F. Castro, and P. Nohama, “Power amplifier circuits for functional electrical stimulation systems,” Res. Biomed. Eng., vol. 33, no. 2, pp. 144–155, 2017.

[4] J. E. Blanco, F. R. Castro, L. H. V. Felão, M. F. R. Urban, R. Gaino, M. A. A. Sanches, and A. A. Carvalho, “Estimulador elétrico funcional utilizando ponte H no estágio de potência,” in Anais do XXV Congresso Brasileiro de Engenharia Biomédica (CBEB), Foz do Iguaçu, PR, Brazil, 2016, pp. 603–606.

[5] M. H. Rashid, Ed., Power Electronics Handbook. San Diego, CA, USA: Academic Press, 2001.

[6] M. N. van Dongen and W. A. Serdijn, “A switched-mode multichannel neural stimulator with a minimum number of external components,” in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), 2013, pp. 1877–1880.

[7] S. C. Huerta, M. Tarulli, A. Prodić, M. R. Popović, and P. W. Lehn, “A universal functional electrical stimulator based on merged flyback-SC circuit,” in Proc. 15th Int. Power Electron. Motion Control Conf. (EPE-PEMC ECCE Europe), Novi Sad, Serbia, 2012, Art. no. LS5a.3.

[8] N. Mohan, T. M. Undeland, and W. P. Robbins, Power Electronics: Converters, Applications, and Design, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2003.

[9] K. Ogata, Modern Control Engineering, 5th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[10] PSIM Professional, version 2025.1.0. Troy, MI, USA: Altair Engineering Inc., 2025.

[11] R. G. Barelli, V. F. Avelino, and M. C. F. Castro, “STIMGRASP: A home-based functional electrical stimulator for grasp restoration in daily activities,” Sensors, vol. 23, no. 1, Art. no. 10, 2023.

[12] V. J. Bolfe, S. I. Ribas, M. I. L. Montebelo, and R. R. J. Guirro, “Comportamento da impedância elétrica dos tecidos biológicos durante estimulação elétrica transcutânea,” Rev. Bras. Fisioter., vol. 11, no. 2, pp. 153–159, 2007.

[13] M. C. C. Gaiotto, G. N. Nogueira-Neto, E. Krueger, C. E. F. Magnani, and P. Nohama, “Uma contribuição ao projeto de saída de estimuladores elétricos neuromusculares bifásicos,” in Anais do XXIII Congresso Brasileiro de Engenharia Biomédica (CBEB), Porto de Galinhas, PE, Brazil, 2012, pp. 670–674.

[14] Equipamento eletromédico — Parte 2-10: Requisitos particulares para a segurança básica e o desempenho essencial de estimuladores de nervos e músculos, ABNT NBR IEC 60601-2-10:2014, 2014.

[15] LTspice, version 26.0.1. Wilmington, MA, USA: Analog Devices Inc., 2026.

Development of a Digital Algometer for Application in Musculoskeletal Interfiber Counterirritant Stimulation (MICS)

1st Andé Fabiano Alves Alflen
Department of Electronic Engineering
Universidade Tecnológica Federal do Paraná - UTFPR
Campo Mourão, PR, Brazil.
0009-0009-3896-4618

2nd Leandro Procópio Alves
Department of Biomedical Engineering
Universidade Anhembi Morumbi – UAM
São José dos Campos, SP, Brazil.
0000-0002-2880-0507

3rd Livia Helena Moreira
Department of Biomedical Engineering
Universidade Anhembi Morumbi – UAM
São José dos Campos, SP, Brazil.
0000-0001-6432-7596

4th Mariano Rocabado
Department of Rehabilitation
Universidad Andrés Bello – UNAB
Santiago, RM, Chile.
0000-0003-3898-030X

5th Rodrigo Alexis Lazo Osório
Department of Biomedical Engineering
Centro de Inovação, Tecnologia e Educação – CITÉ
São José dos Campos, SP, Brazil.
0000-0002-8317-3219

6th Henrique Cunha Carvalho
Department of Electronic Engineering
Universidade Tecnológica Federal do Paraná - UTFPR
Campo Mourão, PR, Brazil.
0000-0003-0806-4132

***Abstract*— Pain is a subjective experience that varies from individual to individual and has a direct impact on quality of life. This study describes the development of a portable digital algometer to standardize pressure application during the Musculoskeletal Interfiber Counterirritant Stimulation (MICS) technique. Given the subjective nature of pain and the variability in trigger-point sensitivity, the device aims to provide an objective, reliable measurement method. The system integrates a strain gauge into a quarter-Wheatstone bridge, an HX711 amplifier with 24-bit resolution, a PIC16F874A microcontroller and an LCD interface for real-time feedback. Calibration with standard weights showed a strong linear relationship between mass and voltage (R^2 = 0.98734), confirming accurate (mean 89%), repeatable measurements within the elastic region. The device enables precise quantification of applied pressure, supporting treatment consistency and potential home use. Future work will focus on usability validation and miniaturization to improve portability and scalability.**

Keywords— algometer, MICS, pain, strain gauge.

I. INTRODUCTION

Pain is a response to actual or potential tissue damage [1] and is one of the most important clinical symptoms, a subjective experience influenced by physical and psychological factors [2]. Therefore, it is necessary the development of methods to evaluate it, since pain is considered an important clinical symptom [3].

In this context, Myofascial Pain Syndrome (MPS) is the most common form of musculoskeletal pain. This condition can reduce tissue elasticity and decrease an individual's mobility. These symptoms occur due to Trigger Points (TPs), which generate referred pain, that is, pain perceived in a location different from its point of origin [4]. In addition to pain, these TPs may also cause sensations of cold or heat in the affected region [5]. The treatment of MPS focuses on pain control with analgesics and anti-inflammatory drugs, and on deactivating TPs, which can be achieved through invasive and non-invasive methods, such as needling and ischemic compression [6].

In this context, the innovative Musculoskeletal Interfiber Counterirritant Stimulation (MICS) technique emerges as a proposed non-invasive treatment. By using the MICS to apply ischemic pressure to the TP, modulation of the autonomic nervous system is induced, resulting in reduced MPS pain in patients with joint hypermobility [7].

Although several approaches exist for treating MPS, it is essential to establish a standardized pain measurement scale to improve efficiency. In this regard, several methods are available to assess pain, including visual, verbal, and behavioral scales, as well as algometers [2].

Algometers offer an advantage over other measurement methods by providing a more precise scale with less subjectivity in patient pain perception [2]. They are used to measure the pressure applied at the moment pain is perceived, providing an objective measure of pain sensitivity [8]. This is particularly useful for evaluating MPS, since TPs may vary in sensitivity. With algometers, healthcare professionals can quantify pain more precisely, consistently, and in a standardized manner, which is essential for monitoring the progression of the pain condition. This, this study aimed the development of a portable digital algometer to standardize pressure application using the MICS technique.

II. METHODOGOLY

The digital algometer developed was based on the component strain gauge. When subjected to pressure or weight, this sensor deforms, changing its electrical resistance. This effect can potentially allow the innovative quantification of the pressure applied at the tip of the MICS. When bonded to MICS, the strain gauge can stretch or compress, altering its internal resistance, which is measured to determine stress.

To measure the variation in applied mechanical pressure, a Wheatstone bridge configured as a quarter-bridge (1/4) was used. In this configuration, the resistance variation is converted into a voltage variation of about millivolts (mV).

Afterward, the generated signal must be amplified, since the voltage variation is very small. Some microcontrollers (MCUs) operate with supply voltages of 5 V or 3.3 V and include 10-bit or 12-bit ADC converters, resulting in resolutions of approximately 4.88 mV and 0.8 mV, respectively. Therefore, signal amplification was required.

The block diagram in Fig. 1 illustrates the system's overall process, its relationships, and the operating flow.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

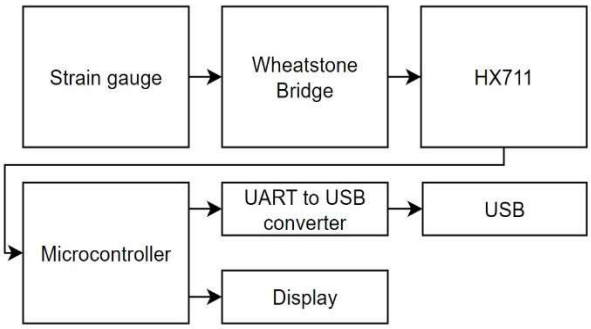


Fig. 1. Block Diagram of the proposed digital algometer.

A. Strain gauge

The precision resistance strain gauge used was a commercial model, BX120-3AA (Ningbo Ane Electric Co., Ltd., Ningbo, China). It was configured in a quarter-bridge (1/4), which, although not completely immune to temperature-related variations, compensates for them using a copper–aluminum alloy.

B. Amplifier/Analog-Digital-Converter

The amplifier used was the HX711 module (Avia Semiconductor, Xiamen, China), which, after signal amplification, performs analog-to-digital (A/D) conversion with 24-bit resolution, however, the effective resolution is lower in practice due to the influence of electronic noise and other non-idealities inherent to the measurement system. Still, it is significantly higher precision than that available in conventional MCUs. The HX711 gain was set to 128, and the clock to 10 Hz.

C. Microcontroller

The selected microcontroller (MCU) was the PIC16F874A, a CMOS flash 8/14-bit family member featuring 14-bit instruction words and a Harvard architecture. It was chosen for its cost-effectiveness (low power and high performance), the adequate number of GPIO (General Purpose Input/Output) pins, and support for UART and I²C communication protocols. This MCU has 192 bytes of SRAM and 128 bytes of EEPROM, which is not a limitation for the chosen display or for reading data from the HX711 module, since each measurement can be stored using 3 bytes.

D. User Interface

A blue 16-character × 2-line Liquid Crystal Display (LCD) module with the HD44780 controller was used. Communication was performed using the 4-bit protocol, reducing the number of required MCU ports. Through this display, the pressure value can be visually presented to the algometer user, allowing the patient undergoing the MICS treatment to use the device to train the standard mechanical pressure application pattern for therapy at home [7].

The digital algometer includes three buttons: one to start and stop the measurement, another to change the pressure scale, and a third to transmit the sampled data via USB at 9600 baud.

E. Power Supply

The power supply is provided by a 9-V Li-Ion battery, with the voltage regulated to 5 V by the LM7805 positive voltage regulator. The choice of a 9 V battery was made due to its practicality and ease of replacement. The 100 µF capacitors filter voltage fluctuations at the output of the voltage regulator.

F. Calibration

To calibrate the strain gauge, a precision digital balance with 1 decimal place resolution was used. The MICS device was positioned on a tiny silicone rubber base to ensure stability. Next, traceable weights (standards) were placed on the device, and their weights were measured using the digital scale. The calibration was performed using three measurements (triplicate), with the mean used to improve accuracy.

With the circuit's resolution and supply voltage, it was possible to relate the A/D converter's output (mV) to reference electrical voltages measured with a digital multimeter, thereby enabling device calibration.

To establish the device's transfer function, a linear regression was applied to the calibration data. The relationship between the applied mass and the resulting electrical voltage is described by the following linear model (Equation 1):

$$V(mV) = S \cdot m(g) + V_0 \quad (1)$$

Where:

- $V(mV)$ is the output voltage measured in millivolts;
- S represents the sensitivity of the system (slope of the calibration curve);
- $m(g)$ is the applied mass in grams;
- V_0 is the intercept, corresponding to the voltage offset at zero load.

The sensitivity (S) was calculated as the ratio of the change in voltage to the change in mass ($\Delta V/\Delta m$). This mathematical model was subsequently implemented in the PIC16F874A firmware to convert the 24-bit digital signal from the HX711 module into real-time mass or pressure measurements, readings displayed on the LCD. The system resolution was determined by correlating the sensitivity with the A/D converter's 24-bit depth.

Pressure values were obtained by converting the reference masses into force using the standard gravitational acceleration (9.81 m/s²). Assuming a contact area of 1 cm² at the device tip, the applied pressure was calculated from the force-to-area relationship. Subsequently, the same calibration and regression procedures previously described were applied to establish the pressure calibration curve. This linear model is described by the following linear model (Equation 2):

$$V(mV) = Sa \cdot P(kPa) + V_1 \quad (2)$$

Where:

- $V(mV)$ is the output voltage measured in millivolts;
- S represents the sensitivity of the system (slope of the calibration curve);
- $P(kPa)$ is the applied pressure in kilopascal;
- V_1 is the intercept, corresponding to the voltage offset at zero load.

G. Softwares

For the development, the electronic schematic design was carried out using KiCad v6.00 (KiCad EDA, GitLab Community, Global), while SimulIDE v1.0.0 (SimulIDE Community, Global) was used for circuit simulation. The source code was written and compiled in MPLAB X IDE

v6.00 (Microchip Technology Inc., Chandler, AZ, USA). Finally, the MCU was programmed using the PICkit 3 v3.01 (Microchip Technology Inc., Chandler, AZ, USA). The data analysis and graphs were created using OriginPro software (v. 8.5, OriginLab, Northampton, MA, USA).

III. RESULTS AND DISCUSSION

The MCU circuit was designed to allow external connections to the LCD and a UART–USB converter, support communication with the HX711, and provide an I²C interface for future use. Fig. 2 shows the MCU circuit.

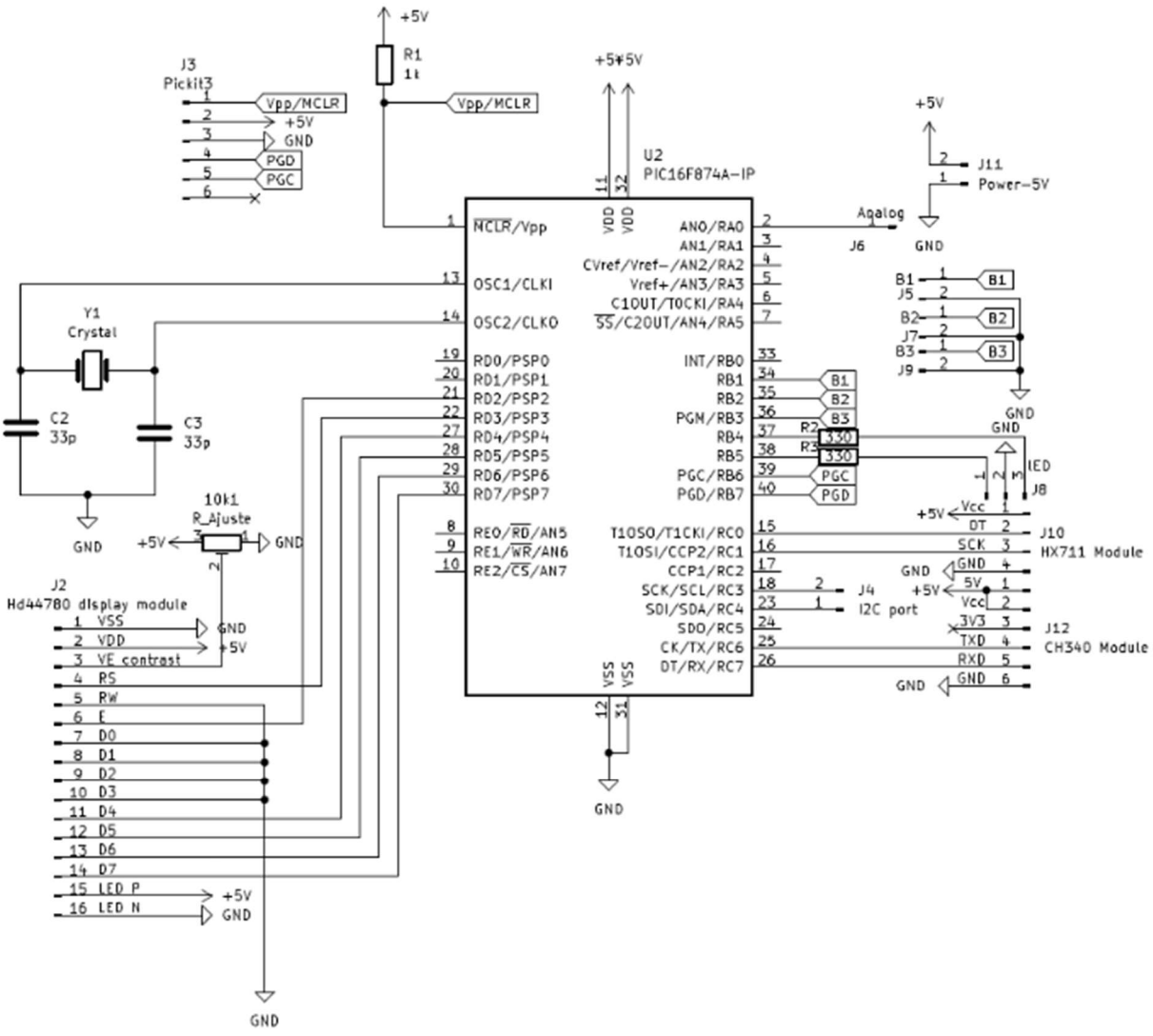


Fig. 2. MCU circuit of the digital algometer.

The pressure measurement circuit schematic was developed separately from the MCU to facilitate a possible strain gauge replacement. Fig. 3 presents this schematic circuit.

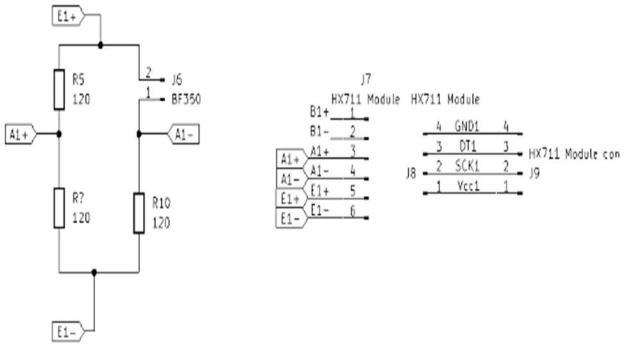


Fig. 3. Circuit HX711 and Wheaston bridge.

Based on the schematics shown in Fig. 2 and using the LM7805 power supply, the Printed Circuit Board (PCB) layouts for the MCU were developed. The jumpers were implemented as traces on the top layer (red lines); however, the PCB was designed to operate as a single-layer board. Fig. 4 shows the MCU PCB layout.

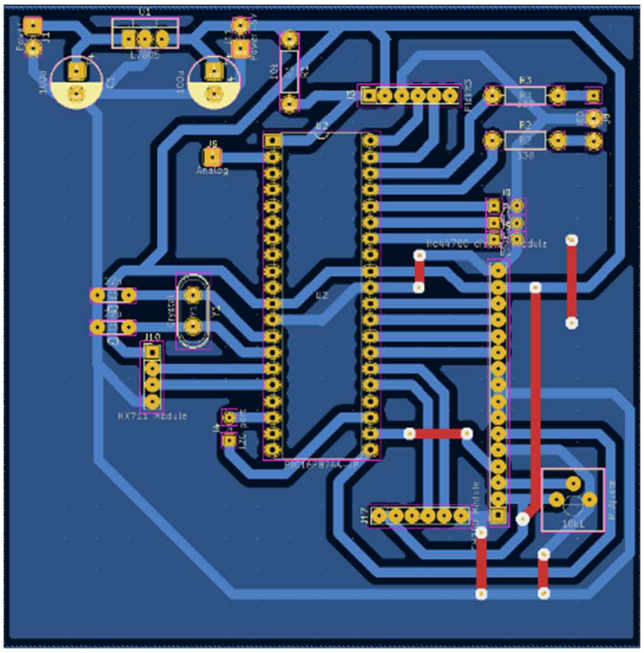


Fig. 4. PCB layout of the MCU.

The PCB layouts for the HX711 and the Wheatstone bridge were developed based on the schematics shown in Fig. 3. The single-layer PCB layout is shown in Fig. 5.

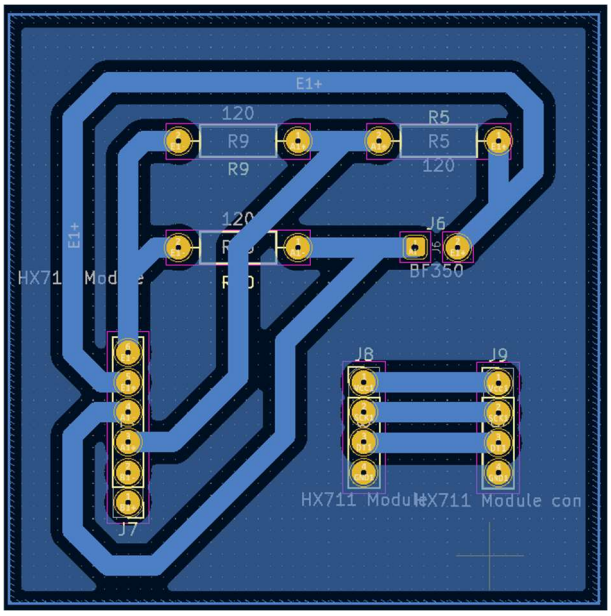


Fig. 5. Circuit HX711 and Wheaston bridge electronic layout.

Figures 6 and 7 show, respectively, the PCB phenolic prototypes developed, which comprise the MCU, and the HX711 and Wheatstone bridge.

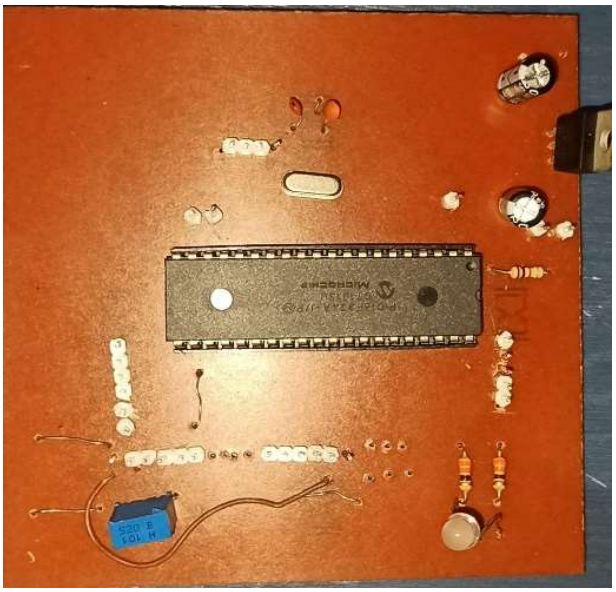


Fig. 6. PCB with a 10 x 10 cm prototype of the MCU.

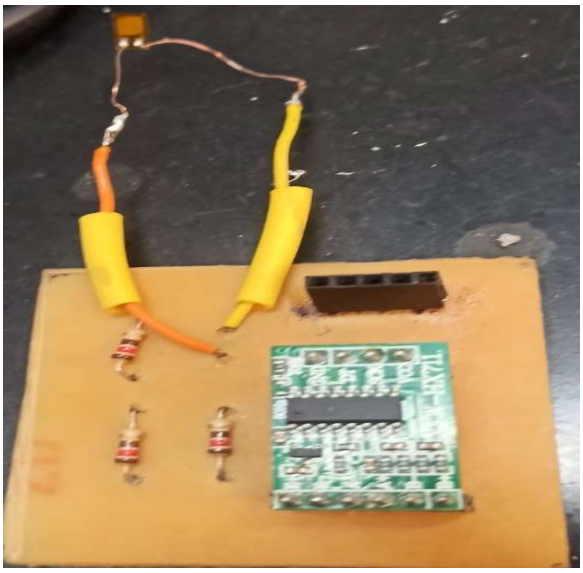


Fig. 7. HX711 and Wheatstone PCB with 5x5 cm prototype connected to the BX120-3AA strain gauge (upper).

The experimental data, f-function and the resulting regression fit are presented in Fig. 8. The calibration curve exhibited high linearity across the clinical range of 100 g to 1000 g, with a coefficient of determination (R^2) of 0.98734. The calculated sensitivity (S) was 0.763 $\mu V/g$. The maximum mass was 1000 g due to the MICS clinical protocol. The low dispersion across triplicates (mean ± standard deviation) indicates high repeatability and confirms that the BX120-3AA strain gauge operated within its elastic region, with no evidence of hysteresis or nonlinearity. This linear response is critical for the MICS protocol, as it ensures that the pressure applied to the trigger point is quantified with precision. The largest absolute error was 61.64 g for a mass of 900 g, and the mean accuracy was 89%.

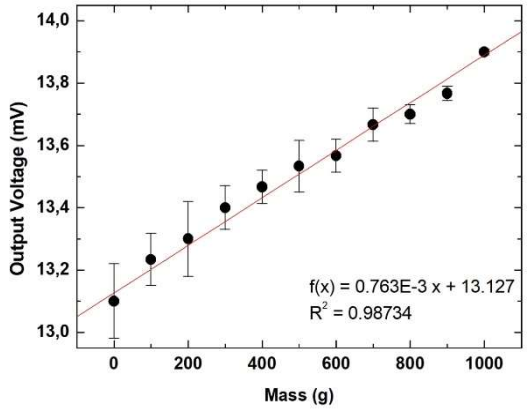


Fig. 8. Experimental data with a regression fit of the strain gauge calibration.

To comply with SI units, the applied loads were converted into pressure values assuming a contact area of 1 cm² at the algometer tip. Consequently, the calibration range corresponded to 0–98.1 kPa, which encompasses the operating range required by the MICS protocol, shown in Fig. 9.

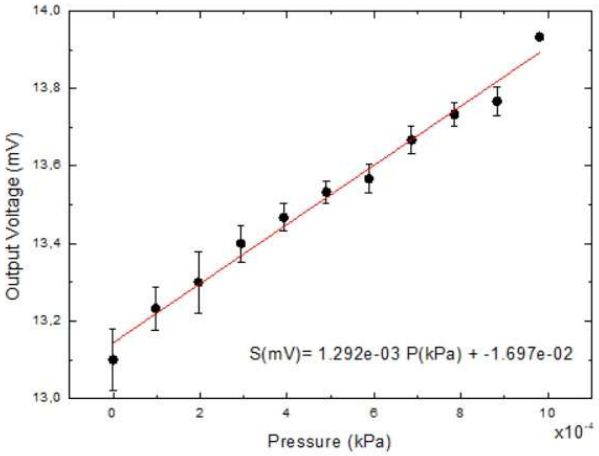


Fig. 9. Strain gauge calibration curve showing the experimental data and the corresponding regression fit of pressure (kPa).

Therefore, the developed digital algometer shown in Fig. 9 was calibrated, and the LCD displays a readout in grams (bottom right). Also, in Fig. 10 a standard weight of 100 g was used (bottom left), and the MICS was positioned in this sample holder. To obtain the signal reading, a linear regression model was applied to the experimental data, relating the output voltage to the applied mass. The average measurement error was 0.015635mV, with a maximum deviation of 0.050600 mV, resulting in a mean relative error of 4.2640%. The system demonstrated a precision of approximately 0.295402 mV.

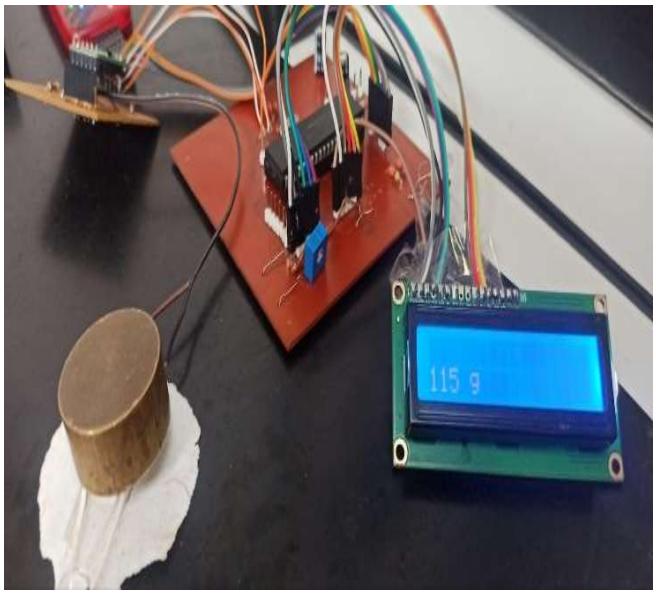


Fig. 10. The digital algometer with the MICS device.

Several studies proposed the development of algometers and used components from this work, such as strain gauges [9] and the HX711 module [9,10], as well as the Wheatstone bridge topology [9].

The differential of this innovative digital algometer compared to other referenced works is its direct focus on the MICS protocol and its use of a MCU from the PIC16 family, rather than the Arduino environment, as in Gimeno (2020)

[9]. An electronic circuit using a more advanced development environment, MPLAB X IDE, provides greater control of the hardware, allowing better optimization and, consequently, the use of a simpler and lower-cost MCU than the ATmega328p (Microchip Technology Inc., Chandler, AZ, USA), as applied in the commercial Arduino Uno R3 and V3.

For the next steps of the project, it will be necessary to conduct intra- and inter-rater analyses to validate the digital algometer's usability. This further analysis also requires ethics & research committee evaluation for tests on humans aiming for usability.

Another important improvement would be a version using SMD (Surface-Mount Device) components, which would significantly reduce its size. This miniaturization using SMD components would offer advantages, such as reduced circuit footprint, improved high-frequency performance, and easier large-scale manufacturing. Other improvements include Bland–Altman agreement assessment, intraclass correlation coefficient (ICC), thermal drift analysis, uncertainty propagation, long-term stability evaluation, and direct comparison with commercially validated algometers.

CONCLUSION

With the development of a portable digital algometer, the possibility arises to standardize the pressure applied to TPs, including in the home care context, which aligns with the MICS treatment approach. This standardization may be a crucial factor in the treatment's effectiveness and in minimizing patient discomfort.

Furthermore, the possibility of self-treatment with the proposed digital algometer may provide patients with greater convenience and autonomy. Future developments will include loading–unloading sensor comparison, redesigning the device using surface-mount technology (SMT/SMD), and enabling a more compact, robust, and user-friendly implementation. However, its clinical applicability, usability, and therapeutic effectiveness still require validation through human-centered studies. In addition, inter-rater and intra-rater reliability assessments must be conducted to evaluate the consistency and reproducibility of the measurements obtained with the device across different evaluators and repeated assessments.

ACKNOWLEDGMENT

The authors would like to thank Federal Technological University of Paraná (UTFPR) for providing all resources for the project development, the National Council for Scientific and Technological Research (CNPq) for a scholarship to A. F. A. A., the Ânima Institute (AI) for the support to L. P. A. and L. H. M., and the Fundação Araucária (FA) & Superintendência de Ciência, Tecnologia e Ensino Superior (SETI) to the conference grants to H. C. C.

REFERENCES

[1] V. F. de Carvalho, M. C. Ferreira, S. A. T. Vieira, and T. Ueda, "Limiar de sensibilidade cutânea dos pés em pacientes diabéticos através do pressure specified sensory device: uma avaliação da neuropatia," *Revista da Associação Médica Brasileira*, vol. 55, pp. 29–34, 2009, doi: https://doi.org/10.1590/S0104-42302009000100011.J

[2] L. Lorusso *et al.*, "Autoalgometry: An important tool for pressure pain threshold evaluation," *Journal of Clinical Medicine*, vol. 7, no. 9, p. 273, 2018, doi: https://doi.org/10.3390/jcm7090273.

[3] W. Mota, "Algômetro por estimulação elétrica transcutânea," Campinas, Universidade Estadual de Campinas – UNICAMP, Faculdade de Engenharia Elétrica e de Computação, 2021.

[4] J. S. Batista, A. M. Borges, and L. M. Wibelinger, "Tratamento fisioterapêutico na síndrome da dor miofascial e fibromialgia," *Revista Dor*, vol. 13, no. 2, pp. 170–174, 2012, doi: https://doi.org/10.1590/s1806-00132012000200014.R.

[5] R. Silva, "Avaliação da eficácia do agulhamento de pontos-gatilho miofasciais (seco, 0,5% e 1% de lidocaína) em pacientes portadores de dor miofascial na musculatura mastigatória," Bauru, Universidade de São Paulo – USP, Faculdade de Odontologia de Bauru, 2007.

[6] M. Culpi and A. B. M. C. Martinell, "Desativação de pontos-gatilho no tratamento da dor miofascial," *Revista Brasileira de Medicina de Família e Comunidade*, vol. 13, no. 40, pp. 1–9, 2018, doi: https://doi.org/10.5712/rbmfc13(40)1777

[7] H. C. Carvalho *et al.,* "Autonomic nerve regulation in joint hypermobility patients with myofascial trigger points by Musculoskeletal Interfiber Counterirritant Stimulation (MICS)," *Medical Engineering & Physics*, vol. 109, pp. 103903–103903. 2022, doi: https://doi.org/10.1016/j.medengphy.2022.103903.

[8] M. Hill, "Design, development, and verification of an automated pain measurement device," Boulder, University of Colorado – CU, 2013.

[9] S. M. Gimeno, "Diseño y realización de un dispositivo electrónico para la medida del umbral de dolor en pacientes con lesiones musculares," València, Universidade Politècnica de València – UPV, Escuela Técnica Superior de Ingeniería del Diseño, 2020.

[10] C. G. Nim *et al.,* "Pressure pain thresholds in a real-world chiropractic setting: topography, changes after treatment, and clinical relevance?," *Chiropractic & Manual Therapies*, vol. 30, no. 1, pp. 25, 2022, doi: https://doi.org/10.1186/s12998-022-00436-2

Development of a Dynamic Electro-Pneumatic Actuator for MRE Applications

1st Akihiro Souza Seido
Graduate Program in Electrical and Computing Engineering
Federal University of Goiás
Goiânia, Goiás 74605-010, Brazil
akihiro@discente.ufg.br

2nd Jeveson Cardozo da Silva
Institute of Physics
Federal University of Goiás
Goiânia, Goiás 74690-900, Brazil
jeveson@ufg.br

3rd Juracy Leandro dos Santos Junior
Graduate Program in Electrical and Computing Engineering/Institute of Physics
Federal University of Goiás
Goiânia, Goiás 74690-900, Brazil
juracyleandro@ufg.br

4th Sílvio Leão Vieira
Institute of Physics
Federal University of Goiás
Goiânia, Goiás 74690-900, Brazil
slvieira@ufg.br

***Abstract*—This work presents the development of a novel microcontroller-based dynamic electro-pneumatic actuator for Magnetic Resonance Elastography (MRE) to overcome the frequency and phase limitations of current clinical systems. The proposed actuator integrates a refrigeration compressor, a cylindrical reservoir, and an electroacoustic modulator for high-speed airflow switching. A 3D-printed pneumatic driver with a latex membrane converts air pulses into mechanical waves, reaching up to 500 Hz in water (electroacoustic modulator) and validated in a paraffin gel phantom over 10-200 Hz. Experimental characterization included cylinder loading, duty cycle, static force, pressure-displacement, and wave propagation in a paraffin gel phantom. Results show a linear force-pressure response (R^2 = 0.9998), frequency-dependent viscoelastic dispersion (shear modulus from 3.66 to 22.44 kPa over 10–200 Hz), and modulator sensitivity with a nonlinear frequency-PWM relationship. The system achieved stable operation up to 20 L/min, with a duty-cycle margin and a displacement sensitivity of approximately 0.19 mm/kPa within the recommended 40–50 kPa range. A compact, mobile enclosure with electromagnetic shielding and a touchscreen interface was developed with future MRI integration in mind, though in-scanner compatibility remains unestablished. The actuator generates controlled shear waves in tissue-mimicking media, offering a versatile, low-cost benchtop platform to advance MRE research toward higher-frequency clinical applications.**

***Index Terms*—electro-pneumatic actuator, magnetic resonance elastography (MRE), microcontroller-based control, shear-wave generation, viscoelastic properties.**

I. Introduction

Manual palpation has served as a fundamental diagnostic tool in medical practice for centuries. The ability to distinguish healthy from diseased tissue based on tactile assessment of stiffness reflects significant changes in tissue mechanical properties associated with pathological processes such as inflammation, fibrosis, neoplasms [1] and gel dosimetry [4]. Malignant tumors, for example, are typically stiffer than the surrounding tissue. However, palpation is limited to superficial organs, is inherently subjective, and depends on the examiner's experience. To address these limitations, this work proposes a novel electropneumatic actuator that integrates the advantages of compressed air with the precision of electronic control.

Conventional imaging modalities (CT, ultrasound, MRI) do not directly quantify tissue mechanical properties: their contrast relies on electron density, acoustic impedance, or magnetic relaxation, which vary on much smaller scales than the elastic modulus [2], while CT and MRI parameters vary by factors of two to three, the shear modulus of soft tissues can differ by more than five orders of magnitude [1], [3].

To address this, elastography has emerged over the past three decades as an imaging technique that quantitatively maps tissue elasticity, complementing US, MRI, and CT interpretation [5]. Among these, Magnetic Resonance Elastography (MRE), introduced by Muthupillai et al. in 1995, enables "virtual palpation" of deep-seated organs with high spatial resolution [6] by propagating harmonic shear waves whose wavelength and damping relate directly to viscoelastic parameters such as shear modulus and viscosity [7].

To achieve this quantification, a mechanical actuation system is necessary to generate these waves in a controlled manner, to synchronize them with the gradient pulses of the MRI scanner, and to ensure compatibility with the high-magnetic field environment [8]. Historically, three main actuator families have been explored: electromagnetic, piezoelectric, and pneumatic [9].

Electromagnetic actuators are capable of generating significant forces; however, they are prone to electromagnetic interference with imaging gradients, produce magnetic susceptibility artifacts, and present safety risks due to conductive metallic structures exposed to intense oscillating fields [10].

Piezoelectric actuators are inherently MRI-compatible and offer high temporal precision and a broad frequency range (0-300 Hz). However, they are limited to small displacements, rarely exceeding 1 mm, and often possess a large footprint, complicating their integration and use in clinical applications [11], [12].

Pneumatic systems offer the simplest and safest compatibility, using air displacement to generate vibration, thereby eliminating the need for electronic components near the patient [8]. Nevertheless, these systems face persistent challenges, including phase delay and a frequency range typically limited to less than 100 Hz [12], [13].

This comparative analysis indicates that the fundamental challenges identified decades ago remain largely unresolved: the only MRE actuator approved for clinical use, the Active Driver^{TM} of the Resoundant system (Rochester, MN, USA), used with GE Healthcare scanners [12], is effective and safe but limited to 40-90 Hz, restricting detailed characterization of stiff tissues or small anatomical structures that require higher frequencies [14].

This scenario reflects both technical and economic barriers, including regulatory approval costs, that keep versatile alternatives such as piezoelectric actuators confined to academic laboratories, while an older, more limited technology remains the only clinically available option — highlighting the need for a solution that directly addresses these unresolved frequency and phase limitations.

This work seeks to combine the safety, compatibility, and flexibility of novel pneumatic systems with the ability to operate in a wide frequency range (10 to 500 Hz), thus directly addressing the phase and frequency limitations of current systems. Our research encompasses the entire process, from mechanical design and 3D component printing, including material selection and flow parameter optimization, to the development of control electronics, microcontroller programming and bench testing measurement.

II. Materials

Here, we describe the materials, components, and methods used to develop the microcontroller-based dynamic electropneumatic system, following a functional hierarchy that details first the pneumatic plant instrumentation and then the control and automation system, to support understanding and reproducibility.

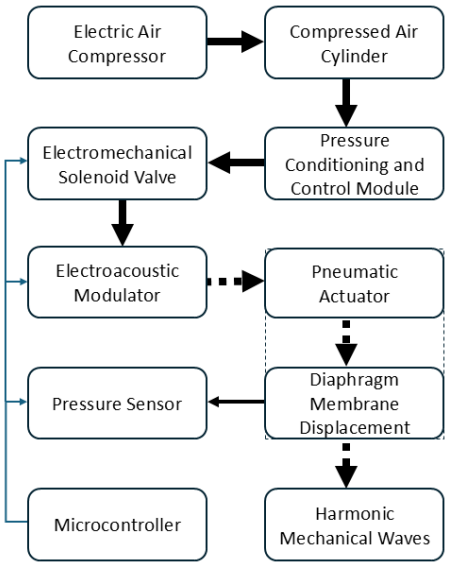
The electropneumatic system consists of an integrated set of devices operating in a cascade to generate, condition, modulate, and apply compressed-air flow as mechanical pressure waves. Figure 1 illustrates the functional block diagram of the system, showing airflow and interconnection between its main modules. Each of these blocks is detailed in the following subsections, with an emphasis on technical characteristics, system function, and selection criteria.

A. Compressed Air Generation and Storage Module

The compressed air generation and storage module captures ambient air, compresses it, and stores it for later use as the system's pneumatic energy source.

1) Electric Air Compressor: The active element of this module is a hermetic reciprocating compressor originally designed for refrigeration by Embraco (Joinville, SC, Brazil), chosen over conventional workshop models for its low noise and vibration — essential in clinical or lab settings. It features a 1/4 HP motor (186 W), operates on 110-220 VAC (50/60 Hz) with a nominal current of 1-4 A, and reduces air volume within a cylinder to increase pressure: the suction (low-pressure) side operates between 0-15 PSI (0-103 kPa), while the discharge (high-pressure) side reaches 100-150 PSI (0.69-1.03 MPa), depending on compression ratio and ambient temperature.

Fig. 1: Functional block diagram of the microcontroller-based dynamic electropneumatic system.



2) Compressed Air Cylinder: The compressed air is stored in a carbon steel cylinder originally designed for refrigerant gas storage (EOS brand, RS, Brazil). The reservoir has a dual purpose: damping pulsations generated by the compressor and providing air volume to meet peak demands without requiring continuous compressor operation. Its internal geometric capacity is 12 liters. The maximum working pressure adopted for the system is 150 PSI (approximately 10.3 bar or 1.03 MPa). When pressurized to 10 bar, the reservoir stores the equivalent of 120 liters of air at standard atmospheric pressure (1 bar), according to Boyle-Mariotte's Law ($P_1V_1 = P_2V_2$), thereby ensuring sufficient capacity for the tests.

B. Pressure Conditioning and Control Module

Before use, the stored air must be reduced to a pressure and stabilized at a safe level suitable for the actuator. This module consists of a pressure regulating valve and a safety and control pressure switch.

1) Pressure Regulator Filter: A diaphragm-type pressure regulating filter (AFR-2000) is installed at the reservoir outlet to reduce the line pressure to a constant working pressure regardless of inlet or downstream flow fluctuations, with a dual-scale (PSI/bar) analog gage for visual monitoring. A rotary knob compresses a spring against the diaphragm to set the valve-opening force, letting the operator select the magnitude of force applied to the actuator.

2) Safety and Control Pressure Switch: To automate reservoir filling and ensure safety, a pressure switch from Evotek Hydraulic Technology Co., Ltd. (Shanghai, China) continuously monitors the cylinder's internal pressure with adjustable hysteresis, typically turning the compressor on at 80 PSI (550 kPa) and off at 120 PSI (830 kPa). Internally, a sensing element (bellow or diaphragm) moves an electrical contact that opens at the upper limit, interrupting the compressor circuit, and closes at the lower limit to restart it.

C. Airflow Modulation Module

This subsystem is responsible for converting the continuous airflow into a high-frequency pulsatile flow. It consists of a solenoid valve and a high-speed airflow switch.

1) Electromechanical Solenoid Valve: A direct-acting, normally closed On/Off solenoid valve (Ar Fusion Brasil Pneumática, Americana, SP, Brazil) serves as the system's master switch, allowing air to flow from the conditioning module to the modulator when activated. It features a 12 VDC, 16 W coil, a brass body, and a 1/4" flow orifice: energizing the coil displaces a ferromagnetic plunger to open the passage, while a spring reseals it on de-energization, with a millisecond-scale response time sufficient for on/off control.

2) Electroacoustic Modulator: We engineered a structure, called an electroacoustic modulator, to manipulate the airflow delivered to the next module, specifically designed to control airflow interruption and featuring high switching speeds between open and closed states. The device operates by discharging high-speed airflow in pulses. It was designed using SolidWorks CAD software (Dassault Systèmes SOLIDWORKS Corp., Waltham, Massachusetts, USA) and fabricated via precisely additive manufacturing (3D printing) with PETG filament. PETG was selected for its high mechanical strength, rigidity, and low moisture absorption.

D. Pneumatic Air Actuator Module

This is the final stage of the pneumatic system, in which the energy of compressed air is converted into periodic mechanical motion for application to biological tissue.

1) Diaphragm Membrane Displacement: The actuator is an electromechanical device designed to transmit air pressure waves to the tissue of interest. The actuator body was designed using CAD (SolidWorks) and fabricated via 3D printing using PETG. Its geometry consists of a cylindrical cavity with a rear inlet for the pulsatile air. The front face of the cavity is sealed by a diaphragm latex membrane (or another low-stiffness, biocompatible elastomer) which acts as a mechanical wave driver. The cyclic increase in internal cavity pressure, caused by the arrival of the air pulses, deforms the membrane outward in a parabolic shape. This deformation (linear motion) constitutes the mechanical wave applied to the tissue in contact with the membrane. The actuator body features a small additional port connected to a flow splitter. This port allows controlled outflow of a fraction of the air and provides a connection to a pressure sensor for monitoring the internal cavity pressure.

2) Pressure Sensor: Real-time pressure acquisition used an MPX2202GP gage pressure sensor (NXP Semiconductors, Eindhoven, The Netherlands; 0-200 kPa range), a low-cost piezoresistive sensor with 0.2 mV/kPa typical sensitivity and a 10-16 VDC supply. A silicon diaphragm with piezoresistors in a Wheatstone bridge deforms under pressure, unbalancing the bridge and generating a differential output voltage, which is amplified and conditioned by a dedicated driver for microcontroller readout. A small-diameter orifice (flow orifice) is connected to the secondary outlet of the flow splitter.

E. User Interface

In this work, a 3.4-inch Nextion TFT LCD touchscreen (Shenzhen, Guangdong, China) interfaces with an ESP32 microcontroller (Espressif Systems, Shanghai, China) for automation and control. This Human-Machine Interface combines a display-integrated processor, 16 MB of flash memory, and Nextion Editor software for GUI development; the display handles graphics (screens, buttons, charts) while the microcontroller updates progress and status via simple ASCII-based text instructions.

F. Power Source

The power supply system for electronic control boards and actuators consists of a 360 W high-performance power supply (SMPS- Switched-Mode Power Supply). This source provides a stabilized 12 VDC voltage and a maximum current of 30 A. It has galvanic isolation between the primary (mains) and secondary (circuits), as well as protection against overcurrent, short circuit, and overvoltage.

The compressor, in turn, is powered directly by the 220 VAC and 60 Hz electrical network, with its own power circuit that includes a protection fuse sized for the starting current and a manual override switch.

G. System Cabinet

All components (pneumatic modules, power supply, motherboard) were integrated inside a metal case, chosen for mechanical shock resistance and Faraday-cage electromagnetic shielding, and mounted on four 360° casters for multidirectional mobility.

III. Methods

A. Cylinder Air Loading Evaluation

The initial test measured the time to pressurize the initially empty reservoir from 0 to 120 PSI in 10-PSI increments, to evaluate the compressor's ability to maintain stable pressure under varying demand. Instantaneous pressure readings were obtained by direct observation of the gauge needle, recorded via smartphone in video-stopwatch mode and tabulated for analysis.

B. Compressor Duty Cycle Evaluation

For compressor duty cycle evaluation, the system was configured with the pressure switch set to a 40 PSI hysteresis, operating within an 80 to 120 PSI range. The system was monitored in various air consumption patterns at constant flow rates of 5, 10, 15, 20, and 25 L/min, with flow rates adjusted using a pressure regulator valve. The compressor's operating times (t_{on}) and rest times (t_{off}) were recorded for each condition.

C. Static Force Calibration

The force generated by the pneumatic actuator's membrane was evaluated with the driver in the horizontal position. Its diaphragm was faced downward, resting on the weighing pan of a precision scale. We then applied increasing static pressures, ranging from 0 to 50 kPa in 10 kPa increments (0.1 bar). The resulting force was displayed on the scale and recorded for subsequent analysis.

D. Pressure-Displacement Relationship

The displacement amplitude of the actuator diaphragm membrane was measured using an FPGA-based development board equipped with two 14-bit, 125 MS/s RF analog channels (ADCs/DACs). Signal processing used a custom algorithm to estimate micrometer-scale vibration motion through coherent phase tracking of ultrasound pulses. A dual-element transducer was used, with a center frequency of 5.8 MHz and a pulse repetition frequency (PRF) of 20 kHz. This transducer incorporates two piezoelectric elements within a single casing, separated by an acoustic barrier: one element transmits the ultrasound signal. The other receives echoes from the medium. During testing, the pneumatic actuator body was partially immersed in a circular container filled with room-temperature water. The transducer was positioned diametrically opposite the actuator at a distance of 50 mm, with its face oriented toward the center of the circular diaphragm. The actuator's excitation frequency was set to 20 Hz to maximize the echo signal-to-noise ratio (SNR). Pressures ranging from 0 to 50 kPa were applied in 2 kPa increments, and the axial displacement of the diaphragm membrane was recorded at each increment.

E. Propagation in Gel

To evaluate the system's ability to generate mechanical waves in media that mimic the properties of biological tissues, we prepared a cylindrical paraffin gel block with a diameter of 10 cm and a height of 2 cm. The medium's temperature was kept constant at 25°C throughout the experiment. This controlled thermal condition is crucial because the mechanical and acoustic properties of paraffin, such as the elastic modulus and the speed of sound propagation, are temperature-dependent.

The pneumatic driver actuator vibrating membrane was positioned on the top surface of the block, in its central region. This actuator was programmed to impart small displacements to the material with an amplitude of around 2 mm, operating at frequencies of 10, 20, 30, 40, 50, 100 and 200 Hz.

A dual ultrasound transducer was coupled to the bottom face of the block, diametrically opposite the actuator; this transducer emitted ultrasonic pulses that passed through the material and detected the reflected or transmitted mechanical waves. The measurement principle relies on the fact that the displacement of moving structures induces a phase shift in the transmitted or backscattered ultrasound signals.

To accurately estimate vibration, an algorithm based on coherent phase tracking was employed; this method uses phase correlation between ultrasound signals acquired at different time points to determine the micrometric displacement induced by vibration.

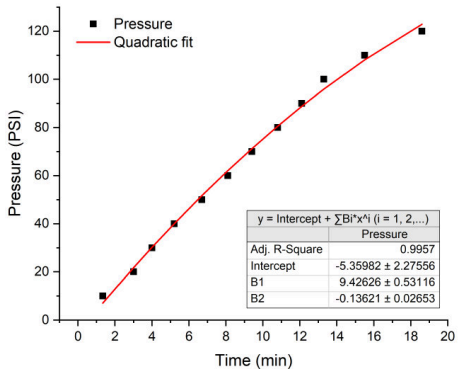
Ultrasound signals were acquired at each excitation frequency. Based on the analysis of phase shift and sound speed in the material (approximately 1448 m/s for paraffin gel at 25 °C). It was possible to determine the material's vibration velocity (m/s) in response to the applied excitation.

IV. Results and Discussion

A. Cylinder Air Loading Evaluation

The air loading of the cylinder was modeled using a quadratic fit, as shown in Fig. 2, with a coefficient of determination (R^2) = 0.997. The concavity, defined by the coefficient B2 = -0.13621, indicates that the curve gradually decelerates. The time required to increase the pressure inside the cylinder increases as the pressure approaches 120 PSI. This deceleration reflects the thermodynamic effects of adiabatic compression, where the air mass inside the reservoir increases, requiring progressively more work to achieve each subsequent pressure increment. As noted by Uffmann and Ladd [8], pneumatic systems face inherent challenges in maintaining stable pressure under varying demand, and our results confirm that the reservoir design effectively dampens compressor pulsations while providing sufficient stored energy for peak demands.

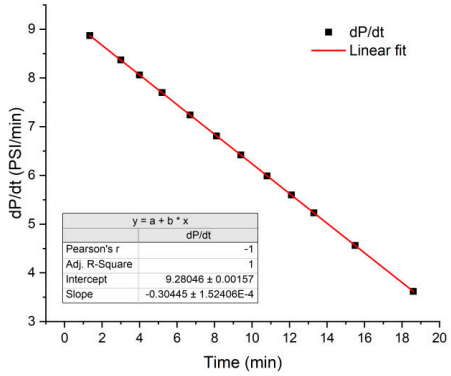
Fig. 2: Air cylinder load curve as a function of the time the compressor remains on, t_{on}



The derivative, representing the average compression rate (dP/dt), is shown in the curve in Fig. 3, fitted by linear regression, with an excellent coefficient of determination (R^2) = 0.9999. We observe that the average compression rate decreases linearly over time, as described by the fitted equation: $dP/dt = 9.2804 - 0.3044\,t$ (with t in minutes

and dP/dt in PSI/min), where the negative slope indicates that the compression rate drops by 0.3044 PSI/min for each additional minute of operation. This means that the medium is "hardening" or compacting, requiring more time for each increment in pressure.

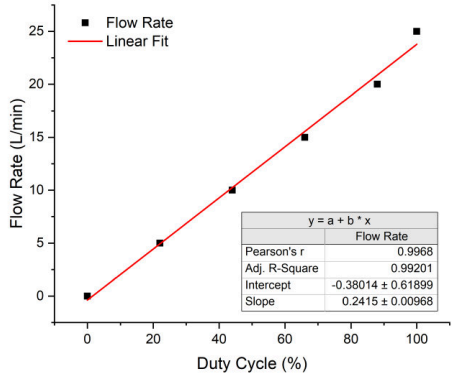
Fig. 3: Air cylinder compression rate curve versus charging time.



B. Compressor Duty Cycle

The results presented in Fig. 4 show an inverse correlation between the consumption flow rate and the compressor's operating time (t_{on}) and the rest time (t_{off}), with the duty cycle (C) varying linearly with pneumatic demand. The system showed excellent agreement with the theoretical model $C = t_{\text{on}}/(t_{\text{on}} + t_{\text{off}})$, with a maximum deviation of only 0.1% between the measured and calculated values.

Fig. 4: The compressor duty cycle is determined by the compressor's operating times (t_{on}) and rest times (t_{off}) for each condition.

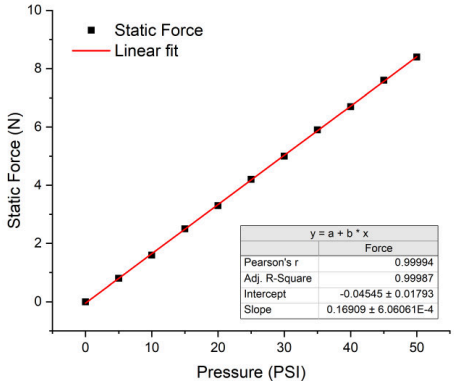


The effective flow rate of the compressor was determined to be 22.7 L/min, which sets the maximum consumption limit for cyclic operation. For demands of 5, 10, 15, and 20 L/min, the operating times were 7.68, 10.71, 17.66, and 50.37 min, respectively, with corresponding duty cycles of 22%, 44%, 66%, and 88%. Operating at a demand of up to 20 L/min is recommended to ensure a safety margin. Demands exceeding 22.7 L/min make cyclic operation unfeasible, causing the compressor to run continuously and leading to a steady drop in pressure. The results underscore the importance of properly sizing the pneumatic system to avoid both undersizing, which can lead to overheating, and oversizing, which can reduce energy efficiency—a trade-off often neglected in laboratory-constructed MRE actuators [9].

C. Static Force Calibration

Fig. 5 shows the static calibration results, showing a linear relationship between applied pressure and the force generated by the pneumatic actuator's membrane, with a coefficient of determination of $R^2 = 0.9998$. The force ranged from 0 N at 0 kPa to 8.4 N at 50 kPa, resulting in an average sensitivity of approximately 0.168 N/kPa.

Fig. 5: Static force calibration: force versus applied pressure.



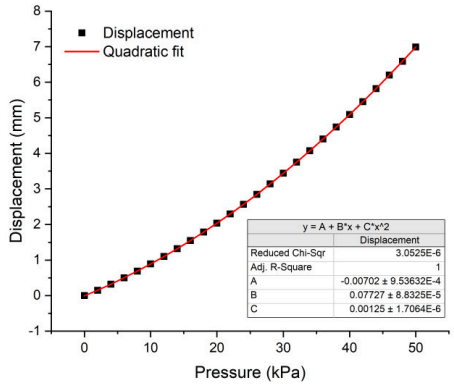
The observed linearity, with no evidence of hysteresis or saturation, confirms that the membrane operates within the linear-elastic regime described by the generalized Hooke's Law for elastomeric materials, indicating negligible friction or leakage losses. This linear pressure-force response is crucial for MRE, enabling precise, repeatable control of the force — and thus the shear-wave amplitude — applied to the tissue by regulating the supply pressure [1].

D. Pressure-Displacement Relationship in Water

The displacement as a function of pressure is illustrated in Fig. 6. This curve was obtained at a fixed, low excitation frequency (20 Hz, Section III-D) to maximize the ultrasound-tracking SNR, isolating the effect of pressure from frequency-dependent effects; the complementary amplitude-frequency dependence, at fixed pressure, is characterized below (Velocity-Frequency Relationship in Water), where amplitude falls from 3.34 mm at 100 Hz to 1.81 mm at 500 Hz. Pressure and frequency thus act as largely independent control parameters — pressure sets the quasi-static displacement amplitude, while frequency modulates it at higher drive rates — though a full joint characterization was not performed and is left for future work. The system operates predominantly in the linear-elastic regime (R^2 = 0.9997 for the linear fit); however, the increments $\Delta d_\mu/\Delta P$ (0.075-0.200 mm/kPa) reveal a slightly quadratic, upward-concave deviation ($C > 0$).

Although the quadratic fit (R^2 = 0.9999) is slightly superior, the operating range considered strictly linear (deviation $\leqslant$ 5%

Fig. 6: Relationship between the pressure generated by the actuator's diaphragm membrane and its corresponding displacement in water.



relative to the linear model) is limited to 40–50 kPa. This narrowing of the linear range is attributed to the system's progressive compliance, possibly due to an increase in the effective area of the diaphragm or to the settling of elastomeric seals under pressure. For MRE applications that require maximum linearity, it is recommended to operate the transducer within the 40–50 kPa range, where the sensitivity is nearly constant, approximately 0.19 to 0.20 mm/kPa. For the full range of 0 to 50 kPa, the quadratic model best describes the behavior, with an error of less than 0.1% throughout. The ability to generate displacements of several millimeters exceeds the typical requirements of MRE (tens of micrometers) [1], providing ample margin for tissue coupling and ensuring sufficient wave amplitude even with attenuation through tissue.

E. Velocity-Frequency Relationship in Water

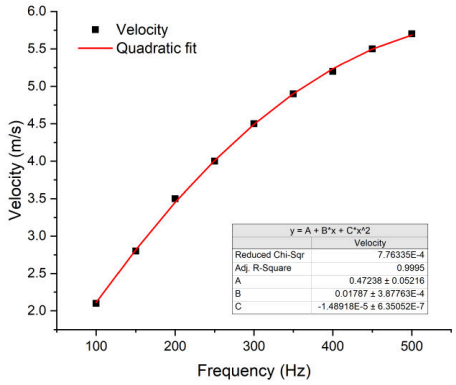
The velocity as a function of frequency is illustrated in Fig. 7. This test was conducted with the driving pressure held fixed, so that the resulting curve isolates the effect of frequency on wave velocity and amplitude. The results show a sublinear increase in velocity with frequency, with a gain decreasing from 0.014 to 0.004 m/s·Hz and a tendency toward saturation at approximately 6.0 m/s. This behavior is attributed to a reduction in vibration amplitude as frequency increases, from 3.34 mm at 100 Hz to 1.81 mm at 500 Hz.

The 100-300 Hz range shows more linear gain and is recommended for operation, while above 350 Hz the system enters a strongly saturated regime; the estimated natural frequency below 100 Hz suggests further measurements would be needed to fully characterize the system's dynamics in that band. This sublinear response is typical of pneumatic systems, which Triolo et al. [12] report as limited above 100 Hz by phase delays and transmission-line compliance — a range our system significantly extends, particularly relevant for MRE of stiff tissues requiring shorter wavelengths [14].

F. Electroacoustic Modulator Control

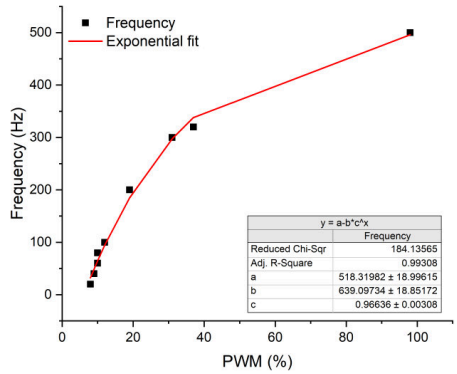
The pulsation-frequency response of the actuator membrane is governed by the duty cycle of the electroacoustic

Fig. 7: Velocity of the mechanical wave propagation in water at different frequencies.



modulator control driver, as shown in Fig. 8. Experimental results demonstrate a non-linear relationship between the duty cycle (PWM) and the pulsation frequency of the pneumatic actuator's membrane. For PWM values between 8% and 12%, the frequency remains in a relatively low range of 20 to 100 Hz. Starting at 19% PWM, an abrupt rise to 200 Hz is observed, followed by a progressive increase to 500 Hz at 98% PWM, with intermediate points at 300 Hz (31%), 320 Hz (37%) and 500 Hz (98%).

Fig. 8: The actuator membrane's pulsation-frequency response as a function of PWM duty cycle.



This behavior highlights the sensitivity of the electroacoustic modulator to pulse width, indicating that the electronic airflow control system directly influences the membrane dynamics. The non-monotonic relationship and observed jumps suggest saturation zones or non-linear response characteristics inherent to the actuator's electromechanical coupling; the 19-37% PWM range exhibits the highest gain and linearity and is the recommended operating region, while above 98% PWM the system approaches saturation, consistent with high-speed pneumatic systems where valve switching times limit the maximum frequency [8].

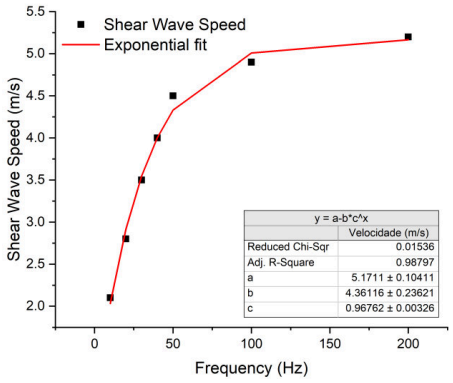
This PWM-frequency mapping underlies open-loop frequency selection via the user interface; however, its run-to-run stability and the error between programmed and actual pulsation frequency have not yet been systematically quantified, and

are expected to be largest near the non-linear transition regions (e.g., 12-19% PWM) rather than the more linear 19-37% band. A lookup-table-based closed-loop calibration, together with this stability/error assessment, is identified as a priority for future work.

G. Wave Propagation in Paraffin Gel

Fig. 9 shows the relationship between the excitation frequency and the propagation velocity of transverse waves on paraffin gel. This material is considered a linear, isotropic, and homogeneous elastic medium; the transverse wave velocity (v_T) is related to the shear modulus by $G = \rho v_T^2$, where ρ is the density. For paraffin, that value is around 0.82 g/cm^3. The results show that the shear modulus increases with frequency, ranging from 3.66 kPa (at 10 Hz) to 22.44 kPa (at 200 Hz).

Fig. 9: Shear modulus of paraffin gel as a function of excitation frequency.



This behavior is the classic signature of viscoelastic dispersion, in which the material exhibits an increase in apparent stiffness with increasing strain rate [7]. The paraffin gel exhibited linear viscoelastic behavior in the 10-200 Hz range. The shear modulus increases with frequency due to viscoelastic dispersion, reaching saturation above 50 Hz. The values (3.66–22.44 kPa) are consistent with those observed in low-stiffness materials, such as those used in soft-tissue-mimicking ultrasound phantoms [3], [5]. Since the temperature was kept constant (25 °C), the variation in shear modulus is attributed exclusively to frequency dependence, which confirms viscoelastic behavior. The observed behavior is consistent with that described in shear-wave dispersion studies of gel systems [3].

This phantom characterization was limited to 10-200 Hz due to the phase-tracking algorithm's sensitivity to the smaller amplitudes and greater attenuation expected at higher frequencies in this stiffer medium, and is distinct from the modulator's up-to-500 Hz response demonstrated separately in water. Extending gel-phantom measurements to 500 Hz, with amplitude, attenuation, and repeatability reported per frequency, is identified as a priority for future work to confirm the full 10-500 Hz bandwidth in tissue-mimicking media rather than in water alone.

H. Drives and User Interface

The user interface, implemented through the Nextion display, proved to be intuitive and efficient. The screen allowed for navigation between the main menu, the standby screen, and the main control screen. In the latter case, the operator could set the desired frequency and start or stop the system. Real-time feedback, which displays the measured frequency and pressure in kgf, was clearly presented, validating open-loop control under operator supervision.

The control architecture proved satisfactory. A key observation was the need for proper processing of pressure sensor signals; a moving average was implemented in the code to smooth the readings and prevent display fluctuations caused by high-frequency noise from the airflow modulation module and air turbulence itself. This highlights the importance of proper signal conditioning in pneumatic systems, where turbulent airflow and valve switching can introduce significant noise [8].

I. Power Supply System and Enclosure

The power system was designed and built with a focus on safety and resistance to mechanical shock. The 12 V/30 A switching power supply proved capable of powering all low-voltage components and electronic boards, even under maximum load. The dual-pole input fuse protection (rated at 10 A) operated correctly during a test-induced short circuit, cutting power before any components were damaged.

The metal enclosure housed all components in an organized, secure manner. The final dimensions (74 cm high, 38 cm long, and 38 cm deep) and a total weight of 20 kg resulted in a compact, stable, and mobile unit, thanks to the inclusion of wheels, meeting ergonomic and transport requirements for clinical or laboratory use.

The selection of a metal enclosure was a wise decision, providing both mechanical protection and electromagnetic shielding. This is intended to protect sensitive electronics from general external interference and to prevent control signals (PWM, pulses) from radiating and interfering with other equipment [10]. The 30 A power supply's extra current headroom ensures operation within a high-efficiency, low-thermal-stress range, thereby extending the system's service life. The layered safety design (AC fuses, galvanic isolation of the power supply and fuses or PTCs on the control board) adheres to engineering best practices and mitigates the risks of electrical shock and fire.

It should be emphasized that this metallic enclosure was conceived as a Faraday cage for general electromagnetic compatibility (EMC), containing the electronics' own emissions and attenuating ambient interference during benchtop use, which, by itself, does not constitute demonstrated MRI compatibility. Establishing MR compatibility would require dedicated in-scanner testing (gradient synchronization, image-artifact assessment, RF-induced heating, and static/gradient-field behavior), none of which has been performed; the actuator has so far been operated and validated exclusively on

the benchtop and in a tissue-mimicking phantom, outside the scanner environment.

J. Pneumatic Actuator, Flow Splitter, and Pressure Sensor

The pneumatic actuator, the heart of the transduction system, was fabricated according to the design. The cylindrical cavity of the 3D-printed actuator was perfectly sealed when coupled to the latex membrane. The airflow splitter, integrated into the actuator body, directed a portion of the pressurized air to the MPX2202GP pressure sensor, while the remaining air was vented through a small-diameter capillary in the outlet regulator filter. This configuration enabled continuous monitoring of internal pressure without compromising the system's dynamics.

It is important to distinguish two pressure quantities: the internal driving pressure applied to the actuator cavity (0-50 kPa, characterized above) and the pressure actually transmitted to the contact surface, attenuated by the membrane and contact area. This latter, contact-surface pressure was verified with a calibrated pressure analyzer (Handy Press Sphygmomanometer Analyzer, PHPR0004, Protec, Cotia, SP, Brazil), yielding a mean of 2.25 kPa (~16.9 mmHg). Low and consistent with pressures generally tolerable for skin contact, encouraging for eventual *in vivo* use, though this single bench measurement does not by itself establish clinical or biocompatibility suitability. Adjusting the internal driving pressure via the compressor's regulating valve allows the stimulus intensity to be tailored to different tissue types or depths, a feature not available in many commercial systems [12]. The MPX2202GP sensor responded well to pulsatile pressure variations; however, its reading is currently a single-point/linear estimate that requires finer, multi-point calibration against a reference instrument to be accurately converted to kgf and to compensate for non-linearities, planned as future work.

These results demonstrate that the proposed actuator addresses key limitations of current MRE actuators, combining the general safety advantages of pneumatic technology with the control precision and extended frequency range needed for advanced MRE; all results reported here correspond to benchtop/phantom conditions, and in-scanner MRI compatibility remains to be established.

V. Conclusions

This work successfully designed, built, and characterized a novel electro-pneumatic actuator for MRE, addressing key limitations of existing commercial systems. The prototype demonstrated a 10-200 Hz operational range validated in a paraffin gel phantom, while the electroacoustic modulator itself reached up to 500 Hz in water, alongside linear force control and a well-characterized, slightly non-linear pressure-displacement response, all established on the benchtop. The modular, low-cost design (3D-printed components, touch-screen interface, EMC-shielded metal enclosure) confirms the system's potential for MRE research beyond current commercial pneumatic drivers, though these results correspond strictly to benchtop/phantom conditions, and clinical or in-scanner MRI compatibility remains a separate, unaddressed step.

Four directions are prioritized for future work: (i) extending gel-phantom propagation measurements to 500 Hz, reporting amplitude, attenuation, and repeatability per frequency; (ii) quantifying PWM-frequency mapping stability and the programmed-versus-measured frequency error, especially across the non-linear transition regions; (iii) implementing multi-point pressure-sensor calibration against a reference instrument; and (iv) integrating and testing the actuator inside an MRI scanner (gradient synchronization, image artifacts, RF heating, field interaction), alongside closed-loop control and *in vivo* validation.

References

[1] Y. K. Mariappan, K. J. Glaser, and R. L. Ehman, "Magnetic resonance elastography: a review," *Clinical Anatomy*, vol. 23, no. 5, pp. 497–511, 2010.

[2] S. L. Vieira, L. N. de Oliveira, and A. A. O. Carneiro, "Princípios físicos da elastografia por ressonância magnética," *Caderno Brasileiro de Ensino de Física*, vol. 36, no. 2, 2014.

[3] G. Low, S. A. Kruse, and D. J. Lomas, "General review of magnetic resonance elastography," *World Journal of Radiology*, vol. 8, no. 1, pp. 59, 2016.

[4] S. L. Vieira, L. N. de Oliveira, and A. A. O. Carneiro, "Quantitative magnetic resonance elastography for polymer-gel dosimetry phantoms," *Medical Engineering & Physics*, Vol. 66, no. 1, pp. 102–106, 2019.

[5] T. Z. Pavan, S. L. Vieira, and A. A. O. Carneiro, "Elastography: A New Ultrasound and MRI Procedure," in *AIP Conference Proceedings*, vol. 1032, no. 1, pp. 51–59, American Institute of Physics, 2008.

[6] R. Muthupillai, D. J. Lomas, P. J. Rossman, J. F. Greenleaf, A. Manduca, and R. L. Ehman, "Magnetic resonance elastography by direct visualization of propagating acoustic strain waves," *Science*, vol. 269, no. 5232, pp. 1854–1857, 1995.

[7] S. L. Vieira, *Imagens das propriedades viscoelásticas por ressonância magnética e ultrassom*. Ph.D. dissertation, FFCLRP, Universidade de São Paulo, 2009.

[8] K. Uffmann and M. E. Ladd, "Actuation systems for MR elastography," *IEEE Engineering in Medicine and Biology Magazine*, vol. 27, no. 3, pp. 28–34, 2008.

[9] Z. T. H. Tse, H. Janssen, A. Hamed, M. Ristic, I. Young, and M. Lamperth, "Magnetic resonance elastography hardware design: a survey," *Proceedings of the Institution of Mechanical Engineers, Part H*, vol. 223, no. 4, pp. 497–514, 2009.

[10] J. Braun, K. Braun, and I. Sack, "Electromagnetic actuator for generating variably oriented shear waves in MR elastography," *Magnetic Resonance in Medicine*, vol. 50, no. 1, pp. 220–222, 2003.

[11] E. R. Gizewski, O. Koeze, K. Uffmann, A. de Greiff, M. E. Ladd, and M. Forsting, "Cerebral activation using a MR-compatible piezoelectric actuator with adjustable vibration frequencies and in vivo wave propagation control," *NeuroImage*, vol. 24, no. 3, pp. 723–730, 2005.

[12] E. R. Triolo, O. Khegai, E. Ozkaya, N. Rossi, A. Alipour, L. Fleysher, P. Balchandani, and M. Kurt, "Design, construction, and implementation of a magnetic resonance elastography actuator for research purposes," *Current Protocols*, vol. 2, no. 3, pp. e379, 2022.

[13] Y. Zheng, G. Li, M. Chen, Q. C. Chan, S.-G. Hu, X.-N. Zhao, R. L. Ehman, E. Y. Lam, and E. S. Yang, "Magnetic resonance elastography with twin pneumatic drivers for wave compensation," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2611–2613, IEEE, 2007.

[14] I. Sack, "Magnetic resonance elastography from fundamental soft tissue mechanics to diagnostic imaging," *Nature Reviews Physics*, vol. 5, no. 1, pp. 25–42, 2023.

Development of a Low-Cost Apparatus with Monitoring Sensors for the Controlled Exposure of Rats to Cigarette Smoke

1st Larissa Hitomi Morine Cortegana
Department of Science and Technology
Institute of Science and Technology
UNIFESP
São José dos Campos, Brazil
hitomi.larissa29@unifesp.br

2nd Alyne Riani Moreira
Department of Medicine
Laboratory of Experimental Therapeutics (LIM20)
University of São Paulo
São Paulo, Brazil
alyne.riani@gmail.com

3rd Fernanda Degobbi Tenorio Quirino dos Santos Lopes
Department of Medicine
Laboratory of Experimental Therapeutics (LIM20)
University of São Paulo
São Paulo, Brazil
fernandadtqsl@gmail.com

4th Fabio Gava Aoki
Department of Science and Technology
Institute of Science and Technology
UNIFESP
São José dos Campos, Brazil
fgaoki@unifesp.br

***Abstract*—The inhalation of cigarette smoke over extended periods exposes the lungs to toxic substances, including carbon monoxide (CO) and particulate matter, thereby contributing to the onset and progression of chronic respiratory diseases. Experimental models involving small rodents and controlled exposure chambers are widely used to investigate the biological effects of this exposure. However, characterization of the CO and particulate matter (PM) generated in these chambers typically relies on expensive instrumentation, limiting access for research groups with limited resources. Consequently, this study describes the development of a cigarette smoke exposure system intended for future trials with small rodents; the goal was to build a low-cost chamber capable of monitoring temperature, humidity, the response signal from the MQ-7 sensor and particulate matter (PM2.5 and PM10) using low-cost sensors (DHT11, MQ-7, and SPS30). Tests conducted with varying rest intervals between puffs (5, 10, and 15 seconds) showed no statistically significant variation for CO (ANOVA, p = 0.214), PM2.5 (ANOVA, p = 0.771), and PM10 (ANOVA, p = 0.684). Despite the need for improvements to the experimental procedure, the electronic architecture appeared to perform adequately regarding signal acquisition.**

***Index Terms*—exposure chamber, cigarette smoke, low-cost sensors, particulate matter, carbon monoxide.**

I. Introduction

Chronic respiratory diseases comprise a group of conditions that persistently affect the respiratory system, impairing lung function and increasing vulnerability to infections [1]. Among these diseases, Chronic Obstructive Pulmonary Disease (COPD) stands out; according to the World Health Organization (WHO), it was associated with approximately 3.5 million deaths in 2021, ranking fourth among the leading causes of death globally [2].

The occurrence of these diseases is linked to the interaction of multiple factors—including genetic, physiological, and environmental aspects—as indicated by epidemiological studies [1]. In the case of COPD, a significant association is observed with continuous exposure to harmful particles and gases, particularly those derived from cigarette smoke [1]. The World Health Organization (WHO) recognizes that exposure to secondhand smoke is associated with substantial health risks and contributes to preventable mortality among non-smokers [3]. In this context, the smoke resulting from cigarette combustion has a complex composition, containing around 7,000 toxic chemical substances, including carbon monoxide (CO) [4] and particulate matter (PM) — which consists of fine particles capable of penetrating deeply into the respiratory tract [5].

A significant portion of research into the toxicological effects of cigarette smoke relies on experimental models using small rodents. Consequently, the use of controlled exposure chambers is common for assessing and understanding how these substances can physiologically alter the human system. Previous studies, such as those by Barbosa et al. (2020) [6] and Junqueira et al. (2020) [7], demonstrate the effects of exposure to conventional cigarette smoke in experimental models using specialized chambers.

However, the equipment required to characterize the smoke generated in these chambers typically consists of high-cost instruments—such as certified gas analyzers (CO) and laser photometers (PM)—whose high price can pose a barrier for certain research groups [8]. Although low-cost experimental setups do exist, cost reductions generally apply only to the structural and mechanical components, while reliance on ex-

pensive commercial devices remains [9].

Thus, this study proposes the development of a conventional cigarette smoke exposure chamber intended for future experiments with small rodents, with the primary objective of using low-cost sensors to monitor carbon monoxide and particulate matter generated by cigarette combustion. This study entails the development and bench testing of the proposed system, involving neither biological exposure nor the use of animals at this stage. The long-term proposal aims to further advance knowledge regarding the toxicological implications of these emissions and their impact on public health by offering a replicable and more accessible alternative.

II. Material and Methods

In order to enable the analysis of particles emitted by conventional cigarettes, this work proposed the development of an exposure system capable of collecting the smoke generated. The system works as a controlled experimental model, allowing standardization of exposure conditions. Furthermore, its physical structure was divided into two compartments, a strategy adopted to favor the accuracy of the readings obtained.

Thus, the primary chamber houses the MQ-7 sensor, which provides a relative response to the gaseous components of cigarette smoke, with CO as its primary target gas, and the DHT11 sensor, used to monitor temperature and humidity. In this same chamber, an electromechanical vacuum system was incorporated, developed to simulate puffing on a cigarette. The analysis of particulate matter (PM) is carried out in a secondary chamber, equipped with the Sensirion SPS30 sensor, which receives sample from the first compartment to be diluted.

The data acquisition architecture was organized with two independent Arduino microcontrollers: the first responsible for the MQ-7 and DHT11 sensors, used to monitor the relative MQ-7 sensor response to gases, temperature and humidity, and the second dedicated exclusively to the particulate matter (PM) sensor - this division allowed the simultaneous processing of information obtained during the tests. Finally, to visualize the measurements in real time, LCD displays incorporated into the system's two electronic circuits were used.

A. Construction of the Primary and Secondary Exposure Chambers

The construction of the primary chamber was inspired by the exposure chamber developed by the LIM 20 team at the *Hospital das Clinicas* of the Faculty of Medicine of the University of Sao Paulo [6] [7]. Therefore, a plastic organizing box with an approximate volume of 30L was used, where four holes were made: two on the sides and two on the lid (Fig. 1.)

In the lower side wall, the first hole was made to allow the mini air compressor to be coupled, with a flow rate of 1.6 L/min. The second hole was positioned on the opposite upper side face and received a hose that was used as an outlet for air to the atmosphere. In this way, the system maintains pressure balance and promotes continuous air renewal, reducing the risk of CO accumulation in concentrations harmful to animals.

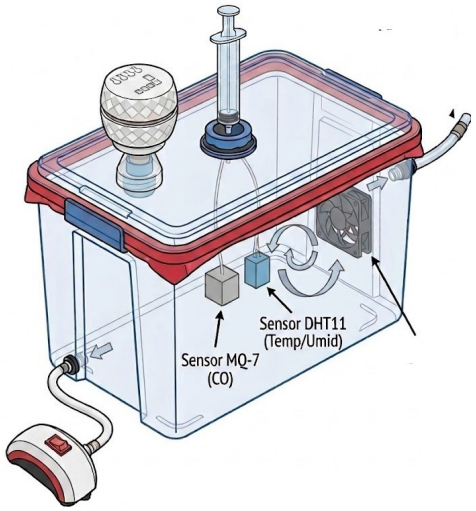


Fig. 1. Arrangement of the Primary Chamber Holes. For more details on the positioning of holes and sensors, see the body of the text.

Finally, in the lid of the box, the other two additional openings were created, one of which was intended for attaching the lung simulator, while the other, located in the central region of the lid, received an elastic rubber seal, used to allow the insertion of the collection syringe.

The MQ-7 and DHT11 sensors were fixed in the central region of the primary chamber to ensure sampling homogeneity, being assisted by an internal microfan that optimizes air circulation and mixing. For the analysis of particulate matter (PM), the implementation of a secondary chamber was essential due to the operational restrictions of the SPS30 sensor, whose reliable measurement range is limited to concentrations of up to 1000 µg/m^3 [10]. Consequently, a 3.1 L glass container was adopted to accommodate the diluted sample (Fig. 2). The lid of this compartment was adapted with two access routes: one dedicated to the controlled injection of the sample using a syringe and the other for the passage of the SPS30 sensor cables.



Fig. 2. Arrangement of the secondary chamber. It is possible to see the SPS30 sensor positioned in the glass jar.

B. Mechanical Cigarette Puff Simulation System

The system responsible for simulating the cigarette puffs was developed with the aim of reproducing the suction and exhaust cycles involved in burning a cigarette. The pneumatic structure of the device was obtained from the adaptation of a breast milk pump, whose silicone membranes and plastic junctions were reused to compose a mechanism similar to a diaphragm associated with an exhaust duct. In this way, the aspirated air is temporarily retained and subsequently conveyed to the main exposure chamber.

The assembly of the cigarette and mechanical components on the box lid was made possible by parts made using 3D printing, allowing better fixation and alignment of the device. Regarding electronic control, the system operates through the alternate activation of a suction motor and a release valve. The drive is coordinated by an Arduino microcontroller connected to a single-channel relay module. During the suction stage, the relay activates the motor, promoting the formation of a vacuum and the aspiration of smoke, while the valve remains closed. In the next step, when the engine is turned off, the valve is opened, allowing the smoke to be released into the chamber.

C. MQ-7 Sensor and Signal Acquisition Circuit

The gaseous response was monitored using the MQ-7 sensor, selected because CO is its primary target gas and due to its widespread use in gas-sensing and environmental monitoring applications [11]. The device's transduction principle is based on modulating the electrical resistance of a tin dioxide film (SnO_2), that is, in environments with clean air, the material presents high resistance, which decreases inversely proportional to the increase in the concentration of the detected gases [12].

Structurally, the component is equipped with heating terminals (H+ and H-) and analog reading (A and B). Prior to the testing phases, the sensor was subjected to a burn-in protocol to volatilize residual manufacturing impurities, applying a continuous voltage of 5 V to the H+ terminal for an uninterrupted period of 48 hours [12].

The MQ-7 presents behavior equivalent to that of a variable resistor sensitive to the presence of gas, with its resistance, called RS, modified according to changes in the conductivity of the sensing material. So that these variations can be interpreted by the microcontroller, the sensor is inserted into a voltage divider, enabling the conversion of the transducer's chemical response into measurable electrical potential differences.

However, it is necessary to respect the thermal cycle indicated in the technical literature. This procedure involves alternating between a high voltage heating phase of 5 V for 60 seconds, aimed at eliminating impurities, and a low voltage phase of 1.4 V for 90 seconds, during which data acquisition occurs. If this cycle was not applied, contaminants could accumulate on the sensitive layer, causing sensor saturation and compromising the accuracy of the results obtained [12].

Each battery of tests was preceded by a stage of conditioning the sensor in clean air, maintained until the reference signal (baseline) stabilized. It was decided not to use the absolute voltage value recorded at the end of each cycle, adopting instead the voltage variation (ΔV) as an analysis parameter. ΔV was obtained by the difference between the maximum voltage value reached at the end of the MQ-7 sensor reading cycle and the initial value, already stabilized in clean air moments before the first puff. This procedure made it possible to disregard the specific reference of each collection and isolate the effect attributable to the entry of cigarette smoke into the primary chamber.

The sensor in question was not calibrated using known CO concentrations; therefore, the presented (ΔV) values are relative measurements of the sensor's response and cannot be interpreted as absolute CO concentrations in ppm. Although the MQ-7 exhibits high selectivity for CO, it also displays cross-sensitivity, albeit to a lesser extent, to other gases (according to the datasheet), which prevents the measured signal from being attributed exclusively to CO.

D. Particulate Matter Detection Circuit

For the detection and quantification of particulate matter (PM) from cigarette smoke, the Sensirion SPS30 optical sensor was used. The selection of this transducer was justified by its metrological reliability and broad validation in the scientific literature for monitoring air quality [13]. The operational principle of the device is based on the scattering of laser light, the process of which begins with the active suction of air into a shielded chamber using an internal microfan. Inside the compartment, the particles interact with a laser beam, generating a light dispersion that is collected by a photodiode. This signal is processed to provide, in real time, the particle mass concentration in µg/m^3 [10].

The measurement of particulate matter (PM) was carried out with the help of a second Arduino microcontroller, powered via a USB connection. This configuration enabled both power supply to the board and direct serial communication with the software used to collect data. The SPS30 sensor was connected according to the manufacturer's specifications, using the I2C bus, while a 16x2 LCD display was added to the circuit to allow viewing of readings in real time, sharing the SCL and SDA pins. The data generated by the SPS30 was recorded in the CoolTerm software, responsible for continuously capturing the information transmitted by the Arduino. The readings were automatically saved in text files (.txt), which they were subsequently used for the statistical treatment of the results.

The SPS30 was kept in a secondary chamber, a strategy adopted to avoid direct exposure to high concentrations of aerosols and, consequently, reduce the risk of sensor saturation and compromise of its optical components. Thus, 40 mL samples of smoke present in the main chamber were collected and injected into the reading chamber, which had an internal volume of 3.1 L.

Due to this procedure, the values initially obtained by the sensor corresponded to the diluted concentration of par-

ticulate matter. To correct this difference and estimate the concentration in the main chamber, a dilution factor (DF) was used, determined by the relationship between the volume of the secondary chamber (3100 ml) and the volume of the introduced sample (40 ml)

$$\mathrm{DF} = \frac{3100\ \mathrm{ml}}{40\ \mathrm{ml}} = 77,5 \tag{1}$$

It should be noted that the dilution factor calculation assumes a homogeneous distribution of the sample introduced into the secondary chamber. Therefore, to approximate this condition, six alternating aspiration and injection cycles were performed using a 60 mL syringe after sample introduction, at a rate of approximately one second per movement. The reported PM2.5 and PM10 values represent concentrations already corrected by the dilution factor.

Suspended particle monitoring and data collection were based on concentration values recorded after the transient phase, aiming to obtain more stable readings following the insertion of the diluted sample. The detection scope focused on particle diameters corresponding to PM2.5 ($\leq$ 2.5 µm) and PM10 ($\leq$ 10 µm).

E. Experimental Procedure and Data Collection Protocol

After the integration of all electronic components, the experimental phase of the work began, involving data acquisition using commercial Marlboro cigarettes ("Red" variant). Prior to testing, the sensors were initially exposed to an environment with clean air and continuous circulation for 30 minutes until they reached thermal and chemical equilibrium, establishing a reliable baseline. After stabilization, the chambers were hermetically closed and the peripheral systems (air compressor and microfan) were activated. Simultaneously, the SPS30 sensor began continuous monitoring, updating values every 1 second.

The sample injection regime was synchronized with the thermal cycle of the MQ-7, that is, during the heating phase (60 seconds under 5 V), the system remained at rest and the activation of the suction simulator occurred exclusively during the sensor measurement phase (90 seconds under 1.4 V). The protocol established cycles of 5 puffs (each puff consisting of a pressure drop developed by a diaphragm of an approximately 30 mL, that forced the smoke in), lasting 2 seconds each, in order to evaluate the dynamics of accumulation of particles and gases. Thus, the cigarettes were reused between trials, with less than one pack being used in total. In this sense, the independent variable stipulated was the rest interval between puffs, adopting periods of 5, 10 and 15 seconds (Fig. 3).

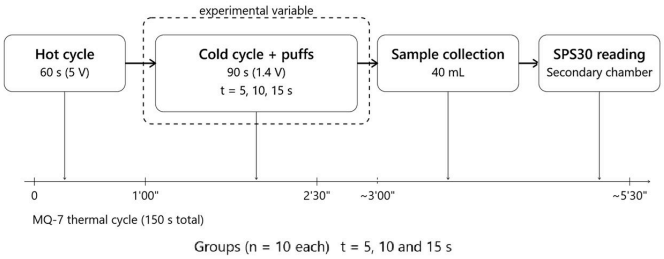


Fig. 3. Temporal protocol of the thermal cycle and synchronization of the cigarette puffs.

At the end of each reading cycle of the MQ-7 sensor, the values corresponding to its relative signal, temperature, and humidity were recorded based on the information displayed on the LCD screen. Next, the particulate material collection stage took place. To do this, a 40 mL sample of the smoke present in the main chamber was extracted with a 60 mL syringe and transferred to the secondary container. After transfer, alternating aspiration and ejection movements were performed with the syringe itself, in order to promote homogenization of the smoke inside the container and ensure better distribution of suspended particles. This procedure allowed the SPS30 sensor to read particulate matter in a more representative way, the data is automatically recorded on the computer using the CoolTerm software.

Then, both boxes were opened, keeping the fan running to remove residual smoke. The cleaning phase was maintained until the sensors returned to values corresponding to clean air, lasting a minimum of approximately 10 minutes per session. The experimental procedure was repeated in blocks by condition, specifically, 10 times for each exposure interval evaluated (5 s, 10 s, and 15 s) in a non-randomized manner, totaling 30 trials. Due to the loss of one experimental record, one observation was excluded from the statistical analyses, resulting in a sample size of 29.

F. Statistical Analysis

Analysis of variance (ANOVA) was performed to evaluate the potential influence of the rest interval between puffs on the mean concentrations of the analyzed pollutants. The normality of the ANOVA model residuals was assessed using the Shapiro-Wilk test to verify adherence to the normality assumption required for ANOVA. Subsequently, the homogeneity of variances across groups was checked using Levene's test; the data were then presented via boxplots showing medians and interquartile ranges, with outliers identified according to Tukey's rule.

III. Results

The cigarette puff simulator presented performance compatible with the expected design during the study. Temperature and humidity were monitored by DHT11 throughout the experiments, resulting in 21.1 °C $\pm$ 0.6 (mean $\pm$ s.d.) and humidity 69.5 % $\pm$ 2.5 (mean $\pm$ s.d.).

Table I presents the medians and standard deviations for the values of particulate matter and (ΔV) obtained under each

evaluated condition, and Table II presents the results of the statistical tests performed for each condition.

TABLE I
MEAN ± STANDARD DEVIATION OF ΔV, PM2.5, AND PM10 FOR DIFFERENT PUFF DURATIONS.

Variable	5 s	10 s	15 s
ΔV (V)	2.601 ± 0.446	2.298 ± 0.560	2.190 ± 0.566
PM2.5 (μg/m^3)	44337.8 ± 15456.7	48096.5 ± 16712.0	49014.4 ± 12452.0[a]
PM10 (μg/m^3)	45880.0 ± 16821.8	50948.5 ± 18631.1	52019.7 ± 13139.0[a]

[a] Nine samples were available for these conditions; all other conditions included 10 samples.

TABLE II
STATISTICAL TESTS FOR ΔV, PM2.5, AND PM10 UNDER DIFFERENT PUFF DURATIONS.

Variable	Shapiro–Wilk[a] p	Levene p	ANOVA p
ΔV (V)	0.336	0.886	0.214
PM2.5 (μg/m^3)	0.676	0.730	0.771
PM10 (μg/m^3)	0.839	0.620	0.684

[a] Shapiro–Wilk test was performed on the model residuals.

Fig. 4 shows the medians and the interquartile range of ΔV obtained for each condition evaluated.

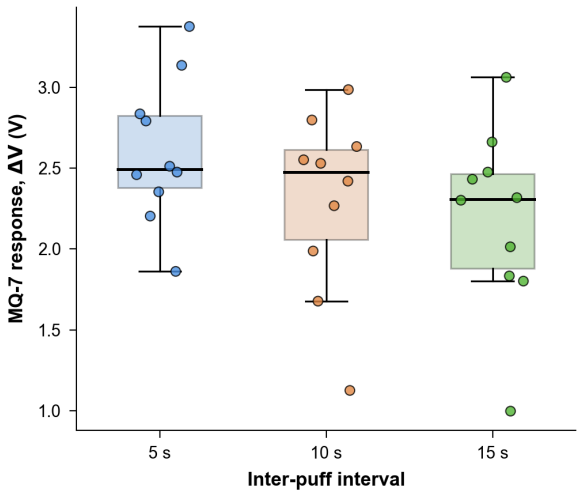


Fig. 4. Boxplot of MQ-7 relative sensor response in relation to the interval between puffs. No statistically significant differences were observed between the intervals (ANOVA, p = 0.214).

The particulate matter values detected after stabilization, expressed as medians and the interquartile range for each time interval studied, are detailed in Figs. 5 and 6.

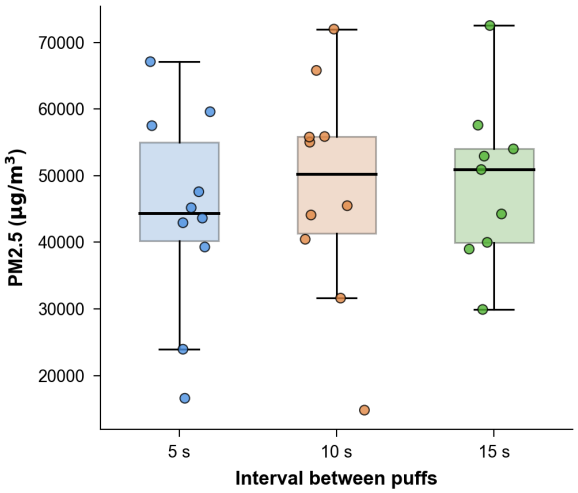


Fig. 5. Boxplot of PM2.5 concentrations in relation to the interval between puffs. No statistically significant differences were observed between the intervals (ANOVA, p = 0.771).

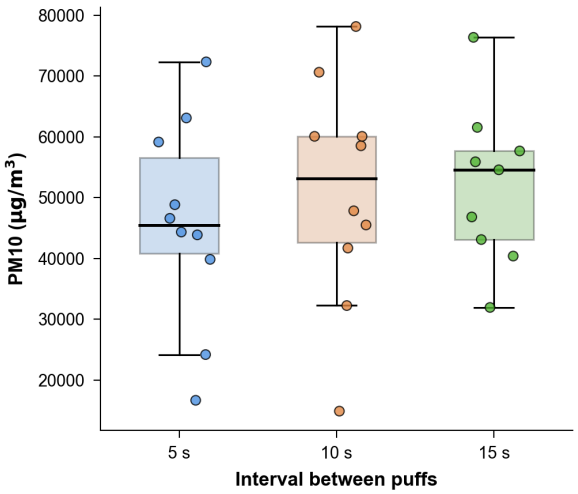


Fig. 6. Boxplot of PM10 concentrations in relation to the interval between puffs. No statistically significant differences were observed between the intervals (ANOVA, p = 0.684).

It is observed that PM10 concentrations are always slightly higher than or equal to PM2.5 in the different experimental groups. This pattern is consistent with the particle size classification of particulate matter, since the PM10 fraction comprises particles with an aerodynamic diameter of up to 10 µm, also covering the fine particles represented by PM2.5. Furthermore, it is important to emphasize that the presented particulate matter values have already been multiplied by the mentioned dilution factor.

IV. DISCUSSION

The purpose of this study was to develop an accessible exposure system, enabling the use of a mechanical structure and low-cost sensors. Data analysis demonstrates the project's

feasibility, as the detected temperature, humidity, MQ-7 sensor signal, and particulate matter (PM) levels confirm the system's viability for monitoring. It was found that, despite being low-cost components, the sensors were capable of capturing data on the experimental conditions.

The MQ-7 sensor selected operates based on changes in the electrical resistance of its chemical layer—composed of tin dioxide—whereby resistance decreases as the ambient CO concentration rises [12]. Consequently, the authors believe that the observed voltage variation values are consistent with the sensor's expected operation, given that these voltages increased upon the entry of cigarette smoke into the primary chamber.

It is worth noting that, although the MQ-7 sensor exhibits a good response to the presence of CO—especially considering its low cost—it also demonstrates documented cross-sensitivity to other substances such as hydrogen, methane, and alcohol compounds [12]. Consequently, since conventional cigarette smoke may contain volatile compounds of this nature [14], the contribution of these gases to the recorded signal cannot be entirely ruled out. Furthermore, MQ-7 relative sensor response was based on the voltage difference read by the microcontroller. Consequently, the values obtained should be interpreted as a relative sensor response rather than an absolute measure of CO concentration.

Regarding particulate matter, the selection of these two ranges—PM2.5 and PM10—aligns with WHO parameters, justified by the high clinical risk associated with the deposition of these fractions directly into the pulmonary alveoli [15]. Furthermore, the PM10 metric encompasses all particles with an aerodynamic diameter of 10 µm or less. Given that the median values for PM2.5 and PM10 were close to one another, it can be inferred that the majority of the particulate matter generated by the cigarette consisted of fine particles capable of penetrating the pulmonary alveoli more easily—a finding that is also consistent with the literature [16].

The sample size used in this study is relatively small, which reduces the statistical power of the tests applied. Therefore, the authors consider that, regarding the relative signal data from the MQ-7 and MP sensors—even though no statistically significant difference was observed for the 5, 10, and 15-second intervals between puffs—it is not possible to assert an absence of equivalence between the groups. It is possible that the air compressor with a flow rate of 1.6 L/min was insufficient to renew and recirculate the air within the chamber within these time intervals. Continuous air renewal is crucial to prevent the asphyxiation of experimental animals caused by CO accumulation; therefore, an airflow rate higher than 1.6 L/min might be more appropriate. Furthermore, it is recommended that future studies increase the sample size and employ intervals between puffs longer than those used in this work, in order to achieve greater control over the relative MQ-7 and PM signals, according to user needs.

V. Conclusion

The present work presented the development of a system that simultaneously integrates a low-cost exposure apparatus and monitoring of temperature, humidity, MQ-7 signal and PM concentrations, using accessible sensors. In this way, it is expected that the model will be replicable and that it will bring together research groups with limited resources, strengthening more accessible science.

Acknowledgment

Generative AI was used to assist with language editing and to support the conceptual development and visual refinement of the figures presented in this manuscript. All scientific content, analyses, interpretations, and the final versions of the figures and manuscript were reviewed, verified, and approved by the authors, who assume full responsibility for the content of this work.

References

[1] M. M. Cruz and M. Pereira, "Epidemiology of Chronic Obstructive Pulmonary Disease in Brazil: a systematic review and meta-analysis", Ciência & Saúde Coletiva, vol. 25, no. 11, pp. 4547–4557, Nov. 2020, doi: 10.1590/1413-812320202511.00222019.

[2] World Health Organization, "Chronic obstructive pulmonary disease (COPD): fact sheet," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd). Accessed: Jun. 8, 2026.

[3] World Health Organization, "Tobacco: fact sheet," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco. Accessed: Jun. 8, 2026.

[4] F. J. R. Paumgartten, M. R. Gomes-Carneiro, and A. C. A. X. de Oliveira, "O impacto dos aditivos do tabaco na toxicidade da fumaça do cigarro: uma avaliação crítica dos estudos patrocinados pela indústria do fumo," Cadernos de Saúde Pública, vol. 33, p. e00132415, 2017, doi: 10.1590/0102-311X00132415.

[5] C. Chung et al., "Fine particulate matter aggravates smoking induced lung injury via NLRP3/caspase-1 pathway in COPD," J. Inflamm. (Lond.), vol. 21, no. 1, p. 13, Apr. 2024, doi: 10.1186/s12950-024-00384-z.

[6] A. P. Barbosa et al., "The deleterious effects of smoking in bone mineralization and fibrillar matrix composition," Life Sci., vol. 241, art. no. 117132, 2020, doi: 10.1016/j.lfs.2019.117132.

[7] J. J. M. Junqueira et al., "Decreased bone type I collagen in the early stages of chronic obstructive pulmonary disease (COPD)," COPD: J. Chronic Obstr. Pulm. Dis., vol. 17, no. 5, pp. 575-586, 2020, doi: 10.1080/15412555.2020.1808605.

[8] K. K. Johnson, M. H. Bergin, A. G. Russell, and G. S. W. Hagler, "Field test of several low-cost particulate matter sensors in high and low concentration urban environments," Aerosol Air Qual. Res., vol. 18, no. 3, pp. 565–578, 2018, doi: 10.4209/aaqr.2017.10.0418.

[9] L. dos R. Izolan, S. Bandiera, R. R. Pulcinelli, M. S. Nin, F. B. Almeida, D. Marques, F. U. Fontella, R. F. de Almeida, M. B. Leal, and R. Gomez, "Low-cost apparatus for cigarette smoke exposure in rats," J. Neurosci. Methods, vol. 366, art. no. 109412, Jan. 2022, doi: 10.1016/j.jneumeth.2021.109412.

[10] Sensirion AG, SPS30 Particulate Matter Sensor for Air Quality Monitoring and Control – Datasheet, 2021. [Online]. Available: https://br.mouser.com/datasheet/2/682/Sensirion_PM_Sensors_Datasheet_SPS30-3539491.pdf. Accessed: Jun. 27, 2026.

[11] M. A. Sugianto and Z. Zulfikar, "Design of a carbon monoxide gas measurement tool using microcontroller-based MQ-7 gas sensor in households," NEWTON: Netw. Inf. Technol., vol. 3, no. 3, pp. 36–47, 2024, doi: 10.32764/newton.v3i3.4881.

[12] Hanwei Electronics Co., Ltd., MQ-7 Semiconductor Sensor for Carbon Monoxide – Datasheet, n.d. [Online]. Available: https://d229kd5ey79jzj.cloudfront.net/1304/MQ-7.pdf. Accessed: Jun. 8, 2026.

[13] J. Kuula, T. Mäkelä, M. Aurela, K. Teinilä, S. Varjonen, Ó. González, and H. Timonen, "Laboratory evaluation of particle-size selectivity of optical low-cost particulate matter sensors," Atmos. Meas. Tech., vol. 13, no. 5, pp. 2413–2423, 2020, doi: 10.5194/amt-13-2413-2020.

[14] Centers for Disease Control and Prevention (US), National Center for Chronic Disease Prevention and Health Promotion (US), and Office on Smoking and Health (US), How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General, ch. 3, "Chemistry and Toxicology of Cigarette Smoke and Biomarkers of Exposure and Harm." Atlanta, GA, USA: CDC, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK53014/. Accessed: Jul. 14, 2026.

[15] World Health Organization, WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. Geneva, Switzerland: WHO, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240034228. Accessed: Jun. 8, 2026.

[16] National Research Council (US) Committee on Passive Smoking, Environmental Tobacco Smoke: Measuring Exposures and Assessing Health Effects. Washington, DC, USA: National Academies Press, 1986, doi: 10.17226/943.

Development of a low-cost graphite-based bioelectrode for electromyographic signal acquisition and stimulation

Maria Clara Bernardes Ferrer
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
maria.ferrer@geb.inatel.br

Rita Elizabeth Santos de Almeida
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
r.elizabeth@inatel.br

Maysa Costa Alves
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

Elisa Rennó Carneiro Déster
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
elisa.renno@inatel.br

***Abstract*—The project aimed to develop a low-cost device for electromyographic signal acquisition and electrical stimulation to evaluate and stimulate muscles. The prototype structure was composed of Ecoflex™ 00-30 silicone, conventional electrode wires with female banana connectors, 3D-printed (PLA) components, and graphite as the conductive material. The pilot model was designed to evaluate the use of graphite combined with a polymeric material as an electrical conductor. The purpose was to assist physical therapists and other healthcare professionals in monitoring patients, with potential future applications in muscle monitoring and rehabilitation.**

***Index Terms*—electrical stimulation, electromyography, graphite, muscle, polymer.**

I. Introduction

Silicone is a highly flexible material that provides improved adaptability and usability to the electrode for the analysis of human bioelectrical signals. As a polymer, it typically exhibits biocompatible properties, causing neither skin irritation nor allergic reactions [1].

Graphite is a mineral that exhibits excellent electrical conductivity. It is essential for providing the electrical properties required for the development of the project proposed in this article. Furthermore, it is a chemically inert material, meaning that it is unlikely to undergo changes when in contact with other substances [2].

In physical therapy, Transcutaneous Electrical Nerve Stimulation (TENS) is performed by delivering low-voltage electrical stimuli through electrodes placed on the patient's skin. It is a safe and non-invasive method [3].

Furthermore, surface electrodes are also used to evaluate muscle electrical activity, that is, the muscle action potential, through electromyography (EMG). This technology has been increasingly explored and applied in a wide range of fields [4].

There is a wide variety of electrodes for different applications, with a specific type required for each use. They can be classified as surface electrodes, which are non-invasive and simply placed on the patient's skin, and implantable electrodes, which are inserted into the region of interest within the body [5].

Furthermore, electrodes can be further classified according to the base material used and their shape. Each type has a specific application. For example, among surface electrodes are metal plate electrodes (ECG, EMG, and EEG), flexible electrodes (wearable devices), suction electrodes (diagnostic ECG), and floating electrodes (biopotential recording). Among implantable electrodes, probe electrodes (body cavities), needle electrodes (intramuscular EMG), fine-wire electrodes (deep muscles), and microelectrodes (neural and cellular recording) are the most prominent [5].

In the book Medical Instrumentation: Application and Design, Chapter 5 – Biopotential Electrodes, Webster and Nimunkar emphasize that biopotential electrodes are responsible not only for conducting electrical current but also for converting the ionic conduction present in biological tissues into the electronic conduction used by measurement equipment [6].

Another important aspect to highlight is the need to reduce the manufacturing costs of these electrodes in order to increase production, simplify their fabrication, and make them more accessible and scalable. This also enables them to be disposable and intended for single use, thereby ensuring more reliable results [7].

Based on this challenge, the objective of this study was to develop a low-cost device intended for future electromyo-

Fundação Instituto Nacional de Telecomunicações (FINATEL).

graphic signal acquisition and electrical stimulation for the evaluation and stimulation of human body muscles, aiming to provide a non-invasive diagnosis and improve accessibility.

II. Methodology

Initially, a survey was conducted on different conductive, flexible materials and their possible combinations, aiming to develop a flexible electrode with reduced dimensions.

This device was required to exhibit sufficient electrical conductivity for the acquisition of muscle action potentials (electromyography) and the delivery of electrical pulses (electrical stimulation) to the muscles. Furthermore, its components needed to possess flexible characteristics to allow coupling and adaptation to different regions of the human body. These characteristics were considered desirable for potential future applications in muscle assessment.

In addition, the objective was to develop a low-cost electrode to improve accessibility for healthcare professionals and their respective patients.

To achieve this objective, research was conducted on the following materials:

- **Hydrogel:** an electrically responsive material; however, due to its consistency, its application would not be feasible;
- **Silver Fiber:** Ag/AgCl electrodes are the most commonly used, providing a more stable interface and relatively lower interference; however, they have a high cost;
- **Disposable conductive defibrillator pad:** used to assist the defibrillator in monitoring cardiac rhythm, enabling the identification of arrhythmias and cardiorespiratory arrest. However, it is a very rigid structure, making it unsuitable for the present application;
- **Disposable conductive return electrode for electrosurgical units:** its function is to provide coagulation and cauterization while ensuring patient safety through electrical current dispersion. However, it cannot be cut or modified in shape;
- **Conductive filament:** a conductive material for 3D printing; however, the specific requirements for proper printing would make the manufacturing process more difficult;
- **Flexible filament:** a flexible material for 3D printing; however, the specific requirements for proper printing would make the manufacturing process more difficult;
- **Silver chloride + silicone:** to make a polymer conductive, a dopant capable of promoting polymer reduction or oxidation is required. However, no specific studies were found involving the addition of silver chloride to silicone to make it conductive;
- **Graphene:** it has lower electrical resistivity than silver (the best electrical conductor among metals) and an electrical conductivity of approximately 2×10^6 S/m (siemens per meter). It also exhibits favorable biocompatibility; however, it requires a high investment;
- **Graphite:** it conducts electricity due to the free electrons present between the layers of carbon atoms and has an electrical conductivity of approximately 10^3 S/m (siemens per meter). Although it is less conductive than some metals (such as silver) and graphene, it is a low-cost material.;
- **Silicone + graphite**: according to the supplier, contaminants that may interfere with silicone curing include sulfur (latex), tin-based catalysts, and epoxy resin hardeners. Therefore, graphite does not interfere with the silicone curing process.

Furthermore, electrical insulation between the healthcare professional's finger and the electrode was required to prevent the electromyograph from acquiring the professional's muscle action potentials. To achieve this insulation, it was necessary to identify a material that was both electrically insulating and flexible, allowing the finger to move as required.

In this regard, suitable materials would include polymers and silicone, allowing the entire finger to be covered. Another alternative would be to coat the side of the sensor that comes into contact with the finger with an electrically insulating material, such as rubber. In this case, material flexibility would not be required, since the entire finger would not be covered by the insulating material.

For the electrode fabrication, molds made of alginate, EVA foam, and 3D-printed PLA were tested to determine which would provide the most uniform and suitable results.

A. Cost Analysis

A cost analysis of the electrodes developed in this study was conducted, considering exclusively the materials used in their fabrication. Costs related to labor, electricity consumption, equipment depreciation, and other indirect manufacturing expenses were not included. In addition, whenever possible, the obtained costs were compared with those of commercially available conductive electrodes.

All prices correspond to values obtained in July 2026.

Mold fabrication, failed batches, labor, electricity consumption, equipment depreciation, and other indirect manufacturing costs were not included.

For the smallest electrode, approximately 0.5 g of graphite was used per unit. Considering an approximate cost of R$40.00 per kilogram of graphite, it was estimated that 2,000 electrodes could be manufactured, resulting in a graphite cost of R$0.02 per unit.

Regarding the silicone, approximately 2/3 g of component A and 2/3 g of component B were used per electrode, totaling 4/3 g of silicone per unit. Considering that 1 kg of silicone (500 g of component A and 500 g of component B) cost R$332.80, it was estimated that approximately 750 electrodes could be manufactured, resulting in a silicone cost of R$0.44 per unit.

The 3D-printed structure cost R$50.98 to manufacture 50 units, corresponding to a unit cost of R$1.02.

In addition, a conventional electrode wire with a female connector was used. Based on the cost of the material employed (conductive carbon fiber heating wire), the estimated cost was approximately R$0.05 per electrode, considering a 100 m

spool priced at R$50.00 and a consumption of approximately 10 cm per unit, allowing the manufacture of 1,000 electrodes.

Thus, the estimated raw-material cost per unit of the smallest electrode was approximately R$1.53 per unit. No commercially available conductive silicone electrodes with equivalent dimensions were identified, making a direct cost comparison impossible.

For the largest electrode, approximately 2.85 g of graphite was used per unit. Considering an approximate cost of R$40.00 per kilogram of graphite, it was estimated that 351 electrodes could be manufactured, resulting in a graphite cost of R$0.16 per unit.

For the fabrication of this electrode, approximately 4.28 g of silicone component A and 4.28 g of silicone component B were used per unit, totaling 8.57 g of silicone. Considering a cost of R$332.80 per kilogram of silicone (500 g of each component), it was estimated that approximately 117 electrodes could be manufactured, corresponding to a silicone cost of R$2.84 per unit.

A conventional electrode wire with a female connector was also used, with the estimated cost remaining at R$0.05 per unit, based on the same criteria described previously.

Thus, the estimated raw-material cost per unit of the largest electrode was approximately R$3.05 per unit. For comparison purposes, a commercially available conductive silicone electrode was found to have a unit cost of R$9.23.

III. Results and Discussion

The feasibility of this project was achieved through the use of the following materials:

- **Graphite powder (Emavi®):** the material responsible for providing the electrical conductivity required for the electrode;
- **Ecoflex™ 00-30:** an extra-soft, platinum-cure, high-performance biocompatible silicone rubber used to provide the flexibility and malleability essential to the device;
- **Electrical conductors:** conventional electrode wires (conductive carbon fiber heating wires) with female banana connectors, enabling the integration of the bioelectrode with the electromyograph or the electrical stimulator;
- **3D-printed (PLA) components:** biocompatible polylactic acid (PLA) polymer structures used to shape and support the wearable device.

After defining the materials required for the device fabrication, the next step consisted of developing the physical structure of the electrode. The geometric characteristics of the prototype were designed considering both the anatomical adaptation and the requirements for incorporating the conductive composite.

The 3D models were designed using the *Autodesk Tinkercad* online platform. For the physical fabrication, a configuration with two walls and 40% infill density was adopted to optimize the mechanical strength of the printed parts.

Different molding approaches were tested to identify the most suitable option for the present project. Therefore, different types of materials were evaluated.

Initially, alginate (a material widely used in dentistry for dental impressions) was employed to produce the mold, using a part manufactured by additive manufacturing (3D printing) to shape it, as shown in Figure 1.



Fig. 1. Alginate mold. Source: Author (2025).

However, it was observed that the alginate dried out very quickly, requiring a new mold to be prepared before each test, which would be impractical for the proposed project.

Therefore, an EVA foam mold was developed, also using a part manufactured by additive manufacturing to cut the material (Figure 2), as well as another 3D-printed part placed inside the electrode to provide structural support.

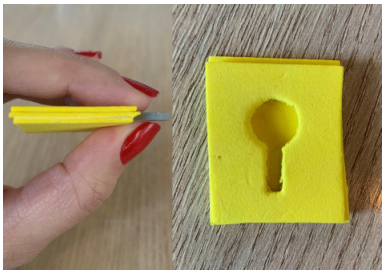


Fig. 2. EVA foam mold. Source: Author (2025).

However, it was evident that, after the silicone had cured in this mold, the resulting electrode did not exhibit the desired appearance (Figure 3). In addition, the 3D-printed PLA part detached easily from the silicone, indicating that PLA was more suitable as a mold material.

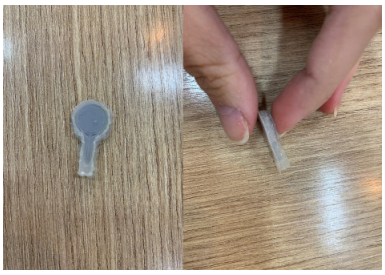


Fig. 3. Pure silicone molded in EVA foam. Source: Author (2025).

Thus, initially, an external mold was designed for the curing of silicone containing graphite, in which the silicone, prior to

hardening, would be deposited, along with an internal support structure with a similar shape but reduced dimensions. This internal structure aims to ensure the mechanical integrity of the electrode and the proper positioning of the conductive components inside the device. These initial test prototypes can be observed in Figures 4 e 5.

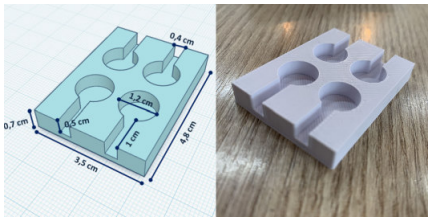


Fig. 4. Initial mold prototype and its dimensions. Source: Author (2025).

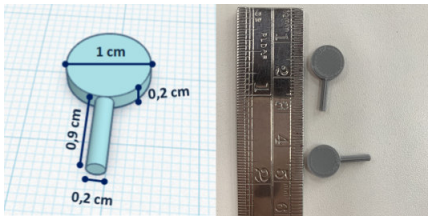


Fig. 5. Initial structure prototype and its dimensions. Source: Author (2025).

In addition to the material tests for modeling, tests regarding the electrode composition were also performed.

Regarding silicone curing and its homogenization with graphite, numerous tests were performed using different percentage concentrations to determine the most suitable ratio. The spreadsheet software *Microsoft Excel* was used to organize and compare the data obtained from each measurement (Table I), in which G represents graphite and S represents silicone.

TABLE I
COMPOSITION OF GRAPHITE SAMPLES. THE FORMULATION IDENTIFIED AS SAMPLE 17 (59.47% GRAPHITE) WAS SELECTED FOR FABRICATION OF THE FINAL PROTOTYPES.

Sample	G (g)	G (%)	S (g)	S (%)	T (g)
1	0.0103	0.95	1.0748	99.05	1.0851
2	0.3968	27.30	1.0566	72.70	1.4534
3	0.4282	29.87	1.0055	70.13	1.4337
4	0.6916	40.07	1.0343	59.93	1.7259
5	0.7111	41.32	1.0100	58.68	1.7211
6	0.7165	41.32	1.0175	58.68	1.7340
7	0.9293	45.89	1.0956	54.11	2.0249
8	1.0225	49.37	1.0486	50.63	2.0711
9	2.0017	49.90	2.0098	50.10	4.0115
10	1.0536	49.91	1.0572	50.09	2.1108
11	1.1999	54.24	1.0122	45.76	2.2121
12	2.0066	57.14	1.5053	42.86	3.5119
13	2.0115	57.15	1.5080	42.85	3.5195
14	1.5039	58.83	1.0523	41.17	2.5562
15	1.5151	59.08	1.0495	40.92	2.5646
16	3.0135	59.26	2.0717	40.74	5.0852
17	1.5030	59.47	1.0244	40.53	2.5274

Based on the preliminary electrical evaluation, the formulation containing 59.47% graphite by mass was selected for the fabrication of the final prototypes. This composition was chosen because it provided the most suitable electrical response among the tested formulations under the experimental conditions.

For the preliminary evaluation of the graphite concentration, three small electrode specimens were fabricated for each tested concentration. The formulation was selected based on the highest measured peak voltage and its numerical proximity to the commercial reference under the same bench conditions was selected for subsequent evaluation. Six additional small electrode specimens were then fabricated using the selected concentration for further characterization and statistical analysis. Since the optimal graphite concentration had already been determined from the preliminary tests with the small electrodes, six large electrode specimens were fabricated exclusively using the selected concentration.

From this perspective, it is important to emphasize that the differences in mass for each experiment were justified not only by the tested concentrations but also by the number of electrodes produced in a single process.

In this context, it was observed that the higher the graphite concentration used in the development of the device, the better the achieved result, meaning that the measured peak voltage increased with graphite concentration under the evaluated bench conditions.

Initially, the graphite and silicone mixture was applied throughout the entire electrode surface, as illustrated in Figure 6, resulting in a fully conductive device.

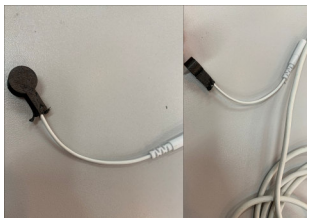


Fig. 6. First graphite-silicone electrode: front and side views, respectively. Source: Author (2025).

However, it was experimentally observed that thinner layers of graphite-doped silicone exhibited a more favorable electrical response under the evaluated bench conditions. This observation cannot be directly attributed to a reduction in electrical resistance based solely on the conductive-layer thickness. According to the resistance relationship, the electrical resistance (R) of a conductor depends on its resistivity (ρ), the effective length of the current path (L), and its cross-sectional area (A):

$$R = \rho \cdot \frac{L}{A} \tag{1}$$

In this equation, ρ represents the electrical resistivity of the material, L is the effective current path length, and A is the corresponding cross-sectional area. Therefore, the effect of reducing the conductive-layer thickness depends on the actual geometry and direction of current flow within the

electrode. Since the current path and the resistivity of the composite were not directly characterized in the present study, the improved electrical response observed for the thinner layer is reported as an experimental observation rather than as a direct consequence of reduced electrical resistance.

Subsequently, a new prototype was fabricated, consisting primarily of pure silicone intended for electrical insulation and structural support, with the addition of only a thin layer of graphite-doped silicone (Figure 7), aiming to optimize contact with the tissue and improve signal conduction efficiency.



Fig. 7. Prototype of the pure silicone mold: front and side views, respectively. Source: Author (2025).

Subsequently, a new mold with larger dimensions and a quadrangular shape was developed to validate the use of graphite-doped silicone in electrodes with different shapes and sizes (Figure 8).

It is noteworthy that the use of the internal support structure manufactured by additive manufacturing became unnecessary for this new mold.

Comparative tests were performed with and without the use of this component. The results did not demonstrate significant improvements when the component was used, making its application unnecessary.

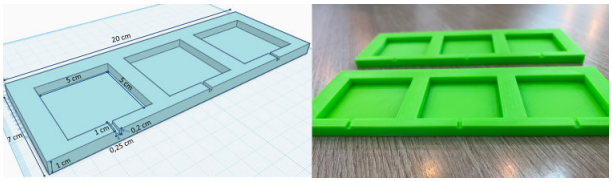


Fig. 8. Square mold prototype and its dimensions. Source: Author (2026).

The final configurations obtained during the prototype development process are summarized in Table II. The table presents the main physical and fabrication parameters of the small and large electrode prototypes, including their dimensions, conductive-layer thickness, graphite concentration, and curing time.

TABLE II
FINAL SPECIFICATIONS OF THE DEVELOPED ELECTRODE PROTOTYPES

Parameter	Small Electrode	Large Electrode
Electrode geometry	⌀12 mm circular region + 10 mm extension	50 × 50 mm
Electrode thickness (mm)	∼5	∼5
Conductive layer thickness (mm)	1	1
Graphite concentration (%)	59.47	59.47
Curing time (h)	12	12

Furthermore, a comparison was performed between the curing of graphite-doped silicone exposed to UV-C light and without UV-C exposure. For four days, measurements were performed by weighing the amount of graphite particles released through the application of a piece of masking tape on the electrodes. The results were recorded in a spreadsheet, and a graph was generated (Figure 9) using the spreadsheet software *Microsoft Excel*.

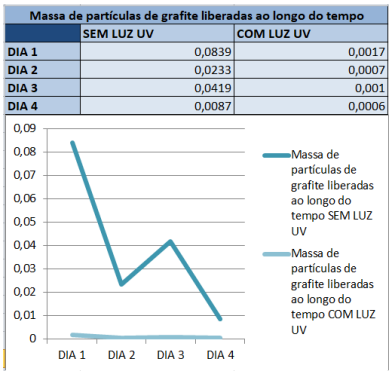


Fig. 9. Comparison graph (UV-C light). Source: Author (2026).

When applied to the electrodes with reduced dimensions, the result was positive, with lower graphite release observed in the electrodes exposed to light.

However, when applied to the larger electrodes, they exhibited similar graphite particle release when exposed and not exposed to UV-C light (slightly lower for the exposed samples). Nevertheless, the electrodes exposed to UV-C light developed surface cracks (Figure 10).

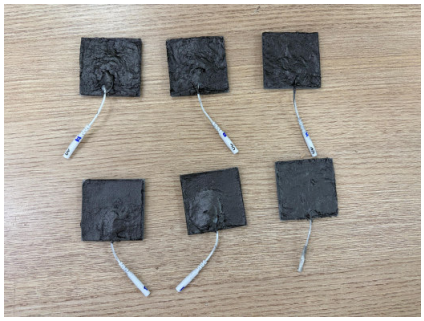


Fig. 10. Quadrangular electrodes: UV-C-exposed electrodes at the top and non-exposed electrodes at the bottom. Source: Author (2026).

Although platinum-cured silicone is not designed to cure under UV-C radiation, UV-C post-curing exposure was associated with reduced graphite particle release in the smaller electrodes under the evaluated conditions. This observation suggests a possible improvement in surface particle retention; however, the underlying mechanism was not directly characterized in the present study.

The samples were exposed to UV-C radiation using the UV-C lamp integrated into a Pachane PCR T2 ECO vertical laminar flow cabinet. The lamp emitted radiation with a wavelength of approximately 254 nm, within the UV-C spectral range. The

irradiance at the sample surface was not measured; therefore, the UV-C dose received by the samples could not be quantified.

The homogenization and curing process of the conductive composite was performed according to the following steps:

1) **Material Preparation:** separation of graphite, silicone rubber (Parts A and B), electrical conductors with female banana connectors, 3D-printed molds and structures, hot glue, beakers, spatulas, and precision balance;
2) **Conductor Preparation:** adjustment of the conductive wire length and stripping of the ends to ensure electrical contact;
3) **Structural Fixation:** positioning of the conductors in the 3D-printed internal structure using hot glue to ensure stability, when applicable (reduced-dimension electrode);
4) **Silicone Preparation:** individual homogenization of Parts A and B. Subsequently, equivalent masses were weighed (1:1 ratio) in a previously tared beaker, followed by mixing for three minutes and subsequent deposition into the external mold;
5) **Assembly of the Set:** positioning of the internal structure (with the wire) over the still slightly liquid pure silicone layer;
6) **New Silicone Preparation:** preparation of a new pure silicone sample (Parts A and B in a 1:1 ratio) for subsequent graphite doping;
7) **Graphite Doping:** addition of graphite (mass calculated according to the desired percentage) to the silicone mixture, followed by homogenization for two minutes to ensure particle dispersion;
8) **Conductive Layer Positioning:** deposition of the silicone/graphite mixture over the internal structure with the wire facing upward;
9) **Curing and Finalization:** resting of the assembly for 12 hours to allow complete polymer crosslinking at room temperature, followed by careful demolding of the electrode (square-shaped);
10) **Post-curing UV-C Light Exposure:** the assembly was exposed to UV-C light for 96 hours to evaluate its effect on graphite particle adhesion and release, followed by careful demolding of the electrode.

After the electrode modeling process was completed and the device was manufactured, preliminary electrical tests were conducted to characterize the prototype under the evaluated bench conditions.

Voltage measurements were performed using an oscilloscope, an electrical stimulator, and a 330 Ω load resistor, allowing a direct comparison with the voltage levels obtained from a commercial conductive rubber electrode.

The peak voltage values generated by each electrode were recorded in the spreadsheet software *Microsoft Excel* for subsequent analysis (Table III).

In the table, the samples correspond to the electrodes developed and tested, Electrode refers to the conventional electrode (commercially available), Voltage represents the peak voltage measurement in volts, and G(%) indicates the graphite concentration used in each sample.

From this perspective, it can be observed that the higher the graphite concentration, the closer the developed electrode becomes, in terms of electrical performance, to the conventional electrode. Except for some cases in which improper positioning during fabrication resulted in undesired conductivity.

TABLE III
VOLTAGE BY GRAPHITE SAMPLE COMPOSITION.

Sample	G (%)	Voltage 1 (V)	Voltage 2 (V)	Voltage 3 (V)
Electrode	-	14.8	14.8	14.8
1	0.95	0.0	0.0	0.0
2	27.30	0.7	0.6	0.6
3	29.87	0.0	0.0	0.0
4	40.07	0.2	0.2	0.3
5	41.32	0.0	0.0	0.0
6	41.32	0.0	0.0	0.0
7	45.89	0.0	0.2	0.0
8	49.37	0.8	1.6	1.4
9	49.90	2.7	2.5	2.9
10	49.91	2.4	2.2	1.8
11	54.24	14.6	14.6	14.4
12	57.14	14.6	13.0	14.4
13	57.15	14.8	14.6	14.6
14	58.83	14.8	14.4	14.4
15	59.08	14.8	14.6	14.6
16	59.26	14.8	14.6	14.8
17	59.47	14.8	14.8	14.8

The experimental setup used for the measurement process is shown in Figure 11.

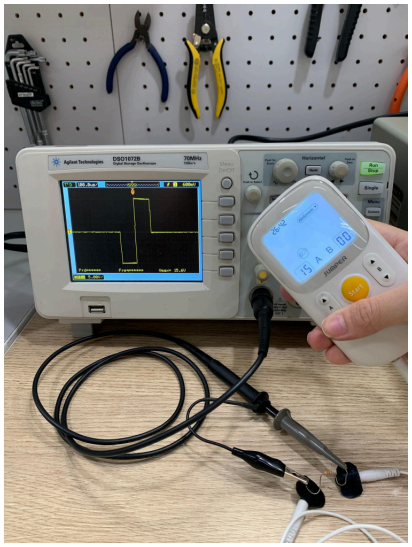


Fig. 11. Oscilloscope measurement. Source: Author (2025).

From this perspective, the prototype — composed of pure silicone with a thin layer of graphite-doped silicone, molded using the additive manufacturing mold (Figures 4 and 8) and supported by the internal structure (Figure 5) — exhibited measurable electrical conduction under the evaluated bench conditions.

The performance surpassed the results obtained in the initial tests with lower graphite concentrations and produced peak

voltage values comparable to those obtained with the commercial reference electrode under the evaluated bench conditions.

It is important to emphasize that the final electrode prototypes were not subjected to experimental biocompatibility tests in the present study. Although Ecoflex™ 00-30 and PLA were selected based on their material properties and suitability for prototype fabrication, the biocompatibility of the complete fabricated electrode, including the graphite-silicone conductive composite and the assembled structure, was not experimentally established. Therefore, no claim regarding the biocompatibility or safety of the final electrode for human use can be made based on the results presented herein.

IV. Conclusion

The development of a low-cost conductive electrode prototype based on graphite and silicone enabled preliminary electrical characterization under controlled bench conditions, using graphite as the conductive material, enabled preliminary electrical characterization under controlled bench conditions. It should be emphasized that the device has not yet been tested in humans, and the current results are limited to preliminary electrical validation.

The silicone rubber provided the necessary flexibility for anatomical positioning, in addition to enabling the electrical insulation of healthcare professionals, optimizing the functionality and application of the bioelectrode.

UV-C post-curing exposure was associated with reduced graphite particle release in the smaller electrodes under the evaluated conditions, whereas the larger electrodes exhibited surface cracking after exposure.

The objective of the project was to develop a low-cost conductive electrode prototype for potential future applications in electromyographic signal acquisition and electrical stimulation.

V. Future Studies

Future work will include real EMG signal acquisition, electrical stimulation tests in biological tissue, biocompatibility assessment, and *in vivo* tests in rats. Subsequently, clinical trials with humans may be conducted to further evaluate the potential application of the device in muscle assessment and rehabilitation.

Additional studies should also investigate the electrode–skin interface impedance, frequency response, temporal stability, signal-to-noise ratio, and repeatability among different fabricated devices. Direct measurements of electrical resistance or resistivity should also be performed to provide a more comprehensive characterization of the conductive composite and its electrical behavior.

Further evaluation should include mechanical-cycle testing, assessment of material durability, irritation and cleaning procedures, as well as tests involving volunteers before considering potential clinical applications. Future studies should also investigate the electrical current delivered to biological tissue during stimulation and evaluate the long-term stability of the electrode under repeated use.

A more comprehensive statistical characterization of the electrical performance and repeatability of the fabricated electrodes will be performed in future work, including the analysis of variability among specimens and additional statistical parameters.

Finally, the fabrication process may be further optimized by evaluating the reproducibility of the selected graphite concentration, conductive-layer geometry and thickness, and the effect of UV-C exposure on graphite particle release and material stability. These additional investigations are necessary to advance the present prototype from preliminary electrical validation toward a more comprehensive evaluation of its performance and potential biomedical application.

Acknowledgment

This work was carried out with the support of the Fundação Instituto Nacional de Telecomunicações - FINATEL, the Instituto Nacional de Telecomunicações - Inatel and the eHealth - Innovation Center Laboratory, which diligently provided the necessary resources for the development of the project.

To Professor Filipe Bueno Vilela, Ph.D., for providing essential research contributions for the development of the artifact.

To Professor Luma Rissati Borges do Prado, for indicating the electrode testing method using the oscilloscope.

To Professor Maysa Costa Alves, Ph.D., for suggesting the use of UV-C light for silicone curing.

To the advisor, Professor Elisa Rennó Carneiro Déster, M.Sc., for all the support, ideas, and knowledge shared.

References

[1] C.-H. Han, S.-U. Kim, K.-S. Lim, Y.-J. Jung, S. Lee, S. H. Kim, and H.-J. Hwang, "Evaluation of a silicone-based flexible dry electrode for measuring human bioelectrical signals," Biomed. Eng. Lett., vol. 15, no. 3, pp. 563–574, May 2025, doi: 10.1007/s13534-025-00471-x.

[2] M. A. Santos, "Desenvolvimento de compósitos de LDPE/grafite para potencial uso em embalagens antiestáticas," Bachelor's thesis, Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, SP, Brazil, 2019.

[3] C. C. S. dos Reis, L. C. Dias, L. B. Carvalho, L. B. Alves Junior, and A. M. Imoto, "Transcutaneous Nerve Electrostimulation (TENS) in Pain Relief During Labor: A Scope Review," Rev. Bras. Ginecol. Obstet., vol. 44, no. 2, pp. 187–193, Feb. 2022, doi: 10.1055/s-0042-1742290.

[4] V. Alcan and M. Zinnuroğlu, "Current developments in surface electromyography," Turk. J. Med. Sci., vol. 53, no. 5, pp. 1019–1031, Oct. 2023, doi: 10.55730/1300-0144.5667.

[5] M. D. F. de Sousa, D. M. B. de Sousa, G. C. S. dos Santos, L. A. de Almeida, L. C. de Sousa, M. A. M. da Silva, and M. M. de Sousa, "Electrotherapy in the Treatment of Children with Neurogenic Bladder: A Systematic Review," Children, vol. 8, no. 8, Art. no. 714, Aug. 2021, doi: 10.3390/children8080714.

[6] J. G. Webster and A. J. Nimunkar, Eds., *Medical Instrumentation: Application and Design*, 5th ed. Hoboken, NJ, USA: Wiley, 2020.

[7] N. Moukri, B. Patella, and R. Inguanta, "Recent Advances in Low-Cost Methods for Electrochemical Sensor Fabrication," Curr. Opin. Electrochem., vol. 51, Art. no. 101771, 2025, doi: 10.1016/j.coelec.2025.101771.

Development of a Low-Cost Portable Multiparameter Monitor Based on ESP32 for Real Time Vital Signals Monitoring

1st Patrick da Silva Santos
Undergraduate Student in Electronic Engineering
University of Brasilia
Biomedical Equipment Technician
Assistive Technology Access and Research Center (CAPTA)
Federal Institute of Brasilia (IFB), Campus Ceilandia
Brasilia, Brazil
patrick.santos@ieee.org

2nd Roberta Viana da Silva
Biomedical Equipment Technician
Federal Institute of Brasilia (IFB), Campus Ceilandia
Brasilia, Brazil
rvianadasilva@gmail.com

3rd Alessandro Pinheiro
Assistive Technology Access and Research Center (CAPTA)
Federal Institute of Brasilia (IFB), Campus Ceilandia
Faculty of Physical Education, University of Brasilia
Brasilia, Brazil
alessandro@ieee.org

4th Rhaíra Helena Caetano e Souza
Assistive Technology Access and Research Center (CAPTA)
Federal Institute of Brasilia (IFB), Campus Ceilandia
Brasilia, Brazil
rhaira.souza@ifb.edu.br

***Abstract*—This paper presents a low-cost IoMT laboratory prototype based on an ESP32 that displays ECG, SpO_2/heart rate, and non-contact temperature on a local Nextion HMI and a Progressive Web App (PWA). The sensing chain uses an AD8232 (ECG), a MAX30102 (PPG/SpO_2 and heart rate), and an MLX90614 (infrared temperature). Work started with software/simulation checks of the data path and then moved to a physical board. In SoftAP (software-enabled access point) laboratory demonstrations on an iPhone 11, Wi-Fi JSON streaming to the PWA was adequate for teaching; Table III records what passed, what was only partial, and what failed. Datasheet figures such as MLX90614 $\pm$0.5°C are design references, not measured system accuracy. Parts cost 140 dollars (Table I). The platform is intended for teaching and demonstration, not clinical use: no volunteer study, ethics protocol, or gold-standard comparison is reported, and the firmware still shows occasional freezes.**

***Index Terms*—Telemedicine, IoMT, ESP32, SpO2, Vital Signals**

I. Introduction

Chronic diseases remain major causes of death worldwide, and continuous vital-sign monitoring is one tool discussed for preventing complications [1]. Commercial multiparameter monitors work well in hospitals, but their price often keeps them out of teaching labs and other low-budget settings. This paper describes a compact ESP32-based sensing setup intended for laboratory use and clinical-engineering teaching—not for clinical care [2], [3], [4].

Supported by Distrito Federal Research Support Foundation (FAPDF), Process n. 23098.000264.2026-74 and Brazilian Ministry of Science, Technology and Innovation (MCTI), TED 973460, Process 01245.019107/2024-05.

IoMT papers often link home care with wireless vital-sign streams [5]. This literature provides useful context, but it is not validated here. An AD8232 (ECG), a MAX30102 (heart rate and SpO_2), and an MLX90614 (temperature) were connected to an ESP32, and the readings were sent to a Progressive Web App (PWA) and a Nextion panel. Abdullah et al. [6] and Islam et al. [7] discuss filtering and feature extraction on biomedical signals; the present work borrows the general idea of embedded conditioning, without claiming clinical-grade pipelines. Before hardware assembly, the data path was exercised in software simulation [4]—a preliminary software verification, not a metrology exercise.

Overall, the paper documents a low-cost ESP32/PWA teaching prototype and qualitative bench results. Patient benefit, institutional rollout, and clinical accuracy are not claimed.

II. Materials and Methods

A laboratory prototype was assembled to acquire ECG, SpO_2/heart rate, and temperature and to display them on-site and on a phone. The AD8232 (Analog Devices, Wilmington, MA, USA) [8] amplifies electrode ECG for demonstration traces [3]. Temperature comes from the MLX90614 (Melexis, Ieper, Belgium) [9] over I^2C; the $\pm$0.5°C figure is from the datasheet only [4].

Peripheral oxygen saturation (SpO_2) and heart rate use the MAX30102 (Maxim Integrated, San Jose, CA, USA) [10], a common PPG module with red/infrared LEDs [4]. The ESP32 runs acquisition, lightweight filtering, and the Wi-Fi link to the PWA and the Nextion HMI (Itead Studio, Shenzhen, China),

TABLE I
UNIT-COST BREAKDOWN (USD) COMPOSING THE REPORTED TOTAL OF 140 DOLLARS FOR SENSORS, NEXTION 2.4" HMI, AND ASSEMBLY EXTRAS (RETAIL/EDUCATIONAL-MODULE PRICES).

Item	USD
ESP32 DEVKIT V1	8
AD8232 ECG module + electrodes (kit)	20
MAX30102 pulse-oximetry module	12
MLX90614 IR temperature module	15
Nextion 2.4" HMI display	45
5 V buzzer, wiring, protoboard, enclosure extras	25
USB power / miscellaneous	15
Total	**140**

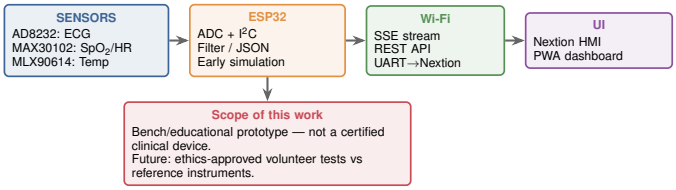


Fig. 1. Simplified system architecture (left-to-right data flow). Sensors feed the ESP32; Wi-Fi serves the PWA; UART drives the Nextion HMI. The bottom box states the intended scope of this paper.

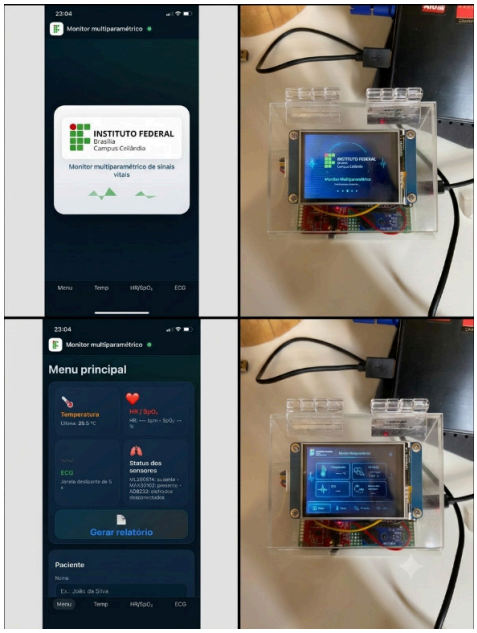


Fig. 2. Laboratory prototype: Nextion 2.4" HMI (right) and PWA dashboard on an iPhone 11 SoftAP demonstration (left). Image documents the setup; it does not imply clinical validation.

whether the source data are software-simulated or from the board.

Bench work focused on noise awareness, shared grounds, and simple checks; IEC 60601 certification or any medical-device clearance was not pursued. Fig. 1 sketches the data flow. The build is intended for teaching and laboratory demonstration, not for delivering a telemedicine service [11], [1]. Parts cost is 140 dollars (Table I).

A. System Overview

The prototype consists of: ESP32 for connectivity and lightweight processing; AD8232 ECG front-end; MAX30102 for SpO_2 and heart rate; and MLX90614 for non-contact temperature; plus a Nextion 2.4" HMI (Itead Studio, Shenzhen, China).

The ESP32 reads analog ECG from the AD8232 (GPIO34), combines SpO_2/heart rate from the MAX30102 (I^2C) and temperature from the MLX90614 (shared I^2C bus), and streams values over Wi-Fi for display. Fig. 2 shows the apparatus used in laboratory demonstrations.

B. Proposed System Architecture

The ESP32 was selected for integrated Wi-Fi/Bluetooth and comparatively larger on-chip resources than simpler microcontroller boards such as the Arduino R4. The architecture is organized into three layers—signal acquisition, processing, and communication:

Signal Acquisition: Sensors use the I^2C bus (MLX90614 and MAX30102) and a dedicated ADC input (Fig. 3 and Table II): GPIO 21 (SDA), GPIO 22 (SCL), AD8232 OUTPUT on GPIO 34, and lead-off on GPIO 32 and GPIO 26.

Processing: Firmware applies basic DSP on ECG sampled at a design rate of 250 Hz, with 5 s windows, simple moving-average or FIR filtering, and JSON serialization. R-peak / beep (prototype concept only): within each ECG window the signal is band-limited; candidate peaks are local maxima above an adaptive amplitude threshold separated by a refractory interval; each accepted peak triggers a short GPIO pulse to the buzzer. Exact thresholds, refractory constants, and firmware version identifiers were not retained for quantitative re-analysis, so the $\leq$100 ms beep-latency target remains an engineering goal rather than a measured result.

Communication: Resource separation on the ESP32 is by peripheral domain, not by sharing the same pins: (i) I^2C + ADC for sensors; (ii) UART0 (GPIO1/3) for the Nextion HMI; (iii) Wi-Fi SoftAP HTTP for SSE (continuous PWA events) and REST (short request/response for demographics and last readings). SSE and REST share the HTTP stack but use distinct routes; UART does not contend for Wi-Fi GPIO. Contention risk is CPU/time-sharing in the cooperative loop that interleaves sensor reads, Nextion updates, and network replies—consistent with intermittent freezes under SSE load (Table III). A forehead-aiming laser module was not included; optical safety classification and clinical necessity were not established.

The PWA provides the remote dashboard and can run as a standalone app on devices such as an iPhone 11 (Apple, Cupertino, CA, USA) in SoftAP laboratory demonstrations.

C. User Protection

High-impedance inputs and common-mode rejection in the ECG front-end are intended to support safer bench handling

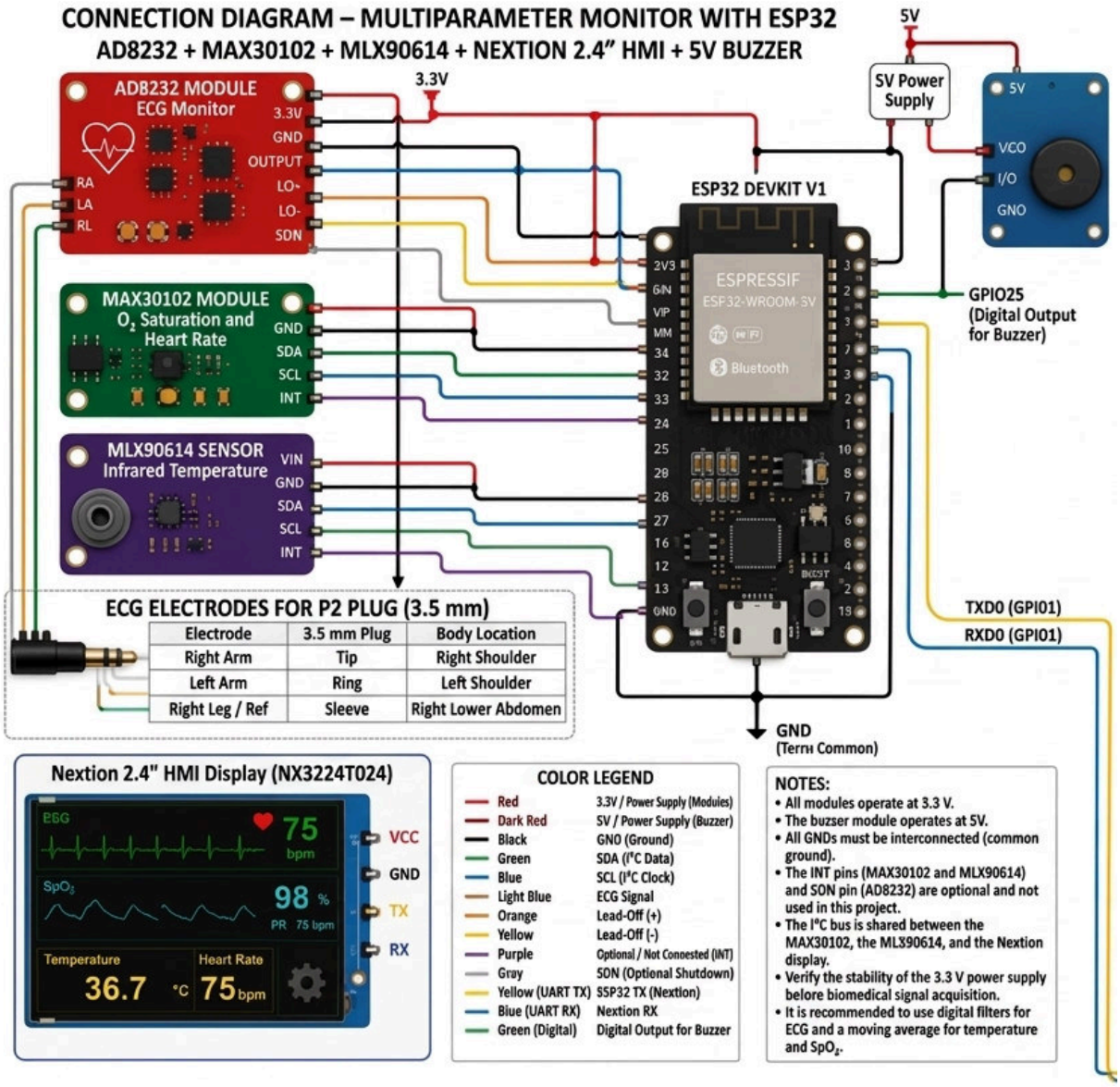


Fig. 3. Connection diagram (enlarged for readability): AD8232 (ECG), MAX30102 (SpO_2/HR), MLX90614 (temperature), Nextion 2.4" HMI, and buzzer on the ESP32 DEVKIT V1. Wire colors match Table II.

and reduce coupling of interference. This design note does not constitute electrical-safety certification (e.g., IEC 60601) or medical-device clearance.

III. Results and Discussion

Three layers of evidence are kept apart:

1) **Software simulation:** early virtual checks of the data path before assembly.
2) **Technical bench tests:** outcomes observed in laboratory demonstrations (Table III).
3) **Human / clinical evaluation:** *not performed*—no ethics protocol, no volunteers, no calibrated clinical reference.

A. Bench outcomes versus acceptance criteria

Table III is a qualitative status summary against the intended targets. "Partial" means the feature worked only sometimes, or missed a numeric goal. Mean±SD error and failure-rate percentages are not reported.

B. Procedures used in development

Heart rate and SpO_2 were checked with a finger on the MAX30102 while Nextion and PWA fields were observed during SoftAP laboratory demonstrations. ECG used three electrodes in a bench demonstration setup (not an ethics-approved study), with a 5 s sliding window on both displays. Temperature used the UI "Measure" action; after disconnecting the MLX90614 and rebooting, the last value retained in NVS

TABLE II

ESP32 PIN MAP (LARGER TYPE). POWER/GND AND SIGNAL LINES PER MODULE; COLORED DASHES FOLLOW THE WIRING LEGEND OF FIG. 3.

(a) AD8232 MODULE – ESP32

Pin	ESP32 Pin	Function
VIN	3.3 V	Power Supply
GND	GND	Ground Reference
OUTPUT	GPIO34	ECG Analog Signal
LO+	GPIO32	Loose Electrode Detection (+)
LO–	GPIO26	Loose Electrode Detection (–)
SDN	Not Connected	Shutdown (Optional)

(b) MAX30102 MODULE – ESP32

Pin	ESP32 Pin	Function
VIN	3.3 V	Power Supply
GND	GND	Ground Reference
SDA	GPIO21	I^2C Data Line
SCL	GPIO22	I^2C Clock Line
INT	Not Connected	Interrupt (Optional)

(c) MLX90614 SENSOR – ESP32

Pin	ESP32 Pin	Function
VIN	3.3 V	Power Supply
GND	GND	Ground Reference
SDA	GPIO21	I^2C Data Line
SCL	GPIO22	I^2C Clock Line
INT	Not Connected	Interrupt (Optional)

(d) NEXTION 2.4" HMI – ESP32 (UART)

Pin	ESP32 Pin	Function
VCC	3.3 V	Power Supply
GND	GND	Ground Reference
RX	TXD0 (GPIO1)	Nextion Reception
TX	RXD0 (GPIO3)	Nextion Transmission

(e) 5 V BUZZER – ESP32

Pin	ESP32 Pin	Function
VCC	5 V	Power Supply
I/O	GPIO25	Digital Output (Buzzer Activation)
GND	GND	Ground Reference

(f) ESP32 GPIO CONNECTIONS SUMMARY

ESP32 GPIO	Function	Type	Connection
GPIO34	ECG Signal Input	Analog	AD8232 OUTPUT
GPIO32	Loose Electrode Detection (+)	Digital	AD8232 LO+
GPIO26	Loose Electrode Detection (–)	Digital	AD8232 LO–
GPIO21	I^2C Data	I^2C	MAX30102 SDA, MLX90614 SDA
GPIO22	I^2C Clock	I^2C	MAX30102 SCL, MLX90614 SCL
GPIO25	Buzzer Output	Digital	5 V Buzzer I/O
GPIO1 (TXD0)	UART Transmission	Digital	Nextion RX
GPIO3 (RXD0)	UART Reception	Digital	Nextion TX

was verified. The intended latency method (timing logs not available for this report) would use 240 fps video or a dual-channel oscilloscope,

$$Delay\ (ms) = \left(\frac{\Delta frames}{240}\right) \times 1000 \qquad (1)$$

C. Limitations

The work is limited to a teaching prototype with qualitative bench results. Clinical accuracy, suitability for home care, and medical-device compliance are not established. Future work includes stabler streaming and reconnection after link loss, saved R-peak constants, timed measurements of SSE/REST/UART load sharing, and—after ethics approval—volunteer tests against reference instruments.

TABLE III

QUALITATIVE BENCH OUTCOMES VERSUS INTENDED ACCEPTANCE CRITERIA. "PASS" INDICATES THE FEATURE OPERATED AS EXPECTED IN LABORATORY DEMONSTRATIONS; RESULTS ARE NOT STATISTICALLY VALIDATED AGAINST A CLINICAL REFERENCE.

Test	Criterion	Status	Note
HR/SpO_2 stream	Continuous UI, no freeze	Partial	Occasional freezes
ECG waveform	Stable 5 s trace	Partial	Trace not always clean
Beep vs R-peak	Delay $\leq$100 ms	Not measured	Method defined; timing logs unavailable
Temperature + NVS	Spot measurement; retain last value	Pass	Datasheet accuracy only
NVS persistence	Survive reboot	Pass	Demographics reload
PWA on iPhone 11	Installable + SSE	Pass	Lab SoftAP demonstration
Hot-reconnect	Status $\leq$2 s	Fail	Slower than target
Color/units UI	Consistent palette	Pass	Nextion and PWA
Clinical reference	Error vs gold standard	Not done	Future work + ethics

IV. CONCLUSION

An ESP32 board was integrated with AD8232 ECG, MAX30102 PPG/SpO_2/heart rate, and MLX90614 temperature, and displayed on a Nextion panel and a PWA for laboratory teaching. Values are transmitted over Wi-Fi to the phone and the HMI. Table III summarizes the evaluation so far: HR/SpO_2 still freezes at times, the ECG trace is not always clean, and reconnection after link loss is slower than the 2 s target—so the system is not a clinically ready monitor.

Development proceeded as follows: early software/simulation checks, perforated-board wiring with a common GND, fixing MAX30102/ESP32 timer library clashes, then checking that NVS kept demographics across reboot. Beep latency was not measured against the $\leq$100 ms goal.

At 140 dollars in parts (Table I), the system is a proof-of-concept for teaching laboratories and further firmware work—not a validated medical device. Next steps are firmware hardening and, only after ethics approval, small volunteer validation studies against reference instruments.

REFERENCES

[1] World Health Organization, The Top 10 Causes of Death, Geneva, Switzerland, 2023.

[2] Espressif Systems, "ESP32 Series Datasheet," Ver. 5.2, Shanghai, China, Nov. 2025. [Online]. Available: https://www.espressif.com/

[3] M. Kokkonda and S.V. Akram, "Wireless IoT-based continuous ECG monitoring using ESP32 and ThingSpeak based performance analysis across age groups," *Sens. Actuators A: Physical*, vol. 399, 2026.

[4] V. Chandani, A. Kotnis and N. R. Kolhare, "Affordable Heart Health Monitoring System Using ESP32, AD8232 and MAX30102," 2025 IEEE Madhya Pradesh Section Conference (MPCON), Jabalpur, India, 2025, pp. 806-811, doi: 10.1109/MPCON66082.2025.11256709.

[5] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, "SLA based healthcare big data analysis and computing in cloud network," J. Parallel Distrib. Comput., vol. 119, pp. 121–135, Sep. 2018, doi: 10.1016/j.jpdc.2018.04.006.

[6] F. Y. Abdullah, M. T. Yaseen, and Y. M. Huseen, "Portable heartbeat rate monitoring system by WSN using LabView," Int. J. Comput. Digit. Syst., vol. 10, pp. 353–360, 2021.

[7] Md. Milon Islam, et al., Internet of things: device capabilities, architectures, protocols, and smart applications in healthcare domain, IEEE Internet Things J. 10(4) (2022) 3611–3641.

[8] Analog Devices, “AD8232 Single-Lead, Heart Rate Monitor Front End,” Data Sheet, Rev. D, Analog Devices, Wilmington, MA, USA, Mar. 2020. [Online]. Available: https://www.analog.com/

[9] Melexis, “MLX90614 Single and Dual Zone Infra Red Thermometer in TO-39,” Datasheet, Rev. 021, Melexis, Ieper, Belgium, Jun. 2026. [Online]. Available: https://www.melexis.com/

[10] Maxim Integrated, “MAX30102 High-Sensitivity Pulse Oximeter and Heart-Rate Sensor for Wearable Health,” Data Sheet, Rev. 1, Maxim Integrated, San Jose, CA, USA, Oct. 2018. [Online]. Available: https://www.analog.com/

[11] R. Bruzamolin, H. Gaspar, and C. F. de Oliveira, “Hospital at Home: A New Model of Home Care in Brazil,” Hosp. a Domic., vol. 7, no. 3, pp. 137–147, 2023.

Development of a Personalized Biosensor-Based Rehabilitation System for Ankle Sprain With Real-Time Corrective Feedback

Tabatha Montoya Estrada
Translational Bioengineering
University of Guadalajara
Guadalajara, Mexico
tabatha.montoya@alumnos.udg.mx

Norma Ramírez Hernández
Translational Bioengineering
University of Guadalajara
Guadalajara, Mexico
https://orcid.org/0000-0002-2022-0444

Sonia Osorio Angel
Computational Sciences
University of Guadalajara
Guadalajara, Mexico
sonia.oangel@academicos.udg.mx

Ángel Tonatiuh Hernández Casas
Computational Sciences
University of Guadalajara
Guadalajara, Mexico
tonatiuh.hernandez@academicos.udg.mx

Abstract*— Home-based rehabilitation following ankle sprain is often limited by poor adherence and the lack of objective biomechanical monitoring, which may compromise exercise execution and contribute to persistent functional deficits and chronic ankle instability. Objective: This study presents the design, implementation, and preliminary bench-level characterization of AnkleLink, an early-stage wearable prototype developed as a proof-of-concept to investigate the feasibility of providing real-time corrective biomechanical feedback during ankle rehabilitation exercises. Methods: The prototype integrates an ESP32-C3 microcontroller with an MPU6050 inertial measurement unit (IMU), a resistive flex sensor configured as an experimental electronic goniometer to provide a complementary measure of angular displacement, and an array of piezoresistive force-sensitive resistors (FSRs) embedded in a prototype insole to monitor plantar loading patterns. A multimodal sensor-fusion approach combines kinematic and kinetic measurements with configurable baseline thresholds to identify predefined deviations and trigger immediate audiovisual feedback. Results: Preliminary laboratory bench tests and exploratory trials demonstrated the technical feasibility of the multimodal sensing architecture for capturing motion-related parameters across multiple axes, obtaining complementary angular measurements, and detecting predefined deviations from reference thresholds in real time. The combined sensor configuration provided a basis for mitigating the limitations associated with relying on a single inertial sensing modality. Conclusion: The developed system demonstrates the preliminary technical feasibility of a multimodal wearable platform for real-time biomechanical monitoring and corrective feedback during ankle rehabilitation exercises. However, the current findings are limited to early-stage laboratory characterization and exploratory testing. Clinical studies involving patients, formal assessment by rehabilitation professionals, and validation against established reference systems, including optoelectronic motion-capture systems, are required before therapeutic efficacy, clinical utility, or generalizability can be established.

Keywords*— *Wearable Prototype, Sensor Fusion, Biofeedback, Ankle Rehabilitation, Proof-of-Concept, Biomechanical Monitoring.

I. INTRODUCTION

Ankle sprains represent one of the most common musculoskeletal injuries in orthopedic and physical therapy consultations. Inadequate rehabilitation protocols or premature cessation of therapy frequently contribute to chronic ankle instability (CAI), recurrent sprains, and long-term functional deficits. While active neuromuscular exercises are critical to restore joint stability, home-based regimens often exhibit high non-compliance and improper movement execution due to the absence of objective supervision [1] [2] [3]

To address unsupervised exercise monitoring, wearable systems based on inertial measurement units (IMUs) have gained significant interest for joint range of motion (ROM) tracking [4] [5]. However, single-modality IMU devices remain vulnerable to cumulative integration drift and magnetic disturbances, while lacking the kinetic capability to assess plantar load distribution or detect improper ground-contact compensation. Conversely, standalone pressure insoles or video-based systems either fail to provide continuous spatial angle tracking or require controlled laboratory environments. Integrating complementary kinematic and kinetic sensors into a synchronized, low-power platform offers a practical approach to overcome these individual sensor limitations [5].

In this context, this paper reports on the design, firmware implementation, and preliminary bench-level evaluation of *AnkleLink*, an early-stage multimodal wearable prototype. The platform combines an MPU6050 IMU, a carbon-polymer flex sensor for absolute angular referencing, and piezoresistive FSR sensors coordinated by an ESP32-C3 microcontroller. Through complementary sensor fusion and configurable threshold routines, the prototype delivers closed-loop audiovisual cues during simulated rehabilitation movements. This study focuses strictly on establishing the baseline technical feasibility, hardware integration, and preliminary sensor-fusion behavior of the prototype, providing a foundational platform for future clinical evaluations with physical therapists and patient cohorts.

A. Background and clinical motivation.

Clinical guidelines emphasize active functional rehabilitation—prioritizing progressive neuromuscular re-

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

education, ROM recovery, and targeted strengthening exercises (such as isometric eversion/inversion and heel raises)—over exclusively passive modalities [2] [3]. Early mobilization and targeted loading restore mechanoreceptor function and strengthen the stabilizing musculature [3].

B. Biomechanical rationale and rehabilitation protocols.

During unsupervised home sessions, patients frequently execute compensatory movements (e.g., hip abduction or lateral trunk tilt instead of pure ankle eversion) or apply premature, excessive plantar loading due to perceived early recovery [2] [3], [4]. Without immediate biofeedback, these errors reinforce dysfunctional motor patterns and increase the risk of recurrence [4]. The integration of synchronized low-power kinematic and kinetic sensors provides an engineering alternative to capture multi-planar motion deviations and ground reaction loading simultaneously, laying the groundwork for objective tele-monitoring platforms [5].

II. METHODOLOGY.

The development of the *AnkleLink* prototype was structured into three functional stages: hardware integration, kinematic/kinetic signal conditioning, and closed-loop biofeedback control logic.

A. Hardware Implementation and Sensor Instrumentation

The wearable architecture is centered on an ultra-compact ESP32-C3 Mini microcontroller, equipped with a 32-bit RISC-V core operating at 160 MHz and integrated 12-bit Analog-to-Digital Converter (ADC) channels [6]. The sensory instrumentation comprises:

Inertial Measurement Unit (IMU): A GY-521 breakout board hosting an MPU6050 sensor (tri-axial accelerometer and tri-axial gyroscope) positioned over the dorsal aspect of the foot. Sensor communication is established via the I^2C protocol at a 400 kHz clock frequency [5].

Resistive Flex Sensor: A 2.2-inch carbon-polymer flex sensor placed longitude finally across the anterior talocrural joint [7] [8]. It functions as a lightweight electronic goniometer to record direct mechanical bending during dorsiflexion and plantarflexion.

B. Signal Conditioning, Kinematic Modeling, and Sensor Fusion

To process the analog inputs, the flex sensor and FSRs were integrated into individual voltage divider circuits connected to the ADC1 channels of the ESP32-C3 [6] [7] [8]. For the flex sensor, the output voltage is modeled by:

$$V_{out} = V_{in} \cdot \frac{R_{flex}}{R_1 + R_{flex}}$$

where $V_{\text{in}} = 3.3$ V and $R_1 = 10$ kΩ is a precision pull-up resistor [6]. The analog voltage V_{out} is digitized and linearized using a second-order polynomial calibration model ($\theta_{\text{flex}} = a \cdot V_{\text{out}} + b$), mapping electrical resistance changes directly to sagittal joint angles (0° to 50°).

For the MPU6050, the static roll (ϕ_{acc}) and pitch (θ_{acc}) angles are estimated from the gravitational acceleration vectors (a_x, a_y, a_z) [9]:

$$\phi_{\text{acc}} = \arctan\left(\frac{a_y}{\sqrt{a_x^2 + a_z^2}}\right) \cdot \frac{180}{\pi}$$

$$\theta_{\text{acc}} = \arctan\left(\frac{-a_x}{\sqrt{a_y^2 + a_z^2}}\right) \cdot \frac{180}{\pi}$$

Relative segment kinematics between the foot and the shank are computed through vector difference:

$$\Delta\theta_{\text{joint}} = \theta_{\text{foot}} - \theta_{\text{shank}}$$

Angular velocity (ω) from the gyroscope is integrated over discrete sampling intervals ($dt = 20$ ms, corresponding to a 50 Hz acquisition loop) [9]:

$$\theta_{\text{gyro}}(t) = \theta_{\text{final}}(t-1) + \left(\frac{\omega}{131}\right) \cdot dt$$

where 131 LSB/(∘/s) denotes the sensitivity scale factor for the configured full-scale range of $\pm 250^\circ/\text{s}$.

To eliminate progressive mathematical drift caused by gyroscope integration while filtering out high-frequency mechanical vibrations from accelerometer signals, a Complementary Filter was implemented [9]:

$$\theta_{\text{final}} = 0.98 \cdot \theta_{\text{gyro}} + 0.02 \cdot \theta_{\text{acc}}$$

The flex sensor measurement (θ_{flex}) acts as an absolute, drift-free physical reference that cross-validates θ_{final}, mitigating cumulative drift during repetitive motion protocols [9].

C. Biofeedback Control Logic and Predefined Biomechanical Thresholds

The firmware executes a continuous monitoring loop running at 50 Hz that evaluates kinematic and kinetic data against predefined clinical safety thresholds established for Phase I ankle rehabilitation [2].

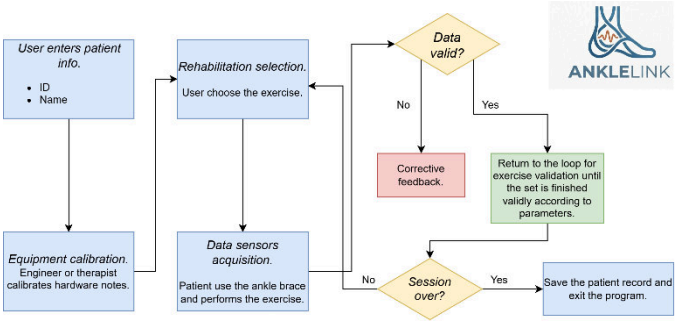


Figure 1. Flowchart represents the operation of the AnkleLink device.

The firmware executes a continuous monitoring loop running at 50 Hz that evaluates multi-axis kinematic and kinetic data against predefined clinical safety envelopes established for Phase I ankle rehabilitation [1] [2].The allowable angular boundary for pure sagittal movement is restricted between $+25^\circ$ (maximum dorsiflexion) and -45° (maximum plantarflexion) to protect healing capsuloligamentous structures while promoting progressive functional mobilization [1] [2] [10].

Concurrently, to isolate true talocrural displacement and prevent maladaptive motor habits, multi-planar compensatory thresholds are enforced: axial rotation (transverse plane) and frontal tilt (inversion/eversion) are bounded within a conservative safety cone of $\pm 15^{\circ}$ [1] [2] [4] [10]. Any rotational excursion exceeding this threshold is classified as a biomechanical compensation—such as unintended external hip rotation, pelvic hiking, or lateral foot rolling—which immediately interrupts the "MOVIMIENTO SEGURO" interface state and triggers synchronized visual and auditory biofeedback alerts to prompt real-time postural correction.

III. Results

The functional performance and baseline technical feasibility of the AnkleLink beta prototype were characterized through bench-level testing under controlled laboratory conditions. The evaluation focused on validating relative multi-axis kinematic tracking, quantifying drift suppression enabled by complementary sensor fusion, verifying discrete calcaneal contact detection, and measuring end-to-end feedback loop latency during simulated rehabilitation protocols.

A. Kinetic Analysis of Plantar Loading (FSR) and Flex Sensor Measurements.

The bench testing protocol was conducted with a convenience cohort of twenty healthy adult volunteers ($N = 20$; 10 males, 10 females; age: 20–25 years). Individuals presenting current joint pain, ligamentous laxity, history of ankle surgical intervention, or acute sprains within the preceding 12 months were excluded [1]. The experimental instrumentation captured multi-segment kinematics and discrete contact transitions through a synchronized multi-transducer arrangement. Segmental kinematics were recorded using two GY-521 breakout boards hosting MPU6050 inertial units: a distal sensor positioned over the dorsal aspect of the midfoot and a proximal counterpart affixed along the anterior distal tibia, enabling the estimation of relative joint displacement through vector subtraction ($\Delta\theta_{\text{joint}} = \theta_{\text{foot}} - \theta_{\text{shank}}$) [9]. To anchor these inertial estimates to an absolute physical baseline, a 2.2-inch carbon-polymer resistive flex sensor was mounted longitudinally across the anterior talocrural joint, directly transducing mechanical bending into sagittal angular displacement [7] [8]. Complementing the kinematic tracking chain, foot-ground interaction was monitored via a single thin-film piezoresistive membrane force sensor (ZD10-100, 500 g nominal limit) embedded within the prototype insole beneath the calcaneus to identify discrete heel-contact and lift-off transitions.

TABLE I. EXPERIMENTAL PROTOCOL AND BIOMECHANICAL THRESHOLDS

Protocol (Plane)	**Biomechanical Monitoring Parameters**		
	Primary Sensors	***Safety Envelope***	***Monitored Compensation***
OP1: Heel Raise *(Sagittal)*	Dual-IMU, Flex, ZD10-100.	$+25^{\circ}$ (DF) to -45° (PF).	Incomplete heel lift-off, premature contact,
		$V_{\text{th}} \geq$ 1.8 V (Heel-off)	sagittal overextension.
OP2: Eversion *(Frontal)*	Dual-IMU (Roll), Flex cross-check	Frontal tilt $\leq \pm 15^{\circ}$; Saggital drift $< 5^{\circ}$	External hip rotation, lateral trunk lean, parasitic flexion.
OP3: Inversion *(Frontal)*	Dual-IMU (Roll), Flex cross-check	Frontal tilt $\leq \pm 15^{\circ}$; Saggital drift $< 5^{\circ}$	Foot supination, parasitic plantarflexion, pelvic tilt

(ROM: Range of Motion; DF: Dorsiflexion; PF: Plantarflexion; V_th: Calcaneal sensor analog voltage threshold. Thresholds follow Phase I clinical functional rehabilitation practice guidelines [1] [2] [10].)

As specified in Table I, three therapeutic exercises were executed: dynamic heel raises (OP1), isometric eversion (OP2), and isometric inversion (OP3) *[2] [3]*.

Each participant completed 3 sets of 8 repetitions per exercise, yielding 100 repetitions per subject and a dataset of 2,000 movement cycles. Transducer signals were acquired at 50 Hz ($dt = 20$ ms) by the ESP32-C3 microcontroller and streamed via USB-serial communication to the custom MATLAB environment (ankleplot) for digital processing, 3-D kinematic reconstruction, and alert logging.

B. Multi-Axis Kinematic Accuracy and Drift Mitigation.

Calibration of the resistive flex sensor demonstrated monotonic impedance variation across the talocrural sagittal arc (0° to 50°). Second-order polynomial linearization yielded a coefficient of determination of $R^2 = 0.984$, with a mean mechanical angular hysteresis of $2.3^{\circ} \pm 0.6^{\circ}$ between bending and unbending phases [7] [8].

TABLE II. KINEMATIC TRACKING ACCURACY, DRIFT SUPPRESSION, AND ERROR ANALYSIS

Modality / Exercise Protocol	**Tracking Error & Stability Metrics**		
	RMSE (deg)	***MAE (deg)***	***Drift Rate (deg/min)***
Standalone Gyroscope Integration	5.84 ± 1.12	4.92 ± 0.96	1.42 ± 0.28
Fused Kinematics: OP1 (Heel Raises)	1.64 ± 0.38	1.28 ± 0.31	0.18 ± 0.07
Fused Kinematics: OP2 (Iso-Eversion)	1.41 ± 0.32	1.09 ± 0.25	0.14 ± 0.05
Fused Kinematics: OP3 (Iso-Inversion)	1.52 ± 0.35	1.18 ± 0.28	0.16 ± 0.06

(Data are expressed as Mean ± Standard Deviation across 20 participants (2,000 cycles total). RMSE: Root Mean Square Error; MAE: Mean Absolute Error.)

As detailed in Table II, uncorrected gyroscopic integration exhibited a drift rate of 1.42°/min, reaching a cumulative offset of 7.10° ± 1.45° at the end of a 5-minute continuous trial. By utilizing the flex sensor as an absolute physical anchor, the complementary filter constrained cumulative drift to 0.18° ± 0.07°/min, achieving an 87.3% reduction in integration error ($p < 0.001$). Across all participants, tracking error remained low, yielding a Root Mean Square Error (RMSE) of 1.64° ± 0.38° during dynamic sagittal motion and below 1.55° during frontal plane isometric tasks.



Figure 2. Graphical user interface (GUI) design illustrating the Contact menu and integrated support functionalities of the proposed application.

Figure 2 illustrates the graphical user interface navigation structure.

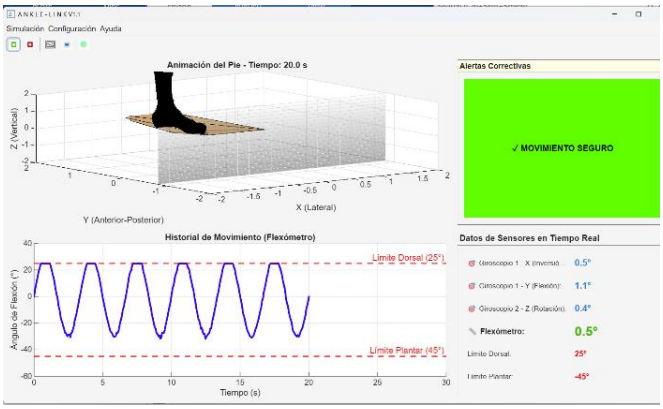


Figure 3. The user interface displays real-time kinematic validation and virtual motion reconstruction by comparing sensor-derived joint angles with actual ankle movement.

Figure 3 shows the real-time operational dashboard in MATLAB during data reception, illustrating virtual segmental animation, angular displacement curves relative to clinical boundaries (+25° dorsal, −45° plantar), and the active feedback alert indicator displaying "MOVIMIENTO SEGURO".

C. Calcaneal Contact and Heel-Off Transition Detection.

The single piezoresistive ZD10-100 sensor under the calcaneus monitored vertical foot-ground contact transitions during heel-raise protocols. The voltage divider circuit yielded a distinct electrical transition between full ground contact and calcaneal lift-off. By calibrating the detection threshold at:

$$V_{\text{th}} \geq 1.8 \text{ V}$$

(Corresponding to an applied vertical unloading threshold of ≈ 15 N), the system identified complete heel-off execution with a true positive accuracy of 98.1%. This thresholding routine discarded false repetitions when participants performed incomplete plantarflexion without lifting the heel off the support surface.

D. Corrective Feedback Latency and Compensation Detection.

The real-time monitoring framework evaluated kinematic and kinetic data against the clinical safety boundaries established in Table I. To validate classification capability, participants performed deliberate compensatory movements during designated sets, including parasitic external hip rotation, lateral trunk leaning, and frontal tilt exceeding ±15°

TABLE III. REAL-TIME BIOFEEDBACK PERFORMANCE AND LATENCY ACROSS EXERCISE PROTOCOLS

Rehabilitation Exercise	**System Evaluation Metrics**		
	Sensitivity (%)	***Specificity (%)***	***Latency (ms)***
OP1: Heel Raises (Sagittal Plane)	94.2 ± 2.1	96.5 ± 1.8	27.4 ± 3.8
OP2: Isometric Eversion (Frontal Plane)	92.1 ± 2.8	95.2 ± 2.2	28.6 ± 4.1
OP3: Isometric Inversion (Frontal Plane)	91.8 ± 3.0	95.4 ± 2.0	29.2 ± 4.5
Combined System Performance	92.7 ± 2.6	95.7 ± 2.0	28.4 ± 4.2

(Values represent classification metrics for detecting intentional compensatory movements (> ±15° roll or premature plantar contact) and end-to-end processing-to-alert latency.)

As reported in Table III, the biofeedback module identified kinematic and kinetic deviations with an overall sensitivity of 92.7% and a specificity of 95.7% . End-to-end feedback latency—defined as the time from physical transducer displacement, on-chip ADC conversion, serial transmission, to MATLAB algorithm decision and visual alert rendering—averaged 28.4 ± 4.2 ms (peak latency: 38.2 ms). This latency falls well within the physiological sensorimotor threshold (< 100 ms), enabling immediate visual and auditory notification prior to motor habituation.

E. Session Data Architecture and Representative Multi-Axis Records

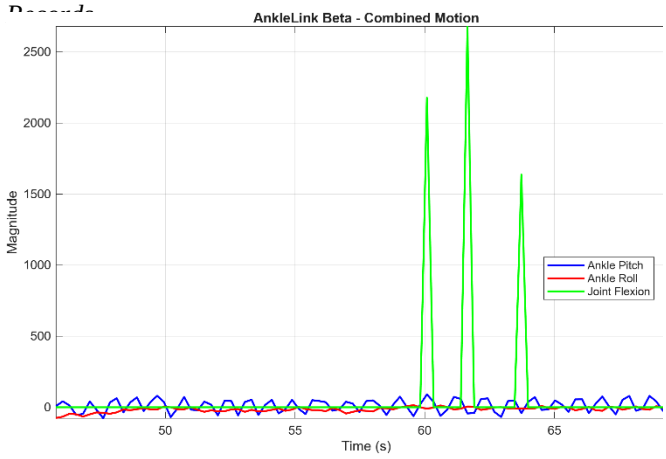


Figure 4. Representative multi-axis experimental output for an individual subject session.

Following protocol execution, the platform executes an automated archiving routine to ensure clinical traceability. Subject metadata and session completion timestamps are logged into a centralized master registry (patients_registry.xlsx). Simultaneously, raw synchronized time-series vectors (kinematics, flex resistance, and calcaneal contact voltage) are exported into individualized spreadsheet workbooks (xlsx) alongside four static diagnostic plots per exercise routine (Figure 4). As illustrated in the representative dataset, these exports provide both a composite tri-axial overview of simultaneous angular excursions and isolated single-plane trajectories for post-session biomechanical verification.

IV. Discussion

Several technical and methodological constraints must be acknowledged when interpreting the performance of the AnkleLink beta prototype. In its current implementation, the system operates strictly as a tethered bench-level testing setup that depends on real-time USB-serial data streaming to a host computer running the program on MATLAB, which limits its immediate deployment as an autonomous wearable telerehabilitation device in domestic environments [3] [4] [5]. Consequently, it is not currently configured as an untethered, standalone wearable medical device suitable for unconstrained home deployment. Furthermore, although internal cross-validation between the dual-IMU architecture and the resistive flex sensor demonstrated high consistency and repeatable drift suppression, the angular estimations and discrete contact events have not yet been benchmarked simultaneously against gold-standard optoelectronic motion capture systems (such as Vicon or Qualisys) or calibrated multi-axis force platforms. Direct comparison against such reference instrumentation represents an indispensable prerequisite before establishing clinical measurement equivalence or asserting therapeutic efficacy [4] [5].

The physical properties of the chosen transducers also introduce specific measurement boundaries. While the second-order polynomial model achieved satisfactory linearity ($R^2 = 0.984$), carbon-polymer flex sensors exhibit inherent mechanical hysteresis ($2.3^\circ \pm 0.6^\circ$) and viscoelastic relaxation under sustained cyclic loading, which can degrade instantaneous angle estimation during rapid ballistic joint reversals [7] [8]. Moreover, securing the sensors to the foot dorsum and distal tibia via elastic straps leaves the setup susceptible to soft-tissue artifacts and subtle sensor migration during repetitive movement sets [5] [9]. Regarding kinetic monitoring, the single piezoresistive ZD10-100 sensor mounted beneath the calcaneus serves strictly as a localized binary contact threshold detector to verify heel-off execution, rather than providing continuous multi-point baropodometric mapping of dynamic plantar pressure gradients across the entire foot [5].

Finally, the preliminary experimental validation was deliberately restricted to a convenience cohort of 20 healthy young adults executing simulated movements within a controlled laboratory environment. While this approach provided a baseline for characterizing sensor fusion mechanics and biofeedback latency, healthy volunteers do not exhibit the altered movement trajectories, antalgic guarding, joint effusion, or restricted ranges of motion characteristic of patients recovering from acute Grade I or II ankle sprains [1] [2] [10]. Future technical iterations will aim to consolidate the discrete breadboard instrumentation into a custom flexible printed circuit (FPC), transfer the processing and thresholding algorithms entirely onto the embedded ESP32-C3 firmware with wireless BLE data transmission [6], and conduct formal usability and validation trials in direct collaboration with physical therapy specialists before evaluating the platform in home-based rehabilitation contexts [3] [5].

V. References

[1] E. a. Ruiz-Sánchez FJ, "Management and treatment of ankle sprain according to clinical practice guidelines: A PRISMA systematic review," *Medicine (Baltimore)*, p. 101(42):e31087, 2022.

[2] A. M. Ariza V, C. C. Salazar B and E. H. Peinado R, "La rehabilitación funcional en el manejo del esguince de tobillo: una revisión sistemática," *Archivos de la Sociedad Chilena de Medicina del Deporte*, vol. 66, no. 1, pp. 3-20, 2021.

[3] D. Chan and F. Can, "Patients' adherence/compliance to physical therapy home exercises," *Fizyoterapi Rehabilitasyon*, vol. 21, no. 3, pp. 132-139, 2010.

[4] M. Attia and M. F. Taher, "A wearable device for monitoring and prevention of repetitive ankle sprain," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, 2015.

[5] I. Ekerete, M. Garcia Constantino, Y. Diaz Skeete, C. Nugent and J. McLaughlin, "Fusion of Unobtrusive Sensing Solutions for Sprained Ankle Rehabilitation Exercises Monitoring in Home Environments," *Sensors (Basel)*, vol. 21, no. 22, p. 7560, 2021.

[6] Espressif, "Espressif Esp32-C3," Febrero 2021. [Online]. Available: https://documentation.espressif.com/esp32-c3_datasheet_en.html.

[7] J. F. García Mejía, J. A. García Mejía, G. Gonzáles Hernández and L. A. Tinoco Monroy, "Obtención de Repetibilidad, Histéresis y Linealidad de un sensor de flexión resistivo por medio de un instrumento virtual," in *Memorias del IV Congreso Nacional de Mecatrónica*, Morelia, 2005.

[8] J. Nagar, A. Panat and R. Gunjan, "Knee Flexion/Extension Measurement System using Wearable Sensor," *International Journal for Research in Engineering Application & Management (IJREAM)*, vol. 5, no. 9, pp. 148-153, 2019.

[9] InvenSense, Inc., "MPU-6000 and MPU-6050 Product Specification," InvenSense, Inc., Sunnyvale, CA, 2013.

[10] E. Delahunt et al, "Clinical assessment of acute lateral ankle sprain injuries (ROAST): 2019 consensus statement and recommendations of the International Ankle Consortium," *Br. J. Sports Med.*, vol. 52, no. 20, pp. 1304-1310, 2018.

Development of a Six-Lead Portable ECG System for Home Health Care and Emergency Alerts

Gustavo Ferreira das Chagas
Biomedical Engineering Program
COPPE, Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
gustavoferreiradaschagas@gmail.com

Marcelo de Carvalho Bossan
Bossan Computação Científica (BCC)
Rio de Janeiro, Brazil
mbossan@bcc-sistemas.com.br

Marcos Vinícius Theodoro P. da Silva
Control and Automation Engineering
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
marcos_theodoro.20221@poli.ufrj.br

Karine Rodrigues Lisboa
Computer and Electronic Engineering
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
karinelisboa.20232@poli.ufrj.br

Jurandir Nadal
Biomedical Engineering Program
COPPE, Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jn@peb.ufrj.br

***Abstract*— The world population is becoming older, with increased prevalence of cardiovascular diseases. The present work aims to develop a safe and reliable telecardiology system consisting of four modules: an ECG acquisition module for remote self-examination; an edge computer for local processing, automatic detection of life-threatening conditions, and alert issuance; a remote server for secure, reliable data reception and data storage; and a smartphone application for doctor use. The ongoing development of the system shows great promise in terms of utility, reliability, and future application. The system was designed to be easy to use, allow patients to self-record ECGs, and incorporate automatic detection of critical conditions to avoid alarm fatigue. Unlike single-channel acquisition systems, the current six-lead frontal-plane system expands diagnostic possibilities. The smartphone application enables near-real-time receipt of notifications, visualization, and analysis of patient data, allowing continuous follow-up and accelerating emergency care when necessary.**

Keywords—Telecardiology, Home Healthcare, Portable ECG, Edge Computer, Internet of Medical Things (IoMT)

I. Introduction

According to the World Health Organization, one-sixth of the world population will be aged 60 years or older in 2030. Along with this, the prevalence of age-related disorders also increases, mainly cardiovascular diseases (CVDs) [1]. CVDs are the leading cause of death and are one of the main causes of the increasing costs of health care. The Brazilian Unified Health System (SUS) alone is expected to spend approximately BRL 10 billion in the coming years [2], while facing 1,000 deaths daily [3]. Globally, spending on CVD could reach 0.7% of global gross product (US$ 1.4 trillion) [4]. Despite current healthcare efforts, it is estimated that there will be an increase of 73.4% in global cardiovascular deaths by 2050 [5], keeping CVD at the top of the list of mortality and disability illnesses.

This scenario points to the need for effective patient monitoring and risk assessment, as well as measures to prevent strain on healthcare facilities. The Internet of Medical Things (IoMT) addresses this gap by monitoring patients in real time outside the hospital environment [6], thereby allowing doctors to make quick interventionist or therapeutic decisions. This approach benefits both urban and rural areas. This will provide equitable access to healthcare, enhanced agility, reduced costs, lower contagion risks, and increased comfort for both doctors and patients [7]. In addition, cardiac IoMT has proven highly effective at detecting various cardiovascular events and diseases across diverse populations, from children to the elderly [7, 8]. However, a major disadvantage of these numerous technologies is the difficulty of exporting captured information to physicians [9], who are ultimately responsible for diagnosis and clinical decision-making, thereby hindering rapid intervention, which in critical situations (such as cardiovascular events) can be fatal.

Just as the invention of the electrocardiograph was revolutionary for the creation of ICUs (Intensive Care Units) [10], implementing robust cardiac telemedicine and home health care systems is expected to reinvigorate modern medicine. With this in mind, the present work aims to develop a safe and reliable telecardiology system with four components: a compact ECG acquisition module, to be operated by the patient; an Edge computer for pre-diagnosis of life-threatening conditions and transmission of cardiac signals; a central server for secure and reliable reception and storage of data; and a medical application for smartphones, which will receive the alerts and allow the doctor to view and analyze the patient's signal. The system is expected to work flawlessly, ensuring that the patient collects his own signal during moments of discomfort or as advised by medical professionals, issues alerts at pre-established severity levels, and guarantees that the alert and the ECG reach the physician in time for quick intervention when needed.

II. Materials & Methods

Structural and behavioral models of the system were developed using the System Modeling Language (SysML). Fig. 1 presents the main Block Definition Diagram (BDD), in which the **HomeECG** block is decomposed into three main blocks: **ClientSystem**, comprising the **PortableECG** and **EdgeComputer** modules; **RemoteServer**, which receives ECG recordings via Wi-Fi, stores them, and distributes them using Apache Kafka streaming technology; and **PhysicianApp**, which delivers fast and reliable smartphone notifications to the physician.

This study received financial support from the Brazilian agencies CNPq (Grant # 309247/2025-0), CAPES (Grant # 001 and DSc scholarship), FAPERJ (Grant # E26/201.247/2021), and the Brazilian Chamber of Deputies (Parliament Amendment #44890020)

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

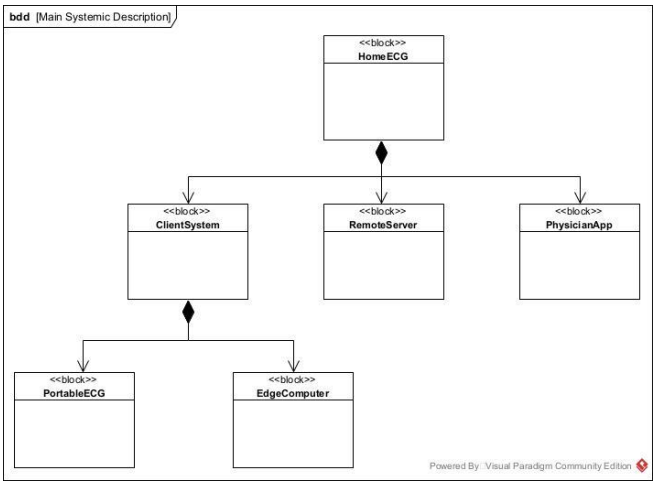
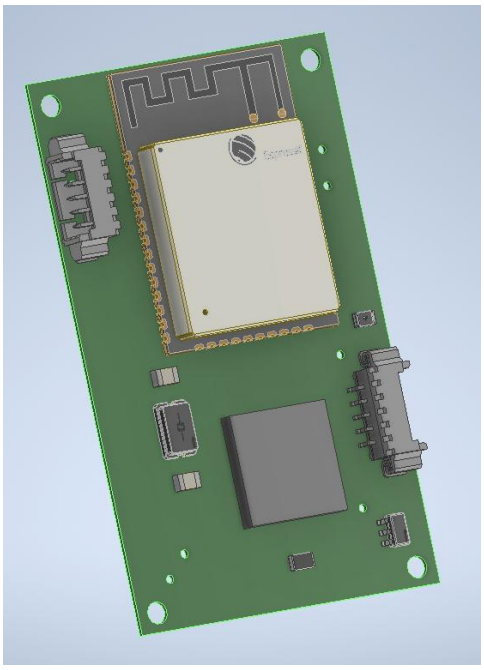


Fig. 1 System block diagram in SysML

A. Portable ECG

The portable ECG module was projected to sample the three bipolar frontal plane ECG Leads I, II and III. The current version of the portable ECG is a stand-alone module with a display, conceived for clinical use by health professionals, testing, and validation of the chosen transducer. It allows a quick view of the patient´s ECG, which can be transmitted to the physician's smartphone or desktop for the issuance of a medical report and recording.

The acquisition modules use two primary electronic components: an ADAS1000-4 3-channel electrocardiographic transducer (Analog Devices) [11] and a low-cost ESP32 microcontroller (Espressif Systems) [12], ideal for IoMT applications. The entire electronic circuit was designed using Altium Designer [13], resulting in a printed board measuring 31 × 50 × 10 mm (Fig. 2).


Fig. 2 Printed circuit board used in portable ECG

The ADAS1000-4 chip operates over the SPI (Serial Peripheral Interface) communication protocol with ESP32. It is configured to acquire signals at a sampling rate of 500 Hz, with a gain of 2,8 V/V (±1% error), providing a resolution of 19 useful bits. Under the manufacturer's test conditions, the integrated circuit exhibits the following typical specifications: 110 dB Common Mode Rejection Ratio with a minimum of 105dB; 1 GΩ in parallel with 10 pF represents the input impedance of the integrated circuit, specified at 10 Hz; and an input-referred noise of 10 µV peak-to-peak for the configured parameters. Its nominal bandwidth is 0.5–150 Hz; in addition, a 3rd-order Butterworth high-pass filter algorithm with a cutoff frequency of 0.5 Hz was also implemented to reduce movement artifacts and baseline fluctuations.

Disposable Ag/AgCl electrodes collect ECG signals via a cable with four vias for the standard Einthoven triangle: Right Arm (RA), Left Arm (LA), Left Leg (LL), and a reference electrode (RLD – Right-Leg Drive). This controls the patient's common-mode voltage and reduces common-mode interference. The ADAS1000- 4 lead-off circuit injects 50 nA DC into each acquisition electrode. When the input voltage falls below or exceeds the typical threshold of 0.2 – 2.4 V, a word indicating the connection or disconnection of a specific electrode is transmitted over the bus [11].

As soon as the ESP32 receives the information captured by the ADAS1000-4 via SPI, it elaborates a data packet as shown in Fig. 3, including a packet counter, the portable ECG battery level, and a CRC (Cyclic Redundancy Check) value. Once structured, this 302-byte packet is transmitted via Bluetooth to the receiver of interest, with latency under 100 ms.

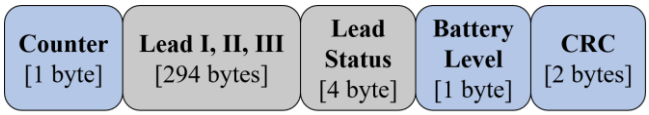


Fig. 3 Data packet transmitted by a portable ECG

The augmented leads (aVR, aVL, aVF) can be easily calculated by the receiver using the following equations:

$$aVR = -\frac{Lead\ I + Lead\ II}{2} \quad (1)$$

$$aVL = \frac{Lead\ I - Lead\ III}{2} \quad (2)$$

$$aVF = \frac{Lead\ II + Lead\ II}{2} \quad (3)$$

Using the Autodesk Inventor 3D modeling software [14], a box was designed to accommodate the printed circuit board and the other electronic components: 3.7 V 1000 mAh battery, USB-C connector, LED, on/off button, 2" ST7789 display, and the connector for the ECG cable. The printed circuit board (Fig. 2) was manufactured in China, and the box was fabricated via additive manufacturing (3D printing) in PET-G, yielding the pilot version (ECGv1, Fig. 4 on the right) of the cardiac signal acquisition module. To evaluate the performance of ECGv1 and ADAS1000-4, we used a Handy Sim HS-30 vital signals generator (Fig. 4 on the left), produced specifically for biomedical equipment assays [15].

Human ECG signal acquisition followed the protocol approved by a Brazilian research ethics committee (CAAE: 97984626.9.0000.5259). For both the simulated ECG signal (Normal Adult, 60 bpm, 1.00 mVpp) and the human ECG signal (from a healthy 28-year-old male), a Bluetooth connection was established with the PEB Workstation ECG analysis software [6]. Five minutes of transmitted data were acquired and stored for each condition to enable post-processing, heart rate variability (HRV) analysis, and comparison between human and simulated ECG signals.



Fig. 4 Integration of the signal generator with the ECGv1

B. Edge Computer

A second version (ECGv2) is being developed for domestic use by patients and operates in integration with an edge computer. In this architecture, the collected ECG is continuously transmitted via the Bluetooth Low Energy (BLE) protocol to the edge computer, which stores the data in an internal buffer. Although the acquisition module uses the same electronics as the ECGv1 (Fig. 2), it differs in its embedded software. It can also be built without a display, since the edge module displays the signal, acquisition warnings, and battery levels (Fig. 5, left). When the edge's push-button is pressed, the system transmits the last 15 seconds of the six-lead ECG signal in JSON (JavaScript Object Notation) format via Wi-Fi to the remote server module using the MQTT (Message Queuing Telemetry Transport) communication protocol. The internal components of the ECGv2 module and edge computer are shown in Fig. 5 (right).

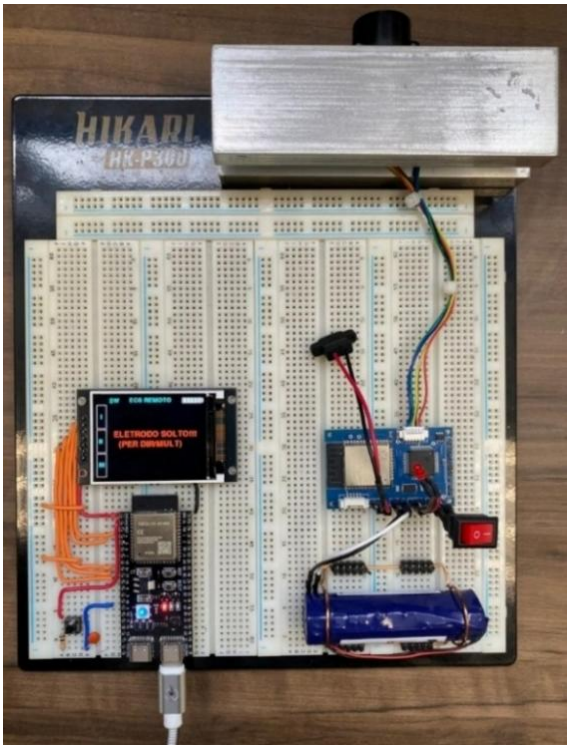


Fig. 5 Edge computer (left) and ECGv2 (right) prototypes under development.

The Edge Computer complements the home ECG. It serves as a visual feedback tool, ensuring the patient has achieved a high-quality ECG signal. Equipped with a 2" ST7789 display, it allows the user to view the collected ECG in real time, displaying warning messages indicating loose electrodes (LA, RA, LL, RLD, or multiple electrodes), acquisition module battery level (ECGv2), connection symbols for Bluetooth (B), Wi-Fi network (W), and MQTT server (T). Finally, a push-button either validates the captured signals and sends them to the server or requests the patient to collect a new, more careful ECG signal.

At the Edge Computer, the ECG will be pre-processed to detect predefined critical conditions, such as atrial fibrillation or ischemic ST changes, including ST-elevation myocardial infarction (STEMI). This automatic detection process screens ECG signals to inform the physician's Mobile Application of an emergency condition [9], which must be urgently confirmed to ensure the patient receives immediate medical care.

C. Remote Server

The server consists of a standard desktop computer running Linux, configured to operate as an MQTT broker. This protocol enables direct communication with many home ECG systems. These systems can operate either as publishers, transmitting ECG data to the server, or as subscribers, receiving guidance messages from the physician. All messages are received in JSON format and stored in an Apache Kafka stream processing broker. This open-source platform, developed by LinkedIn, is currently used worldwide by thousands of large companies [16]. This technology is based on Commit Logs, immutable files that record events in chronological order (date and time). In addition, the Kafka Cluster works by splitting the messages received from its "producers" into partitions and redundantly redistributing them across several different brokers, which can be real or virtual machines, ensuring that its "consumers" have access to all information, even if one of these machines is inoperable [17]. Data anonymization, the development of a proprietary server and the adoption of the SSH (Secure Shell) protocol ensure data encryption and secure access to the acquired clinical and personal databases.

D. Mobile Application

This application will notify physicians on their smartphones in near real time of any cardiac event their patients report. As the ECG signal will be preprocessed, the application will receive both the ECG data and an estimated risk alert indicator, enabling the physician to review and analyze the ECG recording as quickly as possible, only when necessary. When the ECG is considered normal, he can issue a medical report at his convenience. Thus, the automatic pre-diagnosis serves as an information filter for the doctor by pre-processing the signals [9] and avoiding alarm fatigue [18, 19]. This procedure helps speed up decision-making, whether calling emergency contacts or adjusting medication. Consequently, this may increase the chances of a positive prognosis for any event that may occur with patients. The smartphone application will also provide statistical analysis of patient data, ECG history, and emergency contacts. Fig. 6 shows screenshots of the application's main interfaces, written in Portuguese, the only language available now.

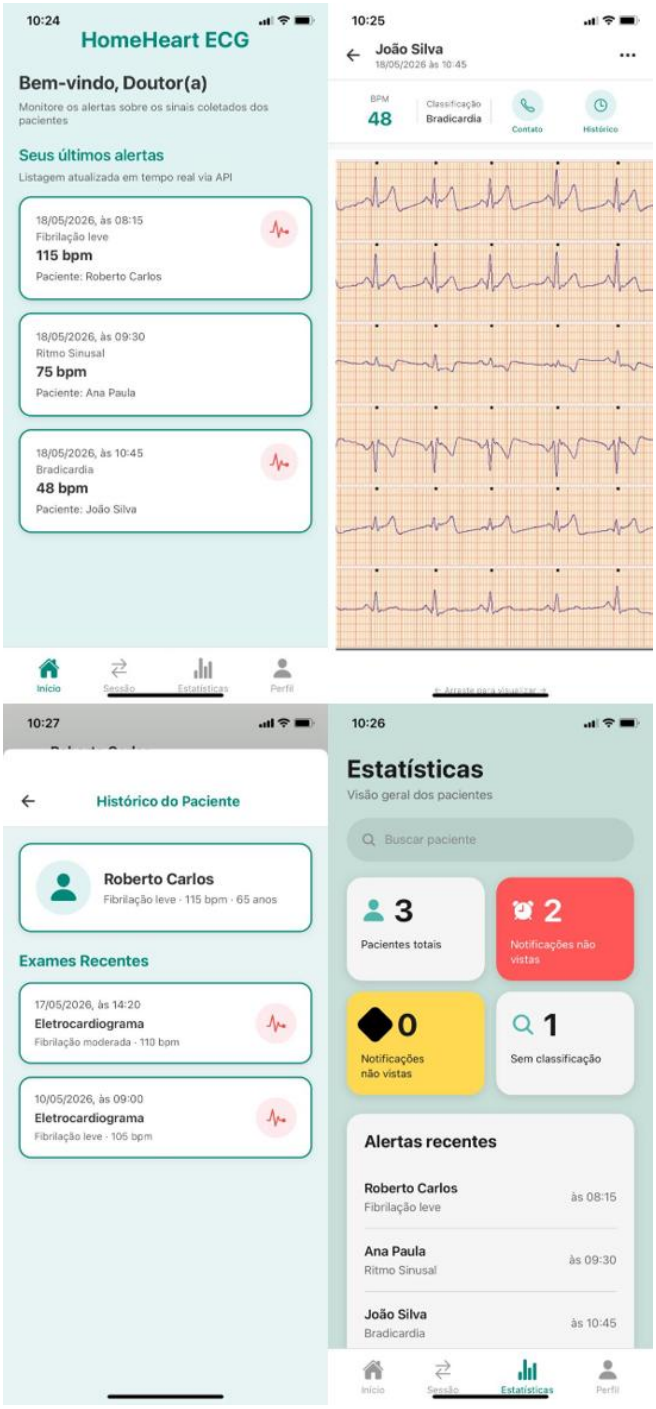


Fig. 6 Screenshots of the app's features

III. Results

The Fig. 7 shows the real-time ECG signals captured by leads I, II, and III by ECGv1, together with the reconstructed augmented leads (aVR, aVL, and aVF), and HRV analysis performed by the Workstation software. At some collection points, 60 Hz noise was observed, consistent with electrical-network coupling, and was easily filtered by applying a notch filter during post-processing or real-time display on the workstation. R-peak detection and HRV values were concise, with markings aligned with the R waves; the simulated signal showed variability tending to zero and an HR of 59.98 bpm; the human signal showed variability and an HR of 60.47 bpm. Although it did not use a data reconstruction algorithm, the CRC protocol proved efficient because all transmissions had zero lost packets. With a consumption of 200 mA, ECGv1 is expected to have an autonomy of 4 to 5 hours, with a conventional Bluetooth communication range of 10 m indoors. The filtering methods adopted are believed to attenuate baseline drift, although this has not been quantified.

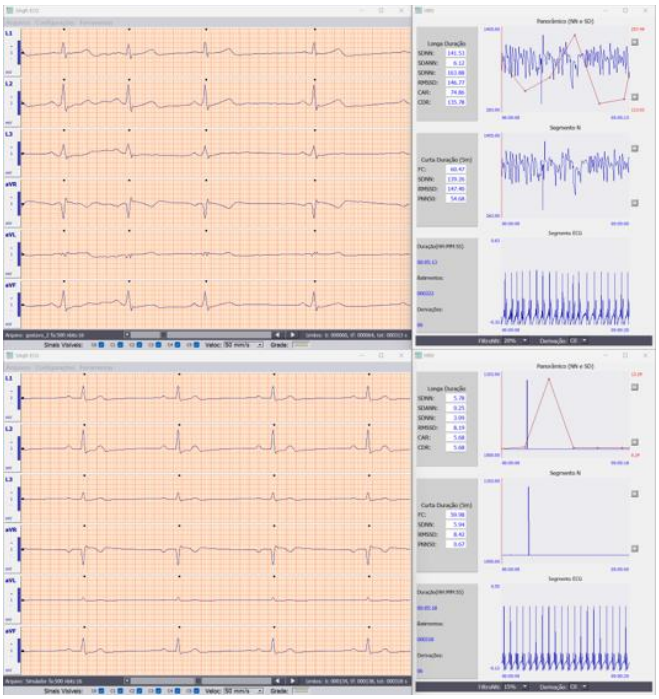


Fig. 7 Workstation interface displaying the six-lead ECG captured by the ECGv1, where the points indicate the detected R_waves, and HRV analysis of a recorded signal. The top image shows the human signal, and the bottom image presents the simulated signal.

IV. Discussion

The developing of this framework shows great promise in terms of utility and reliability. The system provides clean signal acquisitions with little or no interference affecting the signal's morphology. It stands out for its ease of use and possible integration with specialized analysis software and hardware, such as the Workstation [6] and Edge Computer. The current six-lead frontal-plane system expands the possibilities for more comprehensive analyses, including ischemic events such as atrial fibrillation, STEMI, cardioembolic stroke, and branch blocks. The clinical version (ECGv1) allows immediate interpretation by health professionals, while ECGv2, connected to the Edge Computer, allows extending patient treatment to home health care and sending a direct alert to the doctor. By automatically classifying the ECG, it also avoids alarm fatigue [18, 19] and signal-processing overload in other system nodes, since it already exports the analysis results [20]. The Apache Kafka stream service ensures data storage, transmission, and security on its own servers [17], in accordance with the Brazilian LGPD (General Law for the Protection of Personal Data) [21]. The mobile App enables real-time receipt of notifications, as well as visualization and analysis of the patient's signal. In conclusion, the presented system enables patients to perform routine self-monitoring or emergency self-monitoring and accelerates decision-making by doctors and other healthcare professionals.

The limitations of the current study stem from the ongoing development of specific modules, namely the physician application, the automated ECG interpretation algorithms, and

the domestic version of the portable ECG. Therefore, subsequent stages of this research will include bench tests at different rhythms and wave amplitudes, emergency simulations, and home monitoring trials to ensure system robustness and clinical safety. While a specific safety and isolation diagram is omitted, battery safety is ensured by Protection Circuit Module (PCM) that prevents overcharging, overdischarging, overcurrent, and short circuits. Furthermore, the device is protected against operation during USB charging.

In addition, the presented technology has strong potential for incorporation into the Brazilian Unified Health System (SUS), as it already meets most of the criteria established by the National Commission for the Incorporation of Technologies into the SUS (Conitec). According to RDC No. 751/2022, the device is classified as a Class II (medium-risk) medical device. Although the prototype has not yet been registered with the Brazilian Health Regulatory Agency (ANVISA), the Brazilian National Telecommunications Agency (ANATEL) has already certified the ESP32 communication module. Furthermore, its low production cost, ease of use, reliability, and portability are consistent with the ethical and equity principles considered in the Health Technology Assessment (HTA) process adopted by SUS.

ACKNOWLEDGMENT

The authors wish to thank fellow researcher Edgar Norio Taka for providing the calibrated ECG signal generator.

REFERENCES

[1] World Health Organization, "Ageing and Health", *World Health Organization*, 1 October 2024. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (accessed Jun. 28, 2026).

[2] A. Bigoni, I. Oliveira, L. Tôrres, A. C. Padula, and L. Fahham, "Economic Impact of Cardiovascular Diseases in Brazil: Projections of Direct and Indirect Costs for the Next 5 Years," *Epidemiology and Health Services*, vol. 34, no. 1, pp. 221–221, 2025, doi: https://doi.org/10.5327/2237-9622.2025.v34s1.337.

[3] "Cardiometer," *www.cardiometro.com.br*, 2020. http://www.cardiometro.com.br/ (accessed Jun. 28, 2026).

[4] Lishani Ramanayake, "Atherosclerotic cardiovascular disease (ASCVD) costs the world $680 billion a year in healthcare – roughly the size of Sweden's entire economy, new research finds - OHE," *OHE - At the forefront of health economics*, Mar. 31, 2026. https://www.ohe.org/news/atherosclerotic-cardiovascular-disease-ascvd-costs-the-world-680-billion-a-year-in-healthcare-roughly-the-size-of-swedens-entire-economy-new-research-finds/

[5] B. Chong *et al.*, "Global burden of cardiovascular diseases: projections from 2025 to 2050," *European Journal of Preventive Cardiology*, vol. 32, no. 11, Sep. 2024, doi: https://doi.org/10.1093/eurjpc/zwae281.

[6] G. F. Chagas, M. C. Bossan, and J. Nadal, "A Portable Low-Cost Wireless Electrocardiographic System for Home Health Care," *IFMBE proceedings*, vol. 106, pp. 223–232, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-61960-1_22.

[7] Khondker, Mohammad, Zobair , *et al.*, "Systematic review of Internet of medical things for cardiovascular disease prevention among Australian first nations," *Heliyon*, vol. 9, no. 11, pp. e22420–e22420, Nov. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e22420.

[8] J. Medina-Avelino, R. Silva-Bustillos, and J. A. Holgado-Terriza, "Are Wearable ECG Devices Ready for Hospital at Home Application?," *Sensors*, vol. 25, no. 10, p. 2982, May 2025, doi: https://doi.org/10.3390/s25102982.

[9] G. F. Chagas, M. C. Bossan, and J. Nadal, "HRV ANALYST, an Edged Computer to ECG Signal Processing and Emergency Alert for Home Health Care Systems," *IFMBE Proceedings*, vol. 127, pp. 212–220, 2025, doi: https://doi.org/10.1007/978-3-031-94934-0_22.

[10] E. A. Bohula and D. A. Morrow, "The Cardiac Intensive Care Unit at Age 60," *Circulation*, vol. 150, no. 19, pp. 1481–1483, Nov. 2024, doi: https://doi.org/10.1161/circulationaha.124.070542.

[11] "ADAS1000-4," *Analog.com*, 2024. https://www.analog.com/en/products/adas1000-4.html (accessed Jun. 28, 2026).

[12] Espressif, "ESP32 Overview | Espressif Systems," *www.espressif.com*, 2023. https://www.espressif.com/en/products/socs/esp32 (accessed Jun. 28, 2026).

[13] "Altium Designer – PCB Design Software," *Altium Designer – PCB Design Software*, 2019. https://www.altium.com/in/altium-designer (accessed Jun. 28, 2026).

[14] "Autodesk Inventor Software | View prices and buy the official Inventor 2026," *Autodesk.com*, Jan. 15, 2021. https://www.autodesk.com/br/products/inventor (accessed Jun. 28, 2026).

[15] "HS30 ES ECG AND OXIMETRY SIMULATOR - Protec - Hospital Medical Equipment Manufacturer," *Protec - Medical Equipment*, Feb. 19, 2026. https://www.protec.com.br/produto/simulador-ecg-oximetria-hs30-es/ (accessed Jun. 28, 2026).

[16] "Powered By," *Apache.org*, 2026. https://kafka.apache.org/powered-by/

[17] T. P. Raptis and A. Passarella, "A survey on networked data streaming with Apache Kafka," *IEEE Access*, vol. 11, pp. 85333–85350, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3303810.

[18] E. A. M. Michels, S. Gilbert, I. Koval, and M. K. Wekenborg, "Alarm fatigue in healthcare: a scoping review of definitions, influencing factors, and mitigation strategies," *BMC Nursing*, vol. 24, no. 1, Jun. 2025, doi: https://doi.org/10.1186/s12912-025-03369-2.

[19] M. Woo and O. Bacon, *Alarm Fatigue*. Agency for Healthcare Research and Quality (US), 2020. Available: https://www.ncbi.nlm.nih.gov/books/NBK555522/

[20] P. Bhattacharya, A. Mukherjee, B. Bhushan, S. K. Gupta, Thippa Reddy Gadekallu, and Z. Zhu, "A secured remote patient monitoring framework for IoMT ecosystems," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: https://doi.org/10.1038/s41598-025-04774-y.

[21] BRAZIL, "General Law for the Protection of Personal Data (LGPD)," Brazilian *Ministry of Sports*. https://www.gov.br/esporte/pt-br/acesso-a-informacao/lgpd

Development of a Wearable Device Prototype for Parkinsonian Resting Tremor Monitoring

Dério Crisóstomo Oliveira
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
derio-oliveira@hotmail.com

Fábio de Sousa Cardoso
PPGEE/Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
fcardoso@uea.edu.br

Isabela Rogério Cardoso
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
irc.eng25@uea.edu.br

André Luiz Printes
PPGEE/Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
aprintes@uea.edu.br

Afonso Aguiar de Carvalho
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
aguiar94_carvalhoLSE@outlook.com

Bruno da Gama Monteiro
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
bmonteiro@uea.edu.br

Jozias Parente de Oliveira
PPGEE/Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
jpoliveira@uea.edu.br

Danilo da Silva Cunha
Electronic Engineering
State University of Amazonas
Manaus, AM, Brazil
ddscn.eem23@uea.edu.br

Isaque Vilson Batista da Costa
PPGEEL
State University of Amazonas
Manaus, AM, Brazil
ivbdc.mee24@uea.edu.br

***Abstract*—This article presents the development and technical evaluation of a low-cost wearable prototype for acquiring, processing, transmitting, and visualizing wrist-motion signals relevant to Parkinsonian resting-tremor research. The system is based on the T-Watch 2020 V2 platform, integrating an ESP32 microcontroller, BMA423 triaxial accelerometer, AXP202 power-management unit, TFT display, Wi-Fi, and Bluetooth Low Energy (BLE). Firmware implemented in C++ performs calibration, first-order IIR filtering, resultant-magnitude calculation, a 128-point Fast Fourier Transform (FFT), relative spectral-energy analysis in the 4–6 Hz band, and HTTP/BLE transmission to a containerized web dashboard. Evaluation used controlled functional wrist movements and an internal synthetic 5 Hz test signal; no data were acquired from individuals with Parkinson's disease. At 100 Hz sampling, the 128-sample analysis window provides approximately 0.78 Hz frequency resolution. For the synthetic reference, the observed frequency-bin error was 0.31–0.47 Hz (6.2%–9.4%). Average communication latency was 42 ± 12 ms for HTTP/Wi-Fi and 18 ± 5 ms for BLE; in 30-min tests with 1,406 transmissions, packet loss was 0.14% for Wi-Fi and 0% for BLE. These results demonstrate functional integration of sensing, embedded processing, wireless communication, storage, and visualization, while the prototype remains a technical proof of concept requiring formal clinical validation against reference instruments and standardized clinical scales.**

***Keywords*—Parkinson's disease; resting tremor; wearable device; ESP32; digital signal processing; FFT; digital health.**

I. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder associated with impairment of motor and non-motor functions. Recent epidemiological studies indicate a significant increase in the global prevalence of the disease, a fact related to population aging and increased life expectancy [1]-[3]. Among the classical motor symptoms, resting tremor stands out as one of the most characteristic signs, usually manifested as involuntary rhythmic oscillations of the extremities in the 4 Hz to 6 Hz range [2]-[3].

Clinical assessment of these tremors is still frequently based on observations performed by health professionals and on subjective scales, such as the Unified Parkinson's Disease Rating Scale. Although widely used, such approaches may vary among evaluators and do not continuously capture symptom evolution outside the clinical environment [1], [3]-[4]. Objective quantification of tremor frequency, amplitude, and temporal pattern can complement medical assessment and support longitudinal patient monitoring [1]-[2], [4]-[5].

In this context, wearable devices based on inertial sensors constitute a promising alternative for continuous, non-invasive, low-cost monitoring. Recent reviews and representative studies indicate that accelerometers and inertial measurement units have been applied in digital health, rehabilitation, and motor monitoring, enabling signal acquisition in natural environments and the generation of quantitative indicators [1], [3], [5]-[9], [11], [24]. This approach is especially relevant for PD, since tremor may vary throughout the day as a function of medication, fatigue, rest, anxiety, and other conditions [7].

This work presents the development and preliminary functional evaluation of a wearable prototype for resting tremor monitoring in Parkinson's disease, integrating inertial

signal acquisition, processing, wireless communication, and real-time visualization. Evaluation used functional tests and artificially generated signals, without data from individuals with Parkinson's disease. The main contribution is an integrated wearable platform that provides a technical basis for future clinical validation and home-based monitoring.

II. BACKGROUND AND RELATED WORK

A. Resting tremor and quantitative monitoring

Parkinsonian resting tremor is characterized by involuntary, regular, and recurrent oscillations evident at rest [1], [2], arising from cerebral oscillatory networks involving the basal ganglia, thalamus, and motor cortex [2]. It can be analyzed through acceleration, angular velocity, or displacement signals, with frequency being a key parameter for differentiating parkinsonian tremor from other involuntary movements. Inertial sensors can objectively quantify tremor through amplitude, dominant frequency, spectral energy, and temporal variation, with these parameters related to clinical assessment [3], [4], [16], [24]-[26]. Patel et al. demonstrated the potential of wearable sensors for monitoring motor fluctuations outside clinical settings [10]. Previous studies have reported wearable and low-cost embedded approaches for tremor detection, quantification, and visualization [4], [6], [12], [25], [26]. This project extends these approaches by integrating acquisition, processing, HTTP/BLE transmission, sensor-acquisition and synthetic-test modes, CSV storage, and a containerized dashboard. Specifications were compared with representative wearable and inertial monitoring systems [5], [9], [10], [25], [26], as summarized in TABLE I.

TABLE I. PROSPECTIVE COMPARATIVE ANALYSIS WITH REFERENCE WEARABLE SYSTEMS

Parameter	Proposed Prototype	Reference Systems
IMU Sensor	Bosch BMA423 (Accel.)	Representative inertial wearables [5], [9], [10], [25], [26]
Sampling Rate	100 Hz	Typically 50–100 Hz in representative implementations [25], [26]
Connectivity	Wi-Fi and BLE	Bluetooth / Wi-Fi / Cellular
Local DSP	Yes (FFT & Filtering on ESP32)	Yes (Embedded frameworks)
Dashboard	Yes (Flask / Docker Web App)	Yes (Proprietary health app)
CSV Export	Yes (Direct API download)	Restricted / Research APIs
Validation	Technical bench/functional validation; no clinical participants	Clinical/experimental validation reported in representative systems [5], [25], [26]
HW Cost	Low Cost (~USD 60)	High Cost (Commercial devices)

B. Wearable devices and inertial sensors

Wearable devices applied to health combine sensors, microcontrollers, batteries, wireless communication, and visualization software in compact systems capable of monitoring physiological or motor signals. Advances in MEMS sensing and wearable monitoring have made it possible to measure subtle accelerations and movements with low power consumption and reduced cost [5], [9], [11], [23], [24]. In the proposed prototype, the Bosch BMA423 acquires accelerations along the X, Y, and Z axes and supports programmable acceleration ranges and digital interfaces suitable for wearable applications [27]. Integrated into the T-Watch 2020 V2 with an ESP32, display, battery, and power-management circuitry, it provides a compact wearable IoT platform [14]. The ESP32 was selected for its integrated processing, Wi-Fi, Bluetooth Low Energy, I2C interfaces, and low-power modes [13], [22], [28]. The AXP202 power-management IC provides battery charging, power-rail control, monitoring, and protection functions for portable systems [14], [19].

III. SYSTEM ARCHITECTURE AND MATERIALS

The system architecture was organized to enable local signal acquisition, embedded preprocessing, identification of tremor patterns, data transmission, storage, and real-time visualization. The solution was built from commercially available hardware components, aiming to reduce cost, simplify replication, and facilitate future adaptations.

The materials included the T-Watch 2020 V2, USB Type-C cable, development computer, local Wi-Fi network, Visual Studio Code with PlatformIO, and ESP32/T-Watch libraries. The firmware was developed in C++, implementing accelerometer acquisition, calibration, digital filtering, vector magnitude calculation, spectral analysis, tremor detection, HTTP/BLE communication, and operating-state visualization on the device display.

The logical operating sequence is shown in Fig. 1. After initialization, the system configures the sensor, display, communication interfaces, and variables, then verifies connectivity, performs calibration, and selects the operating mode.

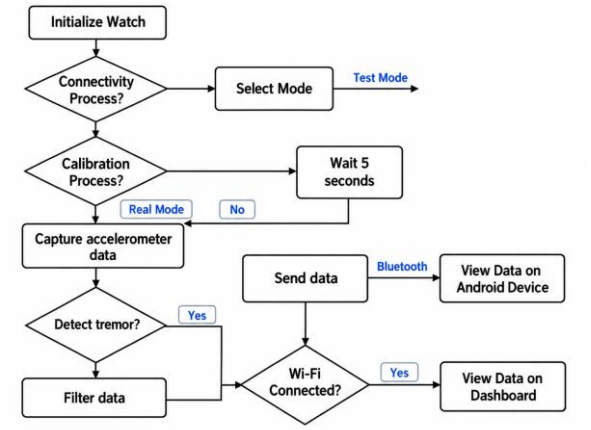


Fig. 1. Logical architecture of the wearable prototype for data acquisition, detection, transmission, and visualization.

In sensor-acquisition mode, acceleration data are acquired from the BMA423, while the synthetic-test mode generates artificial signals for processing validation. In both modes, samples are filtered and processed for magnitude calculation,

spectral analysis, tremor detection, transmission, and visualization.

The smartwatch-like T-Watch 2020 V2 provides a realistic usage scenario by allowing wrist attachment similar to commercial smartwatches, reducing mechanical assembly and improving usability during laboratory tests. During measurements, it was positioned with the display facing upward and consistently oriented relative to the forearm to ensure reproducible inertial signal acquisition. Fig. 2 presents the hardware setup, including the T-Watch 2020 V2, inertial sensing subsystem, and integrated processing and communication resources.



Fig. 2. Hardware elements used in the prototype: wearable platform, inertial sensor, and processing/communication module.

The T-Watch 2020 V2 integrates an ESP32 microcontroller, BMA423 three-axis accelerometer, AXP202 power-management circuit, TFT display, rechargeable battery, and Wi-Fi/Bluetooth Low Energy connectivity, enabling signal acquisition, local processing, user feedback, and wireless transmission. The BMA423 was configured at 100 Hz with a measurement range of ±2 g and 12-bit resolution across the X, Y, and Z axes, with values expressed in g [27]. These settings were maintained throughout the experiments to ensure consistent signal acquisition. TABLE II summarizes the experimental setup and the main prototype specifications.

TABLE II. EXPERIMENTAL SETUP AND SMARTWATCH PROTOTYPE SPECIFICATIONS.

Parameter	Value
Microcontroller	ESP32
Accelerometer	BMA423
Sampling frequency	100 Hz *(derived from SAMPLE_MS = 10 ms)*
Accelerometer range	±2 g
Device mass	58.5 g *(commercial LilyGO T-Watch 2020 V2 nominal mass)*
Battery capacity	350–400 mAh *(integrated 3.7 V Li-Po battery)*
Estimated autonomy	~4 h continuous operation *(1–3 days standby/optimized; USB-powered during tests)*
Acquisition window duration	1.28 s per FFT window *(128 samples at 100 Hz)*
Communication protocols	Wi-Fi (HTTP POST) / BLE (GATT)

The complete system architecture is presented in Fig. 3 and organized into five functional layers. The first layer comprises the wearable embedded device, responsible for acquiring X-, Y-, and Z-axis acceleration signals and executing the main firmware routines. The second performs preprocessing and feature extraction, including calibration, gravitational-component removal, noise reduction, moving-average filtering, FFT computation, and estimation of dominant frequency, amplitude, signal magnitude, and spectral energy. The third layer manages wireless communication. Processed data can be transmitted through Wi-Fi using HTTP requests and JSON messages sent to a REST API, or through BLE notifications based on the GATT protocol. This dual communication strategy allows the prototype to operate both with a local server and with nearby mobile or edge devices [22].

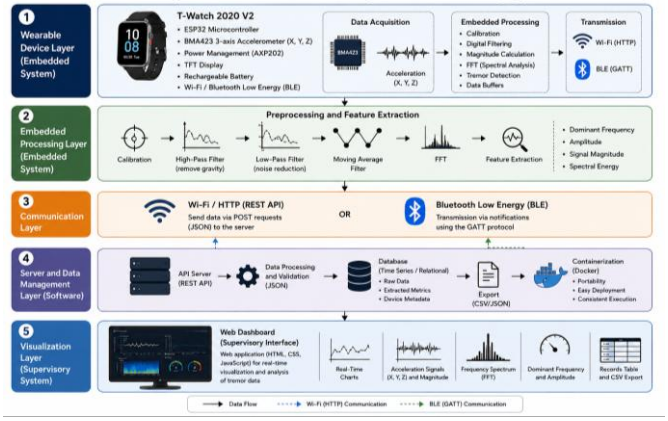


Fig. 3. Proposed layered architecture of the resting tremor monitoring system, integrating the wearable embedded device, local signal processing, wireless communication, server and database infrastructure, Docker containerization, and web-based supervisory interface.

The fourth layer comprises the software infrastructure for data reception and management. A server API receives and validates the transmitted records, which are then stored in a database together with the extracted metrics and device metadata. The backend and its dependencies are organized in Docker containers [17], improving portability, reproducibility, and consistency across different execution environments.

The fifth layer comprises the web-based supervisory system, displaying acceleration signals, vector magnitude, dominant tremor frequency, spectral information, device records, and CSV/JSON exports for real-time monitoring and analysis. The layered architecture separates hardware, firmware, communication, backend, storage, and visualization, facilitating maintenance, component replacement, remote integration, and future adoption of advanced signal-processing or machine-learning methods.

A. Functional firmware structure

The firmware uses independent routines for initialization, communication, acquisition, calibration, processing, detection, and transmission, improving maintainability and modular testing. Initialization configures the display, BMA423, Wi-Fi, BLE, and state variables, while the main loop handles sampling, filtering, magnitude calculation, and tremor detection. Calibration parameters are stored in non-volatile memory for reuse and can be updated to compensate

for wrist-position changes, which affect the accelerometer's gravitational component.

The simplified flow presented in Fig. 4 summarizes the main stages of embedded operation. Although it is an abstract representation, it shows the essential project sequence: initialize the device, collect accelerometer information, process samples, send data, and make visualization available. This view helps explain the transition between hardware, firmware, and web applications.

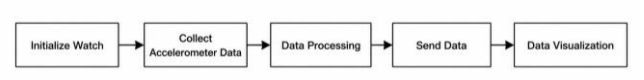


Fig. 4. Simplified flow of the prototype operation stages.

IV. SIGNAL ACQUISITION AND DIGITAL PROCESSING

Acquisition is performed from acceleration components along the three axes. Before analysis, the data undergo calibration to reduce reading deviations and compensate offsets. Preprocessing routines are then applied to attenuate noise, remove static components associated with gravity, and preserve variations compatible with tremor.

To improve repeatability, raw triaxial acceleration first undergoes static-bias calibration. During a stationary phase, the mean of N = 200 samples is estimated independently for each axis and subtracted from subsequent measurements: $a_c,i[n] = a_i[n] - \bar{a}_i$, for $i \in \{x, y, z\}$. The calibrated signals are then processed by first-order recursive filters to reduce the static gravitational component and high-frequency noise before feature extraction.

At the 100 Hz sampling rate (T_s = 10 ms), the firmware uses two first-order Infinite Impulse Response (IIR) stages. The DC-removal/high-pass stage follows $y_HP[n] = \alpha_HP \cdot (y_HP[n-1] + x[n] - x[n-1])$, with $\alpha_HP = 0.95$; this coefficient corresponds to a pole-equivalent corner of approximately 0.82 Hz and is used to suppress quasi-static gravity and slow drift. The smoothing low-pass stage follows $y_LP[n] = y_LP[n-1] + \alpha_LP \cdot (x[n] - y_LP[n-1])$, with $\alpha_LP = 0.20$. For this exact discrete-time implementation, the −3 dB frequency is approximately 3.57 Hz (the frequently used continuous-time approximation $\alpha/(2\pi T_s)$ gives 3.18 Hz). Because this corner lies below the 4–6 Hz target band, the implemented smoother attenuates—but does not eliminate—tremor-band components by approximately 3.5 to 5.8 dB. It is retained here because it is the filter actually used in the evaluated firmware; moving the low-pass corner above 6 Hz or redesigning the band-pass response is identified as a future optimization when amplitude preservation is required.

After filtering, the three spatial components are combined into an orientation-reduced scalar feature using the Euclidean norm $a_r[n] = \sqrt{(a_x,f^2[n] + a_y,f^2[n] + a_z,f^2[n])}$. This resultant dynamic magnitude is the primary time-series feature used for subsequent spectral analysis and detection. The approach reduces dependence on exact sensor orientation while preserving overall oscillatory activity [3], [10].

Resting-tremor-related activity is characterized in the frequency domain using a 128-point Fast Fourier Transform (FFT) applied to a Hamming-windowed frame. Rather than computing a fully normalized power spectral density, the firmware accumulates squared FFT magnitudes over discrete frequency bins. The relative spectral-energy ratio in the target band is defined as $R_4–6 = (\Sigma_\{f_k \in [4,6]\} |X[k]|^2) / (\Sigma_k |X[k]|^2)$.

This embedded formulation avoids additional normalization overhead while preserving the relative distribution of spectral energy needed for the decision logic. With f_s = 100 Hz and N = 128, the FFT-bin spacing is $\Delta f = f_s/N$ = 0.78125 Hz, which limits the resolution of the reported dominant frequency.

The decision rule compares $R_4–6$ with a predefined threshold T = 0.30. A tremor-like event is flagged when $R_4–6 > T$, i.e., when more than 30% of the spectral energy of the filtered resultant signal falls between 4 and 6 Hz. In this study, this threshold is an engineering criterion for functional testing and must not be interpreted as a clinically validated diagnostic threshold.

The embedded implementation prioritizes computational simplicity, ESP32 compatibility, and real-time execution. To evaluate the processing chain under controlled conditions, the runFakeDataMode() routine generates a 5.0 Hz synthetic tremor-like reference. The signal directly supplied to the detection routine is $m(t) = 1.8 + 0.6 \cdot \sin(2\pi \cdot 5t) + \eta(t)$, in g, where η(t) is a bounded pseudo-random term generated by the firmware with approximately ±0.05 g amplitude. Thus, the implemented test source uses sinusoidal modulation plus bounded uniform-like pseudo-random noise, not Gaussian noise. The signal is sampled at 100 Hz and processed in 128-sample windows (1.28 s); the parameters are summarized in Table III.

TABLE III. PARAMETERS OF THE SYNTHETIC SIGNAL GENERATED IN TEST MODE (RUNFAKEDATAMODE)

Parameter	Value	Description
Fundamental frequency (f_0)	5.0 Hz	Reference oscillation within the 4–6 Hz target band
Sampling frequency (f_s)	100 Hz	Sampling period T_s = 10 ms
Window size (N)	128 samples	FFT window duration = 1.28 s; Δf = 0.78125 Hz
Signal composition	Sinusoid + bounded pseudo-random noise	$m(t) = 1.8 + 0.6 \sin(2\pi \cdot 5t) + \eta(t)$
Target band	4.0–6.0 Hz	Spectral-energy band used by decision logic
Magnitude baseline	1.8 g	DC level used by the internal synthetic reference
Oscillation amplitude	0.6 g peak	Amplitude of the 5 Hz sinusoidal modulation
Noise term η(t)	approximately ±0.05 g	Bounded pseudo-random term generated from random(−100,100); not Gaussian
Detection threshold	T = 0.30 (30%)	Tremor-like event flagged when $R_{4–6} > T$

In addition to the sensor-acquisition mode, the firmware includes an internal synthetic-test mode that validates the processing chain without relying on volunteers or patients. This mode generates a controlled 5 Hz reference with known amplitude and bounded pseudo-random noise, allowing magnitude calculation, filtering, spectral analysis, detection logic, transmission, and dashboard updating to be checked under repeatable conditions. It is an engineering test source and does not represent physiological tremor.

V. TECHNICAL VALIDATION METHOD

The prototype uses two communication strategies. The first is based on HTTP over Wi-Fi, allowing measurements to be sent to a local or remote server. HTTP was chosen because of its broad adoption in web applications and because it facilitates integration with APIs, backend services, and dashboards [15], [20], [21]. The second strategy uses Bluetooth Low Energy (BLE), which is suitable for local tests, transmission to mobile devices, and low-power short-range communication supported by the ESP32 platform [13], [28].

The web dashboard was implemented to receive, display, and record the data transmitted by the device. The interface presents tremor magnitude in real time, a record table, detection status, battery level, and a CSV export option. These functions facilitate the organization and inspection of acquired measurements during experimental tests and subsequent analysis. Clear and organized data visualization is particularly relevant for monitoring and interpreting wearable-sensor measurements in digital-health applications [18].

To facilitate environment deployment and replication, the dashboard was containerized with Docker. Containerization allows the application, dependencies, and settings to be packaged into a portable unit, simplifying deployment on different computers and reducing environment incompatibility problems [17]. This approach contributes to the reproducibility of the prototype, particularly in academic and laboratory contexts.

Fig. 5 presents the dashboard interface. The chart allows monitoring of the temporal evolution of signal magnitude, while the table provides visualization of samples and detection states. The CSV export function facilitates subsequent analyses using statistical tools, spreadsheets, or processing scripts. This structure transforms raw accelerometer measurements into organized information for experimental analysis and validation.

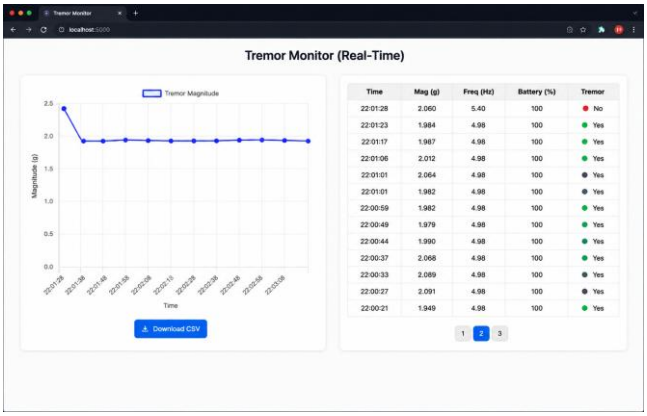


Fig. 5. Web dashboard for real-time visualization, record table, and data export.

A. Functional validation protocol and metrics

The initial validation protocol was designed to verify the consistency of the complete engineering chain, from accelerometer reading to presentation in the web interface. Tests were conducted under controlled conditions using three situations: normal/voluntary wrist activity, manually generated tremor-like wrist oscillations, and the firmware's internal synthetic test mode. The manual tremor-like condition evaluated acquisition and detection through controlled human movement but is treated strictly as a functional movement test, not as physiological or clinical validation. The synthetic mode provided a known 5 Hz reference for evaluating detection logic, transmission, and dashboard updating.

The evaluation considered acquisition stability, dashboard updating, device-state coherence, sample logging, CSV export, display response, and communication continuity. The end-to-end (E2E) path comprises BMA423 acquisition over I²C, ESP32 filtering and FFT processing, HTTP POST or BLE transmission, backend reception, and dashboard rendering. E2E latency was defined as $L_{E2E} = t_{display} - t_{sample}$, where t_{sample} is the timestamp associated with the acquired sample/window and $t_{display}$ is the corresponding dashboard-update timestamp. The 1.28 s FFT accumulation window is reported separately from communication/display latency.

Packet loss was defined as $PL(\%) = 100 \cdot (N_{tx} - N_{rx})/N_{tx}$, where N_{tx} is the number of transmitted records and N_{rx} is the number successfully received and processed. Sequence counters in the transmitted payloads were used to identify missing records. Signal evaluation considered resultant magnitude, temporal behavior, dominant frequency, and FFT spectral concentration in the 4–6 Hz band [2], [3], [10].

The prototype is not intended to replace medical assessment. The present measurements quantify engineering behavior under controlled functional conditions and provide a basis for future comparison with standardized clinical observations and reference devices.

In addition to functional tests, the reproducibility of the software environment was evaluated. PlatformIO facilitates firmware compilation across different computers, while the Docker-based dashboard packages application dependencies into a portable execution environment [17]. This organization supports academic reproducibility by allowing researchers to replicate the architecture, replace components, and evaluate new processing strategies with reduced configuration effort. The modular design integrates wearable hardware, embedded firmware, communication, signal processing, and data visualization into a functional and reproducible system architecture.

Wireless transmission integrity was evaluated by comparing the 5 Hz synthetic reference generated by runFakeDataMode() with the series stored/exported by the dashboard. Root-mean-square error was computed as $RMSE = \sqrt{[(1/N)\cdot\Sigma_i(x_i - y_i)^2]}$, where x_i is the reference sample and y_i is the corresponding received/stored sample.

The resulting RMSE was approximately zero within floating-point precision, indicating that the tested serialization and network path did not introduce measurable amplitude distortion. This result is specific to the synthetic transmission test and does not imply absence of sensor noise,

packet loss under all network conditions, or clinical measurement accuracy.

VI. RESULTS AND DISCUSSION

The wearable prototype was evaluated using BMA423 acceleration data sampled at 100 Hz and analyzed in 128-sample windows (1.28 s) with a 128-point FFT. The detection decision was based on dominant frequency and relative spectral energy in the 4–6 Hz band. Normal/voluntary movement conditions produced frequencies below 2.34 Hz and R_4–6 = 12.4% ± 4.2%. The internal synthetic test signal produced R_4–6 = 88.5% ± 3.1%, whereas manually generated tremor-like wrist movements produced R_4–6 = 41.8% ± 5.6%. Controlled tremor-like movements yielded dominant frequencies of 4.69–5.47 Hz and exceeded the predefined 30% threshold. These measurements demonstrate the behavior of the engineering detection criterion under the tested conditions; they do not constitute physiological or clinical tremor validation.

The signal-processing pipeline was evaluated through static calibration, digital filtering, and spectral analysis. At 100 Hz with N = 128, the FFT frequency resolution is 0.78125 Hz. Relative to the 5.00 Hz synthetic reference, observed dominant-bin errors were 0.31–0.47 Hz (6.2%–9.4%), consistent with the finite FFT resolution. Post-calibration residual noise remained below 0.008 g and the measured signal-to-noise ratio exceeded 31 dB under the reported test conditions.

The communication subsystem was evaluated using HTTP REST over Wi-Fi and Bluetooth Low Energy (BLE), considering processing time, end-to-end latency, packet loss, and data integrity. Embedded processing required less than 4.2 ms per 1.28 s analysis window. Average communication/display latency was 42 ± 12 ms for HTTP/Wi-Fi and 18 ± 5 ms for BLE. During 30-min tests with 1,406 transmissions, packet loss was 0% for BLE and 0.14% for Wi-Fi. Comparison between the synthetic reference and the stored dashboard series yielded RMSE ≈ 0 within floating-point precision, indicating no measurable amplitude distortion attributable to serialization and transport in that test.

The integrated prototype was also evaluated through controlled manual movement tests and continuous operation of the embedded and web infrastructure. Voluntary arm movements produced predominantly low-frequency components below 2.0 Hz and spectral ratios below 15%, whereas manually generated tremor-like wrist movements showed dominant frequencies of 4.69–5.47 Hz and spectral ratios above the 30% detection threshold. Under these controlled conditions, no false-positive triggering was observed for the evaluated voluntary movements. During operation, the display updated every 1.28 s and changed state when the criterion was exceeded, while the detection state was simultaneously transmitted to the backend. The corresponding device states are illustrated in Fig. 6.

The Flask-based web infrastructure successfully received and processed JSON data, enabling visualization of acceleration magnitude, frequency, spectral information, and device status. Overall, the results demonstrate the functional integration of sensing, embedded processing, tremor detection, wireless communication, and remote visualization in a single wearable platform. As the experiments used simulated signals and controlled functional tests rather than data from individuals with Parkinson's disease, the results represent a technical validation of the prototype rather than clinical validation of tremor detection.



Fig. 6. States displayed by the device during tests: normal operation, tremor detected, and simulated mode.

VII. REPLICABILITY ANALYSIS AND APPLICATION POTENTIAL

The proposed architecture was conceived with a focus on replicability. The use of a commercial wearable platform, widely available libraries, and open development tools reduces barriers for reproducing the system in academic activities, embedded systems laboratories, and digital health research. This characteristic differentiates the prototype from closed or proprietary solutions, which often hinder modification of acquisition and analysis algorithms.

The ESP32 supports system expansion through Wi-Fi, BLE, I2C, GPIOs, and low-power modes, enabling new sensors, sampling adjustments, local alerts, smartphone connectivity, and home-based applications [22], [28]. Wi-Fi/HTTP facilitates dashboard and server integration [15], [20], [21], while BLE supports short-range mobile connectivity [13], [28]. The dashboard enables real-time monitoring, detection-state verification, communication-failure identification, and data export, with future extensions including frequency charts, user history, automated reports, and session comparison based on user-centered visualization principles [18].

In digital health, the prototype supports educational and experimental applications by integrating sensing, embedded processing, web visualization, and CSV export for reproducible analysis in Python, MATLAB, spreadsheets, or statistical tools. This architecture enables future filtering, feature extraction, detection/classification, and comparison with clinical scales, subject to ethical, privacy, and data-security requirements.

Although it is a prototype, the integration between wearable hardware, embedded processing, and web application demonstrates a viable path for low-cost solutions. Replacing or complementing the BMA423 with sensors including gyroscopes, increasing battery autonomy, and adding communication encryption may raise the technological maturity of the solution. Likewise, incorporating machine learning may allow distinction among resting tremor, essential tremor, voluntary movements, and artifacts of daily use.

Future development should address sensitivity to voluntary movements, lack of clinical validation, and energy consumption through adaptive filtering, robust feature extraction, machine-learning classifiers, and optimized Wi-Fi, BLE, and display operation. Further work should include clinical validation with volunteers and clinical scales, authentication and encryption, integration with clinical

information systems, and dashboard expansion with history, reports, and session-based metrics.

VIII. LIMITATIONS AND FUTURE WORK

Despite the positive results, the system has limitations that must be considered. Detection based on a fixed spectral-energy threshold may be insufficient in real-world scenarios with complex movements, different watch positions, and individual variability. The implemented $\alpha_LP = 0.20$ smoothing filter has an exact discrete-time −3 dB frequency of approximately 3.57 Hz at 100 Hz sampling and therefore attenuates the 4–6 Hz target band; although detection remained functional in the reported tests, a revised low-pass or band-pass design with a corner above 6 Hz should be evaluated when quantitative amplitude preservation is required. Energy autonomy also requires dedicated prolonged-use measurement with the sensor, display, Wi-Fi/BLE, and local processing simultaneously active.

The current firmware has no local storage buffer or message-retry mechanism; therefore, samples processed during Wi-Fi or BLE disconnections are discarded. Nevertheless, the ESP32 continues acquisition, FFT processing, and real-time detection, updating status and frequency on the T-Watch TFT display. BLE automatically restarts advertising and reconnects after disconnection, whereas Wi-Fi/HTTP only checks connectivity through isWiFiConnected() and skips transmission when unavailable, with no explicit dashboard alert. This limitation is acknowledged as future work, including circular buffering, resilient reconnection, and data retransmission.

Future work will clinically validate the system using individuals with Parkinson's disease and age-matched controls, with standardized resting tests under different medication and postural conditions. Measurements will be compared with UPDRS Part III assessments and a validated medical-grade wearable, analyzing frequency, amplitude, spectral energy, duration, and cross-device agreement through correlation, Bland–Altman analysis, sensitivity, and specificity. This protocol is planned for future validation and is not part of the present experimental results.

Amplitude repeatability metrics (mean, standard deviation, and coefficient of variation) were not collected in the original trials, which focused on real-time mean magnitude, FFT peak frequency, and tremor detection. As these parameters cannot be retrospectively derived without baseline records, dedicated controlled tests are proposed for future evaluation.

As future work, embedded artificial intelligence algorithms are recommended for classifying tremor patterns, including lightweight models capable of running on the ESP32 or edge devices. The dashboard may also be expanded with clinical metrics, session reports, patient history, configurable filters, and integration with digital health platforms. Another improvement involves the use of security and authentication protocols for the transmission of sensitive data, especially in clinical or home-based applications.

Finally, it is suggested to compare the prototype with reference instruments, expand the test dataset, and investigate the combined use of accelerometer and gyroscope. Including contextual data, such as time, medication, posture, and activity performed, may increase the clinical relevance of records. These extensions can transform the prototype into a research platform for biomedical engineering and digital health.

IX. CONCLUSION

This work developed and technically evaluated a wearable prototype based on the T-Watch 2020 V2, integrating ESP32, BMA423, Wi-Fi, BLE, FFT processing, and a containerized web dashboard. The system captured and processed acceleration signals, identified oscillations in the typical 4–6 Hz range under controlled manual and synthetic test conditions, and provided real-time visualization, achieving its objective as a low-cost, portable proof of concept. Nevertheless, the study did not include clinical validation with individuals with Parkinson's disease; functional evaluation was limited to synthetic signals and controlled manually generated tremor-like wrist movements, representing an initial technical validation of the proposed architecture. Future work should include formal clinical validation with patients, improve algorithm robustness against voluntary movement artifacts, optimize power consumption, implement health-data encryption, and investigate intelligent classification methods to support applications in clinical and digital health research.

KNOWLEDGMENTS

The authors thank the State University of Amazonas, the School of Technology, and the Electronic Engineering program for academic support during the development of this work.

REFERENCES

[1] C. E. Tielleman-Wallace et al., "The epidemiology of Parkinson's disease: incidence, prevalence and risk factors," Age and Ageing, vol. 52, no. 6, afad038, 2023.

[2] G. Deuschl et al., "The cerebral oscillatory network of parkinsonian resting tremor," Brain, vol. 126, no. 1, pp. 199-214, 2003.

[3] L. Sigcha et al., "Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review," Expert Systems with Applications, vol. 229, Part A, 120541, 2023.

[4] H. Dai, P. Zhang, and T. C. Lueth, "Quantitative Assessment of Parkinsonian Tremor Based on an Inertial Measurement Unit," Sensors, vol. 15, no. 10, pp. 25055-25071, 2015, doi: 10.3390/s151025055.

[5] A. Antonini et al., "Toward objective monitoring of Parkinson's disease motor symptoms using a wearable device: wearability and performance evaluation of PDMonitor," Frontiers in Neurology, vol. 14, 1080752, 2023.

[6] A. Bougea, "Application of Wearable Sensors in Parkinson's Disease: State of the Art," Journal of Sensor and Actuator Networks, vol. 14, 23, 2025.

[7] M. A. Terry and E. Alvarez Vazquez, "Efficacy of Wearable Devices in Monitoring and Treating Parkinson's Disease Symptoms: A Systematic Review and Meta Analyses," IEEE Access, vol. 13, pp. 84240-84258, 2025.

[8] M. Gregorčič and D. Georgiev, "The Usefulness of Wearable Sensors for Detecting Freezing of Gait in Parkinson's Disease: A Systematic Review," Sensors, vol. 25, no. 16, Art. no. 5101, 2025.

[9] C. Moreau et al., "Overview on wearable sensors for the management of Parkinson's disease," npj Parkinson's Disease, vol. 9, Art. no. 153, 2023.

[10] S. Patel et al., "Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 6, pp. 864-873, 2009.

[11] F. M. Rast and R. Labruyère, "Systematic review on the application of wearable inertial sensors to quantify everyday life motor activity in people with mobility impairments," Journal of NeuroEngineering and Rehabilitation, vol. 17, no. 148, 2020.

[12] M. F. Nozella, "Wearable device prototype using an inertial sensor for real-time detection and visualization of resting tremor in people with Parkinson's disease," Undergraduate thesis, Federal University of Uberlândia, 2022, in Portuguese.

[13] Espressif Systems, "ESP32 Bluetooth architecture," 2024. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32_bluetooth_architecture_en.pdf

[14] LilyGo, "T-Watch-2020 V2 Documentation," 2020. [Online]. Available: https://t-watch-document-en.readthedocs.io/en/latest/introduction/product/2020.html

[15] S. Oliveira, Internet of Things with ESP8266, Arduino, and Raspberry Pi. São Paulo: Novatec, 2017, in Portuguese.

[16] A. P. S. P. B. da Silva, "The use of inertial sensors for characterization and classification of wrist tremor in individuals with Parkinson's disease and correlation with the subjective assessment scale: UPDRS," Ph.D. dissertation, Federal University of Uberlândia, 2018, in Portuguese.

[17] Docker Inc., "Docker overview," 2024. [Online]. Available: https://docs.docker.com/get-started/overview/

[18] S. Sadhu, D. Solanki, L. A. Brick, N. R. Nugent, and K. Mankodiya, "Designing a Clinician-Centered Wearable Data Dashboard (CarePortal): Participatory Design Study," JMIR Formative Research, vol. 7, e46866, 2023, doi: 10.2196/46866.

[19] X-Powers, "AXP202 Single Cell Li-Battery PWM Charger and Power System Management IC," Datasheet, rev. 1.0. [Online]. Available: https://linux-sunxi.org/images/2/20/AXP202_Datasheet_v1.0_en.pdf

[20] MDN Web Docs, "Overview of HTTP," Mozilla Foundation. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/Overview

[21] Cloudflare, "What is HTTP?," Cloudflare Learning Center, 2025. [Online]. Available: https://www.cloudflare.com/learning/ddos/glossary/hypertext-transfer-protocol-http/

[22] J. W. Santos and R. C. Lara Junior, "Low-cost residential automation system controlled by the ESP32 microcontroller and monitored via smartphone," 2019, in Portuguese.

[23] M. Wendling, "Sensors," São Paulo State University, 2010, in Portuguese.

[24] C. Godinho et al., "A Systematic Review of the Characteristics and Validity of Monitoring Technologies to Assess Parkinson's Disease," Journal of NeuroEngineering and Rehabilitation, vol. 13, Art. no. 24, 2016.

[25] P. Pierleoni, A. Belli, O. Bazgir, L. Maurizi, M. Paniccia, and L. Palma, "A Smart Inertial System for 24h Monitoring and Classification of Tremor and Freezing of Gait in Parkinson's Disease," IEEE Sensors Journal, vol. 19, no. 23, pp. 11612-11623, 2019.

[26] D. Lukšys, G. Jonaitis, and J. Griškevičius, "Quantitative Analysis of Parkinsonian Tremor in a Clinical Setting Using Inertial Measurement Units," Parkinson's Disease, vol. 2018, Art. no. 1683831, 2018.

[27] Bosch Sensortec, "BMA423 Digital, Triaxial Acceleration Sensor," Data Sheet, rev. 2.0, Aug. 2019, document BST-BMA423-DS004-00. [Online]. Available: https://media.digikey.com/pdf/Data%20Sheets/Bosch/BMA423_Rev2.0_Aug2019.pdf

[28] Espressif Systems, "ESP32 Series Datasheet," ver. 5.3, 2026. [Online]. Available: https://www.espressif.com/documentation/esp32_datasheet_en.pdf

Development of a Wireless Wearable System for Temporal and Angular Human Gait Analysis Based on the ESP-NOW Protocol

Davi Gabriel dos Santos Oliveira
Faculty of Medicine
Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
oliveira_davi@ufg.br

Ana Carolina Q. Siravenha
Faculty of Electrical and Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Pará, Brazil
siravenha@ufpa.br

Bianca Vitoria Silva Pina
Faculty of Electrical and Biomedical Engineering
Federal University of Pará (UFPA)
Belém, Pará, Brazil
bianca.pina@itec.ufpa.br

Amanda A. Arrais de Garcia
Institute of Physics
Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
amandaalencar@discente.ufg.br

Emerson Nobuyuki Itikawa
Institute of Physics
Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
emersonitikawa@ufg.br

Abstract*—Quantitative gait analysis supports the diagnosis and rehabilitation of motor disorders. The optoelectronic systems taken as the gold standard, however, are expensive and tied to specialized laboratories. This work presents the development and bench validation of a low-cost, wireless, wearable system for angular and temporal gait measurement. Sensor nodes built around the MPU-6050 and the ESP32 talk to a receiver over ESP-NOW in a star topology, so nothing is cabled to the body. The knee flexion–extension angle is reconstructed on-line on the receiver node through a per-segment complementary filter ($\alpha = 0.98$) and streamed to a monitoring application, while gait-cycle segmentation is performed offline by gyroscope-based peak detection. Two independent protocols were applied. Angular accuracy was assessed statically on a rigid bench fixture against a digital protractor ($\pm 0.5^{\circ}$), which characterizes the measurement subsystem in isolation from soft-tissue artifact. A first acquisition showed the prototype underestimating every angle by 2.37°. Inspection traced this to a tilted sensor package; the component was resoldered and the acquisition repeated. The mean absolute error then fell below the resolution of the reference instrument, to 0.025°, with a residual that no longer favoured either direction and an unchanged within-angle repeatability of 0.056°. Both figures bound the residual error rather than quantify it. The swing duration derived from the detected gait events was compared with two-dimensional videogrammetry during slow-cadence walking, yielding a mean absolute error of 0.02 s, of the order of one video frame at 60 frames per second; absolute event instants on a common time base were not compared. The wireless link was characterized separately during walking over 211,007 packets: no packet was lost, and the standard deviation of the inter-arrival interval remained below 4 ms. Link-level latency, measured over 1000 frames per node, had a median one-way value of 1.0–1.2 ms with a 99th percentile below 13 ms, an order of magnitude under the limit specified for medical wireless body area networks. Dynamic knee kinematics were not validated against a kinematic reference; this is stated as a limitation. With a total material cost of approximately BRL 94 (about USD 17), the results support the technical feasibility of the wireless architecture as a basis for future clinical validation.

***Index Terms*—Gait analysis, wearable system, ESP32, ESP-NOW, inertial sensors, wireless body area network.**

I. Introduction

Human gait depends on the integrated action of the central nervous system, the sensory system, and the musculoskeletal apparatus. Changes in the walking pattern are frequently among the first observable indicators of functional decline or underlying pathology, and walking speed has been proposed as an additional clinical indicator in geriatric assessment [1]. Neurological conditions such as stroke and Parkinson's disease, orthopedic disorders, and natural aging alter gait parameters and increase the risk of falls, a relevant public health problem among the elderly [2].

Three-dimensional optoelectronic systems combined with force platforms remain the reference method, and they deliver millimetric precision. They also cost tens of thousands of dollars and need fixed laboratory space, trained operators, and controlled capture conditions [3]. In Brazilian public services and university hospitals, that combination keeps quantitative gait assessment out of reach.

Microelectromechanical systems (MEMS) inertial sensors make compact, low-power wearables possible at a fraction of that cost [4]. Going wireless lifts the spatial constraint of fixed cameras, so data can be collected in corridors, outdoors, or at home [5]. Among the available wireless technologies, the ESP-NOW protocol developed by Espressif Systems operates in the 2.4 GHz band without requiring network infrastructure or association handshakes, allowing direct peer-to-peer transmission with low reported latency and reduced energy consumption [6], [7].

This work presents the development and bench validation of a low-cost, wireless, wearable system for the acquisition of

kinematic gait parameters, based on inertial sensor nodes controlled by ESP32 microcontrollers and communicating through ESP-NOW. In contrast to a purely offline pipeline, the knee angle is reconstructed on-board the receiver node and made available on-line through a dedicated monitoring interface. The primary contribution reported here is the wireless, cable-free node architecture and its metrological characterization on a bench fixture; clinical gait accuracy is explicitly outside the scope of this validation and is identified as the next stage of the work.

II. Materials and Methods

A. System Architecture

The system is a Wireless Body Area Network (WBAN) in a star topology [7]: mobile sensor nodes on the body segments and a fixed concentrator node. Each node samples its inertial sensors and transmits raw packets over ESP-NOW to the master. Two configurations are used, validated independently (Fig. 1).

In the *temporal configuration* (Fig. 1a), a single node is placed on the distal shank (lateral tibia). The master acts as a data concentrator that forwards the raw stream over the serial interface, and the gait-event segmentation is performed offline. Because a single camera cannot record both limbs without occlusion, each lower limb was instrumented and validated separately, one at a time.

In the *angular configuration* (Fig. 1b), two nodes are placed on the thigh (distal femur) and shank (proximal tibia), aligned with the knee flexion axis. Here the master does more than concentrate: it reconstructs the knee flexion–extension angle on-board and streams it to a workstation for live display and recording. This on-board reconstruction is the main architectural change with respect to the earlier, fully offline version.

B. Hardware Development

Each node integrates an ESP32 microcontroller and an MPU-6050 inertial sensor in a compact assembly powered by an autonomous supply composed of a lithium battery managed by a TP4056 charging module and controlled by a physical on/off switch. The circuit was designed in Proteus Design Suite, with the schematic drawn in the ISIS environment and the printed circuit board (PCB) layout defined in the ARES environment. Communication between the ESP32 and the MPU-6050 used the I2C bus operating at 3.3 V, with the clock line (SCL) on GPIO 22 and the data line (SDA) on GPIO 23.

A ground plane covers the empty areas of the board to mitigate electromagnetic interference between the radiofrequency transceiver and the high-sensitivity MEMS sensor [11]. The layout was transferred to the substrate by laser mask ablation, which gives finer track resolution than conventional thermal methods [12], and the exposed copper was removed in ferric chloride. The assembly was encapsulated in heat-shrink tubing, and adjustable hook-and-loop straps fix it to the body segment while limiting motion artifacts [13].

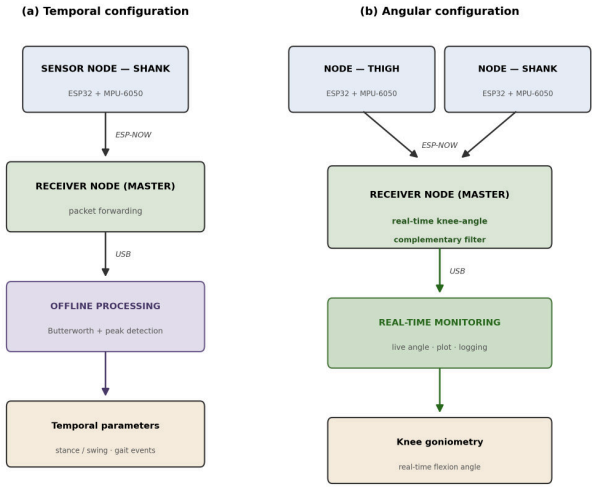


Fig. 1. Block diagram of the two operating configurations: (a) the temporal configuration, with a single shank node and offline event segmentation; (b) the angular configuration, with thigh and shank nodes and on-line knee-angle reconstruction on the receiver node.

C. Firmware and Link Characteristics

Each sensor node configures the MPU-6050 with a full-scale range of ± 2 g for the accelerometer and $\pm 250^\circ$/s for the gyroscope, with the internal digital low-pass filter enabled at approximately 44 Hz as an anti-aliasing mechanism [14]. It should be noted that this cutoff sits close to the 50 Hz Nyquist limit of the 100 Hz output rate, so the attenuation available between 44 and 50 Hz is modest and the anti-aliasing margin is narrow; components above the Nyquist frequency are attenuated rather than suppressed. The ± 2 g range was selected to maximize resolution for the inclination estimate that anchors the complementary filter, since the fusion relies on the gravity vector; this choice, however, may saturate the accelerometer during high-impact transients such as heel strike, and the vertical accelerometer axis is therefore used only as a qualitative corroboration of the detected events, never as a quantitative input to the angle reconstruction. An automatic 3 s calibration routine, executed with the node stationary, estimates the static gyroscope bias, which is subtracted from subsequent readings; the accelerometer output retains the gravity component required for inclination estimation. Each node reads the six inertial axes at a nominal rate of 100 Hz, controlled by timer interrupts, and transmits them as raw data packets through ESP-NOW. This sampling rate is justified by the Nyquist theorem and by the spectral characteristics of human movement, whose kinematic power lies mostly below 6 Hz, providing sufficient oversampling to capture fast transients [15].

The receiver node operates in continuous reception mode. A single packed data structure holds the six inertial axes, a per-node timestamp, and an explicit side identifier that unambiguously routes each packet to the thigh or shank buffer. Nodes do not share a common time base: each node timestamps

its own samples from its local timer, and the master aligns the two segment streams by arrival order, which is adequate for the sagittal difference of (2) at the cadences tested but does not constitute clock synchronization in the strict sense. The receiver firmware includes a characterization mode that timestamps every reception with microsecond resolution and logs it together with the originating node timestamp, which supports the direct measurement of packet delivery, inter-arrival jitter, and relative clock drift reported in Section III-A. The difference between the two timestamps cannot yield absolute latency, since it contains an unknown clock offset. Link-level latency was therefore measured separately through the transmission-complete callback of the ESP-NOW stack, which fires when the frame has been transmitted and acknowledged at the MAC layer: the receiver learns each node address from the incoming stream, then issues 1000 frames of the same 29-byte payload to one node at a time and records the elapsed time on a single clock, avoiding any dependence on inter-node synchronization. The resulting figure is a MAC-level round-trip time and does not include sensor sampling or host-side processing, so it bounds the transport contribution rather than the full acquisition-to-display path.

In the temporal configuration, the master forwards the received stream for offline segmentation. In the angular configuration, the master updates a per-segment complementary filter (Section II-D) and outputs the reconstructed knee angle over the USB serial port at 115200 baud in a lightweight comma-separated format; a single-character serial protocol then allows the operator to reset the angular reference (pose-based zeroing) and to select the sagittal rotation axis without reflashing the device.

D. Signal Processing

The processing was divided into two paths with distinct timing requirements. The angular reconstruction runs on-line on the receiver node, whereas *the temporal event segmentation is performed entirely offline*, in Python (Google Colab), after the acquisition session; no gait event is detected on-line by the present firmware. Sensitivity factors of 16384 LSB/g for the accelerometer and 131 LSB/°/s for the gyroscope were applied to convert raw digital values into physical units [14].

For each segment, a complementary filter fuses the low-frequency stability of the accelerometer with the fast response of the gyroscope, correcting the gyroscope drift [16], [17]:

$$\theta(t) = \alpha\,[\theta(t-\Delta t) + \omega\,\Delta t] + (1-\alpha)\,\theta_{\mathrm{acc}}, \qquad (1)$$

with Δt derived from the packet timestamps. The weighting factor was set to $\alpha = 0.98$, which places the gyroscope–accelerometer crossover near 0.3 Hz at the 100 Hz update rate, below the 0.5–6 Hz band of interest for gait. It was chosen empirically as the lowest weighting that held a flat plateau during static postures without admitting visible accelerometer noise. Drift correction is thus continuous and implicit: the accelerometer term permanently re-anchors the integrated gyroscope estimate, so no high-pass detrending or periodic re-zeroing is applied during a session. The accelerometer inclination



Fig. 2. Monitoring interface during a sustained flexion, showing the instantaneous knee angle, the rolling angle plot and the link status.

θ_{acc} is obtained from the gravity vector through the `atan2` function in the sagittal plane; the accelerometer–gyroscope innovation is wrapped to the $[-180°, 180°]$ interval before fusion, ensuring continuity across the angular discontinuity of `atan2` at large flexion angles. The knee flexion–extension angle is obtained by the algebraic difference between the shank and thigh angles, referenced to a static pose calibration acquired in the neutral position [18], as expressed in (2):

$$\theta_{\mathrm{knee}}(t) = \theta_{\mathrm{shank}}(t) - \theta_{\mathrm{thigh}}(t) - \theta_{\mathrm{offset}} \qquad (2)$$

where θ_{offset} is captured on demand through the serial reset command. Because the fusion preserves the accelerometer anchoring without high-pass detrending, the reconstructed angle sustains static postures without drifting back toward zero, a prerequisite for goniometric use.

For the temporal path, a fourth-order Butterworth low-pass filter with a 6 Hz cutoff was applied offline to the gyroscope signal [15]. A peak-detection algorithm identified the large-amplitude valleys corresponding to mid-swing; from this reference, search windows located the heel strike (zero crossing of angular velocity immediately after the valley) and the toe-off (positive angular-velocity peak immediately before the valley) [19]. Cadence, swing time, and stride time were then computed automatically.

E. Monitoring Interface

A desktop application receives the serial stream, displays the instantaneous knee angle and a rolling time plot, triggers the pose-based zeroing, and logs the data to file. Fig. 2 shows it during a sustained flexion, with the stable plateau characteristic of the on-board complementary filter. It confirmed the feasibility of on-line operation, but the quantitative validation reported below relied on controlled acquisitions rather than on the interface.

F. Experimental Protocols

Experiments were conducted in a controlled indoor environment on a flat, obstacle-free surface under constant artificial lighting. Data were collected through self-experimentation with a single healthy volunteer with no history of musculoskeletal or neurological disorders, in accordance with the

ethical principles of CNS Resolution No. 466/2012. Because the study consisted of technical instrumentation validation without clinical intervention, it was treated as an engineering bench and functionality test.

In Protocol 1 (dynamic, temporal parameters), a single node was fixed to the lateral shank (distal tibia), chosen for its higher angular acceleration during swing [17]. A tripod-mounted camera perpendicular to the sagittal plane at 2.5 m recorded at 60 frames per second [3], and a strong vertical foot stomp before walking synchronized video and sensor signals [20]. The subject walked barefoot along a linear path at a slow, controlled pace, five alternating steps, capturing two complete cycles of the instrumented limb within the camera field of view. This protocol assesses the *duration between detected gait events* only; the knee angle was not reconstructed or compared against any reference during walking.

In Protocol 2 (static, angular parameters), accuracy was assessed on a bench fixture rather than on the limb, so that reference angles could be imposed with known uncertainty. The two modules were rigidly fixed to the arms of a digital protractor ($\pm 0.5^{\circ}$) acting as an articulated ground truth, each aligned to one arm so the inter-arm angle matched the reconstructed joint angle. The fixture is mechanically repeatable but not traceable to a national standard; a calibrated rotary stage is identified as future work. Four target angles spanning the functional flexion range (30°, 60°, 90°, 120°) were set, with five repetitions each and a full zero recalibration between repetitions. Each posture was held stationary and the mean over the stable plateau compared against the protractor reading.

G. Validation Procedure

Validation has two stages that answer different questions and must not be conflated. The static stage characterizes the *angular precision and trueness of the measurement subsystem* on the bench: reconstructed angles were compared with the protractor readings, and absolute and relative errors, within-angle repeatability and a linear regression against the reference were computed. The dynamic stage characterizes the *duration between discrete gait events* during walking, with heel strike and toe-off identified frame by frame in Kinovea [21], [22]; absolute event instants on a common time base were not compared. Dynamic angular accuracy, the agreement of the continuous knee-angle trajectory against a kinematic reference, was not assessed in either stage. Mean absolute error (MAE) is the primary metric in both domains [23], [24].

III. Results and Discussion

A. Wireless Link Characterization

The low-latency premise behind the architecture had never been checked against measurement, so the link was characterized during walking. This is both the condition of interest and the one where body shadowing hurts a 2.4 GHz channel most. Both nodes transmitted at the nominal 100 Hz while the subject walked the path used in Protocol 1. Three repetitions were acquired under normal laboratory conditions and a fourth with a deliberate 2.4 GHz interferer, a mobile device sustaining a large download, about 1 m from the nodes. Packet delivery was inferred from discontinuities in the per-node timestamp stream. Jitter was taken as the standard deviation of the reception intervals, with the intervals adjacent to a discontinuity excluded. Relative clock drift came from the slope of the difference between reception time and node timestamp across the session. In total, 211,007 packets were analyzed across the four acquisitions.

Table I summarizes the results. No packet was lost on either node in any acquisition, giving a packet delivery ratio of 100% under movement and comfortably above the 90% specified for wireless body area networks [8]. Across the three repetitions, inter-arrival jitter was 3.63 ± 0.40 ms for the thigh node and 2.33 ± 0.41 ms for the shank node, an order of magnitude below the 50 ms jitter limit of the same standard [8] and below the 16.7 ms frame interval of the video reference used in Section III-D; over 99.5% of the reception intervals fell within one video frame of the nominal 10 ms period. The interval distribution carries a long upper tail, with 99th percentiles between 16 and 22 ms. That is what one would expect from the slotted, retransmission-based channel access reported for ESP-NOW in field evaluations [9]. Dispersion across repetitions was small, so the link behaves reproducibly under this walking protocol.

Link-level latency was measured over 1000 frames per node, with no transmission failure or timeout in either case. The round-trip time distribution is strongly right-skewed and is best described by its median and percentiles rather than its mean: the median one-way latency was 1.22 ms for the thigh node and 1.04 ms for the shank node, rising to 8.79 ms and 7.38 ms at the 95th percentile and to 12.87 ms and 11.72 ms at the 99th percentile. A pronounced floor near 1.38 ms round trip, corresponding to about 0.69 ms one way, held 32% and 41% of the frames. This is of the same order as the 1.06 ms reported for a single-byte payload in a controlled comparison of ESP-NOW against other protocols [10], although the two figures are not strictly comparable because the reference does not state whether its value is a one-way or a round-trip delay, and the payload sizes differ. Above this floor the distribution decays smoothly rather than in discrete steps, which is compatible with a variable number of channel-access attempts under the slotted access reported for ESP-NOW [9], although the present data do not resolve individual retransmission events. Taking one video frame at 60 frames per second as the reference scale, 99.7% and 100% of the frames respectively had a one-way latency below that interval. All values remain more than an order of magnitude below the 125 ms latency specified for medical applications in wireless body area networks [8].

Relative clock drift was -50.8 ± 2.9 ppm for the thigh node and -67.0 ± 1.4 ppm for the shank node. Each node repeated its own value far more closely than the two nodes agreed with each other, which points to the crystal oscillator rather than to a measurement artifact. This quantifies the absence of a shared time base noted in Section II-C: at 100 Hz, a drift of this magnitude accumulates less than one sampling

TABLE I
ESP-NOW LINK CHARACTERIZATION DURING WALKING (TWO NODES AT 100 HZ)

Metric	Thigh	Shank	Req.
Three repetitions, normal conditions			
Packets analyzed	75558	75538	–
PDR (%)	100.00	100.00	> 90
Jitter σ (ms)	3.63 ± 0.40	2.33 ± 0.41	< 50
Jitter P99 (ms)	20.6–22.5	16.5–18.5	–
Clock drift (ppm)	-50.8 ± 2.9	-67.0 ± 1.4	–
Link-level latency, 1000 frames per node			
Latency median (ms)	1.22	1.04	< 125
Latency P95 (ms)	8.79	7.38	–
Latency P99 (ms)	12.87	11.72	–
Tx failures	0	0	–
Single acquisition, 2.4 GHz interferer at 1 m			
PDR (%)	100.00	100.00	> 90
Jitter σ (ms)	2.72	1.75	< 50

Requirements from IEEE Std 802.15.6 [8]. Mean $\pm$ SD across repetitions; no packet lost in any acquisition. Latency is a MAC-level round-trip time halved (Section II-C).

interval over several minutes, which is compatible with the duration analysis reported here but would become relevant for applications requiring long-term inter-node alignment. Under the interference condition, the packet delivery ratio remained at 100% and the jitter was no higher than in the uncontaminated repetitions (2.72 ms and 1.75 ms); the tested interferer therefore produced no detectable degradation, which should be read as a limitation of the severity of the test rather than as evidence of immunity to radiofrequency interference in general. Taken together, these measurements bound, under movement, the loss and delay contributed by the wireless link, independently of the segmentation results reported in Section III-D.

B. Developed System

The result is two compact wearable devices with no cabling on the sensor nodes. The system is modular: temporal analysis uses a single node on the lateral shank, while knee goniometry uses both nodes aligned with the femur and tibia. Validation was unilateral, but the architecture scales, and bilateral use would enable gait symmetry analysis. Fig. 3 shows both configurations. Each node weighs about 20 g and runs from a 1200 mAh lithium-polymer cell; at roughly 50 mA in active operation the estimated autonomy is around 24 h. Total material cost was approximately BRL 94 (about USD 17 at 2025 exchange rates).

C. Static Bench Validation of the Measurement Subsystem

The angular accuracy of the reconstruction was assessed against the digital protractor at four target angles spanning the functional flexion range, with five repetitions per angle and a full zero recalibration between repetitions. These figures describe the instrument under stationary bench conditions and are not estimates of in-vivo goniometric accuracy.

A first acquisition showed the prototype reading low at every angle. The absolute error stayed between 2.36° and 2.38°, moving by only 0.02° across a 90° span, and it carried the same sign throughout. An error that constant does not come from the angle or from noise; it comes from a fixed source.



Fig. 3. Positioning of the prototypes on the volunteer's lower limb, illustrating the intended body mounting: the single-sensor configuration used in the temporal protocol (walking trials) and the two-sensor configuration for knee-angle reconstruction. The angular accuracy reported in Table II was obtained on the bench protractor fixture of Protocol 2, not on the limb shown here.

Regression against the reference returned a slope of 0.9999, which ruled out a scale error in the fusion algorithm and left misalignment between the sensitive axes and the mechanical reference as the likely cause. Inspecting the assembly, we found the sensor package visibly tilted with respect to its printed circuit board. The component was resoldered and the acquisition repeated in full.

Table II reports the measurements obtained after remounting. The systematic offset is no longer observed: the individual errors are $+0.03^\circ$, -0.05°, 0.00° and $+0.02^\circ$, alternating in sign, with a mean signed error of 0.00° and a mean absolute error of 0.025°. The residual disagreement is therefore no longer systematic, and is consistent with the finite resolution of the reference and the intrinsic dispersion of the device. The regression against the reference gives a slope of 1.0001 with $R^2 \approx 1.000$.

Two observations point back to the mounting as the origin of the offset. The error went from a constant 2.37° bias to a residual with no preferred direction, which is what removing a trueness error looks like, not what a change in noise would produce. Repeatability, meanwhile, barely moved: the mean standard deviation across the four angles was 0.056° before and after remounting, with worst cases of 0.082° and 0.078°. A tilted mount displaces the reading without adding noise, so correcting it should erase the bias and leave the dispersion alone. That is what happened. The evidence remains observational, though. We ran no controlled comparison isolating the remounting from the rest of the handling, so the attribution is plausible rather than proven.

It should be emphasized that the resulting mean absolute error of 0.025° lies well below the $\pm 0.5^\circ$ resolution of

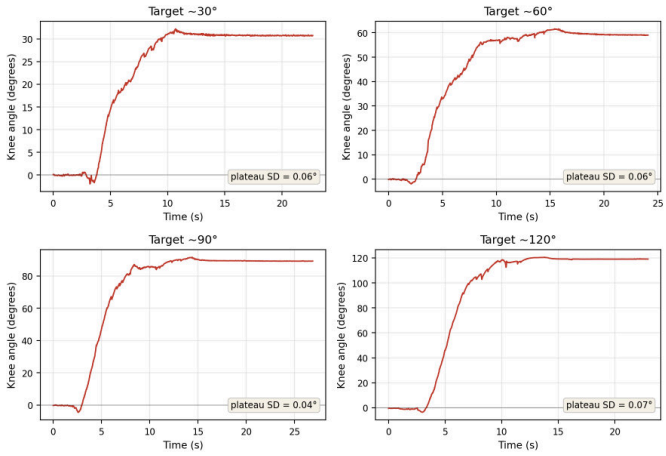


Fig. 4. Representative knee-angle series measured by the prototype during held postures on the bench fixture, showing stable plateaus without drift.

TABLE II
STATIC BENCH VALIDATION AFTER SENSOR REMOUNTING: PROTOTYPE VERSUS DIGITAL PROTRACTOR (FIVE REPETITIONS PER ANGLE)

Reference (°)	Mean (°)	SD (°)	Abs. error (°)	Rel. error (%)
29.98	30.01	0.041	0.03	0.10
60.01	59.96	0.065	0.05	0.08
89.99	89.99	0.078	0.00	0.00
119.97	119.99	0.039	0.02	0.02

Reference: mean digital-protractor reading. SD: within-angle standard deviation over five repetitions. Mean absolute error $0.025°$, below the $\pm 0.5°$ resolution of the reference.

the digital protractor. This figure must therefore be read as evidence that the residual error is no longer resolvable with the present reference, and not as a claim of accuracy at that level; establishing the actual accuracy would require a reference of finer resolution, such as a calibrated and traceable rotary stage, which is identified as future work. The same caveat applies to the reported repeatability, whose dispersion is roughly one order of magnitude finer than the reference resolution. It should also be noted that this repeatability is a measure of internal precision, the dispersion of repeated readings of the device under a nominally fixed angle, and not of agreement with the reference. Both figures remain well within the $2°$–$5°$ inter-examiner variability reported for manual goniometry [24], a comparison developed in Section III-E as context for the intended application. The time series (Fig. 4) remained flat during the held postures, with no observable slope, confirming that the complementary filter sustains static positions without drift over the durations tested.

D. Dynamic Validation of Swing Duration

In the dynamic stage, acquisition covered both lower limbs. Because of the field-of-view restriction of the single-camera 2D videogrammetry, the left and right legs were recorded in independent trials under the same slow, controlled walking protocol. Two-dimensional videogrammetry is itself subject to perspective and out-of-plane errors relative to three-dimensional optoelectronic capture [27], so it is adopted here as an accessible temporal reference rather than as a kinematic gold standard. Fig. 5 shows the raw gyroscope signal, the filtered angular velocity with the automatically detected events, and the raw vertical accelerometer axis, whose impact peaks qualitatively corroborate the detected events.

TABLE III
SWING DURATION: PROTOTYPE VERSUS VIDEOGRAMMETRY

Limb	Sensor mean (s)	Video mean (s)	Abs. error (s)
Left	0.59	0.61	0.02
Right	0.64	0.66	0.02
Overall	0.61	0.63	0.02

Video mean from frame counting between toe off and heel strike. One frame at 60 fps ≈ 0.017 s.

Over two consecutive cycles per limb, mean stance and swing times were 1.08 s and 0.59 s on the left (roughly 65%/35%) and 0.96 s and 0.64 s on the right, the latter close to the 60%/40% ratio reported for normal gait [15]. Cycle durations near 1.6 s are longer than those typical of comfortable-speed walking, reflecting the deliberately slow cadence used to keep the limb within the camera field of view; the values therefore describe this specific condition and should not be compared with normative data acquired at self-selected speed. With a single subject and few cycles, these proportions are illustrative rather than statistically representative.

Table III details the swing-duration comparison against videogrammetry. The quantity compared is the *swing duration*, that is, the interval between the detected toe-off and the subsequent heel strike, and not the absolute instant of either event on a common time base. The mean absolute discrepancy was 0.02 s in both limbs. Considering that the camera operated at 60 frames per second, one frame corresponds to approximately 0.017 s, so the observed discrepancy is of the order of a single frame, that is, at the resolution limit of the reference.

The scope of this result should be stated carefully. Because a constant delay in the acquisition chain affects the detected toe-off and heel strike equally, it cancels in the difference between them; a duration comparison is therefore insensitive to fixed transmission delay and cannot be used to bound it. What the agreement in Table III supports is that the segmentation algorithm recovers the swing interval to within the resolution of the video reference. The delay of the wireless chain is bounded separately and directly by the latency measurement of Section III-A. Nor does the comparison establish that the prototype is more or less precise than the video, since the reference cannot discriminate differences below its own frame interval. Absolute event-timing error, which would require both signals expressed on a common time base through the stomp synchronization, was not quantified and is identified as future work. It should also be noted that the left and right limbs were recorded separately, so bilateral synchrony was not assessed.

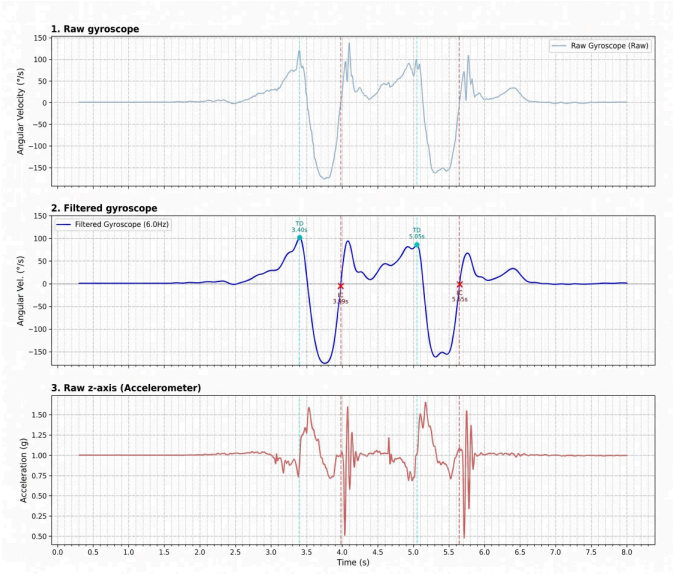


Fig. 5. Gait event detection for one limb: raw gyroscope signal, filtered angular velocity with the automatically detected toe-off and heel-strike events, and raw vertical accelerometer axis. All processing shown was performed offline.

E. Comparison with Previous Studies

The post-remounting bench accuracy and the sub-0.1° repeatability are finer than the 2°–5° inter-examiner variability reported for manual goniometry [24], though both are bounded by the resolution of the reference used here. This says the measurement subsystem is not the limiting factor relative to the manual instrument it might replace. It does not establish clinical equivalence, since the in-vivo error budget should be dominated by soft-tissue artifact rather than by the instrument.

Against other academic prototypes the difference is architectural. Many previous designs wire the lower-limb sensors to a processing unit at the waist, which restricts movement and adds a tripping risk that can alter the very gait pattern under assessment [4], [5], [25]. Independent nodes communicating over ESP-NOW remove that cabling. ESP-NOW is also reported as advantageous over classic Bluetooth, which can introduce variable latency and packet loss in body networks [26]; we adopt that as a design premise and offer no comparative measurement against Bluetooth Low Energy, which remains open.

F. Limitations

The validation was a single-subject case study with a healthy volunteer, a proof of concept focused on the hardware architecture rather than a statistically generalizable evaluation. For the same reason, formal agreement analyses such as Bland–Altman plots, limits of agreement and test–retest reliability were not applied to the angular data: they require a representative sample of paired measurements across subjects, and they remain necessary steps in future multi-subject validation.

The scope of the angular validation is the most consequential limitation. It was performed on a rigid bench fixture, which isolates the measurement subsystem from soft-tissue artifacts but does not reproduce in-vivo fixation; the reported accuracy characterizes the instrument, not the in-vivo joint-angle estimate, where soft-tissue motion should dominate the error budget. Static agreement at held postures also does not validate dynamic knee kinematics. The continuous angle trajectory during walking was never compared against a kinematic reference, and we make no claim about dynamic angular accuracy. Establishing it would require comparison against optical motion capture, or validated commercial inertial units, in multiple participants, reporting bias, limits of agreement, event-timing error and test–retest reliability.

Because the two segment streams are aligned by arrival order rather than by a shared clock, a misalignment of a single sample, that is 10 ms at the 100 Hz rate, is possible. At the segment angular velocities observed during swing in these trials, of the order of 10^2 °/s, such a misalignment would translate into an instantaneous error of a few degrees in the reconstructed knee angle. This reinforces, rather than weakens, the decision not to claim dynamic angular accuracy.

Regarding the wireless link, the latency reported in Section III-A is a MAC-level round-trip time measured on a single clock and halved; it therefore bounds the transport contribution but excludes sensor sampling and host-side processing, and it assumes a symmetric link. The full acquisition-to-display latency was not measured and would require instrumentation at both ends of the chain. The interference test employed a single interferer at close range and produced no detectable degradation, so it does not establish robustness under the denser radiofrequency environments typical of clinical settings; robustness was likewise not evaluated with more than two nodes or at extended range. In addition, ESP-NOW imposes an upper bound on the number of simultaneously registered peers, which constrains the proposed expansion to a full kinematic chain and would require either peer management at run time or a hierarchical topology.

At the sensor level, the ± 2 g range may saturate during heel-strike transients, and the noise floor of the MPU-6050 was not characterized. Without a magnetometer, absolute three-dimensional orientation is unrecoverable, which motivated the sagittal-plane formulation: axial rotations and abductions are not discriminated and may introduce error, and the sagittal axis must be reconfirmed whenever the mounting changes. The system also does not quantify stride length, walking velocity or three-dimensional joint trajectories, whose inertial estimation requires double integration and would demand robust fusion such as Kalman filtering [17], [28].

IV. Conclusion

This work developed and bench-validated a low-cost wireless wearable prototype for gait measurement, removing the cabling that constrains movement in comparable academic prototypes. Pairing the ESP32 with the MPU-6050 yielded small, easily attachable devices, and the ESP-NOW link was stable throughout the acquisitions. Moving the angular recon-

struction on-board the receiver delivered the knee angle online, sustaining static postures without drift.

Three quantities were characterized here, and they answer different questions. Under stationary bench conditions, an initial systematic error of 2.37° was traced to a tilted sensor package and eliminated by remounting the component, after which the mean absolute error fell to 0.025° with unchanged repeatability; both figures lie below the resolution of the reference instrument and therefore bound the residual error rather than quantify it. During walking, the swing duration recovered from the detected events agreed with videogrammetry to within 0.02 s, of the order of one video frame; absolute event timing was not compared on a common time base. The wireless link was characterized under the same movement condition, with no packet loss over 211,007 analyzed packets and an inter-arrival standard deviation below 4 ms, substantiating by measurement the low delay-variation premise of the architecture. Link-level latency had a median one-way value of about 1 ms with a 99th percentile below 13 ms and no transmission failure over 2000 frames. Dynamic knee kinematics, however, were not validated against a kinematic reference, and no result reported here should be read as evidence of clinical gait accuracy.

The contribution reported here is therefore the wireless, cable-free node architecture together with its metrological characterization on the bench. Future work includes physical optimization using surface-mount components, instrumentation of the full acquisition-to-display latency, evaluation under denser radiofrequency environments and with additional nodes, expansion of the topology for full kinematic-chain reconstruction and synchronous bilateral validation [7], characterization of the residual angular error against a traceable rotary stage of finer resolution, comparison against optical motion capture in multiple participants, evaluation in pathological gait such as Parkinson or post-stroke populations, and the incorporation of magnetometers for nine-axis sensor fusion together with Kalman filtering and zero-velocity update (ZUPT) techniques to enable estimation of velocity and distance [4], [28].

Acknowledgment

The authors thank the Faculty of Electrical and Biomedical Engineering of the Federal University of Pará for the support provided during the development of this work.

Declaration on the Use of Artificial Intelligence

For the purposes of academic integrity, the authors declare that artificial intelligence (AI) tools were used in the preparation of this work. Their use was restricted to instrumental support, specifically the formatting of Latex code, the standardization of bibliographic references, the search for some of the cited articles, and grammatical review.

References

[1] S. Fritz and M. Lusardi, "Walking speed: the sixth vital sign," *J. Geriatr. Phys. Ther.*, vol. 32, no. 2, pp. 2–5, 2009.

[2] J. Perry and J. M. Burnfield, *Gait Analysis: Normal and Pathological Function*, 2nd ed. Thorofare, NJ, USA: SLACK, 2010.

[3] D. Levine, J. Richards, and M. W. Whittle, *Whittle's Gait Analysis*, 5th ed. Edinburgh, U.K.: Elsevier, 2014.

[4] W. Tao, T. Liu, R. Zheng, and H. Feng, "Gait analysis using wearable sensors," *Sensors*, vol. 12, no. 2, pp. 2255–2283, 2012.

[5] A. Muro-de-la-Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

[6] Espressif Systems, "ESP32 series datasheet," Rev. 4.0, 2023.

[7] D. Urazayev *et al.*, "Indoor performance evaluation of ESP-NOW," in *Proc. IEEE Smart Inf. Syst. Technol. (SIST)*, 2023, pp. 1–6.

[8] *IEEE Standard for Local and Metropolitan Area Networks — Part 15.6: Wireless Body Area Networks*, IEEE Std 802.15.6-2012, Feb. 2012.

[9] B. Becker, C. Oberli, J. Zobel, R. Steinmetz, and T. Meuser, "ESP-NOW performance in outdoor environments: field experiments and analysis," in *Proc. 20th IEEE/IFIP Wireless On-Demand Netw. Syst. Serv. Conf. (WONS)*, 2025, pp. 41–48.

[10] D. Eridani, A. F. Rochim, and F. N. Cesara, "Comparative performance study of ESP-NOW, Wi-Fi, Bluetooth protocols based on range, transmission speed, latency, energy usage and barrier resistance," in *Proc. Int. Semin. Appl. Technol. Inf. Commun. (iSemantic)*, 2021, pp. 322–328.

[11] H. W. Ott, *Electromagnetic Compatibility Engineering*. Hoboken, NJ, USA: Wiley, 2009.

[12] M. Waqas, S. Ali, and M. Khan, "Fabrication of printed circuit board (PCB) using laser engraving method," in *Proc. Int. Conf. Comput. Math. Eng. Technol. (iCoMET)*, 2018, pp. 1–5.

[13] A. Cappozzo *et al.*, "Position and orientation in space of bones during movement: anatomical frame definition and determination," *Clin. Biomech.*, vol. 20, no. 1, pp. 1–15, 2005.

[14] InvenSense, "MPU-6000 and MPU-6050 product specification," Rev. 3.4, Sunnyvale, CA, USA, 2013.

[15] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 4th ed. Hoboken, NJ, USA: Wiley, 2009.

[16] R. Mahony, T. Hamel, and J.-M. Pflimlin, "Nonlinear complementary filters on the special orthogonal group," *IEEE Trans. Autom. Control*, vol. 53, no. 5, pp. 1203–1218, 2008.

[17] A. M. Sabatini, "Quaternion-based strap-down integration method for applications of inertial sensing to gait analysis," *Med. Biol. Eng. Comput.*, vol. 43, no. 1, pp. 94–101, 2005.

[18] G. Cooper *et al.*, "Inertial sensor-based knee flexion/extension angle estimation," *J. Biomech.*, vol. 42, no. 16, pp. 2678–2685, 2009.

[19] B. R. Greene *et al.*, "An adaptive gyroscope-based algorithm for temporal gait analysis," *Med. Biol. Eng. Comput.*, vol. 48, no. 12, pp. 1251–1260, 2010.

[20] S. Pardoel *et al.*, "Grouping successive freezing of gait episodes has neutral to detrimental effect on freeze detection and prediction in Parkinson's disease," *PLoS ONE*, vol. 16, no. 10, p. e0258544, 2021.

[21] A. Puig-Diví *et al.*, "Validity and reliability of the Kinovea program in obtaining angles and distances using coordinates in 4 perspectives," *PLoS ONE*, vol. 14, no. 6, p. e0216448, 2019.

[22] C. D. Gómez-Carmona *et al.*, "Reliability and validity of Kinovea to analyze spatiotemporal gait parameters," *Front. Psychol.*, vol. 13, 2022.

[23] D. G. E. Robertson *et al.*, *Research Methods in Biomechanics*, 2nd ed. Champaign, IL, USA: Human Kinetics, 2013.

[24] C. C. Norkin and D. J. White, *Measurement of Joint Motion: A Guide to Goniometry*, 5th ed. Philadelphia, PA, USA: F. A. Davis, 2016.

[25] G. Agune *et al.*, "Solução de baixo custo para a análise de marcha utilizando OpenSense" [Low-cost solution for gait analysis using OpenSense], in *Anais do VIII Encontro Nacional de Engenharia Biomecânica (ENEBI 2024)*, Gramado, Brazil, 2024.

[26] M. Olsen, A. V. Pavlova, and E. Papi, "Reliability of Bluetooth inertial sensors for assessing lower limb segment angles and stride length during gait," *J. Mech. Med. Biol.*, vol. 22, no. 6, p. 2250043, 2022.

[27] S. A. Schurr *et al.*, "Comparison of 2D video analysis and 3D motion capture in lower extremity movement assessment," *Int. J. Exerc. Sci.*, vol. 10, no. 5, p. 763, 2017.

[28] O. J. Woodman, "An introduction to inertial navigation," Univ. Cambridge, Cambridge, U.K., Tech. Rep. UCAM-CL-TR-696, 2007.

Development of an Electrical Impedance Tomography System Using an ESP32-Based Data Acquisition Architecture

Laís Moreira de Castro
Department of Graduate Studies - DIPPG
Federal Institute of Ceará (IFCE)
Fortaleza, Brazil
https://orcid.org/0009-0003-4505-2007

Mauri Saraiva dos Santos
Department of Teleinformatics Engineering
Federal University of Ceará (UFC)
Fortaleza, Brazil
https://orcid.org/0000-0002-3987-8000

Luis Gustavo de Sousa
Electrical Engineering Department
Federal University of Ceará (UFC)
Fortaleza, Brazil
https://orcid.org/0009-0003-6881-4000

Antonio Jefferson Mangueira Sales
Department of Physics - LOCEM
Federal University of Ceará (UFC)
Fortaleza, Brazil
https://orcid.org/0000-0002-1970-0547

Daniel Xavier Gouveia
Department of Graduate Studies - DIPPG
Federal Institute of Ceará (IFCE)
Fortaleza, Brazil
https://orcid.org/0000-0003-0305-316X

Antonio Sergio Bezerra Sombra
Department of Physics - LOCEM
Federal University of Ceará (UFC)
Fortaleza, Brazil
https://orcid.org/0000-0001-6461-6793

***Abstract*—Several studies on Electrical Impedance Tomography (EIT) and impedance image reconstruction using the EIDORS framework have been reported in the literature. However, the implementation of complete EIT systems remains challenging due to the complexity of the analog front-end, current injection circuitry, signal acquisition, and image reconstruction. This work presents a noninvasive 16-electrode EIT system for reconstructing electrical impedance images of cylindrical phantoms using open-source software. The proposed architecture comprises a modified Howland current source configured for a nominal 1 mA excitation current at 47.3 kHz, analog signal conditioning, electrode multiplexing, an ESP32-based data acquisition unit, and image reconstruction using the EIDORS framework in GNU Octave. A 16-electrode adjacent stimulation protocol was adopted, providing 208 measurements per acquisition cycle. Experimental results obtained with insulating objects immersed in a conductive phantom demonstrate that the proposed system is capable of detecting conductivity perturbations and qualitatively identifying changes in their relative positions, demonstrating the functional operation of the acquisition and image reconstruction chain under the tested phantom conditions.**

***Index Terms*—Electrical Impedance Tomography, ESP32, biomedical instrumentation, Howland current source, image reconstruction, data acquisition.**

I. Introduction

Electrical Impedance Tomography (EIT) is a non-invasive and radiation-free imaging technique based on the application of low-amplitude alternating electrical currents through electrodes positioned on the boundary of an object and the measurement of the resulting boundary voltages [1], [2]. From these measurements, the internal conductivity distribution, or changes in this distribution, can be estimated through an image reconstruction algorithm. This technique has been applied in areas such as quality control, fault detection, geological exploration, bioimaging, and disease monitoring, particularly of the lungs [3], [4].

Compared with other medical imaging techniques, such as X-ray, computed tomography (CT), and positron emission tomography (PET), EIT offers advantages including the absence of ionizing radiation, portability, and the potential for high temporal resolution. However, EIT has relatively low spatial resolution and an ill-posed inverse problem, making accurate reconstruction challenging and more suitable for repeated functional monitoring than for high-resolution anatomical diagnosis [10].

Several EIT systems have been reported in recent years using embedded platforms, dedicated analog front-ends, and open-source reconstruction software [4]–[7], [12]–[17], [23], [24]. These studies have contributed to the development of compact and reproducible experimental platforms for instrumentation research and algorithm evaluation.

This study proposes an Electrical Impedance Tomography (EIT) system for impedance image reconstruction using 16 electrodes and commercially available electronic components. The proposed architecture comprises a modified Howland current source, analog signal conditioning, electrode multiplexing, an ESP32-based data acquisition unit, and image reconstruction using the open-source EIDORS framework in GNU Octave. Experimental evaluation was performed using a cylindrical phantom filled with tap water, with insulating objects introduced at different positions to assess the ability of the system to detect conductivity perturbations and qualitatively reproduce changes in their relative positions.

II. Theoretical Background

Under the quasi-static approximation, the electric potential in Electrical Impedance Tomography (EIT) is governed by the conductivity equation [2], [12]:

$$\nabla \cdot (\sigma \nabla U) = 0 \tag{1}$$

where σ is the electrical conductivity distribution and U is the electric potential.

The forward problem in EIT can be represented using the Complete Electrode Model (CEM), which accounts for electrode dimensions and contact impedance and provides a realistic representation of practical electrode measurements [20]. Within the conductive domain Ω, and for each electrode e_l, the corresponding equations include

$$\nabla \cdot (\sigma \nabla U) = 0, \qquad \text{in } \Omega, \tag{2}$$

$$U + z_l \sigma \frac{\partial U}{\partial \eta} = U_l, \qquad \text{on } e_l, \tag{3}$$

where U_l is the electrode potential, z_l is the contact impedance, and η denotes the outward normal direction.

The forward problem computes boundary voltages from a known conductivity distribution, whereas the inverse problem estimates the internal conductivity distribution, or changes in this distribution, from boundary measurements. Since the inverse problem is ill-posed, regularization techniques are generally required to obtain stable solutions [9], [11], [21].

In this work, conductivity changes were reconstructed using difference EIT, in which measurements acquired from a homogeneous phantom were used as the reference dataset. Image reconstruction was performed using the EIDORS framework, as described in Section III-D.

III. METHODOLOGY

A. System Overview

The developed EIT system integrates hardware and software modules for data acquisition and image reconstruction. A constant-current excitation strategy was adopted using a modified Howland current source, a topology commonly employed in EIT and bioimpedance instrumentation [8]. The current source was configured for a nominal excitation current of 1 mA at 47.3 kHz during the experiments.

Fig. 1 illustrates the functional architecture of the proposed EIT system. A sinusoidal excitation signal at 47.3 kHz is applied to the modified Howland current source. The excitation current is routed to the phantom through an analog switching network composed of four analog multiplexers, which select the current injection (I_{inj}), current return (I_{ret}), and voltage measurement (V_{meas}) electrode pairs.

The measured differential voltage is amplified by an instrumentation amplifier, filtered by an active band-pass filter, and processed by a logarithmic magnitude detector. A reference signal derived from the excitation oscillator is applied to the detector reference input, allowing the magnitude output (V_{MAG}) to represent a logarithmic amplitude ratio between the measured and reference signals. Although the detector also provides a phase output, only V_{MAG} was acquired and used in the present implementation. Finally, an ESP32-based data acquisition unit digitizes the detector output, controls the multiplexers through GPIO signals, and transmits the acquired data to a host computer, where conductivity-change images are reconstructed using the EIDORS toolbox in the GNU Octave environment.

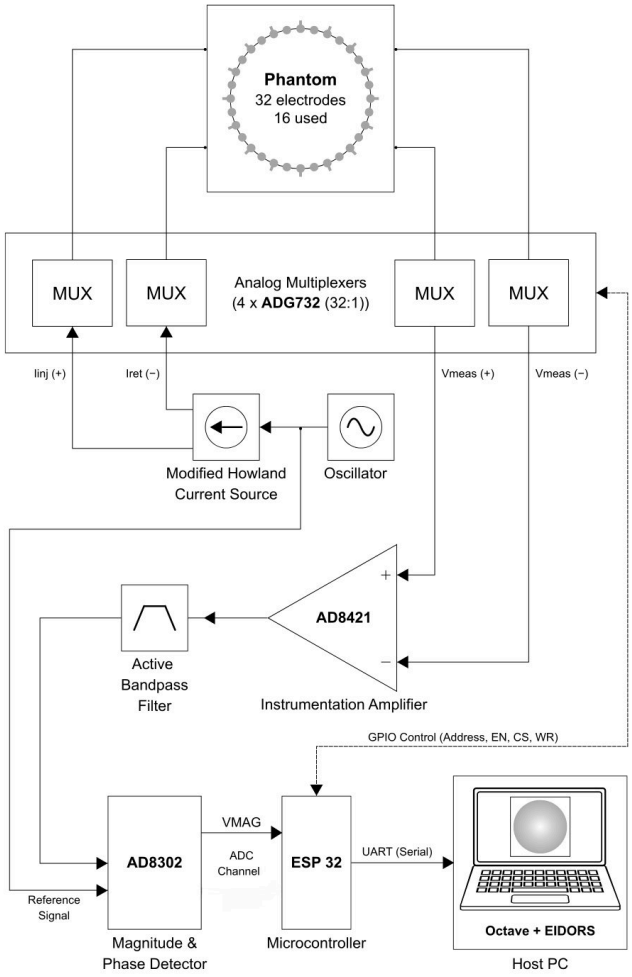


Fig. 1. Block diagram of the developed EIT system.

B. Hardware Design

The analog front-end of the developed EIT system was organized as a modular architecture comprising the excitation signal generator, a modified Howland current source, a voltage acquisition and conditioning circuit, a magnitude detection stage, an analog multiplexing network, and a digital control interface. Figure 1 illustrates the functional architecture of the system, while the following subsections describe the implementation of each hardware module.

The circuit was powered from external $+12$ V and -12 V supply rails. A regulated $+5$ V rail was locally generated from the $+12$ V supply using a linear voltage regulator. This voltage supplies the AD8302 detector, whose specified supply

range is 2.7–5.5 V [23]. The other analog stages operate from the ± 12 V supply.

The excitation signal generator, shown in Fig. 2, produces the sinusoidal reference signal used throughout the EIT system. This signal drives the modified Howland current source and simultaneously provides the reference input for the magnitude detection stage.

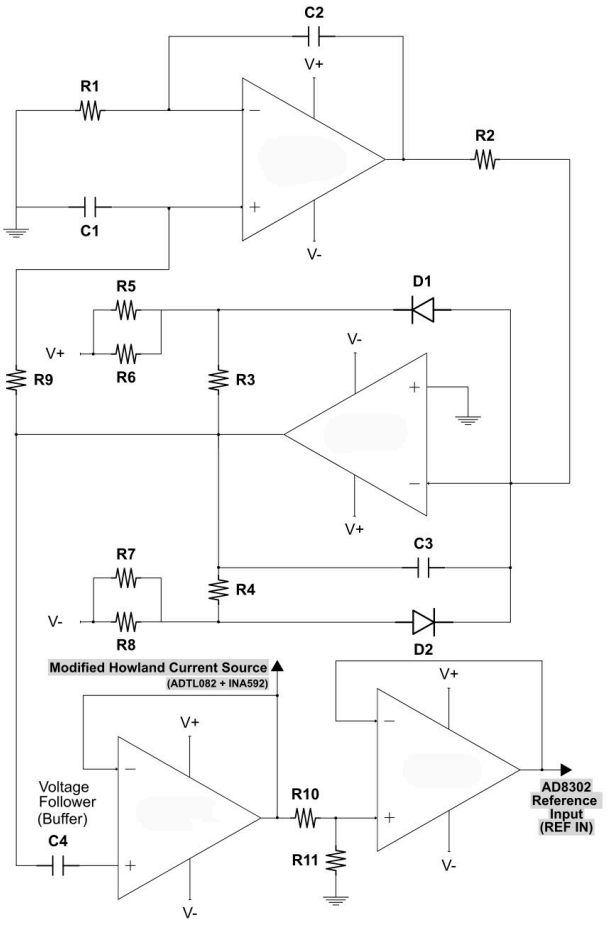


Fig. 2. Simplified schematic of the excitation signal generator based on the ADA4001-2 operational amplifier.

The generator is based on a Wien bridge oscillator, a well-established topology for producing sinusoidal waveforms using a small number of analog components [24], [25]. The oscillator was adjusted to operate at 47.3 kHz, which was adopted as the excitation frequency throughout the experimental measurements.

To stabilize the output amplitude, the oscillator employs a nonlinear amplitude-control network based on diode limiting. As the oscillation amplitude increases, the diodes modify the loop gain, limiting further amplitude growth [25].

The oscillator output is buffered by a voltage follower implemented with an ADA4001-2 operational amplifier. This stage isolates the oscillator from the following analog circuitry and reduces loading of the oscillator output [29].

A second buffered output is routed to the reference input of the AD8302 magnitude and phase detector. Thus, the detector receives a reference signal derived from the same oscillator used to drive the current-injection stage, allowing estimation of the measured signal magnitude [23].

The current injection stage, illustrated in Fig. 3, is based on a modified Howland current source. This topology is commonly employed in bioimpedance and EIT instrumentation, where a high output impedance is desirable to reduce variations in the excitation current caused by changes in load impedance [22], [34].

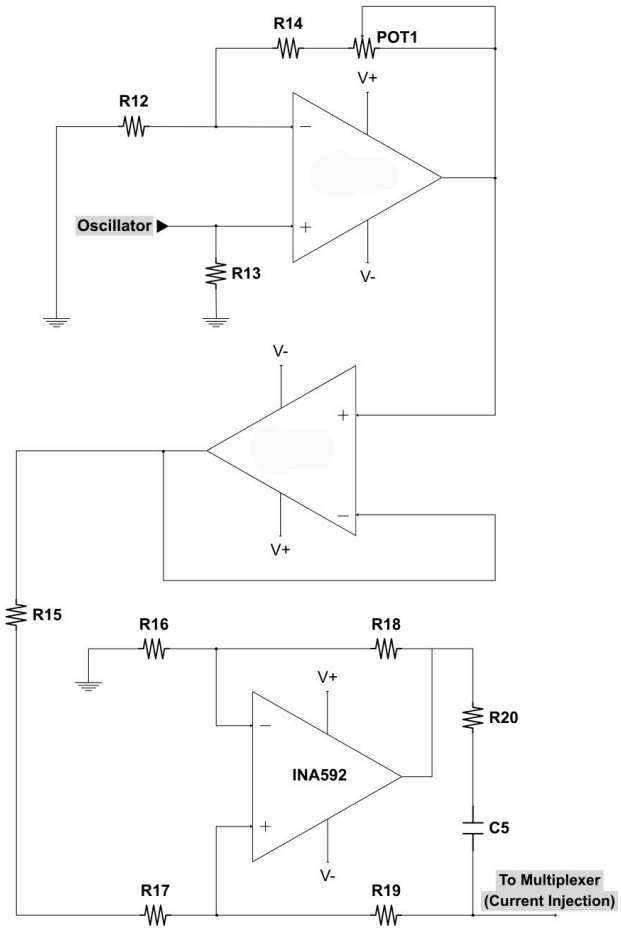


Fig. 3. Simplified schematic of the current injection stage.

The excitation signal generated by the oscillator is first applied to a non-inverting amplifier implemented with an ADA4001-2 operational amplifier. The feedback network, composed of a fixed resistor and the trimming potentiometer POT1, allows adjustment of the excitation amplitude and, consequently, of the nominal injected current.

The amplified signal is subsequently buffered by a second ADA4001-2 operational amplifier configured as a unity-gain voltage follower. This stage isolates the adjustable-gain stage from the following current-source circuitry, reducing loading of the preceding stage.

The buffered excitation voltage drives the modified Howland current source implemented with an INA592 difference amplifier. The resistor network establishes the voltage-to-current conversion and the resistor-ratio matching required by the Howland topology. Ideally, matching the corresponding resistor ratios increases the output impedance of the current source and reduces the dependence of the output current on load variations [22], [34].

The bridge balance condition is expressed as

$$\frac{R_2}{R_1} = \frac{R_4}{R_3}, \tag{4}$$

where matching the corresponding resistor ratios is important for obtaining a high effective output impedance in the Howland topology [34].

The INA592 was selected as the difference-amplifier element of the current-source circuit. According to the manufacturer, the device incorporates a precision matched internal resistor network and provides wide bandwidth and high common-mode rejection [28]. These characteristics are suitable for the implementation of the resistor-ratio relationships required by the adopted circuit topology.

The current source was configured for a nominal output current of 1 mA at 47.3 kHz. However, load-dependent current variation, output impedance, compliance voltage, and harmonic distortion were not independently characterized in the experiments reported here. Therefore, the 1 mA value is reported as the nominal design value rather than as an experimentally validated current amplitude.

The voltage acquisition stage receives the differential voltage selected by the analog multiplexers and conditions it before magnitude detection. As illustrated in Fig. 4, the circuit comprises an input filtering network, an AD8421 instrumentation amplifier, an active filtering stage, and a voltage buffer preceding the AD8302 logarithmic magnitude detector.

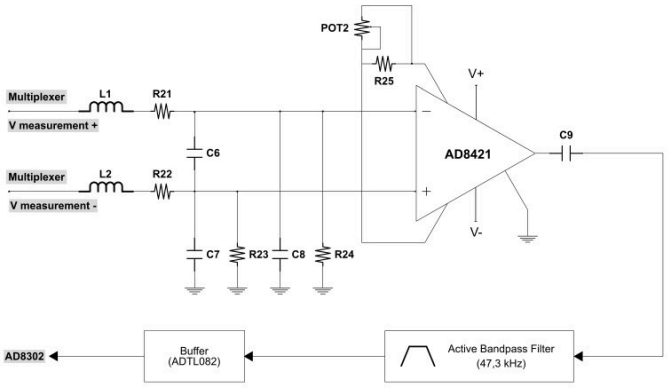


Fig. 4. Voltage acquisition and conditioning stage preceding the AD8302 magnitude detector.

The differential voltage measured across the selected electrode pair is routed from the analog multiplexers to the conditioning circuit. Before amplification, the input signals pass through a passive network composed of ferrite beads, resistors, and capacitors, which provides frequency-dependent attenuation of unwanted signal components before differential amplification.

Signal amplification is performed by an AD8421 instrumentation amplifier, selected for its high common-mode rejection ratio (CMRR), low input voltage noise, and high input impedance [26]. The amplifier provides differential amplification while rejecting common-mode components present at its inputs. Its gain is adjusted through potentiometer POT2 according to the required signal level for the subsequent conditioning stages.

The amplifier output is AC-coupled to an active filtering stage designed to condition the signal around the operating frequency of 47.3 kHz. The resulting signal is then buffered by an ADA4001-2 operational amplifier configured as a voltage follower before being applied to the measurement input of the AD8302 [29]. This buffer reduces loading of the preceding filtering stage.

Logarithmic magnitude detection is performed using the AD8302 gain and phase detector. Although the device provides both magnitude and phase outputs, only the magnitude output was acquired in the present implementation. As illustrated in Fig. 5, the conditioned measurement signal is applied to one detector input, while an amplitude-adjusted reference signal derived from the excitation oscillator is applied to the other.

The reference signal is conditioned by an ADA4001-2 operational amplifier and AC-coupled to the AD8302. Potentiometer POT3 allows adjustment of the reference-signal amplitude, while the conditioned measurement signal is also AC-coupled to the detector. According to the AD8302 transfer characteristic, the magnitude output is a linear function of the logarithmic ratio between the amplitudes of its two input signals [23].

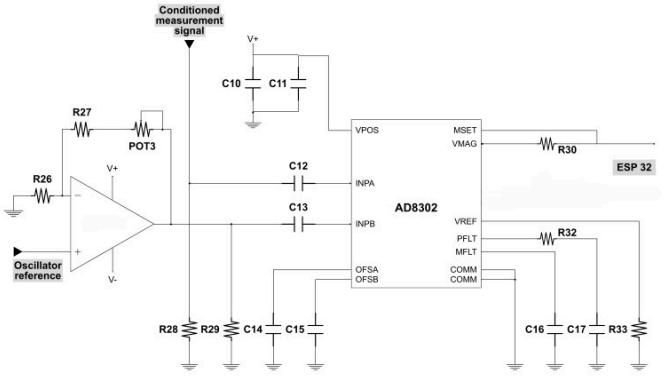


Fig. 5. Simplified schematic of the AD8302-based magnitude detection stage.

The magnitude output is subsequently filtered and digitized by the internal analog-to-digital converter (ADC) of the ESP32 after each electrode measurement. The acquired values are then transmitted to the host computer through the UART interface for subsequent processing and image reconstruction.

Electrode selection and signal routing are performed using four ADG732 analog multiplexers, as shown in Fig. 6. Two multiplexers are used to select the electrode pair for current injection, while the other two select the electrode pair from which the differential voltage is acquired. This arrangement allows the current-injection and voltage-measurement paths to be independently selected under digital control.

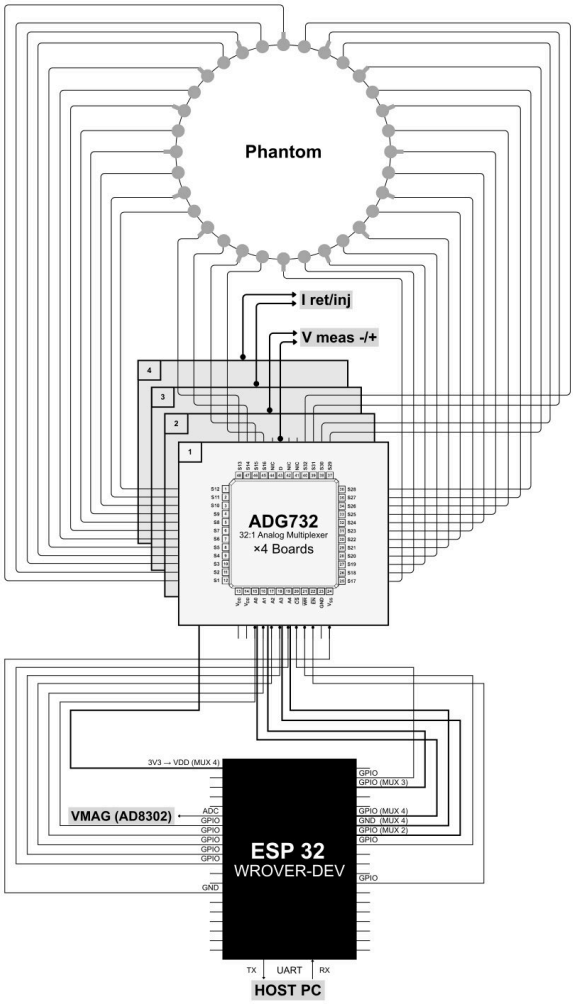


Fig. 6. Simplified architecture of the analog switching network based on four ADG732 multiplexers controlled by the ESP32-WROVER-DEV microcontroller for electrode selection and data acquisition.

The multiplexers are controlled by the ESP32 through five shared address lines (A0–A4), which select one of the 32 available channels. Individual chip-select signals determine which multiplexer is addressed, while the write and enable signals control the switching operation. The ESP32 coordinates the multiplexing sequence, ADC acquisition, and communication with the host computer according to the measurement protocol described in Section III-C.

Figure 7 shows the assembled prototype used in the experimental evaluation. The analog front-end implemented on a custom PCB, the multiplexing circuitry mounted on a dedicated breakout board, the ESP32-based control unit, and the electrode connections were integrated into the prototype employed for the phantom measurements.

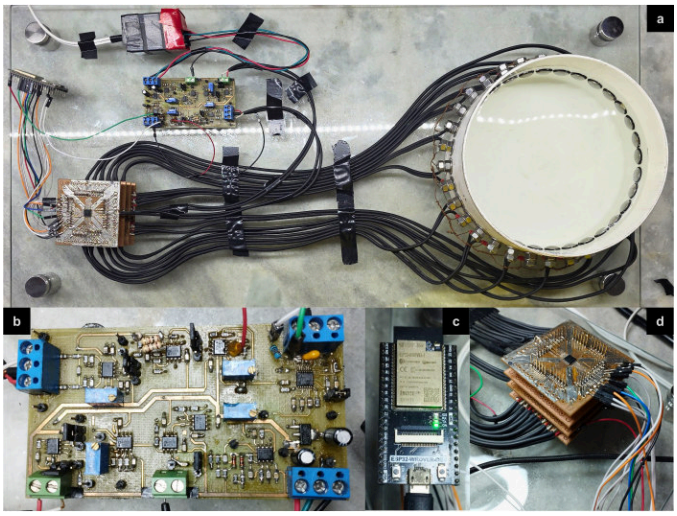


Fig. 7. Photographs of the developed EIT prototype: (a) complete experimental setup; (b) analog front-end PCB; (c) ESP32-WROVER-DEV development board; and (d) ADG732 multiplexer breakout board.

C. Image Acquisition Procedure

After hardware initialization, the ESP32 executes the data acquisition sequence by controlling the four analog multiplexers. The multiplexers share a common five-bit address bus, while the control signals configure the current-injection and voltage-measurement channels according to the acquisition protocol.

For each electrode configuration, the ESP32 sequentially selects the current-injection and differential voltage-measurement pairs. After switching, a programmable settling interval is applied before sampling the VMAG output of the logarithmic magnitude detector with the internal 12-bit ADC. Each measurement is obtained by averaging 1000 ADC conversions, and only the resulting average is retained for subsequent processing. A firmware scale factor of 1.344 is then applied to the averaged ADC output.

Although averaging was employed during acquisition, noise level, signal-to-noise ratio (SNR), and overall measurement uncertainty were not independently characterized in the present experimental evaluation.

The adjacent-adjacent protocol was implemented using 16 electrodes selected alternately from the 32 equally spaced physical electrodes. For each of the 16 adjacent current-injection pairs, voltage measurements were acquired from 13 adjacent electrode pairs that did not include the current-carrying electrodes, resulting in 208 acquired measurements per frame. Considering reciprocity, 104 of these measurements are independent [2], [18].

Considering the programmed delays associated with multiplexer switching and the 20-ms settling interval preceding each

measurement, a complete 208-measurement frame includes approximately 8.35 s of programmed delay. The execution time required for ADC conversions, processing, and serial communication is not included in this value; therefore, the actual acquisition time per frame is greater than 8.35 s.

The acquired values are transmitted to the host computer through the UART interface at 115200 baud. On the host computer, each received magnitude-related value V_i is numerically transformed according to the processing implemented in the acquisition script. Specifically, $q_i = (V_i - 0.9)/0.03$ is first calculated, followed by $d_i = 10^{q_i/10}$. The resulting 208 values are assembled into the measurement vector supplied to EIDORS. A new homogeneous reference dataset was acquired before each experimental configuration. The corresponding reference and perturbed measurement vectors were then used for difference image reconstruction.

D. Image Reconstruction Procedure

Image reconstruction was performed in GNU Octave using the EIDORS toolbox [30]. A two-dimensional circular finite-element inverse model with 16 electrodes was created using the `c2c2` model provided by `mk_common_model`. The adjacent stimulation and measurement patterns were generated using `mk_stim_patterns`, according to the acquisition protocol described in Section III-C.

Difference imaging was performed by comparing the measurement vector acquired from the homogeneous phantom with that obtained after introducing an inclusion. The reference and perturbed measurement vectors were supplied to the EIDORS `inv_solve` function to estimate the corresponding conductivity change.

Because only the magnitude output of the AD8302 was acquired, phase information was not included in the reconstruction. Consequently, the reconstructed images were interpreted as qualitative relative conductivity changes with respect to the homogeneous reference state rather than as calibrated absolute conductivity distributions.

E. Experimental Evaluation

The experimental evaluation was performed using a cylindrical phantom constructed from a PVC pipe. The dimensions of the phantom, electrodes, and test objects reported in this section were measured using a caliper. The phantom has an internal diameter of approximately 183 mm, a total height of 100.22 mm, and a wall thickness of 3.15 mm. It was filled with tap water to a height of approximately 60 mm. The conductivity and temperature of the water were not instrumentally recorded during the experiments. No predefined stabilization interval was adopted before data acquisition.

The electrode array consists of 32 equally spaced stainless-steel carriage bolts, approximately 5.6 mm in diameter and 15 mm in length, mounted around the phantom circumference and fully submerged during the measurements. Electrical connection to each electrode is made using a ring terminal fastened to the external threaded portion of the bolt.

Conductivity perturbations were introduced using a glass cup and a rectangular plastic bar. The glass cup had a height of 92 mm and diameters of approximately 47 mm and 66.6 mm at its narrower and wider sections, respectively, and was partially filled with water during immersion. The rectangular plastic bar measured approximately $135.9 \times 31.10 \times 31.20$ mm.

Two experimental configurations were considered. In the first, the glass cup and plastic bar were simultaneously positioned inside the phantom to evaluate whether two conductivity perturbations could be distinguished in the reconstructed image. In the second, the glass cup was placed at four different positions within the phantom to qualitatively evaluate whether changes in object position were reflected in the reconstructed images.

IV. RESULTS AND DISCUSSION

Figures 8 and 9 present the experimental results obtained with the proposed EIT system. The experiments evaluated whether conductivity perturbations produced by the test objects could be detected and whether changes in their relative positions were reflected in the difference images.

Figure 8 shows the experiment in which the glass cup and the plastic bar were simultaneously positioned inside the phantom. Two spatially distinct perturbation regions are visible in the reconstructed image, corresponding qualitatively to the presence of the two test objects. Although the reconstruction does not reproduce their exact geometry, the result indicates that the measurement and reconstruction chain can distinguish the two perturbations under the tested configuration.

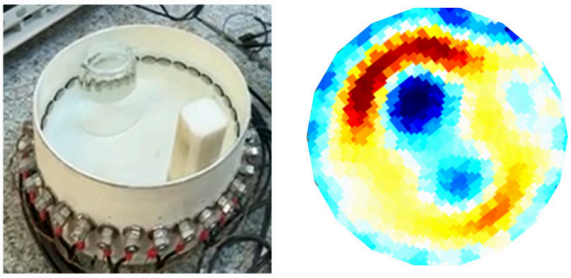


Fig. 8. Experimental setup (left) and corresponding reconstructed difference image (right) for the glass cup and plastic bar simultaneously positioned inside the phantom. Colors represent qualitative relative changes in the reconstructed image; no calibrated conductivity scale is assigned.

Figure 9 presents the reconstructed images obtained with the glass cup placed at four different positions within the phantom. As the cup position changes, the main reconstructed perturbation region changes accordingly, qualitatively reproducing the relative displacement of the object within the imaging domain.

The reconstructed images exhibit spatial spreading and boundary artifacts, which are consistent with the limited spatial resolution and ill-posed nature of the EIT inverse problem [18]. The experimental evaluation was qualitative. The exact coordinates of the test objects and repeated measurement frames for each configuration were not recorded during the

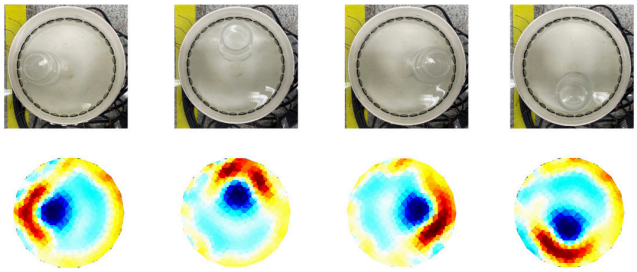


Fig. 9. Glass cup at four different positions within the phantom (top) and the corresponding reconstructed difference images (bottom). Colors represent qualitative relative changes in the reconstructed images; no calibrated conductivity scale is assigned.

original experiments; therefore, quantitative localization error and inter-frame variability could not be retrospectively determined. Accordingly, the results are interpreted as a functional demonstration of perturbation detection and relative position changes rather than as a quantitative validation of localization accuracy or measurement repeatability.

Furthermore, the current acquisition time limits the present implementation for applications requiring high temporal resolution. Future improvements should therefore investigate optimization of the multiplexer switching and settling intervals, ADC averaging, and data-transfer procedure to increase the acquisition rate without compromising measurement quality.

V. CONCLUSION

This work presented the development and experimental evaluation of an Electrical Impedance Tomography (EIT) system based on an ESP32 microcontroller. The proposed architecture integrates a modified Howland current source, analog multiplexing, signal conditioning, logarithmic magnitude detection, and image reconstruction using the open-source EIDORS framework.

Experimental results obtained with a cylindrical phantom showed that the system can detect conductivity perturbations produced by the test objects and qualitatively reflect changes in their relative positions within the imaging domain. These results demonstrate the functional operation of the complete acquisition and reconstruction chain under the tested phantom conditions. The present evaluation remains qualitative, and the acquisition rate and absence of phase information limit the scope of the current implementation.

Future work will focus on quantitative characterization of the current source, repeatability and noise assessment, acquisition-time optimization, and evaluation using phantoms with controlled conductivity properties and tissue-like materials.

Demonstration videos showing the operation of the complete EIT system are available in [32], [33].

ACKNOWLEDGMENT

The authors gratefully acknowledge the support provided by the Brazilian National Council for Scientific and Technological Development (CNPq), IFCE and UFC.

REFERENCES

[1] D. Barber, B. Brown, "Recent developments in applied potential tomography-APT," Inf. Process. Med. Imag., pp. 106-121, 1986.

[2] J. Malmivuo, R. Plonsey, *Principles and Applications of Bioelectric and Biomagnetic Fields*. New York: Oxford University Press, 1995.

[3] R. Bayford, A. Tizzard, "Bioimpedance Imaging: An Overview of Potential Clinical Applications," *Analyst*, vol. 137, pp. 4635-4643, 2012.

[4] T. K. Bera et al., "A Low Cost EIT Instrumentation for Impedance Imaging of Practical Phantoms: A Laboratory Study," in *Soft Computing for Problem Solving*, 2014.

[5] B. Sun et al., "Development of a Miniaturized, Low-cost Electrical Impedance Tomography System," *IEEE Trans. Instrum. Meas.*, 2024.

[6] M. Mohamedali et al., "Design of A Simple and Low Cost Electrical Impedance Tomography System," 2019.

[7] Y. Wu et al., "A 122 fps, 1 MHz Bandwidth Multi-Frequency Wearable EIT Belt for Neonatal Thorax Vital Sign Monitoring," *IEEE Trans. Biomed. Circuits Syst.*, 2019.

[8] M. Rafiei-Naeini, H. McCann, "Low-Noise Current Excitation Sub-System for Medical EIT," *Physiol. Meas.*, vol. 29, p. S173, 2008.

[9] M. Winkler et al., "Model-aware Newton-type Inversion Scheme for Electrical Impedance Tomography," *Inverse Problems*, vol. 31, no. 4, 2015.

[10] M. Takhti, K. Odame, "Structured Design Methodology to Achieve a High SNR Electrical Impedance Tomography," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, pp. 364-375, 2019.

[11] A. Lechleiter et al., "Newton Regularizations for Impedance Tomography: A Numerical Study," *Inverse Problems*, vol. 22, no. 6, 2006.

[12] B. A. Pitaloka et al., "Development of Low Cost EIT Equipment for Educational Purposes," *J. Phys.: Conf. Ser.*, 2019.

[13] M. J. Kraśny, R. Creaven, A. Venugopal, E. Dunne, and N. Colgan, "The Entry-Level Electrical Impedance Tomography (EIT) Device You Can Assemble without Soldering," in *Proc. 23rd Int. Conf. Biomedical Applications of Electrical Impedance Tomography (EIT2023)*, Aachen, Germany, Jun. 2023.

[14] L. Capacho-Valbuena et al., "Prototype of an Embedded Electrical Impedance Tomography System for Coffee Bean Evaluation," 2025.

[15] P. Creegan et al., "A Wearable Open-Source Electrical Impedance Tomography Device," *HardwareX*, 2024.

[16] V. Chitturi et al., "A Low Cost Electrical Impedance Tomography (EIT) for Pulmonary Disease Modelling and Diagnosis," *ICTACE*, 2014.

[17] T. H. Dinh et al., "Application and Investigation on Reconstructed Cross-Sectional Images by an EIT Prototype in Simulated Biological Tissue," *EPiC Series in Engineering*, 2023.

[18] W. R. B. Lionheart et al., "The Reconstruction Problem," in *EIT: Methods, History and Applications*, IOP Publishing, 2005.

[19] J. K. Seo, E. J. Woo, "Electrical Impedance Tomography," *IEEE Signal Process. Mag.*, 2011.

[20] M. Hyvönen, "Complete Electrode Model of Electrical Impedance Tomography: Approximation Properties and Characterization of Inclusions," *SIAM J. Appl. Math.*, 2004.

[21] B. Jin et al., "An Analysis of Electrical Impedance Tomography with Applications to Tikhonov Regularization," *ESAIM: Control, Optim. Calc. Var.*, 2012.

[22] M. Rifai et al., "FPGA-Based Planar Sensor EIT System Characterized by Double Feedback Howland Constant-Current Pump," *IEEE Trans. Instrum. Meas.*, 2024.

[23] Analog Devices, "AD8302 Datasheet: RF/IF Gain and Phase Detector." Analog Devices, Inc., 2002.

[24] A. S. Sedra and K. C. Smith, Microelectronic Circuits, 8th ed. New York, NY, USA: Oxford University Press, 2020.

[25] P. Horowitz and W. Hill, The Art of Electronics, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2015.

[26] Analog Devices, "AD8421: 3 nV/$\sqrt{\text{Hz}}$, Low Power Instrumentation Amplifier," Data Sheet, Rev. A, 2020.

[27] Analog Devices, "ADG726/ADG732: 16-/32-Channel Analog Multiplexers," Data Sheet, Rev. C, 2021.

[28] Texas Instruments, "INA592 High-Precision, Wide-Bandwidth e-trim™ Difference Amplifier," Data Sheet, Rev. F, 2021.
[29] Analog Devices, *ADA4001-2 Low Noise, Low Input Bias Current, Rail-to-Rail Output, JFET Dual Operational Amplifier*, Data Sheet, Rev. C, Norwood, MA, USA, May 2013.
[30] A. Adler and W. R. B. Lionheart, "Uses and Abuses of EIDORS: An Extensible Software Base for EIT," *Physiological Measurement*, vol. 27, no. 5, Art. no. S03, pp. S25–S42, May 2006, doi: 10.1088/0967-3334/27/5/S03.
[31] R. P. Henderson, J. G. Webster, and D. K. Swanson, "Thoracic Electrical Impedance," Annual Conference of the Engineering in Medicine and Biology Society, 1978.
[32] Demonstration video of the proposed EIT system. https://youtube.com/shorts/ymB5hOLzLm8
[33] Demonstration video of the proposed EIT system. https://youtu.be/vd7jVn2Xs2M
[34] A. S. Tucker, R. M. Fox, and R. J. Sadleir, "Biocompatible, High Precision, Wideband, Improved Howland Current Source With Lead-Lag Compensation," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 1, pp. 63–70, Feb. 2013, doi: 10.1109/TBCAS.2012.2199114.

FBG-Based Optical Patch Sensor for Differential Thermal Detection of Skin Cancer

Nazly Rocio Hincapie-Monsalve
Postgraduate Program in Electrical Engineering
Universidade Federal do Espírito Santo
Fundación Universitaria Católica Lumen Gentium
Vitória, Brazil
orcid=0009-0002-6481-9667

Camilo Arturo Rodríguez Diaz
Postgraduate Program in Electrical Engineering
Universidade Federal do Espírito Santo
Vitória, Brazil
orcid=0000-0001-9657-5076

Patrick Marques Ciarelli
Postgraduate Program in Electrical Engineering
Universidade Federal do Espírito Santo
Vitória, Brazil
orcid=0000-0003-3177-4028

Alberto Ferreira De Souza
Applied Computational Intelligence Lab
Universidade Federal do Espírito Santo
Vitória, Brazil
orcid=0000-0003-1561-8447

Teodiano Freire Bastos-Filho
Postgraduate Program in Electrical Engineering
Universidade Federal do Espírito Santo
Vitória, Brazil
orcid=0000-0002-1185-2773

***Abstract*— Skin cancer remains one of the most prevalent malignancies worldwide, and early detection significantly improves patient survival. Malignant skin lesions exhibit altered metabolic activity and vascularization, often leading to localized superficial hyperthermia relative to surrounding healthy tissue. Clinically reported temperature differences range between 0.6 and 2.2 °C. This work investigates the technical feasibility of a Fiber Bragg Grating (FBG)-based optical biosensor for detecting such clinically relevant thermal variations. An experimental calibration was conducted in the temperature range of 20 to 40 °C, resulting in an average thermal sensitivity of approximately 9.84 pm/°C. Based on this sensitivity, expected Bragg wavelength shifts associated with lesion-related temperature increments were calculated to range between 6 and 22 pm. Spectral simulations were performed using experimentally acquired baseline data at 34 °C to model localized temperature increases. The results demonstrate that the predicted wavelength shifts fall within the measurable resolution range of standard optical interrogators. A dual-point differential configuration is proposed, where one FBG is placed over healthy skin as a reference and a second one over the lesion area. Differential measurement enhances robustness against environmental fluctuations and systemic thermal variations. Overall, this proof-of-concept study supports the feasibility of a non-invasive, high-resolution optical patch sensor for detecting localized thermal anomalies associated with suspicious skin lesions.**

Keywords—Fiber Bragg Grating, Optical biosensor, Skin cancer detection, Thermal sensing, Differential measurement, Biomedical instrumentation

I. INTRODUCTION

Skin cancer, including melanoma (MEL), basal cell carcinoma (BCC), and squamous cell carcinoma (SCC), represents one of the most prevalent malignancies worldwide. Early diagnosis significantly increases survival rates, however, current detection methods still rely primarily on visual inspection followed by invasive biopsy confirmation.

According to the Global Cancer Observatory statistics for 2022, a total of 331,722 new cases of MEL were diagnosed globally, resulting in 58,667 deaths [1]. In addition, 1,234,533 new cases of non-MEL skin cancers were reported, with 69,416 associated deaths. The incidence of skin cancer varies considerably across geographical regions, with generally higher occurrence rates observed in men compared to women [2]. In the United States alone, approximately 108,270 new MEL cases were reported, including 64,220 cases in men and 44,050 in women, along with 13,120 deaths attributed to the disease [2].

From a global perspective, Europe accounts for 44.1% of MEL incidence, and 44.6% of MEL-related mortality, whereas North America represents 49.2% of the incidence of non-MEL skin cancers. In contrast, Asia accounts for 46.1% of mortality associated with non-MEL skin cancers [1]. Furthermore, projections indicate a concerning increase in disease burden, with melanoma incidence expected to rise by more than 50%, reaching approximately 510,000 new cases annually by 2040. Mortality is also projected to increase by approximately 68%, reaching 96,000 deaths per year [3]. These trends highlight the urgent need for improved prevention strategies and, critically, the development of reliable early detection technologies.

From a physiological standpoint, malignant tissues exhibit increased metabolic activity due to enhanced glycolysis and angiogenesis [4]. This metabolic rewiring leads to elevated energy consumption and heat generation, producing localized thermal gradients at the skin surface. Infrared thermography studies have reported superficial hyperthermia in MEL lesions, ranging from approximately 0.6 to 2.2 °C relative to adjacent healthy skin [4], [5], [6], suggesting that these temperature differences can serve as a measurable physical biomarker for early detection.

Literature review of dermatological studies provides the clinical foundation for the proposed system. Infrared thermography images, as shown in Fig. 1, consistently reveal localized thermal anomalies in malignant lesions, typically represented using color-coded temperature scales. These scales range from cooler tones (blue/green) to warmer tones (yellow/red), corresponding to increasing surface temperatures. In these images, malignant regions are highlighted by higher intensity colors, indicating localized hyperthermia relative to surrounding healthy tissue.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

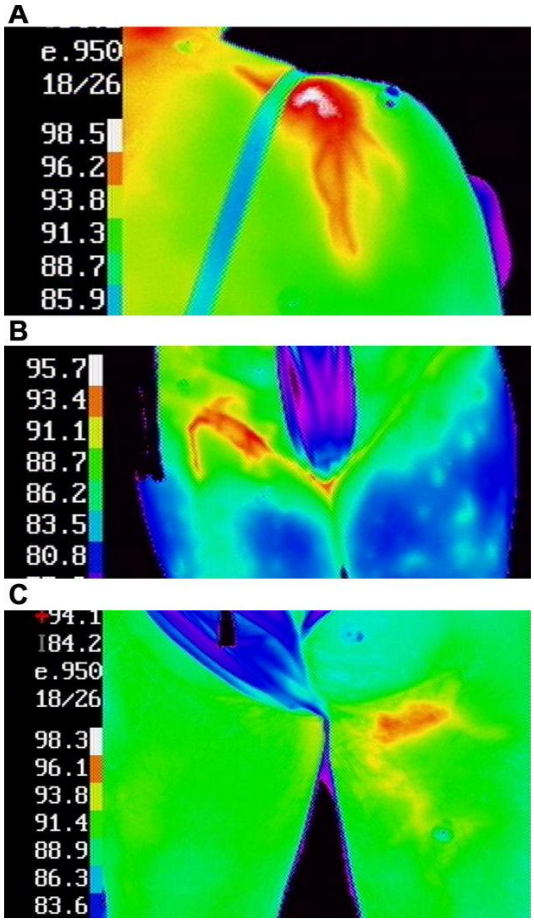


Fig. 1. Representative infrared thermography images of skin lesions, illustrating localized temperature variations using a color-coded scale [6].

Shada et al [6], reported temperature differences, such as shown in Table 1, which compiles affected anatomical regions and corresponding thermal variations. These values were used to define the clinically relevant input range for the proposed sensing approach.

TABLE I. TABLE TYPE STYLES

Case	Lesion Temp (°F)	Normal Skin Temp (°F)	ΔT (°F)	ΔT (°C)
A (Shoulder)	98–99	95–97	2–4	1.1–2.2
B (Groin)	95–96	92–94	1–3	0.6–1.7
C (Thigh)	98–99	96–97	1–3	0.6–1.7

FBG sensors offer an attractive solution for non-invasive thermal monitoring due to their high wavelength sensitivity, immunity to electromagnetic interference, and compatibility with wearable configurations [7], [8], [9], [10].

In recent years, FBG-based sensors have been widely explored in the biomedical field for applications such as body temperature monitoring, respiratory and cardiac activity sensing, minimally invasive surgical instrumentation, and smart textile integration for continuous health monitoring [11], [12]. Their capability to provide real-time, high-resolution measurements, combined with their small size and flexibility, make them particularly suitable for direct contact with biological tissues.

Additionally, FBG sensors have been employed in medical environments where conventional electronic sensors are limited, such as in magnetic resonance imaging (MRI) settings, where electromagnetic immunity is essential. Their multiplexing capability also allows for multiple sensing points along a single optical fiber, enabling distributed sensing systems for spatially resolved physiological measurements.

The Bragg wavelength is defined as

$$\lambda_B = 2n_{eff}\Lambda, \tag{1}$$

where n_{eff} is the effective refractive index, and Λ is the grating period.

Temperature variations modify both parameters, producing a wavelength shift, as shown in Fig. 2, described by

$$\Delta\lambda = S\Delta T, \tag{2}$$

where $S\Delta T$ represents thermal sensitivity.

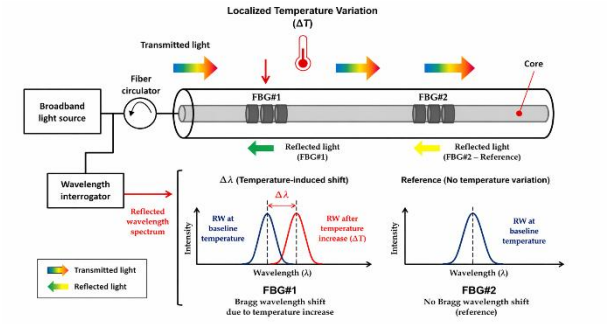


Fig. 2. Physical operating principle of the dual-FBG sensing system under localized temperature variation [13].

In this context, the following research question is formulated:

Can a FBG-based optical patch sensor reliably detect clinically relevant differential temperature variations between healthy skin and malignant lesions through measurable Bragg wavelength shifts suitable for future screening-oriented applications?

This work should be interpreted as a proof-of-concept study aimed at evaluating the technical feasibility of detecting clinically relevant thermal variations using FBG technology. The proposed approach does not assess diagnostic performance in patients and does not replace clinical diagnostic procedures. Instead, it establishes an experimental framework for future validation using biological phantoms, ex vivo tissues, and clinical datasets, while proposing a dual-point differential optical patch architecture to reduce the influence of environmental and systemic thermal fluctuations.

II. METHODOLOGY

The proposed methodology is structured into three main stages: clinical foundation, sensor characterization, and system simulation and design.

A. Clinical Foundation

A comprehensive review of dermatological literature was conducted to identify physiological markers associated with malignant skin lesions, with particular emphasis on localized thermal variations between healthy and cancerous tissues. Infrared thermography studies, such as the work by Shada et al. [6], report temperature differences between malignant lesions and surrounding healthy tissue ranging from 0.6 to 2.2 °C. This range was adopted as the clinically relevant detection interval for the proposed sensing system.

B. Sensor Characterization

The experimental procedure focused on the thermal characterization of the FBG sensor. The Bragg wavelength was continuously monitored using an optical interrogator (Hyperion si255), while the reflected spectral peak was tracked for subsequent analysis.

The optical source was configured using an Anritsu MS9740A optical spectrum analyzer (OSA), as shown in Fig. 3. A broadband super luminescent diode (SLD) was connected to the measurement system through an optical isolator to prevent back reflections and ensure stable emission during spectral acquisition. The OSA wavelength span was set between 1500 and 1600 nm, covering the operational range of the FBG sensor.

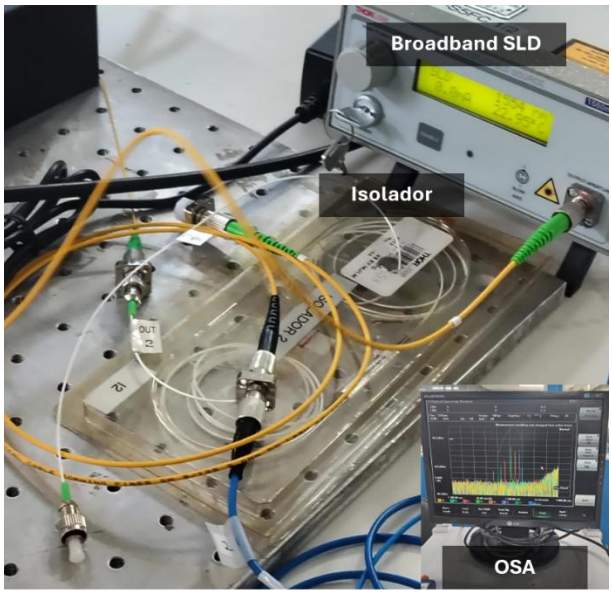


Fig. 3. Experimental setup for optical source characterization. The broadband SLD is connected to the optical spectrum analyzer through an optical isolator to suppress back reflections and ensure stable spectral measurements.

The Bragg wavelengths of the FBG sensors were determined using the reflection-based configuration shown in Fig. 4. In this setup, the broadband SLD was connected to port 1 of an optical circulator, whereas the FBG under test was connected to port 2. The reflected signal from port 3 was directed toward both the OSA and the optical interrogator (Hyperion si255) for spectral acquisition and high-resolution wavelength tracking. This configuration allows for accurate monitoring of the reflected Bragg peak while maintaining system stability.

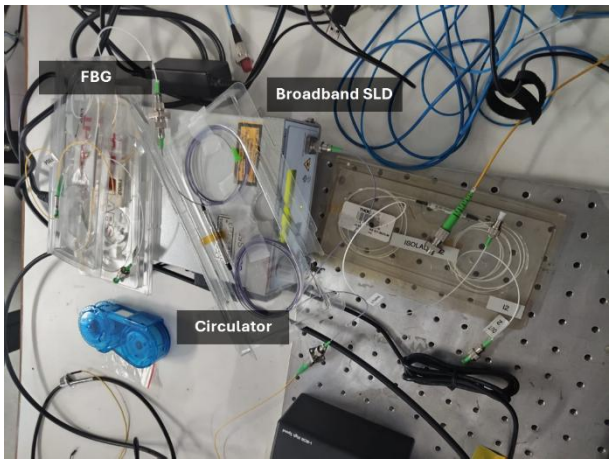


Fig. 4. Experimental setup for Bragg wavelength identification using a reflection configuration. The broadband SLD is connected to port 1 of the optical circulator, the FBG is connected to port 2, and the reflected signal from port 3 is routed to the optical spectrum analyzer OSA.

The thermal calibration was performed using a Peltier module controlled by a temperature driver. As illustrated in Fig. 5, the FBG was placed in direct contact with the Peltier surface using mechanical support to ensure proper thermal coupling and minimize temperature gradients along the grating region.

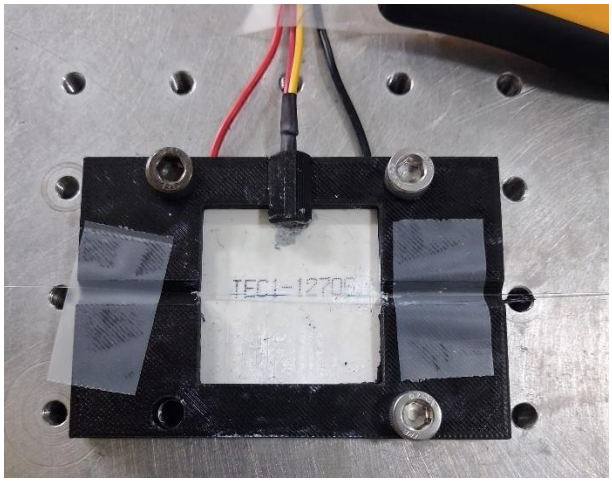


Fig. 5. Peltier module for thermal calibration.

Five complete thermal cycles were conducted. In each cycle, the temperature was increased stepwise from 20 to 40 °C in increments of 2 °C. Subsequently, the temperature was decreased back to 20 °C using the same step size. At each temperature step, sufficient time was allowed to ensure thermal stabilization before acquiring the reflected spectrum. Both heating and cooling phases were recorded to evaluate thermal sensitivity and hysteresis behavior.

The Bragg wavelength was extracted from each spectrum after signal processing, and a linear regression model was applied to establish the relationship between temperature variation and wavelength shift, allowing the determination of the sensor's thermal sensitivity.

To assess the repeatability of the thermal response, the sensitivity values obtained from the five independent heating–cooling cycles were analyzed statistically. For each phase, the mean sensitivity, standard deviation, and

coefficient of variation were calculated. In addition, 95% confidence intervals were estimated using Student's t-distribution, considering the limited number of experimental cycles (n = 5) [14]. The confidence interval was calculated as:

$$CI95\% = \bar{x} \pm t\alpha_{/2} * \frac{s}{\sqrt{n}}, \quad (3)$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the number of experimental cycles, and $t\alpha_{/2}$ is the critical value of the Student's t-distribution. Since five thermal cycles were analyzed, the number of degrees of freedom was n − 1 = 4. For a two-tailed 95% confidence level, the corresponding $t\alpha_{/2}$ value was 2.776. Therefore, this analysis allowed estimating the uncertainty range of the experimentally obtained thermal sensitivity.

C. Simulation and System Design

Using the experimentally measured spectrum as a baseline representing healthy skin conditions, spectral shifts were simulated according to the calibrated thermal sensitivity. These simulations were performed within the clinically relevant temperature range identified in the literature.

Based on this analysis, a dual-point differential sensing architecture was proposed, as shown in Fig. 6. In this configuration, one FBG sensor is positioned over healthy skin to act as a reference, whereas a second sensor is placed over the lesion area. The diagnostic variable is defined as

$$\Delta\lambda = \lambda_{lesion} - \lambda_{healthy}, \quad (4)$$

which enables compensation for global temperature variations and enhances the sensitivity to localized thermal anomalies.

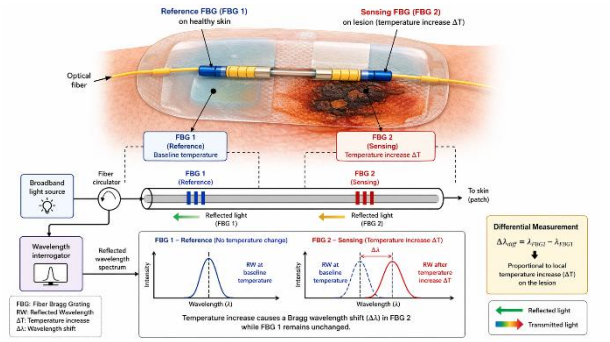


Fig. 6. Proposed dual-point differential FBG configuration for temperature comparison between healthy skin and lesion area [13].

III. Results and Discussion

The experimental results are presented following the methodology stages, including optical source verification, thermal characterization, and system validation through simulation.

A. Optical Source Verification

The measured spectra confirm the expected broadband emission of the SLD within the 1500–1600 nm observation window of the OSA. This spectral range fully covers the operating region of the FBG sensors, ensuring proper interrogation of the Bragg wavelength.

B. Thermal Characterization

Fig. 7 shows the spectral response of the FBG sensor during incremental temperature variation from 20 to 40 °C. A progressive shift of the Bragg wavelength toward higher values is observed as temperature increases, which is consistent with the combined thermo-optic effect and thermal expansion of silica.

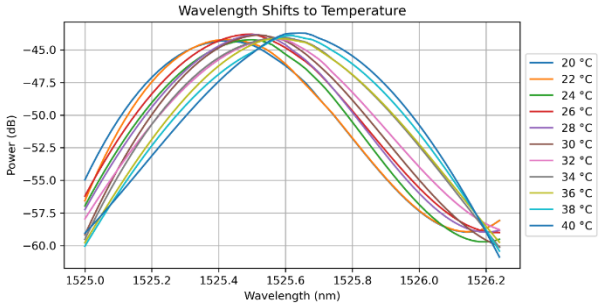


Fig. 7. FBG spectral response during incremental temperature from 20 to 40 °C.

Fig. 8 illustrates the heating–cooling response corresponding to Cycle 1. Clear hysteresis behavior is observed between both phases. For this cycle, the maximum thermal hysteresis reached 40 pm, corresponding to 16.67 % of the full-scale response.

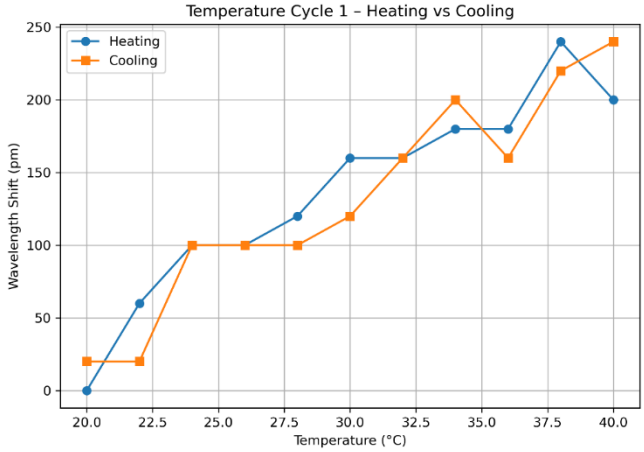


Fig. 8. Hysteresis behavior of the FBG sensor during temperature cycle 1.

Since the measured hysteresis exceeded 5% of the full-scale (FS) response, defined as the total Bragg wavelength shift across the calibrated temperature range, separate linear regression models were fitted for the heating and cooling phases, as shown Fig. 9. This approach provides an initial software-based compensation strategy for the path-dependent thermal response. During heating, the thermal sensitivity was 9.82 pm/°C (R^2 = 0.9014), whereas during cooling it increased to 10.64 pm/°C (R^2 = 0.9237). These values are consistent with the expected contributions of the thermo-optic coefficient and thermal expansion of the fiber.

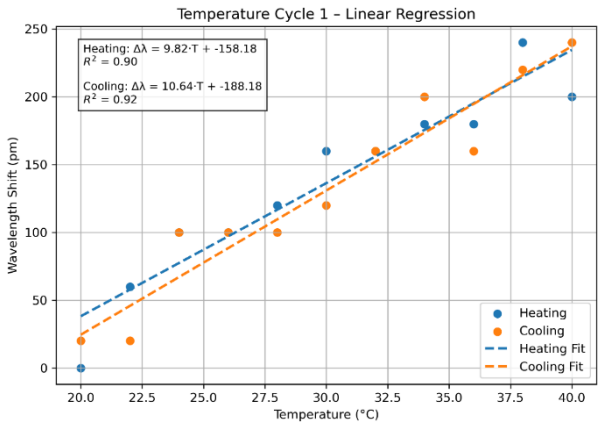


Fig. 9. Linear regression models of the FBG sensor under temperature variation (Cycle 1).

The difference in slopes between heating and cooling phases indicates a path-dependent thermal response, which is consistent with the observed hysteresis. The distinct intercept values further confirm that the temperature trajectory is not fully reversible.

The multi-cycle characterization results are summarized in Table II. Five independent thermal cycles were analyzed to evaluate the repeatability of the FBG sensor under controlled heating and cooling conditions. The heating phase showed an average thermal sensitivity of 9.84 pm/°C, with a standard deviation of 0.92 pm/°C and a coefficient of variation of 9.34%. The cooling phase presented an average sensitivity of 10.15 pm/°C, with a standard deviation of 0.42 pm/°C. These results indicate an acceptable level of repeatability for a proof-of-concept study, although additional long-term stability and drift analyses are still required for wearable deployment.

TABLE II. MULTI-CYCLE TEMPERATURE CHARACTERIZATION RESULTS.

Cycle	Sensitivity Heating (pm/°C)	Sensitivity Cooling (pm/°C)	Max Thermal Hysteresis (pm)
1	9.82	10.64	40.00
2	10.00	9.73	40.00
3	11.27	10.27	40.00
4	8.91	10.45	60.00
5	9.18	9.64	60.00

To further quantify the dispersion of the experimental response, 95% confidence intervals were estimated using Student's t-distribution due to the limited number of cycles. As shown in Table III, the heating phase presented a mean sensitivity of 9.84 pm/°C with a standard deviation of 0.92 pm/°C, resulting in a 95% confidence interval of 8.70–10.98 pm/°C. The cooling phase showed a slightly higher mean sensitivity of 10.15 pm/°C and a lower standard deviation of 0.42 pm/°C, with a 95% confidence interval of 9.63–10.67 pm/°C. These results indicate that both phases exhibited relatively consistent thermal responses, although the lower dispersion observed during cooling suggests greater repeatability in that phase.

TABLE III. CONFIDENCE INTERVALS

Parameter	Mean ± STD	95% CI
Heating sensitivity (pm/°C)	9.84 ± 0.92	8.70–10.98
Cooling sensitivity (pm/°C)	10.15 ± 0.42	9.63–10.67

Considering that the subsequent detectability analysis focuses on clinically relevant temperature increases, the heating phase was selected as the reference for estimating the expected Bragg wavelength shifts. Future real-time implementations should incorporate software-based compensation strategies, including adaptive baseline correction, hysteresis compensation through lookup tables, and dynamic calibration models that account for the heating and cooling trajectory, contact pressure, ambient temperature, and long-term drift. This is particularly relevant because the experimental characterization showed thermal hysteresis values ranging from a minimum of 40 pm to a maximum of 60 pm across the evaluated cycles. These algorithms could improve diagnostic robustness by continuously updating the reference FBG response and reducing false thermal differences caused by environmental or mechanical artifacts.

C. Clinical Range Mapping and Detectability

Based on the calibrated sensitivity obtained from the heating phase, the clinically reported temperature variations (0.6 to 2.2 °C) correspond to Bragg wavelength shifts between approximately 6 and 22 pm.

Considering that the optical interrogator Hyperion si255 provides a resolution on the order of 1 pm or better, these wavelength shifts fall well within the detectable range of the system. This suggests that clinically relevant thermal variations are technically measurable using the proposed FBG-based sensing approach.

D. Spectral Simulation and System Validation

The simulated spectral responses, shown in Fig. 10, demonstrate that temperature increments associated with pathological conditions produce measurable and resolvable Bragg wavelength displacements. The spectral shift is observed as a horizontal translation of the Bragg peak relative to the baseline condition, without distortion of its shape, confirming operation within the linear regime.

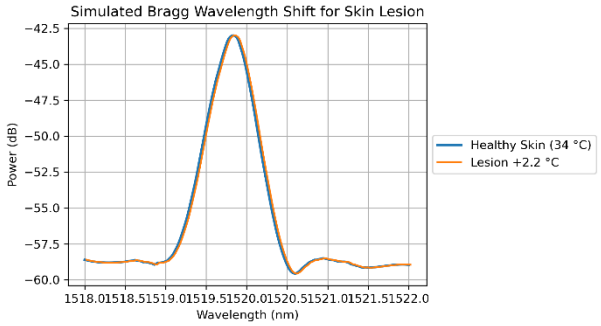


Fig. 10. Simulated spectral response of the FBG sensor under baseline (34 °C) and elevated temperature conditions (+2.2 °C).

Finally, the proposed differential configuration enhances measurement robustness. By comparing the wavelength response of the reference FBG (healthy skin) with that of the sensing FBG (lesion area), the system effectively

compensates for global temperature variations and improves sensitivity to localized thermal gradients.

E. Practical Considerations for Wearable Deployment

In addition to the sensing principle, wearable implementation requires careful consideration of ergonomics, miniaturization, sensor packaging, and long-term usability. In a practical implementation, the FBG sensors could be embedded within a thin, flexible, and biocompatible polymer substrate, such as medical-grade silicone or polyurethane, forming a lightweight and low-profile skin-adherent optical patch. This configuration would ensure conformal contact with curved skin surfaces while minimizing discomfort during use. Proper encapsulation is also necessary to protect the optical fiber from mechanical strain, perspiration, and repeated handling, without significantly increasing the thermal resistance between the skin and the sensing region. The patch could be attached using hypoallergenic medical adhesive, and thin thermal interface layers could be incorporated to improve heat transfer between the skin surface and the sensing element [15], [16]. Such design considerations would support stable placement over the lesion and adjacent healthy tissue during short screening sessions or prolonged monitoring.

IV. Conclusion

This study demonstrated the technical feasibility of using a calibrated FBG sensor to detect clinically relevant differential thermal variations associated with suspicious skin lesions. The experimentally obtained sensitivity, derived from the heating phase of the calibration, enabled the detection of Bragg wavelength shifts between approximately 6 and 22 pm for the clinically reported temperature range of 0.6 to 2.2 °C. These shifts are within the resolution limits of high-precision optical interrogators, and the spectral simulations confirmed that they are resolvable without significant distortion of the spectral response.

The dual-point differential configuration proposed in this work enhances measurement robustness by comparing the wavelength response of a reference FBG positioned over healthy skin with that of a sensing FBG positioned over the lesion area. This strategy allows partial compensation of global temperature variations and improves sensitivity to localized thermal gradients. From an experimental perspective, the FBG sensor exhibited an average thermal sensitivity of approximately 9.84 pm/°C. However, the presence of thermal hysteresis, cycle-to-cycle variability, and moderate dispersion indicates that additional compensation and calibration strategies are required before wearable implementation.

The present study should be interpreted as a proof-of-concept rather than a clinically validated diagnostic system. The validation was performed under controlled laboratory conditions using simulated thermal scenarios derived from published clinical data, rather than biological tissues, tissue phantoms, ex vivo samples, or human subjects. In addition, temperature-based detection does not provide high diagnostic specificity by itself, since localized hyperthermia may also occur in inflammatory processes, infections, wound healing, vascular abnormalities, or other physiological conditions. Therefore, the proposed sensor should be considered as a complementary screening-oriented technology capable of identifying suspicious thermal anomalies that may require further clinical evaluation.

Several practical factors may also influence measurement accuracy in real-world wearable applications, including motion artifacts, variable skin-contact pressure, perspiration, local blood perfusion changes, environmental temperature fluctuations, mechanical deformation, durability, and long-term thermal drift. Although the proposed differential architecture can mitigate part of these effects, future studies should experimentally quantify their impact under realistic operating conditions.

Overall, the results support the potential of FBG-based optical sensing as a non-invasive and high-resolution approach for detecting small differential thermal variations associated with suspicious skin lesions. Future work will focus on validation using tissue phantoms, ex vivo samples, and clinical data, as well as on the integration of the optical patch into wearable platforms. Further investigations should also address software-based compensation strategies, including adaptive calibration and hysteresis correction methods, to improve robustness under real-world wearable conditions. In addition, future studies should evaluate whether different skin cancer types, including melanoma, basal cell carcinoma, and squamous cell carcinoma, exhibit distinguishable thermal signatures associated with their vascularization patterns and metabolic activity.

Acknowledgment

The authors gratefully acknowledge the financial and institutional support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/Brazil) under the MOVE LA AMÉRICA Program [Finance Code 001], the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES/Brazil), and the Instituto de Inteligência Computacional Aplicada (I2CA) [Edital: I2CA – Resolution Nº 285/2021]. The authors also acknowledge the Laboratório de Telecomunicações (LabTel) at UFES for its technical infrastructure and experimental support.

References

[1] Ferlay J *et al.*, “Global Cancer Observatory: Cancer Today,” International Agency for Research on Cancer. Accessed: Apr. 14, 2024. [Online]. Available: https://gco.iarc.fr/today/en

[2] R. L. Siegel, A. N. Giaquinto, and A. Jemal, “Cancer statistics, 2024,” *CA Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, Jan. 2024, doi: 10.3322/caac.21820.

[3] M. Arnold *et al.*, “Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040,” *JAMA Dermatol.*, vol. 158, no. 5, p. 495, May 2022, doi: 10.1001/jamadermatol.2022.0160.

[4] J. Verstockt, S. Verspeek, F. Thiessen, W. A. Tjalma, L. Brochez, and G. Steenackers, “Skin Cancer Detection Using Infrared Thermography: Measurement Setup, Procedure and Equipment,” *Sensors*, vol. 22, no. 9, p. 3327, Apr. 2022, doi: 10.3390/s22093327.

[5] X. Wang, Y. Jiang, S. Xu, H. Liu, and X. Li, “Fiber Bragg Grating-Based Smart Garment for Monitoring Human Body Temperature,” *Sensors 2022, Vol. 22,* vol. 22, no. 11, Jun. 2022, doi: 10.3390/s22114252.

[6] A. L. Shada, L. T. Dengel, G. R. Petroni, M. E. Smolkin, S. Acton, and C. L. Slingluff, “Infrared thermography of cutaneous melanoma metastases,” *Journal of Surgical Research*, vol. 182, no. 1, p. e9, Jun. 2013, doi: 10.1016/j.jss.2012.09.022.

[7] M. Faisal, “Fiber Bragg Grating Temperature Sensor and its Interrogation Techniques,” *Brilliant Engineering*, vol. 4, no. 3, pp. 1–11, Jul. 2023, doi: 10.36937/ben.2023.4840.

[8] W. Wu and X. Liu, “Investigation on high temperature characteristics of FBG sensors,” *Optik (Stuttg).*, vol. 126, no. 20, pp. 2411–2413, Oct. 2015, doi: 10.1016/j.ijleo.2015.06.009.

[9] D. Lo Presti *et al.*, "Fiber bragg gratings for medical applications and future challenges: A review," *IEEE Access*, vol. 8, pp. 156863–156888, 2020, doi: 10.1109/ACCESS.2020.3019138.

[10] M. A. Riza, Y. I. Go, S. W. Harun, and R. R. J. Maier, "FBG Sensors for Environmental and Biochemical Applications - A Review," *IEEE Sens. J.*, vol. 20, no. 14, pp. 7614–7627, Jul. 2020, doi: 10.1109/JSEN.2020.2982446.

[11] M. Pulcinelli, I. Condò, V. Lavorgna, C. Massaroni, E. Schena, and D. Lo Presti, "Combining fiber Bragg grating sensors and artificial intelligence in medicine," *APL Photonics*, vol. 10, no. 3, Mar. 2025, doi: 10.1063/5.0251132.

[12] D. Krizan *et al.*, "Embedding FBG sensors for monitoring vital signs of the human body: Recent progress over the past decade," *APL Photonics*, vol. 9, no. 8, p. 81201, Aug. 2024, doi: 10.1063/5.0226556.

[13] M. Malekzadeh, M. Gul, I. B. Kwon, and N. Catbas, "An integrated approach for structural health monitoring using an in-house built fiber optic system and non-parametric data analysis," *Smart Struct. Syst.*, vol. 14, no. 5, pp. 917–942, Nov. 2014, doi: 10.12989/sss.2014.14.5.917.

[14] Douglas. Montgomery and George. Runger, *Applied Statistics and Probability for Engineers, 5th Edition*, Fifth. John Wiley & Sons, 2010. Accessed: Jun. 06, 2026. [Online]. Available: https://books.google.com/books/about/Applied_Statistics_and_Probability_for_E.html?id=_f4KrEcNAfEC

[15] R. Song, S. Cho, S. Khan, I. Park, and W. Gao, "Lighting the Path to Precision Healthcare: Advances and Applications of Wearable Photonic Sensors," *Advanced Materials*, vol. 37, no. 49, p. 2419161, Dec. 2025, doi: 10.1002/adma.202419161.

[16] J. Liu and H. Liu, "Research on Flexible Sensors for Wearable Devices: A Review," *Nanomaterials*, vol. 15, no. 7, p. 520, Apr. 2025, doi: 10.3390/NANO15070520.

Evaluation of gas vesicles produced by *Anabaena flos-aquae* as contrast agents for ultrasound imaging

Mariana Rocha da Silva
Biomedical Engineering Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
mariana.silva@peb.ufrj.br

Felipe Vianna Garrute
Biomedical Engineering Program/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
felipeviannag@peb.ufrj.br

João Carlos Machado
Biomedical Engineering Program/COPPE and Post-Graduation Program in Surgical Sciences, Department of Surgery, School of Medicine
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
jcm@peb.ufrj.br

Abstract: Ultrasound imaging is widely used for diagnostic applications; however, its image quality is often limited by low contrast and spatial resolution. Gas vesicles (GVs), protein shelled nanostructures naturally produced by microorganisms, have emerged as promising ultrasound contrast agents due to their low cost, biocompatibility, and ability to generate strong ultrasound backscatter signals. In this study, the structural stability and imaging performance of wild-type gas vesicles (GVWT) and modified gas vesicles (Δ GV) were evaluated under different ultrasound incident conditions. Agarose phantoms containing the GVs samples were imaged using a B-mode ultrasound system operating at frequencies of 16 and 21 MHz, with relative transmitted power settings of 1%, 5%, 10%, 20%, and 30%. The optical density of the GV samples was measured at 500 nm (OD500), while ultrasound performance was assessed by measuring grey level intensity from B-mode images. Increasing the ultrasound frequency enhanced the backscatter signal intensity but also promoted gas vesicle collapse. GVWT samples exhibited greater mechanical stability and maintained higher signal intensity than modified ΔGV samples, particularly at higher transmitted relative power. In contrast, the modified gas vesicles showed greater susceptibility to collapse, which was attributed to the absence of the GvpC protein that provides structural reinforcement to the gas vesicle shell. These findings demonstrate that both ultrasound frequency and transmitted power strongly influence the performance of gas vesicles as ultrasound contrast agents and indicate that GVWT exhibits superior imaging performance under the evaluated experimental conditions.

Keywords: ultrasound imaging; gas vesicles; ultrasound contrast agents; ultrasound backscatter; B-mode imaging; GvpC.

I. Introduction

Ultrasonography (US) is a technique widely used in medicine that employs high-frequency sound waves to generate real-time images without causing damage to biological tissues. Commonly used for diagnosing and monitoring health conditions, US operates in a frequency range between 1 and 20 MHz and is an accessible, low-cost, and efficient tool [1] for evaluating orthopedic, cardiac, and ophthalmic lesions. A specific application of this technique is ultrasonic biomicroscopy (UBM), a variation of conventional B-mode ultrasound. UBM employs frequencies ranging from 20 to 100 MHz, enabling the acquisition of high-resolution images at microscopic scales [5], [13]. However, the quality of ultrasound images can be limited by low resolution and contrast, which makes it difficult to visualize fine anatomical structures. To overcome these limitations, ultrasound contrast agents (UCAs), such as microbubbles (MBs), have emerged; these improve image contrast, enabling a more accurate assessment. [5], [9].

Although effective, MBs are expensive, which limits their application in research and clinical practice. A promising alternative comes from gas vesicles (GVs), nanoscale structures produced by microorganisms such as the cyanobacterium *Anabaena flos-aquae* (ANA). These vesicles consist of a gas-filled compartment encapsulated by structural proteins, and when exposed to ultrasound waves, they can act similarly to microbubbles, generating echoes and improving image contrast. GV offer advantages, such as their significantly lower cost and the possibility of structural modifications that enhance their performance as ACUs.[8], [1], [10].

Previous research indicates that modifying GVs, such as removing specific proteins, can increase their response to ultrasound, making them more effective in imaging processes. However, such modifications can also affect the stability of the vesicles, altering their collapse pressure [3] . This study aims to explore the use of modified GVs as contrast agents in ultrasonography by comparing the backscatter signal of the samples when exposed to different ultrasound frequencies (16 MHz and 21 MHz) and relative transmitted US powers (1%, 5%, 10%, 20%, and 30%) in tests conducted with phantoms.

II. Methodology

A. GVs production and purification

The 1403/13F strain of Anabaena flos-aquae was commercially acquired and used as the initial source for GV

production. A 70 mL intermediate culture was established from the original stock and subsequently used to inoculate three parallel cultures in Erlenmeyer flasks. The cultures were maintained in ASM-1 medium [2] under controlled aeration and a 12-hour photoperiod for 20 days. After the growth period, an inoculum from each Erlenmeyer flask was transferred to a new Erlenmeyer flask to renew the culture. For GV production, the culture was transferred to previously sterilized separatory funnels lined on the sides with aluminum foil, allowing light to enter through the top of the funnel. The separatory funnels were kept in a natural environment for 1 to 3 days to allow GV production. Next, the cells were collected in a Falcon tube and a lysis solution (0.5% Tween 20, 500 mM sorbitol, 10 mM Tris, pH 7.5, diluted in PBS) was added. The GVs were then purified by a series of centrifugations at 300 × g at 4 °C for 24 hours. The three parallel cultures were pooled before purification and therefore represented a single batch of GV purification.

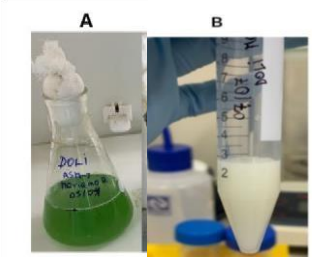


Figure 1. Production and purification of gas vesicles from *Anabaena flos-aquae*. (A) *Anabaena flos-aquae* culture under standard cultivation conditions. (B) Purified gas vesicles (GVs).

B. Optical density measurement

OD measurements were performed using a microvolume UV-Vis spectrophotometer (Thermo Scientific™ NanoDrop™ One C; River Oaks Parkway, San Jose, CA, USA), operating at a wavelength of 500 nm (OD_{500}). For the analysis, 3 µl of the stored GV solution was added, with PBS buffer used as the zero-absorbance standard. The measurement was conducted both to standardize the samples to an OD_{500} of 2 and to verify changes in optical density before and after ultrasound imaging, taking into account the different ultrasonic power levels applied.

C. Removal of the GvpC protein from GVs

The GvpC protein was removed from GVs adjusted to an OD_{500} = 2 with PBS. A solution containing 10 M urea and 100 mM Tris-HCl (pH 8.0) was added in a 3:2 ratio. The suspension was centrifuged at 200 g for 24 h at 4–8°C, resulting in two phases: a white phase (GVs in the supernatant) and a clear phase, which was discarded. The white layer was resuspended in a solution of 6 M urea and 60 mM Tris-HCl (pH 8.0) and centrifuged at 300 g for 4 h at 4–8°C. This process was repeated to remove the GvpC protein from the GVs [4]. The GvpC-free GVs, named ΔGV, were stored in PBS at 4°C for up to one week.

Protein extraction was confirmed by SDS-PAGE [7]. In this case, the proteins were visualized with Coomassie Blue, and the image was recorded using the GelDoc system (BioRad). To enhance the contrast of the bands, ImageJ software (Version 1.53t) was used to process the image, improving the contrast. The absence of the GvpC band (~25 kDa) in the urea treated sample, compared to the control sample, was consistent with GvpC depletion [4].

D. Preparation of Ultrasound Phantoms

The phantoms were prepared using 2% agarose in PBS. The mold, with a diameter of 2 cm and a height of 7 cm, was produced using a 3D printer to form the wells in which the samples were stored, as shown in Fig. 2. Once the phantom had solidified, a GV sample was added in a 1:1 ratio with 2% (m/v) agarose at an OD_{500} of 2. After filling the wells, a layer of agarose was placed over them to seal the openings, and the phantom was left to solidify at room temperature for approximately 1 hour.

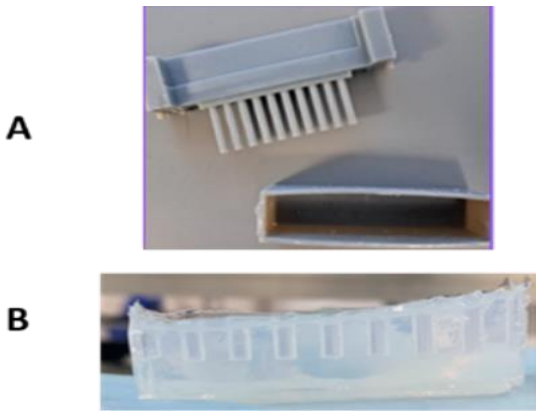


Figure 2: Phantom Mold. In (A) *Phantom* mold created by a 3D printer with a diameter of 2 cm and 7 cm in height. In (B) *Phantom* produced with 2% agarose in PBS, with molded wells.

E. Acquisition and analysis of ultrasound images and quantification by optical density

Ultrasound images of the phantom were acquired using a B-mode ultrasound imaging system (Vevo 2100, isualSonics, Toronto, Canada) equipped with 16- and 21MHz transducer (MS-550D, VisualSonics, Toronto, Canada), operating at relative transmitted power settings of 1%, 5%, 10%, 20%, and 30%. For each experimental condition, a cine loop consisting of 100 B-mode frames was acquired, from which a single image was generated using the Maximum Intensity Projection (MIP) method. Regions of interest (ROIs) were defined on the MIP image, corresponding to the phantom areas containing gas vesicles (GVs) or the negative control (PBS), for quantitative analysis of ultrasound signal intensity.

Following image acquisition, the samples of the control an of GVs in the phantom wells were assessed, a 3-µL aliquot collected and their corresponding OD_{500} was measured using a microvolume UV-Vis spectrophotometer (Thermo Scientific™ NanoDrop™ One C). For each experimental condition, four independent phantom wells were evaluated (n = 4). The four wells were prepared from the same GV culture and purification batch and therefore represent technical/experimental replicates rather than independent biological replicates.

III. Results and Discussion

A. Analysis of OD_{500} for GV_{WT} and ΔGV Concentrations

Figures 3 and 4 show the values for OD_{500} of GV_{WT} and ΔGV, measured by at OD_{500}, for the control group (samples not exposed to ultrasound) and after exposure to relative ultrasound transmit power settings of 1%, 5%, 10%, 20%, and 30% at 16 MHz (Figure 3) and 21 MHz (Figure 4). All experiments were performed in quadruplicate (n = 4), and all samples were standardized to an initial OD_{500} of 2.

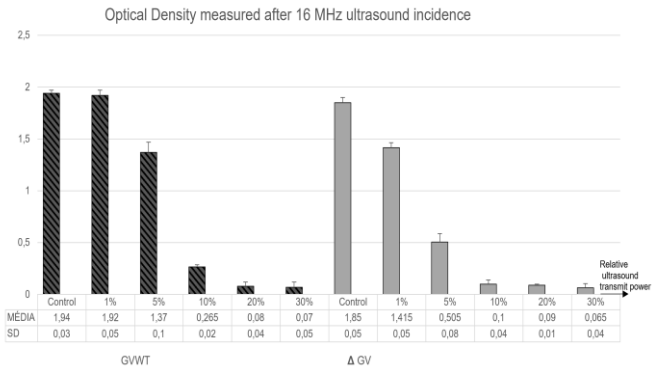


Figure 3. Mean (± SD) OD_{500} for gas vesicles exposed to 16 MHz ultrasound at relative transmit power settings of 1%, 5%, 10%, 20%, and 30%, for control samples and samples exposed to 16 MHz ultrasound at relative transmit power settings of 1%, 5%, 10%, 20%, and 30%. Hatched bars represent GVWT samples, whereas gray bars represent ΔGV samples.

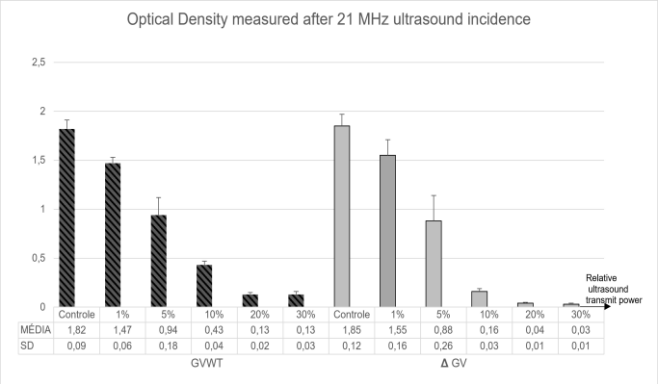


Figure 4. Mean gas vesicle concentration at 21 MHz. Mean optical density at 500 nm (OD_{500} ± SD) for control samples and samples exposed to 21 MHz ultrasound at relative transmit power settings of 1%, 5%, 10%, 20%, and 30%. Hatched bars represent GVWT samples, whereas gray bars represent ΔGV samples.

A reduction in OD_{500} for GV_{WT} relative to the control was observed after exposure to 1% relative transmit power at both 16 MHz (1.03%) and 21 MHz (19.5%). These findings indicate that increasing the ultrasound frequency was associated with a greater reduction in OD500 for GVWT under the conditions evaluated. This trend was consistently observed across all tested power settings, with GVWT exhibiting a higher extent of collapse at 21 MHz than at 16 MHz.

The ΔGV samples exhibited a response that differed from the behavior expected based on previous reports [6], [8]. Nevertheless, the overall results indicate that ΔGV samples are less resistant to ultrasound exposure than GVWT samples, consistent with previous studies. As shown in Table 1, the reduction in OD500 for GVWT relative to the control was evaluated at 16 MHz across different relative power levels (1%, 5%, 10%, 20%, and 30%). Similarly, Table 2 presents the corresponding results obtained at 21 MHz.

	GVwt samples				
	1%	5%	10%	20%	30%
16 MHz	1,03%	29,38%	86%	95,88%	96,40%
21 MHz	19,5%	48,35%	76,37%	92,86%	92,86%

Table 1: Comparison of the reduction in OD_{500} for GVwt relative to the control at frequencies of 16 MHz and 21 MHz at different relative power levels (1%, 5%, 10%, 20%, and 30%)

	Δ GV samples				
	1%	5%	10%	20%	30%
16 MHz	23,24%	73,43%	94,60%	95,14%	96,76%
21 MHz	16,21%	52,43%	91,35%	97,84%	98,38%

Table 2: Comparison of the reduction in in OD_{500} for ΔGV relative to the control at frequencies of 16 MHz and 21 MHz at different relative power levels (1%, 5%, 10%, 20%, and 30%)

ΔGV samples exhibited a distinct response. At a relative transmit power of 1%, their OD_{500} decreased by 23.24% at 16 MHz and by 16.27% at 21 MHz. Unlike GV_{WT} samples, ΔGV samples lack the GvpC protein, which decreases the elasticity of the gas vesicle shell. Consequently, one possible explanation is that these structures can better withstand the resonance induced by ultrasound, thereby shifting-up their collapse pressure.

However, at higher relative transmit power settings, ΔGV samples exhibited a reduction in OD_{500} greater than that for GV_{WT} samples, indicating lower mechanical resistance under these conditions. Furthermore, exposure to relative transmit power levels above 10% resulted in a decrease in the OD_{500} of all gas vesicle samples (Tables 1 and 2). This behavior is likely due to the higher acoustic pressures generated at increased transmit powers, which may be associated with increased destabilization or collapse

B. Ultrasound Image Analysis

Phantom B-mode ultrasound images were acquired at different relative transmitted US power settings and using frequencies of 16 MHz and 21 MHz are presented in Figures 5 and 6, respectively, with PBS used as the negative control. Figures 7 and 8 present the grey level US image intensity of the phantom with gas vesicles at relative transmitted power settings of 1%, 5%, 10%, 20%, and 30% for 16 MHz (Figure 7) and 21 MHz (Figure 8), respectively. All experiments were performed in quadruplicate (n = 4), and the samples were standardized to an initial OD_{500} of 2.

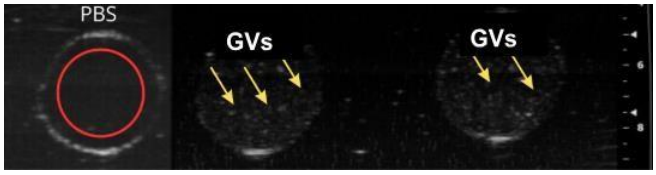


Figure 5. B-mode ultrasound (BMU) images of gas vesicles acquired at 16 MHz. Bmode ultrasound images of PBS (negative control) and GVWT samples (indicated by the yellow arrows) obtained at a relative transmit power of 1% using a B-mode ultrasound system operating at 16 MHz. The red circular outline indicates the region of interest (ROI) selected within the PBS medium.

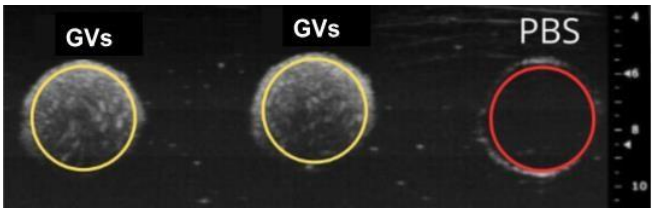


Figure 6. B-mode ultrasound (BMU) images of gas vesicles acquired at 21 MHz. Bmode ultrasound images of PBS (negative control) and GVWT samples obtained at a relative transmit power of 1% using a B-mode ultrasound system operating at 21 MHz. The yellow and red circular outlines indicate the regions of interest (ROIs) selected for the gas vesicle and PBS media, respectively.

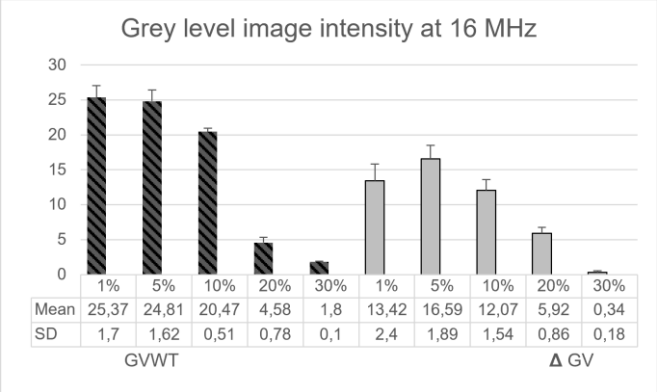


Figure 7. Mean grayscale intensity of US images of phantoms with gas vesicles at 16 MHz. Mean signal intensity (± SD) measured from B-mode ultrasound images of gas vesicle samples exposed to relative ultrasound transmitted power settings of 1%, 5%, 10%, 20%, and 30% using the ultrasound image system operating at 16 MHz. Hatched bars represent the results for GV_{WT} samples, whereas gray bars represent the results for ΔGV samples.

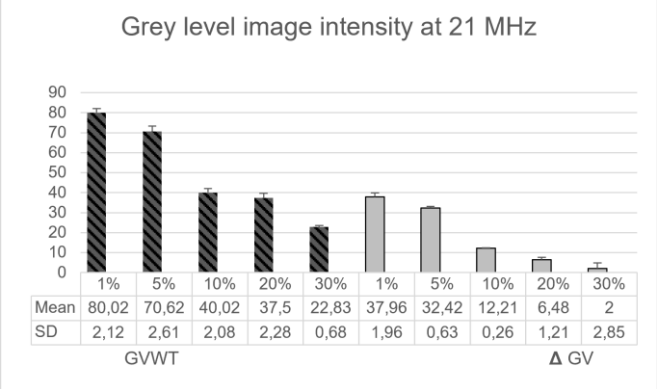


Figure 8. Mean grayscale intensity of gas vesicles at 21 MHz. Mean signal intensity (± SD) measured from B-mode ultrasound images of gas vesicle samples exposed to relative ultrasound transmit power settings of 1%, 5%, 10%, 20%, and 30% using a Bmode ultrasound system operating at 21 MHz. Hatched bars represent the results for GVWT samples, whereas gray bars represent the results for ΔGV samples.

Comparison of the ultrasound image grey level intensity from GV_{WT} samples at 16 MHz and 21 MHz revealed a marked increase in signal intensity at the higher frequency. The mean (±SD) signal intensity was 80.02 (± 2.12) at 21 MHz, compared with 25.37 (± 1.70) at 16 MHz This increase with frequency reflects the greater resonance efficiency of gas vesicles at higher ultrasound frequencies, consistent with the findings of Marsh et al. (1998), who reported that higher frequencies generate stronger backscatter signals, resulting in improved image contrast.

Comparison of signal intensity across the different relative ultrasound transmitted power settings showed that GV_{WT} samples produced the highest grey level intensity at a relative transmitted US power of 1% at both 16 MHz and 21 MHz. Increasing the relative transmitted power to 10%, 20%, and 30% resulted in a substantial reduction in signal intensity, reaching up to 92.91% under some experimental conditions. This marked decrease is consistent with increased GV destabilization or collapse at higher relative transmitted power, supporting previous studies demonstrating that excessive ultrasound pressure can exceed the collapse threshold of gas vesicles [3].

For the ΔGV samples (Figure 7), analysis at 16 MHz showed that the highest signal intensity was obtained at a relative transmitted power of 5%. Increasing the relative transmitted power to 10% and 20% produced significant reductions in signal intensity, with a decrease of up to 64.32% at 20% relative transmitted power. At a relative transmitted power of 30%, the signal intensity was comparable to that of the negative control, suggesting nearly complete collapse of the gas vesicles. At 21 MHz, a similar trend was observed, with the highest signal intensity recorded at 1% relative transmitted power, followed by progressive decreases as the transmit power increased. This effect was more pronounced for ΔGV samples, indicating greater sensitivity to increasing acoustic pressure. The reduced mechanical stability of ΔGV is likely associated with the absence of the GvpC protein, which decreases the structural reinforcement of the gas vesicle shell and increases susceptibility to collapse under elevated ultrasound pressures [11].

IV. Conclusion

The ultrasound backscatter signal was strongly influenced by both ultrasound frequency and relative transmitted power. GV_{WT} samples exhibited a substantially higher mean image grey level signal intensity at 21 MHz (80.02 ± 2.12) than at 16 MHz (25.37 ± 1.70), indicating enhanced resonance efficiency at the higher frequency. This observation is consistent with previous studies reporting that higher ultrasound frequencies produce stronger backscatter signals and improved image contrast [12].

For both frequencies, the highest backscatter signal from GV_{WT} samples was observed at a relative transmitted power of 1%. Increasing the relative transmitted power to 10%, 20%, and 30% progressively reduced signal intensity, with reductions of up to 92.91% under some experimental conditions. These findings indicate that increasing acoustic pressure promotes gas vesicle collapse once the collapse threshold is exceeded, in agreement with previous reports [3].

A similar trend was observed for ΔGV samples. At 16 MHz, the highest signal intensity occurred at a relative transmitted power of 5%, whereas increasing the transmitted power to 10% and 20% resulted in reductions of up to 64.32%. At 30% relative transmitted power, the signal intensity approached that of the negative control, suggesting nearly complete gas vesicle collapse. At 21 MHz, the maximum signal intensity was observed at 1% relative transmitted power, followed by a progressive decrease with increasing transmitted power. Overall, ΔGV samples exhibited greater sensitivity to elevated acoustic pressures than GV_{WT} samples, likely due to the absence of the GvpC protein [11].

ACKNOWLEDGMENTS

The authors thank the Biological Physics Laboratory from UFRJ for their collaboration. The authors gratefully acknowledge the financial support provided by the Coordination for the Improvement of Higher Education Personnel (CAPES), the Research Support Foundation of the State of Rio de Janeiro (FAPERJ), and the National Council for Scientific and Technological Development (CNPq).

REFERENCES

[1] Feng, H., Yang, W., & Hielscher, T. (2008). Power Ultrasound. Food Science and Technology International, 14(5),433–436. https://doi.org/10.1177/1082013208098814

[2] GORHAM, P. R. et al. Isolation and culture of toxic strains of Anabaena flos-aquae (Lyngb.) de Bréb: With Plate 7, 1 figure and 3 tables in the text. SIL Proceedings, 1922-2010, v. 15, n. 2, p. 796–804, fev. 1964

[3] LAKSHMANAN, A. et al. Molecular Engineering of Acoustic Protein Nanostructures. ACS Nano, v. 10, n. 8, p. 7314–7322, 23 ago. 2016.

[4] .LAKSHMANAN, A. et al. Preparation of biogenic gas vesicle nanostructures for use as contrast agents for ultrasound and MRI. Nature Protocols, v. 12, n. 10, p. 2050–2080, 7 set. 2017.

[5] F. S. Foster, C. J. Pavlin, K. A. Harasiewicz, D. A. Christopher, and D. H. Turnbull, “Advances in ultrasound biomicroscopy,” Ultrasound in Medicine & Biology, vol. 26, no. 1, pp. 1–27, Jan. 2000, doi: 10.1016/S0301-5629(99)00096-4.

[6] PFEIFER, F. Distribution, formation and regulation of gas vesicles. Nature Reviews Microbiology, v. 10, n. 10, p. 705–715, out. 2012.

[7] SAMBROOK, J. F.; RUSSELL, D. W. Molecular Cloning: A Laboratory Manual. 3rd. ed. [s.l.] Cold Spring Harbor Laboratory Press, 2001. v. 1,2,3

[8] SHAPIRO, M. G. et al. Biogenic gas nanostructures as ultrasonic molecular reporters. Nature Nanotechnology, v. 9, n. 4, p. 311–316, abr. 2014.

[9] WHEATLEY, M. A.; COCHRAN, M. Ultrasound contrast agents. Journal of Drug Delivery Science and Technology, v. 23, n. 1, p. 57–72, 2013.

[10] A. E. Walsby, “Gas Vesicle,” Microbiological Reviews - American Society for Microbiology, vol. 58, no. 1, pp. 94–144, Mar. 1994, doi: 1 0146-0749/94/$04.00.

[11] D. Maresca et al., “Biomolecular Ultrasound and Sonogenetics,” Annu. Rev. Chem. Biomol. Eng., vol. 9, no. 1, pp. 229–252, Jun. 2018, doi: 10.1146/annurev-chembioeng-060817-084034.

[12] MARSH, J. N. et al. Frequency and concentration dependence of the backscatter coefficient of the ultrasound contrast agent Albunex®. The Journal of the Acoustical Society of America, v. 104, n. 3, p. 1654–1666, set. 1998.

[13] A. Greco et al., “Ultrasound Biomicroscopy in Small Animal Research: Applications in Molecular and Preclinical Imaging,” Journal of Biomedicine and Biotechnology, vol. 2012, pp. 1–14, 2012, doi: 10.1155/2012/519238.

Hall-Effect Multichannel Sensing Architecture for Pelvic Floor Biomechanics: Technical Reassessment and Redesign Proposal

Elio Armando Nunes de Lima
Postgraduate Program in Electrical Engineering (PPGEE)
University of Brasília (UnB)
Federal Institute of Brasília (IFB)
Brasília, DF, Brazil
ORCID: 0000-0002-1516-9648
elio.lima@ifb.edu.br

Adson Ferreira da Rocha
Postgraduate Program in Electrical Engineering (PPGEE)
University of Brasília (UnB)
Brasília, DF, Brazil
ORCID: 0000-0003-4532-1791

Aline Teixeira Alves
Faculty of Ceilândia (FCE)
University of Brasília (UnB)
Brasília, DF, Brazil
ORCID: 0000-0003-0262-7475

***Abstract*—**This paper presents a technical reassessment of the Perin-Mag Hall-effect multichannel sensing architecture for pelvic floor biomechanics and proposes a redesign suitable for quantitative biomedical instrumentation. The original experimental record describes a flexible intravaginal probe with two four-channel arrays, neodymium magnets, A3144 Hall devices, embedded acquisition, and bench loading from 0 to 6 N. No volunteers were recruited. Because raw repeated measurements and a complete circuit schematic were not available, the present work does not claim new bench validation. Instead, the published calibration plot was digitized as a secondary analysis. Eleven points, sampled at 0.5 N intervals over 0.5–5.5 N, were extracted from the plot, yielding an estimated slope of 0.046 mV/N and $R^2 = 0.985$. These metrics are plot-derived estimates and are not traceable measurements with quantified uncertainty. The main technical limitation is the use of the A3144, a Hall-effect switching device with threshold and hysteresis rather than a linear ratiometric analog output. Accordingly, the paper reframes the contribution as a functional architecture review and redesign proposal centered on replacing the switch with a linear ratiometric Hall sensor and implementing a documented acquisition and metrological chain. The available evidence supports feasibility of the concept, but not clinical force measurement or validated eight-channel force distribution.

Keywords—Hall-effect sensor, magnetic transducer, biomedical instrumentation, pelvic floor biomechanics, multichannel sensing, metrology, rehabilitation engineering.

I. INTRODUCTION

Pelvic floor muscle dysfunction is associated with urinary incontinence, pelvic organ prolapse, and impaired quality of life. Objective assessment supports rehabilitation planning, but current methods provide different and complementary measurands: manual palpation provides ordinal clinical grades, perineometry provides intravaginal pressure, electromyography measures electrical activation, and ultrasound provides anatomical or kinematic information [1]–[6]. Multidirectional sensing has therefore been investigated as a means of obtaining localized mechanical information [7].

The Perin-Mag was developed as an intravaginal prototype based on magnets, Hall-effect devices, embedded electronics, and a flexible mechanical interface. The original record reports bench tests using a dynamometer, oscilloscope, multimeter, and regulated power supply, with loads reported from 0 to 6 N [8]. The same record explicitly states that testing was performed without volunteers and that clinical correlation remained a future requirement [8].

The present manuscript responds to a technical reassessment of that work. The objective is not to defend an unsupported analog calibration, but to identify what the available evidence demonstrates, what it does not demonstrate, and how the sensing architecture should be redesigned for quantitative transduction. The contributions are: (1) documentation of the evidence boundary; (2) transparent secondary digitization of the published calibration plot; (3) reconciliation of the reported continuous curve with the electrical behavior of the A3144; and (4) a proposed architecture based on a linear ratiometric Hall sensor and a complete, traceable acquisition chain.

II. TECHNICAL REASSESSMENT OF THE SENSING ARCHITECTURE

A. Measurand and transduction chain

The intended measurand is mechanical loading applied to a compliant probe. The physical transduction concept is magnet displacement relative to a Hall element: deformation changes the magnet-to-sensor geometry, which changes the magnetic field at the sensing location. For a quantitative instrument, the complete chain must therefore be defined as mechanical load → displacement → magnetic field → electrical output → digitization → calibration model → force estimate.

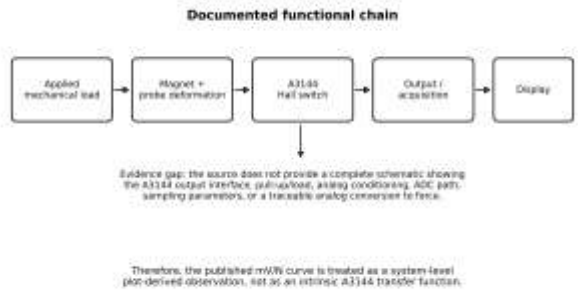


Fig. 1. Documented functional chain of the prototype. The source record does not provide the complete electrical schematic required to establish a traceable analog force conversion.

The available record documents the use of an A3144 Hall-effect device and an Arduino Mega acquisition platform [8]. However, it does not contain a complete circuit showing the A3144 output interface, pull-up resistor or load, analog

conditioning stage, ADC configuration, sampling parameters, or an explicit transfer function from electrical output to force. Consequently, these missing elements cannot be reconstructed without speculation.

B. A3144 device behavior and interpretation

The A3144 is a Hall-effect switch. Its internal architecture includes a Hall element, amplifier, Schmitt trigger, and open-collector output; the output changes state when the magnetic field crosses defined operate and release thresholds, with built-in hysteresis [9]. Allegro distinguishes such switching devices from linear Hall ICs, whose outputs vary continuously and are commonly ratiometric with supply voltage [10]. Therefore, the A3144 should not be treated as an intrinsic linear millivolt-per-newton transducer.

This distinction does not invalidate the mechanical concept. A complete system could, in principle, exploit switching events or a separate analog signal in the acquisition chain. However, the available documentation does not identify such a signal or circuit. The continuous mV-versus-force curve must therefore be treated as a system-level observation from the previous record, not as a direct transfer characteristic of the A3144.

III. SECONDARY ANALYSIS OF THE AVAILABLE CALIBRATION RECORD

A. Data provenance and digitization

The quantitative analysis in this paper is a secondary analysis of Fig. 23 from the authors' previous experimental record [8]. The underlying numerical worksheet was not available. The figure was rendered from the source PDF at 300 dpi and the plotted markers were manually digitized using the labeled axes. Eleven points were extracted at 0.5 N intervals from 0.5 to 5.5 N. The approximate output values read from the plot were 0.22, 0.23, 0.25, 0.26, 0.30, 0.33, 0.36, 0.38, 0.39, 0.42, and 0.43 mV. The extracted values are provided only to reproduce the secondary analysis; they are not raw experimental observations.

TABLE I. POINTS EXTRACTED FROM THE PUBLISHED CALIBRATION PLOT

Force (N)	Output (mV)
0.5	0.22
1	0.23
1.5	0.25
2	0.26
2.5	0.3
3	0.33
3.5	0.36
4	0.38
4.5	0.39
5	0.42
5.5	0.43

B. Fitting procedure

Ordinary least-squares regression was applied to the digitized pairs using $V = aF + b$, where V is the plotted output in millivolts and F is force in newtons. The calculation produced $V = 0.046F + 0.187$ mV, with $R^2 = 0.985$, root-mean-square residual of 0.00896 mV, and maximum absolute residual of 0.01855 mV. The fitting was implemented with a short Python script using NumPy; the script and the point table accompany the revised manuscript package. Because the digitization process introduces reading error and the original repetitions are unavailable, none of these metrics is presented as a traceable metrological result.

IV. AVAILABLE EXPERIMENTAL EVIDENCE

A. Prototype and channel visualization



Fig. 2. Finished Perin-Mag prototype (original Fig. 16).

The completed prototype integrates the flexible probe, cable, and acquisition console. The original record reports that the probe was tested under applied force and that the interface was designed to provide flexible mechanical coupling [8]. The figure establishes the existence of the integrated prototype but does not constitute evidence of clinical suitability.

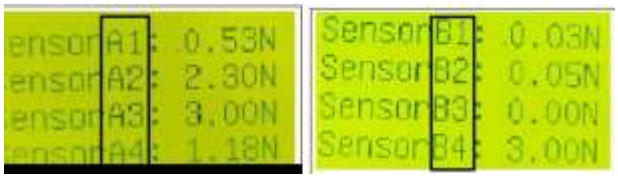


Fig. 3. Representative displayed channel values for Array 1 (original Fig. 19) and Array 2 (original Fig. 20).

The displayed values demonstrate that the prototype can present channel-specific numerical outputs. They should not be interpreted as validated force measurements because channel calibration, reference uncertainty, and the analog acquisition path are not documented sufficiently to establish traceability.

B. Channel response plots

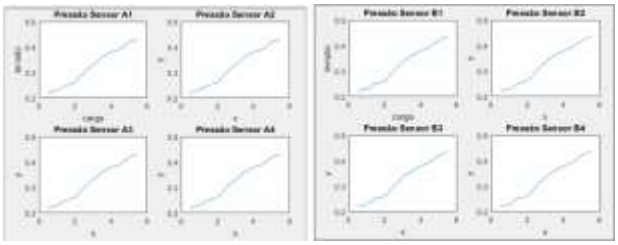


Fig. 4. Previously reported response curves for Array 1 (original Fig. 21) and Array 2 (original Fig. 22).

The eight plotted responses show broadly similar monotonic trends. This observation supports the existence of a common mechanical-sensing behavior, but it does not establish eight-channel quantitative force distribution. Between-channel slope statistics, cross-talk tests, loading/unloading cycles, and repeated measurements were not available.

C. Representative calibration plot

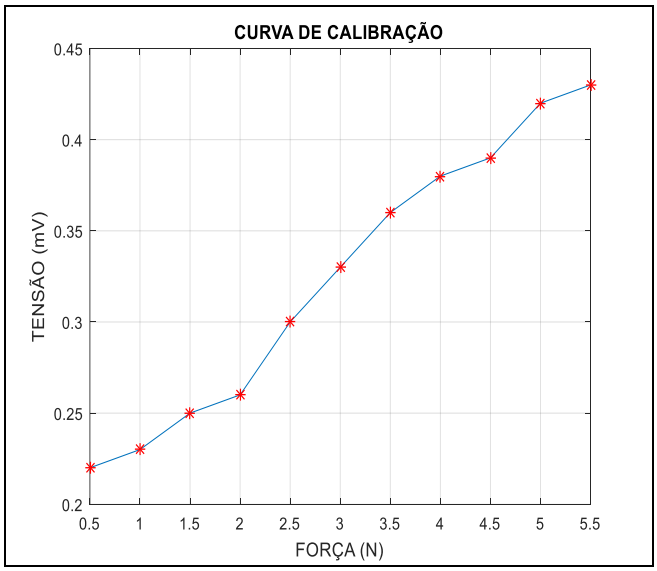


Fig. 5. Previously reported representative calibration curve (original Fig. 23), used only for secondary digitization.

The digitized curve is approximately linear over the plotted interval, but the present analysis must be interpreted as evidence of a previously observed system-level trend. It cannot establish sensor sensitivity, repeatability, hysteresis, resolution, or uncertainty because those quantities require raw repeated observations and a defined measurement chain.

TABLE II. EVIDENCE STATUS OF THE CURRENT PROTOTYPE

Claim	Evidence status
0–6 N bench loading	Reported in previous record
Monotonic channel responses	Supported by plotted curves
Sensitivity 0.046 mV/N	Plot-derived estimate only
$R^2 = 0.985$	Plot-derived estimate only
Repeatability / hysteresis	Not established
Eight-channel force distribution	Not established
Clinical validity	Not assessed

V. REDESIGN FOR QUANTITATIVE BIOMEDICAL INSTRUMENTATION

A. Linear Hall transducer and complete acquisition path

The primary redesign recommendation is to replace the A3144 switching device with a linear ratiometric Hall-effect IC selected according to the magnetic field range generated by the chosen magnet and mechanical geometry. A linear device provides a continuous analog output that can be directly conditioned and digitized, making sensitivity and calibration physically interpretable [10].

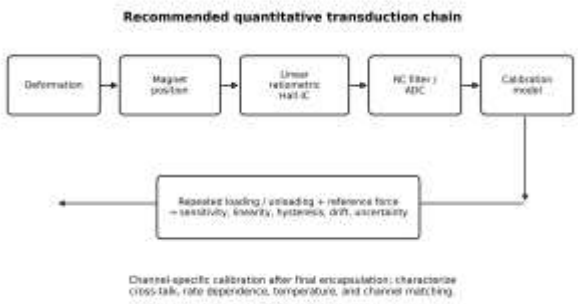


Fig. 6. Proposed redesign for traceable quantitative transduction using a linear ratiometric Hall sensor.

The proposed electrical chain is: linear Hall IC → defined RC conditioning/anti-alias filter → ADC → channel-specific calibration → force estimate. The mechanical stage should first be characterized as magnetic field versus gap, followed by force versus gap and force versus electrical output. This separates the magnetic, mechanical, and electronic transfer functions and permits identification of the dominant source of nonlinearity.

B. Metrological validation protocol

Before any clinical study, each channel should undergo repeated loading and unloading cycles using a traceable force reference. The protocol should quantify sensitivity, linearity, hysteresis, repeatability, drift, resolution, loading-rate dependence, temperature dependence, channel matching, and cross-talk. An uncertainty budget should then combine Type A contributions from repeated measurements with Type B contributions from the reference force instrument, ADC,

geometry, environmental conditions, and calibration model [11].

Only after this characterization should an eight-channel force map be reported. Even then, channel position should be described as a measurement location rather than an anatomical muscle identifier unless independently validated by imaging or clinical examination.

VI. LIMITATIONS AND FUTURE VALIDATION

The principal limitation of the present paper is the absence of raw repeated calibration data and a complete electrical schematic for the original prototype. The numerical values obtained from Fig. 23 are therefore secondary, plot-derived estimates. The second limitation is the A3144 device choice, which prevents a straightforward interpretation of a continuous analog calibration without additional circuit evidence. The third is the absence of volunteer testing; no conclusion regarding pelvic floor muscle force, clinical diagnostic accuracy, or rehabilitation effectiveness can be drawn.

Future work should first implement the linear Hall redesign and document the complete circuit, sensor model, magnet grade and dimensions, supply voltage, ADC resolution, sampling rate, filtering, calibration equations, and uncertainty budget. The next stage should compare the redesigned transducer against a traceable mechanical reference, followed by comparative evaluation against established pelvic floor assessment methods. Ethical approval and clinical validation are required before human measurements are reported.

VII. CONCLUSION

The technical reassessment shows that the Perin-Mag provides a meaningful functional concept for spatially distributed magnetic sensing of mechanical deformation, but the available evidence does not support the original interpretation as a quantitatively validated pelvic floor force instrument. The published calibration curve can be reproduced only as a secondary, plot-derived analysis, yielding approximately 0.046 mV/N and $R^2 = 0.985$ for the extracted points; these values are explicitly not traceable measurements. The decisive engineering correction is to replace the A3144 Hall switch with a linear ratiometric Hall transducer and to implement a fully documented acquisition and metrological chain. This reframing preserves the novelty of the multichannel magnetic architecture while aligning the manuscript's claims with the evidence actually available.

ACKNOWLEDGMENT

The authors acknowledge the University of Brasília and the Federal Institute of Brasília for the infrastructure supporting prototype development and laboratory experimentation.

REFERENCES

[1] S. M. Eickmeyer, "Anatomy and physiology of the pelvic floor," Phys. Med. Rehabil. Clin. N. Am., vol. 28, no. 3, pp. 455–460, 2017.

[2] C. Dumoulin, E. J. C. Hay-Smith, G. Habée-Séguin, and J. Mercier, "Pelvic floor muscle training versus no treatment, or inactive control treatments, for urinary incontinence in women," Cochrane Database Syst. Rev., no. 10, CD005654, 2018.

[3] K. Bø and H. B. Finckenhagen, "Vaginal palpation of pelvic floor muscle strength: Inter-test reproducibility and comparison between palpation and vaginal squeeze pressure," Acta Obstet. Gynecol. Scand., vol. 80, pp. 883–887, 2001.

[4] C. H. J. Ferreira, P. B. Barbosa, F. O. Souza, F. I. António, M. M. Franco, and K. Bø, "Inter-rater reliability study of the modified Oxford Grading Scale and the Peritron manometer," Physiotherapy, vol. 97, pp. 132–138, 2011.

[5] J. Hay-Smith and C. Herbert, "Electromyographic analysis of pelvic floor muscle function in women," Neurourol. Urodyn., vol. 20, pp. 571–582, 2001.

[6] H. P. Dietz, "Pelvic floor ultrasound: A review," Ultrasound Obstet. Gynecol., vol. 37, no. 4, pp. 431–439, 2011.

[7] C. S. Saleme et al., "Multidirectional pelvic floor muscle strength measurement," Ann. Biomed. Eng., vol. 37, no. 8, pp. 1594–1600, 2009.

[8] E. A. N. Lima, "Development of an intravaginal device for the evaluation and training of pelvic floor muscles," M.S. thesis, Univ. Brasília, Brasília, Brazil, 2023.

[9] Allegro MicroSystems, "A3141, A3142, A3143, and A3144 sensitive Hall-effect switches for high-temperature operation," Datasheet, 2024.

[10] Allegro MicroSystems, "Hall-effect sensor ICs: switches, latches, and linear devices," Technical Guide, 2026.

[11] B. N. Taylor and C. E. Kuyatt, "Guidelines for evaluating and expressing the uncertainty of NIST measurement results," NIST Technical Note 1297, 1994.

[12] M. Mirzaei and B. Ebrahimi, "A comprehensive review of Hall effect sensors: Principles, applications, and challenges," Measurement, vol. 175, Art. no. 109155, 2021.

[13] S. Sivaraman and R. Saravanakumar, "Comprehensive review on applications of Hall effect sensors," Int. J. Eng. Adv. Technol., vol. 8, no. 6, pp. 480–484, 2019.

[14] C. Dumoulin, D. Bourbonnais, and M. C. Lemieux, "Development of a dynamometer for measuring the isometric force of pelvic floor musculature," Neurourol. Urodyn., vol. 22, pp. 648–653, 2003.

Hardware for Wearable EEG Systems: A Structured Review on Acquisition, Preprocessing, and Embedded Classification

Camila Tiodista de Lima
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
camila.tiodista@ufpe.br

Jéssica Maria Nogueira de Souza
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
jessica.nogueira@ufpe.br

Ana Beatriz de Souza Sampaio
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
anabeatriz.sampaio@ufpe.br

João Gabriel da Gama Nascimento
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
joao.ggnascimento2@ufpe.br

Andressa Laysa Queiroz Ribeiro
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
andressa.laysa@ufpe.br

Ana Paula Silva de Oliveira
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
paula.silvaoliveira@ufpe.br

Maria Karoline da Silva Andrade
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
karoline.andrade@ufpe.br

Giselle Machado Magalhães Moreno
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
giselle.moreno@ufpe.br

Juliana Carneiro Gomes Cassemiro
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
juliana.cgomes@ufpe.br

Clarisse Lins de Lima
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
clarisse.lins@ufpe.br

Wellington Pinheiro dos Santos
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

Maíra Araújo de Santana
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
maira.araujosantana@ufpe.br

***Resumo*—This structured review maps the development of acquisition, preprocessing, and/or classification systems for EEG signals implemented in hardware. A total of 22 articles published between 2020 and 2026 were analyzed, selected from the ACM Digital Library, IEEE Xplore, Scopus, and PubMed. The studies were categorized into three groups: systems with acquisition (Group 1), systems without acquisition (Group 2), and application-specific integrated circuits (Group 3). Results show a predominance of dry electrodes and the ADS1299 converter in Group 1, the use of FPGAs in Group 2 to balance efficiency and flexibility, and ASICs in Group 3 for ultralow power consumption. The concentration of development in Asian countries highlights Latin America's technological dependence. We conclude that protocol standardization and real-world validation are essential for the clinical consolidation of these devices. Ultimately, the insights from this review will underpin a broader research initiative aimed at developing a hardware device for monitoring individuals with Autism Spectrum Disorder (ASD) and dementia.**

***Index Terms*—electroencephalogram, wearable, portable, hardware, acquisition.**

Foundation for Science and Technology of the State of Pernambuco (FACEPE)

I. Introduction

Wearable electroencephalogram (EEG) devices have established themselves as one of the most promising frontiers in biomedical engineering and neuroscience. These devices are characterized by their portability, light weight, wireless connectivity, ergonomic design, and the use of dry electrodes, which eliminate the need for conductive gels. This simplifies the signal acquisition process outside traditional clinical settings [1], [2]. When integrated with artificial intelligence (AI) algorithms, wearable EEG systems have demonstrated performance comparable to that of clinical equipment in tasks such as epileptic seizure detection, emotional state recognition, screening for neurodegenerative disorders, and the control of brain-computer interfaces (BCIs) [3], [4]. This opens new avenues for applications in mental health, cognitive well-being, and assistive technologies for individuals with severe motor impairments [5], [6].

In this context, convolutional neural networks (CNNs), recurrent models (LSTMs), and classical machine learning (ML) methods have been instrumental in extracting electrophysio-

logical biomarkers from EEG signals, thereby enhancing the accuracy of differential diagnoses among neurodegenerative conditions [5], [6]. The combination of preprocessing techniques (band-pass filters, independent component analysis) with feature extraction in the temporal and spectral domains enables classification accuracies exceeding 90% in tasks such as emotion recognition [7]. Embedded processing in low-power systems enables real-time analysis [8], [9].

This review serves as a theoretical foundation for the development of a wearable EEG system aimed at the assistive and diagnostic monitoring of individuals with Alzheimer's disease and Autism Spectrum Disorder (ASD). Recent advances in the development of wearable EEG devices are systematically reviewed herein, examining acquisition hardware architectures, signal processing strategies, and classifier performance in clinical and neurophysiological applications.

II. Methodology

This is a non-systematic literature review employing a structured search strategy. The guiding research question for this work was: How are wearable embedded systems for EEG signal processing currently being developed? In this context, this review aimed to map the development of hardware-implemented systems designed for the acquisition, preprocessing, and/or classification of these signals. The literature search was conducted across four databases: ACM Digital Library, IEEE Xplore, Scopus, and PubMed. The core search string was: (”wearable”OR ”wireless”OR ”portable”) AND ”EEG”AND (”hardware”OR ”system design”OR ”dry electrodes”OR ”BCI”) AND (”Alzheimer”OR ”autism”OR ”neurodevelopmental”OR ”neurodegenerative”OR ”dementia”OR ”emotion recognition”), applied to titles, keywords, and abstracts. Articles published between January 2020 and May 2026 were included, encompassing both journal articles and conference proceedings, given that the most recent advances in this field are typically disseminated primarily through conferences.

The included studies involved hardware implementation of at least one stage of development, regardless of controller type, communication modality, or the degree of system portability. Review articles, out-of-scope studies, design-only proposals, and software-only implementations were excluded. The review process followed three stages: study identification, title and abstract screening, and full-text evaluation. Finally, the pre-selected studies were reviewed in full, and their hardware and performance characteristics were extracted.

III. Results

Figure 1 presents the main results of the stages of this review. Initially, 179 studies were identified based on the search string and automatic filtering criteria for publication date and venue. Fifty-five duplicate articles were removed manually. The remaining 124 articles proceeded to the screening phase; after abstract review, the following were excluded: 18 review articles, 46 software development studies, 20 out-of-scope studies, and 1 experimental design proposal. This screening yielded 39 articles for full-text assessment, of which 17 were excluded for the following reasons: out of scope (5), development projects (3), software development (4), microcontroller used solely to toggle an LED via external command (1), unclear methodology (1), and unavailability of the full text (3). Finally, 22 full-text articles were analyzed.

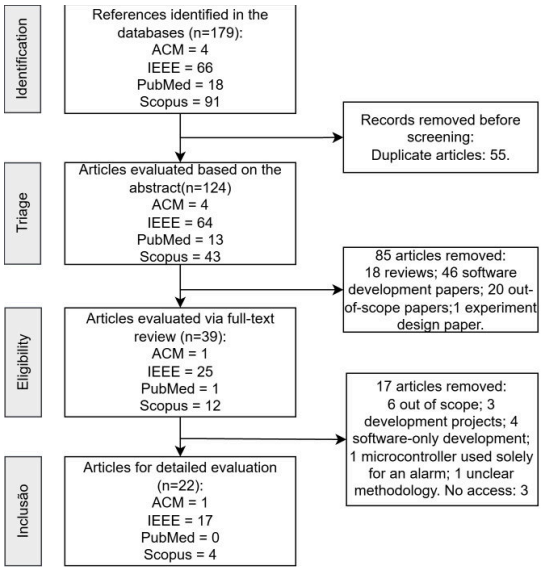


Figura 1. Complete workflow for the article survey.

Regarding the temporal distribution of publications, the year 2020 recorded the highest absolute number of studies (4). However, it is important to note that the data for 2026, comprising 3 publications, cover only the first four full months of the year, suggesting a potential for continued growth. Between 2021 and 2025, the publication counts were, respectively: 3 (2021), 1 (2022), 3 (2023), 3 (2024), and 4 (2025). Regarding geographical origin, based on the institutional affiliation of the first author of each study, Figure 2 indicates that South Korea, China, and India lead the publication ranking with 4 studies each, followed by Pakistan and Germany, with 2 publications each.

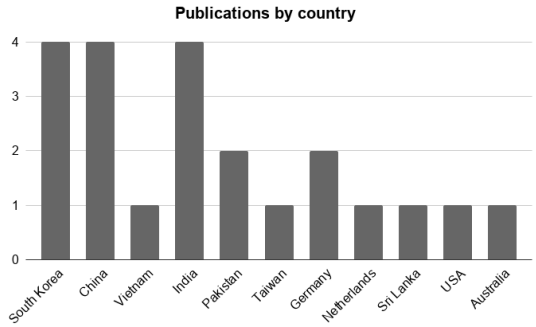


Figura 2. Number of articles by country.

The studies were grouped based on the hardware platform utilized: those that employed commercial prototyping chips (Groups 1 and 2) and those that synthesized and/or fabricated an application-specific integrated circuit (ASIC). Among the

studies based on commercial platforms, those that developed the acquisition system within their embedded pipeline were assigned to Group 1. Group 2, in turn, comprises studies dedicated exclusively to signal preprocessing and/or classification.

Group 1 comprised 10 of the 22 studies [7], [10]–[18], Group 2 included 8 [9], [19]–[25], and Group 3 consisted of the remaining four [26]–[29].

A. Group 1: Systems with Data Acquisition

The main results of this group are summarized in Table I. Initially, emotion recognition emerges as the predominant application, accounting for 60% of the applications in the analyzed studies. Only two studies implemented additional stages beyond signal acquisition. Negi et al. (2025) [14] performed analog filtering of the signals from the ADS1115 converter, employing an AD6020 amplifier and a TL084 filter. In turn, Subashri et al. (2026) [13] were the only researchers to implement the complete embedded pipeline, encompassing acquisition, preprocessing, and classification.

Regarding the acquisition interface, most systems utilized dry electrodes and low-density configurations, ranging from 1 to 8 channels, with the exception of the study by Mannatunga et al. (2020) [16], which employed 128 channels. Furthermore, the use of wireless communication modules, such as Bluetooth, Bluetooth Low Energy (BLE), and Wi-Fi, was predominant.

Regarding hardware validation metrics, only Mannatunga et al. (2020) [16] and Shi et al. (2022) [18] reported specifications beyond power consumption. Mannatunga et al. (2020) [16] detailed the front-end latency ($\sim 0.86\mu s$) and the total system latency, accounting for a 512-sample FIFO (1.07 ms). The input-referred noise averaged $\sim 3\mu V$, with a common-mode rejection ratio (CMRR) of $\sim$ 81 dB. The authors also validated the system's reliability through a one-hour continuous recording using 128 channels at 31.25 kSPS, with no recorded sample loss. In turn, Shi et al. (2022) [18] reported an average input-referred noise of $1\mu V$ peak-to-peak for their system.

B. Group 2: Systems Without Data Acquisition

In this group, studies focused exclusively on the development of preprocessing and/or classification systems. Following the trend observed in the previous group, 62.5% of the studies addressed the task of emotion recognition. Only the study by Nguyen-Duy e Dinh-Duc (2026) [9] did not implement real-time classification; instead, the authors developed an encryption system for EEG signals transmitted via Bluetooth. Notably, FPGA-based platforms were employed in half of the studies.

For validation, a subset of the studies utilized established datasets (SEED, DEAP, DREAMER, and FEXD) [9], [21], [22], whereas the remaining articles [19], [23]–[25] employed commercial devices. Wireless communication remained predominant. Most studies in this group reported detailed hardware performance metrics. Table II details power consumption, energy per classification, latency, and hardware resource utilization, along with other specifications.

C. Group 3: ASIC-Based Systems

Table II also presents the main results of the studies focused on the development of ASICs. This group exhibits greater diversity in the applications addressed. Aslam and Altaf (2020) [26] were the only researchers to design and fabricate a chip encompassing the entire embedded pipeline. However, performance validation for emotion recognition was conducted using well-established datasets from the literature (DEAP and SEED). Consequently, the system's acquisition module was not evaluated in practice. This study was the only one in the group to report system latency. Another relevant metric considered was the silicon area, detailed by all studies in this category, with the exception of de Bruin et al. (2020) [28].

IV. Discussion

The geographical analysis of the publications reveals the dominance of Asian countries in global hardware development, highlighting the absence of contributions from Latin America. This scenario reinforces the region's status as a technology importer, which directly impacts the clinical efficacy of these devices. Designing circuits for distinct climatic conditions and population profiles overlooks physiological variations, such as drastic changes in perspiration and skin conductivity, which affect signal acquisition in tropical environments. Furthermore, a methodological disproportion is evident: of the 124 unique articles screened, 50 focus exclusively on software development. Therefore, there is an urgent need for investments in the prototyping of high-fidelity acquisition hardware that accounts for these operational variables.

From a technical perspective, Group 1 highlights the widespread adoption of dry electrodes in low-density configurations, driven by the demand for wireless wearable systems. However, the inherent imbalance in skin-electrode contact impedance critically degrades the CMRR. Among the 10 studies, only Mannatunga et al. (2020) [16] reported the practical CMRR (approximately 81 dB). This value is consistent with the specifications of the converter used (Intan RHA2132); however, it falls short of the theoretical 110 dB of the ADS1299. Although the ADS1299 component was employed in 50% of the studies due to its optimization for biopotential signals, the authors omitted the measurement of the system's CMRR under operating conditions. The failure to specify the data acquisition conditions precludes test reproducibility and exacerbates the risk of suboptimal performance on skin characteristics differing from those originally tested.

Regarding energy efficiency, reporting only total nominal power can be misleading for systems that do not operate the entire embedded pipeline continuously. To establish more rigorous hardware metrics, the literature should stratify power consumption by task (acquisition, preprocessing, and classification). Another critical omission in the analyzed literature is system latency, precluding direct comparisons. Only Mannatunga et al. (2020) [16] reported this metric; however, given that it involves a 128-channel device with a wired (USB) connection, the result does not reflect the realities of

daily data acquisition using wearable devices in naturalistic environments.

The validation of the acquisition modules in this group was primarily based on the performance achieved during the classification stage. With the exception of Luo et al. (2025) [11], the systems achieved accuracies exceeding 79%. The lower performance reported by Luo et al. (2025) [11] illustrates a potential methodological flaw: the model was trained using the SEED dataset and tested on in vivo recordings from only two participants. Optimizing algorithms on established software datasets and testing them on real-world analog signals frequently results in severe accuracy degradation, as the model is unfamiliar with the specific acquisition artifacts introduced by the hardware. In contrast, Wang et al. (2025) [17] and Shi et al. (2022) [18] employed robust physical validation by correlating the signals from their devices with those from commercial equipment (including the use of gel electrodes). Both studies achieved a correlation exceeding 95%, corroborating the potential viability of low-cost portable systems.

The studies in Group 2 focused primarily on the implementation of embedded classification algorithms. The sole methodological exception was the study by Nguyen-Duy and Dinh-Duc (2026) [9], who proposed an encryption system for data previously acquired using an ADS1299. In this study, the reported metrics specifically reflect the computational cost of the security layer, which consumed only 1.2 KB of flash memory (out of a total of 85.8 KB) and introduced 5.3 ms of latency (within a total pipeline of 47.5 ms).

To evaluate the efficacy of these proposals, it is necessary to categorize the studies into those that integrated processing with commercial acquisition systems [19], [23]–[25] and those that validated their algorithms using pre-recorded datasets from the literature [9], [20]–[22]. In dataset-based studies, critical questions remain regarding the fidelity and real-world utility of these devices. Although the reported latency metrics support the designation of ”real-time”operation, this claim lacks methodological validity when the processing is not subjected to the continuous, noisy data stream of actual biological acquisition.

Regarding energy efficiency, FPGA-based and dedicated processor architectures exhibited extremely low continuous power consumption. The study by Aslam et al. (2021) [20] is particularly notable, reporting a power consumption of only $8.65\mu W$; however, the authors failed to report the energy cost per classification event. In contrast, the studies by Chen et al. (2023) [21] and Gonzalez et al. (2020) [22] detailed the efficiency of their logical operations. The system by Chen et al. (2023) [21] achieved a performance of 9.26 GOPS/W, which is lower than the 11 GOPS/W achieved by Gonzalez et al. (2020) [22], highlighting an architectural trade-off: the latter system attained greater energy efficiency despite operating with more channels, owing to the use of lower sampling rates and resolution.

Regarding clinical efficacy, classification performance plateaued between 75% and 86% accuracy, underscoring the urgent need for more robust feature extraction methods for edge processing. Studies that utilized commercial devices for real-time classification exhibited greater diversity in hardware and electrode configurations. In these instances, the reported latency more accurately reflects the complete processing cycle (from acquisition to classification). However, the systematic omission of fundamental metrics by a significant portion of the authors precludes any fair quantitative comparison. This superficiality in documenting results compromises experimental reproducibility and reflects a severe lack of scientific rigor, thereby hindering the consolidated advancement of the biomedical instrumentation field.

The third group encompasses studies focused on the development of ASICs. Among the four analyzed studies, only Aslam and Altaf (2020) [26] performed the physical fabrication of the chip and integrated the acquisition module. The result was a 6 mm^2 microchip featuring the complete embedded pipeline, exhibiting an energy consumption of 16 μJ per classification and a latency of 48 s. However, the authors conducted validation using the pre-recorded DEAP and SEED datasets. Consequently, the reported energy and time metrics do not reflect the continuous operation of the complete pipeline in practice, highlighting a methodological flaw by isolating the fabricated hardware from real-world analog acquisition conditions. Nevertheless, the merit of the design in unifying the acquisition, preprocessing, and classification stages on a single silicon die is undeniable. The remaining three studies were limited to computational simulations (synthesis or post-layout phases) without physical silicon implementation. Despite lacking physical validation, these proposals stand out for their rigorous optimization, reporting power and area values considerably lower than those of the commercial systems in the previous groups. The most notable achievement in miniaturization is that of Bahadur and Boppana (2025) [27], who designed a noise-removal circuit occupying an area of only 0.004571 mm^2.

It is essential to acknowledge the inherent limitations of this review, which stem from the time and formatting constraints characteristic of a conference paper. The literature search was restricted to four databases and included conference proceedings, which frequently present preliminary or partial research findings. Furthermore, the scope of the search string must be considered, as it may have omitted relevant studies indexed under alternative technological nomenclatures.

V. Conclusion

This review outlines guidelines for the development of low-cost wearable EEG devices aimed at monitoring ASD and dementia syndromes, highlighting the urgent need for complete, rigorously documented, and validated embedded pipelines. The concentration of research in Asia underscores the technological dependence of Latin America. From a technical standpoint, the state of the art consolidates the use of dry electrodes and the ADS1299 converter in the analog front-end, the adoption of FPGAs to balance efficiency and flexibility, and the design of ASICs for ultra-low power consumption scenarios. However, methodological heterogeneity

and the systemic underreporting of hardware metrics preclude direct comparisons and limit reproducibility. Therefore, future research must prioritize the adoption of standardized reporting protocols and the comprehensive validation of these systems in real-world clinical and data acquisition scenarios.

VI. Declaration on the Use of Artificial Intelligence

The authors used AI tools for language polishing. The authors assume full responsibility for all content.

Acknowledgment

This work was supported by the Foundation for Science and Technology of the State of Pernambuco (FACEPE).

Referências

[1] A. Sen, C. R. Newton, and G. Ngwende, "Epilepsy in low-to middle-income countries," *Current Opinion in Neurology*, vol. 38, no. 2, pp. 121–127, 2025.

[2] S. Armand Larsen, L. Klok, W. Lehn-Schiøler, R. Gatej, and S. Beniczky, "Low-cost portable eeg device for bridging the diagnostic gap in resource-limited areas," *Epileptic Disorders*, vol. 26, no. 5, pp. 694–700, 2024.

[3] A. P. S. de Oliveira, C. T. de Lima, J. M. N. de Souza, L. F. C. de Oliveira, L. H. Calixto, A. S. de Oliveira Silva, C. M. de Lima Simões, M. E. S. B. Lins, B. J. A. P. Barbosa, C. L. de Lima, J. C. G. Cassemiro, M. A. de Santana, G. M. M. Moreno, W. P. dos Santos, and L. A. L. Soares, "Gerongnosis: Multimodal machine learning for the differential diagnosis of alzheimer's disease, frontotemporal dementia, and mild cognitive impairment," *AI in Neurology*, 2026.

[4] Y. Chen, Y. Peng, J. Tang, T. Camilleri, K. Camilleri, W. Kong, and A. Cichocki, "Eeg-based affective brain–computer interfaces: Recent advancements and future challenges," *Journal of Neural Engineering*, vol. 22, no. 3, p. 031004, 2025.

[5] H. Zhang, L. Jiao, S. Yang, H. Li, X. Jiang, J. Feng, S. Zou, Q. Xu, J. Gu, X. Wang *et al.*, "Brain–computer interfaces: the innovative key to unlocking neurological conditions," *International Journal of Surgery*, vol. 110, no. 9, pp. 5745–5762, 2024.

[6] X.-Y. Liu, W.-L. Wang, M. Liu, M.-Y. Chen, T. Pereira, D. Y. Doda, Y.-F. Ke, S.-Y. Wang, D. Wen, X.-G. Tong *et al.*, "Recent applications of eeg-based brain-computer-interface in the medical field," *Military Medical Research*, vol. 12, no. 1, p. 14, 2025.

[7] H. Barki, N.-D. Mai, and W.-Y. Chung, "Optimized XGBoost for Multimodal Affective State Classification Using In-Ear PPG and Behind-the-Ear EEG Signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no. 3, pp. 2139–2152, Mar. 2026. [Online]. Available: https://ieeexplore.ieee.org/document/11123688/

[8] J.-Y. Li, Y.-K. Chen, and W.-C. Fang, "An edge ai accelerator design based on lrcn model for real-time eeg-based emotion detection system on the risc-v fpga platform," in *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2023, pp. 89–90.

[9] X.-B. Nguyen-Duy and A.-V. Dinh-Duc, "A Privacy-Preserving Edge-AI Framework for EEG-based Emotion Recognition using Lightweight Authenticated Encryption over BLE," in *Proceedings of the 2025 10th International Conference on Cloud Computing and Internet of Things*, ser. CCIOT '25. New York, NY, USA: Association for Computing Machinery, Feb. 2026, pp. 1–6. [Online]. Available: https://dl.acm.org/doi/10.1145/3785520.3785521

[10] N.-D. Mai, N. M. Hoang Long, and W.-Y. Chung, "1D-CNN-based BCI system for detecting Emotional states using a Wireless and Wearable 8-channel Custom-designed EEG Headset," in *2021 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS)*, Jun. 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9469818

[11] H. Luo, H. Li, W. Tao, Y. Yang, C.-I. Ieong, and F. Wan, "A Portable and Affordable Four-Channel EEG System for Emotion Recognition with Self-Supervised Feature Learning," *Mathematics*, vol. 13, no. 10, 2025.

[12] N.-D. Mai, H. Barki, and W.-Y. Chung, "A Wearable BTE-EEG Embedded Device for Emotion Monitoring with Quantum Machine Learning," in *2024 Tenth International Conference on Communications and Electronics (ICCE)*, Jul. 2024, pp. 562–566, iSSN: 2836-4392. [Online]. Available: https://ieeexplore.ieee.org/document/10634608/

[13] M. Subashri, B. Samyuktha, and G. S. Uthayakumar, "AI-Driven Adaptive Deep Brain Stimulation Using EEG and Emotion Feedback for Multi-Disorder Modulation," in *2026 International Conference on Computing, Electronics & Communications Engineering (ICCECE)*, Feb. 2026, pp. 1–6, iSSN: 2768-0576. [Online]. Available: https://ieeexplore.ieee.org/document/11492927/

[14] A. Negi, S. Apeksha, C. Bharath, and M. Uzair, "Data Acquisition and Pre-processing of EEG Signals Using Brainwave Sensor," in *Lect. Notes Electr. Eng.*, vol. 1297. Springer Science and Business Media Deutschland GmbH, 2025, pp. 1–15, journal Abbreviation: Lect. Notes Electr. Eng.

[15] N.-D. Mai, H.-T. Nguyen, and W.-Y. Chung, "Deep Learning-Based Wearable Ear-EEG Emotion Recognition System With Superlets-Based Signal-to-Image Conversion Framework," *IEEE Sensors Journal*, vol. 24, no. 7, pp. 11 946–11 958, Apr. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10456557/

[16] K. S. Mannatunga, S. H. M. Ali, M. L. Crespo, A. Cicuttin, and J. G. Samarawikrama, "High Performance 128-Channel Acquisition System for Electrophysiological Signals," *IEEE Access*, vol. 8, pp. 122 366–122 383, 2020.

[17] S. Wang, L. W. Yap, Z. Wang, W. Cheng, and W. Chen, "Ultra-Low Power EEG Headband Based on Gold Nanowire Foam Electrode," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2025, pp. 1–5, iSSN: 2694-0604. [Online]. Available: https://ieeexplore.ieee.org/document/11254834/

[18] Z. Shi, X. Han, B. Jiang, J. Zhang, D. Suo, G. Pei, T. Yan, Y. Wang, J. Wu, and J. Wang, "Wearable Multimodule Bio-signal Acquisition System: Brain Multi-Plus," in *2022 16th ICME International Conference on Complex Medical Engineering (CME)*, Nov. 2022, pp. 77–80. [Online]. Available: https://ieeexplore.ieee.org/document/10063300/

[19] X. Li, H. Deng, J. Ouyang, H. Wan, W. Yu, and D. Wu, "Act as What You Think: Towards Personalized EEG Interaction Through Attentional and Embedded LSTM Learning," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 3741–3753, May 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10144694/

[20] A. R. Aslam, N. Hafeez, H. Heidari, and M. A. B. Altaf, "An 8.62 μ W Processor for Autism Spectrum Disorder Classification using Shallow Neural Network," in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, Jun. 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9458412/

[21] Y.-K. Chen, J.-Y. Li, and W.-C. Fang, "An Edge AI Accelerator of LRCN Model with RISC-V Platform for EEG-based Emotion Real-time Detection System," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2023, pp. 1–5, iSSN: 2766-4465. [Online]. Available: https://ieeexplore.ieee.org/document/10389052/

[22] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. A. M. Elfadel, "An Inference Hardware Accelerator for EEG-Based Emotion Detection," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Oct. 2020, pp. 1–5, iSSN: 2158-1525. [Online]. Available: https://ieeexplore.ieee.org/document/9180728/

[23] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel, "BioCNN: A Hardware Inference Engine for EEG-Based Emotion Detection," *IEEE Access*, vol. 8, pp. 140 896–140 914, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9151874/

[24] R. Janapati, S. Bhattacharya, S. R. Gorre, K. Sreedhar, B. K. Verma, C. Adupa, U. Desai, and K. Bikshalu, "Real-Time Emotion Detection System using Emotive and ESP-32," in *2023 IEEE 20th India Council International Conference (INDICON)*, Dec. 2023, pp. 1211–1215, iSSN: 2325-9418. [Online]. Available: https://ieeexplore.ieee.org/document/10440945/

[25] B. U. Demirel, I. Skelin, H. Zhang, J. J. Lin, and M. Abdullah Al Faruque, "Single-Channel EEG Based Arousal Level Estimation Using Multitaper Spectrum Estimation at Low-Power Wearable Devices," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Nov. 2021, pp. 542–545, iSSN: 2694-0604.

[26] A. R. Aslam and M. A. B. Altaf, "An On-Chip Processor for Chronic Neurological Disorders Assistance Using Negative Affectivity

Classification," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 4, pp. 838–851, Aug. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9139281/

[27] I. Bahadur and L. Boppana, "Area and Power Efficient Architecture For Muscular Artifacts Removal From EEG Signal," in *2025 5th International Conference on Intelligent Technologies (CONIT)*, Jun. 2025, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/11167557/

[28] B. de Bruin, K. Singh, J. Huisken, and H. Corporaal, "BrainWave: an energy-efficient EEG monitoring system - evaluation and trade-offs," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 181–186. [Online]. Available: https://dl.acm.org/doi/10.1145/3370748.3406571

[29] Z. Mo, Y. Yuan, W. Shi, J. Mai, D. Zeng, X. Liu, and G. Li, "Design of Low-power EEG Detection System for Emotion Classification," in *2024 7th International Conference on Electronics Technology (ICET)*, May 2024, pp. 1041–1045, iSSN: 2768-6515. [Online]. Available: https://ieeexplore.ieee.org/document/10673238/

Reference	Application	Embedded Pipeline	Hardware Platform	Communication	Additional Components	Electrode Type	Number of channels	Sampling Rate (Hz)	Resolution (bits)	Power (W)	Classification Performance
[10]	Emotion Recognition	Acquisition	Teensy 3.2	Bluetooth	Noise reduction and amplifier (active circuit) + ADC (ADS1299) + Communication (HC-05)	Dry	8	256	24	N/R	0.93 (ACC)
[11]	Emotion Recognition	Acquisition	ESP32-S3	Wi-Fi	Amplifier (KS1092) + ADC (ADS131M08) + Impedance measurement (DAC7311 + TMUX1309 + INA350)	Both	4	200	24	1	0.69 (ACC); 0.59 (MF1)
[12]	Emotion Recognition	Acquisition	Arduino Portenta H7	BLE e Wifi	ADC (ADS1299) + Communication (Murata 1DX)	Dry	3	256	24	N/R	0.92 (ACC; Prec; Rec; F1)
[13]	Adaptive DBS with Emotion Feedback	Acquisition, Preprocessing, and Real-Time Classification	ESP32	Wi-Fi	Acquisition module (ADS8232) + Electronic feedback actuators (LED or Buzzer)	Dry	3	N/R	12	N/R	0.98 (ACC)
[14]	Depression and ASD Severity Detection	Acquisition and Analog Filtering	Arduino uno	USB	Amplifier (AD6020) + ADC (ADS1115) + filters (TL084 + Resistors and capacitors)	Wet	2	N/R	16	N/R	0.78 (ACC)
[15]	Emotion Recognition	Acquisition	Arduino Portenta H7	Bluetooth	ADC (ADS1299)	Dry	3	256	24	N/R	0.92 (ACC)
[16]	Emotion Recognition	Acquisition	Xilinx's Zynq-7020	Ethernet	Amplifier (Intan RHA2132) + ADC (AD7982)	Wet	128	31.25k	18	2.85	N/R
[7]	Emotion Recognition	Acquisition	nRF52840	Bluetooth	ADC (ADS1299)	Dry	1	256	24	44.4m	0.98 (ACC; Prec; Rec; F1)
[17]	Clinical Applications	Acquisition	STM32L452CEU6	BLE	Communication (E104-BT5032A) + ADC (AD4130-8)	Dry	2	2400 (single-channel); 1200 (all channels)	24	9.472m	0.99 (R)
[18]	Neurofeedback for Neurodegenerative Disorders	Acquisition	STM32F405RGT6	Wi-Fi	ADC (ADS1299) + Filter	Dry	32	500	24	2.47	0.96 (R Signal); 0.98 (R FFT)

ADC Analog-to-Digital Converter
ACC Accuracy
MF1 Macro-F1 Score
BLE Bluetooth Low Energy
Prec. Precision
Rec. Recall
F1 F1 Score
DBS Deep Brain Stimulation
N/R Not reported
ASD Autism Spectrum Disorder
USB Universal Serial Bus
R Correlation coefficient

Tabela I
GROUP 1: SYSTEMS WITH DATA ACQUISITION.

Group 2													
Reference	**Application**	**Embedded Pipeline**	**Hardware Platform**	**Communication**	**Acquisition/ dataset**	**Number of channels**	**Sampling Rate (Hz)**	**Resolution (bits)**	**Power (W)**	**Energy per Classification**	**Latency (s)**	**Resource Usage**	**Classification Performance**
[9]	Emotion Recognition	Preprocessing	Nordic nRF52840	BLE	SEED	2	250	24	49.58m	N/R	47.5m (end-to-end); 6.2m (transmission); 5.3m (encryption)	85.8 KB flash memory	0.88 (ACC)
[19]	Brain-Robot Interaction	Preprocessing and Real-Time Classification	Nvidia Jetson TX2	Bluetooth	Brainlink 2.0	3	N/R	N/R	N/R	N/R	3.04	N/R	0.97 (ACC)
[20]	ASD Diagnosis	Preprocessing and Real-Time Classification	Artix-7 FPGA	-	SEED	4	N/R	10	8.62μ	N/R	2.25ms	18361 LUTs; 10627 FFs	0.85 (ACC)
[21]	Emotion Recognition	Preprocessing and Real-Time Classification	FPGA board Xilinx Kintex-7	Bluetooth	DEAP/ FEXD	8	256	24	180m	9.26 GOPs/W	0.014ms	29438 LUTs; 16313 FFs	0.89 (ACC)
[22]	Emotion Recognition	Real-Time Classification	FPGA board Xilinx Altys Spartan-6	UART	DEAP/ DREAMER	14	128	16	150m	11GOps/W	1ms	26228 LUTs; 15180 FFs	0.78 (ACC))
[23]	Emotion Recognition	Real-Time Classification	FPGA board Xilinx Altys Spartan-6	Bluetooth	Emotiv Epoc +	14	128	16	150m	11GOps/W	1ms	26228 LUTs; 15180 FFs	0.83 (ACC valence); 0.77 (ACC arousal)
[24]	Emotion Recognition	Real-Time Classification	ESP32	Wi-Fi/Bluetooth	Emotiv EEG Headset	N/R	N/R	N/R	N/R	N/R	N/R	N/R	0.80 (ACC)
[25]	Arousal Level Estimation	Preprocessing and Real-Time Classification	STM32H743Z	N/R	Grass Technologies Comet XL	19; 256	1000; 5000	N/R	1.83m	N/R	281ms	N/R	0.65 (ACC)
Group 3													
Reference	**Application**	**Embedded Pipeline**	**Hardware Platform**	**Communication**	**Acquisition/ dataset**	**Number of channels**	**Sampling Rate (Hz)**	**Resolution (bits)**	**Power (W)**	**Energy per Classification**	**Latency (s)**	**Area (mm^2)**	**Classification Performance**
[26]	Emotion Recognition for Chronic Neurological Disorders	Acquisition, Preprocessing, and Classification	ASIC (Fabricated)	BLE	DEAP/SEED	8	128	10	2.04m	16(μJ)	48	6	0.73 (ACC DEAP); 0.70 (ACC SEED)
[27]	Artifact Removal for Neurodeve-lopmental Disorder Diagnosis	Preprocessing	ASIC (Post-Layout)	N/R	N/R	N/R	N/R	16	525.4u	N/R	N/R	0.004571	0.85 (Mean R); 2.03 (RMSE)
[28]	Epileptic Seizure Detection	Preprocessing and Classification	ASIC (Synthesized)	BLE	N/R	2	100	32	467.84u	135(μJ)	N/R	N/R	0.89 (Prec.); 0.53 (Rec.); 0.99 (Spec.)
[29]	Emotion Recognition	Preprocessing and Real-Time Classification	ASIC (Synthesized)	N/R	N/R	6	N/R	16	1.59m	1.96 (μJ)	N/R	4.4	0.78 (ACC)

BLE Bluetooth Low Energy
ACC Accuracy
N/R Not reported
LUT Look-Up Table
FF Flip-Flop
FPGA Field-Programmable Gate Array
GOPs/W Giga Operations Per Second per Watt
ASIC Application-Specific Integrated Circuit
R Correlation coefficient
RMSE Root Mean Squared Error
Prec. Precision
Rec. Recall
Spec. Specificity
F1 F1 Score

Tabela II
GROUP 2 (SYSTEMS WITHOUT DATA ACQUISITION) AND GROUP 3 (APPLICATION-SPECIFIC INTEGRATED CIRCUIT DEVELOPMENT).

Instrumental Improvements of the X-QST Module for Thermal Quantitative Sensory Testing in Diabetic Neuropathy Screening

Júlio Cardoso Scorpion
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0009-0002-8669-2249

João Lucas Pereira dos Santos de Paula
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0000-0001-8230-5358

Denny Daniel Collina
Federal Center for Technological Education of Minas Gerais
Belo Horizonte, Brazil
ORCID: 0009-0008-8537-7320

Manoel de Figueiredo Villarroel
Brazilian Society of Clinical Neurophysiology
Belo Horizonte, Brazil
ORCID: 0000-0001-8506-9207

Jefferson Luz Brum Marques
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0000-0001-8910-4735

***Abstract*—Diabetic peripheral neuropathy (DPN) is a common complication of Diabetes Mellitus, and its early diagnosis remains challenging. Quantitative Sensory Testing (QST) is a well-established psychophysical technique for assessing thermal perception and pain thresholds associated with small-fiber dysfunction. This work presents an updated version of the X-QST module integrated into the SDC-X platform for thermal sensory assessment. Compared with the previous prototype, the system was redesigned to improve thermal performance and usability through the incorporation of a contact thermistor, a compact copper thermode with reduced thermal mass, a closed-loop water-cooling system, and a wireless response unit. In addition, support for both the Method of Limits (MLI) and the Method of Levels (MLE) was implemented, expanding the psychophysical testing capabilities of the platform. Experimental characterization showed approximately linear thermal ramps close to $1\,^{\circ}$C/s within the clinically relevant range around the $32\,^{\circ}$C baseline, consistent with the recommendations of the German Research Network on Neuropathic Pain (DFNS). These improvements position the X-QST as an open, low-cost, and customizable platform, whose clinical and metrological performance remains to be validated, with potential applications in clinical research and multimodal screening of diabetic peripheral neuropathy.**

***Index Terms*—QST, DPN, X-QST, Updated.**

I. Introduction

Diabetes Mellitus (DM) represents a major public health challenge due to its high prevalence and the progressive impact of its chronic complications [1]. Among these complications, Diabetic Peripheral Neuropathy (DPN) stands out because of its high incidence and its association with neuropathic pain, sensory loss, plantar ulceration, and an increased risk of lower-limb amputation [2], [3].

The early detection of DPN remains a significant clinical challenge, as impairments in small nerve fibers (Aδ and C fibers) may precede the involvement of large nerve fibers. Consequently, conventional methods such as the Semmes-Weinstein monofilament test and nerve conduction studies may not be sufficiently sensitive to detect early alterations in thermal and pain perception [4]. In this context, Quantitative Sensory Testing (QST) enables the assessment of thermal detection and pain thresholds, constituting an important tool for the functional evaluation of small nerve fibers [5].

Despite its clinical relevance, the widespread application of QST remains limited by the high cost of commercial equipment, limited portability, the need for dedicated infrastructure, and difficulties in integrating with digital health platforms. These limitations have motivated the development of low-cost instrumented systems featuring appropriate thermal control, wireless communication, and automated data acquisition.

In this context, the SDC-X platform has been developed as a modular solution for screening DM-related complications by integrating different physiological assessment modules, including X-CARDIO, X-EYE, and X-ESC [6]–[9]. As part of this platform, X-QST was developed to determine thermal detection thresholds, thermal pain thresholds, and paradoxical thermal sensations, aiming to identify sensory alterations associated with DPN [10]. Furthermore, this approach establishes a foundation for future peripheral neurological assessments, such as X-NCS, which is intended to evaluate large peripheral nerve fibers [11].

A preliminary version of the X-QST system demonstrated the feasibility of a thermal stimulation device based on a Peltier module, embedded data acquisition, and communication with a centralized software platform [12]. The present version improves thermal sensing, stimulus control, firmware, and the acquisition interface: the temperature sensor was replaced with an $F\mu$ thermistor [13], the Method of Levels (MLE) was incorporated for its higher repeatability and lower

bias in detection-threshold assessment [14]–[16], a water-cooling system was added, and the thermode was redesigned with reduced dimensions and a copper contact surface for its superior thermal conductivity [17]. These changes enable its evaluation as a more complete instrumentation prototype for future sensory screening studies.

Therefore, this work presents the evolution of the X-QST system integrated into the SDC-X platform, describing the implemented instrumental and methodological improvements, its thermal characterization, and its potential as a complementary tool for the early assessment of sensory alterations associated with DPN.

A. Related Work

Since the standardization of the protocol by the *German Research Network on Neuropathic Pain* (DFNS), commercial systems have become the reference standard for performing thermal sensory testing [18].

Among the most widely used devices is the *Thermal Sensory Analyzer* (TSA-II), later incorporated into the Pathway platform, both developed by Medoc. These systems employ a Peltier-based *thermode* with closed-loop temperature control and execute all major DFNS-standardized thermal tests—cold and warm detection thresholds (CDT, WDT), cold and heat pain thresholds (CPT, HPT), the thermal sensory limen (TSL), and paradoxical heat sensation (PHS)—using both the Method of Limits (MLI) and the Method of Levels (MLE). Owing to their high accuracy and repeatability, they are considered the reference standard for clinical and research use, although they rely on proprietary architectures and involve high acquisition and maintenance costs [5], [19].

Another widely adopted system is the *MSA Thermal Stimulator*, developed by Somedic SenseLab AB. The device employs a fluid-cooled Peltier-based *thermode*, closed-loop temperature control, interchangeable thermodes of different sizes, and dedicated software for protocol configuration and automated data acquisition. Despite its high metrological performance, its elevated cost limits its use primarily to specialized pain centers and research laboratories [20], [21].

To increase the accessibility of QST, several research groups have developed portable and lower-cost devices. Among them, the *NerveCheck*, proposed by Ponirakis *et al.* [22], performs cold, warm, vibration, and thermal pain assessments using a Peltier-based *thermode*. The device was clinically validated against the Medoc TSA-II, demonstrating comparable diagnostic performance for the assessment of DPN at an approximate cost of US$500. However, like commercial systems, it remains a proprietary solution with limited flexibility for hardware modifications and the implementation of new experimental protocols [22].

II. Materials and Methods

This section describes the methodology adopted for the prototype, including the updates implemented in the current version and the materials used in its development.

A. QST screening method

The clinical QST protocol proposed by the *German Research Network on Neuropathic Pain* (DFNS) standardized the quantitative assessment of thermal, mechanical, vibration, and pressure sensitivity [5]. In the present study, only thermal tests were considered, as they are the most relevant for the early detection of DPN [23]–[25]. The tests were performed only on the dominant hand and foot, since these regions are sufficient for DPN assessment [26].

The DFNS protocol employs the MLI, in which a *thermode* is initially applied to the skin at the baseline temperature of $32°$C. The temperature is then increased or decreased at approximately $1°$C/s to perform the CDT, WDT, CPT, HPT, and PHS tests. During the detection tests (CDT, WDT and PHS), participants indicate the moment they first perceive a cold or warm sensation. During the pain threshold tests (CPT and HPT), the response is given when the thermal stimulus becomes painful, typically by pressing a response button [5], [27]. The MLI procedure is illustrated in Fig. 1(a).

In addition to the MLI, MLE, based on adaptive psychophysical staircase procedures, was incorporated into the X-QST system [28], [29]. The test begins with a stimulus $4°$C above or below the baseline temperature, after which the thermode immediately returns to the adaptation temperature. Participants respond "YES" or "NO", either verbally or using a two-button joystick, indicating whether the stimulus was perceived. Positive responses decrease the intensity of the subsequent stimulus, whereas negative responses increase it. Whenever the response sequence changes (YES–NO or NO–YES), the step size is reduced by half until it reaches $0.2°$C. The initial $4\,°$C step provides rapid convergence toward the threshold, the successive halving at each response reversal refines the estimate following the transformed up–down staircase rule, and the $0.2\,°$C minimum step defines the resolution at which the procedure stops [28], [29]. Table I and Fig. 1(b) illustrate an example of the stimulus sequence considering a baseline temperature of $32°$C.

The MLI and MLE provide highly consistent estimates of thermal thresholds. However, the MLI generally yields slightly higher threshold values because of the participant's reaction time when pressing the response button, an effect that can be mathematically compensated [16].

TABLE I
Example of the MLE progression.

Temp. (°C)	Response	Step	Step Size
36	YES	$\uparrow 4$	Initial $\uparrow 4$
34	NO	$\downarrow 2$	Halved
35	YES	$\uparrow 1$	Halved
34.5	YES	$\downarrow 0.5$	Unchanged
34	NO	$\downarrow 0.5$	Halved
34.2	YES	$\uparrow 0.2$	Final

B. New Equipment

The X-QST prototype underwent updates throughout its hardware architecture. For clarity, the system is divided into

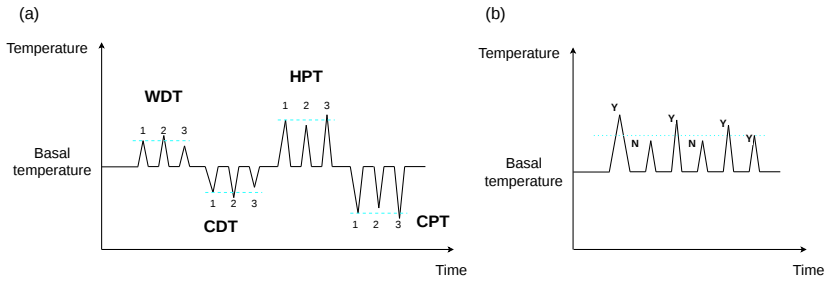


Fig. 1. (a) Method of Limits and (b) Method of Levels.

four modules: the thermal induction module, temperature sensing module, cooling module, and response unit module. The system is powered from either a 127 VAC or 220 VAC mains supply, which is converted to a regulated 12 VDC output through an AC/DC converter. Figure 2 presents the electrical and communication architecture of the complete system.

In addition, the thermode structure was redesigned to improve thermal performance and the execution of QST procedures.

1) Thermal Induction Module: The thermal stimulation module is based on the TEC1-12706 Peltier device [30], which is widely employed in QST systems according to the literature [12]. The module measures 40×40 mm and operates at up to 12 V, as recommended by the manufacturer. Thermal stimulation is achieved through the Peltier effect, in which one surface absorbs heat while the opposite surface dissipates it. Reversing the current direction inverts the heating and cooling surfaces.

Current reversal is performed using a BTS7960 H-bridge driver, allowing bidirectional current control through PWM signals generated by an ESP32 microcontroller. The ESP32 is responsible for controlling the thermal stimulation process as well as the wireless communication of the system via Wi-Fi and Bluetooth. The microcontroller is powered through an L7805 voltage regulator, which converts the 12 V supply from the AC/DC converter to 5 V.

Temperature is regulated by a discrete PID controller running on the ESP32, using the contact-thermistor reading (Eq. 1) as feedback. The controller output (± 255) is mapped to the PWM duty cycle of the H-bridge, with its sign selecting the heating or cooling channel, so that a single loop drives the Peltier device in both directions. The gains were determined empirically during the baseline and stimulation onset phases of the test.

2) Temperature Sensing Module: The previous version of the X-QST employed an infrared temperature sensor, whose measurements were affected by factors such as surface emissivity, sensing distance, field of view, and ambient temperature, compromising the thermal control performance. To improve temperature regulation, the sensor was replaced by a direct-contact Fμ thermistor (223Fμ5183), enabling faster and more accurate thermal feedback required to maintain the heating and cooling rate of approximately 1°C/s specified by the DFNS protocol [5].

The thermistor is attached directly to the thermode, providing improved accuracy, repeatability, and thermal ramp control, while increasing patient safety through reliable monitoring of the operating limits, restricted to temperatures between 0°C and 50°C to prevent skin damage [31]. The selected sensor offers an accuracy of $\pm 0.5\%$, a thermal time constant between 52 and 70 ms, and is specifically designed for medical applications [13].

The thermistor is read through a resistive divider—a 15 kΩ pull-up between the 3.3 V rail and the acquisition node, with the thermistor to ground and a 10 μF filtering capacitor—sampled by the 12-bit analog input of the ESP32. Each temperature value is obtained by averaging 80 consecutive samples, and the resulting signal is streamed at 1 Hz. The thermistor resistance is computed as $R_{\mathrm{NTC}} = R_{\mathrm{p}}\, V/(V_{\mathrm{cc}} - V)$, with $R_{\mathrm{p}} = 15$ kΩ, and converted to temperature by the β-parameter equation

$$\frac{1}{T} = \frac{1}{T_0} + \frac{1}{\beta}\,\ln\!\left(\frac{R_{\mathrm{NTC}}}{R_0}\right), \tag{1}$$

where $R_0 = 14015\ \Omega$ is the nominal resistance at $T_0 = 298.15$ K (25 °C) and $\beta = 3454$ K is the datasheet coefficient, with T in kelvin. A per-sample outlier rejection and an adaptive exponential filter are then applied to reduce electrical noise. The 0–50 °C operating range is enforced in firmware: the temperature setpoint is clamped to these limits and, if the estimated contact temperature reaches either bound, the Peltier drive is switched off and the test is aborted, preventing skin damage.

3) Cooling Module: The cooling module is a new component introduced in the current version of the X-QST. It consists of a closed-loop water-cooling system driven by an SC-300T water pump controlled by the ESP32 through a MOSFET-based switching circuit. A flyback diode is included to suppress the inductive voltage generated during pump switching, protecting the electronic circuitry.

The cooling system comprises a water block, a TEC1-12706 Peltier module, a heat sink, and a cooling fan. The hot side of the Peltier device is coupled to the heat sink and fan for heat dissipation, while the cold side is attached to an aluminum water block. Water circulating through the block is cooled and then pumped to the thermode, which incorporates a

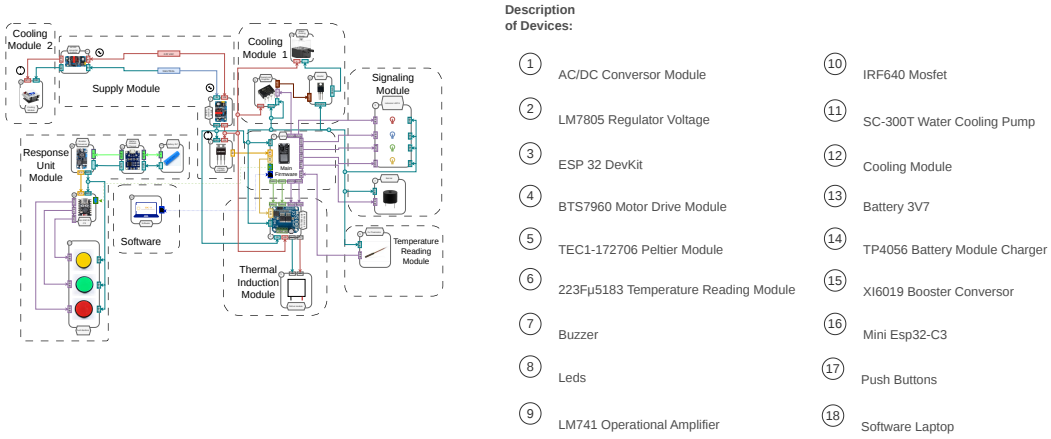


Fig. 2. Electrical and communication diagram of the X-QST system.

similar aluminum water block for heat exchange. After passing through the thermode, the water returns to the reservoir, forming a closed-loop cooling circuit that operates continuously throughout the QST procedure.

The purpose of this module is to improve heat dissipation from the hot side of the thermode Peltier device, thereby increasing cooling efficiency and thermal stability during temperature stimulation. Efficient heat removal minimizes thermal asymmetry between the hot and cold surfaces of the Peltier module, improving the accuracy of thermal stimulation. The cooling fan and the Peltier device are powered by an independent 12 V supply.

4) Response Unit Module: The response unit, already present in the previous version of the X-QST, was updated to support both QST methods. While the previous design included a single response button for MLI, the current version incorporates two additional buttons labeled ”YES” and ”NO” for the MLE, allowing participants to indicate whether the thermal stimulus was perceived.

The main improvement in this module is the replacement of the wired interface with Bluetooth communication. The response unit is based on a mini ESP32-C3 microcontroller, which communicates wirelessly with the main ESP32 whenever a response button is pressed. The module is powered by a rechargeable 3.7 V lithium battery managed by a TP4056 charging circuit and a boost converter that provides the required 5 V supply to the microcontroller.

The response unit also includes four status LEDs indicating power, Bluetooth connection, the MLI operating mode, and the MLE operating mode. The visual and acoustic signaling circuits were maintained from the previous version, as they had already been implemented and described in detail elsewhere [32].

5) Thermode Structure: The thermode was completely redesigned to improve the thermal performance of the X-QST system. The previous version consisted of a 191 $\times$ 234 $\times$ 116 mm enclosure containing a $40 \times 20 \times 40$ mm aluminum block that served as the stimulation surface. A TEC1-12706 Peltier device was attached to the opposite side of the block, while the generated heat was dissipated through a heat sink and cooling fan [10].

In the proposed design, the aluminum block is no longer used as the stimulation interface. Instead, it acts as a water-cooled heat exchanger integrated with the cooling module described in Section II-B3. The hot side of the Peltier device is coupled to this aluminum block, whereas the cold side is attached to a 0.1 mm thick copper plate (40×40 mm), which is placed in direct contact with the participant’s skin. In addition, the thermode enclosure dimensions were reduced to 70 $\times$ 30 $\times$ 50 mm.

This redesign substantially reduces the thermal mass of the stimulation interface, since only the thin copper plate must be heated or cooled instead of an entire aluminum block. Consequently, the Peltier device reaches the desired temperature more rapidly, improving heat transfer, thermal stability, and the ability to maintain the heating and cooling rate of approximately 1°C/s required by the DFNS protocol. Furthermore, the reduced dimensions improve the ergonomics of the device, making it easier to position on different anatomical regions during QST examinations.

C. Firmware and Software

The firmware and software were updated to support the hardware modifications introduced in the new X-QST system. The firmware continued to be developed in C++ using the Arduino IDE for the ESP32 microcontroller, maintaining the software architecture adopted in the previous version [32].

The main firmware updates include the implementation of both the MLI and the MLE, allowing the system to perform different psychophysical testing protocols. In addition, the temperature acquisition routine was modified by replacing the previous I^2C communication with an analog input associated with the new contact thermistor. Bluetooth communication with the ESP32-C3 microcontroller was also implemented to support the wireless response unit.

The host software remained developed in Python 3.6.9 using a graphical interface designed in Qt Designer and implemented through the PyQt5 framework. Communication between the software and the embedded system continues to rely on the *websocket-client* library for transmitting examination parameters and receiving real-time data. The software was updated to support the new testing methods and the wireless response unit while preserving the overall architecture of the previous version.

III. Results

A. New Equipment

The updated X-QST system, incorporating the redesigned thermode, wireless response unit, water-cooling module, and contact temperature sensor, is shown in Fig. 3.

The redesigned thermode (Fig. 3(a)) is substantially smaller and lighter, exposing only a thin copper plate as the skin-contact surface, with heat dissipation transferred to the external water-cooling module. The response unit (b) adds two buttons for the MLE and replaces the wired link with Bluetooth, improving participant mobility. The cooling module (c) circulates cooled water through the thermode, increasing heat dissipation and thermal stability during stimulation. Finally, the concentrator box (d) houses the main circuitry in a redesigned acrylic enclosure with the pump relocated to a separate housing; these changes serve mechanical organization and do not directly affect the thermal performance of the system.

B. Firmware and Software

The firmware architecture remained consistent with the previous version. The main update was the integration of the wireless response unit through the ESP32-C3: depending on the selected method, the main ESP32 enables the MLI or MLE buttons over Bluetooth and the participant's responses are transmitted wirelessly, allowing the test sequence to proceed automatically. The communication flow is illustrated in Fig. 4.

The host software required only minor modifications to support the new functionality. As shown in Fig. 5, two additional controls were incorporated to allow the operator to select either the MLI or the MLE protocol, while the remaining interface and communication structure were preserved.

C. Thermal characterization

The thermal behavior of the redesigned thermode was evaluated from bench runs recorded directly by the contact thermistor at a sampling rate of approximately 1 Hz, all starting from the $32\,^{\circ}$C adaptation baseline. Two cooling and two heating runs are reported. During the adaptation phase the setpoint is held at the $32\,^{\circ}$C baseline (gains $K_p = 3.5$, $K_i = 0.15$, $K_d = 1.0$); during stimulation, the setpoint is advanced at the target rate of $1\,^{\circ}$C/s and an auxiliary rate loop ($K_p = 5.0$, $K_i = 0.5$, $K_d = 1.0$) keeps the ramp close to this value. The gains were tuned empirically; the near-linear ramps and the baseline hold within about $\pm 1\,^{\circ}$C reported in Fig. 6 provide experimental evidence of stable regulation, although a formal stability analysis was not performed.

For each run, the ramp rate was obtained as the slope of a linear regression of temperature versus time over the range of interest around the baseline (32 to $12\,^{\circ}$C for cooling and 32 to $42\,^{\circ}$C for heating). The coefficient of determination (R^2) of this fit is used as a measure of ramp linearity, that is, how close the rate is to a constant value; an R^2 near 1.0 indicates a straight, constant-rate ramp, which is the behavior required for reproducible threshold estimation. Three aspects were assessed: the ramp rate (which should approach the $1\,^{\circ}$C/s recommended by the DFNS protocol), the ramp linearity, and the agreement between repeated runs as a first indication of repeatability.

Figure 6 summarizes the results. Figure 6(a) overlays the two cooling runs: both produced a ramp of $-0.76\,^{\circ}$C/s with $R^2 \approx 1.00$ and differed from each other by less than 1%, so the two curves are almost indistinguishable. Figure 6(b) shows the two heating runs, which followed closely matching slopes and gave average rates of 0.87 and $0.84\,^{\circ}$C/s over the full excursion (32 to $48\,^{\circ}$C), that is, agreement within about 2% with $R^2 \geq 0.99$. Near the baseline the heating rates were slightly more variable (0.78–$0.89\,^{\circ}$C/s over 32–$42\,^{\circ}$C), as one run started somewhat more slowly before settling onto the common slope. For cooling, the rate decreased toward the $0\,^{\circ}$C limit (to about $0.25\,^{\circ}$C/s over the full excursion), as expected from the reduced heat-pumping capacity of the Peltier device near its extremes. The thermode was held near the $32\,^{\circ}$C baseline within about $\pm 1\,^{\circ}$C, and the firmware kept the temperature inside the 0–$50\,^{\circ}$C range throughout.

These results are favorable in the region that matters most for QST. The cooling ramps were close to the $1\,^{\circ}$C/s DFNS target, highly linear ($R^2 \approx 1.00$), and nearly identical between runs (within 1%), while the heating ramps agreed to within about 2% over the full excursion ($R^2 \geq 0.99$), indicating good short-term repeatability in both directions, with the near-baseline heating segment showing slightly larger variability. These values were obtained from a small number of runs: a systematic repeatability study covering both directions, together with overshoot, settling-time, and hysteresis analyses, is still required to fully establish the metrological performance and is left for future work.

IV. Discussion

The modifications implemented in the updated X-QST represent a technical evolution over the previous prototype, both in hardware design and in compliance with international QST protocols. The copper stimulation surface and the Fμ contact thermistor [13], [17] enabled thermal ramp rates of approximately $1\,^{\circ}$C/s, consistent with DFNS recommendations and of the same order of magnitude as the rates targeted by commercial systems [5], [19], [20], whereas the previous prototype exhibited rates from $1\,^{\circ}$C/s to $2\,^{\circ}$C/s. The thermal characterization (Fig. 6) confirmed near-linear ramps of the order of $1\,^{\circ}$C/s in both directions around the baseline, with good short-term repeatability: the two cooling runs agreed to within 1% and the two heating runs to within about 2%

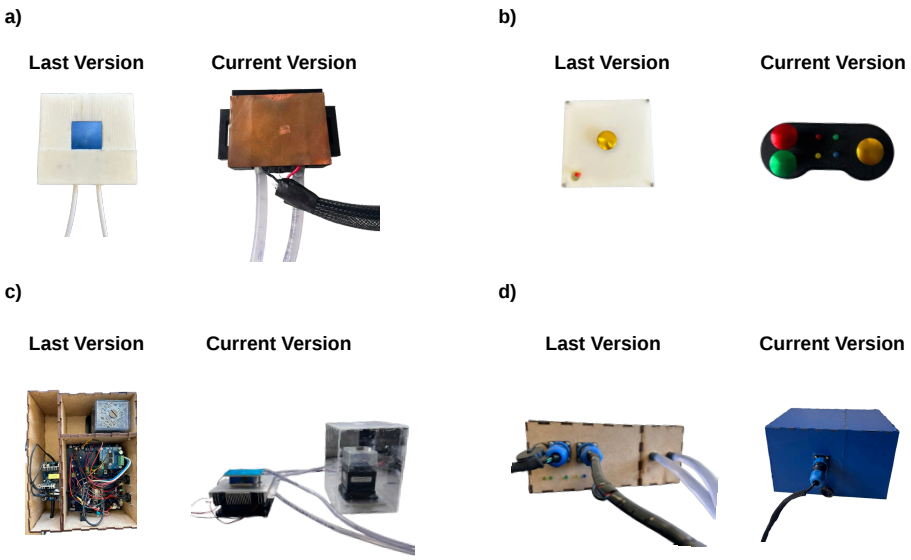


Fig. 3. Previous and current X-QST system: (a) thermode, (b) response unit, (c) cooling module, and (d) concentrator box.

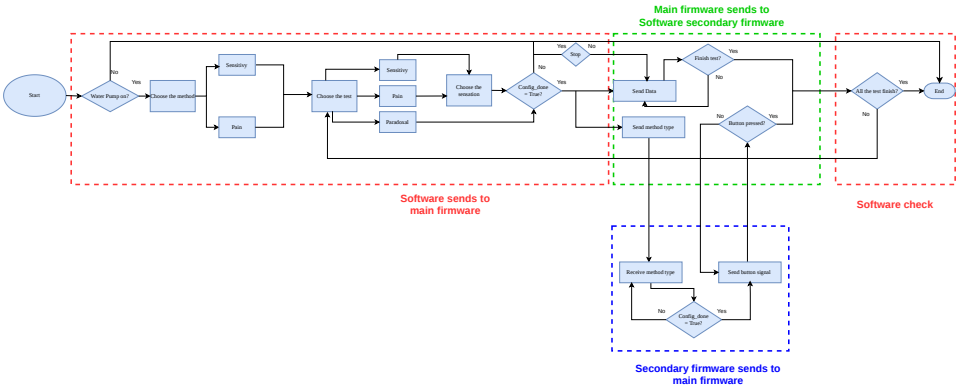


Fig. 4. Communication flow of the X-QST system.

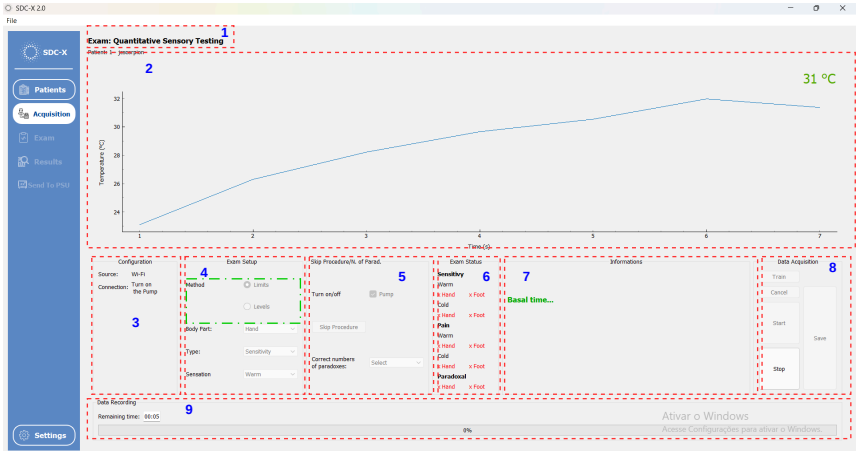


Fig. 5. Acquisition interface of the updated X-QST software.

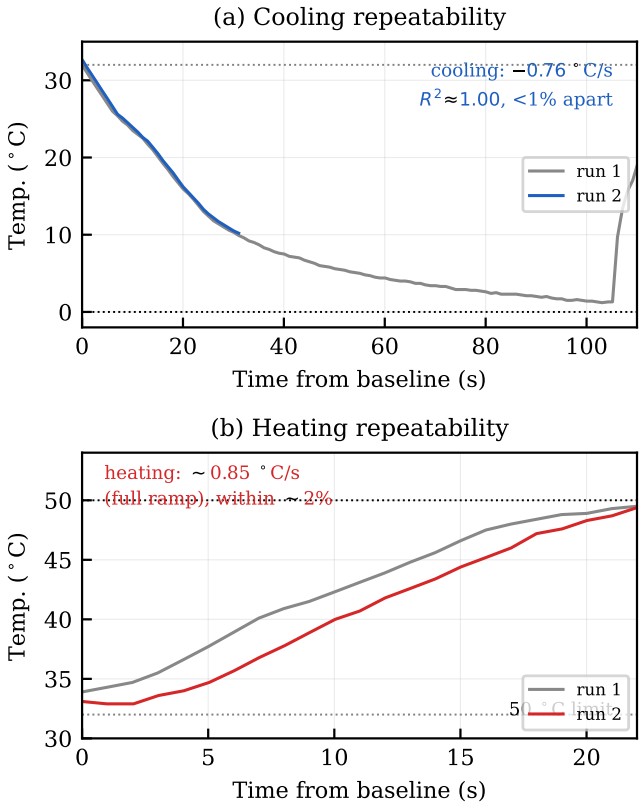


Fig. 6. Thermode temperature recorded from the contact thermistor. (a) Two cooling runs overlaid, showing near-identical ramps ($-0.76\,^{\circ}$C/s, $R^2 \approx 1.00$, agreement within 1%); (b) two heating runs with closely matching slopes (0.87 and $0.84\,^{\circ}$C/s over the full excursion, within about 2%).

over the full ramp. A larger repeatability study across different conditions remains pending.

The redesigned architecture also improved the usability of the system. The reduced thermode dimensions and the introduction of a Bluetooth-based wireless response unit made the device more compact and ergonomic without changing the examination protocol. Furthermore, the incorporation of the Method of Levels (MLE) extends the experimental capabilities of the platform, since this psychophysical procedure is less affected by participant reaction time and provides higher repeatability for thermal detection thresholds than the MLI [14]–[16].

Compared with representative QST systems, the proposed X-QST incorporates the core functionalities found in commercial devices, including bidirectional thermal stimulation, closed-loop temperature control, and automated execution of DFNS-compatible protocols. At the same time, it provides features that are generally unavailable in proprietary systems, such as an open hardware and software architecture, wireless communication, implementation of custom psychophysical protocols, and low-cost development. Compared with the portable NerveCheck system [22], which was primarily designed for clinical screening, the proposed platform offers greater flexibility for research applications through its customizable hardware and software.

A further distinguishing feature is the integration of the X-QST into the SDC-X platform rather than functioning as a standalone device, allowing QST results to be analyzed together with autonomic, pupillometry, sudomotor, and nerve conduction assessments—a multimodal approach uncommon in commercial and portable QST devices, which are generally limited to thermal sensory assessment alone. Overall, these advances establish a low-cost, open, and flexible platform with potential for future clinical research, contingent upon further validation.

Although the proposed system demonstrated improvements in thermal performance and protocol compatibility, additional clinical validation involving larger cohorts of participants with different stages of diabetic peripheral neuropathy is still required. Furthermore, direct comparisons with commercial QST systems under identical experimental conditions will be important to further establish the clinical performance of the proposed platform.

Table II summarizes the key characteristics of representative commercial and low-cost QST systems, highlighting only the functionalities implemented in the proposed X-QST.

V. Conclusion

This work presented an updated version of the X-QST module, integrated into the SDC-X platform, for quantitative thermal sensory assessment. The updated system incorporates a contact temperature sensor for direct monitoring at the thermode, a liquid-cooling module for heat dissipation during Peltier operation, a redesigned low-thermal-mass thermode supporting DFNS-compatible temperature ramps, a wireless response unit that reduces cabling during the examination, and firmware and software supporting both the MLI and the MLE.

Bench characterization showed near-linear heating and cooling ramps close to $1\,^{\circ}$C/s around the baseline, with the two cooling runs agreeing to within 1%. Overall, these updates consolidate the X-QST as a more modular and open research platform within the SDC-X ecosystem, whose clinical and metrological maturity still depends on future work: a full repeatability analysis, clinical validation in larger participant cohorts, and direct performance comparisons with commercial QST systems.

Acknowledgment

The authors would like to thank the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) for the financial support provided for this research under Finance Code 001.

References

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025.

[2] M. G. Savelieff, M. A. Elafros, V. Viswanathan, T. S. Jensen, D. L. Bennett, and E. L. Feldman, "The global and regional burden of diabetic peripheral neuropathy," *Nature Reviews Neurology*, vol. 21, no. 1, pp. 17–31, Jan. 2025, doi: 10.1038/s41582-024-01041-y.

TABLE II
COMPARISON BETWEEN REPRESENTATIVE QST SYSTEMS.

Characteristic	Medoc TSA-II	Somedic MSA	NerveCheck	Proposed X-QST
Peltier-based thermode	✓	✓	✓	✓
Heating and cooling stimulation	✓	✓	✓	✓
Closed-loop temperature control	✓	✓	✓	✓
DFNS-compatible protocols	✓	✓	Partial	✓
MLI	✓	✓	✓	✓
MLE	✓	✓	–	✓
Integrated graphical interface	✓	✓	Limited	✓
Automatic data storage	✓	✓	–	✓
Wireless response unit	–	–	–	✓
Custom protocol implementation	Limited	Limited	–	✓
Open hardware/software architecture	–	–	–	✓
Low-cost implementation	–	–	✓	✓

[3] E. L. Feldman, B. C. Callaghan, R. Pop-Busui, D. W. Zochodne, D. E. Wright, D. L. Bennett, V. Bril, J. W. Russell, and V. Viswanathan, "Diabetic neuropathy," *Nature Reviews Disease Primers*, vol. 5, no. 1, Art. no. 42, Jun. 2019, doi: 10.1038/s41572-019-0097-9.

[4] R. Pop-Busui, A. J. M. Boulton, E. L. Feldman, V. Bril, R. Freeman, R. A. Malik, V. Sosenko, and D. Ziegler, "Diabetic neuropathy: A position statement by the American Diabetes Association," *Diabetes Care*, vol. 40, no. 1, pp. 136–154, Jan. 2017.

[5] R. Rolke, W. Magerl, K. A. Campbell, C. Schalber, S. Caspari, F. Birklein, and R.-D. Treede, "Quantitative sensory testing: A comprehensive protocol for clinical trials," *Eur. J. Pain*, vol. 10, no. 1, pp. 77–88, 2006.

[6] T. G. Terra, "Software de gerenciamento, análise e detecção precoce da disfunção autonômica em indivíduos com Diabetes Mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2020.

[7] M. V. B. Franco, "Sistema remoto distribuído para avaliação da função autonômica em indivíduos com Diabetes Mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2021.

[8] M. Ücker, "Desenvolvimento de sistema point-of-care de pupilometria dinâmica aplicado no screening da Neuropatia Autonômica Diabética," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2020.

[9] K. K. Nakamura, "Sistema para avaliação da função sudomotora aplicado na detecção precoce do Diabetes Mellitus tipo 2," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2022.

[10] J. C. Scorpion, "Sistema de medição de limiares térmicos para integração em plataforma de saúde ubíqua de rastreamento de complicações do Diabetes Mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2024.

[11] J. L. P. S. Paula, "Sistema ubíquo de aquisição de sinais de condução nervosa aplicado ao monitoramento de complicações do Diabetes Mellitus," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2026.

[12] J. C. Scorpion, J. L. P. Paula, L. Umeoka, and J. L. B. Marques, "Concise analysis: Utilization of quantitative sensory testing for the diagnosis of peripheral diabetic neuropathy," in *Proc. Simp. Controle Processos Industriais (SCPI)*, Paranavaí, PR, Brazil, 2023.

[13] SEMITEC Corp., *Fμ Thermistor: Micro Thin Film Thermistor Sensor Probe, 223Fμ5183 / 223Fμ3122*, Datasheet, SEMITEC Corp.

[14] D. Yarnitsky and E. Sprecher, "Thermal testing: Normative data and repeatability for various test algorithms," *J. Neurol. Sci.*, vol. 125, no. 1, pp. 39–45, Aug. 1994.

[15] M. A. Kemler, J. P. Reulen, M. van Kleef, G. A. Barendse, F. A. van den Wildenberg, and F. Spaans, "Thermal thresholds in complex regional pain syndrome type I: Sensitivity and repeatability of the methods of limits and levels," *Clin. Neurophysiol.*, vol. 111, no. 9, pp. 1561–1568, Sep. 2000.

[16] J. P. H. Reulen, M. D. I. Lansbergen, E. Verstraete, and F. Spaans, "Comparison of thermal threshold tests to assess small nerve fiber function: Limits vs. levels," *Clin. Neurophysiol.*, vol. 114, no. 3, pp. 556–563, Mar. 2003.

[17] W. D. Callister Jr. and D. G. Rethwisch, *Materials Science and Engineering: An Introduction*, 10th ed. Hoboken, NJ, USA: Wiley, 2020.

[18] R. Rolke, W. Magerl, K. A. Campbell, C. Schalber, S. Caspari, F. Birklein, and R.-D. Treede, "Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Standardized protocol and reference values," *Pain*, vol. 123, no. 3, pp. 231–243, 2006.

[19] Medoc Ltd., *Pathway System User Manual*. Ramat Yishai, Israel: Medoc Ltd.

[20] Somedic SenseLab AB, *MSA Thermal Stimulator NG User Manual*, Version 6.4, Sweden, 2022.

[21] Somedic SenseLab AB, *MSA Thermal Stimulator NG Brochure*, Sweden, 2024.

[22] G. Ponirakis, I. N. Petropoulos, U. Alam, R. A. Ferdousi, A. Marshall, A. Azmi, H. Fadavi, A. Asghar, A. J. M. Boulton, and R. A. Malik, "NerveCheck: An inexpensive quantitative sensory testing device for patients with diabetic neuropathy," *Diabetes Res. Clin. Pract.*, vol. 113, pp. 101–107, 2016.

[23] M. Al-Khafaji and V. Otti, "Early diagnosis of diabetic neuropathy: A review of current diagnostic tests," *Cureus*, vol. 17, no. 11, Art. no. e97880, Nov. 2025, doi: 10.7759/cureus.97880.

[24] A. C. Themistocleous, J. D. Ramirez, J. Serra, and D. L. H. Bennett, "The clinical approach to small fibre neuropathy and painful channelopathy," *Pract. Neurol.*, vol. 14, no. 6, pp. 368–379, Dec. 2014.

[25] P. R. J. Vas and M. E. Edmonds, "Early recognition of diabetic peripheral neuropathy and the need for one-stop microvascular assessment," *Lancet Diabetes Endocrinol.*, vol. 4, no. 9, pp. 723–725, Sep. 2016.

[26] A. González-Duarte, M. Lem-Carrillo, and L. Guerrero-Torres, "Normative values of quantitative sensory testing in Hispanic Latino population," *Brain Behav.*, vol. 6, no. 7, Art. no. e00466, Jul. 2016, doi: 10.1002/brb3.466.

[27] Medoc Ltd., *TSA-II Operating Manual*. Ramat Yishai, Israel: Medoc Ltd.

[28] T. N. Cornsweet, "The staircase method in psychophysics," *Amer. J. Psychol.*, vol. 75, no. 3, pp. 485–491, Sep. 1962, doi: 10.2307/1419876.

[29] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 467–477, 1971.

[30] Hebei I.T. (Shanghai) Co., Ltd., *Thermoelectric Cooler*, Rev. 2.03, Datasheet.

[31] B. B. Ong and N. Milne, "Injury, fatal and nonfatal: Burns and scalds," in *Encyclopedia of Forensic and Legal Medicine*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2016.

[32] J. C. Scorpion, J. L. P. S. Paula, and J. L. B. Marques, "Prototype Development for Thermal Threshold Tracking in Peripheral Neuropathy Identification in Diabetic Patients Using Quantitative Sensory Testing," in *XXIX Brazilian Congress on Biomedical Engineering – Volume 1: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing*, A. B. Soares, G. C. Cardoso, and R. F. Leoni, Eds., IFMBE Proceedings, vol. 125. Cham, Switzerland: Springer, 2025, pp. 551–560, doi: 10.1007/978-3-031-93646-3_54.

Towards a Low-Cost System for Acquisition and Classification of EMG Signals

1st Atahualpa B. Lelis
Computing Institute
Universidade Federal de Alagoas
Maceió, Brazil
abl2@ic.ufal.br

2nd Gabriel L. B. Germano
Computing Institute
Universidade Federal de Alagoas
Maceió, Brazil
glbg@ic.ufal.br

3rd Thiago D. Cordeiro
Computing Institute
Universidade Federal de Alagoas
Maceió, Brazil
thiago@ic.ufal.br

***Abstract*—Electromyography (EMG) is a non-invasive technique widely used to evaluate muscle electrical activity, supporting clinical assessment, rehabilitation, and human–machine interaction studies. Despite its relevance, the acquisition hardware required for reliable EMG monitoring is often cost-prohibitive for hospitals, clinics, and research laboratories with constrained budgets. This paper presents the design and validation of a low-cost, portable, single-channel testbed for surface EMG acquisition and initial signal classification. The proposed system integrates surface electrodes, an instrumentation amplifier, active analog filtering stages, and a dedicated signal-conditioning circuit connected to a microcontroller for sampling and communication. The system transmits acquired signals to a server for storage, organization, and downstream analysis. A proof-of-concept evaluation used hand-related muscle activations (e.g., fist and grasp), and an unsupervised segmentation approach based on K-Means clustering separated resting and contraction intervals in the absence of precise temporal labels. The results demonstrate the feasibility of an end-to-end pipeline—from acquisition and conditioning to data management and machine-learning-based processing—indicating the platform's potential applicability in physical therapy, sports training, and biofeedback scenarios.**

***Index Terms*—Electromyography (EMG), Low-cost system, Acquisition channel, Machine learning algorithms.**

I. Introduction

Electromyography (EMG) is a technique for evaluating and recording the electrical activity of skeletal muscles, with established applications in medicine, physical therapy, biomedical engineering, and prosthetic device control. By detecting bioelectrical signals from muscle contraction, EMG supports clinical diagnoses, physiological studies, and the development of rehabilitation systems [1].

The non-invasive version of the technique, known as surface electromyography, uses electrodes placed on the skin to capture electrical signals from muscles. This approach has gained prominence for its practicality, safety, and clinical applicability, as well as for enabling functional tests in dynamic situations and providing greater patient comfort. Electromyographic signals can be used for various purposes, such as assessing muscle fatigue, controlling prostheses, biofeedback, ergonomics, and monitoring rehabilitation [2]. The signals obtained represent the sum of action potentials from active motor units and are influenced by factors such as electrode placement, muscle morphology, and tissue electrical characteristics [3].

Despite EMG's advantages, the high cost of commercially available equipment and reliance on imported technology remain significant barriers to its widespread use in Brazil. Many currently available devices are aimed at cutting-edge research, making them economically unfeasible for clinics, public hospitals, and educational institutions with limited resources. The growing demand for accessible technological solutions is driving the development of low-cost, portable, and customizable EMG acquisition systems. In parallel, 3D printing and open-source robotic prosthesis models have gained traction as a promising alternative for producing personalized, low-cost devices. Technologies such as Fused Deposition Modeling (FDM) enable the fabrication of mechanical components tailored to individual needs, contributing to the democratization of prosthetic access. In this context, the InMoov project, conceived by [6], stands out as one of the first robotic hand models developed entirely with open-source code and intended for 3D printing. This project offers advantages such as modularity, replicability, and extensive documentation, making it appropriate for both educational purposes and for prototyping assistive devices.

This work proposes developing and validating a low-cost testbed for acquiring and classifying EMG signals. The proposal aims to integrate electronics, computing, embedded systems, and three-dimensional printing into a functional, easily replicable, and applicable kit for teaching and research. This proposal is expected to contribute to technological accessibility, interdisciplinary training, and the advancement of biomedical engineering.

The rest of this work is organized as follows: the second section presents an in-depth exploration of the background that informs the methodologies used in this research. The third section addresses the electronic component, including circuit development and the methodological steps required to create a multidisciplinary educational kit for evaluating EMG signals. The fourth section presents the initial results of a clustering process toward an intelligent EMG signal classification using machine learning (ML) algorithms. Finally, the concluding chapter presents the main insights and considerations from the research effort.

Identify applicable funding agency here. If none, delete this.

II. Theoretical Background

This section presents the basic concepts essential for understanding this work. This section introduces EMG signals and outlines the steps for constructing a low-cost EMG signal acquisition prototype.

A. Electromyography Signals

Electromyography is a technique that examines muscle function by detecting the electrical potential generated during contraction of specific muscle groups. These bioelectrical signals result from the flow of ions across muscle fiber membranes and are recorded by electrodes on the skin surface or inserted into the muscle [3].

The electromyographic signal allows the estimation of muscle strength, activation time, impulse conduction velocity, and motor coordination across different tasks. The application of electromyography extends from clinical contexts—such as the diagnosis of neuromuscular diseases, fatigue assessment, and biomechanical diagnosis- to engineering applications, including active prostheses and human-machine interfaces.

Beyond its diagnostic use, EMG signal analysis is important in rehabilitation and biofeedback. In therapeutic contexts, the signal is used to monitor muscle activation patterns and train patients to optimize the use of muscle groups affected by injuries or neurological dysfunctions [5].

Electromyographic signals are also strategic in the development of biomechatronic devices. According to [2], myoelectric signals captured on the skin can be processed in real time and used as inputs for prosthesis control, enabling more natural and responsive movements. This principle underlies myoelectric prosthetic systems that respond more accurately to the user's intention.

Another relevant aspect is the stochastic nature of the EMG signal even under controlled conditions. This variability requires robust methods for acquisition, amplification, filtering, and interpretation to ensure data reliability and applicability in clinical or technological contexts.

Therefore, understanding the fundamentals of electromyography is essential not only for correctly interpreting data obtained in research and clinical applications but also for developing integrated solutions that incorporate sensors, embedded systems, and assistive devices.

1) Surface Electromyography: Surface electrodes capture the potential difference generated by muscular electrical activity during contraction, with signal intensity depending on muscle fiber depth, inter-electrode distance, skin impedance, and motor unit geometry [5].

The nature of the EMG signal is complex: it is a low-amplitude wave (generally between 0.1 mV and 5 mV) with a wide frequency range, varying mostly between 10 Hz and 500 Hz. Because it is a low-amplitude signal that is easily contaminated by external noise, acquisition systems need adequate circuits for amplification, filtering, and electrical isolation. The integrity of the captured signal depends directly on electrode quality, cabling type, and the efficiency of the electronic system's signal-conditioning stages [2].

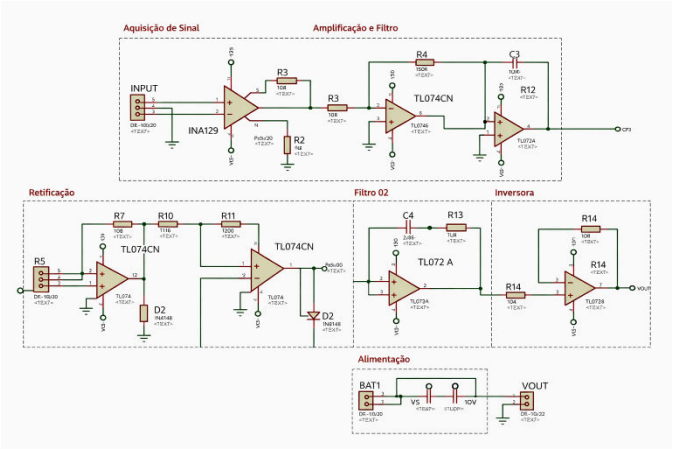


Fig. 1. Diagram of Prototype V03.

The EMG signal amplitude can range from 1 mV to 10 mV (peak-to-peak) or from 0 to 1.5 mV (rms) and contains components up to 10 kHz. The most significant energy in the EMG signal lies between 5 Hz and 500 Hz. Although harmonics exist up to 400 Hz, about 95% of the signal energy is contained within this range, with the harmonic between 400 Hz and 450 Hz being the last one of interest.

B. Electronic Instrumentation

Signal acquisition and conditioning in Prototype V03 rely on operational amplifiers from the TL07x family (TL072, TL074) and the INA129 instrumentation amplifier. These components were selected for their high input impedance (on the order of $10^{12}\,\Omega$, low input-referred noise, and adequate common-mode rejection ratio (CMRR $>$ 90 dB), which are essential to preserve the integrity of low-amplitude EMG signals (typically below 5 mV) [1], [2]. The TL074 integrates four amplifiers into a single package, reducing cost and board complexity compared with multiple discrete TL072 units, at the expense of a higher input offset voltage; this trade-off was mitigated by offset adjustment in the final gain stage.

Signal conditioning follows a classical chain: differential amplification via the INA129, a band-pass filter to isolate the frequency range of interest, active rectification, and low-pass smoothing to extract the signal envelope. Performing most of the noise rejection and envelope extraction in the analog domain reduces the computational load on the microcontroller. It limits the impact of electrode-skin artifacts, power-line interference, and adjacent-muscle crosstalk on the acquired signal.

III. EMG Signal Acquisition

The EMG signal acquisition prototype used in this work, Prototype V03, is a refined version that consolidates lessons learned from previous versions and implements a more complete, integrated approach to acquiring, processing, and using EMG signals (Figure 1).

Efficient PCB routing is fundamental in bioinstrumentation projects, as it reduces parasitic capacitance, trace inductance,

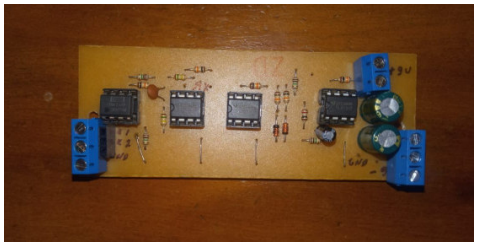


Fig. 2. Image of the final version of the Prototype V03.

and unwanted coupling that can compromise the fidelity of low-amplitude signals such as EMG. Accordingly, the layout of Prototype V03 was carefully planned to separate power, signal, and ground plane traces, minimize noise coupling between analog and digital signals [2], and reduce trace lengths between the amplifiers and the electrode input connectors, thereby limiting the system's vulnerability to ground loops, capacitive reference interference, and radio-frequency/AC power-line noise.

This component configuration was designed to optimize SNR, providing high CMRR and reducing the influence of electromagnetic noise and potential variations between electrodes, following the classic recommendation in biomedical instrumentation projects to separate power, signal, and ground-plane traces.

The Prototype V03 architecture presents well-defined stages for amplification, filtering, and offset adjustment, adequately conditioning the signals before digitization by the Arduino microcontroller. The use of DIP (Dual In-line Package) components also facilitates eventual replacements during the testing phase.

In the Prototype V03 final version, all essential signal-conditioning steps were consolidated onto a single board, including amplification, filtering, rectification, and smoothing. This integration significantly improved the quality of the signal supplied to the microcontroller and, consequently, the prosthesis's functional performance (Figure 2).

The Prototype V03 circuit comprises multiple carefully tuned stages: amplification, performed with low-noise amplifiers, raises the EMG signal to levels suitable for digitization without introducing unwanted distortion or saturation. Active high-pass and low-pass filters were implemented to attenuate content outside a band centered near 106 Hz, reducing external noise and physiological artifacts at the cost of narrowing the pass-band relative to the full theoretical EMG bandwidth. Rectification: the AC EMG signal was converted to a DC component to facilitate interpretation by the prosthesis control system. Smoothing: smoothing techniques, such as moving average filters or integrators, were used to ensure a stable signal free of rapid, non-physiological oscillations.

Integrating all signal-processing stages into a single board resulted in a more compact, less interference-susceptible, and more reliable system for practical applications. The acquired signal is then sent to the Arduino microcontroller, which receives it already amplified and rectified, performs the analog-to-digital conversion, and enables real-time visualization through the serial interface. One practical application is that the acquired EMG signals can be used immediately to control the prosthesis servomotors (Figure 3).

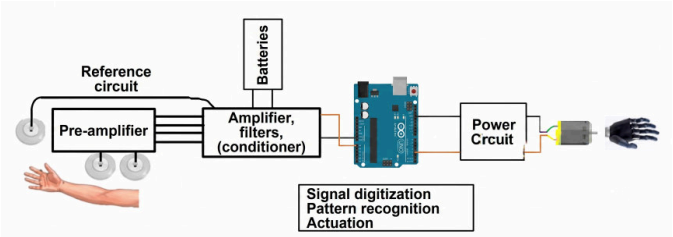


Fig. 3. Complete diagram of the acquisition and prosthesis movement system developed in this work.

The INA129 was used as a differential amplifier to amplify the myoelectric signal. The equation for the applied gain was $G = R_1/R_2$, with $R_1 = 1\text{M}\Omega$ and $R_2 = 1\text{k}\Omega$, with a total gain of 1000. After initial amplification, a non-inverting amplifier with moderate gain was implemented to compensate for potential filtering losses. This gain is $G = 1 + (R_3/R_1)$, with $R_3 = 150\text{k}\Omega$ and $R_1 = 1\text{M}\Omega$, resulting in a gain of approximately 1.15, sufficient to maintain the signal at operational levels without distortion.

Cascading a first-order high-pass and a first-order low-pass stage with an identical corner frequency f_c does not produce a flat pass-band spanning f_c to arbitrarily higher or lower frequencies; instead, it yields a first-order band-pass response peaking at f_c. Solving $|H(j\omega)| = 1/\sqrt{2}$ of the peak value for this cascade gives -3 dB corner frequencies at $\omega_c(\sqrt{2}-1)$ and $\omega_c(\sqrt{2}+1)$, which for $f_c \approx 106$ Hz correspond to approximately $f_{\text{low}} \approx 44\,\text{Hz}, and f_{\text{high}} \approx 256\,\text{Hz}$.

This is substantially narrower than the 10–500 Hz range discussed in Section II-A: signal content below approximately 44 Hz and above approximately 256 Hz is attenuated by more than 3 dB, which, combined with the effective sampling rate limitation discussed in Section V-A, further restricts the fraction of the theoretical EMG bandwidth that is faithfully captured by the current implementation. Assigning independent R/C values to the high-pass and low-pass stages, corner frequencies near 20 Hz and 450 Hz, respectively, would allow the filter to approximate the intended pass-band and is identified as a concrete design revision for future hardware iterations.

Signal rectification was performed with TL074 amplifiers and 1N4148 diodes, forming an active rectification circuit. This configuration eliminates negative half-cycles with minimal voltage loss, an essential characteristic for low-amplitude signals such as myoelectric signals.

This process was fundamental to extracting the signal envelope, which represents muscle intensity over time—a relevant data point for applications in prosthesis control, biofeedback, and clinical analysis.

After rectification, an active low-pass filter was used to smooth the signal and form the envelope. With the values

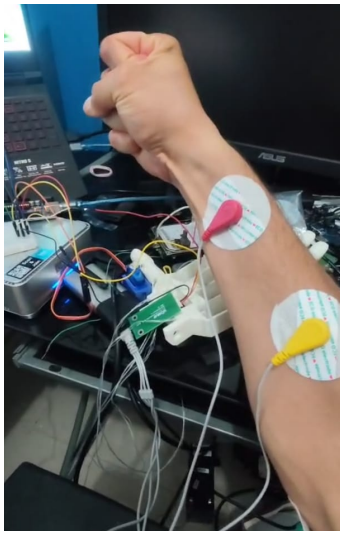


Fig. 4. Electrode position for EMG signal acquisition.

$R = 80.2\text{k}\Omega$ and $C = 1\mu\text{F}$, and applying the following formula, we have:

$$f_c = \frac{1}{2\pi RC} \Rightarrow f_c \approx 1{,}98\,\text{Hz} \quad (1)$$

This value allows the envelope to be obtained without rapid oscillations, facilitating visualization and clinical interpretation.

In the final stage, an inverting amplifier with gain control via a potentiometer was used, which allows the signal amplitude to be adjusted as needed. The formula used was $G = -R_{\text{POT}}R_{14}$, with $R_{\text{POT}} = 20\text{k}\Omega$ and $R_{14} = 1\text{k}\Omega$, resulting in a variable gain of up to -20 times, with the possibility of signal inversion as required by the output circuit.

IV. EMG Signal Acquisition System

Figure 4 illustrates the procedure for positioning the surface electrodes, strictly following the guidelines proposed by [1], which recommend cleaning the skin with isopropyl alcohol and aligning the electrodes along the muscle fibers. Correct skin preparation is critical to signal quality, as skin impedance and the presence of oil or residue can significantly compromise the system's SNR.

The choice of Ag/AgCl surface electrodes with conductive gel is justified by their low contact impedance and widespread use in EMG studies. To ensure accurate data acquisition, an inter-electrode distance of 20 mm was maintained. Furthermore, strategies were adopted to reduce external interference, such as using twisted and shielded cables, in accordance with the recommendations of [4].

After collecting the myoelectric signals described previously, the next step was to read and interpret them using a microcontroller. For this purpose, we developed C/C++ code in the Arduino IDE to capture and process real-time EMG signals.

Figure 5 illustrates a real example of an EMG signal captured by Prototype V03, with defined peaks and variations consistent with muscle contraction and rest cycles. This visualization helps identify noise artifacts, such as tremors or involuntary movements, and adjust the filters and activation thresholds in the embedded code.

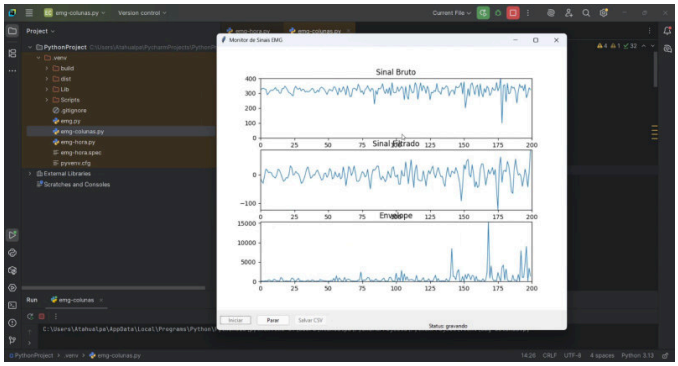


Fig. 5. Image illustrating the screen with the EMG signal capture process using Prototype V03.

Furthermore, this step reinforces hardware-software integration, bringing the system closer to functional biofeedback, where the user can observe the effects of their neuromuscular activity in real time. This approach is highlighted by [7] as promising in assisted rehabilitation and prosthetic control systems using myoelectric signals, especially in contexts that require user adaptation to the device.

The next development stage involved integrating Prototype V03 with the data acquisition system and storage software. Acquired EMG signals were automatically exported to a spreadsheet containing, for each sample, a timestamp and the corresponding captured voltage value, providing a solid foundation for subsequent statistical analysis and digital processing. This structured format enabled both qualitative and quantitative analysis of the signals, allowing verification of interference presence, filter quality, and the effectiveness of the amplification gain applied to the circuit. The interpretation of this signal followed the methodological recommendations described by [2], which guide the analysis of parameters such as dominant frequency, signal energy, and waveform integrity.

V. EMG Signals Identification

The development and implementation of Prototype V03 represented a significant advance in the proposed system's technical and functional trajectory. Integrating EMG acquisition hardware, electronic processing, and actuator control demonstrated the viability of a low-cost testbed for EMG signal acquisition.

The development of an automated system for data export and storage, with spreadsheet generation and graphical visualization of the signals, also contributed to validating the results by providing a consistent basis for subsequent analyses and inter-subject comparisons, in accordance with the methodologies adopted by [7].

To validate the acquired EMG signal, two hand movements were chosen: grasp and fist. The main goal is to evaluate

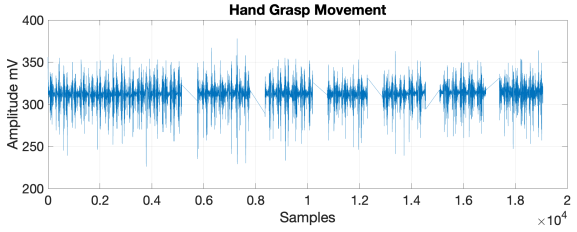


Fig. 6. Time-domain hand grasp movement signal acquired by Prototype V03.

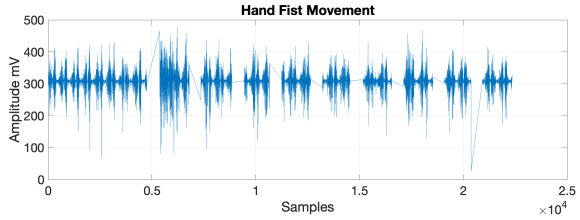


Fig. 7. Time-domain fist grasp movement signal acquired by Prototype V03.

whether ML algorithms can classify them or, at least, identify them. The choice of these two hand movements is explained as follows.

The first movement is grasping, which remains a challenging problem in prosthetics due to complex control requirements. For individuals with prosthetic hands, the ability to perform basic movements, such as grasping objects of various shapes, is vital. EMG signal analysis is the primary method for providing input to prosthetic devices to enable grasping [8]. The time-domain hand grasp movement signal acquired by Prototype V03 can be seen in Fig. 6. The signal consists of sequences of grasps followed by the hand's resting position, to facilitate identifying the moments when the hand is performing a grasp.

The second movement is the hand fist, which is important for evaluating hand force [9]. The time-domain hand fist movement signal acquired by Prototype V03 can be seen in Fig. 7. The signal consists of sequences of fist movements followed by the hand's resting position, to facilitate identifying moments when the hand is performing a fist movement.

Because no prior labels indicate the exact moments of muscle activation, the K-Means algorithm ($k = 2$) was used to segment each recording into Rest and Contraction states. The mean signal envelope was computed over sliding windows with 50% overlap; the window length was set to 150 ms, based on the effective sampling rate measured from the acquired timestamps (approximately 450 Hz), which is consistent with feature-extraction window lengths commonly reported in the surface EMG literature. Class assignment was performed through centroid analysis: the cluster with the lower numerical centroid value was labeled as *Rest*, and the cluster with the higher centroid value as *Contraction*.

This effective rate reflects the throughput of the ASCII-formatted serial logging pipeline rather than a fundamental limitation of the microcontroller's analog-to-digital converter, which can support substantially higher conversion rates. Serializing each sample as human-readable text (timestamp, raw, filtered, and envelope values) introduces per-sample transmission and formatting overhead that caps the achievable logging rate well below the theoretical maximum implied by the serial baud rate alone. The results presented in this section should be interpreted as demonstrating the feasibility of the acquisition-to-classification pipeline under the current logging throughput, rather than as evidence of full-bandwidth signal fidelity. Increasing the effective logging rate via a higher baud rate or a binary transmission protocol in place of ASCII formatting is a direct, low-cost path to closing this gap.

This procedure yielded 291 windows for the hand grasp recording and 303 windows for the fist/pinch recording, with a Rest/Contraction split of 70.4%/29.6% and 88.1%/11.9%, respectively — consistent with contractions occupying a smaller fraction of the total recording time than rest periods in both movements. Fig. 8 illustrates the distribution of these windows over time, showing a clear separation between low-energy (Rest) and high-energy (Contraction) periods for both movements.

To quantitatively assess the separability of the two classes, we trained a K-Nearest Neighbors (KNN, $k = 3$) classifier on the K-Means-derived labels for each movement. Because the feature windows overlap by 50%, a naively shuffled train/test split can result in near-duplicate windows appearing in both sets, artificially inflating accuracy. To avoid this, the dataset was evaluated using (i) a chronological hold-out, in which the last 20% of each recording (in time) was reserved for testing, and (ii) 5-fold *TimeSeriesSplit* cross-validation, in which every fold is contiguous in time, and no fold is trained on windows that occur after the ones it is tested on.

Table I summarizes, for each movement, the overall hold-out and cross-validated accuracy alongside the per-class precision, recall, and F1-score on the hold-out set. The hold-out accuracy reached 98.31% for the hand grasp movement and 100.00% for the fist/pinch movement, with cross-validated accuracies of 99.58% $\pm$ 0.83% and 99.60% $\pm$ 0.80%, respectively. Fig. 9 shows the corresponding confusion matrices.

It should be noted that the labels used to train and evaluate the KNN classifier were derived from the K-Means segmentation itself, rather than from an independently annotated ground truth (e.g., manual onset/offset annotation or a synchronized reference signal). Consequently, these results should be interpreted as a measure of the internal consistency and separability of the extracted feature, rather than as a validation against an external gold standard. Establishing such a gold standard is left as future work.

TABLE I
PER-CLASS CLASSIFICATION METRICS (HOLD-OUT SET)

Movement	Class	Precision	Recall	F1
Hand grasp	Rest	0.978	1.000	0.989
	Contraction	1.000	0.933	0.966
Fist/pinch	Rest	1.000	1.000	1.000
	Contraction	1.000	1.000	1.000

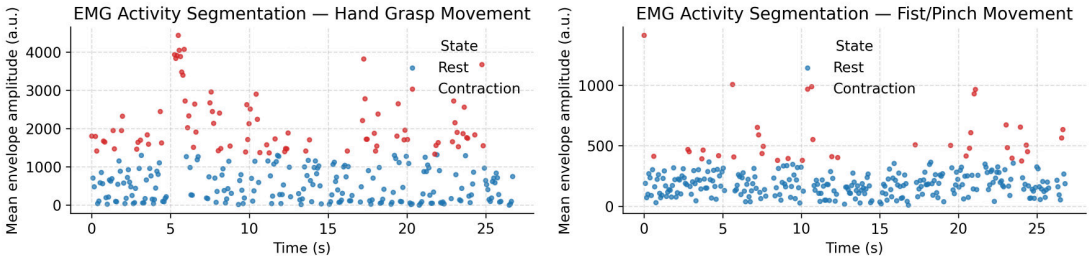


Fig. 8. EMG activity segmentation via K-Means clustering for the hand grasp (left) and fist/pinch (right) movements. Datapoints correspond to temporal windows, color-coded by cluster assignment (Rest vs. Contraction).

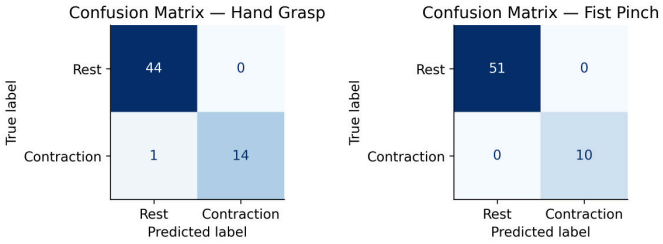


Fig. 9. Confusion matrices for the hand grasp (left) and fist/pinch (right) movements (hold-out set).

VI. Conclusion

This study successfully developed and validated Prototype V03, a low-cost surface electromyography acquisition system designed to democratize access to biomedical instrumentation. By consolidating the amplification, filtering, rectification, and smoothing stages into a single, layout-optimized printed circuit board, the proposed design reduces electromagnetic interference and supports signal conditioning suitable for educational and proof-of-concept human–machine interaction applications; achieving the acquisition bandwidth required for full clinical-grade fidelity will require increasing the effective data logging rate, as discussed below.

Experimental validation using hand grasp and fist movements demonstrated the feasibility of the end-to-end pipeline. Applying the K-Means algorithm showed that the acquired signals exhibit discriminative characteristics that enable automatic segmentation of muscle activity states (contraction versus rest) based on signal envelope energy. The successful clustering of these states indicates that the proposed hardware and processing approach can support myoelectric control use cases. Overall, this work delivers a functional, replicable, and accessible platform that bridges electronics, computing, and biomedical engineering, supporting experimentation and learning in resource-constrained environments. Future work will focus on increasing the effective data logging rate — through a higher serial baud rate or a binary transmission protocol — to meet the $\geq$1 kHz sampling requirement discussed in Section II-A, alongside real-time validation with participants, broader functional performance analysis under diverse conditions, and the integration of more advanced classification models to enable robust multi-gesture recognition. Furthermore, commercial equipment is being acquired in order to compare the performance of the proposed low-cost system.

Acknowledgment

The authors would like to thank CPG/PROPEP/UFAL.

References

[1] P. Konrad, *The ABC of EMG: A practical introduction to kinesiological electromyography*, 1st ed. Scottsdale: NORAXON INC. USA, 2005.

[2] M. Favretto *et al.*, "High density surface EMG system based on ADS1298-front end," *IEEE Latin America Trans.*, vol. 16, pp. 1616–1622, 2018.

[3] C. J. Luca, "Lectromyography," in *Encyclopedia of Medical Devices and Instrumentation*, Boston, EUA, 2006, pp. 98–109.

[4] C. Bischoff, A. Fuglsang-Fredriksen, L. Vendelbo, and A. Sumner, "Standards of instrumentation of EMG," *Clin. Biomech.*, 1999.

[5] Z. Lu *et al.*, "Advanced myoelectric control for robotic hand-assisted training: outcome from a stroke patient," *Front. Neurol.*, vol. 8, p. 107, 2017.

[6] G. Langevin, "Projeto InMoov," 2012. Acesso em: fev. 2016. [Online]. Disponível em: https://inmoov.fr

[7] D. S. Morais, "sEMG-controlled VR serious gaming system for neurorehabilitation and embodiment quantification," 2023. Acessado em: 17 jun. 2025. [Online]. Disponível em: http://hdl.handle.net/10451/64070

[8] M. Coskun et al., "Efficient deep neural network model for classification of grasp types using sEMG signals," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 9, pp. 4437–4450, 2022.

[9] A. Gailey, P. Artemiadis, and M. Santello, "Proof of concept of an online EMG-based decoding of hand postures and individual digit forces for prosthetic hand control," Frontiers in Neurology, vol. 8, p. 7, 2017.

Mapping Conventional and Emerging Sensors for Neonatal Physiological Monitoring: A Scoping Review of the Technological Transition and Validation Gaps

Gabriela Cianciarullo
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo, Brazil
gcianciarullo@fei.edu.br

Salvador Pinillos Gimenez
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo, Brazil
sgimenez@fei.edu.br

Carlos Eduardo Thomaz
Electrical Engineering Department
Centro Universitário FEI
São Bernardo do Campo, Brazil
cet@fei.edu.br

***Abstract*—Continuous physiological monitoring is central to neonatal intensive care, yet conventional NICU sensing, adhesive electrodes, contact pulse oximeters, and wired thermistors carry documented limitations, including skin injury, restricted mobility, and interference with family-centered developmental care. Over the past decade, flexible, wireless, and fully contactless alternatives have proliferated. Still, the resulting evidence remains fragmented across distinct technical communities, with no unified mapping of the conventional and emerging sensor ecosystem. Following the PRISMA-ScR guidelines and a Population, Concept, and Context (PCC) framework (Population, neonates in the NICU; Concept, sensors used in physiological monitoring; Context, the hospital NICU), searches in PubMed, Scopus, and IEEE Xplore retrieved 197 records, reduced to 166 after deduplication; 40 *primary* studies published between 2016 and 2026 were included after two-stage screening. Review articles were excluded from this count and used only as external triangulation, not as primary evidence. Emerging solutions predominate (26 of 40, 65%), optical transduction leads (15 of 40), and flexible or wearable form factors outnumber rigid designs (23 of 40 versus 9 of 40). Activity concentrates on respiratory rate (20 of 40) and heart rate (18 of 40), whereas electroencephalography and transcutaneous bilirubin each appear in only 1 of 40 studies and blood pressure only as a secondary channel (3 of 40). Under an explicit maturity rubric, no emerging paradigm reached multicenter, adequately powered, head-to-head validation, and 85% of studies omit regulatory status. We consolidate these findings into a cross-parameter gap matrix and a research-priority roadmap intended to guide future validation-focused work and enable direct cross-technology comparison.**

***Index Terms*—neonatal monitoring, sensors, neonatal intensive care unit, scoping review.**

I. Introduction

Neonates admitted to Neonatal Intensive Care Units (NICUs), particularly those born prematurely, depend on continuous multiparametric physiological monitoring to detect life-threatening events such as apnea, bradycardia, desaturation, and thermal instability [1]. According to the World Health Organization (WHO), approximately 15 million neonates are born prematurely each year worldwide, and prematurity-related complications remain the leading cause of death among children under five, making monitoring technology a direct determinant of population-level outcomes. Current practice relies predominantly on adhesive ECG electrodes, contact pulse oximeters, and wired thermistors connected to bedside monitors [2]. This architecture, although clinically effective, imposes well-characterized limitations: the structurally immature skin of the preterm neonate is highly vulnerable to adhesive-related injury [3]; cable tangle restricts spontaneous movement and complicates caregiving [4]; and the visual complexity of a wired setup creates psychological barriers to parental skin-to-skin contact and bonding, now recognized as critical to neurodevelopmental outcomes [4], [5]. In response, the past decade has witnessed a proliferation of alternative sensing proposals spanning flexible electronics, textile-integrated sensors, wireless wearables, and fully contactless approaches based on camera imaging and radar [2], [5].

This research, however, remains fragmented across distinct technical communities, biomedical engineering, materials science, computer vision, and clinical neonatology, and further segmented by parameter, with no comprehensive mapping that situates conventional and emerging solutions side by side and identifies which parameters remain disproportionately underserved.

This scoping review systematically maps conventional and emerging sensor technologies for neonatal physiological monitoring across nine parameter categories: peripheral oxygen saturation (SpO_2), electrocardiography and heart rate (ECG/HR), body temperature, respiratory rate, blood pressure, electroencephalography (EEG), transcutaneous bilirubin, blood glucose, and movement and body position. Specific objectives are to characterize transduction principles and form factors per parameter, quantify conventional versus emerging representation, assess validation and regulatory transparency using an explicit maturity rubric, and identify underrepresented parameters as priority targets. The synthesis is deliberately descriptive and quantitative-distributional, as befits a mapping objective [6], and culminates in a cross-parameter gap matrix and a research-

priority roadmap, the review's principal contribution, rather than an efficacy appraisal of any single intervention, unlike recent reviews focused on specific vital-sign subsets [2], [5].

This mapping is restricted to physiological monitoring. A complementary and increasingly prominent axis, behavioral monitoring of pain, clinical state, and neurodevelopmental status, falls outside the PCC Concept as defined here but shares much of the same sensing infrastructure (cameras, microphones, and physiological sensors). As documented in a co-authored companion work [7], AI-based automated neonatal pain assessment faces translational barriers structurally analogous to those identified here.

II. Methods

This section presents the review design, information sources, search strategy, study selection, and data extraction and appraisal procedures.

A. Study Design and Protocol

This scoping review followed the PRISMA-ScR guidelines [6]. The review question and eligibility criteria were structured with the PCC framework: Population, neonates, infants within the neonatal period (up to 28 days of postnatal age), including preterm infants of any gestational age, admitted to the NICU; Concept, sensors employed in physiological monitoring equipment; Context, the hospital NICU setting. No restrictions were imposed on study design, sensor maturity, or validation method, to capture the full spectrum from proof-of-concept proposals to clinically validated devices. The protocol was not prospectively registered; to support reproducibility, we report the a priori PCC criteria and the exact database queries in full below, and the complete per-study extraction dataset is openly available [8]. Consistent with the scoping methodology, we treated review articles (narrative, systematic, or scoping) as secondary sources: they were excluded from the 40 included primary studies, used only to contextualize the Introduction and Discussion and for backward snowballing, and never counted as primary evidence.

B. Information Sources and Search Strategy

Searches were conducted in PubMed, Scopus, and IEEE Xplore, with institutional access via Centro Universitário FEI, on June 16, 2026, to ensure complementary coverage of the biomedical, engineering, and multidisciplinary literature. The strategy combined three Boolean domains joined by AND operators: (1) neonatal population terms, (2) sensor/monitoring technology terms, and (3) NICU context terms. Iterative calibration was required because excessive AND-linked terms yielded near-zero retrieval while insufficient restriction produced unmanageable volumes; the final calibrated string was applied uniformly across databases using platform-appropriate syntax. In Scopus, we restricted the publication year to 2017 onward and the document type to articles and reviews. The exact strings are reported below.

PubMed:
```
("neonatal intensive care unit"[tiab] OR "NICU"[tiab] OR "neonatal ICU"[tiab] OR "Intensive Care Units, Neonatal"[MeSH]) AND (sensor[tiab] OR biosensor[tiab] OR "dry electrode"[tiab] OR "flexible sensor"[tiab] OR "wearable sensor"[tiab] OR "optical sensor"[tiab] OR "MEMS sensor"[tiab] OR "textile sensor"[tiab] OR "printed sensor"[tiab] OR "piezoelectric sensor"[tiab] OR "IMU"[tiab] OR "photoplethysmography"[tiab] OR "PPG sensor"[tiab] OR "infrared sensor"[tiab] OR "electrochemical sensor"[tiab]) AND ("vital signs"[tiab] OR "physiological monitoring"[tiab] OR "patient monitoring"[tiab] OR "bedside monitoring"[tiab] OR "cardiorespiratory monitoring"[tiab] OR "monitoring system"[tiab] OR "medical device"[tiab] OR "clinical monitoring"[tiab])
```

Scopus:
```
TITLE-ABS-KEY( ("neonatal intensive care unit" OR "NICU" OR "neonatal ICU") AND (sensor OR biosensor OR "dry electrode" OR "flexible sensor" OR "wearable sensor" OR "optical sensor" OR "MEMS" OR "textile sensor" OR "printed sensor" OR "piezoelectric sensor" OR "photoplethysmography" OR "PPG sensor" OR "infrared sensor" OR "electrochemical sensor" OR "IMU") AND ("vital signs" OR "physiological monitoring" OR "patient monitoring" OR "cardiorespiratory monitoring" OR "monitoring system" OR "medical device" OR "clinical monitoring" OR "bedside monitoring") ) AND PUBYEAR > 2016 AND DOCTYPE(ar OR re)
```

IEEE Xplore:
```
("neonatal intensive care unit" OR "NICU" OR "neonatal ICU") AND ("sensor" OR "biosensor" OR "dry electrode" OR "flexible sensor" OR "wearable sensor" OR "optical sensor" OR "MEMS" OR "photoplethysmography" OR "PPG sensor" OR "piezoelectric sensor" OR "textile sensor" OR "printed sensor" OR "IMU") AND ("vital signs" OR "physiological monitoring" OR "patient monitoring" OR "monitoring system" OR "medical device" OR "clinical monitoring")
```

The search strategy was iteratively refined to broaden its sensitivity: initial queries proved overly restrictive for the heterogeneous terminology spanning engineering and clinical databases, and were progressively expanded to improve recall of relevant studies.

C. Study Selection

Selection followed a two-stage process. Records were imported into Rayyan (Rayyan Systems Inc., Cambridge, MA, USA) for deduplication; the initial search retrieved 197 records (PubMed 24, Scopus 81, IEEE Xplore 92), reduced to 166 unique records. The first author performed screening and extraction, resolving uncertain cases at both stages in consultation with the senior authors. During title-and-abstract screening against the PCC criteria, 60 records were excluded (reviews without original sensor data, content unrelated to sensors, or context/population outside the inclusion criteria), leaving 106 reports to be retrieved. Of these, 54 could not be retrieved because the full text was not openly accessible under our institutional subscriptions, and the remaining 52 advanced to full-text assessment (per-reason counts in Fig. 1). Full-text assessment excluded a further 12 records for absence of extractable sensor-level data (including reviews and commentaries identified only at full text), population or context outside the neonatal-NICU criteria, or insufficient methodological detail to characterize sensor and validation attributes, yielding 40 studies for synthesis.

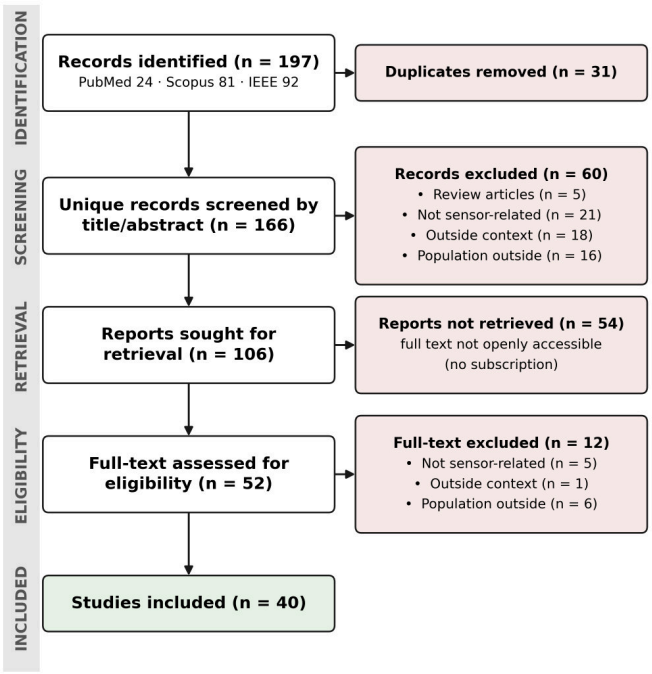


Fig. 1. PRISMA-ScR study selection flow diagram.

D. Data Extraction and Maturity Appraisal

For each of the 40 included studies, a predefined form captured: monitored parameter(s); sensor type and physical transduction principle; form factor (rigid, flexible, textile, wearable, implantable, or camera-based); validation type; population sample; reference standard; performance metrics; regulatory aspects; and author-acknowledged limitations. We counted studies addressing more than one parameter in each relevant category. Because PRISMA-ScR does not mandate formal risk-of-bias appraisal, we instead applied a design-appropriate maturity rubric, assigning each study its highest attained validation stage: Stage 0, theoretical/conceptual (no experimental data); Stage 1, bench, in vitro, phantom, or simulation; Stage 2, in vivo in non-target subjects (adult volunteers or animals); Stage 3, in vivo in neonates, typically single-center and small-sample; Stage 4, multicenter, adequately powered, head-to-head validation against the clinical reference standard. Studies using several approaches were assigned their highest stage to avoid double counting. The complete study-level table (population, sensor, reference standard, sample size, and metrics) is openly available [8]; Table I summarizes the parameter-level synthesis.

III. Results

This section reports study characteristics, technology distribution across parameters, validation status, reported metrics, and author-acknowledged limitations.

A. Overview of Included Studies

The 40 studies were published between 2016 and 2026, concentrated in 2017–2020 and 2022–2023 (Fig. 2), reflecting two waves of activity coinciding with the emergence of flexible epidermal electronics and, later, the post-pandemic acceleration of contactless monitoring. Most propose emerging solutions, experimental, prototype, or alternative to the current standard (26/40, 65%); 8/40 (20%) are hybrid adaptations of established commercial technology; and 6/40 (15%) evaluate the conventional standard directly.

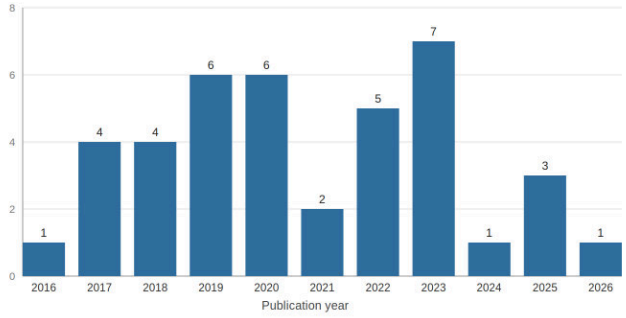


Fig. 2. Temporal distribution of the 40 included studies (2016–2026).

B. Distribution by Monitored Physiological Parameter

Fig. 3 presents the distribution across the parameter categories. Because studies addressing more than one parameter are counted in each, parameter occurrences exceed the number of studies; all percentages use the 40 studies as the denominator. Respiratory rate and ECG/HR together account for nearly half of all occurrences (20/40 and 18/40, respectively), reflecting their central role in detecting apnea, bradycardia, and acute cardiorespiratory events. Temperature (10/40), movement/body position (7/40), and SpO_2 (6/40) follow; blood glucose (4/40), blood pressure (3/40), crying/vocalization (2/40), and incubator-microenvironment variables (humidity and gas composition, 3/40) complete the distribution. In marked contrast, EEG and transcutaneous bilirubin each appear in only 1/40 studies, and blood pressure only as a secondary channel within multiparameter platforms, with no dedicated sensing principle proposed. This asymmetry is clinically significant: EEG is the primary tool for neonatal seizure detection and neuroprotection, and bilirubin measurement is critical for hyperbilirubinemia screening, yet both remain underrepresented relative to their clinical weight.

C. Conventional versus Emerging Technologies

Eleven of the 40 studies integrate three or more parameters into a single platform [9]–[19], generally using flexible wireless wearables that replace multiple cables and adhesive electrodes with a pair of soft, skin-conformable patches. Representative examples include the battery-free binodal epidermal systems of [11], later expanded into the commercial ANNE One platform [9], [19], [20], the e-textile NeoWear system [21], and platforms with textile piezoresistive sensors [12].

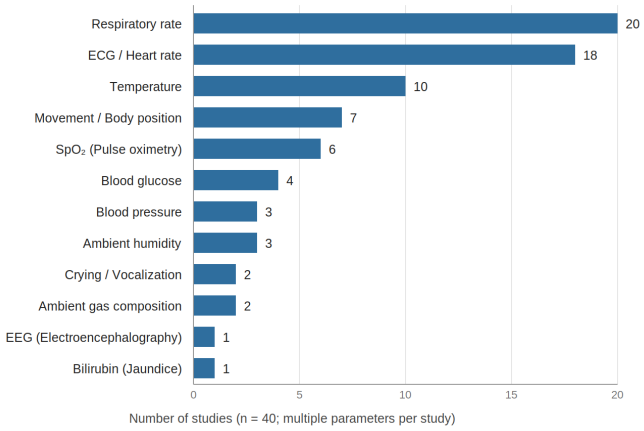


Fig. 3. Distribution of included studies by monitored physiological parameter.

D. Physical Principles and Sensor Forms

Optical sensors predominate (15/40), mainly contact or remote photoplethysmography, whose versatility across heart rate, respiratory rate, and SpO_2 explains this share. Thermal/infrared (9/40) and resistive/piezoresistive sensors (7/40) follow, concentrated in temperature/thermographic respiration and respiration/movement applications, respectively. Electrochemical sensors (6/40) are almost exclusively in glucose monitoring; electrical sensors (5/40) mainly in ECG platforms; and electromagnetic/radar sensors, the most recent frontier, account for 4/40. Regarding form factor, flexible (13/40) and wearable (10/40) sensors together (23/40) already outnumber rigid designs (9/40), while contactless camera (5/40) and radar (4/40) systems form a growing group (9/40). No implantable sensors were identified, consistent with a noninvasive orientation.

E. Technological Mapping by Parameter

1) ECG and heart rate: Beyond the multiparameter wearables already described, dedicated ECG/HR solutions include capacitive and textile electrodes embedded in a smart mattress [22], enabling passive acquisition without skin attachment; a handheld dry-electrode detector for delivery-room resuscitation [23]; and a forehead reflectance-photoplethysmography heart-rate monitor [24]. Among contactless approaches, camera-based imaging photoplethysmography extracts cardiac pulse from spatiotemporal variations in skin reflectance [25]–[27], while Doppler/FMCW radar detects chest-wall microdisplacements unaffected by light or clothing [28], [29]. The coexistence of these two contactless principles reflects a field still diversifying, with no single modality yet established as a preferred alternative to the adhesive-electrode standard.

2) Respiratory rate: Respiratory rate is the most investigated parameter (20/40), reflecting apnea-of-prematurity detection and the signal's accessibility to mechanical, thermal, optical, and electromagnetic transduction. Piezoresistive pressure-sensitive mats beneath the mattress detect thoracic displacement without skin contact [30]. Low-power PVDF-film pyroelectric sensors transduce thermal airflow variations and suit battery-powered wearables [31], [32]. Near-infrared spectroscopy, conventionally used for cerebral oxygenation, has been adapted to extract respiratory rate from the same signal without additional hardware [33]. Long-wave infrared thermal cameras detect the thermal plume of nasal/oral airflow, and camera-based methods track respiratory motion [34], [35], while Doppler/FMCW radar detects sub-millimeter chest-wall displacement, operating through incubator walls and bedding [36]. This breadth reflects intense investment but the absence of a single dominant alternative to conventional impedance pneumography.

3) SpO_2 and pulse oximetry: SpO_2 occupies a distinct position: rather than proposing novel principles, the literature validates the established pulse-oximetry standard in specific neonatal subpopulations [37], having largely converged on conventional contact PPG, with open questions clinical (poor perfusion, motion, pigmentation) rather than technological. When SpO_2 appears in innovation-oriented studies, it is almost exclusively an integrated channel within multiparameter flexible/wearable platforms [9], [11], [13], [16], [19], where innovation resides in system-level miniaturization and packaging rather than the transduction principle, an important distinction under the PCC framework. The relative absence of fundamentally novel SpO_2 principles, such as contactless remote-PPG oxygen estimation, is a noteworthy gap, likely reflecting the difficulty of extracting reliable saturation from non-contact optical signals under variable neonatal perfusion and pigmentation.

4) Temperature: Conventional thermistors in multiparameter incubator systems [14] represent the clinical standard: accurate yet subject to the mobility and contact limitations of wired sensing. Contactless infrared thermopile sensors integrated into flexible wearable patches [16] reduce contact burden while preserving the body-proximate architecture. At the most innovative tier, thermal-and-RGB camera fusion with deep learning enables continuous, contactless temperature estimation across the body surface, approaching contact thermometry accuracy with the added capability of spatial mapping [38], [39]. Computational demand and illumination sensitivity remain the main translational barriers, and none of these studies addresses regulatory pathway or deployment-scale validation.

5) Blood pressure: No study proposed a dedicated noninvasive sensing principle for neonatal blood pressure. It appeared in only three studies, exclusively as a secondary channel within multiparameter systems [9], [11], [17], despite clinical reliance on invasive arterial catheters or intermittent oscillometric cuffs. This is consistent with Senechal et al. [2], who similarly identified only isolated reports. Although cuffless approaches based on pulse transit time, PPG waveform analysis, and ballistocardiography have been extensively investigated in adults, their translation to neonates remains largely unexplored.

6) Blood glucose: The four glucose studies share a common approach: rather than novel neonatal-specific principles, all adopt commercial continuous glucose monitors (Medtronic,

Dexcom) using minimally invasive subcutaneous electrochemical biosensors with 1–5-minute wireless readings [40]–[43]. Recurrent limitations include reduced accuracy during therapeutic hypothermia, when peripheral vasoconstriction impairs interstitial glucose delivery precisely when monitoring is most needed [40]; interstitial–blood discrepancies during rapid glycemic change; and the absence of neonatal-specific calibration in platforms validated against adult physiology. This near-exclusive reliance on repurposed adult CGM technology reflects both a translational gap and an opportunity for neonatal-specific glucose sensing.

7) Electroencephalography: Only one study addressed neonatal EEG sensor technology, evaluating low-cost dry electrodes as an alternative to conventional wet electrodes [44]. Although the system achieved signal quality sufficient for seizure screening, elevated electrode–scalp impedance and increased susceptibility to motion and EMG artifacts remain barriers to replacing the clinical gold standard. Given the central role of EEG in seizure detection, neuroprotection, and neurodevelopmental assessment, this limited engineering activity contrasts sharply with its clinical importance.

8) Transcutaneous bilirubin: Neonatal hyperbilirubinemia affects 60–80% of neonates in the first week of life and, if inadequately monitored, carries risks of kernicterus and permanent neurodevelopmental damage, yet only one study addressed bilirubin sensing. That study describes a prototype wireless transcutaneous jaundice meter using optical reflectance spectroscopy [45], extending established commercial bilirubinometry (e.g., Dräger JM-105, Natus BiliChek) with wireless connectivity for continuous rather than episodic measurement, but reporting only proof-of-concept results without validation against serum bilirubin across the relevant concentration, pigmentation, or gestational-age range. Neither Senechal et al. [2] nor Williams et al. [5] identify bilirubin as an active frontier. This convergent absence flags this parameter as a genuine field-wide deficit.

9) Movement and body position: Movement and body position (7/40) extends beyond motor-activity detection to neurological assessment, pain quantification, and prevention of positional complications. Under-mattress ballistographic sensors detect micro-mechanical vibrations from movement, respiration, and cardiac activity without body contact, though they remain susceptible to environmental mechanical noise [46]; fiber-mat systems similarly enable continuous, unobtrusive movement quantification in preterm infants over long durations [47]. E-textile platforms with integrated inertial measurement units enable spatially resolved posture quantification while remaining skin-contact-free [21], with applications to objective assessment of general movements. Accelerometers, as dedicated channels within broader multiparameter wearables, are the most common implementation [9], [10], [12], [15], [19] and also enable motion-artifact detection. The absence of standardized movement metrics remains the primary barrier to comparative evaluation.

10) Incubator microenvironment (boundary of scope): A small set of studies targets the incubator microenvironment, humidity (three studies) and gas composition (two), overlapping across three distinct studies [14], [15], [48], rather than body-directed physiological signals, using conventional capacitive humidity, electrochemical oxygen, and resistive temperature sensors within microcontroller platforms with wireless transmission. Humidity control limits transepidermal water loss in extremely preterm skin, and gas monitoring maintains FiO_2 targets, since both hyperoxia and hypoxia carry risks of retinopathy of prematurity and bronchopulmonary dysplasia. These components are technologically mature; the engineering challenge is integration, miniaturization, and electromagnetic compatibility rather than transduction innovation.

F. Validation Status under the Maturity Rubric

Under the rubric of II-D, 26/40 studies (65%) reached Stage 3 (in vivo neonatal validation), though typically with small, single-center samples (Section III-H). Three reached Stage 2 (two adult-volunteer, one animal), seven Stage 1 (five bench/in vitro, two phantom/simulation), and four Stage 0 (theoretical), summing to 40. Critically, no study reached Stage 4: none reported multicenter, adequately powered, head-to-head validation against the clinical reference standard, the explicit threshold we adopt for translational readiness. Regulatory transparency is poor: 34/40 (85%) omit regulatory status; of the remaining six, five mention the FDA and one IEC/ISO, none in sufficient detail to assess deployment readiness. This profile reflects the field's predominantly proof-of-concept maturity and echoes, on a larger scale, the methodological-quality concerns of Senechal et al. [2], reinforcing the case for standardized reporting that discloses regulatory pathway alongside performance metrics.

G. Reported Performance Metrics

Reported metrics are markedly heterogeneous and largely incommensurable: accuracy, Pearson/intraclass correlation, mean absolute (percentage) error, RMSE, Bland–Altman bias and limits of agreement, sensitivity/specificity, F1/AUROC, and signal-to-noise ratio all appear, often within the same parameter category. Absent a standardized protocol specifying reference standard, metric, statistical method, and test conditions, the evidence cannot be synthesized into a coherent comparative ranking of competing technologies. A consensus validation framework analogous to ANSI/AAMI/ISO 81060-2 (noninvasive blood pressure) or ISO 80601-2-61 (pulse oximeters), but designed for neonatal physiology, is a concrete, high-priority need; the CONSORT-adapted checklist proposed by Senechal et al. [2] is a meaningful step toward it. Complete per-study metrics are openly available [8].

H. Limitations Reported by Primary Studies

The limitations acknowledged by primary authors are recurrent and structurally interconnected. Small sample size combined with single-center design is the most pervasive, restricting statistical power and generalizability; Senechal et al. [2] corroborate this at scale, reporting a median sample of only 10 patients. Motion-artifact susceptibility is the second

most common, particularly in contactless camera- and radar-based systems, and is most acute during clinical events (apnea, seizures, handling), when reliability matters most. Reduced accuracy under specific pathophysiological conditions, glucose sensors during therapeutic hypothermia, optical sensors affected by perfusion or pigmentation, creates an inverse relationship between clinical need and reliability.

These limitations map onto distinct translational dimensions that should not be conflated: feasibility (can the signal be acquired?), accuracy (does it agree with the reference?), safety (does it avoid harm, e.g., skin injury or thermal load?), and implementation (can it be deployed and sustained in routine care?). Caregiver and clinician burden, setup time, alarm handling, workflow disruption, and the parental-bonding benefit of contactless designs were rarely quantified, despite being decisive for adoption. Collectively, these point to a need for larger, multicenter, long-duration studies with standardized metrics and head-to-head comparisons as a prerequisite for evidence-based translation.

IV. Discussion

The mapping converges on a field in active technological transition, but the picture is uneven and qualified by important exceptions rather than uniform.

A. An Uneven Technological Transition

The clearest signal is a broad shift away from wired, adhesive-dependent, single-parameter architectures toward two paradigms: integrated flexible wireless multiparameter wearables (ECG/HR, SpO_2, temperature, respiratory rate, and, secondarily, blood pressure via soft chest/limb patches) and fully contactless sensing (camera PPG, infrared thermography, radar), the latter at an earlier stage but advancing rapidly. This converges with two independent contemporaneous reviews: Senechal et al. [2], covering 60 contactless/wireless studies, report respiratory rate (72%) and heart rate (51%) dominating; Williams et al. [5] document the proliferation of wireless/wearable proposals for heart rate and oxygenation. Convergence across three independent reviews with distinct databases suggests a genuine research trajectory rather than a sampling artifact. The transition is not uniform, however, and framing it as a simple "innovation versus standardization" dichotomy risks overstating coherence. Two exceptions matter: for SpO_2, the field has effectively converged on conventional pulse oximetry as a validated standard, so the corresponding studies characterize that standard rather than propose new principles; conversely, many "emerging" proposals remain at bench or theoretical stages (Stages 0–1) and are far from the maturity of the wearables dominating the ECG/HR and respiratory-rate literature. The transition is thus best described as parameter-dependent and concentrated in the most technically accessible signals.

B. Clinically Consequential Gaps and Priority Criteria

We designate a parameter a high-priority gap when it satisfies three explicit criteria: representation at or below $\sim$1/40 among emerging studies (or presence only as a secondary channel); well-established, high clinical relevance; and corroborated underrepresentation across at least two independent external reviews. Transcutaneous bilirubin (1/40; near-universal clinical relevance; absent as a frontier in both external reviews), neonatal EEG (1/40; central to seizure and neuroprotection monitoring), and noninvasive blood pressure (3/40, secondary channel only; no dedicated neonatal principle, corroborated by 1/60 in [2]) each meet all three. We triangulate these designations against reviews whose scope exceeds our date and language limits; a definitive priority ranking would require complementary searches in additional databases and gray literature, and we present the criteria so others can reproduce or contest them.

C. Regulatory Pathways and Clinical Adoption

Translation requires navigating regulatory pathways that the reviewed literature largely ignores. In the United States, most such devices would follow the FDA 510(k) route when a substantially equivalent predicate exists, the De Novo pathway for low-to-moderate-risk devices without a predicate, or premarket approval for higher-risk cases [49]; software-driven and AI-enabled functions additionally fall under Software-as-a-Medical-Device and evolving AI/ML guidance [50]. In the European Union, the Medical Device Regulation (2017/745) governs CE marking and has raised the clinical-evidence bar [51]. Beyond approval, adoption depends on prospective validation in the target population, interoperability with existing monitors and electronic records, reimbursement and health-technology-assessment evidence, and demonstrated reductions in caregiver burden. The near-total absence of engagement with these pathways is itself a marker of proof-of-concept maturity and a barrier that standardized reporting frameworks must address by requiring disclosure of the intended regulatory route alongside performance metrics.

D. A Research-Priority Roadmap

Table I consolidates the cross-parameter analysis into a gap matrix and a pragmatic roadmap, ranking parameters by representation, clinical need, and highest validation stage reached. Two priorities emerge. First, consolidation: for respiratory rate and ECG/HR, where many competing principles already reach Stage 3, the pressing need is head-to-head, multicenter validation to identify preferred modalities, not further principle proliferation. Second, filling structural gaps: EEG, transcutaneous bilirubin, and noninvasive blood pressure combine high clinical need with minimal engineering activity and warrant targeted development. Contactless temperature and remote-PPG SpO_2 occupy an intermediate position, feasible but requiring accuracy and safety validation before adoption.

E. The Physiological–Behavioral Frontier

The PCC framework restricts the Concept to physiological sensing, appropriately excluding a growing parallel domain: behavioral monitoring of pain, clinical state, and neurodevelopmental status, which draws on the same infrastructure

TABLE I
CROSS-PARAMETER GAP MATRIX AND RESEARCH-PRIORITY ROADMAP.

Parameter	n/40	Need	Stage	Priority
Respiratory rate	20	High	3	Consolidate
ECG / heart rate	18	High	3	Consolidate
Temperature	10	Med	3	Validate
Movement/position	7	Med	3	Standardize
SpO_2	6	High	3*	Validate
Blood glucose	4	Med–High	3	Calibrate
Blood pressure	3†	High	3	**Fill gap**
EEG	1	High	2	**Fill gap**
Transcut. bilirubin	1	High	0	**Fill gap**

Representation counts parameter occurrences across the 40 primary studies; "Stage" is the highest validation stage reached by any study for that parameter (Section II-D).
*Conventional pulse oximetry is an established standard.
†Present only as a secondary channel.

(cameras, IMUs, microphones, physiological sensors). As documented in a companion work [7], AI-based neonatal pain assessment, integrating facial coding, body-movement analysis, cry acoustics, and contextual physiological data, faces barriers mirroring those found here: single-center datasets, non-standardized protocols, absent shared benchmarks, and regulatory opacity. Multimodal platforms acquiring physiological and behavioral signals from a shared architecture are a natural extension of this work, with barriers that are less technological than organizational and regulatory.

F. Limitations of this Review

Several limitations qualify these findings. The search was restricted to three databases and English-language indexed literature, potentially underrepresenting specialized vocabularies (e.g., neonatal invasive blood pressure, better indexed in Embase, Cochrane, Engineering Village, or Google Scholar) and early-stage engineering work that first appears in conference proceedings, preprints, or gray literature—an acknowledged publication-bias risk for a fast-moving field, partially mitigated by triangulation with the broader-scope reviews of Senechal et al. [2] and Williams et al. [5], which corroborated rather than contradicted the gaps identified here. Reports not retrieved because the full text was not openly accessible may introduce selection bias if not randomly distributed across parameters or venues. The protocol was not prospectively registered, and screening and extraction were performed principally by the first author with senior-author adjudication rather than fully duplicate independent review. Finally, consistent with the mapping objective, the synthesis is descriptive and quantitative-distributional; formal bibliometric analysis (co-authorship and keyword co-occurrence networks, citation clustering) was beyond scope and is a natural next step.

V. CONCLUSION

This scoping review mapped 40 primary studies on conventional and emerging sensors for neonatal physiological monitoring across nine parameter categories (2016–2026). Emerging solutions predominate (26/40, 65%), flexible/wearable form factors outnumber rigid designs (23/40 vs. 9/40), and optical transduction leads (15/40), reflecting an uneven but real convergence toward integrated wireless multiparameter platforms and contactless sensing, corroborated by two contemporaneous reviews [2], [5]. Against this backdrop, EEG and transcutaneous bilirubin (1/40 each) and noninvasive blood pressure (3/40, secondary only) stand out as externally corroborated, clinically consequential priority gaps. No emerging paradigm has reached multicenter, adequately powered, head-to-head validation, and 85% of studies omit regulatory status, underscoring the field's proof-of-concept maturity. The consolidated gap matrix and roadmap (Table I), together with adoption of standardized validation and reporting frameworks [2] and explicit engagement with regulatory pathways, are offered as concrete steps toward rigorous cross-technology comparison and evidence-based clinical translation.

ACKNOWLEDGMENT

This work was supported in part by the São Paulo Research Foundation (FAPESP) under Grant 2024/14632-3, in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, under Grant 304427/2022-5, and in part by a CNPq doctoral scholarship under Grant 140566/2026-0 granted to the corresponding author.

REFERENCES

[1] N. Kumar, G. Akangire, B. Sullivan, K. Fairchild, and V. Sampath, "Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: Big data to the forefront," *Pediatric Research*, vol. 87, no. 2, pp. 210–220, 2020.

[2] E. Senechal, A. Maximov, E. Jeanne *et al.*, "Next generation of non-contact and wireless vital sign monitoring technology in the neonatal intensive care unit: A systematic review," *Pediatric Research*, 2025.

[3] T. Oranges, V. Dini, and M. Romanelli, "Skin physiology of the neonate and infant: Clinical implications," *Advances in Wound Care*, vol. 4, no. 10, pp. 587–595, 2015.

[4] O. Bonner, K. Beardsall, N. Crilly *et al.*, ""there were more wires than him": The potential for wireless patient monitoring in neonatal intensive care," *BMJ Innovations*, vol. 3, no. 1, pp. 12–18, 2017.

[5] E. Williams, R. Ascherl, V. D. Gaertner *et al.*, "Future perspectives of heart rate and oxygenation monitoring in the neonatal intensive care unit – a narrative review," *Journal of Clinical Monitoring and Computing*, vol. 39, pp. 901–915, 2025.

[6] A. C. Tricco, E. Lillie, W. Zarin *et al.*, "Prisma extension for scoping reviews (prisma-scr): Checklist and explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[7] L. A. Ferreira, L. P. Carlini, G. A. S. Coutrin, L. F. Buzuti, R. G. Magalhães Jr., R. N. Orsi, T. M. Heiderich, G. Cianciarullo, S. P. Gimenez, C. Pirie, C. F. Moreno-Garcia, M. C. M. Barros, R. Guinsburg, and C. E. Thomaz, "Barriers and pathways to clinical translation of ai in automated neonatal pain assessment," in *Proceedings of the 26th Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS) – I Workshop de Grandes Desafios da Computação Aplicada à Saúde*, Ouro Preto, Brazil, Jun. 2026, pp. 86–97.

[8] G. Cianciarullo, "Supplementary tables: Synthesis of included studies and performance metrics — sensors for physiological monitoring in neonatal intensive care units," Aug. 2026.

[9] H. U. Chung, A. Y. Rwei, A. Hourlier-Fargette *et al.*, "Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units," *Nature Medicine*, vol. 26, no. 3, pp. 418–429, 2020.

[10] S. Slattery, S. Pessano, J. Yoo *et al.*, "Continuous wireless sensor monitoring with applied diagnostics: Clinical sensor pain scale and automated sensor pain scale in the nicu," *BMJ Health and Care Informatics*, vol. 32, no. 1, 2025.

[11] H. U. Chung, B. H. Kim, J. Y. Lee *et al.*, "Binodal, wireless epidermal electronic systems with in-sensor analytics for neonatal intensive care," *Science*, vol. 363, no. 6430, 2019.

[12] H. Chen, W. Chen, S. Bao *et al.*, "Design of an integrated wearable multi-sensor platform based on flexible materials for neonatal monitoring," *IEEE Access*, vol. 8, pp. 23 732–23 747, 2020.

[13] S. M. Najib, N. B. Hassan, N. T. A. B. M. Shair *et al.*, "Intelligent neonatal monitoring system based on android application using multi sensors," in *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2019, pp. 131–135.

[14] W. Shalannanda, I. Zakia, E. Sutanto *et al.*, "Design of hardware module of iot-based infant incubator monitoring system," in *2020 6th International Conference on Wireless and Telematics (ICWT)*, 2020, pp. 1–6.

[15] S. Shamsir, O. Hassan, and S. K. Islam, "Smart infant-monitoring system with machine learning model to detect physiological activities and ambient conditions," in *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2020, pp. 1–6.

[16] H. D. Nemomssa and T. B. Alemneh, "Device for remote and real-time monitoring of neonatal vital signs in neonatal intensive care unit using internet of things: Proof-of-concept study," *Journal of Clinical Monitoring and Computing*, vol. 37, no. 2, pp. 585–592, 2023.

[17] C. Tang, F. Lei, J. Liu *et al.*, "Infection prevention and early warning in neonatal intensive care unit based on physiological sensor monitoring," *Frontiers in Bioengineering and Biotechnology*, vol. 11, 2023.

[18] J. D. Kim, W. H. Lee, Y. Lee *et al.*, "Non-contact respiration monitoring using impulse radio ultra-wideband radar in neonates," *Royal Society Open Science*, vol. 6, no. 6, 2019.

[19] E. Senechal, D. Radeschi, L. Tao *et al.*, "The use of wireless sensors in the neonatal intensive care unit: A study protocol," *PeerJ*, vol. 11, 2023.

[20] E. Senechal, D. Radeschi, S. Lv *et al.*, "Wireless skin sensors for electrocardiogram and heart rate monitoring in the neonatal intensive care unit: A prospective feasibility, safety, and accuracy study," *Frontiers in Bioengineering and Biotechnology*, vol. 13, 2025.

[21] G. Cay, D. Solanki, M. A. A. Rumon *et al.*, "Neowear: An iot-connected e-textile wearable for neonatal medical monitoring," *Pervasive and Mobile Computing*, vol. 86, 2022.

[22] H. Dore, R. Aviles-Espinosa, Z. Luo *et al.*, "Characterisation of textile embedded electrodes for use in a neonatal smart mattress electrocardiography system," *Sensors*, vol. 21, no. 3, 2021.

[23] A. Abdou, S. Krishnan, and N. Mistry, "Evaluating a novel infant heart rate detector for neonatal resuscitation efforts: Protocol for a proof-of-concept study," *JMIR Research Protocols*, vol. 12, no. 1, 2023.

[24] S. J. Stockwell, T. C. Kwok, S. P. Morgan *et al.*, "Forehead monitoring of heart rate in neonatal intensive care," *Frontiers in Physiology*, vol. 14, 2023.

[25] H. A. Hashim, "Non-contact neonatal heart rate monitoring using blue filter technique with photoplethysmography imaging under uv illumination," *Journal of Biomedical Photonics & Engineering*, vol. 11, no. 1, 2025.

[26] M. Paul, S. Karthik, J. Joseph *et al.*, "Non-contact sensing of neonatal pulse rate using camera-based imaging: A clinical feasibility study," *Physiological Measurement*, vol. 41, no. 2, 2020.

[27] J. C. Cobos-Torres, M. Abderrahim, and J. Martínez-Orgado, "Non-contact, simple neonatal monitoring by photoplethysmography," *Sensors*, vol. 18, no. 12, 2018.

[28] K. Edanami, M. Kurosawa, H. T. Yen *et al.*, "Remote sensing of vital signs by medical radar time-series signal using cardiac peak extraction and adaptive peak detection algorithm: Performance validation on healthy adults and application to neonatal monitoring at an nicu," *Computer Methods and Programs in Biomedicine*, vol. 226, 2022.

[29] S. Goedicke-Fritz, D. Schmiech, R. Thull *et al.*, "Radar-based monitoring: A proof of principle study in a piglet model for a novel approach in non-contact vital sign monitoring," *Sensors*, vol. 26, no. 7, 2026.

[30] A. Bekele, S. Nizami, Y. S. Dosso *et al.*, "Real-time neonatal respiratory rate estimation using a pressure-sensitive mat," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2018, pp. 1–5.

[31] I. Mahbub, T. Oh, S. Shamsir *et al.*, "Design of a pyroelectric charge amplifier and a piezoelectric energy harvester for an invasive–noninvasive wearable and self-powered respiratory monitoring system," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 105–108.

[32] I. Mahbub, S. Islam, S. Shamsir *et al.*, "A low-power wearable respiration monitoring sensor using a pyroelectric transducer," in *2017 USNC-URSI National Radio Science Meeting*, 2017, pp. 1–2.

[33] N. Hakimi, M. Shahbakhti, J. M. Horschig *et al.*, "Respiratory rate extraction from neonatal near-infrared spectroscopy signals," *Sensors*, vol. 23, no. 9, 2023.

[34] C. B. Pereira, K. Heimann, B. Venema *et al.*, "Estimation of respiratory rate from thermal videos of preterm infants," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 3818–3821.

[35] I. Lorato, S. Stuijk, M. Meftah *et al.*, "Towards continuous camera-based respiration monitoring in infants," *Sensors*, vol. 21, no. 7, 2021.

[36] G. Beltrão, R. Stutz, F. Hornberger *et al.*, "Contactless radar-based breathing monitoring of premature infants in the neonatal intensive care unit," *Scientific Reports*, vol. 12, no. 1, 2022.

[37] B. U. Harris, S. Stewart, A. Verma *et al.*, "Accuracy of a portable pulse oximeter in monitoring hypoxemic infants with cyanotic heart disease," *Cardiology in the Young*, vol. 29, no. 8, pp. 1025–1029, 2019.

[38] S. Lyra, J. Rixen, K. Heimann *et al.*, "Camera fusion for real-time temperature monitoring of neonates using deep learning," *Medical & Biological Engineering & Computing*, vol. 60, no. 6, pp. 1787–1800, 2022.

[39] F. Voss, S. Lyra, D. Blase *et al.*, "A setup for camera-based detection of simulated pathological states using a neonatal phantom," *Sensors*, vol. 22, no. 3, 2022.

[40] L. Giordano, A. Perri, E. Tiberi *et al.*, "The utility and safety of a continuous glucose monitoring system (cgms) in asphyxiated neonates during therapeutic hypothermia," *Diagnostics*, vol. 13, no. 18, 2023.

[41] A. Perri, L. Giordano, M. Corsello *et al.*, "Continuous glucose monitoring (cgm) in very low birth weight newborns needing parenteral nutrition: Validation and glycemic percentiles," *Italian Journal of Pediatrics*, vol. 44, no. 1, 2018.

[42] E. Losiouk, G. Lanzola, A. Galderisi *et al.*, "A telemonitoring service supporting preterm newborns care in a neonatal intensive care unit," in *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*, 2017, pp. 1–6.

[43] K. Braune, M. Wäldchen, K. Raile *et al.*, "Open-source technology for real-time continuous glucose monitoring in the neonatal intensive care unit: Case study in a neonate with transient congenital hyperinsulinism," *Journal of Medical Internet Research*, vol. 22, no. 12, 2020.

[44] M. O'Sullivan, A. Temko, A. Bocchino *et al.*, "Analysis of a low-cost eeg monitoring system and dry electrodes toward clinical use in the neonatal icu," *Sensors*, vol. 19, no. 11, 2019.

[45] A. Hakimi, N. Hassan, K. Anwar *et al.*, "Development of real-time patient health (jaundice) monitoring using wireless sensor network," in *2016 3rd International Conference on Electronic Design (ICED)*, 2016, pp. 404–409.

[46] R. Joshi, B. L. Bierling, X. Long *et al.*, "A ballistographic approach for continuous and non-obtrusive monitoring of movement in neonates," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–10, 2018.

[47] Z. Peng, Y. Zhao, D. Kommers *et al.*, "Continuous movement quantification in preterm infants using fiber mat: Advancing long-term monitoring in hospital settings," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

[48] C. S. Bosoc, G. Suciu, and A. M. Drăgulinescu, "Environmental monitoring system in incubators for newborns," in *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2019, pp. 1–4.

[49] U.S. Food and Drug Administration, "Device approvals and clearances," https://www.fda.gov/medical-devices/products-and-medical-procedures/device-approvals-and-clearances, 2026, accessed: 2026-08-24.

[50] ——, "Artificial intelligence in software as a medical device," https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device, 2026, accessed: 2026-08-24.

[51] European Parliament and Council of the European Union, "Regulation (EU) 2017/745 on medical devices (Medical Device Regulation)," Official Journal of the European Union, L 117, pp. 1–175, 2017, http://data.europa.eu/eli/reg/2017/745/oj. Accessed: 2026-08-24.

Mapping LIG-Based Biomedical Sensing Research: Bibliometric Trends, Thematic Clusters, and International Collaboration

1st Théo Saldanha Teixeira
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
theo.saldanhateixeira@ufpe.br

2nd Guilherme Vinícius Costa Silva
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
guilherme.costasilva@ufpe.br

3rd Paulo Henrique da Silva
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
paulo.hsilva3@ufpe.br

4th Ana Cristina Silveira Martins
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
martinsanaacs@gmail.com

5th Cláudia Sampaio de Andrade Lima
ITCBio
Federal University of Pernambuco
Recife, Brazil
claudia.alima@ufpe.br

6th Ricardo Yara
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
ricardo.yara@ufpe.br

***Abstract*—Since its introduction in 2014, Laser-Induced Graphene (LIG) has gained increasing attention for flexible and biomedical sensing applications. This study presents a quantitative bibliometric mapping of LIG-based biomedical sensing research, focusing on thematic structure, temporal evolution, and international collaboration. A total of 734 records published between 2014 and April 27, 2026, were retrieved from the Web of Science Core Collection and analyzed using VOSviewer. Keyword co-occurrence analysis identified 60 terms organized into six thematic clusters related to sensing properties, electrochemical biosensing, energy storage, hybrid nanostructures, synthesis parameters, and computational intelligence. Occurrence and Total Link Strength indicators complemented the network interpretation. Temporal mapping showed increasing attention to electrochemical detection, nanomaterials, and machine learning, while country co-authorship analysis identified 25 countries, with China, the United States, South Korea, and India accounting for the largest publication volumes. Machine learning appeared as a comparatively recent and less connected theme, suggesting opportunities for greater integration between data-driven methods, sensing, and LIG fabrication. Overall, the study provides an exploratory quantitative overview of the evolution and current structure of LIG-based biomedical sensing research.**

***Index Terms*—Laser-Induced Graphene, Biomedical Sensors, Bibliometrics, VOSviewer**

I. Introduction

In 2014, Lin et al. introduced Laser-Induced Graphene (LIG), a three-dimensional porous carbon material produced by photothermal conversion of polymeric substrates using a CO_2 laser [1]. Its relatively simple fabrication, high electrical conductivity, mechanical flexibility, and tunable surface properties have supported its expansion toward flexible and biomedical sensing applications [2], [3]. These characteristics enable the detection of both physical stimuli, such as strain and pressure, and biochemical analytes, including glucose, lactate, and pH in wearable configurations [4].

As the LIG literature has expanded, reviews have addressed its material properties, fabrication methods, and applications in sensing and healthcare [5]. More recently, machine-learning approaches have also been incorporated into LIG-based sensing systems for tasks such as signal classification and pattern recognition [6]. However, existing studies are predominantly focused on technological and application-oriented reviews [7], providing limited quantitative characterization of how biomedical sensing themes have evolved over time and how research activity is distributed across international collaboration networks.

This study addresses this perspective through a bibliometric mapping specifically focused on LIG-based biomedical sensing. Using network and overlay analyses [8], the study examines thematic clusters, temporal trends, international co-authorship, and the emerging association of machine learning with biomedical sensing. Rather than evaluating device performance, the analysis provides a quantitative overview of the structure and evolution of the literature, identifying established themes, recent research directions, and patterns of international collaboration.

II. Methodology

A. Search Strategy and Data Collection

The Web of Science Core Collection was selected as the sole bibliographic database because it provides standardized bibliographic and citation metadata suitable for reproducible bibliometric analyses and provides consistent coverage across the selected publication period [9]. The exclusive use of WoS may nevertheless result in the omission of records indexed only in other databases, such as Scopus, Dimensions, or IEEE Xplore, and is therefore recognized as a limitation of the study. The search string was designed to capture the

intersection of Laser-Induced Graphene (LIG) and biomedical sensing applications using Boolean operators: (”Laser-induced graphene” OR ”LIG”) AND (”sensor*” OR ”biosensor*” OR ”electrode*”) AND (”health*” OR ”medical” OR ”biomedical” OR ”wearable*”) spanning from 2014 to April 27, 2026. The initial query yielded 744 records (Articles, Review Articles, Proceeding Papers, Early Access, Letters and Editorial Materials) in English and Chinese, which were exported as plain text files containing full bibliographic records and cited references.

B. Data Cleaning and Quality Assurance

To ensure the technical integrity of the dataset, a manual screening of titles and abstracts was conducted by a single researcher. During this stage, 10 records were excluded as false positives because they were not directly related to laser-induced graphene research [10]. These records were primarily retrieved due to the use of the acronym “LIG” in unrelated contexts, such as lignin-based sensor studies. The final refined corpus consisted of 734 documents. No inter-reviewer agreement assessment was performed.

C. Data Normalization and Pre-processing

Data processing was performed with VOSviewer (v.1.6.20) [8]. A custom thesaurus file was used to normalize the metadata before map generation [11]. This involved merging synonymous terms (e.g., unifying ”LIG”, ”laser-induced graphene”, and ”induced graphene” into a single node) and standardizing plural and singular forms. Furthermore, generic terms that did not provide sufficient thematic specificity (e.g., ”performance”, ”fabrication”, ”design”, and ”challenges”) were removed because their high semantic generality could obscure more specific thematic relationships in the co-occurrence networks.

D. Bibliometric Mapping and Visualization

To map the landscape of the field, the analysis was divided into two specific scopes:

1) Keyword Co-occurrence Analysis: A minimum threshold of nine occurrences was selected as a pragmatic criterion to balance network comprehensiveness and visual interpretability. Lower thresholds generated a larger number of low-frequency terms, increasing network density and reducing the readability of the thematic clusters. This resulted in a network comprising 60 keywords. This analysis aims to identify the thematic pillars and technological clusters within the LIG-sensor domain.

2) Country Co-authorship Analysis: A minimum threshold of six publications per country was adopted to retain countries with sufficient publication activity for meaningful representation in the co-authorship network while avoiding excessive network sparsity. This resulted in 25 countries being included in the network. This mapping identifies the countries with the highest publication activity and the structure of international collaboration in the field.

a) Visualization Modes: For both analyses, *Network Visualization* was used to identify structural clusters and connectivity, while *Overlay Visualization* was utilized to track the temporal evolution and identify emerging trends up to 2026 [12] and [13].

The publication counts were also exported from WoS according to the ”Final Year of Publication” field and used to characterize annual scientific production. An exponential regression was fitted to the annual publication counts as a descriptive model of the observed temporal trend. Because records for 2026 were available only through April 27, 2026, this year was treated as incomplete and any extrapolation to the end of 2026 was considered exploratory rather than predictive.

The overall workflow for record selection, data processing, and bibliometric mapping is summarized in Fig. 1.

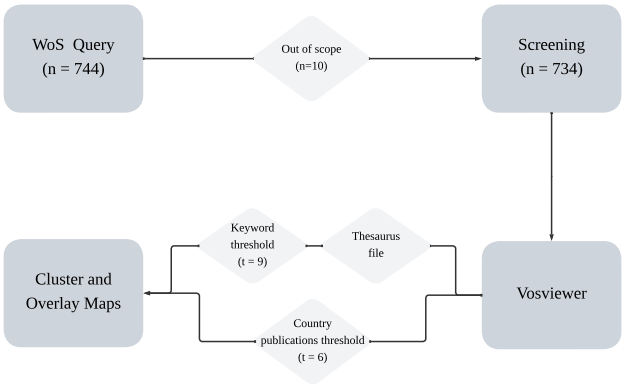


Fig. 1. Flow diagram illustrating the record selection, data processing, and bibliometric mapping thresholds.

III. Results and Discussion

A. Temporal Analysis

Fig. 2 illustrates a marked increase in scientific production on LIG-based biomedical sensing over the analyzed period. Although LIG was introduced in 2014, initial research focused primarily on fundamental materials science and energy-storage applications, such as supercapacitors. Within the scope of the present search, publications specifically addressing biomedical sensors and electrodes became more frequent from 2016 onward, indicating the expansion of LIG toward biomedical sensing applications.

Publication activity increased substantially after 2020, coinciding temporally with the COVID-19 pandemic and the growing interest in flexible point-of-care (PoC) diagnostic and remote health-monitoring technologies [14]. The number of publications increased from 11 in 2019 to 41 in 2020 and continued to rise, reaching 178 publications in 2025. This temporal pattern indicates a marked expansion of research activity in LIG-based biomedical sensing, although the bibliometric data alone do not establish a causal relationship between the COVID-19 pandemic and this increase.

An exponential regression was fitted to the annual publication counts as a descriptive representation of the observed temporal trend. When 2026 was excluded because it represents an incomplete year, the regression yielded $R^2 = 0.95$, indicating a strong fit of the selected function to the observed publication data. This value describes the fit of the model to the available observations but does not demonstrate that exponential growth will continue. Of the 734 documents in the final corpus, 721 contained information in the "Final Publication Year" field, while 13 records did not. For 2026, 90 publications were available as of April 27 and were therefore treated as an incomplete annual observation. An annualized estimate of 270 publications was used only for visualization of the potential full-year output and is not considered an observed publication count. Overall, the temporal analysis indicates rapid growth in research activity while avoiding extrapolation of the incomplete 2026 data as evidence of continued exponential expansion.

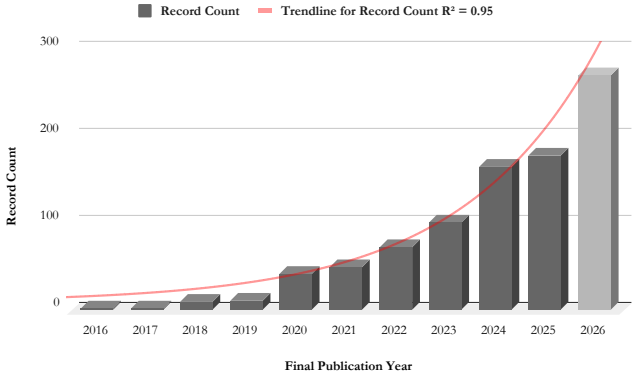


Fig. 2. Temporal evolution of publications on LIG-based biomedical sensing between 2016 and 2026. The 2026 value represents an annualized estimate based on records available through April 27.

B. Keyword Network Analysis

VOSviewer represents keyword co-occurrence as a weighted network in which links indicate the number of documents in which two keywords co-occur. Link strengths were normalized using the association-strength method, and spatial proximity in the map reflects stronger normalized relatedness between terms.

As illustrated in Fig. 3, the 60 selected keywords were organized into six thematic clusters. Table I summarizes representative keywords from each cluster together with their occurrence counts and Total Link Strength (TLS), providing a quantitative complement to the network visualization.

a) Cluster 1 — Sensing Properties: The red cluster is primarily associated with mechanical and wearable sensing applications of LIG. Representative terms include *strain sensor*, *wearable sensor*, and *pressure sensor*, with *strain sensor* showing 50 occurrences and a Total Link Strength (TLS) of 144, followed by *wearable sensor* (27 occurrences; TLS = 78) and *pressure sensor* (23 occurrences; TLS = 67) (Table I). These terms indicate a strong emphasis on the use of LIG as a flexible sensing material for detecting mechanical deformation and pressure, particularly in wearable configurations. The coexistence of these topics within the same cluster suggests an established connection between LIG fabrication and mechanically responsive sensor architectures. Rather than indicating a specific technological transition, the network primarily reflects the strong association of LIG with flexible and wearable sensing applications within the analyzed literature [15].

TABLE I
REPRESENTATIVE KEYWORDS AND BIBLIOMETRIC INDICATORS FOR EACH THEMATIC CLUSTER.

Cluster	Theme	Representative keywords	Occ. / TLS
1	Sensing properties	Strain sensor; wearable sensor; pressure sensor	50/144; 27/78; 23/67
2	Biomarkers and electrochemistry	Electrode; biosensor; electrochemical sensor	85/287; 60/178; 47/143
3	Energy storage	Supercapacitor; energy; energy storage	80/246; 24/73; 11/41
4	Hybrid nanostructures	Nanoparticle; nanocomposite; MXene	41/137; 22/76; 9/24
5	Synthesis parameters	Polyimide; conductivity	42/132; 9/26
6	Computational intelligence	Machine learning	19/36

Note: Occ. = keyword occurrences; TLS = Total Link Strength.

b) Cluster 2 — Biomarkers and Electrochemistry: The green cluster is centered on electrochemical sensing and biosensing applications, with *electrode*, *biosensor*, and *electrochemical sensor* representing prominent terms. Among these, *electrode* presented the highest occurrence and TLS values (85 and 287, respectively), followed by *biosensor* (60 occurrences; TLS = 178) and *electrochemical sensor* (47 occurrences; TLS = 143) (Table I). The strong connectivity of these terms indicates that electrochemical interfaces constitute an important thematic component of LIG-based biomedical sensing research. Their association with terms related to biomarkers and biological targets further reflects the application of LIG electrodes in analytical and biosensing contexts [16]. However, the co-occurrence network identifies thematic associations rather than demonstrating specific analytical performance, selectivity, or clinical applicability.

c) Cluster 3 — Energy Storage: The blue cluster is centered on energy-storage applications and includes *supercapacitor*, *energy*, and *energy storage* as representative terms. Among these, *supercapacitor* showed the highest occurrence and TLS values (80 and 246, respectively), while *energy* and *energy storage* presented 24 and 11 occurrences, with TLS values of 73 and 41, respectively (Table I). Although not

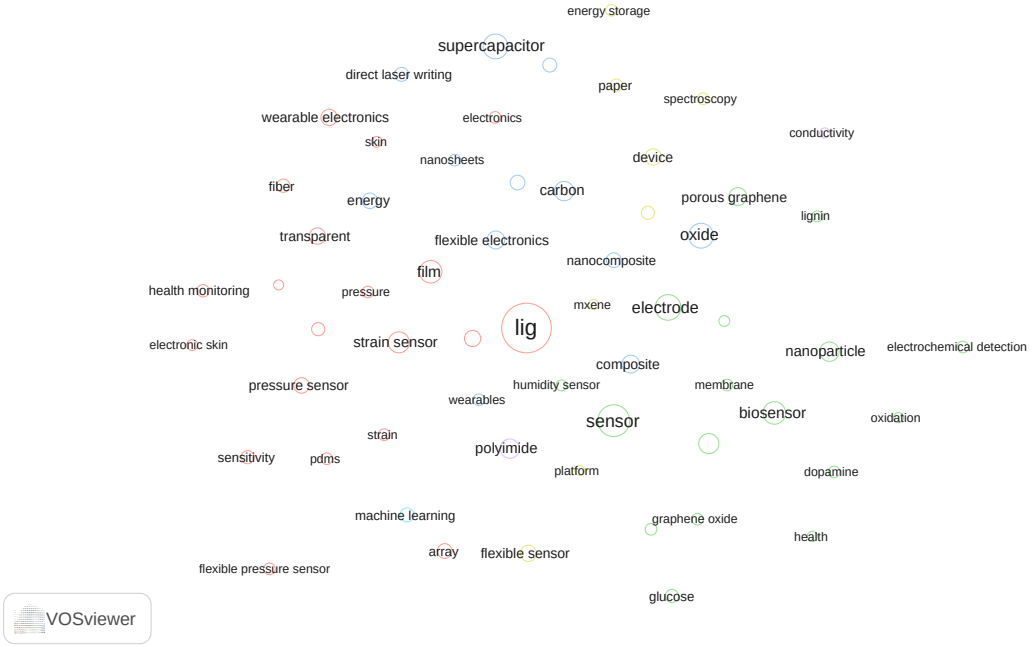


Fig. 3. Keyword co-occurrence map for clusters visualization. Generated using VOSviewer software.

exclusively biomedical, the presence of this cluster indicates that energy-storage research remains an important component of the broader LIG literature represented in the dataset. Similar observations have been reported in broader LIG reviews, where energy devices are described as an important component of the field's technological development [17]. Its coexistence with biomedical sensing themes may indicate potential relevance for self-powered or energy-efficient wearable systems, although the co-occurrence analysis alone does not establish specific technological integration or device performance.

d) Cluster 4 — Hybrid Nanostructures: The yellow cluster is characterized by terms related to the incorporation of nanostructured materials into LIG-based systems. *Nanoparticle*, *nanocomposite*, and *MXene* were selected as representative terms, with 41, 22, and 9 occurrences and TLS values of 137, 76, and 24, respectively (Table I). The relatively strong connectivity of *nanoparticle* and *nanocomposite* indicates that hybrid material architectures constitute a relevant thematic direction within the analyzed literature [18]. The presence of MXene, although less frequent, further indicates the association of LIG research with emerging two-dimensional and composite materials. These associations suggest an expansion of the material combinations investigated for LIG-based sensing, although the bibliometric network alone cannot determine whether such combinations provide superior sensing performance or represent a specific technological progression.

e) Cluster 5 — Synthesis Parameters: The purple cluster is associated with the material synthesis and electrical properties of LIG, with *polyimide* and *conductivity* as representative terms. *Polyimide* presented 42 occurrences and a TLS of 132, whereas *conductivity* showed 9 occurrences and a TLS of 26 (Table I). The prominence of polyimide reflects its strong association with LIG-related research within the analyzed corpus, while the occurrence of conductivity indicates attention to the electrical properties of the resulting material. Together, these terms suggest that substrate selection and material properties remain connected to the development of LIG-based sensing systems [19]. However, the co-occurrence structure does not allow direct conclusions regarding the optimal synthesis conditions or the effect of fabrication parameters on sensor performance.

f) Cluster 6 — Computational Intelligence: The cyan cluster is represented primarily by *machine learning*, which showed 19 occurrences and a TLS of 36 (Table I). Its comparatively lower occurrence and connectivity, together with its concentration around a single representative term, indicate that computational intelligence constitutes a smaller thematic component of the analyzed literature than the more established sensing, electrochemical, and energy-related clusters. Nevertheless, its presence within the keyword network indicates an association between LIG-based sensing and data-driven approaches. The relatively limited connectivity of machine learning with synthesis-related terms may indicate an opportunity for future studies integrating computational methods with material fabrication and sensor optimization [20]. This pattern should be interpreted as a potential research direction rather than evidence that autonomous or closed-loop LIG fabrication is already an established research area.

The combined analysis of the six clusters reveals a progressively interconnected research landscape spanning LIG

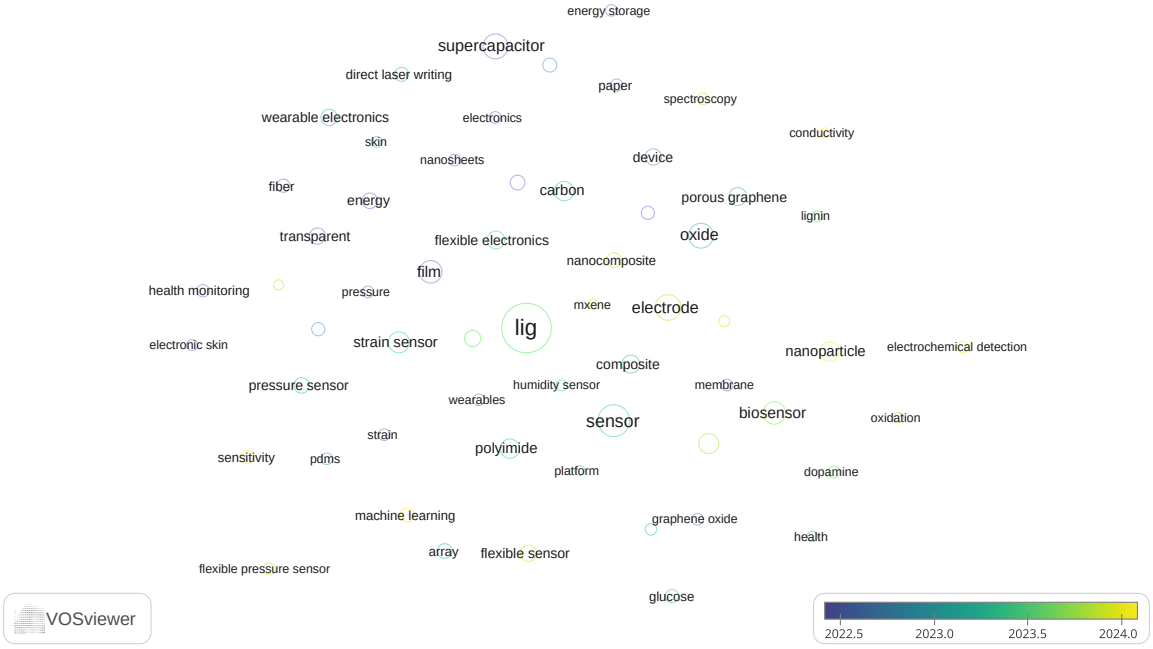


Fig. 4. Keyword co-occurrence map for trend visualization. Generated using VOSviewer software.

synthesis, material engineering, energy-related applications, electrochemical sensing, wearable devices, and computational approaches. Although these thematic groups can be interpreted as complementary stages within the development of LIG-based sensing systems, the co-occurrence network does not establish a strict technological hierarchy among them [2]. The comparatively weaker connection between computational intelligence (Cluster 6) and synthesis-related terms suggests that data-driven approaches remain less integrated with material fabrication than with sensing and signal-processing applications [20]. Rather than representing an established limitation of the field, this pattern may indicate an opportunity for future studies combining computational methods with LIG synthesis and optimization [21].

C. Keyword Overlay Analysis

Overlay visualization was used to examine the average publication year associated with each keyword. As shown in Fig. 4, terms associated with energy storage and material fabrication are concentrated around earlier average publication years (∼2022), whereas terms such as 'machine learning', 'electrochemical detection', and 'nanoparticle' are predominantly associated with later periods (∼2024). This temporal distribution suggests increasing research attention toward intelligent and selective biomedical sensing systems.

The more recent average publication years of terms such as nanoparticle and electrochemical detection are consistent with increasing attention to functionalization and selective electrochemical sensing [22].

Conversely, clusters associated with earlier average years (around 2022), particularly those containing terms such as "supercapacitor", "film", and "energy", reflect the initial focus of LIG research on materials science and energy storage applications. The persistence of these terms as large nodes indicates their foundational role in the field. The co-occurrence of terms such as "health monitoring", "electronic skin", and "wearables" within the same temporal range indicates that biomedical sensing research was already developing alongside established energy-related applications, revealing an early broadening of the LIG research landscape [3].

In summary, the keyword evolution tracks a transition from fundamental material synthesis toward high-level system integration. The recent temporal association of Machine Learning-related terms underscores this transition toward intelligent sensing systems. However, the comparatively weak connectivity between the computational intelligence cluster and the polyimide conversion nodes suggests a potential direction for future research: the development of AI-driven frameworks capable of optimizing both LIG synthesis and sensor performance in a unified manner [21] and [23].

D. Geospatial Collaboration Network

A minimum threshold of 6 publications per country was established for inclusion in the co-authorship network. Of the 61 countries with at least one publication, 25 met this criterion and were included in the final mapping. As shown in Fig. 5, China represents the largest node, with 378 publications and a Total Link Strength (TLS) of 110, followed by the United States (137 publications; TLS = 80), South Korea (91; TLS

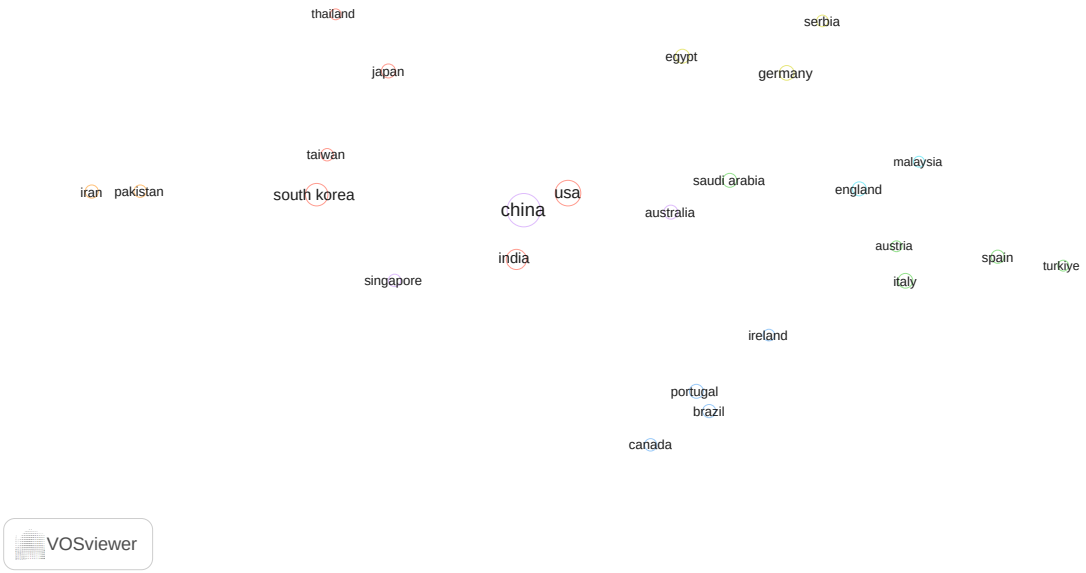


Fig. 5. Country co-authorship map for clusters visualization. Generated using VOSviewer software.

= 48), and India (56; TLS = 16). In contrast, all remaining countries presented fewer than 20 publications, indicating a strong concentration of scientific production within a limited number of nations. The comparatively high TLS values of China, the United States, and South Korea further indicate their strong connectivity within the international co-authorship network. In South America, only Brazil reached the inclusion threshold, with 12 publications, highlighting the comparatively limited regional participation in the analyzed literature. Fig. 5 also reveals the formation of seven distinct collaboration clusters.

Cluster 1 (Red) comprises India, Japan, South Korea, Taiwan, Thailand, and the United States, concentrating the United States and several Asian contributors within the same collaboration group. This cluster reflects a strong concentration of publications and international links involving countries with established activity in flexible electronics and LIG device development [24], [25]. In particular, the United States (TLS = 80) and South Korea (TLS = 48) exhibit the highest connectivity within this group, supporting their prominent positions in the collaboration network. The observed grouping is therefore consistent with substantial research interaction among these countries, although the network topology itself does not establish the technological, institutional, or economic factors underlying these collaborations.

Cluster 2 (Green) groups Austria, Italy, Saudi Arabia, Spain, and Türkiye, forming a collaboration cluster involving countries from Europe and the Mediterranean region. Among them, Saudi Arabia presents the highest Total Link Strength (TLS = 24), despite having 16 publications, followed by Spain (13 publications; TLS = 15) and Italy (18; TLS = 11). These values indicate meaningful international connectivity within the cluster rather than publication volume alone. The observed grouping is also compatible with some degree of regional proximity and shared research interests in nanotechnology and materials science [26], although the bibliometric network itself does not establish the specific factors underlying these collaborations.

Cluster 3 (Blue) comprises Brazil, Canada, Ireland, and Portugal, connecting countries across the Atlantic within the same collaboration group. Portugal presents the highest publication volume in the cluster (16 publications; TLS = 7), while Brazil and Canada show similar connectivity (TLS = 8 each), despite their different publication counts of 12 and 11, respectively. The presence of Brazil and Portugal remains consistent with the possibility that linguistic and historical relationships may contribute to collaborative interactions [27]; however, such factors cannot be inferred directly from the network topology. Overall, this cluster illustrates how countries with comparatively lower publication volumes can still participate in identifiable international collaboration networks within LIG research.

Cluster 4 (Yellow) is composed of Germany, Egypt, and Serbia. Germany presents the highest publication volume and connectivity within the cluster (19 publications; TLS = 18), followed by Egypt (16; TLS = 11) and Serbia (8; TLS = 2). This grouping is consistent with the participation of these countries in nanoscience, laser technologies, and applied materials research [28], although the network itself does not establish a specific bridging role between fundamental and

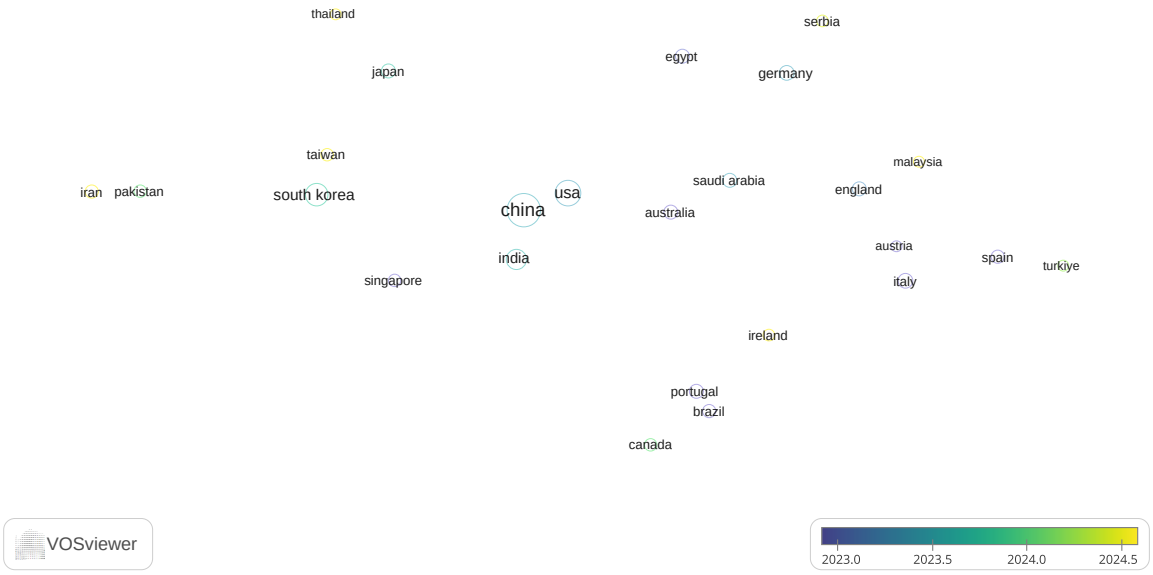


Fig. 6. Country co-authorship map for trend visualization. Generated using VOSviewer software.

applied research.

Cluster 5 (Purple), composed of China, Australia, and Singapore, represents one of the most prominent collaboration groups in the network. China clearly dominates both publication volume and connectivity (378 publications; TLS = 110), while Australia (16; TLS = 21) and Singapore (9; TLS = 9) show smaller but identifiable collaborative participation. This pattern is consistent with the strong presence of the Asia-Pacific region in LIG research, including synthesis, scalability, and interdisciplinary applications [4], [29].

Cluster 6 (Cyan), composed of England and Malaysia, forms a small cross-regional collaboration group. England presents 14 publications and a TLS of 20, whereas Malaysia has 6 publications and a TLS of 3. Their grouping may be compatible with established institutional and international research links [30], although the network itself does not identify the specific factors underlying this collaboration.

Cluster 7 (Orange), composed of Iran and Pakistan, represents a smaller regional collaboration group. Although Iran presents a slightly higher publication volume (12 publications; TLS = 3), Pakistan shows greater network connectivity (9 publications; TLS = 10). The observed grouping is consistent with some degree of regional interaction [31], but geographic or institutional drivers cannot be inferred directly from the co-authorship topology.

Overall, the geospatial network shows a marked concentration of scientific production and international connectivity around China and the United States, which belong to distinct collaboration clusters. This configuration indicates a multipolar collaboration structure rather than a single globally centralized network [32]. Countries with lower publication volumes, such as Brazil, Portugal, and Ireland, also participate in identifiable international connections, indicating that smaller contributors remain integrated into the global collaboration landscape [33]. These patterns describe the structure of scientific cooperation in LIG-based biomedical sensing without establishing the cultural, economic, or institutional factors responsible for the observed associations.

E. Geospatial Collaboration Overlay

As shown in Fig. 6, although China is the largest contributor in terms of publication volume, its average publication year is relatively recent (around 2023), indicating sustained activity across the more recent development of the field rather than predominantly early participation [29]. In contrast, countries such as Brazil, Italy, Australia, Egypt, and Portugal present slightly earlier average publication periods (2022–2023) while remaining comparatively smaller nodes in the network [27], [33].

Countries such as Iran, Malaysia, Ireland, Taiwan, and Thailand present more recent average publication years (around 2024), indicating increasing participation in the recent literature [30]. This temporal distribution suggests a gradual broadening of geographic participation in LIG-based biomedical sensing. The relatively accessible fabrication requirements of LIG, compared with more complex graphene-production techniques such as Chemical Vapor Deposition (CVD), may contribute to this broader participation, since LIG can be produced using comparatively simple CO_2 laser-based systems [17]. However, the bibliometric data alone do not establish

fabrication cost or accessibility as the direct cause of the observed geographic expansion.

IV. Conclusion

This study presented a quantitative and exploratory bibliometric mapping of Laser-Induced Graphene (LIG) research in biomedical sensing from 2014 to April 27, 2026. The results show a marked expansion and thematic diversification of the field, with increasing associations among wearable electronics, electrochemical biosensing, hybrid materials, and computational approaches. Keyword and collaboration analyses further revealed that research activity remains concentrated in a limited number of countries, while smaller contributors also participate in identifiable international networks.

Machine learning emerged as a comparatively recent and less connected theme, with stronger associations with sensing and signal-processing applications than with LIG synthesis. Rather than defining an established technological frontier, this pattern suggests an opportunity for future studies integrating data-driven methods with material fabrication and sensor optimization.

These findings support the continued investigation of LIG for flexible and potentially personalized biomedical sensing applications [34]. Nevertheless, the results should be interpreted considering the exclusive use of Web of Science, the adopted search strategy, the dependence of bibliometric networks on indexed keywords and selected thresholds, the single-researcher screening process, and the continuous updating of bibliographic databases. Future studies may expand database coverage and explore citation, institutional, and translational relationships in greater depth.

References

[1] J. Lin et al., "Laser-induced porous graphene films from commercial polymers," Nature Communications, vol. 5, no. 1, p. 5714, 2014.

[2] Z. Li et al., "Laser-Induced Graphene-Based Sensors in Health Monitoring: Progress, Sensing Mechanisms, and Applications," Small Methods, vol. 8, no. 11, p. 2400118, Nov. 2024

[3] Z. Ma and B. L. Khoo, "Recent advances in laser-induced-graphene-based soft skin electronics for intelligent healthcare," Soft Science, vol. 4, no. 1, p. 20, 2024.

[4] T.-S. D. Le et al., "Laser-induced graphene for biomedical applications: innovations in health monitoring and diagnostics," Nanoscale Horizons, vol. 10, no. 11, pp. 2688–2721, 2025.

[5] R. Ye, D. K. James, and J. M. Tour, "Laser-induced graphene: from discovery to translation," Advanced Materials, vol. 31, no. 1, p. 1803621, 2019.

[6] Y. Li et al., "Stretchable electronic skin using laser-induced graphene and liquid metal with an action recognition system powered by machine learning," Advanced Functional Materials, vol. 34, no. 30, p. 2313824, 2024.

[7] M. J. Islam, "Sustainable laser-induced graphene (LIG) for electrochemical sensing, energy storage and harvesting," Ph.D. dissertation, University College Cork, Cork, Ireland, 2025.

[8] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, 2010.

[9] V. K. Singh et al., "The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis," Scientometrics, vol. 126, no. 6, pp. 5113–5142, 2021.

[10] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," bmj, vol. 372, 2021.

[11] I. Zupic and T. Čater, "Bibliometric methods in management and organization," Organizational Research Methods, vol. 18, no. 3, pp. 429–472, 2015.

[12] C. Chen, "Science mapping: A systematic review of the literature," Journal of Data and Information Science, vol. 2, no. 2, 2017.

[13] Z. Liu, F. Yin, and W. Wang, "Scientometric analysis of textile sensors," The Journal of The Textile Institute, vol. 116, no. 1, pp. 45–60, 2025.

[14] K. Guk et al., "Evolution of wearable devices with real-time disease monitoring for personalized healthcare," Nanomaterials, vol. 9, no. 6, p. 813, 2019.

[15] F. M. Vivaldi et al., "Three-dimensional (3D) laser-induced graphene: structure, properties, and application to chemical sensing," ACS Applied Materials & Interfaces, vol. 13, no. 26, pp. 30245–30260, 2021.

[16] A. C. Bressi et al., "Bioderived laser-induced graphene for sensors and supercapacitors," ACS Applied Materials & Interfaces, vol. 15, no. 30, pp. 35788–35814, 2023.

[17] R. Ye, D. K. James, and J. M. Tour, "Laser-Induced Graphene," Accounts of Chemical Research, vol. 51, no. 7, pp. 1609–1620, 2018.

[18] F. Wang et al., "Laser-induced graphene: preparation, functionalization and applications," Materials technology, vol. 33, no. 5, pp. 340-356, 2018.

[19] H. Wang et al., "A soft and stretchable electronics using laser-induced graphene on polyimide/PDMS composite substrate," npj Flexible Electronics, vol. 6, no. 1, p. 26, 2022.

[20] R. Murray, M. Burke, D. Iacopino, and A. J. Quinn, "Design of experiments and optimization of laser-induced graphene," ACS Omega, vol. 6, no. 26, pp. 16736–16743, 2021.

[21] W.-Y. Wang et al., "Big data-assisted digital twins for the smart design and manufacturing of advanced materials: From atoms to products," Journal of Materials Informatics, vol. 2, no. 1, 2022.

[22] A. T. Lawal, "Advances in nanoparticle-based electrochemical biosensors for bioanalytical applications," Sensing Technology, vol. 4, no. 1, pp. 2630469, 2026.

[23] G. Ghimire et al., "Machine learning optimization of laser-induced graphene parameters for surface-enhanced Raman spectroscopic detection of glucose," Sensors and Actuators B: Chemical, vol. 432, pp. 137466, 2025.

[24] S. Aftab et al., "Laser-induced graphene for advanced sensing: Comprehensive review of applications," ACS Sensors, vol. 9, no. 9, pp. 4536–4554, 2024.

[25] H. Wang et al., "Laser-induced graphene based flexible electronic devices," Biosensors, vol. 12, no. 2, pp. 55, 2022.

[26] A. Kaidarova et al., "Enhanced graphene sensors via multi-lasing fabrication," IEEE Sensors Journal, vol. 21, no. 17, pp. 18562–18570, 2021.

[27] E. C. E. Almeida and P. M. A. R. Correia, "Scientific production between Brazil and Portugal: A comparative study of the periods 1980–90 and 2005–2015," Lex Humana, vol. 12, no. 1, pp. 17–37, 2020.

[28] T. U. Tumkur et al., "A review of laser materials processing paradigms," MRS Bulletin, pp. 1–20, 2025.

[29] M. Zhan et al., "Graphene oxide research: Current developments and future directions," Nanomaterials, vol. 15, no. 7, pp. 507, 2025.

[30] Z. Ismail et al., "Polyphenols as a nanotechnology source for a laser-based chemistry educational activity," Journal of Chemical Education, vol. 102, no. 9, pp. 4166–4174, 2025.

[31] T. H. BANGLANI et al. "Graphene-based nanocomposites for gas sensors: challenges and opportunities," Reviews in Inorganic Chemistry, vol. 44, no. 3, pp. 385-408, 2024.

[32] S. Marginson, "'All things are in flux': China in global science," Higher Education, vol. 83, no. 4, pp. 881–910, 2022.

[33] Z. Chinchilla-Rodríguez et al., "Dependencies and autonomy in research performance: Examining nanoscience and nanotechnology in emerging countries," Scientometrics, vol. 115, no. 3, pp. 1485–1504, 2018.

[34] E. Jovanov, "Wearables meet IoT: Synergistic personal area networks (SPANs)," Sensors, vol. 19, no. 19, pp. 4295, 2019.

Mechatronic Platform Driven by Thoracic Impedance Signals for the Experimental Evaluation of a Radar-Based Respiratory Monitoring System

Julio Cardoso Scorpion*
SENAI Innovation Institute for Embedded Systems
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0009-0002-8669-2249

João Lucas Pereira dos Santos de Paula*
SENAI Innovation Institute for Embedded Systems
Institute of Biomedical Engineering
Federal University of Santa Catarina
Florianópolis, Brazil
ORCID: 0000-0001-8230-5358

Feres Azevedo Salem
SENAI Innovation Institute for Embedded Systems
Florianópolis, Brazil
ORCID: 0000-0002-6223-0034

Elço João dos Santos Junior
SENAI Innovation Institute for Embedded Systems
Florianópolis, Brazil
ORCID: 0000-0001-8295-9677

Flávio Gabriel Oliveira Barbosa
SENAI Innovation Institute for Embedded Systems
Florianópolis, Brazil
ORCID: 0000-0003-2357-4122

Paulo Alberto Macedo Vieira Violada
SENAI Innovation Institute for Embedded Systems
Florianópolis, Brazil
E-mail: pauloav@sc.senai.br

***Abstract*—Short-range radar has emerged as a promising contactless technology for respiratory monitoring. However, experiments with human subjects are affected by inter-subject variability, involuntary motion, and limited repeatability of respiratory patterns. This work presents a controlled mechatronic platform driven by thoracic-impedance waveforms from the publicly available BIDMC dataset for benchtop evaluation of the Acconeer XM132 pulsed coherent radar. Twelve recordings of approximately 240 s were reproduced through servo-driven motion and acquired simultaneously by the radar and an auxiliary MPU6050 inertial sensor. Respiratory rate (RR) was evaluated using different analysis-window lengths and estimation methods. FFT-based estimation with 60 s windows yielded a mean absolute error of 1.94 breaths per minute (bpm) and a median error of 0.09 bpm. For complete 240 s recordings, the mean and median errors were 1.49 and 0.03 bpm, respectively; the mean was dominated by one second-harmonic selection error. These results support the use of the platform as a controlled benchtop environment for preliminary characterization of radar acquisition and RR-processing strategies, while independent displacement measurements and human validation remain necessary before conclusions about mechanical or clinical accuracy can be made.**

***Index Terms*—Contactless respiratory monitoring, pulsed coherent radar, RR estimation, mechatronic platform.**

I. Introduction

Respiratory rate (RR) is considered an important vital sign, as changes in the breathing pattern may indicate variations in a patient's clinical condition and, in intensive care settings, may serve as early indicators of clinical deterioration [1], [2]. Furthermore, dysfunctional breathing comprises a set of abnormalities in the respiratory pattern that may be associated with chronic conditions such as asthma and chronic obstructive pulmonary disease (COPD) [3].

Conventional respiratory monitoring methods often require direct physical contact with the patient through adhesive electrodes, chest belts, nasal cannulas, or other body-mounted sensors. Although these approaches are widely used, they may cause discomfort during prolonged monitoring and are susceptible to motion artifacts, sensor displacement, or poor patient compliance. Consequently, wearable and remote sensing technologies have been increasingly investigated for respiratory monitoring applications [4], [5]. Among remote sensing approaches, contactless sensors have attracted considerable attention because they enable respiratory monitoring without direct attachment to the body, a feature that is particularly advantageous in applications requiring continuous observation or reduced patient instrumentation [6].

Short-range radar sensors have been investigated as an alternative for monitoring thoracic and abdominal movements associated with respiration [7]. Different radar technologies, including frequency-modulated continuous-wave (FMCW), ultra-wideband (UWB), and millimeter-wave radars, have been applied to RR estimation, millimeter-scale displacement tracking, and respiratory or cardiorespiratory monitoring in a variety of experimental and clinical scenarios [8]–[12].

In this context, Acconeer sensors based on pulsed coherent radar operating in the millimeter-wave band provide a compact platform for distance measurement, presence detection, motion sensing, and respiratory monitoring applications [13]–[15]. Despite the availability of reference applications, a controlled experimental platform enables the evaluation of acquisition

*These authors contributed equally to this work.

parameters, signal processing techniques, and RR estimation methods before conducting experiments with human subjects.

Mechatronic test benches and respiratory phantoms complement volunteer-based experiments by providing repeatable and adjustable motion patterns [10], [16]. Existing platforms range from simplified programmed trajectories to more elaborate phantoms driven by recorded physiological signals; notably, Eder et al. used recorded thoracic-impedance waveforms to reproduce cardiopulmonary dynamics [16]. Public databases such as the BIDMC PPG and Respiration Dataset provide reproducible respiratory waveforms acquired from real subjects [17]–[19]. Thoracic impedance has also been used for respiratory monitoring and respiratory signal-quality assessment [20], [21].

Motivated by these considerations, this work proposes a mechatronic platform driven by thoracic impedance signals for the experimental characterization of an Acconeer radar sensor applied to respiratory monitoring. The platform reproduces physiologically inspired breathing patterns through a mechanically actuated structure, allowing the respiratory signal estimated by the radar to be compared with the imposed platform motion and with an auxiliary measurement obtained from an MPU6050 inertial sensor. In this study, the inertial sensor is not regarded as a gold standard for displacement measurement but rather as a complementary channel associated with the platform motion that supports RR estimation [22].

The main contributions of this work are:

- a compact benchtop platform that uses publicly available BIDMC thoracic-impedance recordings to drive respiration-like motion, preserving non-sinusoidal temporal variability in a reproducible input waveform;
- integration of the platform with an Acconeer XM132 pulsed coherent radar, an auxiliary inertial sensor, and dedicated acquisition software;
- characterization of RR recovery from the radar signal across analysis-window lengths and estimation methods, together with tracking-log and temporal waveform-similarity analyses.

II. Related Work

A. Radar-based respiratory monitoring

Radar-based respiratory monitoring relies on the periodic displacement of the thoracic and abdominal walls during the breathing cycle, and different radar technologies have been applied to RR estimation and cardiorespiratory monitoring [8], [9], [11], [12], [23]. The Acconeer sensor belongs to a class of pulsed radars operating around 60 GHz, capable of detecting small distance variations through the reflected response from the target [13]. The objective of this work is not to develop a new radar or signal processing algorithm, but rather to establish a controlled experimental environment for evaluating the performance of this sensor in a specific respiratory monitoring application.

B. Respiratory phantoms and emulated motion

Motion platforms and respiratory phantoms are widely used to reduce experimental variability during radar evaluation [10], [16]. Their designs span simplified programmed trajectories as well as physiologically driven systems. Eder et al., for example, incorporated recorded thoracic-impedance signals into a multi-person FMCW-radar phantom [16]. The contribution of the present work is therefore not the first use of physiological waveforms in a phantom; rather, it is a compact single-target implementation centered on the XM132 pulsed coherent radar, driven by public BIDMC respiratory recordings and coupled to inertial sensing, acquisition software, and a comparative analysis of RR estimators and window lengths. This scope complements more elaborate phantom frameworks while retaining a reproducible physiological input. Chest-wall motion itself is multidimensional and may be non-sinusoidal, which remains only partially represented by the one-dimensional mechanism used here [24], [25].

III. Materials and Methods

The proposed system consists of four main stages: acquisition of the reference respiratory signal, mechanical reproduction of thoracic motion, simultaneous data acquisition by the sensors, and signal processing for RR estimation. To this end, a mobile platform reproduces the motion derived from a thoracic impedance signal, while an Acconeer XM132 radar and an MPU6050 accelerometer simultaneously record the system response. The methodological workflow adopted in this study is illustrated in Fig. 1.

The study focuses on radar-based recovery of the imposed RR while documenting acquisition parameters available in the experimental logs, including the radar-derived tracking distance, region of interest (ROI), effective logging rate, and tracking status. An independent noise-floor experiment, a calibrated saturation threshold, and absolute displacement accuracy were not available and are therefore not claimed. The MPU6050 accelerometer was used only as a complementary measurement of the platform response and was not considered a gold standard for thoracic displacement [22], [26].

A. Respiratory signal database

This study employed the publicly available BIDMC PPG and Respiration Dataset from the PhysioNet platform. The dataset contains recordings from 53 patients, each with approximately 8 minutes of data sampled at 125 Hz. The available signals include electrocardiography (ECG), photoplethysmography (PPG), thoracic impedance respiration, and manually annotated respiratory cycles [17]–[19].

In this work, the dataset was used to obtain real respiratory patterns, which were subsequently employed to reproduce breathing-like motions on the experimental platform.

B. Mechatronic platform

A mechatronic platform was developed using a structure fabricated from PLA and acrylic, a gear transmission mechanism, two MG90S servo motors, and an ESP32 development

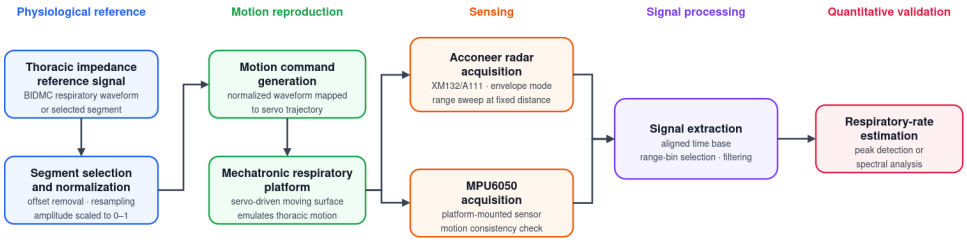


Fig. 1. Overall methodology adopted in this work, including physiological reference, motion reproduction, sensor acquisition, signal processing and respiratory-rate estimation.

board for system control. The gear mechanism converts the rotational motion of the servomotors into an approximately linear displacement, thereby simulating thoracic expansion and contraction. The two servomotors are symmetrically positioned to reproduce the bilateral motion of the rib cage. The acrylic upper plate supports the MPU6050 on its underside, allowing the sensor to move together with the structure, while simultaneously serving as the reference surface for the Acconeer radar, which is mounted in a fixed position to capture the platform motion.

The electronic architecture of the platform, including the connections among the ESP32, the Acconeer XM132 radar, the MPU6050 sensor, the servo motors, and the power supply, is shown in Fig. 2.

The thoracic-impedance respiratory signals from the BIDMC dataset were resampled to 20 Hz, normalized, and mapped to angular position commands for the servo motors. In the experimental logs, the recorded servo command spanned approximately 90–140 degrees across the analyzed trials. The mechanism was designed for an approximate nominal upper-plate travel of 20 mm; however, the displacement produced by the recorded commands was not calibrated with an independent position sensor, so this value is reported only as a nominal mechanical range. Mechanical displacement resolution, hysteresis, gear backlash, and run-to-run repeatability were not independently quantified. The ESP32 received the respiratory command through a serial interface and converted the servo positions into PWM drive signals. During acquisition, the radar's internal tracking reference reported target-distance estimates between approximately 0.394 and 0.460 m across the 12 recordings. These radar-derived distances were used for target tracking and are not treated as an independent measurement of platform displacement.

C. Signal acquisition sensors

RR was estimated from the signals acquired by the radar and the accelerometer, the radar being the primary sensor investigated.

1) Acconeer XM132 radar module: The XM132 module is a radar sensor based on Acconeer's Pulsed Coherent Radar (PCR) technology. The sensor operates in the 60 GHz frequency band, belonging to the class of millimeter-wave (mmWave) radars. Unlike frequency-modulated continuous-wave (FMCW) radars, the A111 radar embedded in the XM132 transmits short-duration electromagnetic pulses and performs coherent reception of the echoes reflected by surrounding objects [13]. This technology combines low power consumption, high spatial resolution, and the capability to detect very small displacements, making it suitable for RR estimation.

In this work, the *Envelope* service, one of the distance measurement methods provided by the XM132, was employed. In this method, the radar records the intensity of the reflected echoes at discrete distance positions (*bins*), generating a one-dimensional profile that represents the reflection amplitude along the monitored range.

Internally, the radar measures a complex signal given by

$$s(r) = I(r) + jQ(r), \tag{1}$$

where $I(r)$ and $Q(r)$ denote the in-phase and quadrature components, respectively. The envelope is obtained as the magnitude of this complex signal:

$$E(r) = \sqrt{I(r)^2 + Q(r)^2} = |s(r)|. \tag{2}$$

The value of $E(r)$ represents the reflection amplitude at each monitored distance. In this study, the platform position was estimated by tracking the dominant reflection within a fixed ROI. The logged ROI width was 70 mm in all 12 recordings, centered on an internally estimated target distance between approximately 0.394 and 0.460 m. The acquisition/playback target rate was 20 Hz; the effective rate of the merged CSV records was 20.00 Hz on average (19.992–20.002 Hz across recordings). The radar module communicates with the laptop through an RS-422 link and a NOVUS USB-I485 converter and is powered by a 12 V supply, as illustrated in Fig. 2.

The analyzed stream was the software-generated `radar_filt` time series stored in the experimental CSV files. The acquisition-side filter coefficients that generated this field were not retained in the merged logs; consequently, the reproducible quantitative pipeline reported here begins from `radar_filt` and applies the offline preprocessing described below. The merged logs contained no missing values in the reference, servo-command, MPU6050 X-axis, or filtered-radar columns. Because radar and ESP32 data were merged asynchronously, repeated or skipped internal frame

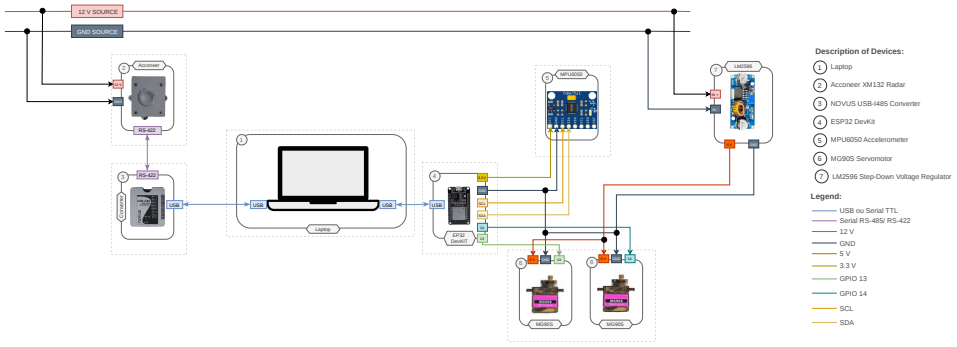


Fig. 2. Hardware architecture of the proposed platform, showing the communication interfaces and electrical connections.

indices in the merged log were not interpreted as a direct measure of source-frame packet loss.

2) MPU6050 accelerometer: In this study, the MPU6050 was used to monitor the mechanical response of the moving platform through acceleration measurements [27]. Although the sensor provides acceleration measurements along three axes, only the X-axis was used because, according to the sensor mounting orientation, it was aligned with the direction of the platform motion.

Before data acquisition, a calibration procedure is performed while the platform remains stationary for $N = 80$ consecutive samples. The baseline of the axis of interest and the calibrated acceleration variation used in the analysis are

$$A_{x0} = \frac{1}{N}\sum_{i=1}^{N} A_x(i), \qquad \Delta_x = A_x - A_{x0}, \tag{3}$$

where A_{x0} represents the sensor baseline (gravitational component plus offset) and Δ_x is the motion-related acceleration along the X-axis; the same procedure applies to the remaining axes.

To reduce noise while preserving the low-frequency components associated with respiratory motion, two low-pass filters are employed. The first is the MPU6050's internal digital low-pass filter, configured with a cutoff frequency of 60 Hz. The second is a first-order infinite impulse response (IIR) filter implemented in the firmware using an Exponential Moving Average (EMA), defined as

$$y[n] = y[n-1] + \alpha\,(x[n] - y[n-1])\,, \tag{4}$$

where $x[n]$ is the current sample, $y[n]$ is the filtered output, and α is the filter coefficient. The cutoff frequency of this filter was set to 2.5 Hz, preserving the typical respiratory frequency range (0.2–0.5 Hz), which may reach approximately 1 Hz in children and neonates [28].

D. Software

A Python-based software application, named *Breathing Platform Monitor*, was developed to integrate the main functionalities of the proposed system, including device configuration, platform control, signal acquisition, real-time visualization, and experimental data storage. The graphical user interface (GUI) was implemented using the PyQt5 library, while signal visualization was performed with PyQtGraph. These libraries were selected because of their suitability for developing graphical applications and their seamless integration with C++-based systems [29].

The software is organized into four tabs, with all communication performed through USB serial interfaces, as illustrated in Fig. 2: (i) serial port detection and device configuration, (ii) loading and playback of the respiratory CSV signal with individual servo testing, (iii) simultaneous acquisition and storage of the radar and MPU6050 signals, and (iv) loading of acquired files for preliminary RR visualization and analysis. The quantitative comparisons reported below were performed offline using the recorded data.

E. Results Analysis

A Google Colab notebook was used for offline analysis of 12 recordings, each approximately 240 s long. The CSV respiratory waveform used to command the servos was treated as the frequency reference; the MPU6050 X-axis and the filtered radar stream were compared with this waveform. Raw and centroid-based radar signals were excluded from the RR tables.

All analyzed signals were placed on a uniform 20 Hz time grid and preprocessed with detrending followed by a third-order Butterworth band-pass filter from 0.10 to 0.70 Hz (6–42 bpm) and z-score normalization. For FFT estimation, a Hann window was applied, the spectrum was zero-padded by a factor of four to the next power of two, and the dominant peak within the respiratory band was refined by parabolic interpolation. Welch estimation used 16 s segments (or the available signal length for shorter records). The other methods were sinusoidal fitting, autocorrelation, positive-going zero crossings, and adaptive peak detection.

The FFT window-length analysis used 15, 20, 30, 45, 60, 90, and 120 s windows with 50% overlap. Method comparison was performed separately on the complete approximately 240 s recordings. For each estimator m, the same estimator was applied to the radar and to the CSV reference, and the absolute error was defined as

$$E_{RR,m} = \left|RR_{\mathrm{radar},m} - RR_{\mathrm{CSV},m}\right|. \tag{5}$$

The same estimator-specific comparison was used for the MPU6050. Uncertainty across recordings was summarized using nonparametric bootstrap confidence intervals; for overlapping-window analyses, resampling was performed at the BIDMC-record level rather than treating overlapping windows as independent. As an exploratory waveform-level comparison, radar and CSV signals were additionally evaluated using absolute zero-lag correlation, maximum absolute cross-correlation within ± 10 s, and respiratory-band spectral coherence. These analyses assess recovery of the imposed respiratory dynamics and do not quantify absolute displacement fidelity.

IV. Results

The results are presented in five parts: (i) system overview, (ii) reproduction of the imposed respiratory frequency, (iii) Acconeer radar tracking, (iv) RR estimation, and (v) agreement analysis.

A. System Overview

A total of 12 respiratory recordings, each approximately 240 s long, were evaluated. The CSV waveform drove the platform motion, the XM132 served as the primary sensor for RR estimation, and the MPU6050 provided a complementary measurement of platform motion. The radar's internal tracking reference reported target-distance estimates of approximately 0.394–0.460 m, with a 70 mm ROI.

Fig. 3 shows the *Breathing Platform Monitor* interface, organized into four tabs corresponding to the experimental workflow. As summarized in the caption of Fig. 3, the tabs provide, respectively: (a) automatic USB port detection and device configuration; (b) loading, previewing and transmitting a respiratory CSV signal to the ESP32, including full-signal and windowed views; (c) starting and stopping the simultaneous acquisition of the radar and the MPU6050, with device status and real-time visualization; and (d) loading the acquired files for automatic RR estimation and visualization of the analysis plots.

B. Respiratory Frequency Reproduction

Table I summarizes the agreement between the reference respiratory signal and the MPU6050 measurements of respiratory rate, in breaths per minute (bpm). Both spectral methods yielded very low estimation errors, suggesting that the mechatronic platform reproduced the dominant respiratory frequency imposed by the reference signal. It should be noted that this agreement concerns the reproduced frequency and not the absolute displacement amplitude, since the MPU6050 is rigidly attached to the moving structure and does not constitute an independent displacement reference.

This low error was expected, since the MPU6050 was fixed to the moving platform and therefore directly measured its mechanical response. The zero Welch error indicates selection of the same discrete Welch frequency bin in the reference and MPU6050 signals; it should not be interpreted as zero physical or displacement error.

TABLE I
Comparison between the reference signal and the MPU6050 measurement.

Method	Mean error (bpm)	Median error (bpm)
FFT	0.08	0.006
Welch	0.00	0.00

C. Acconeer Radar Tracking

The filtered Acconeer radar signal exhibited, for most recordings, a dominant spectral component corresponding to the respiratory frequency imposed by the mechatronic platform. Spectral comparison with the reference signal and the MPU6050 measurement showed good agreement among the dominant peaks within the respiratory frequency range, suggesting that the radar was able to track the imposed respiratory frequency in most recordings under controlled experimental conditions. Representative examples of this spectral behavior are shown in Fig. 4.

An exception was observed, as illustrated in Fig. 4(b), in which both the reference CSV signal and the MPU6050 estimated an RR of approximately 18.5 bpm, whereas the radar FFT selected 36.2 bpm. The corresponding recording retained 100% tracking status with no rejected jumps, supporting interpretation of this event as spectral harmonic selection rather than loss of target tracking. Across all recordings, tracking status was 100% in 11 of 12 trials and 99.98% in the remaining trial; only one rejected-jump event was logged in the complete dataset. No NaNs were present in the key analyzed streams. These observations characterize the stability of the recorded tracking output but do not establish source-frame packet-loss performance or absolute distance accuracy.

D. RR Estimation

RR estimation was evaluated using different analysis window lengths in order to investigate the trade-off between temporal resolution and spectral stability. The results are summarized in Table II.

TABLE II
Radar FFT error versus window length (50% overlap).

Window (s)	N windows	Mean error (bpm)	Median error (bpm)
15	360	2.93	0.54
20	264	2.47	0.31
30	168	2.26	0.19
45	108	1.93	0.12
60	72	1.94	0.09
90	48	1.47	0.05
120	24	2.95	0.03

Median error decreased as window length increased, while the mean error was not monotonic because a small number of harmonic-selection errors produced a heavy upper tail. The lowest mean error occurred at 90 s. The 60 s window is reported as an intermediate time-resolved configuration, providing more frequent RR estimates than 90 s, but it should not be interpreted as an optimal window because no objective



Fig. 3. Breathing platform software: (a) device detection, (b) loading and transmitting respiratory signals from the dataset, (c) platform actuation and simultaneous signal acquisition, and (d) results analysis.

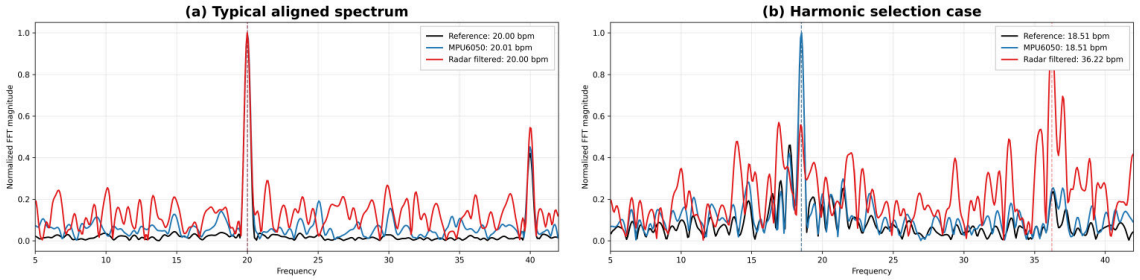


Fig. 4. Normalized spectral comparison between the reference respiratory signal, the auxiliary MPU6050 measurement, and the filtered Acconeer radar signal. FFT magnitudes were normalized to their maximum value to emphasize the dominant frequency components. (a) Representative recording showing agreement among the dominant frequencies of the reference signal, the MPU6050, and the radar. (b) Critical case in which the radar selected a harmonic component at approximately twice the reference respiratory frequency.

criterion combining estimation error and temporal resolution was defined. Record-clustered bootstrap 95% confidence intervals for the mean absolute error were 0.31–4.10 bpm at 60 s and 0.16–3.32 bpm at 90 s; the corresponding median-error intervals were 0.05–0.13 and 0.04–0.08 bpm. Because adjacent windows overlap by 50%, the window counts in Table II must not be interpreted as independent observations. The separate method comparison below uses the complete approximately 240 s recordings.

Different RR estimation methods were also compared using the complete recordings. The corresponding results are presented in Table III.

Using the same estimator on radar and reference signals, FFT and Welch yielded lower radar–reference disagreement than the time-domain approaches, while sinusoidal fitting produced a similarly low mean error. FFT was retained as the primary estimator for consistency with the window-length analysis; no formal computational-complexity benchmark or objective optimization criterion was used to rank the methods.

TABLE III
COMPARISON OF RESPIRATORY-RATE ESTIMATION METHODS FOR THE FILTERED RADAR SIGNAL.

Method	Mean error (bpm)	Median error (bpm)
FFT	1.49	0.03
Welch	1.56	0.00
Sine fit	1.53	0.01
Autocorrelation	3.39	0.21
Zero-crossing	3.59	1.26
Peak detection	5.42	6.01

Autocorrelation, zero-crossing, and peak detection were more sensitive to waveform irregularities in these recordings.

E. Agreement Analysis

Agreement was evaluated across the 12 recordings by comparing estimator-matched RR values from the MPU6050 and filtered radar stream with the CSV command waveform (Tables I and III). For the complete recordings, FFT absolute error was at most 0.038 bpm in 11 of 12 trials; one recording was the exception, with an error of 17.70 bpm caused by selection of a component near twice the reference frequency. Consequently, the 240 s FFT mean absolute error was 1.49 bpm (bootstrap 95% CI: 0.01–4.44 bpm), whereas the median was 0.027 bpm (95% CI: 0.006–0.034 bpm). This wide separation between mean and median emphasizes the sensitivity of the small sample to a single harmonic-selection failure.

The exploratory trajectory analysis provided additional information beyond dominant frequency. Across recordings, the mean and median absolute zero-lag correlations between radar and reference were 0.522 and 0.620; allowing a lag of up to ± 10 s increased the maximum absolute cross-correlation to 0.711 and 0.743, respectively. Mean and median maximum coherence within 0.10–0.70 Hz were 0.821 and 0.842. These values indicate shared respiratory-band temporal structure, but they do not constitute an independent validation of mechanical displacement because the CSV waveform is also the actuator command.

The MPU6050 likewise exhibited low RR error, consistent with its rigid attachment to the moving structure. Overall, the results support preliminary benchtop characterization of RR recovery under the tested conditions, not clinical validation or calibrated displacement accuracy.

V. Discussion

The Acconeer XM132, operated with the *Envelope* service and analyzed after documented respiratory-band preprocessing, recovered the dominant imposed respiratory frequency in most complete recordings. The 0.03 bpm median FFT error should nevertheless be interpreted cautiously: it was obtained from approximately 240 s records, for which long observation time improves spectral discrimination and can obscure transient variation. The 1.49 bpm mean was dominated by a single second-harmonic selection event, despite uninterrupted tracking in the corresponding recording. Thus, target tracking alone is insufficient to guarantee correct identification of the fundamental respiratory component.

FFT and Welch produced lower disagreement than the tested time-domain estimators, while sine fitting performed similarly to the spectral estimators. The window analysis does not support calling 60 s an optimum: 90 s yielded the lowest mean error, whereas 60 s provides more frequent RR estimates. The appropriate window therefore depends on the intended temporal resolution and error tolerance.

The methodological novelty should also be interpreted relative to prior phantom work. Recorded physiological waveforms have been used previously, including thoracic-impedance-driven radar phantoms [16]. The contribution here is the compact integration of public BIDMC waveforms with an XM132 PCR sensor, auxiliary inertial sensing, and an analysis framework that exposes window-length, estimator, tracking, and harmonic-selection behavior. The MPU6050 result (FFT mean error 0.08 bpm) supports reproduction of the imposed frequency, but not absolute displacement fidelity because the sensor is rigidly attached to the mechanism.

Several limitations remain. Motion is one-dimensional, whereas human chest-wall displacement is spatially distributed. No independent linear encoder, laser, or optical reference was available; consequently, the actual motion amplitude, mechanical resolution, hysteresis, backlash, and run-to-run repeatability were not quantified. The nominal 20 mm mechanism range and radar-derived 0.394–0.460 m tracking distances must therefore not be interpreted as calibrated displacement measurements. Although no obvious flat-top behavior was evident in the stored radar-amplitude traces, the logs do not contain a formal saturation flag or full-scale threshold, so saturation cannot be excluded quantitatively. Likewise, the merged asynchronous logs do not permit a defensible estimate of source-frame packet loss. A dedicated stationary acquisition would be needed to quantify a radar noise floor.

The sample comprises only 12 distinct recordings; bootstrap intervals and the per-record behavior reported above illustrate the resulting uncertainty and sensitivity to outliers. In addition, repeated runs of the same waveform were not performed, and the experiments did not constitute dedicated challenges for apnea, tachypnea, irregular breathing, controlled motion artifacts, or human subjects. These are therefore validation gaps rather than demonstrated capabilities. The platform can nevertheless be extended to compare alternative RR estimators, acquisition settings, and radar modules under the same reproducible input framework.

Three practical lessons emerge. First, physiologically derived waveforms do not eliminate spectral ambiguity, as demonstrated by the second-harmonic error. Second, increasing the analysis window generally reduces median error but sacrifices temporal resolution, while mean error can remain non-monotonic because of sparse large errors. Third, close RR agreement does not establish mechanical fidelity when the comparison waveform is also the actuator command. Future work should prioritize independent displacement sensing, formal repeated-run characterization, broader respiratory and motion-artifact protocols, explicit noise/saturation characterization, robust harmonic rejection, and subsequent validation in volunteers or patients.

VI. Conclusion

This work presented a mechatronic platform driven by thoracic impedance signals for the experimental evaluation of the Acconeer XM132 radar for respiratory monitoring. The platform integrates servo motors, an MPU6050 inertial sensor, and dedicated acquisition software, and was evaluated using 12 recordings from the publicly available BIDMC dataset.

Using FFT, the radar yielded a median RR error of $0.09\,\mathrm{bpm}$ and a mean error of $1.94\,\mathrm{bpm}$ with 60 s, 50%-overlapped windows. For complete approximately $240\,\mathrm{s}$ recordings, the mean and median errors were $1.49\,\mathrm{bpm}$ and $0.03\,\mathrm{bpm}$; one second-harmonic selection error accounted for most of the mean error. The study therefore supports the platform as a controlled proof-of-concept environment for preliminary radar and RR-processing characterization. It does not establish calibrated mechanical displacement accuracy or clinical validity. Broader validation requires a larger and repeated signal set, an independent displacement reference, mechanical-amplitude and repeatability characterization, explicit noise/saturation tests, improved harmonic rejection, and subsequent experiments in volunteers or patients.

ACKNOWLEDGMENT

The authors gratefully acknowledge the SENAI Innovation Institute for Embedded Systems for its institutional support and for providing the facilities and infrastructure required to conduct this research. This study was financed in part by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001.

REFERENCES

[1] P. Theerawit, Y. Sutherasan, L. Ball, and P. Pelosi, "Respiratory monitoring in adult intensive care unit," *Expert Review of Respiratory Medicine*, vol. 11, no. 6, pp. 453–468, 2017, doi: 10.1080/17476348.2017.1325324.

[2] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The Importance of Respiratory Rate Monitoring: From Healthcare to Sport and Exercise," *Sensors*, vol. 20, no. 21, Art. no. 6396, 2020, doi: 10.3390/s20216396.

[3] R. Boulding, R. Stacey, R. Niven, and S. J. Fowler, "Dysfunctional breathing: a review of the literature and proposal for classification," *European Respiratory Review*, vol. 25, no. 141, pp. 287–294, 2016, doi: 10.1183/16000617.0088-2015.

[4] D. Vitazkova et al., "Advances in Respiratory Monitoring: A Comprehensive Review of Wearable and Remote Technologies," *Biosensors*, vol. 14, no. 2, Art. no. 90, 2024, doi: 10.3390/bios14020090.

[5] T. Hussain, S. Ullah, R. Fernández-García, and I. Gil, "Wearable Sensors for Respiration Monitoring: A Review," *Sensors*, vol. 23, no. 17, Art. no. 7518, 2023, doi: 10.3390/s23177518.

[6] S. Toften et al., "Contactless and continuous monitoring of respiratory rate in a hospital ward: a clinical validation study," *Frontiers in Physiology*, vol. 15, Art. no. 1502413, 2024, doi: 10.3389/fphys.2024.1502413.

[7] C. Massaroni, A. Nicolò, M. Sacchetti, and E. Schena, "Contactless methods for measuring respiratory rate: A review," *IEEE Sensors Journal*, vol. 21, no. 11, pp. 12821–12839, 2021, doi: 10.1109/JSEN.2020.3023486.

[8] O. Testa, R. Cicchetti, S. Pisa, E. Pittella, and E. Piuzzi, "A Novel FMCW Radar Scheme with Millimeter Motion Detection Capabilities Suitable for Cardio-Respiratory Monitoring," *Sensors*, vol. 25, no. 9, Art. no. 2765, 2025, doi: 10.3390/s25092765.

[9] Z. Hao, Y. Wang, F. Li, G. Ding, and Y. Gao, "mmWave-RM: A Respiration Monitoring and Pattern Classification System Based on mmWave Radar," *Sensors*, vol. 24, no. 13, Art. no. 4315, 2024, doi: 10.3390/s24134315.

[10] A. Fallatah et al., "Monitoring Respiratory Motion during VMAT Treatment Delivery Using Ultra-Wideband Radar," *Sensors*, vol. 22, no. 6, Art. no. 2287, 2022, doi: 10.3390/s22062287.

[11] T. Lauteslager et al., "Validation of a New Contactless and Continuous Respiratory Rate Monitoring Device Based on Ultra-Wideband Radar Technology," *Sensors*, vol. 21, no. 12, Art. no. 4027, 2021, doi: 10.3390/s21124027.

[12] J. D. Kim et al., "Non-contact respiration monitoring using impulse radio ultrawideband radar in neonates," *Royal Society Open Science*, vol. 6, no. 6, Art. no. 190149, 2019, doi: 10.1098/rsos.190149.

[13] Acconeer AB, "Pulsed Coherent Radar Technology Overview," Acconeer Documentation, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/pcr_tech/overview.html

[14] Acconeer AB, "A111 System Overview," Acconeer Documentation, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/a111/radar_data_and_control/system_overview.html. [Accessed: Jun. 17, 2026].

[15] Acconeer AB, "Exploration Tool," Acconeer Documentation, 2026. [Online]. Available: https://docs.acconeer.com/en/latest/exploration_tool/index.html

[16] Y. Eder, E. Zagoury, S. Savariego, M. Namer, O. Cohen, and Y. C. Eldar, "Robust Phantom-Assisted Framework for Multi-Person Localization and Vital Signs Monitoring Using MIMO FMCW Radar," *IEEE Transactions on Mobile Computing*, vol. 25, no. 6, pp. 9120–9136, 2026, doi: 10.1109/TMC.2026.3651442.

[17] M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, and D. A. Clifton, "BIDMC PPG and Respiration Dataset," *PhysioNet*, version 1.0.0, 2018, doi: 10.13026/C2208R.

[18] M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2017, doi: 10.1109/TBME.2016.2613124.

[19] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, doi: 10.1161/01.CIR.101.23.e215.

[20] K. Sel, D. Osman, and R. Jafari, "Non-Invasive Cardiac and Respiratory Activity Assessment From Various Human Body Locations Using Bioimpedance," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 2, pp. 210–217, 2021, doi: 10.1109/OJEMB.2021.3085482.

[21] P. H. Charlton et al., "An impedance pneumography signal quality index: Design, assessment and application to respiratory rate monitoring," *Biomedical Signal Processing and Control*, vol. 65, Art. no. 102339, 2021, doi: 10.1016/j.bspc.2020.102339.

[22] G. B. Drummond, D. Fischer, M. Lees, A. Bates, J. Mann, and D. K. Arvind, "Classifying signals from a wearable accelerometer device to measure respiratory rate," *ERJ Open Research*, vol. 7, no. 2, Art. no. 00681-2020, 2021, doi: 10.1183/23120541.00681-2020.

[23] E.-K. Wu, Q.-G. Fan, M.-C. Li, J.-H. Zhang, J. Jia, T. Qiang, C. Wang, X.-F. Gu, and J.-G. Liang, "Non-contact monitoring of human cardiorespiratory activity during sleep using FMCW millimeter wave radar," *Measurement*, vol. 242, Part D, Art. no. 116144, 2025, doi: 10.1016/j.measurement.2024.116144.

[24] A. U. Singh et al., "Modelling of Chest Wall Motion for Cardiorespiratory Activity for Radar-Based NCVS Systems," *Sensors*, vol. 20, no. 18, Art. no. 5094, 2020, doi: 10.3390/s20185094.

[25] S. Fan and Z. Deng, "Chest Wall Motion Model of Cardiac Activity for Radar-Based Vital-Sign-Detection System," *Sensors*, vol. 24, no. 7, Art. no. 2058, 2024, doi: 10.3390/s24072058.

[26] O. M. Steihaug et al., "Clinical validation of a novel accelerometer-based respiratory rate monitor," *Journal of Clinical Monitoring and Computing*, vol. 40, no. 3, pp. 695–702, 2026, doi: 10.1007/s10877-026-01419-x.

[27] InvenSense, "MPU-6000 and MPU-6050 Product Specification," Rev. 3.4, Aug. 2013.

[28] P. Arlotto, M. Grimaldi, R. Naeck, and J.-M. Ginoux, "An Ultrasonic Contactless Sensor for Breathing Monitoring," *Sensors*, vol. 14, no. 8, pp. 15371–15386, 2014, doi: 10.3390/s140815371.

[29] Riverbank Computing Limited, "PyQt5: Python bindings for the Qt cross platform application toolkit," *Python Package Index (PyPI)*, 2021. [Online]. Available: https://pypi.org/project/PyQt5. [Accessed: Jun. 16, 2026].

Medição Não Invasiva de Pressão Arterial por Ultrassom em Modo A-Scan

1st Maicol Peterson Gandolphi de Almeida
Dep. Engenharia Elétrica
Universidade do Estado de Santa Catarina
Joinville, Brasil
https://orcid.org/0000-0001-8849-7717

2nd Pedro Bertemes Filho
Dep. Engenharia Elétrica
Universidade do Estado de Santa Catarina
Joinville, Brasil
pedro.bertemes@udesc.br

***Resumo*—O monitoramento contínuo e não invasivo de parâmetros cardiovasculares ainda depende na prática clínica de profissionais especializados e equipamentos de alto custo. Este artigo apresenta uma prova de conceito de um sistema embarcado desenvolvido para a medição da distensão vascular utilizando um sensor ultrassônico em modo A-Scan. A arquitetura de hardware emprega um microcontrolador STM32 operando em bare-metal, um conversor analógico-digital de 12 bits a 3 MSPS (ADS7883), um amplificador operacional OPA350 no front-end analógico e um transdutor piezoelétrico de 8 mm $\times$ 1 mm excitado por um driver MOSFET TC4420. Os ecos ultrassônicos captados são transmitidos via comunicação serial para um ambiente Python, onde o envelope do sinal é extraído e armazenado em arquivo CSV para a visualização da variação das paredes da artéria radial. Embora modelos biomecânicos elásticos exponenciais correlacionem o diâmetro vascular à pressão arterial, este estudo delimita-se à validação técnica da eletrônica de aquisição sem calibração clínica. Ensaios técnicos de bancada com diferentes tensões de excitação (5, 10, 15 e 20 V) demonstraram que pulsos de 10 V produzem sinais de eco com melhor relação sinal-ruído, evitando atenuação excessiva e distorções. Os resultados confirmam a viabilidade do hardware proposto como instrumentação de baixo custo e etapa fundamental para o futuro desenvolvimento de sistemas de monitoramento contínuo.**

***Palavras chaves*—Pressão arterial, ultrassom, envelope analógico.**

I. Introdução

Sensores e transdutores atuam como primeira camada de interface entre o organismo vivo e o sistema de aquisição eletrônica, onde a função primordial do transdutor é a conversão de uma grandeza física ou bioquímica em um sinal elétrico analógico equivalente [1], [2].

No monitoramento hemodinâmico por métodos acústicos o elemento piezoelétrico quando submetido a um estresse mecânico sofre deslocamento de cargas, gerando uma diferença de potencial elétrico utilizável (efeito direto). A aplicação de um pulso de tensão elétrica de curta duração deforma mecanicamente o cristal e induz a emissão de uma onda de pressão acústica na pele (efeito inverso) [1], [3].

A escolha geométrica e material do transdutor define sua frequência fundamental de ressonância e o comprimento de onda ideal para obter alta resolução na discriminação de interfaces biológicas milimétricas e superficiais, como as paredes vasculares. O ultrassom em Modo-A (Amplitude Mode) representa a forma mais pura e unidimensional da ecografia pulsada. O princípio operacional baseia-se no método de tempo de voo (Time-of-Flight - ToF), onde o sistema transmite um pulso ultrassônico transiente para o interior do tecido e passa a operar em modo de recepção, aguardando os ecos gerados nas transições entre meios de diferentes impedâncias acústicas [1]–[3].

No monitoramento da artéria radial o processamento desse sinal analógico revela dois picos móveis distintos correspondentes às interfaces da parede anterior e posterior do vaso [2], [4]. O rastreamento dinâmico da distância temporal entre esses picos viabiliza a mensuração micrométrica da oscilação do diâmetro arterial ao longo do ciclo cardíaco, considerando a velocidade nominal do som no tecido mole de 1540 m/s [1], [4].

A artéria radial está localizada superficialmente na região do punho, sendo uma região anatômica ideal para o monitoramento hemodinâmico não invasivo. Durante o ciclo cardíaco a propagação da onda de pulso exerce uma pressão hidrostática contra as paredes do vaso, induzindo uma deformação que devido às propriedades viscoelásticas da parede arterial apresenta uma forte correlação matemática com as variações da pressão intravascular. Consequentemente, a captação dessa variação de diâmetro viabilizada por técnicas de ultrassonografia permite modelar a dinâmica vascular e traduzir o comportamento mecânico da artéria [5].

Para transformar a distensão geométrica em milímetros de mercúrio (mmHg), usam-se modelos baseados na complacência arterial (a relação entre variação de volume e variação de pressão) [1], [3]. Antes de aplicar as equações é preciso aplicar uma janela de calibração, ou seja, no mesmo momento que o esfigmomanômetro mede a pressão do paciente, o software de ultrassom registra os diâmetros da artéria, obtendo-se quatro constantes: pressão sistólica e diastólica medidas pelo esfigmomanômetro manual, diâmetro máximo e mínimo medidos através do sistema ultrassônico.

O modelo exponencial (1) é uma das abordagens matemáticas mais precisas usadas na literatura biomédica para calcular a pressão instantânea a partir do diâmetro instantâneo [5].

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)

$$P(t) = P_d \cdot \exp\left(\beta\left(\frac{D(t)}{D_d} - 1\right)\right) \quad (1)$$

em que $P(t)$ é a pressão arterial instantânea, P_d é a pressão arterial diastólica medida com esfigmomanômetro, β é o índice de rigidez arterial e D_d é o diâmetro mínimo da artéria durante a diástole (medido com o sensor durante a captura dos sinais).

A parede arterial enrijece conforme dilata e esse comportamento é modelado pela relação de Hughes [5] através de um índice de rigidez β calculado em (2).

$$\beta = \frac{\ln\left(\frac{P_s}{P_d}\right)}{\frac{D_s}{D_d} - 1} \quad (2)$$

em que P_s é a pressão sistólica medida com esfigmomanômetro e D_s é o diâmetro maior durante a sístole (medido com o sensor durante a captura dos sinais).

II. Materiais e Métodos

A. Arquitetura de Hardware e Seleção de Componentes

O sistema de aquisição de sinal de ultrassom em modo *A-Scan* foi projetado como uma prova de conceito para capturar a dinâmica geométrica de distensão que simula o comportamento da artéria radial.

Utilizou-se um transdutor na forma de disco piezoelétrico de Titanato Zirconato de Chumbo (PZT) com 8 mm de diâmetro e 1 mm de espessura. A geometria do cristal foi definida visando duas características fundamentais: a espessura de 1 mm determina uma frequência de ressonância natural na faixa de 2 MHz (considerando a velocidade do som no material), enquanto o diâmetro de 8 mm é adequado para focar o feixe acústico.

Para a excitação do transdutor empregou-se o *driver* de MOSFET TC4420 devido à sua capacidade de fornecer picos de corrente elevados com tempos de subida e descida na ordem de 25 ns. Essa característica é crítica para carregar e descarregar abruptamente a capacitância do disco piezoelétrico, convertendo o sinal lógico de baixa potência do microcontrolador em um pulso de excitação mecânica (onda de choque acústica) de banda larga. Um módulo *Step-UP* fornece tensões entre 5 V e 20 V para o TC4420, enquanto os demais componentes são alimentados com 3.3 V do módulo STM32.

Devido ao eco ultrassônico possuir uma amplitude na ordem de milivolts, utilizou-se um amplificador operacional OPA350 com largura de banda 38 MHz e taxa de variação (*slew rate*) de 22 V/μs. Estas especificações garantem a amplificação linear do sinal do ultrassom sem introduzir distorção de fase ou atenuação indesejada dos picos.

A conversão analógico-digital foi realizada pelo circuito integrado ADS7883, um ADC de arquitetura de aproximações sucessivas (SAR) operando a 3 MSPS (*Mega Samples Per Second*) com resolução de 12 bits. Esta taxa de amostragem atende aos requisitos de captura de transientes rápidos do sinal condicionado, enquanto a resolução de 12 bits (4096 níveis de quantização) confere a faixa dinâmica necessária para discriminar sutis variações de impedância acústica.

Para o módulo microcontrolador, optou-se pelo modelo STM32 NUCLEO-F767ZI em detrimento de arquiteturas voltadas à IoT, como o ESP32. O rastreamento de distensão por ultrassom depende estritamente do tempo de voo (*time-of-flight*), onde atrasos de frações de microssegundos inviabilizam a medição. Enquanto o ESP32 executa operações sobrepostas a um Sistema Operacional de Tempo Real (RTOS) encarregado da pilha de rede, o que introduz um *jitter* (variação de latência) inaceitável nas temporizações, o STM32 (Cortex-M7 a 216 MHz) operando *bare-metal* (sem sistema operacional base) assegura o determinismo do sistema. O uso de temporizadores de hardware e recursos de Acesso Direto à Memória (DMA) garante precisão na comunicação SPI e na temporização dos disparos.

B. Validação de Bancada, Aquisição e Processamento de Sinal

Para validar a instrumentação eletrônica e a resposta do algoritmo em um cenário de aplicação, testes preliminares de prova de conceito foram conduzidos por meio de autoexperimentação ($n = 1$). O transdutor foi posicionado usando uma película fina de gelatina neutra sobre a região da artéria radial no pulso do próprio autor e presa através de elástico para manter a mesma pressão em todas as medições, utilizando-a estritamente como um alvo físico para atestar a viabilidade da medição do tempo de voo acústico e das variações geométricas em tempo real. Por se tratar exclusivamente de uma validação técnica de hardware e calibração de sinais, sem caráter de ensaio clínico, análise populacional ou intervenção de risco, o procedimento caracterizou-se como teste de bancada preliminar, dispensando a submissão a um comitê de ética.

A dinâmica operacional do *firmware* no STM32 é estruturada por interrupções e requisições seriais [8]. Ao receber um comando via UART (*Universal Asynchronous Receiver-Transmitter*), o microcontrolador aciona um pino GPIO que controla a emissão do pulso elétrico no TC4420. Após um tempo de supressão (*blanking time*), projetado para ignorar a saturação do pulso principal de emissão (*main bang*), o microcontrolador inicia a leitura no barramento SPI em alta velocidade, coletando um *frame* de 1000 amostras provenientes do ADS7883. Esse pacote de dados é alocado em memória via DMA, envelopado com um cabeçalho hexadecimal de sincronização e transmitido para o computador a uma taxa de transferência de 460.800 bps.

No ambiente Python um programa realiza a recepção em tempo real e outro faz o processamento do arquivo CSV para plotagem [8]. O *script* monitora o fluxo serial, isolando os quadros de 1000 amostras a partir do reconhecimento do cabeçalho de sincronização, garantindo a integridade dos dados sem rupturas ou perdas de alinhamento estrutural (*frame shift*). Uma vez extraído o vetor de 12 bits, o processamento digital de sinais (DSP) ocorre em quatro etapas sucessivas:

1) **Filtragem:** O sinal sofre a aplicação de um filtro passa-baixa Butterworth para mitigação de ruído térmico de alta frequência e suavização da curva.

2) **Demodulação:** Aplica-se a Transformada Analítica de Hilbert, cujo valor absoluto do vetor complexo extrai o envelope de energia acústica, suprimindo as portadoras indesejadas.
3) **Detecção Estrutural:** Um algoritmo de localização de picos avalia o envelope de ecos respeitando restrições projetadas para a anatomia vascular simulada. São exigidas proeminências específicas de amplitude e distâncias morfológicas aceitáveis (1.5 mm a 8.0 mm) para categorizar estritamente o reflexo das superfícies anterior e posterior do alvo.
4) **Mapeamento Gráfico:** A distância entre os alvos é calculada em milímetros com base no vetor calibrado pela velocidade média do som em tecidos biológicos (1540 m/s). Converteu-se esse valor para milímetros por nanossegundos (0.00154 mm/ns) e através do tempo de amostragem do STM32 (692.6 ns) chega-se à constante de distância da amostra (0.5333 mm), significando que cada índice no eixo da profundidade representa o percurso do eco em exatos 0.5333 milímetros. A interface visual, implementada através da biblioteca *Matplotlib*, plota a distância pontual das paredes, renderizando a evolução temporal contínua em 2D na janela de monitoramento, o que permite a visualização dos ciclos de sístole e diástole a cada quadro gerado.

III. Resultados

O teste inicial foi feito com pulsos de 5 V e os dados brutos recebidos pelo programa Python foram processados para identificar os pontos da sístole e diástole. A Fig. 1 mostra o pico máximo (sístole) e mínimo (diástole) da artéria radial entre 3.0 e 5.5 segundos.

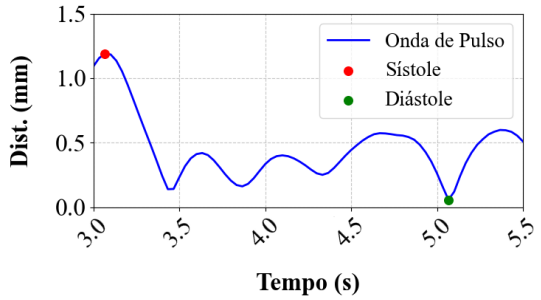


Fig. 1. Representação da sístole e diástole para pulsos de 5V.

O segundo teste foi realizado usando 10 V para emissão dos pulsos ultrassônicos, a Fig. 2 mostra o resultado com os pontos de sístole e diástole.

O terceiro teste usou uma tensão de pulso de 15 V e a Fig. 3 mostra o resultado com os pontos de sístole e diástole para esta tensão.

O quarto e último teste usou uma tensão de pulso de 20 V e a Fig. 4 mostra o resultado com os pontos de sístole e diástole para esta tensão.

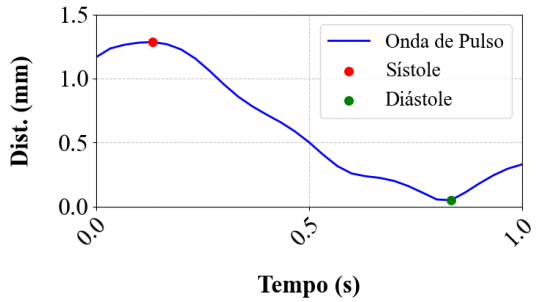


Fig. 2. Representação da sístole e diástole para pulso de 10V.

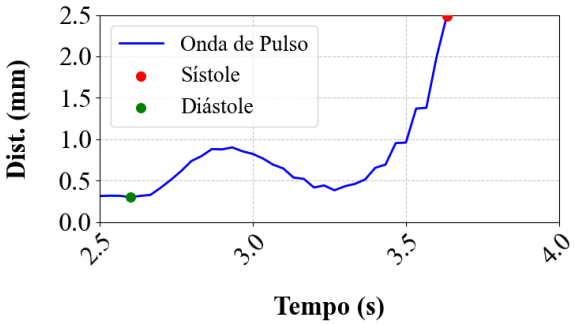


Fig. 3. Representação da sístole e diástole para pulso de 15V.

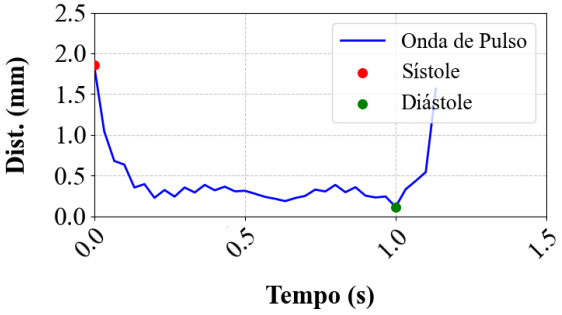


Fig. 4. Representação da sístole e diástole para pulso de 20V.

O gráfico da Fig. 1 não condiz com os gráficos hemodinâmicos encontrados na literatura [6], [7], indicando que uma tensão de 5 V não é adequada para a captação do eco ultrassônico.

Na Fig. 2 temos uma linha representativa de uma hemodinâmica com os pontos de sístole e diástole bem definidos e dentro de um intervalo condizente com a literatura.

Na Fig. 3 e Fig. 4 apesar do intervalo da sístole e diástole estarem dentro do intervalo de aproximadamente 1 segundo, percebe-se que a linha começa a apresentar ruídos e leves variações, possuindo um aspecto diferente da forma representativa de um gráfico ideal.

Para medir a pressão arterial em função do tempo, o valor máximo de D_s e o valor mínimo de D_d devem ser salvos e usando um esfigmomanômetro obtém-se P_s e P_d

padrão do indivíduo, substituindo estes valores em (1) e (2) aplica-se as equações no programa e obtém-se a pressão em função do tempo sem que seja necessário o uso contínuo do esfigmomanômetro.

IV. Discussões e Conclusão

Este artigo apresentou o desenvolvimento e a validação de uma prova de conceito de um sistema embarcado de ultrassom em modo *A-Scan* voltado para o monitoramento da distensão geométrica vascular. A arquitetura de hardware baseada em um microcontrolador STM32 operando *bare-metal* com determinismo temporal, aliada a um condicionamento analógico e ao processamento digital via Transformada de Hilbert, demonstrou ser capaz de rastrear a variação da artéria radial.

A análise comparativa das tensões de excitação do transdutor piezoelétrico revelou que a aplicação de pulsos de 10 V representa o ponto ideal de operação do sistema na bancada. Nessa configuração, o processamento do envelope acústico extraiu a eco-reflexão do alvo, apresentando ciclos com picos de sístole e diástole bem definidos. Constatou-se empiricamente que tensões inferiores (5 V) são insuficientes para uma captação ecográfica adequada do sinal, enquanto tensões superiores (15 V e 20 V) introduzem distorções decorrentes da saturação do front-end analógico.

Os resultados confirmam a viabilidade técnica da instrumentação proposta, validando-a como uma alternativa de baixo custo, não oclusiva e de alta precisão temporal para sistemas de aquisição biomédica. A mensuração bem-sucedida da variação geométrica vascular estabelece a infraestrutura de dados necessária para futuras implementações de modelos biomecânicos de complacência arterial.

Como trabalhos futuros, planeja-se a integração de técnicas avançadas de filtragem e aprendizado de máquina no ambiente Python, bem como a estruturação de ensaios futuros com amostragem ampliada para avaliar a correlação dos sinais obtidos com padrões clínicos de referência.

Agradecimentos

Agradeço à Universidade do Estado de Santa Catarina (UDESC) pelo suporte institucional e à Fundação de Amparo a Pesquisa e Inovação do Estado de Santa Catarina (FAPESC), bem como à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro.

Referências

[1] WELLS, Peter Neil Temple. Biomedical Ultrasonics. IX. London: Academic Press, 1977.

[2] ASSEF, Amauri A. et al. Sistema para geração, aquisição e processamento de sinais de ultrassom. Sba: Controle & Automação Sociedade Brasileira de Automática, v. 20, p. 145-155, 2009.

[3] ZISKIN, Marvin. Applications of ultrasound in medicine: comparison with other modalities. Ultrasound: Medical Applications, Biological Effects and Hazard Potential. New York: Plenum Press, 49-60, 1987.

[4] MEINDERS, Jan M.; HOEKS, Arnold P. Clinical aortic root wall motion measurement using a novel high-resolution tracking system. Ultrasound in Medicine & Biology. v. 30, n. 4, p. 461-471, 2004.

[5] CHIRINOS, Julio A.; SEGERS, Patrick; HUGHES, Timothy; TOWNSEND, Raymond R. Large-Artery Stiffness in Health and Disease: JACC State-of-the-Art Review. Journal of the American College of Cardiology. v. 74, n. 9, p. 1237–1263, 2019. DOI: 10.1016/j.jacc.2019.07.012.

[6] VAPPOU, Jonathan et al. Non-invasive measurement of local pulse wave velocity and arterial stiffness in humans in vivo. Ultrasound in Medicine & Biology. v. 37, n. 1, p. 136-144, 2011.

[7] NITZAN, Meir et al. The variability of the systolic and diastolic blood pressure from the volume pulse characteristics. Medical & Biological Engineering & Computing. v. 40, n. 2, p. 175-179, 2002.

[8] STM32 and Python Source Codes. [Online]. Available: https://drive.google.com/drive/folders/1I3eVWjSzmmbwDByQhDutvDoreevqOYCP

Modelagem de sistema para análise do esforço respiratório natural via sensor de pressão piezoresistivo

1st João Gabriel de Souza Veras
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0000-0003-1351-4812

2nd Jorge Luis Vitorino
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0009-0006-6307-8018

3rd Juliana Lúcia dos S. Valois Corrêa de O.
Dept. de Engenharia Biomédica
Scientifical Solutions
Recife, Brasil
email: juliana.valois@scisol.com.br

4th Emery Cleiton Cabral Correia Lins
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0000-0002-7967-6163

***Abstract*—Alterações na frequência respiratória constituem um dos primeiros sinais de deterioração do estado clínico do paciente, o que evidencia a necessidade de soluções portáteis e não invasivas para seu monitoramento contínuo. Este trabalho teve como objetivo desenvolver um sistema de biofeedback respiratório composto por um sensor de pressão piezoresistivo integrado a um microcontrolador XIAO ESP32 C6 mini com comunicação Bluetooth Low Energy (BLE). A unidade sensora, constituída por microcontrolador, regulador de tensão DC-DC e módulo de sensor barométrico com conversor ADC de 24 bits, processa o sinal respiratório em tempo real e transmite os dados ao programa de análise para visualização da curva respiratória e geração de relatórios com frequência respiratória (FR). A calibração foi conduzida em ambiente controlado, com o protótipo conectado em série a um ventilador mecânico Maquet Servo Air e um pulmão artificial. O sistema demonstrou alta fidelidade na detecção da FR, com desvio padrão máximo de 0,83 rpm em relação aos parâmetros simulados. Os resultados confirmam a viabilidade da unidade sensora para aplicações de saúde preventiva em cenários dinâmicos do cotidiano.**

***Index Terms*—Frequência respiratória, Biofeedback, Sensor Piezoresistivo, pressão**

I. Introdução

A respiração é um processo fisiológico essencial para a manutenção da vida, resultante da ação coordenada dos músculos respiratórios e do centro de controle localizado no cérebro. Alterações no ritmo respiratório podem ocorrer como resposta fisiológica, por exemplo, durante o esforço físico, ou indicar condições clínicas relevantes, como insuficiência respiratória e doenças pulmonares obstrutivas crônicas (DPOC) [1], [2]. Em paralelo a essa demanda clínica, o avanço das tecnologias vestíveis (wearables) tem tornado o acompanhamento dos sinais vitais mais acessível, preciso e contínuo. Esses dispositivos, aliados a biossensores, permitem a captação de variáveis fisiológicas – como temperatura corporal, frequência cardíaca, movimento e fluxo respiratório – transformando-as em sinais elétricos ou ópticos que podem ser analisados por softwares especializados [3]. Esse avanço tecnológico encontra na Internet das Coisas (IoT) um potente elemento integrador, que conecta sensores a plataformas digitais e promove a coleta, transmissão e análise contínua de dados fisiológicos em tempo real mesmo fora do ambiente hospitalar [4]. Dentro desse ecossistema tecnológico, destaca-se o conceito de biofeedback, técnica que visa promover a autorregulação por meio do monitoramento e da devolução de informações fisiológicas em tempo real ao paciente. Essa estratégia é amplamente utilizada em programas de reabilitação física, controle do estresse, manejo de doenças crônicas e aprimoramento do desempenho cardiorrespiratório, inclusive em contextos esportivos [5]. Para a captação do sinal respiratório nesse tipo de aplicação, destacam-se os sensores piezoresistivos, notáveis por sua alta sensibilidade na conversão da tensão mecânica em sinal elétrico. Seu princípio de funcionamento reside na variação da resistência elétrica do material conforme ele é deformado, sendo esse processo reversível após cada ciclo respiratório, o que viabiliza um acompanhamento contínuo e preciso da respiração [6]. Diante do exposto, este trabalho teve como objetivo prototipar e calibrar um sistema de monitoramento e análise do fluxo e esforço respiratório, utilizando um sensor de pressão piezoresistivo para a aquisição de dados e modelado para integração em dispositivo vestível. O sistema visa permitir a avaliação contínua e a análise remota simplificada dos parâmetros respiratórios durante atividades de rotina, fornecendo suporte a soluções preventivas e terapêuticas em indivíduos com diferentes perfis respiratórios, tais como asma, DPOC, desvio de septo e apneia do sono, além de viabilizar a análise comparativa entre as curvas de respiração nasal e oral.

II. Materiais e Métodos

A. Calibração do Sensor

Foi prototipada uma unidade sensora, vista na Figura 1, utilizando o MPS20N0040D-S embutido em um módulo junto ao conversor analógico-digital HX710B de 24bits que utiliza o formato de Complemento de Dois na saída.

O MPS20N0040D-S é um sensor analógico piezoresistivo que converte a pressão do ar aplicado em tensão, a leitura da tensão chega ao conversor adc onde é digitalizada com ganho de 128V/V. Com isso, a equação geral do Bit Menos Significativo (LSB) é dado por:

$$LSB = \frac{2 \times V_{ref}}{Ganho \times 2^{24}} \tag{1}$$

No presente estudo, a tensão de referência (V_{ref}) foi de 5V, com isso, a conversão do dado bruto enviado pelo sensor para tensão de saída (V_{out}) em milivolts é dada por:

$$V_{out} = LSB \times DadoBruto \tag{2}$$

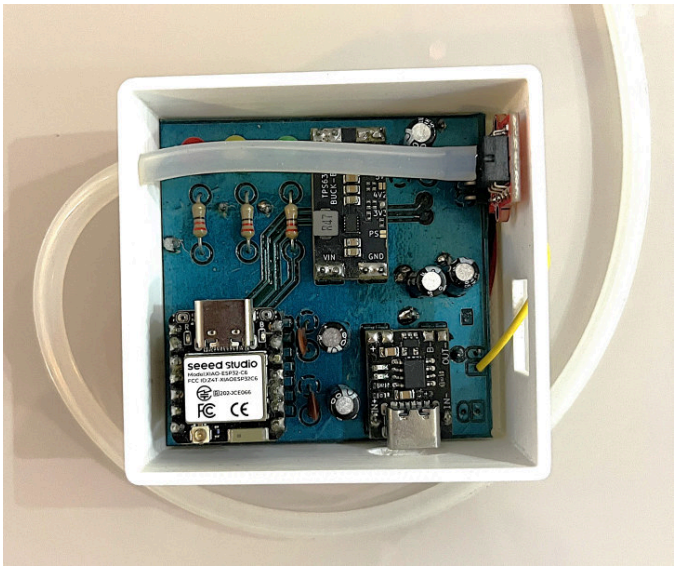


Fig. 1. Unidade sensora (protótipo)

Com essas informações sob registro, foi realizada uma sessão de leituras de pressão do fluxo respiratório utilizando a unidade sensora em série com um ventilador pulmonar MAQUET Servo Air e um pulmão artificial durante ciclos respiratórios de frequência fixa e pico de pressão positiva personalizada que variou de 7cmH2O à 22cmH2O, variando a cada +1cmH2O. A taxa de amostragem do sensor foi de 50ms e foram analisados 951 pontos para cada uma das pressões aplicadas.

Em seguida, foi realizada a identificação dos picos de tensão para cada intervalo associado à pressão aplicada pelo ventilador pulmonar e calculada a média, mediana, moda, variância, desvio padrão, intervalo de confiança, box-plot desses valores e a correlação de Pearson entre a tensão de saída do sensor e a pressão de fluxo aplicada. Após a análise das medidas de tendência central e dispersão, o valor da moda dos picos foi selecionada para melhor representar o comportamento do sensor à pressão associada e utilizado no cálculo de regressão linear para definir o comportamento da tensão de saída em função da pressão aplicada ao sensor.

B. Unidade Sensora

Foi desenvolvido um sistema embarcado autônomo controlado por um XIAO ESP32-C6 Mini responsável por tratar em tempo real as leituras de tensão realizadas pelo módulo do sensor MPS20N0040D-S e por criar um servidor Bluetooth 5.3 que estabelece as regras de comunicação com aplicativo móvel.

O microcontrolador possui dois núcleos de processamento que permite orquestrar de forma assíncrona as tarefas críticas de temporização e gerenciamento da pilha de dados Bluetooth (incluindo o controle físico de rádio, buffers de transmissão e coexistência de rede), o que garante a comunicação sem fio ocorrer sem causar atrasos de sincronização ou perda de pacotes, enquanto o outro se dedica a tratar dos dados do sensor. Isso permite que a comunicação não interfira na precisão e estabilidades das medições.

Visando blindar o sistema contra oscilações de energia, um regulador de tensão buck-boost atua na estabilização contínua da alimentação dos circuitos lógicos e analógicos. Esse regulador assegura que, mesmo com o declínio gradual da carga da bateria, o sensor e o microcontrolador operem sob uma tensão rigorosamente constante, eliminando desvios térmicos ou flutuações que comprometem a calibração e a precisão das leituras de pressão.

C. Modelagem do sistema

O sistema de biofeedback foi desenhado para operar por meio de uma arquitetura de malha fechada, viabilizada pela interação dinâmica entre três agentes principais: o usuário, a unidade sensora(hardware) e o programa de análise(software). O fluxo operacional inicia-se quando o usuário, por meio da interface gráfica do analisador, envia um comando de inicialização via Bluetooth Low Energy para a unidade sensora. Ao receber a instrução, o microcontrolador ativa a rotina de aquisição de dados, realizando a leitura analógica contínua do sensor de pressão respiratória.

Após o tratamento digital local dos sinais de tensão – que envolve a filtragem de ruídos e calibração básica –, as informações de pressão são empacotadas e transmitidas assincronamente de volta ao sistema de análise. No programa, os dados são processados para a renderização em tempo real da curva respiratória e para o cálculo instantâneo dos parâmetros associados.

O ciclo de biofeedback se fecha quando o usuário visualiza o gráfico atualizado continuamente na tela. Esse retorno visual imediato permite que o paciente regule de forma consciente e voluntária o seu esforço, frequência ou padrão respiratório, adaptando sua mecânica pulmonar aos alvos terapêuticos estabelecidos no aplicativo. Vale destacar que a curva positiva

remete ao ciclo expiratório (pressão positiva aplicada ao sensor) e a curva negativa remete ao ciclo inspiratório (pressão negativa).

III. Resultados e Discussão

A modelagem desenvolvida visa ser um método alternativo, confortável e versátil de avaliação da respiração natural – fluxo de ar pelas vias nasais, responsáveis por filtrar, aquecer e umedecer o ar inspirado para propiciar a oxigenação eficiente do organismo – em atividades cotidianas de baixa e média intensidade, cuja estabilidade e precisão permite, à posteriori, o estudo observacional de diferentes perfis respiratórios. Sua aplicação almeja apoiar soluções preventivas e terapêuticas que melhorem a qualidade de vida das pessoas sob diferentes quadros clínicos de restrição respiratória.

A. Calibração (Pressão x Tensão)

A resposta do sensor às diferentes pressões aplicadas pelo ventilador pulmonar foram estáveis, com exceção do pico de pressão à 22cmH2O em que ocorreu a saturação do sensor e sua leitura foi dispensada como dado censurado e excluída da análise. As três medidas de tendência central (média, mediana e moda) convergiram para valores praticamente idênticos em todos os níveis de pressão, com diferenças inferiores a 0,01 mV. Essa concordância indica que a distribuição dos picos de tensão dentro de cada intervalo é aproximadamente simétrica e livre de valores extremos (outliers).

Quanto à dispersão, o coeficiente de variação manteve-se abaixo de 0,14% em todos os 15 níveis de pressão testados, o que evidencia uma repetibilidade muito alta do sensor: ciclo a ciclo, sob uma mesma pressão fixa, a tensão de saída variou, no máximo, cerca de 0,14% em torno de sua média. Os intervalos de confiança de 95%, por sua vez, mostraram-se estreitos em todos os níveis (amplitude inferior a 0,02 mV), reforçando que a média estimada para cada pressão é precisa em relação ao valor esperado de tensão.

A relação entre a pressão aplicada e a tensão de saída do sensor foi modelada por regressão linear simples, resultando na equação de calibração: P = 1.5007V - 0,0181 (P em cmH2O e V em mV), com coeficiente de determinação $R^2 = 0,9997$ e coeficiente de correlação de Pearson $r = 0,9999$ ($p = 2,22 \times 10^{-24}$). A faixa de medição do sensor foi observada entre -90.95cmH2O (≈ -8.9kPa) à +23.30cmH2O(≈+2.3kPa), sendo um intervalo de operação aprovado na literatura para a avaliação da respiração nasal até das vias mais obstruídas e de pacientes com maior esforço de acordo com o Comitê Internacional de Padronização da Rinomanometria que estipula um intervalo útil de ± 0.3kPa [7].

Complementarmente à faixa de medição observada, a resolução prática do sistema de aquisição foi caracterizada por meio de uma captura com a cânula de pressão vedada, isolando o ruído de fundo do conversor sob entrada fisicamente constante. O desvio-padrão das leituras nessa condição foi de $\sigma \approx$ 0,0009–0,0010 cmH2O, correspondendo a uma resolução efetiva de aproximadamente 16,9 bits sobre a faixa total do conversor (nominal de 24 bits, ganho 128). Adotando o critério de resolução livre de ruído (6,6 σ) — que representa a menor variação de pressão distinguível do ruído com margem estatística adequada [8] — obteve-se uma resolução de ≈ 14,1–14,3 bits, equivalente a cerca de 0,006 cmH2O (≈ 0,6 Pa). Esse resultado é consistente com o comportamento documentado para a família de conversores sigma-delta de 24 bits HX711/HX710B, que tipicamente perde 8 a 10 bits nominais para o ruído térmico do amplificador mesmo em condições ideais de medição.

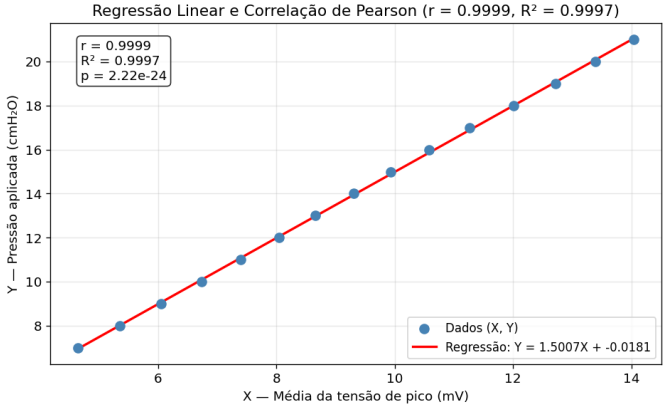


Fig. 2. Função estimada por regressão linear da Pressão x Tensão

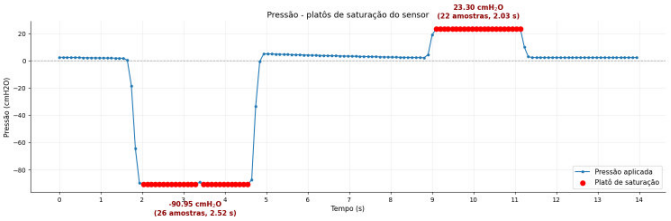


Fig. 3. Platôs de saturação do intervalo de medição do sensor

B. Esforço respiratório natural

A curva do ritmo respiratório obtido no programa de análise após a calibração do sensor foi eficaz para o cálculo da pressão instantânea, análise da forma da curva, cálculo da frequência respiratória e da relação I:E por via nasal, o que pode ser observado na figura4. Tal resultado está coerente com estudos médicos e fisiológicos de procesamento de dados da pressão nasal [9] e permite uma avaliação minuciosa do esforço respiratório do usuário.

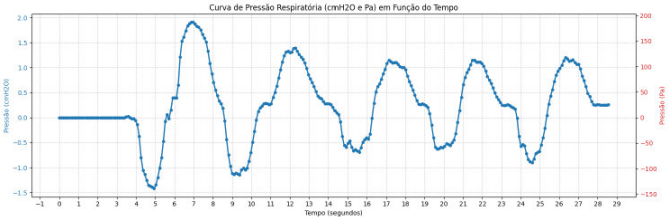


Fig. 4. Curva do esforço respiratório via nasal

IV. Conclusão

Este trabalho demonstrou a viabilidade técnica de um sistema portátil, não invasivo e de baixo custo para o monitoramento contínuo do fluxo e esforço respiratório, utilizando um sensor de pressão piezoresistivo integrado a um microcontrolador com comunicação sem fio. Os ensaios de calibração comprovaram a elevada precisão e repetibilidade do dispositivo na captação dos sinais de pressão e no rastreamento da frequência respiratória. A arquitetura em malha fechada viabilizou a aquisição e a análise remota das curvas de respiração nasal em tempo real, consolidando a solução como uma ferramenta promissora para estratégias de biofeedback e integração em dispositivos vestíveis. Essa abordagem abre caminho para o acompanhamento contínuo e preventivo de indivíduos com diferentes condições respiratórias diretamente em suas atividades cotidianas e na ausência de supervisão clínica presencial.

References

[1] J. E. Brinkman, F. Toro, and S. Sharma, “Physiology, Respiratory Drive,” Treasure Island, FL, USA: StatPearls Publishing, 2026. [Online]. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK482414/

[2] H. Liu, J. Allen, D. Zheng, and F. Chen, “Recent development of respiratory rate measurement technologies,” Physiological Measurement, vol. 40, no. 7, Art. N. 40, Ago. 2019, DOI: 10.1088/1361-6579/ab299e.

[3] N. L. Kazanskiy, S. N. Khonina, and M. A. Butt, “A review on flexible wearables - Recent developments in non-invasive continuous health monitoring,” Sensors and Actuators A: Physical, vol. 366, Art. N. 114993, Fev. 2024, DOI: 10.1016/j.sna.2023.114993.

[4] A. Shamsabadi, Z. Pashaei, A. Karimi, P. Mirzapour, K. Qaderi, M. Marhamati, A. Barzegary, A. Fakhfouri, E. Mehraeen, S. A. SeyedAlinaghi, and O. Dadras, “Internet of things in the management of chronic diseases during the COVID-19 pandemic: A systematic review,” Health Science Reports, vol. 5, no. 2, Art. N. e557, Mar. 2022, DOI: 10.1002/hsr2.557.

[5] K. Malik and A. Dua, “Advancing Patient Care With Biofeedback,” Treasure Island (FL): StatPearls Publishing, Jan. 2025 [Online]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/31971716/

[6] R. De Fazio, M. Stabile, M. De Vittorio, R. Velázquez, and P. Visconti, “An Overview of Wearable Piezoresistive and Inertial Sensors for Respiration Rate Monitoring,” Electronics, vol. 10, no. 17, Art. N. 2178, Set. 2021, DOI: 10.3390/electronics10172178

[7] P. A. Clement and F. Gordts, ”Consensus report on acoustic rhinometry and rhinomanometry,” Rhinology, vol. 43, no. 3, pp. 169–179, Sep. 2005.

[8] W. Kester, ”ADC input noise: The good, the bad, and the ugly. Is no noise good noise?,” Analog Dialogue, vol. 40, no. 1, pp. 1–5, Feb. 2006. [Online]. Available: https://www.analog.com/media/en/analog-dialogue/volume-40/number-1/articles/adc-input-noise.pdf

[9] URNER, Lorenz M.; KOHLER, Malcolm; BLOCH, Konrad E., ”Automatic processing of nasal pressure recordings to derive continuous side-selective nasal airflow and conductance.,” Frontiers in physiology, vol. 9, pp. 1814, 2005.

Modelling of a strain gauge-based transducer for Neonatal Respiratory Flow sensor

Maria Gabriela Medeiros von Adamek
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
https://orcid.org/0009-0001-7644-3815

André Paganotti
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
https://orcid.org/0000-0002-6366-2306

Gustavo José Giardini Lahr
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
https://orcid.org/0000-0002-0403-991X

Celso Moura Rebello
Pediatric Department
Hospital Israelita Albert Einstein
São Paulo, Brazil
https://orcid.org/0000-0002-4637-6295

***Abstract*—Manual positive pressure ventilation (PPV) is the simplest and most common intervention during neonatal resuscitation, yet excessive pressures and volumes can injure the immature lung. Continuous flow monitoring mitigates this risk, but commercial solutions based on hot-wire anemometry remain expensive and proprietary. This work presents a purely computational model of an open-source, low-cost respiratory flow transducer based on an impact disc mounted on a thin cantilever, whose bending is read by a strain gauge. The fluid–structure response was characterized with a sequential workflow: computational fluid dynamics for force extraction, finite element analysis for strain, and an analytical model for the signal-conditioning chain. Under standard neonatal ventilation conditions, over the target range of 0.2 to 9.0 liters per minute, the aerodynamic force on the selected disc geometry reaches about 166 μN, producing a peak bending strain of only about 0.57 $\mu\varepsilon$ at the cantilever root. A demonstration that the base pressure cancels in the net disc force allowed the simulations to be run at zero gauge outlet pressure, removing numerical noise. Linear-elastic behavior let the full operating range be derived from a single reference simulation. A metallic foil gauge was selected for the transducer on grounds of cost, availability and robustness; with this gauge and the present cantilever, the predicted bridge output stays near the practical detection floor of foil technology across the entire range, and falls below it at low flow. This sensitivity shortfall, most severe precisely where the smallest patients are ventilated, is identified as the principal design constraint exposed by the model. No physical prototype has yet been built or calibrated, and the results should be read as a computational feasibility study: they define the geometric and electronic constraints – and motivate a thinner cantilever neck, a full-bridge layout, or a higher-gauge-factor transducer – for a future prototype and its experimental validation.**

***Index Terms*—neonatal ventilation, respiratory flow sensor, strain gauge, computational fluid dynamics, finite element analysis, fluid-structure interaction**

I. Introduction

Positive pressure ventilation (PPV) is the single most effective intervention for a newborn that fails to breathe effectively after birth, yet it can also injure the immature lung: excess volume and pressure cause volutrauma, barotrauma, and biotrauma, driving bronchopulmonary dysplasia [1], [2], while too little volume causes atelectrauma, hypoxemia, and hypercarbia; excessive ventilation additionally produces hypocarbia, linked to white-matter brain injury in preterm infants [3]. This therapeutic window is narrowest precisely in the smallest patients. Yet PPV in the delivery room is most often delivered with self-inflating bags or T-piece resuscitators that give no feedback on delivered volume and flow, so operators routinely misjudge tidal volume, with wide variability between providers [4]. Real-time flow, volume, and pressure monitoring is the practical means of staying within the safe window [5], [6], but is rarely available where resuscitation actually takes place.

Existing flow-measurement technologies carry limitations that are aggravated at neonatal scale [7]. Hot-wire anemometry offers high sensitivity at low flow but is nonlinear, gas-composition-dependent, mechanically fragile, prone to fouling, and supplied only as proprietary hardware. Differential-pressure pneumotachographs are linear only in laminar flow, add dead space and work of breathing, and require correction for oxygen fraction and temperature – hardest precisely at the low flows seen in neonates [8]. Ultrasonic transit-time sensors achieve low dead space but remain sensitive to gas composition and temperature and require costly timing electronics [7]. Across all options, commercial neonatal sensors are proprietary and often single-use, imposing a recurring per-patient cost with no path for open adaptation or local manufacture. What is needed is an open, low-complexity, low-

cost alternative with small dead space and adequate sensitivity at the very low flows characteristic of neonatal ventilation.

This paper models such an alternative: a flow transducer in which an impact disc placed in the airway is deflected by the gas flow, and the resulting bending of a thin cantilever neck is converted into an electrical signal by a strain gauge. The objective is to determine, before any physical prototype is built, whether the aerodynamic forces in the neonatal flow range produce a strain that is measurable with a realistic, low-cost conditioning chain. The analysis follows a one-way fluid-structure workflow: computational fluid dynamics (CFD) extracts the force on the disc, finite element analysis (FEA) converts that force into strain, and an analytical model maps the strain into a bridge output voltage.

II. Methods

The transducer is analyzed with a sequential, one-way fluid–structure workflow: computational fluid dynamics (CFD) yields the aerodynamic force on the impact disc, finite element analysis (FEA) converts that force into the bending strain at the cantilever root, and an analytical model maps the strain into a bridge output voltage. Accordingly, this section first defines the sensing geometry (Section II-A) and the target flow range (Section II-B), and then presents each stage of the chain in turn: force extraction by CFD (Section II-C), strain computation by FEA (Section II-D), and the strain-to-voltage signal model (Section II-E).

A. Drag-disc principle and geometry

The sensing element is a drag-disc, or target, meter: a bluff body placed in the flow whose drag force is measured to infer the flow rate [9]. It consists of a circular disc of area A_s centred in a tube of cross-sectional area A_t, supported by a thin cantilever neck anchored to the tube wall, as shown in Fig. 1. The fraction of the cross-section occupied by the disc is described by the area-restriction figure of merit

$$\phi = \frac{A_s}{A_t}, \tag{1}$$

and, for circular geometries, by the diameter ratio

$$\beta = \sqrt{\phi} = \frac{D_s}{D_t}. \tag{2}$$

Note that ϕ is defined here as the blocked fraction of the cross-section, so that β grows with the size of the obstruction. A larger disc intercepts a greater fraction of the momentum flux and therefore develops a larger force, increasing the sensitivity, but it also raises the flow resistance and intensifies both upstream and downstream disturbances; β must consequently be chosen as a compromise between these competing effects. This trade-off runs opposite to that of an orifice plate, in which the restriction is the aperture rather than the body, and sensitivity therefore grows as β decreases. No standard prescribes β for drag-disc elements at neonatal flow rates: industrial target meters operate at much lower blockage, and the well-established β ranges are those tabulated for orifice-plate meters [10], whose topology is the inverse of the present one. The orifice-plate range was therefore adopted here as a heuristic starting point rather than as a validated criterion, restricting the design space to $0.5 < \beta < 0.7$. The comparatively high blockage that follows is a deliberate consequence of the application: at the lower bound of the neonatal range the aerodynamic force is of the order of a micronewton, so a smaller disc would not develop a measurable signal, and sensitivity must be bought at the cost of pressure drop and dead space.

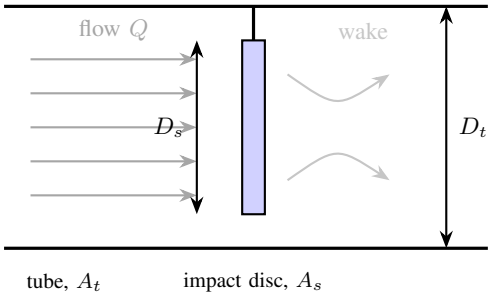


Fig. 1. Schematic longitudinal section of the drag-disc sensing element. A circular disc of diameter D_s and area A_s is centred in a tube of diameter D_t and cross-sectional area A_t, and is supported by a thin cantilever neck anchored to the tube wall. The airflow impinges on the upstream face of the disc, and the resulting axial force bends the neck. The ratio $\beta = D_s/D_t$ determines both the sensitivity and the flow disturbance introduced by the element.

Three candidate disc radii were evaluated: a baseline of $3.5\ \mathrm{mm}$ (denoted r_1), whose β lies in the middle of this range; a radius $20\,\%$ larger, $4.2\ \mathrm{mm}$ (r_2); and a radius $20\,\%$ smaller, $2.8\ \mathrm{mm}$ (r_2). Since the disc area scales with the square of the radius, r_2 and r_3 correspond to about 1.44 and 0.64 times the baseline face area, respectively. For r_2, the face area is $A_s \approx 0.795\ \mathrm{cm}^2$ with $\beta \approx 0.71$, at the upper edge of the design space. The three geometries were simulated independently and compared in Section III-A.

B. Target flow range

The clinical range follows from the smallest and largest patients to be ventilated. Taking the volumetric flow as the tidal volume delivered over an inspiratory time of $\Delta t = 0.5\ \mathrm{s}$, the lower bound corresponds to a 400 g neonate receiving $4\ \mathrm{mL/kg}$, giving $Q_{\min} \approx 0.2\ \mathrm{L/min}$, and the upper bound to a 5000 g neonate receiving $15\ \mathrm{mL/kg}$, giving $Q_{\max} \approx 9.0\ \mathrm{L/min}$. The sensor must therefore operate over 0.2 to $9.0\ \mathrm{L/min}$, a span of roughly $45{:}1$.

C. Simulation for force extraction

The airflow around the disc was simulated with Autodesk CFD 2027 using the finite volume method [11]. The fluid was modelled as air at body-normothermic conditions ($101.325\ \mathrm{kPa}$, $37\ ^\circ\mathrm{C}$), with density $\rho = 1.138\ \mathrm{kg/m^3}$ and dynamic viscosity $\mu = 1.817 \times 10^{-5}\ \mathrm{Pa \cdot s}$, consistent with humidified gas delivered during neonatal resuscitation. The flow was solved as laminar over the entire target range: at the tube diameter and flow rates considered, the Reynolds number

remains well below the transitional threshold ($Re \approx 17$–85 across 0.2–9.0 $\mathrm{L/min}$), which justifies the laminar assumption and avoids the need for a turbulence closure model whose empirical constants are not validated at this scale. The computational domain was discretized using Autodesk CFD's automatic (autosize) meshing algorithm; a single mesh density was used for all reported results, and no mesh independence study was performed (Section IV-D).

Boundary conditions were prescribed as a volumetric flow rate at the inlet (two surfaces) and a fixed static pressure of 0 Pa (gauge) at the outlet (eight surfaces), consistent with the base-pressure cancellation argument below. No inlet velocity profile was imposed manually; the solver distributes the prescribed volumetric flow rate across the inlet surfaces internally.

Convergence was governed by Autodesk CFD's automatic convergence criterion rather than a fixed iteration count. The solver was configured with an iteration budget of 300, but under the automatic criterion execution consistently terminated well before this limit. At 0.4 $\mathrm{L/min}$, for instance, the solver stopped at iteration 110, with residuals of $V_x \approx -5.40\times10^{-3}$, $V_y \approx -1.81$, $V_z \approx -1.02 \times 10^{-1}$, and pressure $\approx 3.56 \times 10^{-2}$; temperature, turbulent kinetic energy, and turbulent energy dissipation residuals were negligible under the laminar, isothermal assumption. While the automatic criterion reflects Autodesk CFD's internal assessment that the solution had stabilized, the y-velocity residual remained roughly an order of magnitude higher than the other velocity and pressure channels at the point of termination; it is not established whether this reflects a converged-but-naturally-noisier component or a channel that would have continued to decay had the automatic stop not triggered.

Standard neonatal PPV parameters were adopted as operating conditions, in accordance with World Health Organization and American Heart Association guidance [12] [13]: positive end-expiratory pressure (PEEP) of 5 $\mathrm{cmH_2O}$, peak inspiratory pressure (PIP) of 20 $\mathrm{cmH_2O}$, respiratory rate of 30 to 60 breaths per minute and tidal volume of 4 to 6 $\mathrm{mL/kg}$. Nineteen steady operating points spanning the target range were solved for each of the three candidate radii (r_1, r_2 and r_3), each geometry being set up and solved independently.

a) Aerodynamic model: The axial force on a bluff body is, to first order,

$$F = \tfrac{1}{2}\, C_d\, \rho\, A_s\, v^2, \tag{3}$$

where ρ is the gas density, v a reference velocity and C_d a drag coefficient. Because the mean velocity scales linearly with the volumetric flow, $v \propto Q$, the ideal scaling is $F \propto Q^2$. The CFD results follow a slightly softer law in the transitional regime,

$$F \propto Q^{1.7}, \tag{4}$$

approaching the quadratic limit only at the highest flows.

b) Base-pressure cancellation: The net axial force on the disc is the difference between the load on the front (impact) face and the load on the rear face,

$$F_{\mathrm{net}} = \int_{A_s} p_f \,\mathrm{d}A - \int_{A_s} p_r \,\mathrm{d}A. \tag{5}$$

For an incompressible flow the absolute pressure level is arbitrary: only the pressure gradient produces force. This can be shown formally by adding an arbitrary uniform base pressure p_0 (representing, physically, the PEEP offset) to both faces:

$$\int_{A_s} (p_f + p_0)\,\mathrm{d}A - \int_{A_s} (p_r + p_0)\,\mathrm{d}A = F_{\mathrm{net}}, \tag{6}$$

since the projected areas of the two faces are equal, the $p_0\, A_s$ terms cancel exactly regardless of the value of p_0. The net force is therefore, by construction, independent of the base pressure level – including the physical PEEP baseline the transducer would actually operate against clinically.

This result has a direct numerical consequence: since F_{net} is provably independent of p_0, the outlet boundary condition can be prescribed at the numerically convenient value of 0 Pa (gauge) instead of at the physical PEEP level, without altering the physical quantity of interest. This choice reads the dynamic pressure directly on the impact face and avoids computing F_{net} as the difference of two much larger, nearly equal absolute-pressure integrals, which would otherwise leave the physical signal degraded by round-off and iterative solver noise. Equation (6) therefore justifies the boundary condition choice on physical grounds first; the reduction in numerical noise is a beneficial byproduct of an already-correct simplification, not the basis for it.

c) Empirical verification: The cancellation was verified for one point: the r_3 geometry at 3.0 $\mathrm{L/min}$, re-run with the outlet at the physical PEEP value (5 $\mathrm{cmH_2O} \approx 490$ Pa) instead of 0 Pa. The front- and rear-face forces read -2140.9 and $+2140.13$ dyne ($\approx \mp 21{,}400$ μN), each dominated by the ≈ 490 Pa base pressure – three orders of magnitude above the expected signal. Their sum, -0.77 dyne ≈ 7.7 μN, agrees with the 7.19 μN obtained under the standard 0 Pa-gauge convention (Table I; $7\,\%$ difference, within run-to-run variation, Section IV-D), confirming Eq. (6): the base pressure cancels in the net force even though it dominates each face individually.

d) Wake correction: The rear-face suction is not resolved cleanly by the wall integration and contributes about $10\,\%$ of the total drag, estimated from the rear-face pressure valley. The net force is therefore taken as

$$F_{\mathrm{net}} = 1.1\, F_{\mathrm{front}}. \tag{7}$$

Forces are reported in micronewtons; the native CFD output in dyne is converted with 1 dyne $= 10$ μN.

D. Simulation for strain from the disc force

The structural response was solved in Autodesk Fusion 360 (Static Stress), with the cantilever modelled in Tough2000 photopolymer resin ($E = 2210$ MPa, $\nu = 0.4$, neck thickness $t = 1$ mm, printed by masked stereolithography). A fixed constraint was applied to the outer plate face ($r = 20$ mm) and the net aerodynamic force to the disc face – equivalent to the resultant acting at the centre of pressure by axisymmetry, avoiding stress singularities. A second-order (parabolic) mesh with absolute element sizing was used.

a) Analytical bending model: Treating the neck as a cantilever loaded near its tip, the bending moment at the root is $M = F\,\ell$ (ℓ: lever arm from disc centre to root), giving a root surface strain of

$$\varepsilon = \frac{\sigma}{E} = \frac{M\,c}{E\,I} = \frac{6\,F\,\ell}{E\,w\,t^2}, \tag{8}$$

with w the neck width, $c = t/2$, and $I = w\,t^3/12$. Strain scales as $1/t^2$ with thickness but only linearly with ℓ, so thinning the neck is the dominant sensitivity lever.

b) Linearity and scaling: Deflections are nanometric, so the response is linear-elastic and a single reference run at maximum flow suffices; strain at any other point follows by scaling,

$$\varepsilon_i = \varepsilon_{\mathrm{ref}}\,\frac{F_i}{F_{\mathrm{ref}}}. \tag{9}$$

A new FEA run is needed only if the geometry changes.

E. Signal model: strain to voltage

A strain gauge converts strain into a fractional resistance change $\Delta R/R = \mathrm{GF}\cdot\varepsilon$ [14], with GF the gauge factor. A single active gauge in a Wheatstone bridge (quarter-bridge) gives, for excitation V_{ex},

$$V_{out} = \frac{V_{ex}}{4}\,\mathrm{GF}\,\varepsilon. \tag{10}$$

The cantilever presents tension on one face and equal compression on the other; placing one gauge on each face in adjacent bridge arms (reading $+\varepsilon$ and $-\varepsilon$) doubles the output to

$$V_{out} = \frac{V_{ex}\,\mathrm{GF}}{2}\,\varepsilon. \tag{11}$$

This half-bridge doubles the quarter-bridge output (10) and, since both gauges share the same thermal environment, cancels common-mode drift – important given the resin substrate's poor thermal conductivity.

III. Results

The results follow the same sequence as the analysis. The three candidate geometries are first compared by the aerodynamic force extracted from CFD, which fixes the disc radius (Section III-A); the selected geometry is then carried into the FEA to obtain the root bending strain (Section III-B), and finally into the signal model, which determines whether the resulting bridge output is measurable over the target range (Section III-C).

A. CFD forces and disc-radius selection

The three candidate radii were compared by their net disc force, summarised in Table I and Fig. 2. At the top of the range (9.0 L/min) the baseline r_1 produces about 88 μN, the larger r_2 about 166 μN, and the smaller r_3 about 46 μN. The net force increases monotonically with the disc frontal area, so r_2 yields roughly 1.9 times the baseline force and about 3.6 times that of r_3. Because a larger force produces a proportionally larger bending strain and therefore a stronger output, r_2 gives the best response of the three while keeping $\beta \approx 0.71$ at the upper edge of the acceptable restriction range; r_3 keeps resistance lowest but its signal is too weak, and r_1 sits in between. The disc radius r_2 was therefore selected, and all subsequent analysis uses this geometry. Figure 3 shows the velocity field and the static-pressure distribution on the disc for the r_2 geometry at 9 L/min. The jet impinges on the upstream face and forms a central stagnation region of maximum pressure that decays towards the rim; integrating this distribution over the face gives the net axial load reported below.

TABLE I
NET DISC FORCE (μN) FOR THE THREE CANDIDATE RADII (OUTLET AT 0 PA GAUGE, NET FORCE = 1.1× IMPACT FACE)

Q (L/min)	r_1 = 3.5 mm	$r_2 = 4.2\,mm$	r_3 = 2.8 mm
0.2	0.30	**0.56**	0.14
0.4	0.76	**1.34**	0.34
0.6	1.23	**2.31**	0.59
0.8	1.84	**3.46**	0.90
1.0	2.54	**4.77**	1.26
1.5	4.70	**8.70**	2.36
2.0	7.30	**13.50**	3.73
2.5	10.30	**19.10**	5.34
3.0	13.80	**25.50**	7.19
3.5	17.70	**32.70**	9.28
4.0	22.00	**40.60**	11.60
4.5	26.70	**49.20**	14.10
5.0	31.90	**58.60**	16.80
5.5	37.40	**69.00**	19.80
6.0	43.30	**80.30**	22.90
6.5	49.70	**92.40**	26.20
7.0	56.50	**105.40**	29.80
7.5	63.80	**119.20**	33.60
8.0	71.50	**133.90**	37.50
8.5	79.80	**149.30**	41.70
9.0	88.40	**165.70**	46.10

For the selected geometry r_2 the net force grows from about 0.56 μN at 0.2 L/min to about 166 μN at 9.0 L/min, following the $F \propto Q^{1.7}$ law of (4), a gain consistent with the area dependence in (3). The impact-face force is recovered as the net value divided by the 1.1 wake factor of (7).

B. FEA strain

At the reference point ($F_{\mathrm{ref}} = 1.657 \times 10^{-4}$ N, i.e. 9.0 L/min), the normal strain along the flow axis at the root reaches $+5.659\times10^{-7}$ (about $+0.566$ $\mu\varepsilon$) on the tension face and -5.78×10^{-7} (about -0.578 $\mu\varepsilon$) on the compression face. The two values are nearly equal and opposite across the 1 mm thickness, which is the signature of pure bending and confirms that the chosen topology is correct.

The analytical estimate from (8) is about 0.72 $\mu\varepsilon$, and the FEA value is about 0.57 $\mu\varepsilon$; the two agree within about 21 % (ratio 0.79). The difference is attributed to the simplifying assumptions of a slender prismatic beam, whereas the real geometry includes the disc and the neck-to-plate transition, which add stiffness. The agreement is close enough to validate the model and to rule out unit errors. By (9), the strain at every other flow point is obtained by scaling the reference value with the force ratio; the resulting strain curve is listed in Table II and plotted in Fig. 4.

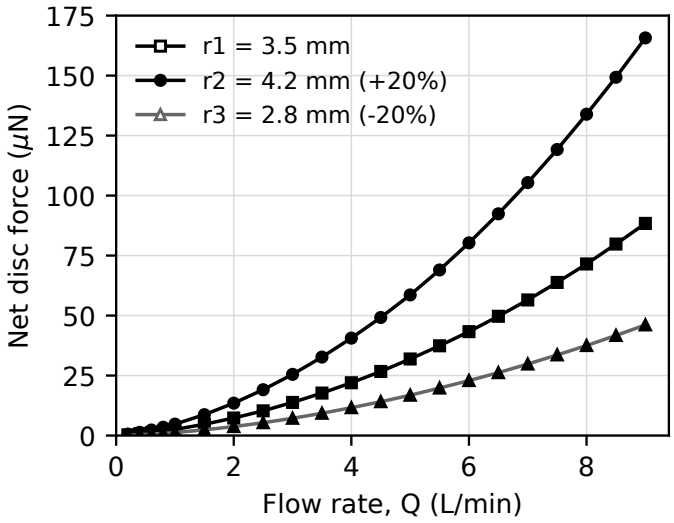


Fig. 2. Net disc force as a function of flow rate for the three evaluated disc radii. The generated force increases nonlinearly with flow rate due to the greater momentum transfer between the airflow and the sensing element. Increasing the disc radius increases the exposed area and enhances the flow-induced force, resulting in higher sensor sensitivity. The $r_2 = 4.2$ mm configuration ($+20\%$ relative to the baseline) produced the highest force response over the entire investigated flow range and was therefore selected as the final design due to its superior sensitivity. The baseline ($r_1 = 3.5$ mm) and reduced radius ($r_3 = 2.8$ mm) configurations exhibited lower force outputs due to their smaller interaction area with the flow.

TABLE II
FEA ROOT STRAIN AND FOIL HALF-BRIDGE OUTPUT FOR r_2 ($t = 1$ mm, GF $= 2$, $V_{ex} = 5$ V)

Q **(L/min)**	F_{net} (μN)	ε ($\mu\varepsilon$)	V_{out} **(foil)** (μV)
0.2	0.56	0.0019	0.010
0.4	1.34	0.0046	0.023
0.6	2.31	0.0080	0.040
0.8	3.46	0.0119	0.060
1.0	4.77	0.0165	0.082
1.5	8.70	0.0300	0.150
2.0	13.50	0.0466	0.233
2.5	19.10	0.0659	0.330
3.0	25.50	0.0880	0.440
3.5	32.70	0.1129	0.564
4.0	40.60	0.1402	0.701
4.5	49.20	0.1698	0.849
5.0	58.60	0.2023	1.011
5.5	69.00	0.2382	1.191
6.0	80.30	0.2772	1.386
6.5	92.40	0.3190	1.595
7.0	105.40	0.3638	1.819
7.5	119.20	0.4115	2.057
8.0	133.90	0.4622	2.311
8.5	149.30	0.5154	2.577
9.0	165.70	0.5720	2.860

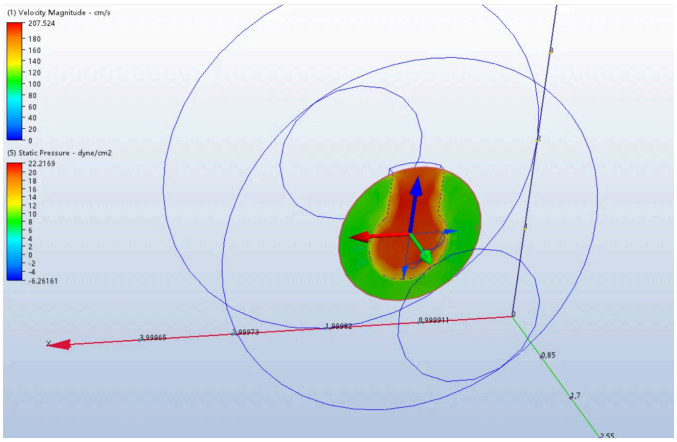


Fig. 3. CFD solution for the r_2 geometry at the upper bound of the neonatal range, 9 L/min. The blue streamlines trace the airflow around the impact disc and are coloured by velocity magnitude (top scale, cm/s), while the disc surface is coloured by static pressure (bottom scale, Pa; converted from the native Autodesk CFD output in dyne/cm^2, where 1 dyne/cm^2 $= 0.1$ Pa – the color-scale legend embedded in the image itself is still labelled in dyne/cm^2 and will be re-exported in Pa in a future revision). As the incoming jet strikes the upstream face, it decelerates and forms a central stagnation region where the static pressure reaches its maximum (≈ 2.2 Pa, red), decreasing towards the rim (≈ 0.6 Pa, green). Integrated over the face, this non-uniform pressure distribution yields the net axial load that deflects the cantilever and is transduced by the strain gauge.

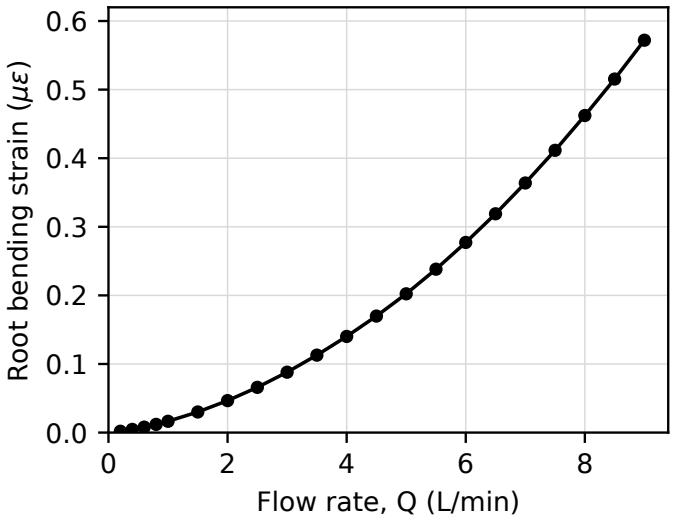


Fig. 4. FEA root bending strain versus flow rate for the r_2 geometry ($t = 1$ mm), obtained by linear scaling of the reference simulation. The predicted increase in root strain with flow rate highlights the greater mechanical sensitivity of r_2, which was selected due to its larger deformation response under the same flow conditions.

C. Signal viability

The transducer uses a metallic foil gauge (GF $= 2$), selected for its low cost, ready availability and robustness. With the half-bridge (11) at $V_{ex} = 5$ V the output reduces to $V_{out} = 5\,\varepsilon$ (μV per $\mu\varepsilon$); this linear strain-to-voltage transfer is shown in Fig. 5. Because the conversion is a constant factor, the foil output versus flow has the same $F \propto Q^{1.7}$ shape as the strain curve of Fig. 4, simply rescaled, and its values are listed in

the last column of Table II. The voltage-to-flow calibration is therefore non-linear.

The peak strain over the whole range is only about $0.57\ \mu\varepsilon$, comparable to the roughly $1\ \mu\varepsilon$ practical detection floor of metallic foil gauges. Even with the half-bridge, the foil output peaks at $2.86\ \mu\mathrm{V}$ and falls to about $10\ \mathrm{nV}$ at the lowest flow; below roughly $2.5\ \mathrm{L/min}$ it approaches the thermal noise floor of the bridge itself, a physical limit that no circuit architecture can overcome. The high-flow portion of the range is therefore the readable region for the present foil-based design, while the low-flow end is the limiting case, addressed in Section IV-D.

IV. Discussion

A. Design levers

With the gauge fixed as foil ($\mathrm{GF} = 2$), the available means to raise the signal are geometric and circuit-level. Geometrically, the neck thickness dominates: since $\varepsilon \propto 1/t^2$, reducing t from 1.0 to $0.5\ \mathrm{mm}$ would quadruple the strain (peak near $2.3\ \mu\varepsilon$), and $t = 0.3\ \mathrm{mm}$ would raise it to about $6.4\ \mu\varepsilon$, well above the foil detection floor. At circuit level, moving from the half-bridge to a full-bridge doubles the output again, and AC carrier excitation with synchronous demodulation, bandwidth limiting to the physiological band, oversampling and ratiometric measurement provide further, smaller gains. A complementary route is to replace the foil gauge with a semiconductor gauge, whose gauge factor is one to two orders of magnitude higher and would lift the output across the entire range, including the low-flow end that geometry alone cannot fully rescue; this was not adopted here on grounds of cost and availability. None of these options is applied to the present design, whose geometry is frozen and whose gauge is fixed as foil.

B. Conditioning chain

With the geometry frozen, the conditioning chain is fully analogue and discrete, and the microcontroller reads only the conditioned voltage at the end. The chain is shown in Fig. 6, and each stage is dictated by the nature of the signal: a sub-microvolt, near-DC voltage produced on a thermally poor resin substrate. The peak foil output is only about $2.86\ \mu\mathrm{V}$ (Fig. 5), so mapping it onto a useful ADC window of a few hundred millivolts requires a total gain near 10^5. The whole design follows from getting that gain without amplifying offset and drift along with the signal [15], [16].

Half-bridge. Two active gauges are placed on opposite faces of the neck, sensing equal and opposite strain. This doubles the sensitivity relative to a quarter-bridge (11) and, because both gauges share the same temperature, it cancels the common-mode thermal response to first order. A full bridge would double the output again but requires four matched gauges on a neck only a few millimetres wide, which the geometry does not allow; the half-bridge is the best compromise between sensitivity and the available area.

Instrumentation amplifier ($G_1 \approx 100$). The first stage must present a very high input impedance so as not to load the bridge, and a high common-mode rejection ratio so that the bridge common-mode voltage and supply noise are removed rather than amplified. A zero-drift (chopper-stabilized) instrumentation amplifier is used because the signal is essentially DC: at these frequencies the dominant error is the amplifier input offset and its drift, not broadband noise, and chopper devices reduce that offset to the nanovolt range [15], [16]. The first-stage gain is held to about 100 so that any residual bridge offset is not driven into saturation before it can be removed.

Offset cancellation. Even a $0.1\,\%$ mismatch between the two gauges produces a bridge offset of the order of millivolts, which at the full gain of 10^5 would saturate the chain long before the microvolt signal appeared. The offset is therefore nulled before the high-gain stage, either with a multi-turn trim potentiometer at the amplifier reference pin or with a DC servo that integrates the output and feeds the slow average back to the reference, high-pass filtering the chain so that only the dynamic (breathing) component is amplified. This is the single most critical requirement of the front-end.

Second-order low-pass filter ($f_c \approx 40$ Hz). The neonatal flow signal and its harmonics lie below a few tens of hertz, so the bandwidth is limited to about 40 Hz. Because broadband noise grows with the square root of bandwidth, restricting the band improves the signal-to-noise ratio, and the filter also serves as the anti-aliasing stage ahead of the ADC. A Sallen-Key topology is chosen because it realises a second-order response with a single op-amp and few passive components and has low sensitivity to component tolerances [17]; an equal-component unity-gain Butterworth alignment gives a maximally flat passband. A corner at 40 Hz passes the signal while attenuating mains interference at $50/60$ Hz and above.

Second gain and level shift (G_2). With the offset removed and the band limited, the remaining gain (about 10^3) is applied in a second stage, again with a zero-drift op-amp, and split from the first stage to keep each amplifier within its gain-bandwidth limit and stable. A level-shift network then moves the bipolar signal, which swings about zero on the $\pm$ rails, into the unipolar 0 to 3.3 V window of the ESP32 input.

Supply and grounding. The front-end runs on symmetric ($\pm$) rails so the bridge signal can swing about zero, while the ESP32 uses its own 3.3 V supply; the two share a single star ground to avoid ground loops, which at the microvolt level would otherwise inject more noise than the signal itself. Finally, because the force follows $F \propto Q^{1.7}$, the voltage-to-flow mapping is non-linear and is implemented as a calibration in firmware after digitisation.

C. Dead space and pressure drop

a) Dead space (disc and neck only): The physical volume displaced by the sensing element within the airway can be estimated from its two components inside the flow path: the impact disc and the cantilever neck. For the selected geometry ($r_2 = 4.2$ mm, disc thickness 1 mm), the disc volume is

$$V_{\mathrm{disc}} = \pi r_2^2 t_{\mathrm{disc}} \approx \pi (4.2\ \mathrm{mm})^2 (1\ \mathrm{mm}) \approx 55.4\ \mathrm{mm}^3. \quad (12)$$

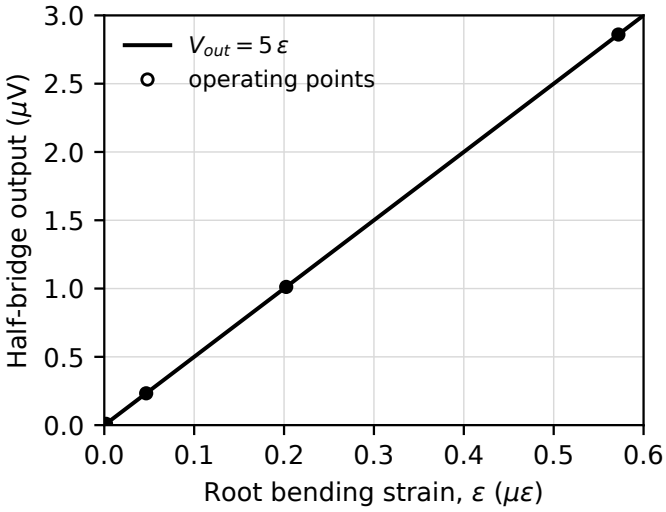


Fig. 5. Strain-to-voltage transfer characteristic of the foil half-bridge used for flow-induced deformation measurement. The bridge output voltage (μV) increases linearly with strain, providing the conversion factor between mechanical deformation and electrical signal.

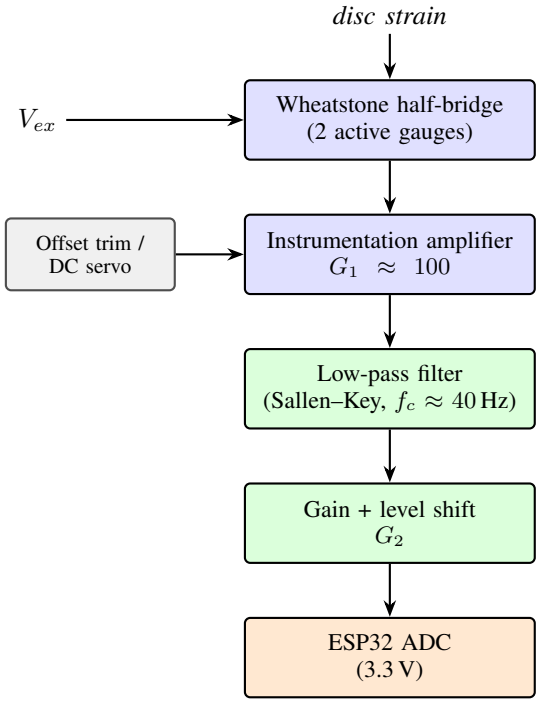


Fig. 6. Conditioning chain.

The cantilever neck, of width $w = 5.5$ mm, thickness $t = 1$ mm, and free length $\ell - r_2 \approx 7.56 - 4.2 = 3.36$ mm (from the disc rim to the root), adds

$$V_{\text{neck}} = w\,t\,(\ell - r_2) \approx (5.5)(1)(3.36)\ \text{mm}^3 \approx 18.5\ \text{mm}^3. \tag{13}$$

Together, $V_{\text{disc}} + V_{\text{neck}} \approx 73.9\ \text{mm}^3 \approx 0.074$ mL, well within the ≤ 1 mL dead-space target set for this sensor. This estimate, however, covers only the disc and neck; it excludes the volume of the outer mounting plate ($r = 20$ mm) and any housing structure within the airway, which were not analyzed for flow displacement in the present CFD study (only the force on the impact disc was extracted). The dead space of the complete assembly therefore remains to be confirmed once the full housing geometry is included.

b) Pressure drop: The pressure drop across the sensor was extracted directly from the CFD solution at one operating point: the r_3 geometry (the smallest of the three candidate radii, $\beta \approx 0.57$) at 3.0 L/min, using the same PEEP-outlet run reported in Section II-C for the base-pressure verification. The area-weighted average pressure at the inlet was 490.705 Pa, against an outlet condition fixed at 50 $\text{mmH}_2\text{O} \approx 490.33$ Pa (using 1 $\text{mmH}_2\text{O} \approx 9.80665$ Pa). The resulting pressure drop is

$$\Delta P = P_{\text{inlet}} - P_{\text{outlet}} \approx 490.705 - 490.33 \approx 0.37\ \text{Pa} \tag{14}$$

at this flow and geometry. This single point is far too limited to characterize the pressure-drop–flow relationship over the full 0.2–9.0 L/min range or to compare candidate radii, since ΔP is expected to scale with flow and with disc blockage β in the same manner as the axial force (Section II-C). Extracting the inlet average pressure at additional operating points, and ideally for the selected r_2 geometry used throughout the rest of this work, is needed to report a complete pressure-drop curve and to assess whether the added flow resistance of the disc is clinically acceptable for the intended application.

D. Limitations

1) **Low-flow sensitivity.** With the foil gauge (GF = 2) and the present 1 mm cantilever, peak strain over the whole range is only 0.57 $\mu\varepsilon$, near the practical detection floor of foil technology; below roughly 2.5 L/min the output approaches the bridge's own thermal noise floor, a physical limit no amplification recovers. Because the smallest patients are ventilated precisely at these low flows, this is the principal barrier to clinical usefulness.
2) **No physical validation.** The study is entirely computational; no prototype has been built or calibrated against a reference flowmeter, so the predicted forces, strains, and noise floor remain unconfirmed experimentally. Masked-stereolithography print tolerances – neck thickness, disc flatness, gauge placement and bonding – are not captured by the model.
3) **One-way fluid-structure coupling.** The aerodynamic force is transferred manually from CFD to FEA with no feedback of the deformed geometry on the flow;

justified by the nanometric, linear-elastic deflections, but the $10\,\%$ wake correction (Eq. (7)) is a fixed estimate rather than resolved independently at each flow, and a single reference FEA point anchors the linear force–strain scaling of Eq. (9) for the entire range.

4) **Convergence.** Autodesk CFD's automatic criterion halted runs before the configured 300-iteration budget (e.g. iteration 110 at $0.4\ \mathrm{L/min}$) while the y-velocity residual (-1.81) remained an order of magnitude above the other channels; whether this component was truly converged is unconfirmed.
5) **Mesh.** Only Autodesk CFD's automatic (autosize) meshing was used, at a single density; no mesh-independence study was performed, so under-resolution of the disc's stagnation region cannot be ruled out as a contributor to the forces in Table I.
6) **Flow directionality.** All nineteen operating points are unidirectional (inspiratory). The half-bridge can in principle sense a sign reversal for expiratory flow, but the aerodynamic force model and wake correction were neither derived nor verified for reversed flow, where the front and rear disc faces exchange roles.
7) **Steady state only.** The nineteen operating points are all steady, but neonatal ventilation is cyclic; the transient response of the disc–cantilever system and the $40\ \mathrm{Hz}$ conditioning bandwidth are not validated dynamically. The added flow resistance and dead space of the disc are only partially quantified (Section IV-C: a single pressure-drop point and a disc-and-neck-only dead-space estimate).

V. Conclusions

A sequential CFD-FEA-signal model of an impact-disc transducer on a thin cantilever, read by a strain gauge, identifies the feasibility constraints of this design as a candidate for an open-source, low-cost neonatal flow sensor. Over the clinical range of 0.2 to 9.0 L/min, the aerodynamic force reaches about 166 μN and produces a peak root strain of only about 0.57 $\mu\varepsilon$ in pure bending. Demonstrating that the base pressure cancels in the net force enabled clean, noise-free simulations, and the linear-elastic response allowed the whole range to be derived from a single reference run. A metallic foil gauge was adopted for cost, availability and robustness; with the present geometry, its predicted sensitivity at low flow is the principal limitation identified by the model, falling below the practical detection floor of foil technology precisely where the smallest patients are ventilated. Addressing this shortfall will require thinning the cantilever neck, moving to a full-bridge topology, or adopting a higher-gauge-factor transducer. As the analysis is entirely computational, the natural next step is a physical prototype, printed in the selected geometry and calibrated against a reference flowmeter over the target range, which would also expose the transient response and the thermal drift of the resin substrate to direct measurement. These results establish a computational foundation and the geometric and electronic constraints for that prototype and its experimental validation, rather than evidence that the sensor concept itself is already viable.

References

[1] R. Kalikkot Thekkeveedu, A. El-Saie, V. Prakash, L. Katakam, and B. Shivanna, "Ventilation-induced lung injury (VILI) in neonates: evidence-based concepts and lung-protective strategies," *J. Clin. Med.*, vol. 11, no. 3, art. 557, 2022.

[2] A. H. Jobe and E. Bancalari, "Bronchopulmonary dysplasia," *Am. J. Respir. Crit. Care Med.*, vol. 163, no. 7, pp. 1723–1729, 2001.

[3] G. R. Polglase, S. L. Miller, S. K. Barton, A. A. Baburamani, F. Y. Wong, J. D. S. Aridas, A. W. Gill, T. J. M. Moss, M. Tolcos, M. Kluckow, and S. B. Hooper, "Initiation of resuscitation with high tidal volumes causes cerebral hemodynamic disturbance, brain inflammation and injury in preterm lambs," *PLoS ONE*, vol. 7, no. 6, art. e39535, 2012.

[4] M. A. Bassani, F. Mezzacappa Filho, M. R. C. Coppo, and S. T. M. Marba, "An evaluation of peak inspiratory pressure, tidal volume, and ventilatory frequency during ventilation with a neonatal self-inflating bag resuscitator," *Respir. Care*, vol. 57, no. 4, pp. 525–530, 2012.

[5] G. M. Schmölzer, C. J. Morley, C. Wong, J. A. Dawson, C. O. F. Kamlin, S. M. Donath, S. B. Hooper, and P. G. Davis, "Respiratory function monitor guidance of mask ventilation in the delivery room: a feasibility study," *J. Pediatr.*, vol. 160, no. 3, pp. 377–381.e2, 2012.

[6] G. M. Schmölzer, C. J. Morley, and P. G. Davis, "Respiratory function monitoring to reduce mortality and morbidity in newborn infants receiving resuscitation," *Cochrane Database Syst. Rev.*, no. 9, art. CD008437, 2010.

[7] E. Schena, C. Massaroni, P. Saccomandi, and S. Cecchini, "Flow measurement in mechanical ventilation: a review," *Med. Eng. Phys.*, vol. 37, no. 3, pp. 257–264, 2015.

[8] E. Schena, G. Lupi, S. Cecchini, and S. Silvestri, "Linearity dependence on oxygen fraction and gas temperature of a novel Fleisch pneumotachograph for neonatal ventilation at low flow rates," *Measurement*, vol. 45, no. 8, pp. 2064–2071, 2012.

[9] R. C. Baker, *Flow Measurement Handbook: Industrial Designs, Operating Principles, Performance, and Applications*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2016, sec. 8.4.

[10] R. W. Miller, *Flow Measurement Engineering Handbook*, 3rd ed. New York, NY, USA: McGraw-Hill, 1996.

[11] H. K. Versteeg and W. Malalasekera, *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*, 2nd ed. Harlow, U.K.: Pearson, 2007.

[12] World Health Organization, *Guidelines on Basic Newborn Resuscitation*. Geneva, Switzerland: WHO, 2012.

[13] K. Aziz *et al.*, "Part 5: Neonatal resuscitation: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care," *Circulation*, vol. 142, no. suppl. 2, pp. S524–S550, 2020.

[14] A. L. Window, Ed., *Strain Gauge Technology*, 2nd ed. London, U.K.: Elsevier Applied Science, 1992.

[15] P. Horowitz and W. Hill, *The Art of Electronics*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[16] W. Kester, Ed., *Practical Design Techniques for Sensor Signal Conditioning*. Norwood, MA, USA: Analog Devices, 1999.

[17] R. P. Sallen and E. L. Key, "A practical method of designing RC active filters," *IRE Trans. Circuit Theory*, vol. 2, no. 1, pp. 74–85, 1955.

Onboard Gait-Based Anatomical Axis Identification for Stride Detection with a Low-Cost Ankle Wearable

Leonardo Monteiro de Sousa
Dept. of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
leonardoms157@gmail.com

Edilson Rocha Porfirio Filho
Dept. of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
edilson.filho@ufc.br

***Abstract*—Wearable inertial monitoring for amateur sport is largely restricted to commercial devices whose cost places them out of reach in lower-income settings. Cheap hardware closes part of that gap, but without a technician to position the sensor the device cannot assume a known mounting orientation. Functional sensor-to-segment calibration solves this in the biomechanics literature, at the price of supervised protocols with dedicated postures and offline processing. This work implements such a calibration entirely onboard an ankle-worn wearable built from retail components for about 21 US dollars, with no operator and no movement beyond walking. The vertical axis is recovered from the resting gravity vector and the mediolateral axis from the gyroscope component of highest variance during walking, when sagittal plane rotation dominates; the anteroposterior axis follows by the right-hand rule. Across sixteen mounting configurations, eight rotation angles crossed with two device faces, the assignment matched the mounting geometry and an offline recomputation in all sixteen, and the confidence ratio governing acceptance varied with geometry as predicted, from 1.26 to 21.83 when recomputed over the full recording. A stride detector on the identified axis reproduced a manual count exactly over six gait recordings spanning slow walking to running, 113 strides; under leave-one-out selection the held-out total deviated by at most three strides, 2.7 percent, with a worst case of 12.5 percent on a single recording. This is a single-participant proof of concept.**

***Index Terms*—Wearable sensors, Inertial Measurement Unit (IMU), Gait analysis, Stride detection, Embedded systems, Low-cost instrumentation**

I. Introduction

Regular physical activity lowers the risk of cardiovascular, metabolic and musculoskeletal disease, and is recommended by national and international public health bodies [1]. In Brazil, adherence is both low and unequally distributed. The 2019 National Health Survey found the odds of engaging in leisure-time physical activity roughly three times higher among adults in the highest per capita income bracket than among those in the lowest, with lower participation also among Black respondents and among women [2]. Activities that depend on equipment or on gym membership concentrate in the higher income brackets. Soccer is the exception, the only activity in the 2013 edition of the same survey more prevalent among lower income and among Black and mixed-race respondents, and the leading one among adults aged 18 to 29; outdoor walking, the most prevalent activity in the country, and street running likewise demand little or no infrastructure [3].

These activities are accessible as practice but not as measurement. Tracking cadence, intensity and progression over time requires instrumentation whose cost falls precisely on the population that most relies on low-barrier activities. The gap addressed here is not access to sport, but access to quantitative information about one's own practice. Commercial reference devices illustrate the distance: at prices consulted in July 2026, foot and ankle mounted inertial trackers were listed at 199 and 249 US dollars [4], [5], and entry-level team tracking systems at 320 US dollars [6].

Recent work shows that much of the accuracy penalty of inexpensive hardware can be recovered through signal processing and calibration rather than costlier components. Gascón *et al.* [7] report 97.8 percent accuracy in combined pass and kick detection from an insole-mounted microelectromechanical systems (MEMS) inertial measurement unit sampled at 50 Hz, using an adaptive threshold pipeline. Bernhart *et al.* [8] detect stride events during running from a chest-worn sensor. Chavarría-Fernández *et al.* [9] integrate a microcontroller-class inertial unit for real-time external load monitoring, framed explicitly in a Latin American resource context.

A requirement common to this body of work remains largely implicit: the sensor is assumed to sit in a known orientation relative to the body segment. An adjustable strap fastened by the user offers no guarantee that the sensor axes align with the anatomical axes of the limb, and a device that silently depends on correct placement will fail in exactly the unsupervised, low-resource setting that motivated building it cheaply in the first place.

The biomechanics literature addresses this through functional, or sensor-to-segment, calibration, in which anatomical axes are recovered from the sensor signals themselves rather than assumed from placement [10], [11]. These procedures are well established and accurate. They are also designed for supervised use: they prescribe dedicated calibration postures,

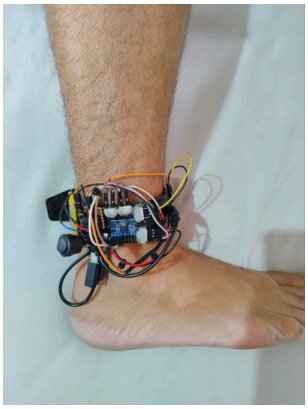


Fig. 1. Prototype strapped to the ankle over the lateral malleolus.

are executed under the guidance of an operator, and are computed offline on a workstation. Each of those assumptions is unavailable here, so the gap this paper addresses is one of implementation rather than of method: what is missing is not a way to recover anatomical axes, but a way to recover them on the device itself, without an operator, without dedicated postures, and within the computational budget of a microcontroller costing a few dollars.

The contributions are a functional ankle-worn prototype assembled from Brazilian retail components at a component cost of 105 Brazilian reais, about 21 US dollars; a fully onboard implementation of gait-based functional calibration, which assigns the vertical, mediolateral and anteroposterior axes to the sensor axes once at the start of each session, with no operator, no calibration posture beyond standing still and no movement beyond walking; an evaluation of that procedure across sixteen mounting configurations, obtained by combining eight strap rotation angles with two device faces, eight of them inverted; and a stride detector built on the identified mediolateral axis, evaluated against a manual reference count over six gait recordings spanning slow walking to running.

II. System Design

The system has two parts: a wearable device that acquires and stores data autonomously during the activity, and desktop tools that retrieve and analyze those data afterwards. This section describes the device.

A. Hardware

The microcontroller is an ESP32-C3 Super Mini, selected for its low cost, compact footprint, internal flash large enough for the recorded sessions, and a USB interface used for both programming and data retrieval. Its radios are not used. The inertial sensor is an MPU6050 module, a MEMS device combining a three-axis accelerometer and a three-axis gyroscope with an inter-integrated circuit (I2C) interface [12], chosen for its low unit cost and wide availability in Brazilian retail. It is configured for full-scale ranges of 16 g and 2000 degrees per second and sampled at 50 Hz. A buzzer and a push button on general-purpose input/output (GPIO) pins 3 and 10 provide

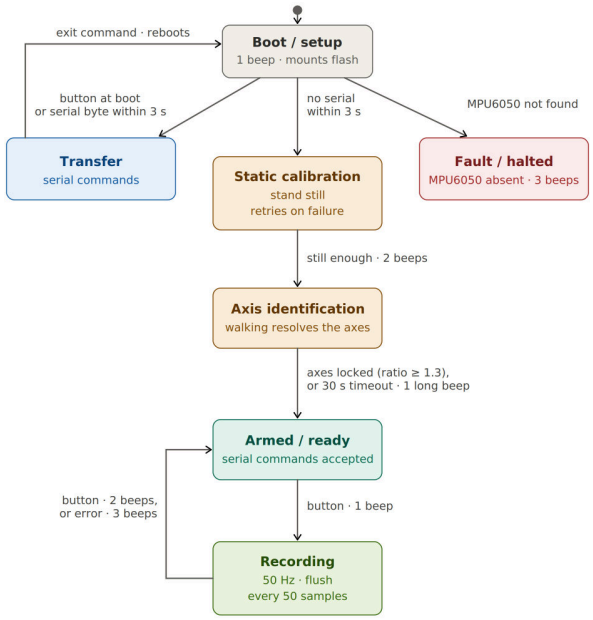


Fig. 2. Firmware state machine. Static calibration and axis identification are mandatory before recording is armed, and the buzzer signals each transition.

the entire user interface, and Fig. 2 shows the firmware state machine.

The assembly is mounted on a repurposed section of leather belt, adjustable by its buckle, and sits over the lateral malleolus, as Fig. 1 shows, the same region instrumented in external load work with runners [9]. Power comes from a single 3.7 V, 250 mAh lithium polymer cell, managed by a combined charger and boost converter module based on the 134N3P integrated circuit. Estimated from component datasheets, the microcontroller and sensor together draw about 48 mA during acquisition [12], [13], which places continuous recording autonomy on the order of five hours. The estimate excludes losses in the boost converter, so it is an upper bound on autonomy; direct measurement remains future work.

B. Recording Architecture

The sensor delivers acceleration and angular velocity to the microcontroller at 50 Hz. At startup the microcontroller runs the static calibration and the orientation identification of Section III, and from then on converts every sample to the body frame with gravity removed. Processed samples are written in binary to internal flash through a LittleFS file system, in sessions delimited with the push button. Each file begins with a fixed 64 byte header carrying the sampling rate, a timestamp, the inferred axis assignment and the resting gravity vector, followed by sample records of 32 bytes each, so a session grows by about 96 kB per minute. The metadata in the header makes each file self-contained, so the raw signal can be reconstructed during analysis without external logs.

The 50 Hz figure is the target of a fixed-rate scheduler, not the rate achieved. Writes to flash occasionally block the loop for longer than one sampling period, and the firmware then re-anchors the schedule rather than emitting the delayed samples

back to back. Measured over the recordings reported here, the achieved rate ranged from 47.8 to 49.5 Hz, with isolated gaps reaching 160 ms during running; each sample carries its own timestamp, so those gaps are visible to the analysis.

The acquisition path runs without external communication. Retrieval happens only after the session ends: the device is connected to a computer over the serial interface, and a script transfers the files and decodes them to comma-separated values (CSV).

C. Cost

At Brazilian retail prices consulted in June 2026, converted at the July 2026 rate of 5.07 reais per dollar, the same month as the commercial prices of Section I, the components total 105.00 reais, about 20.70 US dollars: 35.00 for the ESP32-C3 Super Mini, 25.00 for the lithium polymer cell, 20.00 for the MPU6050 module, 13.00 for buzzer, switch, button and wiring, and 12.00 for the charger and boost module. The strap is repurposed from a belt and is not counted.

The total places the prototype an order of magnitude below the commercial devices of Section I. It covers single-unit retail components only, so quantity sourcing would lower it while an industrialized enclosure and certification would raise it.

III. Automatic Orientation Identification

A strap fastened by the user gives no guarantee that the axes of the sensor align with the anatomical axes of the limb. Without that correspondence the measured components carry no mechanical meaning: there is no way to say which channel describes vertical motion, which rotation in the sagittal plane, and which motion along the direction of travel. The effect of sensor placement and displacement on wearable activity recognition is well documented [14], [15], and three families of response exist.

The first constrains placement, which most sport wearable work implicitly assumes. The second removes the dependence on orientation by transforming the raw signals into quantities invariant to it, discarding the orientation information itself [16], [17]. The third recovers the orientation, either continuously through attitude estimation or once through a functional sensor-to-segment calibration [10], [11].

The procedure described here belongs to the third family and follows its established structure. Carcreff *et al.* [11] recover the vertical axis from gravity in a standing posture, the mediolateral axis from the principal axis of angular velocity during gait, and the sign of that axis from a further seated posture; the first two steps are the ones adopted here. What differs is where the procedure runs. The published methods execute offline, from data recorded under a supervised protocol, whereas the implementation below runs in firmware on the microcontroller, unattended, and asks the wearer for nothing beyond standing still and then walking.

A. Static Calibration

The procedure requires the device to be at rest and averages each axis over one hundred consecutive samples, about two seconds. For the gyroscope, whose expected reading at rest is zero, the per-axis mean is the bias to subtract from subsequent samples. For the accelerometer, the mean is the gravity vector in the sensor frame, used both for gravity removal and for the axis assignment that follows. The same window corrects a scale error: the measured gravity magnitude deviated from the expected value by about 3 percent on the sensor used, so the firmware derives a scale factor forcing the resting magnitude to 9.807 m/s^2, compensating the tolerance of an inexpensive part in software.

Two quality checks accompany the calibration. The gyroscope standard deviation over the window must remain below 0.05 rad/s, about 2.9 degrees per second; above that the wearer moved and the bias estimate is untrustworthy, so the firmware silently discards the window and collects another until a still one is obtained, the beeps that end calibration being emitted only then. The acceleration magnitude is compared against the expected gravity with a tolerance of 0.5 m/s^2; a deviation beyond it points to a scale or mounting problem and is reported, but does not block the sequence, since the scale factor normalizes the magnitude in any case. Parameters are recomputed at every power-up and not persisted, which keeps the calibration representative of the conditions of use.

B. Axis Assignment

The vertical axis follows directly from the static calibration. Among the three sensor axes, the one carrying the largest component of the resting gravity vector is taken as vertical, and the sign of that component fixes its positive direction. This resolves one axis but leaves the other two ambiguous, since a static reading cannot distinguish rotation about the vertical.

Resolving the remaining two exploits a property of human gait: during walking, rotation of the lower limb concentrates in the sagittal plane, about the mediolateral axis of the ankle. That axis is therefore the one of largest angular velocity variance during walking, while the others remain comparatively quiet. This is the principle underlying the gait-based step of functional calibration [10], [11]; the contribution here is in reducing it to a form that a microcontroller can execute online during the session itself. After the static calibration the device enters a phase in which the user walks or swings the leg for a few seconds. During that phase the firmware accumulates angular velocity statistics only over samples with effective motion, identified by an angular velocity magnitude above a threshold of 30 degrees per second. Of the two candidate axes orthogonal to the known vertical, the one of larger variance is identified as mediolateral. The third axis, anteroposterior, follows by the right-hand rule, which also fixes a consistent triad. Continuous attitude estimation would also resolve the axes [18], but a stride detector needs the assignment, not the instantaneous attitude, and resolving it once costs one pair of accumulators per axis and a single comparison. Counted from the firmware sources, the identification holds 62 bytes of persistent state and the static calibration 44 more, so the procedure occupies 106 bytes of RAM, a constant number of operations per sample, and no filter state that could drift. For

scale, the complete firmware image occupies 330 kB of the 1.25 MB application partition and reserves 14.5 kB of the 320 kB of RAM at link time, so the procedure accounts for under 1 percent of an allocation that the linker reports for static memory and not for the heap and stack.

Two safeguards constrain the decision. The first is a confidence criterion: the mediolateral assignment is accepted only when the ratio between the largest and the second largest variance reaches 1.3, below which the difference is too small to trust and the firmware keeps accumulating samples. The second is a timeout of thirty seconds, counted from the first sample of the identification phase, which bounds that wait and forces an outcome. Section III-C describes the three outcomes and what each one writes to the session file.

The sequence is conveyed to the wearer by the buzzer alone, without a screen or a paired application: one beep at power-up to stand still, two beeps when calibration succeeds and walking should begin, one long beep when the assignment is locked and recording is armed.

C. Fallback Behavior

The identification phase ends in one of three ways, and the firmware records which one in the header of the session file. It locks on merit once at least one hundred samples above the motion threshold have accumulated, about two seconds of walking, and the variance ratio has reached 1.3; the session carries no flag. It locks on timeout with partial evidence when thirty seconds elapse with at least fifty such samples: the mediolateral axis is still assigned by largest variance, and the session is flagged as having uncertain orientation when the ratio at that moment is below 1.3. A ratio above 1.3 leaves the session unflagged, since the criterion the flag records is the confidence in the assignment and not the route by which the phase ended. It locks on timeout with no usable evidence when the wearer does not walk: the vertical axis is kept, since gravity alone determines it, the remaining two fall back to the sensor axes orthogonal to it taken in cyclic order, and the session is flagged. The flag occupies one bit of the 64 byte header, so the analysis can hold flagged sessions apart, and no file can lack an assignment, since the button that starts a recording stays inert until the phase ends.

Slipping of the strap is a separate case that none of the three covers. Calibration and identification run once, at power-up, and the firmware neither re-estimates the body frame afterwards nor monitors the device for movement on the leg, so a device that slips after the lock keeps reporting the frame identified at the start, with nothing in the data marking it stale.

D. Sign Ambiguity

Variance is signless, so selecting the mediolateral axis by largest variance determines which sensor axis lies along that direction but not which way it points. Rotating the device by 180 degrees on the strap makes the same forward leg swing produce angular velocity of opposite sign on the same axis, and nothing in the accumulated statistics distinguishes the two cases. The body frame is therefore determined only up to a sign on the mediolateral and anteroposterior axes. The consequences are bounded here, since every quantity used in this work is invariant to that sign: stride detection operates on the magnitude of the sagittal gyroscope component. The ambiguity would matter for a directional quantity, such as telling a forward from a backward kick.

A published solution exists. Carcreff *et al.* [11] resolve the sign with a seated posture with the leg extended, in which the direction of the measured acceleration disambiguates the axis. That step was deliberately not adopted, because it reintroduces what the implementation sets out to avoid: a prescribed posture, and therefore an operator or a screen to instruct it. Whether the sign can instead be recovered from the direction of travel during the identification walk is left as future work.

IV. Experimental Protocol

All data reported in this paper were collected from a single participant, an adult male, the first author of this work. That scope conditions how every result below should be read: the evaluation is a controlled study of the procedure across mounting configurations, not a study of its performance across bodies.

All recordings were made at 50 Hz with the device on the right ankle over the lateral malleolus. Static calibration was performed standing, in a posture close to that of the activity that followed, which keeps the resting gravity reference representative of the recording conditions.

A. Mounting Experiment

The mounting configuration was treated as a controlled variable with sixteen levels: eight rotation angles of the device on the strap crossed with the two faces on which it can be fastened.

Rotation is about the long axis of the shank: the strap remains closed and the device travels around the ankle. The reference of 0 degrees is the buckle positioned at the front of the leg, which places the device at the back of the leg. Angles increase in steps of 45 degrees in a single direction over a full turn: 0, 45, 90, 135, 180, 225, 270 and 315 degrees.

The two faces differ by turning the device over on the strap, which reverses the sign of the sensor axis that reads gravity. Eight of the sixteen configurations are inverted in this sense. The angular reference was set again after the device was turned over, so 0 degrees in the inverted set is not the same physical position as in the upright set, and angles are comparable only within a set. No attempt was made to align the sensor axes with the anatomical axes; the strap was simply fastened at the intended angle.

Each configuration was exercised over an identical route: a few steps forward, a 180-degree turn, a few steps back, a second 180-degree turn in the opposite direction, returning to the starting position and heading. The route provides the walking motion the identification requires, and the two turns add rotation about the vertical axis, which exercises the assignment under motion that is not gait. Two quantities were recorded per configuration: the axis the firmware assigned to each

anatomical direction, and the magnitude of the resting gravity vector. The number of steps was neither fixed nor counted, since these recordings exercise the orientation procedure rather than the detector.

B. Gait Recordings

A second set of six recordings, made in the canonical mounting, covers the range of locomotion in scope: one slow walk, two regular walks of different duration, one fast walk, one jog and one run. Together they span estimated stride periods from about 1.9 s at the slowest to about 0.7 s at the fastest, and total 113 strides. The labels record the pace the participant intended, not a measured one, and the fast walk turned out close in period to the shorter of the regular walks.

The reference for each recording is a manual count performed during execution, one stride per forward swing of the instrumented leg, so each swing corresponds to one detection. Total bilateral cadence follows by doubling, under an assumption of symmetry between legs. Each recording is delimited by a button press at its start and end, and the movement of reaching for the button is treated as part of the signal rather than trimmed away, since a detector intended for unsupervised use has to tolerate it.

C. Kick Recordings

Two further recordings, of ten and five maximum-force kicks, were made with the same device and mounting to probe the upper end of the acceleration range. They carry no stride reference and are used only in the discussion of sensor saturation in Section VI-B. Impacts were located as peaks of the acceleration magnitude and checked axis by axis against the full-scale limit on the raw signal, reconstructed by adding the resting gravity vector back to the recorded dynamic acceleration.

V. Results

A. Orientation Identification Across Mountings

Table I reports the assignment and the confidence ratio for each configuration.

The vertical axis was resolved as expected in all sixteen configurations. In every case the firmware assigned it to the sensor x-axis, with the sign tracking the face on which the device was mounted: positive when upright and negative when inverted.

The mediolateral assignment tracked the rotation of the device, as geometry requires: rotating it about the long axis of the leg exchanges the roles of the two sensor axes orthogonal to the vertical, so the axis nearest the mediolateral direction changes every 90 degrees. With no external motion reference, the assignment is checked against two expectations. It follows that alternation, in pairs within each set, and a post hoc recomputation of the variance on the two candidate axes, over the samples above the motion threshold, selected the same axis in all sixteen configurations. The results below report agreement with these two checks, not an angle measured against an anatomical reference.

TABLE I
Axis Assignment and Confidence Ratio Across the Sixteen Mountings

	Upright ($V = +X$)		Inverted ($V = -X$)	
Angle (deg)	Mediolat. axis	Confidence ratio	Mediolat. axis	Confidence ratio
0	$+Y$	6.14	$+Y$	3.38
45	$+Y$	1.30	$+Z$	21.83
90	$+Z$	18.94	$+Z$	1.52
135	$+Z$	**1.26**	$+Y$	8.06
180	$+Y$	7.39	$+Y$	1.50
225	$+Y$	1.54	$+Z$	10.84
270	$+Z$	13.32	$+Z$	2.03
315	$+Z$	2.12	$+Y$	9.24

Signs follow the right-hand convention and carry no anatomical meaning (Section III-D).

The confidence ratio is the more informative column, and its variation is itself a result. It ranged from 1.26 to 21.83, following the geometry of the mounting: when a sensor axis nearly coincides with the mediolateral direction the variance concentrates on it and the ratio is large, and when the mediolateral direction falls between two sensor axes the variance splits and the ratio approaches unity. The ratio therefore measures how well posed the assignment is, rather than acting as an arbitrary constant, which supports its use as an acceptance criterion.

Two observations qualify this. First, the two sets of eight configurations are offset from each other by about 45 degrees. This follows from the angular reference having been set again after the device was turned over (Section IV-A): the upright set shows high ratios at 0, 90, 180 and 270 degrees, the inverted set at 45, 135, 225 and 315 degrees. The two sets therefore do not sample the same physical angles, and the angle column is to be read as relative within each set. Second, one configuration fell below the acceptance threshold of 1.3, shown in bold in Table I, and a second sits exactly at it, yet neither was flagged as uncertain by the firmware. This follows the firmware logic rather than contradicting it. The two ratios are not the same quantity: Table I reports the ratio over the full recording, whereas the firmware evaluates it over the shorter identification window and, before the timeout, does not lock until that window ratio reaches 1.3. Both sessions locked well inside the timeout, as the timestamps below show, so the window ratio had reached the threshold, which the full recording, with its two turns and the handling at the edges, need not reproduce. The window value is not logged, so this account rests on the firmware logic; recording it at the instant of the lock would settle the point directly and is left as future work.

The session timestamps also constrain how each assignment was reached. They start with the identification phase, the instant from which the timeout is counted, and the button stays inert until the assignment locks, so the first recorded sample bounds the locking instant from above. In fifteen of the sixteen sessions it falls between 10.9 and 18.9 s, inside the 30 s bound, so those locked on merit, the outcome that leaves a session unflagged (Section III-C). The remaining one, at 45 degrees inverted, begins at 40.0 s, so a lock on merit followed by a

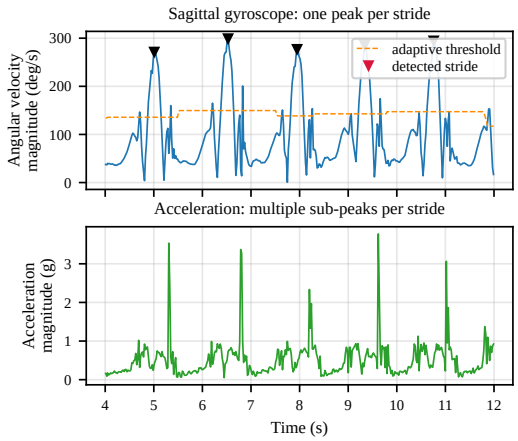


Fig. 3. Candidate input signals over the same walking segment. Triangles mark detected strides, the dashed line the adaptive threshold.

long wait before the button press cannot be told from a lock on timeout. Both routes assign the axis from the variance, and the recorded triad, $+Z$ and $+Y$ rather than the cyclic default $+Y$ and $-Z$, excludes a lock with no usable evidence.

The magnitude of the resting gravity vector was 9.807 m/s^2 in all sixteen configurations, to within 0.001. This confirms that the calibration completed in every mounting, but is not independent evidence of correct decoding, since the calibration derives a scale factor that forces precisely this magnitude.

B. Choice of Input Signal

Two candidate input signals were compared on the same recorded data. The first is the magnitude of the acceleration vector, the immediate candidate because it summarizes effort in a single scalar independent of sensor orientation. The second is the component of angular velocity in the sagittal plane, associated with rotation of the shank during the swing of the leg. Fig. 3 contrasts the two over the same walking segment.

The acceleration magnitude shows several sub-peaks per stride, corresponding to the successive phases of foot support: heel strike, midstance and push-off. These have no one-to-one correspondence with strides and lead to overcounting. The sagittal gyroscope component shows a single well-defined peak per stride, the forward swing of the instrumented leg, which permits direct counting. The choice is an empirical finding, not an inherited convention: no fixed threshold on acceleration magnitude resolved the sub-peaks in these recordings, which are comparable in amplitude to the peak itself and are genuine features of the signal rather than noise.

C. Stride Detection

The detector operates on the magnitude of the sagittal gyroscope component and combines three elements. The first is an adaptive threshold: a sample qualifies as a stride peak candidate only if it exceeds a fraction of the largest amplitude observed over a two second window centered on that sample, set to 0.5, which scales the threshold to the local intensity of the movement, following the adaptive threshold formulation of [7]. The second is a global amplitude floor: beyond the local threshold, a sample must also exceed a fraction of the largest peak of the entire recording, set to 0.3. This discards low-amplitude spurious peaks such as those generated by handling the device at the start and end of a session, which is why the edges of each recording could be left untrimmed. The third is an adaptive refractory period: after each detection, a minimum interval is imposed before the next, set to 0.6 of the stride period. That period is estimated automatically by autocorrelation of each recording [19], with the search restricted to lags between 0.30 and 3 s, which brackets the cadences in scope and keeps the zero-lag peak out of the estimate. When two candidates fall closer than the refractory interval, the smaller one is discarded.

Two of these elements look ahead in the signal: the local threshold reads a window centered on the sample, and the period is estimated over the whole recording. The detector as evaluated here is therefore not causal, consistent with running it after transfer but not during the activity. Moving it into the firmware, listed among the directions in Section VII, requires backward-looking equivalents of both, a trailing window and a period estimated from the intervals already detected, which would alter the operating point and has to be evaluated on its own terms.

The adaptive refractory period is what allows a single parameter set to operate from slow walking to running, specializing for the sagittal gyroscope signal at the ankle an idea established for inertial stride detection during running [8].

TABLE II
STRIDE COUNTS BY RECORDING, WITH PARAMETERS TUNED ON ALL SIX RECORDINGS AND UNDER LEAVE-ONE-OUT SELECTION

Recording	Period (s)	Manual count	Tuned count	Held-out count	Worst error strides (%)
Slow walk	1.90	17	17	17	0
Regular walk	1.42	16	16	16	0
Regular walk	1.16	23	23	23 to 24	1 (4.3)
Fast walk	1.14	16	16	16	0
Jog	0.76	24	24	21 to 24	3 (12.5)
Run	0.66	17	17	17 to 19	2 (11.8)
Total		113	113	110 to 116	3 (2.7)

Worst error is the largest deviation of the held-out count from the manual count; ranges follow from ties at zero training error.

Table II reports the detected and manual counts for each gait recording. With a single parameter set, the detector reproduces the manual count exactly in all six recordings. The range of cadence covered is wide: the estimated stride period in the table varies from 1.90 s in the slow walk to 0.66 s in the run, and the same configuration remains valid across that entire range.

The three parameters were selected over a grid of 36 combinations, and the values above are its optimum on the six recordings; the same grid is the one the leave-one-out selection of Section V-D searches. A wider sweep over 210 combinations locates this operating point on a plateau rather than a spike: 8 of the 210 reproduce the total exactly, 46 miss at most one stride of the total and 81 miss at most three.

The global amplitude floor was the element that brought the count to exact agreement, by suppressing the peaks produced by handling the device at the edges of each session.

This result measures the consistency of the method over the collected set, since the parameters were calibrated on the same recordings used to evaluate them. Section V-D separates tuning from evaluation.

D. Held-Out Evaluation

The counts of Table II come from parameters chosen on the same six recordings they were scored against. To separate tuning from evaluation, the selection was repeated in a leave-one-out scheme: for each recording, the three parameters were chosen on the remaining five over the same grid of 36 combinations of Section V-C, and applied to the recording held out.

Training error reaches zero in every fold, and at more than one grid point: between four and ten of the 36 combinations tie at zero on the five training recordings. Four of the 36 reproduce the manual count exactly in every one of the six recordings, a condition stricter than matching the total, and are among the tied optima of every fold, so a fold resolving the tie inside that set is exact on the held-out recording. The remaining tied combinations, present in three of the six folds, are what produce error.

Table II reports the bounds this leaves and the worst-case error per recording, in strides and in percent. The largest is three strides on the jog, 12.5 percent of its reference count. Summed across folds, the held-out total ranges from 110 to 116 against a reference of 113, an absolute error of at most three strides on the total and a percentage error of at most 2.7 percent.

Ranges are reported instead of single values because the tie among optimal training configurations is real, and any fixed convention would hide it. This is the figure to read for the method, and it remains restricted to one participant.

VI. Discussion and Limitations

A. Scope of the Claim

What was varied across sixteen levels is the mounting configuration, not the wearer, so the evidence supports invariance of the procedure to how the device is fastened, on the gait of one person, and not yet invariance across gaits. The distinction matters because the mechanism has an identifiable dependency. Identifying the mediolateral axis by largest angular velocity variance assumes that rotation of the lower limb concentrates in the sagittal plane during walking. This holds for typical gait, and the margin by which it holds is what the confidence ratio of 1.3 measures. A wearer with pronounced gait asymmetry, an altered gait from injury, or a markedly different limb geometry could reduce that margin. Whether the procedure then degrades gracefully or fails is an open question only multi-subject data can answer, and the timeout fallback exists because that failure mode was anticipated rather than observed.

B. Limitations

The sample is a single participant, the principal limitation: the work is a preliminary validation of a procedure, not a characterization of its performance across a population.

The exact agreement of Table II states internal consistency, since those parameters were chosen on those same recordings; the leave-one-out evaluation of Section V-D is bounded rather than exact and is the figure to read for what to expect elsewhere. Both are limited by the size of the set: six recordings leave each fold with five for tuning, which is why several parameter combinations tie at zero training error.

The orientation is identified once and not re-estimated, so the body frame becomes stale if the device shifts on the strap. The recordings bound the effect only loosely. Over the still intervals they contain, one second with peak angular rate below 7 degrees per second and mean acceleration magnitude within 0.2 m/s^2 of gravity, the measured direction departs from the gravity vector stored at calibration by 2 to 9 degrees, the larger values in the more vigorous recordings. A session in which the wearer only stood shows 2 to 5 degrees, so part of that spread is posture and noise rather than slip, and nothing in the data separates the two. Stride detection is unaffected, since a constant tilt offset does not reach the gyroscope, but any quantity derived from the direct-current (DC) level of gravity-removed acceleration would inherit it.

The invariance of stride detection to mounting is inferred rather than measured. The mounting experiment establishes that the axis assignment is recovered regardless of how the device is fastened, and the gait recordings establish that the detector is accurate on the identified axis in the canonical mounting. The step between the two is reasonable, since the detector consumes only the identified axis, but the mounting recordings carry no reference count against which detection could be scored. Repeating them with counted steps would close this gap at low cost.

The sign of the mediolateral and anteroposterior axes remains unresolved, as discussed in Section III-D.

Three quantities are estimated or counted statically rather than measured: autonomy follows from datasheet currents, execution time was not profiled, and the ratio the firmware evaluates over the identification window is not written to the session file, which leaves the account of Section V-A resting on the firmware logic rather than on a logged value. The last two ask for instrumenting the firmware, the first for a current measurement at the cell.

The accelerometer full-scale range of 16 g proved insufficient for the most intense event of the target sport, the instep kick [20]. Of the fifteen kicks performed in the recordings of Section IV-C, fourteen were located as distinct peaks, and twelve of those drove at least one axis against the limit, where the signal stops rising and flattens at about 15.5 g.

That plateau is a clipping artifact and not a measurement: the sensor clips at its hardware limit, and the static calibration then scales every sample by a fixed factor, close to 0.97 on the unit used, which yields the 15.5 g observed. Since that factor is recomputed at every power-up, the plateau differs

between sessions, 15.50 g in one recording and 15.56 g in the other, which is the signature clipping predicts. The true peak is therefore unknown. Widening the range would require a different part, since 16 g is the largest setting the MPU6050 offers [12]. Clipping is confined to the accelerometer: the gyroscope, set to 2000 degrees per second, peaked at 1330 on a single axis on the same kicks and at 641 across the gait recordings of Section IV-B, both compared against a full scale that is itself per axis, so for explosive events the angular velocity remains the valid quantity.

The validation reference is a manual count performed during execution, subject to human error at high cadence, and it carries no timing: a total per recording, not the instant of each stride. Accuracy is therefore a count error, which Table II reports, and not precision and sensitivity on individual events, since neither a true positive nor a false positive can be defined without a per-event reference, which a synchronized video would supply at low cost. Recordings were short, tens of strides each, so behavior over sessions of realistic duration remains untested.

VII. Conclusion

This paper presented, as a proof of concept, an ankle-worn inertial wearable assembled from retail components at about 21 US dollars, which recovers its own anatomical axes from the gait of the user: the vertical axis from the resting gravity vector, the mediolateral axis from the gyroscope component of largest variance during walking, and the anteroposterior axis by the right-hand rule. The principle is that of functional sensor-to-segment calibration, established in the biomechanics literature; the contribution is an implementation that runs entirely in firmware, requires no operator, no mounting instructions and no calibration posture beyond standing still, and completes once at the start of each session.

Evaluated across sixteen mounting configurations, including eight in which the device was inverted, the assignment matched both checks in every case, and the confidence ratio governing acceptance varied with mounting geometry in the manner geometry predicts. A stride detector built on the identified mediolateral axis, using an adaptive threshold with an adaptive refractory period, reproduced a manual reference count exactly over six gait recordings in the canonical mounting, spanning slow walking to running. Obtained from a single participant, these results establish feasibility within the computational budget of the target microcontroller, and not accuracy across wearers, across gait patterns, or over sessions of realistic duration.

Three directions follow. The first and most important is multi-subject data collection, recorded against a synchronized video so that detections can be scored event by event instead of against a manual total. It would turn invariance to mounting on one gait into invariance across gaits, and is a precondition for any performance claim. The second is resolving the sign ambiguity of the mediolateral and anteroposterior axes without a prescribed posture, plausibly from the direction of travel during the identification walk. The third is moving the detector into the firmware, so that stride counts become available during the activity rather than after transfer, which requires the causal reformulation of Section V-C.

References

[1] F. C. Bull, S. S. Al-Ansari, S. Biddle, K. Borodulin, M. P. Buman, G. Cardon, C. Carty, J.-P. Chaput, S. Chastin, and R. Chou, "World Health Organization 2020 guidelines on physical activity and sedentary behaviour," *British Journal of Sports Medicine*, vol. 54, no. 24, pp. 1451–1462, 2020.

[2] D. K. A. Cruz, K. S. d. Silva, M. V. V. Lopes, F. R. Parreira, and H. M. Pasquim, "Iniquidades socioeconômicas associadas aos diferentes domínios da atividade física: resultados da Pesquisa Nacional de Saúde 2019," *Epidemiologia e Serviços de Saúde*, vol. 31, no. spe1, 2022.

[3] M. G. Lima, D. C. Malta, C. N. Monteiro, N. F. d. S. Sousa, S. R. Stopa, L. d. P. B. Medina, and M. B. d. A. Barros, "Leisure-time physical activity and sports in the Brazilian population: A social disparity analysis," *PLOS ONE*, vol. 14, no. 12, p. e0225940, Dec. 2019.

[4] Playermaker, "Playermaker: AI soccer performance tracker," 2026, accessed: Jul. 14, 2026. [Online]. Available: https://www.playermaker.com/

[5] Stryd, "Stryd: Run with power," 2026, accessed: Jul. 14, 2026. [Online]. Available: https://www.stryd.com/

[6] STATSports, "STATSports Academy: GPS performance tracker," 2026, accessed: Jul. 14, 2026. [Online]. Available: https://statsports.com/

[7] A. Gascón, Á. Marco, D. Buldain, J. Alfaro-Santafé, J. V. Alfaro-Santafé, A. Gómez-Bernal, and R. Casas, "A diagnostic and performance system for soccer: Technical design and development," *Sports*, vol. 13, no. 1, p. 10, Jan. 2025.

[8] S. Bernhart, E. Harbour, S. Kranzinger, U. Jensen, and T. Finkenzeller, "Wearable chest sensor for stride and respiration detection during running," *Sports Engineering*, vol. 26, no. 1, Mar. 2023.

[9] V. Chavarría-Fernández, D. Rojas-Valverde, R. Gutiérrez-Vargas, C. Meza, and M. Méndez-Solano, "Wearable sports technology development in Costa Rica: Inertial measurement unit integration for real-time external load monitoring," *MHSalud: Revista en Ciencias del Movimiento Humano y Salud*, vol. 20, no. 1, pp. 1–13, Jan. 2023.

[10] M. Nazarahari and H. Rouhani, "Semi-automatic sensor-to-body calibration of inertial sensors on lower limb using gait recording," *IEEE Sensors Journal*, vol. 19, no. 23, pp. 12 465–12 474, 2019.

[11] L. Carcreff, G. Payen, G. Grouvel, F. Massé, and S. Armand, "Three-dimensional lower-limb kinematics from accelerometers and gyroscopes with simple and minimal functional calibration tasks: Validation on asymptomatic participants," *Sensors*, vol. 22, no. 15, p. 5657, Jul. 2022.

[12] InvenSense Inc., *MPU-6000 and MPU-6050 Product Specification Revision 3.4*, InvenSense Inc., Sunnyvale, CA, USA, 2013, document Number: PS-MPU-6000A-00.

[13] Espressif Systems, "ESP32-C3 series datasheet," 2025, accessed: Jun. 14, 2026. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32-c3_datasheet_en.pdf

[14] O. Banos, M. Toth, M. Damas, H. Pomares, and I. Rojas, "Dealing with the effects of sensor displacement in wearable activity recognition," *Sensors*, vol. 14, no. 6, pp. 9995–10 023, Jun. 2014.

[15] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, no. 3, pp. 1–33, Jan. 2014.

[16] A. Yurtman and B. Barshan, "Activity recognition invariant to sensor orientation with wearable motion sensors," *Sensors*, vol. 17, no. 8, p. 1838, Aug. 2017.

[17] A. Yurtman, B. Barshan, and B. Fidan, "Activity recognition invariant to wearable sensor unit orientation using differential rotational transformations represented by quaternions," *Sensors*, vol. 18, no. 8, p. 2725, Aug. 2018.

[18] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, "Estimation of IMU and MARG orientation using a gradient descent algorithm," in *Proc. IEEE Int. Conf. Rehabil. Robot. (ICORR)*, Jun. 2011, pp. 1–7.

[19] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2010.

[20] Z. Zhou, Z. Gao, F. Li, D. Wang, Y. Wang, G. Fekete, and Y. Gu, "Comparison of interlimb coordination during soccer instep kicking between elite and amateur players," *European Journal of Sport Science*, vol. 25, no. 9, Aug. 2025.

Performance tests of silicone- and graphite-based surface electrodes, developed for use in electromyography and electrical stimulation equipment

Maria Clara Bernardes Ferrer
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
maria.ferrer@geb.inatel.br

Rita Elizabeth Santos de Almeida
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
r.elizabeth@inatel.br

Elisa Rennó Carneiro Déster
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
elisa.renno@inatel.br

Maysa Costa Alves
Biomedical Engineering
Instituto Nacional de Telecomunicações (Inatel)
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

***Abstract*—The project aimed to organize an experimental procedure for evaluating the functionality of electrodes for electromyography and muscle electrical stimulation. The methodology employed includes output-tracking agreement tests, as well as laboratory assays to detect the presence of visual electromagnetic noise. The model was developed to analyze the device's ability to transmit electrical impulses and capture the electrical potential of the muscle. The purpose was to provide a protocol to standardize the evaluation of electrode performance, contributing to the development and validation of more reliable devices.**

***Index Terms*—electrical stimulation, electrode, electromyography, experimental testing, performance evaluation, validation.**

I. Introduction

Surface electrodes are non-invasive devices that are applied to the patient's skin and are used to acquire electrical activity generated by motor units or to deliver electrical stimulation to the targeted muscle region [1].

The action potential is a rapid change in the membrane electrical potential of neurons and muscle cells, triggered by an adequate stimulus capable of activating these cells. It is responsible for initiating muscle contraction [2].

The summation of the action potentials recorded by a surface electrode over a muscle after its stimulation results in the compound muscle action potential (CMAP) [3].

The ABNT NBR IEC 60601 standard establishes general requirements for the basic safety and essential performance of medical electrical equipment. It defines criteria and tests for assessing device compliance, contributing to the safety of patients and healthcare professionals while ensuring proper device operation [4].

Fundação Instituto Nacional de Telecomunicações (FINATEL).

In the book Medical Instrumentation: Application and Design, Chapter 5 – Biopotential Electrodes, Webster and Nimunkar emphasize that electrode performance directly affects the quality of acquired biopotential signals. The authors highlight the importance of evaluating skin–electrode interface impedance, polarization, half-cell potential, and susceptibility to noise and artifacts to ensure reliable physiological signal acquisition [5].

In 1848, Emil du Bois-Reymond, considered the German founder of modern electrophysiology, demonstrated in his work Untersuchungen über thierische Elektricität that muscle contraction is accompanied by changes in muscle electrical activity, a phenomenon known as "negative variation," providing evidence that muscles and nerves generate electrical signals [6].

Surface electromyography is widely used for the assessment of skeletal muscle electrical activity through electrodes positioned on the patient's skin. It provides information regarding muscle control by the central and peripheral nervous systems. Its application has become increasingly relevant in several areas, such as obstetrics and neurorehabilitation [7].

Neuromuscular electrical stimulation produces contractions on the muscle surface, being responsible for increasing oxidative capacity and reducing inflammation [8].

Electrotherapy can be applied in the treatment of fractures, nerve fiber repair, soft tissue regeneration, treatment of peripheral nervous system injuries, cancer treatment, and may also influence cellular behavior [9], [10].

Electromagnetic interference can be emitted by equipment and can compromise the operation of electronic devices and electrical systems [11].

Based on this issue, the objective of this study was to develop a document containing test guidelines for the validation of devices with electromyographic signal acquisition for human muscle assessment and electrical stimulation, aiming to provide more accurate diagnosis and improved patient recovery.

II. Methodology

In order to evaluate the functional capability of the electrode, a survey was conducted on the different feasible tests for application in this device.

Experimental tests were developed for two types of electrodes, which differ only in shape and dimensions. That is, both present the same graphite- and silicone-based composition.

The main characteristics to be evaluated were:

- Conductivity: ability to conduct electrical current;
- Electromagnetic noise: unwanted signals that may impair device operation;
- Output-tracking agreement: changing the intensity of the electrical stimulator and analyzing the sensor response.

After evaluating these parameters, it is determined whether the electrode meets the necessary requirements for proper operation, that is, providing a preliminary assessment of its electrical performance under the evaluated experimental conditions.

III. Results and Discussion

The employed methodology resulted in the following steps for the analysis and evaluation of the device:

A. Electrical conductivity

The evaluation of electrical conductivity is essential to ensure that the electrical stimulator delivers the electrical signal through the electrode to the patient's muscle at the desired intensity, as previously programmed.

For this approach, an electrical stimulator, a 330 Ω resistor, and an oscilloscope were employed.

Figure 1 shows the adopted positioning.

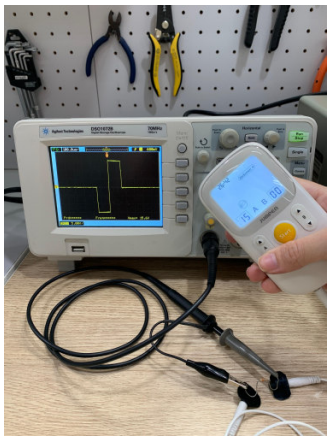


Fig. 1. Oscilloscope measurement. Source: Author (2025)

In the oscilloscope, the horizontal scale (time represented by each horizontal division of the oscilloscope grid) was set to 100.0 μs/div. Additionally, the vertical scale (voltage represented by each vertical division) was set to 5 V/div.

Subsequently, the oscilloscope probe was connected to the oscilloscope, with one clip connected to one terminal of the resistor and the other clip connected to the opposite terminal.

The electrical stimulator was connected to a pair of electrodes, one of which was under evaluation. Each terminal of the resistor was positioned on one of the electrodes.

The electrical stimulator was initially configured to reach its maximum voltage to be applied to the electrode.

For the individual evaluation of each device, a reference electrode with previously validated performance was used. The reference electrode was connected to one terminal of the electrical stimulator, while the electrode under evaluation was connected to the opposite terminal.

Furthermore, to assess electrical consistency across different regions of the electrode surface, meaning uniform current conduction throughout the electrode area, the Cartesian Plane proposed by René Descartes was used, in which each quadrant represented an area to be analyzed (Figures 2 and 3).



Fig. 2. Cartesian Plane markings for the smaller electrode. Source: Author (2025)

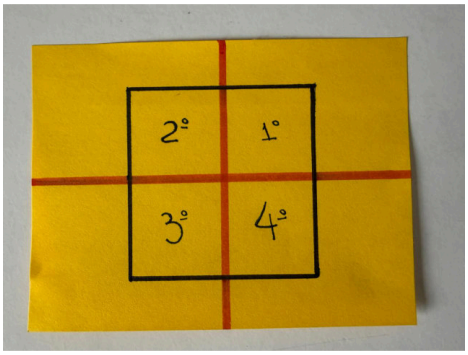


Fig. 3. Cartesian Plane markings for the larger electrode. Source: Author (2026)

The positioning employed for the analysis of homogeneous current distribution in the smaller electrode can be observed in Figures 4, 5, 6, and 7. The positioning for the larger electrode can be observed in Figures 8, 9, 10, and 11.

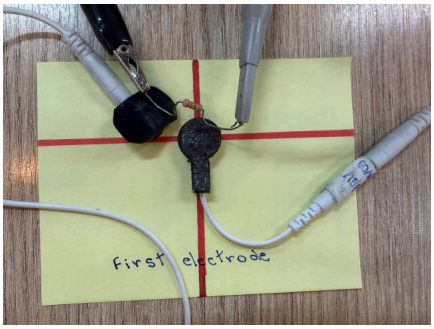


Fig. 4. Positioning in the first quadrant of the Cartesian Plane for the smaller electrode. Source: Author (2025)

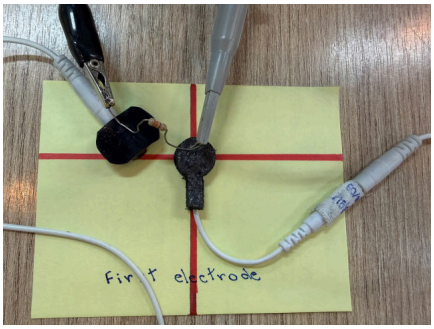


Fig. 5. Positioning in the second quadrant of the Cartesian Plane for the smaller electrode. Source: Author (2025)

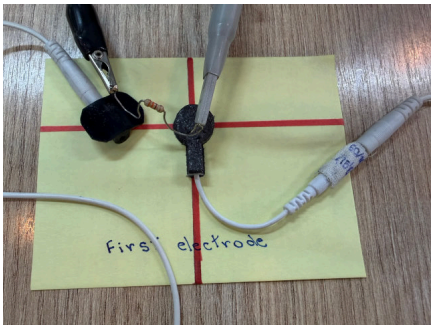


Fig. 6. Positioning in the third quadrant of the Cartesian Plane for the smaller electrode. Source: Author (2025)

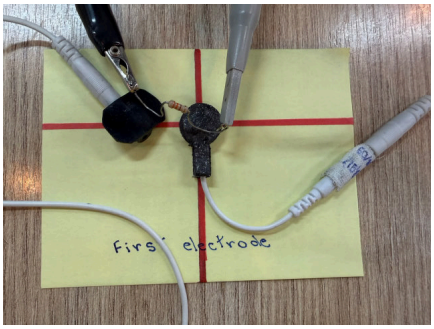


Fig. 7. Positioning in the fourth quadrant of the Cartesian Plane for the smaller electrode. Source: Author (2025)

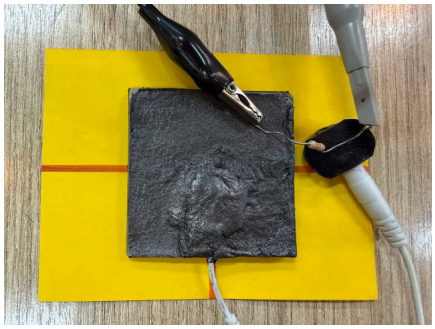


Fig. 8. Positioning in the first quadrant of the Cartesian Plane for the larger electrode. Source: Author (2026)

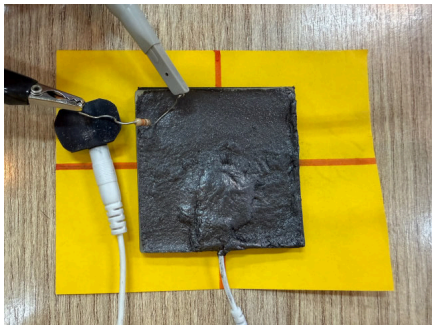


Fig. 9. Positioning in the second quadrant of the Cartesian Plane for the larger electrode. Source: Author (2026)

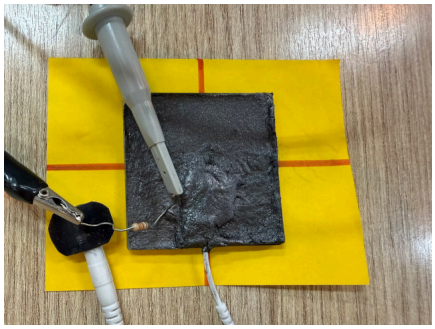


Fig. 10. Positioning in the third quadrant of the Cartesian Plane for the larger electrode. Source: Author (2026)

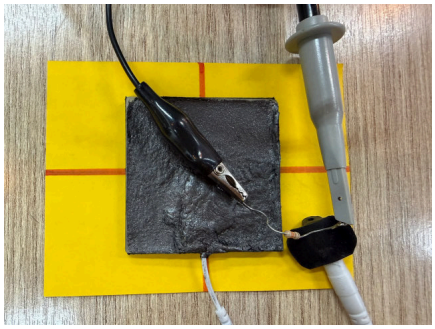


Fig. 11. Positioning in the fourth quadrant of the Cartesian Plane for the larger electrode. Source: Author (2026)

The results obtained for each quadrant were recorded in a table using the electronic spreadsheet software Microsoft Excel for subsequent evaluation and comparison with the gold standard observed in a previously approved electrode.

The obtained results can be seen in Table I, which showed the same measured voltage at the four tested positions under the evaluated conditions, both for the smaller and larger electrodes.

TABLE I
CURRENT HOMOGENEITY TEST.

Position:	**1**	**2**	**3**	**4**
Voltage (V):	14.8	14.8	14.8	14.8

Furthermore, when compared with the voltage value obtained from the analysis of a previously approved electrode, the results were the same. That is, when the electrical stimulator supplied its maximum applied voltage to the electrode, both the electrodes under evaluation and the gold standard presented a voltage of 14.8 V on the oscilloscope.

B. Visual Inspection for Electromagnetic noise

The visual inspection for electromagnetic noise test is necessary because it evaluates whether there is interference from unwanted signals that may affect the electromyography and electrical stimulation of the patient's muscle.

For this test, the same oscilloscope settings, components, and positioning proposed in the previous test (electrical conductivity) were used.

However, the objective was to verify the presence of visual noise in the waveform generated by the oscilloscope, performing a comparison with the image obtained from a previously approved electrode (Figure 12).

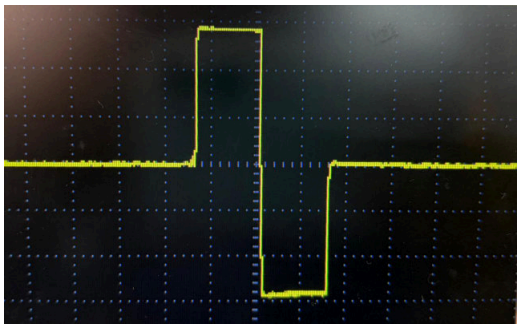


Fig. 12. Waveform without evident visual electromagnetic noise from a gold standard electrode. Source: Author (2025)

The analysis of the presence of electromagnetic noise was performed solely through visual comparison.

As can be observed in Figures 13 (smaller electrode) and 14 (larger electrode), the waveforms generated by the oscilloscope for both electrodes showed no evident electromagnetic noise upon visual inspection when compared with the waveform of the gold standard electrode.

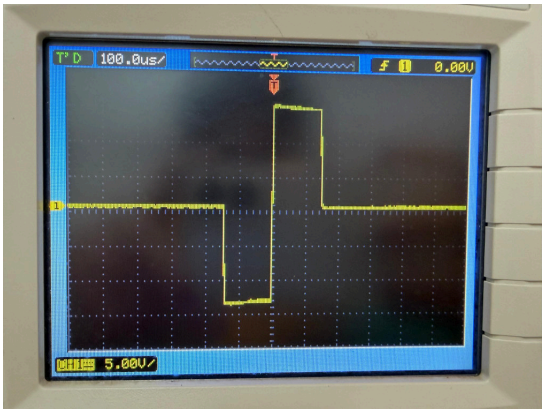


Fig. 13. Waveform without evident electromagnetic noise upon visual inspection from the smaller electrode. Source: Author (2025)

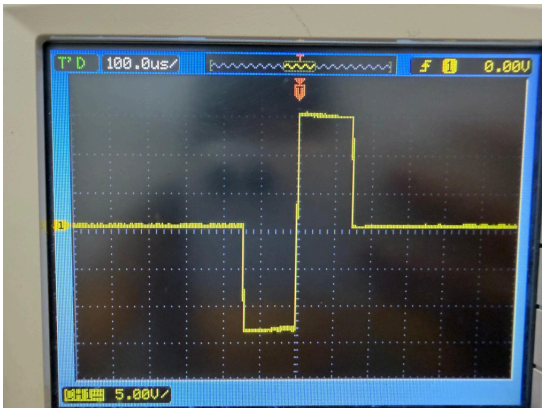


Fig. 14. Waveform without evident electromagnetic noise upon visual inspection from the larger electrode. Source: Author (2026)

C. Quantitative Noise and RMS Analysis

In order to provide a quantitative evaluation of the signal characteristics beyond visual waveform inspection, the Root Mean Square voltage (V_{RMS}) was selected as a key metric. This parameter accounts for the overall voltage magnitude, including baseline fluctuations, and provides a measure of the effective electrical energy delivered to the load under identical operating conditions.

To quantitatively evaluate signal magnitude, Root Mean Square voltage (V_{RMS}) values were extracted from the oscilloscope waveforms. The conventional reference electrode exhibited a V_{RMS} of 6.94 V, whereas the developed small and large electrodes yielded 6.87 V and 6.96 V, respectively, indicating comparable effective voltage levels across all tested geometries under identical experimental conditions.

The experimental setup utilized for the V_{RMS} quantitative measurement was identical to that previously described for the visual noise inspection test, maintaining the same oscilloscope settings, load resistor, probe connection, and component positioning.

Figures 15, 16, and 17 display the resulting oscilloscope trace captures for the conventional, small, and large electrodes,

respectively. In each waveform image, the measured V_{RMS} value is highlighted within a red outlined box for direct comparison.

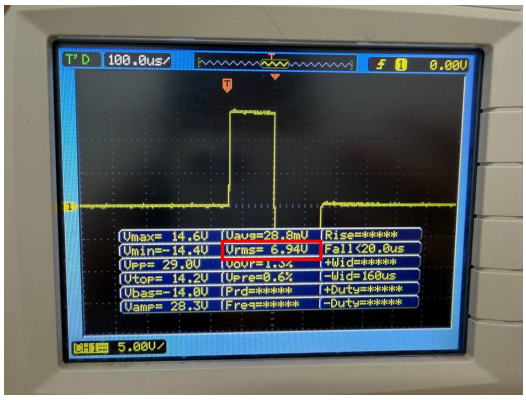


Fig. 15. Output voltage waveform from the conventional electrode. Source: Author (2026)

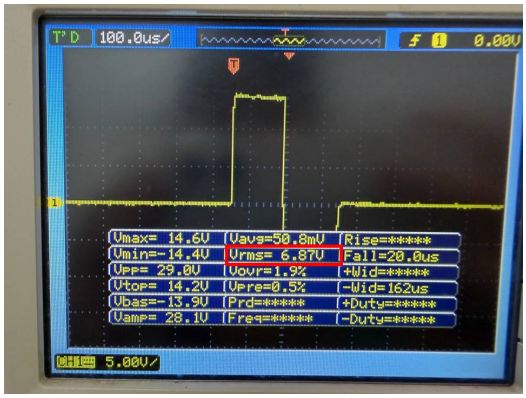


Fig. 16. Output voltage waveform from the smaller electrode. Source: Author (2026)

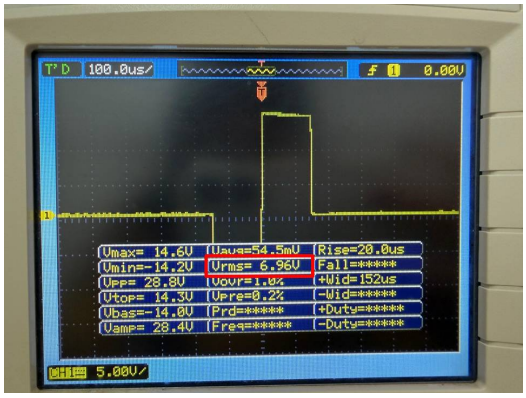


Fig. 17. Output voltage waveform from the larger electrode. Source: Author (2026)

D. Output-tracking agreement

The output-tracking agreement test ensures that, at different voltages, the electrode responds as expected. That is, for different parameters programmed in the electrical stimulator, the electrode must respond according to the applied settings and produce the corresponding electrical output.

For the experiment, the oscilloscope was configured in the same way as in both previous tests, and the positioning of the components for testing remained the same.

However, the only difference was the modification of the electrical stimulator parameter, in which all available voltage levels of the equipment used (modes 1 to 15) were applied in order to compare the results of each mode with those of a previously approved electrode and analyze whether the peak voltage value displayed on the oscilloscope was different for each mode.

The expected trend was that increasing the mode number would lead to an increase in the peak voltage measured by the oscilloscope.

All data obtained from the tests were recorded in spreadsheets using the Microsoft Excel spreadsheet software for subsequent analysis and comparison.

The voltage values obtained during the evaluation of the gold standard electrode are presented in Table II.

TABLE II
OUTPUT-TRACKING AGREEMENT TEST OF THE CONVENTIONAL ELECTRODE.

Mode:	**1**	**2**	**3**	**4**	**5**
Voltage (V):	1.2	2.2	3.2	4.2	5.2
Mode:	**6**	**7**	**8**	**9**	**10**
Voltage (V):	6.2	7.0	7.8	8.8	9.6
Mode:	**11**	**12**	**13**	**14**	**15**
Voltage (V):	10.6	11.8	12.6	13.6	14.6

The voltage values obtained from the smaller electrode are presented in Table III.

TABLE III
OUTPUT-TRACKING AGREEMENT TEST OF THE SMALLER ELECTRODE.

Mode:	**1**	**2**	**3**	**4**	**5**
Voltage (V):	1.2	2.2	3.2	4.2	5.2
Mode:	**6**	**7**	**8**	**9**	**10**
Voltage (V):	6.2	7.0	7.8	8.8	9.8
Mode:	**11**	**12**	**13**	**14**	**15**
Voltage (V):	10.6	11.8	12.6	13.6	14.6

The voltage values obtained from the larger electrode are presented in Table IV.

TABLE IV
OUTPUT-TRACKING AGREEMENT TEST OF THE LARGER ELECTRODE.

Mode:	**1**	**2**	**3**	**4**	**5**
Voltage (V):	1.0	2.2	3.0	4.0	5.0
Mode:	**6**	**7**	**8**	**9**	**10**
Voltage (V):	6.2	7.0	7.8	8.6	9.8
Mode:	**11**	**12**	**13**	**14**	**15**
Voltage (V):	10.8	11.8	12.6	13.6	14.8

The waveform images generated by the oscilloscope from the performed tests can be visualized in Figures 18 (gold

standard electrode), 19 (smaller electrode), and 20 (larger electrode).

The organization of the waveforms in each figure was based on the information order collected from the previously presented tables (Tables II, III, and IV).

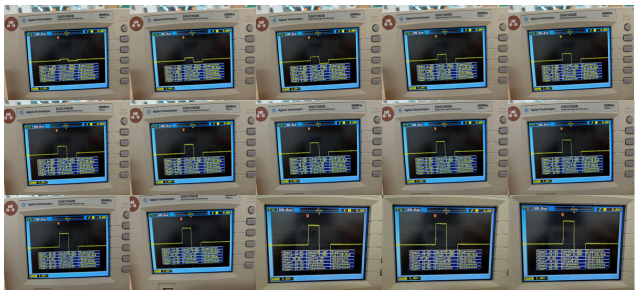


Fig. 18. Output-tracking agreement test waveforms from the gold standard electrode. Source: Author (2025)

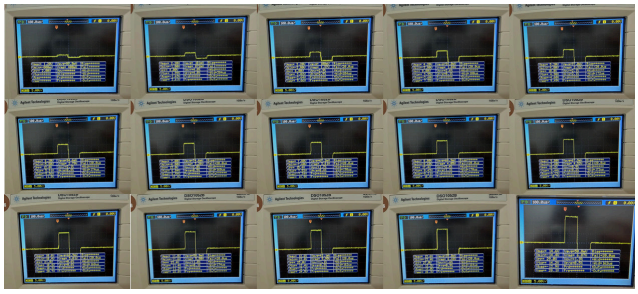


Fig. 19. Output-tracking agreement test waveforms from the smaller electrode. Source: Author (2025)



Fig. 20. Output-tracking agreement test waveforms from the larger electrode. Source: Author (2026)

It can be observed that the values recorded for the three analyzed electrodes showed only small variations, with differences of up to 200 mV. Therefore, it can be concluded that the tested electrodes exhibit output-tracking agreement comparable to those of a previously approved electrode.

E. Experimental Setup Specifications

To ensure test setup reproducibility, the bench evaluation was configured as a fixed resistive load circuit. Signal acquisition was performed using an Agilent Technologies DSO1072B Digital Storage Oscilloscope (70 MHz bandwidth, 1 GSa/s sampling rate) equipped with standard passive voltage probes.

Electrical stimulation was supplied by a portable Jumper TENS/EMS unit operating across 15 output intensity levels. The electrical pulses observed on the oscilloscope exhibited symmetrical or asymmetrical biphasic rectangular waveforms. The oscilloscope was configured with a horizontal time scale of 100.0 μs/div and a vertical voltage scale of 5 V/div for waveform acquisition.

The load circuit was constructed using axial carbon-film resistors with a $\pm 5\%$ tolerance (330 Ω and 510 Ω). The reference electrode consisted of a commercially available, pregelled, self-adhesive electrode. All physical contacts between the resistor leads and the electrode surfaces were maintained at fixed mechanical positions to minimize contact variability.

It is important to emphasize that this bench test setup evaluates voltage transfer exclusively within a simplified, fixed resistive circuit. The experimental design does not account for complex physiological and electrode–skin interface parameters, such as variations in skin impedance, delivered current density, electrode half-cell potential, or motion artifacts. Consequently, these measurements should be interpreted primarily as a preliminary electrical screening rather than as a comprehensive clinical validation.

F. Statistical Analysis

For the electrical conductivity tests, measurements were performed in quintuplicate for each electrode sample under the maximum output condition of the electrical stimulator. Three samples of each developed electrode geometry, as well as a previously approved commercial electrode, were evaluated. Subsequently, the data were analyzed using Microsoft Excel, and the mean values and standard deviations of the five measurements obtained for each test condition were calculated.

Additionally, two resistors with different resistance values, 330 Ω and 510 Ω, were used to evaluate the electrical performance of the electrodes under different load conditions. Each electrode was tested using both resistance values, following the same experimental setup and oscilloscope measurement procedure adopted in the electrical conductivity tests.

The voltage values obtained for the three types of electrodes—conventional, small, and large—were measured in quintuplicate. The results of the five measurements performed for each electrode type are presented in Tables V, VI and VII, respectively. The analysis of these results enabled the assessment of measurement consistency and the comparison of the electrical performance of the developed electrodes with that of the conventional electrode.

TABLE V
VOLTAGE MEASUREMENTS OF THE CONVENTIONAL ELECTRODE USING 330 Ω AND 510 Ω RESISTORS

Test	**330 Ω**	**510 Ω**
1	14.8	22.2
2	14.6	22.2
3	14.8	22.4
4	14.8	22.4
5	14.6	22.2

TABLE VI
VOLTAGE MEASUREMENTS OF SMALL ELECTRODES USING 330 Ω AND 510 Ω RESISTORS

Test	Small Electrode 1		Small Electrode 2		Small Electrode 3	
	330 Ω	510 Ω	330 Ω	510 Ω	330 Ω	510 Ω
1	14.6	22.4	14.8	22.4	14.6	22.2
2	14.8	22.2	14.8	22.4	14.8	22.4
3	14.6	22.4	14.8	22.2	14.6	22.2
4	14.6	22.2	14.6	22.4	14.6	22.2
5	14.8	22.2	14.6	22.2	14.6	22.4

TABLE VII
VOLTAGE MEASUREMENTS OF LARGE ELECTRODES USING 330 Ω AND 510 Ω RESISTORS

Test	Large Electrode 1		Large Electrode 2		Large Electrode 3	
	330 Ω	510 Ω	330 Ω	510 Ω	330 Ω	510 Ω
1	14.6	22.2	14.6	22.2	14.8	22.4
2	14.8	22.4	14.8	22.4	14.8	22.2
3	14.8	22.2	14.6	22.2	14.6	22.2
4	14.6	22.4	14.6	22.2	14.6	22.4
5	14.8	22.4	14.8	22.2	14.6	22.2

Tables VIII, IX, X, XI, XII and XIII present the results obtained using the electronic spreadsheet software Microsoft Excel, including the mean voltage values and their respective standard deviations calculated from the five measurements performed for each electrode.

TABLE VIII
MEAN ± STANDARD DEVIATION OF CONVENTIONAL ELECTRODES USING A 330 Ω RESISTOR

Electrode:	Conventional
Voltage (V):	14.72 ± 0.11

TABLE IX
MEAN ± STANDARD DEVIATION OF SMALL ELECTRODES USING A 330 Ω RESISTOR

Electrodes:	Small 1	Small 2	Small 3
Voltage (V):	14.68 ± 0.11	14.72 ± 0.11	14.68 ± 0.11

TABLE X
MEAN ± STANDARD DEVIATION OF LARGE ELECTRODES USING A 330 Ω RESISTOR

Electrodes:	Large 1	Large 2	Large 3
Voltage (V):	14.72 ± 0.11	14.68 ± 0.11	14.72 ± 0.11

TABLE XI
MEAN ± STANDARD DEVIATION OF CONVENTIONAL ELECTRODES USING A 510 Ω RESISTOR

Electrode:	Conventional
Voltage (V):	22.32 ± 0.11

TABLE XII
MEAN ± STANDARD DEVIATION OF SMALL ELECTRODES USING A 510 Ω RESISTOR

Electrodes:	Small 1	Small 2	Small 3
Voltage (V):	22.28 ± 0.11	22.32 ± 0.11	22.28 ± 0.11

TABLE XIII
MEAN ± STANDARD DEVIATION OF LARGE ELECTRODES USING A 510 Ω RESISTOR

Electrodes:	Large 1	Large 2	Large 3
Voltage (V):	22.32 ± 0.11	22.28 ± 0.11	22.28 ± 0.11

For the output-tracking agreement analysis, three independent measurements were performed for each operating mode (Modes 1–15) using the conventional reference electrode, the small electrode, and the large electrode. For each operating mode, the mean voltage (μ), standard deviation (σ), and percentage relative error (RE%) with respect to the conventional reference electrode were calculated.

The relative error values were calculated using the original mean values before rounding.

The percentage relative error was calculated using the mean voltage of the conventional reference electrode as the reference value, according to:

$$\mathrm{RE}(\%) = \frac{\left|\mu_{\mathrm{electrode}} - \mu_{\mathrm{conventional}}\right|}{\mu_{\mathrm{conventional}}} \times 100 \qquad (1)$$

Where $\mu_{\mathrm{electrode}}$ represents the mean voltage measured for the small or large electrode, and $\mu_{\mathrm{conventional}}$ represents the mean voltage measured for the conventional reference electrode.

Table XIV presents the mean (μ), standard deviation (σ), and percentage relative error ($\mathrm{RE}\%$) for each operating mode.

TABLE XIV
STATISTICAL ANALYSIS OF OUTPUT VOLTAGE ACROSS 15 OPERATING MODES (MEAN ± STANDARD DEVIATION AND RELATIVE ERROR)

Mode	Conventional (V)	Small (V)	Error (%)	Large (V)	Error (%)
1	1.13 ± 0.12	1.13 ± 0.12	0.00%	1.13 ± 0.12	0.00%
2	2.20 ± 0.00	2.20 ± 0.00	0.00%	2.20 ± 0.00	0.00%
3	3.07 ± 0.12	3.13 ± 0.12	2.17%	3.07 ± 0.12	0.00%
4	4.13 ± 0.12	4.07 ± 0.12	1.61%	4.07 ± 0.12	1.61%
5	5.13 ± 0.12	5.13 ± 0.12	0.00%	5.13 ± 0.12	0.00%
6	6.20 ± 0.00	6.20 ± 0.00	0.00%	6.20 ± 0.00	0.00%
7	7.00 ± 0.00	7.00 ± 0.00	0.00%	7.00 ± 0.00	0.00%
8	7.80 ± 0.00	7.80 ± 0.00	0.00%	7.80 ± 0.00	0.00%
9	8.73 ± 0.12	8.67 ± 0.12	0.76%	8.60 ± 0.00	1.53%
10	9.67 ± 0.12	9.73 ± 0.12	0.69%	9.73 ± 0.12	0.69%
11	10.67 ± 0.12	10.67 ± 0.12	0.00%	10.73 ± 0.12	0.63%
12	11.80 ± 0.00	11.80 ± 0.00	0.00%	11.80 ± 0.00	0.00%
13	12.60 ± 0.00	12.60 ± 0.00	0.00%	12.60 ± 0.00	0.00%
14	13.60 ± 0.00	13.60 ± 0.00	0.00%	13.60 ± 0.00	0.00%
15	14.73 ± 0.12	14.73 ± 0.12	0.00%	14.73 ± 0.12	0.00%

The results demonstrated high measurement consistency for both developed electrode geometries, with a maximum standard deviation of 0.12 V. The relative error was 0.00%

for most operating modes, with maximum deviations of 2.17% for the small electrode in Mode 3 and 1.61% for the large electrode in Mode 4. These results indicate comparable output-voltage agreement between the developed and conventional electrodes under the evaluated conditions.

Based on these findings, the proposed procedure provided a preliminary assessment of the electrical performance of the developed electrodes. The electrodes demonstrated electrical behavior comparable to the conventional reference under the evaluated conditions. However, further testing with a larger number of devices is required to assess their safety, reliability, and applicability to electromyography and electrical stimulation.

IV. Conclusion

The development of the experimental procedure for the evaluation and validation of electrodes for human muscle electromyography and electrical stimulation enabled a preliminary assessment of the electrical performance of these devices.

The electrical conductivity test enabled the verification of the electrode's ability to conduct and transmit electrical signals from the electrical stimulator through the electrode.

The visual inspection for electromagnetic noise made it possible to verify that no evident undesirable noise was observed during the experimental conditions evaluated.

The output-tracking agreement experiment enabled the preliminary assessment of the electrode under different electrical stimulator settings, allowing its electrical response to be compared with that of the reference electrode.

The objective of this project was to develop a document compiling the essential tests required for the evaluation of electrodes used in human muscle electromyography and electrical stimulation.

Although the results demonstrate preliminary electrical functioning and the ability of the electrodes to conduct electrical signals, they are not sufficient to confirm that the devices are suitable or fully validated for electromyography or electrical stimulation applications. Furthermore, the experiments performed do not allow conclusions regarding patient safety, clinical efficacy, or suitability for use in humans. Following the completion of the test specification stage, future work will involve conducting tests on a larger number of devices and performing further evaluations to strengthen the validation of the proposed procedure for large-scale application. This may contribute to the future development of safer and more reliable electrodes for muscle assessment and electrical stimulation.

V. Future Work

Future studies will focus on expanding the experimental characterization of the developed electrodes by investigating their behavior under a broader range of operating and application conditions. Additional analyses will be performed to further investigate signal integrity, electrical stability, and the interaction between the electrodes and the application environment.

The study will also be extended to a larger number of devices to obtain a more representative assessment of the reproducibility of the results. Further investigations involving electromyographic signal acquisition and electrical stimulation will be conducted, together with evaluations related to device safety and performance. These additional studies will provide a more comprehensive basis for further evaluating the potential applicability of the developed electrodes in electromyography and electrical stimulation.

Acknowledgment

This work was supported by the Fundação Instituto Nacional de Telecomunicações (FINATEL), the Instituto Nacional de Telecomunicações (Inatel), and the eHealth - Innovation Center Laboratory, which diligently provided the resources necessary for the development of this project.

To Professor Filipe Bueno Vilela, Ph.D., for providing essential research contributions for the development of the artifact.

To Professor Luma Rissati Borges do Prado, for indicating the electrode testing method using the oscilloscope.

To Professor Elisa Rennó Carneiro Déster, M.Sc., for her valuable guidance and teachings in anatomy, physical therapy, and electrode application.

To the advisor, Professor Maysa Costa Alves, Ph.D., for all the support, ideas, and knowledge shared.

References

[1] D. A. Winter, "Electromyography in Biomechanics," in *Biomechanics and Motor Control of Human Movement*, 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2009, ch. 8, pp. 191–212.

[2] A. Fletcher, "Action Potential: Generation and Propagation," Anaesth. Intensive Care Med., vol. 23, no. 3, pp. 183–187, Mar. 2022, doi: 10.1016/j.mpaic.2021.11.014.

[3] P. E. Barkhaus, S. D. Nandedkar, M. de Carvalho, M. Swash, and E. V. Stålberg, "Revisiting the compound muscle action potential (CMAP)," *Clin. Neurophysiol. Pract.*, vol. 9, pp. 176–200, 2024.

[4] *Medical electrical equipment – Part 1: General requirements for basic safety and essential performance*, IEC Standard 60601-1:2005+AMD1:2012+AMD2:2020 CSV, 2020.

[5] J. G. Webster and A. J. Nimunkar, Eds., *Medical Instrumentation: Application and Design*, 5th ed. Hoboken, NJ, USA: Wiley, 2020.

[6] E. du Bois-Reymond, *Untersuchungen über thierische Elektricität*, vol. 1. Berlin: G. Reimer, 1848.

[7] V. Alcan and M. Zinnuroğlu, "Current developments in surface electromyography," *Turk. J. Med. Sci.*, vol. 53, no. 5, pp. 1019–1031, Oct. 2023.

[8] R. Gutiérrez-Arias, Y. Jalil, R. Fuentes-Aspe, and P. Seron, "Effectiveness of neuromuscular electrostimulation in COPD subjects on mechanical ventilation: A systematic review and meta-analysis," *Clinics*, vol. 77, Art. no. 100108, 2022.

[9] Z. Zuo, Y. Wang, Y. Fang, Z. Wang, Z. Yang, B. Jia, and Y. Sun, "Electrostimulation: A promising new treatment for psoriasis," *Int. J. Mol. Sci.*, vol. 25, no. 23, Art. no. 13005, Dec. 2024.

[10] Y. Bezsmertnyi, V. Shevchuk, Y. Jiang, and O. Shevchuk, "Bone healing during electrostimulation," *Med. Glas. (Zenica)*, vol. 20, no. 2, pp. 255–262, Aug. 2023.

[11] L. Omana, P. A. Chandran, S. M. Simon, B. P. J. de Souza, S. Thomas, and N. Kalarikkal, "Recent advances in polymer nanocomposites for electromagnetic interference shielding: A review," *ACS Omega*, vol. 7, no. 30, pp. 25921–25947, Jul. 2022.

Multimodal Physiological Measurement Method Validation in a Flight Simulator: EDA/GSR, ECG, HRV and Pupillometry

Leandro Baran
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID 0009-0005-5932-6632

Pedro Lucas dos Reis Silva
Undergraduate Student in Computer Engineering
Federal University of Technology – Paraná (UTFPR)
Apucarana, Brazil
ORCID 0009-0001-9985-510X

Azir Regina Barreto Pires
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID 0009-0003-0186-4783

Amauri Amorin Assef
Department of Electronics Engineering
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID 0000-0002-2001-5387

Daniel Prado de Campos
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná (UTFPR)
Apucarana, Brazil
ORCID 0000-0001-6233-6077

Cristhiane Gonçalves
Postgraduate Program in Biomedical Engineering
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID 0000-0002-3894-2971

***Abstract*—Physiological monitoring for aviation human factors research is imperative due to human errors being the most prominent influences of aviation incidents and accidents. This paper presents the validation of a multimodal acquisition setup comprising a Daheng camera with Sony CMOS for pupillometry and a BITalino (r)evolution Board Kit with electrodes for electrodermal activity (EDA),), a.k.a. galvanic skin response (GSR), and electrocardiography (ECG) with heart rate variability (HRV) derivation, deployed in a flight simulator of a 737-800 during manual approach and touch-and-go landing (TGL) sequences. We enrolled 8 pilots (mean age 26.4, SD$\pm$7.7) and assessed pre/post-event changes comparing physiological responses across flight phases; measuring reliability using intraclass correlation coefficients (ICC), Bland–Altman analysis, standard error of measurement (SEM), and minimal detectable change (MDC); and Spearman correlations against Likert subjective self–evaluation ratings. ICC values were considered poor between the first pair of TGLs for each pilot, but large effect-sizes ($\eta_p^2 > 0.14$) were observed for skin conductance level (SCL), root mean square of successive RR-interval differences (RMSSD), and pupil/iris ratio. These results support the feasibility and physiological sensitivity of the setup, but test–retest reliability was not established for single-landing interpretation.**

***Index Terms*—pupillometry, EDA, GSR, ECG, HRV, flight simulator, method validation, aviation human factors, biosignals.**

Sponsors: EPA Training Center, www.epa.aero (Curitiba, Brazil) and UTFPR, Universidade Tecnológica Federal do Paraná.

I. Introduction

Considering the need for capturing pilots' autonomic and cognitive state changes during different phases of flight—particularly approach and landing, the most critical in terms of safety in conjunction with human factors [1], [2]—affordable and reliable means of measuring biosignals are required for research spreading. There are gaps for improvement, mainly in general aviation, where higher accident numbers also have prominent causal and contributing factors as human errors [3], [4]. Cognitive workload, attentional allocation, and autonomic arousal fluctuate markedly within intervals linked to each flight phase. Specifically, Instrument Landing System (ILS) final approach can be considered the most attentionally demanding flight phase [5].

Physiological measurement offers continuous, non-intrusive proxies for biosignals during pilots' states. Pupil diameter, modulated by the locus coeruleus-norepinephrine system, is a sensitive indicator of phasic arousal and mental effort [6]. Electrodermal activity (EDA), a.k.a. galvanic skin response (GSR), reflects sympathetic nervous system activation and is widely employed in stress and engagement/workload research [7]. Electrocardiography (ECG) and its derived heart rate variability (HRV) metrics—particularly Standard Deviation of the Normal-to-Normal Intervals (SDNN), Root Mean Square of Successive Differences (RMSSD), Low Frequency/High Frequency (LF/HF) ratio, and percentage of NN

intervals that differ by more than 50 milliseconds (pNN50)—provide indices of autonomic balance and have demonstrated sensitivity to pilot workload [8].

Considering the proliferation of lower-cost wearable biosignal platforms [9], the validation of their psychometric properties in simulation environments is welcome for literature representation and research use. Flight simulators may introduce electromagnetic interference, motion artifacts from control inputs and pilots' movement, lighting variability from screens and projectors, and temporal complexity from discrete task phases—all of which can interfere with signal integrity and measurement consistency.

This study helps address this gap by systematically evaluating the reliability and validity of a BITalino (r)evolution board (PLUX Biosignals, Portugal) for ECG and EDA/GSR acquisition, and a Daheng Venus vision camera for pupillometry, deployed in a Boeing 737NG Flight Training Device during manual touch-and-go landings (TGLs), mainly with student pilots.

The objectives are: (1) evaluate the reliability of pupillometric, EDA, ECG, and HRV measures across consecutive TGL trials within the same simulation session; (2) assess construct validity by comparing physiological responses between the approach phase and landing phase; (3) determine concurrent validity through correlation with subjective workload via Likert scale survey [10]; and (4) report signal quality metrics with artifact rates for each modality in the simulator environment.

II. Theoretical Background

Physiological validation with a standardized assessment protocol, particularly for wearables, ideally should be addressed in three levels: by signal, by parameter, and by event [11].

A. Pupillometry in Cognitive Research

Pupil diameter has been established as a sensitive index of phasic arousal and cognitive effort, driven by norepinephrine release from the locus coeruleus. Significant pupil dilation during higher workload contexts is expressively reported [12]. Measurement reliability across repeated trials with a fixed remote, non-corneal-reflection camera can also be formally validated. Pupil size is susceptible to ambient luminance changes in the cockpit, corneal reflection artifacts, and partial blink occlusion, which require robust preprocessing.

B. Electrodermal Activity (EDA) During Cognitive Tension

EDA (or GSR) decomposition into tonic (skin conductance level, SCL) and phasic (skin conductance response, SCR) components has improved discriminability in dynamic, movement-rich environments. Studies employing EDA sensors have demonstrated satisfactory results for pilot workload analysis during real flight tests [13], despite various limitations due to equipment and signals [14].

C. ECG and HRV During Cognitive Tension

A 2024 systematic review identified HRV as the most prevalent physiological modality in pilot mental workload studies (reported in 29 reviewed articles), with RMSSD, SDNN, and LF/HF ratio as the most consistently reported metrics [15]. Reliability of short-segment HRV is critically dependent on ectopic beat removal and artifact preprocessing. BITalino sensors have been evaluated against reference devices and are considered reliable tools for signal acquisition [16], supporting open-source processing.

III. Materials and Methods

A. Participants

Eight pilots (seven males, mean age 26.4, SD$\pm$7.7) were recruited as volunteers. Seven were students or low-experienced pilots (less than 300 flight-hours) and one commercial pilot with more than 1000 flight hours. Inclusion criteria: valid Brazilian medical certificate [17] (one private pilot's certificate was expired at the date of the test), at least 18 years of age, and being able to maintain attitude-power coordination and safely perform takeoffs and landings in a real airplane. All pilots had no familiarity with Boeing 737-800 procedures and handling. Exclusion criteria: no practical flying time or no medical certificate ever issued. All participants provided written informed consent. The study was approved by the UTFPR Ethics Committee [18], Protocol CEP_8289986.

B. Equipment and Data Acquisition

The flight simulator is a static Cabin Procedures Trainer of a Boeing 737-800X (New Generation), custom made with X-Plane 11 and FS-FlightControl 1.7.3 for external control. Post-flight data in CSV (comma—separated values) format were extracted, with timestamps of flight phases automatically recorded at approximately 1 Hz in the simulation management station. Time information was synchronized with an external notebook via Windows Network Time Protocol (NTP), with a tolerance below ± 1 second observed before every flight. An internal IP Dome camera captured the flight deck view (with audio) for every flight.

For ECG and EDA, a BITalino (r)Evolution Plugged Kit BLE/BT board (2021 version) with an ECG 3-lead electrode cable and an EDA 2-lead electrode cable was used, with Medix SF22 disposable electrodes. Performing the Einthoven Lead [19], the positive electrode was placed on the participant's left collarbone, the negative on the right collarbone, and the reference on the right side of the iliac crest. EDA electrodes were placed on the medial phalanges of the index and middle fingers of the non-dominant hand. Figure 1 contains the sensors/electrodes position indication. The board (battery supplied) was connected via Bluetooth to the external notebook using OpenSignals v2.2.7 at 100 Hz.

The camera for pupillometry is a Daheng Imaging VEN-042-121U3M/C monochrome camera (Sony IMX297 CMOS) with HN-P-1228-6M-C1/1.8 lens, mounted on the left side of the instrument panel at eye level, approximately 60 cm from the participant's face, connected via USB 3.0 and recorded using GalaxyViewer 3.2 (H.264 MP4, up to 99 fps). A custom illumination plate with 3$\times$ Epistar InfraRed LEDs (850 nm,

3 W, 120°), USB-powered, was front-mounted at eye level, 50 cm from the participant's face.

Camera and BITalino softwares live recorded signals in a Samsung SB20-FE notebook, Intel Core i5 10210U, 12 GB RAM, Windows 10 22H2. Absolute timestamps from synchronized clocks were used for multimodal alignment, as no shared hardware trigger was available (Fig. 1).

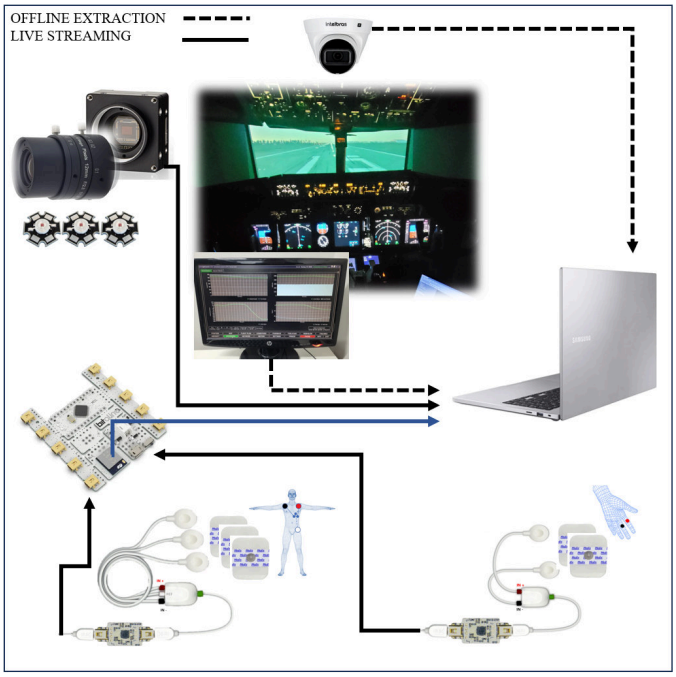


Fig. 1. Data transfer schematic diagram.

C. Experimental Protocol

Following the initial briefing and cabin familiarization, each participant performed consecutive manual TGL sequences (median 5, range 2–6) and a final landing at Curitiba airport, with Instrument Landing System (ILS) of Runway 15 under Visual Flight Rules (VFR), Visual Meteorological Conditions (VMC), and no wind. Scenario order or circuit was the same for all participants (departure, crosswind leg, downwind leg, base leg, final approach). Workload was manipulated and leveled by instructor inputs, considering the number and intensity of deviations/errors (altitude, speed, bank angles, etc) observed for each pilot performance. The final dataset comprised 37 verified TGL events across eight flights.

After the simulation session, participants completed a Likert scale form (based on NASA-TLX [20]) focusing on the hardest approach maneuver in retrospect, rating 10 questions across three dimensions: Cognitive Demand (CD), Perceived Performance (PP), and Emotional Response (ER), on a 0–5 scale. For CD, 2 questions about cognitive effort; for PP, 4 about self-performance and aircraft control; and for ER, 4 about perceived stress and fear. The overall "stimulus" score (0–100%) was computed expecting stronger biosignal responses for higher CD and ER grades, and smaller responses for PP.

D. Signal Preprocessing

1) ECG and HRV: ECG signal preprocessing was calibrated to millivolts, cleaned with Hampel-based artifact repair, detrended, and bandpass-filtered (0.5–40 Hz). R-peaks were detected from a differentiated, squared, and integrated ECG representation and refined on the filtered waveform. RR intervals outside physiological limits or deviating more than 20% from the local median were corrected by interpolation. Heart rate (HR) and RMSSD were estimated in 30-second moving windows, yielding second-by-second indices of chronotropic activation and short-term HRV.

2) EDA: EDA signal preprocessing was calibrated to microsiemens (μS) using the manufacturer transfer function, screened for rail saturation, and low-pass filtered at 2.5 Hz. The tonic component (SCL) was estimated with a 0.05 Hz low-pass filter; the phasic component was obtained as the residual and summarized by its root mean square. Event windows with excessive saturation were excluded from electrodermal inference.

3) Pupillometry: Videos preprocessing was done offline using a custom Python pipeline. The algorithm detects face and eye landmarks, selects the eye region with the best pupil evidence, and tracks this region over time with template-based updates and automatic re-detection when tracking confidence decreases. Inside the eye region of interest, contrast enhancement, Gaussian filtering, inverted thresholding, morphological filtering, and geometric validation are applied to segment the pupil. Frames affected by blinking, poor tracking, abrupt temporal changes, or weak pupil evidence are rejected. The pupil radius is normalized by the estimated iris radius, producing a relative pupil-to-iris percentage signal at 1 Hz. Event annotations were analyzed with pre-, during-, and post-event windows, generating per-event quality metrics, normalized pupil variation, and visual audit frames for methodological validation.

E. Event Windows, Quality Criteria, and Interpretation

Signals were resampled onto a common 1 Hz multimodal timeline. Each verified TGL was analyzed with a pre-event baseline window from -30 to -5 s and a post-event response window from 0 to $+30$ s. At least 20 valid baseline samples and 24 valid post-event samples were required. Valid event responses were averaged within each flight before inferential testing, ensuring that the unit of analysis was the flight rather than the individual landing.

Signal-quality limitations are noted explicitly as they materially affect interpretation. The RR correction fraction ranged from 25.4% to 56.4% across flights, indicating substantial ECG artifact burden in the simulator environment. Electrodermal saturation excluded two flights from EDA inference, whereas 1 Hz pupil availability remained comparatively high (86.1%–100.0%).

F. Statistical Analysis

The primary comparison was the pre/post-event change within each flight, expressed in native units and as normalized

percentage change ($\Delta\% = 100 \times (\text{post} - \text{pre})/\text{pre}$). Phase effects were summarized with paired tests, confidence intervals, and effect sizes, with false-discovery-rate (FDR) adjustment across the five primary metrics. TGL01 and TGL02 denote the first two valid events within each flight, selected a priori for test-retest reliability analysis.

Reliability: ICC(2,1) was computed [21], translated into Standard Error of Measurement (SEM) and Minimum Detectable Change at the 95% confidence level (MDC_{95}). SEM $= \text{SD} \times \sqrt{1 - \text{ICC}}$ and $\text{MDC}_{95} = 1.96 \times \text{SEM} \times \sqrt{2}$.

Bland–Altman analysis compared each flight-level post-event mean with its own pre-event baseline mean and summarized agreement through bias and 95% limits of agreement (LoA).

Likert scores were linearly transformed from the original 1-5 range to a 0-100 scale, averaged within each domain, and then averaged across domains to obtain an overall stimulus score. Because these scores were derived from ordinal responses and the sample size was small, questionnaire-physiology associations were tested at the item level using exact two-sided Spearman rank correlations with the mean pre/post physiological percentage change for each metric. Kendall tau-b was calculated as a sensitivity analysis for tied ranks, and false-discovery-rate correction (pFDR) was applied across the full family of questionnaire-physiology comparisons.

IV. Results

A. Physiological Effects: Pre/Post-TGL Comparison

The clearest group-level tendency was observed in the tonic electrodermal component. SCL increased by 0.91 μS on average, compatible with enhanced sympathetic arousal during landing-related perturbation. This tendency reached nominal significance ($p = 0.026$) but not FDR correction ($p_{\text{FDR}} = 0.129$), and should be interpreted as the strongest signal trend rather than as definitive isolated proof. HR increased modestly (0.53 bpm) and RMSSD tended to decrease (-3.07 ms), compatible with reduced vagal modulation under increased task demand. Phasic EDA changed very little at group level (< 0.001 μS), whereas the pupil-to-iris ratio increased slightly (0.94 percentage points), again in a physiologically plausible direction but with greater heterogeneity. Full results are shown in Table I.

TABLE I
Level of Pre/Post TGL Effects

Metric	n	Δ [95% CI]	F	p_{FDR}	η_p^2
Heart rate (bpm)	8	0.53 [−5.47, 6.53]	0.04	0.991	0.01
RMSSD (ms)	8	−3.07 [−7.41, 1.27]	2.80	0.346	0.29
SCL (μS)	6	0.91 [0.16, 1.65]	9.83	0.129	0.66
Phasic EDA (μS)	6	−0.00 [−0.12, 0.12]	0.00	0.991	0.00
Pupil/iris (%)	8	0.94 [−1.15, 3.03]	1.13	0.537	0.14

Pre: HR 98.79±9.17, RMSSD 59.60±6.51, SCL 15.84±4.25, Phasic EDA 0.34±0.17, Pupil 38.04±5.66. FDR adjusted.

The normalized pre/post distributions showed non-uniform responses across flights. The broadest interquartile spread was observed for Phasic EDA Root Mean Square/RMS (≈49.5 percentage points), indicating that landing-related physiology depends on pilot-specific regulation, procedural adaptation across repetitions, and measurement conditions (Fig. 2).

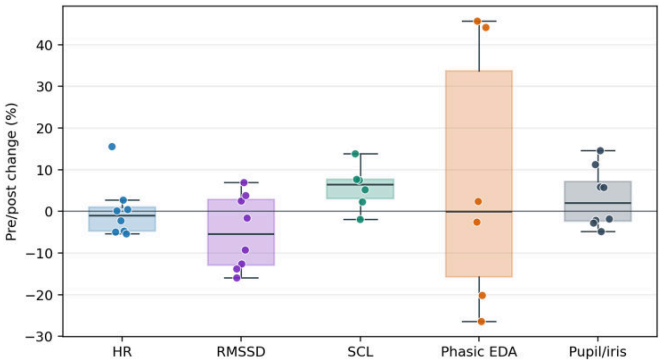


Fig. 2. Boxplots of percentage level change (pre to post-event) for each primary metric, illustrating interindividual variability.

B. Reliability Results

Reliability was weak when the first two valid events were treated as repeated measures. The highest ICC(2,1) was -0.01 for pupil-to-iris ratio, which falls in the poor range (< 0.40) [21]. Bland–Altman plots show broad event-to-event disagreement for several metrics, especially HR and RMSSD, whereas SCL displayed comparatively narrower limits (Fig. 3). These results do not negate physiological sensitivity, but indicate that the proposed setup should not be considered a reliable assessment of individual landing events, instead of a group-level analysis. Full reliability metrics are presented in Tables II and III.

TABLE II
First 2 TGLs Reliability Results – ICC(2,1), SEM, MDC_{95}

Metric	n	ICC(2,1) [95% CI]	SEM	MDC_{95}
Heart rate (bpm)	7	−0.08 [−0.77, 0.24]	12.72	35.25
RMSSD (ms)	7	−0.16 [−0.63, 0.15]	19.24	53.32
SCL (μS)	5	−0.11 [−1.16, 0.57]	1.20	3.32
Phasic EDA (μS)	5	−0.04 [−0.65, 0.47]	0.10	0.29
Pupil/iris (%)	7	−0.01 [−0.53, 0.49]	4.82	13.35

95% CI via bootstrap.

C. Likert Scale Correlations

Post-flight ratings indicated a moderate-to-high perceived challenge, with an overall stimulus score of 56.0 ± 16.3 on the 0-100 scale. Cognitive demand showed the highest average score (70.3), whereas emotional response was more heterogeneous across flights. No questionnaire-physiology association survived false-discovery-rate correction, but the most coherent nominal pattern involved RMSSD. Higher overall stimulus scores were associated with more negative RMSSD changes

TABLE III
BASELINE-REFERENCED BLAND–ALTMAN AGREEMENT

Metric	n	Bias	LoA low	LoA high
Heart rate (bpm)	8	0.53	−13.53	14.59
RMSSD (ms)	8	−3.07	−13.24	7.10
SCL (μS)	6	0.91	−0.48	2.30
Phasic EDA (μS)	6	−0.00	−0.23	0.23
Pupil/iris (%)	8	0.94	−3.95	5.83

LoA = bias $\pm$ 1.96 $\times$ SD_{diff}.

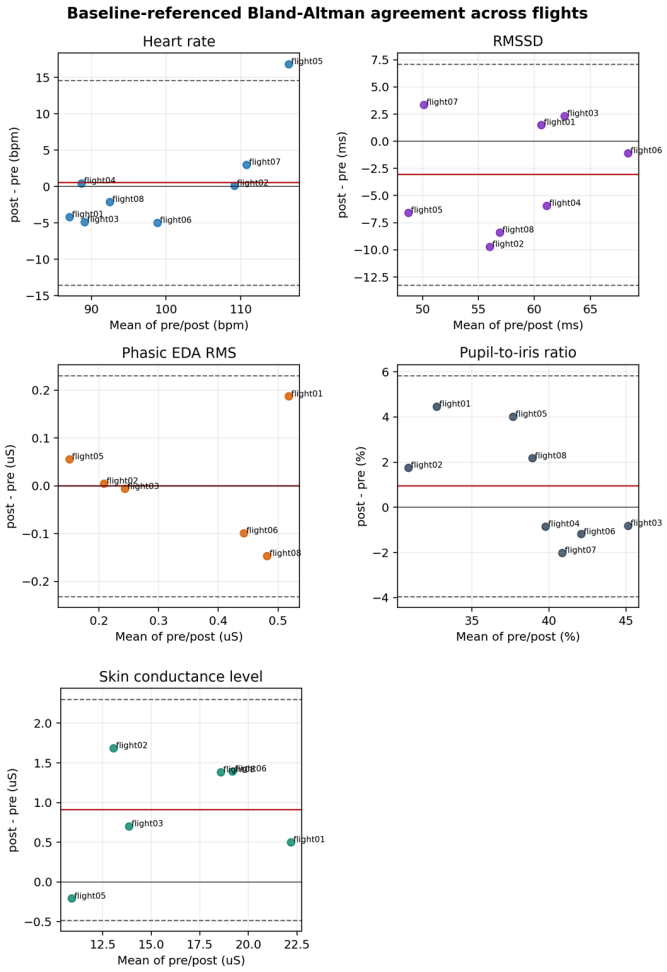


Fig. 3. Baseline-referenced Bland-Altman agreement for the primary physiological outcomes.

(rho=-0.81, exact p=0.022, pFDR=0.278; tau-b=-0.64), with similar trends for perceived performance difficulty (rho=-0.78) and cognitive demand (rho=-0.69). Electrodermal trends were directionally plausible but imprecise, with the strongest nominal EDA association observed between emotional response and SCL (rho = 0.75, n = 6). Correlations with physiological metrics are presented in Table IV. The strongest nominal trend for each flight, with RMSSD versus stimulus score, are presented in Fig. 4.

TABLE IV
SPEARMAN CORRELATIONS BETWEEN METRICS AND LIKERT FORMS

Metric	n	ρ	p-value	p_{FDR}	Kendall τ-B
Heart rate (bpm)	8	0.38	0.360	0.654	0.29
RMSSD (ms)	8	−0.81	0.022	0.278	−0.64
SCL (μS)	6	0.54	0.297	0.654	0.33
Phasic EDA (μS)	6	−0.26	0.658	0.823	−0.20
Pupil/iris (%)	8	0.14	0.752	0.836	0.00

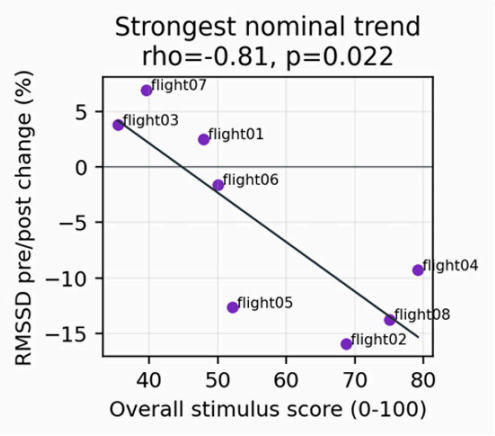


Fig. 4. Baseline-referenced scatter plot for RMSSD and Likert score.

V. DISCUSSION

A. General Validation

Quality control is still required before claiming standalone validation. The clearest support for the study hypotheses comes from the combination of directional plausibility and multimodal consistency rather than from a single overwhelmingly significant marker. Tonic EDA behaved most consistently with the expected autonomic activation, whereas HR, RMSSD, and pupil responses added complementary but more variable information. This pattern is physiologically credible for a simulator setting in which the landing task is repeated, partially anticipated, and not identical across trials. Similar physiological responses are expected to be observed in full-motion simulators or real flights in more challenging situations for experienced pilots.

The large between-flight spread is also scientifically informative: physiological response amplitude depends not only on the characteristics of each final approach, but also on pilot-specific regulation, strategy, familiarity across repetitions, and measurement conditions. The present findings are therefore better framed as evidence that the instrumentation can reveal meaningful physiological modulation than as proof of a single universal physiological signature.

The Likert scale findings suggest partial convergent validity rather than definitive concurrent validation. Flights perceived as more demanding tended to exhibit larger reductions in RMSSD, which is physiologically consistent with reduced vagal modulation under higher workload or stress. However, the estimates remained imprecise after correction for multiple

testing, and the retrospective questionnaire referred to the most difficult flight approach only. This scale mismatch probably attenuated the observed associations and should be resolved in future acquisitions by pairing each subjective rating with a single objectively defined landing event.

B. Instrumentation and Signal Quality

The BITalino platform demonstrated acceptable signal quality in the flight deck environment, without perceivable Bluetooth connection dropouts. Signal instabilities occurred during TGLs, attributed to electrode cable movement during pilots' motion when seeking the runway and alternating focus between panel and external view. Higher hand hidrosis and subsequent electrode detachment or signal saturation were observed during yoke input motion. Higher artifact rates during final approach and landing phases should be considered in future arrangements.

Pupillometry presented a challenge due to head movements around flight deck and the static camera frame limitation. Despite that, no illumination losses were observed due to stable ambient light. The absence of active gaze tracking meant that pupil size estimates could be confounded—a limitation addressable by integrating a wearable eye-tracking system (e.g., Tobii Glasses) in future work.

C. Limitations

The main limitations arise from data quality and synchronization. High RR correction fractions reduce confidence in short-window HRV [22], and electrodermal saturation in two flights shows that sensor range remains a practical bottleneck. Clock-based synchronization was sufficient for 30-second event windows, but would not support claims about sub-second onset latency.

Additional limitations include: *Sample size*—a larger, more experienced cohort would aid generalizability. *Single simulator model*—results may not generalize to other aircraft types or fidelity levels. *Retrospective Likert survey*—memory decay may affect ratings; in-sequence collection after each TGL is recommended. *No gold-standard comparison*—no clinical-grade ECG or laboratory-grade eye-tracker was used simultaneously. *Hand hidrosis*—EDA signal saturation or dropout from sweat accumulation on electrodes. *Pilot motion*—all biosignals are affected by cockpit movement, especially during runway sighting.

VI. Conclusion

This study provides the first systematic validation analysis of a combined BITalino ECG/EDA and fixed Daheng camera pupillometry setup in a commercial airliner flight simulator. The methodology established here—phase-aligned signal segmentation, Hampel-based ECG artifact correction, tonic/phasic EDA decomposition, normalized pupil-to-iris tracking, and ICC/Bland–Altman reliability analysis—provides a replicable framework for psychophysiological signals research in simulations, supporting the use of the described setup, with attention to the limitations identified, for within-pilot workload monitoring during approach and landing operations. The method cannot assure validation of operational fatigue or safety monitoring.

Future work should address: (1) replication in actual flight operations with lightweight, certified wearable devices; (2) integration of gaze tracking; (3) machine learning classification of workload states using the validated multimodal feature set; and (4) multi-session reliability across training events.

Acknowledgment

The authors thank EPA Training Center and UTFPR for the support. The data, screenshots, and equipment information used in this research belong to their respective companies and are used strictly for academic purposes. The authors declare no conflicts of interest.

References

[1] NTSB, "Case Analysis and Reporting Online (CAROL)," NTSB.gov, 2026. [Online]. Available: https://www.ntsb.gov/Pages/AviationQueryv2.aspx [Accessed: Jun. 23, 2026].

[2] D. Ison, "Tracking changes in aviation safety narratives: A text mining study of NTSB reports (2008–2025)," *Aerosp. Traffic Saf.*, vol. 2, no. 3–4, pp. 196–202, Dec. 2025, doi: 10.1016/j.aets.2025.12.006.

[3] E. Sheffield, S.-Y. Lee, and Y. Zhang, "A systematic review of general aviation accident factors, effects and prevention," *J. Air Transp. Manag.*, vol. 128, p. 102859, Sep. 2025, doi: 10.1016/j.jairtraman.2025.102859.

[4] E. A. Okine, E. Zarei, B. J. Roggow, and N. Dehghan, "Evolution of human factors research in aviation safety: A systematic review and bibliometric analysis," *J. Saf. Sci. Resil.*, vol. 7, no. 1, p. 100249, Mar. 2026, doi: 10.1016/j.jnlssr.2025.100249.

[5] V. Peysakhovich, W. Ledegang, M. Houben, and E. Groen, "Classification of flight phases based on pilots' visual scanning strategies," in *Proc. 2022 Symp. Eye Tracking Res. Appl.*, Seattle, WA, USA: ACM, Jun. 2022, pp. 1–7, doi: 10.1145/3517031.3529641.

[6] S. Mathôt, "Pupillometry: Psychology, physiology, and function," *J. Cogn.*, vol. 1, no. 1, p. 16, Feb. 2018, doi: 10.5334/joc.18.

[7] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, no. 2, p. 479, Jan. 2020, doi: 10.3390/s20020479.

[8] J. Koskelo et al., "Cardiac autonomic responses in relation to cognitive workload during simulated flight," *Appl. Ergon.*, vol. 121, p. 104370, Nov. 2024, doi: 10.1016/j.apergo.2024.104370.

[9] I. E. Stan, D. D'Auria, and P. Napoletano, "A systematic literature review of innovations, challenges, and future directions in telemonitoring and wearable health technologies," *IEEE J. Biomed. Health Inform.*, vol. 30, no. 3, pp. 2630–2645, Mar. 2026, doi: 10.1109/JBHI.2025.3598056.

[10] M. Koo and S.-W. Yang, "Likert-Type Scale," *Encyclopedia*, vol. 5, no. 1, p. 18, Feb. 2025, doi: 10.3390/encyclopedia5010018.

[11] H. G. Van Lier et al., "A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor," *Behav. Res. Methods*, vol. 52, no. 2, pp. 607–629, Apr. 2020, doi: 10.3758/s13428-019-01263-9.

[12] A. Lucchese, A. Padovano, and F. Facchini, "Comprehensive systematic literature review on cognitive workload: Trends on methods, technologies and case studies," *IET Collab. Intell. Manuf.*, vol. 7, no. 1, p. e70025, Jan. 2025, doi: 10.1049/cim2.70025.

[13] J. R. S. Scarpari et al., "Quantitative assessment of pilot-endured workloads during helicopter flying emergencies: An analysis of physiological parameters during an autorotation," *Sci. Rep.*, vol. 11, no. 1, p. 17734, Sep. 2021, doi: 10.1038/s41598-021-96773-y.

[14] G. Belloni and P. L. Curşeu, "Real-time cognitive state monitoring via physiological signals in commercial aviation: A systematic literature review with reasoned snowballing expansion," *Safety*, vol. 12, no. 2, p. 56, Apr. 2026, doi: 10.3390/safety12020056.

[15] Z. Lv, M. Li, Y. Zheng, and W. Feng, "Detecting and predicting pilot mental workload using heart rate variability: A systematic review," *Sensors*, vol. 24, no. 12, p. 3723, Jun. 2024, doi: 10.3390/s24123723.

[16] D. Batista, H. Silva, and A. Fred, "Experimental characterization and analysis of the BITalino platforms against a reference device," in *Proc. 39th Annu. Int. Conf. IEEE EMBC*, Seogwipo, Korea: IEEE, Jul. 2017, pp. 2418–2421, doi: 10.1109/EMBC.2017.8037344.

[17] ANAC, "RBAC67 – EMD05," ANAC.gov.br, 2021. [Online]. Available: https://www.anac.gov.br/assuntos/legislacao/legislacao-1/rbha-e-rbac/rbac/rbac-67 [Accessed: Jun. 23, 2026].

[18] UTFPR, UTFPR.edu.br, 2026. [Online]. Available: https://www.utfpr.edu.br/comissoes/permanentes/comite-de-etica-em-pesquisa [Accessed: Jun. 23, 2026].

[19] G. D. Gargiulo et al., "On the Einthoven triangle: A critical analysis of the single rotating dipole hypothesis," *Sensors*, vol. 18, no. 7, p. 2353, Jul. 2018, doi: 10.3390/s18072353.

[20] S. G. Hart and L. E. Staveland, "Development of NASA-TLX: Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, P. A. Hancock and N. Meshkati, Eds. Amsterdam: North-Holland, 1988, pp. 139–183.

[21] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2016.02.012.

[22] J. M. A. Alcantara et al., "Impact of using different levels of threshold-based artefact correction on the quantification of heart rate variability in three independent human cohorts," *J. Clin. Med.*, vol. 9, no. 2, p. 325, Jan. 2020, doi: 10.3390/jcm9020325.

Piezoelectric-Based Acquisition System for Acquisition of Mechanical Signals Associated with Simulated Mandibular Clenching and Grinding Movements

1st Francisco G. D. da Silva Filho
Department of Electrical Engineering
Federal University of Ceará
Fortaleza, Brazil
gdias27@alu.ufc.br

2nd Larissa M. Rodrigues
Department of Electrical Engineering
Federal University of Ceará
Fortaleza, Brazil
larissamendesr@gmail.com

3rd Ênio R. Viana
Department of Electrical Engineering
Federal University of Ceará
Fortaleza, Brazil
eniocc@gmail.com

4th Victor H. C. de Albuquerque
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
victor.albuquerque@lesc.ufc.br

***Abstract*—Polysomnography is the reference method for assessing sleep bruxism, but it is costly and requires specialized facilities. This paper presents the development and experimental characterization of a low-cost instrumentation system for acquiring mechanical signals associated with temporomandibular joint (TMJ) activity using a piezoelectric disk as the primary transducer. The system includes an analog signal-conditioning circuit and digital filtering. Signals were acquired from volunteers performing controlled clenching and grinding sequences. Time-domain analysis showed differences in the waveforms associated with the evaluated events, while spectral analysis revealed differences in their frequency distributions. The results demonstrate the system's ability to acquire mechanical signals associated with controlled mandibular events under laboratory conditions. The proposed system has potential for integration into a future portable device for the assessment and monitoring of sleep bruxism. Clinical validation and overnight recordings were beyond the scope of this study.**

***Index Terms*—Sleep bruxism, piezoelectric sensor, biomedical instrumentation, signal conditioning, temporomandibular joint.**

I. Introduction

Sleep bruxism is the involuntary grinding or clenching of teeth during sleep, a condition that affects a significant portion of the population [1]. Prevalence estimates vary, ranging from 22% to 30% in adults, but in clinical observations with children and adolescents, the numbers have frequently exceeded 40%, consistent with previous epidemiological reports [2]. Patients are often unaware of the condition, since the episodes occur during sleep without causing arousal [3]. In practice, the clinical burden is not limited to tooth wear. Temporomandibular joint pain, morning headaches, and masseter muscle hypertrophy are frequently observed in patients referred to clinical services, often as secondary findings during routine dental or neurological examinations [1].

The gold standard for the definitive diagnosis of sleep bruxism is polysomnography, an examination that monitors multiple physiological variables during sleep, including electromyography of the masseter muscles, electroencephalography, electro-oculography, airflow, and oxygen saturation [4]. However, access to this examination within the public healthcare system is severely restricted. Access to polysomnography within the Brazilian public healthcare system is limited: only 50 sleep centers are accredited by the Brazilian Unified Health System (SUS), with 10 states lacking any mapped diagnostic facilities for sleep disorders [5].

Even when accessible, polysomnography presents technical limitations. Surface electromyography is sensitive to movement artifacts, incorrect electrode placement, and electrical noise from the environment. Studies have shown that configurations relying exclusively on electromyographic signals, without the simultaneous support of audio and video channels, are considered unreliable for the accurate scoring of sleep bruxism events [6].

The limitations of polysomnography and surface EMG motivated the investigation of an alternative transduction principle. Previous studies have compared piezoelectric and surface electromyography sensing approaches for physiological and muscle activity monitoring, highlighting differences in characteristics such as response time, durability, and reusability [12].

This paper details the development of a piezoelectric-based instrumentation system designed to prioritize portability, electrical robustness, and low component cost for the acquisition of mechanical signals associated with mandibular clenching and grinding events characteristic of bruxism.

II. Background Literature

A. Sleep Bruxism

In clinical practice, sleep bruxism manifests in two distinct patterns of mandibular activity [7]. Clenching consists of isometric contraction of the masticatory muscles without dental sliding. Grinding, in turn, involves isotonic movements with lateral or protrusive mandibular sliding. During grinding episodes, mandibular movements often exceed the canine edge-to-edge position, accompanied by elevated masseter muscle activity [8]. These movements account for the characteristic tooth wear and mechanical overload on the temporomandibular joint. Both clenching and grinding are mechanical events. They involve force application, temporomandibular joint deformation, and tissue vibration. The physical quantity directly associated with these events is mechanical deformation, not the preceding electrical activity measured by electromyography.

This distinction guides the choice of transduction principle for the acquisition of mechanical events associated with bruxism. Surface electromyography measures muscle activity indirectly through electrical potentials detected at the skin surface. The technique is sensitive to movement artifacts, electromagnetic interference, and contamination by electrical signals from other muscles or even from the electrocardiogram, whose rhythmic activity may overlap with the signal of interest in the thoracic and cervical regions. The piezoelectric sensor, in turn, responds directly to the mechanical deformation caused by both clenching and grinding. It captures the physical quantity of interest without the interferences that affect conventional electromyographic methods [9].

B. Piezoelectricity

Piezoelectric materials generate an electrical potential difference when mechanically deformed, a property first described by the Curie brothers in 1880 [10]. The same materials also deform when subjected to an electric field. This bidirectional conversion between mechanical and electrical energy makes them suitable for detecting forces, pressures, vibrations, and displacements.

Piezoelectric transducers are well suited for biomedical instrumentation, especially for surface applications. Their dynamic response allows direct transduction of mechanical vibrations without requiring external excitation simplifying the conditioning circuit and reducing power consumption. These transducers also exhibit low drift compared to resistive or capacitive alternatives, which is advantageous for long-term monitoring [9]. Piezoelectric materials range from natural crystals like quartz to ceramics such as PZT and polymers including PVDF. Polymer-based devices, such as PVDF, offer mechanical flexibility and compact form factors.

Each type presents different trade-offs in terms of sensitivity, flexibility, and cost, influencing its suitability for a given application [10]. Polymer-based transducers offer high sensitivity and a broad linear range, capable of detecting sub-Newton forces [9].

III. Methodology

This section describes the components and procedures adopted for the instrumentation of the mechanical signal acquisition system for the temporomandibular joint, covering from the piezoelectric transducer to analog conditioning, digitization, and digital filtering.

A. System Overview

The proposed system, named Morpheus, is illustrated in Fig. 1. The hardware acquisition unit is the focus of the present article, including the analog signal conditioning and data acquisition stages.

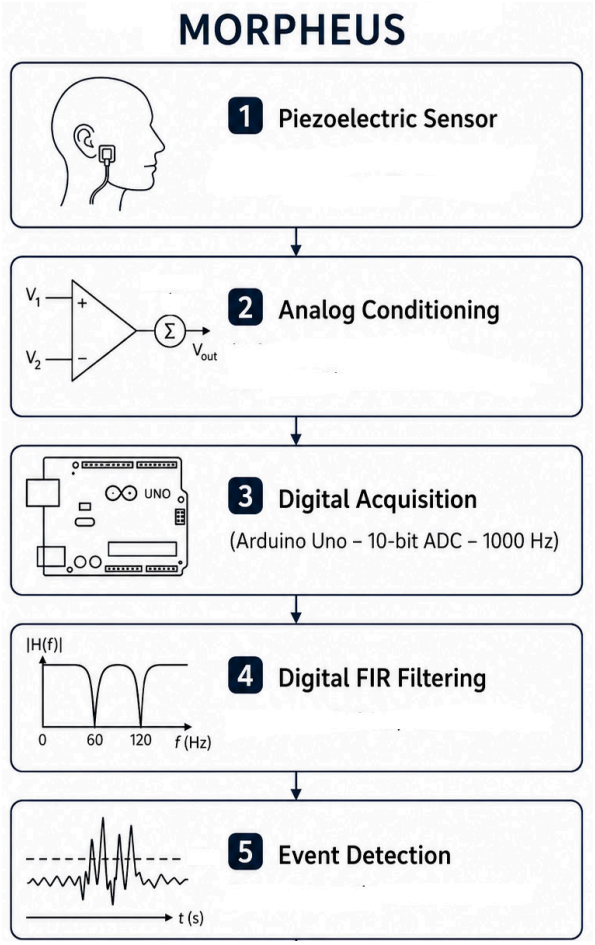


Fig. 1. Architecture of the instrumentation system.

B. Piezoelectric Transducer

The sensing element is a 27 mm piezoelectric disk with a brass substrate and ceramic coating, resonant at 4.5 kHz. Before placement, the transducer was wrapped with adhesive tape and then positioned over the temporomandibular joint, on the lateral region of the face, to capture mechanical deformations resulting from mandibular movements during clenching and grinding. The sensor output consists of millivolt-level signals with bipolar polarity. Since the microcontroller's analog-to-digital converter accepts a 0 to 5 V input range with 4.9 mV

resolution, a conditioning circuit was designed to shift and amplify the signal to the appropriate range.

C. Analog Conditioning Circuit

The conditioning circuit was designed in two stages: an initial amplifier and a summing amplifier with DC level shifting to offset the signal reference.

The mechanical deformation sensed by the piezoelectric transducer depends on the attachment method, including contact pressure, alignment, and fixation technique, which may affect the signal amplitude and morphology for the same mechanical event. To account for such variability, the conditioning circuit was designed to provide signal amplification, while the determination of the optimal gain and detection thresholds was left for a later stage.

The first stage of signal conditioning is presented in Fig. 2.

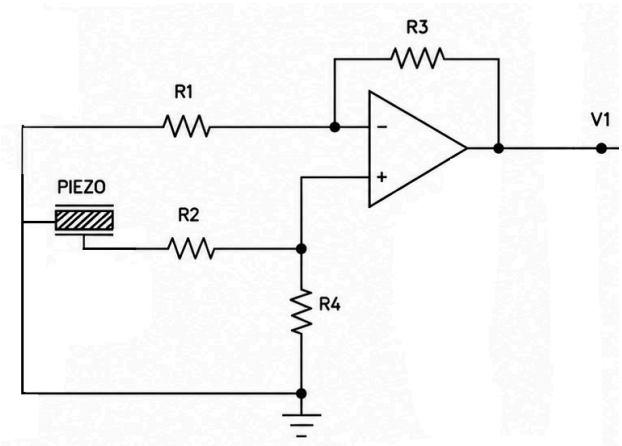


Fig. 2. Piezoelectric transducer amplifier configuration.

The potential difference generated by the piezoelectric sensor directly represents the relative mechanical deformation, making a single transducer sufficient for signal acquisition. Therefore, an additional input for establishing a signal reference is not required.

Applying Kirchhoff's current law at the amplifier input nodes yields the following relationship between the output voltage V_1 and the voltages at the sensor terminals V_{piezo}:

$$V_1 = \frac{R_4 \cdot (R_3 + R_1)}{R_1 \cdot (R_2 + R_4)} \cdot V_{piezo} \tag{1}$$

The second stage, presented in Fig. 3.

In addition to amplifying the signal, adds a DC voltage to the amplified signal, shifting its reference. Since the microcontroller's analog-to-digital converter operates in a range between 0 and 5 V, the added DC voltage must be 2.5 V, so that both positive and negative variations of the biological signal can be acquired by the microcontroller. The output equation of the summing amplifier is given by:

$$V_{out} = -\frac{R_f}{R5} \cdot V_1 - \frac{R_f}{R_6} \cdot V_2 \tag{2}$$

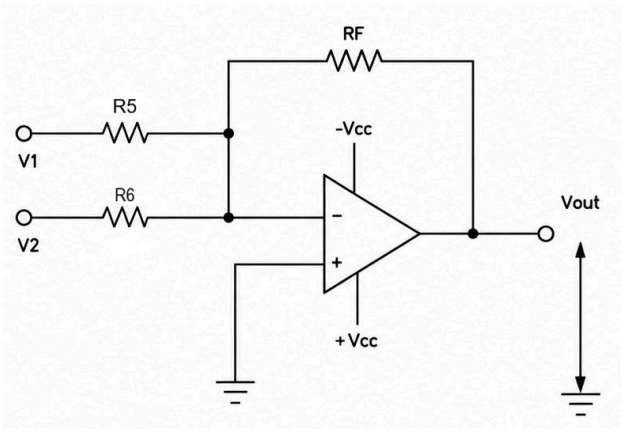


Fig. 3. Summing circuit for DC level shifting.

The signal from the initial amplifier is applied to one of the inputs of the inverting summing amplifier. Thus, the circuit output voltage can be obtained by combining the contribution of the amplified piezoelectric signal with the second input of the summing amplifier. Substituting the expression for V_1, given by Equation (1), into Equation (2), we obtain:

$$V_{out} = -\frac{R_f R_4 (R_3 + R_1)}{R_5 R_1 (R_2 + R_4)} V_{piezo} - \frac{R_f}{R_6} V_2 \tag{3}$$

Since V_2 is the DC voltage used to add a reference level to the conditioned signal, the ratio between resistors R_f and R_6 must be adjusted to $1/2$, so as to position the reference at the midpoint of the circuit's operating range. As the configuration used is an inverting summing amplifier, the contribution of V_2 undergoes polarity inversion at the output. Therefore, to produce a positive offset of $2.5\,\mathrm{V}$ at the output, the input V_2 is connected to a $-5\,\mathrm{V}$ reference.

Considering the condition

$$\frac{R_f}{R_6} = \frac{1}{2}, \tag{4}$$

Equation 3 can be rewritten as

$$V_{out} = -\frac{R_f R_4 (R_3 + R_1)}{R_5 R_1 (R_2 + R_4)} V_{piezo} - \frac{1}{2} V_2. \tag{5}$$

Since $V_2 = -5\,\mathrm{V}$, we obtain

$$V_{out} = -\frac{R_f R_4 (R_3 + R_1)}{R_5 R_1 (R_2 + R_4)} V_{piezo} + 2.5. \tag{6}$$

Therefore, the signal from the piezoelectric sensor is amplified and inverted, while a DC level of $2.5\,\mathrm{V}$ is added to the output, ensuring that the signal remains centered within the desired operating range.

D. Acquisition and Digitization

For digitization of the conditioned signal, the Arduino Uno microcontroller was used, which features the following relevant characteristics:

- 10-bit analog-to-digital converter (resolution of approximately 4.9 mV)
- 5 V reference voltage

As a starting point for selecting the resistors and sampling rate, information reported in the literature on electromyographic signals will be used. These signals exhibit amplitudes on the order of $10^{-3}\,\mathrm{V}$, reaching values close to $5\,\mathrm{mV}$, and frequencies between 5 and 500 Hz [12]. These parameters will be adjusted as the experimental analyses progress.

The digital value read by the ADC, V_r, is converted back to actual voltage and reference using the following expression:

$$V_p = -\left(V_r \cdot \frac{5}{1024} - 2.5\right) \tag{7}$$

E. Participants and Data Collection Protocol

Data were collected from four volunteers to evaluate the proposed signal acquisition and conditioning system. All participants and, where applicable, their legal guardians agreed to participate voluntarily in the data collection. For the minor participant, participation was authorized by the legal guardian, and the participant provided assent. P1 was a 23-year-old male who used an orthodontic appliance; P2 was a 17-year-old female without an orthodontic appliance; P3 was a 23-year-old male without an orthodontic appliance; and P4 was a 24-year-old female who used an orthodontic appliance.

The recordings were performed in a controlled environment. The participants were instructed to perform clenching, teeth grinding, and temporomandibular joint rest. Each event lasted approximately 1 to 3 s. A total of 210 clenching events and 210 grinding events were recorded. P1 contributed 60 events from each class, while P2, P3, and P4 contributed 50 events from each class. The recordings from P1 and P2 were used for training, and those from P3 and P4 were used for validation.

The data were collected on three different dates. The sensor was repositioned after every 10 acquisitions, resulting in repeated sensor placements during the recordings. This procedure introduced changes in the sensor position across the collected events while maintaining the same acquisition protocol. The signals obtained from the different recordings showed the main characteristics described in the subsequent time- and frequency-domain analyses.

IV. Results

This section presents and discusses the results obtained during the development of the proposed instrumentation system. The presentation follows the same sequence adopted in the methodology, beginning with the definition and implementation of the signal conditioning circuit, followed by frequency-domain analysis, evaluation of noise effects and filtering performance, time-domain analysis of the acquired signals, and finally the initial systematic statistical treatment of the data, including variability analyses and signal-to-noise ratio.

The results presented focus on demonstrating the functionality of the signal acquisition and conditioning hardware, with examples of the temporal and spectral signals obtained. Classification was not performed, since the scope of this study is the development and initial characterization of the circuit, which can be integrated into a new device for monitoring.

A. Conditioning Circuit

For the first stage, resistance values were experimentally selected to provide adequate gain without saturation, resulting in $R_1 = 50\,\Omega$, $R_2 = 50\,\Omega$, $R_3 = 100\,\Omega$, $R_4 = 200\,\Omega$. Thus, Equation (1) reduces to:

$$V_1 = 2.4 \cdot V_{piezo} \tag{8}$$

In the second stage, resistors R_2 and R_f were initially implemented using potentiometers for gain adjustment during bench tests. After calibration and validation tests, the final values of $R_5 = 1\,k\Omega$, $R_6 = 20\,k\Omega$, and $R_f = 10\,k\Omega$ were defined, with $V_1 = V_o$ and $V_2 = -5\,V$. The ratio $R_2/R_f = 2$ was maintained to ensure the application of an offset of approximately 2.5 V, while simultaneously providing the necessary amplification to fit the signal within the acquisition system's operating range.

Thus, combining Equations (8) and (2), the final equation of the conditioning circuit is obtained:

$$V_{out} = -24 \cdot V_{piezo} + 2.5 \tag{9}$$

The circuit was assembled on a breadboard, allowing the necessary adjustments and tests to ensure the correct operation of the signal conditioning system. After defining the components and testing the circuit, its implementation was transferred to the printed circuit board shown in Figure 4.

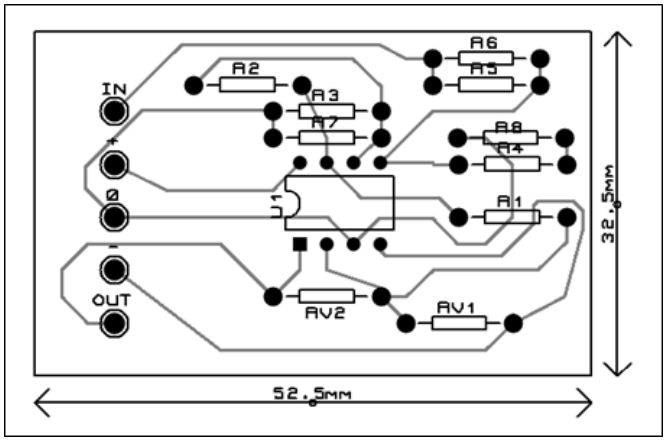


Fig. 4. Printed circuit board developed for the signal conditioning system.

The total gain of 24× was selected considering the 10-bit ADC resolution (approximately 4.9 mV per digital step) and the amplitude of the piezoelectric signals observed during the preliminary acquisitions. The 2.5 V DC offset allows the conditioned signal to vary around the midpoint of the ADC input range, providing an effective signal range of approximately ± 2.5 V around the reference level. The gain was therefore selected to increase the amplitude of the acquired signals while keeping the conditioned output within the 0–5 V ADC range.

B. Frequency-Domain Analysis

After data collection using the using a bench symmetrical power supply, spectral analysis of the signals corresponding to clenching and grinding movements was performed. For this purpose, the normalized Fast Fourier Transform (FFT) was applied to the acquired signals.

Figure 5 presents the normalized spectra of both movements. Clenching events concentrate most of their spectral energy below 20 Hz. Grinding events show a broader spectral distribution with a pronounced peak near 35 Hz.

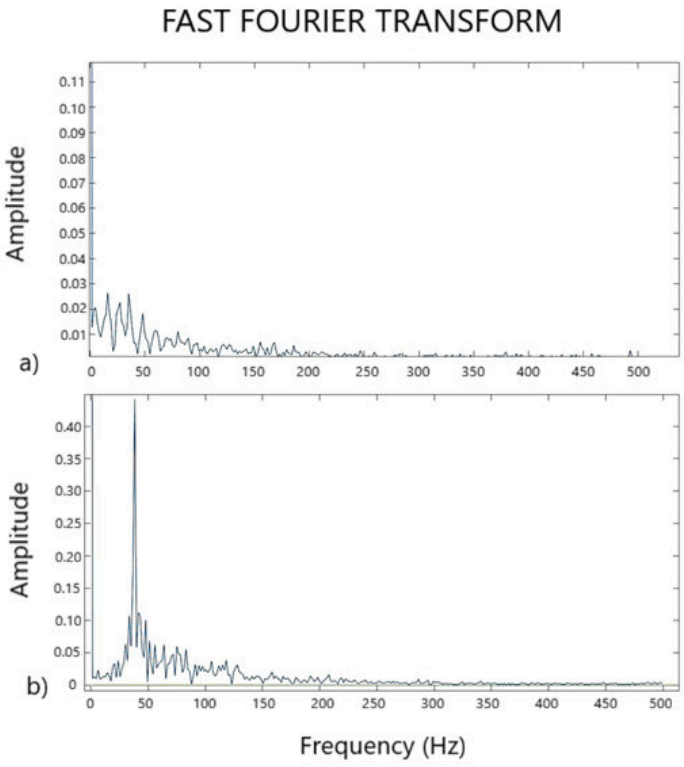


Fig. 5. Normalized spectra of clenching and grinding events obtained with a bench symmetrical power supply.

During the development of the portable version of the device, the circuit power supply was replaced by a symmetrical supply built from cell phone chargers. Although this solution allowed for greater system portability, the presence of interferences that hindered visual analysis of the acquired signals was observed.

To investigate this behavior, the frequency spectra of the same clenching event were compared using two power supply configurations: a bench symmetrical power supply (Figure 6) and a power supply based on cell phone chargers (Figure 7).

When comparing the spectra, the emergence of additional components near 60 Hz and 120 Hz is observed when using the portable power supply. Since the analysis of the characteristic movements indicated that most of the useful signal information lies below 45 Hz. These components are attributed to power supply noise.

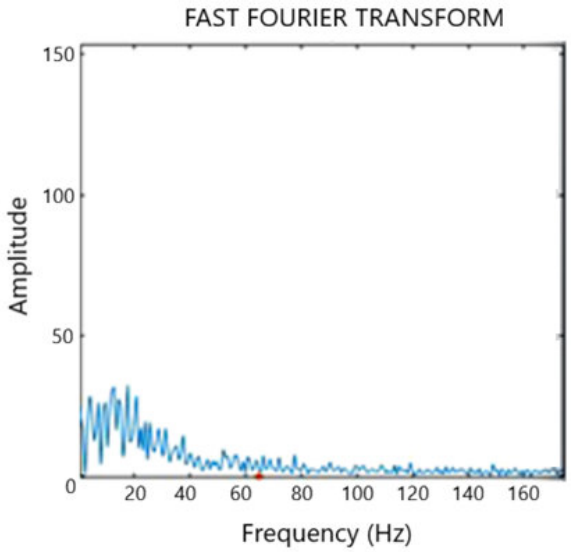


Fig. 6. Frequency spectrum of the clenching event using a bench symmetrical power supply.

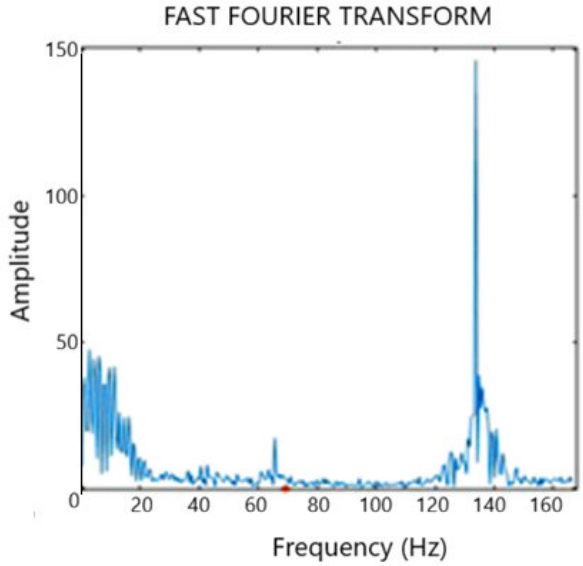


Fig. 7. Frequency spectrum of the clenching event using a power supply based on cell phone chargers.

A digital FIR low-pass filter was implemented to attenuate the noise components while preserving the movement-related frequencies. The filter was designed in MATLAB software using the designfilt() function, with a passband frequency of 50 Hz, stopband frequency of 57 Hz, maximum passband ripple of 1 dB, and minimum stopband attenuation of 60 dB.

After defining the parameters, the FIR filter coefficients were obtained, whose impulse response is shown in Figure 8. The obtained coefficients were then employed in digital signal filtering through the convolution operation.

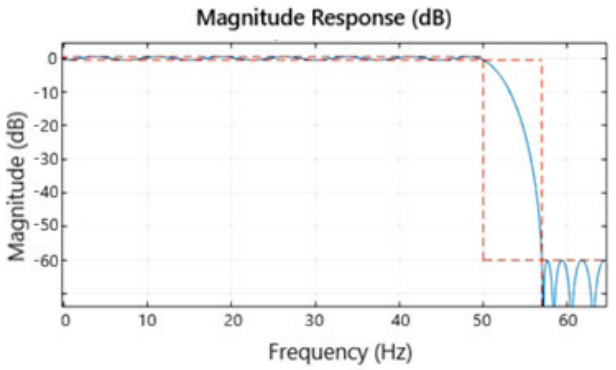


Fig. 8. Impulse response of the designed FIR filter.

C. Time-Domain Signal Analysis

Figure 9 presents a typical clenching event recorded by the system. A temporal pattern characterized by a negative amplitude rise associated with mandibular closure and a positive peak associated with relaxation of the masticatory musculature is observed. The noisy signal between these two moments is the vibration during tooth clenching.

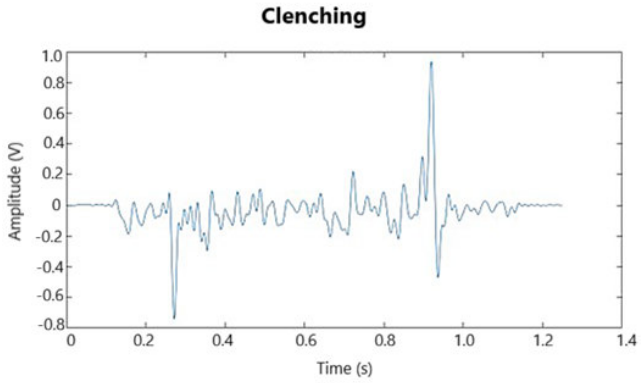


Fig. 9. Temporal signal corresponding to a clenching event.

Figure 10 presents a grinding event. Multiple successive oscillations appear, associated with lateral mandibular sliding.

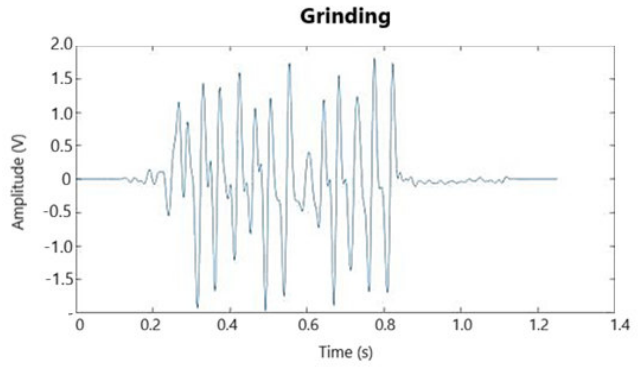


Fig. 10. Temporal signal corresponding to a grinding event.

This pattern differs from clenching due to the continuous displacement involved in grinding, rather than a single closing-opening cycle. Using the same dataset, previous work [11] reported that these signal features enabled automatic classification by neural networks.

D. Signal Quality and Variability

To quantitatively characterize signal quality, the signal-to-noise ratio (SNR) was calculated for each acquisition after DC-offset removal, using a 100-sample noise reference. The reference noise presented an RMS value of 0.0424.

Grinding events presented an SNR of 27.03 $\pm$ 2.75 dB and a CV of 10.16%. Clenching events presented an SNR of 12.97 $\pm$ 7.29 dB and a CV of 56.18%. Thus, grinding signals exhibited higher SNR and lower variability than clenching signals.

The greater variability observed for clenching may be related to differences in event intensity, sensor coupling conditions (e.g., contact pressure, positioning, and attachment), and the intrinsic characteristics of the clenching signal. Unlike grinding, which presents a more continuous oscillatory pattern, clenching signals may contain low-amplitude mechanical components in addition to the main compression and relaxation phases highlighted in Fig. 9, contributing to greater amplitude variability across acquisitions.

E. Cost Estimate

An estimate of the component-level hardware cost for a single bench prototype was performed based on retail prices from national suppliers in the first half of 2026. The 27 mm piezoelectric disk costs between R$ 0.79 and R$ 3.00; the LM358 operational amplifier costs between R$ 1.11 and R$ 2.20; and the Arduino Uno R3 development board costs approximately R$ 120.00. Including the printed circuit board, resistors, connectors, and power supply, the estimated bill of materials (BOM) for the laboratory prototype is approximately R$ 150.00 per unit [13]–[15].

For context, this value represents less than 2% of the retail price of a portable polysomnography system available on the Brazilian market, which costs R$ 14,700.00 [16]. Although this direct price comparison does not account for regulatory certification, large-scale manufacturing overhead, or the additional physiological sensors present in a full PSG (EEG, ECG, EMG, airflow, oximetry), it suggests that the proposed piezoelectric front-end could potentially be integrated into a low-cost, dedicated bruxism screening device at a fraction of the cost of multi-parameter PSG systems.

V. Conclusion

This paper described the design and benchtop characterization of a piezoelectric instrumentation system for TMJ signal acquisition. The conditioning system and digital filter captured signals with distinct features for clenching and grinding events under controlled conditions. The results validated the device's functionality under controlled laboratory conditions.

Although the piezoelectric sensor employed has characteristics that may offer practical advantages over surface electromyography, such as a simpler signal conditioning circuit and the use of a single sensing element, no direct experimental comparison between the two approaches was performed in this study. The results indicate that the piezoelectric approach can acquire and characterize mechanical events associated with clenching and grinding, supporting further investigation of its use in bruxism related monitoring applications.

The clinical validation, including comparison with polysomnography, is beyond the scope of this work, which focuses on hardware development and initial characterization. Future studies with larger samples and in real sleep environments are required to establish the sensitivity, specificity, and diagnostic accuracy of the system. The estimated hardware production cost aligns with the primary objective of developing a low-cost platform. The software cost will be estimated.

The current results are constrained by limitations that will be addressed in future work. The sample size precludes statistical generalization. Data were collected in a laboratory setting with participants awake; performance during sleep, including the effects of head movement and pillow contact, remains unevaluated.

Future work should prioritize:

A systematic metrological characterization of the acquisition system using a calibrated mechanical reference, including calibration and quantification linearity, and frequency response;

- A battery-powered, wireless revision to reduce conducted noise and improve wearability;
- A controlled evaluation of sensor attachment methods and geometries to assess their impact on signal consistency across different facial anatomies;
- Expanded data collection involving a larger number of participants and overnight recordings to evaluate the behavior of the acquired signals under natural sleep conditions;
- A clinical study with simultaneous PSG, surface EMG of the masseter muscle, and sensor recordings to compare the acquired mechanical signals with clinically scored sleep bruxism events. Such a study would provide the basis for subsequent evaluation of diagnostic performance using sensitivity, specificity, and ROC analysis.

REFERENCES

[1] F. Alóe, L. R. Gonçalves, A. Azevedo, and R. C. Barbosa, "Bruxismo do sono," *Revista Neurociências*, vol. 11, no. 1, pp. 4–17, Mar. 2003.

[2] M. Maluly, C. Dal-Fabbro, L. G. S. de Alencar, T. B. de Oliveira, and L. B. C. de Souza, "Prevalence of sleep bruxism in children and adolescents: a systematic review," *Sleep Medicine*, vol. 72, pp. 123–132, Aug. 2020.

[3] J. A. Cruz, "Percepção e consciência do bruxismo do sono em pacientes odontológicos," *Revista Brasileira de Odontologia*, vol. 78, p. e2345, 2021.

[4] D. V. G. Silva and S. D. Bonfa, "Bruxismo e suas implicações na saúde bucal e geral: diagnóstico, tratamento e abordagens preventivas – revisão de literatura," *RevistaFT*, 2024.

[5] Câmara dos Deputados, "Comissão aprova criação no SUS de política de atenção aos distúrbios do sono," Agência Câmara Notícias, Brasília, Brazil, Jul. 7, 2025. [Online]. Available: https://www.camara.leg.br/noticias/1121281-comissao-aprova-criacao-no-sus-de-politica-de-atencao-aos-disturbios-do-sono/

[6] T. Stenberg, K. Myllymaa, A. Muraja-Murro, S. Westeren-Punnonen, T. Hukkanen, J. Töyräs, R. Lappalainen, E. Mervaala, K. Sipilä, and S. Myllymaa, "Polysomnographic scoring of sleep bruxism events is accurate even in the absence of video recording but unreliable with EMG-only setups," *Sleep and Breathing*, vol. 24, pp. 893–904, Jun. 2020.

[7] A. U. Yap and A. P. Chua, "Sleep bruxism: current knowledge and contemporary management," *Journal of Conservative Dentistry*, vol. 19, no. 5, pp. 383–389, Sep.–Oct. 2016.

[8] T. Kato, S. Yamaguchi, M. Okura, A. Yoshida, and G. Lavigne, "Mandibular movement during sleep bruxism associated with current tooth attrition," *Journal of Oral Rehabilitation*, 2024.

[9] P. Liu, J. Li, X. Zhang, and Y. Wang, "Design and evaluation of a novel technology for ambulatory monitoring of bruxism events using a piezoelectric sensor," *Sensors and Actuators A: Physical*, vol. 295, pp. 532–540, 2019.

[10] J. F. Shackelford, *Introduction to Materials Science for Engineers*, 8th ed. Upper Saddle River, NJ, USA: Pearson, 2016.

[11] F. G. D. da Silva Filho, V. P. Pinto, M. A. B. Amora, R. N. C. Almeida, and J. O. F. Moura Filho, "Morpheus - ambiente de coleta, processamento e pré-diagnóstico do bruxismo com base em inteligência computacional," in *XVI Simpósio Brasileiro de Automação Inteligente e X Simpósio Brasileiro de Sistemas de Elétricos*, Fortaleza, Brazil, 2023, pp. 990–996.

[12] I. M. C. Leite, "Estudo de caso e viabilidade de transdutores miográficos," Trabalho de Conclusão de Curso (Graduação), Dept. Eng. Elétrica, Univ. Federal do Ceará, Fortaleza, Brazil, 2018.

[13] Eletrogate, "Pastilha Piezoelétrica 27 mm." [Online]. Available: https://www.eletrogate.com/pastilha-piezoeletrica-27mm-. Accessed: Jun. 21, 2026.

[14] Eletrogate, "Amplificador Operacional LM358." [Online]. Available: https://www.eletrogate.com/amplificador-operacional-tl741. Accessed: Jun. 21, 2026.

[15] Eletrogate, "Amplificador Operacional LM358." [Online]. Available: https://www.eletrogate.com/ci-lm358. Accessed: Jun. 21, 2026.

[16] Meditron MedSupply, "PSG Finder Reduzido." [Online]. Available: https://meditronmedsupply.com.br/produto/psg-finder-reduzido/. Accessed: Jun. 21, 2026.

Scoping Review of Non-Contact Capacitive Sensing for Human Movement Monitoring in Healthcare

1st Anelise Costa dos Santos
Master's Program in Biomedical Engineering
University of Brasilia (UnB)
Brasilia, Brazil
froane@gmail.com

2nd Rodrigo Ortega Tierno
Biomedical Engineering Graduate Program
Federal University of Uberlandia (UFU)
Uberlandia, Brazil
rodrigotierno.dev@gmail.com

3rd Wanghley Soares Martins
Department of Electrical and Computer Engineering
Duke University
Durham, NC, USA
me@wanghley.com

4th Rhaíra Helena Caetano e Souza
Assistive Technology Access and Research Center (CAPTA)
Federal Institute of Brasilia (CAPTA/IFB)
Brasilia, Brazil
rhaira.souza@ifb.edu.br

5th Adriano de Oliveira Andrade
Centre for Innovation and Technology Assessment in Health (NIATS)
Federal University of Uberlandia (UFU)
Uberlandia, Brazil
adriano@ufu.br

6th Cristiano Jacques Miosso
Postgraduate Program in Biomedical Engineering
University of Brasilia (UnB)
Brasilia, Brazil
miosso.unb@gmail.com

7th Fábio Henrique Monteiro Oliveira
Applied Computing Research Group
Federal Institute of Brasilia (GPCA/IFB)
Brasilia, Brazil
fabio.oliveira@ifb.edu.br

***Abstract*—Non-contact capacitive sensing (NCC) has emerged as a promising approach for unobtrusive human movement monitoring in healthcare and healthcare-related contexts. Unlike contact-based sensors, NCC technologies can detect movement, proximity, posture, gesture, gait, physiological signals, and body-related activity without requiring direct skin contact, which may improve comfort and support remote or continuous monitoring. This scoping review organized the literature on NCC technologies applied to human movement monitoring. A total of 452 records were imported from six scientific databases, and 72 qualified studies were included after screening and eligibility assessment. The studies were analyzed according to application domain, sensing configuration, data acquisition strategy, signal processing method, artificial intelligence and machine learning approaches, reported advantages, limitations, and open research questions.**

***Index Terms*—non-contact capacitive sensing, scoping review, human movement monitoring, healthcare, biomedical engineering, signal processing, machine learning, Parkinson's disease.**

I. Introduction

Over the past two decades, sensing technologies have become increasingly present in devices attached to, worn on, or placed near the human body. In particular, wearable, flexible, and stretchable sensors have been explored for human activity recognition, body signal monitoring, and health-related applications. In this context, capacitive sensors have received growing attention because they may offer low power consumption, fast response time, simple structure, great linearity, and the possibility of integration into wearable and unobtrusive systems [1].

Movement sensing plays an integral role in several aspects of daily life, spanning from academic research to healthcare institutions. Its applications range from monitoring periodic physical activities and human activity recognition to aiding in rehabilitation, clinical assessment, and long-term follow-up of patients. Particularly within the realm of health technology, movement sensing serves as a valuable technique for understanding human movement and its implications across diverse applications, including motor assessment, gait analysis, gesture recognition, and patient monitoring [2], [3].

In recent years, the rapid advancement of diagnostic, prognostic, assistive, and human-machine interaction technologies has emphasized the need for sensors capable of accurately detecting human movement in a comfortable and reliable way. This need is evident in both industrial and academic contexts, fueling demand for precise, low-cost, scalable, and non-contact sensing technologies. Consequently, there has been a growing interest in alternatives to traditional inertial and contact-based sensing methods, such as capacitive and electric field sensing, which constitute the primary focus of this study [3].

A non-contact capacitive (NCC) sensor belongs to a category of capacitive or electric potential sensing technologies capable of remotely detecting disturbances in electric fields surrounding the sensor when affected by nearby bodies, ob-

jects, or body segments. These technologies can detect changes associated with proximity, displacement, posture, gesture, and motion without requiring direct skin contact [2], [3]. NCCs have been explored across various domains, including human movement recognition, gesture recognition, physiological monitoring, gait analysis, activity recognition, and patient monitoring [3]–[5].

Specifically focusing on motion and gesture recognition, several sensor technologies have been employed to track position, displacement, or movement patterns over time. Inertial sensors, for example, are extensively used in wearable devices and research applications. However, unlike inertial sensors, which include accelerometers and gyroscopes, NCC sensors do not require direct physical contact with the body to operate. This characteristic may improve comfort, reduce preparation time, avoid skin irritation, and enable unobtrusive monitoring, especially in healthcare applications involving older adults, patients with sensitive skin, or individuals with motor disorders [6], [7].

Despite these advantages, NCC systems still face important technical and methodological challenges. Their performance can be affected by environmental noise, proximity to the target, low coupling capacitance, interference from nearby objects, motion artifacts, and variations in body position. These limitations highlight the need for robust signal processing strategies, noise reduction methods, and systematic benchmarking against established technologies such as inertial sensors, gyroscopes, optical motion capture systems, camera-based systems, and clinical assessment scales [5], [6].

Related works have addressed specific NCC applications in Parkinson's disease, hand motor task assessment, postural tremor analysis, human motion recognition, physiological monitoring, and the physical and material aspects of capacitive sensor design [6], [7]. However, the literature remains heterogeneous regarding sensor architecture, acquisition settings, target applications, comparison methods, performance metrics, and the use of artificial intelligence and machine learning for signal processing. Although AI and machine learning approaches have been increasingly applied to capacitive and electric field sensing data, their effectiveness varies according to the task, dataset, feature extraction procedure, and validation strategy [3].

This study was structured as a scoping review of the literature. This approach is appropriate for organizing a broad and heterogeneous research field, identifying trends, classifying evidence, and highlighting methodological and technological gaps. In this scoping review, 452 records were imported from ACM Digital Library, IEEE Digital Library, ISI Web of Science, ScienceDirect, Scopus, and Springer Link.

This scoping review provides a comprehensive examination of non-contact capacitive sensing and its application to human movement monitoring in healthcare and healthcare-related contexts. It investigates sensor systems, target applications, signal processing techniques, artificial intelligence and machine learning approaches, advantages, limitations, validation strategies, and open research questions. By expanding the number of analyzed studies and organizing the evidence through an analytical taxonomy, this work aims to support future research on NCC-based systems for rehabilitation, gait analysis, tremor assessment, fall detection, remote patient care, and diseases that affect movement, such as Parkinson's disease.

II. Materials and Methods

This study was conducted as a scoping review of the literature on non-contact capacitive sensing technologies applied to human movement monitoring in healthcare and healthcare-related contexts. The review process was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) [9]. The scoping review approach was adopted because the literature in this field is heterogeneous in terms of sensor architecture, acquisition protocol, target application, signal processing strategy, validation method, and performance metrics.

The review was guided by the following research questions: RQ1) What healthcare and healthcare-related applications have employed non-contact capacitive sensing for human movement monitoring? RQ2) Which sensing configurations and data acquisition strategies have been proposed? RQ3) Which signal processing techniques have been adopted? RQ4) Which artificial intelligence or machine learning approaches have been used? RQ5) What advantages, limitations, and open research questions remain in this field?

The scope of the review was structured according to the Population–Concept–Context (PCC) framework. The population comprised human subjects and human movement monitoring systems. The concept was non-contact capacitive sensing, including capacitive proximity, electric field, and electrostatic sensing approaches. The context included healthcare and healthcare-related applications, such as rehabilitation, motor assessment, physiological monitoring, assistive technologies, and human movement analysis.

The methodological workflow comprised four main stages: planning, database search, screening and eligibility assessment, and data extraction and synthesis. Differently from a conventional systematic review focused on estimating a pooled effect, this scoping review aimed to identify, classify, and analyze the main technological and methodological trends related to non-contact capacitive sensing for movement monitoring.

The selection process was conducted in sequential stages, including duplicate removal, title and abstract screening, eligibility assessment, and final inclusion. The records were organized in a spreadsheet and screened according to the inclusion and exclusion criteria. Five reviewers participated in the title and abstract screening stage, and two to three reviewers assessed the studies selected for full-text reading and eligibility confirmation. Disagreements or uncertain cases were resolved through discussion among the authors, considering the relevance of each study to non-contact capacitive sensing and human movement monitoring.

Figure 1 presents the PRISMA-ScR flow diagram of the identification, screening, eligibility, and inclusion process adopted in this scoping review.

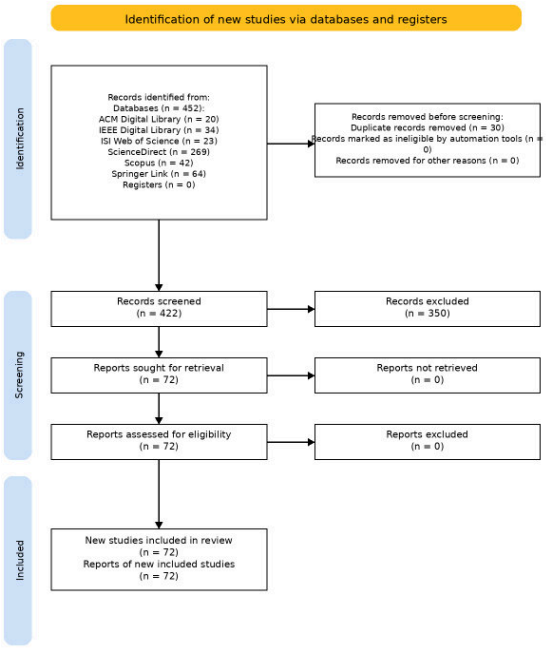


Fig. 1. PRISMA-ScR flow diagram of the identification, screening, eligibility, and inclusion process adopted in this scoping review.

A. Search Methods

The search strategy was developed based on terms commonly used in the literature on capacitive sensors, capacitive proximity sensing, electric potential sensing, electrostatic sensing, and human movement monitoring [1]–[3]. Searches were conducted in six scientific databases: ACM Digital Library, IEEE Digital Library, ISI Web of Science, ScienceDirect, Scopus, and Springer Link. The records were imported into Parsifal on November 14, 2025.

The search terms combined expressions related to the sensing technology and the target application. The general search strategy included the following terms: “capacitive”, “sensor”, “sensing”, “non-contact”, “non contact”, “noncontact”, “motion”, “movement”, “gesture”, “gait”, “human activity”, “rehabilitation”, and “healthcare”. Boolean operators and database-specific filters were adapted according to the search resources available in each database, as summarized in Table I.

A total of 452 records were imported from the selected databases. Titles and abstracts were screened to verify whether each study was related to non-contact capacitive sensing, electric field sensing, capacitive proximity sensing, or closely related approaches applied to movement detection, activity recognition, gesture recognition, gait analysis, physiological monitoring, rehabilitation, or healthcare-related monitoring. After duplicate removal, screening, and eligibility assessment, 72 qualified studies were included in the final review.

TABLE I
DETAILED SEARCH STRATEGY

Database Source	**Search Details**		
	Records[a]	***Specific Filters***	***Search Strategy***
ACM Digital Library	20	Conference and journal papers; English language	capacitive sensing AND movement OR gesture OR human activity
IEEE Digital Library	34	Conference and journal papers; English language	capacitive sensor OR capacitive sensing AND non-contact OR noncontact AND motion OR gait OR gesture
ISI Web of Science	23	Article and proceedings paper; English language	capacitive sensing AND non-contact OR noncontact AND movement OR healthcare
Science Direct	269	Research articles; English language	capacitive sensor OR capacitive sensing AND motion OR movement OR rehabilitation OR healthcare
Scopus	42	Title, abstract, and keywords; article and conference paper; English language	capacitive AND sensor OR sensing AND non-contact OR non-contact AND motion OR gesture
Springer Link	64	Conference paper, book chapter, and article; English language	capacitive sensing AND non-contact OR noncontact AND motion OR movement OR healthcare

[a]Total imported records: 452.

B. Eligibility Criteria

Studies were included when they met all the following criteria: (i) addressed non-contact capacitive sensing, capacitive proximity sensing, electric potential sensing, electrostatic sensing, or closely related capacitive approaches; (ii) involved movement detection, gesture recognition, gait analysis, activity recognition, physiological monitoring, rehabilitation, or healthcare-related monitoring; (iii) were published as full papers in journals, conference proceedings, or book chapters; and (iv) were written in English.

Studies were excluded when they were unrelated to capacitive or electric field sensing, focused exclusively on conventional contact-based sensors without relevance to non-contact capacitive approaches, were abstracts, editorials, posters, protocols, or incomplete records, were duplicates, or did not provide sufficient methodological information for analysis.

C. Data Extraction and Analysis

Data extraction was conducted using a standardized protocol adapted from the previous version of the review and refined to support a scoping review perspective. For each included study, information was collected regarding bibliographic data, sensor type, sensing principle, application domain, target movement or physiological variable, acquisition setup, population or experimental object, number of participants when reported, reference system, validation strategy, validation across individuals, acquisition environment, signal processing methods,

artificial intelligence or machine learning approaches, advantages, limitations, and open research questions.

To characterize the level of experimental and clinical evidence, the extraction also considered whether studies reported sample size, acquisition environment, reference system, validation across individuals, and population type. This assessment was used descriptively and was not applied as an exclusion criterion.

The extraction process resulted in a consolidated spreadsheet used to organize and compare the selected studies. The extracted variables were defined to capture the diversity of the field, including studies on hand movement detection [4], physiological signal monitoring [5], Parkinson-related motor assessment [6], [7], locomotion and gait analysis [10]–[14], hand gesture recognition with machine learning [15], [16], and textile-based capacitive sensing for pose tracking [17].

D. Data Synthesis and Analytical Taxonomy

The included studies were synthesized through descriptive and analytical mapping. The analysis did not aim to calculate a pooled effect size, but to classify the evidence and identify research trends, methodological patterns, and open challenges. The studies were grouped according to their main application domains, including gesture recognition, gait and locomotion analysis, physiological monitoring, Parkinson-related assessment, activity recognition, and wearable or textile sensing.

An analytical taxonomy was defined to organize the evidence according to: (i) sensor type and sensing configuration; (ii) application context; (iii) target movement or physiological variable; (iv) data acquisition setup; (v) signal processing strategy; (vi) artificial intelligence or machine learning method; (vii) validation protocol; and (viii) reported advantages, limitations, and research opportunities.

III. Results and Discussion

The scoping review resulted in the organization of the selected literature according to application domain, sensing configuration, data acquisition strategy, processing method, use of artificial intelligence or machine learning, reported advantages, limitations, and open research questions. A total of 72 qualified studies were included in the final review. Tables II and III present a condensed synthesis of selected representative studies. The corresponding supplementary extraction matrix, including sample size, population, acquisition environment, reference systems, validation strategies, cross-subject validation, AI/ML methods, limitations, and descriptive evidence level, is openly available on Zenodo [29]. Although the final review included 72 qualified studies, the complete extraction matrix is not reproduced in the article due to space limitations. The full set of studies was used to define the application domains discussed in this section, while the tables highlight studies with sufficient methodological detail to support comparative synthesis. The complete extraction matrix was also used to support the descriptive synthesis of sample size, reference systems, validation strategies, validation across individuals, and acquisition environments when reported.

A. Overview of the Included Studies

The analyzed studies show that non-contact capacitive sensing is not restricted to a single type of movement monitoring application. Instead, the technology has been explored across different domains, including Parkinson's disease assessment, gait and locomotion analysis, gesture recognition, human-computer interaction, textile-based monitoring, eye and facial movement tracking, physiological monitoring, and indoor presence detection.

B. Emerging Application Domains

The first domain involves Parkinson's disease and motor assessment, where NCC systems were used to detect hand movements, assess hand motor tasks, and analyze postural tremor [4], [6], [7]. In the mapped literature, the studies identified for Parkinson-related movement assessment using NCC are concentrated in works associated with GPCA/IFB and NIATS/UFU. This concentration emphasizes both the innovative character of the technology in this clinical context and the need for broader validation by other research groups, with larger samples, standardized protocols, and comparison with clinical reference measures.

Figure 2 illustrates the relevance of NCC systems for Parkinson-related hand motor assessment.

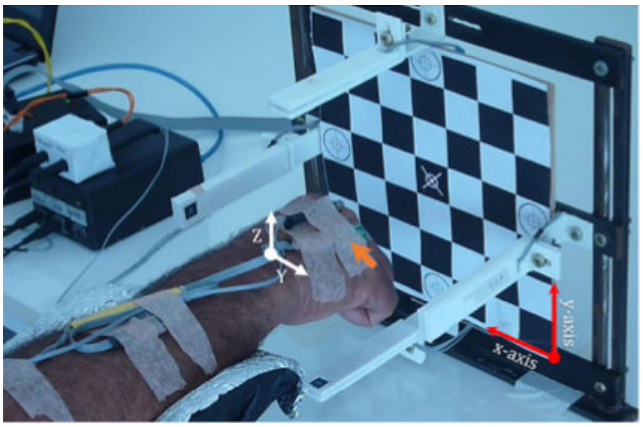


Fig. 2. Example of a non-contact capacitive sensing setup for hand motor assessment in Parkinson-related applications. The system illustrates the use of NCC sensors for detecting hand and wrist movements without direct skin contact. Adapted from [6].

The second domain includes gait, locomotion, and assistive devices. In this group, capacitive sensing was applied to locomotion mode recognition, gait phase detection, gait phase estimation, and temporal gait parameter estimation [10]–[14].

The third domain comprises gesture recognition and human-computer interaction. Studies in this category explored capacitive sensor arrays, textile capacitive sensors, distance-informed features, and machine learning classifiers to recognize hand gestures and support interaction with digital systems [15], [16], [24], [27], [28].

The fourth domain is related to wearable and textile-based body monitoring. Textile capacitive sensing was used

for upper-body pose tracking and body-integrated breathing monitoring [17], [19].

The fifth domain includes eye, blink, and facial movement monitoring. In this group, flexible electrostatic interfaces, capacitive skin glass, and carbon nanotube-paper composites were used for eye tracking, eye expression decoding, wearable eye movement monitoring, and blink detection [18], [20], [21].

The sixth domain includes physiological monitoring and indoor presence detection. Studies in this group used electric potential sensors, displacement current sensors, capacitively coupled electrode arrays, low-noise amplifiers, and OpenCapSense-based systems for ECG, respiration, EEG, EMG, bio-activity detection, in-bed ECG monitoring, and indoor presence detection [5], [22], [23], [25], [26].

C. Data Acquisition and Processing Strategies

The acquisition setups varied substantially among the included studies. Some systems used custom capacitive sensors and analog front-end circuits, while others used commercially available capacitive-to-digital converters, flexible electrodes, textile electrodes, printed sensors, or integrated amplifiers. Sampling frequencies also varied according to the application, ranging from low-frequency monitoring in wearable and indoor sensing systems to higher sampling rates in motor assessment, gait analysis, and physiological signal acquisition.

This heterogeneity limits direct performance comparison across studies, since sensor configurations, acquisition protocols, populations, reference systems, processing methods, and reported metrics vary substantially among application domains.

The most common data processing strategies included filtering, moving average smoothing, signal envelope extraction, nonlinear detrending, feature extraction, threshold-based detection, frequency-domain analysis, and classification. In studies involving gait and locomotion, sliding windows and feature-based classifiers were frequently used [10]–[12]. In studies involving tremor and motor assessment, methods such as moving average filtering, nonlinear detrending, empirical mode decomposition, and Hilbert spectrum analysis were used to characterize signal behavior [6], [7]. In physiological monitoring, filtering, amplification, signal conditioning, and waveform extraction were essential due to the weak coupling between the body and the sensing electrode [5], [25], [26].

D. Use of Artificial Intelligence and Machine Learning

Artificial intelligence and machine learning were not uniformly adopted across the included studies. Several papers did not report AI or machine learning methods, especially those focused on hardware design, sensing principles, or signal acquisition. However, when AI/ML was used, it played an important role in classification, feature interpretation, personalization, and movement recognition.

Machine learning methods identified in the analyzed papers included QDA, LDA, GMM, ECOC-SVM, KNN, SVM classifiers, decision trees, deep neural networks, LSTM models, explainable AI approaches, and personalized classifiers [10]–[12], [15]–[17], [21], [22], [24], [27]. Gesture recognition studies were the most strongly associated with AI/ML, especially because gesture patterns depend on user behavior, sensor placement, and signal variability. Gait and locomotion studies also used classifiers to distinguish locomotion modes and gait phases, indicating the relevance of machine learning for assistive devices and rehabilitation technologies.

Despite these advances, the review indicates that many studies still rely on small datasets, controlled experiments, or subject-specific calibration. Therefore, open questions remain regarding model generalization, cross-user validation, explainability, real-time performance, and the availability of standardized datasets for NCC-based movement monitoring.

E. Limitations of the Review

This scoping review has some limitations. The search was restricted to six scientific databases and to studies written in English. Grey literature, patents, theses, and technical reports were not included. In addition, no formal critical appraisal of methodological quality was conducted, since the objective was to map applications, methods, and research gaps rather than estimate intervention effects. Finally, the synthesis tables present selected representative studies rather than the complete set of 72 included papers, due to space restrictions in the conference format.

IV. Conclusion

This scoping review organized the literature on non-contact capacitive sensing for human movement monitoring in healthcare and healthcare-related contexts. From 452 imported records, 72 qualified studies were included and analyzed according to application domain, sensing configuration, data acquisition strategy, signal processing method, use of artificial intelligence or machine learning, reported advantages, limitations, and open research questions. The results show that NCC technologies are applied across multiple domains, including Parkinson's disease assessment, gait and locomotion analysis, gesture recognition, human-computer interaction, textile-based monitoring, eye and facial movement tracking, physiological monitoring, and indoor presence detection. Across the analyzed studies, signal processing and artificial intelligence appear as central components for extracting movement-related information from capacitive signals, especially through filtering, signal envelope extraction, feature extraction, frequency-domain analysis, and classification methods.

The review also revealed that research on Parkinson's disease assessment using non-contact capacitive sensing is currently concentrated within a limited number of studies associated with GPCA/IFB and NIATS/UFU, emphasizing both the novelty of this application and the need for broader validation by the scientific community. Despite the potential of NCC technologies for accessible, comfortable, and scalable monitoring, the field still faces challenges related to environmental noise, body position, sensor distance, low coupling capacitance, motion artifacts, electrode placement,

limited benchmarking, and reduced validation in real healthcare environments. Future research should focus on standardized acquisition protocols, larger and more diverse datasets, comparison with reference systems, interpretable machine learning, and long-term usability studies for rehabilitation, remote monitoring, and assistive healthcare systems.

Acknowledgment

The authors acknowledge the support of the Applied Computing Research Group of the Federal Institute of Brasília (GPCA/IFB), the Graduate Program in Biomedical Engineering of the University of Brasília (PPGEB/UnB), and the Centre for Innovation and Technology Assessment in Health of the Federal University of Uberlândia (NIATS/UFU).

This work was supported by the National Council for Scientific and Technological Development (CNPq; Grant Nos. 444437/2024-0, 442150/2023-7, and 405365/2023-3), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Foundation for Research Support of the State of Minas Gerais (FAPEMIG; Grant No. APQ-05015-24), and the Foundation for Research Support of the Federal District (FAPDF; Grant No. 00193-00002192/2023-37).

F.H.M.O. is supported by a FAPDF postdoctoral fellowship (Grant No. 23098.001922.2025-64). A.O.A. is a CNPq Research Productivity Fellow (Grant No. 302942/2022-0).

References

[1] A. J. Cheng, L. Wu, Z. Sha, W. Chang, D. Chu, C. H. Wang, and S. Peng, "Recent advances of capacitive sensors: Materials, microstructure designs, applications, and opportunities," *Advanced Materials Technologies*, vol. 8, no. 11, Art. no. 2201959, 2023, doi: 10.1002/admt.202201959.

[2] Y. Ye, C. Zhang, C. He, X. Wang, J. Huang, and J. Deng, "A review on applications of capacitive displacement sensing for capacitive proximity sensor," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2977716.

[3] S. Bian, M. Liu, B. Zhou, P. Lukowicz, and M. Magno, "Body-area capacitive or electric field sensing for human activity recognition and human-computer interaction: A comprehensive survey," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, Art. no. 4, 2024, doi: 10.1145/3643555.

[4] F. H. M. Oliveira, T. F. Zaruz, and A. O. Andrade, "Use of non-contact capacitive sensors to detect hand movements," in *XXVI Brazilian Congress on Biomedical Engineering*, Singapore: Springer Singapore, 2019, doi: 10.1007/978-981-13-2517-5_78.

[5] X. Tang, W. Chen, S. Mandal, K. Bi, and T. Ozdemir, "High-sensitivity electric potential sensors for non-contact monitoring of physiological signals," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3150587.

[6] F. H. M. Oliveira, D. F. Cunha, A. G. Rabelo, L. M. D. Luiz, M. F. Vieira, A. A. Pereira, and A. O. Andrade, "A non-contact system for the assessment of hand motor tasks in people with Parkinson's disease," *SN Applied Sciences*, vol. 3, no. 1, Art. no. 11, 2021, doi: 10.1007/s42452-020-04001-5.

[7] F. H. M. Oliveira, A. G. Rabelo, L. M. D. Luiz, A. A. Pereira, M. F. Vieira, and A. O. Andrade, "On the use of non-contact capacitive sensors for the assessment of postural hand tremor of individuals with Parkinson's disease," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, doi: 10.1109/EMBC.2019.8856746.

[8] B. Fang, N. D. Lane, M. Zhang, and F. Kawsar, "HeadScan: A wearable system for radio-based sensing of head and mouth-related activities," in *Proceedings of the IEEE Conference*, 2016.

[9] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. J. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, O. Tuncalp, and S. E. Straus, "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation," *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, 2018, doi: 10.7326/M18-0850.

[10] E. Zheng, L. Wang, Y. Luo, K. Wei, and Q. Wang, "Non-contact capacitance sensing for continuous locomotion mode recognition: Design specifications and experiments with an amputee," in *2013 IEEE International Conference on Rehabilitation Robotics (ICORR)*, Seattle, WA, USA, 2013, doi: 10.1109/ICORR.2013.6650410.

[11] E. Zheng, L. Wang, K. Wei, and Q. Wang, "A noncontact capacitive sensing system for recognizing locomotion modes of transtibial amputees," *IEEE Transactions on Biomedical Engineering*, 2014, doi: 10.1109/TBME.2014.2334316.

[12] E. Zheng, N. Vitiello, and Q. Wang, "Gait phase detection based on non-contact capacitive sensing: Preliminary results," in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*, 2015, doi: 10.1109/ICORR.2015.7281173.

[13] E. Zheng, S. Manca, T. Yan, A. Parri, N. Vitiello, and Q. Wang, "Gait phase estimation based on noncontact capacitive sensing and adaptive oscillators," *IEEE Transactions on Biomedical Engineering*, 2017, doi: 10.1109/TBME.2017.2672720.

[14] M. Li, P. Li, S. Tian, K. Tang, and X. Chen, "Estimation of temporal gait parameters using a human body electrostatic sensing-based method," *Sensors*, 2018, doi: 10.3390/s18061737.

[15] W. K. Wong, F. H. Juwono, and B. T. T. Khoo, "Multi-features capacitive hand gesture recognition sensor: A machine learning approach," *IEEE Sensors Journal*, 2021, doi: 10.1109/JSEN.2021.3049273.

[16] H. Lee, J. Bae, and H. Lee, "Toward utilization of explainable hand motion recognition in 3-D capacitive proximity sensing," *IEEE Sensors Journal*, 2024, doi: 10.1109/JSEN.2024.3388466.

[17] T. C. Yu, M. M. Zhang, P. He, C.-J. Lee, C. Cheesman, S. Mahmud, R. Zhang, F. Guimbretiere, and C. Zhang, "SeamPose: Repurposing seams as capacitive sensors in a shirt for upper-body pose tracking," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, doi: 10.1145/3654777.3676341.

[18] Y. Shi, P. Yang, R. Lei, Z. Liu, X. Dong, X. Tao, X. Chu, Z. L. Wang, and X. Chen, "Eye tracking and eye expression decoding based on transparent, flexible and ultra-persistent electrostatic interface," *Nature Communications*, 2023, doi: 10.1038/s41467-023-39068-2.

[19] J. Abounasr, M. El Gharbi, R. Fernandez-Garcia, and I. Gil, "Flexible body-integrated breathing monitoring system based on near-field coupling printed sensor," *Measurement*, 2026, doi: 10.1016/j.measurement.2025.119371.

[20] V. Sakthivelpathi, Z. Qian, T. Li, S. Ahn, A. B. Dichiara, R. Soetedjo, and J.-H. Chung, "Capacitive eye tracker made of fractured carbon nanotube-paper composites for wearable applications," *Sensors and Actuators A: Physical*, 2022, doi: 10.1016/j.sna.2022.113739.

[21] H. Wang, S. Zhang, H. Zhang, L. Song, Y. Li, and Z. Wang, "A capacitive skin glass for non-contact eye blink detection in complex environments," *Sensors and Actuators A: Physical*, 2025, doi: 10.1016/j.sna.2025.116410.

[22] L. Wang *et al.*, "Human presence and motion detection through electrostatic sensing," *Measurement: Sensors*, 2025, doi: 10.1016/j.measen.2024.101705.

[23] A. E. Mahdi and L. Faggion, "New displacement current sensor for contactless detection of bio-activity related signals," *Sensors and Actuators A: Physical*, 2015, doi: 10.1016/j.sna.2014.11.019.

[24] Y. Ye, Q. Zhang, X. Li, Y. Liu, Y. Bo, and Y. Lu, "Improving static gesture recognition with capacitive sensor arrays by incorporating distance measurements," *Measurement*, 2026, doi: 10.1016/j.measurement.2025.118697.

[25] K. Wang, B. Feng, D. Zhao, K. Xiao, G. Wu, B. Shi, R. Wang, J. Chen, K. Li, L. Li, G. Zuo, and C. Shi, "Noncontact in bed measurements of electrocardiogram using a capacitively coupled electrode array based on flexible circuit board," *IEEE Transactions on Instrumentation and Measurement*, 2023, doi: 10.1109/TIM.2023.3325864.

[26] Y. M. Chi, C. Maier, and G. Cauwenberghs, "Ultra-high input impedance, low noise integrated amplifier for noncontact biopotential sensing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2011, doi: 10.1109/JETCAS.2011.2179419.

[27] A. Nelson, G. Singh, R. Robucci, C. Patel, and N. Banerjee, “Adaptive and personalized gesture recognition using textile capacitive sensor arrays,” *IEEE Transactions on Multi-Scale Computing Systems*, 2015, doi: 10.1109/TMSCS.2015.2495100.

[28] Y. Ye, C. He, B. Liao, and G. Qian, “Capacitive proximity sensor array with a simple high sensitivity capacitance measuring circuit for human-computer interaction,” *IEEE Sensors Journal*, 2018, doi: 10.1109/JSEN.2018.2840093.

[29] A. Costa dos Santos, R. Ortega Tierno, W. Soares Martins, R. H. Caetano e Souza, A. de Oliveira Andrade, C. Jacques Miosso, and F. H. Monteiro Oliveira, “Supplementary extraction matrix for the scoping review of non-contact capacitive sensing for human movement monitoring in healthcare,” Zenodo, Dataset, ver. 1.0.0, 2026, doi: 10.5281/zenodo.22717919.

TABLE II
CONSOLIDATED DATA FROM THE ANALYZED PAPERS

Ref.	Emerging Application	Advantages	Limitations	AI/Machine Learning	Data acquisition	Data Processing Techniques	Open Research Questions
[6]	Hand motor assessment in Parkinson's disease	Patient comfort, no skin contact, useful for sensitive skin and motor disorders	Specific positioning, noise sensitivity, limited clinical sample	Statistical analysis	3 kHz sampling rate; custom non-contact capacitive system	Signal envelope, moving average filter, time-domain analysis	Long-term stability, clinical validation, quantification of PD motor signs
[10]	Locomotion mode recognition in amputees	Non-contact sensing inside prosthetic socket, comfortable, useful for assistive devices	Tested with limited participants and controlled locomotion modes	QDA classifier	Capacitive electrodes in prosthetic socket; pressure insole reference; 100 Hz sampling	Sliding windows, feature extraction, classification	Improve real-world prosthesis control and long-term reliability
[17]	Textile-based upper-body pose tracking	Comfortable, embedded in clothing, no camera occlusion, wearable design	Noise, durability, stitch type effects, need for fine-tuning	Deep learning, LSTM	Eight capacitive sensing seams; custom board; 32 Hz sampling	Median pooling, SMPL-based pose estimation, median filtering, MAE evaluation	Improve durability, add more seams, handle noise and garment variability

Source: original research results. Additional representative studies and NR details are available in [29].

TABLE III
SELECTED REPRESENTATIVE STUDIES FROM THE ANALYZED PAPERS

Ref.	Emerging Application	Advantages	Limitations	AI/Machine Learning	Data acquisition	Data Processing Techniques	Open Research Questions
[22]	Human presence and motion detection	Passive, privacy-preserving, contactless, useful for indoor monitoring	Cannot operate through walls, affected by nearby devices and environment	SVM classifier	OpenCapSense board; 50 Hz sampling	Low-pass and band-pass filtering, FFT, PSD, histogram features	Scalability, long-term reliability, and elderly care deployment
[25]	In-bed ECG monitoring	Noncontact ECG acquisition, suitable for bed-based patient monitoring	Sensitive to body movement, electrode placement, and noise	NR	Capacitively coupled electrode array on flexible circuit board	ECG filtering, signal conditioning, waveform extraction	Improve sleep monitoring, robustness, and clinical validation
[27]	Personalized textile gesture recognition	Wearable, adaptive, personalized, suitable for gesture interfaces	Requires user adaptation and textile sensor calibration	Personalized ML classifiers	Textile capacitive sensor arrays	Feature extraction, personalization, classifier training	Improve generalization, comfort, and long-term textile durability

Source: original research results. NR: not explicitly reported in the review source; see the supplementary matrix [29].

Termometria 3D por Ultrassom

Nilton Assugeni Neto
Department of Physics, FFCLRP
University of São Paulo (USP)
Ribeirão Preto, Brazil
nilton_aneto@usp.br

Lucas Murilo da Costa
Department of Physics, FFCLRP
University of São Paulo (USP)
Ribeirão Preto, Brazil
lucasdacosta@usp.br

Ricardo Ricci Bordonal
Department of Physics, FFCLRP
University of São Paulo (USP)
Ribeirão Preto, Brazil
ricardo.bordonal@usp.br

David Alejandro Collazos-Burbano
Department of Physics, FFCLRP
University of São Paulo (USP)
Ribeirão Preto, Brazil
david.collazos@usp.br

Theo Zeferino Pavan
Department of Physics, FFCLRP
University of São Paulo (USP) /
National Institute of Science and
Technology in Innovative Research in
Health Sciences: from Nanotechnology
to Artificial Intelligence (INCT)
Ribeirão Preto, Brazil
theozp@usp.br

Antonio Adilton Oliveira Carneiro
Department of Physics, FFCLRP
University of São Paulo (USP) /
National Institute of Science and
Technology in Innovative Research in
Health Sciences: from Nanotechnology
to Artificial Intelligence (INCT)
Ribeirão Preto, Brazil
adilton@usp.br

***Abstract*—Conventional thermal therapy monitoring often relies on high-cost modalities like MRI or is limited to 2D ultrasound planes, which may overlook critical off-plane hotspots. This work presents the implementation of an automated theranostic platform designed for 3D volumetric hyperthermia monitoring. The system integrates an open-architecture ultrasound scanner (ArtUs EXT-1H) with a custom-built mechanical scanning module and a 927 kHz high-intensity focused ultrasound (HIFU) transducer. By applying the Thermal Strain Imaging (TSI) method and the Lucas-Kanade optical flow algorithm to raw radiofrequency (RF) data, the platform reconstructed the induced thermal field across a 1.62° angular sweep on a paraffin-gel phantom. Experimental results demonstrated a localized heating pattern consistent with the focal geometry, reaching a maximum temperature of 41.6 °C. The volumetric mapping allowed for the visualization of discrete thermal slices, providing a new path to analyze a discret thermal map volume. This establishes a robust framework for non-invasive therapeutic monitoring in public health settings.**

Keywords—3D Ultrasound, HIFU, Hyperthermia, Volumetric Reconstruction, Thermal Strain Imaging, Theranostics.

I. INTRODUÇÃO

O agravamento global na incidência do câncer consolidou a doença como um dos principais desafios de saúde pública contemporânea [1], [2]. Projeções do Global Cancer Observatory (GCO) indicam um incremento nos diagnósticos de 60,2% em homens e 48,5% em mulheres até 2045 [3]. No cenário brasileiro, o INCA estima a ocorrência de aproximadamente 781 mil novos casos anuais para o triênio 2026-2028, sendo o câncer de mama a neoplasia mais frequente na população feminina (excluindo o de pele não melanoma) [4], [5], com previsão de alta de 40% na mortalidade específica até 2050 [6]. A gravidade deste quadro é acentuada em regiões de menor renda, onde disparidades de infraestrutura fazem com que cerca de 40% das detecções ocorram em estágios avançados, reduzindo drasticamente as chances de sobrevida relativa [7], [8].

Detectado o câncer, o protocolo clínico convencional envolve etapas que variam de biópsias a tratamentos agressivos como cirurgia e quimioterapia [9]. Como uma alternativa minimamente invasiva, a terapia por hipertermia utiliza o Ultrassom Focalizado de Alta Intensidade (HIFU) para elevar a temperatura do tecido alvo a aproximadamente 43 °C, sensibilizando as células tumorais e potencializando a eficácia de tratamentos adjacentes [10]. A tecnologia HIFU permite a destruição tumoral precisa com preservação de tecidos sadios, tornando o pós-operatório menos complexo [11], [12]. No entanto, a segurança e a eficácia desta modalidade dependem estritamente da deposição correta da dose térmica no volume planejado [12].

Atualmente, o monitoramento térmico padrão-ouro baseia-se na Ressonância Magnética (MRI), cujo alto custo e complexidade de infraestrutura limitam sua adoção, especialmente em países de média e baixa renda [13]. A termometria baseada em ultrassom via Imagem de Deformação Térmica (Thermal Strain Imaging - TSI) surge como uma alternativa acessível e não invasiva [14]. Entretanto, o monitoramento por ultrassom convencional é majoritariamente restrito a planos bidimensionais (2D). Essa limitação impede a captura da distribuição espacial total do calor, podendo ignorar pontos de aquecimento (hot spots) situados fora do plano de imagem, o que compromete a avaliação precisa dos mapas térmicos [15].

Visando superar essa lacuna, o Grupo de Inovação em Instrumentação Médica e Ultrassom (GIIMUS) tem atuado no desenvolvimento de scanners automatizados para reconstrução 3D [16], [17]. O presente trabalho apresenta a transição tecnológica dessas plataformas para uma

configuração dedicada à hipertermia volumétrica. O sistema proposto integra o tratamento por HIFU à reconstrução tridimensional de mapas térmicos a partir de varreduras angulares automatizadas. Esta abordagem visa avaliar a viabilidade de mapear a distribuição espacial de calor em três dimensões, estabelecendo uma prova de conceito para o monitoramento térmico volumétrico.

II. Material e Métodos

A. Instrumentação da Plataforma Teranóstica

A plataforma experimental (Fig. 1a e Fig. 1b) é estruturada em torno de uma cuba de acoplamento acústico contendo um motor DC controlado por uma ponte-H para a execução da varredura rotacional. O sistema é gerenciado por um microcontrolador responsável por automatizar a rotação e emitir sinais de *trigger* para a sincronização temporal entre os módulos. O subsistema de imageamento utiliza um scanner de arquitetura aberta ArtUs EXT-1H (Telemed Ultrasound) integrado a um transdutor linear L12-5N40-A4 de campo de visão de 36,8 mm. As aquisições ocorrem a uma taxa de 60 FPS em modo *Standard*. A sincronização entre o manipulador e o disparo do ultrassom é feita via hardware pela porta ScanStart Input, com conversão de nível lógico para 3,3 V para proteção do equipamento [17].

O sistema de hipertermia consiste em um transdutor de HIFU monoelemento focalizado (raio de concavidade de 88,89 mm e frequência central de 927 kHz [17]). A caracterização focal, calculada pelo método *Full Width at Half Maximum* (FWHM), identificou uma região focal acústica de 10 mm de profundidade por 2 mm de lateralidade e elevação (Fig. 2). Para a validação, utilizou-se um mimetizador de tecido (phantom) de parafina-gel (98 mm de diâmetro e 180 mm de altura), produzido conforme formulação do GIIMUS [18] (Fig. 1c).

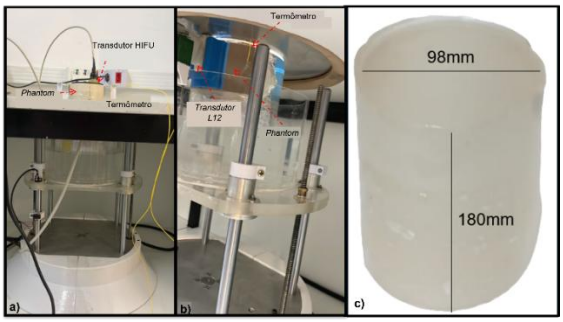


Fig. 1. **Montagem do sistema experimental. (a)** Vista frontal da plataforma mecânica; **(b)** detalhe do acoplamento acústico e transdutores; e **(c)** phantom de parafina-gel.

B. Etapa Experimental e Protocolo de Hipertermia

O protocolo iniciou-se com a aquisição de um frame de referência (pré-aquecimento) com o sistema submerso em água. Para a indução térmica, o gerador de funções foi configurado para uma saída senoidal de 2 Vpp a 927 kHz, elevada para aproximadamente 20 Vpp após amplificação para alimentar a cerâmica piezoelétrica. Essa tensão elétrica foi mantida constante para parametrizar indiretamente a deposição de energia acústica e assegurar a reprodutibilidade entre os ciclos. O período de aquecimento foi de 90 segundos.

Imediatamente após o aquecimento, realizou-se a coleta dos dados de RF pelo scanner ArtUs e a rotação automatizada do phantom em 0,54°. O sistema foi então freado, respeitando-se um intervalo de 2 minutos para resfriamento antes de reiniciar o ciclo, chegando à temperatura *baseline* inicial. Este procedimento (aquisição, aquecimento e rotação) foi repetido para 4 pontos angulares, totalizando uma varredura volumétrica discreta de 1,62°.

A varredura angular foi delimitada em 1,62° (quatro fatias discretas com passo de 0.54º) em conformidade com a extensão espacial do foco do transdutor de HIFU (2 mm de largura em elevação). Em aberturas angulares superiores a esse intervalo, a ausência de deposição acústica direta resultou em variações térmicas residuais (causadas apenas por difusão) insuficientes para induzir deformação térmica mensurável. Consequentemente, as alterações de fase no sinal de RF ficaram abaixo do limite de sensibilidade do algoritmo de Lucas-Kanade, restringindo a detecção de aquecimento a esse volume focalizado.

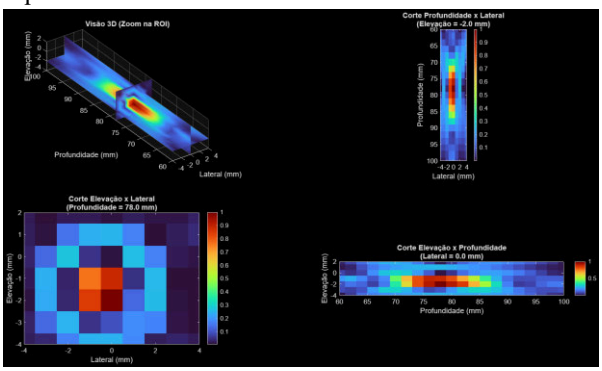


Fig. 2. **Mapeamento do campo acústico do transdutor de HIFU.** Visão 3D e fatias ortogonais 2D (profundidade, elevação e lateral) na região focal. A escala de cores indica a intensidade normalizada.

C. Processamento de Sinal e Reconstrução 3D

Os dados de RF pré e pós-aquecimento foram processados em ambiente MATLAB utilizando o método de TSI [14]. O algoritmo (Fig. 3) realiza o pré-processamento para seleção da região de interesse (ROI) e remoção de ruídos de fundo. O rastreio de microdeslocamentos axiais foi executado pelo algoritmo de fluxo óptico de Lucas–Kanade 2D [19]seguido pelo cálculo do gradiente de deformação (*strain*).

Após a suavização espacial do *strain*, realizou-se um ajuste linear (*fit*) utilizando a temperatura de referência obtida por um termômetro de fibra óptica inserido no *phantom*, determinando-se a constante de sensibilidade (K) do cálculo do TSI. Por fim, as fatias térmicas foram espacialmente coregistradas para a reconstrução do volume 3D discreto da temperatura.

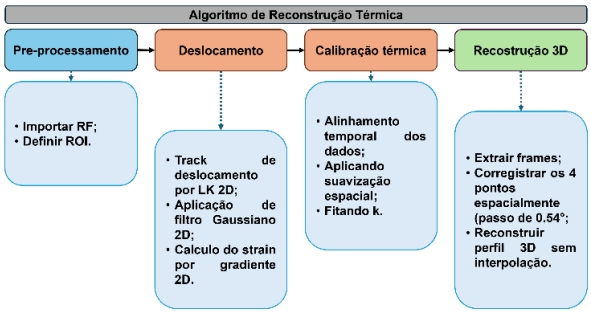


Fig. 3. **Fluxograma do algoritmo de reconstrução térmica.** Principais etapas do processamento dos dados de radiofrequência (RF), englobando o cálculo de deslocamento, calibração e montagem volumétrica 3D.

III. Resultados

Os resultados demonstraram um padrão de aquecimento espacialmente condizente com a geometria acústica do transdutor de HIFU, registrando uma temperatura máxima estimada de 41,6 °C no foco. A estimativa obtida via TSI apresentou concordância com a medição pontual da fibra óptica utilizada na etapa de calibração. Na caracterização volumétrica, a integração das quatro fatias angulares (varredura total de 1,62°) permitiu mapear o perfil térmico e identificar a presença de regiões aquecidas fora do plano principal de imagem (Fig. 4). A opção por uma visualização discreta sem interpolação preservou a resolução espacial nativa dos sinais de RF, evitando a suavização digital dos gradientes térmicos e delimitando o volume efetivo em que a variação de temperatura superou o limiar de detecção do algoritmo.

IV. Conclusão

Este trabalho apresentou uma prova de conceito para uma plataforma teranóstica de arquitetura aberta, demonstrando a viabilidade do mapeamento térmico 3D por ultrassom durante a aplicação de HIFU. A reconstrução volumétrica discreta permitiu a identificação de regiões aquecidas fora do plano principal de imagem, registrando uma variação térmica máxima de 41,6 °C. Contudo, o estudo possui limitações: a varredura esteve restrita a 1,62º (4 fatias) devido à dimensão focal do transdutor (2 mm) e ao limite de sensibilidade do algoritmo TSI para variações puramente difusivas; o ensaio utilizou um phantom sintético homogêneo (sem perfusão sanguínea ou atenuação biológica complexa); e a estimativa dependeu de um único ponto de calibração via fibra óptica, o que limita a validação quantitativa independente do erro térmico no volume. Como trabalhos futuros, pretende-se empregar arranjos independentes de múltiplos sensores de temperatura para mapeamento quantitativo do erro 3D, implementar algoritmos de compensação de movimento e realizar validações em tecidos biológicos, visando ao desenvolvimento contínuo de sistemas acessíveis de monitoramento por hipertermia.

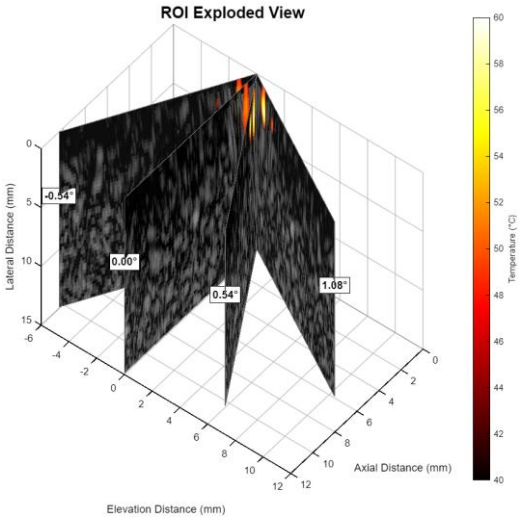


Fig. 4. **Mapeamento espacial 3D da temperatura induzida por HIFU.** A visualização explodida da região de interesse destaca as quatro fatias angulares discretas obtidas na varredura de 1,62°, preservando a resolução nativa do algoritmo TSI.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoa de Nível Superior (CAPES) pelos fomentos de código 001 e código 88887.228455/2025-00, à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP: 2018/16939-8, 2022/07463-5), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: 311224/2021 e 421531/2025-8) e à emenda de bancada do Senado 40940011. Também agradecemos aos técnicos e colaboradores do programa FAMB.

Referências

[1] World Health Organization (WHO), “World health statistics 2019: monitoring health for the SDGs,” Geneva: WHO, 2019.

[2] H. Sung *et al.*, “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.

[3] International Agency for Research on Cancer (IARC), “Cancer Tomorrow – Global Cancer Observatory,” Lyon: WHO, 2024. [Online]. Disponível em: https://gco.iarc.fr.

[4] Instituto Nacional de Câncer (INCA), “INCA estima 781 mil novos casos de câncer por ano no Brasil entre 2026 e 2028,” Rio de Janeiro: Ministério da Saúde, 2026.

[5] W. Primo, “National Cancer Institute and the 2023–2025 estimate - cancer incidence in Brazil,” *Rev. Bras. Ginecol. Obstet.*, vol. 45, no. 1, pp. 1-2, 2023.

[6] Brasil, Instituto Nacional de Câncer (INCA), “Mortalidade por câncer deve aumentar 98,5% até 2050,” Rio de Janeiro: INCA, 2025.

[7] O. Ginsburg *et al.*, “Breast cancer early detection: a phased approach to implementation,” *Cancer*, vol. 126, no. S10, pp. 2379–2393, 2020.

[8] N. P. D. de Oliveira *et al.*, “A multilevel assessment of the social determinants associated with the late stage diagnosis of breast cancer,” *Sci. Rep.*, vol. 11, Art. no. 2712, 2021.

[9] G. Menon, F. M. Alkabban, and T. Ferguson, "Breast cancer," in *StatPearls*, Treasure Island, FL, USA: StatPearls Publishing, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482286/

[10] A. L. Oei *et al.*, “Molecular and biological rationale of hyperthermia as radio- and chemosensitizer,” *Adv. Drug Deliv. Rev.*, vol. 163–164, pp. 84–97, 2020.

[11] N. M. Duc and B. Keserci, “Emerging clinical applications of high-intensity focused ultrasound,” *Diagn. Interv. Radiol.*, vol. 25, no. 5, pp. 398–409, 2019.

[12] V. Bachu *et al.*, “High-intensity focused ultrasound: a review of mechanisms and clinical applications,” *Ann. Biomed. Eng.*, vol. 49, no. 9, pp. 1975–1991, 2021.

[13] S. Murali *et al.*, “Bringing MRI to low- and middle-income countries: directions, challenges and potential solutions,” *NMR Biomed.*, vol. 37, no. 7, Art. no. e4992, 2024.

[14] Y. Hadadian *et al.*, “A novel theranostic platform: integration of magnetomotive and thermal ultrasound imaging with magnetic hyperthermia,” *IEEE Trans. Biomed. Eng.*, vol. 68, no. 1, pp. 68–77, 2021.

[15] C. J. Trujillo Romero and D.-L. Flores, Eds., *Diagnosis and Treatment of Cancer Using Thermal Therapies: Minimal and Non-invasive Techniques*, 1st ed. Boca Raton, FL, USA: CRC Press, 2023. doi: 10.1201/9781003342663.

[16] V. A. S. Luccas *et al.*, “3D automated breast ultrasound imaging: equipment and image reconstruction,” *Res. Biomed. Eng.*, vol. 37, pp. 519–533, 2021.

[17] N. Assugeni Neto, “Implementação de um sistema teranóstico de mama guiado por ultrassom,” Dissertação de Mestrado, FFCLRP-USP, Ribeirão Preto, 2025. [Online]. Disponível em: https://teses.usp.br/teses/disponiveis/59/59135/tde-04122025-183950/pt-br.html.

[18] S. L. Vieira et al., “Paraffin-gel tissue-mimicking material for ultrasound guided needle biopsy phantom,” Ultrasound Med. Biol., vol. 39, no. 12, pp. 2477–2484, Dec. 2013, doi: 10.1016/j.ultrasmedbio.2013.06.008.

[19] D. A. Collazos-Burbano *et al.*, "Beyond axial symmetry: tracking vectorial motion enhances nanoparticles mapping with magnetomotive ultrasound," *IEEE Access*, vol. 13, pp. 62704–62716, 2025, doi: 10.1109/ACCESS.2025.3557679.

Tetrapolar Electrode Geometries for Skin Phantoms Using Electrical Impedance Spectroscopy

Elkin Roman-Gomez
Electrical Engineering Department
Santa Catarina State University, UDESC
Joinville, Brazil
0009-0000-3159-7450

Pedro Bertemes-Filho
Electrical Engineering Department
Santa Catarina State University, UDESC
Joinville, Brazil
0000-0002-5264-4874

Marcos Antonio Rebello
Electrical Engineering Department
Santa Catarina State University, UDESC
Joinville, Brazil
0000-0003-4910-1833

***Abstract*—This work presents the fabrication and electrical characterization of multilayer gelatin-based phantoms designed to reproduce the passive electrical properties of the epidermis, dermis, and hypodermis, and to evaluate the sensitivity of tetrapolar electrode geometry to subsurface impedance changes. A temperature-resistivity calibration was first performed, comparing agar- and gelatin-based formulations; agar exhibited excessive intrinsic conductivity and a limited resistivity range, whereas gelatin provided both adequate mechanical stability and the target resistivity values needed for multilayer fabrication. Using an empirical regression relating temperature, NaCl concentration, and conductivity, three phantoms (P1, P2, P3) were built with identical epidermis and hypodermis resistivities but distinct dermis resistivities, allowing isolated evaluation of subsurface sensitivity. Electrical Impedance Spectroscopy (200 Hz–100 kHz) was performed on each phantom using five tetrapolar electrode configurations (three linear, two coplanar), and the resulting spectra were analyzed with Relative Difference (RD) and Bland–Altman statistics. Coplanar arrangements showed moderate, stable sensitivity to the middle layer variations, with impedance differences that gradually converged at higher frequencies. Among the linear configurations, Linear Arrangement 1 was the most sensitive, with relative differences exceeding 100% at low frequencies for the phantom with the greatest resistivity change, while Linear Arrangements 2 and 3 produced smaller but more consistent responses across the spectrum. Bland–Altman analysis confirmed that the observed differences were systematic rather than random, verifying that electrode geometry governs current distribution and sensing depth. The results demonstrate the potential of the proposed phantoms as a controlled experimental platform for assessing electrode configurations and electrical variations in subsurface skin layers, with educational applications using biological phantoms.**

***Index Terms*—Multilayer measurement, customized skin phantoms, EIS, electrode geometries**

I. Introduction

The skin, the largest organ of the human body, acts as a protective barrier against physical, chemical, and biological agents while maintaining essential homeostatic functions [1] [2]. It is composed of three main layers, namely the epidermis, dermis, and subcutaneous tissue, each exhibiting distinct structural and electrical characteristics [3]. Owing to its continuous exposure to harmful agents associated with skin cancer [3], considerable efforts have focused on developing minimally invasive techniques for early diagnosis, reducing unnecessary biopsies and improving clinical decision-making [4] [5] [6] [7] [8].

Electrical Impedance Spectroscopy (EIS) has emerged as a promising non-invasive technique for differentiating healthy and malignant skin tissues by exploiting changes in their passive electrical properties [9] [10] [11]. Variations in cell morphology, membrane integrity, and tissue composition modify tissue conductivity and capacitance. These alterations produce measurable changes in the impedance spectrum and have been associated with lesions such as Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and melanoma. [9] [12] [11] [13]

The multilayered structure of the skin strongly influences the measured impedance response. Since each layer presents different passive electrical properties, the measured spectrum results from the combined contribution of the epidermis, dermis, hypodermis, and embedded anatomical structures [14] [15] [13]. Moreover, the sensitivity of EIS depends on the distribution of the injected alternating current which is governed by electrode geometry, inter-electrode spacing, measurement configuration, and excitation frequency. [16] [17] [18] [12].

Direct investigation of these factors in vivo is limited by biological variability, ethical constraints, and the difficulty of reproducing controlled experimental conditions [19] [13] [20]. Consequently, biological phantoms have become an important alternative for evaluating bioimpedance systems under reproducible conditions [21] [22] [23] [24]. By reproducing the passive electrical properties of biological tissues, multilayer phantoms enable systematic investigations of current distribution, electrode sensitivity, and impedance variations while providing a low-cost platform for instrumentation development and experimental validation [16].

Commercial EIS systems, such as Nevisense®, have demonstrated promising clinical performance for melanoma detection [25]. However, their high cost and proprietary electrode configurations limit their use in research and educational environments, restricting experimental reproducibility and the development of new measurement methodologies [19]. Therefore, low cost multilayer skin phantoms can provide a controlled and accessible tool for teaching EIS, testing electrode configurations, and studying electrical variations in skin layers. These models may support the development of measurement

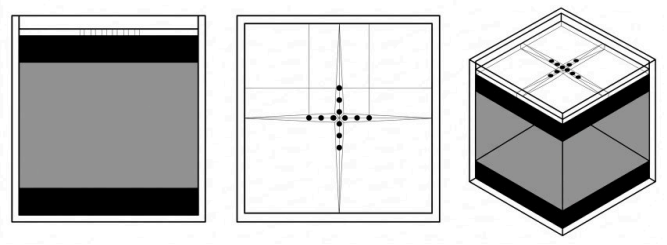


Fig. 1. Side, top, and isometric elevational views of the gelatin phantom. The visualization highlights the multilayered matrix and the corresponding perforated cover, prior to the final positioning and fixation of the measuring electrodes.

methods that may contribute to future diagnostic applications, including cancer detection.

This work investigates the sensitivity of a three layered gelatin skin phantom to variations in tetrapolar electrode geometry using Electrical Impedance Spectroscopy. The study focuses on controlled changes in dermis resistivity and does not evaluate cancer detection.

II. Materials and Methods

A. Creation of the Phantoms

Three multilayer skin phantoms were fabricated using three layers representing the epidermis, dermis, and hypodermis. The layers has thicknesses of 12.15 mm, 56.70 mm, and 12.15 mm, respectively, corresponding to 15%, 70%, and 15% of the 81 mm phantom depth. Each layer was prepared using distilled water and commercial gelatin with documented physicochemical specifications, while controlled amounts of sodium chloride (NaCl) were added to obtain the target electrical conductivity of each skin mimicking layer [26], [27]. The gelatin label indicates a pH of 5.2 in a 1% solution, moisture below 12%, ash below 2%, nitrogen below 15%, and a particle size of 30 mesh. The electrical properties were determined experimentally through the calibration procedure described below.

The phantom geometry was designed to approximate a homogeneous, isotropic, and semi-infinite medium only in the region adjacent to the electrode array where the electric field is established during the measurements. Each phantom was prepared in a cubic acrylic mold with internal dimensions of 81x81×8 mm and stored for 8 hours before measurements. A cover with twelve circular holes held 1.6 mm diameter polished stainless steel electrodes (AISI 308L) in a cross arrangement. The center to center spacing was 5 mm, with a maximum interelectrode distance of 25 mm. The cover weighes 40 g, resulting in a contact pressure of approximately 16 kPa.

B. Conductivity Temperature Calibration

Because the electrical properties of gelatin-based materials are temperature-dependent, controlled cooling ramps were performed prior to phantom fabrication to characterize the relationship between temperature, NaCl concentration, and electrical conductivity. One gelatin-agar and three gelatin-only

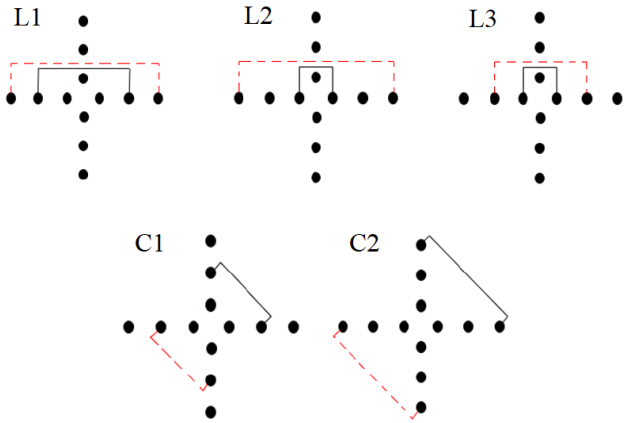


Fig. 2. The five different configurations for linear and coplanar interelectrode arrays. The dotted lines correspond to the current electrodes and the solid lines to the voltage electrodes respectively.

formulations were evaluated for their temperature-dependent electrical behavior and mechanical stability. The resulting calibration curves were used to estimate resistivity at the target measurement temperature (10°C) and to guide formulation selection for multilayer phantom fabrication [27].

TABLE I
Temperature-dependent resistivity of the evaluated phantom formulations.

T(°C)	A(Ω·m)	B(Ω·m)	C(Ω·m)	D(Ω·m)
60.0	19.75	4.74	7.39	13.33
57.5	19.75	5.24	7.64	13.50
55.0	21.20	5.30	7.86	13.88
52.5	21.55	5.54	8.08	14.03
50.0	22.16	5.74	8.40	14.77
47.5	22.94	5.92	8.64	15.15
45.0	23.57	6.15	8.97	15.76
42.5	24.00	6.37	9.25	16.32
40.0	25.33	6.89	9.55	16.92
37.7	25.66	7.04	9.79	17.61
35.0	26.53	7.18	10.40	18.35
32.5	27.27	7.44	10.76	18.98
30.0	28.89	7.88	11.39	19.95

A: 4 g agar and 96 mL water; **B**: 8 g gelatin and 92 mL water; **C**: 6 g gelatin and 94 mL water; **D**: 3 g gelatin and 97 mL water. Resistivity values are expressed in Ω·m.

The formulation containing 3 g of gelatin and 97 mL of water was selected as the epidermis-equivalent layer. Based on the temperature–resistivity calibration, the resistivity at the measurement temperature (10 °C) was estimated using Eqs. (1)–(3), yielding an extrapolated value of approximately 23.84 Ωm. This value was adopted as the reference resistivity for the epidermis-equivalent phantom, while the target resistivities for the remaining layers were determined according to the relative ratios reported in the literature.

$$\rho(T) = -0.2227\,T + 26.0690 \tag{1}$$

$$\rho(10) = -0.2227(10) + 26.0690 = 23.84\ \Omega\text{m} \quad (2)$$

$$\rho_{epi} = 23.84\ \Omega\text{m} \quad (3)$$

Rather than reproducing the absolute resistivity values reported for human skin, the multilayer phantom was designed to preserve the relative electrical contrasts among the epidermis, dermis, and hypodermis described in the literature [28], [29], [13]. Consequently, the target resistivities of the remaining layers were obtained by scaling the resistivity of the epidermis-equivalent formulation according to the relative ratios reported for biological skin tissues.

For the middle layer (dermis):

$$\rho_{der} = \rho_{epi}\left(\frac{20}{100}\right) = 23.84\left(\frac{20}{100}\right) = 4.77\ \Omega\text{m} \quad (4)$$

For the bottom layer (hypodermis):

$$\rho_{hyp} = \rho_{epi}\left(\frac{40}{100}\right) = 23.84\left(\frac{40}{100}\right) = 9.54\ \Omega\text{m} \quad (5)$$

TABLE II
TARGET ELECTRICAL PROPERTIES ADOPTED FOR THE MULTILAYER SKIN PHANTOM BASED ON RELATIVE RESISTIVITY SCALING.

Layer	Relative Ratio	Resistivity (Ωm)	Conductivity (S/m)
Epidermis	1.00	23.84	0.0419
Dermis	0.20	4.77	0.2096
Hypodermis	0.40	9.54	0.1048

Once the target apparent resistivities had been defined, two additional gelatin formulations consisting of 3 g of gelatin and 97 mL of water were prepared by incorporating different amounts of sodium chloride (NaCl), namely 0.010 g and 0.020 g. Similar to the previous experiments, both formulations were subjected to a descending temperature ramp from 60°C to 30°C, while conductivity measurements were acquired at regular intervals. These data were used to establish an empirical relationship among temperature, NaCl concentration, and electrical conductivity, enabling the estimation of the amount of salt required to achieve a target resistivity at a specified temperature. The resulting conductivity profiles are presented in Table II.

Controlled additions of NaCl were subsequently used to adjust the conductivity of each layer, resulting in target resistivities of 23.84 Ωm, 4.77 Ωm, and 9.54 Ωm for the top, middle, and bottom layers, respectively. To address this issue, an empirical multiple regression model was developed using gelatin formulations containing different sodium chloride concentrations and conductivity measurements acquired over a range of temperatures. Since the phantoms are fully solidified at the measurement temperature of 10°C, direct conductivity measurements at this stage are difficult to perform. Therefore, the model was employed to estimate both the amount of NaCl required to achieve a target electrical resistivity and the expected resistivity of the phantom at 10°C. The resulting relationship is given by:

TABLE III
TEMPERATURE-DEPENDENT ELECTRICAL PROPERTIES OF GELATIN PHANTOMS WITH DIFFERENT NACL CONCENTRATIONS.

T (°C)	A		B	
	σ (S/m)	ρ (Ω·m)	σ (S/m)	ρ (Ω·m)
60.0	0.1047	9.55	0.2096	4.77
57.5	0.1024	9.76	0.2054	4.87
55.0	0.1005	9.95	0.2015	4.96
52.5	0.0974	10.27	0.1952	5.12
50.0	0.0950	10.53	0.1902	5.26
47.5	0.0911	10.98	0.1826	5.48
45.0	0.0895	11.17	0.1791	5.58
42.5	0.0876	11.42	0.1754	5.70
40.0	0.0864	11.57	0.1730	5.78
37.7	0.0845	11.83	0.1691	5.91
35.0	0.0828	12.08	0.1656	6.04
32.5	0.0818	12.23	0.1636	6.11
30.0	0.0801	12.48	0.1602	6.24

A: Gelatin phantom prepared with 3 g gelatin, 97 mL distilled water, and 0.010 g NaCl. **B**: Gelatin phantom prepared with 3 g gelatin, 97 mL distilled water, and 0.020 g NaCl.

$$\sigma(T, m_{\text{NaCl}}) = 0.03908 + 2.87\times10^{-4}T + 8.3845\, m_{\text{NaCl}}, \quad (6)$$

where σ is the electrical conductivity (S/m), T is the temperature (°C), and m_{NaCl} is the mass of sodium chloride added to the formulation (g).

Since all impedance measurements were carried out at 10°C, Eq. (6) simplifies to

$$m_{\text{NaCl}} = \frac{\dfrac{1}{\rho_{\text{obj}}} - 0.04195}{8.3845}. \quad (7)$$

C. Measurements and Data Collection

Three different phantoms (P1, P2, and P3) were independently constructed to investigate the influence of the middle layer resistivity on the measured impedance. The top and bottom layers were kept unchanged across all phantoms, while only the middle layer resistivity was varied to represent different electrical conditions.

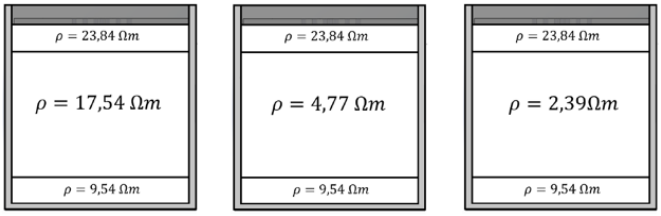


Fig. 3. Phantoms P1, P2, and P3 with the respective resistivity values of each layer.

EIS measurements were done by using the EmStat4R potentiostat (PalmSens BV) operating in a tetrapolar configuration.

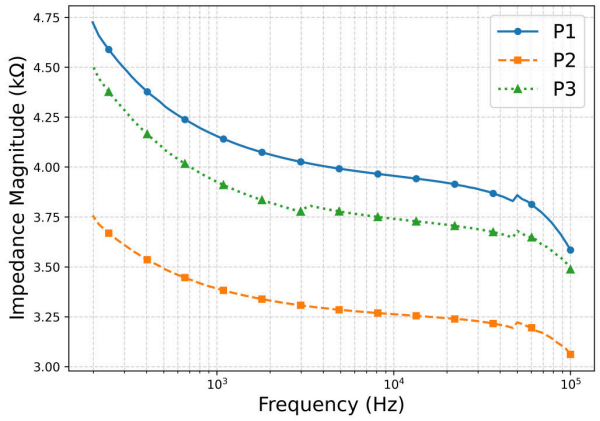


Fig. 4. Impedance magnitude ($|Z|$) spectra of Phantoms P1, P2, and P3 obtained using the Coplanar Arrangement 1 electrode configuration over the frequency range from 200 Hz to 100 kHz. The figure illustrates the influence of variations in the resistivity of the middle layer on the surface impedance response of the multilayer phantom..

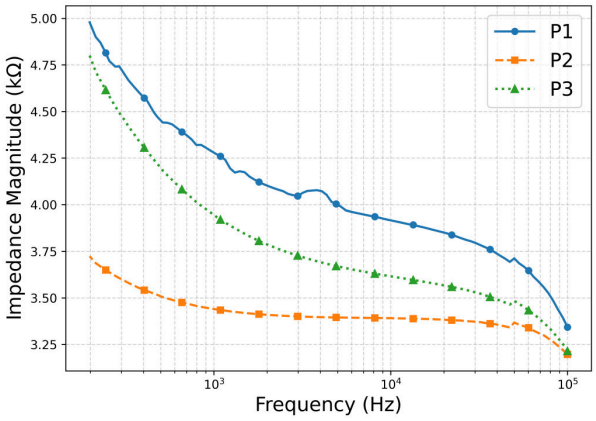


Fig. 5. Impedance magnitude ($|Z|$) spectra of Phantoms P1, P2, and P3 obtained using the Coplanar Arrangement 2 electrode configuration over the frequency range from 200 Hz to 100 kHz. The figure illustrates the influence of variations in the resistivity of the middle layer on the surface impedance response of the multilayer phantom.

Impedance magnitude and phase were measured from 0.2 to 100 kHz, recording. Each phantom was measured using the five electrode configurations, resulting in fifteen experimental datasets. In addition, each layer was individually characterized before depositing the subsequent layer to validate its electrical properties.

The sensitivity of each electrode configuration to middle layer resistivity variations was evaluated by comparing the impedance spectra of the three phantoms. Relative Difference (%RD) and Bland–Altman analyses were employed to quantify the differences between the reference phantom (P2) and the modified phantoms (P1 and P3), following the methodology described in [30].

III. Results

Figures 4–8 present the impedance magnitude ($|Z|$) and real impedance (Z') spectra obtained using the three linear and two coplanar electrode configurations. For each configuration, the responses of Phantoms P1, P2, and P3 are presented over the frequency range from 200 Hz to 100 kHz, enabling a direct comparison of the sensitivity of each electrode configuration to the variations introduced in the middle layer resistivity.

As observed in Figures 4 and 5, the Coplanar Arrangement 1 and Coplanar Arrangement 2 electrode configurations exhibited higher impedance magnitude and real impedance values for Phantoms P1 and P3 than for Phantom P2 over most of the analyzed frequency range. These differences remained nearly constant throughout the spectrum, with Phantom P1 consistently exhibiting slightly higher values than Phantom P3. However, the responses of the three phantoms gradually converged as the frequency increased, becoming more similar at frequencies above 100 kHz.

For the Linear Arrangement 1 electrode configuration, a nearly constant difference of less than 1000 Ω was observed between Phantoms P1 and P2 over the entire frequency range for both the impedance magnitude and the real impedance.

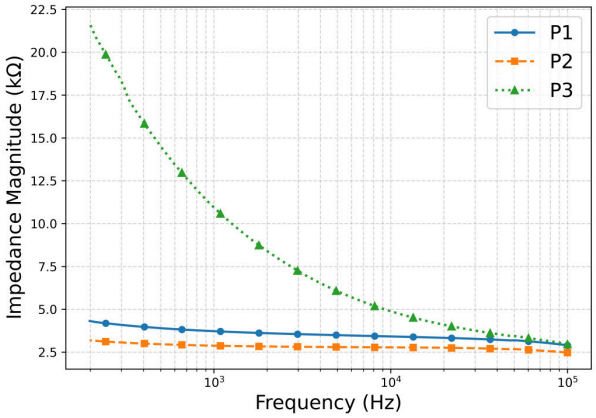


Fig. 6. Impedance magnitude ($|Z|$) spectra of Phantoms P1, P2, and P3 obtained using the Linear Arrangement 1 electrode configuration over the frequency range from 200 Hz to 100 kHz. The figure illustrates the influence of variations in the resistivity of the middle layer on the surface impedance response of the multilayer phantom.

In contrast, Phantom P3 exhibited a markedly different behavior, with impedance magnitude and real impedance values approximately five to ten times higher than those of Phantoms P1 and P2 at low frequencies. As the frequency increased, both parameters decreased progressively, converging toward the responses of P1 and P2 at approximately 100 kHz.

For the Linear Arrangement 2 electrode configuration, higher impedance magnitude and real impedance values were observed for Phantom P1 at the lower end of the frequency spectrum (200 Hz). In contrast, Phantoms P2 and P3 exhibited similar responses, maintaining an approximately constant separation throughout most of the analyzed frequency range despite the gradual decrease in impedance with increasing frequency. As the frequency approached the upper limit of the investigated spectrum, the responses of all three phantoms converged rapidly toward lower impedance values.

For the Linear Arrangement 3 electrode configuration, a

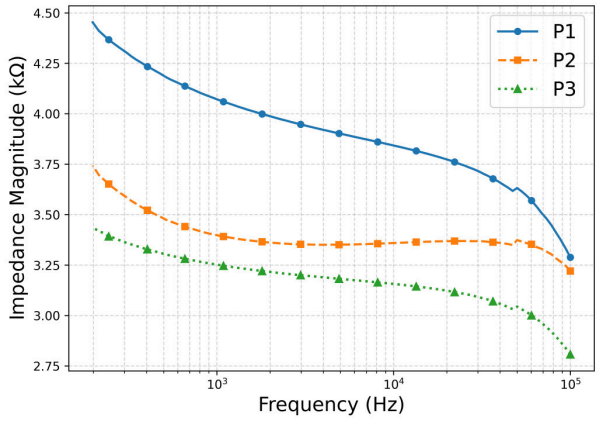


Fig. 7. Impedance magnitude ($|Z|$) spectra of Phantoms P1, P2, and P3 obtained using the Linear Arrangement 2 electrode configuration over the frequency range from 200 Hz to 100 kHz. The figure illustrates the influence of variations in the resistivity of the middle layer on the surface impedance response of the multilayer phantom.

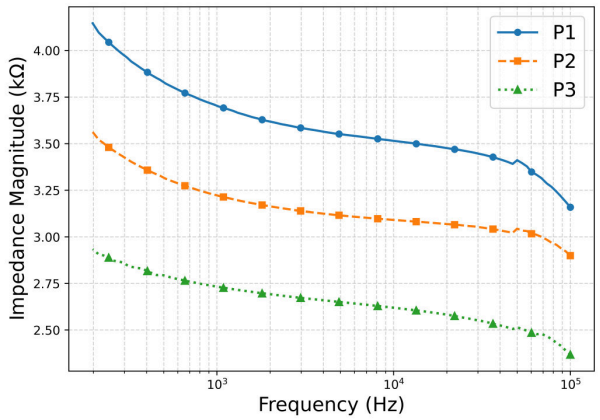


Fig. 8. Impedance magnitude ($|Z|$) spectra of Phantoms P1, P2, and P3 obtained using the Linear Arrangement 3 electrode configuration over the frequency range from 200 Hz to 100 kHz.The figure illustrates the influence of variations in the resistivity of the middle layer on the surface impedance response of the multilayer phantom.

nearly constant separation between the responses of the three phantoms was observed over the entire analyzed frequency range. Both the impedance magnitude and the real impedance exhibited a gradual decrease from 200 Hz to 100 kHz, with the three spectra following approximately parallel trends and maintaining a similar difference throughout the spectrum.

Since the top layer resistivity was identical in all three phantoms, these results suggest that the Linear Arrangement 3 configuration is predominantly sensitive to variations in the electrical properties of the deeper layers. The nearly uniform separation between the spectra indicates that changes in middle layer resistivity produce a proportional variation in the measured impedance over the investigated frequency range.

To complement the analysis of the impedance spectra, a statistical evaluation was subsequently performed using the relative difference (%RD) and Bland–Altman analysis. These methods allowed an objective quantification of the differences between the reference phantom (P2) and Phantoms P1 and P3, as well as an assessment of the agreement between the impedance measurements obtained with each electrode configuration [30].

A. Statistical Analysis

Figures 9–13 present the Relative Difference (%RD) and Bland–Altman plots for the five electrode configurations evaluated. %RD was calculated using P2 as the reference according to:

$$\%\mathrm{RD} = \frac{|Z_i - Z_{\mathrm{P2}}|}{Z_{\mathrm{P2}}} \times 100, \tag{8}$$

where Z_i represents the impedance magnitude of P1 or P3 at each frequency.

The Bland–Altman analysis was used to characterize the distribution and bias of both reference (P2) and modified resistivity phantoms. The absolute impedance difference was calculated as $\Delta Z (= |Z_i - Z_{P2}|)$ in order to avoid interpretation based only on the reference magnitude.

Values above 100% indicate that the absolute impedance difference exceeded the P2 reference value. Because the calculation uses the reference impedance as the denominator, then small reference values produce large %RD values. Therefore, %RD is better interpreted together with the absolute impedance difference.

The Relative Difference plots show the percentage variation in impedance magnitude as a function of frequency for the P1–P2 and P3–P2 comparisons. This analysis identifies the frequency intervals in which changes in middle layer resistivity produced the largest deviations from the reference phantom, as well as the regions where the electrical response remained comparatively stable.

The Bland–Altman plots present the impedance differences as a function of the mean values, allowing the bias and distribution of the differences evaluated for each electrode configuration. Together, these analyses quantify the sensitivity of each electrode arrangement to changes in the middle layer resistivity in the multilayer phantoms [30].

It is important to emphasize that, in the context of the present study, the objective of the %RD analysis is not to minimize the relative difference, but to evaluate the sensitivity of each electrode configuration to changes middle layer resistivity. Consequently, a %RD value close to zero indicates that the electrode configuration is relatively insensitive to the modifications introduced in the inner layer. Conversely, a high and consistent %RD indicates that the configuration is capable of detecting changes in resistivity of the middle layer from surface impedance measurements. Likewise, a stable %RD over the analyzed frequency range indicates a consistent response of the electrode configuration to variations in the electrical properties of the middle layer.

Unlike the relative difference analysis, which quantifies the sensitivity of each electrode configuration to changes in resistivity of the middle layer. The Bland–Altman analysis evaluates the distribution and consistency of these differences over the analyzed frequency range. It is better to emphasize

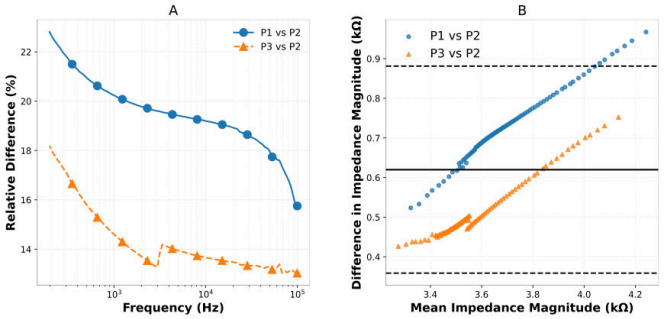


Fig. 9. (A) %RD in impedance magnitude for the P1–P2 and P3–P2 comparisons obtained with the Coplanar Arrangement 1 electrode configuration. (B) Bland–Altman analysis of the corresponding measurements, with a mean bias of 619.83 Ω and limits of agreement defined as ±1.96 SD.

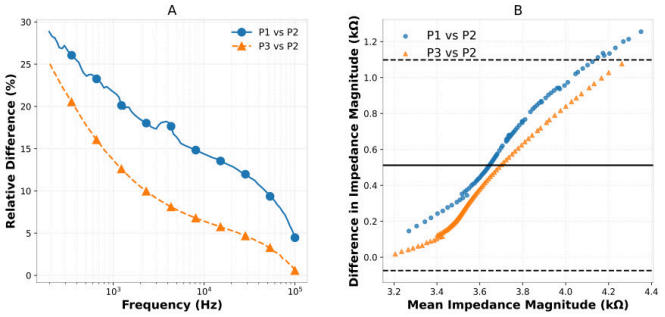


Fig. 10. (A) %RD in impedance magnitude for the P1–P2 and P3–P2 comparisons obtained with the Coplanar Arrangement 2 electrode configuration. (B) Bland–Altman analysis of the corresponding measurements, with a mean bias of 511.89 Ω and limits of agreement defined as ±1.96 SD.

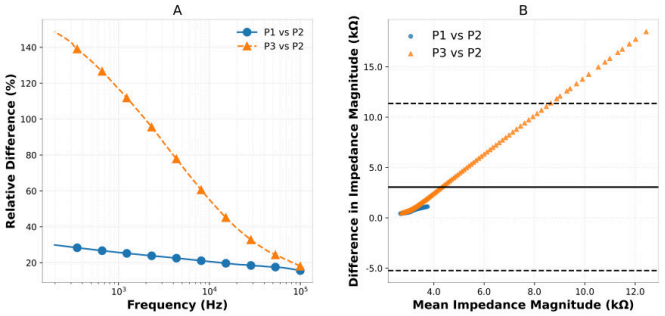


Fig. 11. (A) %RD in impedance magnitude for the P1–P2 and P3–P2 comparisons obtained with the Linear Arrangement 1. (B) Bland–Altman analysis of the corresponding measurements, with a mean bias of 3062.96 Ω and limits of agreement defined as ±1.96 SD.

that the objective here was not intended to obtain a bias close to zero, since the phantoms were intentionally prepared with different resistivities for the middle layers. Therefore, a consistent bias with a compact distribution of the differences indicates a stable response to the introduced changes in resistivity of the middle layers.

For the Coplanar Arrangement 1 electrode configuration, the Relative Difference (%RD) between Phantoms P1 and P2 decreased from approximately 22% at 200 Hz to 15% at 100 kHz. Similarly, the comparison between Phantoms P3 and P2 exhibited a gradual decrease from approximately 18% to 13% over the same frequency range. The smooth reduction in %RD with increasing frequency indicates that the impedance responses of the three phantoms became progressively more similar at higher frequencies while maintaining a consistent separation throughout the spectrum.

The corresponding Bland–Altman plot shows two well-defined ascending clusters associated with the P1–P2 and P3–P2 comparisons, with differences ranging from approximately 400 Ω to more than 1000 Ω. The calculated bias of 619.83 Ω indicates a systematic difference between the reference phantom (P2) and the modified phantoms. Furthermore, the compact distribution of the data around each trend suggests that these differences remain consistent over the analyzed frequency range, demonstrating that the Coplanar Arrangement 1 configuration is capable of reliably distinguishing variations in middle layer resistivity.

For the Coplanar Arrangement 2 electrode configuration, a behavior similar to that observed for Coplanar Arrangement 1 was obtained. The Bland–Altman analysis yielded a bias of 511.89 Ω, indicating a systematic difference between the reference phantom (P2) and the modified phantoms. The data points formed two closely spaced ascending clusters, with differences ranging from approximately 0 Ω to 1200 Ω, demonstrating a consistent trend over the analyzed frequency range.

The %RD also exhibited a gradual decrease with increasing frequency for both the P1–P2 and P3–P2 comparisons. As the frequency approached the upper limit of the analyzed spectrum, the relative differences tended to decrease toward zero, indicating that the electrical responses of the three phantoms became progressively more similar above 100 kHz.

For the Linear Arrangement 1 electrode configuration, the %RD between Phantoms P1 and P2 remained relatively low and nearly constant, decreasing slightly from approximately 30% to 20% over the analyzed frequency range. In contrast, the P3–P2 comparison exhibited a pronounced descending trend, with the %RD decreasing from approximately 140% at 200 Hz to values close to 20% at 100 kHz, where it converged toward the P1–P2 comparison.

The corresponding Bland–Altman plot clearly distinguishes the two phantom comparisons. The P1–P2 data points form a compact cluster below the bias, concentrated at lower mean impedance values and characterized by relatively small differences. In contrast, the P3–P2 comparison follows a marked ascending trend, with the differences increasing from approximately 0 Ω to values exceeding 15 kΩ. The calculated bias of 3062.96 Ω, together with the wide separation between the two clusters, indicates that the Linear Arrangement 1 configuration is highly sensitive to the modifications introduced in the middle layer resistivity, particularly for Phantom P3.

For the Linear Arrangement 2 electrode configuration, the %RD between Phantoms P1 and P2 remained relatively low, decreasing from approximately 17% to 2% over the analyzed frequency range. An even smaller relative difference was

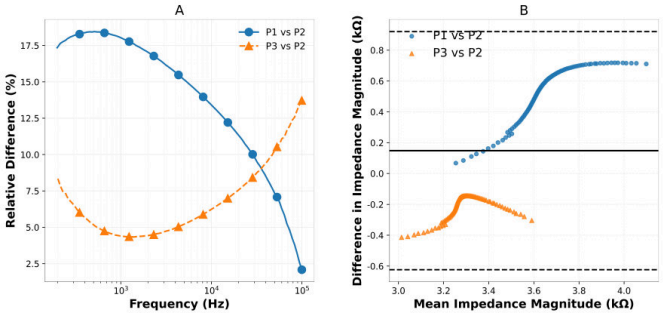


Fig. 12. (A) %RD in impedance magnitude for the P1–P2 and P3–P2 comparisons obtained with the Linear Arrangement 2. (B) Bland–Altman analysis of the corresponding measurements, with a mean bias of 148.28 Ω and limits of agreement defined as ±1.96 SD.

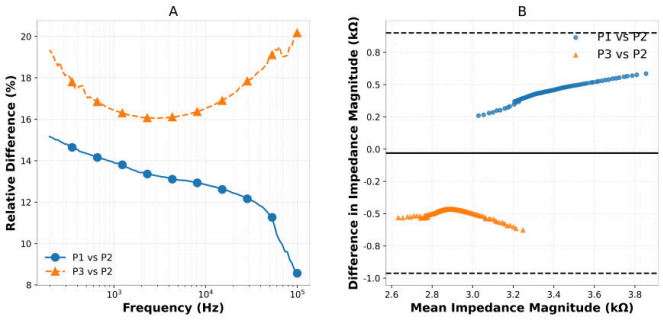


Fig. 13. (A) %RD in impedance magnitude for the P1–P2 and P3–P2 comparisons obtained with the Linear Arrangement 3. (B) Bland–Altman analysis of the corresponding measurements, with a mean bias of -30.95 Ω and limits of agreement defined as ±1.96 SD.

observed for the P3–P2 comparison, varying between approximately 8% and 14%. These results indicate that the Linear Arrangement 2 configuration exhibits a limited sensitivity to the variations in middle layer resistivity, particularly for the P3–P2 comparison.

The corresponding Bland–Altman plot shows a clear separation between the two phantom comparisons. The P3–P2 data are predominantly distributed below the zero-difference line and, consequently, below the calculated bias of 148.28 Ω, reflecting the lower impedance of Phantom P3 resulting from its higher middle layer conductivity. In contrast, the P1–P2 comparison is located almost entirely above the bias and is concentrated over mean impedance values ranging from approximately 3200 Ω to 4000 Ω. The distinct distribution of both clusters indicates that the Linear Arrangement 2 configuration consistently differentiates between increases and decreases in the middle layer resistivity with respect to the reference phantom.

For the Linear Arrangement 3 electrode configuration, the relative difference remained below 20% throughout the analyzed frequency range. The P1–P2 comparison exhibited a gradual decrease, tending toward zero at higher frequencies, whereas the P3–P2 comparison showed the opposite behavior, with the relative difference increasing above approximately 10 kHz. This divergence between the two comparisons at higher frequencies indicates that the Linear Arrangement 3 configuration is capable of distinguishing changes in middle layer resistivity, particularly in the upper portion of the analyzed spectrum.

The corresponding Bland–Altman plot yielded a bias of -30.95 Ω. Although the overall bias was close to zero, this result reflects the compensation between the positive differences observed for the P1–P2 comparison and the negative differences associated with the P3–P2 comparison, rather than the absence of systematic differences. The P1–P2 data form a nearly constant positive cluster, with differences ranging from approximately 250 Ω to 600 Ω, whereas the P3–P2 comparison remains consistently below the zero-difference line, with negative differences spanning approximately 150 Ω. The clear separation between the two clusters demonstrates that the Linear Arrangement 3 configuration consistently discriminates between increases and decreases in middle layer resistivity relative to the reference phantom.

IV. Discussion and Conclusions

The calibration procedure revealed substantial differences between agar and gelatin based formulations. Agar exhibited high intrinsic conductivity, which limited the attainable resistivity range, particularly for epidermis-mimicking layers. Increasing the agar concentration improved mechanical strength but further increased conductivity. In contrast, gelatin based formulations simultaneously provided adequate mechanical stability and the target resistivity range, making them suitable for multilayer phantom fabrication.

The experimental results demonstrated that electrode geometry strongly affects the impedance response of multilayer structures. Although all configurations distinguished the three phantoms, the magnitude and frequency dependence of the measured differences varied considerably. These variations indicate that the electrode arrangement modifies the current distribution and consequently the contribution of each layer to the measured impedance.

The coplanar configurations exhibited moderate sensitivity and progressive convergence of the impedance spectra with increasing frequency. In contrast, the linear arrangements showed markedly different performances. Linear Arrangement 1 produced the highest sensitivity to middle layer resistivity variations, exhibiting relative differences exceeding 100% at low frequencies. Linear Arrangements 2 and 3 displayed lower sensitivities but more stable responses across the analyzed spectrum.

The combined use of relative difference (%RD) and Bland-Altman analysis provided complementary information regarding the performance of each electrode configuration. The %RD quantified the sensitivity to changes in middle layer resistivity, whereas Bland-Altman analysis demonstrated that the observed differences were systematic and reproducible rather than arising from random variability.

The use of tissue mimicking phantoms limits direct extrapolation to biological tissue. Electrode skin contact, tissue heterogeneity, hydration, temperature fluctuations, and motion

were not evaluated. Complementary in vivo measurements using different electrode types were reported separately in [19]. The present study evaluated a defined set of electrode configurations without numerical simulations or geometry optimization.

From a physical perspective, the observed behavior can be attributed to the current density distribution imposed by each tetrapolar geometry. Small modifications in electrode spacing and arrangement produced substantial changes in sensing depth and sensitivity to subsurface electrical properties.

Overall, the proposed multilayer phantoms provided a controlled platform for assessing electrode configurations and variations in the electrical properties of subsurface mimicking skin layers. The methodology enables controlled modifications of individual layers for superficial bioimpedance measurements.

ACKNOWLEDGMENT

The authors acknowledge the Fundacao de Amparo á Pesquisa e Inovacao do Estado de Santa Catarina (FAPESC) for providing the scholarship that supported this study. The authors also acknowledge the Coordenacao de Aperfeicioamento de Pessoal de Nivel Superior (CAPES), Brazil, for the financial support for research supplies and materials

REFERENCES

[1] G. Swann, "The skin is the body's largest organ," *Journal of Visual Communication in Medicine*, vol. 33, no. 4, pp. 148–149, Dec. 2010.

[2] P. M. Elias and M. Steinhoff, "Epidermal barrier function and systemic diseases," *Clinical Reviews in Allergy and Immunology*, vol. 45, no. 2, pp. 217–238, Oct. 2013.

[3] E. Calonje *et al.*, *McKee's Pathology of the Skin*. Elsevier, 2012.

[4] C. G. Watts, A. E. Cust, S. W. Menzies, G. J. Mann, *et al.*, "Specialist surveillance for individuals at high risk of melanoma: A cost-effectiveness analysis," *Journal of Investigative Dermatology*, 2021.

[5] A. C. Aitken, P. H. Youl, D. R. English, M. A. Baade, and A. J. Elwood, "Clinical whole-body skin examination reduces the incidence of thick melanomas," *International Journal of Cancer*, vol. 126, no. 2, pp. 450–458, 2010.

[6] L. Gordon, C. Olsen, D. C. Whiteman, T. M. Elliott, M. Janda, and A. Green, "Prevention versus early detection for long-term control of melanoma and keratinocyte carcinomas: a cost-effectiveness modelling study," *BMJ Open*, vol. 9, no. 4, p. e034388, 2019.

[7] J. M. Jung, J. Y. Cho, W. J. Lee, S. E. Chang, M. W. Lee, and C. H. Won, "Emerging minimally invasive technologies for the detection of skin cancer," *Journal of Personalized Medicine*, vol. 11, no. 10, p. 951, Oct. 2021.

[8] P. McMullan, "Minimally invasive modalities for keratinocyte carcinomas," *Journal of the American Academy of Dermatology*, vol. 92, no. 4, pp. 1002–1004, 2025.

[9] R. P. Braun, J. Mangana, S. Goldinger, L. French, R. Dummer, and A. A. Marghoob, "Electrical impedance spectroscopy in skin cancer diagnosis," *Dermatologic Clinics*, vol. 35, no. 4, pp. 489–493, Oct. 2017.

[10] S. Gladstein, D. Damania, L. M. Almassalha, L. T. Smith, V. Gupta, H. Subramanian, D. K. Rex, H. K. Roy, and V. Backman, "Correlating colorectal cancer risk with field carcinogenesis progression using partial wave spectroscopic microscopy," *Cancer Medicine*, vol. 7, no. 5, pp. 2109–2120, May 2018.

[11] C. Liebich, J. N. Bartsch, I. Schubert, M.-L. von Brühl, and C. Sander, "Electrical Impedance Spectroscopy Improves Skin Cancer Detection and Reduces the Number of Biopsies," *Dermato*, vol. 2, no. 2, pp. 21–29, 2022.

[12] P. Mohr, U. Birgersson, C. Berking, C. Henderson, U. Trefzer, L. Kemeny, C. Sunderkötter, T. Dirschka, R. Motley, M. Frohm-Nilsson, U. Reinhold, C. Loquai, R. Braun, F. Nyberg, and J. Paoli, "Electrical impedance spectroscopy as a potential adjunct diagnostic tool for cutaneous melanoma," *Skin Research and Technology*, vol. 19, no. 2, pp. 75–83, 2013.

[13] T. J. C. Faes, H. A. van der Meij, J. C. de Munck, and R. M. Heethaar, "The electric resistivity of human tissues (100 Hz–10 MHz): a meta-analysis of review studies," *Physiological Measurement*, vol. 20, no. 4, pp. R1–R10, Nov. 1999.

[14] W. Montagna, A. M. Kligman, and K. S. Carlisle, *Atlas of Normal Human Skin*. New York, NY, USA: Springer, 1992.

[15] P. H. McKee, "Normal histology of the skin," in *Pathology of the Skin with Clinical Correlations*. Philadelphia, PA, USA: JB Lippincott Company, 1989, pp. 1–18.

[16] Ø. G. Martinsen and A. Heiskanen, *Bioimpedance and Bioelectricity Basics*, 4th ed. London, U.K.: Elsevier, 2023.

[17] S. Abasi, J. R. Aggas, G. G. Garayar-Leyva, B. K. Walther, and A. Guiseppi-Elie, "Bioelectrical impedance spectroscopy for monitoring mammalian cells and tissues under different frequency domains: A review," *ACS Measurement Science Au*, vol. 2, no. 6, pp. 495–516, 2022, doi: 10.1021/acsmeasuresciau.2c00033.

[18] S. Grimnes and Ø. G. Martinsen, "Sources of error in tetrapolar impedance measurements on biomaterials and other ionic conductors," *J. Phys. D: Appl. Phys.*, vol. 40, no. 1, pp. 9–14, 2006, doi: 10.1088/0022-3727/40/1/S02.

[19] E. Roman-Gomez and P. Bertemes-Filho, "Consistency measurements of skin impedance using different electrode types," in *Proc. IEEE Medical Measurements & Applications (MeMeA)*, 2026, pp. 1–5, doi: 10.1109/MeMeA69746.2026.11537313.

[20] A. Lo Presti, L. Tedesco, and M. F. Santoro, "Fundamentals of skin bioimpedances," *Advanced Materials*, vol. 35, no. 27, p. 2302127, 2023, doi: 10.1002/adma.202302127.

[21] A. Mirbeik-Sabzevari and N. Tavassolian, "Ultrawideband, stable normal and cancer skin tissue phantoms for millimeter-wave skin cancer imaging," *IEEE Trans. Biomed. Eng.*, vol. 66, pp. 176–186, 2019.

[22] M. Kavitha and M. R. Reddy, "Electrical impedance characterization of tissue-equivalent phantoms," in *Proc. IEEE Int. Conf. Imaging Systems and Techniques (IST)*, 2012, pp. 553–557.

[23] E. L. Madsen *et al.*, "Tissue mimicking materials for ultrasound phantoms," *Medical Physics*, no. 5, pp. 391–394, Sep. 1978.

[24] M. M. Burlew *et al.*, "A new ultrasound tissue-equivalent material," *Radiology*, no. 2, pp. 517–520, Feb. 1980.

[25] S. Ollmar and S. Grant, "Nevisense: Improving the accuracy of diagnosing melanoma," *Melanoma Management*, vol. 3, no. 2, pp. 93–96, 2016, doi: 10.2217/mmt-2015-0004.

[26] D. Dutra, A. M. R. Pinto, and P. Bertemes Filho, "Finite element modeling of gelatin phantom from measured impedance spectra," in *World Congress on Medical Physics and Biomedical Engineering*, Toronto, ON, Canada, 2015, vol. 1, doi: 10.13140/RG.2.1.2366.7049.

[27] D. Dutra, "Modelagem Numérica do Efeito de Aplicação de Força em Fantomas Biológicos Usando EIS," Ph.D. dissertation, Universidade do Estado de Santa Catarina (UDESC), Joinville, Brazil, 2018.

[28] B. Tsai *et al.*, "Dielectrical properties of living epidermis and dermis," *Journal of Electrical Bioimpedance*, vol. 10, pp. 14–23, 2019, doi: 10.2478/joeb-2019-0003.

[29] Y. Abe and M. Nishizawa, "Electrical aspects of skin as a pathway to engineering skin devices," *APL Bioengineering*, vol. 5, no. 4, p. 041509, 2021, doi: 10.1063/5.0064529.

[30] C. Tronstad and A. H. Pripp, "Statistical methods for bioimpedance analysis," *Journal of Electrical Bioimpedance*, vol. 5, pp. 1–27, 2014, doi: 10.5617/jeb.830.

Thermal Heating Evaluation Of Orthopedic Plates Implanted In Ultrasonic Phantoms Under Therapeutic Ultrasound

Lidio Toledo de Araujo Filho
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil.
0009-0001-9670-9318

Aldo José Fontes Pereira
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0003-3584-3342

Eduardo Moreno
Instituto de Cibernética
Matemática y Física-ICIMAF
La Habana, Cuba
0000-0003-3689-6163

Wagner Coelho de A. Pereira
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
O0000-0001-5880-3242

***Abstract*— Therapeutic ultrasound is a widely utilized modality in physical therapy, yet its safety in the presence of metallic orthopedic implants remains a subject of clinical debate. This study evaluates the thermal response of multilayer phantoms simulating fat, muscle, and bone tissues when subjected to 1 MHz continuous ultrasound at 1.0 W/cm² for 180 seconds. The methodology integrated experimental infrared thermography with numerical modelling using COMSOL® Multiphysics to analyze heat distribution across three groups: a control group (no plate), a stainless steel 316L plate group, and a titanium Ti-6Al-7Nb plate group. Results indicated that the control group exhibited the highest overall heating, reaching 56.31°C. Stainless steel plate demonstrated the lowest surface bone mean temperature (23.98°C), representing a minimal rise of approximately 1°C, while the titanium plate promoted a slightly surface bone mean temperature elevation (25.05°C).The COMSOL® simulation corroborated the experimental findings, highlighting how acoustic impedance mismatch and material properties influence energy redistribution. The study concludes that implant material significantly dictates thermal safety profiles, with stainless steel providing a more conservative thermal behavior under the tested conditions. Limitations are also considered.**

Keywords— therapeutic ultrasound, orthopedic implants, thermal effects, COMSOL® simulation, titanium, stainless steel.

I. Introduction

Therapeutic ultrasound (TUS) is extensively applied in clinical rehabilitation to promote tissue healing and pain relief through thermal and non-thermal mechanisms [1], [2]. However, the interaction between ultrasonic waves and metallic orthopedic implants introduces complexities regarding energy absorption and reflection at the tissue-implant interface. High acoustic impedance mismatches between metal and biological tissue can lead to localized hot spots, potentially causing thermal injury to peri-implant tissues [3].

To address these concerns, this research employs multilayer phantoms that mimic the acoustic and thermal properties of human tissue layers. A hybrid approach was adopted, combining experimental measurements via high-resolution thermography with computational modelling in COMSOL® Multiphysics. This dual methodology allows for a comprehensive visualization of the internal thermal fields that are otherwise inaccessible during experimental procedures.

This study specifically shows the thermal behavior of stainless steel (316L) and titanium alloy (Ti-6Al-7Nb) implants (Fig. 1). By establishing a clear thermal hierarchy between these materials and a control group, this work aims to provide evidence-based guidance for the safe application of ultrasound in patients with internal fixation devices.

.

II. Theoretical Foundations

The propagation of ultrasonic waves through heterogeneous media is governed by the acoustic impedance (Z) of each layer, which is the product of the material density (ρ) and the speed of sound (c) within that medium (eq. 1).

$$Z = \rho \times c \qquad (1)$$

When the wave encounters an interface between two media with impedances Z_1 and Z_2, a portion of the energy is reflected, according to the reflection coefficient R (eq. 2):

$$R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1}\right)^2 \qquad (2)$$

As the wave travels through a medium, its initial intensity ($I0$) decreases as a function of distance ($I(x)$) and attenuation coefficient α (eq. 3):

$$I = I_0 e^{-2\alpha x} \qquad (3)$$

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

III. Literature Review

The biophysical effects of ultrasound are well documented, with absorption being the primary driver of thermal elevation in protein-rich tissues like muscle and bone [1], [2]. Previous studies by Huang et al. [8] and Sun et al. [9] using simple models and hydrogel phantoms demonstrated that metallic inserts significantly alter the peak temperature location. Zawislak *et al*. [12] investigated titanium implants, suggesting that heating occurs, but may remain within safe limits under low-intensity protocols. However, data from different alloys in a multilayer physiological context remains scarce, needing advanced simulation tools to validate experimental trends.

Previous research also supports the use of phantoms as reliable tissue-mimicking materials for imaging and therapy studies [10]. The structural properties of these materials are essential for reproducing the acoustic and thermal conditions in vivo [3], [4], 6].

IV. Materials and Methods

A multilayer phantom was constructed using tissue-mimicking materials (TMM) to represent the human limb structure. The layers were measured precisely to ensure consistency across all experimental trials. The phantom configuration was based on previously established methodologies [9], [10]. The orthopedic plates in this study were manufactured by Depuy Synthes-Johnson®.

Three experimental configurations were considered: one composed of fat, muscle, and bone phantoms (control), another composed of the aforementioned layers with a TAN alloy plate (Titanium-1 Aluminum-6 Niobium-7) added between the muscle and bone layers, and another similar to the previous one with a stainless-steel plate (316L) (Fig.1).

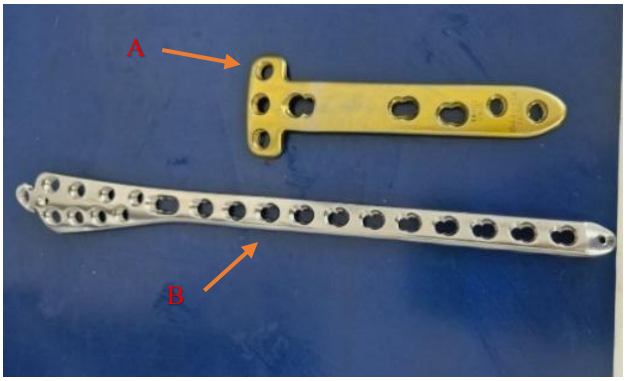


Fig. 1. A: Titanium plate; B: Stainless Steel
Source: the author.

For each experimental set-up, 30 repetitions were performed. The initial temperatures of the phantoms and plates were equal to the room temperature, which ranged between 23 and 24°C (as shown in Tables 1, 2, and 3). Thermography was obtained from the upper face of each of the layers.

The authors wish to thank the Brazilian agencies CNPQ, CAPES and FAPERJ for the financial support.

The following dimensions were used:

- Fat layer: 13.26 mm
- Muscle layer: 12.10 mm
- Bone layer: 1.27 mm
- Titanium plate (Ti-6Al-7Nb): 3.24 mm
- Stainless steel plate (316L): 4.15 mm

Therapeutic ultrasound (SONOPULSE III, IBRAMED, BRAZIL) was applied using a device at a frequency of 1 MHz with a spatial-average temporal-average intensity (I_{sata}) of 1.0 W/cm² in continuous mode for 180°s. The transducer was positioned perpendicularly to the phantom surface using a coupling gel (Fig 2). Each configuration receives 30 sessions of 180 s. The time between the sessions was 30 minutes.

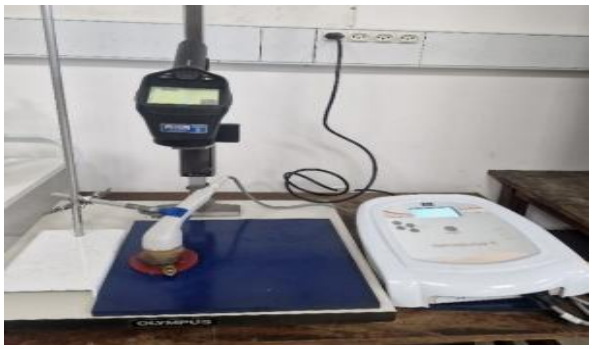


Fig. 2. Experimental Setup.
Source: the author.

Experimental data were captured using a FLIR® E40 infrared camera with 60 Hz (Fig.3). Acquisition rate, thermal range from -20°C to 650°C, sensitivity of 0.07°C, and emissivity adjusted to 0.98.

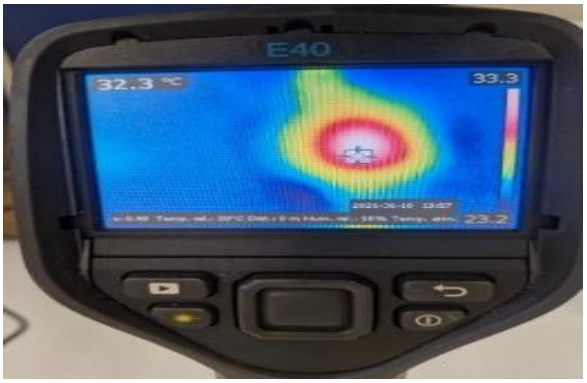


Fig. 3. FLIR® E40 infrared camera.
Source: the author.

In parallel, a 3D numerical model was developed in COMSOL® Multiphysics. In this simulation Helmholtz' and the Penne's bioheat equations are used to model the acoustic and heat propagations in a medium composed of fat/muscle/metal/bone (Fig. 4). The simulation presents a heat distribution map derived from ultrasonic absorption.

For the acoustic model, the radiation zone was defined using a Dirichlet condition in the area labeled "Transducer." The region outside each PML was assigned a free boundary or Neumann condition. For the Penne's model, a Dirichlet condition with a constant temperature of 24°C was applied, assumed as the lab. temperature. A λ/8 mesh was defined for

the acoustic case, assuming a 1 MHz frequency. The same mesh was used for the heating model. Acoustic parameters obtained from the COMSOL Library. Radiation pattern as the real experimental case.

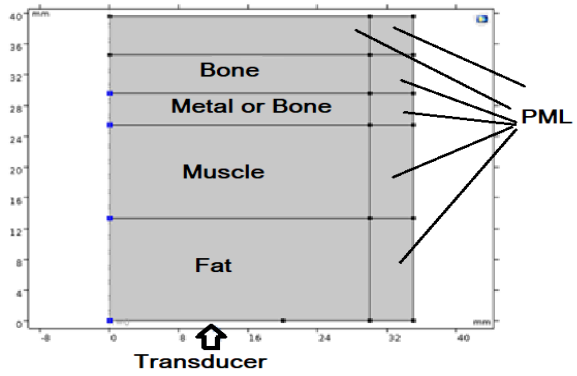


Fig. 4. COMSOL Model (PML – Perfec matching layer
Source: the author.

V. Result

A. Descriptive Statistics

The thermal response of the multilayer phantoms was recorded at the end of the 180-second ultrasound applications (1 MHz, 1.0 W/cm²). Tables 1, 2 and 3 summarize the mean (M), standard deviation (SD), and coefficient of variation (CV) for each experimental group. Figs 5, 6 and 7 show the temperatures reached for each layer in the respective experiment.

TABLE I. DESCRIPTIVE STATISTICS FOR THE CONTROL GROUP.

Interface	Mean [°C] +SD [°C]	CV [%]
Fat (G)	29.02 ± 1.18	4.05
Muscle (MU)	37.16 ± 2.98	8.02
Bone (O)	56.31 ± 5.84	10.36
Room (S)	23.84 ± 0.40	1.68

TABLE II. DESCRIPTIVE STATISTICS FOR THE TITANIUM PLATE GROUP (TP).

Interface	Mean [°C] ± SD [°C]	CV [%]
Plate (P)	25.46 ± 1.08	4.24
Fat (G)	30.10 ± 1.13	3.75
Muscle (MU)	30.53 ± 1.70	5.57
Bone (O)	25.05 ± 0.92	3.67
Room (S)	23.67 ± 0.26	1.10

TABLE III. DESCRIPTIVE STATISTICS FOR THE STAINLESS-STEEL PLATE GROUP (SP).

Interface	Mean [°C] ± SD [°C]	CV [%]
Plate (P)	24.55 ± 2.25	9.16
Fat (G)	27.41 ± 2.40	8.87
Muscle (MU)	28.39 ± 2.32	8.17
Bone (O)	23.98 ± 2.34	9.76
Room (S)	23.12 ± 1.96	8.48

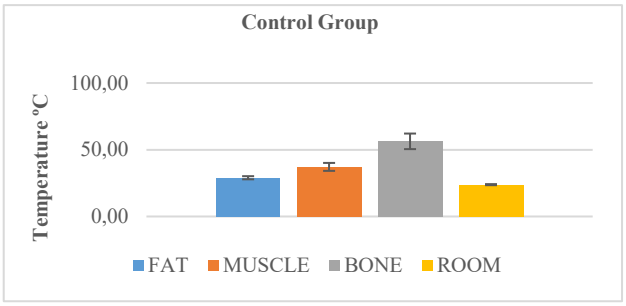


Fig. 5. Layers' temperatures for the Control Group.
Source: the author.

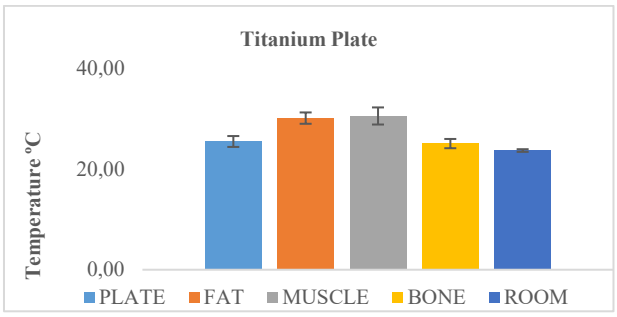


Fig. 6. Layers' temperatures for the Titanium Plate Group
Source: the author.

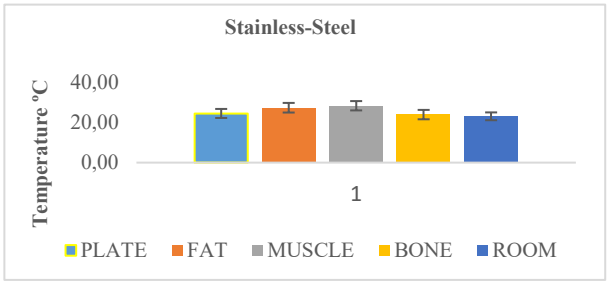


Fig. 7. Layers' temperatures for the Stainless Sterl Group.
Source: the author.

B. Thermal Dynamics Analysis

The temperature rise (ΔT) was significantly higher in the Control Group, where the bone layer reached a critical mean of 56.31°C. In contrast, the presence of metallic plates acted as a thermal heatsink. For the Titanium Group (PT), the bone temperature remained at 25.05°C, representing a minimal rise relative to the initial state. The Stainless-Steel Group (SP) showed the lowest bone temperature at 23.98°C.

C. Inferential Analysis

A Student's T-test was performed to compare bone temperatures between the control group and each of the two groups with metallic implants. The difference between the Control Group and each of implanted groups (TP and SP) was statistically significant ($p < 0.001$).

Regarding the use of the t-test, what is clinically relevant is to determine whether there is a statistical difference between the normal physiological condition (without implant) and the condition with implant. Therefore, the comparison between the set-ups with implants was removed,

and the t-test was maintained for the aforementioned conditions.

D. COMSOL® Simulation Results

The numerical data from COMSOL® Multiphysics corroborated the experimental findings. In the Control Group simulation (Fig.8), the bone surface reached 56 °C at 180 s. For the Titanium Plate (Fig.9), the simulation showed a bone surface temperature of approximately 26.5°C with a rise of 2.5°C. The Steel Plate simulation (Fig.10) showed a similar pattern but with a slightly lower rise (ΔT 1°C) in the metal itself, while keeping the bone surface at 24°C.

The interface profiles in the simulation revealed a gradient where the transducer/fat interface reached 32.5°C, decreasing through the muscle (31°C) to the bone (25°C).

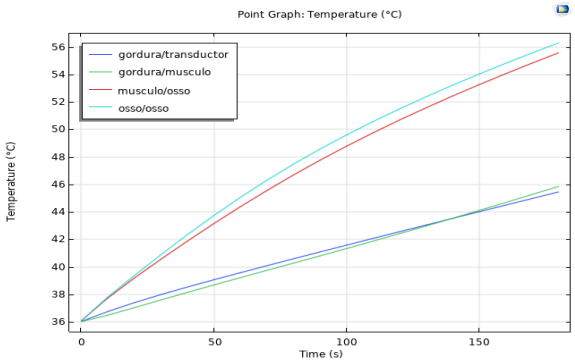


Fig. 8. Control Group COMSOL® Simulation.
Source: the author.

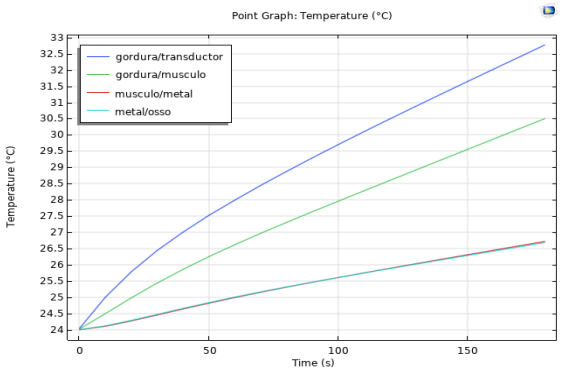


Fig. 9. Titanium COMSOL® Simulation
Source: the author.

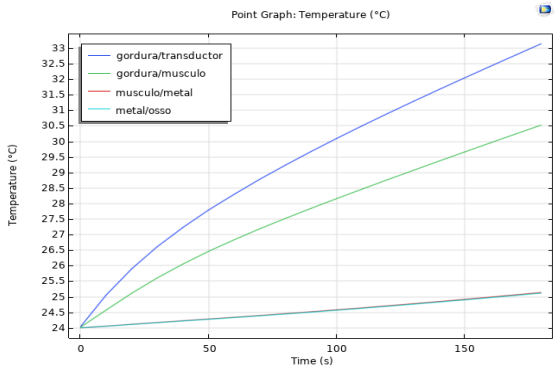


Fig. 10. Stainless Steel COMSOL® Simulation.
Source: the author.

VI. Discussion

A. Thermal Behavior Interpretation

The most significant finding is the "heat sink" effect provided by the metallic implants. In the Control Group, the acoustic energy is highly absorbed by the bone phantom due to its high absorption coefficient, leading to potentially necrotic temperatures (> 50°C). However, the metallic plates, possessing high thermal conductivity (k), rapidly redistribute the absorbed energy. The acoustic impedance mismatch (Z) at the muscle-metal interface also causes a portion of the wave to reflect, reducing the direct energy deposition on the bone.

B. Material Comparison

Stainless-steel (SP) and Titanium (TP) exhibited distinct thermal profiles. Although Titanium is often preferred for its biocompatibility, the Stainless-Steel plate resulted in a lower bone temperature (23.98°C vs 25.05°C). This is attributed to the higher thermal diffusivity of the steel alloy used, which allowed for faster heat dissipation into the surrounding soft tissue phantoms, as evidenced by the higher CV values in the SP group, indicating more dynamic thermal fluctuations.

C. Interface Analysis

The temperature distribution across interfaces was non-uniform. In the TP group, the muscle reached 30.53°C, which was higher than the bone (25.05°C). This suggests that the metal-bone interface is protected by the plate's ability to conduct heat laterally and back toward the muscle layer. The COMSOL® profiles confirmed this, showing that the highest thermal stress occurs at the fat/muscle and muscle/metal boundaries rather than the bone itself.

D. Clinical Implications

The results *in vitro* seem to challenge the traditional contraindication of therapeutic ultrasound over metallic implants. Under the parameters of 1 MHz and 1.0 W/cm², the bone temperature remained far below the threshold for thermal damage (45°C). The presence of the implant actually protected the bone from the excessive heating observed in the control group. However, the increased temperature in the muscle and fat layers (reaching ~30°C) suggests that clinicians should monitor soft tissue sensitivity rather than bone pain.

E. COMSOL® Validation

The good degree of correlation between the experimental average values and the COMSOL® numerical curves validates the multilayer phantom model. The simulation's ability to predict the ~1°C rise for Titanium(Fig. 11) and the ~3°C rise for Steel (Fig. 12) confirms that the physical properties (density, specific heat, and conductivity) assigned to the materials accurately represent the experimental reality. While the Control Group exhibited a continuous linear rise, the implanted groups reached a thermal plateau, suggesting efficient heat dissipation through the metallic medium.

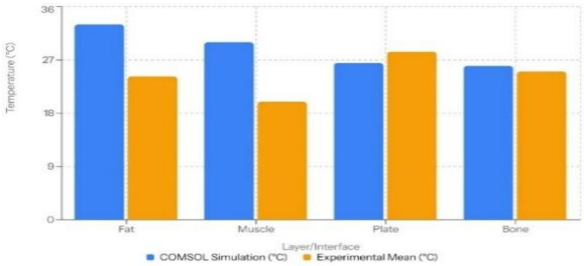


Fig. 11. Titanium Plate COMSOL® Simulation x Experimental Mean.
Source: the author.

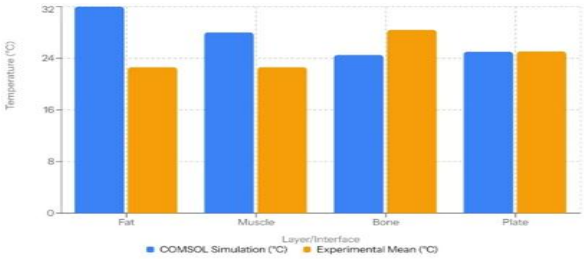


Fig. 12. Stainless Steel Plate: COMSOL® Simulation x Experimental Mean.
Source: the author.

At this point, it is important to note that the objective of the experiments was to simulate conditions close to the real case. Thus, the phantom thicknesses are similar to the real case, and the implants are of the same type and size as those used in practice. Therefore, the behaviour of the implants is expected to approximate the real case. In the simulation, however, the implants have the same size and thickness, which allowed us to observe that the heating at the metal-bone interface was much lower than at the muscle-bone interface in the control case.

VII. Conclusion

This study evaluated the thermal effects of therapeutic ultrasound on phantoms with orthopedic implants. The main findings are:

1. Safety Hierarchy: The Control Group reached dangerous temperatures (56.31°C), while Titanium (25.05°C) and Stainless-Steel (23.98°C) remained safe.

2. Heat Sink Effect: Metallic plates effectively protect bone tissue by dissipating thermal energy and reflecting acoustic waves.

3. Material Performance: Stainless steel provided a slightly higher cooling effect for the bone layer compared to titanium under the tested conditions. Maybe because it is a larger piece.

4. Clinical Recommendation: The results suggest that therapeutic ultrasound at 1.0 W/cm² appears safe for patients with internal fixators, provided the transducer is moved continuously to avoid soft tissue hotspots. Complementary experiments *in vivo* are needed.

Limitations include the use of static phantoms without blood perfusion. Future research should incorporate vascular cooling simulations and evaluate pulsed ultrasound modes to further refine safety protocols. Smaller implant sizes can have a different behavior and should be tested to see if the temperature has similar comportment.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

[1] P. J. Fish, *Physics and instrumentation of diagnostic medical ultrasound*, 1st ed. New York: John Wiley & Sons, 1990.
[2] K. G. Baker, V. J. Robertson, and F. A. Duck, "A review of therapeutic ultrasound: biophysical effects," *Physical Therapy*, vol. 81, pp. 1351–1358, 2001.
[3] L. E. Maggi, "Characterization of acoustic properties of new materials for the fabrication of ultrasonic phantoms," Ph.D. dissertation, COPPE/UFRJ, 2011. *In Portuguese*
[4] D. P. Oliveira, "Study of the ultrasound thermal field in a phantom of the middle third of the arm with and without metallic implants," M.S. thesis, COPPE/UFRJ, 2013. *In Portuguese*
[5] C. K. B. F. Nogueira, "Thermal study at the muscle-bone interface in specimens and ex vivo swine tissue in the presence of a metallic implant after therapeutic ultrasound application," M.S. thesis, Universidade Federal do Acre, 2021. *In Portuguese*
[6] R. B. Alves, "Thermographic images of the heating pattern in a triple-layer phantom irradiated by therapeutic ultrasound at 1 and 3 MHz," M.S. thesis, COPPE/UFRJ, 2013. *In Portuguese*
[7] J. D. N. Cheeke, *Fundamentals and Applications of Ultrasonic Waves*, 2nd ed. Boca Raton, FL: CRC Press, 2012.
[8] C. W. Huang, "Simplified theoretical model for temperature evaluation in tissue–implant–bone systems during ultrasound diathermy," *Applied Sciences*, vol. 10, p. 1306, 2020.
[9] M. K. Sun et al., "Effects of an implant on temperature distribution in tissue during ultrasound diathermy," *Ultrasonics Sonochemistry*, vol. 32, pp. 44–53, 2016.
[10] C. K. McGarry et al., "Tissue mimicking materials for imaging and therapy phantoms: a review," *Physics in Medicine & Biology*, vol. 65, p. 23TR01, 2020.
[11] C. L. Simplício et al., "Extracorporeal shock wave therapy mechanisms in musculoskeletal regenerative medicine," *Journal of Clinical Orthopaedics and Trauma*, vol. 11, pp. S309–S318, 2020.
[12] T. Zawislak et al., "The impact of ultrasound waves on changes in temperature around titanium orthopaedic implants – experimental studies," *European Journal of Medical Technologies*, vol. 2, no. 11, pp. 28–36, 2016.
[13] DePuy Synthes, "TOMOFIX® Osteotomy System Surgical Technique," 2014.
[14] H.-H. Huang, C.-P. Wu, Y.-S. Sun, and T.-H. Lee, "Improvements in the corrosion resistance and biocompatibility of biomedical Ti–6Al–7Nb alloy using an electrochemical anodization treatment," *Thin Solid Films*, vol. 528, pp. 157–162, 2013.

Development of a Universal Device for Photobiomodulation in Dairy Cows

1st Lívia Helena Moreira
Dept. Biomedical Engineering
Anhembi Morumbi University - UAM
São José dos Campos, Brazil
lh.medicinaveterinaria@gmail.com

2nd Maycon Crispim Carvalho
Dept. Biomedical Engineering
Anhembi Morumbi University - UAM
São José dos Campos, Brazil
mayconcarvalho20@gmail.com

3rd Henrique Cunha Carvalho
Dept. Electronic Engineering
Federal Technology University of Paraná - UTFPR
Campo Mourão, Brazil
hccarvalho@utfpr.edu.br

4st Renato Amaro Zângaro
Dept. Biomedical Engineering
Anhembi Morumbi Universit - UAM
São José dos Campos, Brazil
razangaro@gmail.com

5st Leandro de Lima Azevedo
Dept. Biomedical Engineering
Anhembi Morumbi University - UAM
São José dos Campos, Brazil
leandrolazevedo@gmail.com

6st Carlos José de Lima
Dept. Biomedical Engineering
Anhembi Morumbi University - UAM
São José dos Campos, Brazil
cdcflima@gmail.com

***Abstract*— Mastitis is the primary disease affecting dairy cows, causing significant economic losses due to reduced production and a decline in the quality of milk intended for human consumption. Among conventional therapies, antibiotics are widely recommended but can foster the emergence of resistant strains. This study aimed to develop and evaluate a universal photobiomodulation (PBM) prototype for coupling with commercial milking equipment. The device was designed using modeling software and fabricated via 3D printing; it was then deployed on a farm with Jersey cows, which were mechanically milked twice daily. Over three consecutive weeks, the front teats of each cow were subjected to PBM through an energy density (ED) of 2.76 J/cm². Milk samples were collected weekly before, during, and after the intervention for quality analysis. The results demonstrated that the prototype was incorporated into the milking system without compromising the seal, enabling safe PBM application without altering milk quality parameters.**

Keywords— Cow, mechanical milking, photobiomodulation, prevention.

I. Introduction

Photobiomodulation (PBM) is a non-invasive technique that uses red or near-infrared light to stimulate healing, alleviate pain, and reduce inflammatory processes. In veterinary medicine, PBM has been applied in various fields, including rehabilitation, sports medicine, and general practice [1]. In dairy animals, studies report its use for the treatment and prevention of teat hyperkeratosis [2] and mastitis [3–7].

Among the most significant diseases in dairy farming, mastitis directly compromises productivity, milk quality, and farm profitability. Beyond the economic losses associated with reduced production and the disposal of contaminated milk, the disease poses a public health risk due to antibiotic residues and potential pathogens in the final product. Therefore, mastitis control and prevention are essential to ensuring the sustainability, competitiveness, and food safety of the dairy supply chain [8,9].

Subclinical mastitis, which is highly prevalent in dairy herds, represents the most concerning form of the disease; it presents no visible clinical signs and manifests only through changes in milk composition, making its monitoring and control indispensable [10,11].

Despite the availability of portable PBM equipment, one of the major challenges is adapting it to veterinary routines, particularly for large animals, to prevent mammary gland diseases [12]. This is because the devices must be used for extended periods, multiple times a day, in alignment with farm milking schedules, which often leads to premature equipment fatigue.

Given this context, this study aimed to develop a PBM prototype adaptable to commercial milking equipment, enabling its daily use on dairy farms.

II. Material and Methods

A. Device setup

To develop a PBM prototype compatible with commercially available mechanical milking systems, Computer-Aided Design (CAD) software (v. 2025, Autodesk Inc., Mill Valley, CA, USA) was used for 2D and 3D design. The optoelectronic system was fabricated using a 3D printer (Replicator Z18, MakerBot Print, China) and black polylactic acid (PLA) filament.

B. Device and PBM beam parameters

Four flanges were produced based on the 3D-printed PBM optical system. The device has a diameter of 9 cm and a height of 2.4 cm. Near-infrared (940 nm) LEDs (5 mm diameter package, GaAlAs, L-51AEIR1BC, China) were installed on each flange, totaling 50 units geometrically distributed over a circular active area of 39 cm² (Fig. 1a and b).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

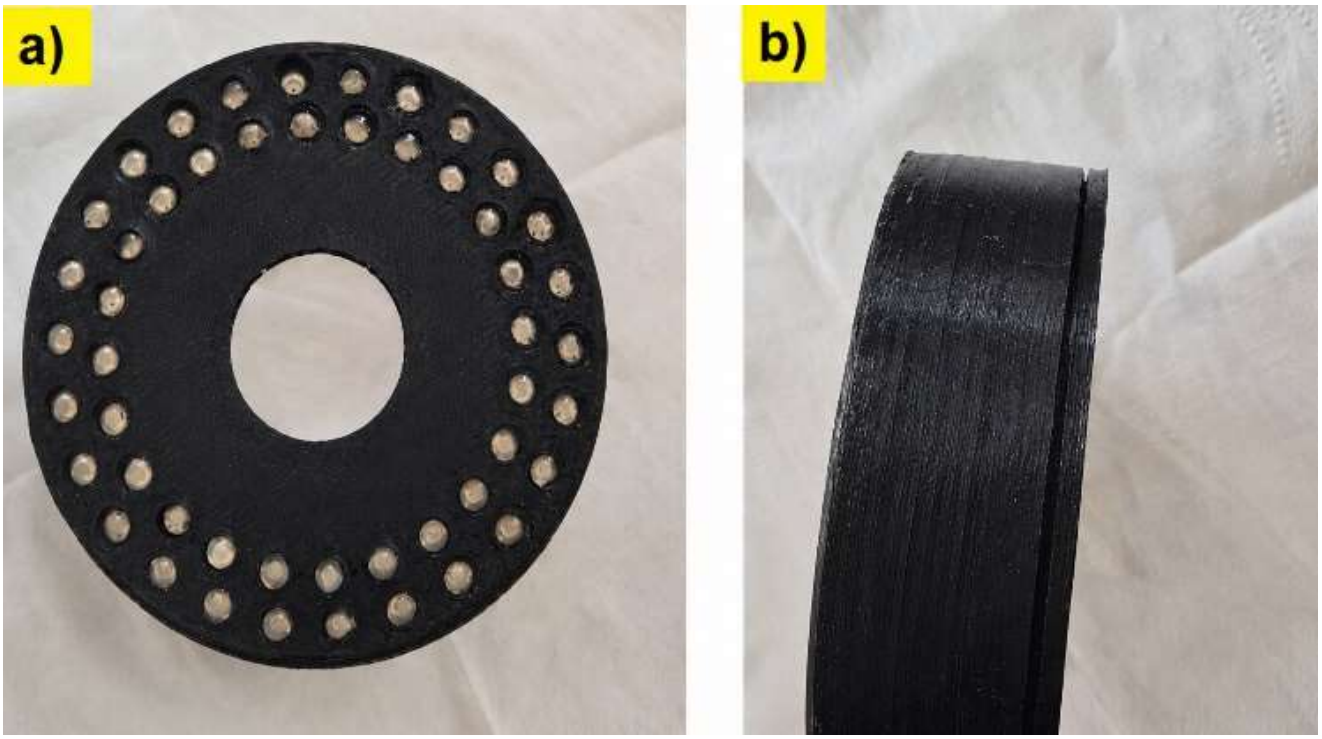


Fig. 1: PBM prototype developed using a 3D printer; front (a) and side (b) views with the near-infrared LEDs.

Each LED-equipped flange emitted a total optical power of 0.9 W, resulting in a power density of 23 mW/cm². The total incident optical energy, considering the four LED flanges (Fig. 2 – a), which are to be individually coupled (Fig. 2 – b) to each of the animal's four teats, corresponds to a radiant energy per session of approximately 650 J.

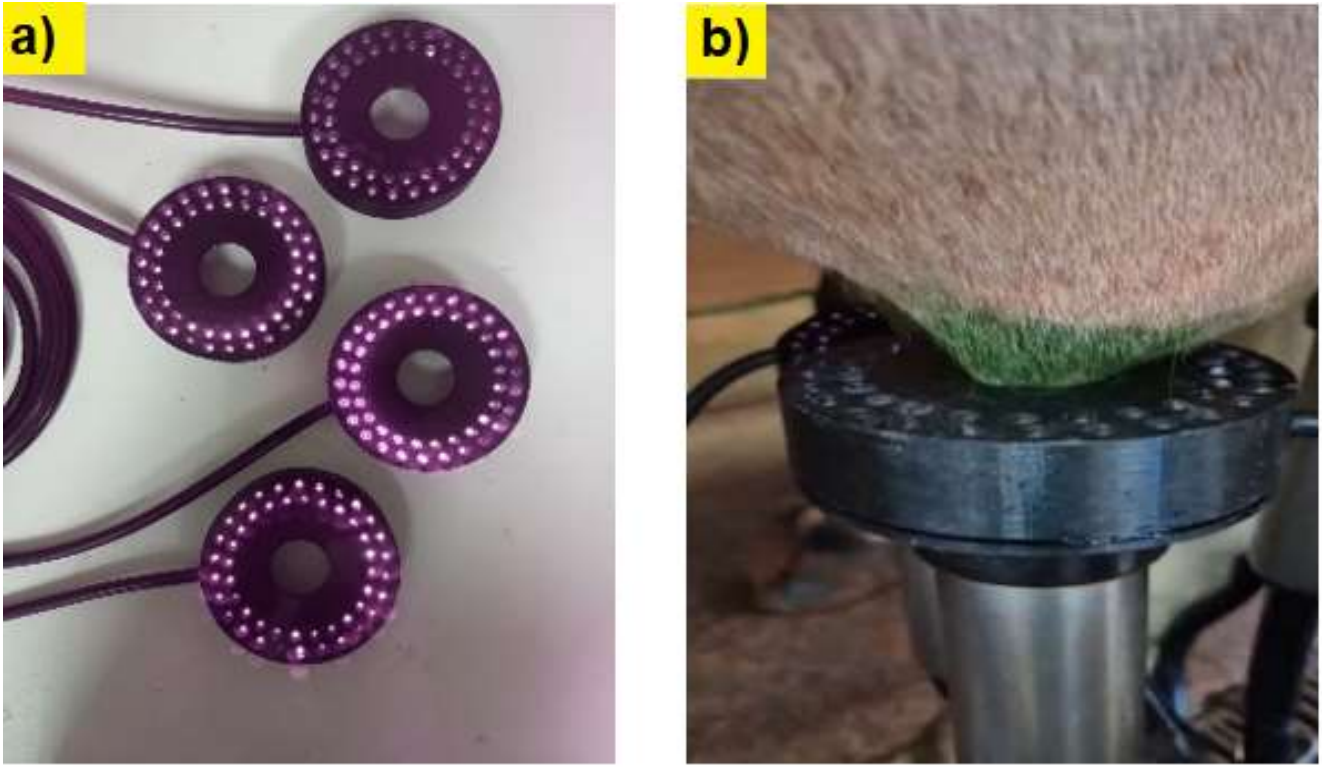


Fig. 2: Prototype assembled with LEDs (a) and coupled to the mechanical milking equipment (b) in operation.

C. Experimental setup

Jersey cows healthy (n=4) and without mastitis (n=14 functional teats), were mechanically milked twice daily at the Universidade Estadual de Londrina (UEL) Teaching-Farm, received PBM following project ethical approval (CEUA No. 031.2023).

The milking system used at UEL Farm is a compact system (mod. P100, DeLaval Ltda., Jaguariúna-SP, Brazil) with an in-line parlor, allowing for the simultaneous milking of two animals. Vacuum calibrations and the replacement of silicone parts are carried out every six months, in accordance with the manufacturer's recommendations.

System cleaning and sanitization take place twice daily, following these steps:

- Pre-milking: closed-circuit cleaning for 10 minutes using a commercial sanitizing product diluted in 40 L of water at 70°C, followed by two rinses with water at ambient temperature.
- Post-milking: closed-circuit cleaning for 10 minutes using an alkaline detergent diluted in 40 L of water at 70°C, followed by two 5-minute rinses with water at ambient temperature.

Additionally, a more intensive cleaning process is performed once a week during the post-milking phase. In this step, a commercial acid detergent is used prior to the alkaline detergent, followed by two 5-minute rinses with water heated to 70°C, after which the farm's standard routine is resumed.

The protocol spanned three consecutive weeks, including every milking session during that period. The optoelectronic PBM system was integrated into the milking equipment; PBM was applied only to the front teats of each cow, while the rear teats served as internal controls.

To evaluate milk quality, three time points were compared: before the intervention (time zero / control – Pre), during the intervention (three consecutive weeks, without interruption), and after the intervention (two weeks without the optical system in use (Post)). An ED of 2.76 J/cm² was evaluated during every milking session throughout the three weeks, as shown in Table 1.

TABLE 1: EXPERIMENTAL PBM PROTOCOL CONDUCTED ON THE COWS MILKED.

Stage	Experimental stage setup
Pre-PBM	Milk samples *(n=14)* were collected from each teat of every animal for initial monitoring.
During-PBM	PBM was applied for 2 min in each session and three consecutive weeks during every milking session (twice daily, following the farm's routine) at an ED of 2.76 J/cm². Milk samples *(n=34)* were collected weekly from each teat of every animal for monitoring.
Post-PBM	The cows went two consecutive weeks without PBM light stimulation while being milked according to the farm's routine; milk samples were collected weekly for assessment. Milk samples *(n=28)* were collected from each teat of every animal for the final monitoring stage.

D. Milk laboratory analysis

Milk samples were collected weekly after cleaning the teats, and before milking (n=76*)*. The samples were collected from each cow's teat, stored in sterile vials, refrigerated for 24 hours, and packed in isothermal boxes with recyclable ice and shipped to the official laboratory of the Ministério da Agricultura e Pecuária (MAPA) in Curitiba-PR, Brazil.

Milk samples were analyzed for content/composition (protein, fat, lactose, total solids, % w/w), Somatic Cell Count (SCC, mil/mL), and Total Bacterial Count (TBC, 10^3 CFU/mL).

E. Milk statistical analysis

Milk data were analyzed using GraphPad Instat software (v. 3.0; GraphPad Software Inc., San Diego, CA, USA). Differences in milk parameters measured before (Pre-PBM), during (During-PBM), and after (Post-PBM) PBM were analyzed using a one-way repeated-measures ANOVA with Dunnett's multiple-comparison test, with Pre-PBM values as the control. Statistical significance was set at $p < 0.05$.

III. Results and Discussion

No signals of discomfort, restlessness, irritability, or burns were observed on the teats during or after the protocol, indicating that the device is potentially safe and may be applied even in dairy animals, which corroborates previous studies [4,6,7,12] demonstrating the PBM potential in veterinary medicine, particularly in dairy cows.

Paudyal *et al.* (2026) [13] highlight in their review that PBM presents a ‘therapeutic window’ between 600 and 1100 nm (red and near-infrared). The authors emphasize that optical dosage must account for penetration depth and animal size; specifically, cattle require higher energy levels than smaller animals, underscoring the need to make specific adjustments to commercial devices for veterinary use.

Adapting equipment originally developed for human use to large-animal use presents a significant challenge. Key barriers include the need to adjust optical dosages, compatibility issues between different types of optical fibers, and limitations regarding applicator probes, which often fail to meet the demands of veterinary routines involving production animals. In many cases, these difficulties render the technique impractical, underscoring the need to develop specialized devices that are suitable for daily use and focused on disease prevention [5, 6, 12].

In previous studies evaluating the application of PBM in dairy cows [5, 6, 12], adapted equipment exhibited several limitations, such as poor durability under continuous handling, ineffective sealing in humid environments, and an inability to accommodate teats with anatomical malformations or greater thickness; additionally, vacuum loss during milking frequently caused the device to fall to the ground. These issues were overcome in the present study, as the photonic device, externally attached to universal-type milking equipment, proved compatible with any commercially available system. It

did not compromise the vacuum and ensured animal comfort during milking, making it an efficient and safe solution.

Regarding the optical dosages applied, Moreira *et al.* (2022, 2024) [5, 6] tested energy levels of 2.5 and 5 J/cm², while Valongo *et al.* (2024) [12] evaluated a dose of 10 J/cm² in dairy cows. In the present study, a similar ED was applied; these dosages had been previously investigated in other studies, with no reports of side effects during short-term application periods. Unlike those studies, the applications here were performed twice daily for 3 consecutive weeks, totaling 42 sessions per optical dosage, allowing evaluation of safety and efficacy under a prolonged-use regimen.

Favaretto *et al.* (2024) [14], in a study on clinical mastitis in cows, applied laser at wavelengths of 808 nm (1.4 W) and 637 nm (1 W), corresponding to power densities of 1.4 and 1 W/cm², respectively, with irradiation of 2 minutes per wavelength for 5 consecutive days. The authors demonstrated that the combination of these options was safe and effective as an adjuvant therapy to antibiotic therapy. Complementarily, Lage *et al.* (2014) [2] used a red-infrared LED device applied three times a week for six weeks and observed that, although it did not reverse the hyperkeratosis lesions of the cows' teats, it caused the incidence of subclinical mastitis (SCC > 250 mil cells/mL). Overall, these results suggest that even relatively low doses of LED can modulate inflammatory processes and prevent the progression of teat lesions, highlighting the potential of PBM for mammary gland health.

Regarding the assessment of milk quality, the results showed that PBM applied using 940 nm infrared LEDs at an ED of 2.76 J/cm² resulted in no statistically significant differences ($p > 0.05$) in milk quality between treated and control teats regarding protein, fat, and lactose content, total solids, SCC, and TBC. Thus, milk quality was maintained throughout and remained within current national regulatory standards for human consumption [15, 16], as illustrated in Figure 3.

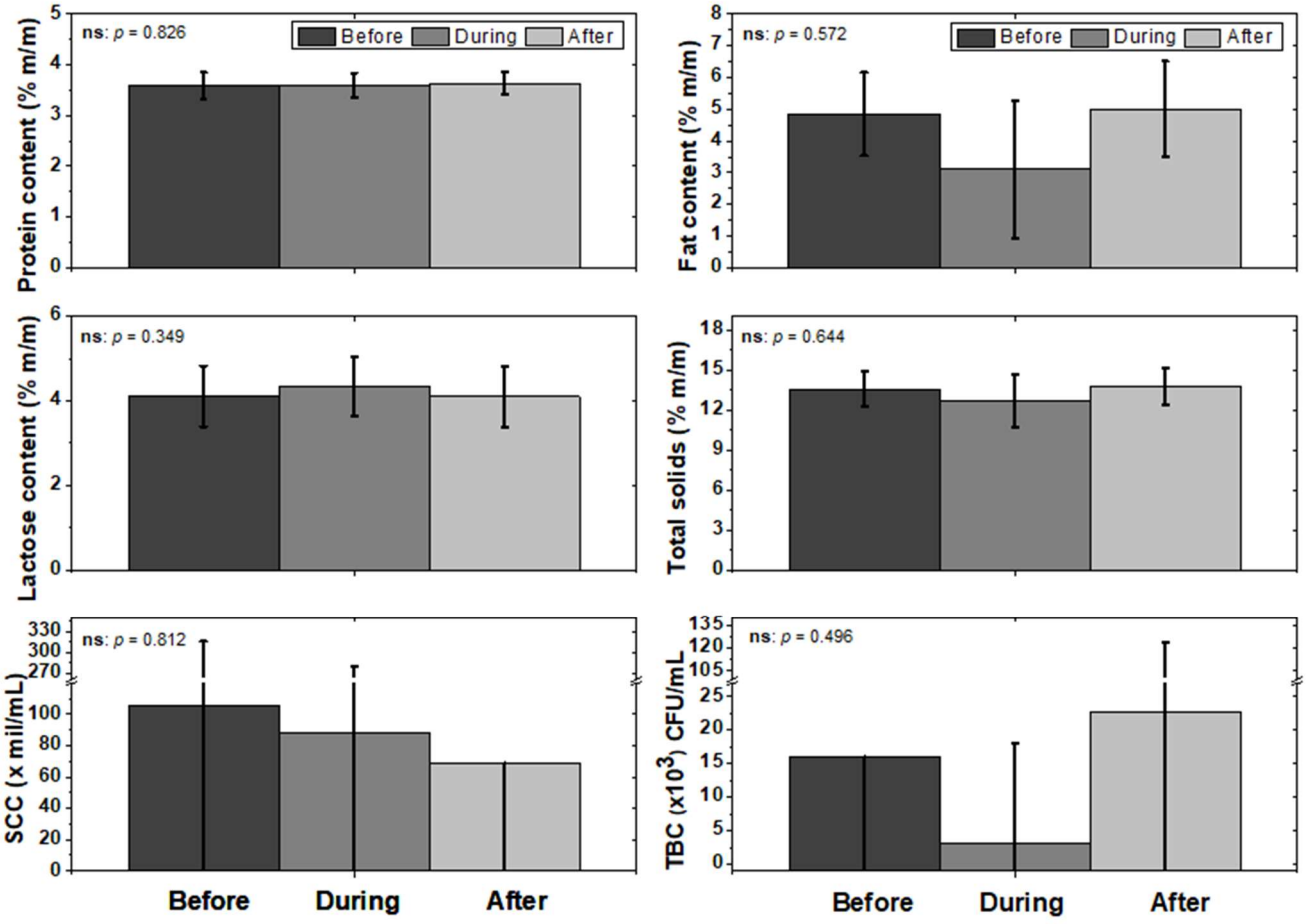


Fig. 3: Evaluation before, during and after PBM of milk quality with daily application of an energy density of 2.76 J/cm². The PBM device was applied twice daily for 21 consecutive days, for a total of 42 sessions. Data expressed as Mean and Standard Deviation (SD) values. ns: non-significant.

These findings reaffirm that PBM can be safely employed without compromising the quality of milk intended for human consumption, corroborating evidence from previous studies [5, 6, 12-14].

The cows included in this preliminary study were healthy, with no cases of clinical or subclinical mastitis, and produced milk within the parameters recommended for human consumption. The use of the PBM device, both during and after PBM, did not cause teat lesions that could predispose the animals to subclinical mastitis in the event of aseptic failure, as described in other studies [8,10,11].

The data presented reinforce that PBM is a sustainability-aligned strategy that can reduce antibiotic reliance and promote animal welfare, thereby contributing to mastitis

prevention in dairy cows. However, further studies are needed, considering different milking practices, incorporating optimized protocols with longer application periods, and including larger sample sizes aiming to consolidate and expand the applicability of this technology.

IV – CONCLUSION

A universal 3D-printed PBM device compatible with commercial milking systems was successfully developed and validated. The prototype maintained vacuum integrity during routine milking, resisted daily sanitation procedures, and enabled safe PBM application without affecting milk quality. These findings support the feasibility of integrating PBM into routine dairy farming as a practical strategy for mastitis prevention.

ACKNOWLEDGEMENT

L. H. M., R. A. Z., and C. J. L. would like to thank Instituto Ânima (IA) for the research grant. H. C. C. thanks Fundação Araucária (FA) & Superintendência de Ciências, Tecnologia e Ensino Superior (SETI) for the conference grant.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] Bunch J. “Photobiomodulation (therapeutic lasers): An update and review of current literature. Vet Clin North Am Small Anim Pract. 53(4), pp. 783-799, 2023. https://doi.org/10.1016/j.cvsm.2023.02.010

[2] Lage P. G., Araújo A. R., Teixeira A. G., Pinotti M. & Faleiros R.R. “Photobiomodulation device for prevention and treatment of teat hyperkeratosis in dairy cows”. Pesq Vet Bras. 34(6), pp. 515-522, 2014.

[3] Moreira, L. H. S., Moretti, A. B. F., Lima, C. J., et al. “Efeitos da aplicação intra-mamáriano tratamento de mastite em bovinos utilizando a ozonioterapia”. In: XXIV Congresso Brasileiro de Engenharia Biomédica, CBEB, Uberlândia, pp. 2143–2146, 2014.

[4] Moreira, L. H., de Souza, J. C. P., de Lima, C. J. et al. “Use of photodynamic therapy in the treatment of bovine subclinical mastitis”. Photodiagnosis Photodyn Ther 21, pp. 246–251, 2018. https://doi.org/10.1016/j.pdpdt.2017.12.009

[5] Moreira, L. H., Lima, C. J., Azevedo, L. et al. “Desenvolvimento do dispositivo de fotobioestimulação para prevenção e tratamento de mastite em vacas lactantes”. In: XXVII Congresso Brasileiro de Engenharia Biomédica, pp. 1648–1649, Vitória, 2020.

[6] Moreira, L. H., Carvalho, H. C., da Silva Mendes, et al.: “Photobiomodulation in the treatment of bovine subclinical mastitis”. In: Marques, J. L. B., Rodrigues, C. R., Suzuki, D. O. H., Marino Neto, J., García Ojeda, R. (eds.) IX Latin American Congress on Biomedical Engineering and XXVIII Brazilian Congress on Biomedical Engineering. CLAIB CBEB 2022. IFMBE Proceedings, vol 99. pp. 624–631. Springer, Cham, 2024. https://doi.org/10.1007/978-3-031-49404-8_60

[7] Silva, L. O., Souza, K. L., Beloti, L., et al. “Use of photodynamic therapy and photobiomodulation as alternatives for microbial control on clinical and subclinical mastitis in sheep”. Lasers Med. Sci. 37, pp. 2305–2310, 2022. https://doi.org/10.1007/s10103-022-03506-2.

[8] Eleodoro, J. I., Muraga, L., Vanot, R. L., Fagnani, R. “Identification and antimicrobial susceptibility of milk isolates from cows with subclinical mastitis in the northwest of Paraná State, Brazil”. Vet Ital. 59(1), 2023. https://doi.org/10.12834/VetIt.2606.16215.1

[9] FAO. Dairy Market Review – Emerging trends and outlook in 2023. Rome, 2023. Available at: https://openknowledge.fao.org/items/011f4397-be2b-40bb-a84b-c620171f8e7e Access: 09/20/2026.

[10] de Oliveira R. P., da Silva J. G., Aragão B.B., et al. “Diversity and emergence of multi-resistant *Staphylococcus* spp. isolated from subclinical mastitis in cows in of the state of Piauí, Brazil”. Braz J Microbiol. 53(4), pp. 2215-2222, 2022. https://doi.org/10.1007/s42770-022-00822-1.

[11] Gonçalves J. L., Kamphuis C., Vernooij H., et al. “Pathogen effects on milk yield and composition in chronic subclinical mastitis in dairy cows”. Vet J. 262, pp. 105473, 2020. https://doi.org/10.1016/j.tvjl.2020.105473

[12] Valongo, V., Fagnani, R., Muniz, C. A. S. D., et al. Photobiomodulation by infrared LEDs in a milking machine: Preliminary *in vivo* study”. In: Soares, A.B., Cardoso, G.C., Leoni, R.F. (eds) XXIX Brazilian Congress on Biomedical Engineering - Volume 1: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing. CBEB. 2025. IFMBE Proceedings, vol 125. pp. 496–502. Springer, Cham. https://doi.org/10.1007/978-3-031-93646-3_51

[13] Paudyal, S., Higgins, M., Meister, K. et al. “Photobiomodulation of the mammary gland for the treatment of mastitis in dairy cattle”. Animal Diseases 6, 34 (2026). https://doi.org/10.1186/s44149-026-00248-4

[14] Favaretto, M.,Fraga, D. R., Cidral-Filho, F. J. et al. “Efeito da fotobiomodulação com laser na composição, produção e qualidade do leite de vacas com mastite clínica”. Rev Gest Soc Ambient, 18(3), pp. 1-13, e04773, 2024. https://doi.org/10.24857/rgsa.v18n3-054

[15] BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA). In nº 76, de 26 de novembro de 2018.

[16] BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA). In nº 77, de 26 de novembro de 2018.

Wireless Wearable Multiparametric System for Neonatal Monitoring and Support for the Kangaroo Method

1st Cristoffer Lincon de Oliveira
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology – Paraná (UTFPR)
Curitiba, PR, Brazil
ORCID: 0009-0003-1449-8743

2nd Cristhiane Gonçalves
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology – Paraná (UTFPR)
Curitiba, PR, Brazil
ORCID: 0000-0002-3894-2971

3rd Henrique Cunha Carvalho
Graduate Program in Biomedical Engineering (PPGEB)
Federal University of Technology – Paraná (UTFPR)
Curitiba, PR, Brazil
ORCID: 0000-0003-0806-4132

***Abstract*— Continuous monitoring of physiological parameters in neonatal patients requires technological solutions that combine mobility, connectivity, and low cost. This study presents the development and evaluation of a proof-of-concept for a wearable, wireless, multiparameter system. The prototype integrates pulse oximetry (MAX30102) and body temperature (AHT25) sensors into an architecture based on ESP32 microcontrollers, which handle the acquisition, processing, and transmission of physiological data. Data are displayed locally on a screen and sent to the ThingSpeak® platform for remote monitoring. Testing demonstrated the integrated operation of the data acquisition, processing, visualization, and transmission modules, as well as the execution of alert routines implemented in the firmware. Additionally, the system operated for approximately 2.6 h. The results confirm the technical feasibility proposed and its potential for future applications in neonatal monitoring.**

Keywords*— Edge Computing, ESP32, Neonatal Monitoring, Internet of Things.

I. INTRODUCTION

This study presents the open architecture and benchtop evaluation of a low-cost, wireless multiparametric wearable prototype aimed at continuous data acquisition, serving as a preliminary step toward supporting non-invasive neonatal monitoring. [2,3].

In Brazilian hospital context, devices predominantly used to monitor newborns' vital signs are bedside multiparametric monitors [2]. Although accurate, such systems pose significant operational limitations when applied in the reality of neonatology [4]. An example is pulse oximetry, where most devices and peripheral sensors were originally designed to meet the ergonomic parameters and tissue characteristics of adult patients [4,5]. This morphological incompatibility can hinder proper sensor placement on newborns' limbs, especially in premature and low-birth-weight infants, compromising the quality of the photoplethysmographic (PPG) signal and increasing susceptibility to motion artifacts and inconsistent readings [4,5].

Additionally, hospital-grade sensors placed on the fingers are typically ineffective at mitigating motion artifacts from active infants [4]. Due to reflexes and spontaneous limb movements, these rigid devices easily detach, triggering recurrent false alarms that cause alarm fatigue among nursing staff or, more seriously, the actual interruption of monitoring [6]. In parallel, the large volume of cables required to connect the sensors can restrict the newborn's mobility and hinder care procedures in neonatal units [3]. This limitation can also act as a barrier to performing the Kangaroo Method and skin-to-skin contact between parents and children [7,8]. Outside the intensive care ecosystem, the vulnerability scenario is exacerbated by the lack of wireless, low-cost home monitoring systems adapted to neonatal anatomy [2,9].

Various approaches in the literature propose developing wireless architectures to mitigate the impact of cables on neonatal monitoring. Recent studies have investigated the use of wearable sensors and remote monitoring systems in both hospital settings and at-home follow-up of newborns [5,6,10]. However, many of these solutions focus on the acquisition and transmission of physiological signals. At the same time, integration with multiparametric monitoring, local alert processing, real-time wireless communication, and remote follow-up support remains limited.

To address this gap, this work aims to develop and evaluate a proof-of-concept for a wireless multiparametric system adapted to neonatal anatomy. In this way, it seeks to expand the capacity for continuous monitoring and to promote the early identification of predefined, potentially critical conditions in both hospital and home environments.

II. METHODOLOGY

From a methodological perspective, this work was structured around identifying project requirements based on the specialized literature.

The system's electronic architecture was centered on the ESP32-S3 Touch LCD 1.69 board, which features a dual-core microcontroller (MCU), integrated Wi-Fi and Bluetooth connectivity, and a 1.69-inch LCD screen (240x280 pixels) with a touchscreen interface (ST7789V2 driver) [11]. The touchscreen enabled the system to be organized into multiple navigation screens, facilitating switching between parameter viewing and real-time data monitoring, as shown in Fig. 1. Additionally, bench tests were conducted in a controlled environment to evaluate the prototype's functional

performance indicators.



Fig. 1. Displays of the ser interface into multiple navigation screens accessible via touchscreen.

A. Specification of Requirements

The first phase of development consisted of mapping the system's functional and non-functional requirements, based on international recommendations for neonatal care proposed by the World Health Organization (WHO) [12] and on evidence from recent literature on neonatal monitoring and wireless wearable devices [2,9,10].

Among the functional requirements established for the system, the capacity for synchronous acquisition of the PPG signal to estimate peripheral oxygen saturation (SpO2) and heart rate (HR), as well as continuous monitoring of body temperature (T), was highlighted. Associated with crying detection and the local triggering of humanized audible alerts, the system utilizes a miniature, square-shaped magnetic micro-speaker (mod. 1609, China) with an impedance of 8 Ω and a maximum power of 1 W [13].

As a non-functional requirement, it was established that the system should operate without wired connections between the newborn and external equipment during use, employing wireless communication for physiological data transmission and alert generation, thereby potentially supporting patient mobility and reducing limitations associated with conventional monitoring systems.

The mechanical requirements were also designed to minimize newborn handling and to reduce the presence of rigid or excessively heavy elements that might restrict mobility. In this way, the design sought to potentially reduce the risk of injuries resulting from mechanical pressure, friction, or improper sensor positioning, in line with the principles of neuroprotective care and the preservation of neonatal skin integrity [2,9,10,14].

B. Hardware Architecture and Selection of Components

The processing core is centered on the ESP32-S3 Touch LCD 1.69 platform (Espressif Systems Co. Ltd., Shanghai, China) [11], selected for its embedded processing capabilities, integrated graphical interface, and wireless connectivity. The system architecture is complemented by independent peripheral modules based on the ESP32-C3 Super Mini microcontroller (Espressif Systems Co. Ltd., Shanghai, China) [15], responsible for physiological signal acquisition and wireless communication with the central unit.

For the acquisition of the monitored physiological parameters, two high-sensitivity digital sensors were integrated, interconnected to the MCU via the I2C (Inter-Integrated Circuit) communication bus:

Oximetry and PPG Module: A CJMCU development module based on the MAX30102 sensor (Analog Devices Inc., Wilmington, USA) was used. It incorporates light-emitting LEDs in the red (660 nm) and infrared (880 nm) wavelengths, along with a high-sensitivity photodetector, an 18-bit analog-to-digital converter (ADC), and integrated ambient-light interference-rejection mechanisms [16].

Thermometry Module: For measuring neonatal body temperature, the AHT25 digital sensor (Aosong Electronics Co., Ltd./ASAIR, Guangzhou, China) was used, selected for its high resolution and accuracy in physiological monitoring applications, as well as its compatibility with low-power embedded systems [17].

The electrical power supply circuit, shown in Fig. 2, uses 3.7 V, 600 mAh lithium-ion batteries integrated with the TP4056 integrated circuit (Nanjing Top Power ASIC Corp., Nanjing, China) [18] to manage charging. The power, regulated at 3.3 V, is distributed independently among the three functional peripheral modules, based on the ESP32-C3-Super Mini MCU (Espressif Systems Co. Ltd., Shanghai, China) [15]. The first module performs the acquisition of SpO_2 and HR via the MAX30102 PPG sensor (Analog Devices Inc., Wilmington, USA) [16]; the second monitors temperature with the AHT25 digital sensor [17]; and the third comprises the acoustic system with the MAX4466 microphone (Maxim Integrated, San Jose, USA) [19] and the DFPlayer Mini player (DFRobot, Shanghai, China) [20], associated with the speaker [13]. Data collected by these nodes is transmitted via the ESP-NOW protocol to an ESP32-S3 Touch LCD central concentrator [11], which is responsible for real-time processing, local display, sending data to the cloud, and remote activation of the DFPlayer's audible alerts.

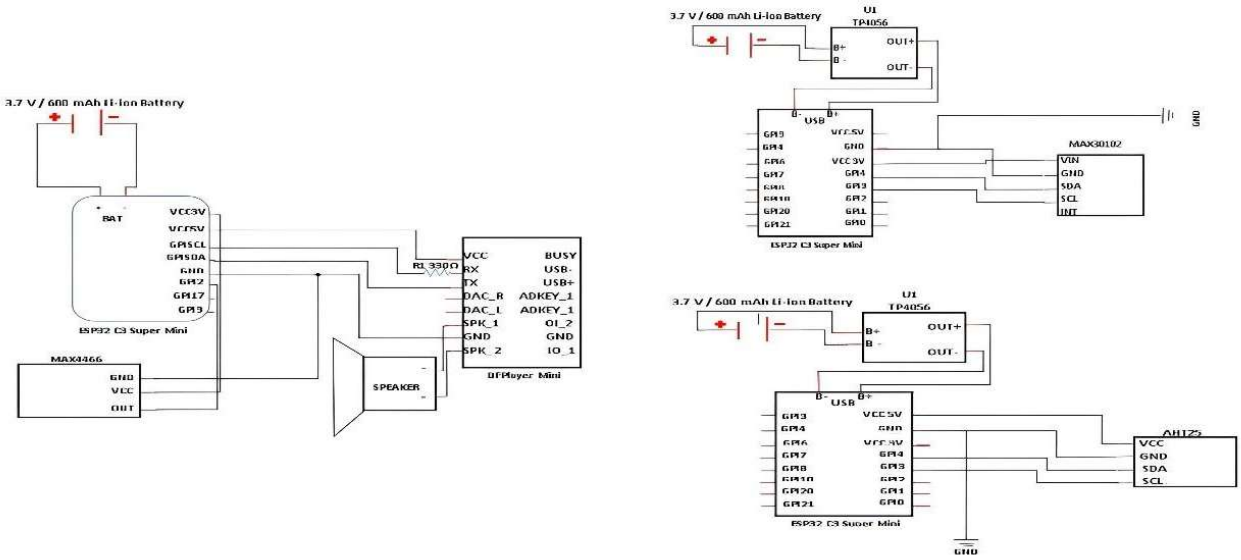


Fig. 2. Schematic diagram of the circuit for power supply and energy management of the neonatal multiparametric device.

C. Firmware, Edge Processing, and IoT Protocol

The ESP32-S3 firmware was developed in C++ and structured on FreeRTOS. Wireless communication between the peripheral nodes (ESP32-C3) and the central unit (ESP32-S3) uses the proprietary ESP-NOW protocol—a communication technology operating at the data link layer (MAC) without a router, enabling rapid packet transmission without burdening traditional connections. This technology was selected for its low transmission latency. The system performs continuous local sampling and applies a digital moving average filter to attenuate noise and transient fluctuations during signal acquisition, aiming to yield more consistent SpO_2 and HR results.

The calculation of SpO_2 is based on the ratio-of-ratios method, widely used in pulse oximetry systems to estimate arterial oxygen saturation from differential light absorption at red and infrared wavelengths.

Locally processed data are transmitted via Wi-Fi to the ThingSpeak® platform using HTTP requests, sending telemetry at 15-second intervals. This architecture enables cloud storage and remote monitoring of physiological variables, thereby characterizing an Internet of Things (IoT) application for continuous monitoring of neonatal parameters [21].

D. Demonstration and Bench Test Protocol

The prototype's operation was initially tested on a bench. The evaluation of thermal readings was performed by coupling the AHT25 sensor to a temperature reference system, enabling analysis of the sensor's behavior over the range of 34.0 to 42.0 °C.

To evaluate the robustness of wireless communication and the software architecture, a continuous physiological data transmission test to the monitoring platform was conducted via Wi-Fi. These tests aimed to analyze communication stability, data integrity and system behavior over long operating periods. The data recorded on the ThingSpeak® platform were used to analyze the temporal behavior of the monitored variables, thereby verifying the system's operational stability and the consistency of measurements during experimental testing.

III. RESULTS AND DISCUSSION

A. Device

Based on the established methodological workflow, the prototype is presented in Fig. 3a. It illustrates the oximetry (MAX30102) and temperature (AHT25) sensors integrated into bands made of elastic strap and Velcro fasteners for personalized contact with the neonatal limb and skin. Fig. 3b shows the encapsulated modules operating independently and wirelessly, communicating directly with the central unit, which is equipped with an LCD for data visualization.



Fig. 3. Functional prototype of the developed system in two views: (a) sensors exposed in the bands; (b) encapsulated modules and central unit.

Fig. 4 presents the system algorithm's flowchart, illustrating its continuous operating cycle. Initially, the SpO_2, HR, T sensors, and the dedicated microphone for cry detection are read. Next, the raw data are filtered and signal-averaged to attenuate noise, followed by local calculation and processing of biometric variables. After updating the display, the algorithm executes a conditional structure to check the neonatal behavioral state ("*Is the baby crying?*"). If affirmative, the system activates the audio to reproduce white noise or the mother's pre-recorded voice. Simultaneously with sending data to the cloud via the ThingSpeak® platform, the system evaluates whether parameters are outside normal limits. If anomalies are detected, an audible alert is emitted, categorized by thermal zones based on WHO guidelines [12]: 37.5 °C for feverish state, 38.0 °C for established fever, and 40.0 °C as a critical level of attention, then returning to the beginning of the cycle for a new evaluation of vital signs.

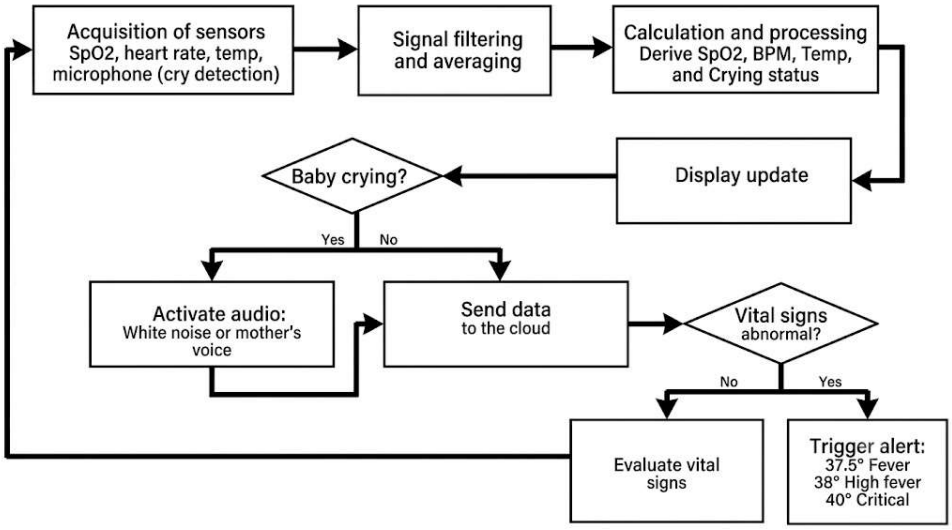


Fig. 4. Flowchart of the system architecture for neonatal multiparametric monitoring. Diagram adjusted using the Nano Gemini® AI-tool.

B. Functional Bench Tests

The prototype underwent functional bench testing to evaluate the functionality of the sensors, embedded processing routines, alert mechanisms, and wireless communication, as well as to evaluate system performance through oximetry indicators (%SpO_2), heart rate (BPM), body temperature (°C), wireless communication stability, and energy autonomy.

The developed prototype had an approximate weight of 32 g, including the battery, MCU, sensors, and the 1.69" ESP-S3 LCD. The system architecture was organized into functional modules responsible for physiological signal acquisition, data processing, power management, user interface, and remote communication. The sensors were distributed in specific modules as described below:

- Pulse Module: Oximetry module consisting of the MAX30102 sensor and a compact housing developed by additive manufacturing (3D printing) using white PLA (Polylactic Acid) filament. The housing was designed to mechanically accommodate the sensor, facilitate its positioning within the monitoring region, and provide greater

stability during PPG signal acquisition. The use of 3D printing enabled the fabrication of a structure with customized dimensions that meet project requirements, allowing geometric adjustments and adaptations during prototype development and tailoring the device to the specific needs identified for neonatal monitoring.

- Temperature Module: Thermal monitoring module featuring the AHT25 digital sensor in its original housing, designed to be positioned in future tests in the neonatal axillary region so that the metal sensing capsule remains in direct contact with the skin for data acquisition.

Benchtop limitations: tests were limited to simulated adult resting-state parameters (65–75 BPM and standard SpO2 values). These preliminary evaluations do not establish measurement accuracy in neonatal physiological ranges (120–160 BPM), nor do they account for motion artifacts, skin temperature variations, low peripheral perfusion, or long-term biocompatibility.

C. Evaluation of Performance Indicators

The evaluation phase aimed to validate the prototype's functionality and its capacity to acquire, process, and transmit physiological parameters through bench tests, assessing communication between electronic modules, the system's operational stability, and data transmission to the remote monitoring platform.

- Operational Stability of the System: The evaluation of the system occurred by monitoring the history of requests sent to ThingSpeak®, allowing the acquisition, processing, and transmission of data from the sensors integrated into the prototype. Each test lasted 60 minutes, with sampling every 15 seconds, enabling analysis of the operation of the distributed architecture comprising ESP32-C3-based acquisition modules and an ESP32-S3-based central unit, as well as communication between devices and the transmission of information to the monitoring platform.

- Oxygen Saturation (% SpO_2): The oximetry module based on the MAX30102 sensor was evaluated on the bench using an electronic patient simulator capable of simulating physiological signals compatible with HR and SpO_2 parameters. The procedure validated the LEDs' physical activation and the photodiode's ability to capture the reflection of the optical signal generated by the simulator. At the same time, the programming interface was used to verify the raw data reading. As shown in Fig. 5, the system continuously acquired and transmitted information to the central unit throughout the 60-minute trial, operating stably and without communication bus failures.

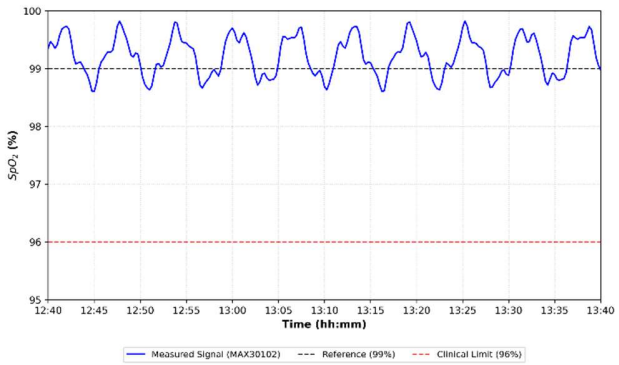


Fig. 5. SpO_2 (%) data obtained with the MAX30102 sensor. *Note:* Benchtop validation was performed using adult/resting-state simulation parameters (65–75 BPM) to evaluate baseline signal acquisition stability

- Heart Rate (BPM): Bench tests evaluate the acquisition of PPG signals, from which the firmware algorithm calculates HR by identifying pulse peaks. It should be noted that during these initial benchtop tests, simulation parameters were set to adult resting-state heart rates (65–75 BPM) to evaluate basic signal processing and transmission stability before clinical trials in neonatal ranges (120–160 BPM). As presented in Fig. 6, the module operated continuously during the 60-minute test, maintaining stable readings in the range of 65 to 75 BPM, corresponding to the condition simulated during the tests.

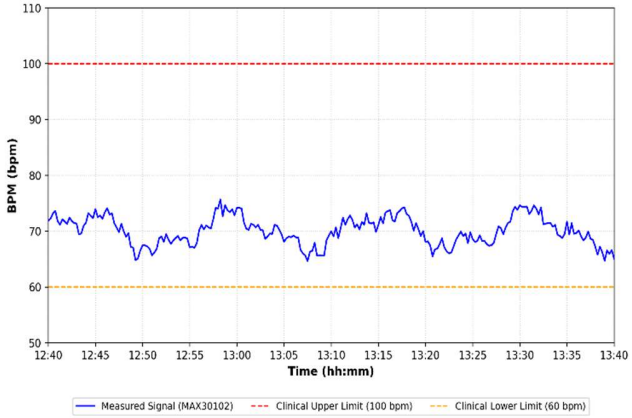


Fig. 6. HR (BPM) logs obtained by the MAX30102 sensor.

- Body Temperature: Fig. 7 presents the temperature data obtained by the AHT25 sensor. The curve records the initial room temperature reading at 20 °C before the start of the test. To evaluate the system's dynamic response, the sensor was subjected to an abrupt thermal variation that simulated the target temperature. The component required 0.45 s to register the thermal transition and stabilize the reading at around 36.5 °C, with residual oscillations of less than ±0.2 °C. Under the conditions evaluated, these results enabled observation of the sensor response and the alert routine's processing time in response to the applied thermal variation.

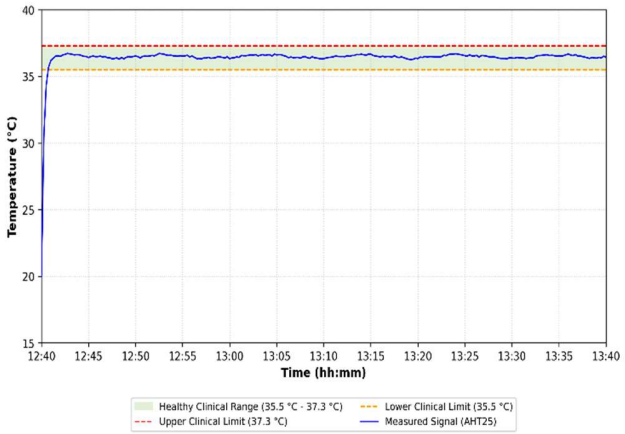


Fig. 7. Reading and thermal stability (°C) of the AHT25 sensor during functional system validation.

- Network Reliability and IoT Data Integrity: During the tests, the data acquired and validated by the acquisition modules were transmitted to the ThingSpeak® platform and made available for remote monitoring. The system employs a verification step to detect blood pulsation using the MAX30102 sensor's infrared (IR) signal before transmitting data. Consequently, only readings deemed valid, obtained when stable blood pulsation is detected over the sensor, are transmitted via ESP-NOW, reducing the risk of inconsistent records and improving remote monitoring reliability. Bench tests indicated proper reception of the transmitted information, contributing to the reliability of remote monitoring.

- Energy Efficiency and Autonomy: The prototype was powered by a lithium polymer (LiPo) battery with a nominal capacity of 600 mAh. During system operation, the acquisition, processing, and data transmission modules functioned continuously, demonstrating the technical feasibility of operation with a portable power supply. However, the resulting runtime is insufficient to support prolonged use of the prototype for the intended application. was verified using a digital oscilloscope (mod. TDS 2024C, Tektronix, Inc., Beaverton, OR, USA), with an average demand of 230 mA when the MAX30102 oximetry sensor was active and periodic data transmission via the wireless network was enabled. Communication periods account for the system's main energy demand due to the integrated radio's activity. Based on these electrical parameters, the module's nominal continuous-discharge autonomy is approximately 2.6 hours.

- Alert Latency and Evaluation of Intervention Logic: To evaluate the system speed, a critical anomaly was simulated through abrupt temperature variation and crying identification. The algorithm processed the variables and triggered the alert in 1.8 s after detecting the events. This response time demonstrates the system's ability to trigger an alert shortly after event detection; future studies should evaluate its potential contribution to reducing this time under the monitored conditions.

- Local Alert System Operation: During the tests, the system identified situations where sensor signals were absent and automatically executed the alert routine implemented in the firmware. Under this condition, the sound reproduction module was activated to emit the preconfigured voice message, providing a local response from the device.

Bench tests validated the MCU firmware interrupt routine for triggering the audible temperature alert via the micro-speaker. Local activation occurred immediately upon detecting the critical threshold, set to 37.5-40.0 °C during programming, and operated autonomously on the embedded unit without relying on external processing or cloud communication. The integrity of the audio signal was verified through logical monitoring of the MCU output pins and direct acoustic validation in the test environment, confirming correct execution of the alert code.

Future studies in neonatal patients may evaluate system performance under different motion and usage conditions. Additionally, the literature highlights the clinical relevance of continuous body temperature monitoring, especially in premature newborns, who are more susceptible to thermal disturbances due to immaturity of thermoregulatory mechanisms [9,22].

D. Kangaroo Method and Neonatal Care

Using a wireless architecture based on ESP32 MCUs eliminates the need for cables to transmit physiological data between system modules throughout the monitoring period, reducing limitations associated with the use of conventional wired connections; furthermore, the ESP32 family modules used in this project hold ANATEL (Agência Nacional de Telecomunicações) national regulatory certification [23] regarding radiofrequency and telecommunications requirements. However, this certification does not equate to approval of the system as a medical device. Consequently, any future application in a hospital setting would require additional evaluations concerning safety, performance, electromagnetic compatibility, and applicable regulatory requirements. In the literature, the Kangaroo Method is recognized as an important strategy for the care of low-birth-weight infants, based on skin-to-skin contact between the baby and their caregivers [7,8]. In this context, wireless monitoring solutions have been investigated as alternatives to reduce physical interference during neonatal care and evaluate the integration of continuous monitoring and humanized care [3,9,10].

Thus, the proposed system features characteristics consistent with current trends in wireless neonatal monitoring reported in the literature, enabling remote acquisition of physiological parameters without requiring permanent physical connections between device modules.

E. Comparative Cost and Functionality Analysis

Table I presents a comparison among the proposed system, hospital neonatal monitoring equipment and a commercial home care monitor. The developed system presents a lower estimated material cost than the final retail value of the other devices, which included certification and commercialization costs. In addition, the prototype integrates

SpO_2, HR, and T monitoring into a single platform, combining wireless connectivity, open cloud storage, and easily replaceable components.

TABLE I. COMPARATIVE COST AND FUNCTIONALITY ANALYSIS

Category	Hospital Monitor (Masimo/Philips)	Owlet Dream Sock	Proposed System
Cost (BRL / R$)	15,000 – 40,000	1,600 – 4,000	350 – 500
Measured Parameters	SpO_2, HR, Temperature	SpO_2, HR	SpO_2, HR, Temperature
Connectivity	Wired	*Bluetooth*	Wi-Fi
Response	Local Alarm	Smartphone Alerts	Audible Alarm, Mobile Alert
Data Storage	Restricted	Company	ThingSpeak®
Maintenance	Specialized Service	Device Replacement	Commercial Off-the-Shelf Components
Weight (g)	1,360 - 5,900	25 - 30	20 - 32
Battery Life (h)	3 - 4	8 – 16	2.6
Dimension (H x W x L) (cm)	22.9 x 10.2 x 16.5	4.1 x 2.3 x 0.8-15.5	4 x 2.8 x 2.2

Hospital multiparametric monitors used in neonatal units, such as the Masimo Rad-97 (Masimo Corporation, Irvine, CA, USA) [24] and the Philips IntelliVue MX450 (Philips Healthcare, Amsterdam, Netherlands) [25], integrate several physiological parameters into a single monitoring platform, including SpO_2, heart rate, temperature, and other vital signs. However, these solutions generally incur higher acquisition and maintenance costs than devices designed for home use or research prototypes. On the other hand, wearable devices intended for residential monitoring, such as the Owlet Dream Sock (Owlet Technologies Inc., Lehi, UT, USA) [26], employ optical sensors for physiological monitoring and utilize proprietary platforms for data storage and visualization. In comparison, the proposed system uses a wireless communication architecture and an open platform for remote monitoring, combined with low-cost, easily replaceable components, as presented in Table I.

One feature of the prototype is the reproduction of pre-recorded audio messages when certain conditions configured in the system are met. Unlike conventional audible alarms widely used in monitoring systems, this functionality was conceived to provide an alternative signaling method for caregivers. The literature highlights that frequent exposure to alarms can pose a challenge in neonatal environments, contributing to the phenomenon known as alarm fatigue [6]. In this context, the implemented feature offers a complementary approach to user interaction. In future versions, its potential use in a hospital setting could be evaluated, including its relationship to more humanized approaches to neonatal monitoring.

Data from the prototype were sent to the ThingSpeak® platform for storage and visualization. Although the platform has its own access control mechanisms, a specific assessment of the system's overall cybersecurity, including data protection during transmission, storage, and processing, was not conducted in this study phase, and will be analyzed in future versions.

Thus, future evaluations should consider various movement and positioning conditions, as well as aspects related to cleaning and contamination control, material biocompatibility, and device safety, before any clinical application.

Although the battery life was preliminarily estimated at around 2.6 h, which limits prolonged home monitoring or support for the Kangaroo Method, future research will focus on firmware optimization to reduce the system's duty cycle (active time). Since the prototype already uses the ESP-NOW protocol for low-latency direct communication, the intention is to implement the ESP32-C3's Deep Sleep low-power mode between sampling intervals. This approach will allow powering down the radio and peripherals almost instantaneously after data packet transmission, drastically reducing average current and extending operational autonomy to levels that could significantly extend the prototype's operational autonomy.

Additionally, further development of the prototype will require obtaining applicable ethical and regulatory approvals in future stages involving neonates. This includes comparative clinical studies against gold-standard reference equipment and quantitative agreement analyses (e.g., Bland–Altman plots, mean absolute error, root mean square error) between prototype measurements and those from certified devices, aiming to evaluate the performance of monitoring algorithms under various physiological conditions. Finally, transitioning from an experimental prototype to a product intended for clinical application evidently requires additional development steps, including electrical and thermal safety testing, mechanical risk assessment, human factors and usability evaluation, electromagnetic compatibility testing, and compliance with relevant regulatory requirements.

Comprehensive evaluations regarding electrical safety, thermal dissipation, mechanical choking hazards from small enclosure components, infection control protocols, and telemetry data encryption (cybersecurity over HTTP to ThingSpeak®) remain to be addressed in future clinical-grade iterations.

IV. CONCLUSION

The system for wireless multiparametric monitoring developed in this study demonstrated the technical feasibility of integrating sensors into a wearable architecture. The results showed proper functioning of the acquisition, processing, visualization, and remote transmission of physiological data, utilizing low-cost components and an open monitoring platform.

Wireless communication eliminated the need for cables to transmit data between devices, potentially providing greater flexibility during newborn care. Additionally, the system correctly executed the local processing and alert routines implemented in the firmware.

It should be emphasized that the results obtained

correspond to a proof-of-concept evaluated exclusively under benchtop conditions. The open, low-cost architecture represents the main contribution of this work; extensive clinical validation, quantitative accuracy assessment, risk analysis, and regulatory approval are mandatory prior to any healthcare or home application.

ACKNOWLEDGEMENTS

The authors thank the Federal University of Technology – Paraná (UTFPR) for making resources available for the project development, and the Araucária Foundation (FA) and the Superintendency of Science, Technology and Higher Education (SETI) for the conference grant awarded to H. C. C.

REFERENCES

[1] M. D. C. M. Senna, "Mortality Information System (SIM)," in *A experiência brasileira em sistemas de informação em saúde*, vol. 2, Brasília, Brasil: Ministério da Saúde, OPAS, Fiocruz, 2009, pp. 87–105.

[2] B. A. Krbec, X. Zhang, I. Chityat, A. Brady-Mine, E. Linton, D. Copeland, B. W. Anthony, E. R. Edelman, and J. M. Davis, "Emerging innovations in neonatal monitoring: A comprehensive review of the progress and potential of noncontact technologies," *Front. Pediatr.*, vol. 12, art. 1442753, 2024, doi: 10.3389/fped.2024.1442753.

[3] H. U. Chung et al., "Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units," *Nature Medicine*, vol. 26, no. 3, pp. 418–429, 2020.

[4] S. Fouzas, K. N. Priftis, and M. B. Anthracopoulos, "Pulse oximetry in pediatric practice," *Pediatrics*, vol. 128, no. 4, pp. 740–752, 2011.

[5] M. Thomas et al., "Accuracy of a wireless pulse oximeter in preterm or low-birth-weight infants <2.5 kg," *Am. J. Perinatol.*, vol. 41, no. S01, pp. e1606–e1612, 2024.

[6] Y. F. Stiglich, P. H. B. Dik, M. S. Segura, and G. L. Mariani, "The challenge of alarm fatigue in the neonatal intensive care unit: A before-and-after study," *Am. J. Perinatol.*, vol. 41, no. S01, pp. e2348–e2355, 2024.

[7] Z. C. Lamy, M. A. D. S. M. Gomes, N. O. M. Gianini, and M. D. A. E. S. Hennig, "Atenção humanizada ao recém-nascido de baixo peso – Método Canguru: a proposta brasileira," *Ciência & Saúde Coletiva*, vol. 10, no. 3, pp. 659–668, 2005.

[8] M. C. Spehar and E. M. F. Seidl, "Maternal perceptions in the Kangaroo Method: skin-to-skin contact, breastfeeding, and self-efficacy," *Psicologia em Estudo*, vol. 18, no. 4, pp. 647–656, 2013.

[9] E. Senechal, D. Radeschi, E. Jeanne, A. S. Ruiz, B. Dulmage, W. Shalish, R. E. Kearney, and G. Sant'Anna, "Wireless skin sensors for electrocardiogram and heart rate monitoring in the neonatal intensive care unit: A prospective feasibility, safety, and accuracy study," *Front. Bioeng. Biotechnol.*, vol. 13, art. 1555882, 2025, doi: 10.3389/fbioe.2025.1555882.

[10] E. Senechal, D. Radeschi, E. Jeanne, A. S. Ruiz, B. Dulmage, W. Shalish, R. E. Kearney, and G. Sant'Anna, "A framework to evaluate feasibility, safety, and accuracy of wireless sensors in the neonatal intensive care unit: Oxygen saturation monitoring," *Sensors*, vol. 25, no. 18, art. 5647, 2025, doi: 10.3390/s25185647.

[11] Waveshare Electronics, *ESP32-S3 1.69inch Touch LCD Development Board User Manual*, rev. 1.0, Shenzhen, China, 2024.

[12] World Health Organization (WHO), *Recommendations on Maternal and Newborn Care for a Positive Postnatal Experience*, Geneva, Switzerland: WHO, 2022.

[13] CUI Devices, *CDM-16095 Miniature Magnetic Speaker Datasheet*, rev. 1.0, 2023.

[14] L. Altimier and R. M. Phillips, "The neonatal integrative developmental care model: Advanced clinical applications of the seven core measures for neuroprotective family-centered developmental care," *Newborn Infant Nurs. Rev.*, vol. 16, no. 4, pp. 230–244, 2016.

[15] Espressif Systems, *ESP32-C3 Series Datasheet*, rev. 2.1, Shanghai, China, 2024.

[16] Analog Devices, *MAX30102 High-Sensitivity Pulse Oximeter and Heart-Rate Sensor for Wearable Health Datasheet*, rev. 2, Wilmington, MA, USA, 2021.

[17] Aosong Electronics, *AHT25 Type Humidity & Temperature Sensor Datasheet*, rev. 1.0, Guangzhou, China, 2020.

[18] Top Power ASIC Corp., *TP4056: 1 A Standalone Linear Li-Ion Battery Charger with Thermal Regulation Datasheet*, rev. 1.0, Nanjing, China, 2019.

[19] Analog Devices, *MAX4466 Low-Cost, Micropower Microphone Preamplifier Datasheet*, rev. 2, San Jose, CA, USA, 2012.

[20] DFRobot, *DFPlayer Mini MP3 Player Module for Arduino — Datasheet and User Manual*, Shanghai, China, 2016.

[21] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[22] E. A. Dunne, C. P. F. O'Donnell, B. Nakstad, and L. K. McCarthy, "Thermoregulation for very preterm infants in the delivery room: A narrative review," *Pediatric Research*, vol. 95, no. 6, pp. 1448–1454, 2024.

[23] UL do Brasil Certificações, "Certificate of Conformity: Certificate UL-BR 19.1464," *Agência Nacional de Telecomunicações (ANATEL)*, rev. 17 jun. 2024.

[24] Masimo Corporation, *Rad-97 Pulse CO-Oximeter Product Information*, Irvine, CA, USA, 2024.

[25] Philips Healthcare, *IntelliVue MX450 Patient Monitor Specifications*, Amsterdam, Netherlands, 2024.

[26] Owlet Technologies Inc., *Owlet Dream Sock User Guide and Product Specifications*, Lehi, UT, USA, 2024.

A Computational Tool for Technical Evaluation of Endoscopes

Mariana Gonçalves do Amaral
Biomedical Engineering
Universidade Federal do ABC
São Paulo, Brazil
m.amaral@aluno.ufabc.edu.br
ORCID: 0009-0003-9816-9591

Martha R. Ortiz-Posadas
Electrical Engineering
Universidad Autónoma Metropolitana Iztapalapa
Mexico City, Mexico
posa@xanum.uam.mx
ORCID: 0000-0001-5032-266

***Abstract*— The objective of this study was to develop a computational tool to prioritize preventive maintenance of endoscopes based on their technical evaluation. A predictive model was developed to estimate equipment failure risk based on the Random Forest Regressor algorithm. Two partial indicators were defined: P_1. Global relevance score, which aggregates operational variables related to equipment usage and maintenance conditions and, P_2. Weighted failure risk, which considers different types of recorded failures and their respective criticality weights. From these indicators, a global indicator called Integrated Risk was developed, representing the technical condition of the equipment. A qualitative scale with three categories was defined: good, attention and critical technical conditions, representing operational criticality levels of the medical devices. The model was applied to a sample of 39 video-gastroscopes from the Endoscopy Department of Hospital Beneficência Portuguesa in São Paulo, Brazil. Results were distributed as follows: six devices in good condition, 17 in attention condition, and 16 in critical condition. These preliminary results suggest that the proposed approach may support technical prioritization processes and preventive maintenance planning in hospital engineering management.**

Keywords—Random Forest regressor, health technology assessment, preventive maintenance.

I. INTRODUCTION

The management of hospital medical equipment is a fundamental component of healthcare administration, ensuring availability, traceability, and proper maintenance of technological resources. In the endoscopy sector, characterized by intensive use of high-cost and complex equipment, the adoption of computerized management systems contributes to process optimization, risk reduction, and improved patient safety. In this context, programs aimed at organizing and monitoring the equipment lifecycle represent strategic tools to support decision-making, promoting operational efficiency and sustainability in hospital practice.

The objective of this work was to develop and preliminarily evaluate a computational tool to support preventive maintenance prioritization of endoscopes based on operational and maintenance indicators derived from a previously developed mathematical model [1]. The Random Forest Regressor was explored as a predictive approximation approach to estimate equipment criticality from historical operational variables [2, 3]. This algorithm operates through the combination of multiple decision trees trained from random samples of the dataset. The goal was to provide a criterion to support prioritization of preventive maintenance, considering operational history and characteristics specific to each type of endoscope. The analyzed dataset included a sample of 39 video gastroscopes from the endoscopy sector of Hospital Beneficência Portuguesa in São Paulo, Brazil.

II. METHODOLOGY

The proposed framework combines a deterministic technical risk formulation with a machine learning predictive stage. Initially, the integrated risk indicator was mathematically constructed from operational and maintenance variables using predefined weighting factors. Subsequently, the Random Forest Regressor was trained to learn nonlinear relationships between the operational variables and the integrated risk score. Therefore, the machine learning stage acts as a predictive approximation model capable of estimating equipment criticality patterns from historical operational behavior.

Two partial indicators and one final criticality indicator were defined:

P_1. *Global relevance score*, which brings together operational variables related to the use and maintenance conditions of the equipment.

P_2. *Weighted failure risk*, which considers the different types of failures recorded and their respective criticality weights.

These two indicators were combined to form the Integrated Risk, representing a consolidated measure of the technical condition and criticality of the equipment.

Integrated Risk Indicator. Defined as the target variable (Y) of the predictive model. The Random Forest Regressor is trained using independent variables (X) related to operational and historical characteristics. During the training process, the algorithm learns the patterns present in historical data and the relationship between these variables and the previously calculated integrated risk.

A. Random Forest Regressor Algorithm

The operation of the Random Forest Regressor (RFR) is based on three key principles: bootstrap aggregation (bagging), which generates random training subsets to reduce variance; random feature selection, which increases diversity among individual trees; and prediction aggregation, where the final

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

output is obtained by averaging the predictions of all trees [2, 3]. As we said, in this study the RFR was used to estimate an integrated risk score associated with the criticality of medical equipment based on technical and operational variables. The resulting risk score enables the classification of equipment into priority levels, supporting preventive maintenance planning, technical interventions, and replacement decision-making processes.

The RFR model was developed in Python using the Scikit-learn library to predict the target variable based on the selected operational and technical indicators. Data preprocessing and analysis were performed with the support of the Pandas and NumPy libraries, while visualization of the results was carried out using Matplotlib and Seaborn. The model was configured with 200 decision trees and no restriction on tree depth, allowing the algorithm to fully capture complex nonlinear relationships within the dataset. To ensure reproducibility, a fixed random seed (random_state = 42) was adopted.

B. Mathematical Model

Five variables were selected with associated relevance weights (ρ_i) (Table I). These variables were considered representative of the technical and operational characteristics of the equipment. The variables considered most relevant were corrective maintenance and days out of service, given that they represent how many failures the equipment presented and how many days the equipment was out of use, respectively.

TABLE I. TECHNICAL VARIABLES

Variavel (x_i)	Weight (p)
x_1 Corrective maintenance (C)	1.0
x_2 Days out of service (D)	1.0
x_3 Preventive maintenance (P)	0.8
x_4 Equipment age (I)	0.7
x_5 Type of failure (F)	

It is important to emphasize that the variable x_2 is defined from the ranges of days the equipment is out of operation. Each range has an associated value as shown in Table II.

TABLE II. INTERVALS FOR VARIABLE X_2. DAYS OUT OF SERVICE

Days interval	Relevance(p)
[0, 30]	0,2
[31, 90]	0,7
[> 91]	1,0

*Global relevance score (Gr***)**. This partial indicator is defined by (1).

$$Gr = \sum_{i=1}^{4} \rho_i \, x_i \quad (1)$$

Weighted Failure Risk (Fr). This partial indicator is formed from the failures that the gastroscope presented. The failures were classified into five main types, and the rest into "other failures" (Table III). 1) Control buttons, are used to operate and control the different functions of the device. 2) Leaks, are usually caused by microcracks or damage to the external sheath, allowing fluids to enter the interior of the device. If not detected in time, they can lead to electronic failures, internal corrosion, and accelerated deterioration of the device. 3) Control knobs, are used to maneuver the distal tip of the insertion tube during a procedure. They provide fine control of angulation, which is essential for proper visualization, targeted examination, biopsy, and therapeutic interventions. 4) Images, related to the image quality display at the monitor. 5) Perforations, refer to tears or holes in the insertion tube, bending section, or internal channels. These defects can cause several serious failures: loss of insufflation, mechanical instability or infection risk, among others. Each type of failure has a specific value (M_5) according to the degree of relevance. This variable is basically formed by the sum of the failures with their respective weights.

TABLE III. TYPE OF FAILURE

x_5	Q_5	M_5
1	Control buttons (CB)	0.25
2	Leaks (L)	0.5
3	Control knobs (CK)	0.75
4	Images (I)	1.0
5	Perforations (P)	1.0
6	*Other failures* (OF)	0.5

The Fr indicator is defined by the mathematical function (2) formed by the failures presented in each piece of equipment, with the weighting factor corresponding to each failure.

$$Fr = \sum_{j=1}^{6} \rho_j \, x_5 \quad (2)$$

Where, ρ_j is the value of each type of failure, j = {1, ..., 6}

Integrated Risk (RI). The construction of this indicator aims to more accurately represent the actual criticality, that is, the technical condition of the equipment, considering not only its failures but also its other operational characteristics. The Integrated Risk (RI) is represented by a mathematical function that incorporates the two partial indicators previously described, along with their respective relevance factors, as shown in (3).

$$R_I = (0.6)\ Gr_k + (0.4)\ Fr_k \quad (3)$$

Where: k = {1, ..., 39}

C. Application of the Random Forest Regressor Algorithm

After constructing the partial indicators and calculating the Integrated Risk, defined as the target variable (Y) of the model, the predictive modeling stage begins using the Random Forest Regressor algorithm, as illustrated in Fig. 1. The independent variables (X) (Table 1) were used for the algorithm training. Additionally, other variable was included: equipment label. This variable originally categorical, was converted into a numerical format using label encoding, allowing its use in the model without introducing any hierarchical relationship among categories.

Qualitative Scale. A qualitative scale was defined to facilitate the interpretation of the results generated by the predictive model. This scale allows the numerical value of the integrated risk to be interpreted in terms of the technical and operational condition of the equipment. To this end, three main evaluation categories were established: good condition,

attention, and critical, which represent different levels of operational criticality identified from the risk score calculated by the model (Table IV).

TABLE IV. QUALITATIVE SCALE FOR THE NUMERICAL VALUE OF RI

R_I Interval	Technical condition
[0.0, 0.3]	Good
[0.31, 0.6]	Attention
[0.61, 1.0]	Critical

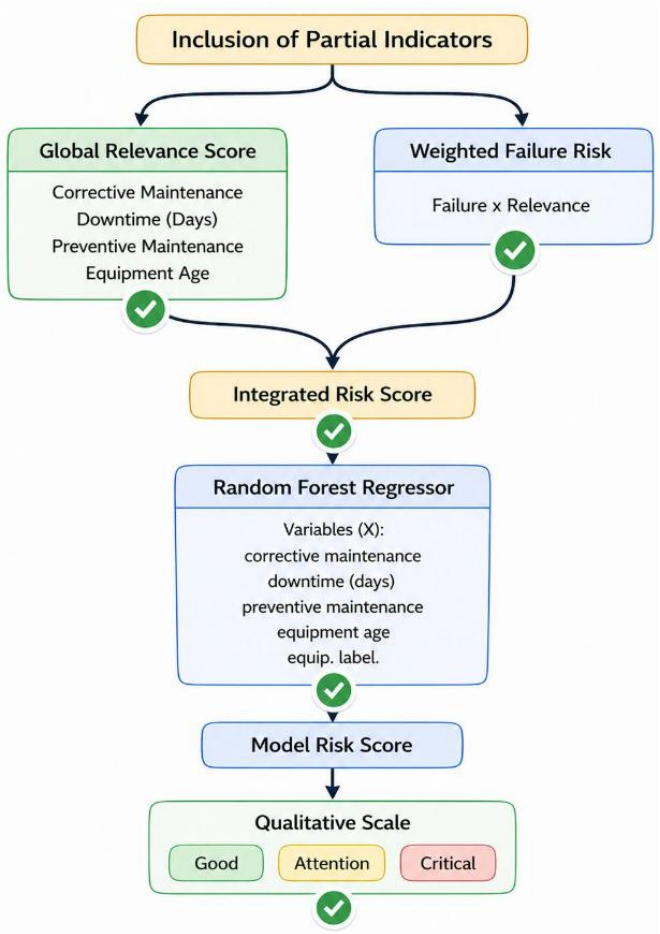


Fig. 1. Algorithm flowchart

D. Statistical Validation of the Predictive Model

The dataset was divided into training and testing subsets, where 75% of the observations were used for model training and 25% were reserved for independent performance evaluation. Due to the limited sample size (n = 39), k-fold cross-validation was additionally considered to reduce overfitting effects and improve the robustness of model evaluation. Model performance was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). Feature importance analysis was also performed to identify the variables with the greatest influence on the estimated integrated risk.

III. RESULTS

The model was applied for technical evaluation to a sample of 39 videogastroscopes: 36 standard gastroscopes and 3 ultrasound gastroscopes (US). The results were distributed as follows:

Good technical condition. Six video-gastroscopes were classified in this condition (Table V). They presented low values of downtime, as well as absence or low occurrence of corrective maintenance and, in some cases, records of preventive maintenance. These factors contributed to lower integrated risk scores in the model. Observe equipment G5 and G6 presented the highest risk values within this group. This behavior can be explained primarily by downtime and the occurrence of failures. For example, G5 registered just one failure, while G6 presented two occurrences. In contrast, the remaining equipment showed no downtime or recorded failures.

TABLE V. RESULTS OF THE INTEGRATED RISK SCORE OF 39 GASTROSCOPES

Video gastroscope	Integrated risk	Qualitative status
G1	0.03	
G2	0.10	
G3	0.15	Good
G4	0.16	
G5	0.25	
G6	0.25	
G7	0.37	
G8	0.37	
G9	0.37	
G10	0.37	
G11	0.37	
G12	0.42	
G13	0.44	
G14	0.45	
G15	0.46	Attention
G16	0,49	
G17	0.49	
G18	0.49	
G19	0.49	
G20	0.49	
G21	0.49	
G22	0.49	
G23+US	0.58	
G24	0.60	
G25	0.61	
G26	0.64	
G27	0.66	
G28	0.66	
G29	0.68	
G30	0.71	
G31	0.72	
G32	0.72	Critical
G33	0.72	
G34	0.75	
G35	0.76	
G36	0.78	
G37	0.79	
G38+US	0.82	
G39+US	0.89	

Attention technical condition. There were classified 17 video-gastroscopes in this condition. They presented intermediate Integrate Risk (RI) scores [0.37, 0.58] (Table V). In general, these devices showed an absence of preventive maintenance and/or the presence of corrective maintenance. Observed that equipment G23 exhibited the highest RI score (RI = 0.58) within this group. This may reflect occurrence of corrective maintenance has a significant influence on the increase of the risk score in this classification condition.

Critical condition. 16 Video-gastroscopes were classified in this category. They presented the highest RI scores (Table V). The RI is mainly associated with the combination of prolonged periods of downtime and a higher occurrence of corrective maintenance. For example, equipment G39 exhibited the highest risk score in the sample (RI = 0.89). This result is primarily associated with a high value of the corrective maintenance variable.

A. Statistical Validation of the Predictive Model

We analyzed the dataset using a Random Forest Regressor (RFR) with Integrated Risk (RI) as the target variable. The model performance is shown in Table VI. Observe that R^2 indicates that approximately 75.2% of the variance in the RI score was explained by the selected predictors. Five-fold cross-validation produced a mean R^2=0.630, suggesting moderate generalization capability despite the limited sample size.

TABLE VI. MODEL PERFORMANCE

Metric	Value
R^2 (Test Set)	0.752
MAE	0.108
RMSE	0.142
Mean Cross-Validation R^2 (5-fold)	0.630

Feature importance analysis revealed that days out of service was the most influential predictor of the Integrated Risk score (Table VII). This indicates that prolonged equipment downtime has the strongest impact on the estimated technical criticality of video-gastroscopes. Failure type and corrective maintenance history also contributed substantially to the model predictions, while preventive maintenance showed a moderate influence. In contrast, equipment age exhibited a relatively low contribution, suggesting that operational performance indicators may provide more relevant information regarding equipment criticality than chronological age alone. These findings support the use of reliability and maintenance-related variables as primary criteria for prioritizing preventive maintenance interventions.

TABLE VII. FEATURE IMPORTANCE

Variable	Importance
Days Out of Service	0.574
Failure Type	0.144
Corrective Maintenance	0.138
Preventive Maintenance	0.123
Equipment Age	0.021

To further evaluate the predictive performance of the RFR, an Observed versus Predicted plot was generated (Fig. 2). Observe a strong agreement between the actual and estimated Integrated Risk values, with most observations distributed near the identity line. That is, RFR successfully captured the relationship between operational variables and equipment criticality. Although a moderate dispersion was observed for some high-risk cases, no evident systematic prediction bias was identified. These findings are consistent with the obtained performance metrics (R^2 = 0.752, MAE = 0.108, RMSE = 0.142) and support the applicability of the model as a preliminary decision-support tool for maintenance prioritization. Nevertheless, further validation with larger datasets is required to assess the robustness and generalizability of the proposed methodology.

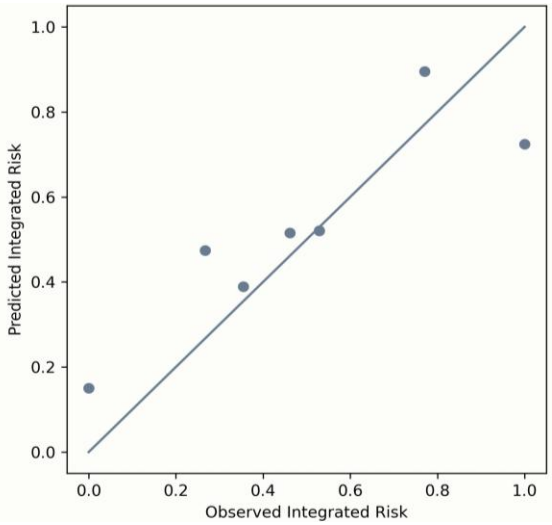


Fig. 2. Observed versus Predicted plot

IV. DISCUSSION

Compared with traditional maintenance prioritization approaches based exclusively on equipment age or corrective maintenance frequency, the proposed framework integrates multiple operational dimensions into a unified risk representation. This multidimensional approach allows identification of equipment presenting hidden operational vulnerability even when isolated indicators remain within acceptable ranges.

Although the Random Forest model demonstrated the ability to reproduce patterns associated with operational criticality, it is important to recognize that the integrated risk score itself was constructed from predefined engineering criteria. Therefore, the predictive model should not be interpreted as an autonomous diagnostic system, but rather as a computational mechanism capable of approximating complex maintenance relationships and supporting prioritization processes. In addition, the strong influence of corrective maintenance and downtime observed in the results suggests that operational instability may play a more relevant role in equipment criticality than age alone. This finding may contribute to maintenance planning strategies focused on reliability and availability indicators.

From a biomedical engineering perspective, the proposed framework may contribute to the evolution of intelligent hospital maintenance systems aligned with Industry 4.0 principles [4]. The integration of predictive maintenance models with hospital management platforms, and digital equipment lifecycle management may improve operational efficiency and reduce equipment downtime.

A. Limitations of the Study

This study presents some limitations that should be considered when interpreting the results. First, the dataset was obtained from a single hospital institution and included a reduced sample size (39 video-gastroscopes). This limitation may affect the generalizability and robustness of the predictive model and increase the risk of overfitting.

Second, the weighting coefficients used in the mathematical formulation were partially based on expert-defined relevance assumptions. Although these coefficients were designed to represent technical criticality, different institutional contexts may require coefficient recalibration.

Finally, the study should be considered exploratory due to the reduced sample size. Future multicenter studies involving larger datasets and different categories of endoscopic equipment are necessary to validate the external generalizability of the proposed framework.

V. CONCLUSIONS

The proposed computational tool demonstrated the potential to assist the prioritization of preventive maintenance actions for endoscopic equipment by integrating operational and maintenance-related variables into a unified risk indicator.

The results suggest that corrective maintenance history and prolonged downtime are strongly associated with increased integrated risk scores, while preventive maintenance may contribute to mitigating equipment criticality. Therefore, the current framework should be interpreted as an exploratory predictive support tool rather than a definitive maintenance decision system.

Future work should incorporate artificial intelligence techniques to improve transparency and interpretability, as well as, multicenter datasets and external validation procedures to improve robustness, scalability, and integration with intelligent hospital technology management systems. The proposed model could be incorporated into real-time maintenance ecosystems capable of continuously updating equipment criticality according to operational behavior and technical intervention history.

ACKNOWLEDGMENT

The authors of this work would like to thank Eng. Mike V Baroni for authorizing the use of the data presented in his undergraduate work, since without this valuable information, the development of the present study would not have been possible.

REFERENCES

[1] M. V. Baroni, Modelo para priorizar a manutenção preventiva de equipamentos médicos. Trabalho de Graduação. Curso em Engenharia Biomédica. Universidad federal do ABC, São Paulo, Brasil. 2025.

[2] [1] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in Python, 2nd ed. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-38747-0

[4] N. Ilankovic, A. Zelic, M Gubán and L Szabó, “Smart factories. The product of Industry 4.0.”, *Prosperitas,* vol. 7, no. 1, . 2020, pp. 19-30+, https://doi.org/10.31570/Prosp_2020_01_2.

A Multi-Criteria Decision Support Framework for Clinical Sensor Selection

Isabella Pereira Costa
Medical student
Federal University of Minas Gerais
Belo Horizonte, Brazil
isabellacostaufmg@gmail.com

Andrei Bueno Ferraz de Moura Gaspar
Undergraduate student in Electrical Engineering
University of Brasilia
Brasília, Brazil
andrei.gaspar@aluno.unb.br

Zilma Silveira Nogueira Reis
Dept. of Gynecology and Obstetrics
Federal University of Minas Gerais
Belo Horizonte, Brazil
ORCID: 0000-0001-6374-9295

Matheus Tiveron
Undergraduate student in Electrical Engineering
University of Brasilia
Brasília, Brazil
matheus.tiveron@aluno.unb.br

Enzo Camara de Oliveira Farias
Undergraduate student in Electrical Engineering
University of Brasilia
Brasília, Brazil
enzo.farias@aluno.unb.br

Artur Nogueira de São José
Department of Electrical Engineering
University of Brasilia
Brasília, Brazil
ORCID: 0000-0001-6985-394X

Pedro Henrique Braga da Silva
Undergraduate student in Electrical Engineering
University of Brasilia
Brasília, Brazil
silva.braga@aluno.unb.br

Hilton de Oliveira Mota
Department of Electrical Engineering
Federal University of Minas Gerais
Belo Horizonte, Brazil
ORCID: 0000-0001-7404-4936

***Abstract*— Healthcare decision-making often requires balancing clinical performance, economic constraints, and practical implementation requirements. When selecting a technology from among several alternatives, decision-makers must consider not only cost-effectiveness but also factors such as resource requirements, portability, environmental robustness, and ease of use. This study proposes a scenario-based multi-criteria decision analysis framework to support the selection of sensing technologies for human milk analysis. The framework was motivated by the need for a low-cost, portable device capable of measuring fat and protein concentrations in human milk samples. Three candidate sensing technologies were evaluated across three healthcare settings: a Primary Healthcare Unit, a Human Milk Bank, and a High-Complexity Tertiary Hospital. Technology-specific performance scores were combined with scenario-specific criterion weights using a normalized weighted-sum model. Results indicate that the relative suitability of the candidate technologies within the proposed framework depends on the operational conditions and clinical priorities of the intended healthcare setting.**

***Keywords*— human milk, MCDA, sensors.**

I. Introduction

Healthcare decision-making is a complex process that requires balancing clinical effectiveness, economic constraints, operational feasibility, and technological availability. Such decisions occur at multiple levels of healthcare systems, ranging from the routine selection of diagnostic tools by healthcare professionals [1] to public policy formulation and health technology incorporation [2], [3]. In particular, data science tools provide a promising means of supporting healthcare professionals in daily decision-making by enabling the analysis of the large volumes of patient data available in hospital settings [1]. Given the importance of multi-criteria decision-making for population health and welfare, evidence-based health technology assessment (HTA) procedures have increasingly been adopted by government health agencies worldwide [2], [3]. Researchers have also proposed incorporating HTA approaches into the early stages of medical device development, including the prototyping phase [4].

When selecting a new medical device from among several candidate technologies, clinical performance remains a fundamental requirement. However, practical factors such as affordability, portability, ease of use, interoperability, and the characteristics of a country's healthcare infrastructure [5] frequently determine whether a technology can be successfully implemented in real-world settings. Consequently, the adoption of biomedical devices increasingly requires structured decision-support approaches capable of simultaneously considering multiple, often conflicting evaluation criteria that extend beyond conventional cost-effectiveness considerations [6].

This challenge has become particularly evident with the rapid expansion of portable sensing technologies for point-of-care testing. Advances in biosensors, microelectronics, optical systems, and embedded computing have enabled the development of compact, low-cost diagnostic devices capable of generating clinically relevant information outside conventional laboratory environments. These technologies

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

are especially attractive in low- and middle-income countries, where geographical barriers, financial constraints, and limited specialized infrastructure frequently restrict access to laboratory-based diagnostic services. Ewen et al. [7] conducted a survey in six low- and middle-income countries—Fiji, Indonesia, Kyrgyzstan, Mali, Peru, and Vietnam—to investigate the availability and affordability of devices for self-monitoring blood glucose. In some of these countries, the availability of glucose meters for diabetes management was as low as 0% in public healthcare facilities and 31% in private pharmacies, while prices ranged from the equivalent of 1.5 to 10.9 days' wages. Xu et al. [8] described the potential scalability of low-cost, reusable wireless skin sensors designed to transmit newborn physiological data to smartphones. Such technologies may contribute to reducing mortality among preterm newborns in low-income countries.

In recent years, our research group has been involved in the development of low-cost sensors, with a particular focus on healthcare systems in low-income countries. One notable example is the Preemie-Test device, which uses optical sensing techniques to estimate gestational age [9] and assess pulmonary maturity [10] immediately after birth. The device provides clinically relevant information in settings where advanced diagnostic resources may not be readily available. Our current research investigates the effects of global warming on maternal and child health. In this context, the nutritional composition of human milk represents an important indicator of maternal and infant health. A recent systematic review [11] suggests a declining trend in milk fat and protein concentrations among mammals exposed to heat stress. We therefore aim to design or select, from among candidate technologies, an appropriate low-cost, portable sensing device capable of measuring fat and protein concentrations in human milk.

The present study addresses this practical need by proposing a scenario-based multi-criteria decision-support framework for evaluating and selecting sensing technologies capable of measuring specific macronutrients in human milk. The framework combines technology-specific performance scores with scenario-dependent weighting to support sensor selection across healthcare settings with different operational and clinical requirements.

The remainder of this paper is organized as follows. Section 2 introduces the proposed multi-criteria decision-support framework, including the evaluation criteria, performance-scoring scale (PSS), scenario-specific weighting strategy (W), and normalized weighted-sum procedure. Section 3 presents the healthcare scenarios and section 4 describes the candidate sensing technologies evaluated in this particular study. Section 5 presents and discusses the comparative results for each clinical scenario. Finally, Section 6 summarizes the main contributions, limitations, and directions for future research.

II. Proposed Framework

A. Multi-Criteria Evaluation Procedure

The proposed framework aims to classify quantitatively the candidate sensing technologies according to a performance-scoring scale (PSS), which characterizes the degree of compliance to technical criteria, and a set of weights (W) that specify the importance of each criteria to a specific healthcare scenario.

PSS and W were combined using a normalized weighted-sum model. The final weighted score S was calculated as:

$$S_{ts} = \frac{\sum_{i=1}^{10}(P_{ti} \times W_{is})}{\sum_{i=1}^{10} W_{is}} \quad (1)$$

In the above, P represents PSS, W the weight, t is the index of the sensor technologies, s is the index of the healthcare scenarios and i is the index of the i-th criterion. Therefore, P_{ti} represents the PSS assigned for technology t according to criteria i and W_{is} represents the weight assigned to criteria i in the s-th application scenario. It is expected that the development team will analyze the candidate sensing techniques and assign P and W values, allowing thus the calculation of the final scores given to each candidate technology. Higher values of the score indicator S_{ts} represent greater overall suitability of the sensing technology for the corresponding healthcare scenario.

B. Study Design

The proposed framework was applied to the comparative evaluation of candidate sensing technologies for human milk analysis. The framework was designed to compare the candidate technologies using a consistent evaluation structure while accounting for the operational and clinical requirements of different healthcare settings. Three technologies were selected based on previously reported applications in the assessment of milk properties. For each scenario, each sensing technology was evaluated according to 10 evaluation criteria, each of which have one of 4 performance scores (degree of technology compliance to the criteria) and one of 5 weights (degree of criteria importance for that scenario).

C. Development Team

Our study was conducted by an interdisciplinary team composed of 8 researchers, 3 of them senior or young university professors and 5 undergraduate students. One of them is a medical student and 4 are electrical engineering undergraduate students. The medical team was responsible for the definition of the clinical scenarios and performance criteria, while the engineering team was responsible for the technical analysis of the candidate sensor technologies.

The team worked in a collaborative way in order to avoid biased decisions as much as possible. However, it is advisable to keep in mind that the setting of PSS and W values in this way is not completely free of tendencies. This is a matter of future investigations in our work, *i.e.* to develop systematic data-based procedures to quantify technical relevance and scenario-specific agreement. In the current stage, the framework has been used for definition of which sensor technologies should be considered for in-depth study, actual development and testing.

D. Development of Evaluation Criteria

The set of 10 evaluation criteria was established to represent the main clinical, analytical, operational, safety, and implementation-related dimensions involved in the adoption of sensing technologies in healthcare practice. The criteria took in consideration clinical relevance, accuracy and reliability, usability, response time, cost, environmental robustness, user safety, sample handling practicality, integration with health information systems, and portability.

Each sensing technology was evaluated using a four-level PSS ranging from 0 to 3. A score of 0 indicated unacceptable performance, 1 indicated limited performance associated with substantial implementation constraints, 2 indicated acceptable performance with manageable limitations, and 3 indicated optimal or near-optimal performance for the intended application. Criterion-specific scoring descriptors were defined before the evaluation to improve transparency, consistency, reproducibility, and avoid biased conclusions, as summarized in Table 1. In particular, the Response Time criterion was defined as the total time from milk sample preparation to result availability.

TABLE I. TEN CRITERION-SPECIFIC DESCRIPTORS FOR THE INTRINSIC PSS.

Criterion/PSS	**0**	**1**	**2**	**3**
1. Clinical relevance	No relevant application to the intended clinical purpose	Indirect relevance; limited potential to support decision-making	Supports decision-making in selected situations	Directly and consistently supports decision-making for the intended application
2. Accuracy and reliability	Inconsistent results or unacceptable agreement with reference methods	High variability or frequent disagreement with reference methods	Acceptable agreement with limited variability	High agreement and consistent reproducibility
3. Usability	Impractical; requires highly specialized expertise	Complex procedure or extensive formal training	Moderate operational complexity; brief training required	Intuitive procedure requiring little or no formal training
4. Response time	>24 h or unsuitable for the intended application	>60 min to 24 h	5–60 min	<5 min or real-time results
5. Cost	Economically unfeasible	High acquisition or operational cost; restricted to specialized centers	Moderate cost; feasible in structured services	Low cost; suitable for large-scale implementation
6. Environmental robustness	Highly unstable outside strictly controlled conditions	Frequent interference; strict environmental control required	Minor interference under usual clinical conditions	Stable under common environmental variations
7. User safety	High biological, electrical, chemical, thermal, or occupational risk	Moderate risk; rigorous protective measures required	Low risk; manageable using standard protective measures	Minimal risk; appropriate protective features incorporated
8. Sample handling practicality	Infeasible or highly complex	Multiple preparation steps or specialized equipment required	Simple collection with limited preparation	Direct or nearly direct analysis with minimal preparation
9. Integration with health information systems	No digital record or output	Manual transcription required	Non-automated digital export	Automatic integration with electronic health records or laboratory systems
10. Portability	Fixed equipment or dependent on permanent infrastructure	Difficult to transport; substantial peripheral equipment required	Moderately portable; some auxiliary equipment required	Lightweight, compact, and self-contained

E. Scenario-Based Weighting Strategy

The relative importance of each evaluation criterion varied according to the operational conditions of each healthcare setting. Therefore, a scenario-specific weight was assigned to each criterion, as described in Table 2.

Weights W were assigned using a five-level relevance scale, in which 1 indicated low importance, 2 limited importance, 3 moderate importance, 4 high importance, and 5 critical importance for the adoption of that technology in that healthcare scenario. The weights represented the expected contribution of each criterion to implementation feasibility and clinical usefulness within the corresponding healthcare environment.

TABLE 2. SCENARIO-SPECIFIC WEIGHTS ASSIGNED TO THE EVALUATION CRITERIA

Evaluation criterion	Primary Healthcare Unit	Human Milk Bank	High-Complexity Tertiary Hospital
1. Clinical relevance	5	5	5
2. Accuracy and reliability	4	5	5
3. Usability	5	3	2
4. Response time	5	4	3
5. Cost	5	3	2
6. Environmental robustness	5	4	1
7. User safety	4	5	4
8. Sample handling practicality	5	4	3
9. Integration with health information systems	3	4	5
10. Portability	3	1	1

III. Clinical Scenarios

To demonstrate the proposed framework, three representative healthcare scenarios within the Brazilian public healthcare system were considered: a Primary Healthcare Unit, a Human Milk Bank, and a High-Complexity Tertiary Hospital. These scenarios were selected because they differ substantially in their operational conditions, resource availability, technological infrastructure, workforce expertise, and clinical priorities. Nevertheless, the proposed methodology is not restricted to the Brazilian Unified Health System (*Sistema Único de Saúde*, SUS) and may be adapted to private hospital networks, research institutions, and other healthcare systems requiring structured support for health technology adoption.

A. Primary Healthcare Unit (PHU)

The PHU scenario represents public primary care facilities characterized by high patient volumes, limited financial resources, heterogeneous technological infrastructure, and limited time for healthcare professionals to undergo training or perform complex procedures. Although these facilities provide adequate infrastructure for patient care, temperature and humidity are not usually subject to rigorous control.

In this setting, successful technology adoption depends primarily on clinical usefulness, affordability, operational simplicity, rapid availability of results, environmental robustness, and straightforward sample handling. Accuracy, reliability and user safety also remain important, while digital integration and portability may provide additional implementation benefits.

B. Human Milk Bank (HMB)

The HMB scenario represents specialized facilities responsible for the collection, processing, quality control, storage, and distribution of human milk, particularly for preterm infants and other high-risk newborns. These facilities employ trained personnel, follow established biosafety protocols, and have basic laboratory infrastructure. Environmental conditions are generally more controlled than those in primary care facilities, although less rigorously than in high-complexity hospitals.

Technology adoption in this scenario is primarily guided by the clinical relevance, analytical reliability, and safety of the results because the generated information may directly influence decisions regarding milk processing, quality

control, and distribution. Traceability, sample handling practicality, and timely result availability are also important, whereas portability is less relevant because the equipment is expected to remain within the facility.

C. High-Complexity Tertiary Hospital (HCTH)

The High-Complexity Tertiary Hospital scenario represents university hospitals and referral centers equipped with advanced technological infrastructure, integrated electronic health record systems, climate-controlled environments, specialized personnel, and technical support.

In this setting, technology adoption prioritizes clinical impact, analytical performance, and digital integration. Financial and operational constraints are less restrictive than in the other scenarios, while environmental variability is minimized through infrastructure-based controls. Consequently, greater operational complexity and higher costs may be acceptable when associated with improved clinical or analytical performance. Portability and environmental robustness are comparatively less critical because the equipment can remain in a fixed, controlled environment.

IV. Candidate Technologies

A. Technology #1

Islam et al. [12] describe a 4-electrode cylindrical sensor, as illustrated in Fig. 1, which is filled with milk samples. In this figure, A, B, C, and D represent terminals connected to metallic surfaces. If a voltage is applied between terminals A and D, while grounding terminals B and C, then the cross-conductance per unit thickness (S/m), G_1, can be indirectly measured with a simple inverting operational amplifier circuit. Now, if a voltage is applied between terminals B and C while grounding terminals A and D, the cross-conductance per unit thickness G_2 can be calculated, enabling the calculation of the conductivity of the medium:

$$\sigma = \frac{\pi(G_1+G_2)}{2Lln(2)} \quad (2))$$

where L is the level of liquid inside the cylindrical device. The idea is to prepare multiple milk samples, each having a certain amount of a given element and observe the variations upon the electric quantities. The authors observe these variations under different milk contamination conditions. However, this sensor has a good potential for testing milk samples with different fat and protein contents**.**

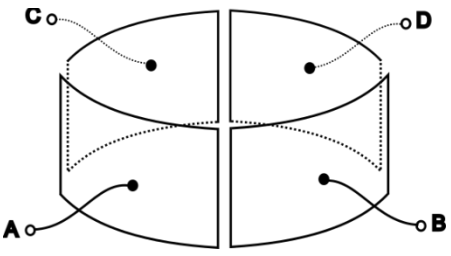


Fig.1 -A simplified illustration of the sensor described in [12]. Milk samples fill the space between the metal sheets.

B. Technology #2

Sheng et al. [13] describe an optical sensor device based on the capture of the transmitted light, after interacting with a milk sample stored within a cuvette, as illustrated in Fig. 2. One of the key aspects of this sensor device is the fact that it is based on a multi-wavelength analysis. The light source is a broadband infrared (IR) light emitting diode (LED) and the photodetector is a multi-spectral sensor capable of detecting six wavelengths: 610, 680, 730, 760, 810, and 860 nm. Such a feature allows the device to detect different absorbance signatures, so it can simultaneously detect fat and protein concentrations. Another key aspect of this technology is the post-processing of the measured signals with the Gradient-Boosted Regression Tree (GBRT) machine learning algorithm. So, an external laptop is required in addition to the elements shown in Fig. 2.

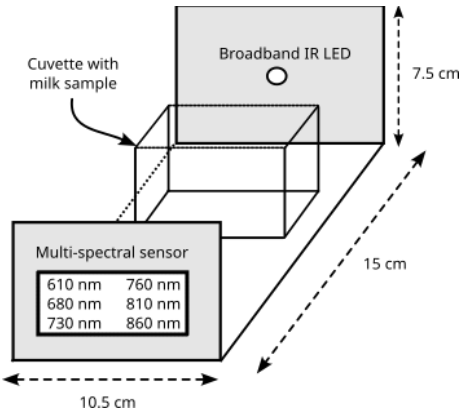


Fig.2 - A simplified illustration of the sensor described in [13].

C. Technology #3

Chakraborty and Biswas [14] describe a sensing technology which, in its first version [15], was designed to identify adulteration in commercial milk. In both versions of this technology, the impedance of a rectangular metallic probe is measured after it is immersed in milk. More specifically, authors found out that, after exciting the metallic probe with low-frequency current (500 Hz), the impedance phase varies as a function of milk fat percentage. The mechanism behind the sensing technology illustrated in Fig. 3 is chemical: a lipase layer is immobilized at the copper surface of the metallic probe. The contact of milk samples with the external lipase molecules triggers an hydrolysis process which, in turn, produces ions, affecting thus the probe impedance. Furthermore, the sensitivity of this device is potentialized with an increase of the surface area for the diffusion of ions. This is achieved after a treatment of the original copper surface with polymethyl methacrylate (PMMA), which makes it porous.

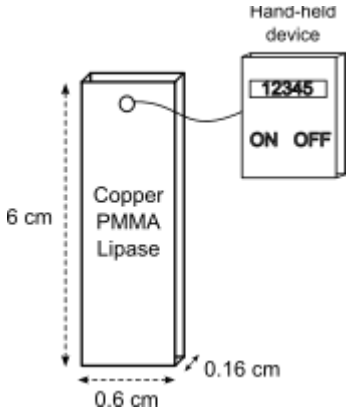


Fig.3 -A simplified illustration of the sensor described in [14].

V. Results and Discussion

Each candidate sensing technology was evaluated against the ten predefined criteria using the four-level ordinal scale described in Section 4.2. The resulting performance scores are presented in Table 3. To understand the comparative analysis, we must take into account the fact that not all information were clearly stated or declared in papers [12]-[14]. For instance, although all authors claim that their devices are cost-effective, only reference [14] brings the actual production cost (0.2 USD). In this situation, one can either make certain inferences or remove this criterion from the comparative analysis.

TABLE 3. PSS ASSIGNED TO THE CANDIDATE SENSING TECHNOLOGIES.

Evaluation criterion	Tec hnology 1 [12]	Tech nology 2 [13]	Tec hnology 3 [14]
1. Clinical relevance	1	3	2
2. Accuracy and reliability	1	2	2
3. Usability	3	2	2
4. Response time	3	3	2
5. Cost	3	2	3
6. Environmental robustness	2	1	1
7. User safety	2	2	2
8. Sample handling practicality	3	3	2
9. Integration with health information systems	1	1	1
10. Portability	3	2	2

Technology 1 [12] achieved the highest possible scores for usability, response time, cost, sample handling practicality, and portability. These scores reflected its simple operation, rapid measurement times, low estimated cost, limited sample preparation requirements, and compact configuration. However, it received lower scores for clinical relevance and accuracy and reliability because this technology was originally conceived for measuring adulteration of commercial milk. Therefore, the sensor was tested for detection of water in excess or even the presence of contaminants such as urea, rather than fat - one of our target milk nutrients. Its integration with health information systems was also limited because automatic data transfer was not demonstrated.

Technology 2 [13] received the highest clinical relevance score because it was designed to quantitatively estimate fat and protein concentrations, which directly corresponded to the intended application. It also achieved favorable scores for accuracy and reliability, usability, response time, and sample handling practicality. Nevertheless, its performance was limited by greater environmental sensitivity, dependence on calibration and computational processing, and limited portability. Automatic integration with health information systems was not demonstrated.

Technology 3 [14] showed moderate performance in clinical relevance, accuracy and reliability, usability, response time, sample handling practicality, and portability. Its categorical assessment of milk fat may provide clinically useful information in selected applications, although it does not quantify protein concentration and does not generate continuous quantitative measurements. The requirement to heat the sample before measurement reduced its environmental robustness and sample handling practicality. The absence of demonstrated digital data transfer resulted in the lowest score for integration with health information systems.

Overall, the three technologies exhibited distinct performance profiles. Technology 1 presented the most favorable operational, economic, and portability characteristics. Technology 2 demonstrated the strongest clinical relevance and a more comprehensive analytical capability. Technology 3 combined a favorable cost profile with moderate clinical and operational performance but was limited by sample heating requirements and the absence of protein measurement.

The suitability of each candidate technology across the application scenario is summarized in Table 4. The results came from the application of Eq. 2 combined with the scenario-specific weights and the PSS values described in Tables 2 and 3. It is interesting to note that there is no universal technology that simultaneously addresses the performance requirements from all application scenarios. For instance, Technology 1 seems to be the most suitable for the PHU scenario but not for the HMB and tertiary hospitals.

TABLE 4. FINAL WEIGHTED SCORES AND RANKINGS OF THE CANDIDATE SENSING TECHNOLOGIES ACROSS THE THREE HEALTHCARE SCENARIOS.

Technology	**PHU score**	**Rank**	**HMB score**	**Rank**	**Tertiary hospital score**	**Rank**
Technology 1 [12]	2.244	1st	2.026	2nd	1.906	2nd

Technology 2 [13]	2.195	2nd	2.132	1st	2.188	1st
Technology 3 [14]	1.976	3rd	1.868	3rd	1.875	3rd

The proposed methodology has certain limitations that should be addressed in a future work. One important limitation is the absence of an assessment of the bias introduced by human subjectivity when the PSS values are assigned (see Table 3). Different evaluators may fill this table differently and, consequently, the final scores of the candidate technologies can be affected. In a future study, we will recruit a set of independent evaluators, generating thus a dataset of PSS values which will, in turn, allow us to perform a statistical analysis. This will allow us to know the impact of small PSS variations from one evaluator to another around the same technology on the final scores assigned to the candidate technologies. The statistical analysis will also allow us to know if two technologies receiving close scores are statistically tied or not.

VI. Conclusion

This study proposed a scenario-based MCDA framework to support the selection of sensing technologies for human milk analysis across healthcare settings with different operational and clinical priorities. The application of the proposed framework showed that the relative suitability of the candidate technologies varied according to the intended healthcare setting.

The findings suggest that selecting a sensing technology solely on the basis of analytical performance or acquisition cost may overlook relevant implementation requirements. Operational simplicity, affordability, rapid response, and sample handling practicality were particularly influential in the PHU scenario. Conversely, clinical relevance, accuracy and reliability, and compatibility with specialized workflows had greater importance in the HMB and High-Complexity Tertiary Hospital scenarios.

The proposed framework provides a transparent and structured approach for comparing candidate technologies across distinct healthcare environments. It may also be adapted to other biomedical devices, clinical applications, and healthcare systems by modifying the evaluation criteria, scoring descriptors, and scenario-specific weights according to the intended context of use.

Nevertheless, the findings should be interpreted in light of the study's limitations. The performance assessment was based on technical evidence reported in the original studies rather than on direct experimental comparison of the technologies using human milk samples. Some relevant characteristics were therefore inferred from the reported operational configurations, and none of the candidate technologies had been comprehensively validated for the intended clinical application. In addition, the performance scores and scenario-specific weights were established through judgment and consensus within the study team and may vary according to institutional priorities, available resources, and stakeholder perspectives.

Future studies should experimentally evaluate the candidate technologies using human milk samples under representative operational conditions, incorporate the perspectives of healthcare professionals and other stakeholders, and perform sensitivity analyses to determine whether variations in performance scores and criterion weights alter the final rankings. These steps may strengthen the robustness and practical applicability of the framework and support evidence-based decisions during the development and adoption of sensing technologies in healthcare.

Acknowledgments

This work was funded by the Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF), through the Edital 09/2024-PROGRAMA PRIMEIROS PROJETOS-PPP-FAPDF-CNPq, grant number 00193-00001402/2024-51.

References

[1] M. W. L. Moreira, J. J. P. C. Rodrigues, V. Korotaev, J. Al-Muhtadi, and N. Kumar, "A comprehensive review on smart decision support systems for health care," IEEE Syst. J., vol. 13, no. 3, pp. 3536–3545, Sep. 2019, doi: 10.1109/JSYST.2018.2890121.

[2] D. Banta, "The development of health technology assessment," Health Policy, vol. 63, no. 2, pp. 121–132, Feb. 2003, doi: 10.1016/S0168-8510(02)00059-3.

[3] A. Angelis et al., "Multiple criteria decision analysis for HTA across four EU member states: Piloting the Advance Value Framework," Soc. Sci. Med., vol. 246, Art. no. 112595, 2020, doi: 10.1016/j.socscimed.2019.112595.

[4] I. G. Gregório, H. P. da Silva, A. F. A. dos Santos, and A. C. L. Vieira, "Early-stage decision-making for wearable and invisible digital health technologies: A staged multi-criteria value framework," IEEE J. Electron Devices Soc., in press, doi: 10.1109/JEDS.2026.3695296.

[5] C. Siriram and R. Harris, "A structured MCDA-based decision framework for funding decisions in South Africa: An illustrative case study," S. Afr. J. Econ., vol. 94, no. 2, Art. no. e70028, 2026, doi: 10.1111/saje.70028.

[6] P. Thokala and A. Duenas, "Multiple criteria decision analysis for health technology assessment," Value Health, vol. 15, no. 8, pp. 1172–1181, Dec. 2012, doi: 10.1016/j.jval.2012.06.015.

[7] M. Ewen et al., "Availability, prices and affordability of self-monitoring blood glucose devices: Surveys in six low-income and middle-income countries," BMJ Public Health, vol. 3, Art. no. e001128, 2025, doi: 10.1136/bmjph-2024-001128.

[8] S. Xu et al., "Wireless skin sensors for physiological monitoring of infants in low-income and middle-income countries," Lancet Digit. Health, vol. 3, pp. e266–e273, 2021, doi: 10.1016/S2589-7500(21)00001-7.

[9] G. L. N. Vitral et al., "Gestational age assessed by optical skin reflection in low-birth-weight newborns: Applications in classification at birth," Front. Pediatr., vol. 11, Art. no. 1141894, 2023, doi: 10.3389/fped.2023.1141894.

[10] Z. S. N. Reis et al., "Respiratory distress syndrome prediction at birth by optical skin maturity assessment and machine learning models for limited-resource settings: A development and validation study," Front. Pediatr., vol. 11, Art. no. 1264527, 2023, doi: 10.3389/fped.2023.1264527.

[11] A. N. de São José et al., "Global warming impacts on lactating mammalian milk: A systematic review," Sci. Total Environ., vol. 1003, Art. no. 180757, 2025, doi: 10.1016/j.scitotenv.2025.180757.

[12] T. Islam, A. Salamat, S. K. Singh, and M. Rehman, "A direct AC cross-conductive sensor for milk quality measurement," IEEE Trans. Instrum. Meas., vol. 71, pp. 1–8, Art. no. 1500208, 2022, doi: 10.1109/TIM.2022.3147313.

[13] T. Sheng, S. Shi, Y. Zhu, D. Chen, and S. Liu, “Analysis of protein and fat in milk using multiwavelength gradient-boosted regression tree,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–10, Art. no. 2507810, 2022, doi: 10.1109/TIM.2022.3165298.

[14] M. Chakraborty and K. Biswas, “Hardware platform to detect fat percent in milk using a lipase immobilized PMMA-coated sensor,” IEEE Trans. Instrum. Meas., vol. 68, no. 11, pp. 4526–4534, Nov. 2019, doi: 10.1109/TIM.2018.2890320.

[15] M. Chakraborty and K. Biswas, “Limit of detection for five common adulterants in milk: A study with different fat percent,” IEEE Sensors J., vol. 18, no. 6, pp. 2395–2403, Mar. 2018, doi: 10.1109/JSEN.2018.2794764.

A Multimodal Framework for Hansen's Disease Monitoring Using Lesion Detection Segmentation Quantitative Feature Extraction and LLM-Based Clinical Reasoning

Francisco Italo Guilherme da Silva
PPGEE, Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brasil
italoguilherme@alu.ufc.br

Marcos José Alves de Barros Monteiro
PPGETI, Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brasil
marcosmonteiro.ufc@gmail.com

Osvaldo Soares Landim Junior
PPGEE, Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brasil
Osvaldo.jrjuniorlinhares@gmail.com

Kauê Lima Sousa
LISIA
Fortaleza, Ceará, Brasil
kauelimasousa02@gmail.com

Leandro Costa Silva
LISIA
Fortaleza, Ceará, Brasil
lcostasilva02@gmail.com

Marcus Vinicius Candido de Figueiredo
Instituto Federal do Ceará (IFCE)
Fortaleza, Ceará, Brasil
mvcf10@gmail.com

Victor Hugo Costa de Albuquerque
DETI, Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brasil
victor.albuquerque@lesc.ufc.br

Luís Fabrício de Freitas Souza
Universidade Federal do Cariri (UFCA)
Juazeiro do Norte, Ceará, Brasil
fabricio.freitas@ufca.edu.br

***Abstract*—Skin lesions pose a significant challenge to the public health system, especially in the case of skin diseases such as Hansen disease, where a delayed diagnosis can lead to serious complications. Given this issue, this study proposes a new approach titled "A Multimodal Framework for Hansen's Disease Monitoring Using Lesion Detection Segmentation Quantitative Feature Extraction and LLM-Based Clinical Reasoning", grounded in a hybrid architecture that integrates lesion detection using YOLOv8, segmentation with the Segment Anything Model (SAM), and feature extraction, followed by clinical interpretation using large language models. The study used the public CO2Wounds-V2 dataset with 764 RGB images from 96 patients with Hansen disease, while the LLMs used were Qwen2.5 and Gemma2. As a result, this study achieved satisfactory detection results with Precision above 90%. This was followed by segmentation, where a Precision above 80% was achieved. Thus, the results obtained were quite satisfactory for CAD systems associated with a multimodal methodology for monitoring skin lesions and supporting clinical medical decisions.**

***Index Terms*—Hansen's disease, detection, segmentation, prediagnosis, large language models (LLMs), artificial intelligence, healthcare, computer-aided diagnosis (CAD), skin lesion analysis.**

I. Introduction

The human skin is one of the most important organs in the human body, as it performs its function exceptionally well, protecting the body against bacterial attacks and dehydration [1]. When a skin injury occurs, the body's defense system triggers several processes to restore the organism's protection. However, wounds in human adults undergo repair processes that lead to scar formation. Nevertheless, in some cases, the natural healing process does not occur properly, giving rise to what are known as chronic wounds [1], [2].

Chronic wounds can be defined as a disruption in the continuity of body tissue. These types of conditions affect approximately 5% of adults in Western countries and impose significant burdens on healthcare systems, as they require prolonged hospitalizations and extensive procedures, with a high recurrence rate [3].

Among the conditions associated with the development of chronic wounds, Hansen disease stands out as a chronic infectious disease caused by *Mycobacterium leprae*, which primarily affects the skin and peripheral nerves and can spread to other organs, causing neurological damage and permanent sequelae that compromise tissue function, leading to the development of skin lesions that are difficult to heal [3], [4]. Transmission occurs through close and prolonged contact between infected individuals and susceptible persons [3]. According to the World Health Organization, Brazil ranks second worldwide in reported Hansen disease cases, with 22,129 new cases recorded in 2024, representing a significant public health concern; if Hansen disease is not properly detected and treated, it can result in serious disabilities, impairing the patient's quality of life due to the social stigma associated with the disease [4], [5].

In this context, early detection of Hansen disease is essential to ensure prompt treatment, clinical follow-up, and the prevention of permanent sequelae [6]. In this regard, with the aim of leveraging technology to improve medical diagnosis, the Federal Council of Medicine published Resolution CFM No. 2,454/2026, which regulates the use of Artificial Intelligence in medical practice in Brazil. This regulation establishes guidelines for the development and application of AI-based systems in healthcare, encompassing diagnosis, prognosis, prevention, and disease management [21].

Consequently, this study address deep learning methods by applying techniques for the detection, segmentation, and progressive quantification of Hansen disease lesions. Recent segmentation models based on foundation models, such as the Segment Anything Model (SAM), enable the monitoring of disease progression and provide a basis for planning and evaluating the applicability of medical treatment [16].

In the context of generalized segmentation in medical imag-

ing, the ability to delineate regions of interest without the need for manual annotation has improved significantly [15].

In addition to the detection and segmentation stages, there remains a gap in the automated clinical interpretation of quantitative features extracted from lesions. In this context, large language models are increasingly being recognized as valuable tools for clinical decision support [12].

Large Language Models (LLMs) are advanced systems based on transformer architecture, capable of processing and integrating large volumes of information, thereby enabling contextualized analysis of clinical data [14]. Recent studies indicate that LLMs have the potential to support clinical decision-making (CDM), particularly in the interpretation of complex data and in the generation of structured recommendations [13].

For this study, the publicly available CO2Wounds-V2 dataset was used. The dataset comprises 764 RGB images from 96 patients diagnosed with Hansen disease, captured using smartphone cameras under varying lighting conditions and image resolutions. Of these, 607 images contain bounding box annotations, which were used during the model training and validation stages. The remaining images, lacking bounding box annotations, were used exclusively for inference, in order to assess the method's performance on unannotated data and evaluate its applicability in clinical contexts.

Accordingly, this study aims to compare different LLMs based on features extracted from lesion images providing a comprehensive perspective on automated detection, segmentation, and LLM-based analysis to support the clinical characterization of Hansen disease lesions, thereby contributing to the monitoring of disease progression in resource-limited settings.

II. RELATED WORK

Traditionally, the diagnosis of infections in dermatological wounds has relied heavily on the assessment of clinical characteristics and reported symptoms. As noted by [20], conventional techniques, such as swabs and biopsies, present critical limitations, as they are invasive and time-consuming, leading to a growing demand for solutions that offer faster and more accurate diagnosis.

In this context, with the advancement of Artificial Intelligence (AI), several studies have proposed methods based on machine learning (ML) and deep learning (DL) for the detection and segmentation of dermatological lesions in medical images. Works such as the SDA-Detection Melanoma framework proposed by Rodrigues et al. [8] demonstrate that convolutional neural networks, including architectures such as U-Net and YOLO, achieve high performance in the identification and segmentation of skin lesions, with reported metrics exceeding 96% accuracy.

The study in [23] focused on skin cancer, where a new approach was developed for the detection, segmentation, and classification of melanomas in dermatoscopic images. The proposed pre-diagnosis using generative AI achieved 99% accuracy, outperforming some models in the literature. Meanwhile, the classification phase relied on an integrated pipeline for feature extraction and selection of the best combination for classifying the melanoma region, achieving an accuracy of 86.0

The study by [24] evaluates a segmentation tool using machine learning methods to aid in the early diagnosis of skin cancer through the semantic segmentation of dermatoscopic images. The results showed average Dice loss values of 0.0971 during training and 0.1724 during validation, indicating good accuracy in image segmentation.

Other studies have made significant contributions to the segmentation of skin lesions, such as the one by [25], in which a computer-aided medical pre-diagnosis system was developed. By combining computer vision techniques with active contour models, the new method contributes to clinical diagnosis, demonstrating good performance even in images with high levels of noise.

Furthermore, Rayed et al. [18] review the advances and challenges of deep learning-based medical image segmentation, highlighting issues such as the scarcity of annotated data, limited model generalization, and sensitivity to variations in clinical images.

Studies applied to other diseases, such as the work presented in [9], demonstrate the potential of advanced computational methods for segmenting tumor regions in magnetic resonance images, achieving high performance rates. Similarly, the work presented in [10] demonstrates the effectiveness of segmentation techniques in computed tomography, with metrics exceeding 99% in controlled settings.

While these studies address different clinical contexts, they share challenges similar to those encountered in the analysis of dermatological lesions, underscoring the need for robust methods focused on detection, segmentation, and quantification in real-world scenarios. In the context of chronic wounds, the study by [11] demonstrates the potential of the YOLO family of architectures for the automatic detection of regions of interest and the prediction of lesion classes and bounding boxes.

Despite these advances, there remains a gap between precise visual segmentation methods and approaches capable of performing structured clinical interpretation of features extracted from medical images. In this context, Large Language Models (LLMs) have emerged as promising tools for supporting the interpretation of medical data and aiding clinical decision-making.

Accordingly, this study proposes the integration of computer vision techniques for lesion segmentation with LLM-based models for medical data analysis, as a diagnostic support technology. Specifically, it employs YOLO-based detection and segmentation architectures to identify Hansen disease lesions, with the aim of supporting clinical interpretation and monitoring disease progression.

III. MATERIALS AND METHODS

A. Dataset

The dataset used in this study was CO2Wounds-V2 [19], comprising 764 RGB images of chronic wounds from 96

patients diagnosed with Hansen disease. The images were captured in a clinical setting using smartphone cameras and varied in terms of lighting, resolution, capture angle, and acquisition conditions.

Among the available images, 607 contain bounding box annotations, which were used during the training and validation stages of the detection model. The remaining unannotated images were used exclusively for inference, allowing for the evaluation of the model's performance in scenarios closer to real clinical practice, where labeled data is scarce. Prior to training, a preprocessing and data cleaning pipeline was conducted, encompassing image integrity checks, format standardization, and partitioning into training, validation, and test subsets.

B. YOLO

YOLOv8 (You Only Look Once, version 8) represents one of the most recent advancements in object detection within the YOLO family of frameworks. Its architecture is organized into four main components: the Backbone, responsible for extracting visual features at multiple scales; the Neck, responsible for fusing and aggregating these features; the Detection Head, responsible for predicting bounding boxes and classifying regions of interest; and the Segmentation Head, an additional module that enables the generation of instance segmentation masks directly from the model.

The model was trained on the 607 annotated images from the CO2Wounds-V2 dataset, partitioned into training, validation, and test subsets. The bounding boxes predicted by the model serve as input for the subsequent segmentation stage, integrating the proposed pipeline in a sequential and automated manner.

C. Segmentation

The purpose of this segmentation step is to precisely delineate the boundaries of Hansen disease lesions identified during the detection stage, thereby enabling the extraction of quantitative features for clinical analysis. To this end, the Segment Anything Model (SAM) [15], developed by Meta AI, was adopted due to its versatility and ability to integrate with other systems. SAM was designed to segment objects of interest in images based on user-provided prompts, which can take the form of a single point, a set of points, a bounding box, or text, and returns a corresponding segmentation mask [16].

In the proposed pipeline, the bounding boxes generated by YOLO during the detection stage were used directly as input prompts for SAM, obtaining segmentation masks as output. This integration enables fully automated segmentation without manual intervention, an approach that proved well-suited to the characteristics of the CO2Wounds-V2 dataset.

D. Large Language Models (LLMs)

Large Language Models (LLMs) are advanced AI systems based on the Transformer architecture, capable of generating and processing human language through probabilistic modeling of sequential tokens. These models are typically trained via self-supervised learning, which enables them to develop broad language comprehension and generalization capabilities, and are increasingly being adapted for clinical tasks. In this study, LLMs were employed to interpret quantitative features extracted from segmented lesions and generate structured clinical descriptions, functioning as clinical decision support tools. To this end, the features of each lesion were organized and standardized prior to being provided as input to the models, which produced structured outputs classifying the lesions and indicating the presence of signs of infection.

IV. Methodology

This section describes the methodological workflow of the proposed approach, as illustrated in Figure 1. It is organized into sequential steps that constitute the complete pipeline, from dataset acquisition and image preprocessing to clinical decision support.

Step 1: Detecting the Lesion Region with YOLO

In the first stage, the YOLOv8 model was employed to automatically detect regions of interest through the prediction of bounding boxes around Hansen disease lesions. The dataset was partitioned into 70% for training, 20% for validation, and 10% for testing. The resulting model returns the detected region along with the corresponding confidence score for each identified lesion, as illustrated in Figure 2.

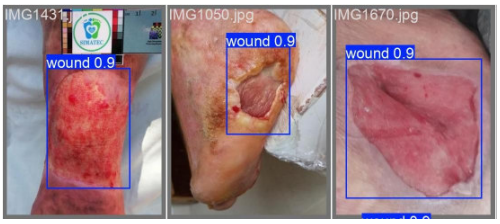


Fig. 2. Predicted detection results.

Step 2: Segmentation with SAM

Following lesion detection, the Segment Anything Model (SAM) was applied to delineate the identified regions. The bounding boxes generated by YOLOv8 were provided directly as input prompts to SAM, enabling more precise mask generation of the wound contours, reducing background interference, and improving the quality of the representation of the affected region, as shown in Figure 3.

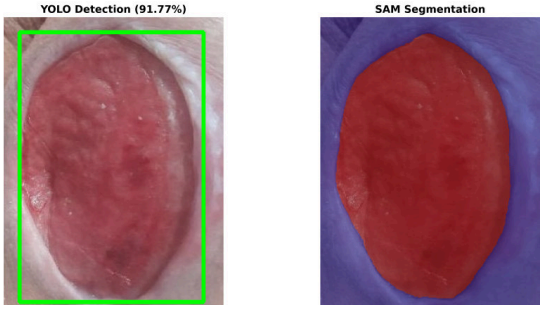


Fig. 3. YOLO Detection and Segmentation with SAM.

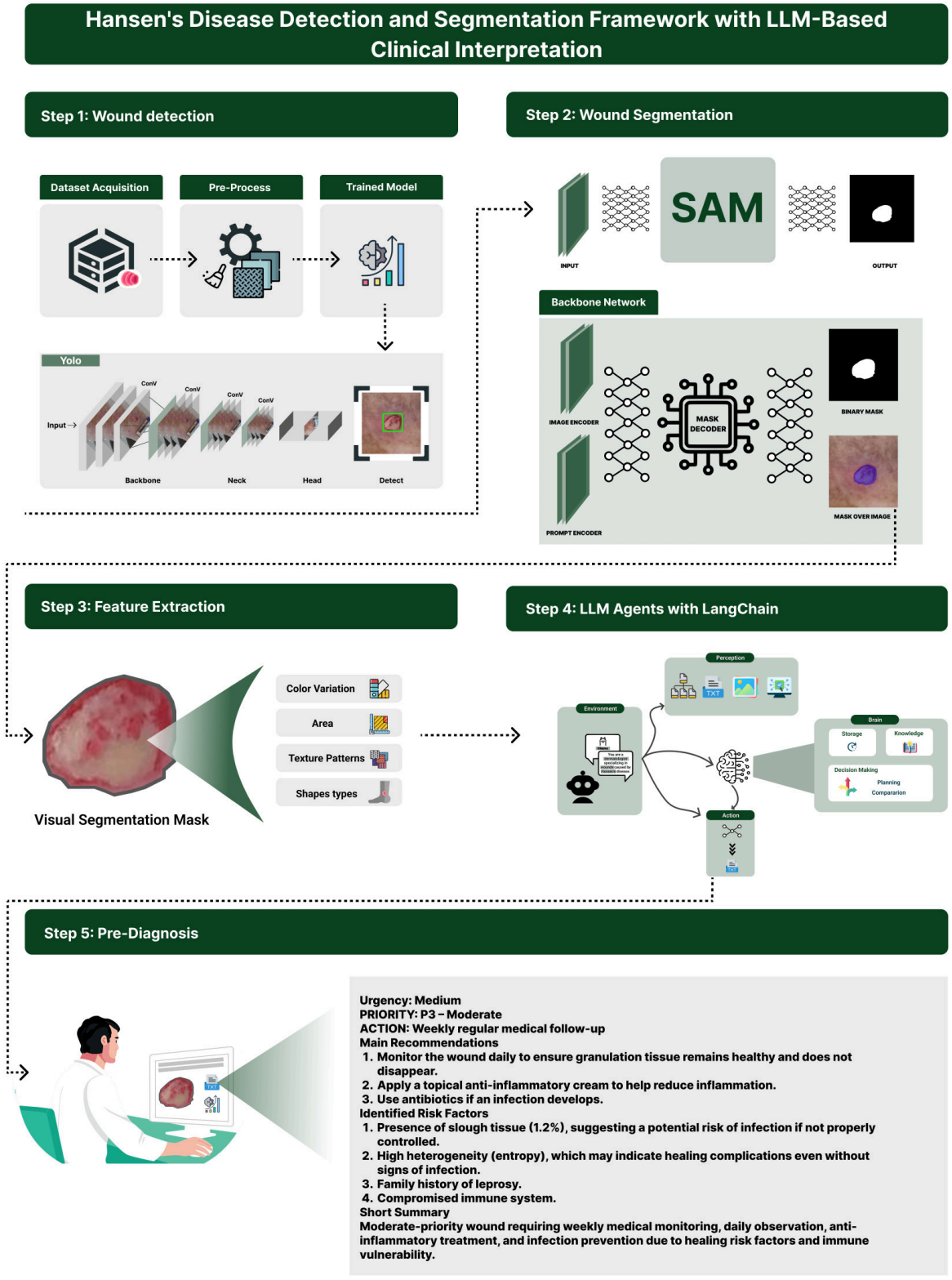


Fig. 1. Proposed methodology divided into five steps: Step1. Wound Detection; Step2. Wound Segmentation; Step3. Feature Extraction; Step4. LLM Agents with LangChain; Step5.Results Consolidation.

Step 3: Extraction of Visual and Geometric Features

After segmentation using SAM and the mask obtained from the process, the next step involved feature extraction from the lesion area, where visual and geometric features were extracted. Some of the features considered relate to size and shape, such as area and perimeter, as well as properties associated with color variation and texture patterns present in the segmented region. This information allows for a quantitative description of the wound, transforming visual aspects into structured data suitable for computational analysis.

Step 4: Analysis Using Language Models

In this step, several natural language models were used to interpret the features that had been extracted, organized into a structured format, and provided to the models. Among the models used were Qwen2.5 and gemma2:2b.

The features extracted from the lesion include the area, perimeter, color distribution, and texture, which were converted into a structured textual format that was also provided as input to the language models. Thus, this quantitative information served as the basis for the LLMs' inference process, enabling the models to correlate visual patterns and possible stages of the lesion with signs of progression and relevant clinical indicators.

The evaluation method adopted for the language models (LLMs) in this study, given the absence of external clinical validation, was based on an internal evaluation strategy relying on inter-model consensus. The two LLMs mentioned above were run locally via Ollama without relying on external infrastructure and were queried independently for each wound.

The diagnoses were compared using agreement metrics, including: classification consensus rate, standard deviation of the severity score, and agreement in the diagnosis of infection. This approach is justified by three main reasons: (i) the exploratory nature of the study, whose objective is to demonstrate the technical feasibility of the YOLO + SAM + LLMs pipeline for the automated analysis of chronic wounds caused by leprosy, rather than to replace clinical judgment; (ii) the lack, to date, of public benchmarks with structured clinical annotations for CO2Wounds-V2, whose test set is deliberately provided without labels; and (iii) the fact that LLMs are fed exclusively quantitative features derived from segmentation, such as color in the RGB, HSV and LAB, the RYKW tissue composition map, LBP and HOG texture descriptors, and geometric attributes such as circularity, solidity, and compactness, thereby eliminating any bias introduced by the raw image and making the analysis reproducible and auditable.

This stage sought to explore the interpretive capacity of the models in the integrated analysis of structured data and to provide clinical contexts, where a comparative analysis of the convergences and divergences among the responses generated by the agents was conducted. This comparison made it possible to assess the consistency of the generated hypotheses, identify patterns of agreement among the models, evaluate the risk of hallucination associated with them, and enable a more accurate and unified diagnosis.

Step 5: Clinical decision support

Finally, based on the diagnostic outputs generated by the LLMs, an interpretive report was produced to support clinical decision-making. The system provides risk stratification and identifies potential indicators of lesion progression. This study does not aim to replace medical evaluation, but rather to serve as a complementary tool supporting healthcare professionals in their decision-making process, offering an AI-based second opinion grounded in computer vision and natural language processing techniques.

V. Results and Discussion

This section presents the results obtained from the proposed YOLO+SAM pipeline applied to the CO2Wounds-V2 dataset, along with an analysis of the implications and limitations of the study and the performance of the LLM-based clinical interpretation stage.

A. Detection

The YOLOv8 model achieved strong detection performance, with a precision of 90.07% and an mAP@50 of 83.44%, indicating robust generalization on the detection task. The recall of 75.63% demonstrates a satisfactory capacity to identify regions of interest within the dataset. Although the primary focus of the evaluation lies on segmentation performance, these results highlight the feasibility of employing detection models in hybrid pipelines, where YOLOv8 serves as a reliable preliminary stage for wound segmentation and downstream analysis.

B. Segmentation

Table I presents a comparison between existing methods for Hansen disease lesion segmentation reported in the literature and the models evaluated in this study on the CO2Wounds dataset. The analysis was conducted using metrics widely used in the literature, including Precision, Recall, F1-score, and IoU.

TABLE I
COMPARISON OF SEGMENTATION METHODS ON THE CO2WOUNDS-V2 DATASET.

Citation	Method	Task	Prec. Value	Rec. Value	F1 Value	IoU Value	Source
This work	YOLO+SAM	Seg.	**80.30%**	74.70%	70.70%	57.80%	This work
[11]	COD2Net	Seg.	84.50%	81.90%	**80.30%**	82.50%	SOTA
[11]	VGG16	Seg.	68.70%	53.90%	50.60%	—	Paper
[11]	SegNet	Seg.	68.40%	51.50%	46.40%	—	Paper
[11]	MobileNetV2	Seg.	65.90%	70.70%	54.10%	—	Paper
[11]	U-Net	Seg.	82.70%	66.70%	64.30%	—	Paper

Among the segmentation methods reported in the literature, COD2Net achieves the highest performance, with an F1-score of 80.30% and an IoU ranging from 80.00% to 85.00%, and was considered state-of-the-art (SOTA) in the original study. Among the remaining architectures, U-Net attains the most competitive performance, with an F1-score of 64.30%, outperforming VGG16, SegNet, and MobileNetV2.

In summary, the proposed YOLO+SAM pipeline demonstrated high precision (80.30%), indicating a strong ability to avoid false positives. The recall of 74.70%, while satisfactory

in absolute terms, was comparatively lower than that of methods focused purely on segmentation, which may account for a higher rate of false negatives; consequently, the F1-score (70.70%) and IoU (57.80%) remained below the SOTA benchmark.

The results indicate that methods dedicated exclusively to segmentation, such as COD2Net, still achieve a better overall balance between precision and recall, whereas the YOLO+SAM pipeline exhibits greater metric variability, which may suggest sensitivity to heterogeneous characteristics across lesion images.

The detection-based strategy employed in this study YOLOv8 demonstrates strong capacity to locate wound regions, which can be leveraged in clinical settings where rapid identification of the lesion site is a priority, particularly in resource-limited environments with restricted access to specialized healthcare services.

Furthermore, the lower IoU achieved by the YOLO+SAM pipeline suggests that incorporating mask refinement strategies or fine-tuning the segmentation components could yield meaningful improvements in overall performance.

From a practical standpoint, the choice among the methods described in the literature will depend on the target application scenario. In clinical settings where minimizing false positives is a priority, approaches with higher precision are preferable. Conversely, for longitudinal monitoring applications, higher recall may be prioritized to avoid missing clinically relevant findings.

C. Robustness Under Varied Acquisition Conditions

Regarding the pipeline's robustness under the various conditions under which the images were acquired and used for model training, it is important to note that the CO2Wounds-V2 dataset subjects the pipeline to heterogeneous image acquisition conditions. The images were collected over approximately one year and nine months using smartphone cameras, resulting in variations in lighting, angle, focal length, and resolution across samples.

Challenges such as uneven lighting and indistinct lesion margins are common in the dataset, where the detection model achieved satisfactory results in detecting skin lesions caused by leprosy. However, it is important to note that the size of the dataset does not adequately represent the clinical variance observed across different populations—such as skin tones, lighting conditions, and disease stages—which can pose a challenge for the pipeline.

D. Evaluation of Large Language Models (LLMs)

The language models were evaluated using the Ollama framework, deployed in a local environment to ensure standardized execution and controlled experimental conditions, including variables such as temperature and maximum token limits. The Qwen2.5 and Gemma2:2b models were assessed, both integrated into the proposed multimodal pipeline.

The evaluation of these models was based on two main criteria: (i) structural agreement with the internally generated consensus, and (ii) textual similarity between the produced descriptions. It is important to emphasize that the consensus used as reference was derived from the system itself, through inferences performed on a separate batch of images; therefore, this evaluation focuses on assessing internal consistency and structural stability, rather than serving as external clinical validation.

Table II presents the results obtained. In summary, the Qwen2.5 model demonstrated greater structural agreement with the internal consensus (Exact Match = 100.00%), whereas Gemma2:2b achieved an average agreement of 76.30%. In terms of textual similarity, both models performed comparably (F1 = 33.00%), indicating substantial lexical variation in the generated clinical descriptions. The low BLEU scores ($<$2.00%) further suggest that differences between the models occur primarily at the level of textual formulation, while the underlying diagnostic structure remains consistent.

TABLE II
QUANTITATIVE COMPARISON OF LLMS RUNNING ON OLLAMA.

Model	Exact Match Value	F1 Value	BLEU Value	Time (ms) Value
gemma2:2b	76.00%	33.00%	2.00%	76185.5
Qwen2.5	**100.00%**	34.00%	2.00%	97724.2

It can also be observed that there was a significant difference in computational cost. Gemma2:2b had a shorter average inference time (approximately 76 s) compared to Qwen2.5 (approximately 98 s). These results indicate a trade-off between greater structural alignment with the internal consensus and higher computational cost, and that depending on where the model is being run, the time tends to be shorter or longer.

VI. CONCLUSION

The proposed study is intended to serve as an auxiliary tool to support clinical decision-making, and not as a standalone diagnostic mechanism, in accordance with Federal Council of Medicine (CFM) Resolution No. 2,454/2026, which regulates the use of Artificial Intelligence in medical practice in Brazil.

The development of the model focused on the detection and segmentation of chronic leprosy wounds yielded interesting results, given the accuracy (80.3%) achieved by the proposed architecture. The integration of LLMs demonstrated the feasibility of building multimodal decision support systems, in which discrepancies between models can be used as an additional indicator of clinical uncertainty.

Furthermore, the choice of model must therefore balance structural performance, computational cost, operational stability, and the application context. The absence of external clinical validation is thus recognized as a limitation of this study, and future work should incorporate specialized clinical annotation to enable assessment of agreement with the human gold standard; the framework requires further study before any clinical application.

Despite the promising results, the study has some limitations related to the size and diversity of the dataset; the

CO2Wounds-V2 dataset contains 764 images from 96 patients, which does not adequately represent the clinical variability observed across different populations, skin tones, lighting conditions, and disease stages. Furthermore, the images were captured by smartphones under various conditions, which may introduce noise and uncontrolled variations in the training of the models.

As future work, we propose (i) to fine-tune the models to address the specific nature of the miscarriage, (ii) to reconstruct the wound in 3D, (iii) to track its development over time, (iv) to include larger multimodal datasets, and (v) to apply clinical RAG.

REFERENCES

[1] R. Zhao, H. Liang, E. Clarke, C. Jackson, and M. Xue, "Inflammation in chronic wounds," International Journal of Molecular Sciences, 2016.

[2] A. Guevara, V. A. Vicente, B. J. F. S. Lima, A. F. Nery, F. Hagen, and R. C. Hahn, "Chromoblastomycosis–leprosy co-infection in Central West Brazil: presentation of three cases and literature review," 2022. Mycopathologia (2022) 187:363–374 doi:/10.1007/s11046-022-00646-5

[3] A. C. Oliveira, D. M. Rocha, S. M. G. Bezerra, E. M. L. R. Andrade, A. M. R. Santos, and L. T. Nogueira, "Quality of life of people with chronic wounds," Revista Brasileira de Enfermagem, 2019, doi: 10.1590/1982-0194201900027.

[4] M. G. Araújo, "Hanseníase no Brasil," Revista da Sociedade Brasileira de Medicina Tropical, vol. 36, no. 3, pp. 373–382, May–Jun. 2003.

[5] World Health Organization, "Global leprosy update," WHO, Geneva, Switzerland, 2023.

[6] D. C. J. Lima and G. O. Paes, "Infection assessment tools for acute and chronic wounds: a scoping review," Revista da Escola de Enfermagem da USP, vol. 59, e20240392, 2025, doi: 10.1590/1980-220X-REEUSP-2024-0392en.

[7] K. Sanchez, C. Hinojosa, O. Mieles, C. Zhao, B. Ghanem, and H. Arguello, "CO2WOUNDS-V2: Extended chronic wounds dataset from leprosy patients," 2023.

[8] Y. O. A. Rodrigues, J. J. C. Nascimento, O. S. L. Junior, A. G. Marques, G. F. B. Severiano, C. M. J. M. Dourado Junior, and L. F. F. Souza, "SDA-Detection Melanoma: Deep approach system for detection and segmentation in melanoma images using fine-tuning," in Proc. Brazilian Symposium on Computing Systems Engineering (SBESC), 2024.

[9] A. G. Marques, "Nova abordagem de tumor cerebral para segmentação através de células computacionais aplicadas à fusão de dados para classificação de imagens de RM com modelo de processamento de linguagem natural," M.S. thesis, Mestrado em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Paracuru, Brazil, 2025.

[10] L. F. F. Souza, "Divisible cell-segmentation: Uma nova abordagem fine-tuning adaptativa para segmentação em imagens médicas por meio de células computacionais sintéticas," Ph.D. dissertation, Programa de Pós-Graduação em Engenharia de Teleinformática, Universidade Federal do Ceará (UFC), Fortaleza, Brazil, 2023.

[11] B. Monroy, K. Sanchez, P. Arguello, J. Estupiñán, J. Bacca, C. V. Correa, L. Valencia, J. C. Castillo, O. Mieles, H. Arguello, S. Castillo, and F. Rojas-Morales, "Automated chronic wounds medical assessment and tracking framework based on deep learning," Computers in Biology and Medicine, 2021.

[12] E. Brügge, S. Ricchizzi, M. Arenbeck, M. N. Keller, L. Schur, W. Stummer, M. Holling, M. H. Lu, and D. Darici, "Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial," 2024. Brügge et al. BMC Medical Education (2024) 24:1391 doi: 10.1186/s12909-024-06399-7

[13] S. Shool, S. Adimi, R. S. Amleshi, E. Bitaraf, R. Golpira, and M. Tara, "A systematic review of large language model (LLM) evaluations in clinical medicine," BMC Medical Informatics and Decision Making, vol. 25, p. 117, 2025, doi: 10.1186/s12911-025-02954-4.

[14] F. Tustumi, N. A. Andreollo, and J. E. Aguilar-Nascimento, "Future of the language models in healthcare: The role of ChatGPT," 2023. doi:10.1590/0102-672020230002e1727

[15] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," Nature Communications, vol. 15, 2024, doi: 10.1038/s41467-024-44824-z.

[16] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment Anything Model for medical image analysis: an experimental study," Medical Image Analysis, vol. 89, 2023, doi: 10.1016/j.media.2023.102918.

[17] J. R. N. Fernandes, A. S. Teles, T. R. S. Fernandes, L. D. B. Lima, S. Balhara, N. Gupta, and S. Teixeira, "Artificial Intelligence on Diagnostic Aid of Leprosy: A Systematic Literature Review,"

[18] M. E. Rayed et al., "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," Informatics in Medicine Unlocked, 2024. doi: 10.1016/j.imu.2024.101504

[19] K. Sanchez, C. Hinojosa, C. V. Correa, P. Arguello, B. Monroy, O. Mieles, F. Rojas, C. Zhao, B. Ghanem, H. Arguello Fuentes, "CO2Wounds-V2: Extended Chronic Wounds Dataset From Leprosy Patients," Mendeley Data, V2, 2024. doi: 10.17632/s2w7rjwz49.2

[20] LI, Shuxin; RENICK, Paul; SENKOWSKY, Jon; NAIR, Ashwin; TANG, Liping. Diagnostics for Wound Infections. Wound Repair and Regeneration, [S. l.], v. 29, n. 2, p. 317-335, 2021.

[21] Conselho Federal de Medicina, "Resolução CFM nº 2.454, de 11 de fevereiro de 2026: Normatiza o uso da inteligência artificial na medicina," Diário Oficial da União, Brasília, Brasil, 27 fev. 2026.

[22] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 2019.

[23] A. G. Marques, M. V. C. de Figueiredo, J. J. da Costa Nascimento, C. T. de Souza, C. M. Jaborandy de Mattos Dourado Junior, V. H. C. de Albuquerque, and L. F. de Freitas Souza, "New Approach Generative AI Melanoma Data Fusion for Classification in Dermoscopic Images with Large Language Model," in *Proc. 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2024, pp. 1–6, doi: 10.1109/SIBGRAPI62404.2024.10716298.

[24] G. V. M. Amaral, "Segmentação semântica de imagens dermatoscópicas de lesões de pele utilizando aprendizado profundo," Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação), Instituto de Computação, Universidade Federal de Alagoas, Maceió, AL, Brasil, 2024.

[25] A. F. de Araujo, J. M. R. S. Tavares, R. B. Oliveira, N. Marranghello, A. S. Pereira, and R. B. Rossetti, "Uma metodologia híbrida para segmentação de lesões de pele," in *Anais do XXIV Congresso Brasileiro de Engenharia Biomédica (CBEB)*, Uberlândia, MG, Brasil, 2014.

A Novel Benchtop Workflow Based in 3D Scanning for Angular Measurement in Femoral Derotational Osteotomies

Lucas P. Acosta
Instituto Superior de Investigaciones Biológicas (INSIBIO) CONICET
National University of Tucumán
Tucumán, Argentina
lacosta@herrera.unt.edu.ar

Milagros Fernandez Jure
Universidad Nacional de Tucumán
Tucumán, Argentina
milagrosfernandezjure@gmail.com

Ramiro Olleac
Hospital Nicolás Avellaneda - Ortopedia Infantil
Tucumán, Argentina
ramiroolleac@hotmail.com

D. G. Goroso
Center for Technological Research, University of Mogi das Cruzes.
Mogi das Cruzes, São Paulo, Brazil.
danielg@umc.br

F. D. Farfan
Instituto Superior de Investigaciones Biológicas (INSIBIO) CONICET
National University of Tucumán
Tucumán, Argentina
ffarfan@herrera.unt.edu.ar

***Abstract*— To address the limitations of high error rates and radiation exposure associated with traditional visual estimation and fluoroscopy in femoral derotational osteotomies (FDO), this study proposes a novel, non-ionizing angular measurement method using structured-light 3D scanning. Utilizing a 3D-printed femur model with predefined pin channels (0° to 35°) and custom-designed 2.5 mm polylactic acid (PLA) cylindrical tracking markers, 40 trial scans were captured using an Artec Eva 3D scanner. The digitized point clouds were reconstructed in Artec Studio, and the axial angular differences were mathematically calculated in Blender 4.2. Bland-Altman analysis against the reference computer-aided design (CAD) model revealed a minor systematic measurement bias of -1.13° ± 2.66°, with 95% limits of agreement ranging from -6.34° to 4.08°. These results demonstrate the feasibility of this benchtop 3D scanning workflow, providing a radiation-free alternative whose repeatability and clinical translation warrant further investigation.**

Keywords—3D scanner, osteotomy, angular measurement.

I. INTRODUCTION

Proper alignment of the lower limbs is fundamental to minimizing muscular work and joint wear during locomotion. Within this scope, three fundamental concepts are distinguished:

Femoral version: Normal and static rotational variation of a specific bone. It is formally defined as the angular difference between the axis of the femoral neck and the transcondylar axis of the knee[1] [2].

Rotation: The result of the dynamic interaction between bones, joints, and muscles.

Bone torsion: Static structural deformation of a bone around its longitudinal axis.

Under normal conditions, the femur exhibits anteversion (the femoral head is rotated anteriorly relative to the condylar axis). At birth, this angle ranges between 25° and 50° and progressively decreases with growth, stabilizing between 10° and 15° around 16 years of age.

Evaluation in the frontal plane distinguishes critical parameters:

Anatomical axis: A straight line from the piriform fossa to the center of the knee.

Mechanical axis: A straight line from the center of the femoral head to the center of the knee.

Neck-shaft angle (cervicodiaphyseal angle): Formed between the axis of the femoral neck and the anatomical axis, with a normal value between 125° and 131°.

A. Femoral Torsional Deformity

Torsional anomalies are clinically defined when femoral version values deviate significantly from the mean (by more than 2 standard deviations).

Types of deformity:

Excessive femoral anteversion (internal torsion): This is the most frequent presentation in pediatric consultations; it causes patellofemoral instability, anterior knee pain, and femoroacetabular impingement.

Femoral retroversion (external torsion): A condition always considered abnormal. It can be physiological (bilateral, symmetrical, and generally self-limiting) or pathological (congenital, or acquired due to conditions such as Perthes disease, cerebral palsy, or poliomyelitis). It is based on clinical history, gait analysis, and the Staheli rotational profile, which measures parameters such as the foot progression angle and hip rotations.

Excessive anteversion is compensated for by an in-toeing gait. Retroversion is compensated for by an out-toeing gait.

For high-precision diagnosis, cross-sectional imaging is utilized (Computed Tomography [CT] and Magnetic Resonance Imaging [MRI]), which are considered the gold standard. MRI is notable for avoiding ionizing radiation and allowing the evaluation of soft tissues and cartilage. Furthermore, image processing software such as OsiriX facilitates digital angle measurement by superimposing axial slices of the femoral neck and condyles.

B. Femoral Derotational Osteotomy (FDO)

When anteversion deformities (>25°–30°) persist beyond 8 years of age, cause frequent falls, or impair daily function, surgical treatment is indicated: femoral derotational

osteotomy (FDO). Its objective is to restore anteversion to a physiological value of approximately 15°.

Distal osteotomies prevent misalignment because they are performed along the mechanical axis, whereas proximal osteotomies are performed transverse to the anatomical axis to avoid unwanted translation of the bone.

C. *Intraoperative Measurement Methods and New Technologies*

The main challenge of FDO is the precise measurement of rotational degrees in the operating room. Traditionally, surgeons have relied on visual estimation and two-dimensional fluoroscopy, methodologies that exhibit a high margin of error.

To overcome these limitations, several intraoperative technological alternatives have been developed:

Computer-Assisted Navigation Systems and Sensors: Computer-assisted freehand tracking using C-arm imaging or electromagnetic sensors coupled to Kirschner wires (K-wires), enabling real-time monitoring of the derotation angle.

Mobile Devices (Smartphones): The use of internal gyroscopic sensors or dedicated applications (such as Navi-Gator) to photographically record the angulation of the pins, proving to be accessible options that are more precise than purely visual estimation.

Robotics and Augmented Reality: Systems such as Skywalker or robot-assisted navigation technologies with augmented reality allow the alignment of 3D preoperative planning with the patient's anatomy in real time, enhancing surgical safety.

Hybrid Operating Rooms: Facilities that combine CT/MRI and real-time 3D reconstruction to guide surgical procedures, optimizing precision but carrying the disadvantages of high cost and radiation exposure.

3D Scanning and Digitization: Based on the principle of stereoscopic triangulation, structured-light scanners are emerging as fast and precise tools for documenting exposed surfaces and instruments in the operating room without resorting to postoperative radiation.

In this work, an analysis performed on a 3D-printed bone model using structured-light-based 3D scanning technology and its contrast with reference values will be presented.

Objective: To propose an innovative angular measurement method using 3D scanning that can be implemented in future surgical practice.

II. Materials and Methods

A. *Artec EVA*

Artec Eva is a high-quality professional 3D scanner used in a wide variety of fields. This is a non-contact scanner that operates on structured-light technology; it is portable, lightweight, and does not emit any harmful radiation.

The scanner projects normal visible light stripe patterns onto the object, and their deformation is captured by three cameras: two cameras on different edges of the scanner and one positioned in the middle, surrounded by flash LEDs to receive texture information. During scanning, Artec Eva captures up to 16 3D frames per second and transfers them to Artec Studio, its proprietary software. The latter creates a texture atlas, and the user receives a color-textured 3D mesh composed of triangles. The final output generated in Artec Studio is a commonly used 3D data format, such as stereolithography (.stl) [3].

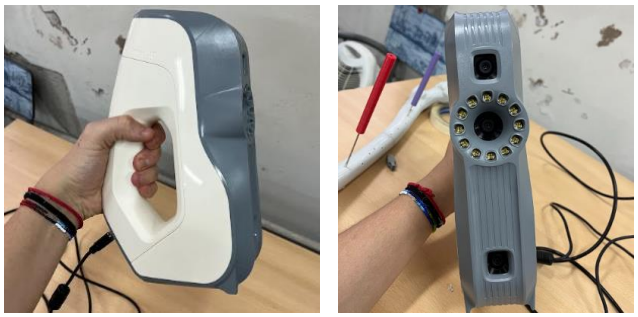


Figure 1. Artec EVA scanner.

Artec Eva has been implemented in various medical fields, such as orthotics and prosthetics for the head and neck region [4], injury documentation during autopsies [5], and facial and body scanning [6]. Furthermore, it has demonstrated superior accuracy in facial scanning applications due to being a mobile rather than a stationary system [3].

B. *3D Printing in the Medical Industry*

3D printing is the process of forming a solid object from a computer-generated digital model. While various technologies exist for manufacturing a three-dimensional object, the fundamental concept lies in the successive addition of layers of material until the final product is obtained.

Acquisition of a three-dimensional model of the object: This can be achieved either through 3D scanning or computer-aided design, typically using CAD software such as SolidWorks, Blender, among others.

Preparation of the model in the appropriate format: Generally STL, where its geometric information is stored.

Transformation of the model file into a format the printer understands: Usually G-code.

Finally, certain printing parameters are specified, and fabrication begins.

C. *Computer-Aided Design*

Blender® software version 4.2 was used for the computer-aided design of all three-dimensional elements in this project: the cylindrical markers and the necessary adaptations to the femoral bones. Additionally, this software offers useful tools to obtain angular measurements with any desired precision.

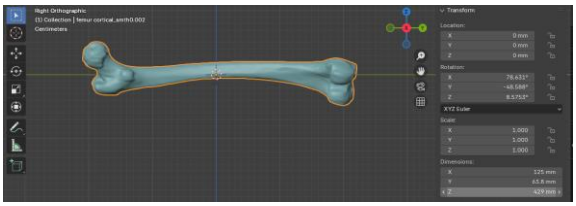


Figure 2. Femur view in Blender ® workspace.

The process begins with the selection of a femur file in STL format, downloaded freely and at no cost from the online platform "GrabCAD". It is imported into Blender, and its characteristics are analyzed. As shown in Figure 2, it has a length of 429 mm and exhibits normal anatomical features. In order to place the pins through the bone, three groups of intersections are planned that differ from each other by 20 and

30 degrees from the axial view of the femur. To achieve this, cylinders of 2.5 mm in diameter are created. From the axial view, these cylinders are parallel, pass through the bone, and are guided by a plane to ensure they remain parallel to one another. Subsequently, using Blender's Boolean operations, a "difference" function is performed between the bone and the cylinders, resulting in holes that pass completely through the bone.

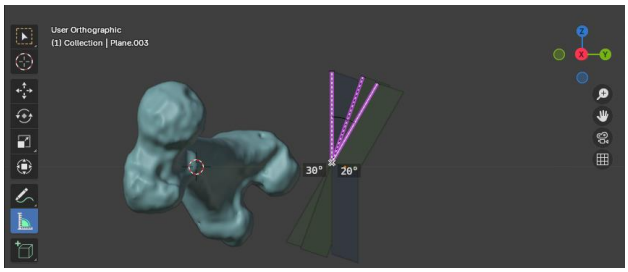


Figure 3. Planes for boolean operations in Blender ®.

According to various authors, the scanner faces certain difficulties when scanning shiny and transparent objects. This gave rise to the idea of implementing colored cylindrical markers to be placed around the pins. These markers are designed to be small enough not to hinder surgical techniques, while facilitating the spatial localization of the pins (which are metallic) in the 3D space. They are manufactured from PLA.

Although the markers are non-invasive, the pins themselves are inserted into the patient's bones, placing them within centimeters of surrounding tissues. Consequently, the markers are also in close proximity. For this reason, we aimed to design the markers with the highest possible precision, taking into account the dimensions of the pins, to ensure they remain securely fixed without risking displacement or falling off while the surgeon moves the patient. Since 2.5 mm diameter pins are used, the markers are designed with an internal diameter of 2.5 mm, an external diameter of 10 mm, and a height of 100 mm.

D. Printing and Scanning

An Ender 3 fused deposition modeling (FDM) printer with a printing precision of 0.1 mm was employed. The fabrication material was PLA filament: white for the femoral bones, and red, lilac, and green for the markers.

Forty random measurements were planned (20 for each bone). Color coding was used on the holes to visually identify which exact angular difference corresponded to each pair of pins.

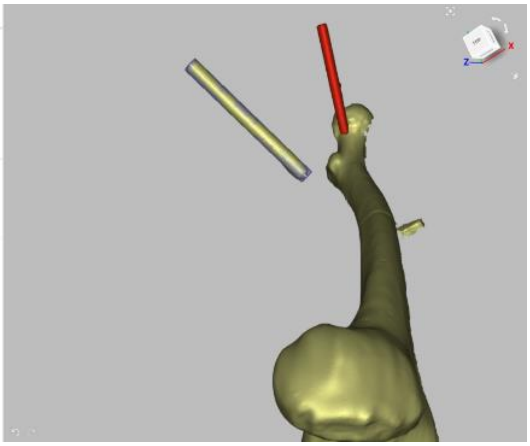


Figure 4. Fussion in Artec Studio. Marker reconstruction.

Each measurement consisted of two scanning sessions from multiple perspectives to achieve a better reconstruction. Processing in Artec Studio involved:

Eraser: Manual selection and removal of noise and the base to isolate the object of interest.

Alignment: Automatic or manual alignment of the point clouds from both sessions by selecting spatial correspondence points.

Fusion: Application of fast mesh simplification and smoothing.

Object Construction: Creation of perfect cylinders over the reconstructed surface of the markers to precisely identify their axis and longitudinal direction.

Translation and Rotation: Reorientation of the femur to align its longitudinal axis with the X-axis, ensuring that the axial measurement plane coincides exactly with the coordinate plane (Y,Z).

E. Angular Measurement and Analysis

The processed scans were imported into Blender. To measure the angular difference in the axial plane without being affected by potential sagittal or coronal tilts, face normal visualization was enabled for the virtual cylinders constructed in Artec Studio. The length of these normal lines was extended until they intersected in the axial plane. Three exact mathematical vertices were created (one on each normal vector and one at the intersection of both normal lines). The "measurement tools" extension was used to automatically register the angle generated by the three vertices with a precision of two decimal places.

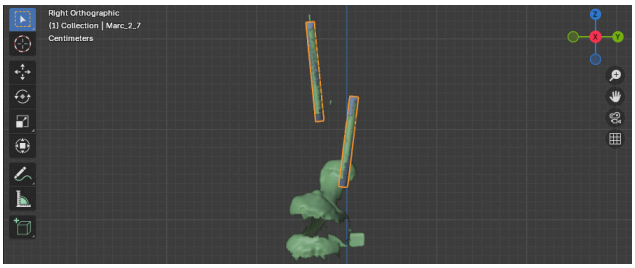


Figure 5. Angle measurement in Blender ®.

An initial descriptive analysis (mean, median, variance, standard deviation, maximum, and minimum) was defined for each angular control group (0°, 10°, 20°, 25°, 30°, and 35°). Data normality was planned to be verified using the Shapiro-Wilk test and histograms. The Bland-Altman statistical methodology was selected to compare the 3D-measured values against the exact theoretical values of the computer model, evaluate precision, identify bias, and calculate the limits of agreement with a significance level α=0.05.

III. Results and Discussion

The markers printed with an internal diameter of approximately 3 mm did not remain securely attached to the pins, whereas those with the same diameter as the pins 2.5 mm, provided a snug fit and remained securely fixed. It is crucial for these markers to remain stationary and for the diameter to be appropriate so that the pins do not easily shift, which would introduce errors in the spatial localization of the markers. Therefore, it is considered essential to know the precise dimensions of the pins to design custom-fit markers.

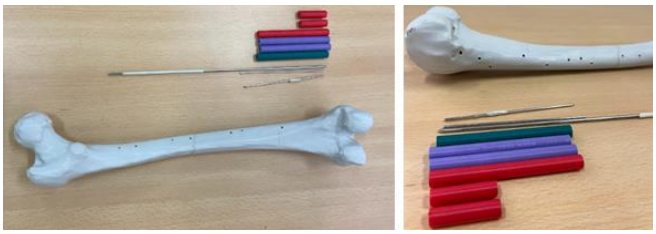


Figure 6. 3D printed femur with markers and wires.

The frequency distribution in Figure 7 shows the distribution of reconstructed angular values across the six target angles alongside their respective Gaussian fits. While the distributions confirm unimodal behavior across trials, the shift in the central tendency relative to nominal references highlights the magnitude-dependent variation summarized in Table 1.

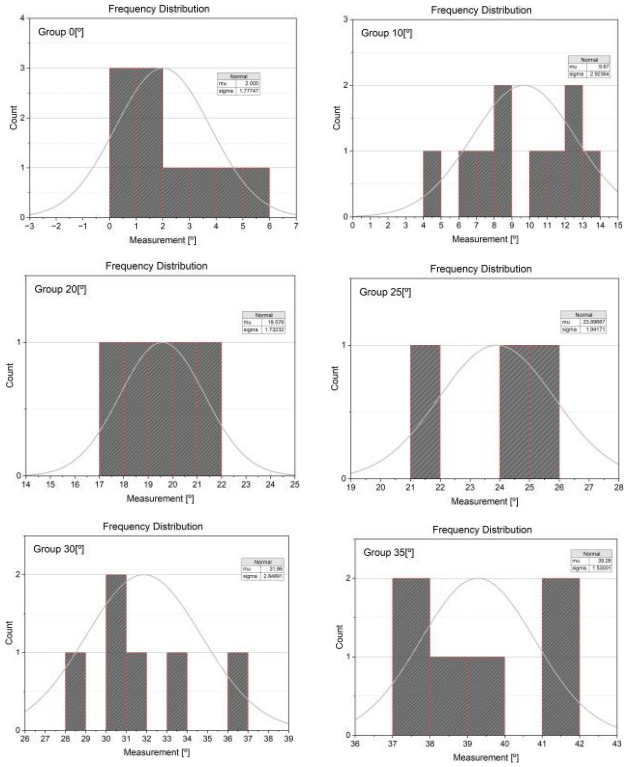


Figure 7. Data distribution histogram with normal curve distribution.

TABLE I. STATISTICS IN GROUPS.

Group	*Mean [°]*	*SD [°]*	*Max Value [°]*	*Min Value [°]*	*MAE [°]*	*RMSE [°]*
0°	2.01	1.78	5.46	0	2.00	2.62
10°	9.67	2.92	13.47	4.28	2.33	2.71
20°	19.58	1.73	21.58	17.09	1.32	1.61
25°	23.90	1.94	25.15	21.66	1.20	1.93
30°	31.86	2.85	36.73	28.89	1.33	1.79
35°	38.78	1.31	41.05	37.59	3.78	3.96

Bland-Altman Plot

To graphically compare the differences between the measured and exact values, the Bland-Altman statistical technique was used. Additionally, this technique helps observe the precision of the measurements obtained with the 3D scanner. First, it was necessary to confirm that the differences between the measured and reference values followed a normal distribution; thus, the Shapiro-Wilk test was applied to them. The hypothesis being tested is whether the differences between the exact and measured values follow a normal distribution; therefore, the null hypothesis H_0 is that they are normally distributed, and the alternative hypothesis H_1 is that they do not follow a normal distribution. The red line represents the bias (the mean of the differences), which has a value of -1.13º. This demonstrates a discrepancy between the measured and exact values. Specifically, the 3D scanning method measures the value as 1.13º less than it actually should be implying a bias that must be taken into account when evaluating the validity of this measurement method. The green and blue lines represent the lower and upper limits of agreement, respectively. These are defined as the mean difference ±1.96 times the standard deviation of the differences. The upper limit (blue line) has a value of 4.08º and the lower limit (green line) is -6.34º. These limits indicate the interval within which 95% of the differences fall, an assertion supported by their normal distribution.

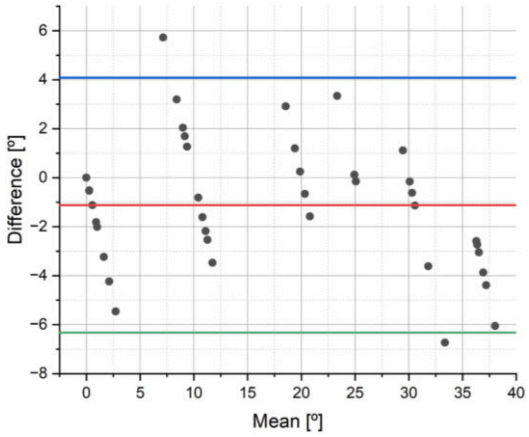


Figure 8. Bland Altman plot.

The progression of the elapsed time across consecutive trials is illustrated in Figure 9, showing a steep reduction from an initial duration exceeding 27 min to a plateau between 7 and 10 min after the 20th trial. This steep downward trend reflects a pronounced learning curve, suggesting that operating speed, manual point-cloud processing in Artec Studio, and vertex-based alignment in Blender are heavily operator-dependent.

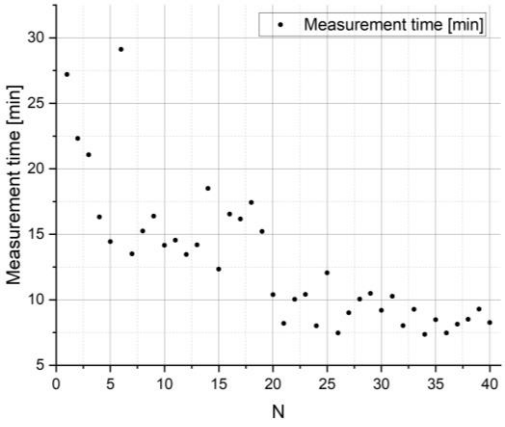


Figure 9. Measurement time.

IV. CONCLUSION

The Artec Eva 3D scanner allows for high precision and accurate marker reconstruction, we succeeded in measuring

the angles between two pins. The Artec Studio and Blender® software were fundamental to performing all measurements. In orthopedic and traumatological applications, the 3D scanner is potentially superior to other technologies due to its portability, ease of use, structured-light operating mode (which is harmless to patients), and excellent resolution, allowing for color-rich, highly textured 3D scans.

The overall measurement duration decreased by approximately 3.5-fold between the initial and concluding trials, demonstrating a clear learning effect throughout the 40 scans. Therefore, the progressive decrease in processing time—along with the variability observed in the angular errors—suggests that operator proficiency across both software environments is a critical variable that directly conditions the reliability and reproducibility of the method, reinforcing the need for standardized multi-observer protocols.

The mean difference between the proposed 3D scanning method and the exact values of the computer model is -1.13º ±2.66º. Comparing the data reported in [7] with the results obtained in this study, we consider the 3D scanning method to be more accurate than visual estimation, and it even yields better standard deviations than the differences measured using the osteotomy template method. However, it does not achieve the high level of precision obtained when using the iPhone 4 gyroscope. Furthermore, the 3D scanning method proposed in this project presents a standard deviation of differences very close to that reported in [8], with values of 2.77º and 2.66º, respectively. This is highly encouraging, especially when compared to a study that implemented measurements based on MRI and CT scans.

For future work, the test using the Artec EVA inside the surgery room in order to obtain real measurements is being analyzed. The aims to evaluate will be: easiness of use during the surgical practice, workflow using the measurement software, reliability of the values in relation with the test results. Although the mean bias of -1.13° reflects low systematic deviation, the limits of agreement (-6.34° to 4.08°) represent a dispersion span exceeding 10°. In the context of femoral derotational osteotomies, where clinical correction margins are typically aimed within ±3° to ±5°, this benchtop variability must be narrowed through automated alignment algorithms before intraoperative application can be pursued.

ACKNOWLEDGMENT

This research work was supported by:

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) Argentina. RESOL-2021-154-APN-DIR#CONICET.

REFERENCES

[1] Viste, A., Fessy, M.-H., & Chaudier, P. (2024). Anomalías de torsión del esqueleto de la extremidad inferior. EMC - Podología, 26(1), 1–10. https://doi.org/10.1016/S1762-827X(23)48674-9

[2] Tamimi Mariño I., Rojas Tomba F. (2020). Alteraciones rotacionales del miembro inferior. In Manual del Residente. Sociedad Española de Ortopedia y Traumatología. Recuperado el 22 de septiembre de 2024, de https://www.secot.es/actualidad-el-manual-del-residente-de-cot-2020

[3] Modabber, A., Peters, F., Kniha, K., Goloborodko, E., Ghassemi, A., Lethaus, B., Hölzle, F., & Möhlhenrich, S. C. (2016). Evaluation of the accuracy of a mobile and a stationary system for three-dimensional facial scanning. Journal of Cranio-Maxillofacial Surgery, 44(10), 1719–1724. https://doi.org/10.1016/j.jcms.2016.08.008

[4] Krajňáková, V., Michalíková, M., Hudák, R., Živčák, J., & Barcalová, M. (2019). Application of the artec eva scanner for orthotics in practice. Lekar a Technika, 49(3), 92–96. https://doi.org/10.14311/CTJ.2019.3.04

[5] Sivanandan J, Liscio E. Evaluación del escaneo 3D con luz estructurada utilizando Artec Eva para la documentación de lesiones durante la autopsia. J Assoc Crime Scene Reconstr. 2016;21:5-13

[6] Hanaphy, P. (12 de Agosto de 2024). Escaneo corporal 3D de la A a la Z. Artec 3D. https://www.artec3d.com/es/learning-center/3d-body-scanner

[7] Graham, D., Suzuki, A., Reitz, C., Saxena, A., Kuo, J., & Tetsworth, K. (2013). Measurement of rotational deformity: Using a smartphone application is more accurate than conventional methods. ANZ Journal of Surgery, 83(12), 937–941. https://doi.org/10.1111/ans.12371

[8] Häller, T. V., Schenk, P., Jud, L., Hoch, A., Götschi, T., & Zingg, P. O. (2021). Consistency of 3D femoral torsion measurement from MRI compared to CT gold standard. BMC Musculoskeletal Disorders, 22). https://doi.org/10.1186/s12891-021-04633-7

An Interactive LLM-Based Approach to Exploratory Data Analysis Evaluated on Open Health Datasets

Fernando Rezende Zagatti
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
Federal University of São Carlos (UFSCar)
São Carlos, Brazil
fzagatti@cti.gov.br

André Gomes Regino
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
Centro Universitário de Jaguariúna (UniFAJ)
Jaguariúna, Brazil
aregino@cti.gov.br

Filipe Loyola Lopes
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
Centro Universitário Max Planck (UniMAX)
Indaiatuba, Brazil
flopes@cti.gov.br

Gilson Yuuji Shimizu
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
gyshimizu@cti.gov.br

Matheus Bernardelli de Moraes
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
mbmoraes@cti.gov.br

Rodrigo Bonacin
Center for Information Technology Renato Archer (CTI Renato Archer)
Campinas, Brazil
rbonacin@cti.gov.br

Daniel Lucrédio
Federal University of São Carlos (UFSCar)
São Carlos, Brazil
daniel.lucredio@ufscar.br

Helena de Medeiros Caseli
Federal University of São Carlos (UFSCar)
São Carlos, Brazil
helenacaseli@ufscar.br

***Abstract*—This paper introduces DFChat, an open-source tool designed to facilitate interactive Exploratory Data Analysis (EDA) over Pandas DataFrames using Large Language Models (LLMs). By integrating Retrieval-Augmented Generation (RAG) and rule-based post-processing, DFChat allows users to ask questions about datasets in natural language and receive insights through text, statistics, and visualizations. A comparative study was conducted using five healthcare datasets and evaluated against two LLM-based tools (Data Formulator and PandasAI) using four statistical measures (Kendall's τ, Spearman's ρ, Kendall's W, and ICC). Results show that DFChat outperforms the compared tools, providing more accurate, complete, and interpretable responses.**

***Index Terms*—large language model, exploratory data analysis, retrieval-augmented generation, natural language**

I. Introduction

Open health datasets play a crucial role in scientific and technological advancement and in public health by facilitating scientific collaboration, enriching research, and expanding analytical capacity to support decision-making [1]. Exploratory Data Analysis (EDA) is a fundamental step in achieving these benefits, responsible for uncovering patterns, identifying anomalies, and providing initial insights [2]. Usually, it is a time-consuming task that requires technical knowledge of programming, as well as proficiency in libraries for data exploration and visualization [3].

We thank the National Council of Technological and Scientific Development (CNPq), Brazil, grant #405352/2021-2.

In recent years, advances in natural language have opened new possibilities for interaction between humans and computational systems with the emergence of Large Language Models (LLMs). These models can interpret and generate natural language with a high degree of fluency and contextual understanding [4], making them ideal candidates to serve as intelligent interfaces between users and data analysis tools.

In this work, we extend and evaluate an approach that builds on the interaction capabilities of LLMs to guide users through the EDA process [5]. The approach enhances LLM effectiveness through Retrieval-Augmented Generation (RAG), which automatically enriches the context with metadata and statistics extracted from the dataset. During each interaction, DFChat dynamically analyzes the response generated by the LLM, returning descriptive summaries, visualizations, and insights in natural language.

The approach presents advances in the use of generative artificial intelligence to boost data analysis by enabling LLMs to execute EDA tasks directly on Pandas DataFrames. Through a benchmark of five real-world health datasets and a multi-criteria evaluation conducted by experts, DFChat consistently outperformed two baseline tools (Data Formulator [6] and PandasAI[1]) in terms of Relatedness, Error presence, Clarity, and Coverage. These findings were supported by statistical correlation and agreement measures, which indicated high

[1]https://github.com/sinaptik-ai/pandas-ai

consistency among annotators. Additionally, we discuss the challenges, limitations, and potential future directions for enhancing the integration between LLMs and DataFrames.

II. Related work

The integration of LLMs into data science pipelines has become increasingly popular, particularly in automating code generation using natural language. This trend has given rise to a variety of systems designed to address the gap between natural language interaction and specialized data manipulation [7], [8]. In this context, the following papers are related work on various applications of LLMs in tasks such as data transformation, standardization, and visualization.

Wang, Thompson and Lee [6] propose the Data Formulator, an approach for creating visualizations through a paradigm called concept binding, which separates the visualization intent (expressed in natural language or examples) from the technical steps of data transformation. The tool utilizes an AI agent to automate these transformations, enabling users to create complex visualizations without requiring technical programming skills. A study with participants demonstrated that the system facilitates the creation of graphs that would typically require structured manipulations with libraries such as Pandas.

Fan et al. [9] propose DataChat, a conversational agent for searching and exploring scientific datasets. The system utilizes knowledge graphs to represent relationships between publications, terms, and institutions, in conjunction with the GPT-3.5-turbo model, which converts questions into Cypher queries for a Neo4j database. This structure enables users to ask questions and receive visual or textual queries as responses.

Jaisinghani and Aggarwal [10] conducted a study to evaluate two LLMs, Llama3 and BambooLLM, applied to five business intelligence tasks in the AdventureWorks dataset. The study investigated the ability of the model to answer analytical questions and generate relevant visualizations using an architecture with PandasAI and Chainlit, while evaluating metrics from both models against predefined expected responses. The results show that Llama3 offers greater efficiency and a lower error rate in tasks that combine structured and interpretive elements. However, the work does not propose any new approach for extracting responses; only a comparative study with LLMs from the literature is carried out.

Finally, Qi, Miao and Wang [11] developed CleanAgent, a framework that combines LLM-based agents with the ``Dataprep.Clean'' library to automate data cleaning and standardization. Unlike approaches that require extensive prompts and multiple interactions, CleanAgent aims for an experience where the user specifies their requirements only once. While it addresses preprocessing tasks similar to those of data analysis, its scope is limited and does not support explainability or extensive discussions about the data.

Overall, our paper applies LLMs to enable exploratory analysis of DataFrames through natural language. While Data Formulator focuses on creating visualizations and CleanAgent on standardizing columns, we propose an approach that combines descriptive and interpretative analysis, using natural language responses from the LLM with statistical analyses and visualizations generated to the DataFrame and user questions. Unlike DataChat, which is integrated with knowledge graphs, our objective is to provide an exploratory analysis experience that enables users to extract insights from a dataset in an interactive, transparent, and natural manner. Our paper also presents a study on the application of three tools in open health datasets, contributing to evaluate the potential contributions of these tools in supporting exploratory data science.

III. DFChat

DFChat is an open-source tool[2] developed to enable EDA by natural language, combining LLM and RAG with complementary rule-based techniques. In this process, the user has to select an LLM from Groq[3] or use the default (`meta-llama/llama-4-maverick-17b-128e-instruct`). The objective of DFChat is to provide a conversational method to ask questions about a Pandas DataFrame and receive answers without requiring explicit programming. Figure 1 shows the execution pipeline.

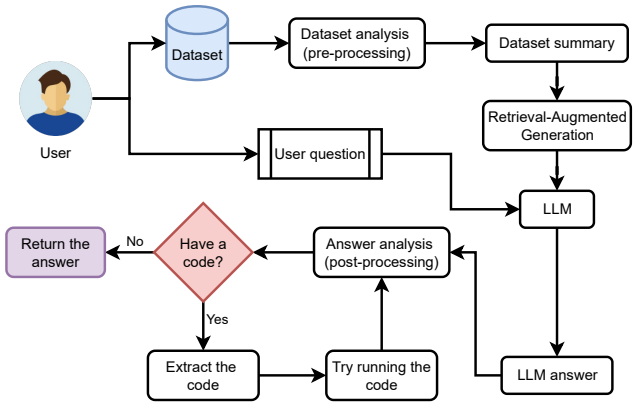


Fig. 1: DFChat execution pipeline

Upon receiving the dataset and the question, the system extracts the columns and their respective data types, which will be used to enrich the LLM context. The RAG mechanism is then activated to structure this knowledge in an interpretable textual format, allowing the model to understand the question more accurately. Based on the LLM response, a post-processing step evaluates whether there is code in the generated text, automatically extracting and executing it. This architecture makes DFChat robust to specific questions, ensuring greater accuracy in the responses generated.

A. Retrieval-Augmented Generation

RAG refers to a strategy that enriches the response generation process of language models by incorporating relevant external information before producing the final output. Rather

[2]https://github.com/fernandozagatti/DFChat/
[3]https://groq.com/

than depending exclusively on internal model parameters, this technique introduces a retrieval stage that collects contextual data from a task-specific knowledge base [12], [13]. In the context of DFChat, RAG is used to structure a textual summary of the dataset that contains relevant information for analysis, including column names, data types, basic statistics, and, when applicable, sample values. This summary is provided as an additional context to the RAG before the user's question.

In addition to enhancing the understanding of the model, the use of RAG also helps reduce incorrect responses, as the model relies on concrete data provided by the user. For example, the LLM will be aware of the columns of the dataset and will perform analyses only on those specific columns. Another significant benefit is modularity: RAG functions as a bridge between data and LLM, enabling DFChat to accurately interpret questions and provide responses without requiring model fine-tuning.

B. Execution and post-processing

Once the LLM generates the response, DFChat proceeds to a controlled execution stage. The LLM returns a response that may include plain natural language text, explanatory reasoning, and/or code intended to be executed over the provided dataset. Before any further action is taken, the system performs a post-processing step to clean the response. Specifically, if the model included internal reasoning (denoted by `<think>...</think>` tags), these tags are removed. Next, the system enters a control loop to determine whether executable code is present. If no code is found in the response, the textual output is returned directly to the user. In the case the response contains one or more code blocks (typically formatted in markdown-style backticks, identified by the markers illustrated in Figure 2), the system attempts to extract the code and execute it.

````
The shape of the dataset is (10, 8), which means it contains 10 rows
(observations) and 8 columns (variables).

To confirm this, we can use the `.shape` attribute of the DataFrame:

```python
import pandas as pd

# Display the shape of the DataFrame
display(df.shape)
```
````

Fig. 2: Code block example

The code is executed in a controlled environment with output redirection and exception handling. If the execution is successful, the results (e.g., printed output, tables, or visualizations) are captured and presented to the user; but if the code fails to execute – for instance, due to syntax errors, undefined variables, or logic faults – the system informs the user that an error occurred during execution. The exact error message is also returned to aid in understanding the failure, as shown in the error output example in Figure 3.

The loop continues until all the codes present in the response have been extracted and executed. This approach ensures that

````
```python
# Summary statistics for numerical columns
display('Summary statistics for numerical columns:')
display(df[['Unnamed:0']].describe())
```
'Summary statistics for numerical columns:'
I tried to run the code, but got the following error:
"None of [Index(['Unnamed:0'], dtype='object')] are in the [columns]"
````

Fig. 3: DFChat error example

the DFChat can handle both purely interpretative responses and those that require active computation. DFChat is integrated directly into the Python code, as shown in Figure 4, where the user imports the dataset and asks questions directly to DFChat.

```python
import pandas as pd
from dfchat import DFChat

df = pd.read_csv('Stress.csv')
chat_llm = DFChat(df)
response = chat_llm.chat("What are the main
characteristics of the dataset?")
```

Fig. 4: DFChat usage example

IV. Evaluation Process

This section investigates whether DFChat provides more accurate, complete, and interpretable responses when performing EDA tasks on real-world open health datasets. To this end, we define an evaluation pipeline that allows us to compare the outputs of different tools when faced with identical analytical questions, measured by human judges. Data Formulator, DFChat, and PandasAI were evaluated using LLaMA 4, applied to five different prompts for each of the five datasets used in the study, resulting in 15 responses per dataset, totaling 75 distinct responses. Then, each response was evaluated by three judges, based on four specific criteria, resulting in a total of 900 distinct scores.

A. Evaluation pipeline

The proposed pipeline involves multiple stages: selecting diverse healthcare datasets, defining benchmark prompts (standardized questions that reflect common EDA tasks), and implementing an evaluation protocol conducted by human experts to assess the quality of each output.

*1) **Datasets and tools**:* To evaluate the effectiveness of DFChat, we conducted a case study using five publicly available healthcare datasets obtained from the Kaggle platform (Human Stress Prediction, Fetal Health Classification, Diabetes Health Indicators Dataset, Lung Cancer Dataset and Suicide and Depression Detection)[4]. We selected the datasets with the intention of covering a range of formats and levels of complexity, including both purely tabular data and data that contain textual information. This diversity enables a more comprehensive evaluation when handling various types of data

[4]https://github.com/fernandozagatti/DFChat/blob/master/Datasets/Datasets.md

through natural language queries. Table I summarizes their characteristics.

TABLE I: Dataset characteristics

Dataset	Instances	Columns	Has text?	Class
1	2838	7	Yes	2
2	2126	22	No	3
3	253680	22	No	3
4	3000	16	No	2
5	232074	3	Yes	2

Thus, we compared DFChat performance with two tools that also use LLMs to assist in data analysis tasks: **Data Formulator** and **PandasAI**. Table II shows the key features of each tool.

TABLE II: Feature comparison between tools

Feature	DFChat	Data Formulator	PandasAI
Natural language interaction	✓	✓	✓
Textual response generation	✓	✓	✓
Graph generation	✓	✓	✓
Code generation	✓	✓*	✗
Internal process visualization	✗	✓	✗
Graphical interface	✗	✓	✗
Jupyter Notebook integration	✓	✗	✓
Dynamic LLM selection	✓	✓	✓

* Only generates SQL code.

Data Formulator is a tool for the automated creation of visualizations from data, enabling users to describe their desired visual intent in natural language. Utilizing a concept-binding architecture, the system translates user intent into data transformations and interpretable graphical components. However, its main objective is to help users express *what* they want to visualize, rather than supporting *why* the data behaves in a certain way. As a result, its analytical capabilities are limited if the inputs provided are open-ended questions and do not identify the columns that should be used.

PandasAI is an open-source library designed to extend the capabilities of the Pandas library with natural language support, using LLMs as a backend. Similar to DFChat, it allows users to ask questions about DataFrames directly in text, and the system interprets these queries by generating and executing Python code.

Since all tools allow you to choose the LLMs that will be used internally, we performed the tests using the same LLM (`meta-llama/llama-4-maverick-17b-128e-instruct`) to ensure a fair comparison.

*2) **Prompts**:* To benchmark the performance of the tools, five standardized prompts were defined that cover different levels of analytical complexity. The prompts were selected to represent everyday tasks in an EDA workflow, from simple to complete analysis of the dataset [14], [15]. This approach enables us to assess the ability of the tools to accurately interpret the intent of the user, provide correct information, and, when applicable, generate visualizations or code. The five prompts used are:

- **"What are the main characteristics of the dataset?"** - Evaluates the ability to identify general information such as the number of columns, data types, number of instances, presence of categorical and numerical variables, among other basic descriptive properties.
- **"Are there any missing values? If so, in which columns and how many?"** - Detect missing values in the dataset, identify the affected columns, and quantify the null values, whether in absolute numbers or percentages.
- **"Can you identify any outliers in the data?"** - Tests statistical methods when dealing with outliers and whether the results presented are consistent.
- **"Generate relevant visualizations that help me understand the distribution of the data"** - Evaluates the generation of useful graphs, such as histograms or boxplots, to help understand the distribution of variables.
- **"Make a complete EDA of my dataset"** - Evaluates the ability to perform a complete exploratory analysis, covering from descriptive statistics to generating visualizations, in a structured and interpretable way.

*3) **Evaluation criteria**:* We conducted a human evaluation involving three data scientists, all of whom are experts with at least six years of experience in artificial intelligence, holding a PhD in Computer Science, Statistics, and Data Science. Each judge was tasked with evaluating the responses provided by the three tools. We opt for an expert-based assessment, as specialists are better equipped to judge the responses, considering that the answers are complex and may have ambiguities. Other methods (e.g., controlled experiments, surveys) are more suitable for later stages of system evaluation. The evaluation was performed independently and anonymously. Judges rated each answer on a Likert scale [16] (from 1 = strongly disagree to 5 = strongly agree) on the following four criteria:

- **Relatedness - "Is the response related to the prompt?":** This criterion evaluates whether the generated content is related to the question.
- **Error-free - "How error-free is the response?":** Raters judged whether the answer exhibited any clear factual, logical, or programming mistakes.
- **Clarity - "Is the response clear and easy to understand?":** This criterion measures the clarity of the response, taking into account structure and overall readability.
- **Coverage - "Does the response show reasoning and cover important aspects of the prompt?":** This criterion evaluates whether the answer provides a thoughtful explanation and addresses the core elements demanded by the prompt.

Each judge assigned a score for each of the criteria in every question, allowing for an aggregated analysis of quality and consistency across tools and criteria. This evaluation design enabled us to compare the general performance of the tools and to identify specific strengths and weaknesses in how each tool interprets and responds to user prompts.

As evaluations are ordinal values (the numbers represent a position in a sequence), the data do not conform to a normal distribution. In this context, we employed two types

of analysis: **correlation metrics** to evaluate the degree of association between pairs of judges (whether they tend to rank items in the same direction, regardless of exact values), and **agreement metrics** to assess the extent to which judges assign the same (or nearby) scores to the same items; this distinction is essential because high correlation does not necessarily imply high agreement [17], [18]. Table III presents the metrics and their suggested interpretations.

B. Correlation metrics

We adopt two non-parametric rank-based coefficients: **Kendall's rank correlation coefficient** (Kendall's τ) and **Spearman's Rank-Order Correlation Coefficient** (Spearman's ρ). Kendall's τ is used to measure the ordinal association between two variables. The τ evaluates the correspondence between the rankings of two sets, specifically the strength and direction of their monotonic association [21], [22].

Like Kendall's τ, Spearman's ρ is used to assess the strength and direction of the monotonic association between two variables [23]; Spearman's ρ method is proper when the data are not normally or contain outliers [22].

C. Agreement metrics

In this study, we use **Kendall's coefficient of concordance** (Kendall's W) and the **Intraclass Correlation Coefficient** (ICC) to measure inter-rater agreement. Kendall's W is a non-parametric statistical measure used to quantify the degree of agreement among multiple raters who rank or score a common set of items. This metric is suitable for ordinal data, and for situations where the goal is to assess the overall consistency of evaluations among three or more judges [24] – Kendall's W takes into account the sum of the ranks assigned to each item and calculates the variation among these totals; the smaller the variation, the greater the agreement among the raters.

ICC is a statistical metric used to quantify consistency or reliability among raters who assign scores to the same set of items. This measure evaluates the proportion of the total variance that can be attributed to actual differences between the items, as opposed to variance arising from subjectivity or inconsistency among the annotators. The ICC is used to assess the agreement among judges, taking into account the variation in the scores assigned by them. We adopt the $ICC(3, k)$, since the judges are fixed – not represent a random sample from the population of possible raters – and the final measure corresponds to the average of the scores by each judge to the same item [25].

V. Results and analysis

The results show that different prompts led to varied performance across the tools, highlighting their strengths and limitations in addressing specific EDA tasks. Both Data Formulator and PandasAI often require multiple code executions to produce results that could be adequately evaluated. In contrast, this issue was significantly reduced with DFChat. Therefore, each tool was executed iteratively until a valid response was obtained, with a maximum of 10 attempts. If all executions resulted in errors, the process was terminated, and the last answer was recorded.

The use of up to ten execution attempts was adopted to account for the stochastic nature of LLM-based systems and the possibility of transient execution failures, such as generation errors, API failures, or unsuccessful code execution. A trial was considered valid when the tool completed the requested task and returned an interpretable response that could be submitted to the evaluation process, regardless of whether the response was ultimately correct. We set the ten-attempt limit as a fixed upper bound to prevent indefinite retries while providing enough opportunities to distinguish occasional failures from systematic execution problems.

A. Correlation between judges

The degree of correlation among the three judges was measured using Kendall's τ and Spearman's ρ. These metrics were applied to evaluate the strength and direction of the scores assigned by each pair of human judges across the four criteria defined in Section IV-A3. Figures 5a and 5b present the results, along with their respective 95% confidence intervals (CI95%) – The CI95% indicates how reliable the estimated value is.

Based on the presented figures, a moderate to good correlation can be observed between the judge pairs. The **"Clarity"** criteria showed the highest levels of agreement between Judge 1 and Judge 3 across both metrics, indicating consistency in the evaluation of this characteristic; in contrast, the **"Coverage"** criteria revealed lower agreement between Judge 2 and Judge 3, suggesting subjectivity or divergence in the interpretation.

B. Agreement between judges

Subsequently, the degree of agreement among the three judges was measured using Kendall's W. This metric was employed to assess the consistency across all human evaluations. Table IV presents the values obtained for each of the evaluated criteria, along with their respective CI95%.

The Kendall's W results indicate an overall level of agreement classified as good among the judges across all criteria. In this way, **"Clarity"** achieved the highest coefficient (W = 0.790), reflecting the greatest consistency in the evaluations. **"Coverage"** showed the lowest value (W = 0.698), although still within the range considered as good agreement. These findings reinforce the overall reliability of the evaluations and corroborate the results of the first two metrics presented in Section V-A. Finally, the degree of agreement among the judges was calculated using $ICC(3, k)$, as shown in Table V, which presents the ICC values and the CI95% obtained.

The $ICC(3, k)$ results showed a good to excellent level of agreement between the judges, with the CI95% ranging from moderate to excellent, indicating that the adopted criteria were interpreted relatively uniformly. However, some variations between judgments were still observed. The **"Error-free"** achieved the highest ICC (0.859), suggesting that the judges

TABLE III: Three suggested interpretations found in the literature

Corr. Range (±)	Interpretation
0.00 - 0.09	Poor
0.10 - 0.39	Weak
0.40 - 0.69	Moderate
0.70 - 0.89	Good
0.90 - 1.00	Excellent

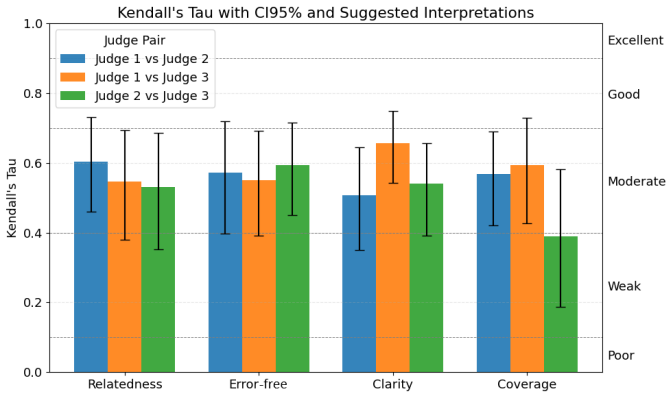
(a) Correlation coefficients [19]

W Range (±)	Interpretation
0.00 - 0.19	Poor
0.10 - 0.39	Weak
0.40 - 0.59	Moderate
0.60 - 0.79	Good
0.80 - 1.00	Excellent

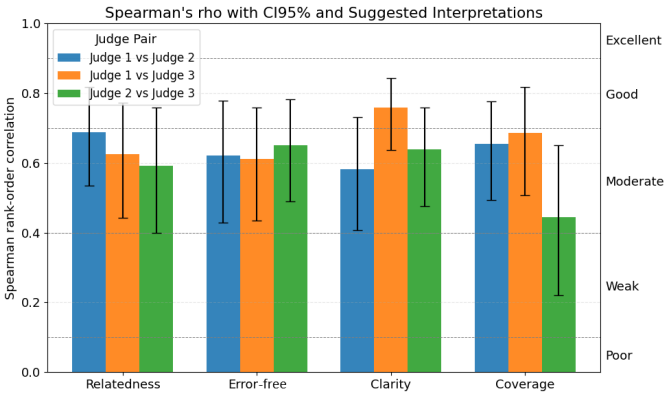
(b) Kendall's W [20]

ICC Range (±)	Interpretation
0.00 - 0.39	Poor
0.40 - 0.54	Weak
0.55 - 0.69	Moderate
0.70 - 0.84	Good
0.85 - 1.00	Excellent

(c) Intraclass correlation [19]



(a) Kendall's τ values with CI95%



(b) Spearman's ρ with CI95%

TABLE IV: Kendall's W Results

Criteria	Kendall's W	CI95% (Range)	Interpretation
Relatedness	0.723	[0.65-0.83]	Good
Error-free	0.748	[0.63-0.83]	Good
Clarity	0.790	[0.67-0.82]	Good
Coverage	0.698	[0.62-0.81]	Good

TABLE V: ICC Results

Criteria	ICC3k	CI95% (Range)	Interpretation
Relatedness	0.813	[0.73-0.88]	Good
Error-free	0.859	[0.79-0.91]	Excellent
Clarity	0.834	[0.76-0.89]	Good
Coverage	0.787	[0.69-0.86]	Good

were able to identify errors with greater objectivity. This behavior is also evident in Kendall's τ and Spearman's ρ, where the correlations of the responses appear more stable than in the other criteria evaluated. The **"Coverage"** criteria showed the lowest agreement (0.787), reflecting the subjectivity involved in this type of judgment.

C. Analysis of the results

Since the judges achieved a high degree of agreement and considering the various aspects evaluated in the experiment conducted, it was decided to present the results as aggregated averages in Table VI rather than listing the raw results individually. In total, 3 tools were evaluated (Data Formulator, DFChat, and PandasAI), all using the same language model (LLaMA 4), applied to 5 different prompts for each of the 5 datasets used in the study – this resulted in 15 responses per dataset, totaling 75 distinct responses. Each of these responses was then evaluated by 3 annotators, based on 4 specific criteria, resulting in a total of 900 distinct scores. Given the large volume of data, presenting the average values allows for the identification of general performance patterns without overwhelming the reader with excessively granular data[5].

Across all prompts, DFChat outperformed the other tools, providing more relevant responses, with greater clarity and coverage, as well as a lower error rate (all LLMs answers are available in Github[6]). The only exception was observed in the fifth prompt, **"Make a complete EDA of my dataset"**, where PandasAI showed a slightly higher average error score (4.2) compared to DFChat (4.1); nonetheless, DFChat showed superior performance in the other three criteria. Figure 6 provides a representation of these results.

Data Formulator, although it also allows interaction with data through natural language, has a distinct focus: it operates by generating SQL queries and presenting a single chart, which limits the analysis. DFChat, even when errors occurred, it often provided the corresponding code and explained their causes. Finally, PandasAI yielded intermediate results: with a similar goal to DFChat, it was capable of generating charts and performing analyses, but with lower transparency and a higher occurrence of failures.

For the fundamental task of identifying missing values, using the Prompt 2 – **"Are there any missing values? If so, in which columns and how many?"** –, DFChat confirmed the integrity of all datasets by reporting zero missing values across every column. When using Prompt 3 – "**Can you identify any outliers in the data?**" –, DFChat utilized the Interquartile

[5]The raw results remain available for detailed analysis on Github.

[6]https://github.com/fernandozagatti/DFChat/tree/master/All-LLM-tools-responses

TABLE VI: Average Performance Evaluation ± Standard Deviation of Tools Across Different Prompts and Criteria

Prompt	Tool	Relatedness	Error-free	Clarity	Coverage
Prompt 1	**Data Formulator**	2.7 ± 0.8	3.9 ± 0.3	2.1 ± 0.1	2.3 ± 0.6
	DFChat	**4.9 ± 0.1**	**5.0 ± 0.0**	**4.5 ± 0.5**	**4.8 ± 0.2**
	PandasAI	2.8 ± 0.5	2.1 ± 0.5	2.5 ± 0.4	2.5 ± 0.6
Prompt 2	**Data Formulator**	1.5 ± 0.5	1.3 ± 0.4	1.6 ± 0.5	1.5 ± 0.4
	DFChat	**5.0 ± 0.0**	**5.0 ± 0.0**	**5.0 ± 0.0**	**5.0 ± 0.0**
	PandasAI	4.3 ± 0.2	3.9 ± 0.3	3.9 ± 0.3	2.9 ± 1.4
Prompt 3	**Data Formulator**	3.3 ± 0.4	3.7 ± 0.6	2.1 ± 0.1	2.3 ± 0.3
	DFChat	**4.5 ± 0.3**	**4.1 ± 0.1**	**4.1 ± 0.1**	**4.5 ± 0.4**
	PandasAI	3.9 ± 0.3	2.7 ± 0.8	2.9 ± 0.9	2.7 ± 1.5
Prompt 4	**Data Formulator**	2.7 ± 1.3	4.0 ± 0.9	2.1 ± 0.6	2.5 ± 0.8
	DFChat	**4.5 ± 0.3**	**4.7 ± 0.3**	**4.1 ± 0.1**	**4.4 ± 0.3**
	PandasAI	3.5 ± 1.6	3.5 ± 0.7	2.5 ± 0.9	2.7 ± 0.8
Prompt 5	**Data Formulator**	2.5 ± 1.0	3.7 ± 0.4	2.1 ± 0.6	2.1 ± 0.9
	DFChat	**4.5 ± 0.3**	4.1 ± 0.1	**4.2 ± 0.2**	**4.5 ± 0.3**
	PandasAI	3.9 ± 0.5	**4.2 ± 0.9**	2.9 ± 0.7	3.1 ± 0.6

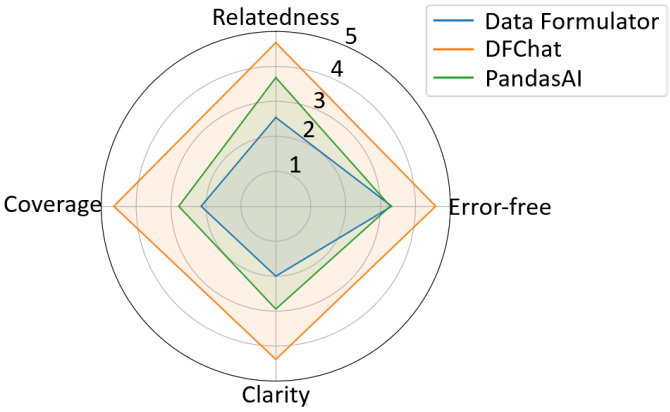


Fig. 6: Radar chart of the general mean scores

Range (IQR) method to identify and quantify anomalous observations in numerical features. Data Formulator often fails or merely suggests visualizing outliers without providing numerical counts. PandasAI gives counts in some cases, but does not do so with the same depth and interpretability as DFChat, typically presenting the graph or outlier count without explaining how it arrived at those results.

DFChat also provided visualizations and relationship mapping to characterize data distributions. Analysis of the "**Suicide and Depression**" dataset revealed a balanced class distribution, with 116037 instances classified as "**suicide**" and an equal number of instances as "**non-suicide**". The "**Human Stress**" dataset identified that the "**ptsd**" (584 entries), "**relationships**" (552 entries), and "**anxiety**" (503 entries) are the most frequent subreddits contributing data. And the "**Fetal Health Classification**" dataset displayed a class imbalance, with 1655 cases classified as "**1.0**" (Normal), 295 as "**2.0**" (Suspect), and 176 as "**3.0**" (Pathological). Data Formulator and PandasAI deliver less detailed results and exhibit a higher frequency of errors when more complex analyses are requested.

VI. Limitations

Although DFChat reduced the number of execution failures compared to competitors, it remains susceptible to occasional misinterpretations and errors, particularly when handling text-based information. The human-based evaluation also introduces a degree of subjectivity, as the evaluation criteria depend on the interpretation of individual judges. Furthermore, as the generated responses are based on a predefined LLM model, the quality of the responses will depend on the size and data used to train this model – therefore, further experiments with additional models are required to investigate the robustness of the observed results.

In parallel, the use of externally hosted LLM APIs also introduces a reproducibility limitation. Model availability, API versions, inference parameters, rate limits, service configurations, and model updates may affect the generated responses over time. Consequently, reproducing the exact experimental results may not always be possible even when using the same prompts and datasets.

Finally, the evaluation was conducted using only five healthcare datasets, which may limit the generalization of the findings to other domains. Although they cover different dataset sizes, numbers of variables, data types, and levels of complexity, this selection represents only a limited subset of the characteristics found in healthcare and biomedical data. Consequently, selection bias may affect the observed results, limiting their generalization to other healthcare datasets and to datasets from domains with different statistical, structural, or semantic properties.

VII. Conclusions and Future Work

This study evaluated DFChat, an interactive and open-source tool for conducting EDA through natural language interaction. By integrating RAG with rule-based execution control, DFChat consistently demonstrated the ability to interpret user questions, generate meaningful visualizations, and deliver accurate insights. Compared to existing tools like Data Formulator and PandasAI, DFChat presented more complete

answers, greater clarity, and lower error rates across all evaluated criteria.

The human evaluation process reinforced the reliability of DFChat. The agreement among judges, measured by Kendall's τ, Spearman's ρ, Kendall's W, and $ICC(3,k)$, reflected moderate to good alignment between the annotators, indicating that the evaluation criteria were well-defined and consistently interpreted. These findings support the conclusion that DFChat provides coherent responses to EDA tasks applied to open healthcare datasets.

For future work, we propose three main directions. First, we aim to evaluate multiple LLMs (both large and lightweight) to assess trade-offs between performance, speed, and resource consumption. This will enable us to determine whether all tools maintain the same quality across different models, or whether there is any gain depending on the LLM used. Second, test the usability of DFChat with non-specialists, such as healthcare professionals, policy analysts, and domain experts without programming experience. Third, integrate DFChat with web-based interfaces to simplify its use; by embedding the tool into intuitive dashboards or interactive platforms (e.g., Streamlit), users will be able to perform data exploration and ask natural language questions without writing code. This direction aims to democratize access to data analysis capabilities, promoting broader adoption in multidisciplinary contexts where data-driven decisions are essential but technical expertise is limited.

Ethics Statement

This work uses publicly available datasets obtained from the Kaggle platform and does not involve direct interaction with human participants.

Disclosure of Generative AI Usage

No generative AI system was used to generate experimental results, fabricate data, or perform the scientific evaluation presented in this work. During the preparation of this manuscript, generative AI tools were used exclusively to assist with language refinement, grammar correction, and text clarity improvement.

References

[1] P. Huston, V. L. Edge, and E. Bernier, "Reaping the benefits of open data in public health," *Can Commun Dis Rep*, vol. 45, no. 11, pp. 252–256, Oct. 2019.

[2] M. B. Courtney, "Exploratory data analysis in schools: A logic model to guide implementation," *International Journal of Education Policy and Leadership*, vol. 17, no. 4, pp. 1–14, 2021.

[3] J. Peng, W. Wu, B. Lockhart, S. Bian, J. N. Yan, L. Xu, Z. Chi, J. M. Rzeszotarski, and J. Wang, "Dataprep. eda: Task-centric exploratory data analysis for statistical modeling in python," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 2271–2280.

[4] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, pp. 1–72, Aug. 2025. [Online]. Available: https://doi.org/10.1145/3744746

[5] F. R. Zagatti, A. G. Regino, F. L. Lopes, G. Y. Shimizu, R. Bonacin, D. Lucrédio, and H. de Medeiros Caseli, "Llm-based solution applied to explore healthcare datasets," in *Advances in Digital Health and Medical Bioengineering II*, H.-N. Costin, R. Magjarevic, and G.-G. Petroiu, Eds. Cham: Springer Nature Switzerland, 2026, pp. 525–530.

[6] C. Wang, J. Thompson, and B. Lee, "Data formulator: Ai-powered concept-driven visualization authoring," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 1128–1138, 2023.

[7] T. Coignion, C. Quinton, and R. Rouvoy, "A performance study of llm-generated code on leetcode," in *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, 2024, pp. 79–89.

[8] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, and B. Myers, "Using an llm to help with code understanding," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[9] L. Fan, S. Lafia, L. Li, F. Yang, and L. Hemphill, "Datachat: Prototyping a conversational agent for dataset search and visualization," *Proceedings of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 586–591, 2023.

[10] B. Jaisinghani and S. Aggarwal, "Comparative performance analysis of large language models in generative business intelligence: Insights from llama3 and bamboollm," *International Journal For Multidisciplinary Research*, vol. 6, no. 6, pp. 1–10, Dec. 2024. [Online]. Available: http://dx.doi.org/10.36948/ijfmr.2024.v06i06.33967

[11] D. Qi, Z. Miao, and J. Wang, "Cleanagent: Automating data standardization with llm-based agents," *arXiv preprint arXiv:2403.08291*, 2024.

[12] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[13] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," *arXiv preprint arXiv:2402.19473*, 2024.

[14] S. V. Daele and G. Janssenswillen, "Identifying the steps in an exploratory data analysis: A process-oriented approach," in *International Conference on Process Mining*. Springer, 2022, pp. 526–538.

[15] S. Paul, R. Singh, M. Bolesnikov, and A. Vuković, "Use of exploratory data analysis and visualization techniques to meet organization strategic goals," in *International Conference on Business, Management, and Economics Engineering*, vol. 1, 01 2025, pp. 456–463.

[16] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, p. 396, 2015.

[17] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: Measures of agreement," *Perspectives in Clinical Research*, vol. 8, no. 4, pp. 187–191, 2017.

[18] C. Edwards, H. Allen, and C. Chamunyonga, "Correlation does not imply agreement: A cautionary tale for researchers and reviewers," *Sonography*, vol. 8, no. 4, pp. 185–190, 2021.

[19] P. Schober, E. J. Mascha, and T. R. Vetter, "Statistics from a (agreement) to z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research," *Anesthesia & Analgesia*, vol. 133, no. 6, pp. 1633–1641, 2021.

[20] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[21] R. Newson, "Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences," *The Stata Journal*, vol. 2, no. 1, pp. 45–64, 2002.

[22] F. Essam, H. El, and S. R. H. Ali, "A comparison of the pearson, spearman rank and kendall tau correlation coefficients using quantitative variables," *Asian J. Probab. Stat*, vol. 20, no. 3, pp. 36–48, 2022.

[23] F. Schmid and R. Schmidt, "Multivariate extensions of spearman's rho and related statistics," *Statistics & Probability Letters*, vol. 77, no. 4, pp. 407–416, 2007.

[24] F. Franceschini and D. Maisano, "Aggregating multiple ordinal rankings in engineering design: the best model according to the kendall's coefficient of concordance," *Research in Engineering Design*, vol. 32, pp. 91–103, 2021.

[25] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

Analysis of orofacial thermal biomarkers at different acquisition of distances: a comparative approach between human analysis and conventional artificial intelligence

A.C.R. Silva, Biomedical Engineering Program - COPPE, Federal Univers. of Rio de Janeiro, Rio de Janeiro,Brazil, carolina.rodrigues@peb.ufrj.br

S.F. Raymundo Biomedical Engineering Program - COPPE Federal University of Rio de Janeiro Rio de Janeiro, Brazil stela.freitas@peb.ufrj.br

W.C.A . Pereira Biomedical Engineering Program - COPPE Federal University of Rio de Janeiro Rio de Janeiro, Brazil wagner@peb.ufrj.br

Abstract— **Infrared thermography is a non-invasive imaging technique increasingly used in the assessment of orofacial functional conditions. However, technical factors, such as acquisition distance, can influence the reliability of measurements. This study evaluates whether human manual analysis and artificial intelligence (AI) computational analysis recognize the same variation in orofacial thermal patterns when the acquisition distance changes from 50 cm to 100 cm across four bilateral regions of interest (ROIs): right and left temporomandibular joints (TMJ) and right and left temporal muscles. Twelve asymptomatic volunteers were assessed via infrared thermography (FLIR E40). Human statistical analysis detected a systematic absolute temperature reduction at 100 cm in 45 of 48 ROI-volunteer pairs (93.75%), with mean reductions ranging from 0.38 °C to 0.75 °C ($p < 0.001$). Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers discriminated acquisition distances with point accuracies of 95.83% and 87.50%, respectively. This investigation constitutes a metrological and computational evaluation rather than a clinical diagnostic study. The findings demonstrate that acquisition distance is a relevant source of technical bias that alters absolute temperature values by a magnitude capable of masking or mimicking physiological asymmetries. Consequently, relational or spatial thermal biomarkers are structurally more robust than isolated absolute measurements. Results remain preliminary (n = 12) and restricted to healthy individuals.**

Keywords— Infrared thermography, Temporomandibular joint, Machine learning, Orofacial analysis, Thermal asymmetry

I. INTRODUCTION

Infrared thermography (IRT) is an imaging technique for orofacial assessment, as it allows mapping of skin surface temperature, reflecting underlying physiological processes such as blood flow and metabolic activity. Previous studies have demonstrated that IRT is sensitive to vascular and inflammatory changes associated with TMD, supporting its use as a complementary diagnostic tool [1], [2]. Haddad et al. demonstrated that thermography can detect subtle thermal asymmetries in the orofacial region, while Dibai-Filho et al. reinforced the association between thermal asymmetry and musculoskeletal dysfunction.

The clinical applicability of IRT remains dependent on methodological standardization. Factors such as environmental conditions, emissivity settings, and acquisition parameters can significantly influence temperature measurements. Acquisition distance represents a critical variable that directly affects the amount of infrared radiation captured by the camera sensor. Variations in distance can lead to systematic changes in measured temperature values, potentially compromising data comparability [3], [4],[5].

Thermal asymmetry has been recognized as an important parameter in thermographic analysis. Recent studies suggest that spatial thermal patterns can provide more robust information than isolated temperature values, particularly when analyzed using advanced computational approaches [6].

In this context, machine learning techniques have gained attention for their ability to identify complex patterns in biomedical data. Algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) have been successfully applied to thermographic datasets, enabling automated classification and enhanced pattern recognition [7],[10]. These approaches complement traditional statistical methods, allowing the assessment of data structure and separability.

Among technical sources of variability, target-to-camera acquisition distance represents a critical parameter that directly affects atmospheric attenuation, optical field of view, and effective spatial resolution per unit area. Despite standard consensus guidelines strongly capture distance control, practical clinical research often exhibits minor variations in patient positioning. The degree to which such distance shifts distort extracted thermal biomarkers has not been fully elucidated in orofacial applications, where anatomical structures are small and thermal variations of 0.3 °C to 0.5 °C are frequently interpreted as clinically significant.

The primary objective of this study is to perform a comparative metrological and computational evaluation between human quantitative analysis and supervised machine-learning algorithms in detecting thermal pattern alterations induced by varying capture distance (50 cm versus 100 cm) across four bilateral orofacial regions. It is explicitly highlighted that this work does not propose a diagnostic tool for temporomandibular disorders (TMD), nor does it evaluate pathological classification. Instead, its core contribution lies in quantifying the technical bias generated by acquisition distance and demonstrating that relational or spatial thermal features provide superior stability compared to raw absolute surface temperatures.

CNPq, CAPES, and FAPERJ for financial support

The presence or absence of temporomandibular disorders was not an inclusion criterion and was not evaluated in this study; the investigation is restricted to the metrological effect of acquisition distance on orofacial thermal patterns.

II. MATERIALS AND METHODS

A cross-sectional observational metrological study was conducted involving 12 asymptomatic adult volunteers (both sexes, aged between 20 and 40 years). The study protocol was approved by the Research Ethics Committee of the Clementino Fraga Filho University Hospital and the Faculty of Medicine (HUCFF) of the Federal University of Rio de Janeiro (UFRJ), under number CAAE: 78605524.3.00005257 and ReBEC number: RBR-10367s7j. A physiotherapist was responsible for capturing, processing, and analyzing the infrared images. All participants signed an informed consent form. Inclusion criteria imply healthy individuals presenting complete natural dentition, absence of facial pain, and absence of temporomandibular signs or symptoms verified through anamnesis. Exclusion criteria encompassed history of facial trauma or surgery, orthodontic treatment within the preceding 12 months, systemic metabolic or rheumatic diseases, use of analgesics, anti-inflammatory drugs, or vasoactive substances in the 72 hours prior to examination, and presence of beard or makeup in the evaluated areas. Image acquisitions occurred in an acclimatized room maintained between 21 °C and 23 °C, with relative humidity at approximately 50%, following a 10-minute stabilization period per subject.

A. Image Acquisition and Human Analysis Protocol

Thermographic images were acquired using an uncooled focal plane array infrared camera (FLIR E40, FLIR Systems Inc., Wilsonville, OR, USA) featuring a native thermal resolution of 160 × 120 pixels, thermal sensitivity (NETD) < 0.07 °C at 30 °C, measurement range of −20 to 650 °C, and measurement accuracy of ± 2 °C or ± 2% of overall reading. Surface emissivity was set at 0.98. For each participant, images were captured perpendicular to the cutaneous surface at two standardized distances: 50 cm and 100 cm, previously marked on the floor. Five consecutive thermograms were captured per region and per distance, generating 20 thermographic images per participant at each distance. Human quantitative evaluation was performed by a specialized physiotherapist with extensive experience in infrared image interpretation using dedicated radiometric software (FLIR Tools). Four anatomical regions of interest (ROIs) were manually delimited bilaterally: Right TMJ, Left TMJ, Right Temporal Muscle, and Left Temporal Muscle. Mean surface temperature (°C), standard deviation, and coefficient of variation (CV) were extracted for each ROI across the five sequential frames per subject and distance.

B. Artificial Intelligence Framework and Tool Differentiation

To ensure methodological clarity and reproducibility, artificial intelligence tools were implemented under two strictly distinct operational modalities. Generative Artificial Intelligence (ChatGPT, OpenAI) was used as a computational assistance tool to implement and execute the conventional machine-learning classifiers (SVM and KNN), generating and running the classification code under the supervision of the authors, and as an auxiliary tool for descriptive summarization, structured textual organization, and comparative interpretation, including the comparison between the computational results and the independent statistical analysis performed by the authors. The generative model was not used for statistical inference: all statistical analyzes (normality assessment, paired comparisons, effect sizes, and confidence intervals) were performed independently by the authors. The conventional machine-learning classifiers (Support Vector Machine and K-Nearest Neighbors) are standard supervised models implemented through reproducible Python scripts (Python 3.9, scikit-learn 1.0). The specifications are:

- Algorithms: SVM (RBF kernel) and KNN (k=5, Euclidean)
- Features: 4 average temperatures per ROI, label: distance (50 vs. 100 cm)
- Split: StratifiedGroupKFold, 5 folds, by participant
- Validation: bootstrap (10,000 iterations), 95% CI
- SVM outperformed KNN (95.83% vs. 87.50% accuracy).

C. Statistical Analysis

Data distribution normality was evaluated using the Shapiro-Wilk test. Continuous variables with Gaussian distribution were analyzed using the paired Student's t-test (Left TMJ and Left Temporal Muscle), while variables exhibiting non-normal distribution were evaluated using the Wilcoxon signed-rank test (Right TMJ and Right Temporal Muscle). Variance homogeneity across distances was verified via Levene's test. Effect sizes for paired samples were quantified using Cohen's d alongside their respective 95% confidence intervals (95% CI). Statistical power was computed post-hoc considering sample size (n = 12), observed effect sizes, and two-tailed alpha of 0.05. Binary directional agreement between human detection and actual physical distance shift was assessed as the proportion of ROI-volunteer pairs in which the temperature decreased at 100 cm relative to 50 cm (reported in Table III). A significance level of $p < 0.05$ was adopted for all analyses.

D. Machine Learning Classification Pipeline

Machine learning techniques were applied to investigate the influence of acquisition distance on thermal patterns. The initial objective was to determine whether temperature patterns in the temporomandibular joint and temporal muscle exhibited thermal variations in thermograms captured at distances of 50 cm and 100 cm. Following the presentation of the image acquisition methodology, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classification algorithms were selected and employed to classify the thermographic data based on the two acquisition distances. Model performance was interpreted as an indicator of thermal patterns across the conditions. In the second stage of the neural network analysis, a table containing independently

CNPq, CAPES, and FAPERJ for financial support.

obtained statistical values was included, and the network was tasked with statistically analyzing and comparing these values.

The computational classification task was designed to determine whether statistical thermal representations could discriminate between capture distances (0 = 50 cm; 1 = 100 cm). The feature vector comprised the four mean ROI temperatures [right ATM, left ATM, right temporal muscle, left temporal muscle] for each participant. So rigorouslyavoid data leakage and performance estimate inflation resulting from repeated captures (pseudoreplication), a participant-level cross-validation strategy was employed. Performance was evaluated using Classification Accuracy, Sensitivity (True Positive Rate for 100 cm), Specificity (True Negative Rate for 50 cm), Positive Predictive Value (PPV), Negative Predictive Value (NPV).

III. RESULTS

Thermographic analysis of the 12 volunteers revealed temperature patterns consistent with the normal physiology of the orofacial region. Table I presents the mean temperatures and standard deviation (SD) for each region of interest (ROI) at both distances, obtained through independent (human) analysis.

In Table I, the average temperatures showed statistically significant reductions in all areas evaluated at a distance of 100 cm, except for the left temporal region. The right temporal region (ATM D) showed a decrease from 34.80 ± 0.65 °C to 34.04 ± 0.64 °C, representing a thermal differential of -0.76 °C. A similar pattern was observed in the left temporal region (ATM E), with a drop from 34.83 ± 0.49 °C to 34.15 ± 0.48 °C.= -0.68°C.ΔT

TABLE I. HUMAN STATISTICAL ANALYSIS BY ROI AT 50 AND 100 CENTIMETERS.

Region of Interest (ROI)	Distance	Mean (°C)	SD (°C)	CV (%)	Median (°C)	Statistical Test	p-value
Left TMJ	50 cm	34.83	0.412	1.18	35.00	Paired t-test	< 0.001
Left TMJ	100 cm	34.17	0.345	1.01	34.20	Paired t-test	< 0.001
Right TMJ	50 cm	34.80	0.751	2.16	35.10	Wilcoxon	< 0.001
Right TMJ	100 cm	34.05	0.582	1.71	34.15	Wilcoxon	< 0.001
Left Temporal Muscle	50 cm	34.83	0.314	0.90	34.90	Paired t-test	< 0.001
Left Temporal Muscle	100 cm	34.45	0.355	1.03	34.60	Paired t-test	< 0.001
Right Temporal Muscle	50 cm	34.73	0.479	1.38	34.90	Wilcoxon	< 0.001
Right Temporal Muscle	100 cm	34.19	0.592	1.73	34.40	Wilcoxon	< 0.001

Source: Author

The effect size (Cohen's d) for the temporomandibular joints ranged from 1.74 to 2.30, indicating a fairly significant effect size. The statistical power of 1.00 for these measures suggests that the observed differences are significant and do not result from random variation.

The 95% confidence intervals (95% CI) for the thermal variation (ΔT) between acquisition distances of 50 cm and 100 cm demonstrate high metrological consistency. The fact that neither interval crosses the zero threshold confirms the statistical significance ($p<0.001$) of the distance-induced thermal displacement.

Regarding the comparative recognition of the distance effect, human evaluation correctly identified a downward thermal shift at 100 cm in 45 out of 48 individual ROI-volunteer pairs, yielding a directional detection accuracy of 93.75%. The only three discordant pairs where temperature increased at 100 cm were Participant #1 (Left Temporal Muscle: 34.60 °C vs 34.62 °C), Participant #5 (Right TMJ: 33.46 °C vs 33.76 °C), and Participant #10 (Left Temporal Muscle: 34.20 °C vs 34.30 °C). Supervised machine-learning models operating on the 4-dimensional ROI thermal vectors demonstrated strong discriminative capacity between capture distances, as shown in Table II.

TABLE II. COMPARATIVE PERFORMANCE METRICS: HUMAN ANALYSIS VS MACHINE LEARNING

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Human Analysis (Decision Rule)	93.75	93.75	93.75	93.75	93.75
Support Vector Machine (SVM)	95.80	92.30	100.00	100.00	91.70
K-Nearest Neighbors (KNN)	87.50	91.70	83.30	84.60	90.90

Source: Author

A. Calculation of the Mean Percentage Error

Based on the average value of -0.7 °C at the average reference temperature (34.6 °C) at 50 cm, the average percentage error caused by the increase in distance is:

$$\text{Erro } \% = \left(\frac{0.7}{34.6}\right) \cdot 100 \approx \mathbf{2.02}\%$$

Although it may seem small in percentage terms, in terms of thermal physiology, a variation can mask or simulate a pathological thermal asymmetry, validating justifying the use of a fixed distance of 50 cm.

According to Table II, the automated model achieved superior levels of accuracy, sensitivity, and specificity. The integration of Support Vector Machine (SVM) algorithms allowed for a more precise identification of thermal asymmetries, minimizing interobserver variability.

B. Confusion Matrix in the SVM and KNN models

The confusion matrix shown in Figure 1 was used to analyze the performance of the SVM (Support Vector Machine) and KNN (K-Nearest Neighbors) machine learning models in classifying thermographic images obtained from different capture distances (50 cm and 100 cm).

The aim of the analysis was to determine whether the algorithms could detect thermal variations related only to the distance from which the thermographic image was captured, and not to physiological changes in the volunteers. The images analyzed were of the same individuals but were captured at two different distances from the thermographic camera.

It was noted that the images obtained at 50 cm showed apparently higher temperatures compared to the images captured at 100 cm, a result that had also been identified previously through human analysis [11].

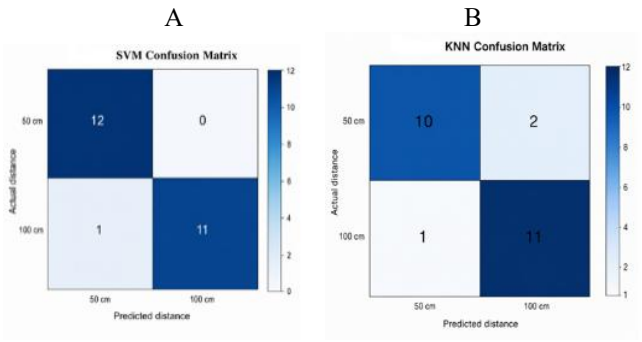


Fig. 1. SVM (A) and KNN (B) Confusion Matrix (Accuracy 95.8%)
Source: Author

In the confusion matrix of the SVM model (Fig. 1 A), the results demonstrated:

- 12 correct classifications for images obtained up to 50 cm;
- 11 correct classifications for images obtained at 100 cm;
- only 1 classification error;
- Absence of false positives for the 100 cm class.

The overall accuracy of the model was approximately:

The accuracy of 95.8% indicates the discriminatorycapacityof the SVM model in recognizing thermal differences resulting from variations in acquisition distance.

C. Positive Predictive Value (PPV)

Positive Predictive Value (PPV) This indicates a 100% probability that images classified as 100 cm (positive class) belong to the 100 cm class.

VPP = 100%

D. Negative Predictive Value (NPV)

Negative Predictive Value (NPV) This indicates a 91.7% probability that images classified as 50 cm (negative class) actually belong to the 50 cm class.

VPN = 91.7%

In Table III, the sensitivity of 92.3% demonstrates efficiency in identifying images acquired at 100 cm (positive class), whereas the specificity of 100% indicates that all images acquired at 50 cm were correctly recognized by the model (negative class). These results indicate that the machine learning algorithm is able to discriminate the two acquisition distances based on the variation of mean absolute temperature, confirming the human analysis reported previously. Thus, the observed thermal variations can be attributed to a technical effect of image acquisition, rather than to physiological alterations in the volunteers

TABLE III. DIAGNOSTIC PERFORMANCE METRICS OF THE SVM AND KNN CLASSIFIERS

Diagnostic Performance Metrics: SVM vs. KNN			
Model	**Sensibility (%)**	**Specificity (%)**	**Acurácia (%)**
SVM (Support Vector Machine)	92,30%	100,00%	95,80%
KNN (K-Nearst Neighbors)	91,70%	83,33%	87,50%

Source: Author

The objective of the KNN analysis was to verify if the model would be able to recognize thermal patterns related to the distance from the thermal imaging camera, distinguishing images acquired at 50 cm from those obtained at 100 cm. The results were also compared to human analysis, which identified apparently higher temperatures in images captured at shorter distances. The confusion matrix of the KNN model (Fig. 1B) showed:

- 11 true positives;
- 10 true negatives;
- 2 false positives;
- 1 false negative.

The overall accuracy of the model was calculated by:

Accuracy = 87.5%

The results indicate that the KNN model correctly classified 87.5% of the evaluated thermographic images but performed worse than the SVM due to a greater number of incorrect classifications. The positive predictive value (PPV = 84.6%) indicates the probability that images classified as 100 cm (positive class) indeed belong to the 100 cm class. The negative predictive value (NPV = 90.9%) indicates the probability that images classified as 50 cm (negative class) indeed belong to the 50 cm class.

The detection of the offset by both SVM and KNN models indicated that the algorithms were able to identify this abrupt temperature variation and classify whether the image was captured at a short or long distance (50 or 100 cm). This confirms that the thermal alteration is a reproducible technical artifact, rather than a physiological alteration in the volunteer, as also evidenced by the human analysis.

Although the KNN presented sensitivity (91.67%), its specificity was 83.33%, the lowest among the tested models, as shown in Table III. This difference indicates that KNN, by relying on neighborhood similarity, is more sensitive to biological variations between individuals. In cases where the volunteer's basal temperature is naturally low, images acquired at 50 cm may "converge" to the 100 cm class of another individual, leading to an increase in the occurrence of False Positives.

IV. DISCUSSION

The results of this research provide information on the technical reliability of infrared thermography (IRT) for orofacial clinical assessment. A key observation is that absolute temperature readings are significantly sensitive to

acquisition distance, as demonstrated by the thermal reduction observed at 100 compared to 50 centimeters.

The findings of this investigation provide evidence that acquisition distance introduces a systematic metrological bias in orofacial infrared thermography. Increasing capture distance from 50 cm to 100 cm resulted in an average absolute temperature reduction of 0.38 °C to 0.75 °C across all four evaluated ROIs. This attenuation is close to the principles outlined in ASTM E2582-17 guidelines standards and fundamental radiation transfer equations, wherein atmospheric absorption, radiation dispersion, and decreased pixel coverage per unit anatomical area degrade thermal readings at extended distances [3] [4].

The implementation of the automated machine learning pipeline resulted in advancements in both diagnostic operational precision and efficiency. SVM showed the highest discriminative performance between the two acquisition distances among the tested models. However, this result reflects a metrological discrimination task and does not support any claim of diagnostic automation or clinical diagnostic robustness. As suggested by Lahiri et al. (2022) [5], machine learning excels at identifying the subtle, non-linear relationships that often remain undetected during manual analysis.

The stability observed consistently with the facial symmetry mapping proposed by Santos et al. (2024) [12], supports the use of relational biomarkers as diagnostic indicators that are more robust than isolated temperature readings. From a physiological perspective, the anterior temporalis muscle and the TMJ regions remain critical to the functioning of the masticatory system, validating the methodology of using bilateral reference points to mitigate environmental or technical noise [4], [7].

The identification of a systematic error of approximately 2.02% (≈0.7 °C) when moving the camera 100 cm away has implications for the validity of thermal asymmetry analysis. In the specialized literature, thermal asymmetry thresholds considered pathological for orofacial structures are frequently set at 0.3°C.

Importantly, the observed thermal discrepancy of up to 0.75 °C at the TMJs is of the same order of magnitude as thermal asymmetry thresholds (frequently defined between 0.3 °C and 0.8 °C) used in clinical literature to suggest unilateral pathology, joint effusion, or localized muscular inflammation. Therefore, unstandardized capture distances in clinical practice or research protocols may either generate false-positive asymmetries if distance varies across bilateral captures or mask true underlying pathological asymmetries if captures are improperly scaled.

The high classification performance achieved by the SVM (95.83%) and KNN (87.50%) algorithms must be interpreted with strict technical caution. These models were trained to perform a metrological discrimination task (distinguishing between 50 cm and 100 cm setups) based on a 4-dimensional vector of mean surface temperatures. The high discriminatory accuracy indicates that the thermal shift is uniform and substantial enough across the face for computational algorithms to readily partition the feature space. However, this does not represent diagnostic accuracy for clinical dysfunction. The algorithms are primarily detecting the gross shift in mean absolute temperature rather than subtle spatial pattern reconfigurations. The classification task distinguishes acquisition distances and does not diagnose temporomandibular disorders; Accordingly, expressions such as diagnostic accuracy, clinical automation, or diagnostic robustness are not supported by the present evidence and were deliberately avoided.

Several methodological limitations must be addressed. First, the sample size was restricted to 12 asymptomatic healthy volunteers, limiting the generalizability of statistical distributions; the presence or absence of temporomandibular disorders was not evaluated in this study. Second, the spatial resolution of the FLIR E40 camera (160 × 120 pixels) results in fewer measurement pixels covering small anatomical targets like the TMJ at 100 cm, increasing susceptibility to partial volume effects and background radiative contamination. Third, inter-individual baseline thermal variability and physiological fluctuations can influence absolute cutoffs, as reflected in the wide 95% confidence intervals for Cohen's d. Fourth, because the features were the mean temperatures per ROI, the present analysis does not provide a voxel-level evaluation of spatial patterns; the stability of the relative thermal organization was inferred from the consistency of the bilateral thermal distribution across distances. Future investigations should incorporate larger cohorts, include patients with clinically confirmed temporomandibular disorders, evaluate spatial texture and gradient features, and implement multi-centric validation.

V. CONCLUSION

The acquisition distance (50 cm versus 100 cm) significantly alters the absolute surface temperature values in orofacial infrared thermography, producing a systematic thermal reduction of 0.38 °C to 0.75 °C in healthy individuals. The magnitude of this technical artifact is sufficient to compromise clinical interpretations by masking or mimicking physiological asymmetries. Both human analysis and machine learning classifiers recognized this distance-dependent variation with high agreement.

Metrological standardization of the patient-to-camera distance is essential to ensure reproducibility in quantitative thermographic protocols. Relational and asymmetry biomarkers demonstrated higher stability than isolated absolute temperatures. Further validation in symptomatic clinical populations and larger cohorts is required.

ACKOWLEDGMENT

We thank CNPq, CAPES, and FAPERJ for financial support.

The authors used an artificial intelligence tool for processing image and statistical support. The authors reviewed and took full responsibility for the content of this manuscript.

REFERENCE

[1] DS Haddad et al., “Infrared thermography in orofacial assessment”, Journal of Bodywork and Movement Therapies, vol. 25, pp. 1–7, 2021.

[2] AV Dibai-Filho et al., “Thermal asymmetry and musculoskeletal correlation in temporomandibular disorders”, Journal of Physical Therapy Science, vol. 34, no. 2, pp. 150–156, 2022.

[3] ASTM International, “ASTM E2582-17: Standard guide for infrared imaging systems for medical applications”, 2017.
[4] Silva, ACR et al. (2025). Evaluation of temperature variation as a function of distance using infrared thermography. In: Soares, AB, Cardoso, GC, Leoni, RF (eds) XXIX Brazilian Congress of Biomedical Engineering - Volume 1: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing. CBEB 2024. Proceedings of IFMBE, vol. 125. Springer, Cham.https://doi.org/10.1007/978-3-031-93646-3_85
[5] BB Lahiri et al., “Deep learning in medical infrared imaging,” Infrared Physics & Technology, vol. 120, 2022.
[6] ASTM International, “ASTM E1934-99a (2021): Standard guide for examining electrical and mechanical equipment with infrared thermography”, 2021.
[7] European Association of Thermology, “Guidelines for Clinical Thermography”, 2020.
[8] AJ Kopczynski et al., “Deep learning segmentation of medical thermal images", Sensors, vol. 21, no. 10, 2021.
[9] K. Song et al., “Automated ROI detection in facial thermography using deep learning,” IEEE Access, vol. 10, pp. 112345–112356, 2022.
[10] T. Pereira et al., “Computer vision methods in facial thermography,” in Proc. IEEE EMBC, 2023, pp. 1–5.
[11] ACR Silva, SF Raymundo and WCA Pereira, "Analysis of the Temperature Pattern of the Temporomandibular Joint and Previous Temporal Muscle as a Function of Acquisition Distance Using Infrared Thermography", 2026 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), Rio de Janeiro, Brazil, 2026, pp. 1-4, doi: 10.1109/GMEPE/PAHCE69420.2026.11550301.
[12] MAB Silva et al., “Normative values of masticatory muscle thermography”, Brazilian Journal of Dentistry, 2023.
[13] AA Santos et al., “Thermal mapping of facial symmetry in healthy adults”, Journal of Oral Rehabilitation, 2024.

Análise Comparativa de Doses e Parâmetros de Aquisição para Diferentes Tecnologias de Mamografia Digital

Alexsandra Pavani Xavier e Homero Schiabel
Depto. de Engenharia Elétrica e de Computação
Escola de Engenharia de São Carlos (EESC/USP)
São Carlos, Brasil

***Resumo*—Este trabalho apresenta uma análise da Dose de Entrada na Pele (DEP) e Dose Glandular Média (DGM) em oito sistemas de mamografia digital utilizando um simulador estruturado desenvolvido em pesquisa prévia pelo grupo e extração automatizada de metadados DICOM. O objetivo foi avaliar e comparar os parâmetros dosimétricos e de aquisição entre os diferentes sistemas de mamografia digital. Os resultados mostraram que o equipamento 4A apresentou os maiores valores médios de DEP, comportamento mantido nas exposições com espessura de 48 mm. Essa tendência pode estar relacionada à estratégia automática de exposição, à combinação alvo/filtro e aos parâmetros de aquisição selecionados pelo sistema. O maior nível de dose não foi acompanhado por maior contraste RMS. Apesar disso, os valores de DGM permaneceram abaixo dos níveis de tolerância adotados como referência da ANVISA . Conclui-se que a extração automatizada de metadados DICOM, associada ao uso de simuladores estruturados, é uma ferramenta útil para o rastreamento e a otimização dosimétrica em mamografia digital.**

***Index Terms*—Mamografia Digital, Dose Glandular Média, Metadados DICOM, Simulador de Mama, Garantia de Qualidade.**

I. Introdução

O câncer de mama é a neoplasia de maior incidência mundial e a principal causa de mortalidade por câncer na população feminina no Brasil [1]. A detecção precoce da doença permanece como o pilar fundamental para a obtenção de prognósticos favoráveis e tratamentos terapêuticos menos invasivos, sendo a mamografia consolidada como método padrão-ouro de rastreamento populacional [1].

Com a evolução tecnológica, os sistemas tradicionais ecrã-filme foram amplamente substituídos pela mamografia digital, subdividida em radiografia computadorizada (CR) e digital direta (DR). Essa transição proporcionou ampla faixa dinâmica de exposição e recursos avançados de pós-processamento, otimizando a detecção de estruturas milimétricas, como as microcalcificações. No entanto, a flexibilidade inerente aos sistemas digitais introduz o risco de sobre-exposição da paciente sem que ocorra uma degradação visível da qualidade da imagem, tornando o monitoramento das doses radiológicas um aspecto crítico da garantia de qualidade [2].

Neste contexto, a avaliação dosimétrica deste estudo fundamenta-se em duas grandezas principais extraídas dos metadados das imagens: a Dose de Entrada na Pele (DEP) e a Dose Glandular Média (DGM). Ambas as grandezas são relevantes para a radioproteção da paciente. A otimização desses níveis de radiação depende do equilíbrio entre os parâmetros técnicos de aquisição, como a tensão do tubo (kVp), a carga (mAs), a combinação alvo/filtro e a lógica do controle automático de exposição implementada por cada fabricante.

Para a avaliação de sistemas mamográficos em ambiente clínico, simuladores mamográficos (*phantoms*) podem ser utilizados para reproduzir características de atenuação do tecido mamário [3]. Somado a isso, os metadados presentes nos cabeçalhos DICOM fornecem parâmetros técnicos e dosimétricos relevantes para o rastreamento dosimétrico dos sistemas mamográficos.

A partir dessa fundamentação, o presente trabalho tem como objetivo avaliar e comparar os parâmetros dosimétricos e de aquisição registrados em oito sistemas de mamografia digital em rotina clínica, a partir dos metadados DICOM, utilizando um *phantom* desenvolvido pelo grupo [4], [5]. Foram analisados os valores de DEP e DGM, os parâmetros técnicos selecionados pelos sistemas automáticos e, de forma complementar, o contraste RMS local e global de imagens obtidas nos equipamentos [6].

Buscou-se, assim, identificar diferenças no comportamento dos sistemas e possíveis associações entre os parâmetros de aquisição e os níveis de dose observados. Além disso, os valores de DGM foram comparados aos níveis de referência estabelecidos pela normativa nacional da Agência Nacional de Vigilância Sanitária (ANVISA) [7].

II. Metodologia

Foram avaliados, de forma individualizada, oito sistemas de mamografia digital direta em rotina clínica, abrangendo a tecnologia Digital de Campo Total (Digital Radiography - DR). A amostragem foi estruturada de modo a analisar o comportamento de cada equipamento, sendo composta por dispositivos de quatro fabricantes distintos: três unidades de um mesmo modelo do Fabricante I, operando respectivamente na Clínica I (equipamento 1A), na Clínica II (equipamento 1B) e na Clínica III (equipamento 1C); uma unidade do Fabricante II na Clínica IV (equipamento 2); uma unidade do Fabricante III no Hospital I (equipamento 3); e três modelos com tecnologias distintas do Fabricante IV, distribuídos entre

o Hospital II (equipamento 4A), a Clínica V (equipamento 4B) e a Clínica VI (equipamento 4C).

Cada um dos oito sistemas possui características próprias, operando com combinações específicas de alvo/filtro e lógicas de controle automático de exposição, incluindo os modos convencional (*Automatic Exposure Control* - AEC), dinâmico de contraste (CNT) e otimização automatizada de parâmetros (AOP).

O número de exposições consideradas variou entre os equipamentos, com amostras de $n = 1$ a $n = 8$. Nos equipamentos 1A, 1B, 1C, 3 e 4B, as aquisições foram realizadas em modo AEC. No equipamento 2, foram consideradas duas aquisições, correspondentes aos modos OPDOSE e AEC. Para o equipamento 4A, foram consideradas as exposições em modo CNT, enquanto no equipamento 4C a aquisição foi realizada em modo AOP.

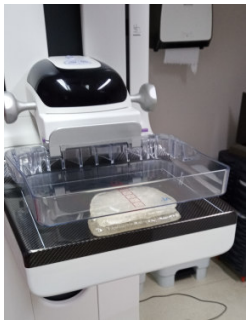
As especificações detalhadas de cada sistema estão apresentadas na Tabela I.

TABELA I
CARACTERÍSTICAS TÉCNICAS E PARÂMETROS DOS SISTEMAS DE MAMOGRAFIA AVALIADOS

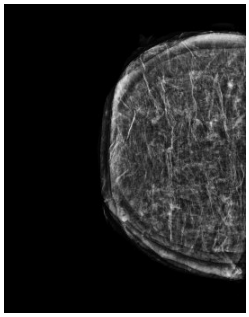
Equipamento	Fabricante	Tecnologia do detector	Ânodo/filtro	Exposições (n)
1A	Fabricante I	a-Se	W/Rh	3
1B	Fabricante I	a-Se	W/Rh	4
1C	Fabricante I	a-Se	W/Rh	4
2	Fabricante II	a-Si+CsI	W/Rh	2
3	Fabricante III	a-Se	W/Rh	1
4A	Fabricante IV	a-Si+CsI	Rh/Rh	8
4B	Fabricante IV	a-Si+CsI	Mo/Mo	1
4C	Fabricante IV	a-Si+CsI	W/Rh	1

Utilizaram-se para a avaliação os dados de testes da exposição de um *phantom* desenvolvido em pesquisa anterior pelo nosso grupo [4], [5]. O *phantom* é composto por camadas de filme de cloreto de polivinila (PVC) submersas em parafina gel, simulando tecidos adiposo e glandular humanos. Um aspecto característico de sua aplicação é que ele é composto por diferentes densidades mamárias que podem ser sobrepostas durante cada exposição, incluída a possibilidade de inserir simuladores de estruturas de interesse clínico, como nódulos e microcalcificações [4], [5].

Para a análise, as exposições foram realizadas utilizado o Molde 9A, cuja configuração específica simula uma mama predominantemente adiposa, para o qual a exposição aos raios X resulta em imagens como a ilustrada na Figura 1(a). Durante os ensaios, o *phantom* foi centralizado e fixado nos mamógrafos de forma padronizada, conforme ilustrado na Figura 1(b), com o objetivo de reduzir variações relacionadas ao posicionamento entre as aquisições.



(a) *Phantom* Molde 9A utilizado para os ensaios experimentais posicionado em um mamógrafo digital.



(b) Imagem mamográfica do *Phantom* Molde 9A obtida após a exposição.

Figura 1: *Phantom* Molde 9A utilizado nos ensaios experimentais.

A coleta de dados baseou-se na extração automatizada de parâmetros técnicos e dosimétricos diretamente dos metadados presentes nos cabeçalhos das imagens no padrão DICOM, utilizando o software Readicom [8]. Os valores de DEP foram obtidos a partir do parâmetro DICOM *Entrance Dose in mGy*, enquanto os valores analisados como DGM foram associados ao parâmetro *Organ Dose*. Além desses parâmetros dosimétricos, foram compiladas as seguintes variáveis:

- Tensão do tubo (kVp);
- Carga do tubo / Exposição (mAs);
- Espessura do objeto comprimido (mm);
- Material do ânodo (alvo) e do filtro;
- Modo de operação do controle de exposição;

A utilização dos metadados DICOM para o rastreamento de doses foi previamente avaliada pelo grupo. Nessas avaliações, os parâmetros dosimétricos registrados nos arquivos foram comparados com medições independentes realizadas com dosímetro calibrado, sendo observada boa concordância entre os resultados [8], [9].

Nessas avaliações dosimétricas prévias, a determinação da Dose Glandular Média foi baseada no formalismo recomendado para mamografia [10], expresso por

$$DGM = K_i \cdot g \cdot c \cdot s, \tag{1}$$

em que K_i representa o kerma no ar incidente e os fatores g, c e s consideram, respectivamente, a conversão para dose

glandular, a composição da mama e a qualidade espectral associada à combinação alvo/filtro.

Os dados foram organizados e processados em cinco frentes distintas de análise:

1) Comparação global de doses: avaliação comparativa das médias de DEP e DGM entre todos os fabricantes utilizando o Molde 9A;
2) Engenharia de aquisição: análise das interações entre os parâmetros de controle (relação kVp vs. mAs) para identificar padrões de compensação e tendências à sobre-exposição;
3) Análise por espessura: análise do subconjunto de exposições com espessura registrada de aproximadamente 48 mm, com o objetivo de reduzir a influência da espessura na comparação entre os sistemas;
4) Influência das inserções simuladas: análise das variações dos parâmetros de exposição associadas às diferentes configurações internas disponíveis do molde, incluindo nódulos redondos, espiculados e microcalcificações;
5) Análise de contraste RMS: avaliação quantitativa complementar das imagens por meio do contraste RMS (*Root Mean Square*) local e global [6].

Os resultados foram analisados de forma descritiva, considerando os valores individuais e, quando aplicável, a média, o desvio-padrão e o coeficiente de variação.

O contraste RMS global foi calculado a partir da razão entre o desvio-padrão e a média das intensidades dos pixels da região correspondente ao *phantom*. Para a análise local, o mesmo cálculo foi realizado em janelas de 64×64 pixels, sendo o resultado final obtido pela média dos valores calculados.

Como as imagens selecionadas apresentavam diferentes inserções, o contraste RMS foi utilizado como uma medida quantitativa complementar de contraste, sem representar isoladamente a qualidade diagnóstica dos sistemas.

Por fim, os valores de DGM foram comparados aos níveis de referência estabelecidos pela Instrução Normativa n° 92/2021 da Agência Nacional de Vigilância Sanitária (ANVISA).

III. Resultados e Discussões

A análise dos dados extraídos de forma automatizada, por meio dos metadados DICOM, permitiu organizar a avaliação dos oito mamógrafos nas cinco frentes metodológicas consideradas. Assim, foram correlacionadas as grandezas dosimétricas com o processo de aquisição de cada sistema e, ainda, com uma análise quantitativa de contraste das imagens.

A. Comparação Global de Doses

Os resultados iniciais analisaram o comportamento dos sistemas diante da atenuação do Molde 9A do *phantom*.

As Figuras 2 e 3 apresentam, respectivamente, os valores de DEP e DGM obtidos nas exposições para cada equipamento, juntamente com as respectivas médias.

Os equipamentos 1A, 1B e 1C apresentaram valores médios de DEP de $1,857 \pm 0,074$ mGy, $2,300 \pm 0,089$ mGy e $2,545 \pm 0,091$ mGy, respectivamente, com coeficientes de variação inferiores a 4%. Os respectivos valores médios de DGM foram $0,570 \pm 0,010$ mGy, $0,665 \pm 0,017$ mGy e $0,718 \pm 0,013$ mGy, também com baixa dispersão relativa.

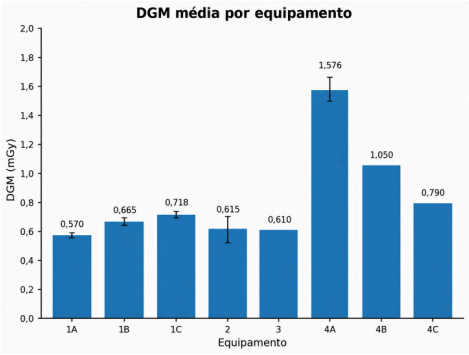


Figura 2: DEP média dos equipamentos analisados.

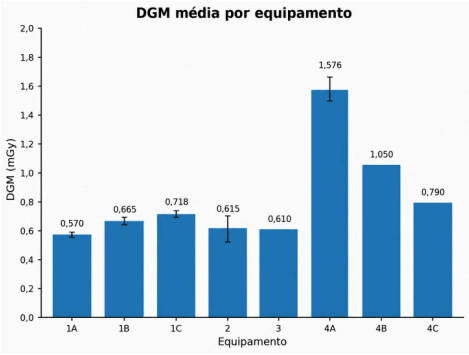


Figura 3: DGM média dos equipamentos analisados.

O equipamento 2 apresentou DEP média de $1,859 \pm 0,309$ mGy e DGM média de $0,615 \pm 0,120$ mGy, correspondendo a coeficientes de variação de 16,62% e 19,55%, respectivamente. Essa maior dispersão pode ser interpretada considerando que estavam disponíveis apenas duas exposições, realizadas em modos automáticos distintos.

O equipamento 4A apresentou os maiores valores médios do conjunto avaliado, com DEP de $5,246 \pm 0,472$ mGy e DGM de $1,576 \pm 0,073$ mGy. Os coeficientes de variação foram de 8,99% para DEP e 4,66% para DGM.

Os equipamentos 3, 4B e 4C foram representados por uma única exposição automática e, portanto, não foram calculadas medidas de dispersão para esses sistemas. Seus valores de DEP foram, respectivamente, 1,950 mGy, 3,760 mGy e 1,721 mGy, enquanto os valores de DGM foram 0,610 mGy, 1,050 mGy e 0,790 mGy.

Como o número de exposições variou entre os equipamentos, a comparação entre os sistemas foi interpretada principalmente com base nos valores observados e nas medidas de dispersão disponíveis.

B. Aquisição e Interação de Parâmetros Técnicos

As diferenças observadas nas doses foram analisadas em conjunto com os parâmetros radiográficos selecionados automaticamente pelos equipamentos.

A Figura 4 apresenta a relação entre a espessura do *phantom*, a tensão do tubo (kVp) e a carga de exposição (mAs) de cada mamógrafo.

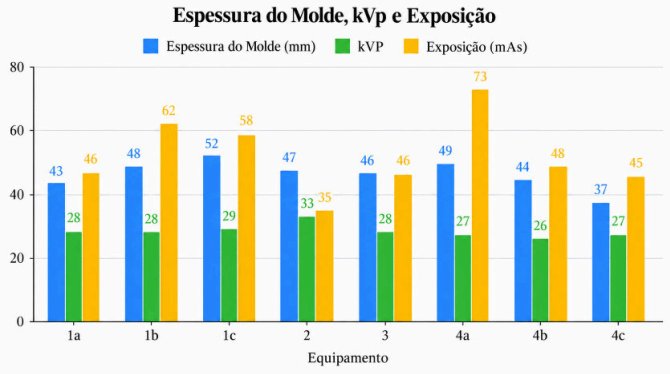


Figura 4: Espessura, kVp e exposição dos *phantoms* molde 9A por equipamento.

O sistema 2 operou com tensão de 33 kVp e carga média de aproximadamente 35 mAs nas exposições. O aumento da tensão do tubo tende a elevar a energia média do espectro de raios X e sua capacidade de penetração, podendo permitir redução da carga necessária para a formação da imagem em determinadas condições de aquisição [11], [12].

Ao contrário desse comportamento, o equipamento 4A operou com uma tensão menor, de 27 kVp, e uma carga de exposição média de aproximadamente 73 mAs, superior à empregada pelo equipamento 2. Essa configuração corresponde ao modo de exposição CNT, associado à combinação alvo/filtro Rh/Rh.

As diferenças observadas também podem estar relacionadas às características espectrais das combinações alvo/filtro utilizadas. Na configuração Rh/Rh, empregada pelo equipamento 4A, o alvo de Ródio ($Z = 45$) produz radiação característica com energias K_α e K_β em aproximadamente 20,2 keV e 22,7 keV, respectivamente, sendo o espectro posteriormente modulado pela filtração de Ródio. Nos sistemas W/Rh, por outro lado, o alvo de Tungstênio ($Z = 74$) produz predominantemente radiação de frenagem (*bremsstrahlung*) nas tensões, formando um espectro contínuo que também é modulado pelo filtro de Ródio [11], [12].

Nas condições avaliadas, essa diferença espectral, associada à maior tensão utilizada pelos sistemas W/Rh, pode ter contribuído para a maior penetração do feixe e para os menores valores de carga do tubo observados nesses equipamentos. Apesar disso, as diferenças de dose não podem ser atribuídas exclusivamente à combinação alvo/filtro, pois também podem estar relacionadas às características próprias de funcionamento e às condições operacionais de cada equipamento.

C. Análise por Espessura Fixa

Com o objetivo de reduzir a influência da espessura sobre a comparação dosimétrica, foi analisado separadamente o subconjunto de exposições com espessura registrada de 48 mm, conforme apresentado na Figura 5.

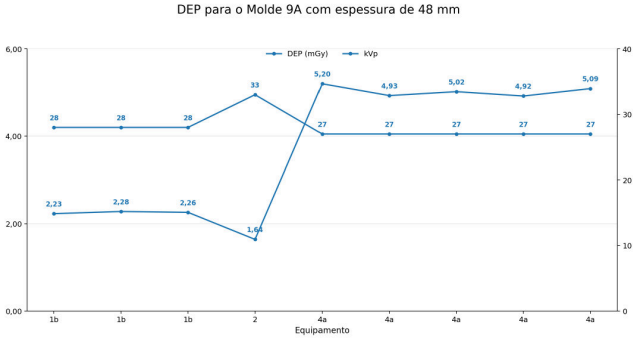


Figura 5: DEP para o *phantom* molde 9A com espessura de 48 mm.

Sob essa condição, o equipamento 1B apresentou três exposições realizadas a 28 kVp e entre 60 e 62 mAs, com valores de DEP entre 2,23 e 2,28 mGy. A DEP média foi de $2,257 \pm 0,025$ mGy, com CV de 1,12%. A DGM média foi de $0,657 \pm 0,006$ mGy, com CV de 0,88%.

O equipamento 2 apresentou uma exposição nessa espessura, realizada a 33 kVp e 31 mAs, com DEP de 1,641 mGy e DGM de 0,530 mGy.

No equipamento 4A, cinco exposições apresentaram espessura registrada de 48 mm, todas realizadas a 27 kVp e com carga entre 68 e 73 mAs. Os valores de DEP variaram de 4,921 a 5,196 mGy, com média de $5,031 \pm 0,116$ mGy, CV de 2,31%. A DGM média foi de $1,538 \pm 0,033$ mGy, com CV de 2,14%.

Mesmo considerando a mesma espessura registrada e o posicionamento padronizado do *phantom*, o equipamento 4A manteve valores de DEP superiores aos observados nos equipamentos 1B e 2 no subconjunto analisado.

D. Resposta do Controle Automático às Inserções de Lesões

O comportamento do controle automático de exposição foi avaliado a partir da inserção de diferentes estruturas internas simuladas no Molde 9A.

Para o equipamento 4A, considerando as exposições com espessura registrada de 48 mm, a DEP foi de 5,196 mGy para a configuração contendo nódulo espiculado e de 4,927 mGy para a configuração contendo nódulo redondo pequeno.

Como observado no conjunto de exposições realizadas nessa espessura, a dispersão dos valores de DEP foi baixa, com CV de 2,31%. As diferentes configurações internas do *phantom* estiveram, portanto, associadas a variações relativamente pequenas da DEP nas condições avaliadas. Esse comportamento é compatível com uma maior influência das características globais de atenuação do objeto, como espessura e composição, sobre a resposta do controle automático de exposição [11], [13].

E. Análise do Contraste RMS

Como complemento à avaliação dosimétrica, foram analisados os valores de contraste RMS local e global obtidos

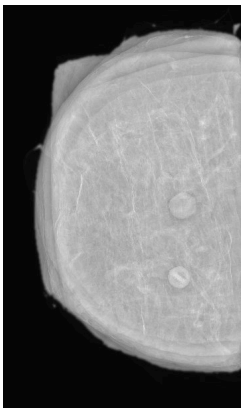
TABELA II
VALORES DE CONTRASTE RMS LOCAL E GLOBAL DAS IMAGENS SELECIONADAS

Equipamento	$\mathbf{C}_{\mathrm{RMS,local}}$	$\mathbf{C}_{\mathrm{RMS,global}}$
1A	0,3932	0,5399
1B	0,4192	0,5353
1C	0,4062	0,5087
2	0,3382	0,4197
3	0,2513	0,3402
4A	0,0516	0,0950
4B	0,0505	0,0791
4C	0,0189	0,0792

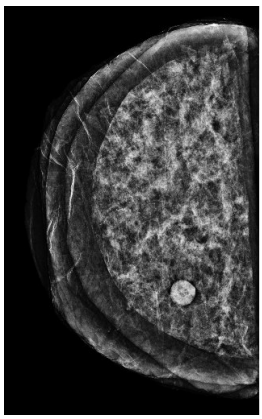
para uma imagem representativa de cada equipamento. Os resultados são apresentados na Tabela II.

Os equipamentos 1A, 1B e 1C apresentaram os maiores valores de contraste RMS global, entre 0,5087 e 0,5399. Os equipamentos 2 e 3 apresentaram valores intermediários, de 0,4197 e 0,3402, respectivamente, enquanto os equipamentos 4A, 4B e 4C apresentaram valores entre 0,0791 e 0,0950. Esse comportamento foi semelhante para o contraste RMS local.

A análise conjunta dos parâmetros dosimétricos e dos valores de contraste mostrou que maiores níveis de dose não estiveram associados a maiores valores de contraste RMS no conjunto avaliado. O equipamento 4A, por exemplo, apresentou a maior DEP média, $5,246 \pm 0,472$ mGy, enquanto os valores de contraste RMS local e global da imagem selecionada foram 0,0516 e 0,0950, respectivamente. Em contrapartida, os equipamentos 1A, 1B e 1C apresentaram menores valores médios de DEP e os maiores valores de contraste RMS entre as imagens analisadas. A Figura 6 apresenta as imagens mamográficas representativas obtidas nos equipamentos 1B e 4A.



(a) Imagem mamográfica do *phantom* molde 9A com nódulos e microcalcificaçõesdo equipamento 4A.



(b) Imagem mamográfica do *phantom* molde 9A com nódulos e microcalcificações do equipamento 1B.

Figura 6: Imagens mamográficas dos equipamentos avaliados utilizadas na comparação quantitativa de contraste.

Esse padrão pode estar relacionado à influência conjunta da tecnologia do detector, do processamento da imagem e dos parâmetros de aquisição. Como as imagens selecionadas não apresentavam exatamente as mesmas inserções, os valores de contraste RMS não devem ser interpretados isoladamente como uma classificação da qualidade diagnóstica dos equipamentos.

F. Comparação Normativa e Diretrizes de Otimização

Diante das diferenças observadas entre os equipamentos mamográficos, os valores de DGM foram situados em relação aos valores de referência estabelecidos pela Instrução Normativa n° 92/2021 da ANVISA [7], apresentados na Tabela III

TABELA III
VALORES DE REFERÊNCIA E TOLERÂNCIA PARA DGM EM MAMOGRAFIA SEGUNDO A IN N° 92/2021 DA ANVISA

PMMA (cm)	**Mama eq. (cm)**	**Referência (mGy)**	**Tolerância (mGy)**
2,0	2,1	0,6	$< 1,0$
3,0	3,2	1,0	$< 1,5$
4,0	4,5	1,6	$< 2,0$
4,5	5,3	2,0	$< 2,5$
5,0	6,0	2,4	$< 3,0$
6,0	7,5	3,6	$< 4,5$

Para as exposições em que a espessura não correspondia exatamente aos valores apresentados na normativa, foi adotada como referência a espessura normativa mais próxima. Considerando esse critério, todos os valores de DGM analisados permaneceram abaixo dos respectivos níveis de tolerância.

Apesar de permanecer abaixo dos níveis de tolerância, o equipamento 4A apresentou valores de dose superiores aos demais sistemas e, como apresentado na análise das imagens, esse aumento não foi acompanhado por maiores valores de contraste RMS. Sendo assim, os resultados indicam que o atendimento aos valores de referência dosimétrica não exclui a possibilidade de otimização dos parâmetros de aquisição, buscando a redução da dose sem comprometer as características da imagem, que são necessárias para o diagnóstico [10], [11], [13].

IV. CONCLUSÃO

Este estudo reforça a eficácia do uso de simuladores estruturados, como o desenvolvido pelo grupo, associado à extração automatizada de metadados DICOM como uma metodologia objetiva para a avaliação e o rastreamento dosimétrico em mamografia digital. A abordagem permitiu realizar análises em ambiente clínico sem a necessidade de exposição de pacientes à radiação, utilizando um simulador com características físicas e anatômicas compatíveis com exames mamográficos.

Os resultados mostraram diferenças entre os sistemas avaliados. O equipamento 4A (Fabricante IV - Hospital II) apresentou os maiores valores de DEP, comportamento que se manteve na análise das exposições com espessura registrada de 48 mm. Nessa condição, os valores de DEP permaneceram superiores aos observados nos equipamentos 1B e 2.

As diferenças observadas podem estar associadas à interação entre o modo de exposição, a combinação alvo/filtro, os parâmetros selecionados automaticamente e as características próprias de funcionamento de cada sistema. Embora a maior dose observada no equipamento 4A possa estar relacionada à combinação Rh/Rh e à estratégia automática de exposição, esse comportamento não pode ser atribuído exclusivamente a esses fatores.

Apesar dos maiores níveis de dose apresentados pelo equipamento 4A, os valores de DGM permaneceram abaixo dos níveis de tolerância adotados como referência pela normativa da ANVISA. A análise complementar das imagens mostrou ainda que o maior nível de dose desse equipamento não foi acompanhado por maiores valores de contraste RMS. Esses resultados indicam que o atendimento aos valores de referência dosimétrica não exclui a possibilidade de otimização dos parâmetros de aquisição, buscando a redução da dose sem comprometer as características da imagem necessárias ao diagnóstico.

Dessa forma, a extração automatizada de metadados DICOM mostrou-se útil como ferramenta de rastreamento e triagem dosimétrica, permitindo identificar equipamentos ou condições de aquisição com comportamento distinto e direcionar avaliações específicas de controle de qualidade, quando necessário.

AGRADECIMENTOS

Os autores agradecem à FAPESP o apoio financeiro e também à Gerência de Ensino e Pesquisa do Hospital Universitário da Universidade Federal de São Carlos (SP, Brasil) - Empresa Brasileira de Serviços Hospitalares (EBSERH), bem como a Oswaldo Jorge Neto e à Dra. Luciana B. Verçosa pelo suporte técnico.

REFERÊNCIAS

[1] Instituto Nacional de Câncer (INCA), *Diretrizes para a Detecção Precoce do Câncer de Mama no Brasil*. Rio de Janeiro, RJ, Brasil: INCA, 2015.

[2] A. C. S. Xavier, V. S. M. Barros, F. Mello, and H. J. Khoury, "Avaliação da dose glandular média em sistemas digitais e convencionais de mamografia," in *Anais da Joint Conference RADIO 2014*, Gramado, RS, Brasil, 2014.

[3] International Atomic Energy Agency (IAEA), *Dosimetry in Diagnostic Radiology: An International Code of Practice*, Technical Reports Series No. 457. Vienna, Austria: IAEA, 2007.

[4] M. A. Z. Sousa, B. R. N. Matheus, and H. Schiabel, "Development of a structured breast phantom for evaluating CADe/Dx schemes applied on 2D mammography," *Biomedical Physics & Engineering Express*, vol. 4, no. 4, Art. no. 045018, 2018.

[5] M. A. Z. Sousa, "Desenvolvimento de um objeto simulador de mama: investigações da percepção visual da imagem e do desempenho de esquemas CADx," Ph.D. dissertation, Depto. de Engenharia Elétrica e de Computação, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP, Brasil, 2017.

[6] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[7] Agência Nacional de Vigilância Sanitária (ANVISA), "Instrução Normativa IN n° 92, de 27 de maio de 2021: Requisitos sanitários para a garantia da qualidade e da segurança em sistemas de mamografia," Brasília, DF, Brasil, 2021.

[8] H. Schiabel, B. Barufaldi, and E. M. Ruberti Filha, "Investigations on a computer application for tracking the mean glandular breast dose profile in mammography," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON 2019)*, vol. 76, pp. 869–873, Coimbra, Portugal, 2019. doi: 10.1007/978-3-030-31635-8_104.

[9] H. Schiabel, E. M. Ruberti Filha, O. Jorge Neto, and L. B. Verçosa, "Radiation dose tracking in digital mammography: Evaluation of population profiles through automatic data extraction from the DICOM header," *London Journal of Medical and Health Research*, vol. 23, no. 11, pp. 1–25, 2023.

[10] D. R. Dance, S. Christofides, A. D. A. Maidment, I. D. McLean, and K. H. Ng, Eds., *Diagnostic Radiology Physics: A Handbook for Teachers and Students*. Vienna, Austria: International Atomic Energy Agency, 2014.

[11] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt, Jr., and J. M. Boone, *The Essential Physics of Medical Imaging*, 4th ed. Philadelphia, PA, USA: Wolters Kluwer, 2021.

[12] J. M. Boone, T. R. Fewell, and R. J. Jennings, "Molybdenum, rhodium, and tungsten anode spectral models using interpolating polynomials with application to mammography," *Medical Physics*, vol. 24, no. 12, pp. 1863–1874, Dec. 1997.

[13] M. S. Ko, H. H. Kim, J. H. Cha, H. J. Shin, J. H. Kim, and M. J. Kim, "Dose reduction in automatic optimization parameter of full field digital mammography: Breast phantom study," *Journal of Breast Cancer*, vol. 16, no. 1, pp. 90–96, 2013.

Análise de Risco de Diabetes: Um Aplicativo Móvel com Aprendizado de Máquina Explicável para Educação em Saúde

Ana Kelly N. Costa*†, Fernanda R. Fernandes*†, Débora F. Costa*†, Náthalee C. de A. Lima‡, Rosana C. B. Rego‡
*Computational Intelligence Laboratory (CiLab), Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
†Information Technology, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
‡Departamento de Engenharias e Tecnologia, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
ana.costa03379@alunos.ufersa.edu.br, fernanda.fernandes36236@alunos.ufersa.edu.br, debora.costa22396@alunos.ufersa.edu.br,
nathalee.almeida@ufersa.edu.br, rosana.rego@ufersa.edu.br

***Resumo*—O Diabetes Mellitus é uma doença crônica de alta prevalência mundial, associada a diversas complicações de saúde e cuja identificação precoce pode contribuir para o acompanhamento adequado dos indivíduos em risco. Este trabalho apresenta o desenvolvimento de um aplicativo móvel, denominado *Análise de Risco de Diabetes*, que integra técnicas de aprendizado de máquina para atuar como ferramenta de triagem do risco de diabetes, associada a funcionalidades voltadas à educação em saúde. A aplicação foi desenvolvida com uma arquitetura cliente-servidor, utilizando Flutter no desenvolvimento móvel e Python com FastAPI na construção da API REST responsável pelo processamento das predições. Foram avaliados cinco algoritmos de classificação, sendo o modelo Multilayer Perceptron (MLP) selecionado para compor a aplicação por apresentar o melhor desempenho geral, alcançando acurácia de 97,08%, precisão de 99,00% e F1-Score de 81,00% para a classe diabética. O modelo foi treinado utilizando seis parâmetros: Índice de Massa Corporal, Hemoglobina Glicada, idade, glicose sanguínea, hipertensão e doença cardíaca. A técnica *Local Interpretable Model-agnostic Explanations* (LIME) foi empregada para analisar localmente as predições do modelo, permitindo identificar os atributos que mais contribuíram para a classificação de instâncias individuais. Os resultados demonstram a possibilidade de integrar aprendizado de máquina e mecanismos de explicabilidade a uma aplicação móvel voltada à triagem e à educação em saúde, ressaltando que as estimativas produzidas possuem caráter complementar e não substituem a avaliação de profissionais de saúde.**

***Index Terms*—Diabetes Mellitus, Aprendizado de Máquina, Multilayer Perceptron, Explicabilidade, Aplicativo Móvel, Educação em Saúde**

I. Introdução

O Diabetes Mellitus (DM) representa um importante problema de saúde pública, e um dos principais desafios para os sistemas de saúde em todo o mundo, apresentando crescimento contínuo no número de casos [1]. A doença está associada ao desenvolvimento de complicações cardiovasculares, renais, neurológicas e visuais, comprometendo significativamente a qualidade de vida dos pacientes [2].

Apesar dos avanços nas estratégias de diagnóstico e monitoramento do diabetes, ainda existe um conhecimento limitado por parte da população em relação aos principais fatores de risco associados à doença. Essa lacuna de informação contribui para a identificação tardia de indivíduos em risco e, consequentemente, para a progressão silenciosa da condição ao longo do tempo [3].

Nesse contexto, fatores metabólicos como obesidade, níveis elevados de hemoglobina glicada (HbA1c) e alterações na glicemia de jejum continuam sendo amplamente reconhecidos como indicadores importantes de risco. No entanto, a ausência de valores críticos nesses marcadores não exclui a possibilidade de desenvolvimento futuro da doença, especialmente quando associados a outros determinantes clínicos e comportamentais [4].

Entre os marcadores clínicos amplamente utilizados na triagem e no rastreamento do diabetes mellitus, destacam-se o Índice de Massa Corporal (IMC), associado à obesidade como um dos principais fatores de risco; a Hemoglobina Glicada (HbA1c), que reflete a média dos níveis glicêmicos dos últimos meses; a idade, uma vez que o risco aumenta progressivamente ao longo da vida; e a glicose sanguínea. Adicionalmente, comorbidades como hipertensão e doenças cardíacas são determinantes agravantes frequentemente associados ao diabetes. Esses seis parâmetros compõem um referencial primário na avaliação de risco da doença [2]. Adicionalmente, o desbalanceamento entre classes positivas e negativas é um desafio recorrente em bases de dados clínicas, sendo a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) amplamente empregada para mitigar esse efeito [5]. Por fim, no que diz respeito à explicabilidade dos modelos, a técnica *Local Interpretable Model-agnostic Explanations* (LIME) [6] tem sido adotada para tornar as predições de modelos complexos mais transparentes, aspecto especialmente relevante em aplicações voltadas à saúde.

Paralelamente, a Inteligência Artificial (IA), especialmente por meio de técnicas de Aprendizado de Máquina (ML), tem demonstrado resultados promissores na área da saúde, possibilitando a identificação de padrões clínicos e o apoio à tomada de decisão [7]. Diversos estudos têm empregado algoritmos capazes de estimar o risco de diabetes com elevada

precisão. [13]

Entretanto, a maioria dessas pesquisas permanece restrita ao ambiente acadêmico e não disponibiliza ferramentas acessíveis diretamente à população. Nesse sentido, torna-se relevante integrar modelos de ML em aplicações que podem contribuir para aproximar essas tecnologias da população e ampliar o acesso a informações relacionadas à prevenção e aos fatores associados ao diabetes. [13]

Diante desse cenário, este trabalho apresenta o desenvolvimento de um aplicativo que integra técnicas de ML para atuar como uma ferramenta de triagem do risco de diabetes, aliada a funcionalidades voltadas à educação em saúde. A aplicação utiliza dados clínicos e características individuais (IMC, HbA1c, glicose, idade, presença de hipertensão e de doenças cardíacas) fornecidas pelo usuário para realizar uma estimativa de risco baseada no modelo treinado. Vale salientar que a saída do aplicativo atua estritamente como uma triagem computacional, não constituindo um diagnóstico médico formal, mas sim um indicativo para auxiliar na identificação de indivíduos que necessitem buscar avaliação profissional. Além da funcionalidade preditiva, o aplicativo disponibiliza informações sobre sintomas, fatores de risco, prevenção e orientações básicas, buscando ampliar o acesso ao conhecimento e incentivar a procura por atendimento qualificado.

II. Trabalhos Relacionados

A aplicação de técnicas de ML para a predição do Diabetes Mellitus tem crescido significativamente nos últimos anos, impulsionada pela disponibilidade de grandes bases de dados clínicos e pela necessidade de desenvolver ferramentas capazes de auxiliar na identificação precoce da doença. Além da busca por modelos cada vez mais precisos, pesquisas recentes também têm enfatizado a importância da Inteligência Artificial Explicável (Explainable Artificial Intelligence - XAI), permitindo compreender os fatores que influenciam as decisões dos algoritmos e aumentando a confiança em sua utilização na área da saúde.

Costa *et al.* [10] desenvolveram e avaliaram modelos de aprendizado de máquina para a predição de diabetes, utilizando cinco algoritmos de classificação: Random Forest, Regressão Logística, KNN, SVC e MLP. O estudo também empregou técnicas de explicabilidade para analisar a influência das variáveis nas predições, destacando a relevância dos níveis de glicose sanguínea e hemoglobina glicada. Os resultados demonstraram o potencial de modelos de aprendizado de máquina aliados a mecanismos de explicabilidade para apoiar sistemas de predição de diabetes. Embora o estudo tenha se concentrado na avaliação dos modelos e na análise de suas predições, o presente trabalho amplia essa abordagem ao integrar o modelo preditivo e o mecanismo de explicabilidade a uma aplicação móvel voltada à triagem e à educação em saúde.

[7] realizaram uma revisão sobre aplicações de IA na área da saúde, destacando o potencial dos algoritmos de ML no auxílio ao diagnóstico e à predição de doenças crônicas, incluindo o Diabetes Mellitus. Os autores discutem o emprego de modelos como Redes Neurais Artificiais, *Random Forest* e Máquinas de Vetores de Suporte, ressaltando que essas técnicas apresentam elevado potencial para apoiar a tomada de decisão clínica. Entretanto, também apontam desafios relacionados à explicabilidade dos modelos e à sua adoção em ambientes clínicos reais.

Em [8] desenvolveram modelos de ML para predição de diabetes utilizando dados do Behavioral Risk Factor Surveillance System (BRFSS). O estudo comparou diferentes algoritmos de classificação e empregou técnicas de XAI para analisar a influência das variáveis utilizadas pelo modelo. Os resultados evidenciaram que fatores como idade, índice de massa corporal e indicadores metabólicos estão entre as características mais relevantes para a classificação dos pacientes, demonstrando que a utilização de métodos explicáveis contribui para aumentar a transparência das predições.

No estudo de [9] foi investigado o uso de modelos explicáveis para avaliação do risco de diabetes, enfatizando que técnicas como LIME e SHAP possibilitam compreender quais atributos mais influenciam as decisões dos algoritmos de ML. Segundo os autores, a explicabilidade representa um requisito essencial para aplicações médicas, pois permite que profissionais da saúde compreendam as justificativas das predições e utilizem essas informações como apoio ao processo de decisão clínica.

Foi proposto por [11] uma abordagem que combina a técnica SMOTE para balanceamento das classes com o método SHAP para explicação das predições realizadas pelos modelos. Os autores compararam diferentes algoritmos de classificação, analisando o impacto do balanceamento de dados sobre as métricas de desempenho. Os resultados indicaram melhora na capacidade de identificação da classe minoritária, além de fornecer explicações detalhadas sobre a contribuição de cada variável para as predições realizadas.

Observa-se que os estudos encontrados na literatura concentram seus esforços principalmente na comparação entre algoritmos de ML e na avaliação de técnicas de XAI, demonstrando elevados índices de desempenho na predição do Diabetes Mellitus. Entretanto, a maioria dessas pesquisas permanece restrita ao ambiente acadêmico, disponibilizando apenas modelos experimentais ou protótipos voltados para pesquisadores e profissionais da área. Poucos trabalhos apresentam aplicações móveis destinadas ao público geral que integrem, em uma única solução, um modelo preditivo, mecanismos de explicabilidade e recursos voltados à educação em saúde.

Nesse contexto, o presente trabalho diferencia-se por propor um aplicativo móvel multiplataforma que integra um modelo de ML baseado em Multilayer Perceptron (MLP) para estimativa do risco de Diabetes Mellitus, empregando a técnica LIME para explicar as predições realizadas. Além da funcionalidade preditiva, a aplicação disponibiliza conteúdo educativo sobre sintomas, fatores de risco, prevenção e controle da doença, buscando aproximar os avanços da IA do público em geral e contribuir para a conscientização e o incentivo à busca por atendimento médico quando necessário.

III. METODOLOGIA

A. Arquitetura do aplicativo

O aplicativo Análise de Risco de Diabetes foi projetado com base em uma arquitetura cliente-servidor de três camadas: (i) Camada de Apresentação, composta pela interface do aplicativo móvel com a qual o usuário interage diretamente; (ii) Camada de Negócio, responsável pelo processamento das requisições por meio de uma API REST; e (iii) Camada de Dados, onde o modelo de ML realiza as predições.

A comunicação entre o aplicativo e o servidor ocorre por meio de requisições HTTP seguras. Ao concluir o formulário de checkup, o aplicativo envia os dados do usuário ao servidor por meio de uma requisição HTTP POST. O servidor processa essas informações utilizando o modelo preditivo e retorna ao cliente o resultado da avaliação de risco (baixo, moderado ou alto), acompanhado das orientações correspondentes.

Esse fluxo reduz o processamento realizado no dispositivo móvel, permitindo que modelos de ML mais complexos sejam executados no servidor. Além disso, centraliza a atualização do modelo preditivo, facilitando sua manutenção e possibilitando sua evolução sem a necessidade de modificar a aplicação instalada no dispositivo do usuário.

B. Casos de Uso

Os casos de uso foram definidos com base nas funcionalidades centrais do aplicativo e no perfil do público-alvo: usuários do público geral, sem necessidade de conhecimento técnico prévio em saúde ou tecnologia. O único ator identificado no sistema é o *Usuário*, que representa qualquer indivíduo que acessa a aplicação. A Tabela I descreve os casos de uso identificados para o sistema.

Tabela I: Casos de Uso do Aplicativo *Análise de Risco de Diabetes*

ID	Caso de Uso	Descrição	Fluxo Principal
UC01	Acessar Aplicativo	O usuário acessa o aplicativo e visualiza a tela inicial, que apresenta informações introdutórias sobre diabetes e orientações gerais de saúde.	(1) Usuário inicia o aplicativo; (2) Sistema exibe a tela inicial; (3) Usuário visualiza as informações e as opções de navegação.
UC02	Realizar Checkup de Diabetes	O usuário realiza a avaliação de risco informando os seis parâmetros clínicos utilizados pelo modelo: IMC, HbA1c, idade, glicose sanguínea, hipertensão e doença cardíaca.	(1) Usuário acessa a opção de checkup; (2) Sistema apresenta o formulário; (3) Usuário informa os dados solicitados; (4) Usuário confirma os dados; (5) Sistema realiza a avaliação de risco.
UC03	Visualizar Resultado de Risco	O usuário visualiza o resultado da avaliação de risco após o processamento dos dados informados.	(1) Sistema processa os dados informados; (2) Sistema apresenta a classificação de risco como baixo, moderado ou alto; (3) Sistema apresenta orientações relacionadas ao resultado.
UC04	Acessar Conteúdo Educativo	O usuário acessa conteúdos informativos relacionados ao diabetes, com o objetivo de ampliar seu conhecimento sobre a doença e seus fatores de risco.	(1) Usuário acessa a seção educativa; (2) Sistema apresenta informações sobre diabetes, sintomas, fatores de risco, prevenção e controle.
UC05	Receber Orientação para Atendimento Profissional	O usuário recebe orientação para procurar atendimento de um profissional de saúde, especialmente após a visualização do resultado da avaliação de risco.	(1) Sistema apresenta orientação para busca de atendimento profissional; (2) Usuário é informado de que a avaliação realizada pelo aplicativo não substitui o diagnóstico clínico.

A Tabela I apresenta as principais interações entre o usuário e o aplicativo, contemplando desde o acesso às informações iniciais e aos conteúdos educativos até a realização da avaliação de risco e a visualização do resultado. Os casos de uso também evidenciam o caráter de triagem da aplicação, uma vez que o resultado apresentado tem como finalidade orientar o usuário quanto à necessidade de buscar avaliação de um profissional de saúde, não substituindo o diagnóstico clínico.

C. Tecnologias utilizadas

Para o desenvolvimento do aplicativo móvel, foi adotado o framework **Flutter**, mantido pelo Google, que permite a criação de aplicações multiplataforma (Android e iOS) a partir de uma única base de código escrita na linguagem Dart. Essa escolha justifica-se pela eficiência no ciclo de desenvolvimento, pelo desempenho nativo em dispositivos móveis e pela facilidade na construção de interfaces responsivas e visualmente agradáveis.

No lado do servidor, foi utilizado **Python** em conjunto com o framework **FastAPI** para a construção da API REST responsável pelo recebimento das requisições do aplicativo, processamento pelo modelo preditivo e retorno dos resultados. O FastAPI foi selecionado por sua alta performance, suporte nativo à programação assíncrona e geração automática de documentação por meio do OpenAPI/Swagger.

D. Modelo de Machine Learning

Para o treinamento e a avaliação dos modelos preditivos, utilizou-se o *Diabetes Prediction Dataset* **[12]**, uma base de dados de domínio público disponibilizada na plataforma

Kaggle. O repositório original conta com 100.000 registros e engloba diversas características demográficas e clínicas dos pacientes. Dentre os atributos disponibilizados, foram selecionados seis parâmetros para compor as variáveis de entrada dos modelos: (i) Índice de Massa Corporal (IMC), obtido a partir do peso e da altura do paciente; (ii) Hemoglobina Glicada (HbA1c); (iii) Idade; (iv) Glicose sanguínea; (v) Hipertensão, representada por um indicador binário; e (vi) Doença Cardíaca, também representada por um indicador binário.

A base apresenta assimetria na distribuição das classes, sendo constituída por aproximadamente 91,5% de instâncias de indivíduos não diabéticos e 8,5% de indivíduos diabéticos. Como etapa de pré-processamento, as variáveis contínuas foram padronizadas por meio do *StandardScaler*. Em razão do desbalanceamento observado, também foi avaliada a aplicação da técnica SMOTE (*Synthetic Minority Over-sampling Technique*). A comparação entre os cenários com e sem balanceamento indicou que, embora a aplicação do SMOTE tenha proporcionado aumento no *recall* da classe positiva, o cenário sem balanceamento apresentou melhor desempenho global, considerando principalmente a acurácia e o F1-Score. Dessa forma, optou-se pela utilização da distribuição original das classes na configuração final dos modelos empregados na aplicação.

Os dados foram divididos em 80% para treinamento e 20% para teste, sendo o conjunto de teste reservado para a avaliação dos modelos. Foram utilizados cinco algoritmos de classificação:*Random Forest*, Regressão Logística, *KNN*, *SVC* e MLP. A comparação entre os modelos considerou as métricas de acurácia, precisão, *recall* e F1-Score. Os principais parâmetros explicitamente definidos na implementação são apresentados na Tabela II.

Tabela II: Parâmetros utilizados nos modelos avaliados.

Modelo	**Parâmetros**
Random Forest	$n_estimators = 100$, $random_state = 42$
Regressão Logística	$max_iter = 1000$
KNN	$n_neighbors = 5$
SVC	$kernel =$ linear
MLP	$hidden_layer_sizes = (50, 50)$, $max_iter = 1000$, $random_state = 42$

Os parâmetros foram definidos diretamente na configuração dos modelos, sem a realização de procedimentos automatizados de otimização, como *Grid Search*. Entre os modelos avaliados, o MLP apresentou o melhor desempenho geral e foi selecionado como modelo final da aplicação. O MLP foi configurado com duas camadas ocultas, cada uma contendo 50 neurônios, totalizando uma arquitetura $(50, 50)$. O treinamento foi limitado a 1000 iterações e foi utilizado *random state*=42 para favorecer a reprodutibilidade dos resultados.

E. Explicabilidade do Modelo

Para conferir explicabilidade às predições do modelo MLP, foi aplicada a técnica LIME. O método fornece explicações locais para predições individuais por meio da aproximação do comportamento do modelo na vizinhança da instância analisada, permitindo identificar quais variáveis exerceram maior influência na classificação de cada paciente.

O LIME foi aplicado de forma *post-hoc*, ou seja, após o treinamento do modelo. A técnica analisa o comportamento do modelo treinado sem alterar seus parâmetros, os dados de entrada ou o pipeline de classificação utilizado pela API. Dessa forma, a predição de risco retornada ao aplicativo é produzida exclusivamente pelo modelo MLP, enquanto o LIME é utilizado como ferramenta de análise das decisões do modelo.

F. Classificação das Faixas de Risco

Para a apresentação do resultado da avaliação de risco no aplicativo, a probabilidade retornada pelo modelo MLP foi categorizada em três faixas: *Risco Baixo*, para probabilidades inferiores a 35%; *Risco Moderado*, para probabilidades entre 35% e 69%; e *Risco Alto*, para probabilidades iguais ou superiores a 70%.

Essa categorização foi utilizada exclusivamente para apresentação do resultado da triagem ao usuário, não correspondendo a categorias clínicas ou critérios de diagnóstico. O objetivo é facilitar a comunicação do resultado e orientar o usuário quanto à necessidade de buscar acompanhamento de um profissional de saúde.

G. Desenvolvimento da interface

A interface do aplicativo foi projetada com foco na acessibilidade e na experiência do usuário do público geral, sem presumir familiaridade prévia com aplicações de saúde. O design segue princípios de usabilidade amplamente reconhecidos, como clareza visual, *feedback* imediato e navegação intuitiva e progressiva.

A tela de início apresenta uma mensagem de conscientização sobre o diabetes, com destaque para a importância do diagnóstico precoce e um incentivo explícito à busca por atendimento médico profissional. O aplicativo deixa claro ao usuário, desde a primeira interação, que a avaliação fornecida representa uma estimativa de risco e não substitui o diagnóstico clínico.

A paleta de cores adota tons de azul e verde, associados à saúde e ao bem-estar, com alto contraste para garantir legibilidade em diferentes condições de luminosidade e acessibilidade para usuários com baixa acuidade visual. O fluxo de navegação foi estruturado de forma linear e progressiva: conscientização $\rightarrow$ checkup $\rightarrow$ resultado $\rightarrow$ conteúdo educativo, minimizando a carga cognitiva e guiando o usuário de forma clara ao longo da avaliação.

A Figura 1 apresenta a tela inicial do aplicativo, enquanto a Figura 2 apresenta a tela utilizada para a realização do checkup.

H. Funcionalidades do aplicativo

O aplicativo *Análise de Risco de Diabetes* disponibiliza as seguintes funcionalidades principais:

- Tela de Início e Conscientização: apresenta ao usuário informações gerais sobre o Diabetes Mellitus, incluindo

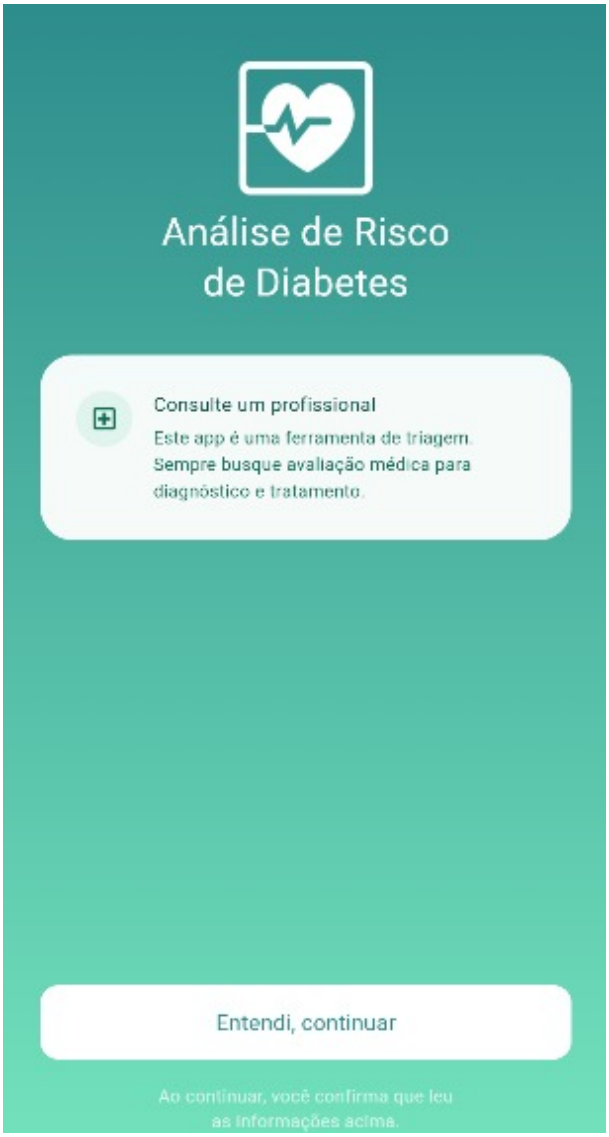


Figura 1: Tela inicial do aplicativo.

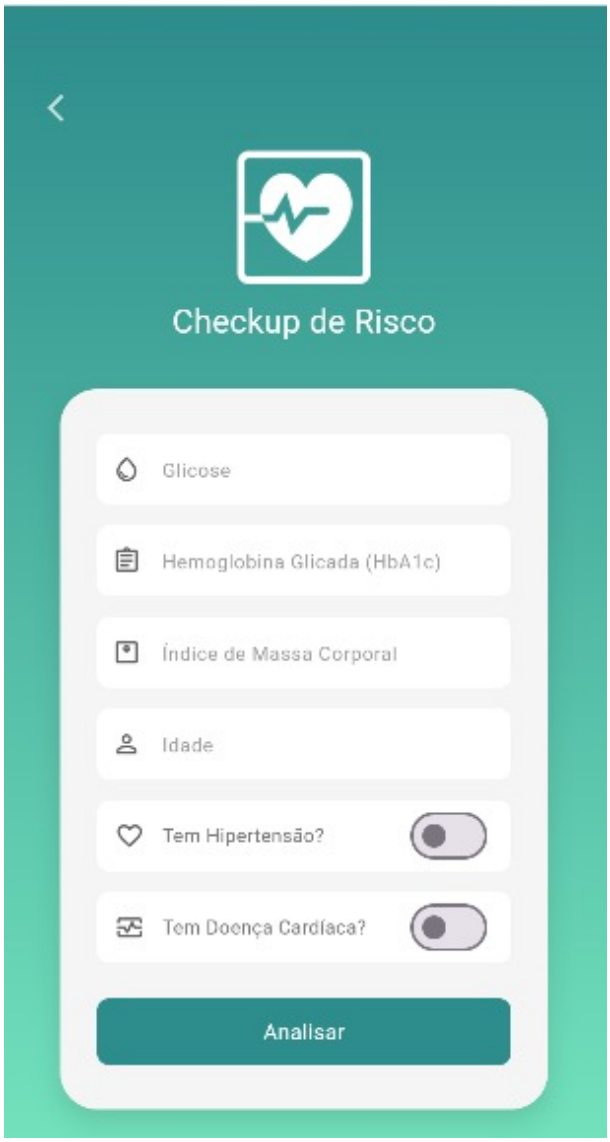


Figura 2: Tela de checkup do aplicativo.

definição, principais sintomas e fatores de risco associados. Uma mensagem de alerta destaca a importância de buscar orientação de um profissional de saúde para confirmação diagnóstica, reforçando o caráter complementar — e não substitutivo — da aplicação.

- Checkup de Diabetes: permite ao usuário preencher um formulário com seis parâmetros de saúde: Índice de Massa Corporal (IMC), nível de Hemoglobina Glicada (HbA1c), idade, nível de glicose sanguínea, e a presença prévia de hipertensão ou doenças cardíacas. O aplicativo orienta o usuário sobre o preenchimento e, ao concluir, os dados são enviados ao servidor para processamento pelo modelo preditivo.
- Resultado da Avaliação de Risco: exibe, de forma clara e acessível, o resultado do checkup, classificando o risco em três categorias: *baixo*, *moderado* ou *alto*. Cada categoria é acompanhada de orientações específicas e recomendações de acompanhamento médico.
- Conteúdo Educativo: disponibiliza informações sobre prevenção, controle e convivência com o diabetes, com linguagem acessível e adaptada ao público geral. O objetivo é promover a educação em saúde e o empoderamento do usuário para a adoção de hábitos mais saudáveis.

IV. Resultados e Discussões

Nesta seção são apresentados e discutidos os resultados obtidos durante o desenvolvimento e a avaliação do aplicativo proposto. Inicialmente, são analisados os desempenhos dos modelos de aprendizado de máquina considerados, incluindo a comparação entre as abordagens com e sem balanceamento das classes. Em seguida, são apresentados os resultados relacionados à explicabilidade das predições por meio do LIME e à integração do modelo ao aplicativo móvel.

A. Desempenho do Modelo

Os experimentos de avaliação foram realizados utilizando o *Diabetes Prediction Dataset*, composto por registros de indivíduos com características clínicas e demográficas relacionadas à ocorrência de diabetes. Os dados foram divididos em conjuntos de treinamento e teste, correspondendo a 80% e 20% dos registros, respectivamente. Os modelos foram treinados utilizando o conjunto de treinamento e, posteriormente, avaliados no conjunto de teste por meio das métricas de acurácia, precisão, revocação, F1-score. Sem a aplicação de técnicas de balanceamento das classes, os modelos apresentaram os resultados apresentados na Tabela III.

Tabela III: Comparação dos modelos de ML

Modelo	Acurácia	Precisão	Recall	F1-Score
Random Forest	96,68	91,00	**69,00**	79,00
Regressão Logística	95,69	85,00	63,00	72,00
KNN	96,32	91,00	65,00	76,00
SVC	95,78	92,00	58,00	71,00
MLP	**97,08**	**99,00**	68,00	**81,00**

A Tabela III apresenta o desempenho dos modelos de classificação avaliados no conjunto de teste. O MLP apresentou a maior acurácia, de 97,08%, além da maior precisão (99,00%) e do maior F1-Score (81,00%) para a classe diabética. O *Random Forest* apresentou desempenho próximo, com acurácia de 96,68% e F1-Score de 79,00%, enquanto KNN, Regressão Logística e SVC apresentaram F1-Scores de 76,00%, 72,00% e 71,00%, respectivamente.

Embora o MLP tenha apresentado o melhor desempenho geral entre os modelos avaliados, observa-se que o recall da classe diabética foi de 68,00%, indicando que parte dos indivíduos pertencentes à classe positiva não foi identificada pelo modelo. Esse resultado deve ser interpretado considerando o desbalanceamento da base, na qual 91,18% das amostras pertencem à classe não diabética e apenas 8,82% à classe diabética. Dessa forma, a análise das métricas específicas da classe positiva é relevante para evitar que a elevada proporção da classe majoritária produza uma percepção superestimada do desempenho do modelo.

De modo geral, os resultados evidenciam que o MLP apresentou uma combinação mais favorável entre as métricas analisadas, especialmente em relação à acurácia, precisão e F1-Score. A diferença observada em relação aos demais modelos reforça sua escolha para as etapas posteriores de análise e explicabilidade. Além disso, a avaliação conjunta das métricas permitiu considerar não apenas o desempenho global, mas também o comportamento do modelo na identificação da classe diabética.

Também foi avaliado o emprego da técnica SMOTE como estratégia de balanceamento das classes, conforme abordagens descritas na literatura [11]. A aplicação do SMOTE contribuiu para o aumento do *recall* da classe positiva, indicando maior capacidade de identificação dos indivíduos pertencentes à classe minoritária. Entretanto, essa melhoria foi acompanhada por redução da acurácia geral e da precisão da classe positiva. Diante desse comportamento, foi adotado o cenário sem balanceamento para o modelo final, por apresentar melhor desempenho global segundo as métricas consideradas.

B. Análise da matriz de confusão

A matriz de confusão do modelo MLP, apresentada na Figura 3, permite observar o desempenho do classificador em cada classe. Dos indivíduos classificados como não diabéticos, 17.503 foram corretamente identificados, enquanto 6 foram classificados incorretamente como diabéticos. Para a classe diabética, 1.166 indivíduos foram corretamente classificados, ao passo que 555 foram classificados como não diabéticos.

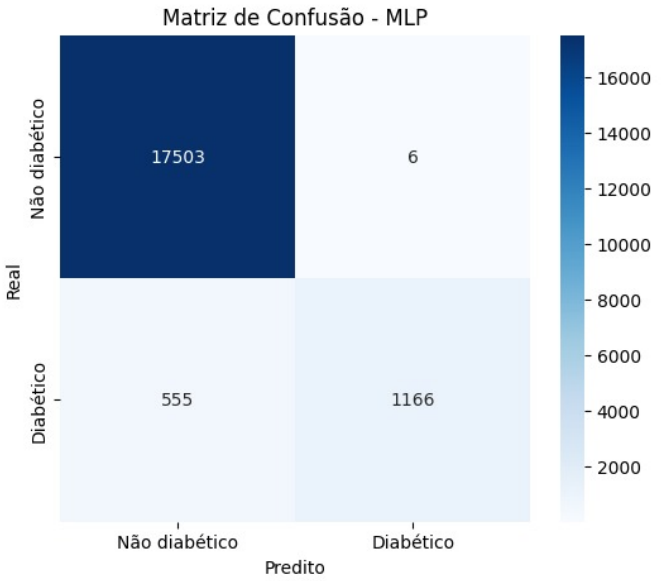


Figura 3: Matriz de confusão do modelo MLP.

Esses resultados evidenciam que, embora o modelo apresente elevada acurácia global, ainda são observados falsos negativos na identificação da classe diabética, aspecto particularmente importante em uma tarefa de avaliação de risco, uma vez que indivíduos diabéticos podem ser classificados como não diabéticos.

C. Análise de Explicabilidade via LIME

Para complementar a explicabilidade do modelo, foi empregada a técnica LIME, com o objetivo de explicar as predições realizadas pelo modelo MLP em instâncias individuais. Diferentemente de métodos de explicabilidade global, o LIME gera explicações locais, identificando quais atributos tiveram maior influência na classificação de cada amostra.

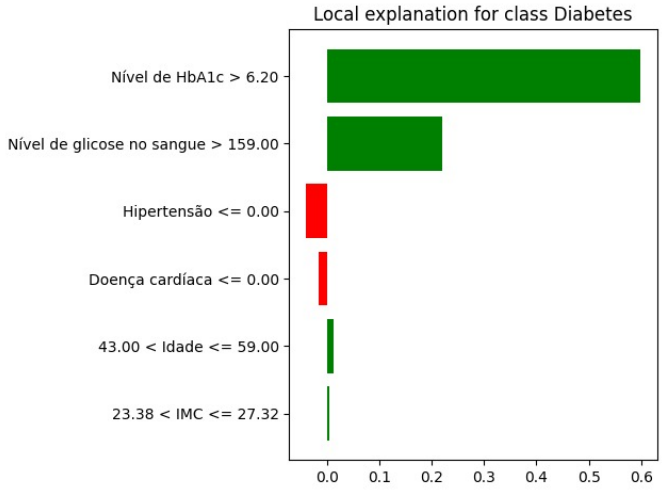


Figura 4: Explicação local da predição do modelo MLP utilizando LIME.

A Figura 4 apresenta uma explicação local para uma instância classificada como pertencente à classe diabetes. Observa-se que o *nível de HbA1c* apresentou a maior contribuição positiva para a classificação, seguido pelo *nível de glicose no sangue*. As condições apresentadas no gráfico, com valores de HbA1c superiores a 6,20 e de glicose no sangue superiores a 159,00, exerceram maior influência na decisão do modelo.

Em contrapartida, as condições de *hipertensão* e *doença cardíaca* apresentaram contribuições negativas para a classificação dessa instância, ambas associadas à ausência dessas condições. A *idade*, representada pelo intervalo entre 43 e 59 anos, e o *IMC*, situado entre 23,38 e 27,32, apresentaram contribuições positivas de menor magnitude em comparação às variáveis laboratoriais.

Esses resultados indicam que, para essa instância específica, o modelo concentrou sua decisão principalmente nos níveis de HbA1c e glicose no sangue, enquanto as demais características apresentaram influência menos expressiva. A explicação local fornecida pelo LIME permite, assim, visualizar quais atributos contribuíram para a classificação realizada pelo modelo, favorecendo maior transparência em seu processo de decisão.

V. Conclusão

Este trabalho apresentou o desenvolvimento do aplicativo *Análise de Risco de Diabetes*, uma solução móvel que integra técnicas de aprendizado de máquina para atuar como ferramenta de triagem do risco de diabetes mellitus, associada a funcionalidades de educação em saúde. A proposta integra um modelo preditivo a uma aplicação móvel, buscando aproximar recursos de aprendizado de máquina do usuário final e possibilitar sua utilização em um contexto aplicado de triagem e educação em saúde.

A arquitetura cliente-servidor adotada, combinando Flutter no cliente e Python com FastAPI no servidor, permitiu separar as responsabilidades entre a aplicação móvel e o processamento do modelo preditivo. Entre os algoritmos avaliados, o modelo MLP apresentou o melhor desempenho geral, alcançando acurácia de 97,08%, precisão de 99,00% e F1-Score de 81,00% para a classe diabética. Apesar do desempenho observado, o *recall* de 68,00% evidencia a ocorrência de falsos negativos, reforçando a necessidade de interpretar os resultados com cautela, especialmente em uma aplicação voltada à triagem de risco. Dessa forma, a saída do aplicativo deve ser compreendida como uma estimativa computacional, não substituindo a avaliação e o diagnóstico realizados por profissionais de saúde.

A aplicação da técnica LIME permitiu analisar localmente os fatores que mais contribuíram para predições individuais do modelo MLP. Na instância analisada, as variáveis *HbA1c_level* e *blood_glucose_level* apresentaram as maiores contribuições positivas para a classificação, possibilitando uma análise mais transparente da decisão do classificador. A integração desse mecanismo de explicabilidade ao processo de análise contribui para uma compreensão mais detalhada das predições realizadas pelo modelo.

Como contribuição, o trabalho apresenta uma aplicação móvel que reúne, em uma mesma solução, uma abordagem de aprendizado de máquina para estimativa de risco, um mecanismo de explicabilidade baseado em LIME e recursos de educação em saúde. Como trabalhos futuros, pretende-se: (i) realizar testes de usabilidade com usuários de diferentes perfis para avaliar a aceitação e a experiência de uso da ferramenta; (ii) investigar modelos alternativos, como XGBoost, buscando melhorar a identificação da classe minoritária e o *recall*, considerando o equilíbrio entre as diferentes métricas de desempenho; e (iii) incorporar a visualização das explicações geradas pelo LIME diretamente na interface do aplicativo, tornando os resultados das predições mais compreensíveis para o usuário final.

Referências

[1] World Health Organization (WHO), “Noncommunicable diseases,” 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Acesso em: 30 mar. 2026.

[2] American Diabetes Association, “Diagnosis and classification of diabetes: Standards of Care in Diabetes—2025,” *Diabetes Care*, vol. 48, no. Suppl 1, pp. S27–S49, 2025.

[3] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 35, no. Suppl 1, pp. S64–S71, 2012.

[4] H. C. A. de Vasconcelos *et al.*, “Fatores de risco para diabetes mellitus tipo 2 entre adolescentes,” *Revista da Escola de Enfermagem da USP*, vol. 44, no. 4, pp. 881–887, 2010.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[7] N. M. Khalifa and M. A. Albadawy, “Artificial intelligence and machine learning applications in healthcare,” *Healthcare*, vol. 12, no. 4, 2024.

[8] X. Liu *et al.*, “Machine learning-based diabetes prediction using BRFSS data and explainable artificial intelligence,” *Scientific Reports*, vol. 14, 2024.

[9] A. Sharma *et al.*, “Explainable machine learning models for diabetes prediction and risk assessment,” *Healthcare Analytics*, vol. 5, p. 100300, 2024.

[10] D. F. Costa, A. K. N. Costa, F. R. Fernandes, N. C. de A. Lima, and G. B. C. e Sá, "Machine Learning na predição de Diabetes: uma abordagem explicável," in *Congresso Brasileiro de Inteligência Computacional*, 2025, doi: 10.21528/CBIC2025-1163557.

[11] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Diabetes prediction using SMOTE and SHAP-based explainable machine learning," *Applied Sciences*, vol. 15, no. 2, 2025.

[12] Kaggle, "Diabetes Prediction Dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. Acesso em: 08 set. 2026.

[13] F. M. Delpino, Â. K. Costa, L. F. Rodrigues, and B. P. Nunes, "ID204 Meta-análise para identificar a capacidade da Machine Learning na predição de diabetes: potencial e impactos da tecnologia na saúde," *Jornal de Assistência Farmacêutica e Farmacoeconomia*, vol. 9, no. s.1, p. 151, 2024.

Extended Resistance-Inertance-Compliance Model in the Characterization of Respiratory Mechanical Abnormalities in Idiopathic Scoliosis

1st Cíntia Moraes de Sá Sousa
Institute of Biology and Faculty of Engineering, Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0003-4777-222X

2nd Luís Eduardo Carelli
Spine surgery
National Institute of Traumatoortopedia
Rio de Janeiro, Brazil
0000-0003-2631-5492

3rd Alderico Girão Campos de Barros
Spine surgery
National Institute of Traumatoortopedia
Rio de Janeiro, Brazil
0000-0002-8337-7676

4th André Luiz de Campos Pessoa
Pedro Ernesto University Hospital
State University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0002-3116-9146

5th Pedro Lopes de Melo
Institute of Biology and Faculty of Engineering, Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0002-2081-6619

***Abstract* — Idiopathic scoliosis is a three-dimensional deformity of the spine that can impair respiratory mechanics due to distortion of the thoracic cage. Respiratory oscillometry has emerged as a noninvasive technique capable of detecting subtle alterations in respiratory system mechanics, which can be further enhanced through electrical modeling approaches. However, no previous studies have applied these methods to the assessment of scoliosis. Therefore, the aim of this study was to analyze respiratory mechanical alterations in patients with idiopathic scoliosis using the Extended Resistance–Inertance–Compliance (eRIC) model applied to respiratory oscillometry. Thirty patients diagnosed with idiopathic scoliosis and thirty healthy volunteers comprising the control group were evaluated. All participants underwent spirometry, body plethysmography, and respiratory oscillometry. Oscillometric parameters were analyzed using the eRIC model, which estimates central (R), peripheral (Rp), and total (Rt=R+Rp) airway resistance, as well as respiratory inertance (I) and dynamic compliance (C). Correlation analyses were also performed between eRIC-derived parameters and pulmonary function indices. The scoliosis group presented a mean Cobb angle of 51.9 ± 20.42°. Compared with controls, patients with scoliosis exhibited significantly increased Rp ($p<0.05$) and Rt ($p<0.01$), whereas I and C were significantly reduced ($p<0.001$). Correlation analysis demonstrated significant negative associations between resistive parameters and spirometric indices, with Rt showing the strongest correlations with $FEF_{25–75}$ ($r=−0.534$; $p < 0.0001$), FEV_1/FVC ($r=−0.489$; $p<0.0001$), and FEV_1 ($r=−0.473$; $p<0.0001$). In contrast, C exhibited the strongest positive correlations with pulmonary function parameters, particularly FEV_1 ($r=0.686$; $p<0.0001$), $FEF_{25–75}$ ($r=0.675$; $p<0.0001$), and FVC ($r=0.642$; $p<0.0001$). Among the plethysmographic variables, only total lung capacity was significantly correlated with C ($r=0.389$; $p<0.05$). These findings indicate increased airway obstruction and reduced respiratory system compliance, reflecting the mechanical constraints imposed by thoracic cage deformity and the consequent impairment of respiratory function. In conclusion, respiratory oscillometry combined with eRIC modeling was able to characterize relevant respiratory mechanical abnormalities associated with idiopathic scoliosis. The significant associations observed between eRIC-derived parameters and conventional pulmonary function indices support the physiological relevance of this model and its potential utility as a complementary tool for the functional assessment of patients with spinal deformities, providing information additional to that obtained via spirometry and oscillometry.**

Keywords— Respiratory oscillometry, idiopathic scoliosis, pulmonary function, mathematical models, spirometry.

I. Introduction

Scoliosis is a condition characterized by a three-dimensional postural deformity of the vertebral column that may lead to alterations in respiratory mechanics, resulting in reduced pulmonary capacity and, in more severe curvatures, contributing to additional systemic dysfunctions [1].

Diagnosis is established through radiographic examination, in which the Cobb angle [2] is measured; scoliosis is defined by a curvature greater than 10 degrees accompanied by vertebral rotation. The condition affects approximately 2–4% of adolescents between 10 and 18 years of age, with a higher prevalence in females [3].

The postural deformity associated with idiopathic scoliosis may lead to respiratory impairment; therefore, pulmonary function tests are routinely employed for the evaluation and longitudinal monitoring of these patients, who often do not present with overt respiratory symptoms [1].

Respiratory oscillometry is a noninvasive and easily applicable technique that enables the assessment of respiratory mechanics through parameters capable of detecting subtle physiological abnormalities that may not be evident in conventional pulmonary function tests. Its principal advantage lies in the superimposition of oscillatory signals during spontaneous tidal breathing, thereby eliminating the need for maximal respiratory maneuvers and reducing patient effort [4].

To improve the interpretation of oscillometric data, modeling approaches based on electrical analogs have been developed to represent the mechanical properties of the respiratory system. Among these approaches, the Extended Resistance-Inertance-Compliance (eRIC) model has been widely used to provide physiologically meaningful estimates of central and peripheral airway resistance, inertance, and respiratory system compliance [5]. Despite its increasing application in respiratory research [6-8], the use of the eRIC model in patients with idiopathic scoliosis has not yet been investigated.

Therefore, the aim of the present study was to evaluate the contribution of the eRIC model to the analysis of respiratory mechanical alterations in patients with idiopathic scoliosis.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

II. Materials and Methods

A. Volunteers

Telephone contact was established with 106 patients diagnosed with thoracic idiopathic scoliosis, of whom 30 were ultimately included in the study. A total of 76 individuals were excluded for the following reasons: three were unable to complete the examinations, six declined participation, eight did not attend the scheduled evaluation, and 59 could not be reached by telephone.

The final sample comprised 60 volunteers divided into two groups. The control group (CG) included 30 healthy individuals with no history of smoking, no prior pulmonary and/or cardiac diseases, no COVID-19 sequelae or post-COVID-19 syndrome, no postural abnormalities, and with spirometry and respiratory oscillometry results within normal ranges. The group of patients consisted of 30 individuals aged 11 to 50 years, diagnosed with thoracic idiopathic scoliosis [2], with no history of smoking, no prior pulmonary and/or cardiac diseases, and no COVID-19 sequelae or post-COVID-19 syndrome. After the participants were selected, they were referred to the Biomedical Instrumentation Laboratory at the State University of Rio de Janeiro.

B. Ethics

All participants provided written informed consent through a Free and Informed Consent Form approved by the Research Ethics Committees of Hospital Universitário Pedro Ernesto and the National Institute of Traumatology and Orthopedics under number 62654422.3.0000.5259.

For volunteers under 18 years of age, an Assent Form was signed by the participant, and the Informed Consent Form was signed by the legal guardian.

C. Spirometric and plethysmographic tests

Pulmonary function testing was carried out by experienced technicians using calibrated instrumentation. Spirometric data were acquired with a pneumotachograph system (nSpire Health, Inc., Longmont, CO, USA). Prior to testing, participants were carefully instructed to perform the respiratory maneuvers according to standardized procedures, ensuring compliance with national quality control criteria for acceptability and reproducibility. The parameters evaluated included forced vital capacity (FVC), forced expiratory volume in the first second (FEV_1), the FEV_1/FVC ratio, and mid-expiratory flow ($FEF_{25–75}$%). Assessment of lung volumes was performed by body plethysmography, following the methodological recommendations established by the American Thoracic Society (ATS)/ERS guidelines [11]. The variables analyzed comprised total lung capacity (TLC), vital capacity (VC), residual volume (RV), and the RV/TLC ratio, which was used as an indicator of air trapping [9, 10]. Plethysmographic variables were compared with predicted reference values rather than measured values obtained from the control group, which should be considered when interpreting these findings.

D. Respiratory oscillometry test

In these analyses, at least three recordings presenting a coherence function greater than 0.90 were required for the test to be considered acceptable. Following the initial assessment, participants were instructed to perform the maneuver in a seated position, breathing quietly through a mouthpiece while wearing a nasal clip for 16 seconds. Volunteers were also instructed to support the cheeks and chin with their palms to minimize signal loss in the upper airways, which could compromise measurement accuracy. Oscillometric measurements were conducted within a frequency range of 4 to 32 Hz, superimposed on the participants' tidal breathing. Respiratory impedance (Zrs) values were calculated using impedance analyzer system developed in our laboratory [11].

E. Electrical model

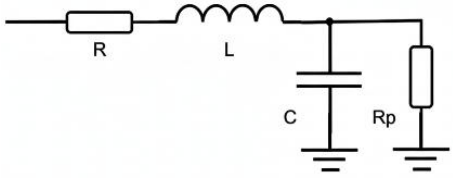
The Resistance–Inertance–Compliance (RIC) model represents the respiratory system using three fundamental mechanical components: resistance, inertance, and compliance. In the present study, the Extended Resistance–Inertance–Compliance (eRIC) model was employed (Figure 1). In this model, respiratory resistance is partitioned into central airway resistance (R) and peripheral airway resistance (Rp), whose sum corresponds to the total respiratory system resistance (Rt). This approach enables a more detailed assessment of respiratory mechanics by allowing the independent evaluation of the contributions of central and peripheral airway compartments. This characteristic is particularly relevant because several respiratory diseases initially affect the small airways, introducing a high potential to detect early mechanical abnormalities that may not be evident using conventional pulmonary function tests [12, 13].

Model parameters were estimated using the Levenberg–Marquardt algorithm, which identified the set of model coefficients that minimized the least-squares error between the measured and predicted data. In addition to the estimated parameters, the analysis provided the total error as an overall measure of model goodness of fit. This metric was calculated as the square root of the sum of the errors associated with the real and imaginary components of respiratory impedance.

Furthermore, the parameters derived from the eRIC model provide physiologically meaningful information that complements the findings obtained from spirometry and body plethysmography. While spirometry primarily evaluates airflow and lung volumes, and plethysmography quantifies static lung volumes and airway resistance, the eRIC model allows characterization of the resistive, inertial, and elastic components of the respiratory system. Consequently, it contributes to a more comprehensive understanding of the mechanical alterations associated with respiratory dysfunction [13].

Respiratory inertance (I) describes the inertial properties of the respiratory system, reflecting the mass of air within the airways and the respiratory tissues that must be accelerated during breathing. Respiratory system compliance (C) represents the elastic properties of the lungs and chest wall, incorporating the contributions of both the pulmonary parenchyma and thoracic cage structures [5, 7, 12, 13] .

Figure 1: Two-compartment integer-order model employed for the analysis of respiratory impedance.



Legend: Resistance (R), inductance (I), and capacitance (C) correspond to the analogs of respiratory resistance, inertance, and compliance, respectively, while Rp denotes peripheral airway resistance.

CAPES; CNPq; FAPERJ.

F. *Statistical analysis*

Statistical analyses and graphical representations were performed using Origin® 8.0 software (Microcal Software Inc., Northampton, MA, USA). Data are presented as mean ± standard deviation (SD) and summarized in tables and box plots. A p-value ≤ 0.05 was considered statistically significant in all analyses.

Data normality was assessed using the Shapiro–Wilk test. Comparisons between groups were performed using the independent-samples Student's t-test for normally distributed variables. When the assumption of normality was not met, the nonparametric Mann–Whitney U test was applied.

Associations between respiratory oscillometry parameters and pulmonary function variables were evaluated through correlation analyses. Pearson's correlation coefficient was used for parametric data, whereas Spearman's rank correlation coefficient was employed for nonparametric data. These analyses were conducted using GraphPad Prism 8 software (GraphPad Software, La Jolla, CA, USA).

The strength of the correlations was interpreted according to the following criteria: coefficients ranging from 0 to 0.25 (or 0 to −0.25) were considered negligible to weak; values between 0.25 and 0.50 (or −0.25 and −0.50) were considered moderate; coefficients between 0.50 and 0.75 (or −0.50 and −0.75) were classified as moderate to strong; and values greater than 0.75 (or less than −0.75) were considered very strong to excellent [14].

III. RESULTS

Table 1 presents the anthropometric and spirometric results. No significant differences were observed in age, body weight, height, or BMI when comparing the control group and the idiopathic scoliosis group.

In the spirometric analysis, the idiopathic scoliosis group demonstrated significantly lower values compared with the control group in FEV_1 (L), FEV_1 (%), FVC (L), FVC (%), FEV_1/FVC (L), FEV_1/FVC (%) e FEF_{25-75} (p<0.0001).

Table 1 - Anthropometric and spirometric measurements of the studied groups.

	Control (n= 30)	Scoliosis (n= 30)	p
Age (years)	22.93 ± 2.75	25.10 ± 9.68	ns
Weight (kg)	58.92 ± 7.90	55.87 ± 11.14	ns
Height (cm)	165.33 ± 7.68	162.63 ± 10.82	ns
BMI (kg/m2)	21.39 ± 2.19	21.19 ± 3.95	ns
Gender (F/M)	23/7	21/9	-
Cobb Angle	0	51.9 ± 20.42	<0.0001
FEV_1 (L)	3.57 ± 0.83	2.46 ± 0.80	<0.0001
FEV_1 (%)	102.34 ± 13.94	75.53 ± 23.14	<0.0001
FVC (L)	4.05 ± 0.93	3.04 ± 0.93	<0.0001
FVC (%)	101.34 ± 13.65	81.76 ± 24.96	<0.0001
FEV_1/FVC (L)	89.81 ± 5.48	80.72 ± 7.92	<0.0001
FEV_1/FVC (%)	101.59 ± 5.17	92.80 ± 9.73	<0.0001
FEF_{25-75}	4.28 ± 1.23	2.53 ± 1.00	<0.0001

Legend: BMI - Body Mass Index; FEV_1 - Forced Expiratory Volume in one second; FVC - Forced Vital Capacity; FEV_1/FVC - Forced Expiratory Volume in one second to Forced Vital Capacity ratio; FEF - Forced Expiratory Flow; ns - not significant;

Table 2 presents the plethysmographic results and predicted values for patients with idiopathic scoliosis. A significant reduction was observed in VC (L) and RV/TLC (p<0.0001, respectively), along with a significant increase in RV (L) (p<0.01). No significant differences were observed in TLC (p= ns).

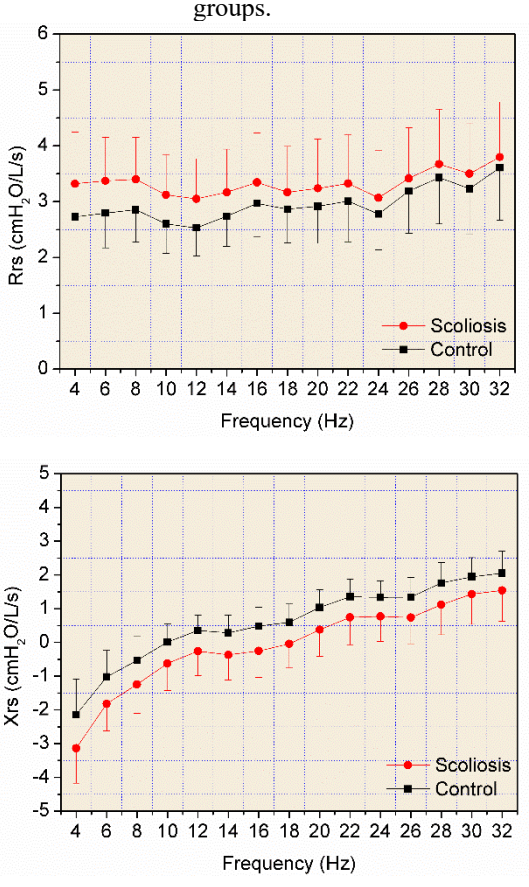
Table 2 – Plethysmography results and predicted values of the studied groups.

	Predicted (n=29)	Scoliosis (n=29)	p
TLC (L)	5.09 ± 1.05	4.62 ± 1.11	ns
VC (L)	3.83 ± 0.73	2.78 ± 0.87	<0.0001
RV (L)	1.32 ± 0.27	1.87 ± 0.80	<0.01
RV/ TLC	25.54 ± 2.50	40.49 ± 12.67	<0.0001

Legend: TLC - Total Lung Capacity; VC – Vital Capacity; RV - Residual Volume; RV/TLC - Ratio between Residual Volume and Total Lung Capacity; n = number of patients evaluated; ns - not significant

Figures 2A and 2B present the mean respiratory resistance (Rrs) and reactance (Xrs) curves as a function of frequency in healthy individuals and patients with idiopathic scoliosis, respectively. Significant differences (p<0.0001) were observed between all resistance curves (Figure A), while non-significant differences were observed in reactance curves (Figure 2B).

Figure 2: Mean respiratory resistance (A) and reactance (B) curves as a function of frequency of the control and Scoliosis groups.

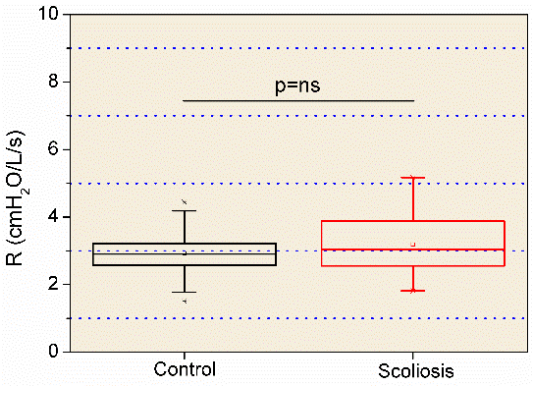


Legend: Rrs - Respiratory resistance curve; Xrs - Respiratory reactance curve; Hz - Hertz.

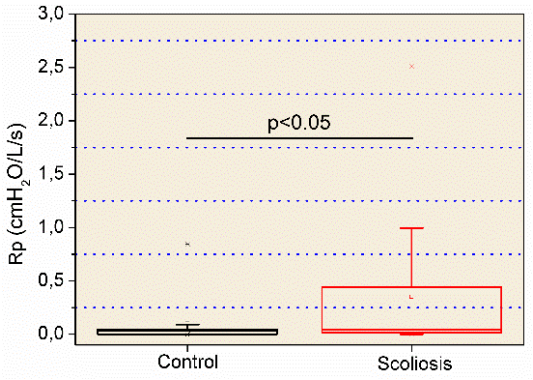
Figure 3 presents the eRIC model analysis. R values, shown in Figure 3A did not demonstrate significant differences compared with the control group (p= ns). In contrast, Rp and Rt, Figures 3B and 3C respectively, were significantly increased in the idiopathic scoliosis group ($p<0.05$).

Respiratory system inertance (Figure 3D), was significantly reduced ($p<0.05$), as was C (Figure 3E, $p<0.001$) when compared with the control group.

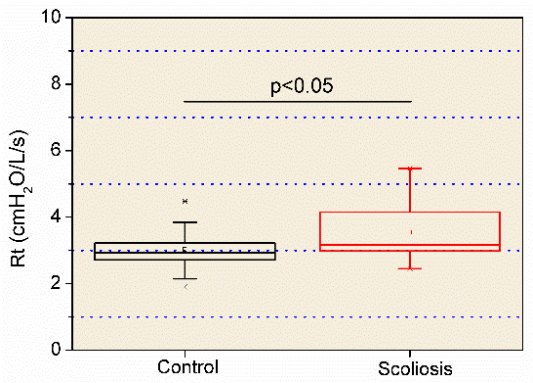
Figure 3: Respiratory Oscillometry results based on extended RIC model analysis



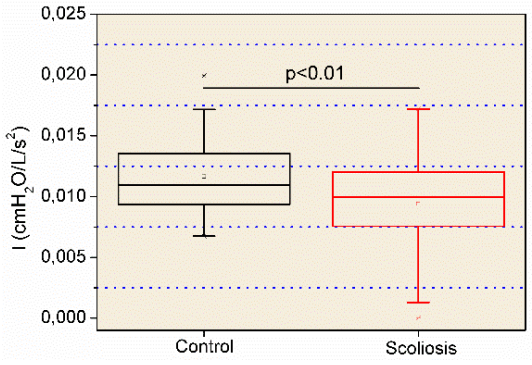
A



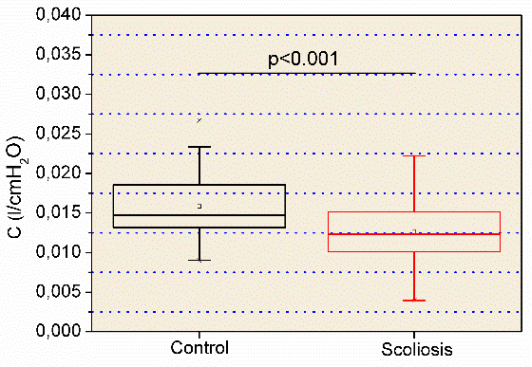
B



C



D



E

Legend: R – Central resistance; Rp – Peripheral resistance; Rt – Total resistance; I – Inertance; C – Dynamic compliance; ns - not significant;

Correlation analysis revealed significant associations between the eRIC model parameters and spirometric indices (Table 3). The resistive parameters exhibited negative correlations, indicating that increased respiratory resistance was associated with reduced pulmonary function. Among these parameters, total respiratory resistance (Rt) showed the moderate to strong associations, particularly with $FEF_{25–75}$ ($r=-0.534$, $p<0.0001$), FEV_1/FVC ($r=-0.489$, $p<0.0001$), and FEV_1 ($r=-0.473$, $p<0.0001$).

In contrast, dynamic compliance (C) demonstrated moderate-to-strong positive correlations with several spirometric parameters, with the strongest relationships observed for FEV_1 ($r=0.686$, $p<0.0001$), $FEF_{25–75}$ ($r=0.675$, $p<0.0001$), and FVC ($r=0.642$, $p<0.0001$). Respiratory inertance (I) did not show significant correlations with any of the spirometric variables evaluated.

Regarding body plethysmography (Table 4), C was the only parameter that exhibited a significant moderate correlation, showing a positive association with TLC ($r=0.389$, $p<0.05$).

Table 3 - Correlation coefficient (r) and correlation significance level (p) of the oscillometry parameters in relation to the spirometry. Significant correlations are highlighted in bold.

		FVC %	FVC	FEV_1 %	FEV_1	FEV_1/FVC %	FEV_1/FVC	FEF_{25-75}	FEF_{25-75}/FVC
R	r	-0.061	**-0.348**	-0.122	**-0.384**	**-0.265**	-0.234	**-0.432**	-0.174
	p	0.644	**<0.01**	0.352	**<0.01**	**<0.005**	0.072	**<0.001**	0.184
Rp	r	-0.145	-0.005	**-0.268**	-0.063	**-0.346**	**-0.293**	-0.138	-0.188
	p	0.270	0.967	**<0.05**	0.631	**<0.01**	**<0.05**	0.292	0.151
Rt	r	-0.127	**-0.375**	**-0.283**	**-0.473**	**-0.489**	**-0.473**	**-0.534**	**-0.311**
	p	0.332	**<0.01**	**<0.05**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.05**
I	r	0.193	0.074	0.239	0.106	0.121	0.127	0.120	0.158
	p	0.140	0.576	0.065	0.421	0.357	0.334	0.362	0.228
C	r	**0.278**	**0.642**	**0.415**	**0.686**	**0.378**	**0.371**	**0.675**	0.179
	p	**<0.05**	**<0.0001**	**<0.001**	**<0.0001**	**<0.01**	**<0.01**	**<0.0001**	0.170

Legend: FEV_1 - Forced Expiratory Volume in one second; FVC - Forced Vital Capacity; FEV_1/FVC - Forced Expiratory Volume in one second to Forced Vital Capacity ratio; FEF - Forced Expiratory Flow; R – Central resistance; Rp – Peripheral resistance; Rt – Total resistance; I – Inertance; C – Dynamic compliance

Table 4 - Correlation coefficient (r) and correlation significance level (p) of the oscillometry parameters in relation to the plethysmography. Significant correlations are highlighted in bold.

		TLC	VC	RV	RV/TLC
R	r	-0.198	-0.035	-0.327	-0.138
	p	0.304	0.857	0.084	0.474
Rp	r	0.149	0.135	0.251	0.105
	p	0.440	0.484	0.188	0.586
Rt	r	-0.076	0.005	-0.233	-0.084
	p	0.696	0.980	0.223	0.666
I	r	0.159	0.279	-0.032	-0.278
	p	0.409	0.143	0.868	0.145
C	r	**0.389**	0.325	0.325	-0.072
	p	**<0.05**	0.086	0.086	0.711

Legend: TLC - Total Lung Capacity; VC – Vital Capacity; RV - Residual Volume; RV/TLC - Ratio between Residual Volume and Total Lung Capacity; R – Central resistance; Rp – Peripheral resistance; Rt – Total resistance; I – Inertance; C – Dynamic compliance

IV. Discussion

The anthropometric results of the scoliosis group showed no significant differences compared with the control group, indicating that the groups were matched and homogeneous (Table 1). As described in the literature, idiopathic scoliosis is more prevalent in females, which is consistent with the findings of the present study [15].

Regarding spirometry, a significant reduction was observed in all parameters ($p<0.0001$), characterizing the pulmonary restriction found in this population and already well described in the literature. Additionally, signs of airway obstruction were also observed in the studied population [15, 16]. These alterations may be explained by the thoracic cage deformity observed in these individuals, which reduces lung volumes due to abnormal rib angulation and decreased diaphragmatic mobility. Furthermore, depending on the severity of scoliosis, airway obstruction may also be observed, often associated with hypokyphosis. In such cases, the spinal column protrudes into the thoracic cavity, altering the anatomical position of the bronchi and thereby contributing to airway obstruction [17, 18].

The results obtained from the plethysmography examination are presented in Table 2. One volunteer was unable to complete the examination due to his inability to perform the forced maneuvers necessary to obtain the measurements. A significant reduction in VC ($p<0.0001$) in comparison with the predicted values was observed. This finding may be attributed to the reduction in lung volume as a consequence of trunk distortion in these patients [19] A significant increase in RV and RV/TLC ($p<0.001$ and $p<0.0001$, respectively) was also observed, which may be explained by air trapping caused by thoracic distortion [20]. TLC did not show statistically significant changes, which may be partially explained by the fact that reductions in TLC are typically observed in patients with Cobb angles greater than 80 degrees or in the presence of air trapping [20, 21].

The analysis of the Rrs curve (Figure 2A) demonstrated significantly higher values ($p<0.0001$) in patients with idiopathic scoliosis compared with healthy controls across the entire frequency range evaluated (4–32 Hz). The increase in Rrs observed in the scoliosis group suggests greater opposition to airflow, likely resulting from structural alterations of the thoracic cage secondary to spinal deformity [22].

Figure 2B shows the respiratory reactance (Xrs) curves. Although no statistically significant differences were observed, patients with scoliosis exhibited visually more negative Xrs values, particularly within the low-frequency range (4–16 Hz). Similar findings have been reported by Van Noord et al. [22], suggesting a tendency toward reduced

respiratory system compliance and increased mechanical loading of the respiratory system [23].

Taken together, these findings are consistent with the restrictive ventilatory pattern commonly described in idiopathic scoliosis, which is characterized by impaired thoracic expansion and increased ventilation heterogeneity. The observed mechanical abnormalities likely reflect the combined effects of thoracic cage distortion on lung volumes, chest wall mechanics, and regional ventilation distribution, thereby contributing to the respiratory dysfunction observed in this population.

The results using the eRIC model are presented in Figure 3. Although clear visual increases may be observed, no statistically significant differences were observed in R (Figure 3A).

Regarding Rp (Figure 3B), the significant increase in peripheral airway resistance suggests that the thoracic deformity associated with idiopathic scoliosis induces mechanical changes capable of compromising small-airway function. Reduced chest wall mobility, combined with heterogeneous ventilation distribution, may contribute to the functional narrowing of distal airways and increased airflow resistance [24].

These mechanisms may foster the development of ventilatory heterogeneity, a phenomenon consistent with the more negative reactance values observed at lower frequencies (Figure 2B). Taken together, these findings indicate that the respiratory consequences of idiopathic scoliosis extend beyond the reduction in lung volumes typically identified by spirometry and body plethysmography, encompassing complex alterations in peripheral airway function and respiratory system mechanics.

In line with these observations, Rt (Figure 3C), which represents the sum of central and peripheral resistances showed a significant increase. This finding suggests that the alterations observed in idiopathic scoliosis include airway obstruction, predominantly affecting the more peripheral airways. This may be attributed to trunk distortion, leading to deformities of the chest wall and lung parenchyma [25].

Studies applying the eRIC model in patients with chronic obstructive pulmonary disease (COPD) have reported Rt values comparable to those observed in the present study, typically ranging from 3.5 to 4.5 $cmH_2O/l/s$ in individuals with mild to moderate disease. This finding suggests that the airway obstruction associated with spinal deformity in scoliosis may reach a magnitude similar to that observed in obstructive pulmonary disorders [26, 27].

Inertance is related to the movement of the mass of gas within the airways. Theoretically, a reduction in lung volume leads to a decrease in the mass of gas that must be displaced during breathing, reducing I. In line with this proposition, Table 2 shows a reduction in lung volumes, while Figure 3D shows that I reduced in scoliosis. This apparent reduction in I suggested the presence of airway obstruction, particularly in distal regions, which limits the propagation of the oscillatory signal and creates "choke points," thereby reducing the measured inertial component of the respiratory system [28]. In addition, the plethysmographic findings of reduced VC and increased RV/TLC suggest air trapping and impaired lung-emptying, conditions that effectively reduce the ventilated lung volume participating in oscillatory flow. Since inertance is influenced by the mass of air set into motion within the airways, a decrease in effective ventilated volume and airflow distribution may further contribute to the observed reduction in I. This relationship highlights the interplay between structural restriction, airway closure, and altered respiratory mechanics in scoliosis, reinforcing the complementary nature of oscillometry and plethysmography in detecting these changes.

The dynamic compliance of the respiratory system was also reduced (Figure 3E). Alterations in thoracic cage dynamics lead to a reduction in dynamic pulmonary compliance in patients with spinal deformity [29]. The presence of "choke points" is also related to a reduction of dynamic compliance in a phenomena similar to what happens in dynamic inertance [28].

The relationship between oscillometric parameters and disease severity has been examined in three previous studies [24, 30, 31]. In these studies, patients were stratified according to Cobb angle severity to evaluate respiratory mechanics across different stages of idiopathic scoliosis progression.

The negative correlations observed between the resistive parameters (R, Rp, and Rt) and spirometric indices suggest that increased respiratory resistance is associated with deterioration in pulmonary function. From a physiological perspective, this relationship is expected, as elevated airway resistance increases opposition to expiratory airflow, leading to reductions in FEV_1, the FEV_1/FVC ratio, and mid-expiratory flow rates, which are recognized markers of airflow limitation [32].

Among the resistive parameters, Rt demonstrated the strongest associations, indicating that the combined assessment of central and peripheral airway resistance may more accurately reflect the overall mechanical impairment of the respiratory system. This observation is consistent with the structural and functional alterations associated with idiopathic scoliosis, including postural deformities and thoracic cage distortion, which may compromise respiratory mechanics and increase airflow resistance [33].

Given that $FEF_{25–75}$ is commonly considered a surrogate marker of small airway function, these findings suggest that the eRIC model is sensitive to mechanical alterations occurring within the peripheral regions of the respiratory system. This observation reinforces the potential value of respiratory oscillometry as a complementary technique to spirometry, providing additional insights into pulmonary mechanical dysfunction that may not be fully captured by conventional pulmonary function testing [34].

Dynamic compliance exhibited the strongest correlations among all eRIC model parameters, particularly with FVC, FEV_1, and FEF_{25-75}. These findings indicate that reduced respiratory system distensibility is associated with impaired pulmonary function, affecting both lung volumes and airflow. From a physiological perspective, compliance reflects the ability of the respiratory system to expand in response to a given pressure change and is determined by the elastic properties of both the pulmonary parenchyma and the thoracic cage. Therefore, the strong associations observed between C and the spirometric indices suggest that this eRIC parameter may reflect global alterations in respiratory mechanics resulting from the postural deformities characteristic of idiopathic scoliosis [34, 35].

The body plethysmography results further support this interpretation. Among all evaluated parameters, dynamic compliance was the only variable significantly correlated with total lung capacity. Given that TLC represents the maximum volume that can be accommodated by the respiratory system, the positive association observed suggests that individuals with greater compliance exhibit a greater capacity for pulmonary expansion. Conversely, the absence of significant correlations between the resistive parameters and static lung volumes indicates that airway resistance primarily reflects airflow-related abnormalities, whereas lung volumes are predominantly determined by the elastic properties of the respiratory system [32].

Another noteworthy finding was the absence of significant correlations involving respiratory inertance (I). Although inertance reflects the mass of air and tissues that must be accelerated during breathing, its contribution to the functional abnormalities observed in the present sample appears to be limited. Similar findings have been reported in previous respiratory modeling studies, in which resistive and elastic parameters generally demonstrated greater clinical relevance than inertial components. A possible explanation is that patients with idiopathic scoliosis typically exhibit reduced FVC values, resulting in lower displacement of gas volume within the lungs and consequently diminishing the influence of inertial properties on respiratory mechanics. These findings are consistent with those reported by Oliveira et al. [32].

Some limitations of this study should be considered. The cross-sectional design precludes establishing causal relationships between spinal deformity and alterations in respiratory mechanics. Furthermore, the lack of analyses stratified by scoliosis severity limits the assessment of the Cobb angle's influence on the eRIC model parameters. Studies involving larger cohorts, independent validation, and longitudinal follow-up are needed to confirm the diagnostic and prognostic performance of these parameters, as well as to evaluate their ability to monitor disease progression and respiratory changes before and after corrective surgery. Comparisons with more advanced respiratory models will also help define the role of eRIC in functional assessment and clinical follow-up. Finally, the results are limited to idiopathic scoliosis and should not be extrapolated to other etiologies.

V. Conclusion

The present study demonstrated that patients with idiopathic scoliosis exhibit significant alterations in respiratory mechanics, characterized by reduced lung volumes, ventilatory restriction, and signs of air trapping. These abnormalities are likely related to thoracic cage deformity and its effects on lung expansion and respiratory muscle mechanics.

The application of the eRIC model provided a comprehensive characterization of respiratory mechanical abnormalities in idiopathic scoliosis. Patients exhibited increased Rp and Rt, together with reduced I and C. Among the eRIC-derived parameters, C showed the strongest associations with spirometric and plethysmographic indices, while Rt also demonstrated consistent relationships with pulmonary function, supporting the potential of eRIC modeling to complement the assessment of respiratory mechanics in this population.

However, the clinical applicability of these findings remains to be established. External validation, analyses stratified by scoliosis severity, investigation of associations with Cobb angle, assessment of measurement repeatability, appropriate control for multiple comparisons, and longitudinal studies are required to confirm the robustness, diagnostic value, and prognostic utility of eRIC-derived parameters in idiopathic scoliosis.

Acknowledgment

The authors would like to thank CNPq, FAPERJ, and CAPES for their financial support. Special thanks to the partnership with the National Institute of Traumatology and Orthopedics (INTO) and all the volunteers.

Conflict of Interest

The authors report no conflict of interest.

References

[1] N. A. Kerna *et al.*, "The pathophysiology of scoliosis across the spectrum of human physiological systems," *J European Journal of MedicalHealth Research,* vol. 2, no. 2, pp. 69-81, 2024.

[2] S. Negrini *et al.*, "2016 SOSORT guidelines: orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth," vol. 13, no. 1, p. 3, 2018.

[3] J. P. Horne, R. Flannery, and S. Usman, "Adolescent idiopathic scoliosis: diagnosis and management," *J American family physician,* vol. 89, no. 3, pp. 193-198, 2014.

[4] I. Loukou, M. Moustaki, A. Deligianni, O. Sardeli, and K. Douros, "Forced oscillation technique for monitoring the respiratory status of children with cystic fibrosis: a systematic review," *J Children,* vol. 8, no. 10, p. 857, 2021.

[5] F. A. d. A. Tuza, P. M. de Sá, H. A. Castro, A. J. Lopes, and P. L. de Melo, "Combined forced oscillation and fractional-order modeling in patients with work-related asthma: a case–control study analyzing respiratory biomechanics and diagnostic accuracy," *J Biomedical engineering online,* vol. 19, no. 1, p. 93, 2020.

[6] P. Bokov, D. Jallouli‐Masmoudi, F. Amat, V. Houdouin, and C. Delclaux, "Small airway dysfunction is an independent dimension of wheezing disease in preschool children," *J Pediatric Allergy Immunology,* vol. 33, no. 1, p. e13647, 2022.

[7] P. Bokov *et al.*, "Beneficial short-term effect of autogenic drainage on peripheral resistance in childhood cystic fibrosis disease," *J BMC Pulmonary Medicine,* vol. 22, no. 1, p. 241, 2022.

[8] P. Bokov *et al.*, "Cross-sectional phenotyping of small airway dysfunction in preschool asthma using the impulse oscillometry system," *Journal of Asthma,* vol. 58, no. 5, pp. 573-585, 2021.

[9] C. Pereira and M. Moreira, "Pletismografia-resistência das vias aéreas," *J Pneumol,* vol. 28, no. Supl 3, pp. S139-S50, 2002.

[10] M. R. Miller *et al.*, "Standardisation of spirometry," *J European respiratory journal,* vol. 26, no. 2, pp. 319-338, 2005.

[11] G. G. King *et al.*, "Technical standards for respiratory oscillometry," *European Respiratory Journal,* vol. 55, no. 2, 2020.

[12] A. C. Faria, J. Veiga, A. J. Lopes, and P. L. Melo, "Forced oscillation, integer and fractional-order modeling in asthma," *J Computer methods programs in biomedicine,* vol. 128, pp. 12-26, 2016.

[13] P. M. de Sá, H. A. Castro, A. J. Lopes, and P. L. d. Melo, "Early diagnosis of respiratory abnormalities in asbestos-exposed workers by the forced oscillation technique," *J PloS one,* vol. 11, no. 9, p. e0161981, 2016.

[14] B. Dawson and R. G. Trapp, "Basic & clinical biostatistics," in *Basic & clinical biostatistics*, 2004, pp. 438-438.

[15] V. Q. Santofimio, A. Clement, D. P. O'Regan, J. S. Ware, and K. A. McGurk, "Identification of an increased lifetime risk of major adverse cardiovascular events in UK Biobank participants with scoliosis," *J Open heart,* vol. 10, no. 1, 2023.

[16] J. Johari *et al.*, "Relationship between pulmonary function and degree of spinal deformity, location of apical vertebrae and age among adolescent idiopathic scoliosis patients," *J Singapore medical journal,* vol. 57, no. 1, p. 33, 2016.

[17] H. Zhang and Y. Hai, "Surgical correction of severe scoliosis leads to changes in central airway resistance evaluated with CT-based 3D reconstruction and impulse oscillometry," *JBJS,* p. 10.2106, 2021.

[18] C. E. Johnston *et al.*, "Correlation of preoperative deformity magnitude and pulmonary function tests in adolescent idiopathic scoliosis," *J Spine,* vol. 36, no. 14, pp. 1096-1102, 2011.

[19] R. Smyth, K. Chapman, T. Wright, J. Crawford, and A. Rebuck, "Pulmonary function in adolescents with mild idiopathic scoliosis," *J Thorax,* vol. 39, no. 12, pp. 901-904, 1984.

[20] T. Tsiligiannis and T. Grivas, "Pulmonary function in children with idiopathic scoliosis," *J Scoliosis,* vol. 7, no. 1, p. 7, 2012.

[21] G. L. McPhail *et al.*, "Obstructive lung disease in children with idiopathic scoliosis," *J The Journal of pediatrics,* vol. 166, no. 4, pp. 1018-1021, 2015.

[22] J. Van Noord, M. Cauberghs, K. Van de Woestijne, and M. Demedts, "Total respiratory resistance and reactance in ankylosing spondylitis and kyphoscoliosis," *J European Respiratory Journal,* vol. 4, no. 8, pp. 945-951, 1991.

[23] A. C. Faria, A. J. Lopes, J. M. Jansen, and P. L. Melo, "Assessment of respiratory mechanics in patients with sarcoidosis using forced oscillation: correlations with spirometric and volumetric measurements and diagnostic accuracy," *J Respiration,* vol. 78, no. 1, pp. 93-104, 2009.

[24] C. Sousa, A. Pessoa, L. Carelli, C. Ribeiro, A. Lopes, and P. Melo, "Respiratory oscillometry and functional analyses in patients with idiopathic scoliosis," *J Brazilian Journal of Medical Biological Research,* vol. 56, p. e12898, 2023.

[25] B. T. Caldas, F. C. V. Ribeiro, J. S. Pereira, W. C. Souza, A. J. Lopes, and P. L. de Melo, "Oscillometry of the respiratory system in Parkinson's disease: physiological changes and diagnostic use," *J BMC Pulmonary Medicine,* vol. 23, no. 1, p. 406, 2023.

[26] C. O. Ribeiro, A. C. D. Faria, A. J. Lopes, and P. L. Melo, "Forced oscillation technique for early detection of the effects of smoking and COPD: contribution of fractional-order modeling," (in English), *International Journal of Chronic Obstructive Pulmonary Disease,* vol. 13, pp. 3281-3295, 2018.

[27] C. O. Ribeiro, A. J. Lopes, and P. L. de Melo, "Oscillation mechanics, integer and fractional respiratory modeling in COPD: effect of obstruction severity," *International Journal of Chronic Obstructive Pulmonary Disease,* vol. 15, p. 3273, 2020.

[28] A. C. Faria, A. R. S. Carvalho, A. R. M. Guimarães, A. J. Lopes, and P. L. Melo, "Association of respiratory integer and fractional-order models with structural abnormalities in silicosis," *J Computer MethodsPrograms in Biomedicine,* vol. 172, pp. 53-63, 2019.

[29] G. J. Redding, "Early onset scoliosis: a pulmonary perspective," *J Spine Deformity,* vol. 2, no. 6, pp. 425-429, 2014.

[30] C. M. d. S. Sousa, L. E. Carelli, A. L. Pessoa, A. J. Lopes, and P. Melo, "Forced oscillations and functional analysis in patients with idiopathic scoliosis," in *Brazilian Congress on Biomedical Engineering*, 2020, pp. 277-283: Springer.

[31] C. Sousa, C. Ribeiro, L. Carelli, A. Pessoa, and P. Melo, "Respiratory Oscillometry in Scoliosis: Changes in Resistive and Reactive Properties with Increasing Deformity Degree," in *Brazilian Congress on Biomedical Engineering*, 2024, pp. 283-291: Springer.

[32] B. F. Oliveira, C. O. Ribeiro, C. M. de Sá Sousa, M. C. Lopes, A. J. Lopes, and P. L. de Melo, "Respiratory abnormalities in sarcoidosis: physiopathology and early diagnosis using oscillometry combined with respiratory modeling," *J BMC Pulmonary Medicine,* vol. 25, no. 1, p. 68, 2025.

[33] C. G. T. Ledonio, B. E. Rosenstein, C. E. Johnston, W. E. Regelmann, D. J. Nuckley, and D. W. Polly Jr, "Pulmonary function tests correlated with thoracic volumes in adolescent idiopathic scoliosis," *J Journal of Orthopaedic Research,* vol. 35, no. 1, pp. 175-182, 2017.

[34] T. Mou, Y. Wang, Y. Fu, Y. Wang, and G. Li, "Analysis of the correlations and inconsistencies between spirometry and impulse oscillometry in the diagnosis of small-airway dysfunction," *J BMC Pulmonary Medicine,* vol. 24, no. 1, p. 619, 2024.

[35] P. Sa, A. Faria, A. Ferreira, A. Lopes, J. Jansen, and P. Melo, "Validation of the Forced Oscillation Technique in the diagnostic of respiratory changes in patients with silicosis," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 398-401: IEEE.

Artificial Intelligence applied to Postural and Functional Ergonomic Analysis of office workers

1st Jalusa Andréia Storch-Díaz
Physiotherapy
Federal University of Goiás
Goiania, Brazil
jastorch@ufg.br

2nd Aldo André Díaz-Salazar
Institute of Informatics
Federal University of Goiás
Goiania, Brazil
aldo.diaz@ufg.br

3rd Ellen Santos Rodrigues
Physiotherapy
Federal University of Goiás
Goiania, Brazil
ellen_rodrigues@discente.ufg.br

4th Débora Eliane Soares de Souza Batista
Computer Engineering
Federal University of Goiás
Goiania, Brazil
deboraeliane@discente.ufg.br

5th Fernanda de Castro Fernandes
Computer Engineering
Federal University of Goiás
Goiania, Brazil
castro.fernanda@discente.ufg.br

***Abstract*—The longevity of healthy workers in corporate environments is one of the goals of ergonomics programs. The aim of this study was to analyze the posture of office workers using Computer Vision and Artificial Intelligence methods in order to identify postural deviations and ergonomic risks and propose ergonomic recommendations. This was an observational, descriptive, quantitative study conducted with 22 office workers in Goiania, Brazil. The instruments used included a medical history questionnaire, as well as the Rapid Upper Limbs Assessment (RULA) and Rapid Entire Body Assessment (REBA) ergonomic postural evaluation protocols. The material used for data collection was an RGB camera mounted on a tripod at a fixed viewpoint. MediaPipe was used to detect keypoints on the human body. The results showed that 40.9% of the workers reported functional limitations in their office work performance. Regarding work-related complaints or conditions, 13.3% reported pain and tension in the lumbar region, 13.3% in the cervical region, and 13.3% reported knee pain. Segmental analysis of the human body revealed that the cervical region was the primary target for physical therapy and ergonomic interventions among office workers in the evaluated dataset, with an overall median RULA score of 3.65 and REBA score of 4.93. We found that the workers' health conditions compromised the proper performance of their work tasks. The use of Computer Vision and AI proved feasible for estimating postural changes in office work.**

***Index Terms*—Physical Therapy, Ergonomics, Computer Vision, Artificial Intelligence.**

I. Introduction

The longevity of workers in office environments is determined by factors that directly or indirectly influence health and quality of life at work. In this context, Occupational Physical Therapy, with an emphasis on Ergonomics, aims to analyze the relationship between the worker and their work, seeking to adapt work conditions to the worker's biopsychosocial characteristics, thereby minimizing occupational hazards and risks within organizations [1].

Office work requires employees to frequently use computers and cell phones, in addition to spending long periods of time in a seated position. Equipment and furniture are not always suitable or adjusted to individual anthropometric measurements. Consequently, poor ergonomic conditions can lead to the development of Repetitive Strain Injuries (RSIs) and/or Work-Related Disorders (WRDs), resulting in musculoskeletal pain, physical and mental fatigue, exhaustion, and stress among workers, thereby increasing workplace accident statistics [2].

Ergonomics has focused on the role of the human workforce in the face of advances in globalization, the Internet of Things, and Industry 4.0 [1]. However, physical therapy-based workplace assessments and ergonomic evaluations have traditionally relied on standardized observational methods, including the Rapid Upper Limb Assessment (RULA) [3] and the Rapid Entire Body Assessment (REBA) [4]. In these protocols, the physical therapist uses observation, photos, or videos, as well as tools such as goniometers and software, to assess posture, joint angles, and biomechanical variables, which are then converted into risk scores [1]. RULA and REBA are methods with satisfactory reliability in measuring joint angles [5]. However, [6] note that the reliance on observation and human judgment in rapid visual estimation methods can lead to subjective responses that hinder the scalability of assessments.

Disruptive technologies related to the Industry 4.0 paradigm have enabled interesting applications in the field of Occupational Physical Therapy and Ergonomics. More specifically, the impact of automated data collection and analysis is shaping a new set of applications for health assessment and intervention processes [7]. From this perspective, the present study proposes an indirect and automated functional and ergonomic assessment based on Computer Vision (CV) techniques and Large Language Models (LLMs) derived from Generative Artificial Intelligence (GenAI). Human pose estimators, such as MediaPipe, are capable, for instance, of extracting joint coordinates of the upper and lower limbs from images or videos, enabling the automation of ergonomic protocols [8].

Using CV algorithms, it is possible to identify and track

the positions of key points on the human body—particularly at joints and on limbs—which are useful for analyzing the body postures and movements used by office workers, thereby creating a digital skeleton that represents the pose of the person being assessed [9]. There are many pose estimation models, among them MediaPipe, which is considered a lightweight, efficient tool compatible with mobile devices such as cell phones, making it ideal for real-world applications such as offices.

The literature still has gaps regarding the use of Computer Vision in ergonomic assessments of office workers. The authors in [10] proposed a method for recognizing the body postures of construction workers, using 3D Computer Vision techniques based on a single 2D camera installed at strategic points on a construction site. In this model, a technique for recognizing improper postures was developed using computer vision techniques based on a single standard RGB (Red-Green-Blue) camera. Few studies simultaneously address the clinical accuracy required by the RULA and REBA protocols in ergonomic assessments, since data collection via video does not respect the privacy of the worker being assessed. Furthermore, few devices are capable of automatically translating angular measurements into terms that are understandable to end users, such as specialized occupational physical therapists and ergonomists.

Considering this, the goal of this study was to analyze the posture of office workers using Computer Vision, MediaPipe, and Artificial Intelligence, in order to identify postural deviations and ergonomic risks and propose ergonomic recommendations. The study also aimed to survey the musculoskeletal pain and complaints of office workers.

II. Materials and Methods

This is an interprofessional project involving faculty and students from the Physical Therapy, Computer Engineering, and Artificial Intelligence programs at the Federal University of Goiás (UFG). The study employed a field-based, observational, descriptive, and quantitative research design conducted with employees at an Educational Technology office located in the city of Goiânia, Brazil.

Participants were selected using the criteria-based sampling method [11]. Eligible workers were those who had been employed at the office for more than one year and who reported some work-related pain or musculoskeletal disorder. In the end, the sample consisted of 22 workers, including 15 women and seven men with a mean age of ± 32 years (range: 18–57 years).

The data collection instrument followed the measurement techniques recommended by [12] and [13], utilizing self-report and observational methods. A medical history questionnaire was used, containing personal information about the workers, work-related complaints or conditions, and an assessment of postural problems. The Rapid Upper Limb Assessment (RULA) and Rapid Entire Body Assessment (REBA) ergonomic postural assessment protocols were also used to evaluate the risk of RSI/WMSDs among the workers.

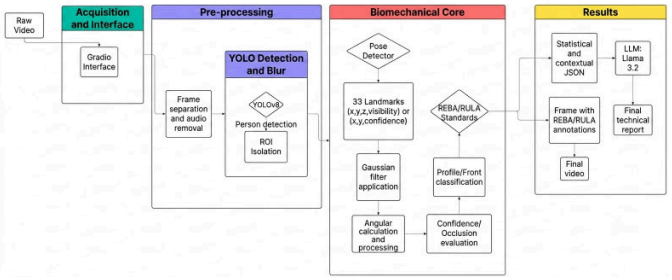


Fig. 1. Proposed framework from raw video to report generation via LLM.

The equipment used included an RGB camera with HD resolution (1024×768 pixels) and a cell phone tripod. Workers underwent video recording to estimate their postures during work, viewed from the right and left sides, as well as from the front and back, with an average recording duration of 15 minutes per session.

MediaPipe was used as the computational tool, employing a single-person estimation approach for human pose detection using Computer Vision to identify and track the positions of keypoints on the human body during the performance of routine office tasks, such as typing, customer service, using a cell phone, and handling documents [9]. The data were tabulated quantitatively using Microsoft Excel 2014 and Python. The study complies with Resolution No. 510/2016 of the National Health Council (CNS) and was approved by the UFG Research Ethics Committee for Studies Involving Human Subjects.

A. Methodology

The proposed framework was structured into four sequential stages, as illustrated in Fig. 1: (i) Acquisition and Interface, in which the raw video was received via the Gradio interface and the audio was suppressed; (ii) Preprocessing, involving person detection and facial anonymization; (iii) Biomechanical Core, involving pose estimation and calculation of RULA and REBA scores; and (iv) Summarization via LLM, synthesizing the data into an actionable textual report.

1) Data collection: The dataset consisted of recordings of 22 workers at office workstations with varied configurations of furniture, monitors, and lighting. Video capture was performed using a Samsung S21FE smartphone mounted on a tripod, with the lens positioned between the worker's iliac crest and the armrests of the chair. The videos were recorded from a side view (profile) and an oblique view (approximately 45° relative to the frontal plane), an effective strategy for capturing subtle postural compensations by workers, such as excessive cervical rotation while monitoring secondary screens.

2) Preprocessing and Privacy: This stage consists of three steps. First, the YOLOv8 model [14] was used to detect people by selecting the most reliable bounding box in the scene. Second, the cropped image was fed to the pose estimator to infer body landmarks. Third, a dynamic Gaussian filter was applied to the detected facial region [15], ensuring that no identifiable images are displayed or stored, in compliance

with the Brazilian General Data Protection Law (LGPD, Law No. 13,709/2018). In addition to visual anonymization, the subject's privacy was also protected by removing the audio tracks to prevent the capture of sensitive verbal information in the workplace, in line with the Privacy by Design approach [16], which represented a significant ethical distinction for workplace monitoring applications.

3) RULA and REBA scores: MediaPipe Pose [17] was evaluated as a pose estimator that utilized a topology of 33 three-dimensional landmarks on the human body. This pose estimation model provided coordinates for each anatomical keypoint, also known as landmarks. Visibility and confidence indicators served as filtering criteria: a landmark was included in the calculation of RULA/REBA scores only if its indicator exceeded the threshold of 0.5. Below this threshold, the system reports a detection failure for the frame in question, preventing spurious angles from contaminating the RULA and REBA scores. Thus, the ergonomic scores were obtained through mapping, in which joint angles (actual values) were quantized into discrete thresholds of 20°, 45°, 90°, in accordance with protocol standards, producing the scores that make up the final risk indices.

The RULA and REBA methods were implemented as lookup tables in direct-search structures that faithfully replicated the original scoring matrices. The joint angle between two adjacent body segments, represented by the vectors $\vec{a}$ and $\vec{b}$ formed from the (x, y) coordinates of the landmarks, was calculated using the Euclidean dot product. The calculations were performed frame by frame, generating a time series of scores. The final score was obtained by a weighted average over the processed time series. The methods were implemented in conjunction: RULA for focusing on the upper limbs and neck, and REBA for whole-body analysis, as comparative studies demonstrate that their sensitivities differ depending on the task dynamics [18].

4) Summarization and automatic generation of technical reports: To translate the RULA and REBA scores into guidance accessible to end users (occupational physical therapists and ergonomists), the Llama 3.2 language model was adopted. This is a quantized model optimized for low computational cost and local inference on embedded computers, as shown in Fig. 2.

The model was adapted to the specific context of physical therapy using the context injection technique: the processed statistical data (angular means, standard deviations, maximum scores, and user metadata—age and gender) were structured into a JSON object and injected as grounding into the prompt. The system prompt incorporated strict guardrails that restrict the output to the identification of biomechanical risks based on the provided data, explicitly refraining from providing clinical diagnoses or prescribing physical therapy treatments. This scope limitation was essential to validate the tool as an ergonomics aid and not as a substitute for a healthcare professional.

ERGONOMIC ANALYSIS REPORT (GENERATED BY LLM)

Analysis ID: video_20251120_081301 | Role: marketing_analyst

1. EXECUTIVE SUMMARY

Moderate risk (Average RULA: 4.08). The cervical region is the main target, partial score 2.85 | the highest among segments.

2. TECHNICAL DETAILS

- Average RULA: 3.99 (Max: 5.0) | Average REBA: 4.65
- Neck: 33.56° (±8.15°) --- Score 2.85 [CRITICAL]
- Trunk: 7.33° (±9.95°) --- Score 1.72
- Arm: 31.23° (±9.65°) --- Score 1.89

3. RECOMMENDATIONS

- Adjust monitor height to eye level.
- Active breaks every 50 min.

NOTE: This is an auxiliary tool. You should consult a professional!

Fig. 2. Example of a technical report automatically generated by the Llama 3.2 model. The system translates vector metrics into actionable natural language, remaining within the scope of an auxiliary tool for Occupational Physical Therapy and Ergonomics.

B. Results and Discussion

The characteristics of 22 workers assessed using a medical history questionnaire are presented in Table I. They are categorized by job position, the presence of functional limitations affecting the workers' ability to perform office work, work-related complaints or illnesses, and the main postural problems reported by the workers.

1) Segmental Analysis and Clinical Validation: Table II presents the analysis of joint angles and the local and global RULA and REBA scores obtained using the proposed Framework. The individual analysis of the joint angles yielded the cervical region as the area of greatest ergonomic concern. Although the neck exhibits a relatively low standard deviation ($\sigma = 8.15°$), indicating static posture, the mean cervical flexion is $33.56°$, which is sufficient to raise the risk score to 2.85 — the highest among all segments. This finding corroborates the literature, which points to a significant prevalence of musculoskeletal disorders in the neck-shoulder region among office workers [19].

The forearm exhibited the greatest angular variability in the study ($\sigma = 16.22°$), reflecting the dynamic nature of typing and mouse use activity. The trunk demonstrated high postural stability after settling into the chair (median of $4.46°$), indicating that the greatest chronic risk factor lies in the position of the head relative to the screen, not in the lumbar posture adopted by the worker while performing office tasks.

In general, the global median scores, RULA of 3.65 and REBA of 4.93, place the participants in the "investigation and potential for imminent change" range, aligning with the typically more conservative sensitivity of the automated RULA [6], [20]. These data are important to help occupational therapists recommend specific measures to address RSI and

TABLE I
RESULTS OF THE MEDICAL HISTORY QUESTIONNAIRE AND THE POSTURAL ASSESSMENT QUESTIONNAIRE ADMINISTERED TO EMPLOYEES AT THE OFFICE LOCATED IN GOIANIA, BRAZIL ($N = 22$).

Characteristics of the workers	**N**	%
Job position		
Support	7	31.8
Management	5	22.7
Executive	1	4.5
Web Designer	1	4.5
Project coordination	1	4.5
Video producer	1	4.5
Data analyst	1	4.5
Consulting	3	13.6
Functional limitations at work?		
yes	9	40.9
no	13	59.1
Work-related complaints or illnesses (*)		
Lumbar disc herniation	2	9
Lumbar osteophytes (6.6%)	2	9
Gout (6.6%)	2	9
Pain and muscle tension in the lumbar spine	3	13.6
Pain and muscle tension in the Cervical spine	3	13.6
Knee pain (13.3%)	3	13.6
Fibromyalgia (6.6%)	1	4.5
Carpal tunnel syndrome (6.6%)	1	4.5
Major postural problems among workers (*)		
Cervical hyperlordosis	10	45.4
Lumbar hyperlordosis (72.7%)	16	72.7
Thoracic hyperkyphosis (36.3%)	8	36.3
Scoliosis (22.7%)	5	22.7
Shoulder asymmetry (36.3%)	8	36.3
Elbow abnormalities (Thales' triangle)	4	18.1
Pelvic anteversion (18.1%)	4	18.1
Knock-knees (27.2%)	6	27.2
Knock-knees (13.6%)	3	13.6

(*) Employees could exhibit more than one physical abnormality.

TABLE II
JOINT ANGLES AND SCORES ($N = 22$).

Variable	μ	σ	**Median**
Joint angle (°)			
Neck (cervical)	33.56	8.15	34.67
Trunk	7.33	9.95	4.46
Arm (upper)	31.23	9.65	32.05
Forearm	114.91	16.22	114.93
Local score			
Knee	2.85	0.43	2.96
Trunk	1.72	0.43	1.97
Legs	1.85	0.34	2.00
Arm	1.89	0.36	2.00
Forearm	1.83	0.29	1.97
Wrist	0.99	0.04	1.00
Global score			
RULA	3.65	0.50	**3.65**
REBA	4.65	0.57	**4.93**

WMSDs, such as adjusting the chair and monitoring height, taking active breaks and performing stretches to minimize pain complaints among affected workers.

2) Comparative assessment of Goniometry against the proposed automated Framework: Table III shows the comparative joint range of motion (ROM) measurements collected by independent professionals using Goniometry against the proposed Framework, together with the corresponding divergence (ϵ).

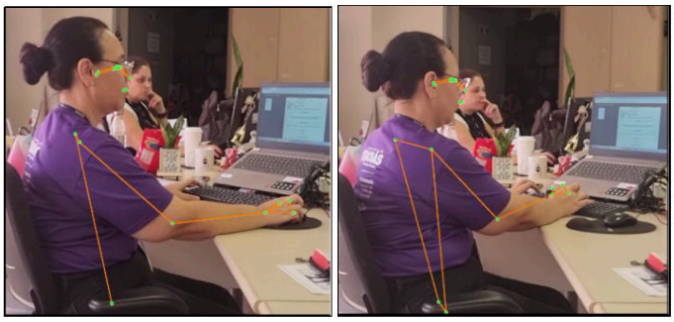


Fig. 3. Keypoints from the MediaPipe tool for estimating the human pose of a female office worker.

Following statistical tolerance criteria, the proposed Framework proved effective in computing joint angle measurements, with errors increasing due to artifacts, such as occluded joints (*e.g.*, ankle, wrist), or artificial landmarks that were inferred by proximal joints (*e.g.*, cervical).

TABLE III
COMPARATIVE JOINT RANGE OF MOTION MEASUREMENTS

Joint angle	**Goniometry (°)**	**Framework (°)**	ϵ (°)
Participant 1			
Neck (cervical)	14	7	7
Shoulder	20	23–25	3–5
Elbow	50–52	49.5–53.5	0.5–3.5
Wrist	12	13.5–18	1.5–6
Lumbar	18	12–14	4–6
Knee	100	89.5–92	8–10.5
Ankle	8–10	7–9	1–2
Participant 2			
Neck (cervical)	36	40–49	4–13
Shoulder	60	33	27
Elbow	10	11–3	1–7
Wrist	8–10	17–30	8–21
Lumbar	34–36	16–17	19
Knee	90	80–75	10–15
Ankle	6	7–10	1–4

3) Automated Framework for Ergonomic Scores Estimation: Regarding the collection of videos for subsequent transcription of the results, the first step was to use MediaPipe Pose to estimate the pose of the anatomical points of interest for assessment in office workers, represented in Fig. 3 by red vectors and green points.

Subsequently, during video data collection, a dynamic Gaussian filter was applied to the detected facial region, ensuring that no image of the worker was identifiable. Next, the system automatically generated summaries and technical reports, with processed statistics (angular means, standard deviations, maximum scores) automatically translated for interpretation by professionals.

From a computational standpoint, this work differs from the state of the art by combining multiple functionalities into a single framework for automated ergonomic assessment,

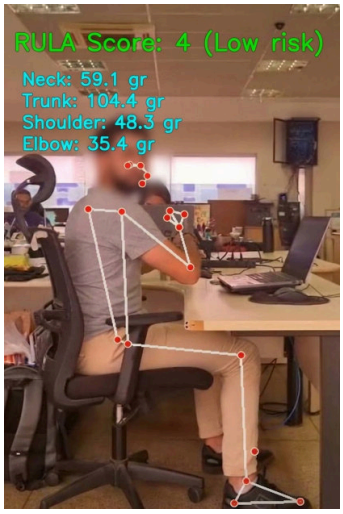


Fig. 4. Representation of the framework in operation: biomechanical skeleton superimposed on the participant with active Gaussian facial anonymization and the RULA score displayed in real time, alongside joint angles by segment.

which integrates: *(i)* a comparative study of pose estimation methods in real-world ergonomic scenarios; *(ii)* a hybrid top-down architecture for pose estimation from videos captured by a single camera; *(iii)* frame-by-frame automation of the RULA and REBA protocols for greater temporal stability; *(iv)* automated anonymization of images and videos compliant with the Brazilian General Data Protection Law (LGPD); *(v)* automatic generation of technical reports through structured prompt engineering using the Llama 3.2 model.

Computer Vision and Generative Artificial Intelligence applications are progressively gaining ground towards the digital transformation of functional and ergonomic assessment of workers. The authors in [21] demonstrated that the implementation of pose detectors on smartphones for classifying physical therapy exercises for the shoulder and lower back achieved 95% accuracy. They also found that a sample size of 7 to 11 patients is sufficient to generate training data that results in efficient classifications.

In [22], it was evaluated the performance of the MediaPipe model for pose estimation in comparison with motion capture (MoCap) systems, concluding that the model's accuracy is highly dependent on the type of exercise and the camera's capture angle, which reinforces the need for standardized data collection protocols in the field of physical therapy.

III. Conclusion

The results revealed in the medical history survey conducted with 22 office workers in the city of Goiania, Brazil, indicated the presence of functional limitations in nearly half of the workers, with evidence of pain and muscle tension in the cervical and lumbar spine and knees.

Regarding the postural abnormalities reported by the workers, a predominance of postural abnormalities in the cervical spine was observed. This study also presented a new proposal for an automated and non-invasive framework for postural and ergonomic analysis of workers in office environments, integrating Computer Vision and Generative Artificial Intelligence. The central contribution lies in the validation of the top-down hybrid architecture in a real-world ergonomic setting, demonstrating that temporal stability is the determining factor for the reliability of the RULA and REBA scores that were automated in this tool. This automated system offers a technical and ethical advantage by implementing facial anonymization prior to biomechanical rendering, ensuring compliance with the Brazilian LGPD legislation.

Segmental analysis revealed that the cervical region was the primary target of physical therapy and ergonomic interventions for office workers in the evaluated dataset, with median overall RULA scores of 3.65 and REBA scores of 4.93.

Future work will focus on hybrid validation using a digital twin protocol to generate a synthetic ground truth with precisely known angles, as well as hand-tracking modules to identify RSI/WMSD risks associated with excessive keyboard and mouse use, and the migration of data to mobile devices.

References

[1] H. A. Couto and D. C. Couto, *Ergonomics 4.0—From Basic Concepts to the Fourth Industrial Revolution*. Belo Horizonte, Brazil: Ergo, 2020.

[2] I. Iida, *Ergonomia: Projeto e Produção*, 2nd ed. São Paulo, Brazil: Blucher, 2005.

[3] L. McAtamney and E. N. Corlett, "RULA: a survey method for the investigation of work-related upper limb disorders," *Applied Ergonomics*, vol. 24, no. 2, pp. 91–99, 1993.

[4] S. Hignett and L. McAtamney, "Rapid Entire Body Assessment (REBA)," *Appl. Ergon.*, vol. 31, no. 3, pp. 201–205, 2000.

[5] J. A. Andrade, V. M. Leite, L. F. Teixeira-Salmela, P. M. P. Araujo, and Y. Juliano, "A comparative study between visual estimation and goniometry methods for assessing the range of motion of the shoulder joint," *Acta Fisiátrica*, vol. 10, no. 1, pp. 12–16, 2003.

[6] A. Krishnan, X. Yang, U. Seth, and C. C. Menassa, "Data-driven ergonomic risk assessment of complex hand-intensive manufacturing processes," *Commun. Eng.*, vol. 4, no. 1, Art. no. 45, 2025.

[7] M. M. Fernández, S. Pérez-Sánchez, and J. M. Gómez, "Ergonomic risk assessment based on computer vision and machine learning," *Comput. Ind. Eng.*, vol. 149, Nov. 2020, Art. no. 106793.

[8] Z. Gao *et al.*, "A systematic survey on human pose estimation: Upstream and downstream tasks, approaches, lightweight models, and prospects," *Artif. Intell. Rev.*, vol. 58, no. 3, Art. no. 68, 2025.

[9] J. E. Estrada, L. A. Vea, and M. Devaraj, "Modeling proper and improper sitting posture of computer users using machine vision for a human–computer intelligent interactive system during COVID-19," *Appl. Sci.*, vol. 13, no. 10, p. 6163, 2023.

[10] X. Zhang, X. Yan, and H. Li, "Vision-based estimation of 3D poses for construction workers from a single 2D camera," *Automation in Construction*, vol. 92, pp. 242–258, Aug. 2018.

[11] J. R. Thomas, J. K. Nelson, and S. J. Silverman, *Research Methods in Physical Activity*, 6th ed. Champaign, IL, USA: Human Kinetics, 2012.

[12] P. Plantard, H. Aubert, and C. PONTINNE, "Validation of an ergonomic assessment method using Kinect data in real workplace conditions," *Appl. Ergon.*, vol. 67, pp. 562–569, Nov. 2017.

[13] X. Yan, H. Li, H. Li, and A. R. Anwer, "Automated ergonomic risk assessment for construction workers using body part localization," *Automation in Construction*, vol. 84, pp. 24–35, Dec. 2017.

[14] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," version 8.0.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[15] M. S. Ahmed and H. Yoshiura, "Dynamic Gaussian filtering for facial privacy preservation in video analytics," *IEEE Access*, vol. 13, pp. 11204–11215, Jan. 2025.

[16] A. Cavoukian, "Privacy by design: The 7 foundational principles," Information and Privacy Commissioner of Ontario, Canada, Tech. Rep., 2009.

[17] V. Bazarevsky, I. Grishchenko, and K. Raveendran, "BlazePose: On-device real-time body pose tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR) Workshops*, 2020.

[18] K. Sirikasemsuk, P. Luangpaiboon, and S. Kaitwanidvilai, "Comparative sensitivity analysis of RULA and REBA assessment methods in varied occupational tasks," *International Journal of Industrial Ergonomics*, vol. 101, p. 103542, May 2024.

[19] Secretaria de Vigilância em Saúde e Ambiente, "DORT e LER: Análise epidemiológica e vigilância em saúde do trabalhador," *Boletim Epidemiológico*, vol. 56, no. 19, 2025.

[20] R. Suwandechochai and P. Klomjit, "Work posture risk comparison of RULA and REBA based on measures of assessment-score variability: A case study of the metal coating industry in Thailand," *Global Journal of Engineering and Technology Advances*, vol. 18, no. 3, pp. 161–169, 2024.

[21] C. Arrowsmith, C. Whyne, and D. Burns, "Physiotherapy exercise classification using single-camera pose detection and machine learning," *Sensors*, vol. 23, no. 1, p. 363, 2023.

[22] S. Dill *et al.*, "Accuracy evaluation of 3D pose estimation with MediaPipe Pose for physical exercises," *Curr. Dir. Biomed. Eng.*, vol. 9, no. 1, pp. 1–4, 2023.

As Fragilidades do Sistema HÓRUS na Gestão de Farmácias no SUS

1st Liana Lysia Cavalcanti Vieira de Brito
Escola de Ciências e Tecnologia (ECT)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0001-2928-7522

2rd Maria Eduarda da Silva Ferreira
Escola de Ciências e Tecnologia (ECT)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0006-5258-7757

3rd Joêmilly Wellen Lima da Silva
Departamento de Engenharia Biomédica (DEB)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0004-3464-0070

4th Antonio de Oliveira Nunes Costa
Mestrando em Engenharia Elétrica e de Computação
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0000-0003-2970-6245

5th Ricardo Alexsandro de Medeiros Valentim
Professor Titular do Departamento de Engenharia Biomédica e LAIS
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0000-0002-9216-8593

Resumo*—A Assistência Farmacêutica no Sistema Único de Saúde (SUS) enfrenta os desafios relacionados à garantia do acesso equânime e o uso racional de medicamentos em um cenário de descentralização da gestão. Nesse cenário, o Sistema Nacional de Gestão da Assistência Farmacêutica (Hórus) foi desenvolvido para integrar informações, qualificar os processos de gestão e fortalecer o planejamento das ações farmacêuticas. Este estudo teve como objetivo analisar as fragilidades operacionais e os impactos da implantação do Hórus na gestão da Assistência Farmacêutica no SUS. Para isso, foi realizada uma revisão bibliográfica, complementada pela análise de documentos normativos e da literatura especializada sobre o tema. Os resultados evidenciam que o Hórus proporcionou avanços importantes, como o aprimoramento do controle de estoque, da rastreabilidade de medicamentos e da qualidade das informações utilizadas na gestão. Entretanto, a literatura aponta que sua efetividade ainda é limitada por problemas relacionados à infraestrutura das unidades de saúde, à insuficiência de capacitação dos profissionais, à fragmentação dos sistemas de informação e à permanência de modelos de assistência predominantemente centrados na logística de medicamentos, em detrimento da atenção farmacêutica. Conclui-se que o Hórus constitui uma importante ferramenta para a gestão da Assistência Farmacêutica no SUS, porém seu pleno potencial depende de investimentos contínuos em infraestrutura, qualificação profissional, integração dos sistemas de informação e reorganização dos processos de trabalho, de modo a fortalecer a atuação clínica do farmacêutico e ampliar a qualidade da assistência prestada à população.

***Palavras-chave*—Hórus; Farmácia; Gestão de Farmácias; SUS.**

I. INTRODUÇÃO

A transformação digital na saúde pública brasileira não representa apenas uma mudança de suporte técnico, mas uma estratégia vital para a eficiência do gasto público e a transparência institucional. Em um cenário de recursos limitados, a transição de modelos de gestão manuais para sistemas informatizados torna-se um imperativo para a resolubilidade das ações de saúde e para a utilização estratégica da Tecnologia da Informação (TI) na gestão governamental [5], [9].

Historicamente, a Assistência Farmacêutica (AF) no Brasil foi caracterizada por ações fragmentadas, centralizadas e verticais — exemplificadas pela antiga Central de Medicamentos (CEME) [7], [15] . A promulgação da Constituição de 1988 e a criação do Sistema Único de Saúde (SUS) estabeleceram a saúde como direito de todos e dever do Estado, impulsionando a descentralização do sistema [18], [5], [14]. Contudo, a descentralização transferiu aos municípios a responsabilidade pela gestão direta dos serviços, revelando assimetrias na capacidade administrativa, orçamentária e operacional das gestões locais [18].

Para superar os desafios da descentralização e mitigar as ineficiências na gestão de estoques, a introdução de Tecnologias de Informação e Comunicação (TICs) nos serviços de saúde assumiu papel estratégico [5], [7]. Sistemas de informação de âmbito local constituem-se como recursos estratégicos para democratizar o acesso, aperfeiçoar o controle do gasto público e subsidiar os processos decisórios [5], [11], [2].

Nesse contexto de inovação, o Ministério da Saúde desen-

volveu e disponibilizou gratuitamente, em 2009, o Sistema Nacional de Gestão da Assistência Farmacêutica (Hórus) [7], [5]. O Hórus foi concebido para atuar em todo o território nacional, permitindo o controle de estoque de medicamentos, a rastreabilidade do envio ao consumo final e o monitoramento dos recursos financeiros investidos nas esferas federal, estadual e municipal [5].

No entanto, a simples disponibilização de uma solução tecnológica não garante a efetividade da gestão pública ou a melhoria da atenção à saúde. A literatura demonstra que a implantação de TICs em órgãos públicos municipais enfrenta barreiras estruturais significativas, como escassez de infraestrutura de conectividade, falta de equipamentos adequados, déficit de pessoal qualificado e resistência cultural às transformações nos processos de trabalho [5], [1], [10]. Ademais, subsiste o desafio de integrar a dimensão logística da gestão à dimensão clínica da atenção à saúde, de modo a assegurar que o registro de dados se converta em indicador de qualidade e no uso racional de medicamentos [18], [11].

Considerando que a garantia do acesso sustentável a medicamentos essenciais é condição fundamental para a efetivação do direito à saúde no SUS, torna-se indispensável analisar criticamente o desempenho das ferramentas informatizadas utilizadas no país [18].

II. METODOLOGIA

Este trabalho fundamenta seu referencial teórico em uma revisão bibliográfica, realizada com o propósito de reunir, sistematizar e analisar estudos previamente publicados sobre o Sistema HÓRUS e sua aplicação na gestão da Assistência Farmacêutica no Sistema Único de Saúde (SUS), contribuindo para a contextualização dos resultados e para a fundamentação da discussão apresentada.

A busca bibliográfica foi realizada nas bases Portal de Periódicos CAPES, Biblioteca Virtual em Saúde (BVS) e Scientific Electronic Library Online (SciELO), complementada pela consulta a documentos oficiais do Ministério da Saúde. Para a identificação dos estudos, foram utilizados os descritores e termos relacionados a “Sistema HÓRUS”, “HÓRUS”, “Assistência Farmacêutica”, “Gestão de Medicamentos”, “Sistemas de Informação em Saúde”, “Atenção Primária” e “Sistema Único de Saúde (SUS)”, combinados por meio dos operadores booleanos AND e OR.

Foram realizadas três estratégias de busca. A primeira utilizou a combinação: (”Sistema Hórus” OR ”Hórus”) AND (”Assistência Farmacêutica” OR ”Gestão de Medicamentos”) AND (”Atenção Primária” OR ”SUS”). Nessa estratégia, foram identificados 1 resultado na SciELO, sendo 1 incluído; 12 resultados no Portal de Periódicos CAPES, considerando publicações do período de 2013 a 2026, dos quais 2 foram incluídos; e 6 resultados na BVS, referentes ao período de 2021 a 2023, dos quais 2 foram incluídos, totalizando 19 registros identificados e 5 estudos selecionados.

A segunda estratégia empregou a combinação (”Sistema Hórus” OR ”Hórus”) AND (”Sistemas de Informação em Saúde” OR ”Assistência Farmacêutica”). Foram identificados 26 resultados no Portal de Periódicos CAPES, com recorte temporal de 2012 a 2023, dos quais 3 foram incluídos, e 26 resultados na BVS, dos quais 2 foram incluídos. Essa estratégia resultou em 52 registros identificados e 5 estudos selecionados.

A terceira estratégia utilizou a combinação (”Sistema Hórus” OR ”Hórus”) OR (”Assistência Farmacêutica” OR ”Gestão de Medicamentos”) AND (”Atenção Primária” OR ”SUS”). Na SciELO, foram identificados 145 resultados, considerando o período de 2000 a 2026, dos quais 10 estudos foram selecionados, sendo identificado 1 registro duplicado entre os resultados recuperados.

Ao todo, as três estratégias de busca resultaram em 216 registros identificados (19 + 52 + 145). Após a aplicação dos critérios de seleção e a análise dos resultados, foram selecionados 20 estudos para avaliação na íntegra. Foram considerados elegíveis estudos que abordassem diretamente o Sistema HÓRUS, sua utilização na Assistência Farmacêutica, sua relação com Sistemas de Informação em Saúde ou sua aplicação na gestão de medicamentos no âmbito do SUS e da Atenção Primária à Saúde. Foram consideradas publicações em português ou inglês que apresentassem informações pertinentes aos objetivos da pesquisa.

Foram excluídos estudos que não apresentassem relação direta com o objeto da pesquisa, publicações incompatíveis com os objetivos estabelecidos e registros duplicados. A seleção dos estudos foi realizada mediante análise dos resultados recuperados e posterior avaliação dos trabalhos selecionados em texto completo. O processo de identificação, seleção e inclusão dos estudos está apresentado no Diagrama do processo de seleção dos estudos (Figura 1). Dos 20 estudos avaliados na íntegra, 2 foram identificados como duplicados, resultando em 18 estudos incluídos no escopo da revisão.

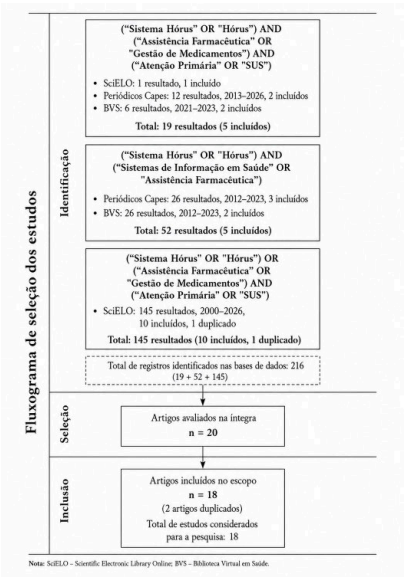


Fig. 1. Fluxograma de seleção dos estudos

Para sistematizar as informações dos estudos incluídos, foi

elaborada uma tabela de caracterização contendo local do estudo, período, amostra, metodologia empregada e principais fragilidades relacionadas ao Sistema HÓRUS identificadas pelos autores (Figura 2). Essa sistematização permitiu comparar os diferentes contextos investigados e identificar aspectos recorrentes relacionados à utilização, implantação e funcionamento do sistema na gestão da Assistência Farmacêutica.

Estudo (Título e Autor)	Local	Amostra	Metodologia	Fragilidades do Hórus Identificadas	Ano
1. Institucionalização da AF em Ananindeua (Torres et al.)	Ananindeua, PA	17 Farmácias Distritais e 3 Especializadas	Pesquisa-ação, exploratória e descritiva	Lacunas no cumprimento de práticas de governança; instabilidade sistêmica.	2021
2. Institucionalização da AF em Benevides (Torres & Soler)	Benevides, PA	31 estabelecimentos de saúde	Exploratória e descritiva (Planejamento Estratégico)	Ausência de ferramentas de controle; sistema operante apenas no componente estratégico.	2023
3. Implantação do HÓRUS nas farmácias do SUS (Dias, K.)	6 municípios (MG, PR, SP)	6 municípios brasileiros	Pesquisa exploratória e de campo (entrevistas/observação)	Falta de internet; déficit de pessoal; lentidão; obrigatoriedade de conexão online.	2013
4. Experiência da estruturação da AF municipal (Azevedo et al.)	Santa Isabel, SP	13 unidades (CS, UBS, ESF, CAF)	Relato de experiência	Lentidão no processamento; necessidade de plataforma offline; problemas de conectividade.	2018
5. HÓRUS: Ferramenta de gestão para AF (Franco et al.)	Iguatu, CE	1 farmácia municipal	Estudo observacional qualitativo	Paralisações frequentes na base nacional em horário comercial; inércia do serviço por falta de preenchimento.	2023
6. Assistência Farmacêutica no SUS (Ribeiro et al.)	Brasil	Revisão bibliográfica	Levantamento bibliográfico	Falta de acesso total à internet nas farmácias e CAF limitando a implantação.	2023
7. Serviço de dispensação de medicamentos (Leite et al.)	Brasil (Nacional)	1.175 farmácias/unidades	Estudo transversal (PNAUM)	Baixa informatização na região Norte (14,3%); falta de integração sistêmica.	2017
8. Caracterização da institucionalização da AF (Souza et al.)	Brasil (Nacional)	507 RAF e 369 SMS	Estudo transversal (PNAUM)	Baixa utilização do Hórus (apenas 16,3% dos que possuem sistema informatizado o utilizam) 11.	2017
9. Gestão da assistência farmacêutica (Gerlack et al.)	Brasil (Nacional)	506 responsáveis pela AF	Estudo transversal (PNAUM)	30% dos municípios sem sistema; menos da metade ligado em rede; falta de uso para controle financeiro.	2017
10. Avaliação da capacidade de gestão da AF (Faraco et al.)	Brasil (Nacional)	600 municípios brasileiros	Quantitativo e transversal (PNAUM)	Fragilidades organizacionais críticas (94% sem CFT); grandes disparidades regionais.	2025
11. AF no Brasil: Análise antropológica (Penaforte, T.)	Brasil (Nacional)	Atas do CNS (1999-2004)	Análise documental antropológica	Tensionamentos entre foco logístico vs. cuidado clínico; lacunas de governança institucional.	2026
12. Avanços e desafios da AF na APS (Costa et al.)	Brasil (Nacional)	Dados gerais do SUS	Estudo avaliativo	Complexidade da gestão descentralizada; disparidades estruturais regionais.	2017
13. Disponibilidade de medicamentos (Mendes et al.)	Brasil (Nacional)	38.812 UBS (PMAQ-AB)	Estudo transversal	Infraestrutura precária (climatização e armazenamento); falta de apoio farmacêutico em unidades.	2014
14. AF na APS: Pactuação interfederativa (Costa et al.)	Brasil (Nacional)	Dados do QUALIFAR-SUS	Estudo avaliativo	Falhas na coordenação entre esferas; necessidade premente de integração de dados nacionais.	2017
15. Acesso a medicamentos e institucionalização (Barros et al.)	Brasil (Nacional)	4.866 usuários e 507 RAF	Estudo transversal (PNAUM)	55,6% dos usuários buscam med. em unidades sem sistema; rede de integração insuficiente 23.	2017
16. Gestão da AF: Controle e farmacoeconomia (Meirelles et al.)	Salto, SP	13 unidades de AF	Relato de experiência	Ineficácia inicial por registros manuais; necessidade de capacitação permanente da equipe.	2015
17. Mudanças na Gestão da AF (Paraíba) (Gomes et al.)	Paraíba (Estadual)	CEDMEX (Componente Esp.)	Estudo descritivo e analítico	Infraestrutura precária e sistema ainda em fase de ajustes operacionais	2013
18. Avaliação do grau de implementação (Inácio, V.)	DSEIs Xavante e Xingu	2 Distritos Indígenas	Estudo de caso múltiplos	Baixa conectividade (internet abaixo do mínimo); falta de RH treinado; banco de dados incompleto para MMH.	2018

Fig. 2. Diagnóstico da AF e fragilidades do Sistema Hórus

III. DESENVOLVIMENTO

A. Histórico e Institucionalização da Assistência Farmacêutica no SUS

A trajetória da Assistência Farmacêutica no Brasil reflete a evolução do próprio sistema público de saúde [16]. A partir da década de 1990, com a aprovação da Política Nacional de Medicamentos em 1998 e da Política Nacional de Assistência Farmacêutica em 2004, a AF passou a ser reconhecida não apenas como um setor de suprimento de insumos, mas como um conjunto de ações voltadas à promoção, proteção e recuperação da saúde [1], [5], [7].

A descentralização das ações de saúde atribuiu novos papéis e responsabilidades aos municípios [1]. Contudo, estudos sobre a caracterização da AF nos municípios brasileiros indicam que a institucionalização desse setor ocorreu de forma heterogênea, desigual e com acentuadas disparidades regionais [16], [6]. A pactuação interfederativa avançou na distribuição de competências orçamentárias e na ampliação da cobertura do fornecimento público de medicamentos, mas a gestão dos serviços nos municípios permaneceu marcada por vulnerabilidades técnicas, operacionais e limitações na autonomia financeira local [18], [8], [6], [4].

O acesso da população aos medicamentos está diretamente associado ao grau de institucionalização, estruturação técnica e organização da AF no nível local [2], [6]. A descontinuidade no suprimento de fármacos, somada às deficiências no armazenamento, no controle de estoques e na distribuição, compromete diretamente a eficácia das diretrizes nacionais de saúde e segurança do paciente. Assim, a informatização e a implantação de sistemas de informação surgiram como propostas centrais para conferir transparência, monitorar os recursos, otimizar a gestão logística e modernizar as estruturas administrativas da saúde pública [5], [7], [11], [3], [9] .

B. O Sistema Hórus: Avanços e Potencialidades de Gestão

O lançamento do Sistema Hórus em 2009 pelo Ministério da Saúde, em parceria com estados e municípios, representou um marco normativo e tecnológico [5], [7]. Conforme destacam Costa e Nascimento Jr [4] e Dias [5], a ferramenta foi concebida como um sistema Web descentralizado e gratuito, destinado a integrar o fluxo de informações da AF em todo o país.

Entre as principais potencialidades e avanços proporcionados pela implantação do Hórus, destacam-se:

- Controle de Estoque em Tempo Real: Permite o rastreamento da entrada e saída de medicamentos nas centrais de abastecimento farmacêutico e nas farmácias das Unidades Básicas de Saúde (UBS) [5], [7], [10].
- Rastreabilidade dos Lotes: Possibilita o acompanhamento detalhado dos lotes e prazos de validade, reduzindo as perdas por vencimento e garantindo a segurança do paciente [1], [5], [10], [12] .
- Visibilidade dos Custos e Agendamento de Atendimentos: Contribui para a transparência do gasto público e para a otimização dos fluxos de atendimento ao usuário [1], [5], [10].
- Apoio ao Planejamento e à Tomada de Decisão: Gera dados agregados que subsidiam a elaboração de estimativas de consumo e a pactuação de compras públicas [5], [7], [10], [12].

Em contextos específicos, como na gestão do Componente Especializado da AF no Estado da Paraíba, a implantação do Hórus demonstrou ganhos significativos no gerenciamento de tratamentos de alto custo, trazendo maior previsibilidade aos estoques estaduais e agilidade no atendimento [9].

C. Fragilidades Estruturais, Operacionais e de Capital Humano

A despeito do arranjo normativo e das avançadas especificações técnicas do Hórus, investigações empíricas realizadas no nível municipal apontam lacunas expressivas entre o modelo ideal concebido e a realidade operacional vivenciada pelos profissionais de saúde nas farmácias do SUS [5], [6], [8], [17], [18].

A análise da implantação do Hórus no Estado do Rio Grande do Sul evidencia que, embora centenas de municípios tenham aderido formalmente ao sistema, a adesão não se traduziu na plena utilização dos seus módulos [1], [5]. As principais fragilidades identificadas compreendem:

- Precariedade da Infraestrutura Tecnológica: A ausência ou instabilidade de conexões de internet nas farmácias públicas inviabiliza o uso continuado do sistema Hórus, que opera na modalidade on-line [1], [5]. A lentidão do sistema e a escassez de computadores adequados geram filas e atrasos no atendimento [1], [5].
- Déficit de Profissionais e Rotatividade: A escassez de farmacêuticos e equipes de apoio e a alta rotatividade de pessoal limitam o alimentação tempestiva dos dados no sistema [1], [5].
- Necessidade de Capacitação Continuada: Os treinamentos ofertados frequentemente revelam-se pontuais e insuficientes frente à complexidade das rotinas de dispensação e às atualizações do software [5], [7], [16].
- Alimentação Incompleta de Dados: A digitação de informações é por vezes negligenciada ou feita com atraso, comprometendo a confiabilidade dos relatórios e anulando a capacidade do Hórus de funcionar como gerador de indicadores gerenciais e assistenciais confiáveis [5].

A dicotomia entre as premissas de planejamento formuladas no nível federal e a capacidade de execução dos governos municipais compromete a consolidação de um modelo unificado de informação assistencial [6], [8].

D. A Tensão entre Tecnologia de Gestão e Tecnologia de Uso: O Papel do Farmacêutico

Uma fragilidade estrutural que perpassa o debate sobre o Hórus refere-se ao foco predominante do sistema na ”Tecnologia de Gestão” (logística, finanças, saldo de estoque) em detrimento da ”Tecnologia de Uso” (cuidado clínico, orientação ao paciente e acompanhamento farmacoterapêutico).

A literatura científica estabelece uma dicotomia crítica entre a ”Tecnologia de Gestão” (logística e abastecimento) e a ”Tecnologia de Uso” (orientação e cuidado clínico) [5], [8]. Por outro lado, enquanto Gomes et al. [9] apontam a eficácia da ferramenta para otimizar controles operacionais de suprimentos, diversos autores alertam para o risco de transformar o profissional farmacêutico em um mero operador de sistema e burocrata do estoque [7], [11].

Quando o ambiente de trabalho da farmácia pública é caracterizado pela alta demanda de usuários, escassez de equipamentos e digitação morosa de cadastros no Hórus, a dimensão relacional do atendimento é sacrificada [1], [5]. A dispensação reduz-se à simples entrega material do fármaco, desprovida da necessária orientação sobre posologia, conservação, efeitos adversos e adesão ao tratamento [11].

Estudos sobre o perfil da atenção primária e o serviço de dispensação na rede básica revelam que orientações clínicas adequadas são indispensáveis para evitar intoxicações, abandono de tratamento e falhas terapêuticas, sobretudo em populações vulneráveis, como idosos e portadores de doenças crônicas [1], [5], [11]. A sobrecarga burocrática provocada por sistemas de informação que não dialogam com a rotina do cuidado atua como uma barreira à consolidação da Atenção Farmacêutica centrada no paciente [11], [14].

Para que os sistemas de informação em saúde no nível local alcancem sua plenitude, a tecnologia deve atuar como meio facilitador do processo de trabalho, promovendo o acolhimento humanizado e a integração interprofissional, e não como um fim burocrático autossuficiente [5], [11], [16].

IV. SÍNTESE CRÍTICA E RECOMENDAÇÕES

A análise da literatura sobre o Sistema Hórus revela que a inovação tecnológica é um elemento necessário, porém insuficiente, para superar as fragilidades históricas da gestão da Assistência Farmacêutica no SUS. A efetividade do sistema depende não apenas de suas funcionalidades, mas também das condições estruturais, organizacionais e humanas existentes nos municípios. Nesse sentido, o aprimoramento da gestão informatizada requer ações articuladas que contemplem infraestrutura, qualificação profissional, qualidade da informação e reorganização dos processos de trabalho.

A Figura 3 sintetiza as principais fragilidades mapeadas e apresenta recomendações de aprimoramento relacionadas a cada uma delas.

DIMENSÃO	FRAGILIDADES IDENTIFICADAS	RECOMENDAÇÕES DE APRIMORAMENTO
Infraestrutura	Instabilidade da internet; falta de computadores e equipamentos nas UBS.	Investimento contínuo em conectividade e renovação do parque tecnológico municipal.
Capacitação	Treinamento pontual; alta rotatividade de pessoal nas farmácias.	Programas permanentes de educação continuada e fixação de profissionais qualificados.
Alimentação de Dados	Subnotificação, registros desatualizados e relatórios não utilizados.	Simplificação da interface do sistema e integração com o Prontuário Eletrônico do Cidadão (e-SUS PEC).
Prática Clínica	Foco exclusivo na logística; sobrecarga burocrática do farmacêutico.	Inclusão de módulos focados no registro do acompanhamento farmacoterapêutico e na dispensação orientada.

Fig. 3. Fragilidades do Sistema Hórus na gestão da Assistência Farmacêutica no SUS e recomendações de aprimoramento. Elaboração dos Autores (2026).

Para conferir maior aplicabilidade prática às diretrizes apresentadas na Figura 2, detalham-se, a seguir, estratégias operacionais de enfrentamento e mitigação articuladas às principais dimensões das fragilidades identificadas. Essas estratégias devem ser compreendidas como proposições de aprimoramento decorrentes da análise realizada, considerando que as limitações observadas são interdependentes e não podem ser enfrentadas exclusivamente por intervenções de natureza tecnológica.

- Infraestrutura e conectividade: Para enfrentar a instabilidade da conexão e a insuficiência de equipamentos nas farmácias públicas, os municípios devem priorizar investimentos em infraestrutura tecnológica e manutenção do parque computacional. A precariedade desses recursos constitui uma barreira direta à utilização contínua do Hórus, podendo ocasionar interrupções, atrasos no atendimento e dificuldades na alimentação tempestiva dos dados. Entre as estratégias possíveis, destacam-se a ampliação e qualificação da conectividade, a manutenção

preventiva dos equipamentos e, quando tecnicamente viável, a adoção de mecanismos que permitam o registro temporário de informações em situações de indisponibilidade de rede, com posterior sincronização. Essas medidas podem contribuir para ampliar a continuidade operacional do sistema e reduzir os impactos das limitações de infraestrutura sobre o atendimento e a qualidade dos registros.

- Capacitação e retenção de profissionais: O enfrentamento dos treinamentos pontuais e da elevada rotatividade exige a institucionalização de programas permanentes de educação continuada, com conteúdos atualizados sobre as funcionalidades do sistema e as rotinas da Assistência Farmacêutica. A literatura aponta que os treinamentos podem ser insuficientes diante da complexidade das rotinas de dispensação e das atualizações do software, enquanto a escassez e a rotatividade dos profissionais dificultam a alimentação tempestiva dos dados. Nesse contexto, recursos de educação a distância podem ampliar o acesso às capacitações e facilitar a atualização das equipes. Paralelamente, estratégias locais de valorização e fixação dos profissionais podem contribuir para preservar o conhecimento acumulado sobre os processos de trabalho e reduzir os impactos da rotatividade sobre a continuidade e a qualidade das atividades desenvolvidas.
- Alimentação e qualidade dos dados: A alimentação incompleta ou tardia das informações constitui uma das principais limitações para o aproveitamento do potencial gerencial do Hórus, uma vez que compromete a confiabilidade dos relatórios e dos indicadores produzidos pelo sistema. O enfrentamento dessa fragilidade requer a padronização dos processos de registro, o acompanhamento sistemático da completude dos dados e a utilização dos relatórios gerenciais como instrumentos efetivos de monitoramento e planejamento. No âmbito tecnológico, a ampliação da interoperabilidade entre sistemas de informação, por meio de interfaces de integração quando tecnicamente viáveis, pode reduzir a duplicidade de registros e a necessidade de digitação manual. Da mesma forma, painéis gerenciais simplificados podem oferecer retorno mais ágil aos gestores, favorecendo a identificação de inconsistências e o uso das informações para subsidiar decisões no nível local.
- Reorientação da prática clínica: A superação do predomínio da dimensão logística e da sobrecarga burocrática constitui um desafio central para que o Hórus contribua efetivamente para a qualificação da Assistência Farmacêutica. Conforme discutido anteriormente, a concentração das funcionalidades na gestão de estoque pode deslocar o trabalho do farmacêutico para atividades predominantemente administrativas, reduzindo o tempo disponível para a orientação e o acompanhamento dos usuários. Nesse sentido, a reorganização dos processos de trabalho e a distribuição de atividades administrativas e operacionais entre os diferentes integrantes da equipe, conforme suas atribuições e competências, podem contribuir para reduzir a sobrecarga do farmacêutico e ampliar sua disponibilidade para atividades de atenção farmacêutica. Adicionalmente, o aprimoramento dos recursos destinados ao registro do acompanhamento farmacoterapêutico e à dispensação orientada pode favorecer a utilização do sistema como instrumento de apoio ao cuidado, fortalecendo a atuação clínica do profissional.

De forma transversal, essas estratégias apontam para a necessidade de superar uma concepção segundo a qual a informatização, por si só, seria suficiente para qualificar a gestão da Assistência Farmacêutica. A tecnologia deve estar articulada às condições concretas de trabalho, à qualificação das equipes e à organização dos serviços, funcionando como instrumento facilitador dos processos de cuidado e de gestão. Essa perspectiva é particularmente importante diante da necessidade de integrar a dimensão logística à dimensão clínica da Assistência Farmacêutica, fazendo com que os dados registrados no sistema sejam convertidos em informações úteis para o planejamento, a tomada de decisão e o acompanhamento dos usuários.

Assim, as recomendações sintetizadas na Figura 1 não devem ser compreendidas como intervenções isoladas, mas como componentes de uma estratégia integrada de aprimoramento do Hórus. O fortalecimento da infraestrutura cria condições para o funcionamento adequado do sistema; a qualificação e a estabilidade das equipes favorecem a continuidade dos registros; a melhoria da qualidade e da interoperabilidade dos dados amplia seu potencial gerencial; e a reorganização dos processos de trabalho permite aproximar a tecnologia da atenção farmacêutica. Em conjunto, essas medidas podem contribuir para que o Hórus ultrapasse sua função de controle logístico e se consolide como instrumento de apoio à gestão e ao cuidado em saúde.

V. CONCLUSÃO

O Sistema Nacional de Gestão da Assistência Farmacêutica (Hórus) representa um inegável avanço na busca por eficiência, rastreabilidade e transparência no uso dos recursos públicos destinados à saúde. As evidências teóricas e empíricas confirmam a sua capacidade de qualificar o controle de estoques e de subsidiar o planejamento em esferas governamentais.

Contudo, a persistência de fragilidades estruturais — sobretudo no nível municipal — demonstra que a tecnologia por si só não resolve os problemas de gestão e de atenção à saúde. A precariedade da infraestrutura física, a escassez de conectividade, a carência de pessoal qualificado e a prevalência de um modelo administrativo burocratizado limitam o potencial do Hórus.

Para que o Hórus e outros sistemas informatizados cumpram integralmente seu papel no SUS, faz-se necessário adotar uma abordagem holística que articule investimentos em infraestrutura de TI, fortalecimento da governança interfederativa e requalificação dos processos de trabalho. Sobretudo, a tecnologia deve ser ressignificada para servir como instrumento de apoio ao cuidado clínico, garantindo que a eficiência

logística resulte, em última instância, no acesso equânime aos medicamentos, na segurança do paciente e na promoção da saúde.

REFERÊNCIAS

[1] AZEVEDO, Ednai Soares de; RAMOS, Elia; COSTA, Joice Vieira da; NOGUEIRA, Maria Helena Haux; FERNANDES, Mayara Serafim; BARBOSA, Nívea Alves. Experiência de estruturação da assistência farmacêutica na atenção básica municipal. In: Prêmio David Capistrano de Experiências Exitosas na Área da Saúde, 8., 2018 (Relato de Experiência de Santa Isabel/SP).

[2] BARROS, Rafael Damasceno de; COSTA, Ediná Alves; SANTOS, Djanilson Barbosa dos; SOUZA, Gisélia Santana; ÁLVARES, Juliana; GUERRA JUNIOR, Augusto Afonso; ACURCIO, Francisco de Assis; GUIBU, Ione Aquemi; COSTA, Karen Sarmento; KARNIKOWSKI, Margô Gomes de Oliveira. Access to medicines. Revista de Saúde Pública, [S.L.], v. 51, 8s, 22 set. 2017. Universidade de São Paulo, Agência USP de Gestão da Informação Acadêmica (AGUIA). http://dx.doi.org/10.11606/s1518-8787.2017051007138.

[3] COSTA, Karen Sarmento; TAVARES, Noêmia Urruth Leão; NASCIMENTO JÚNIOR, José Miguel do; MENGUE, Sotero Serrate; ÁLVARES, Juliana; GUERRA JUNIOR, Augusto Afonso; ACURCIO, Francisco de Assis; SOEIRO, Orlando Mario. Pharmaceutical services in primary health care. Revista de Saúde Pública, [S.L.], v. 51, 2s, 22 set. 2017. Universidade de São Paulo, Agência USP de Gestão da Informação Acadêmica (AGUIA). http://dx.doi.org/10.11606/s1518-8787.201705100supl2ap.

[4] COSTA, Karen Sarmento; TAVARES, Noêmia Urruth Leão; NASCIMENTO JÚNIOR, José Miguel do; MENGUE, Sotero Serrate; ÁLVARES, Juliana; GUERRA JUNIOR, Augusto Afonso; ACURCIO, Francisco de Assis; SOEIRO, Orlando Mario. Pharmaceutical services in the primary health care of the Brazilian Unified Health System. Revista de Saúde Pública, [S.L.], v. 51, 3s, 22 set. 2017. Universidade de São Paulo, Agência USP de Gestão da Informação Acadêmica (AGUIA). http://dx.doi.org/10.11606/s1518-8787.2017051007146.

[5] DIAS, Kelli Engler. A implantação do HÓRUS nas farmácias do SUS: uma proposta de ações para auxiliar esse processo. 2013. 119 f. Dissertação (Mestrado Profissional em Saúde Pública) – Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, 2013.

[6] Faraco EB, Rover MRM, Leite SN. Avaliação da capacidade de gestão da assistência farmacêutica na Atenção Primária à Saúde dos municípios brasileiros. Cad Saúde Colet. 2025;33(3):e33030295. https://doi.org/10.1590/ 1414-462X202533030295.

[7] FRANCO, Amanda Rodrigues; ARAÚJO, Igor Gomes de; MORAIS, Arlandia Cristina Lima Nobre de. HÓRUS: ferramenta de gestão para assistência farmacêutica. Cadernos ESP, v. 17, e1457, 2023.

[8] GERLACK, Letícia Farias; KARNIKOWSKI, Margô Gomes de Oliveira; AREDA, Camila Alves; GALATO, Dayani; OLIVEIRA, Aline Gomes de; ÁLVARES, Juliana; LEITE, Silvana Nair; COSTA, Ediná Alves; GUIBU, Ione Aquemi; SOEIRO, Orlando Mario. Management of pharmaceutical services in the Brazilian primary health care. Revista de Saúde Pública, [S.L.], v. 51, 15s, 22 set. 2017. Universidade de São Paulo. Agência de Bibliotecas e Coleções Digitais. http://dx.doi.org/10.11606/s1518-8787.2017051007063.

[9] GOMES, Fagner Dayan de Lima; SANTOS, Marjorie Diony Abrantes Fernandes dos; SERRANO, Rossana Maria Souto Maior. Mudanças na Gestão da Assistência Farmacêutica dos Componentes Especializados no Estado da Paraíba a partir da Implantação do HÓRUS. Cadernos Ibero-Americanos de Direito Sanitário, Brasília, v. 2, n. 2, p. 530-542, jul./dez. 2013.

[10] INÁCIO, Viviane Aparecida Bruno. Avaliação do grau de implementação do Sistema Nacional de Gestão da Assistência Farmacêutica Hórus Indígena. 2018. 73 f. Dissertação (Mestrado Profissional em Saúde Pública) – Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, 2018.

[11] LEITE, Silvana Nair; BERNARDO, Noemia Liege Maria da Cunha; ÁLVARES, Juliana; GUERRA JUNIOR, Augusto Afonso; COSTA, Ediná Alves; ACURCIO, Francisco de Assis; GUIBU, Ione Aquemi; COSTA, Karen Sarmento; KARNIKOWSKI, Margô Gomes de Oliveira; SOEIRO, Orlando Mario. Medicine dispensing service in primary health care of SUS. Revista de Saúde Pública, [S.L.], v. 51, 11s, 22 set. 2017. Universidade de São Paulo, Agência USP de Gestão da Informação Acadêmica (AGUIA). http://dx.doi.org/10.11606/s1518-8787.2017051007121.

[12] MEIRELLES, Claudia da Costa; OKUMURA, Mauro Takanori; SEMMLER, Tatiane Cristina; NAGATOMI, Maristela Marico; CANDIDO, Zilmara Bassos. Gestão da Assistência Farmacêutica: implantação de controle, rastreabilidade e farmacoeconomia. BIS. Boletim do Instituto de Saúde, v. 16, supl., p. 78-82, 2015.

[13] MENDES, Luiz Villarinho; CAMPOS, Mônica Rodrigues; CHAVES, Gabriela Costa; SILVA, Rondinelli Mendes da; FREITAS, Paula da Silva; COSTA, Karen Sarmento; LUIZA, Vera Lucia. Disponibilidade de medicamentos nas unidades básicas de saúde e fatores relacionados: uma abordagem transversal. Saúde em Debate, [S.L.], v. 38, n. , p. 109-123, out. 2014. GN1 Sistemas e Publicações Ltd.. http://dx.doi.org/10.5935/0103-1104.2014s009.

[14] PENAFORTE, Thais Rodrigues. Assistência farmacêutica no Brasil: uma análise antropológica da formulação da política nacional de assistência farmacêutica. Ciência Saúde Coletiva, [S.L.], v. 31, n. 1, p. 1-10, jan. 2026. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1413-81232026311.15032025.

[15] RIBEIRO, Maria Eduarda de Melo; BAIENSE, Alex Sandro Rodrigues; GONÇALVES, João dos Santos; ANDRADE, Leonardo Guimarães de. ASSISTÊNCIA FARMACÊUTICA NO SUS. Revista Ibero-Americana de Humanidades, Ciências e Educação, [S.L.], v. 9, n. 4, p. 1630-1642, 12 maio 2023. Revista Ibero-Americana de Humanidades, Ciências e Educação. http://dx.doi.org/10.51891/rease.v9i5.9469.

[16] SOUZA, Gisélia Santana; COSTA, Ediná Alves; BARROS, Rafael Damasceno de; PEREIRA, Marcelo Tavares; BARRETO, Joslene Lacerda; GUERRA JUNIOR, Augusto Afonso; ACURCIO, Francisco de Assis; GUIBU, Ione Aquemi; ÁLVARES, Juliana; COSTA, Karen Sarmento. Characterization of the institutionalization of pharmaceutical services in Brazilian primary health care. Revista de Saúde Pública, [S.L.], v. 51, 7s, 22 set. 2017. Universidade de São Paulo, Agência USP de Gestão da Informação Acadêmica (AGUIA). http://dx.doi.org/10.11606/s1518-8787.2017051007135.

[17] TORRES, Andréa Sarmento Figueiredo; BRITO, Michele Lima de; COSTA, Bruno Wesley Bezerra; LIMA, Gustavo Campos de; SOLER, Orenzio. A institucionalização da assistência farmacêutica no município de Ananindeua, Pará. Research, Society And Development, [S.L.], v. 10, n. 14, e576101422368, 14 nov. 2021. Research, Society and Development. http://dx.doi.org/10.33448/rsd-v10i14.22368.

[18] TORRES, Andréa Sarmento Figueiredo; SOLER, Orenzio. Institucionalização da assistência farmacêutica no município de Benevides, Pará. Research, Society And Development, [S.L.], v. 12, n. 4, e10212440992, 1 abr. 2023. Research, Society and Development. http://dx.doi.org/10.33448/rsd-v12i4.40992.

Avaliação Comparativa de Modelos de Aprendizado de Máquina para Classificação da Doença de Alzheimer a partir de Sinais de EEG

Jessica I. C. Castro
Laboratório de Inteligência Computacional - CILAB
Universidade Federal Rural do SemiÁrido - UFERSA
Pau dos Ferros, Brasil
jessica.castro09216@alunos.ufersa.edu.br

Bruno V. Paiva-da-Silva
Departamento de Engenharia de Computação e Automação - DCA
Universidade Federal do Rio Grande do Norte - UFRN
Natal, Brasil
bruno.silva.704@ufrn.edu.br

Wagna M. C. M. Gonçalves
Departamento de Ciências Exatas e Naturais - DECEN
Universidade Federal Rural do SemiÁrido - UFERSA
Pau dos Ferros, Brasil
wagna.maquis@ufersa.edu.br

Huliane Medeiros da Silva
Departamento de Engenharia e Tecnologia - DETEC
Universidade Federal Rural do SemiÁrido - UFERSA
Pau dos Ferros, Brasil
huliane@ufersa.edu.br

***Resumo*—O Alzheimer é uma patologia que causa demência na população idosa, ela afeta gradualmente as funções cognitivas tais como memória, linguagem e atenção. Com isso, a doença apresenta um impacto que vai além do cognitivo, com uma alta presença de sintomas neuropsiquiátricos. Análises recentes que aplicaram a biblioteca PyCaret à detecção da enfermidade inclinam-se na utilização de classificadores fundamentados em árvores sem ponderar os riscos de superajuste ao lidar com sinais biológicos densos, como o eletroencefalograma (EEG). Este estudo utiliza a biblioteca PyCaret para evidenciar esse viés metodológico, contrastando classificadores baseados em árvores, como LightGBM e Random Forest, com a Logistic Regression. Foi empregada uma base pública do Kaggle: de sinais de EEG, proveniente de registros clínicos publicados no repositório OpenNeuro (ds004504) possuindo 848.640 amostras e 17 variáveis, possuiu a execução do treinamento e a validação dos protótipos por meio do PyCaret, com divisão de 70% dos dados para treino e 30% para teste, foi utilizada validação cruzada K-Folds (5 folds) e as métricas de acurácia, precisão, sensibilidade, especificidade, F1-score e AUC. Conclui-se que a disparidade entre a divisão aleatória (>99%) e o GroupKFold (78,87% no *Random Forest* e 60,65% no *LightGBM*) confirma a presença de *data leakage* no particionamento aleatório por amostras em sinais de EEG.**

***Palavras-chave*—Alzheimer's disease; machine learning; biomedical signal processing; pycaret.**

I. Introdução

A doença de Alzheimer (DA) é uma patologia neurodegenerativa progressiva de demência, frequentemente presente na população idosa [1]. De acordo com a Organização Mundial da Saúde [2], o número global de pessoas com demência deve triplicar até 2050, e no Brasil já são registrados aproximadamente 1,2 milhão de casos, com cerca de 100 mil novos diagnósticos por ano. Ainda segundo [2] essa patologia é desencadeada pelo processamento proteico anômalo no sistema nervoso central, cujo acúmulo de fragmentos neurotóxicos resulta em perda neuronal progressiva em regiões como o córtex cerebral e o hipocampo, suprimindo o reconhecimento sensorial, a memória e o raciocínio [3].

Diante disso, torna-se necessário o diagnóstico prematuro da DA, para que se permita realizar tratamentos em seus estágios iniciais com o objetivo de retardar o avanço da doença. Com isso, a tecnologia vem se tornando uma das principais ferramentas aliadas no auxílio do diagnóstico.

Os trabalhos referentes mostram diferenças metodológicas importantes que restringem a comparabilidade direta com este estudo, como por exemplo, o trabalho de [4], no qual utiliza abordagens lineares no *PyCaret* operando sobre características espaciais pré-extraídas, como o *eLORETA*, para detectar biomarcadores clínicos. Outro trabalho que utiliza o *PyCaret* é o de [5], que tem como objetivo demonstrar o potencial do *PyCaret* em atingir acurácias superiores a 98% no diagnóstico do Alzheimer através do uso de técnicas de explicabilidade expost, como o SHAP, as abordagens focam majoritariamente em dados de prontuário ou registros tabulares. Por fim, o trabalho de [6], utiliza do PyCaret, porém, para investigar a presença da Doença de *Parkinson*, em seu estágios iniciais, como por exemplo,os autores reportaram melhor desempenho de algoritmos de ensemble não lineares, como Gradient Boosting e Extra Trees, na classificação de atributos extraídos de gravações vocais nesse contexto específico, sem apresentar métrica quantitativa ou detalhes de validação que permitam avaliar a robustez desse resultado.

Embora estudos recentes demonstrem o potencial de algoritmos de aprendizado de máquina na classificação da doença de Alzheimer a partir de dados biomédicos, a avaliação de modelos aplicados a sinais EEG exige cuidados metodológicos específicos. Como uma única gravação pode gerar milhares de pontos temporais ou segmentos correlacionados, a divisão aleatória dessas amostras pode fazer com que dados pro-

venientes do mesmo participante sejam incluídos tanto no treinamento quanto no teste. Esse procedimento pode favorecer a aprendizagem de padrões individuais, características da sessão de aquisição ou ruídos específicos, produzindo estimativas artificialmente elevadas de desempenho. Apesar de sua relevância, o impacto do protocolo de particionamento sobre classificadores de Alzheimer baseados em EEG ainda precisa ser investigado de forma sistemática.

Diante dessa lacuna, este trabalho tem como finalidade examinar o desempenho de classificadores de árvore e lineares para a classificação da Doença de Alzheimer a partir de sinais de EEG. O estudo realiza uma avaliação comparativa aplicando dois protocolos distintos de validação cruzada: a divisão aleatória tradicional por amostras (K-Folds, $K = 5$) e a divisão agrupada por participante ($GroupKFold$, $K = 5$) utilizando o identificador (ID) dos sujeitos. Com isso, busca-se demonstrar e quantificar o impacto do vazamento de dados ($dataleakage$) e da memorização de padrões na detecção da patologia.

II. Materiais e métodos

O desenvolvimento deste trabalho consiste na utilização de uma base de dados de classificação binária de sinais de EEG obtida na plataforma Kaggle (https://www.kaggle.com/datasets/thngdngvn/openneuro-ds004504). A base reúne amostras de EEG organizadas por canal, derivadas dos registros de repouso com olhos fechados coletados no 2° Departamento de Neurologia do Hospital Geral AHEPA (Tessalonica, Grécia), e originalmente publicados no repositório OpenNeuro (ds004504). Inicialmente, a coorte da base de dados possui um total de 88 participantes, divididos entre diagnósticos de Doença de Alzheimer (DA), Controle e Demência Frontotemporal, na a classificação binária deste estudo, o grupo de Demência Frontotemporal foi descartado, deixando apenas os grupos de Alzheimer e Controle. A base de dados que foi utilizada possui um total de 848.640 amostras, sendo 512.000 amostras com a manifestação da doença (classe 1), e 336.640 amostras sem a doença (classe 0 - controle). Outrossim, contêm-se 17 variáveis numéricas, das quais 16 são eletrodos de escalpo (sistema internacional 10-20) e a última é a variável alvo para a classificação binária.

É importante destacar que, embora a base de dados derivada do repositório OpenNeuro tenha uma estrutura original de separação por sujeitos, estudos prévios realizaram o particionamento com foco nas janelas temporais de forma aleatória. Como apontado na literatura, a falta de um agrupamento explícito por identificador de participante tende a impedir o isolamento total de dependências temporais entre o treino e teste, reforçando ainda mais a justificativa para a validação agrupada $GroupKFold$.

Mediante a análise dos dados, foi realizado o treinamento e a validação dos modelos por meio da biblioteca PyCaret (v3.3.2). O PyCaret atua como um $wrapper$ em Python para diversas bibliotecas de aprendizado de máquina, como *scikit-learn* (v1.4.2) [7].

A partir do uso do PyCaret, a avaliação dos modelos foi estruturada sob dois protocolos distintos: (1) no **Protocolo Aleatório**, realizou-se a divisão dos dados em 70% para treinamento e 30% para teste via amostragem aleatória simples de janelas temporais, utilizando validação cruzada K-Fold ($K = 5$) no conjunto de treino; (2) no **Protocolo GroupKFold**, aplicou-se a validação cruzada agrupada $GroupKFold$ ($K = 5$) utilizando o identificador único (ID) dos participantes como variável de agrupamento, garantindo que nenhuma amostra do mesmo indivíduo aparecesse simultaneamente no treino e no teste. Os valores apresentados para o protocolo agrupado consistem na média aritmética das métricas obtidas ao longo dos 5 folds. Além disso, foram avaliados cinco algoritmos de Machine Learning: Random Forest, Decision Tree, Logistic Regression, Naive Bayes e LightGBM (v4.6.0). Teve o seu desempenho computado através de métricas consolidadas na literatura: acurácia, área sob a curva (AUC), sensibilidade (recall), especificidade e F1-score, complementados pela análise da matriz de confusão.

Para garantir a clareza e a reprodutibilidade do estudo, os experimentos foram realizados em uma máquina local de sistema operacional Linux, com processador *AMD Ryzen* 3500U, com uma GPU integrada e 20 GB de memória RAM, utilizando Python 3.10.12, *PyCaret* na versão 3.3.2 e com o uso de alguns algoritmos de ML da literatura. O projeto foi desenvolvido integralmente em linguagem Python. O código-fonte completo e a configuração dos experimentos podem ser acessados no repositório do GitHub em: https://github.com/brunopaiva1/project-alzheimer

III. Resultados e Discussão

Esta seção exibe a avaliação comparativa dos cinco classificadores de aprendizado de máquina (Logistic Regression, Naive Bayes, Decision Tree, LightGBM e Random Forest) que foram submetidos a dois protocolos de validação cruzada ($K = 5$): a divisão aleatória por amostras e a divisão agrupada por participante ($GroupKFold$), foi utilizado o identificador único (ID) dos 88 participantes da base de EEG OpenNeuro (ds004504). O conjunto total de dados contendo 848.640 amostras foi particionado por meio do framework PyCaret (v3.3.2).

A. Avaliação Comparativa: Divisão Aleatória vs. $GroupKFold$

Para apurar a hipótese de superajuste e a dependência de padrões do participante, verificou-se uma rodada de experimentos aplicando a validação cruzada agrupada ($GroupKFlod$, $K = 5$), utilizando o identificador único (ID) de cada um dos participantes do conjunto de dados como variável de agrupamento. Esse procedimento busca impedir que janelas temporais de um mesmo paciente pertemçam simultaneamente aos conjuntos de treinamento e teste, sendo os valores dispostos na Tabela I a média aritmética das métricas calculadas entre os 5 folds.

A Tabela I sintetiza a comparação direta entre o protocolo de divisão aleatória por amostras e o protocolo $GroupKFold$ para todos os modelos avaliados.

Tabela I
COMPARAÇÃO DE DESEMPENHO DOS CLASSIFICADORES SOB DIVISÃO ALEATÓRIA E VALIDAÇÃO AGRUPADA POR PARTICIPANTE (*GroupKFold*).

Modelo	Protocolo	Acur. (%)	AUC (%)	Rec. (%)	Esp. (%)	F1 (%)
Random Forest	Aleatório	99,92	**100,00**	99,95	99,87	99,93
	GroupKFold	78,87	91,16	79,32	78,18	80,80
LightGBM	Aleatório	99,87	**100,00**	99,90	99,82	99,89
	GroupKFold	60,65	78,17	52,05	73,73	58,84
Decision Tree	Aleatório	99,46	99,43	99,54	99,34	**99,55**
	GroupKFold	55,32	59,12	53,55	58,01	56,14
Logistic Reg.	Aleatório	83,78	**90,61**	87,60	77,97	86,70
	GroupKFold	75,06	83,15	80,50	66,78	78,41
Naive Bayes	Aleatório	66,62	78,49	51,29	**89,93**	64,96
	GroupKFold	64,47	73,74	54,68	79,35	63,29

B. Random Forest

O algoritmo *Random Forest*, fundamentado em agregação por bootstrap (*bagging*) e seleção aleatória de subconjuntos de atributos, obteve o maior desempenho absoluto sob divisão aleatória (acurácia de 99,92% e AUC de 100,00%). A matriz de confusão registrou os menores índices de erro (126 falsos positivos e 80 falsos negativos para 100.866 TN e 153.520 TP). Contudo, sob o protocolo $GroupKFold$, o modelo apresentou uma redução para 78,87% de acurácia, 91,16% de AUC e 78,18% de especificidade, apresentando, ainda assim, o melhor desempenho entre os modelos avaliados sob esse protocolo.

C. LightGBM

O algoritmo de boosting de gradiente *LightGBM* (v4.6.0) alcançou acurácia de 99,87% e AUC de 100,00% na divisão aleatória. Sua matriz de confusão destacou 254.259 acertos na diagonal principal. Sob a validação agrupada $GroupKFold$, contudo, o modelo sofreu uma degradação severa, caindo para 60,65% de acurácia, 78,17% de AUC, 52,05% de sensibilidade e 73,73% de especificidade, evidenciando sensibilidade à variação de sujeitos entre os conjuntos.

D. Decision Tree

O modelo não paramétrico *Decision Tree* atingiu acurácia de 99,46% e AUC de 99,43% na divisão aleatória, com matriz de confusão contabilizando 100.401 TN e 152.953 TP. Ao aplicar o protocolo $GroupKFold$, observou-se a a menor acurácia entre os modelos avaliados sob o protocolo *GroupKFold*, com queda para 55,32% de acurácia, 59,12% de AUC e 58,01% de especificidade.

E. Logistic Regression

O modelo de *Logistic Regression* obteve acurácia de 83,78% e AUC de 90,61% na divisão aleatória. Sua matriz de confusão no conjunto de teste revelou 78.971 TN e 134.731 TP. Sob $GroupKFold$, a Logistic Regression apresentou menor variação em relação ao protocolo aleatório, registrando 75,06% de acurácia e 83,15% de AUC, indicando maior estabilidade do modelo linear sob o isolamento de sujeitos.

F. Naive Bayes

O algoritmo registrou acurácia de 66,62% e AUC de 78,49% na divisão aleatória, apresentando elevada taxa de falsos negativos (74.527 amostras) devido à premissa de independência condicional dos canais de EEG. Sob $GroupKFold$, o modelo obteve 64,47% de acurácia, 73,74% de AUC e 79,35% de especificidade.

G. Discussão Crítica do Impacto do Particionamento

A expressiva disparidade de desempenho observada entre a divisão aleatória (99% nos modelos de árvore) e a validação $GroupKFold$ aponta para uma forte vulnerabilidade dos modelos não lineares ao particionamento por amostras. Como sinais de EEG de repouso constituem séries temporais densas, a divisão aleatória tende a colocar instâncias temporais correlacionadas do mesmo indivíduo simultaneamente no treino e no teste. A avaliação agrupada por $GroupKFold$ isolou essas dependências por participante, fornecendo estimativas mais realistas sobre o desempenho preditivo dos algoritmos para dados de novos indivíduos.

H. Limitações Computacionais

Para assegurar a reprodutibilidade, as simulações foram executadas em ambiente local Linux com processador AMD Ryzen 3500U, GPU integrada e 20 GB de memória RAM, utilizando Python 3.10.12. Em função das restrições de hardware e memória enfrentadas durante o processamento de 848.640 instâncias com 17 variáveis, algoritmos computacionalmente intensivos como XGBoost, SVM, KNN e CatBoost tiveram sua execução inviabilizada, optando-se por algoritmos com menor pegada computacional suportados pelo PyCaret.

IV. CONCLUSÃO

Este estudo avaliou comparativamente classificadores baseados em árvore e modelos lineares na predição da Doença de Alzheimer a partir de sinais de EEG, aplicando a biblioteca PyCaret sob dois protocolos de validação cruzada. Demonstrou-se experimentalmente que a divisão aleatória por amostras produz estimativas de desempenho excessivamente otimistas (99% para modelos de árvore) devido ao vazamento de informações entre treino e teste.

A aplicação do protocolo $GroupKFold$ baseado no identificador do participante forneceu uma estimativa mais realista da capacidade de generalização para participantes não observados durante o treinamento, evidenciando que algoritmos como o *LightGBM* sofrem severa degradação em sujeitos inéditos (acurácia reduzida para 60,65% e especificidade para 73,73%). O *Random Forest* apresentou o melhor equilíbrio de generalização sob isolamento do sujeito (78,87% de acurácia e 91,16% de AUC). Conclui-se que a validação agrupada por participante é um requisito metodológico indispensável para a avaliação confiável de sistemas de auxílio ao diagnóstico da Doença de Alzheimer.

Como trabalhos futuros, pretende-se aprimorar o protocolo de testes com o $GroupKFold$. As etapas incluirão a realização de ajuste fino de hiperparâmetros sob a validação

agrupada, a aplicação de técnicas de extração de características espectrais no domínio da frequência antes do agrupamento e a avaliação de algoritmos mais complexos, como SVM e XGBoost, em ambiente computacional de alta performance, visando otimizar a capacidade de generalização para indivíduos inéditos.

REFERÊNCIAS

[1] Knapskog, A. B., Engedal, K., Selbæk, G., & Øksengård, A. R. (2021). Alzheimers sykdom–diagnostikk og behandling. Tidsskrift for Den norske legeforening.

[2] BRASIL. Alzheimer. Brasilia: Ministério da Saúde; Governo Federal, s. d. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/alzheimer

[3] da Silva, A. F., de Oliveira Felix, H. C., dos Santos, G. M., Mafra, S. L. L., dos Santos, A. P. A., da Costa Barbosa, D. S. A., ... & do Valle Santos, W. M. (2024). DOENÇA DE ALZHEIMER: UMA REVISÃO BIBLIOGRÁFICA. RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218, 5(5), e555225-e555225.

[4] Aoki, Y., Takahashi, R., Pascual-Marqui, R. D., Hata, M., Takahashi, S., Ishii, R., ... & Ikeda, M. (2026). PyCaret machine learning library with three preprocessing steps after eLORETA source estimation predicts Alzheimer's disease. NeuroImage: Reports, 6(1), 100317.

[5] Guillén, P., & Frias-Martinez, E. (2025). Automated Machine Learning Approach (AutoML) to Alzheimer's Disease Diagnosis and Prognosis. Applied Artificial Intelligence, 39(1), 2565166.

[6] Silva, D. H. D., Ribeiro, C. T., Souza, L. R. D. S., & Pereira, A. A. (2025). Application of Open-Source, Low-Code Machine-Learning Library in Python to Diagnose Parkinson's Disease Using Voice Signal Features. Brazilian Archives of Biology and Technology, 68, e25230860.

[7] M. Ali, "PyCaret: An open source, low-code machine learning library in Python,"2020. [Online]. Disponível em: https://www.pycaret.org.

[8] Altshuler, L., Suppes, T., Black, D., Nolen, W., Keck, P., Frye, M., ... & Reed, M. (2014). Agresti A (2002) Categorical Data Analysis, 2nd edn. John Wiley & Sons, New York. Allison PD (2000) Multiple imputation for missing data: a cautionary tale. Sociological Meth-ods and Research 28: 301–309. Allison PD (2002) Missing Data. Sage, Newbury Park, CA. Allison PD (2005) Imputation of Categorical Variables with PROC MI. Proceedings of the. Biometrika, 86, 948-955.

[9] Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. Biometrika, 45(3-4), 296-315.

[10] Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. Chapman and Hall/CRC.

[11] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[12] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Avaliação sistemática de simuladores realísticos na área da saúde: desenvolvimento e análise de um protótipo de baixo custo para reprodução de sons de ausculta

T. A. F. Silva [1], R. E. A. Farias[2]

[1] Universidade Federal do ABC, Depto. CECS - Centro de Engenharia e Ciências Sociais, São Bernardo do Campo - SP, Brasil
[2] Faculdade de Ciências Médicas Santa Casa, SISB - Tecnologia de Sistemas Biomédicos, São Paulo - SP, Brasil

Abstract - **Este artigo apresenta uma avaliação sistemática de simuladores realistas na educação em saúde, com foco em dispositivos de ausculta, e descreve o desenvolvimento e a análise de um protótipo de baixo custo para a reprodução de sons cardíacos, pulmonares, abdominais e traqueais. O objetivo complementar é prjetar e criar o protótipo abrange a arquitetura de hardware, o mapeamento dos pontos de ausculta e as tentativas de obtenção de sons clínicos. Os resultados ficam prejudicados no que diz particularmente respeito à contaminação por ruído e aos procedimentos de obtenção de sons. A discussão enfatiza a relevância social de simuladores de baixo custo em cenários com recursos limitados. Conclui-se apontando limitações de validação educacional, melhoria na qualidade do som e integração com plataformas digitais.**

Palavras Chave - ***Simuladores de Ausculta, Protótipo, Clínica Sistemática e Validação Educacional.***

I. INTRODUÇÃO

A simulação em saúde consolidou-se, nas últimas duas décadas, como eixo estruturante da formação em medicina, enfermagem, fisioterapia e outras profissões, sendo reconhecida internacionalmente como Simulation-Based Education (SBE).[3][8] Nessa denominação, SBE refere-se a estratégias pedagógicas que utilizam cenários simulados, manequins, dispositivos digitais e ambientes virtuais para treinar habilidades técnicas e não técnicas em condições controladas, com potencial para aumentar a segurança do paciente e reduzir erros clínicos [4][8]. Relatórios recentes sobre centros de simulação na América Latina, elaborados segundo critérios de qualidade da ASPIRE (Association for Simulated Practice in Healthcare) e da Society for Simulation in Healthcare (SSH), mostram que 84% dos centros respondentes são vinculados a universidades, que 71% são de porte médio (menos de dez instrutores) e que a maior concentração se dá em países como Chile (37,5%), Brasil (18,1%) e México (12,7%), evidenciando a expansão da simulação em ambientes acadêmicos [1]. Estudos de mapeamento no Brasil indicam que SBE vem sendo incorporada em praticamente todas as fases da formação, da graduação à pós-graduação e à educação continuada, embora ainda haja lacunas importantes de pesquisa aplicada que orientem sua implementação em nível nacional [3][7].

No interior desse movimento, simuladores realísticos ocupam um lugar singular. Por simuladores realísticos, entende-se dispositivos que buscam reproduzir, com alto grau de proximidade, condições clínicas complexas, seja por meio de manequins de alta fidelidade, seja por meio de sistemas híbridos que combinam hardware, software e ambientes virtuais [4][8]. A literatura distingue duas dimensões centrais nesse campo: fidelidade e realismo. Fidelidade refere-se ao nível de detalhamento com que o simulador reproduz aspectos fisiológicos, anatômicos e de resposta dinâmica do paciente; já realismo diz respeito à percepção subjetiva de "semelhança com a realidade" por parte do usuário, incluindo fatores como ambientação, narrativa clínica e plausibilidade das respostas simuladas [4][8]. Um simulador pode, portanto, ser de baixa fidelidade em termos tecnológicos e ainda assim produzir forte sensação de realismo, se estiver bem contextualizado em um cenário clínico; inversamente, um dispositivo de alta fidelidade técnica pode ser percebido como pouco realista se for utilizado de forma desarticulada do contexto educacional [4][8].

Entre os diferentes tipos de simuladores, aqueles voltados à ausculta cardiopulmonar e abdominal desempenham papel particular. A ausculta é componente central do exame físico, exigindo que o profissional reconheça padrões sonoros complexos, muitas vezes em ambientes ruidosos, e os associe a hipóteses diagnósticas precisas [4][8]. Simuladores de ausculta - doravante denominados "simuladores de ausculta" - permitem que estudantes pratiquem sistematicamente a identificação de bulhas cardíacas, sopros, sons respiratórios normais e adventícios e ruídos abdominais em manequins, com repetição controlada e feedback imediato. Revisões de literatura e ensaios clínicos indicam que o uso de simuladores de ausculta está associado à melhoria da acurácia na identificação de sons e ao aumento da confiança dos estudantes em sua própria capacidade de exame físico, sobretudo quando integrado a programas estruturados de treinamento [4][8].

No Brasil, a expansão da SBE tem sido acompanhada por esforços de caracterização de centros de simulação, programas e práticas pedagógicas. Um consenso recente global sobre simulação em saúde, envolvendo 50 sociedades em 67 países, reforça que práticas de simulação devem estar alinhadas a padrões de qualidade e a agendas de pesquisa que vinculem diretamente o uso de simuladores a melhorias

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

em desfechos de pacientes e na formação de profissionais [8]. Autores brasileiros destacam que, apesar do crescimento de centros de simulação e da fundação da Sociedade Brasileira de Simulação na Saúde (ABRASSIM), ainda há baixa produção de pesquisa específica sobre SBE no país, o que dificulta o estabelecimento de parâmetros nacionais de qualidade e de estratégias de avaliação robustas [3][7]. Nesse contexto, projetos que desenvolvem e analisam simuladores de ausculta de baixo custo, especialmente quando articulados a revisões sistemáticas ou narrativas da literatura, podem contribuir para preencher a lacuna entre pesquisa e prática, oferecendo subsídios para decisões de investimento e desenho curricular [3].

Do ponto de vista da engenharia biomédica, há uma tensão permanente entre o desejo de construir simuladores de alta fidelidade, com respostas dinâmicas sofisticadas e integração plena a plataformas digitais imersivas, e a necessidade de desenvolver dispositivos financeiramente viáveis para instituições de ensino com recursos limitados [3][4]. Simuladores de alta fidelidade, embora potencialmente poderosos, são caros, exigem manutenção intensiva e podem ser subutilizados se não forem acompanhados de estratégias pedagógicas claras [4][8]. Por outro lado, simuladores de baixa e média fidelidade, construídos com componentes mais simples, podem alcançar impacto significativo se forem bem alinhados a objetivos de aprendizagem, especialmente em habilidades básicas como ausculta, exame físico geral e raciocínio clínico inicial [4]. É nesse espaço de transição entre baixa e média fidelidade, buscando alto realismo pedagógico com custo reduzido, que se insere o protótipo de simulador de ausculta de baixo custo analisado neste artigo.

Finalmente, cabe destacar que a gravação e reprodução de sons de pacientes reais em simuladores levanta questões técnicas e éticas relevantes. Do ponto de vista técnico, estetoscópio digital – definido aqui como dispositivo eletrônico de ausculta que capta sons mediante microfone, os converte em sinais digitais e permite armazenamento e reprodução – enfrenta desafios associados a ruído ambiental, artefatos de movimento e limitações de hardware [4]. Do ponto de vista ético, a coleta de sons de ausculta em pacientes reais exige consentimento informado, proteção de dados e reflexão sobre o uso secundário desses registros em contextos educacionais e tecnológicos, especialmente quando há possibilidade de identificação do paciente por características sonoras ou pela associação a prontuários [8]. A discussão aprofundada desses temas, articulada ao desenvolvimento de protótipos acessíveis, é fundamental para que a engenharia biomédica contribua de forma responsável para a democratização da simulação em saúde.

II. Objetivos

O objetivo principal principal é apresentar uma avaliação sistemática de simuladores realísticos na educação em saúde, focando o subcampo da ausculta, e descrever o desenvolvimento de um protótipo de baixo custo para reprodução de sons de ausculta abdominal, cardíaca, pulmonar e traqueal, construído com componentes eletrônicos comercialmente disponíveis e embarcado em manequins de simulação.

III. Metodologia e materiais

A metodologia adotada combina dois eixos complementares: uma revisão narrativa da literatura sobre simuladores realísticos na educação em saúde, com ênfase em simuladores de ausculta, e um estudo de caso de desenvolvimento e análise de um protótipo de baixo custo para reprodução de sons cardíacos, pulmonares e abdominais embarcado em manequins. A opção por revisão narrativa, em vez de revisão sistemática, decorre da necessidade de integrar diferentes tipos de estudos – ensaios clínicos, estudos observacionais, editoriais e trabalhos de engenharia biomédica – em um quadro analítico abrangente, e de acomodar a descrição qualitativa do protótipo em diálogo com a literatura [3][4].

A. Revisão de literaturas

A seleção dos estudos ocorreu em três etapas. Na primeira, títulos e resumos foram tirados para verificar pertinência temática e adequação aos critérios básicos de inclusão. Na segunda, textos completos foram lidos para confirmar o alinhamento ao foco da revisão e para identificar contribuições específicas em termos de tipos de simuladores, contextos de uso, resultados educacionais e debates conceituais.

Na terceira, procedeu-se à síntese narrativa, organizando os estudos em núcleos temáticos: (a) caracterização de centros de simulação e de SBE em âmbito nacional e regional; (b) simuladores de ausculta cardíaca e pulmonar de alta fidelidade; (c) simuladores de baixa e média fidelidade, com destaque para dispositivos de baixo custo; (d) discussões sobre fidelidade, realismo e qualidade em simulação; (e) desafios técnicos e éticos na gravação de sons de pacientes reais [1][3][4][5][8]. A figura 1 apresenta os pontos de ausculta pesquisados para definição de patologias.

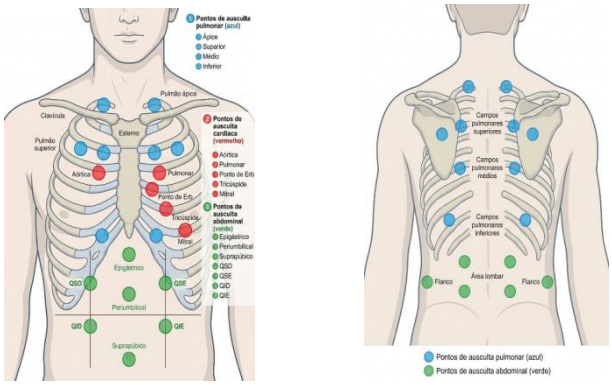


Figura 1 - Pontos de ausculta.

A natureza narrativa da revisão implica que as inferências produzidas têm caráter interpretativo e se baseiam em integração crítica dos estudos, em vez de seguir protocolos de metanálise quantitativa[3][4].

No que diz respeito à caracterização do contexto brasileiro e latino-americano, foram especialmente valorizados estudos que analisam centros de simulação segundo critérios da ASPIRE e da SSH, bem como artigos que discutem a implantação da SBE no Brasil e o papel de associações nacionais como a ABRASSIM [1][3][7]. O trabalho de

caracterização de centros latino-americanos, por exemplo, fornece estatísticas fundamentais: 84% dos centros participantes são universitários, 71% são de porte médio e há distribuição significativa de centros no Brasil (18,1% dos respondentes), o que permite inferir que simuladores realísticos já estão presentes em boa parte dos cursos da área da saúde na região [1]. Um estudo bibliométrico apresentado em reunião da Society for Simulation in Europe, por sua vez, indica que brasileiros são responsáveis por 152 publicações indexadas em simulação em saúde entre 2012 e 2021, sendo a cirurgia o subcampo mais representado (27,3% dos artigos), o que reforça o protagonismo regional e a necessidade de ampliar a produção sobre áreas como ausculta e exame físico [5].

B. Protótipo

O segundo eixo metodológico consiste em um estudo de caso de desenvolvimento de protótipo de simulador de ausculta de baixo custo, cuja concepção foi realizada por conhecimentos da Engenharia Biomédica em parceria com cursos da área da saúde. O dispositivo foi projetado para ser embarcado em manequins de exposição tipo busto e para reproduzir sons de ausculta cardíaca, pulmonar, abdominal e traqueal diretamente na superfície, em pontos pré-definidos. A capacidade de armazenamento é capaz de reproduzir um total de 45 sons, distribuídos em cinco sons abdominais, quinze cardíacos, quinze pulmonares e cinco traqueais, configurados de modo a permitir diferentes cenários clínicos simulados. Cada ponto de ausculta é mapeado em posição anatômica correspondente no manequim - por exemplo, áreas típicas de ausculta das valvas aórtica, pulmonar, tricúspide e mitral, campos pulmonares anteriores e posteriores, quadrantes abdominais e região traqueal cervical – e conectado a mini alto-falantes internos, acionados por lógica de comando central.

Do ponto de vista da engenharia biomédica, o protótipo situa-se, em termos de fidelidade, na transição entre simulador de baixa e média fidelidade. Fidelidade baixa é aqui entendida como dispositivos que reproduzem sons ou sinais fisiológicos de forma estática, sem resposta dinâmica a intervenções do usuário, enquanto fidelidade média se refere a sistemas que, embora não alcancem a sofisticação de manequins de alta fidelidade, incorporam múltiplos pontos de ausculta, diversidade de sons e alguma capacidade de configuração de cenários [4][8]. O núcleo de comando consiste em microcontrolador responsável por gerenciar o acionamento de sons, a seleção de sistemas (cardíaco, pulmonar, abdominal, traqueal), o controle de volume e a interface com usuário por meio de display LCD. As placas de circuito impresso (PCI) foram desenhadas para distribuir sinais aos diferentes alto-falantes internos, conectados a pontos de ausculta através de trilhas e cabos protegidos, garantindo que cada posição anatômica corresponda a um som específico.

O armazenamento dos sons é realizado em placa de memória apropriada, dimensionada para comportar os 45 arquivos de áudio em formato digital, com taxa de amostragem compatível com a reprodução de frequências típicas de sons cardíacos, pulmonares e abdominais.

A figura 2 ilustra o organograma estrutural do projeto.

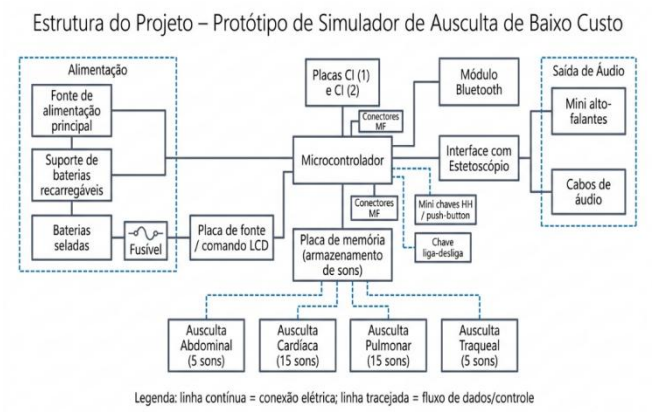
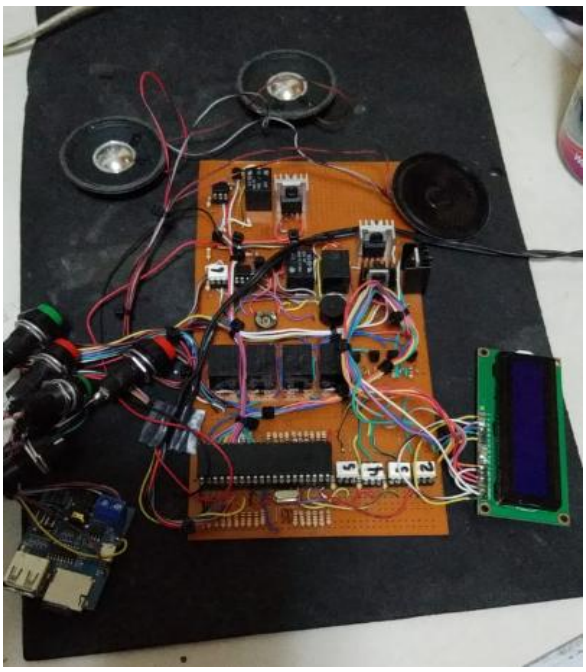


Figura 2 - Organograma da estrutura do projeto

A metodologia de desenvolvimento incluiu etapas de concepção em software de modelagem (CAD), prototipagem física, testes elétricos e acústicos, e ajustes iterativos de posicionamento de alto-falantes e calibração de volume. Foram confeccionados suportes internos para fixação dos alto-falantes em regiões anatômicas dos manequins, com cuidado para evitar vibrações excessivas e para minimizar a propagação indevida de som entre pontos.


Figura 3: Circuito elétrico.

Na concepção do projeto foram considerados a opções de componentes importados e regularmente adquiridos no mercado nacional.

Considerando o objetivo de entrega de um produto de custo reduzido, optou-se para embarcarmos a eletrônica e o sistema de alimentação em um busto manequim plástico adaptado.

O sistema de alimentação de energia utiliza baterias seladas recarregáveis, permitindo uso móvel do simulador em salas de aula e laboratórios sem dependência exclusiva de rede elétrica. Durante os testes, mediu-se temperatura e corrente em pontos críticos da arquitetura para garantir operação segura e evitar aquecimento dos componentes, utilizando multímetros apropriados.

A listagem de material prevê componentes relacionados no Tabela I,

TABELA I - RELAÇÃO DE COMPONENTES POR UNIDADE

ÍTEM	DESCRIÇÃO DO COMPONENTE / EQUIPAMENTO
1	Mini chave HH (Seletora de duas posições)
2	Mini chave push-button (Chave táctil)
3	Chave liga-desliga (Interruptor gangorra comum)
4	Microcontrolador (Arquitetura de mercado de 8 ou 32 bits padrão)
5	Conectores MF (Par de conectores Macho / Fêmea de pinos)
6	Placa de memória (Módulo MicroSD para armazenamento básico)
7	Mini alto-falantes (Pequeno porte para sinalização/áudio básico, 8-9 Ohms
8	Display LCD touch screen (Interface gráfica, tamanho médio 3.5 a 5 16 pol.)
9	Placa de fonte / comando para gerenciamento de display LCD
10	Cabos de áudio blindados (Par de conectores montados por metro)
11	Fusível com caixa (Porta-fusível de proteção com elemento fusível)
12	Fonte de alimentação principal (Estabilizada / Chaveada externa de 12/24V
13	Suporte para baterias recarregáveis (Slot de fixação estrutural)
14	Baterias seladas (Modelo chumbo-ácido regulada por válvula 12V padrão)
15	Componentes e Miscelâneas em geral (Parafusos, espaçadores, fios, solda)
16	Busto manequim (masculino/feminino)
17	Placa de circuito impresso ilhadas (PCBs - parciais)
18	Serviços de mão de obra/fretes/Outros (montagem)

Na Tabela II está relacionado os instrumentos sugeridos e considerados como ideais para realização dos testes e validações dos protótipos.

TABELA II - INSTRUMENTOS PARA TESTES E VALIDAÇÃO

ÍTEM	INSTRUMENTOS SUGERIDOS
1	Multímetro térmico industrial com câmera infravermelha integrada + Kit
2	Licença de software CAD 3D (Assinatura anual corporativa padrão)

IV. RESULTADOS E DISCUSSÃO

Os resultados da revisão narrativa revelam que a simulação em saúde, em suas diferentes modalidades, está amplamente difundida nos cursos da área em âmbito internacional, com crescimento expressivo nas últimas décadas [4][8]. Um levantamento clássico sobre simulação e realidade virtual em educação médica, publicado em periódico brasileiro, indica que, já em meados da década de 2000, uma busca em base internacional com o termo “medical simulation training” retornava mais de quatro mil artigos, com aumento acentuado ao longo dos anos [4]. Esse dado, embora histórico, sinaliza a consolidação da simulação como tema de pesquisa e prática pedagógica em medicina e áreas afins. Mais recentemente, o Global Consensus Statement on Simulation-Based Practice in Healthcare reforça que existem ao menos 50 sociedades de simulação operando em 67 países, apontando para uma infraestrutura global [8].

TABELA III - TABELA COMPARATIVA DE VALORES (05 FABRICANTES - 2026)

Fabricante	Tabela de valores		
	Modelo	Pontos	Preço (R$)
Nasco	Healthcare Auscultation Trainer 2.0 iPad/SmartScope (LF01290)	15 pulmonares 13 cardíacas 5 abdominais	R$ 26.800,00
Sdolf	Simulador c/controle remoto (SD4040)	31 pulmonares 13 cardíacas 0 abdominais	R$ 6.800,00
ProLab	Simulador Avançado de ausculta (SKU-TGD-4025-c)	31 pulmonares 13 cardíacas 0 abdominais	R$ 10.100,00
MedVisioon	Simulador de ausculta MATT	26 pulmonares 26 cardíacas 18 abdominais	R$ 11.000,00
3B Scientific	Simulador de ausculta - SAMBasic - BHS (5022828)	26 pulmonares 37 cardíacas 11 abdominais	R$ 32.000,00
Simulador de "BAIXO CUSTO"	Protótipo	15 pulmonares 15 cardíacas 5 abdominais 5 traqueais	R$ 1.318,00

O desenvolvimento do protótipo foi baseado nas especificações semelhantes aos demais ofertados por outros

fabricantes, guardadas as limitações de custos e modelos. A Tabela III apresenta este comparativo.

No contexto latino-americano, a caracterização de centros de simulação segundo critérios da ASPIRE e da SSH fornece um retrato importante da estrutura e do uso de simuladores realísticos [1]. Entre os centros que responderam ao estudo, 84% são universitários, o que indica forte integração da simulação à formação de graduação e pós-graduação em saúde [1]. A maioria dos centros é de porte médio, com menos de dez instrutores, e há predominância de programas que combinam ensino, pesquisa e educação continuada [1]. Em termos geográficos, Chile, Brasil e México concentram a maior parte dos centros, sugerindo que esses países desempenham papel de liderança regional [1]. Embora o estudo não se concentre exclusivamente em simuladores de ausculta, ele evidencia que recursos realísticos – como manequins de alta fidelidade – fazem parte da infraestrutura desses centros e são usados em treinamento de habilidades clínicas básicas, incluindo exame físico.

No Brasil, artigo recente sobre o avanço da Simulation-Based Education destaca que a simulação foi incorporada à maioria dos cursos de graduação em medicina, enfermagem e outras áreas, bem como a programas de residência e educação continuada [3][7]. Os autores ressaltam, porém, que há déficit de pesquisa nacional robusta que avalie sistematicamente o impacto dos simuladores em desfechos educacionais e clínicos, e que a ABRASSIM, fundada por educadores e pesquisadores em simulação, busca justamente preencher essa lacuna ao incentivar produção de conhecimento aplicada [3][7]. Em particular, poucos estudos brasileiros tratam especificamente de simuladores de ausculta, o que torna relevante a discussão de protótipos de baixo custo nessa área.

Um estudo bibliométrico apresentado em encontro da Society for Simulation in Europe, focado em publicações sobre simulação em saúde na América Latina e Caribe entre 2012 e 2021, indica que a cirurgia é o subcampo mais representado (27,3% dos artigos) e que autores brasileiros respondem por 152 publicações, seguidos por chilenos com 56 [5]. Ainda com resultado, obteve-se o produto final funcionando e ilustrado na figura 5.



Figura 5: Protótipo. Fonte: própria

Como objeto de discussão citamos o fato de a tentativa de efetuar gravações realísticas no Laboratório de Fonoaudioogia da Santa Casa com estetoscópios digitais, porém, não obtivemos resultados audíveis qualitativos com nitidez suficiente para implantarmos nos simuladores para as próximas etapas dos projeto.

V. Conclusões

O protótipo foi concluído e fez parte do fomento de um Projeto do Núcleo de Auxílio a Pesquisa da Faculdade de Ciências Médicas Santa Casa de São Paulo. Esclarecemos que a produção de unidades de montagem de mais unidades para estudos previamente programadas, não ocorreram por razões da não utilização da verba alocada para o projeto e nem termos cronograma de montagem disponível. O produto final entregue ratifica a proposta de produzir um equipamento de baixo custo e com preços finais entre 4% a 20% dos valores dos demais fabricantes.

Referências

[1] GONZÁLEZ, Alberto et al. Characterization of simulation centers and programs in Latin America according to the ASPIRE and SSH quality criteria. Advances in Simulation. Disponível: https://pmc.ncbi.nlm.nih.gov/articles/PMC8588659/. Acesso em: 03 ago. 2026.

[2] FARIAS, Renata; FERREIRA, Maria et al. Clinical simulation strategies for knowledge integration relating to initial critical recognition and management of COVID-19 for use within continuing education and health-related academia in Brazil: a descriptive study. São Paulo Medical Journal. Disponível: https://pmc.ncbi.nlm.nih.gov/articles/PMC9673858/. Acesso em: 03 ago. 2026.

[3] COSTA, Ricardo; SILVA, Ana et al. Advancing Simulation-Based Education in Brazil: Bridging Research and Practice for Healthcare Excellence. Einstein (São Paulo). Disponível: https://journal.einstein.br/article/advancing-simulation-based-education-in-brazil-bridging-research-and-practice-for-healthcare-excellence/. Acesso em: 03 ago. 2026.

[4] PRADO, Cláudio; MASSAD, Eduardo. Medical education: simulation and virtual reality. São Paulo Medical Journal. Disponível: https://www.scielo.br/j/spmj/a/Vyx99cxddwQ9rPhWswHKnNG/?format=pdf&lang=en. Acesso em: 03 ago. 2026.

[5] SOCIETY FOR SIMULATION IN EUROPE. Meeting Abstracts for the Society for Simulation in Europe 2023. Advances in Simulation. Disponível: https://pmc.ncbi.nlm.nih.gov/articles/PMC10905771/. Acesso em: 03 ago. 2026.

[6] GIRARDI, Sônia; CARVALHO, Gilberto. Planning and Sizing of the Health Workforce in Brazil: advances and challenges. Saúde em Debate. Disponível: https://www.scielo.br/j/sdeb/a/KsWVbJJs7pqG7hJrHHZ3WBs/?format=pdf&lang=en. Acesso em: 03 ago. 2026.

[7] ADVANCING SIMULATION-BASED EDUCATION IN BRAZIL. Einstein (São Paulo). Disponível: https://journal.einstein.br/article/advancing-simulation-based-education-in-brazil-bridging-research-and-practice-for-healthcare-excellence/. Acesso em: 03 ago. 2026. [8] GLOBAL CONSENSUS GROUP. Global consensus statement on simulation-based practice in healthcare. BMJ Quality & Safety. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11106913/. Acesso em: 03 ago. 2026.

[9] LIMA, João; OLIVEIRA, Carla et al. Artificial intelligence applied to analyzes during the pandemic: COVID-19 beds occupancy in the state of Rio Grande do Norte, Brazil. Scientific Reports. Disponível: https://pmc.ncbi.nlm.nih.gov/articles/PMC10748397/. Acesso em: 03 ago. 2026.

[10] BARROS, Fernando et al. Cohort Profile: The 100 Million Brazilian Cohort. International Journal of Epidemiology. Disponível: https://academic.oup.com/ije/advance-article-pdf/doi/10.1093/ije/dyab213/42209191/dyab213.pdf. Acesso em: 03 ago. 2026.

Análise comparativa de técnicas de balanceamento na predição de Diabetes

Débora F. Costa*†, Náthalee C. de A. Lima‡, Gabriel C. B. e Sá*
*Computational Intelligence Laboratory (CiLab), Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
†Information Technology, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
‡Departamento de Engenharias e Tecnologia, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
debora.costa22396@alunos.ufersa.edu.br, ana.costa03379@alunos.ufersa.edu.br, fernanda.fernandes36236@alunos.ufersa.edu.br, nathalee.almeida@ufersa.edu.br, gabrielcaldas.sa@gmail.com

***Resumo*—O Diabetes Mellitus é uma das doenças crônicas de maior impacto na saúde pública mundial, tornando essencial o desenvolvimento de métodos que auxiliem sua identificação precoce. Nesse contexto, este trabalho avalia o impacto de diferentes técnicas de balanceamento de classes no desempenho de algoritmos de aprendizado de máquina aplicados à predição de diabetes. Foram comparadas as técnicas SMOTE, ADASYN e RandomOverSampler em conjunto com os classificadores Regressão Logística, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest e Multilayer Perceptron (MLP). Os modelos foram avaliados por meio de validação cruzada estratificada de dez dobras utilizando as métricas de acurácia, precisão, recall, F1-Score e AUC-ROC. Além disso, empregou-se a técnica SHAP para interpretar as decisões do modelo de melhor desempenho. Os resultados demonstraram que as técnicas de balanceamento aumentaram significativamente o recall da classe minoritária, porém ocasionaram redução da precisão e do F1-Score na maioria dos modelos. O Random Forest apresentou o comportamento mais consistente entre os cenários avaliados, enquanto o MLP obteve o melhor desempenho no cenário sem balanceamento. A análise por meio do SHAP evidenciou que o nível de glicose, a hemoglobina glicada (HbA1c) e a idade foram as variáveis de maior influência nas predições do modelo. Os resultados indicaram que, embora as técnicas de balanceamento tenham proporcionado melhorias em algumas métricas, não foram observados ganhos consistentes no desempenho dos classificadores para o conjunto de dados utilizado.**

***Index Terms*—Aprendizado de Máquina, Predição, Diabetes, Balanceamento, SHAP.**

I. Introdução

O Diabetes Mellitus (DM) é uma doença metabólica crônica caracterizada pela elevação dos níveis de glicose no sangue em decorrência da produção insuficiente de insulina ou da incapacidade do organismo de utilizá-la adequadamente. A doença representa um importante problema de saúde pública mundial, estando associada a complicações cardiovasculares, renais, neurológicas e visuais que comprometem significativamente a qualidade de vida dos indivíduos [1], [2].

Segundo a International Diabetes Federation (IDF), aproximadamente 537 milhões de adultos vivem com diabetes em todo o mundo, com projeções indicando um aumento para 643 milhões até 2030 e 783 milhões até 2045 [1]. Nesse contexto, estratégias que possibilitem a identificação precoce da doença tornam-se fundamentais para auxiliar o diagnóstico e reduzir os impactos clínicos associados.

Nos últimos anos, técnicas de Aprendizado de Máquina (*Machine Learning* ML), uma subárea da Inteligência Artificial voltada ao desenvolvimento de algoritmos capazes de identificar padrões complexos nos dados e melhorar seu desempenho com base nas informações disponíveis, e têm sido amplamente empregadas na área da saúde [3]. Essas técnicas possibilitam a análise de grandes volumes de dados e a identificação de padrões relevantes, contribuindo para a predição e o apoio à identificação de diferentes doenças [4], [5]. Nesse contexto, diversos estudos demonstram resultados promissores na predição de diabetes utilizando algoritmos como Regressão Logística (LR), *Support Vector Classifier* (SVC), *K-Nearest Neighbors* (KNN), *Random Forest* (RF) e *Multilayer Perceptron* (MLP) [6], [7], [10].

Entretanto, conjuntos de dados clínicos frequentemente apresentam desbalanceamento entre as classes, situação em que a quantidade de indivíduos pertencentes à classe majoritária é significativamente superior à da classe minoritária. Esse cenário pode comprometer o desempenho dos modelos de classificação, favorecendo a identificação da classe mais frequente e reduzindo a capacidade de detecção dos casos positivos [7], o que pode acarretar em um modelo enviesado.

Para minimizar os efeitos do desbalanceamento, técnicas de balanceamento de dados têm sido amplamente empregadas em tarefas de classificação. Como o *Synthetic Minority Oversampling Technique* (SMOTE) que gera exemplos sintéticos da classe minoritária a partir das relações de vizinhança entre as amostras existentes [8]. De forma semelhante, o *Adaptive Synthetic Sampling* (ADASYN) realiza a geração adaptativa de novas amostras, priorizando regiões do espaço de características nas quais a classificação da classe minoritária é mais difícil [9]. Já o *RandomOverSampler* aumenta a representatividade da classe minoritária por meio da replicação aleatória de suas amostras. Essas abordagens buscam reduzir o viés dos modelos em relação à classe majoritária e ampliar a capacidade de identificação dos casos positivos.

Diante desse contexto, este trabalho tem como objetivo avaliar o impacto de diferentes técnicas de balanceamento de classes no desempenho de algoritmos de ML aplicados à

predição de diabetes. Para isso, foram comparados os métodos SMOTE, ADASYN e *Random Over-Sampling* utilizando os classificadores Regressão Logística, KNN, SVC, Random Forest e Multilayer Perceptron (MLP).

II. Trabalhos Relacionados

No trabalho de [11] foi investigada a aplicação de técnicas de ML para a predição de diabetes utilizando um conjunto de dados contendo informações demográficas, hábitos de vida e condições de saúde autorreferidas, que apresenta desbalanceamento entre as classes. Os autores avaliaram diferentes técnicas de balanceamento, incluindo SMOTE-N, ENN, SMOTE-ENN e SMOTE-Tomek, combinadas com algoritmos como *Logistic Regression*, *Random Forest*, *Gradient Boosting* e *AdaBoost*. Além disso, destacaram a importância de aplicar as técnicas de reamostragem apenas no conjunto de treinamento para evitar vazamento de dados. Os resultados mostraram que todas as técnicas de balanceamento aumentaram o desempenho na identificação da classe minoritária, especialmente em termos de recall. Entre as estratégias avaliadas, o ENN apresentou o melhor desempenho quando combinado ao algoritmo *Gradient Boosting*, evidenciando que o tratamento adequado do desbalanceamento pode melhorar significativamente a capacidade de identificação de pacientes com diabetes.

É apresentado em [12] a proposta de uma estrutura escalável e interpretável para a predição precoce de diabetes utilizando ML. O estudo foi realizado com um conjunto de dados do Kaggle contendo 96.146 registros, após a remoção de duplicatas. Os autores aplicaram o SMOTE para balancear as classes e utilizaram a Eliminação Recursiva de Características (RFE) para selecionar as cinco variáveis mais relevantes: idade, IMC, histórico de tabagismo, HbA1c e glicose no sangue. Também utilizaram o LIME para interpretar as previsões dos modelos. Entre os modelos avaliados, o XGBoost apresentou o melhor desempenho, alcançando 96,8% de acurácia e AUC-ROC de 0,973. Além disso, os autores demonstraram que um modelo utilizando apenas cinco características manteve um desempenho semelhante ao modelo completo, reduzindo o custo computacional e aumentando a interpretabilidade.

Por fim, o estudo [13] apresenta uma abordagem para a predição do risco de diabetes utilizando técnicas de Ml baseadas em métodos de *Bagging, Boosting e Ensemble*. O estudo empregou três conjuntos de dados públicos (CDC Diabetes Health Indicators, Early Stage Diabetes Risk Prediction e PIMA Indian Diabetes), aplicando a técnica SMOTE para tratar o desbalanceamento de classes, seleção de atributos, ajuste de hiperparâmetros por meio do GridSearchCV e validação cruzada de cinco dobras. Entre os modelos avaliados, o CatBoost apresentou o melhor desempenho no conjunto de dados do CDC, alcançando acurácia de 91%, enquanto os métodos de *Ensemble* obtiveram acurácia de 98% no conjunto ESDRP. Os autores concluíram que a combinação das técnicas de balanceamento, seleção de atributos e otimização de hiperparâmetros contribui significativamente para o aumento da robustez e da capacidade preditiva dos modelos de ML aplicados à predição de diabetes.

Os estudos analisados demonstram que técnicas de balanceamento de classes, seleção de atributos, otimização de hiperparâmetros e métodos de ML podem contribuir significativamente para o aumento do desempenho dos modelos de ML na predição de diabetes. Entretanto, cada trabalho concentra-se em aspectos específicos, como a comparação de diferentes técnicas de reamostragem, a redução da dimensionalidade, a seleção de atributos ou o desenvolvimento de modelos de ensemble. Em contraste, o presente trabalho tem como foco comparar o impacto de diferentes técnicas de balanceamento na performance de modelos de ML para a predição de diabetes, bem como explicar os resultados por meio da técnica SHAP, contribuindo para uma análise mais transparente e interpretável.

III. Metodologia

Esta seção detalha os procedimentos metodológicos empregados neste estudo. Primeiramente, são apresentados o conjunto de dados utilizado e as etapas de pré-processamento. Posteriormente, descrevem-se as técnicas de balanceamento de classes, os algoritmos de classificação avaliados, a configuração experimental e as métricas de desempenho dos modelos.

A. Conjunto de Dados

O conjunto de dados empregado neste estudo foi o *Diabetes Prediction Dataset*, disponível na plataforma Kaggle. Esses dados compreendem aproximadamente 100 mil registros e nove atributos, que englobam características clínicas e demográficas dos indivíduos. Tais atributos incluem idade, gênero, hipertensão, doença cardíaca, histórico de tabagismo, índice de massa corporal (IMC), hemoglobina glicada (HbA1c), nível de glicose no sangue e a variável alvo, que indica a presença ou ausência de diabetes.

O conjunto de dados é composto por 100.000 registros, cuja distribuição entre as classes é apresentada na Tabela I. Observa-se que 91,5% das amostras pertencem à classe de indivíduos sem diabetes, enquanto apenas 8,5% correspondem à classe de indivíduos com diabetes, caracterizando um cenário de desbalanceamento de classes. Nesse contexto, a classe majoritária possui cerca de 10,8 vezes mais amostras que a classe minoritária, o que pode levar os modelos de classificação a favorecerem a classe mais frequente e, consequentemente, reduzir a capacidade de identificar corretamente os casos positivos.

Tabela I
Distribuição das classes no conjunto de dados.

Classe	Quantidade	Percentual
Sem diabetes	91.500	91,5%
Com diabetes	8.500	8,5%

B. Pré-processamento dos Dados

Inicialmente, realizou-se a análise exploratória dos dados para identificar inconsistências e verificar a qualidade do conjunto de dados. Foram removidos os registros duplicados e constatou-se a inexistência de valores ausentes.

As variáveis categóricas foram convertidas para o formato numérico utilizando a técnica *Label Encoding*. Em seguida, os atributos numéricos foram padronizados por meio do método *StandardScaler*, com o objetivo de reduzir diferenças de escala entre as variáveis e otimizar o desempenho dos algoritmos de ML.

C. Balanceamento das Classes

Considerando o desbalanceamento entre as classes no conjunto de dados, foram avaliadas três técnicas de *oversampling*, isto é, uma técnica que visa aumentar a quantidade de dados das classes minoritárias [14], e é amplamente utilizada na literatura:

- SMOTE (*Synthetic Minority Oversampling Technique*);
- ADASYN (*Adaptive Synthetic Sampling*);
- RandomOverSampler.

Adicionalmente, um cenário sem a aplicação de técnicas de balanceamento foi considerado, permitindo uma comparação direta entre os resultados obtidos com e sem *oversampling*. As técnicas de balanceamento foram aplicadas exclusivamente aos dados de treinamento, preservando a distribuição original do conjunto de teste para evitar viés na avaliação dos modelos. Essa abordagem garantiu a ausência de vazamento de informações entre os conjuntos de treinamento e teste, resultando em uma avaliação mais confiável dos modelos.

D. Modelos de Classificação

Foram avaliados cinco algoritmos amplamente empregados em tarefas de classificação supervisionada:

- LR;
- KNN;
- SVC;
- RF;
- MLP.

Os hiperparâmetros dos classificadores foram otimizados por meio do algoritmo GridSearchCV, utilizando validação cruzada de cinco dobras (*5-fold cross-validation*) e a métrica F1-score como critério de seleção. Os melhores hiperparâmetros encontrados para cada modelo são apresentados na Tabela II.

Tabela II
MELHORES HIPERPARÂMETROS OBTIDOS POR MEIO DO GRIDSEARCHCV.

Modelo	**Melhores hiperparâmetros**
Random Forest	n_estimators=200; max_depth=20; max_features=sqrt; min_samples_split=5; min_samples_leaf=2
Regressão Logística	C=0.1; penalty=l2; solver=liblinear; max_iter=1000
KNN	n_neighbors=7; weights=uniform; metric=euclidean
SVC	C=10; kernel=rbf; gamma=scale
MLP	hidden_layer_sizes=(50,50); activation=relu; solver=adam; learning_rate_init=0.001; alpha=0.001; max_iter=1000

E. Configuração dos Experimentos

O conjunto de dados foi dividido em treinamento e teste na proporção de 80% e 20%, respectivamente, utilizando amostragem estratificada para preservar a distribuição original das classes. Após a otimização dos melhores hiperparâmetros, os classificadores foram avaliados utilizando validação cruzada estratificada de dez dobras (*10-fold Stratified Cross-Validation*), permitindo uma estimativa mais robusta e confiável do desempenho dos modelos.

A Figura 1 apresenta o fluxo experimental adotado neste estudo, desde a preparação dos dados até a avaliação dos modelos de classificação. Para cada técnica de balanceamento, foi construído um pipeline composto pelas etapas de padronização dos dados, aplicação do método de balanceamento e treinamento do classificador correspondente. Essa estratégia garantiu que o balanceamento fosse realizado apenas nos dados de treinamento, evitando vazamento de informação para o conjunto de teste.

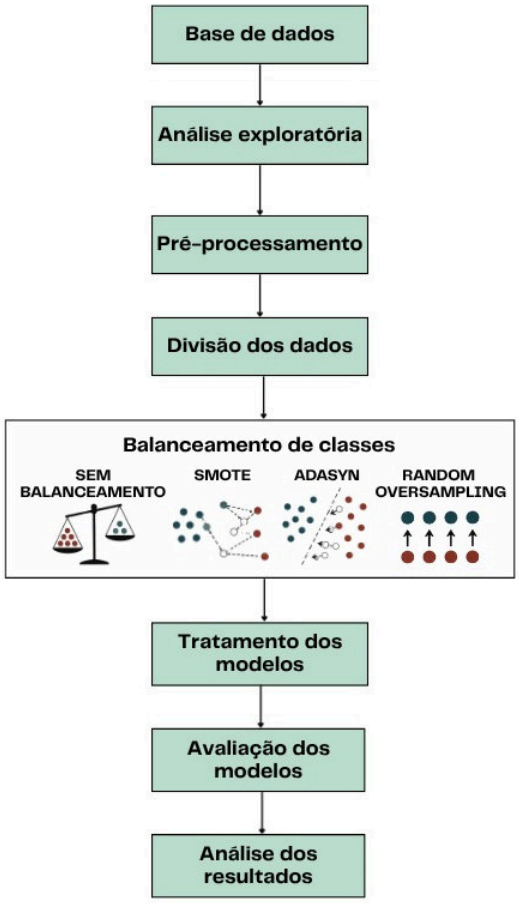


Figura 1. Fluxo experimental da metodologia.

F. Avaliação dos Modelos

O desempenho dos classificadores foi avaliado por meio das métricas de acurácia, precisão, *recall*, F1-Score e área sob a curva ROC (AUC-ROC). Essas métricas foram selecionadas por fornecerem uma avaliação abrangente do desempenho dos modelos, especialmente em cenários de desbalanceamento de classes.

A acurácia mede a proporção total de classificações corretas realizadas pelo modelo. Entretanto, por ser influenciada pela classe majoritária, essa métrica, isoladamente, pode não refletir adequadamente o desempenho em problemas desbalanceados. A precisão indica a proporção de instâncias classificadas como positivas que realmente pertencem à classe positiva, enquanto o *recall* representa a capacidade do modelo de identificar corretamente os indivíduos que possuem diabetes. O F1-Score corresponde à média harmônica entre precisão e *recall*, fornecendo uma medida equilibrada entre ambas. Por fim, a AUC-ROC avalia a capacidade global do modelo em distinguir as classes, independentemente do limiar de decisão adotado.

A análise dos resultados concentrou-se principalmente nas métricas da classe minoritária, uma vez que o objetivo principal é identificar corretamente indivíduos com diabetes, reduzindo a ocorrência de falsos negativos.

G. Explicabilidade do Modelo

A interpretação do modelo de melhor desempenho foi realizada por meio da técnica *SHapley Additive exPlanations* (SHAP), que permite quantificar a contribuição de cada atributo para as predições realizadas pelo modelo. Foram gerados gráficos de importância global das variáveis para identificar os fatores com maior influência na predição de diabetes.

IV. Resultados e Discussão

Esta seção apresenta a comparação entre o desempenho dos modelos nos quatro cenários definidos: sem aplicação de técnicas de balanceamento, com utilização de SMOTE, ADASYN e RandomOverSampler. Como se trata de um problema de predição de diabetes com dados desbalanceados, é importante dar atenção especial à classe minoritária (Classe 1, que é a classe positiva), uma vez que ela representa o principal desafio em problemas de classificação desbalanceada.

Tabela III
Resultados dos modelos sem balanceamento para a classe 1

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1**	**AUC**
Random Forest	0.97	0.96	0.68	0.80	0.97
Regressão Logística	0.95	0.87	0.62	0.72	0.96
KNN	0.96	0.90	0.66	0.76	0.90
SVC	0.96	0.98	0.64	0.77	0.93
MLP	0.97	0.97	0.68	0.80	0.97

A partir dos resultados apresentados na Tabela III, é possível observar que todos os modelos apresentaram acurácia superior a 95% no cenário sem balanceamento para a classe 1. Entretanto, nesse tipo de problema a acurácia isoladamente tende a ser fortemente influenciada pela classe majoritária. Dessa forma, as métricas de precisão, recall e F1-Score da classe minoritária são mais adequadas para a avaliação do desempenho dos modelos.

Os resultados obtidos revelam que o F1-Score dos modelos foi superior sem aplicação de técnicas de balanceamento para a classe minoritária. O modelo MLP apresentou o melhor resultado geral, alcançando F1-Score de 81%, seguido pelo Random Forest com 79% e pelo KNN com 77%. Além disso, a precisão também foi satisfatória para todos os modelos, destacando-se o MLP (98,7%) e o SVC (98,3%). Esses resultados indicam que os modelos foram capazes de identificar corretamente a maioria das instâncias positivas previstas, no entanto, os valores moderados de recall evidenciam uma certa ocorrência de falsos negativos. No caso de cenários médicos, o ideal é minimizar a quantidade de falsos negativos.

Tabela IV
Resultados dos modelos utilizando SMOTE

Modelo	Precisão	Recall	F1	AUC
Random Forest	0.78	0.74	0.76	0.96
Regressão Logística	0.42	0.87	0.57	0.95
KNN	0.49	0.79	0.60	0.90
SVC	0.41	0.88	0.56	0.95
MLP	0.47	0.88	0.61	0.97

A aplicação da técnica SMOTE resultou em um aumento significativo do recall em todos os modelos analisados, conforme apresentado na Tabela IV. Por exemplo, a Regressão Logística aumentou de 63% para 87%, enquanto o SVC passou de 61% para 88%, e o MLP foi de 67% para 88%. Apesar desse ganho na capacidade de detecção da classe minoritária, houve uma redução significativa na precisão e, consequentemente, no F1-Score, indicando que os modelos passaram a ter mais falsos positivos. Esse comportamento é esperado, tendo em vista que geralmente há um trade-off entre a precisão e o recall. Entre os modelos avaliados nesse cenário, o Random Forest apresentou o melhor equilíbrio entre precisão e recall, com F1-Score de 76%, seguido pelo MLP, com 62%.

Tabela V
RESULTADOS DOS MODELOS UTILIZANDO ADASYN

Modelo	Precisão	Recall	F1	AUC
Random Forest	0.71	0.76	0.73	0.96
Regressão Logística	0.31	0.93	0.47	0.95
KNN	0.44	0.82	0.57	0.90
SVC	0.29	0.94	0.45	0.95
MLP	0.38	0.91	0.54	0.97

Os resultados apresentados na Tabela V evidenciam que o balanceamento por ADASYN proporcionou os maiores ganhos de recall entre os métodos avaliados. O recall atingiu 94,6% no SVC, 93,9% na Regressão Logística e 91,9% no MLP, indicando uma elevada capacidade de identificação dos casos positivos de diabetes. Contudo, esse aumento foi acompanhado por uma redução significativa da precisão. O SVC, por exemplo, apresentou queda de precisão de 98,3% para 29,9%, enquanto a Regressão Logística reduziu de 86,7% para 31,5%. Consequentemente, os F1-Scores obtidos com ADASYN foram inferiores aos observados nos demais cenários, indicando que o aumento do número de falsos positivos comprometeu o desempenho global dos classificadores.

Tabela VI
RESULTADOS DOS MODELOS UTILIZANDO RANDOMOVERSAMPLER

Modelo	Precisão	Recall	F1	AUC
Random Forest	0.84	0.71	0.77	0.96
Regressão Logística	0.42	0.87	0.57	0.96
KNN	0.52	0.79	0.63	0.89
SVC	0.42	0.88	0.57	0.96
MLP	0.50	0.85	0.63	0.97

O RandomOverSampler apresentou resultados intermediários entre os métodos avaliados. De forma geral, essa técnica produziu os maiores valores de F1-Score dentre os métodos de balanceamento, superando tanto o SMOTE quanto o ADASYN na maioria dos modelos. O Random Forest teve o maior F1-Score (77,6%), valor próximo ao obtido pelo mesmo modelo sem balanceamento (79%), porém com menos falsos negativos. Resultados semelhantes foram observados para o KNN e para o MLP, cujos F1-Scores foram superiores aos obtidos com SMOTE e ADASYN. Esses resultados sugerem que a simples replicação das amostras da classe minoritária preservou melhor as características originais dos dados quando comparada às abordagens baseadas em geração sintética.

Outro aspecto relevante observado nos experimentos foi o comportamento da métrica AUC-ROC. Apesar das diferenças expressivas entre os valores de precisão, recall e F1-Score, os modelos mantiveram valores elevados de AUC em praticamente todos os cenários avaliados. O Random Forest apresentou AUC-ROC de 0,9646 com SMOTE e 0,9651 com ADASYN, enquanto o MLP alcançou o maior valor observado, com AUC-ROC de 0,9726 utilizando SMOTE. Esses resultados indicam que os classificadores preservaram uma elevada capacidade de distinção entre indivíduos com e sem diabetes, mesmo após as alterações provocadas pelas técnicas de balanceamento.

Ao analisar os diferentes cenários experimentais, observa-se que o Random Forest apresentou o comportamento mais consistente entre os classificadores avaliados. Mesmo após a aplicação das técnicas de balanceamento, o modelo manteve valores elevados de precisão, recall e F1-Score, apresentando menor degradação de desempenho quando comparado aos demais algoritmos. Com SMOTE, o Random Forest obteve F1-Score de 76,4%, enquanto com ADASYN alcançou 73,9% e com RandomOverSampler aproximadamente 77,6%. Esses resultados sugerem maior robustez do modelo frente às alterações na distribuição dos dados. Além de apresentar elevados valores de F1-Score em todos os cenários, o modelo manteve um equilíbrio entre precisão e recall superior ao observado nos demais classificadores.

Os resultados obtidos corroboram estudos da literatura que relatam o aumento do recall após a aplicação de técnicas de *oversampling* em problemas de classificação desbalanceada. Entretanto, também foi observado que o ganho na capacidade de detecção da classe minoritária ocorreu acompanhado de uma redução considerável da precisão, especialmente no cenário com ADASYN.

Esse comportamento evidencia a existência de um compromisso entre a redução de falsos negativos e o aumento de falsos positivos, aspecto que deve ser cuidadosamente considerado em aplicações médicas. Na prática clínica, um aumento no número de falsos positivos pode resultar na realização de exames complementares e em encaminhamentos que, posteriormente, se mostram desnecessários, além de gerar custos adicionais para o sistema de saúde. Por outro lado, em programas de rastreamento do diabetes, esse impacto costuma ser considerado menos preocupante do que a ocorrência de falsos negativos, já que deixar de identificar um paciente com a doença pode atrasar o diagnóstico e o início do tratamento, aumentando o risco de complicações. Dessa forma, a escolha entre priorizar maior sensibilidade ou maior precisão deve levar em conta o objetivo da aplicação do modelo.

De maneira geral, os resultados obtidos indicam que as técnicas de balanceamento não resultaram em melhorias consistentes para o conjunto de dados analisado. Embora tenham aumentado a capacidade dos modelos de identificar exemplos da classe minoritária, esse ganho foi acompanhado por uma redução na precisão, o que impactou negativamente algumas

métricas de desempenho. Esses achados sugerem que, para o conjunto de dados utilizado, o nível de desbalanceamento pode não ter sido suficientemente elevado para justificar a aplicação de técnicas de *oversampling*, uma vez que os classificadores já apresentavam desempenho satisfatório utilizando a distribuição original dos dados.

Tabela VII
MELHORES RESULTADOS OBTIDOS EM CADA CENÁRIO

Cenário	Modelo	F1	Recall
Sem balanceamento	MLP	0.81	0.68
SMOTE	Random Forest	0.76	0.75
ADASYN	Random Forest	0.74	0.76
RandomOverSampler	Random Forest	0.78	0.72

No entanto, como o problema em questão envolve cenários de auxílio à previsão de diagnósticos, é importante tentar reduzir os falsos negativos mantendo um equilíbrio no número de falsos positivos. Desse modo, para situações em que a prioridade seja evitar ao máximo os falsos negativos, cenários com alto recall podem ser mais relevantes. Já quando se busca um equilíbrio entre sensibilidade e precisão, o Random Forest com RandomOverSampler apresenta uma alternativa mais interessante, combinando recall satisfatório com precisão significativamente superior aos modelos com maior sensibilidade.

A. *Análise de Explicabilidade com SHAP*

Com o objetivo de compreender a influência das variáveis nas predições do modelo de melhor desempenho, foi empregada a técnica SHAP. A Figura 2 apresenta a importância global das variáveis, calculada por meio da média dos valores absolutos de SHAP para as duas classes do problema: Classe 0, correspondente aos indivíduos sem diabetes, e Classe 1, correspondente aos indivíduos com diabetes.

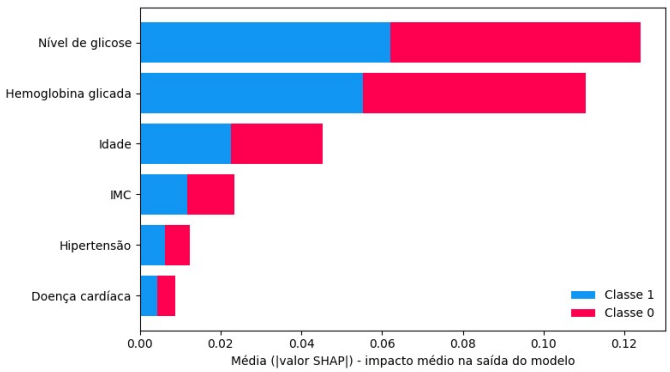


Figura 2. Gráfico SHAP

Conforme apresentado na Figura 2, o nível de glicose no sangue foi a variável com maior impacto nas predições do modelo, seguido pela hemoglobina glicada (HbA1c) e pela idade. Esses atributos apresentaram os maiores valores médios absolutos de SHAP, evidenciando sua elevada contribuição para a classificação dos indivíduos quanto à presença ou ausência de diabetes.

Por outro lado, as variáveis índice de massa corporal (IMC), hipertensão e doença cardíaca apresentaram menor influência nas predições. Embora esses fatores sejam reconhecidos como importantes fatores de risco para o desenvolvimento do diabetes, sua contribuição individual foi inferior à observada para as variáveis diretamente relacionadas ao controle glicêmico, como HbA1c e glicose sanguínea.

Observa-se ainda que nível de glicose no sangue e HbA1c apresentaram elevada contribuição tanto para a classe ”Com diabetes”quanto para a classe ”Sem diabetes”, indicando que o modelo utilizou essas variáveis como principais elementos discriminativos entre as duas classes. Esse comportamento é consistente com a literatura, que reconhece a hemoglobina glicada e a glicose sanguínea como os principais indicadores clínicos para o diagnóstico e acompanhamento do diabetes.

V. CONCLUSÃO

Este trabalho avaliou o impacto de diferentes técnicas de balanceamento de classes no desempenho de algoritmos de ML aplicados à predição de diabetes. Para isso, foram comparados os métodos SMOTE, ADASYN e RandomOverSampler em conjunto com os classificadores Regressão Logística, KNN, SVC, Random Forest e MLP, utilizando validação cruzada estratificada de dez dobras e métricas de classificação adequadas para problemas com desbalanceamento de classes.

Os resultados mostraram que as técnicas de balanceamento aumentaram significativamente a capacidade dos modelos em identificar indivíduos com diabetes, refletida pelo aumento do recall. Entretanto, esse ganho foi acompanhado por uma redução da precisão e, consequentemente, do *F1-Score*, principalmente nos cenários com ADASYN. Ainda assim, em aplicações nas quais a identificação do maior número possível de indivíduos com diabetes seja prioritária em relação à redução de falsos positivos, as técnicas de balanceamento podem representar uma alternativa interessante, uma vez que contribuíram para reduzir significativamente a quantidade de falsos negativos. Entre os modelos avaliados, o Random Forest apresentou o comportamento mais consistente diante das diferentes estratégias de balanceamento, enquanto o MLP obteve o melhor desempenho no cenário sem aplicação de técnicas de *oversampling*.

A utilização da técnica SHAP permitiu interpretar as decisões do modelo, evidenciando que o nível de glicose, a hemoglobina glicada (HbA1c) e a idade foram as variáveis mais relevantes para a predição de diabetes. Esses resultados estão de acordo com o conhecimento clínico disponível na literatura, reforçando a confiabilidade do modelo desenvolvido.

De maneira geral, conclui-se que, para o conjunto de dados utilizado, as técnicas de balanceamento não proporcionaram melhorias no desempenho global dos classificadores, uma vez que o aumento do *recall* ocorreu acompanhado por uma redução significativa da precisão. No entanto, em aplicações médicas nas quais a redução de falsos negativos seja mais importante do que a maximização do desempenho global,

essas técnicas podem representar uma estratégia vantajosa, por favorecerem a identificação de um maior número de indivíduos com diabetes.

Como trabalhos futuros, pretende-se avaliar o desempenho dos modelos em outros conjuntos de dados com diferentes níveis de desbalanceamento, a fim de verificar em quais cenários as técnicas de balanceamento podem proporcionar benefícios. Além disso, pretende-se investigar algoritmos de ML mais recentes, como XGBoost e CatBoost, bem como explorar outras abordagens de explicabilidade, incluindo a técnica LIME, a fim de ampliar a interpretação das predições realizadas pelos modelos.

REFERÊNCIAS

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.

[2] Y. Zhao, X. Wang, and J. Li, "Advances in diabetes mellitus: risk factors, complications and clinical perspectives,"*Frontiers in Endocrinology*, vol. 17, pp. 1–15, 2026.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2022.

[4] N. M. Khalifa and M. A. Albadawy, "Artificial intelligence and machine learning applications in healthcare,"*Healthcare*, vol. 12, no. 4, 2024.

[5] J. Guan et al., "Machine learning in healthcare: Applications and challenges,"*Healthcare Analytics*, vol. 3, p. 100180, 2023.

[6] A. Sharma et al., "Explainable machine learning models for diabetes prediction and risk assessment,"*Healthcare Analytics*, vol. 5, p. 100300, 2024.

[7] X. Liu et al., "Machine learning-based diabetes prediction using BRFSS data and explainable artificial intelligence,"*Scientific Reports*, vol. 14, 2024.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique,"*Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,"*IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.

[10] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Diabetes prediction using SMOTE and SHAP-based explainable machine learning,"*Applied Sciences*, vol. 15, no. 2, 2025.

[11] M. M. Chowdhury, R. S. Ayon, and M. S. Hossain, "An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset,"*Healthcare Analytics*, vol. 5, p. 100297, 2024, doi: 10.1016/j.health.2023.100297.

[12] R. Shetty, A. Paul, A. P. Kumar, and S. Rai, "Early prediction of diabetes using a reduced and interpretable feature set with a scalable machine learning framework,"*Discover Artificial Intelligence*, vol. 6, no. 1, p. 400, 2026, doi: 10.1007/s44163-026-01152-z.

[13] R. Asif, D. Upadhyay, M. Zaman e S. Sampalli, "Enhancing diabetes risk prediction: A comparative evaluation of bagging, boosting, and ensemble classifiers with SMOTE oversampling,"*Informatics in Medicine Unlocked*, vol. 57, p. 101661, 2025, doi: 10.1016/j.imu.2025.101661.

[14] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón e I. García-Rodríguez, Diabetes detection using deep learning techniques with oversampling and feature augmentation, *Computer Methods and Programs in Biomedicine*, vol. 202, p. 105968, 2021. doi: 10.1016/j.cmpb.2021.105968.

Changes in pulmonary mechanics of Endemic Disease Control Agents evaluated by respiratory oscillometry

1st Maria da Penha Laprovita
Institute of Biology and Faculty of Engineering, Biomedical Instrumentation Laboratory
Rio de Janeiro, Brazil
mariadapenhalaprovitao@gmail.com

2nd Cíntia Moraes de Sá Sousa
Institute of Biology and Faculty of Engineering, Biomedical Instrumentation Laboratory
Rio de Janeiro, Brazil
cintiamdesa@gmail.com

3rd Caroline de Oliveira Ribeiro
Institute of Biology and Faculty of Engineering, Biomedical Instrumentation Laboratory
Rio de Janeiro, Brazil
caroline.oliveira@live.com

4th Ariane Leites Larentis
Center for Studies on Workers' Health and Human Ecology, Sergio Arouca National School of Public Health Oswaldo Cruz Foundation
Rio de Janeiro, Brazil
arianelarentis@gmail.com

5th Ana Cristina Simões Rosa
Center for Studies on Workers' Health and Human Ecology, Sergio Arouca National School of Public Health Oswaldo Cruz Foundation
Rio de Janeiro, Brazil
0000-0003-0547-1138

6th Priscila Jeronimo da Silva Rodrigues Vidal
Center for Studies on Workers' Health and Human Ecology, Sergio Arouca National School of Public Health Oswaldo Cruz Foundation
Rio de Janeiro, Brazil
priscilaensp2020@gmail.com

7th Marcos Rogerio da Silva
Projeto Integrador Multicêntrico Center for Studies on Workers' Health and Human Ecology, Sergio Arouca National School of Public Health Oswaldo Cruz Foundation
Rio de Janeiro, Brazil
saude_marcos@yahoo.com.br

8th Pedro Lopes de Melo
Institute of Biology and Faculty of Engineering, Biomedical Instrumentation Laboratory
Rio de Janeiro, Brazil
plopeslib@gmail.com

***Abstract* —** Exposure to pesticides in the work process represents a significant health risk for Endemic Disease Control Agents (EDCAs), potentially leading to alterations in the respiratory system. The present study aimed to analyze respiratory mechanics and pulmonary function in these workers through an integrated assessment, with an emphasis on respiratory oscillometry. This was a cross-sectional, analytical study including 30 individuals, comprising 15 endemic disease control agents and 15 healthy controls. Participants underwent spirometry, body plethysmography, and multifrequency oscillometry (4–32 Hz). Spirometric results demonstrated significant reductions in forced vital capacity (FVC, $p<0.05$) and forced expiratory volume in one second (FEV_1, $p<0.05$) in the EDCA group. Plethysmography revealed a significant reduction in vital capacity (VC, $p<0.01$) and a significant increase in the RV/TLC ratio ($p<0.05$), indicating air trapping. In oscillometric analysis, there was a significant increase in total resistance (R4, $p<0.05$) and central resistance (R20, $p<0.01$), along with a significant reduction in mean reactance (Xm, $p<0.05$) and dynamic compliance (Cdyn, $p<0.05$). Additionally, a significant increase was observed in resonant frequency (fr, $p<0.05$), the area under the reactance curve (Ax, $p<0.01$), and impedance modulus at 4 Hz (Z4, $p<0.05$). These findings contribute to the characterization of respiratory mechanical alterations in this population and are consistent with the underlying pathophysiology. It can be concluded, that respiratory oscillometry is a sensitive tool for identifying alterations in pulmonary mechanics among endemic disease control agents, thereby contributing to improved health monitoring in this workforce.

Keywords: Pesticides; Endemic Disease Control Agents; Oscillometry; Plethysmography; Exposure in the work process.

I. INTRODUCTION

Arboviral diseases, viral pathologies transmitted by arthropod vectors, have become one of the most complex and persistent challenges for global public health in the 21st century [1]. In the Brazilian context, the severity of this scenario is amplified by a combination of rapid and unplanned urbanization, structural deficiencies in basic sanitation, and intense population mobility. These factors, together with anthropogenic climate change, create an ideal environment for the proliferation and spread of the mosquito Aedes aegypti, the biological vector of etiological agents with a high impact on the population [2].

The city of Rio de Janeiro, in particular, has historically stood out as an epicenter of multiple and severe epidemics. Its status as an international tourist destination and its role as a strategic hub within national road and air transport networks facilitate the introduction of exotic viral variants and their subsequent spread into inland regions [3]. Over the past four decades, the state has experienced both endemic and epidemic circulation of Dengue, Zika, Chikungunya, and Yellow Fever, placing a continuous and substantial burden on the Unified Health System (SUS) and requiring uninterrupted, high-cost vector control strategies [3].

The dynamics of these transmission patterns have undergone profound changes due to global warming. Increases in average temperatures and alterations in rainfall patterns accelerate the vector's metabolism and reproductive cycle, expanding its geographic range into previously unaffected areas where viral circulation had not occurred [4]. This phenomenon results in the loss of traditional seasonality, transforming diseases once restricted to warmer periods into year-round threats, thereby requiring continuous surveillance efforts [4]. In response to the spread of arboviral diseases, the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Pan American Health Organization (PAHO) and the Brazilian Ministry of Health recommend integrated strategies that include the extensive use of chemical compounds for vector population control in urban areas [2]. The implementation of these actions, such as fogging/spraying and insecticide application, is governed by technical regulations that establish operational guidelines and health protection measures for workers involved in these activities [2].

In this context, Endemic Disease Control Agents (EDCAs) emerge as the frontline workforce responsible for implementing these public health policies [2]. However, the social relevance of this group contrasts with the chronic precariousness of their working conditions [5]. Historically referred to as "mosquito control workers," these professionals perform their duties under direct exposure to neurotoxic pesticides, such as temephos and other organophosphates and pyrethroids [5, 6]. This exposure often occurs in the absence of adequate provision of personal protective equipment (PPE), without rigorous biological monitoring, and in the context of insufficient and outdated training regarding the toxicokinetics of these compounds [6].

Robust scientific evidence indicates that continuous inhalation and dermal absorption of these toxic agents result in systemic effects [5]. Multicenter studies led by FIOCRUZ in the state of Rio de Janeiro have identified alarming correlations between exposure in the work process and the development of hematological, renal, neurological, and cardiovascular disorders, with the latter affecting up to 46% of the studied population [7]. Within the respiratory system, the prevalence of pathological conditions reaches approximately 19%, suggesting that the inhaled toxic burden compromises airway integrity and pulmonary mechanics in these workers [7, 8].

Considering the pressing need to identify possible respiratory dysfunctions resulting from this exposure, the present study proposes the use of advanced, noninvasive diagnostic methods. In this context, respiratory oscillometry emerges as a highly sensitive tool for assessing respiratory mechanics during spontaneous breathing, enabling the detection of alterations in the resistive and reactive properties of the respiratory system. This technique has been consolidated as a state-of-the-art approach in current clinical practice, allowing the identification of distinct pathophysiological parameters that are complementary to those obtained through conventional tests, such as spirometry and plethysmography [9].

In light of the above, the aim of this study was to analyze respiratory mechanics and pulmonary function through a multimodal approach, integrating respiratory oscillometry, spirometry, and body plethysmography. This assessment sought to identify potential subclinical functional alterations in EDCAs chronically exposed to pesticides, thereby characterizing the possible impact of this toxic burden on pulmonary mechanics.

II. Material and Methods

A. Volunteers

This is a cross-sectional, analytical, and descriptive study. The sample comprised 30 volunteers aged over 18 years, divided into two groups: a control group (CG), consisting of 15 healthy individuals with no history of respiratory or cardiac disease, non-smokers, and without prior COVID-19 or post-COVID-19 sequelae; and an exposed group, consisting of 15 Endemic Disease Control Agents (EDCAs), who were invited to participate through telephone contact in collaboration with Center for Studies on Workers' Health and Human Ecology (CESTEH/ENSP/Fiocruz). Important characteristics of the patient group are observed in Table 1.

Variable	EDCA group (n=15)
Duration of work (years)	26.1 ± 8.5
Duration of direct exposure (years)	24.8 ± 8.3
Primary type of insecticide used	Organophosphates/Pyrethroids
Average frequency of application	5 days/week
Smoking history (pack-years)	9.0 ± 13.8

B. Ethics

All participants provided written informed consent, and the study protocol was approved by the Research Ethics Committee of the Pedro Ernesto University Hospital (HUPE), under CAAE number: 86702825.0.0000.5259.

C. Assessment protocol

Data collection was conducted at the Biomedical Instrumentation Laboratory of the State University of Rio de Janeiro (LIB/UERJ). Participants initially underwent a clinical interview, followed by anthropometric measurements and instructions regarding the procedures for the examinations.

D. Spirometry

All procedures were performed by trained professionals using pre-calibrated equipment. A pneumotachograph-based system was used for spirometric analysis (nSpire Health, Inc., 1830 Lefthand Circle, Longmont, CO 80501). Participants were instructed on the correct execution of respiratory maneuvers, according to the protocols and acceptability and reproducibility criteria established by national guidelines [10]. Forced vital capacity (FVC), forced expiratory volume in the first second (FEV_1), FEV_1/FVC ratio, and forced expiratory flow (FEF_{25-75}%) were analyzed.

E. Plethysmography

For plethysmography examinations, the same equipment used for spirometry was employed, and the examination protocol followed the international guidelines of the ATS [11]. Total Lung Capacity (TLC), Vital Capacity (VC), Residual Volume (RV), and the Ratio between Residual Volume and Total Lung Capacity (RV/TLC) were analyzed. The predicted result for the disease group was used for the statistical test; results from the control group were not included in the comparison.

F. Respiratory oscillometry technique

Respiratory mechanics were further assessed using the respiratory oscillometry technique, with multifrequency analysis ranging from 4 to 32 Hz, using a device developed at the LIB/UERJ, and the results were analyzed using Oscilab 3.0 software [12].

During the examination, participants were instructed to breathe at resting tidal volume while supporting their cheeks

CAPES; CNPq; FAPERJ.

with their hands and wearing a nose clip, in accordance with international standardization guidelines [13]. Respiratory impedance parameters were obtained, including resistance parameters, which assess opposition to airflow in the airways, and reactance parameters, which describe the elastic and inertial properties of the respiratory system.

The resistive parameters analyzed were resistance at 4 Hz (R4), resistance at 20 Hz (R20), and the difference between R4 and R20 (R4–R20). Reactive parameters included mean reactance (Xm), dynamic compliance (Cdyn), resonant frequency (fr), the area under the reactance curve (Ax), and the impedance modulus at 4 Hz (Z4).

All measurements obtained via respiratory oscillometry undergo a quality control procedure based on the coherence function, which allows for the quantification of respiratory variability during measurement. Consequently, only measurements with values above 0.90 are considered valid, while those failing to meet this criterion are discarded.

G. Statistical analysis

Statistical results are reported as mean ± standard deviation (SD) and are presented in tables and box-and-whisker plots. Data analysis was performed using ORIGIN 8.0 software (Microcal Software Inc., Northampton, MA, USA). The distribution of the data was initially assessed using the Shapiro–Wilk test. The statistical analysis employed was the independent samples t-test or the Mann-Whitney U test when comparing two different groups, and the paired t-test or Wilcoxon test when comparing measurements within the same group, as in the case of whole-body plethysmography. A p-value < 0.05 was considered indicative of statistical significance.

Given that this study is exploratory and preliminary in nature, involving a sample of only 15 individuals per group, applying statistical correlation and ROC curve analyses to such a small sample at this stage could yield unstable and unreliable estimates, complicating the interpretation of results; for this reason, we opted not to perform these analyses in the present study, although we recognize that such approaches are relevant and important for future work.

III. RESULTS

Table 1 presents the anthropometric and spirometric data of the studied groups. No statistically significant differences were observed between the EDCA and control groups regarding age, weight, or height; only body mass index (BMI) showed a significant increase ($p < 0.05$).

In the assessment of pulmonary function by spirometry, the EDCA group showed statistically significant reductions in FVC (L) ($p < 0.01$), FVC (%) ($p < 0.05$), FEV_1 (L) ($p < 0.01$), FEV_1 (%) ($p < 0.01$), the FEV_1/FVC (L) ($p < 0.05$), and FEV_1/FVC (%) ($p < 0.0001$). In contrast, $FEF_{25–75}$% did not show a statistically significant difference between groups (p = ns).

Regarding exposure history and lifestyle habits, the exposed group presented a mean duration of direct exposure to pesticides of 26.13 ± 8.49 years ($p < 0.0001$). Concerning smoking status, a mean smoking burden of 9.05 ± 13.79 pack-years was observed in the EDCA group ($p < 0.01$).

Table 1 - Anthropometric and spirometric measurements of the studied groups.

	Control (n= 15)	EDCA (n= 15)	p
Age (years)	59.04 ± 4.86	59.06 ± 8.99	ns
Weight (kg)	80.98 ± 8.80	88.16 ± 15.61	ns
Height (cm)	169.73 ± 5.84	165.13± 6.46	ns
BMI (kg/m2)	27.97 ± 3.43	31.88 ± 4.97	<0.05
Gender (F/M)	5/10	7/8	-
Smoking Load (pack/years)	0	9.05 ± 13.79	-
Exposure period (years)	0	26.13 ± 8.49	-
FEV_1 (L)	3.07 ± 0.44	2.52 ± 0.44	<0.01
FEV_1 (%)	98.54 ± 13.89	84.26 ± 14.37	<0.01
FVC (L)	3.87 ± 0.54	3.19 ± 0.56	<0.01
FVC (%)	98.74 ± 13.48	85.50 ± 15.73	<0.05
FEV_1/FVC (L)	86.32 ± 12.11	78.84 ± 5.54	<0.05
FEV_1/FVC (%)	100.25 ± 7.12	99.01 ± 7.25	<0.0001
FEF_{25-75}(L)	3.14 ± 0.88	2.83 ± 1.17	ns

Legend: BMI - Body Mass Index; FEV_1 - Forced Expiratory Volume in one second; FVC - Forced Vital Capacity; FEV_1/FVC - Forced Expiratory Volume in one second to Forced Vital Capacity ratio; FEF - Forced Expiratory Flow; ns - not significant;

Among the 15 volunteers in the EDCA group, nine were classified as normal, four presented a nonspecific ventilatory pattern, one exhibited a mixed disorder, and one demonstrated an obstructive disorder.

Table 2 presents the plethysmographic parameters and their comparison with predicted values for the EDCA group. A patient was unable to undergo plethysmography due to difficulties understanding how to correctly perform the maneuvers. The results showed a significant increase in RV/TLC ($p < 0.05$) and a significant reduction in VC ($p < 0.001$). Residual volume values were higher in the exposed group compared with predicted values; however, this difference did not reach statistical significance (p = ns). Similarly, TLC values were visually lower, but without a statistically significant difference (p = ns).

Table 2 – Plethysmography results and predicted values of the studied groups.

	Predicted (n=14)	EDCA (n=14)	p
TLC (L)	5.63 ± 0.91	5.33 ± 0.93	ns
VC (L)	3.84 ± 0.53	3.26 ± 0.66	<0.001
RV (L)	1.97 ± 0.32	2.17 ± 0.58	ns
RV/ TLC	33.87± 3.02	40.34± 7.63	<0.05

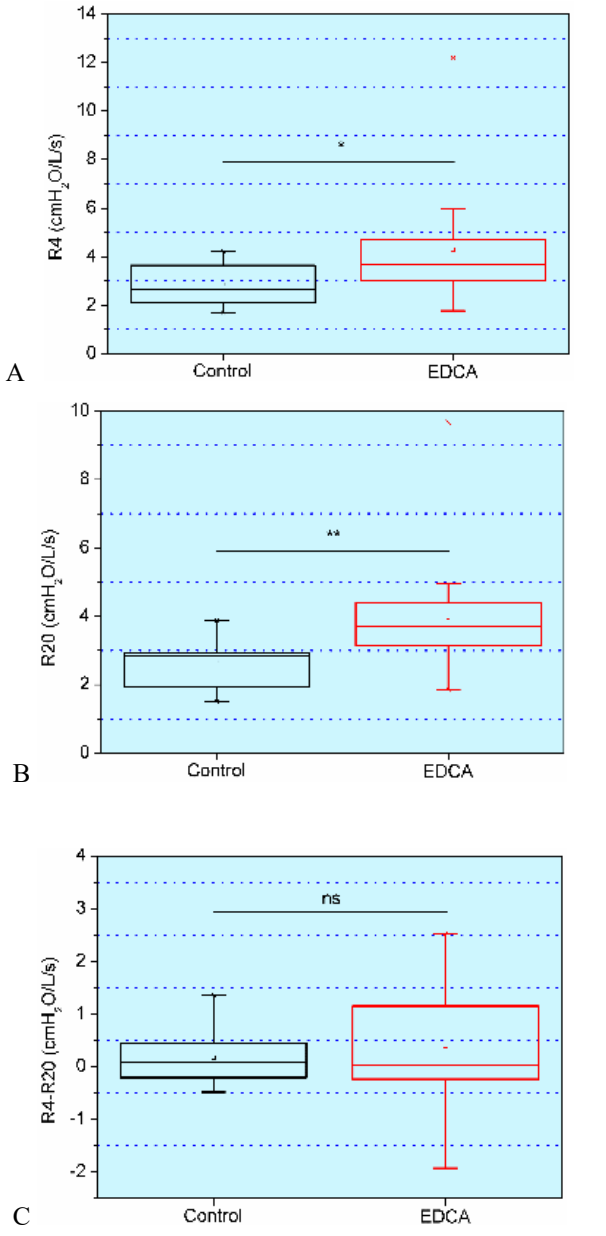
Legend: TLC - Total Lung Capacity; VC – Vital Capacity; RV - Residual Volume; RV/TLC - Ratio between Residual Volume and Total Lung Capacity; n = number of patients evaluated; ns - not significant

Figure 1 presents the resistive results obtained through respiratory oscillometry. A statistically significant increase was observed in R4 ($p < 0.05$; Fig. 1A) and in R20 ($p < 0.01$; Fig. 1B). R4-R20 (Fig. 1C) did not show a statistically significant difference between groups (p = ns).

Figure 2 shows the results related to reactive properties; the EDCAs group showed a significant reduction in Xm ($p < 0.05$; Fig. 2A) and Cdyn ($p < 0.05$; Fig. 2B). In addition, there was an increase in fr ($p < 0.05$; Fig. 2C) and in Ax ($p < 0.01$; Fig. 2D).

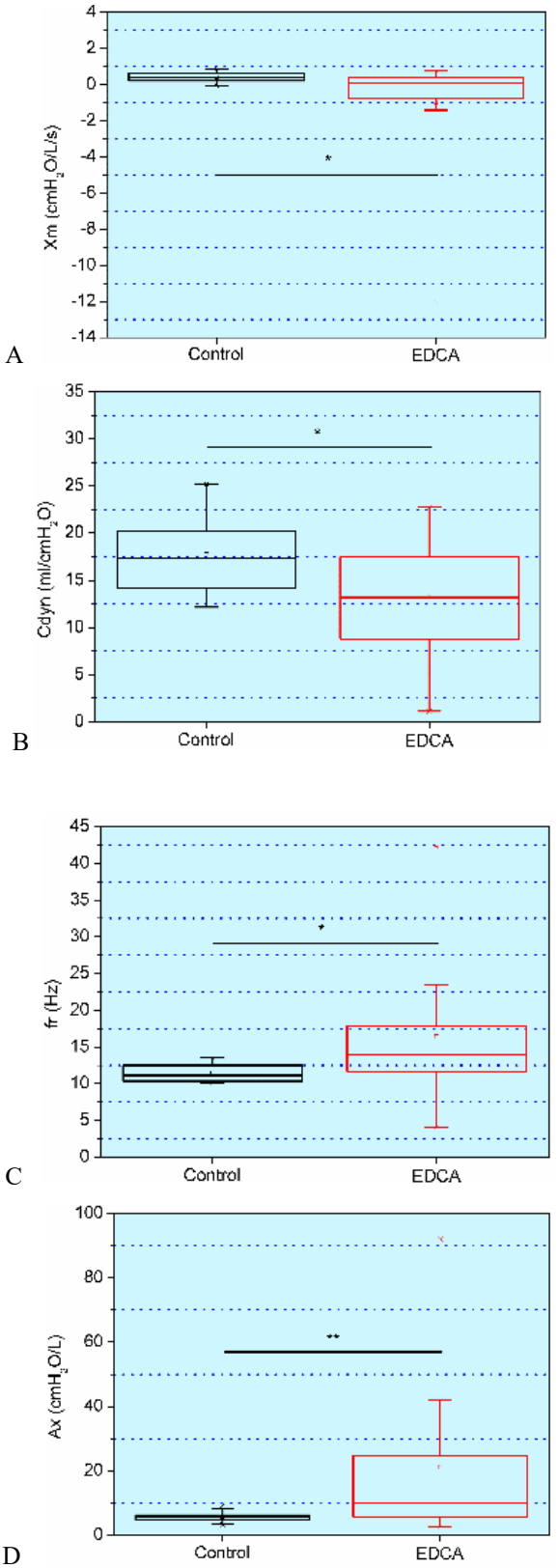
Figure 3 shows that Z4 values (Fig. 3A) were significantly higher in the exposed group compared to the control group ($p < 0.05$).

Figure 1 – Comparative analysis of resistive parameters in oscillometry



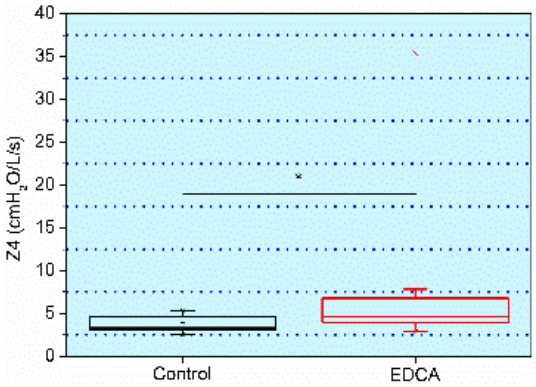
Legend: R4 – total resistance; R20 – central airway resistance; R4-R20 – difference between R4-R20;

Figure 2 – Comparative analysis of reactive parameters in oscillometry



Legend: Xm – mean reactance; Cdyn – dynamic compliance; fr – resonance frequency; Ax – area under the reactance curve.

Figure 3 – Comparative analysis of the impedance modulus in oscillometry



Legend: Z4 – impedance modulus at 4Hz

IV. DICUSSION

In the description of the anthropometric results (Table 1), no statistically significant differences were observed between the EDCA and control groups with respect to age, body weight, or height, indicating group homogeneity. The EDCA group presented a higher BMI ($p<0.05$). According to the Brazilian Thoracic Society (SBPT) guidelines [10], small variations in BMI within this range do not exert a significant impact on pulmonary volume reduction. Therefore, the similarity in the participants' physical profiles supports the interpretation that the reductions observed in FVC and FEV_1 are not attributable to external mechanical influences related to body weight, but rather to exposure in the work process. These findings are consistent with those reported by Hansen et al. whose prospective study of farmers exposed to organophosphates and carbamates [14]. These results are consistent with an increase in airway obstruction.

The absence of a statistically significant difference in $FEF_{25–75}$% is in line with the findings of Quanjer et al. [15]. These authors pointed out that this parameter presents a high variability, consequently, limited utility as an isolated marker, and the presence of functional impairment is necessary to provide a secure diagnosis. This finding suggests that spirometry alone may underestimate functional alterations in exposed populations.

Regarding lifestyle factors, the mean smoking exposure of 9.05 pack-years in the EDCA group is classified as low. According to the literature [16], in individuals with prolonged exposure to chemical agents, the impact of a low smoking burden is considered secondary when compared with the potential inhalational damage associated with exposure in the work process.

The mean duration of exposure in this population was 26 years of direct contact with pesticides, emerging as the primary determinant of the observed alterations in respiratory mechanics and characterizing a condition of cumulative toxicity that outweighs the effects of low-level smoking exposure.

Body plethysmography (Table 2) enabled a more detailed characterization of the respiratory pattern in this population, revealing the presence of air trapping, as indicated by the increased RV/TLC ratio. According to Beachey [17], this phenomenon is associated with small airway closure. Furthermore, exposure to chemical agents with potential for distal deposition may induce bronchiolar inflammation and impair lung emptying, thereby contributing to increased residual volume [18]. Complementarily, Spiro et al. [19], describe that low-solubility chemical agents are capable of penetrating deeply into the respiratory system, promoting bronchiolar inflammation and impairing lung emptying, which may account for the findings observed in the present study.

The observed increase R4 and R20 (Figures 1A and 1B, respectively) indicates a higher resistive load during breathing, suggesting possible obstruction of more central airways and, consequently, an increase in total respiratory system resistance. This finding may be partially attributed to exposure to chemical agents, as described by Hendrick et al. [20], who reported mechanical impairment affecting the conducting airways even at stages in which workers may not yet present evident clinical symptoms.

These findings are in agreement with the spirometric and plethysmographic results (Tables 1 and 2), which demonstrated significant reductions in FVC and FEV_1, as well as decreased VC and an increased RV/TLC ratio, indicative of ventilatory impairment and air trapping. Although $FEF_{25–75}$% showed only a non-significant downward trend, oscillometric analysis was able to detect alterations suggestive of peripheral airway involvement, highlighting its greater sensitivity in identifying early or subtle functional changes not captured by conventional tests.

According to Cavalcante et al. [21] and Libu et al. [22], respiratory oscillometry demonstrates greater sensitivity in detecting early functional alterations that may be underestimated by spirometry alone. Hansen et al. [14], described that chronic exposure to neurotoxic agents induces sustained cholinergic activity, resulting in bronchoconstriction and mucus hypersecretion throughout the tracheobronchial tree, thereby characterizing a state of cumulative toxicity and underdiagnosed pulmonary impairment. Taken together, these two propositions may support findings of Figures 1A and 1B.

Conversely, although a trend toward increased R4–R20 values was observed in the EDCA group (Figure 1C), no statistically significant difference was found between groups (p = ns). These findings, together with the observation in Figures 1A and 1B, suggests that the main changes in EDCAs may happens in more central airway. This finding may also be attributed to a lack of sensitivity of this parameter to the peripheral changes in the EDCA group. The presence of the peripheral airway's involvement and potential alterations in the ventilatory homogeneity should be interpreted in conjunction with reactance data.

The EDCA group exhibited significant reductions in Xm (Figure 2A) and Cdyn (Figure 2B). These findings indicate decreased ventilatory homogeneity, which may be attributed to small airway involvement, where deposition of chemical agents used in vector control activities may occur. Such alterations impair the mechanical performance of the respiratory system, leading to an uneven distribution of ventilation. According to Kaminski et al. [23], these parameters are sensitive markers of distal airway dysfunction, often not detected by conventional methods. Additionally, studies on interstitial lung diseases have reported similar alterations associated with pulmonary fibrosis, which may lead to bronchostenosis, changes in elastic recoil, and

consequently air trapping, findings consistent with those observed in the present population [24, 25].

The alterations observed in reactive properties, particularly the significant increases in fr (Figure 2C) and the Ax (Figure 2D), support the hypothesis that chemical exposure has affected the small airways, resulting in increased pulmonary stiffness and heterogeneous air distribution. These findings are consistent with the plethysmographic evidence of air trapping observed in this population, as indicated by the elevated RV/TLC ratio (Table 2).

This finding is further strengthened when compared with the studies by Kosak et. al [26] and Moitra et. al [27], which highlight Ax as a highly sensitive marker for detecting peripheral disfunction and loss of elasticity, often outperforming resistance parameters in assessing the severity of obstructive diseases and exposures in the work process.

Taken together, the changes in reactive parameters described in Figure 2 provide clear evidence of abnormalities in peripheral airways and ventilatory homogeneity. Therefore, the absence of significant changes in R4-R20 (Figure 1C) may also be attributed to a lack of sensitivity of this parameter to the peripheral changes in the EDCAs group. It can be speculated that the nature of these abnormal changes in EDCAs are not associated with increased obstruction, but related with elastic properties.

The significant increase in Z4 (Figure 3A) reflects a greater overall opposition to airflow, indicating an increased work of breathing in this population. This parameter is related to the total mechanical load of the respiratory system and may therefore be associated with fatigue and dyspnea, symptoms commonly reported by some volunteers in the EDCAs group [28]. Similar findings have been previously described in studies involving respiratory diseases due to exposure in the work process [24].

Multifrequency respiratory oscillometry analysis enabled the identification of subclinical alterations in respiratory mechanics that were not detected by spirometry alone. These findings demonstrate the sensitivity of respiratory oscillometry in assessing pulmonary mechanics, contributing to a more comprehensive characterization of the alterations observed in EDCAs. These results are consistent with that observed by Parker et al. investigating the short-term ventilatory effects in workers exposed to fumes containing zinc oxide [29].

CONCLUSION

This study provides evidence that EDCAs chronically exposed to pesticides present measurable alterations in pulmonary function and respiratory mechanics when compared with non-exposed individuals.

Spirometric and plethysmographic findings indicate a pattern consistent with ventilatory impairment and air trapping, while oscillometric analysis revealed increased resistive load, reduced reactance, and alterations in parameters associated with peripheral airway dysfunction and respiratory work.

Importantly, respiratory oscillometry proved to be a sensitive tool for detecting early and subclinical changes that may not be fully captured by conventional pulmonary function tests alone. The observed increase in total and central resistance, along with changes in reactance-derived parameters such as fr and Ax, suggests the presence of mechanical alterations affecting both central and peripheral airways, likely related to chronic inhalational exposure to toxic agents.

These findings reinforce the relevance of incorporating advanced, noninvasive techniques such as oscillometry into the routine assessment of workers exposed to respiratory hazards. From a perspective of the exposure in the work process, the results highlight the need for improved preventive strategies, including stricter control of exposure, adequate use of personal protective equipment, and continuous health surveillance programs.

Future studies with larger samples and longitudinal designs are warranted to further elucidate the progression of these alterations and to establish oscillometric parameters as reliable biomarkers for early detection and monitoring of lung injury associated with exposure in the work process in this population, as well as validation using multicenter samples.

ACKNOWLEDGMENT

The authors would like to thank CNPq, FAPERJ and CAPES for their financial support.

We also thank all Endemic Disease Control Agents that participated in this study, and in particular to the federal civil servant unions of the category and to Thaís I. L. C. F. Valença that organized the flow of the studied volunteers.

CONFLICT OF INTEREST

The authors report no conflict of interest.

REFERENCES

[1] M. Romanello *et al.*, "The 2024 report of the Lancet Countdown on health and climate change: facing record-breaking threats from delayed action," *J The Lancet,* vol. 404, no. 10465, pp. 1847-1896, 2024.

[2] S. M. Hartinger *et al.*, "The 2023 Latin America report of the Lancet Countdown on health and climate change: the imperative for health-centred climate-resilient development," *J The Lancet Regional Health–Americas,* vol. 33, 2024.

[3] R. Lourenco-de-Oliveira, "Atualidades sobre os vetores e a transmissão de arboviroses no Rio de Janeiro," *J An Acad Nacl Med,* vol. 192, pp. 52-72, 2021.

[4] "MINISTÉRIO DA SAÚDE. Painel de monitoramento da saúde do trabalhador: Agentes de Endemias," SVS/MS, Ed., ed, 2021.

[5] F. F. Carneiro, *Dossiê ABRASCO: um alerta sobre os impactos dos agrotóxicos na saúde*. EPSJV/Expressão Popular, 2015.

[6] L. C. Meirelles *et al.*, "Agentes de combate às endemias: uma população em risco no enfrentamento da Covid-19.," *Observatório Covid-19 Fiocruz,* pp. 321-333, 2022.

[7] ENSP/FIOCRUZ, "Estudo do impacto à saúde de Agentes de Combate às Endemias/Guardas de Endemias pela exposição a agrotóxicos no estado do Rio de Janeiro: resultados do Projeto Integrador Multicêntrico (2018-2023)," 2023.
[8] J. M. S. d. Costa *et al.*, "Exposição ocupacional dos agentes de combate as endemias aos inseticidas," *REVINTER, v. 10, n. 2, p. 134-151,* 2017.
[9] D. Assayag and K. A. Johannson, "Respiratory oscillometry: the new kid on the lung function block," *J BMJ Open Respiratory Research,* vol. 13, no. 1, 2026.
[10] "Sociedade Brasileira de Pneumologia e Tisiologia, Diretrizes para Testes de Função Pulmonar," *Jornal Brasileiro de Pneumologia, Brasília,* vol. 28, pp. 1- 238, 2002.
[11] ATS, "Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories; ATS statement: guidelines for the six-minute walk test," *Am J Respir Crit Care Med,* vol. 166, no. 1, pp. 111-7, Jul 1 2002.
[12] P. L. de Melo, M. M. Werneck, and A. Giannella-Neto, "New impedance spectrometer for scientific and clinical studies of the respiratory system," (in English), *Review of Scientific Instruments,* vol. 71, no. 7, pp. 2867-2872, Jul 2000.
[13] G. G. King *et al.*, "Technical standards for respiratory oscillometry," *Eur Respir J,* vol. 55, no. 2, Feb 2020.
[14] M. R. H. Hansen *et al.*, "Organophosphate and carbamate insecticide exposure is related to lung function change among smallholder farmers: a prospective study," *J Thorax,* vol. 76, no. 8, pp. 780-789, 2021.
[15] P. H. Quanjer and D. J. Weiner, "Measurement of FEF25-75% and FEF75% does not contribute to clinical decision making," *European Respiratory Journal,* vol. 43, no. 4, pp. 1051-8, 2014.
[16] R. Pellegrino *et al.*, "Interpretative strategies for lung function tests," *J European respiratory journal,* vol. 26, no. 5, pp. 948-968, 2005.
[17] W. T. Beachey, "Respiratory Care Anatomy and Physiology: Foundations for Clinical Practice," *Elsevier* vol. 3, 2013.
[18] P. M. Sá, H. A. Castro, A. J. Lopes, and P. L. Melo, "Entropy analysis for the evaluation of respiratory changes due to asbestos exposure and associated smoking," *J Entropy,* vol. 21, no. 3, p. 225, 2019.
[19] S. G. Spiro, G. A. Silvestri, and A. Agustí, *Clinical Respiratory Medicine.* , 4 ed. Philadelfia Saunders Elsevier, 2012.
[20] D. Hendrick, P. S. Burge, W. Beckett, and A. Churg, Elsevier, Ed. *Occupational Disorders of the Lung: recognition, management, and prevention*. London: W B. Saunders, 2002.
[21] J. V. Cavalcanti, A. J. Lopes, J. M. Jansen, and P. L. Melo, "Detection of changes in respiratory mechanics due to increasing degrees of airway obstruction in asthma by the forced oscillation technique," *Respir Med,* vol. 100, no. 12, pp. 2207-19, Dec 2006.
[22] C. Libu, M. R. Otelea, I. A. Arghir, A. Rascu, S. A. Antoniu, and O. C. Arghir, "Challenges in diagnosing occupational chronic obstructive pulmonary disease," *J Medicina,* vol. 57, no. 9, p. 911, 2021.
[23] D. A. Kaminsky *et al.*, "Clinical significance and applications of oscillometry," *Eur Respir Rev,* vol. 31, no. 163, Mar 31 2022.
[24] P. M. de Sá, H. A. Castro, A. J. Lopes, and P. L. d. Melo, "Early diagnosis of respiratory abnormalities in asbestos-exposed workers by the forced oscillation technique," *J PloS one,* vol. 11, no. 9, p. e0161981, 2016.
[25] P. W. Boros, M. Franczuk, and S. Wesolowski, "Value of Spirometry in Detecting Volume Restriction in Interstitial Lung Disease Patients Spirometry in Interstitial Lung Diseases," *J Respiration,* vol. 71, no. 4, pp. 374-379, 2004.
[26] H. Kosak, S. Satar, and P. Ergun, "The relationship of impulse oscillometry results with respiratory muscle strength and exercise capacity in patients with emphysema," *J Northern Clinics of Istanbul,* p. 682, 2025.
[27] S. Moitra *et al.*, "Respiratory morbidity among Indian tea industry workers," *J The International Journal of Occupational Environmental Medicine,* vol. 7, no. 3, p. 148, 2016.
[28] D. MacLeod and M. Birch, "Respiratory input impedance measurement: forced oscillation methods," *J Medical Biological Engineering Computing,* vol. 39, no. 5, pp. 505-516, 2001.
[29] H. Pasker, M. Peeters, P. Genet, J. Clement, B. Nemery, and K. Van de Woestijne, "Short-term ventilatory effects in workers exposed to fumes containing zinc oxide: comparison of forced oscillation technique with spirometry," *J European Respiratory Journal,* vol. 10, no. 7, pp. 1523-1529, 1997.

Prezados Membros da Comissão Temática,

Agradecemos pela oportunidade de submeter uma versão revisada do manuscrito **"*Cold Plasma in Medicine: A Review of Clinical Applications and Underlying Biological Mechanisms*"** para consideração para publicação nos anais do CBEB 2026. Agradecemos também aos revisores pela análise cuidadosa do trabalho e pelos comentários pertinentes, que contribuíram para aprimorar a clareza, a consistência e a apresentação dos resultados. Apresentamos abaixo uma resposta ponto a ponto aos comentários e às questões levantadas pelos revisores, indicando as alterações realizadas no manuscrito e, quando pertinente, esclarecendo as decisões metodológicas adotadas. Também anexamos a versão revisada do manuscrito em um arquivo separado, com todas as alterações destacadas na cor verde. Esperamos que as modificações realizadas tenham esclarecido as questões levantadas durante a revisão e que a versão atualizada atenda aos requisitos do CBEB 2026.

RESPOSTA PARA OS REVISORES

Reviewer #1:

- Uniformizar a classificação metodológica do trabalho como "revisão narrativa" ou "de escopo" em todo o manuscrito — hoje oscila entre "systematic review" (Resumo, seção II) e "narrative review" (Introdução), sem os requisitos de uma revisão sistemática.

  Resposta dos autores: O manuscrito passou a ser classificado, de maneira uniforme, como uma revisão narrativa estruturada. As referências ao trabalho como "revisão sistemática" foram removidas do Resumo, da seção de Métodos e das demais seções do manuscrito.

- Deduplicar os quatro pares de referências repetidas: [7]/[28], [8]/[98], [46]/[90], [47]/[93].

  Resposta dos autores: Os quatro pares de referências duplicadas foram removidos. Posteriormente, todas as citações no texto e as referências bibliográficas foram renumeradas.

- Substituir a citação autor-data solta "(Canady et al.)" pelo número correspondente na lista de referências.

  Resposta dos autores: A citação foi corrigida e substituída pela referência numérica correspondente, atualmente apresentada como [59].

- Detalhar o processo de seleção: data da busca, contagem por base, número de duplicatas e exclusões por etapa (idealmente em fluxograma).

  Resposta dos autores: A seção de Métodos foi ampliada para esclarecer a estratégia de busca, os critérios de elegibilidade, o processo de extração e organização das evidências e a data da última atualização da busca, realizada em 26 de agosto de 2026. Também foi incluído um fluxograma para representar o processo de seleção.

- Reduzir a sobreposição entre Resultados e Discussão, que repetem o mesmo conteúdo por especialidade médica.

  Resposta dos autores: A sobreposição entre Resultado e Discussão foi reduzida.

- Expandir a Conclusão (atualmente 3 frases) com limitações da revisão e direções futuras.

  Resposta dos autores: A conclusão foi expandida.

- Remover o placeholder de template no rodapé.

  Resposta dos autores: o placeholder foi removido.

Reviewer #2:

- Considero apresentar uma tabela de síntese dos 101 estudos, ou pelo menos dos estudos clínicos, contendo aplicação, desenho do estudo, número de participantes, dispositivo/fonte de plasma, gás utilizado, dose ou tempo de exposição, comparador, principais resultados e eventos adversos. Isso permitiria sustentar quantitativamente afirmações como a de que cicatrização de feridas constitui a aplicação clinicamente mais estabelecida, atualmente apresentada principalmente de forma narrativa. Recomendo ainda separar claramente evidências in vitro, animais e humanas, sobretudo em oncologia, pois o próprio manuscrito reconhece que a evidência permanece predominantemente pré-clínica e que ensaios clínicos randomizados de grande escala são escassos. Por fim, incluiria uma avaliação de qualidade metodológica ou risco de viés dos estudos clínicos selecionados e moderaria a conclusão de que CAP possui potencial para "transformar" a prática médica, concentrando-a nas aplicações sustentadas por evidência clínica.

  Resposta dos autores: Após a remoção das referências duplicadas, o manuscrito passou a conter 97 referências únicas. Foram acrescentadas tabelas de síntese dos estudos clínicos com aplicação, desenho, número de participantes, dispositivo, gás, dose ou tempo de exposição, comparador, resultados, eventos adversos e avaliação metodológica. Também foram incluídas tabelas específicas para as principais evidências animais e in vitro. Os Resultados foram reorganizados para separar claramente as evidências humanas, animais e in vitro, especialmente em oncologia, cuja evidência permanece predominantemente pré-clínica. Os estudos clínicos receberam uma avaliação qualitativa do risco de viés. Por fim, a Conclusão foi moderada, destacando que as evidências clínicas mais consistentes estão concentradas no tratamento de feridas e em algumas aplicações dermatológicas, enquanto outras aplicações ainda exigem ensaios clínicos maiores, protocolos padronizados e acompanhamento de segurança em longo prazo.

Comparison Study of Three Different Howland Current Pump Architectures for a Low-Cost EITduino Device

1st Vinícius Zacheo de Goes
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
v.goes@aluno.ufabc.edu.br

2nd Gustavo Pinheiro
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
gustavo.pinheiro@ufabc.edu.br

3rd Mirela de Oliveira Tomazini
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
mirela.tomazini@ufabc.edu.br

4th Jéssica Costa e Silva
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
silva.jessica@ufabc.edu.br

5th Thais H. Samed e Sousa
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
samed.thais@ufabc.edu.br

6th Erick D. L. B. de Camargo
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
erick.leon@ufabc.edu.br

***Abstract*—Electrical Impedance Tomography (EIT) presents a promising, low-cost, and radiation-free alternative to standard anatomical imaging modalities like Computed Tomography. However, ensuring reliable EIT measurements requires a highly stable voltage-controlled current source capable of maintaining high output impedance regardless of load variations. This study evaluates three variations of the Improved Howland Current Pump architecture, a discrete single operational amplifier (TL071), a discrete buffered topology (TL072), and an integrated instrumentation amplifier (AD623), to determine the optimal current injection module for EITduino, an open-source, low-cost EIT device intended for educational and accessible clinical applications. The circuits were modeled using LTspice to assess their frequency response and stability under non-ideal operational conditions across the 10 kHz to 100 kHz frequency spectrum. The evaluation accounted for variable load resistances (100 Ω to 3 kΩ), a 1% resistor mismatch, and a severe parasitic capacitance of 750 pF. Results indicate that the integrated AD623 is severely limited by its 800 kHz gain-bandwidth product, exhibiting significant signal attenuation early in the intended operating range. Furthermore, while the single operational amplifier topology exhibits high-frequency instability under 100 kΩ resistor balancing and a severe cutoff frequency degradation (dropping to the 538 kHz range) under a 20 kΩ balance, the buffered configuration effectively isolates feedback currents and preserves signal integrity. The buffered topology successfully sustained its operational bandwidth near 700 kHz regardless of the balancing resistors utilized and limited maximum current deviation to approximately $\pm 1.5\%$ under resistor tolerance variations. Ultimately, the two-amplifier buffered architecture proved to be the most robust and cost-effective solution, successfully mitigating parasitic degradation and resistor imbalance errors in low-cost EIT instrumentation.**

***Index Terms*—Electrical Impedance Tomography, Howland Current Pump, Bioimpedance, Open-Source Instrumentation, Voltage-Controlled Current Source.**

This work was supported by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001.

I. Introduction

Although Computed Tomography (CT) is the gold standard regarding anatomical imaging [1]–[3], its use is limited by the high cost and the use of ionizing radiation.

The safety of continuous exposure of patients and medical staff to ionizing radiation remains a topic of debate in the literature. Several studies support the claim that tomography is safe and does not lead to adverse health effects for patients [4]. On the other hand, other studies suggest the opposite, indicating a notable increase in the incidence of cancer among individuals who have undergone CT scans [4].

Furthermore, the costs associated with acquiring and maintaining a CT scanner are substantial, often exceeding thousands of reais per month and reaching millions over time [5], [6].

Considering the potential risks imposed on patients due to ionizing radiation in conventional CT scans, as well as the high acquisition and maintenance costs, researchers are exploring alternative imaging methods that are safer and more cost-effective. One such approach is Electrical Impedance Tomography (EIT).

EIT can detect and process variations in electrical impedance between different types of tissues in the body [7]. The applied current is minimal and adheres to safety standards for medical devices [7], [8], ensuring that it does not produce the potentially harmful effects associated with high-energy radiation or any noticeable electrical side effects.

In devices of this nature, an electrical current is injected into the patient through an electrode. Such a signal is typically generated by a voltage-controlled current source. To ensure that the generated current presents a high output impedance and constant current for any load, the Howland Current Pump

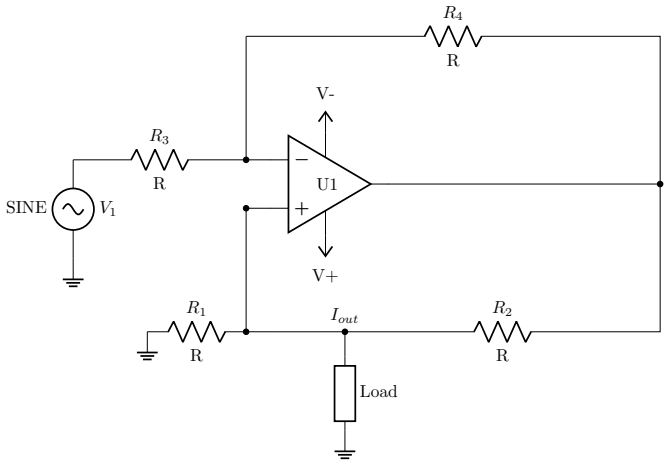


Fig. 1. The diagram shows a generic representation of the Howland pump circuit configuration.

topology is commonly employed [9]–[11], as shown in the diagram of Fig. 1.

The high output impedance is guaranteed by the relationship (1)

$$Z_{\text{out}} = \frac{V_x}{I_{\text{test}}} \tag{1}$$

where Z_{out} represents the output impedance, V_x the voltage at the output node, and I_{test} the theoretical current drained by the external source when $V_{\text{in+}}$ (non-inverting input) and $V_{\text{in-}}$ (inverting input) are grounded. In an ideal source, $I_{\text{test}} \to 0$, resulting in (2)

$$Z_{\text{out}} = \lim_{I_{\text{test}} \to 0} \frac{V_x}{I_{\text{test}}} = \infty \tag{2}$$

which guarantees, therefore, that the output impedance is high.

In this context, the optimization of this circuit is imperative to ensure that, under non-ideal conditions, the device maintains the high output impedance of the current source. Consequently, the present study analyzes three distinct architectures of the Improved Howland Current Pump, aiming at its implementation and improvement in EITduino, an open-source, low-cost EIT equipment, through computational simulations. Such tests were designed to highlight potential signal distortions resulting from resistor imbalance and the presence of parasitic capacitances inherent to the circuit in practical scenarios.

II. Methodology

To ensure the reliability of the current injection module for the EIT system, three voltage-controlled current source topologies were evaluated. The topologies were selected based on technical application notes [12], [13] and were modeled and simulated using LTspice [14]. Additionally, the components selected for the implementation of these circuits prioritized the balance between performance and cost, aligning with the objective of developing a low-cost device.

The evaluated topologies were:

- **Discrete Improved Howland Current Pump:** A single operational amplifier configuration built with discrete components. While effective, this topology relies heavily on precise resistor matching to maintain high output impedance.
- **Discrete Improved Howland Current Pump with Buffer:** A two-amplifier configuration where a buffer is added to the feedback loop. The buffer practically eliminates the feedback current, which simplifies the selection of the feedback resistors, improves output current control, and offers better resilience against parasitic capacitance.
- **Integrated Improved Howland Current Pump:** An implementation utilizing a precision difference amplifier. This topology leverages laser-trimmed internal resistors, virtually eliminating external matching errors and ensuring superior stability with minimal external components.

To assess the robustness of each topology under realistic operating conditions, multiple variables and non-ideal scenarios were simulated:

- **Resistor Balancing Points:** For the two discrete topologies, the performance was evaluated at two distinct nominal balancing resistance values: 20 kΩ and 100 kΩ.
- **Resistor Mismatch and Tolerance:** An intentional imbalance was introduced by changing the value of resistor R2 by $\pm 1\%$ from its nominal value, which, in this case, is considered to be 1 mA or -60 dB.
- **Parasitic Capacitance Analysis:** Parasitic capacitance is recognized as a primary source of error in electrical bioimpedance systems, originating mainly from coaxial cables and electrode-skin interfaces [15]. To evaluate the stability and frequency response of the current pumps under these non-ideal conditions, a severe parasitic capacitance of 750 pF was simulated in parallel with the load. This specific magnitude was chosen based on empirical data from literature, which report average measured parasitic capacitances of 712 ± 128 pF in realistic bioimpedance instrumentation setups [16].
- **Variable Load Resistance:** Tests conducted across 100 Ω, 500 Ω, 1 kΩ, 2 kΩ, and 3 kΩ.

The performance data for each configuration were compiled into comparative tables. The results map each simulated load value against key frequency response metrics: the Limit Frequency (defined as a 1% magnitude drop) and the Cutoff Frequency (-3dB).

III. Results

A. Constructed Circuits

For the implementation of the circuits, resistors with a 1% tolerance were employed, comprising nominal values of 20 kΩ and 100 kΩ. The final configuration of the three assembled circuits is illustrated in Fig. 2, 3, and 4.

For the implementation of the Improved Howland Current Pump topologies, the TL07X family of operational amplifiers was chosen due to its cost-benefit ratio and suitability to the

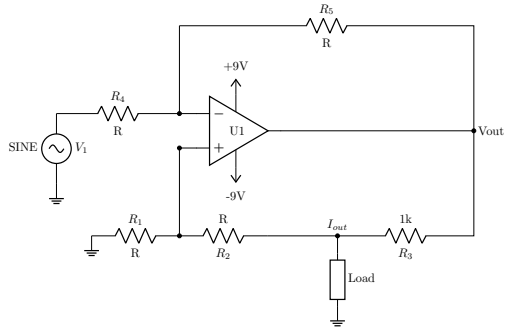


Fig. 2. Schematic of the assembled circuit in the Improved Howland topology using only a single TL071 amplifier.

device's performance requirements. These components present a Gain-Bandwidth Product (GBW) of 5 MHz and a slew rate of 20 V/μs [17], granting them high dynamic characteristics. Such parameters, essential for the stability and precision of said topology, were determining factors for their selection.

On the other hand, the final topology of the Howland Pump incorporated the AD623 instrumentation amplifier. The choice of this component was based on its economic competitiveness against alternatives such as the INA592. However, it should be noted that the AD623 presents technical limitations, having a GBW of 800 kHz (for unity gain) and a slew rate of 0.3 V/μs [18]. Conversely, the differentiator of this component lies in the integration of highly precise internally trimmed resistors, eliminating the need for external components for this purpose, with the exception of the resistor intended to control the output current, denoted by R_S.

For comparative purposes, the topology constructed using the TL071 Improved Howland was altered to accomodate the AD826 to be used as a more robust and ideal standard.

B. Frequency Response (Baseline vs. Parasitic 750 pF)

The frequency response of the Improved Howland Current Pump circuit was evaluated under two distinct resistive balancing conditions: 20 kΩ and 100 kΩ, as detailed in Table I, which specifies the data obtained in the tests for the Improved Howland circuit. A significant divergence between the configurations is observed; the balancing at 100 kΩ exhibits lower attenuation, while the 20 kΩ one presents higher attenuation,

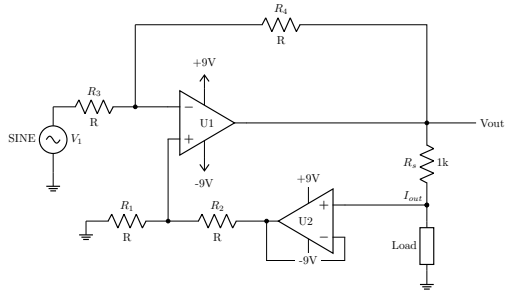


Fig. 3. Schematic of the assembled circuit in the Improved Howland topology using two TL072 amplifiers, one of which is in a buffer configuration for the circuit's feedback.

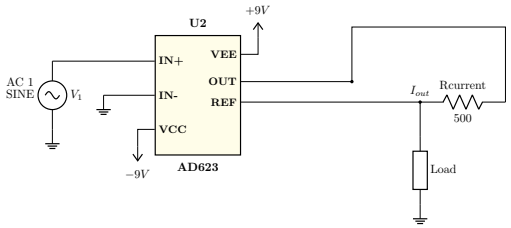


Fig. 4. Schematic of the assembled circuit in the Improved Howland topology using the internal resistors of the AD623 instrumentation amplifier.

this characteristic being particularly prominent at the 1% attenuation limit. Such behavior stems from the use of higher magnitude resistors, which exacerbates the errors associated with bandwidth, instability, and parasitic capacitances, concurrently with the increase in the output impedance of the circuit [12], [13]. Additionally, due to the inherent project needs that determined the use of a less complex amplifier, the limitations of the component become more evident.

As illustrated by Fig. 5, in association with the data from Table II, it is verified that, for lower resistance values, the current magnitude is higher at higher frequencies.

It is observed that the values referring to the 1% gain variation with the balancing at 100 kΩ coincide with the 1% signal attenuation for the circuit balanced at 20 kΩ, around 155 kHz. This fact shows that, in both configurations, the current magnitude variation begins at approximately equivalent frequency thresholds. Additionally, at the frequency of 651.73 kHz, under a load of 100 Ω, a positive variation of 10% in the current magnitude is observed, a characteristic that diverges from the expected ideal behavior for conventional applications of this class of device.

This phenomenon can also be verified when a capacitance is applied in parallel to the load, as can be seen in Table I.

On the other hand, when a more robust amplifier such as the AD826 is used in the circuit, the peaks disappear for both

TABLE I
IMPROVED HOWLAND CIRCUIT: BASELINE AND PARASITIC CAPACITANCE (20 KΩ VS 100 KΩ)

Load	**Baseline**				**Parasitic Capacitance (750 pF)**			
	Bal. 20 kΩ		**Bal. 100 kΩ**		**Bal. 20 kΩ**		**Bal. 100 kΩ**	
(Ω)	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)
100	1.12 M	155.73 k	1.16 M	743.16 k	904.21 k	132.48 k	1.10 M	664.87 k
500	775.71 k	105.54 k	958.10 k	483.28 k	315.50 k	45.63 k	378.30 k	52.14 k
1k	540.18 k	76.05 k	723.78 k	170.51 k	167.75 k	24.03 k	178.40 k	25.19 k
2k	346.79 k	49.14 k	446.21 k	70.33 k	85.92 k	12.30 k	88.02 k	12.55 k
3k	256.21 k	36.40 k	312.92 k	46.30 k	57.68 k	8.28 k	58.58 k	8.39 k

TABLE II
IMPROVED HOWLAND CIRCUIT: CURRENT PEAKS FOR LOW LOADS (BALANCING AT 100 KΩ)

Load	**Peaks**	
(Ω)	1% Gain (Hz)	Peak (Hz)
100	155.85 k	651.73 k
500	275.53 k	348.91 k

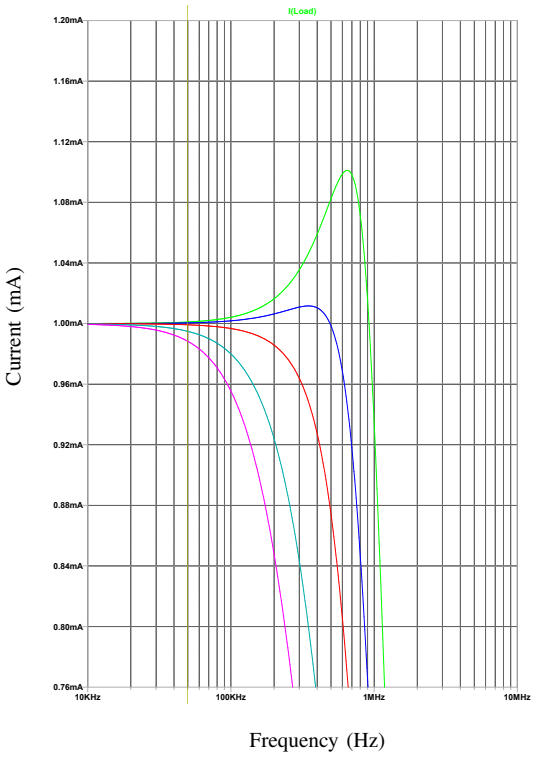


Fig. 5. Graph highlighting the peaks that occur at high frequencies for low load resistances

resistive balancings, as evidenced by Fig. 6, in which no peaks can be seen, even for the lowest load values.

Additionally, this phenomenon is no longer observed upon the insertion of a second amplifier in the positive feedback loop, as evidenced in Table III, which details the values obtained in the test for the Improved Howland topology with Buffer. Such behavior stems from the high input impedance provided by the amplifier in the buffer configuration, which almost entirely suppresses the feedback current [12] [13], resulting in the improvement of the circuit's stability. Furthermore, the feedback current itself acts as a source for several other potential inaccuracies [19] which, in this topology, are mitigated without the need to employ higher magnitude resistors [20]. Thus, the amplification of instabilities that would occur if such resistors were used to perform the same role as the high input impedance provided by the second amplifier in buffer configuration is avoided.

An additional advantage of this configuration, although less expressive, consists of the isolation of the resistor responsible for defining the magnitude of the current directed to the load, a characteristic that considerably facilitates the adjustment of the current if necessary.

In contrast to the other topologies, the Howland Current Pump, when implemented with the AD623 instrumentation amplifier, presents a more pronounced attenuation, besides manifesting itself earlier, even in the base case, as evidenced in Table IV.

This limitation stems from the reduced gain-bandwidth

TABLE III
IMPROVED HOWLAND BUFFER TOPOLOGY: BASELINE AND PARASITIC CAPACITANCE (20 KΩ VS 100 KΩ)

Load	Baseline				Parasitic Capacitance (750 pF)			
	Bal. 20 kΩ		Bal. 100 kΩ		Bal. 20 kΩ		Bal. 100 kΩ	
(Ω)	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)
100	1.61 M	239.30 k	1.61 M	239.95 k	1.16 M	175.38 k	1.16 M	175.67 k
500	1.01 M	145.09 k	1.01 M	145.51 k	338.85 k	48.66 k	339.09 k	48.70 k
1k	692.72 k	99.03 k	694.69 k	88.32 k	175.05 k	25.01 k	175.15 k	25.03 k
2k	428.94 k	61.17 k	430.21 k	61.35 k	88.66 k	12.68 k	88.71 k	12.68 k
3k	311.41 k	44.39 k	312.34 k	44.52 k	59.34 k	8.51 k	59.37 k	8.52 k

product (GBW) of the aforementioned amplifier, which is located at approximately 800 kHz for unity gain. On the other hand, the TL071 model presents a GBW of 5 MHz. Consequently, as the input signal frequency approaches the 800 kHz limit, the AD623 becomes unable to follow the waveform variation, resulting in attenuation and other degradations in

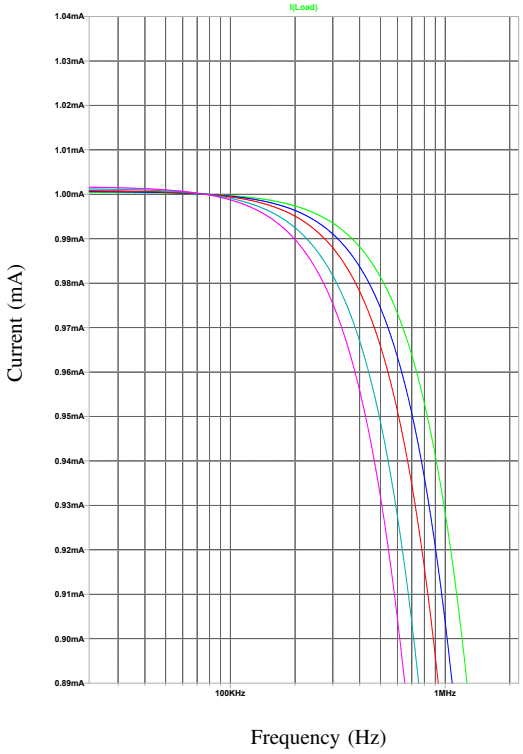


Fig. 6. Attenuation graph of the current generated by the Howland Pump balanced at 100 kΩ using the AD826

TABLE IV
AD623 INTEGRATED TOPOLOGY: BASELINE AND PARASITIC CAPACITANCE

Load (ZL)	Base Condition		Parasitic Capacitance (750 pF)	
	-3dB (Hz)	-1% (Hz)	-3dB (Hz)	-1% (Hz)
100 Ω	694.87 k	325.15 k	660.95 k	269.75 k
500 Ω	497.24 k	106.70 k	286.57 k	41.61 k
1 kΩ	350.69 k	57.38 k	147.84 k	21.04 k
2 kΩ	213.78 k	31.60 k	75.38 k	10.77 k
3 kΩ	153.24 k	22.20 k	50.97 k	7.32 k

signal integrity. Additionally, the AD623 exhibits a limited slew rate of 0.3 V/μs, a value significantly lower than the 20 V/μs observed in the TL071.

Furthermore, when comparing the base case with the parasitic capacitance test, an attenuation of approximately 20% is observed for the lowest load values, and about 80% for the highest values for all studied topologies. This discrepancy is attributed to the fact that the insertion of the capacitance in parallel reduces the circuit's output impedance, causing a decrease in the expected frequency response, as in (3)

$$Z_{\text{out}} = \frac{1}{j\omega C_{\text{parasitic}}} \tag{3}$$

where Z_{out} denotes the circuit's output impedance, ω represents the angular frequency, expressed by $2\pi f$, and $C_{\text{parasitic}}$ corresponds to the parasitic capacitance, simulated by the capacitor added in parallel to perform this test. From this equation, the inversely proportional relationship between the output impedance, the capacitance, and the angular frequency becomes evident, so that the increment of the latter two variables necessarily implies the reduction of Z_{out}. Such a property justifies the observed drop in signal amplitude.

In practice, Electrical Impedance Tomography (EIT) systems typically operate in the range of 10 kHz to 100 kHz [7], [16], [20] for the characterization of materials and biological tissues. In this frequency spectrum, Tables I, III, which compile the simulation data of the Howland circuit in the improved and buffered configurations, demonstrate the stability of these topologies. Within them, reduced or negligible attenuations are observed, even under the strong influence of parasitic capacitances [15].

In contrast, Table IV, which reports the results of the topology implemented with the AD623 amplifier [18], reveals significant attenuations for this same operating range (presenting a 1% drop as early as 21 kHz for a 1 kΩ load). Consequently, such an architecture proves to be a less viable alternative for the characterization of biological tissues in the analyzed frequency spectrum.

Regarding integration into a complete EIT acquisition system, scalability introduces hardware limitations. The routing of signals through analog multiplexers inherently adds series resistance and stray capacitance [15]. The simulated 750 pF parasitic capacitance serves as a worst-case baseline for these effects. While the buffered and improved Howland architectures successfully mitigate these parasitic degradations, keeping the system stable [13], [19], [20], future physical implementations for educational platforms like EITduino must carefully design the PCB layout to minimize trace capacitance and crosstalk across multiple electrode channels [11], [15].

C. Sensitivity Analysis (Resistor Mismatch)

Given the integration of high-precision resistors in the AD623 instrumentation amplifier, performing simulations aimed at analyzing the effects of resistive imbalance becomes dispensable. In contrast, the remaining topologies, implemented with nominal 1% tolerance resistors, are subject to variations that can lead to discrepant results in a benchtop environment, distancing themselves from the ideal behavior predicted in simulation.

TABLE V
EFFECT OF $\pm 1\%$ RESISTOR TOLERANCE ON CURRENT OUTPUT AND BANDWIDTH (1 kΩ LOAD AND 20 kΩ RESISTOR BALANCE)

Topology	Max. Deviation	Cutoff Shift (-3dB)
Discrete (TL071)	$\approx \pm 1.5\%$	538 kHz – 549 kHz
Buffered (TL072)	$\approx \pm 1.5\%$	685 kHz – 700 kHz
AD623*	N/A	N/A

* AD623 utilizes internal laser-trimmed resistors.

For the conducted simulations, the sixth step was considered the nominal value, being 1 mA due to the resistor used to define the current, 1 kΩ. According to Ohm's Law

$$I = \frac{V}{R} \tag{4}$$

where I is the nominal value, R is the current-defining resistor with a value of 1 kΩ and V is the voltage set to 1 V, as is the case in all simulations, it can be deduced that

$$I = \frac{1}{1000} = 1 \text{ mA} \tag{5}$$

This specific target of 1 mA was not chosen arbitrarily; it represents the maximum safe current limit for medical applications operating in the 10 kHz to 100 kHz frequency band, ensuring strict adherence to the IEC 60601-1 safety standard [8].

Thus, the difference in attenuation under various resistor mismatches can be analyzed. Tables V and VI present the results obtained for both the 20 kΩ resistive balance (in Table V) and the 100 kΩ resistive balance (in Table VI). No significant discrepancy between them is observed: the maximum deviation for both analyzed topologies is $\approx \pm 1.5\%$ for the 100 kΩ and 20 kΩ balancing.

The attenuation and gain levels can be easily visualized through the graphs generated for the simulations. All graphs have the same scale on the y and x-axes. Fig. 7 and 8 show the results seen in Table V. The maximum and minimum deviations observed in both topologies proved to be effectively equivalent, standing around 1.015 mA for the maximum gain and 0.985 mA for the maximum attenuation. Both alternatives prove adequate for the purpose of this type of instrument, considering that the current value is a parameter employed in

TABLE VI
EFFECT OF $\pm 1\%$ RESISTOR TOLERANCE ON CURRENT OUTPUT AND BANDWIDTH (1 kΩ LOAD AND 100 kΩ RESISTOR BALANCE)

Topology	Max. Deviation	Cutoff Shift (-3dB)
Discrete (TL071)	$\approx \pm 1.5\%$	721 kHz – 734 kHz
Buffered (TL072)	$\approx \pm 1.5\%$	687 kHz – 702 kHz
AD623*	N/A	N/A

* AD623 utilizes internal laser-trimmed resistors.

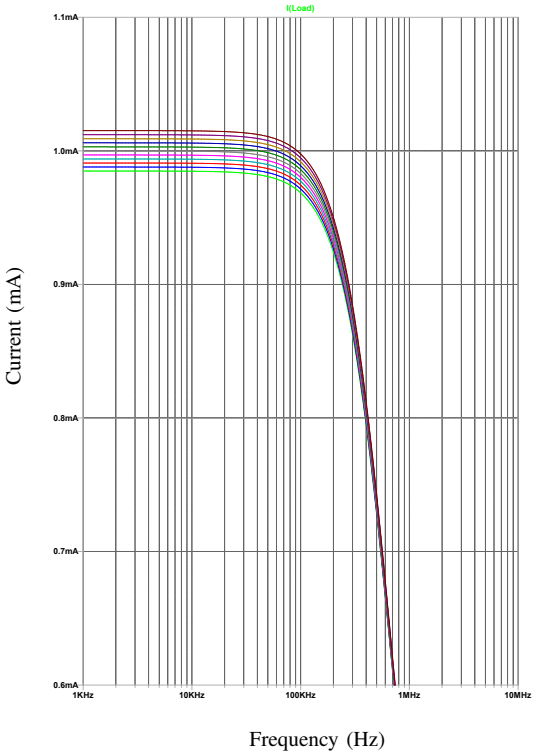


Fig. 7. Sensitivity test graph for the Improved Howland topology utilizing the TL071 Operational Amplifier in a 20 kΩ resistive balance

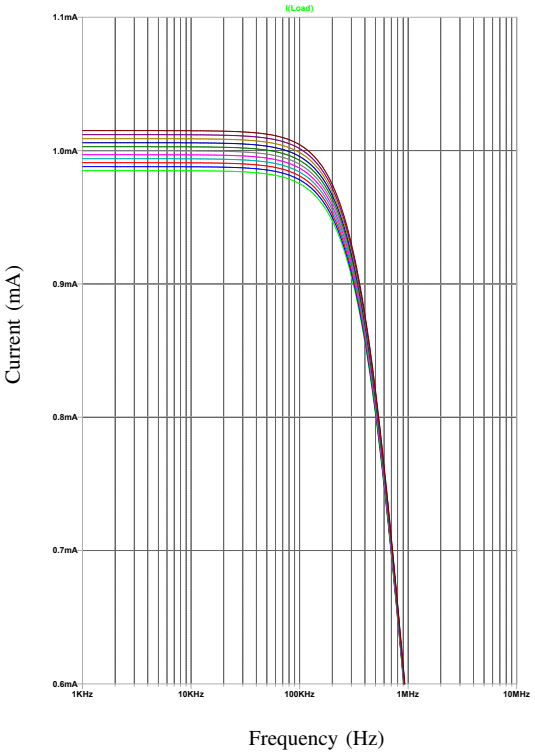


Fig. 8. Sensitivity test graph for the Improved Howland with a buffer topology utilizing the TL072 Operational Amplifier in a 20 kΩ resistive balance

the subsequent image reconstruction steps [7], [20] and the observed variation is not of sufficient magnitude to compromise patient safety, remaining strictly within the normative limits established by the IEC 60601-1 standard [8].

Similarly, Fig. 9 and 10 demonstrate almost the same results. These graphical representations illustrate the data detailed in Table VI. In this scenario, the maximum gain also reached a value of 1.015 mA, and the maximum attenuation was limited to 0.985 mA.

Additionally, a significant difference is observed between the cutoff frequency values of the two resistive balancings evaluated in the topology based on the TL071 operational amplifier. It is noted that this architecture is more susceptible to the low resistance values employed, exhibiting cutoff frequencies in the range of 538 kHz to 549 kHz for the 20 kΩ balancing, and from 721 kHz to 734 kHz for the 100 kΩ balancing.

When comparing these data with the results obtained in the buffered topology (using the TL072 component)–whose values ranged from 685 kHz to 700 kHz for the 20 kΩ balancing, and from 687 kHz to 702 kHz for 100 kΩ–the aforementioned impact becomes evident, corroborating the greater stability of the buffered configuration. This behavior is justified by the introduction of a high-impedance component in the non-inverting feedback loop of this topology [12], [13]. Such an addition acts directly in mitigating parasitic effects and maintaining the output impedance, consequently preventing a decrease in the circuit's cutoff frequency [15], [19], [20].

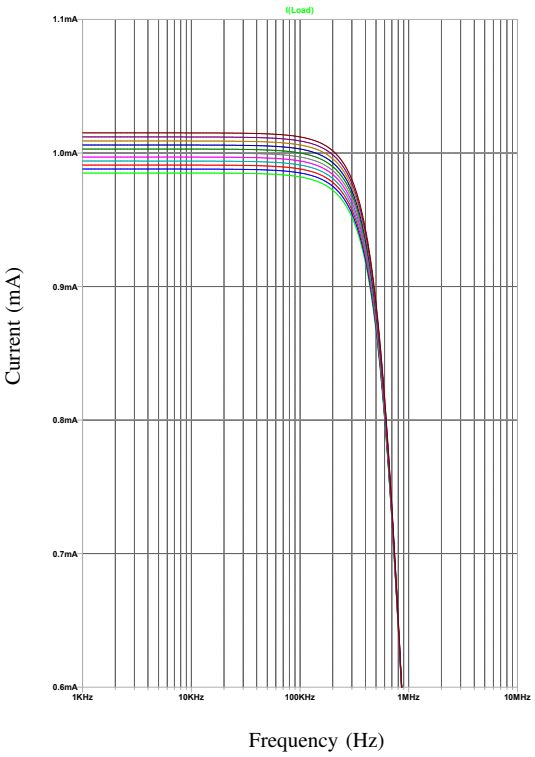


Fig. 9. Sensitivity test graph for the Improved Howland topology utilizing the TL071 Operational Amplifier in a 100 kΩ resistive balance

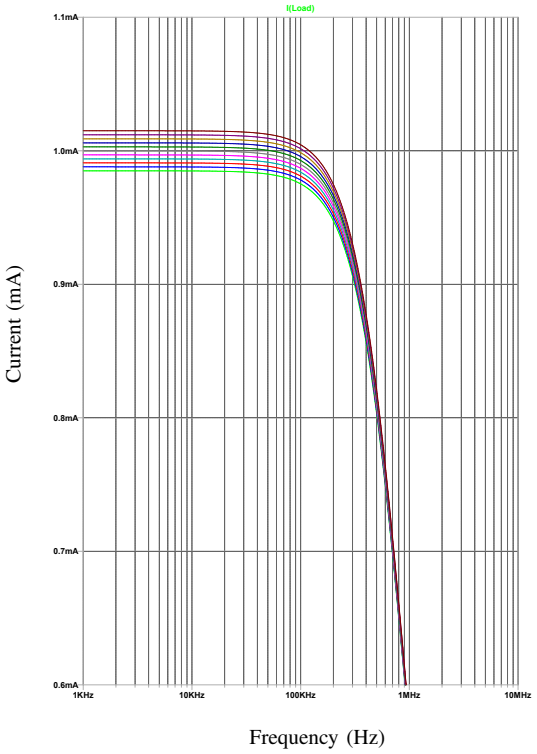


Fig. 10. Sensitivity test graph for the Improved Howland with a buffer topology utilizing the TL072 Operational Amplifier in a 100 kΩ resistive balance

D. Consolidated Results

To facilitate further analysis, a consolidated table was made

TABLE VII
CONSOLIDATED COMPARISON OF THE THREE HOWLAND CURRENT PUMP TOPOLOGIES (1 kΩ LOAD, 20 kΩ RESISTOR BALANCE WHERE APPLICABLE)

Parameter	Discrete (TL071)	Buffered (TL072)	AD623 (Integrated)
GBW / Slew Rate	5 MHz / 20 V/μs	5 MHz / 20 V/μs	0.8 MHz / 0.3 V/μs
-3dB Baseline	540.18 kHz	692.72 kHz	350.69 kHz
-3dB w/ 750 pF	167.75 kHz	175.05 kHz	147.84 kHz
Max. Deviation ($\pm$1% R)	$\approx \pm 1.5\%$	$\approx \pm 1.5\%$	N/A (internal laser-trimmed resistors)
Instability (Peaks)	Yes (+10% at 100 Ω, 651.73 kHz)	No	No (early attenuation)

IV. CONCLUSION

In conclusion, given the design constraints, the measured trade-offs highlight the advantages of the Improved Howland Current Pump topology composed of two operational amplifiers, one of them operating in a buffer configuration. This study evaluated three architectures, a discrete single operational amplifier (TL071), a discrete buffered topology (TL072), and an integrated instrumentation amplifier (AD623), to determine the optimal current injection module for the EITduino device operating within the 10 kHz to 100 kHz frequency spectrum.

This buffered circuit demonstrated high robustness against resistive imbalance and parasitic capacitance, which are critical factors for the proper functioning of the current source. It consistently maintained a highly stable output capable of delivering the nominal 1 mA safe current mandated by the IEC 60601-1 standards for medical applications. The parametric mismatch sweep revealed that while a 1% component tolerance induces an equivalent maximum current deviation of $\approx \pm 1.5\%$ across all resistive configurations, the buffered topology demonstrates superior high-frequency resilience. Unlike the single operational amplifier architecture, which suffers a severe cutoff frequency degradation (dropping to the 538 kHz range) under a 20 kΩ balance, the buffered configuration effectively sustains its operational bandwidth near 700 kHz regardless of the balancing resistors utilized.

Furthermore, the two-amplifier architecture effectively mitigated the problems observed in the other evaluated topologies. Among the overcome limitations, the amplification of high-frequency instabilities and signal peaks associated with the 100 kΩ resistive balancing in the single operational amplifier architecture is notable. The buffered topology successfully isolates feedback currents and preserves signal integrity, even under a severe 750 pF parasitic capacitance representing the stray effects of cables and analog multiplexers.

Additionally, the buffered circuit surpassed the restricted performance of the instrumentation amplifier tested in isolation, whose selection had been primarily guided by its low cost compared to components in the same category. Despite its internal precision, the AD623 was severely limited by its 800 kHz gain-bandwidth product, presenting significant attenuation early in the intended operating range. Ultimately, the buffered Improved Howland topology proved to be the most robust and cost-effective solution for mitigating parasitic degradation and ensuring reliable measurements in low-cost EIT instrumentation.

REFERENCES

[1] D. J. Brenner and E. J. Hall, "Computed tomography—an increasing source of radiation exposure," *New England Journal of Medicine*, vol. 357, no. 22, pp. 2277–2284, 2007.

[2] W. A. Kalender, "X-ray computed tomography," *Physics in Medicine & Biology*, vol. 51, no. 13, p. R29, 2006.

[3] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt Jr, and J. M. Boone, *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 3rd ed., 2011.

[4] H.-M. Shi, Z.-C. Sun, and F.-H. Ju, "Understanding the harm of low-dose computed tomography radiation to the body," *Experimental and therapeutic medicine*, vol. 24, no. 2, pp. 1–6, 2022.

[5] H. M. Pereira, R. Wernke, and C. de Oliveira Ritta, "Aplicação do TD-ABC na avaliação da lucratividade de exames e do nível de ociosidade da tomografia computadorizada do setor de oncologia ginecológica," *Revista Catarinense da Ciência Contábil*, no. 23, p. 31, 2024.

[6] U. de Comunicação Social HU/UFSC/Ebserh, "Hospital universitario investe mais de 3,3 milhoes de reais em novo tomografo," 2021. Last accessed 11 September 2024.

[7] A. Adler and D. Holder, *Electrical impedance tomography: Methods, history and applications*. CRC Press, 2021.

[8] International Electrotechnical Commission, "Medical electrical equipment – Part 1: General requirements for basic safety and essential performance," Standard IEC 60601-1:2005, International Electrotechnical Commission, Geneva, Switzerland, 2005.

[9] Philbrick Researches, *Applications manual for computing amplifiers for modelling, measuring, manipulating & much else*. Boston, MA: Nimrod Press, 1966.

[10] R. A. Pease, "A comprehensive study of the Howland current pump," application note 1515, National Semiconductor, 2008.

[11] S. Franco, *Design with operational amplifiers and analog integrated circuits*. McGraw-Hill Education, 4th ed., 2014.
[12] I. V. Lam, "Analysis of improved Howland current pump configurations," tech. rep., Texas Instruments, 2020.
[13] Texas Instruments, "AN-1515 a comprehensive study of the Howland current pump," tech. rep., Texas Instruments, 2008.
[14] Analog Devices, *LTspice simulator*. Analog Devices, Inc., Norwood, MA, USA, 2024. https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html.
[15] D. W. C. Marcôndes, A. S. Paterno, and P. Bertemes-Filho, "Parasitic effects on electrical bioimpedance systems: Critical review," *Sensors*, vol. 22, no. 22, p. 8820, 2022.
[16] R. E. Dodde, G. H. Kruger, and A. J. Shih, "Design of bioimpedance spectroscopy instrument with compensation techniques for soft tissue characterization," *Journal of Medical Devices*, vol. 9, no. 2, p. 021001, 2015.
[17] Texas Instruments, *TL07xx low-noise JFET-input operational amplifiers*. Texas Instruments, 2017. Datasheet.
[18] Analog Devices, *AD623: Single-supply, rail-to-rail, low cost instrumentation amplifier*. Analog Devices, 2023. Datasheet.
[19] A. S. Tucker, R. M. Fox, and R. J. Sadleir, "Output impedance of broadband Howland current pump," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 1, pp. 63–74, 2013.
[20] A. L. McEwan, A. Romsauerova, R. Trebani, D. S. Holder, and B. H. Brown, "A wideband high output impedance current source for electrical impedance tomography," *Physiological Measurement*, vol. 28, no. 7, p. S61, 2007.

Dispositivo Portátil Multimodal para Aquisição de Sinais Biomecânicos e Bioelétricos Aplicados ao Monitoramento da Doença de Parkinson

José Carlos Gomes da Silva
Departamento de Educação Física
UFRN
Natal, RN, Brasil

Joice Heloise de Oliveira Tomaz
Departamento de Fisioterapia
UFRN
Natal, RN, Brasil

Ana Sofia Varela Siqueira
Departamento de Fisioterapia
UFRN
Natal, RN, Brasil

Antônio Domingos da Silva Neto
Departamento de Fisioterapia
UFRN
Natal, RN, Brasil

Mário André da Cunha Espada
Escola Superior de Educação
Instituto Politécnico de Setúbal
Setúbal, Portugal

Sergio Ferreira Silva Júnior
Instituto Metrópole Digital
UFRN
Natal, RN, Brasil

Luiz Eduardo Cunha Leite
Escola de Ciências e Tecnologia
UFRN
Natal, RN, Brasil

Paulo Moreira Silva Dantas
Departamento de Educação Física
UFRN
Natal, RN, Brasil

***Resumo* — A avaliação motora da mão de pessoas com doença de Parkinson (DP) pode se beneficiar de medidas complementares de força, movimento e atividade muscular obtidas durante a realização de uma determinada tarefa. A integração dessas medidas permitiria investigar sua associação com a gravidade do comprometimento motor do paciente. Este artigo apresenta uma proposta de arquitetura para uma manopla instrumentada destinada à avaliação motora da mão de pessoas com DP. A arquitetura compreende subsistemas de estrutura mecânica, medição de força por dedo, medição da preensão global, sensoriamento inercial do tremor, eletromiografia de superfície, aquisição e comunicação de dados, inteligência artificial e interface com o usuário. A organização da arquitetura define as funções e interfaces de cada subsistema, permitindo a implementação através de diferentes componentes. A implementação de uma manopla baseada na arquitetura proposta encontra-se em andamento, com desenvolvimento da estrutura mecânica e dos subsistemas de força por dedo, preensão global e sensoriamento inercial, incluindo eletrônica, programas embarcados e de computador. Essas implementações apoiarão a caracterização dos subsistemas e a posterior validação da arquitetura integrada. A integração temporal dos sinais, a eletromiografia de superfície e a inteligência artificial estão previstas para etapas posteriores. São apresentadas as escolhas técnicas atuais e o planejamento da caracterização metrológica, da integração dos subsistemas e da avaliação com participantes. A contribuição do trabalho consiste na definição da arquitetura multimodal e de uma estratégia de desenvolvimento e validação por etapas. O desempenho diagnóstico e a validade clínica da plataforma ainda não foram estabelecidos.**

Palavras-chave—doença de Parkinson, arquitetura de sistemas, instrumentação biomédica, preensão palmar, avaliação motora.

***Abstract* — Motor assessment of the hand in people with Parkinson's disease (PD) may benefit from complementary measurements of force, movement, and muscle activity obtained during the same task. Integrating these measurements may enable investigation of their association with the severity of motor impairment. This paper presents the proposed architecture of an instrumented handgrip device intended for manual motor assessment in people with PD. The architecture comprises subsystems for mechanical structure, finger force measurement, overall grip force measurement, inertial tremor sensing, surface electromyography, data acquisition and communication, artificial intelligence, and a user interface. Its organization defines functions and interfaces, allowing alternative component choices. Implementation of the handgrip device is ongoing, with development of the mechanical structure and the finger force, overall grip force, and inertial sensing subsystems, including electronics, embedded software, and experimental interfaces. These implementations will support subsystem characterization and subsequent validation of the integrated architecture. Temporal integration of the signals, surface electromyography, and artificial intelligence are planned for later stages. Current technical choices and plans for metrological characterization, subsystem integration, and evaluation with participants are presented. The contribution of this work lies in defining the multimodal architecture and a staged development and validation strategy. The platform's diagnostic performance and clinical validity have not yet been established.**

Keywords—Parkinson disease, system architecture, biomedical instrumentation, handgrip, motor assessment.

I. INTRODUÇÃO

A doença de Parkinson (DP) é uma condição neurodegenerativa progressiva caracterizada pela perda de células do sistema nervoso central, mais especificamente dos neurônios dopaminérgicos na substância negra do mesencéfalo [1]. Mesmo após mais de 200 anos de sua descrição inicial, o diagnóstico clínico da DP permanece um desafio complexo, com amplas implicações para o manejo clínico [2]. A doença caracteriza-se pela presença de manifestações motoras e não motoras múltiplas, incluindo tremor em repouso, bradicinesia, rigidez muscular e instabilidade postural [3,4]. Para tanto, diversas escalas de avaliação têm sido utilizadas para monitorar a DP; dentre as mais tradicionais estão a de Hoehn e Yahr, para graduar o estágio da doença, e a Escala Unificada de Avaliação da Doença de Parkinson (UPDRS – Unified Parkinson's Disease Rating Scale), para avaliar a incapacidade e o comprometimento [4]. No entanto, embora sejam instrumentos padronizados e amplamente utilizados, essas escalas dependem da avaliação clínica de um examinador

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

experiente, estando sujeitas à variabilidade intra e interavaliador, especialmente na avaliação de determinados sinais motores da doença [5]. A revisão da UPDRS patrocinada pela Sociedade de Distúrbios do Movimento (Movement Disorder Society), denominada MDS-UPDRS, constitui a referência clínica prevista para os estudos de validação deste projeto [6].

Nesse contexto, o monitoramento mais objetivo do desempenho motor tem sido favorecido pelo surgimento de dispositivos instrumentados e de biomarcadores digitais que permitem a obtenção de medidas quantitativas por meio de sensores [1]. O avanço das tecnologias digitais em saúde tem possibilitado a incorporação de medidas instrumentais capazes de complementar as avaliações clínicas tradicionais. Entre essas tecnologias, destacam-se os dispositivos vestíveis (wearables), dotados de acelerômetros e giroscópios, utilizados no monitoramento da frequência e da amplitude dos tremores por meio de sensores inerciais. Ali et al. [7], por exemplo, investigaram esses sensores na estimativa da gravidade do tremor essencial, aplicação que não deve ser confundida com validação para a DP.

Entretanto, a avaliação de variáveis cinemáticas, como tremor e movimento, quando realizada sem integrar informações relacionadas à força muscular e à atividade mioelétrica, limita a caracterização geral do desempenho motor. Dessa forma, a análise do desempenho motor pode permanecer fragmentada, exigindo múltiplos equipamentos para avaliação de diferentes domínios motores, o que dificulta a obtenção de uma caracterização integrada da função neuromuscular. Embora já existam propostas de instrumentação multimodal da mão, como a luva apresentada por Li et al. [8], a integração de medidas complementares em uma mesma tarefa de preensão permanece uma questão de projeto a investigar.

Assim sendo, a progressão dos sintomas motores na DP geralmente é acompanhada da redução da força de preensão palmar manual (FPM), que tem sido apontada como um importante biomarcador [9]. Além de ser complementar às avaliações neurológicas, a mensuração da FPM contribui com a Fisioterapia para estabelecimento do programa de exercícios dos pacientes com DP [10]. Em populações idosas, a força de preensão também tem sido associada à velocidade da marcha, indicando uma possível relação entre a função muscular da mão e o desempenho físico global [11].

Adicionalmente, a eletromiografia de superfície (EMG) pode fornecer informações sobre o declínio do desempenho funcional para atividades de vida diária ao identificar alterações na atividade muscular [12]. Entretanto, os exames que combinam EMG e dinamometria exigem frequentemente equipamentos distintos, de alto custo e operação complexa, dificultando a realização de testes rápidos e padronizados em consultórios ou ambientes domésticos. Além disso, a maioria dos dinamômetros comerciais, como o Jamar, mede apenas a força global de preensão da mão, sem discriminar a contribuição individual de cada dedo. Quando utilizados de forma isolada, estes dispositivos podem mascarar déficits motores específicos e assimetrias finas importantes para o diagnóstico precoce [4].

Diante desse contexto, o presente estudo tem como objetivo apresentar uma proposta de arquitetura para uma manopla instrumentada multimodal para aquisição integrada de dados de força muscular, tremor e oscilação da mão e ativação mioelétrica, visando apoiar a avaliação motora da mão de pessoas com DP. O trabalho descreve a divisão da solução em subsistemas, o desenvolvimento em andamento e as etapas previstas para validação da arquitetura. Este é o primeiro relato científico do projeto, cujo pedido de patente já foi depositado. Apesar de estarem presentes na arquitetura, os subsistemas de sincronização entre os módulos, a eletromiografia e a inteligência artificial permanecem como etapas futuras. Sua aplicação ao diagnóstico ou ao monitoramento clínico dependerá de validação específica futura.

II. Solução Proposta

A. Organização funcional da arquitetura

A arquitetura proposta neste trabalho para o desenvolvimento de uma manopla instrumentada destinada à avaliação motora da mão de pessoas com DP é dividida em oito subsistemas (Fig. 1). A estrutura mecânica define a empunhadura, o apoio dos dedos e a transmissão das cargas. Os subsistemas de força por dedo e preensão global medem, respectivamente, os esforços locais de cada dedo da mão e a reação agregada no suporte do dispositivo. O sensoriamento inercial registra aceleração e velocidade angular, enquanto a EMG fornece informações sobre a ativação muscular. Os subsistemas de aquisição e comunicação associam referência temporal aos sinais e os transmitem ao computador. A interface com o usuário e a gestão de dados organizam as sessões, a visualização e o armazenamento das informações. O subsistema de inteligência artificial (IA) utiliza os registros disponíveis para análises multimodais.

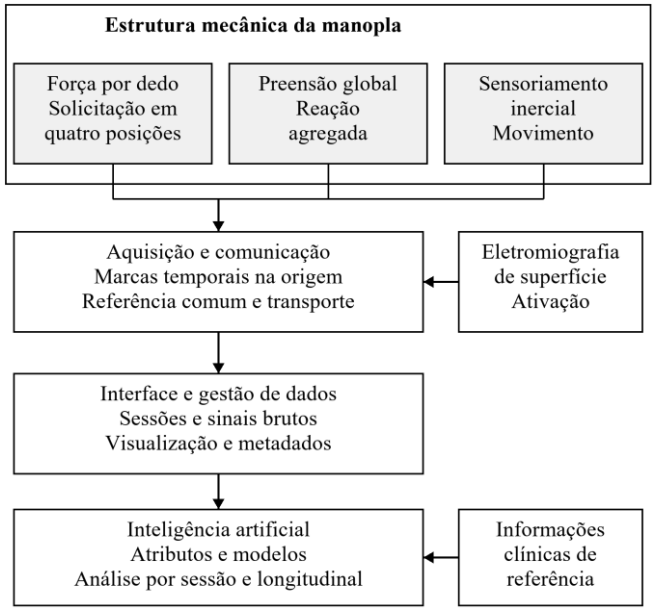


Fig. 1. Arquitetura funcional proposta. As setas representam o fluxo previsto dos dados, não integração já validada. A estrutura mecânica recebe os sensores de força e movimento. O estágio de desenvolvimento está discriminado na Tabela I.

As funções dos subsistemas não dependem de marcas específicas de sensores, conversores ou processadores. Cada módulo deverá informar os canais, grandezas, unidades de medida, taxas, configuração e identificação da calibração. Diferentes transdutores de força, células de carga e unidades inerciais poderão ser utilizados, desde que atendam aos requisitos mecânicos, elétricos e metrológicos. A substituição

de componentes exigirá nova caracterização e verificação das interfaces.

A medição da força por dedo e a medição da preensão global possuem papéis complementares. A primeira tem como objetivo quantificar a contribuição percentual de cada dedo instrumentado para a força total de preensão, tomando como referência a medida fornecida pelo sensor de preensão global, e acompanhar como essas contribuições variam durante a tarefa. Essa análise depende intrinsicamente da calibração dos sensores e da caracterização da transmissão das cargas, pois diferenças entre os caminhos mecânicos e as respostas dos transdutores podem fazer com que a soma das forças individuais difira da força global medida. O sensoriamento inercial por sua vez, fornecerá informações sobre a dinâmica do movimento, enquanto a eletromiografia de superfície (EMG) permitirá investigar sua relação com a ativação muscular. A integração dessas medidas possibilitará estudar a distribuição da força entre os dedos e sua relação com o movimento e a atividade muscular durante a realização de uma tarefa de preensão.

B. Marcação temporal das amostras

Para integrar os sinais dos diferentes subsistemas, a arquitetura contempla uma unidade de aquisição com uma referência de tempo comum, baseada em um contador monotônico, isto é, que avança sem retroceder durante a aquisição. Cada amostra de sinal receberá uma marca temporal nessa unidade, associada ao evento de conversão ou à indicação de que o sensor disponibilizou um novo dado. Nos canais adquiridos sequencialmente por multiplexação, será registrado o instante correspondente a cada conversão, sem considerar simultâneas as medidas de um mesmo conjunto de canais. Quando não for possível identificar diretamente o instante da conversão, será registrado o momento da leitura, e o atraso de conversão, quando caracterizado, será considerado na análise temporal.

Cada registro deverá conter a identificação da sessão, o tipo de sinal e o canal, o número de sequência da amostra, sua marca temporal, o valor bruto e os indicadores de qualidade. As informações de configuração, unidades de medida, calibração e versão do programa serão armazenadas como metadados da sessão. Os dados adquiridos serão mantidos temporariamente em filas, de modo a desacoplar a leitura dos sensores das etapas de transmissão e atualização dos gráficos. Os instantes de chegada dos dados ao computador podem ser utilizados para avaliar a comunicação, mas não substituem a marcação temporal da aquisição, uma vez que estes instantes também refletem os atrasos de transmissão e processamento no computador.

O mecanismo de marcação temporal com referência de tempo comum integra a arquitetura proposta, mas ainda não foi implementado. Sua implementação e validação estão previstas para etapas posteriores ao desenvolvimento descrito neste trabalho.

C. Subsistema de inteligência artificial

O subsistema de inteligência artificial (IA) é concebido como uma camada independente do dispositivo de aquisição, permitindo avaliar diferentes modelos de machine learning sem modificar a cadeia de sensoriamento. Seu funcionamento prevê o uso de dados organizados por sessão de avaliação, incluindo os registros originais dos sensores, as informações de calibração e, quando disponíveis, os resultados de avaliações clínicas de referência. O processamento da informação compreenderá a verificação da qualidade dos registros, a identificação dos trechos correspondentes a cada tarefa e a extração de características dos sinais que servirão como entradas para os modelos. Os resultados por sua vez, deverão ser acompanhados de informações que auxiliem sua interpretação pelo pesquisador ou profissional de saúde. Para o treinamento e a avaliação de modelos supervisionados destinados a estimar aspectos da gravidade motora, serão necessárias avaliações clínicas de referência correspondentes aos registros.

As possibilidades previstas incluem investigar a associação dos sinais com a gravidade do comprometimento motor da mão, identificar padrões de execução das tarefas e acompanhar mudanças entre sessões. A partir dos sinais de força calibrados, poderão ser analisados os valores de pico, a variabilidade da força global, o tempo até atingir o pico e a contribuição percentual de cada dedo instrumentado em relação à força de preensão global, considerando a caracterização da transmissão das cargas. Dos sinais inerciais, poderão ser extraídas medidas de amplitude, frequência dominante e potência em faixas de frequência de interesse. Se os sinais de eletromiografia de superfície (EMG) estiverem disponíveis, poderão ser analisados o envelope do sinal, que representa a variação de sua amplitude ao longo do tempo, e sua relação temporal com a força. As análises conjuntas deverão considerar o alinhamento temporal entre os canais e as limitações de cada subsistema para registrar variações rápidas dos sinais.

Para estimar gravidade motora e identificar padrões de execução, serão comparados modelos de machine learning baseados nas características extraídas dos sinais. Entre os métodos candidatos estão Regularized Regression, Random Forest, Gradient Boosting e Support Vector Machines (SVM), conforme o desfecho e os dados disponíveis. Change-Point Detection poderá apoiar análises entre sessões; Deep Learning dependerá de uma base com quantidade e diversidade adequadas. O método SHAP (SHapley Additive exPlanations) poderá auxiliar a interpretação da contribuição das características para as predições [13], sem estabelecer causalidade.

D. Desenvolvimento atual e etapas futuras

A arquitetura proposta ainda não foi integralmente implementada. O desenvolvimento vem sendo conduzido por subsistema, com montagens experimentais destinadas à caracterização e validação de cada bloco antes da integração do conjunto.

Atualmente estão em desenvolvimento a estrutura mecânica e os subsistemas de força por dedo, preensão global e sensoriamento inercial, com eletrônica, programas embarcados e interfaces experimentais. A leitura e a visualização dos sinais apoiam os testes, mas ainda não demonstram validação metrológica ou integração concluída. A Tabela I apresenta as frentes em andamento e as próximas etapas.

A sequência de desenvolvimento prevê a caracterização e a validação de cada subsistema, seguidas da integração dos fluxos de dados com a referência temporal comum e da avaliação do funcionamento conjunto. A incorporação da eletromiografia de superfície e o desenvolvimento do subsistema de inteligência artificial estão previstos para etapas posteriores.

TAELA I. ESTÁGIO DE DESENVOLVIMENTO DOS SUBSISTEMAS

Subsistema	Situação atual	Próxima etapa
Estrutura mecânica	Montagem experimental em desenvolvimento	Caracterização mecânica e ergonomia
Força por dedo	Quatro canais resistivos e aquisição experimental	Calibração individual e resposta dinâmica
Preensão global	Célula de carga e aquisição experimental	Calibração na montagem e banda útil
Sensoriamento inercial	Aquisição e análise exploratória	Calibração e validação das métricas
Aquisição e comunicação	Programas e placas independentes	Integração e referência temporal comum
Interface e dados	Interfaces específicas em desenvolvimento	Registro multimodal bruto e metadados
Eletromiografia	Subsistema futuro	Projeto da cadeia e segurança
Inteligência artificial	Subsistema futuro	Base rotulada e avaliação de modelos

A validação da arquitetura deverá verificar o atendimento das funções, interfaces e fluxos aos requisitos do sistema integrado. A validação clínica terá objetivos e protocolos próprios e dependerá de estudos com participantes.

III. MATERIAIS E MÉTODOS

A. Abordagem de desenvolvimento

O desenvolvimento adota prototipação e verificação por subsistema, apoiadas por modelos mecânicos, esquemas eletrônicos, programas embarcados e interfaces de aquisição. As configurações a seguir descrevem escolhas experimentais, que não devem ser confundidos com requisitos exclusivos da arquitetura.

B. Construção mecânica

A manopla foi modelada no Fusion 360 com corpo cilíndrico e quatro teclas, correspondentes aos dedos indicador, médio, anular e mínimo. Corpo e teclas foram impressos em plástico ABS por deposição de material fundido através de impressora 3D FDM (do inglês fused deposition modeling). O dedo polegar não possui canal individual de força.

Uma peça de alumínio, também modelada no Fusion 360 e fabricada em centro de usinagem com comando numérico computadorizado (CNC), acomoda os sensores de força por dedo e se acopla à célula de carga de preensão global. Um compartimento interno na base da manopla foi projetado para alojar uma unidade de medição inercial. A Figura 2 apresenta os elementos construtivos e a montagem experimental da manopla.

No interior da manopla, o suporte de alumínio recebe e transmite à célula de carga os esforços aplicados às teclas. A caracterização dessa ligação deverá avaliar a transmissão das cargas, a influência de esforços fora da direção de medição prevista, a interferência entre as posições dos dedos e a repetibilidade das respostas. As dimensões finais, a massa, a faixa admissível de esforço e os mecanismos de proteção contra sobrecarga serão definidos ao longo do desenvolvimento mecânico. A adequação da geometria a diferentes tamanhos de mão também deverá ser avaliada, orientando eventuais ajustes e alternativas de empunhadura a partir de estudos de usabilidade.

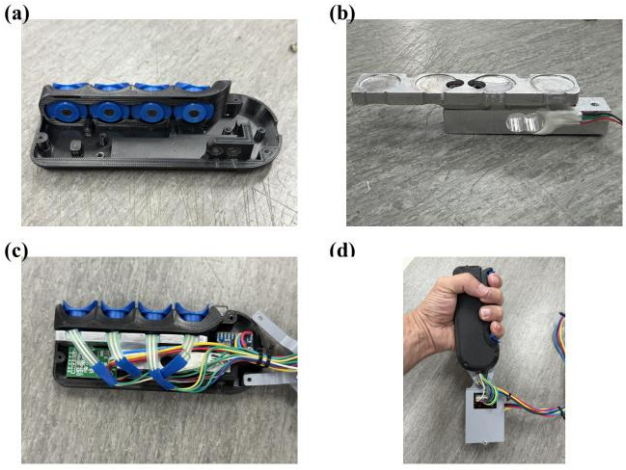


Fig. 2. Montagem experimental: (a) teclas e sensores resistivos; (b) suporte de alumínio e célula de carga; (c) eletrônica e cabeamento; (d) manopla empunhada. Fotografias do desenvolvimento, sem caracterização de ensaio clínico.

C. Aquisição da força por dedo

Para a medição da força exercida por cada dedo, a montagem atual utiliza quatro sensores resistivos de força (FSR, do inglês force-sensing resistors), modelo FSR402. Cada sensor está conectado entre um nó de leitura e o terra, enquanto um resistor de 22 kΩ conecta esse nó à alimentação de 3,3 V, formando um divisor de tensão. Um capacitor de 10 nF liga o nó de leitura ao terra. Um amplificador MCP6004, configurado como seguidor de tensão, reduz a interação entre a impedância do divisor e a entrada do conversor. Sua saída passa por um resistor de 220 Ω, seguido de outro capacitor de 10 nF conectado ao terra, para filtragem do sinal de saída, conforme apresentado na Figura 3.

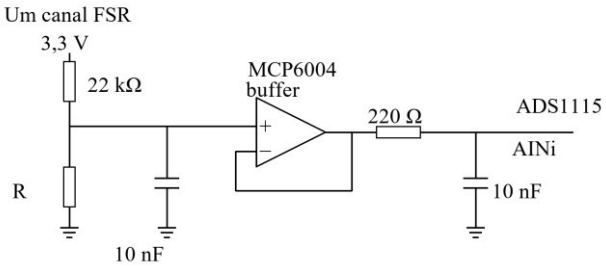


Fig. 3. Condicionamento de um canal FSR na implementação atual. R representa o sensor resistivo. O circuito é repetido para as quatro entradas do conversor ADS1115.

Desprezando o efeito de carga do seguidor de tensão sobre o divisor resistivo, a tensão do canal i é dada por $\mathbf{V_i = V_s R_i/(R_i + R_0)}$, em que $\mathbf{R_i}$ é a resistência do sensor, $\mathbf{V_s = 3{,}3\ V}$ é a tensão de alimentação e $\mathbf{R_0 = 22\ k\Omega}$ é a resistência fixa do divisor. À medida que a força aplicada ao sensor aumenta, sua resistência diminui, reduzindo a tensão medida. A conversão dessa resposta elétrica em força dependerá de curvas de calibração individuais, obtidas com os sensores instalados na manopla, pois a resposta dos FSR depende das condições de aplicação da carga [14].

As tensões dos quatro sensores são digitalizadas por um conversor analógico-digital ADS1115 (ADC, do inglês analog-to-digital converter), que possui resolução nominal de 16 bits e está configurado para uma faixa de ±4,096 V e uma

taxa de 860 conversões por segundo [15]. As quatro entradas são lidas sequencialmente por multiplexação.

Uma placa ESP32 DevKit V1 controla a aquisição e recebe as leituras digitais do ADS1115 pelo barramento I²C (Inter-Integrated Circuit). O ESP32 organiza a aquisição em ciclos nominais de 20 ms, equivalentes a 50 quadros por segundo. A cada ciclo, lê os quatro canais, aplica um filtro de mediana de cinco amostras por canal e converte os resultados em tensão.

O ESP32 envia ao computador, por comunicação serial via USB (Universal Serial Bus), registros CSV (comma-separated values) com o tempo local em milissegundos e as amostras filtradas das quatro tensões. O tempo é obtido no início do ciclo e identifica o quadro, não o instante de cada conversão.

Uma aplicação desenvolvida em Python recebe esses registros e apresenta os valores instantâneos e a evolução temporal dos quatro canais. A implementação utiliza as bibliotecas pySerial para comunicação, PySide6 para a interface gráfica e PyQtGraph para os gráficos. A interface permite visualizar as tensões medidas ou a diferença entre a tensão de repouso de cada canal e sua leitura instantânea. A referência de repouso é estimada durante aproximadamente dois segundos e pode ser redefinida pelo usuário com a manopla em repouso.

A Fig. 4 mostra as diferenças de tensão dos quatro canais durante o acionamento das teclas. Embora a interface use o rótulo “pressão”, os valores estão em volts e ainda não representam força calibrada ou contribuições percentuais para a preensão global.



Fig. 4. Interface Python dos quatro canais FSR. As curvas mostram a diferença entre a tensão de repouso e a tensão medida, em volts, ao longo do tempo.

D. *Aquisição da preensão global*

Complementando as medidas por dedo, uma célula de carga extensométrica registra os esforços transmitidos pelo suporte de alumínio. Seu sinal diferencial em ponte de Wheatstone é digitalizado por um módulo HX711, conversor de resolução nominal de 24 bits, configurado no canal A com ganho 128 [16]. Uma placa ESP32 DevKit V1 recebe as leituras e as transmite ao computador por comunicação serial via USB.

O HX711 está configurado para 10 amostras por segundo. O programa embarcado no ESP32 verifica a disponibilidade de novas leituras e adota intervalo mínimo de 100 ms entre envios. Cada registro contém o tempo local do ciclo, em milissegundos, a leitura bruta e o valor corrigido pela tara, polaridade e fator de escala.

A tara define a referência sem esforço aplicado, e a convenção de polaridade torna positiva a resposta à preensão. Um fator de escala permite converter as leituras numéricas em força, mas deverá ser determinado por calibração com cargas de referência na montagem. Até esse momento, a calibração ainda não foi realizada.

Uma aplicação Python, desenvolvida com as mesmas bibliotecas de comunicação e visualização do subsistema anterior, apresenta na tela do computador: o sinal ao longo do tempo, o valor instantâneo e o pico da sessão. A aplicação permite também selecionar a porta de comunicação serial, executar a tara, ajustar o fator de escala e redefinir o pico. Na Fig. 5, o fator unitário utilizado mantém os valores em contagens corrigidas apenas pela tara e polaridade. O rótulo “unidades calibradas” da interface ainda não representa calibração física concluída, necessária para apresentação do sinal em newtons.

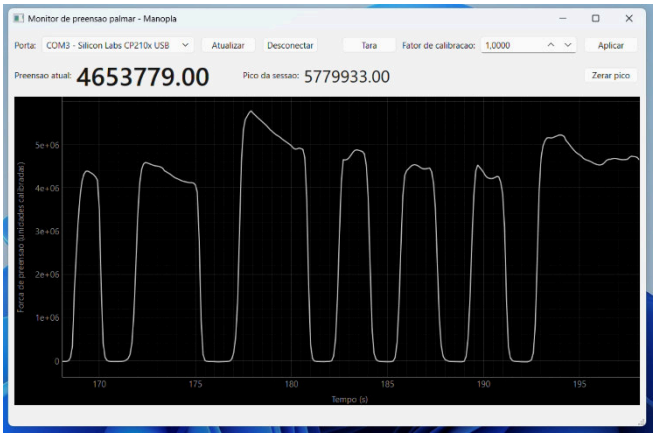


Fig. 5. Interface Python da preensão global, com resposta temporal, valor instantâneo, pico e controles de tara e escala. Os valores são contagens corrigidas, sem conversão para newtons.

E. *Sensoriamento inercial e processamento exploratório*

Complementando as medidas de força, o subsistema inercial tem como objetivo quantificar características dos tremores da mão, como amplitude e frequência, durante as tarefas de preensão. Para isso, registra as acelerações e velocidades angulares da manopla decorrentes dos movimentos transmitidos pela mão. A interpretação desses sinais deverá considerar a influência da empunhadura e do acoplamento mecânico entre a mão e o dispositivo.

A montagem utiliza uma unidade de medição inercial (IMU, do inglês inertial measurement unit) identificada como MPU-9250, da qual são adquiridos três eixos de aceleração e três de velocidade angular. As faixas configuradas são de ±4 g e ±500°/s, respectivamente, com resolução nominal de 16 bits. Uma placa ESP32 DevKit V1, dedicada a esse subsistema na montagem experimental atual, configura o sensor e recebe suas leituras pelo barramento I²C a 100 kHz, com aquisição programada para 100 Hz. Os filtros internos estão configurados para bandas de aproximadamente 44,8 Hz para aceleração e 41 Hz para velocidade angular [17]. Esses valores representam configurações nominais, não resultados de calibração dinâmica do protótipo.

O ESP32 converte as leituras digitais em valores de aceleração, expressos em g, e velocidade angular, em graus

por segundo, utilizando os fatores de escala nominais do sensor. Os dados são transmitidos ao computador por comunicação serial via USB, em registros CSV contendo uma referência de tempo local em milissegundos e os valores dos seis eixos. Essa referência é registrada após a leitura dos dados e ainda não corresponde ao mecanismo de sincronização entre subsistemas previsto na arquitetura.

Assim como nos subsistemas de força, uma aplicação desenvolvida em Python permite acompanhar os sinais durante a aquisição. A implementação utiliza pySerial para comunicação, PySide6 para a interface gráfica, PyQtGraph para visualização e NumPy e SciPy para processamento numérico e filtragem. A aplicação calcula o módulo da aceleração a partir dos três eixos e subtrai o valor de 1 g, produzindo o sinal utilizado na análise exploratória. Sobre esse sinal, aplica um filtro Butterworth passa-faixa de 3 a 12 Hz, configurado com parâmetro de ordem 4 e processamento bidirecional sobre uma janela de até dez segundos.

A partir do sinal filtrado, a aplicação calcula a raiz da média quadrática (RMS, do inglês root mean square) em uma janela de um segundo, utilizada como medida de sua amplitude. Também estima a frequência dominante na faixa analisada por meio da transformada rápida de Fourier (FFT, do inglês fast Fourier transform), aplicada a um trecho de quatro segundos com janela de Hann. Para esse trecho completo, considerando a taxa nominal de 100 Hz, a separação entre os pontos do espectro é de 0,25 Hz.

A interface gráfica apresenta o módulo da aceleração com subtração de 1 g, sua versão filtrada e a evolução do RMS, além dos valores de amplitude e frequência dominante. A Figura 6 ilustra a utilização da manopla com acompanhamento dos sinais na aplicação. As categorias de intensidade exibidas pela interface são baseadas em limiares experimentais definidos no programa e não correspondem a uma classificação clínica validada.



Fig. 6. Manopla em uso experimental e aplicação Python ao fundo, com sinal derivado da aceleração, versão filtrada entre 3 e 12 Hz e amplitude RMS.

F. Organização das unidades de aquisição

Cada um dos três subsistemas anteriormente apresentados utiliza um ESP32 dedicado para facilitar seu desenvolvimento e caracterização independente. Essa escolha permite ajustar circuitos, programas e interfaces sem interferir nos demais módulos, mas não é um requisito da arquitetura. As referências temporais locais descritas anteriormente ainda não implementam sincronização entre subsistemas.

Na integração final, será avaliada a possibilidade de utilização de um único ESP32 para acessar os três subsistemas, centralizar a marcação temporal e comunicar-se com o computador. A decisão dependerá das interfaces disponíveis, dos tempos de leitura e processamento e da capacidade de atender às taxas previstas sem perdas, incluindo a verificação da referência temporal comum.

IV. RESULTADOS E DISCUSSÃO

O principal resultado da etapa pretendida para este trabalho é a definição da arquitetura proposta e sua implementação experimental parcial, composta pela estrutura mecânica e pelos subsistemas de força por dedo, preensão global e sensoriamento inercial, com seus respectivos programas embarcados e interfaces de visualização. Essas implementações permitem acompanhar as respostas dos sensores e apoiar o desenvolvimento de cada subsistema. Entretanto, ainda não estão disponíveis curvas completas de calibração, estimativas de incerteza, medidas de erro de sincronização ou resultados clínicos que permitam considerar o sistema validado.

A organização modular da arquitetura permite avaliar separadamente o comportamento mecânico, a aquisição dos sinais e o processamento dos dados, preparando a integração do conjunto. Conforme descrito na seção anterior, o uso de um ESP32 dedicado a cada subsistema facilita essa etapa, mas os registros atuais ainda não possuem uma referência temporal comum validada. A integração deverá avaliar a possibilidade de centralizar a aquisição em um único ESP32 e implementar o mecanismo de marcação temporal proposto, considerando as diferentes taxas de aquisição e os atrasos de cada canal. A passagem do funcionamento independente para a operação coordenada dos subsistemas será, portanto, uma etapa central da validação da arquitetura.

A contribuição específica da proposta está em reunir, em uma mesma empunhadura, a medição da força por dedo, da preensão global e do movimento da manopla, com previsão de incorporação de EMG e análise por IA. Essa combinação pretende permitir o estudo da contribuição percentual de cada dedo instrumentado em relação à força global e de sua associação com os movimentos observados durante a tarefa. Em relação aos trabalhos sobre modulação da força [18] e luvas multimodais [8], a proposta apresenta uma configuração voltada à investigação da preensão por medidas complementares. Não se afirma, nesta etapa, que essa configuração ofereça maior sensibilidade ou melhor desempenho. O benefício da combinação dos sinais deverá ser investigado por meio da comparação entre análises baseadas em medidas isoladas e conjuntas.

As escolhas da implementação atual também impõem limitações que deverão orientar as próximas etapas. Os FSR exigem calibração individual e avaliação de histerese, fluência e condições de contato mecânico. A medição da preensão global depende da transmissão dos esforços até a célula de carga e da faixa de frequências registrada pelo conversor. As medidas inerciais, por sua vez, podem ser influenciadas pela fixação e orientação do sensor, pelo acoplamento entre a mão e a manopla e pelas opções de processamento. A própria geometria do dispositivo pode alterar a empunhadura e a distribuição dos esforços. Esses fatores deverão ser considerados no projeto mecânico, na caracterização dos sensores e na definição dos protocolos experimentais, sem pressupor que suas influências possam ser compensadas apenas por modelos computacionais [19].

.

V. VALIDAÇÃO PREVISTA E TRABALHOS FUTUROS

A. Caracterização e integração

A validação da proposta será conduzida em etapas, resumidas na Tabela II. Em bancada, cargas de referência serão aplicadas em diferentes níveis e posições, incluindo ciclos de carga e descarga, repetições e manutenção de esforço. Serão estimadas sensibilidade, repetibilidade, histerese, deriva e interferência entre canais. A célula de carga será caracterizada na montagem, com registro da faixa útil e da resposta à sobrecarga prevista em projeto. A IMU será avaliada quanto a escala, orientação e resposta dinâmica com referência apropriada. Cada resultado deverá incluir condições de ensaio e incerteza compatível com a finalidade de medida.

TAELA II. ESTÁGIO DE DESENVOLVIMENTO DOS SUBSISTEMAS

Etapa	Avaliação	Evidência para avançar
Bancada	Carga estática/dinâmica, interferência, deriva, proteção	Curvas e incerteza; faixa de operação documentada
Integração	Referência temporal, sequência, perdas e filtros	Atrasos medidos; lacunas identificadas; dados brutos preservados
Indivíduos Saudáveis	Repetição entre sessões/avaliadores e ergonomia	Confiabilidade e concordância; protocolo executável
Piloto Indivíduos com DP	Estágios I–III, estado medicamentoso, referência clínica	Tolerabilidade e qualidade dos dados; tamanho amostral revisto
Validação	Modelos por participante e avaliação externa	Desempenho com incerteza e domínio de aplicação definido

A integração verificará o contador comum, a correspondência entre marcas temporais e eventos de aquisição, os atrasos de conversão e a continuidade dos registros. Testes específicos de cada cadeia e aquisições prolongadas deverão quantificar defasagens, irregularidades e perdas. Os procedimentos e critérios de aceitação serão definidos antes dos ensaios, conforme as tarefas e características dos sinais que se pretende analisar.

B. Estudos com participantes e análise dos dados

Após os ensaios de bancada, estão previstos estudos de repetibilidade e usabilidade com voluntários saudáveis, seguidos de estudo piloto com pessoas com DP. O planejamento preliminar considera 15 a 30 participantes com DP em estágios I a III de Hoehn e Yahr e um grupo controle, com duas sessões separadas por aproximadamente uma semana. Essa previsão serve à análise de viabilidade; o dimensionamento definitivo dependerá do desfecho primário e da variabilidade observada. O protocolo deverá padronizar postura, mão avaliada, duração, repetições, repouso e instruções, registrando dominância, estado medicamentoso e condições que interfiram na execução.

A avaliação incluirá dinamometria de referência e itens pertinentes da MDS-UPDRS. A repetibilidade será examinada pelo coeficiente de correlação intraclasse; a concordância com a dinamometria, por análise de Bland–Altman, quando as medidas representarem a mesma grandeza, com intervalos de confiança. A associação com as avaliações clínicas será analisada separadamente, preferencialmente com avaliadores cegos aos resultados do dispositivo. Estudos longitudinais exigirão seguimento próprio; diferenças entre pacientes e controles não demonstrarão detecção pré-clínica.

Para a IA, treino e teste serão separados por participante, mantendo suas janelas e sessões no mesmo conjunto. Seleção de atributos, normalização e ajuste dos modelos ocorrerão apenas nos dados de treino, em validação cruzada aninhada. Para regressão serão avaliados erro absoluto médio e raiz do erro quadrático médio; para classificações justificadas pelo protocolo, sensibilidade, especificidade e área sob a curva característica de operação do receptor. Modelos multimodais serão comparados com modalidades isoladas e referências clínicas e demográficas. Os intervalos de confiança deverão respeitar o agrupamento por participante, e conclusões de generalização dependerão de avaliação externa ou temporal.

C. Eletromiografia, segurança e usabilidade

O subsistema de EMG ainda requer desenvolvimento. A cadeia prevista compreende eletrodos de superfície, amplificação diferencial, condicionamento, conversão e integração temporal. Número e posicionamento dos eletrodos, ganho, impedância de entrada, rejeição de modo comum, ruído, banda e taxa de amostragem serão definidos a partir dos músculos e tarefas de interesse. Também serão necessários filtragem antialias, proteção e avaliação da segurança da ligação elétrica com o participante. Nenhuma dessas especificações é apresentada atualmente como implementada.

A análise de riscos será orientada pela norma 14971 da Organização Internacional de Normalização (ISO, do inglês International Organization for Standardization) [20]. Será avaliada a aplicabilidade dos requisitos de segurança elétrica e compatibilidade eletromagnética da série 60601 da Comissão Eletrotécnica Internacional (IEC, do inglês International Electrotechnical Commission) [21], [22]. Não se declara conformidade nesta etapa. Serão considerados fixação, esforço excessivo, aquecimento, limpeza, contato com a pele, isolamento e correntes de fuga, inclusive nas conexões ao computador. A alimentação em baixa tensão não dispensa a avaliação do conjunto.

Avaliações iterativas com profissionais e usuários deverão examinar conforto, diferentes tamanhos de mão, lateralidade, facilidade de uso e higienização. Estudos com participantes dependerão de aprovação ética e consentimento, com pausas e critérios de interrupção por desconforto. Os dados deverão ser protegidos e vinculados a identificadores de pesquisa. O papel pretendido da plataforma é apoiar investigação e avaliação funcional; indicações diagnósticas ou terapêuticas dependerão de validações específicas posteriores.

VI. CONCLUSÃO

Este trabalho apresentou a arquitetura de uma manopla para aquisição multimodal durante tarefas de preensão e descreveu a implementação experimental da estrutura mecânica e dos subsistemas de força por dedo, preensão global e sensoriamento inercial. A organização modular permite avaliar diferentes componentes e desenvolver cada subsistema antes de sua integração.

As próximas etapas compreendem a caracterização metrológica, a integração temporal, a incorporação de EMG e o desenvolvimento da IA. A validação do conjunto e os estudos com participantes deverão estabelecer os limites de medição e a utilidade da plataforma para avaliação motora na DP, ainda não demonstrados nesta etapa.

REFERÊNCIAS

[1] J. Opara, A. Małecki, E. Małecka e T. Socha, “Motor assessment in Parkinson’s disease,” Ann. Agric. Environ. Med., vol. 24, no. 3, pp. 411–415, 2017, doi: 10.5604/12321966.1232774.

[2] R. P. Munhoz, V. Tumas, J. L. Pedroso e L. Silveira-Moriyama, “The clinical diagnosis of Parkinson’s disease,” Arq. Neuropsiquiatr., vol. 82, no. 6, art. s00431777775, 2024, doi: 10.1055/s-0043-1777775.

[3] V. Cabreira e J. Massano, “Doença de Parkinson: revisão clínica e atualização,” Acta Médica Portuguesa, vol. 32, no. 10, pp. 661–670, 2019, doi: 10.20344/amp.11978.

[4] J. Jankovic, “Parkinson’s disease: Clinical features and diagnosis,” J. Neurol. Neurosurg. Psychiatry, vol. 79, no. 4, pp. 368–376, 2008, doi: 10.1136/jnnp.2007.131045.

[5] C. Ramaker, J. Marinus, A. M. Stiggelbout e B. J. van Hilten, “Systematic evaluation of rating scales for impairment and disability in Parkinson’s disease,” Movement Disorders, vol. 17, no. 5, pp. 867–876, 2002, doi: 10.1002/mds.10248.

[6] C. G. Goetz et al., “Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results,” Movement Disorders, vol. 23, no. 15, pp. 2129–2170, 2008, doi: 10.1002/mds.22340.

[7] S. M. Ali et al., “Wearable accelerometer and gyroscope sensors for estimating the severity of essential tremor,” IEEE J. Transl. Eng. Health Med., vol. 12, pp. 194–203, 2024, doi: 10.1109/JTEHM.2023.3329344.

[8] Y. Li et al., “Learning hand kinematics for Parkinson’s disease assessment using a multimodal sensor glove,” Advanced Science, vol. 10, no. 20, art. 2206982, 2023, doi: 10.1002/advs.202206982.

[9] E. K. Saarinen et al., “Hand muscle strength in Parkinson’s disease: A Sarcopenic epiphenomenon or a meaningful biomarker?,” Parkinsonism & Related Disorders, vol. 140, art. 108021, 2025, doi: 10.1016/j.parkreldis.2025.108021.

[10] T. B. M. Vieira, G. F. de Oliveira, L. C. M. Silva, R. A. B. de Sousa, M. F. S. de Oliveira, and A. L. F. de Medeiros, "Reliability and validity of handgrip strength measurement in individuals with Parkinson's disease: A systematic review," Arch. Gerontol. Geriatr., vol. 129, Art. no. 105654, 2024, doi: 10.1016/j.archger.2024.105654.

[11] S. B. Cook, B. A. Chaudhry, C. L. Petersen, T. A. Mackenzie e J. A. Batsis, “Relationship of handgrip strength and asymmetry with walking ability in older adults with excess adiposity,” Advanced Biology, vol. 8, no. 10, art. e2400068, 2024, doi: 10.1002/adbi.202400068.

[12] G. R. Jones, K. P. Roland, N. A. Neubauer, and J. M. Jakobi, “Handgrip strength related to long-term electromyography: Application for assessing functional decline in Parkinson disease,” *Archives of Physical Medicine and Rehabilitation*, vol. 98, no. 2, pp. 347–352, 2017, doi: 10.1016/j.apmr.2016.09.133.

[13] S. M. Lundberg e S.-I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., vol. 30, 2017. Disponível: https://arxiv.org/abs/1705.07874.

[14] Interlink Electronics, “FSR 400 Series Data Sheet.” Disponível: https://www.interlinkelectronics.com/downloads/datasheets/fsr-400-series-datasheet.pdf. Acesso em: 13 set. 2026.

[15] Texas Instruments, “ADS111x ultra-small, low-power, I²C-compatible, 860SPS, 16-bit ADCs with internal reference, oscillator, and programmable comparator,” SBAS444E, rev. E, dez. 2024. Disponível: https://www.ti.com/lit/ds/symlink/ads1115.pdf.

[16] Avia Semiconductor, “HX711: 24-bit analog-to-digital converter (ADC) for weigh scales,” v. 2.0. Disponível: https://e2e.ti.com/cfs-file/__key/communityserver-discussions-components-files/73/HX711.pdf. Acesso em: 13 set. 2026.

[17] InvenSense, “MPU-9250 register map and descriptions,” RM-MPU-9250A-00, rev. 1.6, 2015. Disponível: https://invensense.tdk.com/wp-content/uploads/2015/02/RM-MPU-9250A-00-v1.6.pdf.

[18] S. Pradhan, R. Scherer, Y. Matsuoka e V. E. Kelly, “Grip force modulation characteristics as a marker for clinical disease progression in individuals with Parkinson disease: Case-control study,” Physical Therapy, vol. 95, no. 3, pp. 369–379, 2015, doi: 10.2522/ptj.20130570.

[19] A. Quattrocchi et al., “Handgrip strength in health applications: A review of the measurement methodologies and influencing factors,” Sensors, vol. 24, no. 16, art. 5100, 2024, doi: 10.3390/s24165100.

[20] International Organization for Standardization, “Medical devices—Application of risk management to medical devices,” ISO 14971:2019, 2019.

[21] International Electrotechnical Commission, “Medical electrical equipment—Part 1: General requirements for basic safety and essential performance,” IEC 60601-1:2005+A1:2012+A2:2020, ed. 3.2, 2020.

[22] International Electrotechnical Commission, “Medical electrical equipment—Part 1-2: General requirements for basic safety and essential performance—Collateral standard: Electromagnetic disturbances—Requirements and tests,” IEC 60601-1-2:2014+A1:2020, ed. 4.1, 2020.

E-SUS APS e a Engenharia de Sistemas de Saúde: Desafios e Perspectivas de um dos Maiores Prontuários Eletrônicos de Pacientes do Mundo

1st Joêmilly Wellen Lima da Silva
Departamento de Engenharia Biomédica (DEB)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0004-3464-0070

2nd Liana Lysia Cavalcanti Vieira de Brito
Escola de Ciências e Tecnologia (ECT)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0001-2928-7522

3rd Maria Eduarda da Silva Ferreira
Escola de Ciências e Tecnologia (ECT)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0006-5258-7757

4th Antonio de Oliveira Nunes Costa
Mestrando em Engenharia Elétrica e de Computação
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0000-0003-2970-6245

5th Ricardo Alexsandro de Medeiros Valentim
Professor Titular do Departamento de Engenharia Biomédica e LAIS
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0000-0002-9216-8593

***Resumo*—Este artigo analisa a estratégia e-SUS Atenção Primária (e-SUS APS) sob o paradigma da Engenharia de Sistemas de Saúde (ESS), integrando evidências de uma década de evolução do Prontuário Eletrônico do Cidadão (PEC). Através de um estudo descritivo e exploratório baseado em uma Síntese Técnico-Científica Multimetodológica, investigou-se a trajetória de informatização das Unidades Básicas de Saúde (UBS) no Brasil. Observou-se que a adesão ao PEC saltou de 8.930 unidades em 2017 para 26.091 em 2022, cobrindo 59,93% da rede primária. Os resultados demonstram que a robustez do sistema advém de uma arquitetura escalável e da aplicação de metodologias de User Experience (UX), embora persista a fragmentação com sistemas nacionais de vigilância, onde 15 de 31 sistemas ainda não possuem unificação de interfaces. Conclui-se que a Engenharia Biomédica é o pilar necessário para garantir a interoperabilidade semântica via RNDS e a transição do PEC de um registro burocrático para um ecossistema de suporte à decisão clínica.**

***Palavras-chave*—e-SUS APS, Engenharia de Sistemas de Saúde, Prontuário Eletrônico do Cidadão, Interoperabilidade, Engenharia Biomédica.**

I. Introdução

A criação do Sistema Único de Saúde (SUS), pela Constituição de 1988, consolidou a saúde como um direito universal e dever do Estado, fundamentado nos princípios de universalidade, integralidade e equidade [6]. Dentro dessa estrutura, a Atenção Primária à Saúde (APS) atua como centro ordenador e a porta de entrada preferencial da Rede de Atenção à Saúde (RAS), desempenhando papel crucial na redução de internações por condições sensíveis e na coordenação do cuidado no território [2].

Historicamente, a gestão da informação no SUS foi marcada por uma profunda fragmentação, caracterizada por "ilhas de informática" e dezenas de Sistemas de Informação em Saúde (SIS) heterogêneos e não integrados [4]. Essa desarticulação forçava o registro redundante de dados, gerando retrabalho para os profissionais e resultando em bases de dados consolidadas com baixa qualidade e confiabilidade, o que dificultava a análise epidemiológica e a tomada de decisão gerencial [4] [3].

Para enfrentar esses desafios, o Ministério da Saúde instituiu, em 2013, a estratégia e-SUS Atenção Primária (e-SUS APS) [7]. O objetivo central foi reestruturar o Sistema de Informação em Saúde para a Atenção Básica (SISAB), rompendo com a lógica milenar de dados puramente consolidados para adotar o registro clínico individualizado, fundamentado no Cartão Nacional de Saúde (CNS) [1]. O núcleo tecnológico dessa transformação é o Prontuário Eletrônico do Cidadão (PEC), uma ferramenta projetada para apoiar o processo de informatização das Unidades Básicas de Saúde (UBS), integrar ferramentas de apoio à decisão clínica e permitir o compartilhamento de informações entre serviços [1] [6].

Sob o paradigma da Engenharia de Sistemas de Saúde, o e-SUS APS transcende a definição de um simples software para se tornar um ecossistema sócio-tecnológico complexo [8]. Sua sustentabilidade depende da integração harmônica

entre infraestrutura tecnológica, governança de dados sensíveis e processos assistenciais dinâmicos [3]. Nesse cenário, a Engenharia Biomédica emerge como disciplina estratégica, atuando na mediação entre as demandas clínicas e as soluções tecnológicas [5]. A contribuição da engenharia é vital para transitar da mera "implantação" (disponibilização técnica) para a efetiva "incorporação" tecnológica, processo que envolve a adaptação contínua do sistema às necessidades do usuário e a garantia da segurança do paciente [5].

Ao completar uma década de implementação em 2023, o PEC consolidou-se como a principal ferramenta de registro eletrônico no Brasil, cobrindo a maioria dos municípios [2]. Contudo, a escalabilidade nacional trouxe novos desafios de interoperabilidade e resiliência de sistemas distribuídos [4] [10]. Este artigo analisa essa trajetória sob a ótica da engenharia, discutindo como a aplicação de princípios de requisitos, arquitetura de dados e padrões semânticos pode superar os gargalos que ainda fragmentam o cuidado no SUS [3].

II. Metodologia

Este trabalho caracteriza-se como um estudo descritivo e exploratório, com abordagem mista (quanti-qualitativa), fundamentado em uma síntese técnico-científica multimetodológica [1]. A investigação foi estruturada para analisar o e-SUS APS não apenas como uma ferramenta isolada, mas sob o paradigma da Engenharia de Sistemas de Saúde, que exige a compreensão das interações entre componentes tecnológicos, processos assistenciais e governança de dados [8]. As etapas desse desenvolvimento metodológico do artigo estão detalhadas na figura 1.



Fig. 1. Fluxograma da metodologia de estudo multimetodológica aplicada no artigo. Elaboração dos autores (2026).

A. Coleta de Dados e Fontes de Evidência

A coleta de dados abrangeu o período de 2013 a 2026, de modo a contemplar o histórico de consolidação do e-SUS APS e as diretrizes de interoperabilidade mais recentes estabelecidas para a Rede Nacional de Dados em Saúde (RNDS). O processo baseou-se em três eixos principais :

1. Dados Secundários Oficiais: Foram extraídas informações da base de dados da Secretaria de Atenção Primária à Saúde (SAPS), incluindo indicadores de adesão nacional ao PEC, informatização das Unidades Básicas de Saúde (UBS) e registros de envio de dados ao Sistema de Informação em Saúde para a Atenção Básica (SISAB) [2] [6].

2. Análise Documental com Critérios de Elegibilidade: Para o levantamento documental, adotou-se como critério de inclusão portarias ministeriais, manuais técnicos e normativas oficiais que regem a infraestrutura do e-SUS APS e a interoperabilidade na RNDS publicadas entre 2013 e 2026. Foram excluídos rascunhos de guias e documentos de consulta pública não homologados. A amostra final compreendeu 15 documentos técnicos estratégicos, incluindo as Portarias nº 1.412/2013 e nº 2.983/2019, o Plano Diretor de Tecnologia da Informação (PDTI) e os manuais técnicos do sistema das versões 3.2 a 5.2. [1] [9] [12].

3. Revisão Científica e Seleção de Literatura: Realizou-se uma busca bibliográfica nas bases LILACS, SciELO, PubMed e nos anais do XXIX Congresso Brasileiro de Engenharia Biomédica (CBEB). Os critérios de inclusão contemplaram artigos originais e revisões que abordavam usabilidade, arquitetura, integração ou impacto assistencial do e-SUS APS e do PEC. Foram excluídos trabalhos duplicados e estudos sem foco direto em sistemas de informação da Atenção Primária [13].

Para a análise temporal da adesão ao e-SUS APS, distinguem-se duas métricas complementares extraídas de bases de dados secundárias da Secretaria de Atenção Primária à Saúde (SAPS/MS):

Informatização Efetiva por PEC (2022): Refere-se à proporção de Unidades Básicas de Saúde (UBS) ativas com o Prontuário Eletrônico do Cidadão (PEC) instalado localmente e enviando dados regularmente ao SISAB, totalizando 26.091 UBS (59,93% do parque instalado), segundo dados do CNES/SAPS de dezembro de 2022.

Cobertura do Ecossistema e-SUS APS (2024): Expressa a taxa de adesão institucional global (97,6% das UBS), abrangendo unidades com PEC local, PEC Nuvem e sistemas próprios integrados cadastrados no Programa Informa e-SUS APS, com base no Painel e-Gestor AB (primeiro semestre de 2024).

B. Modelagem Sistêmica, Avaliação LCIM e Análise de Dados

Para traduzir o funcionamento do e-SUS APS sob a ótica da engenharia, aplicou-se a técnica de Modelagem de Modelo Lógico, baseada no referencial de Thurston e Ramaliu [1].Este modelo permitiu decompor o sistema em Estrutura (Insumos), Processos (Atividades) e Resultados e Impactos. Destaca-se

que as propostas de arquitetura e modelos de integração apresentados neste trabalho constituem modelagens conceituais sistêmicas e analíticas, e não validações empíricas testadas em bancada experimental.

- Estrutura (Insumos): Hardware, conectividade e pessoal qualificado [1].
- Processos (Atividades): Registro clínico via SOAP (Subjetivo, Objetivo, Avaliação e Plano) e fluxo de dados assistenciais [1].
- Resultados e Impactos: Redução do retrabalho, unificação de cadastros e ampliação da capacidade clínica [1].

A análise quantitativa foi realizada por meio de estatística descritiva, utilizando a classificação do IBGE para caracterização de espaços rurais e urbanos, visando identificar desigualdades regionais na maturidade digital assistencial [7]. A avaliação da interoperabilidade seguiu os critérios de conformidade com o padrão HL7 FHIR e a integração com os 31 Sistemas Nacionais de Informação em Saúde (SNIS) identificados na literatura [14].

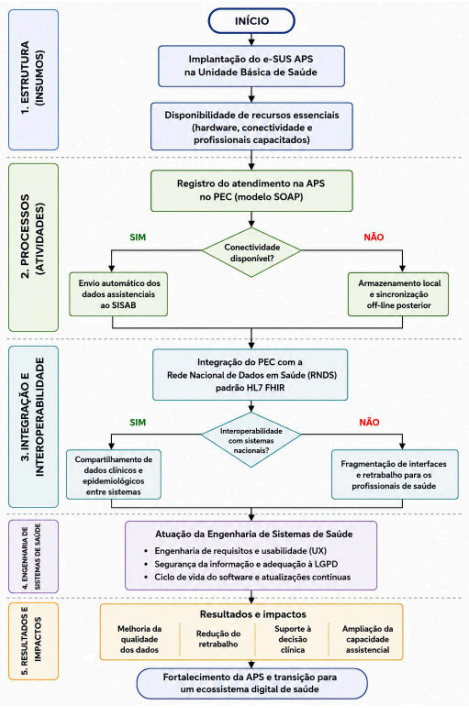


Fig. 2. Modelo lógico e fluxo sistêmico do e-SUS APS sob a perspectiva da Engenharia de Sistemas de Saúde. Elaboração dos autores (2026).

C. Aspectos Éticos e de Integridade

A avaliação do estágio de integração entre o e-SUS APS e os 31 Sistemas Nacionais de Informação em Saúde (SNIS) mapeados na literatura foi realizada de forma objetiva mediante a construção da Matriz de Interoperabilidade baseada no Levels of Conceptual Interoperability Model (LCIM). A pontuação dos sistemas nos quatro níveis do LCIM - Técnico/Fundacional (T), Sintático/Estrutural (S), Semântico (Se) e Organizacional (O). Foi realizada por meio de análise documental sistemática e triangulação de evidências secundárias (manuais de integração, documentação de APIs publicadas pelo DATASUS/MS e literatura técnico-científica), adotando-se uma escala ordinal de 0 a 4:

0 (Inexistente): Ausência de evidências técnicas ou documentais de integração.

1 (Inicial): Integração incipiente, dependente de exportações manuais ou dados em lote.

2 (Parcial): Trocas pontuais e parciais de dados com escopo restrito.

3 (Avançado): Integração estruturada com adoção de padrões e uso regular.

4 (Completo): Integração plena, padronizada e institucionalizada no fluxo de trabalho.

Com base nessas pontuações, calculou-se para cada sistema o Índice de Maturidade de Interoperabilidade (IMI), expresso em percentual (0 a 100 %), conforme a equação (1):

IMD (%) = [(T + S + Se + 0) / 16] * 100

Os sistemas foram subsequentemente categorizados em quatro Níveis de Maturidade: Muito Alto (IMI maior que 80 %), Intermediário (50 % menor que o IMI e menor ingual que 80 %), Baixo (25 % menor que o IMI e menor igual a 50 %) e Muito Baixo (IMI menor igual a 25 %).

TABELA M– Matriz de Interoperabilidade dos 31 Sistemas Nacionais de Informação em Saúde com o e-SUS APS (PEC) segundo o LCIM.

Nº	Sistema Nacional (Órgão Gestor)	Níveis de Interoperabilidade (LCIM)				Índice de Maturidade (0–100)	Nível de Maturidade
		Técnico/ Fundacional	Sintático/ Estrutural	Semântico	Organizacional		
1	PEC / e-SUS APS (DATASUS)	4	4	4	3	93,8	Muito Alto
2	CADSUS Web (DATASUS)	4	4	3	3	87,5	Muito Alto
3	CNES (DATASUS)	4	4	4	3	93,8	Muito Alto
4	CNS (DATASUS)	4	4	4	4	100,0	Muito Alto
5	SISAB (DATASUS)	4	4	4	3	93,8	Muito Alto
6	SINAN (SVSA/MS)	3	3	3	2	68,8	Intermediário
7	SIM (SVSA/MS)	3	3	3	2	68,8	Intermediário
8	SINASC (SVSA/MS)	3	3	3	2	68,8	Intermediário
9	SIH/SUS (SAPS/MS)	3	3	3	2	68,8	Intermediário
10	SIA/SUS (SAPS/MS)	3	3	3	2	68,8	Intermediário
11	SIPNI (SVSA/MS)	2	2	2	1	37,5	Baixo
12	SI-PNI Web (SVSA/MS)	3	3	2	2	62,5	Intermediário
13	e-SUS VE (SVSA/MS)	3	3	2	2	62,5	Intermediário
14	SIVEP-Gripe (SVSA/MS)	3	2	2	1	50,0	Baixo
15	SISVAN Web (SVSA/MS)	3	2	2	2	56,3	Intermediário
16	e-SUS Notifica (SVSA/MS)	2	2	1	1	25,0	Baixo
17	SISREG III (SAPS/MS)	3	3	2	2	62,5	Intermediário
18	e-Gestor AB (SAPS/MS)	2	2	2	1	37,5	Baixo
19	SISPACTO (SAPS/MS)	2	2	2	1	37,5	Baixo
20	SISPRENATAL (SAPS/MS)	2	2	1	1	25,0	Baixo
21	BPA-i (SAPS/MS)	2	2	1	1	25,0	Baixo
22	SCNES (DATASUS)	3	3	3	2	68,8	Intermediário
23	e-SUS Território (DATASUS)	2	2	2	1	37,5	Baixo
24	SAIPS (DATASUS)	2	2	1	1	25,0	Baixo
25	SISRO (DATASUS)	2	2	1	1	25,0	Baixo
26	SISFRONTEIRAS (SVSA/MS)	2	2	1	1	25,0	Baixo
27	SIHFRON (SVSA/MS)	2	2	1	1	25,0	Baixo
28	NOTA TÉCNICA – Influenza (SVSA/MS)	1	1	1	1	12,5	Muito Baixo
29	SI-Malária (SVSA/MS)	1	1	1	1	12,5	Muito Baixo
30	SISLEISH (SVSA/MS)	1	1	1	1	12,5	Muito Baixo
31	SIS-Água (SVSA/MS)	1	1	1	1	12,5	Muito Baixo

LEGENDA DOS NÍVEIS DE INTEROPERABILIDADE (LCIM)

0 – **Inexistente:** Não há evidências de integração.
1 – **Inicial:** Integração incipiente ou pilotos isolados; dependência de processos manuais.
2 – **Parcial:** Integração parcial com trocas limitadas de dados ou sem padronização consistente.
3 – **Avançado:** Integração estruturada com padrões adotados e uso regular.
4 – **Completo:** Integração plena, padronizada e institucionalizada no processo de trabalho.

CÁLCULO DO ÍNDICE DE MATURIDADE (IMI)

$$IMI = \frac{(T + S + Se + O)}{16} \times 100$$

Onde:
T = Técnico/Fundacional;
S = Sintático/Estrutural;
Se = Semântico;
O = Organizacional.

CLASSIFICAÇÃO DO NÍVEL DE MATURIDADE (IMI)

- **Muito Alto (81–100):** Interoperabilidade consolidada e institucionalizada.
- **Intermediário (51–80):** Interoperabilidade em desenvolvimento.
- **Baixo (26–50):** Integração limitada e não sistematizada.
- **Muito Baixo (0–25):** Ausência ou integração incipiente.

BASES E REFERÊNCIAS UTILIZADAS

- Ministério da Saúde. Portaria GM/MS nº 1.412/2013 [1]
- Ministério da Saúde. Portaria GM/MS nº 2.073/2011 [2]
- Ministério da Saúde. Decreto nº 7.508/2011 [3]
- Ministério da Saúde. e-SUS APS – Manual de PEC (versões 3.2 a 5.2) [4,6]
- DATASUS. Documentação técnica dos sistemas [1,4]
- Aguiar et al., 2023 – Interoperabilidade dos sistemas nacionais de informação [9]
- Alfradique et al., 2013 – Avaliação de sistemas de informação em saúde [10]

Elaborado pelos autores (2026).

Por utilizar exclusivamente dados secundários e anônimos de acesso público, este estudo está dispensado de apreciação pelo Comitê de Ética em Pesquisa, conforme a Resolução CNS 510/2016 [1]. A análise técnica manteve a neutralidade frente aos modelos de desenvolvimento (b-thinking) adotados pelos laboratórios parceiros do Ministério da Saúde [6].

O marco fundamental da infraestrutura de interoperabilidade no SUS foi a consolidação da Rede Nacional de Dados em Saúde (RNDS), fundamentada no padrão internacional HL7 FHIR. No entanto, a integração efetiva de interfaces com os demais Sistemas Nacionais de Informação em Saúde (SNIS) ainda enfrenta desafios de maturidade heterogênea. Para avaliar objetivamente o estágio de integração do e-SUS APS com os 31 SNIS mapeados, aplicou-se a matriz baseada no Levels of Conceptual Interoperability Model (LCIM), cujos resultados detalhados e pontuações por nível (Técnico, Sintático, Semântico e Organizacional) estão consolidados na Tabela M.

As propostas arquiteturais apresentadas neste estudo possuem caráter teórico e conceitual, sendo elaboradas a partir da análise da literatura científica e da documentação técnico-institucional selecionada. Dessa forma, os modelos propostos não correspondem a implementações em ambiente real nem foram submetidos a testes ou validações empíricas pelos autores. Sua finalidade é representar cenários possíveis de integração e apoiar a discussão sobre a evolução do ecossistema e-SUS APS. São propostas conceituais derivadas da análise, não resultados de uma implementação experimental.

D. Limitações do Estudo

Reconhecimento da Natureza do Estudo e Limitações:

Cumpre ressaltar, quanto à natureza metodológica do presente trabalho, que a análise do ecossistema e-SUS APS e a mensuração do Índice de Maturidade de Interoperabilidade (IMI) fundamentam-se exclusivamente no levantamento de dados secundários e na análise documental sistemática de portarias, normativas, manuais técnicos e documentações de APIs publicadas pelo Ministério da Saúde e DATASUS. Dessa forma, a classificação de não-integração ou de baixa maturidade atribuída aos sistemas nacionais de informação decorre estritamente do enquadramento desse arcabouço normativo, documental e arquitetural disponível até o presente momento. A caracterização de tais barreiras operacionais não constitui resultado de testes técnicos de interoperabilidade direta, como simulações de rotas de API, análises de pacotes em ambiente de homologação ou testes de bancada experimental.

III. Resultados

A. Evolução do e-SUS APS

O e-SUS APS consolidou-se como a principal ferramenta de registro eletrônico no Brasil. A adesão demonstrou crescimento acelerado após a instituição do programa Informatiza APS em 2019, que vinculou o financiamento à qualificação dos dados enviados ao SISAB [2] [3]. Esse cenário impulsionou a transformação digital do Sistema Único de Saúde (SUS), otimizando a gestão local e o monitoramento das políticas públicas de saúde. Como resultado, o sistema passou a subsidiar de forma mais robusta o planejamento estratégico e a continuidade do cuidado aos usuários na rede pública, fortalecendo a integração e a transparência das informações em saúde em todo o território nacional.

Tabela I
Evolução da adesão ao PEC nas Unidades Básicas de Saúde (2017–2022)

Região	UBS com PEC (2017)	UBS com PEC (2022)	Crescimento (%)
Norte	631	2.247	256,10%
Nordeste	2.618	12.245	367,72%
Centro-Oeste	1.014	2.201	117,06%
Sudeste	2.724	7.002	157,04%
Sul	1.943	2.396	23,31%
Brasil (Total)	8.930	26.091	192,17%

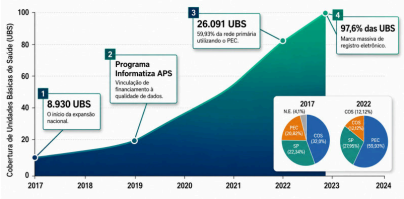


Fig. 3. Curva de crescimento e marcos de adesão nacional ao PEC (2017 - 2024). Elaboração dos autores (2026).

Os dados integrados da Tabela 1 e da Figura 3 revelam que a expansão do PEC não ocorreu de forma imediata ou homogênea no território nacional. Enquanto a região Nordeste apresentou o incremento mais acentuado em termos absolutos e percentuais (367,72%), saltando de 2.618 para 12.245 UBS conectadas, a região Sul manteve um crescimento gradual (23,31%), reflexo de uma cobertura assistencial previamente estabelecida. Esse avanço descentralizado resultou no marco histórico de 2024, atingindo 97,6% de cobertura nacional e consolidando a transição do modelo de Coleta de Dados Simplificada (CDS) para o Prontuário Eletrônico do Cidadão (PEC) sendo considerado o canal principal de melhoria no sistema de saúde da APS.

A evolução observada entre a taxa de 59,93% de informatização efetiva por PEC em 2022 (26.091 UBS) e a cobertura de 97,6% atingida em 2024 reflete uma transição metodológica e estrutural na infraestrutura da Atenção Primária. Em 2022, a métrica refletia primariamente a instalação local do software PEC em servidores físicos municipais, processo limitado por barreiras de conectividade e aquisição de hardware. A aceleração verificada até 2024 justifica-se por três fatores indutores:

(i) a consolidação do Programa Informa e-SUS APS, que vinculou os repasses financeiros federais de custeio ao envio contínuo de dados;

(ii) a expansão da modalidade PEC Nuvem e de sistemas municipais próprios interoperáveis via APIs REST/Thrift; e

(iii) a disseminação de aplicativos móveis com sincronização assíncrona (e-SUS Território), viabilizando o registro digital em áreas remotas.

Dessa forma, o incremento temporal traduz a passagem de um modelo focado na infraestrutura física local para uma cobertura ampliada do ecossistema de dados da APS.

- Mudança de Paradigma (2025): Em junho de 2025, o Sistema de Informação para a Atenção Primária à Saúde (SIAPS) substituiu oficialmente o antigo Sistema de Informação em Saúde para a Atenção Básica (SISAB) como a plataforma central da APS digital. Essa transição visou superar gargalos de performance e latência que limitavam a análise de dados em tempo real nas décadas anteriores. O novo sistema focou principalmente em uma arquitetura de melhor usabilidade e confiabilidade, possuindo integração nativa com a Rede Nacional de Dados em Saúde (RNDS) [6]. Esse ecossistema atingiu cerca de 60 milhões de downloads e 29 milhões de usuários ativos mensais no aplicativo Meu SUS Digital em 2025 [6].
- Perspectiva 2026: O Guia de Informática em Saúde de 2026 já define a interoperabilidade semântica via FHIR R4 (Fast Healthcare Interoperability Resources, Release 4) como o padrão estável e obrigatório para mandatos nacionais, consolidando a RNDS como o repositório de mais de 2,8 bilhões de registros clínicos [11].

B. Arquitetura e Engenharia de Sistemas de Saúde

A arquitetura do PEC evoluiu para suportar a complexidade de 5.570 municípios. Inicialmente baseado em H2, o sistema migrou para PostgreSQL e Oracle, garantindo escalabilidade nacional [6]. O desenvolvimento adota a metodologia b-thinking, que integra desenvolvimento ágil, User Experience (UX) e privacy by design [6]. O registro clínico é estruturado no formato SOAP (Subjetivo, Objetivo, Avaliação e Plano) [1].

C. O Paradoxo da Cobertura em Saúde e o Retrabalho Assistencial na Ponta

A análise integrada dos indicadores de expansão e maturidade revela uma importante assimetria estrutural na saúde digital brasileira: uma alta taxa de cobertura de implantação do PEC (97,6%) não se traduz, necessariamente, em uma interoperabilidade real de dados clínicos na rotina diária[1]. Embora a expansão do ecossistema e-SUS APS tenha sido praticamente universalizada no parque instalado das Unidades Básicas de Saúde (UBS), a troca fluida e automatizada de dados com os demais Sistemas Nacionais de Informação em Saúde (SNIS) permanece altamente fragmentada[1] [6].

A aplicação da matriz LCIM demonstra quantitativamente esse paradoxo: enquanto a cobertura física do sistema atinge patamares universais, a interoperabilidade de alta maturidade (nível Muito Alto, IMI maior que 80%) restringe-se a apenas 5 dos 31 sistemas mapeados (16,1%), limitando-se ao núcleo cadastral básico (como CADSUS e CNES)[cite: 1, 5]. Por outro lado, 16 dos 31 sistemas analisados (51,6%) situam-se nas faixas de maturidade Baixa e Muito Baixa (IMI menor igual que 50%), operando de forma isolada e sem comunicação automatizada com o prontuário eletrônico[1] [5].

Na rotina diária das equipes de Saúde da Família, esse isolamento do bloco de 16 sistemas (que inclui o e-SUS Notifica, SIVEP-Gripe, SI-Malária, SISLEISH, SIPNI e BPA-i) impacta diretamente os profissionais de saúde na ponta, gerando um expressivo gargalo de retrabalho assistencial[1] [5]

- Dupla Digitação e Sobrecarga Burocrática: Ao realizar um atendimento clínico ou procedimento (como a notificação de um caso suspeito de arbovirose ou o registro vacinal), o profissional de saúde é obrigado a registrar a evolução clínica no PEC e, em seguida, reescrever as mesmas informações sociodemográficas e epidemiológicas em portais e sistemas web paralelos da vigilância em saúde[1] [5].
- Carga Cognitiva vs. Tempo de Atendimento: A necessidade de navegar por múltiplas interfaces com diferentes credenciais de acesso consome um tempo precioso do atendimento, reduzindo a disponibilidade de tempo efetivo para a escuta qualificada, anamnese e exame físico do paciente.
- Riscos à Qualidade dos Dados e Subnotificação: A repetição manual do lançamento de dados em interfaces distintas aumenta exponencialmente o risco de erros de transcrição e divergências de cadastro. Ademais, em momentos de sobrecarga de demanda na UBS, a rotina burocrática força os profissionais a postergar ou omitir o preenchimento nos sistemas secundários isolados, gerando subnotificação e comprometendo a fidelidade dos indicadores epidemiológicos nacionais.

Portanto, o avanço da saúde digital na Atenção Primária não pode mais ser medido apenas pela distribuição e cobertura física de software[1] [6]. Superar o retrabalho dos profissionais da ponta exige que a Engenharia de Sistemas de Saúde direcione esforços para a interoperabilidade semântica e sintática via padrões como o HL7 FHIR R4 na RNDS, assegurando que o dado registrado uma única vez no PEC alimente de forma transparente e assíncrona todo o ecossistema público de saúde[1] [5].

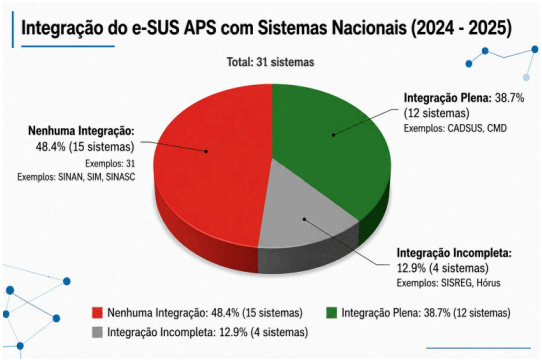


Fig. 4. Situação de integração de interfaces entre o e-SUS APS e os Sistemas Nacionais de Informação em Saúde (SNIS) do ano de 2024 a 2025 [4]. Elaboração dos autores (2026).

Como observado na Figura 4, a respeito do avanço na cobertura técnica, a persistência de 15 dos 31 sistemas nacionais (48,4%) operando com severas barreiras de integração

demonstra que o ecossistema de saúde digital brasileiro ainda opera em silos de informação. Para a Engenharia de Sistemas de Saúde, este dado reforça a urgência da implementação de padrões semânticos e da RNDS para superar o retrabalho dos profissionais na ponta e garantir a integridade dos dados clínicos.

D. Contribuições da Engenharia de Sistemas de Saúde

A Engenharia de Sistemas de Saúde pode contribuir para a sustentabilidade e eficiência dos processos nas instituições de saúde, especialmente por meio da integração entre aspectos clínicos, tecnológicos e de gestão. Nesse contexto, um estudo de caso realizado em 2025 no Hospital Monsenhor Walfredo Gurgel (HMWG), em Natal/RN, relatou um incremento de 23,47% no faturamento por Autorização de Internação Hospitalar (AIH) após a substituição de fluxos analógicos por processos apoiados em prontuário digital [18]. No período analisado, o faturamento passou de 19,4 milhões de reais para 23,9 milhões de reais ao longo de oito meses [18]. Esse resultado sugere uma possível associação entre a qualificação dos registros clínicos e a melhoria dos processos de faturamento; entretanto, por se tratar de um único estudo de caso, não permite estabelecer relação causal nem generalizar esse efeito para outras instituições. A Engenharia de Sistemas atua com suma importância na transição da implantação (instalação técnica) para a incorporação (processo dinâmico de treinamento e adaptação) [5]. As principais contribuições incluem:

- Engenharia de Requisitos: Adaptação das interfaces para cenários de baixa conectividade, resultando em aplicativos móveis com sincronização offline (e-SUS Território) [6].
- Segurança e Resiliência: Implementação de autenticação forte (MFA) e criptografia para proteção de dados sensíveis conforme a LGPD [13].
- Ciclo de Vida de Software: Gestão de atualizações para correção de falhas e adequação a novas políticas como o Previne Brasil [6].

E. Perspectivas Futuras

Como perspectiva futura, a evolução do ecossistema e-SUS APS pode envolver a investigação de recursos de Inteligência Artificial (IA) e modelos de linguagem aplicados aos registros clínicos, visando ampliar o potencial de apoio à decisão clínica. Nesse contexto, Modelos de Fundação (FMs) e Large Language Models (LLMs) constituem possibilidades tecnológicas a serem avaliadas em estudos futuros, especialmente quanto à capacidade de processar informações clínicas não estruturadas [15]. O GatorTron, por exemplo, é apresentado na literatura como um modelo de linguagem clínica desenvolvido para processamento de textos médicos [16]. Entretanto, sua utilização no e-SUS APS não foi implementada nem avaliada experimentalmente neste estudo, sendo apresentada apenas como possibilidade de investigação futura.

A Figura 5 apresenta um cenário conceitual de integração de tecnologias associado às perspectivas da Saúde 4.0 no SUS. Nesse cenário prospectivo, a interoperabilidade semântica, os

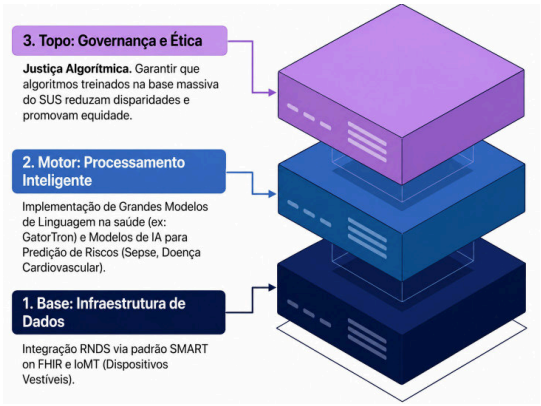


Fig. 5. Cenário conceitual e perspectivas futuras para a Saúde 4.0 no ecossistema do e-SUS APS. Estruturação analítica sem validação empírica direta no presente trabalho. Elaboração dos autores (2026).

padrões de integração e os mecanismos de governança constituiriam elementos fundamentais para eventual incorporação de recursos de IA, LLMs e dispositivos de Internet das Coisas Médicas (IoMT). Padrões como SMART on FHIR poderiam ser investigados como mecanismos de integração entre aplicações e sistemas de informação em saúde [17]. Nesse cenário hipotético, modelos de fundação e LLMs, incluindo o GatorTron, poderiam ser investigados para tarefas de processamento e interpretação de informações clínicas não estruturadas [16]. A efetividade, segurança, desempenho e aplicabilidade desses modelos no contexto do e-SUS APS dependeriam de avaliações experimentais específicas, que não fazem parte do escopo deste estudo.

De forma prospectiva, dispositivos vestíveis e biossensores poderiam ser investigados como fontes adicionais de dados para o ecossistema de saúde digital, incluindo informações de monitoramento contínuo, como glicemia e frequência cardíaca. A eventual integração desses dados ao fluxo de cuidado do PEC dependeria de estudos de viabilidade, interoperabilidade, segurança, privacidade e validação clínica.

Ressalta-se que a arquitetura apresentada na Figura 5 possui caráter conceitual e prospectivo, não constituindo uma arquitetura implementada ou validada experimentalmente pelos autores. Dessa forma, sua utilização como estratégia de apoio à decisão clínica permanece como hipótese para pesquisas futuras, devendo ser precedida por avaliações técnicas, clínicas, éticas e de segurança.

IV. Conclusão

Ao completar sua primeira década, o e-SUS APS consolidou-se como um dos maiores bancos de dados de saúde primária do mundo, atingindo uma cobertura nacional de 97,6% das UBS em 2024 [11] [12]. Entretanto, os resultados deste estudo indicam que a efetividade plena desse ecossistema ainda é limitada pela fragmentação de interfaces, onde 15 de 31 sistemas nacionais de informação permanecem isolados, gerando retrabalho e inconsistência de dados [13] [14]. A Engenharia de Sistemas de Saúde emerge como o elo científico

capaz de converter esse volume massivo de registros em conhecimento clínico aplicado, utilizando a RNDS e padrões semânticos robustos como o HL7 FHIR R4 [4] [11].

Nesse contexto, o papel do Engenheiro Biomédico como 'profissional de interface' é determinante: ele atua na zona de convergência entre as complexas demandas assistenciais e o rigor dos requisitos técnicos, sendo o agente capaz de viabilizar a transição da mera implantação física para a efetiva incorporação tecnológica [15]. Sua intervenção garante que o prontuário eletrônico transcenda a função de repositório burocrático para tornar-se um ecossistema dinâmico de suporte à decisão clínica, atuando como pilar de sustentabilidade institucional e de segurança do paciente [15] [16].

Em última análise, a transformação digital do SUS não deve ser reduzida a um processo de informatização técnica; trata-se de um projeto político-técnico de fortalecimento da cidadania sanitária, orientado pela justiça algorítmica e pelo compromisso inalienável com a equidade em saúde para todos os brasileiros [10].

REFERÊNCIAS

[1] T. L. Gontijo et al., "Análise lógica do Prontuário Eletrônico do Cidadão da estratégia e-SUS Atenção Primária," Revista de APS, vol. 26, p. e262341492, 2023.

[2] I. C. Celuppi et al., "Dez anos do Prontuário Eletrônico do Cidadão e-SUS APS: em busca de um Sistema Único de Saúde eletrônico," Rev. Saude Publica, vol. 58, p. 23, 2024.

[3] A. C. Cielo et al., "Implementation of the e-SUS Primary Care Strategy: an analysis based on official data," Rev. Saude Publica, vol. 56, p. 5, 2022.

[4] G. C. Coelho Neto et al., "Integração entre os sistemas nacionais de informação em saúde: o caso do e-SUS Atenção Básica," Rev. Saude Publica, vol. 55, p. 93, 2021.

[5] I. M. P. Barbalho et al., "Electronic health records in Brazil: Prospects and technological challenges," Front. Public Health, vol. 10, p. 963841, 2022.

[6] Brasil. Ministério da Saúde. Estratégia de Saúde Digital para o Brasil 2020-2028. Brasília, DF: Ministério da Saúde, 2020.

[7] M. Felisberto et al., "Mapping the evaluation of the electronic health system PEC e-SUS APS in Brazil: a scoping review protocol," Systematic Reviews, vol. 13, p. 237, 2024.

[8] L. M. Iahnke, "Saúde Digital na Atenção Primária à Saúde: avanços, desafios e perspectivas," SciELO Preprints, 2025.

[9] R. A. M. Valentim et al., "The relevance of a technology ecosystem in the Brazilian National Health Service's Covid-19 response," Cien Saude Colet, vol. 26, pp. 2035-52, 2021.

[10] M. Lehne et al., "Why digital medicine depends on interoperability," npj Digit. Med., vol. 2, p. 79, 2019.

[11] Brasil. Ministério da Saúde. Portaria nº 1.412, de 10 de julho de 2013. Institui o Sistema de Informação em Saúde para a Atenção Básica - SISAB. Diário Oficial da União, Brasília, DF, 11 jul. 2013.

[12] Brasil. Ministério da Saúde. Portaria nº 2.983, de 14 de novembro de 2019. Altera a Portaria de Consolidação nº 6/GM/MS, de 28 de setembro de 2017. Diário Oficial da União, Brasília, DF, 18 nov. 2019.

[13] Brasil. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

[14] Portal Engenharia Biomédica, "Informática em Saúde: HIS, Prontuário Eletrônico e Interoperabilidade FHIR/HL7," 2026. [Acessado em 2026].

[15] J. Clusmann et al., "The future of medical imaging: foundation models," npj Digit. Med., vol. 6, p. 141, 2023.

[16] X. Yang et al., "A large language model for electronic health records," npj Digit. Med., vol. 5, p. 194, 2022. (GatorTron study).

[17] D. Gotz et al., "SMART on FHIR: a standards-based, interoperable apps platform for electronic health records," Journal of the American Medical Informatics Association, vol. 23, p. 899-908, 2016.

[18] J. V. M. Crisostomo, "Saúde Digital em Ação: A contribuição da Engenharia Biomédica na incorporação do PEP+RN," Trabalho de Conclusão de Curso, Dept. Eng. Biomédica, UFRN, Natal, RN, 2025.

[19] SBEB, "Proceedings of the XXIX Brazilian Congress on Biomedical Engineering," CBEB, 2024.

Towards the Development of a Deep Learning Framework for Neonatal Seizure Detection in Embedded Platforms

Richardson Naves Leão[1,†], Aline Elvina Rodrigues Fernandes[1,†],
Marcos Aurélio Tavares Filho[2], Helton Maia[2]
[1]*Instituto do Cérebro*, Universidade Federal do Rio Grande do Norte, Natal, Brazil
[2]*Escola de Ciências e Tecnologia*, Universidade Federal do Rio Grande do Norte, Natal, Brazil
Corresponding author: helton.maia@ufrn.br

***Abstract*—We present a binary ictal-vs-non-ictal classifier for neonatal scalp EEG, trained and evaluated *cross-patient* with strict Leave-One-Subject-Out (LOSO) over all 79 term neonates of the Helsinki dataset. A compact EEGNet-style CNN is trained on 8 s windows with a clean three-expert consensus ground truth and an aggregation that separates the 39 seizure-bearing subjects from the 40 consensus-negative ones (ranking metrics over the former; specificity and false-alarm rate over all). The full 79-fold study was run at scale on the NPAD supercomputer (NVIDIA V100 GPUs). The detector reaches a pooled AUROC of ≈ 0.81 (best configuration 0.807), a cross-patient baseline; post-hoc event-level scoring with causal, real-time-compatible smoothing already detects $\approx 79\%$ of seizures at ≈ 3.5 false alarms/h, and inference takes ≈ 1.8 ms per 8 s window on a CPU, well within real time for bedside use. As secondary analyses we ablate two input-representation choices usually assumed rather than measured, the montage (referential 19-channel vs. bipolar "double-banana") and per-recording z-score normalization, and add a second, channel-mixing CNN as a control: z-score helps both montages modestly, the two montages reach a comparable pooled AUROC and trade sensitivity for specificity, and the normalization benefit is specific to the compact EEGNet rather than universal. Future work will focus on optimizing this model to run on our custom, low-cost microprocessed EEG hardware within an Artificial Intelligence of Things (AIoT) framework, helping to universalize neonatal neurological care.**

***Index Terms*—neonatal EEG, seizure detection, cross-patient, leave-one-subject-out, EEGNet, convolutional neural network, electroencephalography.**

I. Introduction

In the neonatal intensive care unit (NICU), electrographic-only seizures are among the most frequent neurological manifestations. Because nearly 50% of these events in the first days of life are entirely subclinical, continuous EEG monitoring remains the only reliable method for their identification. This reality strongly motivates the development of automated EEG-based detection systems to support real-time bedside clinical decisions. In this work, we propose a seizure detection classifier grounded in three guiding principles. First, cross-patient generalization: a model is only clinically useful if it performs on babies it has never seen, since neonatal seizure morphology is highly subject-specific. We therefore enforce Leave-One-Subject-Out (LOSO) evaluation, ensuring that no window from the test subject ever appears in training. Second, class-aware metrics: with a rare positive class, a trivial "all-negative" classifier would score high accuracy while detecting nothing, so we report sensitivity, specificity, AUROC (area under the receiver operating characteristic curve), AUPRC (area under the precision–recall curve), and false positives per hour (FP/h), the metric that governs alarm burden in the NICU. Third, comparability: results are benchmarked against established neonatal baselines, specifically Temko *et al.* [3] (support vector machines) and O'Shea *et al.* [4] (fully convolutional networks).

The contribution of this paper is not the proposal of a new architecture, but a leak-free cross-patient neonatal seizure classifier on the full public Helsinki cohort: a compact EEGNet trained and tested under strict 79-fold LOSO, with aggregation separating seizure-bearing from consensus-negative subjects. Around this baseline, we introduce two controlled secondary analyses of input-representation choices that are often assumed rather than empirically measured: the EEG *montage* (referential vs. bipolar) and the input *normalization* (raw vs. per-recording z-score). These are quantified on identical LOSO splits over the public Helsinki neonatal EEG dataset [1], [2], alongside a second architecture included as a generalization control. Finally, further work is currently underway to optimize this model for resource-constrained embedded systems (e.g., using TinyML frameworks), paving the way for low-cost, real-time edge computing at the neonatal bedside.

II. Related Work

Automated neonatal seizure detection has a substantial history. Early systems combined hand-engineered features with classical classifiers: Temko *et al.* [3] trained a support vector machine over spectral, temporal, and information-theoretic features and established a strong patient-independent baseline, and Tapani *et al.* [6] showed that time-varying EEG correlation features further improve detection. The field then moved to deep learning on raw EEG: O'Shea *et al.* [4] proposed a

[†]These authors contributed equally to this work.

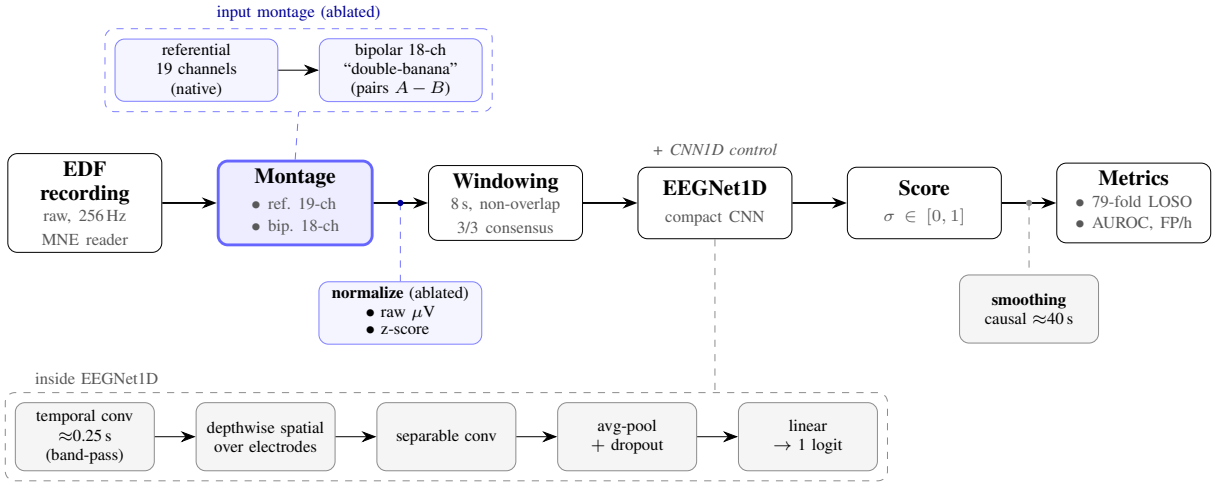


Fig. 1. End-to-end processing pipeline: a 256 Hz EDF recording is converted to the referential or bipolar montage, optionally per-recording z-scored, sliced into non-overlapping 8 s windows, and classified by `EEGNet1D` into a per-window score aggregated under 79-fold LOSO. The insets expand the bipolar “double-banana” derivations (top) and the `EEGNet1D` stages (bottom). Input *montage* and *normalization* are the 2×2 ablated factors.

fully convolutional network that matched or exceeded the feature-based SVM without manual feature engineering, while Ansari *et al.* [7] combined convolutional networks with a heuristic classifier. The clinical viability of such algorithms was assessed by Pavel *et al.* [8] in a multicentre randomised trial. Most recently, Hogan *et al.* [9] showed that scaling a convolutional network to more than 50,000 h of annotated EEG attains expert-level detection, showing that data scale, not only architecture, drives performance. Deep learning on EEG more broadly is surveyed by Roy *et al.* [10]. Compact, EEG-specific architectures such as EEGNet [5], originally designed for brain–computer interfaces, offer a low-parameter alternative that is attractive in the data-limited, leakage-sensitive neonatal setting. A parallel line replaces task-specific supervision with self-supervised pre-training on large unlabelled EEG corpora, through contrastive Transformer encoders [11] or tokenized foundation models pre-trained on thousands of hours of recordings [12]. These are the high-capacity counterpart to the compact model studied here, and Section V returns to the trade-off.

Our work sits in this raw-EEG CNN line, with the emphasis on the evaluation protocol rather than on the network: a cross-patient *baseline* over the full public cohort, around which we measure two *input-representation* choices (the recording montage and the per-recording amplitude normalization) that prior work typically fixes implicitly. To our knowledge this combination, on this dataset, has not been reported.

III. Materials and Methods

Fig. 1 summarizes the full processing pipeline; the stages are detailed below.

A. Dataset

The Helsinki neonatal EEG seizure dataset [1], [2] comprises 79 term neonates recorded in the NICU of Helsinki University Hospital at 256 Hz, with a referential 10–20 montage (∼19 EEG electrodes plus ECG and respiration). Seizures are annotated per second by three independent experts; annotation column j (0-based) corresponds to file `eeg{j+1}.edf`. The data are released under CC BY 4.0; the full set of 79 EDFs is ≈4.33 GB.

Treating a second as ictal only when all three experts agree (the project’s ground truth), 39 of 79 subjects have seizures (46 by majority ≥2/3; 58 by any expert), with ≈10.9 h of total consensus seizure time. The ictal fraction is high relative to scalp-EEG seizure corpora, which makes the positive class comparatively well represented and the cross-patient task tractable. Fig. 2 shows an example recording with its consensus annotation and the two 8 s window classes the model must separate.

a) EDF heterogeneity: The 79 EDF files are not uniform: some are plain EDF (electrode suffix `-REF`), others are declared EDF+C yet contain no “EDF Annotations” signal and use a lowercase `-Ref` suffix, and the physical channel order varies between subjects. The strict reader `pyedflib` rejects the EDF+C-without-annotations files outright, so we read every recording with MNE (`mne.io.read_raw_edf`), verified to reproduce `pyedflib`’s output to $\sim 10^{-13}\,\mu$V on a file both can read. Channel names are normalized case-insensitively and reordered to a fixed canonical order, so per-subject ordering is irrelevant downstream.

B. Labels and windowing

Each annotation CSV is read with its header skipped, so data-row i is second i. The three expert columns are summed per second and thresholded at consensus ($= 3$); contiguous ictal seconds are collapsed into half-open intervals $[s, e)$. Signals are sliced into 8 s windows at 256 Hz (2048 samples), the window length of the dominant neonatal baselines. A window spanning $[t_0, t_1)$ is labelled ictal when it overlaps any

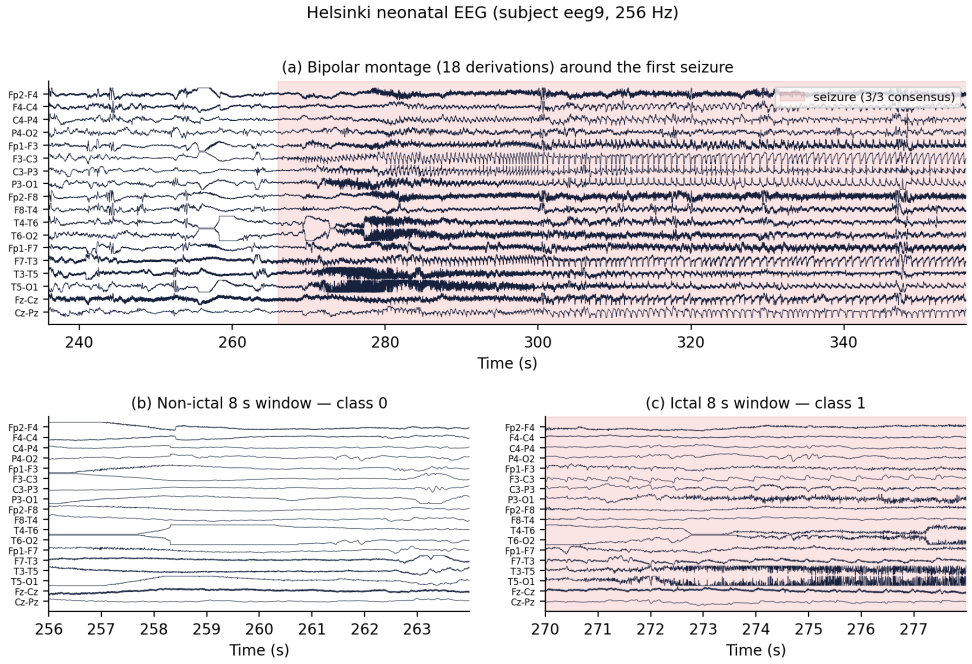


Fig. 2. Example data and the two classes (subject `eeg9`, bipolar "double-banana" montage, 256 Hz; traces per-channel scaled for display). **(a)** A 120 s segment around the first seizure onset, the red span marking the 3/3-consensus ictal interval. **(b)** A non-ictal 8 s window (class 0). **(c)** An ictal 8 s window (class 1); each 8 s window is one input.

consensus interval,

$$y = \mathbb{1}\left[\exists\,[s, e) \in I : t_0 < e \ \wedge\ t_1 > s\right], \tag{1}$$

where I is the set of consensus intervals. The full-cohort study uses non-overlapping windows (stride = window) for both training and evaluation, which avoids double counting at evaluation and, given this corpus's high ictal fraction, makes overlap-as-augmentation unnecessary; the validation gate alone (Table I) used a 4 s stride (50% overlap). Treating the label as intervals rather than a per-second mask makes window labelling independent of window length and robust to CSV padding.

C. Input montages (ablated)

Two montages are compared on identical splits. The referential montage uses the 19 native electrodes (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz) as recorded, the simplest baseline. The bipolar montage is the longitudinal "double-banana" (18 derivations), each a difference of two referential electrodes obtained by a pure re-montage function:

Fp2–F4, F4–C4, C4–P4, P4–O2, Fp1–F3, F3–C3, C3–P3, P3–O1,
Fp2–F8, F8–T4, T4–T6, T6–O2, Fp1–F7, F7–T3, T3–T5, T5–O1,
Fz–Cz, Cz–Pz.

Fig. 3 maps these derivations onto the scalp. This is the representation of clinical practice and the neonatal literature, emphasizing local gradients and attenuating common-reference noise, which is why we include it in the ablation rather than assuming it helps. To keep the montage identical across subjects, a derivation is emitted only when both of its electrodes survive the across-cohort intersection. The same network accepts 19 or 18 channels with no code change. Both montages derive from the same time samples, so they are two views of one recording (Fig. 4).

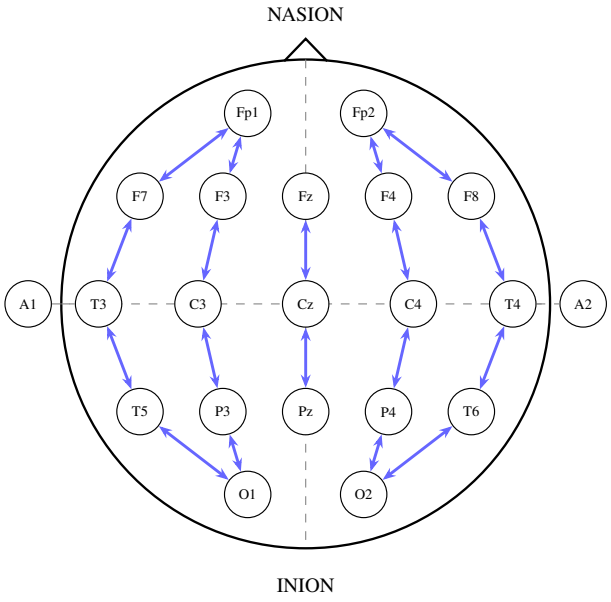


Fig. 3. Scalp map (top view, 10–20 system): the 19 referential electrodes (circles) and the 18 bipolar "double-banana" derivations (blue arrows); A1/A2 are auricular references. Adapted from Fig. 1 of [1].

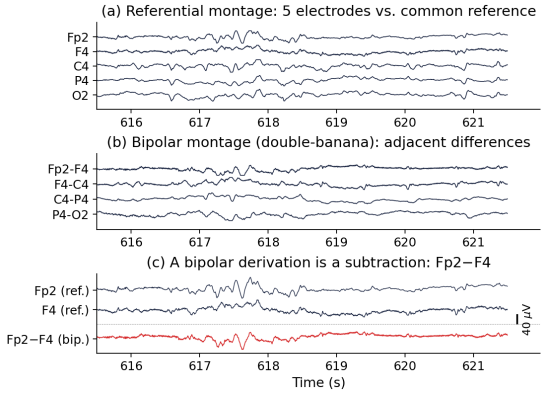


Fig. 4. Referential and bipolar montages of the *same* 6 s ictal segment (`eeg9`). **(a)** Five referential electrodes of the right-medial chain (Fp2, F4, C4, P4, O2). **(b)** Their four bipolar derivations (Fp2–F4, F4–C4, C4–P4, P4–O2). **(c)** The subtraction in real microvolts: Fp2–F4 (red) is the referential Fp2 minus F4, attenuating the common-mode component.

D. Input normalization (ablated)

EEG amplitude scale varies markedly between neonates and recordings (electrode impedance, age, pathology). We ablate two scalings: raw microvolts, and a per-recording, per-channel z-score, in which each channel c of recording r is standardized using only its own statistics,

$$\tilde{x}_{c,t} = \frac{x_{c,t} - \mu_c^{(r)}}{\sigma_c^{(r)}}, \tag{2}$$

where $\mu_c^{(r)}, \sigma_c^{(r)}$ are computed within recording r only (a constant channel, $\sigma \approx 0$, is left centred to avoid division by zero). Being a per-channel affine rescaling, the z-score discards only this nuisance amplitude scale while leaving the waveform *morphology*, the rhythmic shape that signals a seizure, intact. Because the statistics never cross subjects, the z-score introduces no LOSO leakage, in contrast to a global standardization, which would let the held-out subject influence the training scale.

E. Model and training

The classifier, which we refer to in code as `EEGNet1D`, is our PyTorch reimplementation of the published EEGNet architecture [5]; it is not a new network. We keep EEGNet's canonical design and hyperparameters (F_1=8 temporal filters, depth multiplier D=2, dropout 0.25) and adapt only the output to a single binary logit (used with `BCEWithLogitsLoss`); the one data-dependent choice is the temporal kernel width (≈0.25 s, i.e. 64 samples at 256 Hz). An input window (C, T) is reshaped into a single-channel $(1, C, T)$ map so that temporal and spatial filtering separate, and three stages follow (Fig. 1): a *temporal* convolution of width ≈0.25 s applied per electrode, acting as a learned band-pass filter bank; a *depthwise spatial* convolution spanning all electrodes at once, collapsing the electrode axis; and a *separable* convolution for higher-level temporal structure. Average pooling and dropout follow the last two, and a linear layer maps the flattened features to one logit per window. That flattened size is computed dynamically, so the same architecture serves the 19- and 18-channel montages without modification. The whole network has only 2433 parameters (≈10 KB in fp32), small enough for microcontroller memory.

Training uses Adam (lr $= 10^{-3}$, batch 64), 15 epochs per fold. The strong class imbalance is handled automatically by weighting the positive class in the loss,

$$\text{pos_weight} = n_{\text{neg}}/n_{\text{pos}}, \tag{3}$$

computed per fold from the training split, so the minority (ictal) class is up-weighted without resampling or discarding data. Hyperparameters are fixed a priori (no nested cross-validation) to avoid tuning leakage in the LOSO loop, and model initialization is seeded per fold for reproducibility. The model is compute-bound (the depthwise/grouped convolutions do not benefit from tensor cores, so fp32 was empirically faster than mixed precision here); with each of the eight ablation conditions run as one independent GPU job, the full 79-fold study completes in about three hours of wall-clock.

F. Evaluation: LOSO and metric aggregation

For each held-out subject i, the model trains on the concatenation of all other subjects and scores subject i (structural anti-leakage). Per-fold metrics are sensitivity, specificity, AUROC, AUPRC, and false positives per hour (FP/h).

A central methodological choice of this work is *how* these per-subject metrics are aggregated across the cohort, because a naive cohort mean is misleading. Of the 79 neonates, only 39 have seizures under the consensus rule and 40 are consensus-negative. Such a test subject contains a single class, for which a ranking metric such as AUROC (and likewise AUPRC and per-subject sensitivity) is mathematically *undefined*. We therefore partition the aggregation by what each metric can measure:

- *AUROC, AUPRC, sensitivity* are averaged *only* over the 39 seizure-bearing subjects, where the positive class exists; the 40 single-class subjects yield *nan* and are excluded from these means.
- *Specificity and FP/h* are averaged over *all 79* subjects, because the 40 consensus-negative recordings are precisely where false alarms can occur and are thus what make the false-positive rate representative; excluding them would flatter the model.

We additionally report a *pooled* AUROC (every fold's window scores concatenated into one ranking) as a threshold-free headline that does not depend on this partition.

Three estimation levels are kept distinct throughout. *Window-level* metrics score each 8 s window independently and underlie Table II; the threshold-dependent ones (sensitivity, specificity, FP/h) use a fixed operating point of 0.5 on the sigmoid output, with no per-subject or per-fold calibration. *Subject-level* estimates are those per-fold values aggregated as above, each mean carrying its across-fold sample standard deviation (ddof = 1). *Event-level* metrics score whole seizures after causal post-processing and are reported separately (Section IV-B).

G. Reproducibility

The pipeline is the `neoseiz` Python package (windowing, model, training, evaluation, EDF/data I/O, and the LOSO runner) developed test-first with 59 passing unit/integration tests covering windowing, both channel counts, consensus parsing, the real-recording read, z-score behaviour, both architectures, LOSO anti-leakage and aggregation, and the real-time streaming path. For the full study, all 79 subjects are windowed once into a single in-RAM `float32` array (≈9 GB peak resident per job) and each fold gathers mini-batches by vectorized indexing (no per-sample data loader), keeping the GPU saturated. The full study was run on UFRN's NPAD supercomputer, one NVIDIA Tesla V100 (16 GB) GPU per condition under Slurm, with PyTorch 2.6 (CUDA 12.4) and Python 3.12; each condition took about 2–3 h for its 79 folds and all eight ran concurrently; all reported numbers are from these NPAD runs. The pipeline was developed and prototyped on a commodity desktop (NVIDIA RTX 5060 Ti, PyTorch 2.11/CUDA 12.8). MNE, NumPy, SciPy and scikit-learn complete the stack.[1]

IV. Results

A. Validation gate

Before committing GPU time to the full study, a six-subject gate verified pipeline correctness on a deliberately diverse subset: `eeg3` (no consensus seizures, exercising the specificity / *nan* path), `eeg4` (an EDF+C file with the `-Ref` suffix, exercising the MNE reader), and `eeg5`, `eeg7`, `eeg9`, `eeg13` (seizure-bearing), totalling 8567 windows (Table I). Subject leakage was provably absent (per fold, $n_{\text{train}} + n_{\text{test}}$ equals the cohort total, e.g. $8567 - 1102 = 7465$ for `eeg3`); the consensus-negative subject correctly yielded *nan* sensitivity/AUROC while still contributing specificity and FP/h; the heterogeneous EDF headers were read correctly; and all aggregate quantities were finite. The gate was an implementation check only: no hyperparameter, threshold or architectural choice was selected from it, all hyperparameters being fixed a priori from the published EEGNet configuration [5]. The six subjects remain ordinary members of the 79 folds of the full study, so the gate cannot bias the reported estimates.

TABLE I
VALIDATION-GATE SUBSET AFTER 8 S WINDOWING (4 S STRIDE). "HOURS" IS THE RECORDING DURATION USED FOR FP/H.

Subject	Windows	Ictal	Non-ictal	Hours
`eeg3` (no cons. sz.)	1102	0	1102	1.23
`eeg4` (EDF+C)	855	214	641	0.95
`eeg5`	959	794	165	1.07
`eeg7`	912	160	752	1.01
`eeg9`	886	220	666	0.99
`eeg13`	3853	314	3539	4.28
Total	8567	1702	6865	9.53

[1]Source code and reproduction scripts available on Zenodo: https://doi.org/10.5281/zenodo.21014978.

B. Full-cohort results

We then ran the full study over all 79 subjects under strict 79-fold LOSO, sweeping the 2×2 design (montage $\times$ normalization) as four independent runs. The cohort yielded exactly 39 seizure-bearing subjects under the consensus rule, matching Section III-A and confirming the labelling pipeline at scale (Table II).

TABLE II
FULL-COHORT 79-FOLD LOSO, 2×2 ABLATION (15 EPOCHS). CELLS ARE THE ACROSS-FOLD MEAN WITH ITS SAMPLE SD AS A SUBSCRIPT; AUROC/AUPRC/SENSITIVITY OVER THE 39 SEIZURE SUBJECTS, SPECIFICITY AND FP/H OVER ALL 79. POOLED AUROC RANKS EVERY FOLD'S CONCATENATED SCORES AT ONCE AND HAS NO FOLD-LEVEL SD. BEST MEAN PER COLUMN IN **BOLD**.

Montage	Norm.	AUROC	AUPRC	Pooled	Sens.	Spec.	FP/h
ref.	raw	$0.677_{\pm.18}$	$0.458_{\pm.32}$	0.758	$0.371_{\pm.31}$	$\mathbf{0.894}_{\pm.18}$	$\mathbf{43.3}_{\pm80}$
ref.	z-score	$0.731_{\pm.15}$	$0.485_{\pm.33}$	**0.807**	$0.442_{\pm.32}$	$0.872_{\pm.20}$	$48.0_{\pm79}$
bip.	raw	$0.662_{\pm.19}$	$0.437_{\pm.33}$	0.728	$0.400_{\pm.34}$	$0.862_{\pm.23}$	$51.8_{\pm80}$
bip.	z-score	$\mathbf{0.745}_{\pm.16}$	$\mathbf{0.517}_{\pm.35}$	0.786	$\mathbf{0.527}_{\pm.36}$	$0.834_{\pm.20}$	$63.1_{\pm79}$

a) A cross-patient baseline: The headline result is the cross-patient detector itself: a compact EEGNet reaches a pooled AUROC of 0.807 in its best configuration (referential + z-score) and ≈ 0.81 across the standardized conditions (Table II), on all 79 neonates with no window of the test subject ever seen in training. The per-subject spread behind these means is wide, and is detailed below.

b) Effect of normalization: As a secondary analysis, per-recording z-score raises the *pooled* AUROC of both montages, by $+0.049$ for referential ($0.758 \rightarrow 0.807$) and $+0.058$ for bipolar ($0.728 \rightarrow 0.786$), and lowers no ranking metric (Fig. 5; Table II); the largest AUPRC gain is again bipolar's ($0.437 \rightarrow 0.517$). The effect is largest where the raw input is weakest: raw bipolar, whose recording-dependent dynamic range z-score regularizes, is the lowest configuration and gains the most. A subject-paired Wilcoxon test confirms the effect per subject for *both* montages, while neither montage contrast is significant ($p \geq 0.34$; see *Paired significance*). Among these input-representation choices it is thus normalization, not montage, that moves the needle, and even that is a modest effect.

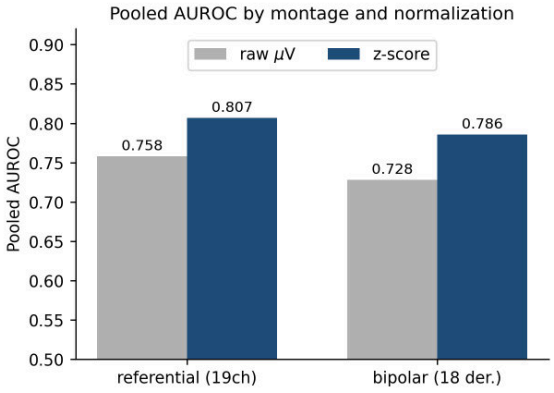


Fig. 5. Per-recording z-score raises the pooled AUROC of both montages.

c) Effect of montage: a sensitivity/specificity trade-off: Once standardized, the two montages reach a comparable pooled AUROC (0.807 referential vs. 0.786 bipolar) and differ mainly in operating character (Fig. 6).

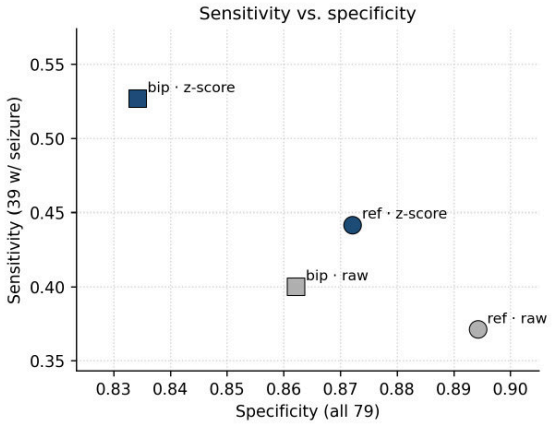


Fig. 6. Under z-score, the montages trade sensitivity (bipolar) against specificity and false-alarm rate (referential).

Bipolar + z-score is the most sensitive (0.527 vs. 0.442) and has the highest seizure-subject mean AUROC (0.745), at the cost of more false alarms (63.1 vs. 48.0 FP/h) and lower specificity. Referential is the more conservative montage (fewer false alarms) and edges the best pooled AUROC. The clinically preferable montage therefore depends on whether the use case rewards catching more seizures or suppressing false alarms.

d) Paired significance: Because the four conditions score every subject on the same LOSO split, the per-subject metrics are paired and admit a subject-level Wilcoxon signed-rank test (Table III), a guarantee the cohort means of Table II cannot give. The normalization effect on per-subject AUROC is significant for *both* montages (median $+0.059$, $p = 0.007$ for bipolar; $+0.034$, $p = 0.020$ for referential), while neither montage contrast is significant under either normalization ($p \geq 0.34$). Normalization is thus the one input-representation factor with a measurable per-subject effect, and even it is modest; the montage differences live within cross-patient noise, a caution against over-reading small differences in cohort-mean tables.

TABLE III
SUBJECT-PAIRED WILCOXON SIGNED-RANK TEST ON PER-RECORDING AUROC (39 SEIZURE SUBJECTS). Δ IS THE PER-SUBJECT DIFFERENCE A − B; p IS TWO-SIDED.

Contrast (A vs. B)	median ΔAUROC	p
Norm., ref.: z-score vs. raw	$+0.034$	**0.020**
Norm., bip.: z-score vs. raw	$+0.059$	**0.007**
Montage, z-score: bip. vs. ref.	$+0.026$	0.42
Montage, raw: bip. vs. ref.	-0.015	0.34

e) Generalization across architectures: To test whether the normalization effect belongs to the input representation or to `EEGNet1D` specifically, we re-ran the full 2×2 ablation with a structurally different compact network: a plain 1D-CNN (`CNN1D`) that mixes all electrodes in its first layer and uses batch normalization throughout. The cross-patient task stays tractable (Table IV), but the normalization effect does *not* carry over: `CNN1D`'s bipolar z-score-minus-raw gap is -0.014 (vs. $+0.058$ for EEGNet), because batch normalization already absorbs the recording-dependent amplitude that z-score otherwise removes. The z-score benefit is therefore *architecture-mediated*, not a universal property of the montage. Notably `CNN1D` at raw referential reaches the highest pooled AUROC of the whole study (0.824, above EEGNet's best 0.807). Repeating the bipolar conditions under a changed seed and epoch budget moved their pooled AUROC by up to 0.05, so a gap of this size does not establish a ranking and the ≈ 0.81 figure is best read as a baseline band.

TABLE IV
GENERALIZATION CONTROL: POOLED AUROC OF THE 2×2 ABLATION FOR `EEGNet1D` AND A STRUCTURALLY DIFFERENT `CNN1D` (CHANNEL-MIXING 1D-CNN), 15 EPOCHS.

Montage	Norm.	EEGNet	CNN1D
ref.	raw	0.758	**0.824**
ref.	z-score	**0.807**	0.786
bip.	raw	0.728	0.782
bip.	z-score	0.786	0.768
bipolar gap (z-score − raw)		$+0.058$	-0.014

f) Pooled ROC curves: Fig. 7 overlays the threshold-free pooled ROC of all four conditions: the two z-score curves reach pooled AUROC $\approx$ 0.79–0.81, above their raw counterparts (0.758 and 0.728), and the montages become close once standardized (0.807 vs. 0.786).

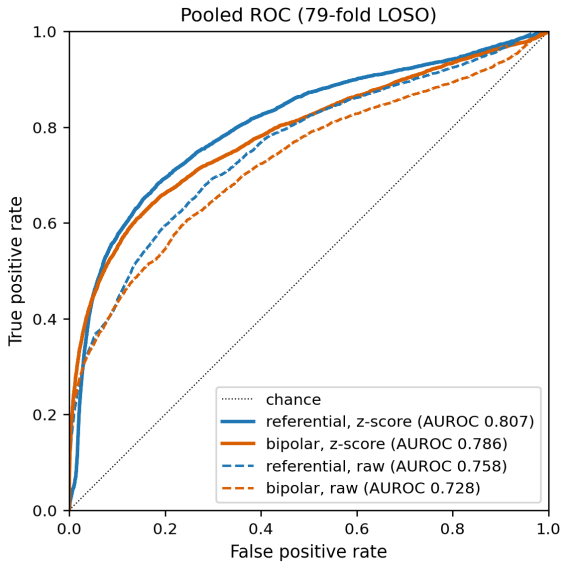


Fig. 7. Pooled ROC of all four conditions (every fold's window scores concatenated): montage by colour (referential blue, bipolar orange), normalization by line style (z-score solid, raw dashed); the legend reports the pooled AUROC.

g) Cross-patient spread: Aggregates hide wide per-subject variation, the defining difficulty of cross-patient neonatal detection. Under referential + z-score, per-subject AUROC ranges from 0.97 (`eeg31`) down to 0.34 (`eeg15`): some neonates' seizure morphology is simply not represented by the other 78, the cost of never training on the test subject.

h) Consensus-negative subjects and the false-alarm rate: The 40 subjects without consensus seizures carry the false-alarm estimate. Under referential + z-score their mean specificity is 0.908 ± 0.181, above the 0.836 ± 0.218 of the 39 seizure-bearing subjects, and their FP/h averages 41.6 with a median of 3.7. The distance between mean and median is itself the finding: over all 79 subjects FP/h has median 13.4 (IQR 1.1–54.3) against a mean of 48.0, and five subjects alone produce 36% of all false positives, so the cohort-mean FP/h is driven by a few outlying recordings and the median is the more representative summary. The estimate is also conservative, because of these 40 subjects only 21 carry no seizure annotation from any expert: the other 19 have seizures marked by one or two of the three (7 of them by a majority), so part of what is counted here as a false alarm is non-consensus seizure activity.

i) Sensitivity by seizure burden: Detection is unevenly distributed across the 39 seizure-bearing subjects and tracks how much seizure each one carries. Grouping them by total consensus ictal time, per-subject sensitivity rises from 0.305 ± 0.281 below 5 min ($n = 12$) to 0.477 ± 0.323 between 5 and 30 min ($n = 19$) and 0.564 ± 0.319 above 30 min ($n = 8$), with per-subject AUROC in the same order (0.693, 0.730, 0.790). The association holds across the cohort (Spearman $\rho = 0.38$, $p = 0.018$ for sensitivity; $\rho = 0.31$, $p = 0.055$ for AUROC). Low-burden neonates, whose seizures are few and brief, therefore account for most of what the detector misses, which is where a larger and more diverse seizure cohort would help first.

j) Event-level detection: The window-level FP/h above is pessimistic: it counts every false 8 s window separately, whereas a clinician cares whether each seizure *event* is caught and how many *separate* alarms occur per hour. We therefore score post-hoc at the event level, with no retraining: a causal (past-only, real-time compatible) 40 s moving average smooths each subject's window scores, detections are merged across gaps < 30 s and short (< 16 s) ones dropped, and a true seizure counts as detected if a predicted event overlaps it within a 30 s collar. Reference events are the contiguous ictal runs of the window labels, so consensus intervals closer together than one window merge and the 343 raw consensus intervals become 315 scored events. For the best configuration (referential + z-score) the detector then flags 78.7% of those 315 events at 3.5 false alarms/h at a score threshold of 0.30, tunable along the same scores (60.6% at 1.9 FP/h at 0.50, or higher sensitivity at more alarms). No threshold search was performed against a clinical protocol; these are two points of one curve. The per-window $\approx$ 48/h thus collapses to a few alarms/h once counted as events, the single largest step toward the neonatal baselines and one needing no change to the model. (The causal smoothing alone also lifts pooled AUROC by $+0.02$ to $+0.03$.)

k) Inference cost: The detector is light enough for the bedside (Fig. 8). A single 8 s window (EEGNet, 19 channels, batch 1) is classified in $\approx$1.8 ms on a commodity CPU and $\approx$1.7 ms on the GPU, far inside the 8 s the window spans. At batch 1 this tiny network is launch-bound rather than compute-bound, so the two are within noise; the GPU's edge is throughput, not single-window latency: $\approx$26,000 windows/s in batch versus $\approx$170/s on the CPU. A single commodity machine can therefore score many neonates in real time and a GPU is optional, which with the causal smoothing above makes real-time NICU monitoring realistic.

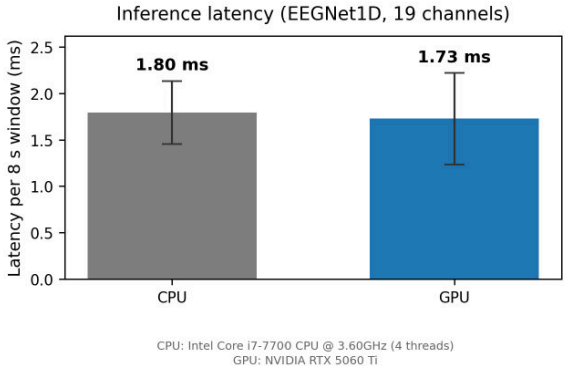


Fig. 8. Inference latency per 8 s window for the 2433-parameter EEGNet1D ($\approx$10 KB; 19 channels, batch 1; mean $\pm$ sd over 200 runs).

V. Discussion

These cross-patient numbers are a baseline rather than a ceiling: several deliberate simplifications leave clear headroom. *(i) Fixed 15-epoch budget:* fixed a priori, with no early stopping; a larger budget would mainly lift the raw configurations toward the standardized ones. *(ii) Post-processing not yet fully tuned:* the event-level threshold and the merge/collar parameters are not calibrated per the protocols of the cited baselines. That remaining tuning, rather than the architecture, is most of the gap to Temko [3] and O'Shea [4] ($\approx$0.95–0.97 AUC). *(iii) Fixed 0.5 operating point:* sensitivity, specificity and FP/h use an uncalibrated threshold; the pooled ROC/AUROC are threshold-free and are the fairer headline. *(iv) No overlap augmentation:* an intentional scope limit for this iteration. *(v) No Transformer or self-supervised comparison:* the foundation-model line [11], [12] is the natural high-capacity counterpart, but it targets a different deployment regime. The constraint here is a detector that fits beside acquisition and video streaming on an ESP32/ARM node, which the 2433-parameter network meets at $\approx$10 KB, and a fair benchmark of pre-trained encoders would need their pre-training corpora rather than the Helsinki set alone. *(vi) No hardware-in-the-loop measurement:* every timing reported here comes from desktop CPU and GPU. Power draw, memory under the full acquisition pipeline and long-term robustness on the target board are not measured, and the $\approx$10 KB figure is a parameter-

storage estimate rather than a deployed footprint. Architecture itself is no longer a single point: the `CNN1D` control above shows the normalization finding is EEGNet-specific, which is itself part of the contribution.

a) Clinical reading: The two best configurations sit at opposite ends of a clinically meaningful axis, referential + z-score the more conservative and bipolar + z-score the more sensitive. Since the event-level false-alarm rate is already within a small factor of the cited baselines, the next iteration should calibrate that operating point rather than chase AUROC.

b) Toward literature comparability: The gap to Temko [3] and O'Shea [4] is largely one of protocol: both smooth the window scores, score detections at the *event* level, and report a calibrated rather than fixed threshold. Applying that post-processing here already closes much of the distance with no change to the model, and a per-protocol calibrated threshold and collar would close more. The `CNN1D` control provides a convolutional comparison on these identical splits; a like-for-like Temko-style SVM on the same splits is the natural next experiment and is left as immediate future work.

VI. Conclusion and Future Work

On all 79 babies of the Helsinki dataset, under leak-free 79-fold LOSO and class-aware aggregation, a compact EEGNet gives a cross-patient seizure-detection baseline at pooled AUROC ≈ 0.81 (best 0.807). Two secondary ablations frame the input representation: z-score helps both montages modestly and significantly per subject, the montages are comparable once standardized, and a channel-mixing CNN gains nothing from z-score, so that benefit is architecture-mediated. For the model itself, future work will: (i) extend the event-level operating point introduced here with per-protocol calibrated thresholds and collars, the cheapest path to literature-comparable numbers; (ii) increase the epoch budget with early stopping; (iii) compare against a classical Temko-style SVM baseline on the identical LOSO splits; and (iv) scale the cohort to larger, ideally multicentre data, since performance is bounded by seizure-subject diversity (39 of 79) rather than by window count [9].

This work is a first step toward a low-cost NICU AIoT video-EEG system built by our group (Center for Child Neuroprotection). At ≈ 1.8 ms per window on a CPU there is headroom for a hybrid ESP32+ARM design, the ESP32 handling low-power acquisition and telemetry while ARM processors run TinyML-compressed pipelines for on-device detection and video streaming. In parallel we are collecting vEEG from preterm neonates in a Brazilian NICU, to adapt the model to a setting with a different disease prevalence and a less stable electrical grid.

Acknowledgment

We thank the High Performance Computing Center at UFRN (NPAD/UFRN) for the computational resources and the Brazilian Unified Health System (SUS) for its support. The Helsinki neonatal EEG dataset is used under CC BY 4.0.

References

[1] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizure annotations," *Scientific Data*, vol. 6, art. 190039, 2019.

[2] N. J. Stevenson *et al.*, "Helsinki neonatal EEG dataset," Zenodo, record 2547147, 2018, doi:10.5281/zenodo.2547147, license CC BY 4.0.

[3] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 464–473, 2011.

[4] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture," *Neural Networks*, vol. 123, pp. 12–25, 2020.

[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, art. 056013, 2018.

[6] K. T. Tapani, S. Vanhatalo, and N. J. Stevenson, "Time-varying EEG correlations improve automated neonatal seizure detection," *International Journal of Neural Systems*, vol. 29, no. 4, art. 1850030, 2019.

[7] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *International Journal of Neural Systems*, vol. 29, no. 4, art. 1850011, 2019.

[8] A. M. Pavel *et al.*, "A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial," *The Lancet Child & Adolescent Health*, vol. 4, no. 10, pp. 740–749, 2020.

[9] R. Hogan, S. R. Mathieson, A. Luca, S. Ventura, S. Griffin, G. B. Boylan, and J. M. O'Toole, "Scaling convolutional neural networks achieves expert-level seizure detection in neonatal EEG," *npj Digital Medicine*, vol. 8, art. 17, 2025.

[10] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, art. 051001, 2019.

[11] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers in Human Neuroscience*, vol. 15, art. 653659, 2021.

[12] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous EEG data in BCI," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

Enhancing Situational Awareness for the Deaf and Hard-of-Hearing: A Systematic Review of Haptic and Bimodal Technologies

Edilson Porfírio Filho
Dept. of Teleinformatics Eng.
Universidade Federal do Ceará
Fortaleza, Brazil
edilson.rocha.porfirio@gmail.com

Diego Lima
Dept. of Teleinformatics Eng.
Universidade Federal do Ceará
Fortaleza, Brazil
diegolimax555@gmail.com

J. Otávio Nascimento
Dept. of Teleinformatics Eng.
Universidade Federal do Ceará
Fortaleza, Brazil
otavio.brasil@alu.ufc.br

José Lessa
Dept. of Teleinformatics Eng.
Universidade Federal do Ceará
Fortaleza, Brazil
josefflessa@gmail.com

Osvaldo Cavalcante Neto
Dept. of Teleinformatics Eng.
Universidade Federal do Ceará
Fortaleza, Brazil
osvaldomedeiros@alu.ufc.br

***Abstract*—This article presents a systematic review on the development of wearable tactile assistive devices for people with hearing impairment (DHH). The research analyzes 41 articles focused on technological solutions to mitigate the absence of sound perception, improving situational awareness and safety. The results indicate a transition from theoretical clinical studies to practical systems based on Edge Computing and microphone arrays. It is concluded that the architecture preferred by users is bimodal, utilizing haptic feedback as an immediate attention alert and visual feedback for semantic and directional confirmation. The study also discusses current technical challenges, such as battery management, latency, and social acceptance.**

***Index Terms*—Deaf and Hard-of-Hearing, Sensory Substitution, Vibrotactile Feedback, Sound Source Localization, Sound Awareness, Assistive Technology, Systematic Review.**

I. Introduction

Hearing plays a fundamental role in human communication and environmental perception, being crucial for cognitive and social development, as well as personal safety. It is estimated that about 20% of the global population lives with some degree of hearing loss, and World Health Organization projections indicate that, by 2050, approximately 2.5 billion people will have some form of hearing impairment [42]. For Deaf and Hard-of-Hearing (DHH) individuals, restricted access to the sound world imposes significant daily challenges. The absence of this perception directly compromises situational awareness, making it difficult to recognize imminent alerts such as sirens, horns, fire alarms, or verbal calls, which substantially increases the risk of accidents in urban environments.

Although traditional assistive technologies, such as Hearing Aids (HAs) and Cochlear Implants (CIs), have revolutionized the rehabilitation of thousands of people, they still present technical limitations. In addition to the high cost and the need for surgical intervention in the case of CIs, users frequently report severe difficulties in understanding speech in noisy environments and in spatially localizing sound sources. The inability to accurately identify the direction of a sound-especially those coming from blind spots, such as behind the user-remains an unresolved problem for these conventional prostheses.

Given this scenario, researchers have explored approaches based on sensory substitution to mitigate such deficiencies. This technique consists of compensating for the impaired sense by using remaining sensory channels, notably vision and touch. The use of haptic technology, which employs mechanical actuators to transmit sound information through vibrations on the skin, has emerged as a promising solution. Due to the tactile system's high sensitivity to intensity variations and temporal rhythmic patterns, wearable devices (such as wristbands, glasses, belts, or vests) can provide vital spatial cues. Furthermore, the literature points out that multimodal integration-combining haptic feedback as an immediate alert and visual feedback as directional or semantic support-is highly accepted by users, assisting in the rapid identification of danger without overloading cognitive load.

With the advancement of electronic miniaturization and the employment of Artificial Intelligence processing directly at the edge (*Edge Computing*), combined with the use of *microphone arrays*, it has become possible to design much more robust assistance systems. Such systems seek not only to alert the user that a noise has occurred but also to inform them about what is sounding and where it is coming from. Given the rapid expansion of these innovations, it becomes necessary to systematically evaluate the technical parameters, usability, and efficacy of these tools in the context of auditory accessibility.

II. Methodology

This Systematic Literature Review (SLR) was conducted in adherence to the PRISMA 2020 (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) guidelines.

1) Information Sources and Search Strategy: The electronic search was executed across major peer-reviewed databases and digital publishers: IEEE Xplore, ACM Digital Library, SpringerLink, MDPI, Elsevier/ScienceDirect, PNAS, and the Journal of the Acoustical Society of America (JASA). The literature retrieval covered all indexed publications up to 2025 (spanning from 1982 to 2025). The unified Boolean search strategy was formulated across three core facets: (i) target population/hearing impairment, (ii) acoustic processing/event detection, and (iii) haptic/vibrotactile feedback:

> *Generic search string: ("hearing aid" OR "hearing assist*" OR "assistive listening device" OR "cochlear implant" OR "hearing impairment" OR "hearing loss" OR "deaf" OR "hard of hearing" OR "auditory prosthesis" OR "sensory substitution" OR "hearing impaired" OR "audiology" OR "otology" OR "hearing disorder" OR "hearing deficits") AND ("sound" OR "acoustic" OR "audio" OR "speech" OR "music" OR "auditory" OR "sound localization" OR "sound source localization" OR "acoustic source localization" OR "direction of arrival" OR "DOA" OR "angle of arrival" OR "AOA" OR "sound source detection" OR "azimuth estimation" OR "sound test" OR "audio fingerprint" OR "omnidirection" OR "feature extraction" OR "speech processing" OR "acoustic perception" OR "auditory perception" OR "auditory scene analysis" OR "binaural hearing" OR "acoustic signal" OR "soundscape") AND (vibrotactile OR tactile OR vibration OR haptic OR "vibrotactile feedback" OR "somatosensory" OR "kinesthetic feedback" OR "vibrotactile stimulation" OR "electrotactile") AND ("wearable" OR "wearable*" OR "body-worn" OR "on-body" OR "body-mounted" OR "wearable device*" OR "wearable sensor*" OR "wearable technology" OR "smart textile*" OR "e-textile*" OR "bracelet*" OR "wristband*" OR "smartwatch" OR "smart-watch" OR "head-worn" OR "earable*")*

2) Deduplication and Screening Pipeline: Initial search yields were aggregated into reference management software (Zotero). Automated deduplication was performed by matching digital object identifiers (DOIs) and normalized title-author metadata, followed by manual adjudication (e.g., resolving duplicate workshop and conference pre-prints).

Eligibility screening followed a two-phase independent double-blind evaluation conducted by two independent reviewers, based on predefined criteria. Inclusion focused on acoustic-to-tactile/haptic technologies for sound source localization (I1), sound classification (I2), and related assistive research (I3, I4).

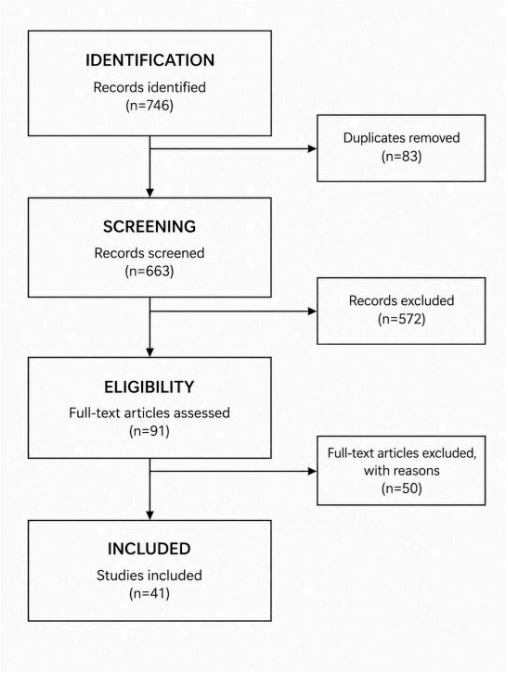


Fig. 1. PRISMA Flow Diagram.

Conversely, studies were excluded if they lacked a haptic component, did not address hearing impairment, or did not use environmental sounds (E1–E3). Non-article documents, outdated technologies, insufficient technical data, irrelevant research, or papers outside English, Portuguese, or Spanish were also excluded (E4–E8).

Disagreements during screening were resolved via structured consensus meetings or adjudicated by a third independent reviewer. Inter-rater reliability was measured using Cohen's kappa coefficient ($\kappa = 0.84$), indicating strong agreement. A final corpus of 41 unique primary studies met all criteria and was synthesized. The detailing of all stages of identification, screening, and eligibility can be visualized in the PRISMA flow diagram (Fig. 1).

3) Limitations and Risk of Bias Assessment: This review faces potential publication bias, as studies reporting positive outcomes in assistive technology are more frequently published than null or negative trials. Furthermore, by restricting the search to peer-reviewed academic literature to ensure scientific rigor, gray literature, proprietary technical reports, and commercial patent databases (e.g., USPTO, WIPO) were excluded, which may omit relevant market-ready implementations and industrial designs.

III. Results

The initial evaluation of the collection of 41 articles reveals that the literature on tactile and visual assistive devices for people with hearing impairment is a field that is still incipient, yet with the potential for rapid expansion.

A clear temporal trend is observed: the vast majority of publications on wearable devices and sound processing is concentrated in the last decade. This growth was strongly driven by the miniaturization and cost reduction of portable

TABLE I
PRISMA SYSTEMATIC ELIGIBILITY AND INCLUSION AUDIT ACROSS PRIMARY STUDIES ($N = 41$)

Reference	Target Condition (DHH/CI/NH)	Acoustic Processing	Tactile/Haptic Modality	Empirical Evaluation	Peer-Reviewed	Eligibility Decision
Chelladurai et al. [1]	DHH	VR Spatial Sound	Head Haptic Ring	Localization Time	Yes	Included (Axis 1)
Buhat et al. [2]	DHH	YAMNet SED	Wristwatch / Headband	Accuracy (Indoor/Outdoor)	Yes	Included (Axis 3)
Michaud et al. [3]	DHH (Targeted)	Microphone Array / DOA	SmartBelt (Waist)	MAE (2.9°) / Selection	Yes	Included (Axis 1)
Sun et al. [4]	Deaf Cyclists	Bandpass Filtering	Smart Helmet	In-situ Functional Test	Yes	Included (Axis 2)
Wong et al. [5]	Able-bodied (Targeted DHH)	Speech Feature Mapping	Forearm Tactile Sleeve	Psychophysical Trials	Yes	Included (Axis 5)
Kim et al. [6]	DHH	SED (BGRNN) + TDOA	Haptic Vibration (Home)	Classification / Conversion	Yes	Included (Axis 4)
Israr et al. [7]	Deaf / NH	Formant Extraction	Multi-finger Display	Absolute Identification	Yes	Included (Axis 5)
Rahman & Fattah [8]	DHH	Active Noise / Speech Rec.	Smart Glass Binaural Cue	Functional Verification	Yes	Included (Axis 3)
Rahul et al. [9]	Deaf-Blind	Speech Recognition	Dynamic Braille Ring	User Translation Trials	Yes	Included (Axis 4)
Singh et al. [10]	DHH / Headphone Users	10-Layer CNN / MFCC	Headphone Vibration	Accuracy (Clean/Noisy)	Yes	Included (Axis 2)
Nakata et al. [11]	DHH	Frequency-to-Vibration	Headphone / Hat Haptics	Localization / Sound Type	Yes	Included (Axis 3)
Orzech et al. [12]	Hearing Loss	Systematic Methodology	Haptic Review	Usability / Feasibility	Yes	Included (Axis 5)
Abhiram et al. [13]	Hearing Impaired Drivers	Sequential CNN	Microvibrator Glove	Real-time Siren Alert	Yes	Included (Axis 2)
Fletcher et al. [14]	CI Users / NH	Tactile Intensity Differences	Tricep / Palmar / Dorsal	Sensitivity (TID Thresholds)	Yes	Included (Axis 5)
Mirzaei et al. [15]	DHH	VR Sound Localization	Haptic + LED Feedback	Localization Speed	Yes	Included (Axis 1)
Fletcher [16]	Hearing-Impaired	Electro-haptic Audio	Wrist Haptics	Perceptual Review	Yes	Included (Axis 5)
Fletcher & Zgheib [17]	CI / Hearing Aid Users	Signal Compression / DSP	Wrist-Worn Haptics	Localization Error (degrees)	Yes	Included (Axis 1)
Fletcher et al. [18]	CI Users	Electro-Haptic Dual Stim.	Wrist Tactors	Localization Task	Yes	Included (Axis 1)
Mirzaei et al. [19]	DHH	3D Spatial Audio	EarVR (Headset Haptics)	Completion Time / Accuracy	Yes	Included (Axis 1)
Goodman et al. [20]	DHH	Multimodal Sound Awareness	Smartwatch Tactons	Wizard-of-Oz In-situ	Yes	Included (Axis 5)
Sakuma et al. [21]	Hearing Impaired	Julius Keyword Engine	Back Haptic Actuators	Recognition / Localization	Yes	Included (Axis 3)
Kim et al. [22]	DHH	Multi-label BGRNN	12-Tactor Sleeve	F1-Score / MOS Trials	Yes	Included (Axis 2)
Yağanoğlu & Köse [23]	DHH / NH	Audio Fingerprint (AF)	Fingertip Vibration	Real-Time Accuracy (94%)	Yes	Included (Axis 2)
Fletcher et al. [24]	Simulated CI	Low-frequency Tactile	Fingertip Transducer	Speech-in-Noise Recovery	Yes	Included (Axis 5)
Yağanoğlu & Köse [25]	DHH / NH	ReliefF + SVM / k-NN	Fingertip Actuators	Direction Accuracy (98%)	Yes	Included (Axis 3)
Desloge et al. [26]	DHH / NH / Simulated	Auditory Gap Thresholds	Vibrotactile (Finger)	Quantitative Psychophysics	Yes	Included (Axis 5)
Russo et al. [27]	DHH / NH	Timbre / Waveform	Voice Coils (Back)	Discrimination Trials	Yes	Included (Axis 5)
Israr et al. [28]	Hearing Assumed	Consonant Contrast Coding	3-Digit Tactual Display	Psychophysical Index	Yes	Included (Axis 4)
Aker et al. [29]	NH	Audio-Tactile Congruence	Haptic Actuator	Preference Rating	Yes	Included (Axis 5)
Chin et al. [30]	DHH (Targeted)	EfficientNet / STFT	Wrist Vibrator	Edge AI / Offline Accuracy	Yes	Included (Axis 2)
Kidd Jr. et al. [31]	CI Users / NH	Triple Beamformer	Binaural Acoustic/Tactile	Speech Identification	Yes	Included (Axis 1)
Spiegel et al. [32]	Tinnitus Observers	Audio-Visual-Somatosensory	Somatosensory	Randomized Control Trial	Yes	Included (Axis 5)
Plant [33]	Profoundly Deaf	Speech / Env. Sounds	Tactile Transducer	Word/Sound Recognition	Yes	Included (Axis 4)
Shimoyama [34]	Hearing Impaired	Sound Tracking / Interaural	Dual Vibrators (Shoulders)	Angle Error ($< 5°$)	Yes	Included (Axis 1)
Paisa et al. [35]	CI Users	Multi-instrument Segregation	Vibrotactile Furniture	Qualitative / Observational	Yes	Included (Axis 4)
Mirzaei et al. [36]	DHH	3D Sound in VR	Haptic Suit / Ear Motors	Task Completion Time	Yes	Included (Axis 1)
Fletcher [37]	CI / Hearing-Impaired	Music Enhancement	Haptic Stimulation	Systematic Evaluation	Yes	Included (Axis 5)
Guilleminot & Reichenbach [38]	NH / Simulated	Syllabic Rate Audio	Vibrotactile (Fingertip)	EEG / Psychophysics	Yes	Included (Axis 5)
Honda et al. [39]	DHH	Acoustic Pressure / DSP	Ontenna (Hair/Ear/Sleeve)	User Experience Trials	Yes	Included (Axis 4)
Sofack et al. [40]	Single-Sided Deafness	Vibrotactile Stimulation	Skin Mechanoreceptors	Cortical EEG / ERP	Yes	Included (Axis 5)
Dupin et al. [41]	Normal Sensing	Haptic Coupling	Bimanual Tactors	Psychophysical Index	Yes	Included (Axis 5)

Note: The eligibility decision indicates the thematic axis (Axis 1 to 5) assigned to each primary study, as defined in Section III.

microcontrollers (such as Raspberry Pi, Arduino, and ESP32), by the advancement of *Deep Learning* algorithms capable of operating locally (*Edge Computing*), and by the popularization of Virtual Reality (VR) environments, which created yet another environment that requires inclusion. Although there are foundational studies from previous decades on the psychophysical capacity of the skin to discriminate frequencies, modern application is mostly focused on the *design* of embedded systems for everyday use.

Regarding technological maturity (*Technology Readiness Level* - TRL), the field is predominantly composed of proofs of concept and academic prototypes tested in controlled environments with small groups. However, there are notable exceptions that have overcome the prototyping barrier. The biggest example identified is the *Ontenna* [39], developed in Japan, which evolved into a mass-manufactured commercial product, adopted in more than 80% of schools for the deaf in the country for teaching rhythm and programming.

The user demands that were the focus of the analyses were divided into two main directions (often complementary): the classification of important sound events (active alerts about sirens, horns, and calls); and the three-dimensional spatial localization of the source (ensuring safety and situational awareness).



Fig. 2. Prototype of the *Ontenna* device, a wearable acoustic sensory interface designed to be attached to the hair, earlobe, or clothing. It uses a condenser microphone to capture ambient sound and converts its characteristics into real-time tactile and visual feedback through 256-stage vibrations and LED light intensities [39].

The analysis of the selected publications revealed five thematic axes of research, categorized by the objective of the interaction between the user and the environment.

The first axis covers devices focused strictly on spatial localization (nine articles identified). The central objective of these projects is situational and navigational awareness (*Direction of Arrival* - DoA), without a focus on sound classification. Examples include instrumented belts (*SmartBelt* [3]), bracelets for spatial enhancement of unilateral cochlear implant users, and spatial mapping in Virtual Reality (*EarVR* [19]).

The second axis encompasses devices that act as sound classifiers (six articles). They monitor the environment to notify the user about specific events, such as doorbells or horns, through pattern recognition and residential IoT monitoring, delegating the spatial identification to the user themselves after the alert.

A more complex category, representing an advanced minority (five studies), focuses on hybrid devices that integrate localization and classification. Projects such as *AICanHear* [2] and *Smart Glass* [8] use AI (*Edge Computing*) to recognize the nature of the sound and microphone arrays to point out its origin, being vital for identifying threats in blind spots, such as the rear approach of vehicles.

The fourth axis moves away from notification systems to explore the direct conversion of sound (six articles). Here fall the "tactile vocoders" and devices for musical appreciation, such as the *Ontenna* [39] and multisensory concert furniture [35]. They do not emit simplified alerts, but translate the raw sound wave or phonemes directly into physical stimuli for the skin.

Finally, the largest portion of the collection (fifteen articles) concentrates on fundamental research and reviews. These are clinical and neuroscientific studies that investigate the psychophysics of touch, vibrotactile perception thresholds, and *cross-modal* plasticity, the reorganization of the auditory cortex in deaf individuals to process somatosensory stimuli, validating the foundational theory for the construction of the aforementioned engineering solutions.

To synthesize these five axes in greater technical depth, the results are organized into three core pillars: (A) Spatial Localization (Axes 1 and 3), (B) Sound Event Detection (Axes 2 and 3), and (C) Sensory Transduction and Response Interfaces (Axes 4 and 5).

A. Spatial Localization and Direction of Arrival (DoA)

Studies within this category (encompassing Axis 1 and hybrid architectures from Axis 3) focus primarily on restoring horizontal spatial awareness and estimating the direction of arrival (DoA) of acoustic sources. Hardware configurations generally rely on multi-microphone arrays (such as wearable belts or glasses) using time-difference-of-arrival (TDoA) and generalized cross-correlation (GCC-PHAT) algorithms, head-mounted binaural interaural level/time difference (ILD/ITD) beamformers, or direct spatial audio integration within virtual reality (VR) engines.

Quantitative performance is reported across heterogeneous paradigms, reflecting distinct physical and virtual setups. In wearable physical devices, accuracy is commonly evaluated via angular error metrics, such as Mean Absolute Error (MAE $= 2.90^\circ$ in *SmartBelt* [3]) or real-time dynamic tracking error below 5° [34]. Clinical trials involving cochlear implant (CI) users quantify performance through root-mean-square (RMS) localization error, demonstrating significant error reductions after training with electro-haptic wrist stimulation (e.g., from 47.2° to 22.7° [18] and 25.2° to 22.6° [17]). Conversely, virtual reality studies assess efficiency through task completion time (e.g., *EarVR* [19], [36], *SoundHapticVR* [1]), showing that localized haptic cues significantly accelerate target search times and spatial identification in immersive environments.

To enable an objective comparison across spatial hearing technologies, Table II synthesizes the quantitative performance of directional systems across both standalone spatial trackers (Axis 1) and hybrid localization-classification architectures (Axis 3), as visually conceptualized in the intersection diagram in Fig. 3. The table reports standardized metrics such as task completion time in VR, root-mean-square (RMS) angular error, directional classification accuracy, and non-standardized outcomes.

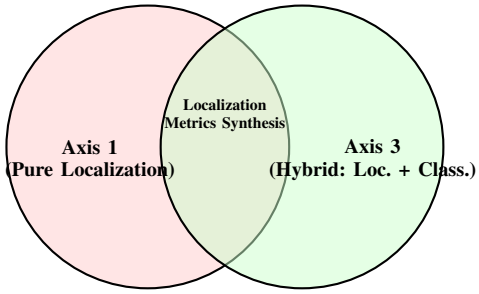


Fig. 3. Venn diagram highlighting the scope of the quantitative synthesis, focusing on spatial localization and direction-of-arrival metrics across standalone spatial trackers (Axis 1) and hybrid localization-classification systems (Axis 3).

B. Sound Event Detection (SED) and Semantic Classification

The second pillar consolidates systems engineered to automatically recognize and classify specific acoustic events (Axis 2) and hybrid systems coupling classification with spatial localization (Axis 3). Evaluating quantitative performance metrics across these implementations is essential for establishing an objective baseline, moving beyond qualitative feasibility to assess real-world viability. Methodological implementations range from classical DSP thresholding and speech keyword-spotting engines to support vector machines (k-NN/SVM) and deep neural architectures (e.g., CNNs, BGRNNs, and edge-deployed EfficientNet ensembles).

To address the functional trade-offs and substantial methodological heterogeneity identified in the literature, Table III synthesizes the quantitative and experimental performance of sound event classification systems, highlighting the intersection between standalone classifiers and hybrid architectures

TABLE II
QUANTITATIVE PERFORMANCE METRICS FOR SPATIAL SOUND SOURCE LOCALIZATION SYSTEMS (AXES 1 AND 3)

Reference	Tracking / Processing Method	Acoustic / Spatial Input	Localization Performance (Metric)	Spatial Setup / Tracking Range	Feedback Modality / Interface	Participants (N and Profile)
Axis 1: Pure Spatial Localization Systems						
Chelladurai et al. [1]	Multi-channel Audio Spatializer	3D Virtual Sound (Meta Quest)	Time: approximately 9 seconds (Accuracy: 89.81%)	360° continuous (4.0 m radius)	LRF Voice-Coil Ring (Head)	$N = 25$ (DHH)
Michaud et al. [3]	GCC-PHAT / TDoA Lookup	8-Microphone Array	MAE = 2.90° (Selection: 92.3%)	360° horizontal (2.0 m distance)	15-Motor SmartBelt (Waist)	$N = 2$ (Bench evaluation)
Mirzaei et al. [15]	Real-time Stereo Energy Balance	3D Spatial Audio (Unreal Engine)	Time: 17.53 ± 1.25 s (Haptic+LED)	Left / Right visual hemifield	2 Ear Vibro-motors + 2 LEDs	$N = 20$ (10 DHH, 10 NH)
Fletcher & Zgheib [17]	Linked Multiband Compression	BTE HRTF (Speech Noise)	RMS: 25.2° (Pre) → 22.6° (Post-train)	11 locations (−75° to +75°)	Dual-wrist HVLab Vibrometers	$N = 32$ (Normal-hearing)
Fletcher et al. [18]	Electro-haptic ILD Extraction	BTE HRTF (Speech Envelope)	RMS: 47.2° (Audio) → 22.7° (Post-train)	11 loudspeakers (−75° to +75°)	Dual-wrist Vibrometers + CI	$N = 12$ (CI users: 7 uni, 5 bi)
Mirzaei et al. [19]	Real-time Stereo Level Tracking	3D Spatial Audio (Google Resonance)	Time: 14.7 ± 1.8 s (EarVR) vs. 29.6 s (None)	6 directions (3D VR enclosed room)	EarVR (2 Ear Coin Motors)	$N = 40$ (20 DHH, 20 NH)
Kidd Jr. et al. [31]	Triple Binaural Beamforming	16-MEMS Microphone Array	Group Mean Error ≈ 15° (Triple 90°)	Horizontal plane (−90° to +90°)	Binaural headphone simulation	$N = 6$ (Normal-hearing)
Shimoyama [34]	Short-time Phase Difference / ITD	Binaural Ear Microphones	Dynamic tracking error < 5°	Omnidirectional (2.0 m platform)	2 Shoulder Transducers (Vest)	$N = 1$ (Simulated deprivation)
Mirzaei et al. [36]	4-Directional Wireless Mapping	3D Spatial Audio (Unreal Engine)	Time: 21.41 s (Ears) vs. 27.05 s (No suit)	4 main positions (Front/Back/L/R)	Haptic VR Suit (2–4 Motors)	$N = 20$ (DHH participants)
Axis 3: Hybrid Systems (Spatial Localization + Sound Event Classification)						
Buhat et al. [2]	Loudness Thresholding / TDoA	4-Electret Array + Grove Sensors	Accuracy (indoors): 56.67% (F/R), 50.00% (B), 53.33% (L)	4 cardinal directions (Indoor/Outdoor)	Wristwatch (LCD + RGB LEDs + Motor)	$N = 13$ (DHH participants)
Rahman & Fattah [8]	Dynamic Binaural Cues (ITD Peak)	5-Microphone Array (1 Front, 4 Sides)	Direction error ≈ 5–15% (with ≥ 3 microphones)	Horizontal plane (0° to 330°, 30° steps)	Smart Glass (3-Color LEDs + 2 Motors)	-
Nakata et al. [11]	4-Channel Micro-Array TDoA	CRESCENT-031 Direction Board	$t(2) = -0.277, p = 0.808, d = -0.204$	6 directions (0°, 60°, . . . , 300°)	Beret Haptuators / Headphone Motors	$N = 3$ (DHH Levels 2 and 3)
Sakuma et al. [21]	CSP Cross-power Spectrum Phase	2-Microphone Array	Accuracy: $87.0 \pm 11.0\%$ (Horn), $86.0 \pm 15.6\%$ (Voice)	5 rear directions (−90° to +90°, 45° steps)	Neck-worn Collar (5 Vibration Motors)	$N = 10$ (Isolated hearing)
Yağanoğlu & Köse [25]	ReliefF + SVM / k-NN Classifier	4 Microphones (Orthogonal)	Accuracy: 98.0% (1 m), 94.0% (4-dir), 91.0% (8-dir)	4 to 8 directions (1–4 m distance)	2 Fingertip Motors + 4 Directional LEDs	$N = 56$ (6 DHH, 50 NH)

depicted in Fig. 4. As detailed in the table, performance metrics vary considerably, encompassing overall accuracy, F1-scores, and error rates—with non-accuracy or qualitative evaluations (where parameters not explicitly disclosed are denoted as NR, *Not Reported*). Furthermore, temporal reporting reveals substantial disparities between input temporal windowing and end-to-end response delays, while participant cohorts diverge across deaf and hard-of-hearing (DHH) individuals, normal-hearing (NH) control groups, and subjective MOS evaluations.

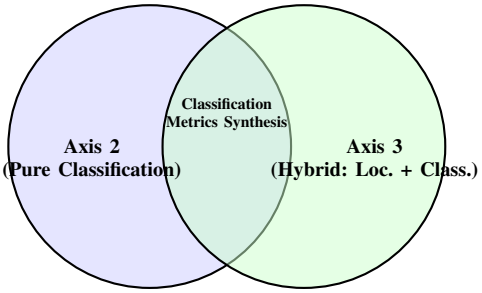


Fig. 4. Venn diagram highlighting the scope of the quantitative synthesis, focusing on classification and sound event detection metrics across standalone classifiers (Axis 2) and hybrid localization-classification systems (Axis 3).

C. Sensory Transduction and Response Interfaces

The third pillar addresses how acoustic information is physically conveyed to the mechanoreceptors of the body, encompassing direct sound-to-haptic substitution (Axis 4) and foundational psychophysical investigations (Axis 5), as synthesized in the taxonomic mapping in Fig. 5. The selection of actuation technology dictates the directional precision, frequency fidelity, and acoustic resolution of the assistive device:

- **Mixed (Visual and Low-Resolution Haptic):** The majority of basic alert systems adopt conventional vibration motors (coin type or Eccentric Rotating Masses - ERM) combined with LEDs or *displays*. Vibration works as an attention trigger, while vision confirms the urgency and directionality (e.g., *SI-LERT* [13]).
- **Exclusively Haptic (High Fidelity):** To relieve the visual load or transmit greater sound complexity, advanced systems adopt *Haptuators* (Linear Actuators) or *Voice-coil* transducers (e.g., *SoundHapticVR* [1]). They mechanically replicate the nuances of the audio wave (frequency, texture, and continuous volume).

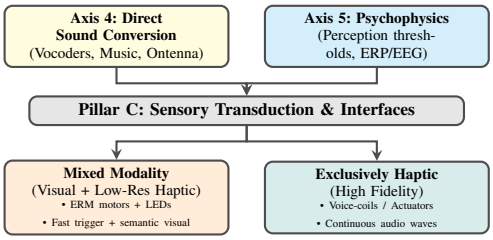


Fig. 5. Taxonomic synthesis of Pillar C (Sensory Transduction and Response Interfaces), mapping Axes 4 and 5 into mixed bimodal and exclusively haptic paradigms.

Beyond discrete alerts, continuous transduction models and psychophysical studies (Axes 4 and 5) explore sensory limits, speech formant and gap discrimination [5], [7], [26], [28], cortical plasticity [38], [40], and spectral compression for musical timbre [27], [29], [35], [37], [39]. However, complex vibratory encoding (*tactons*) often causes sensory overload and high cognitive load. Consequently, bimodal architectures—pairing intuitive haptic triggers with immediate visual confirmation—emerge as the preferred design paradigm for situational awareness and usability.

IV. DISCUSSION

From this research, the understanding is consolidated that haptic feedback (or electro-haptic stimulation) is a powerful resource to compensate for the limitations of traditional medical prostheses, especially in the restoration of spatial awareness, in the improvement of speech rhythm tracking in noisy environments, and in the reintegration of musical perception. However, usability research strongly demonstrates that the architecture preferred by users is not exclusively tactile, but rather **bimodal (haptic and visual)** [15]. Vibration acts masterfully as an immediate *push notification*, while the visual complement (LEDs or screens) informs the exact identity of the noise, minimizing the need to memorize complex vibratory patterns.

This architectural preference is strongly supported by the findings of Goodman et al. [20], in which 13 participants expressed significant concerns regarding visual-only interfaces due to the risk of missing critical acoustic events; as one

TABLE III
QUANTITATIVE PERFORMANCE METRICS FOR SOUND EVENT DETECTION AND CLASSIFICATION SYSTEMS (AXES 2 AND 3)

Reference	Model / Classifier	Feature Extraction	Classification Performance	Temporal Window / Latency	Target Sound Classes	Number of Test Participants
Axis 2: Standalone Sound Event Detection and Classification Systems						
Sun et al. [4]	Amplitude Threshold / DSP	Bandpass Filtering (ADC)	Qualitative bench verification	NR / NR	1 class (Vehicle horns)	-
Singh et al. [10]	10-Layer CNN	MFCC ($O(n \log n)$)	Accuracy: 99.54% (Clean), 91.00% (Noisy NAWED)	1.0 s window (50% overlap) / NR	20 attention/emergency words	-
Abhiram et al. [13]	Sequential CNN	MFCC + Chroma + Mel Spectrogram	NR	7.0 s standardized window / NR	2 classes (Emergency sirens vs. noise)	-
Kim et al. [22]	Multi-label BGRNN	Noise-reduced Spectrogram + TDOA	F1-Score: 84.2%	NR / NR	14 home sound classes (polyphonic)	$N = 14$ (MOS evaluation)
Yağanoğlu & Köse [23]	Audio Fingerprint (AF)	Spectrogram Peak Hashing	Accuracy: 98.0% (Clean), 94.0% (SNR 20 dB), 90.0% (15 dB)	NR / 1.9 s	8 sound classes (indoor and outdoor)	$N = 55$ (5 DHH, 50 NH)
Chin et al. [30]	EfficientNet + Fuzzy Ensemble	STFT Spectrograms	Accuracy: 97.05% (Offline), 95.22% (Edge AI)	1.5 s window / NR	7 classes (4 vocalizations, 3 sirens)	-
Axis 3: Hybrid Systems (Semantic Classification + Spatial Localization)						
Buhat et al. [2]	YAMNet (MobileNetV1)	Spectrogram (Google AudioSet)	Accuracy: 70.48% (Indoor), 63.33% (Outdoor)	NR / NR	8 classes (Sirens, horns, dog, baby...)	$N = 13$ (DHH participants)
Rahman & Fattah [8]	Template Matching	Amplitude/Frequency Envelope	BER analysis	NR / NR	Speech keywords, horns, dog bark	-
Nakata et al. [11]	Frequency Mapping	FFT $\rightarrow$ Linear IFFT	$t(2) = 3.90, p = 0.0599, d = 3.00$	NR / NR	15 daily sound sources (6 categories)	$N = 3$ (DHH participants)
Sakuma et al. [21]	Julius Speech Engine	Keyword Recognition	Accuracy: $68.6 \pm 13.8\%$	NR / NR	5 sound keywords (Table 1 patterns)	$N = 7$ (Participants)
Yağanoğlu & Köse [25]	SVM / k-NN	L_{max} + Zero Crossing Rate (ZCR)	Accuracy: 95.0% (Normal voice), 91.8% (Shouting)	NR / 0.68 s classification response time	Voice levels (Shouting vs. Normal speech)	$N = 56$ (6 DHH, 50 NH)

participant remarked, *"I'm not going to be looking at my watch every two minutes when I'm out and about"* [20]. In contrast, when evaluating a simple vibration feedback mechanism, participants emphasized its efficacy in *"alerting to a situation"* and serving as a stimulus to *"trigger you to look at your watch"* [20], although user feedback indicated that vibration intensity should be customizable [20]. Conversely, for more complex systems encoding specific acoustic events into distinct vibration patterns (*tactons*), critical usability bottlenecks emerge—particularly in time-sensitive or hazardous scenarios [20]. As noted by a participant in the same study: *"I had to wait and wait and then the vibration had finished and by the time I finished decoding what it was [...] I'd missed whatever happened"* [20]. Furthermore, relying exclusively on intricate tactile patterns can substantially increase cognitive load and lead to mental fatigue over extended periods, as highlighted in the systematic review by Orzech et al. [12] (citing Wang et al.): *"If I have to wear it [...] for a long time, I will likely have sensory overload"* [12]. Taken together, these studies corroborate that the most effective and usable design paradigm is a bimodal architecture—employing an adjustable, low-complexity vibratory cue strictly as an asynchronous attention trigger, while delegating semantic classification to an immediate visual display [12], [20].

Furthermore, despite these theoretical and prototypical advances, the transposition of these solutions into daily life encounters critical engineering and usability challenges:

1) **Battery Management and Autonomy:** The continuous processing of audio and the triggering of mechanical motors demand high energy. While the market requires at least 14 hours of daily autonomy, many prototypes hit time limits or rely on heavy *power banks*, although BLE technologies and adaptations in *smartwatches* (such as the *Buzz*) are already starting to overcome this bottleneck. [11], [16]
2) **Processing Time and Latency:** The speed of capture and response is a matter of safety. The latency of mechanical motors added to the processing time of neural networks can exceed several seconds in some models. The use of *Edge Computing* is the main bet to mitigate cloud delay. For continuous stimuli, however, delays in the millisecond range are usually tolerated by the brain thanks to cross-modal temporal recalibration. [23], [30], [37]
3) **Physical Coupling and Comfort:** Vibratory effectiveness depends on constant and firm contact with the skin. The need for mechanical pressure (velcro, elastics) can generate long-term discomfort. Furthermore, flaws in mapping the sound dynamic range can transform a loud ambient noise into an uncomfortable tactile shock. [14], [35], [36]
4) **Cognitive Load and Fatigue:** In noisy environments, the constant reception of tactile notifications generates mental exhaustion and can distract the user from their peripheral visual stimuli. The use of complex *tactons* also requires great decoding effort. The need for highly customizable AI systems becomes evident, allowing the user to apply silencing filters or set alerts only for critical sounds. [12], [20]
5) **Social Acceptance and Aesthetics:** The psychological comfort of the user is a major factor for the adoption of the solution. Devices that are intrusive or too robust suffer strong rejection for stigmatizing the disability in public. Solutions embedded in conventional devices, such as *smartwatches* and *smartphones*, show the highest rates of social acceptance. [12], [16], [20]

A. Conceptual Framework and Proposed Architecture

To bridge the empirical gaps identified in this systematic review—particularly regarding energy efficiency, cognitive load, and usability—and aligned with preliminary field insights from the Ceará Institute of Education for the Deaf (ICES), we propose a conceptual bimodal assistive architecture. It is important to emphasize that this section presents a prospective design proposal aimed at guiding future prototyping and real-world validation, rather than an empirically validated outcome of the literature review itself. The architecture is organized into two decoupled layers: mobile-based Edge AI processing and a low-complexity wearable haptic-visual interface.

1) Processing and Personalization Layer on Mobile Devices: Unlike approaches in the literature that concentrate digital signal processing (DSP) and Machine Learning (ML) inference exclusively on dedicated wearable hardware, which increases energy consumption and device cost, a decentralized approach is proposed. Audio processing is delegated to the user's mobile phone through a dedicated application.

This choice is justified by three critical factors:

- **Hardware Availability:** It leverages the computational capacity of an underutilized device that the user already carries daily;
- **Algorithmic Adjustment and Flexibility:** Artificial Intelligence frameworks for mobile devices are used, allowing the local execution of pre-trained classification models with the flexibility for the user to register and calibrate new personalized acoustic signatures (e.g., specific domestic alerts, their own name...);
- **User Interface (UI):** Facilitates the customization of sensitivity profiles and sound categories in an intuitive way.

2) Embedded Haptic and Visual Actuation Layer: The physical interface with the user consists of a microcontrolled bracelet based on the ESP32-S3 SoC, integrated with the cell phone via the *Bluetooth Low Energy* (BLE) protocol. To mitigate the usability barriers pointed out by ICES experts, the tactile feedback paradigm was reformulated.

Instead of encoding complex messages through multiple spatial or frequency vibration patterns, whose learning curve is steep and prone to interpretation errors if a pulse is missed, the device uses a single eccentric rotating mass (ERM) motor acting strictly as an *attention trigger* (asynchronous alert).

Once the user's attention is awakened by the tactile stimulus, the semantics of the message are transmitted visually by a matrix of four multicolored LEDs (RGB). Each color or luminous pattern represents a sound macrocategory (e.g., danger alerts, domestic events, interpersonal calls), drastically reducing the cognitive load necessary for decoding the stimulus.

Furthermore, in line with the *feedback* obtained from users, the immediate spatial localization mapping (Direction of Arrival - DoA) was temporarily discarded in the hardware. The device prioritizes the immediate alert of the event's occurrence, delegating the search for the sound source to the user's natural visual scanning, which optimizes the global response time of the system and reduces the complexity of the embedded circuit.

3) Engineering Challenges and Future Validation: As real-environment testing with end-users is still in the planning phase, the validation of this architecture presents co-design and engineering challenges that need to be investigated, such as:

1) **Temporal Determinism and Latency:** Evaluate the impact of task scheduling in multitasking mobile operating systems (Android/iOS) on the total pipeline delay (*audio capture* $\rightarrow$ *AI inference* $\rightarrow$ *BLE transmission* $\rightarrow$ *peripheral activation*), ensuring that the total time does not compromise safety in critical scenarios.
2) **Ergonomics and Physical Robustness:** Analyze the acceptance of the bracelet in continuous use, considering biomechanical and environmental factors such as thermal dissipation, moisture (sweat), and mechanical comfort of the vibrations on the skin.

V. Conclusion

This systematic review evidenced that the development of assistive devices for Deaf and Hard of Hearing individuals is in a period of rapid expansion, driven by the miniaturization of *hardware* and local processing with artificial intelligence (*Edge Computing*). The focus of innovations has migrated from purely clinical testing to the creation of practical wearable systems that solve two fundamental demands: the spatial localization of sound sources (using arrays of multiple microphones) and the acoustic classification of urgent events (employing neural networks and sound feature extraction).

For the transmission of information, haptic feedback is consolidated as an effective pathway for sensory augmentation. The literature proves that tactile stimulation restores spatial awareness, assists in urban safety, improves speech comprehension in noisy environments, and returns engagement in musical appreciation. However, the most efficient architecture, and the one preferred by users, is the bimodal one: vibration operates as an immediate alert, while visual feedback provides semantic and directional details, drastically reducing the cognitive load.

Despite the expressive potential for inclusion, the transition from the laboratory to the market requires overcoming practical technical challenges, above all the optimization of battery consumption, the reduction of mechanical latency, and the guarantee of comfort during continuous contact with the skin; but projects like the *Ontenna* [39] show that it is possible to make this transition. Furthermore, for future work, it is essential to prioritize personalization through adjustable algorithmic filters, avoiding sensory overload for the user in noisy environments. It is concluded that integrating these assistive technologies into already consolidated social devices, such as *smartwatches* and smart glasses, appears to be an extremely promising path to minimize stigmas and revolutionize the autonomy and quality of life of the DHH community.

References

[1] P. K. Chelladurai, Z. Li, M. Weber, T. Oh, and R. L. Peiris, "SoundHapticVR: Head-Based Spatial Haptic Feedback for Accessible Sounds in Virtual Reality for Deaf and Hard of Hearing Users," in *Proc. 26th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, New York, NY, USA, 2024, Art. no. 31, pp. 1–17. doi: 10.1145/3663548.3675639.

[2] R. G. A. Buhat et al., "AICanHear: Assistive Alerting Device for Deaf and Hard-of-Hearing People (DHH)," 2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Kuala Lumpur, Malaysia, 2023, pp. 1-6, doi: 10.1109/ICRAIE59459.2023.10468094.

[3] S. Michaud, B. Moffett, A. T. Rousiouk, V. Duda and F. Grondin, "SmartBelt: A Wearable Microphone Array for Sound Source Localization with Haptic Feedback," 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, Korea, Republic of, 2023, pp. 1950-1955, doi: 10.1109/RO-MAN57019.2023.10309323.

[4] P. Sun, J. Luo, C. Wang and D. Wu, "An Improved Smart Helmet for Safe Travel of Deaf People Based on Embedded System," 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2024, pp. 183-188, doi: 10.1109/ICETCI61221.2024.10594072.

[5] E. Y. Wong, A. Israr and M. K. O'Malley, "Discrimination of consonant articulation location by tactile stimulation of the forearm," 2010 IEEE Haptics Symposium, Waltham, MA, USA, 2010, pp. 47-54, doi: 10.1109/HAPTIC.2010.5444681.

[6] G. Y. Kim, S. -S. Shin, J. Y. Kim and H. -G. Kim, "Haptic Conversion Using Detected Sound Event in Home Monitoring System for the Hard-of-Hearing," 2018 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE), Dalian, China, 2018, pp. 1-6, doi: 10.1109/HAVE.2018.8547508.

[7] A. Israr, C. M. Reed and H. Z. Tan, "Discrimination of Vowels with a Multi-finger Tactual Display," 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, Reno, NV, USA, 2008, pp. 17-24, doi: 10.1109/HAPTICS.2008.4479907.

[8] M. L. Rahman and S. A. Fattah, "Smart Glass for Awareness of Important Sound to People with Hearing Disability," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1527-1530, doi: 10.1109/TENSYMP50017.2020.9230986.

[9] A. R, A. A, E. K. V, S. Steephen and N. R. Thomas, "Enhancing Deaf Blind Communication Through Real Time Speech-Text and Braille Translation," 2025 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI), Kannur, India, 2025, pp. 1-7, doi: 10.1109/APCI65531.2025.11137174.

[10] I. Singh, N. Aggarwal and V. P. Singh, "AcouSense: A Deep Learning Based Real-Time Acoustic Triggering System," 2021 IEEE Bombay Section Signature Conference (IBSSC), Gwalior, India, 2021, pp. 1-6, doi: 10.1109/IBSSC53889.2021.9673469.

[11] M. Nakata *et al.*, "A Vibrotactile Device for Enabling Sound Localization and Identification for Deaf and Hard of Hearing Individuals," *APMAR'24: The 16th Asia-Pacific Workshop on Mixed and Augmented Reality*, 2024.

[12] G. Orzech, Y. Luo, and G. Huang, "Haptic Technology for Hearing Loss: A Systematic Review of Technical Feasibility, Usability, and User Experience," Proc. Hum. Factors Ergon. Soc. Annu. Meeting, vol. 68, no. 1, Aug. 2024, doi: 10.1177/10711813241275928.

[13] A. S. P. Abhiram, S. Umamaheswari, M. S. Muhammed Shybin, A. Saju Campbell, V. G. Sreena, and N. R. Raji, "SI-LERT (Assistive Device For Hearing Impaired Drivers)," in Proc. 2024 IEEE 7th Int. Conf. Circuit Power Comput. Technol. (ICCPCT), 2024, doi: 10.1109/ICCPCT61902.2024.10673416.

[14] M. D. Fletcher, J. Zgheib, and S. W. Perry, "Sensitivity to Haptic Sound-Localization Cues at Different Body Locations," Sensors, vol. 21, no. 11, p. 3770, May 2021, doi: 10.3390/s21113770.

[15] M. Mirzaei, P. Kán, and H. Kaufmann, "Head Up Visualization of Spatial Sound Sources in Virtual Reality for Deaf and Hard-of-Hearing People," in Proc. 2021 IEEE Virtual Reality 3D User Interfaces (VR), Mar. 2021, pp. 582-587, doi: 10.1109/VR50410.2021.00083.

[16] M. D. Fletcher, "Using haptic stimulation to enhance auditory perception in hearing-impaired listeners," Expert Rev. Med. Devices, vol. 18, no. 1, pp. 63-74, Jan. 2021, doi: 10.1080/17434440.2021.1863782.

[17] M. D. Fletcher and J. Zgheib, "Haptic sound-localisation for use in cochlear implant and hearing-aid users," Sci. Rep., vol. 10, p. 14171, Aug. 2020, doi: 10.1038/s41598-020-70379-2.

[18] M. D. Fletcher, R. O. Cunningham, and S. R. Mills, "Electro-Haptic Enhancement of Spatial Hearing in Cochlear Implant Users," Sci. Rep., vol. 10, p. 1621, Jan. 2020, doi: 10.1038/s41598-020-58503-8.

[19] M. Mirzaei, P. Kán, and H. Kaufmann, "EarVR: Using Ear Haptics in Virtual Reality for Deaf and Hard-of-Hearing People," IEEE Trans. Vis. Comput. Graph., vol. 26, no. 5, pp. 2084-2093, May 2020, doi: 10.1109/TVCG.2020.2973441.

[20] S. Goodman, S. Kirchner, R. Guttman, D. Jain, J. Froehlich, and L. Findlater, "Evaluating Smartwatch-based Sound Feedback for Deaf and Hard-of-hearing Users Across Contexts," in Proc. 2020 CHI Conf. Hum. Factors Comput. Syst. (CHI '20), Honolulu, HI, USA, Apr. 2020, pp. 1-13, doi: 10.1145/3313831.3376406.

[21] R. Sakuma, Y. Fujita, and K. Zempo, "Audio Substituting Haptic System to Aware the Sound from Backwards," in Adjunct Proc. 32nd Annu. ACM Symp. User Interface Softw. Technol. (UIST '19 Adjunct), New Orleans, LA, USA, Oct. 2019, pp. 87-89, doi: 10.1145/3332167.3357124.

[22] G. Y. Kim, S. S. Shin, J. Y. Kim, and H. G. Kim, "Sound Event Detection and Haptic Vibration Based Home Monitoring Assistant System for the Deaf and Hard-of-Hearing," in Proc. 2018 Workshop Multimedia Accessible Hum. Comput. Interfaces (MAHCI '18), Oct. 2018, doi: 10.1145/3264856.3264857.

[23] M. Yağanoğlu and C. Köse, "Real-Time Detection of Important Sounds with a Wearable Vibration Based Device for Hearing-Impaired People," Electronics, vol. 7, no. 4, p. 50, Apr. 2018, doi: 10.3390/electronics7040050.

[24] M. D. Fletcher, S. R. Mills, and T. Goehring, "Vibro-Tactile Enhancement of Speech Intelligibility in Multi-talker Noise for Simulated Cochlear Implant Listening," Trends Hear., vol. 22, pp. 1-11, Sep. 2018, doi: 10.1177/2331216518797838.

[25] M. Yağanoğlu and C. Köse, "Wearable Vibration Based Computer Interaction and Communication System for Deaf," Appl. Sci., vol. 7, no. 12, p. 1296, 2017, doi: 10.3390/app7121296.

[26] J. G. Desloge, C. M. Reed, L. D. Braida, Z. D. Perez, L. A. Delhorne, and T. J. Villabona, "Auditory and tactile gap discrimination by observers with normal and impaired hearing," J. Acoust. Soc. Am., vol. 135, no. 2, pp. 838-850, Feb. 2014, doi: 10.1121/1.4861246.

[27] F. A. Russo, P. Ammirante, and D. I. Fels, "Vibrotactile discrimination of musical timbre," J. Exp. Psychol. Hum. Percept. Perform., vol. 38, no. 4, pp. 822-826, Aug. 2012, doi: 10.1037/a0029046.

[28] A. Israr, P. H. Meckl, C. M. Reed, and H. Z. Tan, "Controller design and consonantal contrast coding using a multi-finger tactual display," J. Acoust. Soc. Am., vol. 125, no. 6, pp. 3925-3935, Jun. 2009, doi: 10.1121/1.3124771.

[29] S. C. Aker, H. Innes-Brown, K. F. Faulkner, M. Vatti, and J. Marozeau, "Effect of audio-tactile congruence on vibrotactile music enhancement," J. Acoust. Soc. Am., vol. 152, no. 6, pp. 3396-3409, Dec. 2022, doi: 10.1121/10.0016444.

[30] C.-L. Chin, C.-C. Lin, J.-W. Wang, W.-C. Chin, Y.-H. Chen, S.-W. Chang, P.-C. Huang, X. Zhu, Y.-L. Hsu, and S.-H. Liu, "A Wearable Assistant Device for the Hearing Impaired to Recognize Emergency Vehicle Sirens with Edge Computing," Sensors, vol. 23, no. 17, p. 7454, Aug. 2023, doi: 10.3390/s23177454.

[31] G. Kidd Jr., T. R. Jennings, and A. J. Byrne, "Enhancing the perceptual segregation and localization of sound sources with a triple beamformer," J. Acoust. Soc. Am., vol. 148, no. 6, pp. 3598-3611, Dec. 2020, doi: 10.1121/10.0002779.

[32] D. P. Spiegel, T. Linford, B. Thompson, M. A. Petoe, K. Kobayashi, C. M. Stinear, and G. D. Searchfield, "Multisensory attention training for treatment of tinnitus," Sci. Rep., vol. 5, p. 10802, May 2015, doi: 10.1038/srep10802.

[33] G. L. Plant, "Tactile perception by the profoundly deaf. Speech and environmental sounds," Br. J. Audiol., vol. 16, no. 4, pp. 233-244, Nov. 1982, doi: 10.3109/03005368209081468.

[34] R. Shimoyama, "Omnidirectional Haptic Guidance for the Hearing Impaired to Track Sound Sources," Signals, vol. 2, no. 3, pp. 490-507, Aug. 2021, doi: 10.3390/signals2030030.

[35] R. Paisa, D. Cavdir, F. Ganis, P. Williams, L. M. Percy-Smith, and S. Serafin, "Design and Evaluation of a Multisensory Concert for Cochlear Implant Users," Arts, vol. 12, no. 4, p. 149, Jul. 2023, doi: 10.3390/arts12040149.

[36] M. Mirzaei, P. Kán, and H. Kaufmann, "Effects of Using Vibrotactile Feedback on Sound Localization by Deaf and Hard-of-Hearing People in Virtual Environments," Electronics, vol. 10, no. 22, p. 2794, Nov. 2021, doi: 10.3390/electronics10222794.

[37] M. D. Fletcher, "Can Haptic Stimulation Enhance Music Perception in Hearing-Impaired Listeners?," Front. Neurosci., vol. 15, p. 723877, Aug. 2021, doi: 10.3389/fnins.2021.723877.

[38] P. Guilleminot and T. Reichenbach, "Enhancement of speech-in-noise comprehension through vibrotactile stimulation at the syllabic rate," Proc. Natl. Acad. Sci. U.S.A., vol. 119, no. 13, e2117000119, Mar. 2022, doi: 10.1073/pnas.2117000119.

[39] T. Honda, T. Baba, and M. Okamoto, "Ontenna: Design and Social Implementation of Auditory Information Transmission Devices Using Tactile and Visual Senses," in Computers Helping People with Special Needs: ICCHP-AAATE 2022, Lecture Notes in Computer Science, vol. 13342, Lecco, Italy, Jul. 2022, pp. 130-138, doi: 10.1007/978-3-031-08645-8_16.

[40] G. Sofack, K. Raouafi, S. P. Heinrich, A. Aschendorff, T. Wesarg, S. Arndt, P. Sandmann, and I. Speck, "Somatosensory cross-modal activation and changes in cortical somatosensory evoked potential responses in single-sided deafness: an EEG study," Front. Neurosci., vol. 19, Sep. 2025, doi: 10.3389/fnins.2025.1618134.

[41] L. Dupin, V. Hayward, and M. Wexler, "Direct coupling of haptic signals between hands," Proc. Natl. Acad. Sci. U.S.A., vol. 112, no. 2, pp. 619-624, Jan. 2015, doi: 10.1073/pnas.1419539112.

[42] World Health Organization. "Deafness and hearing loss." [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed: Jul. 10, 2026.

Experimental Approach to Optimizing Magnetic Resonance Imaging Protocols with Reduced Specific Absorption Rate

1st Denyel Jefferson Prado de Faria
Graduate Program in Electrical and Computing Engineering
Federal University of Goiás
Goiânia, Goiás 74605-010, Brazil
denyelfaria@gmail.com

2nd Sílvio Leão Vieira
Graduate Program in Electrical and Computing Engineering/Institute of Physics
Federal University of Goiás
Goiânia, Goiás 74690-900, Brazil
slvieira@ufg.br

***Abstract*—The Specific Absorption Rate (SAR) is the primary parameter used to estimate the deposition of radiofrequency energy during magnetic resonance imaging (MRI). It is critical for patient safety, especially in individuals with impaired thermoregulation. This work proposes and experimentally validates a phantom-based approach to optimizing MRI protocols to reduce SAR while maintaining the predefined quantitative image-quality criteria. Initially, a retrospective analysis of clinical SAR distributions was conducted using DICOM header data from brain and lumbar spine examinations to identify sequences with the highest estimated RF energy deposition. The clinical protocols were then optimized by adjusting acquisition parameters such as flip angle, echo train length, Low SAR mode, and, when necessary, sequence replacement. Image quality was quantitatively assessed using the signal-to-noise ratio, geometric distortion, spatial resolution, and percentage of image uniformity. The findings revealed a quadratic relationship between SAR and the flip angle (R^2 = 0.994) and an approximately linear relationship between SAR and the length of the echo train (R2 = 0.999). Six brain sequences that initially exceeded the operational SAR limit were successfully optimized to values below 1.06 W/kg while meeting the predefined quantitative image-quality acceptance criteria. This experimental approach provides a systematic framework for implementing low-SAR MRI protocols. It may support the development of MRI protocols with reduced estimated RF energy deposition for patients who require stricter control of SAR.**

***Index Terms*—Image Quality, Magnetic Resonance Imaging, Phantom, Protocol Optimization, Specific Absorption Rate.**

I. Introduction

Magnetic resonance imaging (MRI) is one of the leading imaging modalities in contemporary clinical practice, widely used because of its high contrast resolution and the absence of ionizing radiation. Despite this favorable safety profile, the interaction between electromagnetic fields used for image acquisition and biological tissues can result in the deposition of energy as heat, mainly from radiofrequency (RF) pulses [1]–[3].

In this context, the Specific Absorption Rate (SAR) is the primary parameter used to indirectly estimate the energy absorbed by tissues during an MRI examination [4]. SAR is used by regulatory bodies such as the Food and Drug Administration (FDA) and the International Electrotechnical Commission (IEC) to define operational safety limits, generally set at 2.0 W/kg for the whole body and 3.2 W/kg for the head in normal operating mode [5], [6]. Although these limits are considered safe for the general population, patients with impaired thermoregulation—such as elderly people, sedated patients, and those with cardiovascular or metabolic disease—may be more susceptible to RF-induced heating, even within regulatory limits [2], [4], [7]–[9].

Since SAR is not measured directly during examination but is estimated from models that incorporate transmitted RF power, coil characteristics, and acquisition parameters, changes to these parameters simultaneously affect both energy deposition and image quality [1], [10]. Among the main factors involved are the angle of flip (FA), repetition time (TR), echo train length (ETL), the type and duration of RF pulses, as well as the electrical properties and volume of the irradiated tissue [2], [3], [7], [11]–[14]. Consequently, SAR-reduction strategies are mainly based on optimizing these parameters by decreasing FA, increasing TR, reducing ETL, using sequences less dependent on refocusing pulses, and employing automated vendor-provided tools [2], [7], [12], [13], [15]. However, such modifications may compromise fundamental image quality parameters, including signal-to-noise ratio (SNR), contrast-to-noise ratio (CNR), spatial resolution, and tissue contrast, thereby making MRI protocol optimization a trade-off between safety and diagnostic performance.

In this context, the use of phantoms is a well-established experimental approach that standardizes acquisition conditions, eliminates the biological variability inherent in *in vivo* studies,

and enables reproducible quantitative assessments of image quality. Phantoms are widely used in quality control programs and in the development of magnetic resonance protocols, allowing an objective evaluation of quality metrics and other characteristics of the physical image [15], [17]–[20].

Although there are studies that address SAR-reduction strategies, few present an experimental phantom-based approach to protocol optimization that combines the characterization of SAR distribution in clinical protocols with a quantitative evaluation of the impact on image quality under controlled experimental conditions [18], [21].

Given this gap, the present work aimed to propose an experimental approach for optimizing MRI protocols with reduced SAR and to validate it experimentally, using a phantom and quantitative image quality metrics. Specifically, the work sought to characterize the distribution of SAR across clinical protocols, identify the sequences that contribute most to energy deposition, evaluate the influence of key acquisition parameters on SAR, and verify whether the proposed optimization can reduce the estimated RF energy deposition reported by the MRI system while preserving objective indicators of image quality.

II. Materials and Methods

A. Materials

The work was conducted in the diagnostic imaging service of a reference hospital, using a 1.5 T MRI system (Canon Medical Systems, Vantage Elan MRT-2020) with gradients up to 33 mT/m and a slew rate of 125 mT/m/ms [22].

The system is equipped with a 16-element quadrature (QD) whole-body coil, with a maximum root-mean-square (RMS) power of 12 kW and a bandwidth of 63.7 MHz ± 275 kHz [22]. A photograph of the equipment is shown in Fig. 1.



Fig. 1. Canon Vantage Elan 1.5 T magnetic resonance imaging system installed at Hospital de Base do Distrito Federal.

The surface coil used were a head coil (Octave Speeder Head, model MJAH-167A). The image of the coil is shown in Fig. 2.

Images were acquired using routine clinical protocols on the manufacturer's MRI homogeneity test phantom and the



Fig. 2. RF coil used head coil.

American College of Radiology (ACR) phantom. The ACR phantom (180 mm, Konex®) is filled with an aqueous solution of sodium chloride and nickel chloride, with electrical properties equivalent to those of biological tissues [20]. The phantom was used as the standard object in all experimental acquisitions. Two images of the ACR phantom are shown in Fig. 3, with the axial view on the left and the profile view on the right.



Fig. 3. 180-mm Konex® ACR MRI phantom used in the work.

The images were processed using ImageJ software (National Institutes of Health, USA).

This work was approved by the Research Ethics Committee (CEP) of the Hospital das Clínicas of the Federal University of Goiás, under approval report number 7.654.997 and Certificate of Presentation for Ethical Appraisal (CAAE) number 86149824.2.0000.5083.

B. Methods

1) Experimental design: This work is a quantitative, experimental, paired-design study in which standard clinical MRI protocols were compared with their SAR-protocol-reduced, optimized versions. All acquisitions were performed on the same MRI system and with the same test object, ensuring control of external variables and enabling direct analysis of changes resulting exclusively from adjustments to acquisition parameters.

It is important to note that the present study was designed as a phantom-based technical validation, aimed at isolating the effect of acquisition parameter adjustments on objective, quantitative image-quality metrics under controlled conditions. Blinded radiologist evaluation, while valuable, applies to the assessment of diagnostic image quality in patient examinations.

2) Retrospective SAR distribution: Before the experimental stage, a retrospective analysis of SAR values extracted from the DICOM header (tag 0018,1316) was performed to characterize the empirical SAR distribution for each clinical protocol (brain and lumbar spine).

This step was used to estimate the mean values and dispersion of the SAR under real clinical conditions. The distribution by patient, protocol, and sequence type was important, since the literature does not provide a systematic characterization of these distributions, a fundamental step to identify the sequences that contribute the most to overall SAR within a specific protocol.

To determine the number of patients per protocol, a pilot collection of 12 examinations per group was first conducted to estimate the variability of the SAR values obtained [16]. From this preliminary sample, descriptive measures—including the mean and standard deviation—were calculated and used to estimate the required sample size. A significance level of 5% ($\alpha = 0.05$) and a statistical power of 80% ($1-\beta = 0.80$) were adopted, in accordance with widely used recommendations in the literature [24]. The sample size was estimated using the two-sample independent means test, as described in [27], indicating a minimum of 30 patients per protocol.

3) Aquisition: The SAR value reported by the system is based on estimates from the internal manufacturer related to the power absorbed by the tissues ($P_{absorbed}$) and the body mass of the patient (m), which can be expressed in simplified form as [10]:

$$SAR = \frac{P_{absorbed}}{m} \tag{1}$$

Routine clinical protocols for the brain and lumbar spine were first performed in standard settings. The sequences selected for optimization were identified. Then, the acquisitions were performed with progressively adjusted parameters to reduce the SAR. However, as much as possible, the equivalent contrast characteristics and spatial resolution were maintained. Each sequence was acquired in triplicate, allowing estimation of intra-system variability and greater statistical consistency of the measurements obtained.

4) Protocol optimization and SAR control: The reduction in SAR was achieved by adjusting the acquisition parameters, including FAs, ETLs, excitation, and reconstruction parameters, and, when necessary, by replacing sequences. The equipment's automatic SAR-reduction function ("Low SAR") was used as an initial step, followed by incremental manual optimizations. The general work workflow is shown in Fig. 4.

5) Operational SAR criterion: The SAR acceptability criteria adopted in this were defined as operational limits specific

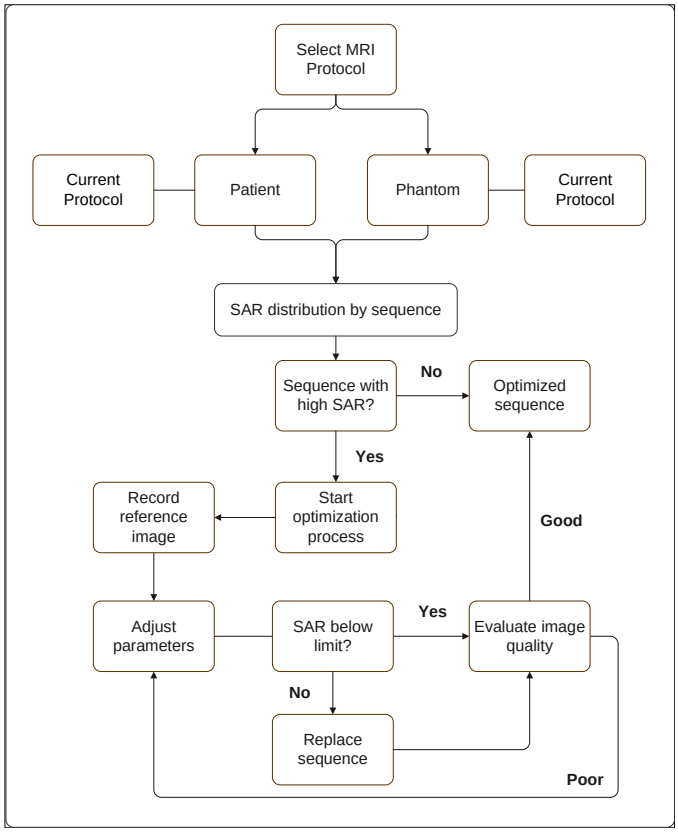


Fig. 4. Workflow for protocol optimization.

to the scenario of impaired thermoregulation. Because the study was motivated by the need to provide an additional thermal safety margin for patients with potentially impaired thermoregulation, a conservative operational SAR criterion was defined a priori.

The physiological literature supports the need for additional caution in patients with impaired thermoregulation, but does not establish a specific percentage reduction in the regulatory SAR limits for this population [7]. Accordingly, rather than being derived from a specific physiological threshold, the 67% reduction adopted in the present study was defined a priori as an operational criterion of the experimental design for the purposes of this study. This criterion was used to establish the operational SAR thresholds adopted for protocol optimization and should not be interpreted as a physiologically established threshold, a regulatory limit, or a universal recommendation. The resulting operational limits are shown in Table I:

TABLE I
OPERATIONAL SAR CRITERION

Protocol	***SAR (W/kg)***
Brain (Head SAR)	$\leq 1,06$
Lumbar Spine (Whole-Body SAR)	$\leq 0,66$

6) Test-object acquisition and experimental control: All acquisitions were performed using the homogeneity MRI

phantom and the ACR phantom, positioned at the equipment isocenter in a standardized manner. The same experimental setup was maintained for both the standard and optimized protocols, ensuring geometric consistency and reproducibility of the measurements.

7) Image quality evaluation: The image quality was quantitatively assessed from DICOM images using ImageJ software. The following parameters were analyzed: SNR, spatial resolution, geometric distortion, and percent image uniformity (PIU). For SNR, equation (2) was used [14], [17], [19], [23], [25]:

$$SNR = 0,655\frac{MPV_{ROI}}{\sigma_{air}} \tag{2}$$

Where MPV_{ROI} is the mean pixel value in a region of interest (ROI) and σ_{air} is the standard deviation within an ROI located in air.

Geometric distortion was obtained using equation (4) [23]:

$$\text{Distortion} = \frac{\text{Distance}_{nominal} - \text{Distance}_{measured}}{\text{Distance}_{measured}} \tag{3}$$

And image uniformity was assessed using equation (5) [20], [23]:

$$PIU = 100\left[1 - \frac{MPV_{\text{Max}} - MPV_{\text{Min}}}{MPV_{\text{Max}} + MPV_{\text{Min}}}\right] \tag{4}$$

Where MPV_{Max} and MPV_{Min} correspond to the mean pixel value in an ROI positioned in the region of highest intensity and in an ROI positioned in the region of lowest intensity, respectively.

ROIs were positioned in a standardized manner, following the ACR phantom guidelines to ensure consistency between acquisitions.

8) Image acceptance criteria: Optimized protocols were considered acceptable when they simultaneously met the image quality criteria defined relative to the reference protocol and the operational SAR criterion, as shown in Table II.

TABLE II
IMAGE QUALITY PARAMETERS AND RESPECTIVE ACCEPTANCE CRITERIA.

Parameter	*Desviation*
SNR	$\geq 75\%$ of reference value
Distortion	$\leq 2\%$
Spatial Resolution	Maintenance of the smallest resolved pattern
PIU	$> 85\%$

9) Statistical analysis: Data were analyzed using descriptive statistics (mean and standard deviation). The design was paired to compare the reference and optimized protocols.

The evaluation of results was based on the percentage variation of the image quality parameters and the compliance with the predefined acceptance criteria.

III. RESULT

A. SAR distribution in clinical protocols

The 30 brain and 30 lumbar spine examinations available for each clinical protocol were analyzed to characterize the distribution of SAR values associated with routinely used clinical protocols. Fig. 5 shows this distribution.

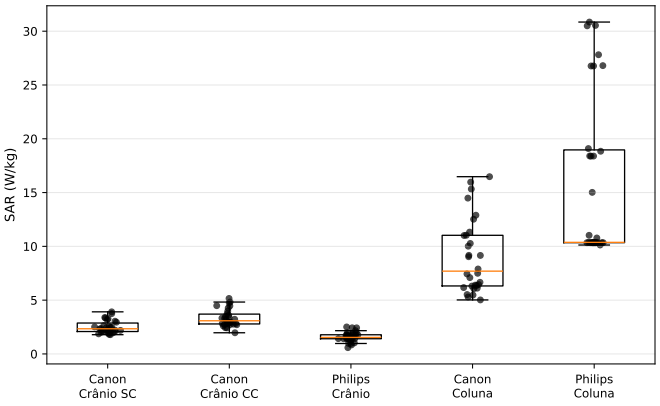


Fig. 5. Statistical distribution of SAR values for brain protocols with and without contrast on the Canon system, the brain protocol on the Philips system, and the lumbar spine protocol on both systems (SC = without contrast; CC = with contrast).

The mean (μ), median, standard deviation (σ), interquartile range (IQR), and minimum and maximum values are shown in Table III. In particular, the SAR recorded in the DICOM header (tag 0018,1316), shown in Fig. 5 and Table III, corresponds to the whole-body SAR (WB SAR).

For the retrospective clinical analysis, SAR values were obtained from DICOM tag (0018,1316), corresponding to whole-body SAR. In contrast, during the experimental optimization stage, the SAR value displayed by the MRI system was recorded according to the anatomical region and acquisition being evaluated.

TABLE III
DISTRIBUTION OF SAR VALUES FOR THE BRAIN AND LUMBAR SPINE PROTOCOLS

Protocol[a]	$\mu \pm \sigma$ **(W/kg)**	**Median (IQR)**	**Min - Max**
Canon Brain SC	2,50 $\pm 0,59$	2,34 (2,08-2,87)	1,79-3,91
Canon Brain CC	3,33 $\pm 0,80$	3,08 (2,79-3,70)	1,97-5,16
Philips Brain	1,59 $\pm 0,46$	1,54 (1,41-1,78)	0,58-2,51
Canon LS[b]	9,01 $\pm 3,42$	7,70 (6,32-11,03)	5,02-16,48
Philips LS[b]	15,98 $\pm 7,60$	10,37 (10,34-18,97)	10,12-30,85

[a]Sample size of 30 patients per protocol
[b]Lumbar Spine.

The largest contributions were observed in the spin echo (SE) and fast spin echo (FSE) based sequences, while the gradient echo (GRE) and fast GRE (FGRE) sequences showed the lowest values, as shown in Figs. 6 and 7.

When grouped by sequence family, the FSE and SE acquisitions showed the highest mean SAR values, as shown in Fig. 8.

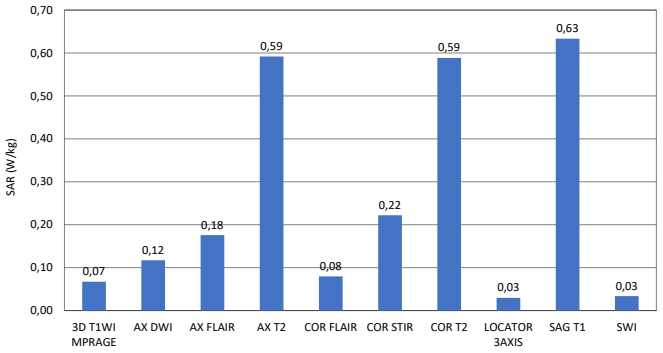


Fig. 6. SAR distribution for the sequences used in the brain protocol.

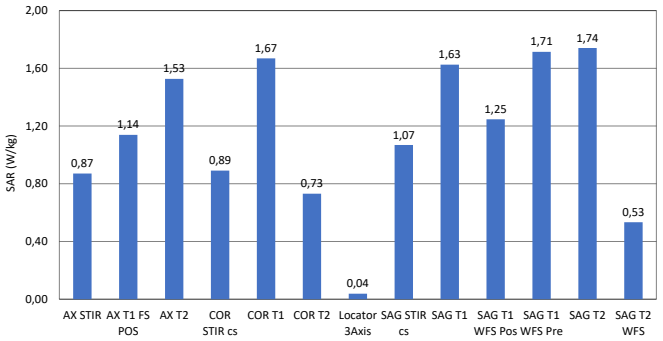


Fig. 7. SAR distribution for the sequences used in the lumbar spine protocol.

B. SAR Reduction strategy

The influence of the main acquisition parameters on SAR was experimentally evaluated using an FSE sequence representative of the clinical brain protocol. SAR exhibited an approximately quadratic dependence on FA ($R^2 = 0.994$) and an approximately linear dependence on ETL ($R^2 = 0.999$).

Activation of the Low SAR mode resulted in reductions ranging from 19% to 63%, depending on the evaluated sequence, as shown in Tables IV.

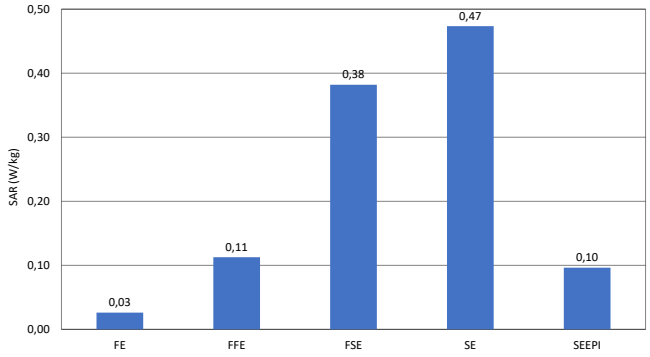


Fig. 8. SAR values grouped by MRI sequence type, for the brain protocol.

TABLE IV
SAR REDUCTION ASSOCIATED WITH THE USE OF THE LOW SAR MODE FOR SPECIFIC SEQUENCES WITHIN THE BRAIN PROTOCOL

Sequence	Original SAR (W/kg)	*Low SAR* (W/kg)	Mean Reduction (%)
FLAIR[a]	0.572	0.462	19.23
3D T1[a]	0.260	0.175	32.69
STIR[a]	1.019	0.771	24.34
T2[a]	2.245	1.639	27.00
T1[a]	1.481	0.548	63.00

[a]Obtained in the axial plane.
[b]WI MPRAGE.

C. Protocol Optimization

Six of the fourteen sequences evaluated in the brain protocol exceeded the adopted operational limit of 1.06 W/kg, as shown in Fig. 9.

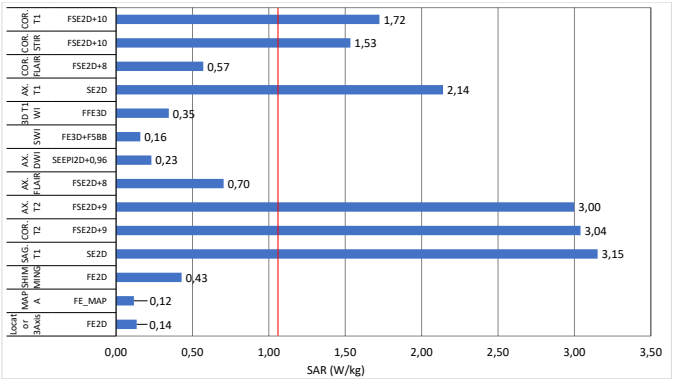


Fig. 9. Estimated SAR values for each sequence of the brain protocol used in clinical routine. The vertical line indicates the SAR = 1.06 W/kg threshold.

After optimizing acquisition parameters, all evaluated sequences achieved SAR values below the operational limit, as shown in Fig. 10.

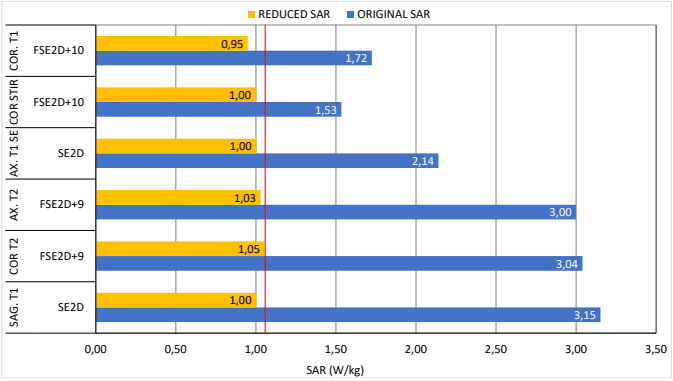


Fig. 10. Estimated SAR values per sequence before optimization (blue) and after optimization (yellow) for the brain protocol.

Table V presents eight sequence configurations. Although only six sequences were optimized, the SAG. T1 and COR. T1 sequences are presented with two optimization alternatives (Adjustments 1 and 2), for a total of eight configurations. For

TABLE V
PARAMETERS OF THE OPTIMIZED SEQUENCES AND RESPECTIVE SAR AND ACQUISITION TIME VALUES. FOR THE SAG, T1, AND COR. THE T1 SEQUENCES, TWO OPTIMIZATION ALTERNATIVES ARE SHOWN (ADJUSTMENTS 1 AND 2).

Name	**SAG. T1**[a]	**SAG. T1**[b]	**COR. T2**	**AX. T2**	**AX. T1**	**COR. STIR**	**COR. T1**[a]	**COR. T1**[b]
SEQ.	SE	FSE	FSE	FSE	SE	FSE	FSE	FSE
FA(°)	100	160	120	120	120	160	120	160
BW(Hz)[c]	122	195	244	244	140	178	178	178
ETL	1	2	19	23	1	13	3	3
SAR(W/kg)	1.000	0.940	1.053	1.029	1.001	0.997	0.949	0.548
$B_1 + RMS(\mu T)$	1.93	1.87	1.99	1.96	1.93	1.93	1.88	1.88
Low SAR	Não	Não	Sim	Sim	Não	Sim	Não	Sim
Time (min:sec)	01:22	02:06	00:42	01:14	01:22	01:23	01:34	03:08

[a]Adjustment 1.
[b]Adjustment 2.
[c]Bandwidth.

each, the original and modified parameters are presented, along with the respective SAR values and acquisition times. In this work, the optimized sequence, "Adjustment 1," was adopted to construct Table VI.

D. Impact on Image Quality

The SNR evaluation results are shown in Fig. 11.

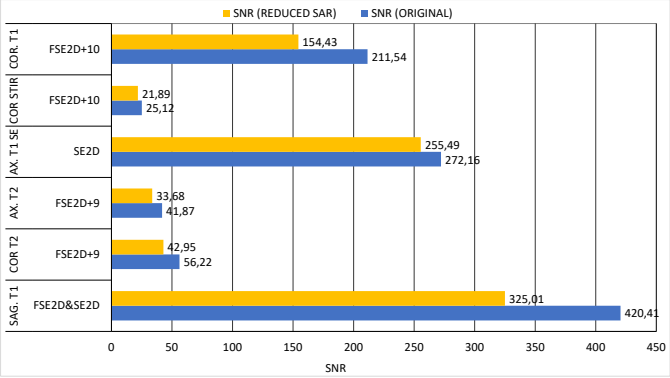


Fig. 11. Signal-to-Noise Ratio (SNR) for the brain protocol sequences before (blue) and after (yellow) SAR reduction.

Since SAR is related to the estimated RF energy deposition and SNR represents the quality of the signal obtained, the SNR/SAR ratio was used as a performance index to estimate the amount of useful signal obtained per unit of estimated RF energy deposition. Although not a standardized metric, this index provides a simple measure of the RF energy efficiency of the evaluated protocols, determined as:

$$\text{Efficiency} = \frac{SNR}{SAR} \tag{5}$$

Table VI presents the efficiency values calculated for the sequences analyzed before and after optimization of the acquisition parameters.

TABLE VI
EFFICIENCY CALCULATED FOR THE SEQUENCES ANALYZED BEFORE AND AFTER OPTIMIZATION

SEQ.	**Original Efficiency**	**Reduced Efficiency**	**Gain**
COR. T1	122.7	162.7	32
COR. STIR	16.4	22.0	34
AX. T1 SE	127.2	255.2	101
AX. T2	14.0	32.7	134
COR. T2	18.5	35.7	93
SAG. T1	133.4	325.0	144

IV. DISCUSSION

Characterization of the SAR distribution in clinical protocols showed that RF energy deposition is not uniform across MRI sequences, a finding also reported in [2], [15], with the brain protocol exhibiting lower SAR values than the other protocol.

The highest contributions to total SAR were observed in SE- and FSE-based sequences for the brain protocol and in FSE for the lumbar spine protocol, which is expected due to the repeated use of refocusing pulses, in agreement with the literature [13]. Although this mechanism is well known from a physical perspective, few studies provide a quantitative assessment of the individual contribution of each sequence within complete clinical protocols.

It was also observed that only a subset of the sequences included in the clinical protocols represented most of the energy deposition. This result demonstrates that optimization can be directed toward the critical sequences while preserving sequences with already low SAR contributions. This approach simplifies the implementation of low-SAR protocols in the clinical routine and minimizes unnecessary changes to the overall quality of the examination.

The observed quadratic behavior confirms the expected physical model, since the transmitted RF power is proportional to the square of the transmit RF field amplitude (B1) and FA is directly related to that amplitude. This result agrees with [13], which identified FA reduction and TR increase as

among the most effective strategies to reduce SAR during clinical examinations, and complements these findings by quantitatively demonstrating this relationship under controlled experimental conditions.

The ETL also had a direct influence on the RF energy deposition, showing approximately linear behavior. This result can be explained by the increase in the number of refocusing pulses applied within each TR, which increases the RF duty cycle and, consequently, the total energy transmitted to the patient. Although effective in reducing SAR, decreasing ETL should be used with caution, as it can increase total acquisition time and alter image weighting characteristics, as demonstrated in [28].

The results showed that the automatic low-SAR mode is an efficient initial strategy to reduce RF energy deposition. Although the internal implementation of the Low SAR mode is proprietary, experimental observations indicate that its activation is associated with changes in RF transmission parameters, such as B1+rms, BW, and TR, resulting in substantial reductions in estimated SAR. However, the observed adjustments varied across sequences. It suggests that the manufacturer's optimization strategy depends on the specific characteristics of each sequence, indicating that its effectiveness varies across sequences. In some sequences, the reduction achieved was sufficient to meet the established operational limit, while in others additional adjustments to acquisition parameters were required.

The results showed that moderate FA reductions of around 120° substantially reduced estimated RF energy deposition while maintaining image quality parameters within the predefined acceptance criteria. Only more intense FA reductions produced significant reductions in SNR, indicating that there is an operational range within which SAR can be optimized without significantly compromising the evaluated quantitative image quality parameters.

Image uniformity remained within the acceptance criteria, while geometric distortion remained within the limits recommended by the MRI quality control guidelines. Similarly, no relevant changes were observed in the spatial resolution of the evaluated sequences. These indicate that the strategies used primarily affected the efficiency of RF excitation while preserving other physical characteristics that determine image quality.

One of the most relevant aspects of this work was the characterization of the distribution of SAR in clinical protocols for the brain and lumbar spine. In contrast to ionizing radiation-based modalities, widely accepted reference values for SAR distribution across MRI protocols are not yet established. Thus, the results presented here can serve as operational indicators for local optimization processes, objectively identifying the sequences that contribute most to RF energy deposition and enabling more efficient targeting of optimization efforts. Furthermore, this characterization establishes a quantitative reference for future comparisons among equipment, protocols, and manufacturers.

A. Energy efficiency of the protocols

In all six evaluated sequences, optimization resulted in increased RF energy SNR-to-SAR efficiency index. This behavior indicates that the reduction in estimated RF energy deposition was disproportionate to the loss in signal intensity, resulting in a higher SNR per unit of deposited RF energy. From a physical perspective, this result demonstrates that the proposed strategy not only reduces SAR but also improves the SNR-to-SAR efficiency index of RF power utilization. An aspect particularly relevant to protocols designed for patients with greater susceptibility to thermal effects.

B. Clinical application

Beyond the implications of the experimental approach presented here, the results have potential applications in clinical routine. As discussed, the proposed approach may support the development of low-SAR protocols for MRI services caring for patients at increased risk of thermal effects. It includes patients who may have increased susceptibility to RF-induced heating, including individuals with impaired thermoregulation or other clinical conditions requiring additional thermal safety considerations. However, clinical diagnostic equivalence was not assessed in the present study.

C. Limitations

This work has some limitations that should be considered when interpreting the results. First, the evaluations were performed on a single 1.5 T MRI system, so generalization of the results to equipment from other manufacturers or with different field strengths should be made with caution. Although the retrospective SAR distribution included equipment from different manufacturers, the experimental optimization stage was conducted exclusively on a Canon system. Future studies should therefore replicate the proposed workflow on MRI systems from different manufacturers and at higher field strengths, particularly 3 T, where RF energy deposition and spatial variations in SAR may represent greater challenges.

Once the technical feasibility of SAR reduction is established under controlled phantom conditions, as demonstrated here, a natural next step would be a clinical validation study involving blinded radiologist scoring of diagnostic image quality and inter-rater agreement in patient examinations.

Finally, the SAR values used in the optimization were estimated by the MRI system rather than directly measured. No direct measurements of phantom temperature or RF energy deposition were performed. Therefore, the proposed approach should be interpreted as an optimization strategy based on the SAR estimates provided by the MRI system, rather than as a direct assessment of tissue heating or thermal dose.

V. Conclusion

This work proposed and experimentally validated an approach for optimizing magnetic resonance imaging protocols using a phantom and quantitative image quality assessment. This experimental strategy was based on characterizing the

SAR distribution in clinical protocols and identifying the sequences that contribute the most to RF energy deposition. The influence of the main acquisition parameters was evaluated, and the performance of the optimized protocols was assessed.

The heterogeneity of the SAR distribution enables objective prioritization of sequences for optimization. The reduction in FA was the most effective strategy. It was complemented by using the Low SAR mode and ETL adjustment. When necessary, sequence replacement complemented optimization, bringing all sequences initially above the operational limit into compliance with the established criteria.

The reduction in estimated RF energy deposition was accompanied by limited quantitative changes in image quality. Although some sequences showed a decrease in SNR, the uniformity, spatial resolution, and geometric distortion values remained within the acceptability criteria. Additionally, the optimized protocols showed greater RF energy efficiency.

The main contribution of this work was to demonstrate that characterizing the SAR distribution in clinical protocols enables the objective identification and optimization of sequences with the highest estimated RF energy deposition, thereby reducing SAR without compromising quantitative image quality criteria. The proposed experimental approach provides a structured and reproducible framework that may be adapted for implementation and further validation across different MRI services and systems.

ACKNOWLEDGMENT

The authors thank Hospital de Base do Distrito Federal and Hospital São Francisco de Assis for providing access to their magnetic resonance imaging systems, the Graduate Program in Electrical and Computer Engineering at the Federal University of Goiás for its academic support, and Cerrad Consultoria em Física Médica e Radioproteção for providing the ACR phantom used in the experimental evaluations.

REFERENCES

[1] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt and J. M. Boone, The Essential Physics of Medical Imaging, Fourth Edition, Philadelphia: Lippincott Williams & Wilkins, 2021.

[2] A. C. C. Murraças, "Estudo exploratório dos valores de SAR em Ressonância Magnética," M.S. dissertation, Universidade de Aveiro, Aveiro, Portugal, 2017.

[3] M. M. Vidotto, K. T. Chaim, D. Papoti, "Análise das diretrizes nacionais e internacionais para sistemas de diagnóstico de Ressonância Magnética e da conformidade de serviços de saúde com a legislação brasileira," Rev. Bras. Fís. Méd., vol. 17, Art. no. 669, 2023, doi: 10.29384/rbfm.2023.v17.19849001669.

[4] F. G. Shellock, Magnetic Resonance Procedures: Health Effects and Safety, Boca Raton, FL, USA: CRC Press, 2001.

[5] American College of Radiology, ACR Manual on MR Safety, Reston, VA, USA, 2024.

[6] U.S. Food and Drug Administration, Testing and Labeling Medical Devices for Safety in the Magnetic Resonance (MR) Environment: Guidance for Industry and Food and Drug Administration Staff, Center for Devices and Radiological Health, Silver Spring, MD, USA, 2023.

[7] E. R. Adair and L. G. Berglund, "Thermoregulatory consequences of cardiovascular impairment during NMR imaging in warm/humid environments," Magn. Reson. Imaging, vol. 7, no. 1, pp. 25–37, Jan. 1989.

[8] M. Díaz and D. E. Becker, "Thermoregulation: physiological and clinical considerations during sedation and general anesthesia," Anesth. Prog., vol. 57, no. 1, pp. 25–32, 2010.

[9] M. Murbach, E. Neufeld, E. Cabot, et al., "Virtual population-based assessment of the impact of 3 Tesla radiofrequency shimming and thermoregulation on safety and B1 + uniformity," Magn. Reson. Med., vol. 76, no. 3, pp. 986-997, 2016, doi: 10.1002/mrm.25986.

[10] Canon Medical Systems Corp., Image Test—Measurement of the SAR Coefficients, document no. 2C900-843EN*D, pp. 247–260.

[11] S. I. C. Silva, "Caraterização dos valores da taxa de absorção específica em exames de ressonância magnética fetal," M.S. dissertation, Universidade Nova de Lisboa, Lisbon, Portugal, 2023.

[12] F. Machado-Rivas, C. Jaimes, J. E. Kirsch, and M. S. Gee, "image quality optimization and artifact reduction in fetal magnetic resonance imaging," Pediatr. Radiol., vol. 50, no. 13, pp. 1830–1838, 2020, doi: 10.1007/s00247-020-04672-7.

[13] J. Allison and N. Yanasak, "What MRI sequences produce the highest specific absorption rate (SAR), and is there something we should be doing to reduce the SAR during standard examinations?," Am. J. Roentgenol., vol. 205, no. 2, p. W140, Aug. 2015, doi: 10.2214/AJR.14.14173.

[14] T. Lowong, "Investigation of parameters for metallic artifact reduction using 3-D SPACE in MR lumbar spine at 1.5 Tesla: A phantom study," M.S. dissertation, Faculty of Medicine, Chulalongkorn Univ., Bangkok, Thailand, 2018, doi: 10.58837/CHULA.THE.2018.344.

[15] J. A. Martinez, K. Moulin, B. Yoo, Y. Shi, H. J. Kim, P. J. Villablanca, and D. B. Ennis, "Evaluation of a workflow to define low specific absorption rate MRI protocols for patients with active implantable medical devices," J. Magn. Reson. Imaging, vol. 52, no. 1, pp. 91–102, 2020, doi: 10.1002/jmri.27044.

[16] S. A. Julious, "Sample size of 12 per group rule of thumb for a pilot study," Pharmaceutical Statistics, vol. 4, no. 4, pp. 287–291, 2005.

[17] G. Laurikaitytė, "Development of Recommendations for the Quality Control of Magnetic Resonance Imaging in Lithuanian Health Care Institutions," M.S. thesis, Kaunas University of Technology, Kaunas, Lithuania, 2025. [Online]. Available: https://hdl.handle.net/20.500.12512/256650.

[18] K. E. Keenan, M. Ainslie, A. J. Barker, M. A. Boss, K. M. Cecil, C. Charles, et al., "Quantitative magnetic resonance imaging phantoms: A review and the need for a system phantom," Magn. Reson. Med., vol. 79, no. 1, pp. 48–61, 2018, doi: 10.1002/mrm.26982.

[19] A. Muzamil and A. H. Firmansyah, "Optimization image of magnetic resonance imaging (MRI) T2 fast spin echo (FSE) with variation echo train length (ETL) on the rupture tendon achilles case," J. Phys.: Conf. Ser., vol. 853, no. 1, pp. 020209, 2017, doi: 10.1088/1742-6596/853/1/020209.

[20] American College of Radiology, "Large and medium phantom test guidance for the ACR MRI accreditation program," American College of Radiology, 2022. [Online]. Available: https://www.acraccreditation.org/

[21] K. E. Keenan, K. V. Jordanova, S. E. Ogier, D. Tamada, N. Bruhwiler, J. Starekova, et al., "Phantoms for quantitative body MRI: A review and discussion of the phantom value," Magn. Reson. Mater. Phys. Biol. Med., vol. 37, no. 4, pp. 535–549, 2024, doi: 10.1007/s10334-024-01181-8.

[22] Canon Medical Systems Corporation, "Vantage Elan MRT-2020 operation manual," Canon Medical Systems Corporation, 2024.

[23] American Association of Physicists in Medicine, "Acceptance testing and quality assurance procedures for magnetic resonance imaging facilities," AAPM Report No. 100, Dec. 2010.

[24] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[25] American College of Radiology, "Magnetic resonance imaging: Quality control manual," American College of Radiology, Reston, VA, USA, 2015.

[26] C. A. C. Alencar, D. C. Oliveira, A. B. M. Teixeira, L. M. G. Lemos, R. C. S. Quesado, I. M. A. Santos, et al., "Tomografia computadorizada e ressonância magnética no Brasil: estudo epidemiológico sobre distribuição dos equipamentos, frequência de realização dos exames e comparação entre setores público e privado," Radiol. Bras., vol. 57, e20230094, 2024, doi: 10.1590/0100-3984.2023.0094.

[27] S. C. Chow, J. Shao, and H. Wang, Sample Size Calculations in Clinical Research, 2nd ed. Boca Raton, FL, USA: Chapman & Hall/CRC, pp. 28, 2008.

[28] S. Demir, B. Clifford, W.C. Lo, A. Tabari, A. Filho, M. Lang, et al., "Optimization of magnetization transfer contrast for EPI FLAIR brain imaging," Magnetic Resonance in Medicine, vol. 87, no. 5, pp. 2380–2387, May 2022, doi: 10.1002/mrm.29141.

Kinovea 2D Analysis: Linear Gait Measurements

Clarissa Cardoso Santos Couto-Paz
Faculty of Sciences and Technology in Health (FCTS)
Federal University of Brasília (UNB)
Brasília, Brazil
Clarissacardososcp@gmail.com

Ranielly Cristina Nunes de Oliveira
Faculty of Sciences and Technology in Health (FCTS)
Federal University of Brasília (UNB)
Brasília, Brazil
Ranielly.cristtina@gmail.com

Lara Leocácio Gomes
Faculty of Sciences and Technology in Health (FCTS)
Federal University of Brasília (UNB)
Brasília, Brazil
Fisio.laragomes@gmail.com

Gabriel de Oliveira Silva
Faculty of Sciences and Technology in Health (FCTS)
Federal University of Brasília (UNB)
Brasília, Brazil
Fisiogabriel.os@gmail.com

Thaís Paulo de Sousa
Faculty of Sciences and Technology in Health (FCTS)
Federal University of Brasília (UNB)
Brasília, Brazil
fisio.thaispaulo@gmail.com

***Abstract*— Human gait analysis is an important technique for assessing, diagnosing, and monitoring certain conditions and biomechanical patterns. Some diseases and disabilities can influence gait dynamics. Individuals with Trisomy 21 (T21) have reduced speed and quality of postural reactions, which leads to compensatory movement patterns. Therefore, analyzing linear gait measurements allows for the objective identification of kinematic patterns that may be altered, enabling comparison of these patterns after therapy and enhancing the therapeutic approach. This study aims to provide a instructions for using the Kinovea software to evaluating linear gait measurements.**

Keywords—Gait Analysis, Trisomy 21, Kinovea

I. Introduction

Gait analysis is an important technique for assessing, diagnosing, and monitoring certain conditions and biomechanical patterns. Despite individual differences, the gait of typically developing individuals is consistent and rarely changes [1].

However, certain genetic conditions, diseases, and disabilities can influence gait dynamics. Individuals with Trisomy 21 (T21) have reduced speed and quality of postural reactions, which leads to compensatory movement patterns during the phases of gait. Therefore, the analysis of linear gait measurements allows for the objective identification of kinematic patterns that may be altered, enabling comparison of these patterns after the therapy, as well as optimizing the therapeutic [2].

Some authors have reported that individuals with T21 may compensate for their congenital mechanical instability by performing specific motor responses, such as increasing muscle activation or executing co-contraction patterns [3][4][5].

Given our understanding of the musculoskeletal changes that modify the pattern of linear gait measurements, as well as joint stiffness in individuals with T21, the objective and quantified analysis of these changes proposed in the manual can assist in the development of therapeutic interventions with specific, measurable, achievable, relevant, and temporally defined objectives, as described in the SMART method [6].

II. Objectives

Provide instructions for using the Kinovea software to evaluate linear gait measurements.

Establish a standardized method for determining the following parameters: stride length and average gait speed.

III. Kinovea: Analiysis Software

This is a free app that provides tools for analyzing 2D motion using spatiotemporal, angular, and linear parameters, as well as distances and frame-by-frame coordinates, of which can be analyzed with or without markers.marcadores.

IV. Filming

A. Camera Position

The camera used for filming must be positioned horizontally, as shown in Figure 1, in a fixed location at an angle of approximately 90°, positioned perpendicular to the participant at a distance of 2.5 meters and at a height of 1 meter above the ground [7]. The camera cannot be moved while recording. In the example that will be presented, a 12 megapixel smartphone camera with a frame rate of 60 Hz was used.

B. Guidelines for Filming the Walk

It is recommended to place a marker on the subject's lateral malleolus, as shown in Figure 2. Preferably, the subject should be wearing their usual shoes or be barefoot. Markers may include: Post-it notes secured with tape, adhesive labels, or dermatographic pens, as shown in Figure 2.



XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Fig. 1. How to position de camera



Fig. 2. Marker Positions



Fig. 3. Calibration Measurement Model

C. *Video Calibration Templates*

We suggest that the calibration measurement be 100 centimeters horizontally. To include this reference in the video, you can use various tools such as adhesive tape, a measuring tape, or two objects positioned to simulate the standard calibration measurement, as shown in Figure 3.

The calibration measurement should be near or part of the path that the individual will walk, as shown in Figure 3.

V. Linear Measurements of the Walk

A. *Video Selection*

To measure linear gait parameters (stride length and speed) through gait video analysis, the video must be recorded in the sagittal plane, and it must be possible to visualize four initial ground contacts and the marked anatomical point: the lateral malleolus.

B. *Analysis of Walking Variables*

The variables analyzed in the walking video are:

COMP-P1: stride length 1.

COMP-P2: stride length 2.

COM-P3: stride length 3.

Δs: distance traveled by the individual from the starting point to the end point.

Δt: time taken by the individual to travel from the starting point to the end point.

V: average walking speed.

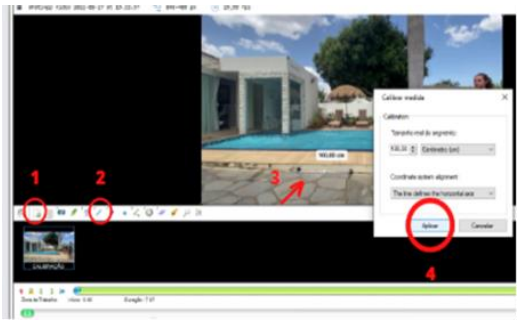


Fig. 4. System Calibration

C. *Calibration of Walking in the Sagittal Plane.*

1- When loading the video, click the button to add a new image frame and select a frame before the first Initial Contact, Figure 4.

2- Select the “line” button. Draw a horizontal line from the beginning to the end of the standard measurement for calibration, which, in Figure 4, is represented by two objects on the ground that represent the actual distance of 100 cm.

3- After drawing the line, right-click on it and select the “Calibrate” button, Figure 4.

4- Set the measurement to 100 centimeters and click “Apply,” Figure 4.

D. *Selection of Initial Contact Frames*

1- Drag the navigation bar and stop at the Initial Contact (IC) phases of the gait, referring to the same lower limb, because we will be working with stride length.

2- Click the icon to add a new image frame and select the frame for the Initial Contact phase, Figure 5.

3- To save the names of the selected frames, simply double-click on the frame to name it, Figure 5.

You must select the frames for the first four Initial Contacts (CI1, CI2, CI3, CI4), Figure 5. This will result in the following: CALIBRATION, CI1, C12, CI3, CI4, Figure 5.

The numbers following the names correspond to the time at which the individual made that initial contact. Therefore, in CI1 the time was 2.07 s, and in CI2 the time was 1.36 s, which is the difference between the times for CI2 and CI1, since the software displays the time continuously. In CI3, the time was 1,27 seconds, that is, the difference between CI3 and CI2, and in CI4, the time was 1,37 seconds, that is, the difference between CI4 and CI3.

The time is automatically saved in the frames, Figure 6. However, it is necessary to calculate these subtractions.

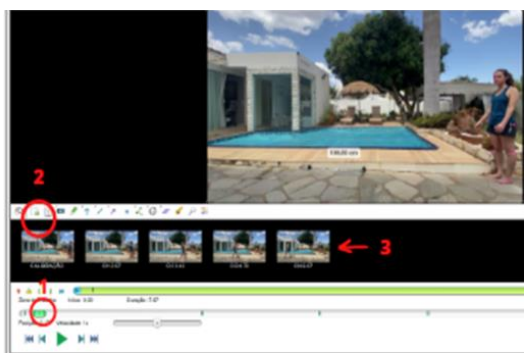


Fig. 5. Frame Selection in the System.



Fig. 6. Frames.



Fig. 7. Calibration Grid Adjustment.

E. Calibration Grid Adjustment

1. After selecting the frames, click on the initial contact frame 1, Figure 7.
2. Click "Tools" and then "Coordinate System"; after these commands, the calibration grid will appear, Figure 7.
3. Select the icon on the navigation bar, position the cursor in the center of the calibration grid, and drag the grid so that the 0 is aligned with the individual's lateral malleolus, Figure 7.
4. Verify that the positioning is correct, Figure 7.

F. Measuring Stride Length 1.

1. To obtain the length of Stride 1, you must be on the CI2 frame, as this frame corresponds to the start of Stride 2 and the end of Stride 1, Figure 8.

2. Select the line button, Figure 8.

3. Draw a horizontal line from the point where the lateral malleolus was located in CI1 (in this case, we adjusted the calibration grid for CI1 to point 0 on the calibration grid) to the point where the lateral malleolus is located in CI2. Then, the line will extend from point 0 on the calibration grid to the individual's lateral malleolus. This will give you the length (COMP) of stride 1 (P1) in centimeters, in Figure 8, this is represented by 140,41 centimeters.

To save the image, right-click and save as: COMP-P1-INDIVIDUALABBREVIATION-MID or MIE (side evaluated).

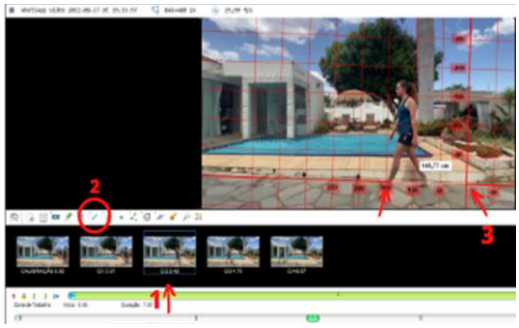


Fig. 8. Measuring Stride Length 1.

G. Measuring Stride Length 2

Para To analyze the length of stride 2, you must first mark the location of the lateral malleolus at the start of stride 2.

To do this, we will use a tool in the application called "marker"

1. Click the "marker" button, Figure 9.
2. Position the marker over the lateral malleolus, Figure
3. Right-click on the marker, click "visibility," and then click "always visible," Figure 9.

This way, when you move to the next frame, the marker will be visible to make it easier to measure the length of the next stride. The markers will be visible in the video as shown in Figure 10.

To Measure Stride Length 2:

1. You must be on the CI3 frame, as this corresponds to the start of stride 3 and the end of stride 2, Figure 11.
2. Select the line icon, Figure 11.
3. Draw a horizontal line from the point where the lateral malleolus was located (as defined by the marker) to the point where the lateral malleolus of CI3 is located, Figure 11.

This will give you the length (COMP) of stride 2 (P2) in centimeters, in Figure 11, this is represented by 134,95 centimeters.

To save the image, right-click and save as: COMP-P2-INDIVIDUALINITIALS-MID or MIE (side evaluated).



Fig. 9. Adding the Marker.



Fig. 10. Markers.

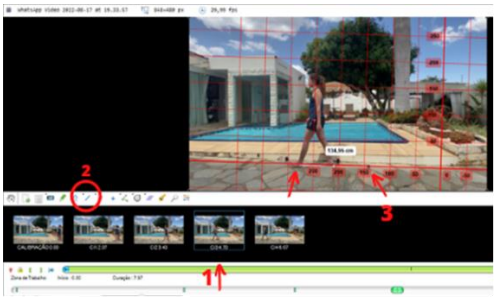


Fig. 11. Measuring Stride Length 2.

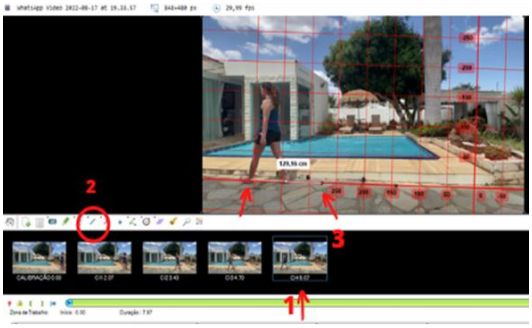


Fig. 12. Measuring Stride Length 3

H. Measuring Stride Length 3

1. To obtain the length of Stride 3, you must be in the CI4 frame, as this frame corresponds to the start of Stride 4 and the end of Stride 3 (Figure 12).

2. Select the line button (Figure 12).

3. Draw a horizontal line from the point where the lateral malleolus was located (as defined by the marker) to the point where the lateral malleolus of CI4 is located, Figure 12.

This will give you the length (COMP) of stride 3 (P3) in centimeters, in Figure 12, this is represented by 129,55 centimeters.

To save the image, right-click and save as: COMP-P3-INDIVIDUALINITIALS-MID or MIE (side evaluated).

I. Walking Speed

To calculate the average walking speed (VM), you need to determine the distance (Δs) traveled by the individual and the time (Δt) taken to move that distance.

J. Calculation of Displacement (Δs)

1. Stay on the CI4 frame, Figure 13.
2. Select the line tool, Figure 13.
3. Draw a line from point "0" on the grid, which marks the start of the displacement, to the individual's lateral malleolus, which marks the end of the displacement, Figure 13.

To save the image, right-click and save as: DISTANCE-INDIVIDUALINITIALS-MID or MIE (side evaluated).

This will give you the (Δs) in centimeters. In our example, the distance traveled was 404.91 centimeters, Figure 14.

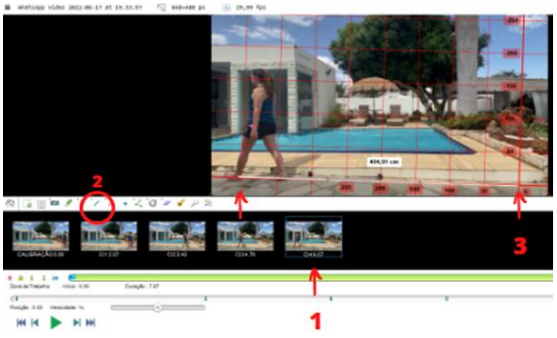


Fig. 13. Guide for determining Displacement (Δs).

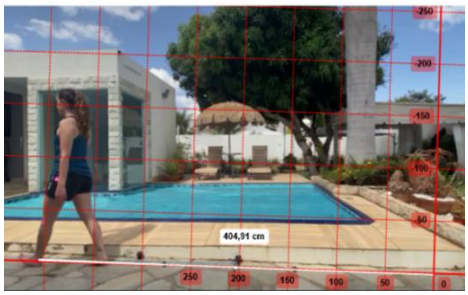


Fig. 14. Obtaining the Displacement (Δs).

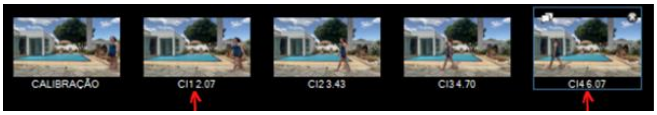


Fig. 15. Obtaining the Time (Δt)

K. Calculating the Time (Δt)

To calculate the time, subtract the time the individual arrived at CI1 from the time they arrived at CI4, as illustrated in Figure 15.

This will give you (Δt) in seconds. In our example, the time was 6,07 s – 2,07 s, or 4,0 s.

L. Calculating Average Speed

To calculate the average speed, we will convert Δs from centimeters to meters, dividing the value in meters by 100, as indicated in equation (1).

$$\Delta s\ (m) = \Delta s\ (cm) / 100 \quad (1)$$

We will divide Δs, measured in meters (m), by Δt, measured in seconds (s), using the following formula:

$$V = \Delta s / \Delta t \quad (2)$$

VI. Conclusion

The use of quantitative protocols for assessing human gait directly impacts the quality of care. The use of Kinovea, a free and accessible tool, for these analyses democratizes this quality. The existence of a detailed protocol helps frontline professionals effectively apply best practices in assessment.

Acknowledgment

Special thanks to all the members of the Study Group on Physical Therapy for Neurodysfunctions (GEFIN); without you, this work would not have been possible. I would also like to thank the University of Brasília (UNB).

References

[1] G. Jarchi D, Pope J, Lee TKM, Tamjidi L, Mirzaei A, Sanei S. A Review on Accelerometry-
Based Gait Analysis and Emerging Clinical Applications.
IEEE Rev Biomed Eng. 2018;11:177-194. doi:
10.1109/RBME.2018.2807182. Epub 2018 Feb 16. PMID: 29994786.

[2] Zago M, Duarte NAC, Grecco LAC, Condoluci C, Oliveira CS, Galli M. Gait and postural control patterns and rehabilitation in
Down syndrome: a systematic review. J Phys Ther Sci. 2020 Apr;32(4):303-314. doi: 10.1589/jpts.32.303. Epub 2020 Apr 2. PMID: 32273655; PMCID: PMC7113426.

[3] Galli M, Rigoldi C, Brunner R, Virji-Babul N, Giorgio A. Joint stiffness and gait pattern evaluation in children with Down syndr

ome. Gait Posture. 2008 Oct;28(3):502-6. doi: 10.1016/j.gaitpost.2008.03.001. Epub 2008 May 2. PMID: 18455922.

[4] Gontijo AP, Mancini MC, Silva PL, Chagas PS, Sampaio RF, Luz RE, Fonseca ST. Changes in lower limb co-contraction and stiffness by toddlers with Down syndrome and toddler s with typical development during the acquisition of independent gait. Hum Mov Sci. 2008 Aug;27(4):610-21. doi: 10.1016/j.humov.2008.01.003. Epub 2008 Jul 22. PMID: 18649965.

[5] Casabona A, Valle MS, Pisasale M, Pantò MR, Cioni M. Functional assessments of the k nee joint biomechanics by using pendulum test in adults with Down s yndrome. J Appl Physiol (1985). 2012 Dec 1;113(11):1747-55. doi: 10.1152/japplphysiol.00960.2012. Epub 2012 Sep 20. PMID: 22995394; PMCID: PMC3544505.

[6] Bjerke MB, Renger R. Being smart about writing SMART objectives. *Eval Program Plann*. 2 017;61:125-127. doi:10.1016/j.evalprogplan.2016.12.009

[7] Fernández-González P, Koutsou A, Cuesta-Gómez A, Carratalá-Tejada M, Miangolarra-Page JC, Molina-Rueda F. Reliability of Kinovea® Software and Agreement with a Three-Dimensional Motion System for Gait Analysis in Healthy Subjects. Sensors (Basel). 2020 Jun 2;20(11):3154. doi: 10.3390/s20113154. PMID: 32498380; PMCID: PMC7308968.

Limitations of RAG Architecture in Compact Medical LLMs: Diabetes Support

Eduardo Borba Neves
Captain's Career School
Brazilian Army
Rio de Janeiro, Brazil
0000-0003-4507-6562

Pablo Gustavo Cogo Pochmann
Captain's Career School
Brazilian Army
Rio de Janeiro, Brazil
0000-0003-3944-7953

Daisy Cristine Albuquerque da Silva
Department of Army Education and Culture
Brazilian Army
Rio de Janeiro, Brazil
0009-0000-4961-9056

***Abstract*—The use of Large Language Models (LLMs) in healthcare opens promising perspectives for clinical decision support; however, privacy and infrastructure constraints drive interest in open-source models executed locally and offline. This experimental study evaluated and compared the impact of Retrieval-Augmented Generation (RAG) architecture on the cognitive performance and computational efficiency of four LLM architectures specialized in the medical domain (Bio-Medical-Llama-3.1-8B, LlaMA-3.1-8B-Medical, Llama-3.2-3B-Instruct, and Qwen2.5-3B-instruct) during the resolution of complex clinical scenarios of Diabetes Mellitus. Inferences were performed in a restricted deterministic computational environment (8 GB VRAM) using the GGUF quantization format (4-bit and 8-bit). Evaluation was based on balanced subsets of 200 questions extracted from the MEDQA and MedMCQA benchmarks. The results revealed a counterintuitive behavior compared to cloud-operated systems: the introduction of RAG, in this setup, caused a degradation of clinical accuracy across almost all scenarios, accompanied by a substantial operational penalty. The LlaMA-3.1-Med-8B model achieved the highest accuracy observed in direct inference (66.5% on MedMCQA; 58.0% on MEDQA), dropping to 58.5% and 49.5% respectively under the RAG condition. From a software engineering perspective, context expansion (up to 4,000 characters) acted as distracting noise for the selective attention mechanisms of compact networks, a pattern consistent with the lost-in-the-middle effect described in the literature, and created a bottleneck in the GPU–CPU bus bandwidth due to partial layer offloading, increasing average latency by up to 34 times and substantially reducing the tokens-per-second (TPS) rate. We conclude that, under the specific configuration evaluated, the standard single-corpus RAG pipeline substantially increased computational cost and frequently reduced clinical accuracy, indicating a trade-off that requires further characterization across alternative retrieval strategies and hardware environments.**

***Index Terms*—Large Language Models, RAG, Diabetes Mellitus, Restricted Hardware, Computer Engineering.**

I. Introduction

Diabetes Mellitus (DM) represents one of the greatest global public health challenges, requiring individualized and multidisciplinary decision-making based on a growing volume of guidelines and therapeutic data [1]–[4]. In parallel, the recent advancement of Large Language Models (LLMs) has opened new horizons for clinical decision support, demonstrating high performance in medical reasoning tasks [5], [6]. However, operational and regulatory constraints regarding the protection of sensitive data, such as LGPD and HIPAA, limit the use of commercial cloud APIs, driving institutional interest in open-source models executed locally and offline.

Despite progress in compact biomedical models, their probabilistic architecture is still prone to factual hallucinations and failures when applying rigid quantitative diagnostic thresholds [7]. To mitigate these flaws, Retrieval-Augmented Generation (RAG) architecture emerges as a strategy to ground answers directly in explicitly validated external documentary sources [8], [9]. The potential of this approach in the medical field and in diabetes education has demonstrated gains in accuracy and reductions in hallucinations [10], [11].

However, gaps remain in the literature regarding the practical impact of RAG on compact medical LLMs executed locally under severe hardware constraints. Injecting additional contexts introduces search and re-ranking stages that directly alter crucial operational metrics such as latency and token generation speed, making it necessary to evaluate the trade-off between cognitive gains and their respective computational costs.

To contribute toward filling this gap, this study aimed to evaluate and compare the impact of RAG architecture on the cognitive performance and computational efficiency of different open-source, offline-executed LLMs specialized in the medical domain during the resolution of structured complex clinical scenarios focused on Diabetes Mellitus and related metabolic disorders.

II. Methodology

This experimental study aimed to evaluate and compare the impact of RAG architecture on the performance of offline-executed LLMs focused on the medical field. Four medical-specialized LLM architectures were evaluated and subjected to subsets of questions extracted from the MEDQA and MedMCQA international benchmarks, which are widely used to assess biomedical knowledge and automated clinical reasoning [12], [13].

A. Execution Environments and Evaluated Models

Inferences were performed locally in a deterministic computational environment (configured with temperature $t = 0.0$ and $top_p = 1.0$) to mitigate the stochastic variability of the responses. The pipeline was structured in Python, using the

llama_cpp library for the orchestration and execution of the quantized models in GGUF format.

To enhance the technical rigor and academic fluidity of the methodology, the selection of the open compact LLM architectures was based on the state-of-the-art models featuring high efficiency and specialized fine-tuning in the biomedical and clinical domains. Four models based on the Llama 3.1, Llama 3.2, and Qwen 2.5 families were selected, operating at scales of 3B and 8B parameters.

To enable strictly deterministic local execution and mitigate the severe video memory (vRAM) constraints typical of consumer hardware environments, all architectures were evaluated using the GGUF (GPT-Generated Unified Format) encapsulation and quantization format. Weight compression was performed using block-based advanced quantization algorithms (k-quants and i-quants), reducing the original 16-bit floating-point representation to hybrid 4-bit and 8-bit representations. This strategy dramatically reduced the memory footprint and the bandwidth required for data transfer without causing substantial degradation in the models' cognitive and semantic capabilities.

The inference pipeline integrated the llama_cpp library with low-level optimizations at the computational kernel level:

- Flash Attention: Implementation of block-based selective attention mechanisms that optimize reading and writing traffic in the GPU cache memory.
- Active KV Cache Offload: Strictly configured with type-2 keys and values in 16-bit floating-point precision, keeping the context history highly faithful and stable.
- Dynamic Layer Allocation (GPU Split): Adaptive management of network topology in the video memory. For smaller-scale models (3B), full layer offloading to vRAM was applied. For larger models (8B), a hybrid allocation restricted to 22 layers on the GPU was implemented, managing the remainder on the host CPU to avoid Out of Memory (OOM) failures.

Detailed technical specifications for each neural network model, including their distribution files, parameter volumes, and clinical specializations, are consolidated in Table I.

The diversity among the chosen quantization methods establishes a robust baseline for evaluating how the RAG architecture behaves against subtle losses induced by weight compression when solving complex clinical cases of Diabetes.

B. Datasets and Thematic Filtering Criteria

The performance evaluation used two established benchmarks in the medical artificial intelligence literature:

- MEDQA: Questions based on the United States Medical Licensing Examination (USMLE).
- MedMCQA: Multiple-choice questions from India's medical entrance examinations (AIIMS and NEET PG).

Since the scope of the paper was delimited to the study of Diabetes, the original datasets underwent a filtering process based on keyword co-occurrence. A question was included in the test subset only if it presented at least two simultaneous terms (co-occurrence ≥ 2) belonging to the following extended English list: *diabetes, glycemia, insulin, glucose, glycated hemoglobin,* and *hba1c*.

After applying the filter, a simple random sampling (with a fixed randomness seed, random_state $= 42$) selected balanced subsets of 200 questions per dataset to compose the final inference test for each model.

C. Inference Pipeline Architecture

The experiment was divided into two distinct experimental conditions: Direct Inference (Without RAG) and Enriched Inference (With RAG).

1) Condition 1: Direct Inference (Zero-Shot Without RAG): In the first stage, prompts were submitted to the models containing only the professional role instruction (*System Prompt*), the clinical question, and the four available options (A, B, C, D). The model was configured with a strict token limit (max_tokens $= 1$) and instructed to return strictly the letter corresponding to the correct alternative.

2) Condition 2: Inference with RAG Support: In the second stage, a RAG pipeline was implemented, consisting of a dense vector search coupled with a re-ranking algorithm:

- Indexing and Vector Database: The knowledge base (comprising International Diabetes Guidelines) was previously chunked and stored in a local vector database using the Qdrant Client.
- Dense Vector Search: For each dataset question, the query was converted into a dense vector using the local embedding model *sentencetransformers/paraphrase-multilingualmpnetbasev2* via the *fastembed* library. The system performed a cosine similarity search in Qdrant, retrieving the 4 most relevant text chunks (chunk_limit $= 4$) while respecting a cap of 4,000 characters.
- Re-ranking: The retrieved fragments were submitted to a local re-ranking model (*ms-marco-MiniLM-L-12-v2* via *flashrank*). The top 2 chunks resulting from the re-ranking were selected and dynamically injected into the final prompt.
- Enriched Prompt: The model received the context from the medical guidelines before the question block, responding under the same token constraints as the previous stage.

D. Knowledge Base for the RAG

The sources composing the knowledge base for the RAG represent some of the most critical, comprehensive, and updated clinical guidelines globally for the diagnosis, management, and treatment of Diabetes Mellitus and its associated comorbidities. This documentary base included the *Standards of Care in Diabetes — 2026* by the American Diabetes Association [2]; the *Clinical Practice Guideline: Developing a Diabetes Mellitus Comprehensive Care Plan — 2022 Update* [14]; the *Consensus Statement: Algorithm for Management of Adults With Type 2 Diabetes — 2026 Update* [15]; the British guideline *NICE Guideline NG28: Type 2 Diabetes in Adults: Management* [16], along with its respective update factsheet [17]; the *Clinical Practice Guideline: The Use of*

TABLE I
TECHNICAL SPECIFICATIONS AND TOPOLOGY OF THE EVALUATED LLM MODELS FOR THE RESOLUTION OF STRUCTURED COMPLEX CLINICAL SCENARIOS FOCUSED ON DIABETES MELLITUS AND RELATED METABOLIC DISORDERS

Model (.gguf file)	Params	Quantization	Architecture / Clinical Specialization
Bio-Medical-Llama-3.1-8B.i1-Q4_K_M	8B	Q4_K_M (i1)	Llama 3.1 base; supervised fine-tuning (SFT) focused on general biomedical knowledge and scientific literature terms.
LlaMA-3.1-8B-Medical.Q4_K_M	8B	Q4_K_M	Llama 3.1 base; fine-tuning geared toward clinical medical reasoning, diagnostics, and practical internal medicine management.
Llama-3.2-3B-Instruct-Medical-Conversational.Q8_0	3B	Q8_0	Llama 3.2 base; ultra-compact model with instructional tuning focused on history-taking and conversational medical interactions.
Qwen2.5-3B-instruct-medical-finetuned.Q4_K_S	3B	Q4_K_S	Qwen 2.5 base; optimized for high-density information medical tasks under severe computational hardware restrictions.

Advanced Technology in the Management of Persons With Diabetes Mellitus [18]; the joint guideline *Guideline for the Diagnosis and Management of Nonalcoholic Fatty Liver Disease (NAFLD)* [19]; and the integrated *Comprehensive Clinical Practice Guidelines for Medical Care of Patients with Obesity* [20]. Together, these documents form the core of modern, multidisciplinary diabetes management, covering prevention, early diagnosis, the intensive use of biotechnologies, and organ protection (heart, kidneys, and liver), managing the disease far beyond blood glucose metrics.

E. Evaluation Metrics

The performance of both approaches was measured and consolidated through three fundamental metrics: Clinical Accuracy (%), Average Latency per Question (s), and Tokens per Second (TPS).

The generated textual responses were normalized using regular expressions (Regex) to extract the predicted letter (A, B, C, or D) for subsequent cross-checking against the official answer key (*correct_answer*). Cases with inconclusive outputs were automatically classified as "INVALID".

III. RESULTS

Experimental results reveal a counterintuitive behavior regarding the use of Retrieval-Augmented Generation (RAG) architectures in compact LLMs executed locally. Contrary to the gains typically observed in large-scale cloud-operated models, the dynamic injection of clinical guidelines was associated with measurable penalties in both cognitive performance (accuracy) and computational efficiency (latency and token throughput) under the configuration evaluated.

A. Cognitive Performance (Accuracy)

Table II consolidates the accuracy data obtained across both benchmarks (MEDQA and MedMCQA) under the two experimental conditions: Direct Inference (Without RAG) and Enriched Inference (With RAG).

The only scenario that showed a marginal gain was the Llama-3.2-3B-Instruct model on the MedMCQA dataset (+1.00%). In contrast, the Qwen2.5-3B model exhibited the sharpest degradation, losing 11.00% accuracy in the same benchmark. A possible interpretation, further examined in the Discussion, is that the long text fragments retrieved via RAG acted as context distraction for the compact 3B and 8B parameter architectures, reducing their ability to extract the correct answer from the expanded prompt.

TABLE II
CLINICAL ACCURACY METRICS ACROSS EXPERIMENTAL CONDITIONS FOR THE RESOLUTION OF STRUCTURED COMPLEX CLINICAL SCENARIOS FOCUSED ON DIABETES MELLITUS AND RELATED METABOLIC DISORDERS

Model	Dataset	w/o RAG (%)	with RAG (%)	Delta (%)
Bio-Medical-8B	MEDQA	55.50	47.00	–8.50
Bio-Medical-8B	MedMCQA	64.00	62.00	–2.00
LlaMA-3.1-Med-8B	MEDQA	58.00	49.50	–8.50
LlaMA-3.1-Med-8B	MedMCQA	66.50	58.50	–8.00
Llama-3.2-3B-Inst	MEDQA	47.50	43.00	–4.50
Llama-3.2-3B-Inst	MedMCQA	56.00	57.00	+1.00
Qwen2.5-Med-3B	MEDQA	39.00	32.50	–6.50
Qwen2.5-Med-3B	MedMCQA	38.00	27.00	–11.00

B. Computational Performance and Efficiency (Latency and TPS)

Infrastructure costs and local processing times were markedly affected by the RAG architecture (Table III). The inclusion of vector search steps in Qdrant and re-ranking via the Ranker (*ms-marco-MiniLM-L-12-v2*) resulted in a substantial increase in response times.

TABLE III
COMPUTATIONAL EFFICIENCY: LATENCY AND TOKEN THROUGHPUT FOR THE RESOLUTION OF STRUCTURED COMPLEX CLINICAL SCENARIOS FOCUSED ON DIABETES MELLITUS AND RELATED METABOLIC DISORDERS

Model / Dataset	Lat. w/o RAG (s)	Lat. with RAG (s)	TPS w/o RAG	TPS with RAG
Bio-Med-8B (MEDQA)	4.65	51.05	0.24	0.05
Bio-Med-8B (MedMCQA)	1.48	49.64	0.86	0.09
LlaMA-8B (MEDQA)	4.75	51.60	0.24	0.05
LlaMA-8B (MedMCQA)	1.53	49.94	0.84	0.09
Llama-3.2-3B (MEDQA)	4.34	39.75	0.26	0.07
Llama-3.2-3B (MedMCQA)	1.21	38.51	1.05	0.12
Qwen2.5-3B (MEDQA)	1.88	24.73	0.61	0.10
Qwen2.5-3B (MedMCQA)	0.57	23.78	2.26	0.17

While direct inference (Without RAG) enabled agile responses (ranging from 0.57s to 4.75s), the enriched pipeline elevated the average latency to thresholds between 23.78s and 51.60s per question. Concurrently, the Tokens per Second (TPS) rate declined consistently across all scenarios, underscoring the impact of processing dense contexts (up to 4,000 characters) on hardware with shared resources. This reinforces the engineering challenge of enabling real-time clinical decision support systems under strictly offline execution.

IV. Discussion

The analysis of the experimental results reveals two highly relevant patterns for medical informatics: (a) a consistent performance hierarchy among the evaluated architectures in the direct inference scenario (Without RAG); and (b) a consistent reduction of clinical accuracy in seven of the eight model–benchmark pairs evaluated following the introduction of the Retrieval-Augmented Generation (RAG) architecture, accompanied by a substantial increment in latency. These phenomena express the complex trade-off involved in the local execution of large language models (LLMs) on consumer hardware with severe video memory limitations (8 GB of VRAM).

A. Performance in Direct Inference (Without RAG)

In the scenario without RAG, the LlaMA-3.1-8B-Medical and Bio-Medical-Llama-3.1-8B models (8 billion parameters, quantized in Q4_K_M) outperformed the smaller architectures, achieving accuracies of 58.0% and 55.5% on MEDQA, and 66.5% and 64.0% on MedMCQA, respectively. These values align with the literature: Singhal et al. [6] documented that the native Llama-3.1-Instruct-8B reaches 58.7% on MedQA and 56.0% on MedMCQA, and demonstrated that parametric scale and medical pre-training density are the primary predictors of accuracy. For instance, Med-PaLM 2 reaches up to 86.5% on MedQA, outperforming its predecessor by more than 19 percentage points. This benchmark, subsequently expanded for resolving complex clinical scenarios with structured justifications at the medical expert level [21], demonstrates that current maximum cognitive proficiency in medicine still depends on robust proprietary cloud architectures.

The Qwen2.5-3B-instruct-medical-finetuned.Q4_K_S model delivered the poorest performance: 39.0% on MEDQA and 38.0% on MedMCQA, values close to a random guess rate (25% for 4 choices). Two mechanisms explain this deficiency. First, the representational capacity limitations inherent to 3-billion-parameter models are insufficient to capture the causal relationships required by complex clinical reasoning [22]. Second, there is evidence that specialized fine-tuning may have induced catastrophic forgetting; studies indicate that refining models on restricted biomedical corpora tends to overfit the network to the domain, sacrificing general logical capabilities and self-correction mechanics required by benchmarks like MEDQA [22].

All models scored 8 to 12 percentage points higher on MedMCQA compared to MEDQA. This asymmetry reflects structural differences between the benchmarks: MEDQA is based on United States Medical Licensing Examination (USMLE) questions, featuring long clinical vignettes and advanced multidisciplinary reasoning [12]; MedMCQA, derived from the Indian AIIMS/NEET PG medical entrance exams, spans over 194,000 questions across 21 disciplines, but features shorter prompts and less logical articulation complexity [13]. The higher rate of tokens per second (TPS) on MedMCQA, alongside latencies of 0.57 to 1.53 s versus 1.88 to 4.75 s on MEDQA, confirms that more compact inference contexts accelerate processing.

B. Performance Degradation with the RAG Architecture

The most prominent finding of this experiment is the reduction in accuracy observed in almost all scenarios following the introduction of RAG: regressions of 6 to 11 percentage points for the 8B models and up to 11 percentage points for the Qwen2.5-3B, with latency rising from under 5 s to 24–51 s per question (a 10- to 34-fold increase). This behavior, counterintuitive relative to large-scale cloud systems, is consistent with mechanisms described in recent literature.

This negative variation may be partially explained by the ”lost in the middle” effect, formalized by Liu et al. [23], which demonstrates that language models exhibit a U-shaped attention bias, processing information located at the extremes (beginning and end) of the prompt more effectively than in the middle region. In the evaluated pipeline, the structural arrangement of the elements (base instruction + context block + question block) was deliberately designed to reflect the standard logical chaining of computer-assisted systems, positioning the retrieved chunks from Qdrant in the center of the prompt. It is plausible that this layout contributed to the observed accuracy degradation through a mechanism analogous to the lost-in-the-middle effect; however, as this was not the primary focus of the present study, the experiment did not vary the position of the retrieved context within the prompt, nor did it directly probe attention weights or internal model representations. This explanation should therefore be regarded as a plausible hypothesis rather than a confirmed causal mechanism, and warrants targeted investigation in future work. Additionally, as observed by Xiong et al. [11] in the MIRAGE benchmark, medical literature indicates that significant accuracy gains in complex clinical scenarios often demand the combination of a diversified corpus and multi-stage retrieval; thus, the methodological choice of a corpus strictly focused on Diabetes guidelines served as an analytical boundary, highlighting operational challenges in providing the discriminative framing necessary for multiple-choice tasks under constrained hardware conditions.

Additionally, Du et al. [24] demonstrated that simply increasing input length substantially degrades the reasoning capacity of LLMs, regardless of retrieval quality and even in the absence of distracting information—an architectural limitation previously unrecognized. Under the employed hardware configuration (8 GB VRAM, 32 GB RAM), the 8B models operated with partial layer offloading between the

GPU and RAM (*GPU split / CPU offload*). The attention window expanded by the RAG context (up to 4,000 characters per injection) imposed heavier matrix operations on weights quantized at reduced precision. The latency profile observed is compatible with a bandwidth bottleneck in the GPU–CPU bus, although this study did not collect hardware telemetry or separate timing measurements for retrieval, re-ranking, and inference stages, and this attribution therefore remains to be confirmed.

The Qwen2.5-3B model manifested the worst-case scenario: declining from 38.0% to 27.0% on MedMCQA (-11 pp) and displaying a residual TPS of 0.10–0.17 tokens/s with RAG. This is compatible with the injected context exceeding its effective attention window under Q4_K_S quantization, although the present design does not allow this mechanism to be isolated from the effects of quantization itself.

One of the main limitations of this experimental study relates to the constraints of the computational hardware used for offline execution (8 GB of VRAM), which necessitated the application of advanced quantization algorithms (such as Q4_K_M, Q4_K_S, and Q8_0) to enable model encapsulation in GGUF format. Although this strategy mitigates Out of Memory (OOM) errors, compressing the original 16-bit floating-point weights to hybrid 4-bit and 8-bit representations can induce subtle losses in the native semantic and cognitive capacities of the architectures. Furthermore, the evaluation was restricted to a RAG pipeline based on a single corpus of clinical guidelines focused strictly on Diabetes Mellitus and metabolic disorders, limiting the generalization of the findings to broader multidisciplinary medical scenarios and preventing analysis of how compact 3B and 8B parameter architectures would respond to mitigation strategies for the lost-in-the-middle effect, such as altering the positioning of the context block within the prompt or diversifying across multiple document retrievers.

V. Conclusion

It can be concluded that, contrary to the benefits widely documented in the literature for large-scale models executed in the cloud, the introduction of Retrieval-Augmented Generation (RAG) architecture in compact local models under severe hardware restrictions (8 GB of VRAM) led to a reduction of clinical accuracy in almost all evaluated scenarios, accompanied by a substantial operational penalty. The LlaMA-3.1-Med-8B model established itself as the top-performing architecture overall in the study, reaching 66.5% accuracy on MedMCQA in direct inference (native systems). However, under the RAG pipeline, this metric declined to 58.5%, a reduction of 8 percentage points.

This behavior is consistent with context expansion via document injection acting primarily as distracting noise for the selective attention mechanisms of smaller-scale networks. From a computational perspective, the approach was associated with a bottleneck in the GPU–CPU bus bandwidth attributable to partial layer offloading, substantially increasing average latency and reducing token throughput.

Given this scenario, under the specific conditions evaluated — a single, restricted document corpus, a fixed prompt structure, and severe hardware constraints (8 GB VRAM) — the standard RAG pipeline substantially increased computational cost and frequently reduced clinical accuracy relative to direct inference. These findings characterize the specific configuration tested and should not be generalized as establishing the categorical infeasibility of RAG-based approaches for real-time care applications in operational environments with resource and connectivity constraints (DIL Environments). Before such a conclusion can be drawn, and before these technologies can become reliable and operable tools in daily medical practice, future research should systematically evaluate: (a) alternative RAG configurations, including the number of retrieved documents, context window size, chunking strategies, hybrid retrieval methods, context compression techniques, caching mechanisms, and reordering of retrieved blocks within the prompt to mitigate peripheral attention bias; (b) approaches that directly test the lost-in-the-middle hypothesis — such as attention-weight analysis — to determine whether it accounts for the observed degradation or whether the underlying architecture itself requires enhancement; (c) alternative quantization schemes, larger-parameter models, and dedicated hardware optimization to enable strictly local execution without sacrificing the complex clinical reasoning required to manage high-information-density pathologies such as Diabetes Mellitus; and (d) validation across other hardware environments and clinical domains beyond Diabetes Mellitus.

Acknowledgment

The authors would like to thank the Department of Army Education and Culture (Departamento de Educação e Cultura do Exército - DECEx), the Coordination for the Evaluation and Development of Military Higher Education in the Army (Coordenadoria de Avaliação e Desenvolvimento da Educação Superior Militar no Exército - CADESM), and the Captain's Career School (Escola de Aperfeiçoamento de Oficiais - EsAO) for their valuable support throughout the course of this work. We also acknowledge Pro-Research Project (Projeto Pró-Pesquisa), from CADESM, for their financial support, which made this research possible.

References

[1] American Diabetes Association, "Standards of care in diabetes—2025," *Diabetes Care*, vol. 48, no. Suppl. 1, 2025.

[2] American Diabetes Association, "Standards of care in diabetes—2026," *Diabetes Care*, vol. 49, no. Suppl. 1, 2026.

[3] D. J. Magliano and E. J. Boyko (eds.), *IDF Diabetes Atlas*, 11th ed. Brussels: International Diabetes Federation, 2025. ISBN 978-2-930229-96-6.

[4] H. Sun et al., "IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, art. 109119, 2022. DOI: 10.1016/j.diabres.2021.109119.

[5] H. Nori et al., "Capabilities of GPT-4 on medical challenge problems," *arXiv*, arXiv:2303.13375, 2023.

[6] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023. DOI: 10.1038/s41586-023-06291-2.

[7] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, art. 248, 2023. DOI: 10.1145/3571730.

[8] Y. Gao et al., "Retrieval-augmented generation for large language models: a survey," *arXiv*, arXiv:2312.10997, 2024.

[9] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[10] D. Wang et al., "Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study," *Journal of Medical Internet Research*, vol. 26, e58041, 2024. DOI: 10.2196/58041.

[11] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," *arXiv*, arXiv:2402.13178, 2024.

[12] D. Jin et al., "What disease does this patient have? A large-scale open-domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, art. 6421, 2021. DOI: 10.3390/app11146421.

[13] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on Health, Inference, and Learning*, Proceedings of Machine Learning Research, vol. 174, pp. 248–260, 2022.

[14] L. Blonde et al., "American Association of Clinical Endocrinology clinical practice guideline: developing a diabetes mellitus comprehensive care plan—2022 update," *Endocrine Practice*, vol. 28, no. 10, pp. 923–1049, 2022. DOI: 10.1016/j.eprac.2022.08.002.

[15] S. L. Samson et al., "American Association of Clinical Endocrinology consensus statement: algorithm for management of adults with type 2 diabetes – 2026 update," *Endocrine Practice*, vol. 32, no. 4, pp. 473–518, 2026. DOI: 10.1016/j.eprac.2026.01.006.

[16] National Institute for Health and Care Excellence (NICE), "Type 2 diabetes in adults: management (NG28)," London: NICE, 2015. Updated Feb 18, 2026.

[17] P. Brown and J. Diggle, "At a glance factsheet: the NICE NG28 type 2 diabetes guideline: management update – what's new?" *Diabetes & Primary Care*, vol. 28, pp. 5–8, 2026.

[18] G. Grunberger et al., "American Association of Clinical Endocrinology clinical practice guideline: the use of advanced technology in the management of persons with diabetes mellitus," *Endocrine Practice*, vol. 27, no. 6, pp. 505–537, 2021. DOI: 10.1016/j.eprac.2021.04.008.

[19] K. Cusi et al., "American Association of Clinical Endocrinology clinical practice guideline for the diagnosis and management of nonalcoholic fatty liver disease in primary care and endocrinology clinical settings," *Endocrine Practice*, vol. 28, no. 5, pp. 528–562, 2022. DOI: 10.1016/j.eprac.2022.03.010.

[20] W. T. Garvey et al., "American Association of Clinical Endocrinologists and American College of Endocrinology comprehensive clinical practice guidelines for medical care of patients with obesity," *Endocrine Practice*, vol. 22, no. Suppl. 3, pp. 1–205, 2016. DOI: 10.4158/EP161365.GL.

[21] K. Singhal et al., "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, pp. 943–950, 2025. DOI: 10.1038/s41591-024-03423-7.

[22] J. Maharjan et al., "OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models," *Scientific Reports*, vol. 14, no. 1, p. 14156, 2024. DOI: 10.1038/s41598-024-64827-6.

[23] N. F. Liu et al., "Lost in the middle: how language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. DOI: 10.1162/tacl_a_00638.

[24] Y. Du et al., "Context length alone hurts LLM performance despite perfect retrieval," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, pp. 23281–23298, 2025. DOI: 10.18653/v1/2025.findings-emnlp.1264.

Mathematical Model to Prioritize Preventive Maintenance of Endoscopes

Mike Valentim Baroni
Biomedical Engineering. Universidad Federal do ABC
São Bernardo, SP, Brazil
mike.denner@aluno.ufabc.edu.br
ORCID: 0009-0006-6970-9159

Martha R. Ortiz-Posadas
Electrical Engineering Departament.
Universidad Autónoma Metropolitana-Iztapalapa
Mexico City, Mexico
posa@xanum.uam.mx
ORCID: 0000-0001-5032-2663

***Abstract*—The Endoscopy Department at the *Beneficência Portuguesa* Hospital in São Paulo, Brazil, has a structure comprising 18 beds for pre- and post-examination procedures, 14 beds for preparation for hospital colonoscopy with the right to an accompanying person, and six examination rooms. This infrastructure was renovated in 2018 and included the acquisition of 97 new endoscopes of different types. Despite annual preventive maintenance performed on these devices, the number of breakdowns has increased significantly: 15% in 2023 and 28% in 2024, resulting in increased costs for corrective maintenance. The objective of the present study was to develop an interpretable and operationally applicable prioritization model for Clinical Engineering practice, applying Multicriteria Decision-Making methodology, using variables routinely available in the hospital maintenance management system. The model aims to prevent possible failures, optimize clinical engineering resources, and avoid interruptions in the use of equipment, increasing its availability. The mathematical model was applied to a sample of 24 video-bronchoscopes from the Endoscopy Department of the *Beneficência Portuguesa* Hospital.**

***Keywords*—Endoscopes evaluation, technical evaluation, economic evaluation, prioritization of preventive maintenance.**

I. INTRODUCTION

Health Technology Assessment (HTA) is a multidisciplinary process that uses explicit methods to determine the value of health technology at different points in its lifecycle. The purpose is to inform decision-making to promote an equitable, efficient, and high-quality health system [1]. Hospital-Based Health Technology Assessment (HB-HTA) is a specific context within and for HTA in hospitals. It is an evidence-based approach that considers a particular technology in a hospital setting, for moving the assessment closer to the point of care, where technologies' impacts and benefits can be directly assessed [2]. HB-HTM encompasses tasks related with health technology management processes, such as preventive maintenance [3].

The *Beneficência Portuguesa* Hospital (BP), located in São Paulo, Brazil, is one of the largest private healthcare centers in the country. The institution has 721 beds and an extensive inventory of medical equipment that supports healthcare services across a wide range of clinical specialties [4]. Among its diagnostic units, the Endoscopy Department stands out, comprising 18 beds for pre- and post-procedure care, 14 beds dedicated to hospital colonoscopy preparation with accommodation for a companion, and six examination rooms [5].

In 2018, this infrastructure underwent a major modernization process through a comprehensive renovation and restructuring, which included the renewal of the technological park of endoscopic equipment. An endoscope is a device used for the direct visualization of internal structures of the human body and play a fundamental role in minimally invasive diagnostic and therapeutic procedures [6]. As part of this modernization initiative, 97 new endoscopes of different types were acquired, including [video] gastroscopes, colonoscopes, bronchoscopes, and duodenoscopes, among others.

Currently, the criterion used to determine whether an endoscope undergoes preventive maintenance is time-based, following the manufacturer's recommendations, which define the annual maintenance interval. Despite the annual preventive maintenance performed on these devices, the number of failures and damages has increased significantly over the years: 15% in 2023, and 28% in 2024. This trend has resulted in a substantial increase in corrective maintenance costs.

In this context, the Hospital's Clinical Engineering Department has intensified efforts to investigate the causes of these failures and to identify solutions aimed at reducing the frequency of corrective maintenance and the associated costs. Therefore, the objective of the present study was to develop an interpretable and operationally applicable model for Clinical Engineering practice, using variables routinely available in the hospital maintenance management system. The proposed model aims to prevent potential failures, optimize clinical engineering resources, and avoid interruptions in equipment use, thereby increasing equipment availability. The model was applied to a sample of 24 video-bronchoscopes from the Endoscopy Department of the *Beneficência Portuguesa* Hospital.

Although several studies have proposed indicators to support maintenance management in healthcare technology, most approaches evaluate technical or economic aspects separately [7–11]. The contribution of the present work lies in the integration of failure frequency, downtime, preventive and corrective maintenance history, equipment age, depreciation, and maintenance costs into a single Global Indicator designed to support preventive maintenance prioritization. The integration of technical reliability and economic burden into a single indicator allows maintenance prioritization beyond conventional time-based scheduling approaches. Unlike conventional time-based maintenance policies recommended by manufacturers, the proposed framework combines operational reliability and economic impact, allowing maintenance decisions to be guided by evidence obtained from real

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

equipment performance data. This approach provides a practical decision-support mechanism for Clinical Engineering departments operating in data-constrained hospital environments.

Unlike predictive maintenance approaches based on large datasets and machine learning algorithms, the proposed model was designed for hospital environments where historical maintenance information is available, but data volume may be insufficient for advanced predictive analytics.

II. Methodology

Multicriteria Decision-Making (MCDM) methodology has been widely applied to support decision-making processes involving multiple and often conflicting criteria. By assigning relative importance to technical, operational, and economic factors, MCDM approaches enable the prioritization of alternatives through a transparent and reproducible evaluation framework [12]. In Health Technology Management, these methodologies have been used to prioritize medical equipment according to variables such as failure frequency, maintenance costs, equipment age, utilization rate, technological obsolescence, and clinical criticality, thereby supporting more efficient allocation of maintenance resources and improving equipment availability [13].

A. Definition of Variables

Variables related to the technical and economic aspects of the endoscopes were defined. It is important to note that these variables have a qualitative domain, which is used by the technical staff of the Hospital's Cinical Engineering Department (CED) to assess medical devices. However, in order to develop the mathematical model proposed in this work, it was necessary to map the qualitative domain into the interval [0,1] to obtain a quantitative domain. Each variable was assigned a weighting factor according to its importance in the evaluation, based on the experience of the CED technical staff. The weighting factors assigned to each variable were defined according to their perceived influence on equipment availability, maintenance burden, and operational continuity. Variables directly associated with service interruption and equipment reliability, such as failure frequency and downtime, received higher weights because they have immediate consequences for clinical workflow and patient care. In contrast, variables with a more indirect influence, such as manufacturer brand, received lower weights.

B. Indicators Development

Once the variables and their corresponding weights were defined, two indicators were established: technical and economic. The resulting numerical value was interpreted through a qualitative scale indicating the technical (or economic) condition of the evaluated medical device. Subsequently, a mathematical expression combining both indicators was developed to obtain a global indicator for an overall assessment of the device.

The qualitative intervals adopted for the interpretation of the indicators were established according to maintenance management practices currently used by the CED. The selected thresholds were intended to distinguish equipment with satisfactory operational performance from those requiring increased preventive attention. Although these intervals are based on institutional experience, they provide a practical framework for decision-making and facilitate the translation of quantitative indicator values into actionable maintenance strategies.

C. Mathematical Model Validation

The mathematical model was validated using a sample of 24 video-bronchoscopes from the Endoscopy Department. All technical and maintenance-related information, including failure history, downtime, preventive and corrective maintenance interventions, and associated costs, was extracted from the Genesis® Hospital Maintenance Management System [14], the official computerized maintenance management system (CMMS) used by the Clinical Engineering Department. The model outputs were then compared with the actual maintenance scenario observed during 2024 to evaluate the consistency and practical applicability of the proposed prioritization framework in a real hospital environment.

Additionally, a sensitivity analysis was performed. The purpose of this analysis was to evaluate the robustness of the prioritization framework against moderate changes in the relative importance assigned to the technical and economic components. The Global Indicator was recalculated considering a ±10% variation in the weighting coefficients of both indicators used in the model.

III. Results

A. Technical Indicator

Six variables (x_i) were defined [15]. Each associated with a relevance factor (ρ_i) as shown in Table 1.

TABLE I. Technical variáables and its relevance factor

x_i	Description	ρ_i
x_1	Failure frequency	1.0
x_2	Days out of use	1.0
x_3	Types of failures	0.9
x_4	Preventive maintenance interventions	0.8
x_5	Equipment age (years)	0.7
x_6	Brand	0.2

The technical indicator (I_T) [15] incorporates the six variables and their corresponding weighting factor into a mathematical function, as shown in (1).

$$I_T = \sum_{i=1}^{6} \rho_i x_i \quad (1)$$

Where: x_i is the variable, i= {1, …, 6}
ρ_i is the weighting factor for each variable.

The result of the technical indicator (I_T) was bounded within the interval [0,1]. A normalization factor (N) was therefore calculated using the expression (2).

$$N = \sum_{i=1}^{6} \rho_i M_{imáx} \quad (2)$$

Therefore, I_T is defined by (3):

$$I_T = \frac{\sum_{i=1}^{6} \rho_i x_i}{N} \quad (3)$$

Qualitative scale. It was defined with three technical conditions corresponding to the overall state of the equipment (Table II). Equipment classified as *Good*, indicates the need for one preventive maintenance intervention per year. Those rated as *Regular*, preferably require two preventive maintenance interventions per year. Finally, equipment classified as *Poor,* due to its less satisfactory technical condition, requires three preventive maintenance interventions per year. The definition of the intervals was established by the Clinical Engineering Department based on its technical expertise and practical experience [15].

TABLE II. QUALITATIVE SCALE FOR I_T INTERPRETATION

Result (I_T)	Technical condition	Preventive maintenance priority	Interpretation (No. interventions)
[0.0, 0.3]	Good	Low	1 a year
[0.31, 0.6]	Regular	Medium	2 a year
[0.61, 1.0]	Poor	High	3 a year

B. Economic Indicator

Variables: Five variables were defined, as presented in Table III, from which two partial economic indicators (I_{E1} e I_{E2}) were derived [16].

TABLE III. ECONOMIC VARIABLES

x_i	Variables	Description
x_7	CPM	Cost of preventive maintenance
x_8	CCM	Cost of corrective maintenance
x_9	CNE	Cost of a new endoscope
x_{10}	EA	Equipment age (years)
x_{11}	CAQ	Endoscope acquisition cost

Partial Economic Indicator 1 (I_{E_1}). This indicator represents the ratio between the total costs of both, preventive maintenance (CPM) and corrective maintenance (CCM), and the cost of a new equipment unit (CNE). It reflects the financial impact of technical interventions relative to the equipment replacement value. It is defined by equation (4) [16].

$$I_{E1} = \frac{(CPM+CCM)}{CNE} \quad (4)$$

Partial Economic Indicator 2 (I_{E2}). It expresses the depreciation of the equipment over time relative to its acquisition cost. Equation (5) represents the accumulated economic value loss as a function of the equipment age [16].

$$I_{E2} = \left(\frac{CNE-\text{CAQ}}{CAQ}\right)\left(\frac{EA}{EA_{highest}}\right) \quad (5)$$

Where: $Ea_{highest}$, is the highest age observed in the sample.

Economic Indicator (I_E). It is obtained through the weighted combination of the two partial indicators (Eq. 6) [16].

$$I_E = \frac{1{,}0\ (I_{E1})+0{,}7(I_{E2})}{1{,}7} \quad (6)$$

The weights 1.0 and 0.7 were defined based on the greater relevance of maintenance cost in the economic analysis, since it is directly influenced by the frequency of preventive interventions, whereas depreciation is considered a secondary factor in this process.

Qualitative scale. The interpretation of the numerical value of I_E was defined in three qualitative ranges, corresponding to the economic condition of the equipment (Table IV). This classification guides the recommended number of annual preventive maintenance interventions, following a criterion analogous to that adopted for the technical indicator (I_T). Thus, higher values of (I_E) indicate a greater likelihood that the equipment will require more preventive interventions, ranging from one to three maintenance actions per year according to the defined intervals [16].

TABLE IV. QUALITATIVE SCALE FOR I_E INTERPRETATION

Interval I_E	Economic condition (Priority)	Interpretation (No. interventions)
[0.0, 0.25]	Low cost	1 a year
[0.26, 0.5]	Medium cost	2 a year
[0.51, 1.0]	High cost	3 a year

C. Global Indicator

In the composition of the Global Indicator (I_G), the two indicators (I_T and I_E) were weighted in order to integrate the operational performance and cost–benefit aspects of each endoscope into a single evaluation. The mathematical expression that defines the I_G is presented in Eq. (7) [16].

$$I_G = 0{,}6\ (I_T) + 0{,}4\ (I_E) \quad (7)$$

The value of I_G ranges within the interval [0,1]. Values approaching 1 indicate high priority for preventive maintenance, as they represent equipment with higher technical and/or economic impact in the context of maintenance management. It is important to note that the interpretation of the I_G results, follows the same criteria defined in Table IV.

D. Mathematical Model Validation

The mathematical model was applied to a sample of 24 video-bronchoscopes, including both adult and pediatric devices. The results for each case are presented below.

E. Technical Evaluation

According to the qualitative scale for the I_T interpretation (Table II) the results indicated that five video-bronchoscopes were in *Good* technical condition, six were classified as *Regular*, and thirteen as *Poor*.

For video-bronchoscopes in *Good* technical condition ($I_T \leq 0.3$), observed that B_1 (Table V), provides excellent image resolution and has a large diameter suitable for adult use. Devices B_2, B_3, and B_4 are used for pediatric examinations. The Endocopy. Considering that the demand for pediatric procedures is relatively low, these devices are usually kept in reserve storage, which explains their high technical rating. B_5 is typically used only when the other devices are unavailable for operation. Its high rating is therefore associated with its low utilization rate during the analyzed period.

TABLE V. RESULTS OF THE TECHNICAL (I_T), ECONOMIC (I_E) AND GLOBAL (I_G) EVALUATIONS OF 24 VIDEO-BRONCHOSCOPES

Video-bronchoscope	I_T	I_E	I_G	Priority
B_1	0.18	0.09	**0.15**	
B_2	0.18	0.09	**0.15**	
B_3	0.18	0.09	**0.15**	**Low**
B_4	0.18	0.05	**0.16**	
B_5	0.23	0.13	**0.16**	
B_6	0.36	0.09	**0.25**	
B_7	0.39	0.15	**0.30**	
B_8	0.40	0.21	**0.32**	
B_9	0.49	0.11	**0.34**	
B_{10}	0.46	0.17	**0.34**	
B_{11}	0.55	0.10	**0.37**	**Medium**
B_{12}	0.63	0.15	**0.44**	
B_{13}	0.63	0.25	**0.48**	
B_{14}	0.77	0.04	**0.48**	
B_{15}	0.63	0.29	**0.49**	
B_{16}	0.73	0.20	**0.52**	
B_{17}	0.73	0.24	**0.53**	
B_{18}	0.73	0.24	**0.53**	
B_{19}	0.73	0.26	**0.54**	
B_{20}	0.75	0.23	**0.54**	**High**
B_{21}	0.73	0.27	**0.54**	
B_{22}	0.73	0.31	**0.56**	
B_{23}	0.75	0.32	**0.58**	
B_{24}	0.73	0.49	**0.63**	

In the *Regular* classification ($0.3 < I_T \leq 0.6$), video-bronchoscopes B_6, B_9, B_{10}, and B_{11} were identified as devices used for pediatric examinations. (with low demand). B_6 obtained a score of 0.36, because of preventive maintenance was not performed. B_9, B_{10}, and B_{11} each underwent one corrective maintenance intervention, which led them also to fall into the same category. B_7, which is used for conventional adult examinations, iobtainde a somewhat atypical classification, because it is the only device of its model within this category. Nevertheless, it received this rating because it required only one corrective maintenance intervention and remained out of service for approximately two months. B_8 is dedicated to procedures that combine video-bronchoscopy with ultrasound diagnostics. It received this classification because it is used in very few cases and, when operated, its handling is performed by specialists. It should also be noted that the device remained out of service for only two days.

Finally, the video-bronchoscopes classified as *Poor* ($I_T > 0.61$) are used for adult patients and all of them underwent at least one corrective maintenance intervention (B_{22} and B_{23} required two corrective interventions), did not receive preventive maintenance. The device with the shortest downtime remained out of service for 137 days, which justifies this classification. Within this group, B_{16} is noteworthy, as it is a pediatric video-bronchoscope. It received this rating because it remained out of service during all year (365 days) and did not undergo preventive maintenance, which explains its low classification. Another case is B_{14}, this device is dedicated to procedures combining bronchoscopy and ultrasound diagnostics, and its prolonged absence from service was due to the need for high-cost imported components.

The remaining video-bronchoscopes classified as *Regular* and *Poor* are consistent with these results, as they include biopsy channels and are widely used in adult examinations, which represent the most common procedures in the department

F. Economic Evaluation

According to the qualitative scale for the I_E interpretation (Table IV), the overall results were as follows: 14 video-bronchoscopes were classified with *Low* cost, seven with *Medium* cost, and one with *High* cost (Table V). It should be noted that the first three devices received a low classification because they required few corrective maintenance interventions, resulting in lower total maintenance costs. It was also observed that device B_9 is the only video-bronchoscope with a larger diameter—typically used exclusively in adult patients—and, due to its limited use, it rarely undergoes corrective maintenance, which consequently results in lower maintenance costs.

Among the devices classified with *Medium* cost ($0.25 < I_E \leq 0.50$), are predominantly video-bronchoscopes with intermediate and large diameters used in adult examinations across different anatomical conditions. For devices classified with *High* cost ($I_E > 0.50$), as expected, a higher corrective maintenance cost corresponds to this index value. In particular, the highest maintenance cost ($ 16,200 USD) was associated with device B_{24}, which was therefore classified with the highest cost. It is also noteworthy that the maintenance cost of this device corresponds to approximately 65% of its acquisition cost, indicating that its replacement should be considered.

In general, it can be observed that there is an inverse relationship between the technical condition and the economic condition. That is, the better the technical condition of the device, the lower the maintenance cost. The same applies to the preventive maintenance priority for the device. The better the device's technical condition, the fewer preventive maintenance interventions it requires.

G. Global Evaluation

To obtain the global evaluation (priority), equation (7) was applied to the 24 video-bronchoscopes. The results reveal a clear stratification of the equipment into three maintenance-priority categories, with 25% of the devices classified as low priority, 37.5% as medium priority, and 37.5% as high priority (Table V). The gradual increase in I_G values and the absence of abrupt fluctuations suggest a consistent prioritization framework capable of distinguishing different levels of maintenance need, as can be seen in the preventive maintenance priority ranking based on the Global Indicator (Fig. 1). Notably, nine devices were classified as high priority ($I_G > 0.50$), indicating a substantial proportion of the fleet with elevated technical and economic maintenance requirements. These findings demonstrate that the proposed model can effectively support preventive maintenance planning by identifying equipment requiring increased attention and enabling a more efficient allocation of Clinical Engineering resources.

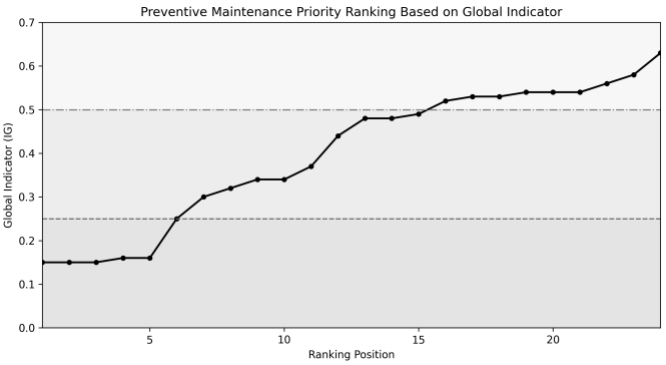


Fig. 1. Preventive maintenance priority ranking based on the Global Indicator.

To better understand the interaction between technical performance and economic impact, the three indicators developed in this study were evaluated jointly. The Technical Indicator (I_T) assesses operational reliability based on maintenance history and equipment condition, whereas the Economic Indicator (I_E) reflects the relative maintenance cost and aging burden of each device. By combining these two dimensions, the Global Indicator (I_G) provides a balanced assessment of preventive maintenance needs This comparison also provides indirect evidence of the consistency of the proposed model, since equipment with poorer technical conditions generally tend to exhibit higher economic impact and, consequently, higher global priority scores.

The values of I_T, I_E, and I_G for the evaluated video-bronchoscopes are compared in Fig. 2. An overall increasing trend can be observed among the three indicators, suggesting a relationship between technical deterioration and economic burden. Devices with higher technical scores generally exhibited higher economic scores, indicating that increased failure frequency, downtime, and maintenance interventions are associated with greater maintenance costs. The Global Indicator follows a similar pattern because it integrates both dimensions through a weighted approach, providing a balanced assessment of equipment condition.

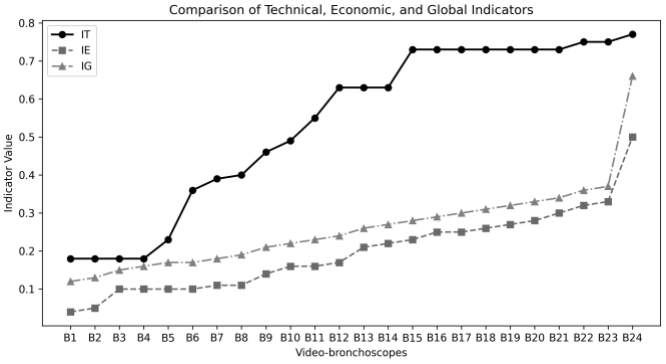


Fig. 2. Comparative profile of the Technical Indicator (IT), Economic Indicator (IE), and Global Indicator (IG) for the evaluated video-bronchoscopes.

These results support the rationale of combining technical and economic factors into a single decision-making index capable of prioritizing preventive maintenance actions more effectively than conventional time-based maintenance schedules.

H. Model Sensitivity

The result of the sensitive analysis is shown in Table VI. In general, the variations in I_G values were proportional and relatively small across all evaluated devices. As expected, increasing the weights (+10%) resulted in a corresponding increase in I_G values, whereas reducing the weights (–10%) produced lower scores. However, these changes did not substantially modify the overall distribution of the equipment, or the prioritization structure generated by the model.

From a maintenance management perspective, the results indicate that the proposed methodology exhibits satisfactory stability under moderate perturbations of the weighting coefficients. Devices with low I_G values, such as B1–B5, remained within the lowest-priority group under both scenarios, while equipment with higher IG values, particularly B22–B24, consistently occupied the highest positions in the ranking. Although some devices located near the classification thresholds experienced slight variations in their scores, the overall prioritization pattern remained largely unchanged. Notably, video-bronchoscope B24 maintained the highest maintenance priority in all scenarios, with IG values ranging from 0.61 to 0.66, confirming the consistency of the model in identifying critical equipment. These findings suggest that the proposed prioritization framework is not excessively dependent on a specific set of weighting coefficients and that the resulting maintenance decisions are reasonably robust to moderate uncertainty in expert-defined parameters.

TABLE VI. SENSITIVITY RESULTS OF THE GLOBAL INDICATOR IG

Video-bronchoscope	I_G	I_G (+10%)	I_G (-10%)
B_1	0,15	0,16	0,14
B_2	0,15	0,16	0,14
B_3	0,15	0,16	0,14
B_4	0,16	0,17	0,14
B_5	0,16	0,17	0,16
B_6	0,25	0,28	0,22
B_7	0,30	0,33	0,28
B_8	0,32	0,34	0,30
B_9	0,34	0,38	0,30
B_{10}	0,34	0,37	0,31
B_{11}	0,37	0,42	0,33
B_{12}	0,44	0,49	0,39
B_{13}	0,48	0,52	0,44
B_{14}	0,48	0,55	0,41
B_{15}	0,49	0,53	0,46
B_{16}	0,52	0,57	0,46
B_{17}	0,53	0,58	0,48
B_{18}	0,53	0,58	0,48
B_{19}	0,54	0,59	0,49
B_{20}	0,54	0,59	0,49
B_{21}	0,54	0,59	0,50
B_{22}	0,56	0,60	0,52
B_{23}	0,58	0,62	0,54
B_{24}	0,63	0,66	0,61

IV. Discussion

The proposed prioritization model differs substantially from the preventive maintenance strategy currently adopted in the institution, which is based primarily on fixed annual intervals recommended by equipment manufacturers. While time-based maintenance schedules assume homogeneous equipment degradation, the proposed model incorporates actual operational performance, maintenance history, and economic impact. Consequently, it enables differentiated maintenance planning according to the specific condition of each device, potentially improving resource allocation and reducing unnecessary interventions.

A. Limitations of the Study

Some limitations of this study should be acknowledged. First, the model was validated using a relatively small sample of 24 video-bronchoscopes from a single hospital department, which limits the generalizability of the findings.

Second, the weighting coefficients and qualitative classification thresholds were established using expert judgment and institutional experience then, future studies should complement using more advanced uncertainty assessment techniques, in order to provide a more comprehensive evaluation of parameter uncertainty and further strengthen the methodological validity of the proposed model.

Third, some variables may present partial correlation, particularly downtime, failure frequency, and corrective maintenance cost. However, these variables were intentionally preserved because each reflects a distinct operational dimension relevant to Clinical Engineering decision-making.

Therefore, the results should be interpreted as evidence of practical applicability rather than definitive predictive performance. Future research should evaluate the model using larger datasets, additional equipment categories, and multicenter environments.

V. Conclusion

The results obtained from the evaluation of 24 video-bronchoscopes suggest that the proposed technical, economic, and global indicators provide a practical framework for supporting preventive maintenance prioritization in Clinical Engineering. The model demonstrated the ability to distinguish equipment with different operational and economic profiles using information routinely available in the hospital maintenance management system. Although the validation was limited to a single equipment category and hospital department, the findings indicate the potential usefulness of indicator-based maintenance planning as a complement to conventional time-based maintenance policies.

The development of the Global Indicator (I_G), with weights of 0.6 assigned to the technical aspect and 0.4 to the economic aspect (Eq. 7), allowed the analysis to balance operational reliability and maintenance cost, prioritizing equipment safety and availability while maintaining financial Sustainability. Therefore, it can be concluded that the Global Indicator model developed in this study, significantly contributes to improve the management of preventive maintenance for medical equipment. Its continuous application may promote a more strategic, indicator-driven management approach, reducing costs, optimizing resources, and increasing the technological availability of hospital units. Furthermore, future studies are recommended to extend the application of the model to other sectors and types of equipment in order to validate its effectiveness in different contexts and enhance its use as a decision-support tool in Clinical Engineering. Future work will also evaluate AI-based predictive maintenance approaches to further improve maintenance planning and resource allocation.

Acknowledgement

The authors would like to thank the technical staff of the Clinical Engineering Department at Beneficência Portuguesa Hospital in São Paulo, whose support made the development of this work possible.

References

[1] World Health Organization: Health technology assessment of medical devices. 2nd edition. Geneva: WHO, 2025 (WHO medical device technical series). Available at: https://www.who.int/publications/i/item/9789240110878#:~:text=This%20document%20describes%20the%20critical,978%2092%204%200110 87%208, last accessed 2025/02/17.

[2] Itaborahy A, Santos VCC, Salomon FCR, Bregman C, Mateus I, Santos M. Hospital-based health technology assessment (HTA) in Latin America and the Caribbean: a scoping review. Rev Panam Salud Pública. 2025 Jan 13;49:e8. https://doi.org/10.26633/RPSP.2025.8

[3] Hernández-López LA, Pimentel-Aguilar AB, Ortiz-Posadas MR: An index to prioritize the preventive maintenance of medical equipment. Health and Technology, **10** (2): 373-375, 2019. DOI: 10.1007/s12553-020-00414-9.

[4] Beneficência Portuguesa de São Paulo. Uma história de valorização da vida. 2025. Disponível em: https://www.bp.org.br/institucional/sobre-a-bp Último acesso em: 23/03/2025 (In Portuguese)

[5] Beneficência Portuguesa de São Paulo. Medicina Diagnóstica inaugura novas instalações do Serviço de Endoscopia. 2018. Disponível em: https://www.bp.org.br/institucional/noticias/bp-medicina-diagnostica-inaugura-novas-instalacoes-do-servico-de-endoscopia Último acesso em: 23/03/2025 (In Portuguese)

[6] Pentax Medical. Vídeo-Broncoscópio PENTAX Medical. Instruções de uso. Série EB-J10. 2019.

[7] Cruz-Hernández, M.d.L., Gutiérrez-Sánchez, B., Ortiz-Posadas, M.R.: A Technical Assessment Tool to Prioritize Medical Devices Replacement. In: Flores Cuautle, J.d.J.A., *et al.* XLVI Mexican Conference on Biomedical Engineering. CNIB **2023**. IFMBE Proceedings, vol 97. Springer, Cham. https://doi.org/10.1007/978-3-031-46936-7_22

[8] Hernández-López LA, Pimentel-Aguilar AB and Ortiz-Posadas MR: An index to prioritize the preventive maintenance of medical equipment. *Health and Technology*, 10(2): 373-375, **2019**. DOI: 10.1007/s12553-020-00414-9

[9] Mora-García TR, Hernández-López LA, Piña-Quintero MF, Pimentel-Aguilar AB and Ortiz-Posadas MR.: "Comparative analysis of two indicators of technical evaluation for the replacement of medical equipment". VIII Latin American Congress on Biomedical Engineering, CLAIB 2019. IFMBE Proceedings, 75: 1365-1372, **2019**.

[10] Guzmán-Canizales, L., Hernández-Martínez, A.A., Ortiz-Posadas, M.R: Cost-Benefit Analysis of Integrating a PET-CT Service in a High Specialty Hospital in Mexico. In: Comas, D.S., Meschino, G.J., Pérez-Buitrago, S.M., Montini Ballarin, F. (eds) Advances in Bioengineering and Clinical Engineering 2025. Springer, Cham. IFMBE Proceedings, vol 136: 384-393, **2025.**

[11] Flores-Estrada A, Bahena-Basave MA and Ortiz Posadas MR.: "Cost-Benefit Analysis of Respiratory Humidifiers Rental in a Second Level Private Hospital". Proceedings 3rd International Clinical Engineering and Health Technology Management Congress. Global Clinical Engineering Journal. Special Issue 2, **2019.**

[12] Tzeng, G. H., & Huang, J. J.: Multiple attribute decision making methods and applications. CRC press. New York (2011).

[13] Mora-García T, Piña-Quintero MF and Ortiz-Posadas MR: "Pattern Recognition for Supporting the Replacement of Medical Equipment at Hospitals". En: *Pattern recognition techniques applied to biomedical problems.* M Ortiz-Posadas**,** (Ed.) Springer. Series STEAM-H (Science, Technology, Engineering, Agriculture, Mathematics and Health*)*. p. 197-215, **2020**.

[14] Genesis®. Solução de Gerenciamento de Manutenção de Equipamentos Hospitalares. Disponível em: https://genesisz.net/hospitalar/ Último acesso em: 06/05/2025.

[15] M. Baroni and M. R. Ortiz-Posadas, "Um indicador de avaliação técnica para equipamentos médicos baseado em evidência". II Congresso Brasileiro Einstein de Engenharia Clínica, Faculdade Israelita de Ciências da Saúde Albert Einstein, São Paulo, 2025. (In Portuguese)

[16] M. Baroni, "Um modelo para priorizar a manutenção preventiva de equipamentos médicos. Trabalho de graduação, 2025. Curso em Engenharia Biomédica. Universidad Federal do ABC. São Bernardo do Campo, São Paulo, Brasil. (In Portuguese)

Monitoring of orofacial thermal patterns after physiotherapy intervention in sequelae of mastoid fracture, using infrared thermography: a case study

A.C.R. Silva
Biomedical Eng. Program - COPPE
Federal University of Rio de Janeiro,
Rio de Janeiro, Brazil
carolina.rodrigues@peb.ufrj.br

T.P. Omena
Federal University of Rio de
Janeiro/Department of Physiotherapy,
Rio de Janeiro, Brazil
thaisomena@gmail.com

J. Oliveira. Jr.
Biomedical Eng. Program - COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
jessejunior@peb.ufrj.br

W.C.A Pereira
Biomedical Eng. Program - COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
wagner@peb.ufrj.br

***Abstract*—Infrared thermography (IRT) is an increasingly used technique to assess the functional conditions of masticatory structures. In this exploratory case study, IRT was used to monitor the orofacial thermal patterns of a 60-year-old patient with sequelae of a right mastoid fracture, before and immediately after a multimodal physiotherapy session. Five consecutively captured thermograms, obtained with a FLIR E40 camera over the temporomandibular joints (TMJs) and anterior temporal muscles, bilaterally, were analyzed. The analysis included arithmetic average of the point temperatures, sample standard deviations, and pre-post differences accompanied by Type A uncertainty. The differences were 0.22 ± 0.14 °C in the right TMJ, 0.50 ± 0.11 °C in the left TMJ, 0.28 ± 0.16 °C in the right temporal muscle, and 0.26 ± 0.10 °C in the left temporal muscle. ± indicates the standard uncertainty Type A (k = 1), estimated under the adopted calculation assumptions. A reduction in the average was observed in the four regions, with similar magnitudes in the temporal regions and distinct magnitudes in the TMJs. The pain score on the Visual Analogue Scale went from 8 to 0 and the mouth opening from 37 to 40 mm. In this case, the monitoring associated the regional description of thermal changes with the variability of the captures and the uncertainty of the differences, adding quantitative information to the clinical documentation of the session.**

Keywords — Infrared thermography; Temporomandibular dysfunction; Mastoid fracture; Measurement uncertainty.

I. INTRODUCTION

Temporomandibular disorders (TMD) are a group of musculoskeletal and neuromuscular conditions that affect the temporomandibular joint (TMJ) and the structures of the masticatory system. Their etiology is multifactorial, involving physical, behavioral, and emotional factors; a history of trauma is part of the clinical evaluation context [1].

Infrared thermography (IT) offers a quantitative and non-invasive record of cutaneous thermal distribution. Studies in TMD investigate temperature differences at facial anatomical points and their relationship to the clinical condition [2]. Bilateral assessment allows describing how temperatures are distributed between corresponding regions, while recordings at different times document their changes throughout the course of treatment.

Infrared thermography (IRT) is a non-invasive tool used to investigate musculoskeletal pain management [3]. Thermal recording of a multimodal session adds regional information to pain and mouth opening measurements. By associating the magnitude of changes with the variability of captures, metrological analysis provides an explicit basis for interpreting the observed differences.

The aim of this study was to monitor orofacial thermal patterns before and immediately after a multimodal physiotherapy session in a patient with sequelae of mastoid fracture, describing the differences between the average point temperatures of the TMJs and anterior temporal muscles and estimating the Type A uncertainty associated with these differences.

The absolute skin temperature can be influenced by environmental factors, such as the use of distances greater than 50 cm between the target and the device in image acquisition protocols [4].

II. MATERIALS AND METHODS

A. Experimental configuration

This exploratory case study, with one participant (N = 1), one intervention session, and two acquisition points, without a control condition, was conducted in accordance with REBEc. Thermographic data collection was approved by the Research Ethics Committee of the Clementino Fraga Filho University Hospital and the Faculty of Medicine of the Federal University of Rio de Janeiro (UFRJ) under CAAE 78605524.3.00005257, and the physiotherapy intervention performed at the HUCFF general outpatient clinic was approved under CAAE 05063218.7.0000.5257. The patient signed informed consent to participate in the study and for the use of her thermographic data.

The 60-year-old participant presented with a history of fracture of the right mastoid process and reported bilateral pain in the temporomandibular joints (TMJs), masticatory muscles, and cervical region. The Fonseca questionnaire showed a score of 95/100, corresponding to a severe TMD classification according to the instrument. The physical examination revealed reduced range of motion in the oral

We thank CNPq, CAPES, and FAPERJ for their financial support.

cavity, lateral deviation, cervicogenic balance deficit, and headache.

She underwent a 60-minute physiotherapy intervention using temporomandibular joint (TMJ) and cervical spine mobilization, cervical traction, and laser therapy (4 J/cm² per point) applied to three points of the masseter muscle and three points of the TMJ, totaling 24 J/cm² of accumulated fluence.

B. Acquisition and definition of the analyzed quantity

The test environment (room) was maintained at a temperature between 21°C and 23°C, in accordance with the International Thermography Protocol (IACT), with a relative humidity of approximately 50%, and thermographic images were acquired at a standardized distance of 50 cm before (Pre) and immediately after (Post) the intervention (Fig. 1).



Fig. 1. Scenario for capturing thermographic images with a distance of 50 centimeters between the camera and the base of the chair.
Source: Author

The experimental setup utilized a FLIR E40 infrared camera (Teledyne FLIR LLC/United States) operating at 60 Hz, with a thermal sensitivity of 0.07 °C, emissivity of 0.98, and an integrated real resolution of 160 x 120 pixels. Five consecutive readings were obtained per region of interest (ROI), before and after the intervention. The immediate post-intervention corresponds to the evaluation at the end of the session; the exact interval until the first capture is not specified in the described protocol.

The image acquisition process respected the 10-minute acclimation period. The patient was instructed to avoid using cosmetics, creams, or lotions on the areas being evaluated on the day of the examination.

The ROIs were right TMJ (TMJ D), left TMJ (TMJ L), right anterior temporal (Temporal D), and left anterior temporal (Temporal L). The anatomical regions were manually delimited, with the cursor reading the point identified as having the highest temperature in each region. The measurand was the average of the five point temperatures selected per ROI and time point. This average summarizes the captures, not all the pixels of the ROI. The participant is the clinical unit; the captures are technical repetitions to describe the variability of acquisition.

In Fig. 2, temperature values were extracted from the standardized anatomical hot spot cursor (white cross) in each thermogram. Pain intensity (Visual Analogue Scale - VAS) and mouth opening amplitude were measured before and after the intervention.

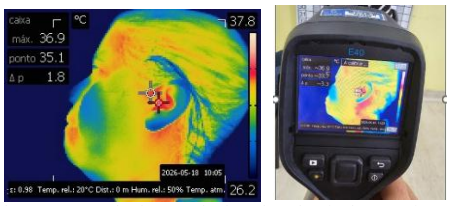


Fig. 2. Thermograms before and after the physiotherapy intervention, respectively.
Source: Author

C. Descriptive analysis and Type A uncertainty

For each ROI and time point j (pre or post), the arithmetic average of nj = 5 temperatures and the sample standard deviation sj were calculated, with divisor nj − 1. The Type A standard uncertainty of the average, uj, was estimated by the dispersion of the catches [5, 6]:

$$\overline{T}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} T_{ij}\,; \quad u_j = \frac{s_j}{\sqrt{n_j}} \quad (1)$$

The difference was defined as $\Delta T = \overline{T}pre - \overline{T}post$; positive values indicate a reduction in average temperature. The comparison was performed between the averages of the acquisition blocks, since the protocol records five captures at each time point, without documented correspondence between pre- and post-acquisition pairs. The calculation hypotheses adopted were independent readings and stability within each block, with uncorrelated pre- and post-acquisition averages.

$$u_A(\Delta T) = \sqrt{u_{pré}{}^2 + u_{pós}{}^2} \quad (2)$$

Eq. (2) combines the variance components of the two time points. In the Type A uncertainty balance, the contribution of each phase was calculated by the ratio between uj² and the sum of the pre and post components, multiplied by 100. The covariance between the pre and post average was considered zero in the calculation model.

The expanded Type A uncertainty, U95,A, and the range ΔT ± U95,A were calculated with an approximate nominal coverage of 95%, using the distribution of the effective degrees of freedom of Welch–Satterthwaite [5]:

$$\nu_{eff} = \frac{(u_{pré}{}^2 + u_{pós}{}^2)^2}{\frac{u_{pré}{}^4}{4} + \frac{u_{pós}{}^4}{4}} \quad (3)$$

$$U_{95,A} = k\,u_A(\Delta T); \quad k = t_{0,975;\,\nu_{eff}} \quad (4)$$

The expansion assumes approximately normal fluctuations, in addition to the previous assumptions. The intervals are individual, referring to the session and restricted to the evaluated Type A components. To complement the regional description, the differences between the right and left averages were calculated. The analyses used the 40 available readings, without excluding any values, with rounding only at the end of the calculations.

III. RESULTS

Average temperatures decreased in all four ROIs (TABLE I and Fig. 3). The reductions were 0.50 °C in ATM E and 0.22 °C in ATM D. During storms, the magnitudes were similar: 0.28 °C on the right and 0.26 °C on the left.

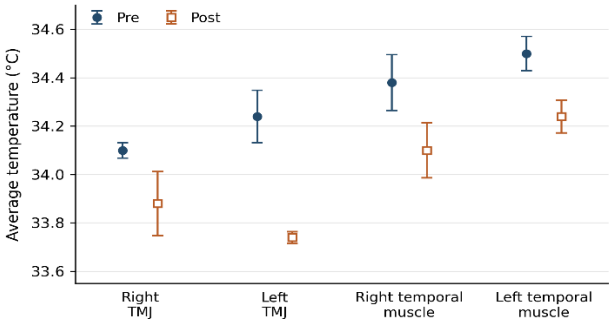


Fig. 3. Average point temperatures before and after the session. Bars: Type A standard uncertainty of the average (k = 1). The pre-post differences and their expanded uncertainties are presented in Table II.
Source: Author

In the TMJs, the side with the higher average shifted from the left in the pre-test to the right in the post-test, maintaining the magnitude of the lateral contrast observed at 0.14 °C. In the temporal lobes, the left average remained higher than the right: a difference of 0.12 °C in the pre-test and 0.14 °C in the post-test.

The dispersion between captures was specific to each ROI and time point. In ATM D, the DP went from 0.071 to 0.295 °C; in ATM E, from 0.241 to 0.055 °C. In the temporal phases, the DPs remained close between pre- and post-capture. In the Type A variance balance of the difference, the post-capture period accounted for 94.6% in ATM D and the pre-capture period for 95.1% in ATM E. In the temporal phases, the contributions of the two phases ranged from 47.9% to 52.1% (TABLE I).

TABLE 1. AVERAGE CAPTURE TEMPERATURES AND TYPE-A COMPONENTS

ROI	Phase	Average (°C)	SD (°C)	$u_A(\bar{T})$ (°C)	Type A Share (%)
TMJ R	Pre	34,10	0,071	0,032	5,4
	Post	33,88	0,295	0,132	94,6
TMJ L	Pre	34,24	0,241	0,108	95,1
	Post	33,74	0,055	0,024	4,9
Temporal R	Pre	34,38	0,259	0,116	50,8
	Post	34,10	0,255	0,114	49,2
Temporal L	Pre	34,50	0,158	0,071	52,1
	Post	34,24	0,152	0,068	47,9

n = 5 per ROI and time point. SD: sample standard deviation. $u_A(\bar{T})$: Type A standard uncertainty of the average. Type A share: contribution of each phase to the estimated variance of the pre-post difference, according to Eq. (2). R: right; L: left
Source: Author

The standard Type A uncertainties of the differences ranged from 0.098 to 0.162 °C, and the expanded uncertainties from 0.23 to 0.37 °C. In the adopted model, the ranges of the left ROIs remained positive; those of the right regions included zero (TABLE II).

In the clinical evaluation, the pain score on the VAS went from 8 to 0 and the mouth opening increased from 37 to 40 mm, a difference of 3 mm.

TABLE II. PRE-POST DIFFERENCES AND TYPE A UNCERTAINTIES

ROI	ΔT(°C)	uA(ΔT)(°C)	U95,A(°C)	Nominal range of 95% (°C)
TMJ R	0,22	0,136	0,36	[−0,14; 0,58]
TMJ L	0,50	0,110	0,30	[0,20; 0,80]
Temporal R	0,28	0,162	0,37	[−0,09; 0,65]
Temporal L	0,26	0,098	0,23	[0,03; 0,49]

U95,A = k × uA(ΔT). Effective degrees of freedom ek, respectively: ATM D, 4.46 and 2.67; ATM E, 4.41 and 2.68; Temporal D, 8.00 and 2.31; Temporal E, 7.99 and 2.31. Individual intervals restricted to the Type A model, under the assumptions of Section II-C.
Source: Author

IV. DISCUSSION

Monitoring documented a reduction in averages across all four regions, with the greatest decrease in the left TMJ and similar magnitudes in the temporal bones. The description by structure locates the changes in the orofacial pattern: in the TMJs, there was a reversal of the side with the higher average; in the temporal bones, the relationship between the sides was approximately preserved. These aspects maintain the anatomical information of the findings within the context of sequelae of mastoid fracture.

The combined presentation of average, dispersions, and uncertainties situates each change in relation to the variability of catches. In the Type A model, the reductions in left-hand ROIs were more clearly defined, with entirely positive intervals. In the right-hand regions, the intervals express greater uncertainty regarding the magnitude and sign of the difference. Thus, the regional interpretation simultaneously considers the observed average reduction and the uncertainty of its estimate.

The balance quantifies the contribution of pre- and post-processing blocks to the estimated variance of the difference. In ATM D, the post-processing includes one reading of 34.4 °C and four between 33.7 and 33.8 °C, all maintained. In ATM E, the pre-processing component predominates, while the temporal components show balanced proportions. Identifying the blocks with the greatest influence on uncertainty guides the documentation of acquisition times and the selection of reading points in future applications. The components describe the variability between captures, without separating instrumental and physiological causes.

The reduction in reported pain and the increase in mouth opening provide clinical context to the thermal records. The literature on TI in TMD examines patterns dependent on the condition and region evaluated [2]. In this study, individual follow-up brings together, in the same session, anatomical information, temperature variation, and clinical measurements. Its contribution is to present this immediate evolution in a quantitative way, maintaining the individuality of the case and the combined nature of the intervention.

The findings describe one participant and one multimodal session, without control or follow-up. Lateral differences are descriptive. The independence and stability of the captures are calculated hypotheses: the timing of the readings, the contours of the ROIs, and the cursor permanence are not

detailed in the described protocol. A positive temporal dependence may increase the uncertainty of the average. Calibration, drift, radiometric adjustments, and repositioning were not quantified separately; U95,A expresses the component evaluated from the captures. Subsequent studies should document these factors and expand the uncertainty balance to assess the consistency of monitoring between sessions.

V. CONCLUSION

In the case study, monitoring documented reductions in average point temperatures of 0.22 to 0.50 °C, with a greater drop in the left atmospheric pressure zone and similar magnitudes during storms. The bilateral description and the Type A uncertainty balance characterized the changes and the contribution of each phase to its uncertainty. Under the adopted assumptions, the standard uncertainties of the differences ranged from 0.10 to 0.16 °C.

Thermal recording added regional information to observations of pain and mouth opening. Continued use in larger samples, with repeated sessions, follow-up, controlled comparisons, and evaluation of isolated techniques, will allow investigation of the relationship with standardized clinical outcomes and the applicability of orofacial monitoring.

RECOGNITION

We thank CNPq, CAPES, and FAPERJ for their financial support.

REFERENCES

[1] PCR Conti, TMD: temporomandibular disorders and orofacial pain — clinical application of scientific evidence. Maringá: Dental Press, 2021.

[2] DS Haddad et al., “Is it possible myogenic temporomandibular dysfunctions change the facial thermal imaging?”, Clinical and Laboratory Research in Dentistry, 2019. DOI: 10.11606/issn.2357-8041.clrd.2019.158306.

[3] C. Tsagkaris et al., “Infrared Radiation in the Management of Musculoskeletal Conditions and Chronic Pain: A Systematic Review,” European Journal of Investigation in Health, Psychology and Education, vol. 12, no. 3, pp. 334–343, 2022. DOI: 10.3390/ejihpe12030024.

[4] ACR Silva, SF Raymundo and WCA Pereira, “Analysis of the Temperature Pattern of the Temporomandibular Joint and Previous Temporal Muscle as a Function of Acquisition Distance Using Infrared Thermography”, GMEPE/PAHCE, Rio de Janeiro, 2026, pp. 1–4. DOI: 10.1109/GMEPE/PAHCE69420.2026.11550301.

[5] JCGM, Evaluation of measurement data — Guide to the expression of uncertainty in measurement. JCGM 100:2008. BIPM, 2008. Sections 4.2, 5.1 and 5.2; Appendix G. Available at: https://www.bipm.org/en/committees/jc/jcgm/publications.

[6] NIST, “Evaluating uncertainty components: Type A”, The NIST Reference on Constants, Units, and Uncertainty. Available at: https://physics.nist.gov/cuu/Uncertainty/typea.html. Accessed on: 10 September. 2026.

Respiratory System Impedance Volume and Flow Loops Areas: Association With Spirometry and Diagnostic Use in Chronic Obstructive Pulmonary Disease

1st Eduarda Martins de Faria
Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
martiinseduarda09@gmail.com

2nd Cíntia Moraes de Sá Sousa
Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
cintiamdesa@gmail.com

3rd Maria Luiza do Nascimento Montenegro
Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
montenegro.marial121@gmail.com

4th Caroline de Oliveira Ribeiro
Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
caroline.oliveira@live.com

5th Agnaldo José Lopes
Pedro Ernesto University Hospital
State University of Rio de Janeiro
Rio de Janeiro, Brazil
agnaldolopes.uerj@gmail.com

6th Pedro Lopes de Melo
Biomedical Instrumentation Laboratory
State University of Rio de Janeiro
Rio de Janeiro, Brazil
plopesuerj@gmail.com

***Abstract*— The assessment of respiratory mechanics using oscillometry-derived parameters has gained increasing interest as a complementary tool to spirometry in chronic obstructive pulmonary disease (COPD). In particular, loop-derived parameters based on resistance and reactance may provide additional insights into intrabreath mechanical behavior. Therefore, the present study aimed to evaluate the areas of respiratory system impedance, volume, and flow loops and investigate their association with spirometric parameters, as well as their potential diagnostic value in patients with COPD. This study analyzed loop-derived parameters obtained from respiratory oscillometry, including area of resistance-volume (ARV), resistance-flow (ARV′), reactance-volume (AXV), and reactance-flow (AXV′) loops. Correlation analyses were performed to assess associations with spirometric parameters. Diagnostic accuracy was evaluated using receiver operating characteristic (ROC) curve analysis. Reactance-derived parameters were significantly higher in COPD patients ($p<0.05$) and showed moderate diagnostic accuracy (AUC=0.751), as well as significant moderate to excellent correlations with spirometric variables (R=0.830). In contrast, resistance-derived parameters did not show significant correlations and demonstrated lower diagnostic performance (p=0.888). From this, reactance-based loop parameters showed stronger associations with spirometric impairment and higher discriminatory performance in the studied sample and may provide additional information on alterations in respiratory mechanics compared to resistance-derived measures. These findings suggest a potential role of oscillometry, particularly reactance-derived indices, as a complementary tool to spirometry in the evaluation of COPD.**

Keywords—COPD, oscillometry, reactance, resistance, loop analysis.

I. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is among the three leading causes of death worldwide, with the majority of deaths occurring in low- and middle-income countries [1, 2]. It is characterized by persistent respiratory symptoms, as well as significant airflow restriction due to abnormalities in the airways and alveoli. These characteristics are usually caused by exposure to harmful gases and particles, with smoking being the main risk factor for developing the disease. It is a preventable and treatable disease, and the most commonly reported symptoms are dyspnea, cough, and discomfort [3].

Airflow limitation is usually measured by spirometry [3]. This pulmonary function test is considered the gold standard, but requires great effort and cooperation from individuals, which can temporarily cause changes in bronchomotor tone, consequently varying the quality of the test results [4, 5].

Respiratory oscillometry involves applying pressure oscillations to the mouth using an external device, allowing the patient to breathe spontaneously while the resistive and elastic properties of the respiratory system are analyzed. The examination requires only passive cooperation from the patient, and forced expiratory maneuvers are not necessary [6]. Thus, it is a method that emerges to complement traditional approaches to assessing pulmonary function, offering additional parameters for the analysis of ventilatory mechanics. Through this technique, it is possible to measure the impedance of the respiratory system (Zrs), composed of two main elements: resistance (Rrs) and reactance (Xrs) [6].

Zrs expresses the total mechanical load imposed on the respiratory system during ventilation, encompassing resistive, elastic, and inertial components of the respiratory system [7, 8]. Rrs reflects the total energy dissipation in the respiratory system, encompassing not only airway resistance but also the properties of lung tissues and the chest wall. Thus, it represents the total respiratory resistance, capturing both the friction related to airflow and the tissue resistive factors. In turn, Xrs expresses the accumulation of energy in the respiratory system [6].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Investigating the dynamic behavior of respiratory impedance throughout the breathing cycle may provide additional physiological information that is not captured by conventional time-averaged oscillometric parameters [9]. Since mechanical properties of the respiratory system vary between inspiration and expiration, particularly in conditions associated with airflow limitation such as COPD [9], the analysis of selected breathing cycles may reveal subtle changes in airway mechanics and ventilatory heterogeneity.

Respiratory system behavior throughout the breathing cycle can be evaluated using loop representations derived from physiological signals, such as flow-volume loops. The flow-volume loop represents the relationship between airflow and lung volume, with inspiratory and expiratory flows displayed on the y-axis and the corresponding lung volumes on the x-axis. Different patterns observed in this curve are associated with obstructive and restrictive respiratory conditions and may assist in their identification and localization [10].

Despite the well-established role of spirometry in the assessment of COPD, its limitations, especially the need for effort and cooperation from the patient, drive the search for alternative or complementary methods. In this context, respiratory oscillometry emerges as a promising technique, allowing the assessment of respiratory mechanics during spontaneous breathing.

Although conventional oscillometric parameters have been widely studied, the analysis of intrabreath dynamics, particularly through loop-derived parameters, remains relatively underexplored. These loop representations may provide additional insights into the nonlinear and dynamic behavior of the respiratory system, especially in conditions such as COPD, where expiratory flow limitation and small airway dysfunction play a central role.

These types of loop-derived parameters have been previously explored in the literature, particularly for the assessment of dynamic changes in respiratory mechanics in different conditions [11-13]. However, the clinical applicability and diagnostic value of these loop-derived parameters are not yet fully established, and studies addressing this topic are limited.

Thus, this preliminary and exploratory study aimed to evaluate the areas of respiratory system impedance, volume, and flow loops and investigate their association with spirometric parameters, as well as their potential diagnostic value in patients with COPD.

II. Materials and Methods

A. Monofrequency respiratory oscillometry

In monofrequency respiratory oscillometry, the assessment was performed by applying pressure oscillations at a single frequency. These analyses were used to evaluate Zrs in real time, enabling the detection of variations in airway caliber during the respiratory cycle [6].

The impedance analyzer was developed in our Laboratory, and followed the most recent international recommendations [6]. A sinusoidal pressure signal at a frequency of 6 Hz *was* applied, generated by a power amplifier coupled to an 8-inch speaker, producing a peak-to-peak pressure amplitude probe signal of 2.0 cmH_2O. Pressure variations (P) were captured by a low-pressure transducer (Honeywell 176PC) [14].

The examination was performed in a seated position, with a nose clip, with the individual breathing normally through a disposable mouthpiece with cheek compression by the individual themselves with their palms. The hands holding the cheeks prevent the upper airways from interfering with the respiratory impedance measurement, ensuring that the examination reflects only the behavior of the intrathoracic respiratory system. During the procedure, participants breathed through a mouthpiece connected to a flow (V′) transducer (PT-36 Jaeger-Tonnies). These signals obtained were processed using Fourier Transform (F) to calculate the total respiratory Zrs [Zrs = F(P)/F(V′)] [14], allowing the measurement of the resistance and reactance at 6 Hz.

The software for data acquisition and Zrs processing was developed on the LabVIEW 2018 platform (National Instruments, Austin, TX). We used a Dell Inspiron 137000 series notebook as the hardware platform, equipped with a touch screen, Intel Core i5 processor, 4 GB of RAM and a 456 GB hard drive [14]. Zrs estimates were obtained at 20 Hz.

Each examination lasted a minimum of 35 s and a maximum of 150 s. Exams containing artifacts or irregular breathing were identified using visual inspection and the criteria described in [15] and excluded from analysis.

At least two acceptable examinations were performed. During the maneuvers, graphs of the resistance and reactance curves were provided, with a signal indicating a Coherence Function (CF) > 0.90 as the criterion for accepting the curve. From the examination performed, 5 stable respiratory cycles were selected to perform the average analysis of the Zrs Loops.

The following loop-derived parameters were analyzed: area of resistance-volume loop (ARV), area of resistance-flow loop (ARV′), area of reactance-volume loop (AXV), and area of reactance-flow loop (AXV′). These parameters were obtained from the integration between resistance (R6) and reactance (X6) at 6 Hz with volume (V) and flow (V′) throughout the respiratory cycle, reflecting intrabreath variations in respiratory mechanics [11].

B. Study population

This case-control study included individuals diagnosed with COPD and healthy control subjects, and was approved by the Research Ethics Committee registered under number 87550225.0.0000.5259.

Selected participants were referred to the Biomedical Instrumentation Laboratory of the State University of Rio de Janeiro, provided they met the eligibility criteria and signed the Informed Consent Form. In addition, volunteers were informed of the need to discontinue the use of cigarettes, alcoholic beverages, short-acting bronchodilators for 6 hours and long-acting bronchodilators for 12 hours prior to the pulmonary function test.

Upon arrival at the Laboratory, they underwent a 10-15 minute rest period for the administration of a standard questionnaire, clinical evaluation, and anthropometric measurements. Following this, respiratory oscillometry and spirometry tests were performed. After this stage, salbutamol aerosol was administered to assess the bronchodilator response, and after 20 minutes, the pulmonary assessment was performed

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

again. For the analyses presented in this study, only post-bronchodilator oscillometric and spirometric measurements were considered.

The control group consisted of individuals ≥ 18 years of age, with no history of smoking, respiratory or cardiac diseases, no history of respiratory infections in the last thirty days, and with spirometry and oscillometry tests compatible with normality. In the COPD group, individuals ≥ 18 years of age were included, with no history of respiratory infections in the last thirty days and with a previous diagnosis of COPD based on the criteria of the Global Initiative for Chronic Obstructive Lung Disease (GOLD), where the FEV_1/FVC ratio is < 0.7 after the use of a bronchodilator [3]. Individuals with a history of mild Coronavirus Disease 2019 (COVID-19) were eligible for both groups, provided they did not have persistent respiratory symptoms and fully met the inclusion criteria established for each group.

C. Spirometry

A computerized system equipped with a pneumotachograph (nSpire Health, Inc., Longmont, CO, USA) was used to perform spirometry. The procedure followed the acceptability and reproducibility criteria established by the Brazilian Thoracic Society (SBPT) and the ATS/ERS [16-18]. For data analysis, reference equations proposed by Pereira were applied for spirometry [19].

Participants were evaluated in a seated position, with the head in a neutral position, breathing through a mouthpiece with a bacterial filter and using a nose clip to occlude the nostrils.

They were instructed to breathe normally for 3 seconds, followed by a maximal inspiration and a forced expiration until the end of expiration. In cases of severe obstruction, forced expiration may last longer than 15 seconds. Therefore, although 6 seconds is the minimum recommended duration, curves lasting longer than 10 seconds are still considered acceptable in the presence of significant obstruction [20].

After obtaining three acceptable maneuvers, with at least two being reproducible, the parameters were analyzed based on the flow–volume and volume–time curves.

The spirometric parameters analyzed in the present study included the forced vital capacity (FVC) that represents total volume of air exhaled during a forced expiration initiated from a maximal inspiration; Forced Expiratory Volume in one second (FEV_1), which represents the volume of air exhaled in the first second of the forced expiratory maneuver; FEV_1/FVC corresponds to the ratio between the volume exhaled in the first second and the forced vital capacity; Forced Expiratory Flow between 25% and 75% ($FEF_{25–75}$) represents the mean forced expiratory flow obtained during the intermediate phase of the forced vital capacity; $FEF_{25–75}/FVC$ corresponds to the ratio between the mean forced expiratory flow during the intermediate portion of the forced vital capacity and the forced vital capacity itself, expressed as absolute or percentage values [20].

D. Statistical analysis

Results were expressed as mean ± standard deviation (SD) with descriptions in tables and graphs. The sample distribution characteristics were evaluated using the Shapiro-Wilk test. For normally distributed data (parametric), the independent t-test was applied. For non-normally distributed data (non-parametric), the Mann-Whitney test was used.

Pearson (parametric method) and Spearman (nonparametric method) correlation analyses were also performed to assess the associations between Zrs loops and pulmonary function parameters. These analyses were conducted using GraphPad Prism 8 (GraphPad Software, La Jolla, CA, USA). Correlation coefficients ranging from 0 to 0.25 (or 0 to −0.25) were considered small or negligible; from 0.25 to 0.50 (or −0.25 to −0.50), fair; from 0.50 to 0.75 (or −0.50 to −0.75), moderate to good; and values exceeding 0.75 (or −0.75) were considered very good to excellent [21].

Diagnostic accuracy, sensitivity, specificity, and cutoff values were evaluated using receiver operating characteristic (ROC) curve analysis. Diagnostic accuracy was assessed by calculating the area under the curve (AUC), where values between 0.50 and 0.70 indicate low accuracy, 0.70 to 0.90 indicate moderate accuracy, and values between 0.90 and 1.00 indicate high accuracy [22].

According to Goedhart et al. [23], a cutoff value of 0.70 was considered appropriate to characterize a clinically useful discriminator. In the present study, a minimum AUC value of 0.75 was adopted as the criterion for adequate diagnostic accuracy. ROC analyses were performed using MedCalc 12 (MedCalc Software, Mariakerke, Belgium).

Differences with $p \leq 0.05$ were considered statistically significant, and the analyses and graphs were performed using the Origin® 8.0 program (Microcal Software Inc., Northampton, Massachusetts, United States).

III. RESULTS

A. Studied volunteers and spirometric evaluations

In total, 30 individuals were included, 15 in the control group and 15 in the COPD group (4 with mild, 3 with moderate, 4 with severe, and 4 with very severe obstruction). The biometric and spirometric parameters of the studied groups are presented in Table 1.

Table 1: Biometric and spirometric parameters of the studied groups.

	Control (n=15)	COPD (n=15)	p
Age (years)	54.86 ± 10.63	66.86 ± 3.88	<0.001
Weight (kg)	79.33 ± 10.04	72.46 ± 9.49	ns
Height (cm)	169.50 ± 10.76	165.33 ± 6.56	ns
BMI (kg/m^2)	27.74 ± 3.78	26.55 ± 3.61	ns
Pack-years	0	77.65 ± 38.70	<0.0001
Gender (F/M)	*7 / 8*	*3 / 12*	-
R6 ($cmH_2O/L/s$)	3.06 ± 1.15	4.62 ± 1.67	0.00666
FVC (L)	3.93 ± 1.02	2.64 ± 1.02	0.00184
FVC (%)	97.06 ± 13.61	69.29 ± 22.88	0.00369
FEV_1 (L)	3.22 ± 0.84	1.56 ± 0.80	<0.0001
FEV_1 (%)	99.47 ± 11.75	52.62 ± 24.73	<0.0001
FEV_1/FVC (%)	81.96 ± 5.30	57.06 ± 10.81	<0.0001

COPD – Chronic Obstructive Pulmonary Disease; BMI - Body Mass Index; FEV_1 - Forced Expiratory Volume in one second; FVC - Forced Vital Capacity; n= number of patients evaluated; ns - nonsignificant.

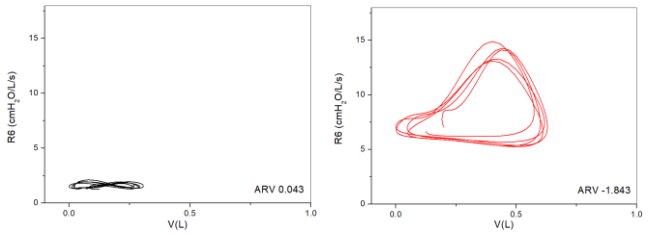
XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Regarding biometric parameters, no statistically significant differences were observed between the groups in terms of body weight, height, and BMI. However, age showed a significant difference ($p<0.001$), with the COPD group being characterized by older individuals.

In addition, the smoking history was significantly higher in the COPD group when compared to the control group ($p<0.0001$). The mean resistance at 6 Hz (R6) was significantly higher in the COPD group compared to the control group ($p<0.01$). All spirometric parameters showed significant differences between the groups, indicating significantly reduced lung function in the COPD group compared to the control group (Table 1).

B. Respiratory System Impedance Volume and Flow Loops Areas

To illustrate the behavior of intrabreath oscillometric loops, representative examples from a healthy control and a patient with COPD are presented. In the control subject, the resistance–volume loops exhibit a compact and nearly closed configuration, with minimal separation between inspiratory and expiratory phases, with limited intrabreath variability in resistance (Figure 1A). In contrast, the COPD patient demonstrates markedly expanded and distorted loops, with a clear separation between inspiratory and expiratory trajectories, resulting in an open-loop configuration. This pattern indicates increased intrabreath variability in resistance during the respiratory cycle (Figure 1B).



(A) **(B)**

Fig. 1 Representative resistance–volume loops in a control subject (A) and in a COPD subject (B).
R6 – Resistance *at* 6 Hz; V – Volume; ARV - Area of Resistance Volume Loop.

Although the COPD group showed a visual tendency toward lower mean values (Figure 2), there was no significant difference in ARV between the groups.

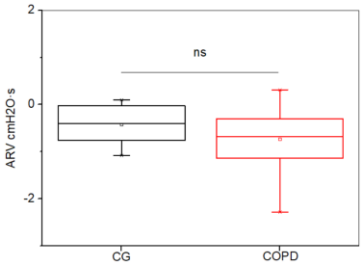
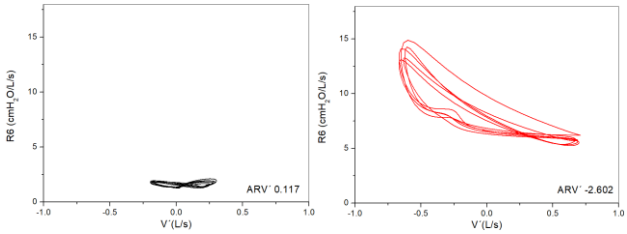


Fig. 2 Comparison of ARV between control group and COPD patients.
ARV - Area of Resistance Volume Loop; CG – Control Group; COPD – Chronic Obstructive Pulmonary Disease; ns – nonsignificant.

Consistent with the patterns observed in resistance-volume loops, the control subject represented in resistance-flow loops also presented a more compact form, with minimal separation between the phases of the respiratory cycle (Figure 3A).

In contrast, this pattern becomes more pronounced in COPD, where the patient exhibits broader and more dispersed loop patterns, with a clear separation between inspiratory and expiratory phases (Figure 3B).



(A) **(B)**

Fig. 3 Representative resistance–flow loops in a control subject (A) and in a COPD subject (B).
R6 – Resistance at 6 Hz; V′ – Flow; ARV′ - Area of Resistance Flow Loop.

Although the COPD group also exhibited a visual tendency toward lower mean values compared to the control group (Figure 4), no statistically significant difference was observed in ARV′ between the groups.

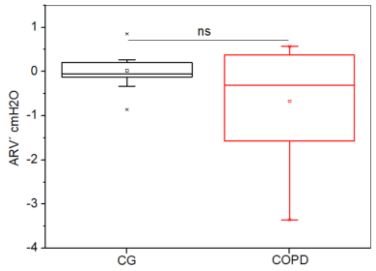
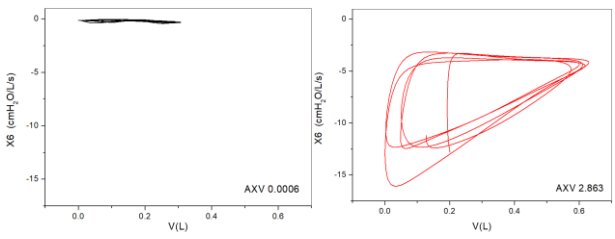


Fig. 4 Comparison of ARV′ between control group and COPD patients.
ARV′ - Area of Resistance Flow Loop; CG – Control Group; COPD – Chronic Obstructive Pulmonary Disease; ns – nonsignificant.

In reactance–volume loops, the control subject exhibited a narrow configuration closely aligned near zero values, indicating minimal variation throughout the respiratory cycle (Figure 5A). Conversely, in COPD, a markedly expanded loop with more negative reactance values and greater separation between inspiratory and expiratory phases was observed (Figure 5B).



(A) **(B)**

Fig. 5 Representative reactance–volume loops in a control subject (A) and in a COPD subject (B).
X6 – Reactance at 6 Hz; V – Volume; AXV - Area of Reactance Volume Loop.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

While resistance-based loops showed limited differences between groups, a distinct pattern emerges when reactance is analyzed. AXV was significantly higher in patients with COPD compared to the CG (Figure 6, $p < 0.01$).

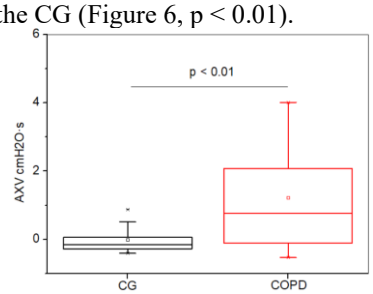
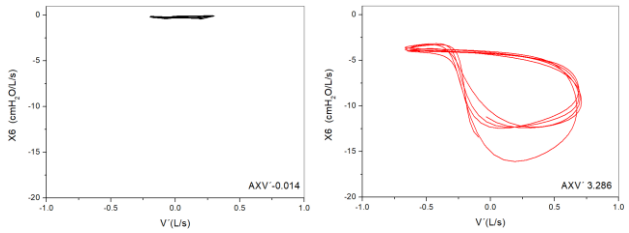


Fig. 6 Comparison of AXV between control group and COPD patients. AXV - Area of Reactance Volume Loop; CG – Control Group; COPD – Chronic Obstructive Pulmonary Disease.

Extending this observation to flow-dependent behavior, reactance–flow loops provide additional insight into the dynamic aspects of respiratory mechanics. In the control subject, the reactance–flow loop remained compact, with a nearly linear trajectory and minimal phase separation between inspiratory and expiratory branches (Figure 7A). However, in COPD, a markedly distorted loop with evident phase separation between inspiratory and expiratory trajectories was observed (Figure 7B), indicating altered dynamic behavior during the respiratory cycle.



(A) **(B)**

Fig. 7 Representative reactance–flow loops in a control subject (A) and in a COPD subject (B).
X6 – Reactance *at* 6 Hz; V′ – Flow; AXV′ - Area of Reactance Flow Loop.

The comparative analysis revealed that AXV′ was significantly higher in patients with COPD compared to the control group (Figure 8, $p < 0.05$).

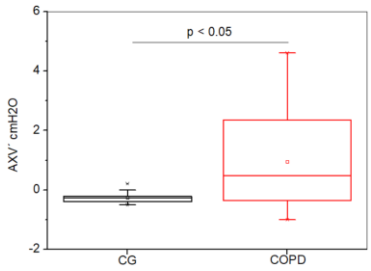


Fig. 8 Comparison of AXV′ between control group and COPD patients. AXV′ - Area of Reactance Flow Loop; CG – Control Group; COPD – Chronic Obstructive Pulmonary Disease.

C. *Correlation Between Impedance Loops and Pulmonary Function Parameters*

Correlation analyses between Zrs loops and spirometry are presented in Table 2. No significant correlations were observed between ARV and any spirometric variables. Similarly, ARV′ did not show significant correlations with spirometry parameters.

In contrast, AXV demonstrated significant moderate to excellent inverse correlations with all spirometric variables. AXV′ also showed significant inverse correlations with most spirometric parameters, except for FEF_{25-75}/FVC %, which showed a significant positive correlation, and for FVC%, which did not show a significant correlation.

D. *Diagnostic Accuracy*

The diagnostic accuracy of Zrs loops is presented in Table 3. Overall, reactance-derived parameters showed higher diagnostic performance in the present sample compared to resistance-derived measures. AXV demonstrated moderate diagnostic accuracy (AUC = 0.791), while AXV′ also showed moderate performance (AUC = 0.751), both with higher sensitivity and specificity.

In contrast, ARV and ARV′ showed lower diagnostic accuracy (AUC = 0.622 and 0.618, respectively), with reduced sensitivity despite high specificity, showing lower discriminatory performance in the present sample.

Table 2: Correlation coefficient (r) and correlation significance level (p) of the oscillometry parameters in relation to the spirometry. Significant correlations are highlighted in bold.

		FVC %	FVC	FEV_1 %	FEV_1	FEV_1/FVC	FEF_{25-75}	FEF_{25-75}/FVC %	FEF_{25-75}/FVC
ARV mean	r	0.305	0.354	0.248	0.297	0.063	0.123	0.090	0.034
	p	0.101	0.055	0.187	0.111	0.741	0.516	0.634	0.860
AXV mean	r	**-0.582**	**-0.830**	**-0.676**	**-0.790**	**-0.534**	**-0.682**	**-0.558**	**-0.541**
	p	**< 0.001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.01**	**< 0.0001**	**< 0.01**	**< 0.01**
ARV′ mean	r	-0.085	-0.027	0.118	0.032	0.134	0.080	0.135	0.107
	p	0.656	0.888	0.534	0.865	0.480	0.675	0.477	0.574
AXV′ mean	r	-0.294	**-0.503**	**-0.460**	**-0.519**	**-0.470**	**-0.527**	**0.504**	**-0.484**
	p	0.114	**< 0.01**	**< 0.05**	**< 0.01**	**< 0.01**	**< 0.01**	**< 0.01**	**< 0.01**

ARV- Area of Resistance Volume Loop; AXV - Area of Reactance Volume Loop; ARV′ - Area of Resistance Flow Loop; AXV′ - Area of Reactance Flow Loop; FEV_1 - Forced Expiratory Volume in one second; FVC - Forced Vital Capacity; FEV_1/FVC - Forced Expiratory Volume in one second to Forced Vital Capacity ratio; FEF - Forced Expiratory Flow.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Table 3: Diagnostic accuracy, sensitivity, specificity, and cutoff point of Zrs loops in patients with COPD. The highest AUC value is shown in bold in the single-frequency oscillometric parameter classes.

	AUC	95% CI	Se (%)	Sp (%)	Cutoff point
Oscillometry					
ARV mean	0.622	0.428-0.792	26.67	100	-1.14
AXV mean	**0.791**	0.604-0.917	66.67	93.33	0.58
ARV′ mean	0.618	0.423-0.788	53.33	86.67	-0.30
AXV′ mean	**0.751**	0.560-0.890	66.67	93.33	-0.004

ARV- Area of Resistance Volume Loop; AXV - Area of Reactance Volume Loop; ARV′- Area of Resistance Flow Loop; AXV′ - Area of Reactance Flow Loop.

IV. Discussion

The results regarding biometric parameters demonstrated that the groups were homogeneous in terms of body weight, height, and BMI (Table 1). However, a significant difference was observed in relation to age, with the COPD group being composed of older individuals when compared to the control group. This finding is consistent with what is already discussed in the literature, where the disease is more prevalent in older individuals, having a higher frequency in people over 65 years of age [24].

The significantly higher tobacco load in the COPD group reinforces the influence of smoking on disease progression (Table 1). Continuous exposure to cigarette smoke triggers a persistent inflammatory process associated with oxidative stress in lung tissue, causing repeated cycles of injury and repair [25]. As a consequence, structural remodeling of the airways occurs, accompanied by stimulation of excessive mucus production, characteristic of chronic bronchitis, in addition to the progressive destruction of alveolar walls, contributing to the development of emphysema [25].

The increased respiratory resistance observed in the COPD group reflects the presence of airflow limitation characteristic of the disease [3] (Table 1). In individuals with COPD, the structural airway alterations already discussed contribute to increased airway resistance [3]. Thus, elevated resistance values obtained by oscillometry are consistent with the obstructive pattern evidenced by the spirometric parameters observed in this study.

With regard to spirometric parameters, the significant reduction in FVC, both in absolute values and as a percentage of predicted, suggests impaired pulmonary ventilation (Table 1). This alteration is mainly due to air trapping, which leads to an increase in residual volume (RV) relative to total lung capacity (TLC), configuring a common pathophysiological characteristic of COPD [26].

The reduction in FEV_1, in liters and as a percentage of predicted, confirms the presence of airflow obstruction (Table 1). This alteration may be mainly related to the impairment of the small airways, resulting from inflammation, excessive mucus production, and structural remodeling [27]. These processes promote narrowing of the bronchial lumen, causing a significant increase in airway resistance and limitation of expiratory airflow, which directly contributes to the decrease in FEV_1 observed in these individuals [27].

Furthermore, the significant reduction in the FEV_1/FVC ratio reinforces this obstructive pattern, being compatible with the diagnostic criteria used for the disease [3] (Table 1).

Intrabreath oscillometry allows continuous monitoring of lung mechanical properties during both inspiratory and expiratory phases, which are graphically represented as loops describing the respiratory cycle [12] (Fig. 1; Fig. 3; Fig. 5; Fig. 7).

Resistance primarily reflects viscous losses in the respiratory system [28], as well as representing opposition to airflow through the airways [29]. The absence of significant differences in both ARV and ARV′ between the control group and COPD patients suggests lower discriminatory performance of resistance-derived curve areas in the present sample (Fig. 2; Fig. 4).

A similar finding has been reported by Fu et al., who evaluated intrabreath oscillometry parameters in different clinical groups and analyzed resistance-derived loop areas, including ARV and ARV′. In their study, ARV did not differ significantly between groups, whereas ARV′ was significantly increased in patients with bronchiolitis obliterans syndrome compared to the other individuals [12].

In the study by Cousins et al., intrabreath oscillometry in school-aged children with prematurity-associated lung disease was investigated. These individuals underwent oscillometry at the beginning of the study, after the maximal cardiopulmonary exercise test, and after post-exercise bronchodilation. No significant differences were observed in either ARV or ARV′ between groups at baseline [13]. These findings are consistent with the results of the present study, in which no significant differences were observed for ARV and ARV′. However, their analysis was conducted in a pediatric population with distinct underlying pathophysiology related to prematurity, which differs from the adult COPD population evaluated in the present study.

As previously discussed, the relationships between Rrs and Xrs with V′ and V exhibit different loop patterns, which can be used to characterize pathological and physiological processes [11].

The reactance-based loop areas (AXV and AXV′) were significantly higher in patients with COPD, consistent with alterations in respiratory system mechanics (Fig. 6; Fig. 8). AXV has been associated with expiratory flow limitation [11], while AXV′ has been related to air trapping during tidal breathing [12]. These features are characteristic of COPD and may explain the significantly higher values observed in affected patients [3].

Similarly, Fu et al. reported that AXV was significantly increased in patients with bronchiolitis obliterans syndrome compared to other groups, while AXV′ values were significantly more negative in these patients [12].

Overall, these findings suggest that reactance-based loop parameters showed more pronounced differences in COPD

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

patients, particularly when compared to resistance-derived measures.

The correlation analysis demonstrated that reactance-derived parameters, particularly AXV, showed moderate to strong correlations with spirometric variables, whereas resistance-derived parameters did not show significant associations (Table 2).

The negative correlations observed between AXV and AXV′ and spirometric parameters suggest that greater alterations in reactance are associated with worse lung function. This finding is particularly relevant in COPD, where airflow limitation and small airway involvement play an important role in the pathophysiology of the disease [3].

Additionally, AXV′ showed significant correlations with most spirometric variables, although with lower magnitude compared to AXV. This finding suggests that volume-based loops may be more closely associated with the functional alterations captured by spirometry in this sample, whereas flow-based loops may be influenced by additional factors related to the breathing pattern.

In contrast, the absence of significant correlations for ARV and ARV′ may suggest a limited ability of resistance-derived parameters to reflect functional impairment. This may be explained by the fact that resistance primarily represents opposition to airflow [29], whereas reactance is related to the elastic and inertial properties of the respiratory system and has been associated with peripheral airway function and dynamic changes throughout the respiratory cycle [27].

The diagnostic performance observed for impedance area loops in the present study suggests that reactance-derived measures showed higher discriminatory performance in COPD when compared to resistance-derived parameters. AXV and AXV′ demonstrated moderate diagnostic accuracy, with higher sensitivity and specificity, indicating their potential to capture disease-related changes in respiratory mechanics (Table 3).

These findings are consistent with the correlation analysis (Table 2), in which reactance-based parameters, particularly AXV, showed significant associations with spirometric variables, suggesting that these measures may reflect functional impairment. In contrast, resistance-derived parameters showed both lower diagnostic accuracy and absence of significant correlations with spirometry, reinforcing their limited discriminatory capacity in this population.

The lower performance of ARV mean and ARV′ mean may be partly related to the physiological characteristics of resistance, which mainly represents opposition to airflow [29], and may be less sensitive to more subtle changes, especially those related to peripheral airways. Conversely, as previously discussed, reactance reflects the elastic and inertial properties of the respiratory system and has been linked to alterations in the peripheral airways, which play a key role in the pathophysiology of COPD [27].

Taken together, these results suggest that reactance-derived loop parameters may provide complementary information regarding functional impairment in COPD, while resistance-derived measures showed lower discriminatory performance.

The present study presents several limitations. First, the small sample size; consequently, the cutoff values derived from the ROC curve should be considered exploratory and require validation in independent samples or through internal resampling. Studies involving larger cohorts and adequate statistical power are needed to confirm these findings, evaluate these parameters longitudinally, and investigate their applicability across different stages of COPD, based on GOLD classification, and analysis in other respiratory diseases. Integrating parameters derived from the loops with machine learning techniques could also enhance their discriminatory and clinical potential.

V. Conclusions

The results of the present study demonstrated that reactance-derived parameters, specifically AXV and AXV′, showed stronger associations with functional impairment and moderate discriminatory performance in the present sample compared to resistance-derived parameters.

These findings suggest that reactance-based parameters may reflect aspects of the functional alterations associated with COPD, particularly those related to airflow limitation and small airway involvement. Although spirometry remains the gold standard for the diagnosis of the disease, the analysis of oscillometry-derived loops, especially those based on reactance, may provide relevant complementary information for the assessment of pulmonary function.

Therefore, the use of these parameters may represent a promising tool for the detection and characterization of respiratory alterations in COPD, contributing to a more comprehensive evaluation of respiratory mechanics.

Acknowledgment

The authors would like to thank CNPq, FAPERJ and CAPES for their financial support.

References

1 Halpin, D., Celli, B., Criner, G., Frith, P., Varela, M.L., Salvi, S., Vogelmeier, C., Chen, R., Mortimer, K., and De Oca, M.M.: ‘The GOLD Summit on chronic obstructive pulmonary disease in low-and middle-income countries’, J International Journal of Tuberculosis Lung Disease, 2019, 23, (11), pp. 1131-1141

2 Meghji, J., Mortimer, K., Agusti, A., Allwood, B.W., Asher, I., Bateman, E.D., Bissell, K., Bolton, C.E., Bush, A., and Celli, B.: ‘Improving lung health in low-income and middle-income countries: from challenges to solutions’, J The Lancet, 2021, 397, (10277), pp. 928-940

3 Global Initiative for Chronic Obstructive Lung Disease: ‘Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: 2025 Report’, in Editor (Ed.)^(Eds.): ‘Book Global Strategy for the Diagnosis, Management, and Prevention

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

of Chronic Obstructive Pulmonary Disease: 2025 Report' (Global Initiative for Chronic Obstructive Lung Disease, 2025, edn.), pp.
4 Di Mango, A.M.G., Lopes, A.J., Jansen, J.M., and Melo, P.L.: 'Changes in respiratory mechanics with increasing degrees of airway obstruction in COPD: detection by forced oscillation technique', Respiratory medicine, 2006, 100, (3), pp. 399-410
5 Cavalcanti, J.V., Lopes, A.J., Jansen, J.M., and Melo, P.L.: 'Detection of changes in respiratory mechanics due to increasing degrees of airway obstruction in asthma by the forced oscillation technique', Respiratory medicine, 2006, 100, (12), pp. 2207-2219
6 King, G.G., Bates, J., Berger, K.I., Calverley, P., De Melo, P.L., Dellacà, R.L., Farré, R., Hall, G.L., Ioan, I., and Irvin, C.G.: 'Technical standards for respiratory oscillometry', European Respiratory Journal, 2020, 55, (2), pp. 1-21
7 MacLeod, D., and Birch, M.: 'Respiratory input impedance measurement: forced oscillation methods', J Medical Biological Engineering Computing, 2001, 39, (5), pp. 505-516
8 Miranda, I.A., Dias Faria, A.C., Lopes, A.J., Jansen, J.M., and Lopes de Melo, P.: 'On the respiratory mechanics measured by forced oscillation technique in patients with systemic sclerosis', PloS one, 2013, 8, (4), pp. e61657
9 Kaminsky, D.A., Simpson, S.J., Berger, K.I., Calverley, P., De Melo, P.L., Dandurand, R., Dellacà, R.L., Farah, C.S., Farré, R., and Hall, G.L.: 'Clinical significance and applications of oscillometry', J European respiratory review, 2022, 31, (163), pp. 1-19
10 Ntima, N.O., and Lumb, A.B.: 'Physiology and conduct of pulmonary function tests', BJA education, 2019, 19, (6), pp. 198-204
11 Lorx, A., Czövek, D., Gingl, Z., Makan, G., Radics, B., Bartusek, D., Szigeti, S., Gál, J., Losonczy, G., and Sly, P.D.: 'Airway dynamics in COPD patients by within-breath impedance tracking: effects of continuous positive airway pressure', J European Respiratory Journal, 2017, 49, (2), pp. 1-11
12 Fu, A., Vasileva, A., Hanafi, N., Belousova, N., Wu, J., Rajyam, S.S., Ryan, C.M., Hantos, Z., and Chow, C.W.: 'Characterization of chronic lung allograft dysfunction phenotypes using spectral and intrabreath oscillometry', Frontiers in physiology, 2022, 13, pp. 1-10
13 Cousins, M., Hart, K., Radics, B., Henderson, A.J., Hantos, Z., Sly, P.D., and Kotecha, S.: 'Intra-breath respiratory mechanics of prematurity-associated lung disease phenotypes in school-aged children', J ERJ Open Research, 2025, 11, (2), pp. 1-13
14 Ribeiro, C., Dantas, N., Lopes, A., and Melo, P.: 'Multiscale sample entropy analysis in the evaluation of respiratory changes due to chronic obstructive pulmonary disease. ': 'XXIX Brazilian Congress of Biomedical Engineering; 2024 02-06 September 2024' (2024)
15 Robinson, P.D., Turner, M., Brown, N.J., Salome, C., Berend, N., Marks, G.B., and King, G.G.: 'Procedures to improve the repeatability of forced oscillation measurements in school-aged children', J Respiratory physiology & neurobiology, 2011, 177, (2), pp. 199-206
16 Tisiologia, S.B.d.P.e.: 'Diretrizes para testes de função pulmonar', J pneumol, 2002, 28, (Suppl 3), pp. S1-S238
17 Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., Van Der Grinten, C., and Gustafsson, P.: 'Standardisation of spirometry', J European respiratory journal, 2005, 26, (2), pp. 319-338
18 Pellegrino, R., Viegi, G., Brusasco, V., Crapo, R.O., Burgos, F., Casaburi, R., Coates, A., Van Der Grinten, C., Gustafsson, P., and Hankinson, J.: 'Interpretative strategies for lung function tests', J European respiratory journal, 2005, 26, (5), pp. 948-968
19 Pereira, C.A.d.C., Sato, T., and Rodrigues, S.C.: 'New reference values for forced spirometry in white adults in Brazil', J Jornal brasileiro de pneumologia, 2007, 33, (4), pp. 397-406
20 Pereira, C.A.C.: 'Espirometria', J Pneumol, 2002, 28, (Supl 3), pp. S1-82
21 Dawson, B., and Trapp, R.G.: 'Basic & Clinical Biostatistics' (Lange Medical Books/McGraw-Hill, 2001, 3rd edn. 2001)
22 Swets, J.A.: 'Measuring the accuracy of diagnostic systems', J Science, 1988, 240, (4857), pp. 1285-1293
23 Goedhart, D.M., Zanen, P., Kerstjens, H.A., and Lammers, J.-W.J.: 'Discriminating asthma and COPD based on bronchodilator data: an improvement of the methods', J Physiological measurement, 2005, 26, (6), pp. 1115-1123
24 Liu, Y., Carlson, S.A., Watson, K.B., Xu, F., and Greenlund, K.J.: 'Trends in the Prevalence of Chronic Obstructive Pulmonary Disease Among Adults Aged ≥18 Years - United States, 2011-2021', MMWR. Morbidity and mortality weekly report, 2023, 72, (46), pp. 1250-1256
25 Su, X., Wu, W., Zhu, Z., Lin, X., and Zeng, Y.: 'The effects of epithelial-mesenchymal transitions in COPD induced by cigarette smoke: an update', Respiratory research, 2022, 23, (1), pp. 225
26 Macklem, P.T.: 'Therapeutic implications of the pathophysiology of COPD', The European respiratory journal, 2010, 35, (3), pp. 676-680
27 Liu, C., Li, P., Zheng, J., Wang, Y., Wu, W., and Liu, X.: 'Role of necroptosis in airflow limitation in chronic obstructive pulmonary disease: focus on small-airway disease and emphysema', Cell death discovery, 2022, 8, (1), pp. 363
28 Hantos, Z.: 'Intra-breath oscillometry for assessing respiratory outcomes', J Current Opinion in Physiology, 2021, 22, pp. 1-7
29 Assayag, D., and Johannson, K.A.: 'Respiratory oscillometry: the new kid on the lung function block', J BMJ Open Respiratory Research, 2026, 13, (1), pp. 1-2

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Performance of Multiple PPG-Based Physiological Measurements Across Different Skin Phototypes

1st Kelly C. Baioco da Silva
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
kellybaioco.123@gmail.com

2nd Yasmim C. B. de Morais
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
11232100397@alunos.umc.br

3rd Ana L. Gomes dos Santos
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
analelegsantos@gmail.com

4th Silvia Helena Bastos de Paula
Research Center for Health Policies and Practices
Health Institute of Sao Paulo
Sao Paulo, Brazil
silviabastos@isaude.sp.gov.br

5th José Luis Puglisi
Center for Technological Research
University North State
California, EUA
Jose.puglisi@cnsu.edu

6th Daniel Gustavo Goroso
Center for Technological Research
University of Mogi das Cruzes
Mogi das Cruzes, Brazil
danielg@umc.br

***Abstract*— Photoplethysmography-based wearable devices enable the acquisition of multiple physiological measurements, but their performance may vary across physiological metrics and can be influenced by the optical properties of the skin. This study evaluated the performance of physiological measurements obtained with the Samsung Galaxy Watch4 and investigated the magnitude of measurement errors across different skin phototypes. A total of 114 participants classified as Fitzpatrick skin phototypes I–IV were evaluated. Systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) measurements obtained with the smartwatch were compared with those obtained using a sphygmomanometer, whereas peripheral oxygen saturation (SpO2) and HR were compared with measurements obtained using a fingertip pulse oximeter. Associations were assessed using Pearson or Spearman correlation coefficients, agreement was evaluated using Bland–Altman analysis, and absolute errors across phototypes were compared using the Kruskal–Wallis test. SBP and DBP showed strong correlations with the comparison device (r = 0.974 and r = 0.980, respectively), small mean biases (+1.39 and +0.39 mmHg), and no significant proportional bias. HR showed strong associations with both comparison devices, whereas SpO2 showed a weak association (Spearman correlation coefficient = 0.192) and significant proportional bias. No statistically significant differences in absolute errors were detected across phototypes I–IV for any of the evaluated metrics (p > 0.05), and no systematic gradient of increasing error was observed across phototypes. Under the conditions evaluated, differences in smartwatch performance were more evident across physiological metrics than across the skin phototypes represented in the sample.**

***Keywords*- photoplethysmography, wearable devices, smartwatch, skin phototype, blood pressure, heart rate, peripheral oxygen saturation.**

I. INTRODUCTION

Wearable devices have become increasingly important tools for the continuous monitoring of physiological parameters, extending health-related data collections beyond traditional clinical settings. The integration of miniaturized sensors with signal-processing algorithms enables the noninvasive, real-time measurement of a range of physiological metrics, supporting applications in remote monitoring, longitudinal assessment, and digital health. As these technologies become increasingly integrated into health monitoring, understanding the factors that may affect measurement performance across populations with diverse individual characteristics becomes equally important [1], [2], [3].

Among the sensing technologies used in wearable devices, photoplethysmography (PPG) plays a prominent role. PPG is an optical technique based on the emission of light into tissue and the detection of signal changes associated with pulsatile variations in peripheral blood volume. This principle enables the derivation of heart rate (HR) and peripheral oxygen saturation (SpO_2) and provides signals that can be incorporated into computational models to estimate other cardiovascular parameters, including systolic blood pressure (SBP) and diastolic blood pressure (DBP). Although multiple physiological metrics may be derived from PPG signals, their estimation relies on distinct physiological principles, optical configurations, and signal-processing steps. Therefore, the performance observed for one metric should not be assumed to apply to others [2], [3], [4].

PPG signal quality depends on the interaction between the light emitted by the sensor, biological tissues, and the pulsatile blood component. This interaction can be influenced by factors such as motion, peripheral perfusion, temperature, contact pressure, sensor positioning, anatomical characteristics, and skin properties. In this context, skin pigmentation has received increasing attention because melanin contributes to light absorption and may alter the amount of optical energy that reaches the underlying tissues and returns to the photodetector. Changes in light absorption and scattering may affect the signal-to-noise ratio and, consequently, the quality of the optical signal processed by device algorithms [2], [4], [5].

This issue has been particularly evident in pulse oximetry. Previous studies have reported differences in pulse oximeter performance among individuals with different levels of skin pigmentation, particularly under conditions of low oxygen saturation. However, the influence of skin characteristics on optical devices should not be assumed to be uniform across physiological metrics. Heart rate, SpO_2, and blood pressure estimates rely on different signal features and processing stages and may therefore respond differently to variations in PPG signal quality [2], [3], [4], [5], [6], [7].

Skin phototype diversity is an important consideration when evaluating PPG-based technologies, as light–tissue interactions may vary among individuals. In this context, the Fitzpatrick scale is a widely used clinical approach for classifying skin phototypes, allowing participants to be stratified according to their skin response to sun exposure. Its application in studies of optical devices makes it possible to investigate whether measurement performance differs across phototypes, thereby contributing to the evaluation of these technologies in more diverse populations. In the present study, the Fitzpatrick scale was used to classify participants into skin phototype groups [8], [9], [10].

Despite the growing body of research on the influence of skin characteristics on optical technologies, studies evaluating multiple physiological metrics within the same population, using the same wearable device, across participants with different skin phototypes remain limited. This approach is particularly relevant because potential differences between phototype groups may not be consistent across the physiological metrics provided by the device.

To address this gap, we hypothesized that the association, agreement, and magnitude of differences between measurements obtained with the smartwatch and the comparison devices would vary across physiological metrics and skin phototypes.

This research was supported by the São Paulo Research Foundation (FAPESP), Brazil, Grant No. 21/14221-0. Kelly C. Baioco da Silva was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil, Grant No. 88887.941749/2024-00.

II. OBJECTIVE

The aim of this study was to evaluate the performance of physiological measurements obtained with a photoplethysmography-based smartwatch by comparing systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) measurements with those obtained using a sphygmomanometer, and peripheral oxygen saturation (SpO_2) and HR measurements with those obtained using a fingertip pulse oximeter. In addition, the performance of these measurements was investigated across different skin phototypes classified according to the Fitzpatrick scale.

III. MATERIALS AND METHODS

A. Sample

The study included 114 adult participants of both sexes, recruited through direct invitations and referrals. Eligible participants were clinically healthy, had no known comorbidities that could interfere with the physiological measurements under investigation, and were able to understand and provide written informed consent. The sample had a mean age of 28.1 ± 9.1 years (range, 22–61 years) and comprised 71 women (62.3%) and 43 men (37.7%). All eligible participants with available data for the planned analyses were retained, with no participants excluded for the purpose of numerically balancing the skin phototype groups.

Participants were classified using a Skin Phototype Classification Questionnaire based on the Fitzpatrick scale and administered electronically. The analysis included participants classified as phototype I (n = 6), II (n = 48), III (n = 42), and IV (n = 18). Sample distribution and demographic characteristics according to skin phototype are presented in Table I.

Data were collected at the Polyclinic of the University of Mogi das Cruzes (UMC). The study was approved by the Research Ethics Committee of the University of Mogi das Cruzes under CAAE 64561022.7.0000.5497. All participants were informed about the study objectives and procedures and provided written informed consent before data collection.

TABLE I.
DEMOGRAPHIC CHARACTERISTICS OF THE SAMPLE ACCORDING TO SKIN PHOTOTYPE

Variable	*I (n=6)*	*II (n=48)*	*III (n=42)*	*IV (n=18)*	*Total (n=114)*
Age, years	23.5 ± 2.3	29.0 ± 10.1	28.5 ± 8.4	26.2 ± 9.0	28.1 ± 9.1
Age range, years	22–27	22–61	22–54	22–51	22–61
Female, n (%)	5 (83.3)	32 (66.7)	27 (64.3)	7 (38.9)	71 (62.3)
Male, n (%)	1 (16.7)	16 (33.3)	15 (35.7)	11 (61.1)	43 (37.7)

Note: Age is presented as mean ± standard deviation and range; sex is presented as n (%).

B. Protocol

Data collection was conducted in a controlled environment using a standardized protocol for all participants. Initially, each participant remained seated in a comfortable position while the study procedures and the informed consent form were explained. This period also provided an opportunity for participants to rest before the measurements. After receiving all relevant information, participants who agreed to take part provided written informed consent, and the experimental procedure was initiated.

During the measurements, participants remained seated with their backs supported against the chair, feet flat on the floor, and legs uncrossed. The Samsung Galaxy Watch4 BT was positioned on the left wrist, ensuring adequate contact between the sensor and the skin, while the cuff of the OMRON HEM-7320-BR automatic sphygmomanometer was placed on the right upper arm. The device uses an Intelli Wrap cuff (HEM-FL31), specified by the manufacturer for arm circumferences ranging from 22 to 42 cm. Throughout the data-collection period, the smartwatch software was checked approximately every 10 days and updated whenever an update was available. Consequently, a single firmware version was not maintained throughout the study.

For systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) measurements, the smartwatch was individually calibrated for each participant, as required for a new user. During calibration, three consecutive measurements were obtained with the sphygmomanometer, with simultaneous smartwatch acquisition during each measurement. The blood pressure values obtained with the sphygmomanometer were entered into the device calibration procedure. Once the three measurements and calibration procedure were completed, the smartwatch performed a subsequent measurement independently, without concurrent cuff measurement. After automatic processing, the final SBP, DBP, and HR values displayed by the smartwatch were recorded in the study database. For the comparative analysis, the arithmetic mean of the three sphygmomanometer measurements was calculated separately for SBP, DBP, and HR, and these mean values were compared with the corresponding final values obtained from

the smartwatch.

After completion of the SBP, DBP, and HR measurements, the sphygmomanometer cuff was removed while the smartwatch remained on the left wrist. For peripheral oxygen saturation (SpO_2) and HR measurements, a Multilaser HC261 fingertip pulse oximeter was placed on the right index finger. During smartwatch acquisition, the left upper limb was maintained at an approximately 45° angle, with the device positioned at heart level. SpO_2 and HR were measured simultaneously by the smartwatch and fingertip pulse oximeter, and the readings were allowed to stabilize before the values displayed by both devices were recorded.

Throughout all measurements, participants were instructed to remain still and refrain from speaking to minimize interference caused by motion or sensor displacement. If motion or improper sensor positioning prevented a valid reading from being obtained, the measurement was restarted.

C. *Statistical Analysis*

Data were organized in an electronic spreadsheet and analyzed using MATLAB® R2017a. Associations between measurements obtained with the smartwatch and the comparison devices were assessed using Pearson's (r) and Spearman's (ρ) correlation coefficients. Pearson's correlation was used for SBP, DBP, and HR comparisons between the smartwatch and the sphygmomanometer, whereas Spearman's correlation was used for SpO_2 and HR comparisons between the smartwatch and the fingertip pulse oximeter. Correlation coefficients are reported with their respective 95% confidence intervals (95% CIs) and p values. Agreement between measurements was assessed using the Bland–Altman method, with differences calculated as smartwatch minus comparison device. Mean bias and 95% limits of agreement (LoA), defined as bias ± 1.96 standard deviations of the differences, were calculated together with their respective 95% CIs. Proportional bias was assessed by linear regression of the differences against the mean of the paired measurements, with the slope coefficient, its 95% CI, and the corresponding p value reported [11].

To investigate differences in error magnitude across skin phototypes, the absolute error between each smartwatch measurement and the corresponding measurement obtained with the comparison device was calculated. Absolute-error distributions across phototypes I, II, III, and IV were compared using the Kruskal–Wallis test and are summarized as median and interquartile range (IQR). Effect size was estimated using epsilon-squared (ε^2), together with its 95% CI. Statistical significance was set at 5% ($p < 0.05$).

In addition, a Monte Carlo simulation-based sensitivity analysis was performed using the Kruskal–Wallis test while preserving the observed sample sizes of the four phototype groups (n = 6, 48, 42, and 18). The simulation considered an ordered location-shift model in which observations for each group were generated according to

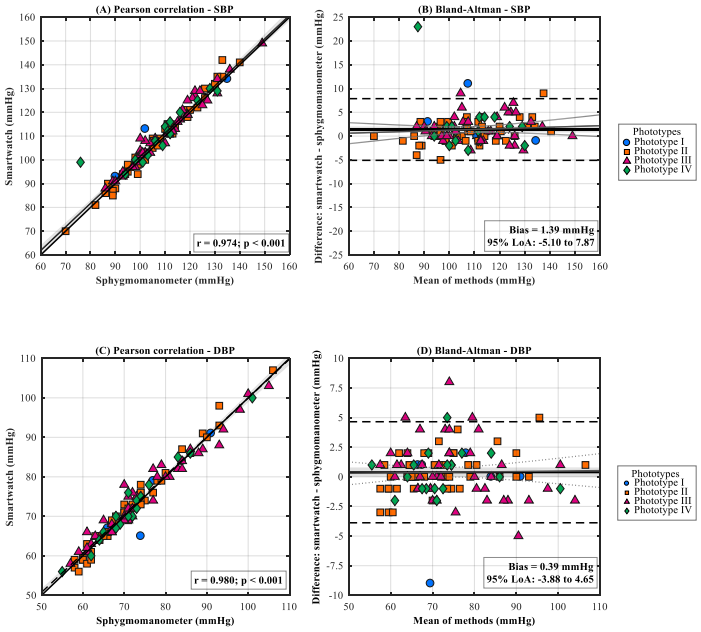
$$Xj \sim N\ [\ (j-1)\ \ \delta,1\ ],\quad j= 1, 2, 3, 4 \qquad (1)$$

where j denotes the phototype group (I–IV) and δ represents the standardized location shift between consecutive groups. Assuming continuous normal distributions with a common variance, a 5% significance level, and a target power of 80%, the minimum detectable standardized shift was defined as the smallest value of δ associated with approximately 80% power. The final power estimate was validated using 100,000 independent simulations.

IV. RESULTS

Blood pressure measurements obtained with the smartwatch showed strong positive linear associations with those recorded by the sphygmomanometer (Fig. 1A–D). For both SBP and DBP, values were closely distributed around the line of identity across the pressure range evaluated, with high correlation coefficients (SBP: r = 0.974; DBP: r = 0.980; both $p < 0.001$). Bland–Altman analysis showed differences predominantly centered around zero, with small positive mean biases of +1.39 mmHg for SBP and +0.39 mmHg for DBP, indicating slight average overestimation by the smartwatch. The 95% limits of agreement ranged from −5.10 to 7.87 mmHg for SBP and from −3.88 to 4.65 mmHg for DBP. Differences remained relatively stable across the observed blood pressure range, with no evidence of proportional bias for either SBP ($p = 0.470$) or DBP ($p = 0.836$).

HR showed a similarly strong positive association between the smartwatch and the sphygmomanometer ($r = 0.939$; $p < 0.001$), although with greater dispersion than observed for blood pressure measurements (Fig. 1E–F). Mean bias was small and positive (+1.45 bpm), while the 95% limits of agreement ranged from −6.79 to 9.69 bpm, indicating greater variability in individual differences. Although regression of the differences showed a positive slope with increasing mean HR, no statistically significant proportional bias was detected ($p = 0.087$).



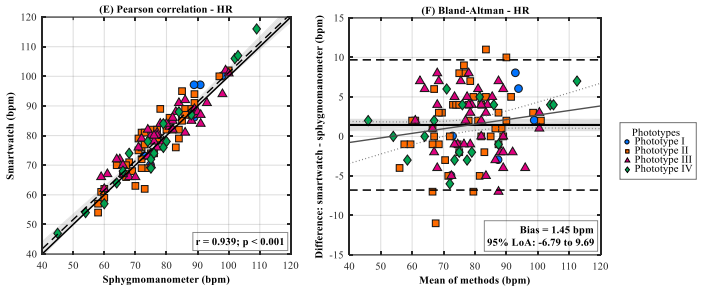


Fig. 1. Association and agreement between smartwatch and sphygmomanometer measurements of SBP (A–B), DBP (C–D), and HR (E–F), stratified by skin phototype. Pearson correlation (A, C, E) and Bland–Altman analysis (B, D, F).

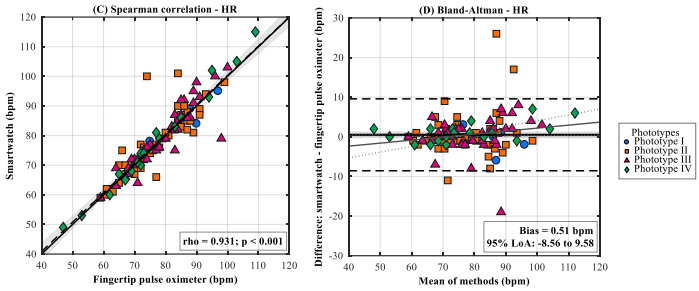
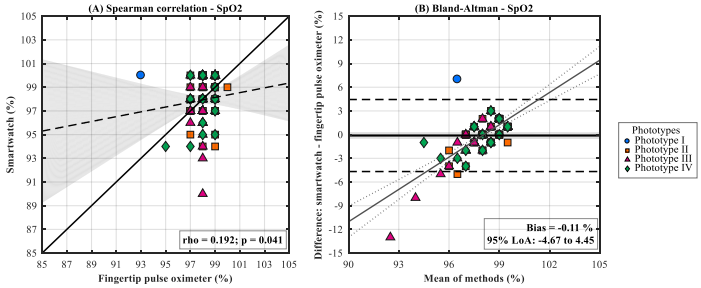


Fig. 2. Association and agreement between smartwatch and fingertip pulse oximeter measurements of SpO_2 (A–B) and HR (C–D), stratified by skin phototype. Spearman correlation (A, C) and Bland–Altman analysis (B, D).

In comparisons with the fingertip pulse oximeter, SpO_2 showed a markedly different pattern from the other physiological metrics (Fig. 2A–B). The association between devices was weak and positive (ρ = 0.192; 95% CI: 0.008–0.363; p = 0.041), with greater dispersion around the line of identity. Although mean bias was close to zero (−0.11%), the 95% limits of agreement ranged from −4.67% to 4.45%. A significant positive proportional bias was also identified (slope = 1.364; 95% CI: 1.119–1.609; p < 0.001), indicating that differences between the smartwatch and the pulse oximeter were not constant across the observed SpO_2 range but varied systematically with measurement magnitude.

In contrast to SpO_2, HR measured by the smartwatch maintained a strong positive association with HR measured by the fingertip pulse oximeter (ρ = 0.931; 95% CI: 0.902–0.952; p < 0.001) (Fig. 2C–D). Mean bias was small (+0.51 bpm); however, the 95% limits of agreement ranged from −8.56 to 9.58 bpm, indicating greater variability in individual differences despite the strong association between devices. Regression of the differences showed a positive slope, with a borderline result for proportional bias (slope = 0.076; 95% CI: 0.000–0.151; p = 0.050). Association and agreement results for all physiological measurements are summarized in Table II.



Stratified analysis of absolute errors showed substantial overlap in the distributions across phototypes I–IV for all evaluated metrics (Fig. 3A–E). No statistically significant differences among phototypes were detected for SBP (H = 1.36; p = 0.715), DBP (H = 4.49; p = 0.213), HR compared with the sphygmomanometer (H = 1.05; p = 0.788), SpO_2 (H = 2.67; p = 0.446), or HR compared with the fingertip pulse oximeter (H = 0.74; p = 0.865). None of the five metrics showed a progressive increase in absolute error across the evaluated phototypes, and effect-size estimates were null or small. Taken together, the error distributions, between-group comparisons, and effect-size estimates provided no evidence that larger errors were systematically concentrated within a specific phototype. Median and interquartile ranges of absolute errors by phototype, together with ε^2 estimates and their respective 95% CIs, are presented in Table III. In the Monte Carlo simulation-based sensitivity analysis, under the specified ordered location-shift model, the minimum detectable standardized shift was 0.405 standard deviations between consecutive groups, corresponding to a confirmed power of 80.7% (Monte Carlo 95% CI: 80.4%–80.9%).

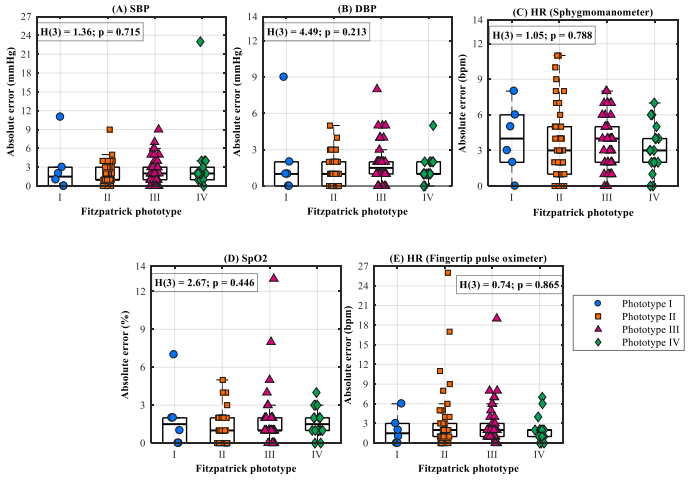


Fig. 3. Absolute errors across skin phototypes for SBP (A), DBP (B), HR compared with the sphygmomanometer (C), SpO_2 (D), and HR compared with the fingertip pulse oximeter (E). Kruskal–Wallis test.

TABLE II:

ASSOCIATION AND AGREEMENT BETWEEN SMARTWATCH MEASUREMENTS AND COMPARISON DEVICES

Metric	Association (95% CI)	P	Mean bias (95% CI)	Lower LoA (95% CI)	Upper LoA (95% CI)	Slope (95% CI)	Slope p
SBP	r = 0.974	<0.001	+1.39	−5.10	7.87	0.016	0.470

(mmHg)	(0.962–0.982)		(0.77–2.00)	(−6.15 to −4.05)	(6.82–8.92)	(−0.027–0.059)	
DBP (mmHg)	r = 0.980 (0.971–0.986)	<0.001	+0.39 (−0.02–0.79)	−3.88 (−4.57 to −3.19)	4.65 (3.96–5.34)	0.004 (−0.034–0.041)	0.836
HR – sphyg. (bpm)	r = 0.939 (0.912–0.957)	<0.001	+1.45 (0.67–2.23)	−6.79 (−8.13 to −5.46)	9.69 (8.35–11.03)	0.058 (−0.009–0.124)	0.087
SpO_2 (%)	ρ = 0.192 (0.008–0.363)	0.041	−0.11 (−0.55–0.32)	−4.67 (−5.41 to −3.93)	4.45 (3.71–5.19)	1.364 (1.119–1.609)	<0.001
HR – fingertip pulse oximeter (bpm)	ρ = 0.931 (0.902–0.952)	<0.001	+0.51 (−0.35–1.37)	−8.56 (−10.03 to −7.09)	9.58 (8.11–11.05)	0.076 (0.000–0.151)	0.050

Note: r, Pearson correlation coefficient; ρ, Spearman correlation coefficient; 95% CI, 95% confidence interval; LoA, 95% limits of agreement; SBP, systolic blood pressure; DBP, diastolic blood pressure; HR, heart rate; SpO_2, peripheral oxygen saturation. Bias was calculated as smartwatch − comparison device. The slope corresponds to the linear regression of the differences against the mean of paired measurements and was used to assess proportional bias.

TABLE III:
ABSOLUTE ERRORS OF SMARTWATCH MEASUREMENTS ACCORDING TO SKIN PHOTOTYPE

Metric	Phototype I (n=6)	Phototype II (n=48)	Phototype III (n=42)	Phototype IV (n=18)	H(3)	p	ε^2 (95% CI)
SBP (mmHg)	1.50 [0.25–2.75]	1.00 [1.00–3.00]	2.00 [1.00–3.00]	2.00 [1.00–2.75]	1.36	0.715	0.000 (0–0.088)
DBP (mmHg)	1.00 [0.25–1.75]	1.00 [0–2.00]	1.50 [1.00–2.00]	1.00 [1.00–2.00]	4.49	0.213	0.014 (0–0.142)
HR – sphyg. (bpm)	4.00 [2.25–5.75]	3.00 [1.00–5.00]	4.00 [2.00–5.00]	3.00 [2.00–4.00]	1.05	0.788	0.000 (0–0.081)
SpO_2 (%)	1.50 [0.25–2.00]	1.00 [0–2.00]	1.00 [1.00–2.00]	1.50 [1.00–2.00]	2.67	0.446	0.000 (0–0.114)
HR – fingertip pulse oximeter (bpm)	1.50 [0.25–2.75]	2.00 [1.00–3.00]	2.00 [1.00–3.00]	2.00 [1.00–2.00]	0.74	0.865	0.000 (0–0.074)

Note: Absolute errors are presented as median [interquartile range]. Absolute error was defined as |smartwatch − comparison device|. H, Kruskal–Wallis test statistic; ε^2, epsilon-squared; 95% CI, 95% confidence interval; SBP, systolic blood pressure; DBP, diastolic blood pressure; HR, heart rate; SpO_2, peripheral oxygen saturation.

V. DISCUSSION

The present study showed that the performance of physiological measurements obtained with the Galaxy Watch4 varied more across the physiological metrics evaluated than across the skin phototypes represented in the sample. SBP, DBP, and HR showed strong associations with their respective comparison devices and small mean biases, whereas SpO_2 exhibited a distinct pattern characterized by a weak association and significant proportional bias. In parallel, analysis of absolute errors identified no statistically significant differences among phototypes I–IV for any of the five comparisons and no systematic gradient of increasing error toward numerically higher phototypes. Taken together, these findings suggest that the performance of a PPG-based wearable device should not be regarded as a single property of the sensor, but rather as the result of interactions among the physiological metric being estimated, optical signal characteristics, and the processing methods used to derive each measurement [1], [2], [4], [12], [13].

This interpretation is consistent with the fundamental principles of photoplethysmography. PPG signals are shaped by light–tissue interactions and the pulsatile component of blood flow and are influenced by optical, physiological, mechanical, and algorithmic factors. Wavelength, light intensity, peripheral perfusion, motion, contact pressure, and sensor geometry can affect signal quality and, consequently, have different effects on the physiological metrics derived from these signals. Previous PPG research has identified hardware characteristics, measurement site, physiological conditions, and signal processing as important determinants of system performance. Accordingly, the strong associations observed for some variables and the less consistent behavior of SpO_2 support the view that physiological metrics obtained with the same device do not necessarily share the same performance profile [2], [4], [14].

Blood pressure measurements showed the most consistent performance among the metrics evaluated. SBP and DBP showed high correlations (r = 0.974 and r = 0.980, respectively), small mean biases (+1.39 and +0.39 mmHg), and no evidence of proportional bias. The limits of agreement were also narrower than those observed for HR, indicating lower dispersion of the differences across the blood pressure range studied. These findings, however, should be interpreted in the context of the measurement approach used by the Galaxy Watch4. Blood pressure estimation requires individual cuff-based calibration, and in the present protocol, the three sphygmomanometer measurements used for calibration also provided, through their arithmetic mean, the comparison value for the final smartwatch measurement. Therefore, the findings characterize device performance after individual calibration and under the experimental conditions employed; they do not constitute an independent evaluation of the calibration procedure itself. This distinction is particularly important when interpreting wearable technologies designed to estimate blood pressure [2], [3], [4], [15].

HR also showed strong associations with both comparison devices, although with greater individual variability. Correlation coefficients remained high relative to both the sphygmomanometer (r = 0.939) and the fingertip pulse oximeter (ρ = 0.931), whereas the limits of agreement were wider than those observed for SBP and DBP. These findings highlight an important methodological distinction: a strong association between two measurement methods does not necessarily imply small differences at the individual level. This consideration is particularly relevant when evaluating PPG-based devices, as the quality of the pulsatile waveform may vary even under resting conditions because of physiological characteristics and the sensor–skin interface. Validation studies and reviews of wearable devices have reported heterogeneous performance across devices and acquisition conditions, reinforcing the importance of interpreting measures of association alongside the distribution of paired differences [12], [14], [16].

SpO_2 showed the greatest divergence between the two measurement methods. Although mean bias was close to zero (−0.11%), the 95% limits of agreement ranged from −4.67% to 4.45%, the association was weak (ρ = 0.192), and Bland–Altman regression showed significant proportional bias. Thus, the near-zero overall mean bias masked a magnitude-dependent pattern, indicating that a small mean bias alone does not imply stable differences between devices. The concentration of observations within a high saturation range limited SpO_2

variability in the sample and may have contributed to the weak correlation despite relatively small absolute differences. Previous studies have shown that pulse oximetry performance may vary according to skin pigmentation, device characteristics, and the saturation range evaluated, with particular concern at lower oxygen saturation levels [2], [4], [6], [7], [13], [14], [17].

Interpretation of the SpO_2 findings should also consider the nature of the comparison device. Both the Galaxy Watch4 and the fingertip pulse oximeter rely on optical principles to estimate peripheral oxygen saturation, whereas studies assessing pulse oximetry accuracy commonly use arterial oxygen saturation determined by CO-oximetry as the reference. Therefore, the differences observed in the present study reflect the relative performance of two optical measurement systems rather than the error of the smartwatch relative to arterial oxygen saturation. This distinction is particularly relevant given the evidence that pulse oximetry performance may vary according to skin pigmentation characteristics [13], [17].

The phototype-stratified analysis provides an additional perspective on these findings. Despite the optical basis for a potential effect of pigmentation on PPG signals, no progressive increase in absolute error was observed across phototypes I–IV. The five comparisons showed substantial overlap in their distributions, all Kruskal–Wallis tests were nonsignificant, and effect-size estimates were null or small. Across SBP, DBP, HR compared with the sphygmomanometer, SpO_2, and HR compared with the fingertip pulse oximeter, ε^2 values ranged from 0 to 0.014, with no evidence that larger errors were systematically concentrated within a single phototype. These results do not demonstrate equivalence among groups, nor do they indicate that pigmentation is irrelevant to optical measurement. Rather, they show that, in this sample and with this device, a potential effect of pigmentation did not translate into a detectable gradient of increasing physiological measurement error across phototypes I–IV [5], [7], [13].

This finding is consistent with the complex relationship between pigmentation and optical technologies described in the literature. Melanin absorption is only one of several factors that determine PPG signal quality, and changes in the raw optical signal do not necessarily produce changes of the same magnitude in the physiological measurement reported by the device. Emitter and photodetector characteristics, wavelength, optical power, sensor geometry, and signal processing may all modify the relationship between skin properties and device output. Available evidence also does not suggest a uniform effect of pigmentation across physiological metrics: recent studies have reported pigmentation-related differences in SpO_2 measurements, whereas systematic or clinically relevant pigmentation-associated bias has not been consistently identified for pulse rate measurements obtained from wearable devices. Therefore, the absence of a common gradient across the five metrics in the present study, together with the differences observed among them, reinforces the need to evaluate performance at the level of individual metrics and devices rather than assuming a uniform effect across PPG-based technologies [2], [4], [6], [13], [18].

Interpretation of the phototype-stratified findings should also account for the classification method used. The Fitzpatrick scale was developed on the basis of skin response to sun exposure and provides a clinical classification of skin phototype rather than a direct quantitative measure of melanin concentration or the optical properties of the skin. Consequently, individuals within the same phototype category may differ in pigmentation characteristics, and biologically relevant differences may not be fully captured by this ordinal classification. Future studies combining phototype classification with instrumental measurements, such as melanin index or spectrophotometry, may provide a more precise characterization of the relationship between skin optical properties and sensor performance. This distinction is important because skin phototype and pigmentation should not be regarded as equivalent concepts [8], [18].

Several characteristics of the sample and study design should be considered when interpreting these findings. Although all eligible participants were retained in the analyses, the distribution across phototypes was unequal, with only six participants classified as phototype I, and phototypes V and VI were not represented. The sensitivity analysis provides further context for this limitation: under the progressive-shift scenario defined for the observed sample sizes, approximately 80% power was achieved to detect a standardized shift of 0.405 standard deviations between consecutive groups; therefore, smaller effects remain compatible with the data and cannot be excluded solely on the basis of statistical nonsignificance. In addition, the study evaluated a single smartwatch model in clinically healthy adults under controlled resting conditions, limiting extrapolation to other devices, clinical populations, and physiological conditions. The smartwatch was not maintained on a single firmware version throughout data collection, which should also be considered when interpreting the findings. Despite these limitations, the study design allowed five physiological comparisons to be evaluated within the same population and under a standardized protocol, combining measures of association, agreement, proportional bias, and absolute errors across phototypes. Under the conditions evaluated, SBP and DBP showed the most consistent performance after individual calibration, HR maintained strong associations despite greater individual variability, and SpO_2 showed magnitude-dependent differences, while absolute errors showed no systematic gradient across phototypes I–IV. Taken together, these findings highlight the importance of evaluating PPG-based wearable technologies at the level of individual physiological metrics rather than assuming a generalized effect of skin phototype on device performance.

VI. CONCLUSION

The results of this study showed that the performance of physiological measurements obtained with the smartwatch varied across the physiological metrics evaluated. SBP and DBP showed strong associations, small mean biases, and no significant proportional bias, whereas HR maintained strong associations with the comparison devices, although with greater individual variability. SpO_2 showed a distinct pattern, with a weak association and significant proportional bias, indicating systematic variation in between-device differences according to measurement magnitude.

Across skin phototypes I–IV, no statistically significant differences in the magnitude of absolute errors were detected for any of the evaluated metrics, and no systematic gradient of increasing error was observed across groups. These findings indicate that, under the conditions evaluated and for the Galaxy Watch4, differences in performance were more evident across physiological metrics than across the skin phototypes represented in this sample. The results underscore the importance of evaluating each physiological metric individually when assessing PPG-based wearable technologies and contribute to a better understanding of their performance across populations with different skin phototypes.

ACKNOWLEDGMENT

K.C.B.S. acknowledges the financial support of the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil (Grant No. 88887.941749/2024-00). The authors gratefully acknowledge the University of Mogi das Cruzes (UMC) and the UMC Polyclinic for providing the facilities and infrastructure necessary to conduct this study. The authors also sincerely thank all the volunteers who generously participated in this research. Finally, the authors express their sincere gratitude to Dr. Daniel Gustavo Goroso for his invaluable guidance, continuous scientific support, and insightful discussions throughout the development of this study.

REFERENCES

[1] A. Ometov *et al.*, "A Survey on Wearable Technology: History, State-of-the-Art and Current Challenges," *Computer Networks*, vol. 193, p. 108074, Jul. 2021, doi: 10.1016/j.comnet.2021.108074.

[2] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey, "Wearable Photoplethysmography for Cardiovascular Monitoring," *Proc. IEEE*, vol. 110, no. 3, pp. 355–381, Mar. 2022, doi: 10.1109/JPROC.2022.3149785.

[3] Z. Sagirova *et al.*, "Cuffless Blood Pressure Measurement Using a Smartphone-Case Based ECG Monitor with Photoplethysmography in Hypertensive Patients.," *Sensors (Basel)*, vol. 21, no. 10, Jun. 2021, doi: 10.3390/s21103525.

[4] K. B. Kim and H. J. Baek, "Photoplethysmography in Wearable Devices: A Comprehensive Review of Technological Advances, Current Challenges, and Future Directions," 2023.

[5] O. E. Okunlola, M. S. Lipnick, P. B. Batchelder, M. Bernstein, J. R. Feiner, and P. E. Bickler, "Pulse Oximeter Performance, Racial Inequity, and the Work Ahead.," *Respir Care*, vol. 67, no. 2, pp. 252–257, 2022, doi: 10.4187/respcare.09795.

[6] P. E. Bickler, "Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation," vol. 102, no. 4, 2005.

[7] J. R. Feiner, J. W. Severinghaus, and P. E. Bickler, "Dark Skin Decreases the Accuracy of Pulse Oximeters at Low Oxygen Saturation: The Effects of Oximeter Probe Type and Gender," *Anesthesia & Analgesia*, vol. 105, no. 6, pp. S18–S23, Dec. 2007, doi: 10.1213/01.ane.0000285988.35174.d9.

[8] S. Sachdeva, "Fitzpatrick skin typing: Applications in dermatology," vol. 75(1), p. 93, Feb. 2009, doi: 10.4103/0378-6323.45238.

[9] A. Icenhower *et al.*, "Investigating the accuracy of Garmin PPG sensors on differing skin types based on the Fitzpatrick scale: cross-sectional comparison study," *Front. Digit. Health*, vol. 7, p. 1553565, Mar. 2025, doi: 10.3389/fdgth.2025.1553565.

[10] H. Sanae Suzuki, "Phototype Comparison Between Caucasian and Asian Skin Types," vol. 3, no. 3, Surgical & Cosmetic Dermatology, Brazil, pp. 193–196, 2011.

[11] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, Sep. 2018, doi: 10.1016/j.tjem.2018.08.001.

[12] F. Sarhaddi *et al.*, "A comprehensive accuracy assessment of Samsung smartwatch heart rate and heart rate variability," *PLoS ONE*, vol. 17, no. 12, p. e0268361, Dec. 2022, doi: 10.1371/journal.pone.0268361.

[13] S. Singh, M. R. Bennett, C. Chen, S. Shin, H. Ghanbari, and B. W. Nelson, "Impact of Skin Pigmentation on Pulse Oximetry Blood Oxygenation and Wearable Pulse Rate Accuracy: Systematic Review and Meta-Analysis.," *J Med Internet Res*, vol. 26, pp. e62769–e62769, Oct. 2024, doi: 10.2196/62769.

[14] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, pp. R1–R39, Mar. 2007, doi: 10.1088/0967-3334/28/3/R01.

[15] M. Yavarimanesh *et al.*, "Assessment of Calibration Models for Cuff-Less Blood Pressure Measurement After One Year of Aging.," *IEEE Trans Biomed Eng*, vol. 69, no. 6, pp. 2087–2093, 2022, doi: 10.1109/TBME.2021.3136492.

[16] P. H. Charlton *et al.*, "The 2023 wearable photoplethysmography roadmap," *Physiol. Meas.*, vol. 44, no. 11, p. 111001, Nov. 2023, doi: 10.1088/1361-6579/acead2.

[17] C. Shi *et al.*, "The accuracy of pulse oximetry in measuring oxygen saturation by levels of skin pigmentation: a systematic review and meta-analysis," *BMC Med*, vol. 20, no. 1, p. 267, Aug. 2022, doi: 10.1186/s12916-022-02452-8.

[18] B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," *J Clin Monit Comput*, vol. 27, no. 3, pp. 313–317, Jun. 2013, doi: 10.1007/s10877-013-9436-7.

Physiological Evaluation of Heat Stress Associated with Mission-Oriented Protective Posture (MOPP) in Brazilian Army Military Personnel

Victor Hugo P. Greenhalgh
Escola de Aperfeiçoamento de Oficiais
Exército Brasileiro
Rio de Janeiro, Brazil
vhgreenhalgh@gmail.com

Eduardo Borba Neves
Escola de Aperfeiçoamento de Oficiais
Exército Brasileiro
Rio de Janeiro, Brazil
neveseb@gmail.com

***Abstract* — This study evaluated the effects of different levels of Mission-Oriented Protective Posture (MOPP) on physiological and perceptual responses in military personnel during simulated Chemical, Biological, Radiological, and Nuclear (CBRN) physical tasks. A quantitative, experimental, within-subjects, repeated-measures design was adopted, with 26 military personnel exposed to MOPP 0, MOPP 2, and MOPP 4 conditions. The protocol included a 4-km march, a 400-m high-intensity sprint, and simulated casualty transport. Heart rate, forehead infrared temperature, hydration-corrected percentage body mass loss, and rating of perceived exertion were analyzed, with emphasis on changes between the pre-exertion and post-physical-task assessments. Depending on data distribution, repeated-measures ANOVA or the Friedman test was applied. Increasing levels of protection produced greater physiological and perceptual strain, with large effect sizes for the change in heart rate and percentage body mass loss, a medium effect for the change in rating of perceived exertion, and a small effect for the change in forehead infrared temperature. Rating of perceived exertion was also significantly higher under MOPP 4 at the pre-exertion assessment. These findings indicate that equivalent physical tasks impose greater physiological and perceptual demands when performed under higher MOPP levels, suggesting that the level of protection should be considered when planning workload, recovery periods, and sustained personnel employment during CBRN operations.**

Keywords—MOPP, heat stress, physiological monitoring, military personnel, protective clothing, thermoregulation, CBRN operations, heart rate, rating of perceived exertion.

I. INTRODUCTION

The use of Personal Protective Equipment (PPE) in Chemical, Biological, Radiological, and Nuclear (CBRN) defense operations is essential for the safety and survival of military personnel on the modern battlefield [1]. However, the use of this equipment, which includes the Permeable Protective Combat Garment (PPCG) and the gas mask, imposes substantial physiological constraints on the operator [2]. The physical barrier required to filter hazardous agents also limits heat exchange with the environment and sweat evaporation, thereby increasing body heat storage and cardiovascular demand during exertion [3].

Body temperature regulation depends on the continuous balance between metabolic heat production and heat dissipation to the external environment [4]. Most of the energy generated by muscular metabolism is converted into heat, which must be dissipated through conduction, convection, radiation, and sweat evaporation. When these mechanisms become insufficient, physiological changes may occur, including increased heart rate, elevated body temperature, and dehydration [2], [4].

In CBRN scenarios, protective clothing adversely affects this balance because of the high thermal insulation and evaporative resistance of the textile materials employed [3], [5]. Military protective garments equipped with activated-carbon filtering layers, although permeable to some degree of gas exchange, substantially restrict sweat dissipation, an effect further aggravated by the use of fully impermeable rubber overboots and gas masks. This process creates a hot and humid microclimate within the protective ensemble, accelerating thermal and cardiovascular fatigue in the operator [2], [6].

The transition from a state of thermal compensation toward exhaustion requires continuous assessment of robust physiological markers. Heart rate (HR) serves as a key indicator of cardiovascular strain, reflecting the increased cardiac effort required to maintain blood flow to the periphery for heat dissipation. At the same time, changes in body mass provide an objective indication of acute fluid loss associated with high sweating rates [7]. In parallel, recent studies, together with advances in measurement technology, have established infrared measurement of surface temperature as a noninvasive and safe method for assessing peripheral thermal responses without significantly interfering with military personnel mobility [8]. In addition to objective parameters, the Rating of Perceived Exertion (RPE) has become a valuable and practical self-assessment tool for identifying increasing levels of discomfort and exertional strain [9].

To manage the balance between protection and thermal risk, military doctrine regulates the use of PPE through Mission-Oriented Protective Posture (MOPP), which organizes protective equipment into progressively higher levels, typically ranging from MOPP 1 to MOPP 4 [1], [10]. Progression to MOPP 4 increases the physiological and perceptual burden imposed on the operator because it combines the highest level of cutaneous and respiratory protection, thereby increasing strain during physical activity and thermal exposure [6].

The literature indicates that the operational safety margin is reduced under these conditions, making environmental and individual monitoring important for preventing heat-related casualties [7]. In this context, the present study aimed to investigate the effects of different levels of protection (MOPP 2 and MOPP 4), compared with a condition without protective measures (MOPP 0), on heart rate (HR), forehead infrared temperature (TIR), percentage body mass loss (ΔMC%), and Rating of Perceived Exertion (RPE), with particular emphasis

Escola de Aperfeiçoamento de Oficiais – Exército Brasileiro.

on changes between the pre-exertion and post-physical-task assessments. The analysis focuses on thermophysiological strain during the active phase of the mission, providing auditable evidence to support operational planning, heat-stress management, and the doctrinal application of protective measures during CBRN operations.

II. Methodology

A. Sample and Experimental Design

The study was characterized as quantitative, experimental, and applied research, using a within-subjects repeated-measures design with a 3 × 2 matrix, since the effects of three MOPP levels were evaluated at two different assessment moments, before and after the physical exercises. This design was adopted to reduce individual variability and allow direct comparison of the physiological responses produced by the different protection levels. The study protocol was approved by a Research Ethics Committee, and followed the Declaration of Helsinki.

Initially, the non-probabilistic purposive sampling design planned for the participation of 30 male military personnel from the 1st Chemical, Biological, Radiological, and Nuclear Defense Battalion (1st CBRN Def Bn). The selection sought to approximately reflect the actual distribution of ranks within the battalion, including officers, sergeants, corporals, and privates, with the aim of ensuring the functional representativeness of a typical Army unit engaged in CBRN operations.

Considering the repeated-measures design, a statistical power analysis was performed using a significance level of 5%, statistical power of 80%, and an expected medium-to-large effect size (d = 0.59), resulting in a minimum estimated sample of 25 participants [11]. The final sample consisted of 26 military personnel, therefore exceeding the minimum estimated sample size.

Participants were physically fit, medically cleared, free from musculoskeletal restrictions or a recent history of heat-related illness, and technically familiar with the PPE evaluated. The loss of four participants resulted from operational factors related to military routine and the application of exclusion criteria, including clinical indisposition, absence from any experimental session, failure to complete the full protocol, or noncompliance with prior instructions regarding hydration, rest, abstinence from alcohol, and avoidance of intense physical exercise during the preceding 24 h.

B. Procedures

The military personnel performed the tests under the three protection conditions on separate days, with the conditions balanced among participants to minimize learning effects or accumulated fatigue. Thus, on each day, one-third of the participants wore each of the three MOPP levels, randomly assigned, so that by the end of the study each participant had completed the experiment three times, once under each MOPP level.

The experimental conditions were defined so that participants were informed of the protective condition assigned to them only at the beginning of each experimental day, before exercise, as follows:

1) MOPP 0: standard military physical training uniform, consisting of a sleeveless shirt, shorts, and athletic shoes, without CBRN protective equipment.
2) MOPP 2: Permeable Protective Combat Garment (PPCG), consisting of a jacket and trousers with an activated-carbon inner layer, without hood or gloves, but with rubber overboots.
3) MOPP 4: complete protective ensemble consisting of a gas mask, PPCG, gloves, and overboots [1].

It is important to emphasize that MOPP 0 does not constitute an operational posture used in contaminated environments and was employed exclusively as the minimum experimental reference condition for quantifying the additional strain imposed by the protective levels. The MOPP levels adopted in the experiment are illustrated in Fig. 1.

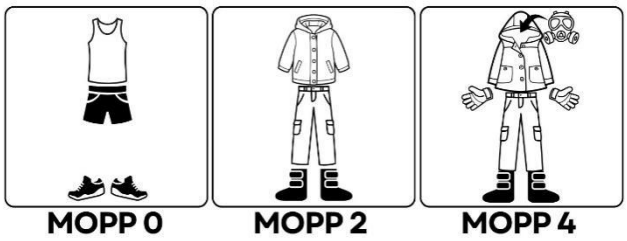


Fig. 1. MOPP levels adopted in the experiment.

C. Data Collection Instruments

For physiological and perceptual monitoring, pre-exertion and post-physical-task assessment points were established, in addition to the initial and final body-mass measurements used to determine percentage body mass loss. The instruments used comprised:

1) Kestrel 3000 portable digital weather station, used for environmental characterization of the experimental sessions, operating over an air-temperature range of −29.0°C to +70.0°C with an accuracy of ±1.0°C through a precision thermistor-based sensor, and over a relative-humidity range of 5% to 95%, with an accuracy of ±3.0% RH using a silicon-polymer capacitive sensor.
2) Digital scale, with a capacity of up to 180 kg and an accuracy of 100 g, used to record initial and final body mass and determine total fluid loss.
3) G-Tech Easy Sensor infrared thermometer, configured in body mode and used to measure forehead infrared temperature (TIR), positioned approximately 15 cm from the central point of the forehead along the midline between the eyebrows. The measurement distance and anatomical location were maintained across all sessions to standardize the procedure.
4) Digital fingertip pulse oximeter, based on Beer–Lambert law spectrophotometry, used to measure heart rate (HR) over a range of 30 to 235 bpm, with an accuracy of ±2% of the measured bpm.
5) Visual analog scale for perceived exertion, based on the Borg scale [9], consisting of a progressive 1-to-7 scale with emojis and descriptors ranging from *Almost Imperceptible* to *Extremely Hard*, as shown in Fig. 2.



Fig. 2. Borg scale adapted with emojis.

D. Physical Exertion Protocol

The simulated workload was divided into three practical tasks performed sequentially to generate aerobic and anaerobic demands under thermal strain. The following steps were performed each day:

1) Recording of environmental conditions.

2) Initial physiological assessment: initial body-mass measurement.

3) Donning of equipment according to the experimental condition assigned for the day.

4) Pre-exertion physiological assessment: measurement of HR, RPE, and forehead infrared temperature (TIR).

5) Task A (March): 4-km movement over flat terrain at a speed of 5 km/h, simulating controlled progression toward a contaminated area and emphasizing aerobic endurance.

6) Task B (Sprint): 400-m sprint at maximum speed and intensity, engaging the lactic anaerobic system and simulating tactical evasion or assault movements.

7) Task C (Casualty Evacuation): tactical evacuation performed in pairs over 200 m while carrying a litter loaded with 60 kg of weight plates placed in military backpacks, requiring submaximal strength and substantial involvement of the upper-body and trunk musculature, thereby simulating casualty evacuation from a contaminated area.

8) Post-physical-task physiological assessment: measurement of HR, RPE, and forehead infrared temperature (TIR).

9) Controlled removal of the protective equipment.

10) Final physiological assessment: final body-mass measurement.

Body-mass measurements were performed with participants wearing only shorts, preventing sweat retained in the protective clothing from being included in the measurement.

Fig. 3 provides a visual representation of the experimental procedures for improved understanding.

E. Investigated Variables

The *independent variables* were MOPP level (MOPP 0, MOPP 2, and MOPP 4) and assessment time (pre-exertion and post-physical-task). MOPP level constituted the main experimental factor, whereas assessment time represented the temporal factor used to quantify the physiological and perceptual changes produced by the physical exertion protocol, which consisted of marching, sprinting, and simulated casualty transport.

Air temperature and relative humidity were recorded as *control variables* and used to obtain a simplified estimate of the Wet Bulb Globe Temperature (WBGT) and characterize the thermal conditions of each experimental session. Because black-globe temperature was not measured, this index was used exclusively for environmental comparison among collection days and was not considered equivalent to the standardized measurement specified by ISO 7243. The starting time of exercise (09:00 h), the 4-km march route and pace (approximately 5 km/h), task sequence, 400-m sprint duration (between 90 and 100 s), and 200-m casualty transport with a 60-kg load were also standardized. Participants received standardized instructions regarding their usual diet, prior hydration, and abstinence from alcohol and intense physical exercise during the preceding 24 h. The balanced distribution of MOPP conditions across collection days was used to reduce the influence of environmental variation on within-subject comparisons.

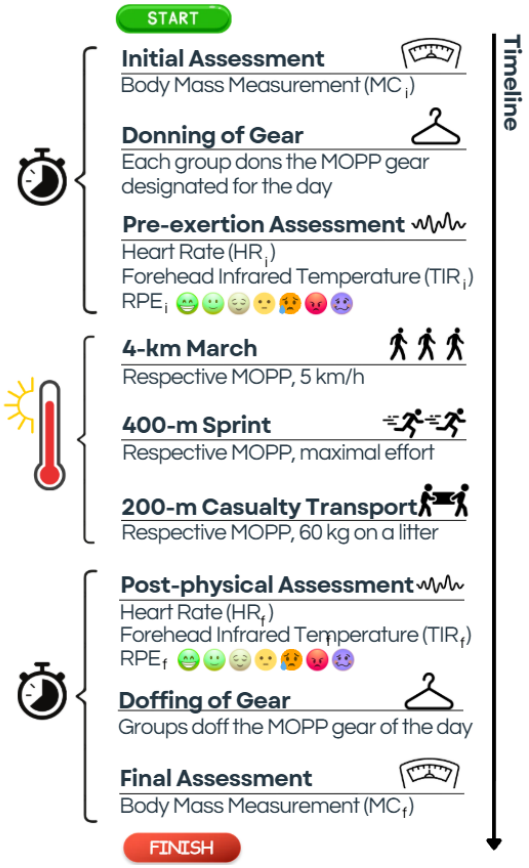


Fig. 3. Experimental and data collection protocol for the Physiological Evaluation of Heat Stress Associated with Mission-Oriented Protective Posture (MOPP) in Brazilian Army Personnel (Rio de Janeiro, 2025).

The *dependent variables* analyzed were:

1) Heart Rate (HR): measured in beats per minute (bpm) using a digital fingertip pulse oximeter at the pre-exertion and post-physical-task assessment points. HR was calculated as the mean of the values recorded during the third and fourth minutes of measurement and was used as an indicator of cardiovascular strain associated with physical exertion and the level of protection.

2) Forehead Infrared Temperature (TIR): measured noninvasively using a G-Tech Easy Sensor infrared thermometer positioned approximately 15 cm in front of the central point of the forehead. The difference between the post-physical-task and pre-exertion assessments (ΔTIR) was used as an indicator of the superficial thermal response associated with the protocol and the different MOPP levels.

3) Percentage Body Mass Loss (ΔMC%): used as an estimate of cumulative fluid loss during the protocol and calculated from initial and final body mass, corrected for the 500 mL of water ingested during the session. Body-mass measurements were performed while participants wore only standardized shorts, thereby reducing the influence of sweat retained in the clothing. The following expression was used:

$$\Delta MC\ (\%) = \frac{M_i - M_f + H}{M_i} \times 100 \quad (1)$$

where M_i represents initial body mass, M_f represents final body mass, and H represents the mass equivalent of the volume of water ingested during the session.

4) Rating of Perceived Exertion (RPE): assessed at the pre-exertion and post-physical-task assessment points using the adapted 1-to-7 visual analog scale based on the Borg scale [9]. The difference between the post-physical-task and pre-exertion assessments (ΔRPE) was used as an indicator of the perceptual response associated with the physical exertion protocol and the different MOPP levels. Higher scores indicated greater perceived exertion.

F. Statistical Analysis

Statistical analyses were performed using the *Statistical Analysis Software for Comparison of Measures Between Groups*, version 1.02, developed in Python [12]. Descriptive analyses were initially performed using means and standard deviations, with 95% confidence intervals when applicable. Distribution normality was assessed using the Shapiro–Wilk and Kolmogorov–Smirnov tests.

Considering the within-subjects design, changes between the post-physical-task and pre-exertion assessments were compared across MOPP 0, MOPP 2, and MOPP 4 using repeated-measures ANOVA followed by pairwise comparisons with Bonferroni adjustment when parametric assumptions were met. For non-normally distributed variables, the Friedman test followed by the Nemenyi post-hoc test was used. When necessary, the Greenhouse–Geisser correction was applied for violations of sphericity.

Effect magnitude was expressed using generalized eta squared ($\eta^2 g$) for parametric analyses and Kendall's W for Friedman analyses. The significance level was set at 5% ($p < 0.05$).

III. Results

The period between the first daily assessment and completion of the physical phase averaged 1 h 22 min. Environmental conditions are summarized in Table I. Day 1 showed the highest simplified WBGT estimate, whereas Days 2 and 3 were lower, with MOPP levels balanced across participants on all collection days.

TABLE I. Atmospheric Conditions of the Experimental and Data Collection Protocol for the Physiological Evaluation of Heat Stress Associated with Mission-Oriented Protective Posture (MOPP) in Brazilian Army Personnel (Rio de Janeiro, 2025).

Day	*Temperature (°C)*	*Relative Humidity*	*Estimated WBGT (°C)*
Day 1	26	81%	24.2
Day 2	24	72%	21.4
Day 3	25	64%	21.6

Table II presents the descriptive statistics for the physiological and perceptual variables at the pre-exertion and post-physical-task assessments across the three experimental conditions. Descriptively, post-physical-task heart rate and rating of perceived exertion values increased with increasing protection level. Forehead infrared temperature values showed a less uniform pattern across conditions.

TABLE II. Pre-Exertion and Post-Physical-Task Physiological Assessments for the Physiological Evaluation of Heat Stress Associated with Mission-Oriented Protective Posture (MOPP) in Brazilian Army Personnel (Rio de Janeiro, 2025)

Measurement	*Operational Condition*		
	MOPP 0	*MOPP 2*	*MOPP 4*
Pre-Exertion Heart Rate $[Avg \pm SD]$[a] (bpm)	79.62 ± 12.04	76.85 ± 14.96	76.73 ± 10.97
Post-Physical-Task Heart Rate $[Avg \pm SD]$[a] (bpm)	100.00 ± 11.08	108.19 ± 10.69	116.23 ± 13.72
Pre-Exertion TIR $[Avg \pm SD]$[a] (°C)	35.78 ± 0.32	35.66 ± 0.54	35.70 ± 0.40
Post-Physical-Task TIR $[Avg \pm SD]$[a] (°C)	35.41 ± 0.78	35.40 ± 0.52	35.75 ± 0.42
Pre-Exertion Rating of Perceived Exertion $[Avg \pm SD]$[a]	1.00 ± 0.00	1.08 ± 0.27	1.85 ± 0.61
Post-Physical-Task Rating of Perceived Exertion $[Avg \pm SD]$[a]	2.46 ± 1.03	3.50 ± 1.14	4.08 ± 0.89

[a.] TIR = forehead infrared temperature; Avg = average; SD = standard deviation.

To assess the comparability of conditions before execution of the physical tasks, measurements obtained at the pre-exertion assessment under MOPP were analyzed, as shown in Table III. No significant differences among conditions were identified for heart rate or TIR. RPE showed a distinct pattern, with higher values under MOPP 4 than under MOPP 0 and MOPP 2, whereas MOPP 0 and MOPP 2 did not differ. It should be noted that this assessment was performed after donning the protective ensemble and therefore represents the pre-exertion condition under MOPP, rather than a baseline measure free from the influence of the equipment.

TABLE III. Statistical Comparison of Pre-Exertion Measurements for the Physiological Evaluation of Heat Stress Associated with Mission-Oriented Protective Posture (MOPP) in Brazilian Army Personnel (Rio de Janeiro, 2025)

Measurement	*Result [$\chi^2(2)$; p-value]*	*Effect Size (Kendall's W)*	*Difference between MOPP (p-value)*
Heart Rate (HR)[b]	1.000; p = 0.607	0.019 (Small)	-
TIR[b]	1.482; p = 0.477	0.029 (Small)	-
Rating of Perceived Exertion (RPE)[bc]	34.421; p < 0.001	0.662 (Large)	MOPP 0-4 (p < 0.001) MOPP 0-2 (p = 0.472) MOPP 2-4 (p < 0.001)

[b.] The global Friedman test was applied to all variables because normality was rejected ($p < 0.05$) by the Shapiro-Wilk and Kolmogorov-Smirnov tests.

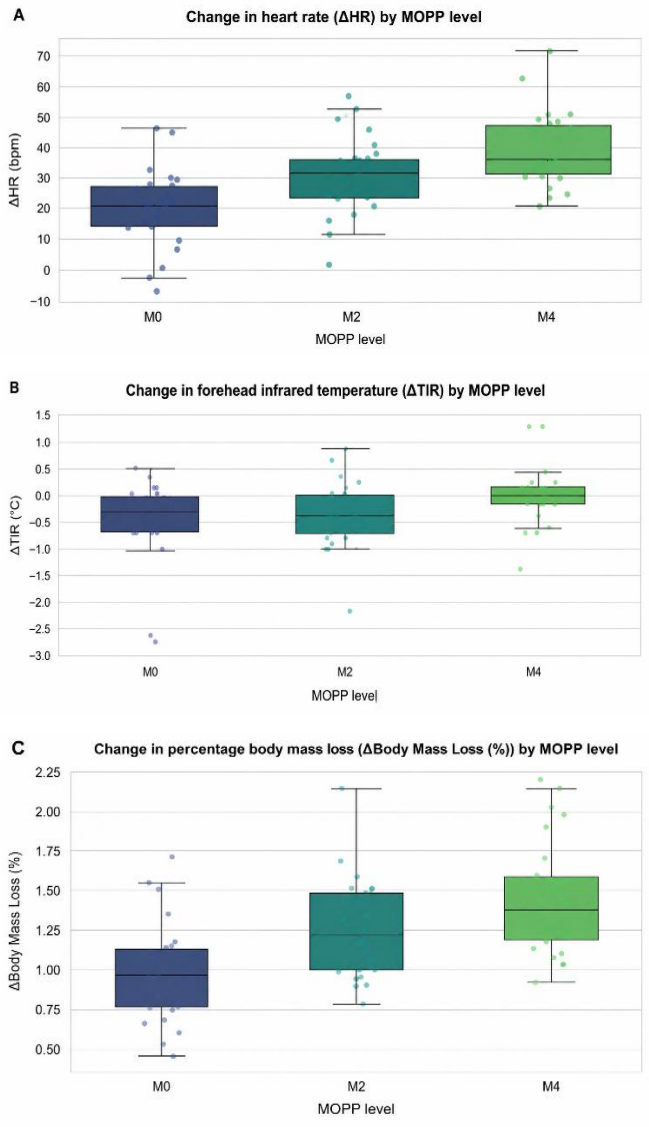
[c.] For pairwise RPE comparisons, the Nemenyi post-hoc test was used.

The response to the physical phase was evaluated primarily using differences between the post-physical-task and pre-exertion assessments. The increase in heart rate was greater under MOPP 2 and MOPP 4 than under MOPP 0, with

no significant difference between MOPP 2 and MOPP 4. The change in TIR also differed among conditions, with MOPP 4 showing a distinct pattern from MOPP 0 and MOPP 2, whereas these two conditions did not differ from each other. For hydration-corrected percentage body mass loss, a progressive increase was observed from MOPP 0 to MOPP 2 to MOPP 4, with differences among all three conditions. The change in RPE was greater under MOPP 2 and MOPP 4 than under MOPP 0, with no difference between the two protected levels.

Although absolute post-physical-task RPE values increased progressively from MOPP 0 to MOPP 2 to MOPP 4, the change relative to pre-exertion did not show the same ordering because pre-exertion RPE was already higher under MOPP 4. This pattern should be considered when interpreting the pre-exertion values together with their respective changes.

Overall, the four indicators differed across protection levels, with distinct magnitudes: a large effect for ΔHR and percentage body mass loss, a medium effect for ΔRPE, and a small effect for ΔTIR. The distributions and physiological and perceptual changes are shown in Fig. 4, while the inferential results and comparisons among MOPP levels are summarized in Table IV.



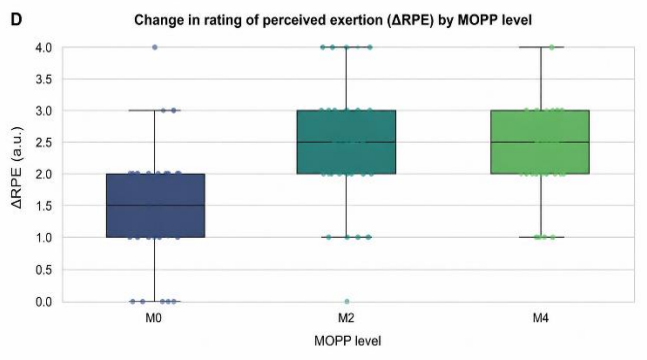


Fig. 4. Thermophysiological and perceptual changes in the sample before and after the physical exertion protocol under the different operational conditions: (A) change in heart rate (ΔHR); (B) change in forehead infrared temperature (ΔTIR); (C) change in percentage body mass loss (ΔMC%); (D) change in rating of perceived exertion (ΔRPE).

TABLE IV. STATISTICAL COMPARISON OF WITHIN-SUBJECT CHANGE (POST-PHYSICAL-TASK − PRE-EXERTION) ASSOCIATED WITH MISSION-ORIENTED PROTECTIVE POSTURE (MOPP) IN BRAZILIAN ARMY PERSONNEL (RIO DE JANEIRO, 2025)

Measurement (Δ)	*Result [Statistic; p-value]*	*Effect Size*	*Difference between MOPP (p-value)*
ΔHR[de]	F(1.537, 38.432) = 15.012 p < 0.001	η^2g = 0.295 (Large)	MOPP 0-2 (p = 0.022) MOPP 0-4 (p < 0.001) MOPP 2-4 (p = 0.174)
ΔTIR[df]	$\chi^2(2)$ = 10.950 p = 0.004	W = 0.211 (Small)	MOPP 0-2 (p = 0.936) MOPP 0-4 (p = 0.008) MOPP 2-4 (p = 0.023)
ΔMC%[de]	F(2,50) = 18.924 p < 0.001	η^2g = 0.260 (Large)	MOPP 0-2 (p = 0.005) MOPP 0-4 (p < 0.001) MOPP 2-4 (p = 0.039)
ΔRPE[df]	$\chi^2(2)$ = 17.167 p < 0.001	W = 0.330 (Medium)	MOPP 0-2 (p = 0.004) MOPP 0-4 (p = 0.003) MOPP 2-4 (p = 0.997)

[d.] ΔHR = post-physical-task minus pre-exertion change in heart rate; ΔTIR = post-physical-task minus pre-exertion change in forehead infrared temperature; ΔMC% = percentage body mass loss; ΔRPE = post-physical-task minus pre-exertion change in rating of perceived exertion.

[e.] For heart rate and percentage body mass loss, which met the parametric assumptions, repeated-measures ANOVA was applied with Bonferroni-adjusted pairwise comparisons and calculation of generalized eta squared (η^2g).

[f.] For TIR and RPE, the nonparametric Friedman test (χ^2) was used with the Nemenyi post-hoc test and Kendall's W for effect size. The significance level was $p < 0.05$.

IV. DISCUSSION

The results show that progression through MOPP levels increases the physiological and perceptual cost required to perform the same sequence of physical tasks. This pattern was not limited to a single marker: heart rate, percentage body mass loss, and rating of perceived exertion showed responses consistent with greater strain under protected conditions, although with different magnitudes. Heart rate and percentage body mass loss showed large effects, whereas the change in RPE showed a medium effect and forehead infrared temperature (TIR) a small effect. This difference in effect sizes reinforces that the assessed indicators represent distinct, although related, components of the response to exertion

under protection and should not be interpreted as interchangeable measures of heat stress [2], [7].

A particularly relevant finding was observed as early as the pre-exertion assessment. While heart rate and TIR showed similar initial conditions across protection levels, RPE was significantly higher under MOPP 4, with a large effect. Because this assessment was performed after donning the protective ensemble, the finding indicates that full protection already alters perceived exertion before physical activity begins. The gas mask, perceived breathing restriction, increased mass and bulk of the equipment, sensory limitation, and the sensation of confinement are components capable of increasing the discomfort associated with PPE use [13], [14]. However, this result does not allow the response to be attributed exclusively to an anticipatory psychological mechanism, since physical, respiratory, and perceptual factors are simultaneously present. Thus, MOPP 4 appears to impose a less favorable perceptual starting condition even before engagement in the tasks.

The cardiovascular response was one of the most consistent manifestations of the strain produced by MOPP. The increase in heart rate was substantially greater under MOPP 2 and MOPP 4 than under MOPP 0, with a large effect size. During exercise, the cardiovascular system must simultaneously meet the metabolic demands of the active muscles and redistribute blood flow to the periphery to facilitate heat dissipation. As protective clothing increases insulation and restricts heat exchange, maintaining this balance becomes progressively more demanding [2], [15]. Sweating and the reduction in circulating volume may further increase cardiovascular demand, favoring a compensatory rise in heart rate [2], [7]. In this sense, the greater ΔHR observed under protection represents not only a statistical difference, but also a relevant increase in the cardiovascular cost required to perform equivalent physical tasks. This pattern is supported by studies showing that heated environments alter cardiovascular responses during prolonged exercise [16] and that skin temperature is related to heart rate during exercise and recovery [17].

TIR showed a different pattern from the other indicators. While MOPP 0 and MOPP 2 showed a mean reduction in forehead temperature after the physical phase, temperature remained approximately stable under MOPP 4, producing a significant difference among conditions but with a small effect size. This result should be interpreted with caution. Surface temperature responses may also vary according to exercise characteristics, as demonstrated in a protocol using different resistance-training volumes [18]. Skin surface temperature is influenced by cutaneous blood flow, sweat evaporation, and environmental conditions and does not directly correspond to core temperature or total body heat storage [8], [19]. Accordingly, the smaller reduction in TIR under MOPP 4 is compatible with an altered peripheral thermal response under greater protection, particularly because of the use of the gas mask over the facial region where the measurement was obtained, but it does not, by itself, allow inference of an equivalent increase in core temperature. The main advantage of this indicator in the present protocol was its rapid and noninvasive measurement, a relevant characteristic in operational settings, although its smaller effect magnitude indicates that it should be interpreted together with other physiological markers [3], [8].

Percentage body mass loss showed a progressive response across MOPP 0, MOPP 2, and MOPP 4, with a large effect size. This pattern is consistent with the increased thermoregulatory demand imposed by protective clothing. In the presence of increased metabolic heat production, sweating is an important heat-dissipation mechanism; however, its effectiveness depends on sweat evaporation rather than on sweat production alone [20]. The increased insulation and evaporative resistance provided by the layers of protective clothing reduce the efficiency of this exchange, such that greater sweat production may coexist with lower cooling efficiency [3]. Therefore, the progressive increase in relative body mass loss observed across MOPP levels complements the cardiovascular findings by showing that maintaining activity under higher levels of protection requires greater mobilization of the physiological mechanisms involved in thermoregulation.

The perceptual response partially followed the objective indicators. After the physical phase, absolute RPE values increased progressively from MOPP 0 to MOPP 2 and MOPP 4. However, when specifically considering the post-physical-task minus pre-exertion difference, MOPP 2 and MOPP 4 showed greater increases than MOPP 0, with no difference between them. This apparent divergence results from the fact that MOPP 4 already showed significantly elevated RPE before the tasks began. Therefore, the absence of a linear progression in ΔRPE does not indicate less discomfort under MOPP 4; rather, it demonstrates the importance of considering both the absolute value and the pre-exertion condition. Taken together, the results suggest that full protection imposes part of its perceptual burden before movement begins and maintains the highest absolute levels of perceived exertion after the activity.

From an operational perspective, these findings indicate that physically equivalent tasks do not impose the same cost on military personnel when performed under different MOPP levels. Large effects on heart rate and percentage body mass loss, together with increased perceived exertion, suggest a reduction in the physiological reserve available for continued activity as the level of protection increases. This finding is particularly relevant for prolonged missions or missions composed of successive tasks, in which accumulated strain may influence the ability to sustain subsequent efforts [2], [6], [7]. Thus, MOPP level should be incorporated into the planning of work and recovery cycles together with task intensity, duration, environmental conditions, hydration availability, and the need for personnel rotation [19]. However, the results of this study were not designed to establish new prescriptive work-rest values; rather, they demonstrate that the level of protection is a variable that should be considered when applying such protocols.

Environmental differences among collection days should also be considered when interpreting the results. Although temperature and humidity varied among sessions, each day included participants under all three MOPP conditions, distributed in a balanced manner, reducing the likelihood that a particular protection condition was systematically associated with the hottest or most humid day. Nevertheless, collection day was not included as a specific factor in the statistical analysis, and therefore the residual influence of environmental conditions on the observed responses cannot be completely excluded. In addition, the WBGT used represented a simplified estimate for comparative characterization of the

sessions because black-globe temperature was not measured. These characteristics warrant caution when extrapolating the observed values to thermal environments different from those encountered during the experiment.

Finally, several limitations define the scope of these findings. Although the sample of 26 participants exceeded the minimum number previously estimated for the study design, it consisted of physically fit military personnel familiar with PPE and assigned to a specialized CBRN unit, limiting direct generalization to untrained individuals, other military populations, or different operational conditions. TIR is a superficial and indirect measure and should not be interpreted as equivalent to core temperature [8]. In addition, post-physical-task heart rate corresponded to the mean obtained during the third and fourth minutes of measurement and therefore represents a standardized measure of early cardiovascular recovery rather than the peak immediately after exertion. These limitations do not invalidate the differences observed among MOPP levels, but they define the context in which the findings should be interpreted and reinforce the need to consider multiple indicators jointly when assessing thermophysiological strain.

V. Conclusion

The results demonstrate that the use of progressively higher MOPP levels during operational physical tasks produces relevant physiological and perceptual changes in military personnel. Compared with the MOPP 0 experimental reference condition, MOPP 2 and MOPP 4 produced greater increases in heart rate, with a large effect size, while percentage body mass loss increased progressively from MOPP 0 to MOPP 2 to MOPP 4, also with a large effect. Rating of perceived exertion showed a medium effect for the post-physical-task minus pre-exertion comparison, in addition to higher pre-exertion values under MOPP 4, indicating that full protection already imposes a perceptual burden before physical activity begins. Change in forehead infrared temperature also differed among protection levels, although with a smaller effect magnitude.

These findings show that increasing individual protection is associated with greater cardiovascular demand, greater relative body mass loss, and greater perceived exertion during the evaluated tasks. Therefore, MOPP level should be considered together with activity intensity and duration and environmental conditions when planning workload, recovery periods, and sustained personnel employment during CBRN operations. The results do not independently establish new work-rest limits, but they provide experimental evidence that progression through protective measures reduces the physiological reserve available for continued exertion.

Acknowledgment

The authors thank the 1st Chemical, Biological, Radiological, and Nuclear Defense Battalion for its operational support and for the voluntary participation of the military personnel who composed the experimental sample of this study. The authors also thank the Specialized Instruction School for the technical collaboration of its military personnel, who served as evaluators during the execution and monitoring of the experimental activities.

References

[1] Comando de Operações Terrestres (COTER), Exército Brasileiro, "Portaria nº 038/COTER, de 14 de junho de 2016," Manual de Campanha Defesa Química, Biológica, Radiológica e Nuclear (EB70-MC-10.233), 1st ed. Brasília, DF, Brazil, 2016.

[2] B. Jacklitsch, W. J. Williams, K. Musolin, A. Coca, J.-H. Kim, and N. Turner, "NIOSH criteria for a recommended standard: Occupational exposure to heat and hot environments," National Institute for Occupational Safety and Health, Cincinnati, OH, USA, Tech. Rep. DHHS (NIOSH) Pub. No. 2016-106, Feb. 2016.

[3] G. Havenith, "Heat balance when wearing protective clothing," *Ann. Occup. Hyg.*, vol. 43, no. 5, pp. 289–296, 1999, PII: S0003-4878(99)00051-4.

[4] Y. de Korte, "Beat the heat: Physiological and perceptual responses to heat stress: From Olympic athletes to healthcare workers," Ph.D. dissertation, Radboud Univ. Nijmegen, Nijmegen, The Netherlands, 2023.

[5] H. Færevik, "Impact of protective clothing on thermal and cognitive responses," Ph.D. dissertation, Dept. Biology, Norwegian Univ. Science and Technology, Trondheim, Norway, 2010.

[6] P. A. Hancock, J. M. Ross, and J. L. Szalma, "A meta-analysis of performance response under thermal stressors," *Human Factors*, vol. 49, no. 5, pp. 851–877, Oct. 2007, doi: 10.1518/001872007X230226.

[7] M. J. Stacey, "Heat stress and susceptibility to exertion-associated illnesses in military personnel," M.D. (Res.) thesis, Dept. Surgery and Cancer, Imperial College London, London, UK, 2016, doi: 10.25560/93218.

[8] V. Stanić and G. Geršak, "Facial thermal imaging: A systematic review with guidelines and measurement uncertainty estimation," *Measurement*, vol. 248, p. 115879, Dec. 2024, doi: 10.1016/j.measurement.2024.115879.

[9] G. Borg, *Borg's Perceived Exertion and Pain Scales*, Champaign, IL, USA: Human Kinetics, 1998.

[10] Headquarters, Department of the Army, "Multiservice tactics, techniques, and procedures for CBRN passive defense," U.S. Department of the Army, Washington, DC, USA, Tech. Rep. ATP 3-11.32, 2016.

[11] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[12] E. B. Neves, *Software de análise estatística para comparação de medidas entre grupos*, version 1.02. Zenodo, 2026. (Software). doi: 10.5281/zenodo.19713014.

[13] E. C. Ritchie, "Psychological problems associated with mission-oriented protective gear," *Military Medicine*, vol. 166, no. suppl_2, pp. 83–84, Feb. 2001, doi: 10.1093/milmed/166.suppl_2.83.

[14] C. H. Wick, J. A. Morrissey, and J. T. Klopcic, "Night reconnaissance operations in mission oriented protective posture," U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, USA, Tech. Rep. BRL-MR-3628, Oct. 1987.

[15] O. R. Ogunleye and W. B. Kwata, "Physiological responses to heat stress: Implications for training and performance among athletes in Nigeria," *Zamfara Int. J. Educ.*, vol. 4, no. 3, pp. 41–47, Oct. 2024, doi: 10.5281/zenodo.13929122.

[16] I. V. P. dos S. Barreto, G. D. Rodrigues, J. de S. Oliveira, E. B. Neves, and P. P. da S. Soares, "Heated environment offsets the cardiovascular responses to prolonged rowing exercise in competitive athletes," *J. Therm. Biol.,* vol. 115, p. 103603, Jul. 2023, doi: 10.1016/j.jtherbio.2023.103603.

[17] E. B. Neves, R. M. Cunha, C. Rosa, N. S. Antunes, I. M. V. Felisberto, J. Vilaça-Alves, and V. M. Reis, "Correlation between skin temperature and heart rate during exercise and recovery, and the influence of body position in these variables in untrained women," *Infrared Phys. Technol*., vol. 75, pp. 70–76, Mar. 2016, doi: 10.1016/j.infrared.2015.12.018.

[18] P. Uchôa, F. Matos, E. B. Neves, F. Saavedra, C. Rosa, V. M. Reis, and J. Vilaça-Alves, "Evaluation of two different resistance training volumes on the skin surface temperature of the elbow flexors assessed by thermography," *Infrared Phys. Technol*., vol. 93, pp. 178–183, Sep. 2018, doi: 10.1016/j.infrared.2018.07.038.

[19] E. B. Neves, "Heat stress control mechanisms in military operations," *Adv. Mil. Technol.*, vol. 17, no. 2, pp. 317–323, 2022, doi: 10.3849/aimt.01740.

[20] D. Gagnon and C. G. Crandall, "Sweating as a heat loss thermoeffector," *Handb. Clin. Neurol*., vol. 156, pp. 211–232, 2018, doi: 10.1016/B978-0-444-63912-7.00013-8.

Preliminary Dosimetric Characterization of an RS 3400 X-Ray Blood Irradiator Using LiF:Mg,Ti Thermoluminescent Dosimeters

Amanda A. Arrais de Garcia
Institute of Physics, Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
amandaalencar@discente.ufg.br

Valdeis Pires de Araújo
Radiotherapy Planning Department, Araújo Jorge Hospital
Goiânia, Goiás, Brazil
valdeispires@gmail.com

Pedro Henrique Dias Coelho
CEBROM, Oncoclínicas Group
Goiânia, Goiás, Brazil
pedrohdcoelho@gmail.com

Ana Paula Rodrigues
Hospital das Clínicas - HU Brasil, Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
ana.rodrigues@ebserh.gov.br

Jonas Oliveira da Silva
Institute of Physics, Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
jonas.silva@ufg.br

Emerson Nobuyuki Itikawa
Institute of Physics, Federal University of Goiás (UFG)
Goiânia, Goiás, Brazil
emersonitikawa@ufg.br

Abstract

Blood-component irradiation prevents transfusion-associated graft-versus-host disease by inhibiting the proliferation of viable donor lymphocytes. Independent dosimetric verification is required to confirm that the prescribed dose is delivered throughout the irradiated volume. This study reports an ongoing, preliminary characterization of an RS 3400 X-ray blood irradiator using 50 lithium-fluoride thermoluminescent dosimeters doped with magnesium and titanium. The experimental program included background evaluation, reader- and crystal-sensitivity sets, preliminary response measurements in a hemodynamic X-ray system, a dose-response experiment using cone-beam computed tomography, and nominal-setting measurements in the RS 3400 from 5 to 25 Gy. Background readings showed a mean of 2370.9 counts and a coefficient of variation (CV) of 3.24%. In the hemodynamic system, mean net response increased from 6132.4 to 21353.0 counts for reference values from 0.9789 to 3.0277 Gy, with $R^2 = 0.9949$. The cone-beam experiment produced a slope of 4186.2 counts/Gy from 2.298 to 22.813 Gy. For the RS 3400, mean net response increased from 19150.5 counts at 5 Gy to 119456.6 counts at 25 Gy, with $R^2 = 0.9984$. However, each nominal level used a different ten-detector group, and the detectors were not completely rotated among all phantom positions. Position-dependent CVs of 18.58–22.75% and relative responses of approximately 72–130% therefore combine detector, position, and possible dose-response effects. The results are reported as relative thermoluminescent response, not as absorbed-dose validation. Replicated irradiations at each level, a balanced detector-rotation design, individual sensitivity correction in a uniform field, and comparison with an independently calibrated reference dosimeter are required before the absorbed-dose distribution can be validated.

Index Terms—blood irradiation, dosimetry, quality assurance, thermoluminescent dosimeter, LiF:Mg,Ti, X-ray irradiator.

I. Introduction

Transfusion of cellular blood components is essential in several clinical settings. In susceptible recipients, however, viable donor T lymphocytes may engraft, proliferate, and attack host tissues, causing transfusion-associated graft-versus-host disease (TA-GVHD). Although uncommon, TA-GVHD is associated with high mortality and limited therapeutic response after clinical manifestation [1].

Irradiation of blood components reduces this risk by inducing irreparable damage to the genetic material of donor lymphocytes, thereby preventing cellular division while preserving the intended therapeutic function of the blood component [1]. The effectiveness of the procedure depends on delivery of an adequate dose throughout the irradiated volume. Insufficient doses may fail to inactivate lymphocytes, whereas excessive doses may impair blood-component quality.

Brazilian regulations establish 25 Gy at the midplane of the irradiated unit, with no point receiving less than 15 Gy or more than 50 Gy [4]. The current Brazilian radiation-safety standard also requires equipment control, dosimetry, and quality assurance for blood and blood-component irradiators [5]. Commissioning and routine quality assurance must therefore verify both dose magnitude and dose distribution.

Blood irradiation has traditionally been performed using self-shielded devices containing cesium-137 or cobalt-60. X-ray tube-based systems constitute an alternative that eliminates radioactive-source replacement, transportation, security, and end-of-life management requirements [6, 7]. However, the output and spatial distribution of kilovoltage irradiators depend on tube voltage, filtration, anode geometry, irradiation distance, sample attenuation, and container motion [8, 9]. Consequently, the absence of a radioactive source does not eliminate the need for independent dosimetric characterization.

The RS 3400 blood irradiator is a self-shielded X-ray device equipped with a cylindrical X-ray emitter and a rotating carousel and canister system intended to improve dose uniformity [10]. The unit installed at Hospital Araújo Jorge operates nominally at 155 kV and 11 mA and delivers a nominal central dose of approximately 25 Gy during a standard irradiation cycle.

Dose mapping is required to identify the minimum, maximum, and central doses within the irradiation container. Previous studies using ionization chambers, alanine, thermoluminescent dosimeters, radiochromic films, and blood-equivalent phantoms have demonstrated that sample geometry and attenuation can produce clinically relevant spatial variations [2, 3, 9, 11].

Radiation dosimetry is the quantitative determination of energy imparted by ionizing radiation to matter [12]. Quantities such as air kerma and absorbed dose have distinct definitions and should not be used interchangeably without an appropriate conversion formalism and clearly specified reference conditions [12, 13]. This distinction is especially important when

measurements from diagnostic-radiology instruments are used to establish a calibration in absorbed dose.

Thermoluminescent dosimeters (TLDs) are suitable for measurements in confined and nonuniform geometries because of their small dimensions, reusability, and ability to integrate dose [12, 14]. In LiF:Mg,Ti, ionizing radiation produces charge carriers that become trapped at metastable defects within the crystalline lattice. Controlled heating releases these carriers, producing a light signal related to the energy absorbed by the material [12].

The response of a thermoluminescent system may vary with individual detector sensitivity, dose, beam quality, thermal history, fading interval, reader stability, and heating profile [15]. LiF:Mg,Ti may also exhibit supralinear response in the high-dose region, requiring calibration over the actual dose interval of interest [16].

This study characterizes the response of a LiF:Mg,Ti system for application to the RS 3400. The current stage evaluates background stability, preliminary linearity, dose-response behavior, and position-dependent thermoluminescent response. The final objective is to validate the delivered absorbed dose and support a routine quality-assurance protocol. That validation remains in progress.

II. Materials and Methods

A. Dosimetric System and Irradiation Equipment

The dosimetric system comprised 50 individually identified LiF:Mg,Ti TLDs, a TLD Cube reader, a TLD Heat furnace, and the TL Studio and Thermosoft software packages. TL Studio was used to configure the reader, define the heating profile and integration region, manage calibration sets, acquire thermoluminescent signals, and export results. Thermosoft controlled the furnace temperature profiles.

Measurements associated with the blood irradiator were performed using an RS 3400 installed in the Blood Bank of Hospital Araújo Jorge. The dosimeters were positioned in a cylindrical solid-water phantom using a dedicated polylactic-acid holder containing ten locations, labeled *a* to *j*. The arrangement sampled central, intermediate, and peripheral regions of the phantom (Fig. 1).

Preliminary response measurements were performed using a Philips Azurion hemodynamic X-ray system installed at Hospital das Clínicas of the Federal University of Goiás. Exposure conditions were independently evaluated using an RTI Piranha multipurpose X-ray quality-assurance meter.

A second dose-response experiment was performed using the cone-beam computed-tomography system coupled to a Varian TrueBeam linear accelerator at Cebrom. All irradiation and measurement systems were included in institutional quality-assurance programs during data acquisition.

B. Thermal Treatment and Readout

Before each irradiation, the dosimeters were annealed in the TLD Heat furnace at $400\,^{\circ}\mathrm{C}$ for 1 h, followed by 2 h at $100\,^{\circ}\mathrm{C}$. The dosimeters were then cooled and stored in a light-protected container until irradiation.

After irradiation, a pre-readout treatment was performed at $100\,^{\circ}\mathrm{C}$ for 15 min. This step reduces the contribution of unstable low-temperature trapping states and the initial fading component. The dosimeters were subsequently read in the TLD Cube using a fixed heating profile (Fig. 2).

The calibration region of the glow curve was selected to exclude unstable fading peaks, following the TLD Cube calibration procedure recommended by the manufacturer [17]. After initiation of the calibration workflow, the heating profile, reader gain, and integration region were maintained unchanged. Detector identification, processing order, heating conditions, and the approximate interval between irradiation and readout were preserved throughout the experiments.

The net response of dosimeter i was calculated as

$$C_{\mathrm{net},i} = C_{\mathrm{raw},i} - C_{0,i}, \tag{1}$$

where $C_{\mathrm{raw},i}$ is the irradiated reading and $C_{0,i}$ is the individual background reading obtained in the zero-count set.

C. Preliminary Response Measurements

Three groups of ten TLDs were exposed in the hemodynamic system to reference values of 0.9789 ± 0.0489, 1.5015 ± 0.0751, and 3.0277 ± 0.1514 Gy. The dosimeters were positioned under reproducible geometric conditions. Tube voltage, tube loading, number of exposures, field size, filtration, and measurement distance were recorded. The Piranha system was placed at the reference measurement location to verify exposure conditions.

The quantity reported by the Piranha must be confirmed as air kerma, absorbed dose, or another application-specific quantity before this experiment can be interpreted as an absorbed-dose calibration. The present analysis therefore treats the Piranha measurements as reference values for a preliminary response evaluation, following the formalism recommended for diagnostic X-ray systems [13, 18].

An exploratory experiment was also performed at Cebrom using tube voltages from 40 to 120 kV. Because tube voltage, tube loading, dose, and beam quality changed simultaneously, those measurements were not used to establish an independent dose-response curve.

A second Cebrom experiment used a fixed nominal condition of 140 kV, 100 mA, and 65 ms. Repeated cone-beam CT acquisitions generated accumulated reference values from 2.298 to 22.813 Gy. Different dosimeters represented the dose levels; thus, dose dependence and detector-to-detector sensitivity were not separated. A common exposure of six dosimeters was also performed to evaluate short-term consistency under an identical acquisition condition.

D. RS 3400 Calibration Sets

The TL Studio workflow comprised four sets: zero counts, device sensitivity, crystal sensitivity, and system/dose calibration, following the manufacturer instructions [17].

The zero-count set was obtained after annealing all 50 non-irradiated TLDs. The device-sensitivity set was acquired after a nominal 25 Gy irradiation cycle in the RS 3400. The crystal-sensitivity set was obtained after exposing the TLDs to nominally equivalent irradiation conditions. The individual response was compared with the group mean to estimate an element correction factor:

$$\mathrm{ECF}_i = \frac{\overline{C}_{\mathrm{net}}}{C_{\mathrm{net},i}}, \tag{2}$$

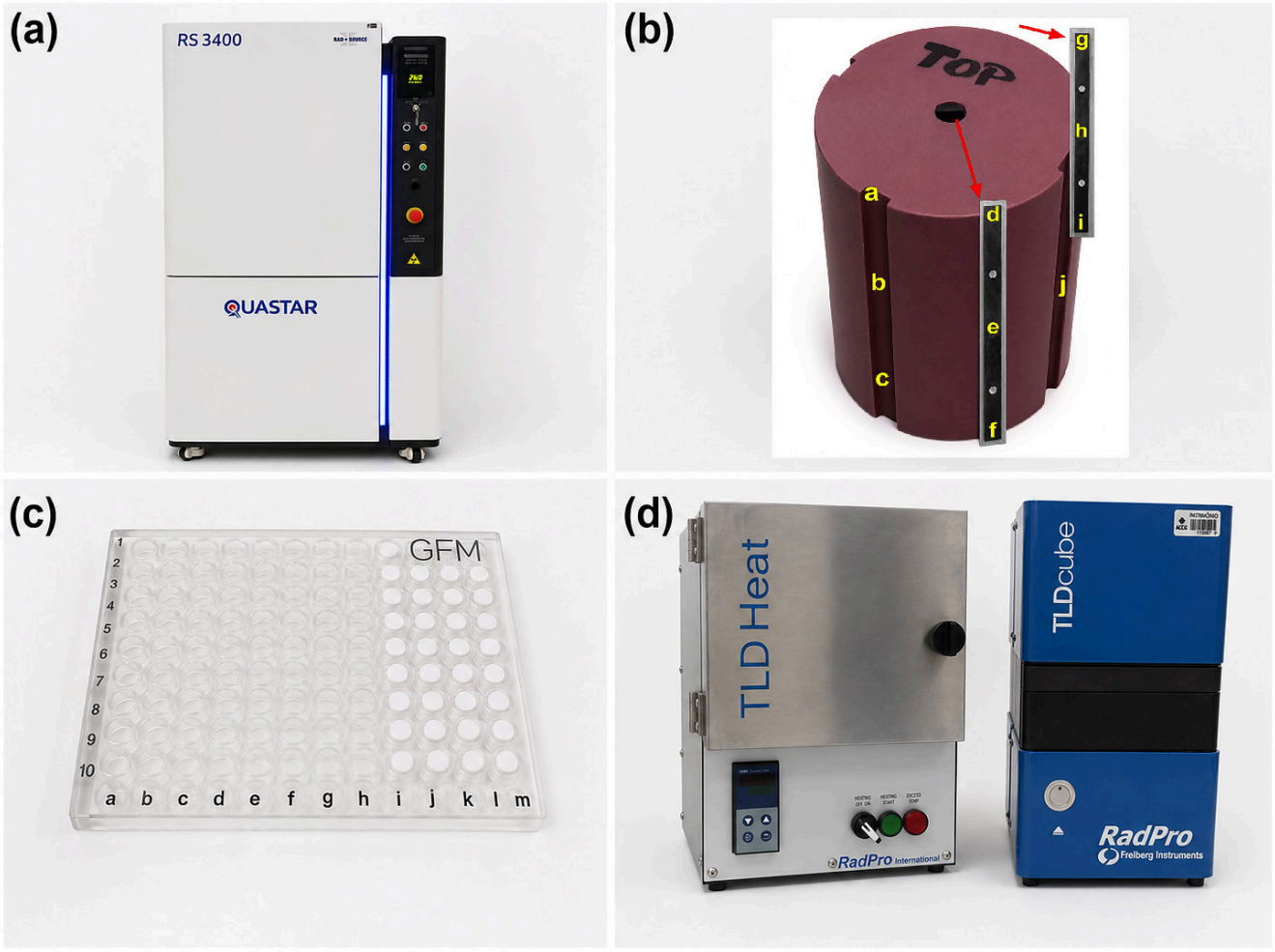


Figure 1. Experimental system used in the study: (a) RS 3400 blood irradiator; (b) cylindrical phantom and dosimeter positions; (c) LiF:Mg,Ti dosimeters; and (d) thermal-treatment and readout equipment.

and the corrected response would be

$$C_{\mathrm{corr},i} = \mathrm{ECF}_i C_{\mathrm{net},i}. \tag{3}$$

Because the TLDs occupied different phantom positions during these sets, the calculated factors may contain both individual-sensitivity and spatial-irradiation components. They were therefore not treated as definitive element correction factors in the present analysis.

For system/dose calibration, the 50 TLDs were distributed into five groups of ten. The groups were irradiated with programmed nominal settings of 5, 10, 15, 20, and 25 Gy by varying irradiation time according to the nominal delivery rate of the RS 3400. Each group occupied the same ten phantom positions. This first-pass design did not rotate each detector through all ten positions and did not include independent replicate irradiation cycles at each nominal level. Consequently, detector, position, and dose-level effects cannot be estimated separately from these data.

For dose level j, mean apparent sensitivity was calculated as

$$S_j = \frac{\overline{C}_{\mathrm{net},j}}{D_j}, \tag{4}$$

where D_j is the programmed nominal setting. A linear model was fitted according to

$$C_{\mathrm{net}} = aD + b. \tag{5}$$

To compare phantom positions across dose groups, each response was normalized to the mean of its respective group:

$$R_{i,j} = \frac{C_{\mathrm{net},i,j}}{\overline{C}_{\mathrm{net},j}} \times 100\%. \tag{6}$$

The resulting quantity was interpreted as relative thermoluminescent response, not as relative absorbed dose.

E. Beam Quality and Final Validation Strategy

A complete characterization of an X-ray field includes tube potential, filtration, half-value layer, homogeneity coefficient, output, spatial uniformity, and measurement uncertainty [13, 18, 19]. The beam-quality stage of the RS 3400 characterization remains in progress.

The final validation protocol will include first and second half-value layers, the homogeneity coefficient, replicated dose-response measurements at every nominal level, individual sen-

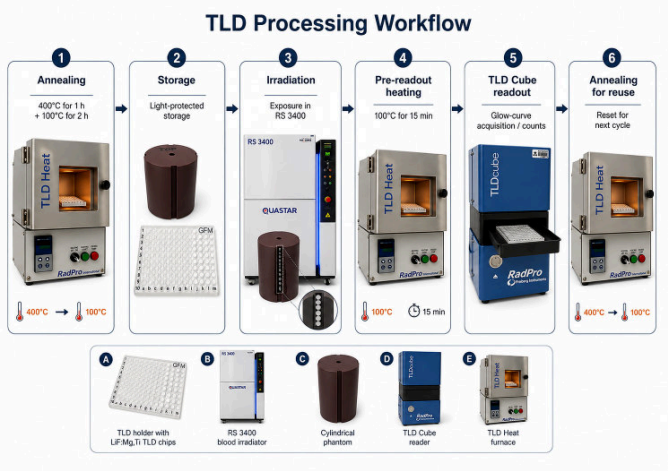


Figure 2. Thermal processing and readout workflow applied to the LiF:Mg,Ti dosimeters.

sitivity correction in a uniform field, and a balanced complete-block rotation in which detectors occupy all phantom positions across repeated cycles. A fixed-dose control group will monitor between-session stability. An independently calibrated reference dosimeter will be used to establish traceability and to compare the central reference value with the TLD-derived result. Only after these steps will corrected responses be converted into absorbed dose and used to determine central, minimum, and maximum values.

F. Statistical Analysis and Scope

For each experimental condition, the mean, sample standard deviation, CV, minimum, and maximum were calculated. Least-squares regression and the coefficient of determination, R^2, were used to evaluate dose-response behavior. Potential outliers were investigated at a significance level of 5% and were not excluded solely by visual inspection. When a potentially discrepant reading was identified, the analysis was performed with and without that result to evaluate its influence.

Because no complete rotation, replicated cycle at each RS 3400 setting, or independent absorbed-dose reference was available, the primary outcome of the present analysis is detector response in counts. No claim of regulatory conformity, dose-uniformity ratio, or validated absorbed-dose distribution is made from this preliminary dataset.

III. Results and Discussion

A. Background and Initial Calibration Sets

The zero-count set presented a mean of 2370.9 counts, an SD of 76.7 counts, and a CV of 3.24%, with values ranging from 2261 to 2511 counts (Table 1). The limited dispersion indicates satisfactory overall background stability; however, a systematic decrease was observed along detector identification and readout sequence. Since identification approximately followed acquisition order, this behavior may reflect reader stabilization, temporal photomultiplier variation, thermal-history effects, or another order-dependent component. Individual background subtraction was therefore preferred to correction by a single batch mean.

The device-sensitivity set presented a mean of 130045.5 counts and a CV of 19.14%. Crystal-sensitivity readings averaged 123166.1 counts, with a CV of 24.73%. One crystal-sensitivity result, 251944 counts for TLD 41, was markedly greater than the remaining readings. Omitting this value only for sensitivity analysis reduced the mean to 120538.0 counts and the CV to 20.23%. The result was not definitively excluded because it has not been repeated under an independently uniform radiation field.

Table 1. Summary of the Initial RS 3400 Calibration Sets

Set	Mean	SD	CV (%)	Range
Zero counts	2370.9	76.7	3.24	2261–2511
Device sensitivity	130045.5	24894.5	19.14	78634–173824
Crystal sensitivity	123166.1	30459.7	24.73	75581–251944
Crystal, TLD 41 omitted	120538.0	24384.0	20.23	75581–165274

Table 2. TLD Response in the Hemodynamic X-Ray System

Reference (Gy)	Mean net counts	SD counts	CV (%)	Sensitivity (counts/Gy)
0.9789	6132.4	231.0	3.77	6264.6
1.5015	8977.0	587.2	6.54	5978.8
3.0277	21353.0	883.4	4.14	7052.6

Excluding the provisional TLD 41 result, the correlation between device- and crystal-sensitivity readings was $r = 0.915$. The response pattern was not purely random; however, the dosimeters occupied different phantom positions during both sets. The correlation therefore includes both intrinsic detector sensitivity and spatial variation. Element correction factors should be determined under uniform and reproducible irradiation conditions [12, 15, 17].

B. Hemodynamic-System Response

Table 2 presents net responses obtained using the hemodynamic system. Within-group CV values ranged from 3.77% to 6.54%, indicating satisfactory repeatability under each condition.

The regression of the three mean responses was

$$C_{\text{net}} = 7582.3D - 1767.0, \tag{7}$$

with $R^2 = 0.9949$. When all 30 individual measurements were included, R^2 remained approximately 0.987. Apparent sensitivity varied from 5978.8 to 7052.6 counts/Gy. Since different dosimeters were used at each reference level and the precise Piranha quantity remains to be documented, these results confirm a consistent response at doses on the order of gray but do not establish a traceable counts-to-absorbed-dose conversion.

C. Cone-Beam Dose-Response Experiment

In the first Cebrom experiment, tube potential, tube loading, reference value, and beam quality changed simultaneously. Although net counts increased with the recorded value, the experiment could not distinguish dose dependence from energy dependence and was therefore treated as exploratory.

Under the fixed nominal condition of 140 kV, 100 mA, and 65 ms, net response ranged from 5509 to 90786 counts for reference values from 2.298 to 22.813 Gy. The linear fit was

$$C_{\text{net}} = 4186.2D - 3943.2, \tag{8}$$

with $R^2 = 0.9996$. The fitted slope differed by approximately 1.8% from the mean manufacturer-derived sensitivity of 4113.4 counts/Gy and was nearly equal to the slope of 4186.1 counts/Gy

Table 3. RS 3400 Response as a Function of Programmed Nominal Setting

Nominal setting (Gy)	Mean net counts	SD counts	CV (%)	Apparent sensitivity
5	19150.5	3591.8	18.76	3830.1
10	42224.1	8958.7	21.22	4222.4
15	65478.8	13437.7	20.52	4365.3
20	92720.1	17225.6	18.58	4636.0
25	119456.6	27178.3	22.75	4778.3

obtained by fitting the manufacturer data through the origin. This agreement supports the order of magnitude of the measured response.

The negative intercept has no direct physical interpretation at zero dose. Apparent sensitivity increased from 2397.3 counts/Gy at 2.298 Gy to approximately 3998.4 counts/Gy at 14.596 Gy and remained close to 3980 counts/Gy at 22.813 Gy. A high R^2 does not demonstrate strict proportionality; the trend may reflect background influence at the lowest levels, detector-to-detector sensitivity, uncertainty in the reference quantity, or supralinearity of LiF:Mg,Ti [12, 15, 16].

A common exposure of six TLDs yielded a mean recorded value of 2.381 Gy with a CV of 0.99% and mean raw counts of 8296.3 with a CV of 1.20%. These results indicate good short-term consistency under a fixed acquisition condition. However, mean net sensitivity was approximately 2436 counts/Gy, lower than the complete-curve slope. An interrupted initial acquisition also produced a different exposure history for TLDs 8 and 9. The curve must therefore be repeated using several TLDs per level and an identical acquisition sequence.

D. RS 3400 Response as a Function of Nominal Setting

Table 3 summarizes the first-pass system/dose calibration after individual background subtraction. Mean net response increased from 19150.5 counts at 5 Gy to 119456.6 counts at 25 Gy. The linear fit shown in Fig. 3 was

$$C_{\mathrm{net}} = 5022.2D - 7526.4, \tag{9}$$

with $R^2 = 0.9984$. A quadratic model increased R^2 to approximately 0.9998, indicating slight upward curvature. However, only five nominal levels were available, and each was represented by a different ten-detector group occupying the ten phantom positions. The present data cannot distinguish supralinearity from detector and position effects.

Mean apparent sensitivity across the five levels was 4366.4 counts/Gy, approximately 6.1% greater than the mean manufacturer-derived value, and increased by 24.8% between 5 and 25 Gy. This 24.8% change cannot be attributed to supralinearity because detector assignment and position changed with dose level. It is therefore described only as a change in apparent sensitivity in this preliminary design.

Mean background represented approximately 11.5% of raw response at 5 Gy and 1.9% at 25 Gy, demonstrating that individualized background correction is particularly relevant at the lower end of the investigated interval.

The programmed levels were derived from RS 3400 irradiation time and were not independently established using a traceable absorbed-dose standard. The fitted curve characterizes thermoluminescent response as a function of programmed nominal setting; it is not an absolute absorbed-dose calibration.

Table 4. Mean Relative Response at the Phantom Positions

Position	Mean (%)	SD (%)	Position	Mean (%)	SD (%)
a	86.1	12.5	f	73.4	4.0
b	119.3	6.9	g	102.9	4.9
c	95.4	4.4	h	129.6	4.7
d	71.7	3.4	i	97.9	4.0
e	101.5	2.1	j	122.2	4.3

E. Position Dependence of the Response

Within-level CV values from 18.58% to 22.75% were substantially greater than the dispersion observed within the hemodynamic groups. Normalization by the mean of each nominal group revealed a recurring position-dependent response pattern (Table 4 and Fig. 4).

Positions *d* and *f* presented the lowest relative responses, 71.7% and 73.4%, respectively. Positions *h*, *j*, and *b* presented the highest responses, 129.6%, 122.2%, and 119.3%. The central position *e* presented 101.5%, close to the group mean. The difference between the lowest and highest position means was approximately 58 percentage points.

Position-specific regressions produced R^2 values from 0.981 to 0.998, while their slopes varied by more than a factor of two. Within the five nominal-setting groups, the normalized means of paired positions *d*/*f* and *c*/*i* differed by 1.7 and 2.5 percentage points, whereas *b*/*h* and *a*/*g* differed by 10.3 and 16.8 percentage points. In the pooled 25 Gy device/crystal sets, relative pair differences for *b*/*h*, *d*/*f*, *a*/*g*, and *c*/*i* were approximately 0.2%, 3.8%, 5.3%, and 6.0%, respectively. Thus, geometric symmetry was only partial and cannot by itself establish irradiator uniformity.

The recurring pattern is compatible with a contribution from irradiation geometry, target orientation, attenuation, and the anode heel effect [8]. Dose distribution also depends on sample composition, container geometry, irradiation motion, and attenuation within the load [3, 9, 11]. Rotation and opposing irradiation directions can reduce, but do not necessarily eliminate, differences between minimum and maximum dose.

Nevertheless, the present result is a spatial variation of thermoluminescent response, not a definitive absorbed-dose map. The same detectors were not systematically rotated through all positions, and the individual correction factors were not determined in an independently uniform field. The observed response combines local dose, individual sensitivity, thermal processing, readout variability, and possible energy dependence. Attribution of the complete variation to irradiator nonuniformity is not justified at this stage.

F. Comparison with Previous Studies

Grasso *et al.* characterized an X-ray blood irradiator using ionization chamber, alanine, and TLD-100 dosimetry and reported a small TLD dose overestimation together with a dose-uniformity ratio consistent with the manufacturer specification [3]. Their use of multiple dosimetric systems illustrates the importance of independent cross-validation.

Lopes *et al.* described physical measurements required for routine quality control of blood irradiation, including absorbed-dose verification, radiation surveys, and documentation of equipment performance [2]. Fazad *et al.* used Gafchromic EBT-XD film and a blood-equivalent phantom to map central, minimum,

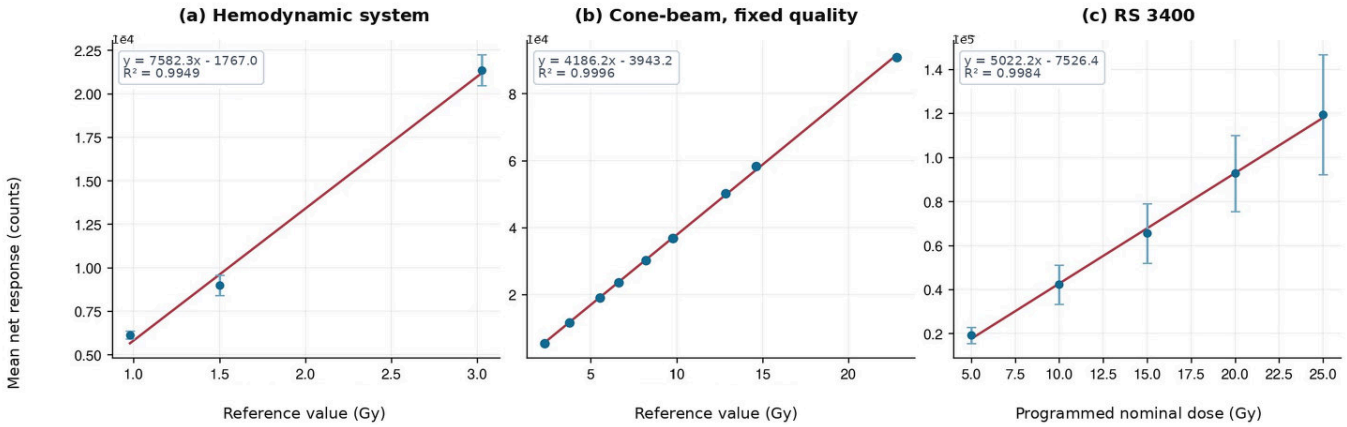


Figure 3. Dose-response results obtained in (a) the hemodynamic X-ray system, (b) the Cebrom cone-beam system under fixed nominal quality, and (c) the RS 3400. Axes, regression notation, decimal format, and equation placement were standardized. Error bars represent one SD where replicate measurements were available. The abscissas in (a) and (b) are reported as reference values because the associated quantity has not yet been confirmed as absorbed dose.

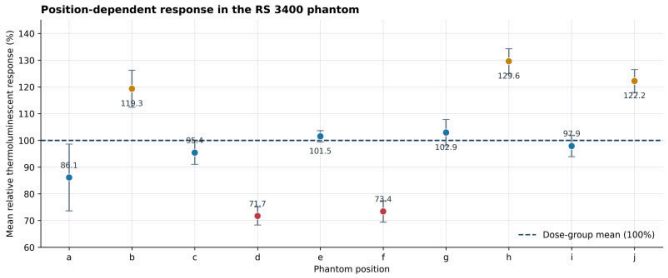


Figure 4. Mean thermoluminescent response at the ten phantom positions, normalized to the mean response of each nominal-setting group. Error bars represent the SD among the five groups. These values combine detector and position effects and must not be interpreted as a definitive absorbed-dose map.

and maximum doses in a cesium-137 irradiator [11]. Gómez-Simuta *et al.* demonstrated that dose mapping can identify reproducible minimum- and maximum-dose regions in an X-ray blood irradiator and emphasized routine dosimetry at a defined reference location [9]. Wagner *et al.* likewise showed that a cylindrical-anode Rad Source irradiator presents position-dependent output and requires experimental characterization of beam quality and dose rate [8].

These studies support the overall methodology but demonstrate that a high R^2 in a dose-response curve is insufficient for validation. Traceability, energy dependence, individual sensitivity, spatial mapping, reproducibility, and uncertainty must be considered jointly.

G. Limitations and Ongoing Work

This study reports an intermediate stage of dosimetric characterization. Its principal limitation is the absence of a complete independent and traceable calibration over the operating range of the RS 3400.

In the Cebrom experiment, each dose level was represented by one TLD, making detector-to-detector sensitivity inseparable from dose dependence. In the RS 3400 experiment, each programmed setting was represented by a different detector group, without independent replicate cycles at each level. The device- and crystal-sensitivity sets were obtained with TLDs distributed among positions that presented substantial response differences. The resulting correction factors may therefore contain both individual detector sensitivity and spatial irradiation effects.

Session-to-session normalization was not performed using a fixed-dose control group in every experiment. Changes in reader gain, thermal processing, fading interval, or irradiation output may contribute to comparisons among measurements performed on different dates and equipment. The precise reference quantities supplied by the Piranha and cone-beam systems must also be documented; air kerma, absorbed dose, and CT-related quantities are not equivalent and may only be converted under defined reference conditions [12, 13].

Beam-quality measurements, including first and second half-value layers and the homogeneity coefficient, have not yet been completed. The analysis also lacks a complete measurement-uncertainty budget. Future work will include contributions from the reference instrument, irradiation reproducibility, reader stability, annealing and pre-readout treatments, positioning, fading, background correction, individual sensitivity, regression coefficients, and beam-quality dependence.

The anomalous reading of 251944 counts for TLD 41 must be repeated before definitive exclusion. A statistical outlier test should be accompanied by investigation of experimental causes, including detector identification, positioning, incomplete annealing, and reader malfunction.

The next experimental stage will use several TLDs and independent irradiation cycles at every nominal setting, with a fixed-dose control group in each session. A balanced rotation will move the same detectors through all ten phantom positions, allowing detector and position effects to be estimated separately. Element correction factors will be established in a uniform reference field. Standard RS 3400 cycles will then be repeated and compared with an independently calibrated, traceable reference dosimeter. Only after completion of these stages will central, minimum, and maximum absorbed doses be compared with the regulatory interval and manufacturer uniformity specification.

IV. Conclusion

The LiF:Mg,Ti system exhibited stable overall background and an increasing response with exposure in three X-ray systems. Hemodynamic-system measurements showed satisfactory within-condition repeatability, and the fixed-quality cone-beam

experiment produced a dose-response slope consistent with the order of magnitude of manufacturer-derived reference data.

In the RS 3400, mean thermoluminescent response was strongly associated with the programmed nominal setting from 5 to 25 Gy. However, apparent sensitivity increased with setting, and each level presented substantial spatial dispersion. A recurring position-dependent pattern was observed, with relative responses from approximately 72% to 130% of the group mean, but geometric symmetry was only partial.

Because dose level, detector group, and position were not independently resolved, this variation cannot be assigned exclusively to irradiator nonuniformity or to LiF:Mg,Ti supralinearity. The results support the feasibility of thermoluminescent dosimetry for characterization and quality assurance of the RS 3400, but they do not constitute absorbed-dose validation or demonstrate regulatory conformity.

Definitive conclusions regarding dose accuracy, distribution, uniformity, and reproducibility require replicated measurements at each level, individual sensitivity correction, complete detector rotation, beam-quality characterization, uncertainty analysis, and independent reference dosimetry.

Acknowledgment

The authors thank the Blood Bank and Radiation Therapy Department of Hospital Araújo Jorge, Hospital das Clínicas of the Federal University of Goiás, and Cebrom for providing access to the irradiation and dosimetric systems used in this study.

References

[1] G. Moroff and N. L. C. Luban, "The irradiation of blood and blood components to prevent graft-versus-host disease: Technical issues and guidelines," *Transfus. Med. Rev.*, vol. 11, no. 1, pp. 15–26, 1997, doi: 10.1016/S0887-7963(97)80006-5.

[2] F. Lopes *et al.*, "Quality control in blood irradiation," *Braz. J. Radiat. Sci.*, vol. 11, no. 1A, pp. 1–17, 2023, doi: 10.15392/2319-0612.2023.2183.

[3] S. Grasso *et al.*, "Absorbed dose evaluation of a blood irradiator with alanine, TLD-100 and ionization chamber," *Appl. Radiat. Isot.*, vol. 200, Art. no. 110981, 2023, doi: 10.1016/j.apradiso.2023.110981.

[4] Ministry of Health, "Consolidation Ordinance GM/MS No. 5 of September 28, 2017: Consolidation of standards on actions and health services of the Brazilian Unified Health System," Brasília, Brazil, 2017, in Portuguese.

[5] Autoridade Nacional de Segurança Nuclear, *Norma ANSN 6.16: Requisitos de Segurança e Proteção Radiológica para Irradiadores de Sangue e Hemocomponentes*, Resolução CNEN 330/24, Brasília, Brazil, Nov. 2024, in Portuguese.

[6] J. Kamen, W.-Y. Hsu, B. Boswell, and C. Hill, "Successful migration from radioactive irradiators to X-ray irradiators in one of the largest medical centers in the US," *Health Phys.*, vol. 117, no. 5, pp. 558–570, 2019, doi: 10.1097/HP.0000000000001095.

[7] K. Mehta and A. Parker, "Characterization and dosimetry of a practical X-ray alternative to self-shielded gamma irradiators," *Radiat. Phys. Chem.*, vol. 80, no. 1, pp. 107–113, 2011, doi: 10.1016/j.radphyschem.2010.08.011.

[8] J. K. Wagner, J. A. Dillon, E. K. Blythe, and J. R. Ford, "Dose characterization of the Rad Source 2400 X-ray irradiator for oyster pasteurization," *Appl. Radiat. Isot.*, vol. 67, no. 2, pp. 334–339, 2009, doi: 10.1016/j.apradiso.2008.10.013.

[9] Y. Gómez-Simuta, A. Parker, C. Cáceres, M. J. B. Vreysen, and H. Yamada, "Characterization and dose-mapping of an X-ray blood irradiator to assess application potential for the sterile insect technique," *Appl. Radiat. Isot.*, vol. 176, Art. no. 109859, 2021, doi: 10.1016/j.apradiso.2021.109859.

[10] Rad Source Technologies, *The RS 3400 Blood Irradiator*. Buford, GA, USA: Rad Source Technologies, product brochure, n.d.

[11] A. S. Fazad, N. M. Noor, and M. A. Nasir, "Dose map blood irradiator dosimetry system: Customized blood equivalent phantom and Gafchromic EBT-XD film," *Asian J. Med. Technol.*, vol. 3, no. 2, pp. 44–54, 2023.

[12] F. H. Attix, *Introduction to Radiological Physics and Radiation Dosimetry*. Weinheim, Germany: Wiley-VCH, 2004.

[13] International Atomic Energy Agency, *Dosimetry in Diagnostic Radiology: An International Code of Practice*, Technical Reports Series No. 457. Vienna, Austria: IAEA, 2007.

[14] A. J. J. Bos, "High sensitivity thermoluminescence dosimetry," *Nucl. Instrum. Methods Phys. Res. B*, vol. 184, no. 1–2, pp. 3–28, 2001, doi: 10.1016/S0168-583X(01)00717-0.

[15] L. C. Matsushima, "Determinação das curvas de isodose e confirmação do planejamento em radioterapia de intensidade modulada empregando as técnicas de termoluminescência, luminescência opticamente estimulada e detectores semicondutores," Ph.D. dissertation, Instituto de Pesquisas Energéticas e Nucleares, Universidade de São Paulo, São Paulo, Brazil, 2015, in Portuguese.

[16] R. Liuzzi, F. Savino, V. D'Avino, M. Pugliese, and L. Cella, "Evaluation of LiF:Mg,Ti (TLD-100) for intraoperative electron radiation therapy quality assurance," *PLOS ONE*, vol. 10, no. 10, Art. no. e0139287, 2015, doi: 10.1371/journal.pone.0139287.

[17] RadPro International GmbH, *Calibration Instructions for TLD Cube with TL Studio: Calibration Procedure—Step by Step*. Wermelskirchen, Germany: RadPro International GmbH, June 2015.

[18] F. M. Bastos, "Caracterização de feixes de raios X de baixa energia com uma câmara de extrapolação," M.S. thesis, Centro de Desenvolvimento da Tecnologia Nuclear, Comissão Nacional de Energia Nuclear, Belo Horizonte, Brazil, 2015, in Portuguese.

[19] International Organization for Standardization, *Radiological Protection—X and Gamma Reference Radiation for Calibrating Dosemeters and Dose Rate Meters and for Determining Their Response as a Function of Photon Energy—Part 1: Radiation Characteristics and Production Methods*, ISO 4037-1:2019. Geneva, Switzerland: ISO, 2019.

Regulação Assistencial Inteligente no SUS: Impactos da Saúde Digital na Arquitetura e Gestão do Acesso

1st Maria Eduarda da Silva Ferreira
Escola de Ciências e Tecnologia (ECT)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0006-5258-7757

2nd Liana Lysia Cavalcanti Vieira de Brito
Escola de Ciências e Tecnologia (ECT)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0001-2928-7522

3rd Joêmilly Wellen Lima da Silva
Departamento de Engenharia Biomédica (DEB)
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0009-0004-3464-0070

4th Antonio de Oliveira Nunes Costa
Mestrando em Engenharia Elétrica e de Computação
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0000-0003-2970-6245

5th Ricardo Alexsandro de Medeiros Valentim
Professor Titular do Departamento de Engenharia Biomédica e LAIS
Universidade Federal do Rio Grande do Norte (UFRN)
Natal, Brasil
https://orcid.org/0000-0002-9216-8593

***Resumo*—A transformação digital tem promovido mudanças estruturais nos sistemas de saúde contemporâneos, especialmente na organização do acesso assistencial. No Sistema Único de Saúde (SUS), a regulação assistencial constitui instrumento estratégico para garantir equidade, integralidade e eficiência na distribuição de recursos públicos. Este estudo analisa os impactos da saúde digital na regulação assistencial brasileira, considerando a Estratégia de Saúde Digital para o Brasil 2020–2028, a Rede Nacional de Dados em Saúde (RNDS) e experiências práticas como o ecossistema RegulaRN. Trata-se de estudo qualitativo, exploratório e descritivo, desenvolvido por revisão narrativa da literatura científica e documental. Os resultados demonstram que a adoção de interoperabilidade clínica, monitoramento em tempo real, algoritmos de priorização e integração de sistemas contribui para reduzir assimetrias informacionais, qualificar decisões regulatórias e ampliar a transparência institucional. Persistem, entretanto, desafios relacionados à infraestrutura digital, capacitação profissional, proteção de dados e desigualdades regionais. Conclui-se que a saúde digital representa elemento estruturante para modernização da regulação assistencial do SUS, com relevante interface com a Engenharia Biomédica.**

***Palavras-chave*—Saúde Digital, SUS, Regulação Assistencial, Interoperabilidade, Engenharia Biomédica, RNDS.**

I. Introdução

O Sistema Único de Saúde (SUS) configura uma das maiores redes públicas universais de saúde do mundo, sendo responsável pela assistência de milhões de brasileiros em diferentes níveis de complexidade [1]. Sua magnitude organizacional exige instrumentos capazes de ordenar fluxos, priorizar demandas e otimizar o uso de recursos assistenciais. Nesse contexto, a regulação assistencial tornou-se componente essencial da governança sanitária, sendo formalmente fortalecida pela Política Nacional de Regulação, instituída pela Portaria nº 1.559/2008 [3]. Tal política busca organizar o acesso a consultas especializadas, exames, internações e leitos críticos, promovendo maior racionalidade distributiva. Historicamente, entretanto, os processos regulatórios brasileiros enfrentaram limitações relacionadas à fragmentação dos sistemas de informação, filas descentralizadas, baixa integração entre níveis assistenciais e dependência de processos administrativos manuais [4]. Essas barreiras impactam diretamente a eficiência operacional e a justiça no acesso.

A saúde digital emerge como eixo estratégico de superação dessas limitações, ao incorporar tecnologias da informação, telessaúde, interoperabilidade clínica, análise de dados e apoio computacional à decisão [5], [6]. No campo regulatório, tais soluções ampliam a rastreabilidade, transparência e capacidade preditiva.

Além disso, a Engenharia Biomédica possui papel central nesse processo, ao integrar conhecimentos de tecnologia, gestão e sistemas clínicos aplicados à saúde pública. Diante disso, este artigo objetiva analisar o impacto da saúde digital na regulação assistencial do Sistema Único de Saúde (SUS), enfatizando aspectos relacionados à arquitetura tecnológica, à interoperabilidade e à gestão estratégica. Como contribuição, o estudo busca sistematizar e discutir os principais elementos tecnológicos e organizacionais que permeiam a transformação digital da regulação assistencial, evidenciando os avanços,

desafios e perspectivas para a construção de sistemas mais integrados, eficientes e orientados às necessidades dos usuários. Dessa forma, pretende-se contribuir para o debate sobre a aplicação estratégica das tecnologias digitais no SUS e para o fortalecimento de processos regulatórios mais qualificados, interoperáveis e equitativos.

II. Metodologia

Trata-se de um estudo qualitativo, descritivo e exploratório, desenvolvido por meio de revisão narrativa da literatura científica e documental. Optou-se pela revisão narrativa por possibilitar uma análise ampla e integrativa da produção científica acerca da saúde digital aplicada à regulação assistencial do SUS, permitindo reunir evidências provenientes de diferentes tipos de estudos, documentos técnicos e normativos.

A busca bibliográfica foi realizada entre março e abril de 2026 nas bases PubMed, Scopus, SciELO e Google Scholar, complementada pela consulta a documentos oficiais do Ministério da Saúde, da Estratégia de Saúde Digital para o Brasil 2020–2028 e da Rede Nacional de Dados em Saúde (RNDS). Os descritores foram combinados por operadores booleanos (AND e OR), utilizando termos em português e inglês relacionados à saúde digital, interoperabilidade, regulação assistencial, RNDS e HL7 FHIR.

Após a leitura dos estudos selecionados por meio dos critérios exibidos na tabela I, realizou-se análise qualitativa do conteúdo, agrupando as evidências conforme similaridade temática. Os achados foram organizados em três eixos analíticos: (i) arquitetura digital e interoperabilidade; (ii) impactos operacionais na regulação assistencial; e (iii) desafios, bioética e perspectivas futuras. Por fim, como complemento à revisão bibliográfica, foi incorporado um estudo de caso descritivo do ecossistema RegulaRN, escolhido por representar uma experiência consolidada de aplicação da saúde digital na regulação assistencial do SUS.

Tabela I
Critérios de inclusão e exclusão utilizados na seleção dos estudos

Critérios de inclusão	**Critérios de exclusão**
Estudos publicados entre 2020 e 2026	Trabalhos duplicados
Artigos completos	resumos sem texto completo disponível
Estudos em português, inglês e espanhol	Estudos sem relação com a regulação assistencial
Saúde digital aplicada ao SUS	publicações opinativas sem base técnica
Documentos normativos oficiais	trabalhos anteriores a 2020 sem relevância histórica central

III. Saúde Digital e Regulação Assistencial

A. Transformação digital no sistema único de saúde

A transformação digital tem promovido mudanças significativas na organização dos sistemas de saúde, possibilitando maior integração entre serviços, otimização dos processos assistenciais e qualificação da tomada de decisão. No Sistema Único de Saúde (SUS), esse processo representa importante estratégia para enfrentar desafios históricos relacionados à fragmentação das informações, à baixa integração entre os níveis de atenção e à elevada demanda por serviços especializados [2], [5], [6].

Nesse contexto, a incorporação de tecnologias digitais favorece a construção de um modelo assistencial baseado no compartilhamento de informações, no monitoramento em tempo real e no uso de dados para subsidiar decisões clínicas e gerenciais. Além de ampliar a eficiência operacional, essas ferramentas contribuem para maior transparência, rastreabilidade dos processos regulatórios e utilização mais racional dos recursos públicos [4], [6].

B. Estratégia de Saúde Digital para o Brasil

A Estratégia de Saúde Digital para o Brasil 2020–2028 estabelece diretrizes nacionais voltadas à transformação digital do setor saúde, promovendo ações relacionadas à integração dos sistemas de informação, fortalecimento da governança digital, ampliação da conectividade e desenvolvimento de serviços centrados no cidadão [2].

No âmbito da regulação assistencial, essa estratégia possibilita que as decisões regulatórias sejam fundamentadas em informações clínicas atualizadas, completas e interoperáveis. Dessa forma, torna-se possível reduzir assimetrias informacionais entre os diferentes pontos da Rede de Atenção à Saúde, minimizar retrabalhos administrativos e aprimorar os mecanismos de priorização do acesso aos serviços especializados [2], [6]. Além disso, o uso de indicadores assistenciais em tempo real favorece o monitoramento da oferta de serviços, permitindo aos gestores identificar gargalos assistenciais, redistribuir recursos e aperfeiçoar continuamente os fluxos regulatórios [6].

C. Rede Nacional de Dados em Saúde (RNDS)

A Rede Nacional de Dados em Saúde (RNDS) constitui a principal infraestrutura nacional para o compartilhamento seguro e padronizado de informações clínicas entre estabelecimentos públicos e privados de saúde [2], [5]. Sua implementação possibilita a interoperabilidade entre diferentes sistemas por meio da adoção de padrões tecnológicos e semânticos, permitindo que dados produzidos em unidades básicas de saúde, hospitais, laboratórios e serviços especializados sejam compartilhados de maneira segura e tempestiva [5], [7], [13].

Essa integração favorece a continuidade do cuidado, reduz a duplicidade de exames, melhora a qualidade das informações disponíveis aos profissionais de saúde e fortalece a capacidade das centrais reguladoras em realizar encaminhamentos baseados em evidências clínicas [5], [7]. Sob a perspectiva da regulação assistencial, a disponibilidade de informações clínicas consolidadas reduz incertezas durante o processo de classificação de pacientes, favorecendo maior equidade na distribuição de consultas, exames e leitos hospitalares [4], [7].

D. Interface com a Engenharia Biomédica

A implementação das tecnologias de saúde digital demanda conhecimentos multidisciplinares relacionados à arquitetura de sistemas, interoperabilidade, integração de dispositivos médicos, segurança cibernética, análise de dados clínicos e desenvolvimento de soluções computacionais para apoio à decisão [5], [7], [13]. Nesse cenário, a Engenharia Biomédica desempenha papel estratégico ao integrar competências tecnológicas e assistenciais voltadas ao desenvolvimento de ecossistemas digitais capazes de melhorar a eficiência da gestão em saúde. O desenvolvimento de arquiteturas interoperáveis, a validação de softwares médicos, a integração entre equipamentos biomédicos e sistemas hospitalares e a aplicação de inteligência artificial representam áreas diretamente relacionadas à modernização da regulação assistencial do SUS [7], [13], [14].

IV. Arquitetura Tecnológica e Interoperabilidade

A consolidação da Saúde Digital no Sistema Único de Saúde (SUS) depende da implementação de arquiteturas tecnológicas capazes de integrar os diferentes níveis da Rede de Atenção à Saúde, promovendo o compartilhamento seguro e oportuno das informações clínicas. A coordenação eficiente dos fluxos assistenciais exige que unidades básicas de saúde, hospitais, laboratórios, serviços especializados e centrais reguladoras atuem de forma integrada, favorecendo a continuidade do cuidado e a qualificação dos processos regulatórios [15].

Nesse contexto, a modernização da regulação assistencial requer arquiteturas escaláveis, seguras e adaptáveis às diferentes realidades regionais. Estudos recentes propõem modelos multicamadas fundamentados na integração de sistemas heterogêneos por meio de interfaces de programação de aplicações (APIs), microsserviços e padrões internacionais de interoperabilidade, possibilitando maior flexibilidade, disponibilidade e compartilhamento de informações entre diferentes plataformas de saúde [7].

Entre os principais padrões tecnológicos utilizados destaca-se o Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR), reconhecido internacionalmente por sua arquitetura baseada em recursos (resources) e APIs RESTful. Essa abordagem permite que diferentes sistemas de informação em saúde troquem dados clínicos de forma padronizada, reduzindo barreiras técnicas e facilitando a integração entre prontuários eletrônicos, sistemas laboratoriais, plataformas de telemedicina e infraestruturas nacionais como a Rede Nacional de Dados em Saúde (RNDS) [13].

Além da interoperabilidade técnica, a qualidade do intercâmbio de informações depende da interoperabilidade semântica, responsável por assegurar que os dados compartilhados possuam o mesmo significado clínico independentemente do sistema utilizado. Para isso, empregam-se terminologias e classificações padronizadas, como SNOMED CT, LOINC, Classificação Internacional de Doenças (CID) e Classificação Internacional de Atenção Primária (CIAP), que favorecem a padronização dos registros clínicos, reduzem ambiguidades e ampliam a consistência das informações utilizadas na assistência e na gestão em saúde [12].

Nesse sentido da regulação assistencial, a integração entre arquiteturas interoperáveis e terminologias padronizadas possibilita maior rastreabilidade dos processos, atualização em tempo real das informações assistenciais e suporte mais qualificado à tomada de decisão pelos gestores. Dessa forma, torna-se possível reduzir retrabalhos, minimizar inconsistências cadastrais, otimizar a ocupação de leitos e aprimorar os mecanismos de priorização clínica, contribuindo para uma gestão mais eficiente, transparente e equitativa dos recursos disponíveis [7], [13].

O desenvolvimento dessas arquiteturas representa um campo estratégico de atuação, envolvendo modelagem de sistemas, integração de dispositivos médicos, desenvolvimento de softwares em saúde, segurança cibernética e governança de dados clínicos. A combinação entre interoperabilidade, infraestrutura tecnológica e análise inteligente de dados constitui um dos principais pilares para a transformação digital da regulação assistencial e para o fortalecimento da gestão pública em saúde [5], [7], [13].

A arquitetura de informação proposta para a regulação assistencial no SUS é apresentada na Figura 1, destacando o fluxo de dados desde o paciente e os estabelecimentos de saúde até a tomada de decisão, com integração entre os diferentes níveis do sistema.



Figura I. Fluxograma da arquitetura de informação para regulação assistencial no sus.

V. Estudo de caso: Ecossistema Regula RN

Como exemplo da aplicação prática dos princípios da Saúde Digital na regulação assistencial, destaca-se o ecossistema

RegulaRN, desenvolvido durante a pandemia da COVID-19 pelo Laboratório de Inovação Tecnológica em Saúde (LAIS/UFRN), em parceria com a Secretaria de Estado da Saúde Pública do Rio Grande do Norte (SESAP/RN). O sistema foi concebido para apoiar a gestão regulatória dos leitos hospitalares, promovendo maior integração entre os estabelecimentos de saúde e fornecendo informações em tempo real para subsidiar a tomada de decisão dos gestores públicos [8].

O desenvolvimento do RegulaRN ocorreu em resposta ao rápido aumento da demanda por leitos de internação, especialmente de terapia intensiva, cenário que evidenciou a necessidade de ferramentas capazes de integrar informações clínicas, reduzir o tempo de resposta das centrais reguladoras e ampliar a transparência dos processos de regulação. Nesse contexto, a plataforma consolidou informações provenientes de diferentes unidades assistenciais, permitindo o acompanhamento contínuo da disponibilidade de leitos e do perfil clínico dos pacientes [8]. Do ponto de vista arquitetural, o ecossistema RegulaRN fundamenta-se na integração entre sistemas de informação em saúde, utilização de bases de dados compartilhadas e disponibilização de painéis gerenciais (dashboards) destinados ao monitoramento da ocupação hospitalar e dos indicadores assistenciais. Essa arquitetura favorece a interoperabilidade entre diferentes serviços de saúde, alinhando-se aos princípios da Estratégia de Saúde Digital para o Brasil e às diretrizes de integração da Rede Nacional de Dados em Saúde (RNDS), discutidas nos capítulos anteriores [2], [7], [13].

Um dos principais diferenciais do sistema foi a implementação do Escore Unificado de Priorização (EUP), desenvolvido para apoiar tecnicamente a classificação de pacientes que necessitavam de internação em leitos críticos. O escore utiliza critérios clínicos padronizados para estabelecer prioridades de atendimento, reduzindo a subjetividade das decisões regulatórias e promovendo maior equidade na alocação dos recursos disponíveis [9].

Segundo Veras [9], a utilização do EUP contribuiu para padronizar os processos de priorização clínica, oferecendo maior segurança aos profissionais responsáveis pela regulação e permitindo decisões fundamentadas em parâmetros objetivos. Essa abordagem favoreceu maior rastreabilidade das decisões e fortaleceu a transparência institucional durante um período de elevada pressão sobre o sistema de saúde. Além da priorização clínica, o RegulaRN proporcionou benefícios relacionados à gestão da informação, ao monitoramento em tempo real da ocupação dos leitos e ao suporte à tomada de decisão pelos gestores estaduais. A consolidação de informações provenientes de diferentes hospitais permitiu maior visibilidade situacional da rede assistencial, contribuindo para otimizar a distribuição dos recursos hospitalares e reduzir o tempo necessário para análise das solicitações de internação [8], [9].

O ecossistema do RegulaRN demonstra como a integração entre arquitetura computacional, interoperabilidade, análise de dados e desenvolvimento de sistemas de apoio à decisão pode contribuir para a modernização da gestão pública em saúde. A experiência evidencia que soluções digitais desenvolvidas com base em princípios de interoperabilidade, governança de dados e inteligência computacional possuem potencial para fortalecer os processos regulatórios, ampliar a eficiência operacional e promover maior equidade no acesso aos serviços especializados do SUS [5], [7], [8].

Embora desenvolvido em um contexto emergencial, o RegulaRN consolidou-se como uma experiência nacional de inovação em saúde digital, demonstrando que ecossistemas tecnológicos integrados podem fortalecer a regulação assistencial e ampliar a eficiência da gestão pública em saúde [2], [6], [8]. Entretanto, a expansão de iniciativas semelhantes para outras regiões do país ainda enfrenta desafios relacionados à infraestrutura tecnológica. A aplicação e os impactos das funcionalidades do RegulaRN podem ser analisados a partir de diferentes aspectos relacionados à gestão, priorização, monitoramento e integração das informações, conforme apresentado na Tabela II.

Tabela II
ASPECTOS AVALIADOS, APLICAÇÃO E IMPACTOS NO REGULARN

Aspecto Avaliado	**Aplicação no RegulaRN**	**Impacto Observado**
Gestão de leitos	Monitoramento em tempo real	Maior visibilidade da ocupação hospitalar
Escore Unificado de Priorização (EUP)	Classificação clínica padronizada	Redução da subjetividade na regulação
Dashboards gerenciais	Acompanhamento de indicadores	Apoio à tomada de decisão
Compartilhamento de informações	Integração entre unidades assistenciais	Maior coordenação da rede
Governança digital	Rastreabilidade das decisões	Transparência institucional

VI. DESAFIOS, BIOÉTICA E PERSPECTIVAS FUTURAS

Ainda que a incorporação de tecnologias digitais tenha promovido avanços significativos na organização da regulação assistencial do Sistema Único de Saúde (SUS), a consolidação desse processo ainda enfrenta desafios de natureza tecnológica, organizacional, ética e regulatória. A superação dessas limitações é fundamental para garantir que os benefícios da transformação digital sejam distribuídos de forma equitativa entre as diferentes regiões do país e contribuam para o fortalecimento da gestão pública em saúde [2], [6]. Um dos principais obstáculos à implementação da Saúde Digital no SUS refere-se às desigualdades estruturais existentes entre as regiões brasileiras. Diferenças relacionadas à infraestrutura tecnológica, disponibilidade de conectividade, informatização dos serviços e acesso à internet comprometem a integração dos sistemas de informação e dificultam a adoção homogênea de soluções digitais em todo o território nacional [10]. Tais limitações podem repercutir diretamente nos processos regulatórios, uma vez que a indisponibilidade ou instabilidade de recursos tecnológicos compromete o registro, o compartilhamento e o acesso oportuno às informações necessárias

à gestão das filas e à tomada de decisão. Dessa forma, os investimentos em infraestrutura devem considerar as particularidades e desigualdades regionais, de modo a evitar que a transformação digital amplie disparidades já existentes no acesso aos serviços de saúde. Além da infraestrutura física, a heterogeneidade na maturidade digital das instituições de saúde também influencia diretamente a capacidade de implementação de arquiteturas interoperáveis e de utilização da Rede Nacional de Dados em Saúde (RNDS), ampliando as disparidades entre diferentes redes assistenciais [2], [10]. A interoperabilidade, nesse contexto, representa um desafio que ultrapassa a simples conexão entre plataformas, envolvendo a adoção de padrões comuns, a padronização das informações e a capacidade dos diferentes sistemas de trocar e interpretar dados de maneira consistente.

A transformação digital exige que gestores, profissionais de saúde e equipes técnicas desenvolvam competências relacionadas ao uso de sistemas de informação, análise de dados, interoperabilidade e governança digital. A incorporação de novas tecnologias sem investimento contínuo em qualificação profissional pode comprometer a efetividade das soluções implementadas e limitar seu potencial de impacto na assistência [6], [11].Ademais, a ausência de capacitação adequada pode contribuir para a subutilização das ferramentas disponíveis, para inconsistências no registro das informações e para dificuldades na adaptação dos processos de trabalho às novas tecnologias. Portanto, programas permanentes de educação e capacitação tornam-se fundamentais para fortalecer a cultura digital nas instituições de saúde, favorecendo a utilização adequada das ferramentas tecnológicas e estimulando processos decisórios fundamentados em evidências científicas [6].

O compartilhamento de informações clínicas em ambientes interoperáveis amplia a necessidade de mecanismos robustos de segurança da informação e proteção da privacidade dos pacientes. A Lei Geral de Proteção de Dados (LGPD) estabelece princípios e diretrizes para o tratamento de dados pessoais sensíveis, exigindo que instituições públicas e privadas adotem práticas voltadas à confidencialidade, integridade, disponibilidade e rastreabilidade das informações em saúde [14]. Além do cumprimento dos requisitos legais, a implementação de políticas de governança de dados, mecanismos de autenticação, criptografia e controle de acesso representa elemento essencial para garantir a confiança dos usuários e assegurar a utilização ética das tecnologias digitais no contexto assistencial [14]. Entretanto, a conformidade com a legislação representa um desafio contínuo, especialmente em ecossistemas caracterizados pelo compartilhamento de informações entre múltiplas instituições, sistemas e profissionais. Nesse cenário, torna-se necessário estabelecer critérios claros para o tratamento, o acesso, o compartilhamento e a responsabilização pelo uso dos dados, conciliando a circulação das informações necessária à continuidade do cuidado com a proteção dos direitos dos usuários.

O avanço das aplicações de Inteligência Artificial (IA) na saúde amplia as possibilidades de apoio à regulação assistencial, especialmente na classificação de pacientes, análise de prioridades, previsão de demanda e otimização da utilização dos recursos disponíveis. Entretanto, a adoção dessas tecnologias deve ser acompanhada por princípios de transparência algorítmica, explicabilidade, supervisão humana e equidade, reduzindo riscos relacionados a vieses computacionais e decisões automatizadas inadequadas [11], [14]. Nesse contexto, torna-se fundamental desenvolver sistemas inteligentes que conciliem elevado desempenho computacional com requisitos éticos, legais e clínicos, garantindo que a inovação tecnológica permaneça centrada nas necessidades dos pacientes e dos profissionais de saúde.

Como perspectivas futuras, a tendência é que a evolução da Saúde Digital no SUS seja impulsionada pela expansão da interoperabilidade nacional, pelo fortalecimento da Rede Nacional de Dados em Saúde (RNDS), pela consolidação de arquiteturas baseadas em padrões internacionais, como o HL7 FHIR, e pela incorporação progressiva de soluções baseadas em Inteligência Artificial para apoio à tomada de decisão [2], [7], [13].

VII. Conclusão

A transformação digital tem redefinido a organização da regulação assistencial no Sistema Único de Saúde (SUS), promovendo avanços na integração das informações, na otimização dos fluxos assistenciais e no apoio à tomada de decisão baseada em evidências. A partir da revisão narrativa realizada, verificou-se que iniciativas como a Estratégia de Saúde Digital para o Brasil 2020–2028, a Rede Nacional de Dados em Saúde (RNDS) e a adoção de padrões de interoperabilidade, como o HL7 FHIR, constituem elementos fundamentais para a construção de um ecossistema de saúde mais integrado, eficiente e centrado no cidadão [2], [5], [7], [13].

O estudo de caso do ecossistema RegulaRN evidenciou que a aplicação de arquiteturas interoperáveis, associadas ao uso de sistemas de apoio à decisão, dashboards gerenciais e critérios padronizados de priorização clínica, pode contribuir significativamente para a qualificação dos processos regulatórios, ampliando a transparência, a rastreabilidade e a eficiência da gestão pública em saúde [8], [9]. Esses resultados demonstram o potencial da transformação digital para fortalecer a coordenação das Redes de Atenção à Saúde e otimizar a utilização dos recursos assistenciais disponíveis [15]. Entretanto, a consolidação dessas iniciativas ainda depende da superação de desafios relacionados à infraestrutura tecnológica, à interoperabilidade entre sistemas, à capacitação dos profissionais, à governança dos dados e à observância dos princípios éticos e legais estabelecidos pela Lei Geral de Proteção de Dados (LGPD). Além disso, a incorporação crescente de soluções baseadas em Inteligência Artificial exige mecanismos de transparência, supervisão humana e avaliação contínua, de forma a garantir decisões seguras, equitativas e alinhadas às necessidades da população [6], [10], [11], [14].

Portanto, a Engenharia Biomédica desempenha papel estratégico ao integrar conhecimentos de engenharia, computação e saúde para o desenvolvimento de arquiteturas

tecnológicas interoperáveis, sistemas inteligentes e soluções inovadoras voltadas à gestão do cuidado. Espera-se que a continuidade dos investimentos em infraestrutura digital, padronização tecnológica e inovação contribua para o fortalecimento da regulação assistencial, promovendo maior eficiência, equidade e sustentabilidade do SUS, além de incentivar novas pesquisas voltadas à avaliação do impacto dessas tecnologias na qualidade da assistência e na gestão dos serviços de saúde [1], [2], [6], [7], [16].

REFERÊNCIAS

[1] Brasil. Constituição da República Federativa do Brasil de 1988.
[2] Brasil. Ministério da Saúde. Estratégia de Saúde Digital para o Brasil 2020–2028. Brasília; 2020.
[3] Brasil. Ministério da Saúde. Portaria nº 1.559, de 1º de agosto de 2008.
[4] Silva LC et al. Regulação assistencial e controle do SUS: revisão de literatura. Europub J Health Res. 2024.
[5] Barbalho IMP et al. Electronic health records in Brazil: prospects and technological challenges. Front Public Health. 2022.
[6] Haddad AE, Lima NT. Saúde Digital no Sistema Único de Saúde (SUS). Interface. 2024.
[7] Cardoso PH et al. Technological architecture for a multi-region solution within the regulation of Brazil's Unified Health System. Front Digit Health. 2026.
[8] Valentim RAM et al. Ecossistema tecnológico no enfrentamento à Covid-19 no SUS: caso do RN. Cien Saude Colet. 2021.
[9] Veras NVR. Impactos do Escore Unificado de Priorização na regulação de leitos do SUS. Tese UFRN. 2024.
[10] Iahnke LM. Saúde Digital na Atenção Primária à Saúde. SciELO Preprints. 2025.
[11] Lopes Júnior JEG et al. Entre algoritmos e territórios: revisão de escopo sobre saúde digital e IA na APS. Rev Panam Salud Publica. 2025.
[12] Galvão MCB. Classificações, terminologias e ontologias no campo da saúde. Asklepion. 2021.
[13] Lins LFA et al. Interoperabilidade entre sistemas de informação em saúde: desafios e tendências com HL7 FHIR. Rev Contemporânea. 2024.
[14] Corrêa HM, Franco MF. Lei Geral de Proteção de Dados (LGPD) na Saúde: tendências, desafios e oportunidades. Anais da ERRC 2025.
[15] SILVA, D. P. et al. Coordenação clínica do cuidado entre Atenção Primária e Atenção Especializada: um estudo transversal. Hygeia, 2025.
[16] COSTA, A. O. N. et al. Saúde digital na regulação assistencial do SUS: revisão integrativa sobre interoperabilidade, inteligência artificial e o papel da Atenção Primária à Saúde. Revista Tópicos, Rio de Janeiro, v. 4, n. 33, p. 1–24, 2026.

Rocket Motion: Desenvolvimento e Validação de um jogo sério para Reabilitação Motora em Crianças Hospitalizadas.

Nayara R. G. de Oliveira
(LAB-E.I.T.A)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
ft.nayrgomes@gmail.com
gmail

Ana Righetto Greco
(LAB-E.I.T.A)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
analuiza.rig@gmail.com

Millena Costa Pimentel
(LAB-E.I.T.A)
Universidade Federal de Goiás
Goiânia, Brasil
gmail

Pedro Schaitl Souza
(LAB-E.I.T.A)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
pedro_schaitl@ufg.br

Pedro Henrique Gonçalves
(LAB-E.I.T.A)
Universidade Federal de Goiás
Goiânia, Brasil
pedrogoncalves@ufg.br

Resumo— A hospitalização infantil é um processo estressante que afeta o bem-estar biopsicossocial da criança. Nesse cenário, jogos sérios e tecnologias assistivas emergem como estratégias valiosas, pois tornam a reabilitação um processo mais lúdico, motivador e menos focado na carga terapêutica. Este estudo teve como objetivo desenvolver um sistema integrado de jogo sério, composto por hardware dedicado, software de interação e plataforma de monitoramento, para estimular a movimentação de crianças hospitalizadas durante sessões de fisioterapia, bem como avaliar sua aplicabilidade clínica inicial em ambiente hospitalar. O sistema "Rocket Motion" foi criado utilizando um controlador inercial (baseado em ESP32-C3 e MPU-6050) e uma *game engine* (Godot). O processo abrangeu o desenvolvimento paralelo de hardware e software, testes de calibração e uma validação clínica preliminar com uma paciente pediátrica. Foi desenvolvida uma ferramenta que combina gamificação com a aquisição objetiva de dados cinéticos. Diferentemente da avaliação visual tradicional, o sistema permite quantificar a qualidade do movimento e armazenar métricas para acompanhamento clínico. A validação indicou alta aceitação e engajamento da paciente, confirmando a viabilidade técnica da solução como suporte à tomada de decisão clínica na reabilitação pediátrica.

Abstract - Child hospitalization is a stressful process that affects a child's biopsychosocial well-being. In this scenario, serious games and assistive technologies emerge as valuable strategies, as they make rehabilitation a more playful, motivating process, less focused on the therapeutic burden. This study aimed to develop an integrated serious game system, composed of dedicated hardware, interaction software, and a monitoring platform, to stimulate the movement of hospitalized children during physical therapy sessions, as well as to evaluate its initial clinical applicability in a hospital environment. The "Rocket Motion" system was created using an inertial controller (based on ESP32-C3 and MPU-6050) and a game engine (Godot). The process encompassed the parallel development of hardware and software, calibration tests, and a preliminary clinical validation with a pediatric patient. A tool was developed that combines gamification with the objective acquisition of kinetic data. Unlike traditional visual assessment, the system allows for quantifying movement quality and storing metrics for clinical monitoring. Validation indicated high acceptance and patient engagement, confirming the technical feasibility of the solution as a support for clinical decision-making in pediatric rehabilitation.

Palavras-chave— Jogos sérios; tecnologia assistiva; reabilitação; hospitalização.

I. Introdução

A hospitalização é um processo crítico, com grandes repercussões para o binômio família/criança. Em decorrência

da internação, a criança é retirada de seu núcleo familiar e social e inserida em um ambiente desconhecido, no qual passa por procedimentos técnicos, muitas vezes invasivos, que podem causar incômodo físico e psicológico. Como reflexo desse cenário, é comum observar, nesses pacientes, manifestações como ansiedade, perda de peso, alterações de humor, declínio cognitivo, baixa autoestima e desenvolvimento de fobias, além de inseguranças e dificuldades de adaptação psicossocial [1].

Em um ambiente inerentemente estressante, repleto de estímulos potencialmente dolorosos e de procedimentos invasivos, garantir a continuidade do desenvolvimento e da qualidade de vida torna-se um desafio crucial para a equipe de saúde. Com isso, ao compreender o brincar como um direito da criança, as Tecnologias Assistivas (TA) apresentam-se como uma estratégia de acessibilidade, uma vez que permitem que as crianças façam parte do seu processo de vida, criando repertório, vínculos, relações e possibilidades de ser e estar, a partir de soluções específicas para crianças em diferentes contextos [2].

Nesse cenário, tecnologias emergentes consolidam-se como recursos valiosos para o desenvolvimento de intervenções digitais lúdicas voltadas ao público infantojuvenil. A implementação de jogos digitais em ambiente hospitalar não apenas proporciona distração durante procedimentos dolorosos, mas também atua como um potente mecanismo para aumentar a motivação dos pacientes e fomentar a interação social entre a criança e a equipe de saúde. Com o avanço tecnológico, surge uma ampla gama de ferramentas que viabilizam a criação de jogos centrados no paciente. Ao utilizar tecnologias avançadas, esses jogos permitem o monitoramento da movimentação em tempo real e o fornecimento de feedback personalizado, tornando a reabilitação um processo mais eficaz e envolvente [3] [4] [5].

O desenvolvimento de jogos sérios (JS) voltados à reabilitação permite a personalização dos níveis de intensidade, bem como o ajuste e o fornecimento de feedback, adaptando-se às necessidades individuais e aos diferentes níveis de comprometimento dos pacientes [6] [7] .

A gamificação, ao integrar elementos como cenários virtuais e atividades lúdicas que exigem movimentação corporal, introduz um componente motivacional à terapia. Isso transforma a sessão de reabilitação em um espaço de descontração e superação pessoal, tornando o processo mais prazeroso e menos focado na carga terapêutica. É importante ressaltar que a criação de tecnologias demanda colaboração interdisciplinar, unindo os conhecimentos de profissionais da saúde e da tecnologia para garantir o sucesso das ferramentas e sua viabilidade na prática clínica [6] [8].

Com isso, partiu-se da hipótese de que a utilização de um jogo sério associado a um controlador inercial personalizado seria capaz de tornar os exercícios fisioterapêuticos mais atrativos, permitindo, simultaneamente, o registro objetivo de parâmetros de desempenho motor durante a execução das atividades.

Portanto, o presente estudo teve como objetivo desenvolver um sistema integrado de jogo sério, composto por hardware dedicado, software de interação e plataforma de monitoramento, para estimular a movimentação de crianças hospitalizadas durante sessões de fisioterapia, bem como avaliar sua aplicabilidade clínica inicial em ambiente hospitalar.

II. Metodologia

O sistema foi concebido como uma arquitetura integrada composta por três subsistemas principais: (i) hardware responsável pela aquisição dos movimentos do paciente; (ii) software de interação baseado em jogo sério; e (iii) módulo de aquisição, armazenamento e visualização de dados clínicos. O desenvolvimento desses subsistemas ocorreu de forma paralela e integrada, garantindo compatibilidade funcional entre todos os componentes. Componentes e a aplicabilidade clínica. Partindo de uma decisão em equipe, foi definido que o jogo precisaria atender às seguintes diretrizes:

- Estimular por meio de recompensa (pontos, estímulo visual e sonoro) exercícios específicos para recuperação e reabilitação dos pacientes;
- Receber os comandos do jogador, por meio de um dispositivo (controle) sem fio;
- Gerar dados para avaliação posterior à sessão de jogo (evolução do quadro clínico);

Esses requisitos foram definidos para contemplar simultaneamente objetivos terapêuticos, aspectos motivacionais e necessidades de monitoramento clínico, permitindo que o sistema funcionasse tanto como recurso de reabilitação quanto como ferramenta de acompanhamento da evolução dos pacientes.

Antes de iniciar o processo de desenvolvimento, realizou-se um levantamento dos softwares e recursos necessários para atender às diretrizes estabelecidas. Por se tratar de um projeto sem fins lucrativos, foram adotados softwares classificados como *"Open Source"*. Dentre os principais softwares, destacam-se o Blender, para modelagem 3D, e o Godot, uma engine de jogos na qual seria desenvolvida a lógica de programação do jogo, bem como a exportação do jogo (aplicativo) para execução. Para o desenvolvimento do hardware (controle), foi adotado o sistema embarcado "ESP32-C3 Super Mini", capaz de ler diversos sensores, em particular, o acelerômetro "MPU-6050".

Quanto ao desenvolvimento do controle, foi dividido em três etapas: programação, construção do protótipo e, por fim, testes e ajustes do controle.

Primeiramente, o desenvolvimento do código precisava garantir que o sistema embarcado executasse as seguintes funções:

1. Inicializar o sistema, definindo terminais e protocolos de comunicação;

2. Verificar a comunicação com o acelerômetro;
3. Verificar a comunicação Bluetooth;
4. Esperar solicitação de conexão Bluetooth;
5. Fazer a leitura do acelerômetro;
6. Calcular o ângulo de inclinação do sensor;
7. Enviar o valor do ângulo para o dispositivo;

Este estudo foi aprovado pelo comitê de ética em pesquisa sob o parecer CAAE: 94597525.5.0000.0237

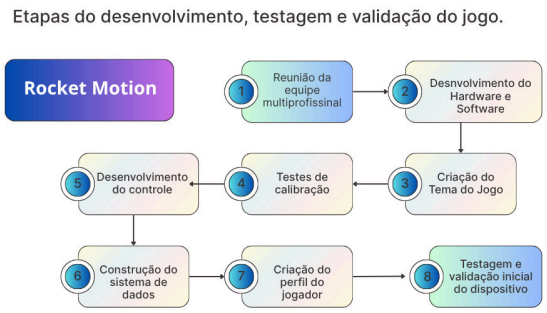


Figura 1: Etapas de desenvolvimento, testagem e validação do jogo.

III. Resultados

O principal resultado deste estudo foi o desenvolvimento de uma plataforma integrada de reabilitação motora, composta por um controlador inercial personalizado, um jogo sério e um sistema automatizado de aquisição de dados. Diferentemente de jogos digitais convencionais, utilizados apenas como recurso motivacional, o sistema desenvolvido incorpora mecanismos objetivos de monitoramento do desempenho motor, permitindo seu uso simultâneo como ferramenta terapêutica e como instrumento de acompanhamento clínico.

Após testes de comunicação e de calibração do acelerômetro, garantindo que toda a parte de programação estava finalizada, iniciou-se a construção do hardware. Nesta etapa, foram estabelecidas características como: tamanho e capacidade de carga da bateria, forma de carregamento da bateria, orientação do acelerômetro, botão de acionamento (liga/desliga), peso máximo e forma de prender o dispositivo ao corpo. A bateria escolhida apresentava tensão nominal de 3,6 volts e capacidade de carga de 450 mAh (miliampères-hora).

O acelerômetro foi posicionado acima do sistema embarcado, o que resultou no conjunto mais compacto possível. Por fim, o conjunto foi acomodado em uma "*case*" feita de PET-G (Polietileno Tereftalato Glicol), impressa em impressora 3D. O material da "*case*" foi escolhido por sua boa resistência mecânica e pela facilidade de higienização.

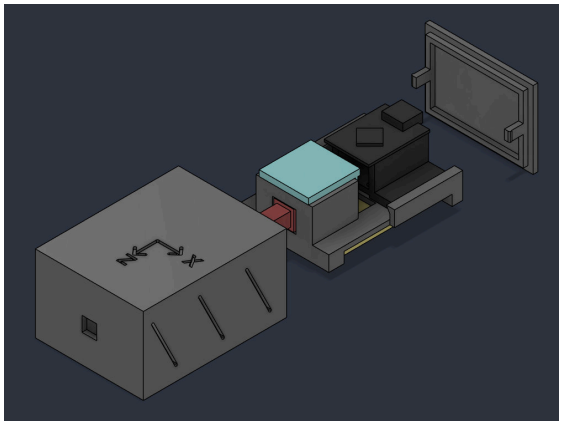


Figura 2: Modelo virtual da montagem do controle.

A última etapa do desenvolvimento do controle foi realizada simultaneamente ao início do desenvolvimento do jogo. Foram feitos pequenos ajustes, tais como a adição de detalhes na tampa da "*case*" para indicar a orientação do sensor, a modificação da taxa de envio e do formato das informações enviadas pelo controle.

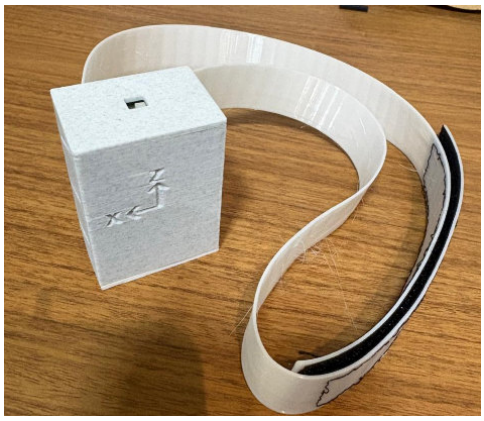


Figura 3: Controle finalizado.

Para iniciar o desenvolvimento do jogo, foi desenvolvido um protótipo no qual o jogador era representado por um bloco, inicialmente localizado no centro da tela, e, conforme o controle era inclinado, o bloco se movia. Essa etapa fez parte do que é conhecido como "blocagem", na qual todos os elementos do jogo são representados da forma mais simples possível. Com o fim desta etapa, foi concluído o desenvolvimento do controle e deu-se início à implementação das mecânicas de jogo.

A escolha da temática espacial teve como objetivo reduzir associações diretas a procedimentos hospitalares, criando um ambiente virtual lúdico capaz de estimular a curiosidade, a exploração e a sensação de progressão durante a execução dos exercícios fisioterapêuticos. Desta forma, foram escolhidos elementos que estimulariam o jogador a agir no jogo: estrelas e

asteroides. Dentro dessa temática, o jogador passa a ser representado por um foguete cujo objetivo é coletar o maior número possível de estrelas, desviando de asteroides.

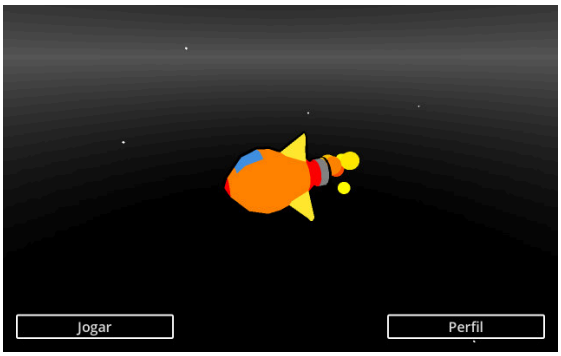


Figura 4: Tela inicial do jogo “Rocket Motion”.

As estrelas são itens coletáveis no jogo que aumentam a pontuação e acionam efeitos visuais e sonoros. Em termos de jogabilidade, é conhecido como “*feedback positivo*” quando uma ação do jogador é recompensada. A pontuação é um fator relevante para aumentar o engajamento do jogador. Se ele apresenta pontuação elevada, o jogador está buscando realizar o exercício proposto; do contrário, o engajamento é baixo e o exercício não está sendo cumprido.

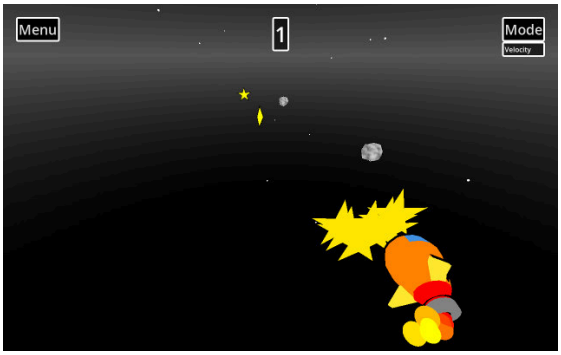


Figura 5: Foguete coletando uma estrela.

Os asteroides funcionam de forma semelhante às estrelas, porém, quando o jogador colide com um asteroide, a pontuação é subtraída e um ruído desagradável é emitido. Trata-se, portanto, de um “*feedback negativo*”. Esse elemento é um estímulo de reforço à ação do jogador, impactando o engajamento dele no jogo.

Depois de implementadas as mecânicas básicas do jogo, passou-se à construção do sistema de dados para análise. O sistema é baseado na criação de um “Perfil” para cada paciente, que contém as seguintes informações: nome, código do paciente e espaço para comentários e observações. O processo de criação do “Perfil” gera uma pasta no dispositivo em que o jogo é executado, na qual são armazenadas as informações de cada exercício.

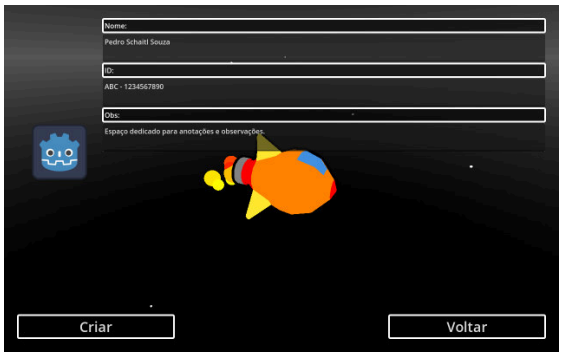


Figura 6: Tela de criação de perfil.

As informações coletadas são utilizadas para analisar o comportamento do paciente durante a execução do exercício, gerando dados sobre o tempo e a intensidade do movimento. Além dos dados numéricos, é gerado um gráfico com três linhas: a branca indica a posição do foguete na tela, a vermelha a inclinação do dispositivo e a amarela a pontuação durante a partida. As linhas pontilhadas são referências fixas para comparação entre os gráficos gerados.

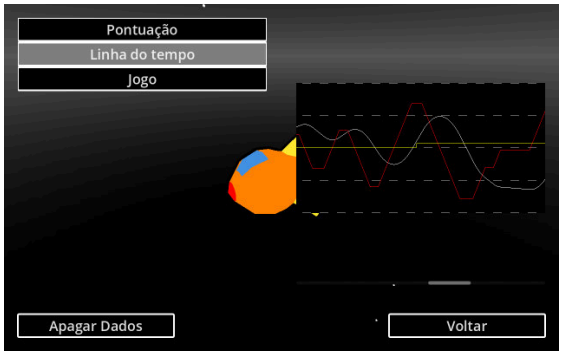


Figura 7: Tela de análise de dados.

A representação gráfica simultânea do movimento executado, da resposta do personagem e da evolução da pontuação fornece ao fisioterapeuta uma visualização integrada do desempenho do paciente, permitindo identificar padrões motores, oscilações de desempenho e possíveis dificuldades durante a execução do exercício.

Desenho experimental, testagem e validação do dispositivo

Foi selecionada uma paciente com diagnóstico de osteomielite, com média de internação superior a 20 dias e que necessitava de descarga de peso do membro inferior. A criança foi convidada a participar da pesquisa juntamente com seu responsável e após o esclarecimento sobre a pesquisa e o objetivo do jogo, foram assinados o termo de consentimento

livre e esclarecido e o termo de assentimento livre e esclarecido.

A criança foi avaliada, antes de iniciar o jogo, quanto à saturação e à frequência cardíaca e, após a sessão, foi aplicada a Escala de Percepção de Esforço da Criança (EPEC). O jogo foi utilizado para trabalhar, de forma lúdica, a movimentação e a descarga de peso bilaterais.

IV. Discussão

A interação no jogo ocorre por meio de enredos lúdicos, como o de guiar um foguete no espaço, no qual a criança o move tanto para coletar estrelas quanto para fugir de asteroides. O cenário interativo ajuda a aumentar a motivação, desvia o foco da dor e da repetição extenuante em um ambiente hospitalar [4].

O principal diferencial do sistema desenvolvido não reside apenas na utilização de elementos de gamificação durante a fisioterapia, mas também na integração entre um controlador inercial personalizado, um ambiente virtual interativo e um módulo automatizado de aquisição de dados clínicos. Corroborando com o estudo de um jogo sério de realidade aumentada desenvolvido para a reabilitação de ombro, que integra métricas específicas para monitorar o desempenho e o progresso do usuário durante as sessões de reabilitação. Os dados coletados são armazenados automaticamente em um banco de dados para permitir a análise clínica e o acompanhamento da evolução ao longo do tempo [9].

Outro fator relevante é o ajuste de velocidade e de dificuldade, o que permite adaptar o desafio ao paciente (jogador) e, consequentemente, mantê-lo engajado no jogo. Para garantir que o paciente se mantenha engajado, os jogos utilizam o Ajuste Dinâmico de Dificuldade (DDA). Esse sistema altera a complexidade do jogo em tempo real com base no desempenho do usuário (ou até mesmo em respostas fisiológicas), garantindo que o desafio não seja tão fácil a ponto de causar tédio, nem tão difícil a ponto de gerar frustração [10] [6].

Essa integração permite que o mesmo sistema utilizado para aumentar o engajamento do paciente também atue como ferramenta objetiva para monitorar a execução dos exercícios e a evolução funcional ao longo do tratamento. Diferentemente da avaliação visual tradicional do terapeuta, o sistema registra métricas precisas que permitem quantificar a qualidade do movimento. Sendo assim, a interação entre dispositivos eletrônicos, jogos digitais e a experiência de análise e compreensão de sistemas biológicos (jogador) propõe um sistema capaz de fornecer dados para o dia a dia dos profissionais da reabilitação, permitindo acompanhar a evolução e regular o processo de reabilitação [11] [12].

A representação gráfica do jogo Rocket Motion permite identificar padrões motores, oscilações de desempenho e possíveis dificuldades na execução do exercício. Isso viabiliza que o profissional modifique, personalize e programe a dificuldade e os parâmetros das futuras sessões (como a velocidade dos alvos ou a tolerância a erros), auxiliando na reabilitação e na tomada de decisão clínica [13].

Na validação, a criança demonstrou fácil compreensão do movimento a ser realizado. Além disso, com o componente de diversão e interação, a criança conseguiu realizar o movimento de inclinação para ambos os lados, com descarga de peso para o lado esquerdo (lado da lesão no joelho). Ao promoverem a motivação intrínseca e a diversão, esses jogos combatem o comportamento sedentário excessivo durante as internações, além de aliviar a ansiedade e a dor [4].

Desse modo, o jogo foi desenvolvido e adaptado para crianças no ambiente hospitalar e, por isso, apresenta menos recursos de interface e sistemas de feedback mais simples. Portanto, reduzir a complexidade, diminuir o número de recursos da interface e incorporar sistemas de recompensa são estratégias recomendadas para minimizar o esforço cognitivo exigido da criança, aumentando as chances de que ela consiga e queira jogar, mesmo estando cansada [3].

V. Conclusão

O presente estudo viabilizou o desenvolvimento e a validação inicial do sistema “Rocket Motion”, um jogo sério integrado a um controlador inercial personalizado, projetado especificamente para a reabilitação motora de crianças hospitalizadas. Ao unir elementos de gamificação,como estímulos lúdicos, recompensas e cenários virtuais, a um sistema de monitoramento de dados cinéticos, a ferramenta demonstrou ser uma alternativa promissora para tornar o processo terapêutico mais atrativo, motivador e menos focado na carga exaustiva do tratamento.

Os resultados obtidos na validação clínica inicial indicaram que a interface simplificada e o feedback em tempo real facilitaram a compreensão e a execução dos movimentos terapêuticos pela paciente, promovendo maior engajamento e auxiliando na descarga de peso. Além do componente lúdico, o sistema mostrou-se uma ferramenta valiosa para o fisioterapeuta, ao fornecer métricas objetivas que permitem quantificar o desempenho, identificar padrões motores e personalizar as intervenções com precisão, superando as limitações da avaliação visual tradicional.

Embora o sistema apresente potencial promissor como recurso assistivo no ambiente hospitalar, este estudo reconhece como limitação a realização de uma única validação preliminar. Dessa forma, pesquisas futuras devem contemplar a

aplicação do “Rocket Motion” em uma amostra maior de pacientes, com diferentes quadros clínicos, bem como a incorporação de instrumentos validados para avaliar a experiência do usuário (usabilidade e aceitação) sob a perspectiva da criança. Mais do que desenvolver um jogo sério, este estudo demonstra a viabilidade técnica de uma plataforma integrada de reabilitação digital capaz de unir gamificação, aquisição objetiva de dados motores e monitoramento clínico em um único sistema. Essa abordagem amplia o potencial dos jogos sérios como ferramentas de apoio à tomada de decisão clínica, além de favorecer sua incorporação em programas de reabilitação pediátrica.

Referências Bibliográficas

[1] Pontes AF, Barros NHC, Rodrigues NA, Albuquerque ML, Cabral MGO, Lucena MCI, et al. O impacto da hospitalização na criança e na família. Res Soc Dev. 2022;11(12):e111111234161. Disponível em: http://dx.doi.org/10.33448/rsd-v11i12.34161

[2] BRASIL, Ministério da Ciência, Tecnologia e Inovações. Plano nacional de tecnologia assistiva / Comitê Interministerial de Tecnologia Assistiva. Brasília, 2021.

[3] Winther L, Stahlhut M, Curtis DJ, Dall CH, Frandsen TL, Sorensen JL. Investigating Challenges in Implementing a Digital Play Intervention in a Complex Organization Across Pediatric Departments: Non-Randomized Controlled Feasibility Trial. JMIR Rehabil Assist Technol. 2025;12:e58019. Disponível em: https://rehab.jmir.org/2025/1/e58019 doi: 10.2196/58019

[4] Jurdi S, Montaner J, Garcia-Sanjuan F, Jaen J, Nacher V. A systematic review of game technologies for pediatric patients. Comput Biol Med. 2018 Jun 1;97:89-112.

[5] Baranyi R, Weindl K, Aigner C, Arnberger C, Hoelbling D, Hoerner W, Grechenig T. Exploring Serious Games Within the Physio and Ergotherapeutic Usage: Insights into Their Acceptance Based on Actual Applications. In: Baumgartner M, et al., editors. dHealth 2025. Amsterdam: IOS Press; 2025. p. 69-74. Available from: https://doi.org/10.3233/SHTI250163

[6] Santos FMK, Eichinger FLF, Domenech SC, Noveletto F, Engster MP, Aquino PAM, et al. Jogo sério para treinamento do equilíbrio em pacientes hemiparéticos por acidente vascular cerebral. Res Soc Dev. 2022;11(14):e571111433600. Disponível em: http://dx.doi.org/10.33448/rsd-v11i14.33600

[7] Mubin O, Alnajjar F, Al Mahmud A, Jishtu N, Alsinglawi B. Exploring serious games for stroke rehabilitation: a scoping review. Disabil Rehabil Assist Technol. 2020;15(7):[página inicial-página final]*. Disponível em: https://doi.org/10.1080/17483107.2020.176830

[8] Noveletto F, Soares AV, Eichinger FLF, Domenech SC, Hounsell MDS, Bertemes-Filho P. Biomedical Serious Game System for Lower Limb Motor Rehabilitation of Hemiparetic Stroke Patients. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2020. [citado em 2026 jul 30]. Disponível em: https://ieeexplore.ieee.org/ doi: 10.1109/TNSRE.2020.2988362.

[9] Viglialoro RM, Turini G, Carbone M, Condino S, Mamone V, Coluccia N, Dell'Agli S, Morucci G, Ryskalin L, Ferrari V, Gesi M. A Projected AR Serious Game for Shoulder Rehabilitation Using Hand-Finger Tracking and Performance Metrics: A Preliminary Study on Healthy Subjects. Electronics. 2023;12(11):2516. Disponível em: https://doi.org/10.3390/electronics12112516

[10] Guzmán DE, Rengifo CF, Guzmán JD, Cena CEG. Virtual reality games for cognitive rehabilitation of older adults: a review of adaptive games, domains and techniques. Virtual Reality. 2024;28(92). Disponível em: https://doi.org/10.1007/s10055-024-00968-3

[11] Grimes RH, Santos AM, Hounsell MS. O Processo de Design de um Sistema Biomédico com Jogo Sério e Dispositivo Especial para Reabilitação Respiratória. In: SBC Proceedings of SBGames; 2018 out. 29-nov. 1; Foz do Iguaçu, Brasil. Porto Alegre: SBC; 2018. p. 215-224.

[12] Shapoval S, García Zapirain B, Mendez Zorrilla A, Mugueta-Aguinaga I. Biofeedback Applied to Interactive Serious Games to Monitor Frailty in an Elderly Population. Appl Sci. 2021;11(8):3502. Disponível em: https://doi.org/10.3390/app11083502

[13] Avola D, Cinque L, Foresti GL, Marini MR. An interactive and low-cost full body rehabilitation framework based on 3D immersive serious games. J Biomed Inform. 2019 Jan;89:81-100. doi: 10.1016/j.jbi.2018.11.012.

Sistema web para gerenciamento de consultas pré-natais de baixo risco

1st Arthur Pinheiro Freitas Silveira
PPG em Engenharia Elétrica
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
arthurpfsilveira@gmail.com

2nd Felipe Jonathan Barros de Oliveira
PPG em Engenharia de Teleinformática
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
felipejonathan@alu.ufc.br

3rd Victor Hugo C. de Albuquerque
Departamento de Teleinformática
Universidade Federal do Ceará (UFC)
Fortaleza, Brasil
victor.albuquerque@ufc.br

***Abstract*—A qualidade dos registros no pré-natal tem impacto direto na promoção da saúde da gestante e do recém-nascido. Diante desse cenário, o registro e o acesso aos dados gerados durante a gestação devem ser realizados de forma consistente para permitir o sucesso do cuidado. Contudo, é possível identificar diversas deficiências associadas ao uso da Caderneta da Gestante que levam à incompletude e à ilegibilidade dos dados. Com o objetivo de superar esses gargalos operacionais, este trabalho desenvolveu e validou a usabilidade de um sistema web capaz de apoiar o acompanhamento pré-natal de baixo risco. O estudo definiu uma metodologia de desenvolvimento voltada ao domínio e centrada no usuário, utilizando *Domain-Driven Design* e Arquitetura Hexagonal. A validação da proposta evidenciou-se na avaliação da usabilidade da ferramenta, que demonstrou alta aceitação por parte dos profissionais de saúde, atingindo nota 77.5 na *System Usability Scale*.**

***Index Terms*—Usabilidade, Engenharia de Software, Desenvolvimento Orientado ao Domínio, Saúde Digital, Saúde da gestante.**

I. Introdução

No Brasil, o acompanhamento pré-natal é uma política pública que engloba iniciativas de promoção de saúde para as gestantes e os recém-nascidos. Para que esta política resulte em ações efetivas, é necessário que os dados clínicos gerados sejam sólidos e rastreáveis [1]. Atualmente, a Caderneta da Gestante atua como instrumento de registro e acompanhamento. Contudo, o modelo impresso apresenta limitações que tornam o processo suscetível a erros de preenchimento, extravio ou deterioração, prejudicando a atuação profissional e a qualidade do cuidado [2].

A associação entre a qualidade dos indicadores de assistência pré-natal e desfechos neonatais ruins é evidenciada na literatura, relacionando a morbimortalidade infantil à descontinuidade da assistência [3]. Em resposta a essa lacuna, o setor da saúde tem intensificado o desenvolvimento de softwares voltados para o apoio à atuação do profissional de saúde.

Com o objetivo de promover a qualificação dos registros, este trabalho desenvolveu um sistema voltado à digitalização da Caderneta por meio da informatização dos fluxos de cadastro e exibição de dados clínicos. O escopo da pesquisa concentra-se no acompanhamento de gestações de baixo risco, situando-se na interseção entre a Engenharia de Software e a Saúde Pública. Para validar a eficácia da ferramenta, foram conduzidos testes de usabilidade com profissionais da área.

II. Fundamentação teórica

A. Desafios no Registro Pré-natal

A Caderneta da Gestante é o principal documento de integração da rede de atenção à saúde [1], [2]. No entanto, a natureza física do documento impõe barreiras à continuidade da informação, como a ilegibilidade e a fragmentação dos dados em prontuários de papel. A digitalização deste processo exige, portanto, não apenas a transposição de formulários, mas uma reengenharia dos fluxos de informação para garantir a integridade dos dados clínicos em um domínio complexo.

B. Saúde Digital no Brasil

A transição para o registro eletrônico de saúde no Brasil é orientada pela Estratégia de Saúde Digital (ESD) 2020-2028. Segundo o 4º Relatório de Monitoramento da ESD, a interoperabilidade e a disponibilidade de dados estruturados são pilares para a continuidade do cuidado na Atenção Primária à Saúde (APS) [13]. O desenvolvimento de soluções que digitalizam a Caderneta da Gestante alinha-se a essa visão, buscando reduzir a fragmentação da informação clínica entre diferentes níveis de atenção.

C. Modelagem e Arquitetura do Sistema

O projeto utilizou *Domain-Driven Design* (DDD), um conjunto de princípios que orientam a modelagem de um domínio de negócio [4] e a sua subsequente transcrição para software. Juntamente, adotou-se a Arquitetura Hexagonal, que visa isolar o núcleo da aplicação de tecnologias externas, como interfaces de usuário e bancos de dados [5].

A escolha pelo DDD se justifica pela complexidade inerente ao domínio obstétrico, onde regras de negócio, como cálculo de idade gestacional, devem ser desacopladas da infraestrutura tecnológica. A Arquitetura Hexagonal complementa essa abordagem ao permitir que as regras de negócio permaneçam agnósticas quanto à detalhes de implementação e tecnologia, facilitando a testabilidade automatizada e a manutenção evolutiva do software.

A Fig. 1 ilustra essa estrutura, evidenciando como as regras de negócio estão protegidas no núcleo, dependendo apenas de interfaces para se comunicar com o mundo externo.

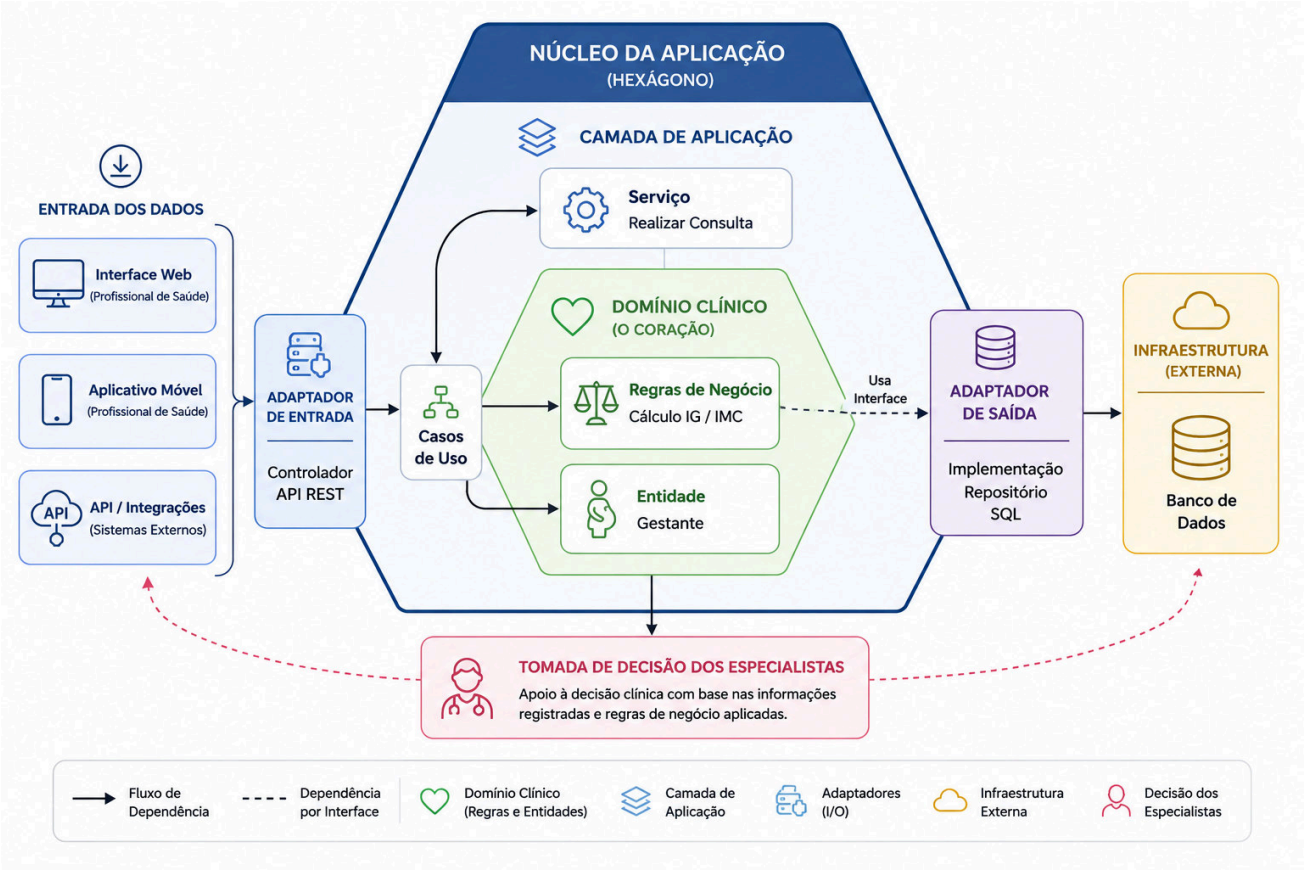


Fig. 1. Modelo arquitetural baseado em Domain-Driven Design e Arquitetura Hexagonal para o sistema proposto de acompanhamento pré-natal.

D. Usabilidade e SUS

A usabilidade é compreendida como o grau em que um sistema pode ser utilizado para alcançar objetivos com efetividade, eficiência e satisfação [7]. Para a avaliação quantitativa, adotou-se a *System Usability Scale* (SUS), um questionário padronizado composto por 10 afirmações mistas (positivas e negativas) avaliadas em uma escala Likert de 5 pontos, variando de ”Discordo Totalmente” (1) a ”Concordo Totalmente” (5) [6].

O cálculo do *score* final do SUS, que varia de 0 a 100, é obtido através de uma normalização das respostas. Para os itens ímpares (positivos), subtrai-se 1 da resposta do usuário ($R - 1$). Para os itens pares (negativos), subtrai-se a resposta de 5 ($5 - R$). A soma resultante é multiplicada por 2,5 para obter a pontuação final, conforme expresso em (1):

$$SUS_{Score} = 2.5 \times \left[\sum_{i=1}^{5} (R_{2i-1} - 1) + \sum_{i=1}^{5} (5 - R_{2i}) \right] \quad (1)$$

Essa metodologia permite mitigar o viés de resposta automática e fornece um indicador único de aceitação do sistema.

III. Metodologia

A. Elicitação de Requisitos e Linguagem Ubíqua

O processo de definição do sistema seguiu uma abordagem centrada no usuário, pautada na técnica de *co-design*. Esta colaboração envolveu sessões de entrevistas e prototipação conjunta com um enfermeiro obstetra especialista na atenção básica. O objetivo central foi o estabelecimento de uma Linguagem Ubíqua, termo do DDD que designa um vocabulário comum compartilhado entre desenvolvedores e especialistas do domínio [4].

Através dessa colaboração, mapeou-se que termos como “Acolhimento”, “Apresentação Fetal” e “Idade Gestacional” deveriam ser refletidos não apenas na interface, mas como entidades centrais no código-fonte. A técnica de *co-design* permitiu validar a sequência lógica de preenchimento dos dados em baixa fidelidade antes da implementação, garantindo que o *software* espelhasse a rotina real de um consultório e reduzisse o hiato entre os requisitos de negócio e o produto final [9].

B. Desenvolvimento da Solução

A arquitetura do servidor fundamenta-se no padrão de Portas e Adaptadores, conhecido como Arquitetura Hexagonal, visando o isolamento rigoroso da lógica de negócio obstétrica frente às tecnologias de infraestrutura [5]. O núcleo da aplicação (*Domain*) centraliza as regras de maior valor da solução, como os cálculos de Idade Gestacional (IG), Data Provável do Parto (DPP) e as validações clínicas essenciais para a integridade dos dados. Operando sob o modelo cliente-servidor, o *backend* expõe esses serviços através de uma API

RESTful, o que assegura que as regras de negócio permaneçam agnósticas à implementação do *frontend*.

Para viabilizar a comunicação desacoplada entre o domínio e o ambiente externo, implementaram-se dois componentes fundamentais:

- Portas (Ports): Interfaces definidas no núcleo que estabelecem o contrato de interação com atores externos, tais como repositórios de dados ou serviços de mensageria.
- Adaptadores (Adapters): Implementações que materializam a comunicação de entrada e saída, subdividindo-se em:
  - *Driving Adapters* (Entrada): Implementados via *Controllers* em Express.js, que traduzem requisições HTTP em comandos para os casos de uso do domínio.
  - *Driven Adapters* (Saída): Artefatos que convertem modelos de domínio em esquemas tecnológicos específicos de forma transparente. Como exemplo, o uso do Prisma ORM para a persistência em banco de dados PostgreSQL foi implementado de modo a permitir que a aplicação gerencie os dados sem acoplamento direto com a tecnologia de armazenamento [8].

Essa estrutura garante que evoluções tecnológicas, como a substituição do banco de dados ou a migração para microsserviços, ocorram sem a necessidade de reescrever o núcleo clínico, elevando a testabilidade e a manutenibilidade da solução.

O desacoplamento proporcionado pela arquitetura facilitou a concepção do sistema sob os princípios do DDD, assegurando que a implementação técnica fosse um reflexo direto das necessidades operacionais dos profissionais de saúde [4].

No nível tático, a consistência dos dados clínicos foi garantida pela implementação de Objetos de Valor (*Value Objects*). Diferente de entidades comuns, esses objetos encapsulam regras de negócio específicas que podem ser reaproveitadas em diversas partes do código, visto que são definidos por seus atributos e não possuem identidade própria. Por exemplo, o objeto de valor responsável pela idade gestacional não apenas armazena o dado, mas valida a cronologia das informações, formata resultados e impede que estados inconsistentes sejam persistidos no núcleo da aplicação, independentemente do contexto ou da função em que é instanciado.

Essa abordagem permitiu que a complexidade dos cálculos obstétricos e das regras de acompanhamento longitudinal fosse isolada em modelos ricos, facilitando a manutenibilidade e garantindo que o sistema atue como uma extensão confiável do raciocínio clínico. A partir desses artefatos, a solução foi organizada em dois contextos delimitados (*Bounded Contexts*) principais:

- Módulo de Consulta: Responsável pela captura estruturada de dados e pela manutenção das invariantes de negócio durante o ato médico. Este módulo centraliza a lógica de validação de parâmetros fisiológicos e automatiza cálculos complexos, como a IG e o IMC, eliminando erros de transcrição e garantindo a integridade imediata do registro clínico. Adicionalmente, o módulo propõe um fluxo estruturado para o atendimento, orientando as etapas necessárias para a realização plena de uma consulta pré-natal. Essa abordagem reduz a carga cognitiva do profissional, permitindo maior foco na interação clínica e nas necessidades da paciente.
- Módulo de Acompanhamento: Voltado à síntese e interpretação do histórico clínico, transformando dados fragmentados em informações estratégicas. Através da agregação de registros longitudinais, o módulo processa a evolução da gestação para gerar projeções visuais de tendência, como curvas de altura uterina e ganho de peso, fundamentais para a detecção precoce de desvios e suporte à decisão clínica baseada em evidências.

C. Stack Tecnológica e Implementação

A materialização dos conceitos arquiteturais utilizou uma pilha tecnológica moderna focada em escalabilidade e robustez. O *frontend* foi desenvolvido com a biblioteca React e o framework Next.js, aproveitando o roteamento baseado em arquivos e a renderização otimizada. Para a gestão de estados de servidor e sincronização de dados, utilizou-se a TanStack Query, garantindo uma interface responsiva mesmo sob condições de rede instáveis.

A camada de persistência foi implementada com o banco de dados relacional PostgreSQL, gerenciado através do Prisma ORM. Esta escolha permitiu o uso de um esquema de dados tipado e migrações estruturadas, refletindo o modelo de agregados definido no DDD. Para garantir a integridade dos dados na entrada, a biblioteca Zod foi aplicada na definição de esquemas de validação tanto no cliente quanto no servidor, assegurando que apenas informações consistentes com o domínio clínico sejam persistidas.

D. Regras de Negócio e Protocolos Clínicos

O sistema foi parametrizado conforme as diretrizes do Ministério da Saúde e as recomendações da Agência Nacional de Saúde Suplementar (ANS), que estabelece a realização mínima de seis consultas durante o período gestacional [14].

O fluxo de dados foi organizado para refletir a jornada da paciente:

- Acolhimento: Registro inicial focado no Cartão Nacional de Saúde (CNS) e na Data da Última Menstruação (DUM), servindo como âncora para os cálculos de IG e DPP.
- Exame Físico Geral e Específico: Interface estruturada para coleta de dados biométricos (peso, altura, IMC) e obstétricos (altura uterina, batimentos cardíacos fetais), onde o sistema aplica validações em tempo real para evitar a incompletude histórica do modelo em papel.

E. Procedimento de Teste

O protocolo de validação seguiu um roteiro estruturado de tarefas que simulavam a rotina real de um consultório. As tarefas foram divididas em três fluxos principais:

1) *Realização de Consulta:* O participante deveria iniciar um atendimento, navegar entre as abas de exame físico e obstétrico e registrar dados clínicos simulados. O sistema deveria validar automaticamente os campos obrigatórios antes de permitir a conclusão.
2) *Gestão de Exames:* Inserção de resultados de exames trimestrais e ultrassonografias, verificando a resposta do sistema quanto à visualização imediata dos dados salvos.
3) *Monitoramento Longitudinal:* Acesso ao histórico da paciente para análise da curva uterina e ganho de peso.

Durante a execução, utilizou-se a técnica *Think Aloud*, onde os participantes verbalizavam suas ações e dificuldades em tempo real. Ao final, aplicou-se o questionário SUS e a entrevista estruturada para quantificar a percepção de usabilidade e obter a percepção dos participantes a respeito do uso da aplicação no contexto de trabalho do profissional de saúde.

Os testes de usabilidade foram validados por profissionais especialistas na atenção primária à saúde vinculados ao contexto do Sistema Único de Saúde (SUS), incluindo enfermeiros e profissionais com experiência em acompanhamento pré-natal de baixo risco. Os participantes possuíam familiaridade prévia com fluxos clínicos obstétricos e com o uso de prontuários eletrônicos em diferentes níveis de maturidade tecnológica. A seleção buscou contemplar profissionais envolvidos na rotina assistencial do pré-natal, garantindo maior aderência entre os cenários simulados e o contexto real de utilização da ferramenta.

IV. Resultados e discussões

A. Resultados Quantitativos

O sistema obteve uma pontuação média de 77.5 na escala SUS (Desvio Padrão = 4.68), como destacado na Tabela **??**, superando o limiar de 68 pontos considerado como média da indústria. Na escala de aceitabilidade, a nota posiciona o sistema na categoria ”Bom” a ”Excelente”.

B. Resultados Qualitativos

Os participantes destacaram a agilidade e a organização das informações. A centralização dos dados na seção de acompanhamento foi citada como uma contribuição significativa para a continuidade do cuidado. Um participante relatou: ”Senti que pude me concentrar mais na paciente e menos em escrever no computador”. Entretanto, foram identificados pontos de melhoria relacionados à curva de aprendizado inicial, refletidos nas notas menores nos quesitos de ”Facilidade de Aprendizagem”.

TABLE I
Pontuação média por componente de usabilidade

Componente Avaliado	Média	Desvio Padrão
Facilidade de Aprendizado	3,8/5,0	0,72
Eficiência de Uso	4,2/5,0	0,65
Consistência do Sistema	4,5/5,0	0,50
Satisfação Geral	4,0/5,0	0,81
Score SUS Total	**77,5**	**4,68**

C. Análise de Usabilidade

A avaliação do protótipo por meio da *System Usability Scale* (SUS) resultou em um escore médio de 77,5 pontos, situando a ferramenta na categoria “B” (Excelente) de acordo com a escala de adjetivos de Bangor [10]. Este índice supera o limiar de mercado de 68 pontos e valida a eficácia da interface na redução de erros de registro clínico.

Os itens com médias mais elevadas estiveram associados à percepção de integridade dos dados e à confiança nos cálculos obstétricos automáticos. Em contrapartida, observou-se uma curva de aprendizado inicial mais acentuada, refletida em escores moderados nos quesitos de facilidade de aprendizado e memorização. A relevância deste escore torna-se evidente ao contrastá-lo com sistemas da APS no Brasil: o aplicativo e-SUS Território apresenta um escore SUS médio de apenas 55,3 [11], enquanto o Prontuário Eletrônico do Cidadão (PEC) atingiu 60,1 [12]. Esta disparidade quantitativa posiciona o protótipo como uma solução superior em aceitação técnica frente aos sistemas governamentais vigentes.

D. Análise Crítica e Sugestões dos Usuários

Durante os testes de usabilidade, a técnica *Think Aloud* revelou que a percepção de utilidade do sistema está diretamente ligada à automação de cálculos repetitivos. Profissionais relataram que a visualização automática da curva de ganho de peso e altura uterina facilita a identificação imediata de desvios de crescimento fetal, algo que na caderneta física exige consulta a gráficos anexos e cálculos manuais.

Contudo, as sugestões coletadas apontam para a necessidade de um módulo de exportação em PDF dos registros digitais. Essa demanda reflete o cenário de transição, onde o registro digital precisa coexistir com a necessidade de a paciente portar um documento físico em casos de emergência hospitalar fora da rede integrada.

V. Discussão

Os resultados indicam que a superioridade nos índices de usabilidade não é acidental, mas sim uma consequência direta da rigidez arquitetural imposta pelo DDD e pela Arquitetura Hexagonal. Ao isolar o domínio, o sistema garantiu que o fluxo de trabalho refletisse a realidade clínica em vez de moldar o médico a uma restrição tecnológica.

Entretanto, a tensão observada entre a ”excelência global” e a ”curva de aprendizado” revela um *trade-off* de design: a aplicação rigorosa de invariantes de negócio e validações cronológicas impede a entrada de dados inconsistentes, mas exige maior esforço cognitivo inicial do usuário. Diferente do e-SUS e do PEC, que apresentam aceitação marginal devido a fluxos genéricos, este protótipo prioriza a segurança do dado sobre a simplicidade trivial, o que justifica a percepção de valor superior relatada pelos profissionais.

Sob o prisma estratégico, a Arquitetura Hexagonal resolve o problema da obsolescência tecnológica na saúde pública. A capacidade de manter o núcleo de regras obstétricas agnóstico permite que futuras integrações com a Rede Nacional de Dados em Saúde (RNDS) ou a substituição de interfaces

web por aplicações móveis ocorram sem reescrever a lógica de cuidado. Por outro lado, a dependência de infraestrutura (internet) e o escopo restrito ao ”Baixo Risco” são riscos críticos que podem limitar a adoção em larga escala em cenários de alta complexidade ou precariedade tecnológica.

Sob a perspectiva de interoperabilidade, a separação entre área clínica e infraestrutura tecnológica favorece a adoção de padrões utilizados na saúde digital, como HL7 FHIR e APIs interoperáveis compatíveis com a RNDS. Essa característica possibilita que o sistema evolua para cenários de compartilhamento de dados clínicos entre diferentes níveis assistenciais do SUS, reduzindo redundâncias de registro e dando continuidade ao cuidado materno-infantil.

Adicionalmente, a arquitetura desacoplada da solução cria condições favoráveis para integração futura com dispositivos de Internet das Coisas (IoT) voltados ao monitoramento materno-fetal. Sensores biomédicos para acompanhamento de frequência cardíaca fetal, pressão arterial materna e glicemia poderiam ser incorporados por meio de adaptadores específicos, sem necessidade de reestruturação do núcleo clínico da aplicação. Essa possibilidade torna a ferramenta proposta promissora para contextos de telemonitoramento e assistência remota em saúde digital.

Diferentemente de abordagens centradas exclusivamente na digitalização de formulários, a proposta apresentada estrutura o domínio clínico como núcleo estratégico da aplicação. Essa característica favorece sustentabilidade tecnológica de longo prazo, reduz dependências de infraestrutura específica e facilita manutenção evolutiva em ambientes públicos de saúde, tradicionalmente impactados por heterogeneidade tecnológica e descontinuidade de sistemas.

VI. Conclusão

Este trabalho apresentou o desenvolvimento e a validação de um protótipo para a digitalização da Caderneta da Gestante, abordando lacunas críticas na continuidade do cuidado pré-natal. A aplicação dos princípios da arquitetura hexagonal e DDD resultou em um artefato de software que não apenas atende aos requisitos funcionais, mas que demonstra alta manutenibilidade e extensibilidade técnica.

Conclui-se que a ferramenta possui potencial significativo para reduzir a carga cognitiva dos profissionais de saúde, automatizando cálculos e centralizando informações vitais que anteriormente estavam dispersas ou ilegíveis.

Além das vantagens operacionais observadas durante a validação de usabilidade, o sistema proposto apresenta alinhamento estratégico com as diretrizes da Estratégia de Saúde Digital para o Brasil, especialmente no que se refere à interoperabilidade, rastreabilidade de dados clínicos e continuidade assistencial. A modularidade arquitetural proposta favorece a integração futura com ecossistemas nacionais de saúde digital, como a RNDS, contribuindo para políticas públicas voltadas à qualificação da atenção pré-natal no SUS.

É importante destacar que a arquitetura proposta permite evolução incremental para cenários mais complexos, incluindo suporte *offline* em regiões com baixa conectividade e expansão do domínio clínico para acompanhamento de gestação de alto risco. Essas possibilidades reforçam a capacidade da solução de adaptação a diferentes realidades assistenciais e tecnológicas do contexto brasileiro.

Para trabalhos futuros, sugere-se:

- A implementação de estratégias *offline* com sincronização através de aplicações *mobile* ou *desktop* para garantir a operação em áreas remotas;
- A expansão do domínio para contemplar protocolos de Gestação de Alto Risco;
- A integração com a RNDS para interoperabilidade com outros sistemas do SUS.

References

[1] MINISTÉRIO DA SAÚDE. “Atenção ao pré-natal de baixo risco,” *Cadernos de Atenção Básica*, vol. 32. Brasília: Editora do Ministério da Saúde, 2012.

[2] L. F. D. Camargos *et al.*, “Avaliação da qualidade dos registros de cartões de pré-natal de mulheres urbanas,” *Escola Anna Nery*, vol. 25, no. 1, p. e20200166, 2021.

[3] E. C. F. Vidal *et al.*, “Assistência pré-natal associada aos desfechos neonatais em maternidades: estudo transversal de base hospitalar,” *Revista da Escola de Enfermagem da USP*, vol. 57, p. e20230145, 2024.

[4] E. Evans, *Domain-driven design: attacking complexity in the heart of software*. Addison-Wesley Professional, 2004.

[5] A. Cockburn, “Hexagonal architecture,” *Alistair Cockburn’s Blog and Articles*, 2005. [Online]. Disponível: https://alistair.cockburn.us/hexagonal-architecture. [Acesso em: 28 jul. 2025].

[6] J. Brooke, “SUS: a retrospective,” *Journal of Usability Studies*, vol. 8, no. 2, pp. 29–40, 2013.

[7] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, “NBR ISO 9241-11: Requisitos ergonômicos para trabalho de escritório com computadores: Parte 11: Orientações sobre usabilidade,” Rio de Janeiro, 2011.

[8] M. GORODNICHEV e K. P. MOSEVA, “Exploring object-relational mapping (orm) systems and how to effectively program a data access model,” *PalArch’s Journal of Archaeology of Egypt/Egyptology*, vol. 17, no. 3, pp. 615–627, 2020.

[9] T. J. NOORBERGEN, M. T. P. ADAM, T. TEUBNER e C. E. COLLINS, “Using co-design in mobile health system development: A qualitative study,” *JMIR Mhealth Uhealth*, vol. 9, no. 11, p. e21929, 2021.

[10] A. BANGOR, P. T. KORTUM e J. T. MILLER, “Determining what individual SUS scores mean: Adding an adjective rating scale,” *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.

[11] R. B. NEVES et al., “Desafios na usabilidade do aplicativo e-SUS território da estratégia de saúde digital do brasil,” *Saúde em Debate*, vol. 49, p. e9774, 2025.

[12] A. SILVA et al., “Usabilidade do Prontuário Eletrônico do Cidadão (PEC) na Estratégia Saúde da Família,” *Revista Brasileira de Informática em Saúde*, 2024.

[13] MINISTÉRIO DA SAÚDE. “4o Relatório de Monitoramento da Estratégia de Saúde Digital no Brasil,” 2023. [Online]. https://bvsms.saude.gov.br/. [Acesso em: 28 jul. 2025].

[14] AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR. “Resolução Normativa RN n o 81, de 2 de setembro de 2004,” 2004. [Online]. Disponível: https://www.gov.br/ans/pt-br.

SolTerm: Solar Water Pasteurization for Humanitarian Emergencies

Gabriel Filadelfo Teles
Production Engineering
National Institute of Telecommunications – Inatel
Santa Rita do Sapucaí, Minas Gerais (MG), Brazil

Lucas Amaral Guimarães
Software Engineering
National Institute of Telecommunications – Inatel
Santa Rita do Sapucaí, MG, Brazil

Maysa Costa Alves
Department of Biomedical Engineering
National Institute of Telecommunications – Inatel
Santa Rita do Sapucaí, MG, Brazil
maysa.alves@inatel.br

***Abstract*—Access to microbiologically safe water may be compromised in humanitarian emergencies. This work presents SolTerm, a solar thermal system concept that may be applied to water pasteurization and that may integrate a parabolic concentrator, a copper receiver, a hydraulic circuit, control, and telemetry. The experimental stage, carried out on June 27, 2026, in Santa Rita do Sapucaí (MG), was limited to the heating of a copper absorber plate coated with graphite paint and enamel, positioned at the focal point of a 500 mm parabolic concentrator, without water circulation through a coil. The plate reached 100 °C in 10 s and 307 °C in 300 s. These values suggest preliminary heating capacity of the optical–absorber assembly, but do not represent the water temperature and do not allow validation of flow rate, residence time, valves, automatic control, or microbiological inactivation. In a separate exploratory trial, condensate formation was observed after 5 min of solar exposure; the sample was inoculated on Brain-Heart Infusion (BHI) agar and evaluated only by visual inspection, a procedure that will not demonstrate microbiological efficacy or potability. The intended future flow rate of 10 L/h remains a nominal target for a future 2 m model and will be verified experimentally. Future validation of SolTerm will take as reference Ordinance No. 888/2021 of the Cabinet of the Minister of Health (GM/MS) and must cover thermal, hydraulic, microbiological, and physicochemical aspects.**

***Index Terms*—solar pasteurization, parabolic concentrator, water treatment, humanitarian emergencies, experimental validation, public health**

I. Introduction

Access to microbiologically safe water may be compromised in humanitarian emergencies associated with floods, droughts, and other extreme events. Damage to infrastructure, interruptions in treatment and distribution, contamination of sources, and unavailability of power may increase the population's exposure to waterborne agents. Vulnerability will likely be more significant in small, decentralized, or temporary systems, in which recovery of conventional supply may require more time [1]. In this context, technologies that use solar energy may contribute to local and emergency treatment strategies, provided that their performance is demonstrated under real operating conditions.

Solar water treatment will not be considered, in this work, as a single technology. Solar water disinfection (SODIS) may jointly exploit ultraviolet radiation and the heating of water exposed to sunlight. Solar pasteurization (SOPAS) will depend predominantly on maintenance of the water at a validated combination of temperature and time. In turn, distillation and certain solar desalination processes will involve evaporation, condensation, or salt separation and will constitute routes distinct from pasteurization. Because these processes will present different mechanisms, operational variables, and performance criteria, results obtained in one of these configurations must not be used in isolation to validate another [2], [3]. Consequently, the observation of condensate in an assembly similar to a distiller will not demonstrate the pasteurization of the water that will circulate through SolTerm in the future.

Recent reviews may expand the state of the art applied to the development of the system. Chaúque, Brandão, and Rott systematized applications of SODIS, SOPAS, and hybrid systems and noted that scale-up will likely depend on collector geometry, flow regime, and water characteristics [2]. Jeon *et al.* compared different point-of-use solar technologies and indicated that target pathogen, water quality, materials, irradiance, location, and season may substantially alter performance [3]. These contributions will suggest that the maximum temperature of an absorber surface will not be sufficient to estimate, by itself, the microbiological quality or the productivity of a treatment system.

Pilot-scale studies may also guide future validation. Chaúque *et al.* evaluated continuous-flow systems that combined heating and ultraviolet radiation for the inactivation of bacteria, *Acanthamoeba* cysts, and spores [4], [5]. These results may demonstrate the potential of hybrid arrangements, but will not allow inactivation to be attributed exclusively to heating. In a predominantly thermal approach, Koninck *et al.* characterized, in an indoor environment, a passive continuous-flow pasteurization process and subsequently carried out outdoor experiments with river water [6], [7]. The use of heat recovery, gravity flow, a thermostatic valve, and thermal-

hydraulic modeling in these studies may serve as a reference for the measurement of flow rate, temperature, and residence time in SolTerm, without direct transfer of their results to the geometry proposed here.

The choice of concentrator and receiver must also be grounded in current research. The review by Paliyal and Mondal may support the use of parabolic concentrators in pathogen inactivation, desalination, and contaminant removal processes, in addition to noting future possibilities for automation, tracking, and improvement of absorber materials [8]. Sizirici, Rachid, and Eniola evaluated a passive system with copper tubes and indicated that temperature, thermosiphon circulation, contact time, and a possible germicidal effect of copper may act jointly [9]. Therefore, the presence of copper in SolTerm must not be treated as a guarantee of inactivation, and its contribution can only be analyzed through controls and quantitative microbiological comparison.

Integrated configurations may also combine complementary mechanisms. Pretorius *et al.* studied a system that brought together pasteurization, solar disinfection, and photosensitization, and considered seasonal variations in irradiance, temperature, and flow rate [10]. In the Brazilian context, Araújo *et al.* carried out field tests with a solar pasteurizer equipped with automated flow control and reported operational limitations associated with valve actuation [11]. These works may help in the specification of SolTerm's control system, but the combination of mechanisms or the use of different geometries will prevent their results from being considered a validation of the present system.

Recent literature will thus indicate that performance will likely depend on the interaction between collection area, solar concentration, heat transfer, flow rate, residence-time distribution, temperature history, initial water quality, and the organism evaluated. A distinction between absorber-surface temperature and the temperature actually experienced by each volume of water will also be necessary. A future microbiological evaluation must compare samples before and after treatment, employ replicates and controls, investigate indicators such as *Escherichia coli* and coliforms, and produce quantitative results through colony-forming units (CFU) and/or most-probable-number (MPN) counts. Visual inspection of growth on non-selective medium will be considered exploratory only and will not allow demonstration of logarithmic reduction, potability, or sanitary compliance.

Given this scenario, SolTerm will be conceived as a proposal for a decentralized solar thermal system that may integrate a parabolic concentrator, a copper receiver or coil, a hydraulic circuit, valves, sensors, microprocessor-based control, and telemetry. The objective of this work will be to present its proposed architecture, record the preliminary thermal characterization of the optical–absorber assembly, and delimit the trials that will still be necessary. The experimental stage already carried out was restricted to the heating, without hydraulic load, of a copper plate positioned at the focal point of a 500 mm concentrator; there was no circulation through the coil, flow control, valve actuation, or microbiological comparison before and after treatment. The measured results will be interpreted exclusively as preliminary evidence of plate heating, while the future effectiveness of SolTerm will depend on the thermal, hydraulic, microbiological, and physicochemical validation of the complete system.

A. Biological and Regulatory Foundations of Water Potability

1) Time–temperature relationship: Thermal inactivation may vary according to the target organism, the matrix, the temperature, the thermal history, and the intended logarithmic reduction. Da Silva, Tiba, and Calazans evaluated five time and temperature conditions for total coliforms and *Escherichia coli*, reproduced in Table I [12]. These conditions belong to the system and protocol of that study; in SolTerm, they will be treated only as a preliminary reference until a specific thermal criterion is validated.

TABLE I
TIME AND TEMPERATURE CONDITIONS EVALUATED IN THE LITERATURE. SOURCE: PREPARED BY THE AUTHORS BASED ON DA SILVA, TIBA, AND CALAZANS [12].

Temperature (°C)	Time (s)	Time (min)
55	3600	60
60	2700	45
65	1800	30
75	900	15
85	15	0.25

2) Regulatory reference: GM/MS Ordinance No. 888/2021 amends Annex XX of GM/MS Consolidation Ordinance No. 5/2017 and establishes the control and surveillance procedures and the potability standard [13]. In the future evaluation, the requirements applicable to the supply form studied will be considered, including absence of *Escherichia coli* in 100 mL, criteria for total coliforms, and relevant physicochemical parameters. Thus, visual inspection, temperature, or a single microbiological result will not be treated as a demonstration of full compliance.

The parameters of the Ordinance will be used as a reference for the validation plan. Any eventual demonstration of compliance will depend on physicochemical and microbiological tests of the complete system and on analysis of the requirements applicable to the supply and distribution strategy adopted.

II. MATERIALS AND METHODS

A. Optical subsystem

In the preliminary trial, a 500 mm parabolic antenna coated with mirror fragments of approximately $5\ \mathrm{cm} \times 5\ \mathrm{cm}$ was used. In a later stage, a 2 m model may be built to expand the collection area. The productivity of the future model must be determined experimentally and cannot be inferred directly from results obtained with other geometries or in other locations [8].

B. Thermal subsystem

In a separate thermal trial, carried out on June 27, 2026, between solar noon and 1 p.m., in Santa Rita do Sapucaí (MG), a thin copper plate coated with graphite paint and enamel was positioned at the focal point of the 500 mm parabolic concentrator. The plate temperature was recorded with a thermocouple as a function of exposure time. The initial temperature recorded was 27 °C, and the plate reached 100 °C in 10 s and 307 °C in 300 s.

This trial was carried out without water, without a coil installed, and without hydraulic load. Therefore, the recorded values correspond exclusively to the temperature of the absorber plate and do not represent the temperature of the water, the vapor, or the condensate. They also do not allow the useful power transferable to the water or the thermal efficiency of the system to be determined in isolation.

The future assembly may employ a 1/2-inch copper coil, with 22 turns and approximately 30 cm in height, combined with thermal insulation. These dimensions will remain preliminary specifications and may be changed after the hydraulic and thermal trials with circulating water.

C. Hydraulic subsystem, sensors, and control

The complete hydraulic system must include a feed reservoir, a coil, piping, flow measurement, temperature sensors at the inlet and outlet, and normally closed solenoid valves. The flow rate of 10 L/h will remain a future design target and must be verified by calibrated measurement during the trials under hydraulic load. The residence-time distribution and the period during which the water remains above the thermal criterion must also be determined experimentally [6], [7].

These components did not take part in the thermal and exploratory condensation trials described here. Consequently, water circulation through the coil, actual flow rate, inlet and outlet temperatures, and integrated valve actuation have not yet been evaluated.

D. Automation, sensing, and control

The control logic will be designed to consider temperature history and residence time, and not only an instantaneous thermal threshold, since irradiance may vary during treatment. The algorithm must stop the count when temperature falls below the defined criterion, block the outlet in the event of sensor or power failure, and record temperature, flow rate, treatment duration, and valve state.

Telemetry based on the Internet of Things (IoT) may assist operational traceability, but will not replace water-quality analysis or sanitary compliance evaluation. Automation, sensors, valves, and control logic must be tested jointly in the complete system. Recent continuous-flow and automated-control systems may guide this stage, without constituting validation of SolTerm [6], [7], [10], [11].

The preliminary parameters that will guide the development are summarized in Table II. The values classified as future targets must still be submitted to experimental validation.

TABLE II
TECHNICAL PARAMETERS AND THEIR CLASSIFICATION. SOURCE: AUTHORS (2026).

Parameter	Value/classification
Application	Solar water pasteurization: design objective
Regime	Batch or controlled flow: will be defined after trials
Prototype tested	500 mm concentrator
Future model	2 m concentrator: may be built
Intended future flow rate	10 L/h: target to be validated
Thermal range	65–70 °C: reference to be reassessed
Proposed receiver	1/2-in. copper, 22 turns, 30 cm
Future control	Arduino platform or ESP32 microcontroller, sensors, and normally closed (NC) valves
Reference standard	GM/MS Ordinance No. 888/2021; compliance to be evaluated

E. Culture Medium Preparation

Brain-Heart Infusion (BHI) agar was used as the culture medium. Petri dishes were prepared in triplicate for the experimental condition. Dehydrated BHI agar was weighed according to the manufacturer's recommendations, dispersed in distilled water, and heated under constant stirring until fully homogenized.

The culture medium and the Petri dishes, previously wrapped in surgical-grade paper envelopes, were autoclaved at 121 °C for 15 min. After sterilization, the medium was allowed to cool inside a previously sterilized laminar flow cabinet until it reached a temperature between 45 and 50 °C. Finally, the medium was poured into sterile Petri dishes and allowed to solidify, forming a uniform agar layer.

F. Sample Collection

A 10 mL volume of contaminated water was transferred to a previously cleaned porcelain crucible positioned at the focal point of the parabolic antenna. The crucible was covered with the lid of a previously sterilized Petri dish, which served as the collection surface for the condensate.

The experimental setup, using the antenna coated with mirror fragments, was subjected to solar heating for 5 min to induce water evaporation and the subsequent condensation of the resulting vapor on the inner surface of the Petri dish lid. The temperature of the water, vapor, and condensate was not measured during this experiment. The value of 307 °C at 300 s refers exclusively to the separate dry-plate thermal experiment described in Subsection B and must not be reported as the condensation or sample temperature.

After the exposure period, the condensate formed on the inner surface of the lid was collected using sterile swabs; these were rubbed across the entire surface containing the condensate for microbiological evaluation purposes. Subsequently, the swabs were used to inoculate plates containing BHI agar using the spread-plate technique. The plates were incubated in a bacteriological incubator at 35 °C for 48 h. Following the incubation period, the plates underwent microbiological

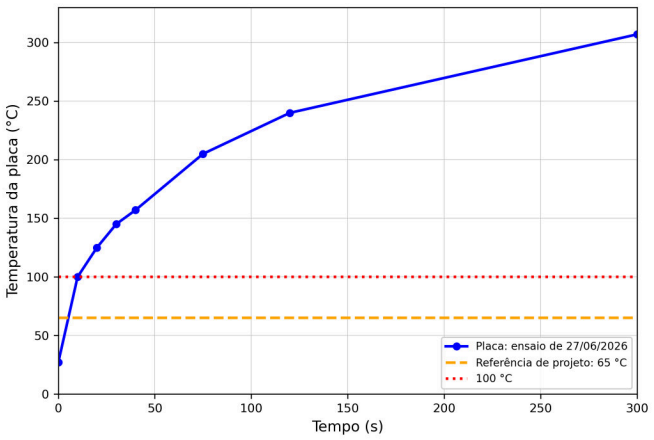


Fig. 1. Experimental temperature curve of the absorber plate. The values do not represent the water temperature. Source: authors (2026).

evaluation based on a visual analysis of the presence or absence of microbial growth.

This procedure was exploratory: no colony-forming-unit (UFC) or most-probable-number (NMP) counts, identification of *E. coli* or coliforms, before-and-after comparison, or positive and negative controls were performed. Therefore, the observation must not be interpreted as evidence of microbial reduction, potability, or validation of the SolTerm system.

III. Results

A. Experimental data

The time and temperature values recorded in the trial are presented in Table III.

The evolution of plate temperature throughout the trial is represented in Figure **??**.

The plate reached 100 °C in 10 s and 307 °C in 300 s. These values suggest that the concentrator–plate assembly may provide relevant heating, but do not represent a water temperature. Without an energy balance and a trial under load, it will not be possible to estimate the useful power transferred to the water, the thermal efficiency, the treatable flow rate, or the residence time.

The apparent average rate was higher at the start and decreased throughout the trial. This behavior may be related to the increase in thermal losses at higher temperatures. However, since there was no measurement of irradiance, repetitions, instrumental uncertainty estimation, or loss balance, no quantitative model will be fitted from this isolated series.

TABLE III
Absorber plate temperature versus exposure time. Source: authors (2026).

Time (s)	Temperature (°C)	Measured quantity
0	27	Absorber plate
10	100	Absorber plate
20	125	Absorber plate
30	145	Absorber plate
40	157	Absorber plate
75	205	Absorber plate
120	240	Absorber plate
300	307	Absorber plate

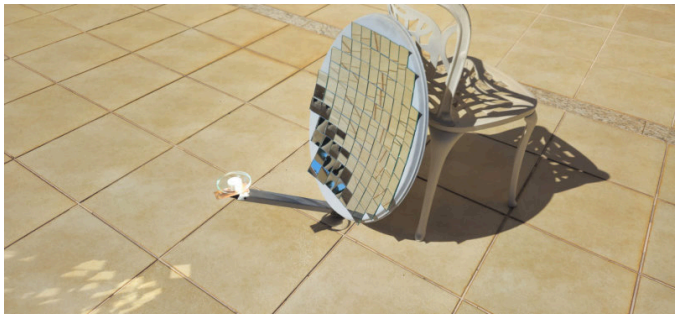


Fig. 2. Assembly of the exploratory condensation trial, with the 500 mm parabolic concentrator and the collection set positioned in the focal region. Source: authors (2026).

B. Exploratory observation of condensate

In the exploratory trial described in Subsection II-F, condensate formation was observed after 5 min of solar concentration. The temperature of the water, the vapor, and the condensate was not measured. The value of 307 °C at 300 s belongs to the separate, water-free trial of the copper plate. The assembly and the sample obtained are presented, respectively, in Figures **??** and **??**. Only visual inspection was performed, without CFU or MPN counts, without identification of *E. coli*/coliforms, without before/after comparison, and without positive and negative controls.

In the visual inspection of the plates prepared in triplicate, no macroscopic changes were observed. This observation will not be treated as a microbiological result: it does not demonstrate logarithmic reduction, does not evaluate water that has passed through the SolTerm coil, and will not support any conclusion about potability or regulatory compliance.

C. Discussion

1) Measured results and projected performance: Only the exposure time and the temperature of the copper plate at the focal point of the 500 mm concentrator were measured. The heating observed suggests that the optical–absorber arrangement may reach elevated temperatures, but will not allow prediction of behavior under hydraulic load. The observation of condensate will be considered only an exploratory record of an assembly similar to a distiller, and not a trial of the SolTerm system.

The intended future flow rate of 10 L/h, the 65–70 °C range, the coil geometry, valve control, and telemetry will remain design parameters or functions. In systems distinct from SolTerm, Koninck *et al.* evaluated thermal characterization and bacterial inactivation in continuous flow, in an indoor environment, and subsequently carried out outdoor trials with river water [6], [7]. Chaúque *et al.* also studied hybrid continuous systems, in which heating and ultraviolet

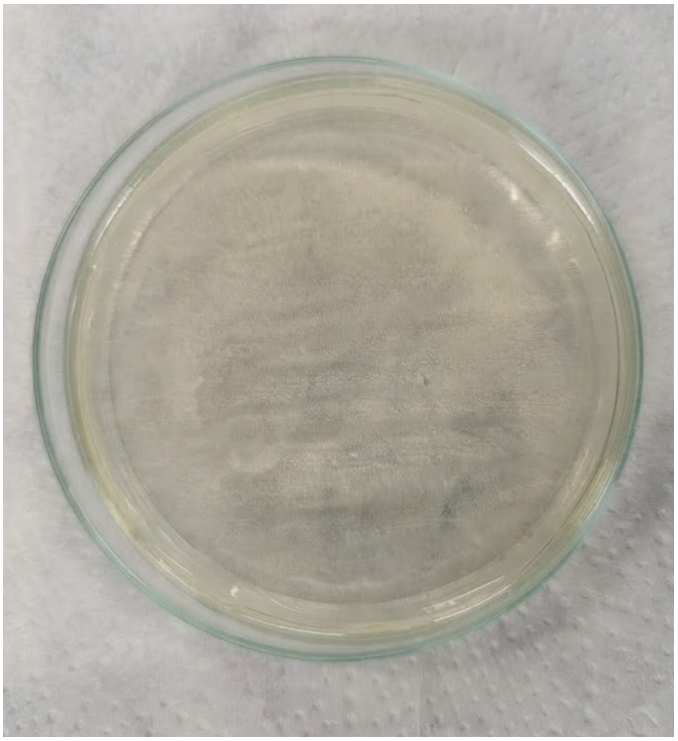


Fig. 3. Condensate sample after 48 h of incubation at 35 °C. Visual inspection does not constitute quantitative microbiological analysis. Source: authors (2026).

radiation acted jointly [4], [5]. Consequently, the productivity and effectiveness of SolTerm must be measured with water and will not be inferred from the no-load plate temperature nor transferred directly from other geometries.

2) Limitations of the study: The current experimental stage did not evaluate: water circulation; actual flow rate; inlet and outlet temperatures; residence-time distribution; period above the thermal criterion; joint valve actuation; performance under hydraulic load; useful thermal efficiency; capacity to reach the future target of 10 L/h; quantitative microbiological reduction; physicochemical parameters; repeatability; measurement uncertainty; or variability with irradiance, wind, and cloud cover.

IV. Conclusion

The preliminary trial was limited to the heating, without hydraulic load, of a copper absorber plate coated with graphite and enamel at the focal point of the 500 mm parabolic concentrator. The plate started at 27 °C, reached 100 °C in 10 s, and reached 307 °C in 300 s. These data may suggest thermal potential for the continuity of development, but must not be interpreted as validation of water pasteurization.

With the current trial, it will not be possible to conclude what temperature the water will reach, what flow rate can be treated, what the residence time will be, how the valves will act, or what microbiological reduction will be obtained. These conclusions will depend on future thermal, hydraulic, and microbiological trials of the complete system.

V. Future Work

Validation of the complete system will be carried out with water, since the current results describe only the heating of the copper plate: 100 °C in 10 s and 307 °C in 300 s, without water circulation through the coil, without valve control, and without evaluation of actual residence time. The next stage must include:

1) the coil, valves, sensors, and control logic will be installed and characterized;
2) the proposed flow rates, including the future target of 10 L/h, will be tested, with calibrated measurement of actual flow rate and of inlet and outlet temperatures;
3) residence time and the period during which each volume of water will remain above the thermal threshold to be validated will be determined;
4) different flow rates will be compared, with repetitions, irradiance measurement, and uncertainty estimation;
5) the preliminary condensate trial will be replaced by microbiological analysis before and after treatment in the complete system, with raw water and water treated by SolTerm;
6) *E. coli* and coliforms will be evaluated by CFU and/or MPN, according to the method selected, with positive controls, negative controls, and replicates; and
7) the applicable physicochemical parameters and the risk of recontamination after treatment will be evaluated.

Acknowledgments

The authors thank the National Institute of Telecommunications (Inatel), the 45th Inatel Technology Fair (FETIN), and Technician Luiz Fernando (Technician in Electrical and Industrial Automation) for the review of the control architecture and for the identification of the limitations of programmable logic controllers (PLCs) and human-machine interfaces (HMIs) in outdoor environments.

References

[1] G. Howard, L. Beevers, K. Charles e A. Nijhawan, "The vulnerability and resilience of drinking water systems to extreme weather events and future climate change," *Current Environmental Health Reports*, vol. 13, no. 1, art. no. 5, 2026, doi: 10.1007/s40572-026-00524-y. Disponível em: https://doi.org/10.1007/s40572-026-00524-y. Acesso em: 21 ago. 2026.

[2] B. J. M. Chaúque, F. G. Brandão e M. B. Rott, "Development of solar water disinfection systems for large-scale public supply, state of the art, improvements and paths to the future—A systematic review," *Journal of Environmental Chemical Engineering*, vol. 10, no. 3, art. no. 107887, 2022, doi: 10.1016/j.jece.2022.107887. Disponível em: https://doi.org/10.1016/j.jece.2022.107887. Acesso em: 21 ago. 2026.

[3] I. Jeon, E. C. Ryberg, P. J. J. Alvarez e J.-H. Kim, "Technology assessment of solar disinfection for drinking water treatment," *Nature Sustainability*, vol. 5, no. 9, pp. 801–808, 2022, doi: 10.1038/s41893-022-00915-7. Disponível em: https://doi.org/10.1038/s41893-022-00915-7. Acesso em: 21 ago. 2026.

[4] B. J. M. Chaúque, A. D. Benetti, G. Corção, C. E. Silva, R. F. Gonçalves e M. B. Rott, "A new continuous-flow solar water disinfection system inactivating cysts of *Acanthamoeba castellanii*, and bacteria," *Photochemical & Photobiological Sciences*, vol. 20, no. 1, pp. 123–137, 2021, doi: 10.1007/s43630-020-00008-4. Disponível em: https://doi.org/10.1007/s43630-020-00008-4. Acesso em: 21 ago. 2026.

[5] B. J. M. Chaúque, P. R. Jankoski, R. L. Doyle, A. S. da Motta, A. D. Benetti e M. B. Rott, "Pilot scale continuous-flow solar water disinfection system by heating and ultraviolet radiation inactivating *Acanthamoeba* cysts and *Bacillus* spores," *Journal of Environmental Chemical Engineering*, vol. 11, no. 3, art. no. 110074, 2023, doi: 10.1016/j.jece.2023.110074. Disponível em: https://doi.org/10.1016/j.jece.2023.110074. Acesso em: 21 ago. 2026.

[6] C. Koninck, V. Goetz, M. Mahboub, J.-M. Mancaux e D. Stitou, "A passive continuous flow pasteurisation process designed for solar application. Part I: Indoor experiments for thermal characterization and bacteria inactivation," *Solar Energy*, vol. 266, art. no. 112177, 2023, doi: 10.1016/j.solener.2023.112177. Disponível em: https://doi.org/10.1016/j.solener.2023.112177. Acesso em: 21 ago. 2026.

[7] C. Koninck, D. Stitou, J.-M. Mancaux e V. Goetz, "A passive continuous flow process for solar pasteurisation of river water. Part II: Modelling and outdoor experiments for performance characterizations," *Solar Energy*, vol. 279, art. no. 112841, 2024, doi: 10.1016/j.solener.2024.112841. Disponível em: https://doi.org/10.1016/j.solener.2024.112841. Acesso em: 21 ago. 2026.

[8] P. S. Paliyal e S. Mondal, "Parabolic solar collectors for sustainable water treatment: A review of applications, advancements and future directions," *Current Opinion in Environmental Science & Health*, vol. 49, art. no. 100691, 2026, doi: 10.1016/j.coesh.2025.100691. Disponível em: https://doi.org/10.1016/j.coesh.2025.100691. Acesso em: 21 ago. 2026.

[9] B. Sizirici, E. Rachid e J. O. Eniola, "Assessing passive thermosyphon solar water heater as low-cost, sustainable water disinfection technology: Leveraging the germicidal effect of copper and thermal convection loops," *Journal of Environmental Chemical Engineering*, vol. 12, no. 2, art. no. 112475, 2024, doi: 10.1016/j.jece.2024.112475. Disponível em: https://doi.org/10.1016/j.jece.2024.112475. Acesso em: 21 ago. 2026.

[10] M. Pretorius, I. Jeon, M. M. Martínez-Fausto, N. Novelli, J. L. Galindo Arevalo, E. Ryberg, M. M. Derby, J.-H. Kim e A. Dyson, "Building-integrated solar water disinfection system for reliable year-round drinking water safety," *npj Clean Water*, vol. 9, art. no. 27, 2026, doi: 10.1038/s41545-025-00539-2. Disponível em: https://doi.org/10.1038/s41545-025-00539-2. Acesso em: 21 ago. 2026.

[11] M. V. M. Araújo, L. C. Fernande, A. R. Albuquerque e T. A. Sorrentino, "Testes de campo de um protótipo de pasteurizador solar com controle de vazão automatizado," em *Anais do X Congresso Brasileiro de Energia Solar*, 2024, doi: 10.59627/cbens.2024.2532. Disponível em: https://doi.org/10.59627/cbens.2024.2532. Acesso em: 21 ago. 2026.

[12] G. C. da Silva, C. Tiba e G. M. T. Calazans, "Solar pasteurizer for the microbiological decontamination of water," *Renewable Energy*, vol. 87, pp. 711–719, 2016, doi: 10.1016/j.renene.2015.11.012. Disponível em: https://doi.org/10.1016/j.renene.2015.11.012. Acesso em: 21 ago. 2026.

[13] Brasil, Ministério da Saúde, "Portaria GM/MS nº 888, de 4 de maio de 2021," *Diário Oficial da União*, 7 maio 2021. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/gm/2021/prt0888_24_05_2021_rep.html. Acesso em: 21 ago. 2026.

Survival and Independent Predictors of Mortality in Patients Undergoing Myocardial Revascularization With and Without Cardiopulmonary Bypass

1º Gabriela Carvalho da Silva
Graduate Program in Biomedical Engineering
Federal University of Technology of Paraná
Curitiba, Brazil
gabrielacarvalhosilva@alunos.utfpr.edu.br

2º Akemi Yagui
Graduate Program in Biomedical Engineering
Federal University of Technology of Paraná
Curitiba, Brazil
akemiyagui@utfpr.edu.br

3º Eduardo Borba Neves
Graduate Program in Biomedical Engineering
Federal University of Technology of Paraná
Curitiba, Brazil
neveseb@gmail.com

***Abstract*—Myocardial revascularization (MR) is the main revascularization strategy for patients with advanced coronary artery disease, and medium-term survival analysis is essential to assess the clinical impact of the procedure. This study aimed to evaluate overall survival for up to one year and identify independent predictors of mortality in patients undergoing MR with and without cardiopulmonary bypass (CPB). This is a single-center retrospective cohort of 104 patients operated on at Hospital Santa Casa de Misericórdia de Curitiba, with data obtained from electronic medical records. Survival was estimated using the Kaplan-Meier method (Log-rank test), and independent predictors were identified using a parsimonious Cox proportional hazards model restricted to three clinical covariates (smoking, systemic arterial hypertension, and type of surgery) with Ridge regularization (penalty factor = 0.025) to mitigate the risk of overfitting resulting from the low number of events (20 deaths). Indication bias between the surgical techniques was controlled using the Inverse Probability of Treatment Weighting (IPTW) method. Smoking was the main independent predictor (HR = 2.320; 95% CI: 1.009–5.338; p = 0.048). Systemic arterial hypertension (HR = 2.155; 95% CI: 0.394–11.781; p = 0.376) and type of surgery (without CPB vs. with CPB: HR = 0.431; 95% CI: 0.125–1.493; p = 0.184) did not show a statistically significant association in the multivariable model. The model showed an apparent C-index of 0.652, which increased to 0.796 when corrected by internal bootstrap (200 iterations), with an AIC of 180.40 and an IBS of 0.2113. These findings indicate that smoking is an independent predictor of mortality in this cohort, whereas the use of CPB did not demonstrate a statistically significant impact on survival within the observed follow-up period of approximately one year, reinforcing the central role of patient risk factors in prognostic stratification.**

***Index Terms*—Myocardial revascularization; Survival analysis; Cox proportional hazards model; Smoking; Cardiopulmonary bypass.**

I. Introduction

Cardiovascular diseases continue to be the leading cause of morbidity and mortality worldwide and in Brazil, with coronary artery disease (CAD) occupying a prominent position in this epidemiological scenario [1], [2]. Population aging and the high prevalence of modifiable risk factors smoking, systemic arterial hypertension, diabetes, and dyslipidemia sustain the high incidence of CAD and its complications, making myocardial revascularization a procedure of major clinical and public health relevance [1], [3].

Myocardial revascularization surgery (MR) is indicated for patients with significant coronary obstruction, with the aim of restoring myocardial perfusion, reducing ischemic events, and prolonging survival. The procedure may be performed with cardiopulmonary bypass (CPB) or without CPB (off-pump), and the comparison between these techniques regarding medium-term outcomes is a subject of intense debate in the literature [6], [7]. Meta-analyses and follow-up studies, including elderly and higher-risk populations, have demonstrated comparable survival between both approaches, suggesting that prognosis is predominantly determined by the patient's risk profile [8]–[12].

The assessment of survival after MR, however, requires methodological rigor. Survival analysis using the Kaplan-Meier estimator and Cox proportional hazards modeling [5] makes it possible to quantify not only the probability of survival over time, but also the independent effect of clinical, demographic, and surgical covariates on the risk of death. Predictive models for post-MR mortality have been proposed in different contexts, incorporating variables such as age, ventricular function, and comorbidities [4], [13]–[15], and factors such as cardiorespiratory fitness and sex have also been associated with medium-term outcomes [16]–[18].

Despite these advances, important gaps remain, particularly the scarcity of data representative of the Brazilian population, whose epidemiological profile and access to health services differ substantially from those of the countries from which most of the evidence originates [3]. In this context, the application of computational tools dedicated to survival analysis [20] to a local cohort contributes to characterizing the prognosis of these patients and identifying modifiable targets for intervention.

Additionally, the scientific novelty of this work lies in

the application of a robust survival analysis study (incorporating IPTW weighting to control indication bias, Ridge regularization to control overfitting due to the low number of events, and rigorous internal validation by bootstrap) in a local Brazilian cohort from the public health system, demonstrating the methodological feasibility of this approach in secondary hospital databases with restricted sample size and number of events. In view of the above, this study aimed to evaluate the impact of clinical, demographic, and surgical factors on the overall survival of patients undergoing myocardial revascularization, identifying independent risk predictors through Cox regression modeling.

II. Methods

A. Study Design, Setting, and Population

This is a retrospective cohort study with a quantitative approach, conducted with data from adult patients (age $\geq$ 18 years) who underwent myocardial revascularization surgery at Hospital Santa Casa de Misericórdia de Curitiba (PR), Brazil. The sample comprised 104 patients with complete electronic medical records containing the clinical information necessary for survival analysis. Patients undergoing elective, urgent, or emergency procedures were included regardless of sex or the presence of comorbidities; those whose medical records did not allow reliable retrieval of the variables of interest or follow-up time were excluded.

The study protocol was approved by the Research Ethics Committee (REC), opinion no. 8.649.194/CAAE: 97691226.5.0000.0177, in accordance with Resolution no. 466/2012 of the National Health Council and the Declaration of Helsinki. As this study involved the analysis of secondary data, the REC waived the requirement for a free and informed consent form, ensuring strict confidentiality, patient anonymization, and data protection in accordance with the General Data Protection Law (LGPD - Law no. 13.709/2018).

Patients who underwent myocardial revascularization surgery during the period from 05/2025 to 05/2026 were evaluated. The data were analyzed according to the following variables: whether the patient was a smoker or not, whether the patient had been diagnosed with SAH, and which type of surgery was performed.

B. Data Collection, Variables, and Outcome

Clinical, demographic, and surgical data were collected retrospectively from the institutional electronic medical record system (TASY) in anonymized form. The covariates initially considered in the univariable modeling were: age (continuous), sex, smoking status, alcohol use, systemic arterial hypertension (SAH), number of affected coronary vessels, preoperative left ventricular ejection fraction (preoperative LVEF %), and type of surgery according to the use of cardiopulmonary bypass (with or without CPB). The outcome of interest was death from any cause, and survival time was defined as the interval, in days, between the date of surgery and the occurrence of the event or the last available follow-up information (censored observations), covering the observed follow-up period of the cohort.

C. Statistical Analysis

Continuous variables were described using measures of central tendency and dispersion, while categorical variables were expressed as absolute frequencies and percentages. The crude comparison of overall survival rates between groups was performed using the Kaplan-Meier estimator, and statistical differences between the curves were assessed using the Log-rank test.

To assess the independent effect of covariates on survival, the Cox proportional hazards model was used. In order to ensure the mathematical stability of the model given the number of available outcomes (20 events) and to respect the principle of parsimony, the final multivariable model was restricted to three clinically relevant covariates (Smoking, Systemic Arterial Hypertension, and Type of Surgery). This selection resulted in an Events per Variable (EPV) ratio of 6.67. As the EPV was slightly below the traditional threshold of 10, Ridge regularization was adopted as a strategy to mitigate the risk of overfitting. The penalty (with a penalizer factor of 0.025) was applied to promote coefficient shrinkage, ensuring greater stability and precision in the estimates of Hazard Ratios (HR - Hazard Ratios).

Considering the observational nature of the study and the possible baseline differences between the compared groups (with and without cardiopulmonary bypass), a causal inference approach using Inverse Probability of Treatment Weighting (IPTW - Inverse Probability of Treatment Weighting) was implemented. Covariate balance between treatment and control groups was assessed using the Standardized Mean Difference (SMD), and was considered adequate after weighting. The final IPTW-adjusted Cox model reported Hazard Ratios using robust standard errors to correct the variance introduced by the weights.

Due to the absence of an external validation cohort and the impossibility of calculating pre-existing clinical scores (such as EuroSCORE II and STS Score) because of limitations in the original database, the generalizability of the model was rigorously subjected to internal validation techniques. Discriminatory power was quantified by the concordance index (C-index). To correct for the optimism inherent in apparent performance in the training sample, internal Bootstrap resampling (200 iterations) was applied, obtaining the corrected C-index (Harrell method), which provided a more realistic estimate of the predictive ability of the model. Additionally, k-fold cross-validation (with 5 partitions) was performed and the overall accuracy of the model was measured using the Integrated Brier Score (IBS).

The proportional hazards assumption intrinsic to the Cox model was statistically verified through the analysis of Schoenfeld residuals, confirming fulfillment of the assumption for all included covariates ($p < 0.05$). The absence of severe multicollinearity was attested by the Variance Inflation Factor ($<$VIF), with all values remaining below 2. This verification

ensured that variance was not inflated, also respecting the stability thresholds of the adjusted GVIF (Adjusted Generalized Variance-Inflation Factor), which is the most robust metric for avoiding overestimation of penalties in strictly categorical variables. Finally, an exploratory subgroup analysis was performed to assess whether the prognostic impact remained consistent across different clinical profiles (SAH and Surgery strata). Effect modification was formally tested by the statistical significance (p-interaction) of the interaction terms between the main exposure and baseline characteristics.

All analyses were performed using the Statistical Software for Survival Analysis (Version 1.0) [20]. Statistical tests were two-tailed, with a significance level of $\alpha = 0.05$ for inferential conclusions.

III. Results

The general clinical and epidemiological profile of the 104 patients evaluated is detailed in Table I. The sample had a mean age of 63.9 $\pm$ 8.5 years, with a predominance of males (71.2%). Regarding baseline comorbidities and lifestyle habits, systemic arterial hypertension was the most prevalent risk factor (90.4%), followed by smoking (35.6%) and alcohol use (12.5%). From an anatomical and functional perspective, coronary involvement was predominantly multivessel, with a mean of 2.8 $\pm$ 0.6 diseased vessels (median 3.0), associated with a mean preoperative left ventricular ejection fraction (LVEF) of 49.4 $\pm$ 17.0%. Regarding the procedure, 79.8% of the interventions were MR with CPB (conventional myocardial revascularization with cardiopulmonary bypass), whereas 20.2% were without CPB (off-pump). During a mean follow-up of 382.7 $\pm$ 173.8 days, a mortality rate of 19.2% was recorded in the sample.

The initial comparison of survival time suggested a statistically significant difference between the crude survival curves of the groups analyzed (p = 0.0432 in the Log-rank test). In practical terms, this indicates that, even before considering other clinical factors, the time to the occurrence of the outcome differed clearly and non-randomly among the patients in the study.

The Kaplan-Meier curve (Figure 1.) was strictly stratified by smoking status (smokers vs. non-smokers), and the significant difference observed (p = 0.0432) refers to this behavioral stratification, and not to the surgical technique.

To understand how multiple clinical variables jointly influence prognosis, a Cox proportional hazards model containing three main covariates was structured: Smoking, Systemic Arterial Hypertension (SAH), and type of Surgery. The model maintained an acceptable ratio of 6.67 events per variable and employed a Ridge regularization technique (penalizer = 0.025) to avoid overfitting, ensuring greater reliability of the findings.

The multivariable analysis (Table II) revealed that Smoking is an independent risk predictor (Hazard Ratio [HR] = 2.320; 95% CI: 1.009–5.338; p = 0.048). From a clinical perspective, a Hazard Ratio (HR) greater than 1 means an increased risk. In this case, individuals with a history of smoking had more than twice the risk of experiencing the event over the period studied, compared with non-smokers, while the other factors remained unchanged.

Table I
Clinical and Demographic Characteristics and Outcomes of Patients Undergoing Myocardial Revascularization Surgery (N = 104), Curitiba, 2025.

Variable	Category / Measure	Result
Age (years)	Mean $\pm$ SD	63.9 $\pm$ 8.5
	Median (IQR)	64.0 (58.8–70.0)
No. of diseased vessels	Mean $\pm$ SD	2.8 $\pm$ 0.6
	Median (IQR)	3.0 (3.0–3.0)
Pre-op. LVEF (%)	Mean $\pm$ SD	49.4 $\pm$ 17.0
	Median (IQR)	54.5 (42.4–61.1)
Follow-up (days)	Mean $\pm$ SD	382.7 $\pm$ 173.8
	Median (IQR)	436.5 (384.2–491.8)
Sex	Male	74 (71.2%)
	Female	30 (28.8%)
SAH	No	10 (9.6%)
	Yes	94 (90.4%)
Smoking	No	67 (64.4%)
	Yes	37 (35.6%)
Alcohol use	No	91 (87.5%)
	Yes	13 (12.5%)
Type of surgery	MR with CPB	83 (79.8%)
	MR without CPB	21 (20.2%)
Death	No	84 (80.8%)
	Yes	20 (19.2%)

SD: standard deviation; IQR: interquartile range (25–75%); LVEF: left ventricular ejection fraction; SAH: systemic arterial hypertension; MR: myocardial revascularization; CPB: cardiopulmonary bypass.

Table II
Univariable and Multivariable Hazard Ratios (HR) by the Cox Model of Patients Undergoing Myocardial Revascularization (N = 104), CURITIBA, 2025.

		Univariable		Multivariable	
Variable	**Cat.**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Smoking	No	1.000 (Ref.)	Ref.	1.000 (Ref.)	Ref.
	Yes	2.414 (0.998–5.840)	0.051	2.320 (1.009–5.338)	0.048
SAH	No	1.000 (Ref.)	Ref.	1.000 (Ref.)	Ref.
	Yes	2.250 (0.301–16.810)	0.429	2.155 (0.394–11.781)	0.376
Surgery	1	1.000 (Ref.)	Ref.	1.000 (Ref.)	Ref.
	2	0.405 (0.094–1.744)	0.225	0.431 (0.125–1.493)	0.184

Note: Multivariable model adjusted with Ridge regularization (penalizer = 0.025). EPV: 6.67. SAH: Systemic Arterial Hypertension

As shown in Table I, among the surgeries evaluated, 79.8% were performed with cardiopulmonary bypass (with CPB) and 20.2% occurred without it (without CPB). This distribution is highly consistent with the national literature; for example, a broad review conducted by Ortolan et al. [21], encompassing data from more than 2.1 million procedures in Brazil, indicated that approximately 73.3% of myocardial revascularization surgeries are performed with CPB support, whereas 26.7% use the technique without CPB. Therefore, although there is a sample imbalance that imposes statistical limitations regarding

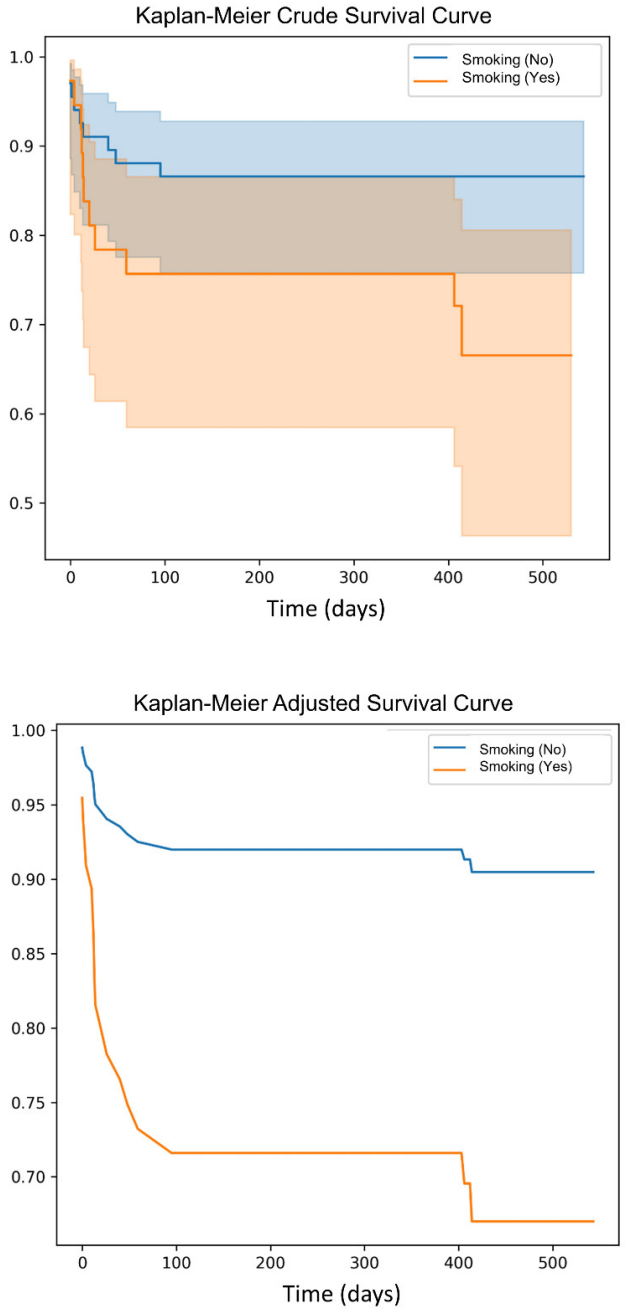


Figure 1. Overall survival curves estimated using the Kaplan-Meier method (Crude and Adjusted), stratified by smoking status (Log-rank test: p = 0.0432).

direct comparisons between the surgical methods, the analysis performed robustly supports future studies in the Brazilian context.

Due to the nature of the study and the possible prior differences between patients operated on with and without cardiopulmonary bypass, Inverse Probability of Treatment Weighting (IPTW) was applied. This technique works as a mathematical 'balance' that corrects imbalances in the groups, simulating a scenario similar to a randomized clinical trial. Covariate balancing was successful, reducing the Standardized Mean Difference (SMD) to values close to zero (SAH from -0.062 to 0.004; Surgery from 0.158 to -0.002). After this rigorous bias correction, the negative impact of smoking remained robust and significant, with an adjusted HR of 2.330 (95% CI: 1.058 – 5.131; p = 0.036). This confirms that the effect found is not caused by baseline differences between the groups.

Seeking to verify whether the risk associated with smoking varied depending on the patient's profile, a subgroup analysis was conducted. In the subgroup of patients with SAH (n=94), the risk remained high and significant (HR = 2.809; 95% CI: 1.126–7.006; p = 0.027). The statistical interaction did not demonstrate significant effect modification between the SAH strata (p-interaction = 0.317) and Surgery (p-interaction = 0.550), suggesting that smoking acts as a consistent risk factor regardless of the clinical profile evaluated.

The mathematical stability of the analysis was confirmed by the absence of multicollinearity (all variance inflation factors [VIF] < 1.65) and by fulfillment of the Schoenfeld proportional hazards assumption ($p > 0.05$ for all variables). The predictive capacity of the model was assessed using the Concordance Index (C-index), which functions in a manner similar to accuracy. After rigorous internal resampling (Bootstrap with 200 repetitions), which corrects the initial 'optimism' of the model, the C-index corrected by the Harrell method reached 0.796. This value suggests that the model has a good ability to discriminate which patients will experience the event first, strengthening the generalizability of the results obtained. The overall accuracy of the model, assessed by the Integrated Brier Score (IBS), ended with a rate of 0.2113.

IV. Discussion

The present study provided an overview of the medium-term postoperative period of patients undergoing Myocardial Revascularization because it was conducted in a cardiology referral hospital in the city of Curitiba. The high prevalence of cardiovascular risk factors in the present sample, with 90.4% systemic arterial hypertension (SAH) and 35.6% smoking, accurately reflects the epidemiological profile of surgical patients in Brazil. This scenario is corroborated by the study by Wagner et al. [22] in a philanthropic cardiology hospital, where an almost identical rate of 91.89% SAH and 37.84% smoking was observed among those undergoing myocardial revascularization, consolidating the national representativeness of our local cohort.

In this retrospective cohort of patients undergoing MR, smoking emerged as the most robust independent predictor of mortality. The multivariable analysis (Table II) revealed that Smoking is an independent risk predictor (Hazard Ratio [HR] = 2.320; 95% CI: 1.009–5.338; p = 0.048). When using SMD, it was also observed that this predictor has a negative impact, with an adjusted HR of 2.330 (95% CI: 1.058 - 5.131; p = 0.036). Thus, using two hazard ratio analysis models, it was observed that the values were close, demonstrating that smoking is an independent risk factor even after adjustment for other relevant clinical variables.

The identification of smoking as the main independent predictor of medium-term mortality (HR = 2.320) finds strong clinical support in the findings of Pietrobon and Barbisan [23]. The authors emphasize that patients who continue smoking in the peri- and postoperative period have significantly higher rates of death and new acute coronary events than those who stop the habit, which justifies the magnitude of the risk observed in our Ridge-stabilized model. This finding is consistent with the central role of smoking as a modifiable cardiovascular risk factor [1], [2] and with the literature identifying smoking among the determinants of mortality after MR. Chua et al. [4] demonstrated, in a medium-term follow-up study after isolated MR, that smoking remains among the main independent predictors of mortality, with a magnitude comparable to the results found in this cohort. The clinical relevance is direct: it reinforces smoking cessation as a priority therapeutic target in perioperative management and in the medium-term follow-up of these patients.

In the univariable analysis, the age variable did not show a statistically significant association with the outcome (HR = 1.041; p = 0.149). Aiming to build a more parsimonious and statistically stable final model, it was decided to limit the multivariable adjustment to only three predictor variables. This restriction was essential to mitigate the risk of overfitting.

A particularly relevant result was the absence of an independent association between the type of surgery (with or without CPB) and survival. This finding aligns with a consistent body of evidence that does not demonstrate superiority of one technique over the other in terms of medium-term survival: the ten-year follow-up of MASS III found no difference between with and without CPB [6]; similar results were observed in left main coronary artery disease [7], in meta-analyses of elderly patients [8], [10], in patients with COPD [11], and in propensity score-matched comparisons with total arterial revascularization [12]. Taken together, these data support the notion that the decision to use CPB should be guided by technical characteristics and the patient's profile, and not by the expectation of a survival benefit.

The implementation of IPTW weighting techniques made it possible to control indication confounding in this retrospective observational sample, balancing the groups with and without CPB (SMD close to zero). The fact that the prognostic impact of smoking remained significant in the weighted model (HR = 2.330, p = 0.036) strengthens the internal validity of the study. In addition, internal validation using 200 Bootstrap iterations (corrected C-index of 0.796) provides a reliable and realistic estimate of the discriminatory power of the local model, surpassing the simplified assessment by the apparent C-index.

It should also be noted that validated clinical scores, such as EuroSCORE II or the STS Score, could not be calculated for comparison due to the scarcity of specific parameters during the collection of data available in the medical records (for example, serum creatinine, NYHA functional class, and history of previous surgeries). This may help justify the need to develop simplified prognostic models focused on the local reality, especially given the absence of complex variables for calculating the aforementioned scores, as supported by the classic epidemiological study by Almeida et al. [24], in which the authors highlight that predictors of mortality in Brazilian cohorts treated by the Unified Health System (SUS) depend on very particular baseline and structural characteristics, reinforcing the incremental value and practical utility of our localized statistical framework, which achieved excellent discriminatory capacity (corrected C-index of 0.796).

This study has limitations related to the assessment of medication, postoperative complications, and socioeconomic data. The retrospective design and the small absolute number of all-cause events (20 deaths) limit the precision of the estimates, as evidenced by the wider confidence intervals. In addition, the marked sample imbalance between the surgical groups (83 vs. 21) reduces statistical power to detect small differences between the revascularization techniques, recommending caution when interpreting the absence of statistical significance as definitive equivalence. In the future, analyses considering these data will be performed and a multicenter study may be conducted with an expanded sample and evaluation of survival time.

V. Conclusion

The present study demonstrated that the medium-term prognosis after myocardial revascularization surgery (MR), in this specific cohort of patients treated at a referral hospital for cardiology care, is predominantly governed by the individual's behavioral risk profile (smoking status) rather than by the surgical technique adopted (with or without cardiopulmonary bypass). Cox proportional hazards modeling with Ridge regularization identified active smoking or recent cessation as the only significant independent predictor of mortality during approximately one year of follow-up, being associated with a significantly elevated risk of death. Internally validated by bootstrap resampling (C-index of 0.796), this framework demonstrates that the absence of cardiopulmonary bypass does not compromise medium-term survival, so technical selection should be individualized based on coronary anatomy. Additionally, this work mitigates the scarcity of representative data in the Brazilian cardiac surgery literature and establishes a robust methodological reference for the analysis of localized hospital databases.

As a suggestion for future studies, increasing the sample size through prospective and multicenter validation of the proposed model is recommended, which will also allow the temporal horizon of survival follow-up to be extended. Second, the inclusion of other predictor variables in the model is suggested, incorporating clinical data of major prognostic relevance that were not available in this secondary cohort. This expansion should encompass the post-discharge therapeutic profile, specific perioperative complications (such as stroke and acute kidney injury), specific causes of death, socioeconomic data, and the incremental role of contemporary cardiac biomarkers, optimizing clinical decision-making algorithms in the national setting.

ACKNOWLEDGMENTS

The authors express their gratitude to Hospital Santa Casa de Curitiba, particularly the Cardiac Surgery team and the institution's research center, for their collaboration and support throughout this study.

REFERENCES

[1] G. M. M. Oliveira *et al.*, "Cardiovascular Statistics – Brazil 2023," *Arq. Bras. Cardiol.*, São Paulo, v. 120, n. 9, supl. 1, e20230079, 2023. DOI: 10.36660/abc.20230079.

[2] GBD 2019 Diseases and Injuries Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, v. 396, n. 10258, p. 1204–1222, 2020. DOI: 10.1016/S0140-6736(20)30925-9.

[3] E. M. Moreira *et al.*, "Five-year mortality in survivors of coronary artery bypass grafting: a retrospective study from a tertiary center," *Arq. Bras. Cardiol.*, São Paulo, v. 123, n. 4, e20250362, 2026.

[4] T. K. T. Chua *et al.*, "Long-term mortality after isolated coronary artery bypass grafting and risk factors for mortality," *J. Cardiothorac. Surg.*, v. 19, n. 1, art. 429, 2024. DOI: 10.1186/s13019-024-02943-0.

[5] D. R. Cox, "Regression Models and Life-Tables," *J. R. Stat. Soc.: Series B*, v. 34, n. 2, p. 187–202, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[6] W. Hueb *et al.*, "Ten-Year Follow-Up of Off-Pump and On-Pump Multivessel Coronary Artery Bypass Grafting: MASS III," *Angiology*, v. 70, n. 4, p. 337–344, 2019. DOI: 10.1177/0003319718804402.

[7] U. Benedetto *et al.*, "Off-Pump Versus On-Pump Bypass Surgery for Left Main Coronary Artery Disease," *J. Am. Coll. Cardiol.*, v. 74, n. 6, p. 729–740, 2019. DOI: 10.1016/j.jacc.2019.05.063.

[8] H. Khan *et al.*, "On- or off-pump coronary artery bypass grafting for octogenarians: a meta-analysis of comparative studies involving 27,623 patients," *Int. J. Surg.*, v. 47, p. 42–51, 2017. DOI: 10.1016/j.ijsu.2017.09.054.

[9] Z. G. Zhu *et al.*, "Comparison of outcomes between off-pump versus on-pump coronary artery bypass surgery in elderly patients: a meta-analysis," *Braz. J. Med. Biol. Res.*, v. 50, n. 3, e5711, 2017. DOI: 10.1590/1414-431X20165711.

[10] R. J. Machado *et al.*, "A systematic review and meta-analysis of randomized controlled studies comparing off-pump versus on-pump coronary artery bypass grafting in the elderly," *J. Cardiovasc. Surg.*, v. 63, n. 1, p. 60–68, 2022. DOI: 10.23736/S0021-9509.21.12012-9.

[11] A. P. Cappellaro *et al.*, "Off-pump versus on-pump coronary artery bypass grafting in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis," *Gen. Thorac. Cardiovasc. Surg.*, v. 73, n. 4, p. 201–208, 2025. DOI: 10.1007/s11748-025-02116-3.

[12] S. Shintomi *et al.*, "Propensity score-matched comparison of total arterial off- and on-pump coronary artery bypass with complete revascularization," *Heart Vessels*, v. 39, n. 2, p. 175–184, 2024. DOI: 10.1007/s00380-023-02317-x.

[13] Z. Hu *et al.*, "An In-hospital Mortality Risk Model for Patients Undergoing Coronary Artery Bypass Grafting in China," *Ann. Thorac. Surg.*, v. 109, n. 4, p. 1234–1242, 2020. DOI: 10.1016/j.athoracsur.2019.08.020.

[14] H. Y. Lin *et al.*, "A Novel Risk Stratification System for Predicting In-Hospital Mortality Following Coronary Artery Bypass Grafting Surgery with Impaired Left Ventricular Ejection Fraction," *Heart Surg. Forum*, v. 23, n. 5, p. E621–E626, 2020. DOI: 10.1532/hsf.3089.

[15] S. H. Sohn *et al.*, "Impact of Myocardial Viability on Long-term Outcomes after Surgical Revascularization," *Thorac. Cardiovasc. Surg.*, v. 72, n. 6, p. 441–448, 2024. DOI: 10.1055/a-2228-7104.

[16] J. F. ter Woorst *et al.*, "Sex Difference in Long-Term Survival After Coronary Artery Bypass Grafting Is Age-Dependent," *J. Cardiothorac. Vasc. Anesth.*, v. 36, n. 5, p. 1288–1295, 2022. DOI: 10.1053/j.jvca.2021.08.104.

[17] J. Ren *et al.*, "Sex differences in long-term survival after total arterial coronary artery bypass grafting," *Eur. J. Cardio-Thorac. Surg.*, v. 65, n. 4, ezae106, 2024. DOI: 10.1093/ejcts/ezae106.

[18] J. Duggan *et al.*, "Long-Term Mortality Risk According to Cardiorespiratory Fitness in Patients Undergoing Coronary Artery Bypass Graft Surgery," *J. Clin. Med.*, v. 13, n. 3, art. 813, 2024. DOI: 10.3390/jcm13030813.

[19] M. Ono *et al.*, "10-Year Follow-Up After Revascularization in Elderly Patients With Complex Coronary Artery Disease," *J. Am. Coll. Cardiol.*, v. 77, n. 22, p. 2761–2773, 2021. DOI: 10.1016/j.jacc.2021.04.016.

[20] E. B. Neves, *Statistical Software for Survival Analysis* (ver. 1.0.1). Zenodo, 2026. DOI: 10.5281/zenodo.20858301.

[21] J. M. Ortolan, L. T. Marcos, D. Sturza, and A. R. Oliveira, "Myocardial revascularization surgery with and without cardiopulmonary bypass. What do the new studies show?," *Vittalle - Health Sciences Journal*, vol. 32, no. 1, pp. 174–184, 2020. DOI: 10.14295/vittalle.v32i1.9716.

[22] J. Wagner *et al.*, "Clinical-demographic characteristics of patients undergoing myocardial revascularization at a philanthropic hospital," *Brazilian Journal of Implantology and Health Sciences*, vol. 4, no. 6, pp. 102–115, 2022. DOI: 10.36557/2674-8169.2022v4n6p102-115.

[23] R. C. Pietrobon and J. N. Barbisan, "Impact of myocardial revascularization surgery on smoking cessation," *Brazilian Journal of Cardiovascular Surgery*, vol. 25, no. 1, pp. 79–84, 2010. DOI: 10.1590/S0102-76382010000100016.

[24] F. F. Almeida, S. M. Barreto, B. R. G. M. Couto, and C. E. F. Starling, "Predictors of hospital mortality and severe perioperative complications in myocardial revascularization surgery," *Brazilian Archives of Cardiology*, vol. 80, no. 1, pp. 41–50, 2003. DOI: 10.1590/S0066-782X2003000100005.

Development and Application of a Form for the Technical Evaluation of Samples in Public Procurement: A Case Study of a Portable Fundus Camera

Patrícia Oliveira de Andrade*
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0000-0001-9781-9140
patricia.andrade@agenciasus.org.br

Rayane de Sousa Sobral*
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0009-0003-7415-925X
rayane.sobral@agenciasus.org.br

Luís Alberto S. Soares de Souza
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0009-0006-1042-7173
luis.souza@agenciasus.org.br

Alex de Oliveira Vasconcelos
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0000-0002-2891-7641
alex.vascoscelos@agenciasus.org.br

Manuela Oliveira da Silva Paz
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0009-0000-8915-5647
manuela.paz@agenciasus.org.br

Caroline Castanho Duarte
UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0000-0001-8348-3586
caroline.duarte@agenciasus.org.br

*Corresponding author

Abstract: This article presents the development and application of a structured form for the technical evaluation of samples in public procurement processes for medical devices, using as a case study the evaluation of a portable fundus camera within a tender process conducted by the Brazilian Agency for Support to the Management of the Unified Health System (AgSUS). The instrument was drawn up on the basis of the technical requirements set out in the Terms of Reference and structured into eliminatory criteria, linked to the mandatory requirements of the contract, and informative criteria intended for recording the operational aspects observed during the practical demonstration. The application of the form made it possible to document non-conformities related to the connectivity requirements set out in the tender documents, as well as to record complementary information on usability, image capture and the operational performance of the equipment evaluated. The analysis highlighted the importance of complementary mechanisms of practical verification for assessing characteristics that cannot always be fully ascertained solely by means of the documentation submitted by suppliers. The results indicate that the instrument contributed to organizing the record of the evidence produced during the demonstration, favouring the traceability of the evaluation process and the substantiation of the technical decision. The experience described suggests that structured approaches of this nature may support health technology assessment processes that require practical analysis complementary to documentary evaluation.

Keywords: HTA; Medical Devices; Technical conformity analysis; Portable fundus camera; Public procurement..

I. INTRODUCTION

The incorporation of medical equipment into the health system requires evaluation processes capable of ensuring that the equipment acquired presents performance compatible with its clinical purpose, operational safety and conformity with the technical requirements established in the tender documents [1], [2]. Especially in the evaluation of equipment that incorporates digital resources, such as diagnostic systems, embedded software and data transmission functionalities, exclusively documentary analysis proves insufficient. In order to ensure the adequate functional validation of the technology offered, the adoption of practical evaluation mechanisms, such as proofs of concept and sample approval procedures, is recommended [3].

In Brazil, the public procurement of medical and hospital equipment traditionally takes into consideration the technical specifications declared by the manufacturers, the regulatory documentation and the sanitary registration issued by the National Health Surveillance Agency. However, part of this information is self-declared in nature, which may limit the capacity for objective analysis of the performance of the equipment under real conditions of use. Although certification carried out by metrological bodies ensures conformity with specific regulatory requirements, it does not necessarily cover in full the operational, clinical and functional aspects that are relevant to the reality of the end user [1], [2], [3].

In addition, it is observed that health technology procurement processes, especially those involving medical and hospital equipment, do not always benefit from the direct participation of professionals with specialized technical knowledge in clinical engineering and/or biomedical engineering [4], [5], [6]. The absence of such technical support may hinder the critical analysis of the specifications presented by manufacturers, increasing dependence on documentation and the risk of incorporating technologies that, although formally compliant, present operational limitations not identified during the judgement phase [3], [7], [8].

In this scenario, the practical evaluation of samples emerges as a tool complementary to documentary analysis, making it possible to observe functional, operational and usability characteristics under conditions of use that cannot

always be fully verified by means of the documentation submitted by the manufacturers. Aspects such as ease of operation, suitability to the workflow, operational stability and performance observed during the use of the equipment may provide additional information to support the technical analysis of health technologies [9], [10], [11].

The present work presents the development of a structured form for technical analysis and the methodology employed for the practical on-site evaluation of equipment submitted to a public procurement process, using as a case study the acquisition of portable fundus cameras. The objective is to describe the structure of the instrument, its application in the evaluation of the sample and its use in systematizing the evidence that supported the technical analysis carried out by the evaluation committee, contributing to greater organization, traceability and transparency of the evaluation process.

II. CASE STUDY

A. Context

The case study presented in this work was developed within the scope of Public Notice No. 5/2026, Electronic Reverse Auction (Price Registration System) No. 5/2026, Process No. AGSUS.001686/2026-01, intended for the price registration for the future and eventual acquisition of portable fundus cameras, expanding Telehealth actions for the Primary Health Care Network, conducted by the Brazilian Agency for Support to the Management of the SUS (UASG: 932133).

Telehealth has been highlighted as a strategic tool for the consolidation of the SUS. Through the use of Digital Information and Communication Technologies (DICT) to expand access to medical care and decentralize health services, Telehealth has managed to connect remote regions to information and support networks, thereby optimizing the efficiency of public care by reducing unnecessary travel and the waiting time of SUS patients and users [12].

The technical evaluation of the sample involved professionals affiliated with AgSUS, the Ministry of Health and the Federal University of Goiás, working in the technical, clinical and information technology areas. This composition was used to record different aspects of the practical demonstration and to support the preparation of the opinion of the Joint Technical Committee.

The tender documents of the said procedure provided for the possibility of submitting samples for technical analysis during the judgement of the bids, expressly establishing the possibility of a practical evaluation of the equipment offered. In accordance with the judgement rules of the procedure, the sample analysis stage is carried out only for the bid summoned during the regular evaluation flow. Accordingly, the present study focused on the analysis of the solution actually submitted to practical demonstration, with no comparative evaluation between different pieces of equipment.

The provision for sample analysis in the tender documents constitutes a relevant complementary mechanism for the technical validation of medical devices, especially in procurement processes for technologies whose verification cannot be fully ascertained solely by means of the technical documentation submitted by the suppliers.

However, it was observed that, although the tender documents provided for the possibility of evaluating samples, there was no prior definition of a structured form or standardized methodology to guide the technical analysis of the equipment submitted for evaluation.

The criteria capable of producing effects on the approval or rejection of the sample were limited to the requirements previously established in Public Notice No. 5/2026 and in the respective Terms of Reference. The process was conducted in accordance with the AgSUS Procurement and Contracting Regulations and with the other rules expressly set out in the tender documents.

Although the process analysed was conducted within the scope of the institutional regulations of AgSUS, the literature on public procurement has emphasized the importance of mechanisms capable of complementing documentary analysis in the evaluation of goods and technologies, especially when certain characteristics of the object cannot be fully verified solely on the basis of the documentation submitted by the suppliers [3].

In this context, the evaluation of the sample constituted a complementary mechanism for verifying the adherence of the solution offered to the technical requirements established for the contract. In the case analysed, its use made it possible to complement the documentary analysis through practical observation of the functional, operational and connectivity characteristics of the equipment, contributing to the technical substantiation of the decision adopted in the evaluation process.

B. Motivation for developing the protocol

In the process analysed, the judgement criterion adopted was the lowest price, conditional upon compliance with the technical specifications and the other requirements established in the tender documents. Under this model, the technical evaluation aimed to verify the conformity of the product offered with the previously defined requirements, rather than to compare or assign relative scores to different technological solutions.

During the initial phase of the analysis of the bids, it was found that the technical documentation submitted by the ranked supplier formally met all the technical specifications required in the tender documents, including the functional characteristics of the equipment, the regulatory documentation and the sanitary registration issued by the National Health Surveillance Agency. From an exclusively documentary perspective, no nonconformities were identified that would justify the technical disqualification of the bid.

However, the technical analysis carried out by the joint technical committee identified the need for complementary validation of the device, considering that the equipment evaluated was a technology recently introduced into the Brazilian market, still without a consolidated history of use in care settings.

In addition, it was found that the equipment did not yet present consolidated evidence of operational validation in telehealth environments, particularly with regard to the transmission of information to the systems of the National Telehealth Network for tele-diagnosis based on the examinations performed. This condition introduces a level of technical uncertainty regarding the functional performance of the device under real conditions of use.

Because of this, the joint technical committee decided to trigger the sample analysis stage provided for in the tender documents, understanding that documentary conformity, although necessary, would not be sufficient to ensure full compliance with the established technical requirements.

In this context, it became necessary to build a structured technical evaluation protocol aimed at the practical on-site verification of the performance of the medical device,

covering aspects related to functional performance, quality of image acquisition, usability, interoperability, operational reliability and full adherence to the previously defined technical specifications.

The instrument was designed to meet the needs of the sample evaluation in the process analysed. Its structure may serve as a reference for adaptations in future evaluations, provided that the criteria are defined in accordance with the object and with the requirements previously established in the respective tender documents. Although it was developed for a specific situation, the form was structured so as to favour its adaptation to similar evaluation processes, and it may be improved and applied in different contexts of health technology acquisition.

III. METHODOLOGY

A. Elaboration of the evaluation instrument

A structured form was developed for the practical evaluation of samples of portable fundus cameras, intended to provide technical support for the process of acquiring medical equipment for use in Primary Health Care, within the context of Telehealth actions.

The eliminatory requirements of the form were derived from the technical specifications and the conditions for the presentation of the sample set out in Public Notice No. 5/2026 and in the Terms of Reference. Informative fields were also included, relating to usability, observed clinical performance, operational stability, connectivity, battery autonomy, biosafety, patient comfort, learning curve, productivity, alerts and additional features. These informative fields were used to record the demonstration and did not, in isolation, constitute grounds for rejecting the sample.

The construction of the instrument involved organizing the requirements of the Terms of Reference and including fields for recording technical, clinical and operational observations. Professionals from AgSUS, the Ministry of Health and the Federal University of Goiás (UFG) took part in monitoring the evaluation, including professionals working in the areas of engineering, information technology and ophthalmology. The preparation of the instrument relied on the participation of professionals from the areas involved in the evaluation, but did not include a formal stage of content validation according to specific methodologies for instrument development.

The instrument was designed to organize the recording of the requirements assessed, the observations made and the evidence produced during the demonstration. In accordance with the rules of the procedure, the practical demonstration was carried out only for the solution summoned to the sample analysis stage, following the regular flow of bid judgement. Accordingly, the instrument was applied to the only sample actually submitted to practical evaluation in the process analysed, without simultaneous comparison between different pieces of equipment.

B. Structure of the evaluation instrument

The instrument was organized into 14 thematic blocks, structured so as to allow the systematic verification of the technical requirements set out in the Terms of Reference, as well as the recording of operational aspects observed during the practical demonstration. It should be noted that each item of the form was associated with the respective regulatory or technical source used in its preparation, ensuring traceability between the evaluative criteria and the requirements of the procurement process. Table 1 details the structure of this instrument and the origin of the evaluation criteria.

TABLE I. TABLE 1 – STRUCTURE OF THE INSTRUMENT AND ORIGIN OF THE EVALUATION CRITERIA

Evaluation block	Item in the form	Purpose	Main source	Nature
Technical and regulatory conformity	5.1, requirements 13 and 14; and 5.6	Verify the mandatory synchronization and protocol requirements set out in the Terms of Reference and record complementary observations on export and integration	Terms of Reference	Eliminatory only for the requirements expressly set out in the Terms of Reference; the other records are informative
Maintenance and support	5.2	Verify the mandatory warranty and support requirements	Terms of Reference	Eliminatory
Usability	5.3	Assess the user's interaction with the equipment	Clinical Engineering	Informative
Clinical performance	5.4	Assess the practical quality of the examinations	Ophthalmology and Clinical Engineering	Informative
Operational Stability	5.5	Assess functioning during real use of the equipment	Clinical Engineering	Informative
Connectivity and interoperability	5.6	Verify communication with external systems	Terms of Reference, Clinical Engineering and Telehealth integration requirements	Eliminatory when provided for in the Terms of Reference
Battery autonomy	5.7	Assess the operating time of the equipment	Clinical Engineering	Eliminatory when provided for in the Terms of Reference
Biosafety	5.8	Assess safety and hygiene aspects	Good practices	Informative, unless specifically provided for in the Terms of Reference
Patient Comfort	5.9	Assess the experience during the performance of the examination	Practical evaluation	Informative
Learning Curve	5.10	Assess the ease of training and operation	Practical evaluation	Informative
Productivity	5.11	Assess performance in real use	Practical evaluation	Informative
Alerts and error systems	5.12	Assess the capacity to identify and signal errors or abnormal conditions	Practical evaluation	Informative

Evaluation block	Item in the form	Purpose	Main source	Nature
Additional features/AI	5.13	Record complementary functionalities	Practical evaluation	Informative
Types of examination	6.0	Assess the performance of the equipment in carrying out the different types of examination, considering image quality, ease of capture, stability, clinical adequacy and the need for repetition	Practical evaluation	Informative, unless specifically provided for in the Terms of Reference

An anonymized version of the form used in the evaluation is provided as supplementary material, preserving the structure of the blocks, the response options, the decision rules and the linkage of the mandatory requirements to the Terms of Reference. Personal data, company identification, contacts, serial numbers, signatures, images of participants and other identifiable information are not available. The provision of the instrument is intended for methodological transparency and does not represent formal validation of its reproducibility.

The Terms of Reference required the equipment to be capable of synchronizing examinations in DICOM format with remote servers using FTP, SFTP, SMB and DICOMweb. These requirements were incorporated into item 5.1 of the form, specifically requirements 13 and 14. Item 5.6 was used to record complementary information on transmission, export, connection, integration and traceability of the examinations. During the demonstration, it was recorded that the equipment provided local export in DICOM, PDF and JPEG, but did not support the mandatory remote transmission protocols. Owing to the absence of these functionalities, no tests of transmission, failure rate, traceability or integration with an external system were carried out. Therefore, the evaluation demonstrated noncompliance with the connectivity requirements set out in the Terms of Reference, but did not constitute a formal conformity test with the DICOM standard, nor a validation of interoperability in a reference environment.

HL7 FHIR and IHE profiles are not included as requirements in the Terms of Reference, in the form or in the final opinion. Consequently, they were not assessed and must not be presented in this article as standards tested or demonstrated. Likewise, no reference server, endpoint, PACS, protocol version or conformity testing tool was documented.

TABLE II. TABLE 2 – MAPPING OF THE CONNECTIVITY AND INTEROPERABILITY CRITERIA

Requirement or functionality	Source	Nature	What was verified	Documented result
Synchronization of examinations in DICOM with remote servers	Terms of Reference; form 5.1, item 13	Eliminatory	Declared/ observed functional availability	The form records the item as met, but without a description of a test with a remote server
Support for FTP, SFTP, SMB and DICOMweb	Terms of Reference; form 5.1, item 14	Eliminatory	Availability of the required tools	Not met; the equipment provided only local DICOM
Export in DICOM, PDF and JPEG	Form 5.6, item 2	Informative	Local file export	Recorded as available
Wi-Fi or wired connection	Form 5.6, item 3	Informative	Network Connectivity	Not assessed
Transmission failures and time	Form 5.6	Informative	Transmission performance	Not assessed
Integration with a real system	Form 5.6, item 5	Informative, unless linked to a mandatory requirement	External functional integration	Not carried out
Traceability of the transmitted examination	Form 5.6, item 6	Informative	Retrieval and identification of the examinations	Not assessed

[a.] Note: "local DICOM" or the export of a DICOM file is not equivalent to a demonstration of full conformity with the DICOM standard, nor does it prove interoperability with external systems.

For the purposes of interpreting the results, the records were organized into four analytical categories:

1. Mandatory requirement of the tender documents: a requirement expressly provided for in the Public Notice or in the Terms of Reference, non-compliance with which could substantiate the rejection of the sample;

2. Observed technical or clinical quality: a result obtained during the demonstration, such as image quality, ease of capture and the need for repetition, recorded in order to characterize performance in the case studied;

3. Preference or functional characteristic: a potentially useful attribute, but one not defined as a mandatory requirement of the contract;

4. Complementary information or item not assessed: data recorded without any effect on the decision, or whose verification was not carried out.

For the administrative decision, only the first category had an eliminatory character. The remaining categories could not, in isolation, substantiate the rejection of the sample. This classification sought to ensure adherence to the principles of binding to the tender documents, objective judgement and equal treatment of bidders, while at the same time allowing the documentation of information relevant to future technology assessments.

In addition, any requirement whose absence would imply noncompliance with the mandatory specifications of the Terms of Reference, or would render unfeasible functionalities considered essential to the purpose of the solution evaluated, was regarded as an eliminatory criterion. The informative criteria were used exclusively to document the performance observed during the practical demonstration and could not, in isolation, justify the rejection of the sample.

In addition to the classification of the criteria, the instrument incorporated formal requirements of technical traceability. The decisions of the evaluation committee were to be supported by objective evidence, associated with the respective requirements of the Terms of Reference, so as to allow the auditing and reconstitution of the decision-making process.

The eliminatory requirements were linked to the requirements of the Terms of Reference, and the non-

conformities were recorded with an indication of the respective evidence. This structure favoured the documentary traceability of the decision in the process analysed. Although the structure of the instrument was designed to favour its application by other teams, the present study did not include a formal assessment of reproducibility or of inter-rater agreement. Furthermore, the requirements assessed were derived from the technical specifications previously defined for the technological solution, and no new criteria were introduced during the practical demonstration phase.

The specifications used in the preparation of the instrument were derived from the technical requirements previously established for the contract and subsequently incorporated into the Terms of Reference, preserving the correspondence between the requirements of the procurement process and the items actually assessed during the practical demonstration.

C. Application of the instrument

The evaluation of the sample was carried out in a real clinical environment, on the premises of UFG, during a practical demonstration conducted by a multiprofessional technical committee composed of specialists from the fields of ophthalmology, nursing, clinical engineering and information technology.

During the practical demonstration, the fundus photography modalities provided for in the Terms of Reference were performed, including colour, red-free, infrared and stereo optic disc images. The tests were conducted on different patients, allowing the observation of the performance of the equipment under distinct conditions of use. Aspects related to image capture, quality for clinical interpretation, operational stability, usability, connectivity, data transmission, availability of the functionalities provided for and other parameters covered by the evaluation instrument were observed.

The entire evaluation process was documented by means of photographic and video records, reports issued on the basis of the examinations performed, and the completion of the evaluation form by the technical team taking part in the demonstration. The evidence produced supported the preparation of the technical opinion of the joint committee, contributing to the traceability, transparency and substantiation of the decision regarding the approval or rejection of the sample.

The evaluation form was completed collectively by the technical team taking part in the demonstration, involving the joint consolidation of the evidence observed during the execution of the tests and of the technical discussions held throughout the evaluation.

The multiprofessional committee acted in accordance with its respective areas of competence, with the criteria of the instrument previously organized into eliminatory and informative categories, together with specific guidance for recording evidence and applying the decision rule. Before the practical demonstration, the participants were briefed on the classification of the criteria, the definition of the eliminatory and informative requirements and the procedures for recording evidence, with a view to promoting greater consistency in the evaluation process.

Before the practical demonstration, the participants received guidance on the classification of the criteria, the definition of the eliminatory and informative requirements, the procedures for recording evidence and the rules for conducting the evaluation. The role of the participants present was also clarified, namely that of observers only, including the representatives of the companies, who could not interfere in the execution of the tests or in the conduct of the evaluations, with a view to promoting greater consistency and impartiality in the evaluation process.

Once the practical demonstration had been concluded, the records of the form and the evidence collected were analysed internally by the evaluation committee. Any divergences of interpretation were discussed until a technical understanding was consolidated, and the final opinion was issued by the Joint Technical Committee by consensus, on the basis of the objective criteria provided for in the evaluation instrument and in the Terms of Reference.

The final result of the evaluation was based exclusively on the eliminatory criteria linked to the requirements of the Public Notice and the Terms of Reference. The informative criteria were used only to record the practical, operational and clinical performance of the equipment and could not, in isolation, justify the rejection of the sample. Accordingly, limitations related to usability, productivity or clinical quality supported the general technical analysis, while the administrative decision to reject the sample remained linked to non-compliance with the previously established mandatory requirements.

IV. RESULTS AND DISCUSSION

The application of the structured form made it possible to systematically assess technical, regulatory, functional and operational aspects of the portable fundus camera submitted to practical demonstration, extending the analysis beyond the mere documentary verification of the specifications set out in the tender documents [1], [2], [10], [11].

The organization of the instrument into eliminatory and informative criteria proved appropriate for distinguishing requirements that are indispensable to compliance with the demands of the tender process from those related to the operational performance of the technology. This structure made it possible to base the decision of the joint committee on objective, previously defined criteria, preserving the principles of binding to the tender documents, equal treatment and objective judgement, while at the same time recording information relevant to supporting future technology incorporations [3], [13], [14].

Among the eliminatory criteria, a non-conformity was identified in relation to support for the connectivity protocols required in the Terms of Reference. The document required the synchronization of examinations in DICOM format with remote servers using FTP, SFTP, SMB and DICOMweb. During the evaluation, it was recorded that the equipment provided only local DICOM functionality and did not offer the required protocols. For this reason, it was not possible to execute the remote transmission flows envisaged for the solution. Documentary omissions were also recorded at the time of the evaluation, including the technical manual and other documents indicated in the form and in the opinion. The rejection was based on non-compliance with these mandatory requirements, and not on the informative criteria of usability or clinical quality.

In addition to the eliminatory criteria, the form made it possible to record operational aspects which, although they did not determine the rejection of the sample, had a potential impact on the use of the equipment in a care environment. A frequent need for focus and positioning adjustments during

image capture was observed, along with a high rate of examination repetition and dependence on the operator for obtaining images of diagnostic quality. Of the seven examinations performed during the evaluation, four (57.1%) were considered adequate for clinical interpretation, while three (42.9%) presented insufficient quality for diagnostic purposes.

The findings indicate that the form was able to record, in the case studied, aspects of usability and image quality that were not sufficiently characterized in the documentary analysis. Although the observations made are restricted to the specific context of the evaluation, they made it possible to document operational characteristics relevant to understanding the performance of the equipment under conditions of use. In technologies intended for Primary Health Care and Telehealth, factors such as ease of operation, learning curve and stability during image acquisition become particularly relevant, since the equipment is often used by professionals with different levels of experience [9], [10].

Another relevant aspect identified during the evaluation was the difficulty of validating the interoperability of the equipment in a telehealth environment. In the case analysed, the unavailability of FTP, SFTP, SMB and DICOMweb prevented the execution of the envisaged remote flows. No formal conformity tests with interoperability standards or with specific reference environments were carried out. Accordingly, the evaluation focused on verifying the connectivity requirements set out in the Terms of Reference and on their practical demonstration during the evaluation process.

Although the equipment presented functionalities related to the storage and export of examinations, the impossibility of fully demonstrating communication using the protocols established in the tender documents highlighted the importance of including practical connectivity tests during sample evaluation processes [15]. Documentary verification alone proved insufficient to ensure that the integration requirements were effectively met under real conditions of use, since many devices declared to be compatible with the DICOM standard or with PACS systems frequently fail to provide full functional conformity when inserted into the institution's network, requiring the execution of practical proofs of concept in order to mitigate the risk of acquiring inoperable platforms [3], [13], [15].

During the practical demonstration, it was found that some functionalities described in the technical documentation submitted by the manufacturer could not be reproduced or demonstrated under the real conditions of use adopted during the evaluation, especially with regard to interoperability and examination transmission tests. This result shows that exclusively documentary analysis may not be sufficient to fully verify the expected functional performance of medical equipment, reinforcing the importance of carrying out practical evaluations, such as proofs of concept and clinical validation, as complementary stages of public procurement and Health Technology Assessment processes [3], [11], [13].

As a positive aspect, the equipment presented additional technological resources, including embedded artificial intelligence to support image analysis, autofocus functionalities and export of examinations in multiple formats. Although these resources add potential value to clinical practice, their presence does not replace compliance with the essential technical requirements established in the procurement process, reinforcing the need for structured evaluations that simultaneously consider regulatory, functional and operational aspects.

The results obtained indicate that the form contributed to organizing the record of the evidence produced during the practical demonstration, favouring the documentation of the findings observed by the evaluation committee and increasing the traceability of the technical analysis process. The distinction between eliminatory and informative criteria proved particularly relevant in making explicit the grounds for the decision adopted in the evaluation process, allowing the operational aspects observed during the demonstration to be recorded without compromising the binding of the decision to the mandatory requirements set out in the Public Notice and in the Terms of Reference.

Finally, the experience described in this study indicates that the adoption of structured evaluation instruments may contribute to qualifying the processes of incorporation and acquisition of health technologies within the Brazilian Unified Health System, favouring evaluation processes that are more transparent, better documented and more closely aligned with care needs.

A. Limitation of the study

The present study has limitations that should be taken into account when interpreting the results. It is a case study based on a single tender process and on the practical evaluation of a single technological solution submitted to the sample demonstration stage, in accordance with the judgement rules applicable to the procedure analysed. Accordingly, the results should not be generalized to all categories of medical devices.

In addition, the evaluation was conducted in a single institutional environment and did not include multicentre validation. The evaluation covered only the solution summoned to the sample analysis stage, owing to the judgement system adopted in the procedure, and it was therefore not possible to make comparisons between different pieces of equipment, suppliers or competing technologies. Formal measures of interrater reliability, internal consistency of the instrument and quantitative indicators capable of measuring impact on the efficiency of the procurement process, economic outcomes or clinical results were also not assessed. The evaluation process was conducted collectively, with the consolidation of the evidence by the technical team and by the Joint Technical Committee; no independent application of the instrument by different teams and no formal assessment of agreement between evaluators were carried out. Thus, the results obtained should be interpreted as evidence of the applicability of the methodology in the context studied, and not as definitive validation of the instrument for different scenarios of use.

V. CONCLUSION

The development and application of a structured form for the evaluation of samples demonstrated the feasibility of using a systematized approach to support the technical analysis of medical equipment in public procurement processes.

In the analysed case, the instrument contributed to organizing the record of the requirements assessed, of the evidence produced during the practical demonstration and of the elements considered in substantiating the technical decision.

The practical application showed that documentary analysis, although indispensable, may not be sufficient to fully verify certain functional and operational characteristics of health technologies. The practical demonstration made it

possible to identify limitations related to connectivity, image quality and the operational performance of the equipment evaluated, highlighting the relevance of complementary verification mechanisms when critical aspects cannot be fully ascertained solely by means of the documentation submitted by suppliers.

The organization of the criteria into eliminatory and informative categories proved appropriate for reconciling the objectivity required in tender processes with a more comprehensive evaluation of the performance of the technology. This structure made it possible to make explicit that the administrative decision was based exclusively on noncompliance with previously established mandatory requirements, while the other findings were recorded as complementary information characterizing the performance observed during the demonstration.

The experience described in this study indicates that structured evaluation instruments may contribute to making technical analysis processes more transparent, better documented and more traceable, especially in situations that require practical verification complementary to documentary analysis.

Finally, it should be stressed that the preparation and application of instruments of this nature require the participation of professionals with technical knowledge compatible with the technology being evaluated, since the definition of the criteria, the interpretation of the evidence and the conduct of the analysis depend on an adequate understanding of the regulatory, operational and clinical requirements involved. In this context, the experience reported reinforces the importance of the work of multiprofessional teams in the technical evaluation of health technologies.

VI. FUTURE PERSPECTIVES AND INSTRUMENT VALIDATION

Although it was developed to meet a specific need of the process analysed, the instrument has the potential to be adapted to similar evaluations of health technologies, provided that its criteria are defined in accordance with the characteristics of the object and with the requirements established for each contract.

Following the application described in this study, the methodology has come to be used as a reference for structuring instruments intended for the evaluation of other technologies, including portable ultrasound equipment assessed in a different institutional context. These experiences suggest the possibility of adapting the approach to different care and technological scenarios.

As indicated by the limitations of the present study, future applications may involve different evaluation teams, institutions and categories of medical devices, contributing to the methodological improvement of the instrument and to a more robust assessment of its reproducibility and applicability in different contexts.

In addition, future investigations may incorporate quantitative indicators aimed at assessing reliability, agreement between evaluators, reproducibility of results and the capacity of the instrument to identify relevant non-conformities during practical evaluation processes of health technologies.

TABLE III. TABLE 3 – INDICATORS PROPOSED FOR FUTURE STUDIES OF VALIDATION AND IMPROVEMENT OF THE INSTRUMENT

Dimension	Proposed Indicator	Form of future evaluation
Inter-rater agreement	Cohen's kappa, weighted kappa or Fleiss' kappa, according to the number of evaluators and the nature of the response	Independent application of the same instrument by two or more evaluators
Reproducibility between teams	Percentage of concordant decisions and chance-adjusted agreement	Re-evaluation of the same cases by independent teams
Decision traceability	Proportion of decisions with an explicit link between requirement, evidence and conclusion	Audit of the forms and opinions
Completeness of the instrument	Proportion of items completed, justified or classified as not assessed	Audit of the completion
Operational efficiency of the evaluation	Total application time, time per stage and number of complementary clarification requests	Prospective recording of times
Complementary detection capacity	Number and proportion of mandatory non-conformities identified in the demonstration and not identified in the documentary analysis	Prospective comparison between the documentary and practical stages
Internal consistency	Cronbach's alpha, only where there is a set of items measuring the same construct	Specific psychometric study
Content validity	Content validity index by a panel of specialists	Structured assessment of the items by specialists
External validity	Consistency of the results across different institutions, equipment and teams	Multicentre study
Quality of the decision	Previously defined indicators of completeness, adherence to the tender documents and support by evidence	Independent evaluation of the decision-making processes
Economic impact or risk reduction	Specific prospective indicators, with a comparison group or historical series	Not to be inferred from the present study

[b.] NOTE: The indicators above constitute a proposal for future studies and were not calculated in the present study. "Complementary detection of non-conformities" should not be presented as a synonym for proven risk reduction.

ACKNOWLEDGEMENTS

The authors thank the Federal University of Goiás (UFG) for providing the infrastructure and physical space required to conduct the practical demonstration and technical evaluation described in this study.

The authors further acknowledge the institutional partnership established with the Brazilian Ministry of Health, through the Secretariat of Primary Health Care (SAPS) and the General Coordination of Primary Health Care Programming and Financing (CGFAP), which contributed to enabling the activities related to the development of the project.

Finally, the authors thank the Brazilian Agency for Support to the Management of the Unified Health System (AgSUS) for its institutional partnership, support, and encouragement in carrying out this research.

REFERENCES

[1] H. P. Silva, C. A. Petramale, and F. T. S. Elias, “Avanços e desafios da Política Nacional de Gestão de Tecnologias em Saúde,” Revista de Saúde Pública, São Paulo, vol. 46, suppl. 1, pp. 83–90, 2012.
[2] D. V. Araújo, M. de S. C. Distrutti, and F. T. S. Elias, “Priorização de tecnologias em saúde: o caso brasileiro,” JBES, vol. 9, suppl. 1, pp. 4–40, 2017.
[3] F. V. Gonçalves Filho, “Nova Lei de Licitações: a padronização, a pré-qualificação e a análise de amostras como mecanismos para aquisições mais vantajosas para a administração,” Zênite Fácil, Doctrine category, Dec. 8, 2021.
[4] M. A. Hossain et al., “Necessity of clinical engineers to improve present health technology management in developing countries,” Modern Applied Science, vol. 9, no. 12, pp. 220–230, 2015.
[5] M. D. Bonfim and A. M. Malik, “The role of clinical engineering in management and decision-making in Brazilian hospitals,” Global Clinical Engineering Journal, vol. 8, no. 1, pp. 114–136, 2026.
[6] H. J. Teixeira, F. N. Silva, and S. M. Salomão, “A prática das compras públicas nos estados brasileiros: a inovação possível,” in Proc. 7th Congresso CONSAD de Gestão Pública, Brasília, DF, 2014. (Cited by C. C. C. Fernandes, “Compras Públicas no Brasil,” 2019.)
[7] M. Verga et al., “The digitization process and the evolution of clinical risk management concept: the role of clinical engineering in the operational management of biomedical technologies,” Frontiers in Public Health, vol. 11, no. 1121243, 2023.
[8] Y. David and T. Judd, “Evidence-based impact by clinical engineers on global patients outcomes,” Health and Technology, vol. 10, pp. 517–535, 2020.
[9] E. Liljegren, “Usability in a medical technology context: assessment of methods for usability evaluation of medical equipment,” International Journal of Industrial Ergonomics, vol. 36, pp. 345–352, 2006.
[10] M. S. G. Roma and E. de V. Garcia, “Medical device usability: literature review, current status, and challenges,” Research on Biomedical Engineering, vol. 36, 2020.
[11] J. C. Goldsack et al., “Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs),” npj Digital Medicine, vol. 3, no. 55, pp. 1–15, 2020.
[12] L. R. A. Cezário et al., “Telessaúde no Brasil: uma revisão de escopo,” Revista Baiana de Saúde Pública, Salvador, vol. 48, no. 2, e4010, 2024, doi: 10.22278/2318-2660.2024.v48.n2.a4010. [Online]. Available: https://rbsp.sesab.ba.gov.br/index.php/rbsp/article/view/4010. Accessed: Jun. 22, 2026.
[13] L. K. C. Silva and V. R. P. Nerio, “Qualidade versus conformidade nas compras públicas: avaliação de amostras de materiais médicohospitalares no Sistema Único de Saúde (SUS),” Revista Saúde em Redes, vol. 12, no. 1, 2026.
[14] V. R. R. Pinto, “Um breve histórico sobre inovações em compras e licitações públicas no Brasil,” Brazilian Journal of Development, Curitiba, vol. 6, no. 8, pp. 63378–63397, Aug. 2020.
[15] N. Gim, A. N. Ferguson, M. Blazes, C. S. Lee, and A. Y. Lee, “The march to harmonized imaging standards for retinal imaging,” Progress in Retinal and Eye Research, vol. 107, p. 101363, Jul. 2025.
[16] C. C. C. Fernandes, “Compras públicas no Brasil: tendências de inovação, avanços e dificuldades no período recente,” Administração Pública e Gestão Social, Viçosa, vol. 11, no. 4, Sep.–Dec. 2019. (Note: this is the source that cites the work of Teixeira, Silva and Salomão.)

Technological Analysis of Patient Positioning and Tracking Systems in Radiotherapy and Medical Imaging Equipment

Brenno Oliveira Canellas da Costa
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
brenno.oliveira@geb.inatel.br

Heloisa Marcela de Faria Paula
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
heloisa.marcela@geb.inatel.br

Monique Freire dos Santos
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
monique.freire@geb.inatel.br

Filipe Bueno Vilela
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
filipe.bueno@inatel.br

Luma Rissatti Borges do Prado
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
luma.prado@inatel.br

***Abstract*—This work presents a comparative analysis of emerging technologies for monitoring patient positioning in medical imaging and radiotherapy, focusing on computer vision systems, laser-based optical sensors, and radar sensors. A literature review was conducted considering monitored-area coverage, detection accuracy, susceptibility to interference, and technological versatility. In addition, a computer-vision-based solution was developed for the automatic identification of patient body displacement. The Mask R-CNN and Keypoint R-CNN models were implemented in Python using Google Colab to extract spatial parameters, including centroid position, body area, and anatomical keypoints. The results demonstrated the feasibility of these techniques for monitoring patient positioning and supporting clinical decision-making. The comparative analysis also highlighted the promising potential of computer-vision-based systems compared with the other technologies evaluated.**

***Index Terms*—Computer Vision; Image-Guided Radiotherapy; Patient Monitoring; Mask R-CNN; Keypoint R-CNN.**

I. Introduction

Proper patient positioning is a critical factor for the quality and safety of medical procedures that depend on high spatial accuracy, such as radiotherapy, computed tomography, magnetic resonance imaging, and other imaging examinations. Positioning consists of accurately aligning the body with the equipment coordinate system so that the region of interest is correctly located within the image acquisition field or therapeutic procedure. This process involves not only the initial setup but also continuous tracking and maintenance of the position throughout the clinical procedure, preventing voluntary or physiological movements from compromising image quality or treatment accuracy [1].

In computed tomography examinations, misalignments of only a few centimeters relative to the equipment isocenter can significantly increase radiation dose and degrade image quality. Displacements of 2, 4, and 6 cm were associated with increases of up to 13.5%, 33.3%, and 51.1% in surface dose, respectively, in addition to an average increase of 23% in total dose and 7% in image noise [2]. In radiotherapy, even systematic errors of approximately 0.6 mm may cause variations of approximately 2% in the equivalent dose delivered to the target volume, emphasizing the high sensitivity of treatments to small geometric inaccuracies [3]. These findings demonstrate the importance of accurate patient-position monitoring.

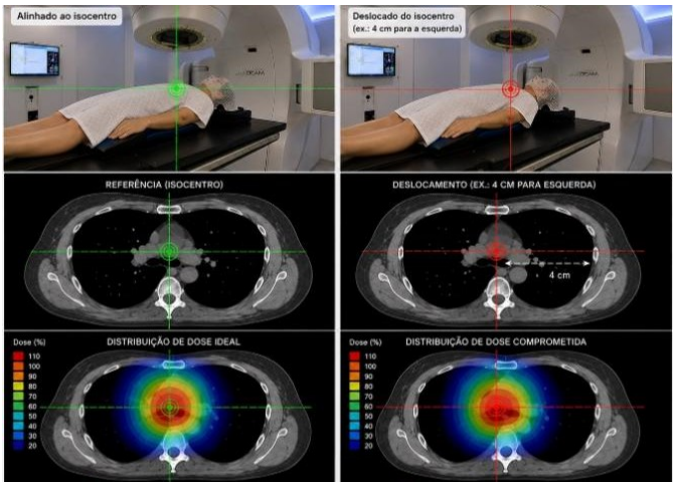


Fig. 1. Schematic representation of the effect of positioning deviations in radiotherapy. Correct alignment with the isocenter (left) is compared with a displaced condition (right).

To ensure appropriate patient alignment during radiotherapy treatments, different immobilization methods have been used, many of which are applied manually. These resources include

thermoplastic masks, used for head and neck tumors to immobilize the patient during treatment. In breast radiotherapy, a *breast board* is used to ensure proper alignment; this adjustable platform allows different inclinations and can be adapted to each patient [4]. Sedation has also become a common practice, particularly for pediatric patients, because movement during imaging procedures may compromise image quality [5].

To mitigate these challenges, technological advances in radiotherapy have enabled the development of systems for real-time monitoring and verification of patient positioning. One example is *ExacTrac Dynamic (BrainLab)*, which combines infrared cameras and X-ray images to ensure patient alignment during treatment. The system correlates external anatomical landmarks with internal structures visible in radiographic images, enabling automatic table adjustments and continuous position verification, thereby reducing deviations during the procedure [6].

Another example is *IDENTIFY (VARIAN)*, a *Surface Guided Radiation Therapy (SGRT)* system based on optical monitoring of the body surface. The system uses cameras to capture images and measure the distance to the patient's surface, reconstructing a three-dimensional body model in real time. It projects light patterns and analyzes their deformation to calculate surface geometry and continuously compare it with a previously acquired reference, allowing even small displacements to be identified [7].

Despite the effectiveness of systems currently used in clinical practice to monitor patients during radiotherapy and diagnostic imaging procedures, new technological approaches are being investigated to expand the availability of continuous and noninvasive methods. These technologies aim to reduce limitations associated with conventional techniques, such as the need for physical markers, implantable devices, or complex image acquisition systems.

Among the main current research areas are advances in computer-vision-based systems. Computer vision uses methods for image and video acquisition, processing, and interpretation to extract relevant information from the environment. Through cameras and algorithms such as three-dimensional reconstruction, image segmentation, and pattern tracking, these systems identify and monitor structures of interest noninvasively. In this context, SGRT has evolved through more robust algorithms for three-dimensional reconstruction of the patient's surface and real-time motion tracking, improving monitoring accuracy and reliability. Consequently, computer vision has enabled more accurate estimates of body position and markerless approaches to patient positioning [8].

Another promising approach involves millimeter-wave radar sensors for noninvasive monitoring of body position and movement during radiotherapy. These sensors emit and analyze signals in the 24 GHz range reflected by the body surface, enabling the detection of distance variations and micromovements associated, for example, with respiration or involuntary patient displacement.

Optical distance sensors have also been investigated for monitoring respiratory motion during medical examinations. In these systems, a laser sensor is directed toward the patient's chest or abdomen to measure distance variations caused by body expansion and contraction during breathing. Experimental studies have shown that this approach can accurately detect respiratory patterns during magnetic resonance imaging, with good agreement relative to conventional respiratory sensors used as references [9].

In this context, this study aims to comparatively evaluate emerging technologies for monitoring and positioning patients in medical equipment, particularly in applications requiring detection of physiological motion and small displacements, such as medical imaging and radiotherapy systems. The study also proposes a practical solution for monitoring patients positioned on imaging and radiotherapy tables, serving as a decision-support tool for identifying dynamic deviations from a reference position.

II. Methodology

A literature review was initially conducted to identify technologies used for monitoring and tracking patient positioning in clinical environments. The search was performed in *Google Scholar*, *PubMed*, and *IEEE Xplore*, considering publications from 2019 to 2026. Search terms related to sensors and software included "optical sensors," "radar sensors," and "computer vision," whereas clinical assessment terms included "electromagnetic interference," "interference in medical equipment," "advantages," and "disadvantages." Additional specific descriptors included "patient motion tracking," "motion monitoring in imaging examinations," "radiotherapy positioning," "surface-guided radiation therapy (SGRT)," and "motion correction in magnetic resonance imaging."

The selected studies addressed the following technological categories: first, computer-vision-based systems, particularly techniques used in SGRT, totaling two studies [8]; second, laser-based optical sensors for motion detection through reflection and the Doppler effect, comprising eight studies [10]–[14], [20]; and third, radar sensors used for vital-sign detection and contactless motion monitoring, totaling seven studies [15]–[18]. Studies related to electromagnetic interference and the compatibility of these technologies with clinical environments were also considered [13], [19]. Fig. 2 presents the classification of the main emerging technologies investigated in this study.

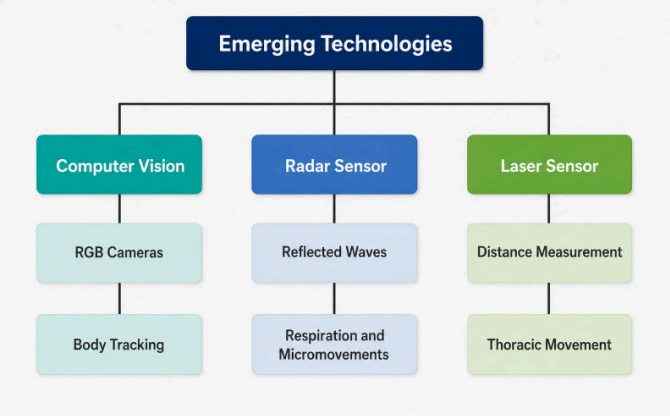


Fig. 2. Emerging technologies analyzed in this study for patient-position monitoring, including computer-vision-based systems, radar sensors, and laser sensors.

The technologies were compared according to the following parameters: monitored-region coverage, susceptibility to interference, technological versatility in examination environments, and motion-detection accuracy. A comparative table was then developed in which each technology was qualitatively classified as "poor," "fair," or "good," based on the synthesis of the literature findings and the parameters established for comparison.

A. Monitored-Region Coverage

Monitored-region coverage refers to the technology's ability to adequately cover the area of interest during an examination. This parameter was analyzed using different approaches from the literature, including laser tracking systems that depend on detector position [20] and body-surface monitoring through optical systems used in SGRT [8]. Studies on millimeter-wave motion tracking [15] and the operational parameters described in radar-sensor manuals [16] were also considered.

B. Susceptibility to Interference

Susceptibility to interference concerns the influence of external factors such as noise and data distortion. This parameter was analyzed from different perspectives reported in the selected studies. First, the influence of electromagnetic interference on image acquisition systems was investigated [19]. Interference associated with laser-based displacement measurements in civil structures [10] was then examined, along with a study assessing respiratory rate using a line laser and an RGB (*Red, Green and Blue*) camera [11]. Finally, radiofrequency-based systems were evaluated through a study on respiratory-motion monitoring using ultra-wideband radar [17].

C. Technological Versatility

Technological versatility in examination environments refers to the ability to adapt to different clinical contexts and operational configurations. This parameter was analyzed by considering operation in environments subject to electromagnetic interference and mitigation strategies such as filtering unwanted noise and signals [19]. Laser-based solutions were also investigated, including self-monitoring applications in breast-cancer radiotherapy [12], LiDAR (*Light Detection and Ranging*) technology for detecting individuals in indoor environments [20], its integration with detectors for displacement identification [8], and laser-tracking systems whose accuracy depends on detector position [20]. Applications of millimeter-wave sensors for identifying human activities in telemedicine were also analyzed [18].

D. Motion-Detection Accuracy

Motion-detection accuracy refers to the ability to correctly identify motion or physiological signals during monitoring. Different perspectives described in the literature were considered. First, accurate displacement tracking using optical systems in magnetic resonance imaging environments was analyzed [13]. Laser-based applications were then examined, including spectroscopy for respiratory-rate monitoring [11], [14], calibration and validation of prototypes for respiratory-motion detection using tracking techniques [14], and optical systems for measuring structural displacement in civil engineering [10]. Finally, millimeter-wave solutions were considered, including sensor-based applications [18] and investigations of motion tracking in radiotherapy using this frequency range [15].

E. Practical System for Preliminary Validation

For preliminary validation of the positioning-monitoring solution, four images were used to simulate different patient-positioning conditions. The images were produced with the assistance of the generative artificial intelligence tool *ChatGPT* (*Chat Generative Pre-trained Transformer*), using a reference scene representing a patient lying on a radiotherapy equipment table. Additional versions were subsequently generated with small changes in body position.

The solution was implemented in a cloud-computing environment using *Google Colab*, which enabled model execution with GPU support to accelerate image processing. The processing *pipeline* was developed in *Python* using libraries such as *Detectron2* for model detection and inference in segmentation and pose estimation, *OpenCV* for image reading and manipulation, *NumPy* for numerical operations and metric calculation, *Matplotlib* for graphical visualization of the results, and *Pandas* for tabular data organization. The source code was made publicly available in a *GitHub* repository at: https://github.com/fssmonique-debug/patient-position-monitoring.

Two approaches were tested and evaluated: instance segmentation using *Mask R-CNN*, which allowed the patient's silhouette to be extracted from each *frame*, and anatomical keypoint estimation using *Keypoint R-CNN*. Both models were pretrained on the COCO (*Common Objects in Context*) *dataset*. The parameters obtained from both models included detection *score*, centroid coordinates, segmented patient area, centroid displacement relative to the reference image, *bounding box* area, number of valid anatomical keypoints, and detected keypoint positions.

Parameters obtained in pixels, such as area and displacement, were converted into physical units in millimeters using a scale factor equivalent to the *Pixel Spacing* attribute in medical images. Based on the displacement values calculated in millimeters, tolerance thresholds were defined for automatic motion classification into three categories: mild, for displacements below 1 mm; moderate, for displacements between 1 and 2 mm; and critical, for displacements above 2 mm.

III. Results

A. Monitored-Region Coverage

Computer vision systems provide broad coverage of the body surface through three-dimensional reconstruction and continuous monitoring of extensive regions. This characteristic is observed both in SGRT systems using optical and infrared cameras [8] and in optical tracking systems applied to magnetic resonance imaging [13]. Radar sensors also provide broad coverage, determined by the sensor field of view and the spatial configuration of the system [15], [18]. The analyzed sensors have conical detection regions with range and angular aperture capable of covering extensive areas or the entire body surface [16]. Laser-based systems provide more localized coverage because measurement depends on beam incidence, detector position, and the selected region [11]. In LiDAR sensors, although the horizontal field of view is broad, the vertical field of view of approximately 33.2° may limit acquisition when the sensor is positioned near the patient, allowing only a body strip to be observed [20]. This characteristic reduces system flexibility in environments with positioning restrictions.

B. Susceptibility to Interference

In computer vision systems, the main sources of interference are associated with electromagnetic fields in magnetic resonance imaging environments, which may distort image-sensor data [19]. However, measures such as grounding, electromagnetic control [19], and the use of adapted optical systems [13] make this limitation less critical. Radar sensors are more sensitive to the physical conditions of the environment because they operate by transmitting and receiving electromagnetic waves. Metallic surfaces may produce multiple reflections, multipath propagation, and unwanted echoes [15], [16]. Manufacturers also report installation restrictions associated with metallic objects and moving elements [16]. In radiotherapy, gantry motion may significantly reduce the system signal-to-noise ratio and impair respiratory-motion detection [17]. In laser-based systems, interference is associated with obstacles in the beam path, sensor alignment, and the optical properties of the monitored surface. Rapid changes in patient position or orientation may require continuous updating of the spatial tracking model, increasing tracking complexity and potentially reducing measurement accuracy [11], [20].

C. Technological Versatility

Computer vision provides greater clinical versatility and has established applications in radiotherapy [8]. In magnetic resonance imaging, adaptations are required to ensure sensor electromagnetic compatibility [13], [19]. Even so, the technology can be used in different clinical environments when these requirements are met. Radar sensors provide lower versatility in clinical environments because of electromagnetic-wave emission and susceptibility to reflections and echoes in rooms containing shielding or metallic structures [15], [16]. These characteristics make installation more challenging in environments with electromagnetic and structural restrictions. Laser-based systems adapt well to the physical environment because they do not depend on broad electromagnetic emission [11], [20]. However, beam incidence must be controlled in sensitive regions such as the eyes, face, and areas of injured or fragile skin [11], [14]. Therefore, these systems provide good environmental versatility but require greater attention to potential patient risks.

D. Motion-Detection Accuracy

Computer vision systems provide high accuracy. In magnetic resonance imaging, reported accuracies were approximately 0.1 mm for translation and 0.15° for rotation [13]. In radiotherapy, these systems enable continuous monitoring of millimeter-scale deviations of the body surface [8]. Radar sensors can detect body-surface micromovements and respiratory motion in real time [15]–[17]. One study reported an accuracy of 94% even under the critical condition of apnea and in the presence of environmental interference [18], indicating high accuracy and robustness for respiratory monitoring. Laser-based systems provide high accuracy in the directly monitored region, particularly for respiratory tracking and displacement measurement [10], [11], [14]. However, movement outside the illuminated region may not be detected with the same efficiency [11], [20]. Accuracy also depends on the distance between the sensor and the target. Although some systems operate at distances of up to 150 m, the study in [20] used a distance of 8 m to increase the number of points projected onto the target and improve measurement reliability.

E. Comparative Summary

Computer vision provided the most balanced set of characteristics, including broad coverage, high accuracy, and greater clinical versatility. Despite the need for adaptations in magnetic resonance imaging environments, it was considered the most appropriate technology for the proposed application and was therefore selected for practical testing. Radar sensors also provided good coverage and accuracy but showed more relevant limitations related to environmental interference and adaptation to clinical rooms. Laser-based systems provided high accuracy for localized measurements but had more restricted spatial coverage and greater dependence on positioning, alignment, and system configuration. Based on these findings, Table I, proposed in the methodology, was completed.

Table I
COMPARISON OF THE TECHNOLOGIES ACCORDING TO THE ANALYZED PARAMETERS.

Parameter	Computer Vision	Laser	Radar
Monitored-region coverage	Good	Poor	Good
Susceptibility to interference	Fair	Fair	Poor
Technological versatility	Fair	Fair	Poor
Motion-detection accuracy	Good	Fair	Good

F. Preliminary Validation

After implementing the code in the *Google Colab* cloud-computing environment, the images were imported. A binary mask was generated for each image: pixels belonging to the patient were assigned a value of 1 and represented as the white region, whereas the remaining pixels were assigned a value of 0 and represented as the black background. This binary mask enabled identification of the area occupied by the patient and served as the basis for calculating spatial parameters.

The main patient spatial parameters were also calculated. The centroid was determined from the mean coordinates of the pixels belonging to the mask. Centroid coordinates were expressed as row (L) and column (C), corresponding to pixel positions in the image. The masks with centroid markings were displayed using the *Matplotlib* library. For the reference mask (a), the centroid was located at $L = 446$ and $C = 474.5$; for mask (b), it was located at $L = 568$ and $C = 494.1$; for mask (c), at $L = 560$ and $C = 495.7$; and for mask (d), at $L = 685.4$ and $C = 613.2$.

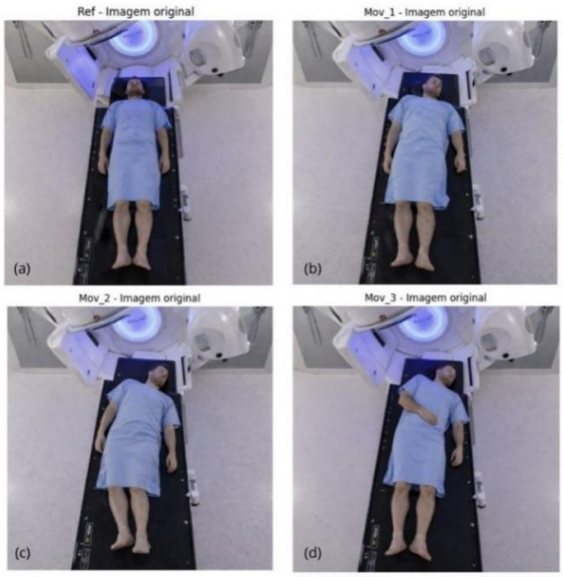


Fig. 3. Set of images used in the models: patient in the initial position (ref.), patient with a slight lateral displacement relative to the table axis (mov_1), patient with slight trunk rotation relative to the longitudinal table axis (mov_2), and patient with lateral displacement combined with slight upper-limb flexion (mov_3).

F.1 – Instance Segmentation (Mask R-CNN): After implementation in *Google Colab*, the images were imported and converted into binary masks. Pixels corresponding to the patient were assigned a value of 1 and represented by the white region, whereas the background was assigned a value of 0 and represented by the black region. These masks enabled identification of the area occupied by the patient and calculation of spatial parameters.

The centroid was calculated from the mean coordinates of the pixels belonging to the mask and represented by a red point in the images. Its coordinates were expressed as row (L) and column (C), in pixels. The masks were displayed using the *Matplotlib* library.

For the reference mask (a), the calculated centroid was $L = 446$ and $C = 474.5$. For the masks with displacement, the following values were obtained: (b) $L = 568$ and $C = 494.1$; (c) $L = 560$ and $C = 495.7$; and (d) $L = 685.4$ and $C = 613.2$.

For each image, the *Mask R-CNN* detection *score* was obtained, with values ranging from 0 to 1. Values closer to 1 indicate greater confidence in identifying the person class. Table II presents the main parameters obtained in pixels.

Table II
PARAMETERS OBTAINED, INCLUDING THE DETECTION *score* FROM THE *Mask R-CNN* MODEL IN THE *Detectron2* LIBRARY AND THE CENTROID ROW AND COLUMN COORDINATES.

Frame	Score	Centroid (row)	Centroid (column)
Ref.	0.997	446.01	474.45
Mov_1	0.997	567.96	494.11
Mov_2	0.998	565.96	495.67
Mov_3	0.996	685.37	613.19

The displacement values obtained for the motion images were expressed in pixels and subsequently converted into millimeters, as was the area. A scale factor of 0.1 mm was used to convert the measurements into physical units. This factor is equivalent to the *Pixel Spacing* commonly available in DICOM (*Digital Imaging and Communications in Medicine*) images. Table III was then constructed.

Table III
PARAMETERS OBTAINED, INCLUDING THE PATIENT AREA IN THE IMAGE FROM THE *Mask R-CNN* MODEL IN THE *Detectron2* LIBRARY AND CENTROID DISPLACEMENT ACROSS THE IMAGES IN PIXELS AND MILLIMETERS.

Frame	Area (px)	Area (mm^2)	Displacement (px)	Displacement (mm)
Ref.	86683	866.83	0.00	0.00
Mov_1	149656	1496.56	123.52	12.35
Mov_2	150761	1507.61	121.81	12.18
Mov_3	208186	2081.86	276.66	27.67

The displacement values in millimeters were classified according to tolerance thresholds used in high-precision radiotherapy positioning systems: mild for values below 1 mm, moderate for values between 1 and 2 mm, and critical for values above 2 mm. These thresholds were based on the

approximately 1–2 mm tolerances adopted in systems such as *ExacTrac Dynamic* [21]. Table IV presents the classification obtained for each image.

Table IV
CENTROID DISPLACEMENT IN MILLIMETERS, MOTION CLASSIFICATION, AND SIMULATED SYSTEM ACTION FOR DIFFERENT PATIENT POSITIONS RELATIVE TO THE REFERENCE IMAGE.

Frame	Displacement (mm)	Classification	System action
Ref.	0.00	Mild	Normal operation
Mov_1	12.35	Critical	Immediately interrupt acquisition / beam
Mov_2	12.18	Critical	Immediately interrupt acquisition / beam
Mov_3	27.67	Critical	Immediately interrupt acquisition / beam

F.2 – Keypoint Detection (Keypoint R-CNN): The same four images were imported for a different analysis approach using the *Keypoint R-CNN* model. Each image was processed individually to detect the person in the *frame* and extract anatomical body keypoints. First, an image containing the keypoints, centroid marker, and body *bounding box* was generated, as shown in Fig. 4. The keypoints were represented as green points at specific anatomical regions such as the shoulders, arms, hips, and legs. The *bounding box* was represented by a red outline delimiting the region occupied by the patient.

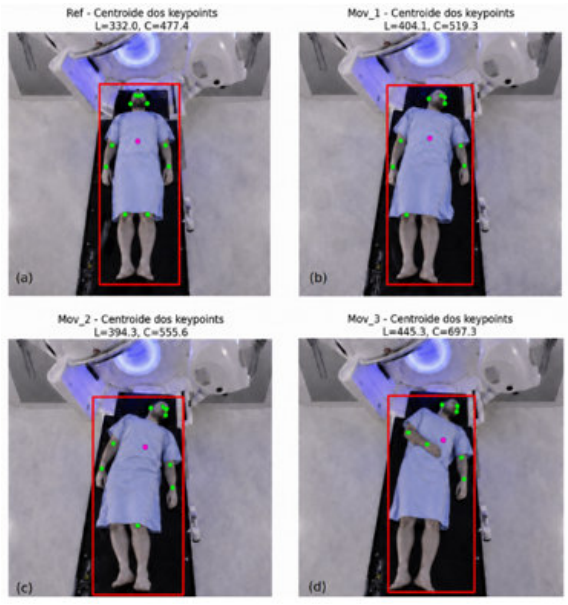


Fig. 4. *Keypoint R-CNN* detection results for the four analyzed images. Images (a), (b), (c), and (d) show the anatomical keypoints (green points), the *bounding box* (red rectangle), and the calculated body centroid (central pink point). Centroid coordinates are expressed as row and column.

Spatial parameters were calculated from the detected *keypoints*. The model used a confidence threshold of 0.7 for person-instance detection and an additional threshold of 0.5 for validating individual keypoints, and only anatomical points above this value were considered. The centroid was obtained from the mean coordinates of the valid points and expressed as row (L) and column (C). The number of valid keypoints and the main detection *score* were also recorded. The results are presented in Table V.

Table V
QUANTITATIVE PARAMETERS EXTRACTED BY THE *Keypoint R-CNN* MODEL FOR THE FOUR IMAGES, INCLUDING THE PERSON-INSTANCE DETECTION *score*, CENTROID COORDINATES IN PIXELS, AND THE NUMBER OF VALID DETECTED KEYPOINTS.

Frame	*Score*	Centroid (row)	Centroid (column)	No. of valid *keypoints*
Ref.	1.000	331.99	477.37	11
Mov_1	1.000	404.08	519.30	9
Mov_2	1.000	394.25	555.61	11
Mov_3	1.000	445.26	697.27	7

The area of the patient-associated *bounding box* was calculated from the dimensions of the delimited region. Patient displacement was calculated as the difference between the centroid of the anatomical keypoints in the current image and the centroid in the reference image. Displacement values were initially obtained in pixels and then converted into millimeters, considering that 1 pixel corresponds to 0.1 mm. Area values were converted in the same manner, enabling a physical interpretation closer to a real clinical scenario. Table VI was constructed from these parameters.

Table VI
PARAMETERS OBTAINED WITH THE *Keypoint R-CNN* MODEL, INCLUDING THE PATIENT-ASSOCIATED *bounding box* AREA AND DISPLACEMENT OF THE ANATOMICAL-KEYPOINT CENTROID ACROSS THE IMAGES, EXPRESSED IN PIXELS AND MILLIMETERS.

Frame	Box area (px^2)	Box area (mm^2)	Displacement (px)	Displacement (mm)
Ref.	142689.06	1426.89	0.00	0.00
Mov_1	243481.83	2434.82	83.40	8.34
Mov_2	251717.70	2517.18	99.99	10.00
Mov_3	363800.53	3638.01	247.36	24.74

The displacement values in millimeters were then classified as mild, moderate, or critical using the same tolerance thresholds adopted for the previous *Mask R-CNN* model. Table VII presents the results obtained from the image analysis.

Table VII
DISPLACEMENT OF THE KEYPOINT CENTROID IN MILLIMETERS, MOTION CLASSIFICATION, AND SIMULATED SYSTEM ACTION FOR DIFFERENT PATIENT POSITIONS.

Frame	Displacement (mm)	Classification	System action
Ref.	0.00	Mild	Normal operation
Mov_1	8.34	Critical	Immediately interrupt acquisition / beam
Mov_2	10.00	Critical	Immediately interrupt acquisition / beam
Mov_3	24.74	Critical	Immediately interrupt acquisition / beam

IV. Discussion

A. General Analysis

The comparative analysis showed that the technologies investigated in this study perform differently when applied to patient-motion monitoring in clinical environments according to the defined parameters. Overall, computer vision provided the most balanced set of characteristics among the evaluated approaches, particularly because of its broader monitored-region coverage, good motion-detection accuracy, and greater versatility across different clinical contexts.

B. Discussion of the Comparative Parameters

Coverage was analyzed according to the extent of the observed region and the ability to track movement along the patient's body. In this respect, computer vision and radar provided better performance, whereas laser-based systems offered more localized coverage [11], [20].

Susceptibility to interference was evaluated according to its impact on detection reliability in clinical environments with physical and electromagnetic restrictions, such as shielded diagnostic imaging and radiotherapy rooms. Under these conditions, radar sensors were more vulnerable to reflections, echoes, and environmental interference [15]–[17].

Versatility was assessed according to the ability of each technology to adapt to different rooms, equipment, and procedures without disrupting clinical workflow. Accuracy was interpreted together with the operating principle of each technology, distinguishing broad measurements from localized detection. Therefore, the most appropriate solution must be selected through an integrated analysis of the parameters, because good performance in an isolated criterion does not necessarily indicate the best clinical application.

C. Considerations Regarding the Selection of Computer Vision

Some factors that could initially be interpreted as limitations of computer vision were not sufficient to prevent its application. Low-light environments, for example, do not necessarily constitute a critical obstacle for optical systems because many clinical arrangements use infrared cameras or acquisition methods adapted to these conditions. Similarly, although magnetic resonance imaging environments impose additional electromagnetic-compatibility challenges, the analyzed studies describe adaptation and mitigation strategies for these scenarios [13], [19].

Among the evaluated technologies, computer vision therefore presented the most manageable limitations and the most favorable balance for monitoring patient positioning. The results reinforce the importance of this project by demonstrating the feasibility of investigating these technologies in clinical environments and by emphasizing the need for more specific and quantitative studies of their practical application. Such progress may support the development of systems capable of improving the safety, accuracy, and reliability of clinical procedures.

D. Proposed Practical Solution

The results demonstrated the feasibility of using computer vision to monitor patient positioning in simulated diagnostic imaging and radiotherapy scenarios. However, because suitable public clinical *datasets* were unavailable, artificial-intelligence-generated images were used to represent different positional changes. The sequence of these images simulated the basic behavior of a video composed of successive *frames*.

In the model comparison, *Mask R-CNN* provided greater stability in silhouette segmentation and body-area estimation. *Keypoint R-CNN* was more sensitive to postural changes because the calculated centroid directly depends on the positions of the detected anatomical points. Consequently, movements of the arms, legs, or head produced more significant variations in the calculated displacement.

Both models were pretrained on the COCO *dataset* and were not specifically developed for clinical applications, which contributed to the limitations observed in segmentation and anatomical-point detection.

This study reinforces the importance of research on patient positioning for improving procedural safety and reducing risks such as unnecessary radiation exposure to healthy tissues [2], [3]. It also demonstrates the potential to develop more accurate and accessible solutions for continuous monitoring without necessarily increasing operating costs [8], [9], [15].

V. Conclusion

The results demonstrated that computer vision, laser-based sensors, and radar sensors have potential for monitoring patient positioning and movement in imaging examinations and radiotherapy, although they provide different levels of suitability. Computer vision showed the most balanced set of characteristics, with broad coverage, good accuracy, and greater versatility, and was therefore considered the most promising approach. Radar provided good coverage and accuracy but was more susceptible to interference, whereas laser-based systems were suitable for localized measurements but offered more restricted coverage and greater dependence on system configuration. The practical stage confirmed the technical feasibility of using convolutional neural networks in simulated scenarios: *Mask R-CNN* was more stable in segmenting the body region, whereas *Keypoint R-CNN* was more sensitive to postural changes. Limitations associated with the scarcity of studies, public clinical data, and quantitative information on the effects of improper positioning were also identified. Although this study did not develop a system validated for hospital use, it demonstrated the computational feasibility of the proposal and established a basis for future research on patient monitoring, computer vision, and decision support in clinical procedures.

References

[1] W. Guo, N. Meier, R. Müller-Polyzou, A. Georgiadis, and Z. Chen, "Patient positioning in radiotherapy," in *Proc. IEEE Int. Symp. Medical Measurements and Applications (MeMeA)*, Bari, Italy, 2020, doi: 10.1109/MeMeA49120.2020.9137355.

[2] M. A. Habibzadeh, M. R. Ay, A. R. Kamali Asl, H. Ghadiri, and H. Zaidi, "Impact of miscentering on patient dose and image noise in X-ray CT imaging: Phantom and clinical studies," *Physica Medica*, vol. 28, no. 3, pp. 191–199, 2012.

[3] J.-H. Cai, X. Peng, and J.-Y. Lu, "Dosimetric and radiobiological impact of patient setup errors in intensity-modulated radiotherapy for esophageal cancer," *Technology in Cancer Research & Treatment*, 2025, doi: 10.1177/15330338241311136.

[4] Brazil, Ministry of Health, National Cancer Institute, *Update for Radiotherapy Technicians*. Rio de Janeiro, Brazil: INCA, 2010, pp. 39–44.

[5] V. R. Badveli *et al.*, "Effective MRI practices for pediatric patients: Optimizing imaging protocols for safety and quality," *Advanced Radiology Imaging*, vol. 2, no. 2, pp. 24–31, 2025, doi: 10.4274/AdvRadiolImaging.galenos.2025.02486.

[6] V. S. Mendes *et al.*, "ExacTrac Dynamic workflow evaluation: Combined surface optical/thermal imaging and X-ray positioning," *Journal of Applied Clinical Medical Physics*, vol. 23, 2022, doi: 10.1002/acm2.13754.

[7] E. L. Covington *et al.*, "Commissioning and clinical evaluation of the IDENTIFY surface imaging system for frameless stereotactic radiosurgery," *Journal of Applied Clinical Medical Physics*, vol. 24, 2023, doi: 10.1002/acm2.14058.

[8] P. Freislederer *et al.*, "Surface guided radiation therapy (SGRT): Review of clinical applications," *Radiation Oncology*, vol. 15, 2020, doi: 10.1186/s13014-020-01629-w.

[9] J. W. Krug, R. Odenbach, A. Boese, and M. Friebe, "Contactless respiratory monitoring system for magnetic resonance imaging applications using a laser range sensor," *Current Directions in Biomedical Engineering*, vol. 2, no. 1, pp. 719–722, 2016, doi: 10.1515/cdbme-2016-0156.

[10] K.-H. Kim and H.-K. Jung, "Development of a remote displacement measuring laser system for bridge inspection," *Sensors*, vol. 22, no. 5, art. 1963, 2022, doi: 10.3390/s22051963.

[11] Y. Jeong *et al.*, "For a clinical application of optical triangulation to assess respiratory rate using an RGB camera and a line laser," *BMC Medical Imaging*, vol. 24, art. 274, 2024, doi: 10.1186/s12880-024-01448-5.

[12] D. W. Kim *et al.*, "Study to improve reproducibility and accuracy of DIBH for breast cancer radiotherapy: A laser-based self-monitoring system," *International Journal of Radiation Research*, vol. 23, pp. 659–664, 2025, doi: 10.61882/ijrr.23.3.21.

[13] M. Zaitsev *et al.*, "Motion artifacts in MRI: A complex problem with many partial solutions," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 4, pp. 887–901, 2015. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3523280/

[14] I. N. Islami *et al.*, "Calibration and volunteer testing of a prototype contactless respiratory motion detection system based on laser tracking," *Journal of Applied Clinical Medical Physics*, vol. 26, art. e14607, 2025, doi: 10.1002/acm2.14607.

[15] J. Olick-Gibson *et al.*, "Feasibility study of surface motion tracking with millimeter wave technology during radiotherapy," *Medical Physics*, vol. 47, pp. 1229–1237, 2020, doi: 10.1002/mp.13980.

[16] Seeed Technology Co., Ltd., *24GHz mmWave Sensor: Human Static Presence Module Lite Datasheet*, 2023.

[17] A. Fallatah, M. Bolic, M. MacPherson, and D. J. La Russa, "Monitoring respiratory motion during VMAT treatment delivery using ultra-wideband radar," *Sensors*, vol. 22, no. 6, art. 2287, 2022, doi: 10.3390/s22062287.

[18] A. Soumya, C. Krishna Mohan, and L. R. Cenkeramaddi, "Recent advances in mmWave-radar-based sensing, its applications, and machine learning techniques: A review," *Sensors*, vol. 23, no. 21, art. 8901, 2023, doi: 10.3390/s23218901.

[19] Y. Liu *et al.*, "Subject grounding for electromagnetic interference reduction in MRI systems," *Magnetic Resonance in Medicine*, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12681293/

[20] J. Gómez, O. Aycard, and J. Baber, "Efficient detection and tracking of human using 3D LiDAR sensor," *Sensors*, vol. 23, no. 10, art. 4720, 2023, doi: 10.3390/s23104720.

[21] M. Poltorak, K. Benisz, A. Spyra, M. Szwast, and I. Walecka, "Optimizing isocenter accuracy in ExacTrac Dynamic: Daily verification and quality assurance for reliable patient positioning," *Polish Journal of Medical Physics and Engineering*, vol. 32, no. 1, 2026, doi: 10.2478/pjmpe-2026-0002.

Technological Mapping of Minimally Invasive Devices for Blood Collection: A Systematic Review from the Perspective of Biomedical Engineering and Product Engineering

1st Thaise Silveira dos Santos Apolinário
dept. of Program in Production Engineering
UFRN
Natal, Brasil
thaise.apolinario.049@ufrn.edu.br

2nd Helio Roberto Hekis
dept. of Biomedical Engineering
UFRN
Natal, Brasil
helio.hekis@ufrn.br

3nd Mario Oreste Aguirre Gonzalez
dept. of Program in Production Engineering
UFRN
Natal, Brasil
mariogonzalez@ufrn.br

4nd Francisca Thais Queiroz Rêgo
dept. of Biomedical Engineering
UFRN
Natal, Brasil
thais.rego.092@ufrn.edu.br

***Abstract*—Blood collection is an essential procedure for the diagnosis, monitoring, and therapeutic management of various clinical conditions. However, conventional venipuncture presents limitations regarding pain, discomfort, vascular access difficulties, and occupational risks. In this context, minimally invasive technologies have been developed to enhance safety, improve the patient experience, and increase the efficiency of diagnostic processes. This study conducted a Systematic Literature Review (SLR), following PRISMA 2020 guidelines, aiming to identify and analyze the main technologies employed in minimally invasive blood collection from the perspectives of Biomedical Engineering and Product Engineering. Searches were performed in the PubMed, Scopus, and Web of Science databases, covering studies published between 2015 and 2026. The selected works were grouped into technological categories, encompassing microneedle-based devices, capillary collection systems, blood microsampling technologies, and biosensors, as well as aspects related to usability and regulation. The results highlighted the advancement of these technologies, driven by developments in biomaterials, microfabrication techniques, and integration with smart systems. Despite the significant potential for pain reduction, improved sample quality, and expanded remote monitoring, challenges regarding clinical validation, industrial scalability, and regulatory compliance remain. It is concluded that Biomedical Engineering plays a fundamental role in developing innovative, safe, and user-centered solutions, contributing to the evolution of devices designed for minimally invasive blood collection.**

***Keywords*—Blood collection; Minimally invasive technologies; Biomedical engineering; Product engineering.**

I. Introduction

Blood collection is one of the most frequently performed procedures in clinical practice, essential for the diagnosis, therapeutic monitoring, and management of various diseases. Although conventional venipuncture remains the most widely used method, it is associated with pain, discomfort, vascular access difficulties in specific patient groups, and occupational risks arising from the use of sharp instruments.

Despite their widespread use, traditional collection methods still present significant limitations such as pain, discomfort, venipuncture failure, and variable sample quality which can compromise test results and necessitate repeat procedures in healthcare settings [1].

In recent years, the development of minimally invasive technologies has provided alternatives capable of mitigating these issues; notable examples include microneedle-based devices, microfluidic systems, needle-free technologies, wearable biosensors, and blood microsampling platforms.

Hoffmann, point out that the perception of pain is related not just to the method used, but also to the technique applied, the professional's experience, and the patient's venous conditions and the systems they use [6].

According to data from ANVISA, approximately 15,501 interruptions to blood collection procedures were recorded in Brazil (with 306 in Rio Grande do Norte), with venipuncture difficulty being one of the primary associated causes [3].

Despite the significant growth in scientific output in this field, challenges remain regarding clinical validation, standardization, regulation, and the integration of these technologies into healthcare services. Thus, it is relevant to gather and critically analyze available evidence, identifying technological trends and opportunities for the development of new biomedical devices.

The incorporation of advanced technologies in the healthcare sector has brought about profound transformations in how clinical processes are conceived, organized, and executed.

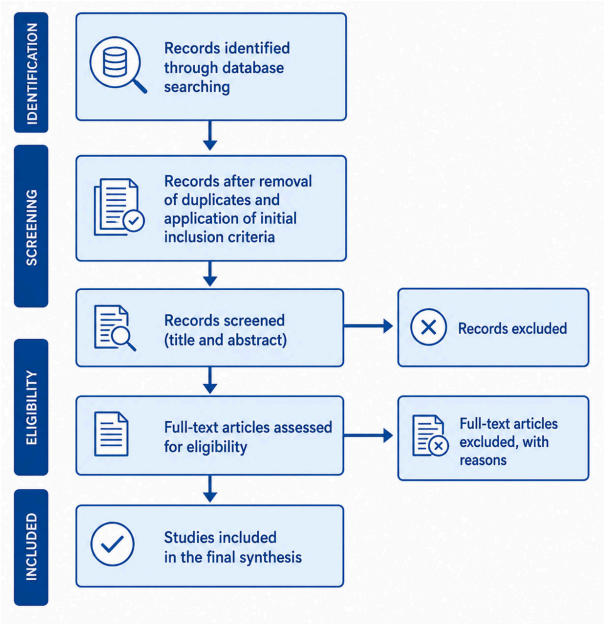


Fig. 1. Study search and selection strategy. Flowchart of the phases of the PRISMA method. Source: adapted from Moher et al. (2009)

These innovations not only expand diagnostic and therapeutic capabilities but also reshape the roles of patients and healthcare professionals in establishing a more integrated and personalized model of care [4].

Accordingly, this study aims to map and analyze minimally invasive blood collection technologies, identifying their technical characteristics and discussing how Product Engineering principles contribute to the development of devices that are safer, more ergonomic, reliable, and user-centered.

II. Materials and Methods

This study is a Systematic Literature Review (SLR) conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. The objective was to identify, analyze, and synthesize scientific evidence regarding minimally invasive technologies used for blood collection, considering technological, clinical, and product development aspects.

A. Search strategy

The bibliographic search was conducted in the Scopus, PubMed, and Web of Science databases in 2026, covering publications from 2015 to 2026. The search strategy combined Medical Subject Headings (MeSH) and free-text terms using Boolean operators. The core search string employed was:

(”blood collection” OR ”blood sampling” OR ”capillary blood”) AND (”microneedle” OR ”microsampling” OR ”minimally invasive” OR ”microfluidic” OR ”biosensor”) AND (”product engineering” OR ”usability” OR ”ergonomics” OR ”device design” OR ”point-of-care”)

English descriptors combined with Boolean operators were used, including the terms blood collection, microneedles, microfluidics, capillary blood collection, minimally invasive, point-of-care, wearable biosensors, and product engineering. The searches covered publications from 2015 to 2026, a period characterized by the growth of research related to the development of minimally invasive biomedical devices.

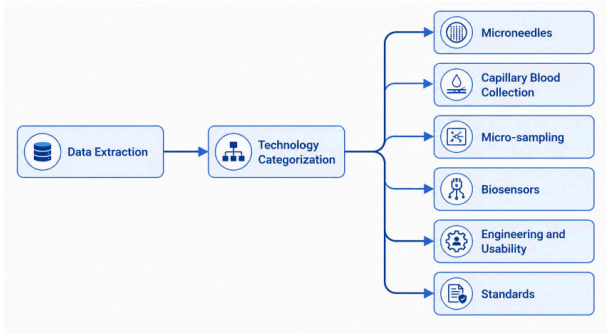


Fig. 2. Search and study selection strategy.Source: Author (2026)

B. Eligibility criteria

Original articles published in journals or scientific conference proceedings were included if they were available in full text, written in English or Portuguese, and related to the development, evaluation, or application of technologies for minimally invasive blood collection.

Duplicate studies, literature reviews, editorials, extended abstracts lacking full text, and works that did not directly address technologies for obtaining blood samples were excluded.

C. Selection of studies

The selection of studies was carried out in successive stages. Initially, duplicate records were removed. Subsequently, titles and abstracts were reviewed to assess their relevance to the topic. Potentially eligible articles underwent full-text evaluation, resulting in the final selection of studies included in the systematic review. The process of identification, screening, eligibility assessment, and inclusion was organized according to the PRISMA 2020 flow diagram.

D. Data extraction and analysis

Information from the selected studies was extracted in a standardized manner, covering authors, year of publication, technology employed, objective, methodology, main results, advantages, limitations, and clinical applications.

Subsequently, the articles were grouped into technological categories to facilitate comparative analysis:

- microneedle-based devices;
- minimally invasive capillary collection systems;
- blood microsampling technologies;
- biosensors and smart devices;
- product engineering and usability aspects;
- applicable standards and regulations.

The synthesis of the results was conducted qualitatively, identifying technological trends, development challenges, and

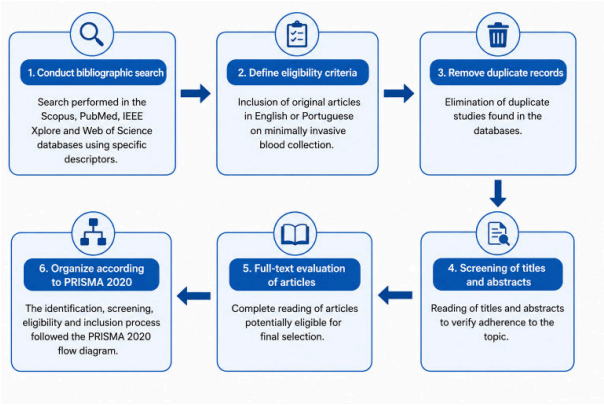


Fig. 3. Data Extraction and Analysis Process.Source: Author (2026)

opportunities for innovation in the field of Biomedical Product Engineering.

III. Results

The search strategy resulted in the identification of studies related to the development of devices for minimally invasive blood collection. After applying eligibility criteria and removing duplicate records, the studies considered most relevant were selected for inclusion in the systematic review.

The included studies demonstrated significant growth in research over the last decade, driven primarily by advances in microfabrication, biomaterials, microfluidics, and biosensors.

The analysis enabled the classification of the technologies into six main categories, presented in Table I.

TABLE I
Criteria used for the analysis of the selected studies

Criterion	Description
Technology	Type minimally invasive device used for blood collection.
Clinical application	Purpose of the technology (diagnosis, monitoring, or laboratory tests).
Degree and level	Invasiveness the procedure (minimal, low, or moderate).
Sample volume and amount	Blood required for collection (μL or mL).
Pain reduction	Evidence regarding the reduction pain or discomfort during collection.
Usability, ergonomics, ease	Use, comfort,interaction between user and device. Safety and aspects related to the risk of infection accidents, and to patient and professional protection.
Technological maturity	Technology development stage.
Limitations	Main technical, clinical, regulatory implementation restrictions identified in the studies.

Source: Author (2026)

A predominance of studies focused on the development of microneedle-based devices was observed, highlighting the technology's potential to reduce pain during collection, improve the patient experience, and enable integration with microfluidic platforms and diagnostic systems.

Studies related to capillary sampling demonstrated advances in obtaining samples with lower blood volumes, thereby reducing patient discomfort and facilitating applications in home monitoring, pediatrics, and chronic disease management. Research on blood microsampling highlighted the reduction in sample volume, simplified sample transport logistics, and the potential for remote monitoring, all of which contribute to the decentralization of laboratory services.

For the Quantitative data for the PRISMA Flowchart:

TABLE II
Quantitative data from the PRISMA flowchart.

PRISMA stage	n	Description
Identification	7.130	Records retrieved from the databases.
Duplicates removed	2.150	Duplicate records removed before screening.
Screening	4.980	Records screened by title and abstract.
Records excluded	4.100	Records excluded after title and abstract screening.
Eligibility assessment	880	Full-text articles assessed for eligibility.
Full-text articles excluded	760	Articles excluded due to ineligible study type ($n = 320$), insufficient data ($n = 260$), or out-of-scope content ($n = 180$).
Studies included	120	Articles included in the final qualitative synthesis.

Source: Author (2026)

The study selection process was conducted in successive stages. Initially, 7,130 records were retrieved from the selected databases, followed by the identification and removal of 2,150 duplicate records, resulting in 4,980 records for screening. Subsequently, the records were screened based on their titles and abstracts according to the predefined inclusion and exclusion criteria, and 4,100 records were excluded at this stage. The remaining 880 articles were assessed for eligibility through full-text reading. At this stage, 760 articles were excluded because they presented an ineligible study type (n = 320), insufficient information for analysis (n = 260), or content outside the scope of the research (n = 180). Finally, 120 studies that met all predefined eligibility criteria were included in the qualitative synthesis. To reduce subjectivity throughout the selection process, the eligibility criteria were established in advance and applied consistently to all retrieved records, considering the relevance of the study to minimally invasive technologies for blood collection and its alignment with the objectives of the research. Product Engineering also contributes to defining requirements for performance, reliability, manufacturability, and regulatory compliance factors

essential for the effective transfer of technologies developed in research settings to clinical practice and the industry.

IV. DISCUSSION

The search and selection process identified studies related to the development and evaluation of technologies for minimally invasive blood collection. After removal of duplicate records, titles and abstracts were screened according to the eligibility criteria. The potentially eligible studies were subsequently assessed in full text, resulting in the final set of studies included in the systematic review.

The characteristics of the included studies are presented in Table II, including the technological category, device characteristics, sample characteristics, blood volume, pain or discomfort outcomes, analytical performance, validation stage, and limitations. The included studies demonstrated continued development of minimally invasive blood collection technologies, particularly in the areas of microneedles, capillary sampling, blood microsampling, microfluidic systems, and biosensor-integrated platforms.

Of the 120 studies included in the systematic review,19 were highlighted for a more specific analysis of the main technological approaches identified. The selection of these studies aimed to demonstrate the diversity of solutions proposed for minimally invasive blood collection, encompassing different collection strategies, operating principles, and device configurations. Table III presents these studies and their respective technologies or approaches, allowing the recurring trends in the analyzed literature to be identified.

The process of identifying, screening, assessing eligibility, and including studies was carried out according to the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020).

Table III shows a slice of the 120 studies included in the systematic review, selected to illustrate the main technologies and approaches identified.

TABLE III
SELECTED STUDIES AND THEIR MAIN CHARACTERISTICS.

No.	Year	Authors	Technology/Approach
1	2017	Lee et al.	3D beveled microneedle
2	2018	Xue et al.	Microneedle-based blood collection
3	2021	Hoang et al.	Microneedles, button actuation and vacuum
4	2021	Ates et al.	Integrated diagnostic platforms
5	2021	Delahaye et al.	Dried blood microsampling devices
6	2023	Ribet et al.	Microneedle patch
7	2023	Van Hileghem et al.	Hollow microneedles and microfluidic patch
8	2023	Noble et al.	Onflow capillary device
9	2023	Hoffman et al.	Minimally invasive capillary sampling
10	2023	Thangavelu et al.	Blood microsampling technologies
11	2023	Verboven et al.	Hollow microneedles and microfluidic patch
12	2023	Hoffman & Taberner	Capillary flow strategies
13	2024	Poland & Cobbaert	Capillary self-collection
14	2024	Bossi et al.	Microsampling technologies
15	2024	Poland & Cobbaert	Capillary self-collection devices
16	2025	Temme-Lattin et al.	Needle-free capillary collection
17	2026	Tuttle et al.	Patient-centric microsampling
18	2026	Duschean et al.	Tasso+ SST microsampling
19	2026	Getreu et al.	Capillary blood collection

The analysis of the highlighted studies reveals the presence of different technological strategies aimed at reducing invasiveness, discomfort, and dependence on conventional venipuncture. Among the main patterns identified are microneedles, microsampling technologies, capillary blood collection devices, and microfluidic-based solutions.

Microneedles constitute one of the most recurrent approaches, being used to penetrate the skin superficially and enable the collection of small amounts of blood with reduced invasiveness. The studies analyzed present different configurations, including solid and hollow microneedles, as well as systems associated with patches and collection mechanisms.

Another identified pattern corresponds to microsampling technologies, which use small blood volumes and may facilitate simpler and more accessible collection strategies suitable for different use contexts. This approach is also associated with the possibility of performing blood collection outside conventional laboratory settings.

Capillary blood collection devices represent another relevant technological pathway. These solutions use blood obtained through superficial skin puncture and incorporate mechanisms designed to facilitate blood flow, sample capture, and storage. The strategies identified include self-collection devices, capillary-based systems, and auxiliary mechanisms for sample acquisition.

The integration of microfluidic systems is also noteworthy, particularly in solutions in which blood collection is associated with sample transport, processing, or analysis within miniaturized structures. This integration expands the potential of these devices by bringing the collection stage closer to subsequent sample processing and diagnostic steps.

Overall, the identified patterns indicate a development trend toward solutions that go beyond simply replacing the conventional needle. The technologies analyzed present different levels of integration among sample collection, transport, processing, and analysis, reflecting a product development perspective focused on reducing invasiveness, improving usability, and integrating different stages of the process. Thus, the selected studies contribute to understanding the main technological directions in minimally invasive blood collection and provide a basis for analysis from a Product Engineering perspective.

The reported technologies varied in their degree of invasiveness, sample volume, level of technological development, usability, and clinical validation. From a Product Engineering perspective, the studies were analyzed according to functionality, ergonomics, usability, materials, manufacturing, safety, reliability, sustainability, regulatory aspects, and industrialization potential.

Among the technologies identified, devices based on microneedles stand out as one of the main trends, presenting the potential to significantly reduce pain during collection, minimize tissue damage and allow obtaining samples suitable for clinical analysis. Furthermore, the integration of these structures with microfluidic systems and biosensors favors the development of portable platforms for diagnosis and continuous monitoring.

Regarding the analysis from a Product Engineering perspective, the devices identified in this review can be examined in terms of a development process spanning various design stages—from the definition of functional requirements to validation for production and integration into healthcare services. This approach reveals that innovation in minimally invasive devices depends not only on clinical efficacy but also on the integration of factors such as ergonomics, usability, material selection, manufacturing processes, safety, reliability, sustainability, and regulatory compliance.

In this context, microneedles stand out for meeting various design requirements, such as pain reduction, minimized tissue damage, ease of application, and the potential for integration with microfluidic systems and biosensors. From a Product Engineering standpoint, these characteristics reflect a user-centered development process that prioritizes comfort, safety, and functional performance.

Blood technologies demonstrate a significant contribution from Product Engineering by simplifying the collection process and sample transport logistics. These devices are designed to obtain low-volume capillary blood or interstitial fluid with minimal tissue trauma, avoiding deep venipuncture [5], [6]. From this perspective, these solutions are developed to meet functionality, usability, and operational efficiency requirements, reducing costs, favoring the decentralization of laboratory tests, and expanding the feasibility of self-collection [5].

These solutions reduce operational costs, facilitate the decentralization of laboratory testing, and expand the potential for self-collection—features that add value to the product and ease its integration into various care settings.

Thus, it is evident that Product Engineering serves not merely as support for technological development but as a strategic element in translating clinical requirements into innovative solutions, balancing technical performance, user experience, industrial feasibility, and regulatory compliance. This integration strengthens the innovation process and enhances the potential for these technologies to be adopted by healthcare systems.

To compare criteria related to Product Engineering, scores contributing to the classification were assigned in the table.

Note: A = High; M = Medium; L = Low. The qualitative classification was established based on a comparative analysis of the studies included in the Systematic Literature Review and was linked to product engineering criteria, such as functional requirements, ergonomics, usability, material selection, manufacturing processes, safety, reliability, sustainability, regulations, and industrialization potential. The technologies were qualitatively classified according to Product Engineering criteria, considering the evidence reported in the studies included in the systematic review. The following classification scale was adopted:

- **A – High:** the criterion was strongly addressed or demonstrated in the analyzed technology.
- **M – Medium:** the criterion was partially addressed or demonstrated, with moderate evidence.
- **B – Low:** the criterion was weakly addressed or demonstrated, with limited evidence.

Note: Technological categories are defined as follows:
T1: Microneedle-based collection systems;
T2: Minimally invasive capillary collection devices;
T3: Blood microsampling dried blood spot (DBS) technologies;
T4: Wearable biosensors and smart diagnostic platforms.
Scale: A = High (criterion strongly fulfilled/demonstrated); M = Medium (partially fulfilled/moderate evidence); B = Low (weakly addressed/limited evidence).

The criteria considered were functionality, ergonomics, usability, materials, manufacturing, safety, reliability, sustainability, regulatory aspects, and industrialization potential. The classification was based on a comparative analysis of the technological characteristics reported in the selected studies.

TABLE IV
COMPARISON OF TECHNOLOGIES BASED ON PRODUCT ENGINEERING CRITERIA

Criterion	T1	T2	T3	T4
Functionality	A	M	A	A
Ergonomics	A	M	A	A
Usability	A	A	A	M
Materials	A	M	M	A
Manufacturing	B	A	A	B
Safety	A	M	A	A
Reliability	A	A	M	M
Sustainability	M	M	A	M
Regulation	M	A	M	B
Industrialization	M	A	A	M

Note: A = High; M = Medium; B = Low.

Studies on blood microsampling have also shown promising results, particularly in applications requiring small blood volumes, such as remote monitoring, personalized medicine, and chronic disease management. These solutions reduce the need for complex laboratory infrastructure, fostering the decentralization of healthcare services and expanding access to diagnostics.

Microneedles (MNs) emerge as a promising alternative to transform this scenario; they are microscopic structures, typically less than 1 mm in length, capable of penetrating the skin's outermost layer the stratum corneum without reaching nerve endings or deep blood vessels. This allows for virtually painless application, with minimal risk of infection and rapid skin healing after use. The stratum corneum is the outermost layer of the skin, acting as a robust barrier against microorganisms, skin dehydration, and drug permeation [6].

From the perspective of Product Engineering, the development of these devices involves specification, design, manufacturing and validation stages, aiming to guarantee technical performance, usability and safety. Product engineering is responsible for activities ranging from project specification and details of its operations to the generation of the necessary documentation for its production [2].

From a product engineering perspective, it has been found that factors such as ergonomics, ease of use, safety, reliability, and user-centered design directly influence the clinical acceptance of these technologies. Consequently, device development must consider not only technical performance but also aspects related to the patient experience, usability by healthcare professionals, and the feasibility of large-scale manufacturing.

However, despite the advances observed, significant challenges remain regarding clinical validation, standardization of collection methods, industrial scalability, and compliance with national and international regulatory requirements. Overcoming these limitations will be crucial to expanding the incorporation of these technologies into clinical practice and strengthening their application in minimally invasive systems.

V. Final Remarks

This systematic review identified the key technologies employed in minimally invasive blood collection, highlighting the growth of research focused on the development of microneedle-based devices, microsampling systems, capillary collection, and platforms integrated with biosensors. The results demonstrate that these technologies hold significant potential to reduce pain, enhance procedural safety, improve sample quality, and expand the possibilities for remote monitoring and decentralized diagnostics.

From the perspectives of biomedical and product engineering, the integration of technological innovation, usability, and regulatory compliance is a key factor in developing more efficient, user-centered clinical devices.

Although challenges regarding clinical validation, large-scale production, and regulation persist, the evidence indicates a trend toward these technologies becoming established as promising alternatives to conventional blood collection methods. Thus, this study contributes to mapping the state of the art and provides a foundation for future research and the development of innovative solutions for healthcare diagnostic and monitoring systems.

References

[1] Brazilian Society of Clinical Pathology/Laboratory Medicine (SBPC/ML), *Recommendations of the Brazilian Society of Clinical Pathology/Laboratory Medicine for Venous Blood Collection*, 2nd ed. Barueri, SP, Brazil: Minha Editora, 2010.

[2] A.-M. Šimundić *et al.*, "Joint EFLM-COLABIOCLI recommendation for venous blood sampling," *Clin. Chem. Lab. Med.*, vol. 56, no. 12, pp. 2015–2038, Dec. 2018, doi: 10.1515/cclm-2018-0602.

[3] Brazilian Health Regulatory Agency (ANVISA), *Hemovigilance and Hemotherapy Production Data Report in Brazil*. Brasília, DF, Brazil: Ministry of Health, 2025.

[4] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[5] P. Xue, L. Zhang, Z. Xu, J. Yan, Z. Gu, and Y. Kang, "Blood collection using microneedles as a minimally invasive platform for biomedical diagnostics," *J. Pharm. Anal.*, vol. 8, no. 4, pp. 225–237, Aug. 2018, doi: 10.1016/j.jpha.2018.07.001.

[6] M. S. F. Hoffman, J. W. McKeage, J. Xu, B. P. Ruddy, P. M. F. Nielsen, and A. J. Taberner, "Minimally invasive capillary blood sampling methods," *Expert Rev. Med. Devices*, vol. 20, no. 4, pp. 317–331, Apr. 2023, doi: 10.1080/17434440.2023.2170783.

[7] D.-S. Lee, C. G. Li, C. Ihm, and H. Jung, "A three-dimensional beveled microneedle for minimally invasive and painless blood collection," *Sens. Actuators B Chem.*, vol. 245, pp. 543–551, Jun. 2017, doi: 10.1016/j.snb.2017.01.196.

[8] L. Van Hileghem *et al.*, "Innovative fabrication of hollow microneedle arrays enabling blood sampling with a self-powered microfluidic patch," *Micromachines*, vol. 14, no. 3, p. 615, Mar. 2023, doi: 10.3390/mi14030615.

[9] L. D. Noble *et al.*, "Painless capillary blood collection: A rapid evaluation of the Onflow device," *Diagnostics*, vol. 13, no. 10, p. 1754, May 2023, doi: 10.3390/diagnostics13101754.

[10] M. Temme-Lattin *et al.*, "Painless, needle-free capillary blood collection for fertility hormone monitoring," *Fertil. Steril.*, vol. 123, no. 2, pp. 310–319, Feb. 2025.

[11] F. Ribet *et al.*, "Microneedle patch for painless intradermal collection of interstitial fluid enabling multianalyte measurement of small molecules, SARS-CoV-2 antibodies, and protein profiling," *Adv. Healthc. Mater.*, vol. 12, no. 13, p. 2202564, May 2023, doi: 10.1002/adhm.202202564.

[12] L. Delahaye, H. Veenhof, B. C. P. Koch, J.-W. C. Alffenaar, R. Linden, and C. Stove, "Alternative sampling devices to collect dried blood microsamples: State-of-the-art," *Ther. Drug Monit.*, vol. 43, no. 3, pp. 310–321, Jun. 2021, doi: 10.1097/FTD.0000000000000864.

[13] M. U. Thangavelu, B. Wouters, A. Kindt, I. K. M. Reiss, and T. Hankemeier, "Technologies for blood microsampling: Innovations and applications," *Anal. Sci. Adv.*, vol. 4, no. 7–8, pp. 1–15, Aug. 2023, doi: 10.1002/ansa.202300011.

[14] K. Duschean *et al.*, "Potential remote measurement of C-peptide using a minimally invasive and convenient sampling technique," *Bioanalysis*, 2026, doi: 10.1080/17576180.2026.2690265.

[15] International Electrotechnical Commission, *Medical devices – Part 1: Application of usability engineering to medical devices*, IEC Standard 62366-1:2015, 2015.

[16] Brazilian Health Regulatory Agency (ANVISA), *Resolution of the Collegiate Board – RDC No. 751, September 15, 2022: Risk classification, notification and registration regimes for medical devices*. Brasília, DF, Brazil: Official Gazette of the Federal Government, 2022.

Telehealth Equipment Infrastructure in Brazil: Regional Distribution, Administrative Availability, Population-Adjusted Inequalities, and Connectivity

Bryan Kalel Araujo Lima, Jessica Silva Gama, Thaisse Dias Paes,
Bárbara Vitória das Neves Cecim, Fayga Karolyne da Silva de Souza, and Antônio Pereira Júnior
Institute of Technology, Federal University of Pará (UFPA), Belém, Pará, Brazil
ORCID: B.K.A. Lima 0009-0007-5562-6106; J.S. Gama 0009-0008-7110-2086; T.D. Paes 0009-0006-2429-413X;
B.V.N. Cecim 0009-0002-7048-9852; F.K.S. Souza 0009-0000-2863-736; A. Pereira Júnior 0000-0002-0808-1058
E-mail: bryan.lima@itec.ufpa.br; jessica.gama@ics.ufpa.br; thaisse.paes@itec.ufpa.br;
barbara.cecim@itec.ufpa.br; fayga.reis@itec.ufpa.br; apereira@ufpa.br

***Abstract*—This ecological cross-sectional study evaluated the geographic distribution and administrative availability of eight telehealth-related equipment categories registered in the Brazilian National Registry of Health Establishments (CNES). Municipal- and state-level analyses combined population-adjusted indicators, global and local spatial autocorrelation, and municipal sensitivity models including connectivity, household income, and urbanization. Nationwide, 11,745 devices were identified, of which 11,285 were registered as "in use", corresponding to a national administrative in-use ratio (AIR) of 96.08%. Among 5,571 municipal territorial units, 4,138 had no equipment registered as in use. Municipal Global Moran's I for equipment registered as in use per 100,000 inhabitants was 0.193 (permutation $p = 0.0001$). At nominal $p < 0.05$, Local Moran analysis classified 352 municipalities as High–High, 51 as Low–High, and 42 as High–Low, with no Low–Low clusters. No municipality met the Benjamini–Hochberg or Bonferroni multiplicity criteria, whereas a permutation-based maxT familywise-error analysis retained four High–High local associations in Mato Grosso do Sul. In the fully adjusted municipal negative-binomial model, the IBC estimate included the null (IRR=1.039, 95% CI 0.957–1.127, $p = 0.361$). These findings support municipal-scale spatial structure but more limited municipality-specific localization, and indicate that the connectivity association is sensitive to socioeconomic adjustment. The CNES "in use" field was interpreted as administrative availability rather than direct evidence of clinical utilization, productivity, or uptime.**

***Index Terms*—Telehealth, CNES, Spatial Analysis, Clinical Engineering, Health Inequalities.**

I. Introduction

Telehealth has become an important strategy for mitigating geographical barriers and expanding access to specialized healthcare [1], [2]. Its effectiveness, however, depends on a chain of mutually dependent resources, including suitable medical devices, reliable connectivity, interoperable information systems, trained personnel, maintenance capacity, and integration with clinical workflows [2]–[4].

In Brazil, the Telessaúde Brasil Redes Program and other digital-health initiatives have contributed to the institutionalization of telehealth within the Unified Health System (SUS) [5]–[8]. Nevertheless, large territorial distances and persistent socioeconomic inequalities create substantial heterogeneity in technological capacity, with previous studies reporting regional concentration of telemedicine services and barriers related to connectivity, professional training, and organizational readiness [9], [10].

Absolute equipment counts alone are insufficient for evaluating equitable access. Highly populated territories are expected to contain larger equipment stocks even when population-adjusted availability is comparatively low. Likewise, the proportion of equipment registered as in use describes the administrative status of installed stock but does not measure clinical service use. These dimensions should therefore be analyzed separately.

Accordingly, this study characterizes the distribution of telehealth-related equipment at municipal and state levels, evaluates spatial autocorrelation in population-adjusted administrative availability, and examines the robustness of the association between connectivity and equipment availability after adjustment for selected municipal socioeconomic characteristics.

II. Methodology

A. Study Design and Data Sources

A quantitative, ecological, and cross-sectional design was adopted. Equipment data were extracted from the CNES physical-resources database maintained by DataSUS [11]. The SCNES filling guide defines "in use" as the number of registered units effectively being used at the establishment and instructs that equipment under maintenance be excluded from this quantity [34]. Because this is a registry field rather than a direct measurement of device activity, we use it as an indicator of administrative availability. The analytical dataset comprised eight telehealth-related equipment categories: facial-recognition cameras, dermatoscopes, portable fetal detectors, medical audiological diagnostic kits, digitizing tablets, portable multiparameter telehealth monitors, portable retinographs, and portable ultrasound systems. The equipment reference period was January 2026.

Population estimates with a reference date of July 1, 2025 were obtained from the Brazilian Institute of Geography

and Statistics (IBGE) [12]. Municipal household income was obtained from the 2022 Demographic Census income results (SIDRA Table 10295), while urban and rural population counts were obtained from the 2022 Census municipal tables and used to calculate the urbanization percentage [30], [31]. Connectivity was represented by the 2025 Brazilian Connectivity Index (IBC), developed by Anatel [13]. The 2025 IBGE Municipal Digital Mesh was used to define municipal and state boundaries for the spatial analysis [32].

The descriptive and spatial analysis comprised 5,571 municipal territorial units. The socioeconomic sensitivity models included 5,570 units with complete IBC, income, urbanization, population, and equipment information. Boa Esperança do Norte (MT), created after the 2022 Census reference structure, was present in the 2025 population/equipment database but had no 2022 socioeconomic covariates and was therefore excluded only from the adjusted municipal models. Complementary analyses were aggregated to the 27 Brazilian federative units.

B. Outcome Definitions

For each territorial unit, the existing quantity ($Q_{existing}$) and the quantity registered as in use (Q_{use}) were summed across the equipment set. The administrative in-use ratio (AIR) was defined as

$$AIR = \frac{Q_{use}}{Q_{existing}} \times 100\%, \tag{1}$$

provided that $Q_{existing} > 0$. Population-adjusted availability was defined as

$$D_{100k} = \frac{Q_{use}}{P} \times 100{,}000, \tag{2}$$

where P is the 2025 population estimate.

The CNES "in use" field was treated as an administrative availability indicator. Although CNES instructs establishments to classify as "in use" only equipment effectively being used and to exclude units under maintenance [34], the registry does not provide direct measurements of uptime, examination or consultation volume, clinical productivity, intensity of use, or integration into routine care. AIR and D_{100k} therefore characterize registered infrastructure rather than realized clinical service production.

C. Statistical and Spatial Analysis

Descriptive statistics were calculated for all indicators. At the municipal level, Spearman rank correlations were used to examine the bivariate associations of population-adjusted equipment availability with IBC, household income, urbanization, and population. Two-sided p-values were calculated and the Benjamini–Hochberg procedure was used to control the false discovery rate across these municipal correlation tests [14]. Because of the large municipal sample, interpretation emphasized both the magnitude of the correlation coefficient and statistical significance. State-level Spearman correlations were retained as a scale-comparison analysis and were likewise reported with Benjamini–Hochberg-adjusted q-values.

Global Moran's I was calculated for equipment registered as in use per 100,000 inhabitants using row-standardized Queen contiguity and 9,999 random permutations [28], [29]. Local Moran's I classified nominal High–High, Low–Low, Low–High, and High–Low patterns at $p < 0.05$ [29]. Two island units had no Queen neighbors, leaving 5,569 local tests. Municipality-specific inference was then re-evaluated using Benjamini–Hochberg FDR ($q < 0.05$) [14], Bonferroni ($\alpha = 0.05/5569$), and permutation maxT FWER control [35]. The upper-tail maxT distribution retained the maximum positive Local Moran statistic across municipalities within each common permutation and therefore targeted positive High–High/Low–Low cluster-type association; negative spatial outliers remained exploratory. The same Global Moran framework was applied at state level. Municipal boundaries came from the 2025 IBGE Digital Mesh [32]. A simulation sensitivity analysis used contiguous clusters of 10, 25, 50, or 100 municipalities, signals of 1–5 observed SD, and approximately 2,000 replicates per scenario to compare Global Moran detection with maxT localization and FWER.

D. Municipal Socioeconomic Modeling

The municipal adjusted analysis used the equipment count registered as "in use" as the dependent variable rather than the rate itself. The count exhibited marked overdispersion (mean $= 2.03$, variance $= 153.64$; variance-to-mean ratio $= 75.83$); therefore, negative-binomial regression with NB2 variance specification was selected instead of a Poisson model [33]. The 2025 municipal population was included as the exposure term, which is equivalent to including $\ln(\text{population})$ as an offset with its coefficient fixed at one.

Four nested sensitivity models were estimated: Model 1 included IBC alone; Model 2 included IBC and per-capita household income; Model 3 included IBC and urbanization; and Model 4 simultaneously included IBC, income, and urbanization. For interpretability, IBC was scaled per 10-point increase, income per R$1,000 increase, and urbanization per 10-percentage-point increase. Results are reported as incidence-rate ratios (IRRs), 95% confidence intervals, and two-sided p-values. In the fully adjusted model, variance-inflation factors were 2.02 for IBC, 1.56 for income, and 1.49 for urbanization, providing no indication of problematic multicollinearity. Because the municipal outcome showed significant spatial autocorrelation, these models were treated as sensitivity analyses. No spatial count regression was fitted; therefore, the regression coefficients were not interpreted as causal or spatially independent effects.

III. Results

A. National and Regional Distribution

Across Brazil, 11,745 telehealth-related equipment items were identified, of which 11,285 were registered as "in use" in CNES. The national administrative AIR was 96.08%, and the population-adjusted availability was 5.29 equipment items registered as in use per 100,000 inhabitants. Among the 5,571

municipal-level territorial units included in the spatial analysis, 4,138 (74.28%) had no equipment registered as in use.

Regional variation was observed in absolute stock, administrative AIR, and population-adjusted availability. The Northeast had 6,601 existing equipment items and 6,404 registered as in use, corresponding to 11.19 per 100,000 inhabitants. The Southeast had 2,517 existing items and 2,423 registered as in use, corresponding to 2.73 per 100,000 inhabitants, while the North had 561 existing items and 550 registered as in use, corresponding to 2.93 per 100,000 inhabitants. Administrative AIR ranged from 85.85% in the Central-West to 98.04% in the North, whereas population-adjusted availability ranged from 2.73 to 11.19 per 100,000 inhabitants. The two indicators are reported separately because they represent distinct features of the registered equipment infrastructure. Regional values are summarized in Table I.

TABLE I
DISTRIBUTION OF TELEHEALTH-RELATED EQUIPMENT BY BRAZILIAN MACRO-REGION.

Region	Existing	In use	Admin. AIR (%)	In use/100k
North	561	550	98.04	2.93
Northeast	6,601	6,404	97.02	11.19
Central-West	763	655	85.85	3.80
Southeast	2,517	2,423	96.27	2.73
South	1,303	1,253	96.16	4.00
Brazil	11,745	11,285	96.08	5.29

B. Equipment-Specific Distribution

The composition of the existing stock also varied across equipment categories. Facial-recognition cameras accounted for 8,132 units (69.24% of the national stock), followed by portable fetal detectors with 1,582 units (13.47%). Dermatoscopes accounted for 491 units (4.18%), digitizing tablets for 402 (3.42%), portable multiparameter telehealth monitors for 327 (2.78%), portable ultrasound systems for 295 (2.51%), portable retinographs for 277 (2.36%), and medical audiological diagnostic kits for 239 (2.03%). Table II presents the distribution across all eight categories.

TABLE II
EXISTING TELEHEALTH-RELATED EQUIPMENT STOCK BY EQUIPMENT CATEGORY.

Equipment category	Existing stock	Share (%)
Facial-recognition camera	8,132	69.24
Portable fetal detector	1,582	13.47
Dermatoscope	491	4.18
Digitizing tablet	402	3.42
Portable multiparameter telehealth monitor	327	2.78
Portable ultrasound system	295	2.51
Portable retinograph	277	2.36
Medical audiological diagnostic kit	239	2.03
Total	11,745	100.00

C. Municipal Spatial Autocorrelation

For equipment registered as in use per 100,000 inhabitants, municipal Global Moran's I was 0.193 (permutation $p = 0.0001$; Queen contiguity; 9,999 permutations), providing evidence of positive global spatial autocorrelation. Among the 5,569 non-isolated municipalities, nominal Local Moran inference identified 445 local associations: 352 High–High, 51 Low–High, and 42 High–Low; no Low–Low municipality met the nominal $p < 0.05$ criterion. The two Queen isolates, Fernando de Noronha (PE) and Ilhabela (SP), were not tested locally (Table III).

Multiplicity correction substantially restricted municipality-specific inference. No municipality satisfied BH-FDR at $q < 0.05$ (minimum $q = 0.0796$), and none satisfied Bonferroni correction. The Bonferroni threshold was 8.98×10^{-6}, below the minimum attainable permutation p-value of 0.0001 with 9,999 permutations; Bonferroni non-rejection therefore cannot be interpreted as evidence that local structure is absent. In contrast, the permutation maxT FWER procedure retained four High–High local associations, all in Mato Grosso do Sul: Pedro Gomes ($p_{FWER} = 0.0047$), Taquarussu ($p_{FWER} = 0.0141$), Jateí ($p_{FWER} = 0.0173$), and Alcinópolis ($p_{FWER} = 0.0400$).

TABLE III
MUNICIPAL LOCAL MORAN CLASSIFICATIONS BEFORE AND AFTER MULTIPLICITY CONTROL.

LISA class	Nominal	BH-FDR	Bonf.	maxT FWER
High–High	352	0	0	4
Low–Low	0	0	0	0
Low–High	51	0	0	0
High–Low	42	0	0	0
No local rejection	5,124	5,569	5,569	5,565
Not tested (isolate)	2	2	2	2
Total	5,571	5,571	5,571	5,571

Simulation confirmed substantially greater global than local detection power: a 25-municipality, 3-SD cluster yielded 78.2% Global Moran power but 2.3% probability of at least one true maxT localization. Global power reached approximately 100% in stronger scenarios, whereas maximum maxT localization power was 52.9%; the mean probability of any outside-cluster maxT detection was 4.84%, consistent with 5% FWER.

D. State-Level Scale Analysis

After aggregation to the 27 federative units, Global Moran's I was 0.166 with $pseudo\text{-}p = 0.0632$, which did not meet the prespecified 5% significance threshold. At this scale, IBC was positively correlated with equipment registered as in use per 100,000 inhabitants ($\rho = 0.589$, BH-adjusted $q = 0.0197$). This state-level association is reported as a scale-specific descriptive result and is interpreted alongside the municipal analysis below.

E. Municipal Associations and Socioeconomic Sensitivity Analysis

The municipal bivariate correlations with equipment registered as in use per 100,000 inhabitants were small in magnitude. The Spearman coefficients were $\rho = -0.060$ for IBC

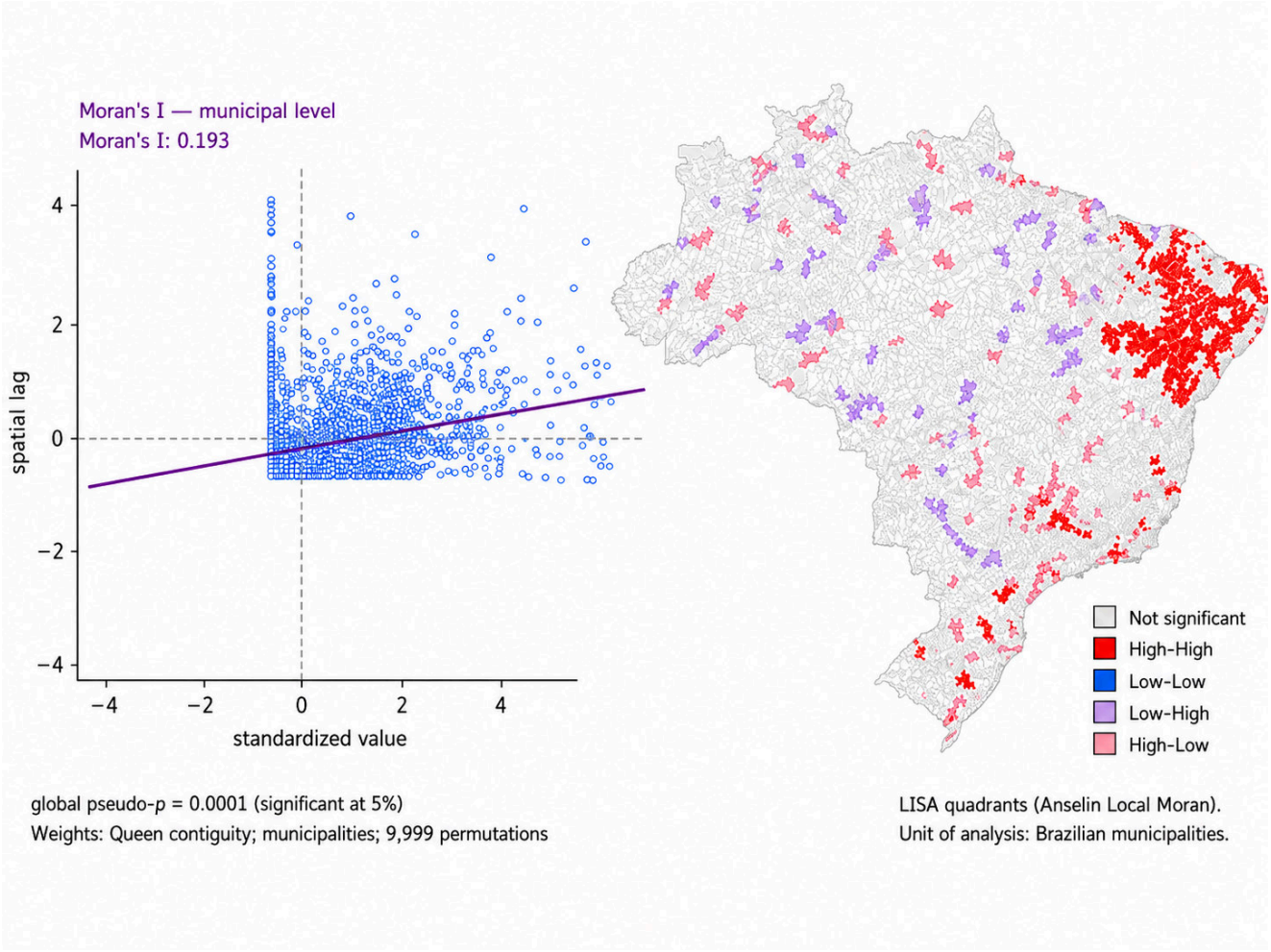


Fig. 1. Municipal spatial analysis of telehealth-related equipment registered as in use per 100,000 inhabitants. The left panel shows the Moran scatterplot and the right panel shows nominal Local Moran (LISA) quadrants at $p < 0.05$. Global Moran's $I = 0.193$ (permutation $p = 0.0001$), based on Queen contiguity and 9,999 permutations. Multiplicity-corrected local results are reported in Table III.

TABLE IV
COMPARISON OF GLOBAL SPATIAL AUTOCORRELATION AT MUNICIPAL AND STATE LEVELS.

Spatial scale	**N**	**Moran's I**	*pseudo-p*
Municipal	5,571	0.193	0.0001
State	27	0.166	0.0632

($q < 0.001$), $\rho = -0.138$ for household income ($q < 0.001$), $\rho = -0.060$ for urbanization ($q < 0.001$), and $\rho = 0.029$ for population 2025 ($q = 0.033$), after Benjamini–Hochberg adjustment across the four municipal correlation tests.

In the negative-binomial models, the IBC estimate differed across model specifications. In Model 1, the IRR for IBC was 0.806 (95% CI 0.759–0.856; $p < 0.001$). With household income included (Model 2), the IBC IRR was 0.943 (95% CI 0.876–1.016; $p = 0.122$). With urbanization included (Model 3), the IBC IRR was 0.886 (95% CI 0.826–0.950; $p < 0.001$). In the fully adjusted Model 4, the IBC IRR was 1.039 (95% CI 0.957–1.127; $p = 0.361$), while the IRRs for household income and urbanization were 0.506 (95% CI 0.420–0.611; $p < 0.001$) and 0.877 (95% CI 0.835–0.920; $p < 0.001$), respectively. AIC values were 13,960.7, 13,915.9, 13,935.5, and 13,889.3 for Models 1–4, respectively.

IV. DISCUSSION

The principal spatial finding is municipal-scale geographic structure rather than the nominal count of LISA locations. Global Moran's I was significant, whereas municipality-specific localization was far more restrictive: 445 local associations appeared nominally, but only four High–High municipalities in Mato Grosso do Sul survived maxT FWER control. These are best interpreted as high-confidence local association cores under the specified Queen weights, not as independent clusters or complete cluster boundaries. The simulations showed the same global/local asymmetry; therefore, failure to survive a local multiplicity criterion is not evidence that local spatial structure is absent. The BH and maxT rejection sets need not be nested because BH was applied to discrete local

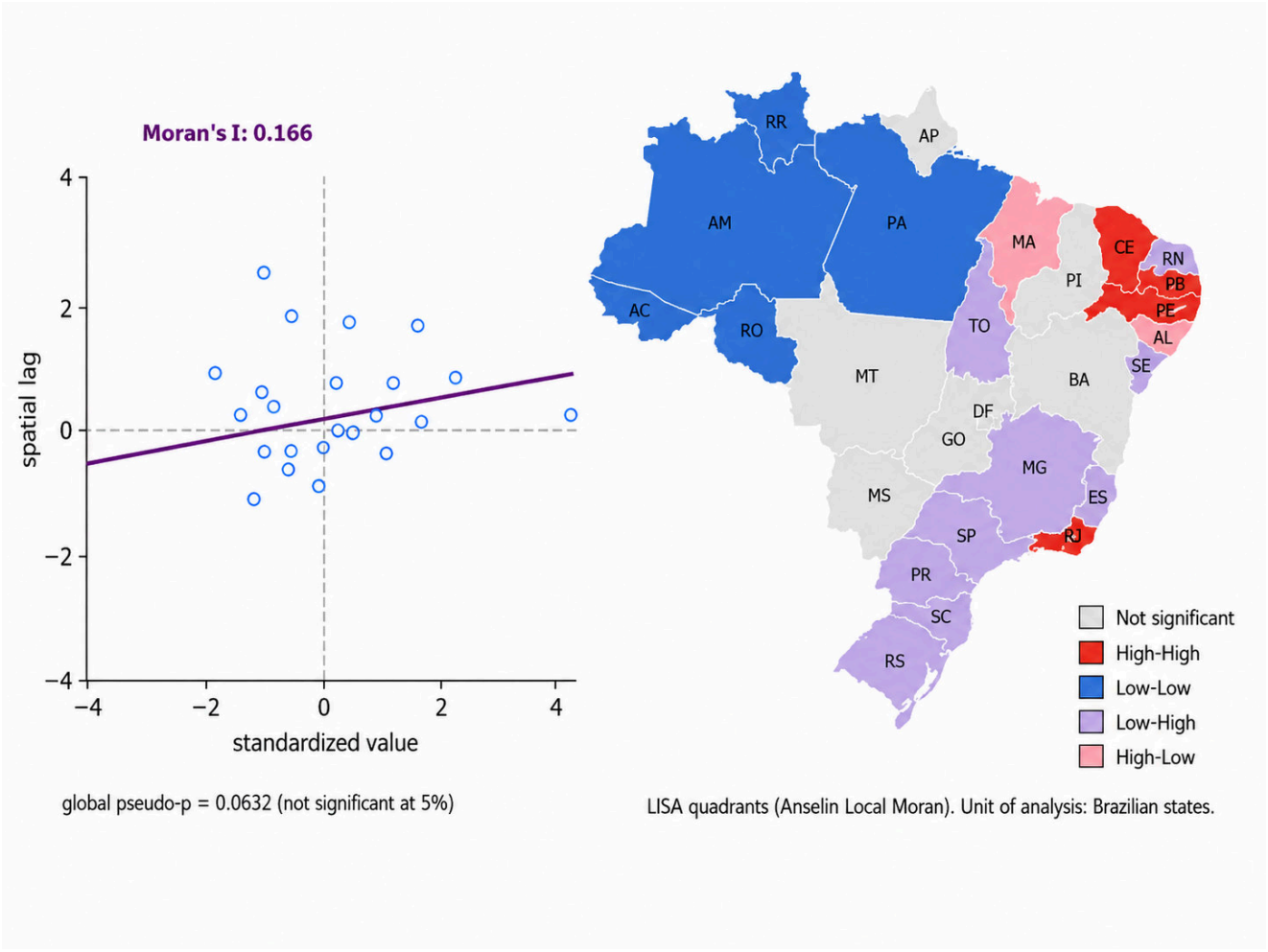


Fig. 2. State-level spatial analysis of telehealth-related equipment registered as in use per 100,000 inhabitants. Global Moran's $I = 0.166$ and $pseudo\text{-}p = 0.0632$. Because the state-level global test did not meet the 5% significance threshold, local state patterns are shown for descriptive and exploratory comparison only.

TABLE V
MUNICIPAL NEGATIVE-BINOMIAL SENSITIVITY MODELS FOR EQUIPMENT REGISTERED AS IN USE.

Model	**Predictor**	**IRR**	**95% CI**	***p*-value**
M1	IBC (per 10 points)	0.806	0.759–0.856	< 0.001
M2	IBC (per 10 points)	0.943	0.876–1.016	0.122
	Income (per R$1000)	0.507	0.419–0.613	< 0.001
M3	IBC (per 10 points)	0.886	0.826–0.950	< 0.001
	Urbanization (per 10 pp)	0.879	0.837–0.923	< 0.001
M4	IBC (per 10 points)	1.039	0.957–1.127	0.361
	Income (per R$1000)	0.506	0.420–0.611	< 0.001
	Urbanization (per 10 pp)	0.877	0.835–0.920	< 0.001

Outcome: count of equipment registered as "in use" in CNES. The 2025 municipal population was included as the exposure term. M1: IBC only; M2: IBC + income; M3: IBC + urbanization; M4: IBC + income + urbanization. The model sample contained 5,570 municipal units with complete covariate data. The models did not explicitly model spatial dependence and are therefore reported as sensitivity analyses rather than causal estimates.

permutation p-values, whereas maxT used the joint maximum-statistic permutation distribution.

The connectivity association was likewise sensitive to scale and adjustment. The positive state-level IBC correlation ($\rho = 0.589$, $q = 0.0197$) was not reproduced municipally, and the IBC coefficient included the null after adjustment for household income and urbanization. Connectivity was therefore not an independent determinant of equipment allocation in this municipal sample and is better interpreted as one dimension of a broader territorial context.

Municipal disaggregation also changes the interpretation of regional inequality. State averages can mask heterogeneity, particularly in extensive Northern states. Municipal analysis reduces this aggregation problem, although municipality-wide averages still conceal urban, rural, remote, and riverine within-municipality differences.

Equipment stocks may reflect historical investment, regional development, health-service density, management capacity, workforce, and private-sector participation [17], [22], [26], [27]. Income and urbanization captured only part of this context, and their inverse adjusted coefficients are treated as ecological associations, not evidence that either factor reduces equipment availability.

Training, organizational readiness, maintenance, consumables, and clinical-engineering support are plausible determinants described in the literature [16], [18], [20], [23], [24], but none was measured here and they are therefore hypotheses rather than study results. Likewise, CNES defines "in use" as equipment reported to be effectively used and not under maintenance [34], but the field does not quantify continuous uptime, clinical throughput, productivity, or realized patient use.

V. Limitations

CNES is an administrative registry; "in use" does not directly measure uptime, throughput, productivity, service use, or quality. The ecological, cross-sectional design does not support causality, and municipal averages may still mask intramunicipal inequality, especially in large Northern municipalities. Residual confounding by historical investment, health-system capacity, workforce, private-sector participation, maintenance, and clinical engineering remains possible. Local inference depends on the Queen weights; with 9,999 permutations the pseudo-p grid could not attain the Bonferroni threshold, and the simulations showed limited maxT localization power. Thus, absence of a corrected local rejection does not establish absence of local structure. The negative-binomial models did not explicitly model spatial dependence, and the municipal–state comparison is not a formal decomposition of the modifiable areal unit problem.

VI. Conclusion

Telehealth-related equipment registered as in use in CNES showed substantial municipal heterogeneity and significant positive global spatial autocorrelation. Formal local inference was more selective: four High–High municipalities in Mato Grosso do Sul remained significant under maxT FWER control, while the broader nominal LISA pattern should be interpreted exploratorily. The IBC association was strongly dependent on geographic scale and socioeconomic adjustment and included the null in the fully adjusted municipal model. The findings therefore support a spatially structured but multifactorial interpretation of equipment allocation, not a connectivity-only mechanism. Future studies should combine longitudinal infrastructure data with direct measures of service production, historical investment, health-system capacity, and spatial regression to clarify the mechanisms underlying these geographic inequalities.

Conflict of Interest

The authors declare that they have no competing financial or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. O. Alenoghena et al., "Telemedicine: A Survey of Telecommunication Technologies, Developments, and Challenges," *J. Sens. Actuator Netw.*, vol. 12, no. 2, p. 20, 2023.

[2] F. Saigí-Rubio et al., "The Current Status of Telemedicine Technology Use Across the World Health Organization European Region: An Overview of Systematic Reviews," *J. Med. Internet Res.*, vol. 24, no. 10, p. e40877, 2022.

[3] A. Garavand et al., "Acceptance of telemedicine technology among physicians: A systematic review," *Informatics in Medicine Unlocked*, vol. 30, p. 100943, 2022.

[4] "Challenges of Implementing Telemedicine Technology: A Systematized Review," 2022.

[5] K. O. Lisboa et al., "A história da telemedicina no Brasil: desafios e vantagens," 2023.

[6] R. C. de Carvalho and R. C. O. S. Castro, "A telemedicina no processo de democratização da saúde no Brasil: uma revisão integrativa," *Brazilian Journal of Implantology and Health Sciences*, vol. 6, 2024.

[7] I. S. de Lima and E. C. de Oliveira, "Avanço da telemedicina no Brasil no período de pandemia da COVID-19: uma revisão sistemática da literatura," *Brazilian Journal of Health Review*, vol. 5, no. 3, pp. 10505–10525, 2022.

[8] D. de C. Bueno and E. Tavares, "Cartografia das controvérsias da telemedicina no Brasil," *Cadernos EBAPE.BR*, vol. 21, 2023.

[9] G. F. S. Silva et al., "Distribuição geográfica da telemedicina no Brasil a partir da produção ambulatorial," *Research, Society and Development*, vol. 11, no. 15, p. e18111535982, 2022.

[10] A. A. Rocha and A. M. M. Lopes, "Desafios sociais da telemedicina no Brasil: uma revisão de literatura," *Fisioter. Bras.*, vol. 26, no. 2, pp. 2147–2153, 2025.

[11] Brazilian Ministry of Health, Department of Informatics of the Unified Health System, "CNES–Physical Resources: Equipment," 2026. [Online]. Available: https://datasus.saude.gov.br/cnes-recursos-fisicos/. Accessed: Jul. 17, 2026.

[12] Brazilian Institute of Geography and Statistics, "Estimates of Resident Population for Municipalities and Federation Units: Reference Date July 1, 2025," 2025. [Online]. Available: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html. Accessed: Jul. 17, 2026.

[13] Brazilian National Telecommunications Agency, "Brazilian Connectivity Index (IBC)," 2025. [Online]. Available: https://www.gov.br/anatel/pt-br/regulado/universalizacao/indice-brasileiro-conectiviadade-ibc. Accessed: Jul. 17, 2026.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[15] N. M. P. Arruda, A. G. Maia, and L. C. Alves, "Inequality in access to health services between urban and rural areas in Brazil: A disaggregation of factors from 1998 to 2008," *Cadernos de Saúde Pública*, vol. 34, no. 6, p. e00213816, 2018.

[16] A. C. F. Bernardes, L. C. Coimbra, and H. O. Serra, "Use of telehealth as a tool to support continuing health education," *Pan American Journal of Public Health*, vol. 42, p. e134, 2018.

[17] A. Bigoni, A. R. Cunha, and J. L. F. Antunes, "Provision of health resources in the intermediate regions of Brazil, 2018," *Cadernos Saúde Coletiva*, vol. 31, no. 4, p. e31040468, 2023.

[18] C. M. Canabrava, A. A. Nunes, and A. Pazin-Filho, "The importance of the medical imaging technology park: Access, equity, and balance," *Radiologia Brasileira*, vol. 57, p. e2, 2024.

[19] A. da Cruz, J. M. S. de V. Maldonado, and C. A. G. Gadelha, "Telemonitoring and business dynamics in health: Challenges and opportunities for SUS," *Revista de Saúde Pública*, vol. 54, p. 65, 2020.

[20] R. F. Damasceno and A. P. Caldeira, "Factors associated with the non-use of telehealth consultancy by physicians of the Family Health Strategy," *Ciência & Saúde Coletiva*, vol. 24, no. 8, pp. 3089–3098, 2019.

[21] M. N. P. Dantas, D. L. B. de Souza, A. M. G. de Souza, K. M. Aiquoc, T. A. de Souza, and I. R. Barbosa, "Factors associated with poor access to health services in Brazil," *Revista Brasileira de Epidemiologia*, vol. 24, p. e210004, 2021.

[22] C. S. R. Machado and A. C. C. Lima, "Spatial distribution of Brazilian public health system and the determinants of public health expenditure," *Revista Econômica do Nordeste*, vol. 52, no. 4, pp. 121–145, 2021.

[23] M. S. Nascimento, "Analysis of the underutilization of biomedical equipment in a University Hospital (UH): Causes, impacts and optimization proposals," B.S. thesis, Dept. Biomed. Eng., Federal University of Rio Grande do Norte, Natal, Brazil, 2025.

[24] R. C. S. R. Pacheco, D. R. Boente, and F. A. Bezerra, "Depreciation and maintenance of high complexity equipment and the impact on the efficiency of federal university hospitals," *Administração Pública e Gestão Social*, vol. 17, no. 1, 2025.

[25] N. C. Palmeira, J. P. Moro, F. A. Getulino, Y. P. Vieira, A. O. Soares Junior, and M. O. Saes, "Analysis of access to health services in Brazil according to sociodemographic profile: National Health Survey, 2019," *Epidemiologia e Serviços de Saúde*, vol. 31, no. 3, p. e2022966, 2022.

[26] J. M. B. Santos, "Analysis of equipment distribution in Brazil according to data from the National Registry of Health Establishments (CNES)," B.S. thesis, Dept. Biomed. Eng., Federal University of Uberlândia, Uberlândia, Brazil, 2021.

[27] D. B. Tomasiello, J. Bazzo, J. Parga, L. M. Servo, and R. H. M. Pereira, "Racial and income inequalities in access to healthcare in Brazilian cities," Discussion Paper No. 2832, Institute for Applied Economic Research (Ipea), Brasília, Brazil, 2023.

[28] P. A. P. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, nos. 1–2, pp. 17–23, 1950.

[29] L. Anselin, "Local indicators of spatial association—LISA," *Geographical Analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[30] Brazilian Institute of Geography and Statistics (IBGE), "Censo Demográfico 2022: População residente por situação do domicílio – Municípios – 2022," 2024. [Online]. Available: https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html. Accessed: Aug. 28, 2026.

[31] Brazilian Institute of Geography and Statistics (IBGE), "Censo Demográfico 2022: Trabalho e Rendimento – valor do rendimento nominal médio mensal domiciliar per capita, SIDRA Table 10295," 2025.

[32] Brazilian Institute of Geography and Statistics (IBGE), "2025 Municipal Digital Mesh (Malha Municipal Digital)," 2025. [Online]. Available: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html. Accessed: Aug. 28, 2026.

[33] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2013.

[34] Brazilian Ministry of Health, "SCNES – Guia de Preenchimento: Equipamentos," Cadastro Nacional de Estabelecimentos de Saúde. [Online]. Available: https://wiki.saude.gov.br/cnes/index.php/SCNES_-_Guia_de_Preenchimento. Accessed: Sep. 2, 2026.

[35] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York, NY, USA: Wiley, 1993.

Terapia Fototérmica Assistida por Nanopartículas Plasmônicas na Região do Infravermelho: Investigação Numérica e Experimental

1ª Mariana J.B.S. Crispim
PPG em Engenharia Elétrica
Universidade Federal de Pernambuco
Recife, Brasil
mariana.crispim@ufpe.br

2ª Gabrielli M. F. de Oliveira
Laboratório de Óptica Biomédica e Imagens
Universidade Federal de Pernambuco
Recife, Brasil
gabrielli.oliveira@ufpe.br

3º Túlio de L. Pedrosa
Universidade Federal Rural de PE
Cabo de Santo Agostinho, Brazil
pedrosatulio@gmail.com

4ª Marcilene S. da Silva
Laboratório de Farmacologia e Cancerologia Experimental
Universidade Federal de Pernambuco
Recife, Brasil
marcilene.souzasilva@ufpe.br

5º Renato E. de Araujo
Laboratório de Óptica Biomédica e Imagens
Universidade Federal de Pernambuco
Recife, Brasil
renato.earaujo@ufpe.br

***Resumo*—A terapia fototérmica (PTT) assistida por nanopartículas plasmônicas constitui uma estratégia promissora para o tratamento localizado de tumores. Neste trabalho é apresentada uma metodologia integrada que combina simulações eletromagnéticas, modelagem térmica, estimativa de dano tecidual e validação experimental para o planejamento racional da PTT. Nanobastões e nanocascas de ouro foram avaliados por meio do método de elementos finitos visando identificar nanoaquecedores eficientes na região do infravermelho próximo. Os resultados mostram que a eficiência fototérmica depende não apenas da seção de choque de absorção, mas também do volume da nanopartícula, sendo o Número de Joule uma métrica adequada para sua seleção. A distribuição de temperatura em tecido biológico foi obtida utilizando a equação de bioaquecimento de Pennes, enquanto o dano térmico irreversível foi estimado pelo modelo de Arrhenius. As predições numéricas foram validadas experimentalmente em um modelo *in vivo* de sarcoma 180 implantado em camundongos, utilizando nanobastões de ouro irradiados com laser de 808 nm. A metodologia proposta fornece uma abordagem abrangente para otimização de terapias fototérmicas mediadas por nanoaquecedores plasmônicos.**

***Index Terms*—Óptica biomédica; Nanomateriais; Terapia óptica; Modelagem computacional.**

I. Introdução

A terapia fototérmica (PTT) tem se consolidado como uma abordagem minimamente invasiva para o tratamento localizado de tumores, explorando a conversão da energia luminosa em calor para induzir dano térmico seletivo em tecidos malignos [1], [2]. Essa estratégia apresenta elevada seletividade espacial, baixa toxicidade sistêmica e potencial para combinação com outras modalidades terapêuticas, como quimioterapia, radioterapia e imunoterapia [2]. A eficiência da PTT depende da interação entre o agente fototérmico, as propriedades ópticas do tecido biológico e os parâmetros de irradiação, como comprimento de onda, irradiância e tempo de exposição [3].

Nesse contexto, nanopartículas metálicas, particularmente as de ouro, têm despertado grande interesse como agente fototérmico devido à elevada estabilidade química, biocompatibilidade, facilidade de funcionalização superficial e intensa absorção óptica associada à ressonância localizada de plasmons de superfície (LSPR) [4], [5]. Além disso, a geometria dessas nanopartículas permite ajustar a posição da ressonância plasmônica para a janela biológica do infravermelho próximo (650–950 nm), região espectral em que os tecidos apresentam menor absorção e espalhamento óptico, favorecendo uma maior profundidade de penetração da luz [6]. Os primeiros estudos utilizando nanocascas de ouro demonstraram a viabilidade da destruição térmica de tumores sob irradiação no infravermelho próximo [1], [7]. Trabalhos posteriores consolidaram os nanobastões de ouro como agentes fototérmicos promissores para aplicações em PTT [8]. Mais recentemente, nanoaquecedores plasmônicos têm sido empregados em plataformas teranósticas multifuncionais e funcionalizados com ligantes específicos para aumentar a seletividade tumoral [9]–[11].

Em aplicações fototérmicas, a seleção de nanoaquecedores normalmente é baseada na magnitude da seção de choque de absorção (σ_{abs}). Entretanto, elevados valores de σ_{abs} não implicam necessariamente na maior eficiência de aquecimento, uma vez que a geração e a dissipação de calor também dependem do volume e da geometria da nanopartícula. Assim, foi demostrado que o Número de Joule (Jo), definido como a razão entre a seção de choque de absorção e o volume da nanopartícula, constitui uma métrica mais apropriada para avaliar a eficiência fototérmica de nanoestruturas plasmônicas [12].

A resposta óptica de nanopartículas com geometrias arbitrárias pode ser determinada por simulações eletro-

magnéticas baseadas no Método dos Elementos Finitos (MEF), que resolvem numericamente as equações de Maxwell [10]. Além da seleção do nanoaquecedor, o planejamento racional da terapia fototérmica requer a previsão da distribuição espacial da temperatura e do dano térmico produzido durante a irradiação. A distribuição de calor no tecido biológico pode ser obtida por meio da equação de bioaquecimento de Pennes, que considera os efeitos combinados da condução térmica, perfusão sanguínea, metabolismo e absorção óptica [13]. Finalmente, a extensão do dano térmico irreversível pode ser estimada pelo modelo cinético de Arrhenius, que relaciona a evolução temporal da temperatura aos processos de desnaturação proteica e morte celular [14], [15].

Embora avanços significativos tenham sido alcançados na terapia fototérmica assistida por nanopartículas plasmônicas, ainda permanecem desafios relacionados à seleção de nanoaquecedores mais eficientes, à definição dos parâmetros de irradiação e à previsão quantitativa do aquecimento e do dano térmico em tecidos biológicos. Nesse contexto, este trabalho propõe uma metodologia integrada para o planejamento racional da terapia fototérmica assistida por nanopartículas plasmônicas. A metodologia combina simulações eletromagnéticas para seleção de nanoaquecedores, modelagem térmica baseada na equação de Pennes, estimativa do dano térmico pelo modelo de Arrhenius e validação experimental em camundongos portadores de Sarcoma 180 tratados com nanobastões de ouro irradiados por laser de 808 nm.

II. Metodologia

Foram investigados nanobastões (NBs) e nanocascas (NCs) de ouro com ressonância plasmônica próxima de 800 nm, visando identificar nanoaquecedores com elevada eficiência fototérmica. As propriedades ópticas das nanopartículas foram obtidas por simulações tridimensionais utilizando o Método dos Elementos Finitos no software COMSOL Multiphysics, considerando uma nanopartícula imersa em água e excitada por uma onda plana linearmente polarizada. A permissividade dielétrica do ouro foi descrita pelo modelo de Drude modificado com correção para efeitos de tamanho [16], [17]. Na seleção do agente fotoaquecedor plasmônico, foram consideradas nanoestruturas cuja ressonância plasmônica ocorre em aproximadamente 800 nm. Para os nanobastões, diferentes orientações em relação ao campo elétrico incidente foram consideradas, permitindo calcular o valor médio angular da eficiência fototérmica.

A seleção dos nanoaquecedores foi realizada por meio do Número de Joule, definido como [18]:

$$Jo = \frac{\lambda_{ref}}{2\pi}\frac{\sigma_{abs}}{V_{np}}, \tag{1}$$

em que V_{np} representa o volume da nanopartícula e $\lambda_{ref} = 1240$ nm corresponde ao comprimento de onda associado à energia de 1 eV.

A distribuição espaço-temporal da temperatura foi obtida resolvendo numericamente a equação de bioaquecimento de Pennes [13], dada por:

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot (k\nabla T) + Q_r + Q_b + Q_m, \tag{2}$$

em que ρ, c_p e k correspondem, respectivamente, à densidade, ao calor específico e à condutividade térmica do tecido, enquanto Q_r, Q_b e Q_m representam as contribuições da absorção óptica, da perfusão sanguínea e do metabolismo. O termo fonte associado à absorção óptica (Q_r) foi obtido explorando simulações de transporte de fótons utilizando o método de Monte Carlo [19], [20]. O modelo numérico reproduziu um tumor de Sarcoma 180 implantado em camundongos, considerando as diferentes camadas do tecido biológico. Os parâmetros ópticos, térmicos e fisiológicos empregados nas simulações foram obtidos da literatura [21]–[26]. Foram avaliadas diferentes concentrações de nanopartículas e irradiâncias do laser.

O dano térmico irreversível foi estimado utilizando o modelo cinético de Arrhenius, dado por [14]:

$$\Omega(t) = \int_0^t A \exp\left[-\frac{E_a}{RT(\tau)}\right] d\tau, \tag{3}$$

em que A é o fator pré-exponencial, E_a a energia de ativação, R a constante universal dos gases e T a temperatura absoluta. Valores de $\Omega \geq 1$ foram considerados indicativos de dano térmico irreversível.

As predições numéricas foram validadas experimentalmente explorando camundongos Swiss portadores de Sarcoma 180. O manejo dos animais durante o experimento foi aprovado pelo Comitê de Ética em Experimentação Animal da Universidade Federal de Pernambuco (UFPE), sob o processo nº 0088/2023. Os animais receberam injeção intratumoral de nanobastões de ouro PEGilados (41 × 10 nm) na concentração de 100 μg/mL, correspondendo a uma concentração numérica de 1.4×10^{17} NP/m^3 no volume tumoral, seguida de irradiação com laser pulsado de 808 nm (140 fs, 80 MHz) e irradiância de 1 W/cm^2. A evolução da temperatura foi monitorada por câmera termográfica (FLIR E4) e comparada aos resultados das simulações.

III. Resultados e Discussão

A Figura 1 apresenta os valores calculados do Número de Joule para nanocascas de ouro com núcleo de sílica (indicado o diâmetro total da nanopartícula) e para nanobastões de ouro (indicado o comprimento do nanobastão), Em ambos os casos, as dimensões das nanoestruturas foram ajustadas de modo a posicionar a LSPR em aproximadamente 800 nm. A curva azul representa a média angular de Jo, considerando todas as possíveis orientações entre o vetor de onda da radiação incidente e o eixo longitudinal do nanobastão. Observa-se que o nanobastão otimizado para fotoaquecimento em 800 nm apresenta dimensões aproximadas de 41×10 nm, correspondendo ao maior valor médio de Jo (160,7). Diferentemente dos nanobastões, as nanocascas de ouro exibem resposta fototérmica independente da orientação da nanoestrutura. A curva vermelha apresenta os valores de Jo para diferentes diâmetros de nanocascas. Verifica-se que a nanocasca de ouro

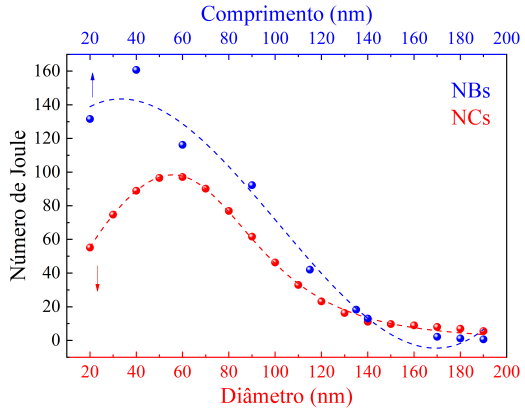


Fig. 1: Número de Joule em função do comprimento do nanobastão (curva azul) e diâmetro total da nanocasca (curva vemerlha).

com núcleo de diâmetro 53 nm e espessura de casca de $3,5$ nm apresenta o maior valor de Jo (97). Apesar do elevado valor de Jo obtido para as nanocascas, seu desempenho fototérmico ainda é inferior ao dos nanobastões. Isso ocorre porque, para uma dada concentração de nanopartículas e irradiância I, o termo de geração volumétrica de calor associado às nanopartículas é proporcional ao Número de Joule, de modo que $Q \propto JoI$. Assim, o maior valor de Jo do nanobastão de 41×10 nm implica maior geração de calor sob a mesma irradiância, justificando sua seleção como nanoaquecedor para as etapas subsequentes da análise térmica.

Com base na equação de bioaquecimento de Pennes, foram construídos mapas da distribuição de temperatura no tecido tumoral alvo. As Figuras 2(a) e 2(b) apresentam a distribuição térmica prevista para o tecido na ausência e na presença de nanobastões de ouro, respectivamente, considerando uma concentração de 1.4×10^{17} NP/m^3, intensidade luminosa incidente de 1 W/cm^2 e tempo de irradiação de $600s$. Observa-se que a elevação da temperatura do tecido depende diretamente da concentração de nanoaquecedores empregada. A Figura 2(c) mostra que a temperatura máxima no tecido aumenta monotonicamente com o aumento da concentração de nanopartículas. Entretanto, para concentrações mais elevadas, esse aumento torna-se menos pronunciado, indicando uma tendência à saturação.

A análise da dinâmica de aquecimento do tecido permite estimar, com base no modelo de Arrhenius, a extensão dos danos térmicos irreversíveis fotoinduzidos. A Figura 3 apresenta a evolução da área de dano irreversível após 600 s de irradiação laser para diferentes intensidades de iluminação e concentrações de nanobastões de ouro. Como esperado, a área normalizada de dano irreversível aumenta com a intensidade da irradiação laser e com a concentração de nanopartículas. Para uma mesma irradiância, concentrações mais elevadas de nanobastões promovem regiões mais extensas de dano térmico irreversível, em decorrência da maior deposição de energia óptica no interior do tumor. Esses resultados evidenciam a

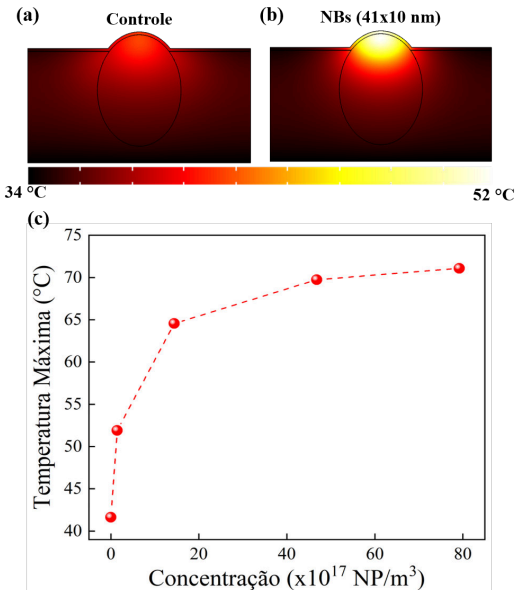


Fig. 2: Distribuição térmica prevista para o tecido (a) na ausência de NBs e (b) na presença de NBs de ouro de 41x10 nm. (c) Alcance térmico em função da concentração de NBs.

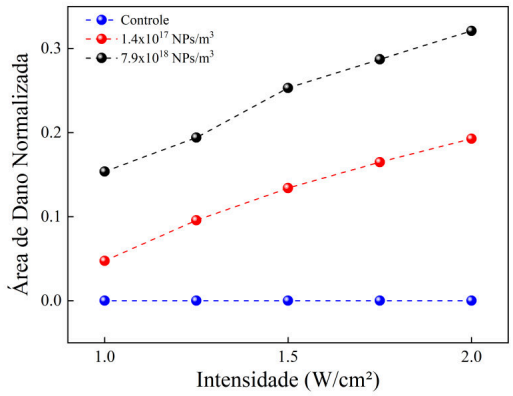


Fig. 3: Área do dano térmico irreversível normalizada na ausência de NBs (curva azul), contendo NBs de ouro de 41x10 nm com concentração de 1.4×10^{17} NP/m^3 (curva vermelha) e 7.9×10^{18} NP/m^3 (curva preta).

importância da otimização conjunta da intensidade do laser e da concentração de nanopartículas, de modo a maximizar a eficiência da ablação tumoral, evitando simultaneamente o emprego de doses excessivas de nanopartículas e exposições desnecessariamente elevadas à radiação laser.

A validação da metodologia foi realizada por meio da análise experimental da elevação de temperatura em camundongos submetidos ao fotoaquecimento com laser de intensidade 1 W/cm^2 durante 10 min. A Figura 4 apresenta a evolução da temperatura superficial da pele dos animais na ausência e na presença de nanopartículas, considerando uma concentração de 1.4×10^{18} NP/m^3. As linhas sólidas representam as predições do modelo teórico proposto, evidenciando

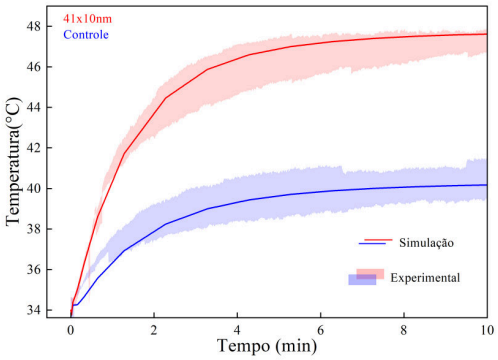


Fig. 4: Evolução temporal da temperatura na região tumoral obtida computacionalmente e experimentalmente.

a concordância com os resultados experimentais ao longo de todo o período de irradiação. O modelo apresentou um erro absoluto médio de 1,3 °C para a condição de controle e de 3,7 °C para a condição contendo nanopartículas.

IV. Conclusão

Foi apresentada uma metodologia integrada para o planejamento da terapia fototérmica assistida por nanopartículas plasmônicas, combinando simulações eletromagnéticas, modelagem térmica, estimativa do dano tecidual e validação experimental. Os resultados mostraram que o Número de Joule é uma métrica eficiente para a seleção de nanoaquecedores, identificando os nanobastões de ouro como a geometria de melhor desempenho para excitação em aproximadamente 800 nm. Além disso, verificou-se que a eficiência da terapia depende da otimização conjunta da concentração de nanopartículas e da irradiância do laser. A boa concordância entre as predições numéricas e os resultados experimentais demonstra o potencial da metodologia como ferramenta para o planejamento e a otimização de protocolos de terapia fototérmica assistida por nanoagentes fototérmicos.

Agradecimento

Os autores agradecem o suporte do CNPq, da CAPES e da FACEPE.

Referências

[1] L. R. Hirsch, R. J. Stafford, J. Bankson, S. R. Sershen, B. Rivera, R. Price, J. D. Hazle, N. J. Halas, and J. L. West, "Nanoshell-mediated near-infrared thermal therapy of tumors under magnetic resonance guidance," *Proceedings of the National Academy of Sciences*, vol. 100, no. 23, pp. 13 549–13 554, 2003.

[2] F. Oudjedi and A. G. Kirk, "Near-infrared nanoparticle-mediated photothermal cancer therapy: a comprehensive review of advances in monitoring and controlling thermal effects for effective cancer treatment," *Nano Select*, vol. 6, no. 3, p. e202400107, 2025.

[3] Y. Li, X. Zhu, Y. Dong, Y. Yang, D. Shen, Z. Li, R. Li, X. Dang, Z. Qin, and K. Fan, "Nanomedicine-enabled mild photothermal therapy strategies for enhanced antitumor treatment," *Advanced NanoBiomed Research*, vol. 4, no. 3, p. 2300094, 2024.

[4] X. Huang, P. K. Jain, I. H. El-Sayed, and M. A. El-Sayed, "Gold nanoparticles: interesting optical properties and recent applications in cancer diagnostics and therapy," *Nanomedicine*, vol. 2, no. 5, pp. 681–693, 2007.

[5] D. Jaque, L. M. Maestro, B. Del Rosal, P. Haro-Gonzalez, A. Benayas, J. Plaza, E. M. Rodriguez, and J. G. Sole, "Nanoparticles for photothermal therapies," *Nanoscale*, vol. 6, no. 16, pp. 9494–9530, 2014.

[6] E. C. Dreaden, L. A. Austin, M. A. Mackey, and M. A. El-Sayed, "Size matters: gold nanoparticles in targeted cancer drug delivery," *Therapeutic Delivery*, vol. 3, no. 4, pp. 457–478, 2012.

[7] D. P. O'Neal, L. R. Hirsch, N. J. Halas, J. D. Payne, and J. L. West, "Photo-thermal tumor ablation in mice using near infrared-absorbing nanoparticles," *Cancer Letters*, vol. 209, no. 2, pp. 171–176, 2004.

[8] X. Huang, I. H. El-Sayed, W. Qian, and M. A. El-Sayed, "Cancer cell imaging and photothermal therapy in the near-infrared region by using gold nanorods," *Journal of the American Chemical Society*, vol. 128, no. 6, pp. 2115–2120, 2006.

[9] S. M. Mousavi *et al.*, "Multifunctional gold nanorod for therapeutic applications and pharmaceutical delivery considering cellular metabolic responses, oxidative stress and cellular longevity," *Nanomaterials*, vol. 11, no. 7, p. 1868, 2021.

[10] T. L. Pedrosa, G. M. Oliveira, A. C. Pereira, M. J. Crispim, L. A. Silva, M. S. Silva, I. A. Souza, A. M. Melo, A. S. Gomes, and R. E. de Araujo, "Tailoring plasmonic nanoheaters size for enhanced theranostic agent performance," *Bioengineering*, vol. 11, no. 9, p. 934, 2024.

[11] J. Domingo-Diez *et al.*, "Effect of photothermal therapy using gold nanoparticles conjugated with hyaluronic acid in an intracranial murine glioblastoma model," *International Journal of Nanomedicine*, pp. 9327–9346, 2025.

[12] T. L. Pedrosa, S. Farooq, and R. E. de Araujo, "Selecting high-performance gold nanorods for photothermal conversion," *Nanomaterials*, vol. 12, no. 23, p. 4188, 2022.

[13] H. H. Pennes, "Analysis of tissue and arterial blood temperatures in the resting human forearm," *Journal of Applied Physiology*, vol. 1, no. 2, pp. 93–122, 1948.

[14] J. A. Pearce, "Comparative analysis of mathematical models of cell death and thermal damage processes," *International Journal of Hyperthermia*, vol. 29, no. 4, pp. 262–280, 2013.

[15] P. N. Prasad, *Introduction to Biophotonics*. John Wiley & Sons, 2004.

[16] J. Lerme, C. Bonnet, M.-A. Lebeault, M. Pellarin, and E. Cottancin, "Surface plasmon resonance damping in spheroidal metal particles," *Journal of Physical Chemistry C*, vol. 121, no. 10, pp. 5693–5708, 2017.

[17] S. Berciaud, L. Cognet, P. Tamarat, and B. Lounis, "Observation of intrinsic size effects in the optical response of individual gold nanoparticles," *Nano Letters*, vol. 5, no. 3, pp. 515–518, 2005.

[18] G. Baffou and R. Quidant, "Thermoplasmonics," in *World Scientific Handbook of Metamaterials and Plasmonics*. World Scientific, 2018, vol. 4, pp. 379–407.

[19] S. L. Jacques and B. W. Pogue, "Tutorial on diffuse light transport," *Journal of Biomedical Optics*, vol. 13, no. 4, p. 041302, 2008.

[20] H.-J. Kim *et al.*, "Laser–tissue interaction simulation considering skin-specific data to predict photothermal damage lesions during laser irradiation," *Journal of Computational Design and Engineering*, vol. 10, no. 3, pp. 947–958, 2023.

[21] J. L. Sandell and T. C. Zhu, "A review of in-vivo optical properties of human tissues and its impact on pdt," *Journal of Biophotonics*, vol. 4, no. 11-12, pp. 773–787, 2011.

[22] L. Azzi, M. El-Alfy, C. Martel, and F. Labrie, "Gender differences in mouse skin morphology and specific effects of sex steroids and dehydroepiandrosterone," *Journal of Investigative Dermatology*, vol. 124, no. 1, pp. 22–27, 2005.

[23] C. P. Sabino *et al.*, "The optical properties of mouse skin in the visible and near infrared spectral regions," *Journal of Photochemistry and Photobiology B: Biology*, vol. 160, pp. 72–78, 2016.

[24] P. Wongchadakul and P. Rattanadecho, "Mathematical modeling of multilayered skin with embedded tumor through combining laser ablation and nanoparticles," *International Journal of Heat and Technology*, vol. 39, no. 1, 2021.

[25] Y. Ren, H. Qi, Q. Chen, and L. Ruan, "Thermal dosage investigation for optimal temperature distribution in gold nanoparticle enhanced photothermal therapy," *International Journal of Heat and Mass Transfer*, vol. 106, pp. 212–221, 2017.

[26] Z. Li, S. Tang, B. Wang, Y. Li, H. Huang, H. Wang, P. Li, C. Li, P. K. Chu, and X.-F. Yu, "Metabolizable small gold nanorods: size-dependent cytotoxicity, cell uptake and in vivo biodistribution," *ACS Biomaterials Science & Engineering*, vol. 2, no. 5, pp. 789–797, 2016.

VitaCheck: Portable Triage Case for Indigenous Communities

1st Joyce Hemily Amorim Vieira
Biomedical Engineering Student
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
joyce.vieira@geb.inatel.br

2nd Larissa Maria Santos de Almeida
Biomedical Engineering Student
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
larissa.almeida@geb.inatel.br

3rd Matheus Henrique Gonçalves Faria
Biomedical Engineering Student
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
matheus.hf@geb.inatel.br

4th Lucas Molina Bruno
Biomedical Engineering Student
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
lucas.bruno@geb.inatel.br

5th Raphael Cardoso Mota Pereira
Department of Production Engineering
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
raphaelpereira@inatel.br

6th Maysa Costa Alves
Department of Biomedical Engineering (Advisor)
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

***Abstract*—Access to healthcare services in Brazilian Indigenous communities is often limited by geographic distance, a shortage of professionals, and the lack of measurement equipment at the point of care. This work presents VitaCheck, a low-cost, portable kit to potentially aid health triage, designed for use by health workers in primary care contexts with limited infrastructure. The device integrates an ESP32 microcontroller with sensors for heart rate and peripheral oxygen saturation (MAX30102), non-contact infrared body temperature (MLX90614), and contact temperature (DS18B20), as well as an LED-based colorimetric module for reading urine reagent strips, extending triage to parameters such as protein and glucose. A piezoelectric actuator signals alerts, and data are formatted into JSON and transmitted via Bluetooth to a mobile device, enabling record-keeping and care prioritization. The entire system is housed in a 3D-printed enclosure with a footprint equivalent to an A4 sheet. The prototype is currently under development, and the proposed experimental evaluation will investigate its accuracy, repeatability, and applicability as a triage support tool.**

***Index Terms*—health triage, indigenous health, telehealth, biomedical instrumentation, ESP32, pulse oximetry, colorimetric urinalysis, low-cost devices.**

I. Introduction

Ensuring equitable access to healthcare services is a persistent challenge for populations living in remote regions, among which Brazilian Indigenous communities stand out. Within the scope of the Unified Health System (SUS), the Indigenous Health Care Subsystem and the National Health Care Policy for Indigenous Peoples were created to ensure differentiated care for these peoples, considering their sociocultural, epidemiological, and logistical particularities [1]. Despite advances, major challenges persist, such as long distances to health units, care discontinuity, high turnover of healthcare professionals, and the lack of basic measurement equipment at the point of care [1]. Primary healthcare for Indigenous populations is operationalized by the Special Indigenous Sanitary Districts and executed at the local level by multidisciplinary teams that often work in villages far from reference health units [1]. In this context, simple instruments that assist health workers in the early identification of cases requiring referral can potentially contribute to comprehensive care and to the rational use of emergency transport and removal resources, which are historically costly [1].

In this scenario, triage — the process of prioritizing care based on objective clinical signs — plays a central role in the rational use of available resources and is aligned with the principles of primary healthcare [2]. Measuring vital signs such as heart rate, peripheral oxygen saturation (SpO_2), and body temperature provides relevant information for this classification. Urine analysis using reagent strips, in turn, is a rapid and low-cost test widely used for general health screening, aiding in the identification of alterations such as proteinuria and glucosuria, which may indicate kidney disease, diabetes, or other conditions [3], [4]. However, these tests often rely on multiple standalone, high-cost devices that are not always available in the field.

Furthermore, alongside all the difficulties in healthcare delivery, public investment in Indigenous health fluctuates depending on the governing administration, and mortality indicators for various diseases remain concerning. For instance, the risk of an Indigenous child dying prematurely is 2.44 times higher than in the non-Indigenous population, with neonatal and post-neonatal mortality remaining 55% above the national average [5], [6]. Thus, the need for alternative measures to access initial information is reinforced — information that is often obtained through advanced triage protocols in emergency

services. Studies indicate that triage can potentially reduce care time and costs [7], and a low-cost, portable triage support kit will likely assist the Special Secretariat of Indigenous Health (SESAI) workforce of 22,000 healthcare professionals — 52% of whom are Indigenous — in potentially serving more than 762,000 Indigenous people living in villages across Brazil [8].

The diffusion of affordable microcontrollers, integrated sensors, and wireless communication has enabled low-cost digital health solutions targeting resource-limited environments, where the point of care often coincides with the patient's home or village [9], [10]. Solutions of this type do not intend to replace clinical examination, but are expected to offer objective triage and data recording that enhance the quality of referral decisions. It is within this gap — the lack of a single, portable, and low-cost instrument that brings together objective measurements to support field triage — that this work is situated.

The main contributions of this work are: (i) the design and integration of a low-cost ESP32-based hardware platform for multiparametric acquisition of vital signs; (ii) the development of an LED-based colorimetric module for automated reading of urine reagent strips, aiming to reduce the subjectivity of visual inspection; (iii) the definition of a triage logic based on configurable rules, with audible alerts and JSON data structuring; and (iv) the proposal of a 3D-printed portable enclosure intended for field use.

This paper proposes VitaCheck, a portable triage kit that combines, into a single low-cost device, automated vital sign acquisition, colorimetric reading of urine reagent strips, a rule-based classification logic, and wireless data transmission to a mobile application. The objective of this study is to offer health workers an accessible, replicable, and easy-to-operate tool suited to the infrastructure constraints found when delivering care to Indigenous communities.

The remainder of this article is organized as follows. Section II discusses related work. Section III details the materials and methods, which will be discussed in Section IV. Section V presents the conclusions, and Section VI outlines future work.

II. Related Work

Recent literature describes several vital signs monitoring devices based on low-cost microcontrollers. Platforms such as the ESP32 are widely used due to their availability, low cost, processing capability, and native Wi-Fi and Bluetooth connectivity [9], [11]. The MAX30102 sensor, based on photoplethysmography (PPG), is broadly employed for measuring heart rate and SpO_2 in wearable and low-cost devices, which are frequently combined with temperature sensors and 3D-printed enclosures for use in resource-limited contexts [9], [12]. Such platforms typically integrate the vital signs sensor via the I^2C bus, a sound actuator for alarms, and a wireless communication interface, potentially offering a lower-cost alternative to commercial monitors that could be deployed in field scenarios.

In the field of urinalysis, reagent strips function as solid-phase colorimetric sensors, in which pads impregnated with reagents change color upon reacting with urinary analytes such as glucose, protein, and nitrite [4]. However, visual reading of these strips can be subject to errors resulting from illumination variations and color perception differences among observers [10]. To mitigate this issue, several studies propose optical readers: Yang et al [3] present a Point-of-Care Testing (POCT) device that illuminates the strip inside a dark chamber and analyzes color through color space conversion; Laucher et al [10] propose smartphone-based colorimetric analysis for home prenatal care; and Çelik et al [4] review point-of-care urinalysis approaches supported by smartphones and artificial intelligence that could shape the future of decentralized diagnostics.

Compared with existing solutions, the distinct advantage of VitaCheck lies in combining vital signs and colorimetric urinalysis with embedded triage logic within a single portable, low-cost device. This integrated architecture is expected to provide an explicit focus on the context of healthcare delivery to Indigenous populations in hard-to-reach regions, potentially improving healthcare coverage and significantly optimizing field decision-making in future deployments.

III. Materials and Methods

A. Design Requirements

The design of VitaCheck stemmed from a set of requirements derived from the usage context: (i) low cost, enabling replication; (ii) portability, for transport between villages; (iii) low power consumption and the option of battery operation, given the unavailability of a stable power grid; (iv) robustness for transport and adverse environmental conditions; (v) simple operation, allowing use by community health workers without specialized technical training; and (vi) data logging and communication for referrals. These requirements guided the selection of components and the architecture described below.

B. Selection of Triage Parameters

The adopted parameters were chosen for their triage value and the feasibility of measurement using low-cost sensors. SpO_2 and heart rate are objective indicators of cardiorespiratory status and routinely form part of risk classification protocols. Body temperature aids in identifying febrile states, which are frequently associated with infectious processes. Urinalysis using reagent strips adds a low-cost biochemical axis: the presence of protein may signal renal impairment, whereas the presence of glucose may indicate hyperglycemia; leukocytes and nitrite may suggest a urinary tract infection [3], [4]. Combining these two axes — physiological and biochemical — is expected to broaden the scope of triage without significantly increasing equipment costs.

C. System Overview and Architecture

VitaCheck comprises a central processing unit (ESP32 microcontroller), a sensing layer for vital signs, a colorimetric module for urinalysis, an audible actuator for alert signaling, and a wireless communication module. The ESP32 was chosen because it natively integrates Wi-Fi and Bluetooth

connectivity, possesses adequate processing capacity for the application, and is low-cost [11]. In terms of architecture, the ESP32 occupies the central position and concentrates the control of the entire system. Connected to its inputs are the vital signs sensors (MAX30102, MLX90614, and DS18B20) and the colorimetric urinalysis module (composed of LEDs and a photodetector); connected to its outputs are the piezoelectric actuator, responsible for audible alerts, and the wireless communication interface (Bluetooth), through which data will be sent to the mobile device. Sensors using the I^2C interface share the same bus, whereas the DS18B20 and the colorimetric module utilize dedicated input and output pins of the microcontroller.

The vital signs sensing layer integrates three sensors connected to the ESP32:

- **MAX30102** — an optical sensor that measures heart rate and SpO_2 via photoplethysmography by emitting red and infrared light and detecting variations in arterial blood volume; communicates via I^2C interface [12].
- **MLX90614** — an infrared thermometer allowing non-contact body temperature measurement, also via I^2C [13].
- **DS18B20** — a digital contact temperature sensor used as a reference/complement, with 1-Wire communication [14].

According to their respective datasheets, the MLX90614 presents a typical accuracy of ± 0.5 °C in the body temperature range, the DS18B20 an accuracy of ± 0.5 °C between -10 °C and $+85$ °C, and the MAX30102 adequate resolution for HR and SpO_2 measurement in wearable devices [12]–[14].

The MAX30102 estimates heart rate (HR) from the time interval Δt between successive peaks of the PPG signal, as:

$$HR = 60/\Delta t \tag{1}$$

with Δt in seconds and HR in beats per minute. Peripheral oxygen saturation is estimated by the ratio-of-ratios R between the alternating (AC) and direct (DC) components of red and infrared light:

$$R = (AC_{\text{red}}/DC_{\text{red}})/(AC_{\text{IR}}/DC_{\text{IR}}) \tag{2}$$

which is related to SpO_2 through an empirical calibration curve of the form $SpO_2 \approx a - b \cdot R$, whose coefficients a and b will be obtained through calibration [12]. Since the PPG signal is sensitive to noise and motion artifacts, filtering and averaging will be applied to the readings prior to calculation. The two I^2C sensors share the same bus, distinguished by their addresses, while the DS18B20 uses a dedicated 1-Wire bus.

D. Colorimetric Module for Urinalysis

To extend triage beyond vital signs, VitaCheck incorporates an optical reading module for urine reagent strips. The operating principle is based on reflectance: a set of LEDs illuminates each reagent pad on the strip, and a photodetector measures the reflected light, whose color

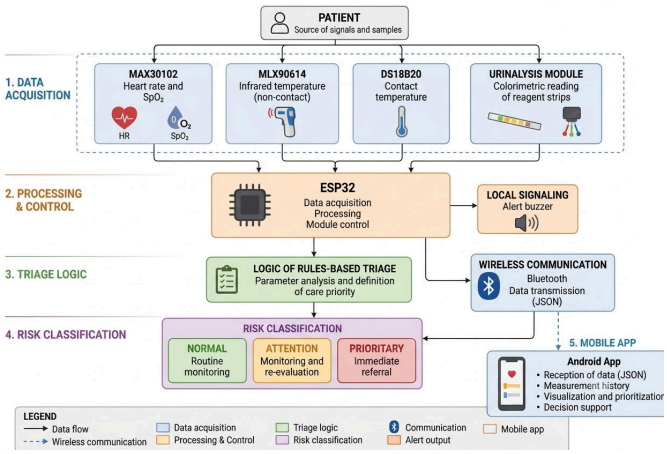


Figure 1. Assembly of the ESP32 with the MAX30102 sensor. Source: The authors (2026)

varies according to analyte concentration [3], [10]. The measured intensity/color is then converted into a concentration range using a calibration curve constructed from the manufacturer's reference color chart.

To reduce environmental light interference — one of the primary sources of error in colorimetric reading [10] — the reading was designed to take place inside a dark chamber with controlled internal illumination, a strategy also adopted in optical urine readers reported in the literature [3]. In this work, the TCS34725 RGB color sensor (I^2C interface, with an integrated white LED) is adopted to capture reflected light; alternatively, discrete LEDs in red (∼625 nm), green (∼525 nm), and blue (∼470 nm) associated with a photodiode could be employed. Reading each pad must be performed after the reaction time recommended by the strip manufacturer, typically between 30 and 60 s.

The reading is normalized against a white reference to mitigate variations in the light source. For each pad, relative reflectance is defined as:

$$R_\lambda = I_\lambda / I_{\lambda,\text{ref}} \tag{3}$$

where I_λ is the measured intensity in color channel λ and $I_{\lambda,\text{ref}}$ is the reference intensity. Channel reflectances will then be mapped to the analyte concentration range via a calibration curve combining red, green, and blue (RGB) channel reflectances, fitted based on the manufacturer's color chart levels. The measurement procedure consists of: (1) dipping the strip into the sample; (2) waiting for the reaction time specified by the manufacturer, a critical factor for result accuracy; (3) inserting the strip into the chamber; (4) reading and normalization; and (5) classification by range.

Multi-parameter reagent strips allow screening for several analytes. In this work, a commercially available multi-parameter reagent strip is used.

E. Triage Logic

Following acquisition, parameter values will be compared against configurable reference ranges, generating a priority classification (e.g., normal, caution, and priority). When at least one parameter exceeds the defined threshold, the piezoelectric actuator is triggered to emit an audible alert, and the priority level will be included in the output message. The decision flow follows four steps: (i) acquisition of sensor measurements; (ii) comparison of each measurement against configured reference ranges; (iii) classification of the priority level based on the worst condition found among the parameters; and (iv) triggering of the audible alert, when applicable, and transmission of the JSON message.

F. Firmware and Processing

The firmware embedded in the ESP32 is responsible for sensor initialization, periodic measurement acquisition, basic signal filtering (especially the PPG signal from the MAX30102), application of the triage logic, and data packetization and transmission. The firmware was developed in C/C++ using the Arduino IDE, employing the SparkFun MAX3010x (MAX30102), Adafruit MLX90614, OneWire, and DallasTemperature (DS18B20) libraries, along with BluetoothSerial for Bluetooth communication. To reduce PPG signal noise, moving averages and rejection of invalid readings are applied before calculating HR and SpO_2. The organization of the firmware into well-defined stages — initialization, acquisition, processing, classification, and transmission — is expected to facilitate maintenance and component swapping, enabling adjustments to parameters (such as averaging windows and triage thresholds) without altering the core program structure.

G. Communication and Data Structuring

Data from each measurement are structured into JSON format and transmitted via Bluetooth to a mobile device, potentially enabling visualization, record-keeping, and care prioritization. The use of JSON aims to facilitate interoperability with mobile applications and health information systems. An example message is shown below:

```
{ "hr": 78, "spo2": 97,
  "temp_ir": 36.5, "temp_contact": 36.7,
  "urine": { "protein": "negative",
             "glucose": "negative" },
  "priority": "normal" }
```

H. Mobile Application and Data Logging

On the paired mobile device, an application receives the JSON messages, displays measured values and priority classification, and maintains a measurement history. This record is expected to allow longitudinal patient monitoring and provide structured information for referral decisions by the healthcare team. Using a standardized, machine-readable format can support future integration with health information systems and organized data export. The application is developed for the Android platform and displays received values, priority classification, and measurement history, with local storage of records on the device. Figure 2 below shows the main application screen featuring the risk classification.



Figure 2. Main application screen displaying risk classification. Source: The authors (2026)

I. Enclosure (Kit)

The electronic assembly was housed in a 3D-printed enclosure with a base footprint equivalent to an A4 sheet, designed to protect components, organize wiring, accommodate the dark chamber for strip reading, and potentially facilitate field transport. The enclosure was printed in PLA, with approximate dimensions of 210 $\times$ 300 $\times$ 50 mm (footprint area equivalent to an A4 sheet) and a sliding drawer-style lid, which should allow access to components and proper positioning and fixing of sensors for measurement.

J. Field Workflow

The intended operation workflow will follow these steps: (1) the health worker turns on the device and pairs it with the mobile application via Bluetooth; (2) places the MAX30102 sensor on the patient's finger and measures HR and SpO_2; (3) performs temperature measurement

with the MLX90614 and/or DS18B20; (4) when indicated, inserts the urine reagent strip into the colorimetric module; (5) the device applies the triage logic, emits an audible alert if priority status is met, and sends the JSON message to the application; and (6) the result is logged to potentially support the referral decision. The process was designed to be completed in a few minutes with minimal training requirements.



Figure 3. Screen displaying patient registration. Source: The authors (2026)

K. Prototype Cost

A core requirement of the project is low cost. The total estimated cost of the components is approximately R$ 200; however, any Point-of-Care Testing (POCT) equipment properly regulated by ANVISA (RDC 302/2005; RDC 786/2023; RDC 978/2025) will ultimately have to comply with various requirements, even if these have been made more flexible in recent years.

L. Evaluation Procedure

The evaluation will consist of comparing VitaCheck measurements with calibrated commercial equipment used as references: a fingertip pulse oximeter for HR and SpO_2, and a clinical thermometer for temperature. For urinalysis, the colorimetric module's automated classification will be compared against visual inspection using the strip manufacturer's color chart. Measurements will be conducted on healthy volunteers under controlled environmental conditions, repeating each reading to verify measurement stability. For each parameter, mean error, standard deviation, and agreement relative to the reference will be calculated to rigorously assess overall accuracy.

IV. Discussion

The interpretation of the results must consider both the accuracy of each measurement and the usability of the overall assembly under real field conditions, including total care time and ease of operation for the health worker. Among the expected limitations are the accuracy of low-cost sensors, the sensitivity of the MAX30102 to motion and peripheral perfusion, the dependence of urinalysis on reaction time and controlled illumination [10], and the need for calibration and clinical validation prior to clinical deployment.

The context of use in Indigenous communities also imposes specific requirements, such as energy autonomy in locations without a stable power grid, transport robustness, simple maintenance, an adapted user interface, and respect for sociocultural diversity [1].

Integrating vital signs and urinalysis into a single device tends to reduce the amount of equipment to transport and maintain, which is desirable in the field; on the other hand, it increases the need for calibration and care with consumables (reagent strips), which have specific shelf lives and storage requirements.

The sustainability of the solution depends, therefore, not only on the cost of the hardware, but also on the supply chain for consumables and the training of health workers for operation and alert interpretation.

In relation to related work, devices based on ESP32 and MAX30102 have already demonstrated performance comparable to reference oximeters for SpO_2 and HR in resource-limited scenarios [9], [12], while optical urine strip readers report good agreement with clinical inspection when illumination is controlled [3], [10].

VitaCheck seeks to combine these two capabilities into a single field device; verifying that this integration preserves the accuracy of each isolated measurement is precisely the objective of the experimental evaluation presented in Section IV.

The potential consolidation of VitaCheck as a triage tool requires a staged validation roadmap: bench evaluation with reference measurements and solutions, comparison against calibrated commercial equipment, repeatability tests, and, subsequently, clinical environment studies with appropriate ethical approval. The scalability of the solution benefits from the low cost and replicability of the hardware, which allow the production of multiple units,

provided that the supply of consumables and training for healthcare teams are secured.

However, the durability of low-cost components, particularly sensors, may influence the overall product lifecycle. Thus, field tests are required to verify the resistance and durability of components in the specific forest environment, characterized by high humidity and heat.

It is worth highlighting ethical and regulatory considerations: devices intended to support health decisions require, for actual clinical use, clinical validation and compliance with applicable standards, such as ANVISA regulatory requirements for medical devices and Brazilian ethical standards for research involving human subjects (CNS Resolution No. 466/2012), as well as specific provisions for research involving Indigenous peoples.

VitaCheck, in its current stage, is a triage support prototype rather than a certified medical device, and it does not replace clinical evaluation.

V. Conclusion

This work presented VitaCheck, a low-cost, portable health triage kit designed for Indigenous communities, integrating multi-parametric acquisition of vital signs, colorimetric reading of urine reagent strips, rule-based classification logic, and wireless data transmission in JSON format.

The total of 22,000 healthcare professionals dedicated to servicing the Indigenous population highlights the economic potential for large-scale production of the equipment, underscoring its potential impact on the quality of life of a population of 762,000 Indigenous individuals through efficiency gains (time and cost) and improved satisfaction in advanced triage within emergency services. VitaCheck presents a promising architecture to support triage in limited-infrastructure settings. However, its accuracy, repeatability, and practical applicability still remain to be investigated through experimental evaluations and future testing.

VI. Future Work

We will conduct standard laboratory measurements, which limits confidence in its diagnostic precision. Additional studies involving representative patient populations, repeatability analyses, environmental robustness, battery autonomy, and usability evaluations under real field conditions will be performed to demonstrate its practical applicability.

We will discuss interoperability with electronic health record standards and data security during wireless transmission, aiming to strengthen the platform's translational perspective.

We will report ethical and consent criteria, reference device calibration, repeated measurements per individual, collection time, and individual paired errors.

We will validate heart rate, SpO2, and temperature across clinically relevant ranges and under adverse conditions



Figure 4. Screen displaying registered patients. Source: The authors (2026)

using agreement analysis. For colorimetry, all target analytes, strip lots, reaction times, illumination levels, camera variations, and interfering substances will be tested against an objective laboratory standard or a validated reader.

Community partnerships, cultural appropriateness, Indigenous data governance, maintenance, consumables, and connectivity will be incorporated. No actual field deployment should be inferred at this stage.

References

[1] A. M. Mendes, M. S. Leite, E. J. Langdon, and M. Grisotti, "O desafio da atenção primária na saúde indígena no Brasil," *Rev. Panam. Salud Publica*, vol. 42, p. e184, 2018.

[2] World Health Organization, "Declaration of Alma-Ata," in *Int. Conf. Primary Health Care*, Alma-Ata, USSR, Sep. 1978.

[3] Z. Yang, G. Cai, J. Zhao, and S. Feng, "An optical POCT device for colorimetric detection of urine test strips based on Raspberry Pi imaging," *Photonics*, vol. 9, no. 10, art. 784, 2022.

[4] H. Çelik, B. B. Caf, C. Geyik, G. Çebi, and M. Tayfun, "Enhancing urinalysis with smartphone and AI: a comprehensive review of point-of-care urinalysis and nutritional advice," *Chem. Pap.*, vol. 78, pp. 651–664, 2024.

[5] G. L. Marinho et al., "Mortalidade infantil de indígenas e não indígenas nas microrregiões do Brasil," *Revista Brasileira de Enfermagem*, vol. 72, pp. 57–63, 2019.



Figure 5. Screen displaying the analyses to be performed.
Source: The authors (2026)

[6] E. Souza et al. *Desigualdades em saúde de crianças indígenas* [e-book] / Comitê Científico do Núcleo Ciência Pela Infância. São Paulo: Fundação Maria Cecilia Souto Vidigal, 2024.

[7] C. B. Soster et al., "Protocolos de triagem avançada no serviço de emergência: revisão sistemática e metanálise," *Revista Latino-Americana de Enfermagem*, vol. 30, p. e3511, 2022.

[8] Saúde Indígena, Ministério da Saúde. Disponível em: https://www.gov.br/saude/pt-br/composicao/sesai/saude-indigena. Acesso em: 9 jul. 2026.

[9] ESP32-powered heart rate and pulse oximetry monitoring with MAX30102. Espressif Systems. 2021. Disponível em: https://www.espressif.com/en/news/ESP32_and_MAX30102. Acesso em: 12 jul. 2026.

[10] M. Flaucher et al., "Smartphone-based colorimetric analysis of urine test strips for at-home prenatal care," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–9, 2022.

[11] Espressif Systems, *ESP32 Series Datasheet*. Espressif Systems, 2023.

[12] Analog Devices (Maxim Integrated), *MAX30102 High-Sensitivity Pulse Oximeter and Heart-Rate Sensor for Wearable Health*, Datasheet.

[13] Melexis, *MLX90614 Family: Single and Dual Zone Infrared Thermometer in TO-39*, Datasheet.

[14] Analog Devices (Maxim Integrated), *DS18B20 Programmable Resolution 1-Wire Digital Thermometer*, Datasheet.

A Collaborative Brain-to-Brain Interface for θ/β Modulation During Sustained Attention Training of Two Individuals

1st Samara Tenório Oliveira Duarte Santana
Postgraduate Program in Neuroengineering (PPGN)
Edmond and Lily Safra International Institute of Neurosciences (ELS-IIN)
Rio Grande do Norte, Brazil
samara.duarte@edu.isd.org.br

2nd Rebeca Campos Morais Albuquerque
PPGN
ELS-IIN
Rio Grande do Norte, Brazil
rebeca.albuquerque@edu.isd.org.br

3rd Yuzen Hatano
Faculty of Advanced Science and Technology (FAST)
Kumamoto University (KU)
Kumamoto, Japan
yuzen2055@st.cs.kumamoto-u.ac.jp

4th Tomohiko Igasaki
FAST
KU
Kumamoto, Japan
iga@cs.kumamoto-u.ac.jp

5th Teodiano Bastos-Filho
Postgraduate Program in Electrical Engineering
Federal University of Espírito Santo
Espírito Santo, Brazil
teodiano@gmail.com

6th André F. Oliveira de Azevedo Dantas
PPGN
ELS-IIN
Rio Grande do Norte, Brazil
andre.dantas@edu.isd.org.br

7th Denis Delisle-Rodriguez
PPGN
ELS-IIN
Rio Grande do Norte, Brazil
denis.rodriguez@edu.isd.org.br

***Abstract*—Sustained attention is a crucial cognitive function for daily living activities, and its enhancement through neurofeedback or brain-computer interface interventions has gained great interest. This study presents the preliminary validation of an electroencephalography based brain-to-brain collaboration interface with a serious car-racing game for attention training. The system establishes an adaptive subject-specific collaborative neurofeedback model for two individuals jointly modulate their theta and beta cortical rhythms. We hypothesized that the participant with superior attentional skill may serve as a reference to dynamically guide the brain wave modulation of the other user. The collected brain signals during pilot test were processed, and preliminary results with two participants demonstrated high classification accuracy between 0.82 and 0.86 during the calibration phase. Throughout the 10-min collaborative training session, the participant with lower-performance successfully reduced the own theta/beta power ratio, stabilizing it closer to the reference participant's levels. Additionally, a significant negative correlation between the attention ratio and the game speed confirmed effective modulation. These findings suggest that collaborative brain-computer interfaces offer a promising and interactive approach to support sustained attention training.**

I. Introduction

Attention is a cognitive function very important in life living and in interaction day by day, allowing the communication between people, and is important to understanding and execution of professional, social and domestic activities that require alertness, perception, focus, selectivity, memorization, and learning. Furthermore, it is a fundamental mechanism to select relevant information in the brain, prioritizing pertinent stimuli while suppressing information considered irrelevant to cognitive processing.

The phenomenon of attention can be classified according to the way in which percepts are attended to. This research will analyze sustained attention, which consists of the voluntary maintenance of attention over time, while ignoring potential distractors and avoiding mind-wandering [1]. Sustained attention is composed of two fundamental aspects: (1) the duration for which performance is maintained and (2) the consistency of performance throughout that period. This attentional dimension is heavily dependent on inhibitory control and is therefore the most commonly impaired in individuals with attention deficits [2].

The neural architecture of attention comprises multiple large-scale networks that interact dynamically. This work primarily focus on the frontoparietal control network (FPN), comprising the lateral prefrontal cortex and inferior parietal lobule, supports working memory and flexible control of goal-directed behavior, along with the cingulo-opercular network, including the medial superior frontal cortex and anterior insula, sustaining stable task-set maintenance over longer periods [3], [4].

This work was partially supported by CNPq projects (Processes: 403247/2025-0, 445300/2024-8), Santos Dumont Institute, Brazilian Ministry of Education (MEC), Coordination for the Improvement of Higher Education Personnel (CAPES) and CNPq scholarships (Process: 313426/20252).

During specific activities involving cognition, motor function, or even resting states, neural oscillations are generated within functional networks, arising from the summation of excitatory and inhibitory postsynaptic activity in the brain. Theta oscillations (4-7 Hz) are associated with cognitive control and performance monitoring, signaling the need to increase attentional control when errors are detected or focus is lost. Beta oscillations (13-30 Hz) indicate an actively engaged and alert brain, with deficits in this frequency consistently associated with difficulty maintaining stable concentration [1], [5]. These cortical oscillations play distinct and fundamental roles in maintaining and regulating sustained attention. Consequently, any alteration in brain wave oscillations within neuronal networks, such as those cognitive performance, leading to effects ranging from memory impairment to significant behavioral impacts.

The neurofeedback and brain-computer interfaces (BCIs) have been successfully applied for attention training often using the ratio between theta (4-8 Hz) and beta (13-30 Hz) rhythms as an attentional biomarker [6]. It is worth noting that attention training using neurofeedback is expanding beyond clinical treatment; it is being applied to neurotypical individuals in various contexts to enhance cognitive performance in professional and sports activities that place high demands on attention [7]–[9]. These findings indicate that attention training interventions can serve not only as therapeutic tools but also as strategies for cognitive enhancement, offering alternatives to pharmacological treatments and their associated side effects [10].

More recently, brain-to-brain collaboration interfaces (BBCIs) have explored synchrony between two or more brains to potentialize performance in joint tasks, with applications ranging from collective decision making to the detection of affective states [11]–[13]. However, the application of BBCIs specifically directed at modulating brain rhythms in attentional cognitive networks in collaboration between individuals remains underexplored in the literature, which focuses mainly on motor and individual performance.

Additionally, integration of serious game into BCIs has proven to be an effective strategy to enhance user's engagement and adhesion during training [14]–[16], while there are emerging concerns regarding the security and integrity of transmitted brain data, since the wireless communication employed by these interfaces exposes them to risks of interception and tampering [17].

In this current work, we proposed to develop and validate an EEG-based brain-to-brain collaboration interface integrated into a serious game, aiming to teach two individuals to jointly modulate theta and beta rhythms in regions associated with the attention network, using performance of the participant with superior modulation as a reference and establishing an adaptive collaborative neurofeedback model between two users.

II. Materials and Methods

A. Overview of the brain-to-brain collaborative interface

EEG signals are acquired by two amplifiers V-Amp 16 (Brain Products), each connected to a cap with 16 positioned electrodes according to the international 10/20 system, covering: Fp1, Fp2, Fz, F3, F4, F7, F8, C3, Cz, C4, Pz, P3, P4, O1, Oz and O2. A desktop computer and a laptop run OpenViBE and Python scripts for the real-time acquisition, processing, classification and generation of the attentional biomarker. The serious game, developed in Unity, receives commands from both participants via MQTT protocol, which guarantee the continuous and secure transmission of attentional indices between the two processing environments.

The system operates in two sequential and uninterrupted stages. At the first moment, during 2 min, each participant controls a virtual car using only the computer keyboard, maintaining a base speed of 40 km/h. This period provides the calibration data for obtaining the attention-level classification models and the neurofeedback model-based in θ/β power ratio variations, usually represented by speed control. In the second stage, lasting 10 min, the car's both participants control their respective cars simultaneously, with speed modulated by both the keyboard and the attentional effort estimated from the θ/β biomarker, such that high sustained attention increases the vehicle's speed, while low attention reduces it. Furthermore, a trigger was implemented at the start and end of the game to temporally synchronize the EEG signals acquired by the two amplifiers, given the data acquisition had to begin before the game.

EEG data of 2 min in length are used to calibrate the classification system, dividing them into a set composed of adjacent 5s segments. This set is later randomly divided into a training (T) set (80% of trials) and testing (E) set (20% of trials) by a k-folds cross-trial validation ($k = 10$). A subset V from T is obtained for validation. Each segment is then divided into 1s windows with 50% overlap, which are pre-processed by a second order Butterworth band-pass filter between 0.1 and 50 Hz, followed by a Common Average Reference filter to remove physiological and non-physiological interferences.

The filtered windows are after transformed into the frequency domain using the Fast Fourier Transform, extracting feature vectors containing the normalized average powers for θ (4–7 Hz), α (8–12 Hz), low β (13–20 Hz), high β (20–30 Hz), and low γ (30–45 Hz) frequency bands from all electrodes.

A Linear Discriminant Analysis (LDA) classifier is then trained exclusively on T and validated on V across $k = 10$ cross-validation folds, retaining the model that yields the highest classification accuracy. For binary decision purposes, the moderate and high attention classes are combined into a single high-attention class.

In the game, the keyboard provides a constant base speed of 40 km/h, while θ/β modulation contributes an additional speed of up to 60 km/h, reaching a maximum of 100 km/h. Vectors classified as low attention force the additional speed contribution to 0 km/h.



Fig. 1. Experimental protocol used during the collaborative BCI session. (A) Participants performing the attention training task. (B) Timeline of the experimental session, including initial preparation (30 min), calibration phase (2 min), and neurofeedback training (10 min).

B. Experiment Protocol

This study was approved by the ethics committee (code number 85123524.6.0000.0129) of the Santos Dumont Institute. Pilot experiments were conducted with two healthy adult participants who were not taking psychiatric medications known to adversely affect cognitive function and had no history of depression, attention-deficit/hyperactivity disorder (ADHD), or other psychiatric or neurological disorders. Neither participant presented any motor impairment affecting the upper limbs. One participant had mild myopia and astigmatism; however, corrective eyeglasses were not worn during the experiment because they were incompatible with the EEG cap. They were recruited conveniently from the state of Rio Grande do Norte in Brazil. A male of 27 years old (Participant 1) and other female of 23 years old (Participant 2).

To acquire the brain signals and use the proposed system with serious game, the participant's scalp was cleaned with 70% alcohol, followed by the application of an abrasive paste all electrically conductive gel to ensure the skin-electrode impedance below 10 kΩ. This preparation stage delayed a total period of 30 min. Subjects comfortably seated in front of a computer screen displaying the serious game, and were instructed in a familiarization stage on the basic keyboard commands required to control their virtual car along the track. The first phase of the experiment EEG data to calibrate the brain-computer interface collected for a period of 2 min. During this stage, participants were oriented to focus their attention on the car while simultaneously controlling the virtual vehicle for a period of 2 min. Afterwards, the participants continued controlling their car for a period of 10 minutes.

C. Evaluation

The performance of the LDA classifier during the calibration phase was evaluated using standard machine learning metrics. Specifically, accuracy, precision, recall, and F1-score were calculated to assess model reliability. Furthermore, Cohen's Kappa coefficient was computed to measure the classification agreement beyond chance level, and confusion matrices were generated to visualize the correct predictions and misclassifications between the "LOW ATTENTION" and "HIGH ATTENTION" classes. During the 10-minute collaborative neurofeedback session, the temporal dynamics of the θ/β power ratio and the corresponding virtual car speeds were continuously recorded and analyzed for both participants. To evaluate the effectiveness of the neurofeedback control model in linking cognitive effort to game performance, a Pearson linear correlation was calculated between each participant's θ/β ratio and their respective vehicle speed. Finally, the spatiotemporal evolution of the attentional biomarker was visually assessed using topographic maps of the θ/β ratio. These maps were segmented into continuous 2-minute time windows to track dynamic changes in cortical engagement, particularly across frontopolar and central brain regions.

III. Results and Discussion

Figure 2 shows that both participants achieved metrics significantly above chance level (0.5), indicated by the red dashed line, in the calibration phase. The LDA classifier achieved mean accuracy of 0.86 (±0.075) and 0.82 (±0.079) for Participants 1 and 2, respectively, with F1-scores of 0.909 and 0.88, demonstrating reliable discrimination between low and high/moderate attention states. The Kappa coefficients of 0.64 and 0.54 indicate substantial and moderate agreement beyond chance, consistent with results reported in previous EEG-based attention classification studies [18]. Furthermore, the lower Kappa for Participant 2 and the larger deviations reflects the pilot nature of the study ($n = 2$, one calibration session), these values are expected to improve with additional data collection sessions and more calibration data.

The confusion matrix illustrates the actual correct predictions and errors between the "LOW ATTENTION" and "HIGH ATTENTION" (high/moderate attention) classes in the calibration phase. For Participant 1, the classifier correctly identified 100% of the LOW ATTENTION samples and 94% of the HIGH ATTENTION samples, with only 6% of the HIGH ATTENTION samples being misclassified as LOW. For participant 2, the classifier correctly identified 83% of the LOW ATTENTION samples and 94% of the HIGH ATTENTION samples. Misclassification consisted of 17% of the LOW ATTENTION samples being classified as HIGH ATTENTION and 6% of the HIGH ATTENTION samples being classified as LOW ATTENTION.

Figure 3 shows the performance of both participants driving the car using the calibrated system for a period of 10 min. Figure 3(a) shows that Participant 1 (winner) maintained a consistently low and stable θ/β power ratio (fluctuating between approximately 2 and 4) throughout the session. On the other hand, Participant 2 began the car control with θ/β power ratios reaching values up to 14 during the first 100s. Later, this participant managed reduced θ/β, stabilizing its

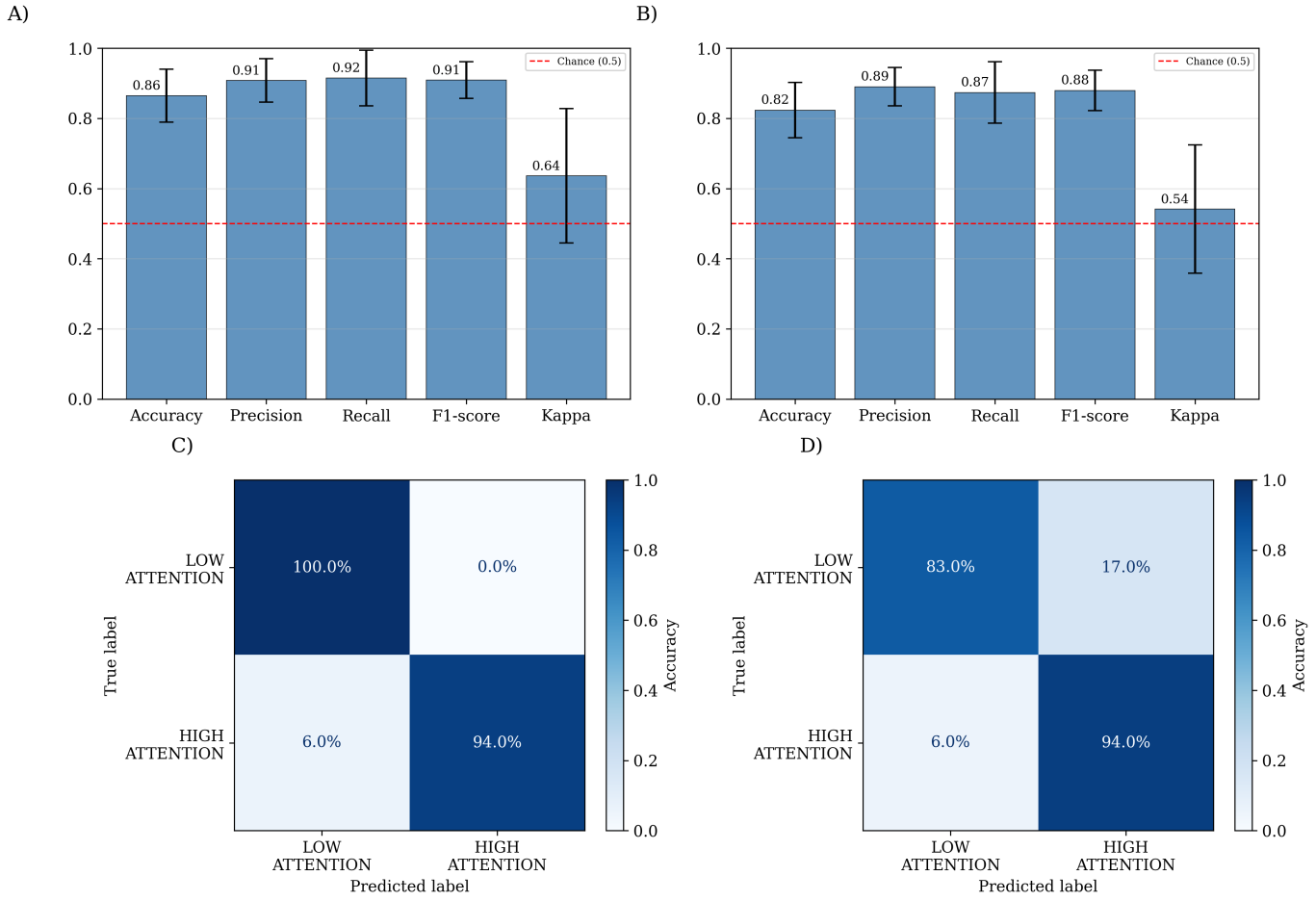


Fig. 2. The system performance during the calibration phase. A) and B) show the mean and standard deviation values for Participant 1 and Participant 2, respectively. C) and D) show confusion matrices corresponding to the best-performance for each Participant 1 and Participant 2, respectively

values much closer to those of the winner player during the last 5 min of the experiment.

Observing Figure 3(b), it is evident that the speed of both cars remained consistently above 40 km/h baseline default speed via keyboard, fluctuating around 60 km/h. The winner's average speed reached 63.2 km/h, while the follower's average speed was around 61.5 km/h. Furthermore, it can be noted that the follower's drops in attention at the start of the run resulted in speed drops, approaching 40 km/h, however the participant managed to recover their performance.

Figure 3(c) illustrates the relationship between the θ/β ratio and virtual car speed for both participants. Analyzing the results, Participant 2 showed a moderate negative correlation (r = -0.449, $p < 0.001$), indicating that lower θ/β ratios were associated with higher car speeds, consistent with the neurofeedback model design. Participant 1 (winner) presented a weak positive correlation (r = 0.208, $p < 0.001$), however, the winner's data were highly concentrated within a narrow θ/β range.

The asymmetric correlation observed among the participants reflects their distinct roles within the collaborative paradigm. The fact that Participant 2 exhibited a negative correlation confirms that the proposed system effectively linked θ/β modulation to car speed in real-time, rewarding attentional engagement with superior performance, consistent with previous findings in BCI-based attention training [19]. In contrast, the weak positive correlation observed in Participant 1 likely reflects the signal's range of variation: unlike Participant 2, whose θ/β ratio fluctuated widely ($\approx$ 2–14), Participant 1 remained within a narrow range ($\approx$ 2–4) for most of the time.

This scenario is consistent with the hypothesis that Participant 1 operated predominantly within a relatively stable attentional state, close to the optimal level for game performance, leaving little residual variance in the biomarker to explain variations in speed. Such variance may be more closely associated with non-attentional factors, such as collisions with obstacles along the course, rather than the attentional regulation mechanism itself. If Participant 1's collisions occurred during periods of higher θ/β ratios, these isolated events might have influenced the regression trend, reflecting specific course characteristics rather than the actual relationship between regulation of θ/β ratios and speed. Furthermore,even with this positive correlation, the winner did not lose engagement during the training.

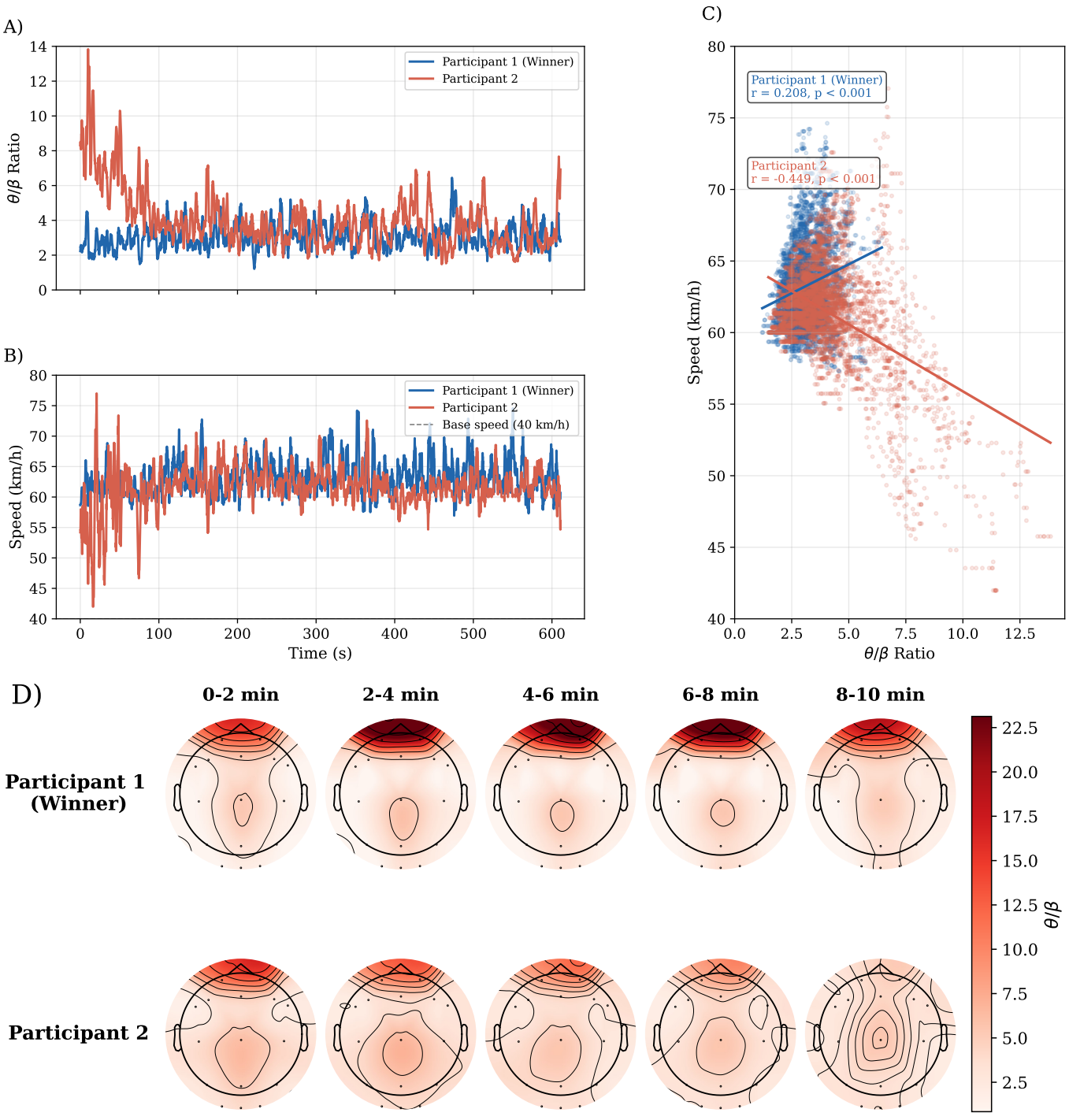


Fig. 3. Brain-to-brain collaboration dynamics during training session and topographic maps of the θ/β ratio. (a) θ/β ratio over time for the leader (Participant 1, blue) and follower (Participant 2, red), showing higher and more variable ratio values for the follower throughout the session. (b) Virtual car speed over time for both participants, computed as the sum of the 40 km/h keyboard base speed and the brain-modulated contribution. (c) Scatter plot showing the Pearson linear correlation between the θ/β ratio and speed. (d) Topographic maps illustrating the spatial distribution of the θ/β ratio segmented into continuous 2-minute time windows for both participants

These interpretations also align with findings indicating that the diagnostic validity and magnitude of the relationship between θ/β and attentional performance vary considerably across individuals and measurement protocols; this variability has led to heterogeneous and sometimes contradictory results in recent literature regarding this biomarker [20].

Taken together, these results suggest that the neurofeedback control model was particularly effective for the participant

with greater attentional variability. Studies with larger samples are needed to determine whether the positive correlation observed in Participant 1 represents a consistent individual pattern or a finding specific to this session.

Figure 3(d) presents the spaciotemporal evolution of the θ/β ratio for both participants throughout the BCI paradigm, segmented into continuous 2 minutes-windows, Visual analysis reveals distinct and dynamic topographic patterns between the winner (Participant 1) and Participant 2. Participant 1 exhibited persistently high θ/β ratio activity localized sharply in the frontopolar region (channels Fp1 and Fp2) throughout the 10-min task, In contrast,Participant 2 demonstrated a globally diffuse spatial distribution with a low θ/β ratio from the start of the task (0–8 minutes). However, starting from the 8–10 minute window, a generalized and progressive increase in relative power was observed in central and parietal areas. We noted that both participants exhibited a θ/β reduction over a large portion of the scalp during the BCI session, particularly across fronto-central regions. This behavior is consistent with a lower relative predominance of theta activity compared to low-beta activity during the neurofeedback task.

Besides, the reduction in the θ/β ratio observed suggests that the proposed neurofeedback paradigm promoted the expected modulation of brain activity associated with attentional processes, because the θ/β ratio reflects the balance between slower theta oscillations and faster low-beta activity, lower values indicate a shift toward a pattern commonly associated with greater cortical engagement and sustained attention. These observed changes in theta and beta oscillations are consistent with previous studies reporting an inverse relationship between the frontal θ/β ratio and attentional performance [21]. Furthermore, these findings align with research identifying cognitive improvements in individuals with attention deficit who reduced their θ/β power ratios following neurofeedback or BCI-based interventions [22]–[24].

Although these findings are encouraging and indicate that the proposed approach was able to induce the intended EEG modulation, they should be interpreted with caution. The present analysis is based on qualitative topographic observations from only two participants, further studies with larger samples and statistical analyses are required to confirm the consistency and significance of these findings.

IV. Conclusion

This pilot study demonstrated the technical feasibility of an EEG-based brain-to-brain collaborative interface for sustained attention training. The proposed adaptive neurofeedback paradigm effectively translated the user's cognitive modulation into real-time game control, allowing a high-performing reference participant to guide the training of a follower. The observed reductions in the θ/β power ratio during the collaborative task suggest an active shift toward greater cortical engagement, confirming the hypothesis that a higher-performing reference participant can effectively guide the brain wave modulation of a collaborative partner. While these preliminary findings support the potential of collaborative interface for cognitive enhancement, they are limited by the small sample size. Future work will implement this validated experimental protocol in a larger cohort, incorporating robust statistical validations, topographic connectivity analyses, and standardized cognitive assessments to comprehensively determine the system's clinical and practical efficacy.

Acknowledgment

The authors would like to thank the two volunteers who participated in our study.

References

[1] M. S. Clayton, N. Yeung, and R. Cohen Kadosh, "The roles of cortical oscillations in sustained attention," *Trends in Cognitive Sciences*, vol. 19, no. 4, pp. 188–195, 2015.

[2] M. Sarter, B. Givens, and J. P. Bruno, "The cognitive neuroscience of sustained attention: where top-down meets bottom-up," *Brain Research Reviews*, vol. 35, no. 2, pp. 146–160, 2001.

[3] A. L. Proskovec, E. Heinrichs-Graham, A. I. Wiesman, T. J. McDermott, and T. W. Wilson, "Oscillatory dynamics in the dorsal and ventral attention networks during the reorienting of attention," *Human Brain Mapping*, vol. 39, no. 5, pp. 2177–2190, 2018.

[4] W. D. Casagrande, E. M. Nakamura-Palacios, and A. Frizera-Neto, "Electroencephalography Neurofeedback Training with Focus on the State of Attention: An Investigation Using Source Localization and Effective Connectivity," *Sensors*, vol. 24, no. 18, p. 6056, 2024.

[5] S. Siuly, Y. Li, and Y. Zhang, *EEG Signal Analysis and Classification: Techniques and Applications*, 2016.

[6] H. Chiu, C.-K. Sun, H.-Y. Fan, R. Tzang, M.-Y. Wang, Y.-C. Cheng, Y.-S. Cheng, P.-Y. Yeh, and W. Chung, "Surface electroencephalographic neurofeedback improves sustained attention in ADHD: a meta-analysis of randomized controlled trials," *Child and Adolescent Psychiatry and Mental Health*, vol. 16, no. 1, p. 104, 2022.

[7] T.-T. Chen, K.-P. Wang, W.-H. Chang, C.-W. Kao, and T.-M. Hung, "Effects of the function-specific instruction approach to neurofeedback training on frontal midline theta waves and golf putting performance," *Psychology of Sport and Exercise*, vol. 61, p. 102211, 2022.

[8] K.-P. Wang, C. Frank, T.-M. Hung, and T. Schack, "Neurofeedback training: Decreases in Mu rhythm lead to improved motor performance in complex visuomotor skills," *Current Psychology*, vol. 42, no. 24, pp. 20 860–20 871, 2023.

[9] I. Alreshidi, I. Moulitsas, and K. W. Jenkins, "Multimodal Approach for Pilot Mental State Detection Based on EEG," *Sensors*, vol. 23, no. 17, p. 7350, 2023.

[10] S. Chaplin, "Attention deficit hyperactivity disorder: diagnosis and management," *Progress in Neurology and Psychiatry*, vol. 22, no. 3, pp. 27–29, 2018.

[11] S. Bhattacharyya, D. Valeriani, C. Cinel, L. Citi, and R. Poli, "Anytime collaborative brain–computer interfaces for enhancing perceptual group decision-making," *Scientific Reports*, vol. 11, no. 1, p. 17008, 2021.

[12] T. Jia, H. Long, L. Ji, and X. Guan, "EEG-based Spatial-Channel Interaction Attention Neural Networks for Detecting Empathy in Motor Collaboration," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–11, 2025.

[13] Y. Si, Z. Wang, G. Xu, Z. Wang, T. Xu, T. Zhou, and H. Hu, "Group-member selection for RSVP-based collaborative brain-computer interfaces," *Frontiers in Neuroscience*, vol. 18, p. 1402154, 2024.

[14] D. Marshall, D. Coyle, S. Wilson, and M. Callaghan, "Games, gameplay, and BCI: The state of the art," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 5, no. 2, pp. 82–99, 2013.

[15] R. Cózar and J.-M. Sáez-López, "Game-based learning and gamification in initial teacher training in the social sciences: an experiment with minecraftedu," *International Journal of Educational Technology in Higher Education*, vol. 13, 2016.

[16] M. A. L. Nicolelis, E. J. L. Alho, A. R. C. Donati, S. Yonamine, M. A. Aratanha, G. Bao, D. S. F. Campos, S. Almeida, D. Fischer, and S. Shokur, "Training with noninvasive brain–machine interface, tactile feedback, and locomotion to enhance neurological recovery in individuals with complete paraplegia: a randomized pilot study," *Scientific Reports*, vol. 12, no. 1, p. 20545, 2022.

[17] S. Rajesh, V. Paul, V. G. Menon, S. Jacob, and P. Vinod, "Secure brain-to-brain communication with edge computing for assisting post-stroke paralyzed patients," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2531–2538, 2020.

[18] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, Jun. 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aab2f2

[19] D. Delisle-Rodriguez, H. L. De Oliveira, J. C. Da Silva, M. L. De Souza, T. Bastos, E. M. Nakamura-Palacios, and A. Frizera-Neto, "Multi-channel EEG-based BCI using regression and classification methods for attention training by serious game," *Biomedical Signal Processing and Control*, vol. 85, p. 104937, 2023.

[20] M. Arns, C. K. Conners, and H. C. Kraemer, "A decade of eeg theta/beta ratio research in adhd: A meta-analysis," *Journal of Attention Disorders*, vol. 17, no. 5, pp. 374–383, 2013.

[21] A. Angelidis, W. van der Does, L. Schakel, and P. Putman, "Frontal eeg theta/beta ratio as an electrophysiological marker for attentional control and its test-retest reliability," *Biological Psychology*, vol. 121, pp. 49–52, 2016.

[22] X. Qian, B. R. Y. Loo, F. X. Castellanos, S. Liu, H. L. Koh, X. W. W. Poh, R. Krishnan, D. Fung, M. W. L. Chee, C. Guan, T.-S. Lee, C. G. Lim, and J. Zhou, "Brain-computer-interface-based intervention re-normalizes brain functional network topology in children with attention deficit/hyperactivity disorder," *Translational Psychiatry*, vol. 8, no. 1, p. 149, 2018.

[23] C. G. Lim, X. W. W. Poh, S. S. D. Fung, C. Guan, D. Bautista, Y. B. Cheung, H. Zhang, S. N. Yeo, R. Krishnan, and T. S. Lee, "A randomized controlled trial of a brain-computer interface based attention training program for adhd," *PLOS ONE*, vol. 14, no. 5, p. e0216225, 2019.

[24] M. Z. Raza, M. Omais, H. M. E. Arshad, M. Maqsood, and A. A. Nadeem, "Effectiveness of brain-computer interface (bci)-based attention training game system for symptom reduction, behavioral enhancement, and brain function modulation in children with adhd: A systematic review and single-arm meta-analysis," *NeuroRegulation*, vol. 12, no. 1, pp. 51–78, 2025.

An Integrated Open-Source Platform for Robot-Assisted Personalized Transcranial Magnetic Stimulation

1st Renan H. Matsuda
IIN-ELS, Santos Dumont Institute
Macaíba, Brazil
NBE, Aalto University
Espoo, Finland
renan.matsuda@isd.org.br

2nd Victor H. Souza
Cortisys Oy
Espoo, Finland
victor.souza@cortisys.fi

3rd Thais C. Marchetti
DF, FFCLRP, University of São Paulo
Ribeirão Preto, Brazil
thaismarchetti@usp.br

4th Márcio A. de Campos Jr.
DF, FFCLRP, University of São Paulo
Ribeirão Preto, Brazil
marcio.campos6@usp.br

5th Lucas S. Betioli
DF, FFCLRP, University of São Paulo
Ribeirão Preto, Brazil
lucasantoniobetioli@usp.br

6th Ana M. Soto
NBE, Aalto University
Espoo, Finland
ana.sotodelacruz@aalto.fi

7th Paulo H. J. Amorim
DITPS, CTI Renato Archer
Campinas, Brazil
paulo.amorim@cti.gov.br

8th Thiago F. Moraes
DITPS, CTI Renato Archer
Campinas, Brazil
totonixsame@gmail.com

9th Oswaldo Baffa
DF, FFCLRP, University of São Paulo
Ribeirão Preto, Brazil
baffa@usp.br

***Abstract*—Neuronavigation enables subject-specific targeting for transcranial magnetic stimulation (TMS), while advances in brain parcellation, connectivity analysis, electric field modeling, robotics, and neurophysiological monitoring have expanded opportunities for personalized neuromodulation. However, these components are commonly distributed across independent software environments. This work presents an open-source, modular platform based on InVesalius that connects multiple stages of personalized image-guided TMS within a common neuronavigation environment. The platform integrates automated brain subparcellation, multimodal connectivity visualization, electric field estimate visualization, neuronavigation, communication with external robotic control software, dual-coil navigation, and motor mapping visualization. Structural MRI-based subparcellation enables cortical and subcortical regions to be used for stimulation planning, while diffusion MRI tractography and fMRI support connectivity-informed target selection. Electric field estimates generated by specialized modeling tools can be incorporated as an additional planning layer. Imaging-derived targets can be communicated to the external *tms-robot-control* framework for robot-assisted coil positioning, including dual-robot workflows. Motor evoked potential visualization further supports construction of motor maps during TMS sessions. The main contribution is the integration of these heterogeneous components into an extensible workflow rather than the independent development or validation of each underlying method. Quantitative benchmarking of the complete platform remains necessary to establish its accuracy, efficiency, usability, and reproducibility relative to existing workflows.**

***Index Terms*—Transcranial magnetic stimulation; neuronavigation; brain parcellation; diffusion MRI; functional MRI; structural connectivity; functional connectivity; robotics; motor mapping; open-source software.**

I. Introduction

Transcranial magnetic stimulation (TMS) has emerged as a powerful non-invasive technique for investigating brain function and modulating neural activity in both research and clinical settings. Applications of TMS range from functional brain mapping and studies of cortical excitability to the treatment of neurological and psychiatric disorders [1], [2]. As the field advances toward personalized neuromodulation, there is growing recognition that stimulation outcomes depend not only on the stimulation parameters themselves but also on individual variations in brain anatomy, connectivity, and network organization [3]. Consequently, the development of image-guided approaches capable of identifying subject-specific stimulation targets has become a major focus of contemporary TMS research.

Neuronavigation has become an essential component of modern TMS systems by enabling stimulation targets to be localized directly on subject-specific anatomical images [4]. Unlike conventional approaches based on scalp measurements or anatomical landmarks, neuronavigation uses three-dimensional reconstructions derived from magnetic resonance imaging (MRI), together with tracking systems, to continu-

FAPESP grants 2013/07699-0 and 2025/26818-7 (OB), 2022/14526-3 (RHM), 2024/07674-1 (TCM), 2025/06042-4 (MC). CAPES, grant no. 88887.197686/2025-00. CNPq, grant no. 408417/2024-2. Jenny and Antti Wihuri Foundation grant no. 00250241 (RHM).

ously estimate the position and orientation of the stimulation coil relative to the individual's brain anatomy. This capability improves targeting accuracy, reduces inter-operator variability, and facilitates reproducibility across sessions and subjects [5], [6].

Motor mapping is another important application of navigated TMS, allowing the characterization of corticospinal excitability and the spatial organization of motor representations. By recording motor evoked potentials (MEPs) elicited by stimulation at different cortical locations, TMS can be used to identify motor hotspots, estimate cortical excitability, and construct spatial maps of motor responses. When combined with neuronavigation, motor mapping enables MEP responses to be interpreted in relation to subject-specific cortical anatomy, supporting more accurate functional localization and reproducible assessment of stimulation effects [7], [8].

Recent developments in neuroimaging have expanded the role of neuronavigation beyond anatomical guidance. Automated brain parcellation [9], diffusion MRI tractography, functional MRI (fMRI), and network-based analyses now enable stimulation targets to be defined according to individual anatomical, structural connectivity, and functional connectivity profiles [10], [11]. Diffusion MRI tractography can be used to estimate white matter pathways and structural connections between brain regions, including in real-time tractography-assisted neuronavigation workflows for TMS [12]. In contrast, fMRI can be used to identify task-related activation patterns or resting-state functional connectivity networks [13], [14]. These advances create opportunities for personalized neuromodulation strategies in which stimulation is tailored to subject-specific brain organization.

Electric field modeling provides an additional layer for personalized TMS planning by estimating how coil position, coil orientation, head anatomy, and tissue properties influence the spatial distribution and magnitude of the induced cortical field. This information can complement anatomical and connectivity-informed targeting by providing an estimate of the stimulation dose delivered to cortical tissue. Recent work from our group has investigated electric field modeling for TMS, supporting its use as part of personalized stimulation planning and interpretation [15].

At the same time, robotic technologies have emerged as a promising solution for improving the accuracy and reproducibility of neuronavigated TMS [16]. While conventional neuronavigation systems provide visual guidance to the operator, robotic systems can actively position and maintain the stimulation coil at predefined targets, supporting compensation for subject motion and reducing operator-dependent variability [17]. Our recent robotic and electronic targeting approaches have demonstrated the feasibility of autonomous and accurate TMS targeting, including robotic–electronic platforms for autonomous transcranial magnetic stimulation targeting and precise multi-locus TMS stimulation workflows [17], [18]. The combination of neuronavigation and robot-assisted positioning enables imaging-derived targets to be translated into more reproducible physical stimulation procedures. Dual-site TMS further extends this concept by enabling two cortical targets to be stimulated or monitored within the same session, supporting the investigation of interactions between distributed brain regions [19], [20].

Despite these advances, personalized TMS workflows remain fragmented across software for image processing, connectivity analysis, electric field modeling, neuronavigation, robotic control, and neurophysiological assessment, often requiring manual data exchange. To address this limitation, we extended the open-source InVesalius environment [4], [21] to connect these stages within a modular workflow. The main contribution of this work is not the independent introduction of each underlying method, several of which rely on established or previously reported tools, but their integration with image-guided targeting and neuronavigation. Specifically, the platform incorporates FastSurfer-based subparcellation, visualization of multimodal imaging-derived information, communication with external robotic control, dual-coil navigation, dual-robot support, and motor-map visualization.

II. System Overview

The proposed platform extends InVesalius, an open-source medical imaging and neuronavigation environment, to connect anatomical reconstruction, brain subparcellation, multimodal data visualization, electric field estimates, neuronavigation, motor mapping, and external robotic control. These components have different levels of integration: some are implemented directly in InVesalius, whereas others rely on externally generated data or specialized control software. The platform also supports dual-site workflows involving two tracked coils and two robotic systems. The integration follows a modular data-flow architecture rather than a monolithic processing pipeline. Subject-specific anatomical data provide the common spatial reference in InVesalius, to which externally processed imaging information and stimulation targets can be registered. Neuronavigation then relates these image-space targets to tracked physical coordinates, while robot commands are exchanged with the external control framework. This separation allows specialized processing and control tools to remain independent while their outputs are incorporated into the navigation workflow.

The workflow starts with subject-specific structural MRI data, which are used for anatomical reconstruction and brain subparcellation. Segmented cortical and subcortical regions can be converted into masks or three-dimensional surfaces for visualization, target planning, and neuronavigation. Diffusion MRI tractography and fMRI can be incorporated to support structural and functional connectivity-informed targeting, while externally generated electric field estimates can provide complementary information about the induced cortical field for candidate coil positions and orientations [15].

After target definition, the neuronavigation module monitors the position and orientation of the TMS coil relative to the subject-specific anatomy. In dual-site mode, Fig. 1, two coils and two targets can be managed concurrently, with independent coil identification, target association, and

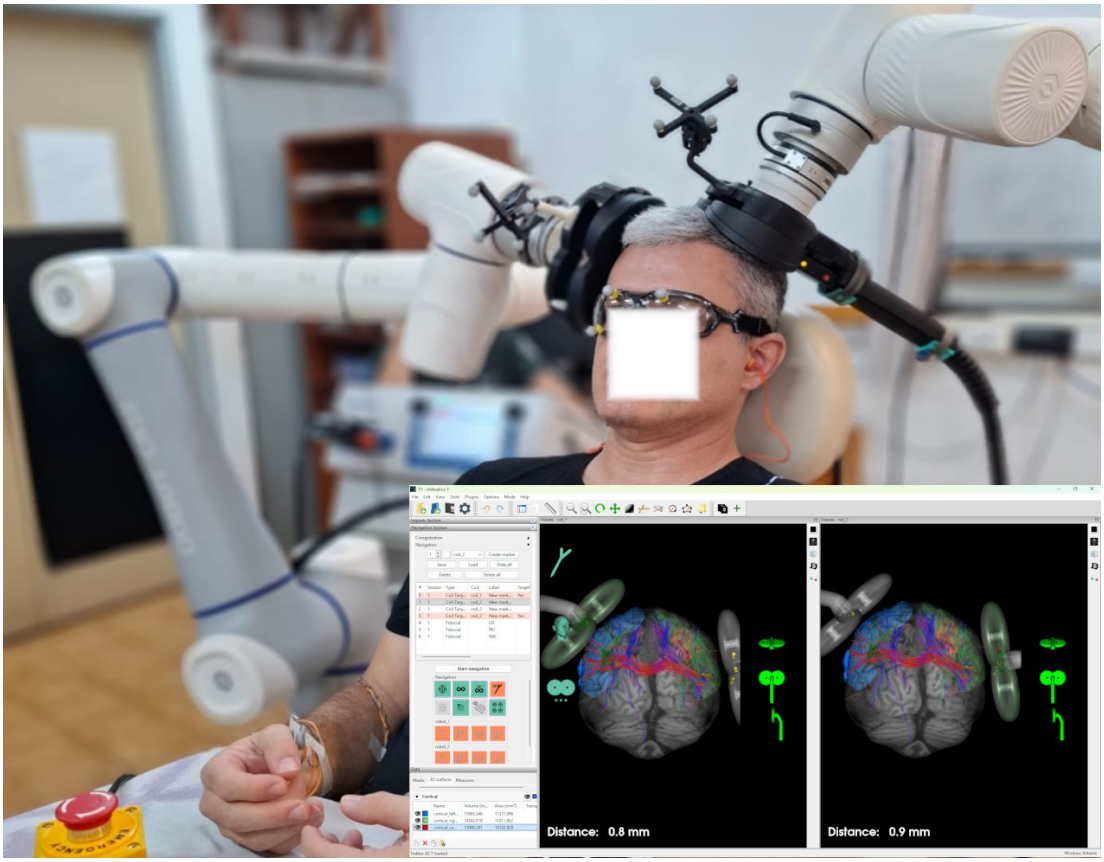


Fig. 1. InVesalius: integrated open-source platform for personalized TMS. The setup illustrates dual-site robot-assisted coil positioning, while the InVesalius interface supports image-guided target planning, brain subparcellation, tractography visualization, and navigation monitoring within a modular workflow.

navigation views. Robot-assisted positioning is implemented through communication with the external *tms-robot-control* framework, which handles robot communication and motion control while InVesalius remains responsible for image-based planning, visualization, tracking, and neuronavigation.

During stimulation sessions, motor evoked potential values can be projected onto reconstructed brain surfaces as heatmaps. Together, these components connect anatomical modeling, connectivity-guided targeting, electric field-informed planning, single- and dual-site neuronavigation, robot-assisted positioning, and motor response visualization.

III. Brain Subparcellation Module

A central component of the platform is the automated brain subparcellation module. This module was developed to segment anatomical subregions of the brain, including cortical and subcortical structures such as the precuneus, superior frontal cortex, hippocampus, and other regions commonly used in neuroimaging-based analysis and TMS planning.

The segmentation pipeline is based on FastSurfer, an open-source convolutional neural network framework for rapid brain segmentation and cortical parcellation [9]. To improve portability and facilitate integration into InVesalius, the module supports inference using both PyTorch and ONNX-compatible models. The FastSurfer models were converted to ONNX primarily to enable lightweight inference backends. A systematic quantitative equivalence analysis between the original PyTorch and converted ONNX models was not performed in the present study and remains necessary to characterize possible backend-dependent differences. The pretrained weights are available at https://github.com/invesalius/weights.

The implemented pipeline includes the following steps:

1) loading the structural MRI volume from the active InVesalius project;
2) applying the required image conformation and preprocessing operations;
3) performing neural network inference using PyTorch or ONNX-compatible execution;
4) post-processing the labelled segmentation volume;
5) mapping numerical labels to anatomical region names;
6) generating binary masks from selected labelled regions;
7) organizing the generated masks into anatomical categories.

The generated masks are automatically categorized in the InVesalius data panel. Cortical parcellation masks are grouped under a cortical category, while subcortical and other anatomical regions are organized into their respective categories. This organization improves usability, particularly when working

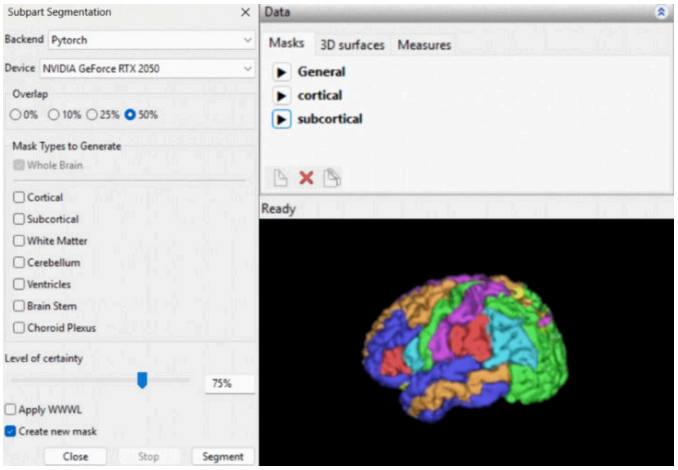


Fig. 2. Brain subparcellation module integrated into InVesalius. The interface allows the user to select the inference backend, processing device, mask types, overlap level, and segmentation confidence threshold. Generated masks are organized into anatomical categories, such as general, cortical, and subcortical regions, and can be converted into three-dimensional surfaces for visualization and TMS target planning.

with a large number of segmented brain regions.

Additional interface tools were implemented to support visualization and interaction. Category headers in the surface panel include hide/show and select-all controls, allowing users to manage multiple structures efficiently. A batch surface generation feature enables users to create three-dimensional surfaces from multiple selected masks simultaneously. These features allow anatomical subregions to be rapidly transformed into objects that can be used for visualization, target definition, and neuronavigation. Fig. 2 illustrates the brain subparcellation module, which generates anatomically categorized masks and surfaces from structural MRI.

IV. Multimodal Connectivity Visualization

The platform supports visualization of externally processed diffusion MRI tractography and functional MRI data rather than performing the underlying diffusion or fMRI processing pipelines. Diffusion MRI tractography provides information about structural connectivity by estimating white matter pathways between brain regions. Recent work has demonstrated the feasibility of real-time tractography-assisted neuronavigation for TMS [12]. Within InVesalius, these data can be visualized together with anatomical models and neuronavigation information to support connectivity-informed targeting. Fig. 3 illustrates this integration.

Functional MRI data can also be incorporated into the targeting workflow. Task-based fMRI activation maps may be used to identify cortical areas involved in specific motor, cognitive, sensory, or affective processes. Resting-state fMRI can be used to estimate functional connectivity between brain regions, supporting network-informed stimulation planning, as illustrated in Fig. 4. For example, a cortical target may be selected based on its functional relationship with a distributed network or with a clinically relevant region [22].

Within the integrated workflow, brain parcellation, structural connectivity, and functional connectivity are complementary. Parcellation provides anatomically defined regions of interest, diffusion MRI provides information about white matter organization, and fMRI provides information about functional activation or network coupling. The combined visualization of these data may assist in selecting cortical targets associated with specific anatomical structures, structural pathways, or functional networks.

V. Integration of Electric Field Modeling and Visualization

Electric field modeling provides subject-specific estimates of the spatial distribution and magnitude of the cortical field induced by TMS as a function of coil configuration, head anatomy, and tissue properties. In the present architecture, the underlying electric field simulations are not computed by InVesalius; estimates generated using specialized modeling tools can instead be incorporated and visualized together with subject-specific anatomy, as illustrated in Fig. 5. This integration provides a complementary biophysical layer for evaluating candidate targets and coil configurations alongside anatomical and connectivity information. Recent work from our group describes the underlying electric-field neuronavigation methodology [15]; the contribution considered here concerns its integration and visualization within the broader workflow.

VI. Robotic Neuronavigation and Robot Control

The neuronavigation component translates imaging-derived targets into physical stimulation procedures. Targets selected from anatomical models, subparcellation masks, structural connectivity maps, or fMRI-derived functional regions can be used to guide coil placement. Robot-assisted TMS is supported through communication with an external robotic control framework rather than by embedding low-level robot control directly inside InVesalius. In this architecture, InVesalius is responsible for image-based target definition, anatomical visualization, tracking, and neuronavigation, while the external control software handles robot communication, motion control, and automated coil positioning. The framework currently supports the Dobot CR10 (6-DoF, 10-kg payload), Elfin E5 (6-DoF, 5-kg payload), and Universal Robots UR10 (6-DoF, 10-kg payload), providing a common interface for robot-assisted TMS across different hardware platforms. The robotic control component used in this work is the open-source *tms-robot-control* framework developed by our group and available at https://github.com/biomaglab/tms-robot-control. Communication between InVesalius and *tms-robot-control* allows imaging-derived targets and tracking information to be translated into robot-assisted positioning commands while preserving modularity between image-guided planning and low-level robot control.

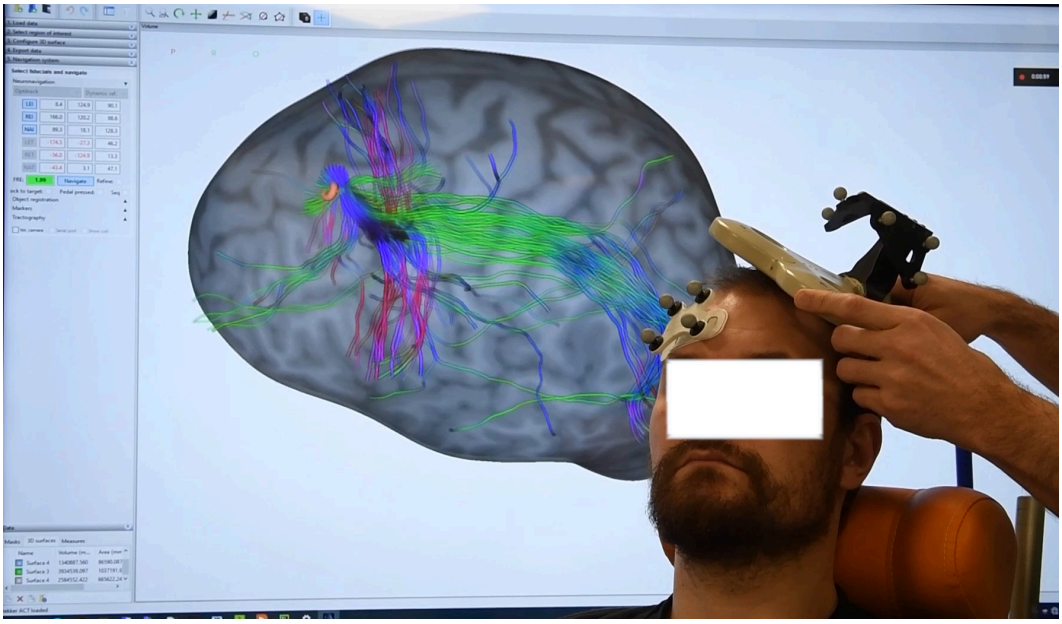


Fig. 3. Tractography-assisted neuronavigation for transcranial magnetic stimulation. Diffusion MRI-derived white matter streamlines are visualized together with the subject-specific cortical surface, while the tracked TMS coil is positioned over the participant's scalp. This visualization allows the operator to inspect the relationship between the stimulation site, cortical anatomy, and underlying structural connectivity during image-guided TMS procedures. Adapted from [12].

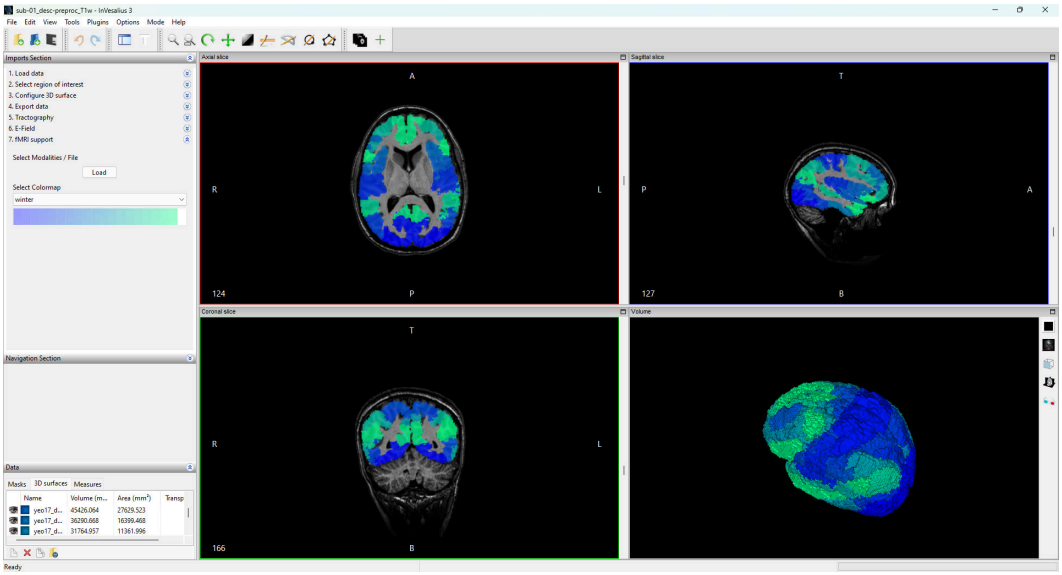


Fig. 4. Multimodal visualization of fMRI-derived information in InVesalius. Functional maps are overlaid on structural MRI slices in axial, sagittal, and coronal views, while the corresponding three-dimensional cortical representation is shown in the volume view. This visualization supports inspection of spatial relationships between anatomical structures and functional or connectivity-related information for image-guided TMS planning.

VII. Dual-Coil Navigation and Dual-Robot Control

In addition to single-coil neuronavigation, the platform has been extended to support simultaneous navigation of two TMS coils for dual-site stimulation protocols, Fig. 1. This development was motivated by the need to investigate and modulate distributed brain networks rather than isolated cortical regions [19]. While conventional TMS workflows typically focus on one stimulation target at a time, dual-site TMS enables the exploration of interactions between two brain areas, supporting studies of effective connectivity, network dynamics, and multi-region neuromodulation strategies [20].

The dual-coil navigation module adapts the InVesalius navigation architecture to manage two stimulation coils and two stimulation targets concurrently. Each coil can be independently identified, named, registered, and associated with a specific target. Coil naming can be performed through the

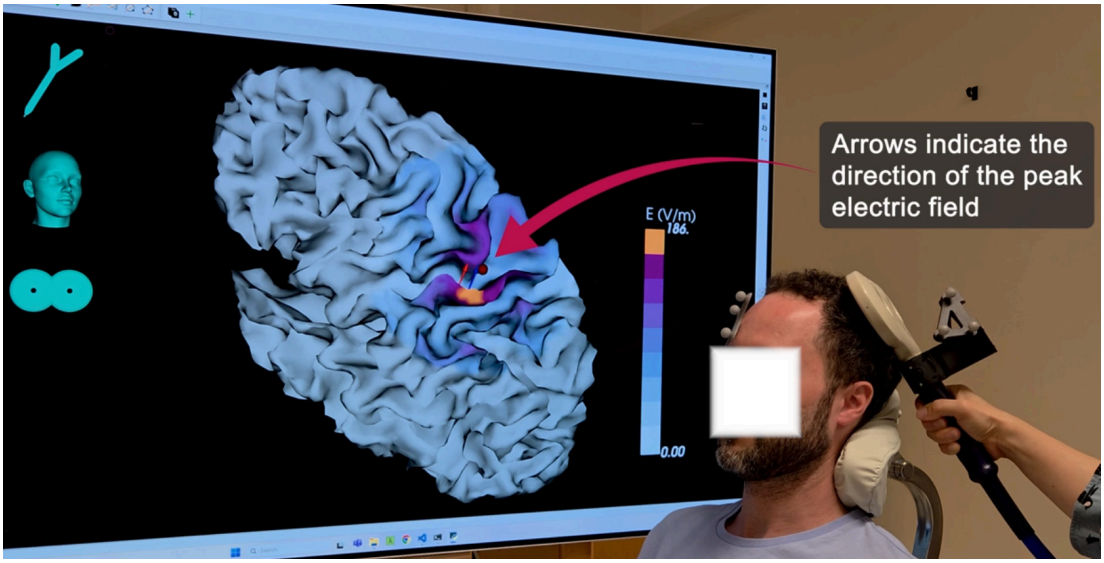


Fig. 5. Electric field modeling for personalized TMS planning. The predicted electric field distribution is visualized over subject-specific cortical anatomy, allowing candidate stimulation sites and coil configurations to be evaluated according to the estimated spatial distribution and magnitude of the induced cortical field. This information can complement anatomical, structural connectivity, and functional connectivity-based target selection. Adapted from [15].

coil registration window or through the coil selection panel, improving clarity when multiple tracked coils are present. The navigation panel was also extended to allow target creation to be associated with a selected coil, and the marker list displays the relationship between each coil and its corresponding target. Simultaneous navigation is activated through a dedicated multiple-target control in the navigation panel. When enabled, the interface provides two independent navigation scenes, each associated with a single coil. This allows the user to visualize and monitor the position and orientation of both coils relative to their respective anatomical targets. The display of tracking information was adapted to represent each coil with independent visual indicators, facilitating interpretation during dual-coil procedures. For dual-robot operation, two independent instances of the external robot-control framework communicate with InVesalius through a local server. A message-splitting mechanism routes commands and tracking information according to the robot identifier, allowing each robot-coil assembly to be controlled independently. The preferences window associates each robot with a specific coil. A preliminary collision-avoidance strategy approximates each coil as a sphere and exponentially reduces robot speed as the inter-coil distance approaches a predefined threshold. This mechanism should be regarded as a preliminary software safeguard rather than a validated safety system. Dual-robot synchronization, simultaneous positioning accuracy, communication latency, and behavior under head motion were not quantitatively evaluated in the present study and require dedicated validation. This extension enables manual and robot-assisted dual-site TMS workflows within the InVesalius-based neuronavigation environment.

VIII. Motor Mapping Visualization

The motor mapping module provides visualization and spatial association of MEP amplitudes. Each stimulation point is associated with its individual MEP value and corresponding stimulation coordinate. The motor map is generated using distance-based Gaussian interpolation, such that the influence of each measured response decreases with increasing distance from its stimulation location. The Gaussian kernel width therefore defines the spatial scale of the interpolation. Users can adjust the selected surface, colormap, display range, and Gaussian interpolation parameters, and the MEP values and visualization settings can be stored with the session.

IX. Discussion

The proposed platform addresses the fragmentation of tools used in personalized and image-guided TMS. Structural MRI processing, parcellation, connectivity analysis, electric field modeling, neuronavigation, robotic positioning, and neurophysiological analysis are commonly distributed across specialized environments. FreeSurfer [23], SimNIBS [24], commercial neuronavigation systems such as Localite TMS Navigator, and motor mapping tools such as TMSMap [25] provide established solutions for individual stages. The proposed platform does not aim to replace these specialized tools. Its principal contribution is an open and modular architecture that brings their outputs or complementary functions closer to image-guided targeting, navigation, robotic positioning, and response visualization within the InVesalius ecosystem.

Importantly, integration should not be interpreted as independent implementation or validation of every component. Brain subparcellation is incorporated through FastSurfer-based inference; tractography and fMRI information may originate from external processing; electric-field estimates rely on specialized modeling methods; and low-level robot control is handled by the external *tms-robot-control* framework. In this context, interoperability refers to modular exchange of imaging-derived information, targets, tracking data, and

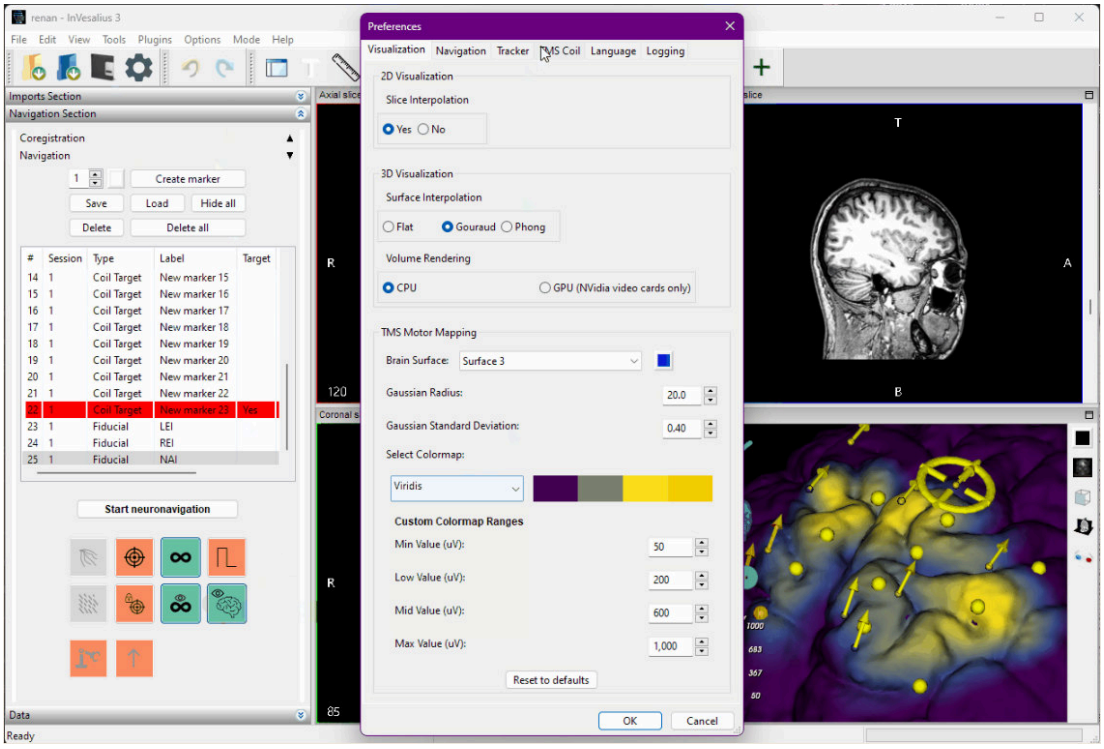


Fig. 6. Motor mapping visualization in the InVesalius-based platform. Motor evoked potential values are projected onto the reconstructed brain surface as a heatmap, allowing spatial exploration of corticospinal responses during TMS sessions.

control messages between supported components, rather than compliance with a formal clinical interoperability standard or demonstrated compatibility with arbitrary hardware.

A major limitation of the present study is the absence of a systematic quantitative evaluation of the complete platform. We did not perform a head-to-head comparison with commercial neuronavigation systems or conventional fragmented workflows, nor quantify end-to-end registration error, positioning accuracy, communication latency, failure rates, usability, workflow efficiency, or session-to-session reproducibility. Consequently, the present results demonstrate software integration and technical feasibility but do not establish superior accuracy, efficiency, reproducibility, usability, or safety relative to existing approaches.

The development process also highlighted practical lessons and research gaps. Maintaining specialized external components can preserve modularity, but heterogeneous imaging, tracking, robotic, and physiological data require consistent coordinate systems and interfaces. Dual-device operation further introduces synchronization and safety requirements beyond single-coil navigation. More broadly, multimodal visualization makes complementary information accessible within a common environment but does not by itself demonstrate improved targeting or clinical efficacy. Future work should therefore benchmark navigation and robotic accuracy, latency, reproducibility, ONNX segmentation equivalence, usability, and workflow efficiency; improve collision avoidance; and evaluate compatibility across additional hardware and multicenter settings. Clinical adoption will additionally require hardware-specific safety validation, robust error handling, standardized deployment, user training, and compliance with applicable medical-device requirements. The open-source nature of the platform facilitates inspection, modification, and extension of the software according to experimental needs. Future extensions will also investigate incorporation of TMS-evoked EEG responses, allowing neurophysiological information beyond motor responses to be explored within the same architecture.

X. Conclusion

We presented an open-source and modular extension of the InVesalius ecosystem aimed at connecting multiple stages of personalized image-guided TMS. The platform combines anatomical reconstruction and subparcellation with multimodal imaging visualization, electric-field estimates, neuronavigation, external robotic control, dual-coil and dual-robot workflows, and motor-response visualization. Its principal contribution is the integration of these heterogeneous components within a common image-guided environment rather than the independent development of all underlying methods.

The current implementation demonstrates the architecture and technical feasibility of this approach but does not establish improvements in accuracy, reproducibility, efficiency, usability, or safety relative to existing systems. Quantitative benchmarking, formal safety assessment of multi-robot operation, broader hardware evaluation, user studies, and multicenter validation remain necessary to establish the performance and translational value of the platform.

Code Availability

The platform is developed as part of the open-source InVesalius project and is available at https://github.com/invesalius/invesalius3. The robotic control framework is available at https://github.com/biomaglab/tms-robot-control.

Acknowledgment

The authors acknowledge the contributions of the InVesalius open-source community and the Google Summer of Code contributors (Shuvam Pal, Mohamed Gaafar, and Michael Chan) involved in the development of the brain subparcellation, motor mapping and functional MRI visualization modules.

References

[1] M. Hallett, "Transcranial magnetic stimulation and the human brain." *Nature*, vol. 406, no. 6792, pp. 147–150, 2000.

[2] J.-P. Lefaucheur, N. André-Obadia, A. Antal, S. S. Ayache, C. Baeken, D. H. Benninger, R. M. Cantello, M. Cincotta, M. de Carvalho, D. De Ridder, H. Devanne, V. Di Lazzaro, S. R. Filipović, F. C. Hummel, S. K. Jääskeläinen, V. K. Kimiskidis, G. Koch, B. Langguth, T. Nyffeler, A. Oliviero, F. Padberg, E. Poulet, S. Rossi, P. M. Rossini, J. C. Rothwell, C. Schönfeldt-Lecuona, H. R. Siebner, C. W. Slotema, C. J. Stagg, J. Valls-Sole, U. Ziemann, W. Paulus, and L. Garcia-Larrea, "Evidence-based guidelines on the therapeutic use of repetitive transcranial magnetic stimulation (rTMS)," *Clinical Neurophysiology*, vol. 125, no. 11, pp. 2150–2206, 2014.

[3] R. F. Cash and A. Zalesky, "Personalized and Circuit-Based Transcranial Magnetic Stimulation: Evidence, Controversies, and Opportunities," *Biological Psychiatry*, vol. 95, no. 6, pp. 510–522, 3 2024.

[4] V. H. Souza, R. H. Matsuda, A. S. Peres, P. H. J. Amorim, T. F. Moraes, J. V. L. Silva, and O. Baffa, "Development and characterization of the InVesalius Navigator software for navigated transcranial magnetic stimulation," *Journal of Neuroscience Methods*, vol. 309, no. June, pp. 109–120, 2018.

[5] J. Ruohonen and J. Karhu, "Navigated transcranial magnetic stimulation." *Neurophysiologie clinique = Clinical neurophysiology*, vol. 40, no. 1, pp. 7–17, 2010.

[6] P. Lioumis and M. Rosanova, "The role of neuronavigation in TMS–EEG studies: Current applications and future perspectives," *Journal of Neuroscience Methods*, vol. 380, p. 109677, 10 2022.

[7] E. M. Wassermann, L. M. McShane, M. Hallett, and L. G. Cohen, "Noninvasive mapping of muscle representations in human motor cortex," *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials*, vol. 85, no. 1, pp. 1–8, 1992.

[8] G. P. Tardelli, V. H. Souza, R. H. Matsuda, M. A. C. Garcia, P. A. Novikov, M. A. Nazarova, and O. Baffa, "Forearm and Hand Muscles Exhibit High Coactivation and Overlapping of Cortical Motor Representations," *Brain Topography*, vol. 35, no. 3, pp. 322–336, 5 2022.

[9] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "FastSurfer - A fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, p. 117012, 10 2020.

[10] E. J. Cole, A. L. Phillips, B. S. Bentzley, K. H. Stimpson, B. S. Nejad, F. Barmak, C. Veerapal, N. Khan, K. Cherian, E. Felber, R. Brown, E. Choi, S. King, H. Pankow, J. H. Bishop, A. Azeez, J. Coetzee, R. Rapier, N. Odenwald, D. Carreon, J. Hawkins, M. Chang, J. Keller, K. Raj, C. DeBattista, B. Jo, F. M. Espil, A. F. Schatzberg, K. D. Sudheimer, and N. R. Williams, "Stanford Neuromodulation Therapy (SNT): A Double-Blind Randomized Controlled Trial," *American Journal of Psychiatry*, vol. 179, no. 2, pp. 132–141, 2 2022.

[11] B. Luber, S. W. Davis, Z. D. Deng, D. Murphy, A. Martella, A. V. Peterchev, and S. H. Lisanby, "Using diffusion tensor imaging to effectively target TMS to deep brain structures," *NeuroImage*, vol. 249, 4 2022.

[12] D. B. Aydogan, V. H. Souza, R. H. Matsuda, P. Lioumis, and R. J. Ilmoniemi, "Real-Time Tractography-Assisted Neuronavigation for Transcranial Magnetic Stimulation," *Human Brain Mapping*, vol. 46, no. 1, 1 2025.

[13] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar mri," *Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[14] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle, "The human brain is intrinsically organized into dynamic, anticorrelated functional networks," Tech. Rep., 2005.

[15] A. M. Soto, M. Stenroos, R. H. Matsuda, T. Mylläri, K. Jyrkinen, H. Sinisalo, M. Laine, I. Granö, R. J. Ilmoniemi, and V. H. Souza, "Real-time electric-field neuronavigation on realistic head models for conventional and multi-locus TMS," *Brain Stimulation*, vol. 19, no. 3, p. 103113, 5 2026.

[16] S. M. Goetz, I. C. Kozyrkov, B. Luber, S. H. Lisanby, D. L. K. Murphy, W. M. Grill, and A. V. Peterchev, "Accuracy of robotic coil positioning during transcranial magnetic stimulation," *Journal of Neural Engineering*, vol. 16, no. 5, p. 054003, 9 2019.

[17] R. H. Matsuda, V. H. Souza, T. C. Marchetti, A. M. Soto, O.-P. Kahilakoski, M. Laine, H. Sinisalo, D. Kicic, P. Lioumis, R. J. Ilmoniemi, and O. Baffa, "Characterizing an electronic–robotic targeting platform for precise and fast brain stimulation with multi-locus transcranial magnetic stimulation," *Physics in Medicine & Biology*, vol. 70, no. 16, p. 165015, 8 2025.

[18] R. H. Matsuda, V. H. Souza, T. C. Marchetti, A. M. Soto, O.-P. Kahilakoski, A. Zhdanov, V. H. Malheiro, M. Laine, M. Nyrhinen, H. Sinisalo, D. Kicic, P. Lioumis, R. J. Ilmoniemi, and O. Baffa, "Robotic–electronic platform for autonomous and accurate transcranial magnetic stimulation targeting," *Brain Stimulation*, vol. 17, no. 2, pp. 469–472, 3 2024.

[19] M. Campos, V. H. Malheiro, T. Marchetti, L. Bettioli, V. H. Moraes, C. Rondinoni, R. Matsuda, and O. Baffa, "Simultaneous Dual-Robot Control for Transcranial Magnetic Stimulation on the InVesalius Platform," in *2026 Global Medical Engineering Physics Exchanges/ Pan American Health Care Exchanges (GMEPE/PAHCE)*. IEEE, 4 2026, pp. 1–4.

[20] G. Koch, M. Franca, M. Fernandez Del Olmo, B. Cheeran, R. Milton, M. Alvarez Sauco, and J. C. Rothwell, "Time Course of Functional Connectivity between Dorsal Premotor and Contralateral Motor Cortex during Movement Selection," *The Journal of Neuroscience*, vol. 26, no. 28, pp. 7452–7459, 7 2006.

[21] P. Amorim, T. Moraes, J. Silva, and H. Pedrini, "InVesalius: An Interactive Rendering Framework for Health Care Support," ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9474, pp. 45–54.

[22] S. Parvin, M. Laine, A. M. Soto, T. Nurmi, M. Stenroos, I. Rissanen, H. Sinisalo, O. P. Kahilakoski, I. Grano, T. Roine, H. Piitulainen, U. Ziemann, R. J. Ilmoniemi, and V. H. Souza, "Feasibility Study of Supplementary Motor Area to Primary Motor Cortex Facilitation using Multi-locus Transcranial Magnetic Stimulation," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2025, pp. 1–6, 7 2025.

[23] B. Fischl, "FreeSurfer," pp. 774–781, 8 2012.

[24] A. Thielscher, A. Antunes, and G. B. Saturnino, "Field modeling for transcranial magnetic stimulation: A useful tool to understand the physiological effects of TMS?" in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 8 2015, pp. 222–225.

[25] P. A. Novikov, M. A. Nazarova, and V. V. Nikulin, "TMSmap – Software for Quantitative Analysis of TMS Mapping Results," *Frontiers in Human Neuroscience*, vol. 12, 7 2018.

An Integrated Optogenetic and Operant Conditioning System for Instrumental Memory Research in Rats

1st Bruno Ferreira da Silva
Prog. de Iniciação Científica
IIN-ELS
Macaíba, Brazil
bruno.silva@edu.isd.org.br

2nd Aurea Estella de Araújo Silva
Prog. Pós-Graduação em Neuroengenharia
IIN-ELS
Macaíba, Brazil
aurea.silva@edu.isd.org.br

3rd Alan Michel Bezerra Furtunato
Prog. Pós-Graduação em Neuroengenharia
IIN-ELS
Macaíba, Brazil
alan.furtunato@isd.org.br

4th Rafael Augusto de Oliveira Guedes
Prog. Pós-Graduação em Neuroengenharia
IIN-ELS
Macaíba, Brazil
rafael.guedes@isd.org.br

5th Ramón Hypolito Lima
Prog. Pós-Graduação em Neuroengenharia
IIN-ELS
Macaíba, Brazil
ramon.lima@isd.org.br

***Abstract*—Studying the neural basis of instrumental memory requires behavioral platforms capable of delivering temporally precise circuit manipulations while continuously recording operant behavior. This paper describes the design, construction, and integration of an optogenetic stimulation system coupled to an automated operant conditioning apparatus for instrumental memory research in rats. The system combines a microcontroller-based trigger circuit, a fiber-coupled LED driver, custom and commercial optical fiber cannulas, and a rotary joint compatible with freely moving animals, synchronized in closed loop with a commercial operant chamber via TTL communication and custom Python scripts. We describe the fabrication and validation of the optical components, the surgical and viral-vector strategy for opsin delivery, the firmware logic for event-triggered and scheduled stimulation, the hardware/software synchronization architecture, and a four-stage behavioral protocol (pre-training, training, reactivation, and test) capable of probing distinct temporal windows of instrumental memory processing. The system was implemented and operated across 18 animals over a multi-month protocol without fiber failure or trigger malfunction, demonstrating its feasibility as a reusable and reproducible platform for combining optogenetics with operant behavioral assays.**

***Index Terms*—optogenetics, instrumental conditioning, behavioral neuroscience instrumentation, closed-loop stimulation, operant chamber**

I. Introduction

Instrumental memory, the association between an action and its consequence, is central to adaptive behavior and depends on a distributed neural circuit involving the prefrontal cortex, hippocampus, amygdala, and striatum [1]. Understanding the causal contribution of specific nodes within this circuit to distinct temporal phases of memory processing — acquisition, consolidation, and reconsolidation — requires experimental tools capable of manipulating neural activity with high temporal precision while simultaneously recording operant behavior [2].

Pharmacological and lesion-based approaches have historically been the main tools for probing these circuits, but their relatively slow onset and long duration of action limit the temporal resolution with which distinct memory phases can be dissociated [2]. Optogenetics addresses this limitation by enabling reversible, cell-type-specific control of neuronal activity through light-sensitive opsins delivered by viral vectors [3]. Its main experimental advantage is temporal resolution: light delivery can be restricted to a few seconds around a specific behavioral event, such as a lever press or a memory-reactivation session, allowing dissociation between processing phases that occur only minutes apart [3], [4].

Realizing this advantage in practice, however, is non-trivial. Most commercial operant-conditioning platforms are designed as closed, self-contained systems for behavioral data acquisition and are not natively equipped for closed-loop optogenetic triggering [5]. Integrating optogenetic stimulation therefore typically requires custom hardware bridges and software synchronization layers built around the commercial platform, a step that is often under-documented in the neuroscience literature despite being a substantial practical bottleneck for laboratories adopting these techniques.

This paper addresses that gap by describing, in engineering detail, an integrated system developed to combine optogenetic manipulation with automated operant conditioning for the study of instrumental memory in rats. We detail the optical and electronic components and their validation, the surgical and viral-vector strategy for opsin delivery, the firmware logic used for event-triggered and scheduled stimulation, the synchronization architecture between the stimulation and behavioral-recording subsystems, and a four-stage behavioral protocol de-

signed to probe acquisition, consolidation, and reconsolidation. Rather than reporting the neuroscientific findings obtained with the platform, the present contribution focuses on the system itself, with the goal of providing a reproducible technical reference for laboratories seeking to implement similar closed-loop optogenetic-behavioral platforms.

The remainder of this paper is organized as follows. Section II briefly situates the system with respect to related closed-loop behavioral-optogenetic approaches. Section III describes the full system architecture and experimental methods. Section IV reports validation results obtained from operating the platform across a full experimental cohort. Section V discusses design trade-offs and limitations, and Section VI concludes the paper.

II. Related Work

Closed-loop optogenetic stimulation, in which light delivery is made contingent on a real-time behavioral or physiological event rather than delivered on a fixed schedule, has been increasingly adopted in behavioral neuroscience because it allows stimulation to be precisely anchored to the trial structure of a task [5]. Compared with open-loop protocols, in which light is delivered independently of ongoing behavior, closed-loop designs reduce temporal jitter between the neural manipulation and the behavioral event of interest, which is particularly important when probing short-lived processes such as post-reactivation memory destabilization [2].

A recurring practical challenge across laboratories implementing such designs is that commercial behavioral platforms (e.g., operant chambers, mazes, or head-fixation rigs) and optogenetic hardware (lasers or LED drivers) are typically supplied as independent systems, each with its own proprietary control software. Bridging them commonly relies on general-purpose microcontrollers or data-acquisition boards operating on the digital input/output (TTL) lines exposed by the behavioral platform [5]. The system described in this paper follows this general strategy, using an inexpensive microcontroller as the integration layer between a commercial operant chamber and a fiber-coupled LED driver, while adding a rotary-joint-based optical delivery path suited to chronic use in freely moving animals.

This pattern is reflected in the literature: several recent studies report similarly bespoke integrations built around commercial platforms such as Med Associates, each solving the bridging problem independently rather than following a shared, documented approach [6], [7], [8]. This fragmentation motivates presenting the present system's hardware and firmware design in enough engineering detail to be replicated directly.

III. System Description and Methods

A. Ethics and Subjects

All procedures were approved by the Animal Ethics Committee of Instituto Santos Dumont (CEUA-ISD, protocol 01/2020 ISD) and followed Brazilian federal law 11,794/2008 and the guidelines of the Conselho Nacional de Controle de Experimentação Animal (CONCEA). The system was implemented and operated using 18 adult male Wistar rats (*Rattus norvegicus*, $\sim$3 months old, 300–400 g), housed in groups of up to three under a 12h light/dark cycle, with controlled temperature ($\sim$22 °C) and humidity. Seven days before the behavioral protocol, animals were placed on a food-restriction schedule (16 g of chow per day, with water *ad libitum*) to increase motivation for the sucrose reward; restriction was maintained throughout the protocol.

B. Safety and Handling Procedures

Standardized handling procedures were established and followed by all operators: (i) the fiber-optic patch cable and implanted cannula were handled with forceps and connector sleeves rather than by hand, given the fragility of the fiber core; (ii) personal protective equipment (gloves, laboratory coat, long trousers, and closed shoes) was mandatory whenever an animal was being tethered or stimulated; (iii) all cables and connector sleeves were stored properly after each session to avoid mechanical damage; and (iv) only trained operators performed the tethering/untethering procedure to minimize torque on the implant.

C. System Overview

The platform is organized into three functional subsystems: (1) an optical stimulation subsystem, responsible for generating and delivering calibrated light to the implanted cannula; (2) a behavioral-recording subsystem, a commercial operant chamber responsible for stimulus presentation, reward delivery, and event logging; and (3) an integration layer, implemented in a microcontroller and a host computer, responsible for synchronizing the two other subsystems in closed loop. Table I summarizes the main hardware components and their function within the platform.

TABLE I
Main Hardware Components of the Optical Stimulation Subsystem

Component	Function
Microcontroller (Arduino)	Processes trigger logic and generates the digital signal that drives light delivery.
LED driver (Plexon LD-1)	Receives the digital trigger and supplies a constant, calibrated current to the light source.
Fiber-coupled LED light source	Converts electrical current into light at a specific wavelength (589 nm).
SMA connector	Precision, threaded mechanical coupling between the light source and the patch fiber, minimizing insertion loss.
Integrated rotary joint	Allows 360° fiber rotation without torquing or breaking the fiber, required for freely moving animals.
Interconnect / ceramic sleeve	Mechanically aligns and couples the patch fiber to the chronically implanted cannula.
Fiber-optic cannula	Surgically implanted interface, terminating in a short fiber segment delivering light to the target region.

D. Optical and Electronic Components

Two microcontroller firmware configurations were used, corresponding to the two stimulation modes required by the behavioral protocol.

The first configuration operates in *closed loop*: it monitors a digital input line (TTL) driven by the operant-chamber controller and, upon detecting a rising edge corresponding to a correct lever press, immediately sets the trigger output high, driving the LED for a fixed 20-second window at a constant output power of 16 mW; a software flag prevents retriggering while a stimulation window is already active, and the output is set low automatically once the window elapses. This logic was implemented as simple edge-detection combined with a non-blocking, `millis()`-based timer, avoiding the use of blocking delay calls so that the trigger circuit remains responsive throughout the session.

The second configuration operates *independently* of the behavioral platform, for use in the open-field post-reactivation stimulation stage. Upon receiving a start command over a serial connection, it executes a fixed pulsed schedule (5 s on / 5 s off) for a total duration of 30 minutes, after which it automatically returns to an idle state awaiting a new start command. This design allows the same hardware to be reused for both an event-contingent (closed-loop) protocol and a fixed-schedule (open-loop) protocol simply by loading the corresponding firmware.

E. Fiber-Optic Cannula Fabrication and Validation

Optical fiber cannulas (200 μm core diameter, 0.39 NA, 9 mm free-fiber length) were either sourced commercially (Plexon Inc.) or fabricated in-house by inserting and epoxy-fixing the fiber into 1.25-mm stainless-steel ferrules compatible with standard FC/PC connectors. Free-fiber length was mechanically polished and verified under a microscope to 9 ± 0.1 mm, following stereotaxic coordinates derived from a standard rat brain atlas. Every fabricated unit was validated prior to implantation by passing continuous-wave 589-nm light through the assembly and visually confirming transmission integrity and the absence of fiber breaks or excessive insertion loss; units that failed this test were discarded and not implanted.

F. Viral Vector Strategy

Opsin expression was achieved with the recombinant adeno-associated viral vector AAV-EF1a-NpHR3.0-EYFP, encoding the inhibitory opsin halorhodopsin (NpHR3.0) fused to an EYFP reporter. This vector class was selected for its low immunogenicity and stable long-term neuronal expression [3]. Following stereotaxic infusion, a three-week interval was allowed for adequate opsin expression before optical fiber implantation and the start of behavioral testing, consistent with standard AAV expression timelines.

G. Stereotaxic Surgical Procedures

Viral infusion and optical fiber implantation were performed as two separate stereotaxic procedures under general anesthesia, three weeks apart, to allow both surgical recovery and adequate opsin expression before optical access was established. Standard post-operative analgesia and individual recovery housing were used following each procedure. As the surgical protocol itself is not the focus of the present system-oriented description, full anesthetic, stereotaxic, and post-operative details are reported in a companion paper from our laboratory dedicated specifically to the surgical procedure.

H. Operant Chamber Integration and Data Pipeline

Behavioral sessions were conducted in sound-attenuated MedAssociates operant chambers equipped with two retractable levers, a central reward magazine with infrared nose-poke detection, a house-light, and cue lights positioned above each lever. The chamber controller and the optogenetic trigger circuit were synchronized through custom Python scripts communicating over the chamber's TTL input/output lines, ensuring that light delivery was tightly time-locked to specific behavioral events without requiring manual intervention during the session. Behavioral event logs exported by the chamber software were subsequently organized into structured tabular datasets and processed using a standard Python-based data-analysis pipeline, providing an end-to-end path from raw event timestamps to analysis-ready behavioral variables suitable for the statistical characterization reported separately.

I. Habituation Procedure

During habituation sessions, the entire optical stimulation subsystem remained powered off; a damaged (non-functional) patch cable was connected to the implanted cannula through the ceramic sleeve, reproducing the mechanical experience of tethering without any light delivery. This step was intended to dissociate the animals' behavioral response to physical tethering from any response driven by light stimulation.

J. Behavioral Protocol

The protocol implemented on this platform comprises four sequential stages. (1) *Pre-training*: a single 15-minute session with free reward delivery at fixed intervals and levers retracted, to familiarize animals with the reward magazine. (2) *Training*: up to ten daily 30-minute sessions under a fixed-ratio 1 (FR1) schedule, in which a correct response on the active lever delivers a sucrose pellet and, depending on the experimental condition, triggers closed-loop optogenetic stimulation. (3) *Reactivation*: a single session, 24 hours after the last training session, under a variable-ratio 20 (VR20) schedule, a reinforcement schedule associated with destabilization of previously consolidated instrumental memories; this session may be optionally followed by post-reactivation stimulation delivered in an open-field arena. (4) *Test*: a 30-minute extinction session (levers available, no reward delivered) to assess memory retention. This modular structure allows optogenetic manipulation to be selectively restricted to the training window, the post-reactivation window, or both, depending on the specific scientific question being addressed by the experimenter.

IV. System Validation

A. Optical and Mechanical Reliability

The platform was operated continuously across 18 animals over a multi-month experimental timeline encompassing up to ten training sessions, one reactivation session, and one test session per animal. All fabricated and commercial fiber-optic cannulas passed the pre-implantation light-transmission test described in Section III-D. No fiber failures or cable disconnections were recorded during the behavioral phase, indicating that the rotary-joint and interconnect-sleeve arrangement reliably tolerated repeated tethering and untethering cycles in freely moving animals across the full protocol duration.

B. Trigger Timing Reliability

The closed-loop firmware reliably delivered a stimulation pulse contingent on each qualifying lever press throughout training, with the non-blocking timer design preventing missed or overlapping trigger events during sustained bouts of responding. The independent pulsed-mode firmware reliably executed the fixed 30-minute open-field stimulation protocol without requiring manual timing by the experimenter, reducing a common source of human error in fixed-duration stimulation protocols.

C. Operational Throughput

Across the full cohort, the platform supported the complete four-stage protocol — pre-training, up to ten training sessions, reactivation, and test — without requiring hardware reconfiguration between animals or between stimulation modes, other than reflashing the appropriate firmware configuration onto the microcontroller. A full quantitative behavioral characterization of instrumental learning obtained with this platform, including group-level comparisons under different optogenetic manipulation conditions, is beyond the scope of the present technical description and is reported elsewhere.

D. Behavioral Performance Under Tethered Conditions

To assess whether the tethering and implant hardware itself interferes with normal instrumental behavior, we examined performance in animals that were fully connected to the optical stimulation subsystem but received no light delivery (Control condition, n = 9). These animals reliably acquired the operant task, reaching the learning criterion ($\geq$60 correct responses) within 4.67 $\pm$ 1.00 days on average, with 100% of animals reaching criterion within the 10-day training window, and discrimination indices approaching ceiling (¿0.95) by the fifth training day (Fig. 1). This performance is consistent with expected acquisition timelines for this task, indicating that the physical presence of the fiber-optic tether and rotary-joint assembly does not measurably impair normal instrumental learning.

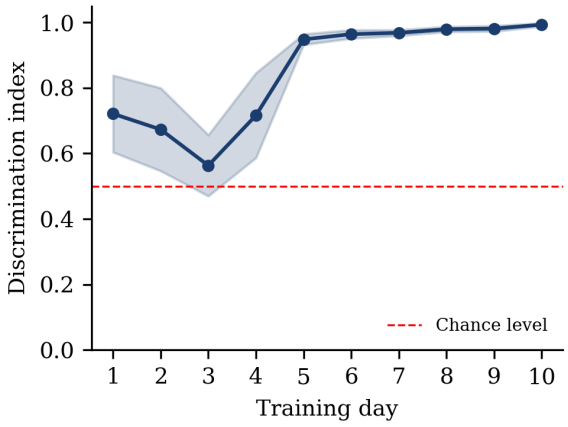


Fig. 1. Discrimination index across the 10-day training protocol for tethered Control animals (n = 9, mean $\pm$ SEM), showing normal task acquisition under the physical presence of the fiber-optic tether and rotary-joint assembly.

V. Discussion

The system described here addresses a practical bottleneck in circuit-level studies of instrumental memory: commercial operant-conditioning platforms do not natively support closed-loop optogenetic triggering, and integrating the two typically requires an *ad hoc* hardware/software bridge [5]. By using an inexpensive microcontroller as the trigger layer and TTL-based communication with the chamber controller, the present architecture keeps the behavioral platform largely unmodified while adding precise event-contingent optical control — a design choice intended to make the system easier to replicate in laboratories that already operate commercial operant chambers, without requiring proprietary hardware modifications.

The rotary-joint and sleeve-based fiber-coupling strategy was a key element enabling multi-week use in freely moving animals without mechanical fiber failure, which is a common practical limitation in chronic optogenetic experiments [3]. Separating viral infusion from fiber implantation by three weeks, following established practice for AAV-mediated opsin expression, further reduced the risk of implanting before adequate expression was achieved, avoiding wasted implantation surgeries on animals with insufficient opsin availability.

Using two interchangeable firmware configurations on the same microcontroller hardware — one event-contingent and one schedule-based — allowed a single low-cost hardware design to serve both stimulation modes required by the protocol, rather than requiring separate dedicated devices. This design choice reduces both hardware cost and the complexity of maintaining multiple stimulation circuits in parallel.

A limitation of the present configuration is the anatomical offset between the viral infusion coordinates and the fiber implantation coordinates, introduced to accommodate the angled implantation trajectory used in this protocol. Laboratories targeting small or anatomically subdivided nuclei should validate, on a pilot basis, that the illuminated tissue volume adequately overlaps the transduced region for their specific target, since a mismatch between expression and illumination sites could reduce the effective manipulation even when both individually appear successful. Future iterations of the system could

incorporate automated optical-power calibration and real-time logging of stimulation events alongside behavioral timestamps within a single unified data stream, further reducing manual verification steps and simplifying post-hoc alignment between neural manipulation and behavior. Additionally, the present validation did not include independent electrical or optical measurement of trigger latency and jitter (e.g., via oscilloscope or photodiode), nor systematic logging of expected-versus-delivered stimulation counts; these measurements would further strengthen claims of synchronization precision and are recommended for future iterations of the system.

VI. Conclusion

We described the design, construction, and validation of an integrated optogenetic and operant-conditioning system for instrumental memory research in rats, combining a microcontroller-based closed-loop trigger, a fiber-coupled LED driver, a rotary-joint fiber-delivery path suited to freely moving animals, and a four-stage behavioral protocol capable of probing acquisition, consolidation, and reconsolidation. The system operated reliably across 18 animals over a multi-month protocol without fiber failure or trigger malfunction, supporting its use as a reproducible, low-cost platform for laboratories investigating the neural circuit basis of instrumental behavior. Future work will incorporate automated light-power calibration and unified event logging to further streamline data alignment between neural manipulation and behavioral output.

Acknowledgment

I express my sincere gratitude to all my colleagues at the Instituto Santos Dumont (ISD) for their support. Special thanks go to my advisor, Ramón H. Lima, for his outstanding guidance and mentorship, and to M.Sc. candidate Aurea E. A. Silva, who led the broader research project for which this experiment was planned and conducted.

References

[1] B. W. Balleine and J. P. O'Doherty, "Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action," *Neuropsychopharmacology*, vol. 35, no. 1, pp. 48–69, 2010.

[2] L. de Oliveira Alvares and F. H. Do-Monte, "Understanding the dynamic and destiny of memories," *Neuroscience & Biobehavioral Reviews*, vol. 125, pp. 592–607, 2021.

[3] A. Guru, R. J. Post, Y.-Y. Ho, and M. R. Warden, "Making sense of optogenetics," *International Journal of Neuropsychopharmacology*, vol. 18, no. 11, p. pyv079, 2015.

[4] M. S. Kang and J.-H. Han, "Optogenetic inhibition of medial entorhinal cortex inputs to the hippocampus during a short period of time right after learning disrupts contextual fear memory formation," *Molecular Brain*, vol. 14, p. 2, 2021.

[5] L. Grosenick, J. H. Marshel, and K. Deisseroth, "Closed-loop and activity-guided optogenetic control," *Neuron*, vol. 86, no. 1, pp. 106–139, 2015.

[6] S. Miller, J. C. Slack, and A. P. Yadav, "Low-cost, open-source, full-stack software and Arduino-based hardware for control of commercially available animal behavior systems," *Behavior Research Methods*, vol. 58, no. 8, p. 211, 2026.

[7] S. K. T. Kapanaiah, B. Van Der Veen, D. Strahnen, T. Akam, and D. Kätzel, "A low-cost open-source 5-choice operant box system optimized for electrophysiology and optophysiology in mice," *Scientific Reports*, vol. 11, no. 1, p. 22279, 2021.

[8] N. S. Chandra, N. K. Birenbaum, L. Green, W. Z. Ray, and M. R. MacEwan, "A custom interface for the joint operation of Med Associates and Tucker-Davis Technologies hardware in a rodent behavioral model," *Journal of Neuroscience Methods*, vol. 371, p. 109528, 2022.

BCI-FES Integration for Motor Rehabilitation: A Scoping Review

Melyssa Júlia Neres Faria
Graduate Program in Electrical Engineering
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627, 31270-901
Belo Horizonte, MG, Brazil
ORCID: 0000-0002-1440-5733

Ana Amantea Leal Marques Afonso
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
ORCID: 0009-0008-3524-5565

Felippe Veloso Marinho
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
ORCID: 0009-0003-6910-9143

Gustavo Eulalio de Souza Ferreira
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
ORCID: 0009-0008-0254-4112

Samuel Souza Marques
Faculdade de Medicina
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
ORCID: 0009-0007-3270-3471

Henrique Resende Martins
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
ORCID: 0000-0002-4879-1345

***Abstract*—Operational integration of brain-computer interfaces (BCIs) and functional electrical stimulation (FES), in which real-time neural signals trigger muscle activation, offers a neurobiologically grounded route to motor recovery, yet its technical scope has not been systematically mapped. We conducted a scoping review following PRISMA-ScR, searching PubMed, Scopus, and IEEE Xplore through March 2026. Eighty-two studies (2010–2025) were included. Stroke accounted for 72% of the clinical population and spinal cord injury for 28%; non-invasive EEG was dominant (94%), with motor imagery (48%) and movement execution or attempt (40%); surface FES was used in 93%. Thirty-two percent measured only system performance, without a clinical outcome. Of the 23 studies (28%) with a separate control group, 15 compared groups on a clinical outcome, with BCI-FES superior in 9 (8 of 14 independent cohorts), a vote count across mostly underpowered studies (median n=7) that should be read as a descriptive signal, not a measure of efficacy. Randomized trials numbered 15 (18%). FES frequency and pulse width were unreported in 41% and 39% of studies. Technical feasibility is suggested mainly in chronic upper-limb stroke, while efficacy evidence rests on a few small controlled studies and FES parameter reporting remains variable.**

***Index Terms*—brain-computer interface, functional electrical stimulation, motor rehabilitation, electroencephalography, scoping review**

I. Introduction

Stroke and spinal cord injury (SCI) are among the leading causes of permanent motor disability worldwide, imposing a substantial burden on individuals, caregivers, and health systems [1], [2]. Conventional rehabilitation approaches (physiotherapy, occupational therapy, task-oriented training) have shown limited efficacy in patients with severe deficits, motivating the development of neurotechnology-assisted interventions [3].

Brain-computer interfaces (BCIs) capture neural signals in real time, most commonly through electroencephalography (EEG), and translate them into commands for external devices, bypassing injured motor pathways [4]. Functional electrical stimulation (FES) delivers controlled electrical current to peripheral nerves or muscles, eliciting functional contractions in otherwise paralyzed limbs [5]. Both modalities have been investigated independently as rehabilitation tools; each, however, addresses only part of the neural-motor circuit.

When decoded cortical signals trigger the delivery of FES in real time, the result is a closed neuromuscular loop grounded in the principles of Hebbian plasticity. The simultaneous activation of the motor cortex and the peripheral muscle during intentional motor effort is hypothesized to strengthen corticospinal connections and to support long-term motor recovery [6], [7]. This rationale has sustained a growing body of experimental work over the past 15 years.

Existing systematic reviews and meta-analyses have examined the efficacy of BCI-FES in upper-limb rehabilitation after stroke [8]–[10]. This scoping review complements that evidence by characterizing the broader landscape: the populations studied, the technical configurations used, and the outcomes measured across all study designs, providing context for interpreting and extending the evidence from controlled trials. Previous reviews, including systematic and narrative syntheses focused on efficacy in specific populations [8]–[10], concentrated on the subset of controlled trials; to our knowledge, no prior scoping review has characterized the broader technical landscape of BCI-FES integration, spanning all study designs and populations, under the PRISMA-ScR framework.

This review aims to characterize study designs, populations, technical BCI and FES parameters, and reported outcomes in the BCI-FES integration literature, identifying patterns and gaps that may inform future research and reporting guidelines.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

II. Methods

A. Study Design

This scoping review was conducted following the PRISMA extension for scoping reviews (PRISMA-ScR) [11], with the protocol, eligibility criteria, and search strategy defined a priori.

B. Search Strategy

Three electronic databases were searched through March 20, 2026. PubMed was selected for its clinical and biomedical coverage; Scopus for its interdisciplinary breadth; and IEEE Xplore for its engineering and neurotechnology publications, where a substantial share of BCI research appears and is not fully indexed in the other databases. The search combined terms related to brain-computer interfaces ("brain-computer interface", "BCI", "brain-machine interface", "BMI"), functional electrical stimulation ("functional electrical stimulation", "FES", "neuromuscular electrical stimulation", "NMES"), and motor rehabilitation ("motor rehabilitation", "motor recovery", "neurorehabilitation"). No language restriction was applied, and the strategy was adapted to the syntax of each database. The full search string is available from the authors on request.

C. Inclusion and Exclusion Criteria

Studies were included if they: (I1) involved adults ($\geq$18 years) with motor deficits secondary to a neurological condition; (I2) implemented real-time BCI-FES integration, defined as central neural signals captured by a BCI directly and continuously modulating FES delivery; (I3) reported at least one quantitative outcome from an implemented system; and (I4) were published between January 2010 and March 20, 2026, as full papers.

Studies were excluded if they: (E1) involved exclusively healthy participants; (E2) used BCI and FES in temporally separate phases, without real-time integration; (E3) controlled FES through electromyography, movement sensors, or manual control without an operational neural BCI; (E4) used a powered robotic exoskeleton as the primary effector without FES; (E5) were exclusively computational studies, simulations, or animal experiments; or (E6) were reviews, meta-analyses, editorials, protocols without collected data, or conference abstracts without a full paper.

D. Screening and Data Extraction

Titles and abstracts of 1,161 unique records were screened by a single reviewer; 168 full texts were assessed for eligibility. Single-reviewer screening is an acknowledged limitation, mitigated by requiring a supporting verbatim quotation from the source article for the six classification fields most subject to interpretation (design, control group, BCI paradigm, BCI-FES integration, outcome, and between-group superiority). Data were extracted using a standardized form with 52 fields across eight domains: study identification, participants, study design, BCI characteristics, FES characteristics, BCI-FES integration, intervention protocol, and outcomes. Categorical fields followed a closed vocabulary; any field without explicit information in the source article was coded "NR" (not reported), and inference was not permitted. The unit of analysis was the individual publication. Some groups contributed multiple articles, and a subset reports overlapping or identical participant cohorts: the Remsik et al. series (2018, 2019, 2021, 2022) [55], [60], [73], [80], derived from a common cohort; two publications from the same Krueger et al. cohort (2022, 2024) [15], [78]; two publications on the same patient by Bastos-Filho et al. [92], [93]; and two publications describing the same nine-patient cohort by Tan et al. (2010, 2011) [20], [22]. Participant totals should therefore be read at the report level, without assuming strictly independent samples. A post-hoc reliability check was subsequently conducted for all 82 included studies: a second author independently re-extracted the six critical fields listed above, blind to the original classification, and discrepancies were adjudicated by the author team; the figures reported below reflect this adjudication.

E. Classification Criteria

To reduce subjectivity, each categorical field followed an explicit rule. Design was classified hierarchically: randomized trial ($\geq$2 groups with random allocation), non-randomized controlled trial ($\geq$2 groups without randomization), single case (n=1), and, for single-arm samples with n$\geq$2, feasibility/pilot study (when the stated primary outcome was feasibility, accuracy, or safety) or case series (when it described clinical outcomes). The control-group field was coded as present only when a separate, prospectively allocated participant group existed; within-subject comparisons (e.g., A-B-A-B or condition-crossover designs) and historical controls were treated as the absence of a control group. The outcome was labeled according to what the study set out to measure: studies without any pre/post clinical comparison were coded "no clinical outcome", regardless of technical performance; the remainder were classified as positive, mixed, or negative by the prespecified clinical outcome, based on the Results section rather than the Discussion. Superiority was assigned only when a study reported a between-group statistical test on a clinical outcome.

III. Results

A. Study Selection

The database searches identified 1,301 records (PubMed: 118; Scopus: 981; IEEE Xplore: 202). After deduplication, 1,161 unique records were screened at the title and abstract level; 194 were sought for retrieval, of which 26 were not retrieved, yielding 168 articles assessed in full text. Of these, 82 met all inclusion criteria (Fig. 1). The most common reasons for full-text exclusion were reviews or records without primary data (n=24), the inclusion of healthy participants only (n=23), and FES driven by electromyographic, sensor, or manual control without an operational BCI (n=12); the full breakdown is shown in Fig. 1. Table III summarizes the 82 studies individually.

B. Publication Trends

The included studies were published between 2010 and 2025. Thirty-five percent of the corpus (n=29) appeared in the last five years (2021–2025), reflecting sustained growth in publication volume across the review period.

C. Study and Population Characteristics

Study designs were heterogeneous (Table I). Designs without a separate control group predominated: single-case studies (n=19, 23%), feasibility or pilot studies without control (n=16, 20%), and case series (n=12, 15%). Among designs with a comparator, non-randomized controlled trials accounted for 18 studies (22%), some using within-subject or historical comparison, and randomized trials for 15 (18%, including 2 randomized pilots); two studies had other designs (a technical parameter report and a case study). In total, 23 studies (28%) included a separate (between-subjects) control group, the denominator adopted for the superiority analyses below.

Stroke was the predominant clinical condition (n=59, 72%); SCI accounted for 28% (n=23). Among the 72 studies that reported the clinical phase, participants in the chronic phase were the most common (n=47, 65%), followed by the subacute phase (n=17, 24%), mixed-phase populations (n=7, 10%), and the acute phase (n=1, 1%).

Upper-limb rehabilitation was the primary focus in 78% of studies (n=64); lower-limb targets were addressed in 16% (n=13) and both segments in 5% (n=4). Sample sizes, defined as the number of participants recruited, ranged from 1 to 64 (median: 7; mean: 10.3). Twenty studies (24%) involved a single participant. Table I summarizes the corpus characteristics.

TABLE I
CHARACTERISTICS OF THE 82 INCLUDED STUDIES

Characteristic	Category	n (%)
Design	Single case	19 (23%)
	Feasibility/pilot, no control	16 (20%)
	Case series	12 (15%)
	Non-randomized controlled	18 (22%)
	Randomized trial	15 (18%)
	Other	2 (2%)
Clinical condition	Stroke	59 (72%)
	Spinal cord injury	23 (28%)
Clinical phase[a]	Chronic	47 (65%)
	Subacute	17 (24%)
	Mixed	7 (10%)
	Acute	1 (1%)
Target limb	Upper	64 (78%)
	Lower	13 (16%)
	Both	4 (5%)
	Not reported	1 (1%)
BCI type	Non-invasive EEG	77 (94%)
	Invasive (intracortical)	5 (6%)
BCI paradigm	Motor imagery	39 (48%)
	Execution/attempt	33 (40%)
	ERD/MRCP, attention, P300	8 (10%)
	Other	1 (1%)
	Not reported	1 (1%)
FES type	Transcutaneous	76 (93%)
	Peripheral nerve	1 (1%)
	Percutaneous	1 (1%)
	Not reported	4 (5%)
Outcome	Positive	44 (54%)
	Mixed	12 (15%)
	No clinical outcome	26 (32%)

[a]Phase percentages computed over the 72 studies reporting phase.

D. BCI Characteristics

Non-invasive EEG was the dominant BCI modality, used in 77 of the 82 studies (94%). Fully implanted intracortical arrays appeared in 5 studies (6%), aimed mainly at restoring grasp and reach in SCI [14]. Motor imagery (MI), in which participants imagined limb movement without physical execution, was the most common paradigm (n=39, 48%), followed by movement execution or attempt (ME), in which participants actively tried to perform the movement (n=33, 40%). MRCP/ERD-based paradigms without explicit MI/ME classification were used in 3 studies (4%), generally aimed at inducing associative neuroplasticity [18]; attention/concentration indices and P300 appeared in the remainder.

Recording channels were reported in 90% of studies (n=74), with a median of 16; the highest values correspond to the invasive intracortical studies. Electrode placement predominantly targeted the sensorimotor cortex: C3, C4, and Cz (the canonical hand and arm motor cortex positions in the 10-20 system) appeared in the large majority of configurations, often supplemented by frontocentral and centroparietal electrodes; frontal-pole electrodes (Fp1/Fp2) appeared in only 8 studies (10%). EEG frequency bands were reported in 90% of studies (n=74); the mu rhythm (8–13 Hz) and the beta band (13–30 Hz) were the primary features, with most studies applying broadband 8–30 Hz filtering that captures both. Classification algorithms were reported in 94% of studies (n=77), with LDA, threshold-based ERD detection, and SVM among the most frequent.

E. FES Characteristics

Transcutaneous FES was used in 76 studies (93%), intramuscular percutaneous FES in 1 (1%), peripheral nerve stimulation in 1 (1%), and 4 studies (5%) did not report the stimulation type. Among the 48 studies that reported the stimulation frequency, the median was 35 Hz; the parameter was not reported in 41% of studies (n=34). Pulse width had a median of 250 μs among the 50 studies that reported it, and was absent in 39% of studies (n=32). The most frequent target muscles were the wrist and finger extensors, followed by the flexors; in lower-limb studies, the ankle dorsiflexors, as in post-stroke foot-drop correction systems [17]. Devices were varied commercial stimulators (e.g., Hasomed RehaStim, Compex Motion, g.tec), unnamed in 16 studies. Table II summarizes the FES parameters.

TABLE II
REPORTED FES STIMULATION PARAMETERS

Parameter	Reported, n (%)	Median	Most frequent
Stimulation frequency	48 (59%)	35 Hz	n/a
Pulse width	50 (61%)	250 μs	n/a
Stimulation type	78 (95%)	n/a	Transcutaneous (n=76)
Device	66 (80%)	n/a	Varied

F. Intervention Protocols

The number of sessions was reported in 84% of studies (n=69), with a median of 12. Session duration had a median of 60 min, reported in 71% (n=58).

G. Outcomes and Results

The Fugl-Meyer Assessment for the Upper Extremity (FMA-UE) was the most used primary motor outcome (n=21 studies); the Action Research Arm Test (ARAT) appeared as a secondary measure in several additional studies.

Clinical outcome measurement varied with the developmental stage of the studies. Thirty-two percent (n=26) assessed only BCI system performance metrics, such as classification accuracy and signal quality; many are engineering or feasibility studies, in which technical performance is the outcome pertinent to that stage. Among the 56 studies that assessed clinical outcomes, 44 (54% of the full corpus) reported statistically significant improvements in at least one primary motor measure; 12 (15%) reported mixed or inconclusive results; and no study reported exclusively null or negative motor findings (the boundary between positive and mixed depends on the criterion of significant improvement in at least one primary motor measure, and the absence of negative studies is consistent with a publication filter).

Formal efficacy evaluation was infrequent. Of the 82 studies, 23 (28%) had a separate control group, but only 15 statistically compared the groups on a clinical outcome; the remaining 8 reported within-group results only. In those 15 studies, BCI-FES exceeded the control in 9 and did not differ significantly in the remaining 6. As two of the nine favorable reports describe the same trial (Krueger et al., 2022 and 2024), superiority corresponds to 8 of 14 independent cohorts.

This figure is a vote count of statistical significance across small and heterogeneous studies (median n=7) and should be read as a descriptive signal. Most reported significance only, without quantifying effect magnitude, which therefore could not be pooled; the absence of significance may reflect low statistical power, and publication bias favoring positive outcomes cannot be excluded.

IV. Discussion

A. Technical Profile and Methodological Fragmentation

The included studies describe a technically convergent field: non-invasive EEG capturing sensorimotor signals (predominantly at C3, C4, and Cz) to drive transcutaneous FES of the upper limb in stroke survivors. The predominance of the upper limb (78%) follows the clinical profile of the dominant population, since stroke survivors (72%) most commonly present hand and arm deficits. SCI accounted for 28% of studies, some using intracortical recording to restore grasp and reach, as in the demonstration by Ajiboye et al. [14]. The corpus thus ranges from non-invasive feasibility systems to invasive functional-restoration interfaces, and from designs in which the neural signal continuously modulates stimulation to those in which a classifier's discrete output triggers a preconfigured FES pattern, a heterogeneity worth weighing when interpreting aggregate counts under the single label "BCI-FES".

On this technical convergence rests a methodological fragmentation. Motor imagery (MI, 48%) and movement execution or attempt (ME, 40%) appear in similar proportions and rarely with explicit differentiation, although they engage distinct neural circuits [95] (MI activates cortico-cortical motor-simulation networks; ME recruits residual corticospinal pathways) and generate different EEG signatures in topography, latency, and frequency content. In plegic patients, the boundary between imagery and attempted movement is clinically thin, since the attempt may not produce observable movement. Treating neurophysiologically distinct paradigms as equivalent, together with the variability in FES parameter reporting discussed below, makes it difficult to pool results and compare protocols that, on the surface, appear similar.

B. Efficacy: A Narrow Evidence Base

Only 23 of the 82 studies (28%) included a separate control group and, of these, 15 conducted a between-group statistical comparison on a clinical outcome, with superiority reported in favor of BCI-FES in 9 (60% of those tested; 39% of controlled studies), or 8 of 14 independent cohorts (57%) after merging the two reports of the same Krueger trial. This result accompanies the positive randomized trials in the corpus: Biasiucci et al. [7] documented lasting arm motor recovery after stroke, Krueger et al. [15] attributed the gain to Hebbian plasticity induced by the temporal coincidence of intention and stimulation, and Chen et al. [16] reproduced the effect in chronic stroke; no SCI study in the corpus reported a formal between-group comparison favoring BCI-FES, leaving the favorable evidence concentrated in stroke. The meta-analyses of Ren et al. [9] and Liang et al. [10] estimate a moderate effect magnitude in upper-limb rehabilitation after stroke, compatible with the proportion observed here.

The broad reading, however, qualifies this signal. The meta-analyses synthesize only the controlled stroke subset, whereas the full mapping shows that such trials are a small fraction of the published work: most controlled studies had a small sample (median n=7), the 15 randomized trials (18%) follow the same size pattern, and 72% of studies included no comparator at all. The efficacy evidence remains preliminary: the 60% proportion results from vote counting of significance in small studies, which limits its interpretation as a measure of efficacy. Effect magnitude was rarely quantified: most controlled studies reported statistical significance only, and few reported a standardized effect size (for example, Biasiucci et al. [7], Cohen's d = 1.03; Krueger et al. [15], η^2 = 0.57; Ramirez-Nava et al. [86], Cohen's d = 0.99 on the Fugl-Meyer), which prevents a pooled magnitude estimate. Publication bias favoring positive outcomes also cannot be excluded.

C. FES Parameter Reporting: A Documented Variability

A contrast stands out in FES reporting. While nearly all studies described the mechanism by which the BCI triggers stimulation, the frequency was absent in 41% of articles and

the pulse width in 39%, proportions that remain relevant even among the controlled designs. The studies thus document how they integrate BCI and FES, but not always with which stimulation parameters. The omission has a practical consequence: a recent meta-analysis associated, in exploratory subgroup analyses, specific ranges with larger motor effects (frequency $>$50 Hz, pulse width 150 μs), although the between-subgroup comparison did not reach statistical significance [10], so that positive results obtained without recording these settings become difficult to reproduce or compare.

D. Scope of the Evidence: Motor BCIs and Cognitive Outcomes

Within the studies included in this review, FES control relied on sensorimotor-cortex signals; frontal-pole electrodes (Fp1/Fp2) rarely served as the primary source (8 studies, 10%), and cognitive outcomes appeared only as secondary endpoints (13 studies, 16%), never as a primary target. Some prefrontal attention-based systems (Fp1) drive rehabilitation through a powered pedaling robot rather than FES [12], [13]; they fall outside this review but point to a prefrontal attention-based approach (Fp1/Fp2) not yet combined with FES in the mapped corpus.

E. Limitations

The acknowledged limitations include: single-reviewer screening, partly mitigated by mandatory verbatim quotation for six critical fields and, post hoc, by an independent second-author re-extraction of those fields across the full corpus (Section II-D); the absence of a formal methodological quality assessment, consistent with scoping-review methodology [11]; a search restricted to three bibliographic databases, without grey literature or trial registries; probable publication bias favoring positive outcomes; cohort overlap among some publications from the same group, which inflates participant totals relative to the number of independent samples; and, in at least two randomized trials [7], [81], allocation by covariate minimization rather than simple randomization, a boundary case retained as "randomized trial" under the pre-specified rule but worth noting when interpreting the randomized-trial count.

F. Implications and Future Directions

Three directions emerge from this mapping. First, the variability in FES parameter reporting (Section IV-C) supports the adoption of a minimum reporting set for BCI-FES studies, a checklist containing stimulation frequency, pulse width, intensity, target muscles, and session regimen; without these items, protocols cannot be compared or replicated, and the ranges associated with larger effects by Liang et al. [10] become actionable only if these parameters are systematically recorded. Second, although technical feasibility is suggested mainly in chronic upper-limb stroke, with sparser demonstrations in other populations, the consolidation of therapeutic evidence depends on larger randomized trials, particularly in SCI, the lower limb, and the acute phase, underrepresented in the corpus and relevant for populations with gait deficits. Third, the prefrontal attention-based BCI approach (Fp1/Fp2), already explored with robotic effectors [12], [13], has not yet been combined with FES within the corpus mapped here; given the role of attention in engagement and in the effective training dose during rehabilitation, this integration represents a promising and as-yet-unexplored direction.

V. Conclusion

Eighty-two primary studies published between 2010 and 2025 characterize a field with a consistent technical profile: non-invasive EEG, motor-intention paradigms, and surface FES targeting predominantly the upper limb in stroke survivors. Among the studies that formally compared BCI-FES to a control, superiority was reported in about half of the between-group comparisons (9 of 15; 8 of 14 independent cohorts), a proportion qualified by small sample sizes (median n=7); randomized trials total 15 studies (18%). Substantial variability in FES parameter reporting is documented across study designs, and 32% of studies assessed system performance without measuring any clinical motor outcome. The corpus mapped here suggests technical feasibility concentrated in chronic upper-limb stroke, with sparser demonstrations in other populations and configurations and mostly in small studies, while protocol standardization and the replication of larger controlled evidence remain infrequent.

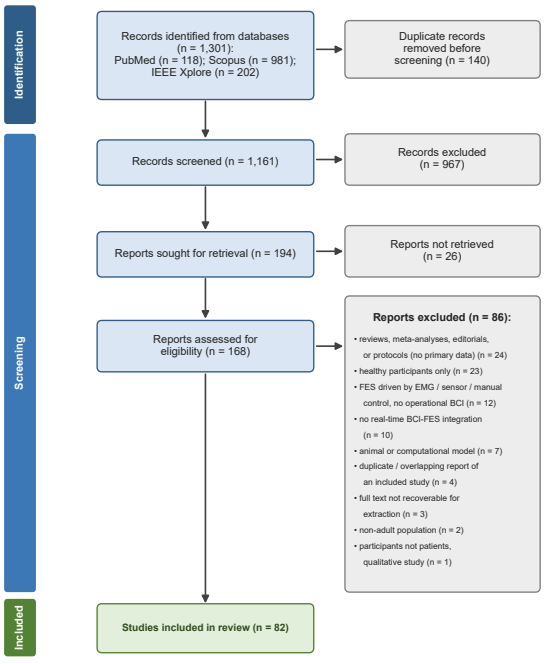


Fig. 1. PRISMA-ScR flow diagram of study selection (1,301 records identified; 1,161 after deduplication; 194 sought; 168 full texts assessed; 82 included).

References

[1] V. L. Feigin et al., "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, 2022, doi: 10.1177/17474930211065917.

TABLE III

INCLUDED STUDIES (N=82): CHARACTERISTICS AND OUTCOMES. COND.: STROKE; SCI (SPINAL CORD INJURY). DESIGN: RCT; PRCT (PILOT RCT); NRC (NON-RANDOMIZED CONTROLLED); SC (SINGLE CASE, N=1); FP (FEASIBILITY/PILOT); CS (CASE SERIES); OTH (OTHER). PARAD.: MI (MOTOR IMAGERY); ME (EXECUTION/ATTEMPT); ERD; ATT (ATTENTION); P300. FES: FREQUENCY (HZ) / PULSE WIDTH (μS). CTRL: SEPARATE CONTROL GROUP. OUTCOME: POS (POSITIVE); MIX (MIXED); NCO (NO CLINICAL OUTCOME); "SUP." INDICATES BETWEEN-GROUP SUPERIORITY FAVORING BCI-FES; "N.S." INDICATES A NON-SIGNIFICANT BETWEEN-GROUP DIFFERENCE.

Ref	Study	Cond.	Design	Parad.	FES Hz/μs	n	Ctrl	Outcome
[20]	Tan et al.	Stroke	NRC	MI	25/250	9	No	NCO
[21]	Tam et al.	Stroke	CS	MI	40/300	5	No	MIX
[22]	Tan et al.	Stroke	FP	MI	25/250	9	No	NCO
[17]	Do et al.	Stroke	SC	ME	NR/NR	1	No	NCO
[23]	Takahashi et al.	Stroke	SC	ME	50/1200	1	No	POS
[24]	Liang et al.	Stroke	FP	MI	NR/NR	8	Yes	NCO
[25]	Ono et al.	Stroke	SC	ME	20/100	1	No	POS
[26]	Pedrocchi et al.	SCI	FP	P300	22.5/225	5	No	NCO
[27]	Rohm et al.	SCI	SC	MI	16/NR	1	No	POS
[28]	Tam et al.	Stroke	FP	MI	NR/NR	5	No	NCO
[29]	Wang et al.	Stroke	NRC	MI	NR/NR	10	Yes	NCO
[30]	Chen et al.	Stroke	NRC	MI	NR/NR	10	Yes	NCO
[31]	Ibáñez et al.	Stroke	NRC	ME	30/250	5	No	NCO
[32]	King et al.	SCI	SC	ME	NR/NR	1	No	NCO
[33]	Li et al.	Stroke	RCT	MI	60/150	15	Yes	POS, sup.
[34]	Mukaino et al.	Stroke	SC	ERD	NR/NR	1	No	MIX
[35]	Young et al.	Stroke	SC	ME	NR/NR	1	No	POS
[36]	Zhang et al.	Stroke	NRC	MI	NR/NR	8	Yes	NCO
[37]	Chung et al.	Stroke	pRCT	Att	35/250	10	Yes	POS, n.s.
[38]	Jiang et al.	Stroke	CS	MI	NR/NR	4	No	NCO
[39]	King et al.	SCI	SC	ME	NR/NR	1	No	NCO
[40]	McCrimmon et al.	Stroke	CS	ME	NR/NR	9	No	POS
[41]	Vuckovic et al.	SCI	CS	MI	30/300	2	No	MIX
[42]	Jure et al.	Stroke	FP	MI	NR/NR	NR	No	NCO
[43]	Kim et al.	Stroke	RCT	Att	60/150	30	Yes	POS, sup.
[18]	Mrachacz-Kersting et al.	Stroke	NRC	ME	NR/1000	22	Yes	POS
[19]	Osuagwu et al.	SCI	NRC	ME	26/200	12	Yes	POS
[14]	Ajiboye et al.	SCI	SC	ME	12.5/100	1	No	POS
[44]	Bayon-Calatayud et al.	SCI	Oth	ME	NR/NR	1	No	POS
[45]	Friedenberg et al.	SCI	SC	MI	20/500	1	No	POS
[46]	Ibáñez et al.	Stroke	CS	ME	30/350	4	No	POS
[47]	Irimia et al.	Stroke	CS	MI	NR/NR	2	No	POS
[48]	Marquez-Chin et al.	Stroke	SC	ERD	NR/NR	1	No	POS
[49]	Poboroniuc et al.	Stroke	CS	MI	50/255	3	No	POS
[7]	Biasiucci et al.	Stroke	RCT	ME	23/255	27	Yes	POS, sup.
[50]	Cho et al.	Stroke	NRC	MI	50/300	7	No	POS
[51]	Gant et al.	SCI	FP	MI	NR/200	14	No	NCO
[52]	Hashimoto et al.	Stroke	NRC	ME	33.3/300	4	No	MIX
[53]	Irimia et al.	Stroke	CS	MI	NR/300	5	No	POS
[54]	Lupu et al.	Stroke	NRC	MI	NR/NR	7	No	NCO
[55]	Remsik et al.	Stroke	RCT	ME	60/150	21	Yes	POS, n.s.
[56]	Trincado-Alonso et al.	SCI	NRC	ME	40/350	4	No	MIX
[57]	Bockbrader et al.	SCI	SC	ME	NR/NR	1	No	POS
[58]	Cho et al.	Stroke	CS	MI	50/300	2	No	POS
[59]	Jovanovic et al.	Stroke	SC	ERD	40/250	1	No	POS
[60]	Remsik et al.	Stroke	RCT	ME	60/150	21	Yes	POS
[61]	Selfslagh et al.	SCI	CS	MI	30/300	2	No	POS
[62]	Choi et al.	Stroke	FP	MI	NR/NR	8	No	NCO
[63]	Chung et al.	Stroke	pRCT	Att	35/250	25	Yes	POS, sup.
[64]	Miao et al.	Stroke	RCT	MI	NR/300	16	Yes	POS
[65]	Murovec et al.	Stroke	FP	MI	50/350	3	No	POS
[66]	Sebastián-Romagosa et al.	Stroke	FP	MI	50/300	51	No	POS
[67]	de Castro-Cros et al.	Stroke	FP	MI	NR/NR	16	No	NCO
[68]	Broniera Junior et al.	SCI	SC	MI	35/123	1	No	NCO
[69]	Cajigas et al.	SCI	SC	MI	NR/NR	1	No	MIX
[16]	Chen et al.	Stroke	RCT	MI	25/NR	32	Yes	POS, sup.
[70]	Jovanovic et al.	SCI	FP	ME	40/250	8	No	NCO
[71]	Jovanovic et al.	SCI	CS	ME	40/250	5	No	POS
[72]	Jovanovic et al.	SCI	FP	ME	40/250	8	No	NCO
[73]	Remsik et al.	Stroke	NRC	ME	NR/NR	16	No	MIX
[74]	Sinha et al.	Stroke	NRC	ME	60/150	23	No	POS
[75]	Zulauf-Czaja et al.	SCI	FP	ME	33/250	8	No	NCO
[76]	Hashimoto et al.	Stroke	CS	ME	33.3/300	4	No	MIX
[77]	Kim et al.	Stroke	NRC	Oth	NR/NR	34	No	NCO
[78]	Krueger et al.	Stroke	RCT	ME	NR/NR	12	Yes	POS, sup.
[79]	Kumari et al.	Stroke	FP	ME	33/200	30	No	NCO
[80]	Remsik et al.	Stroke	Oth	ME	60/150	3	No	NCO
[81]	Wang et al.	Stroke	NRC	MI	30/300	7	Yes	POS, n.s.
[82]	Zhan et al.	Stroke	RCT	MI	NR/NR	24	Yes	POS, n.s.
[83]	Bigoni et al.	Stroke	SC	NR	NR/NR	1	No	POS
[84]	Cantillo-Negrete et al.	SCI	FP	MI	30/300	6	No	MIX
[85]	Fiorin et al.	SCI	NRC	MI	20.5/400	2	No	NCO
[86]	Ramirez-Nava et al.	Stroke	NRC	P300	30/375	14	Yes	POS, sup.
[87]	Sebastián-Romagosa et al.	Stroke	NRC	MI	50/350	25	No	POS
[88]	Brunner et al.	Stroke	RCT	MI	50/300	40	Yes	POS, n.s.
[89]	Herring et al.	SCI	SC	ME	NR/NR	1	No	POS
[15]	Krueger et al.	Stroke	RCT	ME	NR/NR	12	Yes	POS, sup.
[90]	Kumari et al.	SCI	RCT	ME	33/230	10	Yes	MIX, n.s.
[91]	Luo et al.	Stroke	RCT	MI	NR/NR	64	Yes	POS, sup.
[92]	Bastos-Filho et al.	Stroke	SC	MI	45/250	1	No	MIX
[93]	Bastos-Filho et al.	Stroke	SC	MI	20/200	1	No	MIX
[94]	Mansour et al.	Stroke	FP	ME	NR/NR	7	No	POS

[2] S. L. James et al., "Global, regional, and national burden of traumatic brain injury and spinal cord injury, 1990–2016," *Lancet Neurol.*, vol. 18, no. 1, pp. 56–87, 2019, doi: 10.1016/S1474-4422(18)30415-0.

[3] P. Langhorne, J. Bernhardt, and G. Kwakkel, "Stroke rehabilitation," *Lancet*, vol. 377, no. 9778, pp. 1693–1702, 2011, doi: 10.1016/S0140-6736(11)60325-5.

[4] J. R. Wolpaw, J. del R. Millán, and N. F. Ramsey, "Brain-computer interfaces: Definitions and principles," in *Brain-Computer Interfaces*, Handbook of Clinical Neurology, vol. 168, 2020, doi: 10.1016/B978-0-444-63934-9.00002-0.

[5] P. H. Peckham and J. S. Knutson, "Functional electrical stimulation for neuromuscular applications," *Annu. Rev. Biomed. Eng.*, vol. 7, pp. 327–360, 2005, doi: 10.1146/annurev.bioeng.6.040803.140103.

[6] J. N. Sanes and J. P. Donoghue, "Plasticity and primary motor cortex," *Annu. Rev. Neurosci.*, vol. 23, pp. 393–415, 2000, doi: 10.1146/annurev.neuro.23.1.393.

[7] A. Biasiucci et al., "Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke," *Nat. Commun.*, vol. 9, p. 2421, 2018, doi: 10.1038/s41467-018-04673-z.

[8] A. R. Alashram, E. Padua, and G. Annino, "Effects of brain-computer interface controlled functional electrical stimulation on motor recovery in stroke survivors: A systematic review," *Curr. Phys. Med. Rehabil. Rep.*, 2022, doi: 10.1007/s40141-022-00369-0.

[9] J. Ren et al., "The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: A systematic review and meta-analysis," *Front. Hum. Neurosci.*, vol. 18, 2024, doi: 10.3389/fnhum.2024.1438095.

[10] F. Liang, X. Chen, B. Li, and B. Yang, "Brain-computer interface combined with functional electrical stimulation for post-stroke upper limb motor recovery: A systematic review and meta-analysis," *Clin. EEG Neurosci.*, 2026, doi: 10.1177/15500594261441055.

[11] A. C. Tricco et al., "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," *Ann. Intern. Med.*, vol. 169, no. 7, pp. 467–473, 2018, doi: 10.7326/M18-0850.

[12] Y. Yuan et al., "Effect of BCI-controlled pedaling training system with multiple modalities of feedback on motor and cognitive function rehabilitation of early subacute stroke patients," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2569–2577, 2021, doi: 10.1109/TNSRE.2021.3132944.

[13] J. Wan et al., "Effects of dual-task mode brain-computer interface based on motor imagery and virtual reality on balance and attention in patients with stroke," *J. Neuroeng. Rehabil.*, vol. 22, 2025, doi: 10.1186/s12984-025-01730-9.

[14] A. B. Ajiboye et al., "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration," *Lancet*, vol. 389, no. 10081, pp. 1821–1830, 2017, doi: 10.1016/S0140-6736(17)30601-3.

[15] J. Krueger et al., "Hebbian plasticity induced by temporally coincident BCI enhances post-stroke motor recovery," *Sci. Rep.*, vol. 14, 2024, doi: 10.1038/s41598-024-69037-8.

[16] L. Chen et al., "EEG-controlled functional electrical stimulation rehabilitation for chronic stroke: system design and clinical application," *Front. Med.*, vol. 15, 2021, doi: 10.1007/s11684-020-0794-5.

[17] A. H. Do et al., "Brain-computer interface controlled functional electrical stimulation device for foot drop due to stroke," in *Proc. 34th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2012, doi: 10.1109/EMBC.2012.6347462.

[18] N. Mrachacz-Kersting et al., "Efficient neuroplasticity induction in chronic stroke patients by an associative brain-computer interface," *J. Neurophysiol.*, vol. 115, no. 3, pp. 1410–1421, 2016, doi: 10.1152/jn.00918.2015.

[19] B. A. Osuagwu et al., "Rehabilitation of hand in subacute tetraplegic patients based on brain computer interface and functional electrical stimulation: a randomised pilot study," *J. Neural Eng.*, vol. 13, no. 6, 2016, doi: 10.1088/1741-2560/13/6/065002.

[20] Tan et al., "Post-Acute Stroke Patients Use Brain-Computer Interface to Activate Electrical Stimulation," in *32nd Annu. Int. Conf. IEEE EMBS*, Buenos Aires, Argentina, 2010, doi: 10.1109/IEMBS.2010.5627381.

[21] Tam et al., "A Minimal Set of Electrodes for Motor Imagery BCI to Control an Assistive Device in Chronic Stroke Subjects: A Multi-Session Study," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2011, doi: 10.1109/TNSRE.2011.2168542.

[22] Tan et al., "EEG controlled neuromuscular electrical stimulation of the upper limb for stroke patients," *Front. Mech. Eng.*, 2011, doi: 10.1007/s11465-011-0207-1.

[23] Takahashi et al., "Event related desynchronization-modulated functional electrical stimulation system for stroke rehabilitation: A feasibility study," *J. NeuroEng. Rehabil.*, 2012, doi: 10.1186/1743-0003-9-56.

[24] Liang et al., "A Frequency Boosting Method for Motor Imagery EEG Classification in BCI-FES Rehabilitation Training System," in *ISNN 2013, Part II, LNCS 7952*, Springer, 2013, doi: 10.1007/978-3-642-39068-5_35.

[25] Ono et al., "Functional recovery in upper limb function in stroke survivors by using brain-computer interface: A single case A-B-A-B design," in *35th Annu. Int. Conf. IEEE EMBS*, 2013, doi: 10.1109/EMBC.2013.6609488.

[26] Pedrocchi et al., "MUNDUS project: MUltimodal Neuroprosthesis for daily Upper limb Support," *J. NeuroEng. Rehabil.*, 2013, doi: 10.1186/1743-0003-10-66.

[27] Rohm et al., "Hybrid brain-computer interfaces and hybrid neuroprostheses for restoration of upper limb functions in individuals with high-level spinal cord injury," *Artif. Intell. Med.*, 2013, doi: 10.1016/j.artmed.2013.07.004.

[28] Tam et al., "Channel selection in chronic stroke rehabilitation," in *3rd IFAC Int. Conf. Intelligent Control and Automation Science (ICONS)*, 2013, doi: 10.3182/20130902-3-CN-3020.00048.

[29] Wang et al., "Causal Neurofeedback Based BCI-FES Rehabilitation for Post-stroke Patients," in *ICONIP 2013, Part I, LNCS 8226*, Springer, 2013, doi: 10.1007/978-3-642-42054-2_52.

[30] Chen et al., "Classification of Stroke Patients' Motor Imagery EEG with Autoencoders in BCI-FES Rehabilitation Training System," in *ICONIP 2014, Part III, LNCS 8836*, Springer, 2014, doi: 10.1007/978-3-319-12643-2_25.

[31] Ibáñez et al., "Upper-limb muscular electrical stimulation driven by EEG-based detections of the intentions to move," in *36th Annu. Int. Conf. IEEE EMBS*, 2014, doi: 10.1109/EMBC.2014.6943921.

[32] King et al., "Brain-Computer Interface Driven Functional Electrical Stimulation System for Overground Walking in Spinal Cord Injury Participant," in *Proc. IEEE EMBS*, 2014, doi: 10.1109/EMBC.2014.6943821.

[33] Li et al., "Neurophysiological substrates of stroke patients with motor imagery-based brain-computer interface training," *Int. J. Neurosci.*, 2014, doi: 10.3109/00207454.2013.850082.

[34] Mukaino et al., "Efficacy of brain-computer interface-driven neuromuscular electrical stimulation for chronic paresis after stroke," *J. Rehabil. Med.*, 2014, doi: 10.2340/16501977-1785.

[35] Young et al., "Case report: post-stroke interventional BCI rehabilitation in an individual with preexisting sensorineural disability," *Front. Neuroeng.*, 2014, doi: 10.3389/fneng.2014.00018.

[36] Zhang et al., "An Iterative Method for Classifying Stroke Subjects' Motor Imagery EEG Data in the BCI-FES Rehabilitation Training System," in *Found. Pract. Appl. Cognitive Systems and Information Processing, AISC 215*, Springer, 2014, doi: 10.1007/978-3-642-37835-5_32.

[37] Chung et al., "Effects of brain-computer interface-based functional electrical stimulation on balance and gait function in patients with stroke: preliminary results," *J. Phys. Ther. Sci.*, 2015, doi: 10.1589/jpts.27.513.

[38] Jiang et al., "Application of BCI-FES System on Stroke Rehabilitation," in *7th Int. IEEE EMBS Conf. Neural Engineering*, Montpellier, France, 2015, doi: 10.1109/NER.2015.7146823.

[39] King et al., "The feasibility of a brain-computer interface functional electrical stimulation system for the restoration of overground walking after paraplegia," *J. NeuroEng. Rehabil.*, 2015, doi: 10.1186/s12984-015-0068-7.

[40] McCrimmon et al., "Brain-controlled functional electrical stimulation therapy for gait rehabilitation after stroke: a safety study," *J. NeuroEng. Rehabil.*, 2015, doi: 10.1186/s12984-015-0050-4.

[41] Vuckovic et al., "Hybrid Brain-Computer Interface and Functional Electrical Stimulation for Sensorimotor Training in Participants With Tetraplegia: A Proof-of-Concept Study," *J. Neurol. Phys. Ther.*, 2015, doi: 10.1097/NPT.0000000000000063.

[42] Jure et al., "BCI-FES system for neuro-rehabilitation of stroke patients," *J. Phys.: Conf. Ser.*, vol. 705, 012058, 2016, doi: 10.1088/1742-6596/705/1/012058.

[43] Kim et al., "Effects of Action Observational Training Plus Brain Computer Interface-Based Functional Electrical Stimulation on Paretic Arm Motor Recovery in Patient with Stroke: A Randomized Controlled Trial," *Occup. Ther. Int.*, 2016, doi: 10.1002/oti.1403.

[44] Bayon-Calatayud et al., "Usability of the Combination of Brain-Computer Interface, Functional Electrical Stimulation and Virtual Reality for Improving Hand Function in Spinal Cord Injured Patients," in *Converging Clinical and Engineering Research on Neurorehabilitation II, Biosystems & Biorobotics 15*, Springer, 2017, doi: 10.1007/978-3-319-46669-9_56.

[45] Friedenberg et al., "Neuroprosthetic-enabled control of graded arm muscle contraction in a paralyzed human," *Sci. Rep.*, 2017, doi: 10.1038/s41598-017-08120-9.

[46] Ibáñez et al., "Low Latency Estimation of Motor Intentions to Assist Reaching Movements along Multiple Sessions in Chronic Stroke Patients: A Feasibility Study," *Front. Neurosci.*, 2017, doi: 10.3389/fnins.2017.00126.

[47] Irimia et al., "Brain-Computer Interfaces With Multi-Sensory Feedback for Stroke Rehabilitation: A Case Study," *Artif. Organs*, 2017, doi: 10.1111/aor.13054.

[48] Marquez-Chin et al., "Brain-Computer Interface and Functional Electrical Stimulation for Upper Limb Rehabilitation After Stroke," in *IEEE GlobalSIP*, 2017, doi: 10.1109/GlobalSIP.2017.8308678.

[49] Poboroniuc et al., "FES&BCI Based Rehabilitation Engineered Equipment: Clinical Tests and Perspectives," in *6th IEEE Int. Conf. E-Health and Bioengineering (EHB)*, 2017, doi: 10.1109/EHB.2017.7995365.

[50] Cho et al., "Motor rehabilitation for hemiparetic stroke patients using a brain-computer interface method," in *IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2018, doi: 10.1109/SMC.2018.00178.

[51] Gant et al., "EEG-controlled functional electrical stimulation for hand opening and closing in chronic complete cervical spinal cord injury," *Biomed. Phys. Eng. Express*, 2018, doi: 10.1088/2057-1976/aabb13.

[52] Hashimoto et al., "Development of Rehabilitation System with Brain-Computer Interface for Subacute Stroke Patients," in *IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2018, doi: 10.1109/SMC.2018.00018.

[53] Irimia et al., "High Classification Accuracy of a Motor Imagery Based Brain-Computer Interface for Stroke Rehabilitation Training," *Front. Robot. AI*, 2018, doi: 10.3389/frobt.2018.00130.

[54] Lupu et al., "BCI and FES Based Therapy for Stroke Rehabilitation Using VR Facilities," *Wirel. Commun. Mob. Comput.*, 2018, doi: 10.1155/2018/4798359.

[55] Remsik et al., "Behavioral Outcomes Following Brain-Computer Interface Intervention for Upper Extremity Rehabilitation in Stroke: A Randomized Controlled Trial," *Front. Neurosci.*, 2018, doi: 10.3389/fnins.2018.00752.

[56] Trincado-Alonso et al., "A Pilot Study of Brain-Triggered Electrical Stimulation with Visual Feedback in Patients with Incomplete Spinal Cord Injury," *J. Med. Biol. Eng.*, 2018, doi: 10.1007/s40846-017-0343-0.

[57] Bockbrader et al., "Clinically Significant Gains in Skillful Grasp Coordination by an Individual With Tetraplegia Using an Implanted Brain-Computer Interface With Forearm Transcutaneous Muscle Stimulation," *Arch. Phys. Med. Rehabil.*, 2019, doi: 10.1016/j.apmr.2018.07.445.

[58] Cho et al., "Feasibility of Brain-Computer Interface Triggered Functional Electrical Stimulation and Avatar for Motor Improvement in Chronic Stroke Patients," in *ICNR 2018, Biosystems & Biorobotics 21*, Springer, 2019, doi: 10.1007/978-3-030-01845-0_219.

[59] Jovanovic et al., "Restoration of upper-limb function after chronic severe hemiplegia: A case report on the feasibility of a brain-computer interface controlled functional electrical stimulation therapy," *Am. J. Phys. Med. Rehabil.*, 2019, doi: 10.1097/PHM.0000000000001163.

[60] Remsik et al., "Ipsilesional Mu Rhythm Desynchronization and Changes in Motor Behavior Following Post Stroke BCI Intervention for Motor Rehabilitation," *Front. Neurosci.*, 2019, doi: 10.3389/fnins.2019.00053.

[61] Selfslagh et al., "Non-invasive, Brain-controlled Functional Electrical Stimulation for Locomotion Rehabilitation in Individuals with Paraplegia," *Sci. Rep.*, 2019, doi: 10.1038/s41598-019-43041-9.

[62] Choi et al., "Functional Electrical Stimulation Controlled by Motor Imagery Brain-Computer Interface for Rehabilitation," *Brain Sci.*, 2020, doi: 10.3390/brainsci10080512.

[63] Chung et al., "Therapeutic effects of brain-computer interface-controlled functional electrical stimulation training on balance and gait performance for stroke: A pilot randomized controlled trial," *Medicine (Baltimore)*, 2020, doi: 10.1097/MD.0000000000022612.

[64] Miao et al., "BCI-Based Rehabilitation on the Stroke in Sequela Stage," *Neural Plast.*, 2020, doi: 10.1155/2020/8882764.

[65] Murovec et al., "Preliminary Results of a Brain-Computer Interface System based on Functional Electrical Stimulation and Avatar Feedback for Lower Extremity Rehabilitation of Chronic Stroke Patients," in *IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2020, doi: 10.1109/SMC42975.2020.9283200.

[66] Sebastián-Romagosa et al., "Brain Computer Interface Treatment for Motor Rehabilitation of Upper Extremity of Stroke Patients - A Feasibility Study," *Front. Neurosci.*, 2020, doi: 10.3389/fnins.2020.591435.

[67] de Castro-Cros et al., "Effects of Gamification in BCI Functional Rehabilitation," *Front. Neurosci.*, 2020, doi: 10.3389/fnins.2020.00882.

[68] Broniera Junior et al., "EEG-FES-Force-MMG closed-loop control systems of a volunteer with paraplegia considering motor imagery with fatigue recognition and automatic shut-off," *Biomed. Signal Process. Control*, 2021, doi: 10.1016/j.bspc.2021.102662.

[69] Cajigas et al., "Implantable brain-computer interface for neuroprosthetic-enabled volitional hand grasp restoration in spinal cord injury," *Brain Commun.*, 2021, doi: 10.1093/braincomms/fcab248.

[70] Jovanovic et al., "Characterizing the stimulation interference in electroencephalographic signals during brain-computer interface-controlled functional electrical stimulation therapy," *Artif. Organs*, 2021, doi: 10.1111/aor.14059.

[71] Jovanovic et al., "Brain computer interface-triggered functional electrical stimulation therapy for rehabilitation of reaching and grasping after spinal cord injury: a feasibility study," *Spinal Cord Ser. Cases*, 2021, doi: 10.1038/s41394-020-00380-4.

[72] Jovanovic et al., "KITE-BCI: A brain-computer interface system for functional electrical stimulation therapy," *J. Spinal Cord Med.*, 2021, doi: 10.1080/10790268.2021.1970895.

[73] Remsik et al., "Ipsilesional Mu Rhythm Desynchronization Correlates With Improvements in Affected Hand Grip Strength and Functional Connectivity in Sensorimotor Cortices Following BCI-FES Intervention for Upper Extremity in Stroke Survivors," *Front. Hum. Neurosci.*, 2021, doi: 10.3389/fnhum.2021.725645.

[74] Sinha et al., "Brain-Computer Interface Training With Functional Electrical Stimulation: Facilitating Changes in Interhemispheric Functional Connectivity and Motor Outcomes Post-stroke," *Front. Neurosci.*, 2021, doi: 10.3389/fnins.2021.670953.

[75] Zulauf-Czaja et al., "On the way home: a BCI-FES hand therapy self-managed by sub-acute SCI participants and their caregivers: a usability study," *J. NeuroEng. Rehabil.*, 2021, doi: 10.1186/s12984-021-00838-y.

[76] Hashimoto et al., "Portable rehabilitation system with brain-computer interface for inpatients with acute and subacute stroke: A feasibility study," *Assist. Technol.*, 2022, doi: 10.1080/10400435.2020.1836067.

[77] Kim et al., "Brain-computer interface-based action observation combined with peripheral electrical stimulation enhances corticospinal excitability in healthy subjects and stroke patients," *J. Neural Eng.*, vol. 19, no. 3, 036039, 2022, doi: 10.1088/1741-2552/ac76e0.

[78] Krueger et al., "Functional electrical stimulation driven by a brain-computer interface in acute and subacute stroke patients impacts beta power and long-range temporal correlation," in *IEEE Int. Conf. Metrology for Extended Reality, AI and Neural Engineering (MetroXRAINE) / COMPENG*, 2022, doi: 10.1109/COMPENG50184.2022.9905448.

[79] Kumari et al., "Short term priming effect of brain-actuated muscle stimulation using bimanual movements in stroke," *Clin. Neurophysiol.*, 2022, doi: 10.1016/j.clinph.2022.03.002.

[80] Remsik et al., "BCI-FES With Multimodal Feedback for Motor Recovery Poststroke," *Front. Hum. Neurosci.*, 2022, doi: 10.3389/fnhum.2022.725715.

[81] Wang et al., "Multimodal Neural Response and Effect Assessment During a BCI-Based Neurofeedback Training After Stroke," *Front. Neurosci.*, 2022, doi: 10.3389/fnins.2022.884420.

[82] Zhan et al., "EEG-Based Brain Network Analysis of Chronic Stroke Patients After BCI Rehabilitation Training," *Front. Hum. Neurosci.*, 2022, doi: 10.3389/fnhum.2022.909610.

[83] Bigoni et al., "Novel personalized treatment strategy for patients with chronic stroke with severe upper-extremity impairment: The first patient of the AVANCER trial," *Med (Cell Press)*, 2023, doi: 10.1016/j.medj.2023.06.006.

[84] Cantillo-Negrete et al., "Upper Limb Recovery in Cervical Spinal Cord Injury After a Brain-Computer Interface Controlled Functional Electrical Stimulation Intervention," *J. Med. Biol. Eng.*, 2023, doi: 10.1007/s40846-023-00824-w.

[85] Fiorin et al., "The Learning Curve of People with Complete Spinal Cord Injury Using a NESs-FESs Interface in the Sitting Position: Pilot Study," *Eng*, 2023, doi: 10.3390/eng4020097.

[86] Ramirez-Nava et al., "Functional electrical stimulation therapy controlled by a P300-based brain-computer interface, as a therapeutic alternative for upper limb motor function recovery in chronic post-stroke patients. A non-randomized pilot study," *Front. Neurol.*, 2023, doi: 10.3389/fneur.2023.1221160.

[87] Sebastián-Romagosa et al., "Brain-computer interface treatment for gait rehabilitation in stroke patients," *Front. Neurosci.*, 2023, doi: 10.3389/fnins.2023.1256077.

[88] Brunner et al., "Brain computer interface training with motor imagery and functional electrical stimulation for patients with severe upper limb paresis after stroke: a randomized controlled pilot trial," *J. NeuroEng. Rehabil.*, 2024, doi: 10.1186/s12984-024-01304-1.

[89] Herring et al., "Reconnecting the Hand and Arm to the Brain: Efficacy of Neural Interfaces for Sensorimotor Restoration After Tetraplegia," *Neurosurgery*, 2024, doi: 10.1227/neu.0000000000002769.

[90] Kumari et al., "Motor priming to enhance the effect of physical therapy in people with spinal cord injury," *J. Spinal Cord Med.*, 2024, doi: 10.1080/10790268.2024.2317011.

[91] Luo et al., "Effects of motor imagery-based brain-computer interface-controlled electrical stimulation on lower limb function in hemiplegic patients in the acute phase of stroke: a randomized controlled study," *Front. Neurol.*, 2024, doi: 10.3389/fneur.2024.1394424.

[92] Bastos-Filho et al., "Upper-Limb Rehabilitation in Chronic Stroke Using Brain-Computer Interface Based on Motor Imagery plus Non-Invasive Brain and Muscular Stimulation," in *IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2025, doi: 10.1109/SMC58881.2025.11342488.

[93] Bastos-Filho et al., "Rehabilitation of Chronic Stroke Using Neurofeedback, Functional Electrical Stimulation and Cerebrospinal Direct Current Stimulation," in *Int. Conf. Rehabilitation Robotics (ICORR)*, 2025, doi: 10.1109/ICORR66766.2025.11063073.

[94] Mansour et al., "A clinical trial evaluating feasibility and acceptability of a brain-computer interface for telerehabilitation in patients with stroke," *J. NeuroEng. Rehabil.*, 2025, doi: 10.1186/s12984-025-01607-x.

[95] S. Hétu, M. Grégoire, A. Saimpont, M. P. Coll, F. Eugène, P. E. Michon, and P. L. Jackson, "The neural network of motor imagery: An ALE meta-analysis," *Neurosci. Biobehav. Rev.*, vol. 37, no. 5, pp. 930–949, 2013, doi: 10.1016/j.neubiorev.2013.03.017.

Comparison of Oxidative Stress-Induced Damage in Neuronal and Neuro-Astrocytic Models of Alzheimer's Disease

Luiza de Andrade Giraldi
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0009-0009-1305-3084

Geisa Rodrigues Salles
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0000-0003-4336-1634

Erick José Nogueira de Andrade
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0009-0007-9504-4856

Cristina Pacheco-Soares
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0000-0002-0572-074X

Abstract*— Alzheimer's disease (AD), the leading cause of dementia worldwide, has its underlying mechanisms incompletely understood. Although *in vitro* neuronal models have been widely employed to investigate AD pathophysiology, they fail to reproduce the complexity of neuron-glia interactions that play a critical role in disease progression. In this study, we compared two *in vitro* models of AD to investigate the contribution of astrocytes to the cellular response to AD-related oxidative stress, (i) a neuron-like monoculture composed of differentiated human neuroblastoma cells (SH-SY5Y) and (ii) a neuro-astrocytic co-culture consisting of differentiated SH-SY5Y and human glioblastoma cells (U87-MG), respectively. Oxidative stress was induced by hydrogen peroxide (H_2O_2) exposure to mimic a key pathological feature of AD. Cellular responses were evaluated by cell viability assays, bright-field morphology, and Congo Red staining to qualitatively assess amyloidogenic aggregate deposition. H_2O_2 exposure significantly compromised cell viability and induced morphological alterations in both models. However, the neuro-astrocytic co-culture exhibited greater preservation of cellular viability, cellular organization, and intercellular connectivity, as well as a qualitative tendency toward reduced Congo Red-positive aggregate deposition compared with the neuronal monoculture. These findings suggest that the presence of astrocyte-like cells mitigates oxidative stress-induced cellular damage, reinforcing the importance of neuron-glia interactions in AD pathophysiology. Furthermore, the proposed co-culture model provides a more physiologically relevant *in vitro* platform for investigating disease mechanisms and may facilitate the development of more representative experimental models for evaluating potential therapeutic approaches for AD.

Keywords*— *Alzheimer's disease, oxidative stress, SH-SY5Y, U87-MG, co-culture.

I. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide [1]. As with other age-related neurodegenerative diseases, the increasing life expectancy of the global population is expected to contribute to a rise in its incidence over the coming years [2].

AD is characterized by a series of neuropathological alterations that progressively impair brain structure and function. Among the most well-established pathological features are oxidative stress, resulting from an imbalance between reactive oxygen species (ROS) production and the cellular antioxidant defense system, and the deposition of senile plaques composed of amyloid-beta (Aβ) aggregates [3]. These events are closely interconnected, as Aβ accumulation triggers inflammatory responses and enhances ROS generation, while increased oxidative stress further promotes Aβ aggregation. This self-perpetuating relationship contributes to mitochondrial dysfunction, impaired energy metabolism, and neuronal degeneration, establishing a sustained pathological cycle [4], [5], [6].

In this context, experimental models capable of reproducing the main pathophysiological features of AD are essential for investigating disease mechanisms and evaluating potential therapeutic strategies. Among the available *in vitro* models, differentiated human neuroblastoma SH-SY5Y cells represent a widely used neuronal model. Upon differentiation into a neuron-like phenotype through retinoic acid exposure [7], these cells reproduce key aspects of AD when subjected to oxidative stress-inducing agents, such as hydrogen peroxide (H_2O_2), while also offering the advantages of easy culture and manipulation [8], [9], [10].

However, despite their widespread use, neuron-like monocultures do not fully reproduce the complexity of brain tissue. Astrocytes, traditionally regarded as passive support cells, play an active role in AD neurobiology by modulating neuronal energy metabolism, maintaining ionic homeostasis, participating in neurotransmission, and contributing to the neuroinflammatory response. They also play a central role in the clearance of Aβ deposits [11], [12]. Nevertheless, under chronic stress conditions and during AD progression, astrocytes acquire a reactive phenotype that is closely associated with chronic inflammation and cognitive decline. Reactive astrocytes, particularly those surrounding Aβ

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) (Grants 152384/2024-3, 110869/2025-7, and 404952/2025-5), and the Coordination for the Improvement of Higher Education Personnel (CAPES) (Grant 88887.271538/2026-00).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

plaques, may contribute to the pathological accumulation of Aβ by promoting its production and deposition [13], [14].

Comparing neuronal monocultures with neuro-astrocytic co-cultures may therefore provide important insights into the contribution of astrocytes to oxidative stress-induced cellular damage and amyloidogenic processes involved in AD. In this regard, the development of a neuro-astrocytic co-culture composed of differentiated SH-SY5Y cells as a neuronal model and U87-MG cells as a human astrocyte-like model provides a more physiologically relevant *in vitro* microenvironment, allowing the investigation of neuron-glia interactions involved in AD pathogenesis [15].

Therefore, the present study aimed to characterize two *in vitro* models of AD: (i) a neuron-like monoculture composed of differentiated SH-SY5Y cells and (ii) a neuro-astrocytic co-culture consisting of SH-SY5Y and U87-MG cells, both exposed to H_2O_2 to induce oxidative stress characteristic of AD. The physiological responses of these models to oxidative stress were evaluated through cell viability, morphological analysis, and assays indicative of Aβ deposition, providing a methodological basis for the development of more complex and physiologically relevant models of AD.

II. METHODOLOGY

A. Cell Culture

SH-SY5Y cells were cultured in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F12), whereas U87-MG cells were maintained in Dulbecco's Modified Eagle Medium (DMEM). Both culture media were supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Cells were maintained in 25 cm^2 culture flasks at 37 °C in a humidified atmosphere containing 5% CO_2, and the culture medium was replaced every 2–3 days to ensure adequate cell growth and nutrient supply.

B. Cell Seeding and Differentiation

Upon reaching confluence, cells were trypsinized and seeded into 24-well plates for the following models.

For the monoculture, SH-SY5Y cells were seeded in DMEM/F12 supplemented with 10% FBS and 1% penicillin/streptomycin at a final density of 1×10^5 cells per well.

For the co-culture, SH-SY5Y and U87-MG cells were seeded in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin at a 1:1 ratio, corresponding to 5×10^4 cells of each cell line per well, resulting in a total density of 1×10^5 cells per well.

Cells were incubated for 24 h at 37 °C in a humidified atmosphere containing 5% CO_2. After this period, the culture medium was replaced with differentiation medium.

For the monoculture, neuronal differentiation was induced using DMEM/F12 supplemented with 0.5% FBS and 0.01% retinoic acid (RA).

For the co-culture, differentiation toward a neuro-astrocytic phenotype was performed using DMEM supplemented with 0.5% FBS and 0.1% RA.

Cells were maintained in differentiation medium for 3–5 days until complete differentiation.

The monoculture and co-culture models were established using different culture media and RA concentrations according to the differentiation conditions adopted for each experimental model.

C. Oxidative stress induction and experimental design

Following differentiation, cultures were exposed or not to 200 µM H_2O_2 diluted in their respective differentiation media for 1 h to induce oxidative stress. Control groups consisted of differentiated cells maintained in the corresponding differentiation medium without H_2O_2 exposure.

The experimental groups were as follows:

- Neuron-like
- Neuron-like + H_2O_2
- Neuron-like + Astrocyte-like
- Neuron-like + Astrocyte-like + H_2O_2

The assays were performed in two independent experiments, with each experimental condition evaluated in triplicate.

D. Cell viability assay

Cell viability was assessed using the Alamar Blue assay. Following oxidative stress induction, the differentiation medium was replaced with a solution containing 10% resazurin prepared in the corresponding differentiation medium. Cells were incubated for 2 h at 37 °C in a humidified atmosphere containing 5% CO_2, protected from light.

After incubation, the solution from each well was transferred to a new microplate, and fluorescence was measured using a microplate reader (bottom reading mode; excitation: 544 nm; emission: 590 nm). A solution containing 10% resazurin without cells was incubated under identical conditions and used as the negative control. Results were expressed as the percentage of the respective untreated control.

E. Morphological analysis

To evaluate cellular differentiation and the morphological alterations induced by oxidative stress, cells were fixed with 4% paraformaldehyde (PFA) in phosphate-buffered saline (PBS) for 10 min at room temperature. Photomicrographs were then acquired using a bright-field and phase-contrast inverted optical microscope to qualitatively assess neurite distribution and cellular morphology.

F. Amyloid-beta detection

To evaluate Aβ accumulation, previously fixed cells were incubated with 0.5% Congo Red solution prepared in PBS for 15 min. Subsequently, cells were washed three times with PBS to remove excess dye.

Bright-field images were acquired using an inverted phase-contrast microscope, allowing the visualization of Congo Red-positive regions indicative of Aβ deposits. Aβ accumulation was qualitatively assessed based on the stained areas.

G. Statistical analysis

Data are expressed as mean ± standard deviation (SD). Statistical analyses were performed using GraphPad Prism 5.0. Differences between the control and H_2O_2-treated groups within each experimental model were assessed using an unpaired Student's t-test. A p-value < 0.05 was considered statistically significant. Graphs were generated using OriginPro 9.

III. Results and Discussion

A. Effects of oxidative stress on cell viability

Cell viability, assessed using the Alamar Blue assay (Fig. 1), was significantly reduced following exposure to 200 µM H_2O_2 in both experimental models compared with their respective untreated controls. This result supports the use of H_2O_2 as an oxidative stressor capable of impairing cellular metabolic activity and inducing cytotoxicity [8], [9].

An important methodological consideration is that the neuronal monoculture and neuro-astrocytic co-culture were established using different culture media and RA concentrations. Therefore, the potential contribution of these experimental conditions to the differences observed between the models cannot be completely excluded. In the neuronal monoculture, viability decreased to approximately 56% of the control value ($p = 0.0018$), whereas the neuron–astrocyte co-culture retained approximately 74% viability after H_2O_2 exposure ($p = 0.0161$). Although oxidative stress affected both models, the smaller numerical reduction observed in the co-culture suggests protection to H_2O_2-induced cytotoxicity in the presence of astrocyte-like cells. This observation is consistent with studies indicating that neuronal monocultures are highly susceptible to oxidative and neurotoxic stressors when deprived of glial support [15].

The greater viability retained by the neuron–astrocyte co-culture may be associated with the physiological functions of astrocytes in redox homeostasis, antioxidant defense, metabolic support, and maintenance of neuronal survival under moderate stress conditions [16], [17]. However, because the Alamar Blue assay reflects the overall metabolic activity of the co-culture, the contribution of each cell type cannot be determined separately. Therefore, the results indicate increased resilience of co-culture. However, under sustained or severe oxidative stress, astrocytes may also undergo redox and metabolic dysfunction, which can compromise their homeostatic and neuroprotective functions and contribute to pathological processes associated with AD [13].

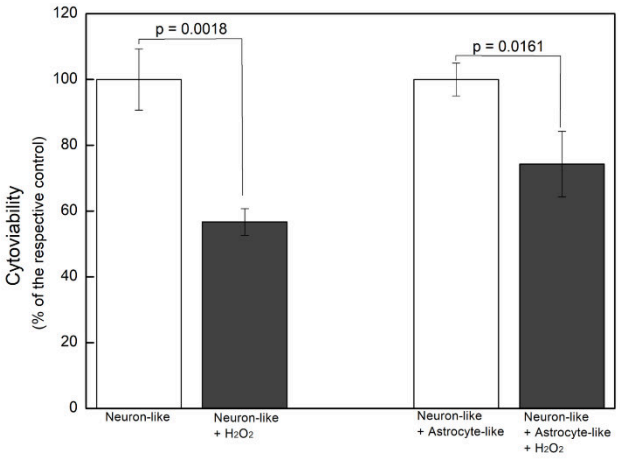


Fig. 1. Cell viability of differentiated SH-SY5Y neuron-like monocultures and SH-SY5Y/U87-MG neuron–astrocyte co-cultures following exposure to 200 µM H_2O_2, assessed by the Alamar Blue assay. Data are presented as mean ± SD and expressed as the percentage of the respective control group.

B. Morphological alterations induced by oxidative stress

Architectural integrity and the maintenance of intercellular connectivity are fundamental determinants of central nervous system function. Bright-field microscopy qualitatively (Fig. 2) revealed morphological differences between the experimental models following H_2O_2 exposure. In neuron-like monocultures, control cells exhibited elongated neuritic processes connecting adjacent cell clusters, indicating preserved neuronal morphology. Conversely, H_2O_2 exposure promoted extensive morphological disruption, characterized by neurite retraction, loss of intercellular connectivity, increased cell rounding, and the formation of compact cellular aggregates.

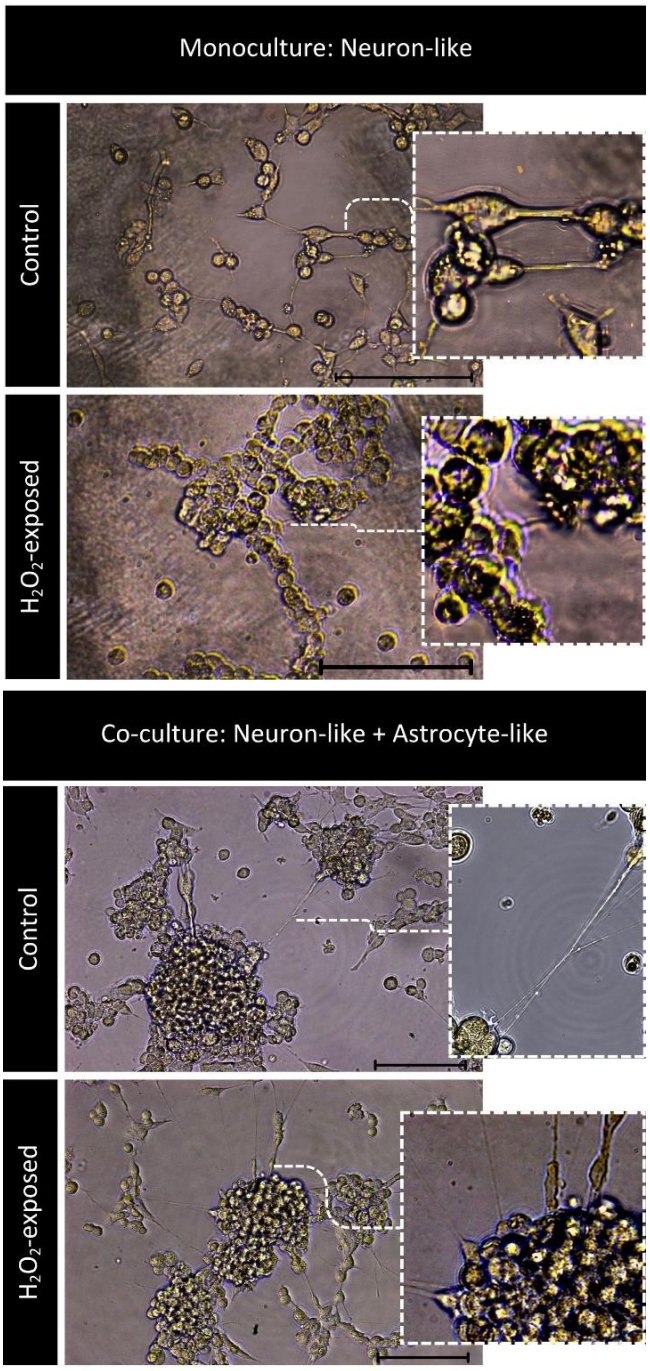


Fig. 2. Representative bright-field images of SH-SY5Y neuron-like monocultures and SH-SY5Y/U87-MG neuron–astrocyte co-cultures under control conditions and after exposure to 200 µM H_2O_2. Magnified panels highlight representative morphological features of each experimental model, including neurites in the monoculture and cellular extensions in the co-culture. Scale bars = 50 µm.

This pattern of degeneration reflects the intrinsic vulnerability of isolated neuronal populations to oxidative

stress. Under excessive reactive oxygen species (ROS) production, neurons undergo pronounced cytoskeletal depolymerization, leading to neurite collapse and disruption of neuronal networks [15]. Furthermore, acute oxidative stress severely impairs mitochondrial bioenergetic homeostasis and compromises cytoskeletal integrity in neuronal monoculture models [8], [9].

In contrast, neuron-like/astrocyte-like co-cultures displayed a more organized cellular architecture under basal conditions, with extensive cellular networks connecting large cell aggregates. Following oxidative stress, although slight morphological alterations were observed, the overall structural organization and intercellular connectivity remained largely preserved. Compared with monocultures, co-cultures exhibited reduced morphological damage after H_2O_2 exposure, suggesting that astrocyte-like cells confer structural protection against oxidative stress.

The preservation of neuritic processes and intercellular connectivity observed in the co-culture following oxidative stress is consistent with the well-established role of astrocytes in maintaining cytoskeletal homeostasis and neuronal structural integrity under cytotoxic conditions [17], [18]. Astrocytes provide extracellular matrix components and secrete neurotrophic factors that support neurite extension, synaptic organization, and the establishment of mature neuronal networks. Furthermore, astrocytes actively remodel the extracellular matrix and mitigate microenvironmental cytotoxicity, thereby contributing to the preservation of intercellular connections and limiting structural damage induced by oxidative stress [16], [19].

Overall, the comparison between the two experimental models indicates that the morphological disruption and loss of intercellular connectivity induced by H_2O_2 are attenuated in the presence of astrocyte-like cells. These findings are consistent with the established role of astrocytes in preserving cytoskeletal integrity, maintaining tissue organization, and promoting neuroprotection under oxidative stress.

C. Qualitative assessment of amyloidogenic aggregate deposition by Congo Red staining

Congo Red staining (Fig. 3) revealed qualitative differences in amyloidogenic aggregate deposition between the experimental groups. Under control conditions, both monocultures and co-cultures exhibited only sparse Congo Red-positive structures with weak staining intensity. Following H_2O_2 exposure, neuron-like monocultures showed a marked increase in Congo Red-positive aggregates, characterized by larger and more intensely stained deposits distributed throughout the cellular clusters. In contrast, although Congo Red-positive structures were also observed in H_2O_2-treated co-cultures, the staining appeared less extensive and more localized than in monocultures. These findings qualitatively suggest that astrocyte-like cells may attenuate oxidative stress-associated amyloidogenic aggregate accumulation while preserving overall cellular morphology.

The qualitatively reduced Congo Red-positive staining observed in the neuro-astrocytic co-culture is consistent with the well-established role of astrocytes in amyloid homeostasis [16]. Astrocytes produce and secrete proteolytic enzymes involved in extracellular Aβ degradation, whose expression is frequently upregulated in reactive astrocytes surrounding amyloid deposits as part of the attempt to restore the cerebral microenvironment [11], [16]. In addition, astrocytes possess an intrinsic capacity to internalize and degrade amyloid fragments through lysosomal pathways, thereby contributing to amyloid clearance [11], [20].

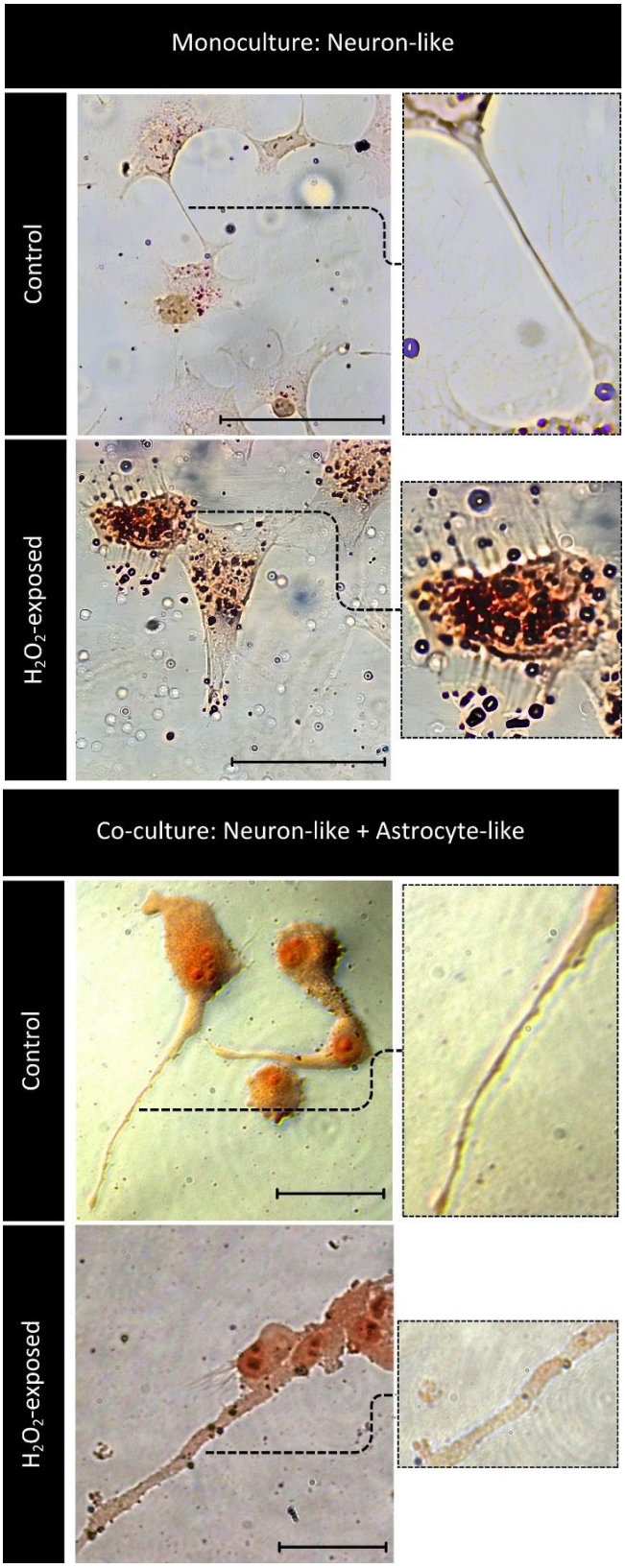


Fig. 3. Representative bright-field images of Congo Red-stained SH-SY5Y neuron-like monocultures and SH-SY5Y/U87-MG neuron–astrocyte co-cultures under control conditions and after exposure to 200 μM H_2O_2. Magnified panels highlight representative Congo Red-positive deposits observed in each experimental condition. Scale bars = 50 μm.

Beyond the direct removal or non-accumulation of extracellular aggregates, astrocytes also provide antioxidant precursors and neurotrophic factors that support neuronal homeostasis. By reducing oxidative stress within the cellular microenvironment, astrocytes may indirectly limit the oxidative processes that favor the formation and accumulation of new amyloidogenic aggregates [11], [13].

Congo Red staining was evaluated qualitatively in the present study, demonstrating aggregates morphology and arrangement within the groups. However, quantitative measurements of the aggregates, including number, location, average size and staining intensity are under analysis to provide an objective assessment of possible amyloidogenic aggregate deposition, as well as Aβ-specific antibody approaches to confirm the identity of the Congo Red-positive structures.

Collectively, this study demonstrates the first characterization steps of a novel neuro-astrocytic co-culture, compared with a conventional commonly used neuron-like monoculture, in order to produce a more physiologically relevant experimental platform. Importantly, as the co-culture system is being designed, both evaluated models differed in some cultivation conditions, to attend to both cellular requirements. Moreover, further assays are under development to validate and elucidate neurodegenerative pathways of the model and to specifically determine each cellular type response and activity within the co-culture model.

IV. Conclusions

The present study demonstrated distinct responses to oxidative stress in neuronal monoculture and neuro-astrocytic co-culture models of AD. Following H_2O_2 exposure, the co-culture showed a numerically smaller reduction in overall metabolic activity and greater preservation of cellular morphology and intercellular connectivity, together with a qualitative tendency toward reduced Congo Red-positive staining. Furthermore, by incorporating neuron–astrocyte interactions, the proposed neuro-astrocytic co-culture represents a more physiologically relevant experimental platform than conventional neuronal monocultures, representing an useful tool for investigating AD-related mechanisms and evaluating potential therapeutic strategies *in vitro*.

Acknowledgment

The authors gratefully acknowledge Prof. Dr. Leandro José Raniero and his research team for providing access to the microplate reader and for their technical assistance throughout this study.

References

[1] World Health Organization, "Dementia," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed: Jul. 26, 2026].

[2] Alzheimer's Disease International, *World Alzheimer Report 2023: Reducing dementia risk: Never too early, never too late*. London, U.K.: Alzheimer's Disease International, 2023.

[3] C. R. Jack *et al*., "NIA-AA research framework: Toward a biological definition of Alzheimer's disease," *Alzheimers Dement.*, vol. 14, no. 4, pp. 535–562, Apr. 2018, doi: 10.1016/j.jalz.2018.02.018.

[4] Z. Liu, T. Zhou, A. C. Ziegler, P. Dimitrion, and L. Zuo, "Oxidative stress in neurodegenerative diseases: From molecular mechanisms to clinical applications," *Oxid. Med. Cell. Longev.*, vol. 2017, Art. no. 2525967, 2017, doi: 10.1155/2017/2525967.

[5] A. C. Rego and C. R. Oliveira, "Mitochondrial dysfunction and reactive oxygen species in excitotoxicity and apoptosis: Implications for the pathogenesis of neurodegenerative diseases," *Neurochem. Res.*, vol. 28, no. 10, pp. 1563–1574, Oct. 2003, doi: 10.1023/A:1025682611389.

[6] L. Zuo, T. Zhou, B. K. Pannell, A. C. Ziegler, and T. M. Best, "Biological and physiological role of reactive oxygen species—the good, the bad and the ugly," *Acta Physiol.*, vol. 214, no. 3, pp. 329–348, Jul. 2015, doi: 10.1111/apha.12515.

[7] F. M. Lopes *et al*., "RA differentiation enhances dopaminergic features, changes redox parameters, and increases dopamine transporter dependency in 6-hydroxydopamine-induced neurotoxicity in SH-SY5Y cells," *Neurotox. Res.*, vol. 31, no. 4, pp. 545–559, May 2017, doi: 10.1007/s12640-016-9699-0.

[8] G. R. Salles, L. A. Giraldi, N. S. da Silva, M. A. Porcionatto, and C. Pacheco-Soares, "2D and 3D models of Alzheimer's disease: Investigating neuron-like cells in oxidative environments," *ACS Omega*, vol. 10, no. 25, pp. 27501–27514, Jul. 2025, doi: 10.1021/acsomega.5c03306.

[9] L. A. Giraldi, G. R. Salles, E. J. Nogueira de Andrade, and C. Pacheco-Soares, "Neurotherapeutic potential of photobiomodulation in a 3D model of Alzheimer's disease induced by oxidative stress," *In Vitro Models*, pp. 1–8, May 2026, doi: 10.1007/S44164-026-00122-0.

[10] J. Kovalevich and D. Langford, "Considerations for the use of SH-SY5Y neuroblastoma cells in neurobiology," *Methods Mol. Biol.*, vol. 1078, pp. 9–21, 2013, doi: 10.1007/978-1-62703-640-5_2.

[11] M. Rodríguez-Giraldo, R. E. González-Reyes, S. Ramírez-Guerrero, C. E. Bonilla-Trilleras, S. Guardo-Maya, and M. O. Nava-Mesa, "Astrocytes as a therapeutic target in Alzheimer's disease: Comprehensive review and recent developments," *Int. J. Mol. Sci.*, vol. 23, no. 21, 2022, doi: 10.3390/ijms232113630.

[12] C. R. Figley and P. W. Stroman, "The role(s) of astrocytes and astrocyte activity in neurometabolism, neurovascular coupling, and the production of functional neuroimaging signals," *Eur. J. Neurosci.*, vol. 33, no. 4, pp. 577–588, Feb. 2011, doi: 10.1111/j.1460-9568.2010.07584.x.

[13] Z. Cai, C. Q. Wan, and Z. Liu, "Astrocyte and Alzheimer's disease," *J. Neurol.*, vol. 264, no. 10, pp. 2068–2074, Oct. 2017, doi: 10.1007/s00415-017-8593-x.

[14] G. R. Frost and Y. M. Li, "The role of astrocytes in amyloid production and Alzheimer's disease," *Open Biol.*, vol. 7, no. 12, 2017, doi: 10.1098/rsob.170228.

[15] S. Park, H. N. Oh, and W. K. Kim, "Human coculture model of astrocytes and SH-SY5Y cells to test the neurotoxicity of chemicals," *Ecotoxicol. Environ. Saf.*, vol. 269, Jan. 2024, doi: 10.1016/j.ecoenv.2023.115912.

[16] S. F. Carter, K. Herholz, P. Rosa-Neto, L. Pellerin, A. Nordberg, and E. R. Zimmer, "Astrocyte biomarkers in Alzheimer's disease," *Trends Mol. Med.*, vol. 25, no. 2, pp. 77–95, Feb. 2019, doi: 10.1016/j.molmed.2018.11.006.

[17] R. E. González-Reyes, M. O. Nava-Mesa, K. Vargas-Sánchez, D. Ariza-Salamanca, and L. Mora-Muñoz, "Involvement of astrocytes in Alzheimer's disease from a neuroinflammatory and oxidative stress perspective," *Front. Mol. Neurosci.*, vol. 10, Dec. 2017, doi: 10.3389/fnmol.2017.00427.

[18] A. Ahtiainen, B. Genocchi, J. M. A. Tanskanen, M. T. Barros, J. A. K. Hyttinen, and K. Lenk, "Astrocytes exhibit a protective role in neuronal firing patterns under chemically induced seizures in neuron-astrocyte co-cultures," *Int. J. Mol. Sci.*, vol. 22, no. 23, 2021, doi: 10.3390/ijms222312770.

[19] P. Edison, "Astroglial activation: Current concepts and future directions," *Alzheimers Dement.*, vol. 20, no. 4, pp. 3034–3053, Apr. 2024, doi: 10.1002/alz.13678.

[20] Q. Yue and M. P. M. Hoi, "Emerging roles of astrocytes in blood-brain barrier disruption upon amyloid-beta insults in Alzheimer's disease," *Neural Regen. Res.*, vol. 18, no. 9, pp. 1890–1902, Sep. 2023, doi: 10.4103/1673-5374.367832.

Deep-learning for prediction of awake Slow-wave occurrences in EEG

1st Maria Fernanda Oliveira de Figueiredo
Centro de Matemática, Computação e Cognição
Universidade Federal do ABC
São Bernardo do Campo, Brasil
figueiredo.maria@ufabc.edu.br

2nd Yossi Zana
Centro de Matemática, Computação e Cognição
Universidade Federal do ABC
São Bernardo do Campo, Brasil
yossi.zana@ufabc.edu.br

***Abstract*—Awake slow-waves are transient, sleep-like cortical dynamics associated with cognitive lapses. Proposing noninvasive neuromodulation strategies for experimental procedures intended to address these events requires accurate real-time prediction from electroencephalography (EEG). This study presents the development and evaluation of two deep-learning frameworks for awake slow-wave prediction. Using a public EEG dataset comprising resting-state and cognitive task recordings, a unidimensional CNN-based architecture and a comparable model augmented with a multihead attention mechanism were implemented. Results showed that both models performed well, with the CNN-attention model yielding an equivalent performance (test accuracy: 0.82, AUC: 0.90) compared to the CNN-only model (test accuracy: 0.81, AUC: 0.89), though with higher computational demand. Statistically significant but small differences between models were observed for effect sizes. Both architectures achieved higher sensitivity than specificity. Cross-condition generalization performances were better for training under resting-state than cognitive tasks. Finally, both models may be considered promising for awake slow-wave prediction, with training time constraints leveraging the cnn-only architecture.**

***Index Terms*—EEG, slow-waves, deep learning**

I. Introduction

Although electroencephalography-detected brain waves of low frequency and high amplitude are characteristic of sleep patterns, they are also observed during wakefulness for brief moments and are conceptualized as transient intrusions of non-REM-like cortical dynamics into the waking brain, reflecting local neuronal "OFF-states" within otherwise active networks. At the cellular level, slow waves arise from bistable membrane dynamics characterized by alternating depolarized (ON) and hyperpolarized (OFF) periods, a mechanism traditionally associated with sleep homeostasis but now observed locally during wakefulness [1].

During sleep, slow-wave activity is responsible for plasticity-related neuronal processes such as strengthening of connections between neuronal networks that, in practice, contribute to memory consolidation and acquisition of skills [2], [3]. Therefore, high rates of slow-wave episodes during sleep are associated with healthy quality of sleep in adults [3]

However, local slow-waves also occur during wakefulness in a use- and state-dependent manner, increasing with time awake and reflecting regional fatigue of cortical circuits, consistent with the framework of "local sleep" [4]. Functionally, their occurrence in wakefulness has been linked to fluctuations in vigilance, attentional lapses, and mind-wandering, suggesting that they transiently disrupt task-relevant processing [5].

Recent works are introducing investigations about where such ON/OFF effects are observed during wakefulness and how sleep-like brain activity patterns affect practical contexts in daily life. For instance, Pinggal et al. [6] found that slow-waves during vigilance are more frequent in neurodivergent subjects afflicted with ADHD when compared to typical individuals. The authors propose that this effect is correlated with common factors that pose impairments to the execution of daily activities due to cognitive drawbacks experienced by such individuals.

The work by Bernasconi et al. [7] also explores the occurrence of these events in Parkinson's disease, where it is proposed to relate to hallucinations and cognitive impairment. Together, these findings support the view that wake slow waves might indicate a mesoscale marker of cortical instability and local disengagement, bridging sleep homeostasis and moment-to-moment variability in cognitive performance.

Since, in wakefulness, slow-waves usually represent events associated with lapses in cognitive performance, it is hypothesized that mechanisms for their disruption during activities of intense information processing may provide potential interventions for the symptoms provoked in ADHD and Parkinson's. This proposition is supported by experiments that evaluate the opposite effect, where non-invasive neuromodulation strategies have successfully been employed to increase slow-wave activity during NREM sleep and subsequently improve processes of memory consolidation and learning [8], [9].

The development of an experimental setting aimed at evaluating the effects of awake slow-wave disruption via noninvasive neuromodulation is a future prospect of the current work. However, in order to carry out this endeavor, the development of a system for slow-wave forecasting in electroencephalography (EEG) is made necessary. This is due to the transient nature of awake slow-waves, which differ from sleep patterns and usually occur in a very brief and localized manner.

Currently, the employment of deep-learning models for EEG event prediction represents a widely explored field of investigation. Works such as Godoy's [10] aim at predicting

pathological episodes and compare the effectiveness of various techniques — such as convolutional neural networks (CNN), recurrent networks and attention-based Transformer models — in this example, designed to forecast epilepsy episodes in EEG.

In general, slow-wave-related works tend to use neural networks to assess sleep quality and predict pathological episodes, such as the following: Halder et al. [11], who Developed combined multi-resolution feature extraction and attention mechanisms to automatically detect A-phases of the cyclic alternating pattern from single-channel EEG; Ganglberger et al. [12] who trained a hybrid CNN-LSTM model on sleep EEG from large cohorts to predict cognitive function, neurodegenerative disease, and mortality risk, aiming to derive a brain health biomarker; and Lou et al. [13] who designed a multiscale attention deep learning model to both classify sleep quality and potentially guide sleep modulation strategies based on EEG signals.

In face of such demand for a reliable predictor of awake-specific slow-wave occurrences in EEG, the current work presents the development and evaluation of two deep-learning-based models for slow-wave detection. In this sense, two frameworks are analyzed under a public EEG dataset of cognitive load: a simple CNN architecture and a more complex CNN-Attention network.

II. Methodology

A. EEG data

The EEG dataset currently employed is provided by Zyma et al. [14]. It consists of 72 recordings from 36 participants in a resting-state setting and during the process of solving arithmetic tasks of serial subtractions of numbers. From the recordings that are made accessible, each rest condition lasts three minutes, while the cognitive task lasts one minute. Consequently, in addition to the previously described class imbalance, an asymmetry in EEG conditions is also present.

Participants were in good physical and psychological conditions. The volunteers also had matched ages and were all students of the Educational and Scientific Centre “Institute of Biology and Medicine”, National Taras Shevchenko University of Kyiv.

The provided data has a sampling frequency of 500 Hz and was originally preprocessed with band-pass filters and noise reduction using standard measures. Additionally, for the present proposal, the data was filtered again to a narrower band of frequencies (1-40 Hz), aiming to better account for the slow oscillations of interest. The data was also re-referenced to channels TP9 and TP10.

EEG recordings were subdivided into 1-second windows with an overlap of 0.5 seconds and a prediction horizon of 0.5 seconds. These values are in accordance with standard procedures of automated EEG prediction for neuromodulation, which usually apply 0.5- to 1-second windows to forecast the following 0.5 sec. [15], [16].

B. Slow-wave identification

Procedures for event detection were employed in accordance with the standard literature [6], [7]. First, a Chebyshev Type II band-pass filter limited signals to 1–10 Hz. Then, the occurrence of waves in the data was defined as two consecutive zero-crossings, and peak-to-peak (PTP) amplitude was calculated. A wave was then classified as ”slow” if it: (a) presented frequency under 7 Hz and (b) PTP amplitude above the 80th percentile for a given electrode, thus defining a personal amplitude threshold for each individual.

Slow-wave events were then rejected if they occurred within 1 second of any high-amplitude event ($>$150 μV) or if the positive peak exceeded 75 μV.

It is anticipated that cortical neurons do not engage in slow-wave-related bistability states as frequently as other brain activity dynamics during wakefulness. Thus, an imbalance of identification of events and their lack in a given set of windows is expected. When the total number of slow-wave events is quantified for each context, the assumption is fulfilled. Indeed, a class imbalance is identified for both the cognitive load setting (average of 44 events in a mean of 115 windows) and resting-state (average of 90 events in 349 windows).

C. Neural Networks

In the present proposal, two deep learning models were implemented and assessed: a simpler CNN architecture and an equivalent model with the addition of an attention mechanism. This double-framework evaluation provides comparable scenarios where it is possible to factor both model performance for the current task and relative demand for computational resources.

Both neural network models were implemented in Python/Tensorflow and are composed of a depthwise 1D convolution layer, a separable 1D convolution, and two standard 1D convolution layers. All of which followed by batch normalization and dropout. Finally, two blocks composed of global average pooling followed by dense layers were added. The CNN-attention network also included a multihead attention layer and normalization before pooling.

To mitigate class imbalance and improve model generalization, an online augmentation strategy was also implemented. It generates balanced mini-batches during training while applying lightweight EEG-specific transformations on-the-fly. This approach ensures that each training batch contains an equal number of positive and negative samples, addressing the inherent class imbalance in slow-wave detection. The augmentation pipeline applies four complementary transformations to each sample with certain probability: (1) random amplitude scaling to simulate inter-subject variability in signal amplitude, (2) additive Gaussian noise to enhance robustness to measurement noise, (3) circular time-shift (up to 8% of window length) to simulate minor latency variations across recording sessions, and (4) channel dropout (10% probability of zeroing a random subset of channels) to prevent overfitting to specific electrode configurations. Additionally, temporal masking (4% of window duration) was applied stochastically

to simulate brief signal artifacts or missing data segments. The augmentation probability was set to 0.75, allowing 25% of training samples to pass through unaugmented, maintaining a balance between regularization and training signal fidelity.

For the current implementation, model training was restricted to frontal channels, which correspond to the local foci of slow-wave formation across the cortical surface [3] and therefore also represent the main target regions for noninvasive neuromodulation procedures. All models were trained until the validation AUC stabilized at the same value for at least 5 epochs.

D. Evaluation procedures

For model training and performance assessment, a subject-wise split was performed using the group K-fold method (k = 6), which is intended to mitigate learning biases associated with windows of the same subject appearing in both training and testing sets. Similarly, to reduce condition-related biases, a cross-condition validation was also performed, aiming to evaluate the different models performances under isolated brain-state settings. Analyzed metrics were accuracy, ROC-AUC, sensitivity, specificity and F1. Additionally, in order to account for the usage of computational resources, the training time of each architecture was also quantified on the group K-fold method.

To assess whether the inclusion of the attention layer effectively enhanced model performance, a standard McNemar statistical test was implemented, considering correct and incorrect predictions from the k-fold test sets (training and testing included both task and rest). Furthermore, effect sizes were estimated via paired bootstrap resampling (10,000 iterations) of test samples to quantify inter-model (cnn-att − cnn-only) differences in accuracy and AUC, with 95% confidence intervals computed from the empirical percentiles of the bootstrap distribution.

Finally, the code developmentss of the current project was aided by various AI agents.

III. Results

In general, the neural network models succeeded in predicting slow-wave events for both resting-state and cognitive task settings when both conditions were present during training. Model performances were comparable for both deep-learning architectures. General metrics are displayed in Table I

TABLE I
MEAN TEST METRICS ACROSS FOLDS FOR CNN-BASED MODELS

Model	Accuracy	AUC	Sensitivity	Specificity
CNN-Attention	0.758	0.889	0.952	0.704
CNN-only	0.782	0.882	0.852	0.707

As observed from Table I, the absence of the attention mechanism yielded slightly superior testing performance in absolute accuracy values, while a relative tradeoff in sensitivity was indicated. However, the McNemar test for model comparison did not reveal significance under $\alpha = 0.05$ ($\chi^2 = 3.161$ p = 0.075). Although the attention-aided model yielded fewer false-positives and the cnn-only architecture missed fewer slow-wave occurrences in general, the absolute differences in model performances did not result in relevant effect sizes. Effect size evaluations for group K-fold showed a marginal difference between model accuracies (mean difference: 0.019; 95% CI: [0.0130, 0.0256]; bootstrap p-value: 0.000), and AUC displayed some improvement but overall maintained a similar trend (mean difference: 0.010; 95% CI: [0.006, 0.013]; bootstrap p-value: 0.000). Although the observed effect may be considered statistically significant, the corresponding magnitude of metric differences may not be relevant for practical purposes.

Both models also performed better when trained under resting-state and evaluated under the cognitive task condition than the opposite. Specifically, model training only on the task condition was shown to be impractical, since it was not possible to generalize the slow-wave detection task to the resting-state setting. Meanwhile, model training on resting-state displayed a slight enhancement in performance metrics when compared to the mixed-condition setting (group K-fold), which might indicate a preferable methodology for practical purposes. Test metrics are specified in Table II.

TABLE II
CROSS-CONDITION TESTING METRICS OF CNN-BASED MODELS

Condition and model		Accuracy	AUC	Sensitivity	Specificity
Rest→task	CNN-Att	0.820	0.898	0.802	0.816
	CNN	0.811	0.894	0.717	0.883
Task→rest	CNN-Att	0.259	0.334	1.000	0.001
	CNN	0.260	0.847	0.999	0.002

It is also possible to visualize the confusion matrices of all model evaluations in Fig. 1

Finally, total training time differed considerably for both architectures. While the CNN-only architecture varied between 7 to 12 minutes across folds, the inclusion of the attention mechanism increased training time to a range of 19 to 43 minutes. Differences in training time are highlighted in Fig. 2

IV. Discussion and conclusion

The present proposal describes the development and evaluation of two deep-learning frameworks for awake slow-wave prediction in EEG data. The first model, a unidimensional CNN-based architecture, provided satisfying results with a rapid training process and modest demand for computational resources. The second network, similar to the first one but with the addition of an attention mechanism, yielded an equivalent performance with marginal gains in sensitivity and was accompanied by a noticeably higher demand for resources and training time.

The neural network models were set to forecast slow-wave occurrences in two EEG acquisition scenarios: resting-state and cognitive task, where model training and testing were

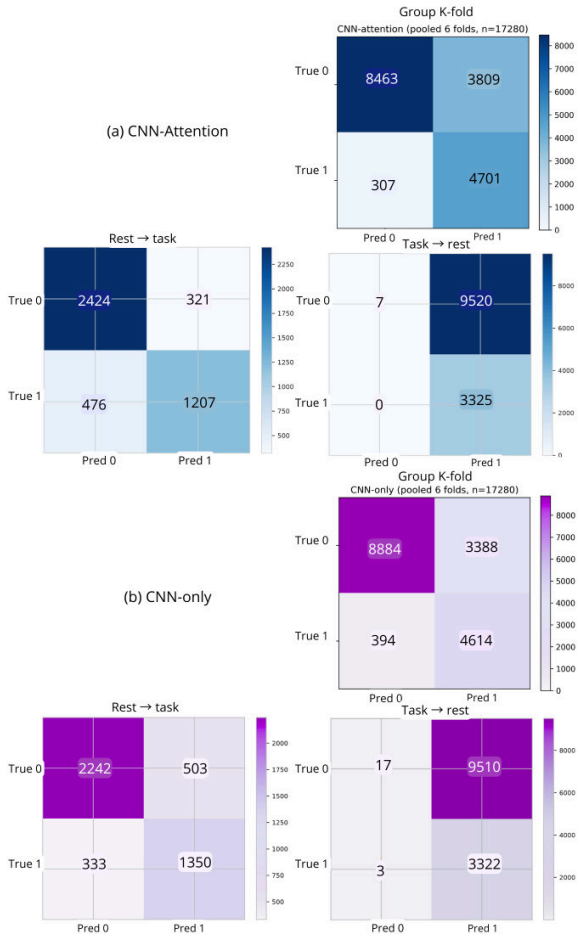


Fig. 1. Confusion matrices of both trained models and EEG contexts

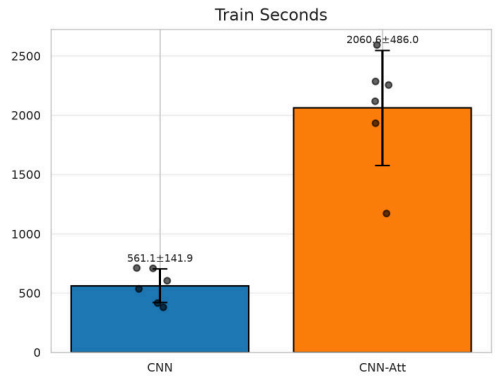


Fig. 2. Group K-fold training times

evaluated under cross-conditions. While the former scenario provides a baseline of brain activity patterns for model training, the latter represents the most relevant context for awake slow-wave prediction, since potential future applications are intended for the disruption of such events in order to attempt to reduce episodes of cognitive lapses.

In this regard, only the models that were trained on resting-state were able to generalize to the other condition. This methodology of model training further indicated superior suitability for slow-wave detection when compared to training under mixed conditions.

The observed deviation in model performances for training under the different contexts may be attributed to the overall profile of brain activity that characterizes each brain state (i.e., general low-frequency-high-amplitude signals for rest and the opposite for the cognitive task). Additionally, there is the influence of an asymmetry in sample size for the provided EEG settings since the cognitive task condition has $1/3$ the duration of resting-state. This issue was not addressed in the current project version in order to avoid hindering the training process by providing insufficient samples or biasing results by excessive data augmentation, but the employment of a larger dataset with more cognitive task samples may be considered a relevant future prospect for research.

Another potentially relevant aspect that may be considered for future enhancements in model performance is the window length provided for training. Currently, the window length has been set in accordance with the literature concerning general works of deep-learning models for EEG time-series prediction. However, an extension of each window may be considered in order to provide a more adequate foundation for model training in this particular context, since slow-wave detection might be hindered in single-second spans.

Since noninvasive neuromodulation therapies are generally safe and have no adverse effects, it is plausible to argue that, for this particular instance, the incidence of false-positives does not pose considerable harm risks if safety measures are carefully employed for experimental sessions. In this regard, it is possible to interpret the developed models' superior sensitivity metrics in comparison to specificity as an acceptable feature, although this consideration is not applicable to the task-to-rest training condition.

When comparing to performance reports of deep-learning models for EEG from the literature [10]–[13] (from 80 to around 90% accuracy, for example), the implemented models showed relatively inferior metrics for mixed conditions and comparable for rest-to-task training, which may be due to the limited sample size, especially under two conditions with considerable differences in brain activity patterns. Results could be subsequently improved in order to seek achievement of optimum reported values (accuracy around 98%) despite the differences in addressed scenarios.

In summary, this work demonstrates that both the standard CNN and the attention-augmented CNN are viable approaches for predicting awake slow-waves, with the choice between

them ultimately depending on the specific priorities of the application, namely, computational efficiency versus potential performance gains, although the present work did not display significant differences or noticeable effect sizes for the latter. The models' ability to generalize from resting-state to task-related EEG, despite the inherent challenges of class and context imbalance, underscores their potential for real-world use, in particular for the proposition and experimentation of non-invasive neuromodulation protocols. Future prospects aimed at enhancing the proposed predictors include further refining model architecture, implementing hyperparameter optimization, increasing the length of training windows, augmenting the sample size, and, mainly, evaluating performance on more datasets in order to better assess generalization capacity and mitigate the reported context asymmetry issue, especially by including more cognitive task-centered data.

ACKNOWLEDGMENT

The authors acknowledge UFABC for providing the necessary structure to carry out academic research.

REFERENCES

[1] F. Siclari and G. Tononi, "Local aspects of sleep and wakefulness," Curr. Opin. Neurobiol., vol. 44, pp. 222–227, 2017. doi: 10.1016/j.conb.2017.05.008.

[2] Rasch B, Born J. About sleep's role in memory. Physiol Rev. 2013 Apr;93(2):681-766. doi: 10.1152/physrev.00032.2012. PMID: 23589831; PMCID: PMC3768102.

[3] G. Bernardi and F. Siclari, "Local patterns of sleep and wakefulness," in *Handbook of Behavioral Neuroscience*, vol. 30, Elsevier, 2019, pp. 33–47.

[4] E. Pinggal, P. M. Dockree, R. G. O'Connell, M. A. Bellgrove, and T. Andrillon, "Pharmacological manipulations of physiological arousal and sleep-like slow waves modulate sustained attention," J. Neurosci., vol. 42, no. 43, pp. 8113–8124, 2022. doi: 10.1523/JNEUROSCI.0836-22.2022.

[5] T. Andrillon, J. Windt, T. Silk, S. P. A. Drummond, M. A. Bellgrove, and N. Tsuchiya, "Does the mind wander when the brain takes a break? Local sleep in wakefulness, attentional lapses and mind-wandering," Front. Neurosci., vol. 13, p. 949, 2019. doi: 10.3389/fnins.2019.00949.

[6] E. Pinggal, J. Jackson, A. Kusztor, D. Chapman, J. Windt, S. P. Drummond, et al., "Sleep-like slow waves during wakefulness mediate attention and vigilance difficulties in adult attention-deficit/hyperactivity disorder," The Journal of Neuroscience, vol. 46, no. 15, p. e1694252025, 2026.

[7] F. Bernasconi, T. Andrillon, J. C. Farah, I. Arnulf, S. Schwartz, J. Potheegadoo, S. Martinez-Horta, J. Pagonabarraga, J. Kulisevsky, and O. Blanke, "Sleep-like slow waves during wakefulness uncover a malignant form of Parkinson's disease," medRxiv, p. 2025.08. 29.25334276, 2025.

[8] S. Fattinger, T. de Beukelaar, K. Ruddy, et al., "Deep sleep maintains learning efficiency of the human brain," Nat. Commun., vol. 8, p. 15405, 2017. doi: 10.1038/ncomms15405.

[9] L. Marshall, H. Helgadóttir, M. Mölle, et al., "Boosting slow oscillations during sleep potentiates memory," Nature, vol. 444, pp. 610–613, 2006. doi: 10.1038/nature05278.

[10] R. Vilela de Godoy, "Predição de crises epilépticas utilizando técnicas de Deep Learning," 2021. doi: 10.11606/D.18.2021.tde-31012022-114038.

[11] B. Halder, T. Anjum, and M. I. H. Bhuiyan, "An attention-based multi-resolution deep learning model for automatic A-phase detection of cyclic alternating pattern in sleep using single-channel EEG," Biomed. Signal Process. Control, vol. 83, p. 104730, 2023.

[12] . W. Ganglberger, et al., "Brain health from sleep EEG: A multicohort, deep learning biomarker for cognition, disease, and mortality," NEJM AI, vol. 3, no. 3, p. AIoa2500487, 2026.

[13] Y. Lou, et al., "A novel sleep modulation and quality classification strategy-based EEG signals and multiscale attention deep learning," IEEE Internet Things J., vol. 12, no. 11, pp. 17377–17386, 2025. doi: 10.1109/JIOT.2025.3536200.

[14] Zyma I, Tukaev S, Seleznov I, Kiyono K, Popov A, Chernykh M, Shpenkov O. Electroencephalograms during Mental Arithmetic Task Performance. Data. 2019; 4(1):14. https://doi.org/10.3390/data4010014.

[15] H. Liu, Z. Qi, Y. Wang, Z. Yang, L. Fan, N. Zuo, and T. Jiang, "A novel real-time phase prediction network in EEG rhythm," Neurosci. Bull., vol. 41, no. 3, pp. 391–405, 2025.

[16] K. H. Madsen, A. N. Karabanov, L. G. Krohne, M. G. Safeldt, L. Tomasevic, and H. R. Siebner, "No trace of phase: Corticomotor excitability is not tuned by phase of pericentral mu-rhythm," Brain Stimul., vol. 12, no. 5, pp. 1261–1270, 2019. doi: 10.1016/j.brs.2019.05.005.

Deep Learning-Based U-Net for Urine Volume Estimation in Rats with neurogenic bladder Using Freehand Ultrasound Images

*

1st Ricardo Rique de Almeida Cantídio
Instituto Internacional de Neurociências Edmond e Lily Safra Instituto Santos Dumont (ISD)
Macaíba, Brasil
ricardo.rique@edu.isd.org.br

2nd Amanda Geórgia Diniz de Campos
Instituto Internacional de Neurociências Edmond e Lily Safra Instituto Santos Dumont (ISD)
Macaíba, Brasil
famandadiniz@gmail.com

3rd Anibal Cotrina Atencio
Departamento de Computação e Eletrônica da Universidade Federal do Espírito Santo Universidade Federal do Espírito Santo (UFES)
Vitória, Brasil
anibal.atencio@ufes.br

4th Fernando da Silva Fiorin
Instituto Internacional de Neurociências Edmond e Lily Safra Instituto Santos Dumont (ISD)
Macaíba, Brasil
fernando.fiorin@isd.org.br

5th Radan Elvis Matias de Oliveira
Departamento de Ciências Agrárias (CCA) Universidade Federal da Paraíba (UFPB)
Areia, Brasil
radan.oliveira@ufersa.edu.br

6th André Felipe Oliveira de Azevedo Dantas
Instituto Internacional de Neurociências Edmond e Lily Safra Instituto Santos Dumont (ISD)
Macaíba, Brasil
andre.dantas@isd.org.br

7th Caroline Cunha do Espirito Santo
Instituto Internacional de Neurociências Edmond e Lily Safra Instituto Santos Dumont (ISD)
Macaíba, Brasil
caroline.santo@isd.org.br

***Abstract*—Background: Neurogenic bladder (NB) after spinal cord injury (SCI) contributes to urinary tract complications, including infections. Wistar rats are widely used as experimental models for translational studies of the NB. However, the accurate assessment of bladder urine volume remains challenging. Although methods such as the Credé maneuver and catheterization using intravenous catheters can be employed to estimate bladder urine volume, both approaches have important limitations. Therefore, noninvasive approaches integrating freehand ultrasonography and artificial intelligence-based image analysis, such as U-Net deep learning models, may provide a promising strategy for accurate and reproducible estimation of bladder urine volume in NB conditions.**

Objective: This study proposes a novel, accessible (non-invasive) alternative by adapting U-Net Deep Learning to analyze 2D free-hand ultrasound images of Wistar rat bladders, estimating bladder volume and generating 3D visualizations.

Results: The pre-trained U-Net achieved high segmentation accuracy across validation images within 250 training epochs, reaching a Dice Similarity Coefficient(Measures the spatial overlap between the manual segmentation and the U-Net prediction.) of 0.8615 (Dice Loss of 0.23). The 3D Voxel Interpolation (Convex Hull) model demonstrated superior volumetric accuracy compared to the pure prolate ellipsoid model, effectively preventing severe volume overestimation in distended neurogenic bladders, with absolute errors up to 3.63 mL vs. 7.33 mL, respectively.

Conclusion: This study demonstrated the feasibility of adapting a CT-trained U-Net convolutional neural network to segment 2D freehand ultrasound images and reconstruct 3D bladder surface models in Wistar rats with NB after SCI.

***Index Terms*—Neurogenic Bladder, Spinal Cord Injury, Deep Learning, Non-invasive, Ultrasound, Wistar Rats.**

I. INTRODUCTION

Traumatic spinal cord injury (SCI) refers to spinal cord damage resulting from falls and traffic accidents. In general, SCI results in the complete or incomplete loss of sensory and/or motor function below the level of the injury, as well as various forms of autonomic dysfunction [1].

One of the consequences of SCI is neurogenic lower urinary tract dysfunction, also referred to as neurogenic bladder (NB) (Spinal Cord Injury 2016). NB remains a clinical challenge, especially in relation to the high risk of urine storage frequently observed in the presence of detrusor underactivity or overactivity combined with detrusor sphincter dyssenergia in humans and animal models of SCI [2] [3].

In SCI animal models, the Credé maneuver is necessary for bladder voiding [4], once animals develop a urine-retention bladder in the first two weeks post-injury [5]. This method can also be used to assess bladder urine volume while simultaneously performing bladder emptying required to maintain animal health. However, this method may be imprecise and may lead to variability among experimenters.

Although transurethral characterization can measure the real post-residue urine [6] in rodents, it presents a challenge to researchers because of their small size and fragility during invasive procedures [7]. Moreover, daily characterization may increase the risk of urinary tract infection, tissue injury, and pain or discomfort in animals. In this context, translational studies using rats are the most widely used models for investigating NB in SCI models [8] and new tools are needed to measure urine storage in rodents.

Ultrasonography (US) is a highly efficient, cost-effective, and non-invasive method for monitoring the urinary tract. It serves to assess bladder wall thickness, trabeculations, and the identification of sediment and vesical calculi [9]. In humans, US is useful for classifying bladder wall deformities associated with NB [10]. However, these morphological alterations do not reflect the degree of bladder filling or urinary retention. US as a measurement of vesical volume is a relatively new approach in the literature [12].

Bladder volume is conventionally estimated using geometric formulas, most notably the prolate ellipsoid model [11]. However, these formulas assume a regular geometry, which is anatomically inaccurate given the continuous morphological changes of the organ. Consequently, reconstruction methods capable of accounting for its irregular shape are necessary for reliable volume measurements [13].

In this context, the present study proposes the use of Deep Learning techniques for measuring vesical volume in the NB condition. For this purpose, a U-NET (Neural Network), originally pre-trained with human Computed Tomography (CT) scans, was adapted to work with 2D free-hand US images of Wistar rat bladders to estimate bladder volume and generate preliminary 3D visualizations.

This approach allows individuals with less professional experience (in emergencies or when a qualified professional is unavailable) or researchers with limited resources to achieve similar results, making US examination more accessible.

II. Methods

A. Ethical aspects

This preclinical study was approved by the Ethics Committee on Animal Use (CEUA) of the Santos Dumont Institute (ISD) under protocol number 2025-01-ISD.

B. Animals

Eight adult female Wistar rats (approximately 4 months old) were used. The rats were obtained from the institutional ISD animal facility. The animals were housed (3 per cage) in standard rodent cages under a 12-hour light/dark cycle, at a controlled temperature of 22 ± 2°C and 55% relative humidity, with food and water provided *ad libitum*.

C. Experimental design

After 24-hours, the animals that underwent the surgical procedure and presented normal locomotion (sham group) and those without or with minimal locomotor recovery (SCI group), according to the BBB scale [16], were included in the next steps of the study. Only for animals in the SCI group, the bladder was emptied twice a day using the credé maneuver until 21 postoperative day (POD). At 21 POD, an abdominal US examination was performed on the five sham and three SCI female rats. Following US, image data of the urinary bladders were pre-processed using ImageJ software. These images were then fed into a convolutional neural network (U-Net with CNN), originally designed to analyze bladder CT scans, to adapt it for interpreting bladder ultrasound images. Model training and evaluation were performed using a leave-one-animal-out cross-validation strategy on the base animal dataset. The overall workflow for ultrasound image processing and AI-based bladder volume estimation is illustrated in Fig1. 1.

D. Spinal Cord Injury Animal Model

Firstly, Sham and SCI animals were administered pre-anesthetic medication, including ketamine (70 mg/kg; IP) and atropine (3 mg/kg; IM). Subsequently, anesthesia was induced and maintained with inhaled isoflurane 1–5% in 2 L/min). For the SCI and Sham animals, the dorsal region was tricotomized once anesthetized. Each animal was placed in the prone position, and surgery started with a cranio-caudal incision along the dorsal region. After retracting and incising part of the thoraco-lumbar fascia, the vertebral column was exposed at the level of the eighth thoracic vertebra (Th8). The lamina of the Th8 was carefully sectioned to avoid injuring the spinal cord. In sham animals, the surgical procedure was concluded after completion of the laminectomy. In SCI animals, spinal cord compression was induced using an aneurysm clip with a calibrated closing force of 70 g applied for 30 seconds, following a method adapted from [14]. Twenty-four hours after surgery, treatment was started with the antibiotic enrofloxacin (5.0 mg/kg; SC BID), the opioid analgesic tramadol (5–10 mg/kg; SC BID), and the non-steroidal anti-inflammatory drug meloxicam (1.0 mg/kg; SC SID) for three consecutive days. In addition, the Credé maneuver began on the first POD.

E. Ultrasound image acquisition

Two days before the US examination, all animals were submitted to trichotomy in the abdominal region. To perform the US examination, all animals were anesthetized with 1-5 isoflurane % in 2 L/min). Images from the sham and

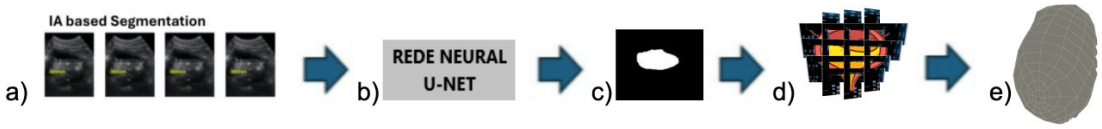


Fig. 1: Ultrasound image processing workflow using AI and expected results: a) selection of ultrasound images, b) input of images into the neural network, c) image binarization and bladder region selection, d) image overlay and organization, e) generation of a 3D bladder model for volumetric calculation and shape prediction

SCI animals were acquired at 21 POD. For recording, all animals were positioned in supine, and a linear probe (7.5 MHz) connected to a Z50 veterinary ultrasound unit (Mindray, Shenzhen, China) was applied to the abdominal region. A water-based transducer gel was used to eliminate air gaps between the probe and the skin surface, resulting in improved image resolution and allowing for accurate interpretation of anatomy, physiology, and diagnosis [15]. For each animal, seven ultrasonographic images of the urinary bladder were acquired3, including four in the transverse plane and three in the longitudinal plane. The first image in each plane was obtained at the region corresponding to the greatest bladder height and served as a reference for subsequent acquisitions. From this reference image, the remaining images were acquired at equidistant positions. In the longitudinal plane, two additional images were obtained, one parallel to the right and the other parallel to the left of the central reference image. In the transverse plane, one image was acquired cranial to the reference image, and two additional images were acquired caudal to it 2. To optimize bladder visualization, it was necessary to use sham animals with low urine volume at the time of the examination. One milliliter of saline solution was administered into the bladder via catheterization using 22 Gauge intravenous catheter (blue Jelco®. To assess post-ultrasound urinary retention. In SCI animals, transurethral catheterization was performed immediately following the US examination using a 22-Gauge intravenous catheter (Jelco®). The retained urine volume was evacuated and recorded to serve as a gold-standard reference for volumetric validation).



Fig. 2: Representation of image sections of the rat bladder.

F. Database Organization

Following US, image data (800x600 JPGs) of the urinary bladders of five Sham animals and three SCI-Animals were



Fig. 3: Seven ultrasound images of the bladder (800 × 600 pixels, RGB color format, .JPG) were used to reconstruct the three-dimensional (3D) bladder model

.

pre-processed using ImageJ software. Following the US examination, images of the bladder complete with calculated metrics—were obtained by taking screenshots of the ultrasound display using the built-in software tool. To organize the database, three folders were created on Google Drive: ”image base” (for the raw screen-shots) and ”mascara base” (for the images edited and binarized in ImageJ), with each image assigned to three-digit number based on its upload order

G. Image Editing

The US screenshots had a pixels and RGB color formatting, image editing and metric was performed using the *ImageJ* software (National Institutes of Health- NIH). Initially, a segmentation selection tool was used to measure the selected area (ROI)—defined by the bladder contour—in pixels for subsequent performance evaluation. Next, the ”flatten” command was applied to merge the segmentation and the original image into a single layer, which was then saved to a ”contour” folder used solely for comparison after neural network training. The mask was created using the ”fill” option to turn the selection interior white and the ”clean outside” function to delete the image content outside the selection. All images were saved in .jpg format 4.

H. Neural Network Training

The U-Net neural network based employed: Semantic segmentation with PyTorch library, can be used to be a previously trained image segmentation tool on urinary tract computed tomography images. To adapt it, the dataset was configured to use US image inputs and corresponding masks, aiming to produce an output contour as close as possible to the mask. To this end, 56 US base images were used; data

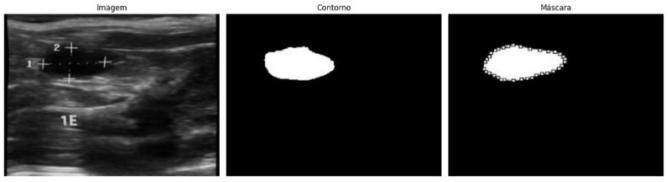


Fig. 4: (Left) Ultrasound image of the base of the rat bladder; (Center) Binarized image generated by the neural network; (Right) Image manually outlined and binarized in ImageJ for neural network feedback.

augmentation techniques were applied to expand the dataset, generating the elastic-deformed and rotated images mentioned earlier, resulting in a total of 280 images. The training process involved 250 epochs, with a Dice loss of 0.23, The Dice Similarity Coefficient (DSC) was used to evaluate the spatial overlap between the manual ground-truth segmentation by the expert and the prediction generated by the U-Net model. A DSC value closer to 1 indicates higher segmentation accuracy and agreement with the original image. To evaluate the model's accuracy and performance across all subjects, a leave-one-animal-out cross-validation approach was applied to the base animal dataset. Model performance was evaluated using the Dice coefficient, alongside a manual visual comparison between the generated contours and reference masks.

I. Statistic

The volume measured by transurethral characterization served as the gold standard for comparison against the volumes predicted by the models. In Sham animals (animals 1 to 5), the volume was standardized to 1.0 mL during ultrasound acquisition. Conversely, in SCI animals (animals 6 to 8), the volume was measured via post-ultrasound transurethral characterization to quantify retained urine (2.50 mL, 1,00 and 5.00 mL, respectively).

Height 3D reconstruction models were evaluated: (1) a **3D Convex Hull Interpolation**, in which the algorithm envelops the point cloud by creating a continuous 3D polygonal mesh (composed of multiple connected triangles) spanning the contours of neighboring and orthogonal slices; and (2) a **Pure Ellipsoid** model, in which the code extracts the absolute extreme values of the point cloud along the three spatial axes ($\max - \min$) to determine the total length, width, and height of the structure (D_x, D_y, D_z). using the Fórmula:(1):

$$V_{\text{ellipsoid}} = \frac{4}{3}\pi \cdot r_x \cdot r_y \cdot r_z = \frac{4}{3}\pi \cdot \left(\frac{D_x}{2}\right) \cdot \left(\frac{D_y}{2}\right) \cdot \left(\frac{D_z}{2}\right) \quad (1)$$

The calculated volumes from both models were converted to milliliters (mL) and evaluated against the transurethral catheterization gold standard for each animal. Performance was quantified by calculating both the absolute error (mL). Across the cohort, the 3D Voxel interpolation (Convex Hull) demonstrated superior accuracy compared to the traditional pure prolate ellipsoid model. While both methods showed low absolute errors in low-volume conditions (Sham group, A1–A5), the pure ellipsoid model significantly overestimated or distorted volume calculations in high-volume, neurogenic bladders (SCI group, A6–A8), reaching absolute errors up to 7.33 mL. In contrast, the 3D Convex Hull approach maintained higher fidelity to the actual geometry of the bladder, achieving lower absolute and percentage errors across all subjects.

III. Results and Discussion

A. Mitigating Small Dataset Limitations via Transfer Learning and Data Augmentation

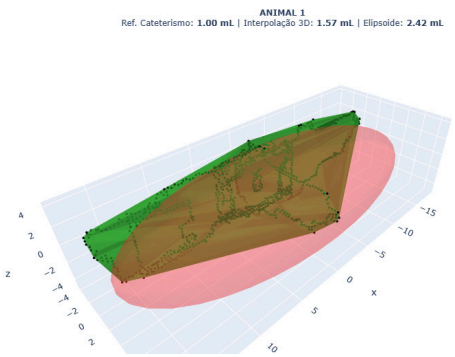
The acquisition of high-quality ultrasound image datasets in preclinical rodent models of NB presents significant logistical and anatomical challenges. In this study, the initial dataset was constrained to red 56 raw 2D freehand ultrasound images obtained from eight animals. In deep learning applications, such limited training data typically leads to severe overfitting, poor boundary segmentation, and high generalization error.

To overcome this limitation without increasing animal usage, we applied transfer learning by adapting a U-Net architecture originally pre-trained on human abdominal computed tomography (CT) scans. The pre-trained features enabled the network to leverage low-level edge and shape representations, significantly reducing the required convergence time for ultrasound image segmentation. Second, data augmentation techniques—specifically non-linear elastic deformations and random spatial rotations—were applied to expand the base dataset threefold to 280 images. This synthetic expansion effectively simulated real-world acoustic variations, transducer pressure artifacts, and anatomical morphological shifts during scanning. Consequently, the network achieved stable segmentation accuracy across unencountered validation images, reaching a Dice Loss below 0.23 within 250 training epochs.

Although this synthetic expansion successfully mitigated initial overfitting in this pilot proof-of-concept study ($N = 8$), a larger dataset remains essential to further reduce prediction errors during 2D ultrasound slice segmentation. Improving 2D boundary segmentation accuracy directly minimizes propagation errors in 3D point cloud generation, allowing more precise volume formulas and mathematical corrections when converting 2D slice contours into continuous 3D reconstructions. Future work will incorporate larger animal cohorts to evaluate broader morphological heterogeneity and refine slice-to-volume transformation algorithms.

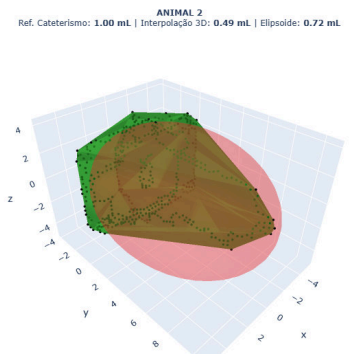
The 3D volumetric estimations derived from the two surface reconstruction approaches yielded distinct representations of the bladder across all analyzed subjects (Fig. 5):

- **3D Voxel Interpolation (Convex Hull):** Wraps the segmented point cloud in a continuous bounding polygon. While it abstracts complex surface contours (lacking exact local geometric fidelity), it maintains high volumetric fidelity by tightly enclosing the segmented slices without overestimating unobserved spaces.
- **Pure Prolate Ellipsoid Model:** Extracted extreme axial dimensions (D_x, D_y, D_z) to calculate volume, assuming a perfectly symmetrical geometry.



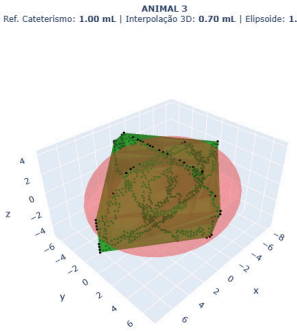
(a) Animal 1 (Sham)

Vol. Hull: 0.00 mL — Vol. Ellip: 0.72 mL



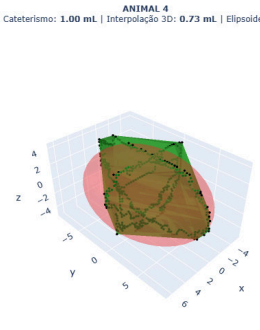
(b) Animal 2 (Sham)

Vol. Hull: 0.01 mL — Vol. Ellip: 0.78 mL



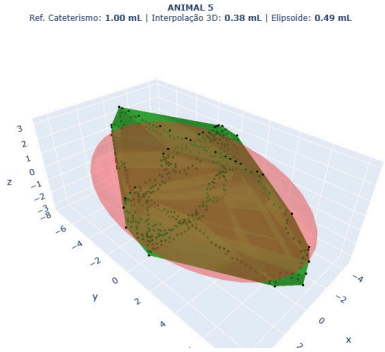
(c) Animal 3 (Sham)

Vol. Hull: 0.00 mL — Vol. Ellip: 0.55 mL



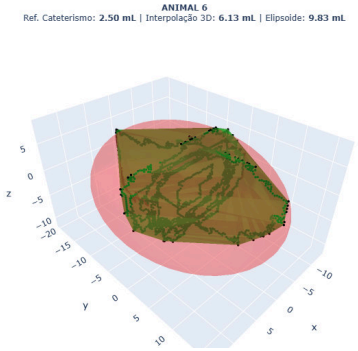
(d) Animal 4 (Sham)

Vol. Hull: 0.03 mL — Vol. Ellip: 0.72 mL



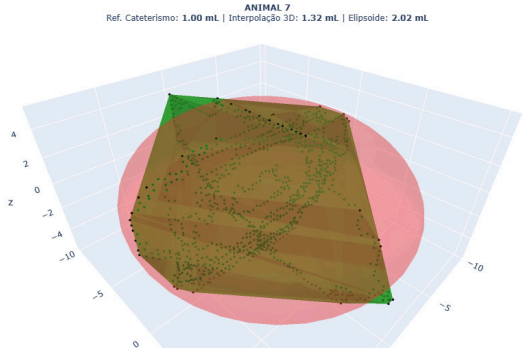
(e) Animal 5 (Sham)

Vol. Hull: 0.08 mL — Vol. Ellip: 0.38 mL



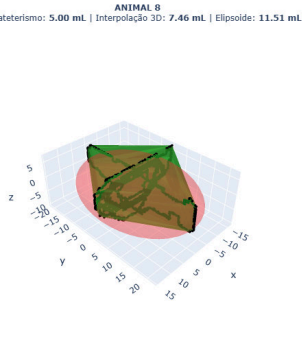
(f) Animal 6 (SCI)

Vol. Hull: 0.58 mL — Vol. Ellip: 0.73 mL



(g) Animal 7 (SCI)

Vol. Hull: 1.88 mL — Vol. Ellip: 1.10 mL



(h) Animal 8 (SCI)

Vol. Hull: 0.80 mL — Vol. Ellip: 0.44 mL

Fig. 5: 3D continuous surface reconstructions of rat bladders for Sham (Animals 1–5) and SCI groups (Animals 6–8), showing calculated Convex Hull (Vol. Hull) and Ellipsoid (Vol. Ellip) volumes.

TABLE I: Raw Comparative Absolute Error (mL) Across Animal Subjects

Animal	Ref. Vol. (mL)	Interp. Vox. (mL)	Ellipsoid (mL)	Vox. Err. (mL)	Ellip. Err. (mL)
Animal 1	1.0	1.57	2.42	0.57	1.42
Animal 2	1.0	0.49	0.72	0.51	0.28
Animal 3	1.0	0.70	1.04	0.30	0.04
Animal 4	1.0	0.73	1.00	0.27	0.00
Animal 5	1.0	0.38	0.49	0.62	0.51
Animal 6	2.5	6.13	9.83	3.63	7.33
Animal 7	1.0	1.32	2.02	0.32	1.02
Animal 8	5.0	7.46	11.51	2.46	6.51

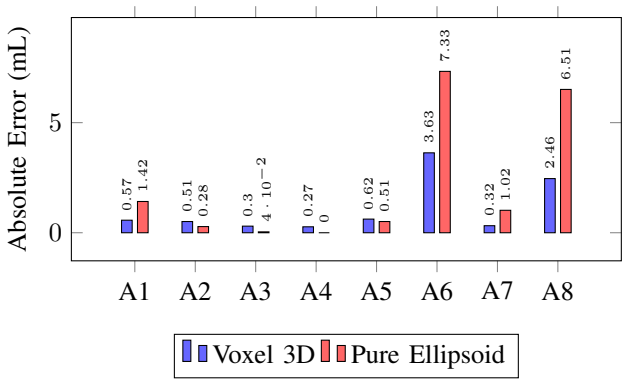


Fig. 6: Comparison of Absolute Errors (mL) between Voxel 3D Interpolation and Pure Ellipsoid methods per animal.

1) Comparison with Transurethral Catheterization (Gold Standard): Transurethral catheterization remains the clinical and experimental gold standard for measuring exact post-void residual urine volume. However, in small animal models such as Wistar rats, catheterization requires anesthesia, expert handling, and carries severe risks of urethral trauma, tissue scarring, and secondary urinary tract infections (UTIs)—factors that directly compromise longitudinal studies.

The Deep Learning-assisted ultrasound reconstruction method developed in this study provides a non-invasive, highly repeatable alternative. By mapping 2D ultrasound slices into standardized 3D continuous surface meshes ($1\ \text{cm}^3 = 1\ \text{mL}$), the proposed pipeline achieves volume estimations that closely track physical fluid infusion values without exposing fragile post-SCI rodents to invasive catheter insertion. This workflow democratizes rodent bladder evaluation, enabling reliable longitudinal monitoring even when operated by non-specialist researchers.

IV. CONCLUSION

In this study, we presented a non-invasive pipeline combining deep learning-based segmentation (U-Net) with 3D surface reconstruction algorithms (3D Voxel Interpolation via Convex Hull and Pure Prolate Ellipsoid models) to estimate rat bladder volume from 2D freehand ultrasound slices. The proposed workflow demonstrated high volumetric fidelity relative to the catheterization gold standard, effectively overcoming the severe volume overestimation limitations associated with conventional geometric formulas in distended neurogenic bladders.

These findings indicate that automated 3D ultrasound reconstruction offers a reliable, non-invasive alternative for longitudinal assessment of neurogenic bladder in rodent preclinical models. However, acquiring a larger dataset of 2D ultrasound images is essential for future work to reduce slice-level segmentation prediction errors, enabling fine-tuned mathematical volume corrections during 3D point cloud generation and further validating the model across broader anatomical variations.

REFERENCES

[1] World Health Organization. (2023). *Spinal cord injury*. WHO Fact Sheets.

[2] Krebs, J., Wöllner, J., & Kessler, T. M. (2016). Lower urinary tract dysfunction in patients with spinal cord injury. *Current Opinion in Urology*, *26*(4), 328–334.

[3] Wada, N. (2023). Animal models for neurogenic lower urinary tract dysfunction. *International Journal of Urology*, *30*(2), 112–120.

[4] S. Bauer *et al.*, "Translational Pitfalls in SCI Bladder Research: The Hidden Role of Urinary Drainage Techniques in the Rat Model," *Biology*, vol. 14, no. 8, p. 928, July 2025, doi: 10.3390/biology14080928.

[5] David, S., & Stewart, A. (2010). Bladder management and urinary retention during acute phase spinal cord injury in rats. *Journal of Neurotrauma*, *27*(8), 1501–1509.

[6] P. P. Smith, E. Hurtado, C. P. Smith, T. B. Boone, and G. T. Somogyi, "Comparison of cystometric methods in female rats," *Neurourology and Urodynamics*, vol. 27, no. 4, pp. 324–329, Sept. 2007, doi: 10.1002/nau.20512.

[7] N. KILIÇ, N. KILIÇ, B. BAYDOĞAN, C. YİĞİT, and S. YİĞİT, "Comparative Evaluation of the Effects of Two Anesthetic Combinations and Administration Routes on Lacrimal Secretion in Gerbils.," *Medycyna Weterynaryjna*, vol. 82, no. 11, pp. 554–564, 2026, doi: 10.21521/mw.7187.

[8] K. Bai *et al.*, "A New Rat Model of Sacral Cord Injury Producing a Neurogenic Bladder and Its Functional and Mechanistic Studies," *Biomolecules*, vol. 14, no. 9, p. 1141, Sept. 2024, doi: 10.3390/biom14091141.

[9] Matsuoka, Kanako *et al.* Noninvasive monitoring of lower urinary tract dysfunction in rats with bladder outlet obstruction using uroflowmetry and ultrasonography. International Urology and Nephrology, v. 58, n. 3, p. 873–880, 25 ago. 2025.

[10] Zanini, Jovita Lane Soares Santos *et al.* Dynamic ultrasound evaluation of morphological and functional urinary tract characteristics in newborns and young infants with myelomeningocele. Pediatric Radiology, v. 56, n. 2, p. 403–414, 6 nov. 2025.

[11] B. Padmapriya, T. Kesavamurthy, M. S. Thamalari Selvi, and A. Gayathri, "An Intensive Review on the Noninvasive Methods to Measure the Urinary Bladder Volume Using Ultrasound Images," *Applied Mechanics and Materials*, vol. 573, pp. 803–807, June 2014, doi: 10.4028/www.scientific.net/amm.573.803.

[12] T. Ogawa, "Bladder Deformities in Patients with Neurogenic Bladder Dysfunction," *Urologia Internationalis*, pp. 59–62, Feb. 2010, doi: 10.1159/000282252.

[13] Z. Peng *et al.*, "Weakly supervised learning-based 3D bladder reconstruction from 2D ultrasound images for bladder volume measurement," *Medical Physics*, vol. 51, no. 2, pp. 1277–1288, July 2023, doi: 10.1002/mp.16638. .

[14] do Espírito Santo, C., Silva, R., & Oliveira, M. (2019). Clip-compression model of spinal cord injury in Wistar rats: Surgical standardizations. *Journal of Neuroscience Methods*, *320*, 41–50.

[15] T. Uehara *et al.*, "Feasibility study of volumetric modulated arc therapy with HalcyonTM linac for total body irradiation," *Radiation Oncology*, vol. 16, no. 1, Dec. 2021, doi: 10.1186/s13014-021-01959-3. .

[16] Basso, D. M., Beattie, M. S., & Bresnahan, J. C. (1995). A sensitive and reliable locomotor rating scale for open field testing in rats. *Journal of Neurotrauma*, *12*(1), 1–21.

Dense recordings in the Posterior Parietal Cortex

1st Anne Beatriz Vieira
Department of Biomedical Engineering
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
anne.vieira@ufpe.br

2nd Johseph P. G. Souza
Department of Biomedical Engineering
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
johseph.souza@ufpe.br

3rd Rubem C. A. Guedes
Department of Nutrition
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
rubem.guedes@ufpe.br

4th Sonia Carolina Guerrero Prieto
Department of Nutrition
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
0000-0003-1145-4350

5th Pedro V. Carelli
Department of Physics
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
ORCID: 0000-0002-5666-9606

6th Leandro A. A. Aguiar
Department of Physics
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
ORCID: 0000-0002-8195-4601

7th Nivaldo A. P. de Vasconcelos
Department of Biomedical Engineering
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brazil
ORCID: 0000-0002-0472-8205

***Abstract*—The posterior parietal cortex (PPC) is an associative cortical region involved in sensorimotor integration, spatial processing, and decision-related activity. Here, we characterized single-unit and population spiking activity in the anesthetized rat PPC using high-density extracellular recordings with a multi-shank silicon probe. Spike sorting and manual curation were used to identify single-unit activity (SUA) and multi-unit activity (MUA) clusters. For SUA units, we quantified firing rate, inter-spike interval metrics, burstiness, and waveform features. Population activity was analyzed across shanks by pooling SUA and MUA spikes, binning activity at 50 ms resolution, and computing the coefficient of variation (CV) of spike counts over time. A total of 186 SUA units and 703 SUA+MUA sorted units were included. SUA firing rates were broadly distributed, with a median firing rate of 0.62 Hz, and waveform analysis identified 78 narrow-spiking and 100 wide-spiking SUA units. Population CV analyses revealed temporally structured variability across shanks and animals, with example raster plots showing patterns consistent with UP/DOWN-like activity patterns under anesthesia. These results provide an initial quantitative description of unit-level diversity and population-level dynamics in the rat PPC and support the use of high-density recordings to characterize associative cortical circuits.**

***Index Terms*—cortical dynamics, high-density recordings, neuronal population dynamics, extracellular electrophysiology**

I. Introduction

Information processing in the mammalian brain emerges from the coordinated activity of neuronal populations distributed across hierarchically organized cortical circuits [31], [32]. Along this hierarchy, cortical areas differ in their functional order: lower-order regions process specific sensory or motor signals, whereas higher-order associative areas combine them to support flexible behavior [34].

Within the cortical hierarchy, the posterior parietal cortex (PPC) serves as a critical associative node that actively synthesizes incoming sensory, spatial, and motor variables [35], [36]. By acting as a functional bridge between sensory processing streams and motor command networks, the PPC plays a central role in attention, spatial representation, action planning, and decision-making, serving as a critical interface between sensory and motor networks. In rodents, PPC activity has been shown to encode spatial location, body position, movement intentions, tactile sensorimotor information, and choice-related signals that support goal-directed behavior [6], [7]. Crucially, these diverse signals are processed through distributed computations that unfold across thousands of interconnected neurons. Consequently, moving beyond single-cell properties to map how these population-level representations are structured in neural state-space is essential to deciphering how the PPC transforms sensory inputs into goal-directed motor outputs.

Finep, CNPq

Beyond advancing basic systems neuroscience, decoding this structured population activity is of profound practical interest for neuroengineering. Neurofeedback provides a closed-loop framework in which subjects learn to volitionally regulate their own neural activity from real-time feedback of a target signal [37]. Given its integrative role in intention and movement planning, the PPC is an attractive target for such closed-loop control, as its population activity can, in principle, be decoded and fed back to support self-regulation and neuroprosthetic applications.

Accessing this population-level organization, however, depends on the available recording methods. While classical studies have greatly advanced our understanding of single-neuron properties, recent work in extracellular electrophysiology has shown that many neural functions depend on the collective dynamics of large neuronal ensembles [1], [2]. High-density silicon probes now enable the simultaneous recording of dozens to hundreds of neurons, offering unprecedented access to population-level dynamics and the functional organization of cortical networks [8], [9], while advances in spike-sorting methods have improved the reliable identification and classification of neuronal units from large-scale recordings [10], [11].

Together, these advances make high-density silicon probes particularly well suited to studying the PPC, where the coordinated activity of large neuronal populations is thought to underlie sensory integration, decision-making, and motor planning. In the present study, the recordings primarily sampled deep cortical layers (V and VI), which contain projection neurons involved in corticocortical, corticothalamic, and subcortical communication, making them particularly relevant for investigating population activity in this associative cortical region [12]–[14].

Therefore, the present study aimed to characterize neuronal population activity in the rat posterior parietal cortex using high-density silicon probe recordings. By combining extracellular electrophysiological recordings with spike-sorting procedures, we sought to provide a detailed description of neuronal activity across deep cortical layers, contributing to a better understanding of neuronal dynamics within this important associative cortical region.

II. Materials and methods

A. Subjects

A total of six male Wistar rats (*Rattus norvegicus*) aged 3–4 months and weighing between 350 and 500 g were included in the study.

The rat was obtained from the Central Vivarium of the Federal University of Pernambuco (Recife, Brazil). The animal was housed using a 12-h light/dark cycle and had free access to water and standard laboratory chow. The project was approved by the Animal Use Ethics Committee (CEUA) of the UFPE (Protocol N^{o} 0014/2025) and all procedures were conducted in accordance with established guidelines for the care and use of laboratory animals.

B. Electrophysiology

In vivo experiments were performed to record neural activity from the posterior parietal cortex in animals anesthetized. Anesthesia was induced with ketamine (150 mg/kg) and xylazine (5 mg/kg), using concentrations of 100 mg/ml (ketamine) and 20 mg/ml (xylazine). An initial dose was administered, followed by supplemental doses of ketamine throughout the surgical procedure as required.

The animals were placed in a stereotaxic frame to ensure stability during the surgical procedure. Lidocaine (4%, s.c.) was applied locally to the scalp at the beginning of the surgery to provide additional analgesia. After confirming adequate anesthesia and analgesia, the scalp was incised, and the exposed area was carefully cleaned and dried prior to initiating stereotaxic procedures. Glucose (4 ml, 5%, s.c.) was administered to rehydrate the animals and maintain metabolic stability prior to probe insertion.

A craniotomy (2.5 mm in diameter) was performed at stereotaxic coordinates referenced to Bregma: anteroposterior (AP) -4.2 mm, mediolateral (ML) 2.5 mm, and dorsoventral (DV) 1.4 mm, targeting the posterior parietal cortex, corresponding to the parietal association cortex (PtA) as defined in the Paxinos and Watson atlas [18].

Electrophysiological recordings were performed using a 64-channel silicon probe (BuzsakiA64sp, sharp-tip configuration, NeuroNexus). The probe geometry consisted of 64 recording sites distributed across 6 shanks, with 10 sites per shank. The shanks were spaced 200 μm apart, and each recording site had an area of 160 μm^2, with an impedance ranging from 1 to 3 MΩ at 1 kHz. The recording sites were arranged in a staggered configuration along each shank, with an inter-site spacing of 20 μm, allowing for high-density sampling of neuronal activity across cortical layers. This configuration enables simultaneous recordings of local field potentials (LFPs) and neuronal spiking activity across multiple spatial scales [1], [11].

The probe was positioned over the craniotomy and aligned along the anteroposterior axis. The probe was maintained parallel to the horizontal plane and gradually advanced at a controlled rate of 100 μm/min using a micromanipulator, ensuring precise targeting and minimizing tissue damage. The probe was lowered to a final depth of approximately 1400 μm from the pia mater, allowing recordings across cortical layers. Each animal underwent a single recording session with an average duration of approximately 90 minutes. Analyses focused primarily on signals recorded between 900 and 1200 μm in depth, corresponding predominantly to cortical layers V and VI. These deeper cortical layers are characterized by a high density of pyramidal neurons with long-range projections, playing a central role in corticofugal output and integration of sensorimotor and associative information [12], [14].

The probe was carefully inserted into the target cortical area and connected to a NeuroNexus probe adapter. The signal was then routed via an Omnetics connector to a headstage (Intan, RHD2164), which performed amplification, filtering, and digitization of the analog signals. The digitized data were

transmitted to an Open Ephys acquisition board at a sampling rate of 30 kHz per channel with 16-bit resolution per sample. Data acquisition, synchronization, and real-time visualization of electrophysiological signals were performed using the Open Ephys Project software interface [20].

C. Spike sorting and unit classification

Spike detection and sorting were performed with Kilosort 4.0. Clusters were manually curated in phy 2.1.0rc1 [22] by inspecting waveform consistency, principal-component distributions, autocorrelograms (to assess refractory-period violations), and temporal stability across the session. Each cluster was classified as single-unit activity (SUA), multi-unit activity (MUA), or noise; noise clusters were excluded from all subsequent analyses [11]. Units were classified as SUA if the fraction of inter-spike intervals (ISIs) below 2 ms was less than 1 %; all other units were classified as MUA.

D. Waveform-based classification

SUA waveforms were extracted from the curated spike-sorting output using the SpikeInterface framework [27]. To standardize the profiles, mean waveforms were aligned to their respective troughs and normalized by their trough amplitudes. We then calculated the trough-to-peak duration for each unit, defined as the time interval between the global minimum (trough) of the waveform and its subsequent local maximum (peak). These duration values were subsequently fitted with a two-component Gaussian Mixture Model (GMM) to classify the neuronal population into narrow-spiking (NS) and wide-spiking (WS) units based on the GMM decision boundary. Finally, we compared the baseline firing-rate distributions between the classified NS and WS populations [41].

E. Population activity analysis,

Spike times from units classified as single-unit activity were used to quantify the temporal dynamics of neuronal population activity in the posterior parietal cortex. These metrics followed commonly used extracellular electrophysiology descriptors, including metrics implemented in CellExplorer [23]. Inter-spike intervals (ISIs) were computed from consecutive spike times, and the median ISI was used to summarize the typical temporal spacing between spikes. The proportion of long ISIs was defined as the fraction of ISIs longer than 3 s. Burstiness was quantified from the ISI distribution as

$$B = \frac{\sigma_{\mathrm{ISI}} - \mu_{\mathrm{ISI}}}{\sigma_{\mathrm{ISI}} + \mu_{\mathrm{ISI}}}, \tag{1}$$

where σ_{ISI} and μ_{ISI} are the standard deviation and mean of the ISI distribution, respectively.

Population activity analyses were performed using spike times from units classified as SUA or MUA. Following the approach used to characterize cortical states through spiking variability [21], [38]–[40], the recording was divided into non-overlapping 10 s time windows. Within each 10 s window, spikes from all SUA and MUA units belonging to the same shank were pooled and binned using 50 ms bins. For each shank and time window, the population firing rate was computed as

$$FR = \frac{N_{spikes}}{T}, \tag{2}$$

where N_{spikes} is the total number of spikes in the window and $T = 10$ s. Firing rate was expressed in Hz.

The coefficient of variation (CV) of the population activity was computed from the distribution of spike counts across the 50 ms bins within each 10 s window [21], [38]–[40]:

$$CV = \frac{\sigma}{\mu}, \tag{3}$$

where σ is the standard deviation and μ is the mean of the binned spike-count distribution. Thus, CV quantified the temporal variability of the summed population activity within each shank. Higher CV values indicate more irregular or burst-like population activity, whereas lower values indicate more regular activity across time bins. For the example population plots, CV was computed using non-overlapping 5 s windows over the first 60 min of each recording, with fixed 50 ms bins within each window. Example raster plots and population firing-rate traces were selected to represent windows with the lowest CV, highest CV, and median CV. For the across-animal summary, CV distributions were computed using non-overlapping 10 s windows and fixed 50 ms bins.

For the population firing-rate traces, spikes from all included SUA and MUA units were pooled, and spike counts in each 50 ms bin were divided by the bin duration and smoothed using a Gaussian kernel with a standard deviation of 150 ms. Therefore, population firing rate represented summed activity and was not normalized by the number of units. In contrast, unit-level firing rates were calculated separately by dividing the total number of spikes emitted by each unit by the corresponding recording duration.

To assess whether the pooled distributions were disproportionately influenced by individual recording sessions or by differences in unit yield, we additionally summarized population metrics at the shank level. For each animal and shank, the median population CV across non-overlapping 10-s windows was calculated from the summed SUA+MUA activity. Mean firing rate per shank was calculated by averaging the recording-wide firing rates of the SUA+MUA units assigned to that shank. These summaries preserved the nested structure of the dataset, with units grouped within shanks and shanks grouped within animals.

III. Results

A. Single-unit activity metrics

We first characterized the baseline spike-train properties of single units recorded in the PPC (Fig. 1). The experimental setup, including the stereotaxic targeting of the PPC and the high-density multi-shank silicon probe layout, is illustrated in (Fig. 1a). Across six recording sessions, our curated dataset yielded a total of 186 high-quality SUA clusters (Table I);

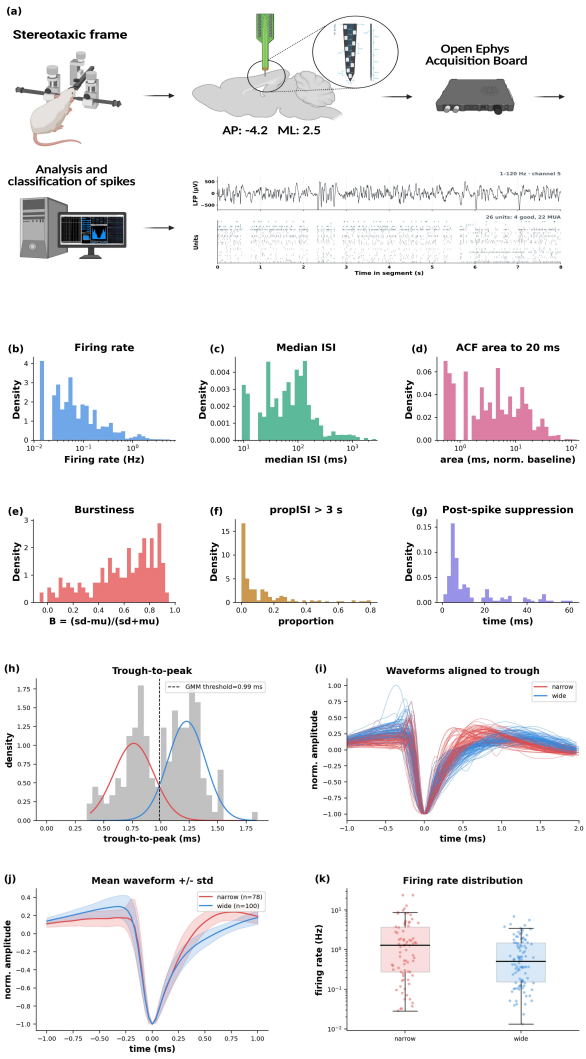


Fig. 1. **Single-unit spike-train metrics in rat PPC under anesthesia**. **(a)** Experimental workflow and setup. Extracellular recordings were obtained from PPC using a high-density multi-shank silicon probe. Wideband signals were acquired via an Open Ephys acquisition system and processed offline for spike sorting. A representative local field potential (LFP) trace (bandpass filtered at 1–120 Hz) and corresponding spike raster plot illustrate the recorded neural activity, highlighting the distinction between classified single units (SUA/"good" units, in blue) and multi-unit activity (MUA, in grey). **(b–g)** Distributions across SUA ($n = 186$) of: **(b)** mean firing rates, **(c)** median ISIs, **(d)** normalized ACF area (20 ms lag), **(e)** burstiness index $B = (\mathrm{SD} - \mu)/(\mathrm{SD} + \mu)$, **(f)** proportion of ISIs > 3 s, and **(g)** post-spike suppression. (**h**) Distribution of trough-to-peak durations (ms) across units. A two-component Gaussian mixture model was fit to the trough-to-peak values, separating narrow-spiking (red; n = 78) from wide-spiking (blue; n = 100) units; The dashed line indicates the GMM classification thresh-old at 0.99 ms. Colored curves show the two fitted Gaussian components. (**i**) All mean waveforms, amplitude-normalized and aligned to the trough, colored by class (red, narrow; blue, wide). (**j**) Class-averaged waveforms (mean ± SD) for narrow (red) and wide (blue) units. (**k**) Firing-rate distribution of units classified as narrow-spiking and wide-spiking. Each point represents one SUA unit, boxes indicate the interquartile range, horizontal lines indicate the median, and whiskers show the data range excluding outliers (Hz; log scale). *Created in BioRender. BEATRIZ VIEIRA COUTO DE LIMA, A. (2026). https://BioRender.com/hzntz9y*

when multi-unit activity (MUA) clusters were also included, a total of 703 sorted units were available for analysis (Table II).

Firing rates were broadly distributed and right-skewed, with most units firing at low rates and a smaller subset showing higher activity levels. The median firing rate was 0.62 Hz (interquartile range, IQR: 0.17–1.82 Hz).

Consistent with this firing-rate profile, the ISIs also showed a broad distribution across units, with a median of 200.9 ms (IQR: 108.3–823.1 ms). The proportion of long ISIs (>3 s) varied across units, with a median of 0.044 (IQR: 0.012–0.154), indicating heterogeneous silent-period structure across the SUA population. Burstiness values were mostly positive, with a median burstiness index of 0.64 (IQR:0.43–0.80), indicating that many SUA spike trains showed temporally clustered firing patterns.

B. Waveform-based unit classification

A total of 178 extracellular waveforms from single-unit activity were extracted and analyzed (Fig. 1). Waveforms were aligned to their trough and normalized by trough amplitude to enable comparisons across units. The trough-to-peak duration exhibited a bimodal distribution and was classified using a two-component Gaussian mixture model, separating units into narrow-spiking and wide-spiking populations (Fig. 1h.).

The GMM classification thresh-old at 0.99 ms. This classification identified 78 narrow-spiking units and 100 wide-spiking units. Narrow-spiking units exhibited shorter trough-to-peak durations (median: 0.78 ms), whereas wide-spiking units showed longer trough-to-peak durations (median: 1.23 ms) (Fig. 1i.). To illustrate the characteristic waveform shape of the recorded population, the mean normalized waveform was overlaid on the individual waveforms for each group (Fig. 1j.). Finally, firing-rate distributions were compared between waveform classes, showing that NS units tended to exhibit higher firing rates than WS units (Fig. 1k).

C. Population activity dynamics

Population activity dynamics were examined based on the variability of the neuronal population spiking activity over time [33], [38], [40], as illustrated in Fig. 2. As detailed in Methods (Sec. II-E), this variability was estimated using the coefficient of variation, CV(t), computed separately for each shank (Fig. 2a) and for the pooled population across all shanks (Fig. 2c). We have previously demonstrated strong coupling in neuronal population spiking dynamics within and among primary sensory cortices in rodents [38], but whether such coupling exists within the PPC remains unknown. Shank-wise CV(t) values showed temporal fluctuations throughout the recording, and the CV(t) correlation matrix summarized the similarity of population variability time series across shanks (Fig. 2b). For the pooled population (Fig. 2d), CV values ranged from 0.38 to 0.90, with a median of 0.63. Example raster plots and population firing-rate traces were selected from windows corresponding to the lowest CV, highest CV, and median CV(Fig. 2e-g).

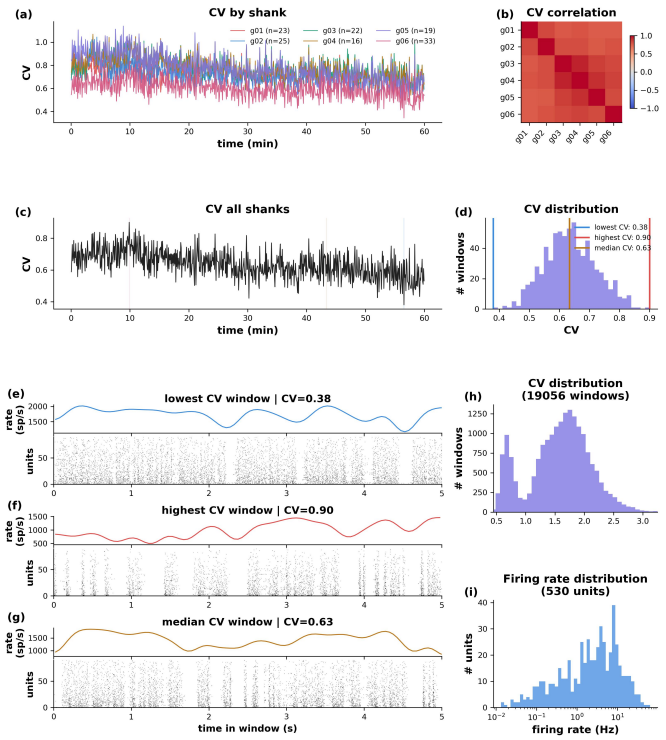


Fig. 2. **Population spiking variability in PPC recording.** (**a**) Coefficient of variation, CV(t), of population activity computed separately for each probe shank. Spikes from SUA and MUA clusters were pooled within each shank, binned using fixed 50 ms bins, and CV was computed in non-overlapping 5 s windows over the first 60 min of the recording. (**b**) Correlation matrix between shank-wise CV(t) time series. (**c**) CV(t) computed after pooling activity across all shanks. Colored vertical bands indicate the example windows selected for raster and firing-rate visualization. (**d**) Distribution of CV values across time windows for the pooled population activity. Vertical lines indicate the selected lowest, highest, and median CV windows. (**e-g**) Population firing-rate traces and raster plots for example windows corresponding to the lowest CV, highest CV, and median CV, respectively. In raster plots, each row represents one SUA or MUA unit and each tick corresponds to one spike. Example shown for animal 20260514. n = 138 units. (**h**) Distribution of population CV values computed from pooled SUA+MUA activity across all animals recorded from 2026-04-06 onward. For each shank, spikes were binned using fixed 50 ms bins, and CV was computed from non-overlapping 10 s windows. (**i**) Distribution of firing rates across SUA+MUA units, computed as the total number of spikes divided by the recording duration and shown on a logarithmic scale. n = 19,056 CV windows and 530 SUA+MUA units

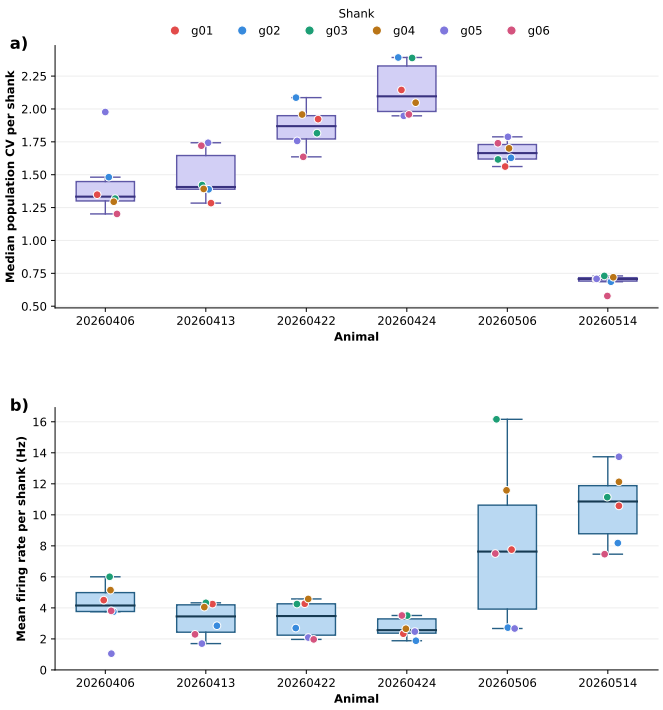


Fig. 3. **Population activity variability and firing rate across animals.** (**a**) Distribution of the population coefficient of variation across shanks for each animal. For each shank, spikes from all curated single-unit and multi-unit activity were pooled, binned at 50 ms within non-overlapping 10-s windows, and the median CV across windows was calculated. (**b**) Distribution of mean firing rates across shanks. For each shank,firing rates were calculated for individual SUA+MUA units over the recording session and then averaged across units. In both panels, each colored point represents one shank, with colors identifying the corresponding shank. Boxes show the interquartile range, horizontal lines indicate the median across shanks, and whiskers extend to 1.5 times the interquartile range.

TABLE I
NUMBER OF SINGLE-UNIT (SUA) CLUSTERS PER SHANK AND ANIMAL.

Animal	Sh1	Sh2	Sh3	Sh4	Sh5	Sh6	Total
2026-04-06	10	4	0	5	8	12	39
2026-04-13	7	12	7	6	9	4	45
2026-04-22	3	6	1	2	6	4	22
2026-04-24	5	0	0	1	5	0	11
2026-05-06	2	14	0	0	4	3	23
2026-05-14	7	13	4	4	5	13	46
Total	**34**	**49**	**12**	**18**	**37**	**36**	**186**

In addition, population variability and firing-rate distributions were summarized across all recordings (Fig. 2). Across six recordings, the SUA+MUA dataset included 530 units. The number of units per recording ranged from 61 to 138 units.

The distribution of population CV values was computed across shanks and recordings using fixed 50 ms bins within non-overlapping 10 s windows. Across 19,056 CV windows, the median CV was 1.62 (IQR:1.25–1.92), with values ranging from 0.41 to 4.64. Firing rates across SUA+MUA units were broadly distributed on a logarithmic scale, with a median firing rate of 2.68 Hz (IQR: 0.61–7.69 Hz). Together, these distributions summarize the range of population variability and unit firing rates observed across the recorded animals.

We further examined the distribution of population CV and firing rate using the shank as the summary level (Fig. 3). Each point therefore represented one shank rather than an individual time window or unit. Population CV and mean firing rate varied both within and across animals, indicating that the broad pooled distributions were not attributable to a single recording. Nevertheless, the visible between-animal variability suggests that recording identity and local sampling should be considered when interpreting pooled population metrics.

IV. DISCUSSION

In this study, we characterized SUA and population activity recorded from the rat PPC using a high-density multi-shank silicon probe.

TABLE II
TOTAL NUMBER OF SORTED UNITS (SUA AND MUA) PER SHANK AND ANIMAL.

Animal	Sh1	Sh2	Sh3	Sh4	Sh5	Sh6	Total
2026-04-06	29	16	6	21	14	28	114
2026-04-13	25	35	20	22	25	17	144
2026-04-22	13	18	8	11	20	16	86
2026-04-24	25	9	4	10	20	4	72
2026-05-06	17	38	4	8	25	11	103
2026-05-14	30	38	26	20	24	46	184
Total	**139**	**154**	**68**	**92**	**128**	**122**	**703**

Different recording modalities provide complementary views of population dynamics. Here, we focused on spiking activity because its cellular resolution allows us to characterize individual neuronal dynamics and population coding, whereas LFPs primarily capture broader network-level activity. LFP-based spectral analyses may therefore provide a complementary perspective on the network organization underlying the dynamics observed here.

Rather than examining behavior-related neural coding, our recordings capturedmspontaneous neuronal activity under ketamine/xylazine anesthesia, allowing us to characterize single-unit firing patterns and the temporal organization of population activity in the PPC. This baseline characterization provides a quantitative framework for understanding the intrinsic dynamics of neuronal ensembles in this associative cortical region and establishes a reference for future studies of cortical population dynamics and neural coding [1], [2].

Because anesthesia strongly shapes cortical dynamics, the anesthetic regime is a central consideration when interpreting the population activity reported here [21], [28]. Recordings were obtained under ketamine/xylazine rather than the urethane protocol previously adopted for acute recordings in our laboratory [30]. This choice was motivated by the associative nature of the PPC: as a higher-order region that relies on polysynaptic cortico-cortical and thalamo-associative transmission, the PPC is disproportionately sensitive to anesthetics that promote slow-wave states. Urethane, in particular, favors prolonged slow oscillations with extended DOWN states that attenuate precisely the polysynaptic transmission on which associative processing depends [28], [29]. Ketamine, an NMDA-receptor antagonist, combined with the α_2-adrenergic agonist xylazine, instead produces a comparatively more activated, desynchronized cortical state, with reduced low-frequency power and better preserved cortico-cortical engagement than urethane [28]. The broad, right-skewed firing-rate distributions and the wide range of population CV values observed here are consistent with such an activated regime, rather than with the stereotyped, highly synchronous slow-wave activity typical of deep urethane anesthesia.

The distribution of SUA firing rates was broad and right-skewed, with most units exhibiting low firing rates and a smaller subset displaying substantially higher firing rates. This heterogeneous distribution is consistent with previous observations in cortical and hippocampal populations, where neuronal firing rates span several orders of magnitude and often follow skewed or log-normal-like distributions [24]. Such heterogeneity likely reflects differences in intrinsic excitability, synaptic connectivity, neuronal subtype, and participation in local network events. Likewise, the broad distribution of ISIs and the predominance of positive burstiness values indicate that PPC neurons exhibit diverse temporal firing patterns rather than homogeneous discharge activity.

Waveform analysis further revealed two major groups of extracellularly recorded units based on trough-to-peak duration. This waveform-based classification has been widely used to describe the physiological diversity of neurons recorded using extracellular electrophysiological techniques [25]. In the present dataset, narrow-spiking and wide-spiking units exhibited consistent differences in waveform duration and firing-rate distribution, suggesting that PPC recordings captured electrophysiologically distinct neuronal populations. However, because extracellular waveform characteristics alone cannot unequivocally identify neuronal cell types, these groups should be interpreted as electrophysiological categories rather than definitive classifications of excitatory or inhibitory neurons.

At the population level, coefficient of variation analyses revealed pronounced temporal fluctuations in collective spiking activity across animals and recording shanks. Periods of lower CV were associated with more regular population firing, whereas higher CV values corresponded to more clustered or burst-like spiking patterns. These alternating activity regimes exhibit features consistent with UP/DOWN-like cortical dynamics, in which neuronal populations alternate between states of higher and lower activity. Similar state-dependent fluctuations have been reported during spontaneous cortical activity under anesthesia [21], [26], suggesting that the anesthetized PPC retains a temporally structured organization of population activity despite the absence of sensory stimulation or behavioral engagement.

The shank-level analysis provided an additional view of the hierarchical structure of the dataset. Population CV and mean firing rate showed variability both among shanks from the same animal and across animals. This indicates that the pooled distributions should not be interpreted as arising from a homogeneous population of independent observations. Differences in unit yield, local probe sampling, and animal-specific physiological conditions may all contribute to the observed dispersion. Summarizing the data at the shank level within each recording makes this variability explicit and reduces the disproportionate influence of recordings containing more units or time windows on the overall distributions.

Across animals, the distributions of CV values and firing rates demonstrated that PPC population activity varied across recordings. This variability may reflect differences in local circuit sampling across shanks, the number of recorded units per animal, and the functional state of the cortical network under anesthesia. Given that the PPC serves as an important associative hub responsible for integrating sensory and motor-related information, such variability may be particularly relevant for understanding how this region transitions between internally generated activity patterns and those associated

with information processing during behavioral tasks. Future studies comparing spontaneous activity under anesthesia with recordings obtained from awake animals during behavioral tasks may help clarify how these population patterns relate to the functional roles of the PPC.

Overall, these results provide an initial quantitative characterization of single-unit activity properties and the population dynamics of combined SUA and MUA recordings in the anesthetized rat posterior parietal cortex. The findings indicate that PPC activity exhibits heterogeneous single-unit firing patterns, waveform diversity, and temporally structured population-level variability. These features are consistent with the concept that PPC function depends on the coordinated dynamics of neuronal ensembles and provide a foundation for future investigations into cortical state organization within this associative cortical region.

V. Conclusion

This study provides an initial characterization of single-unit and population activity in the anesthetized rat posterior parietal cortex. SUA units displayed heterogeneous firing-rate, ISI, burstiness, and waveform properties, while SUA+MUA population activity showed temporally structured variability across shanks and animals. CV-based analyses and raster plots revealed epochs compatible with UP/DOWN-like dynamics, although further analyses combining spiking activity with additional state markers are required to confirm this interpretation. Together, these findings support high-density extracellular recordings as a useful approach to describe unit-level diversity and population-level dynamics in PPC circuits.

Acknowledgment

The authors thank the Federal University of Pernambuco (UFPE) for institutional support, specifically: the Neuroscience Laboratory (Department of Physics) for providing the recording facilities; the Neurorobotics Laboratory and the High-Performance Computing infrastructure (PADMEC) at the Petroleum and Energy Research Institute (i-LITPEG) for data analysis support; and the Naíde Teodósio Physiology of Nutrition Laboratory (Department of Nutrition) for providing reagents and technical assistance.

This work was supported by the following funding agencies and grants: CNPq (Grant No. 314094/2023-7; Grant No. 444500/2024-3, Grant No. 444713/2024-7); FACEPE (ARC-0115-3.13/25); FADE/UFPE (Grant No. 64/2024); INCT-NeuroComp (CNPq Grant No. 408389/2024-9); and Finep-Neurobeep (Grant No. 01.24.0124.00).

References

[1] G. Buzsáki, "Large-scale recording of neuronal ensembles," *Nature Neuroscience*, vol. 7, pp. 446–451, 2004.

[2] M. A. L. Nicolelis and M. A. Lebedev, "Principles of neural ensemble physiology: From single neurons to brain–machine interfaces," *Nature Reviews Neuroscience*, vol. 10, no. 7, pp. 530–540, 2009.

[3] R. A. Andersen and C. A. Buneo, "Intentional maps in posterior parietal cortex," *Annual Review of Neuroscience*, vol. 25, pp. 189–220, 2002.

[4] R. A. Andersen and H. Cui, "Intention, action planning, and decision making in parietal-frontal circuits," *Neuron*, vol. 63, no. 5, pp. 568–583, 2009.

[5] J. R. Whitlock, "Posterior parietal cortex," *Current Biology*, vol. 27, no. 14, pp. R691–R695, 2017.

[6] C. D. Harvey, P. Coen, and D. W. Tank, "Choice-specific sequences in parietal cortex during a virtual-navigation decision task," *Nature*, vol. 484, no. 7392, pp. 62–68, 2012.

[7] H. Mohan, R. de Haan, H. D. Mansvelder, and C. P. J. de Kock, "Functional architecture and encoding of tactile sensorimotor behavior in rat posterior parietal cortex," *The Journal of Neuroscience*, vol. 39, no. 37, pp. 7332–7343, 2019.

[8] J. J. Jun *et al.*, "Fully integrated silicon probes for high-density recording of neural activity," *Nature*, vol. 551, no. 7679, pp. 232–236, 2017.

[9] N. A. Steinmetz *et al.*, "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *Science*, vol. 372, no. 6539, p. eabf4588, 2021.

[10] C. Rossant *et al.*, "Spike sorting for large, dense electrode arrays," *Nature Neuroscience*, vol. 19, no. 4, pp. 634–641, 2016.

[11] K. D. Harris and R. Q. Quiroga, "Spike sorting," *Neuron*, vol. 91, no. 2, pp. 298–313, 2016.

[12] K. D. Harris and G. M. G. Shepherd, "The neocortical circuit: Themes and variations," *Nature Neuroscience*, vol. 18, pp. 170–181, 2015.

[13] A. M. Thomson, "Neocortical layer 6, a review," *Frontiers in Neuroanatomy*, vol. 4, p. 13, 2010.

[14] G. M. G. Shepherd, "Corticostriatal connectivity and its role in disease," *Nature Reviews Neuroscience*, vol. 14, pp. 278–291, 2013.

[15] National Research Council, Guide for the Care and Use of Laboratory Animals, 8th ed. Washington, DC: National Academies Press, 2011.

[16] N. P. du Sert et al., "The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research," Nature Protocols, vol. 15, pp. 3793–3801, 2020.

[17] P. Flecknell, Laboratory Animal Anaesthesia, 4th ed. London: Academic Press, 2015.

[18] G. Paxinos and C. Watson, The Rat Brain in Stereotaxic Coordinates, 6th ed. Amsterdam: Elsevier, 2007.

[19] R. Sitaram et al., "Closed-loop brain training: the science of neurofeedback," Nature Reviews Neuroscience, vol. 18, pp. 86–100, 2017.

[20] J. H. Siegle et al., "Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology," Journal of Neural Engineering, vol. 14, no. 4, 2017.

[21] A. J. Fontenele et al., "Criticality between cortical states," Physical Review Letters, vol. 122, no. 20, p. 208101, 2019.

[22] Cortex Lab, "phy: interactive visualization and manual spike sorting of large-scale ephys data," GitHub repository, 2026. [Online]. Available: https://github.com/cortex-lab/phy

[23] P. C. Petersen, J. H. Siegle, N. A. Steinmetz, S. Mahallati, and G. Buzsáki, "CellExplorer: A framework for visualizing and characterizing single neurons," *Neuron*, vol. 109, no. 22, pp. 3594–3608.e2, 2021.

[24] G. Buzsáki and K. Mizuseki, "The log-dynamic brain: how skewed distributions affect network operations," *Nature Reviews Neuroscience*, vol. 15, no. 4, pp. 264–278, 2014.

[25] P. Barthó et al., "Characterization of neocortical principal cells and interneurons by network interactions and extracellular features," *Journal of Neurophysiology*, vol. 92, no. 1, pp. 600–608, 2004.

[26] A. Luczak, P. Barthó, S. L. Marguet, G. Buzsáki, and K. D. Harris, "Sequential structure of neocortical spontaneous activity in vivo," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 347–352, 2007.

[27] A. P. Buccino, S. Hurwitz, S. Garcia, J. Magland, J. H. Siegle, R. Hurwitz, and M. H. Hennig, "SpikeInterface, a unified framework for spike sorting," *eLife*, vol. 9, p. e61834, 2020.

[28] V. Sorrenti et al., "Understanding the effects of anesthesia on cortical electrophysiological recordings: a scoping review," *International Journal of Molecular Sciences*, vol. 22, no. 3, p. 1286, 2021.

[29] E. A. Clement et al., "Cyclic and sleep-like spontaneous alternations of brain state under urethane anaesthesia," *PLOS ONE*, vol. 3, no. 4, p. e2004, 2008.

[30] N. A. P. de Vasconcelos et al., "Coupled variability in primary sensory areas and the hippocampus during spontaneous activity," *Scientific Reports*, vol. 7, p. 46077, 2017.

[31] R. J. Douglas and K. A. C. Martin, "Neuronal circuits of the neocortex," *Annual Review of Neuroscience*, vol. 27, pp. 419–451, 2004.

[32] K. D. Harris and T. D. Mrsic-Flogel, "Cortical connectivity and sensory coding," *Nature*, vol. 503, no. 7474, pp. 51–58, 2013.

[33] K. D. Harris and A. Thiele, "Cortical state and attention," *Nature Reviews Neuroscience*, vol. 12, no. 9, pp. 509–523, 2011.

[34] J. A. Harris, S. Mihalas, K. E. Hirokawa, *et al.*, "Hierarchical organization of cortical and thalamic connectivity," *Nature*, vol. 575, no. 7781, pp. 195–202, 2019.

[35] R. A. Andersen, L. H. Snyder, D. C. Bradley, and J. Xing, "Multimodal representation of space in the posterior parietal cortex and its use in planning movements," *Annual Review of Neuroscience*, vol. 20, no. 1, pp. 303–330, 1997.

[36] H. Mohan, R. de Haan, R. Broersen, A. W. Pieneman, F. Helmchen, J. F. Staiger, H. D. Mansvelder, and C. P. J. de Kock, "Functional architecture and encoding of tactile sensorimotor behavior in rat posterior parietal cortex," *Journal of Neuroscience*, vol. 39, no. 37, pp. 7332–7343, 2019.

[37] R. Sitaram, T. Ros, L. Stoeckel, S. Haller, F. Scharnowski, J. Lewis-Peacock, N. Weiskopf, M. L. Blefari, M. Rana, E. Oblak, N. Birbaumer, and J. Sulzer, "Closed-loop brain training: the science of neurofeedback," *Nature Reviews Neuroscience*, vol. 18, no. 2, pp. 86–100, 2017.

[38] N. A. P. de Vasconcelos, C. Soares-Cunha, A. J. Rodrigues, S. Ribeiro, and N. Sousa, "Coupled variability in primary sensory areas and the hippocampus during spontaneous activity," *Scientific Reports*, vol. 7, art. no. 46077, 2017.

[39] E. V. de Paula, R. M. Jungmann, A. J. Fontenele, L. A. A. Aguiar, P. V. Carelli, F. S. Matias, M. Copelli, and N. A. P. de Vasconcelos, "Revealing stimulus-dependent dynamics through statistical complexity," *PRX Life*, vol. 4, no. 2, art. no. 023018, 2026.

[40] A. Renart, J. de la Rocha, P. Barthó, L. Hollender, N. Parga, A. Reyes, and K. D. Harris, "The asynchronous state in cortical circuits," *Science*, vol. 327, no. 5965, pp. 587–590, 2010.

[41] M. D. Best, K. Takahashi, A. J. Suminski, C. Ethier, L. E. Miller, and N. G. Hatsopoulos, "Comparing offline decoding performance in physiologically defined neuronal classes," *Journal of Neural Engineering*, vol. 13, no. 2, Art. no. 026004, 2016, doi: 10.1088/1741-2560/13/2/026004.

Desenvolvimento e Padronização de um Procedimento Operacional Padrão para Cirurgia Estereotáxica de Infusão Viral e Implante de Cânula em Ratos Wistar

1st Lívian Cerqueira Mousinho C. de Sousa
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
livian.mousinho@edu.isd.org.br

2nd Aurea Estella de Araújo Silva
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
aurea.silva@edu.isd.org.br

3rd Bruno Ferreira da Silva
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
bruno.silva@edu.isd.org.br

4th Ana Helena Silva Moura
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
ana.moura@edu.isd.org.br

5th Isabella Figueiredo da Silva
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
isabella.silva@edu.isd.org.br

6th Rebeka Nogueira da Silva
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
rebeka.silva@edu.isd.org.br

7th Giovana Teixeira Moreira
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
giovana.moreira@edu.isd.org.br

8th Ramón Hypolito Lima
Laboratório de memória e Cognição
Instituto Santos Dumont
Macaíba, Brasil
ramon.lima@edu.isd.org.br

***Abstract*—Os ratos *Wistar* são amplamente utilizados em pesquisas de neurobiologia e engenharia biomédica, especialmente em estudos que empregam a optogenética como ferramenta de neuromodulação. A entrega de vetores virais e o implante de cânulas por meio de cirurgia estereotáxica são etapas críticas desse tipo de estudo, mas a literatura carece de um protocolo estruturado e padronizado que seja aplicável de forma consistente entre diferentes laboratórios. Este trabalho teve como objetivo desenvolver e descrever um Procedimento Operacional Padrão (POP) para a cirurgia estereotáxica de infusão viral e implante de cânula em ratos *Wistar*, dividido em quatro etapas: preparação pré-operatória e anestesia, procedimento cirúrgico estereotáxico, implante da cânula com infusão viral, cuidados pós-operatórios.O protocolo foi aplicado em 28 animais, dos quais 13 foram submetidos a cirurgia combinada (implante e infusão na mesma sessão) e 15 a cirurgias separadas. A taxa de sobrevivência geral foi de $85{,}7\%$, com os óbitos concentrados predominantemente na fase anestésica. Os resultados indicam que o POP proposto é tecnicamente viável e reprodutível, contribuindo tanto para a padronização experimental quanto para o bem-estar animal. A acurácia do posicionamento e a expressão viral foram validadas por confirmação histológica.**

***Index Terms*—Optogenética; Cirurgia estereotáxica; Ratos Wistar; Infusão viral; Procedimento Operacional Padrão; Neuromodulação.**

I. Introdução

A estimulação optogenética estabeleceu-se como uma ferramenta indispensável na pesquisa em neuromodulação em roedores de laboratório, permitindo o controle preciso e a investigação detalhada de circuitos neurais específicos por meio da infusão de vetores virais [1]. Para que essa abordagem alcance a eficácia desejada, o sucesso metodológico depende criticamente da acurácia na cirurgia estereotáxica e na implantação de cânulas-guia para a entrega precisa do material genético no tecido cerebral de interesse [2]. Esse tipo de abordagem cirúrgica já foi descrito para diversas estruturas encefálicas, como a área tegmental ventral, o locus coeruleus e o núcleo dorsal da rafe, mas cada grupo de pesquisa tende a adaptar o procedimento às particularidades de sua própria linha de trabalho [3].

Apesar da ampla utilização dessas técnicas, a literatura aponta uma lacuna importante: a inexistência de um protocolo estruturado e padronizado que seja aplicável de forma consistente entre diferentes modelos animais e diferentes laboratórios. Revisões sistemáticas sobre boas práticas cirúrgicas em roedores mostram que há grande variabilidade na execução de procedimentos como assepsia, manejo perioperatório e cuidados pós-cirúrgicos, o que compromete tanto a reprodutibilidade científica quanto o bem-estar animal [4]. Estudos que buscaram refinar tecnicamente a cirurgia estereotáxica em roedores demonstraram que o aprimoramento e a padronização de etapas cirúrgicas reduzem significativamente o número de animais necessários por grupo experimental, além de diminuir

complicações e melhorar a recuperação pós-operatória [3].

A padronização por meio de um Procedimento Operacional Padrão (POP) é fundamental pois assegura a reprodutibilidade dos resultados entre diferentes execuções e operadores, reduzindo variáveis técnicas que poderiam confundir os desfechos experimentais, e protocolos bem definidos e refinados estão diretamente associados a menor morbidade, menor tempo cirúrgico e melhor manejo da dor, contribuindo para o bem-estar animal [5].

Diante disso, este artigo tem como objetivo apresentar e descrever um Procedimento Operacional Padrão (POP) para a cirurgia estereotáxica de infusão viral via cânula em ratos Wistar, com foco na padronização das etapas pré-operatórias, intraoperatórias e pós-operatórias, visando aumentar a reprodutibilidade experimental e assegurar o bem-estar animal em pesquisas de simulação optogenética.

II. Metodologia

Este estudo foi aprovado pelo Comitê de Ética no Uso de Animais do Instituto Santos Dumont (CEUA-ISD; protocolo n.º 01/2020 ISD), seguindo estritamente a Lei n.º 6.638/1979 e as normas regulamentadas pelo Conselho Nacional de Controle de Experimentação Animal (CONCEA).

A. Preparação Pré-Operatória e Anestesia

Foram utilizados ratos Wistar (*Rattus norvegicus*) machos, com aproximadamente 3 meses de idade e peso corporal entre 300 e 400 g. Os animais estão alojados no biotério do Instituto Santos Dumont (ISD), sob supervisão da médica veterinária e bioteristas. As caixas-moradia, feitas de acrílico transparente, abrigarão até três animais e serão preenchidas com maravalha autoclavada, garantindo a qualidade de vida para um bom procedimento.

Ao ser retirado do biotério, o animal é transferido para uma caixa de indução anestésica acoplada a um vaporizador calibrado de isoflurano. A indução tem início com uma concentração de $3,5\%$, sendo progressivamente ajustada conforme a resistência fisiológica apresentada por cada animal, até que seja atingido o plano anestésico adequado (perda do reflexo de endireitamento). Durante todo o procedimento, o animal permanece sob fluxo contínuo de oxigênio, mantido em 1 L/min, como observado na Figura 1 abaixo.

Uma vez adormecido, o animal recebe, por via intraperitoneal, dois fármacos calculados individualmente conforme o peso corporal: primeiro, a atropina, na dose calculada por peso (g) $\times\ 0,02/100$ mg, empregada para reduzir secreções e prevenir bradicardia reflexa; em seguida, a cetamina, na dose de peso (g) $\times\ 0,07/100$ mg. No intervalo de 5 minutos entre a aplicação dos dois fármacos, a tricotomia da região a ser operada é realizada utilizando como referência anatômica o Triângulo de Einthoven, delimitado pelos vértices superior direito, superior esquerdo e inferior, o que garante uma área de tricotomia padronizada entre os animais [6].

Após a administração da cetamina, o animal é transportado até o centro cirúrgico para o início do procedimento. A manutenção da anestesia é então realizada com isoflurano (2–3%) em fluxo contínuo de oxigênio.



(a) (b) (c)

Fig. 1: Sistema de anestesia inalatória e componentes preparatórios. (a) Bancada experimental contendo seringas preparadas para indução anestésica e câmara de indução de acrílico conectada ao sistema. (b) Fluxômetro acoplado à rede de oxigênio medicinal calibrado para fluxo contínuo. (c) Vaporizador calibrado específico para administração e controle de isoflurano.

Um cuidado adicional diz respeito à água do umidificador do sistema de oxigênio, que deve ser sempre substituída por água previamente tratada em centrífuga, tanto no oxigênio do centro cirúrgico quanto no da sala de infusão, onde é realizada a indução anestésica. Além disso, antes do início de qualquer procedimento anestésico, é fundamental que todo o maquinário do centro cirúrgico já esteja ligado e que as doses dos fármacos já estejam calculadas e preparadas.

B. Procedimento Cirúrgico (Estereotaxia)

Antes do início da manipulação do animal, o cirurgião e o auxiliar realizam a paramentação cirúrgica completa,envolvendo o avental estéril, degermação das mãos com esponja cirúrgica estéril de dupla face e luvas cirúrgicas de tamanho adequado, garantindo os cuidados de higiene e prevenção de contaminação. Os materiais utilizados nessa etapa (Figura 2) incluem:

- Avental cirúrgico;
- Campo estéril para a mesa cirúrgica;
- Campo estéril para a mesa acoplada onde colocam os materiais;
- Esponja cirúrgica estéril de dupla face;
- Luvas cirúrgicas;
- Cloreto de sódio (cinco unidades de NaCl de 10 mL);
- Uma caixa de amostras de swab;
- Saco de lixo para resíduos infectantes;
- Seis unidades de seringas de 1,0 mL);

O animal, já tricotomizado, é posicionado sobre uma cama térmica, a qual deve estar previamente aquecida no momento da chegada do animal, prevenindo a hipotermia intraoperatória. A cabeça do animal é então fixada ao aparelho estereotáxico por meio das barras auriculares e da barra incisiva, sendo essencial observar se os dentes incisivos centrais superiores estão corretamente posicionados no equipamento, de modo a permitir a expiração adequada e evitar o risco de sufocamento



Fig. 2: Equipamentos de proteção individual (EPIs) e insumos cirúrgicos estéreis organizados em bancada antes da etapa de paramentação e início do protocolo experimental.

do animal, como apresentado na Figura 3. Previamente ao início do procedimento, o aparelho estereotáxico deve estar conectado e calibrado por meio do software NeuroStar, responsável pela movimentação controlada do equipamento durante a cirurgia.

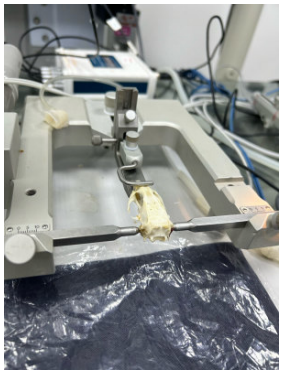


Fig. 3: Alinhamento e fixação estereotáxica do modelo animal. A imagem demonstra o posicionamento correto da cabeça no aparelho estereotáxico, evidenciando a fixação bilateral e simétrica através das barras auriculares inseridas nos canais auditivos externos.

Para a proteção ocular, é aplicada uma gota de NaCl em cada olho do animal, prevenindo o ressecamento da córnea, e um esparadrapo umedecido com NaCl é posicionado sobre os olhos durante toda a cirurgia. A temperatura corporal é monitorada continuamente por meio de termômetro retal, inserido com o auxílio de gel lubrificante íntimo.

Uma vez posicionado o animal no aparelho estereotáxico, é realizada a seguinte sequência padronizada de etapas: (1) limpeza da região do escalpo com solução degermante; (2) preparo da pele com solução antisséptica; (3) aplicação de uma dose de lidocaína diluída em NaCl, para anestesia local complementar; (4) abertura da pele com lâmina de bisturi; e (5), quando necessário, uso de água oxigenada para limpeza e hemostasia local, em casos de sangramento mais intenso. Ao longo de todo o procedimento, mantêm-se disponíveis NaCl, esparadrapos e swabs para conter eventuais sangramentos, conforme representado na Figura 4 abaixo.



Fig. 4: Etapa cronológica e sequencial de preparação para o início do procedimento cirúrgico.

C. Implante da Cânula e Infusão Viral

Ao todo, 28 ratos foram submetidos aos procedimentos cirúrgicos. Desses, 13 animais foram submetidos a uma cirurgia combinada, na qual o implante da cânula e a infusão do vetor viral foram realizados na mesma sessão cirúrgica, enquanto os 15 animais restantes foram submetidos a duas cirurgias separadas, uma exclusivamente para a infusão viral e outra, realizada posteriormente, para o implante da cânula de fibra óptica.

Após a exposição da superfície craniana, os pontos anatômicos de referência, bregma e lambda, são identificados e utilizados para verificar o nivelamento do crânio e para o cálculo das coordenadas estereotáxicas (anteroposterior, mediolateral e dorsoventral) relativas ao bregma, como apresentado na Figura 5 abaixo. A partir das coordenadas calculadas, é realizada a perfuração craniana com broca odontológica, tomando-se cuidado para preservar a integridade da duramáter.



Fig. 5: Alinhamento do crânio para o posicionamento correto da cabeça com o equipamento suporte de alinhamento.

Com o crânio já exposto e nivelado, o direcionamento das coordenadas no aparelho estereotáxico é realizado por meio do software NeuroStar, seguindo a sequência: *Set Bregma* (definição do bregma como ponto de origem) $\rightarrow$ inserção dos valores de AP e ML desejados $\rightarrow$ *Go To* (comando que desloca o braço do equipamento até a posição definida). Uma vez direcionadas as marcações de acordo com a coordenada desejada, são fixados dois parafusos na porção posterior do crânio, que auxiliam na sustentação das cânulas, necessários pois a superfície craniana do animal é lisa. Esses parafusos são posicionados sempre abaixo das cânulas, nunca sobre as suturas cranianas, a fim de evitar sangramentos desnecessários. Ao longo de toda a cirurgia, desde o seu início, os sinais vitais do animal são monitorados a cada 20 minutos.

Foram conduzidos dois experimentos independentes: (1) infusões virais intra-NAc e (2) infusões virais intra-HPCd. No primeiro experimento, foi realizada cirurgia estereotáxica para infusão bilateral do vetor viral contendo opsinas inibitórias AAV-EF1a-NpHR3.0-EYFP (VectorLabs) no núcleo accumbens (NAc), nas coordenadas em relação ao bregma (bregma = 0): AP $+1,4$ mm, ML $\pm 1,4$ mm e DV $-6,8$ mm. Para o posicionamento da agulha, a descida foi realizada gradualmente, em velocidade de 100 μm a cada 10 segundos, até atingir a profundidade determinada. O vírus foi então administrado com seringa Hamilton de 1 μL, a uma taxa de infusão de $0,2$ μL/min, com volume total de $0,5$ μL por hemisfério, conforme ilustrado na Figura 6. Ao final da infusão, a agulha permaneceu no local por 5 minutos, a fim de favorecer a difusão do vetor viral e minimizar o refluxo pelo trajeto da agulha, sendo então removida lentamente. O procedimento foi finalizado com sutura do tecido.

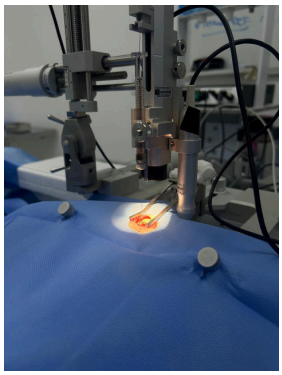


Fig. 6: Infusão do vírus por meio da seringa Hamiltin de 1 μL.

Três semanas após a infusão viral no NAc, foram implantadas bilateralmente cânulas de fibra óptica, com deslocamento das coordenadas para ML $\pm 1,6$ mm, AP $= 0$, em um ângulo de 10°, e DV $-6,9$ mm, permitindo a estimulação optogenética dos terminais axonais. Após o posicionamento, foi utilizada resina dental para manter a cânula fixada no local, complementando a fixação com o parafuso previamente instalado. Em seguida, a janela craniana é fechada com pontos de sutura, deixando a cânula exposta na parte externa da cabeça do animal. Ao término do procedimento, a concentração de isoflurano é reduzida gradualmente, permitindo que o animal desperte de forma lenta e controlada.

D. Cuidados Pós-Operatórios

Após os procedimentos cirúrgicos de implante da cânula ou fibra óptica, os animais foram alojados individualmente por uma semana, a fim de garantir a recuperação pós-cirúrgica e evitar lesões decorrentes de interação social. A analgesia pós-operatória foi realizada com meloxicam na dosagem de $0,5$ mg/kg. A primeira alimentação oferecida após a cirurgia foi umedecida, a fim de facilitar a hidratação, a deglutição e a mastigação do animal durante a fase inicial de recuperação. Todos os procedimentos de manipulação animal, incluindo os cuidados pós-operatórios, ocorreram exclusivamente durante o ciclo claro, entre 8 e 17 horas, minimizando o estresse associado à manipulação fora do período de atividade natural dos animais.

Após o período de uma semana de recuperação, os animais permaneceram com as fibras implantadas até o final do protocolo experimental e foram gradualmente habituados à estimulação optogenética, conectados ao estimulador, sem anestesia, com movimento livre em aparato de campo aberto e no aparato comportamental utilizado para o estudo de memórias instrumentais, antes do início efetivo dos experimentos.

RESULTADOS

Em síntese, 28 ratos *Wistar* foram submetidos aos procedimentos cirúrgicos descritos no protocolo, dos quais 13 animais (46,4%) foram submetidos à cirurgia combinada (implante de cânula e infusão viral na mesma sessão) e 15 animais (53,6%) foram submetidos a duas cirurgias separadas (infusão viral e, posteriormente, implante de cânula).

Do total da amostra, 24 animais (85,7%) sobreviveram ao procedimento cirúrgico, enquanto 4 animais (14,3%) foram a óbito, 3 (10,7%) durante a indução ou manutenção anestésica e 1 (3,6%) durante o próprio ato cirúrgico.

A taxa de sobrevivência foi semelhante entre os dois grupos:

- **Grupo submetido à cirurgia combinada ($n = 13$):** 11 animais (84,6%) sobreviveram, com 1 óbito por complicação anestésica (7,7%) e 1 óbito durante a cirurgia (7,7%);
- **Grupo submetido a cirurgias separadas ($n = 15$):** 13 animais (86,7%) sobreviveram, com 2 óbitos por complicação anestésica (13,3%) e nenhum óbito durante o ato cirúrgico propriamente dito.

Não houve, portanto, diferença expressiva na taxa de sobrevivência entre as duas abordagens cirúrgicas, embora a maior parte dos óbitos em ambos os grupos tenha ocorrido durante a fase anestésica, e não durante a manipulação cirúrgica em si, achado que reforça a importância do monitoramento contínuo dos parâmetros fisiológicos (oximetria, temperatura retal e frequência cardíaca) descrito na seção de metodologia.

A distribuição dos desfechos por tipo de procedimento está representada na Figura 7 abaixo.

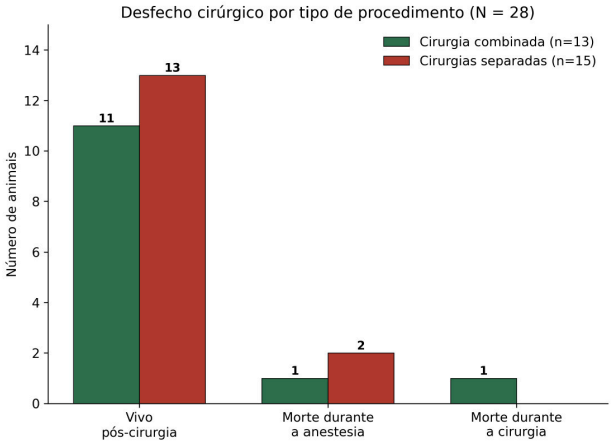
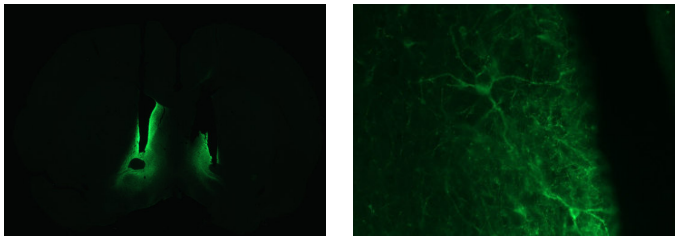


Fig. 7: Número de animais por desfecho cirúrgico (vivo pós-cirurgia, morte durante a anestesia ou morte durante a cirurgia), comparando o grupo submetido à cirurgia combinada (n = 13) e o grupo submetido a cirurgias separadas (n = 15).

Apesar da taxa de sobrevida similar, a avaliação técnica e operacional evidenciou vantagens expressivas da cirurgia combinada frente às intervenções separadas. O procedimento combinado apresentou um tempo cirúrgico médio de $204{,}09 \pm 21{,}38$ minutos ($SEM = 6{,}45$ min; intervalo de 160 a 234 minutos) em sessão única. Essa abordagem eliminou a necessidade de um segundo procedimento cirúrgico, dispensando uma nova montagem de campo estéril, nova indução anestésica e recalibração completa das coordenadas estereotáxicas em dias distintos. Além disso, a intervenção única reduziu o estresse tecidual cumulativo ao evitar a reabertura cirúrgica e a manipulação sobre tecido cicatricial craniano, concentrando o período inflamatório e de analgesia pós-operatória em uma única janela de recuperação.

Para validar o sucesso técnico e a precisão neuroanatômica do POP, os encéfalos dos animais sobreviventes foram submetidos à análise histológica pós-experimento por microscopia de fluorescência. A análise dos cortes coronais confirmou a acurácia estereotáxica do procedimento, evidenciando o trajeto de penetração e a terminação da cânula e agulha de infusão centrados nas coordenadas planejadas para o núcleo accumbens (NAc; AP +1,4 mm; ML ±1,4 mm; DV -6,8 mm), sem desvios laterais acentuados ou hemorragias ventriculares extensas.

Conforme ilustrado na Figura 8a, observa-se uma área de fluorescência bilateral compatível com o halo de difusão de 0,5 µL do vetor AAV-EF1a-NpHR3.0-EYFP restrito à citoarquitetura do NAc. Em aumento de 40× (Figura 8b), constatou-se transdução neuronal robusta e integrada, marcada pela fluorescência na membrana somática e ramificações dendríticas/axonais, demonstrando viabilidade celular e ausência de necrose focal excessiva no sítio de infusão. Esses achados confirmam que o protocolo padronizado não apenas preserva a sobrevida dos animais, mas atinge acurácia estereotáxica satisfatória e expressão gênica funcional no alvo de interesse.



(a) (b)

Fig. 8: Confirmação histológica do posicionamento e expressão viral no núcleo accumbens (NAc). (a) Corte coronal em aumento de 5× evidenciando o trajeto das cânulas e o halo de infusão do vetor AAV-EF1a-NpHR3.0-EYFP. (b) Fotomicrografia em 40× demonstrando neurônios expressando robustamente o fluoróforo EYFP.

Conclusão

O presente estudo teve como objetivo apresentar e padronizar um Procedimento Operacional Padrão (POP) para a cirurgia estereotáxica de infusão viral e implante de cânula em ratos Wistar, voltado à padronização de etapas até então descritas de forma heterogênea na literatura. A aplicação do protocolo em 28 animais resultou em uma taxa de sobrevivência de 85,7%, com desfechos semelhantes entre as abordagens de cirurgia combinada (84,6%) e cirurgias separadas (86,7%).

A concentração dos óbitos na fase anestésica, e não durante a manipulação cirúrgica propriamente dita, indica que os pontos críticos de atenção do protocolo estão relacionados sobretudo ao ajuste individualizado da anestesia e ao monitoramento fisiológico contínuo, orientando refinamentos futuros no cálculo das doses e no manejo anestésico. A comparação entre as abordagens demonstrou que a cirurgia combinada representa a estratégia metodologicamente superior: com duração média de $204,09 \pm 21,38$ minutos em etapa única, otimiza o tempo total de bancada, diminui o consumo de insumos cirúrgicos e anestésicos, e restringe a dor e o estresse pós-operatório a um único ciclo de recuperação.

Aliada à sobrevida, a confirmação histológica do sítio cirúrgico e da expressão do vetor viral no núcleo accumbens valida o sucesso neuroanatômico e experimental do POP. Dessa forma, o protocolo proposto comprova sua reprodutibilidade e eficiência técnica, oferecendo um guia confiável para estudos de neuromodulação e estimulação optogenética em roedores.

References

[1] ENNO, L.; YIZHAR, O.; DEISSEROTH, K. Optogenetics: 10 years of microbial opsins in neuroscience. Annual Review of Neuroscience, v. 34, n. 1, p. 389-412, 2011.

[2] DAVID, O. et al. Stereotaxic surgery for implantation of guide cannulas for microinjection into the brain in rats. Journal of Visualized Experiments, n. 150, e59833, 2019.

[3] Improving Stereotaxic Neurosurgery Techniques and Procedures Greatly Reduces the Number of Rats Used per Experimental Group — A Practice Report. PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8465152/

[4] Protocol for a systematic review of good surgical practice guidelines for experimental rodent surgery. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC9644723/
[5] Implementing the 3Rs in Laboratory Animal Research. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC10571647/
[6] CHU, V.; HUANG, C. L.; JOU, S. Noninvasive electrocardiography in mice and rats. Nature Protocols, v. 1, n. 2, p. 583–586, 2006.

EEG-based classification model for predicting tDCS responders in the treatment of neuropathic pain

1st Hanna Barbara Rodrigues Morais
Santos Dumont Institute
Macaíba, Brazil
0009-0000-4847-6081

2nd Marília da Costa Paiva
Santos Dumont Institute
Macaíba, Brazil
0000-0002-5047-6010

3rd Uliane Monteiro da Silva
Uninassau University Center
Natal, Brazil
0009-0004-3329-0611

4th Lucivanio Varela da Silva
Federal University of Rio Grande do Norte
Natal, Brazil
0009-0005-6930-5878

5th Fernando da Silva Fiorin
Santos Dumont Institute
Macaíba, Brazil
0000-0002-6517-7160

6th Caroline Cunha do Espírito Santo
Santos Dumont Institute
Macaíba, Brazil
0000-0001-8657-9532

Abstract*—Introduction: Neuropathic pain (NP) is a complex clinical condition that affects approximately 58% of patients with spinal cord injury (SCI) and is often refractory to treatment. In such cases, transcranial direct current stimulation (tDCS) offers a potential alternative to improve symptoms. With this non-invasive neuromodulation modality, therapeutic effects can be enhanced through multiple sessions with short intervals between them; however, this schedule can limit access to and continuity of treatment for this patient population. Remotely supervised tDCS (RS-tDCS) has emerged as a solution, allowing home-based application monitored remotely by a trained professional; it holds promise for modulating central pain pathways and fostering better adherence and continuity of treatment. Given its applicability, individual neurophysiological factors—such as alterations in cortical activity measured by electroencephalography (EEG)—may influence the therapeutic response. Consequently, investigating potential cortical biomarkers could serve not only to characterize pain, but also to predict response to remote NP treatment. Objective: This exploratory study aimed to develop a classifier using supervised machine learning to identify predictive biomarkers within the theta, alpha, and beta frequency bands capable of distinguishing between responders and non-responders to RS-tDCS in individuals with SCI and neuropathic pain. Methodology: The cohort study included 14 participants with chronic neuropathic pain following SCI. The treatment regimen consisted of 20 RS-tDCS sessions (10 active and 10 sham) administered over seven weeks. Analgesic response was assessed using the Numerical Rating Scale (NRS), with a reduction of $\geq$ 1.8 points defining a responder. Resting-state 16-channel EEG was recorded before the intervention. Following preprocessing and Power Spectral Density (PSD) extraction, Linear SVM, RBF SVM, and LDA classifiers were used to predict treatment response based on the alpha band. Results: The mean PSD in the responder group was observed to be higher in the low-alpha frequency range. However, the classification models performed poorly: Linear SVM and LDA achieved only 33% accuracy (F1-score of 0.5), while RBF SVM achieved 16.7%. Confusion matrices revealed that all models struggled to correctly discriminate non-responders. Conclusion: Although RS-tDCS reduced pain intensity in the majority of participants, alpha-band EEG characteristics, when analyzed in isolation, did not demonstrate sufficient predictive accuracy to distinguish responders from non-responders. The limited performance of the classifiers was attributed to the small sample size, class imbalance, and high inter-individual variability in EEG signals. Future studies should utilize larger samples and incorporate additional metrics, such as peak frequency, to develop more robust predictive models.

***Index Terms*—Spinal Cord Injury (SCI), Neuropathic Pain (NP), Transcranial Direct Current Stimulation (tDCS)**

I. Introduction

Spinal cord injury (SCI) is a complex clinical condition associated with multifaceted challenges, including the management of neuropathic pain, resulting from dysfunction of somatosensory system [1]–[3]

The global prevalence of neuropathic pain in SCI is estimated at 58%, often presenting as refractory and difficult to manage [4]. Neuromodulation strategies, such as transcranial direct current stimulation (tDCS), have been employed to modulate pain processing within central pathways [5], [6].

tDCS is a noninvasive technique of cortical modulation that involves the application of anodal stimulation to the primary motor cortex (M1) [7] and has gained popularity because of its ease of application, safety profile, and the fact that side effects are mild and transient [8].

Although tDCS has shown potential for pain management [9], [10], both external and internal factors may contribute to poor treatment response. Poor adherence is an external factor that can be partially addressed through remote supervision (RS), which facilitates home-based tDCS administration [11], [12]. However, less attention has been given to individual neurophysiological factors that may influence RS-tDCS response and limit expected pain-relief outcomes. For example, alterations in cortical activity, assessed by electroencephalography (EEG), have been investigated as potential markers of therapeutic response [9].

EEG allows for the investigation of changes in cortical activity frequency bands and the identification of potential biomarkers linked to neuromodulation responses [13], [14].

This work was supported by the National Council for Scientific and Technological Development (CNPq) through its postgraduate program (Process No. 131182/2026-9).

Thus, identifying electrophysiological characteristics associated with the analgesic response to RS-tDCS can contribute to both understanding the neural mechanisms involved in neuropathic pain treatment and to decisions regarding the continued use of this approach [9], [10].

Studies have shown a tendency for increased low-alpha frequency (8–10 Hz) in patients with neuropathic pain, whereas high-alpha activity typically exhibits the opposite pattern [13]. Given this potential correlation between pain and alterations in this alpha-frequency sub-band, alpha activity may serve as a significant biomarker for characterizing neuropathic pain and predicting treatment response.

Consequently, machine learning-based methods are required to identify complex EEG patterns capable of predicting which individuals might achieve effective outcomes with RS-tDCS for neuropathic pain. In this context, this exploratory study aimed to develop a classifier using supervised machine learning to identify potential predictive biomarkers within the theta, alpha, and beta frequency bands able to distinguishing individuals with SCI between responders and non-responders to RS-tDCS based on the reduction in pain intensity in a protocol to management neuropathic pain.

II. Methods

A. Study type and ethical aspects

This cohort study was conducted using a non-public database derived from a single-group quasi-experimental study previously approved by the Research Ethics Committee of the Santos Dumont Institute (ISD) under the registration number CAAE 69887123.8.0000.0129.

B. Participants and clinical information

The data set consisted of 14 participants with neuropathic pain after SCI. The sample was recruited by convenience sampling from a list of users at rehabilitation centers located in the 7th Health Region of Rio Grande do Norte, Brazil, between October 2024 and January 2026. The eligibility criteria were as follows:

- Inclusion criteria: individuals of both sexes (aged over 18 years) diagnosed with chronic neuropathic pain—defined as a score of ≥ 4 on the Douleur Neuropathique 4 questions (DN4) scale—located at or below the level of the SCI (classified according to the International Standards for Neurological Classification of Spinal Cord Injury - ISNCSCI); pain intensity of at least 4 out of 10 on the Numeric Rating Scale (NRS); not currently undergoing pharmacological treatment or, if undergoing treatment, maintaining a stable therapeutic regimen for at least two weeks before the study start; access to the internet and basic proficiency with digital tools (smartphone, computer, and/or tablet, provided at no cost or personally owned); and having a companion or caregiver who commits to being present at all stages of the study.
- Exclusion criteria: individuals with severe pain of non-neuropathic origin (NRS ≥ 7 with high interference in daily activities); intolerance to the target tDCS current of 2mA during the in-person training period (NRS ≥ 7 regarding equipment tolerance); alcohol or other substance abuse; history of traumatic brain injury with residual sequelae; pregnancy or breastfeeding; epilepsy; metallic implants in the head and/or neck; use of a pacemaker; compromised skin integrity at or near the current application site; changes to pharmacological treatment during the study; and inability to adhere to the pre-established data collection schedule.

C. RS-tDCS treatment with remote supervision protocol

Individuals with neuropathic pain underwent intensive treatment lasting seven weeks, comprising 20 sessions of RS-tDCS: 10 sessions of sham RS-tDCS and 10 sessions of active RS-tDCS. To ensure safe administration of tDCS in the home environment, in-person training was conducted with participants and caregivers during the first week of intensive treatment. Participants were assessed using NRS and EEG before intervention (Assessment 1 - A1), post-sham (Assessment 3 - A3), and post-active RS-tDCS. The change in NRS scores was calculated as the delta (Δ) between post-treatment (A3) and pre-treatment (A1) scores ($\Delta\text{NRS} = \text{A3} - \text{A1}$). For the prediction model to explore responders and non-responders to RS-tDCS, only data from the A1 assessment were used.

D. Outcomes measures

- Numerical rating scale: Participants were considered RS-tDCS responders if they demonstrated a satisfactory analgesic effect, defined as a reduction in neuropathic pain intensity equal to or greater than the minimal clinically important difference (MCID) for the NRS in the SCI population (a reduction of ≥ 1.8 points) [15] at A3 when compared to A1.

- Electroencephalography (EEG): Brain rhythms were assessed using a 16-channel EEG connected to the V-Amp system (Brain Products), with wet adhesive electrodes (actiCAP Xpress, Brain Products) positioned according to the international 10-20 system over the FP1, FP2, F3, F4, FC1, FC2, C4, C3, Cz, CP2, CP1, P4, Pz, P3, O2, and O1 regions. The EEG electrodes were attached to the scalp by applying conductive gel at each electrode site, following prior cleansing of the scalp with alcohol and abrasive gel. Impedance was maintained below 10 kΩ, and the sampling rate was set at 512 Hz. The protocol included recording resting-state EEG with eyes open for a period of 2 minutes.

E. Signal Preprocessing

The EEG data were band-pass filtered within the 4–35 Hz frequency range, and a 60 Hz notch filter was applied to remove low-frequency components and high-frequency noise. Subsequently, spatial filtering was performed using the Common Average Reference (CAR) method, which involves calculating the average across all channels and subtracting the

mean from the signal at each electrode at every time point, thereby generating a re-referenced EEG signal.

The frequency bands of interest were theta (4–7 Hz), alpha (8–12 Hz), and beta (13–30 Hz) [16]. A 10-second windowing process was applied; the first two windows were discarded to account for the EEG stabilization period, and the final window was removed to standardize the data duration. Next, Welch's method was applied using 1-second Hamming windows with 50% overlap. Finally, Power Spectral Density (PSD) curves were extracted by averaging the windowed data from both groups—responders and non-responders—across the three target frequency bands.

F. Prediction model

After extracting features from the theta, alpha, and beta bands for each window, a matrix representing the subjects' averages was constructed for each subject. In this matrix, features from all windows belonging to the same individual were grouped, and the mean value for each feature was determined. Consequently, each row of the matrix represented a subject, while the columns represented the frequency power values for all analyzed channels. After this analysis, only the low-alpha and high-alpha bands were considered for the processing stage of the prediction model, since these frequencies showed the greatest potential to differentiate between responders and non-responders.

Next, the subjects were divided into a training group comprising four randomly selected responders and four non-responders and a test group consisting of the remaining subjects. Before the classifier training began, the data, newly split into training and test sets, passed through the StandardScaler to prevent data leakage.

Three types of classifiers were employed: Linear Support Vector Machine (SVM), Radial Basis Function (RBF) SVM, and Linear Discriminant Analysis (LDA); the performance of each classifier was evaluated using accuracy and F1-score.

Finally, confusion matrices were generated for each classifier, allowing for distinction between true positives, true negatives, false positives, and false negatives.

III. Results

Fourteen participants (11 men and 3 women), with a mean age of 45.7 years, were included in the study. Of these 14 participants, 9 were classified as responders to RS-tDCS, and 5 were classified as non-responders (Figure 1). A summary of the participant demographics is presented in Table I.

A. Global PSD characteristics during resting-state EEG differed between responders and non-responders

We observed that mean PSD in the responder group was higher than in the non-responder group in the Low Alpha frequency (Figure 2) at A1. Mean PSD values in the High Theta and High Beta frequencies were also slightly higher for responders, although the difference was not as pronounced as that observed in the Alpha band. Based on this result, to confirm that the Low Alpha frequency can serve as a biomarker for individuals likely to respond positively to RS-tDCS for NP, we used three classifiers: linear SVM, SVM RBF, and LDA.

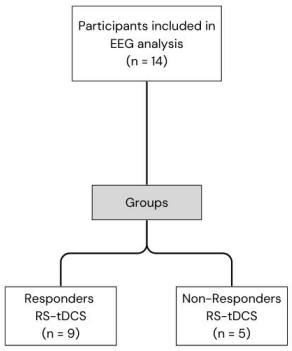


Fig. 1. Participants flowchart

TABLE I
DEMOGRAPHIC DATA

Variable		Mean	Std. deviation
Sex	Female	3.00 (21.42%)	
	Male	11.00 (78.57%)	
Age		45.7 (years old)	± 14.6
ASIA (ISNCSCI)	Neurological level		
	C1-C4	6.00 (42.8%)	
	C5-C6	1.00 (7.14%)	
	C7-C8	0	
	T1-T8	3.00 (21.42%)	
	T9-T12	4.00 (28.57%)	
	AIS		
	Complete	10.00 (71.42%)	
	Incomplete	4.00 (28.57%)	

Note: Abbreviations: ASIA (ISNCSCI), American Spinal Injury Association (International Standards for Neurological Classification of Spinal Cord Injury); AIS, ASIA Impairment Scale.

B. Associating EEG cortical activity as a predictive model of responders or non-responders to RS-tDCS

Among the classifiers trained using the subject-average model, the Linear SVM and LDA achieved the best results, with 33% accuracy and an F1-score of 0.5 for both (Table II). The RBF SVM achieved 16.7% accuracy and an F1-score of 0.28. These values indicate that, when analyzed in isolation, the features extracted from the alpha band were insufficient to discriminate between responders and non-responders.

The confusion matrix for the three classifiers presented in figure 3 showed that all models struggled to discriminate non-responders, as they misclassified data belonging to this group. Regarding the responder group, the Linear SVM and LDA classifiers correctly identified two out of five subjects, whereas the RBF SVM correctly identified only one.

IV. Discussion

We found that RS-tDCS is able to reduce the neuropathic pain intensity in most investigated participants, but unfortu-

TABLE II
COMPARISON BETWEEN THE CLASSIFIERS

Classifier	Accuracy	F1-score
SVM Linear	0.333	0.500
SVM RBF	0.167	0.286
LDA	0.333	0.500

nately the EEG features did not demonstrate sufficient predictive accuracy to distinguish between responders and non-responders to RS-tDCS for NP treatment.

Identifying patterns in cortical activity could contribute to the personalization of interventions, thereby increasing the precision and efficacy of neuromodulation treatment [10]. In this context, machine learning techniques have shown promise in recognizing patterns and identifying biomarkers capable of predicting protocol success based on each patient's individual neurophysiological characteristics.

According to Mussigmann et al. (2022) [13], one of the most consistent biomarkers for neuropathic pain in the literature is a Peak Frequency Shift (PFS) within the theta-alpha bands (4–12 Hz), characterized by a "slowing" of the alpha rhythm. This shift occurs when the brain, instead of oscillating at the normal (faster) alpha rhythm, begins to oscillate predominantly at slower frequencies (theta and low alpha), which is considered a hallmark of thalamocortical dysrhythmia, a process that involves a disruption of the harmony between the brain and the thalamus. This alters brain oscillation patterns and contributes to changes in pain perception. Studies have also observed a possible association between neuropathic pain and reduced activity in the upper-alpha and lower-beta bands; however, the relationship between pain intensity and alterations in the alpha band remains controversial [13].

Given this background, the present study investigated the spectral characteristics of the low-alpha frequency band derived from resting-state EEG as potential predictive biomarkers for the RS-tDCS response in individuals with SCI suffering from neuropathic pain. Although exploratory analysis revealed differences in alpha-band PSD between groups, the classification models (linear SVM, RBF SVM, and LDA) performed poorly when analyzing mean PSD values per subject. This

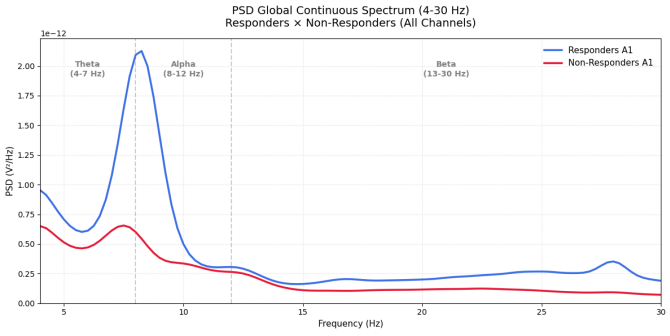


Fig. 2. Power Spectral Density (PSD) global mean for responders (blue) and non-responders (red) in Theta, Alpha and Beta frequencies.

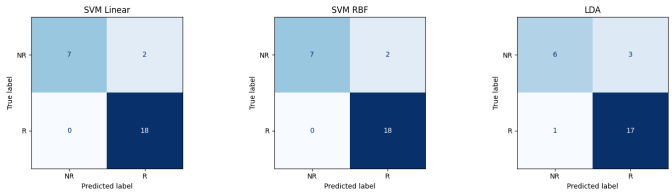


Fig. 3. Confusion matrix. NR, non-responder; R, responder; SVM Linear, Linear Support Vector Machine; RBF SVM, Radial Basis Function Support Vector Machine; LDA, Linear Discriminant Analysis

suggests that, when investigated in isolation, these characteristics were not sufficiently consistent to enable classification between responders and non-responders to tDCS. The linear SVM served as the starting point for the analysis, allowing for the establishment of a performance baseline before testing new classifiers [17]. As a simpler model requiring the tuning of only one hyperparameter—the soft-margin constant (C)—it offers less flexibility and a lower risk of overfitting compared to non-linear SVM models with more tuning parameters, such as RBF. Consequently, the linear SVM classifier is presumed to have outperformed the RBF SVM because the sample's features and classes are predominantly linear.

In addition to SVM, LDA was also employed, which is suitable for data classification and dimensionality reduction. It relies on analyzing between-class and within-class variance to establish a distance that allows for a clear separation between groups to be visualized [18]. When applied to this sample, the model achieved an accuracy similar to that of the Linear SVM (33%), which is below the chance level. Thus, it can be inferred that the features extracted based solely on the alpha-band PSD were insufficient to serve as a predictive biomarker.

Therefore, to develop a more robust predictive model, more complex models that incorporate other features, such as PFS, must be used. Furthermore, it is crucial to utilize a larger dataset with more balanced classes, which would enable classifiers to be trained by considering not only different data features but also each individual's neurophysiological response to neuromodulation, thereby optimizing the models' ability to accurately discriminate between groups.

A. Limitations

The relatively small number of participants constitutes a significant limitation of this study, as it may reduce statistical power and limit the generalizability of the findings. The imbalance between the Responder (9) and Non-responder (5) classes hindered the classifier's ability to identify the smaller group. Given the lack of additional available participants, the results should be considered preliminary and require confirmation in larger, independent cohorts.

Furthermore, EEG datasets often contain multiple segments from the same participant; this can lead to overly optimistic performance estimates if segments from the same individual are included in both the training and testing sets, potentially resulting in data leakage. Therefore, participant-level validation, such as leave-one-subject-out cross-validation or group K-fold,

may be preferable for assessing generalization to participants not included in the classifier's training. In the present study, processing using LOSO was performed (data not shown), and the results indicated significant data leakage. Thus, although other strategies may be more suitable for a more precise analysis of prediction data, the sample size limitation remains an issue that needs to be addressed in future studies.

V. Conclusion

The trained classifiers struggled to distinguish between responders and non-responders due to the small sample size and the high heterogeneity of cortical responses related to neuromodulation. Future studies require a more robust sample, as well as the analysis of new characteristics, such as peak frequency. Therefore, the present analyses do not allow us to determine which frequency bands are specifically involved in the response to treatment for neuropathic pain.

References

[1] T. S. Jensen *et al.*, "A new definition of neuropathic pain," *Pain*, vol. 152, no. 10, pp. 2204–2205, Oct. 2011.

[2] N. B. Finnerup *et al.*, "Neuropathic pain in spinal cord injury: from pathophysiology to treatment," *Pain*, vol. 157, no. 3, pp. 340–349, Mar. 2016.

[3] S. N. Raja *et al.*, "The revised IASP definition of pain: concepts, challenges, and compromises," *Pain*, vol. 161, no. 9, pp. 1976–1982, Sep. 2020.

[4] C. Hunt *et al.*, "Prevalence of chronic pain after spinal cord injury: a systematic review and meta-analysis," *Regional Anesthesia and Pain Medicine*, vol. 46, no. 4, pp. 328–336, Apr. 2021.

[5] L. Garcia-Larrea *et al.*, "At-home cortical stimulation for neuropathic pain: a feasibility study with initial clinical results," *Neurotherapeutics*, vol. 16, no. 4, pp. 1198–1209, Oct. 2019.

[6] N. E. O'Connell *et al.*, "Non-invasive brain stimulation techniques for chronic pain," *Cochrane Database of Systematic Reviews*, no. 3, 2018.

[7] A. J. Woods *et al.*, "A technical guide to tDCS, and related non-invasive brain stimulation tools," *Clinical Neurophysiology*, vol. 127, pp. 1031–1048, 2016, doi: 10.1016/j.clinph.2015.11.012.

[8] A. C. M. Piovesan, H. Z. Borba, I. D. P. Cadan, G. S. B. Nogueira, E. S. B. Nogueira, T. C. Vilhena, L. R. de Rosse, S. C. Mathias, and N. F. P. Dircksen, "Estimulação transcraniana em dores neuropáticas: Uma revisão sistemática," *Brazilian Journal of Implantology and Health Sciences*, vol. 7, no. 8, pp. 1362–1405, Aug. 2025, doi: 10.36557/2674-8169.2025v7n8p1362-1405.

[9] J. Ma, Z. Wang, T. Cheng, Y. Hu, X. Qin, W. Wang, G. Yu, Q. Liu, T. Ji, H. Xie, D. Zha, S. Wang, Z. Yang, X. Liu, L. Cai, Y. Jiang, H. Hao, J. Wang, L. Li, and Y. Wu, "A prediction model integrating synchronization biomarkers and clinical features to identify responders to vagus nerve stimulation among pediatric patients with drug-resistant epilepsy," *CNS Neuroscience & Therapeutics*, vol. 28, no. 11, pp. 1838–1848, Nov. 2022, doi: 10.1111/cns.13923.

[10] A. Vučković, V. J. F. Gallardo, M. Jarjees, M. Fraser, and M. Purcell, "Prediction of central neuropathic pain in spinal cord injury based on EEG classifier," *Clinical Neurophysiology*, vol. 129, no. 8, pp. 1605–1617, Aug. 2018, doi: 10.1016/j.clinph.2018.04.750.

[11] S. Carvalho *et al.*, "Feasibility and safety of home-based transcranial direct current stimulation for neuropathic pain in patients with spinal cord injury: a pilot study," *Journal of NeuroEngineering and Rehabilitation*, vol. 15, no. 1, p. 24, 2018.

[12] L. E. Charvet *et al.*, "Remotely supervised transcranial direct current stimulation increases the benefit of at-home cognitive training in multiple sclerosis," *Neurology*, vol. 90, no. 24, pp. e2105–e2111, 2018.

[13] T. Mussigmann, B. Bardel, and J.-P. Lefaucheur, "Resting-state electroencephalography (EEG) biomarkers of chronic neuropathic pain: A systematic review," *NeuroImage*, vol. 258, Art. no. 119351, 2022, doi: 10.1016/j.neuroimage.2022.119351.

[14] A. Vučković, M. K. H. Altaleb, M. Fraser, C. McGeady, and M. Purcell, "EEG correlates of self-managed neurofeedback treatment of central neuropathic pain in chronic spinal cord injury," *Frontiers in Neuroscience*, vol. 13, Art. no. 762, 2019, doi: 10.3389/fnins.2019.00762.

[15] M. A. Hanley *et al.*, "Clinically significant change in pain intensity ratings in persons with spinal cord injury or amputation," *The Clinical Journal of Pain*, vol. 22, no. 1, pp. 25–31, Jan. 2006.

[16] C. S. Nayak and A. C. Anilkumar, "Normal EEG waveforms," in *StatPearls [Internet]*, Treasure Island (FL): StatPearls Publishing, 2026. [Updated Aug. 3, 2025]. Available: https://www.ncbi.nlm.nih.gov/books/NBK539805/

[17] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data Mining Techniques for the Life Sciences*, O. Carugo and F. Eisenhaber, Eds., *Methods in Molecular Biology*, vol. 609. New York, NY, USA: Humana Press, 2010, pp. 223–239, doi: 10.1007/978-1-60327-241-4_13.

[18] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis: A brief tutorial," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, MS, USA, Tech. Rep., 1998.

EEG-Based Robotic Hand Control During Motor Execution and Motor Imagery

Sousa, M. L. L.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
maria.lisboa.sousa22@gmail.com

Carneiro, H. A. B.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
heloisacarneiro70@gmail.com

Souza, A. O.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
arthursouzaoliveira250@gmail.com

Figueiredo, L. E. T. C.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
cassianolynnda@gmail.com

Fonseca, M. C. P.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
conceicao@ufpa.br

Miziara, I. M.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
immiziara@ufpa.br

Abstract*—Loss of hand motor function substantially compromises independence and quality of life, making prosthetic and orthotic devices important resources for individuals with amputations, spinal cord injuries, stroke, and neuromuscular disorders. However, conventional myoelectric control depends on the presence of functional residual musculature, which limits its use in these populations. As an alternative, electroencephalography (EEG)-based brain-computer interfaces (BCIs) enable non-invasive detection of movement intention directly from cortical activity. This study presents a low-complexity and accessible EEG-based system capable of controlling an anthropomorphic robotic hand using signals acquired during motor execution and motor imagery tasks. The proposed architecture is organized into two sequential classification stages. The first stage distinguishes rest from motor activity and provides the command used to actuate the robotic hand. The second stage, applied to segments classified as movement, discriminates motor execution from motor imagery using features derived from *Common Spatial Patterns* (CSP) and is evaluated *offline*. In the current processing pipeline, training uses 1 s sliding windows with a 0.1 s step, spatial preprocessing is kept consistent between training and validation, and temporal decision stabilization is applied. In addition to demonstrating robotic-hand actuation, an event-based quantitative evaluation was performed using the LED synchronization channel exclusively as a temporal reference for comparison with classifier output. The first stage achieved an accuracy of 91.08%, whereas the second stage achieved 71.43%. In the event-based evaluation, 26 of 30 activation commands and 28 of 30 return-to-rest commands were detected, corresponding to detection rates of 86.7% and 93.3%, respectively, and an overall detection rate of 90.0%. A total of 61 isolated false positives were recorded, equivalent to 11.6 false positives per minute of recording. Computational latency between data-block availability and classification output had a mean of 128.5 ms, a median of 112.9 ms, and a 95th percentile of 209.8 ms. These results demonstrate the feasibility of the proposed approach while also highlighting the need to reduce false activations and optimize latency for assistive and motor-rehabilitation applications.

***Index Terms*—brain-computer interface, EEG, motor imagery, motor rehabilitation, real-time processing, robotic hand**

I. Introduction

This study presents a real-time EEG-based system for controlling a 3D-printed anthropomorphic robotic hand during ME and MI tasks. Unlike recent BCIs that rely on high-density EEG acquisition and advanced decoding techniques, the proposed framework prioritizes accessibility, lower hardware complexity, and reproducibility by combining a clinical EEG acquisition system with an 11-channel montage based on the international 10-20 system.

The proposed framework combines established preprocessing, feature-extraction, and classification methods in a reproducible two-stage architecture. The first stage distinguishes rest from movement by grouping ME and MI into a single movement class and provides the command used to actuate the robotic hand. The second stage discriminates ME from MI using *Common Spatial Patterns* (CSP), with performance evaluated *offline*. Training is performed using sliding windows synchronized with experimental events, whereas continuous operation uses only EEG data for decision making. Operational performance is additionally evaluated by quantifying command detection, false activations, and computational latency, with the LED synchronization channel used only as a temporal reference for validation and not as an input to the classifier.

This work represents an initial step toward validating the proposed approach; with further testing across diverse conditions and populations, the framework may eventually contribute to rehabilitation programs and to the development of prosthetic and orthotic devices.

The objectives of this study are: (i) to develop and implement a two-stage EEG-based classification architecture capable of distinguishing rest from movement and, subsequently, motor execution from motor imagery, using a reduced 11-channel montage; (ii) to evaluate the *offline* performance of both classification stages; and (iii) to assess the operational feasibility of the system in a real-time (or near-real-time,

sequential-replay) setting by quantifying command detection rates, false activations, and computational latency during control of a 3D-printed robotic hand.

II. Materials and Methods

The development of this study comprised complementary experimental stages. The first stage involved EEG acquisition under three conditions: motor execution (ME), action observation with simultaneous execution, and motor imagery (MI). ME and MI recordings were used to characterize movement related cortical activity and to train the proposed classification model, whereas the action observation and execution condition served as a familiarization stage before the motor imagery task. The system was subsequently used for *online* control of an anthropomorphic robotic hand based on motor activity decoded from EEG. Finally, a complementary quantitative evaluation was performed through sequential replay of an independent recording, preserving block by block processing and using the LED synchronization channel exclusively as a temporal reference for measuring command detection, missed commands, false activations, and computational latency.

Fig. 1 illustrates the overall workflow of the proposed system, from EEG acquisition to real time robotic hand control, and summarizes the methodological stages described in the following sections.

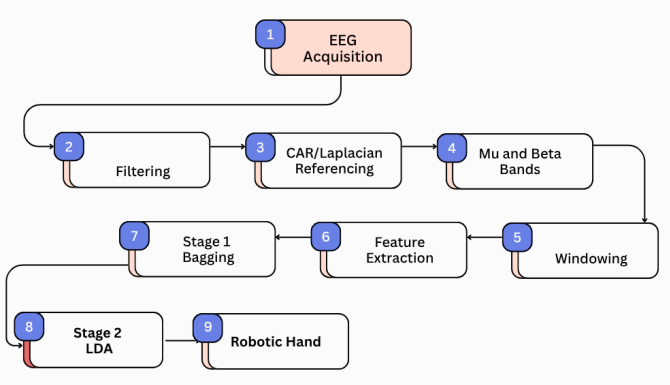


Figure 1. Overview of the proposed methodology, including EEG acquisition, preprocessing, frequency band decomposition, temporal windowing, feature extraction, two stage classification, robotic hand actuation, and event based quantitative evaluation.

A. Participants and Experimental Setting

A total of three healthy male volunteers, aged between 20 and 50 years, were recruited to participate in the experimental procedures. Participants were selected by convenience sampling based on their availability, willingness to participate, and self reported good health. Participation was voluntary.

Before data acquisition, each volunteer received a detailed explanation of the study objectives, experimental procedures, EEG recording protocol, and the tasks that would be performed during the experiment. Participants were given the opportunity to clarify any questions before the beginning of the recordings. After agreeing to participate in the study, all volunteers provided written informed consent by signing the Informed Consent Form. Personal information and participant identities were kept confidential throughout the study, and the data were handled in a manner that preserved participant privacy.

All recording sessions were conducted in a controlled indoor environment. Participants were invited to sit comfortably in a chair and were positioned in a relaxed posture throughout the experiment. The room temperature was controlled to provide stable and comfortable recording conditions. Before each acquisition session, participants were instructed to remain as relaxed as possible, avoid unnecessary body movements, and follow the visual cues presented during the experimental protocol.

EEG data were acquired across multiple recording sessions. In each session, each participant completed three experimental conditions synchronized with a visual stimulus provided by a light emitting diode (LED). The experimental sequence was designed to investigate cortical activity associated with actual hand movement and imagined hand movement while maintaining a standardized timing structure across conditions.

During the first condition, the participant observed and simultaneously performed repeated hand opening and closing movements. During the second condition, the participant observed the same movement while performing motor imagery without physical execution. During the third condition, the participant performed motor imagery without visual observation of the movement.

The first condition involved actual motor execution and therefore provided reference epochs with precisely defined movement periods, allowing characterization of movement related cortical activity. The second condition served as a familiarization and imagery induction stage, allowing participants to associate the observed movement with its mental representation before performing motor imagery without visual guidance during the third condition.

B. EEG Signal Acquisition

EEG signals were acquired at a sampling rate of 500 Hz using an EMSA clinical electroencephalograph. A reduced 11 electrode montage was used, with electrodes positioned according to the international 10–20 system at scalp locations Fp1, Fp2, F3, F4, Fz, C3, C4, Cz, P3, P4, and Pz. The reference electrode was placed on the earlobe, and the ground electrode (GND) was positioned over the mastoid process.

The selected montage covered frontal regions, represented by Fp1, Fp2, F3, F4, and Fz; central sensorimotor regions, represented by C3, C4, and Cz; and parietal regions, represented by P3, P4, and Pz. The central electrodes were selected because of their proximity to cortical areas involved in hand motor control and their relevance for detecting event related desynchronization and synchronization associated with motor execution and motor imagery.

The frontal electrodes were included primarily to support the identification of activity potentially associated with ocular artifacts, whereas the parietal electrodes provided complementary information regarding the spatial distribution of sensori-

motor activity. This configuration was intended to preserve relevant cortical information while maintaining a reduced number of acquisition channels.

Before electrode placement, the participant was positioned comfortably in the chair, and the electrodes were arranged according to the international 10–20 system. After preparation of the EEG montage, signal quality was visually inspected before the beginning of the experimental protocol. Participants were instructed to minimize head, facial, and body movements during the recordings, except for the hand movements explicitly required during the motor execution condition.

Task timing was controlled by an Arduino Uno, which generated a digital signal synchronized with the visual LED stimulus. Logic high (*HIGH*) represented the active task period, whereas logic low (*LOW*) represented the resting period. This digital signal was connected to the FOTO trigger channel of the EEG acquisition system, allowing the onset and offset of each experimental task to be recorded simultaneously with the EEG signals.

The FOTO channel was used exclusively for temporal synchronization and subsequent identification of the experimental events. It was not used as an input feature for the EEG classification models. Therefore, classification decisions were based exclusively on EEG activity.

C. Experimental Protocol

Each experimental recording began with an initial resting baseline of 10 s. During this period, the participant remained seated in a relaxed position without performing the task. The baseline was followed by 30 task and rest cycles. Each cycle consisted of 3 s of active task followed by 7 s of rest, resulting in a cycle duration of 10 s.

Considering the initial 10 s resting baseline and the 30 experimental cycles, each condition had a total duration of 310 s, corresponding to 5 min 10 s. The same temporal structure was maintained for all participants and experimental conditions.

Three experimental conditions were performed:

- **Motor execution (ME):** when the LED was turned on, the participant performed a hand closing movement and maintained the hand in the closed position during the 3 s active period. When the LED was turned off, the participant relaxed the hand and returned to the resting state for 7 s.
- **Action observation and execution:** the participant observed a visual demonstration of repeated hand opening and closing movements while simultaneously reproducing the same movement. This condition was used as a familiarization stage to facilitate the association between visual observation, motor execution, and the mental representation of the movement. The data acquired during this condition were inspected as part of the experimental protocol but were not included as a separate class during classifier training.
- **Motor imagery (MI):** when the LED was turned on, the participant was instructed to vividly imagine performing the hand closing movement without producing any voluntary physical movement. During the 3 s active period, the participant maintained attention on the imagined movement. When the LED was turned off, the participant discontinued the imagery task and returned to a relaxed resting state for 7 s.

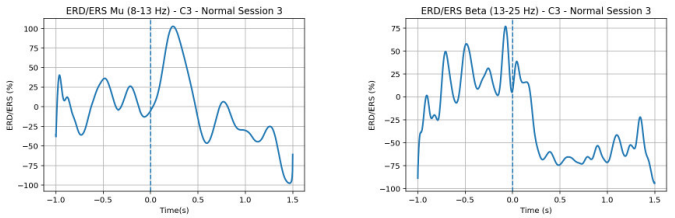
The same instructions were provided to all participants before each condition to maintain consistency across recording sessions. Participants were reminded to avoid unnecessary movements, facial contractions, excessive blinking, and changes in posture whenever possible because these factors could introduce artifacts into the EEG recordings.

The synchronization signal generated by the Arduino Uno enabled precise temporal alignment between the experimental protocol and the EEG recordings. The transition from *LOW* to *HIGH* indicated the beginning of each active period, whereas the transition from *HIGH* to *LOW* indicated the return to the resting period. These temporal markers were subsequently used to segment the EEG signals and identify the corresponding periods of motor execution, motor imagery, and rest.

For classifier development, recordings obtained during the motor execution and motor imagery conditions were used to generate the EEG segments associated with motor activity and rest. The action observation and execution condition was retained as a familiarization stage and was not treated as an independent classification class. This organization allowed the proposed system to focus on distinguishing resting cortical activity from movement related activity and, subsequently, on discriminating motor execution from motor imagery.

D. Preprocessing

A preliminary ERD/ERS analysis was performed in the mu (8 to 13 Hz) and beta (13 to 25 Hz) frequency bands. As illustrated in Fig. 2 for channel C3, the beta band exhibited desynchronization and synchronization patterns that were more consistent with those reported in the literature for motor tasks [1]. Consequently, subsequent analyses focused primarily on this frequency band, although both bands were retained during feature extraction.



(a) Mu band (8 to 13 Hz) (b) Beta band (13 to 25 Hz)

Figure 2. ERD/ERS relative to the resting baseline in channel C3 for the motor execution and motor imagery conditions. Time $t = 0$ corresponds to the onset of the visual stimulus. Negative values indicate ERD, whereas positive values indicate ERS. The figure illustrates the temporal evolution of the ERD/ERS response and does not correspond to the analysis windows used during classification.

Task timing was recovered from the *trigger* channel, FOTO, by detecting the rising edges of the LED synchronization

signal, which defined the onset of each experimental cycle. The sampling rate was estimated from the median interval between consecutive *triggers*, considering the known 10 s duration of each cycle. For the recordings used in this study, estimated values remained close to 500 Hz.

EEG signals were preprocessed using fourth-order Butterworth filters applied with zero-phase forward and reverse filtering. A 59 to 61 Hz band-stop filter attenuated 60 Hz power-line interference, followed by a 1 to 50 Hz band-pass filter to preserve the frequency range of interest. The signals, originally referenced to the earlobe, were subsequently rereferenced using the common average reference (CAR) and spatially filtered using a surface Laplacian to enhance local activity over the sensorimotor cortex.

Let $v_i[n]$ denote the signal recorded at channel i and sample n. The CAR subtracts from each channel the instantaneous mean across the M recording channels, according to Eq. (1),

$$v_i^{\mathrm{CAR}}[n] = v_i[n] - \frac{1}{M}\sum_{j=1}^{M} v_j[n], \tag{1}$$

thereby attenuating spatially widespread activity common to all electrodes. The surface Laplacian was then applied to emphasize local cortical activity by subtracting the average of neighboring channels, according to Eq. (2),

$$v_i^{\mathrm{LAP}}[n] = v_i^{\mathrm{CAR}}[n] - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} v_j^{\mathrm{CAR}}[n], \tag{2}$$

where $\mathcal{N}_i$ represents the set of neighboring electrodes associated with channel i, and $|\mathcal{N}_i|$ represents its cardinality. The adopted neighborhood sets are listed in Table I. Although electrodes Fp1, Fp2, P3, P4, and Pz were not used in the subsequent feature-extraction stage, they were retained in the acquisition montage because they contribute to computation of the Laplacian spatial filter.

Table I
NEIGHBORING CHANNEL SETS USED IN THE LAPLACIAN FILTER.

Channel	**Neighbors ($\mathcal{N}_i$)**	$\lvert\mathcal{N}_i\rvert$
F3	Fp1, C3, Fz	3
F4	Fp2, C4, Fz	3
Fz	F3, F4, Cz	3
C3	F3, P3, Cz	3
C4	F4, P4, Cz	3
Cz	Fz, C3, C4, Pz	4

The spatially filtered signals were then decomposed into the mu (8 to 13 Hz) and beta (13 to 25 Hz) frequency bands using band-pass filters.

To quantify the ERD/ERS response, the first 10 s of each recording were adopted as the resting baseline, from which the reference power in each frequency band was estimated using the Hilbert envelope. Let $x_b[n]$ denote the signal filtered in frequency band b, and let $\mathcal{H}\{\cdot\}$ denote the Hilbert transform. Instantaneous power was obtained from the envelope of the corresponding analytic signal according to Eq. (3),

$$P_b[n] = \left|x_b[n] + j\,\mathcal{H}\{x_b[n]\}\right|^2, \tag{3}$$

whereas the ERD/ERS response was calculated as the percentage change relative to the mean baseline power, P_b^{ref}, according to Eq. (4),

$$\mathrm{ERD/ERS}[n] = \frac{P_b[n] - P_b^{\mathrm{ref}}}{P_b^{\mathrm{ref}}} \times 100\%, \tag{4}$$

such that negative values indicate ERD and positive values indicate ERS.

Finally, signal segmentation was performed using 1 s windows with a 0.1 s step. For the movement class, windows were extracted from the interval between 0.5 and 2.0 s after each *trigger* to reduce the influence of the visual response immediately following LED onset and to concentrate the analysis on the period associated with sensorimotor modulation. For the rest class, windows were extracted from the interval between 5 and 7 s after each *trigger*. The same segmentation procedure was used for the motor execution and motor imagery recordings.

E. Sliding-Window Classification

EEG signals were classified in two sequential stages: rest versus movement, followed by ME versus MI. In the first stage, windows from the motor execution and motor imagery conditions were combined into a single movement class and compared with rest windows. In the second stage, only movement windows were used to discriminate motor execution from motor imagery.

The *offline* evaluation was performed using stratified five-fold cross-validation. Because 1 s windows were shifted in 0.1 s steps, consecutive windows had 90% temporal overlap. This characteristic increases the number of available samples but may also introduce dependence between neighboring windows. For this reason, accuracy values obtained with this strategy should be interpreted as a characterization of performance within the available recording set, while event-grouped or session-grouped validation remains important for evaluating generalization under more stringent conditions.

F. Feature Extraction

Different feature sets were adopted for the two decision stages.

For the first stage, rest versus movement, each 1 s window was represented by 11 features extracted from each of the six sensorimotor channels after filtering in the mu and beta bands: root mean square (RMS), variance, mean absolute value (MAV), waveform length (WL), energy, integrated EMG (IEMG), zero crossings (ZC), slope sign changes (SSC), ERD response estimated from the Hilbert envelope relative to the resting baseline, beta-band energy, and the ratio between mu- and beta-band power. Each window was therefore described by a 66-dimensional feature vector (6×11), combining conventional time-domain descriptors with ERD-related spectral features.

For the second stage, feature extraction was based on CSP, calculated independently for the mu and beta bands. Spatial filters were estimated from the mean covariance matrices of the two classes, normalized by their trace. Let X_e denote the data matrix of epoch e. The mean covariance matrix for class c is given by Eq. (5),

$$\Sigma_c = \frac{1}{|\mathcal{E}_c|} \sum_{e \in \mathcal{E}_c} \frac{X_e X_e^\top}{\operatorname{tr}(X_e X_e^\top)}, \qquad c \in \{1, 2\}, \tag{5}$$

and the spatial filters $\mathbf{w}$ were obtained by solving the generalized eigenvalue problem

$$\Sigma_1 \mathbf{w} = \lambda (\Sigma_1 + \Sigma_2) \mathbf{w}, \tag{6}$$

whose eigenvectors associated with the largest and smallest eigenvalues maximize the variance of one class while minimizing that of the other. For each band, the first two and last two spatial filters were retained, resulting in four spatial projections. Log-variance, log-energy, and RMS were extracted from each projected signal, and the features from both bands were concatenated into a single feature vector.

G. Two-Stage Classification

The proposed framework consisted of two sequential classification stages. The first discriminated rest from movement, whereas the second distinguished ME from MI. The classifiers adopted at each stage were selected to balance classification performance and computational efficiency. Computational efficiency was particularly important because the EEG system does not provide direct real-time data transmission. Instead, EEG data are continuously retrieved from a proprietary `.plg` file, introducing an inherent acquisition delay. Therefore, computationally demanding classifiers that could further increase latency and compromise robotic-hand actuation were avoided.

The first stage used a *bagging* ensemble of 200 decision trees, with a maximum depth of 20 levels per tree, whereas the second stage used Linear Discriminant Analysis (LDA) with a least-squares solver and automatic *shrinkage* regularization. Tree depth was limited to control ensemble complexity and reduce computational cost without changing the input feature structure.

LDA assumes that class-conditional feature distributions are Gaussian and share a common covariance matrix Σ. Under this assumption, the discriminant function for class c is given by Eq. (7),

$$\delta_c(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^\top \Sigma^{-1} \boldsymbol{\mu}_c + \ln \pi_c, \tag{7}$$

where $\boldsymbol{\mu}_c$ is the class mean vector and π_c is its prior probability. The feature vector $\mathbf{x}$ is assigned to the class that maximizes $\delta_c(\mathbf{x})$.

Although the dimensionality of the feature space was substantially lower than the number of training windows, the 90% temporal overlap between consecutive windows introduced dependence among observations, while some of the extracted features were potentially correlated. These factors can reduce the effective amount of independent information available for covariance estimation and affect the stability of the covariance matrix used by LDA. Therefore, *shrinkage* regularization was applied to improve covariance estimation, numerical stability, and classifier generalization according to Eq. (8),

$$\hat{\Sigma}_\gamma = (1 - \gamma)\hat{\Sigma} + \gamma \frac{\operatorname{tr}(\hat{\Sigma})}{p} \mathbf{I}, \tag{8}$$

where $\hat{\Sigma}$ is the sample covariance matrix, p is the dimensionality of the feature space, and $\mathbf{I}$ is the identity matrix. The *shrinkage* coefficient $\gamma \in [0, 1]$ was estimated automatically using the Ledoit-Wolf method.

LDA was selected because it combines low computational complexity with robust performance and has been widely adopted in EEG-based brain-computer interface applications.

H. Real-Time Operation and 3D-Printed Robotic-Hand Actuation

The trained model was used in continuous operation to control the robotic hand. Inference was performed as new EEG data became available from the acquisition system. After an initial 10 s resting baseline, the signals underwent the same preprocessing used during training, including filtering, common average reference (CAR), spatial Laplacian filtering, and decomposition into the mu and beta bands. First-stage classification was performed on 1 s windows updated every 100 ms.

To improve temporal stability, consecutive decisions were subjected to majority voting. In the configuration used for the quantitative evaluation, the five most recent classifications were considered, and the movement state was assumed when at least three of them indicated movement. This procedure reduces the influence of isolated decisions and makes the output more suitable for controlling a physical end effector.

The hand was controlled by an ESP32 microcontroller, with one SG90 servomotor assigned to each finger. Communication between the computer and the microcontroller was established through a serial interface at 115200 *baud*. The servomotors were powered by an external 5 V, 3 A supply, and a 1000 μF electrolytic capacitor was connected in parallel with the power supply to reduce voltage drops during simultaneous finger movements.

Because the first-stage classifier groups motor execution and motor imagery into a single movement class, both conditions can generate the hand-closing command. The second-stage classifier, responsible for distinguishing motor execution from motor imagery, remained restricted to *offline* evaluation and was not integrated into the actuation loop.

I. Event-Based Quantitative Evaluation

To complement the physical demonstration of the system, a quantitative evaluation was performed through sequential replay of an independent recording in `.ASC` format. Data were supplied to the pipeline in blocks corresponding to approximately 100 ms, preserving the temporal order of the recording and reproducing the logic of continuous acquisition.

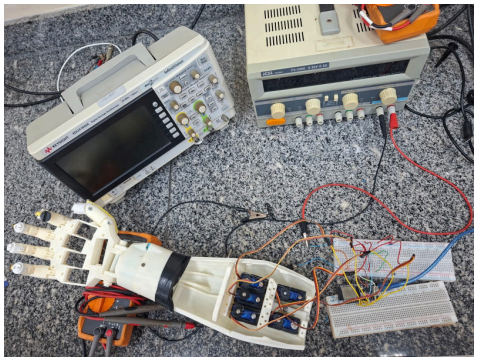


Figure 3. Robotic-hand actuation circuit.

The same preprocessing and the same trained classifier were applied at each update. The first 10 s were used to compute the baseline.

In this evaluation, the FOTO channel was used exclusively as a temporal reference, or *ground truth*, to determine the expected activation and rest periods. The channel was not included in the feature vector and was not provided to the classifier. Thus, the decision continued to be obtained exclusively from EEG signals, whereas the FOTO channel enabled subsequent comparison of the stabilized classifier output with the expected commands.

Each transition to the active LED state was treated as an ON command associated with hand closing, whereas each transition back to the inactive state was treated as an OFF command associated with return to rest and hand opening. For each command type, the total number of events, detected events, and missed events were counted. The detection rate for each condition was calculated as

$$\text{Detection rate} = \frac{N_{\text{detected events}}}{N_{\text{expected events}}} \times 100\%. \tag{9}$$

Episodes of unintended activation outside the expected periods were also counted and reported as isolated false positives. The false-positive rate per minute was calculated relative to the total recording duration,

$$FP/\text{min} = \frac{N_{FP}}{T_{\text{recording}}/60}, \tag{10}$$

where N_{FP} represents the number of false-positive episodes and $T_{\text{recording}}$ is the total duration in seconds. In addition to point estimates, 95% confidence intervals were obtained for the event-detection rates.

Computational latency was measured in each classification cycle as the interval between data-block availability for processing and classifier-output generation. This measure includes preprocessing applied to the available block, feature extraction, standardization, classifier inference, and decision stabilization implemented in the evaluated processing pipeline.

III. Results

A. Offline Classification Performance

The dataset used in the first stage contained 4046 windows, each represented by 66 features. Of these, 2618 windows belonged to the rest class and 1428 to the movement class, corresponding to an approximate class ratio of 1.83:1. For the second stage, 1428 movement windows were used, represented by 24 features after CSP application, including 726 motor-execution windows and 702 motor-imagery windows.

Using stratified five-fold cross-validation, the first classification stage, responsible for distinguishing rest from movement, achieved an accuracy of 91.08%. The second stage, designed to discriminate motor execution from motor imagery, achieved an accuracy of 71.43%. The similar numbers of motor-execution and motor-imagery samples in the second stage reduce the effect of class imbalance in this comparison, whereas the first stage shows a moderate predominance of rest windows.

B. Event-Based Command Detection

In the quantitative evaluation, 30 ON events and 30 OFF events were analyzed, totaling 60 expected state transitions. Among the ON events, 26 were detected and four were missed, resulting in a detection rate of 86.7% and a 95% confidence interval from 74.5% to 98.8%. Among the OFF events, 28 were detected and two were missed, corresponding to a detection rate of 93.3% and a 95% confidence interval from 84.4% to 100.0%.

Considering ON and OFF transitions together, 54 of the 60 expected events were detected, resulting in an overall detection rate of 90.0% and a miss rate of 10.0%. The 95% confidence interval for overall detection ranged from 82.4% to 97.6%. A total of 61 isolated false positives were also recorded throughout the recording, equivalent to 11.6 false positives per minute when normalized by the total recording duration.

Table II
Command-detection performance in the event-based evaluation.

Metric	Value
Expected ON events	30
Detected ON events	26
Missed ON events	4
ON detection rate	86.7%
95% CI for ON detection	[74.5%; 98.8%]
Expected OFF events	30
Detected OFF events	28
Missed OFF events	2
OFF detection rate	93.3%
95% CI for OFF detection	[84.4%; 100.0%]
Overall detection rate	90.0%
Overall miss rate	10.0%
Isolated false positives	61
False positives per minute	11.6

C. Computational Latency

A total of 3045 classifications were performed during sequential replay of the recording used in the quantitative

evaluation. Mean computational latency was 128.5 ms, with a median of 112.9 ms and a standard deviation of 44.3 ms. The minimum recorded latency was 81.5 ms, and the maximum was 766.1 ms. The 95th percentile was 209.8 ms, indicating that 95% of processing cycles were completed within this value, whereas the 99th percentile was 274.1 ms.

Table III
COMPUTATIONAL-LATENCY STATISTICS MEASURED BETWEEN DATA-BLOCK AVAILABILITY AND CLASSIFICATION OUTPUT.

Metric	Value
Number of measurements	3045
Mean	128.5 ms
Median	112.9 ms
Standard deviation	44.3 ms
Minimum	81.5 ms
Maximum	766.1 ms
95th percentile	209.8 ms
99th percentile	274.1 ms

The difference between the mean and median, together with the observed maximum value, indicates the presence of processing cycles with execution times above the typical range. For this reason, percentile values provide complementary information to the mean and are particularly useful for characterizing the temporal stability of the pipeline.

D. Real-Time Operation and Robotic-Hand Actuation

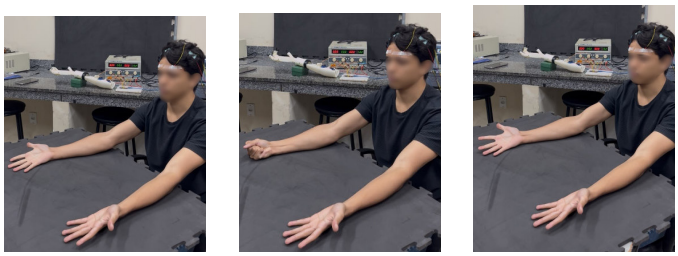
During validation with the physical end effector, the system continuously processed EEG data without using the *trigger* as a decision input. After temporal stabilization of the classifications, commands were transmitted to the ESP32, which actuated the five SG90 servomotors. During rest, the robotic hand remained open. During motor execution, identification of the movement class caused the hand to close. During motor imagery, the hand could also close because the first stage groups execution and imagery into the same movement class.

Fig. 4 presents the three states used in the physical demonstration. This demonstration confirms the integration of acquisition, processing, classification, and actuation. However, the quantitative results of the event-based evaluation show that the system response should not be described as invariably correct because missed commands and false activations were observed. The physical demonstration and the quantitative evaluation should therefore be interpreted as complementary assessments.

IV. DISCUSSION

Table IV compares the proposed approach with recent studies on EEG-based BCIs for end-effector control. The comparison should be interpreted as indicative rather than as a direct equivalence among studies because they differ in participant numbers, tasks, channel counts, classifiers, and validation protocols.

The 91.08% accuracy obtained in the first stage demonstrates that the reduced montage and 66-feature set preserve discriminative capability for separating rest from motor activity. This performance is close to the 91.5% reported by Araujo *et al.* [2] for robotic hand-exoskeleton control. These results reinforce the feasibility of using a reduced number of electrodes for motor-state detection, supporting systems with lower instrumentation complexity.



(a) Rest (b) Motor execution (c) Motor imagery

Figure 4. Robotic-hand response under the three evaluated conditions: (a) rest, with the hand open; (b) motor execution; and (c) motor imagery. Both motor conditions are associated with the hand-closing command when the first stage identifies the movement class.

Table IV
COMPARISON WITH STUDIES ON EEG-BASED BCI SYSTEMS WITH END-EFFECTOR ACTUATION.

Study	Channels	Partic.	Task	Acc.
Araujo *et al.* [2]	16	1	Rest *vs.* MI	91.5 %
An *et al.* [3]	16	4	MI (*offline*)	78 %
An *et al.* [3]	16	4	MI (operation)	62 %
Ding *et al.* [4]	128	21	MI, 2 fingers	80.56 %
Ding *et al.* [4]	128	21	MI, 3 fingers	60.61 %
Yong and Menon [5]	N/A	12	Same-limb MI	66.9 %
This work	**11**	**3**	Rest *vs.* mov.	**91.08 %**
This work	**11**	**3**	ME *vs.* MI	**71.43 %**

The second stage achieved an accuracy of 71.43% in discriminating motor execution from motor imagery, which is consistent with results reported for tasks involving related motor states [5], [6]. This classification is inherently more complex because of the overlap between the cortical networks involved and the similarity of ERD patterns in the mu and beta bands [7]. Even approaches using higher electrode density and deep-learning methods face similar challenges [4]. Thus, the obtained results indicate that the proposed architecture can distinguish these states with meaningful performance, while future optimization may support integration of this stage into real-time physical control.

The event-based quantitative evaluation added an important operational perspective to system performance. Of the 30 ON commands, 26 were detected, corresponding to 86.7%, whereas 28 of the 30 OFF commands were recognized, resulting in 93.3%. Considering both conditions, 54 of the 60 expected events were identified, yielding an overall detection rate of 90.0%. These results show that the system recognized most state transitions during continuous operation, with slightly better performance in identifying the return to rest.

Continuous-behavior analysis also identified activation episodes outside the expected intervals. A total of 61 isolated false positives were recorded, corresponding to 11.6 false positives per minute when normalized by total recording

duration. This metric complements the command-detection rate and contributes to a broader characterization of operational stability. Future work should investigate adjustments to temporal voting and state-transition criteria to further improve system response during continuous use.

Computational latency had a mean of 128.5 ms and a median of 112.9 ms, with a 95th percentile of 209.8 ms and a 99th percentile of 274.1 ms. These values indicate that most classifications were completed within short intervals compatible with the proposed continuous-processing framework. The difference between the mean and median, together with the occasional occurrence of higher processing times, suggests that future computational optimization may help make processing more uniform.

Comparison with An *et al.* [3] reinforces the importance of complementing *offline* results with evaluations that more closely reflect continuous operating conditions. In the present study, sequential replay of the recording enabled quantification of event detection and computational latency, whereas physical actuation demonstrated functional integration with the robotic hand. Together, these evaluations provide a broader characterization of system behavior.

Finally, the reduced electrode montage is a relevant feature of the proposed approach because it lowers acquisition complexity without preventing identification of patterns associated with motor activity. Recent studies indicate that appropriate spatial electrode distribution may be more important than simply increasing channel density [3], [4]. Increasing the number of participants and sessions, together with refinement of event detection and temporal response, represents a natural continuation of this work and may help consolidate the proposed system as an accessible alternative for brain-computer interface and motor-rehabilitation applications.

V. Conclusion

This study presented a methodology for EEG acquisition, preprocessing, classification, and application to the control of an anthropomorphic robotic hand using a reduced 11-channel montage based on the international 10-20 system. The architecture was organized into two decision stages. The first distinguished rest from movement and provided the command used for robotic-hand actuation. The second discriminated motor execution from motor imagery using CSP-based features and remained restricted to *offline* evaluation. The current processing pipeline preserved the same filtering and spatial-processing chain between training and validation, used 1 s windows updated every 0.1 s, and incorporated temporal decision stabilization.

The first stage achieved an accuracy of 91.08%, whereas the second achieved 71.43%. In addition to window-based evaluation, the system was analyzed at the event level to assess its ability to recognize state-change commands. A total of 26 of 30 ON events and 28 of 30 OFF events were detected, corresponding to detection rates of 86.7% and 93.3%, respectively. The overall detection rate was 90.0%, with six missed events among the 60 expected events. The analysis also identified 61 isolated false positives, equivalent to 11.6 false positives per minute of recording, showing that reducing unintended activations remains an important requirement for practical applications.

Temporal analysis provided an additional quantitative characterization of the pipeline. Across 3045 classification cycles, mean computational latency was 128.5 ms, with a median of 112.9 ms, a 95th percentile of 209.8 ms, and a 99th percentile of 274.1 ms. These values correspond to the interval between data-block availability and classification-output generation and should not be interpreted as the complete latency from motor intention to physical hand movement. Validation with the robotic end effector demonstrated functional integration among signal processing, decision making, ESP32 communication, and actuation of the five servomotors.

Overall, the results demonstrate the feasibility of an accessible, low-complexity brain-computer interface for translating cortical motor activity into robotic commands while objectively identifying the main aspects that require improvement. Future work should increase the number of participants and sessions, use event-grouped or session-grouped validation, reduce the frequency of false positives, optimize computational cost, measure end-to-end latency from intention to physical actuation, and improve the second classification stage to enable its integration into real-time control and support more specific commands.

References

[1] G. Pfurtscheller, "Functional brain imaging based on ERD/ERS," *Vision Research*, vol. 41, no. 10–11, pp. 1257–1260, 2001.

[2] R. S. Araujo, C. R. Silva, S. P. N. Netto, E. Morya, and F. L. Brasil, "Development of a low-cost EEG-controlled hand exoskeleton 3D printed on textiles," *Frontiers in Neuroscience*, vol. 15, p. 661569, 2021.

[3] Y. An, D. Mitchell, J. Lathrop, D. Flynn, and S.-J. Chung, "Motor imagery teleoperation of a mobile robot using a low-cost brain-computer interface for multi-day validation," 2024, arXiv:2412.08971.

[4] Y. Ding, C. Udompanyawit, Y. Zhang, and B. He, "EEG-based brain-computer interface enables real-time robotic hand control at individual finger level," *Nature Communications*, vol. 16, p. 5401, 2025.

[5] X. Yong and C. Menon, "EEG classification of different imaginary movements within the same limb," *PLOS ONE*, vol. 10, no. 4, p. e0121896, 2015.

[6] K. Jayalaksshme Srinivasan, P. Periasamy, and S. Gunasekaran, "Motor imagery and motor execution: A narrative review of electroencephalographic (EEG) signatures, methodological consistency, and translational applications," *Cureus*, vol. 17, no. 9, p. e93011, 2025.

[7] K. Xu, Y.-Y. Huang, and J.-R. Duann, "The sensitivity of single-trial mu-suppression detection for motor imagery performance as compared to motor execution and motor observation performance," *Frontiers in Human Neuroscience*, vol. 13, p. 302, 2019.

Efeito do Exercício Físico sobre a Preservação Neuronal em Fatias de Hipocampo de Rato Submetidas a *Status Epilepticus in vitro*

Andre Luis Fonseca Furtado
Programa de Pós-graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
andre.furtado@ifsudestemg.edu.br

Antônio-Carlos Guimarães de Almeida
Laboratório de Neurociência Experimental e Computacional
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
acga@ufsj.edu.br

Antônio Márcio Rodrigues
Laboratório de Neurociência Experimental e Computacional
Universidade Federal de São João del-Rei
São João del-Rei, Brazil
amr@ufsj.edu.br

Carla Alessandra Scorza
Departmento de Neurologia e Neurocirurgia
Universidade Federal de São Paulo/Escola Paulista de Medicina
São Paulo, Brasil
carla.scorza@unifesp.br

Fabricio Neves Mendonça
Programa de Pós-graduação em Neurologia Neurociências
Universidade Federal de São Paulo/Escola Paulista de Medicina
São Paulo, Brasil
fabricio.mendonca@ifmg.edu.br

Fulvio Alexandre Scorza
Departmento de Neurologia e Neurocirurgia
Universidade Federal de São Paulo/Escola Paulista de Medicina
São Paulo, Brasil
scorza.nexp@epm.br

Resumo*—A utilização de fatia hipocampal para simular os danos causados no tecido nervoso pelo estado de crise contínua, através da indução das atividades epileptiformes não sinápticas (AENS) in vitro, encontra similaridade com a injúria provocada pelo status epilepticus (SE) em modelos experimentais in vivo. Vários estudos apontam que o SE in vivo pode causar danos neuronais em regiões do sistema límbico, tanto em modelos experimentais quanto em humanos. A utilização do protocolo de zero $Ca2^{+}$ e alto K^{+} para induzir atividades epileptiformes (AE) se constitui, por excelência, em um modelo de atividades ictais que reproduzem o estado tônico-clônico das crises epilépticas. Trata-se, portanto, de um modelo de SE devido ao fato de apresentar recorrência dos eventos ictais em períodos com duração superior a 30 minutos e com potencial de gerar danos celulares similares aos observados nas condições in vivo. Neste trabalho foi avaliada a influência do exercício físico, praticado por ratos Wistar durante o desenvolvimento cerebral, sobre a capacidade de preservação neuronal diante dos danos provocados pela indução de AENS nas fatias hipocampais. Os resultados encontrados neste estudo demonstram que o exercício físico durante o desenvolvimento cerebral proporcionou maior preservação do tecido neuronal da camada granular (CG) do giro dentado (GD) do hipocampo.

***Palavras-chave*—epilepsia; exercício físico; histologia**

I. INTRODUÇÃO

O estado de mal epiléptico ou SE é caracterizado por uma convulsão continua ou convulsões múltiplas com duração acima de 30 minutos, sem recuperação da consciência nos intervalos entre as crises [1]. O SE é uma condição de emergência médica, pois o prolongamento das crises convulsivas pode levar o paciente a sérios prejuízos neurológicos que, além de aumentar o risco de desenvolver epilepsia, pode também levá-lo a óbito [2]. Há uma incidência de SE em torno de 20 para cada 100.000 pessoas e a taxa de mortalidade varia entre 3% a 40% dependendo da etiologia, idade, tipo e duração do SE [3]. Fortes evidências apontam que o SE pode causar danos neuronais principalmente em regiões do sistema límbico, tanto em pacientes humanos quanto em modelos experimentais [4, 5].

A utilização de modelos animais para simular o estado de crise contínua, seja *in vivo* ou *in vitro*, envolve manipulações experimentais nas quais a condição de sustentação prolongada do hipersincronismo resulta em danos neuronais similares aos observados como consequência de um episódio agudo de SE. De acordo com [6] em modelos experimentais *in vivo* de SE, a morte neuronal na CG do GD está associada ao insulto neurológico inicial e não com as crises espontâneas ocorridas após o processo de epileptogênese que segue ao SE.

Nesse sentido, um estudo desenvolvido por [7] constatou a ocorrência de morte neuronal por necrose em diversas regiões do sistema límbico de ratos adultos submetidos a SE induzido por ácido caínico. No referido estudo, os neurônios necróticos apresentaram formação de núcleos picnóticos, condensação de cromatina e DNA em escada [7]. O processo de morte celular induzida por SE é iniciado pela liberação pré-sináptica de glutamato, que é responsável por ativar os receptores NMDA (N-metil-D-aspartato) dos neurônios pós-sinápticos, desencadeando assim um excessivo influxo de Na^{+} e Ca^{2+} e caracterizando um quadro de excitotoxicidade. Essa dinâmica resulta na ativação intracelular de proteases e síntese de óxido nítrico, com geração de radicais livres. Em consequência, observam-se danos nas membranas celulares, na estrutura de proteínas e enzimas essenciais, culminando em necrose celular [8].

Um importante aspecto a ser considerado sobre as condições experimentais do SE *in vitro* é a recorrência de eventos que se assemelham às crises *in vivo*. A utilização de fatias de cérebro são as mais convenientes preparações para o estudo *in vitro*, uma vez que a preservação dos circuitos neuronais inerentes à fatia a torna adequada ao estudo das AE sob condições de controle da hiperexcitabilidade. Um importante aspecto a ser considerado sobre as condições experimentais *in vitro* é a ocorrência cíclica de eventos que se assemelham às crises [9]. Fatias hipocampais perfundidas com fluido cerebroespinhal artificial com alta concentração de potássio podem deflagrar e sustentar por horas e de forma cíclica, disparos de potenciais tipo tônico-clônicos [10]. Esses eventos cíclicos e autorregenerativos, se repetidos por tempo suficiente, tem o potencial de reproduzir o risco de lesão

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

neuronal causado pelo SE *in vivo* [11], fazendo com que a indução de AE persistente no hipocampo se caracterize, de fato, como um modelo *in vitro* de SE. Nesse sentido, desde 1982, manobras experimentais utilizando fatias de hipocampo têm mostrado que AE podem ser geradas na ausência de Ca^{2+} [12, 13]. As AENS registradas no GD, quando induzidas por meio do protocolo de alto K^+ e zero-Ca^{2+}, são caracterizadas por *bursts* espontâneos apresentando período ictal prolongado, variação negativa e lenta da linha de base do potencial extracelular (PE) e a ocorrência de *population spikes* (PS) com grandes amplitudes [14]. O modelo de alto K^+ e zero-Ca^{2+}, utilizando fatias de hipocampo é, por excelência, um modelo de atividades ictais, portanto associadas ao chamado período tônico-clônico das crises de epilepsia. A indução deste tipo de atividade neuronal em fatias de hipocampo se apresenta como uma ferramenta muito importante para o entendimento das possíveis influências promovidas pelo exercício físico.

Investigações têm demonstrado que o exercício físico praticado durante o período de desenvolvimento é capaz de promover plasticidade no hipocampo [15]. Outros estudos mostram que o exercício físico pode ser capaz de proteger o hipocampo contra danos advindos do estresse oxidativo, reduzindo a ocorrência de morte celular [16, 17]. As possíveis mudanças promovidas pela prática regular de exercício físico parecem também estar relacionadas com maior preservação do tecido nervoso. Em pesquisa realizada em nosso grupo, foi demonstrado que a prática de exercício físico favoreceu neurogênese em ratos submetidos a SE, numa situação em que a perda neuronal é sabidamente acentuada [18].

Contudo, o conhecimento dos mecanismos celulares e moleculares pelos quais o exercício físico é capaz de produzir ainda são muito limitados. Tendo isso em vista e, considerando-se a abordagem translacional para a condição humana, uma investigação sobre os benefícios da prática regular de exercício físico frente a um episódio agudo de SE, pode contribuir para o entendimento de como o cérebro de um indivíduo fisicamente ativo pode reagir a um insulto com condições de levá-lo a desenvolver uma desordem neurológica, tal como a epilepsia.

Um estudo recente demonstra que a realização de um protocolo de exercício físico exerce forte influência na frequência e intensidade dos sintomas motores induzidos no modelo experimental *in vivo* [19]. Porém, poucos são os trabalhos que relacionam o efeito que o exercício físico exerce na capacidade do cérebro de resistir aos danos provocados por um episódio agudo de SE [19, 20]. O conhecimento dos mecanismos celulares e moleculares pelos quais o exercício físico é capaz de produzir neuroproteção ainda são muito limitados [21, 22]. No presente trabalho, a indução prolongada de AENS foi utilizada como modelo *in vitro* de SE com o objetivo de verificar a capacidade de preservação da CG do GD em razão do exercício físico durante o desenvolvimento cerebral contra os danos observados nas fatias de hipocampo submetidas ao SE não sináptico *in vitro*.

II. Métodos

A. Animais

O estudo seguiu o protocolo aprovado pela Comissão de Ética em Pesquisa Envolvendo Animais da UFSJ – protocolo 08/2017 (CEUA-UFSJ). Ratos machos da linhagem *Wistar* foram mantidos em condições controladas de luz, temperatura e umidade, com acesso livre à água e comida. Os animais foram submetidos a um protocolo de exercício físico em esteira por 40 dias, iniciando aos 21 dias de idade. Dois grupos foram estabelecidos: o Grupo Controle (Ctrl), que não recebeu exercício, e o Grupo Exercício (Exc), que foi submetido ao programa de exercício físico.

B. Programa de exercício físico

O protocolo de exercício físico seguiu o modelo proposto por [15] consistindo em duas fases: adaptação à esteira e programa de exercício físico. Ambos os grupos, Ctrl e Exc, passaram por familiarização com a esteira por três dias: P21 a P23, com critérios de treinabilidade para inclusão. O programa de exercícios ocorreu dos dias P24 a P60, aumentando gradualmente em duração e intensidade para minimizar o estresse. O grupo Ctrl foi submetido às mesmas condições ambientais que o grupo Exc durante o experimento.

C. Eletrofisiologia - SE in vitro

Os animais foram eutanasiados por decapitação em guilhotina. O hipocampo foi isolado e seccionado em fatias de 400 µm de espessura. Após 40 minutos em uma câmara de perfusão, as fatias foram transferidas para uma câmara de interface para indução das AENS. Os registros do PE foram realizados com eletrodos com fio de prata cloretada, inseridos em pipetas de boro-silicato, preenchidas por 2M de NaCl (5-10 MΩ) e inseridas na CG do GD. O eletrodo foi conectado a uma *head-stage* (Modelo AI 402x50, ULTRALOW NOISE AMPLIFIER – AXON INSTRUMENTS) para aquisição do sinal. Os registros eletrofisiológicos foram realizados no GD em fatias de hipocampo inicialmente mantidas em solução de *Ringer*, com a seguinte composição: NaCl 127mM, KCl 2mM, $MgSO_4$ 1.5 mM, $NaHCO_3$ 26 mM, KH_2PO_4 1.1 mM, glicose 10 mM e $CaCl_2$ 2mM. A indução das AE foi promovida por meio de solução de *Ringer* modificada, com alta concentração de K^+ (8.1 mM) e sem adição de Ca^{2+}, solução capaz de deflagrar AE no GD de fatias de hipocampo de ratos. As soluções foram mantidas sob constante borbulhamento de mistura carbogênica (95% de O_2/ 5% de CO_2) para ajuste do pH em torno de 7.4. Após a indução das AENS o registro era iniciado. A figura 1 ilustra imagens de lesão em tecido neuronal causadas pelo SE *in vivo* e pelo SE *in vitro*.

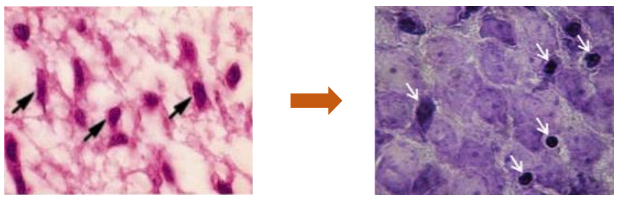


Figura 1 – As duas imagens mostram a formação de núcleos picnóticos (apontadas pelas setas) como resultado da morte celular causada pelo SE tanto no modelo *in vivo* como no modelo *in vitro*. Adaptado de [23].

D. Método de Nissl para análise histológica

Após os registros eletrofisiológicos, as fatias hipocampais utilizadas nos registros (uma fatia por animal) foram coletadas com o intuito de avaliar o perfil histológico de morte celular provocada pelo SE não sináptico *in vitro*. Para isso, as fatias foram acondicionadas em tubos de *Eppendorf* e fixadas em solução de paraformol 4%, temperatura em 4ºC.

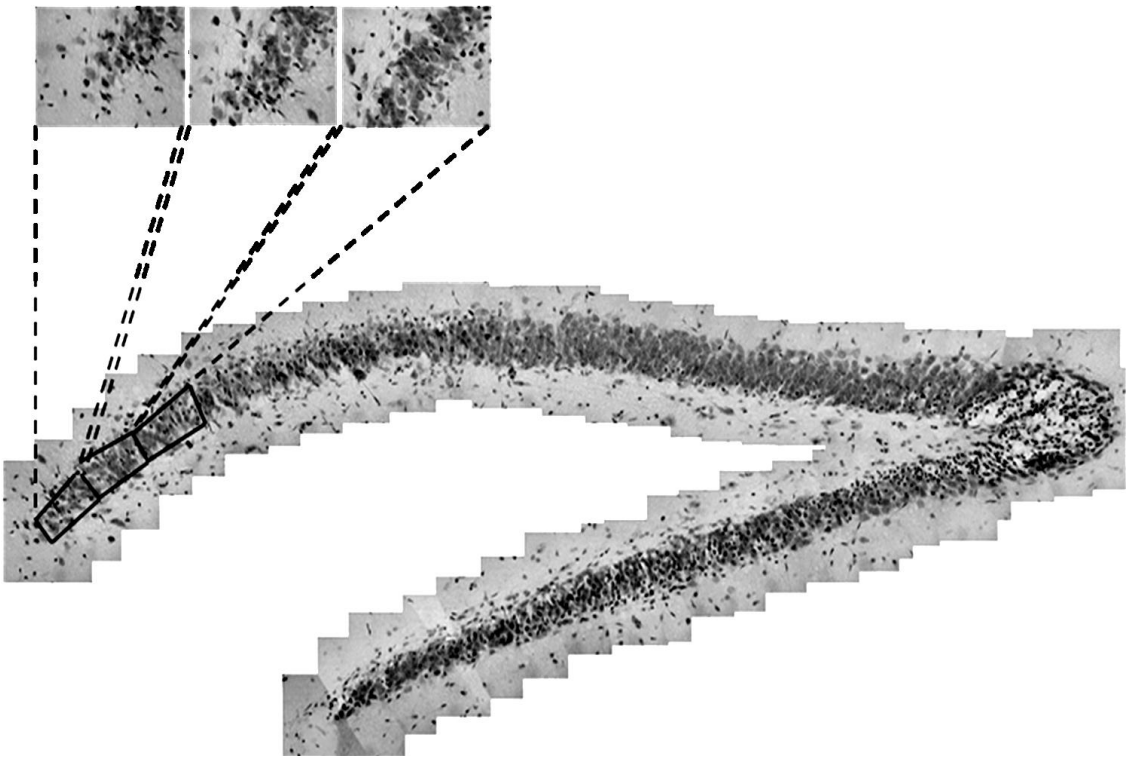


Figura 2 - Marcação de *Nissl* com trechos da camada granular do giro dentado capturados por câmera digital e abaixo a imagem, reconstruída e com a ampliação de 400x utilizada para a quantificação da área preservada.

Após período de duas semanas, as fatias de 400 µm foram resseccionadas em fatias de 15 µm, por meio de um vibrátomo (*Leica* VT 1000S). Para o procedimento de ressecção em fatias de 15 µm foi utilizado um emblocamento em ágar (1,25g de Agar-Agar diluídos em 25 ml de água destilada), com fixação no porta blocos do vibrátomo e imersão em solução salina (9g de NaCl são diluídos em 1000 ml de água destilada) resfriada.

As fatias resseccionadas (6 fatias de 15 µm por animal) foram, então, posicionadas sobre duas lâminas de vidro previamente gelatinizadas (3 fatias por lâmina). Após a montagem das lâminas, estas permaneceram na capela de exaustão (*Sppencer Scientific Equipament*) para secagem e posterior coloração. A figura 2 mostra o resultado da coloração de uma fatia, onde podem ser observados os núcleos celulares.

O método de *Nissl* foi aplicado imergindo-se as lâminas por 5 minutos em água destilada (hidratação) e 35 minutos em violeta de cresila (*Synth*) para coloração. O processo de desidratação consistiu na imersão das lâminas por 2 vezes em álcool 70%, 7 vezes em álcool 95%, 8 vezes em álcool 100%. O processo de diafanização foi concluído imergindo-se as lâminas em solução xilol 1 (1 minuto) e em solução xilol 2 (5 minutos).

E. Quantificação da área preservada

Para cada animal, foi utilizada uma única fatia de 400 µm (uma fatia por animal), da qual foram obtidas seis seções de 15 µm, fixadas em paraformol (4%) e coradas pela técnica de *Nissl.* O valor de área preservada de cada animal correspondeu à média aritmética dos percentuais de área preservada das seis seções analisadas, de modo que cada animal contribuiu com uma única observação independente na análise estatística (N=10 por grupo). Após a fixação das lamínulas com *Entellan* (*Merck*), as lâminas foram levadas ao microscópio (*Axio Scope*.A1 – *Zeiss*), que, acoplado a uma câmera digital (*PowerShot* A650IS, *Canon*), permitiu a captura de partes das imagens da CG com capacidade de ampliação de 400x.

As imagens de toda a área da CG do GD de cada fatia de 15 µm foram reconstruídas utilizando-se um Sistema convencional (*Adobe® Photoshop* CS2, versão 9.0) que integra cada uma das partes capturadas e permite obter uma imagem única de toda a camada, como mostrado na figura 2. Para a realização do cálculo da área preservada das fatias, foi utilizado o método de delimitação das regiões da CG com o auxílio de um Sistema computacional MATLAB®.

Cada imagem montada foi quantificada traçando-se uma poligonal ao redor do GD do hipocampo, conforme a figura 3. O sistema computacional utilizado permitiu estimar a área delimitada pela poligonal, em número de *pixels*. Para fins de normalização, o percentual da área preservada foi calculado em relação à área total da CG do GD de cada fatia. Ela corresponde à extensão de tecido que permaneceu íntegro após a indução do SE não sináptico *in vitro*, de modo que valores mais altos indicam menor dano à camada.

F. Análise estatística

Após a aplicação do teste de normalidade de *Shapiro-Wilk* para verificação de distribuição normal dos dados, procedeu-se à comparação da área preservada total da CG do GD entre os dois grupos experimentais, por meio do teste t de Student não pareado para amostras independentes, sendo o animal a

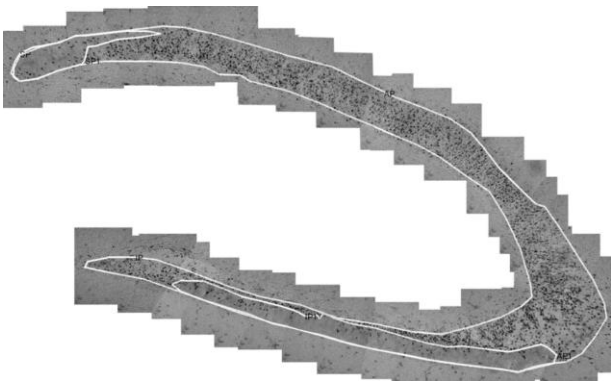


Figura 3 – Representação das poligonais traçadas ao redor das áreas preservadas da CG do GD.

unidade experimental (N=10 por grupo). A análise envolveu uma única variável de desfecho, não havendo necessidade de ajuste para comparações múltiplas. O nível de significância de probabilidade de erro utilizado foi de 5%.

III. Resultados

A. *Validação do protocolo de lesão por SE não sináptico in vitro*

O padrão de lesão observado nas células das fatias de hipocampo submetidas às AENS *in vitro* encontra similaridade ao padrão de lesão observado no SE *in vivo*, conforme mostra a figura 1.

B. *Quantificação celular pela técnica de Nissl*

Utilizando as técnicas de histologia descritas, nosso trabalho investigou a hipótese de que o exercício físico exerce influência benéfica na preservação das células da CG do GD do hipocampo de ratos de fatias submetidas a SE não sináptico *in vitro*. Assim, os cortes histológicos dos hipocampos dos animais dos grupos Ctrl e Exc foram analisados e os valores de área preservada por fatia analisada foram obtidos.

A análise estatística do parâmetro área preservada demonstrou um valor significativamente maior no grupo Exc, quando comparado ao grupo Ctrl (Ctrl: 29.10 ± 13.40 média % área preservada N=10; Exc: 44.04 ± 14.32 % área preservada N=10; t= 2.335; df= 18; p= 0.0313). A diferença média entre os grupos foi de 14,94 pontos percentuais, com IC 95% de 1,50 a 28,38 pontos percentuais.

IV. Discussão

Mesmo com o tempo médio total de duração do SE não sináptico *in vitro* no grupo Exc sendo maior do que o observado no grupo Ctrl medido em [24], a análise estatística apontou para alterações significativas induzidas pelo exercício físico, com o grupo Exc apresentando um perfil de maior preservação do tecido quando comparado com as fatias dos animais do grupo Ctrl (figura 4). Essa evidência mostra que o treinamento na esteira foi capaz de promover neuroproteção, mitigando os danos provocados pelo SE não sináptico *in vitro*. Essa hipótese, defendida na presente investigação, se fundamenta no argumento de que o exercício físico pode exercer influência benéfica simultânea sobre a demanda e o metabolismo energético durante as AENS. Conforme estudo prévio do nosso grupo de pesquisa [24], apesar do maior tempo total de sustentação do SE *in vitro*, a redução na intensidade e duração dos eventos epileptiformes no grupo Exc, tende a favorecer a redução da demanda energética para o reestabelecimento dos gradientes iônicos da célula neuronal. Por outro lado, uma maior demanda das reações oxidativas para liberação de energia, pode favorecer o aumento na produção de espécies reativas ao oxigênio que tem o potencial de degradar a estrutura da Na^+/K^+ ATPase, e comprometendo sua funcionalidade, agravando o quadro de excitabilidade e morte neuronal [25, 26]. Tais mecanismos não foram medidos aqui diretamente, sendo uma hipótese sustentada por evidências apontadas pela literatura científica [24, 25, 26]. Portanto, a potencial redução da demanda

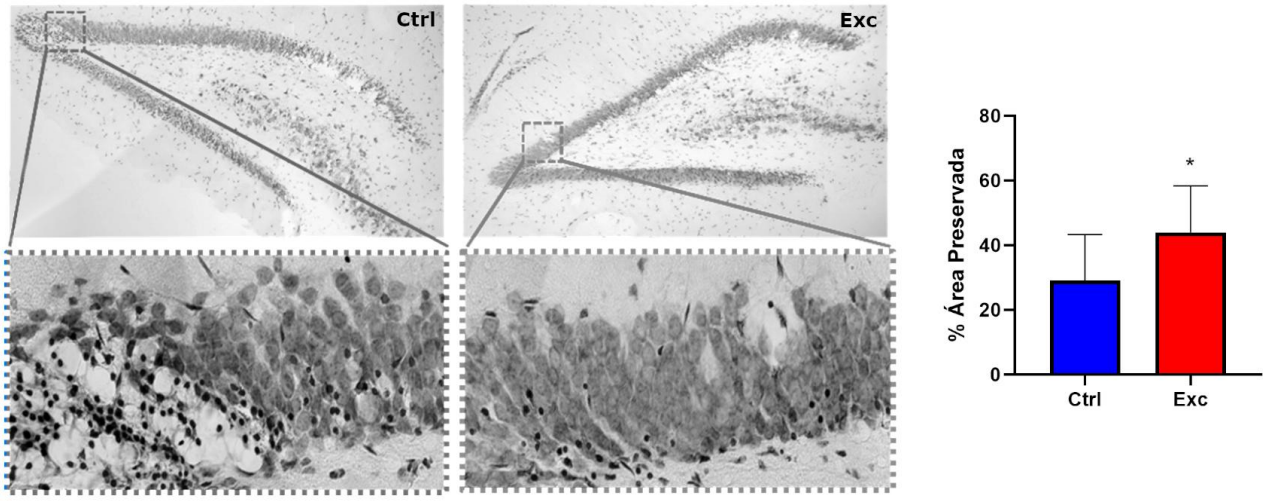


Figura 4 – Aspectos morfológicos da CG do GD dos grupos Ctrl e Exc (coloração pelo método de *Nissl*). As imagens superiores possuem ampliação de 50 vezes e as inferiores possuem ampliação de 400 vezes, destacando células edemaciadas e núcleos picnóticos na região de registro. O asterisco, quando indicado, representa diferença significativa ao nível de 5% pelo teste T não pareado. Os valores dos gráficos estão expressos no formato de media ± erro padrão. Número de amostras por grupo (Ctrl = 10; Exc = 10).

energética requerida nas AENS do grupo Exc durante toda a duração do SE não sináptico *in vitro*, pode ter favorecido uma menor produção de radicais livres nas reações envolvidas na produção de ATP pelas células da CG do GD, concedendo uma maior resiliência ao tecido neuronal quando comparado ao grupo Ctrl (figura 5). Tal condição pode contribuir para a manutenção da funcionalidade de proteínas essenciais, como a Na^+/K^+ ATPase, favorecendo a manutenção da homeostase iônica, aumentando a resiliência do tecido neuronal frente a estados de hiperexcitabilidade sustentada e configurando-se como um potencial mecanismo de neuroproteção em condições experimentais de epilepsia. Nesse sentido, estudos relacionados à formação de reserva neural por meio da atividade física em idade precoce corroboram nossos achados [21]. Segundo [21], essa reserva neural seria capaz de proteger o cérebro contra danos decorrentes do estado sustentado de hiperexcitabilidade do tecido nervoso, como observado durante as crises convulsivas e as AENS.

V. Conclusão

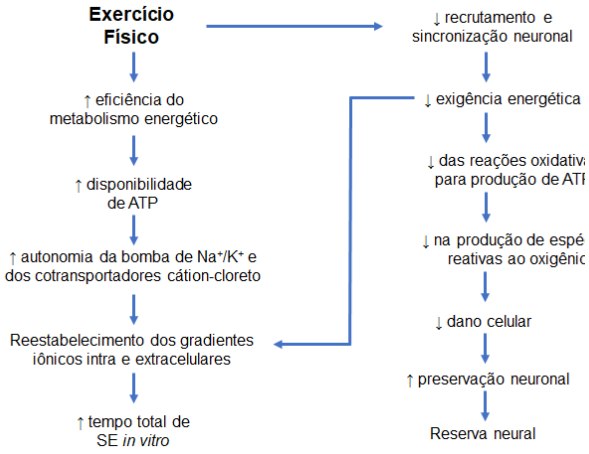


Figura 5 – Representação esquemática da influência do exercício físico na preservação da CG do GD hippocampal.

Estudos posteriores com marcadores complementares de viabilidade celular e correlação entre carga eletrofisiológica e preservação tecidual podem esclarecer de forma mais aprofundada os achados aqui descritos. Contudo, mesmo considerando exclusivamente os dados obtidos por meio da coloração de *Nissl* e o tempo de duração diferente do SE não sináptico in vitro entre os grupos, tomando em conjunto a preservação da camada granular associada ao exercício físico observada em nossos resultados e os achados eletrofisiológicos previamente descritos em [24], observa-se que o exercício físico exerce efeito modulador sobre a dinâmica das AENS e seus impactos sobre a integridade do tecido neuronal no GD, promovendo neuroproteção na região de ocorrência do SE não sináptico *in vitro*.

Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), ao Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG), ao Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais (IF Sudeste MG), à Universidade Federal de São João del Rei (UFSJ) e à Universidade Federal de São Paulo (UNIFESP) pelo apoio financeiro e institucional.

Referências

[1] D. S. Reddy and R. Kuruba, "Experimental Models of Status Epilepticus and Neuronal Injury for Evaluation of Therapeutic Interventions," *Int. J. Mol. Sci.*, vol. 14, no. 9, pp. 18284–18318, 2013.

[2] S. Knake, H. M. Hamer, and F. Rosenow, "Status epilepticus: A critical review," *Epilepsy and Behavior*, vol. 15, no. 1, pp. 10–14, 2009.

[3] A. Coeytaux, P. Jallon, B. Galobardes, and A. Morabia, "Incidence of status epilepticus in French-speaking Switzerland: (EPISTAR)," *Neurology*, vol. 55, no. 5, pp. 693–697, 2000.

[4] Ebert U, Brandt C, Löscher W. (2002). Delayed sclerosis, neuroprotection, and limbic epileptogenesis after status epilepticus in the rat. Epilepsia, 43(Suppl 5):86-95. DOI 10.1046/j.1528-1157.43.s.5.39.x.

[5] G. Giovannini and S. Meletti, "Fluid biomarkers of neuro-glial injury in human status epilepticus: a systematic review," Int. J. Mol. Sci., vol. 24, no. 15, art. 12519, 2023.

[6] L. S. Deshpande and R. J. DeLorenzo, "Basic Mechanisms of Neuronal Injury from Seizures and Status Epilepticus," 2009.

[7] D. G. Fujikawa, S. S. Shinmei, and B. Cai, "Kainic acid-induced seizures produce necrotic, not apoptotic, neurons with internucleosomal DNA cleavage: implications for programmed cell death mechanisms," *Neuroscience*, vol. 98, no. 1, pp. 41–53, 2000.

[8] D. G. Fujikawa, "Prolonged seizures and cellular injury: Understanding the connection," *Epilepsy & Behavior*, vol. 7, no. Suppl 3, pp. 3–11, 2005.

[9] L. L. Antonio, M. L. Anderson, E. A. Angamo, S. Gabriel, Z. J. Klaft, A. Liotta, S. Salar, N. Sandow, and U. Heinemann, "In vitro seizure like events and changes in ionic concentration," *J. Neurosci. Methods*, vol. 260, pp. 33–44, 2016.

[10] S. F. Traynelis and R. Dingledine, "Potassium-induced spontaneous electrographic seizures in the rat hippocampal slice," *J. Neurophysiol.*, vol. 59, no. 1, pp. 259–276, Jan. 1988.

[11] U. Heinemann and K. J. Staley, "What is the clinical relevance of in vitro epileptiform activity?" *Adv. Exp. Med. Biol.*, vol. 813, pp. 25–41, 2014.

[12] J. G. Jefferys and H. L. Haas, "Synchronized bursting of CA1 hippocampal pyramidal cells in the absence of synaptic transmission," *Nature*, vol. 300, pp. 448–450, 1982.

[13] C. P. Taylor and F. E. Dudek, "Synchronous neural afterdischarges in rat hippocampal slices without active chemical synapses," *Science*, vol. 218, no. 4574, pp. 810–812, 1982.

[14] A. C. G. Almeida, A. M. Rodrigues, and M. A. Duarte, "Biophysical Aspects of the Nonsynaptic Epileptiform Activity," in *Underlying Mechanisms of Epilepsy*, F. S. Kaneez, Ed., 2011, pp. 189–218.

[15] S. Gomes da Silva, F. Doná, M. J. S. Fernandes, F. A. Scorza, E. A. Cavalheiro, and R. M. Arida, "Physical exercise during the adolescent period life increases hippocampal parvalbumin expression," *Brain & Development*, vol. 32, no. 2, pp. 137–142, 2009.

[16] J. I. Reiss, R. K. Dishman, H. E. Boyd, J. K. Robinson, and P. V. Holmes, "Chronic activity wheel running reduces the severity of kainic acid-induced seizures in the rat: Possible role of galanin," *Brain Res.*, vol. 1266, pp. 53–60, 2009.

[17] S. Vaynman, F. Gomez-Pinilla. "License to Run: Exercise Impacts Functional." *Neurorehabil Neural Repair*, v. 19, n. 4, p. 283-295. 2005.

[18] F. N. Mendonça, L. E. C. Santos, and A. M. Rodrigues, "Physical Exercise Restores the Generation of Newborn Neurons in an Animal Model of Chronic Epilepsy," *Front. Neurosci.*, vol. 11, art. 98, pp. 1–9, 2017.

[19] M. Shishmanova-Doseva, K. Georgieva, Y. Uzunova, L. Ioanidu, M. Atanasova, Z. Nenchovska, and J. Tchekalarova, "Pre- and post-endurance training mitigates the rat pilocarpine-induced status epilepticus and epileptogenesis-associated deleterious consequences," Int. J. Mol. Sci., vol. 23, no. 21, art. 13188, 2022.

[20] H. Yu, M. Shao, X. Luo, C. Pang, K.-F. So, J. Yu, and L. Zhang, "Treadmill exercise improves hippocampal neural plasticity and relieves cognitive deficits in a mouse model of epilepsy," Front. Neurosci., vol. 17, art. 1254680, pp. 1–12, 2023.

[21] R. M. Arida, A. C. G. Almeida, E. A. Cavalheiro, and F. A. Scorza, "Experimental and clinical findings from physical exercise as complementary therapy for epilepsy," *Epilepsy & Behavior*, vol. 26, no. 3, pp. 273–278, 2013.

[22] F. Yousaf, S. Kao, S. Ishaq, and S.-D. Lee, "Molecular regulation of exercise training on hippocampal neuroprotection in post-traumatic brain injury: A systematic review," Exp. Neurol., vol. 394, pp. 1–12, 2025.

[23] D. G. Fujikawa, S. S. Shinmei, B. CAI. "Lithium-pilocarpine-induced status epilepticus produces necrotic neurons with internucleosomal DNA fragmentation in adult rats". *European Journal of Neuroscience*, v. 11, n. 5, p. 1605-1614, 1999.

[24] A. L. F. Furtado, F. N. Mendonça, A. M. Rodrigues, F. A. Scorza, C. A. Scorza, and A. C. G. Almeida, "Exercício Físico e Atividades Epileptiformes em Ratos Wistar: Um Estudo Eletrofisiológico," in Anais do XXIX Congresso Brasileiro de Engenharia Biomédica, Ribeirão Preto: SBEB, 2024, vol. 1, pp. 84–89.

[25] M. A. Souza, M. S. Oliveira, A. F. Furlan, L. M. Rambo, L. R. Ribeiro, F. D. Lima, L. C. Dallacorte, L. F. Silva, L. T. Retamoso, C. L. Dalla Corte, G. O. Puntel, D. S. de Ávila, F. A. Soares, M. R. Fighera, C. F. de Mello, and L. F. Royes, "Swimming training prevents pentylenetetrazol-induced inhibition of Na+, K+-ATPase activity, seizures, and oxidative stress," *Epilepsia*, vol. 50, no. 4, pp. 811–823, 2009.

[26] L. F. A. Silva, M. S. Hoffmann, R. R. Gerbatin, F. S. Fiorin, F. Dobrachinski, B. C. Mota, A. T. B. Wouters, S. P. Pavarini, F. A. A. Soares, M. R. Fighera, and L. F. F. Royes, "Treadmill Exercise Protects Against Pentylenetetrazol-Induced Seizures and Oxidative Stress after Traumatic Brain Injury," *J. Neurotrauma*, vol. 30, no. 14, pp. 1278–1287, 2013.

Efeitos da estimulação vibrotátil na regulação de forças máximas e submáximas sem realimentação visual

Giovanna Ogata Pedrassani
Departamento de Eletrônica e Engenharia Biomédica
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brasil
g221085@dac.unicamp.br

Willian A. Guedes da Silva
Departamento de Eletrônica e Engenharia Biomédica
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brasil
w298991@dac.unicamp.br

José O. Alves Filho
Departamento de Eletrônica e Engenharia Biomédica
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brasil
j290880@dac.unicamp.br

Carina Marconi Germer
Departamento de Eletrônica e Engenharia Biomédica
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brasil
cgermer@unicamp.br

***Resumo*—Este artigo investiga os efeitos da estimulação vibrotátil na regulação de forças isométricas máximas e submáximas do dedo indicador. O estudo contou com onze voluntários saudáveis que realizaram tarefas de contração voluntária máxima (MVC) e contrações isométricas com fases de rampa e platô a 30% e 45% da MVC, nas condições com e sem estimulação vibrotátil. Nas tarefas submáximas, a realimentação visual da força produzida foi retirada após os 2 segundos iniciais da rampa, enquanto a força-alvo permaneceu visível. Na amostra analisada até o momento, não foi detectada diferença significativa na força máxima entre as condições. Por outro lado, para a taxa de incremento da força, foi identificado efeito significativo da estimulação vibrotátil durante a fase de rampa. Já durante o platô, não foram observados efeitos significativos da estimulação sobre a força média ou o coeficiente de variação. Esses resultados fornecem evidências preliminares de que, nas condições avaliadas, a estimulação vibrotátil pode afetar de maneira específica a fase dinâmica de incremento da força, sem alterações significativas nas variáveis analisadas durante a fase de contração isométrica.**

***Palavras-chave*—estimulação vibrotátil, controle motor, variabilidade da força**

I. Introdução

A capacidade de modular e regular a força muscular é essencial para a realização de atividades motoras em diferentes contextos cotidianos. Essa regulação depende de uma complexa integração entre os sistemas sensoriais e motores, sendo continuamente influenciada por mecanismos neurais centrais e periféricos. Como parte desse processo, a atividade do núcleo motor é modulada por diferentes tipos de entradas sensoriais, incluindo estímulos táteis, proprioceptivos e visuais.

Nesse contexto, a informação tátil desempenha um papel essencial na orientação de movimentos precisos. Em especial, a estimulação vibrotátil aplicada à pele ativa receptores cutâneos de adaptação rápida, cujos sinais influenciam diretamente a informação trafegada nos circuitos neurais da medula espinhal, do tronco encefálico e de estruturas corticais. Dessa forma, diversas evidências sugerem que esse tipo de estimulação favorece o recrutamento de unidades motoras [1]–[3], aumenta a excitabilidade motora [4] e promove um ganho de força em contrações isométricas [5], além de atuar na redução da variabilidade da força em contrações de baixa intensidade [6]–[8].

Além da estimulação tátil, a realimentação visual é um fator determinante na modulação do controle motor, pois permite correções e ajustes em tempo real do movimento. Quando esse feedback está ausente, o sistema nervoso enfrenta um desafio maior, o que geralmente altera a variabilidade e aumenta o erro na força produzida [9], [10]. Isso, por sua vez, pode interferir na interpretação dos efeitos de intervenções sensoriais, como a tátil. Diante disso, investigar os efeitos da estimulação vibrotátil, especificamente na ausência de realimentação visual, é fundamental para compreender seu efeito sobre desempenho motor.

Dessa forma, este artigo propõe investigar os efeitos da estimulação vibrotátil na capacidade de geração de força máxima e na regulação de intensidades submáximas em tarefas de contração isométrica, concentrando-se especificamente no comportamento motor sob a ausência de realimentação visual. Com isso, pretende-se contribuir para o entendimento dos mecanismos sensoriais e neurais envolvidos no controle neuromuscular e para o desenvolvimento de estratégias não invasivas de modulação sensório-motora com potencial aplicação em contextos de reabilitação e treinamento.

II. Metodologia

A. Comissão de Ética

O estudo foi conduzido em estrita conformidade com a Resolução 466/2012 do Conselho Nacional de Saúde e recebeu aprovação prévia do Comitê de Ética em Pesquisa da Universidade Estadual de Campinas (UNICAMP) sob o número de CAAE 88364525.6.0000.5404.

Até o estágio atual do desenvolvimento da pesquisa, participaram onze voluntários destros e saudáveis, sendo seis mulheres e cinco homens, com idade média de $26,18 \pm 4,67$ anos e faixa etária entre 19 e 40 anos. Dessa forma, indivíduos que não fossem destros ou que não atendessem aos critérios de saúde não eram elegíveis para participar do estudo. Além disso, para a caracterização da amostra, os participantes também foram questionados sobre histórico de fraturas na mão, prática de atividades esportivas e uso de óculos; essas informações foram registradas, mas não constituíram critérios de exclusão. Por fim, antes do início dos procedimentos, todos receberam orientações sobre a dinâmica das tarefas e assinaram voluntariamente o Termo de Consentimento Livre e Esclarecido.

B. Aparato Experimental

Durante as sessões experimentais, os participantes permaneceram sentados em uma cadeira, em frente a um monitor digital que exibia a força produzida e a força-alvo por meio de um marcador móvel e de uma barra móvel, respectivamente.

Para assegurar o isolamento do movimento voluntário, o antebraço e a mão correspondentes ao lado dominante do voluntário foram devidamente fixados e imobilizados em um dispositivo mecânico especialmente projetado para aprimorar a ergonomia e a experiência geral do usuário (Fig. 1).

A mensuração da força exercida pelo músculo primeiro interósseo dorsal foi realizada por uma célula de carga multiaxial (Nano25, ATI Instruments), cujos sinais analógicos foram digitalizados por um sistema de aquisição de dados multifuncional (USB-6343, National Instruments) sob uma taxa de amostragem de 2000 Hz. Adicionalmente, os sinais de força foram submetidos a uma filtragem digital passa-baixas com frequência de corte em 30 Hz.

O estímulo tátil foi aplicado próximo à articulação metacarpofalângica no dedo indicador por meio de um atuador ressonante linear (C10-100, Precision Microdrives). Este atuador operou em sua frequência de ressonância de 175 Hz e aceleração de 0,9 G, sendo o controle de ativação da vibração gerenciado pela saída do sistema de aquisição de dados multifuncional (USB-6343, National Instruments). Um programa desenvolvido em LabVIEW foi utilizado para controlar as condições experimentais, fornecer a realimentação visual e registrar os sinais de força.

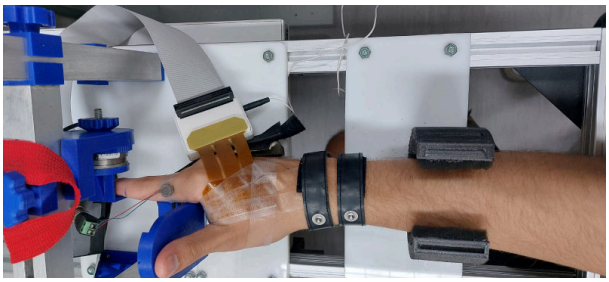


Fig. 1. Aparato Experimental

C. Protocolo Experimental

O protocolo experimental foi estruturado de forma sequencial e iniciou-se obrigatoriamente com a determinação da Força Voluntária Máxima (MVC, do inglês, *maximum voluntary contraction*) de cada participante. Esse parâmetro é fundamental e utilizado como a referência para as etapas subsequentes do experimento. Para o estabelecimento dessa referência, os voluntários realizaram três tentativas de contração isométrica máxima em cada condição experimental: sem vibração (Controle) e com vibração (Vibração). Cada tentativa teve duração de 10 segundos, com intervalo de repouso de 1 minuto entre as repetições. Em cada condição, o maior valor obtido entre as três tentativas foi adotado como a respectiva MVC. Para a normalização das tarefas submáximas, contudo, utilizou-se o maior valor de MVC obtido na condição Controle. Dessa forma, uma única referência foi mantida para ambas as condições, evitando que o uso de diferentes valores de normalização influenciasse a comparação entre Controle e Vibração. A ordem de apresentação das condições foi aleatorizada individualmente para cada participante.

Antes do início das tentativas experimentais, os participantes foram familiarizados com a tarefa de força trapezoidal por meio de dez repetições realizadas com realimentação visual da força produzida e da força-alvo. As repetições incluíram níveis-alvo entre 5% e 70% da MVC, abrangendo as intensidades de 30% e 45% analisadas neste estudo. Durante essa fase, a força produzida e a força-alvo foram apresentadas visualmente, sendo essa trajetória-alvo composta por fases de rampa de subida e descida a uma taxa de 5% MVC/s e por um platô de 15 s.

Em seguida, os participantes realizaram contrações trapezoidais em diferentes intensidades (5%–70% da MVC) com a realimentação visual, cujos resultados não são objeto de análise neste trabalho. Essas etapas foram fundamentais e suficientes para viabilizar a execução da fase experimental sem a realimentação visual.

Na etapa analisada neste estudo, os participantes realizaram contrações nos níveis de 30% e 45% da MVC, calculados exclusivamente a partir da MVC obtida na condição Controle. A trajetória-alvo apresentou taxa de incremento de 5% da MVC/s, sendo a mesma referência utilizada nas condições Controle e Vibração. Nessa fase, a realimentação visual da força produzida foi apresentada apenas nos 2 s iniciais da rampa de subida, enquanto a força-alvo permaneceu visível durante toda a tarefa. Após esse período, os participantes passaram a depender exclusivamente de informações sensório-motoras para continuar o incremento da força e manter o platô o mais estável possível por 15 s. Ao término de cada tentativa, um sinal visual indicava o encerramento da tarefa.

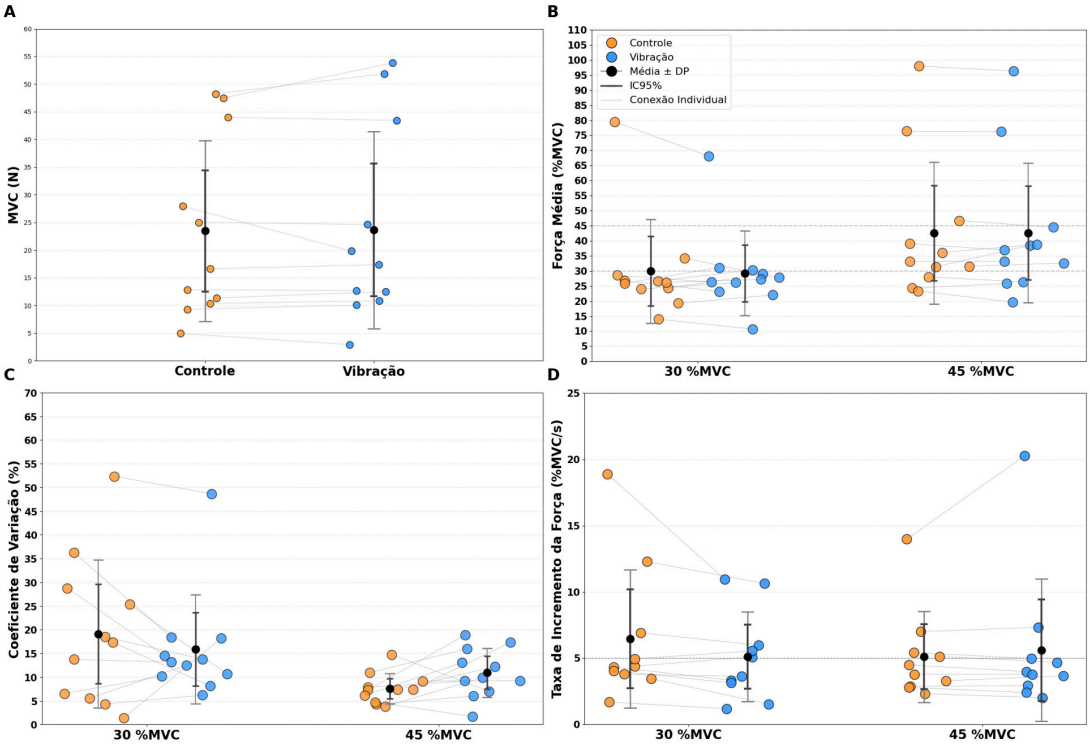


Fig. 2. (A) Força voluntária máxima (MVC) nas condições Controle e Vibração. (B) Força média, (C) coeficiente de variação da força e (D) taxa de incremento da força durante a rampa nas condições Controle e Vibração para as duas intensidades de força-alvo. Os pontos representam os participantes, as linhas conectam os resultados do mesmo participante e os marcadores pretos representam as médias do grupo, acompanhadas das respectivas barras de desvio-padrão e de intervalo de confiança de 95%.

Foram realizadas três repetições para cada combinação entre estimulação (Controle e Vibração) e intensidade de força (30% e 45% da MVC), totalizando doze tentativas por participante. A ordem de execução dessas tentativas foi definida previamente por meio de uma tabela de randomização elaborada em planilha eletrônica, que atribuiu uma sequência individual de apresentação a cada participante. Foi adotado um intervalo de repouso de 1 min entre as tentativas para reduzir possíveis efeitos de fadiga.

D. Análise de Dados

Todos os dados de força coletados foram armazenados digitalmente e processados offline por meio de rotinas e algoritmos desenvolvidos em Python. As variáveis dependentes extraídas para a análise da MVC, a taxa de incremento de força ao longo da rampa após a retirada da realimentação visual, a força média sustentada no platô e a variabilidade da força no platô, medida pelo coeficiente de variação (CoV).

A taxa de incremento de força foi estimada pelo coeficiente angular de uma regressão linear ajustada ao sinal de força durante o período da rampa sem realimentação visual. As variáveis associadas ao platô foram calculadas a partir dos 14 s finais dessa fase, excluindo-se o primeiro segundo de análise devido à presença de efeitos transitórios.

Para a análise estatística, a comparação da MVC entre as condições experimentais foi realizada por meio de uma ANOVA de um fator com medidas repetidas. Já para as variáveis extraídas nas tarefas isométricas sem realimentação visual (força média, CoV e taxa de incremento), empregou-se a ANOVA de medidas repetidas de dois fatores, adotando o estímulo (Controle e Vibração) e a intensidade da força como fatores (30% e 45% MVC). O nível de significância empregado foi de $0{,}05$. Adicionalmente, foi calculado o eta-quadrado parcial (η_p^2) para estimar o tamanho do efeito. Valores entre 0,01 e 0,05 será considerado como efeito pequeno. Entre 0,06 e 0,13 será considerado efeito médio e acima de 0,13 efeito grande. Na Fig. 2, os marcadores pretos representam as médias do grupo, acompanhadas pelas respectivas barras de desvio-padrão e de intervalo de confiança de 95%. Na Tabela I, os resultados são apresentados como média $\pm$ margem do intervalo de confiança de 95%.

III. Resultados

A. Força Máxima Voluntária (MVC)

A Fig. 2A revela uma heterogeneidade entre indivíduos na capacidade de gerar força, com registros variando de 5 a 55 (N), o que justifica o elevado desvio padrão observado na tendência central do grupo. A análise estatística não indicou diferença significativa entre as condições avaliadas ($\eta_p^2 = 0{,}003$, $p = 0{,}875$).

B. Força Média e Coeficiente de Variação

A maioria dos voluntários mostrou-se treinada para realizar a tarefa sem realimentação visual, uma vez que conseguiu

Tabela I

Média ± margem do intervalo de confiança de 95% das variáveis analisadas nas condições Controle e Vibração ($n = 11$).

	30% MVC		45% MVC	
Variável	Controle	Vibração	Controle	Vibração
Força média (%MVC)	$29{,}94 \pm 11{,}57$	$29{,}24 \pm 9{,}43$	$42{,}54 \pm 15{,}82$	$42{,}63 \pm 15{,}53$
CoV da força (%)	$19{,}10 \pm 10{,}49$	$15{,}85 \pm 7{,}73$	$7{,}60 \pm 2{,}14$	$10{,}94 \pm 3{,}48$
Taxa de incremento (%MVC/s)	$6{,}34 \pm 3{,}33$	$5{,}38 \pm 2{,}24$	$5{,}95 \pm 2{,}88$	$6{,}21 \pm 3{,}67$

reproduzir a trajetória de força relativamente bem. Isso é comprovado pelas forças médias próximas ao valor-alvo (Fig. 2B). Ainda assim, observa-se certa dispersão entre os sujeitos, o que revela a dificuldade da tarefa na retirada do auxílio visual. A análise estatística revelou um efeito significativo do fator intensidade da força ($\eta_p^2 = 0{,}341$, $p = 0{,}046$), conforme esperado, mas não do fator estímulo ($\eta_p^2 = 0{,}011$, $p = 0{,}744$), o que demonstra que a aplicação da estimulação vibrotátil não alterou a capacidade do grupo de manter a força média próxima ao alvo. Por fim, a ausência de uma interação significativa entre a força e o estímulo ($\eta_p^2 = 0{,}019$, $p = 0{,}667$) indica que esse comportamento nulo da vibração foi consistente e independente do nível de esforço dos participantes.

Em relação à variabilidade da força, a análise estatística revelou dependência significativa da intensidade de contração ($\eta_p^2 = 0{,}335$, $p = 0{,}048$), mas não da estimulação vibrotátil ($\eta_p^2 < 0{,}001$, $p = 0{,}982$). Da mesma forma, não foi observada interação estatisticamente significativa entre a intensidade de força e a estimulação vibrotátil ($\eta_p^2 = 0{,}293$, $p = 0{,}068$). Entretanto, observa-se, na Fig. 2C, um comportamento com redução do CoV a 30% da MVC e aumento a 45% da MVC na condição de Vibração. O tamanho do efeito foi considerado grande, mas a interação não atingiu significância estatística.

C. Taxa de Incremento

A Fig. 2D evidencia, novamente, elevada variabilidade interindividual, refletida pelos amplos desvios-padrão observados. Esse comportamento sugere diferentes níveis de controle motor entre os participantes ao executarem uma tarefa desafiadora e fortemente dependente da percepção sensório-motora. A análise estatística revelou efeito significativo do fator estímulo ($\eta_p^2 = 0{,}352$, $p = 0{,}042$), mas não da intensidade de força ($\eta_p^2 = 0{,}006$, $p = 0{,}809$). Esse resultado é consistente com o fato de que a taxa de incremento da força-alvo foi mantida constante entre as condições de 30% e 45% da MVC, além dos participantes não terem conhecimento prévio da intensidade-alvo a ser executada. Adicionalmente, a ausência de interação entre os fatores estímulo e intensidade de força ($\eta_p^2 = 0{,}068$, $p = 0{,}413$) indica que o efeito da vibração sobre a taxa de incremento da força foi semelhante nos dois níveis de esforço avaliados.

IV. Discussão

Na amostra analisada, não foi detectada diferença significativa na MVC entre as condições Controle e Vibração. Esse resultado indica ausência de evidência de alteração da força máxima nas condições avaliadas, mas não permite concluir sobre recrutamento ou frequência de disparo das unidades motoras.

Por outro lado, nas tarefas submáximas, o protocolo incluiu uma fase dinâmica de incremento de força (rampa de 5% MVC/s), durante a qual a remoção da realimentação visual associada à aplicação do estímulo vibrotátil reduziu significativamente a taxa de incremento da força. A realimentação visual é reconhecidamente um fator importante para os ajustes motores em tempo real, e sua ausência tende a comprometer a precisão do controle motor e a produção de força [9], [10]. Os resultados obtidos indicam que, nas condições experimentais avaliadas, a presença da estimulação vibrotátil esteve associada a uma menor taxa de incremento da força. Esse achado sugere que a vibração pode influenciar mecanismos sensoriais envolvidos na estimativa da força produzida, levando os participantes a adotar uma estratégia motora mais conservadora à medida que a força aumenta. Entretanto, a possibilidade de a vibração ter influenciado a percepção da força produzida permanece limitado às condições avaliadas neste estudo e como uma hipótese a ser investigada.

A esse respeito, Pearcey et al. [5] observaram um efeito oposto, no qual a estimulação cutânea promoveu um aumento agudo da força durante rampas de flexão plantar. Entretanto, diferenças metodológicas importantes devem ser consideradas, incluindo a natureza do estímulo aplicado (elétrico versus mecânico), a disponibilidade de realimentação visual e o grupo muscular avaliado. De fato, os próprios autores não observaram efeito da estimulação cutânea durante tarefas de dorsiflexão plantar. Assim, os resultados do presente estudo corroboram a ideia de que os efeitos da estimulação sensorial sobre o controle motor dependem fortemente do contexto experimental e da tarefa executada. Em particular, em nossos achados a vibração esteve associada a mecanismos envolvidos no controle dinâmico da força, reduzindo a taxa de aumento da força dos indivíduos quando guiados predominantemente por informações sensório-motoras internas, ou seja, sem realimentação visual, nas condições experimentais avaliadas.

Durante a fase isométrica, foi observado efeito significativo da intensidade de contração sobre o coeficiente de variação. Entretanto, não foi identificado efeito significativo da estimulação vibrotátil nem interação estatisticamente significativa entre estimulação e intensidade. Dessa forma, os resultados obtidos até o momento não permitem afirmar que a estimulação vibrotátil tenha modificado a variabilidade da

força durante o platô ou produzido efeitos diferentes entre as intensidades avaliadas.

A. Limitações do estudo

Os resultados devem ser interpretados considerando algumas limitações. Para exemplificar isso, o reduzido tamanho amostral ($n = 11$) limita o poder estatístico das análises e pode ter dificultado a identificação de efeitos de menor magnitude. Essa limitação é relevante diante da elevada variabilidade interindividual observada nas variáveis analisadas, indicando que os efeitos da estimulação vibrotátil podem não ocorrer de maneira uniforme entre os participantes. Além disso, a inclusão somente de voluntários saudáveis restringe a generalização dos resultados para populações clínicas. Portanto, os achados devem ser considerados preliminares e não permitem estabelecer recomendações definitivas.

B. Implicações práticas

O efeito significativo da estimulação vibrotátil foi identificado somente na taxa de incremento da força, enquanto não foram observados efeitos significativos sobre a força média ou o coeficiente de variação durante o platô. Esse padrão constitui evidência preliminar de um possível efeito seletivo da estimulação nas condições avaliadas, sem permitir, contudo, estabelecer sua relevância funcional. Na prática, esse resultado pode ser relevante para tarefas que exigem um aumento gradual e controlado da força. Entretanto, a redução observada não deve ser interpretada necessariamente como melhora ou prejuízo do desempenho, pois sua aplicabilidade dependerá do objetivo funcional e das exigências da tarefa.

C. Principais aprendizados

O presente estudo demonstrou a importância de analisar separadamente as diferentes fases da produção de força. Ademais, destacou-se a importância de controlar a disponibilidade da realimentação visual da força produzida, permitindo analisar a contribuição da estimulação vibrotátil em uma condição mais dependente das informações sensório-motoras internas. Por último, a elevada variabilidade interindividual mostrou que a resposta ao estímulo não é homogênea e que a análise dos valores individuais deve complementar a interpretação dos resultados médios do grupo.

D. Lacunas de pesquisa

A continuidade da pesquisa, com a ampliação da amostra, permitirá avaliar a consistência dos resultados observados até o momento e reduzir as limitações associadas à variabilidade interindividual. Além disso, estudos futuros deverão incluir populações clínicas, visto que o presente protocolo considera apenas voluntários saudáveis, a fim de verificar a aplicabilidade dos achados em contextos de reabilitação motora. Por fim, pode ser necessário investigar diferentes parâmetros da estimulação vibrotátil, como frequência, amplitude, duração e local de aplicação, além de outras intensidades de contração e tarefas motoras.

V. Conclusão

Na amostra analisada até o momento, não foi detectada diferença significativa na força máxima entre as condições controle e vibração. Nas tarefas submáximas, foi identificado efeito significativo da estimulação vibrotátil sobre a taxa de incremento da força durante a fase de rampa sem realimentação visual da força produzida. Durante o platô, não foram observados efeitos significativos sobre a força média ou o coeficiente de variação. Esses resultados constituem evidências preliminares de um possível efeito seletivo da estimulação vibrotátil sobre a fase dinâmica da produção de força nas condições avaliadas. Entretanto, os dados obtidos não permitem determinar os mecanismos sensoriais ou neurais responsáveis pelo resultado. Assim, a continuidade do estudo, com a ampliação da amostra, e futuras investigações serão necessárias para confirmar os achados e esclarecer os mecanismos envolvidos.

Agradecimentos

Os autores agradecem a bolsa de iniciação científica concedida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio do Programa Institucional de Bolsas de Iniciação Científica (PIBIC), e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo 2024/08318-4).

Referências

[1] Y. Masakado, G. Kamen, and C. J. De Luca, "Effects of percutaneous stimulation on motor unit firing behavior in man," *Exp. Brain Res.*, vol. 86, no. 2, pp. 426–432, 1991.

[2] R. Garnett and J. A. Stephens, "Changes in the recruitment threshold of motor units produced by cutaneous stimulation in man," *J. Physiol.*, vol. 311, pp. 463–473, 1981.

[3] J. A. Stephens, R. Garnett, and N. P. Buller, "Reversal of recruitment order of single motor units produced by cutaneous stimulation during voluntary muscle contraction in man," *Nature*, vol. 272, no. 5651, pp. 362–364, 1978.

[4] R. Garnett and J. A. Stephens, "The reflex responses of single motor units in human first dorsal interosseous muscle following cutaneous afferent stimulation," *J. Physiol.*, vol. 303, pp. 351–364, 1980.

[5] G. E. P. Pearcey, Y. Sun, and E. P. Zehr, "Plantarflexion force is amplified with sensory stimulation during ramping submaximal isometric contractions," *J. Neurophysiol.*, vol. 123, no. 4, pp. 1427–1438, 2020.

[6] C. M. Germer, L. S. Moreira, and L. A. Elias, "Sinusoidal vibrotactile stimulation differentially improves force steadiness depending on contraction intensity," *Med. Biol. Eng. Comput.*, vol. 57, no. 8, pp. 1813–1822, 2019.

[7] C. M. Germer, L. S. Moreira, and L. A. Elias, "Assessment of force control improvement induced by sinusoidal vibrotactile stimulation in dominant and non-dominant hands," *Res. Biomed. Eng.*, vol. 37, no. 1, pp. 95–103, 2021.

[8] C. M. Germer, A. Del Vecchio, F. Negro, D. Farina, and L. A. Elias, "Neurophysiological correlates of force control improvement induced by sinusoidal vibrotactile stimulation," *J. Neural Eng.*, vol. 17, no. 1, Art. no. 016043, 2020.

[9] H. S. Baweja, D. M. Kennedy, J. Vu, D. E. Vaillancourt, and E. A. Christou, "Greater amount of visual feedback decreases force variability by reducing force oscillations from 0–1 and 3–7 Hz," *Eur. J. Appl. Physiol.*, vol. 108, no. 5, pp. 935–943, 2010.

[10] H. S. Baweja, B. K. Patel, J. D. Martinkewiz, J. Vu, and E. A. Christou, "Removal of visual feedback alters muscle activity and reduces force variability during constant isometric contractions," *Exp. Brain Res.*, vol. 197, no. 1, pp. 35–47, 2009.

Effects of Transcranial Photobiomodulation on Cognitive Function Following Ischemic Stroke: A Randomized Clinical Trial Assessed by the Mini-Mental State Examination (MMSE)

Fernanda de Souza Silva
Graduate Program in Biomedical Engineering
Universidade Anhembi Morumbi
São José dos Campos, Brazil
fisio.fer@yahoo.com

Pâmela Camila Pereira
Graduate Program in Biomedical Engineering
Universidade Anhembi Morumbi
São José dos Campos, Brazil
pam_milaf@yahoo.com.br

Renato Amaro Zângaro
Graduate Program in Biomedical Engineering
Universidade Anhembi Morumbi
São José dos Campos, Brazil
razangaro@gmail.com

***Abstract* — Ischemic stroke is frequently associated with cognitive deficits that negatively impact patients' functionality and life quality. Transcranial photobiomodulation (tPBM) has emerged as a non-invasive therapeutic approach with potential to modulate neural activity and promote neuroplasticity. The aim of this study was to evaluate the effects of tPBM on cognitive function in post-stroke patients using the Mini-Mental State Examination (MMSE) as the primary outcome measure. This randomized, controlled, single-blind, longitudinal clinical trial was conducted in a home-based setting. The sample consisted of 16 patients allocated into two groups: an intervention group (active tPBM, n = 8) and a control group (Sham, n = 8). The protocol was applied three times per week for four weeks. Statistical analyses were performed using the Wilcoxon and Mann–Whitney tests, adopting $p < 0.05$. Significant intragroup improvements in MMSE scores were observed both in the tPBM group ($p = 0.0078$) and the Sham group ($p = 0.0384$). In the intergroup comparison, the tPBM group demonstrated superior cognitive improvement compared with the control group ($p = 0.0382$). These findings suggest a possible benefit of tPBM on the global cognitive function of post-ischemic stroke patients. However, these findings should be interpreted as preliminary evidence, considering the small sample size, the baseline difference between the groups, and the exclusive use of the MMSE.**

Keywords — Photobiomodulation; Stroke; Cognition; Mini-Mental State Examination; Neurorehabilitation.

I. INTRODUCTION

Stroke remains one of the leading causes of disability worldwide, with the ischemic subtype accounting for most cases. In addition to motor deficits, cognitive impairments are highly prevalent and directly impact functional independence, quality of life, and long-term prognosis. Alterations in domains such as attention, memory, and executive functions are frequently observed, highlighting the need for therapeutic strategies capable of acting not only on motor recovery but also on the functional reorganization of cognitive systems [1].

In recent years, Transcranial Photobiomodulation (tPBM) has emerged as a promising non-invasive approach in neurological rehabilitation. This technique is based on the application of light in the red to near-infrared spectrum, capable of interacting with intracellular chromophores, particularly cytochrome c oxidase, promoting increased adenosine triphosphate (ATP) production and improving cellular bioenergetic efficiency. Additionally, tPBM may modulate cerebral blood flow, oxidative stress, and inflammatory processes, contributing to stabilization of neuronal activity and functional dynamics of neural networks involved in cognition [1–2].

From this perspective, experimental and clinical evidence suggests that tPBM may act as a modulator of neuroplasticity processes following ischemic stroke, promoting neuronal survival, synaptic function, and cortical reorganization. Although early studies indicate benefits regarding cognitive function, this area remains underexplored, especially in randomized clinical trials using standardized and measurable cognitive outcomes [1–5].

Despite these advances, an important gap remains in the literature regarding the systematic evaluation of the effects of tPBM on cognitive function in post-ischemic stroke patients, particularly in controlled studies using validated clinical instruments. The Mini-Mental State Examination (MMSE) stands out as a widely used tool for global cognitive assessment, being sensitive to alterations in different cognitive domains and suitable for monitoring therapeutic interventions, however, because it is a global screening instrument, it presents limited sensitivity to detect specific changes in executive functions, attention, processing speed, and memory [5-7].

Therefore, the present study aimed to evaluate the effects of transcranial photobiomodulation on global cognitive function in patients with ischemic stroke through a randomized pilot clinical trial, using the MMSE as the primary outcome measure.

II. MATERIALS AND METHODS

A. Ethical Considerations

This study was approved by the Research Ethics Committee of Centro Universitário de Itajubá – FEPI (CAAE: 74363523.6.0000.5094) and registered in the Brazilian Registry of Clinical Trials (ReBEC) under registration number U1111-1304-8858. Patients who agreed to participate in the study signed an informed consent form before the beginning of treatment.

This was a randomized, controlled, single-blind, longitudinal clinical trial conducted in a home-based setting.

B. Sample Characterization

The sample of the present study consisted of 16 patients clinically diagnosed with ischemic stroke, of both sexes, aged

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

between 50 and 70 years. Inclusion criteria included a clinical diagnosis of ischemic stroke, clinical stability, and minimum cognitive capacity to understand and perform the proposed assessments. Exclusion criteria included the presence of other associated neurological diseases, a history of neoplasms, photosensitivity disorders, severe psychiatric disorders, or any clinical condition that could interfere with the application of photobiomodulation or cognitive assessment.

It is important to note that participants were not necessarily undergoing physiotherapeutic follow-up during the study period. Before the beginning of the protocol, a comprehensive initial evaluation was performed, including the collection of personal, sociodemographic, and clinical data, as well as the assessment of vital signs. This stage aimed to characterize the sample and verify possible baseline differences between groups, contributing to initial homogeneity and the appropriate interpretation of the results throughout the study.

Cognitive function was assessed using the Mini-Mental State Examination (MMSE), a widely validated instrument commonly used in clinical practice and research for global cognitive evaluation. The instrument was administered at two distinct time points: before the intervention (pre-intervention assessment) and after completion of the protocol (post-intervention assessment). This longitudinal approach allowed for a direct comparison of the obtained scores, enabling analysis of the evolution of cognitive performance and identification of possible changes associated with the proposed intervention. Assessments were conducted by a previously trained evaluator following a standardized administration protocol. It should be noted, however, that the MMSE is a global screening instrument and does not replace a comprehensive neuropsychological battery.

C. Cognitive Function Assessment (MMSE)

Mini-Mental State Examination (MMSE)

The MMSE is an instrument used for global cognitive assessment, allowing the evaluation of different domains such as temporal and spatial orientation, immediate memory, attention and calculation, recall, language, and visuoconstructive ability. In the present study, the MMSE was used as the primary cognitive outcome measure to verify possible changes in participants' cognitive performance before and after the tPBM protocol.

The test was administered in a quiet environment, with the participant comfortably seated and free from external interference that could compromise attention or performance. Before the beginning of the assessment, the patient was instructed regarding the procedure and asked to respond to the tasks proposed by the evaluator according to their cognitive capacity at the time of application.

During the assessment, tasks related to temporal and spatial orientation, word registration and recall, attention, calculation, naming, repetition, comprehension of verbal commands, reading, writing, and figure copying were administered. The total MMSE score ranges from 0 to 30 points, with higher scores indicating better global cognitive performance.

The instrument was administered at two distinct time points: before the intervention and after completion of the therapeutic protocol. Comparison between pre- and post-intervention scores allowed analysis of participants' cognitive function evolution throughout the study and verification of possible effects of transcranial photobiomodulation on global cognitive performance in patients with ischemic stroke.

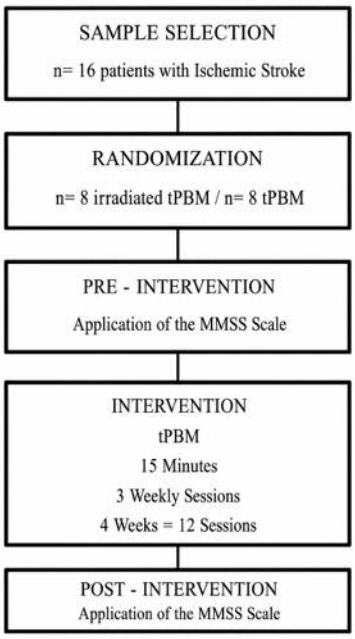
D. Group Allocation

Sixteen patients clinically diagnosed with ischemic stroke were selected to participate in the study and randomly allocated into two groups: the Intervention Group (irradiated tPBM), composed of 8 patients who received active tPBM with the helmet turned on, and the Control Group (Sham), composed of 8 patients submitted to the same protocol, although with the inactive device. This was a pilot study designed to evaluate protocol feasibility and estimate the effect of the intervention in patients with ischemic stroke.

Participant allocation was performed through randomization, ensuring equivalent distribution between groups and reducing possible selection biases. The study followed a single-blind design in which participants were unaware of the group to which they belonged.

Interventions were performed three times per week for four consecutive weeks, totaling 12 sessions per participant. The tPBM protocol was applied in a standardized manner in all sessions, respecting the same parameters for the intervention group and simulated conditions for the control group. Figure 1 presents the experimental design of the study, including pre-intervention assessment, protocol application, and post-intervention reassessment.

Figure 1 – Experimental design of the randomized clinical trial.



Source: Personal Archive.

E. tPBM Protocol for Cognitive Function in Post-Stroke Patients

The tPBM intervention was performed using a helmet composed of gallium-aluminum-arsenide (GaAlAs) light-emitting diodes (LEDs), arranged to cover cortical regions related to motor control. The device was designed to provide homogeneous distribution of light radiation over the scalp, favoring light penetration into encephalic structures.

The helmet emitted near-infrared radiation at a wavelength of 808 nm, a value widely described in the literature due to its high tissue penetration capacity and interaction with cytochrome c oxidase, favoring cellular bioenergetic effects. The device presented an acrylic encapsulation and an approximate diameter of 5 mm per emitter. The system had a total optical power of 6 W and a power density of 3.41 mW/cm², allowing application for 900 seconds, with an approximate energy density of 3.07 J/cm² per session.

During the application, the patient remained in a comfortable position with the helmet properly adjusted to the cranial region, ensuring adequate contact and uniform light distribution. To ensure appropriate biosafety conditions, the device was disinfected with 70% alcohol after each session.

F. Statistical Analysis

Statistical analysis was performed using GraphPad Prism 8.0 software (GraphPad Software Inc., La Jolla, CA, USA). Considering the reduced sample size and the ordinal nature of the MMSE scores, non-parametric statistical tests were adopted.

For intragroup comparison of MMSE scores between pre- and post-intervention periods, the paired Wilcoxon test was used. For intergroup comparison (Intervention Group vs Sham Group), the Mann–Whitney test was employed on the individual variations of the scores (Δ = post-intervention − pre-intervention). The data was presented as mean and standard deviation. A significance level of $p < 0.05$ was adopted. To complement the p-values, the mean difference between individual variations and the rank-biserial correlation (r_rb) were estimated as a measure of effect size for the intergroup comparison, with 95% confidence intervals obtained by bootstrapping with 10,000 resamples.

III. RESULTS

The study sample consisted of 16 patients clinically diagnosed with ischemic stroke, distributed into two groups: the irradiated tPBM group (n = 8) and the Sham group (n = 8). Table 1 presents the initial characterization of participants regarding age, sex, and time since injury. Descriptive data are presented as mean and standard deviations. Statistical analysis demonstrated no significant differences between groups regarding age ($p > 0.05$), sex distribution ($p > 0.05$), and time since injury ($p > 0.05$). However, the absence of statistical significance does not prove equivalence between the groups. In addition, a baseline difference was observed in MMSE scores, with a mean of 24.4 ± 2.5 in the irradiated tPBM group and 20.0 ± 3.8 in the Sham group, an aspect that must be considered in the interpretation of the results.

Table 1 - Initial characterization of participants in the irradiated tPBM and Sham groups.

Variables , avarage (DP)	Irradiated tPBM	tPBM Sham	*p*-value
Age (years)	58,1 (6,7)	60,5 (7,4)	p > 0,05
Sex (F/M)	4/4	4/4	p >0,05
Time Since Injury (years)	2,0 (1,1)	3,0 (2,5)	p >0,05

Caption: Female Sex; M: Male Sex
Source: Personal Archive

Considering the reduced sample size and the ordinal nature of MMSE scores, non-parametric statistical tests were adopted. Table 2 presents MMSE scores of the irradiated tPBM and Sham groups in the pre- and post-intervention periods, as well as p-values related to intragroup and intergroup analyses, expressed as mean and standard deviations.

Intragroup analysis demonstrated statistically significant improvement in cognitive performance after the intervention in both groups, with significant differences observed in the irradiated tPBM group ($p = 0.0078$) and in the Sham group ($p = 0.0384$), although with different magnitudes.

Regarding comparison between groups, the tPBM group demonstrated superior cognitive improvement compared with the Sham group, and this difference was statistically significant ($p = 0.0382$). Descriptively, the mean variation was 2.13 ± 1.13 points in the irradiated tPBM group and 0.88 ± 0.83 points in the Sham group. The mean difference between the variations was 1.25 points (95% bootstrap CI: 0.38 to 2.13), and the rank-biserial correlation was $r_rb = 0.61$ (95% bootstrap CI: 0.16 to 0.92).

Considering the pilot nature of the study and the reduced sample size, the findings suggest a positive effect of tPBM on cognitive function in patients with ischemic stroke, indicating a potential benefit of the intervention. However, these results should be interpreted as preliminary evidence, and larger samples are necessary to confirm the current findings. Although the direction and magnitude of the observed effect favored the irradiated tPBM group, the width of the confidence intervals highlights imprecision in the estimates, especially in light of the small sample size and the baseline difference in MMSE scores. Thus, the results do not allow establishing the definitive clinical efficacy of the intervention.

Table 2 - Evaluation of MMSE scores in the irradiated tPBM and Sham groups, pre- and post-intervention, with intragroup and intergroup analyses.

Group	Pre, mean (SD)	Post, mean (SD)	*Intragroup p-value*
Irradiated tPBM	24,4 (2,5)	26,5 (2,5)	0,0078

FBMT Sham	20,0 (3,8)	20,9 (3,7)	0,0384
Intergroup comparison of pre-post variations (*Mann-Whitney Test*)			$p = 0,0382$

Caption: SD: Standard Deviation
Source: Personal Archive

Improvement in MMSE scores was observed in both groups, although it was more meaningful in the group submitted to tPBM.

IV. DISCUSSION

Scientific literature demonstrates that ischemic stroke is frequently associated with cognitive deficits that significantly compromise individuals' functionality and quality of life. Alterations in domains such as memory, attention, and executive functions reflect dysfunctions in the organization and dynamics of neural networks, making cognitive recovery one of the main challenges in neurological rehabilitation [12–13]. tPBM has emerged as a promising approach, acting in the modulation of neural activity through bioenergetic and neurovascular mechanisms. In general, the interaction between near-infrared light and intracellular chromophores is associated with increased neuronal metabolic efficiency, modulation of oxidative stress, and improvement in cerebral perfusion, creating a favorable environment for cerebral neuroplasticity and cortical reorganization [1–3]. According to Gottesman and Hillis (2020), post-stroke cognitive alterations are strongly associated with reduced functional independence and worse long-term clinical prognosis. Likewise, Cramer (2008) highlights that functional cerebral reorganization represents one of the main mechanisms related to neurological recovery after ischemic stroke.

The results of the present study demonstrated significant improvement in MMSE scores in the group submitted to tPBM, with greater improvement than that observed in the Sham group. Despite the pilot nature of the study and the reduced sample size, the findings suggest positive effects of the intervention on cognitive function in patients with ischemic stroke. In the intergroup comparison of pre-post variations, the mean difference was 1.25 points (95% CI: 0.38 to 2.13), and the rank-biserial correlation was 0.61 (95% CI: 0.16 to 0.92), indicating an effect magnitude favorable to the irradiated tPBM group. However, the width of the confidence intervals highlights imprecision in the estimates.

These findings are aligned with recent studies pointing to the potential of tPBM in improving cognitive performance in neurological populations, demonstrating gains in functions such as memory, attention, and processing speed following tPBM application, reinforcing the consistency of the results observed in the present study. Similar findings were described by Naeser et al. (2020), who observed improvements in cognitive performance after transcranial tPBM application, especially in functions related to memory and attention. In addition, Li, Wong, and Ng (2024) suggest that tPBM presents therapeutic potential in post-stroke functional recovery due to its ability to modulate neurovascular and metabolic mechanisms involved in cortical reorganization. However, differences between populations, protocols, and assessment instruments require caution in comparing results.

Despite the improvement observed in both groups, possibly associated with factors such as learning effects or spontaneous recovery, the statistically significant intergroup difference suggests that tPBM may have exerted a specific effect on cognitive function [5–7]. Nevertheless, since this was a pilot study with a reduced sample size, these findings should be interpreted with caution and more appropriately considered preliminary evidence. In addition, the baseline difference in MMSE scores, with a higher initial score in the irradiated tPBM group, may have influenced the magnitude of the observed variation and favored the occurrence of a ceiling effect.

The improvement observed in the Sham group also deserves attention, since part of this result may be related to non-specific intervention effects, such as placebo effects, in addition to familiarization with the assessment instrument itself. Reapplication of the MMSE may facilitate participants' performance simply due to prior contact with the tasks. Furthermore, post-stroke patients frequently present some degree of spontaneous recovery over time, which may also influence the results. These aspects reinforce the importance of studies with greater methodological control, longer follow-up periods, and the use of complementary cognitive instruments [1,6]. Arevalo-Rodriguez et al. (2021) emphasize that repeated cognitive testing may be influenced by learning effects, especially in global cognitive screening assessments such as the MMSE.

Overall, the findings of the present study are consistent with recent literature suggesting tPBM as potentially useful strategy in modulation of cognitive function in neurological populations. Studies conducted by Liebert et al. (2021), Hamblin (2021), and Mitrofanis (2020) reported improvement in parameters related to neural activity and cognitive performance following tPBM application in neurological populations, supporting the plausibility of the results observed in the present study. However, the heterogeneity of the studied populations, irradiation protocols, assessment instruments, and methodological designs limits direct comparison between studies.

From a clinical perspective, the magnitude of improvement observed in the group submitted to tPBM highlights the potential of the technique as a complementary resource in post-stroke cognitive rehabilitation. This is a non-invasive intervention, relatively simple to apply. In this sample, no relevant adverse events were observed, other than the report of heating during application. These characteristics broaden its possibilities for use, including in home-based settings. This aspect is especially relevant when considering the functional limitations frequently present in this population and the difficulties in accessing specialized services [1,4,8]. However, statistical significance alone does not demonstrate clinical relevance. Since a minimal clinically important difference has not been defined for the MMSE and functional or quality-of-life outcomes were not evaluated, it is not possible to state that the difference observed between the groups represents a relevant clinical benefit [1,4,8].

Regarding biophysical aspects, it is possible that the observed effects are related to the irradiation parameters used, such as wavelength, energy density, and exposure time. These factors may influence the neuronal response to tPBM,

suggesting a possible dose-dependent behavior [2–4]. However, the response to tPBM depends on the combination of parameters, including emission mode, irradiated area, and total energy, making it impossible to attribute the result to a single parameter. The heterogeneity of devices and protocols available in the literature also limits direct comparisons. Therefore, this interpretation should be approached with caution and still requires further investigation in future studies.

Finally, some limitations must be considered, such as the small sample size, which limits statistical power and increases the possibility of unstable estimates of the observed effect. The relatively wide confidence intervals reinforce the imprecision of these estimates. The sequential allocation of participants, performed according to the order of acceptance, may also have introduced selection bias and contributed to baseline differences between the groups. The short intervention period and the absence of longitudinal follow-up prevent verifying the maintenance of the results. Furthermore, the use of a single cognitive assessment instrument may not address all dimensions of cognition, especially executive functions, attention, processing speed, and memory [13,14]. The absence of functional or quality of life outcomes also prevents determining whether the improvement observed in the MMSE translates into relevant clinical benefits. Another limitation corresponds to the incomplete characterization of the sample regarding the location and extent of the lesions, initial severity of the stroke, education level, current medications, and concomitant rehabilitation interventions. Therefore, new studies with larger samples, randomized allocation, better control of baseline differences, longer follow-up time, more detailed clinical characterization, and multimodal cognitive assessment are needed to confirm the findings and clarify the role of tPBM in post-stroke cognitive rehabilitation.

V. Conclusion

The results of the present study suggest a potential effect of tPBM in the modulation of cognitive function in patients with ischemic stroke, evidenced by greater gains in MMSE scores compared with the control group. However, the findings must be interpreted as preliminary evidence, considering the small sample size, the baseline difference between the groups, and the exclusive use of the MMSE.

Due to its non-invasive nature, low risk, and ease of application, tPBM presents potential as a complementary approach in post-stroke cognitive rehabilitation. However, confirming these results in studies with larger samples, a more comprehensive neuropsychological assessment, functional outcomes, and longitudinal follow-up is fundamental before its clinical application can be recommended.

ACKNOWLEDGMENTS

This study was supported by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES). The authors also thank Instituto Ânima for the support provided for the development of this study.

CONFLICT OF INTEREST

The authors declare that there will be no conflict of interests.

REFERENCES

[1] Li S, Wong TWL, Ng SSM. Potential and challenges of transcranial photobiomodulation for the treatment of stroke. CNS Neurosci Ther. 2024;30(12):e70142. doi:10.1111/cns.70142.

[2] Tang L, Li X, Kang J, Huang Y, Wen Y, Tang M. Application of non-invasive transcranial photobiomodulation in ischemic stroke: mechanisms and current insights. iScience. 2025;29(1):114254. doi:10.1016/j.isci.2025.114254.

[3] Yan B, Zhou J, Yan F, Gao M, Tang J, Huang L, et al. Unlocking the potential of photobiomodulation therapy for brain neurovascular coupling: biological effects and medical applications. J Cereb Blood Flow Metab. 2025;45(5):800-830. doi:10.1177/0271678X241311695.

[4] Salehpour F, Mahmoudi J, Kamari F, Sadigh-Eteghad S, Rasta SH, Hamblin MR. Brain photobiomodulation therapy: a narrative review. Mol Neurobiol. 2018;55(8):6601-6636. doi:10.1007/s12035-017-0852-4.

[5] Nagaratnam JM, Sharmin S, Diker A, Lim WK, Maier AB. Trajectories of Mini-Mental State Examination scores over the lifespan in general populations: a systematic review and meta-regression analysis. Clin Gerontol. 2022;45(3):467-476. doi:10.1080/07317115.2020.1756021.

[6] Arevalo-Rodriguez I, Smailagic N, Roqué-Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, et al. Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment. Cochrane Database Syst Rev. 2021;7(7):CD010783. doi:10.1002/14651858.CD010783.pub3.

[7] Truong QC, Cervin M, Choo CC, Numbers K, Bentvelzen AC, Kochan NA, et al. Examining the validity of the Mini-Mental State Examination (MMSE) and its domains using network analysis. Psychogeriatrics. 2024;24(2):259-271. doi:10.1111/psyg.13069.

[8] Naeser MA, Zafonte R, Krengel MH, Martin PI, Frazier J, Hamblin MR, et al. Significant improvements in cognitive performance following transcranial photobiomodulation. Photobiomodul Photomed Laser Surg. 2020;38(6):345-356. doi:10.1089/photob.2019.4738.

[9] Cassano P, Cusin C, Mischoulon D, Hamblin MR, De Taboada L, Pisoni A, et al. Near-infrared transcranial radiation for major depressive disorder: proof-of-concept study. Mol Psychiatry. 2016;21(2):169-175. doi:10.1038/mp.2015.36.

[10] Hennessy M, Hamblin MR. Photobiomodulation and the brain: a new paradigm. J Opt. 2017;19(1):013003. doi:10.1088/2040-8986/19/1/013003.

[11] Jokinen H, Melkas S, Ylikoski R, Pohjasvaara T, Kaste M, Erkinjuntti T. Post-stroke cognitive impairment is common even after mild stroke. Eur J Neurol. 2020;27(7):1232-1239. doi:10.1111/ene.14209.

[12] Gottesman RF, Hillis AE. Predictors and assessment of cognitive dysfunction resulting from ischemic stroke. Lancet Neurol. 2020;19(10):860-871. doi:10.1016/S1474-4422(20)30256-9.

[13] Cramer SC. Repairing the human brain after stroke: mechanisms of spontaneous recovery. Ann Neurol. 2008;63(3):272-287. doi:10.1002/ana.21393.

[14] Nudo RJ. Recovery after brain injury: mechanisms and principles. Front Hum Neurosci. 2019;13:344. doi:10.3389/fnhum.2019.00344.

[15] Hamblin MR. Photobiomodulation for Parkinson's disease: mechanisms and therapeutic perspectives. J Neurosci Res. 2021;99(1):8-20. doi:10.1002/jnr.24705.

[16] Mitrofanis J, Henderson LA, Morries L. Could photobiomodulation be used as a non-invasive therapy for Parkinson's disease? Neurophotonics. 2020;7(4):040901. doi:10.1117/1.NPh.7.4.040901

Engineering principles of Machine Learning-Driven Adaptive Deep Brain Stimulation: A narrative review

1st João Victor Barbosa de Macedo
Instituto Internacional de Neurociências Edmond e Lily Safra (IIN-ELS)
Instituto Santos Dumont (ISD)
Macaíba, Brazil
joaovictorbm2001@gmail.com

***Abstract*—Deep brain stimulation (DBS) is an established therapy for the motor symptoms of Parkinson's disease (PD), but conventional systems deliver stimulation continuously and at fixed parameters, independent of the patient's fluctuating neurological state. Closed-loop or adaptive DBS (A-DBS) addresses this limitation by sensing neural activity and adjusting stimulation in real time. This paper reviews the engineering foundations of closed-loop DBS and the role of machine learning (ML) in personalizing therapy. This review explains Parkinson's disease as a control problem, summarizes the neural and imaging biomarkers used as feedback signals, particularly subthalamic beta-band local field potentials (LFPs) and diffusion/functional imaging correlates of stimulation response, and examines supervised and reinforcement learning approaches to real-time parameter selection. We further discuss clinical evidence supporting A-DBS, the regulatory and hardware constraints specific to implantable devices, and open challenges related to personalization, safety, and generalization across patients. The review concludes that the integration of neural biomarkers, wearable sensing, and machine learning represents a methodological template that is likely to extend beyond Parkinson's disease to other conditions treated by invasive neuromodulation.**

***Index Terms*—deep brain stimulation, machine learning, Parkinson's disease, Adaptative neuromodulation.**

I. Introduction

Parkinson's disease (PD) is among the most common neurodegenerative disorders, being characterized by tremor, rigidity, bradykinesia, and postural instability, alongside other motor and non-motor symptoms, arising from disrupted electrical signaling within the basal ganglia. Degeneration of dopaminergic neurons in the substantia nigra pars compacta disturbs the balance of excitation and inhibition throughout the motor circuit, producing abnormal, pathologically synchronized oscillatory activity that locks the motor system into a dysfunctional state.

Deep brain stimulation (DBS) has been used to treat the motor symptoms of PD since the pioneering work of Benabid and colleagues in the early 1990s [1]. By implanting electrodes in specific subcortical targets and delivering continuous high-frequency electrical pulses, DBS disrupts pathological oscillatory patterns and produces substantial motor improvement. It is now the standard of care for patients whose symptoms can no longer be adequately controlled with medication alone or suffer from the side effects of said treatment.

Despite this success, conventional DBS remains an open-loop system ,where stimulation parameters are set by a clinician during periodic outpatient visits and then held constant, regardless of how symptoms fluctuate over the course of the day. This mismatch between a static controller and a dynamic physiological system motivates closed-loop, or adaptive DBS (A-DBS), in which neural signals recorded from the implanted electrode are used to continuously adjust stimulation to the patient's current state. Machine learning is increasingly central to this transformation, providing the means to extract patient-specific control policies from neural and behavioral data. This review examines the engineering basis of DBS, the biomarkers available for closed-loop control, the machine learning methods being applied to this problem, and the clinical and regulatory context in which they must operate. This work aims to clarify the engineering principles underlying machine learning strategies for A-DBS.

II. Search strategy

The selection of literature in this narrative review was primarily informed by the author's clinical and research background in deep brain stimulation for Parkinson's disease, rather than a predefined systematic search protocol. This initial selection was complemented, during manuscript revision, by targeted searches in PubMed and Google Scholar using combinations of "adaptive deep brain stimulation," "machine learning," and "closed-loop," conducted to verify claims, identify recent evidence, and address specific gaps regarding quantitative comparison across studies, cross-patient generalization and reinforcement-learning safety. As a narrative rather than a systematic review, no PRISMA protocol or formal quality appraisal was applied, and the coverage of evidence reflects the reviewer's judgment rather than an exhaustive database query, a limitation acknowledged here.

III. Parkinson's Disease as a Control Problem

From an engineering standpoint, PD can be framed as the failure of a biological feedback control system. The motor system normally operates through parallel loops connecting the cortex, basal ganglia, and thalamus, which refine and gate motor commands to enable smooth, coordinated movement. Dopamine released from the substantia nigra acts as a neuromodulator that calibrates these loops, enhancing intended movements and suppressing unwanted ones.

As dopaminergic neurons degenerate, this balance shifts toward excessive thalamic inhibition and reduced cortical drive for voluntary movement, and neural firing across the basal ganglia becomes abnormally synchronized, particularly within the beta frequency band (13–30 Hz). Under normal conditions, beta oscillations are brief and transient, but in PD they become persistent and pathologically amplified, locking the motor circuit into a dysfunctional attractor state [3]. Because beta power fluctuates with symptom severity, diminishing with dopaminergic medication and rising during symptom worsening, it is a natural candidate feedback signal for closed-loop control.

IV. DBS Hardware and Stimulation Mechanism

A DBS system consists of an implanted electrode lead, an implantable pulse generator (IPG), and an extension cable. The lead, typically 1.27 mm in diameter and carrying four to eight contacts, is stereotactically placed in the target nucleus, most often the subthalamic nucleus (STN) or globus pallidus interna (GPi) for PD. The IPG, implanted subcutaneously below the clavicle, delivers charge-balanced biphasic pulses (A primary pulse followed by a charge-recovery phase of opposite polarity) preventing net charge accumulation and electrochemical damage on the neural tissue around the lead.

Four stimulation parameters can be adjusted on the therapy: frequency (typically 130–180 Hz is standard for PD but in some cases can be as low as 60 Hz), amplitude (the amount of electric current discharged by each pulse, generally 1.0 – 3.6 mA), pulse width (the duration of each pulse, usually 60 – 90 microseconds), and contact configuration (which contacts are active and their polarity). Newer directional leads with segmented contacts allow the stimulation field to be steered in three dimensions, improving selectivity for the target nucleus while sparing adjacent structures, therefore avoiding areas that could cause unwanted effects. Programming these systems to find an optimal parameter combination is time-consuming and represents a significant clinical and economic burden, motivating algorithm-assisted programming, as well as fully closed-loop control.

The mechanism underlying the therapeutic effect of high-frequency stimulation has been debated, but current evidence favors a disruption model in which DBS overrides pathological low-frequency synchronized oscillations with high-frequency, temporally irregular firing, decoupling the circuit from its dysfunctional state rather than simply inhibiting or exciting it [2].

TABLE I
Candidate Biomarkers for Closed-Loop DBS Control

Signal	What It Reflects	Maturity
STN beta power (13–30 Hz)	Overall pathological synchronization	Clinically validated [4]
Beta bursts	Duration/magnitude of transient beta episodes	Research stage [6]
Gamma / HFO (60–90 Hz)	Correlates with therapeutic stimulation effect	Research stage
DTI tractography	Electrode-target connectivity, contact placement	Preoperative planning
Resting-state fMRI	Whole-brain connectivity correlates of DBS response	Research stage [11]
Wearable IMU	Tremor, gait, bradykinesia (peripheral)	Clinically available

V. Neural and Imaging Biomarkers for Closed-Loop Control

A closed-loop system requires a reliable, real-time signal reflecting symptom severity. The most extensively studied candidate is the beta-band local field potential (LFP) recorded from the stimulation electrode itself. LFPs reflect the summed synaptic activity of neuron populations near the contacts and, unlike single-unit recordings, are robust and stable over years, allowing the same macroelectrode contacts to be used for both stimulation and sensing.

Landmark work by Little and colleagues demonstrated that STN beta power correlates with motor symptom severity within individual patients, a relationship stable enough to serve as a control signal in human trials [4]. Subsequent analysis by Tinkhauser and colleagues showed that it is not overall beta power but the duration and magnitude of discrete beta "bursts", episodes of elevated activity lasting roughly 100 to 500 ms, that correlates most specifically with motor impairment, with long-duration bursts being pathological and brief transient bursts appearing to be part of normal motor processing [6]. A well-designed closed-loop controller should ideally distinguish between the two.

Beyond electrophysiology, neuroimaging contributes a pre-operative dimension to biomarker-driven DBS. Diffusion tensor imaging (DTI) of white matter tracts can help predict which electrode contacts are optimally positioned relative to therapeutic targets, reducing the empirical parameter search after implantation, while resting-state functional MRI connectivity patterns have been associated with DBS response, offering a surgery-independent predictor of outcome. Several groups have demonstrated the feasibility of integrating such imaging biomarkers with neural signal classifiers to build patient-specific therapy models [11]. Gamma-band activity and high-frequency oscillations have also been associated

TABLE II
COMPARISON OF CONTROL STRATEGIES FOR CLOSED-LOOP DBS

Approach	Principle	Advantage	Limitation / Risk
Threshold control	Fixed beta-power threshold triggers on/off stimulation	Simple, low power, easy to validate	Threshold drifts over time; single-feature
Supervised learning	Classifier (SVM, RF, shallow NN) trained on labeled LFP/kinematic states	Multivariate; higher classification accuracy	Requires labeled clinical data; may not generalize across patients
Reinforcement learning	Agent learns policy from reward signal (e.g., reduced beta, tremor score)	No pre-labeled data; adapts online	Exploration risk; requires constrained/safe action space

with therapeutic stimulation effects, and wearable inertial sensors and electromyography provide complementary peripheral measures of tremor, gait, and bradykinesia. Integrating these heterogeneous data streams into a single control signal is an active engineering challenge that is well suited to machine learning methods.

VI. MACHINE LEARNING IN THE LOOP

Threshold-based controllers, which switch stimulation on or off according to a fixed beta-power threshold, were an important first step but have clear limitations: the optimal threshold varies between patients and may drift as the disease progresses, and a single-feature controller discards most of the information present in the LFP and peripheral sensors. Machine learning approaches address both limitations by learning patient-specific, multivariate control policies directly from data.

Supervised learning algorithms, including support vector machines, random forests, and shallow neural networks, trained on combined LFP and kinematic features have demonstrated reliable classification of on/off medication states, bradykinesia severity, and dyskinesia presence. Once trained on clinical session data, such classifiers can in principle be deployed onboard an implantable pulse generator to drive real-time parameter selection.

The most ambitious approach applies reinforcement learning (RL) directly to stimulation parameter optimization: an agent explores the parameter space and receives a reward signal based on outcomes such as reduced beta power or improved accelerometer-derived tremor scores, learning a policy that maps brain state to optimal stimulation without requiring pre-labeled training data. The principal safety concern with RL in an implanted device is exploration, since an agent probing unfamiliar regions of the parameter space could transiently cause side effects; constrained RL formulations that restrict the agent to a pre-validated safe envelope are an active research direction, conceptually related to the ”safety layer” and ”shielding” approaches developed for safety-critical RL outside neuromodulation, where the policy's proposed action is projected back into a feasible set before being executed. Recent DBS-specific simulation work such as SEA-DBS [15] addresses a related concern, unstable exploration in the discrete action spaces typical of implantable hardware, using a predictive reward model to reduce reliance on live physiological feedback during learning; however, this and comparable approaches remain validated in biologically realistic simulation rather than in implanted human subjects. A clinically deployable system would also need an explicit fail-safe hierarchy: continuous monitoring of signal quality, automatic reversion to the clinician's last validated conventional settings whenever the policy proposes an out-of-envelope action, and a clinician override capable of halting autonomous adaptation at any time or of triggering fallback to safe conventional stimulation whenever sensing is unavailable or signal quality is poor. Typical supervised pipelines can extract spectral power, beta burst duration, entropy measures, and wearable-derived kinematic features.

TABLE III
COMPARISON OF MACHINE LEARNING ALGORITHMS FOR CLOSED-LOOP DBS

Algorithm	Input Data	Strengths	Limitation
Support Vector Machine (SVM)	LFP features IMU signals	High accuracy with small datasets; computationally efficient; robust to high-dimensional data	Limited adaptability; requires offline retraining when disease characteristics change
Random Forest (RF)	LFP, wearable sensors, clinical variables	Interpretable; resistant to overfitting; handles heterogeneous features	Larger memory footprint; limited online adaptation
Convolutional Neural Network (CNN)	Raw LFP time series, spectrograms	Automatic feature extraction; captures complex temporal and spectral patterns	Requires large datasets and greater computational resources; limited interpretability
Reinforcement Learning (RL)	Continuous neural feedback and symptom-related reward	Learns personalized stimulation policies; adapts online without labeled data	Exploration may induce unsafe stimulation; requires constrained learning and extensive validation

Models are generally evaluated using cross-validation, with performance reported through accuracy, sensitivity, specificity, F1-score, or area under the ROC curve.

Beyond algorithmic performance, several practical barriers still limit the widespread clinical adoption of machine learning-driven adaptive DBS. First, inter-patient variability remains substantial. Neural biomarkers such as beta oscillations differ in amplitude, spatial distribution, and temporal stability across individuals, making models trained on one population difficult to generalize without patient-specific calibration. Disease progression, medication adjustments, electrode migration, and long-term changes in neural tissue may further alter signal characteristics, requiring periodic model updates throughout the patient's lifetime.

Hardware limitations also impose significant constraints. Implantable pulse generators operate under strict power budgets, limited onboard memory, and low-latency requirements. Although deep learning models may achieve superior classification performance, their computational cost can exceed the capabilities of current implantable hardware. Consequently, lightweight classifiers such as SVMs or decision-tree ensembles remain attractive for near-term clinical deployment, while more complex models may rely on edge computing or hybrid architectures in which intensive processing is performed externally before transmitting optimized stimulation parameters to the implant.

From a regulatory perspective, adaptive algorithms introduce challenges that extend beyond conventional medical devices. Regulatory agencies increasingly require evidence not only of safety and efficacy but also of algorithm robustness, explainability, cybersecurity, and reliability under unexpected operating conditions. Machine learning models capable of updating themselves over time raise additional concerns regarding software validation, version control, and post-market surveillance. Future clinical systems will therefore likely combine interpretable AI methods with strict safety constraints, allowing autonomous adaptation while preserving clinician oversight and the ability to revert immediately to conventional continuous stimulation whenever abnormal system behavior is detected.

A recurring challenge across ML-based DBS approaches is the tension between personalization and generalization: models trained solely on one patient's data are precise but require extensive per-patient data collection, while population-level models are more readily deployable but miss individual variation. Transfer learning—pretraining on a large multi-patient dataset and fine-tuning on limited data from a new individual, is a promising strategy for addressing this tension.

Validation strategy strongly shapes how these accuracy figures should be read. Many supervised classifiers reviewed above are evaluated with random k-fold cross-validation, which allows samples from the same patient to appear in both training and test folds and can inflate apparent performance through within-subject correlation. Leave-one-subject-out (LOSO) cross-validation, in which an entire patient's data is withheld from training, provides a more realistic estimate

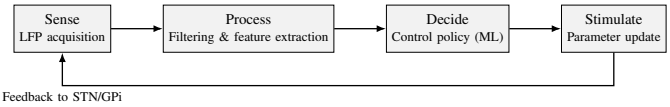


Fig. 1. Functional stages of a closed-loop DBS system. The stimulate stage feeds back into the sense stage, forming a continuous control loop operating at approximately 10–20 Hz.Reported closed-loop latencies vary with the control strategy: threshold-based beta-power controllers typically operate with control-loop update rates around 10–20 Hz (50–100 ms per cycle, as depicted in Fig. 1), while burst-aware controllers must additionally resolve individual beta bursts, which last roughly 100–200 ms under physiological conditions and 200–1,000 ms when pathological [17]. Across published closed-loop implementations, end-to-end latency between biomarker detection and stimulation adjustment has generally been reported below 400 ms, with some systems achieving under 800 ms [17]; this places an upper bound on how fast a burst the controller can meaningfully act on before it has already ended.

of how a model will perform on a new individual and is increasingly treated as the minimum standard for claims of generalizability in neural decoding [5]. External validation, testing a model on a dataset collected at a different center, with different hardware or recording montages, goes a step further and remains rare in the closed-loop DBS literature specifically, in contrast to adjacent fields such as EEG-based PD detection, where multi-site external validation is becoming more common. Transfer learning offers a practical middle ground: a population-level model is pretrained on a multi-patient dataset and then fine-tuned on a short calibration recording from the new individual, reducing the data-collection burden of a fully patient-specific model while retaining much of its accuracy. Merk et al. [**?**] formalize a set of good practices for training and testing in this context, including subject-wise data splitting and reporting of both within- and across-subject performance, which future ML-based aDBS studies, including the clinical closed-loop trials summarized in Table IV, should adopt to make claims of generalization verifiable. Notably, none of the ML classifiers discussed in Table III have yet been validated in closed-loop, in-device deployment at comparable scale to the threshold- and burst-based trials in Table IV; classifier work remains largely offline and retrospective.

VII. Clinical Evidence and Current Devices

Clinical validation of closed-loop DBS has accelerated substantially. In pivotal work by Little and colleagues, beta-triggered adaptive DBS reduced motor symptoms by approximately 50% relative to conventional continuous stimulation in a within-patient crossover study, while also reducing total stimulation energy by roughly 40%, demonstrating that adaptive stimulation can be simultaneously more effective and more efficient [4]. Arlotti and colleagues subsequently showed that eight hours of continuous adaptive DBS was safe, well tolerated, and produced consistent motor benefit under more clinically realistic conditions [9], and Herron and colleagues demonstrated a wearable closed-loop architecture using cortical recordings to trigger adaptive stimulation in an essential tremor patient, showing that the control loop need not be entirely implanted [10].

TABLE IV
KEY STUDIES UNDERPINNING CLOSED-LOOP AND ML-BASED DBS

Study	N (patients)	Signal(s)	Validation	Outcome / Performance	Compute / Device
Little *et al.* 2013 [4]	8 STN	Beta power, fixed threshold	Within-patient crossover (aDBS vs. cDBS vs. off)	∼50% greater motor improvement; ∼40% less stimulation energy vs. continuous	Threshold detector; externalized research platform
Tinkhauser *et al.* 2017 [6]	13 Patients (16 Hemispheres)	Beta burst duration/amplitude	Within-patient, off-medication	Burst amplitude scales with duration; long bursts track motor impairment	Offline analysis; externalized platform
Arlotti *et al.* 2018 [9]	13 STN	Beta power	Perioperative, 8 h continuous	Safe, well tolerated, consistent motor benefit	Externalized aDBS device
Herron *et al.* 2017 [10]	1 Patient (essential tremor)	Cortical ECoG + wearable IMU	Single-case	Feasibility of a partly-external closed loop	Wearable hub + cortical sensing
Oehrn *et al.* 2024 [16]	4 Patients	Stimulation-entrained gamma (STN or motor cortex)	Blinded randomized crossover	Personalized gamma-based aDBS outperformed conventional DBS	Chronic sensing-enabled IPG (Percept-class)
H. M. Bronte-Stewart et al. 2025 [14]	68 Patients	Single-/dual-threshold beta control	Multi-site nonrandomized trial	Feasibility of algorithm-driven single/dual-threshold control at scale	Embedded closed-loop IPG

The 2020 regulatory approval of the first commercially available bidirectional sensing-capable IPG, Medtronic's Percept PC, marked the transition of closed-loop DBS from research tool to clinical product, enabling gradual accumulation of patient-specific LFP datasets during normal use that can support the training and validation of personalized algorithms. DBS systems remain classified as high-risk implantable medical devices, and any autonomous algorithm that adjusts stimulation must demonstrate robust performance across failure modes, patient populations, and device interactions before clinical deployment a regulatory burden that is appropriate given the consequences of malfunction in a device implanted intracranially. Reviews of the closed-loop DBS field more broadly have emphasized that data streams from neural and wearable sensors do not fully capture a patient's moment-to-moment experience, arguing for hybrid systems that combine neural and wearable biomarkers with structured patient-reported outcomes, and for clinical decision-support tools that inform rather than replace clinician judgment [7], [8].

VIII. FUTURE DIRECTIONS AND CHALLENGES

Several converging developments are likely to shape the next generation of closed-loop DBS systems. Wearable integration continuous data from smartwatches, in-ear accelerometers, and smartphone-based gait analysis, can supplement implanted LFP sensing with a richer picture of daily motor function, updating patient-specific models on a rolling basis as the disease progresses. Architectures combining implanted sensing, edge processing on a wearable hub, and cloud-connected model updating, as proposed for epilepsy management [12], illustrate a general design pattern that is directly applicable to PD.

Miniaturization and power efficiency remain central engineering constraints, since every onboard computation draws from the same battery that powers stimulation; neuromorphic and ultra-low-power analog signal processing architectures may eventually allow more sophisticated onboard inference within existing power budgets. The therapeutic reach of closed-loop neuromodulation is also expanding beyond PD to essential tremor, dystonia, and, under active investigation, treatment-resistant depression, obsessive-compulsive disorder, and epilepsy [13]; the underlying engineering problem—identifying a reliable biomarker, designing a safe control policy, and validating it within a regulatory framework, is structurally similar across indications, making PD-focused closed-loop algorithms a methodological template for the broader field.

IX. CONCLUSION

Closed-loop DBS reframes Parkinson's disease as a control engineering problem, replacing static, clinician-programmed stimulation with therapy that continuously adapts to the patient's neurological state. Beta-band LFP activity remains the most validated feedback signal, but neuroimaging biomarkers, wearable sensing, and multimodal machine learning models are increasingly complementing it, moving the field from simple threshold controllers toward patient-specific, multivariate control policies. Realizing this potential at scale will depend as much on solving hardware, power, and regulatory constraints as on further algorithmic development, and on treating machine learning as a tool that supports, rather than replaces, clinical judgment. Ultimately, the future of adaptive DBS will likely rely on hybrid approaches combining patient-specific machine learning, multimodal biomarkers, and clinician-guided optimization, balancing personalization with safety, computational efficiency, and regulatory compliance.

REFERENCES

[1] A. L. Benabid, P. Pollak, C. Gervason, *et al.*, "Long-term suppression of tremor by chronic stimulation of the ventral intermediate thalamic nucleus," *Lancet*, vol. 337, no. 8738, pp. 403–406, 1991.

[2] S. Chiken and A. Nambu, "Mechanism of deep brain stimulation: inhibition, excitation, or disruption?," *Neuroscientist*, vol. 22, no. 3, pp. 313–322, 2016.

[3] C. Hammond, H. Bergman, and P. Brown, "Pathological synchronization in Parkinson's disease: networks, models and treatments," *Trends Neurosci.*, vol. 30, no. 7, pp. 357–364, 2007.

[4] S. Little, A. Pogosyan, S. Neal, *et al.*, "Adaptive deep brain stimulation in advanced Parkinson disease," *Ann. Neurol.*, vol. 74, no. 3, pp. 449–457, 2013.

[5] Merk, T., Peterson, V., Köhler, R., Haufe, S., Richardson, R.M., Neumann, W.-J. "Machine learning based brain signal decoding for intelligent adaptive deep brain stimulation." *Experimental Neurology*, 351:113993, 2022.

[6] G. Tinkhauser, A. Pogosyan, S. Little, *et al.*, "The modulatory effect of adaptive deep brain stimulation on beta bursts in Parkinson's disease," *Brain*, vol. 140, no. 4, pp. 1053–1067, 2017.

[7] W. Bouthour, P. Mégevand, J. Donoghue, *et al.*, "Biomarkers for closed-loop deep brain stimulation in Parkinson disease and beyond," *Nat. Rev. Neurol.*, vol. 15, no. 6, pp. 343–352, 2019.

[8] A. C. Meidahl, G. Tinkhauser, D. M. Herz, *et al.*, "Adaptive deep brain stimulation for movement disorders: the long road to clinical therapy," *Mov. Disord.*, vol. 32, no. 6, pp. 810–819, 2017.

[9] M. Arlotti, S. Marceglia, G. Foffani, *et al.*, "Eight-hours adaptive deep brain stimulation in patients with Parkinson disease," *Neurology*, vol. 90, no. 11, pp. e971–e976, 2018.

[10] J. A. Herron, M. C. Thompson, T. Brown, *et al.*, "Chronic electrocorticography for sensing movement intention and closed-loop deep brain stimulation with wearable sensors in an essential tremor patient," *J. Neurosurg.*, vol. 127, no. 3, pp. 580–587, 2017.

[11] P. Khanna, D. Totten, L. Novik, J. Roberts, R. J. Morecraft, and J. Bhattacharya, "Neurofeedback control in Parkinson's disease," *J. Neurol. Neurosurg. Psychiatry*, vol. 88, no. 8, pp. 744–754, 2017.

[12] V. Kremen, B. H. Brinkmann, I. Kim, *et al.*, "Integrating brain implants with local and distributed computing devices: a next generation epilepsy management system," *IEEE J. Transl. Eng. Health Med.*, vol. 6, p. 2500112, 2018.

[13] P. A. Starr, "Totally implantable bidirectional neural prostheses: a flexible platform for innovation in neuromodulation," *Front. Hum. Neurosci.*, vol. 12, p. 619, 2018.

[14] H. M. Bronte-Stewart, M. Beudel, J. L. Ostrem, *et al.*, "Long-term personalized adaptive deep brain stimulation in Parkinson disease: a nonrandomized clinical trial," *JAMA Neurol.*, vol. 82, no. 11, pp. 1171–1180, 2025.

[15] H. Ravivarapu, G. Bagwe, X. Yuan, C. Yu, and L. Zhang, "Sample-efficient reinforcement learning controller for deep brain stimulation in Parkinson's disease," arXiv:2507.06326, 2025.

[16] C. R. Oehrn, S. Cernera, L. H. Hammer, *et al.*, "Chronic adaptive deep brain stimulation versus conventional stimulation in Parkinson's disease: a blinded randomized feasibility trial," *Nat. Med.*, vol. 30, no. 11, pp. 3345–3356, 2024.

[17] S. Wang, G. Zhu, L. Shi, C. Zhang, B. Wu, A. Yang, F. Meng, Y. Jiang, and J. Zhang, "Closed-loop adaptive deep brain stimulation in Parkinson's disease: procedures to achieve it and future perspectives," J. Parkinsons Dis., vol. 13, no. 4, pp. 453–471, 2023.

Estimulação transcutânea auricular do nervo vago por corrente senoidal de 2 kHz e 100 Hz: estudo exploratório

1st Katherine Daniela Tacuri-Alvarez
Programa de Engenharia Biomédica
COPPE/UFRJ
Rio de Janeiro, Brasil
dntacuri@peb.ufrj.br

2nd Pablo Fernando Cevallos-Larrea
Universidad Politécnica Salesiana
(UPS) - GIIB
Cuenca, Ecuador
pcevallosl@ups.edu.ec

3rd Carlos Julio Tierra-Criollo
Programa de Engenharia Biomédica
COPPE/UFRJ
Rio de Janeiro, Brasil
carjulio@peb.ufrj.br

***Resumo*—O presente estudo teve como objetivo explorar as possíveis alterações nos índices da VFC durante a taVNS com corrente senoidal, utilizando 100 Hz (Condição 1) e 2 kHz (Condição 2). Após a exclusão dos registros com ruído significativo, foram analisados nove participantes saudáveis em cada condição, dos quais cinco participaram de ambos os protocolos. O eletrocardiograma foi registrado durante 15 minutos, divididos em períodos de pré-estimulação, estimulação e pós-estimulação. Foram calculados os índices SDNN, RMSSD, LF, HF e a razão LF/HF. As comparações temporais pareadas foram realizadas pelo teste de Wilcoxon, com correção de Bonferroni ($\alpha = 0{,}017$), e o tamanho do efeito foi calculado para cada comparação. Na condição de 100 Hz, o SDNN apresentou aumento durante a estimulação, com tamanho de efeito grande ($r = 0{,}64$), mas sem significância estatística. Na condição de 2 kHz, o RMSSD apresentou discreto aumento, também com tamanho de efeito grande ($r = 0{,}64$), enquanto a razão LF/HF apresentou redução durante a estimulação. Entretanto, essas alterações não foram estatisticamente significativas. A comparação exploratória entre as condições, realizada nos cinco participantes submetidos a ambos os protocolos, também não identificou diferenças significativas para SDNN, RMSSD ou LF/HF. Em relação à percepção sensorial, foram observados perfis distintos entre as condições, com predominância de descritores previamente associados a fibras finas em 100 Hz e a fibras grossas em 2 kHz, sem que esses achados permitam inferir recrutamento seletivo de fibras. Conclui-se que os resultados devem ser interpretados como exploratórios e podem contribuir para o planejamento de estudos futuros com um maior número de participantes, delineamento cruzado, condição placebo e maior padronização dos parâmetros de estimulação.**

***Palavras-chave*—neuromodulação autonômica, estimulação transcutânea auricular do nervo vago, corrente senoidal, variabilidade da frequência cardíaca**

I. Introdução

O nervo vago (NV), décimo par craniano, constitui uma importante via parassimpática de comunicação entre o sistema nervoso central e os órgãos viscerais, participando da regulação autonômica cardiovascular, gastrointestinal e imunológica. Por essa razão, tornou-se um alvo relevante para intervenções de neuromodulação [1]–[3]. Enquanto a estimulação vagal invasiva possui aplicações clínicas estabelecidas em condições como epilepsia refratária e depressão resistente ao tratamento [4], existe uma alternativa não invasiva destacada que é investigada para a abordagem de condições associadas à disfunção autonômica. A estimulação transcutânea auricular do nervo vago (*transcutaneous auricular vagus nerve stimulation* – taVNS). A taVNS se baseia na aplicação de estímulos elétricos em regiões da orelha que apresentam inervação total ou parcial pelo ramo auricular do nervo vago, como a concha, o trago, a fossa triangular e outras regiões auriculares [2]. Os protocolos de taVNS são geralmente configurados para estimular fibras grossas aferentes mielinizadas Aβ, cujos sinais podem ser transmitidos a estruturas do tronco encefálico, incluindo o núcleo do trato solitário, e alcançar circuitos envolvidos na regulação autonômica [2], [3].

Nesse contexto, a variabilidade da frequência cardíaca (VFC) tem sido empregada em estudos de taVNS para avaliar possíveis alterações na modulação autonômica cardíaca e inferir, de forma indireta, na modulação vagal [5], [6]. A VFC corresponde às variações temporais entre intervalos NN consecutivos, obtidos a partir dos intervalos RR após a identificação e a correção de artefatos e batimentos ectópicos, refletindo a influência dinâmica simpática e parassimpática [5], [7].

A VFC pode ser analisada nos domínios do tempo e da frequência [6], [7]. No domínio do tempo, o SDNN (*standard deviation of normal-to-normal intervals*) representa a variabilidade global dos intervalos NN durante o período analisado, enquanto o RMSSD (*root mean square of successive differences*) apresenta maior relação com a modulação parassimpática cardíaca [5]–[7]. No domínio da frequência, o componente de alta frequência (HF; 0,15–0,40 Hz) está predominantemente relacionado à modulação vagal associada à arritmia sinusal respiratória. O componente de baixa frequência (LF; 0,04–0,15 Hz), por sua vez, envolve mecanismos barorreflexos e influências simpáticas e parassimpáticas. Da mesma forma, a razão LF/HF deve ser interpretada com cautela, pois não representa diretamente o equilíbrio entre os ramos simpático e parassimpático [5]–[7].

Estudos que avaliaram os efeitos da taVNS sobre a modulação autonômica cardíaca apresentaram resultados het-

erogêneos. Alguns trabalhos relataram alterações em índices temporais e espectrais da VFC após a estimulação, sugerindo que a taVNS pode influenciar a regulação autonômica cardiovascular [8]–[12]. Enquanto estudos identificaram modificações em índices como SDNN ou RMSSD, outros não observaram diferenças significativas entre a estimulação ativa e o placebo [13]–[15].

Essas divergências podem estar relacionadas à elevada heterogeneidade metodológica dos protocolos, incluindo o local e a lateralidade da estimulação, a intensidade da corrente, a frequência, a largura de pulso, a duração da estimulação e a forma de onda utilizada [2], [14], [16]. A ausência de padronização desses parâmetros limita a comparação entre os estudos e dificulta a identificação das condições mais adequadas para produzir alterações consistentes na modulação autonômica.

A maior parte dos protocolos descritos na literatura utiliza pulsos retangulares ou bifásicos, enquanto os efeitos de formas de onda contínuas, como a senoidal, tem sido pouco explorados [2], [14]. A estimulação cervical do nervo vagal com uma senoide de 5 kHz tem sido investigada em pessoas sadias [17] e utilizada em investigação do COVID-19 [18]. Entretanto, ao nosso melhor conhecimento, não foram encontrados trabalhos com taVNS.

A estimulação elétrica senoidal caracteriza-se pela variação contínua da amplitude ao longo do tempo e pela presença de uma frequência fundamental definida. Estudos realizados em nervos cutâneos periféricos sugerem que a frequência da corrente senoidal pode influenciar o padrão de ativação das fibras sensoriais. Frequências elevadas, particularmente 2 kHz e 3 kHz, têm sido associadas a respostas predominantemente compatíveis com a ativação de fibras Aβ, que apresentam maior diâmetro e grau de mielinização. Frequências mais baixas, por outro lado, podem produzir respostas relacionadas a diferentes combinações de fibras Aβ, Aδ e C. Entretanto, essa possível neuroseletividade permanece controversa e pode variar conforme a intensidade do estímulo, o local de aplicação e as características individuais [19]–[21].

Martins et al. (2013) observaram que a estimulação senoidal entre 1 e 3 kHz produziu predominantemente descritores sensoriais e tempos de reação compatíveis com maior contribuição de fibra grossa mielinizadas Aβ. Por outro lado, os autores, em frequências entre 1 e 500 Hz obtiveram maiores descritores relacionados a fibra fina (Aδ e C), sendo o tempo de reação maior do que as altas frequências.

O presente estudo teve como objetivo investigar a taVNS com senoidal de frequências de 100 Hz e 2 kHz por meio das alterações dos índices da VFC.

II. Materiais e métodos

Participaram deste estudo 17 voluntários saudáveis, com idade entre 18 e 38 anos. Todos os participantes foram previamente informados sobre os procedimentos experimentais e assinaram o Termo de Consentimento Livre e Esclarecido conforme o comitê de Ética em Pesquisa aprovado nº 64138022.4.1001.5257. Os participantes foram orientados a evitar o consumo de café, bebidas energéticas e outras substâncias estimulantes 4 horas antes ao experimento. Não foram incluídos no estudo indivíduos que faziam uso de medicamentos que afetam o sistema nervoso e aqueles com histórico de doenças cardiovasculares. O participante foi acomodado em uma cadeira confortável, em um ambiente controlado 25 ± 1 °C , e orientado a permanecer em repouso durante 5 minutos. Inicialmente, as regiões de contato dos eletrodos descartáveis com a pele foram higienizadas com álcool. Para o registro do eletrocardiograma (ECG) foi utilizada uma configuração bipolar correspondente à derivação I com o eletrodo ativo foi posicionado abaixo da clavícula esquerda, o eletrodo referência abaixo da clavícula direita e o eletrodo terra na última costela esquerda. Os eletrodos para estimulação elétrica foram colocados na orelha esquerda na região da fossa triangular. Foi utilizado gel condutor para reduzir a impedância eletrodo–pele. Os sinais de ECG foram registrados, utilizando o sistema de aquisição de sinais MADQ [22], em duas condições experimentais:

- Na primeira condição participaram 11 indivíduos. Se utilizou um estímulo senoidal de 100 Hz, aplicado em ciclos compostos por 500 ms de estimulação (ON) seguidos por 800 ms sem estimulação (OFF), totalizando 1,3 s por ciclo [23], por meio de um eletrodo tipo spike (cátodo) e um eletrodo de ouro (ânodo). O sinal de ECG foi coletado a 1 kHz.
- Na segunda condição participaram 11 indivíduos, sendo que 5 também participaram da etapa anterior. Aqui, se adotou um estímulo senoidal de 2 kHz, aplicado em ciclos compostos por 30s de estimulação (ON) seguidos por 30 s sem estimulação (OFF) [23], por meio de dois eletrodos de ouro tipo disco (diâmetro de 5mm). O sinal de ECG foi coletado a 8 kHz.

Primeiro, foi determinado o limiar sensorial (LS) para cada participante [23]. A seguir, aplicou-se um procedimento para caracterizar a resposta cognitiva pela estimulação no LS. Para isso, foi solicitado ao participante realizar uma escolha forçada de três palavras dentre as oito palavras apresentadas ao azar associadas às sensações de ativação da fibra fina (calor, agulhada, beliscada, picada) e fibra grossa (contração, formigamento, vibração e pressão) [23]. A intensidade da estimulação foi ajustada 1,5 vezes acima do valor sensorial determinado individualmente, com ausência de desconforto.

Após a determinação do LS, foi concedido um período adicional de 5 minutos de repouso para o retorno ao estado basal. O registro dos sinais ECG ocorreu de maneira contínua durante 15 minutos, diferenciando três períodos denominados como: pré-estimulação (5 minutos sem estimulação); estimulação (5 minutos) e; pós-estimulação (5 minutos sem estimulação).

A. Processamento dos sinais de ECG

O processamento dos sinais de ECG foi realizado no software MATLAB. Inicialmente, foi removido o componente DC do sinal e aplicado um filtro notch em 60 Hz para atenuar interferências elétricas. O ECG foi filtrado em uma

banda de 5 Hz a 20 Hz [24]. A detecção dos picos R foi realizada utilizando o algoritmo de Pan-Tompkins, com um limiar mínimo correspondente a 0,3 vezes o desvio-padrão mais média do ECG e um limiar máximo correspondente a 4,3 desvios-padrão acima da média do ECG [25], [26]. Em seguida, foram calculados os intervalos RR para a obtenção do tacograma cardíaco. Para a remoção de artefatos e de possíveis batimentos ectópicos, foi utilizado um algoritmo baseado na mediana móvel, considerando uma janela de cinco amostras dos intervalos RR [25], [26]. Os intervalos que apresentaram desvio superior a 20% em relação à mediana local foram removidos. Posteriormente, os valores faltantes foram interpolados por meio de spline cúbica, gerando a série de intervalos NN (normal-to-normal). Para o monitoramento dos índices da VFC no domínio do tempo, SDNN e RMSSD, e da frequência, LF e HF, foram utilizadas janelas de 60 s com deslocamento de 30 s, resultando em uma sobreposição de 50%. Para a obtenção dos índices LF e HF, os intervalos NN foram interpolados, reamostrados a 10 Hz, e a densidade espectral de potência (*power spectral density* – PSD) foi estimada por meio de um modelo autorregressivo (AR) de ordem 13, utilizando o método de Yule-Walker [27].

B. Análise Estatística

Para a análise estatística dos índices da VFC, os intervalos NN foram agrupados em três janelas de 300 s (5 min.), correspondentes aos períodos experimentais de pré-estimulação (Janela 1), estimulação (Janela 2) e pós-estimulação (Janela 3). Foram obtidas estatísticas descritivas de média, mediana e quartis Q1-Q3 para cada período. Para a comparação entre as condições, os valores de cada índice foram transformados por meio do logaritmo natural, com o objetivo de reduzir a assimetria e a dispersão dos dados. Utilizou-se o teste de Wilcoxon para amostras pareadas, com aplicação da correção de Bonferroni às três comparações temporais realizadas para cada índice. Dessa forma, adotou-se um nível de significância ajustado de α=0,017. Além disso, calculou-se o tamanho do efeito para cada comparação por meio da expressão $r = Z/\sqrt{N}$, em que Z corresponde ao valor padronizado do teste de Wilcoxon e N ao número de pares analisados, permitindo avaliar a magnitude das diferenças observadas [29]. Para a interpretação da magnitude do tamanho do efeito, foram adotados os seguintes critérios: efeito pequeno para $0{,}10 \leq |r| < 0{,}30$, efeito moderado para $0{,}30 \leq |r| < 0{,}50$ e efeito grande para $|r| \geq 0{,}50$ [29].

Os resultados para 100 Hz e 2 kHz, dos cinco participantes que realizaram os experimentos para ambas as frequências, foram comparados. Para tanto, o valor durante a estimulação foi normalizado (em cada frequência) com o valor da pré-estimulação e o logaritmo foi aplicado. Foi utilizado o teste de Wilcoxon para amostras pareadas e calculado o tamanho do efeito. Devido ao limitado tamanho amostral, os resultados foram interpretados de forma exploratória.

III. Resultados

Os sinais de ECG de 2 participantes em cada condição apresentaram presença de ruído significativo causado pelo excessivo movimento dos participantes, sendo excluídos na apresentação de resultados. Assim o número de participantes em cada condição foi de n=9, sendo que 5 participaram de ambas as condições. A media do limiar de sensibilidade (LS) foi 89,71 ± 118,40 µA e 1445,45 ± 528,4 µA, para o estímulo de 100 Hz e 2 kHz, respectivamente. Em relação à sensação percebida pelo participante, a porcentagem do número de palavras relacionadas a fibra fina foi de 74% (20 de um total de 27) para o estímulo de 100 Hz, enquanto que para o estímulo de 2 kHz, a porcentagem para sensações de fibra grossa foi de 96% (26 de um total de 27).

A Figura 1 apresenta o sinal de ECG registrado ao longo do protocolo experimental de um participante ilustrativo (#10). A Figura 1a mostra a detecção dos picos da onda R utilizando o algoritmo de Pan-Tompkins. A Figura 1b apresenta o tacograma dos intervalos normal-normal (NN), obtido após a remoção de artefatos e de batimentos ectópicos, seguida da interpolação por spline cúbica.

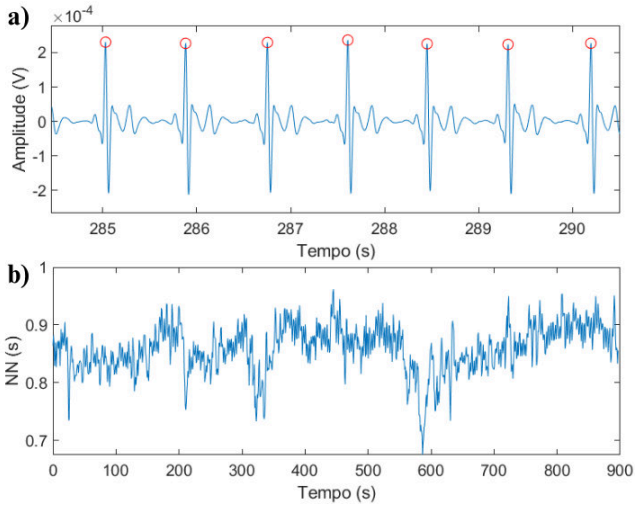


Fig. 1: a) Sinal de ECG e deteção de picos R, b) tacograma NN do participante #10 durante estimulação de 100 Hz

A Figura 2 apresenta a evolução temporal dos índices de variabilidade da frequência cardíaca de um participante ilustrativo (#10) durante o protocolo com estimulação de 100 Hz. A Figura 2a mostra a evolução do SDNN ao longo dos 15 minutos de registro (900 s). Nos primeiros 300 s, correspondentes ao período pré-estimulação, os valores permaneceram entre aproximadamente de 20 a 40 ms. Durante a estimulação (300–600 s), observou-se um aumento do SDNN no primeiro minuto e meio (300–400 s) e novamente no último minuto e meio (500–600 s). Após o término da estimulação, os valores retornaram gradualmente para níveis semelhantes aos observados no período basal.

Comportamento semelhante foi observado para o RMSSD (Figura 2b). Entretanto, nesse caso, os valores permaneceram

elevados durante a maior parte do período de estimulação, retornando progressivamente aos níveis basais após o término do estímulo.

A Figura 2c apresenta a evolução temporal da razão LF/HF. Durante o período pré-estimulação (30–300 s), a razão permaneceu relativamente estável. Após o início da estimulação, observou-se um aumento da razão LF/HF durante o primeiro minuto e meio (300–400 s), seguido de uma redução no intervalo entre 400 e 500 s. Nos 100 s finais da estimulação (500–600 s), a razão voltou a aumentar. Após o término da estimulação, observou-se uma redução gradual da razão LF/HF, aproximando-se dos valores observados no período basal.

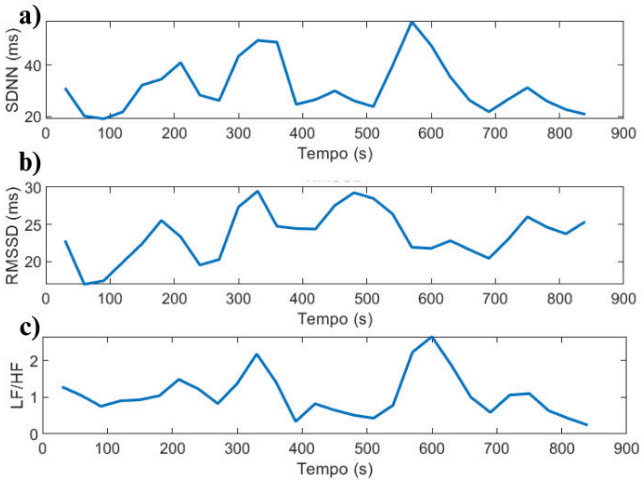


Fig. 2: Evoluçaõ temporal dos índices de VFC do participante #10 na frequência de 100 Hz. a) Índice SDNN b) Índice RMSSD c) Razão LF/HF

As Tabelas I e II apresentam, respectivamente, as estatísticas descritivas e os resultados das comparações pareadas dos índices da VFC entre as diferentes condições experimentais da condição 1 (100 Hz). Para o índice SDNN, os valores observados durante a estimulação apresentaram média e mediana superiores às registradas nos períodos de pré-estimulação e pós-estimulação. Essa diferença, entretanto, não atingiu significância estatística após a aplicação da correção de Bonferroni ($\alpha = 0{,}017$), embora tenha sido identificado tamanho de efeito grande (r = 0,64).

Para o RMSSD, as médias foram semelhantes entre as condições de pré-estimulação e estimulação, enquanto a mediana foi discretamente maior durante a estimulação. Apesar dessa variação descritiva, não foi observada diferença estatisticamente significativa entre as duas condições ($p > 0{,}017$).

No domínio da frequência, a razão LF/HF apresentou média e mediana mais elevadas durante a estimulação em comparação com o período de pré-estimulação. Contudo, essa diferença também não alcançou significância estatística após o ajuste para comparações múltiplas ($p > 0{,}017$).

Nas comparações complementares envolvendo o período de pós-estimulação, o SDNN apresentou diferença entre pré-estimulação e pós-estimulação antes da correção para comparações múltiplas (p = 0,02), acompanhada de tamanho de efeito grande (r = 0,87). Entretanto, esse resultado não permaneceu estatisticamente significativo após a correção de Bonferroni, uma vez que o valor de (p) foi superior ao nível de significância ajustado de (0,017).

Para os demais índices e comparações, não foram observadas diferenças estatisticamente significativas após a correção para comparações múltiplas. Ainda assim, algumas comparações apresentaram tamanhos de efeito de magnitude pequena a moderada.

TABELA I: Estatísticas descritivas dos índices de VFC nas diferentes condições experimentais– Frequência de 100 Hz.

Índice	**Pré-estimulação**	**Estimulação**	**Pós-estimulação**
	Média (DP) Mediana (Q1–Q3)	Média (DP) Mediana (Q1–Q3)	Média (DP) Mediana (Q1–Q3)
SDNN	37,5 ± 10,12 38,9 (31,5–41,2)	44,2 ± 12,7 43,9 (31,9–51,1)	40,9 ± 10,92 41,1 (38,8–48,8)
RMSSD	28,1 ± 12,1 22,1 (18,6–37,3)	28,0 ± 12,26 26,47 (17,8–35,4)	29,1 ± 10,8 25,0 (22,4–40,0)
LF	450,4 ± 222,5 489,1 (246,5–538,2)	472,2 ± 22671 390,3 (347,7–500,0)	561,4 ± 344,9 503,0 (284,3–822,3)
HF	338,3,4 ± 232,0 261,8 (181,2–557,1)	339,6 ± 230,2 263,4 (155,8–534,0)	368,7 ± 209,1 328,8 (205,6–527,5)
LF/HF	1,7 ± 0,9 1,5 (1,3–2,3)	1,81 ± 0,8 2,2 (1,4–2,4)	1,6 ± 0,7 1,7 (1,1–2,5)

TABELA II: Comparação dos índices de variabilidade da frequência cardíaca entre os momentos pré-estimulação (Pré), durante a estimulação (Est) e pós-estimulação (Pós), apresentando os valores de significância estatística (p) e o tamanho do efeito (r). Frequência 100 Hz.

	Pré vs Est		Pré vs Pós		Est vs Pós	
Índice	p	r	p	r	p	r
Log(SDNN)	0,10	0,64	0,02	0,87	0,43	0,33
Log(RMSSD)	1,00	0,02	0,13	0,60	0,30	0,42
Log(LF)	0,50	0,29	0,10	0,64	0,65	0,20
Log(HF)	1,00	0,02	0,10	0,64	0,30	0,42
Log(LF/HF)	0,36	0,38	1,00	0,02	0,65	0,20

A Figura 3 ilustra a evolução temporal dos índices de VFC de um participante ilustrativo (Participante #2) durante a estimulação com onda senoidal de 2 kHz. Na Figura 3a, o índice SDNN apresenta um comportamento variável ao longo do tempo. Durante a estimulação, os valores da SDNN permaneceram predominantemente entre 20 e 40 ms. Entre aproximadamente 400 e 500 s, verificou-se uma redução do SDNN, seguida por um aumento, mantendo-se posteriormente relativamente estável na faixa de 20 a 35 ms.A Figura 3b apresenta a evolução temporal do índice RMSSD. Entre aproximadamente 30 e 400 s, os valores permaneceram relativamente estáveis. A partir de cerca de 500 s, observou-se um aumento do RMSSD, com um pico mais pronunciado nos minutos finais do registro, correspondentes ao período de estimulação e início

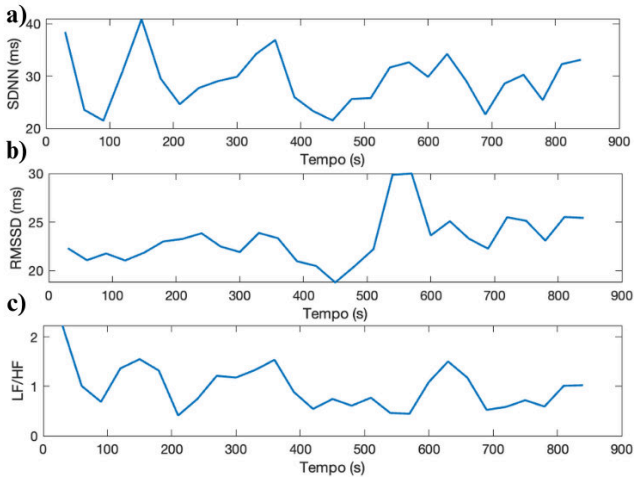


Fig. 3: Evoluçaõ temporal dos índices de VFC do participante #2 na frequência de 2 kHz a) Índice SDNN b) Índice RMSSD c) Razão LF/HF

do período pós-estimulação. Na Figura 3c, referente à razão LF/HF, observam-se valores relativamente estáveis durante os primeiros minutos (30–400 s). Entre aproximadamente 500 e 600 s, período em que o RMSSD apresentou aumento, verificou-se uma redução da razão LF/HF, seguida de um retorno a valores semelhantes aos observados no início do registro.

As Tabelas III e IV apresentam, respectivamente, as estatísticas descritivas e os resultados das comparações pareadas dos índices da VFC entre as diferentes condições experimentais da condição 2 (frequência 2 kHz).

O índice SDNN apresentou redução dos valores médios e medianos durante a estimulação em relação ao período de pré-estimulação. Entretanto, essa diferença não atingiu significância estatística após a aplicação da correção de Bonferroni ($\alpha = 0{,}017$), embora tenha apresentado tamanho de efeito moderado (r = 0,38).

Para o índice RMSSD, verificou-se aumento descritivo da média e da mediana durante a estimulação em comparação com o período de pré-estimulação. Apesar dessa variação descritiva, não foi observada diferença estatisticamente significativa entre as condições ($p > 0{,}017$).

Em relação à razão LF/HF, tanto a média quanto a mediana diminuíram durante a estimulação quando comparadas ao período de pré-estimulação. Contudo, essa redução também não alcançou significância estatística após o ajuste para comparações múltiplas ($p > 0{,}017$).

De modo geral, não foram observadas diferenças estatisticamente significativas entre os períodos de pré-estimulação, estimulação e pós-estimulação para nenhum dos índices analisados. Ainda assim, a análise dos tamanhos de efeito indicou magnitudes grandes em algumas comparações.

Na comparação entre estimulação e pós-estimulação, os índices LF e LF/HF apresentaram tamanhos de efeito grandes (r = 0,64 e r = 0,56, respectivamente), embora sem significância estatística.

TABELA III: Estatísticas descritivas dos índices de VFC nas diferentes condições experimentais– Frequência 2 kHz.

Índice	**Pré-estimulação**	**Estimulação**	**Pós-estimulação**
	Média (DP) Mediana (Q1–Q3)	Média (DP) Mediana (Q1–Q3)	Média (DP) Mediana (Q1–Q3)
SDNN	55,3± 22,5 59,2 (34,2–69,5)	54,7 ± 28,7 51,9 (29,7–68,7)	52,4 ± 23,7 46,1 (32,8–67,2)
RMSSD	36,0 ± 24,5 25,7 (20,0–55,0)	36,7 ± 24,8 26,1 (20,0–47,1)	35,4 ± 23,0 24,1(17,0–51,2)
LF	1044 ± 690,9 1111 (379,3–1233)	1046 ± 857,3 961,6 (235,0–1947)	1162± 924,1 725,0 (361,0–1951)
HF	844,7 ± 959,8 365,6 (218,0-1206)	883,7 ± 1103 322,5(212,3–1128)	761,7 ± 904,7 302,4 (167,1–930,1)
LF/HF	3,0 ± 3,1 2,0 (1,4–2,8)	1,7 ± 0,8 1,7 (1,1–1,9)	2,6 ± 2,0 2,4 (1,7–2,5)

TABELA IV: Comparação pareada dos índices da VFC entre os momentos pré-estimulação (Pré), durante a estimulação (Est) e pós-estimulação (Pós), apresentando os valores de significância estatística (p) e o tamanho do efeito (r) Condição 2.

	Pré vs Est		Pré vs Pós		Est vs Pós	
Índice	p	r	p	r	p	r
Log(SDNN)	0,36	0,38	0,43	0,33	1,00	0,02
Log(RMSSD)	0,10	0,64	0,82	0,11	0,64	0,20
Log(LF)	0,57	0,24	0,82	0,11	0,10	0,64
Log(HF)	0,36	0,38	0,91	0,07	0,50	0,29
Log(LF/HF)	0,50	0,29	0,91	0,07	0,16	0,56

TABELA V: Comparação dos índices da VFC entre as condições de 100 Hz e 2 kHz nos cinco participantes avaliados em ambas as condições, apresentando os valores de significância estatística (p) e o tamanho do efeito (r)

Índice	p	r
SDNN	0,18	0,73
RMSSD	0,79	0,20
LF/HF	0,28	0,60

A Tabela V apresenta a comparação dos índices de variabilidade da frequência cardíaca (VFC) entre as Condições 1 e 2 nos cinco participantes avaliados em ambas as condições. Não foram encontradas diferenças estatisticamente significativas entre as condições para os índices SDNN (p = 0,18), RMSSD (p = 0,79) e LF/HF (p = 0,28), considerando o nível de significância ajustado pela correção de Bonferroni ($\alpha = 0{,}017$). Assim, os resultados não forneceram evidências estatísticas de diferenças entre as Condições 1 e 2. Por outro lado, no tamanho do efeito, foram observados efeitos grandes para os índices SDNN (r = 0,73) e LF/HF (r = 0,60), enquanto o RMSSD apresentou tamanho de efeito pequeno (r = 0,20).

IV. Discussão e Conclusão

Neste estudo, avaliaram-se as respostas da variabilidade da frequência cardíaca (VFC) durante duas condições de estimulação transcutânea auricular do nervo vago com corrente senoidal, utilizando frequências de 100 Hz e 2 kHz. Participaram nove indivíduos em cada condição, dos quais cinco participaram de ambas as condições. De modo geral, não foram encontradas diferenças estatisticamente significativas entre os períodos de pré-estimulação, estimulação e pós-estimulação para nenhum dos índices analisados, considerando o nível de significância ajustado pela correção de Bonferroni . Entretanto, foram observadas alterações descritivas e tamanhos de efeito moderados ou grandes em algumas comparações.Considerando o reduzido tamanho amostral e a ausência de significância estatística, esses resultados devem ser interpretados como achados exploratórios. As condições de estimulação estudadas apresentaram perfis distintos de percepção sensorial. O limiar sensorial foi menor no protocolo de 100 Hz (89,71±118,40 μA) e maior no protocolo de 2 kHz (1445,45±528,4 μA). Na estimulação de 100 Hz, 74% dos descritores selecionados pelos participantes estavam relacionados a sensações previamente associadas a fibras finas, enquanto, em 2 kHz, 96% estavam relacionados a sensações associadas a fibras grossas. Entretanto, os descritores sensoriais são respostas subjetivas e não permitem identificar diretamente quais fibras nervosas foram recrutadas.

Na condição 1, o aumento descritivo do SDNN observado durante o protocolo de 100 Hz indica uma alteração descritiva da variabilidade cardíaca global, mas não necessariamente um aumento específico da modulação vagal, uma vez que não foi acompanhado por alteração significativa do RMSSD. Esse resultado apresenta semelhanças com os achados de Atanackov et al. (2025), que observaram aumento do SDNN em determinadas combinações de frequência e largura de pulso, sem modificações no RMSSD [12]. De Couck et al. (2017) também identificaram alterações principalmente no SDNN, dependentes da duração da estimulação, da lateralidade e do sexo dos participantes [9]. De forma semelhante, Gharabaghi e Keute (2025) encontraram aumento do SDNN em algumas condições experimentais, sem mudanças correspondentes no RMSSD [15].

Deve-se mencionar que a análise considerou apenas os valores médios da SDNN em cada período de estimulação. Nesse sentido, seria interessante avaliar também a dinâmica temporal desse parâmetro, especialmente por meio da comparação entre os 100 segundos iniciais e os 100 segundos finais do período de estimulação,pois podem ocorrer alterações transitórias nos índices da VFC ao longo do período de estimulação. As Figuras 2 e 3 sugerem visualmente essa possibilidade, que deverá ser investigada posteriormente.

Na condição 2, com estimulação senoidal de 2 kHz, o SDNN apresentou redução da média e da mediana durante a estimulação, indicando que não houve aumento da variabilidade global mensurada pelo SDNN. Por outro lado, o RMSSD apresentou discreto aumento e tamanho de efeito grande na comparação com o período de pré-estimulação ($r = 0{,}64$). A razão LF/HF também apresentou redução da média e da mediana. Essas alterações não permitem inferir aumento da modulação vagal ou redução da atividade simpática, uma vez que não permaneceram estatisticamente significativas após a correção para múltiplas comparações, e a razão LF/HF não constitui uma medida direta do equilíbrio simpatovagal.

Os resultados da condição 2 podem ser relacionados aos achados de Borges, Laborde e Raab (2019), que observaram aumento do RMSSD ao longo do tempo, mas sem diferenças entre a estimulação ativa e o placebo. Dessa forma, os autores não puderam atribuir a alteração especificamente à estimulação vagal [13]. Machetanz et al. (2021) também observaram que as alterações no SDNN e no RMSSD dependiam do local de aplicação e dos parâmetros de estimulação utilizados [11].

A redução descritiva da razão LF/HF observada na condição de 2 kHz apresentou direção semelhante à relatada por Clancy et al. (2014). Utilizando taVNS com pulsos de 30 Hz, os autores observaram redução significativa da razão LF/HF e da atividade nervosa simpática muscular, embora as potências LF e HF, analisadas separadamente, não tenham apresentado alterações significativas [28]. Porém, no presente estudo, a ausência de significância estatística para a razão LF/HF e de uma medida direta da atividade nervosa autonômica permite apenas reconhecer uma semelhança descritiva entre os resultados, sem confirmar a ocorrência do mesmo mecanismo fisiológico.

Na comparação entre as duas condições, realizada com os cinco participantes submetidos a ambos os protocolos, não foram encontradas diferenças estatisticamente significativas para SDNN ($p = 0{,}18$), RMSSD ($p = 0{,}79$) e LF/HF ($p = 0{,}28$). Quanto ao tamanho do efeito, foram observados efeitos grandes para os índices SDNN ($r = 0{,}73$) e LF/HF ($r = 0{,}60$), enquanto o RMSSD apresentou tamanho de efeito pequeno ($r = 0{,}20$). Esses resultados devem ser interpretados de forma exploratória, devido ao número reduzido de participantes avaliados em ambas as condições. Os tamanhos de efeito obtidos podem ser utilizados como informação preliminar para auxiliar no planejamento de estudos posteriores, mas não devem ser interpretados, isoladamente, como evidência de diferenças fisiológicas entre as frequências avaliadas.

Além disso, as condições diferiram não apenas quanto à frequência da corrente, mas também quanto ao tipo e à configuração dos eletrodos, ao ciclo de estimulação e à frequência de amostragem do ECG. Portanto, as diferenças observadas entre as condições não podem ser atribuídas isoladamente às frequências de 100 Hz ou 2 kHz. Para investigar especificamente o efeito da frequência da corrente senoidal, estudos posteriores deverão empregar um delineamento cruzado, no qual os mesmos participantes sejam submetidos às duas frequências, mantendo constantes o tipo e a posição dos eletrodos, a duração e o ciclo de estimulação e os parâmetros de aquisição do ECG. Recomenda-se também a randomização da ordem das sessões para reduzir possíveis efeitos relacionados à sequência das condições experimentais.

A ausência de diferenças estatisticamente significativas é

compatível com a heterogeneidade dos resultados encontrados na literatura. Yap et al. (2020) verificaram que alguns estudos relataram aumento de indicadores relacionados à modulação parassimpática, enquanto outros não encontraram diferenças entre a estimulação ativa e o placebo [14]. Farmer et al. (2021) destacaram que a falta de padronização dos parâmetros de estimulação e dos métodos de aquisição e análise dos sinais limita a comparação entre os estudos e dificulta a identificação de respostas autonômicas reprodutíveis [16].

Algumas limitações metodológicas devem ser consideradas na interpretação dos resultados. A exclusão de quatro registros (18,2%) devido à presença de ruídos reduziu a amostra disponível para nove participantes em cada condição. Além disso, a comparação entre os protocolos foi realizada com apenas cinco participantes que participaram de ambas as condições. A quantidade de registros excluídos mostra a importância de se obter sinais de ECG com boa qualidade, especialmente em ambientes menos controlados como na clínica. Outra limitação importante é a ausência de uma condição placebo, o que dificulta distinguir possíveis efeitos específicos da estimulação. Além disso, antes da aquisição do período considerado basal, os participantes foram submetidos à estimulação elétrica durante o procedimento de determinação do limiar sensorial. Dessa forma, não é possível excluir a possibilidade de que essa estimulação prévia tenha produzido efeitos autonômicos residuais e influenciado os valores utilizados posteriormente como referência. Estudos futuros deverão incluir uma condição placebo e estabelecer um período de recuperação entre a determinação do limiar sensorial e a aquisição do período basal, de modo a minimizar possíveis efeitos residuais da estimulação prévia.

Em conclusão, nas condições avaliadas neste estudo exploratório, não foram detectadas alterações estatisticamente significativas nos índices de VFC. Porém, foram identificadas alterações descritivas e tamanhos de efeito moderados e grandes que podem ser um indicativo de possíveis alterações da VFC. Estudos futuros, com maior número de participantes, delineamento cruzado, randomização da ordem das sessões, condição placebo e padronização dos parâmetros de estimulação, são necessários para investigar de forma mais robusta os possíveis alterações autonômicas da estimulação senoidal.

Agradecimentos

As agências financiadoras CAPES (PROEX), CNPq, FAPERJ (E-26/204.393/2024) e FINEP (Convênio 01.24.0122.00).

Referências

[1] H.-R. Berthoud and W. L. Neuhuber, "Functional and chemical anatomy of the afferent vagal system," *Auton. Neurosci.*, vol. 85, nos. 1–3, pp. 1–17, 2000, doi: 10.1016/S1566-0702(00)00215-0.

[2] E. Kaniusas, S. Kampusch, M. Tittgemeyer, et al., "Current directions in the auricular vagus nerve stimulation I—A physiological perspective," *Front. Neurosci.*, vol. 13, Art. no. 854, 2019, doi: 10.3389/fnins.2019.00854.

[3] H. Yuan and S. D. Silberstein, "Vagus nerve and vagus nerve stimulation, a comprehensive review: Part I," *Headache*, vol. 56, no. 1, pp. 71–78, 2016, doi: 10.1111/head.12647.

[4] C. W. Austelle, S. S. Cox, K. E. Wills, and B. W. Badran, "Vagus nerve stimulation (VNS): Recent advances and future directions," *Clin. Auton. Res.*, vol. 34, no. 6, pp. 529–547, 2024, doi: 10.1007/s10286-024-01065-w.

[5] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research: Recommendations for experiment planning, data analysis, and data reporting," *Front. Psychol.*, vol. 8, Art. no. 213, 2017, doi: 10.3389/fpsyg.2017.00213.

[6] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Front. Public Health*, vol. 5, Art. no. 258, 2017, doi: 10.3389/fpubh.2017.00258.

[7] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996, doi: 10.1161/01.CIR.93.5.1043.

[8] V. Wolf, A. Kühnel, V. Teckentrup, J. Koenig, and N. B. Kroemer, "Does transcutaneous auricular vagus nerve stimulation affect vagally mediated heart rate variability? A living and interactive Bayesian meta-analysis," *Psychophysiology*, vol. 58, no. 11, Art. no. e13933, 2021, doi: 10.1111/psyp.13933.

[9] M. De Couck et al., "Effects of short and prolonged transcutaneous vagus nerve stimulation on heart rate variability in healthy subjects," *Auton. Neurosci.*, vol. 203, pp. 88–96, 2017, doi: 10.1016/j.autneu.2016.11.003.

[10] M. K. Jensen, S. S. Andersen, S. S. Andersen, C. H. Liboriussen, S. Kristensen, and M. Jochumsen, "Modulating heart rate variability through deep breathing exercises and transcutaneous auricular vagus nerve stimulation: A study in healthy participants and in patients with rheumatoid arthritis or systemic lupus erythematosus," *Sensors*, vol. 22, no. 20, Art. no. 7884, 2022, doi: 10.3390/s22207884.

[11] K. Machetanz, L. Berelidze, R. Guggenberger, and A. Gharabaghi, "Transcutaneous auricular vagus nerve stimulation and heart rate variability: Analysis of parameters and targets," *Auton. Neurosci.*, vol. 236, Art. no. 102894, 2021, doi: 10.1016/j.autneu.2021.102894.

[12] P. Atanackov, J. Peterlin, M. Derlink, U. Kovačič, N. Kejžar, and F. F. Bajrović, "The acute effects of varying frequency and pulse width of transcutaneous auricular vagus nerve stimulation on heart rate variability in healthy adults: A randomized crossover controlled trial," *Biomedicines*, vol. 13, no. 3, Art. no. 700, 2025, doi: 10.3390/biomedicines13030700.

[13] U. Borges, S. Laborde, and M. Raab, "Influence of transcutaneous vagus nerve stimulation on cardiac vagal activity: Not different from sham stimulation and no effect of stimulation intensity," *PLoS ONE*, vol. 14, no. 10, Art. no. e0223848, 2019, doi: 10.1371/journal.pone.0223848.

[14] J. Y. Y. Yap, C. Keatch, E. Lambert, W. Woods, P. R. Stoddart, and T. Kameneva, "Critical review of transcutaneous vagus nerve stimulation: Challenges for translation to clinical practice," *Front. Neurosci.*, vol. 14, Art. no. 284, 2020, doi: 10.3389/fnins.2020.00284.

[15] A. Gharabaghi and M. Keute, "Effects of stimulation site and protocol on autonomic responses to auricular vagus nerve stimulation," *Front. Neurosci.*, vol. 19, Art. no. 1692530, 2025, doi: 10.3389/fnins.2025.1692530.

[16] A. D. Farmer et al., "International consensus based review and recommendations for minimum reporting standards in research on transcutaneous vagus nerve stimulation (Version 2020)," *Front. Hum. Neurosci.*, vol. 14, Art. no. 568051, 2021, doi: 10.3389/fnhum.2020.568051.

[17] R. Nonis, K. D'Ostilio, J. Schoenen, and D. Magis, "Evidence of activation of vagal afferents by non-invasive vagus nerve stimulation: An electrophysiological study in healthy volunteers," *Cephalalgia*, vol. 37, no. 13, pp. 1285–1293, 2017, doi: 10.1177/0333102417717470.

[18] C. Tornero et al., "Non-invasive vagus nerve stimulation for COVID-19: Results from a randomized controlled trial (SAVIOR I)," *Front. Neurol.*, vol. 13, Art. no. 820864, 2022, doi: 10.3389/fneur.2022.820864.

[19] H. R. Martins, R. Zanetti, C. C. dos Santos, G. M. Manzano, and C. J. Tierra-Criollo, "Current perception threshold and reaction time in the assessment of sensory peripheral nerve fibers through sinusoidal electrical stimulation at different frequencies," *Rev. Bras. Eng. Bioméd.*, vol. 29, no. 3, pp. 278–285, 2013, doi: 10.4322/rbeb.2013.028.

[20] E. P. V. Félix et al., "Sensations and reaction times evoked by electrical sinusoidal stimulation," *Neurophysiol. Clin.*, vol. 39, no. 6, pp. 283–290, 2009, doi: 10.1016/j.neucli.2009.10.001.

[21] M. Langille, J. A. Gonzalez-Cueto, and S. Sundar, "Analysis of the selective nature of sensory nerve stimulation using different sinusoidal

frequencies," *Int. J. Neurosci.*, vol. 118, no. 8, pp. 1131–1144, 2008, doi: 10.1080/00207450701769323.

[22] P. Cevallos-Larrea, L. Guambaña-Calle, D. A. Molina-Vidal, M. Castillo-Guerrero, A. d'Affonsêca Netto, and C. J. Tierra-Criollo, "Prototype of a multimodal and multichannel electro-physiological and general-purpose signal capture system: Evaluation in a sleep-research-like scenario," *Sensors*, vol. 25, no. 9, Art. no. 2816, 2025, doi: 10.3390/s25092816.

[23] D. A. Molina-Vidal, P. Cevallos-Larrea, L. Guambaña-Calle, D. Liquori, and C. J. Tierra-Criollo, "Design and evaluation of a prototype of dual channel electrical stimulator: Application in the assessment of current perception thresholds to multiple stimuli," in *Smart Technologies, Systems and Applications*, ser. Communications in Computer and Information Science, vol. 1705, F. R. Narváez, F. Urgilés, T. F. Bastos-Filho, and J. P. Salgado-Guerrero, Eds. Cham, Switzerland: Springer, 2023, pp. 342–355, doi: 10.1007/978-3-031-32213-6_25.

[24] T. Curiel-Gurza and E. Bojorges-Valdez, "HRV connectivity metrics as indicators of interpersonal agreement in dialogue," in *Proc. Computing in Cardiology (CinC)*, vol. 52, pp. 1–4, 2025, doi: 10.22489/CinC.2025.229.

[25] B. Mali, S. Žulj, R. Magjarević, D. Miklavčič, and T. Jarm, "Matlab-based tool for ECG and HRV analysis," *Biomed. Signal Process. Control*, vol. 10, pp. 108–116, 2014, doi: 10.1016/j.bspc.2014.01.011.

[26] M. A. Peltola, "Role of editing of R-R intervals in the analysis of heart rate variability," *Front. Physiol.*, vol. 3, Art. no. 148, 2012, doi: 10.3389/fphys.2012.00148.

[27] R. Balocchi, A. Macerata, C. Carpeggiani, M. Emdin, A. Benassi, and A. L'Abbate, "A spectral approach for heart rate fluctuations analysis," *Math. Comput. Model.*, vol. 11, pp. 1078–1081, 1988, doi: 10.1016/0895-7177(88)90170-7.

[28] J. A. Clancy, D. A. Mary, K. K. Witte, J. P. Greenwood, S. A. Deuchars, and J. Deuchars, "Non-invasive vagus nerve stimulation in healthy humans reduces sympathetic nerve activity," *Brain Stimul.*, vol. 7, no. 6, pp. 871–877, 2014, doi: 10.1016/j.brs.2014.07.031.

[29] F. F. Peres, "Effect sizes for nonparametric tests," *Biochem. Med. (Zagreb)*, vol. 36, no. 1, pp. 5–16, 2026, doi: 10.11613/BM.2026.010101.

Evaluation of TENS-Evoked Perception for the Restoration of Somatosensory Feedback: A Case Study

Leandra Lima de Almada
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-3655-8850

Victor Henrique Bértoli da Silva
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-4009-6177

Ana Clara Pereira Resende da Costa
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-5533-8880

Alcimar Barbosa Soares
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-1100-3533

***Abstract*—Upper limb amputation is an event that brings major life changes. The use of a prosthesis can help restore some of the function lost, but after the fitting process and training to use it, the lack of somatosensory feedback often leads to its abandonment. Transcutaneous Electrical Nerve Stimulation (TENS) emerges as a promising non-invasive technique to restore this lost sensory information. This case study aims to evaluate the somatosensory perception evoked by TENS in two volunteers with upper limb amputation, comparing the results with data previously collected from healthy subjects. Biphasic balanced rectangular pulses were applied using a custom-developed electrical stimulator. The protocol consisted of identifying the sensation and pain thresholds, followed by incremental and decremental frequency modulation at fixed current amplitudes and pulse width to assess intensity perception, frequency discrimination, and sensory descriptors. Results indicated that the amputees' sensation threshold was similar to that of healthy volunteers, while the pain threshold was higher. Furthermore, the perceived intensity exhibited a logarithmic behavior relative to the stimulation frequency, as expected. Frequencies below 50 Hz were predominantly described as a pulsing sensation. These findings suggest that TENS can evoke distinguishable and comfortable sensations, providing fundamental parameters for optimal stimulus modulation in sensory feedback restoration for upper limb prostheses.**

***Index Terms*—TENS, electrotactile feedback, amputee, stimulus modulation**

I. Introduction

Upper limb amputation results not only in loss of motor capacity but also in loss of sensory abilities important for daily life, drastically impacting the quality of life of the amputee. The rehabilitation process for the use of prostheses is extensive and requires extremely high physical and mental effort. However, the abandonment rates for prostheses remain high, and one of the main factors for this rejection is the absence of sensory feedback [1]. Without tactile feedback, such as naturally provided by the skin's mechanoreceptors, the user becomes highly dependent on visual feedback to monitor object manipulation. This generates a high cognitive load and results in movements that are less natural and more difficult to control. Furthermore, the inability to feel touch impairs the subject's interaction with their environment, which destroys the chance to develop a sense of ownership of the artificial limb to the body. These aspects are pointed out as fundamental for both motor skills and psychological well-being [2].

Given this scenario, the restoration of sensory feedback in amputees becomes a crucial demand for greater adherence and improved usability of upper limb prostheses. By closing the control loop and providing tactile information to the user, manipulation becomes more intuitive [3], [4]. More importantly, continuous sensory feedback directly acts on brain plasticity, facilitating the incorporation of the prosthesis into the individual's body schema (embodiment) and making its use more natural [5].

The skin mechanoreceptors are responsible for providing tactile signals to the nervous system. They are associated with two primary types of afferent fibers: rapidly adapting (RA) and slowly adapting (SA) fibers. RA fibers fire intensely at the onset and offset of a stimulus. In contrast, SA fibers discharge more moderately and continuously throughout the stimulus duration. This mechanism ensures that the brain remains constantly aware of sensory events without becoming cognitively overloaded. For instance, this is why humans can ignore the constant sensation of wearing clothes, yet distinctly feel them when they direct attention to the area [6].

Therefore, to successfully restore sensory feedback in prostheses, a stimulus modulation strategy is necessary to keep the user constantly aware of tactile interactions. In [7], the authors successfully modeled skin mechanoreceptors, enabling the identification of stimulus location on an artificial skin with a resolution approximating that of biological skin. This

was achieved using spiking patterns that mimic real neuronal behavior through the Izhikevich model. The next step involves translating the data captured by these artificial sensors into electrical stimuli, thereby delivering a sensation to the individual that is as close to physiological as possible.

Several approaches have been proposed to restore the lost sensitivity; among them, Transcutaneous Electrical Nerve Stimulation (TENS), a technique with already widely consolidated use in physiotherapeutic practice for pain treatment (analgesia) [8], stands out as a promising non-invasive alternative [9]. By properly modulating the electrical parameters of the stimulus, TENS has the capacity to activate preserved afferent nerve fibers, evoking artificial tactile sensations that can restore lost sensory feedback, whether due to amputation or neurodegenerative diseases [10].

Despite its potential, the challenge lies in mapping the optimal stimulation parameters that are comfortable and generate sensations easily interpreted by the brain [11], [12]. Therefore, this case study aims to evaluate the stimulation parameters that optimize the perception evoked, without causing discomfort, in an arm amputee.

The main contributions of this work are: the evaluation of TENS-evoked somatosensory perception in upper limb amputees; a direct comparison with previously collected able-bodied data using an identical protocol; and empirical evidence on comfortable stimulation parameters for closed-loop sensory feedback design.

II. Methodology

The research of which this work is part was approved by the Ethics Committee of the Federal University of Uberlândia (CAAE: 82612724.4.0000.5152).

A. Materials

The electrical stimulator used was developed at the Biomedical Engineering Laboratory (BioLab) of the Faculty of Electrical Engineering at the Federal University of Uberlândia, being bench-validated in [13] and validated in human subjects in [14]. It was built aiming at the restoration of sensory feedback; thus, the stimulus control parameters must be highly controllable and have high accuracy between the programmed and actual current. In [13], the authors proved through bench validation that the equipment meets these requirements.

The stimuli applied in the experiment consisted of rectangular, balanced biphasic waveforms, featuring a cathodic first phase and a 10 µs interphase interval. These parameters were selected because the rectangular waveform provides a fast rise time, thereby hindering nerve fiber accommodation. Additionally, the balanced biphasic nature drastically reduces the net electrical charge delivered to the skin, which prevents tissue damage when properly applied [15].

The electrode used to stimulate the volunteers was of a concentric circular type, made of siliconized carbon mesh, previously fabricated at the BioLab by [11], with the outer ring functioning as the anode and the inner one as the cathode, as shown in Figure 1.

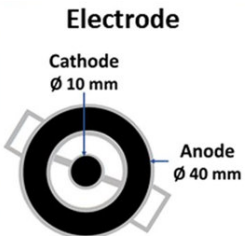


Fig. 1: Electrode used on the stimulation protocol. Figure taken from [7] with the author permission.

B. Stimulation protocol

The protocol consisted of two phases. The first phase was dedicated to determining the sensation threshold — defined as the lowest current amplitude at the lowest stimulation frequency that elicited any identifiable sensation by the volunteer — and the pain threshold, defined as the highest amplitude at the highest stimulation frequency that caused mild discomfort. These measurements were performed in triplicate with a 2-minute interval between each trial. The mean of these results was then used to calculate the amplitudes for the subsequent experimental phase, which were set as follows:

1) 50% above the sensation threshold;
2) 80% of the pain threshold;
3) The mean between the values obtained in steps 1 and 2.

The second phase was divided into three stimulation blocks. In each block, the current amplitude and pulse width were kept constant while the frequency was varied. Each stimulus lasted between 3 and 5 seconds, after which the volunteer answered a four-question survey. The questions were:

1) What was the intensity of the perceived sensation on a scale from 0 to 10?
2) Did the sensation feel more discrete or continuous?
3) Was there any change in the frequency of the last stimulus compared to the previous one (did it increase, decrease, or remain the same)?
4) What sensations were experienced during the application of the stimulus?

To answer question 4, the volunteer had on their hands a list of descriptors to choose from, presented in Table I.

TABLE I: List of descriptors taken from [16].

Tingling	Prickling	Electrical	Shocking
Sharp	Pricking	Stinging	Itching
Stabbing	Pressure	Touch	Cramping
Squeezing	Throbbing	Pounding	Aching
Pinching	Vibrating	Pulsing	Buzzing
Flutter	Tapping	Fizzing	Tickling
Movement	Pulling	Twitching	Contracting
Tensing	Tugging		

The interval between stimuli wasn't standardized since each volunteer responded to the questions in their own time. Also,

the stimulus could be replayed upon the volunteer's request. In this protocol, the response of sense accuracy was vital.

In every block, the stimulation frequency was varied first in an ascending order and immediately followed by a descending order. Throughout the experiment, the volunteer was blinded to the stimulation order, and the pulse width was maintained at 500 μs.

The same stimulation protocol was used in [14] to collect data with able-bodied people.

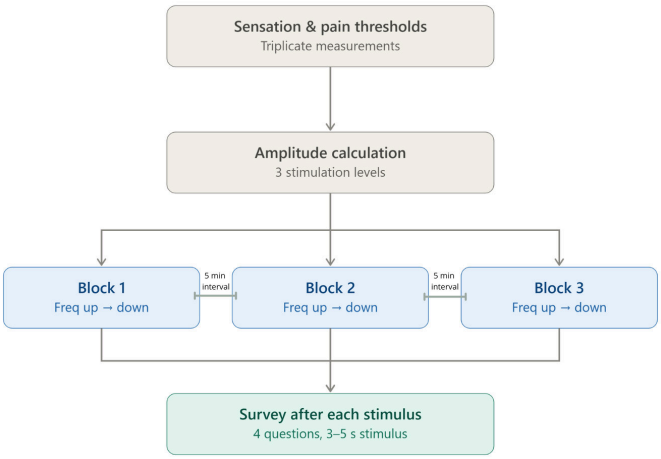


Fig. 2: Stimulation protocol fluxogram.

In this study, the first volunteer was a 55-year-old male with a wrist disarticulation amputation resulting from trauma 6 years prior. When questioned about his phantom limb, he reported a constant sensation of a clenched fist; therefore, it was not possible to map specific digits on the phantom limb. The second volunteer was a 64-year-old male with a transradial amputation, also, resulting from trauma 59 years ago. He reported feeling the phantom limb when moving the healthy arm, but without specificity, which made unable to map specific digits.

III. Results and Discussion

Data from Block 1 from the first volunteer were excluded from the figures and tables, as he reported no discernible sensations during this block — likely because the stimulation amplitude applied was insufficient to evoke clearly distinguishable sensations.

A. Sensation and pain thresholds

The sensation threshold previously found in experiments with healthy volunteers, as described in [14] presented a mean of 0.27 mA with a standard deviation of 0.06 mA. This indicates a high similarity among the volunteers in these measurements, as expected. It was anticipated that the amputee volunteers would exhibit a higher sensation threshold than the healthy group, given that the sensory receptor density at the fingertips is significantly greater than in other regions of the upper limb. However, a sensation threshold of 0.25 mA was identified in both volunteers, which is comparable to that of the healthy volunteers.

Conversely, the mean pain threshold for the healthy volunteers was 1.29 mA with a standard deviation of 0.7. Although this standard deviation might seem relatively broad, a comparison of the absolute values reveals that the comfortable range of current amplitudes varies finely across individuals. The lowest measured value was 0.57 mA, whereas the highest was 2.67 mA. If the stimulation amplitude tailored for the individual with the highest threshold were applied to the one with the lowest, it would cause substantial discomfort. For a system intended for continuous daily use, comfort must be prioritized. In the amputee volunteers, the pain threshold was measured at 5.24 mA for both of them. These results show a higher pain threshold between the amputee volunteers than in the healthy volunteers, which makes the range of possible current amplitudes to use on amputees bigger. These are a desired result, as the goal is to elicit as many sensations as possible.

B. Intensity

In the experiment with healthy volunteers conducted in [14], the perceived intensity exhibited the same logarithmic behavior reported by [11]. This is consistent with the electrical model of the skin, which consists of a resistor and a capacitor in parallel acting as a passive high-pass filter. Consequently, the perceived intensity tends to increase during the ascending phase of the stimulation blocks and decrease as the frequency drops during the descending phase. The logarithmic nature of this relationship occurs because this increase is not indefinite; biologically, there is a saturation limit to the recruitment of nerve fibers in the stimulated region. In the data collected from the amputee volunteers, this same behavior can be observed, although with a noisier trend due to the lower data density. This logarithmic behavior is particularly highlighted in Figure 3 (block 3 - decremental), which presented a correlation coefficient of $R^2 = 0.8961$, and Figure 4 (block 02 - incremental) with $R^2 = 0.8104$.

It can be observed that during the decremental phase, the same frequencies tend to receive a lower intensity rating. This is expected, as the stimuli are delivered over a prolonged period. Even at different frequencies, sensory perception varies depending on whether the nerve fibers are in a resting state or have already been excited; consequently, a stronger stimulus is required to reach the depolarization threshold.

C. Perception of varying frequency

Question 3 aimed to evaluate how frequency variations are perceived by the user. Thus, evaluating if, depending on the target sensory response, specific frequencies may be more suitable than others.

Given that the protocol was applied in an ascending order followed by a descending order, while the volunteers remained blinded to this sequence, the objective was to analyze the frequency of iteration ranges that produced a discernible difference in the perceived sensation. In [14], with able-bodied

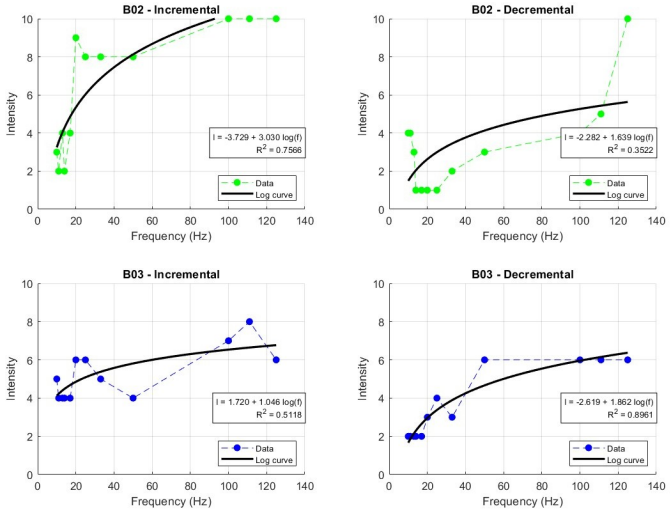


Fig. 3: Logarithmic regression of the relationship between perceived stimulus intensity and applied frequency of volunteer number 1.

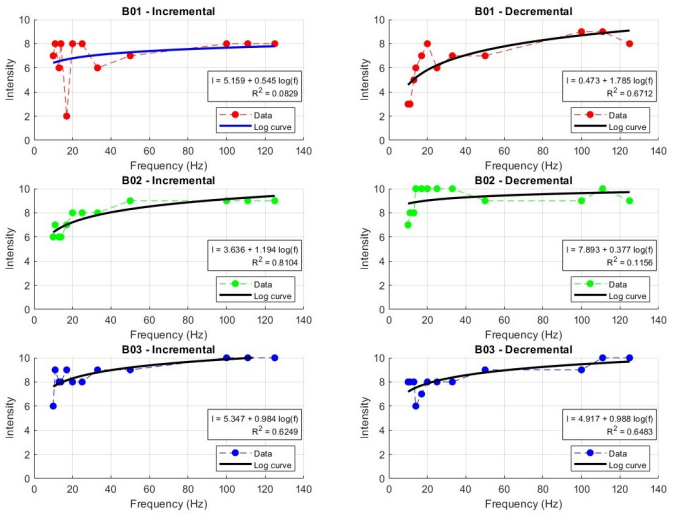


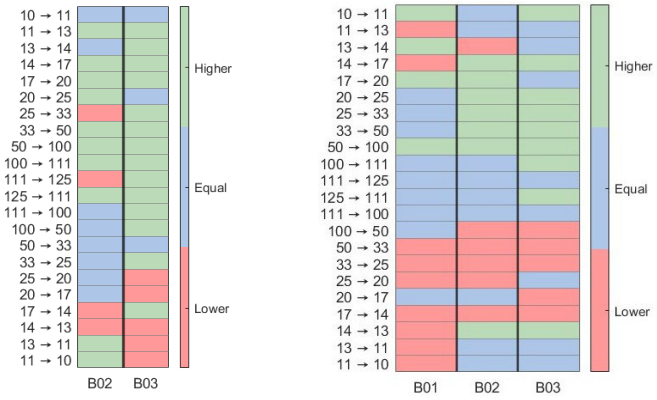
Fig. 4: Logarithmic regression of the relationship between perceived stimulus intensity and applied frequency of volunteer number 2.

volunteers, a decrease in this perception error was observed over the successive blocks. This trend is not well defined in the experiment with amputees. However, the "block 3 - incremental" phase had a significant drop in the error rate for both volunteers. But it didn't continue in the following experiment phase, as shown in Table II.

When seeing the data collected in [14], it is believed that the amplitude value was an essential factor for this error decrease. Additionally, due to the staircase design of the protocol, a learning effect may have occurred; the volunteer might have begun to recognize the stimulation pattern, thereby reducing the error rate. This conclusion cannot yet be extended to amputees, given the limited number of participants in the present study, and further trials with a larger amputee sample are required.

TABLE II: Change in frequency perception error rate. "V1" refers to the responses from volunteer 1 and "V2" refers to the responses from volunteer 2.

Block	V1	V2
Block 01 - incremental	-	63.63%
Block 01 - decremental	-	36.36%
Block 02 - incremental	63.63%	45.45%
Block 02 - decremental	45.45%	54.54%
Block 03 - incremental	18.18%	36.36%
Block 03 - decremental	90.90%	54.54%



(a) Responses collected from volunteer 1.

(b) Responses collected from volunteer 2.

Fig. 5: Changing frequency perception collected during the protocol application. Each column represents the responses from one block, and each line represents a changing frequency.

The results obtained with the amputee volunteer followed a similar trend, showing a decrease in the number of errors during block 3 in the incremental phase for the first volunteer. The second volunteer had similar numbers throughout the experiment.

D. Discrete vs continuous

The first subject reported experiencing only continuous sensations. Therefore, there was no variation in the responses regardless of the experimental block. Therefore, we hypothesize that the volunteer did not fully comprehend the distinction between discrete and continuous stimuli. Consequently, no comparable results could be obtained.

The second subject reported different responses to discrete and continuous stimuli, as shown in Figure 6. The lower frequencies (10-33 Hz) were mostly described as discrete and the higher ones (50-125 Hz) as continuous. These results are similar to the data collected from healthy volunteers in [11].

E. Descriptors

Lower frequencies (10–20 Hz) predominantly elicited comfortable, mechanoreceptive-like sensations, described by the

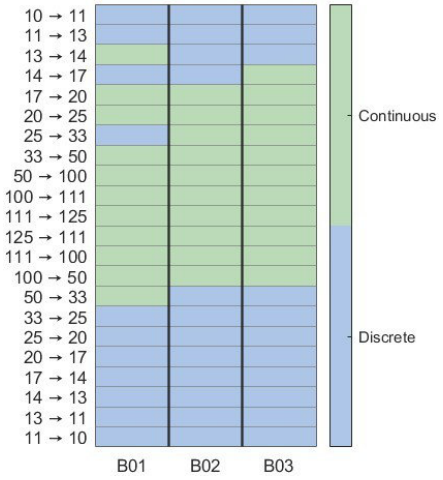


Fig. 6: Color map of responses regarding question 2 from volunteer 2. Columns represent the blocks, and lines represent the frequency change.

volunteers as predominantly ”pulsing” and ”tapping”. However, as the frequency increased towards 100–125 Hz, or when a higher current amplitude was applied, the tactile sensations rapidly transitioned into paresthetic and nociceptive descriptors, such as ”electrical shock”, ”prickling” and ”pinching”. Across all frequencies, ”tingling” was commonly used by both volunteers. Similar results were obtained by [12]. It is noticeable that the staircase protocol highlighted the phenomenon of neural adaptation and sensory hysteresis. For instance, in volunteer 1 Block 2, a 50 Hz stimulus was perceived as a ”shock” during the ascending phase, but the same frequency was described as ”tapping” during the descending phase. This indicates that prior excitation states significantly alter the subsequent perception, emphasizing the need for dynamic stimulus modulation to maintain consistency in prosthetic applications [17].

IV. Conclusion

This study evaluated the somatosensory perception evoked by TENS in two individuals with upper limb amputation, aiming to begin to establish viable parameters for the restoration of sensory feedback in upper limb prostheses. By applying the same psychophysical protocol previously validated in able-bodied individuals, this work provides direct within-protocol comparisons between amputee and able-bodied TENS-evoked perception, offering an empirical basis for parameter selection that aims to restore tactile feedback through electrostimulation.

The results demonstrated that the amputees' sensation threshold (0.25 mA) is comparable to that of healthy individuals, indicating the preservation of sensitivity to electrical stimulation. Conversely, the pain threshold was found to be considerably higher (5.24 mA), which is a positive finding as it provides a wider and more comfortable dynamic range of stimulation for daily user application.

Furthermore, it was observed that the perception of stimulation intensity follows a logarithmic behavior concerning frequency variation, respecting the biological limits of nerve fiber recruitment as seen in related works. It was also noted that stimulation frequencies below 50 Hz generate more distinguishable sensations, predominantly described as ”pulsing”.

As a study limitation, we observed the first volunteer's difficulty in distinguishing between discrete and continuous stimuli, which precluded a direct comparison of these approaches and indicates the need to refine prior training in future protocols. Test-retest reliability was not assessed in this study, as each volunteer completed the protocol in a single session; future work should increase sample size for statistical validation and repeat the protocol across multiple sessions to evaluate the stability of the measured thresholds. Additionally, the fixed ascending-then-descending frequency order may have introduced learning or adaptation effects, so a randomized, counterbalanced presentation is needed to isolate true frequency discrimination thresholds. Finally, the stimulation parameters identified here should be validated through closed-loop functional tasks — including stimulus recognition accuracy, grip force control, and object manipulation performance.

Despite these limitations, this case series contributes an initial, empirically grounded set of stimulation parameters for individuals with upper limb amputation, narrowing a gap between protocols validated only in able-bodied populations and the clinical population that would ultimately use them.

The ultimate goal will be to translate information from artificial tactile sensors into electrostimulation patterns that closely approximate the physiological response, thereby promoting a sense of limb ownership and reducing prosthesis abandonment rates.

Acknowledgment

The authors gratefully acknowledge the financial support provided by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) and Financiadora de Estudos e Projetos (FINEP).

References

[1] S. Salminger, H. Stino, L. H. Pichler, C. Gstoettner, A. Sturma, J. A. Mayer, M. Szivak, and O. C. Aszmann, “Current rates of prosthetic usage in upper-limb amputees – have innovations had an impact on device acceptance?” *Disability and Rehabilitation*, vol. 44, no. 14, pp. 1–12, 12 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33377803/

[2] B. K. Jakubiak and B. C. Feeney, “Affectionate touch to promote relational, psychological, and physical well-being in adulthood: a theoretical model and review of the research,” *Personality and Social Psychology Review*, vol. 21, no. 3, pp. 228–252, 5 2016.

[3] J. W. Sensinger and S. Dosen, “A review of sensory feedback in upper-limb prostheses from the perspective of human motor control,” *Frontiers in Neuroscience*, vol. 14, 6 2020.

[4] Nebojša Malešević and C. Antfolk, “Sensory feedback in upper limb prosthetics: advances and challenges,” *Nature Reviews Neurology*, 6 2024. [Online]. Available: https://doi.org/10.1038/s41582-024-00987-3

[5] D. M. Page, J. A. George, D. T. Kluger, C. Duncan, S. Wendelken, T. Davis, D. T. Hutchinson, and G. A. Clark, “Motor control and sensory feedback enhance prosthesis embodiment and reduce phantom pain after long-term hand amputation,” *Frontiers in Human Neuroscience*, vol. 12, 9 2018.

[6] J. E. Hall and M. E. Hall, *Sensory Receptors, Neuronal Circuits for Processing Information*, 14th ed. Philadelphia, PA: Elsevier, 2020, ch. 47.

[7] A. Clara, M. Filosa, A. B. Soares, and C. M. Oddo, "Type ii mechanoreceptors and cuneate spiking neuronal network enable touch localization on a large-area e-skin," *Nature Machine Intelligence*, 8 2025. [Online]. Available: https://www.nature.com/articles/s42256-025-01076-w

[8] C. G. Vance, D. L. Dailey, B. A. Rakel, and K. A. Sluka, "Using tens for pain control: the state of the evidence," *Pain Management*, vol. 4, no. 3, pp. 197–209, 5 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4186747/

[9] P. Svensson, U. Wijk, A. Björkman, and C. Antfolk, "A review of invasive and non-invasive sensory feedback in upper limb prostheses," *Expert Review of Medical Devices*, vol. 14, no. 6, pp. 439–447, 5 2017.

[10] K. Kaczmarek, J. Webster, P. Bach-y Rita, and W. Tompkins, "Electrotactile and vibrotactile displays for sensory substitution systems," *IEEE Transactions on Biomedical Engineering*, vol. 38, no. 1, pp. 1–16, 1991.

[11] A. C. P. R. da Costa, A. L. D. Siqueira Junior, F. A. de Castro Oliveira, S. R. de Jesus Oliveira, and A. B. Soares, "Electrodermal stimulation variant analysis for optimum perception in somatosensory feedback protocols," *Research on Biomedical Engineering*, vol. 38, no. 2, pp. 451–463, 1 2022.

[12] J. N. Mello, M. R. Garcia, A. C. P. R. da Costa, and A. B. Soares, "Towards optimum amplitude and frequency of electrotactile stimulation in amputee for effective somatotopic sensory feedback," *Research on Biomedical Engineering*, vol. 41, no. 1, 11 2024.

[13] V. Silva, L. Almada, and A. Soares, "Functional validation of a tactile feedback system for upper limb prostheses users based on electrodermal stimulation," in *XXIX Brazilian Congress on Biomedical Engineering - Volume 1: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing*, vol. 125. Cham: Springer, 2025. [Online]. Available: https://doi.org/10.1007/978-3-031-93646-3_98

[14] L. L. de Almada, V. H. B. da Silva and A. B. Soares, "Validação de um sistema de eletroestimulação cutâneo para retroalimentação sensorial," presented at the XVI Simpósio de Engenharia Biomédica (SEB 2025), Zenodo, Oct. 2025. doi: 10.5281/zenodo.17363854.

[15] D. R. Merrill, M. Bikson, and J. G. Jefferys, "Electrical stimulation of excitable tissue: design of efficacious and safe protocols," *Journal of Neuroscience Methods*, vol. 141, no. 2, pp. 171–198, 2 2005.

[16] E. L. Graczyk, B. P. Christie, Q. He, D. J. Tyler, and S. J. Bensmaia, "Frequency shapes the quality of tactile percepts evoked through electrical stimulation of the nerves," *The Journal of Neuroscience*, vol. 42, no. 10, pp. 2052–2064, 1 2022.

[17] E. L. Graczyk, B. P. Delhaye, M. A. Schiefer, S. J. Bensmaia, and D. J. Tyler, "Sensory adaptation to electrical stimulation of the somatosensory nerves," *Journal of Neural Engineering*, vol. 15, no. 4, p. 046002, 5 2018.

Determination of the Absorption Coefficient of Mouse Brain Tissue: Analysis of the Influence of Hypoxic-Ischemic Injury and Aging

1st Lívia Lemes Soares
Center for Engineering, Modeling, and Applied Social Sciences
Federal University of ABC
São Bernardo do Campo, Brazil
livia.lemes@aluno.ufabc.edu

2nd Victor Augusto Generoso
Center for Mathematics, Computing, and Cognition
Federal University of ABC
São Bernardo do Campo, Brazil
v.generoso@aluno.ufabc.edu.br

3rd Nasser Ali Daghastanli
Center for Engineering, Modeling, and Applied Social Sciences
Federal University of ABC
São Bernardo do Campo, Brazil
nasser.daghastanli@ufabc.edu.br

4th Silvia Honda Takada
Center for Mathematics, Computing, and Cognition
Federal University of ABC
São Bernardo do Campo, Brazil
silvia.honda@ufabc.edu.br

***Abstract*—Transcranial photobiomodulation is a non-invasive therapeutic technique that uses light in the red and near-infrared spectrum to modulate cellular processes, especially those related to mitochondrial function, showing therapeutic potential in neurological diseases and brain injuries. Among its applications, hypoxic-ischemic encephalopathy (HIE) stands out, a condition frequently associated with perinatal asphyxia and characterized by alterations in neonatal neurological development. However, the optimization of the technique's parameters is still limited by insufficient knowledge of the optical properties of target brain tissues, which are essential for determining light distribution within the tissue.**

This study aimed to determine the optical absorption coefficient (μa) of brain tissue from mice SAMP8 at neonatal periods P4 and P8, comparing control and anoxia-subjected animals, in order to investigate changes in optical properties induced by hypoxic insult and development. Measurements of incident and transmitted light intensity were obtained by optical spectrophotometry, allowing the estimation of the absorption coefficient at 808 nm, a wavelength commonly used in transcranial photobiomodulation protocols.

The results demonstrated a significant difference among the evaluated groups (Kruskal-Wallis, H = 17.15; p $<$ 0.001), with a large effect size (ϵ^2 = 0.7859), indicating that approximately 78.6% of the variance in μa values was attributed to the experimental condition. Post hoc analysis (Dunn's test) revealed a significant difference between the Anoxia P4 and Anoxia P8 groups (p = 0.001), evidencing a temporal change in optical properties after anoxia exposure, with a higher absorption coefficient at the earlier period (0.0444 vs. 0.0197 mm^{-1}). Furthermore, the Anoxia P4 group differed significantly from the Control P8 group (p = 0.022), suggesting that anoxic tissue at P4 presents optical properties distinct from those observed in physiological development at P8. In contrast, no significant differences were observed between Control P4 and Control P8 (p = 0.370), indicating that physiological neonatal development did not promote expressive changes in the absorption coefficient within the evaluated interval. Likewise, there was no significant difference between Anoxia P8 and Control P8 (p = 1.000), suggesting possible recovery or adaptation of tissue optical properties in periods later after injury.

Taken together, these results indicate that neonatal anoxia modifies the temporal trajectory of brain optical properties during development, promoting significant changes in the early period (P4) that do not persist in the later period (P8), when anoxic tissue presents characteristics similar to the control. The characterization of these properties is essential to improve the application of transcranial photobiomodulation, allowing greater precision in defining therapeutic parameters, such as dose and application time, for conditions such as hypoxic-ischemic encephalopathy and other neonatal brain injuries, especially considering that therapeutic windows at P8 may offer greater predictability in light delivery to the tissue.

***Index Terms*—Transcranial Photobiomodulation, Absorption Coefficient, Perinatal Anoxia, Near-Infrared Laser.**

I. Introduction

A. Photobiomodulation

Photobiomodulation consists of the use of light in the red or near-infrared bands to regenerate or protect injured or at-risk tissues [2]. Transcranial photobiomodulation (t-PBM), in turn, presents great potential as a non-invasive therapeutic method for brain injuries [3]. t-PBM has proven effective both in animal models and in clinical trials for the treatment of stroke, traumatic brain injury (TBI), psychiatric disorders such as anxiety and depression, and even neurodegenerative diseases such as Alzheimer's and Parkinson's [1].

There are two mechanisms associated with the cellular effect of photobiomodulation. One is specific, which is due to the resonant interaction of light with chromophores present in the cell, such as mitochondria, nitrosyl cytochrome complexes of the electron transfer chain, hemoglobin, etc. On the other

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)

hand, the non-resonant interaction of light with cell membrane proteins represents a non-specific mechanism. However, it is important to emphasize that the non-specific mechanism does not pose a risk to the cell, whereas the specific mechanism can damage the membrane or cellular behavior through what is known as photodynamic damage. Therefore, it is important to determine the appropriate light dose for photobiomodulation, as very high intensities cause cell damage [4].

There is still no consensus on the best parameters and protocols for photobiomodulation application. Nevertheless, some parameters are essential for the good effectiveness of photobiomodulation: wavelength, energy, power, illuminance, irradiance, pulse mode and duration, and treatment repetition. All these parameters are still highly variable in the literature. Furthermore, [5] demonstrated that the number of mitochondria also influences the efficacy of photobiomodulation, such that cells with a greater number of mitochondria respond better to the technique than those with fewer, probably due to the specific mechanism of the technique, which, as mentioned earlier, interacts with cell chromophores.

Finally, the optical properties of the target tissue are also relevant for the action of photobiomodulation, since they interfere with photon transport through the tissue, which are the particles that actually interact with the molecules that compose it [6].

B. Light-Tissue Interaction Mechanisms

As mentioned earlier, light interacts with tissue in two ways to generate the photobiomodulation effect: resonantly, the specific mechanism of the technique, and non-resonantly, the non-specific mechanism. The effect of these interactions can be considered specific or non-specific depending on the optical phenomenon each causes [4].

Resonant interaction occurs when the frequency of light (or the photon energy) corresponds to one of the characteristic energy levels of a substance. When light interacts resonantly with biological material, it can excite molecules or atoms to a higher energy level. In other words, light has energy that corresponds to an energy transition of molecules that compose biological tissue [27].

The characteristics of this type of interaction include: high efficiency; since light has the same energy as the energy transitions of the biological system, the absorption or emission of energy is very efficient. Specificity; the interaction occurs in a highly specific manner due to the correspondence between light frequencies and the energy transitions of molecules or atoms present in the tissue [27].

Finally, the optical phenomena that can be observed from a resonant interaction are: light absorption by chromophores; biological molecules such as proteins or pigments (e.g., hemoglobin, chlorophyll, or melanin) absorb light at specific wavelengths. When the light frequency exactly matches the electronic transition frequency of the chromophore, resonant absorption occurs. Fluorescence: after photon absorption, one or more electrons of the molecule can be excited to a higher quantum state. When they return to their lower energy state, they emit a photon with a different frequency (and longer wavelength), a phenomenon known as fluorescence [27].

On the other hand, in non-resonant interaction, light does not directly correspond to any specific energy transition within the biological tissue. Instead, light can interact with molecules in a more general way, without exciting a specific energy state [27].

The characteristics of this type of interaction are: lower efficiency; there is no absorption or emission of energy as efficient as in resonant interaction, since there is no correspondence between light frequency and the material's energy transitions. Greater diffusion; the interaction tends to be less specific, affecting molecules more generally without causing direct excitation [27].

Finally, the optical phenomena observed in this case are: scattering effect; resulting from light scattering by molecules or structures of biological tissue without resonant absorption. This phenomenon occurs when molecules or structures do not absorb light but deviate its path, as happens with Rayleigh or Mie scattering. Thermal effects: occur when light is absorbed, but not resonantly. Thus, part of the energy is converted into heat, resulting in an increase in tissue temperature. This is typical in techniques such as phototherapy, where temperature increase can be used to treat certain types of tissue or cells [27].

C. Perinatal Asphyxia

Perinatal asphyxia (PA) consists of the transient interruption of body oxygenation and results from the low level of oxygen available to cells when the supply-demand balance of the gas is unbalanced. With an incidence of 1% to 6% in live births, one of its consequences is mitochondrial degeneration, impairing cellular energy metabolism [13]. In a study with cardiac cells, [14] demonstrated that the mitochondrial respiratory chain was damaged by the hypoxia-ischemia process and, although some recovery was observed after reoxygenation, mitochondrial activity will not return to its normal level.

The brain, especially during its developmental period, is particularly sensitive to oxygen deprivation. Thus, PA is directly related to cognitive sequelae such as learning deficits and impaired spatial memory. The condition, responsible for 23% of newborn deaths worldwide, can cause multiple organ failure and results in hypoxic-ischemic encephalopathy [15; 13].

D. Hypoxic-Ischemic Encephalopathy and t-PBM

Hypoxic-ischemic encephalopathy (HIE) is a clinical condition that affects, as mentioned earlier, newborns. It is characterized by moderate to severe metabolic acidosis at birth and is associated with global impairment of neonatal neural function [9]. Worldwide, the incidence of HIE is 0.15% in full-term infants, and when considering premature infants, the incidence can increase by up to 60%. Thus, HIE demonstrates a significant risk for newborns worldwide [10].

t-PBM shows potential as a treatment for HIE, as it increases adenosine triphosphate (ATP) production by cell mitochondria.

Cytochrome c oxidase, an enzymatic complex that plays an important role in the mitochondrial energy production chain, absorbs light at wavelengths near 830 nm and stimulates ATP production. Additionally, photobiostimulation activates secondary cell signaling pathways that also contribute to the normalization of ATP levels in cells [11].

Thus, cellular energy metabolism, impaired by the ischemic process of HIE, is improved by t-PBM, preventing the spread of cell death in the penumbra region, where cells are at reversible risk of death. There is also evidence that t-PBM increases neurogenesis and the production of endogenous neurotrophic factors (molecules responsible for neuronal growth and survival of mature neurons), contributing to the recovery of brain function lost due to HIE [11].

Many cases of HIE result from perinatal asphyxia (PA) and, even with the application of already established PA treatments (e.g., therapeutic hypothermia), about 45% of newborns with HIE die or present significant sequelae, such as epilepsy, neurological and/or cognitive deficits, and cerebral palsy [12].

E. Senescence-Accelerated Mouse Prone 8 (SAMP8)

SAMP8 is an animal model of accelerated senescence developed by [16]. When compared to a model resistant to accelerated senescence (SAMR1), the differences are evident: the average lifespan of SAMR1 is 18 months, while SAMP8 lives only 12 months on average. Furthermore, SAMP8 presents, at the 8th month of life, a higher score than SAMR1 on the senescence grading scale, which evaluates cellular aging through rates of irreversible cell activity arrest and production of substances that contribute to the development of aging-related diseases [17; 18].

Additionally, [19] demonstrated that energy deposition in brain tissues by t-PBM decreases with increasing age. Thus, appropriate parameters depend on the patient's age, so they should be adapted to the individual's developmental and/or aging stage [20]. Therefore, SAMP8 proves to be an interesting model for evaluating laser parameters for t-PBM, as it allows assessing the impact of age over a short period.

F. Optical Properties of Cerebral Tissue

The development and improvement of phototherapy techniques, such as photobiomodulation, is due to advances in biophotonics studies, a field of physics that aims to elucidate the mediators of the interaction between light and biological systems. By determining the optical properties of tissues, it is possible to estimate the direction of photon propagation in the irradiated tissue, as well as the light fluence rate distribution. Therefore, it is essential to know these properties to predict both absorption and energy distribution throughout the tissue, which shape the therapeutic effects of irradiation [7].

One of the main optical properties that affects how light interacts and propagates through a tissue is its absorption coefficient [7]. In order to determine it for biological tissues, such as brain tissue, it is possible to use, with reservations, the Lambert-Beer Law. To do so, it is necessary to determine the spectral absorbance of mouse brain slices, for example. From the absorbance, it is possible to estimate the absorption coefficient (μa) using the relation derived from the Lambert-Beer law:

$$\mu_a = \frac{ln(10).A}{L}$$

where A is the measured absorbance and L is the sample thickness. This step will provide a preliminary estimate of μa, acknowledging that optical scattering effects may contribute to the attenuation of transmitted light [24]. The use of light in the near-infrared (NIR) band, on the other hand, is due to its high tissue penetration power, since the chromophores water and hemoglobin absorb little light in the wavelength range of 600 to 900 nm, which encompasses part of the NIR [8].

Therefore, this project proposes to measure the optical absorption of brain tissue samples from SAMP8 mice, using optical absorption spectrophotometry, to determine the absorption coefficient of brain tissue, in order to allow a better understanding of the brain-light interaction and, consequently, contribute to the definition of the most appropriate t-PBM parameters for the treatment of HIE and PA.

II. Objectives

Determination of the optical absorption coefficients of brain tissue in SAMP8 mice at ages P4 and P8, in control conditions and after submission to neonatal anoxia at stage P3. To identify whether there is an influence of age and mitochondrial numbers on the optical properties of brain tissue in SAMP8 mice.

III. Methods

A. Experimental Animals

All experiments were carried out in accordance with the Ethical Principles of Animal Experimentation adopted by the Brazilian College of Animal Experimentation (COBEA) and approved by the Ethics Committee on the Use of Animals of the Federal University of ABC (protocols 8852240517 and 3572091023).

SAMP8 mice, an animal model of accelerated senescence, were used at postnatal days 4 and 8. For each age, there were two different groups: animals subjected to neonatal anoxia and control animals. The animals were housed in the animal facility of the Federal University of ABC, São Bernardo do Campo campus, under a 12-hour light/dark cycle (with the light phase starting at 7:00 AM), maintaining a constant temperature of (22 ± 1°C. They had unrestricted access to water and food. Both males and females were used for breeding, and groups were formed based on their litters.

Based on literature data [25], a minimum detectable difference of 2 mm^{-1} and a standard deviation of 0.8 mm^{-1} were estimated. Sample size calculation for ANOVA with α= 0.05 and 80% power indicated the need for 5 to 6 animals per group. Thus, n = 6 per group was adopted, totaling 24 animals, based on international guidelines that recommend considering adequate sample sizes to ensure greater statistical robustness, compensate for possible experimental losses, and ensure result reproducibility, as indicated by [26].

B. Neonatal Asphyxia Experimental Model

To simulate the perinatal asphyxia process, the neonatal anoxia model previously described by Takada et al. [21] was adopted. P3 pups were exposed to a continuous flow of 10.5 L/min of pure nitrogen (100%) for 12 minutes, at a temperature of 37°C, in a semi-hermetic chamber. The control group underwent the same procedure, but without the nitrogen flow.

C. Transcardial Perfusion

The animals, after being anesthetized with urethane (1000 mg/Kg), were subjected to transcardiac perfusion with 0.9% saline solution (pH 7.4). Perfusion was maintained until whitening of peripheral tissues was observed, which characterizes complete blood removal. Subsequently, the brains were dissected and post-fixed in 4% PFA solution for 24 hours at 4°C. This is because Moy et al.[22] demonstrated that the presence of blood in the tissue compromises the accuracy of optical absorption measurements, and its complete removal is recommended for reliable spectral analyses. The brains were stored in phosphate-buffered saline (PBS) at 4°C until the time of measurement, preventing dehydration and enzymatic degradation [23].

D. Spectrometry

Each sample (brain) of the mice was placed in a holder to ensure stable fixation and was irradiated by the laser, in continuous wave mode. Transmittance measurements were acquired using the Ocean Optics spectrometer with the OceanView software. To ensure repeatability, each measurement was performed three times, and the average value was used for the final analysis. The reduced scattering coefficient was assumed as $\mu s'$ = 1 mm^{-1}, a typical average value reported in the literature for brain tissues [29].

E. Experimental Setup

- Microscope slides with samples were carefully positioned and aligned on the spectrophotometer, and 3 scanning measurements (400–900 nm) of the tissue optical absorption were performed.
- A dental diode laser source (808 nm) was positioned perpendicularly to the slide with the sample, and the optical fiber was positioned off-axis from the incident beam, in order to capture only diffuse components and avoid direct light. The respective diffuse transmittances, information required to determine the absorption coefficients, were measured separately.

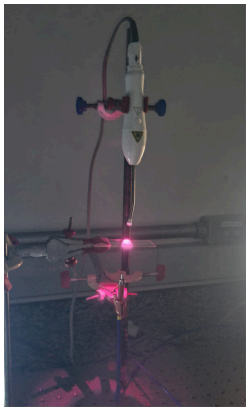


Fig. 1. Experimental Setup.

IV. Results

The analysis of the optical properties of brain tissue was performed by determining the optical absorption coefficient (μa) at the wavelength of 808 nm. After exploratory data analysis, an outlier value was identified in the Control P8 group, which was excluded from subsequent statistical analyses.

The descriptive statistics of the absorption (Table I) coefficient (μa) revealed differences in the distribution of values among the experimental groups. The Anoxia P4 group presented the highest mean absorption coefficient (0.04440 mm^{-1}), followed by the Control P4 group (0.03638 mm^{-1}), Control P8 group (0.03318 mm^{-1}), and Anoxia P8 group (0.01971 mm^{-1}). The Anoxia P4 group also showed a relatively narrow distribution, with values ranging from 0.03834 to 0.05623 mm^{-1}, whereas the Control P8 group exhibited the greatest variability, with values ranging from 0.02238 to 0.06569 mm^{-1}. The median absorption coefficients were 0.04104 mm^{-1} for Anoxia P4, 0.01748 $mm^{-1}$for Anoxia P8, 0.03749 mm^{-1} for Control P4, and 0.02626 mm^{-1} for Control P8. Overall, these descriptive results indicate a lower absorption coefficient in the Anoxia P8 group compared with the other experimental groups, particularly when compared with Anoxia P4.

TABLE I
DESCRIPTIVE STATISTICS OF THE ABSORPTION COEFFICIENT (μ_a) FOR EACH EXPERIMENTAL GROUP.

Group	n	Mean	SD	Median	Minimum	Maximum
Anoxia P4	6	0.04440	0.00712	0.04104	0.03834	0.05623
Anoxia P8	6	0.01971	0.00668	0.01748	0.01222	0.02803
Control P4	6	0.03638	0.00447	0.03749	0.02891	0.04144
Control P8	5	0.03318	0.01825	0.02626	0.02238	0.06569

The μa values showed differences among the experimental groups. Statistical analysis using the Kruskal-Wallis test indicated a significant difference among the four evaluated groups (H = 17.15; p = 0.00066), with an effect size of ϵ^2 = 0.7859, indicating a high effect of the experimental condition on the absorption coefficient.

Since the data did not meet the assumptions of parametric statistical tests, non-parametric methods were used for the

analysis. When a significant overall difference was identified by the Kruskal–Wallis test, post hoc pairwise comparisons were performed using Dunn's test to determine which groups differed significantly from each other. To account for the increased risk of Type I error resulting from multiple pairwise comparisons, p-values were adjusted using the Benjamini–Hochberg procedure, which controls the false discovery rate (FDR). This approach was selected as it is appropriate for non-parametric data and allows for the identification of significant differences between specific groups while limiting the proportion of false-positive findings among the statistically significant results.

TABLE II
DUNN'S POST HOC PAIRWISE COMPARISONS WITH BENJAMINI–HOCHBERG CORRECTION.

Comparison	**Adjusted p-value**
Anoxia P4 vs. Anoxia P8	0.003299*
Anoxia P4 vs. Control P4	0.728374
Anoxia P4 vs. Control P8	0.027198*
Anoxia P8 vs. Control P4	0.055263
Anoxia P8 vs. Control P8	1.000000
Control P4 vs. Control P8	0.226556

* $p < 0.05$.

Multiple comparisons among groups (Table II) revealed significant differences between the Anoxia P4 and Anoxia P8 groups (p = 0.003299), as well as between the Anoxia P4 and Control P8 groups (p = 0.027198). The Anoxia P4 group presented a higher absorption coefficient (mean = 0.0444 mm^{-1}) compared with both the Anoxia P8 group (mean = 0.0197 mm^{-1}) and the Control P8 group (mean = 0.0269 mm^{-1}). No statistically significant difference was observed between the Control P4 (mean = 0.0364 mm^{-1}) and Control P8 groups (p = 0.226556). Similarly, no significant differences were observed between the Anoxia P4 and Control P4 groups (p = 0.728374), or between the Anoxia P8 and Control P8 groups (p = 1.000000). The comparison between Anoxia P8 and Control P4 also did not reach statistical significance (p = 0.055263).

Also, the distribution of the absorption coefficient (μ_a) across the experimental groups is illustrated in Fig. 2. The boxplot shows a higher central tendency of μ_a in the Anoxia P4 group compared with the other groups, whereas the Anoxia P8 group exhibited the lowest values and greater variability. The Control P4 and Control P8 groups showed intermediate absorption coefficients, with a relatively narrower distribution in the latter group. The Anoxia P4 group also presented a broader range of values and a high individual value, while the Anoxia P8 group displayed a wider dispersion of measurements. Overall, the observed distribution is consistent with the statistical analysis, particularly the significant differences involving the Anoxia P4 group, and suggests that the optical absorption properties of brain tissue may vary according to both developmental stage and exposure to hypoxic-ischemic insult.

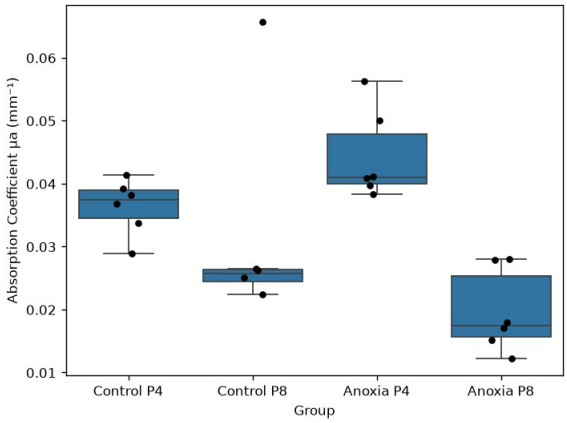


Fig. 2. Boxplot of Results.

These preliminary results suggest that hypoxic-ischemic insult may be associated with changes in the optical properties of brain tissue during neonatal development, particularly at the P4 developmental stage, as indicated by the higher absorption coefficient observed in the Anoxia P4 group.

V. DISCUSSION

The results demonstrated significant differences in the optical absorption coefficient (μa) of brain tissue among the experimental groups of SAMP8 mice. The Kruskal–Wallis test indicated an overall effect of experimental group on μa(H = 17.15; p $<$ 0.001), with a large effect size (ϵ^2 = 0.7859). This indicates that a substantial proportion of the variability observed in the absorption coefficient was associated with the experimental groups. However, the overall result should be interpreted together with the pairwise comparisons and the uncertainty associated with the estimated group values.

Descriptive statistics showed that the highest mean $\mu a$was observed in the Anoxia P4 group (0.0440 ± 0.00712 mm^{-1}; n = 6), followed by Control P4 (0.03638 ± 0.00447 mm^{-1}; n = 6), Control P8 (0.03318 ± 0.01825 mm^{-1}; n = 5), and Anoxia P8 (0.01971 ± 0.000668 mm^{-1}; n = 6). The approximate 95% confidence intervals for the mean μawere 0.0365–0.0515 mm^{-1} for Anoxia P4, 0.0317–0.0411 mm^{-1} for Control P4, 0.0105–0.0558 mm^{-1} for Control P8, and 0.0190–0.0204 mm^{-1} for Anoxia P8. These confidence intervals are presented as descriptive estimates of uncertainty around the group means and should be interpreted alongside the non-parametric statistical analysis.

Dunn's post hoc analysis with Benjamini–Hochberg correction showed that the principal significant difference occurred between the Anoxia P4 and Anoxia P8 groups. The Anoxia P4 group presented a significantly higher μathan the Anoxia P8 group (adjusted p = 0.00329). In contrast, no statistically significant differences were observed between control and anoxia groups at the same postnatal age. At P4, the difference between Anoxia P4 and Control P4 was not significant (adjusted p = 0.72834), despite the numerically higher mean observed in Anoxia P4. Similarly, Anoxia P8 and Control P8 did not differ significantly (adjusted p = 1.00000).

These results indicate that the main pairwise difference identified in the present experiment was between the two anoxia groups at different postnatal ages, rather than between anoxia and control animals at the same age. Therefore, the data do not demonstrate a statistically detectable difference in μaattributable to anoxia at either P4 or P8 when animals are compared within the same developmental stage.

The difference between Anoxia P4 and Anoxia P8 indicates that μavaried between the two postnatal ages within the anoxia condition. However, this finding cannot be interpreted as evidence of recovery, adaptation, or attenuation of the effects of the hypoxic-ischemic insult. Because the study has a cross-sectional design, the comparison between Anoxia P4 and Anoxia P8 simultaneously involves differences in postnatal age and experimental condition. Consequently, the observed difference may reflect developmental changes, effects associated with the hypoxic-ischemic condition, or a combination of both. Additional longitudinal measurements would be necessary to determine whether μachanges over time within the same animals.

The comparison between Anoxia P4 and Control P8 was also statistically significant (adjusted p = 0.027198). However, these groups differ in both experimental condition and postnatal age. Therefore, this comparison does not allow the independent contribution of anoxia to be distinguished from that of developmental stage. The remaining pairwise comparisons were not statistically significant, including Anoxia P4 versus Control P4 (adjusted p = 0.72834), Anoxia P8 versus Control P8 (adjusted p = 1.00000), and Control P4 versus Control P8 (adjusted p = 0.226556).

The confidence intervals provide additional information regarding the precision of the estimated group means. The relatively narrow interval observed for Anoxia P8 (0.0190–0.0204 mm^{-1}) contrasts with the substantially wider interval observed for Control P8 (0.0105–0.0558 mm^{-1}). This difference in interval width is consistent with the greater variability observed in the Control P8 group (SD = 0.01825 mm^{-1}) and its smaller sample size (n = 5). In comparison, the Anoxia P4, Anoxia P8, and Control P4 groups each included six animals. Thus, the small sample sizes, particularly in Control P8, introduce substantial uncertainty into the estimation of $\mu a$and may limit the statistical power to detect relatively small differences between groups.

Importantly, the absence of statistically significant differences between control and anoxia groups at the same postnatal age should therefore not be interpreted as evidence of equivalence between the groups. With only five to six animals per group, the study may have limited ability to detect subtle effects, particularly in the presence of biological variability. Conversely, the significant difference between Anoxia P4 and Anoxia P8 should not be interpreted independently of the sample size and variability of each group. Larger experimental groups would provide more precise estimates and greater statistical power to determine whether the observed differences are reproducible.

The biological interpretation of the observed differences should also remain restricted to the present SAMP8 mouse model. Changes in μamay potentially be associated with differences in tissue composition or in the concentration and distribution of absorbing chromophores, such as hemoglobin, water, and mitochondrial chromophores including cytochrome c oxidase [24,28]. However, these variables were not directly measured in the present experiment. Therefore, specific biological mechanisms cannot be established from the optical absorption measurements alone.

The observed differences in $\mu a$are relevant to the optical characterization of brain tissue in the SAMP8 model because absorption contributes to light attenuation within biological tissue. However, the present results do not establish whether the observed differences in μaresult in different biological responses to transcranial photobiomodulation, nor do they support the identification of a specific developmental stage as a more favorable therapeutic window. Such conclusions would require independent experimental measurements of treatment response and biological outcomes.

A further limitation is that $\mu a$is an integrative optical parameter and does not allow the individual contribution of specific chromophores to be determined. In addition, the present approach does not provide an independent simultaneous estimate of tissue scattering. Therefore, an important next step is to validate the estimated μausing an optical method specifically suited to turbid biological media. In particular, an integrating sphere combined with the Inverse Adding-Doubling (IAD) method would allow absorption and reduced scattering coefficients to be estimated simultaneously. Independent validation using this approach would strengthen the reliability of the optical characterization and determine whether the differences observed among the SAMP8 experimental groups are reproducible when both absorption and scattering are considered.

Future studies should therefore include larger sample sizes to improve statistical power and reduce the uncertainty of the estimated optical properties. Longitudinal experimental designs, when feasible, together with histological, immunohistochemical, or biochemical measurements, would also help distinguish developmental changes from alterations associated with the hypoxic-ischemic condition. These approaches would provide a stronger basis for interpreting the differences in optical absorption observed between P4 and P8 in the SAMP8 mouse model.

REFERENCES

[1] Hamblin MR. Shining light on the head: Photobiomodulation for brain disorders. BBA Clin. 2016 Oct 1;6:113-124. doi: 10.1016/j.bbacli.2016.09.002. PMID: 27752476; PMCID: PMC5066074.

[2] Mester E, Ludany G, Selyei M, Szende B, Total GJ. The stimulating effect of low power laser rays on biological systems. Laser Rev. 1968 Mar;1:3-8.Mester E, Ludany G, Selyei M, Szende B, Total GJ. The stimulating effect of low power laser rays on biological systems. Laser Rev. 1968 Mar;1:3-8.

[3] Salehpour F, Cassano P, Rouhi N, Hamblin MR, De Taboada L, Farajdokht F, Mahmoudi J. Penetration profiles of visible and near-infrared lasers and light-emitting diode light through the head tissues in animal and human species: a review of literature. Photobiomodul Photomed Laser Surg. 2019 Oct;37(10):581-595. doi: 10.1089/photob.2019.4676. PMID: 31553265.

[4] Liu TCY, Jiao JL, Xu XY, Liu XG, Deng SX, Liu SH. Photobiomodulation: phenomenology and its mechanism. Proc SPIE. 2005;5630:185-191. doi: 10.1117/12.581165.

[5] Zein R, Selting W, Hamblin MR. Review of light parameters and photobiomodulation efficacy: dive into complexity. J Biomed Opt. 2018 Dec 1;23(12):1-17. doi: 10.1117/1.JBO.23.12.120901. PMID: 30550048; PMCID: PMC8355782.

[6] Pitzschke A, Lovisa B, Seydoux O, Haenggi M, Oertel MF, Zellweger M, Tardy Y, Wagnières G. Optical properties of rabbit brain in the red and near-infrared: changes observed under in vivo, postmortem, frozen, and formalin-fixated conditions. J Biomed Opt. 2015 Feb;20(2):025006. doi: 10.1117/1.JBO.20.2.025006. PMID: 25706688.

[7] Shanshool AS, Lazareva EN, Hamdy O, Tuchin VV. Optical properties and fluence distribution in rabbit head tissues at selected laser wavelengths. Materials (Basel). 2022 Aug 18;15(16):5696. doi: 10.3390/ma15165696. PMID: 36013828; PMCID: PMC9413642.

[8] Hamdy O, Abdel-Salam Z, Abdel-Harith M. Optical characterization of biological tissues based on fluorescence, absorption, and scattering properties. Diagnostics (Basel). 2022 Nov 17;12(11):2846. doi: 10.3390/diagnostics12112846. PMID: 36428905; PMCID: PMC9689259.

[9] Gunn AJ, Soul JS, Vesoulis ZA, Ferriero DM. The importance of not increasing confusion around neonatal encephalopathy and hypoxic-ischemic encephalopathy. Pediatr Res. 2024 Mar;95(4):871-872. doi: 10.1038/s41390-023-03001-6. PMID: 38158415.

[10] Kurinczuk JJ, White-Koning M, Badawi N. Epidemiology of neonatal encephalopathy and hypoxic-ischaemic encephalopathy. Early Hum Dev. 2010 Jun 16;86(6):329-338. doi: 10.1016/j.earlhumdev.2010.05.010. PMID: 20554402.

[11] Lapchak PA. Taking a light approach to treating acute ischemic stroke patients: transcranial near-infrared laser therapy translational science. Ann Med. 2010 Dec;42(8):576-586. doi: 10.3109/07853890.2010.532811. PMID: 21039081; PMCID: PMC3059546.

[12] Annink KV, de Vries LS, Groenendaal F, van den Heuvel MP, van Haren NEM, Swaab H, van Handel M, Jongmans MJ, Benders MJ, van der Aa NE. The long-term effect of perinatal asphyxia on hippocampal volumes. Pediatr Res. 2019 Jan;85(1):43-49. doi: 10.1038/s41390-018-0115-8. PMID: 30254237.

[13] Iribarren I, Hilario E, Álvarez A, Alonso-Alconada D. Neonatal multiple organ failure after perinatal asphyxia. An Pediatr (Engl Ed). 2022 Oct;97(4):280.e1-280.e8. doi: 10.1016/j.anpede.2022.08.010. PMID: 36115781.

[14] Yang HM, Yao YJ, Li WR. Injury of myocardial mitochondria in neonatal swines with hypoxic-ischemic brain damage. Sichuan Da Xue Xue Bao Yi Xue Ban. 2004 Jan;35(1):39-41. PMID: 14981810.

[15] Takada SH, Motta-Teixeira LC, Machado-Nils AV, Lee VY, Sampaio CA, Polli RS, Malheiros JM, Takase LF, Kihara AH, Covolan L, Xavier GF, Nogueira MI. Impact of neonatal anoxia on adult rat hippocampal volume, neurogenesis and behavior. Behav Brain Res. 2016 Jan 1;296:331-338. doi: 10.1016/j.bbr.2015.08.039. PMID: 26416672.

[16] Takeda T, Hosokawa M, Higuchi K. Senescence-accelerated mouse (SAM): a novel murine model of accelerated senescence. J Am Geriatr Soc. 1991 Sep;39(9):911-919. doi: 10.1111/j.1532-5415.1991.tb04460.x. PMID: 1885867.

[17] Pačesová A, Holubová M, Hrubá L, Strnadová V, Neprašová B, Pelantová H, Kuzma M, Železná B, Kuneš J, Maletínská L. Age-related metabolic and neurodegenerative changes in SAMP8 mice. Aging (Albany NY). 2022 Sep 16;14(18):7300-7327. doi: 10.18632/aging.204284. PMID: 36126192; PMCID: PMC9550245.

[18] Raj V, Charles S, Goenka L, Ramamoorthy T, Marimuthu C, Emmanuel C, Mala K, Kumarasamy S, George M. Assessment of peripheral blood mononuclear cells senescence and endothelial dysfunction among adults with high cardiovascular risk. Arq Bras Cardiol. 2021 Jan;116(1):37-47. doi: 10.36660/abc.20190409. PMID: 33566963; PMCID: PMC8159491.

[19] Yuan Y, Cassano P, Pias M, Fang Q. Transcranial photobiomodulation with near-infrared light from childhood to elderliness: simulation of dosimetry. Neurophotonics. 2020 Jan;7(1):015009. doi: 10.1117/1.NPh.7.1.015009. PMID: 32118086; PMCID: PMC7039173.

[20] Rodríguez-Fernández L, Zorzo C, Arias JL. Photobiomodulation in the aging brain: a systematic review from animal models to humans. GeroScience. 2024 Dec;46(6):6583-6623. doi: 10.1007/s11357-024-01231-y. PMID: 38861125; PMCID: PMC11493890.

[21] Takada SH, Sampaio CAG, Allemandi W, Ito PH, Takase LF, Nogueira MI. A modified rat model of neonatal anoxia: development and evaluation by pulseoximetry, arterial gasometry and Fos immunoreactivity. J Neurosci Methods. 2011 May 15;198(1):62-69. doi: 10.1016/j.jneumeth.2011.03.009. PMID: 21439321.

[22] Moy AJ, Capulong BV, Saager RB, Wiersma MP, Lo PC, Durkin AJ, Choi B. Optical properties of mouse brain tissue after optical clearing with FocusClear™. J Biomed Opt. 2015 Sep 21;20(9):095010. doi: 10.1117/1.JBO.20.9.095010. PMID: 26388460; PMCID: PMC4963466.

[23] Ramos MAG, Leivas FG, Missio D, Cibin FWS, Guimarães ACG, Brum DS. Medium and time conservation affect follicular morphology and superoxide dismutase enzyme activity of bovine preantral follicles during storage at 4ºC. Semina Cienc Agrar. 2020 May;41(4):1227-1236. doi: 10.5433/1679-0359.2020v41n4p1227.

[24] Jacques SL. Optical properties of biological tissues: a review. Phys Med Biol. 2013 Jun 7;58(11):R37-R61. doi: 10.1088/0031-9155/58/11/R37. PMID: 23666068.

[25] Al-Juboori SI, Dondzillo A, Stubblefield EA, Felsen G, Lei TC, Klug A. Light scattering properties vary across different regions of the adult mouse brain. PLoS One. 2013 Jul 9;8(7):e67626. doi: 10.1371/journal.pone.0067626. PMID: 23874433; PMCID: PMC3706487.

[26] Festing MFW, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J. 2002 Oct 1;43(4):244-258. doi: 10.1093/ilar.43.4.244. PMID: 12391400.

[27] Nelson DL, Cox MM. Lehninger Principles of Biochemistry. 3rd ed. New York: W. H. Freeman; 2000.

[28] Karu TI. Primary and secondary mechanisms of action of visible to near-IR radiation on cells. J Photochem Photobiol B. 1999 Mar;49(1):1-17. doi: 10.1016/S1011-1344(98)00219-X. PMID: 10365442.

[29] Giacometti P, Diamond SG. Correspondence of electroencephalography and near-infrared spectroscopy sensitivities to the cerebral cortex using a high-density layout. Neurophotonics. 2014 Oct;1(2):025001. doi: 10.1117/1.NPh.1.2.025001. PMID: 25558462; PMCID: PMC4280681.Giacometti P, Diamond SG. Correspondence of electroencephalography and near-infrared spectroscopy sensitivities to the cerebral cortex using a high-density layout. Neurophotonics. 2014 Oct;1(2):025001. doi: 10.1117/1.NPh.1.2.025001. PMID: 25558462; PMCID: PMC4280681.

Assessment of gCKC, FastICA, and Swarm-Contrastive Decomposition in HD-sEMG

José O. Alves Filho
Department of Electronics and Biomedical Engineering
University of Campinas (UNICAMP)
Campinas, SP - Brazil
j290880@dac.unicamp.br

Willian A. Guedes da Silva
Department of Electronics and Biomedical Engineering
University of Campinas (UNICAMP)
Campinas, SP - Brazil
w298991@dac.unicamp.br

Giovanna Ogata Pedrassani
Department of Electronics and Biomedical Engineering
University of Campinas (UNICAMP)
Campinas, SP - Brazil
g221085@dac.unicamp.br

Leonardo Abdala Elias
Department of Electronics and Biomedical Engineering
University of Campinas (UNICAMP)
Campinas, SP - Brazil
leoelias@unicamp.br

Carina Marconi Germer
Department of Electronics and Biomedical Engineering
University of Campinas (UNICAMP)
Campinas, SP - Brazil
cgermer@unicamp.br

***Abstract*—High-density surface electromyography (HD-sEMG) decomposition provides a non-invasive approach to estimate motor unit spike trains, but algorithm comparisons can be influenced by method-specific post-processing and quality criteria. This study compared Gradient Convolution Kernel Compensation (gCKC), Fast Independent Component Analysis (FastICA), and Swarm Contrastive Decomposition (SCD) using the same recordings from eleven healthy participants performing trapezoidal isometric contractions for a broad range of intensities. A unified post-processing framework was applied to all methods, including CKC-based innervation pulse train (IPT) reconstruction, minimum discharge count, pulse-to-noise ratio (PNR), silhouette value (SIL), duplicate removal, and cross-method motor unit matching. Across visual-feedback trapezoidal tasks, FastICA identified a higher number of motor units, followed by gCKC and SCD. Accepted units showed high PNR and SIL values for all methods, with SCD presenting the highest average quality metrics after the common reconstruction procedure. Linear mixed-effects models showed significant force-level effects on MUAP amplitude and discharge rate, whereas neither measure differed significantly among decomposition methods. Overlap analysis showed that only a subset of recovered motor units was shared by all three approaches, indicating that algorithm choice can influence both motor unit yield and the motor unit sample recovered from the same HD-sEMG recordings.**

***Index Terms*—HD-sEMG, motor unit decomposition, Gradient Convolution Kernel Compensation, FastICA, Swarm Contrastive Decomposition**

This work was supported by the São Paulo Research Foundation (FAPESP; grants 2024/08318-4 and 2025/16519-2), the National Council for Scientific and Technological Development (CNPq) through a Technological and Industrial Development Fellowship (DTI-B, grant 420228/2023-3), and was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Additional support was provided by the Institutional Scientific Initiation Scholarship Program (PIBIC), and the Teaching, Research and Extension Support Fund (FAEPEX, grant 3284/25). Leonardo A. Elias is a CNPq Research Productivity Fellow, process no. 316320/2023-4.

I. Introduction

High-density surface electromyography (HD-sEMG) decomposition has become one of the most powerful techniques for investigating the neural control of movement [1]. By separating interference EMG into the discharge times of individual motor units (MUs), HD-sEMG provides non-invasive access to the output of spinal motor neurons during voluntary contractions. This approach has supported detailed studies of motor unit recruitment, discharge behavior, neural drive, fatigue, aging, and neurological disorders [2].

Several computational approaches have been proposed for HD-sEMG decomposition, each relying on different mathematical principles to separate the contributions of individual motor units from multichannel EMG recordings. Among the most established methods are Gradient Convolution Kernel Compensation (gCKC) [5] and Fast Independent Component Analysis (FastICA) [8], both based on blind source separation but employing different optimization strategies [5], [9]. More recently, Swarm Contrastive Decomposition (SCD) was proposed as an adaptive source-separation approach that combines contrast-function optimization with a peel-off strategy. Using simulated data with excitation levels ranging from 10 to 100% MVC, the method was shown to improve the detection of small or deep motor units [10], [11].

The performance of HD-sEMG decomposition algorithms is traditionally assessed using the number of identified motor units, and, more directly, measures of the quality of the estimated sources, including the Pulse-to-Noise Ratio (PNR) and the Silhouette value (SIL) [6]–[8]. These metrics provide information about decomposition quality and reliability and can be used to compare algorithms. However, considerably less attention has been paid to investigating how different decomposition methods behave when applied to the same recordings acquired under varying contraction conditions.

This question is relevant for submaximal trapezoidal contractions performed at different force levels. As contraction intensity increases, interference EMG becomes progressively more complex because of additional motor unit recruitment, changes in discharge behavior, and greater action potential overlap [3]. Consequently, algorithms may exhibit different sensitivities to signal complexity, potentially affecting motor unit yield, decomposition quality, and agreement between methods.

Therefore, this study compared gCKC, FastICA, and SCD using the same HD-sEMG recordings collected during isometric trapezoidal contractions at different force levels. The methods were compared in terms of accepted motor units, PNR, SIL, and cross-method overlap. In addition, motor unit action potential (MUAP) amplitude and mean discharge rate were analyzed to investigate whether the choice of decomposition algorithm influenced the characteristics of the identified motor units. By using identical experimental conditions and acceptance criteria, this study aims to support the interpretation of HD-sEMG decomposition results across algorithms.

II. Materials and Methods

A. Participants

Eleven healthy volunteers (6 females and 5 males; age: 26.2 ± 4.5 years; body mass: 68.1 ± 11.3 kg; height: 1.67 ± 0.07 m) participated in this study. None of the participants reported a history of neurological or musculoskeletal disorders affecting the upper limbs. All participants provided written informed consent prior to participation. The experimental protocol was approved by the Research Ethics Committee of the University of Campinas (CEP/UNICAMP; CAAE: 88364525.6.0000.5404).

B. Experimental Protocol

Participants performed isometric trapezoidal contractions of the First Dorsal Interosseous (FDI) muscle at six force levels corresponding to 5, 15, 30, 45, 60, and 70 % of their maximum voluntary contraction (MVC). Each visual-feedback contraction began with a 2-s baseline period, followed by a linear increase in force at a constant rate of 5% MVC/s, a 15-s constant-force plateau, and a linear decrease at the same rate. Consequently, the total duration of each visual-feedback contraction was 21, 25, 31, 37, 43, and 47 s for the 5, 15, 30, 45, 60, and 70 %MVC conditions, respectively. Real-time visual feedback of the target force trajectory was displayed throughout these trials to assist participants in accurately following the desired force profile. A 90-s resting period was provided between consecutive visual-feedback contractions to minimize fatigue.

C. HD-sEMG Acquisition

HD-sEMG signals were acquired from the FDI muscle using a Muovi+ acquisition system (OT Bioelettronica, Turin, Italy) equipped with a HD08MM1305 high-density electrode array. The recordings consisted of 64 monopolar channels sampled at 2000 Hz. The same HD-sEMG recordings were independently processed by each decomposition algorithm, ensuring that all methods were evaluated using identical input signals.

D. HD-sEMG Decomposition

Three decomposition algorithms were evaluated: gCKC, FastICA, and SCD. The gCKC decomposition was performed using a MATLAB implementation based on the original gCKC formulation [5], [12]. The FastICA-based implementation followed the MUedit framework [9], whose algorithm is available at https://github.com/simonavrillon/MUedit2. The SCD implementation followed the particle swarm-optimized independence estimator and HD-sEMG SCD framework [10], [11], with code available at https://github.com/AgneGris/swarm-contrastive-decomposition.

All decomposition outputs were subjected to the same post-processing pipeline and motor unit acceptance criteria.

E. Motor Unit Acceptance Criteria

Identical acceptance criteria were applied to all detected motor units, regardless of method. Specifically, motor units were required to satisfy three quality criteria: a minimum number of detected discharges, a PNR above the predefined threshold, and a minimum SIL.

Motor units with fewer than 21 detected discharges were discarded. This operational threshold was used to exclude motor units that were active only briefly during the trapezoidal contraction, often near their recruitment or derecruitment thresholds, where discharge behavior tends to be more variable. Retaining only units with at least 21 discharges also provided a minimum number of interspike intervals (ISI) for estimating discharge rate and reduced the influence of unstable quality metrics and spike-triggered MUAP estimates derived from very short discharge trains. This criterion was considered complementary to PNR and SIL, since these metrics do not fully prevent the inclusion of motor units with irregular discharge behavior near activation thresholds.

The PNR was computed from the innervation pulse train (IPT) as the logarithmic ratio between the average energy of the IPT at the detected discharge instants and the average energy of the remaining signal after excluding a ±10-ms window around each discharge [6]. The PNR expression is shown in Equation (1).

$$PNR = 10\log_{10}\left(\frac{\frac{1}{N}\sum_{i=1}^{N} IPT^2(t_i)}{\frac{1}{M}\sum_{j=1}^{M} IPT^2(t_j)}\right), \tag{1}$$

where t_i denotes the detected discharge instants, t_j corresponds to the samples considered as noise, and N and M are the numbers of discharge and noise samples, respectively. Prior to the computation, the IPT was normalized by the maximum amplitude at the detected discharge instants.

The SIL was used to quantify the separation between spike and noise clusters. Let μ_s and μ_n denote the centroids of the spike and noise samples, respectively. The within-cluster and between-cluster distances are defined in (2).

$$W = \sum_{i=1}^{N} (s_i - \mu_s)^2, \qquad B = \sum_{i=1}^{N} (s_i - \mu_n)^2. \tag{2}$$

The silhouette value was then computed using (3).

$$SIL = \frac{B - W}{\max(B, W)}. \tag{3}$$

To avoid differences in the internal source representation of each decomposition algorithm influencing PNR and SIL, the IPT was reconstructed for every accepted candidate using a Convolution Kernel Compensation (CKC)-based formulation derived from the convolutive blind source separation framework [5]. This reconstruction used the common HD-sEMG input signal, rather than algorithm-specific source estimates. For each recording, the raw HD-sEMG was band-pass filtered using a Butterworth filter between 20 and 500 Hz, and a 60-Hz notch filter was applied. The filtered signal was then temporally extended and whitened to form $\mathbf{X}_e$. For motor unit j, the correlation vector was estimated from the detected discharge instants $\tau_i^{(j)}$ using (4)

$$\mathbf{c}_{\bar{y}t_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{X}_e(\tau_i^{(j)}). \tag{4}$$

where N_j denotes the number of detected discharges of motor unit (j).

A separation vector was obtained from the inverse covariance matrix of the extended observations and normalized to unit norm, as shown in (5),

$$\mathbf{w}_j = \mathbf{C}_{yy}^{-1} \mathbf{c}_{\bar{y}t_j}, \qquad \mathbf{w}_j \leftarrow \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}. \tag{5}$$

The linear source estimate and CKC-based IPT were computed using (6).

$$\hat{t}_j = \mathbf{w}_j^T \mathbf{X}_e, \qquad \mathrm{IPT}_j = \hat{t}_j |\hat{t}_j|. \tag{6}$$

Finally, IPT_j was normalized by $\max(|IPT_j|)$ before PNR and SIL computation. This procedure was applied to gCKC, FastICA, and SCD using the discharge times detected by each method, so that quality metrics were computed from a common IPT representation.

A PNR threshold of 28 dB was adopted based on Holobar et al. [6], who associated values above 28 dB with discharge-identification sensitivities above approximately 85%. The SIL threshold was set to 0.90, consistent with prior EMG decomposition work [7]. Together with the minimum discharge count, these thresholds were used to retain reliable motor units while preserving yield.

This unified procedure reduced the influence of noisy or irregular sources and focused the comparison on motor units satisfying the same minimum quality criteria across algorithms.

F. Motor Unit Matching and Overlap Analysis

To evaluate overlap between methods and avoid counting the same underlying motor unit multiple times, all pairs of motor unit discharge trains were compared using spike-train cross-correlation.

For each motor unit pair, a coincidence histogram was computed within a temporal search window of ± 20 ms. A normalized spike-train similarity index was then calculated as

$$S = 100 \frac{\sum_{k=-L}^{L} C(k)}{\max(N_1, N_2)}, \tag{7}$$

where $C(k)$ denotes the number of coincident discharges at lag k, N_1 and N_2 correspond to the numbers of detected discharges in the two motor units, and L represents a tolerance window of ± 10 ms around zero lag.

Motor unit pairs with $S > 60\%$ were considered to correspond to the same underlying motor unit. When duplicates were identified, only the motor unit with the highest PNR was retained.

The same matching procedure was used to remove within-method duplicates and quantify cross-method overlap.

G. Outcome Measures

The decomposition methods were compared using the following outcome measures:

1) number of accepted motor units;
2) number of motor units excluded by the acceptance criteria;
3) PNR and SIL of the accepted motor units;
4) overlap between decomposition methods;
5) MUAP peak-to-peak amplitude;
6) mean discharge rate.

MUAP templates were estimated by spike-triggered averaging of the original HD-sEMG recordings using the discharge times identified by each decomposition method [4]. The channel exhibiting the largest peak-to-peak MUAP amplitude was selected to represent each motor unit. Mean discharge rate was calculated from the inter-spike intervals of each accepted motor unit [3], [4].

H. Statistical Analysis

MUAP peak-to-peak amplitude and mean discharge rate were quantified for each accepted motor unit. Descriptive analyses were performed at the motor-unit level. Statistical comparisons for MUAP amplitude and discharge rate were performed using linear mixed-effects models with decomposition method, force level (%MVC), and their interaction as fixed effects, while participant was included as random intercept. When a significant main effect of decomposition method was detected, pairwise contrasts between methods were conducted with Bonferroni correction.

Task-dependent changes in PNR and SIL were assessed within each algorithm using Friedman tests across force levels.

When significant differences were detected, pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons. To assess the robustness of cross-method matching, overlap analysis was repeated using similarity thresholds of 50%, 60%, and 70%. Statistical significance was established at $p < 0.05$.

III. Results

A. Motor Unit Yield

Across the six visual-feedback trapezoidal tasks, FastICA identified on average 7.42 accepted motor units per subject-task, followed by gCKC (5.09) and SCD (3.70) (Table I). FastICA also had the highest number of excluded units, indicating a larger initial candidate pool and a larger number of units that did not satisfy the unified quality thresholds.

TABLE I
MOTOR UNIT YIELD AND QUALITY AFTER UNIFIED ACCEPTANCE CRITERIA. VALUES ARE MEAN ± SD ACROSS VISUAL-FEEDBACK TRAPEZOIDAL TASKS. N_{MU} INDICATES THE TOTAL NUMBER OF ACCEPTED MOTOR UNITS ANALYZED FOR EACH METHOD.

Method (N_{MUs})	Accepted MUs	Excluded MUs	PNR (dB)	SIL
gCKC (336)	5.09 ± 1.42	11.03 ± 1.25	33.48 ± 1.12	0.942 ± 0.005
FastICA (490)	7.42 ± 1.30	46.03 ± 7.19	33.91 ± 0.60	0.930 ± 0.004
SCD (244)	3.70 ± 0.87	2.44 ± 0.63	36.21 ± 1.97	0.944 ± 0.006

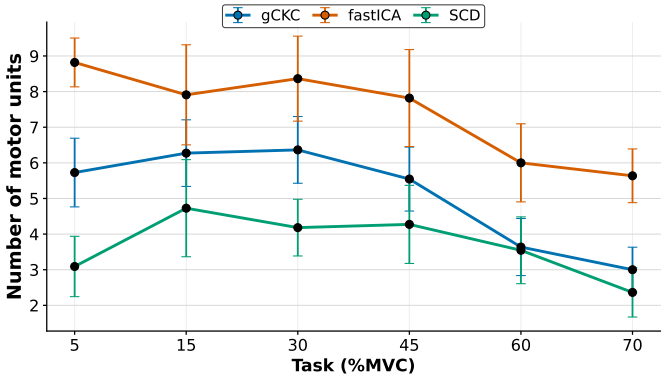


Fig. 1. Number of accepted motor units after applying the unified acceptance criteria and within-method duplicate removal. Error bars represent SEM across participants ($N = 11$).

B. Decomposition Quality

Mean PNR was highest for SCD, followed by FastICA and gCKC (Table I; Fig. 2). Mean SIL values were above the acceptance threshold for all methods, with SCD and gCKC showing slightly higher average SIL than FastICA (Fig. 3). Task-dependent changes in these quality metrics are reported in the next subsection.

C. Task-Dependent Quality Patterns

Task-dependent changes in quality metrics were assessed using Friedman tests performed separately within each decomposition method. For PNR, a significant force-level effect

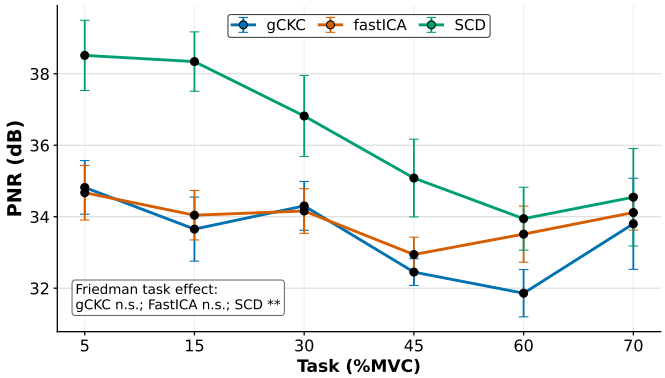


Fig. 2. Mean PNR of accepted motor units across trapezoidal tasks for each decomposition method. Error bars represent SEM across participants with accepted MUs; the number of contributing participants ranged from 9 to 11 across force levels and methods.

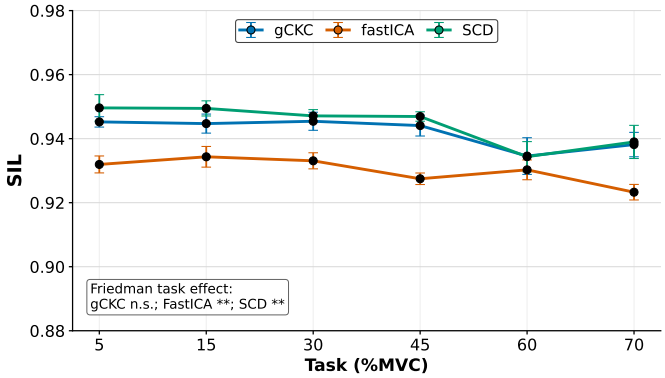


Fig. 3. Mean SIL of accepted motor units across trapezoidal tasks for each decomposition method. Error bars represent SEM across participants with accepted MUs; the number of contributing participants ranged from 9 to 11 across force levels and methods.

was observed for SCD ($\chi^2 = 16.76$, $p = 0.005$), but not for gCKC ($\chi^2 = 9.69$, $p = 0.085$) or FastICA ($\chi^2 = 4.83$, $p = 0.437$; Fig. 2). For SIL, significant force-level effects were observed for FastICA ($\chi^2 = 19.07$, $p = 0.002$) and SCD ($\chi^2 = 20.10$, $p = 0.001$), but not for gCKC ($\chi^2 = 3.49$, $p = 0.625$; Fig. 3). Because these tests were performed within each method, they describe method-specific task effects rather than a formal method-by-force-level interaction.

D. MUAP Amplitude and Discharge Rate

Accepted motor units were further characterized by best-channel MUAP peak-to-peak amplitude and mean discharge rate. MU-level median MUAP amplitudes were 0.56 for gCKC, 0.56 for FastICA, and 0.68 for SCD (Table II). The mixed-effects model indicated no significant effect of decomposition method ($F_{2,1043.24} = 0.15$, $p = 0.861$), but showed a significant force-level effect ($F_{5,1043.05} = 58.56$, $p < 0.001$) and no significant decomposition method-by-force-level interaction ($F_{10,1042.62} = 1.08$, $p = 0.375$).

Median MU-level discharge rates were 10.50 Hz for gCKC, 10.86 Hz for FastICA, and 10.54 Hz for SCD. The mixed-

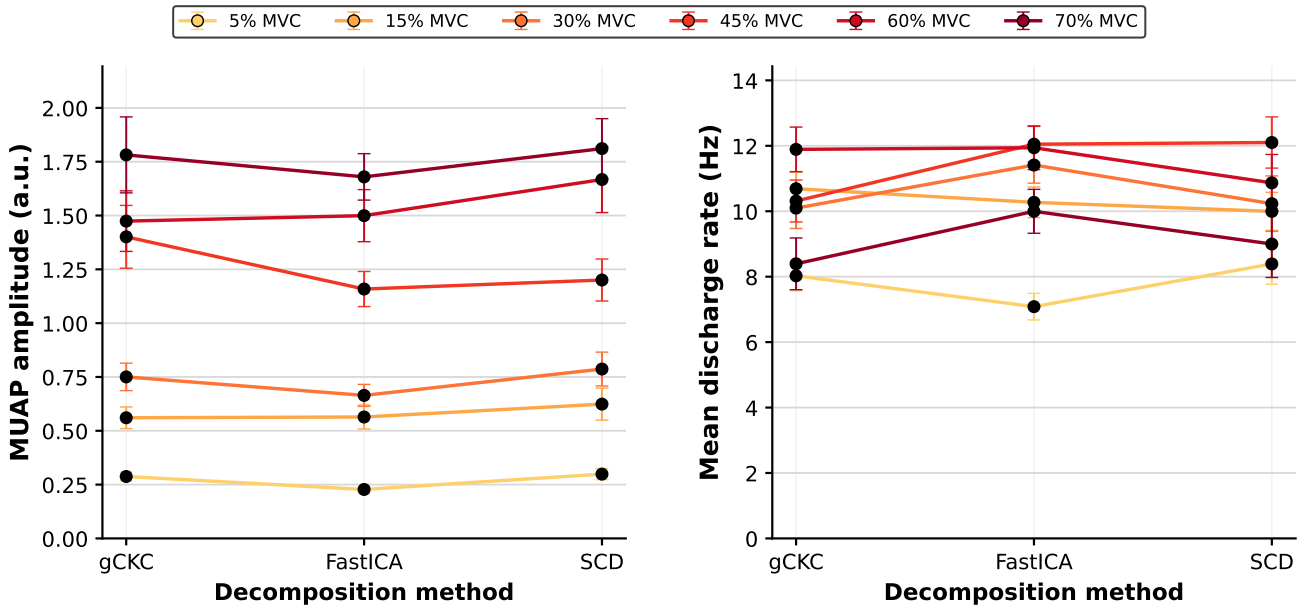


Fig. 4. MU-level MUAP peak-to-peak amplitude and mean discharge rate across decomposition methods for each force level. Values are MU-level means with SEM. Different lines represent force levels, and the x-axis shows the decomposition methods. Statistical comparisons were performed using linear mixed-effects models with decomposition method, and force level (%MVC), and their interaction as fixed effects and participant as a random intercept; no significant decomposition method-by-force-level interaction was observed for either variable.

effects model indicated no significant effect of decomposition method on discharge rate ($F_{2,1046.62} = 1.68$, $p = 0.187$), but showed a significant force-level effect ($F_{5,1045.62} = 3.81$, $p = 0.002$) and no significant decomposition method-by-force-level interaction ($F_{10,1044.25} = 1.40$, $p = 0.176$). The corresponding MU-level force-resolved summaries are shown in Fig. 4.

TABLE II
MUAP AMPLITUDE AND DISCHARGE-RATE CHARACTERISTICS OF ACCEPTED MOTOR UNITS. VALUES ARE MU-LEVEL MEDIANS WITH INTERQUARTILE RANGES.

Method	MUs	MUAP Amplitude	Mean DR (Hz)
gCKC	336	0.56 (0.32–1.30)	10.50 (5.94–13.63)
FastICA	490	0.56 (0.29–1.30)	10.86 (6.28–14.18)
SCD	244	0.68 (0.37–1.53)	10.54 (5.90–14.10)

E. Overlap Between Decomposition Methods

The three algorithms did not recover identical motor unit samples from the same recordings. Using the primary similarity threshold of 60%, an average of 2.61 motor units per subject-task were common to gCKC, FastICA, and SCD. Pairwise-only overlap was most pronounced between gCKC and FastICA (1.56 motor units per subject-task), whereas gCKC-SCD pairwise-only overlap was lower (0.14). Method-exclusive motor units were also observed, with gCKC-only, FastICA-only, and SCD-only components averaging 0.89, 2.56, and 0.28 motor units per subject-task, respectively (Fig. 5). Sensitivity analysis showed that the overall interpretation was stable when the similarity threshold was varied between 50% and 70%. Across these thresholds, the number of motor units shared by all three methods changed only modestly (2.66, 2.61, and 2.56 motor units per subject-task), while FastICA-only estimates remained the largest method-exclusive component (2.42, 2.56, and 2.72 motor units per subject-task).

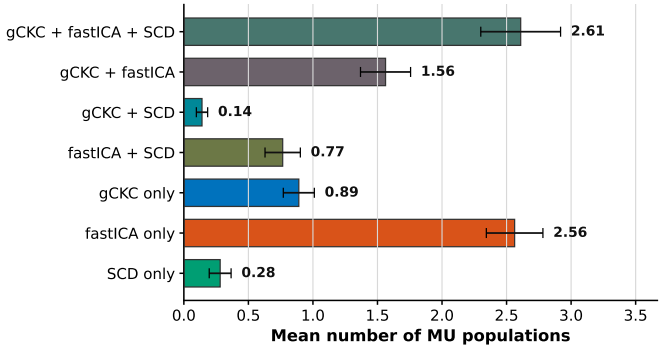


Fig. 5. Group-level overlap patterns among motor units identified by gCKC, FastICA, and SCD after applying the same acceptance and matching criteria.

IV. Discussion

This study compared gCKC, FastICA, and SCD using identical HD-sEMG recordings and a common CKC-based post-processing framework. HD-sEMG decomposition is widely used to estimate motor unit discharge behavior non-invasively, but algorithmic differences in source extraction and post-processing can affect the set of motor units available for

physiological interpretation [1], [4]. The main finding was that the algorithms produced partly overlapping but non-identical motor unit samples. FastICA yielded the largest number of accepted motor units after filtering and duplicate removal, gCKC showed an intermediate yield, and SCD yielded fewer accepted units under the present criteria. At the same time, accepted SCD units showed high PNR and SIL values, indicating that the lower SCD yield reflected the number of motor units passing the current acceptance pipeline rather than poor quality of the retained motor units.

The larger number of excluded FastICA candidate motor unit estimates indicates that the initial number of extracted sources should not be interpreted alone as decomposition performance. This is consistent with the use of signal-based quality metrics, such as PNR, and cluster-separation measures, such as SIL, to screen decomposition outputs before physiological analysis [6], [7]. In this context, the larger set of FastICA candidate motor unit estimates should be interpreted together with the number of motor units that remained after quality filtering and duplicate removal.

The MUAP and discharge-rate analyses indicate that method-related differences in accepted motor unit properties were modest compared with the differences observed in yield, quality metrics, and overlap structure. The mixed-effects models, which accounted for repeated observations within participants, did not identify a significant method effect for either MUAP amplitude or discharge rate, nor a significant decomposition method-by-force-level interaction. Both variables varied across force levels, which is expected because changes in contraction intensity are associated with motor unit recruitment, rate coding, and changes in the complexity of the interference EMG signal [3]. Thus, separating force levels in the analysis was important, but the present results do not indicate that the three algorithms sampled motor units with systematically different MUAP amplitudes or discharge rates.

Although MUAP amplitude is only an indirect proxy for motor unit size or depth, it was examined because SCD has been reported to improve the detection of small or deep motor units, particularly in simulated analyses [10], [11]. In the present data, however, the MUAP amplitude distribution of the units retained by SCD was not lower than those observed for gCKC or FastICA, and the mixed-effects model did not show a significant effect of decomposition method. In addition, SCD showed the lowest accepted-unit yield in the present results, which may indicate that, under the current acceptance criteria, this method behaved more selectively. This interpretation is consistent with the aim of the present pipeline, which prioritized the quality and reliability of the retained motor units over the total number of extracted sources.

The overlap results are particularly important for interpreting algorithm selection. If all methods recovered the same underlying motor units from the HD-sEMG recordings, differences in yield would mainly reflect missed detections, repeated estimates, or differences in quality filtering. Instead, the present results show that each method can contribute exclusive motor unit samples even after quality filtering and matching. This suggests that decomposition algorithms may be sensitive to different signal features or source characteristics, which is plausible given their different blind-source-separation strategies [5], [8]–[10]. The sensitivity analysis indicated that this interpretation was not strongly dependent on the specific 60% similarity threshold used for the primary overlap analysis. For studies that use motor unit samples to infer recruitment, discharge behavior, or neural drive, this finding emphasizes that algorithm choice can influence not only the number of accepted units but also which units are included in downstream analyses.

A. Limitations

This study has some limitations. First, the comparison was restricted to visual-feedback trapezoidal isometric contractions of the FDI muscle in healthy participants, and therefore the findings may not generalize to other muscles, contraction types, or clinical populations. Second, although MUAP amplitude and discharge rate were analyzed using mixed-effects models to account for repeated observations within participants, the yield and overlap analyses still depend on the accepted motor unit samples obtained under the selected quality criteria. Third, PNR and SIL provide signal-based and cluster-separation measures of decomposition quality, but they do not independently verify whether each detected discharge corresponds to the true activity of a motor unit [6], [7]. Validation using simulated signals with known motor unit activity or concurrent intramuscular recordings would provide a stronger independent reference. Finally, although the sensitivity analysis showed stable overlap patterns when the similarity threshold was varied from 50% to 70%, future work should also examine the influence of PNR and SIL thresholds on motor unit yield, quality, and cross-method agreement.

V. Conclusion

Using a unified CKC-based quality assessment framework, gCKC, FastICA, and SCD showed different motor unit yield, quality profiles, and overlap patterns when applied to the same HD-sEMG recordings. FastICA produced the largest number of accepted motor units, whereas SCD yielded fewer units but maintained high PNR and SIL values after the common post-processing procedure. MUAP amplitude and mean discharge rate varied across force levels, but mixed-effects models showed that neither variable differed significantly among decomposition methods, and no method-by-force-level interaction was observed. These findings indicate that, under the present acceptance criteria, differences among algorithms were mainly reflected in the number, quality, and overlap of the recovered motor unit samples rather than in MUAP amplitude or discharge-rate characteristics. Therefore, standardized post-processing and cross-method matching are important for interpreting comparisons among HD-sEMG decomposition algorithms.

Acknowledgment

The authors thank all participants and the research team involved in data collection and signal processing.

REFERENCES

[1] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface EMG," *Journal of Applied Physiology*, vol. 96, no. 4, pp. 1486–1495, 2004, doi: 10.1152/japplphysiol.01070.2003.

[2] J. Duchateau and R. M. Enoka, "Human motor unit recordings: origins and insight into the integrated motor system," *Brain Research*, vol. 1409, pp. 42–61, 2011, doi: 10.1016/j.brainres.2011.06.011.

[3] R. M. Enoka and J. Duchateau, "Rate coding and the control of muscle force," *Cold Spring Harbor Perspectives in Medicine*, vol. 7, no. 10, 2017, doi: 10.1101/cshperspect.a029702.

[4] A. Del Vecchio, A. Holobar, D. Falla, F. Felici, and R. M. Enoka, "Tutorial: Analysis of motor unit discharge characteristics from high-density surface EMG signals," *Journal of Electromyography and Kinesiology*, vol. 53, p. 102426, 2020, doi: 10.1016/j.jelekin.2020.102426.

[5] A. Holobar and D. Zazula, "Multichannel blind source separation using convolution kernel compensation," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4487–4496, 2007.

[6] A. Holobar, M. A. Minetto, and D. Farina, "Accurate identification of motor unit discharge patterns from high-density surface EMG and validation with a novel signal-based performance metric," *Journal of Neural Engineering*, vol. 11, no. 1, 2014.

[7] F. Negro, S. Muceli, A. M. Castronovo, A. Holobar, and D. Farina, "Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation," *Journal of Neural Engineering*, vol. 13, no. 2, 2016.

[8] C. Dai and X. Hu, "Independent component analysis based algorithms for high-density electromyogram decomposition: Experimental evaluation," *Computers in Biology and Medicine*, vol. 108, pp. 42–51, 2019.

[9] S. Avrillon, F. Hug, S. N. Baker, C. Gibbs, and D. Farina, "Tutorial on MUedit: An open-source software for identifying and analysing the discharge timing of motor units from electromyographic signals," *Journal of Electromyography and Kinesiology*, vol. 77, p. 102886, 2024, doi: 10.1016/j.jelekin.2024.102886.

[10] A. Grison, A. K. Clarke, S. Muceli, J. Ibanez, A. Kundu, and D. Farina, "A particle swarm optimised independence estimator for blind source separation of neurophysiological time series," *IEEE Transactions on Biomedical Engineering*, pp. 1–11, 2024, doi: 10.1109/TBME.2024.3446806.

[11] A. Grison, I. Mendez Guerra, A. K. Clarke, S. Muceli, J. Ibanez, and D. Farina, "Unlocking the full potential of high-density surface EMG: novel non-invasive high-yield motor unit decomposition," *The Journal of Physiology*, vol. 603, no. 8, pp. 2281–2300, 2025.

[12] A. Holobar and D. Zazula, "Gradient convolution kernel compensation applied to surface electromyograms," in *Independent Component Analysis and Signal Separation*, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 617–624.

Hemispheric Asymmetry and Sex-Related Differences in Gray Matter Morphometry in Deep Brain Stimulation Candidates with Parkinson's Disease: An Exploratory Voxel-Based Morphometry Study

1st João Victor Barbosa de Macedo
Instituto Internacional de Neurociências Edmond e Lily Safra (IIN-ELS)
Instituto Santos Dumont (ISD)
Macaíba, Brazil
joaovictorbm2001@gmail.com

***Abstract*—Abstract—Parkinson's disease (PD) is characterized by progressive degeneration of subcortical motor circuits, with direct implications for deep brain stimulation (DBS) planning. Two clinical features of PD are particularly relevant to individualized neuromodulation: the well-documented lateralization of motor symptoms and the sex-related differences in prevalence and disease progression. Structural correlates of both dimensions may constitute complementary biomarkers for surgical target selection. Objective: To investigate hemispheric asymmetry and sex-related differences in gray matter volume among DBS candidates with PD, to test whether these two dimensions interact, and to evaluate the influence of age at onset and disease duration on the resulting morphometric profile. Methods: Structural T1-weighted MRI scans from 38 individuals with PD (age: 60.3 ± 7.1 years; 25 men, 13 women; disease duration: 11.5 ± 3.9 years) were obtained from a public dataset and processed with voxel-based morphometry (CAT12). Hemispheric asymmetry was assessed with the paired Wilcoxon test; sex-related differences were assessed with the Mann-Whitney test, both before and after adjustment for total intracranial volume (TIV); a formal sex-by-hemisphere interaction was tested by comparing a per-subject asymmetry index between sexes; all comparisons were corrected for multiple testing (Benjamini-Hochberg). Results: Significant hemispheric asymmetries survived correction in the putamen (left dominance), supplementary motor cortex (left dominance), hippocampus (right dominance), and caudate nucleus (right dominance). Men showed larger unadjusted regional volumes than women across several regions, but TIV was itself 13.9% larger in men ($p<0.001$), and none of the regional sex differences remained significant after TIV adjustment. The formal asymmetry-index-based test found no significant sex-by-hemisphere interaction in any region. Age at onset correlated positively with right globus pallidus volume ($p=0.035$); no correlations were found with disease duration. Discussion: Hemispheric asymmetry is a robust structural feature of this cohort, distinguishable from global brain-size effects, whereas the apparent sex-related regional differences appear largely attributable to global brain-size scaling rather than PD-specific, region-level pathology. Conclusions: DBS candidates with PD present detectable, TIV-independent hemispheric asymmetry in motor and limbic structures, including the pallidum, a primary surgical target, and age at onset relates to pallidal volume. Sex-related regional differences and a sex-by-hemisphere interaction were not supported once appropriate covariate adjustment and formal interaction testing were applied, underscoring the importance of controlling for global brain size in morphometric studies of sex effects.**

***Index Terms*—Parkinson's Disease; Deep Brain Stimulation; Brain Asymmetry; Sex Dimorphism**

I. Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, clinically characterized by resting tremor, rigidity, bradykinesia, and postural instability, arising from progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta. Two clinical features of PD are especially relevant to the planning of individualized surgical treatment. The first is the marked asymmetry of motor symptoms: the majority of patients present with unilateral symptom predominance at disease onset, and this asymmetry tends to persist, albeit attenuated, throughout disease progression. This observation supports the hypothesis that nigrostriatal degeneration in PD does not occur symmetrically between hemispheres, and that this biochemical and functional asymmetry should be paralleled by measurable structural asymmetries on neuroimaging. The second feature is the well-established sex difference in PD: men are diagnosed roughly 1.5 times more often than women, and accumulating evidence points to distinct trajectories of motor and non-motor symptom progression between sexes, potentially reflecting differences in nigrostriatal dopaminergic density, hormonal neuroprotection, and genetic susceptibility.

Deep brain stimulation (DBS) is currently the primary surgical option for medication-refractory motor symptoms in PD, with the subthalamic nucleus (STN) and the internal globus pallidus (GPi) as the most commonly used targets. Target selection and lateralization, that is, the decision of which hemisphere to prioritize in unilateral implants, or how to asymmetrically adjust parameters in bilateral implants.

currently relies primarily on clinical criteria, such as the side of symptom predominance, complemented by stereotactic planning based on population atlases that do not typically account for individual variability in either hemispheric laterality or sex. In this context, individualized structural biomarkers capturing both dimensions have the potential to complement this decision-making process, particularly as voxel-based morphometry (VBM) becomes an increasingly accessible and reproducible tool in clinical neuroimaging workflows.

Prior structural neuroimaging studies in PD have demonstrated volumetric atrophy in basal ganglia nuclei, particularly the putamen, with magnitude associated with disease severity and progression [3,4]. More recent investigations combining functional and structural MRI have reinforced the existence of measurable interhemispheric asymmetry already in early-stage PD, both in functional connectivity metrics and regional volumetry [2]. In parallel, a growing body of literature has begun to characterize sex-specific gray matter correlates of PD, including sex differences in anxiety-related gray matter underpinnings in de novo PD [9] and longitudinal patterns of gray matter change that may differ by sex [10]. However, hemispheric asymmetry and sex dimorphism have generally been studied in isolation, and few studies have examined either dimension specifically in populations of DBS candidates — a clinically relevant subgroup, since these patients are already in an advanced disease stage and are precisely those for whom structural information would have the most direct practical application in surgical planning [6,7].

The present study addresses this gap by jointly investigating hemispheric asymmetry and sex-related differences in gray matter volume within the same cohort of DBS candidates with PD, with particular attention to motor structures (putamen, supplementary motor cortex), the surgical target structure (globus pallidus), and limbic structures (hippocampus, amygdala, caudate nucleus). Because both analyses were performed on the same sample, this design also allows a preliminary, exploratory look at whether the structures implicated by hemispheric laterality overlap with those implicated by sex, a question with direct relevance for how individualized these two axes of variability might need to be in DBS planning.

II. Objectives

The objectives of this study were:
(i) to investigate the presence of hemispheric asymmetry in the volume of gray matter among DBS candidates with PD;
(ii) to investigate differences related to sex in the volume of gray matter within the same cohort;
(iii) to evaluate the influence of age at the time of onset and duration of the disease on the resulting regional morphometric profile.

III. Methods

1) Sample: Structural T1-weighted MRI scans from 38 individuals with a clinical diagnosis of PD (age: 60.3 ± 7.1 years; 25 men, 13 women; disease duration: 11.5 ± 3.9 years) were obtained from the *London Health Sciences Centre Parkinson's Disease Dataset* (LHSCPD), a publicly available dataset hosted on OpenNeuro [1]. Men and women in the sample did not differ significantly in age (62.1 ± 5.6 vs. 59.1 ± 5.4 years), disease duration (12.1 ± 3.9 vs. 10.5 ± 4.1 years), or age at onset (all $p>0.05$), reducing the likelihood that these demographic variables confounded the sex comparisons reported below.

2) Image acquisition and preprocessing: Structural T1-weighted images were processed with the Computational Anatomy Toolbox (CAT12, version 26.0.rc3, revision 2874), running within a standalone SPM/MATLAB Runtime (R2023b) environment. Preprocessing followed the standard CAT12 pipeline: spatial adaptive non-local means (SANLM) denoising, bias-field correction, tissue segmentation into gray matter, white matter, and cerebrospinal fluid, and nonlinear spatial registration to the MNI152NLin2009cAsym template via Geodesic Shooting, resampled to 1.5 mm isotropic resolution. Regional gray matter volumes for the putamen, globus pallidus, caudate nucleus, hippocampus, amygdala, and supplementary motor cortex were extracted bilaterally in native space using the Neuromorphometrics atlas implemented in CAT12, which registers the atlas to each subject's native anatomy to yield absolute regional volumes. Total intracranial volume (TIV) was computed for each subject as the sum of gray matter, white matter, and cerebrospinal fluid volumes. Image quality was assessed automatically via CAT12's built-in quality metrics (noise-to-contrast ratio, inhomogeneity-to-contrast ratio, and the composite Image Quality Rating).

Statistical analysis:

Two complementary analyses were performed on the same cohort. First, hemispheric asymmetry was assessed by comparing homologous left and right ROI volumes using the paired Wilcoxon test, given the expected non-parametric distribution of regional volumes in a moderate-sized sample. Second, sex-related differences were assessed by comparing ROI volumes between men and women using the Mann-Whitney test, with effect size (r) reported alongside p-values to contextualize the magnitude of each difference beyond statistical significance alone. The relationship between regional volumes and the clinical variables age at onset and disease duration was evaluated by correlation analysis (Spearman) . Statistical significance was set at $p< 0.05$ throughout. .

IV. Results

1) Hemispheric asymmetry: Table I summarizes the hemispheric asymmetry findings across the sample as a whole.

Age at onset correlated positively with right globus pallidus volume ($p=0.035$), suggesting that patients with later disease onset present with greater volumetric preservation of this structure relative to those with earlier onset. No statistically significant correlations were observed between disease duration and volume in any of the regions assessed, including the DBS target structures.

TABLE I
HEMISPHERIC ASYMMETRY OF GRAY MATTER VOLUME

Region	**Direction of dominance**	**p-value**
Putamen	Left	0.002
Supplementary motor cortex	Left	0.008
Hippocampus	Right	<0.001
Caudate nucleus	Right	0.03

TABLE II
SEX-RELATED DIFFERENCES IN GRAY MATTER VOLUME (MEN VS. WOMEN), RAW AND TIV-ADJUSTED

Region	**Diff. (M>F)**	**Raw** p	r	**TIV-adj.** p
Putamen, left	+16.4%	0.012	0.51	0.62
Putamen, right	+17.0%	0.008	0.53	0.27
Hippocampus, left	+11.9%	0.002	0.64	1.00
Hippocampus, right	+11.1%	0.004	0.58	0.64
Amygdala, left	—	0.012	—	0.98
Amygdala, right	—	0.023	—	0.71
SMC, right	+11.0%	0.021	0.46	0.67
Caudate, right	+10.4%	0.049	0.40	0.34

2) Sex-related differences : Table II summarizes the unadjusted sex-related differences in gray matter volume observed within the same cohort. Across all regions in which a significant unadjusted difference was found, men presented with greater gray matter volume than women, with effect sizes ranging from small-to-moderate to large.

However, total intracranial volume (TIV) itself differed significantly by sex, with men presenting 13.9% larger TIV than women (1493 ± SD vs. 1311 ± SD cm³, p<0.001). Given that TIV was a strong predictor of regional volume across nearly all structures examined (p<0.001 for all regions except the pallidum), we re-tested each regional comparison after adjusting for TIV: for each region, gray matter volume was regressed on TIV, and the resulting TIV-independent residuals were compared between sexes with the Mann-Whitney test. Table II reports these TIV-adjusted comparisons. None of the twelve regional sex differences remained statistically significant after TIV adjustment, including after Benjamini-Hochberg correction for multiple comparisons; most adjusted p-values exceeded 0.6. A sensitivity analysis using a proportional (volume/TIV ratio) adjustment instead of residualization produced a consistent pattern, with no region showing a stable, method-independent effect. These findings indicate that the unadjusted sex differences reported in Table II are largely attributable to global brain-size scaling rather than to region-specific, PD-related structural divergence between sexes.

3) Formal test of sex-by-hemisphere interaction:: To directly test whether hemispheric asymmetry differs by sex, a per-subject, per-region asymmetry index was computed as AI = 100 × (Right Left)/[(Right + Left)/2], and compared between men and women with the Mann-Whitney test for each of the six regions examined. This approach respects the within-subject, paired structure of hemispheric data, which a conventional two-way ANOVA treating hemisphere as an independent factor would not. No region showed a significant sex-by-hemisphere interaction (all p>0.28; all p>0.77 after Benjamini-Hochberg correction), indicating that, in this cohort, the degree and direction of hemispheric asymmetry did not differ measurably between men and women. This formal result qualifies the descriptive observation that the putamen, hippocampus, and caudate nucleus appeared in both the asymmetry and (unadjusted) sex analyses: rather than reflecting a genuine interaction between these two clinical dimensions, this overlap is better interpreted as coincidental co-occurrence of two independently distributed sources of variability, one of which (the sex-related component) did not survive appropriate covariate adjustment.

V. DISCUSSION

The pattern of asymmetry observed in the putamen and supplementary motor cortex is compatible with the hypothesis of asymmetric nigrostriatal degeneration underlying the classic clinical lateralization of PD. The putamen, receiving the bulk of dopaminergic nigrostriatal projections related to motor control, is one of the structures most consistently associated with progressive atrophy in PD in prior VBM studies, and the magnitude of its atrophy has been proposed as a possible clinical evaluation index for the disease [4,5]. The left-hemisphere dominance observed here for the putamen and supplementary motor cortex is consistent with recent structural and functional asymmetry findings reported in early-stage PD [2], although the specific direction of asymmetry (left- versus right-predominant) varies across studies and populations, likely reflecting the distribution of the clinically predominant side in each cohort.

The right-hemisphere dominance observed in the hippocampus and caudate nucleus falls within a territory distinct from the primary motor circuit, aligning instead with the limbic and associative circuits of the basal ganglia. This finding is compatible with the literature describing non-motor involvement in PD, including cognitive and limbic dysfunction, as a process partially dissociated, in terms of laterality, from motor involvement per se.

The apparent sex-related differences observed in the unadjusted analysis initially seemed consistent with proposed sex differences in nigrostriatal dopaminergic density and disease progression rate reported elsewhere in the literature [10]. However, once total intracranial volume was accounted for, none of these regional differences remained statistically distinguishable from chance (Section IV.2). This pattern is consistent with the well-established observation that men have, on average, larger cranial and brain volumes than women, and it illustrates why covariate adjustment for global brain size is essential before attributing regional volumetric differences specifically to sex, rather than to global scaling. This does not rule out the existence of genuine, region-specific sex effects

in PD-related neurodegeneration; it indicates that this study, at its current sample size and with the covariate adjustment now applied, did not detect them.

The formal test of a sex-by-hemisphere interaction, based on a per-subject asymmetry index, was similarly negative across all regions examined (Section IV.3). This result argues against interpreting the overlap of the putamen, hippocampus, and caudate nucleus across the two analyses as evidence that hemispheric laterality and sex jointly shape structural variability in this cohort. Instead, the overlap is more parsimoniously explained as two separate axes of variability, one robust (hemispheric asymmetry) and one that did not survive proper covariate adjustment (sex), coincidentally implicating an overlapping set of regions. Consequently, the present data support individualizing DBS planning by hemispheric laterality, for which the evidence here is robust to TIV and survives multiple-comparisons correction, but do not currently support individualizing planning by sex on the basis of regional structural differences.

The most direct link to DBS planning practice remains the correlation between age at onset and right globus pallidus volume identified in the asymmetry analysis. As the GPi is one of the most frequently used surgical targets in PD, particularly in patients with prominent dyskinesia or in whom the STN is considered a higher-risk target [6,7], the existence of a systematic association between a readily available clinical variable (age at onset) and the volume of the target structure itself suggests that early- versus late-onset PD phenotypes may translate into measurable anatomical differences in the surgical target. Combined with the robust hemispheric asymmetry findings reported above, this reinforces the broader argument, already present in the DBS patient-and-target-selection literature, that target choice and fine-tuning should be individualized, and that preoperative structural neuroimaging, read jointly with clinical variables such as symptom laterality and age at onset, may add complementary information to the clinical criteria traditionally used [6,7]. The present data do not, however, support extending this individualization to sex on structural grounds, given the absence of TIV-independent regional sex differences reported above.

The absence of a correlation between disease duration and regional volume, including in the target structures, deserves careful discussion, as it may reflect either a genuine limitation in the sensitivity of the cross-sectional design employed or a real feature of the atrophy trajectory in PD, in which chronological disease duration is a weaker predictor of structural atrophy than motor severity or clinical subtype. Without longitudinal data or a concurrent measure of motor severity (e.g., MDS-UPDRS-III score at the time of acquisition), it is not possible to distinguish between these two explanations with the data currently reported.

1) A. Limitations: This study presents limitations that should be explicitly acknowledged. First, the overall sample size (n=38) is modest for voxel-based morphometry analyses, and the sex-comparison subgroup is particularly unbalanced (25 men vs. 13 women), which limits statistical power for the sex analysis and increases the instability of effect-size estimates; this imbalance is especially relevant to the null TIV-adjusted and interaction results reported above, since a modest, unbalanced sample has limited power to detect a genuine but smaller sex-related or interaction effect, were one present. Second, Benjamini-Hochberg false discovery rate correction was applied across the combined set of asymmetry, sex-related, and interaction tests reported above; this procedure is less conservative than family-wise approaches such as Bonferroni, under which the right caudate nucleus asymmetry finding (raw p=0.03) would not survive, and this discrepancy should be kept in mind when weighing the strength of individual findings. Third, the design is cross-sectional and does not include a healthy control group matched for age and sex, which prevents disentangling PD-specific asymmetries from those already described in the general population. Fourth, the absence of information on the clinically predominant side of motor symptoms limits the ability to directly test the central hypothesis of clinical-structural lateral- ization, leaving this association inferred only by analogy with the literature; similarly, no clinical motor subtype (tremor- dominant versus rigidity/bradykinesia-dominant) was available to contextualize the sex-related findings. Fifth, a formal test of the sex-by-hemisphere interaction, based on a per-subject asymmetry index compared between men and women, was performed and yielded no significant interaction in any region (Section IV.3); the modest and unbalanced sample nonetheless limits the power of this test to detect a genuine but smaller interaction effect, were one present. Sixth, in this reanalysis of a general-purpose public dataset, potentially relevant variables, such as medication state during image acquisition, could not be controlled. Finally, regional volumes were extracted via atlas-based parcellation in native space rather than a fully voxel-wise, smoothed analysis; because segmen- tation and registration relied on the asymmetric MNI152NLin2009cAsym template, template-driven left-right registration bias cannot be fully excluded as a contributor to the hemispheric asymmetry findings, and future replications should consider a symmetric template (e.g., MNI152NLin2009cSym) to isolate genuine anatomical asym- metry from registration artifact. Additionally, although TIV was incorporated as a covariate in the sex comparisons via both residualization and a proportional volume/TIV ratio (Section IV.2), this adjustment assumes a linear relationship between TIV and regional volume; nonlinear or region-specific scaling relationships, which residualization does not fully capture, cannot be excluded, and the absence of TIV-independent sex-related regional effects reported here should be interpreted with this caveat in mind.

VI. Conclusion

DBS candidates with Parkinson's disease present detectable, TIV-independent hemispheric asymmetry in motor structures (putamen, supplementary motor cortex) and limbic structures (hippocampus, caudate nucleus), with a left-motor/right-limbic pattern characterizing the sample as a whole; this asymmetry survived correction for multiple comparisons. Of particular

relevance to surgical practice, the volume of the right globus pallidus, one of the primary DBS target structures in PD, correlated with age at onset. In contrast, sex-related regional volume differences observed in unadjusted comparisons did not survive adjustment for total intracranial volume, and a formal test found no evidence of a sex-by-hemisphere interaction; taken together, these results suggest that the apparent sex effects in this cohort largely reflect global brain-size scaling rather than region-specific, PD-related structural divergence. These findings should still be interpreted as exploratory, given the sample size and the unbalanced sex distribution, which limits power to detect genuine region-specific sex effects were they present. Hemispheric asymmetry, robust to TIV adjustment and multiple-comparisons correction, remains a more directly actionable structural biomarker for individualized DBS planning at this stage than the sex-related dimension originally proposed; validation in larger, sex-balanced, TIV-matched cohorts is required before either dimension can be proposed as a clinically applicable biomarker.

REFERENCES

[1] M. Abbass, G. Gilmore, A. Taha, R. Chevalier, M. Jach, T. M. Peters, A. R. Khan, and J. C. Lau, "London Health Sciences Centre Parkinson's Disease Dataset (LHSCPD)," OpenNeuro, 2023. DOI: 10.18112/openneuro.ds004471.v1.0.1

[2] Y. Liu et al., "Exploring brain asymmetry in early-stage Parkinson's disease through functional and structural MRI," *CNS Neurosci. Ther.*, vol. 30, no. 7, p. e14874, 2024. DOI: 10.1111/cns.14874

[3] A. Cherubini et al., "MRI correlates of Parkinson's disease progression: a voxel based morphometry study," *Am. J. Neuroradiol.*, vol. 35, no. 8, pp. 1541–1547, 2014. DOI: 10.3174/ajnr.A3868

[4] K. Kinoshita et al., "Putamen atrophy is a possible clinical evaluation index for Parkinson's disease using human brain MRI," *J. Imaging*, vol. 8, no. 11, p. 299, 2022. DOI: 10.3390/jimaging8110299

[5] C. Gaser, R. Dahnke, P. M. Thompson, et al., "CAT: a computational anatomy toolbox for the analysis of structural MRI data," *GigaScience*, vol. 13, p. giae049, 2024. DOI: 10.1093/gigascience/giae049

[6] K. A. Ehgoetz Martens et al., "New perspectives of deep brain stimulation indications for Parkinson's disease: a critical review," *J. Clin. Med.*, vol. 13, no. 13, p. 3887, 2024. DOI: 10.3390/jcm13133887

[7] M. C. Keuken et al., "Considerations for patient and target selection in deep brain stimulation surgery for Parkinson's disease," in *StatPearls*. Treasure Island: StatPearls Publishing, 2023. PMID: 30855883

[8] J. Ashburner and K. J. Friston, "Voxel-based morphometry — the methods," *NeuroImage*, vol. 11, no. 6, pp. 805–821, 2000. DOI: 10.1006/nimg.2000.0582

[9] Y. Liu et al., "Sex-specific gray matter underpinnings of anxiety and their mediating mechanisms in de novo Parkinson's disease," *Front. Aging Neurosci.*, vol. 18, p. 1769488, 2026. DOI: 10.3389/fnagi.2026.1769488

[10] J. B. Pereira et al., "Longitudinal study of gray matter changes in Parkinson disease," *Am. J. Neuroradiol.*, vol. 36, no. 12, pp. 2219–2226, 2021. DOI: 10.3174/ajnr.A4158

Hybrid EEG-EMG Fusion Comparison for Executed Gesture Dataset Classification

Diego de Sá Dias
Biomedical Engineering Program
Institute of Science and Technology
Federal University of São Paulo (UNIFESP)
São José dos Campos (SP), Brazil
ORCID: 0009-0005-1747-4573

Jean Faber
Department of Physics
Federal University of Sergipe (UFS)
São Cristovão, Brazil
Graduate Program in Neurology and Neuroscience
Federal University of São Paulo (UNIFESP)
São Paulo, Brazil
ORCID: 0000-0002-2129-8251

***Abstract*—Hybrid EEG-EMG interfaces are promising for motor control applications because they combine cortical and peripheral information related to movement. However, it remains unclear whether hybrid models consistently outperform strong unimodal baselines when different fusion strategies are evaluated under the same subject-specific protocol. This study compared EEG-only, EMG-only, and hybrid EEG-EMG pipelines for the classification of seven executed upper-limb gestures using a public dataset with 11 subjects, 8 EEG channels, and 8 EMG channels. EEG was resampled to 200 Hz to match EMG, and the first 5 s of each segment were used as execution epochs. The evaluated approaches included feature-level fusion, decision-level fusion, corticomuscular coherence (CMC)-based features, and Riemannian covariance-based models. All pipelines were assessed using subject-specific six-fold grouped cross-validation, with macro F1-score, balanced accuracy, and classification time per epoch as the main metrics. The highest mean performance was obtained by the EMG Riemannian pipeline with an RBF SVM, achieving a macro F1-score of 0.925 ± 0.043 and balanced accuracy of 0.946 ± 0.030. However, planned post-hoc comparisons did not confirm a significant advantage over feature-level or decision-level fusion after correction for multiple comparisons. Feature-level and decision-level fusion were competitive, reaching macro F1-scores of 0.877 ± 0.072 and 0.863 ± 0.091, respectively, with lower offline classification time per epoch. In contrast, CMC-based models performed close to chance level, although this result should be interpreted with caution because coherence estimation is highly dependent on EEG-EMG temporal synchronization. These findings indicate that EEG-EMG fusion can be useful, but it is not automatically superior to unimodal models. For executed gesture classification, the benefit of hybridization depends strongly on the fusion strategy, feature representation, and computational cost.**

***Index Terms*—Hybrid brain-computer interface, human-machine interface, EEG-EMG fusion, executed gesture classification, corticomuscular coherence, Riemannian geometry**

I. Introduction

Although prosthetic technologies have advanced substantially in both research and industry, many people living with limb amputation still abandon or use their devices only sporadically [1]. This limited adoption is associated with several factors, including cost, comfort, aesthetics, ergonomics, restricted functionality, and the difficulty of achieving intuitive control and neurocognitive embodiment [2], [3]. In this context, Human-Machine Interfaces (HMI) based on physiological signals have been investigated as a way to decode intended or executed movements more reliably. Among these signals, Electromyography (EMG) and Electroencephalography (EEG) are commonly investigated, as EMG directly reflects muscle activation [4] and EEG provides cortical information directly related to the intention of movement [5].

Despite promising results, using only one of these approaches has important limitations. EMG-based interfaces provide direct access to peripheral muscle activation, but their performance can be affected by electrode placement, muscle fatigue, crosstalk, and the availability of residual muscles, which may constrain intuitive control in amputees [6]. On the other hand, noninvasive EEG can provide information related to cortical motor processes, but it is typically noisy, nonstationary, and limited in the number of reliable control commands that can be decoded in real time [7]. In this context, hybrid Brain-Computer Interfaces (hBCIs) aim to combine multiple signals or control pathways to increase reliability, robustness and control dimensionality, thereby mitigating some limitations of single-signal HMIs [7], [8].

Hybrid EEG-EMG interfaces have been explored in motor-related applications, including direct EEG-EMG fusion, probabilistic fusion, and multimodal control of robotic devices [9]–[11]. Nevertheless, the feasibility of multimodal fusion does not guarantee that hybrid models consistently outperform strong unimodal baselines. The benefit of hybridization may depend on the fusion level, feature representation, and classifier architecture. Thus, a direct comparison of feature-level, decision-level, coherence-based, and Riemannian EEG-EMG pipelines under the same subject-specific validation protocol remains relevant [7], [12].

Therefore, this study compares unimodal and different EEG-EMG hybrid pipelines for subject-specific multiclass classification of executed upper-limb gestures using the public dataset *8-channel EMG, EEG upper limb gesture data* [13]. We evaluate EEG-only and EMG-only baselines, feature-level fusion, decision-level fusion, corticomuscular coherence-based

features, and Riemannian covariance-based models under the same grouped cross-validation protocol. In addition to classification performance, we also report classification time per epoch to assess the practical relevance of each approach for future online motor interfaces. By directly comparing multiple fusion strategies against strong unimodal baselines, this work aims to clarify when hybrid EEG-EMG models are actually beneficial for the executed gesture classification in a dataset of healthy participants. These findings may support future studies investigating hybrid approaches for prosthetic control in amputees.

II. Materials and Methods

A. Dataset

We conducted the experiments using the public *8-channel EMG, EEG upper limb gesture data* dataset [13]. The main characteristics of the dataset and the preprocessing decisions adopted in this study are summarized in Table I

TABLE I
MAIN CHARACTERISTICS OF THE EEG-EMG DATASET USED IN THIS STUDY

Characteristic	**Description**
Dataset	*8-channel EMG, EEG upper limb gesture data*
Subjects	11 (7 males, right-handed, 1 male, left-handed, and 3 females, right-handed; ages 25 $\pm$ 5.5 years)
Classes	7 executed upper-limb gestures
Repetitions per class	6
EEG channels	8 channels placed on F3, F4, C3, CZ, C4, P3, PZ and P4 with the earlobes as reference, originally sampled at 250 Hz
EMG channels	8 channels placed on the proximal part of the subject's forearm, originally sampled at 200 Hz
Preprocessing	EEG resampled to 200 Hz to match EMG
Epoch window	First 5 s of each segment
Epoch length	1000 samples at 200 Hz

Two EEG segments (S4_R3_G3, subject 4 at the third repetition of the third gesture, and S11_R6_G2, subject 11 at the sixth repetition of the second gesture) were excluded due to unexpected signal shapes. Since all analyses used paired EEG-EMG epochs, the corresponding EMG segments were also removed from the paired dataset. This resulted in 460 valid EEG-EMG pairs.

B. Classifiers

Three classifiers were adopted as base models: Support Vector Machine, with two types of kernels: linear and Radial Basis Function (RBF); and Multilayer Perceptron (MLP). These models were selected because they are commonly used in EEG and EMG-based classification tasks and represent different levels of model complexity, from linear decision boundaries to nonlinear kernel-based and neural network-based classifiers [4], [14], [15]. All models were implemented in Python using `scikit-learn` for conventional machine learning classifiers, feature selection, preprocessing, cross-validation, and performance metrics [16].

Since the evaluated task involves seven gesture classes, SVM-based models were implemented using a one-vs-rest multiclass strategy. Both SVM configurations used balanced class weights to compensate for small class imbalance caused by excluded segments. The linear SVM used $C = 1.0$ and a maximum of 20,000 iterations, while the RBF SVM used $C = 1.0$, $\gamma = \text{scale}$, and the same maximum number of iterations when applicable.

The MLP classifier was used as a lightweight nonlinear neural model. In the feature-based classification experiments, it consisted of a single hidden layer with 32 neurons, rectified linear unit (ReLU) activation, Adam optimization, L_2 regularization with $\alpha = 10^{-3}$, an initial learning rate of 10^{-3}, and a maximum of 500 training iterations. These hyperparameters were kept fixed across subjects to favor comparability between pipelines rather than extensive model-specific optimization.

The hyperparameters were selected as simple and commonly used configurations for SVM and MLP classifiers, with no subject-specific hyperparameter search. This choice was adopted to avoid introducing an additional optimization layer that could favor particular pipelines or subjects, especially considering the limited number of trials available per participant.

C. Unimodal Baseline Pipelines

Unimodal EEG-only and EMG-only pipelines were first implemented to provide reference performances for the hybrid approaches. Therefore, these baselines are essential because the advantage of multimodal fusion can only be interpreted when compared with strong single-modality models evaluated under the same protocol.

For the feature-based baselines, EEG and EMG epochs were represented by time and frequency-domain descriptors extracted independently from each channel. Time-domain features included mean value, standard deviation, root mean square (RMS), mean absolute value (MAV), variance, waveform length, and peak-to-peak amplitude. Frequency-domain information was calculated using Power Spectral Density (PSD) with Welch's method [17], [18] and then summarized through bandpower features. For EEG, bandpower was computed in the 4–8 Hz, 8–13 Hz, 13–30 Hz, and 30–45 Hz bands [19]. While for EMG, spectral information was summarized in three broad frequency ranges, 10–30 Hz, 30–60 Hz, and 60–90 Hz, to provide a compact representation of the low-to-mid frequency content of the surface EMG signal. Unlike EEG bands, these EMG ranges were not intended to represent canonical physiological rhythms, but rather frequency-domain descriptors for pattern recognition.

Furthermore, features were standardized before classification, and the same feature set was used across subjects and folds. No supervised feature selection was included in the official experiments. This conservative choice was adopted

because each subject had a limited number of trials, and feature selection could increase the risk of overfitting to subject-specific or fold-specific patterns.

A covariance-based Riemannian baseline was also implemented separately for EEG and EMG. In this pipeline, each epoch was represented by its channel covariance matrix, estimated with Oracle Approximating Shrinkage (OAS), and projected to the tangent space using the Riemannian metric. The resulting tangent-space features were standardized and classified using the same base classifiers. This approach was included as a stronger unimodal baseline because Riemannian methods are well established for classification of multichannel physiological signals [12].

D. Fusion Pipelines

After establishing the unimodal baselines, four EEG-EMG fusion strategies were evaluated: feature-level fusion, decision-level fusion, corticomuscular coherence (CMC)-based fusion, and Riemannian covariance-based fusion. All fusion pipelines used the same epochs, labels and classifiers adopted for the unimodal baselines.

- **Feature-level fusion** was implemented by concatenating EEG and EMG features vectors used on each unimodal classifier before classification. The resulting multimodal feature vector was standardized and classified using the same base classifiers described previously, representing an early fusion strategy [8]–[10].
- **Decision-level fusion** was implemented as a late fusion strategy. EEG and EMG classifiers were first trained independently, and their class-wise scores for the seven gestures were concatenated into a 14-dimensional decision vector. This vector was then classified by a meta-classifier, implemented either as logistic regression or MLP. To avoid information leakage, the meta-classifier was trained using out-of-fold predictions generated within the training data [7]–[9].
- **CMC-based fusion** was evaluated as a physiologically motivated representation of EEG-EMG coupling. Magnitude-squared coherence was computed between EEG and EMG channel pairs. Two representations were tested: average coherence in motor-related frequency bands (μ-band: 8–13 Hz and β-band: 13–30 Hz [19]) and the full coherence spectrum between 8 and 30 Hz [20]–[22]. The band-averaged representation provides a compact description of corticomuscular coupling, whereas the spectral representation preserves frequency-specific peaks at the cost of substantially higher dimensionality. Just raw EMG signals were considered, and CMC was computed using a combination taken pair by pair of all EEG and EMG channels.
- **Riemannian covariance-based fusion** was implemented by stacking EEG and EMG channels before covariance estimation. Before covariance estimation, each epoch was globally normalized by its Frobenius norm, as also adopted in the unimodal Riemannian pipelines. The resulting multimodal epochs were then processed with the same Riemannian pipeline used for the unimodal baselines, including OAS covariance estimation, tangent-space projection, standardization, and classification. No additional channel-wise normalization was applied before stacking EEG and EMG channels. [12].

E. Pipelines Validation

All experiments were performed in a subject-specific manner, i.e., models were trained and evaluated independently for each participant. This choice was adopted because EEG and EMG patterns are highly subject-dependent and because the intended hBCI approaches assumes an individual calibration stage for each user.

For each subject, a six-fold stratified grouped cross-validation was used. The repetition identifiers were used as grouping variables, so all gesture trials belonging to the same repetition were assigned either to the training set or to the test set, but never to both. This strategy prevents samples from the same repetition from appearing simultaneously in training and test sets, reducing optimistic bias due to repetition-specific patterns. Stratification was used to preserve the class distribution across folds whenever possible.

The same validation protocol was applied to all unimodal and hybrid pipelines. For feature-based models, all preprocessing steps that depended on the training data, including standardization, were fitted only on the training folds and then applied to the corresponding test fold. In decision-level fusion, the meta-classifier was trained using out-of-fold predictions from the base EEG and EMG classifiers to avoid information leakage.

Model performances were evaluated using macro-averaged F1-score, balanced accuracy, accuracy, Cohen's kappa, and one-vs-rest macro AUC. Macro F1-score and balanced accuracy were considered the main metrics because the task involved multiclass gesture classification and two segments were excluded from the original dataset. In addition, the average classification time per epoch was measured for each pipeline to compare the computational cost of the trained classifiers. This measurement considered offline prediction time in Python.

F. Statistical Analysis

For statistical analysis, each subject was treated as the experimental unit. Therefore, for each pipeline, cross-validation results were first averaged across folds within each subject, and group-level results were then reported as mean $\pm$ standard deviation across subjects. This procedure avoids treating folds from the same subject as independent observations.

Differences among pipelines were assessed using the Friedman test, since the same subjects were evaluated under all pipeline configurations. When a significant global effect was observed, planned pairwise comparisons were performed using the Wilcoxon signed-rank test with Holm-Bonferroni correction for multiple comparisons. These planned comparisons focused on the main methodological questions of the study: whether the best EMG-based unimodal model differed

from the strongest fusion strategies, whether feature-level and decision-level fusion differed from each other, whether band-averaged and full-spectrum CMC representations differed, and whether stacked Riemannian fusion improved over EMG-only Riemannian classification. Effect sizes for pairwise comparisons were reported using Cohen's d_z.

III. Results

Complete results for all classifier configurations were obtained under the same subject-specific validation protocol. Table II summarizes the best-performing model for each evaluated pipeline, reporting macro F1-score, balanced accuracy, and classification time per epoch. In addition, Fig. 1 provides a visual summary of these selected models by showing their overall performance across multiple metrics and the trade-off between macro F1-score and classification time. All values are reported as mean $\pm$ standard deviation across subjects, after averaging the cross-validation folds within each subject. Therefore, the reported variability reflects inter-subject variability, not independent trial-level variability.

A. Unimodal Baselines

Among the feature-based unimodal baselines, EMG-based features achieved higher mean performance than EEG-based features. The best EEG-only model was obtained with the MLP classifier, reaching a macro F1-score of 0.708 ± 0.059 and balanced accuracy of 0.767 ± 0.049. In contrast, the best EMG-only feature-based model was obtained with the RBF SVM, reaching a macro F1-score of 0.858 ± 0.099 and balanced accuracy of 0.889 ± 0.087.

The Riemannian representation further improved the EMG-only performance. The EMG Riemannian pipeline with an RBF SVM achieved the highest mean performance among all evaluated pipelines, with a macro F1-score of 0.925 ± 0.043 and balanced accuracy of 0.946 ± 0.030. Conversely, the EEG Riemannian pipeline remained close to the EEG feature-based baseline, reaching a macro F1-score of 0.711 ± 0.096 and balanced accuracy of 0.768 ± 0.085. These results suggest that, for executed upper-limb gesture classification, the covariance structure of multichannel EMG contains highly discriminative information.

B. Hybrid Fusion Performance

Feature-level and decision-level fusion achieved competitive performance when compared with the feature-based unimodal baselines. Feature-level fusion achieved the best hybrid performance among the non-Riemannian fusion pipelines. The best model was the MLP classifier, with a macro F1-score of 0.877 ± 0.072 and balanced accuracy of 0.904 ± 0.058.

Decision-level fusion also performed well, but slightly below feature-level fusion. The best configuration combined an EEG RBF SVM, an EMG MLP, and an MLP meta-classifier, reaching a macro F1-score of 0.863 ± 0.091 and balanced accuracy of 0.898 ± 0.072. Although both hybrid strategies improved over the EEG-only baselines and were comparable to the EMG-only feature-based model, neither exceeded the mean performance of the EMG Riemannian baseline.

C. Coherence-Based Fusion

CMC-based fusion showed substantially lower performance than the other evaluated pipelines. The band-averaged CMC representation, computed from the average coherence in the μ and β bands, achieved its best result with a linear SVM, reaching a macro F1-score of 0.168 ± 0.056 and balanced accuracy of 0.214 ± 0.071. The full-spectrum CMC representation between 8 and 30 Hz performed even worse, with a macro F1-score of 0.102 ± 0.029 and balanced accuracy of 0.143 ± 0.046.

The planned comparison between both CMC representations indicated that the band-averaged representation achieved higher macro F1-score than the full-spectrum representation after nonparametric correction. However, both approaches remained close to chance level. Therefore, preserving the full coherence spectrum did not compensate for the low discriminative capacity of CMC in this dataset. This suggests that the poor CMC performance was unlikely to be explained only by the loss of narrowband peaks caused by band averaging, but may also reflect a mismatch between CMC as a physiological coupling marker and the multiclass gesture identification problem evaluated here.

D. Riemannian Covariance-Based Fusion

The best stacked Riemannian model was obtained with an RBF SVM, reaching a macro F1-score of 0.804 ± 0.092 and balanced accuracy of 0.846 ± 0.082. Although this result was higher than the EEG-only Riemannian baseline, it remained below the EMG-only Riemannian model.

This finding suggests that simply stacking EEG and EMG channels before covariance estimation did not provide additional discriminative information beyond the EMG covariance structure alone. Although joint covariance matrices can theoretically capture cross-modal relationships, EEG and EMG have different physiological origins, amplitude ranges, and signal-to-noise ratios. Therefore, when EMG is already highly discriminative, adding EEG channels may introduce structured variability into the covariance matrix rather than useful complementary information. This may explain why Riemannian methods were highly effective for EMG-only classification, but stacked EEG-EMG covariance was not the most effective fusion strategy in this dataset.

E. Statistical Comparison and Class-wise Analysis

The statistical analysis supported the existence of differences among pipelines for macro F1-score, balanced accuracy, and classification time (Friedman test, $p < 0.001$ for all metrics). However, planned post-hoc comparisons showed that the EMG Riemannian pipeline, despite achieving the highest mean macro F1-score and balanced accuracy, was not significantly superior to feature-level fusion, decision-level fusion, or EMG-only features after correction for multiple comparisons. In contrast, classification time differed more consistently across pipelines. Feature-level fusion was significantly faster than EMG Riemannian, decision-level fusion, and EMG-only features. For the coherence-based approaches, band-averaged

TABLE II

BEST PERFORMANCE OBTAINED BY EACH EVALUATED PIPELINE. F1: MACRO-AVERAGED F1-SCORE; BACC: BALANCED ACCURACY. VALUES ARE REPORTED AS MEAN ± STANDARD DEVIATION ACROSS SUBJECTS.

Pipeline	Best model	F1	BAcc	Time (ms/epoch)
EEG-only features	MLP	0.708 ± 0.059	0.767 ± 0.049	0.042 ± 0.003
EMG-only features	RBF SVM	0.858 ± 0.099	0.889 ± 0.087	0.280 ± 0.014
EEG Riemannian	RBF SVM	0.711 ± 0.096	0.768 ± 0.085	2.079 ± 0.899
EMG Riemannian	RBF SVM	0.925 ± 0.043	0.946 ± 0.030	1.987 ± 0.771
Feature-level fusion	MLP	0.877 ± 0.072	0.904 ± 0.058	0.073 ± 0.012
Decision-level fusion	EEG RBF SVM + EMG MLP + meta-MLP	0.863 ± 0.091	0.898 ± 0.072	0.338 ± 0.045
CMC band mean	Linear SVM	0.168 ± 0.056	0.214 ± 0.071	54.138 ± 2.236
CMC spectrum	Linear SVM	0.102 ± 0.029	0.143 ± 0.046	53.612 ± 7.318
Riemannian fusion	RBF SVM	0.804 ± 0.092	0.846 ± 0.082	2.782 ± 1.363

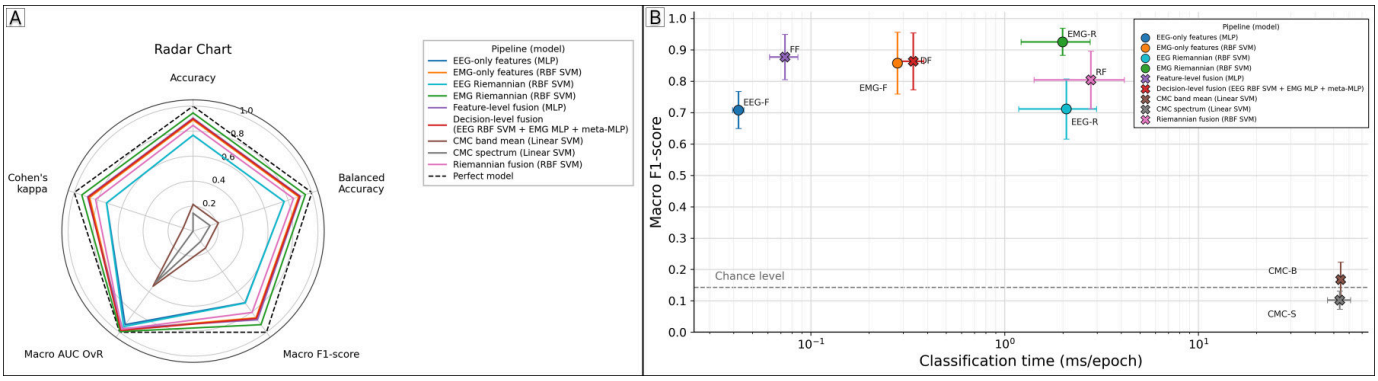


Fig. 1. Summary of the best-performing configuration for each evaluated pipeline. (A) Radar chart showing the mean performance across subjects for accuracy, balanced accuracy, macro F1-score, macro one-vs-rest AUC, and Cohen's kappa. The dashed line represents the ideal value for all metrics. (B) Macro F1-score versus classification time per epoch for the best model in each evaluated pipeline. The unimodal baselines are represented by circles and the hybrid models are represented by x-shaped markers. Error bars indicate one standard deviation across subjects, after averaging cross-validation folds within each subject. Vertical bars were clipped to the valid macro F1-score range [0, 1]. Overall, the results indicate that EMG provided stronger discriminative information than EEG for executed gesture classification, and that hybrid fusion strategies were competitive but did not necessarily outperform the best unimodal baseline.

CMC achieved significantly higher macro F1-score than full-spectrum CMC, although both remained close to chance level.

The class-wise recall analysis in Fig. 2 showed that the strongest EMG-based and hybrid pipelines maintained high recall across most gestures, whereas EEG-only and EEG Riemannian models showed more variable class-dependent performance. For the EMG Riemannian model, recall remained high for all gestures, with comparatively lower values for G3 and G4. In contrast, CMC-based models showed low recall across nearly all gesture classes, indicating that their poor performance was not restricted to a single movement, but reflected a broader limitation of the CMC representations for this multiclass task.

IV. DISCUSSION

Overall the results presented, they suggest that fusion can improve some unimodal feature-based models, but it does not automatically outperform a strong unimodal baseline configuration. Both feature-level and decision-level fusion achieved high mean performance and improved over the EEG-only baselines. However, although the EMG Riemannian pipeline achieved the highest mean macro F1-score and balanced accuracy, planned statistical comparisons did not confirm a significant advantage over feature-level fusion, decision-level fusion, or EMG-only features after correction for multiple comparisons. This finding is important because hybrid BCIs are often motivated by the complementarity between cortical and peripheral signals, with the expectation of improved robustness, reliability, or number of control commands [7], [8], [23]. Our results support this rationale only partially: EEG-EMG fusion was competitive and computationally attractive, especially in the feature-level and decision-level pipelines, but its benefit depended strongly on the representation and classifier architecture. Thus, the advantage of hybridization should not be assumed from the use of multiple modalities alone; it must be demonstrated against strong unimodal models evaluated under the same validation protocol.

The highest mean performance was obtained with the EMG Riemannian pipeline. A possible explanation is that the Riemannian representation captures the covariance structure among EMG channels, rather than treating each extracted feature independently. For multichannel physiological recordings, covariance-based representations can encode spatial relationships between sensors and have been shown to be effective in BCI classification problems [12]. In the present task, these channel relationships may reflect coordinated activation patterns across the forearm electrodes, which are highly



Fig. 2. Class-wise recall for the best-performing model of each evaluated pipeline. The heatmap summarizes how performance varied across gesture classes and pipelines, allowing the identification of movements that were more difficult to classify without requiring one confusion matrix for each model.

informative for distinguishing executed gestures. This may also explain why the same approach was less effective for EEG alone and for stacked EEG-EMG fusion. In EEG, the covariance structure may be less directly tied to the final gesture class, whereas stacking EEG and EMG channels before covariance estimation combines signals with different physiological origins, signal-to-noise ratios, and amplitude structures. Moreover, no channel-wise normalization was applied before stacking EEG and EMG channels. Thus, the joint covariance representation may still have been affected by differences in signal-to-noise ratio and discriminative strength between modalities. In this context, adding EEG channels to an already highly discriminative EMG covariance structure may have introduced structured noise rather than complementary information, which may partly explain why Riemannian fusion did not outperform the EMG-only Riemannian pipeline.

Although the hybrid models were not the best overall classifiers, their results should not be dismissed. Feature-level fusion achieved one of the best performance-time trade-offs, reaching high macro F1-score with very low classification time per epoch. Decision-level fusion also performed close to the EMG feature-based baseline while preserving a modular structure in which EEG and EMG classifiers are trained separately before being combined by a meta-classifier. This is relevant for future online interfaces. In this sense, the results suggest that feature-level and decision-level fusion remain promising strategies, especially if future systems are designed to exploit EEG as complementary information when EMG reliability decreases due to fatigue, electrode displacement, or changes in residual muscle condition [6], [7], [9].

The coherence-based pipelines showed substantially lower performance than the other approaches. This result does not imply that corticomuscular coherence is physiologically irrelevant. On the contrary, CMC is a well-established measure of coupling between sensorimotor cortical activity and peripheral muscle activation, especially during sustained contraction and motor control tasks [20], [21]. Moreover, CMC-based hBCIs have shown promising results in rehabilitation-oriented paradigms. For example, de Seta et al. used CMC features in a pseudo-online hBCI to discriminate hand movements from rest in healthy and stroke participants, reporting high detection performance and clinically relevant delays [24]. However, their objective was movement detection relative to rest, whereas the present study evaluated seven-class gesture identification during executed movements. Therefore, the two findings are not contradictory. Instead, they suggest that CMC may be more suitable for detecting motor engagement, task-related corticomuscular coupling, or movement onset than for encoding the fine identity of multiple executed gestures.

This interpretation is also consistent with studies that use CMC or corticomuscular networks as markers of motor control, rehabilitation state, or hybrid BCI activation rather than as direct multiclass gesture descriptors [22], [25]. In the present study, the comparison between band-averaged and full-spectrum CMC also suggests that the low performance was not caused only by the loss of narrow spectral peaks through band averaging. Although the full-spectrum representation preserved more frequency-specific information, it did not improve performance over the compact band-averaged representation. Thus, the poor CMC performance may also reflect physiological and methodological constraints of coherence estimation in this dataset, including the absence of a specific task-versus-rest contrast, the limited number of trials, possible EEG-EMG synchronization uncertainty, and the need for sufficiently stable time windows to estimate corticomuscular coupling [26], [27]. These aspects are particularly relevant for hBCI design,

because protocols that use CMC as a control feature may require adequate contraction duration, temporal alignment, and coupling stabilization before classification. Therefore, the low CMC performance observed here should not be interpreted as evidence against CMC as a physiological marker or as an hBCI activation feature, but rather as evidence that these simple CMC representations were not competitive for multiclass gesture classification in this dataset.

In addition, it is important to highlight that this study has some limitations when it comes to generalizing the results obtained here to real-time BCI or hBCI. First, all analyses were performed offline, using previously segmented EEG and EMG recordings, and therefore do not fully reproduce the constraints of online signal acquisition, continuous preprocessing, decision updating, and feedback delivery. Second, although the subject-specific design is coherent with individual calibration scenarios, the dataset included only 11 subjects with a limited number of trials per class, which restricts the generalization of the observed performances and increases the importance of validating these findings in larger datasets. The small sample size also limits the statistical power of the comparisons among pipelines. Therefore, the absence of statistically significant differences between the strongest models should not be interpreted as evidence that these models are equivalent. Third, because the EEG and EMG signals were provided as segmented files and were originally sampled at different frequencies, it is not possible to guarantee that the two modalities were perfectly synchronized at the sample level. This is a critical point for multimodal EEG-EMG analyses, especially for coherence-based approaches, since CMC estimation depends directly on the temporal relationship between cortical and muscular activity [27], [28]. Therefore, synchronization uncertainty may have contributed directly to the low performance of the CMC-based models, in addition to being a general limitation for multimodal interpretation. Finally, classification time was measured in a Python/offline environment rather than on embedded real-time hardware. Thus, the reported time per epoch should be interpreted as a comparative estimate of computational cost across pipelines, and not as a final measure of real-time system latency.

Taken together, for the specific analysis of the specific dataset showed here, the present results indicate that EEG-EMG fusion is useful, but not universally superior. For executed upper-limb gesture classification, EMG carried the strongest discriminative information, especially when represented through Riemannian covariance features. However, the lack of statistically confirmed superiority over the strongest feature-level and decision-level fusion pipelines suggests that computational cost and implementation simplicity should also guide model selection for future online systems. CMC, despite its physiological relevance, was not suitable as a direct multiclass gesture feature in this dataset. These findings reinforce that hybrid motor interfaces should be evaluated not only by whether they combine EEG and EMG, but by how the signals are represented, fused, validated, and ultimately translated into computationally feasible control decisions.

V. Conclusion

This study provided a comparative analysis between unimodal models and different hybrid EEG-EMG fusion strategies for the classification of seven executed upper-limb gestures. Overall, the results showed that hybrid EEG-EMG models are not automatically superior to unimodal models, since their benefit depends directly on the adopted fusion strategy. Although the highest mean performance was achieved by the unimodal EMG Riemannian pipeline, this advantage was not statistically confirmed over the strongest feature-level and decision-level fusion pipelines after correction for multiple comparisons. Since these fusion approaches achieved competitive performance with lower offline classification time per epoch, they provide relevant insights for future online evaluations of EEG-EMG architectures.

An important negative finding was the substantially lower performance of the CMC-based models. Nevertheless, CMC-based approaches should not be automatically discarded, since corticomuscular coupling has a well-established physiological basis. Instead, the present results suggest that CMC may not be the most suitable feature for distinguishing different executed gestures in a multiclass problem. Rather, it may be more informative as a marker of motor engagement, attention, or task-related corticomuscular interaction during motor activity.

Therefore, although further studies with larger datasets, improved EEG-EMG synchronization, real-time validation, and complete real-time pipeline validation are still necessary, the results obtained in this analysis may help guide future hBCI applications. In particular, they indicate which fusion strategies and unimodal baselines should be prioritized in future online tests for EEG-EMG-based motor interfaces.

VI. Declaration of AI use

The authors declare that AI tools were used exclusively for language editing and proofreading, under strict human oversight. AI was not used to generate scientific hypotheses, define the methodology, perform data analysis, or interpret the results. After using these tools, the authors carefully reviewed and edited the manuscript to ensure accuracy, coherence, and scientific responsibility.

VII. Acknowledgment

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant no. 2024/03628-5, by the Conselho Nacional de Desenvolvimento Científico (CNPq), Universal Call, grant no. 409232/2023-8, and also was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

[1] G. Moisan and C. Z.-H. Ma, "Advances in prosthetics and orthotics," *BMC Musculoskeletal Disorders*, vol. 25, no. 1, p. 135, Feb. 2024.

[2] N. Olaya-Mira, L. M. Gómez-Hernández, C. Viloria-Barragán, and I. C. Soto-Cardona, "Methods to assess lower limb prosthetic adaptation: a systematic review," *Journal of NeuroEngineering and Rehabilitation*, vol. 22, no. 1, p. 100, Apr. 2025.

[3] J. B. Webster, A. Crunkhorn, J. Sall, M. J. Highsmith, A. Pruziner, and B. J. Randolph, "Clinical practice guidelines for the rehabilitation of lower limb amputation: An update from the department of veterans affairs and department of defense," *American Journal of Physical Medicine & Rehabilitation*, vol. 98, no. 9, pp. 820–829, Sep. 2019.

[4] A. Phinyomark, R. N. Khushaba, and E. Scheme, "Feature extraction and selection for myoelectric control based on wearable EMG sensors," *Sensors*, vol. 18, no. 5, p. 1615, May 2018.

[5] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[6] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use," *Journal of Rehabilitation Research and Development*, vol. 48, no. 6, pp. 643–660, 2011.

[7] K.-S. Hong and M. J. Khan, "Hybrid brain-computer interface techniques for improved classification accuracy and increased number of commands: a review," *Frontiers in Neurorobotics*, vol. 11, p. 35, Jul. 2017.

[8] G. Pfurtscheller, B. Z. Allison, C. Brunner, G. Bauernfeind, T. Solis-Escalante, R. Scherer, T. O. Zander, G. Mueller-Putz, C. Neuper, and N. Birbaumer, "The hybrid BCI," *Frontiers in Neuroscience*, vol. 4, p. 3, 2010.

[9] R. Leeb, H. Sagha, R. Chavarriaga, and J. d. R. Millán, "A hybrid brain-computer interface based on the fusion of electroencephalographic and electromyographic activities," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025011, Apr. 2011.

[10] O. Özdenizci, S. Y. Günay, F. Quivira, and D. Erdoğmuş, "Hierarchical graphical models for context-aware hybrid brain-machine interfaces," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Jul. 2018, pp. 1964–1967.

[11] J. Zhang, B. Wang, C. Zhang, Y. Xiao, and M. Y. Wang, "An EEG/EMG/EOG-based multimodal human-machine interface to real-time control of a soft robot hand," *Frontiers in Neurorobotics*, vol. 13, p. 7, Mar. 2019.

[12] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by Riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012.

[13] M. Dere, "8-channel EMG, EEG upper limb gesture data," Mendeley Data, Dec. 2022.

[14] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[15] R. H. Chowdhury, M. B. I. Reaz, M. A. B. M. Ali, A. A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, no. 9, pp. 12 431–12 466, Sep. 2013.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] O. M. J. Solomon, "PSD computations using Welch's method [power spectral density (PSD)]," Sandia National Laboratories, Tech. Rep. SAND1991-1533, 1991.

[18] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[19] C. S. Nayak and A. C. Anilkumar, "EEG normal waveforms," in *StatPearls*. Treasure Island, FL: StatPearls Publishing, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK539805/

[20] S. N. Baker, "Oscillatory interactions between sensorimotor cortex and the periphery," *Current Opinion in Neurobiology*, vol. 18, no. 6, pp. 649–655, Dec. 2008.

[21] J. Liu, Y. Sheng, and H. Liu, "Corticomuscular coherence and its applications: a review," *Frontiers in Human Neuroscience*, vol. 13, p. 100, Mar. 2019.

[22] F. Pichiorri, J. Toppi, V. de Seta, E. Colamarino, M. Masciullo, F. Tamburella, M. Lorusso, F. Cincotti, and D. Mattia, "Exploring high-density corticomuscular networks after stroke to enable a hybrid brain-computer interface for hand motor rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 20, no. 1, p. 5, Jan. 2023.

[23] B. Z. Allison, R. Leeb, C. Brunner, G. R. Müller-Putz, G. Bauernfeind, J. W. Kelly, and C. Neuper, "Toward smarter BCIs: extending BCIs through hybridization and intelligent control," *Journal of Neural Engineering*, vol. 9, no. 1, p. 013001, Jan. 2012.

[24] V. de Seta, J. Toppi, E. Colamarino, R. Molle, F. Castellani, F. Cincotti, D. Mattia, and F. Pichiorri, "Cortico-muscular coupling to control a hybrid brain-computer interface for upper limb motor rehabilitation: a pseudo-online study on stroke patients," *Frontiers in Human Neuroscience*, vol. 16, p. 1016862, Nov. 2022.

[25] E. Colamarino, V. de Seta, M. Masciullo, F. Cincotti, D. Mattia, F. Pichiorri, and J. Toppi, "Corticomuscular and intermuscular coupling in simple hand movements to enable a hybrid brain-computer interface," *International Journal of Neural Systems*, vol. 31, no. 11, p. 2150052, 2021.

[26] T. W. Boonstra and M. Breakspear, "Neural mechanisms of intermuscular coherence: implications for the rectification of surface electromyography," *Journal of Neurophysiology*, vol. 107, no. 3, pp. 796–807, 2012.

[27] F. Artoni, A. Barsotti, E. Guanziroli, S. Micera, A. Landi, and F. Molteni, "Effective synchronization of EEG and EMG for mobile brain/body imaging in clinical settings," *Frontiers in Human Neuroscience*, vol. 11, p. 652, Jan. 2018.

[28] S. Iwama, M. Takemi, R. Eguchi, R. Hirose, M. Morishige, and J. Ushiba, "Two common issues in synchronized multimodal recordings with EEG: jitter and latency," *Neuroscience Research*, vol. 203, pp. 1–7, 2024.

Intelligent System for Cervical Biomechanical Tracking in Speech-Language Rehabilitation of Parkinson's Disease

1st Arthur Silva
Edmond and Lily Safra Neurosci. Inst.
Santos Dumont Institute (ISD)
Macaíba, Brazil
arthur.silva@edu.isd.org.br

2nd Elizama Silva
Edmond and Lily Safra Neurosci. Inst.
Santos Dumont Institute (ISD)
Macaíba, Brazil
elizama_barreto@hotmail.com

3rd Mariselma Xerez
Edmond and Lily Safra Neurosci. Inst.
Santos Dumont Institute (ISD)
Macaíba, Brazil
fgamariselmatrajano@gmail.com

4th Ana Reis
Anita Garibaldi Ctr. for Educ. and Health
Santos Dumont Institute
Macaíba, Brazil
ana.reis@isd.org.br

5th André Dantas
Edmond and Lily Safra Neurosci. Inst.
Santos Dumont Institute (ISD)
Macaíba, Brazil
andre.dantas@isd.org.br

6th Caroline Santo
Edmond and Lily Safra Neurosci. Inst.
Santos Dumont Institute (ISD)
Macaíba, Brazil
caroline.santo@isd.org.br

***Abstract*—Parkinson's disease compromises cervicofacial mobility, affecting swallowing dynamics and phonatory functions. This work presents a proof-of-concept study of a computer vision system for the non-invasive tracking of cervical anatomical landmarks from conventional video. Using the DeepLabCut tool with transfer learning (ResNet-50), a network was trained to locate 17 cervical anatomical landmarks from 24 publicly accessible videos, obtaining a Mean Radial Error (MRE) of 3.86 pixels on the test set after 100,000 iterations. It should be stressed that the 95/5 partition was performed at the frame level rather than at the video level, which implies possible information leakage between training and test and makes this value an optimistic estimate. From the estimated trajectories, the automatic detection of the swallowing reflex onset was evaluated through a benchmark among four classifiers (Random Forest, XGBoost, SVM, and KNN); the dataset for this stage contains only 182 sampled frames, of which 27 (14.8%) correspond to onset. Only Random Forest detected any event, with 25% recall and an F1 of 0.08, meaning that roughly three out of every four onsets went undetected; no class-balancing strategy was applied. The results demonstrate the technical feasibility of markerless cervical tracking, but event-detection performance is insufficient for clinical use. Importantly, the dataset does not include individuals with a confirmed diagnosis of Parkinson's disease, so no conclusion regarding clinical applicability to parkinsonian dysphagia can be drawn at this stage.**

***Index Terms*—DeepLabCut, transfer learning, swallowing, Parkinson's disease, computer vision, proof of concept**

I. Introduction

Parkinson's disease (PD) is a chronic and progressive neurodegenerative condition of the central nervous system, resulting primarily from the loss of dopaminergic neurons in the substantia nigra pars compacta [1]. Clinically, it manifests through motor disorders such as resting tremor, muscle rigidity, postural instability, and bradykinesia. In the speech-language domain, bradykinesia and rigidity affect the orofacial and cervical musculature, culminating in facial hypomimia, dysphonia, and oropharyngeal dysphagia, the latter being directly associated with an elevated risk of aspiration pneumonia and malnutrition [1], [2].

Preserving phonatory and swallowing functions depends on systematic speech-language interventions. In this context, there arises a need to track and map specific anatomical landmarks on the face and neck in order to quantify the patient's mobility and movement patterns. In response to this demand, at least two viable approaches stand out: networks with transfer learning focused on the cervical region, and consolidated tracking solutions such as MediaPipe [3], aimed at facial dynamics.

Systems based on computer vision and deep learning enable the non-invasive extraction of biomechanical metrics using conventional cameras. Although the detection of anatomical landmarks by deep networks is well established in the literature [4], the mapping of the external cervical region from conventional video remains little explored. The focus on the cervical region is justified by its importance in swallowing biomechanics, specifically in monitoring the vertical excursion of the hyolaryngeal complex and in following suprahyoid and stabilizing muscle dynamics. To address this demand, the markerless pose estimation tool DeepLabCut stands out [5], [6].

Modeling through DeepLabCut employs architectures based on adapted Residual Networks (ResNet) [7] coupled with spatial probability decoders (*scoremaps*). Training such networks from scratch requires thousands of annotated images, a scarce resource in the biomedical domain. Transfer learning [8] mitigates this bottleneck by reusing weights previously optimized on generic databases (ImageNet [9], ResNet-50 architecture) and fine-tuning them for cervical anatomy.

Works applying deep learning to swallowing kinematics concentrate on instrumented modalities. Videofluoroscopy (VFSS) remains the gold standard and has already been combined with deep networks for automatic hyoid bone detection [10], but it depends on hospital infrastructure, ionizing radiation, and specialized professionals. Non-invasive alternatives include high-resolution cervical auscultation with deep learning [11] and wearable accelerometry sensors [12], which dispense with radiation but require instrumentation attached to the patient. Conventional video tracking, by contrast, requires neither radiation nor attached sensors, yet it has been little investigated for the external cervical region. The present work is positioned in this gap and is explicitly characterized as a technical proof-of-concept study: it evaluates whether markerless pose estimation is able to track cervical structures in conventional video, without claiming, at this stage, clinical validation in a parkinsonian population.

II. Objectives

A. General Objective

To evaluate, as a proof of concept, the technical feasibility of a computer vision-based system for the markerless tracking of cervical anatomical landmarks relevant to swallowing biomechanics.

B. Specific Objectives

- To optimize the training of the neural network via DeepLabCut using transfer learning for the identification and marking of cervical anatomical landmarks critical to swallowing, based on a public video database.
- To evaluate the feasibility of automatically detecting the onset instant of the swallowing reflex from the spatial trajectories estimated by the network, through a comparative benchmark among supervised machine learning classifiers.
- To analyze the model's performance quantitatively and qualitatively, comparing it with reference values from the literature, so as to delimit its methodological limitations and support guidelines for future work.

III. Methodology

A. Dataset Characterization

To assemble the database, searches for videos of the cervical region were carried out in search engines, restricted to publicly accessible material. The set comprised 24 videos depicting the frontal positioning characteristic of the intended use of the application, a size consistent with recommendations for exploratory pilot studies [13].

The following characteristics and limitations of the dataset should be made explicit, given the exploratory nature of this work:

- **Absence of a clinical population:** the videos depict individuals with no known or confirmed PD diagnosis, and no clinical metadata are available (age, sex, Hoehn and Yahr stage, presence or degree of dysphagia). The dataset therefore does *not* represent the target clinical population, and the performance reported here cannot be extrapolated to parkinsonian patients, who present precisely a reduction in the amplitude and velocity of cervical movements.
- **Acquisition heterogeneity:** the videos were obtained from independent public sources, with no control over camera distance, lighting, or resolution, and with acquisition parameters that were not recorded at collection time. Frame rates were in the region of 30 fps, but spatial resolution varied across the set and was not documented per video. This heterogeneity carries a direct methodological consequence that compounds the interpretability problem discussed in Section IV-A: because a pixel corresponds to a different physical distance in each video, the MRE aggregated over the dataset is not expressed in a single consistent physical unit. Recording resolution and frame rate per clip, and reporting the error stratified by acquisition condition, is a required step in any follow-up study.
- **Absence of spatial calibration:** no physical referent of known scale is present in the videos, which prevents converting errors and displacements from pixels to millimeters.

To avoid computational redundancy, the *K-means* clustering algorithm (integrated into DeepLabCut [5]) was used to automatically select, on average, 20 representative frames per video, covering rest, onset of the reflex, peak of laryngeal excursion, and return to the basal position. This amounted to approximately 480 annotated frames. Python 3.8 was used with DeepLabCut in a Conda environment with TensorFlow, on a Linux (Ubuntu) operating system.

B. Anatomical Landmark Selection Criteria

The elevation and anteriorization of the hyolaryngeal complex constitute the primary biomechanical mechanisms of airway protection during swallowing, and hyoid and laryngeal displacement is a consolidated parameter in the instrumental assessment of dysphagia, with normative values described in the literature [14]. The taxonomy adopted therefore sought to cover three functional domains: (i) the vertical excursion of the hyolaryngeal complex, the primary target of the analysis; (ii) the activity of the suprahyoid musculature, responsible for hyoid traction; and (iii) stable references for displacement normalization.

A protocol of 17 landmarks distributed across the cervical region and the upper thoracic boundary was defined: lateral borders of the neck (superior, medial, and inferior; right and left sides), sternocleidomastoid muscle (SCM) (superior, medial, and inferior; right and left sides), suprahyoid musculature (digastric, mylohyoid, and geniohyoid), larynx (thyroid prominence), and sternum (jugular notch), shown in Fig. 1.

One conceptual distinction must be recorded: the landmarks tracked here are *cutaneous surface landmarks*, not the bony or cartilaginous structures themselves. The "larynx" point corresponds to the cutaneous projection of the thyroid prominence, not to the hyoid bone visualized in VFSS. These are

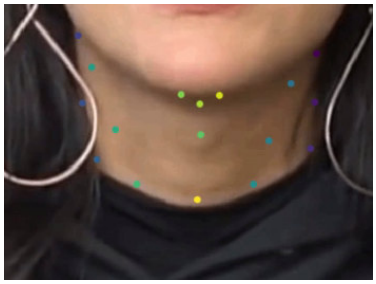


Figure 1. Anatomical landmarks.

therefore indirect measures (*proxies*) whose correlation with actual hyoid displacement has not been established in this work, which constitutes a validation step required before any clinical use.

C. Annotation and Data Partitioning

Manual annotation (*ground truth*) was performed frame by frame in the DeepLabCut Graphical User Interface, defining the Cartesian coordinates (x, y) of each landmark. The marking was carried out **independently** by two specialized speech-language pathologists, neither having access to the other's annotations. It is recorded, however, that **no quantitative measure of inter-rater agreement was computed** (for example, intraclass correlation coefficient or mean pixel error between annotators). Consequently, the variability inherent to the *ground truth* itself remains unknown, which makes it impossible to determine which portion of the reported MRE is attributable to the model and which to imprecision in the reference. Quantifying this agreement is a necessary step in future work.

Since the study exclusively involved public-domain videos, with no recruitment, intervention, or collection of identifiable data, it did not involve procedures subject to review by a Research Ethics Committee under the applicable national regulation [15].

The annotated set was partitioned according to DeepLabCut's *Shuffle* 4, allocating 95% of the frames to training and 5% to testing. It must be stressed, as the central methodological limitation of this work, that **this partition operates at the frame level and not at the video or subject level**. As a consequence, frames from the same video (and therefore from the same individual, with the same anatomy, lighting, and framing) may be simultaneously present in the training and test sets. This information leakage makes the reported error an optimistic estimate of the actual capacity to generalize to new individuals, rather than a measure of between-subject generalization.

D. Network Training

The base model was a residual network pre-trained on ImageNet (ResNet-50), adapted with deconvolutional layers for the generation of *scoremaps*. The main parameters are summarized in Table I.

Performance was quantified by the Mean Radial Error (MRE), the Euclidean distance in pixels between predicted coordinates and the *ground truth*. A spatial probability threshold (*p-cutoff* = 0.1) was additionally applied, restricting the computation to predictions with confidence equal to or greater than 10%.

E. Swallowing Onset Detection Pipeline

Once trained, the network was used to automatically generate the coordinates of the 17 landmarks in every frame of three new swallowing videos, producing, for each one, a CSV file with the spatiotemporal trajectories. Specialized speech-language pathologists performed the manual marking of the onset instant of the reflex. This marking was not performed frame by frame, but every 3 frames, an interval sufficient for visual perception of the difference between rest and onset. It should be noted that this 3-frame labeling resolution establishes a floor for the attainable temporal precision: the maximum label granularity is, by construction, three times coarser than that of the video.

From these markings, a binary column (*onset*) was created and incorporated into the coordinate CSV. The *pipeline* adopted was as follows:

- **Feature vector:** the Cartesian coordinates (x, y) of the 17 anatomical landmarks estimated by the network, augmented with variation/velocity features obtained by first-order differencing between consecutive frames of the sampled series, totaling 54 columns in the final dataset. The spatial probabilities (*likelihood*) produced by DeepLabCut were **not** incorporated into the feature vector.
- **Temporal resolution and windowing:** classification was performed independently on each sampled frame, **with no sliding window** and no aggregation of temporal context. The only dynamic information available to the models is that contained in the first-order variation features, which corresponds to an effective context of two sampled frames.
- **Pre-processing:** no spatial normalization of the coordinates relative to a reference point, no feature scaling, no temporal filtering, and no explicit handling of missing values were applied.
- **Class imbalance handling: no balancing strategy was adopted**. All classifiers were trained on the original class distribution, without class weighting, oversampling, or synthetic example generation. As discussed in Section IV-B, this is a determining factor for the degenerate behavior observed in three of the four models.

Table II presents the composition of the dataset used in this stage. The final set contains 182 sampled frames, of which only 27 (14.8%) correspond to the "onset" class. Note the marked heterogeneity across videos: the proportion of onset frames ranges from 3.9% (*video_2*) to 29.4% (*video_3*), so that the difficulty of the task and the degree of imbalance depend strongly on which video is held out for testing.

Table I
TRAINING PARAMETERS ADOPTED

Parameter	**Specification**
Optimizer	Stochastic gradient descent (SGD) with momentum and learning rate decay.
Learning Rate	*Multi-step* schedule: 0.005 (0 to 10,000 iterations) and 0.02 (10,000 to 100,000 iterations).
Batch Size	`batch_size: 1`
Data Split	Shuffle 4, with a partition of 95% for training and 5% for testing, **at the frame level**.
Total Iterations	100,000 iterations, with *snapshots* every 1,000 iterations.
Processing	Conda environment (`dlc_gui`), Python 3.8, TensorFlow *backend*, Linux (Ubuntu), executed on a central processing unit (CPU).

Table II
DATASET COMPOSITION FOR ONSET DETECTION (FRAMES SAMPLED EVERY 3 FRAMES OF THE ORIGINAL VIDEO)

Video	**Rest**	**Onset**	**Total**	**% Onset**
video_1	21	8	29	27.6%
video_2	98	4	102	3.9%
video_3	36	15	51	29.4%
Total	**155**	**27**	**182**	**14.8%**

Table III
QUANTITATIVE PERFORMANCE OF THE DEEPLABCUT MODEL (RESNET-50, SHUFFLE 4) AT 100,000 ITERATIONS

Metric	**Training**	**Test**
MRE, without *p-cutoff*	4.03 px	3.86 px
MRE with *p-cutoff* = 0.1	4.03 px	3.86 px

A comparative benchmark was conducted among Random Forest [16], XGBoost [17], *Support Vector Machine* (SVM) with a radial basis function (RBF) *kernel* [18], and *K-Nearest Neighbors* (KNN), all subjected to the same decision probability threshold ($p = 0.35$). Of the three videos, two were allocated to training and one was reserved exclusively for testing, so that the evaluation rests on a single video not observed during fitting. Given the distribution in Table II, this implies that the test set contains at most a handful of positive frames, making the minority-class metrics extremely sensitive to individual hits or misses [19].

IV. RESULTS AND DISCUSSION

A. Anatomical Mapping Performance

The evaluation was conducted upon reaching 100,000 iterations in *Shuffle* 4. Table III summarizes the MRE obtained.

Fig. 2 visually illustrates the spatial proximity between the manual markers (circles) and the network predictions (crosses). The frame shown belongs to the **training** set; the result observed there is therefore more optimistic than performance on unseen frames and serves only as a qualitative check.

As a technical reference point, the founding DeepLabCut literature reports that with as few as approximately 200 labeled frames the network attains tracking performance on test frames comparable to human accuracy, and further documents that the trial-to-trial labeling variability of a single expert annotator is itself mostly below 5 pixels [5]. Against that benchmark,

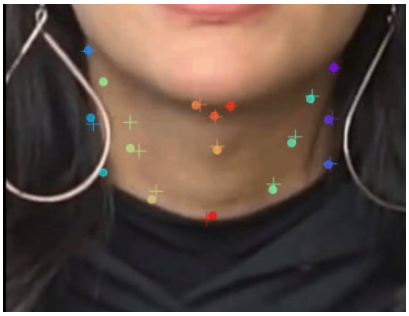


Figure 2. Comparison between the manual markers (*ground truth*, circles) and the neural network predictions (crosses) in a frame from the training set.

the MRE of 3.86 pixels obtained here from approximately 480 annotated frames falls within the range expected for the tool. The comparison is nonetheless only indicative, since those reference figures were obtained on datasets with different image scales and different target-structure sizes.

This value, moreover, requires three qualifications before any interpretation of merit:

1) **Leakage from frame-level partitioning:** as detailed in Section III-C, the 3.86 px error was measured on frames sharing video and individual with the training set. It measures within-video consistency, not between-subject generalization, and should be read as an optimistic upper bound.
2) **Absence of absolute metric meaning:** an error in pixels is not biomechanically interpretable without spatial calibration. To gauge the problem: the measurement error for hyoid and laryngeal displacement in VFSS, the reference modality, lies between 2.48 and 3.06 mm [20]. Without a physical referent in the videos, it is not possible to state whether 3.86 px corresponds to a fraction or a multiple of that range and, consequently, **no clinical meaning can be assigned to the value obtained**.
3) **Aggregated error:** the reported MRE is the mean across the 17 landmarks. As the qualitative analysis below indicates, the dispersion across landmarks is substantial, so the mean masks the poor performance of specific markers. Per-landmark errors were not computed at this stage.

The numerical equivalence of the errors with and without the confidence filter (*p-cutoff* = 0.1) indicates that the network

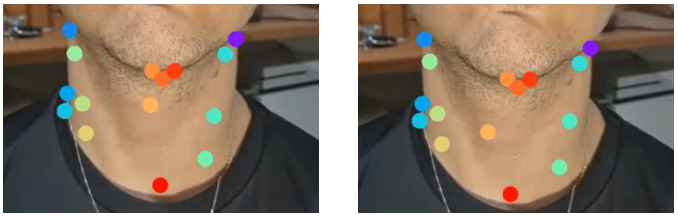


Figure 3. Examples of automatic tracking of the laryngeal point by the network, in the same video: (a) anatomically coherent marking, anchored on the thyroid prominence; (b) marking with deviation relative to the anatomical structure. Visual classification validated by two specialized speech-language pathologists.

kept its spatial probability above the threshold in virtually all estimates. It is worth noting that a threshold of 0.1 is permissive, so this stability should not be read as an indicator of high confidence.

The qualitative analysis indicates heterogeneous adherence of the markers to the underlying structures:

1) **Larynx and suprahyoid region:** the central laryngeal point and the suprahyoid region showed good fidelity in most frames. Oscillation and occasional deviation of the laryngeal marker were nevertheless observed in certain frames, probably associated with moments of higher displacement velocity or with contrast variations at the thyroid prominence. The distinction between coherent frames and frames with deviation was made by visual analysis, validated by two specialized speech-language pathologists, as exemplified in Fig. 3. Since the laryngeal marker is precisely the primary target of the biomechanical analysis, this instability is the most consequential of the tracking limitations observed.
2) **Sternocleidomastoid and lateral borders:** the chains of SCM points maintained adequate alignment. The lateral border points, however, did not remain stable throughout the video, an instability attributed to image noise and to the poor contour definition in that region.
3) **Lower anchoring point:** the sternum marker remained stable, providing a baseline for computing relative vertical elevation distances.

A demonstration video of the tracking in operation was made publicly available[1], allowing direct visual inspection of marker adherence throughout the movement.

B. Swallowing Onset Detection: Classifier Benchmark

Table IV summarizes the overall accuracy and the metrics for the class of interest ("onset").

The results highlight the difficulty of the task under severe class imbalance. The XGBoost, SVM, and KNN models failed to generalize to the rare class, obtaining null *recall* (0.00). In the confusion matrices of Fig. 4, rows represent the actual class and columns the predicted class; for these three classifiers, the entire column corresponding to the predicted "Onset" class remains empty, showing that no frame was classified as onset.

[1]Video available at: https://youtu.be/dnU-ldg7EPM

Table IV
COMPARATIVE BENCHMARK OF THE SWALLOWING ONSET DETECTION CLASSIFIERS (THRESHOLD = 0.35)

Model	Acc.	Prec.	Rec.	F1
Random Forest	77.5%	0.05	0.25	0.08
XGBoost	63.7%	0.00	0.00	0.00
SVM (RBF)	95.1%	0.00	0.00	0.00
KNN ($K = 5$)	76.5%	0.00	0.00	0.00

Prec., Rec., and F1 refer to the "Onset" class.

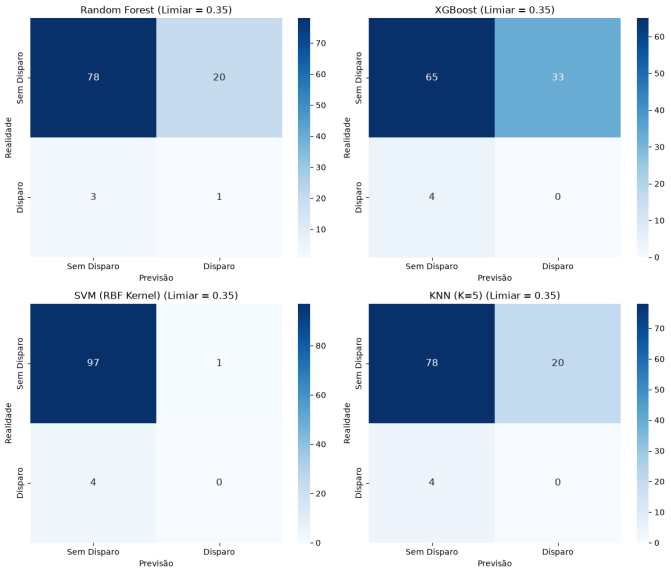


Figure 4. Confusion matrices of the four classifiers.

The SVM, despite achieving the highest nominal accuracy of the experiment (95.1%), classified the entire test video as rest: this is a false accuracy peak, explained solely by the imbalance [19], and it illustrates why overall accuracy is an inadequate metric in this scenario.

Random Forest was the only classifier to overcome the majority-class bias, with 25% *recall* and an overall accuracy of 77.5%. In the corresponding subplot of Fig. 4, this translates into at least one true positive in the "Onset" (actual) × "Onset" (predicted) cell.

Interpreting this result requires candour about its operational meaning. A *recall* of 25% means that **three out of every four swallowing onsets went undetected**. A precision of 0.05 means that, among the frames flagged as onset, approximately 95% were false positives. The resulting F1 of 0.08 is very close to what random flagging would yield under this degree of imbalance. In a dysphagia screening context, the false negative is the more consequential error: failing to detect a swallowing event may mean missing precisely the altered episode one intended to identify, with a risk of false clinical reassurance. For this reason, Random Forest should be understood here only as the most promising baseline classifier among those tested, and **not as a usable detector**. No inference regarding applicability to clinical screening or monitoring is supported

at this performance level.

Contributing to this result, cumulatively and now quantifiable from Table II: (i) the very small absolute number of positive examples, with only 27 onset frames in the entire dataset and a single test video containing just a few of them; (ii) the absence of any class-imbalance handling, which makes constant prediction of the majority class the lowest-cost solution for the models and directly explains the null *recall* of XGBoost, SVM, and KNN; (iii) the independent per-frame classification, which discards the sequential structure of the event and reduces the available temporal context to a first-order difference between two frames; and (iv) the 3-frame labeling granularity, which limits the attainable temporal precision by construction.

One further note of methodological transparency is warranted: the overall accuracy values in Table IV are not fully reconcilable with the per-video partition described in Table II, which suggests a possible divergence between the subsets actually evaluated by each classifier. This verification, together with the inclusion of the specificity requested by the reviewers, is under way and does not alter the qualitative conclusion of this section: detection performance remains well below what clinical use would require.

C. Comparison with Existing Approaches

Comparison with the literature on automatic swallowing event detection situates the present work clearly. In high-resolution cervical auscultation with deep networks, detection accuracies above 95% have been reported, validated on more than 3,000 swallows from 248 patients labeled by experienced speech-language pathologists [11]. Wearable accelerometry approaches have likewise demonstrated hyoid displacement tracking without videofluoroscopy [12], and VFSS combined with deep learning achieves robust automatic hyoid detection [10].

Against these works, the approach presented here is, on the axis of event-detection performance, inferior by a wide margin, explained above all by a difference of two to three orders of magnitude in the volume of annotated events. Its potential contribution lies on a different axis: the barrier to access. Cervical auscultation and accelerometry require sensors attached to the patient and dedicated hardware; VFSS requires a hospital environment and radiation. Conventional video tracking dispenses with all three. The open question, which this work does not answer, is whether the information available at the cervical skin surface is sufficient to support event detection at a clinically useful level.

D. Limitations

The following limitations are consolidated:

1) **Mismatch between motivation and validation:** the motivation is dysphagia in PD, but the dataset contains neither individuals with confirmed PD nor clinical metadata. No conclusion about the target population is supportable.
2) **Frame-level partitioning:** the MRE of 3.86 px is optimistic owing to within-video leakage and does not measure between-subject generalization.
3) **Absence of spatial calibration and undocumented acquisition parameters:** errors in pixels are not convertible to millimeters and are therefore not biomechanically interpretable. Because per-video resolution was not recorded, the aggregated MRE also mixes pixels of differing physical size.
4) **Insufficient event-detection performance:** 25% *recall* and an F1 of 0.08 rule out any clinical use.
5) **Sample size:** 24 videos for tracking and three videos (two for training, one for testing) for event detection, with only 27 positive frames out of 182.
6) **No class-imbalance handling:** the classifiers were trained on the original distribution (14.8% positives), without class weighting or resampling.
7) **No temporal modeling:** independent per-frame classification disregards the sequential nature of the swallowing event.
8) **Inter-rater agreement not quantified:** annotation was independent, but the variability of the *ground truth* was not measured.
9) **Surface *proxies*:** the markers track the skin surface, and their correlation with actual hyoid displacement has not been established.
10) **Aggregated error metrics:** per-landmark errors were not computed.

V. Final Considerations and Future Work

This work demonstrated the technical feasibility of using markerless pose estimation to track external cervical structures in conventional video, with an MRE of 3.86 pixels consistent with the range expected for the tool. This is a proof-of-concept contribution: it establishes that tracking is technically feasible with scarce annotated data, and it delimits precisely what still separates this result from a clinically usable tool.

The swallowing onset detection stage, by contrast, did not reach usable performance, and the evidence obtained does not support application to dysphagia screening or monitoring in any population.

Future work follows directly from the limitations identified:

1) **Independent validation protocol:** adopt partitioning strictly at the video or subject level and *leave-one-subject-out* cross-validation, so as to obtain an unbiased estimate of between-subject generalization.
2) **Spatial calibration:** incorporate a physical referent of known scale into the acquisition protocol, allowing errors and displacements to be reported in millimeters and compared against the error ranges of the reference instrumental assessment [20]. Standardize camera distance and positioning, and compute the individual error of each of the 17 landmarks.
3) **Temporal modeling:** replace per-frame classification with models operating over sequences (e.g., long short-term memory (LSTM), temporal convolutional net-

works (TCN), or *Transformers*), explicitly incorporating marker velocity and acceleration, and substantially increase the number of annotated swallowing events, in addition to adopting explicit class-imbalance strategies (class weighting, resampling, or synthetic example generation), absent in the present configuration.

4) **Validation in a clinical population:** assemble a cohort of patients with confirmed PD, with documented staging and speech-language assessment, and validate the extracted metrics against the instrumental gold standard before any proposal of clinical use.
5) **Quantification of the reference:** compute inter-rater agreement between the annotators, so as to separate model error from the variability of the *ground truth* itself.
6) **Tracking improvements:** expand the dataset, apply temporal smoothing, and use automatic cropping of the region of interest (ROI) to mitigate the instability observed in the laryngeal and lateral-border markers.

REFERENCES

[1] I. Suttrup and T. Warnecke, "Dysphagia in Parkinson's disease," *Dysphagia*, vol. 31, no. 1, pp. 24–32, Feb. 2016.

[2] J. G. Kalf, B. J. M. de Swart, B. R. Bloem, and M. Munneke, "Prevalence of oropharyngeal dysphagia in Parkinson's disease: A meta-analysis," *Parkinsonism Relat. Disord.*, vol. 18, no. 4, pp. 311–315, 2012.

[3] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," 2019, arXiv:1906.08172.

[4] F. Schwendicke *et al.*, "Deep learning for cephalometric landmark detection: Systematic review and meta-analysis," *Clin. Oral Investig.*, vol. 25, no. 7, pp. 4299–4309, 2021.

[5] A. Mathis *et al.*, "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning," *Nat. Neurosci.*, vol. 21, no. 9, pp. 1281–1289, Sep. 2018.

[6] T. Nath *et al.*, "Using DeepLabCut for 3D markerless pose estimation across species and behaviors," *Nat. Protoc.*, vol. 14, no. 7, pp. 2152–2176, Jul. 2019.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[8] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[9] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[10] Z. Zhang, J. L. Coyle, and E. Sejdić, "Automatic hyoid bone detection in fluoroscopic images using deep learning," *Sci. Rep.*, vol. 8, art. no. 12310, 2018.

[11] Y. Khalifa, J. L. Coyle, and E. Sejdić, "Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings," *Sci. Rep.*, vol. 10, art. no. 8704, 2020.

[12] C. Donohue, S. Mao, E. Sejdić, and J. L. Coyle, "Tracking hyoid bone displacement during swallowing without videofluoroscopy using machine learning of vibratory signals," *Dysphagia*, vol. 36, no. 2, pp. 259–269, 2021.

[13] S. A. Julious, "Sample size of 12 per group rule of thumb for a pilot study," *Pharm. Statist.*, vol. 4, no. 4, pp. 287–291, 2005.

[14] Y. Kim and G. H. McCullough, "Maximum hyoid displacement in normal swallowing," *Dysphagia*, vol. 23, no. 3, pp. 274–279, Sep. 2008.

[15] Conselho Nacional de Saúde (National Health Council), "Resolution No. 466 of December 12, 2012: Guidelines and regulatory standards for research involving human beings," *Diário Oficial da União*, Brasília, DF, Brazil, Jun. 13, 2013 (in Portuguese).

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[20] I. Sia, P. Carvajal, G. D. Carnaby-Mann, and M. A. Crary, "Measurement of hyoid and laryngeal displacement in video fluoroscopic swallowing studies: Variability, reliability, and measurement error," *Dysphagia*, vol. 27, no. 2, pp. 192–197, 2012.

Investigation of Theta-Gamma Co-modulation during Sequential Learning as a Biomarker of Psychopathological Dimensions in the HBN Cohort

Naynara de Souza
Postgraduate degree in Biomedical Engineering (PPGEBM)
Federal University of ABC (UFABC)
São Paulo, Brazil
naynarasouza13@gmail.com

Ana Carolina Quintão Siravenha Muller
School of Electrical and Biomedical Engineering(FEEB)
Federal University of Pará (UFPA)
Pará, Brazil
siravenha@ufpa.br

***Abstract*—Theta-gamma coupling (TGC) is a neural mechanism associated with the integration of cognitive processes and has been proposed as a potential biomarker of psychopathological vulnerability. This study investigated the relationship between TGC and the p-factor during a sequence learning task using data from the Healthy Brain Network (HBN). The sample consisted of 141 participants aged 8–16 years. TGC was quantified in the central, frontal, frontopolar, and parietal regions using the Modulation Index between theta (4–8 Hz) and gamma (30–50 Hz) oscillations. Spearman correlations, ANOVA, multiple linear regression, and receiver operating characteristic (ROC) analyses were performed. No statistically significant associations were observed between TGC metrics and the p-factor, and the regression models and ROC analyses showed no significant effects and limited discriminative performance. Exploratory analyses identified weak correlations between PAC_slope and the p-factor, with the largest association in the central region ($\rho = 0.129$), and between Delta_PAC and Attention, also in the central region ($\rho = -0.147$); however, neither association reached statistical significance. Overall, the findings provide limited evidence of an association between TGC and transdiagnostic psychopathological vulnerability. Future studies with larger samples, longitudinal designs, and cognitive tasks targeting specific domains may help clarify the potential of TGC as a neurophysiological marker.**

***Index Terms*—theta-gamma phase-amplitude coupling, p-factor, electroencephalography, visual sequential learning.**

I. Introduction

Understanding the neurobiological architecture of psychopathology has advanced through transdiagnostic approaches, particularly with the introduction of the p-factor, a general construct that represents the shared vulnerability across different mental disorders and consistently emerges in bifactor models applied to the Child Behavior Checklist (CBCL) [1]. Evidence indicates that the p-factor is associated with alterations in the connectivity of brain networks involved in cognitive control, including the default mode network (DMN) and frontoparietal networks, and is further linked to genetic influences and the development of functional brain networks [2]–[4].

Electroencephalography (EEG) is a promising tool for investigating neural biomarkers of psychopathology because of its high temporal resolution [5], [6]. In this context, the Healthy Brain Network (HBN) provides a large multimodal dataset, including high-density EEG recordings and transdiagnostic measures such as the p-factor, enabling the investigation of the relationship between neural activity and vulnerability to psychopathology [7].

Among the tasks available in the HBN dataset, the sequence learning task is particularly relevant because it engages attention, working memory, and executive control processes [8], [9]. These cognitive processes rely on the interaction between theta (4–8 Hz) and gamma (30–80 Hz) oscillations, with theta–gamma phase-amplitude coupling (PAC) being considered a key neural mechanism for the temporal integration of information and learning [10], [11].

Based on the association between the p-factor and alterations in cognitive control networks, we hypothesized that individuals with higher p-factor scores would exhibit distinct patterns of oscillatory power and theta–gamma coupling during sequence learning. Therefore, this study investigated the relationship between the p-factor and neural activity during this task using data from the HBN cohort. Specifically, we aimed to (1) characterize differences in theta and gamma band power and theta–gamma coupling among participants with different levels of the p-factor and (2) evaluate the potential of these neurophysiological markers to predict the general dimension of psychopathology.

II. Theta-Gamma Coupling as an Integrative Mechanism of Cognitive Processes and Psychopathology

Theta-gamma coupling (TGC) has been proposed as a fundamental neural mechanism underlying the integration of higher-order cognitive processes by linking neural oscillations at different frequencies to coordinate communication across distributed brain regions [12]. Theta oscillations (4–8 Hz) are associated with the temporal organization of sequential information and the maintenance of working memory, whereas gamma oscillations (30–80 Hz) reflect local neuronal activity and information integration within specific neural

networks [10]. TGC refers to the modulation of gamma-band amplitude by the phase of theta oscillations, enabling information from multiple sources to be integrated within a common temporal framework [13].

This mechanism has been implicated in cognitive processes such as working memory, selective attention, sequence learning, and executive control—functions (particularly within frontoparietal networks) that are frequently impaired across psychiatric disorders [14]. In this context, TGC represents a unifying mechanism that may account for the shared vulnerability across different forms of psychopathology, as disruptions in this coupling could impair the temporal coordination required for efficient cognitive processing. Previous studies have reported reduced TGC in conditions such as schizophrenia, attention-deficit/hyperactivity disorder (ADHD), and autism spectrum disorder (ASD), suggesting that it may serve as a transdiagnostic marker of neural dysfunction [15], [16].

Likewise, the p-factor represents a general dimension of psychopathology that captures the shared vulnerability across multiple mental disorders and has been linked to alterations in neural circuits underlying cognitive control and emotion regulation. Consequently, alterations in TGC may reflect the functional integrity of these neural systems, supporting its potential as a neurophysiological marker of vulnerability to psychopathology [17], [18].

Previous studies have also demonstrated that TGC is sensitive to individual differences in cognitive performance and can be modulated by interventions such as cognitive training and transcranial direct current stimulation (tDCS), highlighting its potential as a biomarker of neural plasticity and as a therapeutic target in clinical populations [12], [19]. However, the relationship between TGC and psychopathology remains largely unexplored in pediatric populations, particularly in relation to the p-factor. Moreover, few studies have examined this relationship using sequence learning tasks, which rely on the integration of attentional, mnemonic, and executive processes mediated by TGC within frontoparietal networks. The present study addresses this gap by investigating the relationship between TGC and the p-factor in a sample of children and adolescents using an eight-target sequence learning task, with the aim of determining whether neural coupling during the acquisition of motor and cognitive sequences reflects general vulnerability to psychopathology.

III. Methodology

A. Dataset and Study Cohort

This study used data from the Healthy Brain Network (HBN) dataset published by [7], a large-scale pediatric and adolescent biobank collected in New York, USA, including participants aged 5–21 years and containing neurophysiological, behavioral, cognitive, and clinical measures. Participants were included if they: (i) were 8–16 years old; (ii) had available EEG recordings from the Sequence Learning 8 Targets task; and (iii) had available p-factor, attention, externalizing, and internalizing scores. Participants who did not meet these criteria or had insufficient EEG data for phase-amplitude coupling analyses were excluded. After eligibility and EEG quality-control procedures, the final sample comprised **141 participants**, consistently used across the central, frontal, frontopolar, and parietal analyses. The 8–16-year age range was selected to reduce developmental heterogeneity, given the substantial structural and functional brain maturation occurring during childhood and adolescence [20]. Although variability remains within this interval, restricting the age range was intended to reduce broader developmental differences.

The Sequence Learning 8 Targets task was selected because sequential learning involves attention, working memory, performance monitoring, and the formation of learned representations, processes associated with theta-gamma interactions [21]. Participants observed and learned a sequence of eight stimuli repeated across five learning blocks [22], [23]. As the study relied on a single sequential learning paradigm, the task was not considered to capture all cognitive processes potentially associated with psychopathological vulnerability.

The primary transdiagnostic measure was the p-factor, which reflects shared vulnerability across different psychopathological symptoms [24]. Higher scores were interpreted as indicating greater overall psychopathological burden. Attention, externalizing, and internalizing scores were additionally examined to assess more specific cognitive and psychopathological dimensions. Attention reflects attentional and inhibitory-control processes [25], whereas externalizing and internalizing dimensions broadly reflect behavioral dysregulation and symptoms related to anxiety, fear, and emotional distress, respectively [26], [27]. Continuous scores were used as the primary variables in association analyses, while group-based analyses derived from the p-factor and attention scores were treated as complementary exploratory analyses, with categorization procedures described in the statistical analysis subsection.

B. EEG Data Acquisition and Preprocessing

1) Data Acquisition: EEG data were acquired using a high-density 128-channel EGI HydroCel Geodesic Sensor Net (GSN) system, following the standardized Healthy Brain Network (HBN) acquisition protocol [7].

The central region was defined as the primary region of interest (ROI) for theta-gamma phase-amplitude coupling analyses. Complementary analyses were conducted in the frontal (E11, E12, E13, E19, E20, E24, and E124), frontopolar (E8, E21, and E25), and parietal (E61, E62, E67, E68, E77, E78, E83, and E84) regions.

These regions were selected based on their relevance to cognitive processes involved in sequential learning and theta-gamma interactions. The central region was treated as the primary ROI, whereas the frontal, frontopolar, and parietal regions were included as complementary regional analyses. For each multi-electrode ROI, the same predefined electrode sets were applied consistently across participants and analyses.

2) Preprocessing: EEG signal preprocessing was performed using the MNE-Python library. The continuous EEG

signals were band-pass filtered between 1 and 100 Hz to attenuate low-frequency drifts and high-frequency components. A 60 Hz notch filter was subsequently applied to reduce power-line interference. The signals were then re-referenced using the average reference across the available EEG channels.

Experimental events were identified using the `mne.events_from_annotations()` function. Events corresponding to the *learningBlock* markers were used to segment the recordings into five consecutive learning blocks associated with the Sequence Learning 8 Targets task.

Segments shorter than 3 s were excluded from subsequent analyses to ensure that only epochs with sufficient duration for the planned phase-amplitude coupling calculations were retained.

C. Phase-Amplitude Coupling Calculation

Phase-amplitude coupling (PAC) is a neurophysiological measure that quantifies the relationship between the phase of a low-frequency brain oscillation and the amplitude of a higher-frequency oscillation. In the present study, theta oscillations (4–8 Hz) were used as the phase-providing signal, whereas gamma oscillations (30–50 Hz) were used as the amplitude-providing signal.

PAC was computed using the TensorPAC library. The coupling method was defined using the parameter `idpac=(2,0,0)`, in which the first value specifies the PAC estimation method. Thus, the Modulation Index (MI) proposed by [28] was used to quantify theta-gamma coupling. The second value indicates that no surrogate-based computation was applied, whereas the third value indicates that no statistical correction was applied within the PAC computation.

PAC was calculated separately for each participant, brain region, and learning block. The Sequence Learning 8 Targets task was divided into five consecutive learning blocks, allowing the temporal evolution of theta-gamma coupling throughout the task to be examined.

From these values, PAC-derived measures were calculated to characterize the magnitude and temporal evolution of coupling. These measures included: (i) PAC values for each task block (*PAC_block1–PAC_block5*); (ii) the overall mean PAC value (*PAC_mean*); (iii) between-block variability (*PAC_std*); (iv) the temporal trend estimated by the slope coefficient of a linear regression across the five blocks (*PAC_slope*); (v) the difference between the last and first blocks (*Delta_PAC*); and (vi) a normalized learning index (*Learning_Index*), calculated as:

$$Learning_Index = \frac{PAC_5 - PAC_1}{PAC_1} \tag{1}$$

The flowchart presented in Figure 1 summarizes the main stages of the methodology, from EEG preprocessing to the calculation of the PAC-derived measures used in the subsequent analyses.

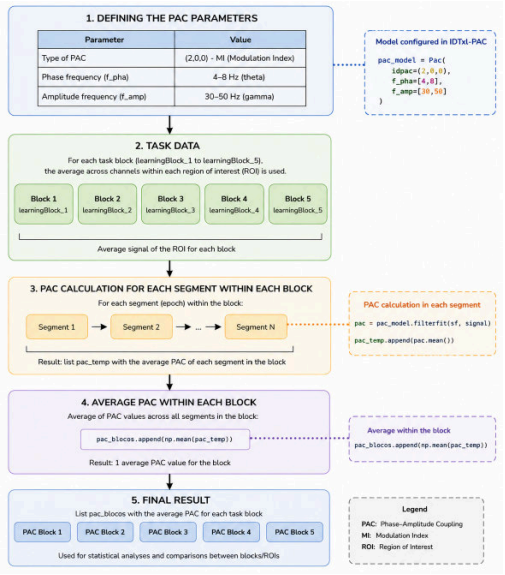


Fig. 1. Flowchart illustrating the main stages of EEG preprocessing and theta-gamma phase-amplitude coupling calculation.

IV. RESULTS

A. Sample Characterization

The final sample comprised 141 participants aged 8-16 years (mean = 11.52 years, SD = 2.09), including 85 males and 56 females. According to the predefined p-factor thresholds, participants were categorized into low (p-factor < -0.5; $n = 46$), intermediate ($-0.5 \leq$ p-factor ≤ 0.5; $n = 47$), and high (p-factor > 0.5; $n = 46$) groups. Table I summarizes the demographic and psychopathological characteristics of the sample.

TABLE I
DEMOGRAPHIC AND PSYCHOPATHOLOGICAL CHARACTERISTICS OF THE SAMPLE ($N = 141$).

Variable	Mean (SD)
Age (years)	11.18 (1.89)
Sex (M/F)	85/56
p-factor	-0.003 (0.92)
Attention	0.02 (0.85)
Internalizing	0.11 (0.76)
Externalizing	-0.08 (0.67)

B. Correlational Analyses

To investigate the relationship between theta-gamma coupling and psychopathological dimensions, Spearman correlation analyses were performed between *PAC_mean* and the p-factor, Attention, Internalizing, and Externalizing scores separately for the central, frontal, frontopolar, and parietal regions.

As shown in Figure 2, correlations between *PAC_mean* and psychopathological measures were consistently weak across all regions, with Spearman coefficients ranging from $\rho = -0.093$ to $\rho = 0.084$. The largest negative association was observed between *PAC_mean* and Attention in the frontopolar

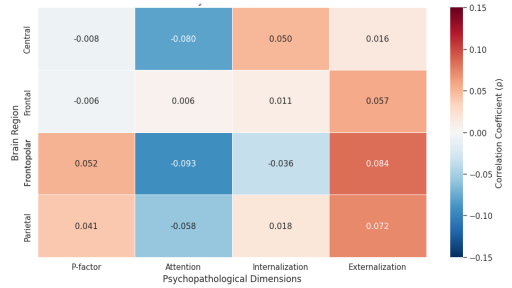


Fig. 2. Correlations between *PAC_mean* and psychopathological dimensions across brain regions.

region ($\rho = -0.093$, $p = 0.274$), whereas the largest positive association was observed between *PAC_mean* and Externalizing in the same region ($\rho = 0.084$, $p = 0.322$). Neither association was statistically significant.

After correction for multiple comparisons, none of the correlations reached statistical significance. Overall, these results indicate that mean theta-gamma phase-amplitude coupling was not significantly associated with general psychopathological vulnerability or with Attention, Internalizing, and Externalizing dimensions in the present sample.

Exploratory analyses involving additional PAC-derived measures also yielded no statistically significant associations. In the central region, *PAC_slope* showed a weak positive correlation with the p-factor ($\rho = 0.129$, $p = 0.126$), whereas *Delta_PAC* showed a weak negative correlation with Attention ($\rho = -0.147$, $p = 0.082$). Because these associations were small and did not reach statistical significance, they were not interpreted as evidence of specific neurophysiological or compensatory mechanisms.

C. Comparison Between p-factor Groups

Participants were divided into three groups according to the predefined p-factor thresholds. Table II presents the mean and standard deviation of *PAC_mean* for each group and brain region, together with the results of the one-way ANOVA and corresponding effect sizes.

TABLE II
MEANS AND STANDARD DEVIATIONS (IN PARENTHESES) OF PAC_MEAN FOR THE P-FACTOR GROUPS IN EACH BRAIN REGION, ALONG WITH THE UNIVARIATE ANOVA RESULTS (F, p, AND η^2) ($N = 141$).

Region	**Low**	**Inter.**	**High**	$F(2, 138)$	p	η^2
Central	0.00204 ± 0.0043	0.00189 ±0.0034	0.00512 ±0.0102	1.67	0.192	0.024
Frontal	0.00158 ±0.0041	0.00171 ±0.0035	0.00325 ±0.0055	1.32	0.270	0.019
Frontopolar	0.00121 ±0.0029	0.00124 ±0.0025	0.00219 ±0.0040	0.98	0.379	0.014
Parietal	0.00130 ±0.0027	0.00113 ±0.0020	0.00229 ±0.0040	1.12	0.329	0.016

As shown in Table II, no statistically significant differences in *PAC_mean* were observed among the p-factor groups in any of the four brain regions analyzed. Effect sizes were small across all regions ($\eta^2 < 0.03$), indicating that group membership accounted for only a limited proportion of the variability in *PAC_mean*.

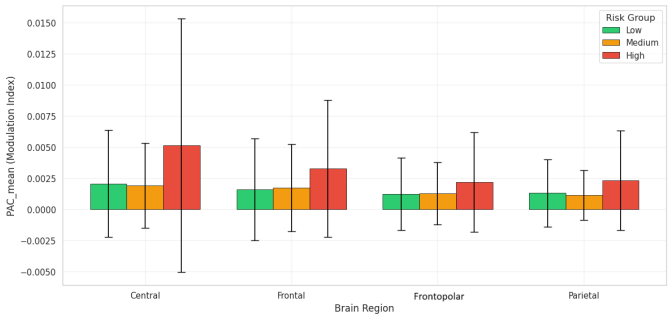


Fig. 3. Theta-gamma coupling (*PAC_mean*) across p-factor groups and brain regions.

Figure 3 shows the descriptive distribution of *PAC_mean* across the three groups. Numerically, the high p-factor group presented higher mean *PAC_mean* values than the low and intermediate groups across all regions. The largest numerical difference was observed in the central region, where mean *PAC_mean* values were 0.00512, 0.00204, and 0.00189 for the high, low, and intermediate groups, respectively.

The high p-factor group also showed larger standard deviations, particularly in the central region. However, because the group differences were not statistically significant and effect sizes were small, these descriptive differences were not interpreted as evidence of distinct neurophysiological mechanisms or greater neurophysiological heterogeneity.

D. Analysis of Temporal Dynamics of Coupling

The sequential learning task consisted of eight stimuli repeated across five consecutive blocks. For each block (*PAC_block1* to *PAC_block5*), theta-gamma phase-amplitude coupling was estimated, allowing the temporal dynamics of PAC to be examined throughout the task. A mixed repeated-measures ANOVA was performed with block (1–5) as the within-subjects factor and p-factor group as the between-subjects factor. Figure 4 presents the PAC trajectories across the five blocks for each brain region.

No statistically significant main effect of block was observed in any of the regions analyzed. Likewise, no significant block-by-group interaction was found, indicating no evidence that the temporal evolution of PAC differed among the p-factor groups.

As shown in Figure 4, descriptive differences in PAC trajectories were observed across regions and groups. In particular, the high p-factor group showed an increasing numerical pattern in the parietal region and a decreasing pattern in the frontopolar region. However, these variations were not statistically significant and were therefore interpreted only as descriptive patterns.

Figure 5 presents the distribution of *PAC_slope*, which summarizes the direction of PAC change across the five learning blocks. Descriptively, the high p-factor group showed

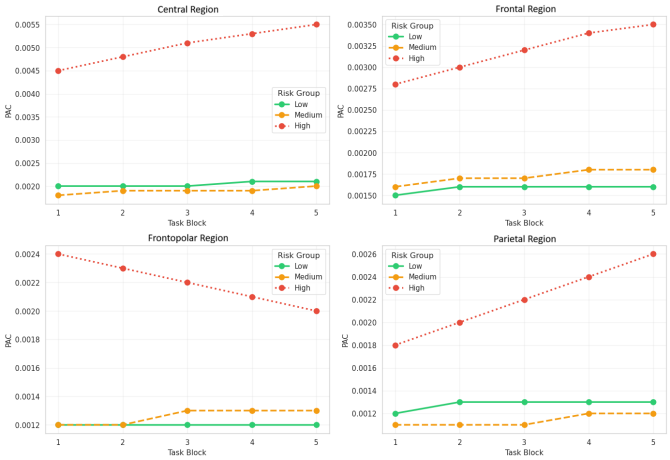


Fig. 4. Evolution of theta-gamma phase-amplitude coupling across the five task blocks.

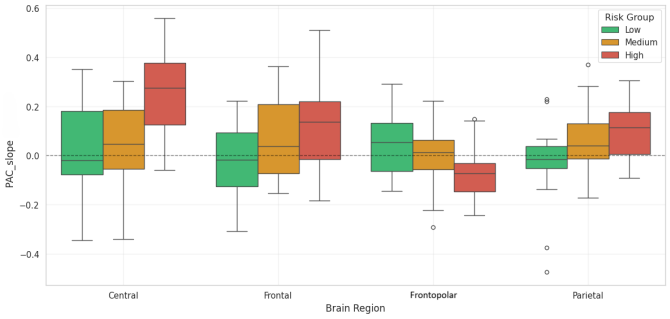


Fig. 5. Temporal slope of theta-gamma coupling (*PAC_slope*) across p-factor groups and brain regions.

positive PAC slopes in the parietal region and negative slopes in the frontopolar region, whereas the low and intermediate groups showed less pronounced changes.

Because these patterns did not reach statistical significance, they were not interpreted as evidence of differential neural recruitment, cognitive impairment, or specific neurophysiological mechanisms. Overall, the temporal analyses provided no statistically significant evidence of systematic changes in theta-gamma coupling across the task or of differences in PAC trajectories among p-factor groups.

E. Multiple Linear Regression Analyses

Multiple linear regression models were fitted to examine whether *PAC_slope* was associated with the p-factor after controlling for age and sex. Separate models were estimated for the central, frontal, frontopolar, and parietal regions. Table III presents the standardized regression coefficients (β), standard errors, t-statistics, p-values, AIC, and adjusted R^2 for each model.

Across all regions, *PAC_slope* was not significantly associated with the p-factor after adjustment for age and sex. The largest standardized coefficient was observed in the central region ($\beta = 0.128$, $p = 0.126$), although this association

TABLE III
LINEAR REGRESSION RESULTS FOR EACH CORTICAL REGION.

Region	**Predictors**	β	**Std. Error**	**t**	p	**AIC**	**Adjusted** R^2
Central	PAC_slope	0.128	0.083	1.54	0.126	412.3	0.015
	Age	-0.045	0.082	-0.55	0.583		
	Sex	0.062	0.082	0.76	0.449		
Frontal	PAC_slope	0.073	0.083	0.88	0.379	415.1	0.005
	Age	-0.048	0.083	-0.58	0.563		
	Sex	0.065	0.083	0.78	0.436		
Frontopolar	PAC_slope	0.052	0.084	0.62	0.539	416.7	0.003
	Age	-0.042	0.084	-0.50	0.618		
	Sex	0.058	0.084	0.69	0.491		
Parietal	PAC_slope	0.069	0.084	0.82	0.414	415.8	0.004
	Age	-0.046	0.084	-0.55	0.583		
	Sex	0.061	0.084	0.73	0.467		

was not statistically significant. Age and sex were also not significant predictors in any of the regional models.

The adjusted R^2 values were consistently low (< 0.02), ranging from 0.003 to 0.015, indicating that the models explained only a small proportion of the variability in the p-factor. Overall, these findings provide little evidence that *PAC_slope*, when considered together with age and sex, has substantial explanatory value for general psychopathological vulnerability in the present sample.

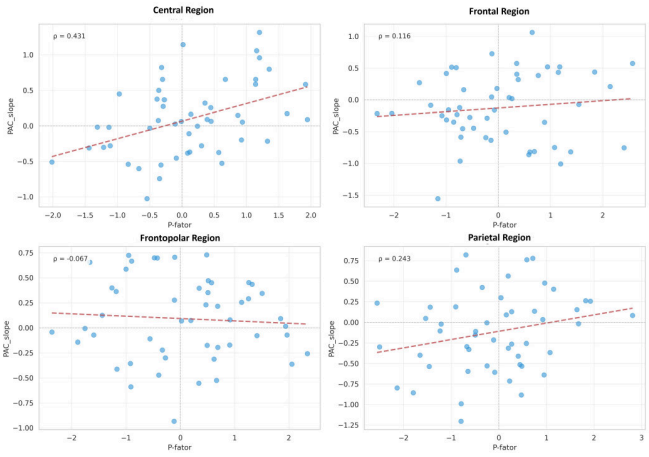


Fig. 6. Relationship between the p-factor and *PAC_slope* across brain regions.

As illustrated in Figure 6, the associations between the p-factor and *PAC_slope* were weak across all brain regions. Small positive associations were observed descriptively in the central and parietal regions, whereas the frontopolar region showed a near-zero association. However, none of these relationships reached statistical significance.

Accordingly, the observed regional differences were not interpreted as evidence of compensatory neural recruitment, altered cognitive efficiency, or region-specific neurophysiological mechanisms. Instead, the regression results indicate that temporal changes in theta-gamma coupling showed limited association with the p-factor when examined in isolation and after adjustment for age and sex.

F. ROC Curve Analysis

As illustrated in Table IV, the ROC curves for all brain regions display trajectories very close to the diagonal line representing a random classifier (AUC = 0.50). This visual pattern indicates that *PAC_mean* does not possess discriminative ability to distinguish between individuals with low and high psychopathological vulnerability. The area under the curve (AUC) values, unsurprisingly, corroborate this behavior.

TABLE IV
DISCRIMINATIVE PERFORMANCE OF *PAC_mean* BETWEEN LOW- AND HIGH-RISK GROUPS ($n = 92$).

Region	AUC	95% CI	Cutoff Point	Sensitivity	Specificity	Youden's Index
Central	0.52	[0.41; 0.63]	0.0012	0.52	0.54	0.06
Frontal	0.51	[0.40; 0.62]	0.0009	0.50	0.52	0.02
Frontopolar	0.55	[0.44; 0.66]	0.0007	0.57	0.53	0.10
Parietal	0.54	[0.43; 0.65]	0.0008	0.54	0.54	0.08

All 95% confidence intervals include the value 0.50, reinforcing the conclusion that the performance of *PAC_mean* as a classifier is not statistically superior to chance. The cutoff points ranged from 0.0007 to 0.0012, depending on the region. However, the overall low accuracy renders these values of little practical utility. The sensitivity (0.50 to 0.57) and specificity (0.52 to 0.54) values were equally low, indicating that *PAC_mean* fails to adequately identify either high-risk or low-risk individuals. Youden's index values were very low (0.02 to 0.10), confirming the reduced accuracy of *PAC_mean* as a discriminative marker.

G. Exploratory Correlation Patterns

Two descriptive correlation patterns were further examined: *PAC_slope* versus the p-factor and *Delta_PAC* versus Attention.

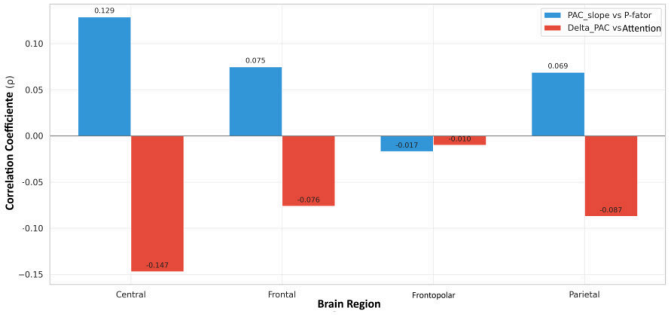


Fig. 7. Spearman correlation coefficients for *PAC_slope* versus the p-factor and *Delta_PAC* versus Attention across brain regions.

1) PAC_slope vs. p-factor: As shown in Figure 7, the largest positive correlation between *PAC_slope* and the p-factor was observed in the central region ($\rho = 0.129$). Smaller positive correlations were observed in the frontal ($\rho = 0.075$) and parietal ($\rho = 0.069$) regions, whereas the frontopolar region showed a near-zero negative correlation ($\rho = -0.017$).

Although these coefficients differed in direction and magnitude across regions, all associations were weak and did not reach statistical significance. Therefore, these regional patterns were interpreted descriptively and were not considered evidence of compensatory neural recruitment, neural desynchronization, or region-specific modulation by psychopathological vulnerability. Overall, the results provide limited evidence for an association between temporal changes in theta-gamma coupling and the p-factor in the present sample.

2) Delta_PAC vs. Attention: The same figure shows that the correlations between *Delta_PAC* and Attention were negative across all four regions. The largest negative correlation was observed in the central region ($\rho = -0.147$), followed by the parietal ($\rho = -0.087$) and frontal ($\rho = -0.076$) regions. The frontopolar region showed a near-zero negative correlation ($\rho = -0.010$).

Despite the similar direction of these coefficients, all associations were weak and did not reach statistical significance. Therefore, the observed pattern was considered descriptive and does not provide evidence that attentional difficulties are associated with reduced temporal variation in theta-gamma coupling. Likewise, the present findings do not support the conclusion that *Delta_PAC* is more sensitive to attentional processes than to general psychopathological vulnerability. Further studies are required to determine whether this pattern can be replicated in independent samples.

V. CONCLUSION

The present study examined how the general dimension of psychopathology, captured by the p-factor, is reflected in neural dynamics during a sequence learning task.

The present findings did not provide evidence of significant associations between theta-gamma coupling metrics and the psychopathological dimensions assessed across the brain regions analyzed. Correlational analyses, group comparisons, regression models, and ROC curve analyses consistently showed small effects, low explained variance, and limited discriminative performance. Taken together, these results indicate that theta-gamma coupling, when assessed in isolation during the sequential learning task, has limited utility as a neurophysiological marker of transdiagnostic psychopathological vulnerability.

The absence of robust associations also highlights the importance of considering psychopathological vulnerability as a complex and multidimensional phenomenon that may not be adequately characterized by a single electrophysiological measure. Therefore, the present findings establish a boundary on the use of theta-gamma coupling as a standalone biomarker in this experimental context rather than supporting specific neurophysiological mechanisms.

Future studies should investigate larger and longitudinal cohorts, employ cognitive paradigms more specifically targeted to relevant cognitive domains, and integrate complementary EEG measures, such as spectral power, functional connectivity, event-related potentials, and network-level features. Multimodal and multivariate approaches may provide a more comprehensive characterization of the neurophysiological mechanisms associated with psychopathological vulnerability.

REFERENCES

[1] Chen Jiang, Xuejun Gu, Sixuan Li, Xudong Zhao, Fang Cai, Yi Zheng, Kun Zhang, Linlin Ding, Sujun Hu, Tingting Zhou, et al., "Measurement properties of the child behavior checklist-10: an ultra-brief screening updated in longitudinal cohort," *Frontiers in Psychiatry*, vol. 17, pp. 1767665, 2026.

[2] Vinod Menon and Mark D'Esposito, "The role of pfc networks in cognitive control and executive function," *Neuropsychopharmacology*, vol. 47, no. 1, pp. 90–103, 2022.

[3] Derek Evan Nee, "Integrative frontal-parietal dynamics supporting cognitive control," *elife*, vol. 10, pp. e57244, 2021.

[4] Emma Sprooten, Barbara Franke, and Corina U Greven, "The p-factor and its genomic and neural equivalents: an integrated perspective," *Molecular Psychiatry*, vol. 27, no. 1, pp. 38–48, 2022.

[5] Evgenia Gkintoni, Apostolos Vantarakis, and Philippos Gourzis, "Neuroimaging insights into the public health burden of neuropsychiatric disorders: a systematic review of electroencephalography-based cognitive biomarkers," *Medicina*, vol. 61, no. 6, pp. 1003, 2025.

[6] Seokho Yun, "Advances, challenges, and prospects of electroencephalography-based biomarkers for psychiatric disorders: a narrative review," *Journal of Yeungnam Medical Science*, vol. 41, no. 4, pp. 261–268, 2024.

[7] Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, Shannon Litke, Bridget O'Hagan, Jennifer Andersen, Batya Bronstein, Anastasia Bui, Marijayne Bushey, Henry Butler, Victoria Castagna, Nicolas Camacho, Elisha Chan, Danielle Citera, Jon Clucas, Samantha Cohen, Sarah Dufek, Megan Eaves, Brian Fradera, Judith Gardner, Natalie Grant-Villegas, Gabriella Green, Camille Gregory, Emily Hart, Shana Harris, Megan Horton, Danielle Kahn, Katherine Kabotyanski, Bernard Karmel, Simon P. Kelly, Kayla Kleinman, Bonhwang Koo, Eliza Kramer, Elizabeth Lennon, Catherine Lord, Ginny Mantello, Amy Margolis, Kathleen R. Merikangas, Judith Milham, Giuseppe Minniti, Rebecca Neuhaus, Alexandra Nussbaum, Yael Osman, Lucas C. Parra, Ken R. Pugh, Amy Racanello, Anita Restrepo, Tian Saltzman, Batya Septimus, Russell Tobe, Rachel Waltz, Anna Williams, Anna Yeo, Francisco X. Castellanos, Arno Klein, Tomas Paus, Bennett L. Leventhal, R. Cameron Craddock, Harold S. Koplewicz, and Michael P. Milham, "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Scientific Data*, 2017.

[8] Pierre Barrouillet and Valérie Camos, "The time-based resource-sharing model of working memory," in *Working Memory: State of the Science*, pp. 85–115. Oxford University Press, 2021.

[9] Sidni A Justus, Patrick S Powell, and Audrey Duarte, "Intact context memory performance in adults with autism spectrum disorder," *Scientific Reports*, vol. 11, no. 1, pp. 20482, 2021.

[10] Ji Seon Ahn, Jaeseok Heo, Jooyoung Oh, Deokjong Lee, Kyungun Jhung, Jae-Jin Kim, and Jin Young Park, "The functional interactions between cortical regions through theta-gamma coupling during resting-state and a visual working memory task," *Brain Sciences*, vol. 12, no. 2, pp. 274, 2022.

[11] Nicolas Roehri, Lucie Bréchet, Martin Seeber, Alvaro Pascual-Leone, and Christoph M Michel, "Phase-amplitude coupling and phase synchronization between medial temporal, frontal and posterior brain regions support episodic autobiographical memory recall," *Brain Topography*, vol. 35, no. 2, pp. 191–206, 2022.

[12] Mohammed Abubaker, Wiam Al Qasem, Kateřina Pilátová, Petr Ježdík, and Eugen Kvašňák, "Theta-gamma-coupling as predictor of working memory performance in young and elderly healthy people," *Molecular brain*, vol. 17, no. 1, pp. 74, 2024.

[13] Dmitry O Sinitsyn, Alexandra G Poydasheva, Ilya S Bakulin, Alfiia H Zabirova, Dmitry Yu Lagoda, Natalia A Suponeva, and Michael A Piradov, "Estimating the frequencies of maximal theta-gamma coupling in eeg during the n-back task: Sensitivity to methodology and temporal instability," *Algorithms*, vol. 16, no. 12, pp. 540, 2023.

[14] Mina Mirjalili, Iryna S Palamarchuk, Heather Brooks, Reza Zomorrodi, Ashley Melichercik, Wei Wang, Sean M Nestor, Daniel M Blumberger, Abhishek Datta, Christopher Bowie, et al., "Individualized frequency and montage tacs to engage theta-gamma coupling and enhance working memory in mild cognitive impairment," *Frontiers in Psychiatry*, vol. 16, pp. 1565881, 2025.

[15] Mauro Scala, Juergen Dukart, Giuseppe De Simone, Giuseppe Fanelli, Rocío Arroyo Iturra, Sara Bordignon, Chiara Fabbri, and Alessandro Serretti, "Magnetoencephalographic signatures of facial emotion processing in schizophrenia and major depressive disorder: a systematic review," *Neuropsychobiology*, vol. 85, no. 3, pp. 180–200, 2026.

[16] Angela G. Koski, "Quantum psychiatric field hypothesis (qpfh): A bio-quantum model of neural dysconnectivity in mental health disorders," *OSF Preprints*, 2025.

[17] Haipeng Wang, Shanruo Xu, Runkun Guo, Jiang Han, and Ming-Chun Huang, "Neurosense: Bridging neural dynamics and mental health through deep learning for brain health assessment via reaction time and p-factor prediction," *Diagnostics*, vol. 16, no. 2, pp. 293, 2026.

[18] Dashiell Sacks, *EEG phase-amplitude coupling during adolescent development: a neurobiological marker for cognition, psychological distress, and mental wellbeing*, Ph.D. thesis, University of the Sunshine Coast, Queensland, 2023.

[19] Dawid Strzelczyk, Charline Peylo, and Nicolas Langer, "Impact of aging on theta-phase gamma-amplitude coupling during learning: A multivariate analysis," *bioRxiv*, pp. 2025–02, 2025.

[20] B. J. Casey, Nim Tottenham, Conor Liston, and Sarah Durston, "Imaging the developing brain: What have we learned about cognitive development?," *Trends in Cognitive Sciences*, vol. 9, no. 3, pp. 104–110, 2005, Special Issue: Developmental Cognitive Neuroscience.

[21] Natalie A. Steinemann, Clara Moisello, M. Felice Ghilardi, and Simon P. Kelly, "Tracking neural correlates of successful learning over repeated sequence observations," *NeuroImage*, vol. 137, pp. 152–164, 2016.

[22] Mauro Ursino and Gabriele Pirazzini, "Theta-gamma coupling as a ubiquitous brain mechanism: implications for memory, attention, dreaming, imagination, and consciousness," *Current Opinion in Behavioral Sciences*, vol. 59, pp. 101433, 2024.

[23] Mohammed Abubaker, Wiam Al Qasem, and Eugen Kvašňák, "Working memory and cross-frequency coupling of neuronal oscillations," *Frontiers in Psychology*, vol. 12, pp. 756661, 2021.

[24] Avshalom Caspi, Renate M. Houts, Daniel W. Belsky, Sidra J. Goldman-Mellor, HonaLee Harrington, Salomon Israel, Madeline H. Meier, Sandhya Ramrakha, Idan Shalev, Richie Poulton, and Terrie E. Moffitt, "The p factor: One general psychopathology factor in the structure of psychiatric disorders?," *Clinical Psychological Science*, vol. 2, no. 2, pp. 119–137, 2014, PMID: 25360393.

[25] Adele Diamond, "Executive functions," *Annual Review of Psychology*, vol. 64, no. Volume 64, 2013, pp. 135–168, 2013.

[26] T.M. Achenbach and L.A. Rescorla, *Manual for the ASEBA School-age Forms & Profiles: An Integrated System of Multi-informant Assessment*, Publicaciones de Psicología Aplicada: Serie Menor. ASEBA, 2001.

[27] Robert F. Krueger and Kristian E. Markon, "Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology," *Annual Review of Clinical Psychology*, vol. 2, no. Volume 2, 2006, pp. 111–133, 2006.

[28] Adriano B. L. Tort, Robert Komorowski, Howard Eichenbaum, and Nancy Kopell, "Measuring phase-amplitude coupling between neuronal oscillations of different frequencies," *Journal of Neurophysiology*, vol. 104, no. 2, pp. 1195–1210, 2010, PMID: 20463205.

Low-Density Electroencephalographic Riemannian Trajectories as Candidate Learning Markers in Motor Imagery Brain–Computer Interfaces

1st Luiz Eduardo Vedoato Almeida Everton
Federal University of Espírito Santo
Vitória, Espírito Santo, Brazil
luiz.vedoato@icloud.com

2nd Teodiano Freire Bastos-Filho
Federal University of Espírito Santo
Vitória, Espírito Santo, Brazil
teodiano.bastos@ufes.br

***Abstract*—Motor imagery brain–computer interface (MI-BCI) performance is commonly summarized by a scalar output that does not describe the evolution of the neural path. This longitudinal within-participant pilot study evaluates candidate markers derived from electroencephalographic (EEG) covariance trajectories on the manifold of symmetric positive definite (SPD) matrices. Eight sessions from a healthy 24-year-old man, collected over 30 days, were analyzed using a 16-channel EEG-cap and continuous pneumatic-glove feedback. Successive covariances in the 8–30 Hz band were compared with the affine-invariant Riemannian metric. Dispersion quantified compaction between the first and third training blocks, as well as persistence between the feedback-free periods before (PRE) and after (POST) training; length, curvature, and straightness complemented the description. Six sessions showed compaction, with a median reduction of 0.119 and a rank-biserial effect size of -0.611, although the result has been inconclusive (p = 0.148). Four sessions showed lower dispersion in POST, without consistent persistence (p = 0.844). In an exploratory analysis, mean controller target aperture was not associated with compaction (Spearman's rho = 0.024, p = 0.977). The results support Riemannian trajectories as complementary descriptors of neural dynamics during the protocol and provide an auditable framework for larger studies.**

***Index Terms*—brain–computer interface, motor imagery, neural manifolds, neural trajectories, Riemannian geometry.**

I. Introduction

Motor imagery brain–computer interfaces (MI-BCIs) convert voluntary modulations of brain activity into external commands [1]. During MI, changes in sensorimotor mu and beta rhythms can be observed with electroencephalography (EEG), particularly over frontocentral and centroparietal regions [2]. Training seeks to produce these patterns voluntarily, stably, and reproducibly, involving adaptation by both the user and the system [3], [4]. Visual feedback or feedback delivered through robotic effectors adds a sensorimotor consequence to practice, which is relevant to neurorehabilitation [5]–[7].

Classification accuracy is frequently used to quantify learning in MI-BCIs [3], [8], although it represents the joint performance of the user and the algorithm and depends on calibration, model choice, decision threshold, and EEG nonstationarity [8]. Decoder adaptation can increase accuracy without a consistent neural change, whereas neural reorganization may not alter a binary decision. Within-class Riemannian changes can accompany skill acquisition and stabilization even when recalibration is infrequent [9]. Classifier-independent stability and separability metrics have therefore been proposed to assess user evolution rather than only the machine decision [10]–[12].

In multichannel EEG, each window describes a joint configuration across the electrodes; ordering these configurations forms a path between states and allows for analysis of both the state reached and the transitions throughout the task. Intracortical studies show that population evolution contains sensorimotor information absent from isolated observations [13], whereas latent-state models represent motor activity as low-dimensional trajectories [14]–[16]. Dynamic mode decomposition also extracts coherent spatiotemporal patterns [17]. These findings do not establish equivalence between intracortical populations and scalp EEG, but support the methodological principle of jointly preserving the spatial organization and temporal ordering of the signals. Explicit descriptors also favor the balance between flexibility and interpretability recommended for data-driven modeling in neuroengineering [18].

Motor learning can reorganize variability within pre-existing neural subspaces [19], [20], and MI-BCI studies have identified low-dimensional representations that are sensitive to training and rehabilitation [21], [22]. In EEG, a covariance matrix represents the multichannel configuration and lies on the manifold of symmetric positive definite (SPD) matrices [23], [24]. Riemannian methods assign one covariance matrix to each data segment for classification [25], [26] or user-stability assessment [11], [12], whereas manifold analyses emphasize state embeddings [21], [22]. Here, successive covariance matrices are linked into a trajectory, preserving both geometry and temporal order.

This study characterizes neural trajectories at three scales: during training, after feedback withdrawal, and across sessions. Each temporal segment is represented as a trajectory on the SPD manifold; dispersion quantifies reproducibility, whereas length, curvature, straightness, geodesic progress, and orthogonal deviation describe path shape and direction. The primary

This work was supported by the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES), Brazil, through research funding and a scholarship.

hypothesis is that greater stabilization of MI corresponds to more compact trajectories. As a complementary hypothesis, trajectories oriented toward the target state should exhibit lower curvature and orthogonal deviation, together with greater straightness and geodesic progress.

A. EEG representation on the SPD manifold

Let a multichannel EEG window be $\boldsymbol{X}_t \in \mathbb{R}^{C\times N}$, with C channels and N samples. After centering, its sample covariance $\boldsymbol{S}_t$ is shrinkage-regularized:

$$\boldsymbol{C}_t = (1-\lambda)\boldsymbol{S}_t + \lambda\frac{\mathrm{tr}(\boldsymbol{S}_t)}{C}\boldsymbol{I}, \qquad \boldsymbol{C}_t \in \mathcal{S}_{++}^{C}. \tag{1}$$

Here, $\boldsymbol{I}$ is the identity matrix, $\mathrm{tr}(\cdot)$ denotes the trace, and $\lambda \in [0,1]$ is the shrinkage intensity, which preserves positive definiteness and numerical stability. Because the SPD cone is not a vector space, Euclidean differences between covariance matrices can distort its structure. We therefore used the affine-invariant Riemannian metric (AIRM) [24], [27]:

$$d_{\mathrm{R}}(\boldsymbol{P},\boldsymbol{Q}) = \left\|\log\!\left(\boldsymbol{P}^{-1/2}\boldsymbol{Q}\boldsymbol{P}^{-1/2}\right)\right\|_{\mathrm{F}}, \tag{2}$$

for $\boldsymbol{P},\boldsymbol{Q} \in \mathcal{S}_{++}^{C}$, where $\log(\cdot)$ is the matrix logarithm and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. This metric preserves distances under invertible congruence transformations. A set of covariance matrices can be summarized by a geometric mean, defined as the minimizer of the sum of squared distances,

$$\bar{\boldsymbol{C}} = \arg\min_{\boldsymbol{C}\in\mathcal{S}_{++}^{C}} \sum_{i=1}^{n} d_{\mathrm{R}}^{2}(\boldsymbol{C},\boldsymbol{C}_i), \tag{3}$$

where n is the number of matrices in the set. This construction yields a centroid consistent with the geometry of the space.

B. Trajectories and geometric markers

The temporal sequence of covariance matrices for segment i is denoted by $\mathcal{T}_i$. For two trajectories aligned across their T windows, we defined

$$d_{\mathrm{T}}(\mathcal{T}_i,\mathcal{T}_j) = \frac{1}{T}\sum_{t=1}^{T} d_{\mathrm{R}}(\boldsymbol{C}_{i,t},\boldsymbol{C}_{j,t}). \tag{4}$$

The dispersion of a block B with n_B segments is the mean distance over all pairs,

$$D_B = \frac{2}{n_B(n_B-1)}\sum_{i<j} d_{\mathrm{T}}(\mathcal{T}_i,\mathcal{T}_j). \tag{5}$$

The three feedback training blocks are denoted B1, B2, and B3, and the feedback-free periods before and after training are denoted PRE and POST. Thus, $\Delta D_{\mathrm{train}} = D_{\mathrm{B3}} - D_{\mathrm{B1}} < 0$ characterizes within-session compaction, whereas $\Delta D_{\mathrm{persist}} = D_{\mathrm{POST}} - D_{\mathrm{PRE}} < 0$ indicates persistence of greater consistency after feedback withdrawal.

To describe path shape, each covariance matrix is projected onto the tangent space at a reference $\boldsymbol{R}$ as $\boldsymbol{z}_t = \mathrm{vec}_{\mathrm{sym}}\big[\log(\boldsymbol{R}^{-1/2}\boldsymbol{C}_t\boldsymbol{R}^{-1/2})\big]$, where $\mathrm{vec}_{\mathrm{sym}}$ denotes the isometric vectorization of the independent elements of a symmetric matrix. Length L and straightness S are

$$L = \sum_{t=1}^{T-1}\|\boldsymbol{z}_{t+1}-\boldsymbol{z}_t\|_2, \qquad S = \frac{\|\boldsymbol{z}_T-\boldsymbol{z}_1\|_2}{L}. \tag{6}$$

With $\Delta\boldsymbol{z}_t = \boldsymbol{z}_{t+1}-\boldsymbol{z}_t$, curvature was defined as the mean, in radians, of the angles between $\Delta\boldsymbol{z}_t$ and $\Delta\boldsymbol{z}_{t+1}$, computed from the normalized cosine. For an expected direction $\boldsymbol{v}$ from the reference to the MI target state, progress and orthogonal deviation are

$$p_t = \frac{\langle\boldsymbol{z}_t,\boldsymbol{v}\rangle}{\|\boldsymbol{v}\|_2^2}, \qquad o_t = \left\|\boldsymbol{z}_t - \frac{\langle\boldsymbol{z}_t,\boldsymbol{v}\rangle}{\|\boldsymbol{v}\|_2^2}\boldsymbol{v}\right\|_2. \tag{7}$$

To visualize the local direction of neural movement, the plane formed by the first two tangent components was divided into a grid. In each cell g, the number of states n_g and the sum $\boldsymbol{s}_g$ of consecutive displacements were accumulated. After discrete Gaussian smoothing G_σ, we defined

$$\widehat{\boldsymbol{f}}(\boldsymbol{x}_g) = \frac{G_\sigma(\boldsymbol{s})_g}{\max[G_\sigma(\boldsymbol{n})_g,\varepsilon]}. \tag{8}$$

Here, $\boldsymbol{x}_g$ is the center of cell g, and $\varepsilon > 0$ is a numerical stabilization constant. The field estimates mean displacement in visited regions without assuming an autonomous system or attractors. Together, these measures separate consistency, path efficiency, and the direction of reorganization.

A second visualization summarizes occupancy through an empirical potential,

$$V(\boldsymbol{x}) = -\log[\widehat{p}(\boldsymbol{x})+\varepsilon] - \min_{\boldsymbol{u}}\{-\log[\widehat{p}(\boldsymbol{u})+\varepsilon]\}, \tag{9}$$

where $\widehat{p}$ is a kernel density estimate and $\boldsymbol{u}$ ranges over the domain points used to define the minimum. Lower values indicate more frequently visited regions, not physiological energy. This deliberately descriptive interpretation distinguishes these representations from methods intended to identify operators, delay models, or governing equations [17], [28]. The protocol sequence, EEG montage, and analytical trajectory pipeline are shown in Figs. 1, 2, and 3, respectively.

II. Materials and Methods

A. Study design and protocol

A longitudinal within-participant pilot study was conducted with a healthy 24-year-old man. Eight sessions (S1–S8) performed over 30 days were analyzed. In every session, the participant received standardized instructions to kinesthetically imagine opening his right hand without executing a voluntary movement, while a pneumatic glove assisted hand opening during feedback. The study was conducted under a protocol approved by the Research Ethics Committee of the Federal University of Espírito Santo, Brazil, under Certificate of Presentation for Ethical Review (CAAE) 91161425.6.0000.5060. The procedures complied with the Declaration of Helsinki, and written informed consent was obtained from the participant before any study-related procedure.

Each session began with 180 s of rest, followed by a 70-s PRE period of motor imagery without feedback. After a short preparation interval, three 70-s training blocks with continuous feedback, B1, B2, and B3, were performed. A 90-s rest interval separated B1 from B2 and B2 from B3. Following B3, a 3–5-min rest period preceded POST, which comprised 70 s of MI without feedback, as shown in Fig. 1.

Before training, a session-specific calibration related sensorimotor changes in the 8–30 Hz band to MI evidence. During B1–B3, this evidence was continuously converted into a target spanning 0–80% of the glove's safe range of motion; stronger evidence produced a larger target aperture, and pneumatic motion provided visual and kinesthetic feedback. The control logic evolved from an adaptive configuration in S1 to an anti-saturation ramp in S2–S6 and a geodesic gate in S7–S8. These changes did not enter the offline geometric computation, but limited direct comparison of controller outputs. PRE and POST retained the same instruction, without glove movement or decoder output. Contrasts between B1 and B3, PRE and POST, and across sessions defined within-session adaptation, feedback-free persistence, and longitudinal evolution, respectively.

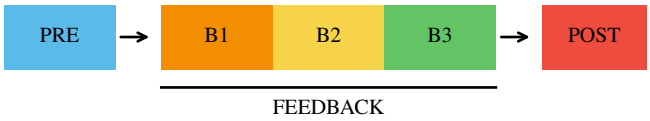


Fig. 1. Analytical block sequence. PRE and POST were performed without feedback, whereas B1–B3 used proportional pneumatic-glove feedback. Rest intervals are omitted from the diagram and described in the text.

B. EEG acquisition

EEG was acquired with an OpenBCI Cyton+Daisy system using 16 channels at a sampling rate of 125 samples/s. FCz was connected to the common SRB2 input: in the referential mode of the ADS1299 converter each channel was digitized directly as a potential difference relative to FCz. AFz was connected to the BIAS input, which was used for common-mode stabilization and interference reduction rather than as a second reference subtracted offline [29]. The montage comprised FC3, FC1, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, and CP4. Geometric analysis used ten frontocentroparietal channels: FC3, FC1, FC2, FC4, C3, Cz, C4, CP3, CPz, and CP4 (Fig. 2).

C. Preprocessing and trajectory construction

Temporal integrity was verified from the sample index and effective sampling rate. A 60-Hz notch filter with quality factor $Q = 30$ and a fourth-order, zero-phase 1–40 Hz Butterworth bandpass filter were applied. For geometric analysis, the ten selected channels were filtered between 8 and 30 Hz. The first 5 s of each 70-s phase were discarded; 64 s of the remaining 65 s were divided into eight contiguous 8-s temporal segments. These segments were standardized analysis units within PRE, B1, B2, B3, and POST rather than independent behavioral trials.

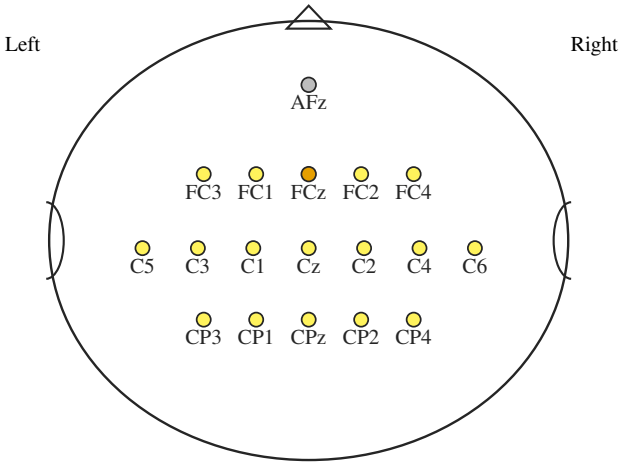


Fig. 2. EEG montage used in the study, redrawn as a vector diagram. The labeled yellow markers indicate the 16 acquisition sites; FCz was the physical reference and AFz the ground electrode.

Channels were evaluated for flatness, extreme amplitude, line-noise content, and derivative peaks. Each temporal segment was checked for robust amplitude, power ratios in the 1–4 and 30–40 Hz bands, nonfinite values, and covariance conditioning. The primary thresholds were a robust peak amplitude of 200 μV, a maximum robust score of 3.5, and a condition number below 10^6. PRE and POST were required to retain at least seven of the eight valid segments.

Each temporal segment was decomposed into 2-s windows with a 1-s step. Covariance matrices were regularized with $\lambda = 0.10$ and compared using the AIRM. Trajectory distance was computed after alignment by temporal index. For visualization and shape markers, the matrices were mapped to the tangent space at a reference estimated from B1–B3 and reduced to three principal components. When computing the distance from one segment to its phase centroid, the centroid was estimated using only the remaining segments from that phase, preventing the segment from contributing to its own reference. On the plane of the first two components, occupancy counts and consecutive displacements were accumulated on a 28×28 grid and smoothed with a Gaussian filter of $\sigma = 1.25$ cells. The occupancy and field defined in (8) were used as descriptive representations.

To obtain a common three-dimensional visualization across all eight sessions, the 2,170 tangent states were first subjected to principal component analysis (PCA), retaining at least 95% of the variance; 33 components preserved 95.3%. Diffusion maps with 25 neighbors were then applied [30], and the first three nontrivial coordinates, ψ_1, ψ_2, and ψ_3, defined the embedding. A principal surface $\psi_3 = f(\psi_1, \psi_2)$ was fitted with radial basis functions, and its color encoded kernel-estimated local density. For each phase of S8, the observed trajectory with the smallest mean distance to the other trajectories in that phase was selected. This representative trajectory was projected onto the surface while preserving temporal order. The colored ribbons follow

the local tangent plane only to aid depth perception; they do not represent geodesics or introduce a new metric.

The stability of this construction was assessed by varying PCA retention from 90% to 98% and the number of neighbors from 15 to 45. We also performed 200 bootstrap resamples stratified by session and phase while preserving complete temporal trajectories, eight leave-one-session-out reconstructions, and three levels of surface smoothing. Comparisons used Procrustes alignment, Spearman correlation between paired distances, and overlap among the ten nearest neighbors.

To compare centroids across sessions, a common log-Euclidean tangent reference and PCA were estimated from B1–B3 in all eight sessions; each session was then centered by its own training centroid. PRE, B1, B2, B3, and POST centroids were computed in this global system. Ninety-five-percent ellipses for the means were obtained from 4,000 bootstrap resamples; the first two axes explained 77.1% of the tangent-space training variance. The pipeline used to construct each trajectory is summarized in Fig. 3.

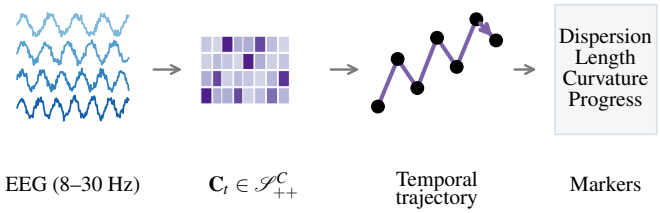


Fig. 3. Schematic representation of a temporal segment. Multichannel EEG is divided into windows; each window yields an SPD covariance matrix, and the ordered sequence preserves the trajectory used to compute the markers.

D. Markers and statistical analysis

The primary marker of within-session adaptation was $\Delta D_{\text{train}} = D_{\text{B3}} - D_{\text{B1}}$; negative values indicated greater compaction by the end of training. Feedback-free persistence was assessed by $\Delta D_{\text{persist}} = D_{\text{POST}} - D_{\text{PRE}}$, likewise oriented toward negative values. POST dispersion across session order described longitudinal evolution. In offline analysis of the acquired data, length, straightness, and curvature were computed as complementary descriptors of trajectory shape. Geodesic progress and orthogonal deviation were formalized as candidate directional markers but were not included in the primary longitudinal contrasts. Event-related desynchronization/synchronization (ERD/ERS) in the mu (8–12 Hz) and beta (18–24 Hz) bands relative to rest was also computed offline as a physiological control.

The session was retained as the inferential unit because temporal segments were nested within each phase. Paired contrasts used the exact two-sided Wilcoxon signed-rank test; the median, paired rank-biserial effect size (r_{rb}), mean, and 95% confidence interval (95% CI) were reported. Trends over session order were evaluated using Spearman correlation with exact permutation and the Theil–Sen slope. Spectral trends and the relationship between compaction and mean controller target aperture were treated as exploratory rather than confirmatory analyses.

III. Results

A. Session integrity

All eight sessions contained the required phases. Out of the 320 expected temporal segments, 310 were valid (96.9%). All 16 channels passed the post-filtering criteria in every session, and the effective sampling rate remained 125 samples/s without breaks in the sample index. In PRE and POST, every session retained at least seven of the eight temporal segments.

B. Within-session compaction and persistence

Six of the eight sessions ended B3 with lower dispersion than B1 (Table I). The difference had a median of -0.119, a mean of -0.088 (95% CI: -0.191 to 0.015), and $r_{\text{rb}} = -0.611$. The pattern was directionally consistent in most sessions but inconclusive (Wilcoxon $p = 0.148$). The largest relative compactions occurred in S8 (13.9%), S1 (13.7%), and S5 (13.4%).

In the PRE–POST contrast, four sessions showed lower dispersion, with a median of -0.021, a mean of -0.008 (95% CI: -0.176 to 0.160), and $r_{\text{rb}} = -0.111$ ($p = 0.844$; Table II). S3 and S8 showed the largest individual reductions, 19.0% and 11.8%, respectively; no consistent persistence after feedback withdrawal was detected.

Table I
WITHIN-SESSION COMPACTION ACROSS THE EIGHT SESSIONS

Session	B1	B3	$\Delta_{\text{B3}-\text{B1}}$
S1	1.553	1.340	-0.213
S2	1.317	1.399	$+0.082$
S3	1.278	1.205	-0.073
S4	1.320	1.265	-0.055
S5	1.350	1.169	-0.180
S6	1.483	1.318	-0.165
S7	1.387	1.484	$+0.097$
S8	1.428	1.229	-0.198
Median Δ; 6/8 decreased			-0.119
Exact two-sided Wilcoxon			$p = 0.148$

Table II
PRE–POST PERSISTENCE ACROSS THE EIGHT SESSIONS

Session	PRE	POST	$\Delta_{\text{POST}-\text{PRE}}$
S1	1.557	1.481	-0.076
S2	1.317	1.331	$+0.014$
S3	1.442	1.168	-0.274
S4	1.126	1.521	$+0.395$
S5	1.348	1.304	-0.044
S6	1.374	1.477	$+0.102$
S7	1.367	1.368	$+0.001$
S8	1.525	1.345	-0.180
Median Δ; 4/8 decreased			-0.021
Exact two-sided Wilcoxon			$p = 0.844$

C. Trajectories across multiple scales

To relate within-phase dispersion, across-session organization, and trajectory shape, the data were examined at three complementary scales. Session S8 is presented as an illustrative,

rather than inferential, case because it combined substantial B1–B3 compaction with a PRE–POST reduction; all inferential contrasts remained based on the eight sessions.

At the first scale, 38 out of the 40 temporal segments from S8 were valid, with 13.9% B1–B3 compaction and an 11.8% PRE–POST reduction. The distance to the phase centroid, estimated without the segment itself, decreased in B3 (Fig. 4); the pairwise distance matrix likewise showed phase organization and greater homogeneity in B3 (Fig. 5). These summaries capture related but nonidentical properties: centroid distance quantifies concentration around a phase-specific reference, whereas the full matrix reveals the relative organization of all trajectories without reducing each phase to a single center. Their agreement indicates that the S8 pattern was not dependent on only one descriptive summary.

At the second scale, the global projection placed the mean B1–B3 centroids in a region distinct from the feedback-free periods, whereas PRE and POST remained close (Fig. 6). Within-session lines and bootstrap intervals indicated heterogeneity around the mean path. The proximity of PRE and POST is consistent with the small PRE–POST effect, whereas the displacement of the feedback blocks suggests a context-dependent organization during glove use. This projection summarizes relationships across sessions, but it does not replace the paired AIRM contrasts or establish a retained physiological state.

At the third scale, trajectories in the common embedding traversed contiguous regions while differing in extent and orientation (Fig. 7). In S8, B3 extended beyond the region shared with B1, whereas POST followed a longer path than PRE. Because temporal order is preserved, this view distinguishes two paths that may occupy nearby regions yet traverse them differently. It therefore complements dispersion by showing how the sequence evolves, without treating visual separation as an inferential result.

The global configuration remained stable in the sensitivity analyses. Median Spearman correlations of the global distances were 0.997 when varying the parameters, 0.986 under bootstrap resampling, and 0.997 when excluding one session. In contrast, local neighborhoods were less stable, with median Jaccard indices of 0.210, 0.125, and 0.244, respectively. Thus, the large-scale ordering was reproducible, whereas the identity of individual nearest neighbors was sensitive to analytical perturbations.

Across the eight sessions, the four PRE–POST reductions occurred among the six sessions that compacted during feedback. However, S4 and S6 compacted between B1 and B3 without retaining lower dispersion in POST. Within-session adaptation and feedback-free persistence were therefore related in this sample but were not equivalent outcomes.

The embedding surface should consequently be read as a descriptive map rather than a mechanistic model. The ribbons preserve temporal order and permit comparison of path extent and orientation, but do not represent geodesics, attractors, or physiological boundaries. Conclusions about fine surface details remain limited by local instability, and the paired AIRM contrasts across all eight sessions remain the primary evidence.

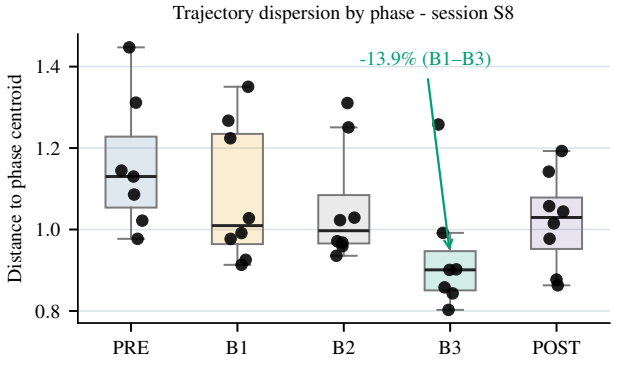


Fig. 4. Distance from each temporal segment to the estimated phase centroid in S8. Boxes show the median and interquartile range.

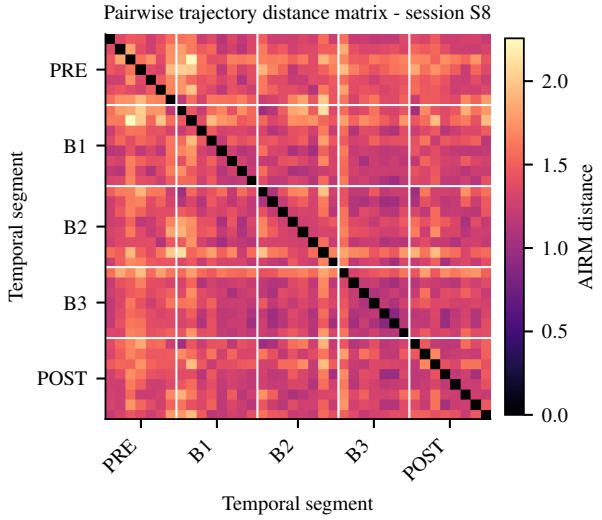


Fig. 5. AIRM distance matrix among the 38 valid trajectories from S8. White lines delimit PRE, B1, B2, B3, and POST.

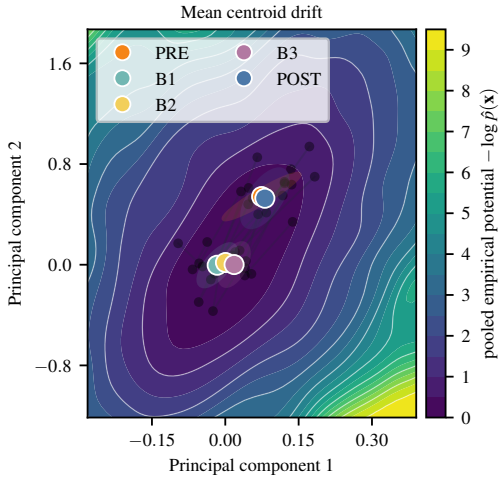


Fig. 6. Centroid drift across the eight sessions after global tangent referencing and PCA. Lines: sessions; points: phase means; ellipses: 95% bootstrap CIs.

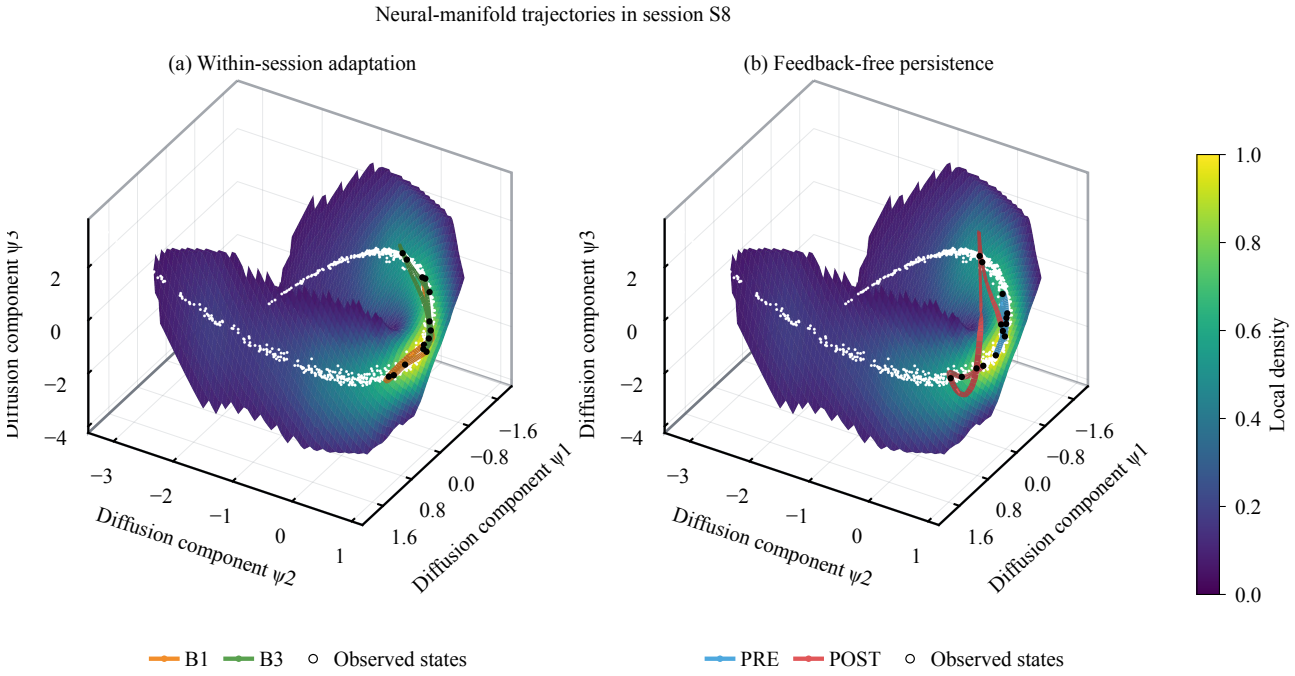


Fig. 7. S8 trajectories on the principal surface estimated in the common embedding of all eight sessions. (a) B1 and B3: within-session adaptation; (b) PRE and POST: feedback-free persistence. Color encodes local density; white points indicate observed states; colored ribbons show medoid-selected trajectories projected onto the surface. Temporal order is preserved, but the paths do not represent geodesics.

D. Physiological drift and specificity controls

There was no monotonic trend in POST dispersion across the eight sessions ($\rho = -0.024$; $p = 0.977$). In an exploratory analysis, beta ERD decreased with session order ($\rho = -0.905$; $p = 0.0046$; Fig. 8), suggesting that compaction, persistence, and physiological drift may operate on different scales and require independent confirmation.

Mean controller target aperture during B1–B3 ranged from 0.161 to 0.416, with no detectable monotonic association with within-session compaction ($\rho = 0.024$; $p = 0.977$; Fig. 9). Thus, a larger scalar controller output did not necessarily correspond to greater neural compaction. This comparison is exploratory and does not evaluate classifier accuracy.

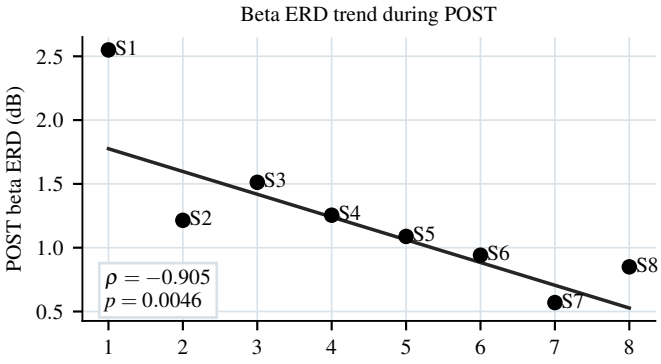


Fig. 8. Beta ERD during POST and its robust trend over session order. The line represents the Theil–Sen slope.

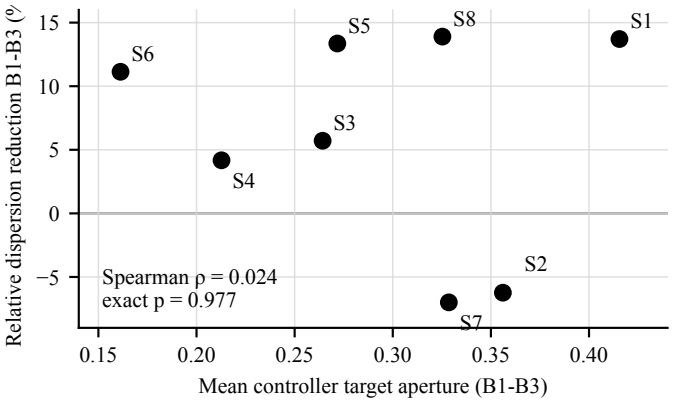


Fig. 9. Exploratory relationship between mean controller target aperture during B1–B3 and within-session compaction. Each point represents one session; positive vertical-axis values indicate lower dispersion in B3. No monotonic association was detected ($\rho = 0.024$; $p = 0.977$).

IV. Discussion

This study examines MI-BCI training through the temporal geometry of neural responses. The main observation was recurrent within-session compaction: six out of the eight sessions showed lower dispersion in B3 than in B1. The repeated direction and moderate rank-biserial effect support trajectory dispersion as a candidate marker for future validation, although the primary contrast remains nonconfirmatory. The central contribution is not a new classifier, but measures that separate practice consistency, feedback-free persistence, and longitudinal drift.

A. Compaction as a candidate marker

Paired dispersion directly measures the reproducibility of the neural path. Two blocks may produce the same scalar output, yet differ in how consistently the user traverses the MI state space. The absence of a detectable association between mean target aperture and compaction (Fig. 9) is compatible with this complementarity, but it does not demonstrate independence or superiority of the geometric metrics. This distinction extends classifier-independent stability metrics [10], [11]: instead of summarizing each segment by a single covariance matrix, the definition in (4) preserves the sequence of states. In longitudinal training, Tortora *et al.* found that within-class Riemannian changes tracked skill acquisition and stabilization better than conventional metrics [9]; the present result extends this principle from covariance distributions to the reproducibility of complete trajectories.

This formulation brings macroscopic EEG analysis closer to the population-dynamics perspective, according to which the temporal evolution of joint activity and latent factors carries relevant motor information [13]–[15], [31]. The extension is conceptual rather than an equivalence of scale: EEG covariances do not directly recover intracortical neural states. Nevertheless, the compaction observed in several sessions is compatible with stabilization during practice and with the hypothesis of manifold-constrained learning [19], [20]. The term *candidate marker* remains necessary because habituation or a nonspecific reduction in variability could also produce consistency.

B. Learning as a multiscale phenomenon

The B1–B3 contrast captures adaptation during feedback; PRE–POST asks whether the change persists without the effector; and session order tests longitudinal stabilization. The lack of equivalence among these contrasts is informative. One session may compact during training and return to its previous regime in POST, suggesting a context-dependent effect, whereas another may preserve the reorganization. The mean proximity of PRE and POST in Fig. 6, together with the displacement of the trained blocks, reinforces this multiscale interpretation. In Fig. 7, both contrasts are visualized on the same data-estimated surface, allowing path extent and orientation to be compared without confusing them with the primary inferential measure. The proposed framework prevents within-session improvement from being labeled automatically as retention or consolidation.

Length, curvature, and straightness add a second interpretive dimension. Lower dispersion indicates that repetitions became closer; it does not establish whether the path became more direct, shorter, or better oriented toward the target. Geodesic progress and orthogonal deviation allow these properties to be tested relative to an a priori direction. Differential geometry and delay models provide related tools for organizing observed trajectories and dynamics [28], whereas modal decomposition methods seek coherent spatiotemporal patterns [17]. Here, flow, potential, and surface were kept descriptive: the contribution is to quantify and visualize the path, not to claim recovery of governing equations.

C. Nonstationarity as part of the phenomenon

The longitudinal beta trend was stronger than the across-session geometric trend, although its exploratory nature requires replication. This contrast illustrates why separation across scales and physiological controls are necessary. Changes in fatigue, attention, strategy, posture, or baseline state can alter covariance structure and spectral power. At another observational scale, Raut *et al.* showed that arousal can organize the evolution of spatiotemporal brain measures [32]; the relevant inference is that global states should be measured rather than conflated with learning. A robust marker should therefore combine geometric consistency, feedback-free persistence, physiological specificity, and stability under global changes in brain state.

D. Limitations

The within-participant design provides longitudinal resolution but does not support population-level generalization. The fixed order does not separate training, habituation, and fatigue. Daily recalibration accommodated EEG nonstationarity; nevertheless, the three controller configurations limit absolute comparisons of controller output and causal attribution of the geometric pattern to feedback. Low-density EEG and the absence of surface electromyography (sEMG) or synchronized video restrict separation among cortical activity, muscle tension, and subtle movement. The session was retained as the inferential unit because segments were nested within blocks. Although the global embedding geometry was stable, the lower stability of exact neighborhoods argues against interpreting surface details as physiological states or attractors.

V. Conclusion

Riemannian covariance trajectories enable neural dynamics to be tracked during MI-BCI training while separating within-session compaction, feedback-free persistence, and longitudinal drift. Lower dispersion from B1 to B3 in six out of the eight sessions, with a moderate directional effect, identifies a promising but still nonconfirmatory candidate marker. Consistent PRE–POST persistence was not detected, and compaction showed no monotonic correspondence with mean controller target aperture. Together, geometric measures, shape descriptors, and ERD/ERS provide a characterization complementary to the system's scalar output and an auditable basis for larger studies using a frozen protocol.

Acknowledgment

The authors thank the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES), Brazil, for research funding and a scholarship. Generative artificial intelligence (AI) tools were used exclusively to support translation, editorial review, and compliance checking. The authors fully reviewed the methods, results, and scientific content and take responsibility for them.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[3] S. Perdikis and J. del R. Millán, "Brain–machine interfaces: A tale of two learners," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 3, pp. 12–19, 2020.

[4] C. Jeunet, E. Jahanpour, and F. Lotte, "Why standard brain–computer interface (BCI) training protocols should be changed: An experimental study," *Journal of Neural Engineering*, vol. 13, no. 3, p. 036024, 2016.

[5] X. Duan, S. Xie, X. Xie, K. Obermayer, Y. Cui, and Z. Wang, "An online data visualization feedback protocol for motor imagery-based BCI training," *Frontiers in Human Neuroscience*, vol. 15, p. 625983, 2021.

[6] J. Cantillo-Negrete, R. I. Carino-Escobar, P. Carrillo-Mora, M. A. Rodriguez-Barragan, C. Hernandez-Arenas, J. Quinzaños-Fresnedo, I. R. Hernandez-Sanchez, M. A. Galicia-Alvarado, A. Miguel-Puga, and O. Arias-Carrion, "Brain–computer interface coupled to a robotic hand orthosis for stroke patients' neurorehabilitation: A crossover feasibility study," *Frontiers in Human Neuroscience*, vol. 15, p. 656975, 2021.

[7] M. Maier, B. R. Ballester, and P. F. M. J. Verschure, "Principles of neurorehabilitation after stroke based on motor learning and brain plasticity mechanisms," *Frontiers in Systems Neuroscience*, vol. 13, p. 74, 2019.

[8] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[9] S. Tortora, G. Beraldo, F. Bettella, E. Formaggio, M. Rubega, A. D. Felice, S. Masiero, R. Carli, N. Petrone, E. Menegatti, and L. Tonin, "Neural correlates of user learning during long-term BCI training for the Cybathlon competition," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 1, p. 69, 2022.

[10] F. Lotte and C. Jeunet, "Defining and quantifying users' mental imagery-based BCI skills: A first step," *Journal of Neural Engineering*, vol. 15, no. 4, p. 046030, 2018.

[11] N. Ivanov and T. Chau, "Riemannian geometry-based metrics to measure and reinforce user performance changes during brain–computer interface user training," *Frontiers in Computational Neuroscience*, vol. 17, p. 1108889, 2023.

[12] ——, "Riemannian geometry metric-based visual feedback for BCI user training: Towards exploratory learning of motor imagery," *Brain-Apparatus Communication: A Journal of Bacomics*, vol. 3, no. 1, p. 2366193, 2024.

[13] W. Truccolo, L. R. Hochberg, and J. P. Donoghue, "Collective dynamics in human and monkey sensorimotor cortex: Predicting single neuron spikes," *Nature Neuroscience*, vol. 13, no. 1, pp. 105–111, 2010.

[14] M. Aghagolzadeh and W. Truccolo, "Inference and decoding of motor cortex low-dimensional dynamics via latent state-space models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 2, pp. 272–282, 2016.

[15] C. Pandarinath, K. C. Ames, A. A. Russo, A. Farshchian, L. E. Miller, E. L. Dyer, and J. C. Kao, "Latent factors and dynamics in motor cortex and their application to brain–machine interfaces," *Journal of Neuroscience*, vol. 38, no. 44, pp. 9390–9401, 2018.

[16] D. Xing, M. Aghagolzadeh, W. Truccolo, E. Bezard, G. Courtine, and D. Borton, "Low-dimensional motor cortex dynamics preserve kinematics information during unconstrained locomotion in nonhuman primates," *Frontiers in Neuroscience*, vol. 13, p. 1046, 2019.

[17] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz, "Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition," *Journal of Neuroscience Methods*, vol. 258, pp. 1–15, 2016.

[18] B. W. Brunton and M. Beyeler, "Data-driven models in human neuroscience and neuroengineering," *Current Opinion in Neurobiology*, vol. 58, pp. 21–29, 2019.

[19] P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, B. M. Yu, and A. P. Batista, "Neural constraints on learning," *Nature*, vol. 512, no. 7515, pp. 423–426, 2014.

[20] B. Feulner and C. Clopath, "Neural manifold under plasticity in a goal driven learning behaviour," *PLOS Computational Biology*, vol. 17, no. 2, p. e1008621, 2021.

[21] M. Frosolone, R. Prevete, L. Ognibeni, S. Giugliano, A. Apicella, G. Pezzulo, and F. Donnarumma, "Enhancing EEG-based MI-BCIs with class-specific and subject-specific features detected by neural manifold analysis," *Sensors*, vol. 24, no. 18, p. 6110, 2024.

[22] T. Liu, Z. Wang, S. Shakil, and R. K.-Y. Tong, "Uncovering low-dimensional manifolds of neural dynamics for motor-imagery based stroke rehabilitation: An EEG-based brain–computer interface study," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 3281–3292, 2025.

[23] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012.

[24] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain–computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.

[25] F. Yger, M. Bérar, and F. Lotte, "Riemannian approaches in brain–computer interfaces: A review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, 2017.

[26] Z. Shuqfa, A. N. Belkacem, and A. Lakas, "Decoding multi-class motor imagery and motor execution tasks using riemannian geometry algorithms on large EEG datasets," *Sensors*, vol. 23, no. 11, p. 5051, 2023.

[27] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[28] S. M. Hirsh, S. M. Ichinaga, S. L. Brunton, J. N. Kutz, and B. W. Brunton, "Structured time-delay models for dynamical systems with connections to the Frenet–Serret frame," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 477, no. 2254, p. 20210097, 2021.

[29] OpenBCI, "Setting up for EEG, EMG, and ECG at the same time," OpenBCI Documentation, 2026, acesso em: 29 jul. 2026. [Online]. Available: https://docs.openbci.com/GettingStarted/Biosensing-Setups/ExGSetup/

[30] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[31] W. Truccolo, "From point process observations to collective neural dynamics: Nonlinear Hawkes process GLMs, low-dimensional dynamics and coarse graining," *Journal of Physiology–Paris*, vol. 110, no. 4, pp. 336–347, 2016.

[32] R. V. Raut, Z. P. Rosenthal, X. Wang, H. Miao, Z. Zhang, J.-M. Lee, M. E. Raichle, A. Q. Bauer, S. L. Brunton, B. W. Brunton, and J. N. Kutz, "Arousal as a universal embedding for spatiotemporal brain dynamics," *Nature*, vol. 647, no. 8089, pp. 454–461, 2025.

Microdrive construction and implantation for chronic extracellular recordings targeting the dorsal hippocampus in freely moving rats

1st Isabella Figueiredo da Silva
Institute Santos Dumont
Macaíba, Brazil
isabella.silva@edu.isd.org.br

2st Bruno Ferreira da Silva
Santos Dumont Institute
Macaíba, Brazil
bruno.silva@edu.isd.org.br

2nd Ramon Hypolito Lima
Institute Santos Dumont
Macaíba, Brazil
ramon.lima@isd.org.br

***Abstract*—Chronic extracellular electrophysiological recordings in freely moving animals require implantable systems that integrate mechanical stability, compact design, adjustable electrode positioning, and compatibility with electrophysiological acquisition systems. In this work, we describe the fabrication and stereotaxic implantation of a shuttle-drive microdrive designed for chronic extracellular recordings in the dorsal hippocampus of adult Wistar rats. The device accommodates up to eight independently movable tetrodes connected to a 32-channel electrode interface board. The fabrication workflow comprises the assembly of the microdrive body and movable shuttles, preparation of the adjustment screws, construction of a silica guide-tube matrix, fabrication of nickel-chromium tetrodes, integration of the electrodes with the interface board, and stereotaxic implantation. The surgical procedure and postoperative management required for device implantation are also described. The resulting system provides a modular and adjustable platform for the implantation of multiple tetrodes in a compact configuration. By detailing the materials, fabrication procedures, component assembly, electrode organization, and implantation steps, this work provides a reproducible methodological framework for the construction and implementation of microdrive-based systems for chronic extracellular electrophysiological recordings in freely moving rats.**

***Index Terms*—microdrive, electrophysiology, hippocampus, Wistar rats.**

I. Introduction

Chronic extracellular electrophysiological recordings in freely moving animals require implantable systems that combine mechanical stability, compact dimensions, reliable electrical connections, and the possibility of maintaining or adjusting electrode position after surgery. The performance of these systems therefore depends not only on the recording electronics, but also on the mechanical design of the implant, the materials used in its construction, the organization of the electrodes, and the surgical procedures required for implantation [1], [2].

The dorsal hippocampus is a frequent target for chronic electrophysiological recordings because of its involvement in learning, memory, and spatial processing. Recordings from this structure require accurate stereotaxic placement and, in many experimental configurations, the ability to advance the electrodes after implantation. Movable tetrode systems are well suited to this purpose because the depth of individual recording elements can be adjusted independently while the animal remains implanted.

Several microdrive architectures have been developed to improve the accessibility and adaptability of chronic electrophysiological implants. These systems include lightweight and open-source designs, recoverable components, 3D-printed structures, modular head-mounted assemblies, and configurations capable of supporting independently movable electrodes [3]–[7]. Other designs have emphasized the integration between implant construction, surgical implantation, and multichannel electrophysiological acquisition [8]–[11]. These developments provide different solutions to common technical requirements, including electrode positioning, mechanical protection, component organization, and adaptation to different experimental configurations.

Among these approaches, the shuttle-drive architecture described by Voigts et al. provides independent mechanical control of multiple tetrodes through movable shuttles coupled to adjustment screws [12]. This arrangement allows individual electrodes to be advanced after implantation while maintaining multiple recording channels within a compact assembly. Reproducing such a system, however, requires coordination of several fabrication steps, including preparation of the drive body, assembly of the movement mechanism, construction of the guide-tube matrix, tetrode fabrication, connection to the electrode interface board, and stereotaxic implantation.

Detailed documentation of these procedures is important because changes in component dimensions, guide-tube organization, electrode routing, or implantation strategy can affect the final configuration and usability of the device. Methodological descriptions that integrate mechanical assembly, electrode preparation, surgical implantation, and electrophysiological acquisition can therefore facilitate reproduction and adaptation of microdrive systems under different laboratory conditions.

In the present study, we describe the construction and stereotaxic implantation of a shuttle-drive microdrive based on the architecture reported by Voigts et al. [12]. The device was configured to accommodate eight independently movable tetrodes connected to a 32-channel electrode interface board and was implanted targeting the dorsal hippocampus of an

adult Wistar rat. The workflow includes microdrive assembly, preparation of the adjustment mechanism, construction of the silica guide-tube matrix, fabrication and organization of nickel–chromium tetrodes, connection to the electrode interface board, and surgical implantation. The functionality of the implanted system was further assessed through multichannel extracellular signal acquisition and local field potential spectral characterization. The aim was to provide a technically detailed and reproducible description of the construction and implementation of this recording system while defining the electrophysiological validation within the scope of the data obtained in the present experiment.

II. Materials and Methods

A. Experimental Design

This study describes the fabrication and stereotaxic implantation of a shuttle-drive microdrive for chronic extracellular electrophysiological recordings in freely moving rats, followed by electrophysiological data acquisition and offline LFP analysis. The experimental procedures included microdrive fabrication and assembly, tetrode preparation, stereotaxic implantation, electrophysiological recording, electrode impedance assessment, and LFP signal processing.

One adult Wistar rat (*Rattus norvegicus*; $n = 1$) was used in the present study. The microdrive was designed to accommodate eight independently movable tetrodes, providing a total of 32 recording channels [12]. The main stages of device assembly and stereotaxic implantation are illustrated in Fig. 1.

B. Animals

One adult male Wistar rat (*Rattus norvegicus*), aged 5–7 months and weighing approximately 480 g, was used in the present study. The animal was housed in a high-ceiling cage under controlled environmental conditions, with temperature maintained between 22 and 26 °C and a 12-h light/dark cycle. Food and water were provided according to the experimental protocol.All procedures were performed by trained personnel and approved by the local Ethics Committee for Animal Use, following institutional and national guidelines for animal experimentation. High-ceiling cages were used after implantation to reduce the risk of mechanical contact between the implant and the cage lid. The animal were monitored daily during the postoperative period for body weight, grooming, locomotor activity, wound healing, and signs of discomfort.

C. Microdrive Assembly

We used a microdrive originally described by [12]. The device was designed to accommodate up to eight independently movable tetrodes, corresponding to 32 recording channels, in small rodents. Its fabrication followed the main steps described in the original protocol, including assembly of the microdrive body, construction of the silica matrix, tetrode fabrication, connection to the electrode interface board, and stereotaxic implantation. Initially, the microdrive body was positioned on an assembly support to ensure mechanical stability and safe handling. The shuttles, which are the movable components responsible for the independent displacement of each tetrode, were inserted into the guide channels of the device. Custom M0.6 screws were then coupled to the corresponding shuttles. The upper portion of each screw was secured with light-curable dental resin, forming a mechanism that allowed controlled and independent vertical displacement of each tetrode. The silica matrix was constructed using polyimide-coated fused-silica tubing (Molex Polymicro TSP100170). Twelve tube segments were arranged into a matrix approximately 4 mm in length and bonded together with epoxy resin. After the matrix had been assembled, eight silica guide tubes were inserted to establish the connection between the shuttles and the matrix. At the end of this stage, a correspondence map was created to document the position of each tube within the matrix, as well as its associated shuttle and screw. This mapping allowed each tetrode to be individually identified during post-implantation adjustments.

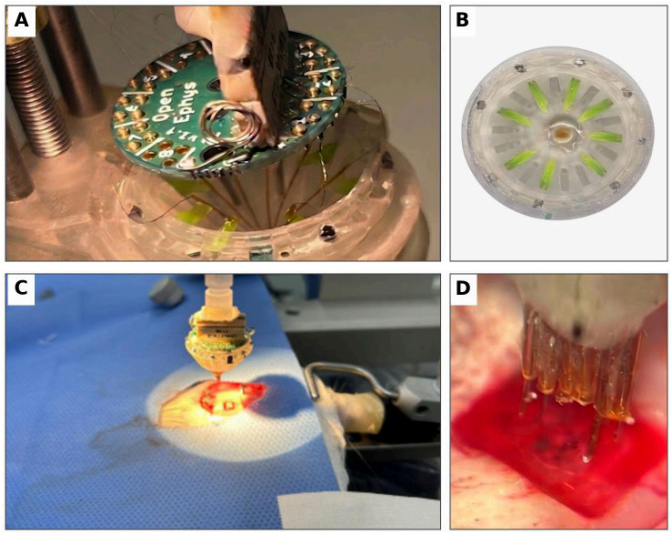


Fig. 1. Assembly and stereotaxic implantation of the shuttle-drive microdrive. (A) Connection of the tetrode wires to the 32-channel electrode interface board. (B) Top view of the assembled microdrive, showing the radial arrangement of the movable shuttles. (C) Stereotaxic positioning of the microdrive above the craniotomy. (D) Magnified view of the guide-tube matrix and tetrode tips aligned with the implantation site.

The tetrodes were fabricated using nickel-chromium alloy wire (Sandvik), with a diameter of approximately 0.012 mm, and coated with polyimide insulation. For each tetrode, four wires were aligned and placed in a high-speed tetrode twister operating at 200 revolutions per minute. After twisting, the tetrodes had a final length ranging from 3.5 to 4.0 cm. Each tetrode was inserted into its corresponding movable guide tube under stereomicroscopic visualization and subsequently secured with epoxy resin. The tetrode wires were connected to a custom 32-channel electrode interface board equipped with an Omnetics connector. The board was positioned on the upper portion of the microdrive and rotated between 90° and 180°, as needed, to accommodate the wires and reduce mechanical tension on the tetrodes. After appropriate positioning, the

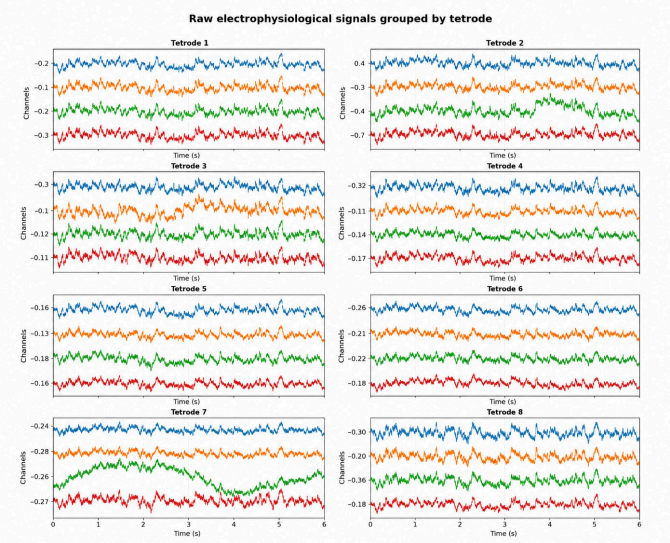


Fig. 2. Representative raw electrophysiological signals acquired from the 32 recording channels. Signals are grouped into eight tetrodes, with four simultaneously recorded channels per tetrode. A representative 6-s recording segment is displayed, and the traces were vertically offset within each panel to improve visualization.

electrode interface board was secured to the microdrive body with epoxy resin. During this stage, the wires were inspected to ensure that they remained properly organized and did not interfere with shuttle movement. For surgical implantation, the animals were anesthetized with isoflurane delivered in oxygen, using a concentration of 4% for induction and 1.5–2% for maintenance. Following the shaving of the cranial region and the antiseptic preparation of the surgical field, the animals were positioned in a stereotaxic apparatus. The skull was exposed and leveled using the bregma and lambda as anatomical reference points. A unilateral craniotomy was performed over the right dorsal hippocampus using the following stereotaxic coordinates relative to bregma: anteroposterior, -4.2 mm; mediolateral, +2.5 mm; and dorsoventral, -2.4 mm relative to the skull surface. During implantation, the tetrodes remained fully retracted in their uppermost position to minimize the risk of mechanical damage to the brain tissue. The microdrive was positioned over the craniotomy and secured to the skull using stainless-steel screws and methyl methacrylate-based dental cement. At the end of the implantation procedure, the tetrode tips were positioned approximately 1 mm above the cortical surface, allowing their gradual advancement toward the target region during the postoperative period. Following surgery, the animal received meloxicam at a dose of 1 mg/kg, administered subcutaneously, for postoperative analgesia. They were then individually housed in high-ceiling cages to prevent mechanical contact between the implant and the cage lid and to allow adequate postoperative recovery.

D. Electrode Impedance

Prior to implantation,electrode impedance was assessed for all 32 recording channels before electrophysiological analysis using the impedance-measurement function of the acquisition system. Impedance magnitude was recorded individually for each channel.

E. Electrophysiological Acquisition

Neural signals were acquired using an Open Ephys acquisition system connected to the implanted 32-channel electrode interface board. The recording configuration comprised eight tetrodes, with four channels assigned to each tetrode. Electrophysiological activity was continuously recorded during the experimental sessions and subsequently exported for offline signal processing.Signal quality was initially assessed through visual inspection of the raw electrophysiological traces.A channel was considered functional when a continuous electrophysiological signal without saturation or loss of acquisition was present during the analyzed interval. The power spectral density was then estimated for all recording channels to characterize the distribution of signal power across frequencies and to assess possible contamination by 60-Hz power-line noise and its 120-Hz harmonic. Time-frequency representations were also calculated for selected channels to examine changes in spectral power throughout the recording interval. Signal processing and visualization were performed in Python using NumPy, SciPy, and Matplotlib. When digital filtering was required, zero-phase filtering was applied to minimize phase distortion and preserve the temporal structure of the electrophysiological signals.

III. Results

A. Electrophysiological Recording

Electrophysiological signals were successfully acquired from all 32 recording channels connected to the implanted microdrive, corresponding to a functional-channel yield of 32/32 channels (100%) during the analyzed recording session. A channel was considered functional when a continuous electrophysiological signal without saturation or loss of acquisition was present during the analyzed interval. Fig. 2. Each tetrode comprised four simultaneously recorded channels. Continuous voltage fluctuations were observed across the analyzed 6-s segment, with channel-dependent differences in signal amplitude and waveform characteristics. The traces were vertically offset within each panel to facilitate comparison among the channels of the same tetrode.

B. Electrode Impedance

Impedance measurements were obtained from all 32 recording channels. Across channels, impedance magnitude ranged from 0.277 to 18.40 MΩ, with a median of 1.12 MΩ. Most channels showed impedance values below 3 MΩ; however, Channel 5 exhibited a markedly higher impedance of 18.40 MΩ. When this channel was considered separately as an extreme high-impedance channel, the remaining 31 channels showed a mean impedance of 1.29 ± 0.73 MΩ, ranging from 0.277 to 2.99 MΩ.

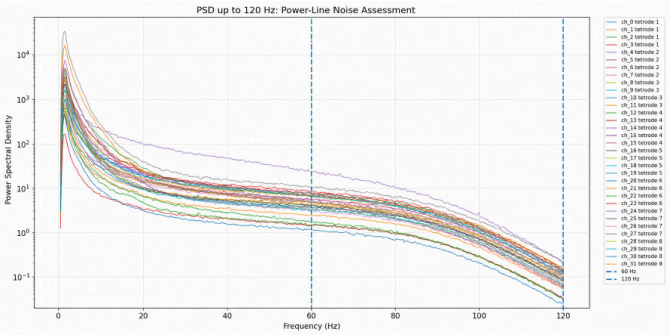


Fig. 3. Power spectral density estimates for the 32 extracellular electrophysiological recording channels from 0 to 120 Hz. Dashed vertical lines indicate the 60-Hz power-line frequency and its 120-Hz harmonic. Spectral power was predominantly concentrated at lower frequencies and decreased toward the higher-frequency range.

C. Power Spectral Density

The frequency-domain characteristics of the recordings were evaluated by estimating the power spectral density of all 32 channels. As shown in Fig. 3, spectral power was predominantly concentrated at lower frequencies and progressively decreased as frequency increased. Although differences in absolute power were observed among channels, the overall spectral profiles were comparable. Visual inspection did not reveal a dominant narrowband peak at the 60-Hz power-line frequency or at its 120-Hz harmonic, indicating that power-line interference did not dominate the analyzed recording segment.

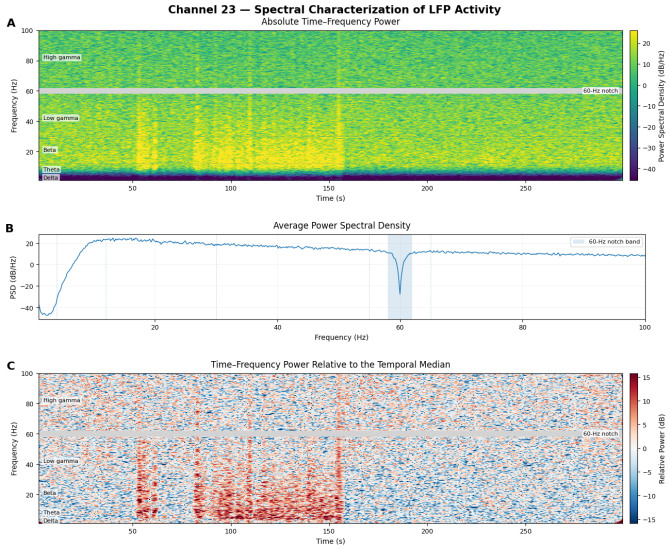


Fig. 4. Spectral characterization of local field potential activity from Channel 23, shown as an illustrative example of the time–frequency analysis. (A) Absolute time–frequency power over an approximately 300-s recording interval. (B) Average power spectral density from 0 to 100 Hz. (C) Time–frequency power expressed relative to the temporal median.

D. Representative LFP Spectral Characterization

A detailed spectral characterization was performed for Channel 23, selected as a representative recording channel. During the recording session, the animal was observed alternating between spontaneous locomotion and immobility. As shown in Fig. 4, the recorded activity remained continuous throughout the approximately 300-s interval and exhibited a structured distribution of spectral power. The absolute time–frequency representation showed greater power at low and intermediate frequencies, with a progressive reduction toward higher frequencies (Fig. 4A). The average power spectral density displayed a broadband profile and a marked attenuation centered at 60 Hz, resulting from the notch filter applied to suppress power-line interference (Fig. 4B). After normalization relative to the temporal median, transient increases in spectral power became more evident, particularly below approximately 30–40 Hz and between approximately 50 and 160 s (Fig. 4C). These variations indicate that the spectral composition of the signal was not temporally uniform and may reflect changes in the functional state of the recorded neural network. These results demonstrate that the spectral composition of the recorded signal varied over the analyzed interval. Because behavioral events were not temporally synchronized with the electrophysiological recording, the physiological origin of these variations was not determined.

IV. Discussion

The present study describes the fabrication and implantation of a shuttle-drive microdrive configured for eight independently movable tetrodes and 32 recording channels. The use of independently movable tetrodes is an important feature of the device because electrode depth can be adjusted after implantation. This configuration is advantageous when gradual electrode advancement is required and is consistent with the mechanical principles of the shuttle-drive architecture described by Voigts et al. [12]. The present implementation further demonstrates that this architecture can be assembled using the materials and procedures described here and coupled to a 32-channel Open Ephys acquisition system. The electrophysiological validation presented in this study should,however, be interpreted within its methodological scope. Recordings were obtained from a single animal ($n = 1$), and the 32 channels therefore represent recording sites within the same experimental subject rather than independent biological replicates. In addition, the detailed time–frequency characterization was performed on one representative channel. Consequently, the present results provide a technical proof-of-principle rather than an assessment of between-animal reproducibility. A further limitation is the absence of histological verification of the final tetrode positions. Therefore, anatomical localization in the present study is based on the stereotaxic implantation coordinates and electrode advancement procedure rather than postmortem confirmation of individual recording sites.

V. Conclusion

This article describes the fabrication and surgical implantation of a shuttle-drive microdrive for chronic extracellular electrophysiological recordings in freely moving rats. The device enables independent adjustment of multiple tetrodes

and is suitable for targeting the dorsal hippocampus during behavioral and resting-state recordings. This methodological framework may support reproducible studies of hippocampal oscillations, neuronal activity, and spatial memory in rodents.

REFERENCES

[1] D. Yi, Y. Yao, Y. Wang, and L. Chen, "Manufacturing processes of implantable microelectrode array for in vivo neural electrophysiological recordings and stimulation: A state-of-the-art review," *Journal of Micro- and Nano-manufacturing*, vol. 10, no. 4, p. 041001, 2022.

[2] ——, "Design, fabrication, and implantation of invasive microelectrode arrays as in vivo brain machine interfaces: A comprehensive review," *Journal of manufacturing processes*, vol. 126, pp. 185–207, 2024.

[3] M. Vöröslakos, H. Miyawaki, S. Royer, K. Diba, E. Yoon, P. C. Petersen, and G. Buzsáki, "3d-printed recoverable microdrive and base plate system for rodent electrophysiology," *Bio-protocol*, vol. 11, no. 16, pp. e4137–e4137, 2021.

[4] M. Vöröslakos, P. C. Petersen, B. Vöröslakos, and G. Buzsáki, "Metal microdrive and head cap system for silicon probe recovery in freely moving rodent," *Elife*, vol. 10, p. e65859, 2021.

[5] M. Brosch, A. Vlasenko, F. W. Ohl, and M. T. Lippert, "Tetrodrive: an open-source microdrive for combined electrophysiology and opto-physiology," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046030, 2021.

[6] C. Sun, Y. Cao, J. Huang, K. Huang, Y. Lu, and C. Zhong, "Low-cost and easy-fabrication lightweight drivable electrode array for multiple-regions electrophysiological recording in free-moving mice," *Journal of neural engineering*, vol. 19, no. 1, p. 016003, 2022.

[7] M. Guardamagna, R. Eichler, R. Pedrosa, A. Aarts, A. F. Meyer, and F. P. Battaglia, "The hybrid drive: a chronic implant device combining tetrode arrays with silicon probes for layer-resolved ensemble electrophysiology in freely moving mice," *Journal of Neural Engineering*, vol. 19, no. 3, p. 036030, 2022.

[8] R. J. Pendry, L. D. Quigley, L. J. Volk, and B. E. Pfeiffer, "A lightweight drive implant for chronic tetrode recordings in juvenile mice," *Journal of visualized experiments: JoVE*, no. 196, pp. 10–3791, 2023.

[9] E. Ferreira-Fernandes, M. Laranjo, T. Reis, B. Canijo, P. A. Ferreira, P. Martins, J. Vilarinho, M. Tavakoli, C. Kunicki, and J. Peça, "In vivo recordings in freely behaving mice using independent silicon probes targeting multiple brain regions," *Frontiers in Neural Circuits*, vol. 17, p. 1293620, 2023.

[10] J. P. Hartner, D. Yi, H. L. Zhu, B. O. Watson, and L. Chen, "Three-dimensional-printed headcap with embedded microdrive system for customizable multi-region brain recordings with neural probes," *Frontiers in Neuroscience*, vol. 18, p. 1478421, 2024.

[11] R. J. I. Alcalá, A. Y. Macias, C. N. Heaton, R. S. Jurado, A. A. Salcido, N. F. Reyes, S. A. Batson, L. D. Davila, D. W. Beck, L. I. Rakocevic *et al.*, "A modular, adaptable, and accessible implant kit for chronic electrophysiological recordings in rats," *Cell Reports Methods*, vol. 5, no. 9, 2025.

[12] J. Voigts, J. P. Newman, M. A. Wilson, and M. T. Harnett, "An easy-to-assemble, robust, and lightweight drive implant for chronic tetrode recordings in freely moving animals," *Journal of neural engineering*, vol. 17, no. 2, p. 026044, 2020.

Método de Preparação de Fatias Hipocampais para Avaliação Integrada de Eletrofisiologia e Imunofluorescência

Fabricio Neves Mendonça
Programa de Pós-graduação em Neurologia e Neurociências
Universidade Federal de São Paulo/Escola Paulista de Medicina
São Paulo, Brasil
fabricio.mendonca@ifmg.edu.br

Andre Luis Fonseca Furtado
Programa de Pós-graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
andre.furtado@ifsudestemg.edu.br

Antônio-Carlos Guimarães de Almeida
Laboratório de Neurociência Experimental e Computacional
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
acga@ufsj.edu.br

Antônio Márcio Rodrigues
Laboratório de Neurociência Experimental e Computacional
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
amr@ufsj.edu.br

Carla Alessandra Scorza
Departamento de Neurologia e Neurocirurgia
Universidade Federal de São Paulo/Escola Paulista de Medicina
São Paulo, Brasil
carla.scorza@unifesp.br

Fulvio Alexandre Scorza
Departamento de Neurologia e Neurocirurgia
Universidade Federal de São Paulo/Escola Paulista de Medicina
São Paulo, Brasil
scorza.nexp@epm.br

Heloisa Beatriz Silva
Programa de Pós-graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
heloisabeatrizsilva@aluno.ufsj.edu.br

Jessica Naiara Lara
Programa de Pós-graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
jessicanaiaralara.jnl@aluno.ufsj.edu.br

Mayra Consuelo Aarão
Programa de Pós-graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, Brasil
mayra@ufsj.edu.br

***Resumo*— Modelos animais, em especial os modelos *in vitro* de fatias de cérebro e hipocampo, são fundamentais para a pesquisa em neurociências, pois permitem elucidar mecanismos neuronais e suas bases patológicas em condições controladas. Contudo, é essencial desenvolver estratégias que reduzam o número de animais utilizados, em conformidade com os princípios dos 3Rs (*Replacement, Reduction, Refinement* - Substituição, Redução, Refinamento). Neste estudo, apresentamos um método que combina a perfusão transcardíaca utilizando fluido cerebrospinal artificial protetor modificado contendo sacarose (SP-mACSF) à ressecção de fatias de hipocampo. O sucesso dessa abordagem reside na capacidade do SP-mACSF em promover uma lavagem vascular eficiente, permitindo tanto a viabilidade tecidual para a realização de gravações eletrofisiológicas quanto a obtenção de imagens de imunofluorescência de alta qualidade para análises de histologia, a partir do mesmo tecido. Para o desenvolvimento e validação desse método, foram utilizadas fatias de 400 µm de espessura obtidas de ratos machos da linhagem *Wistar*. As fatias provenientes de animais perfundidos com SP-mACSF exibiram atividades epileptiformes não-sinápticas no giro dentado indistinguíveis das dos controles em todos os aspectos avaliados. As ressecções de 40 µm produzidas a partir das fatias de 400 µm, após fixação em paraformaldeído proporcionaram marcações anti-GFAP bem definidas, isentas de artefatos e com baixa autofluorescência no grupo SP-mACSF quando comparadas às do grupo Controle. Este método inovador representa um avanço significativo para a neurociência, consolidando a aplicação dos princípios dos 3Rs ao permitir a correlação direta entre dados funcionais e estruturais, com a extração de dados multifatoriais de um único animal, promovendo uma prática científica mais sustentável, ética e metodologicamente integrada.**

Palavras-chave — fatias de hipocampo, atividade epileptiforme, imunofluorescência, autofluorescência, 3Rs

I. INTRODUÇÃO

A epilepsia é uma condição neurológica crônica e complexa, caracterizada por descargas neuronais excessivas e hiperssincronizadas recorrentes, conhecidas como crises epilépticas espontâneas. Estima-se que aproximadamente 1 em cada 10 indivíduos apresentará ao menos uma crise epiléptica ao longo da vida, enquanto cerca de 1 em cada 6 desenvolverá epilepsia [1]. Com aproximadamente 50 milhões de pessoas afetadas no mundo, a epilepsia destaca-se como uma das doenças neurológicas mais prevalentes no planeta [2]. Apesar dos avanços no desenvolvimento de fármacos anticrise, aproximadamente um terço dos pacientes apresenta epilepsia refratária ao tratamento medicamentoso [3], evidenciando a necessidade de uma compreensão mais profunda dos mecanismos fisiopatológicos subjacentes à doença. Essa eficácia limitada da terapia medicamentosa [4] sugere que a origem dessas descargas ultrapassa as conexões puramente sinápticas, envolvendo também o fluido iônico, a matriz extracelular e as comunicações não-sinápticas da rede neuronal. A relevância da modulação não-sináptica na epilepsia foi confirmada experimentalmente nos anos 1980, quando estudos pioneiros demonstraram a viabilidade de induzir atividades epileptiformes *in vitro*, em fatias de hipocampo, apenas pela alteração das concentrações extracelulares de cálcio e potássio [5, 6]. Tais interações não-sinápticas exercem profundos efeitos fisiológicos no tecido neuronal, sendo categorizadas em quatro mecanismos principais [7]: o acoplamento eletrotônico, o efeito de campo elétrico, o acoplamento efáptico e as flutuações iônicas extracelulares.

Diante da elevada prevalência global e do elevado índice de pacientes refratários aos tratamentos farmacológicos

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

atuais, a compreensão dos mecanismos fisiopatológicos da epilepsia torna-se imperativa para o desenvolvimento de novas terapias. As preparações de fatias de cérebro e, particularmente, de hipocampo preservam aspectos essenciais da organização neuronal e da circuitaria sináptica, tornando-se ferramentas amplamente empregadas para investigar mecanismos celulares e de rede em condições experimentais rigorosamente controladas [8, 9]. Em fatias de hipocampo, modelos *in vitro* de atividade epileptiforme permitem investigar, com elevada resolução temporal e espacial, os mecanismos envolvidos na geração, propagação e modulação das descargas neuronais, constituindo uma plataforma consolidada para estudos fisiopatológicos e farmacológicos [10]. Contudo, as implicações éticas constituem um fator significativo de complexidade no emprego de tecido humano vivo para estudos de natureza eletrofisiológica [11].

Em decorrência disso, ao longo dos anos, o uso de fatias de cérebro e de hipocampo obtidas a partir de modelos animais tornou-se uma abordagem bem estabelecida na pesquisa em neurociências [9, 12, 13, 14]. Nesse contexto, modelos animais *in vitro* são amplamente utilizados em investigações experimentais de atividades epileptiformes, reforçando sua relevância na pesquisa translacional em epilepsia [15, 16, 17]. Na pesquisa em epilepsia, essa preparação é particularmente valiosa, pois permite a investigação detalhada dos mecanismos celulares e de rede subjacentes à geração, propagação e modulação da atividade epileptiforme sob condições mais acessíveis à manipulação eletrofisiológica e farmacológica do que em modelos *in vivo* [18].

Paralelamente aos estudos eletrofisiológicos com fatias de hipocampo, a análise histológica e a marcação por imunofluorescência de roedores são ferramentas fundamentais na neurociência, pois permitem a identificação de alterações celulares, de astrogliose reativa e de remodelação sináptica dentro de redes epileptogênicas [19]. Contudo, para realizar adequadamente essas avaliações microscópicas, é mandatório fixar quimicamente o tecido cerebral, geralmente com paraformaldeído. Esse processo de fixação promove a reticulação de proteínas e interrompe a degradação tecidual, preservando a citoarquitetura das fatias e garantindo a retenção espacial precisa dos marcadores moleculares necessários para investigar a fisiopatologia das crises [20].

A integração entre abordagens eletrofisiológicas, histológicas e de imunofluorescência constitui um dos pilares da neurociência experimental contemporânea, permitindo correlacionar alterações funcionais e estruturais envolvidas em processos fisiológicos e patológicos do sistema nervoso. No entanto, apesar da relevância complementar das gravações eletrofisiológicas e das análises histológicas e de imunofluorescência em neurociência, essas abordagens geralmente não são aplicadas ao preparado tecidual do mesmo animal. A presença de sangue residual no tecido constitui uma importante fonte de autofluorescência [21, 22]. Essa autofluorescência constitui um fator de interferência conhecido na marcação por imunofluorescência, podendo comprometer a especificidade e a qualidade das imagens obtidas por microscopia confocal [21, 22]. Por esse motivo, a análise de histologia com imunofluorescência utiliza, rotineiramente, a perfusão transcardíaca com PBS, seguido de paraformaldeído, para fixação química com vistas a preservar a arquitetura tecidual por meio da estabilização estrutural [13, 18, 20]. Mas, uma vez que o tecido nervoso tenha sido fixado com paraformaldeído, a atividade neuronal se torna ausente, uma vez que o tecido e as células neuronais não são mais viáveis, o que os torna inadequados para investigações eletrofisiológicas da excitabilidade neuronal ou da atividade epileptiforme, pois gravações eletrofisiológicas requerem tecido nervoso vivo com células neuronais viáveis [8, 23].

Assim, embora registros eletrofisiológicos e análises histológicas sejam complementares, essas abordagens permanecem, na maioria dos protocolos experimentais, habitualmente dissociadas. Enquanto a eletrofisiologia depende da manutenção da viabilidade neuronal em tecido vivo, a imunofluorescência requer fixação química para preservar a arquitetura tecidual e os antígenos de interesse. Isso torna as duas abordagens, em grande parte dos desenhos experimentais, mutuamente exclusivas para o mesmo animal [13, 18, 20]. Como consequência, avaliações funcionais e estruturais costumam ser realizadas em animais distintos, impedindo a correlação direta entre os achados experimentais, aumentando a variabilidade biológica e ampliando o número de animais necessários para responder à mesma hipótese científica.

Diante desse cenário, o presente estudo propôs desenvolver e validar uma estratégia metodológica baseada na perfusão transcardíaca com fluido cerebrospinal artificial modificado contendo sacarose (SP-mACSF), combinada à preparação de fatias de hipocampo. O método foi concebido para permitir a realização de registros eletrofisiológicos e análises por imunofluorescência utilizando tecido proveniente do mesmo animal, preservando a funcionalidade neuronal e reduzindo a autofluorescência decorrente de resíduos sanguíneos. Dessa forma, a nossa abordagem buscou ampliar a integração entre análises estruturais e funcionais, ao mesmo tempo em que otimiza o aproveitamento experimental de cada animal em conformidade com os princípios dos 3Rs. O método desenvolvido e testado neste estudo foi adaptado como uma combinação modificada que integra duas abordagens previamente descritas. A primeira foi descrita por [24], que empregou uma solução à base de sacarose para lavar o sangue do tecido antes da dissecção cerebral, preservando a viabilidade tecidual e minimizando artefatos para visualização em um microscópio confocal de disco giratório BX51WI *Olympus*, destinado a experimentos simultâneos de fluorescência e eletrofisiologia. A segunda foi descrita por [25], na qual fatias utilizadas para registrar atividade epileptiforme não sináptica em câmara de interface foram pós-fixadas em paraformaldeído, embutidas em gel de ágar e resseccionadas em vibrátomo para coloração por meio do método de *Nissl*. A Figura 1 mostra um resumo gráfico esquemático do método desenvolvido e validado.

II. MÉTODOS

A. Animais

Neste trabalho, foram utilizados 10 ratos *Wistar* machos com 4 semanas de idade, com um tamanho amostral de 5 animais por grupo. O estudo foi conduzido de acordo com os procedimentos aprovados pelo Comitê de Ética no Uso de Animais da Universidade Federal de São João del Rei (CEUA/UFSJ), sob o protocolo 8518220426/2026, em conformidade com a legislação brasileira vigente e com as diretrizes do CONCEA para experimentação animal.

B. Perfusão transcardíaca e preparação de fatias de hipocampo

Para a coleta de tecido, os animais do grupo SP-mACSF foram anestesiados por meio de injeção intraperitoneal de ketamina (100 mg/kg) e xilazina (10 mg/kg). Após a indução anestésica, foi realizada uma incisão horizontal mediana abdominal, estendendo-se da região genital e subindo verticalmente de ambos os lados, a fim de acessar a cavidade torácica e o coração. A seguir, o diafragma foi seccionado, as costelas de ambos os lados foram bissectadas até o nível axilar e o esterno foi elevado e preso nessa posição por meio de uma pinça hemostática para expor o coração. Uma vez realizada a exposição cardíaca, uma agulha hipodérmica (0,6 x 25 mm) conectada por meio de um tubo extensor a uma bomba de perfusão, foi inserida no ventrículo esquerdo (Figura 1). Em seguida, os animais foram perfundidos com 100 ml de SP-mACSF para depurar o sistema circulatório e garantir a remoção dos resíduos de sangue tecidual. Previamente ao início da perfusão, realizou-se uma incisão na artéria renal esquerda para permitir a drenagem sanguínea, sendo a bomba de perfusão conFigurada para uma taxa de fluxo de um litro por hora. Utilizou-se um volume total de 200 ml de SP-mACSF por animal, dos quais 100 ml foram destinados à perfusão transcardíaca propriamente dita, 50 ml para o preenchimento do sistema de alimentação da bomba de perfusão e 50 ml para o armazenamento e transporte do encéfalo, bem como para a posterior dissecação e seccionamento dos hipocampos.

A solução de perfusão SP-mACSF consistiu em um fluido cerebrospinal artificial com baixas concentrações de sódio, potássio e cálcio, combinadas com concentrações elevadas de magnésio e glicose, modificado a partir da formulação empregada por [24]. Assim como no protocolo original descrito por [24], o cloreto de sódio foi integralmente substituído por uma concentração equimolar de sacarose. Essa substituição tem como objetivo inibir a atividade neuronal e, consequentemente, limitar a excitotoxicidade, sem, contudo, alterar a osmolaridade do mACSF. A elevação da concentração de magnésio foi empregada para reduzir a ativação glutamatérgica de receptores NMDA, amplificando as propriedades neuroprotetoras do SP-mACSF. As concentrações reduzidas de potássio e cálcio foram utilizadas com o objetivo de diminuir a excitabilidade neuronal e a liberação de neurotransmissores, respectivamente, reduzindo os danos excitotóxicos durante a preparação do tecido. Adicionalmente, com o intuito de atenuar lesões induzidas por hipóxia e otimizar o fluido cerebrospinal artificial (ACSF) proposto por [24], a concentração de glicose no SP-mACSF foi ajustada para 20 mM. Essa concentração, superior aos 10 mM empregados por [24], fundamenta-se em evidências de que valores mais elevados de glicose atenuam os danos isquêmicos associados a condições hipóxicas [26]. Assim, a composição do SP-mACSF utilizado para perfusão transcardíaca neste estudo foi (em mM): 210 sacarose, 26 $NaHCO_3$, 2.5 KCl, 1 $CaCl_2$, 4 $MgCl_2$, 1.25 NaH_2PO_4 e 20 glicose. A solução SP-mACSF foi preparada em volume estoque de 2 litros, borbulhada com mistura carbogênica (95% O_2, 5% CO_2) por 30 minutos e congelada em tubos *falcon* de 50 ml para armazenamento. Para a perfusão transcardíaca, o SP-mACSF foi descongelado e mantido em 2 a 4 °C até o momento de início da perfusão. Após o descongelamento, o SP-mACSF destinado ao armazenamento do encéfalo e à dissecção e fatiamento dos hipocampos foi resfriado novamente até atingir uma consistência semissólida.

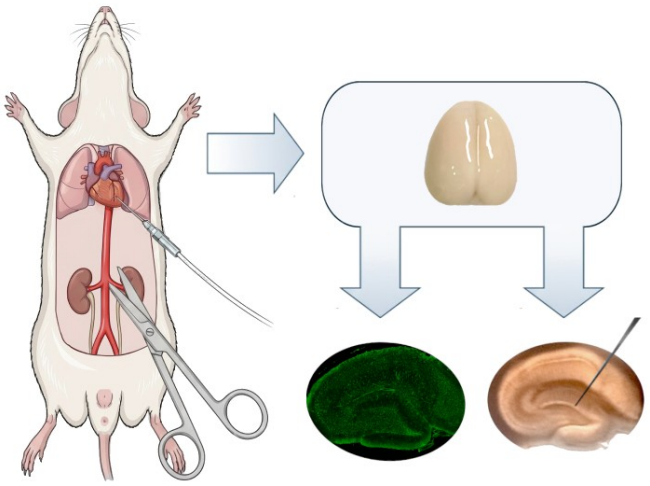


Figura 1. Resumo gráfico do método descrito no artigo. Detalhe metodológico da técnica de perfusão modificada evidenciando a secção da artéria renal esquerda e o uso de agulha hipodérmica inserida no ventrículo esquerdo. Esse método de perfusão, desenvolvido originalmente em nosso laboratório para utilização em gerbils da Mongólia, é uma alternativa à tradicional canulação da aorta com incisão no átrio direito.

Concluída a perfusão transcardíaca, procedeu-se a decapitação com guilhotina de aço inoxidável para roedores (*INSIGHT*) para facilitar o manuseio. O crânio foi então aberto e o encéfalo cuidadosamente removido. Após a dissecção, o encéfalo foi armazenado em SP-mACSF semicongelado por cerca de um minuto. Em seguida, os hipocampos foram isolados e fatiados em 400 µm utilizando um fatiador do tipo *chopper*. Para isso, após separar os hemisférios cerebrais com bisturi, enquanto um hemisfério foi dissecado para separar o primeiro hipocampo para fatiamento, o segundo hemisfério permaneceu submerso em SP-mACSF resfriado a 2 a 4 °C, aguardando a conclusão do processo de fatiamento do primeiro hipocampo. Uma vez concluído o primeiro fatiamento procedeu-se à dissecção e fatiamento do hipocampo do hemisfério contralateral. Cada hipocampo produziu de 10 a 12 fatias, sendo metade da quantidade de seções (5 a 6 fatias por hipocampo) fixada em paraformaldeído e a outra metade reservada para registros eletrofisiológicos.

No grupo Controle, os animais foram decapitados sem anestesia e tiveram o encéfalo removido e armazenado em ACSF semicongelado. A seguir, a dissecção e o fatiamento dos hipocampos foram realizados com ACSF resfriado a 2 a 4 °C. Além disso, o período de recuperação na câmara de perfusão das fatias do grupo Controle, foi de 60 minutos. Todos os procedimentos subsequentes foram idênticos ao grupo SP-mACSF.

C. Registro dos Potenciais de Campo durante Atividade Epileptiforme no Giro Dentado

Previamente às gravações eletrofisiológicas, as fatias foram mantidas submersas a 30 °C em uma câmara de perfusão preenchida com fluido cerebrospinal artificial normal oxigenado (*Ringer* ACSF) por 120 a 150 minutos, permitindo a recuperação de traumas decorrentes do preparo e a eliminação dos agentes anestésicos. Após esse período inicial de recuperação, as fatias foram transferidas para uma câmara de interface e mantidas em repouso por 60 minutos adicionais antes da exposição à solução de modificada de ACSF (mACSF) indutora de atividade epileptiforme. A composição do *Ringer*-ACSF normal foi (em mM): 127 NaCl,

2 KCl, 1.5 $MgSO_4$, 1.1 KH_2PO_4, 26 $NaHCO_3$, 2 $CaCl_2$ e 10 glicose [27]. A composição do mACSF utilizado para induzir atividade epileptiforme no giro dentado foi (em mM): 127 NaCl, 7 KCl, 1.5 $MgSO_4$, 1.1 KH_2PO_4, 26 $NaHCO_3$ e 10 glicose (sem cálcio adicionado) [27]. Todas as soluções foram continuamente borbulhadas com carbogênio (95% O_2 5% CO_2), mantendo o pH em 7.4. A temperatura da câmara de interface foi mantida constante em 32 °C durante todo o período de repouso das fatias e também durante a indução das atividades epileptiformes.

Na câmara de interface, as fatias foram mantidas sobre uma membrana microporosa, submetida a uma perfusão contínua de solução *Ringer*-ACSF/mACSF por baixo, aquecida em banho-maria a temperatura de 32 °C. A mistura de carbogênio (95% O_2, 5% CO_2), após ser borbulhada na água destilada do banho-maria, portanto sendo umidificada e aquecida, foi continuamente liberada na parte superior da membrana, formando a interface gás/líquido da câmara, onde as fatias foram mantidas. Durante a indução de atividades epileptiformes, as fatias foram visualizadas utilizando um estereomicroscópio *Nikon* SMZ1500.

Os registros dos potenciais de campo foram realizados utilizando eletrodos de prata cloretados e inseridos em micropipetas de vidro preenchidas com a solução mACSF indutora de atividades epileptiformes. As micropipetas foram fabricadas a partir de capilares de vidro com paredes finas, estiradas em um puxador universal. Os eletrodos foram conectados a um *headstage* de baixo ruído, ligado a um amplificador conFigurado para ganho total de 500x e filtro passa-baixa com corte de 1600 Hz. Os sinais amplificados foram digitalizados em frequência de amostragem de 5 kHz via plataforma desenvolvida em *LabVIEW*. Os eletrodos de registro foram posicionados na camada granular do giro dentado, precisamente na região de transição entre o ápice e a lâmina supra piramidal (Figura 2. D). As gravações eletrofisiológicas foram obtidas de duas fatias por animal, totalizando 20 gravações (10 para cada grupo), e as marcações de imunofluorescência foram realizadas utilizando o mesmo número de fatias dos mesmos animais. As fatias foram escolhidas aleatoriamente, tanto para o registro eletrofisiológico quanto para a histologia.

D. Obtenção dos valores de parâmetros do Potencial de Campo (FP)

Os parâmetros duração do evento (DE), eventos por minuto (Ev/min) e amplitude média de *population spikes* (Amp Med) foram obtidos, para cada um dos eventos epileptiformes analisados, utilizando-se um *script* desenvolvido em plataforma MATLAB®. A Transformada Discreta de Fourier (DFT) foi utilizada para decompor o sinal nos componentes da corrente contínua (DC) e o sinal contendo os PS. A partir do cálculo da DFT do sinal registrado, foi realizado o corte da frequência baixa abaixo de 10 Hz e que corresponde à componente DC separando-a da frequência alta, que corresponde aos *population spikes* (PS). A seguir, através da Transformada Inversa de Fourier, foi obtido o sinal contendo apenas a componente DC (sem os PS). Uma vez obtida a componente DC, a linha de base foi determinada como a média de um trecho do potencial elétrico antes do evento. Assim, a amplitude da componente DC foi obtida como sendo a variação máxima do sinal em relação à linha de base. O parâmetro DE foi obtido por meio da fórmula: DE(s) = Tfinal (s) - Tinicial (s). O instante de tempo "Tinicial" foi considerado como quando a componente DC do sinal atingiu

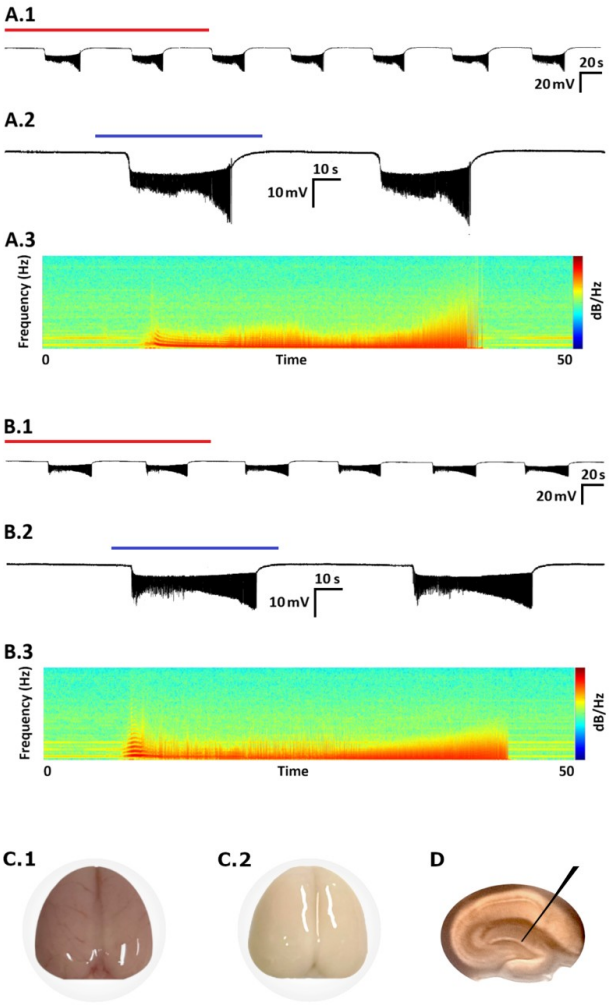


Figura 2. (A, B) Morfologias das atividades epileptiformes registradas na camada granular do giro dentado dos grupos Controle e SP-mACSF. (A1) Traços brutos de eventos epileptiformes: a linha vermelha indica o segmento do traço expandido em A2. (A2) Detalhe temporal do evento: a linha azul indica o período de registro utilizado para calcular o espectro de potência apresentado em A3. (A3) Espectro de potência do segmento destacado, com escala de frequência (*Frequency)* de 0 a 1600 Hz e escala de potência de –30 a 30 dB/Hz. (B) Gravações correspondentes de uma fatia obtida de um animal perfundido com SP-mACSF (B1–B3), seguindo a mesma sequência e parâmetros de escala de A1-A3). Figura gerada em MATLAB®. (C) Encéfalos obtidos de animais do grupo controle (C.1) e SP-mACSF (C.2). (D) Local de posicionamento do eletrodo para registro das

uma variação de 20% da amplitude máxima em relação à linha de base e o instante de tempo "Tfinal", quando esta retornou a 20% da amplitude máxima em relação à linha de base. O parâmetro Ev/min foi obtido por meio da fórmula: Ev/min = total de eventos registrados/tempo total de registro (min). A Amp Med corresponde ao componente de alta frequência dos eventos epileptiformes e também foi separada do sinal do PE utilizando-se o cálculo da DFT do sinal elétrico, eliminando frequências inferiores a 10 Hz (componente DC), restando um sinal cujas componentes em alta frequência (acima de 10 Hz) correspondem aos PS. O valor do parâmetro Amp Med foi obtido pela valor médio das amplitudes dessas variações de alta frequência. A Figura 2 ilustra exemplos típicos de eventos epileptiformes registrados nos dois grupos, bem como o aspecto dos encéfalos dos grupos Controle e SP-mACSF e o local de registro das atividades epileptiformes não-sinápticas na camada granular do giro dentado.

E. Histologia

Após o corte em fatiador do tipo *chopper*, as fatias de tecido provenientes de ambos os grupos experimentais destinadas à análise de histologia foram coletadas e armazenadas temporariamente em solução ACSF (Controle) ou SP-mACSF, resfriada (2 a 4 °C) pelo tempo necessário para realizar o fatiamento dos hipocampos. Imediatamente após a conclusão do fatiamento, as fatias foram coletadas da solução ACSF (Controle) ou SP-mACSF resfriada (utilizando-se um conta gotas plástico descartável com diâmetro de 5 mm) e fixadas por imersão em solução de paraformaldeído a 2% resfriada (2 a 4 °C), onde foram mantidas (resfriadas em 2 a 4 °C) por 5 a 7 dias para pós-fixação. Em seguida, as fatias fixadas pelo paraformaldeído foram embutidas em gelatina ágar e resseccionadas em 6 a 8 seções seriadas, de 40 µm de espessura, utilizando vibrátomo *Leica* e armazenadas em solução crioprotetora anticongelante (solução *antifreeze*) no congelador. A solução crioprotetora anticongelante foi composta por 300 g de sacarose, 500 ml de tampão fosfato (PBS) e 300 ml de monoetileno glicol, ajustada ao volume final de 1 litro com PBS. Para a realização das marcações de imunofluorescência, as secções permaneceram *overnight* a 4 °C em uma solução de soro albumina bovina (BSA), contendo 0,1% de Triton X-100, 10% de BSA em PBS 0,1 M. Em seguida, as secções foram incubadas por 48 horas, a 4 °C, em solução (2% de BSA em PBS 0,1M e 0,1% Triton X-100) contendo o anticorpo primário anti-GFAP (glial fibrillary acidic protein - proteína do citoesqueleto de astrócitos – *Abcam*, produzido em coelho, na concentração 1:1000). Após o processo com os anticorpos primários, as secções foram incubadas por duas horas, em temperatura ambiente, em solução com o anticorpo secundário anti-IgG de coelho (anticorpo policlonal, produzido em cabra, conjugado com *DyLight*® 488; 1:250, Abcam).

Eritrócitos exibem autofluorescência endógena em múltiplos comprimentos de onda, que promove sobreposição direta com os espectros de excitação e emissão de fluoróforos comumente empregados, interferindo no sinal específico da marcação [31, 32, 34]. Nesse contexto, a imunomarcação fluorescente anti-GFAP foi utilizada neste estudo tanto para a evidenciação da autofluorescência de fundo quanto por sua relevância anatômica, uma vez que os astrócitos, mesmo estando presentes em toda a fatia hipocampal, se distribuem de forma peculiar ao redor dos vasos sanguíneos no hipocampo. Essa característica anatômica dos astrócitos torna particularmente evidentes, na marcação para GFAP, os artefatos decorrentes de sangue residual no interior dos capilares. A marcação por imunofluorescência foi realizada segundo protocolos padrão descritos em estudos anteriores, com aquisição de imagens em microscopia confocal utilizando um sistema confocal *Zeiss* LSM710 acoplado a um microscópio invertido *Observer* Z1, conforme descrito previamente por [28, 29].

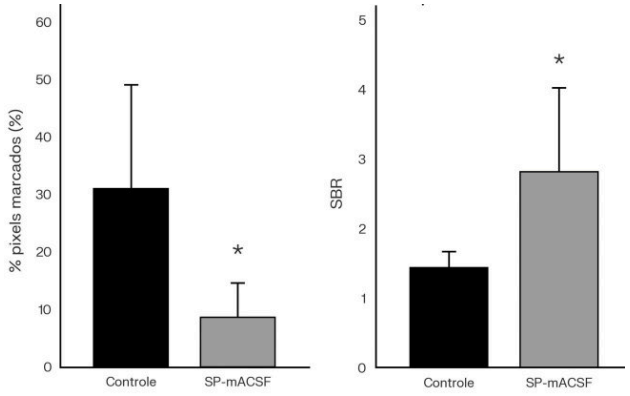


Figura 3. Marcação imunofluorescente anti-GFAP utilizado limiar fixo. Valor de % de pixels marcados significativamente menor no grupo SP-mACSF quando comparado ao controle, em decorrência de identificação de marcação inespecífica por autofluorescência e elevado *background* no grupo controle. Valor de SBR significativamente maior no grupo SP-mACSF.

Para a quantificação do % de pixels marcados, as micrografias confocais foram binarizadas aplicando-se um limiar fixo de detecção com intensidade de brilho igual a 75 (em uma escala de 0 a 255), calculando-se a proporção de pixels com intensidade acima desse corte em relação à área total da micrografia. A intensidade corrigida (BCI) foi calculada pela diferença entre a intensidade média da região de interesse de sinal (IS) e a intensidade média da região de fundo (IB): BCI = IS - IB. A razão sinal/fundo (SBR) foi calculada pela razão entre a IS da região de interesse (ROI) e a intensidade média da ROI de fundo (IB): SBR = IS / IB. A SBR normalizada (SBRN) foi calculada como: SBRN = (IS - IB) / IB. Essa medida foi utilizada como uma forma complementar para facilitar a interpretação do contraste relativo entre sinal e fundo.

F. Análise estatística

Para a análise estatística dos registros de potencial de campo, foram avaliados os parâmetros DE, Ev/min e AmpMed. Nas imagens de microscopia confocal, foram avaliados % de pixels marcados, IC, SBR e a SBRN. As fatias foram mantidas como unidades observacionais individuais. Como duas fatias de cada animal foram utilizadas para cada análise, a dependência entre observações provenientes do mesmo animal foi considerada por meio de modelos lineares de efeitos mistos. Para cada variável, o grupo experimental foi incluído como efeito fixo, enquanto o animal foi incluído como intercepto aleatório. Foram estimadas as diferenças entre os grupos, os intervalos de confiança de 95% e os valores de *p* associados ao efeito do grupo. Os tamanhos de efeito corrigidos de *Hedges* foram calculados como medida complementar da magnitude das diferenças. A normalidade dos dados foi avaliada pelo teste de *Shapiro–Wilk*. O nível de significância adotado foi de 5%, e as análises foram realizadas no software JAMOVI [30], utilizando o módulo GAMLj para os modelos lineares de efeitos mistos.

III. Resultados

A. Viabilidade eletrofisiológica após perfusão com a solução SP-mACSF

Todos os experimentos de indução de atividade epileptiforme no giro dentado utilizando fatias do grupo perfundido com SP-mACSF geraram eventos epileptiformes de forma consistente, assim como no grupo controle. Os registros das atividades epileptiformes foram realizados simultaneamente em duas fatias por experimento, escolhidas aleatoriamente dentre as 10 a 12 dispostas na câmara de interface, respeitando os critérios de viabilidade tais como: regularidade dos eventos e nível de shift DC no entorno de 5 mV e PS no entorno de 10 mV. A Figura 2 mostra sinais típicos registrados em experimentos do grupo controle e do grupo perfundido com SP-mACSF. A comparação qualitativa dos sinais mostra padrões morfológicos similares entre os grupos, confirmando a preservação da viabilidade eletrofisiológica após perfusão com SP-mACSF.

A análise quantitativa dos parâmetros do FP confirma a compatibilidade entre os registros realizados em fatias provenientes de ambas as preparações. Os valores do parâmetro DE foram semelhantes entre os grupos. O grupo Controle apresentou média de 41,46 ± 6,38 s, enquanto o grupo SP-mACSF apresentou 41,20 ± 5,97 s. A diferença estimada entre Controle e SP-mACSF foi de -0,262 s, com IC95% de -6,264 a 5,740 e p=0,932. O tamanho de efeito foi desprezível, com *Hedges' g*=-0,041. Ev/min também foi semelhante entre os grupos, sendo os valores médios 0,772 ± 0,087 no grupo Controle e 0,803 ± 0,196 no grupo SP-mACSF. A diferença estimada entre Controle e SP-mACSF foi de 0,031 evento/min, com IC95% de -0,141 a 0,204 e p = 0,720. O tamanho de efeito foi pequeno, com *Hedges' g* = 0,199. O valor médio de Amp Med foi de 0,911 ± 0,196 mV no grupo Controle e 0,903 ± 0,362 mV no grupo SP-mACSF. A diferença estimada entre os grupos Controle e SP-mACSF foi de -0,008 mV, com IC95% de -0,272 a 0,256 e p = 0,952. O tamanho de efeito foi desprezível, com *Hedges' g* = -0,026. Em conjunto, os modelos mistos não indicaram diferenças estatisticamente significativas entre os grupos para os parâmetros do FP avaliados.

B. Qualidade do Sinal de Imunofluorescência

A análise qualitativa de imunofluorescência revelou que as secções histológicas do grupo Controle apresentaram elevado nível de marcação decorrente da autofluorescência, o que resultou em uma marcação inespecífica acentuada. Na Figura 3, ilustra-se a análise quantitativa da porcentagem de área marcada, obtida mediante a aplicação de um limiar (*threshold*) de intensidade de fluorescência fixo para ambos os grupos. O grupo SP-mACSF exibiu um valor médio significativamente menor em comparação ao Controle. Entretanto, astrócitos são um tipo celular de formato característico e expressam GFAP de uma forma que possibilita facilmente a identificação qualitativa de marcação inespecífica, o que propiciou a avaliação das micrografias (Figura 4) e confirmou que os índices de marcação elevados no grupo Controle correspondiam a marcação inespecífica gerada por autofluorescência e não o resultado de uma ligação genuína do complexo anticorpo primário secundário direcionado à proteína glial fibrilar ácida GFAP (Figura 3). Assim, realizamos a quantificação da relação sinal / ruído (SBR) e constatamos quantitativamente o ganho de qualidade de marcação nas fatias do grupo SP-mACSF.

A análise estatística dos dados quantitativos da histologia apontou que a porcentagem de pixels marcados foi maior no grupo Controle do que no grupo SP-mACSF. A média foi de 31,00 ± 18,10% no grupo Controle e 8,61 ± 6,02% no grupo SP-mACSF. A diferença estimada entre Controle e SP-mACSF foi de -22,396 pontos percentuais, com IC95% de -37,706 a -7,086 e p = 0,004. O tamanho de efeito foi grande, com *Hedges' g* = -1,590. A BCI foi menor no grupo Controle do que no grupo SP-mACSF. Os valores médios foram 38,56 ± 15,48 no grupo Controle e 77,76 ± 22,69 no grupo SP-mACSF. A diferença estimada entre Controle e SP-mACSF foi de 39,200 unidades, com IC95% de 22,176 a 56,224 e p < 0,001. O tamanho de efeito foi grande, com *Hedges' g* = 1,933. A SBR foi menor no grupo Controle do que no grupo SP-mACSF. A média foi de 1,44 ± 0,23 no grupo Controle e 2,82 ± 1,21 no grupo SP-mACSF. A diferença estimada entre Controle e SP-mACSF foi de 1,380, com IC95% de 0,617 a 2,142 e p < 0,001. O tamanho de efeito foi grande, com *Hedges' g* = 1,519. A SBRN apresentou média de 0,44 ± 0,23 no grupo Controle e 1,82 ± 1,21 no grupo SP-mACSF. A diferença estimada entre Controle e SP-mACSF foi de 1,380, com IC95% de 0,617 a 2,142 e p < 0,001, com *Hedges' g* = 1,519. Sendo a SBRN matematicamente equivalente à SBR -

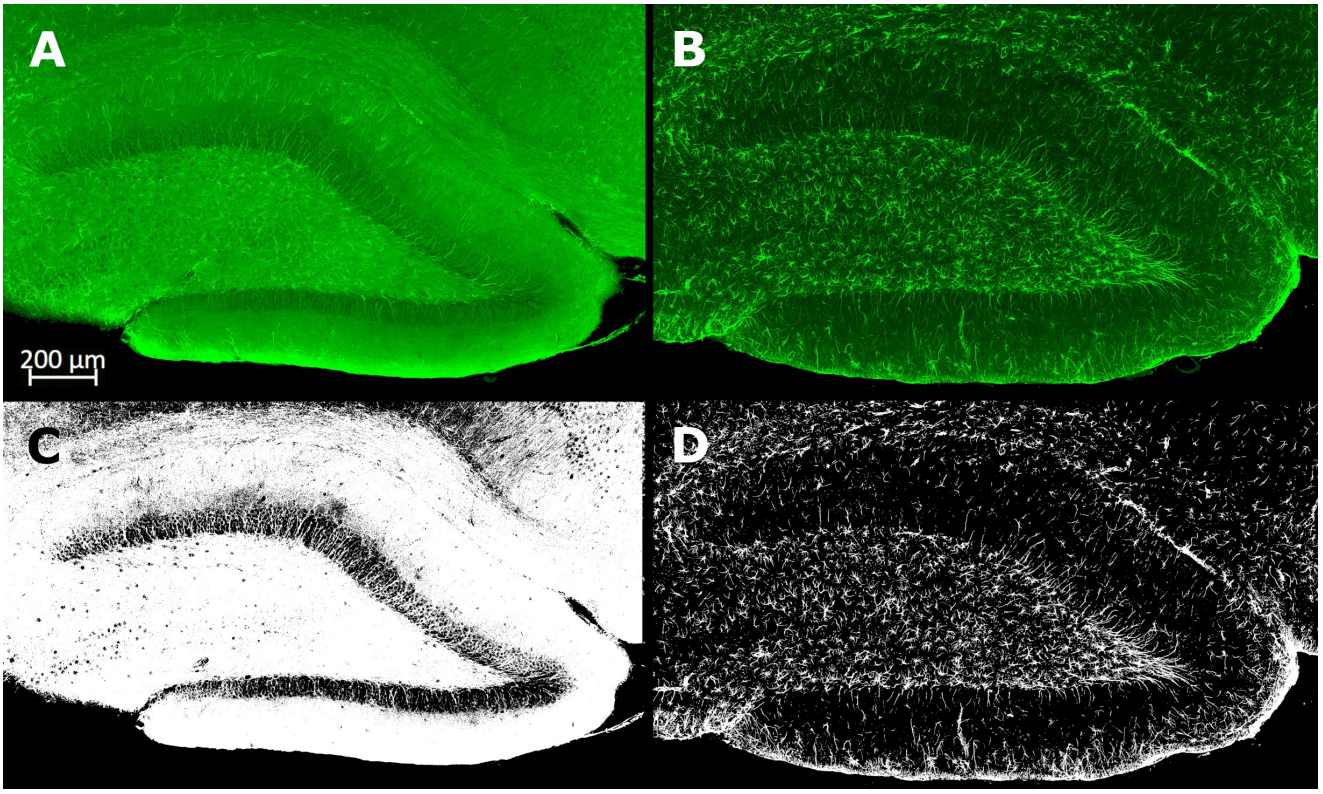


Figura 4. Micrografias representativas da marcação anti-GFAP no giro dentado de fatias de hipocampo: Grupo Controle (A) e grupo SP-mACSF (B). (C, D) Imagens binarizadas para quantificação com limiar fixo em fatia do grupo Controle (C) e fatia do grupo SP-mACSF (D). Identificação de marcação inespecífica causada pela autofluorescencia de *background* nas imagens do grupo Controle (A, C).

1, os valores de diferença, IC95%, *p* e tamanho de efeito foram iguais aos observados para a SBR.

Os modelos lineares de efeitos mistos indicaram que os grupos Controle e SP-mACSF apresentaram parâmetros eletrofisiológicos semelhantes, sem diferenças detectáveis em DE, Ev/min ou Amp Med. Em contraste, o grupo Controle apresentou maior % de pixels marcados, enquanto o grupo SP-mACSF apresentou maiores BCI, SBR e SBRN. Esse conjunto de resultados sugere que a perfusão com SP-mACSF preservou a viabilidade funcional das fatias e, simultaneamente, melhorou a qualidade relativa da marcação por imunofluorescência. A maior porcentagem de pixels marcados no grupo Controle, quando acompanhada de menor BCI, menor SBR e SBRN, é compatível com maior contribuição de fluorescência inespecífica e autofluorescência de fundo. Assim, o grupo SP-mACSF apresentou menor área acima do limiar, mas maior contraste relativo entre o sinal da ROI e o fundo.

IV. Discussão

O principal achado deste estudo foi demonstrar que a perfusão transcardíaca com SP-mACSF, associada ao método de preparação de fatias de hipocampo desenvolvido, permite realizar, de forma eficiente, análises eletrofisiológicas e morfológicas utilizando tecido proveniente do mesmo animal, sem comprometer a viabilidade funcional das fatias e, ao mesmo tempo, melhorar a qualidade das marcações por imunofluorescência. Mais do que uma adaptação técnica, essa estratégia representa um avanço metodológico ao proporcionar a possibilidade de correlação direta entre propriedades funcionais e estruturais do tecido nervoso.

Nesse contexto, a análise histológica demonstrou que, em contraste com as secções do grupo Controle, que exibiram acentuada autofluorescência de fundo, o tecido submetido à perfusão com SP-mACSF demonstrou uma marcação imuno-fluorescente nítida. A análise das micrografias confocais corroborou essas observações qualitativas, revelando uma proporção significativamente menor de área marcada no grupo perfundido quando aplicado um limiar fixo de detecção. Tendo em vista que o valor médio elevado de % de pixels marcados no grupo Controle derivou predominantemente de fluorescência inespecífica, o que ficou evidenciado pelos valores médios de IC, SBR e SBRN, conclui-se que o uso do SP-mACSF foi eficaz em eliminar a autofluorescência reduzindo a quantificação de artefatos, permitindo a detecção fidedigna dos fluoróforos secundários efetivamente acoplados à proteína-alvo, a GFAP. Nesse sentido, a literatura estabelece amplamente que os eritrócitos retidos na microvasculatura são importantes fontes de autofluorescência endógena no tecido nervoso [31, 32, 33, 34]. Dessa forma, os resultados evidenciam que o protocolo de perfusão transcardíaca com SP-mACSF foi eficaz na lavagem vascular, mitigando a autofluorescência de forma considerável e garantindo uma detecção mais precisa e confiável dos antígenos-alvo.

No aspecto eletrofisiológico, não é possível afirmar, a partir dos dados obtidos no nosso estudo, se a glicose aumentada (de 10 mM para 20 mM) em relação ao protocolo original de [24], exerceu, de fato, efeito protetor contra danos hipóxicos, conforme demonstrado por [26]. Entretanto, as fatias obtidas de animais submetidos à perfusão com SP-mACSF apresentaram viabilidade equivalente para indução de atividade epileptiforme não sináptica no giro dentado quando comparadas a fatias controle não perfundidas com SP-mACSF. Assim, apesar da variabilidade na morfologia dos eventos entre fatias, comum na indução de atividades epileptiformes não-sinápticas [27, 35, 36, 37] a análise qualitativa das gravações de FP revelou padrões morfológicos de atividade epileptiforme indistinguíveis entre os grupos Controle e SP-mACSF. A semelhança entre os registros é evidenciada por traços brutos sobrepostos e ilustrada pela densidade espectral de potência, apresentadas na Figura 2. Além disso, a análise e a comparação dos parâmetros quantitativos Ev/min, DE e Amp Med dos eventos não demonstrou diferença estatisticamente significativa nos valores obtidos para o grupo Controle e para o grupo SP-mACSF, corroborando o que foi constatado por meio da análise qualitativa dos registros do FP. Esses resultados indicam que a perfusão transcardíaca com SP-mACSF foi capaz de preservar a função neuronal mantendo as fatias de hipocampo em um estado de viabilidade capaz de deflagrar e sustentar eventos epileptiformes, mesmo diante do procedimento invasivo de lavagem vascular, corroborando os achados de [24].

Além disso, a equivalência qualitativa e quantitativa em termos dos parâmetros avaliados no FP, entre as gravações eletrofisiológicas obtidas em fatias perfundidas e fatias não perfundidas sugere que o período prolongado de equilíbrio na câmara de perfusão (120-150 min), somado ao tempo de descanso na câmara de interface (60 minutos), ambos sob perfusão de solução Ringer-ACSF normal, foi capaz de promover a extração completa (*washout*) dos anestésicos residuais do tecido, impedindo qualquer interferência na indução da atividade epileptiforme. Esse tempo é compatível com o tempo de 4 a 5 horas utilizado para tecido de hipocampo humano proveniente de ressecção cirúrgica de pacientes com epilepsia [38].

Em conjunto, os dados apresentados evidenciam que o método desenvolvido e validado neste estudo constitui um avanço metodológico significativo para a neurociência integrativa. A técnica viabiliza, a partir de um mesmo espécime animal, a obtenção de registros eletrofisiológicos de alta fidelidade e ensaios de imunofluorescência que produzem imagens de alta qualidade. Ao promover a lavagem vascular com SP-mACSF, a abordagem garante a preservação da integridade e excitabilidade neuronal, eliminando a autofluorescência inerente às células sanguíneas retidas no tecido. Desse modo, o método atinge uma dupla funcionalidade: sustenta a viabilidade tecidual necessária para a indução e o registro da atividade epileptiforme, além de assegurar a preservação estrutural para análises microscópicas de alta resolução.

V. Conclusão

Os resultados demonstram que a perfusão transcardíaca com SP-mACSF preserva a viabilidade eletrofisiológica das fatias de hipocampo, ao mesmo tempo em que reduz significativamente a autofluorescência e melhora a qualidade das análises por imunofluorescência. Dessa forma, o método permite a integração confiável de avaliações funcionais e morfológicas, utilizando tecido proveniente do mesmo animal e superando uma limitação metodológica recorrente na neurociência experimental. Além das vantagens técnicas, a estratégia proposta maximiza o aproveitamento científico de cada animal experimental e tem potencial para reduzir a variabilidade entre indivíduos e fortalecer a correlação entre alterações estruturais e funcionais, em consonância com os princípios dos 3Rs. Embora validado neste estudo em um

modelo de atividade epileptiforme não sináptica, o protocolo apresenta potencial para aplicação em diferentes paradigmas eletrofisiológicos e modelos experimentais, constituindo uma plataforma metodológica promissora para pesquisas integrativas em neurociência.

AGRADECIMENTOS

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), ao Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG), ao Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais (IF Sudeste MG), à Universidade Federal de São João del Rei (UFSJ) e à Universidade Federal de São Paulo (UNIFESP) pelo apoio financeiro e institucional.

REFERÊNCIAS

[1] T. A. Milligan, "Epilepsy: a clinical overview," *Am. J. Med.*, vol. 134, no. 7, pp. 840–847, 2021.

[2] V. L. Feigin *et al.*, "Global, regional, and national burden of epilepsy, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet Public Health*, vol. 10, no. 3, pp. e203–e227, 2025.

[3] J. A. French, "Refractory epilepsy: clinical overview," *Epilepsia*, vol. 48, pp. 3–7, 2007.

[4] P. Kwan and M. J. Brodie, "Early identification of refractory epilepsy," *N. Engl. J. Med.*, vol. 342, no. 5, pp. 314–319, 2000.

[5] J. G. R. Jefferys and H. L. Haas, "Synchronized bursting of CA1 hippocampal pyramidal cells in the absence of synaptic transmission," *Nature*, vol. 300, no. 5891, pp. 448–450, 1982.

[6] C. P. Taylor and F. E. Dudek, "Synchronous neural afterdischarges in rat hippocampal slices without active chemical synapses," *Science*, vol. 218, no. 4574, pp. 810–812, 1982.

[7] J. G. Jefferys, "Nonsynaptic modulation of neuronal activity in the brain: electric currents and extracellular ions," *Physiol. Rev.*, vol. 75, no. 4, pp. 689–723, 1995.

[8] L. L. Antonio *et al.*, "In vitro seizure-like events and changes in ionic concentration," *J. Neurosci. Methods*, vol. 260, pp. 33–44, 2016.

[9] C. G. Dulla *et al.*, "How do we use in vitro models to understand epileptiform and ictal activity? A report of the TASK1-WG4 group of the ILAE–AES Joint Translational Task Force," *Epilepsia Open*, vol. 3, no. 4, pp. 460–473, 2018.

[10] S. S. Jahromi, M. R. Pelletier, P. J. McDonald, H. Khosravani, and P. L. Carlen, "Antiepileptic efficacy of topiramate: Assessment in two in vitro seizure models," *Brain Res.*, vol. 872, no. 1–2, pp. 20–28, 2000.

[11] I. Hyun, J. C. Scharf-Deering, and J. E. Lunshof, "Ethical issues related to brain organoid research," *Brain Res.*, vol. 1732, Art. no. 146653, 2020.

[12] M. Avoli and J. G. R. Jefferys, "Models of drug-induced epileptiform synchronization in vitro," *J. Neurosci. Methods*, vol. 260, pp. 26–32, 2016.

[13] D. M. Magalhães *et al.*, "Ex vivo model of epilepsy in organotypic slices—a new tool for drug screening," *J. Neuroinflammation*, vol. 15, no. 1, Art. no. 203, 2018, doi: 10.1186/s12974-018-1225-2.

[14] J. V. Raimondo *et al.*, "Methodological standards for in vitro models of epilepsy and epileptic seizures. A TASK1-WG4 report of the AES–ILAE Translational Task Force of the ILAE," *Epilepsia*, vol. 58, pp. 40–52, 2017.

[15] J. T. Ting, T. L. Daigle, M. A. Kozur, and K. J. Bender, "Preparation of acute brain slices using an optimized N-methyl-D-glucamine protective recovery method," *J. Vis. Exp.*, no. 132, Art. no. e57039, 2018.

[16] E. S. Lein, X. Zhao, and F. H. Gage, "Acute hippocampal slice preparation and study of synaptic transmission using whole-cell patch clamp," *Cold Spring Harb. Protoc.*, vol. 2011, no. 10, Art. no. pdb.prot5564, 2011, doi: 10.1101/pdb.prot5564.

[17] Y. Buskila *et al.*, "Extending the viability of acute brain slices," *Sci. Rep.*, vol. 4, no. 1, Art. no. 5309, 2014, doi: 10.1038/srep05309.

[18] G. Losi *et al.*, "A brain slice experimental model to study the generation and the propagation of focally-induced epileptiform activity," *J. Neurosci. Methods*, vol. 260, pp. 125–131, 2016.

[19] X. Wang *et al.*, "Changes in the excitability of the medial parabrachial nucleus neurons during the chronic phase of pilocarpine-induced epilepsy in mice," *Front. Pharmacol.*, vol. 16, Art. no. 1474254, 2025.

[20] X. Lu and J. Lichtman, "Preserving extracellular space for high-quality optical and ultrastructural studies of whole mammalian brains," *Cell Rep. Methods*, vol. 3, no. 7, Art. no. 100542, 2023, doi: 10.1016/j.crmeth.2023.100542.

[21] I. Korshunova *et al.*, "A multispectral LED array for the reduction of background autofluorescence in brain tissue," *J. Neurosci. Methods*, vol. 216, no. 2, pp. 171–179, 2013.

[22] A. McGowan *et al.*, "Autofluorescence: A potential pitfall in immunofluorescence-based inflammation grading," *Histopathology*, vol. 72, no. 5, pp. 926–936, 2018.

[23] U. W. E. Heinemann, O. Kann, and S. Schuchmann, "An overview of in vitro seizure models in acute and organotypic slices," in *Models of Seizures and Epilepsy*, 2006, pp. 35–44.

[24] S. Jalini, H. Ye, A. A. Tonkikh, M. P. Charlton, and P. L. Carlen, "Raised intracellular calcium contributes to ischemia-induced depression of evoked synaptic transmission," *PLoS One*, vol. 11, no. 3, Art. no. e0148110, 2016.

[25] L. E. C. Santos *et al.*, "Alcohol abuse promotes changes in non-synaptic epileptiform activity with concomitant expression changes in cotransporters and glial cells," *PLoS One*, vol. 8, no. 11, Art. no. e78854, 2013.

[26] G. F. Tian and A. J. Baker, "Protective effect of high glucose against ischemia-induced synaptic transmission damage in rat hippocampal slices," *J. Neurophysiol.*, vol. 88, no. 1, pp. 236–248, 2002.

[27] J. S. Schweitzer, H. Wang, Z. Q. Xiong, and J. L. Stringer, "pH sensitivity of non-synaptic field bursts in the dentate gyrus," *J. Neurophysiol.*, vol. 84, no. 2, pp. 927–933, 2000.

[28] F. N. Mendonça *et al.*, "Physical exercise restores the generation of newborn neurons in an animal model of chronic epilepsy," *Front. Neurosci.*, vol. 11, p. 98, 2017.

[29] G. S. Nogueira *et al.*, "Enhanced nonsynaptic epileptiform activity in the dentate gyrus after kainate-induced status epilepticus," *Neuroscience*, vol. 303, pp. 59–72, 2015.

[30] The jamovi project, *jamovi*, Version 2.6, Computer software, 2024.

[31] P. Gandhi and R. Khare, "A unique immunofluorescence protocol to detect protein expression in vascular tissues tacking a long standing pathological hitch," *Turk. J. Pathol.*, vol. 34, no. 1, 2018.

[32] S. Lowndes, "Blood interference in fluorescence spectrum experiment, analysis and comparison with intraoperative measurements on brain tumor," 2010.

[33] S. Wang *et al.*, "Blocking autofluorescence in brain tissues affected by ischemic stroke, hemorrhagic stroke, or traumatic brain injury," *Front. Immunol.*, vol. 14, Art. no. 1168292, 2023.

[34] N. C. Whittington and S. Wray, "Suppression of red blood cell autofluorescence for immunocytochemistry on fixed embryonic mouse tissue," *Curr. Protoc. Neurosci.*, vol. 81, no. 1, pp. 2–28, 2017.

[35] J. Lian, M. Bikson, J. Shuai, and D. M. Durand, "Propagation of non-synaptic epileptiform activity across a lesion in rat hippocampal slices," *J. Physiol.*, vol. 537, no. Pt 1, p. 191, 2001.

[36] E. Pan and J. L. Stringer, "Burst characteristics of dentate gyrus granule cells: evidence for endogenous and nonsynaptic properties," *J. Neurophysiol.*, vol. 75, no. 1, pp. 124–132, 1996.

[37] H. L. Haas and J. G. Jefferys, "Low-calcium field burst discharges of CA1 pyramidal neurones in rat hippocampal slices," *J. Physiol.*, vol. 354, no. 1, pp. 185–201, 1984.

[38] S. Z. Reyes-Garcia *et al.*, "Different patterns of epileptiform-like activity are generated in the sclerotic hippocampus from patients with drug-resistant temporal lobe epilepsy," *Sci. Rep.*, vol. 8, Art. no. 7116, 2018.

Clinical Response after Transcranial Photobiomodulation on Parkinsonian Tremor: Clinical and Functional Outcomes in a Case Report

Fernanda de Souza Silva
Graduate Program in Biomedical Engineering
Universidade Anhembi, Morumbi
São José dos Campos, Brazil
fisio.fer@yahoo.com

Pâmela Camila Pereira
Graduate Program in Biomedical Engineering
Universidade Anhembi, Morumbi
São José dos Campos, Brazil
pam_milaf@yahoo.com.br

Renato Amaro Zângaro
Graduate Program in Biomedical Engineering
Universidade Anhembi, Morumbi
São José dos Campos, Brazil
razangaro@gmail.com

Abstract* — Parkinson's Disease (PD) is a neurodegenerative condition characterized by progressive motor impairment, with tremor being one of its most disabling manifestations. Transcranial Photobiomodulation (tPBM) has been investigated as a non-invasive therapeutic strategy with potential neuromodulatory effects on neural activity. The aim of this study was to describe the clinical response observed after the isolated application of tPBM on tremors in a patient with PD, using the MDS-UPDRS scale as a clinical outcome measure. This was a prospective longitudinal case study conducted in a home-based setting. The participant underwent a tPBM protocol five times per week for 12 weeks, totaling 60 sessions. The intervention was performed using a near-infrared light-emitting device (808 nm) applied for 20 minutes per session. Motor assessment was conducted before and after the intervention using the MDS-UPDRS scale. A reduction in the total score from 35 to 17 points was observed, corresponding to an approximate 51% reduction in the evaluated motor scores, in addition to a consistent decrease in resting, postural, and kinetic tremor scores, as well as rigidity. Functional improvements in activities of daily living, especially fine motor coordination, were also observed. The findings describe a clinical response temporally associated with tPBM in this participant. However, due to the case study design, the results should be interpreted with caution, without allowing inferences of efficacy or causality, and controlled studies are required to confirm these findings.

Keywords — Parkinson's Disease; Tremor; Neurorehabilitation; MDS-UPDRS.

I. INTRODUCTION

Parkinson's Disease (PD) is the second most prevalent neurodegenerative disorder associated with aging and is characterized by progressive and disabling motor alterations, including resting tremor, rigidity, bradykinesia, and postural instability. Among these manifestations, tremor stands out as one of the most disabling symptoms, significantly impairing functionality and life quality. Despite advances in pharmacological and surgical therapies, limitations related to sustained efficacy and adverse effects reinforce the need for complementary therapeutic approaches [1–2].

In this context, Transcranial Photobiomodulation (tPBM) has emerged as a promising strategy for modulating neural activity. This approach is based on the application of light at specific wavelengths capable of interacting with intracellular chromophores, particularly cytochrome c oxidase, potentially promoting increased ATP production and improving neuronal bioenergetic efficiency. Additionally, studies suggest that tPBM may be associated with reduced oxidative stress and modulation of inflammatory processes, mechanisms directly involved in the pathophysiology of PD [1–5].

From a functional perspective, these effects may contribute to the reorganization of neural circuits and improvements in the dynamics of systems responsible for motor control. Experimental and clinical studies suggest potential neuroprotective effects of tPBM and a possible enhancement of neuroplasticity, with observed associations between the intervention and reductions in motor symptoms, including tremor [2–5].

However, despite promising findings, important gaps remain in the literature, especially regarding the isolated application of tPBM and its ability to specifically modulate motor parameters assessed by standardized clinical scales. Most available studies combine the technique with other interventions, making it difficult to understand its direct effects.

Therefore, the present study aimed to describe the clinical response observed after the isolated application of tPBM on tremors in a patient with PD, using the MDS-UPDRS as the primary outcome measure, thereby contributing to the understanding of the clinical and functional effects of this intervention.

II. MATERIALS AND METHODS

A. Ethical Considerations

This study was approved by the Research Ethics Committee of Centro Universitário de Itajubá – FEPI (CAAE 73135023.5.0000.5094). The participant was properly informed about the objectives and procedures of the study and signed an informed consent form prior to the beginning of treatment.

This was a prospective, descriptive, and longitudinal case study conducted in the participant's home environment.

B. Clinical Description of the Patient

The present study was conducted with a 62-year-old male participant diagnosed with PD who was under regular clinical follow-up and on regular use of the medication Prolopa® BD 100/25 mg, at a dose of one tablet at 6:00 AM, 11:30 AM, and 5:00 PM. The medication regimen remained stable during the 12 weeks of the protocol. The patient presented predominant tremor in the Right Upper Limb (RUL), with functional impairment in activities requiring fine motor control.

Eligibility criteria for inclusion in this case study included a clinical diagnosis of PD, the presence of motor

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

tremor, and stable medication use. Exclusion criteria included the presence of other associated neurological diseases, a history of recent neurosurgical procedures, photosensitivity disorders, or clinical conditions that could interfere with tPBM application.

An initial evaluation was performed, including the collection of personal, anthropometric, and clinical data, such as a body weight of 86 kg, a height of 1.72 m, and a Body Mass Index (BMI) of 29.1 kg/m².

Motor function assessment was performed using the MDS-UPDRS (Movement Disorder Society–Unified Parkinson's Disease Rating Scale), applied before and after the intervention period. The analysis focused primarily on items related to resting tremor, postural tremor, kinetic tremor, and rigidity, allowing comparison of the participant's clinical progression throughout the treatment. The pre- and post-intervention evaluations were performed in the ON state, approximately 60 minutes after the administration of Prolopa®, maintaining the relative timing of medication.

C. Motor Function Assessment

The MDS-UPDRS is a widely used scale for assessing the severity and progression of PD, encompassing both motor and non-motor aspects of the disease. In the present study, the analysis focused primarily on motor function items, with emphasis on resting tremor, postural tremor, kinetic tremor, and rigidity.

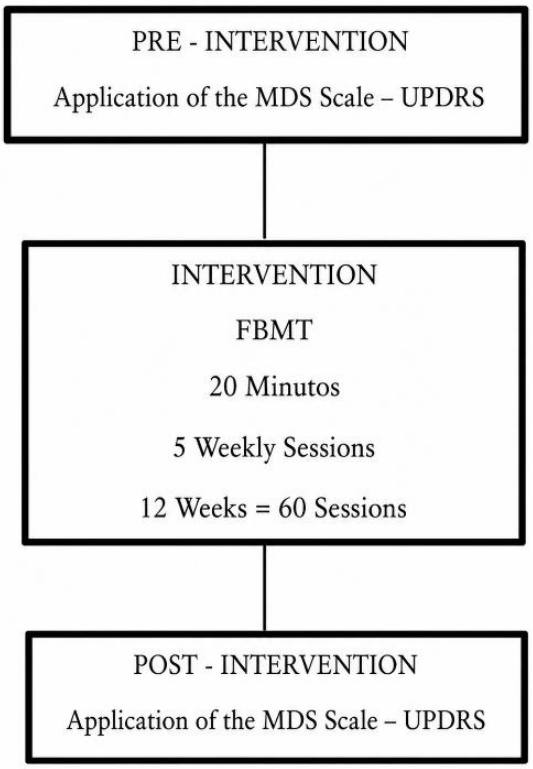
The scale was administered by another author of the study, who was previously trained and distinct from the researcher responsible for the intervention sessions, in a controlled environment, with the patient under standardized resting conditions and during the execution of specific motor tasks. During the assessment, the participant was instructed to remain seated for resting tremor evaluation and subsequently perform voluntary upper limb movements to assess action tremors and motor control. Scores were assigned according to the intensity and frequency of the observed clinical signs, ranging from 0 (no impairment) to 4 (severe impairment). The evaluator was aware of the pre- or post-intervention timing, with no blinding. The scale was applied at two distinct time points: before the intervention and after completion of the tPBM protocol, allowing comparison of clinical findings before and after the treatment period.

D. Intervention Protocol

As this was a case study, no group allocation was performed, and an individualized approach was adopted to analyze the effects of the intervention. The participant underwent exclusive tPBM application, without association with other therapeutic interventions, allowing the isolated evaluation of the effects of the technique. However, the absence of a control group or Sham condition prevents attributing the observed changes exclusively to tPBM.

The protocol was conducted five times per week for 12 consecutive weeks, totaling 60 sessions. tPBM was applied using a device positioned over the cranial region containing light-emitting diodes (LEDs) operating in the near-infrared range (808 nm), with a duration of 20 minutes per session. The application was performed with the patient in a comfortable position, with the helmet covering the entire cranial region, without targeting specific cortical regions.

Figure 1 – Experimental design of the case study.



Source: Personal Archive

E. tPBM Protocol in PD

The tPBM intervention was performed using a helmet composed of gallium-aluminum-arsenide (GaAlAs) LED emitters arranged to cover the entire cranial region, without selective targeting of a specific cortical target. The device was designed to provide homogeneous distribution of light radiation over the scalp, favoring light penetration into encephalic structures.

The helmet emitted near-infrared radiation at a wavelength of 808 nm, a value widely described in the literature for its tissue penetration capacity and interaction with cytochrome c oxidase, favoring cellular bioenergetic effects. The device featured an acrylic encapsulation with an approximate diameter of 5 mm per emitter. The system presented a total optical power of 6 W and a power density of 3.41 mW/cm², allowing application for 1200 seconds per session, corresponding to 20 minutes, with an approximate energy density of 4.09 J/cm² per session.

During application, the participant remained in a comfortable position with the helmet properly adjusted to the cranial region, ensuring adequate contact and uniform distribution of light. No additional anatomical landmarks were used for cortical targeting, and the same fitting procedure was repeated across sessions. To ensure

appropriate biosafety conditions, the device was disinfected with 70% alcohol after each session.

F. Statistical Analysis

As this was a case study, no inferential statistical analyses were performed. Data analysis was conducted descriptively, comparing pre- and post-intervention findings. The analysis focused primarily on MDS-UPDRS scores related to resting tremor, postural tremor, kinetic tremor, and rigidity, allowing the evaluation of the participant's clinical progression throughout the tPBM protocol. No intermediate or follow-up assessments were conducted, making it impossible to characterize the trajectory or persistence of the observed changes.

Results were presented using absolute values and graphical representations, enabling visualization of the changes observed after the intervention period. Data analysis was conducted in an integrated manner, considering both quantitative changes and the clinical relevance of the observed findings.

III. RESULTS

After completion of the 60 tPBM sessions, a descriptive reduction in the total MDS-UPDRS score was observed, decreasing from 35 to 17 points, representing an approximate 51% reduction in the evaluated motor parameters. Descriptive improvement was observed in resting tremor, kinetic tremor, and articular rigidity.

The present study involved a single participant diagnosed with PD presenting predominant tremor in the RUL. During the initial MDS-UPDRS assessment, the participant presented a total score of 35 points, indicating mild-to-moderate motor impairment, with emphasis on the presence of resting, postural, and kinetic tremor in the RUL, as well as rigidity in the same segment and cervical spine.

Regarding specific items, reductions were observed in scores related to resting tremor, postural tremor, and kinetic tremor, in addition to decreased articular rigidity and improvement in functional tasks such as handwriting, which is associated with fine motor control. Parameters related to gait and postural stability remained stable throughout the intervention period.

Table 1 presents the comparison of scale scores in the pre- and post-intervention periods, while Figure 2 graphically illustrates the reduction in the evaluated motor parameters.

In the MDS-UPDRS, scores range from 0 (no impairment) to 4 (severe impairment), representing increasing levels of motor symptom severity. A score of 2 corresponds to mild-to-moderate impairment, characterized by evident clinical manifestations already interfering with movement quality and precision, whereas a score of 1 indicates mild impairment, with subtle signs and lower functional impact.

In this patient, the reduction from 2 to 1 suggests a transition from a motor pattern with noticeable impairment to a milder condition in which symptoms remained present but with lower intensity and reduced interference in motor task execution. Since accelerometry or IMUs were not used, it was not possible to objectively quantify the amplitude or frequency of the tremor.

Table 1 – Comparison of MDS-UPDRS Scores Pre- and Post-tPBM Intervention.

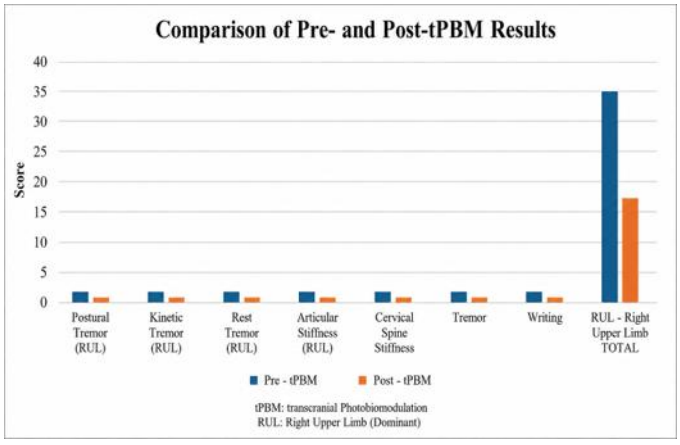
Variable	Pre-tPBM	Post-tPBM
Postural Tremor (RUL)	2	1
Kinetic Tremor (RUL)	2	1
Resting Tremor (RUL)	2	1
Articular Rigidity (RUL)	2	1
Cervical Spine - Articular Rigidity	2	1
Tremor	2	1
Handwriting	2	1
Total Score	**35**	**17**

Caption: Right Upper Limb; tPBM: Transcranial Photobiomodulation
Source: Personal Archive

From a clinical perspective, this change may be associated with improved motor control, greater movement fluidity, and reduced tremor amplitude observed in the participant, potentially contributing positively to functional capacity. Furthermore, although numerically representing only a one-point difference on the scale, this variation has important clinical relevance, especially in patients with PD, in whom small score changes may be associated with meaningful functional gains in activities of daily living, particularly those requiring coordination and manual dexterity, such as handwriting. However, a minimal clinically important difference (MCID) has not been defined for these items, and the absence of instrumental measures limits the interpretation of clinical relevance.

Figure 2 presents the comparison of MDS-UPDRS scores before and after the intervention, demonstrating a descriptive reduction across all evaluated motor parameters.

Figure **2** – Scores Pre- and Post-tPBM Intervention.



Source: Personal Archive

A homogeneous reduction pattern was observed in the items related to tremor (postural, kinetic, and resting) and rigidity, with scores decreasing from 2 to 1 in all these domains. Since this is an ordinal clinical scale applied at only two distinct time points, this pattern does not allow demonstrating response stability or low intra-individual variability.

Additionally, functional items such as tremor and handwriting also showed score reductions, suggesting a possible association between the intervention and improvement in fine motor task execution. This finding reinforces the potential relationship between improvements in clinical parameters and participant functionality.

A reduction in the total scale score from 35 to 17 points was also observed, representing an approximate decrease of 51%. This magnitude of change suggests a potential association between tPBM and improvement in the participant's global motor performance. In the absence of EEG, neuroimaging, or other neurophysiological measures, it is not possible to assert that these changes resulted from the modulation of neural circuits.

IV. DISCUSSION

The results observed in the present study, assessed through the MDS-UPDRS, were accompanied by consistent reduction in motor scores following tPBM application. An approximate 51% decrease in the total scale score was observed, from 35 to 17 points, suggesting a potential association between the intervention and improvement in the evaluated motor parameters. This pattern was accompanied by a homogeneous reduction in scores related to resting tremor, postural tremor, kinetic tremor, and rigidity, in addition to improvements in functional tasks associated with fine motor control, such as handwriting. These findings describe a change between two time points, without allowing the determination of its trajectory, persistence, or causal relationship with tPBM.

The findings observed in the present study are consistent with results described in the literature. Liebert et al. (2021) reported improvements in motor parameters in patients with PD undergoing tPBM, including reductions in tremor-related manifestations and functional improvement. Similarly, Ahrabi et al. (2022) described a potential neuroprotective effect of tPBM on motor circuits involved in movement control, suggesting an association between neuronal bioenergetic modulation and functional improvement observed in neurodegenerative diseases [1,5].

From a clinical perspective, tremor represents one of the most disabling motor manifestations of PD and is frequently associated with functional limitation, loss of independence, and impaired quality of life. According to Abusrair, Elsekaily, and Bohlega (2022), tremor-related motor alterations may significantly impact activities of daily living dependent on fine motor coordination, such as handwriting, eating, and object manipulation. In this context, the reduction observed in tremor-related scores may suggest important functional relevance, even with small variations of the clinical scale. However, the absence of an objective measure of tremor and of additional functional evaluations limits this interpretation.

From a physiological and biophysical perspective, tPBM may act as an energetic stimulus potentially capable of modulating cellular processes related to neuronal activity. The application of near-infrared light (808 nm) allows greater tissue penetration, reaching cortical structures and potentially influencing neuronal activity [11]. The interaction of light with intracellular chromophores, especially cytochrome c oxidase, may promote increased mitochondrial activity and ATP production, resulting in possible improvement in neuronal bioenergetic efficiency [5–6]. These mechanisms are hypotheses proposed by the literature and were not measured in the present study.

Additionally, the reduction in rigidity identified in the present study may be associated with modulation of neural pathways related to muscle tone, possibly influenced by reduced central inflammatory processes. Experimental studies suggest that tPBM may reduce microglial activation and pro-inflammatory cytokine release, potentially contributing to improved neuronal function and motor control [5,12].

Regarding functional items of the scale, such as tremor and handwriting, the observed improvement suggests a possible alteration in the execution of fine motor tasks. Although the literature presents hypotheses related to the integration between motor planning and execution, the present study did not directly evaluate the signal-to-noise ratio or the activity of motor circuits. Therefore, it is not possible to confirm neurophysiological mechanisms from the clinical outcomes obtained [5,10].

From a biophysical standpoint, the observed effects may also be related to the irradiation parameters used, such as wavelength, energy density, and exposure time. However, as this is a single case and there is no comparison between different doses, it is not possible to establish a dose-dependent behavior of the response to tPBM.

Although the identified findings are promising, it should be considered that case studies present important limitations regarding generalization of the results, since they do not allow control of external variables or comparison with a control group. In addition, factors such as individual clinical variability, placebo effects, and possible natural disease fluctuations may influence the reported findings. Therefore, the results should be interpreted with caution, reinforcing the need for controlled studies with greater methodological robustness to confirm the effects of tPBM [3].

Another relevant aspect is that the intervention was applied in isolation, which allows a possible association between the identified clinical response and isolated tPBM application without interference from other therapeutic variables [4,10]. However, the absence of a control group does not allow attributing the clinical response exclusively to tPBM [4,10]. Furthermore, the inclusion of only a single participant limits the analyses of interindividual variability. Electrophysiological or imaging parameters that could provide greater support for interpretation of the involved mechanisms were not evaluated [10].

Therefore, the results of the present study describe clinical changes temporally associated with the tPBM protocol in this participant. However, the data do not allow demonstrating modulation of neural circuits, therapeutic efficacy, or a causal relationship. Future studies with greater methodological robustness are necessary to confirm these findings.

V. CONCLUSION

Near-infrared tPBM applied in isolation in the present case report was associated with a reduction in the motor scores assessed by the MDS-UPDRS in a patient with PD. Decreases in tremor and rigidity-related parameters were observed, in addition to improvements in functional tasks related to fine motor control, such as handwriting.

The observed findings suggest a potential association between the application of tPBM and the observed clinical changes. However, the study does not allow asserting functional modulation of neural circuits, clinical efficacy, or a causal relationship.

Although the results are limited by the case study design, the clinical response observed justifies further investigations of tPBM as a non-invasive strategy in the rehabilitation of patients with PD. However, future controlled studies with a larger sample size, blinded evaluators, serial assessments, a follow-up period, and objective instrumental measures are necessary before conclusions regarding its clinical efficacy can be supported.

ACKNOWLEDGMENTS

This study was supported by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES). The authors also thank Instituto Ânima for the support provided for the development of this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest

REFERENCES

[1] Liebert A, Bicknell B, Laakso EL, Heller G, Jalilitabaei P, Tilley S, Mitrofanis J, Kiat H. Improvements in clinical signs of Parkin-son's disease using photobiomodulation: a prospective proof-of-concept study. BMC Neurol. 2021 Jul 2;21(1):256. doi: 10.1186/s12883-021-02248-y. PMID: 34215216; PMCID: PMC8249215.

[2] Ahrabi B, Tabatabaei Mirakabad FS, Niknazar S, et al. Photobi-omodulation therapy and cell therapy improved Parkinson's dis-ease by neuro-regeneration and tremor inhibition. J Lasers Med Sci. 2022;13:e28. doi:10.34172/jlms.2022.28.

[3] McGee C, Liebert A, Bicknell B, Pang V, Isaac V, McLachlan CS, Kiat H, Herkes G. A Randomized Placebo-Controlled Study of a Transcranial Photobiomodulation Helmet in Parkinson's Di-sease: Post-Hoc Analysis of Motor Outcomes. J Clin Med. 2023 Apr 13;12(8):2846. doi: 10.3390/jcm12082846. PMID: 37109183; PMCID: PMC10146323.

[4] McGee C, Liebert A, Herkes G, Bicknell B, Pang V, McLachlan CS, Kiat H. Protocol for randomized controlled trial to evaluate the safety and feasibility of a novel helmet to deliver transcranial light emitting diodes photobiomodulation therapy to patients with Parkinson's disease. Front Neurosci. 2022 Aug 17;16:945796. doi: 10.3389/fnins.2022.945796. PMID: 36061601; PMCID: PMC9428720.

[5] Hamblin MR. Mechanisms and applications of the anti-inflammatory effects of photobiomodulation. AIMS Biophys. 2017;4(3):337–361. doi: 10.3934/biophy.2017.3.337.

[6] Hamblin MR. Photobiomodulation for Parkinson's disease: mechanisms and therapeutic perspectives. J Neurosci Res. 2021;99(1):8–20. doi: 10.1002/jnr.24705.

[7] Kalia LV, Lang AE. Parkinson's disease. Lancet 2015;386(9996):896–912.

[8] Obeso JA, Stamelou M, Goetz CG, et al. Past, present, and future of Parkinson's disease. Lancet Neurol. 2017;16(8):659–673.

[9] Hamblin MR. Shining light on the head: photobiomodulation for brain disorders. BBA Clin. 2016;6:113–124.

[10] Salehpour F, Rasta SH. The potential of transcranial photobio-modulation therapy for treatment of brain disorders. Neurosci Biobehav Rev. 2020;111:117–134.

[11] Mitrofanis J, Henderson LA, Morries L. Could photobiomodula-tion be used as a non-invasive therapy for Parkinson's disease? Neurophotonics. 2020;7(4):040901.

[12] Hirsch EC, Hunot S. Neuroinflammation in Parkinson's disease: a target for neuroprotection? Lancet Neurol. 2009;8(4):382–397.

[13] Abusrair AH, Elsekaily W, Bohlega S. Tremor in Parkinson's disease: from pathophysiology to advanced therapies. Tremor Other Hyperkinet Mov (N Y). 2022;12:29. doi:10.5334/tohm.712.

Photobiomodulation potential in a neuroglial co-culture model exposed to the SARS-CoV-2 Spike protein.

Erick José Nogueira de Andrade
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, University of Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0009-0007-9504-4856

Geisa Rodrigues Salles
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, University of Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0000-0003-4336-1634

Luiza de Andrade Giraldi
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, University of Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0009-0009-1305-3084

Cristina Pacheco-Soares
Laboratory of Cell and Tissue Biology – Cellular Compartment Dynamics
Institute of Research and Development, University of Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil
ORCID: 0000-0002-0572-074X

***Abstract*—Neurodegenerative disorders have been increasingly associated with the neurological complications observed following Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infection, particularly those related to the neurotoxic effects of the Spike protein. Although photobiomodulation (PBM) has emerged as a promising neuroprotective strategy, its effects in neuroglial co-culture models exposed to the Spike protein remain poorly understood. This study evaluated the neuroprotective potential of PBM in an *in vitro* SH-SY5Y/U87MG neuroglial co-culture challenged with the SARS-CoV-2 Spike protein. Cell metabolic activity was assessed using the Alamar Blue assay at 2 and 24 h after irradiation, while brightfield microscopy and Congo Red staining were used to evaluate the structural organization of the neuroglial network and protein aggregation, respectively. Spike exposure reduced metabolic activity and disrupted the structural organization of the neuroglial network, accompanied by alterations in the distribution of Congo Red-positive material. PBM attenuated the metabolic impairment induced by Spike, preserved neurite-like intercellular connections, and modified the distribution pattern of Congo Red-positive protein aggregates. These findings suggest that PBM exerts neuroprotective effects by preserving cellular metabolism and maintaining the structural integrity of neuroglial networks following Spike-induced injury, supporting its potential as a therapeutic strategy for neuroinflammatory and neurodegenerative conditions.**

Keywords—Neurotoxicity, Oxidative Stress, Mitochondrial Dysfunction, Protein Aggregation, Neuron-glia Interaction, Neuroprotection, Cell Metabolism.

I. INTRODUCTION

Neurodegenerative disorders represent one of the greatest global public health challenges nowadays. These pathologies are characterized by the progressive loss of neuronal functions and, classically, by the accumulation of abnormal protein aggregates in brain tissue, with a particular emphasis on amyloid-β (Aβ) peptides.

Recently, the scientific literature has increasingly investigated the correlation between systemic viral infections and the onset or acceleration of central nervous system (CNS) dysfunctions. In the context of the Coronavirus Disease 2019 (COVID-19) pandemic, SARS-CoV-2 has demonstrated neurotropic potential and has been associated with persistent neurological manifestations. Pioneering studies indicate that the isolated Spike protein of SARS-CoV-2 can compromise cellular homeostasis, induce oxidative stress, and activate inflammatory pathways in brain parenchyma cells, thereby mimicking the early stages of neurodegenerative processes [1]. Furthermore, mitochondrial dysfunction has emerged as one of the earliest events associated with Spike-induced neurotoxicity, resulting in impaired Adenosine Triphosphate (ATP) production, increased Reactive Oxygen Species (ROS) generation, and disruption of neuronal bioenergetics, all of which contribute to the loss of cellular homeostasis and viability [2].

A thorough investigation of these neurotoxic mechanisms and the screening of new therapeutic strategies require *in vitro* models that more faithfully reproduce the brain microenvironment. While monocellular models provide valuable data, they fail to mimic the complex bidirectional communication between neurons and glial cells. In this context, the co-culture of neuron-like and astrocyte-like cells serves as a robust tool for simulating neuroglial interactions. Astrocytes-like cells play a critical role in metabolic support and inflammatory response, directly modulating neuronal viability and vulnerability in the face of toxic insults such as exposure to the Spike protein. Beyond their metabolic role, astrocytes actively regulate neurotransmitter recycling, antioxidant defenses, synaptic homeostasis, and neuronal survival through continuous bidirectional communication with neurons. Consequently, alterations in neuron–glia interactions have been recognized as key events during the

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) (Grants 152384/2024-3, 110869/2025-7, and 404952/2025) and the Coordination for the Improvement of Higher Education Personnel (CAPES) (Grant 88887.271538/2026-00).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

development and progression of several neurodegenerative disorders, making neuroglial co-culture models more physiologically relevant than conventional neuronal monocultures [3-6].

In addition to direct cellular damage, evidence suggests that the Spike protein can interact with proteins prone to misfolding, accelerating fibrillization and deposition of Aβ aggregates [7]. To evaluate the amyloidogenic profile and the formation of these β-sheet-rich structures in the co-culture model, the histological dye Congo Red is widely used, serving as a classical and reliable biomarker for the detection and quantification of these protein aggregates [8]. Moreover, alterations in protein aggregation are frequently accompanied by cytoskeletal remodeling and disruption of the neuroglial network, reinforcing the importance of simultaneously evaluating metabolic activity, cellular organization and amyloidogenic responses following Spike exposure.

Given the neurotoxic potential of the Spike protein, the search for non-invasive and bioenergetic therapies becomes a priority. Photobiomodulation (PBM), using light-emitting diodes (LEDs), stands out for its neuroprotective properties. Acting predominantly in the red to near-infrared range, PBM is absorbed by the mitochondrial chromophore cytochrome c oxidase, optimizing the electron transport chain, increasing ATP synthesis, and modulating ROS [9]. In addition to improving mitochondrial bioenergetics, previous studies have demonstrated that PBM contributes to the preservation of neuronal morphology, maintenance of neurite integrity, modulation of oxidative stress, and regulation of protein homeostasis, highlighting its potential as a therapeutic strategy for neurodegenerative disorders [10-13]. However, the effectiveness of PBM in mitigating stress and the amyloidogenic profile, specifically in co-culture models challenged with the SARS-CoV-2 Spike protein, still needs to be thoroughly elucidated.

Considering the above, the objective of this study is to evaluate the neuroprotective effects of PBM in an *in vitro* SH-SY5Y/U87MG neuroglial co-culture model exposed to the SARS-CoV-2 Spike protein. Specifically, metabolic activity, structural organization of the neuroglial network, and protein aggregation assessed by Congo Red staining were investigated to determine whether PBM attenuates the early cellular alterations induced by Spike exposure.

II. METHODOLOGY

A. Cell Culture and Establishment of Co-culture

The human neuroblastoma (SH-SY5Y) and human glioblastoma (U87MG) cell lines were cultured individually in flasks under controlled incubation conditions (37°C and 5% CO_2). SH-SY5Y cells were maintained in Dulbecco's modified Eagle's medium/nutrient mixture F-12 (DMEM/F12), while U87 cells were maintained in DMEM, both supplemented with 10% Fetal Bovine Serum (FBS) and 1% Antibiotic-Antimycotic (AA).

After reaching appropriate cell confluence, both cell lines were trypsinized, centrifuged, and resuspended in DMEM. For the establishment of co-culture, the cells were seeded in 96-well plates at a final density of $1x10^5$ cells/mL, with a cell ratio of 70:30, consisting of $7x10^4$ cells/mL of SH-SY5Y and $3x10^4$ cells/mL of U87.

B. Cell Differentiation and Experimental Design

After 24 h of initial seeding, the culture medium for all experimental conditions was replaced with a differentiation medium composed of DMEM supplemented with 0.5% FBS, 1% AA, and 0.1% retinoic acid (RA). Cellular differentiation and maturation were maintained for five days.

Following the differentiation period, the cultures were assigned to four experimental groups to evaluate the effects of Spike protein challenge and PBM: Control, maintained in differentiation medium without intervention; Spike, exposed to the recombinant SARS-CoV-2 Spike protein (0.5 µg/mL) for 1 h; LED Control, subjected only to the LED irradiation protocol (660 nm, 3 J/cm²); and LED + Spike, which received PBM in combination with Spike protein exposure.

The recombinant SARS-CoV-2 Spike protein used in this study was the full-length glycoprotein (amino acids 1-1208) produced by the Cell Culture Engineering Laboratory (LECC, COPPE/UFRJ, Rio de Janeiro, Brazil; batch 0151021). The protein was expressed in HEK293 cells cultured in chemically defined medium, purified by affinity chromatography, and characterized by SDS-PAGE and Western blot. The stock solution (0.25 mg/mL) was supplied in phosphate-buffered saline (PBS) containing biotin and 0.02% sodium azide and stored at -80 °C until use. The working concentration of 0.5 µg/mL was selected based on previous studies demonstrating biologically relevant effects of the recombinant Spike protein in central nervous system-derived *in vitro* models [1], [14].

PBM was performed using a custom-built red LED system (660 nm) housed in a dedicated irradiation chamber (Biotable). Before each experiment, the device was calibrated with a power meter to ensure reproducible irradiation conditions. The irradiation parameters consisted of an irradiance of 36.8 mW/cm², a total output power of 115.34 mW, and a fluence of 3 J/cm², delivered over 82 s. Culture plates were positioned at a fixed distance of 4 cm from source to ensure homogeneous light distribution. Temperature monitoring confirmed that thermal variation remained below 0.5 °C throughout irradiation, excluding photothermal effects.

C. Cell Viability Assay

Cellular metabolic activity was quantified 2h and 24 h post-treatment using the Resazurin-based Alamar Blue® assay (Sigma-Aldrich, USA). First, the incubation media were replaced with a working solution composed of differentiation medium and 10% Alamar Blue reagent.

Next, the cocultures were incubated for 2 h under standard conditions (37ºC and 5% CO_2) in the absence of light. Subsequently, 100 µL supernatants from each well were transferred to a 96-well microplate. Then, fluorescence intensity was detected using a microplate spectrophotometer (Excitation: 544 nm; Emission: 590 nm). The same analysis were repeated for the 24 h incubation group. Data were normalized and expressed as a percentage relative to the untreated control group. All experimental conditions were executed in independent triplicates.

D. Cell Fixation and Morphological Analysis

Following the 24 h treatment period, the culture medium was removed, and the cocultures were fixed with 4% paraformaldehyde (PFA, dissolved in phosphate-buffered saline, PBS) for 10 min at room temperature. For the assessment of cellular integrity and neurite outgrowth, digital

photomicrographs were captured using a brightfield and phase-contrast optical microscope.

Concurrently, to evaluate the formation and accumulation of protein aggregates, fixed cells were subjected to a Congo Red binding assay by incubation with a 0.5% Congo Red working solution (w/v, in PBS) for 30 min. After rinsing three times with PBS to eliminate excess dye, the protein deposition was recorded via optical microscopy.

E. Statistical analysis

The assays were performed in two independent biological experiments, with each experimental condition evaluated in triplicate. Data are expressed as mean ± standard deviation (SD). Statistical significance between groups was evaluated using one-way analysis of variance (ANOVA) followed by Dunnett's post hoc test for multiple comparisons between each group and the control, using GraphPad Prism 5.0. Graphs were generated using OriginPro 8.5.

III. Results and Discussion

A. Metabolic response of neuroglial cocultures following Spike exposure and photobiomodulation

Figure 1 shows that exposure of the SH-SY5Y/U87MG neuroglial coculture to the SARS-CoV-2 Spike protein reduced metabolic activity, as measured by the Alamar Blue assay, to approximately 77% and 75% of the control at 2 and 24 h, respectively. In contrast, PBM partially preserved metabolic activity at 2 h and promoted a marked increase after 24 h, with values reaching approximately 130% of the untreated control. PBM alone maintained metabolic activity close to the control after 2 h, whereas a lower metabolic response was observed after 24 h.

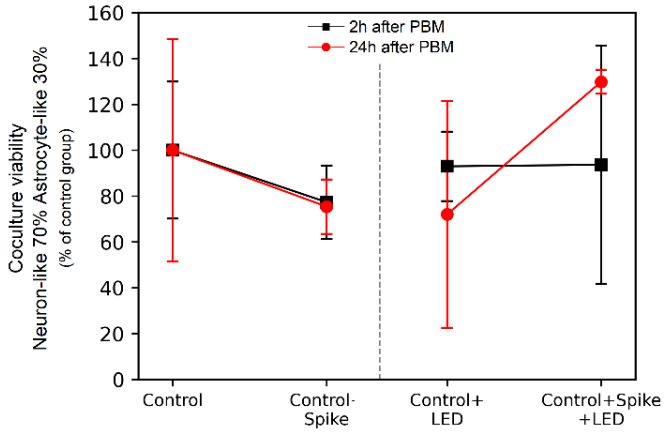


Fig. 1. Effect of photobiomodulation on the metabolic activity of SH-SY5Y/U87MG neuroglial cocultures exposed to the SARS-CoV-2 Spike protein. Metabolic activity was assessed by the Alamar Blue assay at 2 and 24 h after PBM. Data are expressed as a percentage of the untreated control and presented as mean ± SD.

The reduction in metabolic activity induced by the Spike protein is consistent with previous studies demonstrating that this viral protein impairs mitochondrial function, increases oxidative stress, and compromises neuronal metabolism [2]. Since the present model consists of a neuroglial coculture, the observed response likely reflects alterations in neuron–astrocyte interactions. Astrocytes are fundamental for neuronal metabolic support and redox homeostasis, but under pathological conditions they may acquire reactive phenotypes that contribute to neuronal dysfunction through the release of inflammatory mediators and ROS [3, 4].

PBM has been extensively described as a mitochondrial modulator capable of enhancing oxidative phosphorylation, ATP synthesis, and redox signaling, mechanisms that are largely attributed to the activation of cytochrome c oxidase [10]. These mechanisms may explain the recovery observed in the Spike + PBM group, particularly after 24 h, suggesting that PBM promotes a delayed metabolic adaptation following cellular injury.

Furthermore, in a differentiated SH-SY5Y neuronal model, PBM attenuated the deleterious effects induced by the SARS-CoV-2 Spike protein, preserving cell viability and reducing cellular damage [11]. As well, PBM demonstrated neuroprotective effects in a three-dimensional Alzheimer's disease model subjected to oxidative stress, reinforcing its ability to improve cellular responses under neurodegenerative conditions [12]. Together, these findings suggest that the beneficial effects of PBM are maintained in more physiologically relevant neuroglial coculture models.

Although the Spike + PBM group presented metabolic activity above the control after 24 h, this result should be interpreted cautiously. Alamar Blue primarily measures cellular reducing capacity rather than absolute cell number, indicating that the increased signal may reflect enhanced mitochondrial metabolism, increased proliferation, or an adaptive metabolic response rather than exclusively improved cell survival [15].

Overall, these findings suggest that PBM mitigates the metabolic impairment induced by the SARS-CoV-2 Spike protein in a neuroglial coculture model, supporting its potential as a therapeutic strategy for neuroinflammatory conditions associated with mitochondrial dysfunction.

B. Evaluation of protein aggregation by Congo Red staining

Congo Red staining performed 24 h after irradiation revealed distinct distribution patterns of Congo Red-positive material among the experimental groups (Figure 2). The untreated control exhibited predominantly intracellular and perinuclear staining, with preserved cellular morphology and elongated neurite-like processes. In contrast, qualitative inspection of the Spike-exposed co-cultures revealed discrete Congo Red-positive regions distributed along cellular processes and in pericellular areas (indicated by green arrows), accompanie by an apparent loss of neurite-like continuity.

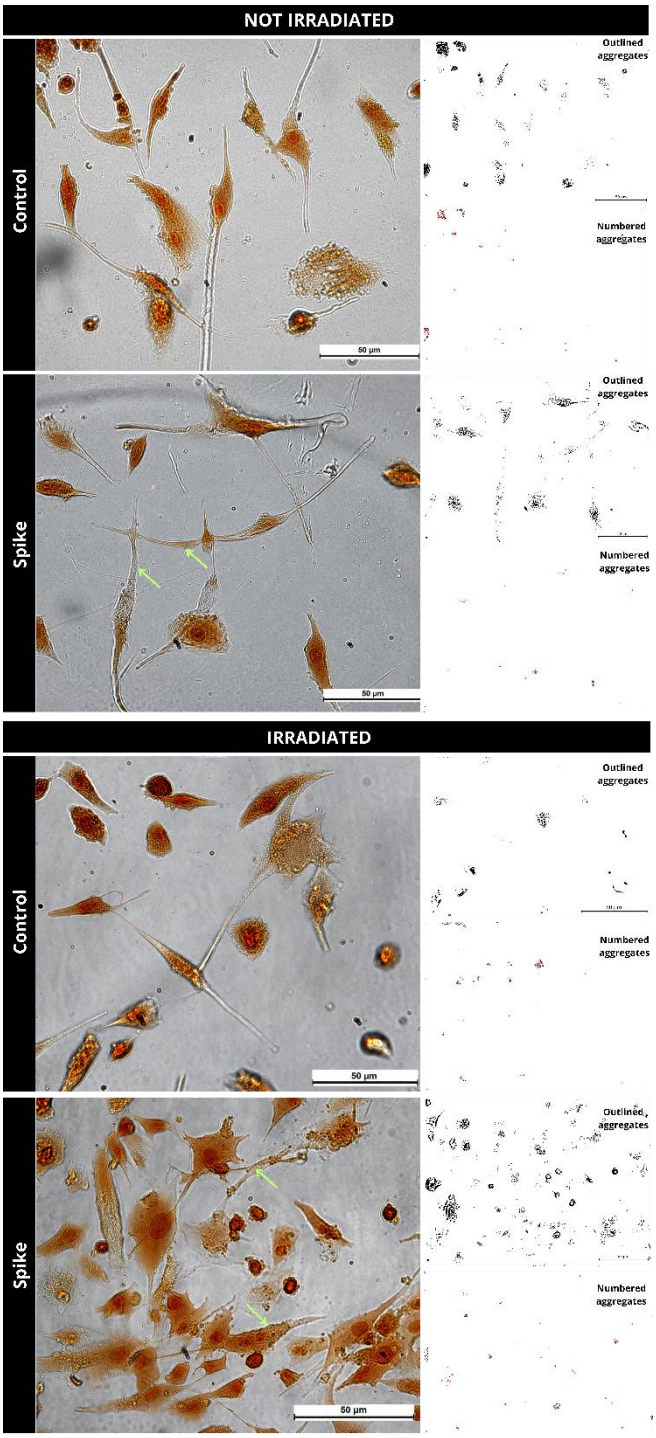


Fig. 2. Representative Congo Red staining of SH-SY5Y/U87MG neuroglial cocultures exposed to the SARS-CoV-2 Spike protein and treated with PBM. Representative brightfield images were acquired 24 h after PBM. Green arrows indicate representative Congo Red-positive regions, which are qualitatively compatible with the presence of protein aggregates. Scale bars = 50 µm

The PBM-treated control displayed Congo Red staining mainly associated with the cell bodies, without evident accumulation along cellular processes. In the Spike + PBM group, Congo Red-positive material remained detectable; however, the staining was predominantly localized around the cell bodies and in discrete pericellular regions, unlike the Spike group, which exhibited more pronounced deposits along neurite-like processes.

To quantitatively complement the citological findings, morphometric analysis of Congo Red-positive deposits was performed across the experimental groups (Figure 3). Evaluation of the total number of aggregates revealed no statistically significant differences among the groups (Figure 3A), despite a numerical trend toward increased aggregate counts in the Spike + PBM condition ($109.8 \pm \mathrm{SD}$; $p = 0.2025$ vs. Spike). Similarly, assessment of the total aggregate area per field demonstrated comparable coverage across all conditions (Figure 3B; $p = 0.6660$ for Control vs. PBM; $p = 0.9958$ for Control vs. Spike; and $p = 0.5724$ for Spike vs. Spike + PBM). Furthermore, analysis of individual aggregate dimensions indicated that average particle size remained statistically unaltered among groups (Figure 3C; $p > 0.05$).

Taken together, these morphometric analyses indicate that neither Spike exposure nor PBM treatment significantly altered the gross accumulation, total area coverage, or size profile of Congo Red-positive material at 24 h post-irradiation. Instead, the differences observed microscopically (Figure 2) reflect a spatial redistribution—shifting aggregates away from vulnerable neurite-like extensions toward somatic and pericellular compartments in the Spike + PBM group-rather than an immediate bulk clearance of protein deposits.

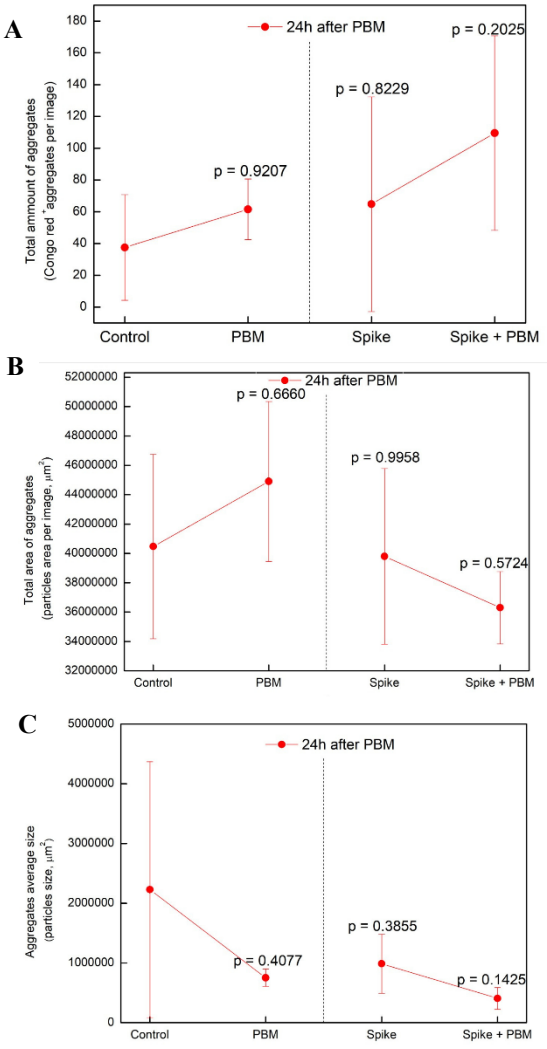


Fig. 3. Morphometric quantification of Congo Red-positive aggregates in SH-SY5Y/U87MG neuroglial cocultures. Cocultures exposed to the SARS-CoV-2 Spike protein and treated with photobiomodulation (PBM) were evaluated 24 h post-irradiation. Quantitative particle analyses include: (A) total number of Congo Red-positive aggregates per image field; (B) total aggregate area per image (µm²); and (C) average aggregate size (µm²).

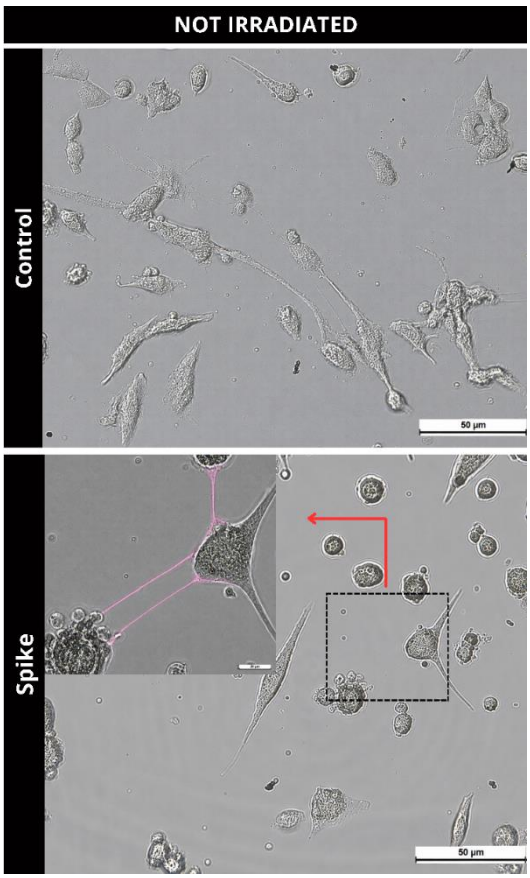
The SARS-CoV-2 Spike protein contains amyloidogenic segments capable of self-assembling into β-sheet-rich fibrillar aggregates following proteolytic cleavage [16]. Furthermore, Spike-derived fragments have been shown to interact with amyloid- β , accelerating aggregation and enhancing neurotoxicity in experimental models [17]. Consistent with these mechanisms, the presence of Congo Red-positive material in Spike-treated cultures is compatible with amyloid-like aggregation.

Consistent with the present findings, our previous work demonstrated that PBM protects neuronal cells against the deleterious effects of the SARS-CoV-2 Spike protein [11] and mitigates oxidative stress in a three-dimensional neurodegenerative model [12]. These findings are consistent with literature demonstrating that PBM improves mitochondrial function, cellular redox balance and protein homeostasis [10], mechanisms that may be associated with the cellular response to protein aggregation.

Importantly, while quantitative assessment revealed that PBM did not significantly alter the overall number of Congo Red-positive aggregates at 24 h post-irradiation (Figure 3), microscopic examination revealed a clear spatial redistribution: aggregates shifted from vulnerable neurite-like processes in the Spike group toward somatic and pericellular compartments in the Spike + PBM group. This suggests that the protective effects of PBM in this model may be related to cellular compartmentalization and preservation of neurite integrity rather than an immediate clearance of amyloid aggregates at this early time point.

C. Structural organization of the neuroglial network

Brightfield microscopy performed 2 h after PBM revealed distinct differences in the structural organization of the neuroglial network among the experimental groups (Figure 4).



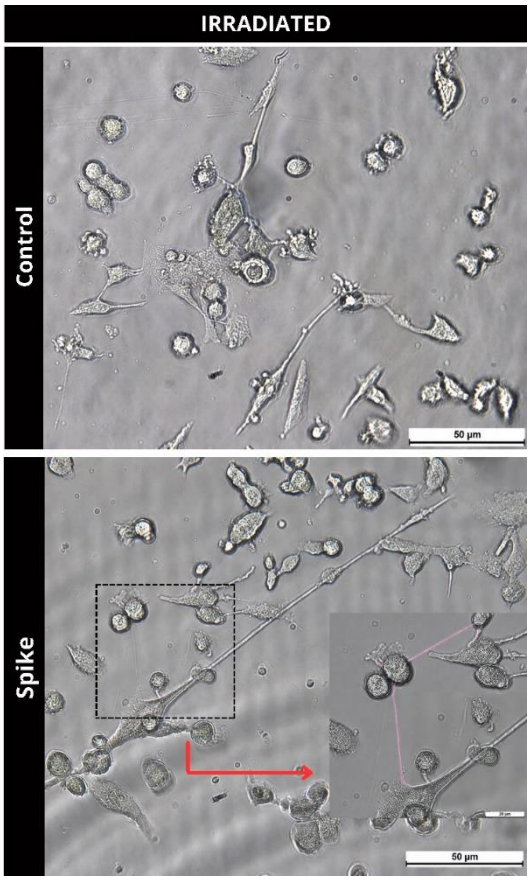


Fig. 4. Representative brightfield images showing the structural organization of SH-SY5Y/U87MG neuroglial cocultures exposed to the SARS-CoV-2 Spike protein and treated with photobiomodulation. Images were acquired 2 h after PBM. Enlarged regions highlight representative neurite-like processes establishing physical contacts between neighboring cells. Scale bars = 50 µm

In the control group, the coculture exhibited an organized cellular architecture composed of elongated neuron-like cells interconnected by thin neurite-like processes, together with cells displaying astrocyte-like morphology. Representative enlarged regions highlighted numerous fine cellular projections establishing physical contacts between neighboring cells, Suggesting the presence of neuron–glia-like cellular interactions within the SH-SY5Y/U87MG coculture.

These observations are consistent with previous studies demonstrating that neuroglial coculture models promote the formation of complex cellular networks through reciprocal interactions between neurons and astrocytes. Astrocytes provide structural and metabolic support, regulate extracellular homeostasis, and contribute to neurite stabilization, resulting in a more physiologically relevant in vitro microenvironment than neuronal monocultures [5, 6].

Exposure to the SARS-CoV-2 Spike protein Qualitative inspection of the brightfield images revealed apparent alterations in the structural organization of the neuroglial network in the Spike-treated group. The Spike-treated group qualitatively exhibited greater cell rounding, an apparent reduction in neurite-like continuity, and fewer visible intercellular contacts compared with the untreated control compared with the untreated control. These findings are consistent with previous reports demonstrating that the SARS-CoV-2 Spike protein promotes cytoskeletal remodeling, mitochondrial dysfunction, and oxidative stress, ultimately impairing neuronal morphology and cellular homeostasis [2].

Interestingly, Qualitative inspection suggested that the PBM-treated coculture retained a structural organization broadly comparable to that of the untreated control, preserving elongated neurite-like processes and extensive intercellular contacts. Likewise, although morphological alterations remained evident in the Spike + PBM group, representative enlarged regions demonstrated the persistence of thin neurite-like processes establishing physical contacts between neighboring cells. The apparent maintenance of these structural connections is consistent with a possible role of PBM in preserving aspects of neuroglial network organization during the early response to Spike exposure of the neuroglial network during the early response to Spike-induced injury.

The maintenance of neuron–glia interactions is essential for neuronal survival and tissue homeostasis. Astrocytes actively regulate synaptic function, neurotransmitter recycling, metabolic support and antioxidant defenses, while neurons provide reciprocal signals that modulate astrocytic activity [3-5]. Therefore, disruption of this structural organization may compromise the coordinated cellular responses required for maintaining neural tissue integrity. Conversely, preservation of the neuroglial network, even in the presence of cellular injury, may represent an important component of the neuroprotective effects associated with PBM.

These observations are supported by previous evidence showing that PBM preserves mitochondrial bioenergetics, reduces oxidative stress, and contributes to cytoskeletal maintenance, thereby favoring neuronal morphology and neurite integrity [10, 13]. Collectively, these mechanisms may explain the greater preservation of the neuroglial network architecture observed in the Spike + PBM group.

Overall, qualitative morphological observations suggest that exposure to the SARS-CoV-2 Spike protein is associated with alterations in the apparent structural organization of SH-SY5Y/U87MG neuroglial cocultures, whereas PBM appears to partially preserve aspects of the cellular network organization and visible intercellular contacts during the early period following Spike exposure , at least partially, the physical intercellular connections that characterize the neuroglial network during the early stages following Spike-induced.

IV. Conclusion

Overall, the present findings indicate that exposure to the SARS-CoV-2 Spike protein is associated with metabolic impairment and qualitative alterations in the structural organization and distribution of Congo Red-positive material in SH-SY5Y/U87MG neuroglial co-cultures. PBM partially attenuated the metabolic impairment induced by Spike exposure and was associated with the preservation of apparent neurite-like processes and intercellular contacts, particularly during the early response. Although the morphological and Congo Red findings were evaluated qualitatively in the present study, quantitative analysis of the Congo Red-positive area will be performed to further characterize the observed staining pattern.

Future studies should also incorporate quantitative assessment of morphological parameters, including the number and length of cellular processes, connection density, and other relevant structural features, together with appropriate statistical analyses. These complementary approaches may provide a more comprehensive understanding of the effects of Spike exposure and the potential neuroprotective mechanisms of PBM in neuroglial co-culture models.

Acknowledgment

The authors gratefully acknowledge Prof. Dr. Leandro José Raniero and his research team for providing access to the microplate reader and for their technical support. The authors also thank Prof. Dr. Juliana Ferreira Strixino and her research team for granting access to the LED irradiation chamber.

References

[1] T. P. Buzhdygan et al., "The SARS-CoV-2 spike protein alters barrier function in 2D static and 3D microfluidic in-vitro models of the human blood–brain barrier," Neurobiol. Dis., vol. 146, p. 105131, Dec. 2020, doi: 10.1016/j.nbd.2020.105131.

[2] C. A. Pileggi et al., "The SARS-CoV-2 spike glycoprotein interacts with MAO-B and impairs mitochondrial energetics," Current Research in Neurobiology, vol. 5, p. 100112, 2023, doi: 10.1016/j.crneur.2023.100112.

[3] D. Pathak and K. Sriram, "Neuron-astrocyte omnidirectional signaling in neurological health and disease," Front. Mol. Neurosci., vol. 16, Jun. 2023, doi: 10.3389/fnmol.2023.1169320.

[4] R. Patani, G. E. Hardingham, and S. A. Liddelow, "Functional roles of reactive astrocytes in neuroinflammation and neurodegeneration," Nat. Rev. Neurol., vol. 19, no. 7, pp. 395–409, Jul. 2023, doi: 10.1038/s41582-023-00822-1.

[5] S. Park, H.-N. Oh, and W.-K. Kim, "Human coculture model of astrocytes and SH-SY5Y cells to test the neurotoxicity of chemicals," Ecotoxicol. Environ. Saf., vol. 269, p. 115912, Jan. 2024, doi: 10.1016/j.ecoenv.2023.115912.

[6] J. Kim, S. Yang, and I. S. Choi, "Neutralization of Cannabidiol Neurotoxicity in Neuron-Astrocyte Sandwich Coculture," Adv. Biol., vol. 7, no. 10, Oct. 2023, doi: 10.1002/adbi.202300090.

[7] D. Idrees and V. Kumar, "SARS-CoV-2 spike protein interactions with amyloidogenic proteins: Potential clues to neurodegeneration," Biochem. Biophys. Res. Commun., vol. 554, pp. 94–98, May 2021, doi: 10.1016/j.bbrc.2021.03.100.

[8] E. I. Yakupova, L. G. Bobyleva, I. M. Vikhlyantsev, and A. G. Bobylev, "Congo Red and amyloids: history and relationship," Biosci. Rep., vol. 39, no. 1, Jan. 2019, doi: 10.1042/BSR20181415.

[9] M. R. Hamblin, "Shining light on the head: Photobiomodulation for brain disorders," BBA Clin., vol. 6, pp. 113–124, Dec. 2016, doi: 10.1016/j.bbacli.2016.09.002.

[10] M. R Hamblin, "Mechanisms and applications of the anti-inflammatory effects of photobiomodulation," AIMS Biophys., vol. 4, no. 3, pp. 337–361, 2017, doi: 10.3934/biophy.2017.3.337.

[11] E. J. N. de Andrade, G. R. Salles, L. de Andrade Giraldi, M. Y. R. Sudo, L. de Paula Ramos, and C. Pacheco-Soares, "From neurotoxicity to neuroprotection: photobiomodulation against the effects of the SARS-CoV-2 spike protein in a 2D neurodegenerative model using differentiated SH-SY5Y cells," In vitro models, May 2026, doi: 10.1007/s44164-026-00129-7.

[12] L. de Andrade Giraldi, G. Rodrigues Salles, E. J. Nogueira de Andrade, and C. Pacheco-Soares, "Neurotherapeutic potential of photobiomodulation in a 3D Model of Alzheimer's disease induced by oxidative stress," In vitro models, May 2026, doi: 10.1007/s44164-026-00122-0.

[13] F. dos S. Cardoso, F. Salehpour, N. C. Coimbra, F. Gonzalez-Lima, and S. Gomes da Silva, "Photobiomodulation for the treatment of neuroinflammation: A systematic review of controlled laboratory animal studies," Front. Neurosci., vol. 16, Sep. 2022, doi: 10.3389/fnins.2022.1006031.

[14] E. Clough et al., "Mitochondrial Dynamics in SARS-COV2 Spike Protein Treated Human Microglia: Implications for Neuro-COVID," Journal of Neuroimmune Pharmacology, vol. 16, no. 4, pp. 770–784, Dec. 2021, doi: 10.1007/s11481-021-10015-6.

[15] S. N. Rampersad, "Multiple Applications of Alamar Blue as an Indicator of Metabolic Function and Cellular Health in Cell Viability Bioassays," Sensors, vol. 12, no. 9, pp. 12347–12360, Sep. 2012, doi: 10.3390/s120912347.

[16] S. Nyström and P. Hammarström, “Amyloidogenesis of SARS-CoV-2 Spike Protein,” J. Am. Chem. Soc., vol. 144, no. 20, pp. 8945–8950, May 2022, doi: 10.1021/jacs.2c03925.

[17] S. Cao et al., “Spike Protein Fragments Promote Alzheimer’s Amyloidogenesis,” ACS Appl. Mater. Interfaces, vol. 15, no. 34, pp. 40317–40329, Aug. 2023, doi: 10.1021/acsami.3c09815.

Proposal for Robotic Gait Training with a Motor Imagery-Based Brain–Computer Interface and Transcutaneous Spinal Cord Stimulation

1st Willyenne Gomes de Lima
Postgraduate Program in Neuroengineering
Edmond and Lily Safra International Institute of Neurosciences
Macaíba, Brazil
willyenne.lima@edu.isd.org.br

2nd Nathaly Vitória Rodrigues Basílio
Department of Electrical Engineering
Federal University of Rio Grande do Norte
Natal, Brazil
nathaly.basilio@edu.isd.org.br

3rd Rafael Augusto de Oliveira Guedes
Santos Dumont Institute
Macaíba, Brazil
rafael.guedes@isd.org.br

4th Caroline Cunha do Espírito Santo
Santos Dumont Institute
Macaíba, Brazil
caroline.santo@isd.org.br

5th Denis Delisle Rodríguez
Santos Dumont Institute
Macaíba, Brazil
denis.rodriguez@isd.org.br

***Abstract*—Spinal cord injury can result in persistent motor, sensory, and autonomic deficits, and functional recovery remains limited, particularly in individuals with chronic complete injuries. Strategies combining different levels of the sensorimotor system have been investigated as alternatives to increase neural engagement during rehabilitation. This study presents and discusses the neuroanatomical, neurophysiological, and experimental foundations supporting the integration of motor imagery, brain-computer interfaces, transcutaneous spinal direct current stimulation, and robotic locomotor training. Available evidence indicates that individuals with complete spinal cord injury may retain cortical activity related to motor imagery, that this activity can be identified and used by brain-computer interfaces during locomotor training, and that transcutaneous spinal direct current stimulation can modify measures related to spinal excitability. Studies also indicate the feasibility of combining spinal neuromodulation with locomotor training. However, this evidence is mainly derived from investigations of the individual components or partial combinations and does not demonstrate that temporally conditioning stimulation to cortical activity produces effects superior to stimulation applied independently of cortical state. Therefore, the synchronization of motor intention, spinal modulation, and locomotor afferent input represents a neurophysiologically plausible hypothesis whose contribution to plasticity and functional recovery after spinal cord injury still requires specific experimental evaluation.**

***Index Terms*—transcutaneous spinal cord stimulation; motor imagery; brain–computer interface; spinal cord injury; paraplegia; robotic gait; neurorehabilitation.**

I. Introduction

Spinal cord injury (SCI) is associated with motor, sensory, and autonomic impairments that may persist throughout life and impose important limitations on functional independence. Global Burden of Disease data indicate that the global age-standardized incidence was estimated at 7.1 cases per 100,000 population and prevalence at 183.6 cases per 100,000 population, with heterogeneous distribution across regions, age groups, and sexes [1]. In developing countries, traumatic SCI occurs predominantly in young men, with motor vehicle accidents and falls being its main etiologies [2]. In Brazil, Faleiros et al. observed a predominance of traumatic injuries in a national sample, with traffic accidents accounting for 40.8% of cases, followed by firearm or knife injuries, at 17.5%, and falls, at 10.7% [3]. Locomotor rehabilitation after SCI remains challenging, particularly in the presence of severe neurological impairments. Robotic technologies have been employed to provide repetitive and standardized training of the lower limbs, with the possibility of controlling assistance, body-weight support, and gait speed. However, systematic reviews indicate that the benefits of robotic training vary according to the population, outcome, and protocol employed, and the available evidence does not support consistent superiority of robotic training across all functional domains assessed [4], [5]. These findings have encouraged the investigation of strategies capable of combining locomotor training with interventions acting at different levels of the nervous system. Among these strategies are brain-computer interfaces (BCIs), which use neural signals to identify states related to movement intention or representation and convert them into information capable of controlling external systems and/or devices. Electroencephalography (EEG) is widely used in this context because it enables noninvasive acquisition with high temporal resolution. During motor imagery (MI), changes in sensorimotor rhythms can be used as features for classification, although the performance of these systems depends on the protocol employed, signal quality, and interindividual variability [6]. In individuals with complete SCI, MI remains capable of recruiting cortical sensorimotor regions, providing neurophysiological support for the use of this activity as an input signal for BCIs even in the absence of observable voluntary movement [7].

Studies conducted specifically in the context of locomo-

tor training also demonstrate the possibility of integrating MI and cortical activity with the use of the Lokomat. Serafini et al. investigated visual neurofeedback associated with gait imagery during passive Lokomat training in individuals with complete SCI, observing modulation of sensorimotor rhythms and electrophysiological changes throughout training [8]. Subsequently, Blanco-Diaz et al. developed a multichannel EEG-based BCI capable of discriminating passive walking without MI from passive walking associated with gait imagery, achieving a mean classification accuracy of 74.4% in individuals with motor-complete SCI [9]. These results demonstrate the feasibility of accessing and classifying cortical states associated with gait imagery during robotic training, although they do not, in isolation, constitute evidence of the clinical efficacy of the BCI. In parallel, noninvasive spinal neuromodulation techniques have been investigated as a means of modifying the functional state of circuits located below the injury. Among them, transcutaneous spinal direct current stimulation (tsDCS) applies low-intensity direct current for neuromodulatory purposes. Studies conducted in individuals with SCI have demonstrated changes in measures of spinal excitability after tsDCS, whereas studies combining the technique with locomotor training have indicated the feasibility of repeated application, although there is still insufficient evidence to establish consistent functional benefits [10], [11].

The possibility of relating voluntary cortical activity, spinal modulation, and sensorimotor afference produced by locomotor training therefore represents a relevant question for neuroengineering applied to SCI. Although these components have been investigated individually or in different combinations, it remains uncertain whether establishing a temporal association between the cortical state related to motor intention and spinal neuromodulation produces effects different from those observed when stimulation is applied independently of the cortical state. In this context, this study presents and discusses the neuroanatomical, neurophysiological, and experimental foundations supporting the integration of motor imagery-based BCI, tsDCS, and robotic locomotor training, considering the available evidence, its limitations, and the questions that remain open for investigation.

II. Literature Review

1) Neuroanatomical and neurophysiological bases of sensorimotor integration: Motor imagery corresponds to the internal representation of an action without its physical execution and can recruit components of the cortical network involved in movement planning and control. In the kinesthetic modality, this representation emphasizes the sensations associated with task execution, including proprioceptive and kinesthetic aspects. In individuals with complete SCI, neuroimaging studies demonstrate that MI remains capable of activating sensorimotor regions. Wang et al. observed, in individuals with complete SCI, greater activation of sensorimotor areas during MI, associated with changes in effective connectivity between sensorimotor regions and areas related to cognitive processing [7]. These findings indicate that components of the cortical networks related to movement representation remain functionally recruitable after injury, although their organization may differ from that observed in neurologically intact individuals. The clinical classification of an SCI as complete should not be interpreted as evidence of absolute anatomical interruption of all connections across the injury site. Sherwood et al. identified, through neurophysiological assessments, residual brain influence below the level of injury in individuals clinically classified as having complete SCI, a condition termed "discomplete SCI" by the authors [12]. The high frequency observed in that study should not be generalized to the entire population with complete SCI, considering the specific characteristics of the sample investigated. Nevertheless, the findings demonstrate that the absence of clinically detectable voluntary movement and the absolute absence of descending influence are not equivalent concepts. The spinal cord, in turn, is not merely a passive pathway for the transmission of supraspinal commands. Spinal interneuronal circuits integrate descending and afferent information and participate in the organization of motor activity. Particularly during locomotion, sensory information from muscles, tendons, joints, and skin continuously modifies the activity of these networks. Experimental evidence highlights proprioception as a relevant component of motor circuit reorganization after SCI, influencing the plasticity of spinal networks and movement recovery [13]. However, an important part of this mechanistic knowledge derives from experimental models, and caution should therefore be exercised when extrapolating it directly to individuals with complete SCI. Sensory afferents produced by movement also participate in reorganization at supraspinal levels. Moreno-López and Hollis highlight that SCI can induce remodeling of sensory circuits at different levels of the nervous system and that the interaction between somatosensory and motor processing is relevant to the reorganization associated with movement recovery [14]. Thus, peripheral feedback produced during locomotor training should be understood as an active component of sensorimotor dynamics rather than merely as a mechanical consequence of movement.

In this context, tsDCS represents a means of modifying the excitability state of spinal circuits without necessarily directly producing action potentials. Hubli et al. demonstrated that the application of anodal tsDCS over the spinal region in individuals with motor-complete SCI increased spinal reflex amplitude and reduced its threshold, whereas equivalent effects were not observed under the other conditions evaluated [10]. The authors themselves emphasized that the mechanisms responsible for the effect remained partially unknown. Thus, these results support the ability of tsDCS to modify measures related to spinal circuit excitability but do not allow its effects to be considered uniformly excitatory. This caution is reinforced by subsequent studies. Powell et al. investigated the effects of different tsDCS polarities on corticospinal excitability in individuals with chronic incomplete SCI and did not observe consistent significant differences in motor-evoked potential amplitude between conditions [15]. More recently, Sayed Ahmad and Knikou observed muscle-

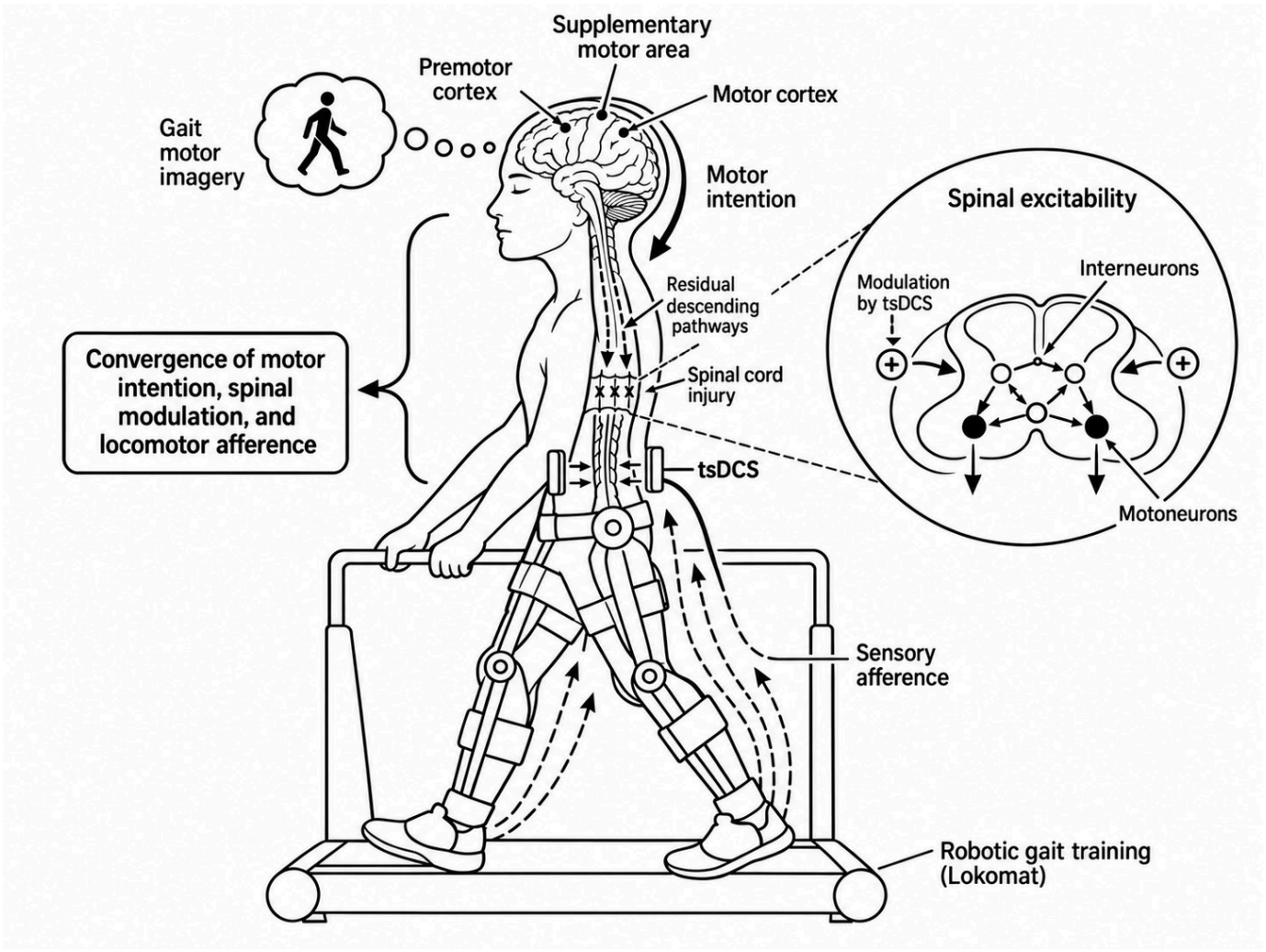


Fig. 1. Neuroanatomical and neurophysiological bases of sensorimotor integration. Gait motor imagery recruits cortical areas involved in movement planning and representation, while possible residual descending pathways maintain influence over spinal circuits below the injury. Transcutaneous spinal direct current stimulation modulates the excitability of spinal networks, and robotic gait training provides task-related sensory afferent input. The convergence of motor intention, spinal modulation, and sensorimotor feedback constitutes the physiological basis for investigating activity-dependent neuromodulation strategies.

and circuit-dependent responses after repeated applications of transspinal direct current stimulation in individuals with SCI [16]. Taken together, these findings suggest that polarity, electrode placement, neurological characteristics, and the neurophysiological variable analyzed may influence the response to stimulation, making it inappropriate to attribute a simple relationship between polarity and functional effect to tsDCS. The integration of these mechanisms provides a physiological basis for investigating interventions that bring voluntary cortical activity, spinal modulation, and sensory afference into closer temporal association. MI enables the recruitment of cortical networks related to movement representation; tsDCS can alter the excitability state of spinal networks; and locomotor training produces repetitive task-related afferent input. These components act at different but functionally interdependent levels of the sensorimotor system [7], [10], [13], [14]. The possibility of bringing them into temporal proximity constitutes a physiologically plausible hypothesis of activity-dependent modulation, but this plausibility should not be confused with demonstration of the clinical superiority of the combination.

2) neurofeedback and brain-computer interfaces in motor rehabilitation: Neurofeedback (NFB) and MI-based BCIs use cortical activity in complementary ways. In NFB, information derived from neural activity is presented to the participant in order to enable training in the self-regulation of specific electrophysiological patterns. In BCIs, neural signals are processed to identify different states and produce a control output. In EEG- and MI-based systems, changes in sensorimotor rhythms, particularly in the mu and beta bands, can contribute to discrimination between states related to the motor task [6], [9] . The performance of these interfaces varies among participants, making training and calibration relevant components. Zhou et al. investigated NFB based on relative alpha-band power in neurologically intact individuals and observed an increase in this power associated with an overall improvement

in the performance of an MI-based BCI [17]. However, the effects showed interindividual variability, and not all comparisons related to classification performance showed significant improvement. These results demonstrate that the ability to self-regulate cortical rhythms can be trained, but they also reinforce that the effects of NFB are not uniform across individuals. In people with SCI, the possibility of using cortical activity associated with MI is supported by both neuroimaging studies and BCI applications. Preservation of sensorimotor area recruitment during MI [7] provides the neurophysiological basis for this type of interface. In the specific context of gait, Blanco-Diaz et al. used multichannel EEG recordings to classify passive walking without MI and passive walking associated with gait imagery during Lokomat use. The BCI employed participant-specific calibration and achieved a mean accuracy of 74.4% [9]. The study also demonstrated changes in features related to class separability and cortical self-regulation over the sessions. These results should be interpreted primarily as evidence of the technical feasibility of classification and NFB, rather than as proof of functional recovery resulting from the BCI. Serafini et al. also investigated gait MI and NFB during robotic locomotor training in individuals with complete SCI. In a study involving two participants who underwent 12 sessions, changes in sensorimotor rhythm power, increases in measures of brain connectivity, and changes in clinical assessments were observed after the intervention [8]. However, the small sample size, the absence of a parallel control group in the published study, and the multimodal nature of the intervention prevent the clinical findings from being attributed specifically to NFB or MI. The study is particularly relevant because it demonstrates that this paradigm can be implemented in individuals with complete SCI during Lokomat training and provides preliminary data on cortical modulation in this context. Cortical activity classified by the BCI can also be used as a signal to condition an external intervention. Kumari et al. evaluated an EEG-based BCI associated with functional electrical stimulation before physical practice of the upper limbs in people with SCI. The mean BCI accuracy was approximately 85%, demonstrating the feasibility of cortical activity-dependent control [18]. Nevertheless, no significant differences were identified between groups in the main physical outcomes, and the authors emphasized limitations related to sample size. Therefore, the study supports the feasibility of temporally conditioning an external intervention to cortical activity but does not allow the conclusion that such conditioning is, by itself, sufficient to produce greater clinical recovery. When the output of a BCI is used to directly control another intervention, classifier performance assumes additional importance. A classification error no longer affects only the quality of neurofeedback and may determine whether an intervention occurs or fails to occur at an inappropriate moment. In this scenario, individual calibration, discriminative ability, false activations, and the temporal relationship between identification of the neural state and device response become relevant parameters for characterizing the control architecture. The study by Blanco-Diaz et al. provides an important technical basis by demonstrating individual calibration and classification of gait imagery [9], but the use of this signal as a controller of a neuromodulatory intervention represents a different purpose from that originally evaluated.

III. Integration of cortical activity, spinal neuromodulation, and robotic locomotor training

The combination of BCI, tsDCS, and robotic locomotor training enables the integration of cortical, spinal, and peripheral components of the sensorimotor system. In this arrangement, cortical activity related to MI can be used not only to provide feedback to the participant but also as a variable to determine the timing of an external intervention. tsDCS may act on the excitability state of spinal circuits, while locomotor training provides task-related afferent input. Thus, the association does not imply replacement of corticospinal conduction but establishes a functional relationship between the detection of a cortical state and an intervention applied at the spinal level. The concomitant application of tsDCS and locomotor training has already been investigated independently of BCI. Hawkins et al. evaluated the safety and feasibility of this combination in individuals with chronic incomplete SCI. Following an initial single-session study, the authors conducted a 16-session intervention comparing tsDCS combined with locomotor training with sham stimulation combined with the same training. The combination was considered safe and feasible, and gait outcomes varied among participants; three of the four participants in the active group showed improvements in speed above the minimal clinically important difference, compared with one of the four participants in the sham group, whereas endurance gains above this threshold occurred in one participant in each group [11]. Given the small sample size, the authors themselves considered further investigation of the efficacy of the approach necessary. Thus, the conceptual difference between continuous tsDCS application and its application conditioned by the BCI lies mainly in dependence on the cortical state. In the first situation, stimulation occurs independently of activity related to motor intention. In the second, detection of a cortical pattern associated with MI can determine when spinal modulation is applied. The hypothesis of interest is therefore whether the functional state of the nervous system at the time of stimulation influences the response produced by the combination of neuromodulation and locomotor practice. The available literature provides support for the components of this integration, but at different levels of evidence. Activity related to gait imagery can be detected during Lokomat training [9]; individuals with SCI can use cortical activity to control external interventions [18]; tsDCS can modify measures of spinal excitability [10], [15], [16] and the combination of tsDCS and locomotor training has been shown to be feasible in individuals with incomplete SCI [11] . However, none of this evidence demonstrates that temporally conditioning tsDCS to MI during robotic locomotor training produces effects superior to those obtained when stimulation

is applied independently of cortical state. Investigation of this difference therefore remains an open experimental question.

IV. Discussion

The evidence analyzed supports the possibility of accessing different components of the sensorimotor system after SCI, but it also demonstrates that these components are at different levels of scientific maturity. Cortical activity associated with MI may persist after clinically complete injuries [7], and studies by Serafini and Blanco-Diaz demonstrate that this activity can be incorporated into robotic locomotor training paradigms using EEG and NFB [8], [9]. At the same time, studies using tsDCS indicate an ability to modify spinal neurophysiological measures, although the direction of the effects varies across experiments and depends on methodological and individual characteristics [10], [15], [16]. Therefore, the integration of these technologies has an experimental basis but cannot be presented as an intervention whose efficacy has already been established. The studies by Serafini et al. and Blanco-Diaz et al. are particularly relevant because they were conducted in the same functional context considered in this line of investigation: gait imagery during passive Lokomat training in people with severe SCI [8], [9]. Serafini et al. demonstrated that participants with complete SCI were able to perform NFB training related to mu and beta rhythms during assisted movement, whereas Blanco-Diaz et al. demonstrated that passive walking states with and without MI can be discriminated by a multichannel BCI, with a mean accuracy of 74.4% [8], [9]. Together, these studies support the feasibility of using brain activity related to gait imagery during robotic training. However, their results do not demonstrate that using this classification to control tsDCS produces an additional benefit, since conditioned spinal stimulation was not part of these protocols. Another relevant source of evidence comes from the randomized pilot trial by Nicolelis et al. [19]. The study included eight men with chronic traumatic SCI classified as AIS A, allocated to either assisted locomotor training alone or locomotor training combined with BMI, virtual reality, and tactile feedback. Both groups showed sensory and motor changes throughout the protocol, whereas the group receiving the multimodal intervention showed greater mean changes in measures such as Lower Extremity Motor Score, light touch, and pinprick [19]. Three of the four participants in the multimodal group and one of the four participants receiving locomotor training alone were classified as AIS C at the end of the protocol. However, the authors emphasized the small sample size and other limitations, including the absence of blinding of participants and caregivers. This study is important to the present discussion, but its interpretation requires caution. The experimental group in Nicolelis et al. simultaneously received BMI, visuohaptic feedback, and virtual reality, in addition to locomotor training, making it impossible to attribute the results specifically to the BMI. Furthermore, the BMI and locomotor training sessions did not constitute an architecture in which cortical classification temporally controlled spinal stimulation. The study provides evidence that a multimodal intervention incorporating cortical participation may be associated with greater neurological recovery than locomotor training alone in that pilot trial, but it does not test the specific hypothesis of synchronization between motor intention and tsDCS [19]. In the field of spinal neuromodulation, the available results also warrant a moderate interpretation. Hubli et al. demonstrated modulation of spinal reflexes after anodal tsDCS in individuals with motor-complete SCI [10]. Powell et al. found less consistent results when evaluating corticospinal excitability in individuals with incomplete SCI [15], and Sayed Ahmad and Knikou subsequently demonstrated that responses to transspinal direct current stimulation can vary according to muscles, circuits, and neurological characteristics [16]. Hawkins et al., in turn, provided evidence of safety and feasibility for repeated application of tsDCS during locomotor training, but with a sample insufficient to establish clinical efficacy [11]. These studies converge in demonstrating the ability of stimulation to modify the neurophysiological state of the spinal system, but they do not support a uniform response nor do they allow, based solely on the current literature, the definition of a universally superior configuration. This heterogeneity reinforces the relevance of separating two potentially distinct effects: the effect related to the application of tsDCS itself during training and a possible additional effect associated with its temporal conditioning to cortical activity. If a conditioned intervention is studied without an active nonconditioned condition, differences relative to sham would not allow determination of whether the outcome results from synchronization with motor intention or simply from the presence of spinal stimulation. Consequently, the conceptual comparison between cortical state-dependent and cortical state-independent stimulation is important for specifically investigating the contribution of temporal coupling, without assuming in advance that such coupling is superior. The plausibility of this hypothesis derives from the interdependent organization of the sensorimotor system. MI provides a cortical state related to the voluntary representation of movement; residual descending signals may remain present in some individuals clinically classified as complete [12]; tsDCS may modify the responsiveness of spinal circuits [10] ; and assisted movement provides task-related sensory afferents [13], [14]. However, the simultaneous presence of these mechanisms does not demonstrate that their synchronization necessarily produces functionally relevant plasticity. The temporal relationship between stimuli constitutes, in this context, a mechanistic hypothesis to be tested rather than an established effect. In addition to physiological uncertainties, there are important technical limitations. The performance of EEG-based BCIs varies among individuals, and an overall accuracy above chance does not fully describe the reliability required to control another intervention. In a conditioned architecture, false positives may trigger stimulation outside the intended cortical state, whereas false negatives may prevent its application when MI is present. For this reason, classifier performance metrics, calibration stability, and temporal characterization of the system become relevant components

in evaluating the integration. Blanco-Diaz et al. demonstrated that individual calibration and online classification are feasible during robotic gait [9], but transforming this output into a neuromodulation trigger introduces requirements that were not evaluated in that study. It is also necessary to consider that neurophysiological changes do not necessarily correspond to clinical recovery. Changes in EEG rhythms, connectivity, spinal excitability, or muscle recruitment may provide information about the response of the nervous system before or even in the absence of detectable changes in clinical scales. This distinction is particularly relevant in individuals with chronic complete SCI, in whom voluntary motor gains may be limited. Thus, future studies on the integration of BCI, tsDCS, and locomotor training should jointly interpret technical, electrophysiological, and clinical measures, avoiding the use of biomarker changes as automatic surrogates for functional recovery. Taken together, the literature analyzed allows a sequence of complementary evidence to be established: motor imagery can recruit cortical networks after complete SCI [7]; its activity can be detected and trained during assisted gait [8], [9]; cortical signals can control external interventions [18]; tsDCS can modify measures related to spinal excitability [10], [15], [16]; and its association with locomotor training is technically feasible [11] . The element that has not yet been demonstrated is whether temporally linking these components produces additional effects compared with stimulation applied independently of cortical state. This gap more precisely delineates the scientific contribution of the approach and avoids interpreting results from partial or multimodal interventions as direct evidence of the efficacy of a conditioned system.

V. Conclusion

The available evidence supports the neurophysiological plausibility of integrating MI-based BCI, tsDCS, and robotic locomotor training after SCI. These resources act on complementary components of the sensorimotor system: MI enables the recruitment of cortical networks related to movement representation, the BCI allows this activity to be identified and used as a control variable, tsDCS may modify the excitability state of spinal circuits, and locomotor training provides task-related sensory afferents. Previous studies demonstrate that people with complete SCI can exhibit cortical activity related to motor imagery and use NFB and BCI systems during locomotor training, whereas tsDCS studies provide evidence of neurophysiological modulation and the feasibility of combining it with locomotor training [8]–[11]. Multimodal interventions involving BMI and locomotor training have also shown relevant neurological outcomes in pilot studies [19]. However, none of this evidence directly demonstrates that temporally conditioning tsDCS to cortical activity produces effects superior to stimulation applied independently of cortical state. Thus, the main question arising from the literature is not whether each of these components can act on the sensorimotor system, but whether the temporal relationship between cortical motor intention, spinal modulation, and locomotor afference modifies the response produced by their combination. Experimental evaluation of this hypothesis, accompanied by objective measures of BCI performance, synchronization, safety, neurophysiological responses, and clinical outcomes, may contribute to clarifying the role of activity-dependent conditioning in future neuromodulation strategies after SCI.

References

[1] Y. Lu, Z. Shang, W. Zhang, X. Hu, R. Shen, K. Zhang, Y. Zhang, L. Zhang, B. Liu, M. Pang, and L. Rong, "Global, regional, and national burden of spinal cord injury from 1990 to 2021 and projections for 2050: A systematic analysis for the Global Burden of Disease 2021 study," *Ageing Res. Rev.*, vol. 103, p. 102598, 2025.

[2] A. Golestani, P. Shobeiri, M. Sadeghi-Naini, S. B. Jazayeri, S. F. Maroufi, Z. Ghodsi, M. A. Dabbagh Ohadi, E. Mohammadi, V. Rahimi-Movaghar, and S. M. Ghodsi, "Epidemiology of traumatic spinal cord injury in developing countries from 2009 to 2020: A systematic review and meta-analysis," *Neuroepidemiology*, vol. 56, no. 4, pp. 219–239, 2022.

[3] F. Faleiros, M. Marcossi, O. Ribeiro, A. Tholl, G. Freitas, and M. Riberto, "Epidemiological profile of spinal cord injury in Brazil," *J. Spinal Cord Med.*, vol. 46, no. 1, pp. 75–82, 2023.

[4] G. La Rosa, M. Avola, T. Di Gregorio, R. S. Calabrò, and M. P. Onesta, "Gait recovery in spinal cord injury: A systematic review with metanalysis involving new rehabilitative technologies," *Brain Sci.*, vol. 13, no. 5, p. 703, 2023.

[5] A. R. Alashram, G. Annino, and E. Padua, "Robot-assisted gait training in individuals with spinal cord injury: A systematic review for the clinical effectiveness of Lokomat," *J. Clin. Neurosci.*, vol. 91, pp. 260–269, 2021.

[6] T. Majoros and S. Oniga, "Overview of the EEG-based classification of motor imagery activities using machine learning methods and inference acceleration with FPGA-based cards," *Electronics*, vol. 11, no. 15, p. 2293, 2022.

[7] L. Wang, X. Li, W. Zheng, X. Chen, Q. Chen, Y. Hu, L. Cao, J. Ren, W. Qin, J. Lu, and N. Chen, "Motor imagery evokes strengthened activation in sensorimotor areas and its effective connectivity related to cognitive regions in patients with complete spinal cord injury," *Brain Imaging Behav.*, vol. 16, no. 5, pp. 2049–2060, 2022.

[8] E. R. S. Serafini, C. D. Guerrero-Mendez, T. F. Bastos-Filho, A. Cotrina-Atencio, A. F. O. de Azevedo Dantas, D. Delisle-Rodriguez, and C. C. do Espírito-Santo, "Gait training-based motor imagery and EEG neurofeedback in Lokomat: A clinical intervention with complete spinal cord injury individuals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 1896–1905, 2024.

[9] C. F. Blanco-Diaz, E. R. da Silva Serafini, T. Bastos-Filho, A. F. O. de Azevedo Dantas, C. C. do Espirito Santo, and D. Delisle-Rodriguez, "A gait imagery-based brain–computer interface with visual feedback for spinal cord injury rehabilitation on Lokomat," *IEEE Trans. Biomed. Eng.*, vol. 72, no. 1, pp. 102–111, 2025.

[10] M. Hubli, V. Dietz, M. Schrafl-Altermatt, and M. Bolliger, "Modulation of spinal neuronal excitability by spinal direct currents and locomotion after spinal cord injury," *Clin. Neurophysiol.*, vol. 124, no. 6, pp. 1187–1195, 2013.

[11] K. A. Hawkins, L. A. DeMark, A. Vistamehr, H. J. Snyder, C. Conroy, C. Wauneka, G. Tonuzi, D. D. Fuller, D. J. Clark, and E. J. Fox, "Feasibility of transcutaneous spinal direct current stimulation combined with locomotor training after spinal cord injury," *Spinal Cord*, vol. 60, no. 11, pp. 971–977, 2022.

[12] A. M. Sherwood, M. R. Dimitrijevic, and W. B. McKay, "Evidence of subclinical brain influence in clinically complete spinal cord injury: Discomplete SCI," *J. Neurol. Sci.*, vol. 110, pp. 90–98, 1992.

[13] A. Takeoka, "Proprioception: Bottom-up directive for motor recovery after spinal cord injury," *Neurosci. Res.*, vol. 154, pp. 1–8, 2020.

[14] Y. Moreno-López and E. R. Hollis, "Sensory circuit remodeling and movement recovery after spinal cord injury," *Front. Neurosci.*, vol. 15, p. 787690, 2021.

[15] E. S. Powell, C. Carrico, E. Salyers, P. M. Westgate, and L. Sawaki, "The effect of transcutaneous spinal direct current stimulation on corticospinal excitability in chronic incomplete spinal cord injury," *NeuroRehabilitation*, vol. 43, pp. 125–134, 2018.

[16] A. M. Sayed Ahmad and M. Knikou, "Transspinal direct current stimulation as targeted therapy to increase motor neuron output and restore inhibition in human spinal cord injury," *Front. Neurol.*, vol. 17, p. 1773679, 2026.

[17] Q. Zhou, R. Cheng, L. Yao, X. Ye, and K. Xu, "Neurofeedback training of alpha relative power improves the performance of motor imagery brain-computer interface," *Front. Hum. Neurosci.*, vol. 16, p. 831995, 2022.

[18] R. Kumari, A. Dybus, M. Purcell, and A. Vučković, "Motor priming to enhance the effect of physical therapy in people with spinal cord injury," *J. Spinal Cord Med.*, vol. 48, no. 2, pp. 312–326, 2025.

[19] M. A. L. Nicolelis, E. J. L. Alho, A. R. C. Donati, S. Yonamine, M. A. Aratanha, G. Bao, D. S. F. Campos, S. Almeida, D. Fischer, and S. Shokur, "Training with noninvasive brain–machine interface, tactile feedback, and locomotion to enhance neurological recovery in individuals with complete paraplegia: A randomized pilot study," *Sci. Rep.*, vol. 12, p. 20545, 2022.

Reaction Time During a Manual Dexterity Task in Parkinson's Disease: A Pilot Study

Alice de Oliveira Barreto Suassuna
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-1304-737X

Valton Costa
Physical Therapy Department
Federal University of São Carlos
São Carlos, Brazil
ORCID: 0000-0002-2356-7523

Lavínia Samuele Araújo Menezes
Faculty of Biomedical Engineering
Federal University of Rio Grande do Norte
Natal, Brazil
ORCID: 0009-0003-7733-528X

Lucas Tomaz de Moura
Faculty of Biomedical Engineering
Federal University of Rio Grande do Norte
Natal, Brazil
ORCID: 0009-0000-3369-9530

Edgard Morya
Edmond and Lily Safra International
Neuroscience Institute
Santos Dumont Institute
Macaíba, Brazil
ORCID: 0000-0003-0954-5317

Adriano de Oliveira Andrade
Faculty of Electrical Engineering
Federal University of Uberlândia
Uberlândia, Brazil
ORCID: 0000-0002-5689-6606

***Abstract*— Parkinson's disease (PD) is a neurodegenerative disease of the Central Nervous System (CNS) characterized by motor symptoms. Bradykinesia is one of the most disabling manifestations of PD, generally defined as slowed movement execution; however, it is also associated with reduced spontaneous movements, decreased range of motion, and increased reaction time (RT). RT is a quantitative physiological marker widely used as an indicator of motor function in individuals with PD. This pilot study aimed to analyze RT in patients with PD in their ON and OFF medication states and in healthy controls. RT was calculated while participants performed a gross manual dexterity task with five interruption blocks. An inertial sensor with an accelerometer was placed on the dorsum of participants' hands to detect movement onset after auditory stimulation. Non-parametric tests were used given the small sample sizes, with Benjamini-Hochberg correction for multiple comparisons. Results indicate that RT was significantly longer in both PD groups compared to controls at Block 1. The PD-OFF group showed additional significant differences in Blocks 4 and 5, suggesting that without dopaminergic support, RT impairment persists under higher cumulative motor demand. The PD-ON group converged toward control performance across subsequent blocks, consistent with a preserved short-term adaptive effect. No significant ON vs OFF differences were found after correction.**

Keywords — Parkinson's disease, Bradykinesia, Reaction Time, Manual Dexterity, Inertial Sensor.

I. Introduction

Parkinson's disease (PD), first described by the physician James Parkinson (1755–1824) in his seminal "An Essay on the Shaking Palsy" [1] in 1817, is a chronic neurodegenerative disease of the Central Nervous System (CNS) that specifically affects the basal ganglia. It involves the degeneration of dopaminergic neurons in the substantia nigra, leading to decreased dopamine levels in the striatum and abnormal motor control [2]. PD is characterized by primary motor symptoms: tremor (involuntary oscillatory movements), bradykinesia (slowness and hesitation of movement), rigidity (increased muscle tone) and postural instability [3].

Bradykinesia is one of the most disabling manifestations of PD, broadly defined as slowness of movement execution [4]. It is also associated with reduced spontaneous movements, decreased range of motion, and increased Reaction Time (RT) [5, 6]. Evidence suggests that hypoactivation of the primary motor cortex (M1) or premotor areas may underlie the difficulties in preparing and initiating voluntary movements in individuals with PD [7–9]. Studies have shown this hypoactivation leads to a delay in neural activation onset, manifesting as increased premotor RT, a longer time to reach peak velocity, and prolongation and variability in movement duration [10].

Reaction time is a quantitative physiological marker widely used as an indicator of motor function in PD [11]. This variable reflects the central neural processing preceding movement execution, being prolonged in complex tasks and reduced in the presence of preparatory signals. Difficulty initiating volitional movements is a hallmark of PD [12]. In complex multi-joint upper limb movements such as grasping or picking up objects, bradykinesia becomes even more prominent [13, 14].

To quantitatively characterize motor symptoms in PD, several studies have used wearable technologies such as gyroscopes, accelerometers, magnetometers, electromyography, force sensors, and smartphone applications [15]. A previously developed protocol using inertial sensors investigated the role of movement termination in difficulty initiating movement and bradykinesia in PD during complex manual dexterity tasks [16].

Although RT is a well-established marker of motor dysfunction in PD, most protocols assess it in single, brief trials, offering limited information on how RT behaves under sustained or repeated motor demand, and whether dopaminergic medication modulates this pattern differently over time. To address this gap, the present pilot study tested the hypothesis that the RT impairment in PD would be more pronounced and sustained in the OFF state of medication, using the inertial-sensor protocol described in [16] during a manual dexterity task in patients with PD (ON and OFF medication states) and healthy controls, employing a repeated-block protocol, thereby contributing to a more quantitative characterization of motor function in PD.

II. Materials and Methods

A. Participants

Participants were divided into three groups: PD under dopaminergic medication (PD-ON, n = 7), PD without medication (PD-OFF, n = 7), and healthy controls group (CG, n = 7). Patients diagnosed with PD without history of bronchoaspiration or musculoskeletal sequelae in the upper limbs were included. Controls comprised individuals without upper limb musculoskeletal sequelae. Individuals with cognitive impairments, inability to understand instructions, or severe hearing impairments were excluded from both groups. One Control Group (CG) participant was subsequently excluded due to a hearing impairment identified during the experimental session that directly interfered with the auditory protocol, yielding a final CG of n = 6.

This study was conducted in accordance with the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Instituto de Ensino e Pesquisa Alberto Santos Dumont (CAAE 61005322.3.0000.0129). All participants provided written informed consent.

B. Inertial Sensor

Upper limb movement was monitored using an accelerometer (acquisition rate: 148 Hz) from the Trigno™ wireless system (Delsys Inc., USA). A sensor was placed on the dorsum of the participant's most-affected hand during the manual dexterity task (Fig. 1).

C. Experimental Protocol

Participants performed the Box and Blocks Test (BBT), in which they transferred 2.5 cm wooden cubes between two compartments of a box using their more-affected hand as quickly as possible. The task requires coordinated multi-joint movements and demands gross motor skills.

The protocol used two auditory stimuli, chosen to allow unrestricted movement. A high-frequency tone (2000 Hz) signaled task initiation; a low-frequency tone (1000 Hz) signaled interruption. Five blocks were performed in randomized order: Block 1 (30 s execution / 60 s pause), Block 2 (55 s / 55 s), Block 3 (60 s / 30 s), Block 4 (45 s / 15 s), and Block 5 (40 s / 10 s). Randomization prevented anticipation of the stimulus.

D. Data Analysis

Data were collected using EMGWorks software (Delsys) and processed in R. To ensure precise temporal alignment between the auditory stimulus and the accelerometer signal, the synchronization unit was connected to an electromyography (EMG) electrode channel. The onset of each auditory stimulus produced a detectable electrical artifact in this channel, providing a hardware-level timestamp of stimulus delivery independent of software latency. Accelerometer signals (x, y, z axes) were filtered using a band-pass filter (0.5–10 Hz). RT was computed as the interval between this stimulus-onset timestamp and movement onset, the latter identified from the vector magnitude of the three axes as the point exceeding the baseline mean plus two standard deviations.

E. Statistical Analysis

All estimated parameters were averaged for each participant for subsequent statistical analyses. All statistical analyses were performed using R. Given the small sample sizes (n = 6–7 per group), non-parametric tests were adopted regardless of normality screening. The Mann-Whitney U test was used for independent group comparisons (CG vs PD-ON and CG vs PD-OFF), and the Wilcoxon signed-rank test was applied for the paired comparison between medication states (PD-ON vs PD-OFF), as the same patients were assessed under both conditions.

To control for Type I error inflation across the five blocks, p-values were adjusted using the Benjamini-Hochberg (BH) false discovery rate procedure. Effect sizes were calculated as r = |Z| / N, and interpreted as small (r < 0.3), moderate (0.3 r < 0.5), or large (r 0.5). Given the small sample sizes, 95% confidence intervals (CIs) for the group medians were estimated using a non-parametric bootstrap (10,000 resamples with replacement, percentile method) and are reported alongside the median and IQR in Table 2. A significance level of 5% was used for all analyses.

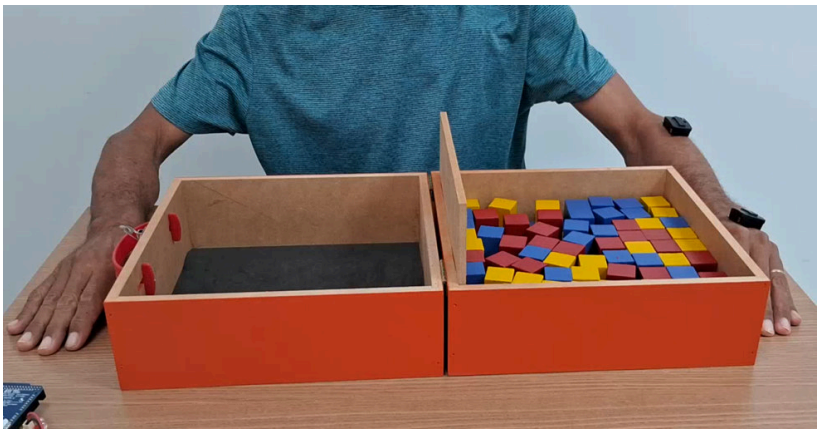


Fig. 1. Participant from the PD group with the inertial sensor placed on the dorsum of the hand to perform the Box and Blocks Test (BBT).

III. Results

A. Demographic and clinical characteristics

Table 1 presents participant characteristics. The PD group was predominantly male (F = 2, M = 5), with normal cognitive status Mini-Mental State Examination (MMSE) (27 ± 1.73) and mild motor impairment in both ON and OFF states (Hoehn & Yahr (H&Y) Stage 2). One CG participant was excluded due to hearing impairment, yielding CG n = 6.

TABLE 1. Demographic and clinical characteristics of the participants.

	PD-ON (n=7)	PD-OFF (n=7)	CG (n=6)
Age	60.29 ± 6.95	60.29 ± 6.95	60 ± 5.39
Female/Male	2/5	2/5	2/4
Time Disease (year)	8.57 ± 6.95	8.57 ± 6.95	-
H&Y Scale	2	2	-
MMSE Score	27 ± 1.73	27 ± 1.73	-

B. Reaction Time in BBT

Table 2 presents detailed data on the median reaction times of the groups across blocks. In Block 1, the first time the stimulus was presented requiring participants to initiate the task, reaction time was significantly longer in both PD-ON and PD-OFF groups compared to healthy controls (p = 0.027 for both comparisons), with large effect sizes (r = 0.792). In the subsequent blocks, the PD-ON group no longer differed significantly from controls, suggesting a progressive reduction in the RT gap over repeated stimulus presentations.

In contrast, the PD-OFF group continued to show significantly longer RT compared to controls in Blocks 4 and 5 (p = 0.031 for both, r = 0.674), and a trend toward significance in Block 2 (p = 0.067, r = 0.555). This pattern indicates that without dopaminergic medication, the impairment in movement initiation persists and is more sustained across the task. No significant differences were found between the PD-ON and PD-OFF states in any block after BH correction, although Block 5 showed a trend (p = 0.035, p_BH = 0.173, r = 0.799), suggesting a possible late-task effect of medication on RT that warrants investigation in larger samples. The graph in Figure 2 shows the reaction times of the participants in all groups across the five blocks of the protocol during the BBT task.

As expected for a small sample, the bootstrap CIs for the group medians were wide, particularly in later blocks, visually reinforcing the limited precision of the estimates and reinforcing the need for larger samples in future confirmatory studies.

IV. Discussion

This study aimed to analyze reaction time in PD patients under the influence of dopaminergic medication, without the effect of medication, and healthy patients during the performance of a gross manual dexterity task, Box and Block Test, which requires the execution of more complex multi-joint movements.

TABLE 2. Median (Scale e interquartile range - IQR) [95% bootstrap CI] of reaction time in seconds for each group across the five blocks of the experimental protocol, with pairwise comparisons and Benjamini-Hochberg correction. (* p_BH < 0.05; # p_BH < 0.10; ns = non-significant).

Block	PD-ON Median (IQR) [95% CI]	PD-OFF Median (IQR) [95% CI]	CG Median (IQR) [95% CI]	CG vs ON (p_BH)	CG vs OFF (p_BH)	ON vs OFF (p_BH)
1	0.730 (0.242) [0.567, 0.931]	0.806 (0.341) [0.709, 1.166]	0.318 (0.135) [0.240, 0.432]	0.027*	0.027*	0.342ns
2	0.754 (0.385) [0.516, 0.932]	0.948 (0.231) [0.772, 1.062]	0.531 (0.134) [0.420, 0.791]	0.432ns	0.067#	0.554ns
3	0.806 (0.184) [0.664, 0.880]	0.678 (0.539) [0.515, 1.353]	0.528 (0.123) [0.365, 0.695]	0.185ns	0.134ns	0.554ns
4	0.541 (0.478) [0.481, 1.379]	0.878 (0.403) [0.540, 1.151]	0.445 (0.303) [0.320, 0.725]	0.281ns	0.031*	0.342ns
5	0.902 (0.286) [0.535, 1.036]	1.012 (0.256) [0.553, 1.034]	0.489 (0.270) [0.378, 0.872]	0.223ns	0.031*	0.173ns

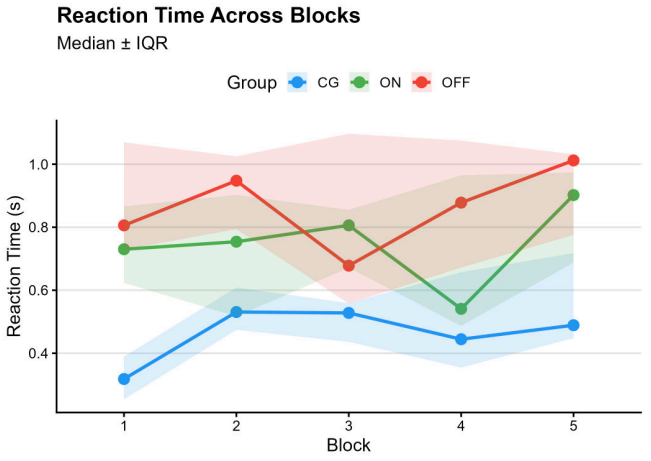


Fig. 2. Reaction times in second of the participants across the five blocks.

Our results showed that reaction times in patients with Parkinson's disease are longer than those in participants without the disease, both under the influence of dopamine medication and not.

These findings corroborate the literature, as studies indicate that bradykinesia and difficulty initiating movements are more prominent without the effect of dopaminergic activation, especially in novel motor tasks [5]. Furthermore, another important factor to note is that the BBT task requires rapid grasping, visual-motor coordination, and transportation, factors that depend on efficient activation of the motor cortex and the integrity of dopaminergic circuits [17]. According to Lorek (2024) [18], impaired manual dexterity in PD is associated with decreased dopamine in the striatum.

A notable finding of the present analysis is the divergent pattern between PD-ON and PD-OFF across blocks. While PD-ON patients showed significantly longer RT only in Block 1, converging toward control performance in subsequent blocks, PD-OFF patients exhibited sustained and significant differences in Blocks 4 and 5 as well. In the absence of dopaminergic support, this increased cumulative motor demand may have been sufficient to unmask and sustain the RT impairment, whereas medicated patients retained the capacity to maintain performance. This interpretation aligns with evidence that dopaminergic medication is particularly effective in attenuating bradykinesia under conditions of greater motor load [5].

It is important to note that this convergence of PD-ON patients toward control performance occurred within a single testing session; while it may reflect short-term adaptive processes, it should not be interpreted as motor learning in the strict sense, which requires evidence of retention beyond the testing session. The cross-sectional, single-session design of this pilot study does not allow this distinction, and future studies incorporating retention tests are needed to confirm this hypothesis.

This interpretation is further supported by the concept of motor fatigability in PD, an objectively measurable decline in motor performance during repeated or sustained tasks, distinct from the subjective perception of fatigue [19]. Its pathophysiology is multifactorial, involving disrupted dopaminergic signaling across fronto-striatal circuits, and dopamine depletion has been directly linked to fatigability in PD [20]. Without dopaminergic support, the basal ganglia may therefore be less capable of compensating for fatigue-related decrements in motor drive, resulting in the sustained RT impairment observed in PD-OFF across the later blocks. The absence of a statistically significant direct comparison between PD-ON and PD-OFF, despite large effect sizes ($r = 0.799$ at Block 5), likely reflects the limited statistical power of the paired analysis with $n = 7$, and should be interpreted with caution.

According to Trevarrow (2024) [21], antiparkinsonian medications reduce the impaired effects of PD by improving motor function; however, medication does not improve all aspects of movement equally, being more effective in improving movement speed than in improving the range of motion in tasks involving visual guidance. Together, these findings highlight the importance of multimodal approaches combining pharmacological and rehabilitative strategies.

Future studies should expand the sample size to increase statistical power, particularly for paired ON vs OFF comparisons, which showed large effect sizes but did not reach significance likely due to limited power, and control for disease duration, medication dose, and time since last administration. Test-retest sessions would help establish the stability of RT measures over time. Incorporating direct kinematic parameters such as movement duration, amplitude, and peak velocity alongside RT, would further strengthen the characterization of motor function in PD; these analyses are currently being conducted on the same dataset as part of a broader doctoral research program and will be reported in future publications.

V. Conclusion

The results of this study indicate that reaction times in patients with Parkinson's disease are longer than those in participants without the disease, both under the influence of dopamine medication and not. This difference was statistically significant at Block 1 for both medication states, and persisted across Blocks 4 and 5 exclusively in the PD-OFF group, highlighting the partial but meaningful role of dopaminergic medication in attenuating RT impairments. Over the course of the task, PD-ON patients showed a convergence toward control-level performance within the testing session, a pattern that may reflect short-term adaptive processes rather than motor learning in the strict sense, since the cross-sectional, single-session design of this pilot study does not allow assessment of long-term retention; PD-OFF patients did not exhibit the same convergent pattern.

Limitations include the small sample size, which may have limited the statistical power to detect significant differences between PD-ON and PD-OFF states, particularly in later blocks where large effect sizes were observed; the lack of control over medication dose and time since last administration; the absence of test-retest sessions; and the absence of independent validation (e.g., video or electromyography) of the accelerometer-based movement-onset detection threshold. Future studies should expand the sample, address these methodological points, and investigate range of motion, movement duration, and velocity alongside reaction time.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The present work was carried out with the support of the National Council for Scientific and Technological Development (444437/2024-0, 442150/2023-7, and 405365/2023-3), Coordination for the Improvement of Higher Education Personnel (CAPES), Foundation for Research Support of the State of Minas Gerais (FAPEMIG APQ-05015-24). AOA (302942/2022-0) is fellow of CNPq.

References

[1] J. Parkinson, "An essay on the shaking palsy," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 14, no. 2, pp. 223–236, 2002.

[2] S. J. Chia, E.-K. Tan, and Y.-X. Chao, "Historical perspective: models of parkinson's disease," *International journal of molecular sciences*, vol. 21, no. 7, p. 2464, 2020.

[3] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of neurology, neurosurgery & psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[4] D. Berg, C. H. Adler, B. R. Bloem, P. Chan, T. Gasser, C. G. Goetz, G. Halliday, A. E. Lang, S. Lewis, Y. Li, *et al.*, "Movement disorder society criteria for clinically established early parkinson's disease," *Movement Disorders*, vol. 33, no. 10, pp. 1643–1646, 2018.

[5] A. Berardelli, J. C. Rothwell, P. D. Thompson, and M. Hallett, "Pathophysiology of bradykinesia in parkinson's disease," *Brain*, vol. 124, no. 11, pp. 2131–2146, 2001.

[6] J. Gauntlett-Gilbert and V. J. Brown, "Reaction time deficits and parkinson's disease," *Neuroscience & Biobehavioral Reviews*, vol. 22, no. 6, pp. 865–881, 1998.

[7] B. Haslinger, P. Erhard, N. Kämpfe, H. Boecker, E. Rummeny, M. Schwaiger, B. Conrad, and e. A. Ceballos-Baumann, "Event-related functional magnetic resonance imaging in parkinson's disease before and after levodopa," *Brain*, vol. 124, no. 3, pp. 558–570, 2001.

[8] U. Sabatini, K. Boulanouar, N. Fabre, F. Martin, C. Carel, C. Colonnese, L. Bozzao, I. Berry, J. Montastruc, F. Chollet, *et al.*, "Cortical motor reorganization in akinetic patients with parkinson's disease: a functional mri study," *Brain*, vol. 123, no. 2, pp. 394–403, 2000.

[9] K.-M. Lee, K.-H. Chang, and J.-K. Roh, "Subregions within the supplementary motor area activated at different stages of movement preparation and execution," *Neuroimage*, vol. 9, no. 1, pp. 117–123, 1999.

[10] V. Cutsuridis and S. Perantonis, "A neural network model of parkinson's disease bradykinesia," *Neural Networks*, vol. 19, no. 4, pp. 354–374, 2006.

[11] P. H. Gordon, Q. Yu, C. Qualls, H. Winfield, S. Dillon, P. E. Greene, S. Fahn, R. E. Breeze, C. R. Freed, and S. L. Pullman, "Reaction time and movement time after embryonic cell implantation in parkinson disease," *Archives of neurology*, vol. 61, no. 6, pp. 858–861, 2004.

[12] T. Warabi, K. Fukushima, P. M. Olley, S. Chiba, and N. Yanagisawa, "Difficulty in terminating the preceding movement/posture explains the impaired initiation of new movements in parkinson's disease," *Neuroscience letters*, vol. 496, no. 2, pp. 84–89, 2011.

[13] B. G. Farley, S. Sherman, and G. F. Koshland, "Shoulder muscle activity in parkinson's disease during multijoint arm movements across a range of speeds," *Experimental brain research*, vol. 154, no. 2, pp. 160–175, 2004.

[14] M. Bologna, G. Paparella, A. Fasano, M. Hallett, and A. Berardelli, "Evolving concepts on bradykinesia," *Brain*, vol. 143, no. 3, pp. 727–750, 2020.

[15] I. Vanmechelen, H. Haberfehlner, J. De Vleeschhauwer, E. Van Wonterghem, H. Feys, K. Desloovere, J.-M. Aerts, and E. Monbaliu, "Assessment of movement disorders using wearable sensors during upper limb tasks: A scoping review," *Frontiers in Robotics and AI*, vol. 9, p. 1068413, 2023.

[16] A. Suassuna, L. Cavalcanti, C. Silva, E. Morya, and A. Andrade, "Protocol for bradykinesia characterization through the relationship between the movement initiation and the termination of the preceding movement.," in *Anais do XIV Simpósio de Engenharia Biomédica (ISSN: 2358-3568).*, Zenodo, Dec. 2022.

[17] M. O. Klein, D. S. Battagello, A. R. Cardoso, D. N. Hauser, J. C. Bittencourt, and R. G. Correa, "Dopamine: functions, signaling, and association with neurological diseases," *Cellular and molecular neurobiology*, vol. 39, no. 1, pp. 31–59, 2019.

[18] K. Lorek, J. Mączewska, L. Królicki, M. Chalimoniuk, K. Markowska, S. Budrewicz, M. Koszewicz, Ł. Szumowski, and J. Marusiak, "Motor cortex activation mediates associations between striatal dopamine depletion and manual dexterity in parkinson's disease," *Parkinsonism & Related Disorders*, vol. 125, p. 107049, 2024.

[19] L. Bojtos, J. Rodríguez-Antigüedad, J. Pagonabarraga, S. Martínez-Horta, and J. Kulisevsky, "Fatigue and neuropsychiatric symptoms in parkinson's disease: a narrative review," *Frontiers in Neurology*, vol. 16, p. 1670644, 2025.

[20] D. d. L. Scheffer, F. C. Freitas, A. S. Aguiar Jr, C. Ward, L. G. A. Guglielmo, R. D. Prediger, S. J. Cronin, R. Walz, N. A. Andrews, and A. Latini, "Impaired dopamine metabolism is linked to fatigability in mice and fatigue in parkinson's disease patients," *Brain Communications*, vol. 3, no. 3, p. fcab116, 2021.

[21] M. P. Trevarrow, M. J. Munoz, Y. M. Rivera, R. Arora, Q. H. Drane, G. D. Pal, L. Verhagen Metman, L. C. Goelz, D. M. Corcos, and F. J. David, "Medication improves velocity, reaction time, and movement time but not amplitude or error during memory-guided reaching in parkinson's disease," *Physiological reports*, vol. 12, no. 17, p. e16150, 2024.

Carta Resposta aos Revisores

Manuscrito: Reaction Time During a Manual Dexterity Task in Parkinson's Disease: A Pilot Study.

Os autores agradecem aos revisores pelos comentários detalhados e construtivos, que contribuíram significativamente para a clareza e o rigor do manuscrito. Abaixo, respondemos ponto a ponto a cada comentário.

Revisor 1

Comentário 1.1

"A avaliação utiliza 7 pacientes com doença de Parkinson e seis controles, o que reduz o poder estatístico e dificulta determinar se as diferenças observadas entre os estados ON e OFF representam efeito da medicação ou variabilidade individual. Sugiro ampliar a amostra e controlar fatores como duração da doença, dose e tempo desde a última administração da medicação, além de repetir as avaliações em sessões diferentes para verificar a estabilidade do tempo de reação."

Resposta:

Os autores agradecem o comentário e concordam que a ampliação da amostra, o controle detalhado da dose e do tempo desde a última administração da medicação, e a realização de sessões de teste-reteste representam melhorias metodológicas relevantes. Tratando-se de um estudo piloto (pilot study, conforme indicado no título) com amostra já coletada sob aprovação ética específica (CAAE 61005322.3.0000.0129), a implementação dessas sugestões exigiria um novo recrutamento de paciente com Doença de Parkinson no estado ON e OFF e controle, nova coleta e análise de dados, o que não é exequível no prazo de revisão deste manuscrito. Essas limitações foram explicitamente incorporadas à seção de Conclusão e os autores destacam esses pontos como prioridades metodológicas para a próxima fase do estudo, atualmente em desenvolvimento com amostra ampliada.

Comentário 1.2

"O início do movimento é definido automaticamente quando a magnitude da aceleração ultrapassa a média basal acrescida de dois desvios padrão, sem validação desse limiar contra

uma referência independente. Acrescentaria uma validação do instante de início do movimento por vídeo, EMG ou outro método de referência."

Resposta:

Os autores esclarecem que o método de detecção por limiar (média basal + 2 desvios-padrão sobre a magnitude vetorial do acelerômetro) é amplamente utilizado na literatura de sensores inerciais para detecção de onset de movimento em tarefas motoras. Cabe esclarecer também que o momento do estímulo sonoro já era validado por hardware: a unidade de sincronização estava conectada a um canal de eletromiografia (EMG), e cada estímulo auditivo produzia um artefato elétrico detectável nesse canal, fornecendo um timestamp do disparo do estímulo independente de qualquer latência de software. Esse detalhe metodológico, antes implícito, foi agora explicitado na seção de Análise de Dados do manuscrito revisado. Reconhecemos, no entanto, que o instante de início do movimento em si (detectado unicamente pelo limiar no acelerômetro) ainda careceria de uma validação cruzada com referência independente (vídeo ou EMG do próprio movimento) para confirmar sua acurácia. Essa limitação foi adicionada explicitamente ao texto (seção de Limitações/Conclusão) e será incorporada ao desenho do próximo estudo.

Comentário 1.3

"Avaliaria também duração, velocidade e amplitude do movimento."

Resposta:

Os autores concordam que a análise de duração, velocidade e amplitude do movimento complementaria de forma relevante os achados de tempo de reação. Esses parâmetros cinemáticos já vêm sendo extraídos dos dados coletados no mesmo protocolo experimental e estão sendo analisados em maior profundidade no âmbito da tese de doutorado associada a este estudo, com submissão planejada para periódicos de maior escopo. Optou-se por não incluir essa análise neste manuscrito de congresso para preservar a originalidade e o ineditismo dos dados nas publicações futuras (tese e artigo de periódico), evitando sobreposição de conteúdo entre publicações. Uma referência a essa análise em andamento foi adicionada à seção de trabalhos futuros.

Comentário 1.4

"Os resultados mostram diferenças entre pacientes e controles em alguns blocos, mas nenhuma diferença significativa entre os estados ON e OFF após correção para múltiplas comparações."

Resposta:

Concordamos com essa observação, que já era discutida no manuscrito original. Reforçamos no texto revisado que a ausência de diferença significativa entre ON e OFF após correção, apesar de tamanhos de efeito grandes em alguns blocos (ex.: r = 0,799 no Bloco 5), provavelmente reflete o poder estatístico limitado da amostra pareada (n = 7), e não a ausência de um efeito real da medicação. Essa interpretação foi explicitada tanto na Discussão quanto na Conclusão.

Revisor 2

Comentário 2.1

"Reestruturar a introdução para focar diretamente na lacuna e hipótese do estudo."

Resposta:

Atendido. O parágrafo final da Introdução foi reescrito para apresentar explicitamente a lacuna na literatura (avaliação de RT majoritariamente em ensaios únicos e breves, sem caracterizar o comportamento sob demanda motora sustentada) e a hipótese do estudo (maior e mais sustentado comprometimento do RT no estado OFF sob demanda motora cumulativa).

Comentário 2.2

"Ajustar a seção final para pontuar brevemente caminhos futuros para expandir o tamanho da amostra e alcançar poder estatístico adequado em comparações pareadas de medicação, incorporar parâmetros cinemáticos diretos, como duração do movimento, amplitude de movimento, tempo de reação entre outros."

Resposta:

Atendido. O parágrafo de direções futuras, ao final da Discussão, foi expandido para mencionar explicitamente a necessidade de ampliação amostral para comparações pareadas ON vs OFF, a incorporação de parâmetros cinemáticos diretos (duração do movimento, amplitude e velocidade de pico), e a validação independente da detecção de início do movimento.

Comentário 2.3

"Definir explicitamente todas as siglas na primeira ocorrência no texto: CG, MMSE, H&Y e IQR."

Resposta:

Atendido. Todas as siglas foram definidas por extenso na primeira ocorrência no corpo do texto: Control Group (CG), Mini-Mental State Examination (MMSE), Hoehn & Yahr (H&Y) Scale e interquartile range (IQR).

Comentário 2.4

"Incluir os intervalos de confiança nas tabelas para aumentar a precisão estatística."

Resposta:

Atendido. Intervalos de confiança de 95% para as medianas de cada grupo em cada bloco foram calculados por bootstrap não paramétrico (10.000 reamostragens com reposição, método dos percentis), consistente com o uso de testes não paramétricos ao longo do manuscrito, e foram incluídos na Tabela 2 do manuscrito revisado, ao lado da mediana e do IQR já reportados. Como esperado para amostras pequenas (n = 6–7 por grupo), os ICs obtidos são amplos, o que é discutido no texto como reforço visual da limitação de precisão associada ao tamanho amostral reduzido.

Comentário 2.5

"Moderar as conclusões sobre 'aprendizado motor', reconhecendo as limitações impostas pelo tamanho reduzido da amostra, pelo design transversal, visto que o aprendizado motor real exige retenção a longo prazo."

Resposta:

Atendido. A linguagem foi revisada tanto na Discussão quanto na Conclusão para não caracterizar a convergência do grupo PD-ON em direção ao desempenho do grupo controle como "aprendizado motor" no sentido estrito. O texto agora indica que esse padrão pode refletir processos adaptativos de curto prazo dentro de uma única sessão, e que o desenho transversal e de sessão única do estudo não permite distinguir isso de retenção a longo prazo, que caracterizaria aprendizado motor verdadeiro.

Ressonância magnética na investigação de biomarcadores associados à identificação antecipada do transtorno do espectro autista: uma revisão de escopo

Thiago Padre da Silva
Programa de Pós-Graduação em Engenharia Biomédica – PPGEB
Universidade de Brasília – UnB
Brasília, DF
thiagopadre1@gmail.com

Cristiano Jacques Miosso
Programa de Pós-Graduação em Engenharia Biomédica – PPGEB
Universidade de Brasília – UnB
Brasília, DF
miosso@unb.br

***Resumo*—**O Transtorno do Espectro Autista (TEA) é uma condição do neurodesenvolvimento caracterizada por déficits persistentes na comunicação e interação social, associados a padrões restritos e repetitivos de comportamento. Como o diagnóstico permanece baseado predominantemente em avaliações clínicas, a identificação de biomarcadores objetivos constitui um desafio relevante. Nesse contexto, a ressonância magnética (RM), incluindo técnicas estruturais, funcionais e por tensor de difusão, tem sido investigada para caracterizar alterações cerebrais associadas ao TEA. Foi realizada uma revisão de escopo da literatura conforme as princípios do PRISMA ScR, com buscas nas bases PubMed, Scopus, Embase, realizadas até 30 de maio de 2026. Foram identificados 642 registros, dos quais 36 estudos foram incluídos após aplicação dos critérios de elegibilidade.Os estudos incluídos demonstraram alterações associadas ao TEA em aspectos estruturais, funcionais e de conectividade cerebral, envolvendo medidas como volume cerebral, espessura cortical, integridade da substância branca e padrões de conectividade funcional. Entretanto, os achados apresentaram heterogeneidade relacionada às características das amostras, métodos de aquisição e estratégias de análise, limitando a comparação entre estudos e a aplicação clínica dos biomarcadores identificados. Conclui-se que a RM apresenta potencial como ferramenta complementar para investigação dos mecanismos neurobiológicos do TEA e identificação de biomarcadores candidatos, mas as evidências atuais ainda não sustentam sua utilização isolada para diagnóstico clínico. Estudos futuros com maior padronização metodológica, amostras ampliadas e validação multicêntrica são necessários para determinar sua aplicabilidade clínica.

Palavras-chave: Transtorno do Espectro Autista; Ressonância Magnética; Diagnóstico Precoce; Neuroimagem; Biomarcadores; Engenharia Biomédica.

I. Introdução

O Transtorno do Espectro Autista (TEA) é uma condição do neurodesenvolvimento caracterizada por déficits persistentes na comunicação e interação social, associados a padrões restritos e repetitivos de comportamento [3]. Devido à sua elevada heterogeneidade clínica e etiológica, o diagnóstico permanece predominantemente baseado em critérios clínicos e comportamentais, o que frequentemente retarda a identificação do transtorno e o início de intervenções antecipadas, período considerado crítico em razão da elevada neuroplasticidade cerebral [35]. Neste contexto, cresce o interesse pela identificação de biomarcadores objetivos capazes de complementar a avaliação clínica e contribuir para o diagnóstico antecipado do TEA [33].

Entre as tecnologias de neuroimagem disponíveis, a ressonância magnética (RM) destaca-se por fornecer imagens de alta resolução do sistema nervoso central sem a utilização de radiação ionizante, permitindo a investigação não invasiva de alterações estruturais e funcionais do cérebro [22]. A RM estrutural possibilita avaliar modificações no volume cerebral, na espessura cortical e em regiões como cerebelo, corpo caloso, amígdala e hipocampo [4]. Por sua vez, a ressonância magnética funcional (fMRI) utiliza o sinal *Blood Oxygen Level Dependent* (BOLD) para inferir a atividade neuronal por meio das respostas hemodinâmicas, permitindo a análise da conectividade funcional em repouso ou durante a execução de tarefas específicas [31]. Complementarmente, a imagem por tensor de difusão (Diffusion Tensor Imaging – DTI) possibilita avaliar a integridade da substância branca por meio da difusão das moléculas de água ao longo dos feixes nervosos, fornecendo informações relevantes sobre a conectividade estrutural cerebral [37].

Nas últimas décadas, os avanços no processamento digital de imagens e na inteligência artificial ampliaram significativamente a capacidade de análise dos dados obtidos por ressonância magnética [17]. Técnicas de aprendizado de máquina e aprendizado profundo têm sido empregadas para integrar informações provenientes das modalidades estrutural, funcional e de difusão, possibilitando a identificação de

padrões complexos e de potenciais biomarcadores associados ao TEA [30]. Entretanto, a elevada heterogeneidade dos estudos, as diferenças nos protocolos de aquisição e processamento das imagens, os reduzidos tamanhos amostrais e a ausência de validação multicêntrica ainda representam obstáculos importantes para a incorporação dessas abordagens na prática clínica [7].

Sob a perspectiva da engenharia biomédica, a integração entre neuroimagem, processamento de sinais, inteligência artificial e análise computacional representa uma estratégia promissora para o desenvolvimento de ferramentas capazes de apoiar o diagnóstico antecipado e ampliar a compreensão dos mecanismos neurobiológicos envolvidos no TEA [21]. Apesar dos avanços observados, as evidências disponíveis ainda são insuficientes para estabelecer biomarcadores de imagem com aplicabilidade clínica consolidada, tornando necessária a realização de estudos metodologicamente mais robustos e padronizados [10].

Diante desse cenário, o presente estudo tem como objetivo sintetizar criticamente as evidências científicas acerca da utilização da ressonância magnética estrutural, funcional e por tensor de difusão na investigação de biomarcadores associados ao Transtorno do Espectro Autista [1]. Por meio de uma revisão de escopo conduzida conforme as recomendações do PRISMA ScR, busca-se analisar o potencial dessas modalidades de neuroimagem para a identificação antecipada do TEA, discutir suas limitações metodológicas e apontar perspectivas futuras para sua aplicação na prática clínica e no desenvolvimento de tecnologias em Engenharia Biomédica [19].

II. Materiais e Métodos

Esta revisão de escopo foi conduzida e relatada de acordo com as recomendações do Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR), conforme as diretrizes propostas por Tricco et al. [32].

A. Estratégia de Busca

A busca bibliográfica foi realizada nas bases PubMed/MEDLINE, Scopus e Embase, utilizando descritores MeSH e DeCS relacionados ao TEA, ressonância magnética, neuroimagem, biomarcadores e diagnóstico antecipado, combinados por operadores booleanos e adaptados à sintaxe de cada base.

B. Critérios de Elegibilidade

O recorte temporal de 2015 a 2026 foi definido com o objetivo de contemplar a produção científica mais recente e representativa sobre o tema investigado, refletindo os avanços conceituais, metodológicos e tecnológicos ocorridos na última década. O ano de 2015 foi adotado como marco inicial por representar um período em que houve maior consolidação das abordagens contemporâneas relacionadas ao objeto de estudo, permitindo a inclusão de evidências atuais sem comprometer a abrangência da revisão.

Embora 2026 ainda esteja em andamento, as publicações desse ano foram incluídas por estarem disponíveis nas bases de dados até o momento da realização da estratégia de busca. Para garantir transparência e reprodutibilidade, a busca bibliográfica foi realizada até 30 de maio de 2026, de modo que somente os estudos indexados e disponíveis nas bases consultadas até essa data foram considerados elegíveis.

C. Processo de Seleção

A busca eletrônica nas bases PubMed/MEDLINE, Embase e Scopus identificou inicialmente 642 registros. Após a remoção de 156 registros duplicados, permaneceram 486 estudos para triagem por títulos e resumos. Nessa etapa, 381 publicações foram excluídas por não atenderem aos critérios de elegibilidade previamente estabelecidos. Os 105 artigos potencialmente relevantes foram avaliados em texto completo, dos quais 69 foram excluídos por diferentes motivos metodológicos. Ao final, 36 estudos preencheram todos os critérios de inclusão e compuseram a síntese qualitativa desta revisão de escopo [32].

D. Extração dos Dados

A seleção dos estudos foi realizada por um único revisor, que conduziu independentemente a triagem dos títulos e resumos, a avaliação dos textos completos e a extração dos dados. Os estudos potencialmente elegíveis foram avaliados segundo os critérios de inclusão e exclusão previamente estabelecidos. Para garantir a padronização da coleta das informações, foi elaborado um instrumento de extração de dados contendo as principais variáveis analisadas em cada estudo incluído na revisão. Esse instrumento foi utilizado durante a leitura completa dos artigos para registrar informações relacionadas às características metodológicas, população estudada, modalidade de ressonância magnética empregada, biomarcadores investigados, principais resultados e limitações dos estudos. A Tabela 1 apresenta as variáveis que compuseram esse processo de extração de dados, não correspondendo aos resultados individuais dos estudos incluídos.

TABLE I
Variáveis planejadas para extração de dados dos estudos incluídos na revisão de escopo.

Variável	Descrição
Autor	Primeiro autor.
Ano	Ano de publicação.
País	Local do estudo.
Tipo de estudo	Coorte, transversal, caso-controle, ensaio clínico etc.
Amostra	Número de participantes.
Faixa etária	Idade dos participantes.
Modalidade de RM	Estrutural, fMRI ou DTI.
Biomarcadores avaliados	Região cerebral, conectividade ou outro parâmetro avaliado.
Principais resultados	Achados relevantes do estudo.
Limitações	Limitações metodológicas.

E. Síntese dos Dados

A síntese dos estudos foi conduzida por meio de síntese narrativa, em razão da expressiva heterogeneidade entre os trabalhos incluídos. Os estudos foram inicialmente agrupados de acordo com a modalidade de neuroimagem empregada, compreendendo: ressonância magnética (RM) estrutural, ressonância magnética funcional (fMRI) em estado de repouso, fMRI baseada em tarefas, imagem por difusão (Diffusion Tensor Imaging – DTI e técnicas correlatas) e abordagens multimodais, que combinaram diferentes técnicas de aquisição e análise de neuroimagem [5].

Dentro de cada grupo, os estudos foram comparados considerando aspectos metodológicos, incluindo características da amostra (tamanho, faixa etária, critérios diagnósticos e características clínicas), protocolos de aquisição das imagens, métodos de processamento e análise, regiões cerebrais investigadas e principais desfechos [11]. Essa organização permitiu identificar padrões de convergência e divergência entre os achados, bem como avaliar a consistência das evidências produzidas por cada modalidade de imagem [20].

A heterogeneidade metodológica foi considerada durante a interpretação dos resultados [15]. Diferenças relacionadas às características das populações estudadas, aos delineamentos, aos parâmetros de aquisição de imagem, aos métodos analíticos e às medidas de desfecho foram utilizadas para contextualizar variações entre os estudos. Quando resultados conflitantes foram identificados, estes foram apresentados de forma explícita e discutidos à luz de possíveis explicações metodológicas e clínicas, sem atribuição de maior peso a um estudo isolado, mas considerando a qualidade metodológica, a consistência dos achados entre estudos semelhantes e a plausibilidade biológica das interpretações propostas.

A condução da síntese narrativa seguiu os princípios recomendados pelas diretrizes para síntese narrativa em revisões de escopos científicos, envolvendo a organização dos estudos em grupos conceitualmente homogêneos, a descrição analítica das características e dos resultados, a exploração das fontes de heterogeneidade e a elaboração de uma interpretação crítica das evidências disponíveis [27].

Não foi realizada metanálise porque os estudos apresentaram heterogeneidade substancial quanto aos delineamentos, às populações avaliadas, às modalidades de neuroimagem, aos protocolos de aquisição, às estratégias de processamento dos dados e aos desfechos analisados. Além disso, os resultados foram frequentemente apresentados por meio de diferentes métricas, regiões de interesse e métodos estatísticos, impossibilitando a obtenção de estimativas quantitativas comparáveis e a combinação estatística dos achados de forma metodologicamente robusta. Dessa forma, a síntese narrativa foi considerada a abordagem mais apropriada para integrar criticamente as evidências disponíveis [31].

F. Limitações

Esta revisão apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. A triagem dos estudos, a avaliação dos textos completos e a extração dos dados foram realizadas por um único revisor, o que pode aumentar o risco de viés de seleção e de possíveis erros na extração ou interpretação das informações [32]. Embora tenham sido empregados critérios de elegibilidade previamente definidos e procedimentos padronizados ao longo dessas etapas, a ausência de um segundo revisor independente constitui uma limitação metodológica relevante. Estudos futuros poderão minimizar esse risco mediante a participação de revisores independentes em todas as etapas de seleção e extração dos dados, com resolução de eventuais discordâncias por consenso.

Também devem ser consideradas as limitações relacionadas ao recorte temporal adotado (2015-2026), aos idiomas de publicação e às bases de dados consultadas, fatores que podem ter restringido a identificação e a inclusão de estudos potencialmente relevantes. Adicionalmente, não se pode descartar a possibilidade de viés de publicação, uma vez que estudos com resultados negativos, inconclusivos ou menos expressivos podem estar sub-representados na literatura disponível.

Por fim, os estudos incluídos apresentaram considerável heterogeneidade quanto às características das amostras, modalidades e protocolos de ressonância magnética, estratégias de processamento das imagens e métodos de análise empregados. Essa variabilidade metodológica inviabilizou a realização de uma metanálise e justificou a adoção de uma síntese narrativa dos achados. Somada às diferenças na qualidade metodológica dos estudos primários, essa heterogeneidade limita a comparabilidade entre os resultados e deve ser considerada ao interpretar a generalização das conclusões.

III. Resultados

A estratégia de busca realizada nas bases PubMed/MEDLINE, Embase e Scopus resultou na identificação de 642 registros. Após a remoção de 156

TABLE II
Comparação das modalidades de Ressonância Magnética aplicadas ao diagnóstico do Transtorno do Espectro Autista.

Modalidade	**Principal aplicação**	**Biomarcadores investigados**	**Vantagens**	**Limitações**
RM Estrutural	Avaliação anatômica cerebral	Volume cerebral, espessura cortical, cerebelo e amígdala	Alta resolução espacial	Baixa especificidade para o TEA
fMRI	Avaliação da conectividade funcional	Redes neurais, Default Mode Network (DMN) e conectividade social	Avalia a atividade cerebral em tempo real	Sensível a movimentos e à variação dos protocolos
DTI	Avaliação da integridade da substância branca	Anisotropia fracionada e difusividade	Detecta alterações microestruturais	Requer processamento avançado das imagens

TABLE III
VISÃO GERAL DO PROCESSO DE SELEÇÃO DOS ESTUDOS CONFORME O PRISMA 2020.

Etapa do PRISMA 2020	Quantidade (n)
Registros identificados na PubMed	182
Registros identificados na Embase	247
Registros identificados na Scopus	213
Total de registros identificados	**642**
Registros duplicados removidos	156
Registros após remoção de duplicatas	486
Estudos selecionados para leitura de título e resumo	486
Estudos excluídos na triagem	381
Artigos avaliados em texto completo	105
Artigos excluídos após leitura completa	69
Estudos incluídos na revisão de escopo	**36**

duplicatas, permaneceram 486 estudos para a etapa de triagem por títulos e resumos. Desses, 381 foram excluídos por não atenderem aos critérios de elegibilidade previamente estabelecidos. Foram selecionados 105 artigos para leitura na íntegra, dos quais 69 foram posteriormente excluídos por não responderem à questão de pesquisa. Ao final do processo de seleção, 36 estudos foram incluídos na síntese qualitativa desta revisão de escopo, seguindo os processos de um fluxograma PRISMA-ScR [32]. A Tabela IV sintetiza as informações centrais dos 36 artigos selecionados, incluindo o tipo de desenho experimental, o tipo de ressonância utilizado e o foco de cada artigo.

Os estudos incluídos foram publicados entre 2015 e 2026 e apresentaram diferentes delineamentos metodológicos, contemplando principalmente estudos transversais, de coorte e caso-controle. As pesquisas envolveram indivíduos com diagnóstico ou suspeita de Transtorno do Espectro Autista, predominantemente crianças e adolescentes, e utilizaram diferentes modalidades de ressonância magnética, incluindo ressonância magnética estrutural, ressonância magnética funcional (fMRI), imagem por tensor de difusão (DTI) e abordagens multimodais associadas a técnicas de inteligência artificial. Os principais biomarcadores investigados compreenderam alterações no volume cerebral, espessura cortical, conectividade funcional, integridade da substância branca e organização de redes neurais [34].

Os estudos que utilizaram ressonância magnética estrutural identificaram alterações em diferentes regiões encefálicas, incluindo modificações no volume cerebral, na espessura cortical, no cerebelo, no corpo caloso, na amígdala e no hipocampo. De modo geral, esses achados indicaram diferenças morfológicas entre indivíduos com TEA e controles, embora os resultados tenham apresentado variabilidade entre as populações estudadas [37]. Nos estudos que empregaram ressonância magnética funcional, foram observadas alterações na conectividade de redes neurais relacionadas à comunicação, cognição social, linguagem e processamento emocional [23]. As diferenças identificadas envolveram tanto a conectividade em estado de repouso quanto durante a execução de tarefas cognitivas, variando conforme as características clínicas das amostras e os protocolos utilizados [30].

Os estudos baseados em imagem por tensor de difusão demonstraram alterações na integridade da substância branca, evidenciadas por diferenças em parâmetros relacionados à organização dos feixes neurais e à conectividade estrutural cerebral [9]. Além disso, alguns trabalhos combinaram diferentes modalidades de neuroimagem com algoritmos de aprendizado de máquina e aprendizado profundo, buscando identificar padrões capazes de auxiliar na classificação de indivíduos com TEA. Essas abordagens multimodais apresentaram resultados promissores na identificação de biomarcadores candidatos, embora a maioria dos estudos tenha destacado a necessidade de validação em amostras independentes antes de sua aplicação clínica [8].

De maneira geral, os estudos incluídos evidenciaram alterações estruturais, funcionais e de conectividade cerebral associadas ao transtorno do espectro autista, indicando o potencial da ressonância magnética para a investigação de biomarcadores relacionados ao transtorno [31]. Entretanto, também foi observada elevada heterogeneidade entre os estudos quanto às características das populações avaliadas, modalidades de imagem empregadas, protocolos de aquisição, métodos de processamento e estratégias analíticas, aspecto que dificultou a comparação direta entre os resultados e impossibilitou a realização de metanálise [34].

IV. DISCUSSÃO

Os achados desta revisão de escopo demonstram que a ressonância magnética tem desempenhado papel relevante na ampliação do conhecimento acerca das alterações neurobiológicas associadas ao transtorno do espectro autista [26]. A análise dos estudos incluídos evidencia que as diferentes modalidades de neuroimagem, incluindo a ressonância magnética estrutural, a ressonância magnética funcional e a imagem por tensor de difusão, são capazes de identificar alterações na morfologia cerebral, na conectividade funcional e na integridade da substância branca [16]. Esses resultados reforçam a hipótese de que o TEA está associado a alterações em redes neurais distribuídas, envolvendo diferentes regiões cerebrais e múltiplos mecanismos do neurodesenvolvimento, ao invés de uma única alteração anatômica ou funcional específica [9], [28].

Embora as alterações estruturais tenham sido observadas de forma recorrente em regiões como córtex cerebral, cerebelo, amígdala, hipocampo e corpo caloso, os estudos analisados apresentaram resultados heterogêneos quanto à localização, magnitude e direção dessas alterações [13]. Essa variabilidade pode ser explicada pela própria natureza multifatorial do TEA, bem como pelas diferenças entre as populações estudadas, incluindo faixa etária, gravidade clínica, presença de comorbidades, critérios diagnósticos utilizados e tamanhos amostrais. Além disso, diferenças nos protocolos de aquisição das imagens e nas metodologias empregadas para processamento e análise contribuem para a inconsistência dos achados descritos na literatura [10].

TABLE IV: Artigos analisados e utilizados no escopo.

Nº	Autor/Ano	Desenho/Tipo	Modalidade de RM	Nº amostral	Foco principal
1	APA (2022)	Documento técnico	N/A	N/A	Transtornos neurocognitivos
2	Andrews et al. (2021)	Longitudinal	DTI	NI	Conectividade da substância branca
3	APA (2023)	Manual diagnóstico	N/A	N/A	Critérios diagnósticos do TEA
4	Augusto et al. (2024)	Observacional	RM	NI	Diagnóstico tardio do TEA
5	Carvalho Macruz et al. (2024)	Revisão	RM / neuroimagem	N/A	IA aplicada à neurorradiologia
6	Courchesne et al. (2018)	Revisão	RM estrutural	N/A	Desenvolvimento cerebral no TEA
7	Dichter (2017)	Revisão	fMRI	N/A	Função cerebral no TEA
8	Ding et al. (2024)	Meta-análise	RM + IA	N/A	Deep learning no diagnóstico
9	Donovan & Basson (2017)	Revisão	RM estrutural	N/A	Neuroanatomia do TEA
10	Ecker et al. (2015)	Revisão	RM estrutural/fMRI	N/A	Neuroimagem no TEA
11	Edmondson et al. (2020)	Estudo clínico	MRS	NI	GABA e desempenho visual
12	Fatima et al. (2025)	Estudo original	RM + IA	NI	Classificação do TEA
13	Hazlett et al. (2017)	Coorte	RM estrutural	NI	Desenvolvimento cerebral precoce
14	Hazlett et al. (2017)	Coorte	RM estrutural	123	Desenvolvimento cerebral precoce
15	Heinsfeld et al. (2018)	Estudo original	RM/fMRI + IA	1.035	Deep learning e ABIDE
16	Heinsfeld et al. (2017)	Estudo original	RM/fMRI + IA	1.035	Classificação do TEA
17	Just et al. (2019)	Revisão	fMRI	N/A	Conectividade cortical
18	Le Bihan et al. (2016)	Metodológico	DTI	N/A	Difusão e conectividade
19	Leote et al. (2020)	Estudo original	fMRI	NI	fMRI navegada
20	Maguire-Zeiss et al. (2016)	Revisão	N/A	N/A	Neurodegeneração
21	Mendes (2025)	Qualitativo	N/A	NI	TEA na atenção primária
22	Ministério da Saúde (2023)	Documento oficial	N/A	N/A	Linha de cuidado do TEA
23	Mottron (2021)	Artigo de opinião	N/A	N/A	Diagnóstico do TEA
24	Nelson & Valakh (2015)	Revisão	N/A	N/A	Equilíbrio excitação/inibição
25	Nielsen et al. (2020)	Revisão	RM/neuroimagem	N/A	IA em psiquiatria
26	Page et al. (2022)	Diretriz	N/A	N/A	PRISMA 2020
27	Pearson et al. (2021)	Revisão	N/A	N/A	Metodologia em pesquisa sobre TEA
28	Randall et al. (2018)	Revisão Cochrane	N/A	N/A	Testes diagnósticos
29	Sharma et al. (2024)	Revisão	RM estrutural	N/A	Processamento de imagens
30	Thomson et al. (2024)	Revisão sistemática/meta-análise	MRS	N/A	Neuroquímica cerebral
31	Tricco et al. (2018)	Diretriz	N/A	N/A	PRISMA-ScR
32	Vakadakis et al. (2024)	Revisão sistemática	RM + IA	N/A	Biomarcadores diagnósticos
33	Varoquaux & Cheplygina (2022)	Metodológico	RM + ML	N/A	Machine learning médico
34	WHO (2022)	Classificação	N/A	N/A	CID-11
35	Wu et al. (2023)	Revisão	RM estrutural	N/A	Neuroimagem estrutural
36	Zeidan-Chulia et al. (2024)	Revisão	RM/conectividade	N/A	Biomarcadores e conectividade

Nota: ABIDE = Autism Brain Imaging Data Exchange; DTI = imagem por tensor de difusão; fMRI = ressonância magnética funcional; IA = inteligência artificial; ML = aprendizado de máquina; MRS = espectroscopia por ressonância magnética; N/A = não se aplica; NI = não informado; RM = ressonância magnética; TEA = transtorno do espectro autista.

Resultados semelhantes foram observados nos estudos que empregaram ressonância magnética funcional [29]. Embora diversas pesquisas tenham identificado alterações na conectividade de redes neurais relacionadas à comunicação social, linguagem, funções executivas e processamento emocional, não foi possível estabelecer um padrão universal de conectividade funcional aplicável a todos os indivíduos com TEA [24]. Essa diversidade de resultados evidencia que a organização funcional cerebral apresenta grande variabilidade entre os pacientes, reforçando a complexidade biológica do transtorno e a necessidade de abordagens individualizadas para sua investigação [11].

As pesquisas baseadas em imagem por tensor de difusão também demonstraram alterações na organização microestrutural da substância branca, sugerindo comprometimento da conectividade estrutural entre diferentes regiões cerebrais [18]. Entretanto, assim como observado nas demais modalidades de neuroimagem, diferenças metodológicas relacionadas aos parâmetros de aquisição, métricas utilizadas e estratégias de processamento limitaram a comparação direta entre os estudos, dificultando a consolidação desses achados como biomarcadores confiáveis para utilização clínica [6].

Outro aspecto que merece destaque é o crescente emprego de métodos de inteligência artificial aplicados aos dados de ressonância magnética [25]. A combinação entre neuroimagem e algoritmos de aprendizado de máquina ou aprendizado profundo tem possibilitado a identificação de padrões complexos dificilmente detectáveis por métodos estatísticos convencionais, aumentando o potencial para classificação automatizada de indivíduos com TEA [14]. Apesar dos resultados promissores, a maior parte dos estudos ainda apresenta limitações relacionadas ao reduzido número de participantes, ao risco de sobreajuste dos modelos computacionais, à limitada interpretabilidade dos algoritmos e à ausência de validação externa em populações independentes, fatores que restringem sua utilização rotineira na prática clínica [13].

De forma geral, os resultados desta revisão indicam que a ressonância magnética constitui uma importante ferramenta para investigação dos mecanismos neurobiológicos do Transtorno do Espectro Autista e para a identificação de potenciais biomarcadores de imagem [36]. No entanto, as evidências atualmente disponíveis ainda não sustentam sua utilização como método diagnóstico isolado, uma vez que a elevada heterogeneidade metodológica dos estudos, associada à ausência de protocolos padronizados e de validação multicêntrica, reduz a reprodutibilidade dos achados e limita sua aplicação clínica [12].

Sob a perspectiva da Engenharia Biomédica, os resultados reforçam a importância do desenvolvimento de protocolos padronizados de aquisição e processamento de imagens, bem como da integração entre neuroimagem, inteligência artificial e outras modalidades de biomarcadores biológicos [10]. A utilização de abordagens multimodais poderá contribuir para aumentar a precisão diagnóstica e favorecer a identificação precoce do TEA. Entretanto, a consolidação dessas tecnologias dependerá da realização de estudos multicêntricos, com amostras representativas, validação externa e metodologias harmonizadas, permitindo que os biomarcadores identificados possam ser incorporados de forma segura e eficaz à prática clínica [2].

Em síntese, os estudos analisados evidenciam avanços significativos na utilização da ressonância magnética como ferramenta de investigação científica do TEA. Contudo, a tradução desses achados para a prática assistencial ainda requer o fortalecimento das evidências disponíveis, especialmente por meio de pesquisas colaborativas de grande escala, capazes de reduzir a heterogeneidade metodológica atualmente observada e estabelecer biomarcadores robustos, reprodutíveis e clinicamente aplicáveis [2].

V. Conclusão

A presente revisão sintetizou os achados publicados entre 2015 e 2026 sobre a aplicação da ressonância magnética estrutural, da ressonância magnética funcional, da imagem por difusão e de abordagens multimodais na investigação do Transtorno do Espectro Autista (TEA). Em conjunto, os estudos identificaram alterações recorrentes na morfologia cerebral, na conectividade funcional e na integridade da substância branca em indivíduos com TEA. Entretanto, os padrões observados não foram uniformes entre as pesquisas, variando conforme a faixa etária, as características das amostras, os protocolos de aquisição e processamento das imagens e os métodos analíticos empregados.

A análise das evidências indica que as técnicas de neuroimagem apresentam potencial para ampliar a compreensão dos mecanismos neurobiológicos do TEA e para identificar biomarcadores candidatos. Contudo, a heterogeneidade metodológica e clínica observada entre os estudos, aliada à ausência de protocolos padronizados e à limitada reprodutibilidade de alguns achados, reduz a consistência das evidências atualmente disponíveis e impede a definição de biomarcadores de imagem com aplicabilidade clínica estabelecida. Dessa forma, os resultados desta revisão devem ser interpretados como evidências de potencial diagnóstico e investigativo, e não como demonstração de utilidade clínica consolidada.

As abordagens que integram diferentes modalidades de neuroimagem e métodos computacionais, incluindo técnicas de processamento de imagens e inteligência artificial, mostraram resultados promissores em alguns estudos para a discriminação entre indivíduos com TEA e controles. No entanto, esses achados ainda necessitam de validação em estudos multicêntricos, com amostras maiores, protocolos harmonizados e validação externa antes de sua incorporação à prática clínica.

Em síntese, os estudos revisados sustentam que a ressonância magnética constitui uma importante ferramenta de pesquisa para a caracterização das alterações cerebrais associadas ao TEA, mas as evidências atualmente disponíveis ainda apresentam certeza limitada a moderada, principalmente em decorrência da heterogeneidade dos estudos incluídos. Assim, os dados disponíveis não sustentam o uso da neuroimagem como substituta da avaliação clínica nem permitem concluir que sua utilização, isoladamente, resulte em melhora dos

desfechos clínicos ou da qualidade de vida dos indivíduos com TEA.

REFERENCES

[1] American Psychiatric Association Publishing. Neurocognitive disorders supplement: Updated excerpts for delirium codes, major and mild neurocognitive disorders. 2022.

[2] D. S. Andrews et al. Longitudinal development of white matter connectivity in autism spectrum disorder. 2021. Estudo longitudinal; dados bibliográficos a completar.

[3] Associação Psiquiátrica Americana. *Manual Diagnóstico e Estatístico de Transtornos Mentais: DSM-5-TR*. Artmed, Porto Alegre, 5 edition, 2023. Edição revisada.

[4] Janaína Aparecida De Oliveira Augusto, Carolina Kulcsar Caravieri, Rodrigo Genaro Arduini, Décio Brunoni, Sylvia Maria Ciasca, Maria Cristina Triguero, and Veloz Teixeira. Delayed diagnosis in children with autism spectrum disorder or intellectual disability. 2024.

[5] F. B. Carvalho Macruz, A. L. M. P. Dias, C. S. Andrade, M. P. Nucci, C. M. Rimkus, L. T. Lucato, A. J. D. Rocha, and F. C. Kitamura. The new era of artificial intelligence in neuroradiology: Current research and promising tools. *Arquivos de Neuro-Psiquiatria*, 82(6):1–12, June 2024.

[6] Eric Courchesne et al. Brain development in autism spectrum disorder. 2018. Dados bibliográficos a completar.

[7] Gabriel S. Dichter. Functional magnetic resonance imaging of autism spectrum disorders. *Dialogues in Clinical Neuroscience*, 14(3):319–351, September 2012.

[8] Gabriel S. Dichter. Functional magnetic resonance imaging of autism spectrum disorders. *Dialogues in Clinical Neuroscience*, 14(3):319–351, September 2017.

[9] X. Ding et al. Deep learning for the diagnosis of autism spectrum disorder: A meta-analysis. 2024. Meta-análise; dados bibliográficos a completar.

[10] A. P. Donovan and M. A. Basson. The neuroanatomy of autism – a developmental perspective. *Journal of Anatomy*, 230(1):4–15, January 2017.

[11] Christine Ecker, Susan Y. Bookheimer, and Declan G. M. Murphy. Neuroimaging in autism spectrum disorder: Brain structure and function across the lifespan. *The Lancet Neurology*, 14(11):1121–1134, 2015.

[12] D. A. Edmondson et al. Gaba and visual performance in autism spectrum disorder. 2020. Estudo clínico utilizando espectroscopia por ressonância magnética.

[13] S. Fatima et al. Classification of autism spectrum disorder using magnetic resonance imaging and vgg20. 2025. Estudo original; dados bibliográficos a completar.

[14] Heather Cody Hazlett et al. Early brain development in infants at high risk for autism spectrum disorder. 2017. Estudo de coorte; dados bibliográficos a completar.

[15] Heather Cody Hazlett, Hongbin Gu, Brent C. Munsell, Sun Hyung Kim, Martin Styner, Jason J. Wolff, Jed T. Elison, Meghan R. Swanson, Hongtu Zhu, Kelly N. Botteron, D. Louis Collins, John N. Constantino, Stephen R. Dager, Annette M. Estes, Alan C. Evans, Vladimir S. Fonov, Guido Gerig, Penelope Kostopoulos, Robert C. McKinstry, Juhi Pandey, Sarah Paterson, John R. Pruett, Robert T. Schultz, Dennis W. Shaw, Lonnie Zwaigenbaum, Joseph Piven, and IBIS Network. Early brain development in infants at high risk for autism spectrum disorder. *Nature*, 542(7641):348–351, 2017.

[16] Anibal S. Heinsfeld, Pedro L. Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. 2018. Estudo original utilizando aprendizado profundo e dados ABIDE.

[17] Anibal Sólon Heinsfeld, Rosa Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23, 2017.

[18] M. A. Just et al. Cortical connectivity and underconnectivity in autism spectrum disorder. 2019. Revisão sobre conectividade cerebral.

[19] Denis Le Bihan, Jean-François Mangin, Cyril Poupon, Chris A. Clark, Sabina Pappata, Nicolas Molko, and Hugues Chabriat. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546, April 2016.

[20] J. Leote, R. Loução, C. Viegas, M. Lauterbach, A. Perez-Hick, J. Monteiro, R. G. Nunes, and H. A. Ferreira. Impact of navigated task-specific fmri on direct cortical stimulation. *Journal of Neurological Surgery Part A: Central European Neurosurgery*, 81(6):555–564, November 2020.

[21] Kathleen A. Maguire-Zeiss, Bogdan A. Stoica, Varun Kesherwani, Robert I. Richards, Sarah A. Robertson, Louise V. O'Keefe, Dani Fornarino, Andrew Scott, Michael Lardelli, and Bernhard T. Baune. The enemy within: Innate surveillance-mediated cell death, the common mechanism of neurodegenerative disease. *Frontiers in Neuroscience*, 10:193, 2016.

[22] M. Mendes. Autism spectrum disorder in primary health care: Professional training and care. 2025. Estudo qualitativo; dados bibliográficos a completar.

[23] Ministério da Saúde. Linha de cuidado para a atenção às pessoas com transtornos do espectro do autismo e suas famílias na rede de atenção psicossocial do sus, 2023. Documento oficial.

[24] Laurent Mottron. Diagnostic prototypes and autism spectrum disorder. 2021. Artigo de opinião; dados bibliográficos a completar.

[25] S. B. Nelson and V. Valakh. Excitation-inhibition balance and neuronal homeostasis in autism spectrum disorder. 2015. Revisão; dados bibliográficos a completar.

[26] A. Nielsen et al. Artificial intelligence in psychiatry and neuroimaging. 2020. Dados bibliográficos a completar.

[27] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Laura A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. A declaração prisma 2020: diretriz atualizada para relatar revisões sistemáticas. *Revista Panamericana de Salud Pública*, 46:e112, 2022.

[28] A. Pearson et al. Autism research: Epistemological and methodological considerations. 2021. Revisão; dados bibliográficos a completar.

[29] M. Randall et al. Diagnostic tests for autism spectrum disorder. 2018. Revisão Cochrane; dados bibliográficos a completar.

[30] A. Sharma et al. Structural magnetic resonance imaging and digital image processing techniques in autism spectrum disorder: A comprehensive review. *NeuroImage: Clinical*, 41:103560, 2024.

[31] A. R. Thomson, D. Pasanta, T. Arichi, and N. A. J. Puts. Neurometabolite differences in autism as assessed with magnetic resonance spectroscopy: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 162:105728, July 2024.

[32] Andrea C. Tricco, Erin Lillie, W. Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D. J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jessica Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. *Annals of Internal Medicine*, 169(7):467–473, 2018.

[33] V. Vakadakis et al. Artificial intelligence and mri in autism spectrum disorder: A systematic review of diagnostic and predictive biomarkers. *Brain Sciences*, 14(2):145, 2024.

[34] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: Methodological considerations. 2022. Dados bibliográficos a completar.

[35] World Health Organization. International classification of diseases 11th revision (icd-11), 2022. Classificação internacional de doenças.

[36] X. Wu et al. Structural neuroimaging and molecular basis of autism spectrum disorder. 2023. Dados bibliográficos a completar.

[37] F. Zeidan-Chulia et al. Epigenetics, brain connectivity, and neuroimaging biomarkers in autism spectrum disorder: Current evidence and future directions. *Journal of Clinical Medicine*, 13(4):1012, 2024.

Temporal Processing and Decoder Architecture in Pseudo-Online BCIs

José Vanderlei da Silva Júnior
Institute of Science and Technology
Federal University of São Paulo (UNIFESP)
São José dos Campos, Brazil
jose.vanderlei@unifesp.br

Seidi Yonamine Yamauti
Graduate Program in Neurology and Neurosciences
Federal University of São Paulo (UNIFESP)
São Paulo, Brazil
seidi.yonamine@unifesp.br

Jean Faber
Department of Physics
Federal University of Sergipe (UFS)
São Cristóvão, Brazil
Graduate Program in Neurology and Neurosciences
Federal University of São Paulo (UNIFESP)
São Paulo, Brazil
jean.faber@unifesp.br

***Abstract*—Motor imagery brain-computer interfaces (BCIs) require rapid, stable decisions for real-time operation, yet decoding pipelines are typically selected based solely on offline accuracy. To address this, we applied a causal pseudo-online evaluation framework on a public database, incorporating a within-session protocol and a rest class to investigate how decoder architecture and temporal parameters (window length and overlap) jointly affect BCI usability. Comparing Common Spatial Pattern (CSP) and Riemannian Tangent Space (TGSP) approaches, mixed-effects modeling revealed that decoder architecture primarily dictates decision quality and confidence, whereas temporal stability is strictly governed by temporal processing. Larger windows consistently improved classification performance, confidence, and stability, but increased detection latency. Also, smaller steps between windows accelerated responsiveness and increased prediction confidence in nonlinear Riemannian decoders. These findings demonstrate that real-time BCI configuration is a multiobjective design problem: rather than relying on offline accuracy alone, temporal and algorithmic parameters must be tailored to the latency and stability requirements of specific clinical applications.**

***Index Terms*—Brain-computer interface (BCI); Motor imagery; Electroencephalography (EEG); Pseudo-online evaluation; Temporal processing.**

I. Introduction

Brain-computer interfaces (BCIs) establish a closed-loop interaction between neural activity, user intention, external devices, and sensory feedback. In motor imagery (MI)-based neurorehabilitation, this interaction is particularly important because the feedback generated by the system can reinforce task-related cortical activity, promote neuroplasticity, and facilitate motor recovery. However, once a BCI operates in real time, the user's neural activity continuously adapts to the feedback provided by the decoder, making online operation fundamentally different from the static conditions typically assumed during offline model development [1], [2].

Despite this distinction, decoding pipelines are still predominantly selected according to offline classification performance. While offline accuracy is essential for evaluating discriminability, it does not necessarily reflect how a decoder behaves during continuous operation [3]–[6]. Practical BCIs must simultaneously satisfy several additional requirements, including rapid response, reliable decision making, high prediction confidence, and temporal stability. These objectives are often competing, meaning that maximizing a single offline metric does not necessarily produce the best real-time behavior.

Pseudo-online evaluation has emerged as an intermediate step between conventional offline analysis and fully online experiments [4]. By processing continuous EEG recordings using causal sliding windows, pseudo-online frameworks approximate many characteristics of real-time operation while avoiding the logistical complexity of repeated online experiments. Nevertheless, most BCI studies remain focused on comparing decoding algorithms, providing limited insight into a more fundamental question: which design choices actually determine different dimensions of online usability?

In particular, the temporal organization of signal processing has received considerably less attention than decoder architecture. Decisions such as window length and temporal sampling density directly determine how much neural information is integrated before each prediction and how frequently new decisions become available. These parameters may therefore influence system responsiveness, confidence, and stability independently of the decoding algorithm itself. Understanding the relative contribution of temporal processing and decoder architecture is essential for designing BCIs whose behavior matches the requirements of different clinical and assistive applications.

In this work, we adapt a pseudo-online evaluation protocol to investigate how temporal processing parameters and decoder architecture jointly shape multiple dimensions of BCI performance. Continuous EEG recordings are processed causally using sliding windows and an additional rest class to reproduce realistic operating conditions. Classical CSP-based pipelines are compared with tangent-space Riemannian approaches under different temporal configurations, while performance is evaluated through complementary measures of classification performance, detection latency, prediction confidence, and temporal stability.

Rather than identifying a single optimal decoding pipeline, our objective is to characterize how different design parameters influence different aspects of pseudo-online behavior. This analysis provides a multiobjective view of decoder evaluation, allowing pipeline configurations to be selected according to application-specific requirements instead of offline accuracy alone.

Specifically, this study investigates: (i) whether offline classification performance predicts online responsiveness and stability; (ii) the relative contribution of decoder architecture versus temporal processing to each dimension of usability; and (iii) whether these dimensions can be jointly optimized or require application-specific trade-offs

The remainder of this paper is organized as follows. Section II describes the proposed pseudo-online analysis, the evaluated decoding pipelines, and the performance metrics. Section III presents the experimental results, emphasizing how temporal processing and decoder architecture affect different dimensions of system behavior. Finally, Section IV discusses the implications of these findings for the design and evaluation of real-time BCIs and concludes the paper.

II. Materials and Methods

A. Computational setting

The analysis was conducted on a CPU Google Colaboratory environment, using Python as programming language. The decoder pipelines used standard implementations available on *sklearn* [7] and *pyriemann* [8] (for riemannian manifold based methods). Preprocessing was done with *MNE* [9] library. Data analysis and visualization were made with help of *Pandas* [10], *matplotlib* [11] and *seaborn* [12] libraries.

B. Dataset

The analysis was conducted using the BNCI2014-001 motor imagery database from the BCI Competition IV [13]. The dataset comprises EEG recordings from nine healthy participants performing four different motor imagery tasks following visual cues: left hand, right hand, both feet and tongue. Each subject completed two recording sessions acquired on the same day, each containing six runs of 48 trials. EEG was recorded from 22 scalp electrodes at a sampling frequency of 250 Hz.

Unlike conventional offline analyses based on random cross-validation, the objective of this study is to approximate the conditions encountered during online operation. Therefore, all evaluations were performed using a causal within-session protocol that preserves the temporal order of the recorded signals.

C. Pseudo-online evaluation framework

The pseudo-online framework was inspired by previous continuous motor imagery decoding approaches, but introduces modifications designed to better reproduce realistic online operation while allowing the evaluation of multiple dimensions of decoder usability.

The only preprocessing step was the band-pass filtering in the sensorimotor frequency range (8–30 Hz) of the continuous EEG recording. Instead of segmenting the data into isolated motor imagery epochs, the entire recording was processed continuously using causal sliding windows. Each prediction therefore relied exclusively on information available up to the current instant, reproducing the causal constraint of an online BCI.

To emulate continuous operation, windows were extracted throughout the recording, including both motor imagery and non-imagery periods. Windows outside the motor imagery interval were assigned to an additional rest class, allowing the decoder to distinguish intentional control from background activity. When a window contained samples belonging to multiple classes, its label was assigned according to the majority of contained samples; ties were resolved in favor of the most recent label, reflecting the user's latest motor intention.

Although the overall framework follows the pseudo-online strategy proposed by Carrara and Papadopoulo, two important modifications were introduced. First, decoder training and evaluation followed a causal within-session protocol, in which training was performed using the initial 70% of each session recording while testing used exclusively subsequent data, avoiding cross-validation and eliminating future information leakage. Second, the evaluation was extended beyond conventional classification performance by incorporating metrics specifically designed to characterize online usability.

Each subject completed two same-day recording sessions, which were processed independently under the causal within-session protocol described above, yielding one train/test fold per session. For each subject, the resulting performance metrics were averaged across the two sessions to obtain a single subject-level estimate, which was then used as the unit of observation in the mixed-effects models

D. Decoding pipelines

Five decoding pipelines were evaluated to investigate the relative influence of decoder architecture and temporal processing.

The baseline pipeline consisted of Common Spatial Patterns (CSP) followed by Linear Discriminant Analysis (CSP+LDA), representing one of the most widely adopted approaches for motor imagery BCIs.

To isolate the effect of the classifier, CSP features were also combined with both linear and radial basis function (RBF) Support Vector Machines (CSP+SVM).

Finally, covariance matrices were projected onto the Riemannian tangent space prior to classification using linear and RBF SVMs (TGSP+SVM). By exploiting the geometry of covariance matrices, tangent-space representations provide richer descriptions of sensorimotor activity than conventional CSP features [14].

Classifier hyperparameters were optimized independently for each pipeline using a randomized search with internal k-fold cross-validation on the training set, without any test data, maximizing the offline normalized Matthews Correlation Coefficient (nMCC). Selected hyperparameters include number of selected CSP components, ranging from 4 to 8; SVM regularization, ranging from 0.1 to 10; and LDA shrinkage coefficient, ranging from 0.1 to 0.5 and also comparing the value obtained from Ledoit-Wolf lemma, as implemented in *sklearn* package.

E. Experimental factors

The objective of this study is not only to compare decoding pipelines, but to investigate how different design choices influence pseudo-online behavior.

Three experimental factors were considered. Decoder architecture, represented by the evaluated pipelines. Window length, ranging from 0.5s to 2.0s in 0.5s intervals, corresponding to the temporal interval integrated before each prediction. Temporal sampling density, or step size, namely 0.15s, 0.3s and 0.5s, controlled through the step between consecutive windows. The step size controls overlap between windows and increasing overlap produces more frequent predictions while maintaining the same window duration.

By varying these factors independently, the proposed analysis enables the investigation of how temporal processing and decoder architecture contribute to different dimensions of system performance.

F. Performance Metrics

Performance was evaluated using complementary metrics describing different aspects of pseudo-online usability.

1) Classification performance: Classification quality was measured using the normalized Matthews Correlation Coefficient (nMCC). Unlike conventional accuracy, nMCC is robust to moderate class imbalance and provides more information as it summarizes all four elements of the confusion matrix, which is particularly important due to the inclusion of the additional rest class [15].

2) Detection latency: Detection latency was defined as the elapsed time between cue onset and the first correctly classified window. This metric represents the responsiveness perceived by the user during online operation.

3) Prediction confidence: Prediction confidence was computed as the posterior probability assigned to the predicted class. Higher confidence indicates greater certainty in the generated decisions.

4) Temporal stability: Rather than evaluating isolated correct predictions, temporal stability was quantified from continuous blocks of consecutive correct classifications.

Predictions were first partitioned into maximal contiguous blocks according to both trial identity and classification correctness. Consecutive correctly classified windows belonging to the same trial were grouped into a single stability block. For each block, its proportion relative to the entire trial was computed.

G. Statistical Analysis

The influence of decoder architecture and temporal processing on each performance metric was investigated using linear mixed-effects models.

For every dependent variable, fixed effects included decoder architecture, window length, step between windows, and all interaction terms. Subject identity was included as a random intercept to account for repeated observations within participants.

This modeling strategy allows the independent contribution of each design parameter to be quantified while accounting for inter-subject variability, enabling the identification of which experimental factors govern different dimensions of pseudo-online behavior.

Residual diagnostics were inspected to assess model adequacy; while some deviations from normality and homoscedasticity were observed, model estimates were considered robust given the large sample sizes involved. Reported effects were also verified against Benjamini-Hochberg false discovery rate correction for multiple comparisons. All findings remained significant after this correction, with the exception of the CSP+SVM (RBF) effect on nMCC ($p = 0.015$, $p_FDR = 0.063$), which is interpreted with corresponding caution.

III. Results

A. Responsiveness

The first objective of this study was to determine how different design choices influence system responsiveness. Contrary to the common assumption that response time is governed solely by temporal windowing, mixed-effects modeling revealed that decoder architecture also plays a critical role.

Using the conventional CSP+LDA pipeline as the reference, all Support Vector Machine (SVM)-based approaches significantly reduced the time to the first correct detection (all $p < 0.001$). The largest reductions were observed for CSP+SVM with RBF kernel and TGSP with linear kernel ($\beta = -0.71$ and $\beta = -0.73$, respectively), while TGSP with RBF SVM and CSP with linear SVM showed a smaller reduction ($\beta = -0.55$ and $\beta = -0.68$, respectively) indicating that SVM classifiers consistently produce faster responses regardless of the feature extraction strategy.

Temporal processing parameters also exerted a strong influence on responsiveness. Increasing window length systematically delayed the first correct detection ($\beta = 0.40$, $p < 0.001$), while increasing the temporal step between consecutive windows, which corresponds to reducing the sampling density,

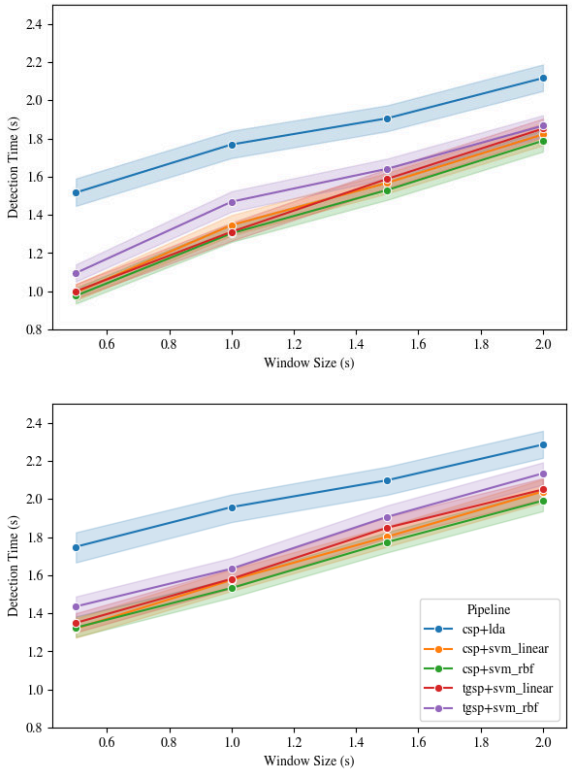


Fig. 1. Effect of window size on system detection time under two temporal sampling rates. Top: window step of 0.15 s (higher sampling density). Bottom: window step of 0.50 s (lower sampling density).

also increased detection latency ($\beta = 0.70$, $p < 0.03$). The interaction between window length and temporal step was not significant indicating that increasing window overlap (reducing step size) does not compensate for the latency penalty introduced by larger windows. Fig. 1 shows the effects of temporal parameters in latency.

B. Decision quality

Unlike temporal sampling density, which primarily affected response dynamics, decision quality was jointly determined by decoder architecture and the amount of neural information integrated before each prediction.

In terms of classification performance, measured by the normalized Matthews Correlation Coefficient (nMCC), SVM-based pipelines consistently outperformed the conventional CSP+LDA approach. Significant improvements were observed for CSP+SVM RBF ($\beta = 0.03$, $p = 0.015$), TGSP+SVM Linear ($\beta = 0.03$, $p = 0.003$), and TGSP+SVM RBF ($\beta = 0.03$, $p = 0.004$). Window length also produced a consistent increase in nMCC across all pipelines ($\beta = 0.04$, $p < 0.001$), whereas temporal sampling density and classifier linearity showed no significant effects. Fig. 2 shows how nMCC is impacted by pipeline and window size given a window step.

Prediction confidence followed a similar pattern. Relative to CSP+LDA, the Riemannian pipelines TGSP+SVM Linear

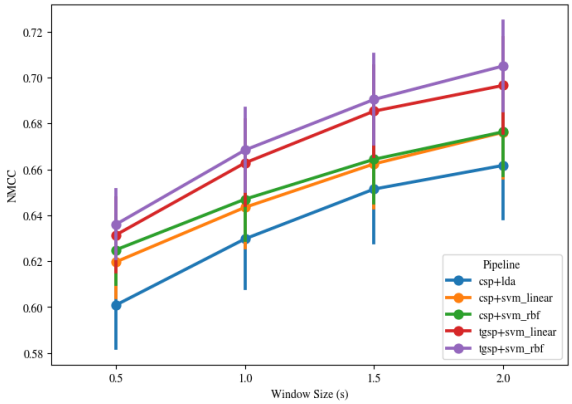


Fig. 2. Effect of window size in nMCC with 0.3s window step.

($\beta = 0.035$, $p < 0.001$) and TGSP+SVM RBF ($\beta = 0.27$, $p < 0.001$) generated substantially higher classification probabilities, indicating more confident decisions. Larger temporal windows also increased prediction confidence ($\beta = 0.04$, $p < 0.001$). Sparser temporal sampling (bigger temporal steps) had significant impact only on tangent space based classification, reducing its confidence, particularly for TGSP+SVM RBF ($\beta = -0.62$, $p < 0.001$). This effect was partially mitigated by increasing window length ($\beta = 0.35$, $p < 0.001$), demonstrating that the benefits of richer feature representations depend on an appropriate temporal parameterization. Fig. 3 shows the effect of window size and pipeline on classification probability.

C. Temporal stability

Unlike responsiveness and decision quality, temporal stability was determined almost exclusively by the temporal organization of signal processing.

Temporal stability was quantified as the proportion of each trial occupied by contiguous blocks of correct classifications, providing a measure of how consistently the decoder maintained correct decisions over time. Mixed-effects analysis revealed no significant differences among any of the evaluated decoding pipelines or classifier architectures (all $p > 0.74$), indicating that improving the classifier alone does not substantially increase the continuity of correct predictions.

Instead, stability was strongly governed by the temporal processing parameters. Increasing window length significantly increased the proportion of sustained correct classification blocks ($\beta = 0.022$, $p < 0.001$), indicating that longer temporal integration favors more persistent decoding. Likewise, increasing step between windows also produced a substantial improvement in temporal stability ($\beta = 0.092$, $p < 0.001$), suggesting that sparser temporal sampling promotes longer uninterrupted periods of correct classification.

Importantly, no significant interactions were observed between temporal parameters and decoder architecture, indicating that these improvements were consistent across all evaluated pipelines.

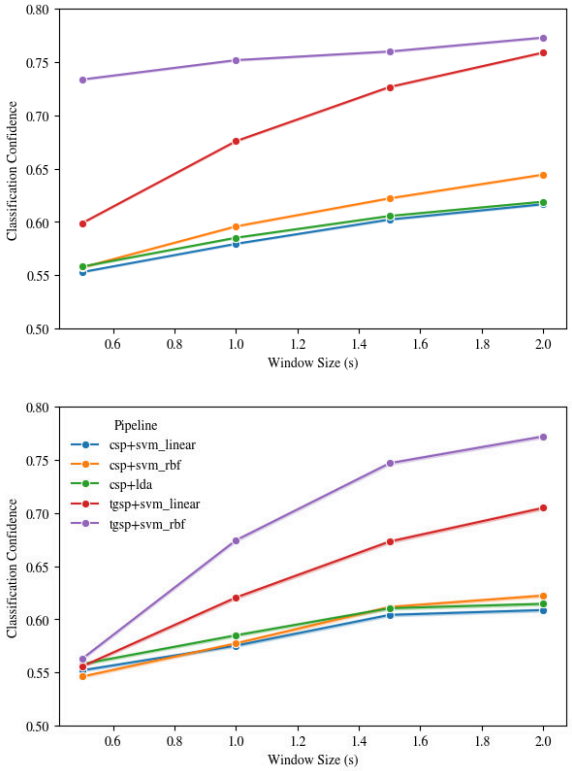


Fig. 3. Effect of window size on classification confidence under two temporal sampling rates. Top: window step of 0.15 s (higher sampling density). Bottom: window step of 0.50 s (lower sampling density).

D. Decision and responsiveness trade-off

The combined analysis of these results demonstrates that configuring a pseudo-online BCI requires balancing competing objectives rather than optimizing a single performance metric.

Maximizing decision quality—expressed by higher nMCC values and promoting earlier convergence to stable decoding both require longer temporal integration windows. However, these same windows substantially increase detection latency.

Conversely, increasing window overlap accelerates system responses by producing more frequent predictions, but delays the emergence of stable decoding and reduces the confidence of the Riemannian decoder. These competing effects define the Pareto frontier shown in Fig. 4., illustrating that improvements in one performance dimension generally require compromises in another. Each point in the analysis represents the average performance of a unique combination of classification pipeline, window size, and window step across all subjects. A configuration was considered Pareto-optimal if no other configuration simultaneously achieved a higher nMCC and a lower mean latency.

Table I summarizes the Pareto-optimal configurations considering the trade-off between responsiveness and classification performance (nMCC). Interestingly, all Pareto-optimal solutions employ the smallest evaluated window step (0.15 s), indicating that a high temporal sampling density is required to achieve the best latency–accuracy compromise. Although

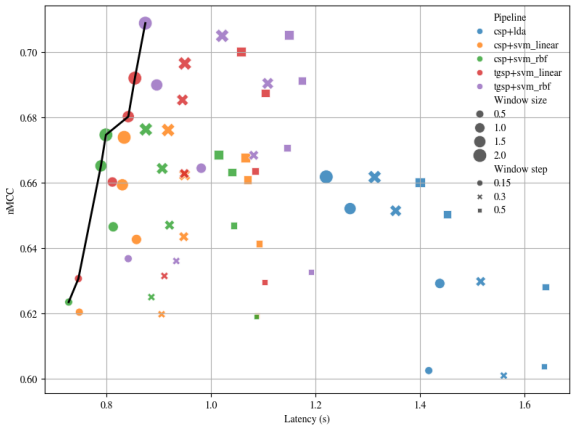


Fig. 4. Pareto frontier for conflicting objectives: nMCC and latency.

increasing the sampling density does not significantly improve nMCC by itself, it consistently reduces detection latency, allowing these configurations to dominate alternatives with larger temporal steps.

The frontier also illustrates that no single pipeline dominates the entire operating range. CSP+SVM (RBF) provides the fastest operating points, whereas TGSP-based pipelines progressively replace CSP-based solutions as larger temporal windows become available, yielding higher classification performance at the expense of increased response time. In particular, TGSP+SVM (RBF) achieves the highest nMCC on the frontier, but only with the longest evaluated window (2.0 s), highlighting that the most expressive decoder also requires the largest amount of temporal information before producing reliable predictions.

Therefore, rather than identifying a universally optimal decoder, the Pareto frontier defines a family of optimal operating points from which the appropriate configuration should be selected according to the latency constraints and accuracy requirements of the intended application.

Importantly, the Pareto frontier only considers latency and classification performance. Other metrics analyzed in this work, such as prediction confidence and temporal stability, exhibit different dependencies on temporal parameterization. Consequently, a configuration that is Pareto-optimal in the latency–accuracy space is not necessarily optimal when additional performance dimensions are considered, reinforcing that pseudo-online BCI design should be formulated as a genuinely multiobjective optimization problem.

E. Decoder error profile

The error distribution of each decoding pipeline provides critical insight into their operational failure modes under continuous evaluation. Table II summarizes the performance metrics, false positive rate (FPR, active classifications during background rest), and false negative rate (FNR, rest classifications during motor imagery intervals) for a fixed 1.0 s window.

While CSP+LDA exhibits a remarkably low FPR ($11.8\% \pm 6.3\%$), it suffers from an excessively high FNR ($77.4\% \pm$

TABLE I
PARETO-OPTIMAL CONFIGURATIONS CONSIDERING LATENCY AND NMCC AVERAGED OVER SUBJECTS.

Pipeline	Window Size (s)	Window Step (s)	Latency (s)	nMCC
CSP+SVM (RBF)	0.5	0.15	0.728	0.623
CSP+SVM (RBF)	1.5	0.15	0.789	0.665
CSP+SVM (RBF)	2.0	0.15	0.799	0.675
TGSP+SVM (Linear)	0.5	0.15	0.746	0.631
TGSP+SVM (Linear)	1.5	0.15	0.842	0.680
TGSP+SVM (Linear)	2.0	0.15	0.854	0.692
TGSP+SVM (RBF)	2.0	0.15	0.874	0.709

TABLE II
CLASSIFICATION PERFORMANCE, LATENCY, AND ERROR TRADE-OFFS ACROSS EVALUATED PIPELINES (1.0 S WINDOW SIZE)

Pipeline	Window step (s)	nMCC	Latency (s)	FPR	FNR
CSP+LDA	0.15	0.629 ± 0.066	1.438 ± 0.477	0.119 ± 0.063	0.771 ± 0.148
CSP+LDA	0.50	0.628 ± 0.065	1.640 ± 0.458	0.118 ± 0.063	0.774 ± 0.146
CSP+SVM (BOTH)	0.15	0.645 ± 0.053	0.835 ± 0.210	0.570 ± 0.100	0.543 ± 0.125
CSP+SVM (BOTH)	0.50	0.644 ± 0.053	1.068 ± 0.212	0.573 ± 0.100	0.544 ± 0.124
TGSP+SVM (BOTH)	0.15	0.662 ± 0.052	0.896 ± 0.219	0.385 ± 0.119	0.586 ± 0.125
TGSP+SVM (BOTH)	0.50	0.667 ± 0.051	1.115 ± 0.183	0.385 ± 0.102	0.575 ± 0.116

14.6%). This asymmetry reveals a severe class collapse toward the predominant rest state, as the linear discriminant struggles to establish a balanced decision boundary. Consequently, the low latency responsiveness observed in SVM pipelines is absent in CSP+LDA, as transitioning out of the rest state probably requires fully motor imagery windows.

Conversely, combining CSP with SVM alleviates this bias, reducing the FNR to $54.3\% \pm 12.5\%$, but at the expense of the highest false alarm rate across all candidates ($57.0\% \pm 10.0\%$ FPR). Projecting covariance matrices onto the Riemannian tangent space (TGSP+SVM) effectively regularizes this trade-off, retaining a low FNR ($57.5\% \pm 11.6\%$) while substantially curbing spurious activations down to $38.5\% \pm 10.2\%$, achieving the highest overall discriminability ($nMCC = 0.667 \pm 0.051$).

Notably, varying the window step from 0.15 s to 0.50 s systematically penalizes detection latency across all decoders without altering their underlying error profiles (FPR and FNR remain invariant within error margins). This indicates that the operational balance between false alarms and missed intentions is dictated primarily by the feature extraction geometry and classifier loss formulation.

IV. DISCUSSION

The present study demonstrates that pseudo-online BCI performance emerges from the interaction between decoder architecture and temporal processing rather than from the choice of a single decoding pipeline. More importantly, different design parameters influence different dimensions of system behavior, revealing that pseudo-online decoder evaluation should be treated as a multiobjective design problem instead of a conventional classifier comparison.

One of the main findings is the distinct role played by temporal window length and temporal sampling density. Although both parameters define the temporal processing strategy, they affect different aspects of decoder behavior. Larger windows consistently improved classification performance, prediction confidence, and the proportion of contiguous stable blocks, suggesting that integrating longer EEG segments provides a more informative and resilient representation of motor imagery patterns. This improvement, however, comes at the expense of increased response latency, illustrating the classical trade-off between information integration and responsiveness.

Temporal sampling density exhibited a more complex, double-edged behavior. Unlike window length, modifying the overlap by changing step size between consecutive windows had little influence on overall classification performance (nMCC). Instead, increasing the sampling density accelerated system responses but also decreased the proportion of stable blocks, as a denser temporal tracking may produce more confusing windows. Also, sparser sampling introduced computational uncertainty, significantly reducing the prediction confidence of highly expressive decoders (such as TGSP+SVM RBF). This distinction suggests that window length and temporal sampling density should not be regarded as interchangeable parameters, despite both belonging to the temporal processing stage.

The role of decoder architecture was also more nuanced than simply identifying a ”best” classifier. Tangent-space representations combined with SVMs consistently produced higher classification performance and confidence, whereas SVM-based classifiers reduced detection latency relative to the conventional CSP+LDA pipeline. However, decoder architecture had no significant effect on temporal stability (measured as the proportion of continuous correct classifications), indicating that replacing the classifier alone is insufficient to obtain more sustained online behavior. This behavior might be explained by neural dynamics. In any neural task, subject attention may not be fully in the task all the time, making the signal more confusing to classifiers. Also, the rest class introduced in this pseudo-online processing is not homogeneous. It might contain some motor imagery task and many other neural states, making it not very separable from the other classes. Further work should investigate neural representations associated with each

class to provide explanation on temporal stability. Also, many other decoder architectures could be used in order to achieve higher stability, such as ones possessing memory (e.g recurrent neural networks). However, this work suggests that stable decoding is mostly determined by temporal integration rather than by the discriminative power of a memoryless classifier.

In continuous asynchronous operation, the inclusion of a rest class exposes critical vulnerabilities that remain concealed under balanced, epoch-based evaluations. The conventional CSP+LDA pipeline exhibits an apparent selectivity (FPR $\approx 12\%$) that in reality represents a failure mode as the classifier collapses toward the majority rest class, rejecting active motor imagery commands at a high rate (FNR $\approx 77\%$). Mitigating this collapse by pairing CSP with SVM architectures accelerates responsiveness, but introduces an operational penalty of over-activation (FPR $\approx 57\%$). In functional settings such as robotic navigation or prosthetic manipulation, a spontaneous false activation every other window can lead to safety hazards and user frustration. In contrast, projecting spatial covariance matrices onto the Riemannian tangent space (TGSP+SVM) effectively regularizes the decision boundary, reducing false activations to 38.5% while sustaining high sensitivity. This confirms that capturing the intrinsic geometry of EEG signal manifolds provides richer class separability without succumbing to majority-class attraction.

Although smaller step sizes inevitably introduce strong serial correlation among contiguous evaluation epochs, the marginal error rates (FPR and FNR) remain virtually unchanged within each pipeline family. Crucially, the invariance of FPR and FNR across different temporal steps demonstrates an architectural decoupling as the operational bias between missed detections and false alarms is determined by feature extraction and classifier, whereas temporal sampling density strictly dictates the latency with which transitions are registered.

Because performance metrics were averaged across each subject's two same-day sessions prior to statistical modeling, potential within-subject, between-session variability, such as due to electrode impedance changes, fatigue, or short-term nonstationarity in sensorimotor rhythms, was not explicitly modeled and may have been partially masked by this averaging step. Future analyses could treat session as an additional random-effects level to explicitly quantify this source of variability and its relationship with signal nonstationarity.

From a practical perspective, these findings reinforce that different BCI applications require different operating points. Applications demanding rapid interaction, such as assistive device control, may benefit from shorter windows and denser temporal sampling despite a moderate reduction in decision quality and confidence. Conversely, neurorehabilitation protocols, where reliable and sustained feedback is generally more important than minimum response time, may benefit from longer temporal windows that favor confidence and stability. The Pareto analysis illustrates that these objectives cannot generally be optimized simultaneously, emphasizing that decoder configuration should be guided by application-specific priorities rather than by offline accuracy alone.

Finally, this study is limited by its pseudo-online nature. Although the proposed analysis reproduces causal processing and continuous decision making more faithfully than conventional offline evaluation, it does not capture user adaptation to decoder feedback, which is an inherent characteristic of fully closed-loop BCIs. Future work should therefore validate whether the design principles identified here remain consistent during real-time experiments and across different motor imagery paradigms and clinical populations.

Ultimately, rather than identifying a universally optimal decoder, these results provide a principled way to navigate the design space of pseudo-online BCIs by explicitly relating design parameters to different dimensions of online usability.

V. Conclusion

This study demonstrates that pseudo-online BCI performance cannot be adequately characterized by offline classification accuracy alone. Instead, practical system behavior emerges from the combined influence of decoder architecture and temporal processing, with each design parameter governing distinct dimensions of online usability.

Our results show that decoder architecture primarily determines the quality of neural decisions. In contrast, temporal processing plays a broader role. Window length influences every evaluated performance dimension, improving classification performance, confidence, and temporal stability while inevitably increasing detection latency. Temporal sampling density, on the other hand, primarily affects the dynamics of decoder operation, accelerating system responses at the expense of temporal stability and, for Riemannian decoders, prediction confidence.

These findings demonstrate that configuring a pseudo-online BCI inherently involves balancing competing objectives. Improving one performance dimension generally requires compromises in another, indicating that no single configuration simultaneously optimizes responsiveness, decision quality, confidence, and stability. Consequently, decoder selection should not be formulated as the search for the most accurate offline model, but rather as a multiobjective design problem whose optimal solution depends on the intended application.

The adapted pseudo-online evaluation protocol employed here offers a design-oriented methodology for characterizing these trade-offs under realistic causal processing conditions. Rather than identifying a universally optimal decoding pipeline, it enables researchers and engineers to understand how different design choices shape online system behavior, supporting the selection of operating points that better match the clinical and functional requirements of applications ranging from rapid assistive control to highly stable neurorehabilitation protocols.

Despite offering a latency-aware and causally constrained perspective, this study is inherently bounded by the pseudo-online nature of its evaluation. Static offline recordings cannot capture real-time user co-adaptation, voluntary strategy shifts,

or the closed-loop neuroplasticity elicited by immediate sensory feedback. Future work should validate whether the design principles and error dynamics identified here hold during fully closed-loop implementations across longitudinal sessions and diverse clinical populations. Additionally, future efforts should pursue the development of composite usability metrics that formally integrate latency, stability and decision into a unified real-time benchmark.

ACKNOWLEDGMENT

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant no. 2024/03628-5. This work was also supported by the Conselho Nacional de Desenvolvimento Científico (CNPq), Universal Call, grant no. 409232/2023-8.

REFERENCES

[1] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012, doi: 10.3390/s120201211.

[2] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz, "Machine-learning-based coadaptive calibration for brain-computer interfaces," *Neural Computation*, vol. 23, no. 3, pp. 791–816, Mar. 2011, doi: 10.1162/NECO_a_00089.

[3] A. Vavoulis, P. Figueiredo, and A. Vourvopoulos, "A Review of Online Classification Performance in Motor Imagery-Based Brain–Computer Interfaces for Stroke Neurorehabilitation," *Signals*, vol. 4, no. 1, pp. 73–86, 2023, doi: 10.3390/signals4010004.

[4] I. Carrara and T. Papadopoulo, "Pseudo-online framework for BCI evaluation: a MOABB perspective using various MI and SSVEP datasets," *Journal of Neural Engineering*, vol. 21, no. 1, 2024, doi: 10.1088/1741-2552/ad171a.

[5] N. Mrachacz-Kersting and S. Aliakbaryhosseinabadi, "Comparison of the efficacy of a real-time and offline associative brain-computer interface," *Frontiers in Neuroscience*, vol. 12, Art. 455, 2018, doi: 10.3389/fnins.2018.00455.

[6] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, Art. 031005, 2018.

[7] A. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] P. Ablin, A. Barachant, A. Bonnet, S. Congedo, and collaborators, "PyRiemann: A Python package for Riemannian geometry in brain-computer interfaces," *Journal of Open Source Software*, vol. 7, no. 76, Art. 4067, 2022.

[9] A. Gramfort *et al.*, "MNE software for processing MEG and EEG data," *NeuroImage*, vol. 86, pp. 446–460, 2014.

[10] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proc. 9th Python in Science Conference (SciPy)*, 2010, pp. 56–61.

[11] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[12] M. L. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, vol. 6, no. 60, Art. 3021, 2021.

[13] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI Competition IV," *Frontiers in Neuroscience*, vol. 6, Art. 55, 2012, doi: 10.3389/fnins.2012.00055.

[14] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain–Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[15] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Art. no. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

Time-frequency characterization of the cVEMP signal using Wavelet Transform in participants with vestibular complaints

1st Geraldo V. de Medeiros Xavier
Edmond and Lily Safra
Int. Inst. of Neuroscience (IIN-ELS)
Santos Dumont Institute (ISD)
Macaíba, Brazil
geraldo.xavier@edu.isd.org.br

2nd Gizele F. F. do Nascimento
Federal University of Rio Grande
do Norte (UFRN), Natal, RN, Brazil
Santos Dumont Institute (ISD)
Macaíba, RN, Brazil
gizele.nascimento@isd.org.br

3rd Caroline C. do Espirito Santo
Edmond and Lily Safra
Int. Inst. of Neuroscience (IIN-ELS)
Santos Dumont Institute (ISD)
Macaíba, RN, Brazil
caroline.santo@isd.org.br

***Abstract*—Cervical Vestibular Evoked Myogenic Potential (cVEMP) recordings were evaluated in five patients with otoneurological symptoms (Experimental Group [EG]; 46.6 $\pm$ 14.88 years) and three healthy subjects (Control Group [CG]; 26 $\pm$ 7.21 years) using 500 Hz tone-burst stimuli at 95 dB nHL. Unfiltered signals were processed in Python to evaluate P13 and N23 peak latencies, peak-to-peak amplitudes, and pre-stimulus background EMG RMS. Continuous Wavelet Transform (CWT) was implemented for time-frequency signal decomposition, assessing multi-scale energy distribution across slow, transition, and fast muscle fiber bands. The CG demonstrated larger P13–N23 amplitudes and shorter latencies compared to the EG, with attenuated EG responses attributed to sacculocollic pathway hyporesponsiveness rather than reduced tonic muscle contraction. Wavelet scalograms revealed distinct time-frequency patterns: CG participants exhibited concentrated low-frequency energy (25–50 Hz) at 20–35 ms, whereas the EG displayed signal desynchronization, prolonged inhibitory bands, elevated fast-fiber energy distribution ($p = 0.0007, r = 0.358$), and significantly lower dominant frequencies ($p = 0.0001$). These findings suggest vestibulocollic reflex dysregulation in patients with vestibular complaints and demonstrate that Wavelet-based spectral analysis effectively complements conventional cVEMP assessment by resolving temporal and spectral dynamics obscured in standard time-domain averaged waveforms.**

***Index Terms*—Cervical Vestibular Evoked Myogenic Potentials (cVEMP), Wavelet Transform, Time-Frequency Analysis, otoneurology.**

I. Introduction

Vestibular function assessment relies on methods capable of objectively characterizing the integrity of the sensory pathways involved in balance control. The vestibular system, specifically its otolithic component, plays a role in detecting linear acceleration and maintaining postural stability[1]. Alterations in this system are associated with symptoms such as vertigo, postural instability, and oculomotor deficits, which are frequently observed in various neurological and vestibular conditions[2].

Cervical Vestibular Evoked Myogenic Potential (cVEMP) is used to assess the vestibulocollic pathway, specifically the function of the saccule and the inferior vestibular nerve. The signal is elicited via acoustic or vibratory stimuli and recorded using surface electromyography (sEMG) of the sternocleidomastoid (SCM) muscle. It reflects an inhibitory vestibulocollic reflex, manifesting as a transient reduction in SCM tonic activity[3]. The P13 and N23 potentials correspond to a transient suppression of SCM tonic activity resulting from a decrease in the firing rate of a subset of alpha motoneurons.

In clinical practice, cVEMP interpretation relies on time-domain parameters, such as P13 and N23 wave latencies and peak-to-peak amplitude[2]. Conversely, sEMG spectral information, such as frequency (Hz, is not utilized in conventional cVEMP clinical metrics. Frequency-domain parameters are influenced by inter-subject variability (e.g., age, sex, clinical condition and muscle length) and the level of tonic muscle contraction required during the test, all of which can induce subtle changes in signal dynamics[4].

Furthermore, individuals with vestibular complaints may exhibit responses within normal ranges, suggesting that conventional analysis may fail to detect subtle physiological alterations[5]. Such limitations demonstrate the need for methods capable of simultaneously evaluating the temporal and spectral information of the signal while preserving local variations relevant to its characterization. The Wavelet Transform enables the time-frequency decomposition of non-stationary signals, allowing for the extraction of coefficients associated with transient components that are not captured by global time-domain metrics[6, 7]—unlike the Fourier Transform or Fast Fourier Transform, which do not provide information regarding the specific time instants at which these frequencies occur[8].

In this context, applying the Wavelet Transform to raw cVEMP EMG signals can enhance the characterization of the vestibulospinal response by enabling the identification of subtle changes in the temporal dynamics of SCM activity, including modifications in the onset, duration, and evolution of the inhibitory response, that might go undetected by conventional latency (P13 and N23) and amplitude measurements. Despite this potential, this approach remains under-explored in cVEMP analysis.

Therefore, this study proposed the application of Wavelet

Transform to the raw cVEMP signal in patients with otoneurological complaints to explore the time-frequency feature extraction of the signal as a complementary tool to conventional analysis.

II. Materials and methods

This study was approved by the Research Ethics Committee of the Alberto Santos Dumont Institute of Education and Research (CEP/ISD), under number 7.890.476. The analyzed data were derived from routine diagnostic procedures performed at the clinical facility. This pilot observational study used a convenience sample of eight participants, who were assigned to two groups: an experimental group (EG) and a control group (CG).

A. Sample characterization

The EG consisted of five participants: three females and two males, aged 26 to 65 years (mean: 46.6 ± 14.88 years). The main complaints reported included dizziness, noise-induced vertigo, imbalance, and difficulties with activities involving linear acceleration and vertical body movement, such as climbing and descending stairs, using elevators, and traveling in vehicles. The CG comprised two females and one male, aged 20 to 32 years (mean: 26 ± 7.21 years), with no otological complaints.

B. Test execution and data acquisition

The tests were performed using an EP25-Eclipse system (Interacoustics®). The participants remained seated, and to activate the target SCM, all individuals were instructed to rotate their heads contralaterally to the muscle being tested. Muscle contraction levels were monitored using electromyographic biofeedback provided by the cVEMP module, targeting electromyographic activity between 50 and 150 μV[2]. Electrodes were positioned bilaterally: active electrodes on the middle third of the SCM, reference electrode at the sternoclavicular joint, and ground electrode on the forehead[2, 9]. cVEMP responses were recorded via air conduction using ER-3A insert earphones, employing a 500 Hz tone-burst stimulus (2-0-2 cycles) at a presentation rate of 5.1 stimuli per second and an intensity of 95 dB nHL[2]. Data acquisition was performed within a -20 to 80 ms time window, using a 10 Hz high-pass filter (6 dB/octave) and a 1 kHz low-pass filter, with an average of 200 stimuli[2, 9]. Two independent and consecutive runs were performed for each ear, labeled L and L1 for the left ear and R and R1 for the right ear. Within each run, four responses were recorded: two ipsilateral (IPSI A and IPSI B) and two contralateral responses (CONTRA A and CONTRA B). Point-by-point averaging was applied to combine IPSI A with IPSI B and, CONTRA A with CONTRA B, yielding one representative ipsilateral and one contralateral response per run. These repeated acquisitions were conducted to assess the trace consistency and potential reproducibility. The cVEMP amplitudes were normalized to the background electromyographic (EMG) activity using the equipment software. The normalized amplitude was calculated as the ratio between the raw peak-to-peak response amplitude and the root mean square (RMS) value of the baseline muscle EMG activity, according to the following equation:

$$\text{Normalized Amplitude} = \frac{\text{Raw Peak-to-Peak Amplitude } (\mu\text{V})}{\text{Baseline EMG RMS } (\mu\text{V})} \quad (1)$$

This calculation is used to minimize the influence of variations in muscle contraction levels on the recorded amplitudes. Conventional cVEMP parameters were analyzed from the waveforms, including the P13 and N23 peak latencies and peak-to-peak amplitude.

C. Data preprocessing

The cVEMP recordings were exported via the EP25 system's Research module in XML format and subsequently processed in a computational environment. Signals were acquired at a sampling rate of approximately 3.000 Hz with a 100-ms recording window, resulting in 467 samples per segment. The signals were organized and exported to a spreadsheet format (Microsoft Excel®) for subsequent processing and analysis.

The time axis of each response was reconstructed based on the sampling frequency and number of pre-stimulus samples specified in the XML file. Following the consolidation of responses A and B, as previously described, the ipsilateral and contralateral waveforms for each averaged trace were retained for subsequent processing steps. Baseline correction was performed by subtracting the signal median calculated over the pre-stimulus window from -20 to -5 ms. The signals were arranged in a standardized structure containing the time axis and the corresponding ipsilateral and contralateral waveforms for acquisitions L, L1, R, and R1.

For each participant, the parameter values extracted from acquisitions L, L1, R, and R1 were subsequently consolidated by computing their mean, yielding a single final value for each analyzed parameter.

D. Data processing

The cVEMP signals were processed using a Python code executed in the Google Colab environment. XML files exported by the equipment were automatically imported, and the two recording replications corresponding to the IPSI A Raw and IPSI B Raw vectors were extracted.

The time axis was reconstructed based on the sampling frequency and the number of pre-stimulus samples recorded in each file. Subsequently, the average waveform of the two replications was calculated, while the individual signals were preserved to assess response reproducibility. Baseline correction was performed by subtracting the signal median within the pre-stimulus interval of -20 to -5 ms.

Polarity orientation was defined according to the expected cVEMP morphology, with P13 designated as a positive deflection and N23 as a negative deflection. Component identification was conducted using the "find peaks" function. For the participants, the search was conducted within broad windows: 8 to 20 ms for P13 and 14 to 32 ms for N23.

These windows allowed for the identification of responses with latency variations, while normative intervals of approximately 13 to 16 ms for P13 and 21 to 25 ms for N23 were considered for the clinical interpretation of the results[9].

Within each window, the algorithm identified local extrema based on a minimum distance of 2 ms between peaks and a prominence criterion that was adjusted to signal variability. To identify N23, the component was required to occur after P13, with an interpeak interval of 3 to 20 ms. When multiple peaks were identified within the window, the peak with the highest combination of amplitude and prominence was selected. In the absence of a peak meeting these criteria, the extremum with the greatest magnitude within the window was selected. After determining the latencies, the P13 and N23 amplitudes were obtained at the respective signal points. The peak-to-peak amplitude was calculated as the difference between the P13 and N23 amplitudes.

The response was classified as absent when it was not possible to identify a P13–N23 complex with a positive amplitude and appropriate temporal ordering. The pre-stimulus RMS value provided by the equipment was analyzed separately as an indicator of SCM muscle contraction level during the examination.

For raw signal processing, the Wavelet Transform was applied to decompose the signal in the time-frequency domain and characterize its multiscale structure, as represented by the following expression:

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt \qquad (2)$$

where x(t) represents the analyzed signal, ψ^* denotes the complex conjugate of the wavelet function, a corresponds to the scale parameter, and b indicates the temporal position.

This approach allowed for the observation of energy distribution over time and across the different frequency bands of the cVEMP. The acquired data were processed individually, enabling an assessment of response consistency both within and across participants, as well as the identification of variations in the morphology and pattern of the evoked response.

III. Results and discussion

Two trials were performed for each acquisition (ipsilateral and contralateral), resulting in four trials per ear. The results of the conventional cVEMP analysis are presented in Table 1, which shows the P13 and N23 latency values and peak-to-peak amplitudes obtained from the sample.

The CG exhibited shorter latencies and larger amplitudes of the P13–N23 complex, particularly in the right ear. In the same ear, three of the five participants in the EG showed an absent response; consequently, mean latency and amplitude values were calculated using only the valid responses. In the left ear, latencies were similar between the groups, whereas amplitudes remained larger in the CG.

The lower amplitudes observed in the EG do not appear to stem solely from a low level of muscle contraction, given that the mean pre-stimulus RMS values were similar across groups and fell within the baseline range for cVEMP testing. This finding may be explained by the reduced responsiveness of the sacculocollic pathway in the EG, potentially involving differences in the excitability of saccular sensory cells, conduction via the inferior vestibular nerve, or reflex activation of the SCM muscle, a hypothesis further supported by the higher number of absent responses in the right ear.

TABLE I
Results of the conventional cVEMP analysis for the evaluated participants, divided into the experimental group (EG) and CG (CG).

Right Ear				
	EG (n=5)		**CG (n=3)**	
	Mean	**SD**	**Mean**	**SD**
P13 Latency (ms)	20.33	± 1.41	15.89	± 0.51
P13 Amplitude (μV)	12.42	± 5.69	48.65	± 22.43
N23 Latency (ms)	30.50	± 0.71	25.33	± 1.76
N23 Amplitude (μV)	−7.14	± 1.16	−64.16	± 15.51
Peak-to-Peak Amplitude P13–N23 (μV)	19.56	± 6.84	112.81	± 35.25
Pre-stimulus RMS (μV)	82.00	± 28.94	87.33	± 12.86
Left Ear				
	EG (n=5)		**CG (n=3)**	
	Mean	**SD**	**Mean**	**SD**
P13 Latency (ms)	18.87	± 3.69	18.44	± 2.41
P13 Amplitude (μV)	8.58	± 12.47	39.20	± 38.65
N23 Latency (ms)	26.40	± 2.51	27.22	± 3.34
N23 Amplitude (μV)	−27.80	± 17.52	−49.98	± 67.79
Peak-to-Peak Amplitude P13–N23 (μV)	36.38	± 19.07	89.18	± 98.29
Pre-stimulus RMS (μV)	86.00	± 26.90	95.67	± 14.57

Data are presented as mean ± standard deviation (SD). **EG**: Experimental Group; **CG**: CG; **cVEMP**: Cervical Vestibular Evoked Myogenic Potential; **n**: number of participants ; **SD**: Standard Deviation; **ms**: milliseconds; **μV**: microvolts; **RMS**: Root Mean Square.

N23 latency in the left ear showed a different pattern than that observed in the right ear, being slightly longer in the CG. However, the small difference between means and the data variability suggest that this result may represent sampling variation rather than indicating vestibular dysfunction in CG participants, especially given their lack of comorbidities or vestibular complaints.

The Wavelet Transform, when applied to sEMG analysis, allows for the assessment of energy distribution across different frequency bands, providing insights into muscle recruitment patterns. Frequencies below 100 Hz are known to be associated with the combined activity of slow-twitch (Type I) and fast-twitch (Type II) muscle fibers [8]. Thus, this analysis can serve as an indirect indicator of motor unit recruitment patterns. Characterizing these patterns has been employed in the study of motor tasks such as walking and running [8]. Time-frequency analysis of the signal obtained during cVEMP testing can contribute to the assessment of motor fiber recruitment in various vestibular pathologies; in the future, alterations observed in spectral distribution and muscle recruitment dynamics could be investigated as potential functional biomarkers for different clinical conditions.

Furthermore, individual analysis of the recordings allows for the examination of intra- and inter-individual variability, considering aspects such as amplitude, morphology, energy distribution, coherence between replications, and inter-aural differences. Such information is often reduced or obscured

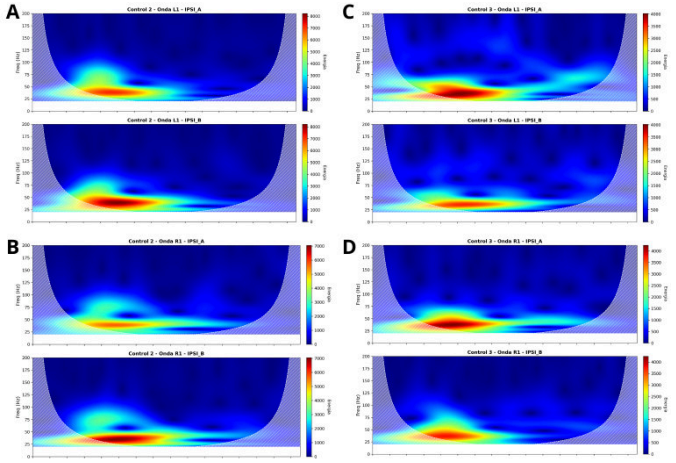


Fig. 1. Comparison of the Wavelet Scalogram across individuals in the CG. **A** – Scalogram of Wave L1 for Participant 2 of the CG; **B** – Scalogram of Wave R1 for Participant 2 of the CG; **C** – Scalogram of Wave L1 for Participant 3 of the CG; **D** – Scalogram of Wave R1 for Participant 3 of the CG.

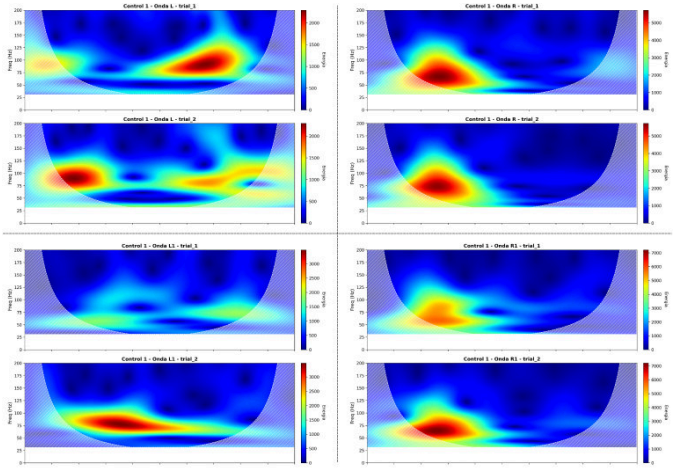


Fig. 2. Wavelet scalogram showing intra-individual comparison for the CG. **A** – L-wave scalogram for CG Participant 1, showing differing responses between the IPSI A and B recordings; **B** – L1-wave scalogram for CG Participant 1; **C** – R-wave scalogram for CG Participant 1, showing the appearance of contralateral responses in both recordings.

when the assessment is limited to the average waveform, as the averaging process attenuates variations present in individual responses[10].

Participants 2 and 3 in the CG exhibited, in recordings L, L1, R, and R1, a higher concentration of energy between 20 and 35 ms (predominantly in the 25–50 Hz range) in both the IPSI A and IPSI B replications, alongside an absence of response in the contralateral channels. This finding may be related to a greater contribution from motor units with slower conduction characteristics (10–90 Hz), as shown in Figure 1.

For CG Participant 1, acquisition L showed temporal inconsistency between the IPSI A and IPSI B replications (Figure 2A). Although both displayed regions of high energy, desynchronization was observed in the timing of their responses. In the IPSI A replication, the highest-intensity activity occurred at a later latency, with energy concentrated predominantly in the 50–65 Hz range. In contrast, the main energy concentration in the IPSI B replication was observed at the beginning of the recording. These differences indicate low temporal and morphological reproducibility between replications, which may be attributable to instability in the level of muscle contraction, differences in motor unit recruitment between acquisitions, or the presence of artifacts not synchronized with the stimulus. In acquisition L1, IPSI A replication showed a reduction in background energy throughout the recording. Conversely, IPSI B replication revealed a continuous horizontal band of high energy concentrated at approximately 50 Hz across the entire analyzed window, suggesting sustained inhibition of the SCM muscle fibers in this frequency range (Figure 2B). In the right-side acquisitions (R and R1), the pattern was similar to that of the other two controls. However, foci of energy were also detected in the contralateral channels, occurring at time intervals coinciding with the ipsilateral responses (Figure 2C).

Analysis of the Wavelet scalograms for Participant 1 of the EG revealed transient and prolonged energy patterns, with differences observed between the acquisitions and their respective replications. In acquisition L, the IPSI B replication displayed a phasic pattern characterized by a delimited region of higher energy between approximately 40 and 80 ms, predominantly in the 50–75 Hz range. Conversely, in the IPSI A and IPSI B replications of acquisition L1, the highest energy concentration occurred at the beginning of the recording (Figures 3C and 3D).

In acquisition R1, both ipsilateral replications showed two distinct energy foci (Figures 3A and 3B). Additionally, the IPSI A replication exhibited a continuous, prolonged horizontal band around 50 Hz (Figure 3A), a pattern also observed in Participant 5 of the EG (Figure 3D). This shift in response suggests two possibilities: a continuous inhibitory signal or sequential firing of inhibitory events via the vestibulocollic pathway.

A continuous energy pattern throughout the analyzed interval suggests the sustained inhibition of SCM motor units. Notably, no EG participant reported pain, muscle weakness, or limited cervical mobility, which reduces the likelihood that the observed differences were related to peripheral musculoskeletal alterations. Regarding the sequential firing of inhibitory events, this could manifest as peaks and troughs in conventional cVEMP analysis when observing variations over time.

The findings observed in the EG suggest that the output of the vestibulocollic pathway may not elicit the expected motoneuronal suppression in response to the auditory stimulus. This raises the hypothesis of a dysregulation in the processing of this reflex, which aligns with the primary complaints of patients regarding dizziness in situations requiring postural maintenance against linear acceleration stimuli, a function mediated by the otolithic organs [9].

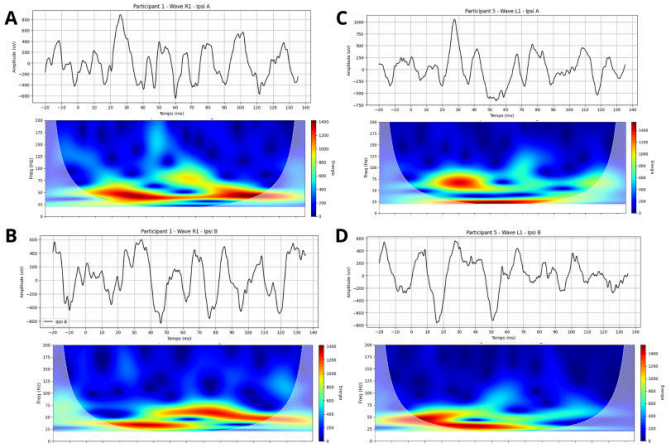


Fig. 3. Comparison between Wavelet Scalogram and conventional intra-individual signals from the Experimental Group. **A** – Raw time-domain cVEMP signal and Wavelet Scalogram of Wave R1 - IPSI A (Participant 1, Experimental Group); **B** – Raw time-domain cVEMP signal and Wavelet Scalogram of Wave R1 - IPSI B (Participant 1, Experimental Group); **C** – Raw time-domain cVEMP signal and Wavelet Scalogram of Wave L1 - IPSI A (Participant 5, Experimental Group); **D** – Raw time-domain cVEMP signal and Wavelet Scalogram of Wave L1 - IPSI B (Participant 5, Experimental Group).

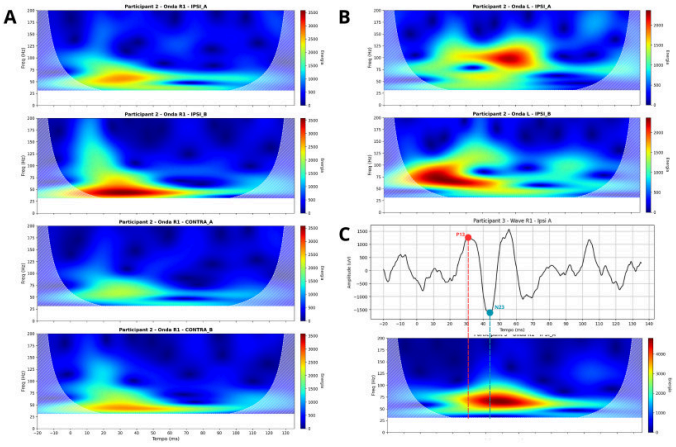


Fig. 4. Conventional signal results and wavelet scalogram for Participant 2 of the GE group. **A** – Wavelet scalogram of the R1 wave for Participant 2 of the GE group, showing contralateral responses; **B** – Wavelet scalogram of the L wave for Participant 2 of the GE group, showing responses above 75 Hz; **C** – Comparison between the raw signal in the frequency domain and the wavelet scalogram.

One argument that could be raised to challenge these assumptions is the presence of various peaks and troughs between -20 and 0 ms that are not clearly defined in the wavelet analysis. However, no sound was emitted during this period. Therefore, it was impossible to trigger a vestibular response. From t = 0 ms, coinciding with the onset of the auditory stimulus, there was an increase in wave amplitude and the appearance of energy clusters, occurring immediately after the triggering of the initial P1 response at the 10 ms mark (Figure 3).

Participant 2 stood out due to the appearance of frequencies above 75 Hz for the L-wave (IPSI A and B; Figure 4B) and contralateral activation for the R and R1 waves (Figure 4A). Participant 3 demonstrated a pattern closer to normal on the left side, with a longer latency for the appearance of the energy focus in the right ear and frequencies above 50 Hz. This result resembled the conventional analysis of the test, which indicated a delayed latency.

Finally, Participant 4 (data not shown) demonstrated a pattern similar to that of the controls on the left side; on the right side (R), a patch extended from 15 to 110 ms within the 25–50 Hz frequency range for IPSI A at R1, while IPSI B showed a patch extending from -10 to 110 ms, with more distinct energy foci at 15 and 80 ms (also in the 25–50 Hz range) and a contralateral response (CONTRA B) with similar morphology. Quantitative analysis evaluated the percentage of energy in slow (10–60 Hz), transition (60–90 Hz), and fast (90–270 Hz) muscle fibers, based on the classification by Koenig et al. (2018)[1], with an adaptation for transition fibers, in addition to the dominant frequency and peak energy latency. Following the consistency assessment between responses A and B across all conditions, no statistically significant differences were observed (p ¿ 0.05). Consequently, IPSI A and IPSI B responses were averaged, as were CONTRA A and CONTRA B. The parameter values obtained from runs L, L1, R, and R1 were subsequently averaged to yield the final value for each parameter per participant (Table 2). Both groups exhibited a predominance of energy in slow fibers, accounting for over 95% of the total energy, with approximately 4% attributed to transition fibers and less than 1% to fast fibers, demonstrating similar behavior between the sides. In the EG, the mean dominant frequency was 39 Hz in the ipsilateral condition, with a latency of approximately 71 ms, and 25 Hz in the contralateral condition, with a duration near 10 ms. In the CG, similar values were observed between the ipsilateral and contralateral conditions, with dominant frequencies of approximately 39 Hz at 55 ms and 38 Hz at 56 ms, respectively.

Statistical analysis was performed using the Mann–Whitney U test as a non-parametric method to compare groups, given the data distribution. The level of significance was set at 5% (p ¡ 0.05), and effect size r was calculated to assess the magnitude of the observed differences.

Energy allocation analysis across muscle fiber types revealed no statistically significant differences between the EG and CG for slow (U = 1874, p = 0.8228, r = 0.024) or transition fibers (U = 1776, p = 0.4800, r = 0.075). However, a statistically significant difference was observed for fast fibers (U = 2607, p = 0.0007), with a moderate effect size (r = 0.358). Median comparison demonstrated values of 0.77 for the EG and 0.40 for the CG, indicating a higher proportion of energy attributed to fast fibers in the EG. This finding suggests that a higher activation pattern of fast fibers may be associated with otoneurological alterations, analogous to observations in other neuromuscular pathologies[10]. Given that motor fiber activation patterns are precisely coordinated through distinct firing frequencies, disruption of this pattern may serve as

TABLE II
SPECTRAL ENERGY DISTRIBUTION, DOMINANT FREQUENCY, AND LATENCY DERIVED FROM CONTINUOUS WAVELET TRANSFORM.

Group	Condition	Run	Slow Fibers (%)	Transition Fibers (%)	Fast Fibers (%)	Dominant Freq. (Hz)	Latency (ms)
CG ($n = 3$)	IPSI	L	93.62 ± 6.95	5.74 ± 6.39	0.62 ± 0.58	41.74 ± 14.20	62.39 ± 15.22
		L1	95.21 ± 2.84	4.20 ± 2.55	0.58 ± 0.32	39.06 ± 5.90	59.94 ± 21.42
		R	96.44 ± 1.62	3.25 ± 1.42	0.31 ± 0.21	37.14 ± 4.91	51.06 ± 14.45
		R1	97.49 ± 0.95	2.25 ± 0.84	0.26 ± 0.14	37.58 ± 0.43	46.72 ± 3.01
	CONTRA	L	94.19 ± 4.12	4.91 ± 3.46	0.87 ± 0.71	40.33 ± 10.90	65.33 ± 17.49
		L1	95.44 ± 3.35	3.82 ± 2.86	0.72 ± 0.51	38.78 ± 10.08	57.17 ± 15.21
		R	96.29 ± 2.10	3.27 ± 1.86	0.43 ± 0.25	34.75 ± 6.57	56.06 ± 15.82
		R1	96.81 ± 1.31	2.80 ± 1.21	0.39 ± 0.14	37.82 ± 1.23	46.00 ± 0.17
EG ($n = 5$)	IPSI	L	93.48 ± 5.39	5.38 ± 4.58	1.11 ± 0.93	38.28 ± 12.65	67.60 ± 31.65
		L1	96.38 ± 1.66	2.77 ± 1.34	0.83 ± 0.56	31.13 ± 7.05	68.97 ± 17.76
		R	94.69 ± 2.28	4.49 ± 2.02	0.81 ± 0.34	36.22 ± 5.42	64.20 ± 15.58
		R1	96.30 ± 1.91	3.04 ± 1.61	0.65 ± 0.35	38.14 ± 4.34	84.30 ± 34.60
	CONTRA	L	94.78 ± 4.71	4.09 ± 3.64	1.08 ± 1.05	25.98 ± 9.36	92.87 ± 34.70
		L1	97.48 ± 1.60	1.73 ± 1.08	0.75 ± 0.60	16.32 ± 3.96	119.53 ± 27.56
		R	94.33 ± 2.07	4.24 ± 1.49	1.40 ± 0.79	29.03 ± 3.21	83.20 ± 16.01
		R1	96.61 ± 2.64	2.58 ± 2.05	0.79 ± 0.71	27.33 ± 3.64	99.97 ± 24.42

Data are presented as mean $\pm$ standard deviation (SD).
EG: Experimental Group; CG: CG; cVEMP: cervical Vestibular Evoked Myogenic Potential; n: number of participants; Hz: Hertz; ms: milliseconds.

TABLE III
MANN–WHITNEY U TEST FOR ENERGY DISTRIBUTION ACROSS FIBER CATEGORIES

Fiber Category[a]	*U*-Statistic	Effect Size (*r*)	*p*-Value
Slow	1874.0	0.024	0.8228*
Transition	1776.0	0.075	0.4800*
Fast	2607.0	0.358	0.0007**

[a] Fiber frequency classification adapted from Koenig et al. (2018).
* Non-significant ($p \geq 0.05$).
** Statistically significant ($p < 0.05$).

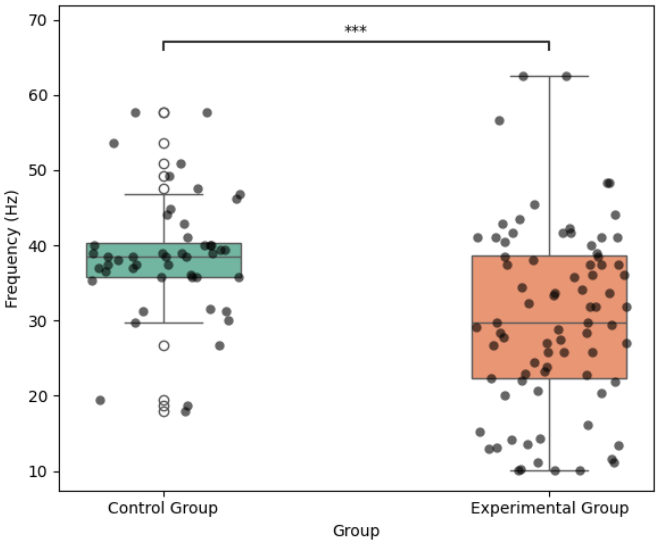


Fig. 5. Boxplot of dominant frequency distribution between the CG (CG) and Experimental Group (EG). Individual data points are overlayed as jittered scatter points. The horizontal line inside each box represents the median frequency. *** p = 0.0001 (Mann–Whitney U test).

a potential biomarker for diagnosing and evaluating clinical neurodegenerative and neuromuscular disorders[9]. Dominant frequency analysis demonstrated a statistically significant difference between the groups (U = 1134.0, p = 0.0001), with the CG exhibiting a higher median (38.46 Hz) than the EG (29.70 Hz). When evaluating the slow and transition frequency bands, the CG also displayed consistently higher medians.

Taken together, the findings regarding dominant frequency and energy percentage across fiber categories revealed a trend toward a higher concentration of frequencies below 90 Hz in the CG, whereas the EG presented a larger proportion of energy attributed to fast fibers. Furthermore, the mean latency of the dominant frequency was prolonged in both ipsilateral and contralateral conditions for the EG compared to the CG. Based on these observations, we hypothesize that individuals without otological complaints exhibit a preferential activation pattern of lower frequencies along with a shorter latency of the dominant frequencies.

The median comparison demonstrated that the EG had a higher energy proportion attributed to fast fibers than the CG, raising the hypothesis of distinct motor unit recruitment patterns between groups. In this context, age-related structural and functional modifications of muscle tissue are well recognized. Previous literature highlights a selective reduction of type II muscle fibers in elderly individuals, while type I fibers remain relatively preserved [10].

This process is closely tied to age-related motor unit remodeling, characterized primarily by the denervation of type II fibers and subsequent collateral reinnervation by motoneurons

supplying type I fibers [10]. However, because our sample was not homogeneous regarding advanced age, aging alone cannot account for the increased fast-fiber activation observed in the EG.

Although statistically significant differences and a moderate effect size for fast fibers were identified, these results should be interpreted with caution. Limitations such as the small sample size, lack of age and sex matching, heterogeneity of otoneurological complaints, and unequal group sizes may influence the findings. Additionally, the absence of a more rigorous methodological design restricts definitive conclusions regarding the observed recruitment patterns.

Consequently, the hypotheses derived from time-frequency characteristics do not allow, at this preliminary stage, for the establishment of specific recruitment profiles or precise clinical associations. Confirmation in future studies with larger, age-matched, and clinically stratified cohorts is required.

Despite these limitations, our findings highlight the potential of time-frequency analysis as a complementary tool in cVEMP signal evaluation. Ongoing research aims to deepen the understanding of time-frequency signal properties and their relationship with conventional cVEMP parameters.

Future investigations featuring larger, clinically stratified cohorts should quantitatively assess the presence and distribution of energy across distinct frequency bands and their correlation with standard cVEMP metrics. Identifying consistent spectral signatures may clarify the utility of time-frequency analysis in characterizing and differentiating otoneurological conditions (e.g., Meniere's Disease, Vestibular Hypofunction, Tinnitus) from central nervous system disorders (e.g., Parkinson's Disease, Multiple Sclerosis) as an adjunct to standard assessment. Furthermore, a relative attenuation of low-frequency components accompanied by a spectral shift toward higher frequency bands across the recording window warrants further investigation regarding transient alterations in sternocleidomastoid motor unit activity during vestibular stimulation.

IV. Conclusion

Time-frequency analysis using the Wavelet Transform revealed differences in energy distribution, temporal organization, and response reproducibility of cVEMPs between participants with vestibular complaints and control subjects. In the EG, greater variability between replications, absent responses, and distinct energy patterns were observed. These findings indicate that Wavelet-based analysis can complement conventional cVEMP assessment by providing additional insights into the signal's temporal and spectral dynamics, which may not be detected solely through averaged waveforms.

Thus, this study offers a complementary approach to cVEMP signal processing, suggesting that time-frequency analysis may aid in identifying response patterns related to alterations in vestibulo-collic pathway modulation and sternocleidomastoid neuromuscular recruitment. However, as this is a preliminary study, further validation in larger cohorts using standardized criteria and quantitative metrics is necessary to confirm its clinical applicability and diagnostic contribution to otoneurology.

References

[1] McDonnell MN, Hillier SL. Vestibular rehabilitation for unilateral peripheral vestibular dysfunction. *Cochrane Database Syst Rev*. 2015; 2015(1): CD005397.

[2] Rosengren SM, Colebatch JG, Young AS, Govender S, Welgampola MS. Vestibular evoked myogenic potentials in practice: Methods, pitfalls and clinical applications. *Clin Neurophysiol Pract*. 2019; 4: 47–68.

[3] Rosengren SM. Effects of muscle contraction on cervical vestibular evoked myogenic potentials in normal subjects. *Clin Neurophysiol*. 2015; 126(11): 2198–2206.

[4] Nguyen KD, Welgampola MS, Carey JP. Test-retest reliability and age-related characteristics of the ocular and cervical vestibular evoked myogenic potential tests. *Otol Neurotol*. 2010; 31(5): 793–802.

[5] Santos CC, et al. Comparação das respostas do VEMP cervical e ocular em indivíduos com e sem doenças otoneurológicas. *Distúrb Comun*. 2020; 32(3): 437–446.

[6] Wu HT. Current state of nonlinear-type time–frequency analysis and applications to high-frequency biomedical signals. *Curr Opin Syst Biol*. 2020; 23: 8–21.

[7] Hecker DJ, Lohscheller J, Schorn B, Litschel R, Weber KP. Electromotive triggering and single sweep analysis of vestibular evoked myogenic potentials (VEMPs). *IEEE Trans Neural Syst Rehabil Eng*. 2014; 22(1): 158–167.

[8] Koenig I, Eichelberger P, Blasimann A, Steiner L, Hauswirth H, Baeyens JP, et al. Wavelet analyses of electromyographic signals derived from lower extremity muscles while walking or running: A systematic review. *PLoS One*. 2018; 13(11): e0206549.

[9] Garcia-Retortillo S, Romero-Gómez C, Ivanov PC. Network of muscle fibers activation facilitates inter-muscular coordination, adapts to fatigue and reflects muscle function. *Commun Biol*. 2023; 6(1): 891.

[10] Miljkovic N, Lim JY, Miljkovic I, Frontera WR. Aging of skeletal muscle fibers. *Ann Rehabil Med*. 2015; 39(2): 155–162.

A Non-Intrusive IoMT Prototype for Duty-Cycle Monitoring of a Laboratory Air Compressor

Vinícius Rodrigues Oviedo*, Lourdes Maria Muraro Favarin*,
Marcos Vinicius Pasqualoto Prior*, Daniel Moro Druzian*, William Leonardo da Silva*

* *Applied Nanomaterials Research Group (GPNAp)*
Franciscan University (UFN)
Santa Maria, RS, Brazil
vinicius.oviedo@ufn.edu.br; lourdes.favarin@ufn.edu.br;
marcos.prior@ufn.edu.br; daniel.moro.druzian@gmail.com; w.silva@ufn.edu.br

***Abstract*—The Internet of Medical Things (IoMT) is driving a profound digital transformation in healthcare, interconnecting medical devices, sensors, and software applications to enable data-driven, proactive asset management in Clinical Engineering. This paper presents a low-cost, non-intrusive IoMT prototype designed to estimate the operational status of a hospital air compressor without altering its original circuitry. This paper presents a low-cost, non-intrusive Internet of Medical Things (IoMT) prototype designed to estimate the operational status of a hospital air compressor without altering its original circuitry. Utilizing an ESP32 microcontroller paired with an MPU6050 accelerometer, the system captures triaxial vibration data and computes Root Mean Square (RMS) values at the edge to classify equipment status as active or idle. Collected data is transmitted via HTTP POST requests to a cloud pipeline comprising Google Apps Script and Google BigQuery, paired with a companion interface for real-time visualization. A preliminary evaluation was conducted on a single FANEM 089/CAL compressor using manually logged start/stop events as ground truth across five observation sessions: three 10-minute sessions for the compressor alone and 10-minute sessions for each of two external interference sources. The results demonstrate that this non-intrusive approach effectively tracks device duty cycles, despite the simplicity of using a single fixed RMS threshold. Overall, this work presents an early-stage architecture and its initial validation, laying the groundwork for future multi-device studies, frequency-domain analysis, and tiny machine-learning-based (tinyML) fault detection.**

***Index Terms*—Clinical Engineering, Internet of Medical Things, medical devices, data-driven**

I. Introduction

The modern healthcare ecosystem is undergoing a profound digital transformation driven by the integration of the Internet of Medical Things (IoMT)—a specialized domain of the Internet of Things (IoT) that encompasses interconnected medical devices, sensors, and software applications designed to communicate within clinical environments [1]. In this context, ensuring the continuous availability and operational reliability of medical equipment is essential to maintaining patient safety and optimizing healthcare workflows [2]–[4].

Currently, Clinical Engineering teams face substantial operational challenges in managing decentralized support assets, such as laboratory air compressors. The lack of automated real-time monitoring hinders precise tracking of actual equipment running hours, forcing reliance on manual, error-prone logs or periodic maintenance schedules that do not reflect true usage patterns [5]. Furthermore, Clinical Engineers must comply with stringent regulatory frameworks that prohibit alterations to the electrical circuitry of certified medical equipment, preventing the adoption of standard current- or power-monitoring solutions. Consequently, key reliability metrics—such as the *Mean Time Between Failures* (MTBF)—are frequently distorted because they rely on rough estimations rather than precise operational data [6]. This lack of visibility limits proactive management, complicates the identification of overloaded or underutilized assets, and elevates the risk of unexpected downtime in critical clinical applications.

To address these limitations, this paper introduces a low-cost telemetry prototype specifically tailored for Clinical Engineering applications. Built around an ESP32 microcontroller and an MPU6050 accelerometer, the system continuously acquires and processes triaxial vibration signals from a laboratory air compressor–used as a testing model. By executing edge-processing algorithms, the device calculates the Root Mean Square (RMS) to estimate the operational state ("*On*" or "*Off*") and dynamically monitor the equipment's duty cycle.

A central design requirement of this architecture is complete non-intrusiveness. By relying exclusively on an external vibration sensor attached to the chassis, the system captures operational indicators without requiring physical or electrical modifications to the device. This approach aligns with regulatory directives, such as ANVISA's frameworks (e.g., RDC 509/2021 [7] and RDC 665/2022 [8]), which restrict changes to an approved device's manufacturing profile. However, non-intrusive mounting does not automatically imply full regulatory compliance: a dedicated risk assessment evaluating stability and potential environmental interference remains necessary prior to clinical deployment. Within these constraints, the proposed system provides automated, per-compressor running-time telemetry to support asset management in decentralized environments.

The acquired data is transmitted via HTTP POST requests to a cloud backend, where Google Apps Script manages ingestion

and streams records into Google BigQuery for persistent storage. Because BigQuery is a general-purpose, SQL-based Data Warehouse, the resulting historical log is compatible with external business-intelligence platforms (e.g., Looker Studio or Power BI), which could be connected in future deployments to provide Clinical Engineers with interactive operational dashboards; the present prototype, however, relies on a lightweight custom web interface for real-time monitoring, described in the Methodology. Ultimately, this paper presents the architectural design and preliminary single-device validation of the proposed prototype, offering an accessible, low-cost solution to track real-time operational cycles of decentralized medical compressors.

II. Methodology

The development of the non-intrusive IoMT telemetry prototype was structured into three primary layers: (*i*) edge hardware data acquisition, (*ii*) cloud-based middleware ingestion, and (*iii*) a data warehouse.

A. Hardware Setup and Non-Intrusive Data Acquisition

The data acquisition node consists of an ESP32 microcontroller integrated with an MPU6050 triaxial accelerometer. The sensor was externally attached, via mechanical coupling (with a double-sided tape) to the chassis of a single FANEM 089/CAL laboratory compressor, a model often used in hospital/clinical settings for vacuum and pressure supply. The sensor attachment did not require any electrical or physical modification of the equipment's original internal circuitry. The system was configured to sample raw acceleration data along the X, Y, and Z axes through an I^2C communication interface at a sampling frequency adjusted via firmware. The proposed prototype is illustrated in Figure 1. To evaluate potential interferences, the same tests were ran with a magnetic stirrer Fisatom (Model 752A, 650 W, 220 V) and a water distiller Gnatus Aquaclean (750 W, 220 V). Both magnetic stirrer and water distiller were placed about 12 cm next to the laboratory compressor, in distinct moments. Hence, manual log and automated responses from the device were obtained.

B. Edge Signal Processing and Feature Extraction

To isolate mechanical vibrations from the static gravitational acceleration, the ESP32 samples triaxial acceleration data at a frequency defined by a 5 ms interval. For each sampling cycle, the system collects a sequence of $N = 60$ readings. The vibration intensity is calculated by determining the magnitude of the acceleration vector (A_{mag}) and subtracting the constant gravitational component ($g = 9.81$ m/s^2). The Root Mean Square (RMS) value is then computed to represent the vibration energy, as defined by (1):

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(A_{mag,k} - g)^2} \tag{1}$$

where $A_{mag,k} = \sqrt{a_{x,k}^2 + a_{y,k}^2 + a_{z,k}^2}$ is the magnitude of the acceleration vector at sample k. This RMS value was compared against an empirically defined operational threshold of 1.195 m/s^2 (Section III-C discusses about this choice). If the calculated RMS exceeds this value, the status is updated to "*Em operacao*"; otherwise, it is flagged as "*Desligado*".

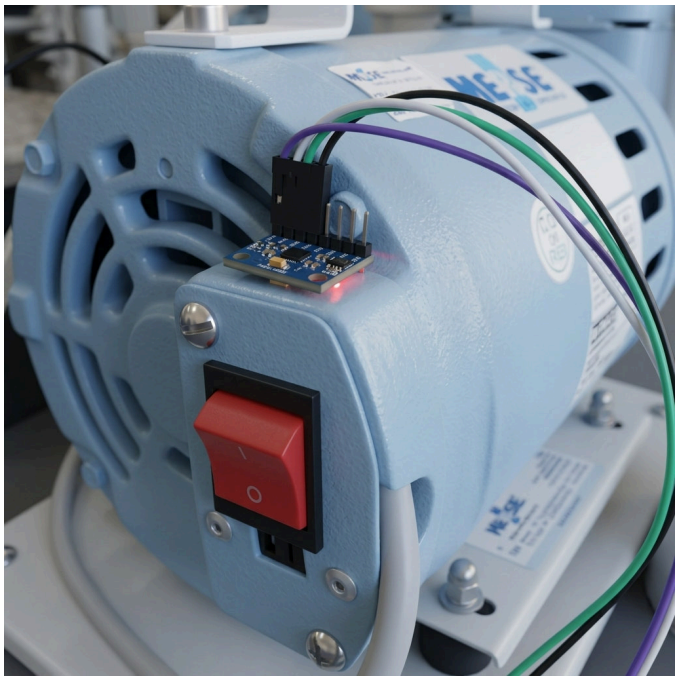


Fig. 1. Proposed IoMT-based prototype.

C. Cloud Ingestion and Data Warehousing Architecture

The ESP32 processes vibration data in batches of 60 samples and transmits the resulting status via an HTTP POST request to the cloud-based architecture, typically achieving a reporting frequency of approximately 2 samples per second. To maintain asset traceability, the system utilizes a relational mapping between the hardware's unique *id_iot*, the equipment's official asset number (*patrimonio*), and its physical location (*localizacao*), so that telemetry data is attributed to the specific compressor unit.

To ensure authenticity and integrity of the telemetry, each payload is signed on the device using HMAC-SHA256 over a canonical string composed of the device identifier, the RMS value, and the operational status, with a pre-shared secret key. The cloud script recomputes this signature upon receipt and discards any payload whose signature does not match, mitigating spoofed or tampered readings. Confidentiality in transit is provided by the HTTPS endpoint of the proposed architecture.

Upon receiving and validating a payload, the cloud script performs two tasks:

- **Real-time Caching:** it stores the raw JSON in the global *CacheService* to feed the client-side user interface, at the full sampling resolution of the device.
- **Historical Logging:** it writes to Google BigQuery (table: *tb_monitoramento_iot*) using an event-driven, throttled

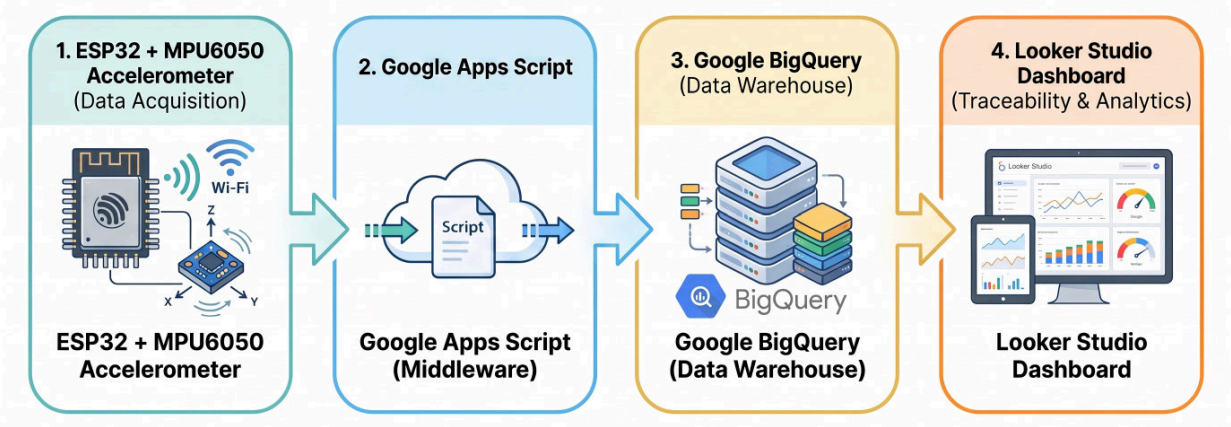


Fig. 2. Architectural framework of the proposed IoMT system.

persistence strategy: a record is written immediately whenever the operational status changes (preserving precise transition timestamps for duty-cycle analysis), and, while the status remains unchanged, at most once every 5 seconds as a heartbeat. This reduces the insertion rate by an order of magnitude compared to logging every sample, while still reconstructing a continuous "on/off" history – including idle periods, which the previous status-gated design did not persist.

The pipeline of the developed prototype is shown in Figure 2. It is important to note that the pre-shared HMAC key is currently embedded in the firmware and manually mirrored in the script's configuration, with no key-rotation mechanism, replay protection, or local buffering strategy for network outages yet implemented. These aspects are discussed further in Section III-C.

D. User Interface

The user interface was developed using HTML5, CSS3, and JavaScript, embedded within the Google Web App environment. The data architecture mapping between the clinical asset and its telemetry records is shown in the Entity-Relationship Diagram (ERD) in Figure 3.

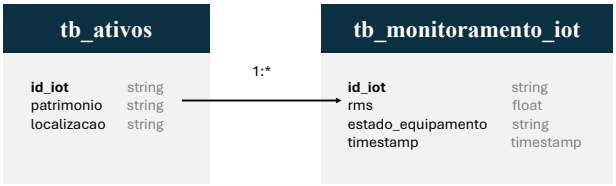


Fig. 3. Entity-Relationship Diagram illustrating the data architecture mapping between the clinical asset and its telemetry monitoring records.

E. Evaluation Protocol

To assess the system's ability to detect compressor operational status, a preliminary evaluation was conducted on the single FANEM 089/CAL unit, as described above. Ground truth was established through a manually recorded log of start and stop events, kept in parallel with the system's automated readings across five observation sessions: three 10-minute sessions for the compressor alone and 10-minute sessions for each of two external interference sources. For each condition, the events were compared between the manual log and the system log to compute sensitivity.

III. Results and Discussion

The proposed IoMT prototype is presented through a web interface (Figure 4) that visualizes the current status of the monitored compressor and its recent vibration telemetry.



Fig. 4. Custom web interface displaying the real-time operational status and RMS vibration reading of the monitored compressor.

TABLE I
RESULTS OBTAINED DURING OPERATIONAL MONITORING

Condition	State	N	Mean	Std	Min	Max	Q1	Q2	Q3
Normal	Off	158	0.456	0.031	0.398	0.582	0.427	0.468	0.477
	On	338	2.127	0.124	1.808	2.923	2.000	2.135	2.218
Magnetic stirrer interference	Off	29	0.446	0.010	0.432	0.492	0.441	0.445	0.447
	On	69	2.197	0.069	2.044	2.351	2.145	2.193	2.240
Water distiller	Off	25	0.442	0.020	0.421	0.520	0.434	0.437	0.520
	On	44	2.252	0.067	2.089	2.373	2.200	2.261	2.373

N: sample count, Std: Standard Deviation, Min: minimum, Max: maximum, Q1: first quartile, Q2: second quartile, Q3: third quartile.

TABLE II
COMPARISON BETWEEN COMMERCIAL IOT PLATFORMS AND THE PROPOSED ARCHITECTURE.

Criterion	Commercial platforms (Blynk, ThingSpeak, Arduino IoT Cloud)	Proposed architecture (ESP32 + BigQuery + custom Web App)
Cost	Low, typically with limited free tiers	Low, variable cost based on data volume (serverless pricing model)
Data ownership	Data hosted on third-party provider infrastructure	Data stored in the institution's own cloud project
Analytical granularity	Pre-configured dashboards, limited customization	Custom SQL queries over raw data (BigQuery), enabling ad-hoc analysis
Integration with hospital systems	Limited, typically via generic APIs	Potential direct integration with hospital BI tools via SQL
Scalability	Adequate for few devices; paid tiers required at scale	Serverless architecture natively scalable to hundreds of assets
Initial setup effort	Low – optimized for rapid prototyping	Moderate – requires cloud project configuration and custom scripting
Security and governance	Depends on provider; limited customization	Requires custom configuration; not yet fully implemented in the current prototype
Maturity / validation	Mature, widely tested platforms	Early-stage prototype, validated in a restricted scope (single device)

A. Operational Monitoring

The statistical summary of the RMS acceleration measurements (m/s^2) obtained by the IoMT monitoring device under various operating conditions is presented in Table I. Across all evaluated scenarios, the proposed system demonstrated a clear and consistent separation between the operational states of the FANEM laboratory air compressor (*Off* vs. *On*).

In the *Normal* condition, the inactive state (*Off*) yielded a mean RMS acceleration of $0.456 \pm 0.031\,\mathrm{m/s^2}$, whereas active compressor operation (*On*) produced a mean RMS acceleration of $2.127 \pm 0.124\,\mathrm{m/s^2}$. The introduction of external vibrational noise sources—specifically a magnetic stirrer and a water distiller—induced negligible variations in the baseline acceleration. Under magnetic stirrer interference, the mean *Off* value for RMS was $0.446 \pm 0.010\,\mathrm{m/s^2}$, while under water distiller operation, it reached $0.442 \pm 0.020\,\mathrm{m/s^2}$. Noteworthy, the recordings displayed a low variability, once coefficient of variation (CV) values ranged from 2–7%.

Crucially, the maximum RMS acceleration recorded across all inactive states remained bounded at $0.582\,\mathrm{m/s^2}$, whereas the minimum acceleration recorded during active compressor operation was $1.808\,\mathrm{m/s^2}$. This established an unpopulated decision margin between $0.582\,\mathrm{m/s^2}$ and $1.808\,\mathrm{m/s^2}$, confirming that external laboratory interference does not trigger false positives and supporting the observed 100% classification sensitivity.

B. Comparison with Commercial IoT Platforms

Table II summarizes how the proposed architecture compares to general-purpose commercial IoT platforms commonly used for rapid prototyping, such as Blynk, ThingSpeak, and Arduino IoT Cloud. The comparison highlights that the main advantage of the proposed architecture lies in data ownership and analytical flexibility, rather than in ease of initial setup, where commercial platforms remain more accessible.

C. Limitations

The present study has several limitations that should guide interpretation of the results and future work.

Scope of validation. The system was evaluated on a single compressor unit (FANEM 089/CAL) over a limited observation period (Section II-E), using manually logged events as ground truth. Also, the current sample size does not capture variability across different compressor models, ages, or mounting conditions. Evaluation across multiple compressors and longer time frames would be interesting before generalizing these results.

Thresholding approach. Operational status was determined by comparing the RMS of vibration against a single, empirically chosen threshold (1.195 m/s^2). Within the tested conditions, this simple approach proved robust: the two external interference sources evaluated (a magnetic stirrer and a water distiller, both placed 12 cm from the compressor) induced only minor shifts in baseline vibration (Table I), and the resulting decision margin remained wide enough that no false activations were observed. This robustness, however, should not be generalized beyond the specific sources, distances, and intensities tested here: a single static threshold calibrated for one installation may not transfer directly to other compressor models, mounting surfaces, or nearby equipment without recalibration. Data loss due to Wi-Fi oscillation or outages was also not evaluated in the present study and is left for future work.

Security and data governance. The cloud ingestion pipeline (Google Apps Script and BigQuery) was implemented as a functional prototype and has not undergone a formal security review. Payload authenticity and integrity were enforced via HMAC-SHA256 signing with a pre-shared key, and confidentiality in transit relied on the Web App's HTTPS endpoint; however, the shared key is embedded in the firmware without a rotation mechanism or secure on-device storage, and the current design does not include replay protection (e.g., nonces or timestamp-based expiry), leaving it potentially vulnerable to replayed payloads. Data-retention policy and local buffering behavior during network outages also remain to be developed and tested before deployment in a production hospital environment, for example.

Regulatory claims. The non-intrusive nature of the sensor attachment is consistent with the intent of ANVISA RDC 665/2022 and RDC 509/2021, in that it avoids modification of the device's original circuitry; however, this work does not constitute a formal regulatory compliance assessment or risk analysis, which would be required before clinical deployment.

IV. Conclusion

This study presented the design and a preliminary, single-device evaluation of a low-cost, non-intrusive IoMT prototype for monitoring the operational status of a laboratory air compressor. By combining edge-based vibration sensing with a cloud-based data pipeline, the proposed architecture avoids any physical or electrical modification to the monitored equipment's circuitry, addressing a practical constraint faced by Clinical Engineering teams operating under strict device-integrity regulations.

Testing on a single FANEM 089/CAL unit over five sessions indicated that the RMS-thresholding approach could distinguish active and idle states with 100% sensitivity. When working next to magnetic stirrer and water distiller, the baseline RMS suffered a little variation, however, such variations did not affect the device's ability to track on/off states of the FANEM unit. These findings suggest that the architecture is technically feasible as a low-cost alternative to manual duty-cycle estimation, while they also showed that a single, static RMS threshold worked well to guarantee operational classification of laboratory air compressor.

Is worthy pointing out that this work should be understood as an initial architectural proof of concept for duty-cycle tracking rather than a validated, fleet-scale operational-monitoring solution. Future work should prioritize: (*i*) multi-device and multi-site data collection; (*ii*) lightweight machine learning (e.g., TinyML) to move beyond simple thresholding; and (*iii*) a formal security and data-governance review of the cloud pipeline, including authentication strength, encryption policy, and retention rules.

Acknowledgment

The authors would like to thank the Franciscan University (UFN) for providing the infrastructure used in this research. This work was also developed within the scope of the telehealth service, linked to the funding project *Tech Futuro*, from the Secretaria de Inovação, Ciência e Tecnologia (process number 23/2500-0000819-7).

References

[1] C. Huang, J. Wang, S. Wang, and Y. Zhang, "Internet of medical things: A systematic review," *Neurocomputing*, vol. 557, pp. 1–18, 2023.

[2] H. Al-Momani, G. Magableh, O. Al Meanazel, H. Maaitah, and M. Obaidat, "Advancing healthcare equipment reliability: the role of cmms and rbem methodologies in medical equipment maintenance," *Journal of Quality in Maintenance Engineering*, vol. 32, no. 2, pp. 391–417, 2026.

[3] S. Alkhatib, R. Katmah, D. Kosaji, S. U. B. Afzal, M. Tariq, M. C. E. Simsekler, and S. Ellahham, "Ai-driven decision support framework for preventing medical equipment failure and enhancing patient safety: A new perspective," *Journal of Multidisciplinary Healthcare*, vol. Volume 18, pp. 6299–6313, 2025.

[4] M. Moleda, B. Malysiak-Mrozek, W. Ding, V. Sunderam, and D. Mrozek, "From corrective to predictive maintenance—a review of maintenance approaches for the power industry," *Sensors*, vol. 23, no. 13, pp. 1–47, 2023.

[5] A. Aminzadeh, S. Sattarpanah Karganroudi, S. Majidi, C. Dabompre, K. Azaiez, C. Mitride, and E. Sénéchal, "A machine learning implementation to predictive maintenance and monitoring of industrial compressors," *Sensors*, vol. 25, no. 4, pp. 1–25, 2025.

[6] S. Taghipour, D. Banjevic, and A. K. S. Jardine, "Reliability analysis of maintenance data for complex medical devices," *Quality and Reliability Engineering International*, vol. 27, no. 1, pp. 71–84, 2011.

[7] ANVISA - Agência Nacional de Vigilância Sanitária, "Resolução da Diretoria Colegiada - RDC n. 509, de 27 de maio de 2021," Dispõe sobre a avaliação de impacto regulatório de atos normativos de interesse da Anvisa, 2021, accessed: June 15, 2026. [Online]. Available: https://www.in.gov.br/en/web/dou/-/resolucao-rdc-n-509-de-27-de-maio-de-2021-322634358

[8] ANVISA - Agência Nacional de Vigilância Sanitária, "Resolução da Diretoria Colegiada - RDC n. 665, de 24 de março de 2022," Dispõe sobre os requisitos de Boas Práticas de Fabricação de produtos médicos e produtos para diagnóstico de uso in vitro, 2022, accessed: June 15, 2026. [Online]. Available: https://www.in.gov.br/web/dou/-/resolucao-rdc-n-665-de-24-de-marco-de-2022-389146200

[9] V. N. Mohan, "Intelligent preventive maintenance: The rise of usage-based scheduling," *Journal of Computer Science and Technology Studies*, vol. 7, no. 6, pp. 1090–1095, 2025.

[10] C. Li, J. Wang, S. Wang, and Y. Zhang, "A review of iot applications in healthcare," *Neurocomputing*, vol. 565, pp. 1–12, 2024.

[11] P. Guo, S. Xu, and W. Liang, "A cloud-assisted anonymous and privacy-preserving authentication scheme for internet of medical things," *Computers & Security*, vol. 157, pp. 1–16, 2025.

[12] G. Shahrour, A. Junejo, and A. Al-Said Ahmad, "Security and privacy challenges in healthcare iot systems with cloud, edge, and fog computing: A systematic literature review," *Computer Science Review*, vol. 62, pp. 1–28, 2026.

[13] S. S. Salehi, H. Saadatfar, S. S. Oyelere, S. Hussain, J. Hassannataj Joloudari, M. Taheri Ledari, E. Arslan, and B. Barzegar, "Enhancing healthcare outcome with scalable processing and predictive analytics via cloud healthcare api," *Frontiers in Digital Health*, vol. 7, pp. 1–18, 2026.

[14] I. C. Panagou, S. Katsoulis, E. Nannos, F. Zantalis, and G. Koulouras, "A comprehensive evaluation of iot cloud platforms: A feature-driven review with a decision-making tool," *Sensors*, vol. 25, no. 16, pp. 1–58, 2025.

Accessibility in Medical Devices: Perceptions and Strategies in Clinical Engineering

Mariana Ribeiro Brandão
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
marianaribeirobrandao@gmail.com

Anderson Alberto Ramos
Department of Electronics and Biomedical Engineering (DEEB)
Universidade Estadual de Campinas
Campinas, Brazil
a123236@dac.unicamp.br

Renato Garcia Ojeda
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
renato.garcia.ojeda@ufsc.br

***Abstract*—Medical devices are designed for the benefit of the population; however, they must be accessible to meet the diverse needs of the entire population, considering different ethnicities, races, genders, disabilities, as well as diverse contexts and locations. This study generates real-world evidence on accessibility barriers in medical devices within healthcare settings from a Clinical Engineering perspective, proposing strategies to enhance inclusivity. An online questionnaire administered to 68 clinical engineering professionals revealed a gap in Healthcare Technology Management (HTM); 55.6% of respondents were unaware of accessible technologies within their equipment inventory. Among the remaining 44.4% who confirmed such availability, the inventory was concentrated in mobility-related assistive technologies, such as wheelchairs and walkers (17.5%), followed by diagnostic imaging equipment (9.5%). Qualitative analysis indicates that the conceptualization of "accessible medical devices" remains narrowly restricted to mobility aids. To ensure patient safety and functional autonomy, HTM frameworks must incorporate accessibility and Universal Design as a criteria in procurement processes. Adopting this equity-driven approach is important to transform medical devices into tools that guarantee inclusive care and equal access in healthcare systems.**

Keywords—Accessibility; Medical Devices; Clinical Engineering; Health Technology Management.

I. Introduction

Health technologies consist of medical devices, medications, vaccines, procedures, and systems developed to solve a health problem and improve quality of life. Due to the increasing incorporation of technology, Health Technology Management (HTM) becomes essential to assist in its safe and reliable use. One area that can support the development, implementation, and monitoring of the HTM program in healthcare settings is Clinical Engineering (CE), a field within Biomedical Engineering [1]. Due to the complexity and interdisciplinary nature of the involvement, interconnectivity among the various actors in the healthcare field becomes essential for the safety and comprehensiveness of the appropriate use and accessibility of these technological solutions for health

HTM can be understood as a set of activities that ensure quality in technological processes within healthcare facilities. The Institute of Biomedical Engineering (IEB-UFSC) bases its management model on three main pillars: infrastructure, technology, and human resources, as shown in Fig 1. The interaction between them aims to analyze the entire context in which the technology is embedded, and the pursuit of quality in the technological process results in improved safety, reliability, and effectiveness in the use of these technologies [2]. This work presents an additional perspective on the importance of accessibility in the quality of the technological process in healthcare and presents improvement strategies to make technologies more inclusive for the population.

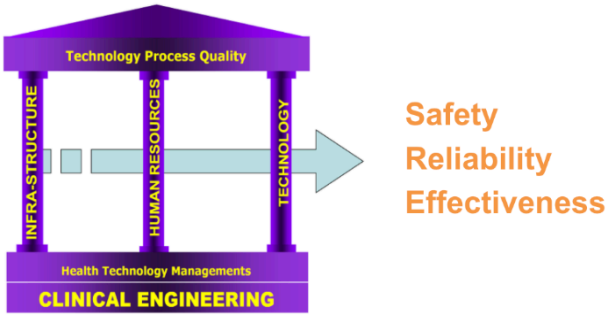


Figure 1. Health Technology Management Model based on the pillars of infrastructure, technology, and human resources of the IEB-UFSC

Medical devices are designed for the benefit of the population, however, they must be accessible to meet the diverse needs of the entire population, considering different ethnicities, races, genders, disabilities, as well as diverse contexts and locations. According to data from the World Health Organization, it is estimated that 1.3 billion people suffer from significant disabilities. This represents 16% of the world's population. Considering accessibility aspects in the development of health technologies is essential to ensure inclusion and improve usability [3].

Accessibility, as defined by ABNT NBR 17060:2022, consists of "the extent to which products, systems, services, environments, and facilities can be used by people from a population with the widest variety of characteristics and abilities, to achieve a specific goal in a specific context of use" [4]. The goal of accessibility is to expand the target

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

population, ensuring that products, systems, services, environments, and facilities are usable by a wider range of individuals in diverse contexts of use.

Accessibility is not only required by law, but is also crucial for reducing inequalities and ensuring the inclusion of all people in the use of healthcare technologies [5][6]. Accessibility should be considered throughout the entire lifecycle of healthcare technology, from product development to the design of customizable technology for people with different limitations and characteristics, to post-commercialization stages [7][8]. Inclusive and human-centered design focuses on developing solutions that consider human diversity beyond people with disabilities, the elderly, and physical accessibility. It takes into account psychosocial dimensions and values quality of life and user experience. When investigating barriers to inclusion, the goal is to involve diverse user profiles and create innovative solutions that benefit everyone [9]. Therefore, it is recommended that the design be adjustable and customizable [10].

Although usability is discussed in medical devices, accessibility, specifically from the perspective of those who manage the technology, is rarely explored. Medical devices are often not designed to meet the needs of some specific users [11]. Therefore, this work aims to generate real-world evidence on accessibility barriers in medical devices in healthcare settings from a Clinical Engineering perspective, as well as to elucidate improvement strategies to make healthcare more inclusive for all people.

II. Methodology

This research is characterized as an exploratory and descriptive study, with a mixed, quantitative and qualitative approach, based on the collection of real-world data, with the objective of generating evidence of accessibility problems in medical devices from the perspective of Clinical Engineering. One of the authors of this work is a clinical engineer and has aphasia, contributing significantly to the analysis and establishment of strategies for the inclusive design of Medical Devices.

The target population comprised professionals working directly in Clinical Engineering departments across hospital institutions (public, private, and philanthropic) in Brazil. The inclusion criterion was being a mid-level or higher-level professional (engineers, technologists, technicians, and managers) with an active role in hospital Clinical Engineering services. Responses with incomplete data were excluded. The final valid sample consisted of 68 participants.

Data collection took place between August 2025 and April 2026. Data were collected using a structured online questionnaire administered through an electronic research platform. The survey link was strategically disseminated via professional networks, social media, and technical forums dedicated to Clinical and Biomedical Engineering nationwide. The research instrument was constructed based on human factors engineering, usability guidelines, and multidimensional accessibility frameworks.

Content validation was conducted prior to deployment through expert review by Clinical Engineering researchers to assess item clarity, thematic relevance, and comprehensiveness. The final questionnaire comprised multiple-choice blocks (demographic characteristics, institutional profile, team composition, and technology incorporation practices) and open-ended items exploring accessibility barriers in medical devices and management strategies.

Quantitative data were analyzed using simple descriptive statistics (absolute frequencies and percentages). Inferential statistics were not applied due to the non-probabilistic nature of the sample. Qualitative responses from open-ended items were subjected to thematic content analysis. All open-ended responses were independently read and assigned to initial thematic categories aligned with multidimensional accessibility domains (e.g., operational, spatial, informational, economic, sociocultural, and interoperability).

The research protocol was approved by the Research Ethics Committee of the Federal University of Santa Catarina (UFSC) under approval No. 7.706.804 (CAAE: 87849125.4.0000.0121). Electronic informed consent was obtained from all participants prior to survey completion, ensuring anonymity and data confidentiality.

III. Results

The exploratory study with Clinical Engineering professionals on accessibility included the participation of 68 professionals working in Clinical Engineering in public, private, and philanthropic hospitals. Regarding education, 43 were professionals with engineering degrees, 13 with technical training, 10 technologists, 1 administrator, and 1 health professional.

The research highlighted a gap in hospital technology management, where 55.6% of participants reported being unaware of the existence of specific accessible technologies in their equipment inventory. Among the 44.4% who confirmed availability, there was a concentration on assistive technology items to aid mobility, such as wheelchairs and walkers (17.5%), followed by diagnostic imaging equipment (9.50%). Although 44.4% of institutions report having accessible technologies, the qualitative analysis reveals that the concept of 'accessible medical device' is still strongly linked to mobility items such as wheelchairs. Adapted diagnostic technologies, such as mammograms and specific scales, are emerging, demonstrating a vast field for Clinical Engineering to manage more inclusive technologies throughout the lifecycle.

Table 1 illustrates the types of accessible technologies present in the hospital as reported by those working in the Clinical Engineering sector. Because respondents were able to identify multiple types of devices within their inventories, the technology categories are not mutually exclusive. Percentages in Table I are calculated relative to the total valid study sample (n=68) to reflect overall population prevalences. Another factor analyzed in the research was related to the process of incorporating medical devices in hospitals. It was observed that 44% of professionals reported not considering accessibility as a decision-making factor. The contrast with the scarcity of accessible devices reported previously suggests that the assessment of accessibility at the time of purchase is often subjective or insufficient, lacking standardized technical criteria to guarantee Universal Design.

TABLE I - REPORT ON ACCESSIBLE MEDICAL DEVICES IN HOSPITALS AS REPORTED BY PROFESSIONALS WORKING IN CLINICAL ENGINEERING.

Technological Category	Examples of Devices and Resources Mentioned	Sample Prevalence (%)
Mobility and Support	Wheelchairs, walkers, hospital beds, patient lifts, adapted stretchers, and beds with minimum height adjustment.	**17,50%**
Diagnostic Imaging	CT scanners, MRI, Mammography (height adjustment), ceiling stands and layout adaptation (wide doors/ramps).	**9,50%**
Weight Supported	ICU beds, stretchers and surgical tables reinforced for obese patients (focus on load and ergonomics).	**8,20%**
Communication and Sensory	Medvoice (speech/hearing aid), sound amplification devices, screen readers, adapted keyboards, intercoms with video and subtitles.	**7,80%**
Scale and Triage	Scale for wheelchair users (low-profile platforms).	**6,30%**
Other / Innovation	Automated instrument carts, AI-powered equipment and adapted audiometry booths; Pediatric autorefractors (rapid measurement for agitated patients or those with difficulty concentrating).	**6,40%**
Absence / Lack of Knowledge	Professionals who do not identify or claim not to possess specific devices.	**55,60%**

The main barriers reported by the participants were categorized in the pillars of Health Technology Management of IEB-UFSC related to infrastructure, human resources and technology (Table 2). Regarding infrastructure, the presence of old hospitals that prevent modern adaptations, lack of accessible elevators and ramps with inadequate inclination, narrow doors, high counters and lack of adapted furniture stand out. Related to technology, the lack of adapted equipment, incompatibility of technologies that meet the needs of people with disabilities, absence of braille, tactile floors and audio-visual resources; lack of sign language interpreters or assistive technologies to mediate care. Regarding human resources, the following stand out: professional unpreparedness, lack of training to deal with diversity and specific situations involving people with disabilities; prejudice and lack of prioritization, perception that accessibility is not a management priority or is not seen as an essential part of clinical care; poor budget management and lack of investment directed towards adaptations.

TABLE II - ACCESSIBILITY BARRIERS IN MEDICAL DEVICES CATEGORIZED ACCORDING TO THE GTS PILLARS.

Pillar (Frequency of citations)	Barrier identified	Description of the report
Infrastructure (56%)	Obsolete Buildings Access Barriers Environmental Signage	*"Old hospitals", "Physical structure not designed for accessibility". "Steep ramps", "Lack of elevators", "Narrow doors", "Access only by stairs". "Lack of tactile paving", "Non-adapted bathrooms", "High countertops"*
Human Resources (28%)	Lack of Technical Preparation Communication Failure Organizational Culture Lack of Prioritization	*"Unqualified professionals for the service", "Lack of knowledge of the team in general". "Lack of people who know sign language", "Assistance and support team without training". "Prejudice", "Lack of management priority", "Lack of perception of the needs of others" "Lack of effective institutional policies" "Corruption"*
Technology (16%)	Inaccessible interface Obsolescence Economic barrier Interoperability of use Non-compliance with regulations	*"Outdated technology park," "Equipment not adapted (button height, interface)." "High cost of good devices," "Lack of renewal of the technology park." "Incompatibility of equipment and suitable furniture," "Lack of equity solutions." "Failure to minimally comply with current standards and legislation" "Not communicating by voice, no audiovisual equipment, the height to operate the buttons is too high"*

Another result analyzed was the composition of Clinical Engineering teams with people with disabilities. A low representation rate was found, with only 17.6% of institutions having professionals with disabilities on their technical staff (Fig 2). This data is concerning when combined with the perception of technological barriers. The absence of professionals with disabilities in the management of health technologies can result in the neglect of usability and accessibility requirements because they are not part of the direct experience of decision-makers. Promoting the inclusion of people with disabilities in Clinical Engineering teams is not just a matter of diversity, but a co-design strategy, essential for the specification and evaluation of medical devices to be genuinely inclusive and user-centered.

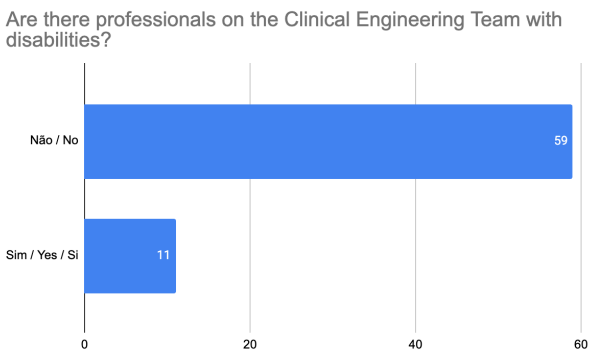


Figure 2. Representation of professionals with disabilities in the Clinical Engineering Team in Hospitals

The suggestions for accessibility improvements in medical devices reported were categorized from the perspective of the domains based on the accessibility framework for medical devices [12]. The domains are: operational, economic, informational, infrastructure and spatial accessibility, connectivity and interoperability, and sociocultural accessibility, as presented in Table 3. The strategies elucidated were multidimensional to ensure safety and equity in care.

In the operational domain, the reports highlight the urgency of inclusive interfaces including audio and braille resources. In the infrastructure domain, a concern emerges regarding physical safety in emergency situations, exemplified by the need for universal fixation supports so that portable equipment is not positioned over the patient. The evidence in the sociocultural domain reinforces the importance of participatory design, suggesting that the inclusion of people with disabilities from the design phase is the fundamental path to converting technological innovation into functional autonomy and dignified care.

IV. Discussion

The integrated analysis of the data reveals a divergence between theoretical awareness and institutional practice in Clinical Engineering. Although a majority of professionals (56%) stated that accessibility is a relevant consideration during device incorporation, 55.6% reported either an absence or a lack of identification of accessible technologies in their institutions. This finding suggests that beyond physical availability, professionals themselves may lack sufficient awareness or technical training to recognize accessibility features already embedded in medical equipment. Furthermore, this scenario suggests that accessibility is still treated as an infrastructure add-on and not as an intrinsic safety requirement, which contrasts with the view of Panagiotopoulou (2020), who argues that health services should be equally accessible to all, regardless of present or future disabilities [13].

This technological gap is underpinned by a crisis of representation, where only 17.6% of Clinical Engineering teams have professionals with disabilities. Without the user's perspective in the incorporation process, essential devices may be being neglected in favor of traditional metrics of cost and technical performance. These findings converge with the barriers identified by Bonanno (2025), who points to accessibility disparities, training needs, and limited user participation in design as obstacles to an inclusive health technology assessment (HTA) [14].

The suggestions for improvement collected indicate future perspectives in the convergence between emerging technologies and the principles of Universal Design. As proposed by Kaminska (2023) and Brandão (2026), the design should be useful and accessible to people with diverse abilities; accommodate a wide range of individual preferences and abilities; be easy to understand and use, regardless of the user's experience, knowledge, language skills, or level of concentration; provide information in multiple ways to accommodate different abilities and sensory preferences, as well as tolerance for errors. Universal design is essential because it promotes inclusion, autonomy, and social engagement for people of all abilities [15][16].

Table III - Improvement Strategies categorized into accessibility domains to make Medical Devices more Accessible

Domain	Improvements and Suggested Resources	Participant Reports (Evidence)
Operational	Inclusive interface; audio and braille requirements; contrast and size adjustments on display; support technology resources; height adjustment, customization; intuitive interfaces, ambidextrous design and remote controls; technologies that support the patient's weight	*"Ambidextrous design is a necessity." "Braille, audio command" "Inclusive user interface, configuration customization, visual instructions on the equipment itself and accessible communication" "Adaptations to cables and accessories to measure basic parameters in patients without limbs" "More visible alarms for people with hearing problems, clearer and larger images for people with visual impairments" "Height adjustments, sound effects and support features for people with reduced mobility." "Specific stretcher or surgical table for weight capacity and physical structure of the patient"*
Infrastructure and Space	Height adjustment, universal hooks and transport ergonomics	*"A hook to attach a monitor to the bed... for defibrillator and portable ventilator"; "Shorter stature, difficulty accessing monitoring." "Doors with automatic locks, disabled people need to open them and sometimes end up falling."*
Informational	Audio/Braille manuals, redundant visual alarms, and instructions on the equipment itself; staff training	*"More visible alarms for people with hearing problems"; "Written and dictated instructions"; "Braille manual... voice translation." "Accessible keys and buttons and clear and objective information." Specific training for the care team.*
Interoperability and Connectivity	Self-positioning cameras, remote interface tablets and adapted sensors; AI voice command, features accessible via tablets	*"Devices having a portable tablet for these people"; "Self-positioning cameras on the equipment". "Integrate voice command with the use of AI to enable/program the device"*
Sociocultural	Co-design and humanization of care.	*"Bringing people with disabilities into the area so that they can express their vision from the design stage"; "Listening to accounts from people with disabilities"; "There should be more respect and care for people with disabilities."*
Economical	Cost reduction and increased availability of adapted accessories from manufacturers.	*"High cost of good devices"; "Suppliers increasingly addressing the issue of accessibility."*

In this context, Clinical Engineering plays a crucial role in generating data to establish strategies that make technological use universal. The implementation of co-design strategies in Living Labs environments presents itself as the path to a collaborative innovation ecosystem, as discussed in previous works [12][17]. This approach is aligned with Meinert's research (2024), where co-production places people with disabilities and their caregivers at the center of the process as co-researchers. By integrating universities, industry, government and, fundamentally, end users, Brazilian Clinical Engineering can move from passive management to the creation of solutions that guarantee, as the literature and field data suggest, autonomy and dignity in access to health. Regarding the limitations of the work, due to the lack of discussions and evidence on accessibility in medical devices, there is a lack of understanding on the part of clinical engineering on this topic, which may have interfered with the participants' responses. There is also a limited view on accessibility. The lack of representation of people with disabilities in Clinical Engineering also reflects a bias in the data.

V. Conclusion

In conclusion, this work makes an original scientific contribution by presenting exploratory, real-world evidence regarding the incorporation and perception of accessibility within clinical engineering practice in Brazil. By highlighting how professionals in the field experience this issue, the study establishes a novel empirical foundation to guide the development of procurement guidelines, management policies, and future research on the subject.The results of this study demonstrate that accessibility in medical devices from a Clinical Engineering perspective is still underexplored. Although most professionals recognize the relevance of the topic, the persistence of technological gaps indicates that accessibility is still often perceived as an additional cost or merely an adaptation of infrastructure, rather than a requirement for patient safety.

This study has limitations that must be taken into account when interpreting the findings. First, the cross-sectional design and non-probabilistic restrict the generalizability of results across all Brazilian hospital settings. Second, the data rely exclusively on self-reported perceptions, which are susceptible to response bias, and varying levels of individual technical literacy regarding accessibility. The lack of representation of people with disabilities in Clinical Engineering also reflects a bias in the data. The conceptual definition of accessibility provided in the instrument may have framed participant responses. Finally, the lack of statistically matched subgroups by hospital size, region, and type precluded inferential tests of association, highlighting the need for future studies with robust stratification and objective device audits.

The consolidation of the data reveals a complex management scenario. While the low representation of professionals with disabilities points to limited diversity within Clinical Engineering teams, this factor alone does not directly establish that their absence causes the neglect of usability or accessibility requirements. Nevertheless, the reported demands for assistive communication, adapted sensory interfaces, and artificial intelligence integration signal a growing awareness and a potential paradigm shift in technology incorporation.

The authors consider it relevant to consider the incorporation of usability techniques throughout the life cycle of technologies involving people with disabilities, to explore accessibility flaws and propose effective strategies for inclusive design. Another recommendation is to incorporate Human Factors Engineering indicators into the HTM process to ensure better usability and safety for all people involved with medical devices. Accessibility is evolving to become an active and fundamental element in technical specification. Clinical Engineering must, therefore, assume its role as an integrating link, overcoming infrastructure and human resource barriers to ensure that the technology is, in fact, usable and operable by everyone.

In conclusion, while the findings reveal a limited understanding of accessibility within Brazilian Clinical Engineering, this study represents an innovative contribution to the field. By filling a critical evidence gap in HTM, it establishes a valuable foundation and offers a promising framework for inclusive procurement, design, and management practices. Nevertheless, broader normative recommendations will still require future studies supported by larger representative samples, formal validation and involvement of users with disabilities.To guarantee safety and functional autonomy, it is recommended that standard operating procedures in HTM incorporate accessibility and Universal Design as a criteria in the acquisition processes. Only through a vision based on equity will it be possible to transform medical devices into an instrument for democratizing care and human dignity in healthcare.

References

1. World Health Organization (WHO), Global Atlas of Medical Devices: WHO Medical Devices Technical Series, Geneva, Switzerland, 2017
2. M. R. Signori and R. Garcia, “Clinical engineering incorporating human factors engineering into risk management,” in World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany (IFMBE Proceedings, vol. 25/7), O. Dössel and W. C. Schlegel, Eds. Berlin, Heidelberg: Springer, 2009, pp. 446–449, doi: 10.1007/978-3-642-03885-3_125.
3. World Health Organization (WHO), “Disability, Key facts,” 2023. [Online]. https://www.who.int/health-topics/disability#tab=tab_1 [retrieved: May, 2026]
4. Brazilian Association of Technical Standards (ABNT), Accessibility on Mobile Devices, ABNT NBR 17060:2022, 2022.
5. Americans with Disabilities Act (ADA), Access to Medical Care for Individuals with Mobility Disabilities: Guidance & Resources, USA, 2020.
6. Brasil, “Lei Nº 13.146, de 6 de julho de 2015,” Presidência da República, Jul. 06, 2015. Available: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm
7. M. Brandao and R. Garcia, “Analysis of accessibility problems in medical devices,” International Journal on Advances in Life Sciences, vol. 16, no. 3-4, pp. 178–187, 2024. Available: https://www.thinkmind.org/library/LifSci/LifSci_v16_n34_2024/lifsci_v16_n34_2024_7.htm
8. L. Heinemann, D. Drossel, G. Freckmann, and B. Kulzer, “Usability of medical devices for patients with diabetes who are visually impaired or blind,” Journal of Diabetes Science and Technology, vol. 10, no. 6, pp. 1382–1387, 2016, doi: 10.1177/1932296816666536.
9. I. Leason, N. Longridge, M. Mathur et al., “An opportunity for inclusive and human-centred design,” British Dental Journal, vol. 233, pp. 607–612, 2022, doi: 10.1038/s41415-022-5101-1.
10. D. A. Norman, The Design of Everyday Things, revised and expanded ed. New York, NY, USA: Basic Books, 2013.
11. C. H. Fung et al., “Human factors/usability barriers to home medical devices among individuals with disabling conditions: In-depth interviews with positive airway pressure device users,” Disability and Health Journal, vol. 8, no. 1, pp. 86–92, Jan. 2015, doi: 10.1016/j.dhjo.2014.06.002.
12. M. R. Brandão et al. Accessibility in Medical Devices: Proposal for a Conceptual Framework and Strategies for Universal Design. Proc. 11th Int. Conf. Universal Accessibility in the Internet of Things and Smart Environments (SMART ACCESSIBILITY 2026), Veneza, Italy. https://www.thinkmind.org/library/SMART_ACCESSIBILITY/SMART_ACCESSIBILITY_2026/smart_accessibility_2026_1_10_88002.html

13. M. Panagiotopoulou and M. Tsirintani, “Accessibility and web quality information for people with disabilities in healthcare structures,” Studies in Health Technology and Informatics, 2020, doi: 10.3233/shti200556.
14. M. Bonanno et al., “Assistive technologies for individuals with a disability from a neurological condition: A narrative review on the multimodal integration,” Healthcare, vol. 13, no. 13, p. 1580, 2025, doi: 10.3390/healthcare131315802.
15. D. Kamińska, G. Zwoliński, L. Pinto-Coelho e R. Raposo, "Universal design and empathic design for engineers," Med. Pr., vol. 74, no. 3, pp. 211–225, 2023. doi: 10.13075/mp.5893.01388.
16. M. R. Brandão, J. V. Fogaça, J. L. B. Marques e R. G. Ojeda, "Perception of users with visual impairments regarding adverse events in the use of technologies for diabetes," J. Accessibility Design All, vol. 16, no. 1, pp. 96–105, mai. 2026. [Online]. https://www.jacces.org/index.php/jacces/article/view/639
17. M. R. Brandão, J. L. B. Marques e R. G. Ojeda, "Analysis of Adverse Events in Medical Devices for Diabetes," *J. Diabetes Sci. Technol.*, set. 2025, Art. no. 19322968251368860. doi: 10.1177/19322968251368860.

AI-based Fall and Movement Detection System for Hospital Beds

João Carlos Corrêa Gomes
Biomedical Engineering
Inatel
Santa Rita do Sapucaí, Brazil
joao.carlos@geb.inatel.br

Mizael Domingos Contini
Biomedical Engineering
Inatel
Santa Rita do Sapucaí, Brazil
mizael.domingos@geb.inatel.br

Samuel Mangia Maciel
Biomedical Engineering
Inatel
Santa Rita do Sapucaí, Brazil
samuel.mangia@geb.inatel.br

Yasmim Silva Ramos
Biomedical Engineering
Inatel
Santa Rita do Sapucaí, Brazil
yasmim.silva@geb.inatel.br

Carlos Alberto Ynoguti
Inatel
Santa Rita do Sapucaí, Brazil
ynoguti@inatel.br

***Abstract*—Falls, in general, can cause various damages to individuals, such as abrasions, bruises, fractures, and physical and psychological trauma. When they occur in the hospital environment, in addition to the effects already mentioned, the patient's clinical condition may worsen, especially in cases where it takes time to identify these events. Consequently, a fall can result in higher financial expenses for the hospital due to the increased length of stay. Based on this scenario, the present study proposes a patient monitoring system for hospital beds using YOLOv8, a state-of-the-art real-time object detection algorithm. Unlike binary fall/no-fall approaches reported in the literature, the proposed system classifies five distinct patient behaviors, enabling preventive alerts before fall events occur.**

***Index Terms*—Artificial Intelligence, Hospital Fall, Image Monitoring, Pattern Recognition**

I. Introduction

A patient fall is an undesirable and harmful occurrence, an unpredictable event for which there is still little incentive directed toward prevention and/or immediate assistance. Studies addressed by the Secretariat of the Federal District State [1] indicate that the patient fall rate in hospitals in developed countries ranged between 3 to 5 falls per 1,000 patients/day.

According to Lima et al. [2], in Brazil, 22.5% of deaths of elderly people that occurred in the period from 1996 to 2005 resulted from falls. That study presents an analysis of AIHs (Autorizações de Internação Hospitalar), carried out with data from the SIH (Sistema de Informações Hospitalares) of the SUS (Sistema Único de Saúde), and found that 1,746,097 AIHs due to falls in elderly people were recorded in the period from 2000 to 2020, corresponding to a cost of R$ 2,315,395,702.75 for the health system.

In addition, the study by de Frutos et al. [3] emphasizes the problem related to falls, especially among the elderly. The document states that falls are the second leading cause of death from unintentional injuries worldwide, and the highest number of fatal falls occurs mainly among adults over 60 years of age. In the United States, about 700,000 to 1,000,000 fall cases are reported annually in hospitals. These cases are considered *never events* (events that should never occur), meaning that insurance companies do not cover the expenses related to these incidents. As a result, hospitals in the USA spend about 50 billion dollars per year to cover these expenses [3].

The consequences of a fall range from skin injuries, dislocations, and fractures to impairment of physical mobility, progressive reduction of functional capacity, and worsening of the previous clinical condition. Falls can also generate psychological consequences such as anxiety, insecurity, social isolation, and loss of confidence due to fear of falling again, as well as loss of autonomy to perform basic activities of daily living. Furthermore, falls can increase the length of hospital stay and compromise the clinical condition and recovery of the individual, negatively impacting healthcare costs.

As a result of the numerous complications and serious damages that a fall can cause, together with the high cost of prolonged hospitalization, it is necessary to use methods and techniques that enable the identification of fall patterns, allowing the responsible clinical staff to be informed as quickly as possible about these situations and enabling faster and more effective intervention.

The present work aims to present a simpler method that is as effective as existing approaches for detecting fall situations in hospital beds. This environment was selected as the focus because it is where individuals are found, generally with reduced physical capacity, under the effect of medication and adverse conditions that may lead to fall events.

Currently, similar works exist focused on fall detection. However, in the literature, there is a greater focus on binary systems that identify patients as either in bed or out of bed. In this type of analysis, other patterns beyond falls, such as risk situations and movement out of bed, tend to be overlooked.

The system proposed in this study addresses different patient behaviors, enabling the issuance of alerts to the hospital clinical staff in different situations — such as the patient

leaving the bed or being in a near-fall condition — and not being limited only to fall events. In this way, it is possible to provide better monitoring and medical intervention to patients whenever necessary. The analysis of patient behavior patterns brings more efficiency in fall prevention, reducing the damage caused by this event.

II. Related Work

Several camera-based approaches to fall detection have been proposed in the literature, predominantly relying on binary classification schemes. This section reviews three representative systems whose methodologies and results serve as baselines for the proposed work.

Bian et al. [4] proposed a system based on a depth camera, which enables image acquisition even in low-lighting environments. The methodology corrects the image position to a vertical orientation to enable better analysis of the torso. Articulable body points — such as the neck, head, elbows, and knees — are identified using a Randomized Decision Tree, a machine learning algorithm that generates decision nodes. Head orientation is then analyzed using a Support Vector Machine (SVM) for pattern recognition, enabling the identification of body trajectory in fall situations. A fall recovery analysis mechanism concludes the pipeline by comparing the height position of the hip and head over time to confirm whether a fall trajectory has been completed. The system was trained on a dataset with 100 thousand data points of head trajectory patterns, and its results are shown in Table I.

TABLE I
Results of Model 1 [4]

Metrics	Results
True Positives	188
True Negatives	104
False Positives	86
False Negatives	2
Sensitivity	98.8%
Specificity	54.7%
Accuracy	76.8%
Error Rate	23.2%

Nguyen et al. [5] proposed a camera-based surveillance model aimed at elderly supervision. The system acquires 30 frames per second, applies joint segmentation, and uses a moving average filter to attenuate noise, followed by background removal. Objects are then identified through comparison with predefined detection probability thresholds, and the resulting images are classified using a rule-based system that distinguishes fall from non-fall movements based on body proportion and angulation. The model was evaluated on 3,996 frames — 2,873 fall frames and 1,123 non-fall frames — and its results are shown in Table II.

De Miguel et al. [6] presented a home-based fall detection system using 2D and 3D cameras. After image acquisition and background removal, the frames pass through a Kalman filter that reduces noise and absorbs repetitive changes in detected action behaviors by comparing noisy and clean images. This filter also enables object tracking and position prediction at each frame. The OpenCV library is used alongside the Kalman filter, applying optical flow techniques for precise tracking of items of interest. Classification is performed by the K-Nearest Neighbors (KNN) algorithm, which assigns class labels based on neighboring data sets. The system was trained on frames from 53 videos recorded in two different environments, and its results are shown in Table III.

TABLE II
Results of Model 2 [5]

Metrics	Results
True Positives	2587
True Negatives	1111
False Positives	11
False Negatives	286
Sensitivity	90.0%
Specificity	99.0%
Accuracy	92.5%
Error Rate	7.4%

TABLE III
Results of Model 3 [6]

Metrics	Results
True Positives	24
True Negatives	40
False Positives	1
False Negatives	1
Sensitivity	96.0%
Specificity	97.6%
Accuracy	96.9%
Error Rate	3.0%

All three systems described above adopt a binary classification approach, distinguishing only between fall and non-fall events. The present work extends this paradigm by introducing a five-class model capable of recognizing intermediate and preventive behaviors, enabling intervention before a fall event occurs.

III. Theoretical Background

A. Evaluation Criteria

For the evaluation and comparison of results, metrics similar to those of other studies were used, following the quality criteria for the analysis of fall detection systems suggested by Noury et al. [7]. The evaluation parameters consist of data provided by the confusion matrix, a tool used for performance analysis of classification systems, and its derived metrics:

- **True Positives (TP)**: The model correctly predicts the class being predicted, when it is actually present;
- **False Positives (FP)**: The model incorrectly predicts the class being predicted, when it is not present;
- **True Negatives (TN)**: The model correctly predicts the class that is not being predicted, when it is not present;
- **False Negatives (FN)**: The model incorrectly predicts the class that is not being predicted, when it is present.

Based on these parameters, the following metrics can be derived:

- **Accuracy**: Provides the rate of correct predictions in relation to the total number of predictions.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

- **Sensitivity**: Ability to correctly identify results belonging to the positive class.

$$\frac{TP}{TP + FN}$$

- **Specificity**: Ability to correctly identify results that belong to the negative class of the model.

$$\frac{TN}{TN + FP}$$

- **Error Rate**: Provides the rate of incorrect predictions in relation to the total number of predictions.

$$\frac{FP + FN}{TP + TN + FP + FN}$$

IV. Materials and Methods

A. Image Collection and Processing

Initially, a corpus of images was created from videos and photos of various individuals and situations to ensure that the dataset was diverse and representative of the target scenarios.

These images and videos were used to perform detailed annotations, marking specific and relevant characteristics in each frame. This labeling stage is essential for the AI model to learn to recognize patterns, behaviors, and distinct characteristics of individuals captured in each scenario. The diversity of both individuals and environments allows the model to learn more effectively and better generalize its predictions, expanding its applicability in different contexts. Dataset preparation is a key factor for the success of AI training and the accuracy of the final results.

The captured videos were processed and split into frames, allowing the collection of a greater variety of scenarios and capture conditions, making the dataset more robust and representative for future analyses.

Another fundamental aspect is the proper positioning of the camera. It must be strategically adjusted to ensure that the entire relevant environment is within the field of view. This includes framing the bed, the user, and the surrounding floor. This positioning is crucial to ensure that important events, such as falls or user movements, are clearly captured, providing a comprehensive view of the space and user interactions with the environment. In this work, the camera was always positioned facing the beds, also capturing the floor to enable fall event identification.

Once the image/video upload is completed, the next step is labeling and classification. This process was carried out using the *Roboflow* platform [9], which provides facilitating tools such as contour selectors and class definitions, enabling the demarcation of an area through multiple points that form a polygon around the object of interest, in this case, the individuals.

Annotation is performed image by image, selecting the object or person present and assigning it a class.

The project uses five classes:

- *person-in-bed*: shown in Figure 2, refers to the moment when the individual is in the bed;
- *person-moving-out*: shown in Figure 3, refers to the moment when the individual is moving out of the bed;
- *person-out*: shown in Figure 4, refers to the moment when the individual is outside the bed but standing;
- *person-near-fall*: shown in Figure 5, refers to a situation of imbalance or risk, a near-fall condition;
- *person-fall*: shown in Figure 6, refers to the moment when the individual experiences a fall.

Figure 1 shows an example of frame classification.

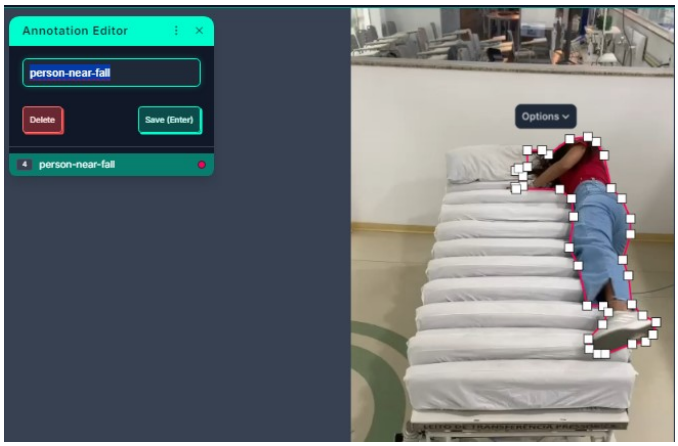


Fig. 1. Image classification process.



Fig. 2. Person-in-bed class.

Training was performed using pretrained weights from YOLOv8 *small* [8], with 300 epochs and an input resolution of 640×640 pixels.

Currently, the dataset contains 1,201 images divided into the five previously mentioned classes, obtained from videos recorded in five different environments, with 7 individuals, including 2 women and 5 men aged between 20 and 24 years. All individuals depicted in the dataset are the authors of this



Fig. 3. Person-moving-out class.

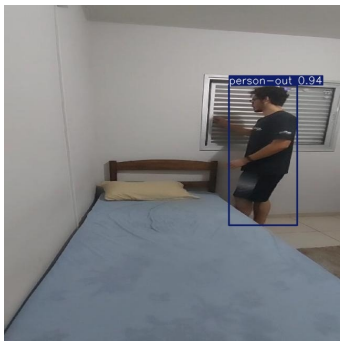


Fig. 4. Person-out class.

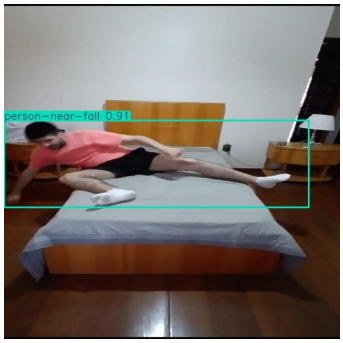


Fig. 5. Person-near-fall class.

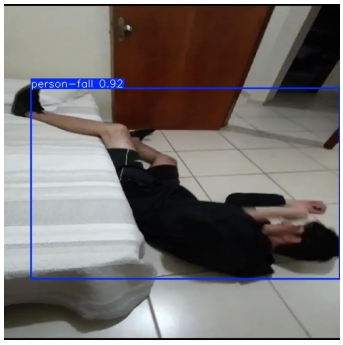


Fig. 6. Person-fall class.

work, who provided informed consent for image capture and publication. These images were used for model training. From the dataset, 841 images were used for training, 240 for testing, and 120 for validation. Figure 7 shows some sample images.

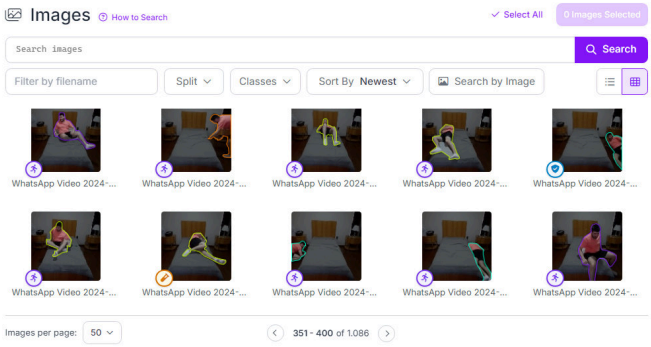


Fig. 7. Some images from the dataset with their respective classes.

B. YOLO

To develop a system capable of detecting different situations, a real-time image detection system was used. Such systems use computer vision algorithms to identify and locate objects, patterns, or features in images or videos. YOLO (*You Only Look Once*) [8] was chosen due to its fast processing, which is ideal for real-time detection applications.

Using a different approach from traditional methods, YOLO processes the entire image at once, making it faster and more efficient than two-stage methods. The image is passed through a Convolutional Neural Network (CNN), which extracts features and distributes them across a grid of cells. Each cell is responsible for predicting the presence of an object. The result is a set of bounding boxes, which are filtered to remove overlaps and provide class labels for each object in the image, resulting in an output similar to that shown in Figure 8.

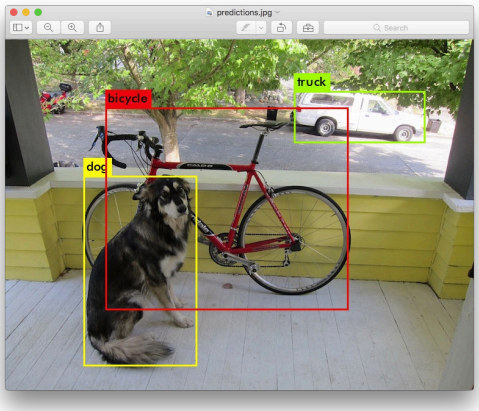


Fig. 8. Class identification using YOLO.

V. Developed Application

A. API

The system developed for this study consists of a camera with internet connectivity integrated with software responsible for image processing, which returns results obtained from inference based on the trained model weights.

The HTTP (*Hypertext Transfer Protocol*) [10] enables communication between systems through the exchange of requests and responses over the internet between clients and servers. In the developed application, the API (Application Programming Interface) acts as the server, while the hospital information systems act as clients.

Based on this protocol, integration between the camera and the application is carried out through an API, which exposes a POST HTTP endpoint allowing image submission to the system. The API was developed in Python [11], a high-level programming language with a wide range of libraries and resources focused on Artificial Intelligence and web systems.

The Flask micro-framework [12] was used, consisting of a set of functions and components for web application development, aiming to enable a system with internet connectivity. Through this tool, the API became exposed to the network, allowing it to receive requests.

Requests containing images go through the YOLO inference process, which includes computer vision capabilities. The inference function performs object detection based on the trained weights, which are preloaded into the model. The system does not store any images; it only processes them once. This process returns the detected class, if any, along with its confidence score, ranging from 0 to 100%.

The data is returned in key-value format following the JSON (JavaScript Object Notation) standard [13], which enables system integration across different platforms due to its simple and compact structure.

API responses follow the HL7 FHIR (Health Level 7 - Fast Healthcare Interoperability Resources) standard, established for healthcare data exchange, enabling system interoperability [14]. Thus, the developed application is ready for integration with other systems used in hospital environments.

Additionally, the system uses the Twilio library [15], which provides programmable communication functionalities such as email sending, phone calls, and Short Message Service (SMS) delivery. In this project, SMS messaging was used to notify the responsible clinical staff. These alerts are triggered based on the confidence level returned by YOLO when identifying a class in the processed image.

For system validation in a controlled environment, a webcam positioned facing the bed was integrated, sending one image per second to the API. In addition, a control center interface was developed to simulate a monitoring hub, displaying real-time images from the camera. When an alert was received, the system was configured to emit a visual and audible warning to the staff, in addition to the SMS notification.

Figure 9 shows the basic flow of information received and sent by the API.

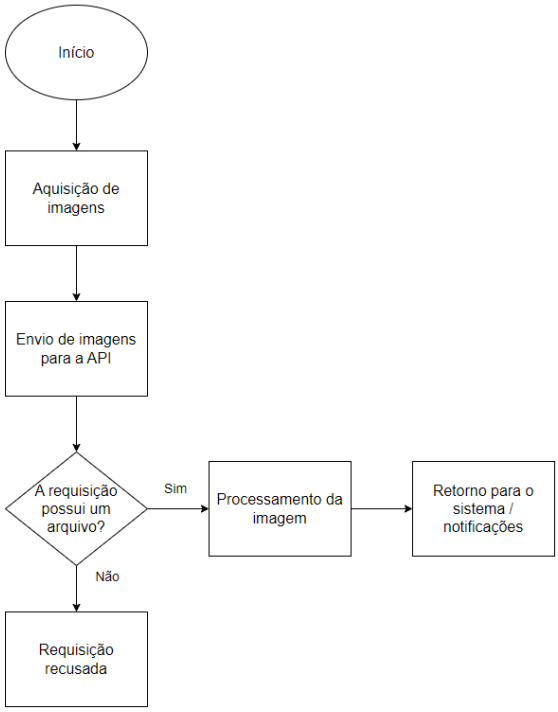


Fig. 9. Basic communication flow diagram.

B. Cloud Hosting

Cloud computing consists of making applications, programs, and data available over the internet instead of relying on local physical infrastructure, allowing users in different locations to access and use available resources and content [16]. Cloud hosting follows the same concept and involves deploying systems such as APIs through allocated resources on virtual machines [17] or similar environments.

Based on this concept, the application was structured so that hospital information systems can use its functionalities without needing to add a new service to their infrastructure, accessing system resources via HTTP requests.

Thus, cloud hosting was implemented using a virtual machine with a Linux operating system provided by Microsoft Azure [18], which offers sufficient computational resources for fast processing of information and images, ensuring efficient system response given that execution occurred in a single test environment.

Furthermore, aiming at future portability and scalability, Docker container technology [19] was used. It packages the application into a self-contained environment that can run on different machines and services, facilitating replication and distribution across servers through a set of instructions known as images.

The diagram shown in Figure 10 summarizes the system infrastructure.

C. Alarm Emission

The developed application includes a notification system that informs both the patient and the clinical staff. The alarms

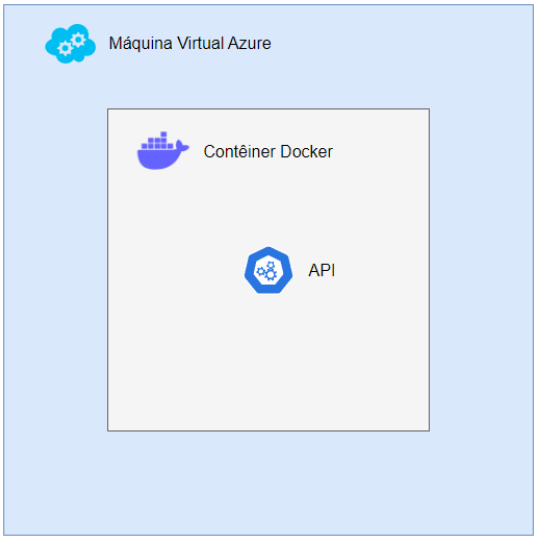


Fig. 10. Infrastructure diagram.

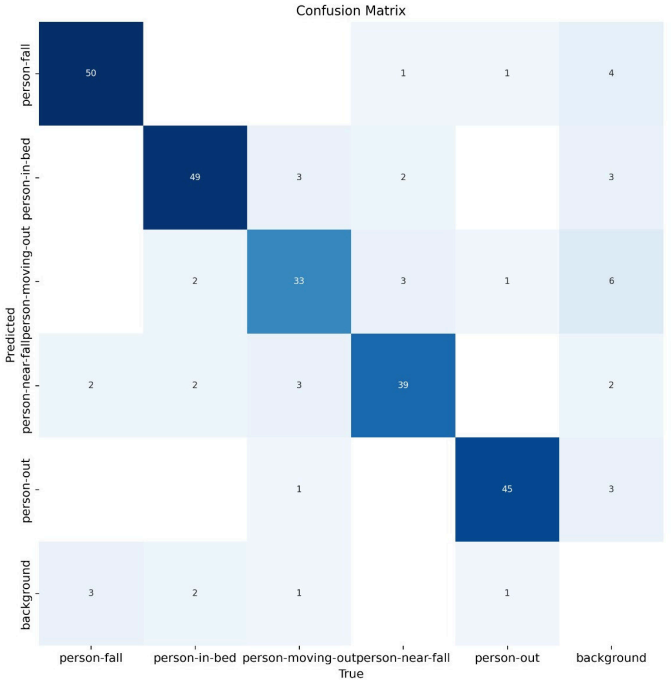


Fig. 11. Confusion matrix.

are customizable and adaptable to each patient's context through an interface that allows parameter adjustments. Their main goal is to mitigate harm caused by falls through faster intervention and to prevent falls by alerting responsible staff in potentially dangerous situations.

The basic configuration includes an audible alarm for the patient when *person-moving-out* and *person-out* situations are detected, reinforcing the importance of remaining in bed. In cases of *person-near-fall* and *person-fall* events, an alert is sent to the clinical staff through a notification to the central monitoring system.

In addition, the system uses the Twilio library [15], enabling SMS messaging to hospital staff describing whether a fall or other incident has been detected. SMS alerts also serve as a redundancy mechanism, ensuring that clinical staff are notified even when not present at the monitoring center at the time of the event.

VI. Results

To obtain the results, the test set was used, containing 240 images and representing 20% of the dataset.

The confusion matrix shown in Figure 11, obtained through YOLO, displays the actual classes on the Y-axis and the predicted classes on the X-axis. By analyzing the main diagonal, it is possible to obtain the True Positive (TP) results for each class, as well as derive the other metrics based on the information contained in the rows and columns of the matrix.

Based on the analysis of the confusion matrix, it can be observed that the *near-fall* class, together with the *person-moving-out* class, presented the highest number of False Negatives and False Positives, representing, together, 45.65% and 30.43%, respectively, of the total of these metrics.

It is observed that the confusion matrix presents information related to the *background* class, which represents frames in which no class was identified. The absence of detections in this pseudo-class indicates a need to more clearly address situations in which the patient is not in bed.

The *near-fall* class presented the lowest performance during the conducted tests, which is likely due to the fact that the near-fall position can vary significantly depending on the patient's situation. To improve the metrics of this class, which is considered one of the most important for the developed application, it is necessary to expand the dataset by capturing more scenarios. More images provide more learning possibilities for the AI model and, consequently, a greater ability to achieve more reliable detection.

The overall results of the model developed in this work are shown in Table IV.

TABLE IV
Results of the Developed Model

Metrics	Results
True Positives	216
True Negatives	1264
False Positives	46
False Negatives	46
Sensitivity	82.44%
Specificity	96.49%
Accuracy	94.15%
Error Rate	5.85%

When performing a numerical comparison with the model proposed by Bian et al. [4], which used a training dataset of 100 thousand data points, the proposed model performs better in terms of accuracy, specificity, and error rate. Sensitivity, however, shows a lower result, which may be related to the greater number of classes in this study. It is important to

note that such comparisons carry methodological limitations, since the datasets, environments, and class definitions differ substantially between works.

All studies used as comparison present a computer vision approach with binary output, where the systems distinguish whether a fall occurred or not. In this study, five output classes were addressed, allowing not only fall detection, but also the recognition of different patient behaviors, such as moving out of bed, being out of bed, and near-fall conditions.

Analyzing different patterns, without being restricted to a binary fall/no-fall decision, provides a more precise reading of patient behavior; thus, fall events can be not only detected but also potentially prevented.

VII. Conclusion

The system developed for the detection of fall events and related conditions achieved an accuracy of 94.15% in controlled environments.

The use of the YOLOv8 algorithm together with cloud-based systems enables fast processing of images captured by the camera, which are quickly interpreted and classified as "Patient in bed", "Patient moving out of bed", "Patient out of bed", "Patient in a near-fall situation", and "Patient fallen". In this way, the situation at the hospital bed can be monitored and reported in real time.

The implementation of this system in hospital beds can help identify falls in a shorter period of time, without the need for wearable devices or structural adaptations to the hospital bed. The real-time image transmission, combined with fast image processing, can identify the patient in a near-fall situation or moving out of bed and issue an alert to the care team, allowing for faster and more efficient medical intervention.

The model showed results comparable to those found in the literature, such as those reported by Nguyen et al. [5] and de Miguel et al. [6], given that five identification classes were used and the dataset is more limited when compared to those works.

Limitations and Future Work

The current work presents limitations that should be addressed in future studies. The dataset was composed exclusively of young adults aged 20 to 24 years, whereas the target population in clinical practice consists primarily of elderly patients, who may exhibit substantially different movement patterns. Furthermore, the validation was conducted exclusively in a controlled environment; experiments in real hospital settings are needed to assess the system's robustness against variable lighting, partial occlusions, and different bed configurations. Future work should also explore per-class performance optimization, particularly for the *near-fall* class, and evaluate system latency end-to-end (camera to alert) under real network conditions.

Systems for detecting patient conditions in hospital beds are gaining recognition in the research community. The developed project has proven to be efficient for detection in controlled environments and may serve as a viable alternative for the proposed problem, presenting potential for future integrations and improvements that expand its applicability to broader hospital contexts.

References

[1] Secretaria de Saúde do Distrito Federal, "Patient Safety – Fall Prevention," 2022. Available: https://www.saude.df.gov.br/documents/37101/87400/Segurana+do+Paciente++Preveno+de+Quedas.pdf. Accessed: Sept. 6, 2024.

[2] J. S. Lima, D. V. Quadros, S. L. C. Silva, J. P. Tavares, and D. D. Pai, "Custos das autorizações de internação hospitalar por quedas de idosos no Sistema Único de Saúde, Brasil, 2000-2020: um estudo descritivo," Epidemiologia e Serviços de Saúde, vol. 31, no. 1, p. e2021603, 2022, doi: 10.1590/S1679-49742022000100012.

[3] J. G. de Frutos, M. S. N. Siemon, and K. Nasrollahi, "Fall Detection in Hospital Rooms with Probabilistic Graphical Models," in Proc. International Conference on Image Processing Theory, Tools and Applications, IEEE, 2024.

[4] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Fall Detection Based on Body Part Tracking Using a Depth Camera," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 2, pp. 430–439, 2015, doi: 10.1109/JBHI.2014.2319372.

[5] V. D. Nguyen, M. T. Le, A. D. Do, H. H. Duong, T. D. Thai, and D. H. Tran, "An efficient camera-based surveillance for fall detection of elderly people," in Proc. 9th IEEE Conference on Industrial Electronics and Applications, pp. 994–997, 2014, doi: 10.1109/ICIEA.2014.6931308.

[6] K. De Miguel, A. Brunete, M. Hernando, and E. Gambao, "Home Camera-Based Fall Detection System for the Elderly," Sensors, vol. 17, no. 12, Art. no. 2864, 2017, doi: 10.3390/s17122864.

[7] N. Noury, A. Fleury, P. Rumeau, A. K. Bourke, G. O. Laighin, V. Rialle, and J. E. Lundy, "Fall Detection – Principles and Methods," in Proc. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1663–1666, 2007, doi: 10.1109/IEMBS.2007.4352627.

[8] Ultralytics, *Ultralytics YOLOv8*. Available: https://pypi.org/project/ultralytics/. Accessed: Oct. 12, 2024.

[9] Roboflow, *Roboflow: Give Your Software the Power to See Objects in Images and Video*. Available: https://roboflow.com. Accessed: Oct. 12, 2024.

[10] Mozilla, *An Overview of HTTP*. Available: https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview. Accessed: Oct. 13, 2024.

[11] Python Software Foundation, *Python Language Reference, Version 3.11*. Available: https://www.python.org. Accessed: Oct. 12, 2024.

[12] Pallets Projects, *Flask Documentation, Version 3.0.x*. Available: https://flask.palletsprojects.com/en/3.0.x/. Accessed: Oct. 12, 2024.

[13] Mozilla, *Working with JSON*. Available: https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Objects/JSON. Accessed: Oct. 12, 2024.

[14] D. Bender and K. Sartipi, "HL7 FHIR: An Agile and RESTful Approach to Healthcare Information Exchange," in Proc. 26th IEEE International Symposium on Computer-Based Medical Systems, pp. 326–331, 2013, doi: 10.1109/CBMS.2013.6627810.

[15] Twilio, *Twilio Programmable Messaging*. Available: https://www.twilio.com/en-us/messaging. Accessed: Oct. 12, 2024.

[16] A. Rashid and A. Chaturvedi, "Cloud Computing Characteristics and Services: A Brief Review," International Journal of Computer Sciences and Engineering, vol. 7, no. 2, pp. 421–426, 2019.

[17] Microsoft, *What is a Virtual Machine and How Does It Work — Microsoft Azure*. Available: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-virtual-machine. Accessed: Oct. 13, 2024.

[18] Microsoft, *Cloud Computing Services — Microsoft Azure*. Available: https://azure.microsoft.com/en-us/. Accessed: Oct. 13, 2024.

[19] Docker, *Docker: Accelerated Container Application Development*. Available: https://www.docker.com. Accessed: Oct. 13, 2024.

Análise do Impacto da Pandemia de COVID-19 no Consumo Hídrico de um Hospital de Referência em Pernambuco (2020–2024)

Augusto César Alves de Souza
Dept. de Engenharia Biomédica
Univ. Federal de Pernambuco
Recife, Brasil
augustocesar.souza@ufpe.br

Paulo Lucas Albuquerque de Aguiar
Dept. de Engenharia Biomédica
Univ. Federal de Pernambuco
Recife, Brasil
paulo.lucasa@ufpe.br

Marcus Paulo Bezerra Silva
Hospital das Clínicas HC-UFPE / HU Brasil
Recife, Brasil
marcus.silva@hubrasil.gov.br

Maurício Francisco de Oliveira
Hospital das Clínicas HC-UFPE / HU Brasil
Recife, Brasil
mauricio.oliveira@hubrasil.gov.br

Frederico Dias Nunes
Dept. de Energia Nuclear
Univ. Federal de Pernambuco
Recife, Brasil
fnunes@ufpe.br

Emery Claiton Cabral C. Lins
Dept. de Engenharia Biomédica
Univ. Federal de Pernambuco
Recife, Brasil
emery.lins@ufpe.br

Patricia Silva Lessa
Dept. de Eletrônica e Sistemas
Univ. Federal de Pernambuco
Recife, Brasil
patricia.lessa@ufpe.br

Resumo — A gestão eficiente da água em hospitais é essencial para a sustentabilidade ambiental, a redução de custos e o cumprimento dos Objetivos de Desenvolvimento Sustentável (ODS) 6 (Água Potável e Saneamento) e 12 (Consumo e Produção Responsáveis). Este estudo analisou o consumo hídrico, o volume de esgoto e os impactos financeiros no Hospital Universitário Oswaldo Cruz (HUOC/UPE), em Recife-PE, durante e após o período da pandemia COVID-19 de 2020 a 2024. Os dados mensais de consumo e custos, fornecidos pela Companhia Pernambucana de Saneamento (COMPESA), foram avaliados de forma quantitativa. No intervalo de cinco anos, registraram-se 129.173 m³ de água consumida e 132.377 m³ de esgoto gerado, com despesa total de R$ 2.776.107,50. As médias anuais foram de 25.834,6 m³ de consumo de água, 26.475,4 m³ de volume de esgoto e R$ 555.221,50 em custos. Os resultados demonstraram o forte impacto da pandemia de COVID-19 nas operações e nos custos hospitalares entre 2020 e 2022, seguido por um período de estabilização do consumo em 2023 e 2024, evidenciando a relevância do monitoramento sistemático e de auditorias operacionais para a sustentabilidade de hospitais públicos.

Palavras-chave — Gestão Hídrica Hospitalar; Sustentabilidade; Saúde Pública; ODS; Eficiência de Recursos.

I. INTRODUÇÃO

O gerenciamento dos recursos ambientais e financeiros em estabelecimentos assistenciais de saúde é um componente crítico da governança contemporânea em saúde pública. Entre esses insumos, a água se destaca como um recurso estratégico vital, indispensável não apenas para a subsistência humana e higiene básica, mas também para a viabilização de procedimentos de alta complexidade, tais como diálise, esterilização de materiais cirúrgicos, processamento de roupas hospitalares e climatização de ambientes críticos[1,2,3].

O Hospital Universitário Oswaldo Cruz (HUOC), vinculado à Universidade de Pernambuco (UPE), atua como um importante pólo assistencial do Sistema Único de Saúde (SUS) do Estado de Pernambuco. Caracterizado como um hospital universitário integrante do Complexo Hospitalar Enio Lustosa Cantarelli, o HUOC desempenha funções assistenciais de média e alta complexidade, além de exercer papel estratégico na formação acadêmica e no desenvolvimento científico regional[4]. Sua infraestrutura horizontal, localizada no bairro de Santo Amaro em Recife-PE, estende-se por uma área de 18.000 m² distribuída em 10 pavilhões, abrigando uma capacidade instalada de 487 leitos conforme registro no Cadastro Nacional de Estabelecimentos de Saúde - CNES[4,5].

Dada a magnitude de sua operação, que abrange desde consultas ambulatoriais especializadas até transplantes de órgãos e terapias oncológicas complexas, o consumo de água e a consequente geração de esgoto representam importância no orçamento operacional da instituição[4,5]. O consumo hídrico é amplamente afetado por flutuações na taxa de ocupação dos leitos, variações sazonais de patologias e emergências sanitárias, como a pandemia da COVID-19[6].

Embora diversos estudos tenham avaliado o consumo hídrico hospitalar durante a pandemia[6,7,8], poucos investigaram simultaneamente a dinâmica operacional da infraestrutura hídrica, os efeitos dos reajustes tarifários e a existência de mudanças na série temporal por meio de métodos estatísticos de detecção de rupturas[9,10,11].

Diante disso, este artigo analisou o perfil do consumo hídrico, do volume de esgoto e do correspondente impacto financeiro no HUOC durante e após o período da pandemia COVID-19 de 2020 a 2024. A relevância desta análise alinha-se diretamente com os Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU), especificamente o ODS 6 (Água Potável e Saneamento) e o ODS 12 (Consumo e Produção Responsáveis), além de repercutir no ODS 3 (Saúde e

Bem-estar) ao promover a infraestrutura necessária para a segurança do paciente[2].

Com base na literatura e nas mudanças operacionais observadas durante a pandemia, este estudo testou as seguintes hipóteses:

H1: O período da pandemia esteve associado ao maior consumo de água em decorrência da intensificação dos protocolos de higienização, da ampliação da capacidade assistencial e da expansão dos leitos destinados ao tratamento de pacientes infectados.

H2: A evolução dos custos hídricos hospitalares não foi determinada exclusivamente pelo volume consumido, sendo influenciada também pelos reajustes tarifários e pelas alterações regulatórias aplicadas ao serviço público de abastecimento de água e esgoto sanitário.

H3: A redução progressiva da demanda assistencial associada à COVID-19 promoveu uma reorganização da dinâmica operacional da infraestrutura hídrica hospitalar, resultando em uma mudança do consumo de água e na transição para um regime operacional pós-pandemia estatisticamente distinto daquele observado durante os períodos de maior pressão assistencial.

As principais contribuições deste estudo são:
- Apresentar uma caracterização temporal do comportamento hídrico hospitalar, identificando diferentes estágios operacionais associados às mudanças epidemiológicas e regulatórias observadas entre 2020 e 2024.
- Demonstrar a influência conjunta dos fatores assistenciais e tarifários sobre a evolução dos custos de abastecimento de água em hospitais públicos, contribuindo para modelos de planejamento orçamentário e gestão sustentável dos recursos hídricos.
- Disponibilizar evidências que possam subsidiar auditorias operacionais, políticas públicas de sustentabilidade hospitalar e estratégias de gestão de infraestrutura, alinhadas aos Objetivos de Desenvolvimento Sustentável (ODS) 3 (Saúde e Bem-estar), ODS 6 (Água Potável e Saneamento) e ODS 12 (Consumo e Produção Responsáveis).

II. MATERIAIS E MÉTODOS

Este trabalho constitui-se como um estudo de caso quantitativo, descritivo e retrospectivo, focado na avaliação dos padrões de consumo de água potável, descarte de efluentes e evolução dos custos hídricos do HUOC durante e após a pandemia COVID-19. A unidade de análise corresponde ao imóvel institucional cadastrado sob a matrícula número 057774649 junto à Companhia Pernambucana de Saneamento (COMPESA), concessionária pública responsável pelo abastecimento hídrico e coleta de esgoto na região metropolitana do Recife.

A série temporal compreende 60 meses observados entre janeiro de 2020 e dezembro de 2024. Os dados secundários relativos aos volumes medidos/faturados em metros cúbicos (m³) e aos valores financeiros expressos em Moeda Corrente Nacional (R$) foram extraídos diretamente das faturas de serviços emitidas mensalmente pela concessionária. A análise englobou a caracterização do perfil operacional por meio do balanço hídrico, definido pela relação percentual entre o volume de esgoto registrado e o consumo de água faturado.

Adicionalmente, investigamos a evolução da estrutura tarifária aplicada ao perfil da instituição durante o quinquênio, mapeando os reajustes regulatórios autorizados e suas respectivas fórmulas de cálculo. Para o processamento estatístico e avaliação de sazonalidade, os dados foram agregados e segmentados por quadrimestres (Primeiro Quadrimestre: janeiro-abril; Segundo Quadrimestre: maio-agosto; Terceiro Quadrimestre: setembro-dezembro), permitindo isolar a volatilidade decorrente de fatores sazonais e operacionais de acordo com os respectivos períodos anuais. A agregação quadrimestral foi adotada para reduzir oscilações mensais sem comprometer a resolução temporal necessária para identificação das tendências anuais.

Foram levantados e analisados os principais indicadores assistenciais do HUOC ao longo do período de estudo, incluindo a evolução da capacidade instalada de leitos destinados ao atendimento da COVID-19, a distribuição por nível de complexidade assistencial, o número de hospitalizações relacionadas à doença e indicadores de desempenho hospitalar, tais como a Taxa de Ocupação Hospitalar (TOH), o Tempo Médio de Permanência (TMP) e a Taxa de Mortalidade Hospitalar (TMH) obtidos através de informações operacionais publicadas nos Anuários Estatísticos do Complexo Hospitalar da UPE[12,13,14]. Esses indicadores foram utilizados para contextualizar as alterações observadas no comportamento do consumo hídrico e interpretar sua relação com a dinâmica operacional da instituição durante e após a pandemia.

A validação das hipóteses foi conduzida por meio de diferentes métodos estatísticos, selecionados de acordo com a natureza de cada hipótese investigada[15].

Para a Hipótese H1, que avaliou a diferença no consumo hídrico entre os períodos crítico e pós crítico da pandemia, empregou-se o teste *t* de Welch, apropriado para comparação entre médias de grupos independentes com variâncias potencialmente distintas.

A Hipótese H2, referente à relação entre consumo de água e impacto financeiro, foi investigada inicialmente por meio do coeficiente de correlação de Pearson e, posteriormente, por um modelo de regressão linear múltipla, incorporando simultaneamente o volume consumido e os diferentes períodos tarifários. Para a Hipótese H3, que analisou a reorganização do padrão operacional de consumo após a pandemia, aplicou-se o Teste de Pettitt, para identificação de mudanças estruturais na série temporal, seguido de um modelo de Regressão Segmentada por Séries Temporais Interrompidas *(Interrupted Time Series – ITS)*, utilizado

para estimar a mudança imediata no nível de consumo, a alteração da tendência temporal e a estabilização do novo regime operacional. Em todas as análises adotou-se nível de significância de 5% ($\alpha = 0,05$), sendo os resultados apresentados com respectivos valores de p e intervalos de confiança de 95% quando aplicáveis.

A ruptura estatística considerada foi identificada pelo teste de Pettitt (maio/2022). Contudo, para avaliar a consolidação do regime pós-pandemia utilizamos janeiro de 2023 como marco correspondente ao encerramento das medidas emergenciais.

Antes da aplicação do coeficiente de correlação de Pearson, a relação entre consumo hídrico e custo mensal foi analisada por meio de gráfico de dispersão, evidenciando comportamento aproximadamente linear. A análise de Pearson forneceu evidências da associação linear positiva extremamente forte ($r = 0,9801$; $p < 0,001$), sustentando a utilização de modelos lineares para investigar essa relação. Para avaliar os pressupostos da regressão linear múltipla e do modelo de Séries Temporais Interrompidas *(Interrupted Time Series – ITS),* foram realizados diagnósticos adicionais. A independência dos resíduos foi investigada pelo teste de Durbin-Watson (DW = 1,382), cujo resultado indicou autocorrelação serial positiva de baixa a moderada. A homocedasticidade foi avaliada pelo teste de Breusch–Pagan, que indicou evidências de heterocedasticidade (LM = 10,89; p = 0,0279), demonstrando que a variância dos resíduos não pode ser considerada constante.

A normalidade dos resíduos foi examinada pelos testes de Jarque-Bera (JB = 13,205; p = 0,0014) e Shapiro-Wilk (W = 0,919; p = 0,0007), que evidenciaram desvio da normalidade. Embora esses resultados indiquem violações dos pressupostos clássicos da regressão linear, não foram aplicadas técnicas de correção da inferência, como erros-padrão robustos ou procedimentos de reamostragem (*bootstrap*). Contudo, considerando o tamanho amostral (n = 60) e a robustez assintótica dos estimadores de mínimos quadrados ordinários, os modelos foram mantidos para fins de inferência estatística.

A multicolinearidade foi avaliada pelo Fator de Inflação da Variância (VIF), cujos valores variaram entre 1,41 e 2,44, indicando baixa colinearidade entre as variáveis explicativas e estabilidade dos coeficientes estimados.

A tendência temporal anterior apresentou declínio significativo (β=-50,16; p=0,001). Após a intervenção institucional de janeiro de 2023 observou-se alteração significativa da inclinação da série (β=69,82; p=0,023), indicando mudança na dinâmica temporal do consumo. Entretanto, a redução imediata do nível médio não foi estatisticamente significativa (β=-682,70; p=0,155). Considerando especificamente o período pós-pandemia (2023–2024), a tendência estimada de 19,66 m³/mês não diferiu significativamente de zero (p=0,456), caracterizando estabilização do consumo em um novo regime operacional.

A metodologia adotada neste estudo está sintetizada no fluxograma apresentado na Fig. 1, que ilustra as etapas sequenciais de coleta, processamento, análise e interpretação dos dados.

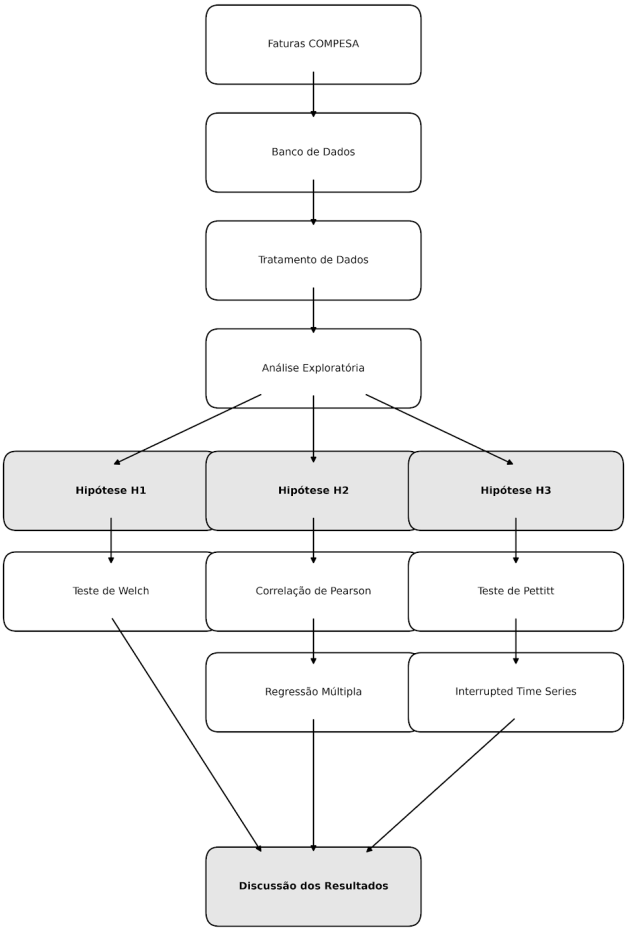


Fig. 1 - Etapas da pesquisa

III. RESULTADOS E DISCUSSÃO

A. Panorama Geral Quinquenal

No acumulado do período investigado (2020–2024), o HUOC contabilizou um consumo hídrico total de 129.173 m³ de água e uma geração de efluentes mensurada em 132.377 m³ de esgoto. Esse fluxo de consumo e descartes resultou em um dispêndio financeiro global de R$ 2.776.107,50 junto à concessionária. Em termos de médias anuais, a instituição demandou 25.834,6 m³ de água, gerou 26.475,4 m³ de esgoto e incorreu em custos anuais médios da ordem de R$ 555.221,50.

A importância do HUOC como centro de referência para o tratamento de patologias infectocontagiosas no Estado de Pernambuco gerou a mobilização e reestruturação de sua capacidade para o fornecimento de leitos dedicados à COVID-19. Em 2020, sob demanda da Secretaria de Saúde do Estado (SES/PE), a instituição disponibilizou 183 leitos para tratamento da pandemia. A distribuição dessas vagas ocorreu sendo 135 leitos de enfermaria/urgência e 48 leitos

de Unidade de Terapia Intensiva (UTI), conforme observado na Fig. 2.

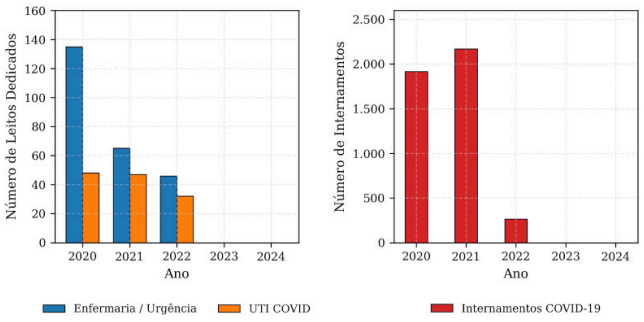


Fig. 2 - Enfrentamento à Pandemia por SARS-CoV-2 no HUOC UPE, fonte SCH/UPE

Esse arranjo de contingência sofreu um processo de desmobilização e redução progressiva ao longo dos anos subsequentes. Esse recuo foi impulsionado pela inserção de novos serviços de saúde na rede de assistência do Estado, pelas flutuações dos cenários epidemiológicos locais e pelo avanço expressivo da cobertura vacinal. Como resultado, a infraestrutura de enfermarias/urgência foi reduzida para 65 leitos em 2021 e 46 leitos em 2022. Os leitos críticos de UTI foram reduzidos para 47 em 2021 e 32 em 2022, até a completa desmobilização física nos anos posteriores.

No ano de 2020, o serviço registrou 3.650 atendimentos na triagem/emergência direcionados a pacientes com suspeita ou confirmação de infecção por COVID-19. Desses, 1.914 indivíduos necessitaram de admissão e internamento hospitalar. O índice de positividade laboratorial foi de 42,8%, consistindo em 819 internamentos com diagnóstico confirmado para COVID-19. A série de internamentos sob suspeita ou confirmação registrou seu pico em 2021, atingindo 2.170 ocorrências, antes de sofrer uma expressiva redução de 87,8% em 2022, quando computou apenas 265 internamentos, conforme demonstra a Fig. 3.

Devido à ausência de um serviço de pronto-atendimento de "porta aberta" convencional, o hospital operou por meio de seu serviço de triagem de doenças infecciosas para adultos e crianças. Consequentemente, o fluxo de admissão de pacientes foi realizado ou por encaminhamentos regulados externos, intermediados pela Central de Regulação Hospitalar da SES/PE, ou por meio de demanda interna do próprio hospital.

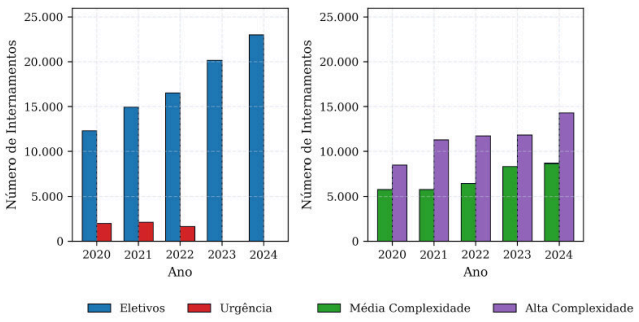


Fig. 3 - Números de internamento e nível de complexidade, fonte SCH/UPE

O ano de 2022 marcou o início da descompressão do sistema de saúde após as ondas críticas da pandemia de COVID-19. Observou-se a continuidade do crescimento das internações impulsionada por uma retomada assistencial programada. O HUOC registrou um aumento de 11% nas internações eletivas de pacientes regulados, um reflexo direto da reativação de leitos clínicos e cirúrgicos operados em regime de rotina.

Concomitantemente, identificou-se uma redução nas internações por urgência motivada pela queda sustentada do número de casos suspeitos e confirmados de COVID-19. Sob a perspectiva de severidade, a alta complexidade expandiu-se em 4%.

O comportamento de fluxo em 2023 revelou uma assimetria nas taxas de crescimento entre os níveis de complexidade. O HUOC apresentou uma variação positiva expressivamente maior no quantitativo de internamentos de média complexidade (+29%) em comparação aos de alta complexidade (+1%). Este desalinhamento temporário de crescimento foi determinado por uma resposta emergencial a um surto epidemiológico de Síndrome Respiratória Aguda Grave (SRAG) ocorrido no primeiro semestre de 2023. O hospital absorveu os casos expandindo em 50% os internamentos pediátricos de média complexidade.

Para além da resposta à crise sanitária infantil, o segmento de média complexidade foi sustentado pela reativação da demanda reprimida de procedimentos regulados, registrando incremento de 37% nos internamentos cirúrgicos e 19% nos clínicos. Em 2024, o HUOC atingiu 23.008 internamentos (alta de 14% em relação ao ano anterior). Por não possuir um serviço de urgência/emergência de "porta aberta" e amparada por uma queda substancial de internamentos regulados para as triagens de Doenças Infectocontagiosas adulto e infantil, o HUOC optou por considerar 100% dos internamentos como eletivos.

Considerando que pacientes acometidos por quadros de Síndrome Respiratória Aguda Grave (SRAG), sejam suspeitos ou confirmados para COVID-19, apresentaram disfunções ventilatórias que exigiram um tempo médio de permanência (TMP) hospitalar prolongado para reabilitação, analisamos os indicadores de tempo de ocupação, permanência e mortalidade considerando o período de 2019 (pré-pandemia) a 2024(pós-pandemia) conforme apresentado na Fig. 4.

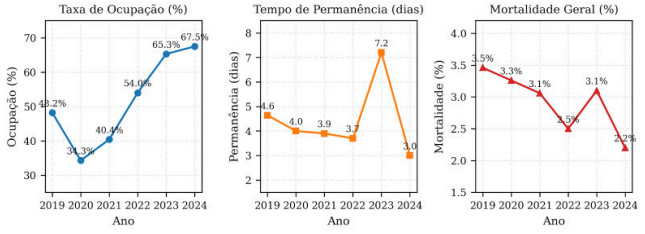


Fig. 4 - Indicadores de ocupação, permanência e mortalidade do HUOC, fonte SCH/UPE

No período pré-pandemia (em 2019), o hospital registrou uma Taxa de Ocupação Hospitalar (TOH) de 48,2%. Com o advento da emergência sanitária em 2020, o indicador

sofreu uma redução significativa, atingindo o mínimo de 34,27%. Esse recuo de 13,93% decorreu da necessidade de bloqueio de leitos de rotina, da suspensão temporária de cirurgias eletivas e da alocação de áreas físicas exclusivas para o isolamento de doenças infectocontagiosas. A partir de 2021, ocorreu uma recuperação gradual, com a TOH subindo para 40,4% em 2021 e 54,0% em 2022. Esse movimento de expansão acelerou progressivamente em 2023, alcançando 65,3%, impulsionado pelo acréscimo de 11% nos internamentos hospitalares totais e pela expansão de 22% nos procedimentos cirúrgicos eletivos. Em 2024, a taxa atingiu seu ponto máximo com 67,5% de ocupação.

O Tempo Médio de Permanência (TMP) apresentou um padrão de estabilidade durante os primeiros quatro anos da série histórica, oscilando de 4,64 dias em 2019 para 4,00 dias em 2020, 3,90 dias em 2021 e 3,70 dias em 2022. No entanto, em 2023 o indicador subiu para 7,20 dias, representando um acréscimo de 3,5 dias em relação ao ano anterior. A reversão desse indicador ocorreu em 2024, quando o TMP registrou uma queda acentuada de 4,2 dias frente a 2023, estabelecendo o menor índice de 3,00 dias. A redução no tempo de internação elevou o giro de leitos e permitiu absorver o aumento da demanda sem sobrecarregar a capacidade física da instituição, demonstrando eficiência dos processos de regulação interna.

A Taxa de Mortalidade Hospitalar (TMH) manteve uma trajetória de declínio sustentado ao longo de quase todo o período analisado, evoluindo de 3,46% em 2019 para 3,26% em 2020, 3,06% em 2021 e atingindo 2,50% em 2022. Esses resultados demonstraram a resiliência dos protocolos assistenciais do hospital mesmo durante os anos mais críticos da crise sanitária. Em 2023, observamos uma discreta elevação do indicador para 3,10%. Em 2024, a TMH registrou uma redução de 0,9%, alcançando o menor valor com 2,20%. A convergência entre a maior taxa de ocupação do período (67,5%), o menor tempo de permanência (3,00 dias) e o menor índice de mortalidade geral (2,20%) indicam que o ganho de produtividade e a aceleração do giro de leitos ocorreram acompanhados pela melhoria da qualidade e da segurança do cuidado prestado aos pacientes.

B. Estágios Operacionais do Balanço Hídrico

A análise da relação entre o consumo de água e o esgoto sanitário permitiu identificar três estágios operacionais distintos na série histórica, evidenciando modificações no comportamento do consumo e nas metodologias de faturamento da concessionária:

1) Estágio 1 (Janeiro de 2020 a Agosto de 2021): Caracterizou-se por elevados volumes absolutos de consumo de água induzidos pelo ápice da crise sanitária da COVID-19. Nesse estágio, o balanço hídrico manteve-se estritamente estável, com o volume de esgoto faturado fixado em equivalência exata a 100% do consumo de água medido. Foram registrados picos expressivos de consumo associados à intensificação das atividades de higienização, desinfecção e abertura de leitos críticos de UTI, destacando-se os meses de novembro de 2020 (4.245 m³) e julho de 2021 (4.331 m³), contrapondo-se a períodos de menor intensidade assistencial como março de 2020 (1.649 m³) e fevereiro de 2021 (1.374 m³).

2) Estágio 2 (Setembro de 2021 a Dezembro de 2022): Este estágio foi marcado por uma mudança metodológica no registro dos efluentes pela concessionária, rompendo o equilíbrio de 100% e gerando a ocorrência de volumes de esgoto superiores ao volume de água consumido. Em setembro de 2021, o consumo hídrico foi de 3.429 m³ frente a 3.617 m³ de esgoto. Em outubro do mesmo ano, registrou-se o teto máximo de consumo do ano, com 4.618 m³ de água e 4.694 m³ de esgoto. Em dezembro de 2021, o consumo de 3.422 m³ gerou 3.577 m³ de esgoto. Concomitantemente, o ano de 2022 representou o início de um declínio acentuado no consumo físico de água, registrando mínimos históricos em maio (658 m³) e julho (799 m³) e máximos em janeiro (3.866 m³) e março (2.907 m³).

3) Estágio 3 (Anos de 2023 e 2024): Consolidou a estabilização de um novo patamar operacional pós-pandêmico. Observou-se a manutenção do excedente de esgoto em relação à água em todos os meses, porém sob patamares de consumo de água drasticamente inferiores aos anos iniciais. Em 2023, o consumo hídrico oscilou entre o mínimo atípico de 303 m³ (abril) e o máximo de 1.692 m³ (agosto). No ano de 2024, verificou-se comportamento assemelhado, com o consumo variando de 846 m³ (março) a 1.828 m³ (agosto).

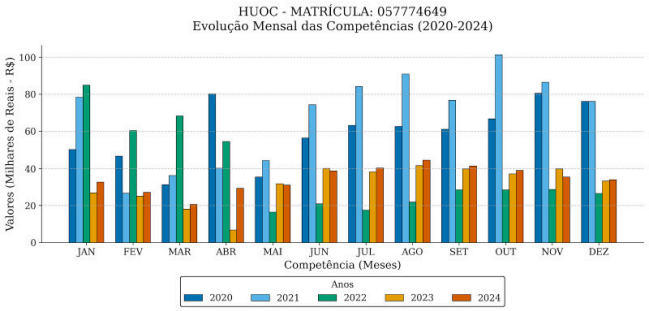


Fig. 5 - Evolução do desempenho hídrico do HUOC 2020/2024, fonte COMPESA

C. Evolução da Estrutura Tarifária

O impacto financeiro sofrido pela instituição não decorre exclusivamente do volume de água/esgoto demandado, mas também da evolução da estrutura de tarifas aplicadas pela concessionária. O modelo de faturamento comercial da COMPESA baseia-se em uma taxa fixa de disponibilidade de serviço aplicada aos primeiros 10 m³ consumidos, acrescida de um valor variável indexado por metro cúbico excedente.

Ao longo da série histórica, identificaram-se três reajustes tarifários. No ano de 2020, as tarifas vigentes estipulavam R$ 62,67 para a faixa de consumo básico (até 10 m³) e R$ 9,50 para cada metro cúbico excedente. Em janeiro de 2021, o primeiro reajuste elevou a taxa básica para R$ 64,17 e a tarifa excedente para R$ 9,73 por m³. Em agosto de 2021, um segundo reajuste elevou os parâmetros para R$ 71,81 (faixa básica) e R$ 10,89 (excedente), coincidindo com o período imediatamente anterior à mudança metodológica do cálculo de esgoto.

Esse modelo permaneceu inalterado ao longo de todo o ano de 2022. Por fim, em maio de 2023, foi implementado o terceiro reajuste, fixando a taxa básica em R$ 79,86 e a tarifa excedente em R$ 12,11 por m³, valor que permaneceu vigente até o encerramento de 2024.

VIGÊNCIA INICIAL	TARIFA BÁSICA (ATÉ 10 M³)	TARIFA EXCEDENTE (POR M³)
Janeiro de 2020	R$ 62,67	R$ 9,50
Janeiro de 2021	R$ 64,17	R$ 9,73
Agosto de 2021	R$ 71,81	R$ 10,89
Maio de 2023	R$ 79,86	R$ 12,11

Tabela 1 - Evolução cronológica da estrutura tarifária (COMPESA)

A flutuação financeira também sofre influência da composição tributária e regulatória inerente ao faturamento de serviços de saneamento básico. Em termos de tributação federal, as faturas de água e esgoto estão sujeitas à incidência do Programa de Integração Social (PIS) e da Contribuição para o Financiamento da Seguridade Social (COFINS). Estes tributos incidem sobre a receita bruta da concessionária e são repassados ao consumidor.

No âmbito estadual, o fornecimento de água tratada por concessionárias de serviço público possui natureza jurídica de serviço público essencial, não configurando circulação de mercadoria. Dessa forma, amparado por jurisprudência consolidada do Supremo Tribunal Federal (STF RE 607.056/RJ), não há incidência do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS) sobre o valor da fatura, o que difere a matriz de custos hídricos da matriz de custos com energia elétrica.

D. Sazonalidade Quadrimestral e Volatilidade Financeira

A análise dos custos mensais refletiu a interação entre a demanda assistencial hospitalar e os reajustes tarifários descritos, revelando uma flutuação sazonal quando segregada por quadrimestres, conforme apresentado na Fig. 6.

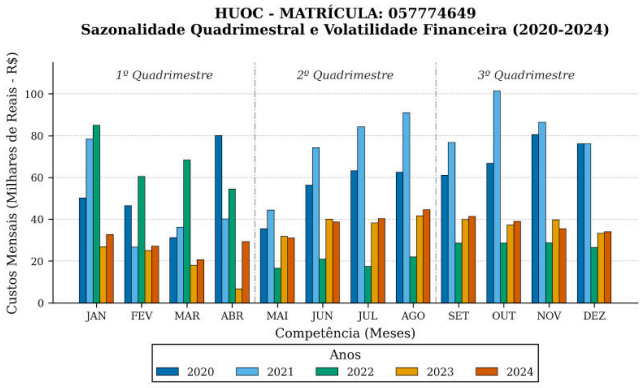


Fig. 6 - Sazonalidade quadrimestral e volatilidade financeira, fonte COMPESA

A análise quadrimestral dos custos evidenciou elevada variabilidade durante o período da pandemia, seguida de estabilização no biênio pós-pandemia. O primeiro quadrimestre apresentou as maiores oscilações, com picos entre 2021 e 2022, associados ao aumento da demanda assistencial decorrente da COVID-19 e de surtos sazonais de arboviroses, seguidos por redução expressiva dos custos em 2023 e estabilização em 2024. O segundo quadrimestre exibiu comportamento semelhante, com custos elevados em 2020–2021, redução acentuada em 2022 e posterior estabilização, refletindo a diminuição da pressão assistencial e a retomada gradual das atividades eletivas. O terceiro quadrimestre apresentou menor variabilidade relativa, embora tenha concentrado o maior custo mensal de toda a série histórica, registrado em outubro de 2021 (R$ 101.333,50), resultado da combinação entre o elevado consumo durante a pandemia e os reajustes tarifários implementados pela concessionária. Nos anos de 2023 e 2024, os custos permaneceram relativamente estáveis em todos os quadrimestres, caracterizando um novo padrão operacional pós-pandemia.

E. Desempenho hídrico anual

Nos dois primeiros anos analisados, a instituição manteve um patamar de consumo acima de 37.000 m³ anuais, atingindo o pico de 39.328 m³ em 2021. Este volume acentuado correlaciona-se diretamente com o período de enfrentamento crítico da pandemia de COVID-19, onde os protocolos de higienização, desinfecção e operação contínua de blocos de internação demandaram uma carga hidráulica máxima da infraestrutura predial conforme demonstrado na Fig. 7.

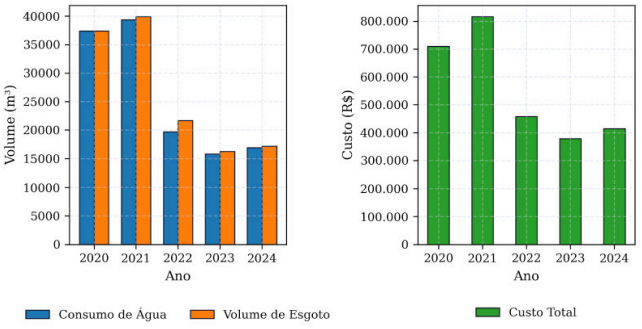


Fig. 7 - Desempenho anual do volume de água, esgoto e custos, fonte COMPESA

A partir de 2022, observou-se uma redução abrupta de 49,9% no consumo de água (decréscimo para 19.703 m³), estabilizando-se em uma média de 16.546 m³ no biênio subsequente (2023–2024).

Um aspecto relevante observado neste estudo foi a divergência entre o volume de água medido no hidrômetro principal e o volume de esgoto faturado pela concessionária. Embora, em condições usuais, o volume de esgoto corresponda a uma fração da água consumida (fator de retorno), verificou-se que, a partir de setembro de 2021, o esgoto faturado passou a superar o consumo de água, alcançando diferença de 9,95% (1.959 m³) em 2022. Essa divergência é compatível com a utilização de fontes alternativas de abastecimento, como poços artesianos e caminhões-pipa destinados a setores críticos (como

Hemodiálise e CME), cujos efluentes são lançados na rede pública sem que a água correspondente seja registrada pelo hidrômetro principal. Esse cenário possui impacto para a governança contratual e tarifária do serviço de saneamento, justificando a implantação de hidrometria setorizada nas fontes alternativas e a revisão, junto à COMPESA, do fator de retorno de efluentes e dos critérios de faturamento do volume de esgoto, visando aprimorar a confiabilidade do balanço hídrico do hospital.

O desempenho hídrico também foi avaliado por meio dos indicadores L/leito-dia e m³/m²/ano. Como os dados de consumo referem-se à unidade consumidora nº 057774649 da COMPESA, enquanto a área construída (18.000 m²) e os 487 leitos correspondem ao complexo do HUOC, esses parâmetros foram utilizados como *proxies* institucionais. Embora a comparação dos indicadores normalizados entre as fases pré-pandêmica, crítica e pós-pandêmica seja desejável, a ausência de dados de consumo hídrico anteriores a 2020 impossibilitou o cálculo desses indicadores para a fase pré-pandêmica. Assim, a análise dos indicadores normalizados foi restrita ao período de 2020 a 2024. Na fase crítica da COVID-19, o consumo específico atingiu 558,48 L/leito-dia (2020), 961,99 L/leito-dia (2021) e 692,02 L/leito-dia (2022), evidenciando que o consumo hídrico não reduziu na mesma proporção da capacidade assistencial. O indicador de intensidade de consumo por área diminuiu de 2,078 para 0,940 m³/m²/ano entre 2020 e 2024, correspondendo a uma redução de 54,8%. Como a área construída permaneceu constante ao longo do período analisado, esse indicador mostrou-se mais apropriado para comparações temporais e para *benchmarking* com instituições que utilizem métricas padronizadas e critérios metodológicos comparáveis. Os resultados, entretanto, devem ser interpretados considerando as limitações inerentes ao uso de *proxies* institucionais.

Para avaliar a hipótese H1, investigamos o impacto da pandemia de COVID-19 sobre a carga operacional da infraestrutura hídrica hospitalar por meio da comparação do consumo mensal de água entre o período da pandemia (2020–2022; $n = 36$ meses) e o período pós-pandemia (2023–2024; $n = 24$ meses). A comparação das médias foi realizada utilizando o teste t de Welch, adequado para amostras independentes com variâncias potencialmente distintas. Os resultados evidenciaram diferença estatisticamente significativa entre os dois períodos ($t = 6{,}128$; gl = 43,11; $p < 0{,}001$), demonstrando que o consumo hídrico durante a pandemia foi substancialmente superior ao observado após a retomada das condições operacionais regulares. Esse comportamento é consistente com a intensificação dos protocolos de controle de infecção, ao aumento da frequência dos procedimentos de higienização, à ampliação temporária da capacidade assistencial e às adaptações operacionais impostas pelo enfrentamento da COVID-19. Diante disso, os resultados fornecem evidências quantitativas que corroboram a hipótese H1, indicando que a pandemia exerceu impacto significativo sobre a demanda da infraestrutura hídrica hospitalar.

Para testar a hipótese H2, a relação entre o consumo hídrico e os custos operacionais foi inicialmente avaliada por meio do coeficiente de correlação de Pearson, o qual revelou correlação linear positiva extremamente forte entre o volume de água consumido e o custo mensal correspondente ($r = 0{,}980$; $p < 0{,}001$). Embora esse resultado indique que o aumento do consumo está diretamente associado à elevação das despesas, a elevada correlação, isoladamente, não permite distinguir a contribuição dos reajustes tarifários ocorridos ao longo do período de estudo. Dessa forma, foi desenvolvido um modelo de regressão linear múltipla incorporando simultaneamente o consumo de água e os diferentes períodos tarifários estabelecidos pela concessionária. O modelo apresentou elevado poder explicativo ($R^2 = 0{,}993$), indicando que aproximadamente 99,3% da variabilidade observada nos custos foi explicada pelas variáveis independentes consideradas. A análise dos coeficientes indicou que tanto o consumo de água ($\beta = 21{,}17$; $p < 0{,}001$) quanto os reajustes tarifários permaneceram estatisticamente significativos, evidenciando que a evolução dos custos operacionais resulta da interação entre fatores físicos, representados pelo volume consumido, e fatores regulatórios, associados às alterações na estrutura tarifária. Esses achados confirmam a hipótese H2 e demonstram que o comportamento financeiro do sistema hídrico hospitalar não pode ser explicado exclusivamente pela demanda de água, sendo também influenciado pelas políticas tarifárias adotadas ao longo do período.

Para avaliar a hipótese H3, aplicamos inicialmente o Teste de Pettitt para identificar mudanças estruturais na série temporal de consumo hídrico. O teste detectou uma ruptura altamente significativa em maio de 2022 ($K = 864{,}0$; $p = 2{,}77 \times 10^{-9}$), evidenciando a transição para um novo regime estatístico de consumo. Embora o teste não identifique relações causais, o ponto de ruptura coincide temporalmente com a redução das medidas emergenciais implementadas durante a pandemia de COVID-19, sugerindo uma reorganização operacional da infraestrutura hospitalar.

A hipótese H3 foi avaliada por meio da combinação do teste não paramétrico de Pettitt e da Regressão Segmentada de Série Temporal Interrompida (ITS). O teste de Pettitt foi empregado exclusivamente para detectar a ruptura estatística, enquanto janeiro de 2023 foi definido *a priori* como marco operacional do período pós-pandemia.
O teste de Pettitt identificou uma mudança estrutural altamente significativa na série temporal ($K = 864{,}0$; $p = 2{,}77 \times 10^{-9}$), localizada na observação 29 (maio de 2022), indicando alteração da dinâmica do consumo hídrico antes do marco operacional adotado para a ITS. Considerando janeiro de 2023 como início do período pós-pandemia e marco operacional da intervenção na ITS, a regressão segmentada evidenciou tendência decrescente significativa no período pré-intervenção ($\beta_1 = -50{,}16$ m³/mês; $p = 0{,}001$). A mudança imediata no nível médio após a intervenção não foi estatisticamente significativa ($\beta_2 = -682{,}70$ m³; $p = 0{,}155$). Em contrapartida, observou-se alteração significativa da tendência temporal ($\beta_3 = 69{,}82$ m³/mês; $p = 0{,}023$), indicando mudança na dinâmica de consumo após a reorganização operacional. Adicionalmente, a análise restrita ao período de 2023–2024 evidenciou tendência de 19,66 m³/mês, sem significância estatística ($p = 0{,}456$),

caracterizando estabilização do consumo em um novo regime operacional.

IV. CONSIDERAÇÕES FINAIS

A análise integrada dos indicadores operacionais, assistenciais, financeiros e estatísticos demonstrou que as mudanças observadas no consumo hídrico acompanharam a reorganização da infraestrutura hospitalar e da capacidade assistencial ao longo do período investigado.

Os resultados confirmaram a hipótese H1, evidenciando que o período pandêmico apresentou consumo hídrico significativamente superior ao período pós-pandemia (teste *t* de Welch: $t = 6{,}128$; $p < 0{,}001$). Esse comportamento é consistente com a ampliação temporária da capacidade instalada, a intensificação dos protocolos de higienização, o aumento das internações relacionadas à COVID-19 e a maior utilização da infraestrutura hospitalar durante a emergência sanitária.

A hipótese H2 também foi corroborada. Embora tenha sido observada correlação linear extremamente forte entre consumo de água e custos operacionais ($r = 0{,}980$; $p < 0{,}001$), a regressão linear múltipla indicou que a evolução das despesas não depende exclusivamente do volume consumido. O elevado poder explicativo do modelo ($R^2 = 0{,}993$) evidenciou que tanto o consumo de água/esgoto quanto os sucessivos reajustes tarifários exerceram influência estatística sobre os custos, ressaltando que o planejamento orçamentário hospitalar deve considerar simultaneamente fatores operacionais e regulatórios.

Em relação à hipótese H3, o Teste de Pettitt identificou uma ruptura altamente significativa na série temporal em maio de 2022 ($K = 864{,}0$; $p < 0{,}001$), indicando a transição entre o regime operacional associado à pandemia e um novo padrão de funcionamento. A Regressão Segmentada forneceu evidências da redução imediata do nível médio de consumo após essa ruptura, enquanto a análise específica do período de 2023–2024 indicou ausência de tendência temporal significativa ($p = 0{,}456$), caracterizando a estabilização do consumo em um novo patamar operacional. Os resultados sugerem a consolidação de um novo padrão operacional no período observado (2023–2024).

Além dos aspectos quantitativos, a identificação de três estágios operacionais distintos do balanço hídrico evidenciou alterações relevantes tanto no comportamento do consumo quanto no método de faturamento dos serviços de água e esgoto. A divergência observada entre os volumes de água consumida e de esgoto faturado após 2021 sugere a necessidade de auditorias técnicas mais aprofundadas sobre os critérios de medição e cobrança adotados pela concessionária, bem como sobre possíveis fontes alternativas de abastecimento utilizadas pela instituição.

Do ponto de vista da gestão hospitalar, os resultados fornecem evidências que podem subsidiar estratégias de eficiência hídrica, planejamento orçamentário, auditorias operacionais e políticas institucionais de sustentabilidade, contribuindo para a otimização do uso dos recursos naturais em hospitais públicos de média e alta complexidade. As evidências produzidas também reforçam o alinhamento da gestão da infraestrutura hospitalar aos Objetivos de Desenvolvimento Sustentável, especialmente os ODS 3 (Saúde e Bem-Estar), ODS 6 (Água Potável e Saneamento) e ODS 12 (Consumo e Produção Responsáveis).

Como limitações, este estudo corresponde a um estudo de caso desenvolvido em um único hospital universitário, utilizando dados administrativos de faturamento e consumo. Embora os resultados apresentem consistência, sua generalização para outras instituições deve ser realizada com cautela. Pesquisas futuras poderão ampliar esta abordagem por meio de análises multicêntricas envolvendo hospitais de diferentes portes e perfis assistenciais, bem como incorporar indicadores de eficiência hídrica, *benchmarking*, técnicas de *Data Envelopment Analysis* (DEA), modelos de previsão por séries temporais e sistemas de monitoramento inteligente para apoiar a gestão da infraestrutura hospitalar.

REFERÊNCIAS

[1] World Health Organization, *Global Progress Report on Water, Sanitation and Hygiene in Health Care Facilities: Fundamentals First*. Geneva, Switzerland: World Health Organization, 2020. Available: https://www.who.int/publications/i/item/9789240017542

[2] United Nations, "Sustainable Development Goals (SDGs)," United Nations Brazil, Brasília, Brazil, 2015. [Online]. Available: https://brasil.un.org/pt-br/sdgs

[3] P. M. Buss *et al.*, "Governança em saúde e ambiente para o desenvolvimento sustentável," *Ciência & Saúde Coletiva*, vol. 17, no. 6, pp. 1479–1491, Jun. 2012. Available: https://pesquisa.bvsalud.org/sms/resource/pt/ens-29494

[4] Universidade de Pernambuco, "UH HUOC – Hospital Universitário Oswaldo Cruz," Recife, Brazil: Universidade de Pernambuco, 2023. [Online]. Available: https://upe.br/complexo-hospitalar/huoc/

[5] Ministério da Saúde, "Cadastro Nacional de Estabelecimentos de Saúde (CNES): ficha do estabelecimento de saúde – Hospital Universitário Oswaldo Cruz (HUOC)," DATASUS, Brasília, Brazil, 2025. [Online]. Available: https://cnes.datasus.gov.br/pages/estabelecimentos/ficha/identificacao/2611600000477

[6] A. V. Estrada, A. Kalbusch, and E. Henning, "Indicadores de consumo de água e influência da COVID-19 em hospitais de Joinville-SC," in *Proc. XXIV Simpósio Brasileiro de Recursos Hídricos (XXIV SBRH)*, Belo Horizonte, Brazil, Nov. 2021. [Online]. Available: https://files.abrhidro.org.br/Eventos/Trabalhos/142/XXIV-SBRH0222-2-0-20210803-220110.pdf

[7] F. Anayah, I. A. Al-Khatib, and B. Hejaz, "Assessment of water and sanitation systems at Palestinian healthcare facilities: pre- and post-COVID-19," *Environmental Monitoring and Assessment*, vol. 193, Art. no. 41, 2021, doi: 10.1007/s10661-020-08791-4. Available: https://link.springer.com/content/pdf/10.1007/s10661-020-08791-4

[8] V. Jiménez-Lacarra, E. Martínez-Cámara, J. Santamaría-Peña, E. Jiménez-Macías, and J. Blanco-Fernández, "Healthcare in the time of COVID-19: an environmental perspective on the pandemic's impact on hospitals," *Applied Sciences*, vol. 14, no. 10, Art. no. 4007, 2024, doi: 10.3390/app14104007. Available: https://www.mdpi.com/2076-3417/14/10/4007/xml

[9] L. Santana and D. Sant'Ana, "Water use in healthcare facilities: consumption patterns, indicators and demand forecasting models – a literature review," *Architectural Engineering and Design Management*, published online Feb. 12, 2026, doi: 10.1080/17452007.2026.2626983. Available: https://www.tandfonline.com/doi/full/10.1080/17452007.2026.2626983

[10] L. Santana and D. Sant'Ana, "Análise dos usos finais de água em hospital público de Brasília, Brasil," *Engenharia Sanitária e Ambiental*, vol. 28, Art. no. e20220228, 2023, doi: 10.1590/S1413-41522022028. Available: https://www.scielo.br/j/esa/a/Bvy6dVdSdJy7NV3cHvHDKMd/?format=html&lang=pt

[11] G. Benetti, A. Kalbusch, and E. Henning, "Previsão de consumo de água utilizando modelagem com séries temporais: estudo de caso em um hospital universitário," in *Proc. XVI Simpósio Nacional de Sistemas Prediais (SISPRED)*, Porto Alegre, Brazil, 2025, Art. no. e8137, doi: 10.46421/sispred.v4.8137. Available: https://eventos.antac.org.br/sispred/article/view/8137

[12] Universidade de Pernambuco, Superintendência do Complexo Hospitalar, *CH em Números: Anuário Estatístico do Complexo Hospitalar da Universidade de Pernambuco - Ano 2023*. Recife, Brazil: SCH/UPE, 2023. Available: https://drive.google.com/file/d/1SQnk1TYiqSUWT5b_8JEiFYYFdoovWFL5/view

[13] Universidade de Pernambuco, Superintendência do Complexo Hospitalar, *CH em Números: Anuário Estatístico do Complexo Hospitalar da Universidade de Pernambuco - Ano 2020*. Recife, Brazil: SCH/UPE, 2020. Available: https://drive.google.com/file/d/1ox2PfjIryv9jkiMRQ8sf0IL-AbJmqdpjx/view

[14] Universidade de Pernambuco, Superintendência do Complexo Hospitalar, *CH em Números: Anuário Estatístico do Complexo Hospitalar da Universidade de Pernambuco - Ano 2024*. Recife, Brazil: SCH/UPE, 2024. Available: https://drive.google.com/file/d/1lGJHoBdnDk9oubO2EOsuwxHNmllo_-Wp/view

[15] M. Santos, A. Borges, D. Carneiro, and F. Ferreira, "Time series analysis for anomaly detection of water consumption: A case study," in *Innovations in Industrial Engineering*, J. Machado, F. Soares, J. Trojanowska, and V. Ivanov, Eds., Lecture Notes in Mechanical Engineering. Cham, Switzerland: Springer, 2022, pp. 234–245, doi: 10.1007/978-3-030-78170-5_21. Available: https://link.springer.com/chapter/10.1007/978-3-030-78170-5_21

Avaliação de Pontos de Contaminação em Cabines de Segurança Biológica Classe II

1st Thales Correa Taranto
Programa de Engenharia Biomédica
UFRJ/COPPE
Rio de Janeiro – RJ, Brasil
thales.taranto@peb.ufrj.br

2nd Carlos Roberto Tavares de Miranda
COGIC
Fundação Oswaldo Cruz
Rio de Janeiro – RJ, Brasil
carlos.miranda@fiocruz.br

3rd Luiz Rogério Guimarães Taranto
COGIC
Fundação Oswaldo Cruz
Rio de Janeiro – RJ, Brasil
luiz.taranto@fiocruz.br

4th Diogo Antonio Tschoeke
Programa de Engenharia Biomédica
UFRJ/COPPE
Rio de Janeiro – RJ, Brasil
diogoat@peb.ufrj.br

5th Roberto Macoto Ichinose
Programa de Engenharia Biomédica
UFRJ/COPPE
Rio de Janeiro – RJ, Brasil
macoto@peb.ufrj.br

***Resumo*—As cabines de segurança biológica (CSB) de Classe II são fundamentais para conter agentes biológicos e proteger os operadores, o ambiente e as amostras. O objetivo deste estudo foi avaliar, por meio de cultivo biológico, se existem pontos com maior potencial de contaminação na estrutura das CSBs, não apenas em áreas que são acessadas pelo usuário mas também por profissionais que realizam a manutenção destes equipamentos. A metodologia consistiu no mapeamento de 25 pontos, definidos com base na literatura, e na coleta microbiológica com placas de contato RODAC contendo Ágar Triptona de Soja, incubadas para fungos e bactérias. Foi utilizado como critério de indicação das superfícies contaminadas, um limite de crescimento de 2,5 UFC/cm^2. Dos 25 pontos avaliados, 10 apresentaram crescimento abaixo do limite, enquanto os 15 restantes apresentaram um crescimento desenfreado das colônias, impossibilitando a contagem. Estes resultados reforçam a hipótese de que alguns pontos são potencialmente mais contaminantes, e portanto seu mapeamento pode estabelecer as áreas que requerem maior atenção nos processos de descontaminação, aumentando a segurança ocupacional dos usuários e da equipe de manutenção.**

***Palavras-chave*—cabines de segurança biológica, biosegurança, microbiologia, engenharia clínica**

I. Introdução

Em ambientes laboratoriais nos quais há manipulação de agentes infecciosos capazes de representar risco ocupacional e ambiental, o conceito de biossegurança está relacionado à implementação de medidas de contenção destinadas a reduzir ou eliminar a exposição a esses agentes potencialmente perigosos. Nesse contexto, destacam-se três elementos fundamentais: a adoção de boas práticas laboratoriais, a seleção adequada de equipamentos de proteção e contenção, e o correto projeto das instalações, conforme o nível de risco envolvido. [1]

Entre os equipamentos de contenção, as cabines de segurança biológica (CSB) constituem o principal dispositivo destinado à proteção contra aerossóis e respingos gerados durante procedimentos laboratoriais. Essas cabines são classificadas em Classe I, II e III, sendo sua escolha determinada pelo tipo de procedimento realizado e pela classe de risco dos agentes biológicos manipulados. As CSB de Classe II possuem fluxo de ar controlado e filtros de alta eficiência (HEPA), responsáveis por reter partículas potencialmente contaminantes, protegendo simultaneamente o operador, a amostra e o ambiente [2].

No contexto da avaliação dos riscos associados ao uso deste equipamento, a literatura descreve pontos específicos para a verificação da eficácia da descontaminação ambiental. [3]–[11] Nesse sentido, o presente estudo tem como objetivo avaliar, por meio de cultivo biológico, se existem pontos com maior potencial de contaminação em áreas que são acessadas pelo usuário e por profissionais que realizam a manutenção destes equipamentos.

II. Materiais e Métodos

O presente trabalho apresenta a seleção do equipamento a ser avaliado, o levantamento e a avaliação dos pontos de interesse descritos na literatura. Adicionalmente, foi definido o protocolo de aplicação do meio para cultivo. Por fim, foi estabelecida uma metodologia para análise e quantificação dos dados obtidos, permitindo a avaliação objetiva dos resultados obtidos após incubação do material.

A. Delineamento do estudo e equipamentos utilizados

Para a identificação dos potenciais pontos de contaminação, foi realizada uma busca na literatura por estudos que abordassem procedimentos de descontaminação ambiental utilizando agentes químicos vaporizados, tais como formaldeído e peróxido de hidrogênio. Foram selecionados trabalhos que avaliassem a eficácia desses procedimentos por meio da utilização de indicadores biológicos posicionados em regiões previamente definidas dos equipamentos. Essas foram as regiões consideradas de interesse para o estudo.

Foi realizado o levantamento do parque de equipamentos de uma instituição pública com laboratórios em diversos estados do Brasil, com o objetivo de quantificar e caracterizar as

Cabines de Segurança Biológica (CSB) existentes. A coleta de dados abrangeu a identificação do setor, bem como a classe e o modelo de cada equipamento, utilizando-se uma planilha eletrônica para o armazenamento, análise e seleção dos atributos. Adotou-se como critério de inclusão exclusivamente os equipamentos de Classe II, visto que eles são maioria na instituição, e a literatura científica consultada fundamenta-se majoritariamente neste tipo [3]–[11]. Essa delimitação justifica-se, também, pela baixa prevalência de equipamentos da Classe I e III nos laboratórios locais. Da mesma forma, os sistemas de Classe III, além de incomuns, possuem estrutura totalmente enclausurada, característica que inviabilizaria a condução do experimento proposto. Além disso, utilizou-se como critério de exclusão os equipamentos que estivessem instalados fora do município do Rio de Janeiro.

A estrutura interna da CSB pode ser dividida em duas áreas: as áreas limpas e as áreas contaminadas (ou sujas). [3]–[11] As áreas limpas compreendem as regiões de circulação de ar, previamente submetido à filtragem por meio dos filtros HEPA, englobando os filtros de insuflação e exaustão. Por sua vez, as áreas sujas delimitam as regiões internas de trânsito do ar provido de carga biológica ou particulados.

B. Procedimento para detecção de microrganismos em superfície

Foi utilizado um meio específico para detecção e enumeração de microrganismos presentes em superfícies. Trata-se de uma placa de petri RODAC (do inglês *Replicate Organism Detection and Counting*), apelidada de ”carimbo”, de 65x6mm com inativante irradiada, para garantir esterilidade, contendo Ágar Triptona de Soja [12], como alternativa aos métodos tradicionais baseados na detecção de ATP por fluorescência.

A demarcação dos pontos na estrutura do equipamento foi realizada com o auxílio de marcador permanente. Com o uso de luvas estéreis, cada placa RODAC foi pressionada contra a sua respectiva área delimitada pelo período de 5 segundos, garantindo o alinhamento concêntrico entre a numeração registrada na superfície do equipamento e o disco de coleta para cultura. Para os pontos localizados nos filtros plissados, a identificação foi efetuada contornando, com caneta, a placa imediatamente após o posicionamento.

A Figura 1 retrata as placas numeradas pelas posições de coleta no equipamento. Após as medições, aplicação do ”carimbo” nas áreas identificadas, o material foi incubado em estufa à 22,5ºC por 5 dias para cultura de fungos, e 32,5ºC por 2 dias para cultura de bactérias. O processo envolve a contagem das colônias formadas após a incubação, e o resultado é expresso em unidades formadoras de colônias (UFC).

C. Análise dos dados obtidos

Considerando que para cada observação quantificou-se o número de UFC, a análise consistiu em verificar se alguma região ultrapassou limites recomendados de presença de microrganismos em superfícies. Embora não exista um limite universalmente aceito foi definido, baseado na literatura, que superfícies que registraram crescimento menor que 2,5 UFC/cm^2 estão em níveis aceitáveis, para isso o resultado obtido em UFC foi dividido por 25 que representa a área, em cm^2, da superfície de contato da placa RODAC [12], [13].

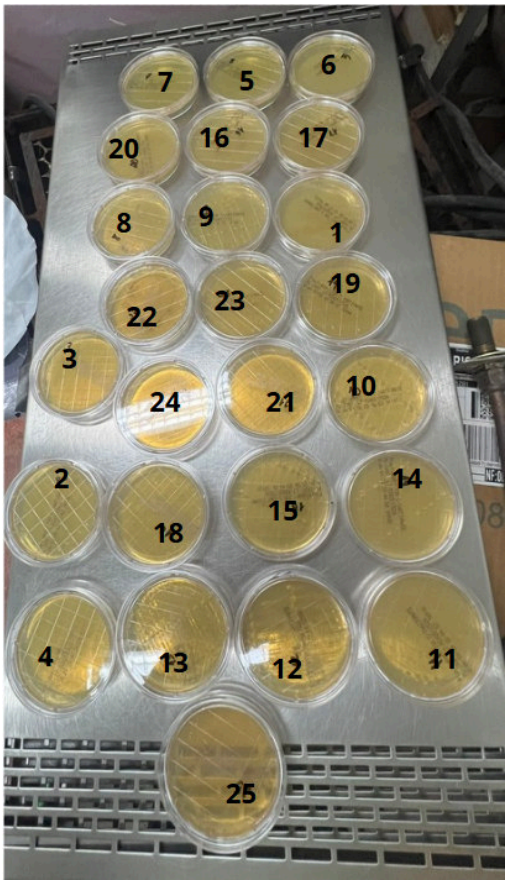


Figura 1. Placas RODAC identificadas pelas posições das superfícies coletadas.

III. Resultados e Discussão

As informações solicitadas referentes às quantidades, localização e características das CSBs foram disponibilizadas, pela instituição, por meio de uma planilha com 734 registros e 12 atributos categóricos. Dos 734 registros do banco de dados analisados, um total de 709 se tratavam de CSB, sendo 3 classificadas como Classe I, 705 como Classe II, sendo estas as de interesse do trabalho, e 1 como Classe III. Destas 709 CSBs, 490 unidades estavam localizadas no Rio de Janeiro. A distribuição quantitativa das 5 marcas com mais equipamentos encontram-se apresentadas na Tabela I.

Tendo em vista que os equipamentos da marca 1 representam mais de 50% do parque de Cabines de Segurança Biológica (CSBs) analisado, a avaliação dos pontos com maior potencial de contaminação concentrou-se nos modelos desta marca. A análise foi realizada em uma CSB de um laboratório que trabalhava com manipulação de Vírus da Imunodeficiência

Tabela I
CINCO MARCAS COM MAIS EQUIPAMENTOS NA INSTITUIÇÃO

Marca	Quantidade
Marca 1	267
Marca 2	39
Marca 3	26
Marca 4	24
Marca 5	22

Humana (HIV), que são virus com baixa sobrevida fora de seu habitat natural.

Foram avaliados 25 pontos do equipamento, os quais encontram-se mapeados e ilustrados na Figura 2. A descrição detalhada e a respectiva identificação de cada um desses pontos estão sistematizadas na Tabela II.

Tabela II
LOCALIZAÇÃO DOS PONTOS DE ANÁLISE

Índice	Descrição
1	Parede lateral do lado direito do exaustor
2	Meio da mesa de trabalho na parte frontal
3	Lado direito da mesa de trabalho na parte frontal
4	Lado esquerdo da mesa de trabalho na parte frontal
5	Meio da bandeja coletora de resíduos na parte frontal
6	Lado direito da bandeja coletora de resíduos na parte frontal
7	Lado esquerdo da bandeja coletora de resíduos na parte frontal
8	Superfície direita do lado limpo do filtro de insuflação
9	Superfície esquerda do lado limpo do filtro de insuflação
10	Superfície direita do lado sujo do filtro de insuflação
11	Superfície esquerda do lado sujo do filtro de insuflação
12	Meio da superfície do lado limpo do filtro de exaustão
13	Fundos canto direito da superfície do lado limpo do filtro de exaustão
14	Meio da superfície do lado sujo do filtro de insuflação
15	Meio da superfície do lado sujo do filtro de exaustão
16	Parede fundo da área de trabalho, no nível mais alto
17	Parede do fundo da área de trabalho na metade da altura do equipamento
18	Pleno positivo, no meio da chapa de cobertura, ao lado do difusor de ar
19	Parede direita do pleno positivo
20	Parede esquerda do pleno positivo
21	Meio da parede do fundo do pleno positivo
22	Pleno negativo inferior, parede do fundo da bandeja coletora de resíduos
23	Parede lateral do lado esquerdo do exaustor
24	Carcaça do exaustor
25	Meio da mesa de trabalho na parte do fundo

Para facilitar a visualização dos resultados, os pontos indicados na Figura 2 foram coloridos conforme a quantidade de UFC/cm^2 encontrada. As regiões com valores abaixo de $2,5$ UFC/cm^2 foram destacadas como hexágonos na cor verde. Nos casos em que não foi possível contabilizar as colônias formadas, utilizaram-se círculos na cor cinza. Além disso, os resultados visuais das colônias formadas em cada placa foram agrupados e divididos de acordo com a localização no equipamento: áreas limpas (Figura 3) ou áreas sujas (Figura 4).

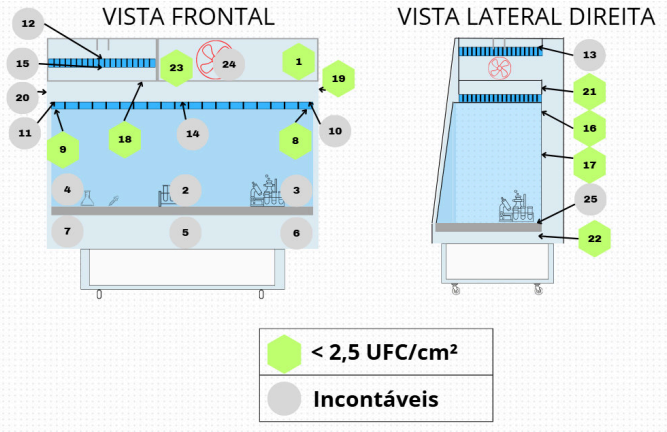


Figura 2. Mapeamento dos pontos de coleta representados nas posições do equipamento em vista frontal e vista lateral, os pontos com crescimento abaixo de 2,5 UFC/cm^2 estão pintados em verde, enquanto os pontos que representam incontáveis colônias formadas estão em cinza.

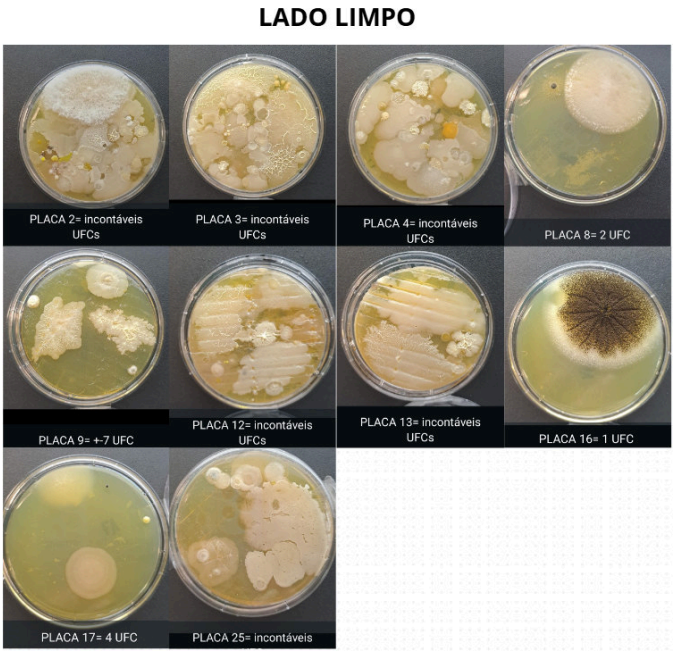


Figura 3. Resultado visual do crescimento de colônias após incubação agrupados por pontos que ficam localizados lado limpo. (Placas: 2, 3, 4, 8, 2, 9, 12, 13 ,15, 17 e 25 respectivamente da esquerda para direita de cima para baixo).

O padrão de crescimento dos microrganismos pode ser associado à forma como o ar circula na Cabine de Segurança Biológica (CSB). O ar do ambiente é puxado principalmente pelas frestas da mesa de trabalho em direção à bandeja de resíduos. Por causa da pressão negativa, ele corre pelo fundo do equipamento até o motor na parte superior, que sopra o ar para o pleno positivo, onde ficam os filtros HEPA. Como no pleno negativo o ar está sendo aspirado, é provável que não

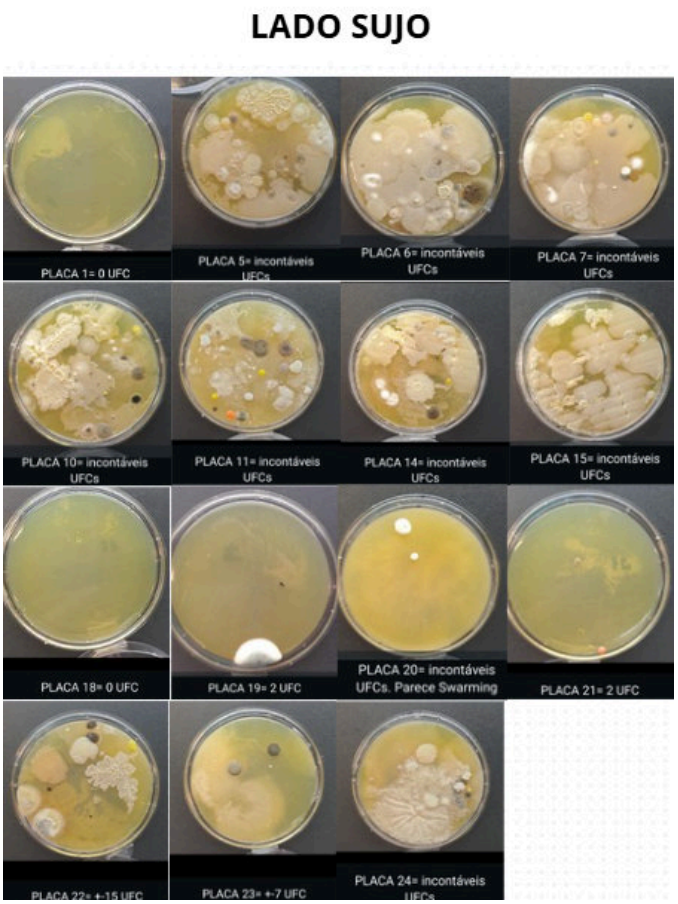


Figura 4. Resultado visual do crescimento de colônias após incubação agrupados por pontos que ficam localizados lado sujo (Placas: 1, 5, 6, 7, 10, 11, 14, 15, 18, 19, 20, 21, 22, 23 e 24 respectivamente da esquerda para direita de cima para baixo).

haja muita fixação nas suas superfícies, o que poderia explicar a ausência ou baixa quantidade de colônias nas regiões 1 e 23 (paredes laterais do pleno negativo superior).

Como era de se esperar, os pontos avaliados após o filtro de insuflação não apresentaram crescimento de UFC acima do limite. Por outro lado, no "lado limpo" do filtro HEPA de exaustão (ponto 12), houve um crescimento grande e diversificado de microrganismos, isto por que nesta região havia um acúmulo grande de poeira, o que pode justificar o grande crescimento de fungos.

A análise das imagens e dados mostraram ausência de crescimento nos pontos 1 e 18. Como mencionado, é provável que o resultado obtido no ponto 1 venha da própria característica construtiva das cabines, onde o fluxo de ar no pleno negativo é aspirado, e portanto haveria pouca aderência de aerossóis dos microrganismos nas paredes deste compartimento.

Já o ponto 18 fica no pleno positivo, no meio da chapa de cobertura e ao lado do difusor. Como essa é uma região de pressão positiva, onde há mais chance de turbulência do ar sobre as superfícies, o esperado era um crescimento acima do limite, porém não houve crescimento. É provável que o resultado nesse ponto tenha sido causado por uma aplicação do "carimbo" em uma pressão insuficiente para adesão dos microrganismos presentes na superfície ao meio TSA da placa ou o deslizamento da placa RODAC na hora da coleta. [12]

Também houve divergência entre pontos vizinhos. Nas laterais do pleno positivo, os pontos 19 e 20 tiveram comportamentos opostos: o ponto 19 apresentou formação de colônias abaixo do limite recomendado, enquanto no ponto 20 o crescimento foi tão denso que impediu a contagem (colônias incontáveis). Sendo que, o ponto 19 é o mais próximo da saída do motor de insuflação, e portanto deveria ter resultado semelhante ao 20 que fica mais distante.

De acordo com a norma ABNT NBR 17095: Cabines de segurança biológica classe II (CSB) - Projeto, características e ensaios de desempenho, [14] as superfícies da área limpa devem ser frequentemente descontaminadas com desinfetantes validado para o processo, portanto, a contaminação residual desses pontos deveria ser baixa. Porém, o resultado obtido foram incontáveis colônias em diversas placas, especialmente em regiões localizadas no lado superior das superfícies ou próximas a vértices, locais mais propensos a acúmulo de material particulado. Uma provável explicação pode ser o acúmulo de poeira, pois o equipamento estava fora de operação, em razão de uma necessidade de descontaminação ambiente para troca dos filtros. Também podem ter contribuído fatores como: distribuição dos materiais utilizados no interior do equipamento, turbulência gerada por movimentações realizadas pelo operador, procedimento inadequado de descontaminação, concentração e natureza dos agentes manipulados, posição do equipamento no layout do laboratório e frequência de uso do equipamento.

Após o isolamento microbiano nas placas RODAC, é possível identificar os microrganismos por diversas técnicas, que incluem a análise da morfologia da colonia formada, a técnica de coloração de Gram para bactérias e coloração por azul de lactofenol para fungos, a realização de provas bioquímicas e sequenciamento 16S para bactérias, e ITS para fungos. No trabalho só foi possível fazer uma análise morfológica visual, embora essa abordagem permita distinguir fungos de bactérias, ela não é suficiente para determinar a taxonomia exata das colônias. Apesar dessa limitação, a inspeção visual indicou que a maior parte do crescimento foi de fungos, identificados pelo aspecto cotonoso das colônias, que chegaram a dominar algumas placas. Além disso, observou-se a presença de bactérias pigmentadas que fica evidente na placa 11 da Figura 4.

O método utilizado traz algumas limitações que precisam ser consideradas. As placas RODAC são sensíveis e indicadas para avaliar superfícies limpas (pós-higienização) [15]. Quando usadas em superfícies sabidamente contaminadas, o crescimento de bactérias pode ser tão intenso que forma uma massa única (um "tapete"), dificultando a contagem individual das colônias.

Outro fator é que foi feita apenas uma coleta por ponto, o que impede uma análise estatística dos dados. Nos pontos onde não houve crescimento, fica a dúvida se a superfície estava realmente limpa ou se ocorreu algum erro na hora de

aplicar o ”carimbo” na superfície.

O cronograma também influenciou: para marcar os pontos, as tampas traseiras do equipamento precisaram ser abertas. Como houve um atraso de cerca de 5 dias até a chegada das placas RODAC, a cabine ficou destampada e exposta ao ambiente durante esse período. Por fim, a identificação se limitou à análise visual do cultivo; para maior precisão, seria importante realizar testes de coloração, provas bioquímicas ou sequenciamento genético (16S/ITS).

O procedimento com as placas RODAC será repetido logo após a descontaminação da cabine para se verificar a eficácia do processo e observar o desenvolvimento ou não de colônias. O protocolo será revisado a partir das limitações identificadas, prevendo, por exemplo, coletas em triplicata por ponto avaliado, para garantir uma melhor avaliação no processo aplicado a cada ponto, e a redução do tempo de abertura da cabine durante a marcação, a fim de minimizar a interferência do ambiente. Também está prevista a realização de sequenciamento molecular (16S/ITS) para identificar com precisão as espécies presentes nos pontos críticos, além da avaliação de CSBs que manipulam microrganismos patogênicos.

IV. Conclusão

Conclui-se que, embora as limitações metodológicas tenham gerado resultados parciais e, em alguns pontos, inconclusivos, este estudo foi capaz de descartar pontos com baixo potencial de contaminação, e alguns pontos que necessitam atenção, onde houve um crescimento desenfreado de colônias. Estes resultados indicam que há uma distinção de risco de contaminação entre algumas regiões do equipamento, oferecendo subsídios para adoção e revisão de práticas que possam aumentar a segurança ocupacional dos usuários e da equipe de manutenção.

Agradecimentos

Os autores agradecem as agências de fomento de pesquisa CNPq, CAPES e FAPERJ. Agradecemos também a Dra. Ana Paula Pereira Alcides do Setor de Produtos Estéreis do INCQS-FIOCRUZ, que nos auxiliou no processo de coleta do material e incubação das placas RODAC.

Referências

[1] S. Y. M. UEKI *et al.*, “Cabine de segurança biológica: efeito da luz ultravioleta nas micobactérias,” *Revista do Instituto Adolfo Lutz*, 2006.

[2] G. A. JAGTAP *et al.*, “The role of the biosafety cabinet in preventing infection in the clinical laboratory.” *Cureus*, 2023.

[3] K. E. JONES *et al.*, “T1 bacteriophage as an indicator for decontamination of laminar-flow biological safety cabinets,” *Applied and environmental microbiology*, 1981.

[4] K. MUNRO *et al.*, “A comparative study of methods to validate formaldehyde decontamination of biological safety cabinets,” *Applied and environmental microbiology*, 1999.

[5] L. HALL *et al.*, “Use of hydrogen peroxide vapor for deactivation of mycobacterium tuberculosis in a biological safety cabinet and a room,” *Journal of clinical microbiology*, 2007.

[6] ——, “Deactivation of the dimorphic fungi histoplasma capsulatum, blastomyces dermatitidis and coccidioides immitis using hydrogen peroxide vapor,” *Medical mycology*, 2008.

[7] D. NGABO *et al.*, “Cabinet decontamination using formaldehyde,” *Applied Biosafety*, 2017.

[8] G. FREY *et al.*, “Decontamination validation of a class ii type a2 biosafety cabinet during laboratory fumigation,” *Applied Biosafety*, 2020.

[9] J. STUART *et al.*, “Comparing the efficacy of formaldehyde with hydrogen peroxide fumigation on infectious bronchitis virus,” *Applied Biosafety*, 2020.

[10] T. POTTAGE *et al.*, “Hazard group 3 agent decontamination using hydrogen peroxide vapour in a class iii microbiological safety cabinet,” *JOURNAL OF APPLIED MICROBIOLOGYGREAT CLARENDON ST*, 2020.

[11] J. P. e. a. TEIFKE, “Dry-fog decontamination of microbiological safety cabinets after activities with sars-cov-2: cycle development and process validation for dry fogging with peroxyacetic acid,” *GMS hygiene and infection control*, 2021.

[12] *Placa Rodac (p/Bact) S/TTC Irr. Emb. Tripla*, Plastlabor, acessado em: 01 jul. 2026. [Online]. Available: https://plastlabor.com.br/produtos/placa-rodac-p-bacts-ttc-irr-emb-tripla/

[13] O. P. FROTA *et al.*, “Eficiência da limpeza e desinfecção de superfícies clínicas: métodos de avaliação,” *Revista Brasileira de Enfermagem*, 2020.

[14] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, *ABNT NBR 17095: Cabines de segurança biológica classe II (CSB) — Projeto, características e ensaios de desempenho*, Associação Brasileira de Normas Técnicas, Rio de Janeiro, 2025.

[15] B. LEAS *et al.*, “Environmental cleaning for the prevention of healthcare-associated infections: Ahrq publication no. 15-ehc020-ef,” Agency for Healthcare Research and Quality, Tech. Rep., 2015.

Challenges in Adapting Dental Care Services to ANVISA Regulation RDC No. 1,002/2025: The Clinical Engineering Experience at a Brazilian Oncology Hospital

1st Maria Rafaella Cardoso Delgado Santos
Clinical Engineering
TECSAÚDE
Recife, Brazil
maaria.rafaella@gmail.com

2nd Anne Fernandes Stegmann D'Antona
Clinical Engineering
TECSAÚDE
Recife, Brazil
dantona.anne@gmail.com

***Abstract*—Brazilian National Health Surveillance Agency (ANVISA) RDC No. 1,002/2025 establishes new Good Practice requirements for dental care services, mandating formal health technology management, medical device reprocessing, and risk management. Applying these requirements reveals technical and metrological challenges that remain scarcely described in the literature. This paper presents a practical pre-operational diagnostic and regulatory-adaptation guide developed by the Clinical Engineering team responsible for the Dentistry service at the Pernambuco Cancer Hospital (Hospital de Câncer de Pernambuco, HCP). This is a descriptive case study based on the systematic mapping of the requirements set out in the regulation as they relate to Clinical Engineering activities, encompassing a technology asset inventory, a gap-analysis matrix, and the need to redesign workflows, highlighting the regulatory decision to centralize sterilization at the hospital's Central Sterile Supply Department (CSSD) in light of local physical constraints. The results show that Clinical Engineering acts as an indispensable technical link in pre-operational planning, overcoming market bottlenecks such as the scarcity of metrological parameters from manufacturers and of accredited calibration laboratories. As its main contribution, this work provides a replicable implementation model to support transition and audit processes in other dental care services.**

***Index Terms*—clinical engineering; RDC 1,002/2025; dental services; health technology management; equipment lifecycle**

I. Introduction

The publication of ANVISA Resolution RDC No. 1,002/2025 represents a regulatory milestone for dental services by consolidating requirements for good operating practices, biosafety, reprocessing of health technology and management of health technology [1].

Although many of these demands are already familiar to hospital facilities through technology management standards [2], applying them to the dental sector presents specific challenges, particularly for smaller facilities or those historically not integrated into formal clinical engineering structures.

This article presents a descriptive case study and a Framework for pre-operational diagnostic and regulatory compliance for the dental service of the Hospital de Câncer de Pernambuco (HCP) regarding ANVISA RDC No. 1,002/2025.

This study highlights the practical challenges encountered by the authors, who work at TECSAÚDE and provide Clinical Engineering services, including management and consulting of technological assets, to the hospital in question. The purpose of this paper is to present a compliance methodology to serve as a practical model/guide for dental services that must adapt to the new RDC regulations.

II. Regulatory and scientific background

ANVISA Resolution RDC No. 1,002/2025 establishes operational requirements for dental services and specifies obligations related to health technology management, medical device reprocessing, and patient safety [1]. In practice, the standard shifts dental compliance from an approach focused solely on physical infrastructure and biosafety to an integrated approach that includes equipment, documentation, traceability, and technical responsibility.

The requirements with the greatest impact on Clinical Engineering activities include: the development and implementation of a health technology management plan [2]; calibration and qualification of reprocessing equipment; equipment inventory; planned maintenance schedule for all equipment; evidence of preventive and corrective maintenance execution; mandatory equipment identification, among others.

In this scenario, clinical engineering plays a technical translation role: it transforms regulatory requirements into plans, schedules, acceptance criteria, execution records, and auditable evidence [3]. This role becomes even more relevant when manufacturers do not readily provide qualification parameters or when testing and calibration laboratories lack an accredited scope compatible with dental equipment.

III. Methodology

This work is a descriptive, applied, pre-operational case study conducted in the dental service of the Hospital de

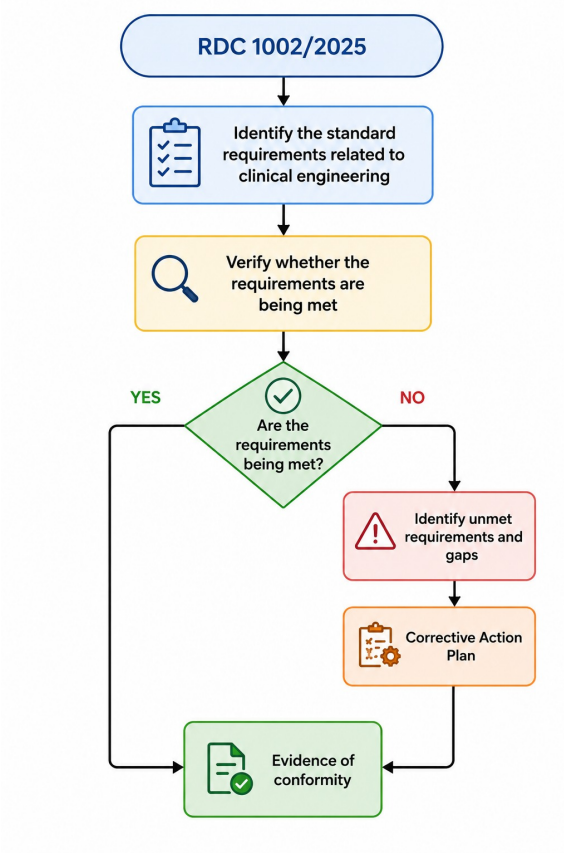


Fig. 1. Methodological design of the study

Câncer de Pernambuco (HCP) during its architectural and regulatory adaptation phase to ANVISA Resolution RDC No. 1,002/2025. This approach is justified by the opportunity to map and resolve—within a real hospital environment—the existing physical, metrological, and documentary constraints involved in the implementation of the new regulation.

The evaluation was conducted by the TECSAÚDE Clinical Engineering team working at HCP between March and May 2026, which included a documentary audit, an architectural project review and an inventory of existing technological assets.

The study was structured into four linked stages (Fig. 1):

- Normative Mapping and Structuring: systematic analysis of RDC No. 1,002/2025 and identification of requirements relevant to Clinical Engineering activities.
- Field Diagnosis and Audit: verification of compliance with the identified requirements.
- Gap Analysis Matrix [4]: identification of non-conformities and market constraints (such as the scarcity of manufacturer parameters and accredited calibration laboratories).
- Action Plan and Evidence of Conformity: formulation of corrective actions and workflow readjustments, where necessary, alongside the presentation of auditable evidence of conformity.

IV. HCP Case Study

A. Characterization Of The HCP Dental Service

Hospital de Câncer de Pernambuco (HCP) is a private, non-profit institution and a state reference center for the diagnosis and treatment of oncology patients within the Brazilian Unified Health System (SUS) [5]. The dental service constitutes an essential and mandatory stage in the care continuum, particularly for patients undergoing head and neck radiotherapy or chemotherapy.

The dental department is currently classified as a collective dental clinic, comprising: (i) three complete dental units; (ii) a dedicated room for medical device cleaning and reprocessing; and (iii) a reception area. Concurrently, the service is undergoing physical expansion to scale up patient care. The new facility will house a panoramic X-ray machine, an individualized office for patients requiring isolation, and the infrastructure for the future installation of a wall-mounted intraoral X-ray device.

Given the ongoing construction, an opportune moment to prevent regulatory non-conformities, the Clinical Engineering team performed a pre-operational diagnosis by auditing the architectural project, operational workflows, and existing assets. Mapping mandatory regulatory requirements and verifying compliance with ANVISA Resolution RDC No. 1,002/2025 was fundamental for adapting workflows and infrastructure, ensuring that once the department transitions to the new facility, it will operate in full compliance with health and legal standards.

B. RDC Compliance Experience

Given this scenario and following a detailed analysis of ANVISA Resolution RDC No. 1,002/2025, the Clinical Engineering team structured a regulatory diagnostic and gap analysis matrix. This tool correlates mandatory regulatory requirements with the findings from the pre-operational audit. In addition, it defines corrective actions, responsible personnel, legal grounds, and evidence verification methods [4], as presented in Table I.

V. Results and Discussion

Applying the regulatory diagnostic matrix enabled the Clinical Engineering team to map the operational status of the HCP dental care service during its pre-operational phase and guide targeted intervention actions. The primary findings and outcomes are detailed below.

A. Technical Inventory, HTMP, and Maintenance Schedule

The initial assessment confirmed that the Dentistry service was already covered by HCP's Health Technology Management Plan (HTMP), so no corrective action was required for this item.

After the sector inventory was formed, it was found that some existing equipment was not registered in the Clinical Engineering management software, nor was it included in the planned maintenance schedule. The updated inventory is shown in Table II.

TABLE I
GAP-ANALYSIS MATRIX AND ACTION PLAN FOR ADAPTATION TO ANVISA RDC NO. 1,002/2025

Axis/Requirement Assessed	Regulatory Basis	Current Situation	Corrective Action	Responsible Party	Evidence Method
Health Technology Management Plan (HTMP)	RDC 509/2021 [2] / Art. 131 RDC 1,002/2025 [1]	Compliant. The dental sector was already integrated into the institutional HTMP.	Maintain sector routines inclusion and periodically validate workflows.	Clinical Engineering	Current, approved institutional HTMP document.
Technology inventory	Art. 130 RDC 1,002/2025 [1]	Full integration of the dental technology park into the software was lacking.	Carry out a sector inventory and register the equipment not yet recorded in the management software.	Clinical Engineering	Inventory report in the management software and completed physical spreadsheet.
Planned Maintenance Schedule	Art. 130/134 RDC 1,002/2025 [1]	Equipment identified outside the planned maintenance routine.	Include the missing equipment and parameterize the scheduled preventive maintenance program.	Clinical Engineering	Issued service orders (SOs) and annual schedule.
Water Quality Control	Chapter III/IV RDC 1,002/2025 [1]	No sampling points mapped for microbiological control.	Add water sampling points in the sector for laboratory sampling and analysis.	Occupational Health and Safety Service (collection), Infection Control Committee, and Clinical Engineering (result evaluation)	Periodic physicochemical and microbiological analysis reports.
Asset Dimensioning	Art. 37 RDC 1,002/2025 [1]	Need for diagnostic expansion and isolation capability.	Map gaps and request acquisition of the necessary medical equipment.	Clinical Engineering, Dentistry, Management	Acquisition process and Technical Incorporation Report.
Qualification and metrology	Art. 130/134 RDC 1,002/2025 [1] / ISO 17665:2024 [6]	Lack of parameters in manuals and from local suppliers.	Contract external thermal qualification and calibration of peripherals.	Clinical Engineering	Metrological certificates and qualification reports.
Radiation Protection	RDC 1,002/2025 [1] / RDC 611/2022 [7] / RDC 50/2002 [8]	New panoramic and intraoral X-ray rooms under construction.	Design shielding and request a post-construction radiometric survey.	Maintenance (design), Clinical Engineering (imaging tests)	Radiometric survey report and imaging tests.
Continuing education	Art. 113 RDC 1,002/2025 [1]	No formal records of operational training.	Develop a continuing education schedule for care teams.	Continuing Education, Clinical Engineering	Training attendance records.

TABLE II
TECHNOLOGY INVENTORY OF THE HCP DENTISTRY SECTOR

Equipment / System	Qty.	Manufacturer/Model
Dental chair	04	Dentemed/Prime
Dental autoclave	02	Vitale 12/Cristófoli; Digitale/BS Equipamentos
Dental compressor	02	Motomil and FIAC
Micromotor	04	Marathon 3/Talmax
Dental laser	02	Laser duo/MM Optics
Vacuum pump	01	Suctron Eletronic/Schuster
Panoramic X-ray unit	01	Eagle 2D/Dabi Atlante
Ultrasonic bath	01	Cristófoli
Electrosurgical unit	01	HF120/WEM
Sealer	01	Cristófoli
Dental ultrasound scaler	01	Altsonic jet ceramic/Alt Equipamentos
Dental handpiece	02	Torque extra 605/Kavo
Contra-angle handpiece	01	Kavo
Negatoscope	01	Generic
Curing light	02	Bluestar/Microdent
Suction kit	04	Schuster

When integrating this equipment into the maintenance schedule, the team faced significant challenges due to the scarcity of technical data from the manufacturers and the broader market. The key parameters for calibration, thermal qualification, and performance qualification—particularly for sterilization equipment (autoclaves)—were omitted from the instruction manuals.

Although the documentation provided basic guidance on preventive and corrective maintenance, it omitted the metrological and validation requirements necessary for proper compliance with the sole paragraph of Art. 130 of RDC No. 1,002/2025 [1].

B. Metrological Gaps in the Local Market

In addition, a gap was identified in the local market for metrology service providers. The calibration providers consulted (such as LEMPE, Laboratório Medição, and Medserv) did not perform temperature calibration or performance qualification for dental equipment, which made it difficult to obtain the metrological evidence required by the regulation.

To gather more information and meet regulatory requirements, Clinical Engineering had to contact manufacturers and authorized factory representatives directly to build the updated planned maintenance schedule for the sector (Table III).

TABLE III
PLANNED MAINTENANCE SCHEDULE FOR THE HCP DENTISTRY SECTOR

Equipment	Maintenance Type	Frequency
Dental chair	Internal preventive; electrical safety	Annual
Dental autoclave; ultrasonic bath; sealer (decommissioned)	External calibration; external preventive	Annual or after corrective maintenance
Dental compressor	External preventive	Annual
Dental handpiece; contra-angle hand piece; micro motor; ultrasound scaler; curing light; laser	Internal preventive	Annual
Vacuum pump	External preventive	Annual
Panoramic X-ray unit	Image quality test; external preventive	Annual
Electro surgical unit	Internal preventive; internal calibration; electrical safety	Annual or after corrective maintenance

C. Adaptation of the Architectural Project and Centralization of Reprocessing

Before executing the action plan, Clinical Engineering met with a multidisciplinary hospital team (including Nursing, Architecture, Projects, Executive Board, Dentistry, and Civil Engineering) to evaluate and review the architectural design of the new area under construction. The sector's floor plan (Fig. 2) required adjustments to optimize equipment placement and functional area sizing, ensuring full compliance with regulatory requirements.

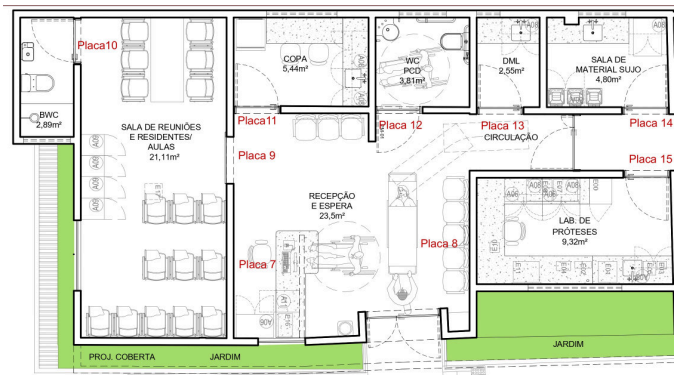


Fig. 2. Conceptual floor plan of the dental sector after architectural adaptation (simplified view)

Since the sector is configured as a collective dental practice, the area intended for medical device reprocessing had to comply with items II or III of Art. 25 of the regulation [1]. After reviewing the project, it was found that although a dedicated room for contaminated material (soiled utility room) was provided with a total area of 4.8 m² — which would strictly meet the lower limits of Item II, Art. 25 of the regulation [1] — the physical spatial arrangement presented severe operational constraints.

Specifically, the dimensions of the workbenches, the space required for existing equipment, and the mandatory regulatory distances between devices made it impossible to establish a compliant, unidirectional workflow from the soiled area to the clean/sterilized area.

Given these physical constraints, the sector's and HCP's management chose to centralize the disinfection and sterilization steps at the institution's Central Sterile Supply Department (CSSD), retaining only the pre-cleaning step within the dental sector, as a technical and regulatory safeguard measure.

Consequently, the sterilization equipment (autoclaves, ultrasonic bath, and sealer) mapped during the initial diagnostic phase were decommissioned from the sector's active maintenance schedule (Table III) and will be reallocated according to institutional workflows.

VI. FINAL CONSIDERATIONS

The adaptation of the Pernambuco Cancer Hospital's dental care service to ANVISA RDC No. 1,002/2025 demonstrates that compliance with the new regulatory requirements demands management-related technical competencies that go beyond the scope of traditional dental practice.

In this pre-operational scenario of sector expansion, Clinical Engineering established itself as a strategic agent for interpreting mandatory requirements, aligning infrastructure projects, and enabling engineering decisions — such as the decision to centralize material reprocessing at the hospital's CSSD to overcome the unidirectional-flow constraints required by the regulation.

The challenges identified throughout this case study—such as the scarcity of data in manufacturers' manuals and the lack of local laboratories with accreditation scope for the dental sector—indicate that while the new regulation promotes good practices and safe technology management, it also exposes market bottlenecks regarding technical reference materials and specialized service providers.

As a practical contribution, this work provides a methodological guide and a diagnostic matrix that can be replicated to assist other Clinical Engineering professionals, or even dental care services themselves, in adapting to the regulatory requirements.

Finally, since this intervention was carried out during the sector's physical restructuring phase, future work is recommended to monitor performance and operational-effectiveness indicators after the new HCP dentistry sector opens.

ACKNOWLEDGMENT

The authors would like to thank the Hospital de Câncer de Pernambuco (HCP) and the TECSAÚDE Clinical Engineering team for their institutional support and technical collaboration throughout the regulatory adaptation process.

REFERENCES

[1] Brasil, Agência Nacional de Vigilância Sanitária, “Resolução da Diretoria Colegiada – RDC nº 1.002, de 15 de dezembro de 2025: dispõe sobre os requisitos de Boas Práticas de Funcionamento para os serviços que prestam assistência odontológica,” (Good operating practice requirements for dental care services), Diário Oficial da União, Brasília, DF, Brazil, no. 239, sec. 1, p. 191, Dec. 16, 2025.

[2] Brasil, Agência Nacional de Vigilância Sanitária, “Resolução da Diretoria Colegiada – RDC nº 509, de 27 de maio de 2021: dispõe sobre o gerenciamento de tecnologias em saúde em estabelecimentos de saúde,” (Health technology management in healthcare facilities), Diário Oficial da União, Brasília, DF, Brazil, no. 101, sec. 1, p. 145, May 31, 2021.

[3] J. F. Dyro, Ed., Clinical Engineering Handbook. Burlington, MA, USA: Elsevier Academic Press, 2004.

[4] World Health Organization, Health Technology Assessment of Medical Devices, WHO Medical Device Technical Series. Geneva, Switzerland: World Health Organization, 2011.

[5] Hospital de Câncer de Pernambuco (HCP), "About Us: Institutional profile and health services," HCP Official Portal, 2024. [Online]. Available: https://www.hcp.org.br. [Accessed: Jul. 14, 2026].

[6] ISO 17665:2024, Sterilization of health care products – Moist heat – Requirements for the development, validation and routine control of a sterilization process for medical devices. Geneva, Switzerland: ISO, 2024.

[7] Brasil, Agência Nacional de Vigilância Sanitária, "Resolução da Diretoria Colegiada – RDC nº 611, de 9 de março de 2022: estabelece os requisitos sanitários para a organização e o funcionamento de serviços de radiodiagnóstico médico ou odontológico," (Establishes the health requirements for the organization and operation of medical or dental radiodiagnosis services), Diário Oficial da União, Brasília, DF, Brazil, no. 51, sec. 1, p. 107, Mar. 16, 2022.

[8] Brasil, Agência Nacional de Vigilância Sanitária, "Resolução da Diretoria Colegiada – RDC nº 50, de 21 de fevereiro de 2002: dispõe sobre o Regulamento Técnico para planejamento, programação, elaboração e avaliação de projetos físicos de estabelecimentos assistenciais de saúde," (Technical Regulation for planning, programming, elaboration, and evaluation of physical projects for healthcare facilities), Diário Oficial da União, Brasília, DF, Brazil, no. 53, sec. 1, p. 52, Mar. 20, 2002.

Clinical Engineering in Brazilian Hospitals: Practice and Challenges - Is There Opportunity for a Conceptual Redefinition of the Field?

Mariana Ribeiro Brandão
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
marianaribeirobrandao@gmail.com

Marcella Wayss Darold
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
marcelladarold@gmail.com

Emanuele Fritzen Philippsen
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
emanuelefp5@gmail.com

Cleverson Natan da Silva
Institute of Psychology
Federal University of Alagoas (IP-UFAL)
Alagoas, Brazil
cleverson.silva@ip.ufal.br

Jefferson Brum Marques
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
jlbmarques@gmail.com

Renato Garcia Ojeda
Institute of Biomedical Engineering
Federal University of Santa Catarina (IEB-UFSC)
Florianópolis, Brazil
renato.garcia.ojeda@ufsc.br

***Abstract*— This study aims to perform a national exploratory mapping of the scope, practices, operational and professional perceptions within Brazilian Clinical Engineering. These empirical findings serve as a foundation for a critical conceptual discussion on the expansion and evolution of the profession's scope in the healthcare ecosystem. A descriptive and exploratory study was conducted using a mixed-methods approach (quantitative and qualitative). Data collection was carried out through an online questionnaire from August 2025 to April 2026, yielding a final sample of 84 professionals working in hospital Clinical Engineering departments across the country. The data reveal an expansion of the sector's scope. Although medical-assistive equipment accounted for 100% of managed assets, CE departments have absorbed the management of support equipment (91.38%), surgical instruments (44.83%), medicinal gases (36.21%), infrastructure (22.41%), and medical information technology systems (15.52%). The main challenges identified were equipment misuse by clinical staff (31.71%) and logistical dependency on manufacturers (18.29%). In light of these bottlenecks, the study proposes a conceptual redefinition of CE as an interdisciplinary, evidence-based field aimed at ensuring reliability, traceability, accessibility, and safety.**

Keywords—Clinical Engineering; Health Technology Management; Biomedical Engineering.

I. Introduction

Health technologies consist of devices, medicines, vaccines, procedures, and systems designed to address specific health problems and contribute to improving population quality of life [1]. As a result of growing technological incorporation, a proportionally greater number of professionals are required to carry out Health Technology Management (HTM) and to support safe and reliable use [2]. One of the fields that can contribute to the development, implementation, and monitoring of HTM programs in health services is Clinical Engineering (CE), a branch of Biomedical Engineering [1][3].

The participation of Clinical Engineering in the Health Technology Process (HTP) encompasses the application of knowledge across all lifecycle stages, contributing to effective and safe HTM in healthcare settings. The CE actions incorporated into the HTM process model of the Institute of Biomedical Engineering (IEB-UFSC) are based on three pillars: technology, human resources, and infrastructure. The interaction among these areas aims to improve the quality of health services through Clinical Engineering actions and the detection and analysis of HTP needs within each institution, thereby developing qualified human resources for the use of technological resources and adapting infrastructure more efficiently, with the objective of promoting reliability, effectiveness, and safety across health technology processes [4].

The definition of Clinical Engineering varies depending on the source, and is recurrently conceptualized based on the responsibilities of the clinical engineer. According to American College of Clinical Engineering (ACCE), this professional supports improvement in patient care by applying engineering knowledge and health technology management skills [5]. Bronzino characterizes the professional as responsible for applying technological and scientific expertise in healthcare environments to support clinical activities aimed at improving patient care [6]. The WHO, in one of its technical reports on medical devices, discusses the importance of updating this professional in light of the new competencies required in health systems. Clinical Engineering demands a more holistic approach, one that aims to support technology users by applying not only engineering knowledge but also

economic, communication, and managerial competencies involving health technologies [1].

There are many activities required of Clinical Engineering professionals; some of the main functions include: Health Technology Management, analysis and supervision of safety and performance testing; inventory management; economic and quality analyses; risk management; Health Technology Assessment (HTA); training; development and implementation of protocols, among others [6]. Given the growing presence of technology in healthcare settings, an HTM program aims to ensure availability, accessibility, cost reduction, and safety [1]. In Brazil, the management of these technologies in healthcare establishments is governed by a set of regulations, as presented in Table I.

TABLE I. ANALYZED REGULATORY AND NORMATIVE DOCUMENTS

Category	***Standard***
Health Technology Management	RDC 509/2021 - Provides for the management of health technologies in healthcare establishments.
	RDC 36/2013 - Establishes actions for patient safety in healthcare services and other provisions.
	ABNT NBR 15943:2011 - Guidelines for a management program for healthcare infrastructure equipment and medical equipment.
	ABNT NBR ISO 9001:2015 - Quality management systems.
	ABNT NBR ISO 14971:2020 - Medical devices – Application of risk management to medical devices.
	ABNT NBR ISO 7101:2025 - Healthcare organization management – Quality management systems in healthcare organizations.
	Law nº 14133/2021- Public Procurement and Administrative Contracts Law.
	Law nº 15210/2025 - Establishes requirements for the purchase of equipment intended for diagnostic or therapeutic procedures within the Brazilian Unified Health System (SUS).
	Portaria GM/MS n° 11694/2026 - Provides planning guidelines and recommendations for the acquisition of equipment intended for diagnostic or therapeutic procedures within SUS.
Infrastructure	RDC 50/2002 - Provides the Technical Regulation for planning, programming, design, and evaluation of physical projects for healthcare facilities.
	ABNT NBR 13534:2008 - Requirements for electrical installations in healthcare facilities.
	ABNT NBR 16651:2019 - Fire protection in healthcare facilities.
	ABNT NBR 12188:2016 - Centralized supply systems for medical gases, gases for medical devices, and vacuum systems for healthcare services.
	ABNT NBR 7256:2022 - Air treatment in healthcare facilities – Requirements for design and installation.
Medical Device Safety	RDC 751/2022 - Provides for risk classification, notification and registration requirements, and labeling and instructions for use of medical devices.
	RDC 579/2021 - Provides for the importation, commercialization, and donation of used and refurbished medical devices.
	ABNT NBR IEC 60601-1:2010 - Medical electrical equipment – Part 1: General requirements for basic safety and essential performance.
	ABNT NBR ISO 17025:2017 - General requirements for the competence of testing and calibration laboratories.
	ABNT NBR IEC 62353:2019 - Medical electrical equipment – Recurrent tests and tests after repair of medical electrical equipment.
	RDC 67/2009 – Provides technovigilance requirements applicable to medical device registration holders in Brazil.
	RDC 925/2024 – Provides requirements for the manufacture, commercialization, importation, and placing into use of personalized medical devices.
Specific Services	RDC 63/2011 - Provides requirements for good operating practices in healthcare services.
	RDC 07/2010 - Establishes minimum requirements for the operation of intensive care units and other provisions.
	RDC 15/2012 - Provides requirements for good practices in the processing of healthcare products and other provisions.
	RDC 611/2022 - Establishes sanitary requirements for the organization and operation of diagnostic and interventional radiology services and regulates the control of medical, occupational, and public exposures arising from diagnostic and interventional radiological technologies.
	RDC 11/2014 - Provides requirements for good operating practices in dialysis services and other provisions.
	RDC 222/2018 - Establishes good practices for the management of healthcare waste and other provisions.

Collegiate Board Resolution (RDC) nº509, of May 27, 2021, which governs health technology management in healthcare establishments, establishes a series of management procedures that are planned, developed, and implemented to ensure the traceability, quality, efficacy, effectiveness, safety, and performance of health technologies used in the provision of health services [7].

Collegiate Board Resolution (RDC) nº63, of November 25, 2011, which establishes Good Practice Requirements for Healthcare Services, also includes the need for technology management as one of the good practice requirements for health services [28]. Thus, the implementation of efficient management throughout the entire technology lifecycle constitutes an essential tool for safety and quality, leading to satisfactory outcomes in health systems [1].

The technology lifecycle encompasses all stages involving a technology, from initial innovation, research, and technological diffusion processes through incorporation, use, and obsolescence. HTM is responsible for developing, implementing, and monitoring management programs across these stages, covering activities such as planning, selection and incorporation, inventory, installation, training, technical interventions, and, finally, obsolescence assessment and final disposal. ABNT NBR 15943:2011 establishes guidelines for implementing equipment management programs for health services, in order to minimize risks associated with technologies in healthcare establishments [9].

There is a significant disparity in Clinical Engineering practice across different healthcare settings, both in activities performed throughout the technology lifecycle and in the profiles of professionals and the scope of internal versus outsourced services, among other factors. In this

context, this study aims to analyze the practice of Clinical Engineering in hospital settings across Brazil through the application of usability techniques in a questionnaire-based survey to identify, analyze, and propose improvements for Health Technology Management and patient safety. These empirical findings serve as a foundation for a critical conceptual discussion on the expansion and evolution of the profession's scope in the healthcare ecosystem.

II. Methodology

A quantitative and qualitative, exploratory and descriptive study with a cross-sectional design was conducted. The mixed-methods approach allowed for the descriptive characterization of structural scenarios in Clinical Engineering and the interpretation of professionals' perceptions and challenges through their written responses.

The target population consisted of professionals working directly in Clinical Engineering departments at hospital institutions in Brazil. The inclusion criterion was being a mid-level or higher-level professional (engineers, technologists, technicians, and managers) with an active role in hospital Clinical Engineering services. Responses with incomplete or duplicate data were excluded. The final valid sample consisted of 84 participants. Because institutional identifiers were not collected, it was not possible to determine whether more than one respondent belonged to the same institution.

Data collection took place from August 2025 to April 2026. A structured online questionnaire was used as the research instrument, developed and administered through an electronic research platform. The survey link was strategically disseminated through social media, professional networks, technical discussion forums related to Clinical Engineering and Biomedical Engineering at the national level, using a non-probabilistic recruitment strategy. As the questionnaire was publicly disseminated and the number of individuals reached could not be determined, a response rate could not be calculated. The instrument comprised blocks of multiple-choice questions (demographic data, team sizing, activities performed, use of computerized tools) and open-ended questions allowing investigation of daily challenges and priority actions to enhance the sector's recognition.

The collected data were exported and subjected to different processing workflows depending on the nature of the variable. Quantitative data were processed using simple descriptive statistics. No inferential statistical analyses were performed. Qualitative data pertaining to daily challenges and priority actions for sector recognition were subjected to content analysis and categorization. All responses to the open-ended questions were read and grouped according to thematic similarity. An initial categorization was performed by one researcher and subsequently discussed by two researchers. Responses containing more than one challenge or priority action could be assigned to more than one thematic category; therefore, the categories were not mutually exclusive. No inter-rater agreement statistic was calculated.

The research protocol was approved by the Research Ethics Committee of the Federal University of Santa Catarina (UFSC) under approval nº 7.706.804 (Nº CAAE 87849125.4.0000.0121). All participants expressly provided their free and informed consent on the initial page of the online platform before completing the questionnaire, ensuring respondent anonymity and the confidentiality of institutional information collected.

III. Results

Regarding the geographic distribution of the 84 respondents, participation was observed from all regions of Brazil. The Southeast region accounted for the highest proportion of responses (47.62%), followed by the Northeast (16.67%) and Center-West (15.48%) regions. The South and North regions corresponded to 13.10% and 3.57% of responses, respectively. Additionally, in 3.57% of cases it was not possible to identify the respondents' region of origin.

To understand the breadth of responsibilities held by hospital Clinical Engineering departments in Brazil, participants were asked about the types of assets under their teams' direct management. The consolidated results reveal that the sector's scope extends significantly beyond the exclusive management of medical equipment, as presented in Figure 1.

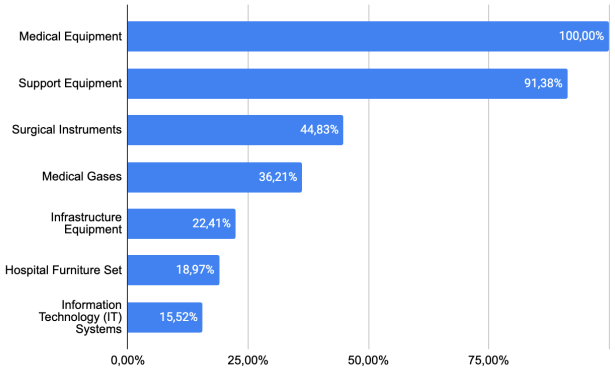


Fig. 1. Health technologies managed by Clinical Engineering departments in Brazil, as reported by participants

All Clinical Engineering departments reported managing Medical-Assistive Equipment. In parallel, the management of Support Equipment (91.38%) reinforces the role of Clinical Engineering as a key contributor to life-support and sanitary safety infrastructure. The expansion into Surgical Instruments (44.83%) and Medicinal Gases (36.21%) reflects a trend toward scope centralization by administrations seeking greater cost control and risk mitigation in high-financial-impact supplies. Finally, involvement in Infrastructure (22.41%) and IT (15.52%) outlines a trend toward integration across different areas.

Regarding the workforce contracting model, the results revealed a scenario distributed between in-house and outsourced service delivery. The in-house team model emerged as the most frequent in the sample, adopted by 42.8% of institutions (n=36), as illustrated in Figure 2.

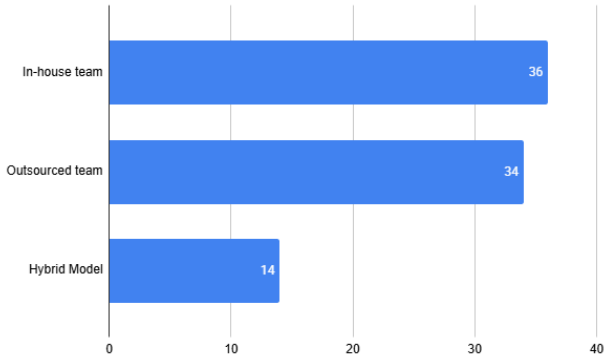


Fig. 2. Clinical Engineering team models in Brazil.

On the other hand, the full outsourcing model demonstrated strong market penetration in Brazil, accounting for 40.5% of the mapped cases (n=34). Finally, the hybrid model characterized by shared operational responsibilities between an in-house management core and companies contracted on demand was reported by 16.7% of respondents (n=14).

Regarding the sizing of Clinical Engineering departments, most institutions have up to 10 professionals, with a predominance of teams with 6 to 10 members (n=29). Next are teams composed of 11 to 20 professionals (n=20), 4 to 5 (n=13), and 1 to 3 professionals (n=12), as presented in Figure 3. Larger teams are less frequent, identified in only seven institutions with 21 to 50 professionals and in two institutions with more than 50 members (Figure 3).

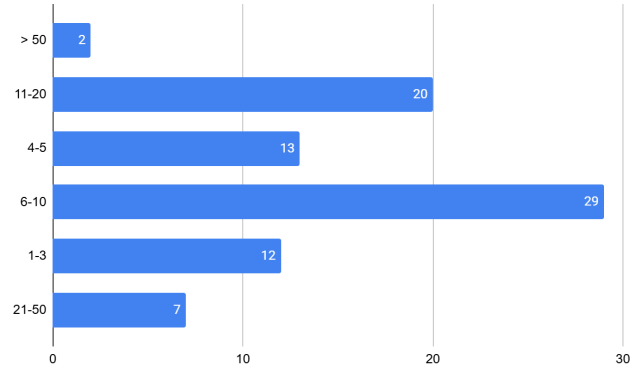


Fig. 3. Size distribution of Clinical Engineering teams in Brazilian hospitals.

Regarding team composition (Figure 4), technicians predominate, accounting for 54.6% of professionals, followed by engineers (22.1%). The category of other professionals, responsible for 18.8% of team composition, includes professionals from different fields, such as medical physicists, nursing technicians, interns, administrative staff, lawyers, advisors, and young apprentices. Technologists represent 4.4% of the total.

.

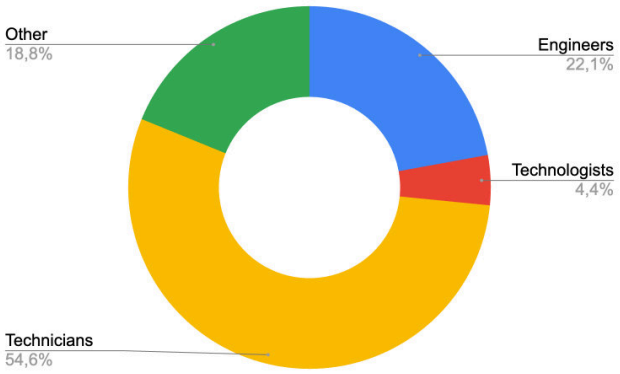


Fig. 4. Professional composition of Clinical Engineering teams

Analyzing the activities performed internally and externally (Figure 5), the results indicate a strong trend toward outsourcing most Clinical Engineering activities. Strategic functions, such as the incorporation of medical equipment (91%) and the implementation of maintenance programs (90%), operational activities such as equipment installation (89%), preventive maintenance (89%), and corrective maintenance (88%), as well as managerial functions such as risk management (66%), indicator analysis (65%), and adverse event management (61%), were predominantly performed in-house. This indicates that organizations maintain direct control over the day-to-day management of technologies.

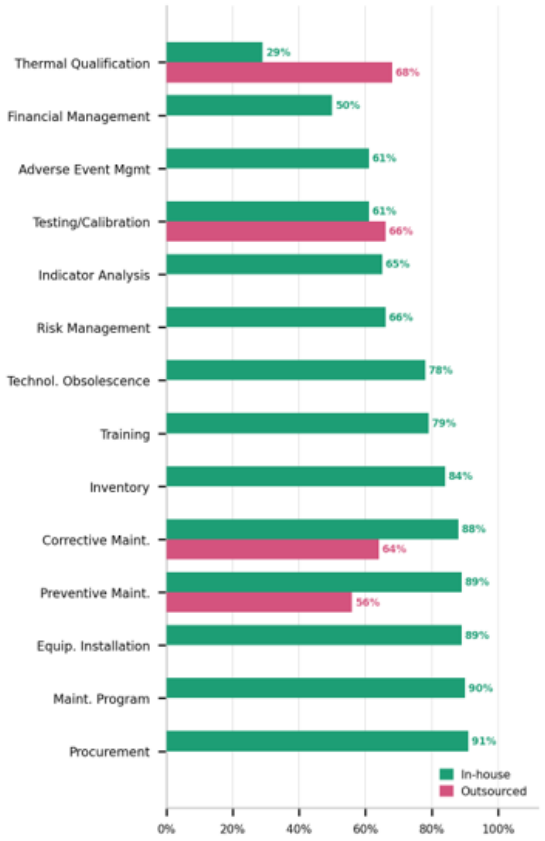


Fig. 5. Activities performed by Clinical Engineering departments in hospitals in Brazil.

In contrast, external activities were concentrated in areas such as thermal qualification, which showed the highest outsourcing rate (68%) and the lowest internal execution rate (29%). Testing and calibrations also showed a high frequency of external execution (66%), although 61% of institutions perform these activities in-house. Similarly, corrective maintenance (64%) and preventive maintenance (56%) present significant rates of external execution.

The overlap between internal and outsourced execution for maintenance and calibration highlights the adoption of a hybrid model, in which routine strategic and operational activities are maintained in-house, while outsourcing is selectively applied for specialized services requiring specific tools, mandatory certifications, or direct manufacturer support.

Regarding the adoption of computerized systems (maintenance/asset management software) by Clinical Engineering departments, results indicate strong consolidation of these digital tools, with 86.9% of institutions (n=73) using them actively for asset management.

Among the 84 participants included in the study, 82 provided valid responses to the question on daily challenges. The analysis of responses revealed a complex landscape of daily challenges faced in Clinical Engineering, highlighting that the greatest obstacles transcend the technical sphere and are concentrated in human and market factors (Table II). Within institutions, the interface with clinical staff emerged as the most frequently reported challenge, with 31.71% of mentions (n=26).

Reports consistently pointed to equipment misuse, difficulties in equipment handling, and low adherence to operational training. These internal challenges were accompanied by financial and budgetary constraints (15.85%) and workforce-related issues, including understaffing, limited availability of qualified professionals, and team turnover (20.73%).

TABLE II. CHALLENGES FACED BY CLINICAL ENGINEERING DEPARTMENTS

Challenge category	*Percentage of respondents (%)*	*Main evidence from reports (examples)*
Clinical Staff Interface / Equipment Misuse / Training	31.71%	Equipment misuse, low adherence to training, and user disorganization.
Spare Parts Logistics and Supplier Relationships	18.29%	Manufacturer delays, lengthy after-sales lead times, high spare-parts costs, and dependence on external technical support
Financial, Budgetary, and Infrastructure Issues	15.85%	Limited resources, insufficient budget for fleet renewal, outdated salaries, and inadequate tools.
Workforce, Staffing, and Team Turnover	20.73%	Chronic overload of beds per technician, lack of specialized staff, and high professional turnover.
Work Overload, High Demand, and Deadline Pressure	15.85%	Pressure to meet repair-time targets, a culture in which “everything is an emergency,” and growing corrective backlog.
Professional Recognition and Gender Challenges	13.41%	Lack of recognition of technical expertise, assignments outside the scope of practice, and gender bias in the field.
Administrative Bureaucracy and Public Management Issues	8.54%	Slow public procurement processes and delays in budget approval procedures.
Other General Operational Challenges	9.76%	Difficulty locating equipment, managing service reports, and surgical center routines.

On the other hand, in the external ecosystem, Clinical Engineering faces a logistical challenge in its relationship with suppliers and manufacturers (18.29%). Professionals reported slow after-sales services, long lead times, and high costs for acquiring spare parts, as well as commercial barriers limiting access to service software.

TABLE III. PRIORITY ACTIONS TO IMPROVE CLINICAL ENGINEERING IN BRAZIL

Priority action category	*Percentage of respondents (%)*	*Core themes identified in responses*
Compensation, Benefits, and Labor Rights	24.62%	Salary adjustments, inclusion of occupational allowances (hazard pay and unhealthy working conditions allowances), and compensation compatible with technical responsibilities.
Presentation of Performance Indicators and Financial Results	18.46%	Demonstrating financial return, savings in procurement processes, and costs avoided due to reduced equipment downtime.
Communication, Relationships, and Internal Dissemination	27.69%	Breaking the sector's isolation, improving dialogue with clinical teams, and providing effective feedback on service orders.
Technical and Operational Training	18.46%	Investment in training on emerging medical technologies and encouragement of continuous professional development.
Strategic Engagement and Integration with Senior Management	26.15%	Participating in decision-making processes, developing a holistic perspective, and aligning activities with hospital leadership.
Physical Infrastructure, Budget, and Resources	10.77%	Ensuring adequately equipped workshops, synchronization with building maintenance, and allocation of dedicated budgets.
Regulatory Oversight and Inspection	6.15%	Active supervision by professional councils regarding job positions, salaries, and irregular hiring practices in the hospital sector.

Among the 84 participants, 65 provided valid responses to the question regarding priority actions to promote the recognition and appreciation of Clinical Engineering. The responses revealed the need for structural labor reforms, combined with an urgent need to strategically reposition teams within the hospital environment.

When investigating priority actions to promote professional recognition of Clinical Engineering, results revealed the need for structural labor reforms, combined with an urgent need to strategically reposition teams within the hospital environment. Communication, relationships, and internal dissemination emerged as the most frequently identified category, with 27.69% of respondents (n=18) emphasizing the importance of strengthening dialogue with clinical teams, disseminating the activities and results of Clinical Engineering, and improving intersectoral relationships.

Strategic engagement and integration with senior management were also frequently identified (26.15%; n=17), particularly through participation in decision-making processes and the presentation of evidence and results to institutional leadership. Compensation, benefits, and labor rights were reported by 24.62% of respondents (n=16), highlighting salary adjustments, benefits, and recognition compatible with professional responsibilities. Additionally, respondents emphasized technical and operational training (18.46%; n=12), particularly continuous training and preparation for new technologies, as well as the presentation of performance indicators and financial results (18.46%; n=12) to demonstrate the economic contribution of Clinical Engineering. Physical infrastructure, budget, and resources (10.77%; n=7) and regulatory oversight and inspection (6.15%; n=4) were also identified as priority areas (Table III).

IV. Discussion

The results show that Clinical Engineering in Brazil constitutes an interdisciplinary structure. Although definitions frequently center the field on the clinical engineer as its central figure, the operational and strategic reality of Health Technology Management demands a shift from this paradigm toward a multiprofessional ecosystem model. In the interest of patient safety and risk mitigation, the analytical structure of contemporary Clinical Engineering is now composed of diverse competencies: engineers from various specialties, technologists, and technicians, health professionals and medical physicists, as well as administrators, lawyers, architects, among others. This convergence of knowledge demonstrates that technological safety in hospital environments is not the product of isolated engineering, but rather the result of collaborative and integrated governance across different sectors.

The unanimous convergence around the management of Medical-Assistive Equipment (100.00%) confirms the sector's identity; however, it is the significant management of Support Equipment (91.38%), Surgical Instruments (44.83%), Medicinal Gases (36.21%), Infrastructure (22.41%), and IT (15.52%) that redraws the boundaries of the sector's scope. However, it is important to emphasize that the fact that Clinical Engineering departments manage multiple asset types does not, by itself, prove an epistemological redefinition of the field. Rather than representing a planned strategic evolution, this expansion of scope may also reflect local administrative centralization, resource scarcity, or an accumulation of operational responsibilities due to institutional demands.

The distribution between in-house teams (42.8%), full outsourcing (40.5%), and the hybrid model (16.7%) indicates the coexistence of different organizational arrangements for Clinical Engineering service. Cruz and Rincon (2012), in a mapping review on the outsourcing of medical device maintenance, identified that the decision to internalize or outsource services is influenced by factors such as asset specificity, uncertainty, and transaction frequency [34].

The concentration of outsourcing in specific activities such as thermal qualification (68%), testing and calibrations (66%), and corrective maintenance (64%) suggests that outsourcing is selectively applied to activities that may require specific expertise, equipment, certification, or manufacturer support. However, Cruz and Rincon (2012) note that this decision frequently occurs without a consolidated methodological basis, given that no empirical model yet exists to integrally guide supplier selection and performance evaluation in the context of medical device maintenance [34].

The adoption of continuous training programs aimed at mitigating operational errors, the development of standard operating procedures (SOPs), the consolidation of performance indicators, and the transition to evidence-based management constitute fundamental technical-operational guidelines for Clinical Engineering, converging with the landscape discussed by Iadanza (2023). However, the results suggest that strengthening Clinical Engineering requires actions that extend beyond technical activities, including improved inter-sectoral communication, greater strategic engagement with senior management, recognition of professional responsibilities and compensation, and continued technical training. These priorities were reflected in the open-ended responses, particularly in relation to communication and relationships (27.69%), strategic engagement with senior management (26.15%), and compensation, benefits, and labor rights (24.62%). An analysis of the profile reveals a gap between traditional academic training and emerging market demands. It is recommended that continuing education programs be aligned with competencies in risk management, human factors engineering, and digital transformation.

Additionally, the need for more robust oversight by governmental bodies and professional councils is highlighted, consolidating a governance framework that integrates regulatory compliance and institutional recognition [35]. These results highlight the need for public policies aimed at strengthening Clinical Engineering, encouraging the creation of further national guidelines for Health Technology Management (HTM) and fostering continuous professional training aligned with regulatory and strategic demands. As main lessons learned, the study demonstrates that Clinical Engineering in Brazil operates under significant maturity heterogeneity across the public, private, and philanthropic sectors, revealing that while technical proficiency in medical equipment maintenance

and fleet management is well-established, a gap persists between traditional academic training and strategic management skills. Conversely, key research gaps were identified, underscoring the need for future studies that correlate Clinical Engineering maturity with patient safety indicators, evaluate the cost-effectiveness of HTM throughout the technology lifecycle, and investigate the impact of human factors engineering and usability on preventing use errors within the Brazilian healthcare context.

V. Conclusion

This study mapped the landscape of Clinical Engineering in hospital institutions in Brazil, revealing the structural, operational, and market complexity involved in the management of medical equipment. Based on a quantitative and qualitative approach with a sample of 84 professionals, the results provided an exploratory characterization of sector digitalization, and challenges that limit its full institutional recognition.

The high rate of adoption of computerized management systems (86.90%) and the significant rates of internal execution of strategic activities, such as maintenance program planning (90%) and the incorporation of new technologies (91%), suggest increasing involvement of Clinical Engineering in strategic activities. However, this technical advancement contrasts with severe behavioral and structural barriers. The interface with clinical staff emerged as the most prevalent challenge (31.71%), manifested through equipment misuse and low engagement with operational training, which may contribute to user-related failures and pose potential risks to patient safety. Additionally, the sector operates under multifactorial internal pressures, including chronic budgetary constraints and understaffed teams relative to bed volume, and external pressures driven by manufacturer dependency, which imposes high costs and long lead times for spare parts.

As a path to overcoming institutional invisibility, the Clinical Engineering community indicates that financial and salary recognition (24.62%) must be accompanied by a proactive change in posture from the teams themselves. This translates into the need for professionals to move beyond purely technical repair activities and to communicate with hospital senior leadership through quantifiable data, numerically demonstrating financial return (ROI), costs avoided through reduced equipment downtime, and the sector's direct impact on the micro- and macroeconomic sustainability of the health institution.

In this context, the findings reinforce the importance of implementing and systematically reviewing the Health Technology Management Plan (HTMP) as a strategic instrument for health technology governance. Its periodic updating enables the incorporation of improvement opportunities, supports the rapid technological evolution of the sector, and addresses changes in the population's epidemiological profile, thereby promoting greater operational efficiency, regulatory compliance, institutional sustainability, and strengthened patient safety.

Regarding limitations, the study was conducted through a self-administered questionnaire and voluntary public dissemination, making the results subject to respondent bias, which may lead to omissions or underreporting of failures in the health technology management process. Although the study achieved national coverage, the representativeness of the sample cannot be established due to the non-probabilistic recruitment and the uneven distribution of participants, with greater concentration in the Southeast region. Because institutional identifiers were not collected, it was not possible to determine whether more than one respondent belonged to the same institution, and the public dissemination of the questionnaire prevented the calculation of a response rate.

Furthermore, the descriptive nature of the study and the absence of inferential statistical analyses limited the investigation of associations between institutional profile, management model, and reported challenges. Stratified and inferential analyses using larger and more representative samples are suggested in future investigations to identify regional and institutional specificities. Nevertheless, this study provides exploratory national evidence on the landscape of Clinical Engineering in Brazil, contributing to discussions on public policies, professional development, and new scientific directions for the field.

These empirical findings serve as a foundation for a critical conceptual discussion on the expansion and evolution of the profession's scope in the healthcare ecosystem. Based on the evidence, a redefinition of the concept of Clinical Engineering is proposed, dissociating it from a purely technical, clinical engineer-centered view, to establish it as a holistic field oriented toward sustainability and safety in healthcare. Under this perspective, Clinical Engineering operates as an interdisciplinary link aimed at managing health technologies throughout the entire lifecycle, ensuring metrological reliability, asset traceability, accessibility, and safety for the mitigation of adverse events. Managing technology must therefore become more strategic and evidence-based.

In conclusion, this study provides important exploratory evidence that Brazilian Clinical Engineering increasingly operates as an expanded, interdisciplinary field encompassing management, risk, technology, infrastructure, and governance. However, a formal conceptual redefinition of the discipline requires future triangulation with institutional databases, formal consultation with professional societies and regulatory bodies, cross-national comparative studies, and empirical validation through larger, more representative, and stratified samples.

References

1. World Health Organization (WHO), Global Atlas of Medical Devices: WHO Medical Devices Technical Series, Geneva, Switzerland, 2017
2. M. R. Signori and R. Garcia, “Clinical engineering incorporating human factors engineering into risk management,” in World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany (IFMBE Proceedings, vol. 25/7), O. Dössel and W. C. Schlegel, Eds. Berlin, Heidelberg: Springer, 2009, pp. 446–449, doi: 10.1007/978-3-642-03885-3_125.
3. World Health Organization (WHO), "Human resources for medical devices, the role of biomedical engineers," WHO Medical Device Technical Series, Geneva, Suíça, ago. 2022. [Online]. Available: https://www.who.int/publications/i/item/9789241565479.
4. R. Garcia et al., "Health care technology management applied to public hospitals in Santa Catarina - Brazil," in *Proc. First WHO Global Forum on Medical Devices*, Geneva, Switzerland, 2010.

5. American College of Clinical Engineering (ACCE), "About ACCE: Clinical Engineer," 1992. [Online]. Available: https://accenet.org/about/Pages/ClinicalEngineer.aspx. [Accessed: Jun. 19, 2026].
6. Bronzino, J.D. (2006) The Biomedical Engineering Handbook, Vol. I: Biomedical Engineering Fundamentals.
7. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 509, May 27, 2021. Provides for the management of health technologies in healthcare establishments, Official Gazette of the Union, Brasília, DF, Brazil.
8. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 36, July 25, 2013. Establishes actions for patient safety in healthcare services and other provisions, Official Gazette of the Union, Brasília, DF, Brazil.
9. Brazilian Association of Technical Standards (ABNT), ABNT NBR 15943:2011 – Guidelines for a Medical Equipment Management Program, Rio de Janeiro, RJ, Brazil, 2011.
10. Brazilian Association of Technical Standards (ABNT), ABNT NBR ISO 9001:2015 – Quality Management Systems - Requirements, Rio de Janeiro, RJ, Brazil, 2015.
11. Brazilian Association of Technical Standards (ABNT), ABNT NBR ISO 14971:2020 – Medical Devices - Application of Risk Management to Medical Devices, Rio de Janeiro, RJ, Brazil, 2020.
12. Brazilian Association of Technical Standards (ABNT), ABNT NBR ISO 7101:2025 – Healthcare Organization Management - Quality Management Systems in Healthcare Organizations, Rio de Janeiro, RJ, Brazil, 2025.
13. Brazil, Law Nº 14,133, April 1, 2021: Public Procurement and Administrative Contracts Law, Official Gazette of the Union, Brasília, DF, Brazil, 2021.
14. Brazil, Law Nº 15,210, 2025: Establishes requirements for the purchase of equipment intended for diagnostic or therapeutic procedures within the Brazilian Unified Health System (SUS), Official Gazette of the Union, Brasília, DF, Brazil, 2025.
15. Brazil, Ministry of Health, Ordinance GM/MS Nº 11,694, 2026: Provides planning guidelines and recommendations for the acquisition of equipment intended for diagnostic or therapeutic procedures within the Brazilian Unified Health System (SUS), Official Gazette of the Union, Brasília, DF, Brazil, 2026.
16. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 50, February 21, 2002. Provides the Technical Regulation for the Planning, Programming, Design, and Evaluation of Physical Projects for Healthcare Facilities, Official Gazette of the Union, Brasília, DF, Brazil.
17. Brazilian Association of Technical Standards (ABNT), ABNT NBR 13534:2008 – Low-Voltage Electrical Installations - Specific Requirements for Installations in Healthcare Facilities, Rio de Janeiro, RJ, Brazil, 2008.
18. Brazilian Association of Technical Standards (ABNT), ABNT NBR 16651:2019 – Fire Protection in Healthcare Facilities, Rio de Janeiro, RJ, Brazil, 2019.
19. Brazilian Association of Technical Standards (ABNT), ABNT NBR 12188:2016 – Centralized Supply Systems for Medical Gases, Gases for Medical Devices, and Vacuum Systems for Use in Healthcare Services, Rio de Janeiro, RJ, Brazil, 2016.
20. Brazilian Association of Technical Standards (ABNT), ABNT NBR 7256:2022 – Air Treatment in Healthcare Facilities - Requirements for the Design and Installation of Systems, Rio de Janeiro, RJ, Brazil, 2022.
21. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 751, September 25, 2022. Provides for risk classification, notification and registration requirements, and labeling and instructions for use of medical devices, Official Gazette of the Union, Brasília, DF, Brazil.
22. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 579, November 19, 2021. Provides for the importation, commercialization, and donation of used and refurbished medical devices, Official Gazette of the Union, Brasília, DF, Brazil.
23. Brazilian Association of Technical Standards (ABNT), ABNT NBR IEC 60601-1:2010 – Medical Electrical Equipment - Part 1: General Requirements for Basic Safety and Essential Performance, Rio de Janeiro, RJ, Brazil, 2010.
24. Brazilian Association of Technical Standards (ABNT), ABNT NBR ISO 17025:2017 – General Requirements for the Competence of Testing and Calibration Laboratories, Rio de Janeiro, RJ, Brazil, 2017.
25. Brazilian Association of Technical Standards (ABNT), "ABNT NBR IEC 62353:2019 – Medical Electrical Equipment—Recurrent Tests and Tests After Repair of Medical Electrical Equipment," Rio de Janeiro, RJ, Brazil, 2019.
26. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 67, December 21, 2009. Provides technovigilance requirements applicable to medical device registration holders in Brazil, Official Gazette of the Union, Brasília, DF, Brazil.
27. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 925, September 19, 2024. Provides requirements for the manufacture, commercialization, importation, and placing into use of personalized medical devices, Official Gazette of the Union, Brasília, DF, Brazil.
28. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 63, November 25, 2011. Provides for the Requirements of Good Operating Practices for Healthcare Services, Official Gazette of the Union, Brasília, DF, Brazil.
29. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 07, February 25, 2010. Establishes Minimum Requirements for the Operation of Intensive Care Units and Other Provisions, Official Gazette of the Union, Brasília, DF, Brazil.
30. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 15, March 15, 2012. Provides Requirements for Good Practices in the Processing of Healthcare Products and Other Provisions, Official Gazette of the Union, Brasília, DF, Brazil.
31. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 611, March 09, 2022 Establishes the Health Requirements for the Organization and Operation of Diagnostic and Interventional Radiology Services, Official Gazette of the Union, Brasília, DF, Brazil.
32. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 11, March 13, 2014. Provides Requirements for Good Operating Practices for Dialysis Services and Other Provisions, Official Gazette of the Union, Brasília, DF, Brazil.
33. Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution - RDC nº 222, March 28, 2018. Establishes good practices for the management of healthcare waste and other provisions, Official Gazette of the Union, Brasília, DF, Brazil
34. A. M. Cruz and A. M. R. Rincon, "Medical device maintenance outsourcing: Have operation management research and management theories forgotten the medical engineering community? A mapping review," *European Journal of Operational Research*, vol. 221, no. 1, pp. 186–197, Aug. 2012
35. Iadanza E, Gonnelli V, Satta F, Gherardelli M. Evidence-based medical equipment management: a convenient implementation. Med Biol Eng Comput. 2019 Oct;57(10):2215-2230. doi: 10.1007/s11517-019-02021-x. Epub 2019 Aug 10. PMID: 31399897; PMCID: PMC6791913.

Conformidade Técnica e Normativa em Contratos de Gestão de Resíduos Hospitalares na Rede Pública Estadual do Tocantins

Paulo Jadir Vieira Medeiros Baia
Curso de Engenharia Biomédica
Universidade Federal do Norte do Tocantins (UFNT)
Araguaína, Brasil
paulo.baia@ufnt.edu.br

Fábio Matos Rodrigues
Departamento de Física e Tecnologias Interdisciplinares (DEFIT)
Universidade Federal do Norte do Tocantins (UFNT)
Araguaína, Brasil
fabio.rodrigues@ufnt.edu.br

***Resumo*—Este estudo analisa a convergência técnica e normativa entre as diretrizes de biossegurança (RDC ANVISA n.º 222/2018 e Resolução CONAMA n.º 358/2005) e a estruturação dos instrumentos convocatórios para gestão de Resíduos de Serviços de Saúde (RSS) na rede pública estadual do Tocantins. Por meio de pesquisa documental qualiquantitativa centrada na análise de conteúdo, escrutinou-se o Termo de Referência do Pregão Eletrônico n.º 90147/2024, que rege a logística de 44 estabelecimentos assistenciais. O escopo restringe-se à avaliação do projeto do edital (*tender design*), não constituindo auditoria de execução prática nas unidades. Os resultados evidenciam alinhamento formal às normas federais, destacando-se a exigência de inativação microbiana de Nível III, incineração para partes anatômicas e requisitos de rastreabilidade digital (balanças calibradas pelo INMETRO, monitoramento via GPS e emissão do Manifesto de Transporte de Resíduos via SINIR). O Instrumento de Medição de Resultado (IMR) estabelece fórmula objetiva de glosa financeira baseada no Fator de Qualidade (FQ). Identificaram-se lacunas relacionadas à alta carga de auditabilidade e à dependência de autodeclaração das calibrações. Conclui-se que o desenho do edital fornece um arcabouço normativo estruturado para a fiscalização da engenharia clínica.**

***Index Terms*—Resíduos de Serviços de Saúde, Engenharia Clínica, Gestão Contratual, Rastreabilidade, Licitações Públicas.**

I. Introdução

O Estado do Tocantins administra uma rede de unidades hospitalares dimensionada para absorver demandas de média e alta complexidade. Polos estratégicos como o Hospital Geral de Palmas (HGP), o Hospital Regional de Araguaína (HRA) e o Hospital Regional de Gurupi (HRG) operam como geradores contínuos de Resíduos de Serviços de Saúde (RSS). A escala operacional dessas unidades implica a necessidade de rígido controle sanitário e ambiental para mitigar riscos à saúde pública.

O alicerce jurídico nacional é estabelecido pela Lei n.º 12.305/2010 [1], que institui a Política Nacional de Resíduos Sólidos (PNRS) e fixa a responsabilidade compartilhada do gerador. No plano operacional, a Resolução da Diretoria Colegiada (RDC) n.º 222/2018 da ANVISA [2] padroniza as etapas de segregação, acondicionamento e tratamento. Complementarmente, a ABNT NBR 10004:2004 [3] estabelece os parâmetros de classificação de periculosidade dos resíduos.

A proteção ambiental e a saúde ocupacional são regidas pela Resolução CONAMA n.º 358/2005 [4] e pelas Normas Regulamentadoras NR-9 e NR-32. A NR-32 estabelece diretrizes específicas para a segurança dos trabalhadores em serviços de saúde. Neste contexto, a Engenharia Clínica atua na interpretação, monitoramento e fiscalização dos fluxos contratuais, assegurando a aplicação dos parâmetros técnicos estipulados nos Termos de Referência (TR).

A investigação centra-se na análise do grau de convergência técnica entre a legislação sanitária vigente e o projeto instrucional do edital de licitação da Secretaria de Estado da Saúde do Tocantins (SES/TO). O presente trabalho avalia especificamente a estrutura normativa do instrumento convocatório (*tender design*), não abrangendo a verificação empírica da execução do serviço em campo.

II. Objetivos

A. Objetivo Geral

Analisar a convergência técnica e normativa entre as diretrizes da RDC n.º 222/2018 da ANVISA e as cláusulas estipuladas no edital público de gerenciamento de RSS da rede estadual do Tocantins para os Grupos A, B e E.

B. Objetivos Específicos

- Identificar as exigências normativas de segregação, acondicionamento e destinação dos Grupos A, B e E estipuladas na legislação federal [2].
- Mapear a estrutura contratual recente elaborada pela SES/TO para coleta e tratamento de RSS [7].
- Escrutinar o Termo de Referência, registrando o nível de detalhamento técnico frente aos requisitos normativos e classificando as provisões contratuais [2], [4].
- Avaliar a viabilidade dos mecanismos de fiscalização previstos no edital para atuação da engenharia clínica.

Os autores agradecem à Universidade Federal do Norte do Tocantins (UFNT) pelo suporte institucional e à Secretaria de Estado da Saúde do Tocantins (SES/TO) pela disponibilização dos documentos públicos que fundamentaram esta pesquisa documental.

III. Referencial Teórico

A. Arcabouço Normativo de Biossegurança

A RDC ANVISA n.º 222/2018 padroniza o gerenciamento de RSS e exige a elaboração do Plano de Gerenciamento de Resíduos de Serviços de Saúde (PGRSS) [2]. A norma impõe critérios para segregação na fonte, identificação e inativação de agentes patogênicos. Todos os normativos citados neste trabalho encontravam-se em plena vigência temporal durante a elaboração e publicação do certame analisado.

B. Rastreabilidade Eletrônica e Nova Lei de Licitações

A Portaria MMA n.º 280/2020 instituiu o Manifesto de Transporte de Resíduos (MTR) nacional via Sistema Nacional de Informações sobre a Gestão dos Resíduos Sólidos (SINIR) [8]. Paralelamente, a Lei Federal n.º 14.133/2021 (Nova Lei de Licitações) formalizou o uso do Instrumento de Medição de Resultado (IMR) para atrelar o pagamento ao cumprimento de metas contratuais objetivas [9].

C. Transporte Rodoviário de Cargas Perigosas

O transporte externo de RSS é regulado pela Resolução ANTT n.º 5.998/2022 [10]. Em nível estadual, a Resolução COEMA-TO n.º 07/2005 exige a Autorização de Transporte de Cargas Perigosas (ATCP) expedida pelo NATURATINS [11].

IV. Metodologia

A. Protocolo de Análise Documental e Fontes de Dados

Realizou-se pesquisa documental, descritiva e exploratória, com abordagem qualiquantitativa baseada em análise de conteúdo. A coleta dos documentos públicos ocorreu em 15 de abril de 2024 via Portal Nacional de Contratações Públicas (PNCP). O objeto de análise foi o Termo de Referência do Pregão Eletrônico n.º 90147/2024 (Processo Administrativo n.º 2023/30550/006554), promovido pela SES/TO para contratação de serviços em 44 estabelecimentos de saúde [7].

B. Unidades de Análise e Matriz de Codificação

O protocolo estruturou-se em quatro fases reprodutíveis:

1) Levantamento normativo: Mapeamento dos requisitos compulsórios das Resoluções RDC 222/2018 e CONAMA 358/2005.
2) Codificação contratual: Extração sistemática das cláusulas técnicas e operacionais contidas no Termo de Referência.
3) Categorização e vínculo: Classificação de cada cláusula em três categorias: (a) Provisão Compulsória Legal (exigência direta de lei ou resolução); (b) Prática Recomendada (especificação de boas práticas operacionais); e (c) Interpretação dos Autores (análise técnica sobre a aplicabilidade).
4) Avaliação do desenho do edital: Confronto estruturado entre os requisitos normativos e o texto contratual (*tender design*).

C. Delimitação do Escopo e Limitações

Este estudo limita-se estritamente à análise do texto e das especificações contidas no instrumento convocatório. A metodologia não inclui auditoria física nas 44 instalações de saúde, entrevista com gestores, nem coleta de dados operacionais passados de execução.

D. Aspectos Éticos e Transparência

Por utilizar dados secundários de acesso público, o estudo é isento de avaliação por Comitê de Ética em Pesquisa (CEP/CONEP). Informa-se que o modelo de inteligência artificial generativa Gemini (Google) foi utilizado exclusivamente para apoio na revisão gramatical, adequação ao estilo IEEE e formatação do código LaTeX. Os autores assumem responsabilidade integral pelo conteúdo, análises e conclusões do manuscrito.

V. Resultados e Discussão

A. Dimensionamento e Estruturação do Licitamento

O Pregão Eletrônico n.º 90147/2024 abrange a prestação de serviços de coleta, transporte, tratamento e disposição final de RSS em 44 unidades de saúde estaduais [7]. A demanda total prevista é de 17.420.400 kg ao longo de 60 meses de contrato, dividida no Lote 1 (Região Norte, 5,8 mil toneladas) e Lote 2 (Região Centro-Sul, 11,5 mil toneladas) [7]. O edital fixa frequências de coleta variáveis conforme a complexidade da unidade, variando de mensal para unidades ambulatoriais até três vezes por semana para hospitais de grande porte (HRA, HGP e HRG) [7].

B. Matriz Auditável de Conformidade do Desenho do Edital

A Tabela I apresenta a matriz de conformidade obtida pelo confronto entre a legislação vigente e as cláusulas do Termo de Referência.

Tabela I
MATRIZ AUDITÁVEL DE CONFORMIDADE DO DESENHO DO EDITAL (PE 90147/2024)

Grupo	Requisito Legal / Normativo	Especificação no Termo de Referência (TR)	Categoria de Análise
Grupo A (Infectante)	Inativação Nível III (A1/A2); Incineração compulsória para A3/A5 (RDC 222/2018 [2]).	Exigência de inativação Nível III e incineração para A3/A5, com Laudo de Queima Semestral (Cláusula 4.2 [7]).	Provisão Compulsória Legal
Grupo B (Químico)	Proibição de efluentes líquidos em aterros; Destinação em Aterro Classe I (RDC 222/2018 [2]).	Vedação de líquidos em aterro; exigência de neutralização prévia e descarte em Aterro Classe I (Cláusula 4.8 [7]).	Provisão Compulsória Legal
Grupo E (Perfurocortante)	Acondicionamento em recipiente rígido e estanque; Inativação compatível (RDC 222/2018 [2]).	Comodato de bombonas rígidas em PEAD e tratamento em Nível III prévio ao aterro (Cláusula 5.1 [7]).	Prática Recomendada / Legal

Fonte: Elaborado pelos autores (2026).

A especificação do TR exige que os subgrupos A3 (peças anatômicas) e A5 (suspeita de príons) sejam destinados estritamente à incineração, solicitando a apresentação semestral

de Testes de Eficiência de Queima emitidos por laboratórios credenciados, em conformidade com a Resolução CONAMA n.º 316/2002 [7].

C. Regras de Subcontratação

O edital estabelece limite máximo de 30% para subcontratação do objeto global, distribuído em até 25% para disposição final em aterro e até 5% para incineração dos subgrupos A3 e A5 [7]. O TR exige que as empresas subcontratadas apresentem Licença de Operação (LO) válida e comprovem regularidade ambiental perante os órgãos competentes [7].

D. Requisitos Metrológicos e Rastreabilidade Eletrônica

O TR exige que a empresa contratada forneça bombonas de polietileno de alta densidade (PEAD) em comodato e instale ao menos duas balanças digitais industriais calibradas por unidade hospitalar [7]. As balanças devem possuir capacidade mínima de 200 kg, plataforma de 60x60 cm, impressora de etiquetas integrada e certificação emitida pela Rede Brasileira de Calibração (RBC/INMETRO) [7].

O registro das pesagens fundamenta a emissão do Manifesto de Transporte de Resíduos (MTR) via sistema SINIR, atendendo à Portaria MMA n.º 280/2020 [8]. Para o transporte rodoviário, exige-se frota com baú estanque, sinalização conforme ABNT NBR 7500, licença ambiental ATCP e monitoramento por GPS [7], [10], [11].

E. Mecanismo do Instrumento de Medição de Resultado (IMR)

O modelo contratual adota a medição de desempenho por meio do Fator de Qualidade (FQ), calculated pela equação:

$$FQ = \left[\frac{(A - D) - C}{(A - D)} \right] \times 100 \tag{1}$$

Em que A representa o total de serviços avaliados, D indica os serviços dispensados ou não aplicáveis no período, e C representa o somatório de infrações técnicas ou operacionais registradas pela fiscalização.

O edital estabelece descontos faturados progressivos conforme o valor de FQ:

- $FQ \geq 90\%$: Pagamento integral (100% do valor mensal).
- $75\% \leq FQ < 90\%$: Glosa de 2% sobre a fatura mensal.
- $50\% \leq FQ < 75\%$: Glosa de 5% sobre a fatura mensal.
- $FQ < 50\%$: Glosa de 10% sobre a fatura mensal e abertura de processo administrativo de rescisão contratual.

F. Análise Crítica de Lacunas e Gargalos do Edital

Apesar da convergência formal identificada, a análise neutra do documento aponta limitações potenciais:

- Carga de auditabilidade: A fórmula do IMR exige verificação frequente de laudos de calibração, certidões ambientais e registros no SINIR. Sem dimensionamento adequado das equipes de fiscalização hospitalar, a identificação da variável C pode ser comprometida.
- Dependência de autodeclaração: O controle metrológico fundamenta-se no envio de certificados emitidos pela própria contratada ou por laboratórios por ela contratados, sem previsão expressa de aferição independente periódica por parte do contratante.

G. Requisitos de Saúde Ocupacional e Gestão do Fator Humano

Em conformidade com a NR-32, o TR exige que a empresa contratada ministre treinamento comprovado aos trabalhadores em até 20 dias após a emissão da Ordem de Serviço ou da contratação de novos funcionários [7]. É exigida a apresentação do Programa de Controle Médico de Saúde Ocupacional (PCMSO), do Programa de Gerenciamento de Riscos (PGR) e do Plano de Prevenção de Riscos de Acidentes com Materiais Perfurocortantes (PPRAMP), acompanhados das Anotações de Responsabilidade Técnica (ART) registradas nos conselhos de classe (CREA/CRM) [7].

VI. Considerações Finais

A presente investigação cumpriu o objetivo de analisar o alinhamento técnico e normativo do Termo de Referência do Pregão Eletrônico n.º 90147/2024 da SES/TO frente às diretrizes federais da RDC ANVISA n.º 222/2018, da Resolução CONAMA n.º 358/2005 e da Lei n.º 14.133/2021.

Os resultados demonstraram que o desenho do edital (*tender design*) apresenta elevado nível de convergência documental, incorporando requisitos metrológicos (balanças aferidas pelo INMETRO), rastreabilidade digital (MTR/SINIR) e critérios objetivos de glosa financeira previstos no IMR. Contudo, reforça-se que a conformidade instrucional do edital não assegura, por si só, a eficácia do cumprimento operacional nas 44 instalações assistenciais da rede estadual.

A. Limitações e Trabalhos Futuros

Por se tratar de uma pesquisa centrada no escrutínio documental do certame, identificam-se oportunidades de ampliação da investigação para trabalhos futuros:

- Auditorias operacionais *in loco*: Realização de inspeções físicas em campo para confrontar as cláusulas do edital com a prática diária de segregação, acondicionamento e frequência de coleta nas unidades de saúde.
- Análise da série histórica do IMR: Levantamento dos relatórios mensais de execução para mensurar a aplicação real das fórmulas de glosa (FQ) e quantificar a taxa de inconformidades operacionais ao longo da vigência contratual.
- Estudos comparativos multicêntricos: Aplicação da matriz de conformidade desenvolvida neste trabalho em editais de gestão de resíduos hospitalares de outros estados brasileiros, permitindo o *benchmarking* da governança pública em saúde.
- Desenvolvimento de ferramenta padronizada de *compliance*: Criação de um *framework* auditável e reprodutível que possa ser adotado por equipes de Engenharia Clínica para a fiscalização contínua de contratos de serviços de apoio hospitalar no Sistema Único de Saúde (SUS).

REFERÊNCIAS

[1] Brasil, *Lei n.º 12.305, de 2 de agosto de 2010. Institui a Política Nacional de Resíduos Sólidos*. Diário Oficial da União, Brasília, DF, 2010.

[2] Agência Nacional de Vigilância Sanitária (ANVISA), *Resolução da Diretoria Colegiada - RDC n.º 222, de 28 de março de 2018. Regulamenta as Boas Práticas de Gerenciamento dos Resíduos de Serviços de Saúde*. Diário Oficial da União, Brasília, DF, 2018.

[3] Associação Brasileira de Normas Técnicas (ABNT), *NBR 10004: Resíduos Sólidos - Classificação*. Rio de Janeiro, 2004.

[4] Conselho Nacional do Meio Ambiente (CONAMA), *Resolução n.º 358, de 29 de abril de 2005. Dispõe sobre o tratamento e a disposição final dos resíduos dos serviços de saúde*. Diário Oficial da União, Brasília, DF, 2005.

[5] Secretaria de Estado da Saúde do Tocantins (SES/TO), *Portaria n.º 494/2022/SES/GASEC. Estabelece competências e atribuições para a atuação dos gestores e fiscais de contratos*. Diário Oficial do Estado do Tocantins, Palmas, TO, 2022.

[6] Secretaria de Estado da Saúde do Tocantins (SES/TO), *Termo de Referência n.º 6/2020/SES/SUHP*. Processo SGD: 2020/30559/087605. Palmas, TO, 2020.

[7] Secretaria de Estado da Saúde do Tocantins (SES/TO), *Edital do Pregão Eletrônico n.º 90147/2024. Processo: 2023/30550/006554*. Palmas, TO, 2024.

[8] Ministério do Meio Ambiente (MMA), *Portaria n.º 280, de 29 de junho de 2020. Institui o Manifesto de Transporte de Resíduos - MTR nacional*. Diário Oficial da União, Brasília, DF, 2020.

[9] Brasil, *Lei n.º 14.133, de 1º de abril de 2021. Lei de Licitações e Contratos Administrativos*. Diário Oficial da União, Brasília, DF, 2021.

[10] Agência Nacional de Transportes Terrestres (ANTT), *Resolução n.º 5.998, de 3 de novembro de 2022. Atualiza o Regulamento para o Transporte Rodoviário de Produtos Perigosos*. Diário Oficial da União, Brasília, DF, 2022.

[11] Conselho Estadual do Meio Ambiente do Tocantins (COEMA-TO), *Resolução n.º 07/2005. Dispõe sobre a Autorização de Transporte de Cargas Perigosas (ATCP)*. Palmas, TO, 2005.

Data quality in a computerized hospital maintenance management system within a multicenter network

1st Alexandre H. Hermini
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0000-1200-4437

2nd Miguel R. Nunes
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0009-1757-5568

3rd Gilberto S. Zunta
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-6006-2142

4th Henrique T. Moriya
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0000-0002-1595-1779

5th Edison Silva Filho
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-5397-088X

***Abstract*—Quality management in healthcare services depends on reliable records and indicators capable of supporting decision-making. This study evaluated the quality of data from closed corrective work orders extracted from a Computerized Maintenance Management System (CMMS) used in a multicenter healthcare network. A descriptive quantitative analysis of 82,431 work orders was performed in R, considering completeness, uniqueness, temporal consistency, informativeness, and the ability to support managerial indicators. Core administrative fields showed high completeness, and no duplicate values were identified in the main work-order identifiers. In contrast, downtime and return-to-operation data were missing in 80.6% of records. This high missingness persisted across the main priorities and periods evaluated and varied among equipment categories. Cost and release-for-clinical-use fields were formally completed but showed limited managerial interpretability. The database therefore provides adequate structural support for basic operational analyses, while limitations in critical managerial fields restrict the reliable generation of more advanced indicators. Effective CMMS use requires data governance, appropriate field definitions, validation rules, category standardization, user training, and continuous monitoring of data-entry quality.**

***Index Terms*—Clinical Engineering, CMMS, data quality, managerial indicators.**

I. Introduction

Quality management in healthcare services relies on documented processes, reliable records, and indicators capable of supporting decision-making. Continuous quality improvement, when applied to healthcare systems, requires structured models, monitoring tools, and the systematic analysis of care-related and support processes [1]. In this context, data availability alone does not guarantee improvement, since decision-making depends on the quality, completeness, and standardization of the available records [2].

In Brazil, RDC 63/2011 establishes good practice requirements for the operation of healthcare services, emphasizing the importance of qualified management and the reduction and control of risks to users [3]. In addition, RDC 509/2021 establishes requirements for the management of health technologies in healthcare facilities, reinforcing the need for structured processes throughout the technological life cycle [4]. More recently, the publication of the technical standard ABNT NBR ISO 7101:2025 reinforced the importance of quality management systems in healthcare organizations, oriented toward the delivery of timely, safe, effective, efficient, equitable, and people-centered care [5]. Furthermore, regulatory and standards-based frameworks establish criteria and guidelines for health technology management, with emphasis on quality, safety, and performance throughout the technological life cycle [4], [6].

In healthcare settings, Computerized Maintenance Management Systems (CMMS) can support the recording of work orders and maintenance histories by documenting failure causes, response times, costs, responsible technical personnel, execution status, and operational indicators [7]. These systems may contribute to equipment traceability, monitoring of corrective and preventive maintenance activities, resource planning, and performance assessment in Clinical Engineering services.

However, the mere existence of a computerized system does not ensure the generation of actionable managerial information. A CMMS may record a large volume of data and still present limitations for decision-making if its fields are incomplete or inconsistently completed. Therefore, the managerial usefulness of the tool depends on the quality of the data entered and on the ability to transform operational records into reliable indicators capable of adding value to the management process.

The publication of ABNT NBR ISO 7101:2025 expanded the possibilities and requirements applicable to management systems. Considering the relevance of corrective maintenance in the routine of healthcare facilities, records associated with these activities may represent an important source of information for assessing and improving management processes. Fields such as work order opening date (Abertura) and work order closing date (Fechamento) may allow, for example,

the calculation of the administrative cycle time of a work order. However, they do not necessarily indicate the actual downtime of the equipment. For this purpose, specific records of downtime (Parada) and return to operation (Funcionamento) would be required. Similarly, fields such as cause (Causa), occurrence (Ocorrência), cost (Custo), evaluation (Avaliação), and release for clinical use (Liberado para Uso) may exist in the database, but their analytical usefulness depends on the completeness and quality of the recorded data.

Thus, this study aimed to evaluate the quality of data recorded in corrective work orders extracted from a hospital CMMS used in a multicenter network, considering completeness, uniqueness, temporal consistency, informativeness, and the ability to support the generation of indicators in Clinical Engineering.

II. Materials and Methods

The study material consisted of a database of closed work orders extracted from a computerized hospital maintenance management system used in a multicenter hospital network. Data processing, organization, and analysis were performed using the R programming language, version 4.5.0 (R Foundation for Statistical Computing, Vienna, Austria), with support from the RStudio Desktop integrated development environment, version 2024.12.1.563 (Posit Software, PBC, Boston, MA, USA). The tidyverse, readxl, janitor, lubridate, xml2, writexl, and scales packages were used for data import, cleaning, transformation, analysis, and export [8].

A. Study Design and Data Source

An observational, descriptive, and quantitative study was conducted using a pre-extracted database of closed work orders obtained from the CMMS of a multicenter healthcare network. The dataset was provided to the authors as a spreadsheet export, with no intervention in care processes and no direct data collection from professionals or patients. The analytical sample corresponded to all records contained in the main work-order tab at the time the file was received. No row-level exclusions were applied to define the analytical sample. The original database query and any filtering procedures performed before delivery of the export were not available for independent verification.

The original spreadsheet contained four tabs: “Análise Fornecedores”, “Chamados Fechados”, “Chamados Corretiva Tempo Exec.”, and “Unidades”. The “Chamados Fechados” tab was used as the primary analytical dataset and contained 82,431 closed work orders and 64 original fields. The remaining tabs were used only for structural and consistency checks of the export and for auxiliary characterization and were not incorporated into the main analytical dataset.

The analysis was conducted at the network level. Individual healthcare facilities were not identified or compared, and unit-level characteristics such as care complexity, local versus centralized Clinical Engineering workflows, and installed-base criticality were not used for institutional stratification. The available organizational fields were therefore not interpreted as direct measures of healthcare-facility performance. The maintenance-type field classified the records into three CMMS categories: “corretiva”, “corretiva pro ativa”, and “corretiva de backup”. For the purposes of this study, these three system-defined categories were treated as corrective maintenance. No records classified in the analytical dataset as preventive maintenance, calibration, or electrical safety testing were identified. Accordingly, the scope of the study was restricted to closed corrective work orders.

The unit of analysis was the closed work order. For each record, administrative, operational, and managerial fields were evaluated, including work order number (OS), work order serial code (Código Serial OS), company (Empresa), maintenance shop (Oficina), maintenance type (Tipo de Manutenção), priority (Prioridade), work order status (Situação da OS), equipment (Equipamento), model (Modelo), manufacturer (Fabricante), department (Setor), cost center (Centro de Custo), opening date, closing date, occurrence, cause, work order notes (Observação), cost, asset tag (Patrimônio), evaluation, downtime, return to operation, and release for clinical use.

B. Data Quality Assessment and Statistical Analysis

Data quality assessment was structured into five main dimensions: completeness, uniqueness, temporal consistency, informativeness, and ability to generate managerial indicators. These dimensions were operationalized for CMMS work-order data considering general concepts used in secondary healthcare data quality assessment, including data presence, completeness, plausibility, and fitness for the intended use [9].

Completeness was assessed by calculating the absolute and relative frequencies of missing values for each variable. A value was considered missing only when it was null, an empty string, or contained only whitespace. Values that were formally present but carried limited analytical information were assessed separately under informativeness and were not counted as missing. For core work-order fields, completeness was used to assess the structural support provided by the database for operational characterization. For managerial fields, completeness was interpreted in relation to the feasibility of generating management indicators.

Uniqueness was assessed by checking duplicate values in the main work-order identifiers, particularly the work order number and the work order serial code. Temporal consistency was assessed from the opening and closing dates by identifying missing values and records outside a logical chronological sequence. Cross-field temporal plausibility was additionally evaluated for the downtime and return-to-operation fields by checking whether downtime preceded work-order opening, whether return to operation occurred after work-order closing, and whether return to operation preceded the recorded downtime.

Informativeness was evaluated separately from completeness. After normalization of case, accents, and excess whitespace, standardized entries corresponding to “Nao informado”, “Sem informacao”, “Nao se aplica”, “N/A”, “NA”, “NULL”,

"NULO", "Indefinido", and "Ignorado" were classified as non-informative. Zero values in the cost field were not considered missing but were analyzed separately because their economic interpretation could not be established from the database alone. Evaluation-note entries containing three or fewer characters were flagged as very short text entries for descriptive assessment. These classifications were performed using predefined computational rules rather than independent manual record-by-record judgment.

Semantic quality was assessed through the distribution and concentration of cause and occurrence categories. Absolute and relative frequencies were calculated for the most frequent categories. For normalization of compound entries, comma-separated terms were separated, whitespace was standardized, repeated terms were removed after text normalization, and the remaining terms were recombined while preserving their original categorical meaning.

To explore heterogeneity in data quality, missingness in the downtime, return-to-operation, and user-evaluation fields, as well as non-informative release-for-clinical-use entries and zero-cost records, were additionally stratified by year of work-order opening, work-order priority, and equipment type. For equipment-level interpretation, emphasis was placed on the 20 equipment categories with the highest number of work orders to reduce the influence of sparse categories. The stratified analysis remained descriptive and was intended to identify variation in completeness and informativeness across operational strata rather than to compare institutional performance.

Descriptive analyses were presented using absolute frequencies, percentages, means, medians, percentiles, and extreme values, according to the nature of each variable. Graphs were generated in the R environment and exported as image files for subsequent insertion into the manuscript.

III. Results and Discussion

A. General Analysis of the Database

The analyzed database consisted of 82,431 closed work orders and 64 original variables. The records had opening dates ranging from April 15, 2022, to June 4, 2025. The temporal analysis showed a higher density of records from 2024 onward, a finding that should be interpreted with caution because it may reflect implementation, migration, or consolidation of system use rather than an actual increase in maintenance demand.

The main identifiers showed good structural consistency. The work order number was filled in all records and showed no duplicates. A similar result was observed for the work order serial code, also with 82,431 unique values. At the identifier level, each row corresponded to a distinct work-order record, supporting the use of the closed work order as the unit of analysis.

The administrative opening and closing dates were filled in all records. The median time between opening and closing was 0.36 day, whereas the mean was 10.03 days. This difference between mean and median indicates an asymmetric distribution, with a large number of work orders closed within shorter intervals and a small set of records with prolonged duration. Fig. 1 shows the distribution of opening-to-closing time on a logarithmic scale for improved visualization.

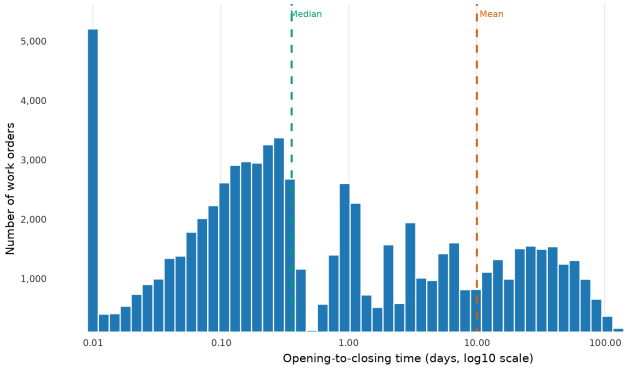


Fig. 1. Distribution of opening-to-closing time for corrective work orders on a logarithmic scale. Dashed lines indicate the median and mean values. Zero-day intervals were displayed as 0.01 day only for logarithmic visualization.

The results summarized in Table I indicate that the database has good structural quality for the operational recording of work orders. The absence of duplicate values in the main identifiers and the completeness of administrative dates allow descriptive analyses of volume, period, and administrative cycle time. Fig. 2 presents the five equipment types with the highest number of corrective work orders. However, this structural quality should not be confused with full capacity to generate advanced managerial indicators.

TABLE I
General Characterization of the Analyzed Service Order Database

Indicator	**Result**
Analyzed service orders	82,431
Original variables in the main tab	64
Opening period	15/04/2022 to 04/06/2025
Work order number filled	82,431
Duplicated work order	0
Duplicated work order serial code	0
Opening filled	100%
Closing filled	100%
Median opening-to-closing time	0.36 days
Average opening-to-closing time	10.03 days

B. Completeness of Core and Managerial Fields

The completeness assessment showed an important difference between core administrative fields and critical managerial fields. Variables such as work order number, work order serial code, company, maintenance shop, maintenance type, priority, work order status, department, cost center, opening date, closing date, occurrence, and cause showed high completeness. This result indicates that the CMMS is used consistently for the administrative and operational recording of work orders.

However, when fields required for management were evaluated, relevant gaps were observed. The overall percentage of empty cells in the main database was approximately 27.8%,

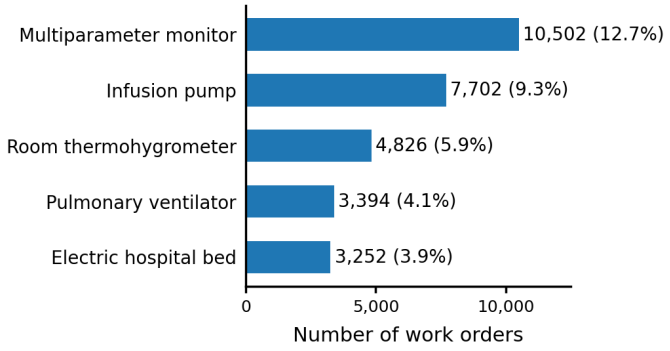


Fig. 2. Equipment types with the highest number of corrective work orders.

but this value should not be interpreted as a uniform property of the database because missingness varied substantially among variables. Missingness was particularly high in the closing-deadline and closing-justification fields and also affected evaluation notes, downtime, return to operation, and external service. Fig. 3 presents, in decreasing order, the percentage of missing records in critical managerial fields. Table II summarizes the managerial interpretation associated with the absence of data in these fields.

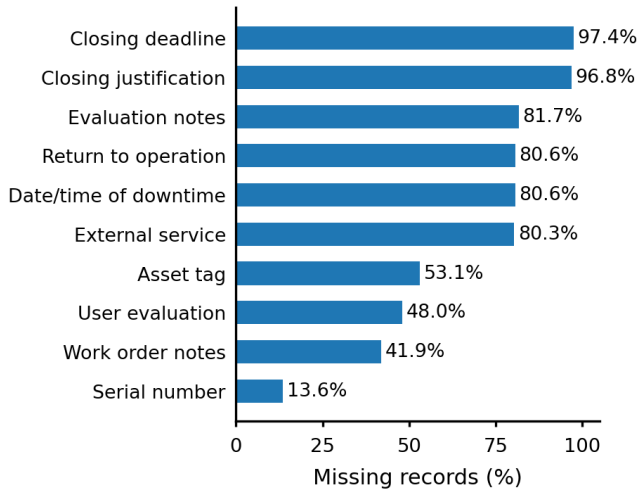


Fig. 3. Percentage of missing records in critical managerial fields of the work order database.

The stratified analysis showed that the high missingness of downtime and return-to-operation data was not restricted to a single work-order priority or period. Missingness in both fields was 81.1% for high-priority work orders, 79.8% for medium-priority work orders, and 85.6% for low-priority work orders. In 2024 and 2025, which together accounted for most of the analyzed records, the corresponding percentages were 79.9% and 82.0%, respectively. Greater variation was observed across the 20 equipment categories with the highest work-order volumes, for which missingness in both fields ranged from 71.0% to 92.5%. These results indicate that the limitation was present across the main operational strata evaluated,

TABLE II
COMPLETENESS OF SELECTED FIELDS FROM THE MAIN DATABASE

Variables	Missing	Interpretation
Serial Number	13.6%	Moderate loss for technical tracking
SO Notes	41.9%	Limited for qualitative text analysis
User evaluation	48.0%	Limited for perceived quality analysis
Asset tag	53.1%	Limited for individual tracking
External service	80.3%	Limited for third-party analysis
Date downtime	80.6%	Limited for unavailability calculation
Return operation	80.6%	Limited for usage return calculation
Evaluation notes	81.7%	Low availability for qualitative analysis
Closing justification	96.8%	Negligible for management analysis
Closing deadline	97.4%	Negligible for management analysis

although its magnitude varied among equipment categories. User-evaluation missingness also varied across the 20 most frequent equipment categories, ranging from 19.6% to 63.9%.

This finding does not mean that the database is unsuitable for analysis. Rather, it shows that the system is capable of recording work orders in a structured manner while also indicating that its use as a managerial intelligence tool depends on the completion of variables that go beyond the simple opening and closing of a work order.

C. Fields Formally Completed but with Limited Informational Value

Data quality analysis should not be limited to checking for empty fields. Some fields may be formally completed and still have limited analytical usefulness. This issue was observed particularly in the cost and release-for-clinical-use fields and, among completed entries, in evaluation notes.

The cost field was completed in all records. However, 75,968 work orders, corresponding to approximately 92.2% of the database, had a cost equal to zero. Only 6,463 records, equivalent to 7.8%, had a positive cost. Although the field showed formal completeness, its use for comprehensive cost analysis should be approached with caution. Without additional validation, it is not possible to determine whether zero values indicate an actual absence of cost, lack of cost entry, internally absorbed cost, non-applicable cost, or another operational criterion used by the institution.

From a managerial perspective, the predominance of zero values represents a loss of opportunity for maintenance cost monitoring, life-cycle cost assessment, replacement planning, and economic prioritization of maintenance actions. Nevertheless, this result should not be interpreted as evidence of absence of cost or as a management error without additional validation of the institution's costing rules and data entry practices.

The release-for-clinical-use field showed a similar pattern. Although formally completed, 98.3% of the records were classified as having limited informational value. This result limits the use of the field to assess the safe return of equipment to clinical use after maintenance.

The percentage of records affected by fields that were formally completed but had limited informational value is

shown in Fig. 4. The highest percentages were related to the release-for-clinical-use field with limited informational value and to zero cost. In addition, 12.0% of all records had very short evaluation notes. Considering that only 18.3% of records contained any evaluation-note information, very short entries represented a substantial proportion of the completed notes. This result should be interpreted together with the low completeness of this field, as shown in Fig. 3.

The stratified analysis also showed that limited informativeness was not confined to a single priority group. Non-informative release-for-clinical-use entries ranged from 97.8% to 98.4% across priority levels, while zero-cost records ranged from 91.9% to 96.1%. Greater heterogeneity was observed among the 20 most frequent equipment categories: non-informative release-for-clinical-use entries ranged from 95.6% to 99.6%, whereas zero-cost records ranged from 67.0% to 99.8%.

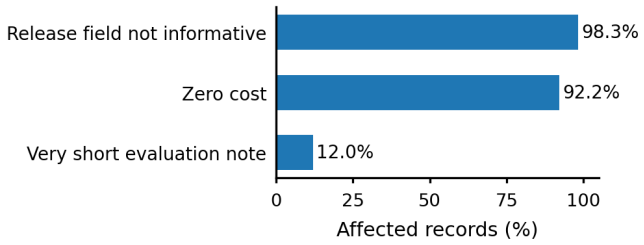


Fig. 4. Percentage of records with selected patterns of limited or uncertain managerial interpretability.

In addition to the fields shown in Fig. 4, the analysis of free-text fields indicated other limitations. The work order notes field was completed in 58.1% of the records, with a median of eight words. The service field was completed in 100% of records but had a median of only one word per completed record, indicating that descriptions were usually very short. The median was used because it is less sensitive to extreme values, since a small number of long texts could increase the mean without representing the predominant data entry pattern.

These results reinforce the need to distinguish formal completeness from analytical usefulness. The presence of a value in a given cell does not necessarily guarantee that the recorded information is sufficient to build reliable indicators. Therefore, data quality assessment should consider not only missing values but also the ability of the recorded content to support audits, managerial analyses for decision-making, and continuous process improvement.

D. Administrative Work Order Time Versus Actual Downtime

As shown in III-A, the database allows analysis of the administrative time between work order opening and closing. This interval was available for all records and showed no chronological inconsistency. This interval was not considered Time to Repair (TTR), because it does not necessarily represent the effective repair time or the actual equipment downtime. Therefore, the time between work order opening and closing should be interpreted as the distribution of the administrative work order cycle, not as the distribution of actual equipment downtime. This distinction is important for Clinical Engineering because opening and closing a work order may be influenced by administrative factors, spare-part availability, vendor availability, delays in data entry, process follow-up, or closure after the equipment has already returned to use.

To assess actual downtime, specific data on downtime and return to operation would be required. The latter was interpreted as the operational return of the equipment. These two fields showed identical availability patterns: both were completed in 15,991 records (19.4%) and both were missing in the remaining 66,440 records (80.6%), as shown in Fig. 3. No record had only one of the two fields completed.

Among the records in which the downtime and return-to-operation fields were completed, 863 cases had a downtime date earlier than the work order opening date. This condition may be operationally plausible, since the equipment may have been identified as unavailable by the care unit before the maintenance team formally opened the work order. In 326 cases, the return-to-operation date was later than the work order closing date, which may indicate a difference between administrative closure of the work order and the recorded operational return of the equipment. Among the completed records, there were no cases in which the return-to-operation date was earlier than the downtime date, indicating temporal coherence between these two fields.

These results show that the database adequately supports measurement of the administrative work order cycle, but it does not allow actual equipment downtime to be calculated for all records. Therefore, indicators such as Mean Time to Repair (MTTR) and operational availability should be considered weak when calculated exclusively from the fields available in the analyzed database.

E. Semantic Quality of Causes and Occurrences

The cause and occurrence fields were fully completed in the analyzed database. This is positive from a completeness standpoint. However, the semantic quality of these fields requires evaluation, since formal completeness does not guarantee sufficient granularity for more robust analysis.

After simple normalization, the most frequent causes were natural wear, equipment failure, damaged accessory, and operational error. These categories accounted for approximately 80.1% of the work orders. Fig. 5 shows the distribution of the most frequent causes.

The concentration of records in a few categories may be useful for Pareto analyses, as it allows the main groups of failures or demands to be identified. However, some categories are broad in nature. For example, the expression equipment failure indicates that a failure occurred, but it does not necessarily describe the mechanism involved, the affected component, or the root cause. Similarly, natural wear may represent an expected condition in some cases, but it may also include different situations that would require greater detail to

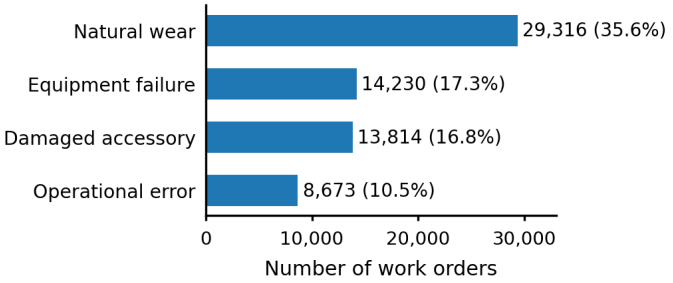


Fig. 5. Distribution of the four most frequent causes recorded in corrective work orders.

distinguish service life, usage pattern, preventive maintenance, and accessory condition.

In the occurrence field, the most frequent categories were third-party service, damaged accessory, malfunction, and alarming. These categories accounted for 64.5% of all occurrences. Fig. 6 shows the most frequent occurrences.

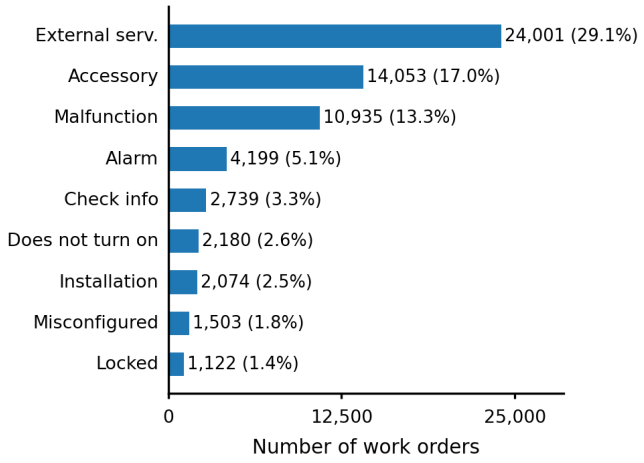


Fig. 6. Most frequent occurrences recorded in corrective work orders.

The predominance of third-party service as an occurrence suggests that the field may be used not only to record the technical symptom or observed problem but also to describe the operational workflow or routing of the work order. This mixture of different types of information in the same field limits more refined analyses, since it combines technical event, request, administrative condition, and operational routing.

Accordingly, the results indicate that cause and occurrence are complete fields, but they still require greater semantic standardization to support root-cause analysis, preventive actions, and evidence-based decision-making. One possible improvement would be to conceptually separate fields intended for observed symptom, probable cause, root cause, action performed, and need for third-party support.

F. Implications for Data Governance in Clinical Engineering

Based on the results obtained, the analyzed CMMS provides adequate structural support for the operational management of work orders. The database contains unique main identifiers, complete administrative opening and closing dates, and structured information that allows analyses of work-order volume, temporal distribution, equipment, departments, causes, occurrences, and cost centers. This operational structure, however, does not imply that all fields required for more advanced managerial indicators have the same level of completeness or informational value.

This distinction is particularly evident for downtime and return-to-operation data. These fields were available in only 19.4% of the work orders, and their missingness remained high across the main priorities and periods evaluated. Greater variation was observed among equipment categories. Therefore, estimates of downtime, MTTR, or related indicators calculated only from completed records may not adequately represent the entire set of work orders. The magnitude and direction of this possible distortion cannot be determined from the available database because the true values for records with missing information are unknown.

From a data-governance perspective, these results suggest that improving CMMS use depends not only on increasing data entry, but also on defining which information is required for each managerial purpose. Previous experience with CMMS database management in Clinical Engineering has also emphasized the need for data assessment, cleaning, validation, and organization before managerial use [10]. Fields such as downtime, return to operation, asset identification, and release for clinical use may require validation at work-order closure when they are applicable to the maintenance process. Controlled vocabularies could improve the consistency of cause and occurrence categories, while automated cross-field checks could identify implausible temporal relationships. Periodic dashboards could also monitor missing or low-information records over time and across relevant operational strata.

The dimensions considered in this study were applied to the structure and intended managerial use of the analyzed CMMS. They overlap with concepts used in broader healthcare data-quality frameworks, including completeness, consistency or concordance, plausibility or validity, uniqueness, timeliness, and accuracy [2], [9]. In the present study, these concepts were adapted to maintenance records, with additional emphasis on informativeness and fitness for managerial use. This suggests that the general assessment logic may be applicable to other CMMS environments, although the specific fields, validation rules, and thresholds would need to be adapted to the structure and operational workflow of each system.

Training, periodic review of field definitions, and monitoring of data-quality indicators may complement these controls. At the same time, making a large number of fields mandatory without considering their operational relevance may increase documentation burden without necessarily improving the analytical usefulness of the database.

Future analyses could extend the stratification to equipment criticality, department, cost center, unit profile, and institutional workflow when such descriptors are reliably available. This may help determine whether missing or low-information

records are concentrated in specific operational contexts or reflect broader network-wide data-entry patterns.

Table III summarizes the main managerial implications of the results. The analyzed database provides good support for basic operational indicators but presents relevant limitations for advanced indicators, particularly those dependent on downtime, cost, release for clinical use, and user evaluation.

TABLE III
COMPUTABILITY OF SELECTED MANAGERIAL INDICATORS

Indicator and rationale	Class.
Work order volume by period Work order ID and opening date were complete.	Computable
Opening to closing time Opening and closing dates were complete and temporally consistent.	Computable
Administrative SLA Depends on the institutional SLA rule and interpretation of related fields.	Caution
Equipment ranking by calls Equipment, model, and manufacturer fields were mostly available.	Computable
Real downtime Downtime and return to operation were missing in 80.6% of records.	Weak
MTTR Depends on downtime and return to operation, which were poorly completed.	Weak
Operational availability Depends on real downtime and a validated installed base.	Weak
Cost per work order or equipment Zero cost was recorded in 92.2% of records.	Weak
User perceived quality User evaluation and evaluation notes were incomplete or poorly informative.	Partial
Safe return to use The field was mostly filled with non-informative content.	Weak

IV. CONCLUSION

The analyzed database showed adequate structural quality for the operational recording of corrective work orders. The 82,431 records had unique main identifiers and complete administrative opening and closing dates, supporting analyses of work-order volume, temporal distribution, equipment, departments, causes, occurrences, and cost centers. These characteristics support the use of the CMMS for basic operational analyses, but they do not by themselves ensure that more advanced managerial indicators can be calculated reliably.

Relevant limitations remained in fields required for managerial interpretation. Downtime and return-to-operation data were missing in 80.6% of the records, and this high missingness persisted across the main priorities and periods evaluated, although its magnitude varied among equipment categories. Cost and release-for-clinical-use fields were formally completed but had limited interpretability, with 92.2% of cost records equal to zero and 98.3% of release-for-clinical-use records classified as non-informative. Cause and occurrence fields were complete, but broad or mixed categories limited more detailed root-cause analysis.

Because the analytical dataset contained only corrective work orders and no institutional or unit-profile stratification was performed, the results should be interpreted as an overall assessment of data quality within the analyzed network.

The results indicate that the managerial value of a CMMS depends not only on record availability, but also on data governance, field definitions, appropriate validation rules, category standardization, user training, and continuous monitoring of data-entry quality.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001 and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (309464/2022-6 to HTM).

REFERENCES

[1] A. Endalamaw et al., "A scoping review of continuous quality improvement in healthcare system: conceptualization, models and tools, barriers and facilitators, and impact," *BMC Health Services Research*, vol. 24, Art. no. 487, 2024, doi: 10.1186/s12913-024-10828-0.

[2] A. Lighterness, M. Adcock, L. A. Scanlon, and G. Price, "Data quality-driven improvement in health care: systematic literature review," *Journal of Medical Internet Research*, vol. 26, Art. no. e57615, 2024, doi: 10.2196/57615.

[3] Agência Nacional de Vigilância Sanitária, *Resolução da Diretoria Colegiada RDC n. 63, de 25 de novembro de 2011: Dispõe sobre os requisitos de boas práticas de funcionamento para os serviços de saúde*. Brasília, Brasil: ANVISA, 2011.

[4] Agência Nacional de Vigilância Sanitária, *Resolução da Diretoria Colegiada RDC n. 509, de 27 de maio de 2021: Dispõe sobre o gerenciamento de tecnologias em saúde em estabelecimentos de saúde*. Brasília, Brasil: ANVISA, 2021.

[5] Associação Brasileira de Normas Técnicas, *ABNT NBR ISO 7101: Gestão de organizações de cuidados à saúde, sistemas de gestão da qualidade em organizações de cuidados à saúde, requisitos*. Rio de Janeiro, Brasil: ABNT, 2025.

[6] Associação Brasileira de Normas Técnicas, *ABNT NBR 15943: Diretrizes para um programa de gerenciamento de equipamentos de infraestrutura de serviços de saúde e de equipamentos para a saúde*. Rio de Janeiro, Brasil: ABNT, 2011.

[7] T. Cohen, "The basics of CMMS," *Biomedical Instrumentation & Technology*, vol. 48, no. 2, pp. 117–121, Mar./Apr. 2014.

[8] H. Wickham and G. Grolemund, *R para Data Science*. Rio de Janeiro, Brasil: Alta Books, 2019.

[9] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, Jan. 2013, doi: 10.1136/amiajnl-2011-000681.

[10] R. Panichelli, "CMMS database migration in clinical engineering: Penn Medicine's experience," *Biomedical Instrumentation & Technology*, vol. 56, no. 2, pp. 41–45, Apr. 2022, doi: 10.2345/1943-5967-56.2.41.

Design and Deployment of Interactive Standard Operating Procedures for CMMS Customization: A Case Study at a University Hospital

1st Sofia K. Ho
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0003-9495-8411

2nd Diego A. O. Rosa
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0004-2422-8453

3rd Danilo M. do Nascimento
University Hospital (HU)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-6885-3782

4th Paulo M. Zanuzzio
University Hospital (HU)
University of São Paulo (USP)
São Paulo, Brazil
0009-0000-6740-5166

5th Edison Silva Filho
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-5397-088X

6th Henrique T. Moriya
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0000-0002-1595-1779

***Abstract*—This study describes the development of interactive Standard Operating Procedures (SOPs) and educational materials using H5P technology to support the implementation and promote standardized data-entry practices of a Computerized Maintenance Management System (CMMS) at the University Hospital of the University of São Paulo. The challenges encountered mainly involved the complexity and robustness of the CMMS and its non-standardized use by the technical team, which hindered reliable data analysis. To address this, flowcharts for critical procedures, such as equipment decommissioning and service requests, were created and integrated with custom video tutorials using an interactive educational content development software. These materials were exported as standalone HTML files to ensure offline access within the hospital environment, overcoming connectivity limitations. The resulting digital repository contains interactive SOPs and visual tutorials, organized by a centralized index that allows the maintenance team to access standardized instructions quickly. In the future, the use of interactive H5P tools can serve as both pedagogical support and a quality control mechanism for the CMMS database. However, the effect on data quality was not quantitatively evaluated. The framework employed may be replicable in other hospital CMMS implementations seeking to standardize local workflows.**

***Index Terms*—CMMS, Clinical Engineering, Standard Operating Procedures, H5P, Equipment Maintenance**

I. Introduction

The management of medical equipment represents one of the most critical dimensions of hospital administration. In the healthcare environment, ensuring that diagnostic and therapeutic devices operate reliably and safely is a prerequisite for quality patient care. According to the Brazilian Ministry of Health, the implementation of structured maintenance programs is fundamental to reduce accident risks, ensure device availability, and guarantee the effectiveness of clinical procedures [1].

Within the framework of hospital quality systems, performance indicators play a central role in guiding management decisions. Indicators such as equipment downtime rates, mean time between failures (MTBF), and mean time to repair (MTTR) allow clinical engineering teams to identify bottlenecks, allocate resources more efficiently, and demonstrate accountability to hospital administration. However, the validity of such indicators depends entirely on the quality and consistency of data recorded by maintenance technicians in every service interaction. When data entry is inconsistent or incomplete, the entire analytical framework built upon it loses its reliability, ultimately undermining the hospital's quality management system.

To support this data-driven approach, Computerized Maintenance Management Systems (CMMS) were developed. A CMMS is a comprehensive software platform that automates the full lifecycle of maintenance management and centralizes records [2].

Despite their benefits, modern CMMS platforms are complex systems with numerous data entry fields, configurable workflows, and interrelated modules, demanding a considerable level of training and familiarity from the technical staff. New users, and even experienced technicians adapting to a new system, face a steep learning curve that, if not adequately supported, can lead to non-standardized data entry and degraded database integrity over time [3].

A critical gap in current health technology management stems from commercial training resources: vendor-provided CMMS manuals and tutorials are inherently generic, covering general system functionality rather than the specific workflows adopted by individual institutions. As a result, they may include procedures that do not apply, omit steps that are locally relevant, or present sequences that differ from the actual

practice of the team. This gap between generic documentation and institutional reality creates a persistent training deficit that static manuals alone cannot bridge.

To bridge this gap, Standard Operating Procedures (SOPs), which have long been recognized as essential instruments for ensuring process consistency in regulated environments, can be combined with interactive digital formats, engaging users more effectively than traditional documents, facilitating knowledge retention and real-time consultation.

In this context, the open-source H5P (HTML5 Package) framework [4] emerges as a particularly suitable technology. H5P is an open-source framework that enables the creation of richly interactive content directly in the browser, without requiring specialized programming skills. Using tools such as an interactive content authoring software (ICAS) (Lumi Education UG, Germany) [5], H5P activities can be authored through an intuitive interface and exported in self-contained formats compatible with offline use.

The primary purpose of this work is to describe the development and implementation of interactive SOPs and educational materials using H5P technology to support the adoption of a specific CMMS at the University Hospital of the University of São Paulo (HU-USP) (Effort, GlobalThings, Brazil) [6]. The approach was designed with the usability needs of the Healthcare Technology Management (HTM) team in mind, prioritizing accessibility, contextual relevance, and independence from internet connectivity.

This study describes the development and implementation of institutional educational materials as part of standard quality improvement processes. All collaboration with the maintenance team occurred through routine institutional engagement. Although formal ethics approval was not sought, as this constitutes quality improvement rather than research, the technical team members were aware of and voluntarily participated in all development and validation activities.

II. Materials and Methods

The development process was guided by a user-centered design philosophy, placing the practical needs and daily workflow of the HTM team at the forefront of every decision. The methodology unfolded across several interconnected stages: observational assessment, procedural mapping, content production, iterative validation, and technical deployment.

A. Observational Assessment and Team Interaction

The first stage consisted of an observation period of two months, during which the actual use of the CMMS by the HTM team was monitored. This phase involved accompanying technicians during their routine activities, identifying the specific steps they performed in the system, and noting points of uncertainty, recurring errors, or inconsistencies in data entry.

Beyond simple observation, this stage also included informal consultations with team members, documented through field notes, to understand their prior experience with digital maintenance systems, their familiarity with the CMMS platform, and the difficulties they most frequently encountered.

These interactions were instrumental in establishing a shared understanding between the research team and the maintenance staff. Consequently, it became clear which procedures should be prioritized to establish a foundation for SOP development.

Two fundamental procedures were selected to be developed in detail as proof-of-concept: service request intake and equipment decommissioning. Notably, these processes do not possess complex, conditional steps or numerous decision points; instead, they are inherently simple. They were chosen because they effectively outline the boundaries of the maintenance workflow, acting as the logical initiation and a potential termination of a work order.

B. Procedural Mapping and Flowchart Development

Based on the insights gathered during the observational phase, detailed procedural flowcharts were developed collaboratively through discussions with two technical leads. Using a graph-diagramming tool (yEd, yWorks, Germany), the flowcharts were designed to represent not only the linear sequence of steps, but also the conditional branches and decision points inherent in each process. Special care was taken to ensure that the diagrams reflected the actual practice of the HU-USP HTM team, rather than a generic or idealized version of the procedure.

The flowcharts were subsequently reviewed in alignment meetings with the maintenance team, where all seven technicians commented and suggested modifications. These meetings proved particularly valuable for capturing tacit knowledge, procedural nuances that technicians perform automatically, but are rarely documented in official manuals.

Three rounds of revision were carried out based on team feedback. They resulted in changes such as the order and the terminology of the workflow, ensuring that the final flowcharts accurately captured the approved institutional workflow before any educational material was produced.

C. Video Recording and Content Production

With validated flowcharts in hand, the next stage involved producing custom tutorial videos tailored to the specific procedures adopted at HU-USP. As one of the experienced technicians executed each procedural step within the CMMS, the screen was captured using an open-source screen recording software (OBS Studio, OBS Project, USA). This approach was intended to produce content that was both authentic and directly applicable to the team's context, differing from the generic tutorials provided by the CMMS vendor.

The raw video recordings were subsequently edited using a video editing tool (Clipchamp, Microsoft, USA) to remove redundant segments, adjust pacing, and ensure that each clip corresponded precisely to a single step of the procedural flowchart. This segmentation was a deliberate pedagogical choice: by presenting each action in isolation, the tutorials allow the user to consult specific steps without navigating through lengthy recordings.

D. H5P Development and SOP Documentation

The edited video segments were then integrated into interactive H5P activities using the Lumi software. The "image hotspots" content type was selected as the primary format, as it allows users to interact with a visual representation of the procedural flowchart by clicking on individual steps to reveal the corresponding video tutorial.

This design directly mirrors the cognitive model of a technician consulting the SOP during task execution: they identify where they are in the process, access the relevant instruction, and proceed.

Concurrently, written SOPs were produced for each procedure. These documents provided a detailed narrative description of every step depicted in the flowchart, complemented by annotated screenshots extracted from the tutorial videos.

The combination of textual description and visual reference was intended to accommodate different learning preferences within the team and to serve as a printable fallback in situations where digital access might be unavailable.

E. Offline Deployment and Index Development

A critical constraint of the hospital environment is the limited or inconsistent internet connectivity in certain areas where maintenance activities are performed. To ensure that the educational materials could be consulted by technicians at the point of care, regardless of network availability, a technical solution was developed to package and distribute the H5P content as self-contained offline files.

Using the ICAS, each completed H5P activity was exported via the path: File → Export → All-in-one HTML file → Export Now. This process generated standalone HTML files capable of being executed directly in any browser without requiring an internet connection.

The resulting files were organized within a dedicated folder ("*fluxogramas*"), alongside the corresponding SOP documents (stored in a "*documentos*" folder).

To facilitate intuitive navigation across the growing collection of materials, a custom index site was coded in HTML. This interface presents all available procedures as interactive cards, organized alphabetically, with direct links to both the H5P interactive flowchart and the corresponding SOP PDF document.

Adding new procedures to the index requires only a straightforward editing step: duplicating the HTML card block (as illustrated in Fig. 1), adjusting the title, description, and file references, and saving the updated file.

This low-barrier update process was intentionally designed so that team members without advanced technical skills could independently maintain and expand the repository over time.

F. Final Validation

Upon completion of all materials, a final series of alignment meetings was held with the HU-USP HTM team. During these sessions, team members navigated the index interface, interacted with the H5P flowcharts, and reviewed the written SOPs. Feedback was collected regarding clarity, completeness, and usability, and minor adjustments were made accordingly.

```
<!-- Card 1 -->
<div class="card">
    <!-- Card title -->
    <h3>Service Request Intake</h3>
    <!-- Brief card description -->
    <p>Critical analysis of service calls made by the University
        Hospital to the Equipment Maintenance Section.</p>
    <div class="button-group">
        <!-- link to flowchart -->
        <a href="fluxogramas/RequisicaoDeServico.html"
        target="_blank" class="btn btn-h5p">View Flowchart</a>
        <!-- link to SOP -->
        <a href="documentos/POP_RequisicaoDeServico.pdf"
        target="_blank" class="btn btn-pdf">View SOP (PDF)</a>
    </div>
</div>
```

Fig. 1. Structural HTML block for adding new H5P content to the index.

This validation step aimed to ensure that the materials were not only technically accurate but also genuinely accessible and useful to the intended audience in their day-to-day activities. Feedback was collected verbally during alignment meetings and documented in field notes. Although no standardized usability instruments or systematic usability metrics were applied, the overall response from the team was positive.

III. Results

Applying the described methodology, a total of two H5P interactive flowcharts and two written SOPs were successfully developed and validated, covering the two foundational procedures selected for the HU-USP HTM sector.

The development process resulted in clear flowcharts that map the straightforward nature of each procedure, capturing the essential, linear workflows that technicians follow to initiate and conclude a work order in practice.

Fig. 2 presents the flowchart for service request intake, which encompasses the steps from initial service demand identification through the proper registration of the service order in the CMMS.

Fig. 3 presents the flowchart for equipment decommissioning, a procedure of particular importance for maintaining an accurate active equipment inventory within the CMMS.

Each H5P activity was built around its respective flowchart, enabling users to click on individual steps to access the corresponding custom video tutorial.

The videos, recorded specifically for the HU-USP context, present the exact sequence of actions performed within the CMMS as practiced by the institutional team, making them immediately applicable rather than requiring adaptation from generic instructional content.

To consolidate all produced materials and provide an accessible query interface for the maintenance team, the index site was developed (Fig. 4).

The interface not only organizes the available procedure cards in alphabetical order, but also offers a search bar, allowing users to quickly locate the desired procedure and

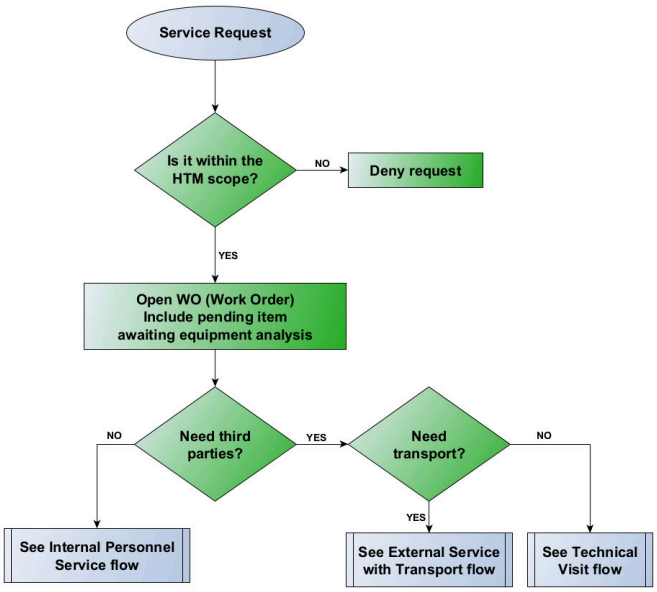


Fig. 2. Flowchart for service request intake.

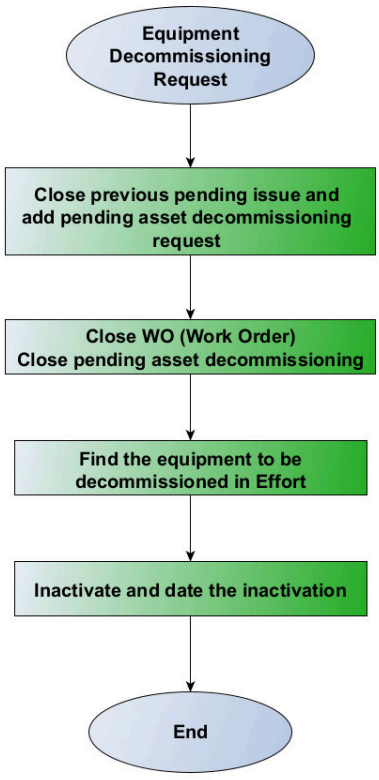


Fig. 3. Flowchart for equipment decommissioning.

open either the interactive H5P flowchart or the corresponding SOP document with a single click.

At the moment, the index is incomplete, with only the Equipment Decommissioning and the Service Request Intake procedures with both the interactive flowchart and SOP. However, all eight institutional CMMS procedures already have the flowchart (without the videos), providing a template for future expansion.

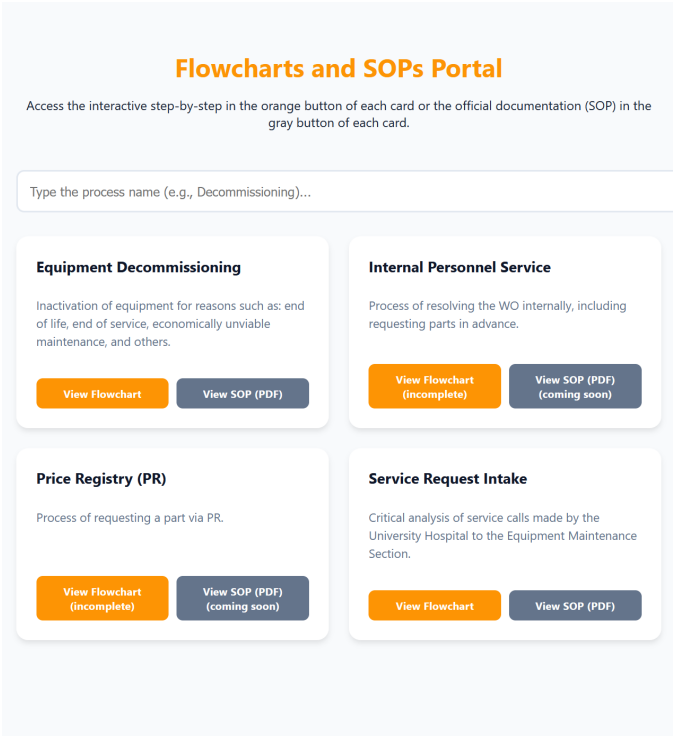


Fig. 4. Graphical interface of the developed index site.

The produced materials are available in Portuguese in the project's digital repository [7], organized as a compressed ZIP file containing the index, flowcharts, and SOP documents. For data protection and institutional privacy, the H5P videos in the shared repository have been replaced with example placeholder videos and the CMMS screenshots have been deleted from the SOPs. The full customized institutional versions are maintained locally within HU-USP systems.

Once the ZIP file is extracted, the entire collection is accessible by simply opening the index file in a browser, with no internet connection required.

This packaging strategy ensures that the full repository (including all interactive H5P content) can be distributed, stored locally, and consulted offline across different devices within the hospital environment.

Whenever new materials are added, updating the distributed folder ensures that all users have access to the most current version.

IV. Discussion

The Computerized Maintenance Management System represents, in contemporary clinical engineering, a management

tool of considerable scope and sophistication. Designed to integrate the full spectrum of equipment lifecycle activities into a single platform, a CMMS encompasses dozens of configurable fields, interrelated modules, and workflow dependencies that collectively demand a high level of familiarity from its users.

The process of adopting such a system in a hospital environment is not merely a technical transition, it is an organizational one, requiring that maintenance technicians internalize new data entry habits, adapt to structured workflows, and consistently apply procedures that may differ substantially from their previous informal practices.

This learning curve is well-documented in the literature and represents one of the primary barriers to successful CMMS implementation, particularly in public hospitals where training resources are often limited [2].

The experience at HU-USP highlighted this challenge. Through discussions with the HTM team, it was identified that the absence of clear, institutionally tailored guidelines contributed to variability in how maintenance records were entered. Different technicians interpreted the same fields differently, optional steps were inconsistently completed, and the classification of maintenance events lacked uniformity.

Over time, this variability accumulated into a database that, while voluminous, was insufficiently reliable for quantitative analysis, rendering performance indicators such as MTBF, MTTR, and preventive maintenance compliance rates either inaccurate or impossible to compute with confidence.

Most commercial CMMS vendors, including the provider of the HU-USP's platform, do supply their users with learning resources such as help menus, user manuals, and video tutorials. However, these materials are inherently designed for a general audience: they demonstrate system functionality in the abstract, without reference to the specific procedures, nomenclature, or institutional conventions of any particular hospital.

At HU-USP, this meant that available vendor tutorials contained steps not performed by the team, omitted institution-specific requirements, and followed sequences that diverged from locally approved workflows.

The generic nature of these resources, while understandable from a commercial perspective, makes them poorly suited as a sole training mechanism for teams with highly specific operational contexts.

By developing custom interactive materials anchored to the actual workflows of the HU-USP HTM team, this project aimed at addressing precisely that gap. The iterative process of observation, flowchart validation, and team alignment ensured that the produced SOPs and H5P tutorials reflected not a theoretical ideal, but the approved institutional reality.

Technicians consulting the materials encounter procedures that match their daily experience and expectations, which is expected to reduce the cognitive friction associated with adapting generic instructions to local practice.

The choice of H5P as the interactive content format brought additional pedagogical advantages beyond mere standardization. Interactive elements, such as the clickable flowchart hotspots linked to step-specific video demonstrations, were designed with the intent to promote active engagement rather than passive reading, supporting better knowledge retention. However, assessment of actual learning retention was outside the scope of this implementation study.

The visual and sequential structure of the materials also accommodates varying levels of prior CMMS experience within the team: newcomers can follow the tutorials step by step, while more experienced technicians can use the materials as a quick reference for less familiar procedures.

A critical technical challenge addressed during implementation was the limited internet connectivity available in several areas of the hospital where maintenance activities are performed. Exporting the H5P content as self-contained HTML files using the ICAS [5] provided an effective and low-cost solution to this constraint, enabling full offline access without requiring any server infrastructure or institutional IT investment. This portability makes the repository immediately deployable and maintainable by the team itself.

It is also worth noting that the optimization of tutorial video duration, achieved by recording procedure-specific clips rather than relying on the vendor's longer generic videos, had a direct impact on the practical viability of offline distribution. Shorter, targeted videos significantly reduced the total size of the compressed repository, making it feasible to store and share the complete collection via conventional digital media within the hospital.

One limitation of the current implementation relates to scalability: as the number of documented procedures grows, so does the volume of videos and PDF documents, and consequently the storage footprint of the standalone folder. This growth must be managed thoughtfully, with periodic reassessment of storage strategies and distribution mechanisms as the repository expands.

Technical limitations also include the lack of compatibility testing across different browsers (e.g., Chrome, Firefox, Safari) and operating systems (Windows, macOS, Linux), which is a critical requirement prior to broader deployment. Additionally, offline HTML distribution presents version control and content currency challenges, as the repository currently records only its last update date. To mitigate the security and reliability risks inherent to standalone files, future releases should incorporate: (1) explicit version numbering, (2) automated outdated-content warnings, (3) a centralized tracking system for active versions, (4) detailed change logs, (5) robust file integrity verification, (6) backup procedures, and (7) disaster recovery protocols. Finally, formal accessibility considerations were not incorporated; while video captions were unnecessary due to the absence of audio instructions, future iterations should include alternative text descriptions for all visual elements.

Another limitation of this study is that no logging or tracking of material utilization was performed, and no quantitative evaluation of CMMS data quality before and after implementation was conducted. Therefore, claims about improvements in database accuracy cannot be made. It can only be said that the implementation of this material was positively perceived by

the HU-USP's HTM team. Thus, comprehensive assessment of the impact of these materials on maintenance data quality and technician performance would require a more rigorous evaluation framework, including pre- and post-implementation data analysis and potentially a controlled comparison with alternative training approaches. Such evaluation is recommended as future work.

A critical aspect of long-term sustainability involves establishing explicit triggers for content review and updates. Specifically, whenever the CMMS interface undergoes significant updates (including new fields, modified workflows, or altered rules) or when institutional maintenance procedures are modified, the corresponding flowcharts and H5P tutorials should be reviewed and updated accordingly. This protocol ensures that the repository remains synchronized with actual institutional practice and continues to serve as an accurate reference over time.

Although this study was limited to a single institutional context and two detailed procedures, it is important to highlight the potential for broader applicability of the methodology developed in this work. The framework (observation, iterative procedural mapping, custom video production, H5P integration, offline packaging, and team validation) is not specific to the CMMS or to HU-USP. It is a potentially transferable framework that may be adapted to other settings; external validation is still required.

In contexts where inconsistent data entry compromises the quality of maintenance databases and, by extension, the reliability of clinical engineering performance indicators, this approach can offer a structured, low-cost pathway toward standardization.

The potential for application in other institutions suggests that this work, while grounded in a specific institutional experience, carries value beyond its original context, contributing to the broader goal of building robust, reliable health technology management systems across the Brazilian healthcare network.

V. Conclusion

This study describes a methodology for developing and implementing customized, interactive SOPs within institutional CMMS contexts. The framework integrating observational assessment, iterative flowchart development, custom video production, H5P-based interactive content, and offline deployment was designed to address known barriers to CMMS adoption in settings with limited connectivity.

While no quantitative evaluation of impact was performed in this work, the approach offers a template for other institutions seeking structured, low-cost solutions to standardize maintenance workflows.

Validation of the framework in another institutional context or department is recommended as future work to assess generalizability. Future work should also include systematic evaluation of adoption rates, database quality improvements, and maintenance performance indicators.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgment

The authors would like to extend their gratitude to the entire team of the HTM Section of the HU-USP for their invaluable collaboration and assistance throughout the development of this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001 and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (137972/2025-3 to SKH and 309464/2022-6 to HTM).

References

[1] Brazil Ministry of Health, Health Investment Management Secretariat, *Medical-hospital equipment and maintenance management*. Brasília: Ministry of Health, 2002. (Series F. Health Communication and Education).

[2] World Health Organization (WHO), *Computerized maintenance management system*. Geneva: WHO, 2011. (WHO Medical device technical series). [Online]. Available: https://apps.who.int/iris/handle/10665/44567

[3] T. Cohen, "The basics of CMMS," *Biomedical Instrumentation & Technology*, vol. 48, no. 2, pp. 117–121, Mar./Apr. 2014.

[4] H5P Group, "H5P: Create, share and reuse interactive HTML5 content," 2024. [Online]. Available: https://h5p.org/

[5] Lumi Education, "Lumi: desktop app for H5P," 2024. [Online]. Available: https://lumi.education/

[6] GlobalThings, "Effort: Sistema de Gestão de Manutenção (CMMS)," 2024.

[7] S. K. Ho, "Supplementary materials: interactive tutorials and technical documentation," 2026. Open Science Framework. [Online]. Available: https://osf.io/avh9c

Development of an Assessment Tool for Health Technology Management in Home Care

1st Marcella Wayss Darold
Biomedical Engineering Institute (IEB-UFSC)
Federal University of Santa Catarina (UFSC)
Florianópolis, Brazil
marcelladarold@gmail.com

2nd Jefferson Luiz Brum Marques
Biomedical Engineering Institute (IEB-UFSC)
Federal University of Santa Catarina (UFSC)
Florianópolis, Brazil
jlbmarques@gmail.com

3rd Renato Garcia Ojeda
Biomedical Engineering Institute (IEB-UFSC)
Federal University of Santa Catarina (UFSC)
Florianópolis, Brazil
rengar99@gmail.com

***Abstract*— The growing demand for home care programs and services has led to an increased incorporation of health technologies into the home environment, expanding care complexity and imposing new challenges related to technology safety and reliability, infrastructure adequacy, and user qualification. In this scenario, Clinical Engineering assumes a strategic role in managing these technologies throughout the technology lifecycle, from initial home assessment to continuous monitoring of home care conditions. This study proposes a home assessment tool for health technology management, grounded in the principles of Clinical Engineering, normative requirements, and the Medical-Hospital Technology Management model of the Institute of Biomedical Engineering of the Federal University of Santa Catarina (IEB-UFSC). The tool was developed through the integration of findings from a rapid literature review, normative requirements, and the IEB-UFSC conceptual model. It is organized into four progressive levels: initial home assessment (N1), multidisciplinary assessment of identified non-conformities (N2), installation and operational validation of health technologies (N3), and continuous care monitoring (N4). This structure integrates infrastructure, health technology, and human resource aspects, enabling a tiered approach. The proposal provides a structured framework for assessing and monitoring home conditions during the implementation and follow-up of home care.**

Keywords*— *home care, clinical engineering, health technology management, medical devices, home assessment

I. INTRODUCTION

The growing decentralization of healthcare has become increasingly established worldwide, and the home is one of the most important care settings. Driven by population aging, the increasing prevalence of chronic diseases, and the need for transition of care, this modality has emerged as a humanized alternative capable of reducing hospital costs and expanding continuity of treatment within the family context [1][2]. Simultaneously, there is a growing incorporation of health technologies in the home environment, including medical devices, life-support equipment, and monitoring systems, expanding care complexity and challenges related to the safety and management of these technologies [3][4].

Unlike healthcare facilities (HCF), homes are not designed to accommodate medical devices and may present limitations related to electrical infrastructure, physical space, environmental conditions, and user qualification [5][6]. Additionally, human factors represent critical challenges that require special attention from managers, particularly those related to user qualification, caregiver overload, and the lack of systematic monitoring and support mechanisms for the use of health technologies [7][8].

In this context, Clinical Engineering assumes a strategic role in managing health technologies used in the home environment, acting throughout the technology lifecycle and encompassing planning, acquisition, installation, training, monitoring, and maintenance processes. In recent decades, this field has evolved from an equipment maintenance-focused approach to a strategic model oriented toward quality, patient safety, and evidence-based decision-making [3]. In the home care context, this role becomes even more relevant due to the structural variability of homes, the use of devices by lay users, and the absence of continuous technical supervision [4][5].

Despite the growing incorporation of health technologies into home care and the existence of regulations such as Brazilian Health Regulatory Agency (ANVISA) Collegiate Board Resolution (RDC) number 11/2006, and ABNT NBR IEC 60601-1-11:2021, there is a limited availability of structured instruments to guide the systematic assessment of home conditions from a Clinical Engineering perspective [10][11][12]. The literature describes challenges associated with infrastructure, human factors, and the use of medical devices in the home, but lacks standardized methods capable of integrating these elements into an assessment process that supports decision-making regarding the installation, use, and monitoring of health technologies [3][9].

Existing approaches address specific aspects of home care, including the safe use and maintenance of medical devices [13] and the assessment of environmental barriers, adaptations, accessibility, and safety risks [14]. However, these approaches have different scopes and do not explicitly integrate technology, infrastructure, and human resources into a Health Technology Management process. The proposed tool seeks to address this gap through a progressive assessment structure encompassing initial screening, multidisciplinary analysis, installation and operational validation, and continuous monitoring.

In this sense, the present study proposes the development of a home assessment tool grounded in the principles of Clinical Engineering and Health Technology Management,

structured in four progressive levels. The tool integrates aspects related to infrastructure, health technologies, and human resources involved in care, providing a structured framework for the assessment and management of conditions associated with the use of health technologies in the home environment.

II. METHODS

This is an applied, qualitative, and descriptive study aimed at developing a home assessment tool for health technology management from the perspective of Clinical Engineering. The study was structured in two parts: the theoretical foundation of the tool, comprising a rapid literature review, normative analysis, and analysis of the conceptual model of the Institute of Biomedical Engineering of the Federal University of Santa Catarina (IEB-UFSC); and the elaboration and structuring of the tool, organized in four progressive assessment levels.

A. Tool Foundation

The rapid literature review was conducted based on the Methodological Guidelines of the Brazilian Ministry of Health for the development of systematic reviews [15] and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations [16]. The review was guided by the research question: "What are the models, strategies, and outcomes reported in the literature regarding health technology management in the home care context?".

The searches were conducted in July 2025 in the PubMed, IEEE Xplore, Scopus, and SciELO databases. The search strategy applied to PubMed, IEEE Xplore, and Scopus combined the terms (“Home Care” OR “Home Health Care” OR “Home Health Services”) AND (“Health Technology Management” OR “Medical Devices” OR “Medical Equipment” OR “Clinical Engineering”). In SciELO, the strategy was adapted to Portuguese using (“assistência domiciliar” OR “atenção domiciliar”) AND (“tecnologia em saúde” OR “equipamento médico” OR “dispositivos médicos”). No publication date restriction was applied.

Eligibility criteria were defined prior to study selection. Studies were included when they addressed health technology management in the home care context and considered at least one of the three pillars adopted in this study: technology, infrastructure, or human resources related to health technology management in home care. Full-text studies were considered, including original and secondary research. Studies unrelated to the home care context, not addressing the proposed scope, unavailable in full-text, or duplicated were excluded.

The review did not include a formal methodological quality assessment. The included studies were instead characterized according to their objectives, context of application, methodological approach, and contribution to the three health technology management pillars.

All retrieved records were imported into Rayyan for reference management and screening. Duplicate records were automatically identified by Rayyan and subsequently checked manually. The screening process was performed by a single reviewer, initially through title and abstract assessment. A total of 1,260 records were identified, of which 907 remained after duplicate removal. Following title and abstract screening, 100 publications were selected for full-text assessment. After full-text screening, 42 studies were excluded because they did not meet the predefined eligibility criteria, resulting in 58 studies included in the qualitative synthesis. The complete process is presented in Figure 1.

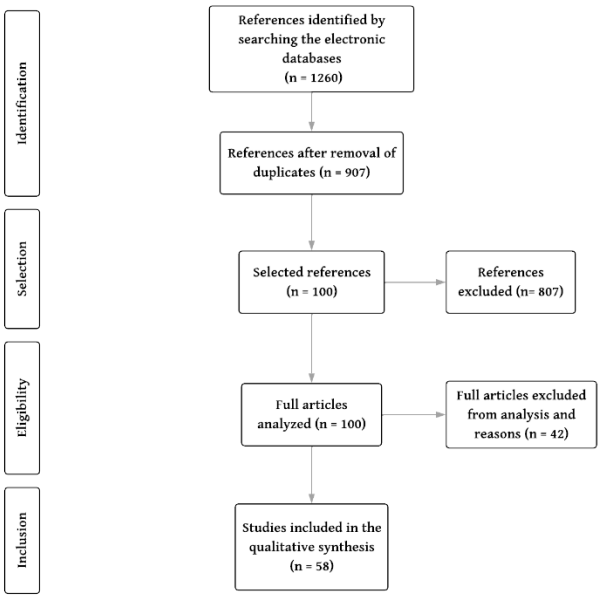


Fig. 1. Flow diagram of the study selection process for the rapid literature review.

The included studies were characterized according to their objectives, context of application, methodological approach, and relationship with the three health technology management pillars. The findings were synthesized into three analytical axes: technology, infrastructure, and human resources. Patient safety, adverse events, regulation, and related cross-cutting issues were considered throughout the synthesis and subsequently incorporated into the development of the assessment tool.

The normative analysis covered ANVISA regulations, national and international technical standards, and home care documents, with the objective of identifying technical and operational requirements for tool construction. Standards specifically related to home use were used as technical references and interpreted in relation to the characteristics of the home environment. Environmental parameters, including temperature, humidity, noise, and illuminance, were incorporated as reference criteria for assessing home conditions and identifying situations requiring further evaluation or corrective measures.

The analysis of the Medical-Hospital Technology Management model of IEB-UFSC, grounded in the pillars of technology, infrastructure, and human resources, was also conducted. This model seeks to promote safety, reliability, and effectiveness of health technology processes through integrated management of these three components

(Figure 2). Considering the particularities of home care, the pillars were adapted to this context and used as the basis for structuring the assessment criteria.

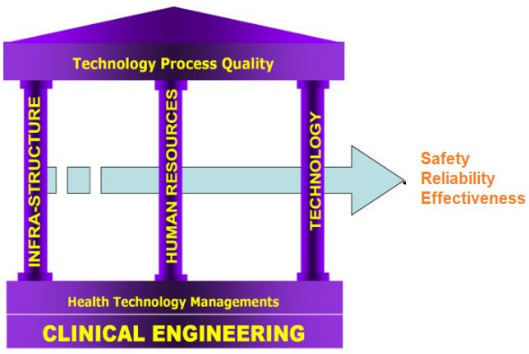


Fig. 2. IEB-UFSC health technology management process model.

B. Structuring the Home Assessment Tool

From the integration of literature findings, identified normative requirements, and the conceptual structure proposed by the IEB-UFSC model, the home assessment tool was developed. The assessment items were organized according to their role in the home care process, considering technology, infrastructure, and human resources, and structured into four progressive levels. This organization follows the sequence from initial assessment to implementation and continuous monitoring, allowing progression between levels according to the conditions identified at each stage.

The process begins at Level 1 (N1), intended for screening home conditions. When no non-conformities requiring further analysis are identified, the flow advances directly to Level 3 (N3). When restrictions or situations requiring further investigation are observed, Level 2 (N2) is triggered to support multidisciplinary assessment of the identified conditions. After defining the necessary adjustments, N3 covers the implementation of home care, including the installation, configuration, and validation of health technologies. Finally, Level 4 (N4) corresponds to the continuous monitoring of care and operational conditions, performed periodically or in response to events requiring reassessment.

The tool uses qualitative classifications rather than numerical weights or scores. Situations requiring further analysis are directed to N2, where the identified conditions can be considered according to the technology, infrastructure, and human resources involved.

III. Results

A. Tool Foundation

1) Rapid Literature Review

The rapid literature review resulted in the inclusion of 58 studies published between 1989 and 2025. The included publications used varied methodological approaches, predominantly observational and descriptive studies, such as case studies, descriptive analyses, and reviews. Results were organized around three analytical axes corresponding to the IEB-UFSC model pillars: technologies, infrastructure, and human resources.

In the technology axis, the literature highlights that the use of medical devices in the home presents challenges related to usability, especially due to operation by lay users and uncontrolled conditions. Issues such as unintuitive interfaces, inadequate instructions, and lack of standardization are pointed out as relevant challenges for safety in the use of these devices [17][18]. Clinical Engineering is identified as a strategic element, acting in device selection, installation, maintenance, and user training, with emphasis on the application of tools such as Failure Mode and Effects Analysis (FMEA) in improving management and safety of home devices [4][19]. Additionally, telemonitoring emerges as a promising alternative for remote monitoring and failure reduction, although challenges related to interoperability, regulation, and information security persist [20][21].

The most frequently reported equipment included mechanical ventilators, oxygen concentrators and cylinders, positive airway pressure devices (PAP/CPAP), infusion pumps, pulse oximeters, sphygmomanometers, multi-parameter monitors, and dialysis equipment [22]. Respiratory support, dialysis, and oxygen therapy devices are the most studied groups, present in 45%, 39%, and 29% of studies included in systematic reviews, respectively [23].

In the infrastructure axis, studies indicate that home adequacy is one of the main challenges of home care, since the residential environment is not designed for clinical purposes. Factors such as unsafe electrical installations, insufficient space, inadequate ventilation and lighting, and precarious hygiene conditions can compromise equipment operation and patient safety [24][25]. The reliability of electricity supply stands out as a critical factor for patients dependent on life-support equipment, with backup systems, contingency plans, and coordination with utility companies recommended for prioritizing these homes [26]. Home preparation should occur before hospital discharge through structured multidisciplinary planning [6].

In the human resources axis, training for the safe use of medical devices is recognized as an essential component of home care. It is noted that technologies originally developed for clinical environments are not always suitable for home users, making it necessary to develop training centered on tasks to be performed daily [7][8]. Family members and caregivers play a fundamental role in equipment operation but may face overload and operational difficulties in the absence of adequate support [27][28]. Management of this dimension also involves adverse event reporting systems and structured communication channels between professionals, patients, and caregivers [5][29].

2) Normative and Regulatory Analysis

The analysis of normative, regulatory, and technical documents aimed to identify relevant technical and operational requirements for tool construction. The analyzed documents covered ANVISA regulations, national and international technical standards, and home

care documents, as presented in Table I. It should be noted that some of these references were originally developed for healthcare facilities and were considered in relation to the specificities of the home environment, considering its structural variability and the absence of an original clinical design. Technical manuals for medical equipment commonly used in home care were also consulted to support operational, environmental, and electrical criteria.

TABLE I. NORMATIVE AND REGULATORY DOCUMENTS ANALYZED

Analyzed Documents		
Category	***Document***	***Contribution to the Tool***
Home Care	RDC 11/2006 [10]	Establishes technical operating regulations for home care services (N1, N4).
	Ordinance GM/MS nº 825/2016 [11]	Redefines Home Care within the Brazilian Unified Health System (SUS) and updates authorized care teams.
	Ministry of Health Home Care Handbook [30] [31]	References for patient eligibility, caregiver role, multidisciplinary work, follow-up, and monitoring of home care.
Health Technology Management	RDC 509/2021 [32]	Governs health technology management in healthcare facilities Reference for the health technology management block at N4.
	ABNT NBR 15943: 2011 [33]	Guidelines for management programs of healthcare infrastructure equipment and health equipment.
Electrical Infrastructure	ABNT NBR 13534:2008 [34]	Requirements for electrical installations in healthcare facilities, adapted for home electrical safety assessment.
	ABNT NBR 5410:2004 [35]	Low-voltage electrical installations; general home electrical safety criteria.
	ABNT NBR 14136:2012 [36]	Standardization of outlets and plugs used for electrical infrastructure assessment.
Environmental Condition	ABNT NBR 10152:2017 [37]	Noise level – reference: ≤ 35dB for healthcare environments; quantitative parameter for N1 and N3.
	ABNT NBR 16401-2:2024 [38]	Thermal comfort: temperature 20–24 °C, humidity 40–60%; quantitative parameters for N1 and N3.
	ABNT NBR ISO/CIE 8995-1:2013 [39]	Reference for illuminance assessment; adopted as a reference criterion in the home care context. Minimum illuminance of 100 lux; quantitative parameter for N1 and N3.
Medical Devices	ABNT NBR IEC 60601-1-11: 2021 [12]	Specific requirements for electro-medical equipment used in home environments, supporting compatibility assessment between technology, user, and environment.
Data Protection	Law nº 13.709/2018 (LGPD) [40]	Reference for evaluating privacy, security, and data processing aspects in telemonitoring technologies and home automation integrated into home care.

B. Development of the home assessment tool

The application of the IEB-UFSC conceptual model made it possible to organize the tool criteria around the technology, infrastructure, and human resources pillars. Adapting these components to the home care context allowed the incorporation of aspects related to environmental conditions, technologies used in care, and the actors involved in care. Figure 3 presents the operational structure of the assessment tool.

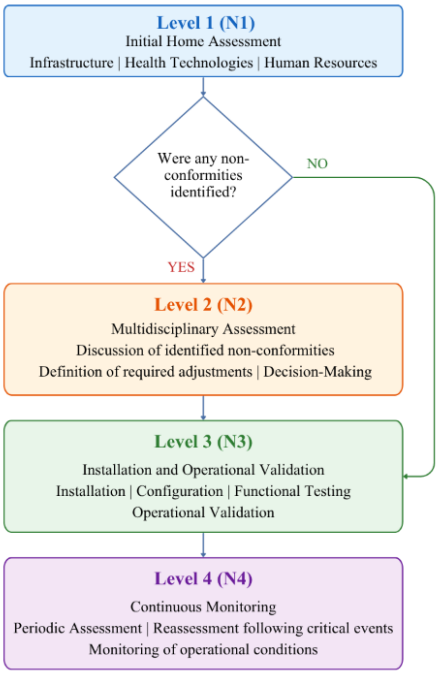


Fig. 3. Operational structure of the home assessment tool

N1 constitutes the initial screening stage and is applied to all homes prior to installation. It assesses physical infrastructure, health technologies, human resources, and quantitative environmental parameters, organized in four blocks (Table II). At the end of the assessment, the home is classified as adequate, with direct advancement to N3; deficient, triggering the N2 multidisciplinary assessment protocol; or critical, indicating the need to reassess the feasibility of implementation with N2. Conditions classified as critical are those with a potential to compromise the feasibility or continuity of care and therefore require immediate multidisciplinary assessment, whereas deficient conditions require corrective measures or further evaluation before progression.

Non-conformities are assessed qualitatively according to their potential to compromise the installation, operation, continuity of care, or safe use of the health technology. The tool does not use numerical scores or weighting; conditions requiring further analysis are directed to N2 for multidisciplinary assessment and definition of the appropriate course of action.

TABLE II. BLOCKS AND ITEMS ASSESSED AT LEVEL 1

Block	*Items to be Assessed*
Physical Infrastructure	Adequacy of physical space (RDC 11/2006); basic sanitation; outlets compliant with NBR 14136:2012; circuit breaker and surge protection device (SPD); electrical grounding; accessibility for wheelchair and stretcher; Wi-Fi or mobile network connectivity.
Health Technologies	Valid ANVISA registration; no active sanitary alerts; technical documentation available; compatibility of equipment with clinical profile and home conditions.
Human Resources	Patient's understanding of equipment use; physical or cognitive limitations of the caregiver; completion of formal training; availability of institutional training protocol.
Environmental Reference Parameters	Electrical voltage (V); reference temperature 20-24° C (NBR 16401-2:2024); reference relative humidity of 40-60% (NBR 16401-2:2024); reference noise level ⩽35 dB (NBR 10152:2017); reference illuminance ⩾100 lux (NBR ISO/CIE 8995-1:2013).

N2 is triggered when N1 identifies deficient or critical conditions requiring further assessment. It consists of a multidisciplinary assessment with mandatory participation from Clinical Engineering representatives, the care team, and service management. Operational responsibilities are shared among Clinical Engineering, the care team, and service management, with all parties participating in the assessment, discussion, decision-making, implementation, and follow-up of identified conditions according to their respective areas of expertise. The assessment is organized in three dimensions (Table III), resulting in a formal decision: installation authorized; authorized subject to adjustments; temporarily contraindicated; or contraindicated, with reassessment after corrective measures when required.

TABLE III. BLOCKS AND NON-CONFORMITIES ADDRESSED AT LEVEL 2

Block	*Non-Conformities Addressed*
Physical Infrastructure	Insufficient physical space; inadequate electrical installation; insufficient number of outlets; absence of grounding; power fluctuation; inadequate ventilation; humidity/infiltration; inadequate hygiene; access difficulties.
Health Technologies	Incompatibility between environment and technology; need for structural adjustment; risk to safe operation; need for complementary technology; risk of therapeutic interruption.
Human Resources	Family context for technology use; training needs; physical/cognitive limitations of the user; difficulty communicating in emergencies.

For example, if N1 identifies an electrical condition that may compromise the operation of a life-support device, the case is directed to N2 for multidisciplinary assessment and definition of corrective measures. After the required adjustments, the technology proceeds to N3 for reassessment and operational validation, while subsequent monitoring is performed through N4, with reassessment triggered when relevant changes or events occur.

N3 covers the installation, configuration, and validation of health technologies before the start of home care, structured in five assessment blocks (Table IV). In cases of home automation and telemonitoring, N3 also includes the assessment of sensors, connectivity, and the telemonitoring system. At the end of this stage, the home setup is classified as Validated, Pending, or Not Validated for service admission.

TABLE IV. BLOCKS AND ITEMS VERIFIED AT LEVEL 3

Block	*Items to be Assessed*
Physical Infrastructure – Physical Space	Room dimensions compatible with the planned technology; minimum clearance for installation and care; equipment access to the care environment.
Physical Infrastructure – Electrical Installation	Outlets near the installation site; voltage measurement per NBR 13534:2008; tested and documented energy backup with adequate autonomy; registration with the utility company as a priority customer; grounding verified; protection devices (SPD, RCD) installed.
Physical Infrastructure – Environmental Conditions	Reference temperature (NBR 16401-2); reference humidity 40–60% (NBR 16401-2); reference noise ≤ 35 dB (NBR 10152:2017); reference illuminance ≥ 100 lux (NBR ISO/CIE 8995-1); no electromagnetic interference identified.
Health Technology	Installation per manufacturer's technical specifications; clinical parameters configured and validated; alarm thresholds tested and documented; operation tested under real conditions; traceability recorded (serial number, date, initial configuration); contingency plan defined for life-support equipment.
Human Resources	Formal training completed for each device; practical competency assessed and recorded; educational materials available in the home; liability form signed by the responsible party.

N4 is intended for continuous monitoring after implementation and may be applied periodically or in response to events requiring reassessment of home conditions, technologies in use, or human resources involved. The assessment is organized in three blocks (Table V), and each item is classified as Compliant, Pending, or Non-Compliant, generating an action plan with responsible party, deadline, and priority.

TABLE V. BLOCKS AND ITEMS MONITORED AT LEVEL 4

Block	*Monitored Items*
Physical Infrastructure	General home conditions (space, hygiene, and ventilation); electrical installation without changes or damage; backup power system functional with maintained autonomy; correct disposal of healthcare waste.
Health Technologies	Equipment free of physical damage and in full operation; parameters and alarms configured per prescription; preventive maintenance performed according to the applicable maintenance program and manufacturer recommendations, with RDC 509/2021 used as a health technology management reference; accessories and supplies within expiration date; absence of recurring failures; no active sanitary alert or recall (ANVISA); adverse event records

Block	*Monitored Items*
	updated; telemonitoring functional (when applicable).
Human Resources	Caregiver present and without changes that compromise care; operational competence maintained; questions identified and clarified during the visit; communication channel with the care team active.

The developed tool integrates criteria related to infrastructure, health technologies, and human resources, organized in four progressive assessment levels. Grounded in the findings of the literature review, normative requirements, and the IEB-UFSC Clinical Engineering model, the tool organizes the assessment process from initial home screening to continuous monitoring of conditions relevant to the use of health technologies.

IV. Discussion

The transfer of technologies to the home environment introduces challenges related to safety, device performance, and the ability of both the environment and users to support their use. The literature indicates that physical infrastructure, environmental conditions, equipment usability, and user preparedness are important factors influencing the safety and effectiveness of home care [7][17][24][29]. In this context, the proposed tool seeks to systematize this assessment through criteria related to infrastructure, health technologies, and human resources.

Although the literature presents evidence on challenges related to infrastructure, medical devices, and human factors in home care, these elements are frequently addressed in isolation [6][7][17]. The findings of the review suggest that conditions affecting the safe use of health technologies may arise from the interaction among these components. For example, the availability of an appropriate technology does not by itself ensure safe care when structural limitations in the home compromise its installation or operation, or when caregivers lack the skills required for its use [28]. Therefore, the assessment tool was structured to consider technology, infrastructure, and human resources together, allowing conditions identified in one dimension to be considered in relation to the requirements of the others.

The tool was constructed by combining home care-specific documents, such as RDC 11/2006, with standards originally developed for healthcare facilities [10]. This approach was adopted because specific references for the technical assessment of home environments are limited, particularly regarding infrastructure and safety requirements. Requirements from these standards were considered as technical references and interpreted in relation to the structural variability of homes and the participation of non-professional users in care. For environmental conditions, the corresponding values were incorporated as reference parameters rather than rigid exclusion criteria. Deviations from these parameters are intended to indicate conditions requiring further assessment according to the technology involved, care requirements, and characteristics of the home.

This integrated approach is aligned with the principles of Clinical Engineering and Health Technology Management, fields that have been expanding their scope beyond traditional maintenance activities [4][9]. The Medical-Hospital Technology Management model developed at IEB-UFSC proposes a systemic view of health technology processes, considering the interaction between technology, infrastructure, and human resources as essential elements for the safety, reliability, and effectiveness of services [41]. Adapting this model to the home context allowed the incorporation of characteristics unique to this environment, such as the structural variability of homes, the active participation of patients, family members, and caregivers, and the limited availability of technical support [3][42]. These characteristics were incorporated into the proposed assessment structure, particularly through criteria addressing home infrastructure, technology–environment compatibility, caregiver competence, and conditions requiring reassessment. In this way, the IEB-UFSC model served as the conceptual basis for organizing the assessment criteria within the home care context.

Existing international approaches address several aspects of safety related to health technologies in the home. The Food and Drug Administration (FDA) provides guidance materials and checklists addressing the safe use and maintenance of medical devices, while Emergency Care Research Institute (ECRI) highlights risks associated with device use in home healthcare [13][43][44]. Other home assessment approaches focus on environmental hazards, functional performance, person-environment fit, and home modifications [14]. Home safety guidance also emphasizes continuous assessment, reassessment, and collaboration among patients, caregivers, and healthcare professionals [45]. Although these approaches address important components of home care safety, they differ in scope and in how assessment activities are organized.

In contrast, the proposed tool brings these dimensions together within a Health Technology Management process structured into four progressive levels. N1 performs initial screening of home conditions; N2 addresses conditions requiring multidisciplinary assessment; N3 focuses on installation and operational validation; and N4 establishes continued monitoring and reassessment. This organization links findings from infrastructure, technology, and human resources to successive decision and follow-up stages, providing an operational structure for documenting conditions and defining subsequent actions.

Despite its theoretical and normative foundation, the proposed tool has not yet undergone formal content validation or evaluation of reliability, usability, and operational applicability in real home care settings. The rapid review also did not include a formal methodological quality appraisal, which limits the assessment of the strength of the evidence synthesized. These aspects are important for determining whether the proposed criteria can be consistently interpreted and applied by different professionals and services. Existing home assessment instruments have been subjected to different reliability and validity assessments, demonstrating the relevance of these procedures for consolidating structured assessment tools

[46]. Therefore, future studies should include expert content validation, inter-rater reliability assessment, usability evaluation, and pilot application in home care services. These studies should also evaluate the time required for application, ease of completion, training needs, acceptability, and the tool's influence on the decision-making process among professionals from different home care services.

The proposed framework also incorporates elements that support future digital implementation, particularly through the assessment of sensors, connectivity, telemonitoring, alert management, and data protection. Future developments may include integration with electronic health records and decision-support platforms to expand data sharing and follow-up of home care conditions.

The development of the tool highlighted the importance of considering technological, environmental, and human factors together when assessing conditions for home care. Future research should expand the evidence base through multicenter studies and the development of indicators to evaluate the operational implications and potential clinical relevance of health technology management in the home environment.

V. CONCLUSIONS

This study aimed to develop a home assessment tool for health technology management from the perspective of Clinical Engineering. Grounded in scientific evidence, normative requirements, and the Medical-Hospital Technology Management conceptual model of IEB-UFSC, the tool was structured into four progressive assessment levels, encompassing aspects related to infrastructure, health technologies, and human resources involved in care.

The proposed tool organizes the assessment of home conditions throughout the implementation and follow-up stages of home care by integrating technology, infrastructure, and human resources into a single assessment structure. Its four progressive levels connect initial screening, multidisciplinary assessment, installation and operational validation, and continuous monitoring.

As a preliminary and theoretically grounded proposal, the tool has not yet undergone expert content validation, reliability or usability assessment, or application in real home care services. Therefore, future studies should include content validation, inter-rater reliability assessment, usability evaluation, and pilot application in real care settings to assess its applicability and operational performance. These studies may support the refinement of the assessment criteria and the further development of the tool for use in different home care contexts.

REFERENCES

[1] F. L. Rajão and M. Martins, "Home care in Brazil: An exploratory study on the construction process and service use in the Brazilian health system," *Ciencia e Saude Coletiva*, vol. 25, no. 5, pp. 1863–1877, 2020.

[2] M. L. S. V. Rangel, "Processo de desospitalização e atenção domiciliar no Brasil e seus fatores associados", Research, *Society and Development*, vol. 12, no. 4, p. e0612440793, 2023.

[3] M. W. Darold and R. G. Ojeda, "Proposal for a Health Technology Management Program for Home Care," in *Proc. Congresso Argentino de Bioingeniería*, Cham: Springer Nature Switzerland, 2025, pp. 343–358.

[4] P. S., de Avelar, F. Sônego and R. Garcia, "A Gestão de Tecnologia Médico-Hospitalar como Estratégia da Engenharia Clínica no Atendimento Domiciliar no Brasil," in *Proc. IV Latin American Congress on Biomedical Engineering*, Margarita Island, Venezuela, 2007, pp. 1203–1206.

[5] C. A. Mestas and S. J. Calil, "Risks Involved With Medical Device Use in a Home Care Program," in *Proc. World Congress on Medical Physics and Biomedical Engineering*, Munich, Germany, 2009, pp. 519–522.

[6] A. Torrent-Vernetta *et al.*, "Arrangement of residence before hospital discharge for children on home-invasive mechanical ventilation," *Pediatric Pulmonology*, vol. 59, no. 8, pp. 2080–2088, 2024.

[7] S. E. McBride *et al.*, "Challenges of training older adults in a home health care context," in *Proc. Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA, 2012, pp. 2492–2496.

[8] P. A. Patterson, "Human factors in labeling and training for home healthcare technology," *Biomedical Instrumentation & Technology*, p. 33, 2010.

[9] C. A. Mestas and S. Calil, "The Clinical Engineering in Home Care: A Strategic Vision," in *Proc. 5th European Conf. International Federation for Medical and Biological Engineering*, Budapest, Hungary, 2012, pp. 953–956.

[10] Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution – RDC nº 11, January 26, 2006: Provides for the Technical Regulations for the Operation of Home Care Services. Federal Official Gazette, Brasília (2006).

[11] Brazil, Health Ministry, Ordinance GM/MS nº 825, April 26, 2016: Redefines Home Care within the Unified Health System (SUS) and Updates the Authorized Home Care Teams, Federal Official Gazette, Brasília, 2016.

[12] Brazilian Association of Technical Standards (ABNT), NBR IEC 60601-1-11: General requirements for basic safety and essential performance - Collateral Standard: Requirements for electromedical equipment and electromedical systems used in domestic health care environments. Rio de Janeiro, 63 p., 2021

[13] U.S. Food and Drug Administration (FDA), "Brochure: Home Healthcare Medical Devices Checklist." [Online]. Available: https://www.fda.gov/medical-devices/home-health-and-consumer-devices/brochure-home-healthcare-medical-devices-checklist. Accessed: Jun. 08, 2026.

[14] J. N. Aroni, P. C. C. Ribeiro, and E. do Nascimento, "Instrumentos para avaliação do ambiente domiciliar de idosos: uma revisão sistemática". Estudos Interdisciplinares sobre o Envelhecimento (Online), pp. 147–170, 2020.

[15] Health Ministry, Methodological Guidelines: development of systematic reviews and meta-analysis of randomized clinical trials, Brasília: Ministry of Health Publisher, 2021.

[16] D. Moher *et al.*, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement," *PLOS Medicine*, vol. 6, no. 7, p. e1000097, Jul. 2009. doi: 10.1371/journal.pmed.1000097.

[17] A. Tase *et al.*, "Usability challenges in the use of medical devices in the home environment: a systematic review of literature," *Applied Ergonomics*, vol. 103, p. 103769, 2022.

[18] M. R. Lemke and R. J. Mendonca, "A question of accessibility: understanding lay users of medical devices," *Biomedical Instrumentation & Technology*, vol. 47, no. s1, pp. 20–25, 2013.

[19] O. Gajdoš *et al.*, "Failure Mode and Effects Analysis Application on a Home Mechanical Ventilation," in *Proc. 2019 E-Health and Bioengineering Conf. (EHB)*, IEEE, 2019, pp. 1–4.

[20] J. Ackrivo, L. Elman and J. Hansen-Flaschen, "Telemonitoring for home-assisted ventilation: a narrative review," *Annals of the American Thoracic Society*, vol. 18, no. 11, pp. 1761–1772, 2021.

[21] E. Parimbelli *et al.*, "Trusting telemedicine: a discussion on risks,

safety, legal implications and liability of involved stakeholders," *International Journal of Medical Informatics*, vol. 112, pp. 90–98, 2018.

[22] M. F. Story, “Medical devices in home health care,” in The Role of Human Factors in Home Health Care: Workshop Summary, Washington, DC, USA: National Academies Press, 2010, pp. 145–172.

[23] I. Ten Haken, S. Ben Allouch, and W. H. van Harten, “The use of advanced medical technologies at home: a systematic review of the literature,” BMC Public Health, vol. 18, no. 1, p. 284, 2018.

[24] D. Kautman-Rivi, J. Collins-Mitchell and R. Jetley, "Design considerations for medical devices in the home environment," *Biomedical Instrumentation & Technology*, p. 21, 2010.

[25] D. R. Milamed, H. Lasthaus and J. Hedley-Whyte, "Highlights of the New International Standard on Home Healthcare Devices," *Biomedical Instrumentation & Technology*, p. 27, 2010.

[26] R. Bean *et al.*, "Keeping the power on to home medical devices," *PLOS ONE*, vol. 15, no. 7, p. e0235068, 2020.

[27] H. J. Barton *et al.*, "Medical device workarounds in providing care for children with medical complexity in the home," *Pediatrics*, vol. 147, no. 5, p. e2020019513, 2021.

[28] B. Munck *et al.*, "District nurses' experiences of advanced medical technology in home care: A qualitative interview study," *Nordic Journal of Nursing Research*, vol. 43, no. 1, p. 20571585221127344, 2023.

[29] E. Lauta *et al.*, "Improving Homecare Risk Management and Patient Safety," *Epidemiology, Biostatistics, and Public Health*, vol. 16, no. 4, 2019.

[30] Brazil, Health Ministry, Secretariat of Health Care, Department of Primary Care, Home Care Handbook, vol. 1, Brasília: Ministry of Health, 2012.

[31] Brazil, Ministry of Health, Secretariat of Health Care, Department of Primary Care, Home Care Handbook, vol. 2, Brasília: Ministry of Health, 2013.

[32] Brazilian Health Regulatory Agency (ANVISA), Collegiate Board Resolution – RDC nº 509, May 27, 2021: Provides for Health Technology Management in Healthcare Establishments, Federal Official Gazette, Brasília, 2021.

[33] Brazilian Association of Technical Standards (ABNT), NBR 15943:2011 — Guidelines for a Management Program for Healthcare Infrastructure Equipment and Medical Equipment, Rio de Janeiro, Brazil, 2011.

[34] Brazilian Association of Technical Standards (ABNT), NBR 13534:2008 — Low-Voltage Electrical Installations: Specific Requirements for Installations in Healthcare Facilities, Rio de Janeiro, Brazil 2008.

[35] Brazilian Association of Technical Standards (ABNT), NBR 5410:2004 — Low-Voltage Electrical Installations, Rio de Janeiro, Brazil, 2004.

[36] Brazilian Association of Technical Standards (ABNT), NBR 14136:2012 — Plugs and Socket-Outlets for Household and Similar Purposes: Standardization, Rio de Janeiro, Brazil, 2012.

[37] Brazilian Association of Technical Standards (ABNT), NBR 10152:2017 Acoustics: Sound Pressure Levels in Indoor Environments of Buildings, Rio de Janeiro, Brazil, 2017.

[38] Brazilian Association of Technical Standards (ABNT), NBR 16401-2:2024 — Conditioning Installations: Central and Unitary Systems. Part 2: Thermal Comfort Parameter, Rio de Janeiro, Brazil, 2024.

[39] Brazilian Association of Technical Standards (ABNT), NBR ISO/CIE 8995-1:2013 Lighting of Work Places. Part 1: Indoor, Rio de Janeiro, Brazil, 2013.

[40] Brazil, Law nº 13.709, de 14 de agosto de 2018: General Data Protection Law (LGPD), Federal Official Gazette, Brasília, 2018.

[41] S. J. A. Garcia *et al.*, "Health care technology management applied to public primary care health," in *Proc. 2011 Pan American Health Care Exchanges*, IEEE, 2011, pp. 250–253.

[42] R. Cyrus, "Home Healthcare: The Next BMET Frontier," *Biomedical Instrumentation & Technology*, p. 57, 2010.

[43]ECRI, *Top 10 Health Technology Hazards for 2024*. Plymouth Meeting, PA, USA: ECRI, 2024. [Online]. Available: https://home.ecri.org/blogs/ecri-news/challenges-with-home-use-medical-devices-for-patients-and-caregivers-tops-ecris-2024-health-tech-hazards. Accessed: Jun. 08, 2026.

[44]ECRI, *Top 10 Health Technology Hazards for 2025*. Plymouth Meeting, PA, USA: ECRI, 2025. [Online]. Available: https://home.ecri.org/blogs/ecri-thought-leadership-resources/techhazards2025. Accessed: Jun. 08, 2026.

[45] K. L. Davis and D. D. Davis, “Home safety techniques,” StatPearls [Internet], Treasure Island, FL, USA: StatPearls Publishing, Jul. 17, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK560539/

[46] Home Modification Information Clearinghouse, Evidence Based Practice Review: Self-assessment Tools of Person-Environment Fit, 1st ed. Sydney, Australia: University of New South Wales, Dec. 2020, doi: 10.26288/5fec08907af77.

Development of Open-Source Software for Automated Accuracy Testing of Infusion Pumps

Mathis Moissaing
Grenoble Institute of Technology – Phelma
University Grenoble Alpes
Grenoble, France
ORCID: 0009-0006-2613-8982

Henrique T. Moriya
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brazil
ORCID: 0000-0002-1595-1779

Diego A. O. Rosa
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brazil
ORCID: 0009-0004-2422-8453

***Abstract*—This work presents an open-source software developed to automate the verification of infusion pump performance according to IEC 60601-2-24:2012. By integrating real-time data acquisition from an analytical balance with automated data processing, the software generates start-up graphs and trumpet curves, eliminating manual spreadsheet calculations. Key features include configurable sampling rates, automatic data storage, and adaptive visualization for flow rates with deviations exceeding the standard plotting range. Experimental validation demonstrated that the proposed system simplifies the testing workflow, reduces manual processing, and facilitates the analysis of infusion pump performance. The software provides a practical tool for post-market surveillance and research applications involving infusion pump accuracy.**

***Index Terms*—Infusion pumps, accuracy, post-market surveillance, automated testing, measurement uncertainty.**

I. Introduction

The accuracy of medical devices used in intensive care units (ICUs) is crucial for patient safety and treatment effectiveness. Therefore, robust regulatory systems and appropriate measurement capabilities are required to monitor the performance of medical devices throughout their operational lifetime.

Over the past decades, medical device regulation has evolved considerably [1], [2]. Regulatory authorities establish requirements that manufacturers must satisfy to demonstrate the safety and performance of medical devices before market authorization [2].

However, regulatory oversight does not end after market authorization. Medical devices must also be monitored during their use in real-world environments through activities commonly referred to as post-market surveillance (PMS) [1]. PMS includes the systematic collection and evaluation of information concerning device performance, safety, and potential failures.

The importance of PMS and maintenance of medical devices became particularly evident during the COVID-19 pandemic, which exposed vulnerabilities in the availability, maintenance, and performance monitoring of medical equipment. Regulatory authorities have consequently established specific requirements and recommendations for post-market surveillance activities [3], [4].

Medical device failures and inappropriate device use can contribute to adverse events in intensive and neonatal care. Reported problems include equipment failures, incorrect device settings, inadequate alarm configuration, and insufficient maintenance [5], [6]. A study conducted in the United Kingdom involving 1,021 incidents associated with medical equipment in an intensive care unit reported 537 incidents related to equipment failure and 358 associated with incorrect staff settings. Harm was reported in 26% of equipment-failure incidents and 39% of incidents involving incorrect settings [7].

Process improvement strategies may also contribute to reducing errors associated with infusion devices. In a study conducted in an intensive care department in the Netherlands, a Lean-based quality improvement approach involving 1,748 infusion pumps reduced reported infusion pump-related errors from 17.7% to 2.3% [8].

Although human errors can be addressed through appropriate training and awareness, device-related errors require appropriate maintenance, calibration, and performance verification. Regular verification is particularly important for infusion pumps because deviations in delivered flow may affect drug administration and consequently patient treatment.

Previous studies have demonstrated that infusion pump performance may depend on factors such as pump design, operating conditions, accessories, and flow rate [10]. Mechanical and electromechanical characteristics of the pumping mechanism can also affect the stability and accuracy of fluid delivery.

Infusion pumps are widely used in intensive care environments and are therefore required to provide safe, reliable, and sufficiently accurate fluid delivery. Over-infusion may result in excessive drug administration and potentially cause toxicity, whereas under-infusion may result in insufficient treatment [13]–[15].

TABLE I
SPECIFICATIONS OF THE ANALYTICAL BALANCE USED IN THE EXPERIMENTS.

Characteristic	Value
Maximum capacity	220 g
Readability	0.1 mg
Repeatability	±0.1 mg
Linearity	±0.2 mg
Stabilization time	approximately 2 s
Calibration	Internal and external
Pan diameter	80 mm
Operating temperature	15 °C to 30 °C

In this context, automated testing tools may facilitate the implementation of systematic post-market performance verification. Automation can reduce manual data processing, standardize calculations, and facilitate the analysis of a large number of devices.

Therefore, this article presents an open-source software developed to automate the acquisition, processing, and analysis of gravimetric measurements obtained during infusion pump accuracy tests according to IEC 60601-2-24:2012 [9]. The software integrates data acquisition from an analytical balance, automated calculations, spreadsheet generation, and graphical visualization of start-up behavior and trumpet curves.

The source code is openly available at:
https://github.com/Mathis-moiss/Infusion-pumps-testing-software.

II. MATERIALS AND METHODS

A. Experimental Setup

The experimental setup consisted of an infusion pump, an analytical balance, and the software developed in this study. The software was designed to support both volumetric infusion pumps using peristaltic mechanisms and syringe pumps.

1) Infusion Pump: A volumetric infusion pump equipped with a linear peristaltic mechanism (Infusomat Compact, B. Braun, Germany) was used to verify the functionality of the software.

2) Analytical Balance: An analytical balance (BP221S, Sartorius, Germany) with a maximum capacity of 220 g was used for the gravimetric measurement of the infused liquid. The balance has a readability of 0.1 mg and was operated according to the manufacturer's instructions.

The main specifications of the analytical balance are presented in Table I.

3) Software: The software was developed in Python version 3.12. It performs data acquisition, processing, analysis, visualization, and automatic storage of the experimental results.

The software communicates directly with the analytical balance through a serial communication interface and records the mass measurements at the selected sampling interval.

B. IEC 60601-2-24 Standard

The experimental procedure was based on IEC 60601-2-24:2012, which specifies particular requirements for the basic safety and essential performance of infusion pumps and controllers [9].

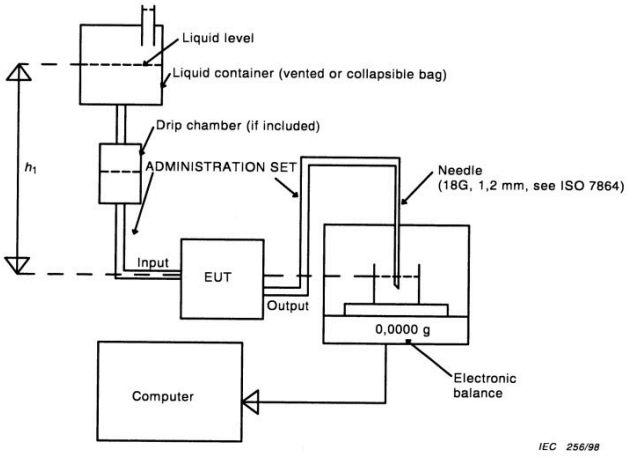


Fig. 1. Experimental apparatus for testing volumetric infusion pumps. Adapted from IEC 60601-2-24:2012 [9]. EUT: equipment under test.

The standard establishes requirements for the evaluation of flow accuracy and specifies the experimental arrangement, acquisition procedure, data analysis, and graphical representation of the results.

Figure 1 illustrates the experimental arrangement used for volumetric infusion pumps.

The software implements the calculations required to obtain the start-up graph, trumpet curves, and overall mean percentage flow errors.

1) Start-Up Graph: The start-up graph represents the flow-rate behavior throughout the initial period of the test.

For each sampling interval, the average flow rate Q_i is calculated from the difference between two consecutive mass measurements according to

$$Q_i = \frac{60(W_i - W_{i-1})}{Sd}(mL/h) \tag{1}$$

where Q_i is the calculated flow rate in mL/h, W_i and W_{i-1} are consecutive mass measurements in $T_x(g)$, S is the sampling interval in minutes, and d is the liquid density in g/mL.

The index i represents the sampling interval, with

$$i = 1, 2 \ldots T_0/S \tag{2}$$

For the start-up graph, the analysis period T_0 corresponds to the first 120 min of the test.

The graph axes follow the requirements of IEC 60601-2-24:2012. The flow-rate axis ranges from $-0.2r$ to $2r$, where r is the programmed flow rate, with increments of $0.2r$. The time axis ranges from 0 to 120 min, with increments of 10 min.

2) Trumpet Curves: Infusion pumps may deliver fluid in a pulsatile manner. Consequently, short-term flow-rate measurements may exhibit substantial deviations from the programmed flow rate, whereas longer observation periods generally provide more stable average values.

Trumpet curves represent the maximum positive and negative percentage errors obtained for different observation-window durations.

The observation-window durations considered:

$$P \in \{2, 5, 11, 19, 31\} \text{ min.} \tag{3}$$

Three analysis periods were considered when applicable: T_0, corresponding to the first 120 min; T_1, corresponding to the second hour of the test; and T_2, corresponding to the last hour of the test.

For an observation window of duration P, the number of windows within an analysis period T_x.

$$m = \left\lfloor \frac{T_x - P}{S} \right\rfloor + 1 \tag{4}$$

where m is the maximum number of observation windows, T_x is the analysis period specified as T_0, T_1, or T_2; P is the duration of the observation window, and S is the sampling interval in minutes.

The maximum positive percentage error:

$$E_{p(\max)} = \max_{j=1,\ldots,m} \left[\frac{S}{P} \times \sum_{i=j}^{j+\frac{P}{S}-1} 100 \times \left(\frac{Q_i - r}{r} \right) \right] (\%) \tag{5}$$

and the maximum negative percentage error:

$$E_{p(\min)} = \min_{j=1,\ldots,m} \left[\frac{S}{P} \times \sum_{i=j}^{j+\frac{P}{S}-1} 100 \times \left(\frac{Q_i - r}{r} \right) \right] (\%) \tag{6}$$

The overall mean percentage flow errors for the analysis period T_1 and T_2:

$$Q_a = \frac{60(W_j - W_k)}{T_1 d} (mL/h) \tag{7}$$

and

$$Q_b = \frac{60(W_j - W_k)}{T_2 d} (mL/h) \tag{8}$$

where W_j and W_k are the final and initial mass measurements of T_1, and T_2 respectively.

The corresponding overall mean percentage flow errors are

$$A = \frac{100(Q_a - r)}{r} (\%) \tag{9}$$

and

$$B = \frac{100(Q_b - r)}{r} (\%) \tag{10}$$

Here, r is the programmed flow rate in mL/h, while A and B are expressed as percentages.

For the trumpet curves, the standard graphical representation uses a vertical axis ranging from -15% to $+15\%$, with increments of 5%, and a horizontal axis ranging from 0 to 31 min.

C. Measurement Uncertainty

Measurement uncertainty was considered in the evaluation of the gravimetric flow-rate measurement. The primary measured quantity used to calculate the flow rate is the mass difference between two consecutive measurements.

For a single sampling interval, the flow-rate measurand:

$$Q = \frac{60(W_j - W_k)}{Td} (mL/h) \tag{11}$$

Assuming that the uncertainties associated with the two mass measurements are independent, the combined standard uncertainty of Q can be estimated using the law of propagation of uncertainty:

$$u^2(Q) = \left(\frac{\partial Q}{\partial W_k} u(W_k) \right)^2 + \left(\frac{\partial Q}{\partial W_j} u(W_j) \right)^2 \tag{12}$$

The corresponding sensitivity coefficients:

$$\frac{\partial Q}{\partial W_k} = -\frac{60}{Td} \qquad \frac{\partial Q}{\partial W_j} = \frac{60}{Td} \tag{13}$$

Assuming equal standard uncertainties for the two mass measurements:

$$u(W_k) = u(W_j) = u(W) \tag{14}$$

Eq. (12) becomes:

$$u(Q) = \frac{60\sqrt{2}}{Td} u(W) \tag{15}$$

The standard uncertainty associated with the balance measurement was obtained from its calibration certificate. The certificate reports an expanded uncertainty of 0.0002 g with a coverage factor of $k = 2$.

$$u(W) = \frac{U(W)}{k} = \frac{0.0002\,\text{g}}{2} = 0.0001\,\text{g} \tag{16}$$

Thus, the uncertainty associated with the analytical balance can be propagated to the calculated flow rate using Eq. (15).

The uncertainty analysis was restricted to the gravimetric flow-rate measurement for T_1 and T_2 and was not extended to the extremal values of the trumpet curves. Consecutive flow-rate estimates Q_i are calculated from overlapping mass measurements. For example, Q_i and Q_{i+1} share the mass measurement W_i and are therefore not statistically independent. Consequently, a rigorous uncertainty evaluation of $E_{p(\max)}$ and $E_{p(\min)}$ would require the consideration of covariance among the underlying measurements, as well as the statistical treatment associated with selecting the maximum and minimum values from multiple observation windows. This analysis is beyond the scope of the present software-development study.

The sampling interval was controlled by the data-acquisition software, and the liquid density was treated as a controlled input in the present simplified model. Therefore, timing and density contributions were not included in the uncertainty estimate reported here.

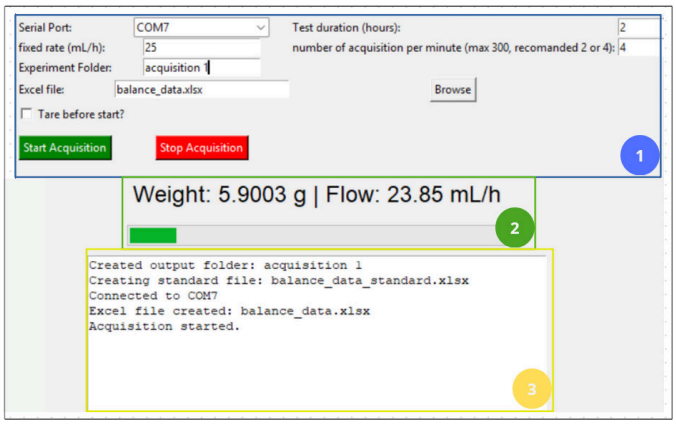


Fig. 2. Software graphical user interface.

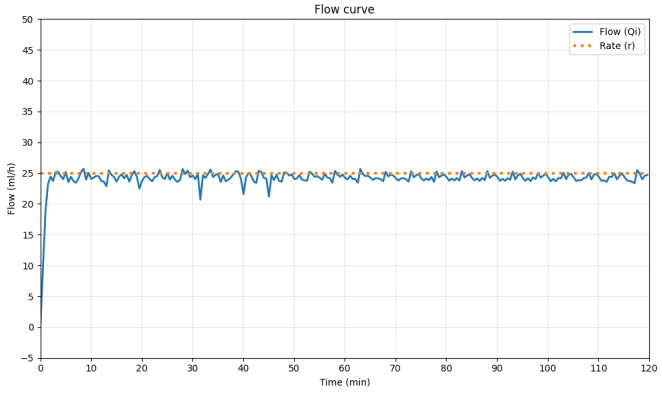


Fig. 3. Start-up graph generated by the software for a programmed flow rate of 25 mL/h.

III. Results

A. Graphical User Interface

Figure 2 presents the software graphical user interface. The interface comprises three main areas: (1) a configuration panel for defining the acquisition parameters; (2) a real-time monitoring area displaying the progress of the acquisition, the measured mass, and the calculated flow rate; and (3) a status window providing information concerning the acquisition process and potential errors.

B. Start-Up Graphs and Trumpet Curves

After completion of the acquisition, the software automatically calculates, plots, and saves the results.

The start-up graph generated by the software is presented in Fig. 3. The trumpet curves are presented in Fig. 4. The infusion pump was programmed at 25 mL/h during the test.

C. Spreadsheet Generation

The software automatically generates a spreadsheet containing the data acquired during each test.

The spreadsheet includes the following parameters: time in hours, time in minutes, mass W_i in grams, calculated flow rate Q_i in mL/h, and programmed flow rate r in mL/h.

The generated data provide a complete record of the acquisition and were used for software verification and validation. The files may also support subsequent verification activities and independent review of the calculations.

D. Sampling Rate Assessment

An excessively high sampling rate caused the data acquisition process to exceed the data transmission capability of the analytical balance. This resulted in intermittent zero readings and consequently generated unrealistic flow-rate values.

A dichotomy-based procedure was used to determine the maximum sampling rate that could be reliably handled by the communication interface. Tests were initially performed at high sampling frequencies and progressively reduced until continuous data acquisition was obtained.

A maximum operational acquisition rate of 300 samples per minute was identified for continuous data transmission without zero-value artifacts, as illustrated in Table II.

Although the communication interface allowed acquisition rates substantially higher than the sampling rate required for the standard analysis, very high sampling frequencies produced flow-rate values that were not representative of the pump's short-term delivery behavior. At very short intervals, the mass difference between consecutive readings approaches the resolution and repeatability limits of the balance, amplifying fluctuations in the calculated flow rate.

Therefore, although the software permits flexible sampling configurations, a sampling rate of two samples per minute was maintained for analyses following IEC 60601-2-24:2012. Higher sampling rates, such as four or six samples per minute, may be useful for exploratory or research applications, provided that their influence on the calculated flow-rate variability is considered.

Figure 5 illustrates the flow-rate behavior obtained with six samples per minute.

IV. Discussion

A. Analytical Balance Communication

The software successfully communicates with the analytical balance to acquire and record mass measurements. However, bidirectional communication was not implemented. Consequently, commands such as remote tare cannot currently be sent from the software to the balance.

The communication module was developed specifically for the balance model used in this study. Different balance models may employ different ASCII command sets or communication protocols. Therefore, modifications to the communication module may be required when another balance model is used.

Future development should include a configurable communication layer that allows the user to select the balance model and communication protocol without modifying the main software code.

B. Data Acquisition

Particular attention was given to the organization and accessibility of the acquired data. The software automatically

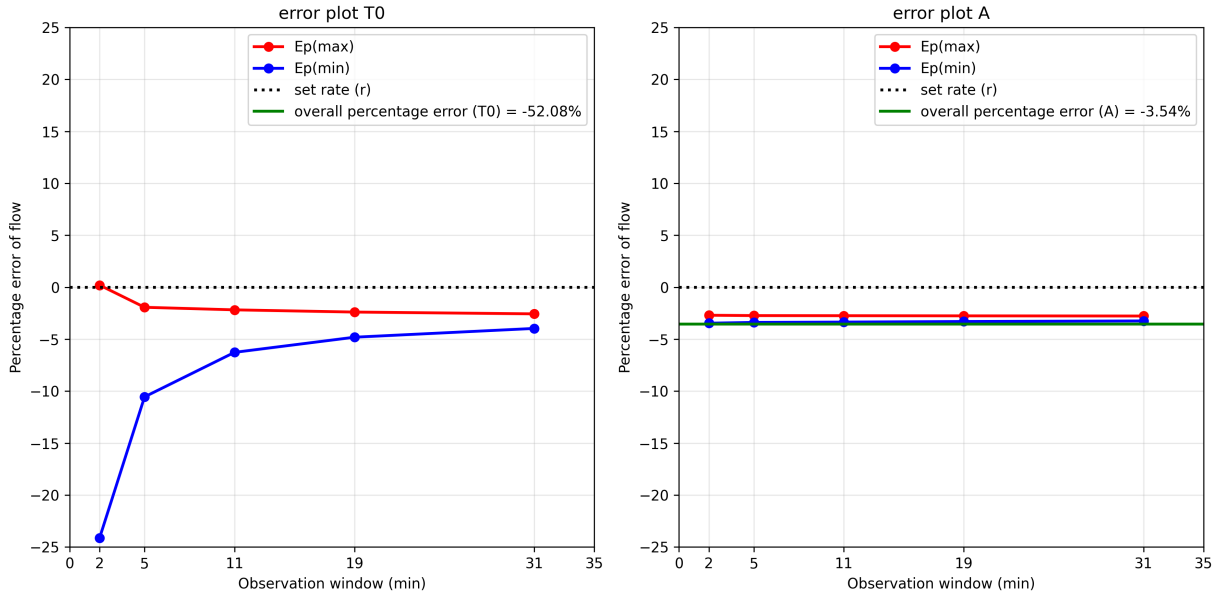


Fig. 4. Trumpet curves generated by the software for a programmed flow rate of 25 mL/h.

TABLE II
EXAMPLES OF DATA ACQUIRED AT HIGH SAMPLING RATES.

Time (h)	Time (min)	W_i (g)	fQ_i	Q_i (mL/h)	r (mL/h)
0.0000	0.00	0.0000	25	0.00	25
0.0000	0.02	64.8766	25	1 367 298	25
0.0001	0.01	64.8776	25	21.09	25
0.0001	0.01	64.8796	25	42.32	25
0.0002	0.01	64.8814	25	38.48	25
0.0002	0.02	0.0000	25	−1 382 503	25
0.0003	0.02	64.8837	25	1 381 220	25
0.0004	0.02	64.8842	25	10.59	25
0.0004	0.02	64.8852	25	21.06	25
0.0004	0.03	64.8856	25	8.51	25
0.0005	0.03	64.8874	25	38.26	25
0.0005	0.03	64.8952	25	160.60	25
0.0006	0.03	0.0000	25	−1 383 895	25

(a) Acquisition at 360 samples per minute.

Time (h)	Time (min)	W_i (g)	fQ_i	Q_i (mL/h)	r (mL/h)
0.0000	0.00	0.0000	25	0.00	25
0.0001	0.00	64.6706	25	922 785.1	25
0.0001	0.01	64.6711	25	7.10	25
0.0002	0.01	64.6723	25	17.10	25
0.0003	0.02	64.6738	25	21.36	25
0.0004	0.02	64.6742	25	9.15	25
0.0004	0.03	64.6750	25	11.35	25
0.0005	0.03	64.6768	25	24.29	25
0.0006	0.04	64.6769	25	1.44	25
0.0006	0.04	64.6778	25	12.81	25
0.0007	0.04	64.6787	25	12.86	25
0.0008	0.05	64.6812	25	35.65	25
0.0009	0.05	64.6825	25	18.48	25

(b) Acquisition at 300 samples per minute.

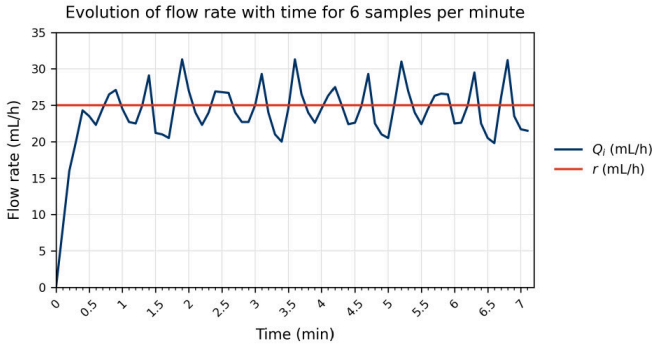


Fig. 5. Flow rate over time obtained with six samples per minute. Q_i represents the calculated flow rate, and r is the flow rate adjusted by the operator.

stores the measurements in spreadsheet format, including the acquisition time, mass measurements, programmed flow rate, and calculated flow rate.

The mass measurements are acquired directly from the analytical balance and timestamped by the acquisition software. The programmed infusion rate r is defined by the user, whereas the flow rate Q_i is calculated automatically according to Eq. (1).

After completion of the acquisition period, the spreadsheet is automatically saved and organized in a dedicated folder. This workflow reduces the amount of manual data handling required during infusion pump testing.

The software also allows the user to configure sampling rates according to the intended application. When a sampling rate different from the two samples per minute used in the IEC analysis is selected, the software can retain the complete data

set while also generating a data set adjusted to the standard sampling interval.

This feature allows the same acquisition to support both normative testing and exploratory research without requiring a new experiment.

C. Data Analysis

The calculations specified by IEC 60601-2-24:2012 were implemented directly in the software. These calculations generate the start-up graph, trumpet curves, and overall mean percentage flow errors.

The automated implementation reduces the possibility of transcription and spreadsheet-calculation errors and provides a consistent analysis procedure across tests.

An additional feature concerns the visualization of trumpet curves. IEC 60601-2-24:2012 specifies a fixed graphical range of $\pm 15\%$. However, experimental results may exceed this range, making part of the curve difficult to visualize.

For this reason, the software dynamically adjusts the vertical axis when the calculated errors exceed $\pm 15\%$. The scale is adjusted in increments of 5% according to the maximum absolute deviation.

When both $E_{p(\max)}$ and $E_{p(\min)}$ remain within $\pm 15\%$, the standard scale is maintained. When the deviation exceeds this range, the software expands the axis to the next appropriate multiple of 5%.

This adaptive visualization does not modify the calculated values; it only improves their graphical representation.

D. Measurement Uncertainty and Scope

The uncertainty analysis was intentionally limited to the gravimetric flow-rate measurement. The balance contribution was propagated to Q at T_1 and T_2 using the law of propagation of uncertainty, based on the expanded uncertainty reported in the calibration certificate.

The trumpet-curve analysis involves multiple overlapping observation windows. Since consecutive Qi values share mass measurements, their uncertainties cannot be treated as statistically independent without accounting for covariance. Furthermore, $E_{p(\max)}$ and $E_{p(\min)}$ are selected as extrema from a set of observation windows, introducing an additional statistical aspect to their uncertainty evaluation.

For these reasons, the present study does not report uncertainty values for the extremal trumpet-curve errors. This approach avoids assigning a simplified uncertainty value that would not adequately represent the correlation structure of the measurements and the statistical selection of the extrema.

A complete uncertainty evaluation of the trumpet-curve quantities may be addressed in future work using a covariance-based model and an appropriate statistical treatment of the extremum-selection process.

E. Software

The software was developed to simplify and accelerate infusion pump performance testing. The user can define important acquisition parameters, including the programmed flow rate, test duration, and sampling frequency.

The flexibility in sampling configuration was intentionally implemented to support future research applications in which higher temporal resolution may be desirable. Nevertheless, when the objective is strict compliance with the IEC 60601-2-24:2012 analysis procedure, the sampling configuration specified by the standard should be maintained.

The open-source nature of the software also allows researchers and clinical engineering laboratories to inspect, adapt, and extend the implementation according to their experimental requirements.

The source code is available at:
https://github.com/Mathis-moiss/Infusion-pumps-testing-software.

V. Conclusion

This work presented the development and experimental verification of an open-source software designed to automate the acquisition, processing, and analysis of infusion pump accuracy tests according to IEC 60601-2-24:2012.

The proposed system integrates communication with an analytical balance, automated data acquisition, flow-rate calculation, spreadsheet generation, and graphical visualization of start-up behavior and trumpet curves.

The implementation reduces manual data processing and provides a standardized workflow for infusion pump performance testing. The time saved by researchers using the program was approximately forty to sixty minutes per infusion pump tested, compared to the previous version of the program, which was built in LabVIEW, in addition to the convenience of being able to use the program on any computer. The software also allows flexible sampling configurations, making it suitable for both normative testing and exploratory research.

The current implementation is limited by the unidirectional communication with the analytical balance and by the dependence of the communication module on the specific balance model. Bidirectional communication, including remote tare and additional balance models, represents an important direction for future development.

The proposed software provides a practical tool for infusion pump performance verification and can support post-market surveillance activities in clinical engineering laboratories. The software is currently being used by the Normative Testing Division of the Biomedical Engineering Laboratory at the Escola Politécnica of the University of São Paulo (DEN–LEB/EPUSP).

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant No. 309464/2022-6, awarded to H. T. Moriya.

References

[1] A. Badnjević, M. Cifrek, R. Magjarević, and Z. Džemić, Eds., *Inspection of Medical Devices for Regulatory Purposes*, 2nd ed. Cham, Switzerland: Springer Nature Switzerland AG, 2024.

[2] C. Amaral, M. Paiva, A. R. Rodrigues, F. Veiga, and V. Bell, "Global regulatory challenges for medical devices: Impact on innovation and market access," *Applied Sciences*, vol. 14, no. 20, Art. no. 9304, 2024, doi: 10.3390/app14209304.

[3] U.S. Food and Drug Administration, "Postmarketing surveillance programs," 2020. [Online]. Available: https://www.fda.gov/drugs/surveillance/postmarketing-surveillance-programs. Accessed: Jun. 2025.

[4] European Union Medical Device Regulation, "Post-market surveillance system," 2017. [Online]. Available: https://eumdr.com/post-market-surveillance-system/. Accessed: Jun. 2025.

[5] ECRI Institute, *2019 Top 10 Health Technology Hazards: Executive Brief*. Plymouth Meeting, PA, USA: ECRI Institute, 2018.

[6] ECRI Institute, *Top 10 Health Technology Hazards for 2015: Key Safety Threats to Manage in the Coming Year*. Plymouth Meeting, PA, USA: ECRI Institute, Nov. 2014.

[7] A. N. Thomas and I. Galvin, "Patient safety incidents associated with equipment in critical care: A review of reports to the UK National Patient Safety Agency," *Anaesthesia*, vol. 63, no. 11, pp. 1193–1197, Nov. 2008, doi: 10.1111/j.1365-2044.2008.05607.x.

[8] A. F. van der Sluijs, E. R. van Slobbe-Bijlsma, A. Goossens, A. P. J. Vlaar, and D. A. Dongelmans, "Reducing errors in the administration of medication with infusion pumps in the intensive care department: A Lean approach," *SAGE Open Medicine*, vol. 7, 2019, Art. no. 2050312118822629, doi: 10.1177/2050312118822629.

[9] International Electrotechnical Commission, *Medical Electrical Equipment – Part 2-24: Particular Requirements for the Basic Safety and Essential Performance of Infusion Pumps and Controllers*, IEC 60601-2-24:2012, Geneva, Switzerland: IEC, 2012.

[10] M. D. S. Silva, J. L. Araújo, G. A. M. A. Nunes, M. F. F. Rosa, G. V. D. S. Luz, S. S. R. F. Rosa, and A. Piratelli-Filho, "Precision and reliability study of hospital infusion pumps: A systematic review," *BioMedical Engineering OnLine*, vol. 22, no. 1, Art. no. 26, Mar. 2023, doi: 10.1186/s12938-023-01088-w.

[11] Elsevier, "Syringe pump," *ScienceDirect Topics*. [Online]. Available: https://www.sciencedirect.com/topics/engineering/syringe-pump. Accessed: Jun. 2025.

[12] Pediatric Oncall, "Syringe pump," [Online]. Available: https://www.pediatriconcall.com/medical-equipment/syringes/3/syringe-pump/18. Accessed: Jun. 2025.

[13] A. Valentin *et al.*, "Errors in administration of parenteral drugs in intensive care units: Multinational prospective study," *BMJ*, vol. 338, Art. no. b814, Mar. 2009, doi: 10.1136/bmj.b814.

[14] A. A. M. Ibey, C. Ciarniello, and S. Gorelik, "Inadvertent overinfusion of norepinephrine using infusion pump loading dose," *Intensive and Critical Care Nursing*, vol. 31, no. 6, pp. 367–369, Dec. 2015, doi: 10.1016/j.iccn.2015.05.003.

[15] X. Gao *et al.*, "A hazard analysis of class I recalls of infusion pumps," *JMIR Human Factors*, vol. 6, no. 2, 2019, Art. no. e10366, doi: 10.2196/10366.

[16] Joint Committee for Guides in Metrology, *Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement*, JCGM 100:2008. Sèvres, France: BIPM, 2008.

Distribuição tecnológica e cobertura assistencial regional dos serviços de diagnóstico por imagem em Araguaína-TO

V. G. Barros[1], M. S. Roxo[1], É. C. Gomes[1]

1. Universidade Federal do Norte do Tocantins (UFNT), Centro de Ciências Integradas (CCI), Araguaína, TO, Brasil.
Emails: victor.barros@ufnt.edu.br, marcelo.roxo@ufnt.edu.br, erica.gomes@ufnt.edu.br

***Resumo*—O mapeamento de parques tecnológicos e o acompanhamento do ciclo de vida de equipamentos médicos constituem atividades estratégicas da Engenharia Clínica para subsidiar o planejamento em saúde. Este trabalho apresenta uma investigação de diagnóstico estrutural e exploratório baseado em bancos de dados administrativos e parâmetros normativos do parque de radiodiagnóstico e medicina nuclear em Araguaína, Tocantins, contrapondo a capacidade instalada declarada aos parâmetros geodemográficos de acessibilidade assistencial das Portarias SAS/MS nº 1.101/2002 e nº 1.631/2015. A metodologia baseou-se na coleta de dados quantitativos (dezembro/2025) via Cadastro Nacional de Estabelecimentos de Saúde (CNES), validados junto à Comissão Nacional de Energia Nuclear (CNEN) para a modalidade de Gama Câmara, além de indicadores demográficos do IBGE. Os resultados evidenciam superávit quantitativo de equipamentos declarados frente às estimativas teóricas de planejamento para a cidade. Contudo, a análise geográfica revela que o fluxo macroregional captado pelo polo ultrapassa os raios de cobertura física idealizados. No panorama estadual, Araguaína centraliza 17,50% do inventário declarado do Tocantins, destacando-se na alta complexidade, na qual retém 31,25% dos mamógrafos, 25,00% dos densitômetros e 24,07% dos tomógrafos computadorizados. Sob o aspecto regulatório, identificou-se uma assimetria de informações: o CNES registra quatro dispositivos de Gama Câmara no município, enquanto a CNEN autoriza a operação de apenas uma instalação. Com isso, conclui-se que os dados coletados evidenciam um aparente excedente quantitativo de equipamentos em relação à população municipal. No entanto, esse achado pode ocultar pressões sobre o sistema de saúde associadas à função de Araguaína como centro de referência macrorregional. Nesse sentido, recomenda-se que estudos futuros incorporem indicadores capazes de avaliar se o aparente superávit de equipamentos se traduz, efetivamente, em maior acesso da população aos serviços de saúde.**

***Palavras-chave*—Engenharia Clínica, Diagnóstico por Imagem, Planejamento em Saúde, Parâmetros Assistenciais.**

I. Introdução

A integração de métodos de diagnóstico por imagem é fundamental na prática médica contemporânea, possibilitando diagnósticos mais rápidos, precisos e a detecção precoce de diversas patologias, o que impacta diretamente na escolha terapêutica e na gestão de recursos em saúde [1], [2]. A disponibilidade e a distribuição adequada desses serviços, que utilizam tanto radiação ionizante (como raios-X, tomografia computadorizada, mamografia) quanto não ionizante (como ultrassonografia e ressonância magnética), são cruciais para garantir o acesso equitativo da população a cuidados de saúde de qualidade.

A relevância deste tema é validada por estudos que demonstram a radiologia diagnóstica como ferramenta indispensável na redução da mortalidade por doenças cerebrovasculares e neoplasias [3]. Nesse sentido, o mapeamento da distribuição geográfica e da infraestrutura declarada dos serviços de diagnóstico por imagem assume grande relevância para o planejamento em saúde pública. Autores demonstram a importância de conhecer a distribuição de equipamentos e a frequência de exames para analisar a adequação da oferta à demanda populacional, visto que esse tipo de levantamento permite identificar disparidades regionais, orientar investimentos, otimizar a alocação de recursos e avaliar a infraestrutura tecnológica disponível [4], [5].

No âmbito da Engenharia Biomédica, esse tipo de trabalho equipara-se a uma das principais atribuições do engenheiro clínico, que consiste em atuar diretamente na gestão do parque tecnológico institucional por meio do cadastro dos equipamentos existentes, mapeando informações essenciais como o custo de aquisição, o histórico técnico, a localização setorial, o registro de falhas e o tempo de utilização [6]. Ademais, a Engenharia Clínica acompanha o ciclo de vida dos equipamentos médico-assistenciais, atividade considerada determinante para garantir a segurança operacional e a otimização dos recursos hospitalares [7], [8].

Nesse sentido, estudos destinados ao mapeamento dos equipamentos de diagnóstico por imagem em regiões interioranas brasileiras constituem instrumentos relevantes para o planejamento da saúde pública, pois permitem caracterizar a distribuição territorial, a disponibilidade e a concentração do parque tecnológico. O levantamento sistemático do parque de diagnóstico por imagem fornece subsídios para reconhecer desigualdades de oferta, áreas potencialmente desassistidas, duplicidades de recursos e dependências intermunicipais. Constata-se, contudo, que a região Norte apresenta uma carência de tais estudos, em virtude da escassez de produção científica voltada para este campo de atuação, mesmo considerando o grande impacto social e econômico da matéria.

Diante desse panorama, a cidade de Araguaína, situada no Centro-Norte do estado do Tocantins, surge como um polo regional que transcende os limites municipais. Como sede

administrativa e centro de referência de alta complexidade, Araguaína recebe uma expressiva demanda por diagnósticos especializados proveniente de quase metade do território estadual, além de estados vizinhos [9]. Essa centralidade foi reforçada legalmente em 2026 com a institucionalização de sua Região Metropolitana, que estabeleceu uma estrutura de integração de larga escala abrangendo 65 municípios [10].

Portanto, o presente trabalho tem como objetivo realizar o levantamento e a classificação de dados referentes ao parque tecnológico de diagnóstico por imagem de Araguaína, avaliando sua infraestrutura face ao seu papel de polo na rede de saúde do Tocantins. Para tanto, realizou-se o mapeamento do contingente de equipamentos e instalações de Medicina Nuclear na modalidade de Gama Câmara em bases de dados federais — como o Cadastro Nacional de Estabelecimentos de Saúde (CNES) [11] e a Comissão Nacional de Energia Nuclear (CNEN) [12] —, seguido de uma análise comparativa entre os parâmetros estimativos teóricos demográficos e geográficos estabelecidos pelas Portarias SAS/MS nº 1.101/2002 e nº 1.631/2015 do Sistema Único de Saúde (SUS) [13], [14].

Cabe destacar que este trabalho não tem como objetivo interpretar os dados com base em indicadores como condição operacional, carga de trabalho, tempo de indisponibilidade, qualificação da equipe, infraestrutura física, manutenção e produção de exames. Tais aspectos deverão ser examinados em pesquisas posteriores, fundamentadas na análise espacial desenvolvida neste estudo. Desse modo, espera-se que a identificação realizada, em um contexto regional ainda pouco investigado, forneça subsídios iniciais para iniciativas voltadas ao desenvolvimento regional.

II. Objetivos

A. Objetivo Geral

Realizar uma análise da infraestrutura de serviços de diagnóstico por imagem na cidade de Araguaína, Tocantins, por meio do mapeamento sistematizado de parques tecnológicos declarados e regulamentados.

B. Objetivos Específicos

1) Mapear o inventário quantitativo e o perfil tecnológico dos dispositivos de diagnóstico por imagem declarados em Araguaína, contrapondo-os com o panorama geral do estado do Tocantins;
2) Aferir a adequação da capacidade instalada local frente às diretrizes geodemográficas das Portarias SAS/MS nº 1.101/2002 e nº 1.631/2015;
3) Avaliar a regularidade operacional das instalações de Medicina Nuclear com o equipamento Gama Câmara mediante o cruzamento de dados do CNES e da CNEN.

III. Materiais e Métodos

A presente investigação iniciou-se com um levantamento de dados quantitativos acerca dos equipamentos de diagnóstico por imagem e de medicina nuclear em Araguaína, estabelecendo uma análise comparativa com o panorama do estado do Tocantins. A coleta dos dados utilizou como fonte o Cadastro Nacional de Estabelecimentos de Saúde (CNES) [11], extraídos na competência de dezembro de 2025.

Visando correlacionar o número de equipamentos declarados pelas instituições de saúde com a realidade sociodemográfica a partir de parâmetros nacionais de cobertura assistencial, integraram-se indicadores extraídos do portal "Cidades" do Instituto Brasileiro de Geografia e Estatística (IBGE) [15]. Foram consideradas as projeções populacionais para o ano de 2025; contudo, para a análise da cobertura dos mamógrafos, utilizou-se o contingente de mulheres na faixa etária de 40 a 59 anos registrado no Censo de 2022, uma vez que não existem projeções oficiais para 2025 estratificadas simultaneamente por sexo e faixa etária no nível municipal. A utilização conjunta dessas bases, referentes a períodos distintos, constitui uma limitação metodológica decorrente da indisponibilidade de dados oficiais mais recentes com o nível de desagregação necessário.

Adicionalmente, realizou-se uma consulta ao guia de instalações autorizadas da Comissão Nacional de Energia Nuclear (CNEN) [12] para validar a regularidade dos serviços de Medicina Nuclear atuantes na região.

Para a classificação e sistematização do inventário, os dispositivos foram segregados nas categorias "Existentes", "Em Uso", "Existentes no SUS" e "Em uso no SUS", aplicando-se uma taxonomia adaptada de trabalhos anteriores [4]. Sob essa perspectiva, cumpre destacar que a classe "Existentes" representa o total de todo parque tecnológico mapeado, funcionando como o indicador bruto que engloba as demais categorias, as quais operam como subconjuntos analíticos. A classificação abrangeu as seguintes modalidades cadastradas no CNES:

- Raio-X: englobando equipamentos de "Raios X de até 100 mA", "Raios X de 100 a 500 mA" e "Raios X mais de 500 mA";
- Fluoroscopia: incluindo equipamentos de "Raios X com Fluoroscopia";
- Densitometria: incluindo equipamentos de "Raios X para Densitometria Óssea";
- Angiógrafo: incluindo equipamentos de "Raios X para Hemodinâmica";
- Mamógrafo: englobando aparelhos com "Mamógrafo com Comando Simples", " Mamógrafo com Estereotaxia" e "Mamógrafo Computadorizado";
- Ultrassonografia: englobando "Ultrassom Convencional", "Ultrassom Ecógrafo" e "Ultrassom Doppler colorido";
- Tomografia: incluindo equipamentos de "Tomografia Computadorizada";
- Ressonância: incluindo equipamentos de "Ressonância Magnética";
- Gama Câmara: dedicada à análise individualizada da infraestrutura de Medicina Nuclear, correlacionando os registros do CNES com as autorizações da CNEN.

Paralelamente, desenvolveu-se uma matriz de dados dedicada à estimativa teórica de equipamentos para o município de Araguaína conforme as diretrizes da Portaria SAS/MS nº 1.631/2015 [14] (para tomógrafo, ressonância magnética,

mamógrafo e ultrassom) e da Portaria SAS/MS nº 1.101/2002 [13] (para radiologia simples e densitometria óssea).

Sob a óptica metodológica, a Portaria SAS/MS nº 1.631/2015 representou um avanço substancial em relação à de 2002, visto que deixou de considerar apenas o cálculo numérico populacional para integrar critérios de regionalização e acessibilidade geográfica, tais como o raio de atendimento físico e o tempo de deslocamento espacial.

Todavia, a utilização combinada de ambas as normativas justifica-se pela ausência de parâmetros específicos para tecnologias de menor complexidade no documento de 2015. Assim, a incorporação da Portaria nº 1.101/2002 viabilizou a análise dos equipamentos de radiologia simples (Raios X) e da densitometria óssea, respaldando-se em sua relevância histórica e técnica. Para as modalidades não contempladas por ambos os documentos, não foram calculadas estimativas por ausência de parâmetros oficiais. Os modelos matemáticos e critérios geográficos aplicados para cada modalidade são descritos a seguir:

A. Tomógrafo Computadorizado

Conforme os parâmetros de planejamento da Portaria SAS/MS nº 1.631/2015 [14], o dimensionamento adota a modelagem demográfica de um equipamento para cada grupo de 100 mil habitantes, associado a um critério de regionalização que delimita um raio de atendimento de 75 km para o dispositivo. O modelo matemático é expresso por:

$$\mathrm{U}_{\text{Tomógrafo}} = \frac{\text{População Total}}{100.000} \tag{1}$$

B. Ressonância Magnética

Orientados pela Portaria SAS/MS nº 1.631/2015 [14], para os critérios de acessibilidade territorial impõe-se um raio de cobertura de até 30 km ou um tempo máximo de deslocamento de 60 minutos. Para a modelagem numérica, primeiramente é calculada a necessidade anual esperada de procedimentos (N_{RM}), considerando o indicador epidemiológico de 30 exames para cada 1.000 habitantes ao ano:

$$N_{\text{RM}} = \frac{\text{População Total} \times 30}{1.000} \tag{2}$$

A estimativa do contingente de equipamentos (U_{RM}) é obtida dividindo-se a demanda anual calculada pela produtividade média padrão estabelecida para o dispositivo, que é de 5.000 exames/ano:

$$U_{\text{RM}} = \frac{N_{\text{RM}}}{5.000} \tag{3}$$

C. Mamógrafo

De acordo com as diretrizes da Portaria SAS/MS nº 1.631/2015 [14], o planejamento geográfico dos equipamentos de mamografia estabelece uma distância de 60 km ou um tempo máximo de viagem de 60 minutos. Para a estimativa do número de equipamentos, inicialmente, efetua-se o cálculo do número de mamografias anuais necessárias (N_{M}) para a população de mulheres na faixa etária de 40 a 59 anos, conforme os seguintes critérios:

- Mulheres de 40 a 49 anos: 10% para indicação diagnóstica (D_1) e 10% para outras indicações (O_u);
- Mulheres de 50 a 59 anos: 50% para indicação de rastreamento (R_1) e $8,9\%$ para indicação diagnóstica (D_2).

$$N_{\text{M}} = D_1 + D_2 + R_1 + O_u \tag{4}$$

Assim, o número de equipamentos necessários ($U_{\text{Mamógrafo}}$) é determinado a partir da produtividade média anual de referência definida pela portaria, sendo 6.758 exames/ano:

$$U_{\text{Mamógrafo}} = \frac{N_{\text{M}}}{6.758} \tag{5}$$

D. Ultrassom

A estimativa fundamenta-se nos indicadores de produção da Portaria SAS/MS nº 1.631/2015 [14]. Diferente das modalidades citadas anteriormente, a normativa não estabelece um raio geográfico ou tempo de deslocamento padronizados para o ultrassom convencional. Desse modo, para a estimativa de equipamentos, primeiramente é calculada a necessidade anual de exames (N_{US}), que assume a razão de 150 exames para cada 1.000 habitantes ao ano:

$$N_{\text{US}} = \frac{\text{População Total} \times 150}{1.000} \tag{6}$$

Logo, a quantidade necessária teórica de equipamentos ($U_{\text{Ultrassom}}$) é calculada com base na produtividade anual de referência por dispositivo, estipulada em 3.024 exames:

$$U_{\text{Ultrassom}} = \frac{N_{\text{US}}}{3.024} \tag{7}$$

Nota-se que os parâmetros supracitados referem-se à modalidade de diagnóstico descrita como "Ultrassom Convencional", constando também no mesmo documento a categoria "Ultrassom Obstétrico" [14]. Porém, a base de dados do CNES estratifica os dispositivos de ultrassonografia em três categorias distintas: "Ultrassom convencional", "Ultrassom ecógrafo" e "Ultrassom com Doppler colorido". A aparente divergência de identificação do equipamento entre a Portaria SAS/MS nº 1.631/2015 [14] e a base de dados do CNES ocorre em função de abordagens distintas. Enquanto a Portaria [14] trata de parâmetros assistenciais e tipologia de exames, a base do CNES foca no inventário físico de equipamentos. Portanto, no presente trabalho, optou-se pela consolidação do contingente tecnológico em uma única categoria analítica de ultrassonografia, semelhantemente à metodologia adotada em levantamentos prévios [4].

E. Aparelho de Radiologia Simples (Raio-X)

O dimensionamento adota os critérios quantitativos da Portaria SAS/MS nº 1.101/2002 [13]. Por se tratar de uma normativa focada na relação demográfica, o documento não estabelece parâmetros geográficos de raio ou tempo de viagem. A proporção de cobertura estipula um equipamento para cada grupo de 25.000 habitantes, expressa por:

$$\mathrm{U}_{\text{Raio-X}} = \frac{\text{População Total}}{25.000} \tag{8}$$

F. Densitometria Óssea

Regulado de forma análoga pela Portaria SAS/MS nº 1.101/2002 [13], o planejamento quantitativo restringe-se a estimativas numéricas de capacidade instalada por volume populacional. A proporção estabelecida é de um dispositivo para cada 140.000 habitantes, portanto $U_{\text{Densitometria}}$ é obtido por:

$$\text{U}_{\text{Densitometria}} = \frac{\text{População Total}}{140.000} \tag{9}$$

Enfatiza-se que tais normativas constituem apenas um referencial teórico de planejamento, não correspondendo diretamente à capacidade operacional real, à produtividade efetiva dos equipamentos ou ao acesso geográfico concreto da localidade estudada.

Em suma, os dados de Araguaína foram consolidados e organizados para apresentação em quatro eixos analíticos sequenciais: o mapeamento quantitativo municipal e sua inserção no SUS, a análise comparativa frente ao contexto estadual, o confronto da capacidade declarada com os parâmetros assistenciais normativos do SUS e, por fim, a verificação do cadastro e da autorização dos equipamentos de Gama Câmara por meio do cruzamento das bases do CNES e da CNEN.

IV. Resultados e Discussão

Nesse tópico estão apresentados dados oriundos da investigação quantitativa realizada. Cabe destacar que tais resultados apresentam apenas um diagnóstico estrutural e exploratório baseado em bancos de dados administrativos e parâmetros normativos e não apresentam evidência de suficiência, equidade ou deficiência na assistência à saúde. Para conclusões a respeito da cobertura assistencial real, de gargalos e de necessidades tecnológicas efetivas é necessário a continuidade do estudo, com a expansão da análise e uso de dados reais como indicadores de produção, tempos de espera, origem dos pacientes, fluxos intermunicipais, dentre outros.

A. Mapeamento Quantitativo e Inserção no SUS

A Tabela I apresenta a distribuição quantitativa de todos os equipamentos de diagnóstico por imagem declarados em Araguaína na competência de dezembro de 2025, exceto os equipamentos de Medicina Nuclear - Gama Câmara - que serão analisados à parte, discriminados pelas categorias de existência e uso geral, bem como sua respectiva inserção no Sistema Único de Saúde (SUS).

Para fins de contextualização da composição tecnológica no município, a Fig. 1 foi gerada para segmentar a participação proporcional de cada modalidade diagnóstica frente ao montante de 172 dispositivos mapeados na base federal.

A análise combinada da Tabela I e da Fig. 1 expõe uma configuração tecnológica registrada fortemente dominada por tecnologias de média complexidade. A ultrassonografia assume a liderança correspondendo a 53,5% do inventário local (92 equipamentos), seguida pela radiologia convencional (Raio-X), que detém 22,1% (38 dispositivos). Somadas, ambas as modalidades representam mais de três quartos (75,6%) de toda a tecnologia de diagnóstico por imagem identificadas.

TABELA I
Número de Equipamentos de Diagnóstico por Imagem em Araguaína - TO (CNES, 2025)

Equipamento	Existentes	Totais em Uso	Existentes no SUS	Em Uso no SUS	% Em Uso SUS
Raio-X	38	35	27	24	68,57%
Fluoroscopia	1	1	1	1	100,00%
Densitometria	5	5	2	2	40,00%
Angiógrafo	2	2	1	1	50,00%
Mamógrafo	15	15	10	10	66,67%
Ultrassonografia	92	90	34	33	36,67%
Tomografia	13	13	9	9	69,23%
Ressonância	6	6	4	4	66,67%
Total	**172**	**167**	**88**	**84**	**-**
(%)	**100,00%**	**97,09%**	**51,16%**	**48,84%**	**50,30%**

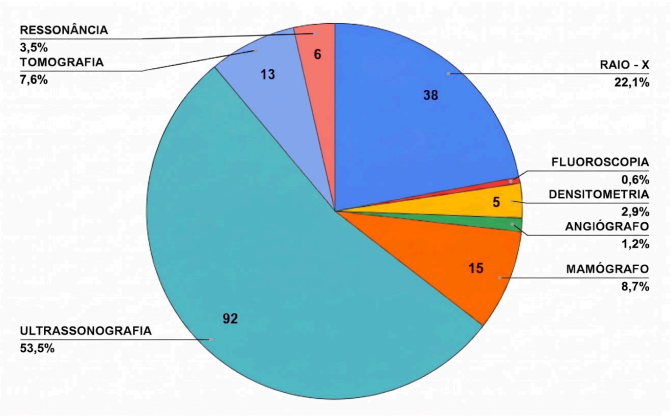


Fig. 1. Segmentos de equipamentos existentes no município de Araguaína - TO (CNES, 2025).

Em uma segunda análise, a Fig. 2 apresenta o quantitativo de equipamentos registrados que atende ao SUS em comparação com o total documentado. Esse comparativo revela a capacidade assistencial oferecida à população dependente da rede pública. O mapeamento dessas diferenças quantitativas constitui o passo inicial para a determinação da real cobertura assistencial. Sob a perspectiva da gestão pública de saúde, a mensuração do inventário tecnológico destinado ao SUS fundamenta as tomadas de decisão alinhadas aos parâmetros das Portarias SAS/MS nº 1.101/2002 [13] e SAS/MS nº 1.631/2015 [14].

Os resultados dispostos na Tabela I e graficamente mapeados na Fig. 2 revelam uma diferença no aproveitamento público de algumas tecnologias. Embora Araguaína possua um total de 172 equipamentos — dos quais 167 encontram-se declarados em uso efetivo (97,09%) —, pouco menos da metade do parque tecnológico atende diretamente à rede pública (48,84% em uso no SUS).

A maior assimetria localiza-se na ultrassonografia, onde, dos 90 dispositivos mapeados na cidade, apenas 33 realizam exames pelo SUS, representando uma retenção de 63,33% da capacidade para esse tipo de diagnóstico. Cenário similar observa-se na densitometria óssea, cuja vinculação pública é

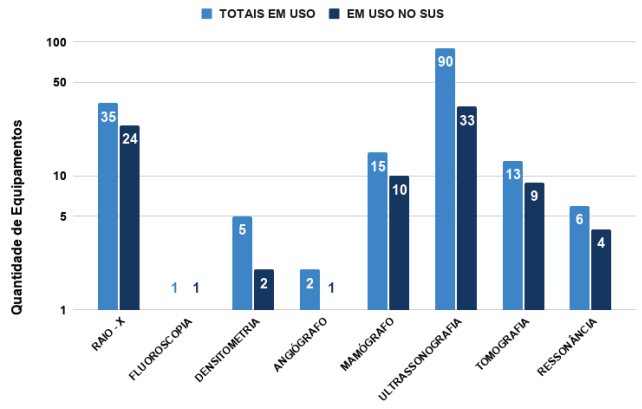


Fig. 2. Comparativo quantitativo de equipamentos em uso geral e em uso ativo na rede SUS em Araguaína - TO. (CNES, 2025)

de apenas 40,00%. Esse arranjo pode indicar um entrave no acesso público ao exame de densitometria óssea, indicativo que será melhor investigado na continuidade desse estudo.

Analisando isoladamente as demais modalidades, evidencia-se uma taxa de inserção no SUS superior a 60% para Raio-X, Mamógrafo, Tomografia e Ressonância. A Fluoroscopia se destaca com uma vinculação pública de 100% dos equipamentos identificados, contudo tal cenário pode mascarar uma vulnerabilidade estrutural, dada a existência de apenas um único aparelho de Fluoroscopia declarado em toda a extensão municipal.

Além disso, tendo em vista que o município se consolida como um polo de saúde macroregional, a aparente equidade observada nesses eixos de maior complexidade deve ser interpretada com cautela. A dispersão geográfica da demanda, que atrai pacientes de toda a região e de estados limítrofes [9], aponta para uma possível pressão de uso sobre a rede assistencial local, indicando a necessidade de maiores investigações acerca da demanda e fluxo intermunicipal.

B. Análise Comparativa Frente ao Contexto Regional

Para contextualizar a centralidade de Araguaína como polo de saúde, a Tabela II expõe a participação percentual do município em relação ao total de equipamentos registrados em todo o estado do Tocantins.

Com o intuito de visualizar essa dispersão, a Fig. 3 apresenta a composição percentual da capacidade instalada entre o polo municipal estudado e os demais municípios do território estadual.

Sob a perspectiva dos dados, a interpretação conjunta da Tabela II e da Fig. 3 contesta, à primeira vista, a expectativa para um polo de saúde, visto que Araguaína concentra apenas 17,50% do total geral de equipamentos de imagem inventariados do estado (172 de 983). Esse percentual relativamente baixo reflete as características geopolíticas do Tocantins: um estado jovem, de vasta extensão territorial e com uma rede urbana fracionada. Tal configuração geográfica impõe a necessidade de descentralizar dispositivos de menor custo

TABELA II
PARTICIPAÇÃO PERCENTUAL DE ARAGUAÍNA FRENTE AO INVENTÁRIO TECNOLÓGICO DO TOCANTINS (CNES, 2025)

Equipamento	Araguaína	Demais Municípios	Total no TO	% de Araguaína
Raio X	38	176	214	17,76%
Fluoroscopia	1	12	13	7,69%
Densitometria	5	15	20	25,00%
Angiógrafo	2	17	19	10,53%
Mamógrafo	15	33	48	31,25%
Ultrassom	92	487	579	15,89%
Tomografia	13	41	54	24,07%
Ressonância	6	30	36	16,67%
Total	**172**	**811**	**983**	**17,50%**

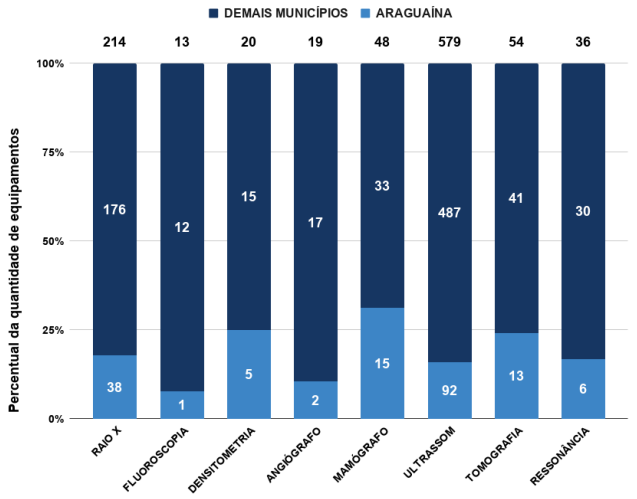


Fig. 3. Fração percentual e quantitativa de equipamentos existentes: Araguaína frente aos demais municípios do Tocantins (CNES, 2025).

— como Raio-X convencional e ultrassons, equipamentos mais numerosos no estado — pelas demais municipalidades, visando garantir a cobertura da atenção primária e evitar vazios assistenciais básicos.

Em contrapartida, o papel de Araguaína como polo regional se manifesta quando a análise é segmentada pelo nível de complexidade tecnológica. O município centraliza quase um terço dos mamógrafos estaduais (31,25%), um quarto de todos os densitômetros ósseos (25,00%), cerca de um quarto dos tomógrafos computadorizados (24,07%) e 16,67% dos equipamentos de ressonância magnética do Tocantins. Desse modo, esses índices revelam que as decisões de investimento tendem a se aglutinar no município quando se trata de alta densidade tecnológica.

Embora, em uma análise inicial, essa centralização indique uma otimização da viabilidade de operação, ela dificulta o acesso dos usuários de cidades periféricas que necessitam de diagnósticos especializados, impondo a necessidade de grandes deslocamentos. A análise espacial dessas distâncias geográficas, fundamentada nos raios normativos de acessibilidade dos equipamentos, será aprofundada na Seção C.

Diante disso, a aparente limitação do índice geral de 17,50% evidencia que a análise isolada dos limites político-administrativos da cidade mostra-se insuficiente para compreender a capacidade tecnológica local. Esse resultado atesta a necessidade de expandir o escopo analítico para além das fronteiras municipais, incorporando a dinâmica integrada de sua Região Metropolitana, instituída por lei estadual [10], ou de sua Região Geográfica Intermediária estabelecida pelo IBGE [16].

C. Aferição da Realidade Declarada com Parâmetros Assistenciais do SUS

A partir da aplicação dos modelos matemáticos determinados pelas Portarias SAS/MS nº 1.101/2002 [13] e nº 1.631/2015 [14], calculou-se a necessidade teórica de dispositivos para atender a população residente de Araguaína, estimada pelo IBGE em 183.024 habitantes para o ano de 2025 [15]. Pondera-se que o cálculo para os equipamentos de mamografia fundamentou-se especificamente no contingente de mulheres na faixa etária de 40 a 59 anos, compreendendo 12.470 mulheres entre 40 e 49 anos e 8.845 entre 50 e 59 anos, conforme o último censo de 2022 feito pelo IBGE [15].

Logo, a Tabela III confronta a necessidade normativa com o número total de equipamentos existentes no município, evidenciando um aparente cenário de superávit geral.

TABELA III
NÚMERO DE EQUIPAMENTOS ESTIMADOS A PARTIR DOS PARÂMETROS DE PLANEJAMENTO DO SUS DAS PORTARIAS DE 2002 E 2015

Equipamento	**N. Exames/Ano**	**N. Equip. Estimados**	**N. Equip. Existentes**
Tomógrafo	-	1,83	13
Ressonância	5490,72	1,10	6
Mamografia	7703,71	1,14	15
Ultrassom	27453,60	9,08	92
Raio X	-	7,32	38
Densitometria	-	1,31	5

Sob a óptica dos indicadores de planejamento, os dados dispostos na Tabela III revelam que o quantitativo de equipamentos ultrapassa as metas teóricas recomendadas pelas normativas do SUS para a população local. Uma ênfase maior se dá na ultrassonografia, cujo total existente (92 dispositivos) é mais de dez vezes superior à necessidade calculada (≈ 9). Variações semelhantes ocorrem nas modalidades de tomografia computadorizada (13 existentes frente a uma meta de ≈ 2) e na ressonância magnética (6 existentes frente a apenas ≈ 1 requerido). Esse fenômeno evidencia que as diretrizes baseadas estritamente na demografia não refletem a dinâmica real de mercado e de alocação de recursos tecnológicos em municípios que atuam como polos de atenção regional.

Diante desse cenário, o superávit observado corrobora as conclusões de estudos anteriores [9] de que Araguaína atua como o principal centro de referência assistencial da rede hospitalar com um raio de até 260 km de alcance. Configura-se, portanto, como um polo de saúde cuja infraestrutura atende a uma demanda que ultrapassa sua população residente e seus limites político-administrativos.

Sob a perspectiva da Portaria SAS/MS nº 1.631/2015 [14], o planejamento prevê raios de atendimento de 75 km para tomografia, 60 km para mamografia e 30 km para ressonância magnética. Para permitir uma visualização comparativa da abrangência do polo, dos limites normativos teóricos e das distâncias geográficas, a Fig. 4 foi elaborada no software QGIS [17], sobrepondo os raios de cobertura assistencial e de alcance da rede hospitalar à malha territorial da Região Metropolitana de Araguaína, composta por 65 municípios do norte tocantinense [10].

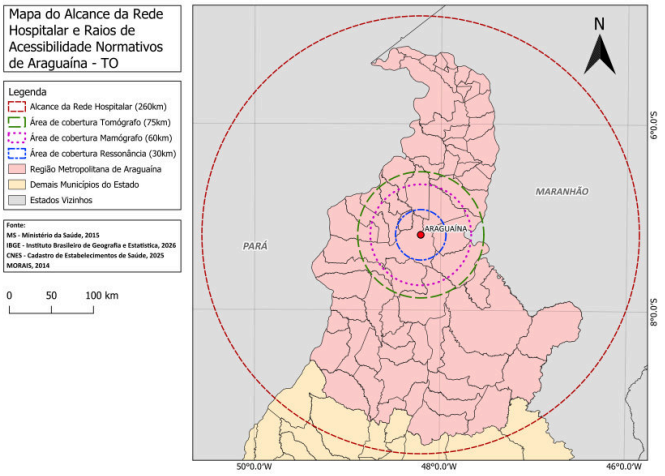


Fig. 4. Comparativo espacial entre os raios de cobertura normativos e o alcance real da rede hospitalar na Região Metropolitana de Araguaína-TO.

A análise da Fig. 4, em conjunto com os dados quantitativos discutidos nesta seção, permite concluir que o fato de o fluxo real captado pelo polo ultrapassar os raios de cobertura previstos normativamente atesta o fenômeno de concentração tecnológica municipal. Ademais, sugere que o município atua como retaguarda de alta complexidade para a região, embora a necessidade de grandes deslocamentos imponha possíveis restrições de acessibilidade aos serviços. Paralelamente, do ponto de vista mercadológico e de gestão, essa centralização indica o potencial econômico de investimentos e alocação de serviços na cidade.

D. Análise Comparativa CNES x CNEN para a Modalidade de Gama Câmara

Como eixo final desta análise, a Tabela IV apresenta o cruzamento de dados entre os diferentes sistemas regulatórios federais (CNES e CNEN) para o dispositivo de Gama Câmara, revelando uma assimetria nas informações cadastradas no âmbito da Medicina Nuclear.

A partir desse panorama, evidencia-se uma incoerência entre os registros do CNES e as autorizações vigentes da CNEN. Enquanto o sistema do Ministério da Saúde aponta a existência de 9 equipamentos de Gama Câmara no estado do Tocantins — sendo 4 deles localizados em Araguaína —, a base de dados da

TABELA IV
MAPEAMENTO COMPARATIVO DAS INSTALAÇÕES DE GAMA CÂMARA NAS BASES CNES E CNEN

Base de Dados	Araguaína	Demais Municípios	Total no Tocantins
Declarados pelo CNES	4	5	9
Autorizados pelo CNEN	1	1	2

Comissão Nacional de Energia Nuclear reconhece a operação de apenas duas instalações em todo o território estadual, com cada uma possuindo um equipamento, reduzindo o contingente do polo municipal a um único dispositivo autorizado.

Essa divergência estatística pode indicar diversas ocorrências, incluindo a existência de dispositivos em trâmite de certificação junto à CNEN. Contudo, sem uma validação em campo, essa diferença não deve ser interpretada automaticamente como indício de operação irregular, necessitando de investigações adicionais.

Desse modo, entende-se que assimetria de informações afeta de forma direta a gestão de tecnologias em saúde e a vigilância radiológica na região. Como a Medicina Nuclear lida com radioisótopos e exige que cada Gama Câmara ativa esteja vinculada a um serviço rigorosamente licenciado — envolvendo radioproteção, blindagem e gerência de rejeitos —, a inconsistência cadastral no CNES tem o potencial de comprometer o planejamento em saúde pública e ocasionar imprecisões na avaliação da distribuição de serviços no norte tocantinense. Reforça-se, portanto, a necessidade de mecanismos de interoperabilidade entre as bases federais, o que permitiria distinguir com clareza os equipamentos inoperantes ou obsoletos daqueles em processo de regularização normativa.

V. CONCLUSÃO

O trabalho apresentou uma investigação estrutural do parque tecnológico de diagnóstico por imagem e Medicina Nuclear no município de Araguaína, TO, fundamentada nos equipamentos declarados e autorizados nos bancos de dados do CNES e da CNEN, evidenciando as complexas dinâmicas entre a capacidade instalada e a cobertura assistencial teórica e efetiva.

Os resultados demonstram que, embora o município exiba um maior número de equipamentos em relação aos parâmetros demográficos teóricos recomendados pelas normativas do SUS para a população residente, potenciais gargalos de acesso para dependentes da rede pública podem estar ocultos e necessitam ser investigados por meio de dados adicionais sobre produção, demanda, deslocamento e utilização dos dispositivos.

Esse cenário decorre, principalmente, devido ao fato de a cidade se consolidar como um polo de saúde macrorregional, atraindo pacientes de uma rede de usuários distribuída em um raio de até 260 km. Essa dinâmica analítica ganha consistência ao se aplicar a abordagem geodemográfica, evidenciando que a demanda real captada pelo polo extrapola os raios de cobertura física e tempos de deslocamento previstos teoricamente para algumas modalidades.

Ademais, o confronto dos dados locais com os números estaduais revela que, embora Araguaína concentre apenas 17,50% do inventário mapeado do Tocantins, esse índice é atenuado pela possível dispersão de dispositivos de média complexidade em municípios vizinhos integrantes de seu fluxo territorial. Em contrapartida, o município se destaca na alta complexidade, com a segregação por densidade tecnológica demonstrando uma centralização de modalidades estratégicas, nas quais o polo retém até 31,25% da capacidade do estado.

No âmbito regulatório, a assimetria de informações identificada nos registros de Gama Câmara expôs uma lacuna na interoperabilidade entre diferentes bases de dados. Este panorama reforça a necessidade de auditorias cadastrais capazes de diferenciar erros operacionais no sistema e eventuais processos pendentes para licenciamento junto à CNEN, garantindo a fidedignidade dos registros nacionais e do próprio planejamento em saúde.

Por fim, ressalta-se que este estudo representa um diagnóstico estrutural inicial restrito à distribuição territorial e baseado exclusivamente em bancos de dados secundários e parâmetros normativos. Embora o objetivo de elaborar um inventário territorializado tenha sido alcançado, a ausência de validação em campo e a não contemplação de indicadores operacionais impedem afirmações definitivas sobre gargalos e necessidades tecnológicas efetivas. Desse modo, como desdobramento futuro, propõe-se a expansão do escopo analítico para o estado do Tocantins, aliada à apuração de dados reais de produção, tempo de indisponibilidade, tempos de espera, origem dos pacientes e fluxos territoriais intermunicipais. A incorporação de tais parâmetros é fundamental para confirmar se o aparente superávit de equipamentos mapeado se traduz, de fato, em um maior e efetivo acesso da população aos serviços de saúde.

VI. AGRADECIMENTOS

Agradecemos à Universidade Federal do Norte do Tocantins (UFNT) pelo apoio institucional e pelo incentivo à pesquisa científica. O desenvolvimento desta pesquisa contou com Apoio à Pesquisa e a Graduação da PROPI/UFNT, Edital n° 47/2026. Este trabalho foi desenvolvido no âmbito do Programa Institucional Voluntário de Iniciação Científica (PIVIC/UFNT), ao qual expressamos nosso reconhecimento pela contribuição à formação acadêmica e científica do estudante envolvido.

REFERÊNCIAS

[1] A. R. Hallak, J. P. B. Schmitt, G. O. L. Silva, e L. F. N. Sousa, "A importância da radiologia no contexto do Sistema Único de Saúde (SUS): Uma revisão da literatura," *Res., Soc. Develop.*, vol. 12, no. 13, p. e54121344176, dez. 2023.

[2] L. R. Goodman, *Princípios de radiologia torácica de Felson*, 3. ed. Rio de Janeiro, Brasil: Elsevier, 2016.

[3] C. Borrás, "The role of diagnostic and therapeutic radiology in the field of public health," *Rev. Panam. Salud Pública*, vol. 20, no. 2-3, pp. 84–86, ago. 2006.

[4] M. B. Freitas e E. M. Yoshimura, "Levantamento da distribuição de equipamentos de diagnóstico por imagem no Estado de São Paulo," *Radiol. Bras.*, vol. 38, no. 5, pp. 347–354, out. 2005.

[5] L. Pozzo, A. L. A. Marinho, e A. S. Lodovici, "O SUS na medicina nuclear do Brasil: avaliação e comparação dos dados fornecidos pelo Datasus e CNEN," *Radiol. Bras.*, vol. 47, no. 3, pp. 141–148, jun. 2014.

[6] L. B. W. Jaskulski, T. Lanzarini, C. Kasmirscki, e D. A. Cicolella, "Inovações tecnológicas na saúde: o papel da engenharia clínica em ambiente hospitalar," *Rev. Cuidado Enferm. – CESUCA*, vol. 4, no. 1, pp. 47–58, jan. 2018.

[7] F. B. Ferreira, "A Engenharia Clínica na Avaliação de Tecnologia em Saúde – Equipamentos Médico-Assistenciais na fase de utilização do ciclo de vida," Dissertação de Mestrado, Fac. Ciências e Tecnologia, Univ. Nova de Lisboa, Lisboa, Portugal, 2013.

[8] J. C. de Araújo Júnior, "Análise e proposta de melhoria no processo de aquisição de equipamentos médicos: uma abordagem baseada na engenharia clínica," Trabalho de Conclusão de Curso, Dept. Engenharia Biomédica, Univ. Federal do Rio Grande do Norte, Natal, RN, Brasil, 2025.

[9] I. A. Morais, "Araguaína (TO): enquanto cidade média no contexto regional," Dissertação de Mestrado, Dept. Geografia, UnB, Brasília, Brasil, 2014.

[10] Tocantins. Lei Complementar nº 172, de 11 de fevereiro de 2026. *Diário Oficial do Estado do Tocantins*, Palmas, TO, no. 6998, pp. 1–5, 11 fev. 2026.

[11] Ministério da Saúde, "Cadastro Nacional de Estabelecimentos de Saúde (CNES)," Departamento de Informática do SUS (DATASUS), Brasília, DF, Brasil, 2025. [Online]. Available: https://cnes.datasus.gov.br/, accessed Jun. 3, 2026.

[12] Comissão Nacional de Energia Nuclear (CNEN), "Instalações Autorizadas," Rio de Janeiro, RJ, Brasil. [Online]. Available: https://appasp2019.cnen.gov.br/seguranca/cons-ent-prof/entidades-aut-cert.asp, accessed Jun. 3, 2026.

[13] Brasil. Ministério da Saúde, Secretaria de Assistência à Saúde, "Portaria SAS/MS nº 1.101, de 12 de junho de 2002. Estabelece os parâmetros de cobertura assistencial no âmbito do SUS," Brasília, DF, 2002. [Online]. Available: [https://bvsms.saude.gov.br/bvs/saudelegis/gm/2002/prt1101_12_06_2002.html], accessed Jun. 3, 2026.

[14] Brasil. Ministério da Saúde, Secretaria de Atenção à Saúde, "Portaria SAS/MS nº 1.631, de 1º de outubro de 2015. Aprova os critérios e parâmetros para o planejamento e programação de ações e serviços de saúde," Brasília, DF, 2015. [Online]. Available: [https://bvsms.saude.gov.br/bvs/saudelegis/gm/2015/prt1631_01_10_2015.html], accessed Jun. 3, 2026.

[15] Instituto Brasileiro de Geografia e Estatística (IBGE), "Portal Cidades" 2025. [Online]. Available: [https://cidades.ibge.gov.br/], accessed Jun. 3, 2026.

[16] Instituto Brasileiro de Geografia e Estatística (IBGE), *Divisão regional do Brasil em regiões geográficas imediatas e regiões geográficas intermediárias: 2017*, Rio de Janeiro, RJ, Brasil: IBGE, 2017, 82 p.

[17] QGIS Development Team, "QGIS Geographic Information System," Open Source Geospatial Foundation Project, 2026. [Online]. Available: https://www.qgis.org/, accessed Sep. 5, 2026.

Electrical Energy Consumption Monitoring System for Public Health Facilities: the case of Child and Adolescent Unit (UCA) at University Hospital of Brasilia (HUB-UnB/HU Brasil)

João Marcus Ferreira B da C Lima
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
0009-0005-3743-9687

Jorge Cormane
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
0000-0001-5263-0985

André Bisinoto Matias
Infrastructure dept.
HU Brasil
Brasilia, Brazil
andre.matias@hubrasil.gov.br

***Abstract*—Electrical energy consumption in Brazilian public hospitals represents a growing challenge to operational sustainability, due to their uninterrupted operation, high demand for air conditioning, and the use of high-power medical equipment. Although initiatives such as migration to the Free Energy Market contribute to cost reduction, the complexity of these environments demands more comprehensive energy management approaches. This study proposes an integrated methodology for monitoring, forecasting, and managing electrical energy consumption at the University Hospital of Brasília (HUB-UnB/HU Brasil), grounded in systematic data collection, Artificial Intelligence techniques, and data-driven strategies. The methodology encompasses a technical survey of the electrical installations, data acquisition and preprocessing, and predictive modeling of time series using statistical and machine learning methods. Expected outcomes include predictive models evaluated by RMSE, MAE, and R^2 metrics, identification of consumption patterns in critical hospital sectors, and data-driven energy efficiency strategies replicable across other HU Brasil network units, contributing to the digital transformation of energy management in the Brazilian hospital sector.**

***Index Terms*—Energy Efficiency, Public Hospital, University Hospital, Healthcare Facility**

I. Introduction

High energy consumption systems from continuous operations involving lighting, heating, ventilation, and air conditioning loads, as well as imaging equipment, operating rooms, and other facilities that must remain available at all times. In Brazil, although most federal university hospitals belong to a single centralized Federal Government network, HU Brasil, there are federal university hospitals with independent administrations, managed directly by their respective universities. These public hospitals simultaneously perform teaching, research, and healthcare functions. They also face the challenge of maintaining patient care amidst budget constraints, as they operate exclusively for the Unified Health System (Sistema Único de Saúde - SUS).

Aspects related to power quality and the quality of the electric power supply in hospital environments have direct implications for patient safety. Failures or instabilities in the power supply network can compromise the operation of life-support equipment as well as the infrastructure in critical areas, such as intensive care units and operating rooms. In this context, continuous monitoring of the electrical infrastructure is not merely a matter of energy efficiency but a requirement with significant operational impact.

The Child and Adolescent Unit (UCA) is the most recent building at the University of Brasília Hospital (HUB-UnB/HU Brasil); it has been in operation since 2023 but was officially inaugurated in 2024 and is managed by HU Brasil [11]. The building was designed to integrate a system for measuring electrical parameters with a photovoltaic generation system. This monitoring infrastructure features 21 installed electronic meters and a dedicated electrical substation, creating a favorable environment for implementing a supervisory system based on open SCADA technologies.

This paper presents a survey of the UCA's electrical monitoring infrastructure and a description of the healthcare services provided on each floor of the building, as well as a proposal for a continuous monitoring system based on the open-source SCADA-LTS software platform. Preliminary results from data collected at measurement points distributed throughout the unit are also presented. The technological development activities described herein aim to integrate the metering infrastructure with the implementation of a supervisory system and the design of hospital energy management models driven by energy consumption data.

II. General Description of Children and Adolescents Unit (UCA)

The UCA building has a capacity of 59 active beds and 80 specialized consultation rooms, distributed across a total area of 6,930 m². The care services provided include a pediatric ward, pediatric ICU, neonatal ICU, rooming-in unit, laboratories, and teaching areas. In total, 61 specialties and subspecialties are available: 37 focused on adult care and

24 dedicated to children and adolescents [11]. The building comprises four levels: a basement, a ground floor, and two upper floors. Each with a specific care-related purpose, as described in Table I.

TABLE I
DISTRIBUTION OF HEALTHCARE SERVICES OFFERED AT THE UCA BUILDING

Floor	Purpose
Basement	Adult Outpatient Clinic
Ground Floor	Outpatient Clinic and Pediatric Rehabilitation
1st floor	Pediatric Ward and Classrooms
2nd floor	Neonatal ICU (NICU), Pediatric ICU (PICU), Maternity Rooming-In Care

III. UCA Electrical Infrastructure

A. Distribution Topology

The UCA is supplied by a substation shared with two other buildings at the HUB-UnB/HU Brasil hospital complex. This substation houses the supply transformers for all three buildings, as well as the emergency generator sets serving the entire hospital, making it a strategic location within the complex's electrical network. HUB-UnB/HU Brasil is classified as a Group A4 consumer, with a supply voltage ranging from 2.3 kV to 25 kV, under the "Blue" three-phase tariff structure (time-of-use and seasonal pricing), operating within the Free Energy Market (Ambiente de Contratação Livre - ACL). Contracted demand levels are 1,090 kW during peak hours and 1,280 kW during off-peak hours [12]. Exceeding these limits results in significant additional charges on the energy bill, making the monitoring and forecasting of demand peaks a matter of direct economic importance.

B. Metering System

The UCA is equipped with 21 WEG electric meters (model MW03-CH [1]) located in distribution panels throughout the building. The MW03-CH is a multi-function energy meter that records electrical parameters such as voltage, current, active power, reactive power, power factor, and accumulated energy, among others. The meter's resolution is suitable for energy management applications in medium- and large-scale facilities.

Communication with the meters takes place via the Modbus TCP protocol over Ethernet and the RS485 serial interface; both are widely supported by industrial monitoring, supervisory, and energy management platforms [9]. The use of these industrial communication protocol standards is crucial to the strategy adopted in this work: fully leveraging the existing metering infrastructure without the need to replace equipment or adopt different protocols.

Due to healthcare service demands, patient care provided on each floor of the building is subject to frequent changes. Thus, it was not possible to associate the distribution of electrical panels with the type of care provided within the coverage area of each circuit. Consequently, it was only possible to determine which meters monitor each floor, as shown in Table II. It can be observed that the meter located in the rooftop panel corresponds to the building's photovoltaic solar generation system, while panels QD-AC-1, QD-AC-2, QD-AC-3, and QD-AC-4 contain the meters associated with the HVAC system. The remaining meters serve the other installed loads.

TABLE II
LOCATION AND IDENTIFICATION OF UCA METERS

Floor	Meter
Basement	QD-S1, QD-S2, QD-S3, QD-S4, QD-S5
Ground Floor	QD-T1, QD-T2, QD-T3, QD-T4, QD-T5, QD-T6
1st floor	QD-1-1, QD-1-2, QD-1-3, QD-AC-1, QD-AC-2, QD-AC-3
2nd floor	QD-2-1, QD-2-2, QD-AC-4
Rooftop	QDG-IT

C. Supervision System

The UCA has a commercial SCADA system responsible for integrating the meters into a single centralized platform. However, the SCADA system is currently inoperative. One of the purposes of this work is the recovery of the SCADA system through the use of the existing infrastructure, including all physical metering infrastructure, communication cabling, and data network, which has been fully preserved. This condition creates an opportunity to implement an alternative supervisory solution without the need for new hardware investments, leveraging the assets already installed. The adoption of an open-source platform, in addition to eliminating dependency on specific vendors, reduces the risk of discontinuation that compromised the legacy system.

IV. Proposed Supervisory System

A. SCADA-LTS Platform

The solution adopted to restore electrical monitoring at the UCA is SCADA-LTS, an open-source supervisory and data acquisition platform built on the Mango Automation architecture. The system runs in a web environment, operating on a Java server with a relational database, and exposes the collected data through a REST API, enabling integration with external analysis and visualization tools without dependence on proprietary software.

The choice of SCADA-LTS is based on three objective criteria. The first is native compatibility with the Modbus TCP protocol used by the WEG MW03-CH meters installed at the UCA, eliminating the need for protocol converters or hardware adaptations. The second is the absence of licensing costs, a factor in the context of a public hospital. The third important is long-term technical sustainability: being open source, the system can be maintained, modified, and expanded by the hospital's own technical team, removing dependence on an external vendor.

B. Architecture

The proposed architecture for the UCA monitoring system consists of three functional layers, illustrated in Fig. 1.

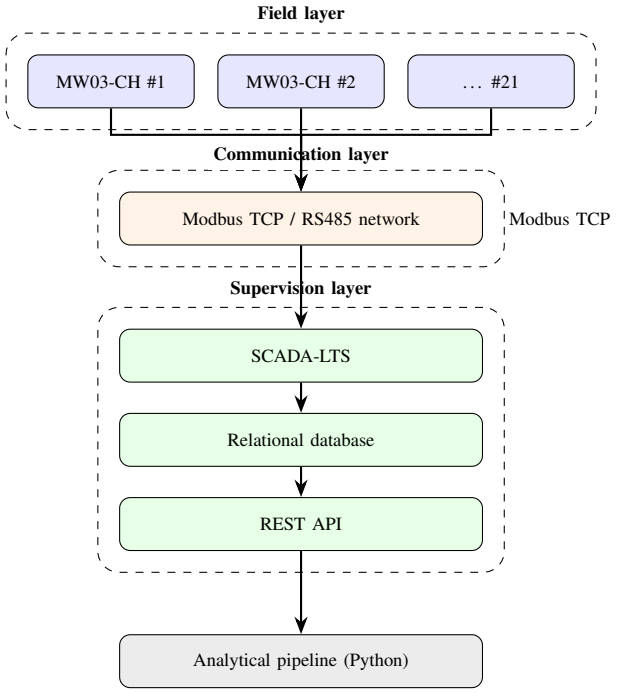


Fig. 1. Three-layer architecture of the proposed supervisory system for the UCA.

The field layer is formed by the 21 WEG MW03-CH [1] meters distributed across the distribution panels. Each UCA's meter operates as a Modbus slave, responding to requests from the supervisory server and providing records of electrical quantities.

The communication layer is formed by the data network already installed in the building, through which Modbus TCP traffic is routed between the meters and the SCADA-LTS server. The RS485 interface available on the meters is used at points where Ethernet communication is not available, routing this data to an RS485/Ethernet converter installed in the main panel on the UCA's ground floor.

The supervision layer is the SCADA-LTS server, responsible for periodically polling the meters, storing the data in a relational database, and making the historical series available through the REST API. The temporal resolution of acquisition is configurable per measurement point; for this work, a 15-minute interval is adopted as the minimum granularity, consistent with the resolution used in similar studies in the literature [2] and sufficient for detecting intraday demand patterns.

C. Data Pipeline and Analytical Integration

The data collected by SCADA-LTS are stored with a *timestamp* and a measurement point identifier, forming continuous time series for each monitored variable. Data extraction for analysis is performed via a REST API, which returns records in JSON format, that are processed directly by Python scripts within the analytical pipeline.

The data treatment pipeline comprises the following steps: identification and interpolation of communication failures; removal of documented maintenance periods; normalization of quantitative variables; and encoding of categorical temporal variables, such as day type and tariff period, are both described in [2]. The treated data feed, the demand forecasting models and the continuous monitoring system based on the CUSUM (*Cumulative Sum of Differences*) method, whose integration with SCADA-LTS constitutes the long-term goal of this work.

V. Results

A. Supervision Dashboard

Fig. 2 illustrates the main monitoring dashboard developed on the SCADA-LTS platform. The interface was designed to consolidate the electrical quantities collected by the 21 meters installed in the UCA (University Care/Health Unit — adjust acronym as needed). The meters were organized by floor, as described in Table II.

The main panel shows the total power of the whole hospital and allows to chose the building for a specific monitoring. Showing the total power of the building, the tension and the frequency measured, showing the limits established for a good energy quality, according to the regulatory standards [13].

In parallel with the integration between SCADA-LTS and the meters, a portable meter was installed at the substation as a strategy for collecting time-series consumption data, aiming at the construction of future data logging databases. This approach enables continuous recording of the electrical quantities across all distribution panels of the UCA over a predetermined period. This measurement aimed to anticipate the characterization of the facility's electrical profile and to provide a dataset with real measurements for validating the analyses proposed in this study, independently of the schedule for restoring the supervisory system.

The recorded quantities include current, voltage, power factor, and frequency, with a sampling granularity compatible with the analysis of intraday demand patterns. The objective was to verify the proper functioning of the meters, as well as to monitor the operation of the UCA from an electrical energy standpoint.

Fig. 3 and Fig. 4 present boxplots of the frequency and voltage values recorded across the 18 operational meters, respectively. The x-axis represents the individual meters, while the y-axis indicates the operational limits (upper and lower) established for each monitored quantity. Meters QD-T1, QD-T4, and QD-T6 were temporarily excluded from the system due to being out of service and were duly reported to the maintenance team.

In both cases, it is observed that frequency and voltage operate in compliance with the limits established by current regulatory standards [13]. For frequency, the values must be between 59.5 Hz and 60.5 Hz, for voltage, between 209 V and 231 V. Despite the presence of outliers, it is noteworthy that 100% of the collected data fall within regulatory limits, represented by the black lines above and under the figures, in-

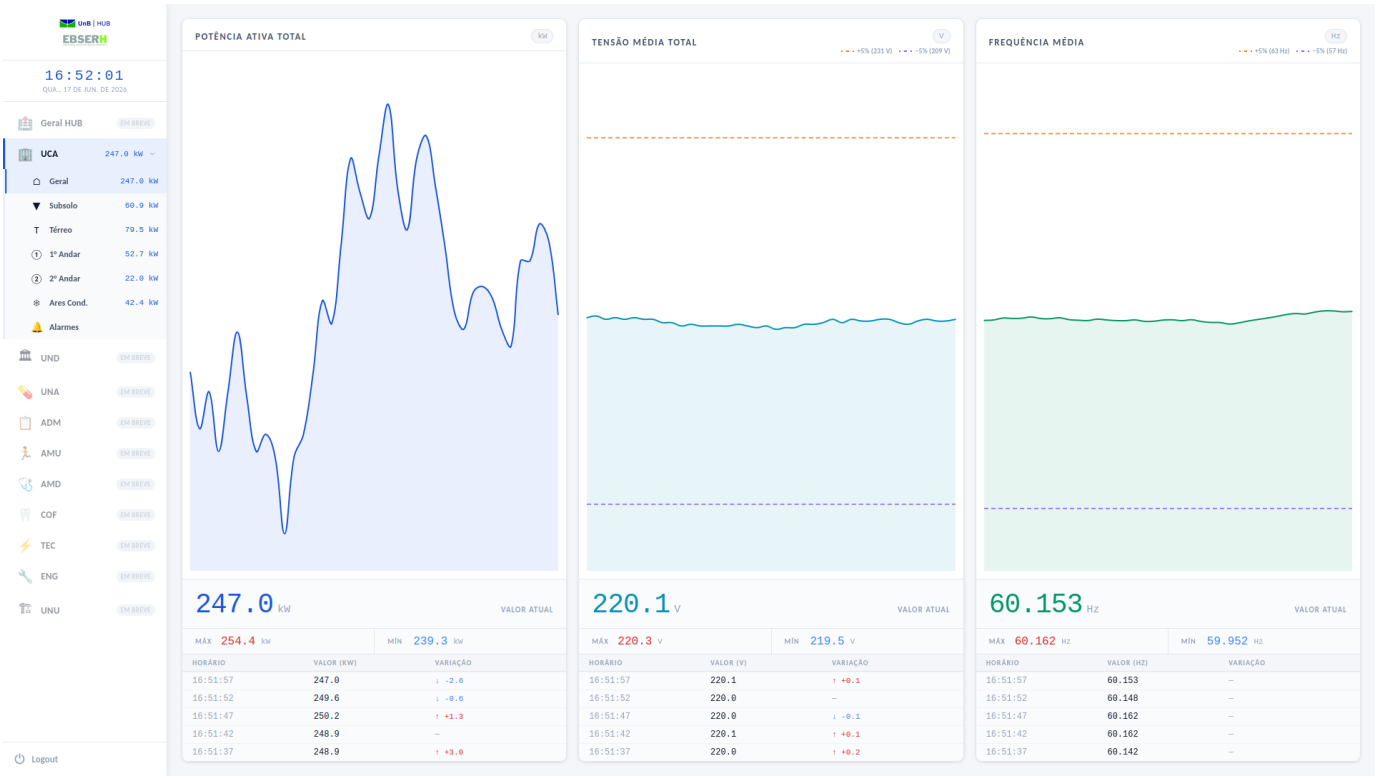


Fig. 2. SCADA-LTS platform main dashboard

dicating that the hospital's electrical system operated correctly throughout the monitored period.

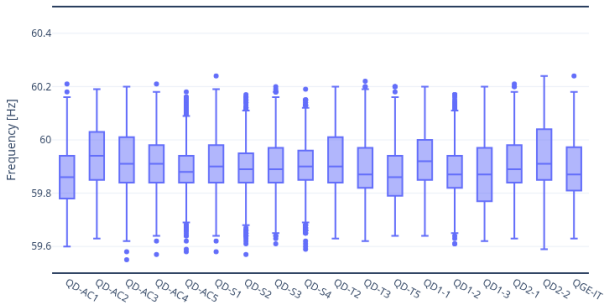


Fig. 3. Frequency monitoring of the UCA

In addition to these data, charts for energy recorded in 15 minutes per meter, system power factor, and average current per phase were generated, represented by Fig. 5, 6, and 7, respectively.

In 5 shows the energy consumption data for the seven monitored days. It can be observed that the meters showing the highest consumption are QD-AC1, QD-AC2, QD-AC3, QD-AC4, and QGE-IT. The first four are associated with the HVAC system distributed across the floors, while the last one

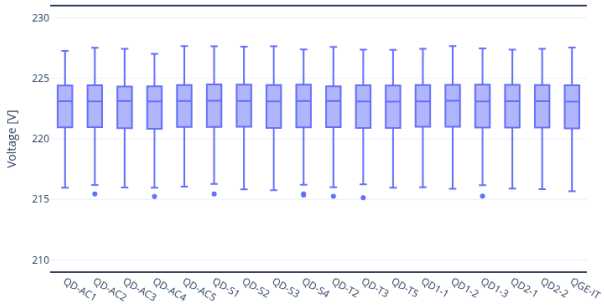


Fig. 4. Voltage monitoring of the UCA

is associated with the solar panel system.

As shown in Fig. 6, the power factor of the electrical installation is generally observed to be inductive (lagging). Given the presence of loads with varying technologies, operating profiles, and consumption patterns, it was to be expected that the electrical distribution system supplying the hospital would perceive it as a predominantly inductive load.

The behavior of the current values shown in Fig. 7 demnostratres that the UCA's electrical installation exhibits slight imbalances. However, from a general perspective, a balanced distribution of installed loads thoughout the building is still

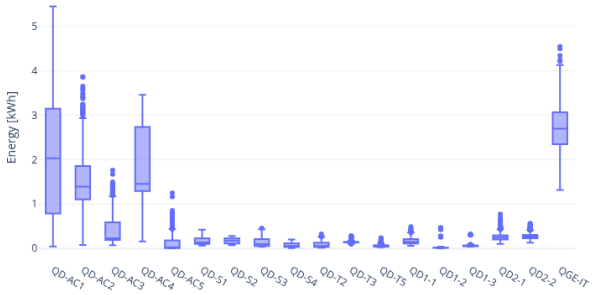


Fig. 5. Energy monitoring of the UCA

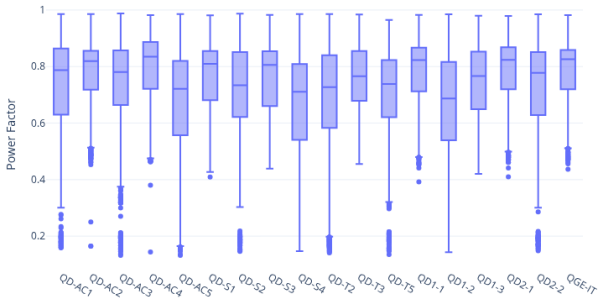


Fig. 6. Power factor monitoring of the UCA

maintained.

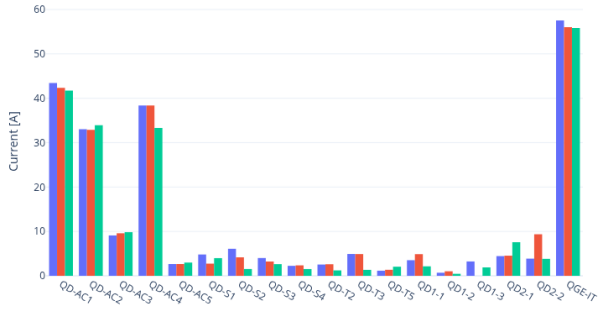


Fig. 7. Current monitoring of the UCA. Phase 1 (blue bar), Phase 2 (red bar) and Phase 3 (green bar)

VI. Conclusion

With full integration between the UCA meters and the SCADA-LTS server, continuous and automated collection of electrical quantities will become feasible. This integration will enable the enhancement of Modbus TCP communication at each measurement point, as well as the consolidation of a structured time-series repository covering the 21 meters installed in the UCA, with a minimum granularity of 15 minutes. Such granularity will allow the characterization of consumption profiles by circuit and by floor over complete seasonal cycles, as well as during sporadic events associated with increased patient flow throughout the year. Full operation of the SCADA-LTS system is further expected to cover broader seasonal variations and enable a more statistically robust characterization of the UCA's electrical profile.

Continuous monitoring of energy consumption and power quality will further enable the distinction between transient events and structural conditions requiring preventive, predictive, or corrective interventions. It is also proposed to evaluate the combined use of power quality monitoring and demand forecasting as inputs to a decision support system for the preventive maintenance of hospital electrical infrastructure, capable of anticipating failures before they compromise the continuity of patient care.

The development of electrical demand forecasting models based on the collected historical series is proposed, with performance evaluated using RMSE, MAE, and R^2 metrics, following the methodology validated in the literature for healthcare facilities [2], [3]. These models should be benchmarked against traditional statistical methods to determine the most suitable approach for each monitored variable. It is further proposed to apply the demand forecasting models already developed for HUB-UnB as a whole [2] to the specific context of the UCA, comparing the performance of statistical and machine learning models on the unit's electrical data. This is expected to validate the forecasting methodology at a more granular scale, per care unit, prior to its eventual expansion to the hospital's other units.

The implementation of a continuous monitoring system will enable the early detection of deviations in consumption patterns arising from operational failures, changes in hospital occupancy, or inefficiencies in HVAC systems.

Documentation of the proposed system will support its replication in other units of HUB-UnB, as well as in hospitals within the HU Brasil network.

Finally, the enhancement of data-driven energy infrastructure management is highlighted as a contribution stemming from the proposed approach, including the development of indicators based on the facility's electrical patterns, combined with other hospital logistics-related data, to support building administration.

For future works, is pretended to make a validation of anomaly detection algorithms, economic impact assessment and replication across mutiple hospital units, aiming to strengthen the scientific contribution and practical applicability of the proposed system.

Acknowledgment

The authors thank the Graduate Program in Biomedical Engineering, the National Council for Scientific and Tech-

nological Development (CNPq) and the University Hospital of Brasília (HUB-UnB/HU Brasil) for their collaboration and institutional support, which made this study possible.

REFERENCES

[1] WEG. *Multimedidor Eletrônico MMW03-CH: manual do usuário*. Documento nº 10006647116, Revisão 01. Jaraguá do Sul: WEG, 2024. Disponível em: https://static.weg.net/medias/downloadcenter/hba/h6c/WEG-multimedidor-MMW03-CH-manual-do-usuario-10006647116-pt.pdf. Acesso em: 26 ago. 2026.

[2] M. Zini and C. Carcasci, "Machine learning-based monitoring method for the electricity consumption of a healthcare facility in Italy," Energy, vol. 262, p. 125576, 2023, doi: 10.1016/j.energy.2022.125576.

[3] A. Ahmed, T. Jakir, M. N. H. Mir, M. A. F. Zeeshan, A. Hossain, A. H. Jui, and M. S. Hasan, "Predicting energy consumption in hospitals using machine learning: a data-driven approach to energy efficiency in the USA," Journal of Computer Science and Technology Studies, vol. 7, no. 1, pp. 199–219, 2025, doi: 10.32996/jcsts.2025.7.1.15.

[4] C. Liu et al., "Intelligent forecasting of electricity consumption for hospital outpatient buildings by FOA-SVR hybrid strategy," Frontiers in Energy Research, vol. 13, p. 1566664, 2025, doi: 10.3389/fenrg.2025.1566664.

[5] S. Mridha, A. Pandey, N. Venkatesan, and M. Chauhan, "AI driven energy consumption & environmental KPI forecasting in hospitals," International Journal of Environmental Sciences, vol. 11, no. 22s, pp. 1663–1676, 2025.

[6] M. J. Eckelman and J. Sherman, "Environmental impacts of the U.S. health care system and effects on public health," PLOS ONE, vol. 11, no. 6, p. e0157014, 2016, doi: 10.1371/journal.pone.0157014.

[7] S.-C. Hu, J. Chen, and Y. Chuah, "Energy cost and consumption in a large acute hospital," International Journal of Architectural Science, vol. 5, no. 1, pp. 11–19, 2004.

[8] ANVISA, "RDC n.º 50: Regulamento técnico para planejamento, programação, elaboração e avaliação de projetos físicos de estabelecimentos assistenciais de saúde," Agência Nacional de Vigilância Sanitária, Brasília, 2002.

[9] Modbus Organization, "Modbus Application Protocol Specification V1.1b3," Modbus Organization, 2006. [Online]. Disponível em: https://modbus.org/docs/Modbus_Application_Protocol_V1_1b3.pdf.

[10] EBSERH, "Perguntas frequentes – HUB-UnB," Empresa Brasileira de Serviços Hospitalares, 2024. [Online]. Disponível em: https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-centro-oeste/hub-unb. Acesso em: abr. 2026.

[11] UnB NOTÍCIAS, "Autoridades inauguram Unidade da Criança e do Adolescente do HUB," Brasília: Universidade de Brasília, jul. 2024. [Online]. Disponível em: https://noticias.unb.br/informes/7427-autoridades-inauguram-unidade-da-crianca-e-do-adolescente-do-hub. Acesso em: abr. 2026.

[12] ANEEL, "Resolução Normativa n.º 1000: Condições gerais de fornecimento de energia elétrica," Agência Nacional de Energia Elétrica, Brasília, 2021.

[13] ANEEL, "Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST, Módulo 8 – Qualidade do Fornecimento de Energia Elétrica". Brasília: Agência Nacional de Energia Elétrica, 2021. Estabelecido pela Resolução Normativa nº 956, de 7 de dezembro de 2021. [Online]. Available: https://www.gov.br/aneel/pt-br/centrais-de-conteudos/procedimentos-regulatorios/prodist

14 de setembro de 2026

“Estimated Annual Maintenance Cost of Life-Support Medical Equipment Registered in the Brazilian National Registry of Health Facilities” - Karin Moreira Santos; Harki Tanaka

Resposta aos/às revisores/as

Prezados/as, prezado/a editor/a,

Agradecemos a revisão atenta do nosso manuscrito, e os valiosos comentários para a complementação do documento, que certamente contribuirá à qualidade da versão final.

No que segue, iremos abordar as preocupações levantadas pelos/as revisores/as e indicar como contemplamos cada uma delas.

Revisor/a 1

O trabalho aborda um tema de grande relevância para a Engenharia Clínica e apresenta uma metodologia reprodutível, baseada em bases públicas (CNES e FNS), para estimar o custo anual de manutenção de equipamentos médicos de suporte à vida. O objetivo central busca contribuir diretamente para o planejamento orçamentário e a gestão de tecnologias em saúde. A metodologia é bem descrita e os autores deixam claras as limitações do estudo relativas às bases de dados, principalmente quanto ao caráter estimativo dos resultados.
Entretanto, o custo para manter a usabilidade, desempenho e a segurança de um equipamento médico não pode ser baseado somente no plano de manutenção. Ele é influenciado por diversos outros fatores, tais como: a tecnologia empregada, a necessidade de calibração, o uso de acessórios, a intensidade de uso, o modelo de manutenção (preventiva, corretiva ou contratual) e a disponibilidade de peças de reposição ou preço de mercado.
Nesse sentido, recomenda-se aprofundar a discussão como alerta ao leitor sobre os custos de calibração ou ensaios de desempenho, uma vez que esses componentes da manutenção apresentam características particulares, exigências regulatórias e periodicidades distintas entre as diferentes tecnologias médicas.

Agradecemos pela avaliação gentil do/a revisor/a, que levanta um ponto que de fato é essencial. Modificamos o nosso texto para contemplar essa ressalva importante. Mais especificamente, inserimos as seguintes frases no final da Introdução:

“Maintenance costs can be defined in several ways, and include or exclude expenses that are obviously related to the use of medical equipment, but for which, to our knowledge, no suitable database exists. Examples are the possible need of periodic calibration, accessories required for functioning, the variable cost of replacement of damaged parts, among other considerations that are hard to map, as these are extremely varied across different types of

medical equipment. There is no systematized database dedicated to these more specific costs of usage of medical devices, and we therefore present the data in this study noting first that these kinds of expenses are not considered in our estimates."

Como mencionado no próprio texto, "não existe hoje no Brasil uma base única consolidada que registre exclusivamente e diretamente o custo real executado de cada ordem de serviço ou manutenção técnica nacional". Contudo, o trabalho abre portas para que, em um futuro próximo, seja possível compor e validar esses custos por meio do cruzamento de dados oriundos de sistemas CMMS (Computerized Maintenance Management System). Esses sistemas registram o histórico real de cada equipamento, incluindo a quantidade de ordens de serviço, os custos de calibração, peças trocadas e acessórios, o valor da mão de obra (própria ou terceira), o tempo de parada (downtime), entre outros indicadores.
Portanto, nesse sentido recomenda-se discutir a utilização dos sistemas CMMS como ferramenta essencial para a coleta e a padronização de dados de manutenção nos estabelecimentos de saúde, permitindo futuramente confrontar os custos efetivamente realizados com as estimativas baseadas no valor de aquisição dos equipamentos.

Agradecemos o/a revisor/a por esta sugestão. De fato, a integração em maior escala de ferramentas digitais de acompanhamento de dispositivos do início até o final do ciclo da vida útil dos diversos equipamentos deve permitir um panorama bem mais acurado no que tange aos objetivos do nosso estudo, que é fornecer uma primeira estimativa do custo de manutenção do parque de equipamentos. É uma consideração importante para uma possibilidade que no entanto deve ser situada no futuro, já que a implementação deste tipo de sistema de acompanhamento ainda está em fase inicial na maioria das instituições. Contemplamos esta consideração através do seguinte novo trecho no manuscrito, alocado no final da seção III. Preliminary Results.

"Many institutions are starting to implement equipment management systems that register a broad set of operations related to the acquisition, functioning and maintenance of the devices that integrate the medical equipment park monitored by professionals of clinical engineering. To the extent these Computerized Maintenance Management Systems (CMMS) are adopted more widely, new methods for a much more accurate and systematized maintenance record will also allow to stipulate more reliable maintenance cost estimates. The implementation of CMM systems is gradual, and we expect this method of estimation to be available only at some point in the future."

Revisor/a 2

Agradecemos pela leitura atenta e pelas anotações do/a revisor/a; Tomamos a liberdade de itemizar os comentários levantados por este revisor para apresentar as respostas de maneira mais ordenada.

1. *Realizar uma análise de sensibilidade por tipo de equipamento e deixar ainda mais explícito que os valores representam cenários potenciais, não gastos reais.*

Entendemos que a possível falta de acurácia em alguns casos possa suscitar preocupações sobre o valor final das estimativas apresentadas. Na nossa análise, incluímos 15

dispositivos diferentes, e uma análise mais aprofundada nas estimativas de manutenção de cada um e os possíveis vieses associados para cada equipamento, individualmente, certamente receberá a atenção da nossa equipe de pesquisa futuramente. Para a finalidade do paper atual, da forma que o *scope* foi definido, acreditamos que este nível de detalhamento ocuparia um espaço desproporcional no texto, e preferimos deferir esta análise para posteriormente. Fizemos as devidas ressalvas no texto, no entanto, onde constam formulações como “The study is explicitly framed as an estimate of *potential* annual maintenance cost, not as a measurement of the real expenditure [...]” (com ênfase original), “exploratory and estimative approach based on open public data is the only viable path to a nationally representative study within the available timeframe” e “it does not report equipment age, usage frequency, failure history, whether preventive maintenance was actually executed and at what periodicity, maintenance contract terms, amounts paid, or equipment downtime. Consequently, the cost figures presented in this paper are best interpreted as an estimated potential maintenance cost associated with the registered park, computed from acquisition-value scenarios, and not as a measurement of real maintenance expenditure incurred by health facilities;”. Portanto, mesmo sem a apresentação de uma avaliação mais detalhada dos possíveis vieses nas estimativas por tipo de equipamento, que não conseguimos incluir no escopo do estudo atual, acreditamos que tenhamos enfatizado suficientemente junto aos interessados no manuscrito que as nossas estimativas não devem ser lidas como apresentando uma acurácia que não conseguimos oferecer ainda, a partir dos métodos utilizados.

2. *Também corrigiria o mapeamento CNES–FNS, pois o flowmeter foi associado ao preço de um Doppler vascular por ausência de correspondência exata, o que pode introduzir viés evitável na estimativa.*

O/a revisor/a certamente levanta um ponto importante, que é uma consequência do uso de referências de catálogo diferentes entre os sistemas existentes para o acompanhamento de aparelhos médicos em uso, que é um dos elementos mais pertinentes nas nossas conclusões. No caso do relacionamento do flowmeter com o Doppler vascular, fizemos uma tentativa para evitar uma lacuna no nosso levantamento através de uma associação que acreditávamos ser a mais plausível. Ao mesmo tempo, se o flowmeter corresponder a um equipamento distinto do equipamento de Doppler vascular, provavelmente seria um de menor complexidade e portanto menor valor do que o levantado no estudo. Como, pelos números que obtivemos, o valor do Doppler vascular só corresponde a 0,1% do valor total do parque de equipamentos médicos (v. Table III), o impacto de incluir ou excluí-lo da contabilização não deve produzir uma grande variação nas estimativas obtidas. Para permitir que os leitores façam uma avaliação deste ponto da relevância numérica limitada do flowmeter, incluímos, no texto, a seguinte frase, no final da Seção II: “According to the results obtained, presented below, the financial weight of the flowmeter in our estimates is minor (in the order of 0,1%), and we believe that alternative mappings have no significant impact on the aggregated numbers we present in this study. ” Nós mantivemos a intenção de produzir novas estimativas em trabalhos futuros quando os códigos forem harmonizados, como já citada no texto original: “this correspondence will be revisited as the FNS catalog is updated”.

3. *Além disso, monitores de pressão invasiva e não invasiva foram excluídos dos custos apesar de representarem aproximadamente 18,4% das unidades em uso, tornando o total uma estimativa inferior ao parque completo.*

Trata-se novamente de uma consequência da falta de harmonização entre os sistemas que usamos para realizar o trabalho. Infelizmente, diferente do caso do flowmeter, não teríamos embasamento para encontrar valores correspondentes. Concedemos que é uma limitação relevante do nosso estudo, e acrescentamos uma frase no final da Discussão para enfatizar que as estimativas, por esta lacuna na base da FNS, naturalmente seriam maiores se estes dispositivos forem incluídos:

“As the systems that can be consulted for the cost estimates currently do not allow for alternative estimates of the pressure monitors, one must consider that the values presented in this study would underestimate total maintenance cost in applications in which they are to be taken into account.” Furthermore, we added the comment “The invasive and non-invasive pressure monitors are set to zero in this graph because their average values are not available in the FNS system.” in the caption of Figure 2.

4. *Recomendo apresentar explicitamente o impacto dessas ausências e, se possível, acrescentar indicadores normalizados por população ou número de leitos para evitar que o ranking regional reflita predominantemente o tamanho dos estados.*

Agradecemos o/a revisor/a por este ponto, que contemplamos na versão atual do manuscrito, como ressalva: “As the maintenance costs are calculated proportionally to the value of the medical equipment park considering the life support items included in this study, the results per macroregion reflect the uneven distribution of both population and medical resources throughout Brazil. For example, using population estimates for july 2026 provided by the Instituto Brasileiro de Geografia e Estatística (IBGE), the maintenance cost according to the 7% benchmark amounts to R$9,95 per capita in the South, but only R$5,66 per capita in the North.”

Accumulated Dose Estimating to the Eye Lens in Fluoroscopic Procedures

Jose CdeC. Lourenco
Londrina State University (UEL)
Londrina - Paraná, Brazil
jcl@uel.br

Hugo R. Schelin
Federal University of Technology (UTFR)
Curitiba - Paraná, Brazil
Schelin2@gmail.com

Sergei Paschuk
Federal University of Technology (UTFPR)
Curitiba - Paraná, Brazil
spaschuk@gmai.com

Interventional radiologists receive large doses of radiation. The International Commission on Radiological Protection recommends an annual occupational dose of 20 mSv to the eye lens to prevent cataracts. We used a cumulative dose tracking system with real-time dosimeter surveillance to inform physicians about the risks of fluoroscopy table radiation. The present study deals with the real-time radiation exposure to the eyes of physicians who are standing 57 to 76 cm away from the primary beam of the fluoroscopy table during 8-, 12-, and 20-min procedures. Interventional radiologists perform 270 simulated procedures per year, exposing them to 8.68, 20.44, and 31.08 mSv at 57 cm and 6.44, 14.84, and 22.12 at 76 cm. We can protect interventional radiologists and their teams by considering the range and duration of the radiation beam. The estimated results are partly consistent with the ICRP criteria.

Keywords— Eye lens dose; Estimating Accumulated Dose; Interventional radiology.

I. INTRODUCTION

An interventional cardiologist uses X-rays all the time to see what's going on inside the body during procedures. Because they are constantly exposed to ionizing radiation, individuals are required to wear the appropriate radiological protection whenever they are near the X-ray beam. Excellent protection against ionizing radiation will prevent acute risks such as cataracts and skin lesions [1-5]. Additionally, long-term exposure can lead to chronic effects, including an increased risk of cancer. This is important for observing safety measures and monitoring radiation exposure during interventional procedures. The interventional physician receives high doses of radiation if his hands are in the region of the X-ray beam. The severity of side effects caused by radiation increases regarding the dose received, and this adaptation does not interfere with the dose, potentially leading to simple erythema or severe burns of the skin. This study shows how important it is to have adequate radiation protection, which minimizes these side effects. By implementing proper safety measures and using advanced shielding materials, we can significantly reduce the risk of exposure. Ensuring that both workers and patients are well-protected is essential for maintaining health and safety in environments where radiation is present. The Brazilian Institute for Radiation Protection and Dosimetry (IRD) has determined that when working with the dose received, and this adaptation does not interfere with ionizing radiation during a medical diagnosis, the occupational doses of radiation to which medical personnel are exposed should be controlled. According to Bingham, the IRD also advises keeping records so that clinicians can obtain a report of their interventional radiology-related activities. It also suggests that routines for measuring occupational doses for staff should be improved by frequent measurements of radiation doses on the hands and lenses, as well as in the assessment of recorded personal doses. Employees working in areas with ionizing radiation typically use thermoluminescence dosimeters (TLDs) in radiology-intensive hospitals. The short distance from the radiation source contributes to the staff's high radiation exposure during interventional work. Often, when working near the patient, it is difficult to use protective screens or even glasses, which are essential for protection against radiation exposure when working near the source. Unfortunately, the interventionist physician does not always use these tools properly, and many find them awkward and uncomfortable to use during long procedures [7]. The topic of ionizing radiation is especially relevant for interventionist physicians since the cumulative dose of interventionist radiography teams represents more than 90% of the collective dose of all radiology personnel. This emphasizes the significance of continual education and training to ensure that all team members are well-equipped to deal with radiation exposure successfully. Intervenors guaranty quality and safety. The dosimeter used by medical staff has the disadvantage of not immediately showing its cumulative dose, which is usually read one month after its use; so, a system for effectively measuring dose rates is essential for the safety of the medical staff. It is important that this system shows radiation doses in real time, as the information can help them be aware of protection against ionizing radiation and therefore lead to a

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

decrease in their personal doses [8]. The real-time system lets an interventionist doctor see the equivalent dose immediately. The information helps them stay away from the source of radiation and move to a position that exposes them to less radiation without affecting the performance of the intervention procedure. To obtain an equivalent cumulative dose real-time system [9], you need to know the factors about the ionization chamber's reference point and the dose rates around the fluoroscopy table, which are shown by the isodose map. The construction of the isodoses map is the result of the dose rate measurements of all points around the fluoroscopy table. The measured points with the same dose rates have adjusted the use of the interpolation technique to construct the isodoses. An ionization chamber located at a distance from the main beam records dose rates in real time. Real-time correction of measured values occurs through conversion factors in the isodose map, which displays dose rates and cumulative dose. The medical staff views these values numerically in relation to their position regarding the main beam. The medical staff [10-12] numerically views these values in relation to their position regarding the main beam. The software development project [13-14] aims to estimate equivalent doses in real time, as shown in figure 1.

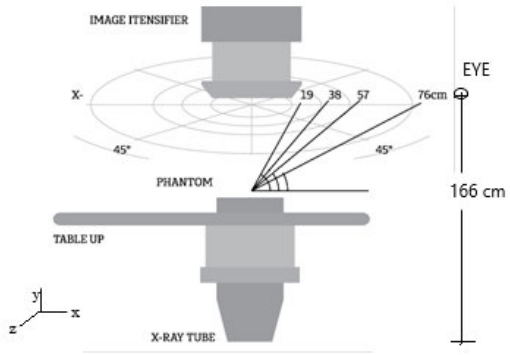


Fig. 1: Design of the development to estimate doses in real time.

The interventionist physician and their team will be able to verify radiation exposure in real time with the visualization system. By comprehending the consequences of ionizing radiation, medical professionals can modify their positions to reduce the dose they receive while maintaining the efficacy of the medical procedure. The annual occupational dose for the eye lens should be reduced from 150 mSv to 20 mSv, as per a recent statement from the International Commission on Radiological Protection (ICRP) [6]. The objective of this decrease in the cumulative dose limit is to mitigate the risk of cataracts after absorption. The main objective of this project is to develop a system that accurately estimates the equivalent dose to the eye lens in real time during interventional medical procedures that utilize fluoroscopy.

II. MATERIALS AND METHODS

The experimental setup consists of a cylindrical phantom (15 cm height, 30 cm width) to simulate the scattered radiation, a Ludlum 9DP ionization chamber connected to a computer using a Ludlum cable to measure the rate, and an automated fluoroscopy system, INNOVA. The real-time monitor is next to the fluoroscopy table. The algorithm estimates cumulative doses near the doctor's view. The real-time system's development was divided into two phases, as illustrated in Figure 2. The first step was a system linearity test. The system linearity test was a measurement of the dose rate as a function of the voltage, amperage, and exposure time. The current was kept constant at 30 mA, while the voltage was between 50 and 90 kV. Moreover, the voltage was constant (50 kV), and the current was changed from 5 to 30 mA. The exposure time was 10 s. Measurements were taken at a height of 166 cm above ground level, specifically at positions of 57 and 76 cm along the x and y axes from 45° to 45°. All measurements were conducted with a current of 30 mA, a voltage of 50 kV, an exposure time of 10 seconds, and 100 cm from the x-ray tube to the table. The ionization chamber was configured to gather continuous data at one-second intervals, aligned toward the center of the scattering object at all points on the plane. After data collection, an algorithm was developed to (1) eliminate data below 10 μSv/h, which identified the data for each position; (2) account for the ionization chamber's dead time by utilizing data above the average for each point; and (3) discard data that did not correlate. The Chauvenet criterion was applied to calculate the mean and standard deviation of the dose rate at each point [15]. We used interpolation techniques to determine isodoses based on data measured from scattered radiation points in each plan. The isodoses of interest corresponded to points (RangeDosesRatesInMap) at 38, 57, and 76 cm. These values were subsequently used to create a map of isodoses. A conversion factor was applied for the isodose on the map. The conversion factor isodose rate is 443 μSv/h. It is the dose rate at the reference point of Eq. (1).

$$\text{Factor}=\text{RangeDosesRatesInMap}/\text{DoseRateReferencePoint} \quad (1)$$

The second part, illustrated in Figure 2, pertains to the real-time system in which the ionization chamber was positioned at a fixed reference point. This system was designed to collect

data at specified intervals and transmit it to the 9DP Logging Spreadsheet software. The estimated real-time isodose rates indicate the dose rates measured at that specific location, where the conversion factor from Equation (2) is applied.

$$\text{Estimate} = \text{Factor} * \text{DoseRateReferencePoint} \quad (2)$$

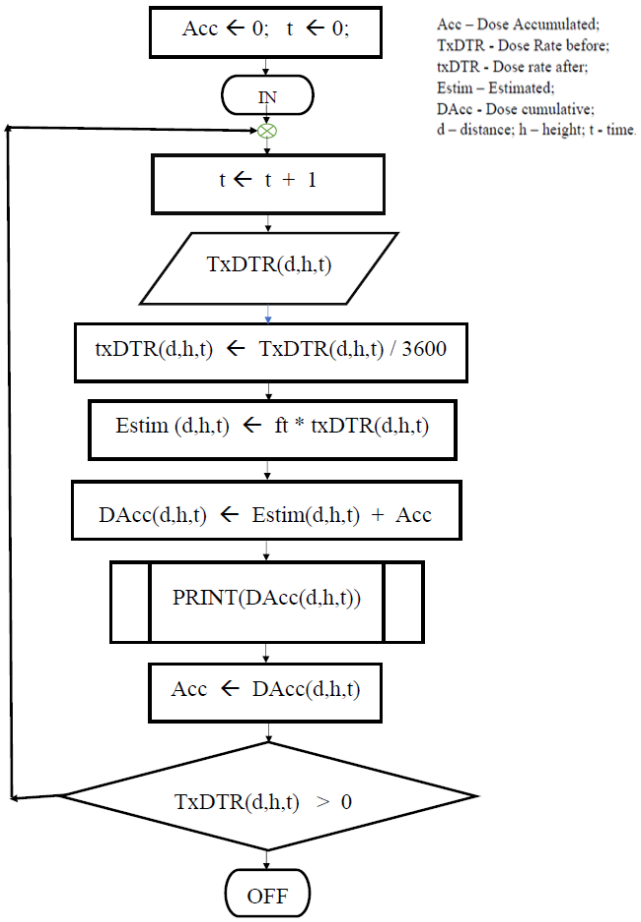


Fig. 2: Computational diagram of software development to estimate doses in real time [13].

The isodose rates (EstimateDoseRate) were initialized every time the x-ray system was triggered. Otherwise, the recording of dose rates referred to background radiation. Equation (3) calculated the estimated cumulative dose (EstimateAcc) for each isodose bin.

$$\text{EstimateAcc} = \text{EstimateAcc} + \text{Estimate} / 3600 \quad (3)$$

All interventionist procedures began with the readings for the accumulated dose of each isodose map at zero. The estimated cumulative dose is calculated in real time, and the results are displayed on a TV monitor for each corresponding radiation range when the fluoroscopy system is activated for a medical angiography procedure. This feature enables the physician to promptly make decisions, thereby ensuring the patient's safety in the vicinity of their eyes. Additionally, the system is designed to alert the physician to any substantial modifications in dose levels, allowing them to make immediate adjustments as required.

III. Results

The real-time dose system was validated from the interventional physician's perspective at distances of 57 cm and 76 cm from the primary beam, at a height of 166 cm from the floor to eye level, and at an angle of 45° from the fluoroscopy table. The system was tested with an irradiation duration of 10 seconds, as outlined in Table 1.

Table 1 - Dose rates were assessed at the interventional physician's position after a 10-second period. The primary beam was oriented at a 45° angle to the fluoroscopy table, and measurements were taken at distances of 56 cm and 76 cm.

Distance (cm)	Measured (µSv/h)	Estimated (µSv/h)	Deviation (%)
57	727,1	728,1	0,14
76	248,5	248,1	0,16

The interventional physician dose-rate tests and the estimated dose-rate tests yielded statistically significant agreement between the computed and measured values (p-values ranging from 0.05 to 0.025). Typically, the duration of an angiography procedure ranges from 30 minutes to 1 hour, contingent upon the examination's complexity. The voltage and current are automatically adjusted during the interventional procedure. The tests were administered for 8, 12, and 20 minutes. Real-time dose rates and accumulated doses in the physician's ocular region are estimated by the prototype for main beam distances of 57 cm and 76 cm. These were 31.20, 72.82, and 110.29 µSv at 57 cm and 22.54, 52.60, and 78.43 µSv at 76 cm per procedure with posteroanterior (PA) tube angles (figure 3-5). The estimated annual cumulative doses were determined by assuming that the physician conducted 270 interventional procedures annually [4]. The experimental conditions have been simulated. During these procedures, the medical professional is exposed to ionizing radiation. At a height of 57 cm, this professional receives 8.68, 20.44, and 31.08 mSv annually, while at 76 cm, they receive 6.44, 14.84, and 22.12. The data were collected

without the use of glasses. The results indicate that fluoroscopy suites are only used for interventional treatments. According to the findings, fluoroscopy allows devoted people to face higher amounts of radiation exposure when performing interventional operations.

to elevated levels of radiation exposure. Eye Protectors and Lead Aprons: provide protection against ionizing radiation. The doctor and their team should minimize the time in primary radiation to ensure their health and safety. However, at times, a brief exposure may be necessary to avoid interruptions during the interventional medical procedure.

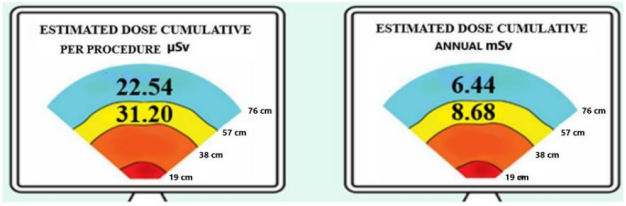


Fig. 3 – The estimate shows the distribution of cumulative fluoroscopy doses per procedure for a duration of 8 minutes. The annual accumulated dose for 270 procedures performed under the same conditions is also estimated.

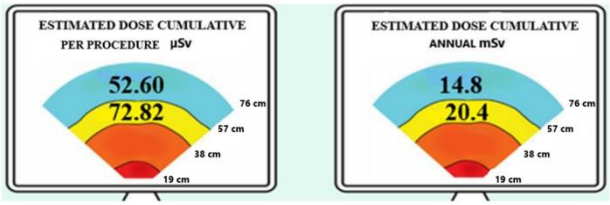


Fig. 4 –. The estimate shows the distribution of cumulative fluoroscopy doses per procedure for a duration of 12 minutes. The annual accumulated dose for 270 procedures performed under the same conditions is also estimated.

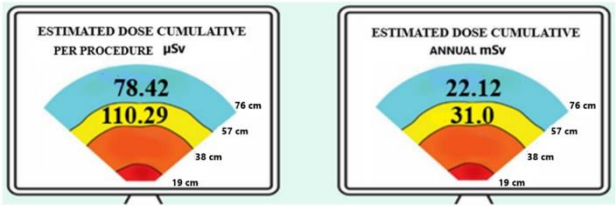


Fig.5 – The estimate shows the distribution of cumulative fluoroscopy doses per procedure for a duration of 20 minutes. The annual accumulated dose for 270 procedures performed under the same conditions is also estimated.

IV. Discussion

The aim of this study was to better understand settings at 8, 12, and 20 minutes, as well as at distances of the dispersion of to the physician's eye, which is particularly vulnerable to radiation exposure due to ocular malignancies. The International Commission on Radiological Protection (ICRP) recommended limiting radiation exposure to 20 mSv per year and the use of personal dosimeters, lead-lined eyeglasses, and other radiation protection measures. The prototype must display real-time estimates of cumulative doses per procedure setting at 8, 12, and 20 minutes, as well as at distances of the accumulated yearly dose on a TV for the interventional physician. In our work, we have studied the effect of ionizing radiation on X-ray generation. The interventional physician used automatic amperage settings at 8, 12, and 20 minutes, as well as at distances of 57 and 76 cm from the primary beam. We estimated 270 interventions per year and found that the physician's eye received the highest cumulative dose for estimated doses per procedure greater than 20 minutes of radiation exposure. The estimates are consistent with ICRP-103 [16-17] and use the same parameters at the various research stages.

V. Conclusions

Monitoring the estimates of cumulative doses during fluoroscopy can help the physician perform the angiography better by controlling the exposure to scattered radiation during the interventional procedure. A TV display is used to present real-time dosage estimates for radiation dispersion near the fluoroscopy table, both per procedure and annually. The results indicate that the best distance is between 57 and 76 cm, and fluoroscopy time beyond 20 minutes of estimated exposure exceeds the accumulated dose estimates of 20 mSv, raising serious concerns about the risk of exposure to the physician's ocular lenses causing cataracts. The fact that the results are partially consistent with the ICRP guidelines allows for considering the approach to the real-time visualization system, which is crucial and provides immediate responses for estimating accumulated doses.

ACKNOWLEDGEMENTS

The authors thank Dr Cesar A. Santos, from the Hospital University of the State University of Londrina, for his generous help and time for discussion.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Valentin, J., 2000. Abstract: Avoidance of radiation injuries from medical interventional procedures, ICRP Publication 85. Ann. ICRP 30, 7–7. https://doi.org/10.1016/S0146-6453(01)00004-5

[2] Bingham, D., Bull, R.K., 2013. A real-time internal dose assessment exercise. Radiat. Prot. Dosimetry 154, 267–275. https://doi.org/10.1093/rpd/ncs241

[3] Bera, Geraldine., *et al*.; Eye lens radiation exposure of workers during medical interventional procedures and surgery. Radiation Protection Dosimetry. p. 323-328, 2018

[4] Valentin, J., 2000. Abstract: Avoidance of radiation injuries from medical interventional procedures, ICRP Publication 85. Ann. ICRP 30, 7–7. https://doi.org/10.1016/S0146-6453(01)00004-5

[5] U O'CONNOR; J M RYAN; et al., Occupational radiation dose to eyes from interventional radiology procedures in light of the new eye lens dose limit from the International Commission on Radiological Protection. The British journal of radiology · March 2015.

[6] Stewart, F.A., Akleyev, A.V., Hauer-Jensen, M., Hendry, J.H., Kleiman, N.J., MacVittie, T.J., Aleman, B.M., Edgar, A.B., Mabuchi, K., Muirhead, C.R., Shore, R.E., Wallace, W.H., 2012. ICRP PUBLICATION 118: ICRP Statement on Tissue Reactions and Early and Late Effects of Radiation in Normal Tissues and Organs — Threshold Doses for Tissue Reactions in a Radiation Protection Context. Ann. ICRP 41, 1–322. https://doi.org/10.1016/j.icrp.2012.02.001

[7] Sandblom, V., Mai, T., Almén, A., Rystedt, H., Cederblad, A., Båth, M., Lundh, C., 2013. Evaluation of the impact of a system for real-time visualisation of occupational radiation dose rate during fluoroscopically guided procedures. J. Radiol. Prot. 33, 693–702. https://doi.org/10.1088/0952-4746/33/3/693

[8] Rigatelli, G., Panin, S., Fiorrevanti, R., Oliva, L., Dario, A., Tiberio, L., Bedendo, E., Pertoldi, D., Chambers, C., 2016. Impact of Operators' Height on Individual Radiation Exposure Measurements during Catheter-Based Cardiovascular Interventions. J. Interv. Cardiol. 29, 83–88. https://doi.org/10.1111/joic.12263

[9] Vano, E., 2011. ICRP and radiation protection of medical staff. Radiat. Meas. 46, 1200–1202. https://doi.org/10.1016/j.radmeas.2011.05.031

[10] Sailer, A.M., Paulis, L., Vergoossen, L., Kovac, A.O., Wijnhoven, G., Schurink, G.W.H., Mees, B., Das, M., Wildberger, J.E., de Haan, M.W., Jeukens, C.R.L.P.N., 2017. Real-Time Patient and Staff Radiation Dose Monitoring in IR Practice. Cardiovasc. Intervent. Radiol. 40, 421–429. https://doi.org/10.1007/s00270-016-1526-8

[11] Baptista, M., Figueira, C., Teles, P., Cardoso, G., Zankl, M., Vaz, P., 2015. Assessment of the occupational exposure in real time during interventional cardiology procedures. Radiat. Prot. Dosimetry 165, 304–309. https://doi.org/10.1093/rpd/ncv052

[12] Hosokai, Y., Matsumoto, K., Kozakai, M., Takahashi, K., Usui, A., Win, T.P., Muroi, K., Saito, H., 2017. Development of real-time radiation exposure dosimetry system using synthetic ruby for interventional radiology. Radiat. Prot. Dosimetry 175, 517–522. https://doi.org/10.1093/rpd/ncw381

[13] De Camargo Lourenco, J. C.; Paschuk, S. A.; Schelin, H. R.; Denyak, V.; Santos, C. A. C. Development of a system to estimate doses in real time. Radiation Physics and Chemistry, v. 167, n. December 2018, p. 108267, 2020. Elsevier Ltd. Disponível em: <https://doi.org/10.1016/j.radphyschem.2019.04.017>.

[14] Principi, Sara, Development of methodologies for estimating the dose to the eye lens in interventional radiology: operational implications of the eye lens new dose limits. Universitas Politècnica de Catalunya, Doctoral thesis, 2017. http://hdl.handle.net/2117/121020

[15] Leon-Garcia, A., 2009. Probability, Statistics, and Random Processes for Engineers, Technometrics. https://doi.org/10.1198/004017004000000400

[16] Vetter, R.J., 2008. ICRP Publication 103, The Recommendations of the International Commission on Radiological Protection. Health Phys. 95, 445–446. https://doi.org/10.1097/01.hp.0000324200.73903.5b

[17] Interventional Cardiology Medical Group & Interventional Cardiology Medical Group. (2025, May 27). How long does an angiogram take? Interventional Cardiology Medical Group. https://www.icardiomg.com/how-long-does-an-angiogram-take/

Evaluation of Open Ecosystems for Digital Prosthetic Workflows in Dentistry within the Brazilian Unified Health System: A Proof of Concept Applied to Public Procurement

Patrícia Oliveira de Andrade*
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0009-0003-7415-925X
patricia.andrade@agenciasus.org.br

Rayane Souza Sobral*
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health Syste, Brasília, DF
0000-0001-9781-9140
rayane.sobral@agenciasus.org.br

Janne Carla Lopes Mendes
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System Brasília, DF
0009-0009-8145-8251

Luís Alberto S. Soares de Sousa
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0009-0006-1042-7173
luis.souza@agenciasus.org.br

Manuela Oliveira da Silva Paz
CMTA/UAPS
Brazilian Agency for Support to the Management of the Unified Health System, Brasília, DF
0009-0000-8915-5647
manuela.paz@agenciasus.org.br

Caroline Castanho Duarte
UAPS
Brazilian Agency for Support to the Management of the , Unified Health SystemBrasília, DF
0000-0001-8348-3586
caroline.duarte@agenciasus.org.br

*Corresponding author

Abstract: The growing incorporation of digital technologies into dentistry imposes new challenges on public procurement processes, since the performance of these solutions depends on the integration between equipment, software and technological infrastructure. This paper reports the application of a Proof of Concept (PoC) for the technical evaluation of a digital dental workflow intended for the production of prosthetic devices within the Brazilian Unified Health System (SUS). The PoC was used as an instrument for the analysis of the interoperability and operational performance of the components of the digital ecosystem, supporting decision-making in a specific context of technology acquisition. As a case study, an intraoral scanner, CAD software and their integration into the clinical-laboratory workflow were evaluated by means of document analysis and a practical demonstration conducted by a multidisciplinary team. The evaluation made it possible to identify aspects related to usability, compatibility, computational infrastructure and operational performance that would not be fully verified through the technical documentation submitted by suppliers alone. The results show that practical validation increases the safety of decision-making and contributes to more robust and better-grounded public procurement processes, as well as allowing better planning of the subsequent stages of the technological densification process. It is concluded that the evaluation of digital dental technologies must consider the technological ecosystem in an integrated manner, incorporating practical validation methods as a complement to documental analysis.

Keywords: Health Technology Assessment; Digital Dentistry; Device compatibility; Proof of concept; Public Procurement.

I. INTRODUCTION

Digital transformation has brought about significant changes in health care processes, driving the replacement of analogue workflows by automated and connected processes supported by integrated computational systems [1, 2]. In dentistry, this movement has become particularly evident in the field of prosthetic rehabilitation, in which the incorporation of digital technologies has profoundly transformed the way dental prostheses are planned, produced and delivered to the patient, increasing clinical accuracy, reducing production time and improving the predictability of rehabilitation treatments [3-5].

The growing incorporation of digital innovations into health systems increases the relevance of Health Technology Assessment (HTA) as a strategic tool to support decision-making processes related to technological densification. In the context of digital health technologies, HTA has taken on an increasing role in the assessment of the value, effectiveness and organizational impacts associated with the adoption of these solutions [6,7]. Traditionally, HTA is structured around the multidimensional analysis of specific technologies, assessing aspects related to safety, clinical effectiveness, technical performance, cost-effectiveness and organizational impacts associated with the adoption of medicines, equipment, medical devices or clinical procedures [6,7].

However, the recent evolution of digital health technologies has introduced challenges that go beyond conventional evaluation models. Unlike technologies assessed as isolated devices, the digital workflows applied to dentistry operate as complex technological ecosystems, in which clinical and operational performance does not depend exclusively on the individual characteristics of each item of equipment, but emerges directly from the functional interaction between multiple interdependent components. In such scenarios, technology assessment requires the verification of interoperability, functional integration and systemic performance along the entire digital workflow,

which are not fully captured by traditional models based on the isolated analysis of the components [7-10].

These workflows may be structured in open or proprietary architectures, integrating different equipment and software responsible for the acquisition, processing and fabrication stages of prosthetic devices [3-5]. In this model, the final outcome of the care process depends simultaneously on the coordinated functioning of the intraoral scanner, the digital planning software, the computational capacity available for processing, the automated manufacturing equipment and, above all, on the interoperability between technologies frequently developed by different manufacturers, a critical element in systems structured under an open architecture [9-10]. Consequently, the individual technical conformity of each component does not necessarily ensure the overall performance of the system when operating as an integrated workflow [9-11]. This characteristic becomes particularly relevant in public procurement processes, in which different technological components may be acquired separately on the basis of independent technical specifications and regulatory documentation. Although they individually meet the established requirements, their functional compatibility and the integrated performance of the system must be verified jointly [7,12].

From the perspective of Health Technology Assessment, this scenario reveals an important methodological gap: how can technological systems whose performance derives from the integrated behaviour of multiple interdependent components, and whose documentary evaluation is not sufficient to predict their performance under real conditions of use, be assessed in a robust manner applicable to public management? Recent discussions on HTA applied to digital technologies and medical devices point to the need for evaluative approaches capable of addressing aspects of interoperability, implementation and performance in real contexts of use [6, 7, 13].

In this context, it becomes necessary to adopt complementary mechanisms capable of assessing in advance the integrated performance of these complex ecosystems before their definitive incorporation. Because of this, carrying out Proofs of Concept (PoC) emerges as a methodological strategy capable of validating the compatibility between system components, reducing the technological risks associated with the public procurement process and producing more robust technical evidence to support incorporation decisions. In this context, the present study describes the application of a proof of concept carried out during the technical evaluation of a digital dental ecosystem intended for the Brazilian Unified Health System, discussing its potential as an instrument complementary to Health Technology Assessment in public procurement processes [11, 14].

II. CASE STUDY

A. Context

Expanding access to oral rehabilitation and modernizing oral health care processes are priorities of the public policies aimed at strengthening Primary Health Care (PHC) within the Brazilian Unified Health System (SUS). In this context, a national strategy for the incorporation of digital technologies was designed with the objective of qualifying care, increasing the resolution capacity of services and promoting greater efficiency in the production of dental prostheses. This initiative is part of the effort to strengthen Primary Health Care, oriented towards expanding the resolution capacity of services and the technological qualification of Basic Health Units (UBS) and Dental Specialty Centres (CEO), in accordance with national guidelines for the modernization of care infrastructure. As part of this strategy, a procurement process was conducted for the implementation of a digital workflow for the fabrication of dental prostheses in public health services. Unlike traditional acquisitions centred on isolated equipment, the contract awarded had as its object the implementation of a complete technological ecosystem, structured to support all stages of the digital prosthetic workflow, from the acquisition of intraoral images to the manufacture of the final prosthetic device. This contracting model introduces additional challenges to technical evaluation, since the individual conformity of the components does not in itself ensure the integrated performance of the system under real conditions of use.

The technological ecosystem comprised intraoral scanners, specialized software for acquisition, planning and digital modelling (CAD/CAM/CAI), compatible computational infrastructure, additive manufacturing equipment, post-processing devices and specific consumables, constituting a complete digital workflow for the production of dental prostheses and configuring a sociotechnical system whose performance depends on the interaction between technological, operational and human factors.

Although each component of the solution individually met the technical requirements set out in the contracting documentation, the adoption of an open architecture made it necessary to verify the functional compatibility between equipment, software and computational infrastructure when operating in an integrated manner. In this context, a Proof of Concept (PoC) was conducted as a complementary stage of technical evaluation, aimed at producing evidence on the compatibility and the performance of the technological ecosystem under conditions representative of use.

It is important to highlight that the acquisition was structured under an open solution model (open ecosystem), in which the different technological components that make up the production workflow do not necessarily belong to the same manufacturer, thereby increasing technological flexibility and competition among suppliers. This architecture, however, transfers to the technological densification process the challenge of ensuring functional compatibility and full compatibility between all the components that make up the system.

B. B. Motivation for developing the Proof of Concept

Although public procurement processes traditionally rely on the documentary analysis of technical specifications, regulatory certifications and demonstrations of conformity of the products offered, this model has limitations when applied to the evaluation of technological ecosystems, whose performance depends on the interaction between multiple components. In the specific case analysed, it was found that the individual technical conformity of the components alone was not sufficient to ensure the performance of the digital workflow when operating in an integrated manner. The quality of the solution depended on the interaction between equipment, software, computational infrastructure and digital files throughout the entire production process.

In addition to compatibility aspects, the evaluation took into account that the different components of the digital workflow were subject to distinct regulatory requirements, which limits the exclusive use of sanitary documentation as sufficient evidence to attest to the performance of the technological ecosystem in an integrated manner, especially in scenarios involving multiple suppliers and interoperable solutions.

The analysis of the technical report prepared by the Federal University of Uberlândia (UFU) showed that factors such as computational capacity, compatibility between software packages, performance of CAD systems and operator experience directly influenced the performance of the digital workflow, confirming that such variables constitute critical determinants of performance and cannot be fully assessed by means of isolated technical documentation. This finding revealed the inadequacy of an evaluation based on the expectation of complete proprietary solutions, reinforcing the need to consider hybrid configurations based on complementary software that is already available and has been functionally tested.

Moreover, it was found that certain native software packages originally supplied by the manufacturers presented functional limitations regarded by the UFU team as relevant to the full execution of the prosthetic workflow, making it necessary to additionally use specialized external software in order to enable specific stages of digital modelling and planning. This observation reinforced the understanding that the isolated evaluation of each component would not be sufficient to predict the overall performance of the system after its implementation in a real environment of use.

On the basis of the integrated analysis of the technical documentation, the evidence presented in the technical report and the statements submitted by the supplier, it was concluded that uncertainties remained regarding the integrated performance of the proposed solution. In view of these limitations, a Proof of Concept (PoC) was planned as a strategy to support administrative decision-making under conditions of greater technical safety, reducing the uncertainties associated with the implementation of the solution. It was conceived as a process of assisted practical demonstration in a controlled environment, conducted with the joint participation of the institutional technical team, academic specialists and representatives of the suppliers, with the objective of validating the compatibility between the acquired components, verifying the operational feasibility of the complete digital workflow and reducing the technological risks associated with the incorporation of the solution in the context of the SUS.

The PoC therefore represented a stage complementary to the traditional process of documentary evaluation, making it possible to increase the technical robustness of the decision-making process and to produce additional evidence capable of supporting the safe incorporation of a complex technological ecosystem on a large scale in the public health system. Furthermore, it made it possible to compare the documentary evidence with the performance observed in an environment representative of use, allowing the empirical validation of the critical aspects identified in the preliminary analysis, especially those related to compatibility, operational performance and infrastructure requirements.

The authors declare that they have an institutional affiliation with the Brazilian Agency for Support to the Management of the SUS (AgSUS), the institution responsible for the acquisition of the equipment and for conducting the proof of concept presented in this study. However, they have no financial or commercial ties with the manufacturers or suppliers of the technologies evaluated. The institutional affiliation did not influence the analysis of the results, their interpretation or the decision to publish.

III. METHODOLOGY

The present study is characterized as a descriptive case study that systematizes the technical evaluation process employed during the analysis of a digital dentistry solution undergoing public procurement. It is an applied study with a qualitative approach, oriented towards the analysis of a real decision-making process in a context of technological densification within the Brazilian Unified Health System (SUS). The evaluation was conducted sequentially, integrating document analysis, specialized technical analysis, statements by the supplier and practical demonstration, thus making it possible to consolidate different sources of evidence to support the administrative decision regarding the technical feasibility of the solution.

Initially, an analysis was carried out of the technical documentation made available during the contracting process, including equipment specifications, the requirements of the tender documents and the technical report prepared by the Federal University of Uberlândia (UFU), which assessed the performance of intraoral scanners and CAD software in the context of the digital prosthetic workflow, applied to the fabrication of single crowns, occlusal splints and complete dentures.

In the following stage, a critical analysis was made of the results presented in the report, considering aspects related to the performance of the equipment, the quality of the digital models produced, the required computational infrastructure, usability, processing time, learning curve and clinical applicability of the digital workflow. The analysis considered technical criteria previously defined on the basis of parameters usually employed in health technology assessment and clinical engineering, focusing on the identification of critical factors of system performance. The technical statements submitted by the supplying company were also assessed, with the aim of identifying complementary elements and possible divergences in relation to the technical evidence previously gathered that might support the analysis of the proposed solution.

Owing to the limitations inherent in an exclusively documentary evaluation, a complementary stage of assisted practical demonstration was carried out in a controlled environment and monitored by a multidisciplinary team composed of representatives of the Brazilian Agency for Support to the Management of the SUS (AgSUS), the Ministry of Health and specialists from the Federal University of Uberlândia (UFU) and the University of Brasília (UnB). The participants in the evaluation performed complementary roles throughout the PoC. The institutional team was responsible for the document analysis, the verification of adherence to the contracting requirements and the consolidation of the evidence to support the decision. The academic specialists contributed to the technical evaluation of the digital workflow, the interpretation of the results and the analysis of the clinical

and laboratory applicability of the solution. The suppliers' representatives carried out the operational demonstration of the equipment and software evaluated, as well as providing the technical support required for the execution of the planned tests. The practical demonstration was structured as a Proof of Concept (PoC), with the objective of assessing the integrated performance of the technological ecosystem under conditions representative of use.

During the practical demonstration, the following were evaluated: the performance of the intraoral scanner, the native software, the integration with complementary CAD software, the required technological infrastructure and the applicability of the digital workflow in different clinical scenarios, including the fabrication of single crowns, occlusal splints and complete dentures. The evaluation covered technical criteria related to the quality of digital image acquisition, interoperability between systems, operational stability of the workflow, usability, computational infrastructure requirements, processing time, reproducibility of results and adherence of the workflow to the care needs of the Brazilian Unified Health System.

The analysis adopted acceptance criteria related to the feasibility of executing the digital workflow, the capacity to exchange files between the different stages of the process, the operational stability of the systems and the clinical adequacy of the devices produced. Whenever applicable, aspects such as marginal fit, the possibility of completing the proposed workflow and the need for corrective interventions to enable the final result were also considered.

The digital models generated during the evaluation were exported, transferred and processed in STL format, used as the reference file for the subsequent stages of CAD modelling, dimensional assessment and fabrication of the prosthetic devices. The compatibility of the workflow was verified through the capacity to exchange these files between the different components of the technological ecosystem evaluated.

The table below presents the technical characteristics assessed and the evaluation matrix.

TABLE I. TABLE I. PROPOSED EVALUATION MATRIX

Characteristics	**Evaluation matrix**	**Evaluation matrix**
Characteristics	Indicator	Evaluation Method
Interoperability	Capacity for transfer between systems	Verification of the export, transfer and import of the files
Integrity of the information	Changes in the geometry of the scanned mesh; fit of the final product	Final product's adaptation
Time	Acquisition time	Timing
Time	Processing time	Timing
Efficiency	Total workflow time	Timing
Infrastructure	Adequacy of the computational requirements	Comparison between requirements and available infrastructure
Compatibility	Opening/import of files	Export, transfer and import test

[a.] Note: The proposed matrix represents an evaluation model applicable to different technological ecosystems. In the case studied, some of the indicators were employed in a qualitative and observational manner, owing to the characteristics of the demonstration carried out.

For the operationalization of the Proof of Concept, the criteria presented in the matrix were applied by means of structured observation of the complete digital workflow, covering tests of file export, transfer and import, verification of the execution of the clinical and laboratory stages, assessment of the operational stability of the systems, analysis of the adequacy of the computational infrastructure and appraisal of the clinical applicability of the devices produced.

Finally, the evidence obtained in all stages was consolidated into an integrated technical analysis, making it possible to compare the documentary results with the observations obtained in the practical demonstration and to substantiate the assessment of the technical feasibility of the solution from the perspective of technological densification within the Brazilian Unified Health System. The methodological sequence employed in the case studied was systematized in an evaluation framework (Fig. 1), structured so as to allow its adaptation to different technological ecosystems while maintaining the stages of analysis, definition of tests, evaluation, recording of results and decision-making. The indicators, test scenarios and acceptance criteria may be adjusted to the characteristics of each solution.

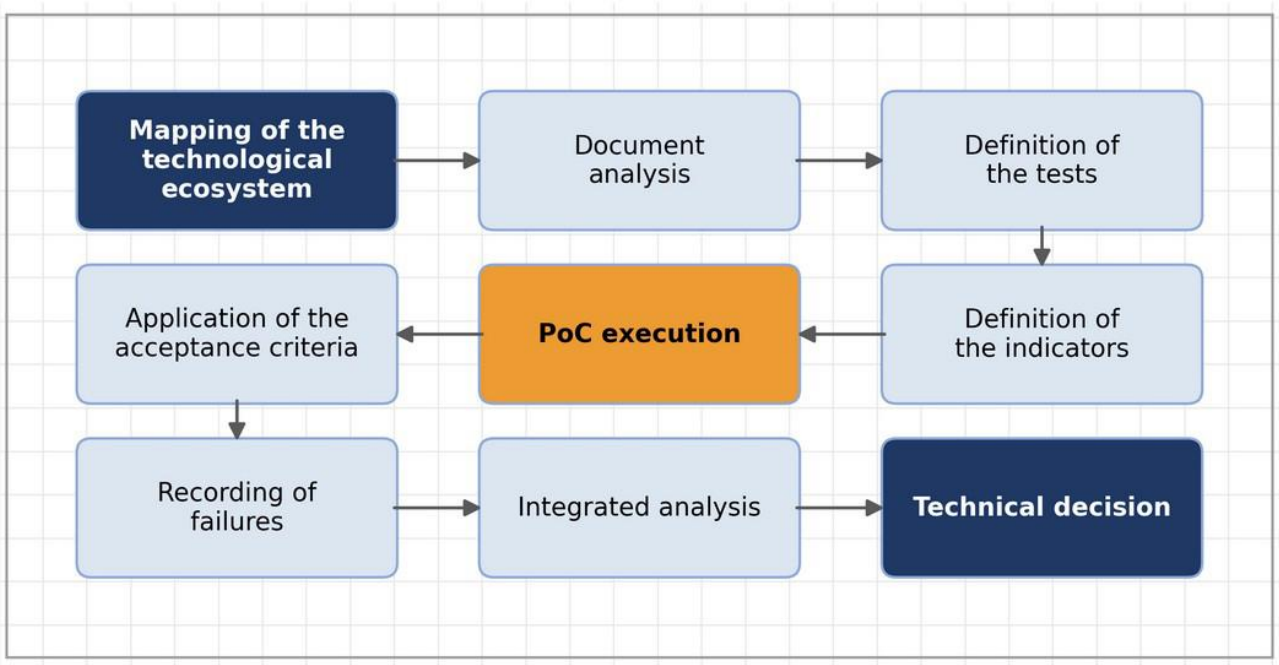


Fig. 1. Fig. 1.Evaluation Framework

This stage allowed for the empirical validation of the main critical factors identified in the preliminary analysis, especially those related to interoperability, systemic performance and operational requirements.

IV. RESULTS AND DISCUSSION

The PoC was conducted in a controlled technical environment with the objective of validating the operational feasibility of the digital dental prosthesis workflow acquired in the public contracting process. The validation focused mainly on assessing the integrated performance of all the technological components under conditions representative of operation, simulating the complete production workflow expected during the implementation of the solution in public health services.

Three levels of interpretation of the findings should be distinguished. At the level of documentary conformity, the components met the specifications of the tender documents and presented the appropriate regulatory documentation. At the level of observed technical behaviour, integrated execution revealed that the native software required external software, and that this software presented relevant limitations of reproducibility, operational stability and dependence on a continuous internet connection. The future organizational consequences discussed below, relating to contracting models, team training, centralization of digital planning and the role of clinical engineering, should be read as hypotheses to be tested.

The evaluation included interoperability tests between the intraoral scanner, the CAD/CAM/CAI software environment, the computational infrastructure, the three-dimensional printing system, the printing software, the post-processing equipment (washing and curing) and the consumables required for production, including dental resins and pigmentation kits. Execution was carried out jointly by technical representatives of the contracting institution, academic specialists, supplying companies and professionals experienced in digital workflows applied to dentistry. This multidisciplinary approach enabled an integrated analysis of the technical, operational and clinical aspects of the system evaluated.

The results showed that, although all the components individually met the technical specifications previously established in the procurement process, relevant operational limitations were identified during the integrated execution of the workflow. Variations in performance were observed, especially in relation to the interaction between different software environments, compatibility between exported and imported digital files, the computational requirements needed for processing the digital modelling stages and dependencies related to the functional limitations of the native software originally supplied by some manufacturers. These findings confirm that individual technical conformity is not sufficient to guarantee the systemic performance of the technological ecosystem when operating in an integrated manner, especially in scenarios dependent on interoperability between heterogeneous systems and on the structured exchange of information between different platforms [8], [10], [11], [15].

TABLE II. TABLE II. ASPECTS EVALUATED AND OBSERVATIONS RECORDED DURING THE POC

Aspect evaluated	Observation recorded during the PoC
Interoperability between components	Workflow executed with integration between scanner, CAD software and digital manufacturing
Native software	Limitations for the complete execution of the prosthetic workflow
Complementary software	Required for certain stages of the process
Dependence on connectivity	Identified in part of the solutions evaluated

Aspect evaluated	Observation recorded during the PoC
Computational infrastructure	Direct influence on stability and performance
Usability	Variable according to the software and the operator's experience
Complete denture	Greater operational complexity and dependence on specialized software

The findings made it possible to distinguish two complementary levels of interoperability. The first corresponds to syntactic interoperability, related to the capacity to export, transfer, import and use the files along the different stages of the digital workflow. The second refers to the preservation of the relevant characteristics of the transferred information, including the quality of the digital mesh, the continuity of the surfaces, the geometry, the spatial alignment of the records and other information necessary for the clinical use of the digital models. It was thus found that the mere possibility of exchanging files does not in itself constitute sufficient evidence of satisfactory performance of the technological ecosystem.

It should be noted that the present protocol does not replace interoperability standards or norms used in health care. Its purpose was to verify the practical operation of the digital dental workflow under conditions representative of use, assessing the interaction between equipment, software, computational infrastructure and digital files. Accordingly, the proposed approach should be understood as complementary to consolidated interoperability references such as HL7 FHIR, DICOM and IHE profiles, since it seeks to verify, in an operational context, aspects related to implementation, functional integration and the behaviour of the technological ecosystem. Within the scope of this study, formal conformity or certification of the systems in relation to these standards was not assessed.

The validation process also showed that certain stages of the workflow required complementary adjustments that could not be identified solely through the document analysis carried out during the contracting phase. In particular, it was found that more advanced digital planning activities require the use of specialized external software in addition to that originally provided with the equipment. This finding reinforces the limitations of evaluations based exclusively on documentary and regulatory conformity for complex digital technologies, in which factors related to implementation, integration between systems and context of use directly influence the observed performance [6], [7], [13], [16].

Despite the interoperability challenges observed, the PoC demonstrated that the proposed technological architecture was operationally feasible when appropriately configured, allowing the successful execution of the digital workflow from the acquisition of intraoral images to the final fabrication of the prosthetic device. The process generated relevant technical evidence confirming the possibility of implementing an open solution model for the digital production of prostheses in the context of the public health system, provided that minimum requirements related to interoperability, technological infrastructure and the use of specialized complementary software are observed [9,11].

The results obtained reinforce a methodological challenge that is increasingly present in contemporary processes of technological densification in health care: the growing mismatch between traditional models of technology assessment and the complexity of the new digital ecosystems applied to care delivery. Recent studies show that digital technologies require evaluative models capable of addressing aspects related to interoperability, implementation, usability, technological integration and value generation under real conditions of use [6, 13, 16].

Historically, public procurement and technology assessment processes have been structured on the basis of the individual analysis of discrete technologies, in which each component is assessed in isolation from technical specifications, regulatory documentation, certifications and its own performance criteria. However, the case analysed showed that, when health technologies operate as interconnected technological ecosystems, the individual conformity of the components ceases to be a reliable predictor of the operational performance of the integrated system. This finding is consistent with recent studies on digital health ecosystems, which identify interoperability and the capacity for integration between systems as critical factors for successful technological implementation [8, 10, 15].

The digital prosthetic workflow evaluated in this study represents a typical example of a multilayer technological ecosystem, in which clinical and operational performance does not derive from the isolated functioning of the individual devices, but from the dynamic interaction established between software, computational infrastructure, compatibility between hardware, compatibility between digital files, manufacturing processes, properties of the consumable materials and the coordinated execution of the clinical and production workflow. This characteristic is consistent with contemporary models of digital health ecosystems, in which the value generated depends on the interaction between multiple technological and organizational components [8,15].

This result reveals an important limitation of the models traditionally employed in the field of Health Technology Assessment, historically oriented towards the evaluation of medicines, isolated medical devices, diagnostic tests or specific therapeutic procedures. In contrast, the growing digitalization of health services requires evaluation approaches at a systemic level, capable of incorporating criteria of interoperability, technological integration and operational performance in a real context of use [6,7,13,16].

The adoption of a contracting strategy based on an open solution further increased this complexity. Although open architectures offer important advantages in terms of competition between suppliers, technological flexibility, lower dependence on single manufacturers and greater sustainability in public procurement processes, this model transfers to the Public Administration the responsibility for ensuring functional compatibility and compatibility between heterogeneous components. Recent studies indicate that medical device procurement processes still face relevant challenges in adequately incorporating aspects related to

interoperability, implementation and organizational impacts into decision-making [12].

From the perspective of engineering applied to health care, the results show that technological densification processes should not be restricted to the individual evaluation of devices when the object under analysis involves integration between hardware, software, digital workflow, additive manufacturing, compatibility between platforms/devices and operational dependence between multiple suppliers. In addition to clinical and economic aspects, the evaluation of such systems requires consideration of organizational and operational factors and of factors related to the interaction between technology and user, particularly in contexts of greater technological complexity [7,17].

In this context, the use of PoC methodologies proved to be a relevant complementary strategy for reducing technological uncertainties in public procurement processes, especially in scenarios involving complex and interdependent technological ecosystems whose overall performance cannot be adequately predicted solely by means of technical documentation or conventional regulatory analysis. In addition, the PoC made it possible to empirically validate critical criteria previously identified, strengthening the robustness of administrative decision-making.

This result reinforces the potential of proofs of concept as complementary instruments for the evaluation of complex technologies prior to their large-scale implementation, contributing to the analysis of interoperability, operational feasibility and the risks associated with technology adoption [9,11]

V. LIMITATIONS

The present study has limitations that should be taken into account when interpreting its results. It is a single case, conducted within a specific public contracting process, which partly restricts the generalization of its results to other institutional contexts. Moreover, the evaluation carried out in this article did not compare different ecosystems and was predominantly qualitative and observational. For this reason, some of the indicators presented in Table I could not be fully measured.

Although the technical reports used in the evaluation stage included specific measurements related to dimensional accuracy, processing times and operational performance of the equipment assessed, these results were not produced in a standardized manner within the context of the present Proof of Concept, and for this reason they were not used as comparable quantitative outcomes in this study.

The demonstration of the operation of the equipment took place in a controlled environment, with on-site support from the suppliers, a condition that does not faithfully reproduce the different realities of the Brazilian Unified Health System. In addition, the demonstration was carried out by professionals already familiar with the equipment. Under these conditions, execution times, the quality of the models and the capacity to solve software-related problems tend to represent a

performance scenario superior to that expected in routine operation. Accordingly, the findings of this article should be interpreted as indicative of the potential performance of the solution, provided that favourable conditions for its operation are in place. Finally, no post-implementation follow-up was carried out, so the longitudinal performance of the solution remains to be assessed.

VI. CONCLUSION

The present study suggests that the incorporation of complex digital health technologies, especially those structured as technological ecosystems made up of multiple interdependent components, may involve challenges that go beyond conventional models of technology assessment and the models traditionally employed in public contracting processes, particularly in the context of administrative decision-making. In the experience analysed, it was observed that the individual technical conformity of the equipment, compliance with regulatory requirements and conventional document analysis are not, in isolation, sufficient to ensure the operational performance of the solution, given its dependence on compatibility between different hardware and software components. These findings must, however, be interpreted in the light of the specific characteristics of the case studied and do not allow fully generalizable conclusions to be drawn for other technology implementation processes.

The results observed in this context indicate the relevance of considering digital health technologies not only as isolated devices, but also as integrated systems in which technological, operational and clinical aspects interact. The use of PoC methodologies made it possible to clarify technological uncertainties, compatibility aspects and operational issues before definitive implementation, producing complementary technical evidence capable of supporting safer and more robust decisions aligned with the real conditions of use in health services. In addition, the approach made it possible to identify and characterize challenges associated with the subsequent stages of incorporating the digital prosthesis workflow, under conditions close to those of its practical application.

A central aspect observed throughout the process was the interdisciplinary character required for the adequate validation of the solution evaluated, involving clinical, technological and management knowledge in the assessment of digital health solutions. The conduct of the PoC relied on the articulation between universities responsible for the production of technical and scientific knowledge and methodological support, dental surgeons responsible for the clinical validation of the digital workflow, supplying companies in charge of providing the technological infrastructure and operational support, and biomedical engineers, whose work contributed to the systemic analysis of the solution, the technical evaluation of the equipment, the verification of compatibility between components and the management of the technological risks associated with the incorporation process.

The experience provides preliminary evidence that the growing incorporation of digital technologies in the health sector may require approaches capable of integrating clinical knowledge, applied engineering, technology assessment, regulatory aspects, digital systems architecture and innovation management, especially in complex care environments. However, the applicability of this approach to other scenarios still requires validation in different institutional, care and technological contexts.

In addition, the results point to the need for the future development of more standardized methodological protocols for conducting PoC processes in public procurement

involving complex technological ecosystems, allowing greater reproducibility and the reduction of uncertainties during technological densification processes, especially in the context of solutions based on open architecture and multiple suppliers. Likewise, it becomes relevant to carry out longitudinal studies capable of assessing the behaviour of these solutions after their implementation in different realities of the Brazilian Unified Health System, considering factors such as operational sustainability, learning curve, clinical performance and technological maintenance over time.

Finally, the experience presented here shows that, in the context of technological densification, the work of Biomedical Engineering contributed to the integration of technical aspects, infrastructure, compatibility, technology assessment and management of technological risks. In this context, the integration between engineering, technology management and health technology assessment proved relevant in supporting the safe and sustainable implementation of the solution evaluated. Nevertheless, considering the specific characteristics of the case, further studies are proposed to assess the applicability of this approach to the contribution of Biomedical Engineering in different contexts of technology acquisition and health care delivery.

VII. FUTURE PERSPECTIVES

The results presented highlight opportunities for improving evaluation methodologies applied to the acquisition of technological ecosystems in health care. As a future perspective, it is recommended that standardized protocols for conducting Proofs of Concept (PoC) be developed and validated in multicentre settings, allowing greater reproducibility and comparability between different technological solutions.

Future studies may incorporate quantitative indicators related to interoperability, integrity of information, operational performance, usability, computational infrastructure and organizational impact, thereby increasing the robustness of the evaluation processes. In addition, it is relevant to investigate the application of the proposed methodology in other public procurement contexts and in different care realities of the Brazilian Unified Health System.

Longitudinal evaluations after the implementation of the solutions are also recommended, covering aspects such as the users' learning curve, operational sustainability, technological maintenance, clinical performance and the evolution of digital workflows over time. Finally, the integration of the proposed methodology with formal models for the evaluation of interoperability may contribute to the development of more comprehensive frameworks for the analysis of digital health ecosystems.

ACKNOWLEDGEMENTS

The authors thank the University of Brasília (UnB) for providing the infrastructure and physical space that made it possible to carry out the Proof of Concept (PoC) and the experimental activities of this study. They also thank the Federal University of Uberlândia (UFU) for the technical collaboration and scientific support provided during the process of evaluating and validating the digital dental workflow.

The authors further acknowledge the institutional partnership established with the Ministry of Health, through the Secretariat of Primary Health Care (SAPS) and the General Coordination of Oral Health (CGSB), which was essential for enabling the activities related to the development of the project.

Finally, the authors thank the Brazilian Agency for Support to the Management of the Unified Health System (AgSUS) for its institutional partnership, support, and encouragement in carrying out this research.

REFERENCES

[1] A. W. K. Yeung et al., "The promise of digital healthcare technologies," Frontiers in Public Health, vol. 11, Art. no. 1196596, 2023, doi: 10.3389/fpubh.2023.1196596.

[2] N. Naik et al., "Transforming healthcare through a digital revolution: A review of digital healthcare technologies and solutions," Frontiers in Digital Health, vol. 4, Art. no. 919985, 2022, doi: 10.3389/fdgth.2022.919985.

[3] S. A. Bernauer, N. U. Zitzmann and T. Joda, "The complete digital workflow in fixed prosthodontics updated: A systematic review," Healthcare, vol. 11, no. 5, Art. no. 679, 2023, doi: 10.3390/healthcare11050679.

[4] J. Abduo and V. Rasaie, "Digital workflows in prosthodontics," Australian Dental Journal, vol. 70, Suppl. 1, pp. S146–S158, 2025, doi: 10.1111/adj.70005.

[5] C. Auduc, T. Douillard, E. Nicolas and N. El Osta, "Fully digital workflow in full-arch implant rehabilitation: A descriptive methodological review," Prosthesis, vol. 7, no. 4, Art. no. 85, 2025, doi: 10.3390/prosthesis7040085.

[6] A. von Huben et al., "Application of a health technology assessment framework to digital health technologies that manage chronic disease: A systematic review," International Journal of Technology Assessment in Health Care, vol. 38, no. 1, Art. no. e9, 2022, doi: 10.1017/S0266462321001665.

[7] F. S. Toscas et al., "Domains and methods of medical device technology evaluation: A systematic review," Public Health Reviews, vol. 45, Art. no. 1606343, 2024, doi: 10.3389/phrs.2024.1606343.

[8] R. Huettemann, B. Sevov, S. Meister and L. Fehring, "How to establish digital health ecosystems from the perspective of health service-organizations: A taxonomy developed based on expert interviews conducted as modified Delphi approach," Digital Health, vol. 10, 2024, doi: 10.1177/20552076241271890.

[9] J. E. W. C. van Gemert-Pijnen, "Implementation of health technology: Directions for research and practice," Frontiers in Digital Health, vol. 4, Art. no. 1030194, 2022, doi: 10.3389/fdgth.2022.1030194.

[10] E. Li, J. Clarke, H. Ashrafian, A. Darzi and A. L. Neves, "The impact of electronic health record interoperability on safety and quality of care in high-income countries: Systematic review," Journal of Medical Internet Research, vol. 24, no. 9, Art. no. e38144, 2022, doi: 10.2196/38144.

[11] L.-G. Lemus-Zúñiga et al., "A proof-of-concept IoT system for remote healthcare based on interoperability standards," Sensors, vol. 22, no. 4, Art. no. 1646, 2022, doi: 10.3390/s22041646.

[12] M. Cangelosi, A. Chahar and S. Eggington, "Evolving use of health technology assessment in medical device procurement—Global systematic review: An ISPOR special interest group report," Value in Health, vol. 26, no. 11, pp. 1581–1589, 2023, doi: 10.1016/j.jval.2023.06.005.

[13] H. Mumtaz et al., "Current challenges and potential solutions to the use of digital health technologies in evidence generation: A narrative review," Frontiers in Digital Health, vol. 5, Art. no. 1203945, 2023, doi: 10.3389/fdgth.2023.1203945.

[14] M. Callihan et al., "Proof of concept testing of safe patient handling intervention using wearable sensor technology," Sensors, vol. 23, no. 12, Art. no. 5769, 2023, doi: 10.3390/s23125769.

[15] N. M. R. Rajkumar, M. R. Muzoora and S. Thun, "Dentistry and interoperability," Journal of Dental Research, vol. 101, no. 11, pp. 1258–1262, 2022, doi: 10.1177/00220345221100175.

[16] J. Segur-Ferrer, C. Moltó-Puigmartí, R. Pastells-Peiró and R. M. Vivanco-Hidalgo, “Methodological frameworks and dimensions to be considered in digital health technology assessment: Scoping review and thematic analysis,” Journal of Medical Internet Research, vol. 26, Art. no. e48694, 2024, doi: 10.2196/48694.

[17] J. Ming et al., “Health technology assessment of medical devices: Current landscape, challenges, and a way forward,” Cost Effectiveness and Resource Allocation, vol. 20, no. 1, Art. no. 54, 2022, doi: 10.1186/s12962-022-00389-6.

Importance of Metrological Performance Testing for New and Post-Corrective Maintenance Medical Equipment

1st Juliano Martins
Federal University of Santa Catarina/Department of Electrical and Electronic Engineering, Biomedical Engineering Institute
Florianópolis - SC, Brazil
[0000-0003-4655-823X]

2nd Renato Garcia
Federal University of Santa Catarina/Department of Electrical and Electronic Engineering, Biomedical Engineering Institute
Florianópolis - SC, Brazil
[0000-0001-7732-970X]

***Abstract*— Clinical Engineering (CE) plays a relevant role in the safety and reliability of medical equipment. However, constant challenges persist, such as metrological control. An example is the assessment of Non-Invasive Blood Pressure (NIBP) modules, particularly in acceptance testing for new equipment and post-corrective maintenance. This descriptive case study evaluated 87 performance test reports of 53 multiparameter monitors across two Emergency Care Units (ECUs) and two hospitals over a three-year period. The analysis revealed that only 20% of the tested devices were fully compliant with normative limits. Notably, none of the 23 tests performed after corrective maintenance achieved full compliance, indicating vulnerabilities in traceability and quality control. Focusing on the 120/80 (93) mmHg reference point-the threshold for hypertension screening 14 out of 33 new devices and 15 out of 23 post-maintenance devices exceeded the ±3 mmHg maximum permissible error. These results suggest that the release of equipment without metrological verification may introduce hidden deviations, thereby reducing diagnostic accuracy. The study advocates for the systematic integration of metrological testing throughout the technology's life cycle as a viable Health Technology Management (HTM) tool to mitigate clinical risks, highlighting the need for future multicenter and inferential studies.**

Keywords—Clinical Engineering, Health Metrology, Metrological Testing, Quality Control, HTM.

I. INTRODUCTION

Metrology in the healthcare sector has gained significant prominence over the years, becoming essential for patient safety and the quality of medical procedures, as evidenced by performance testing. Medical Hospital Technologies (MHT) that perform measurements have been used in medicine for many years; however, it was only a few decades ago that legal metrological control, combined with the quality policies of ISO/IEC 17025 (General requirements for the competence of testing and calibration laboratories), established MHT performance testing as an integral part of medical technology management. Performance test reports facilitate the interpretation and evaluation of the reliability of technologies used in Health Care Facilities (HCF) [1]. Clinical Engineering (CE) serves as the cornerstone for fostering metrological awareness and technovigilance within HCFs, bridging the gap between technical oversight and patient safety [2], aligning with technovigilance principles (a post-market surveillance system designed to identify and evaluate adverse events and technical complaints) among all healthcare professionals. Performance testing is a key indicator of safety for both patients and operators; if equipment does not perform adequately, misinterpretations and diagnostic errors may occur, compromising the therapy or procedure to which the patient is subjected.

One of the most common medical technologies found in HCFs is the non-invasive blood pressure (NIBP) monitor. These devices can be analog (using an aneroid sphygmomanometer and stethoscope), semi-automatic, or automatic. Currently, many HCFs, including those in primary care, utilize automatic equipment, a practice protocolized by the Brazilian Society of Hypertension (SBH). It is important to highlight that while these devices provide static readings, facilitating the operator's task, minor deviations in measurement results can impair equipment performance and compromise the correct approach to treatment or the classification of the patient being attended.

The 2025 Brazilian Guideline on Arterial Hypertension (AH), in chapters 3 and 5, recommends "classifying pre-hypertension," encompassing systolic blood pressure (SBP) values between 120-139 mmHg or diastolic blood pressure (DBP) between 80-89 mmHg in clinical settings, with the goal of early identification of at-risk individuals and encouraging proactive, non-pharmacological interventions to prevent progression to AH [3]. Furthermore, it recommends pharmacological treatment after 3 months for individuals with BP of 130-139/80-89 mmHg and high cardiovascular (CV) risk. The threshold differences between classifying a patient as pre-hypertensive versus hypertensive with CV risk are minimal. The summation of minor procedural and equipment measurement errors can lead to a misdiagnosis when the patient's BP is in the "borderline risk" zone between normal and established hypertension [3]. According to statistics from the Brazilian Ministry of Health, nearly 30% of the adult population in Brazil is hypertensive.

The objective of this article is to demonstrate not only the well-established importance of performance testing but also the critical need for its correct interpretation, specifically emphasizing tests conducted after corrective maintenance and during the acceptance phase for new equipment. This approach aims to strengthen the "safety layer" in the performance and operation of MHT.

II. Materials and Methods

This study was conducted at the Health Metrology Laboratory of the Biomedical Engineering Institute at the Federal University of Santa Catarina (IEB-UFSC), which maintains various protocols for evaluating the performance and safety of diverse Health Technologies (HT). Among these is the Internal Testing Procedure PE-209 - Non-Invasive Blood Pressure Monitor Performance Test (a test validated and audited by INMETRO), which evaluates non-invasive blood pressure (NIBP) measurement technologies, generating the reports of interest for this study.

PE-209, like the laboratory's other procedures, is based on validated protocols from the Emergency Care Research Institute (ECRI). For NIBP specifically, it follows the 454-0595 "Blood Pressure Monitors, Electronic Indirect" and 454-20130812 "Health Devices Inspection and Preventive Maintenance System" guidelines, as well as the technical standard ABNT NBR IEC 80601-2-30 - Particular requirements for the basic safety and essential performance of automated non-invasive sphygmomanometers.

The procedure consists of a qualitative evaluation of the equipment under test, subjecting it to a comprehensive visual and operational inspection of items such as the power cord, housing, cuff, software, and battery, among other elements that could compromise measurement results. It also includes a quantitative evaluation (calibration), which compares the blood pressure (systolic, diastolic, and mean) in millimeters of mercury (mmHg) measured by the equipment with the corresponding blood pressure values selected on a calibrated patient simulator. The test includes 7 adult pressure simulation presets and 4 neonatal pressure simulation presets:

TABLE I. Simulated Adult Parameters

Preset	Blood pressure (mmHg) systolic/diastolic (average)
1	120 / 80 (93)
2	150 / 100 (116)
3	200 / 150 (166)
4	255 / 195 (215)
5	60 / 30 (40)
6	80 / 50 (60)
7	100 / 65 (76)

TABLE II. Simulated Neonatal Parameters

Preset	Blood pressure (mmHg) systolic/diastolic (average)
1	35 / 15 (22)
2	60 / 30 (40)
3	80 / 50 (60)
4	100 / 70 (80)

The tests are conducted under the following environmental and mains voltage conditions (for mains-powered equipment) [6]:

- Temperature: 15°C to 35°C.
- Relative Humidity: 45% to 75%.
- Atmospheric Pressure: 860 hPa to 1060 hPa.
- Mains voltage: V_{mains} (V) ±10 %.

All auxiliary instruments used to verify the test conditions, humidity, temperature, barometric pressure, and mains voltage, as well as the patient simulator, are calibrated with valid traceability as established by the applicable technical standard [7]. The tests take place either at the Healthcare Facility (EAS) or at the IEB-UFSC metrology laboratory. Environmental and mains conditions are measured at the test site before the start and at the end of the test using auxiliary equipment, recording the corresponding values on the datasheet for subsequent inclusion in the test report. The report presents the data and measurements collected and processed according to the procedure, formatted as tables. The qualitative evaluation table uses the concepts of Compliant, Non-compliant, or Not Applicable. "Compliant" indicates that the evaluated item meets the recommendations outlined in PE-209; "Non-compliant" indicates that the equipment does not meet the recommendations; and "Not Applicable" indicates that the equipment lacks the item or the item cannot be evaluated, in which case the reason must be stated in the remarks field. The calibration table includes the measurement average (mean of three readings), the error (E), the uncertainty (U), and the coverage factor (k) [4].

In metrology, the concepts of error, uncertainty, and coverage factor complement measurement reliability, as equipment exhibiting high dispersion correlated with usage time may suggest the need for replacement. By definition, the test report is not "conclusive" [6]; that is, it does not state whether the equipment is "approved" or "rejected." This determination is a subsequent step, falling under the responsibility of the Clinical Engineering (CE) team in conjunction with the clinical staff. The objective of the report is to provide qualitative and metrological data, leaving the decision-making to the end user. When applicable, specific technical standards may be included in this decision. Therefore, as an interpretation aid, the normative limits for NIBP are provided at the end of the report, as established by INMETRO's legal metrology in standards NIT-SEFIQ-026 (verification of electronic sphygmomanometers) and NIT-SEFIQ-025 (verification of mechanical sphygmomanometers). Under these guidelines, the maximum permissible error is ±4 mmHg for periodic testing (equipment in use, voluntary verification) and ±3 mmHg for testing new equipment or post-repair [5].

The case study presented herein results from the analysis of performance test reports of NIBP modules integrated into multiparameter monitors from two Emergency Care Units (ECUs) and two hospitals in the metropolitan region of Florianópolis over a three-year period. Authorization to access the reports was granted by the metrology laboratory, adhering to confidentiality and impartiality policies [6].

The analyzed sample consisted of 53 multiparameter monitors encompassing 9 distinct brands. To preserve the study's commercial neutrality, the manufacturers were not nominally identified. Out of the total equipment, 18 are domestically manufactured and 35 are imported. Regarding technological age and usage time, the sample comprises 33 new devices (subjected to acceptance testing), 9 devices with up to 5 years of use, and 11 devices operating for over 5 years in the healthcare environment.

The objective of the study focuses on the NIBP measurement results from the reports that are used to proactively identify individuals predisposed to arterial hypertension (AH) [4]. Serving as an aid in clinical decision-making for adult patients, these involve measurements close to 120–139 mmHg for systolic BP and 80–89 mmHg for diastolic BP. Within the report, the measurement of interest is the simulated measurement in preset 1 for adult patients: 120/80 (93) mmHg — systolic/diastolic (mean). The selection and prioritization of this preset are justified as it represents the borderline reference value in clinical guidelines for hypertension screening and classification, where metrological deviations may contribute to inadequate classification. The CE database was also accessed to verify, within each equipment's history, the occurrence or absence of performance test requests during equipment incorporation, post-corrective maintenance, and/or on a periodic basis.

III. Results and Discussion

The sample used in the case study comprised a total of 87 performance test reports conducted on equipment belonging to two Emergency Care Units (ECUs) (35 tests) and two medium-complexity hospitals (52 tests) in the metropolitan region of Florianópolis. The sample consists of tests on new equipment, periodic tests, and post-corrective maintenance tests, as presented in the tables below:

TABLE III. Distribution of tests across ECU and Hospitals

Test Type	ECU	Hospitals
Acceptance Testing (New)	3	30
Post-Corrective Maintenance	12	11
Periodic Testing	20	11

Over the three years comprising the sample, some devices naturally underwent more than one test. The number of registered devices in the two ECUs is 12, and in the two hospitals, it is 41, totaling 53 multiparameter monitors. The sample encompasses devices from 9 distinct brands (18 domestically manufactured and 35 imported). Regarding technological age and usage time, the baseline consists of 33 new devices, 9 devices with up to 5 years of use, and 11 devices operating for more than 5 years.

Upon analyzing Table 3, considering the number of registered devices versus the number of tests, subtle divergences in the technological management methodology among the Healthcare Facilities (EAS) become apparent. While the ECUs conducted a higher number of post-corrective maintenance and periodic tests, testing during equipment incorporation (commissioning) showed low adherence. In contrast, the hospitals exhibited a significant number of incorporation tests, but the frequency of periodic and post-maintenance testing was low. This difference does not necessarily prove the existence of distinct institutional "cultures" or negligence; rather, it indicates varying levels of maturity in local protocols and points to a vulnerability in quality control throughout the technology's life cycle.

The reality in the ECUs demonstrates that, of the 12 registered devices, 3 underwent testing upon incorporation, 8 were incorporated without testing, and 1 was already in the unit prior to the analyzed period. Although the volume of post-corrective tests in primary care appears high, these analyses were concentrated on only 4 devices, as they required the most interventions according to the CE history. It was observed that 6 devices were never tested, and there were 4 records of corrective maintenance where the equipment was returned to clinical use without metrological evaluation.

In the hospitals, out of 41 devices, 30 were tested upon receipt. However, 33 devices did not undergo periodic testing during the period, indicating a departure from the guidelines of the NBR IEC 62353 standard, which recommends recurrent and post-repair testing [7]. The 11 post-corrective tests were concentrated on 8 devices, and, similar to the ECUs, there were 13 recorded instances of post-repair return without testing. Another detail found in the database: in one of the hospitals, there were 2 occurrences of periodic testing without linkage to a service order in the history, evidencing flaws in documentary traceability.

The overall approval results for the 87 reports are detailed in Table 4, specifying compliance with the PE-209 parameters:

TABLE IV. OVERALL RESULTS OF THE TESTS IN THE TWO ECUS (EMERGENCY CARE UNITS) AND TWO HOSPITALS

Test Type	Number of Tests	100% Compliant
Acceptance Testing (New)	33	12
Post-Corrective Maintenance	23	0
Periodic Testing	31	7

In Table 4, the operational definition of "approved" requires the absence of qualitative non-conformities and errors strictly within normative limits [6], resulting in an overall compliance rate of only 20%. It is emphasized that the metrological report does not impose automatic rejection; the final decision lies with the CE and the clinical staff. For example: if a device allocated to an Adult ICU presents a failure only in the neonatal simulation, the CE, evaluating

demand and criticality, may keep it in use restricted to adults. However, if this same scenario occurs during the acceptance of a new device, it must be returned to the supplier. From the perspective of operational and economic impact, the systematic adoption of these tests demonstrates viability, given that the cost of a test is lower than the costs associated with the possibility of maintenance rework, equipment downtime, institutional attrition, or the potential for interference in clinical management and complementary exams.

The metrological focus of the study centers on preset 1 of PE-209 for adult patients: 120/80 (93) mmHg - systolic/diastolic (mean). This choice is based on the borderline reference value in the guidelines for arterial hypertension (AH) screening [4]. For the measurement result to be compliant, a maximum permissible error of ±3 mmHg [6] is considered for new and post-maintenance equipment; the sample results are presented in Tables 5 and 6:

TABLE V. OVERALL RESULT: PRESET 1, ADULT, NEW EQUIPMENT

Measurement mmHg	Compliant	Non-compliant
120 systolic	33	0
80 diastolic	19	14
93 average	25	8

TABLE VI. OVERALL RESULTS FOR PRESET 1: ADULT EQUIPMENT POST-MAINTENANCE

Measurement mmHg	Compliant	Non-compliant
120 systolic	10	13
80 diastolic	10	13
93 average	12	11

Consolidating the measurement results for preset 1 from each report, 19 were compliant while 14 exceeded the error limit; for the results in post-corrective test reports, 8 were compliant while 15 exceeded the error limit. Evaluating the magnitude of the deviations, the maximum error found in the systolic measurement was +7 mmHg (post-corrective), in the diastolic measurement +7 mmHg (post-corrective), and in the mean measurement -8.67 mmHg (new). To exemplify the potential interference, an undetected error of +7 mmHg in a clinically borderline patient (e.g., 135/85 mmHg) could result in an inadequate classification for the hypertension stage, potentially leading the medical team to unnecessary pharmacological intervention. Although there are no documented harms to patients in this study, the data substantiate the clinical risk.

The overall total for the measurement results in preset 1 was 45 compliant tests and 42 exceeding the error limit. Isolating only the systolic and diastolic pressures (the primary metrics for AH diagnosis), the number shifts to 64 compliances and 23 exceeding the error limit. Non-compliance in the mean pressure measurement should not be underestimated, as it indicates the effectiveness of blood flow to vital organs (perfusion), which could interfere with the management of critical patients [8].

A limitation identified in the documentary analysis is that the reports certify non-compliance but do not discriminate the root cause (such as failures in the piezoelectric transducer, pneumatic leaks, or component fatigue). The absence of this differentiation in the reports restricts deeper predictive analyses.

Finally, the concern regarding NIBP monitors is validated on a national level. According to INMETRO, periodic inspections are essential to ensure reliability and prevent health impacts [8]. In the latest national action in April 2026, out of 1,994 inspected instruments, 116 presented irregularities [8], corroborating the relevance of the constant vigilance demonstrated in this local case.

IV. CONCLUSION

The primary contribution of this study consists of the integrated analysis of the metrological performance of multiparameter monitors from the perspectives of acceptance, post-corrective maintenance, and periodic monitoring. NIBP measurement technology is constructively complex, encompassing mechanical, pneumatic, and electronic components that require constant vigilance [9], whether by Clinical Engineering (CE) or testing laboratories. Furthermore, the predominance of digital equipment in healthcare facilities may create a false sense of security, given that metrological defects in these devices are not visually apparent unlike the deviation of a pointer in a mechanical sphygmomanometer making the execution of performance testing even more indispensable.

The analysis revealed a high percentage of non-conformities across the three tested conditions. The most sensitive finding points to metrological deviations precisely at the pressure levels used to classify the stages of hypertension. Although this descriptive study does not document adverse events, such errors signal a potential clinical risk, as they may induce diagnostic inaccuracies and undesirably influence therapeutic management.

As the main lesson learned, the study highlights the vulnerability of commissioning based solely on the presumption of factory conformity. Within the scope of this work, the non-conformity of an entire batch comprising 10 new units was observed. Consequently, formal notification to the manufacturer and regulatory agencies (such as ANVISA) is recommended. A preventive action is also recommended for CE: during the incorporation of new technologies, the technical specifications in the bidding process [10] can stipulate the requirement of a performance test report, to be included in the documentation upon delivery as a binding criterion for final acceptance.

The case study evidences subtle variations in institutional metrological awareness. While in some situations metrology is well integrated into Health Technology Management (HTM), in others, a fragility in traceability and quality control is observed. The evolution of this technical culture should be fostered starting from professional training to mitigate risks. It is also noteworthy that hidden failures can stem not only from

the equipment itself but also from the incorrect choice of accessories or patient conditions, demanding a systemic view from the healthcare team.

Finally, the limitations of this research are acknowledged. Among the remaining research gaps, future studies are suggested to investigate the root causes of the non-conformities, conduct multicenter validation of the findings, and apply inferential statistical analyses. The development of economic impact studies will also be fundamental to demonstrate the operational viability of metrology as an essential pillar for patient safety.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] J. Martins, "Metrologia ferramenta de melhoria da segurança e efetividade da oximetria de pulso na atenção básica," in *Metrologia para a indústria 4.0*, 2019, código CBM 87.

[2] J. Martins, "Clinical Engineering approach to the reliability of the performance of technologies used in Home Care," in *X Latin American Congress on Biomedical Engineering (Claib2024)*, 2024.

[3] Diretriz Brasileira de Hipertensão Arterial – 2025, ISSN 0066-782X, ISSN 1678-4170, 2025.

[4] INMETRO, Avaliação de dados de medição: Guia para a expressão de incerteza de medição – GUM, 1ª ed. brasileira, 2012.

[5] INMETRO, Verificação de Esfigmomanômetro Eletrônico, Norma NIT-SEFIQ-026, 2022.

[6] ABNT, Equipamento eletromédico - Ensaio recorrente e ensaio após reparo de equipamento eletromédico, ABNT NBR IEC 62353:2019, 2019.

[7] INMETRO, "Operação Dia Mundial da Saúde: Inmetro fiscaliza medidores de pressão arterial em todo país." Portal INMETRO. [Online]. Available at: https://www.inmetro.gov.br. [Accessed on: Apr. 15, 2026].

[8] D. DeMers and D. Wachs, "Physiology, Mean Arterial Pressure," in StatPearls. Treasure Island (FL): StatPearls Publishing, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK538226/

[9] L. M. Figueroa Gallo, "Evaluación y Mejora de la Gestión de Tecnología Médica en el Sector Privado de Salud," in *XXV Congreso Argentino de Bioingeniería (SABI 2025)*, 2025.

[10] J. Martins, "Medical Equipment Acquisition Methodology in Public Procurement," in The second International Clinical Engineering and Technology Management Congress (II ICEHTMC 2017), 2017.

Integrating Performance Indicators and Quality Tools to Support Decision-Making in Clinical Engineering

Marcelo Perrella
Graduate Program in Biomedical Engineering, USP
São Paulo, Brazil
ORCID: 0009-0007-1456-0040

Simone R. Argenton-Perrella
MBA in Pet Market Management
FMVZ, University of São Paulo
Pirassununga, Brazil
ORCID: 0009-0000-9192-1574

Henrique T. Moriya
Biomedical Engineering Lab.
University of São Paulo (USP)
São Paulo, Brazil
ORCID: 0000-0002-1595-1779

Alexandre H. Hermini
Clinical Engineering Program
Albert Einstein Hospital
São Paulo, Brazil
ORCID: 0009-0000-1200-4437

***Abstract*—Performance indicators and quality management tools are widely used in maintenance assessment; however, their integrated application to support evidence-based decision-making in Clinical Engineering remains limited. This study presents an integrated framework combining performance indicators and quality management tools to transform routine corrective maintenance records into strategic information for healthcare technology management. A retrospective, documentary, observational, and quantitative study was conducted using maintenance records from a Brazilian federal university hospital. During April 2026, the institution recorded 1,349 maintenance work orders, of which 239 corresponded to corrective maintenance and were included in the analysis. Maintenance performance was assessed using the Mean Time to Repair (MTTR), which was 1.92 days. Repair times were characterized using a histogram, followed by Pareto analysis to identify the equipment contributing most to cumulative corrective maintenance time. Five prioritized prolonged-repair cases were investigated using Ishikawa cause-and-effect diagrams, and improvement actions were structured according to the Plan–Do–Study–Act (PDSA) cycle. Recurrent factors associated with prolonged repair times included equipment aging, limited availability of replacement components, administrative process constraints, lack of standardized maintenance procedures, and insufficient technology renewal planning. Within the scope of this single-center, one-month analysis, the workflow demonstrated its potential to organize maintenance information and support prioritization and improvement planning; longitudinal evaluation is still required to determine whether the proposed actions reduce MTTR or improve maintenance performance over time.**

***Index Terms*—Clinical Engineering, Performance Indicators, Mean Time to Repair (MTTR), Quality Management Tools, Decision-Making**

I. Introduction

The increasing incorporation of medical technologies into healthcare systems has substantially increased the complexity of hospital services, making equipment availability, reliability, lifecycle management, and evidence-based decision-making essential components of safe and efficient patient care. In this context, Clinical Engineering has evolved from a predominantly technical support function into a strategic discipline responsible for managing healthcare technologies throughout their entire lifecycle, including planning, acquisition, installation, maintenance, performance evaluation, obsolescence management, and technology replacement [1]–[3].

The widespread implementation of computerized maintenance management systems (CMMS) has enabled the systematic collection of operational information generated during routine maintenance activities. Consequently, maintenance work orders have evolved beyond administrative records to become valuable sources of evidence capable of supporting performance assessment, maintenance planning, resource allocation, risk management, and strategic decision-making in Clinical Engineering [4]–[6]. Among the indicators used to evaluate maintenance performance, the Mean Time to Repair (MTTR) is one of the most widely adopted measures of maintainability.

MTTR estimates the average time required to restore equipment to operational condition following a failure and reflects not only the efficiency of technical repair activities but also organizational factors such as spare-part availability, administrative workflows, logistics, technical support, and maintenance planning [7], [8]. Consequently, MTTR should be interpreted as a system-level maintenance performance indicator rather than simply a measure of repair duration. Despite its broad applicability, MTTR alone does not explain why repair times vary among equipment or which factors contribute most to maintenance delays.

Performance indicators therefore require complementary analytical approaches capable of identifying priority equipment, investigating potential contributing factors, and supporting continuous quality improvement initiatives. In this context, quality management tools—including histograms, Pareto charts, Ishikawa cause-and-effect diagrams, and the Plan–Do–Study–Act (PDSA) cycle—provide complementary perspectives that facilitate systematic analysis of maintenance processes and support evidence-based quality improvement [9]–[12].

Although performance indicators and quality management tools are individually well established, studies integrating these approaches into a structured analytical workflow for

Clinical Engineering remain limited, particularly those based on routinely collected maintenance data from university hospitals. Bridging this gap may enhance the analytical value of maintenance records by transforming operational data into strategic information capable of supporting maintenance management, technology renewal planning, and evidence-based decision-making [2], [4], [5].

Therefore, this study proposes an integrated analytical framework combining the Mean Time to Repair (MTTR), histogram analysis, Pareto charts, Ishikawa cause-and-effect diagrams, and the Plan–Do–Study–Act (PDSA) cycle to evaluate corrective maintenance performance in a Brazilian federal university hospital. By integrating complementary analytical tools into a sequential decision-support workflow, the proposed approach demonstrates how routinely collected maintenance data can be transformed into actionable information to support evidence-based decision-making and continuous improvement in healthcare technology management.

II. Material and Methods

A. Study Design

This retrospective, documentary, observational, and quantitative study was conducted using maintenance records from the Clinical Engineering Department of a Brazilian federal university hospital. The analyses were based on data extracted from the institution's computerized maintenance management system (CMMS), including all maintenance work orders registered during April 2026. During the study period, a total of 1,349 maintenance work orders were recorded. Among these, 239 corresponded to corrective maintenance and were included in the analysis. Preventive maintenance, calibration activities, inspection records, duplicated entries, incomplete work orders, and administrative records without corrective interventions were excluded.

B. Maintenance Performance Indicator

Corrective maintenance performance was evaluated using the Mean Time to Repair (MTTR). In this study, MTTR was operationally defined as the mean elapsed time between the Service Order Request (SOS), which marks the beginning of the corrective repair process, and the Work Order Ready/Completed status (OSP), which marks its completion in the CMMS. This definition encompasses the complete corrective repair process recorded by the institution rather than only hands-on technical intervention time. Individual Time to Repair (TTR) values were expressed in days, and MTTR was calculated for the 239 corrective maintenance work orders included in the final analytical dataset.

Because repair-time distributions may be asymmetric, MTTR was complemented by the median, interquartile range (IQR), 90th and 95th percentiles, observed range, skewness, and a 95% bootstrap confidence interval for the mean. MTTR was retained as the principal maintenance indicator, whereas prioritization was interpreted from the combined distributional and Pareto analyses rather than from the mean alone. MTTR was calculated as

$$\mathrm{MTTR} = \frac{\sum_{i=1}^{N} \mathrm{TTR}_i}{N}, \tag{1}$$

where MTTR is the Mean Time to Repair (days), $\sum \mathrm{TTR}_i$ is the sum of the individual SOS–OSP intervals for all corrective maintenance work orders, and N is the number of corrective maintenance work orders in the final analytical dataset ($N = 239$).

C. Analytical Workflow

The analytical workflow was designed to progressively characterize, prioritize, investigate, and address factors associated with prolonged corrective maintenance times. Initially, the distribution of individual Time to Repair (TTR) values was examined using a histogram to characterize variability and asymmetry. Descriptive statistics were calculated to complement the graphical assessment, including the median, IQR, percentiles, observed range, skewness, and a 95% bootstrap confidence interval.

For equipment prioritization, "SUPPORT" work orders, representing miscellaneous technical services such as accessory replacement, were excluded without affecting the 239 records used to calculate overall MTTR. Among the remaining records, TTR values above the overall MTTR formed an exploratory screening set ranked in descending order. This step was not interpreted as a validated clinical or engineering criticality threshold. A Pareto chart examined the screened records' contribution to cumulative corrective maintenance time [12]. The five highest-ranked cases were selected for qualitative investigation using Ishikawa cause-and-effect diagrams and are reported in Table I. Potential contributing factors were organized into the categories Equipment, Parts and Materials, Methods, Workforce, Management, and Suppliers [9], [10].

The diagrams were developed from information available in the maintenance records and were subsequently reviewed by five Clinical Engineering professionals (three technicians and two clinical engineers) to assess the technical plausibility and operational relevance of the mapped factors. This expert review supported the qualitative interpretation of the diagrams but did not constitute quantitative causal validation. Based on the potential contributing factors identified, improvement actions were structured according to the Plan–Do–Study–Act (PDSA) cycle as a planning framework [11].

In the present study, the PDSA actions were proposed from the analytical findings; their longitudinal implementation and post-intervention effect on MTTR were not evaluated. Therefore, no claim of intervention efficacy was made.

III. Results and Discussion

A. Maintenance Performance

During the study period, 239 corrective maintenance work orders, extracted from a total of 1,349 maintenance records, were analyzed. The calculated Mean Time to Repair (MTTR) was 1.92 days. The median TTR was 1.0 day (IQR: 0–2 days), the 90th and 95th percentiles were 5 and 7 days,

respectively, and the observed range was 0–27 days. The distribution was markedly right-skewed (skewness = 3.98). The 95% bootstrap confidence interval for the mean was 1.56–2.33 days. These statistics demonstrate that the mean was influenced by a relatively small number of prolonged repair intervals and should therefore be interpreted together with robust distributional measures.

In this study, MTTR represents the complete SOS–OSP interval recorded in the CMMS, from the service request that initiates the corrective repair process to the work-order ready/completed status. Accordingly, the indicator reflects the institution's operational repair process rather than only hands-on technical intervention time. This operational definition should be considered when comparing the present MTTR with indicators derived from other maintenance systems or definitions. This broader interpretation extends the analytical value of MTTR beyond repair efficiency alone.

By reflecting the overall performance of the maintenance management system, the indicator provides evidence to support decisions related to resource allocation, technology renewal planning, obsolescence management, and maintenance strategy. Recent studies have similarly emphasized that maintenance performance indicators constitute essential components of evidence-based maintenance management in healthcare institutions [5], [8]. Nevertheless, although MTTR objectively quantifies maintenance performance, it does not explain the factors responsible for prolonged repair times. Therefore, complementary analytical tools were required to characterize repair-time distribution, prioritize critical equipment, investigate potential contributing factors, and structure improvement actions, justifying the integrated analytical workflow adopted in this study.

B. Repair Time Distribution

The distribution of individual Time to Repair (TTR) values was examined using a histogram (Fig. 1), providing an overview of repair-time variability in relation to the calculated MTTR and complementary descriptive statistics. The histogram revealed a pronounced right-skewed distribution, consistent with the numerical skewness of 3.98.

The median TTR (1.0 day) was substantially lower than the mean (1.92 days), while the 90th and 95th percentiles reached 5 and 7 days, respectively. Thus, a relatively small number of prolonged repair intervals exerted considerable influence on the overall MTTR. This finding supports reporting the mean together with median- and percentile-based measures rather than interpreting MTTR in isolation [8], [12]. Beyond describing repair-time variability, the histogram served as an exploratory analytical tool that facilitated the identification of atypical repair durations prior to equipment prioritization. This preliminary assessment provided the basis for the subsequent Pareto analysis.

C. Prioritization of Critical Equipment

The Pareto analysis showed that corrective maintenance time was concentrated in a subset of equipment records (Fig. 2),

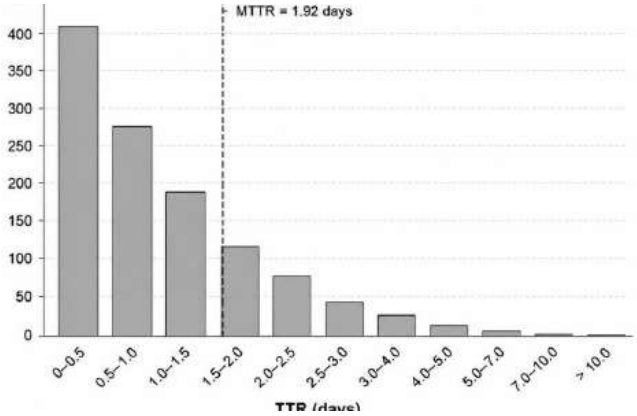


Fig. 1. Distribution of repair times for corrective maintenance work orders.

supporting descriptive prioritization of prolonged repair cases. Given the marked asymmetry of the TTR distribution, the overall MTTR was not interpreted as a validated cutoff for defining equipment criticality. Instead, the Pareto chart was used to visualize the relative contribution of repair intervals to cumulative corrective maintenance time. After excluding SUPPORT work orders from equipment prioritization, records with TTR above the overall MTTR were screened and ranked in descending order.

The five highest-ranked records were selected for qualitative investigation (Table I): three multiparametric monitor cases and two ophthalmic tonometer cases, with TTR values of 27, 17, 17, 8, and 8 days. Prioritization was based on this exploratory screening set; the mean was not interpreted as a validated criticality threshold. Rather than establishing a clinical or engineering criticality threshold, the TTR $>$ MTTR step was used only as an exploratory screening rule. Within that screened set, descending TTR ranking and the Pareto chart provided a descriptive means of concentrating attention on prolonged repair intervals.

This distinction is important because the mean of a markedly right-skewed distribution is not, by itself, a validated cutoff for prioritization. The Pareto stage therefore supported exploratory prioritization for subsequent investigation of potential contributing factors [4], [8]. The Pareto analysis complemented the MTTR and distributional statistics. While MTTR quantified average maintenance performance and the median and percentiles characterized the asymmetric distribution, the Pareto chart visualized the concentration of cumulative repair time, establishing the basis for the subsequent qualitative Ishikawa analysis.

D. Investigation of Potential Contributing Factors

To investigate factors potentially associated with prolonged repair times, Ishikawa cause-and-effect diagrams were developed for the five cases taken forward from the prioritization stage (Figs. 3–5). This qualitative approach organized potential contributing factors into six analytical categories: equipment characteristics, spare parts, maintenance methods, workforce, management practices, and supplier-related factors.

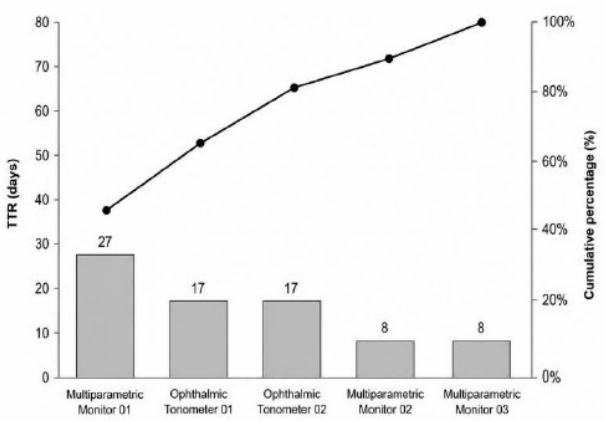


Fig. 2. Pareto chart of the equipment contributing most to cumulative corrective maintenance time.

TABLE I
EQUIPMENT PRIORITIZED FOR INVESTIGATION OF POTENTIAL CONTRIBUTING FACTORS

Equipment	TTR (days)	Main identified factor
Multiparametric Monitor 01	27	Equipment lifecycle limitations and replacement component unavailability
Ophthalmic Tonometer 01	17	Discontinued equipment and spare-part supply restrictions
Ophthalmic Tonometer 02	17	Discontinued equipment and spare-part supply restrictions
Multiparametric Monitor 02	8	Electronic failures associated with administrative constraints
Multiparametric Monitor 03	8	Equipment lifecycle limitations and insufficient spare-part inventory

Across the diagrams, the maintenance records suggested recurrent potential contributing factors, including equipment obsolescence, limited availability of replacement parts, discontinued manufacturer support, incomplete diagnostic procedures, deficiencies in maintenance documentation, communication failures between technical and administrative sectors, and institutional management constraints. These factors should be interpreted as hypotheses generated from the available records rather than as independently verified causes. Among the potential factors represented in the diagrams, equipment obsolescence and discontinued manufacturer support were considered plausible contributors to prolonged downtime.

The mapped factors were reviewed by three technicians and two clinical engineers from the Clinical Engineering team, providing an expert-based assessment of their technical plausibility and operational relevance. Nevertheless, no formal consensus method, independent documentary confirmation of every mapped factor, or quantitative causal assessment was performed. The diagrams should therefore be interpreted as expert-reviewed exploratory tools for hypothesis generation and improvement planning [1], [4].

The Ishikawa diagrams also highlighted potentially modifiable organizational factors, including diagnostic protocols, maintenance procedure standardization, spare-parts inventory, communication between technical and administrative departments, workforce training, and documentation practices [5], [8]. These factors require prospective verification before their effect on MTTR can be established. Taken together, the Ishikawa diagrams suggest that maintenance performance may be influenced by interactions among technical, organizational, and managerial processes. This systems perspective is compatible with the operational definition of MTTR adopted in this study, which covers the complete SOS–OSP repair interval.

The potential contributing factors mapped through the Ishikawa analysis subsequently served as the foundation for structuring proposed improvement actions using the PDSA cycle.

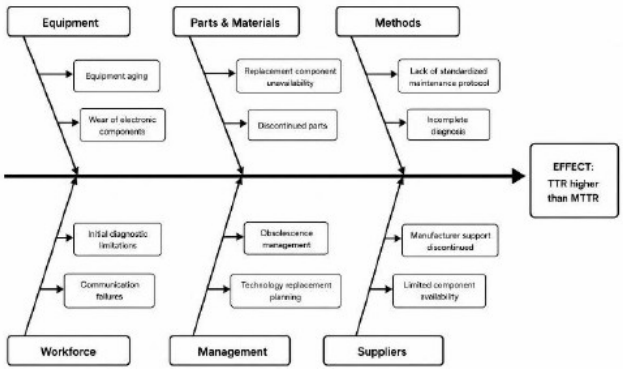


Fig. 3. Ishikawa diagram for Multiparametric Monitor 01.

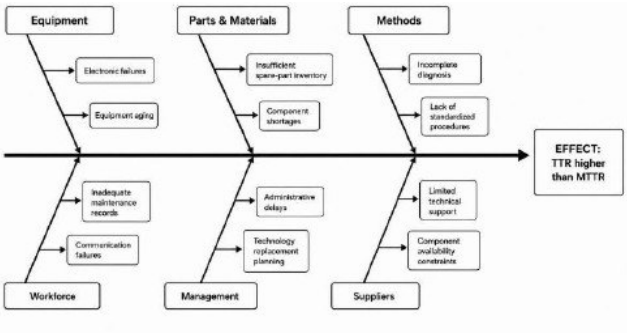


Fig. 4. Ishikawa diagram for Multiparametric Monitors 02 and 03.

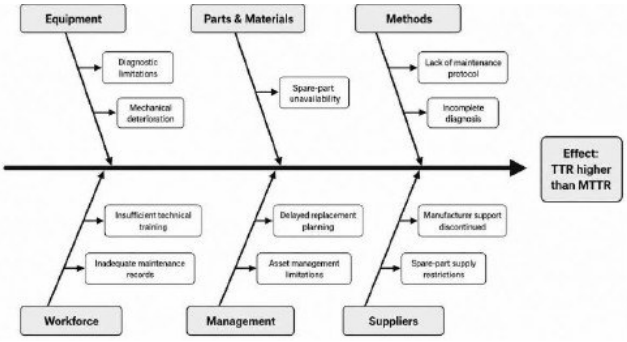


Fig. 5. Ishikawa diagram for Ophthalmic Tonometers 01 and 02.

E. Integration of Performance Indicators and Quality Management Tools

The integrated analytical workflow showed that maintenance performance indicators and quality management tools provide complementary descriptive information when applied sequentially. MTTR quantified average performance; the median, IQR, percentiles, confidence interval, and histogram characterized the asymmetric repair-time distribution; the Pareto chart supported descriptive prioritization; and the Ishikawa diagrams organized potential contributing factors. Finally, the PDSA cycle translated these findings into proposed improvement actions. The workflow therefore supports structured decision-making but does not, in the present cross-sectional analysis, demonstrate improvement efficacy.

The proposed workflow illustrates how maintenance records generated during routine hospital activities can support operational maintenance control and decision-making in Clinical Engineering. Its principal contribution is the sequential application of complementary tools to real CMMS records, allowing descriptive performance measurement, prioritization, exploratory cause mapping, and improvement planning within a single analytical process [4], [5], [8]. Several limitations should be considered when interpreting these findings.

The analysis was conducted at a single university hospital and covered only one month, limiting generalization to institutions with different equipment inventories, maintenance structures, and operational characteristics. The study focused exclusively on corrective maintenance and therefore did not permit comparison with preventive or predictive maintenance. The CMMS field for equipment criticality was incomplete in the analytical dataset, preventing a valid stratified analysis by criticality, and no dedicated variable quantified waiting time for spare parts. Consequently, spare-part waiting time could not be analyzed independently and was outside the scope of this study. Equipment-type stratification was also limited by small numbers within several categories.

The analytical dataset comprised completed corrective work orders; consequently, no right-censored observations were present among the 239 analyzed records. Nevertheless, restricting the analysis to completed work orders may introduce selection effects because repairs that remained open beyond the observation window were not represented in the final analytical dataset. Furthermore, the PDSA cycle was used to structure proposed actions rather than to evaluate longitudinal implementation. Although the Ishikawa factors were reviewed by three technicians and two clinical engineers, this validation remained qualitative and did not establish causality.

No direct comparison was performed with predictive-maintenance, machine-learning, or multicriteria decision-support approaches, as such a comparison was outside the scope of this study. Future studies should include multicenter and longitudinal datasets, explicit tracking of spare-part waiting time, more complete equipment-criticality fields, preventive and predictive maintenance indicators, reliability measures, utilization, lifecycle costs, and prospective assessment of PDSA interventions.

IV. Conclusion

This study showed that integrating MTTR with complementary descriptive statistics and quality management tools provides a structured approach for analyzing corrective maintenance in Clinical Engineering.

Using 239 corrective maintenance work orders from routine CMMS records, the mean SOS–OSP repair interval was 1.92 days, while the median was 1.0 day (IQR: 0–2 days), confirming a markedly right-skewed distribution. Histogram and Pareto analyses characterized variability and supported descriptive prioritization, Ishikawa diagrams organized potential contributing factors, and the PDSA cycle was used to structure proposed improvement actions. Within the scope of this single-center, one-month study, the contribution of the work lies in the sequential application of these tools to real maintenance records as a decision-support workflow.

The findings demonstrate descriptive prioritization and structured improvement planning, not proven improvement in maintenance outcomes. Future multicenter and longitudinal studies are required to evaluate generalizability and to quantify the effect of proposed interventions on MTTR and other maintenance outcomes.

References

[1] A. Badnjević, "Evidence-based maintenance of medical devices: Current shortage and pathway towards solution," *Technology and Health Care*, vol. 31, no. 1, pp. 293–305, 2023, doi: 10.3233/THC-229005.

[2] M. D. Bonfim and A. M. Malik, "The Role of Clinical Engineering in Management and Decision-Making in Brazilian Hospitals," *Global Clinical Engineering Journal*, vol. 8, no. 1, pp. 114–137, 2026, doi: 10.31354/globalce.v8i1.223.

[3] International Federation for Medical and Biological Engineering–Clinical Engineering Division (IFMBE-CED), *Clinical Engineering Global Performance Indicators*. IFMBE-CED, 2024.

[4] E. Iadanza, V. Gonnelli, F. Satta, and M. Gherardelli, "Evidence-Based Medical Equipment Management: A Convenient Implementation," *Medical & Biological Engineering & Computing*, vol. 57, no. 10, pp. 2215–2230, 2019, doi: 10.1007/s11517-019-02021-x.

[5] A. Luschi, G. L. Daino, G. Ghisalberti, V. Mezzatesta, and E. Iadanza, "Empowering Clinical Engineering and Evidence-Based Maintenance with IoT and Indoor Navigation," *Future Internet*, vol. 16, no. 8, Art. no. 263, 2024, doi: 10.3390/fi16080263.

[6] R. M. dos Santos and M. E. L. V. Dallora, "Evaluation of Clinical Engineering Performance Indicators: A Proposal for a University Public Hospital," *Medicina (Ribeirão Preto)*, vol. 52, no. 1, pp. 34–46, 2019, doi: 10.11606/issn.2176-7262.v52i1p34-46.

[7] Associação Brasileira de Normas Técnicas, *ABNT NBR 5462: Confiabilidade e Mantenabilidade*. Rio de Janeiro, Brazil: ABNT, 1994.

[8] Z. Lin, J. Kang, Y. Wei, and B. Zou, "Maintenance Management Strategies for Medical Equipment in Healthcare Institutions: A Review," *BME Horizon*, vol. 2, Art. no. 2024135, 2024, doi: 10.70401/bmeh.2024.135.

[9] K. Ishikawa, *Guide to Quality Control*. Tokyo, Japan: Asian Productivity Organization, 1976.

[10] K. Ishikawa, *What Is Total Quality Control? The Japanese Way*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1985.

[11] W. E. Deming, *Out of the Crisis*. Cambridge, MA, USA: MIT Press, 1986.

[12] D. C. Montgomery, *Introduction to Statistical Quality Control*, 8th ed. Hoboken, NJ, USA: John Wiley & Sons, 2020.

Inventário estruturado desenvolvido em um hospital do agreste alagoano visando à implantação da engenharia clínica institucional

Marília Gabrielle de Jesus Silva
Departamento de Engenharias
IFAL – Campus Arapiraca
Arapiraca, Brasil
0009-0006-4633-9968

Maria Stefhany Francelino Lima
Departamento de Engenharias
IFAL – Campus Arapiraca
Arapiraca, Brasil
0009-0001-2855-8503

Paulyanne Karlla Araújo Magalhães
Departamento de Engenharias
IFAL – Campus Arapiraca
Arapiraca, Brasil
0000-0002-6731-6515

Jean Vitor Vilela dos Santos
Departamento de Engenharias
IFAL – Campus Arapiraca
Arapiraca, Brasil
0009-0007-0073-4441

Paulo Cesar do Nascimento Cunha
Departamento de Engenharias
IFAL – Campus Arapiraca
Arapiraca, Brasil
0000-0002-3935-3474

Elenilda Josefa de Oliveira
Departamento de Engenharias
IFAL – Campus Arapiraca
Arapiraca, Brasil
0009-0009-0384-3445

***Resumo*—A crescente incorporação de tecnologias nos serviços de saúde tem ampliado a necessidade de métodos estruturados para o gerenciamento dos equipamentos médico-hospitalares, especialmente em instituições que ainda não possuem setores formalizados de Engenharia Clínica. A literatura aponta que a ausência de inventários atualizados e de sistemas de gestão tecnológica compromete a rastreabilidade dos ativos, dificulta o planejamento da manutenção e impacta diretamente a disponibilidade operacional. Nesse contexto, o presente trabalho apresenta o desenvolvimento de um inventário estruturado do parque tecnológico de um hospital público municipal da região Agreste de Alagoas, como etapa inicial para a implantação da Engenharia Clínica institucional. A pesquisa, caracterizada como estudo de caso, envolveu o levantamento físico dos equipamentos, a transição dos dados para um sistema informatizado de gerenciamento de manutenção (CMMS) e a estruturação de Ordens de Serviço (OS) com protocolos de segurança elétrica. Os resultados evidenciaram que o hospital operava sob um modelo de manutenção predominantemente corretivo e terceirizado, sem cronogramas preventivos. A informatização do inventário e a gestão via OS permitiram organizar o acervo, criando condições para a rastreabilidade das intervenções e para a mitigação de riscos, estabelecendo as bases para a extração futura de indicadores de desempenho, como o Tempo Médio Entre Falhas (Mean Time Between Failures - MTBF), o Tempo Médio de Reparo (Mean Time to Repair - MTTR) e a disponibilidade operacional. Conclui-se que o inventário estruturado e informatizado representa o pilar fundamental para práticas modernas de Engenharia Clínica em hospitais públicos interioranos, promovendo a transição para a cultura preventiva, o aumento da segurança assistencial e a otimização dos recursos em saúde.**

***Palavras-chave*—Engenharia Clínica, Gestão de Tecnologias em Saúde, Manutenção de Equipamentos Hospitalares, Sistemas de Gestão de Manutenção (CMMS).**

I. Introdução

Os Equipamentos Médico-Hospitalares (EMH) constituem elementos essenciais para a assistência moderna à saúde, participando diretamente dos processos de diagnóstico, monitoramento e tratamento dos pacientes. O aumento da complexidade tecnológica dos hospitais exige métodos estruturados para controle, rastreabilidade e manutenção desses ativos, garantindo maior segurança ao paciente e eficiência operacional dos serviços de saúde [1]–[3].

A Engenharia Clínica surge nesse contexto como área estratégica responsável pelo gerenciamento do ciclo de vida das tecnologias em saúde, abrangendo atividades relacionadas à aquisição, instalação, inventário, manutenção, avaliação de desempenho e descarte. Estudos demonstram que hospitais com processos estruturados de Engenharia Clínica apresentam maior disponibilidade tecnológica e redução dos custos associados à manutenção corretiva [4]–[6].

Pesquisas recentes evidenciam que a implantação de práticas formais de gestão tecnológica promove melhorias significativas na redução de falhas operacionais e no planejamento das intervenções técnicas [7]. Constata-se, contudo, que a ausência de sistemas estruturados de gerenciamento tecnológico e de informações organizadas sobre o acervo constitui uma das principais barreiras para o desenvolvimento da área em hospitais públicos brasileiros, impactando diretamente a eficiência assistencial [8], [9].

No cenário nacional, Cardoso e Waqued [7] observaram que a Engenharia Clínica deixou de atuar apenas como setor de oficina para assumir um papel estratégico, incorporando o gerenciamento de riscos e o controle do ciclo de vida dos equipamentos. Essa evolução reforça a urgência da implantação de sistemas organizados de inventário. Além disso, a experiência observada durante a pandemia da COVID-19 demonstrou a criticidade de informações confiáveis. Segundo Toscas *et al.* [8], dificuldades relacionadas ao controle patrimonial comprometeram a rápida mobilização de ventiladores

pulmonares e outros suportes de vida em diversas regiões do país.

A literatura aponta que a etapa fundamental para a consolidação da Engenharia Clínica é a elaboração de um inventário estruturado. A ausência de dados unificados sobre localização, estado operacional, fabricante e histórico de manutenção inviabiliza a tomada de decisão técnica [3], [10], [11]. Em hospitais localizados em regiões interioranas, os desafios tornam-se ainda maiores devido à limitação financeira, escassez de profissionais especializados e alta dependência de assistências técnicas sediadas em grandes centros urbanos. Essa realidade gera um aumento do tempo de indisponibilidade dos equipamentos, forçando a predominância de ações corretivas em detrimento das preventivas [6], [12].

A integração entre logística hospitalar e manutenção também se mostra determinante. A localização física dos equipamentos, o controle das movimentações internas e o acompanhamento dos fluxos de Ordem de Serviço (OS) influenciam diretamente a prontidão tecnológica das instituições [13]. Para sanar essas lacunas, a utilização de sistemas informatizados de gerenciamento de manutenção (CMMS) tem sido destacada na literatura. A consolidação de bancos de dados estruturados permite identificar padrões de falhas, embasar decisões e gerar indicadores de desempenho confiáveis para os gestores [6].

Diante desse cenário, o presente trabalho tem como objetivo apresentar o desenvolvimento de um inventário estruturado dos equipamentos médico-hospitalares de um hospital público municipal localizado na região Agreste do estado de Alagoas, atuando como etapa inicial para a implantação da Engenharia Clínica institucional. O estudo busca evidenciar como esse mapeamento fornece suporte para a transição de um modelo de controle informal para um sistema informatizado, podendo subsidiar o planejamento da manutenção, estabelecer uma base para a rastreabilidade dos ativos e criar condições para práticas modernas de gerenciamento de tecnologias em saúde.

II. Referencial Teórico

A Engenharia Clínica é definida como a aplicação dos conhecimentos de engenharia, gestão e ciências da saúde para garantir a utilização segura, eficaz e economicamente viável das tecnologias empregadas nos serviços assistenciais. Seu principal objetivo consiste em assegurar que os equipamentos médico-hospitalares mantenham desempenho adequado ao longo de todo o seu ciclo de vida [1], [2].

Segundo Wang [4], a gestão tecnológica hospitalar deve contemplar atividades integradas de planejamento, aquisição, inventário, manutenção, monitoramento e substituição de equipamentos. Essas ações permitem reduzir falhas operacionais e ampliar a confiabilidade dos sistemas utilizados na assistência aos pacientes [4], [5]. No Brasil, a regulamentação relacionada à gestão de tecnologias em saúde é fortalecida pela Resolução da Diretoria Colegiada (RDC) nº 509/2021 da Agência Nacional de Vigilância Sanitária (ANVISA), que estabelece diretrizes rigorosas para o gerenciamento de equipamentos e infraestrutura tecnológica em estabelecimentos assistenciais de saúde [14].

Complementando essa visão regulatória, pesquisadores do Hospital das Clínicas da Universidade de São Paulo (USP) [15] demonstraram que a formalização dos fluxos de trabalho da Engenharia Clínica proporciona maior eficiência na comunicação entre os setores assistenciais, a equipe de manutenção e a administração hospitalar. O mapeamento desses processos operacionais permite identificar gargalos e otimizar significativamente o gerenciamento das tecnologias em saúde.

A Organização Mundial da Saúde (OMS) considera o inventário de equipamentos como a primeira etapa, inegociável, para a implementação de um programa de gerenciamento tecnológico. O inventário fornece a base de dados primária sobre quantidade, localização, características técnicas, fabricante, estado de conservação e criticidade do acervo institucional [10], [11]. De acordo com Calil e Teixeira [3], um inventário adequadamente estruturado permite que os gestores conheçam a real situação do parque tecnológico hospitalar, subsidiando decisões estratégicas relacionadas à manutenção, aquisições e desativações.

A literatura reforça que a inexistência de inventários atualizados inviabiliza a rastreabilidade tecnológica e suprime a capacidade de planejamento das unidades de saúde. Em contrapartida, a implantação de inventários devidamente informatizados destaca-se como um dos principais fatores responsáveis por um avanço significativo na qualidade da gestão de equipamentos [7]. Além do mero controle patrimonial, essa ferramenta fornece suporte para a avaliação de riscos tecnológicos e no planejamento de investimentos futuros, conferindo maior eficiência operacional [6], [16].

A implantação da Engenharia Clínica em hospitais públicos ocorre, habitualmente, de forma gradual, tendo o levantamento físico e patrimonial como o ponto de partida para a construção de um ecossistema de gestão eficiente [7], [15]. Estudos de caso realizados em hospitais brasileiros evidenciam que a ausência dessa base de dados dificulta o planejamento da manutenção preventiva, infla os custos operacionais e prolonga o tempo de indisponibilidade de máquinas vitais. Por outro lado, a implantação de rotinas sistematizadas contribui para a redução das ocorrências de falhas severas, impacto que se torna ainda mais relevante em hospitais interioranos, onde os recursos técnicos e financeiros são frequentemente escassos e a dependência de terceiros é alta [2], [5], [12].

A partir da consolidação do inventário e do histórico de intervenções técnicas (Ordens de Serviço), o referencial contemporâneo aponta para a adoção de sistemas informatizados de gerenciamento de manutenção (CMMS). Pesquisas voltadas para estratégias baseadas em dados [6] demonstraram que a utilização de registros estruturados permite extrair indicadores de desempenho robustos e confiáveis.

Segundo os princípios da Engenharia Clínica contemporânea, a extração de métricas como o MTBF, o MTTR, a disponibilidade operacional e a taxa de manutenção preventiva torna possível identificar tendências de depreciação, estabelecer prioridades de manutenção e otimizar, de forma fundamentada, a alocação de recursos financeiros e humanos da instituição [1], [4], [16].

III. Materiais e Métodos

A pesquisa caracteriza-se como um estudo de caso, de natureza aplicada, com abordagem mista (qualitativa e quantitativa) e objetivos descritivos. O estudo foi desenvolvido em um hospital público municipal localizado na região Agreste do estado de Alagoas, que presta atendimento de média complexidade a diversas cidades circunvizinhas. A instituição possui um parque tecnológico diversificado, englobando setores de urgência, internação, centro cirúrgico, laboratório e diagnóstico. Para preservar a confidencialidade institucional, a identificação do hospital foi suprimida, visto que a pesquisa aborda aspectos sensíveis relacionados à gestão tecnológica e aos processos de manutenção.

Inicialmente, realizou-se uma revisão da literatura nacional e internacional sobre Engenharia Clínica, gerenciamento de tecnologias em saúde, inventário hospitalar, gestão do ciclo de vida dos equipamentos e indicadores de desempenho. Essa etapa forneceu a fundamentação teórica necessária para o delineamento dos procedimentos experimentais. O plano de trabalho contemplou o mapeamento dos setores a serem visitados, a definição das variáveis a serem coletadas e o cronograma de atividades.

A coleta de dados ocorreu por meio de visitas técnicas quinzenais, nas quais procedeu-se ao levantamento físico dos equipamentos. Registrou-se, para cada ativo, informações como localização, patrimônio, fabricante, modelo, número de série, estado de conservação, condição operacional e características técnicas relevantes. Paralelamente, foram realizadas entrevistas técnicas com os gestores da manutenção para compreender o fluxo atual de atendimento. O diagnóstico inicial abrangeu a análise de contratos com empresas terceirizadas, a verificação de cronogramas de manutenção preventiva e a análise da resposta à manutenção corretiva.

Após a coleta, os dados foram processados e inseridos em um sistema informatizado de gerenciamento de manutenção (CMMS), utilizando a plataforma Arkmeds em ambiente educacional. O banco de dados foi estruturado para organizar o parque tecnológico por setor, registrar o histórico de intervenções e permitir a geração de Ordens de Serviço (OS).

Por fim, os dados foram analisados por meio de estatística descritiva e análise qualitativa, buscando identificar o perfil do parque tecnológico e as principais fragilidades na gestão. Os resultados foram confrontados com a literatura especializada para subsidiar a proposta de implantação gradual da Engenharia Clínica, incluindo a padronização das ordens de serviço, a criação de indicadores de desempenho, como MTBF e MTTR, e o estabelecimento de um plano anual de manutenção preventiva.

IV. Resultados e Discussão

A inspeção visual permitiu registrar as características físicas e o estado de conservação do parque tecnológico. Conforme ilustrado na Fig. 1, o painel evidencia a diversidade de ativos críticos presentes na instituição, englobando desde equipamentos de suporte à vida e terapia, como incubadoras de transporte, ventiladores pulmonares e bombas de infusão, até sistemas radiológicos de diagnóstico por imagem, como mamógrafos. Ademais, o registro fotográfico de etiquetas de teste e calibração em equipamentos específicos, observado na Fig. 1(h), reforça a necessidade prática de rastreabilidade técnica.

Com base nessa inspeção física, o inventário estruturado resultou na catalogação de 40 equipamentos médico-hospitalares, destinados ao atendimento básico e de média complexidade. A totalidade do parque tecnológico inventariado encontrava-se com status operacional ativo (operante) no momento da coleta. A distribuição desses ativos revela uma concentração em setores críticos, como Centro Cirúrgico e Estabilização, conforme evidenciado na Tabela I.

Tabela I
Distribuição dos Equipamentos Inventariados por Setor Hospitalar

Setor / Unidade	Quantidade	Status Operacional
Centro Cirúrgico	12	Operante (100%)
Estabilização	12	Operante (100%)
Unidade Mista (UMNSG)	6	Operante (100%)
Maternidade	5	Operante (100%)
Clínica Médica	2	Operante (100%)
Esterilização (CME)	2	Operante (100%)
Radiologia	1	Operante (100%)
Total	**40**	**Operante (100%)**

Além da localização espacial, o mapeamento evidenciou os principais tipos de tecnologias disponíveis na instituição. A compilação quantitativa por categorias de equipamentos é apresentada na Fig. 2.

Durante o levantamento, observou-se que a instituição não possuía histórico consolidado de manutenção nem sistema informatizado para gestão dos equipamentos, evidenciando limitações no controle das atividades de manutenção e na rastreabilidade das intervenções técnicas. Durante as entrevistas realizadas com o responsável pela manutenção dos equipamentos, foi informado que o hospital possui um acordo firmado com uma empresa especializada localizada na capital do estado, Maceió, responsável pela execução dos serviços de manutenção. Entretanto, a instituição não dispõe de informações detalhadas sobre o conteúdo desse contrato.

Quando questionado sobre a existência de um calendário de manutenção preventiva dos equipamentos médico-hospitalares, o responsável informou desconhecer sua existência, afirmando não possuir acesso a qualquer cronograma ou planejamento formal das intervenções preventivas. Curiosamente, o levantamento físico e fotográfico revelou a existência de etiquetas de teste e calibração em ativos específicos, como observado na bomba de seringa (Fig. 1h), o que evidencia um grave descompasso entre as rotinas de campo e o controle administrativo da instituição. Essa situação evidencia fragilidades na gestão tecnológica do hospital, dificultando o acompanhamento das ações de manutenção e o controle da disponibilidade dos equipamentos.

Em relação ao fluxo de atendimento das manutenções corretivas, foi relatado que, quando ocorre uma falha em algum equipamento, o hospital entra em contato com a empresa



(a) Monitor multiparamétrico (Fabricante: COMEN; Modelo: STAR8000E; Localização: Estabilização; Patr.: 20211897; Nº Série: S/N; Condição: Operante).



(b) Incubadora neonatal (Fabricante: Olidef; Modelo: RWT PLUS; Localização: Maternidade; Patr.: 3697; Nº Série: S/N; Condição: Operante).



(c) Ventilador pulmonar (Fabricante: Magnamed; Modelo: OxyMag; Localização: Estabilização; Patr.: 20222181; Nº Série: S/N; Condição: Operante).



(d) Mamógrafo com console (Fabricante: Aurora; Localização: Maternidade; Patr.: S/N; Nº Série: S/N; Condição: Operante).



(e) Bomba de infusão volumétrica (Fabricante: Não ident.; Modelo: SAM-6003; Localização: Centro Cirúrgico; Patr.: 2025005436; Condição: Operante).



(f) Detalhe do gantry do mamógrafo Aurora catalogado na Maternidade (Ativo vinculado à Fig. 1d).



(g) Configuração clínica do ventilador Magnamed OxyMag interligado ao cilindro de oxigênio (Ativo vinculado à Fig. 1c).



(h) Bomba de seringa (Fabricante: CMOS Drake; Modelo: Yonah; Localização: Estabilização; Patr.: 20242504003; Nº Série: 0408006; Condição: Operante).

Figura 1. Painel ilustrativo do parque tecnológico catalogado, apresentando a compilação final dos dados de rastreabilidade extraídos durante o inventário físico institucional, em conformidade com as variáveis definidas no método. Na coluna esquerda: (a) Monitor multiparamétrico de sinais vitais; (b) Incubadora neonatal de transporte; (c) Ventilador pulmonar; e (d) Sistema de mamografia. Na coluna direita: (e) Bomba de infusão volumétrica; (f) Detalhe do sistema de compressão do mamógrafo; (g) Detalhe de instalação do ventilador pulmonar; e (h) Bomba de seringa evidenciando etiqueta de rastreabilidade de manutenção.

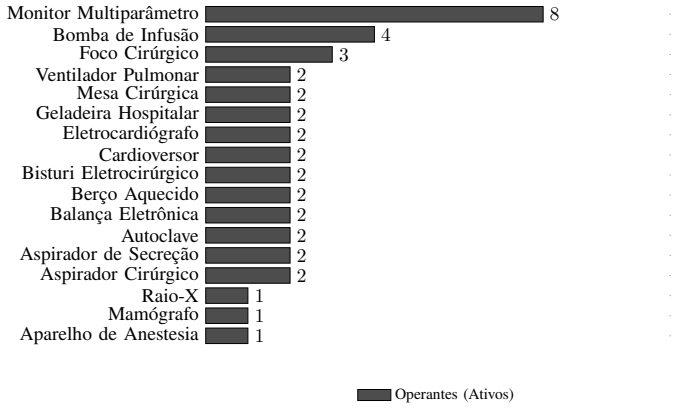


Figura 2. Distribuição quantitativa das categorias de equipamentos médico-hospitalares identificados no inventário. O levantamento constatou que 100% (40) dos ativos registrados encontram-se em status operacional operante (ativos), sem incidência de equipamentos inoperantes ou sucateados.

contratada, que estabelece um prazo de até dez dias para atendimento da solicitação. Esse intervalo pode comprometer significativamente a disponibilidade operacional de equipamentos essenciais de suporte à vida e diagnóstico por imagem, a exemplo dos ventiladores pulmonares e sistemas de mamografia identificados no inventário, sobretudo em hospitais de pequeno porte e localizados no interior, onde a dependência de assistência técnica externa tende a aumentar o tempo de resposta às ocorrências.

Esses achados corroboram a literatura, que aponta que a ausência de um sistema estruturado de Engenharia Clínica, aliado à inexistência de registros organizados de manutenção preventiva, contratos tecnicamente gerenciados e indicadores de desempenho, favorece a predominância de ações corretivas, eleva os custos operacionais e aumenta o tempo de indisponibilidade dos equipamentos médico-hospitalares [6], [7], [12]. A implantação de um sistema informatizado de gestão, associada à formalização dos processos de manutenção e ao acompanhamento sistemático dos contratos de assistência técnica, estabelece uma base para o aumento da confiabilidade tecnológica da instituição.

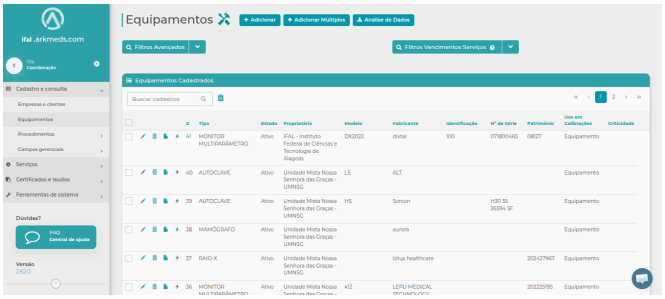
A implantação do inventário estabelece a base para a organização patrimonial dos ativos e cria condições para a rastreabilidade tecnológica, a identificação rápida da localização dos equipamentos, o registro padronizado das características técnicas, a criação de histórico para futuras análises de manutenção e a implementação de indicadores de desempenho, como o MTBF, o MTTR, a disponibilidade operacional e o índice de manutenção preventiva.

Como desdobramento prático do inventário físico, procedeu-se com a inserção dos dados em um software de gerenciamento de manutenção informatizado (CMMS), cuja interface é ilustrada na Fig. 3.

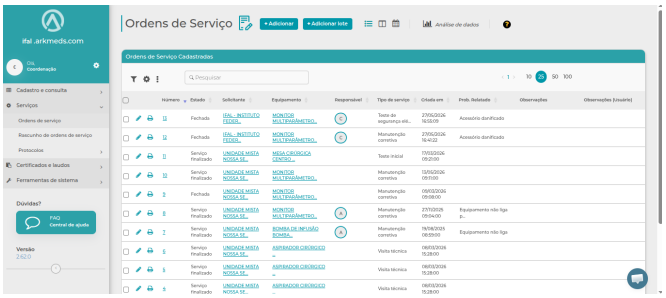
Para demonstrar a aplicabilidade do sistema e exemplificar a execução dos protocolos de qualidade, realizou-se um ensaio prático de segurança elétrica em um dos equipamentos críticos inventariados (Monitor Multiparamétrico Dixtal DX2022). O teste foi parametrizado com base nos limites normativos de aceitação estabelecidos pela norma ABNT NBR IEC 60601-1:2010 (incluindo Emenda 1:2016) para a segurança geral e pelos critérios de ensaios recorrentes de campo preconizados pela ABNT NBR IEC 62353:2019. Para as medições das correntes de fuga (do equipamento e do gabinete), o analisador de segurança elétrica aplicou o Método de Medição Direto. O monitor ensaiado possui classificação geral de proteção contra choque elétrico de Classe I (com aterramento de proteção), e suas partes aplicadas possuem classificações distintas a depender do parâmetro (Tipo BF para PNI e SpO2, e Tipo CF para o canal de ECG). Para a simulação física das correntes de fuga das partes aplicadas registradas na Ordem de Serviço nº 13, conectou-se ao analisador o cabo paciente de ECG de 5 vias, avaliando-se o limite para partes aplicadas Tipo BF/CF.

Os parâmetros coletados durante a intervenção foram inseridos diretamente na plataforma Arkmeds. A Tabela II apresenta os resultados qualitativos e quantitativos do ensaio de segurança. É importante destacar que o monitor foi adequadamente caracterizado como um equipamento de Classe I com partes aplicadas flutuantes, utilizando-se o cabo paciente de ECG como a parte aplicada ativa para o teste físico de corrente de fuga de paciente. Os valores medidos de corrente de fuga do equipamento, do gabinete e das partes aplicadas mantiveram-se significativamente abaixo dos limites de aceitação impostos pelas normas vigentes ABNT NBR IEC 60601-1 e ABNT NBR IEC 62353, o que garantiu a aprovação técnica integral do ativo.

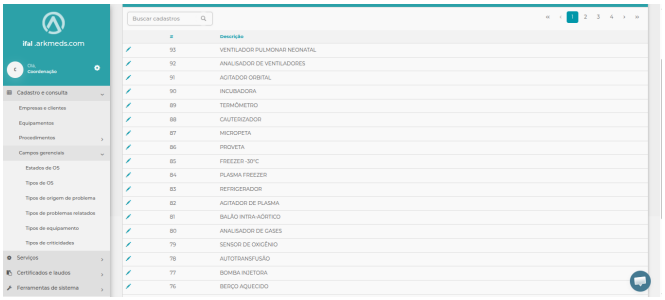
Com base nos resultados obtidos, propõe-se a implantação gradual da Engenharia Clínica por meio da consolidação contínua do inventário no CMMS. A adoção dessa ferramenta viabiliza a estruturação do banco histórico de falhas e a implantação formal do sistema de ordens de serviço. Por fim, o desenvolvimento de Procedimentos Operacionais Padrão (POPs) garantirá a organização e padronização de todos esses processos.



(a) Módulo de cadastro e consulta.



(b) Painel de Ordens de Serviço (OS).



(c) Parametrização de tipos de equipamentos.



(d) Trilha de auditoria e histórico de ações.

Figura 3. Interface do sistema de gerenciamento de manutenção informatizado (CMMS) implementado. (a) Consolidação do inventário estruturado com dados patrimoniais e técnicos; (b) Painel de Ordens de Serviço evidenciando fluxos de manutenção corretiva e testes de segurança; (c) Padronização das nomenclaturas e tipologias dos ativos críticos; e (d) Histórico de ações no sistema, garantindo a rastreabilidade e a segurança das informações gerenciais da Engenharia Clínica.

Tabela II
RESULTADOS DO TESTE DE SEGURANÇA ELÉTRICA (ORDEM DE SERVIÇO Nº 13) BASEADOS NAS NORMAS ABNT NBR IEC 60601-1 E 62353

Descrição do Parâmetro Avaliado	**Valor Medido**	**Critério de Aceitação**	**Avaliação**
Resistência de aterramento (Ω)	0.26	< 0.3	Conforme
Corrente de fuga do equipamento; Conexão Direta; Com Fase; Com Neutro; Classe I (μA)	2.41	< 500	Conforme
Corrente de fuga do equipamento; Conexão Direta; Sem Fase; Com Neutro; Classe I (μA)	34.92	< 1000	Conforme
Corrente de fuga do equipamento; Conexão Inversa; Com Fase; Com Neutro; Classe I (μA)	2.35	< 1000	Conforme
Corrente de fuga do gabinete; Conexão Direta; Com Fase; Com Neutro; Com Terra de proteção; Classe I (μA)	0.0	< 100	Conforme
Corrente de fuga do gabinete; Conexão Direta; Com Fase; Sem Neutro; Com Terra de proteção; Classe I (μA)	0.0	< 500	Conforme
Corrente de fuga do gabinete; Conexão Inversa; Com Fase; Com Neutro; Com Terra de proteção; Classe I (μA)	0.0	< 100	Conforme
Corrente de fuga do gabinete; Conexão Inversa; Com Fase; Com Neutro; Sem Terra de proteção; Classe I (μA)	1.58	< 500	Conforme
Corrente de fuga da parte aplicada (Cabo de ECG Tipo CF/BF) (μA)	1.61	< 100	Conforme
Corrente de fuga de todas as partes aplicadas conectadas; Conexão Inversa (μA)	1.64	< 100	Conforme
Resistência de isolamento entre partes sob tensão para gabinete não aterrado (MΩ)	100.0	> 2	Conforme
Tensão de alimentação (V)	213.1	N/A	Conforme
Corrente Funcional (A)	0.04	N/A	Conforme
Potência aparente (VA)	8.52	N/A	Conforme

V. Conclusão

O desenvolvimento do inventário estruturado permitiu estabelecer as bases técnicas necessárias para a implantação da Engenharia Clínica na instituição estudada. Os resultados evidenciaram que o inventário não se limita ao controle patrimonial, constituindo uma ferramenta estratégica para a gestão tecnológica hospitalar. Sua implementação, aliada ao levantamento físico e visual em campo, forneceu suporte para a organização das informações dos equipamentos, para a futura rastreabilidade dos ativos e para a criação de mecanismos visando o acompanhamento sistemático da manutenção.

Além disso, o estudo demonstrou que a transição de um modelo de manutenção predominantemente corretivo e dependente de assistência técnica externa para uma abordagem preventiva exige a formalização de processos e o engajamento contínuo da gestão hospitalar. A adoção de um sistema informatizado de gerenciamento revelou-se indispensável para a centralização dos dados e para a futura geração de indicadores de desempenho, os quais serão fundamentais para embasar decisões estratégicas e avaliar a eficácia dos contratos de prestação de serviços.

Como trabalhos futuros, recomenda-se a consolidação das ordens de serviço digitalizadas, a elaboração de um plano anual de capacitação contínua para a equipe assistencial e a implantação de um programa de gerenciamento de riscos tecnológicos. Conclui-se, portanto, que a estruturação inicial desse inventário representa o primeiro passo decisivo para a consolidação de práticas modernas de Engenharia Clínica em hospitais interioranos, com potencial para contribuir diretamente para a melhoria da qualidade assistencial, a otimização dos recursos públicos e a construção de uma base sólida para o aumento da confiabilidade tecnológica e da segurança do paciente.

Agradecimentos

Os autores agradecem ao Instituto Federal de Alagoas (IFAL) pelo suporte acadêmico e institucional, bem como à gestão do hospital parceiro pela abertura e colaboração fundamentais para a realização do levantamento de campo e para a execução dos testes de segurança elétrica que viabilizaram este estudo.

Referências

[1] J. D. Bronzino e D. R. Peterson, *The biomedical engineering handbook: Clinical engineering fundamentals*, 4. ed. Boca Raton: CRC Press, 2015.

[2] J. F. Dyro, *Clinical engineering handbook*, 2. ed. Burlington: Academic Press, 2014.

[3] S. J. Calil e M. S. Teixeira, *Gerenciamento de manutenção de equipamentos hospitalares*. São Paulo: Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, 1998.

[4] B. Wang, *Medical equipment management and maintenance*. San Rafael: Morgan & Claypool Publishers, 2012.

[5] W. A. Hyman, "Clinical engineering and medical technology management," *Journal of Clinical Engineering*, vol. 26, no. 3, pp. 218–223, 2001.

[6] R. M. Santos e M. E. L. V. Dallora, "Avaliação de indicadores de desempenho da área de engenharia clínica: uma proposta para um hospital público universitário," *Medicina (Ribeirão Preto)*, vol. 52, no. 1, pp. 34–46, 2019.

[7] J. R. Cardoso e C. A. Waqued, "Engenharia clínica como vetor do avanço na gestão de equipamentos médico-hospitalares: caso de um hospital regional no DF," *Revista de Engenharia e Tecnologia*, vol. 11, no. 4, pp. 86–98, 2019.

[8] F. S. Toscas, E. M. Dias, M. L. Dias, T. R. Santos e E. J. V. Oliveira, "Management of medical equipment in the Brazilian public health system (SUS), historical situation and the context of the pandemic COVID-19: a cut for lung ventilators," *Health and Technology*, vol. 13, pp. 515–521, 2023.

[9] M. D. Bonfim e A. M. Malik, "The role of clinical engineering in management and decision-making in Brazilian hospitals," *Global Clinical Engineering Journal*, vol. 5, no. 1, pp. 1–12, 2026.

[10] World Health Organization, *Medical equipment inventory management*. Geneva: WHO Press, 2011. Disponível em: https://www.who.int/publications/i/item/9789241501392. Acesso em: 06 set. 2026.

[11] World Health Organization, *Medical equipment maintenance programme overview*. Geneva: WHO Press, 2011. Disponível em: https://www.who.int/publications/i/item/9789241501538. Acesso em: 06 set. 2026.

[12] J. S. C. Filho, A. J. M. Caldas e M. L. C. Neto, "A importância da engenharia clínica nas instituições de saúde: experiência em um hospital público federal," *Rev. Pesq. Saúde*, vol. 16, no. 2, pp. 75–79, 2015.

[13] A. A. Souza, A. C. C. Pereira, A. G. Xavier, D. O. Xavier e E. S. Mendes, "Logística hospitalar: um estudo de caso diagnóstico das dificuldades na gestão logística do setor de engenharia clínica," em *Book of Proceedings - Tourism and Management Studies International Conference Algarve*, Algarve, Portugal, 2012, vol. 4, pp. 1074–1084. Disponível em: http://periodicos.unifacef.com.br/rea/article/view/474. Acesso em: 05 set. 2026.

[14] Brasil, Agência Nacional de Vigilância Sanitária, "Resolução RDC nº 509, de 27 de maio de 2021. Dispõe sobre o gerenciamento de tecnologias em saúde em estabelecimentos de saúde," *Diário Oficial da União*, Brasília, DF, 31 maio 2021. Disponível em: https://www.in.gov.br/en/web/dou/-/resolucao-rdc-n-509-de-27-de-maio-de-2021-323002855. Acesso em: 08 set. 2026.

[15] A. F. Gomes *et al.*, "Mapeamento do fluxo de trabalho das atividades em engenharia clínica: a experiência do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo," *Medicina (Ribeirão Preto)*, vol. 48, no. 1, pp. 41–47, 2015.

[16] ECRI Institute, *Health technology management system: planning for medical equipment inventory and maintenance*. Plymouth Meeting: ECRI Institute, 2020. Disponível em: https://www.who.int/publications/b/74278. Acesso em: 10 set. 2026.

Level of Knowledge and Application of ABNT NBR IEC 62353:2019 Among Clinical Engineering Professionals in Brazil

1st Andre Felipe Brescovici Nunes
Postgraduate Program in Biomedical Engineering
Federal Technological University of Paraná - UTFPR
Curitiba, PR, Brazil
0009-0008-2488-6199

2nd Henrique Cunha Carvalho
Postgraduate Program in Biomedical Engineering
Federal Technological University of Paraná - UTFPR
Campo Mourão, PR, Brazil
0000-0003-0806-4132

3rd Frieda Saicla Barros
Postgraduate Program in Biomedical Engineering
Federal Technological University of Paraná - UTFPR
Curitiba, PR, Brazil
0000-0002-3962-1192

Abstract* — This study evaluated the level of knowledge and application of the ABNT NBR IEC 62353:2019 standard among Clinical Engineering professionals in Brazil. The research was conducted using a structured questionnaire administered to 45 professionals from 18 Brazilian states. Results indicate that only 22.2% of participants reported in-depth knowledge of the standard. Regarding test execution, the rates for performing the complete protocol prescribed by the standard (visual inspection, functional testing, and electrical safety tests) were low: 22.2% reported conducting all tests at equipment reception, 31.1% after repair, and 48.9% on a recurrent basis. It was also observed that most participants who perform tests use ABNT NBR IEC 60601-1:2022 as their reference, which is not the most appropriate standard for post-market testing, since ABNT NBR IEC 62353:2019 is specifically designed for this purpose. Larger institutions and those with quality certifications demonstrated higher levels of knowledge and greater adherence to the standard guidelines.

Keywords* — ABNT NBR IEC 62353:2019, electromedical equipment, electrical safety, clinical engineering.

I. Introduction

The interaction of electrical current with the human body can result in irreversible health damage and even death. Medical electrical equipment (ME Equipment) represents a potential source of electrical energy that may interact with the body during normal use; therefore, these risks must be managed throughout the devices' entire service life [1], [2].

The ABNT NBR IEC 62353:2019 standard establishes guidelines for conducting pre-service, recurrent, and post-repair testing on ME Equipment, aiming to ensure electrical safety and functional performance throughout their lifecycle [1], [3]. Despite its relevance, the adoption of this standard within Brazilian healthcare institutions remains poorly characterized.

Clinical Engineering (CE) departments are responsible for managing the performance of medical equipment inventories in hospitals. Activities such as acceptance testing, preventive and corrective maintenance, and periodic safety inspections comprise the core functions of these departments [4]. The inclusion of a specific standard for these tests within hospital routines represents a significant advancement for the sector. Previously, only the ABNT NBR IEC 60601-1:2022 series standards were available [2]; however, these standards focus on equipment manufacturing and require testing in environmentally controlled laboratories [5].

The first edition of IEC 62353 was published in 2007. However, national adoption in Brazil only occurred 12 years later, with the publication of ABNT NBR IEC 62353:2019 [1].

A study conducted at a large public hospital in Curitiba demonstrated the practical importance of these tests: out of 393 tests performed, failures were identified in 9.3% of cases, all of which had the potential to compromise patient and operator safety [6].

The present work aims to characterize, at a national level, the degree of knowledge and application of ABNT NBR IEC 62353:2019 among CE professionals, as well as to identify the main barriers to its implementation in Brazilian healthcare institutions.

II. Methodology

A. Study Design and Participants

This cross-sectional study involved CE professionals working in Brazilian healthcare institutions. The questionnaire was distributed electronically, and responses were collected anonymously and voluntarily. Participants working at more than one institution were instructed to respond based on their primary workplace. The study was approved by the Research Ethics Committee (CEP) of UTFPR (CAAE Nº. 84255024.2.0000.0177, Report Nº. 7.226.777).

Of the 47 responses received, two were excluded: one from a participant outside Brazil (Angola) and one without state-of-practice identification. The final sample comprised 45 valid responses from professionals in 18 of Brazil's 26 states.

B. Questionnaire Structure and Analysis

The structured questionnaire addressed five main areas, as illustrated in Fig. 1. Institutional characterization collected data on the total inventory of ME Equipment, the funding model (public, private, or philanthropic), quality certifications (e.g., ONA, JCI, or equivalent), and the state in which the participant operates. The self-reported knowledge level was assessed by asking participants to rate their familiarity with the ABNT NBR IEC 62353:2019 standard using a proficiency scale. Reference standards mapped the specific technical guidelines that professionals adopt in their daily work routines. Types of testing performed included detailed safety practices across three critical stages of the equipment lifecycle: acceptance (incoming inspection), post-repair, and recurrent testing. Finally, the identified implementation challenges highlighted the primary technical, financial, and cultural barriers to adopting the standard. Responses were analyzed descriptively, and subgroup analyses were conducted to identify correlations among institutional

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

characteristics, knowledge of the standard, and actual testing practices.

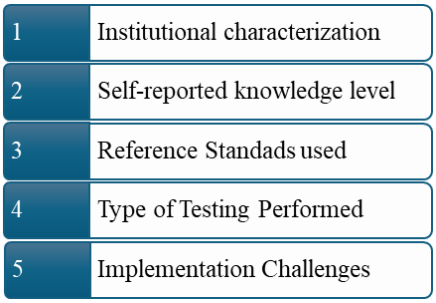


Fig. 1. Main topics addressed in the structured questionnaire.

III. Results and Discussion

A. Participant Characterization

Regarding the number of ME Equipment (institutional size), 66.7% of respondents (*n*=30) worked in hospitals with more than 1,500 ME Equipment, 15.6% (*n*=7) in medium-sized institutions with 501 to 1,500 devices, and 17.8% (*n*=8) in establishments with fewer than 500 ME Equipment. As for the funding model, 53.3% (*n*=24) worked in public institutions, 24.4% (*n*=11) in private ones, and 22.2% (*n*=10) in mixed institutions. Regarding quality certification (ONA, JCI, or equivalent), 35.6% (*n*=16) worked in certified institutions, while 64.4% (*n*=29) did not. Responses were received from 18 states, with the highest concentrations in the Southeast (19 participants), the Northeast (9), and the South (8). Table I summarizes the characterization of the study participants.

TABLE I. Characterization of Survey Participants

Variable	Category	n	%
Institution size	> 1,500 ME Equipments	30	66.7
	501 to 1,500 ME Equipments	7	15.6
	< 500 ME Equipments	8	17.8
Funding model	Public	24	53.3
	Private	11	24.4
	Mixed	10	22.2
Quality certification	Certified	16	35.6
	Not certified	29	64.4
Region	Southeast	19	42.2
	Northeast	9	20.0
	South	8	17.8
	Other regions	9	20.0

B. Knowledge Level of ABNT NBR IEC 62353:2019

The survey results reveal that knowledge of the ABNT NBR IEC 62353:2019 standard among CE professionals in Brazil is still in its early stages.

Of the 45 participants, only 22.2% (n=10) stated that they know the content of the standard deeply, while another 42.2% (n=19) have only superficial knowledge of it.

The most concerning data is that 35.6% of the participants have never heard of the standard or have only heard of it but are unfamiliar with its content, highlighting a significant dissemination gap for a regulation published more than five years ago.

These results may be partially explained by the recent history of CE in Brazil, which is still a maturing field. As the survey itself indicates, IEC 60601-1 remains the most widely used reference for routine testing: 30.6% of participants' institutions use it exclusively, while only 11.1% use IEC 62353. An additional 16.7% reported using both. According to Nascimento *et al.* [10], the IEC 60601 series does not encompass post-market testing, as it was designed to qualify equipment for commercialization in certified laboratories under conditions that are often destructive and incompatible with hospital routines. Laferrara *et al.* [5] corroborate this distinction by identifying IEC 62353 as the most appropriate standard for monitoring ME Equipment service life in hospital maintenance environments. Furthermore, testing according to IEC 62353 requires less time than testing per IEC 60601 [11].

Table II presents the level of knowledge regarding the IEC 62353 standard by institutional size. In institutions with more than 1,500 ME equipment, 76.7% of professionals reported familiarity with the standard, either superficially or in depth, with 30.0% claiming comprehensive knowledge. In medium-sized institutions (501 to 1,500 devices), this rate drops to 57.1%, whereas in institutions with fewer than 500 ME equipment, it reaches only 25.0%, with no participants reporting in-depth knowledge. This gap of over 50 percentage points between the extremes is unsurprising: services managing a higher volume of equipment tend to face greater operational complexity and, consequently, increased pressure to standardize technical processes. Miniati *et al.* [4] point in the same direction, observing that the formal structuring of CE processes generally scales with the institution's operational capacity. In other words, a larger equipment inventory drives a greater need and internal pressure for teams to understand and comply with existing regulatory frameworks.

TABLE II. Knowledge Level of ABNT NBR IEC 62353:2019 By Institution Size

Knowledge Level	> 1,500 MEE (*n*=30)	%	501–1,500 MEE (*n*=7)	%	< 500 MEE (*n*=8)	%
In-depth knowledge	9	30.0	0	0.0	1	12,5
Superficial knowledge	14	46.7	4	57.1	1	12,5
Heard of it, does not know	6	20,0	1	14,3	5	62,5
Does not know	1	3,3	2	28,6	1	12,5
Knows (superf. + in-depth)	**23**	**76.7**	**4**	**57.1**	**2**	**25.0**

The comparison between professionals working in hospitals with and without quality certifications presents even more significant differences, as shown in Table III. Among professionals from certified institutions (ONA, JCI, or similar), 87.5% stated they knew the standard superficially or deeply. In non-certified institutions, this rate drops to 51.7%. Deep knowledge was reported by 37.5% of professionals at certified institutions, compared with only 13.8% at non-certified institutions. This pattern is consistent with the

findings of Costa *et al*. [7], who demonstrated that accredited hospitals present higher operational efficiency and greater compliance with quality standards. Silva *et al*. [12] complement this perspective by highlighting that accreditation improves hospital infrastructure and constitutes an effective strategy to raise the quality of processes related to healthcare equipment. Within the scope of CE, Oliveira *et al*. [8] found that professionals from certified institutions report greater clarity in protocols and a higher perception of workplace safety. From an international perspective, Stern [9] described how JCI-regulated environments favor the systematic implementation of recurring electrical safety testing programs. This result aligns directly with the observations in this study.

TABLE III. KNOWLEDGE LEVEL OF ABNT NBR IEC 62353:2019 BY QUALITY CERTIFICATION

Knowledge Level	Certified (*n*=16)	%	Not Certified (*n*=29)	%
In-depth knowledge	6	37.5	4	13.8
Superficial knowledge	8	50.0	11	37.9
Heard of it, does not know	1	6.3	11	37.9
Does not know	1	6.3	3	10.3
Knows (superf. + in-depth)	**14**	**87.5**	**15**	**51.7**

The results, therefore, indicate that low familiarity with the ABNT NBR IEC 62353:2019 standard, especially in smaller and non-certified institutions, poses an obstacle to its practical adoption. Knowledge of the standard is a necessary condition for testing to be properly structured. Without this foundation, teams will find it difficult to develop internal protocols that comply with regulatory guidelines, thereby directly compromising the quality and traceability of tests performed on ME Equipment.

C. Tests Performed at Healthcare Institutions

The ABNT NBR IEC 62353:2019 standard defines three moments at which tests must be performed: upon receipt of new equipment (before entering service), after any repair, and on a recurrent basis as part of the preventive maintenance program [1]. The survey results indicate that none of these stages fully adheres to the protocol required by the standard, which includes visual inspection, functional testing, and electrical safety.

Upon receipt of the ME equipment, as presented in Table IV, visual inspection and functional testing are widely adopted practices, with 97.8% and 95.6% adherence, respectively. However, when it comes to electrical safety testing, only 22.2% of participants reported performing all three tests together, which is the complete protocol required by the standard. This data is noteworthy because receipt is, by definition, the equipment acceptance stage, and at this point, the device is verified as safe for use. Performing only visual inspection or functional testing, without electrical testing, means approving equipment without a complete functional evaluation under actual conditions of use. The discrepancy between certified and non-certified institutions reinforces this point: while 43.8% of professionals from accredited institutions perform the complete protocol upon receipt, this rate drops to 10.3% among those from non-certified institutions.

TABLE IV. TESTS PERFORMED ON ME EQUIPMENT AT RECEPTION INTO SERVICE (*N*=45)

Test type	Responses	%
No tests performed	1	2.2
Visual inspection	44	97.8
Functional test + Visual inspection	43	95.6
Functional test + visual inspection + electrical safety	10	22.2

Post-repair tests (Table V) follow a similar pattern. Visual inspection and functional testing are performed by 86.7% and 97.8% of respondents, respectively, but the complete protocol is adopted by only 31.1%. This result warrants particular attention: it is precisely at this stage that the risk of electrical failure is greatest, as the device has often been disassembled and reassembled. IEC 62353 [1] is explicit about the mandatory nature of post-repair tests, and data from Nunes et al. [6] reinforce this requirement. 23.7% of devices evaluated after repair did not meet the normative requirements. It was also observed that certified institutions show nearly twice the adherence rate of non-certified ones at this stage: 43.8% versus 24.1%.

TABLE V. TESTS PERFORMED ON ME EQUIPMENT AFTER REPAIR (*N*=45)

Test type	Responses	%
No tests performed	1	2.2
Visual inspection	39	86.7
Functional test + visual inspection	44	97.8
Functional test + visual inspection + electrical safety	14	31.1

Recurrent testing, shown in Table VI, presents the highest adherence rate to the complete protocol among the three stages (48.9%). This is expected, as recurrent tests are performed as part of preventive maintenance, an activity already well established in the routine of most CE departments. Even so, just over half of the participants still do not include electrical safety testing in this stage.

TABLE VI. RECURRENT TESTS PERFORMED ON ME EQUIPMENT (*N*=45)

Test type	Responses	%
No tests performed	5	11.1
Visual inspection	41	91.1
Functional test + visual inspection	41	91.1
Functional test + visual inspection + electrical safety	22	48.9

Table VII shows adherence to the complete protocol by stage and certification status. Certified institutions display higher adherence across all stages, with differences ranging from 20 to 33 percentage points. This outcome aligns with the observed levels of knowledge regarding the standard, suggesting that quality certification not only demands regulatory compliance but also establishes organizational

conditions that foster practical implementation and more structured protocols [7–9].

TABLE VII. ADHERENCE TO FULL PROTOCOL (ALL THREE TESTS) BY STAGE AND CERTIFICATION

Stage	Certified (*n*=16)	%	Not certified (*n*=29)	%
Reception	7	43.8	3	10.3
After repair	7	43.8	7	24.1
Recurrent	10	62.5	12	41.4

D. Challenges to Implementation

Participants were asked to indicate the primary difficulties encountered in implementing the ABNT NBR IEC 62353:2019 standard at their institutions, with multiple-choice options. The results are presented in Figure 2.

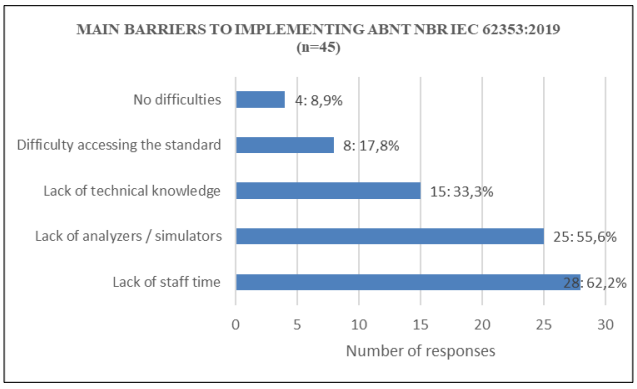


Fig. 2. Main barriers to implementing ABNT NBR IEC 62353:2019 (*n*=45).

The most frequently reported barrier was the lack of time among the technical staff, cited by 62.2% of the respondents. As a reference, Nunes et al. [6] reported a mean execution time of 5.34 ± 1.71 minutes per test (n = 363), measured with a calibrated stopwatch in a large public hospital. That interval corresponds to the complete protocol and not only to the electrical safety measurements: it covers the preparation of the equipment for testing, the visual inspection, the protective earth bond resistance measurement, the leakage current measurements, the functional test, and the completion of the test record form. This value, however, was obtained in a single institution that already had a consolidated CE team, its own analyzers, and a standardized internal operational procedure, and it should therefore be taken only as an order-of-magnitude reference, not as a parameter directly transferable to the heterogeneous institutional realities identified in the present survey, in which staff sizing, inventory size, and degree of process standardization vary widely. In addition, the time per test represents only part of the effort required by the standard, which also demands scheduling, training, metrological traceability, record keeping, and the incorporation of a new activity into an already saturated workload. The lack of time reported by the participants should thus be understood as a real structural barrier, associated with team sizing and work organization, rather than a difficulty relativized by the execution time measured in a specific study.

The lack of analyzers and simulators was reported by 55.6% of participants. This represents a concrete infrastructure barrier, as it is impossible to perform the electrical safety tests required by the standard with metrological traceability without the proper equipment. Campbell [13] lists electrical safety analyzers as essential items for the operation of a CE department, and ANVISA [14] emphasizes that the physical infrastructure of healthcare technology management services must be compatible with the activities performed.

A lack of technical or managerial knowledge was indicated by 33.3% of the participants, which aligns with the low levels of familiarity with the standard observed in this study. Lastly, 17.8% reported difficulties in assessing the standard due to its high acquisition cost, while only 8.9% reported no difficulties implementing it.

IV. CONCLUSION

The results of this study indicate that the ABNT NBR IEC 62353:2019 standard has not yet been fully incorporated into the practices of Brazilian CE departments. Only 22.2% of professionals possess in-depth knowledge of the standard, and adherence to the complete testing protocol remains low across all stages. Furthermore, the persistent reliance on the ABNT NBR IEC 60601-1:2022 series as a reference for routine testing underscores that this gap is not merely operational but also technical.

Larger, quality-certified institutions consistently demonstrated superior results, suggesting that more structured organizational environments facilitate the adoption of technical standards.

The primary barriers identified were a lack of time, the absence of analyzers, and insufficient technical knowledge. These findings demonstrate that widespread implementation of the standard requires targeted training, infrastructure investment, and the formalization of internal protocols. Ultimately, the data from this study fills a critical gap in national characterization and can support capacity-building initiatives and strategic incentives for standard adoption, particularly within smaller institutions.

ACKNOWLEDGMENT

The authors would like to thank the Federal Technological University of Paraná (UTFPR) for providing all resources for project development, and the Fundação Araucária (FA) & Superintendência de Ciência, Tecnologia e Ensino Superior (SETI) for the conference grants to H. C. C. and F. S. B.

REFERENCES

[1] Associação Brasileira de Normas Técnicas. ABNT NBR IEC 62353: Equipamento eletromédico – Ensaio recorrente e ensaio após reparo. Rio de Janeiro: ABNT, 2019.

[2] Associação Brasileira de Normas Técnicas. ABNT NBR IEC 60601-1:2010 + Emenda 1:2016 + Emenda 2:2022 - Equipamento eletromédico - Parte 1: Requisitos gerais para segurança básica e desempenho essencial. Rio de Janeiro: ABNT, 2010, 2016, 2022.

[3] Angelakis, S. C.; Saatsakis, G.; Prionas, D.; Valais, I. Electrical safety of medical equipment: an experimental approach. *e-Journal of Science & Technology (e-JST),* v. 9, n. 5, pp. 57–64, 2014.

[4] Miniati, R.; Iadanza, E.; Dori, F. Clinical Engineering: From Devices to Systems. 1. ed. San Diego: Academic Press, 2016.

[5] Laferrara, F. J. A. *et al. Ensayos recurrentes y ensayos después de reparación del equipo electromédico* (UNE-EN 62353:2009). XVIII Congreso Argentino de Bioingeniería SABI 2011.

[6] Nunes, A. F. B.; Barros, F. S.; Carvalho, H. C. Implementation of IEC 62353:2014: the importance of recurrent and post-repair testing in

electromedical equipment. Research on Biomedical Engineering, v. 42, art. 40, 2026. DOI: 10.1007/s42600-026-00486-y.

[7] Costa, C. A. S. *et al*. Serviços de acreditação hospitalar no Brasil nos últimos 3 anos. Observatório de la Economía Latinoamericana, v. 22, n. 10, e6950, 2024.

[8] Oliveira, J. L. C. de *et al. Influência da acreditação hospitalar na satisfação profissional da equipe de enfermagem*. Rev. Latino-Am. Enfermagem, v. 27, e3109, 2019.

[9] Stern, G. Using a scientific approach to meet Joint Commission AEM standards. Biomedical Instrumentation & Technology, v. 51, n. 2, pp. 147–156, 2017.

[10] Do Nascimento, J. L. G.; Costa-Felix, R. P. B. Post-market evaluation of medical electrical equipment. Research in Biomedical Engineering, v. 37, pp. 105–109, 2021.

[11] RIGEL MEDICAL (2019). A Practical Guide to the IEC 62353 Standard. Disponível em: https://www.rigelmedical.com/gb/support/guides/electrical-safety-testing/18-a-practical-guide-to-iec-62353/ . Acesso em 19/03/2024 16:30.

[12] DA SILVA, A.; MACHADO JUNIOR, C.; MARTINIANO, M. L. G.; BARRETO, L. K. da S.; PORTELLA, D. L. *Estado da arte sobre acreditação hospitalar no Brasil: um ensaio teórico apoiando o ensino*. Revista de Gestão e Secretariado, v. 14, n. 10, pp. 16756–16768, 2023. DOI: 10.7769/gesec.v14i10.2580

[13] Campbell, J. Equipment management. In: Dyro, J. F. (Ed.). Clinical Engineering Handbook. Burlington: Elsevier Academic Press, 2004.

[14] Agência Nacional de Vigilância Sanitária. Resolução da Diretoria Colegiada – RDC nº 509, de 27 de maio de 2021. Brasília: ANVISA, 2021.

Microbiological Evaluation of Contamination Syringe Plunger Handling

1st Brisa Daniely Silva Mendes
Department of Nursing
Centro de Ensino Superior em Gestão Tecnologia e Educação (FAI)
Santa Rita do Sapucaí, Brazil
briisamendes@gmail.com

2nd Elisa Rennó Carneiro Déster
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
elisa.renno@inatel.br

3rd Rita Elizabeth Santos de Almeida
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
r.elizabeth@inatel.br

4th Maysa Costa Alves
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

***Abstract*—Administration of parenteral medications requires strict adherence to aseptic techniques and technical procedures to minimize the risk of microbiological contamination and ensure patient safety. Among the potential sources of contamination, manual contact with the syringe plunger during the preparation of parenteral solutions has been identified as a possible risk factor. This study aimed to evaluate the occurrence of microbiological contamination associated with this practice. An experimental, descriptive and analytical study was conducted under controlled laboratory conditions. Different handling conditions were established, including non-hygienized hands, hands disinfected with soap and water or 70% alcohol, and the use of procedure and sterile gloves. Samples collected from the syringe plunger and the aspirated solutions were inoculated onto Brain Heart Infusion (BHI) agar supplemented with defibrinated sheep blood and incubated at 35°C for 48 hours. No microbial growth was observed in the aspirated solutions under any of the experimental conditions. However, microbiological growth was detected in samples obtained by direct inoculation of the syringe plunger surface, indicating the presence of microorganisms on this structure. These findings suggest that, under the conditions evaluated, manual contact with the syringe plunger did not promote the transfer of microorganisms to the parenteral solutions, although the plunger may serve as a microbiological reservoir. These results reinforce the importance of maintaining aseptic techniques during the preparation and administration of parenteral medications in clinical practice.**

***Index Terms*—nursing, microorganisms, syringe plunger, healthcare-associated Infections (HAIs), aseptic Technique**

I. Introduction

Parenteral medication administration is one of the most frequent and critical procedures in nursing practice, requiring strict adherence to technical standards and aseptic practices to ensure patient safety. As an invasive procedure, failures during medication preparation or administration may contribute to the occurrence of Healthcare-Associated Infections (HAIs), compromising the quality of care and increasing morbidity, mortality, length of hospital stay, and healthcare costs [1].

In this context, the proper handling of devices used during medication administration represents an essential measure for preventing contamination and reducing infection risks. The hands of healthcare professionals are recognized as one of the main routes of microorganism transmission, harboring both resident and transient microbiota that may contaminate surfaces, equipment, and materials used in patient care [2]. Therefore, appropriate hand hygiene and strict adherence to aseptic technique are fundamental practices for maintaining microbiological safety throughout the preparation and administration of medications.

Among the devices used in parenteral therapy, syringes play a central role and require specific precautions to preserve sterility. During nursing education and clinical training, one of the most emphasized recommendations is to avoid direct contact with the syringe plunger, based on the assumption that manual manipulation may facilitate the transfer of microorganisms to the internal components of the syringe and subsequently contaminate the medication solution administered to the patient [1].

Although this recommendation is widely incorporated into healthcare protocols, experimental evidence demonstrating that contact with the syringe plunger effectively results in microbiological contamination of parenteral solutions remains limited. Previous studies have reported conflicting findings, suggesting that the presence of microorganisms on a surface does not necessarily result in their transfer to another environment, as microbial transmission may be influenced by factors such as microbial load, contact time, applied pressure, environmental conditions, and characteristics of the microorganisms involved [3], [4].

Microbiological evaluation of surfaces and medical devices is therefore essential for understanding potential contamination pathways. Culture media capable of supporting the recovery of clinically relevant microorganisms, such as Brain Heart Infusion (BHI) agar supplemented with sheep blood, are

widely used for microbiological investigations due to their high nutritional capacity and ability to support the growth of microorganisms with complex nutritional requirements. Additionally, blood supplementation enables the observation of hemolytic patterns, which may contribute to the preliminary characterization of bacterial isolates [5].

Considering the importance of maintaining sterility during medication preparation and the limited experimental evidence regarding microbial transfer associated with syringe plunger manipulation, further investigations are necessary to evaluate this potential contamination pathway. A better understanding of the mechanisms involved in microbial transmission may contribute to the improvement of medication preparation practices and support evidence-based approaches to patient safety.

Therefore, the present experimental study was conducted to investigate the risk of microbiological contamination associated with manual contact with the syringe plunger during the preparation of parenteral solutions. By comparing different manipulation conditions, this study aimed to evaluate whether contact with the plunger could compromise the sterility of aspirated solutions and provide scientific evidence to support safe medication preparation and administration practices.

II. Materials and Methods

An in vitro experimental study with a qualitative, descriptive, and analytical approach was conducted. The methodological design was structured as a pilot study to evaluate the feasibility of the experimental protocol and generate preliminary data on the investigated phenomenon.

A. Study Setting and Period

The experiment was carried out at the Chemistry and Biosciences Laboratory of the National Institute of Telecommunications (INATEL), in a controlled laboratory environment. All sample handling procedures were performed inside a previously disinfected and sterilized vertical laminar airflow cabinet (Pachane, Brazil) on March 6, 2026. The incubation period lasted 48 hours, and microbiological analyses were performed on March 9, 2026. The laboratory environment was maintained under standardized hygiene and temperature conditions to minimize the influence of external factors on microbial growth.

B. Materials

The following materials were used in the experiment: sterile disposable 5-mL syringes; 25 × 7 mm hypodermic needles; sterile distilled water; sterile swabs with Stuart transport medium; sterile Petri dishes; Brain Heart Infusion (BHI) agar supplemented with defibrinated sheep blood; defibrinated sheep blood (Nova Lab, Brazil); 70% ethanol; vertical autoclave (Prismatec CS-18, 18 L); vertical laminar airflow cabinet (Pachane, Brazil); and bacteriological incubator (Mylabor, 11 L).

C. Culture Medium Preparation

Brain Heart Infusion (BHI) agar supplemented with sheep blood was used as the culture medium. Eighteen Petri dishes were prepared, with three replicates for each experimental condition. Dehydrated BHI agar was weighed according to the manufacturer's recommendations, dissolved in distilled water, and heated under constant stirring until complete homogenization. The culture medium, together with the Petri dishes previously packed in surgical-grade paper envelopes, was sterilized in an autoclave at 121°C for 15 minutes.

After sterilization, the medium was allowed to cool inside the previously sterilized laminar airflow cabinet until it reached a temperature between 45 °C and 50 °C. Defibrinated sheep blood was then added at a final concentration of 5% (v/v), followed by gentle homogenization. Finally, the medium was poured into sterile Petri dishes and allowed to solidify, forming a uniform agar layer.

D. Experimental Design

Six experimental conditions were established to simulate different handling practices during the preparation of parenteral solutions:

- Sterile gloves;
- Examination (non-sterile nitrile) gloves;
- Hands washed with soap and water;
- Hands washed with soap and water followed by 70% alcohol disinfection;
- Unwashed hands;
- Samples obtained by direct inoculation of the syringe plunger surface.

E. Experimental Procedures

Initially, the workbench, supporting tray, and all materials were disinfected with 70% ethanol to minimize environmental contamination. Sterile distilled water was then aspirated into syringes to simulate the preparation of parenteral solutions. During the experiment, intentional manual contact with the syringe plunger was performed under each experimental condition to reproduce different clinical handling practices, including situations with and without aseptic technique. All samples were appropriately labeled for subsequent microbiological analysis.

F. Sample Collection

Following syringe manipulation, samples were collected from the plunger surface using sterile swabs containing Stuart transport medium. In addition, 0.1-mL aliquots of the aspirated solutions were collected and inoculated onto the culture medium. An additional sample obtained by direct inoculation of the syringe plunger surface was prepared under the same experimental conditions. However, instead of sampling the aspirated solution, the plunger surface was directly inoculated after completion of all five experimental handling conditions in order to assess possible microbial contamination of the plunger during the experiment.

G. Microbiological Analysis

Samples were inoculated onto BHI agar plates using the spread plate technique. The plates were incubated in a bacteriological incubator at 35°C for 48 hours. After incubation, microbiological evaluation included the following parameters:

- Presence or absence of microbial growth;
- Number of colonies;
- Macroscopic characteristics of the colonies;
- Presence of agar aggregates.

H. Data Analysis

The collected data were organized into spreadsheets and analyzed descriptively. Comparisons were performed among the experimental groups to identify potential differences in the pattern and frequency of microbiological contamination under the evaluated handling conditions.

I. Ethical Considerations

Because this was an experimental laboratory study that did not involve human participants or identifiable personal data, approval by a Research Ethics Committee was not required according to the applicable national regulations. Nevertheless, all applicable biosafety standards were strictly followed throughout the experimental procedures to ensure researcher safety and the reliability of the study findings.

III. RESULTS AND DISCUSSION

Brain Heart Infusion (BHI) agar was selected because it is an enriched culture medium widely used in microbiology due to its high nutritional capacity. Its composition includes infusions of animal brain and heart tissues, as well as peptones, which provide amino acids, peptides, vitamins, and essential minerals required for bacterial growth.These characteristics support the cultivation of a wide variety of microorganisms, including fastidious (nutritionally demanding) species of clinical and environmental importance, such as bacteria belonging to the genera Streptococcus, Staphylococcus, and Enterococcus [5], [6].

When supplemented with sheep blood, BHI agar becomes even more versatile. Sheep blood provides additional nutrients, such as iron and hemin, which are essential for the growth of many fastidious microorganisms. In addition, it enables the assessment of hemolysin production, as some bacteria produce enzymes capable of lysing the red blood cells (erythrocytes) present in the culture medium. This phenomenon, known as hemolysis, is an important characteristic for the preliminary identification of several bacterial pathogens [7].

Hemolysis may occur in three distinct patterns. Alpha hemolysis is characterized by the partial lysis of red blood cells, resulting in a greenish halo surrounding the bacterial colonies due to the oxidation of hemoglobin. Beta hemolysis corresponds to the complete lysis of red blood cells, producing a clear, transparent zone around the colonies. Gamma hemolysis indicates the absence of red blood cell lysis, with no visible changes observed in the culture medium surrounding the colonies [5].

Because of these characteristics, BHI agar supplemented with sheep blood is widely used in both clinical and experimental microbiology. This medium is employed for the primary isolation of microorganisms from a variety of biological specimens, including blood, secretions, tissues, and other clinical samples. It is also extensively used in scientific research, microbiological contamination studies, and microbiological quality control [8].

After 48 hours of incubation, no microbiological growth was observed in the parenteral solutions (10 mL of sterile distilled water) under any of the experimental conditions evaluated. Regardless of the manipulation strategy employed, including contact with unwashed hands, hands sanitized with soap and water, hands sanitized with 70% alcohol, the use of sterile gloves, or procedure gloves, the samples showed no evidence of bacterial or fungal growth Figs.[1 - 5].

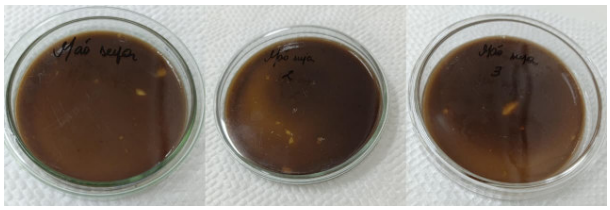


Fig. 1. Absence of microbiological growth in samples obtained under the unwashed hands condition.

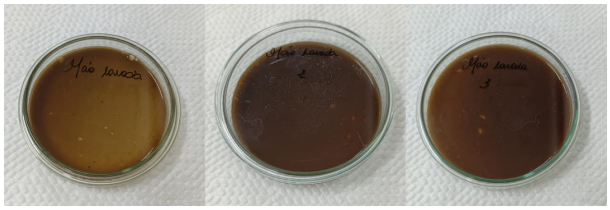


Fig. 2. Absence of microbiological growth in samples obtained after hand hygiene with soap and water.

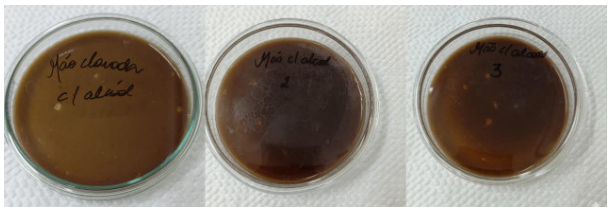


Fig. 3. Absence of microbiological growth in samples obtained after hand hygiene with 70% alcohol. The isolated structure observed was not consistent with microbial colony morphology.

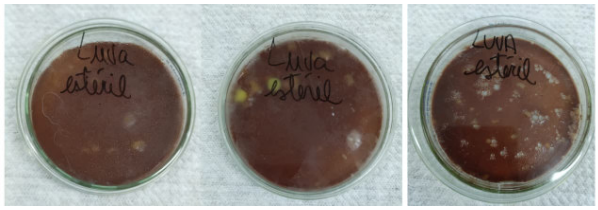


Fig. 4. Absence of microbiological growth in samples handled using sterile gloves.

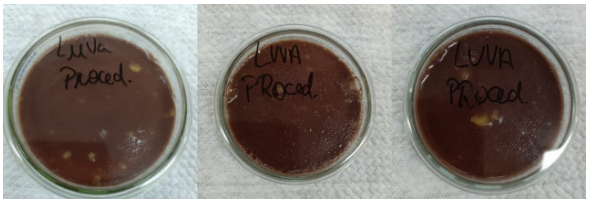


Fig. 5. Absence of microbial growth in samples handled using examination gloves.

Therefore, under the experimental conditions established in this study, manipulation of the syringe plunger did not result in detectable microbial transfer to the aspirated solution, and no differences were observed among the evaluated groups.

During the analysis of the culture plates, some structures with coloration ranging from white to yellowish and irregular morphology were observed. However, these structures did not present characteristics compatible with microbial colonies and were considered aggregates derived from the culture medium itself. Unlike bacterial or fungal colonies, these structures did not exhibit progressive growth, defined edges, spreading patterns, or typical hemolytic characteristics, indicating that they did not represent microbiological growth.

The absence of microbial growth in the parenteral solutions demonstrates the importance of differentiating the presence of microorganisms on device surfaces from the actual contamination of the medication. Surface contamination, such as that occurring on the syringe plunger, represents a potential microbiological risk; however, the transfer of microorganisms to a sterile solution depends on several factors related to the interaction among the microorganism, the contaminated surface, and the receiving environment. Therefore, the mere presence of microorganisms in a specific location does not necessarily imply their transfer to another substrate.

Microbial transfer is a complex process influenced by factors such as the initial microbial load, cellular adhesion capacity, contact time, pressure applied during manipulation, surface moisture, characteristics of the materials involved, and environmental conditions [9]. Thus, even when a device presents microbiological contamination, the migration of microorganisms to a parenteral solution may not occur or may occur at levels below the detection capacity of the analytical method employed.

Similar results were reported by Ferreira et al. (2007) [3], who evaluated contamination associated with syringe handling and observed that contact with the plunger did not necessarily result in significant microbiological contamination of solutions, particularly when basic aseptic conditions were maintained.

Likewise, Bucholz (2019) [4], in an experimental study involving devices used for medication administration, demonstrated that the presence of microorganisms on manipulated surfaces does not necessarily result in their transfer to sterile solutions or culture media, supporting the findings observed in the present study. However, although the parenteral solutions remained sterile under the evaluated conditions, samples obtained by direct inoculation of the syringe plunger surface showed significant microbiological growth Fig 6.



Fig. 6. Microbial growth in samples obtained by direct inoculation of the syringe plunger surface.

This result highlights a fundamental difference between surface contamination and medication contamination, demonstrating that the syringe plunger may act as a reservoir of microorganisms after manual manipulation, even when detectable transfer to the aspirated solution does not occur.

In the samples obtained directly from the plunger, multiple colonies with different macroscopic characteristics were observed, including variations in size, white, yellowish, and gray pigmentation, as well as cottony or velvety structures, compatible with different microbial groups, including possible fungi. However, definitive identification of the microorganisms was not performed, since morphological characterization alone does not allow taxonomic confirmation.

The recovery of microorganisms from the plunger surface reinforces that devices manipulated during medication preparation may act as microbiological reservoirs, representing a potential source of contamination under certain clinical conditions [10]. Nevertheless, these findings should not be interpreted as evidence that all surface contamination will necessarily result in medication contamination, but rather as evidence of the importance of maintaining strict aseptic practices.

From the perspective of nursing practice, the results obtained do not invalidate the recommendation to avoid direct contact with the syringe plunger. On the contrary, they reinforce the need to understand the mechanisms involved in microbial transfer and to maintain preventive measures aimed at interrupting potential contamination pathways. Aseptic technique remains essential to ensure patient safety, particularly in clinical situations where small microbial loads may represent significant risks, such as in immunocompromised patients or those undergoing intravenous therapies [11].

IV. Conclusion

The results of this study demonstrated that, under the experimental conditions evaluated, manual contact with the syringe plunger did not result in detectable microbiological contamination of the parenteral solutions, regardless of the manipulation condition employed. The absence of bacterial or fungal growth in the aspirated solutions indicates that the transfer of microorganisms from the plunger surface to the liquid medium did not occur at a significant level in this experimental model.

However, the recovery of microorganisms from samples obtained by direct inoculation of the syringe plunger surface demonstrates that this component of the syringe may act as a microbiological reservoir after manual manipulation. These findings reinforce the importance of distinguishing surface contamination from actual medication contamination, as the presence of microorganisms on a device does not necessarily imply their transfer to a sterile solution.

Although the results obtained did not indicate impairment of the sterility of the evaluated solutions, they do not eliminate the importance of strict adherence to aseptic practices during the preparation and administration of parenteral medications. Prevention of contamination remains essential in nursing care, particularly considering that clinical, environmental, and microbiological factors may influence the potential for microbial transfer under different healthcare settings.

As a pilot study, this investigation contributes to expanding the understanding of the mechanisms involved in microbial transfer during syringe manipulation and provides evidence to support future research with larger sample sizes, identification of isolated microorganisms, and more comprehensive control of experimental variables. Further studies are required to deepen knowledge regarding the microbiological risks associated with the handling of devices used in parenteral therapy and to strengthen evidence-based practices in healthcare.

V. Future Studies

As a contribution to future studies, the description and monitoring of laboratory environmental conditions will be expanded, specifying the instruments and methods employed, including the use of data-logging thermometers, hygrometers, and microbiological air monitoring techniques through passive and/or active sampling methods. Considering the experimental nature of the study and its focus on the assessment of microbiological contamination, the systematic characterization of environmental conditions and the establishment of quality control procedures may contribute to enhancing the reliability, reproducibility, and validity of the results obtained.

References

[1] P. A. Potter and A. G. Perry, *Fundamentos de Enfermagem*, 11th ed. Rio de Janeiro, Brazil: GEN Guanabara Koogan, 2024.

[2] P. R. G. S. Galvão *et al.*, "Pathogenic microorganisms on the hands of multidisciplinary intensive care teams: a hidden risk," *Caderno Pedagógico*, vol. 22, no. 1, 2025. [Online]. Available:https://ojs.studiespublicacoes.com.br/ojs/index.php/cadped/article/view/13126/7423. Accessed: Apr. 22, 2026.

[3] A. M. Ferreira *et al.*, "To touch or not to touch the syringe plunger," *Revista Enfermagem UERJ*, vol. 15, no. 1, 2007. [Online]. Available: https://www.researchgate.net/publication/317456641_Technique_for_the_preparation_of_parenteral_drug_touching_or_not_touching_the_piston. Accessed: Apr. 6, 2026.

[4] A. O. Bucholz *et al.*, "Contamination of aspirated solutions caused by contact with the syringe plunger: a microbiological analysis," *Saúde e Biociências*, vol. 1, no. 1, 2019. [Online]. Available: https://seer.unifatea.edu.br/index.php/saudebiociencias/article/view/53/39. Accessed: Apr. 12, 2026.

[5] G. J. Tortora *et al.*, *Microbiology: An Introduction*, 12th ed. Porto Alegre, Brazil: Artmed, 2017.

[6] M. T. Madigan *et al.*, *Brock Biology of Microorganisms*, 16th ed. New York, NY, USA: Pearson, 2021.

[7] B. A. Forbes *et al.*, *Bailey & Scott's Diagnostic Microbiology*, 12th ed. St. Louis, MO, USA: Mosby Elsevier, 2007.

[8] J. G. Cappuccino and N. Sherman, *Microbiology: A Laboratory Manual*, 10th ed. Boston, MA, USA: Pearson, 2014.

[9] A. C. de Oliveira *et al.*, "Hospital environmental surfaces as potential reservoirs of resistant bacteria: a review," *Revista da Escola de Enfermagem da USP*, vol. 44, no. 4, pp. 1118–1123, 2010.

[10] L. F. D. R. Bacca and P. B. Gouvea, "Contact contamination of the syringe plunger during the preparation of parenteral solutions: a case-control study," *Revista Científica da Univali*, 2025. [Online]. Available: https://periodicos.univali.br/index.php/asic/article/view/21757. Accessed: Apr. 22, 2026.

[11] Ministry of Health of Brazil, *Patient Safety and Quality in Healthcare Services*. Brasília, Brazil: Ministry of Health, 2017. [Online]. Available: https://www.gov.br/anvisa/pt-br/centraisdeconteudo/publicacoes/servicosdesaude/manuais/cadernos-da-serie-seguranca-do-paciente-e-qualidade-em-servicos-de-saude-2017. Accessed: Apr. 26, 2026.

Mitigando e Caracterizando Relativamente EMI Radiadas em Salas Cirúrgicas de Alta Complexidade

Anderson Alves dos Santos*, William Knob de Souza*, Gabriel Fallavena Chaves*, Rodrigo Lopes Rezer*
*Engenharia Clínica, Hospital Moinhos de Vento, Porto Alegre, RS, Brasil
{anderson.santos, william.knob, gabriel.chaves, rodrigo.rezer}@hmv.org.br

***Resumo* — A incorporação de tecnologias de maior sensibilidade em salas cirúrgicas de alta complexidade reduz a margem de imunidade eletromagnética do ambiente construído. Manifestada a interferência, o serviço de engenharia clínica não dispõe de meio acessível para localizar a fonte, dimensionar a intervenção e comprovar a correção. Propõe-se um ciclo de mitigação — detecção, localização, intervenção e verificação — executável pela equipe própria com analisador portátil de US$ 170, sem câmara anecoica e sem rastreabilidade a padrões nacionais. A grandeza é a elevação do ruído eletromagnético ambiente radiado sobre a condição de referência da própria sala, obtida por subtração pareada: medição relativa, que reduz a dependência de calibração absoluta. O ciclo foi aplicado a uma sala de hemodinâmica com artefato de imagem intravascular (IVUS) e perda de conexão do sistema de reserva de fluxo fracionada (FFR) associados ao acionamento da iluminação LED. A detecção registrou 13,23 dB de elevação em 30 a 150 MHz, e a amostragem em sete pontos indicou emissão distribuída pelo circuito. Após a substituição dos drivers das luminárias, a elevação caiu para 1,04 dB — redução de 12,19 dB no indicador relativo, aprovada segundo o critério interno nos sete pontos; duas salas de referência sem intervenção, medidas no mesmo dia, mantiveram 12,40 e 12,68 dB, fortalecendo a atribuição do fenômeno ao circuito de iluminação e à intervenção realizada. Emergiu um achado metodológico: a faixa de 225 a 575 MHz indicou 2,94 dB e sugeria conformidade operacional, contra 13,01 dB no mesmo ponto e dia sob a faixa crítica. A faixa de análise determina a conclusão e deve ser fixada por caracterização inicial do próprio ambiente.**

***Index Terms* — compatibilidade eletromagnética, mitigação de interferência, emissão radiada, medição relativa, engenharia clínica, salas cirúrgicas, ultrassom intravascular, iluminação LED**

I. INTRODUÇÃO

A prática cirúrgica e intervencionista incorpora, de forma progressiva, tecnologias de maior resolução e sensibilidade. Sistemas de ultrassom intravascular (IVUS) e de reserva de fluxo fracionada (FFR) operam com sinais de amplitude reduzida conduzidos por cateteres de pequeno diâmetro; navegação eletromagnética e plataformas robóticas dependem de sensores igualmente suscetíveis. Cada geração de equipamento amplia a informação disponível ao operador e, ao mesmo tempo, *reduz a margem de imunidade eletromagnética disponível* no ambiente em que opera.

O arcabouço normativo acompanhou essa evolução no que diz respeito ao equipamento. A IEC 60601-1-2 [1] estabelece ensaios de imunidade e limites de emissão para o equipamento eletromédico, verificados pelo fabricante antes da colocação no mercado. A CISPR 15 [2] regula a perturbação radioelétrica de equipamentos de iluminação, e a versão brasileira da norma colateral [3] reproduz esse escopo. Todas atuam sobre o produto isolado, em condição de ensaio, antes da instalação.

O ambiente construído permanece fora desse escopo, e a consequência é operacional antes de ser conceitual. Quando a interferência se manifesta em uma sala em uso, o serviço de engenharia clínica precisa responder a quatro perguntas em sequência — existe interferência? de onde vem? o que substituir? a substituição resolveu? — e não dispõe de método acessível para nenhuma delas. Na ausência de medição, a investigação recai sobre o equipamento médico, que é o elemento visível da cadeia: trocam-se cabos e módulos, o quadro não se altera, e o custo da tentativa se acumula sem que a fonte seja alcançada. A mitigação, quando ocorre, é conduzida por presunção.

A literatura sobre compatibilidade eletromagnética em ambiente hospitalar concentra-se em levantamentos de campo e em fontes móveis. Ishida et al. [4] caracterizaram os campos eletromagnéticos de um hospital antes e após a ocupação, concluindo pela conformidade aos limites normativos sob condições normais de uso. Estudos de caso em unidades de terapia intensiva seguem abordagem semelhante, comparando os campos medidos aos limites da IEC [12]. A iluminação como fonte foi investigada por Ishida et al. [13], que avaliaram a interferência radiada por foco cirúrgico LED sobre sistema de navegação eletromagnética e concluíram por interferência improvável naquele arranjo.

Três aspectos permanecem pouco explorados, e todos os três bloqueiam a mitigação. Primeiro, esses trabalhos medem o *nível absoluto* de campo e o comparam a limites normativos, ao passo que a degradação relatada na prática decorre da *elevação relativa* do ruído eletromagnético ambiente sobre a condição de referência da própria sala — grandeza que um limite absoluto não captura. Segundo, avaliam a sala em um instante, sem tratar a instalação como algo que se altera por reformas ao longo da vida útil. Terceiro, os arranjos empregados dependem de instrumentação e de *expertise* raramente disponíveis ao serviço de engenharia clínica, o que impede a reexecução da medida depois da intervenção — e sem medição posterior não existe mitigação comprovada, apenas serviço executado.

O objetivo deste trabalho é propor e aplicar um ciclo de mitigação eletromagnética de salas de alta complexidade que seja (i) executável pela própria equipe de engenharia clínica, (ii) exequível sem câmara anecoica e sem rastreabilidade a padrões nacionais, (iii) capaz de orientar o escopo da intervenção a partir da medição, (iv) encerrado por critério objetivo de aceitação e (v) de custo compatível com a realidade das instituições brasileiras.

Um resultado colateral mostrou-se tão relevante quanto a proposta: a faixa de análise escolhida pode inverter o veredito sobre uma mesma sala, medida no mesmo dia (Seção V-D, Fig.

4). A faixa é parte do critério, não um parâmetro de configuração.

II. CASO MOTIVADOR

Em fevereiro de 2026, a equipe de engenharia clínica foi acionada para investigar e mitigar artefato de imagem no sistema de orientação multimodal AVVIGO+ (Boston Scientific) [14] de uma sala de hemodinâmica: granulação na imagem IVUS e perda intermitente de conexão do FFR Link. A iluminação não era suspeita; chegou-se a ela por eliminação.

A etapa inicial foi a anamnese técnica junto à equipe assistencial, que delimitou as condições de ocorrência. Como conduta primária, o equipamento foi conectado a ponto de energia isolado do sistema IT médico e o artefato persistiu; a substituição de cabos e módulo também não alterou o quadro. O elemento decisivo veio do relato de que o mesmo equipamento operara sem falha em sala adjacente: confrontadas as duas instalações, identificou-se que a única diferença fora de procedimento era a iluminação, desligada onde o teste fora bem-sucedido. Desligado o circuito de iluminação da sala com a anomalia, a interferência cessou de imediato; religado, retornou. Estabeleceu-se assim que o artefato decorria de ruído eletromagnético originado nas luminárias do ambiente — e nada havia sido medido: nenhum desses passos produz número que justifique intervir na instalação predial, e é essa lacuna que o método a seguir pretende fechar.

III. MATERIAIS E MÉTODOS

A. Instrumentação

Todas as campanhas foram realizadas com o analisador de espectro portátil TinySA Ultra (modelo ZS405, firmware 1.4-104), com antena telescópica na entrada de baixa frequência. O instrumento dispunha de calibração de fábrica — ajuste de nível do fabricante, exatidão de amplitude declarada de ±1 dB [20] —, porém sem rastreabilidade a padrões nacionais (Seção III-D). O ZS405 saiu de linha após as campanhas; o modelo atual, TinySA Ultra+ (ZS406), executa o mesmo procedimento por US$ 170 — custo de referência para quem reproduza o protocolo. A escolha da classe de instrumento é deliberada: a proposta perde sentido se depender de equipamento inacessível ao serviço, porque é a própria equipe que reexecuta a medida após a intervenção.

Os parâmetros de varredura, mantidos constantes em todas as campanhas, estão na Tabela I. Nenhuma calibração de nível absoluto foi aplicada; as amplitudes em dBm referem-se à entrada do analisador e destinam-se exclusivamente à comparação relativa entre condições, conforme discutido na Seção VI-D.

Tabela I
PROTOCOLO DE MEDIÇÃO

Parâmetro	**Configuração**
Instrumento	TinySA Ultra ZS405 (atual: Ultra+ ZS406, ordem de US$ 170)
Infraestrutura	nenhuma — sem câmara nem laboratório
Faixa	0 a 350 MHz, 450 pontos
RBW / VBW	850 kHz
Atenuação	0 dB, nível de referência 0 dBm
Protocolo	A/B — luminárias acesas × apagadas
Amostragem	7 pontos por sala
Execução	equipe própria, ≈ 2 h por sala
Critério	elevação ≤ 3 dB em 30 a 150 MHz (critério interno)

B. Protocolo A/B de detecção

O elemento central do método é a comparação pareada entre duas condições que diferem exclusivamente pelo estado da iluminação. Para cada ponto de medição registra-se uma varredura com as luminárias acionadas e outra com as luminárias apagadas, sem deslocar a antena entre as duas aquisições e mantendo inalterado o estado de todos os demais equipamentos da sala.

A grandeza de interesse é a *elevação*, definida como a diferença ponto a ponto entre as duas condições:

$$\Delta(f) = A_{\text{acesa}}(f) - A_{\text{apagada}}(f) \tag{1}$$

onde A(f) é a amplitude em dBm na frequência f. Por ser uma razão entre potências, Δ é expressa em dB — e não em dBm, unidade reservada às amplitudes absolutas dos eixos.

A subtração pareada reduz a dependência de calibração absoluta: contribuições comuns às duas condições — radiodifusão, telefonia, demais equipamentos — cancelam-se, restando a parcela atribuível à iluminação. Essa propriedade é o que torna o método executável em sala assistencial, sem desmontar a instalação nem interromper o serviço. A Fig. 1 situa a cadeia causal que o protocolo isola dentro da sequência de investigação que a produziu.

Cabe delimitar a grandeza medida. Com antena no interior da sala, o arranjo caracteriza a componente *radiada* da perturbação. A emissão conduzida pela rede, também regulada pela CISPR 15, exigiria rede fictícia de acoplamento e interrupção do circuito — inviável em sala em operação. Ambas compartilham a origem no chaveamento do driver, de modo que a componente radiada serve como indicador acessível do fenômeno.

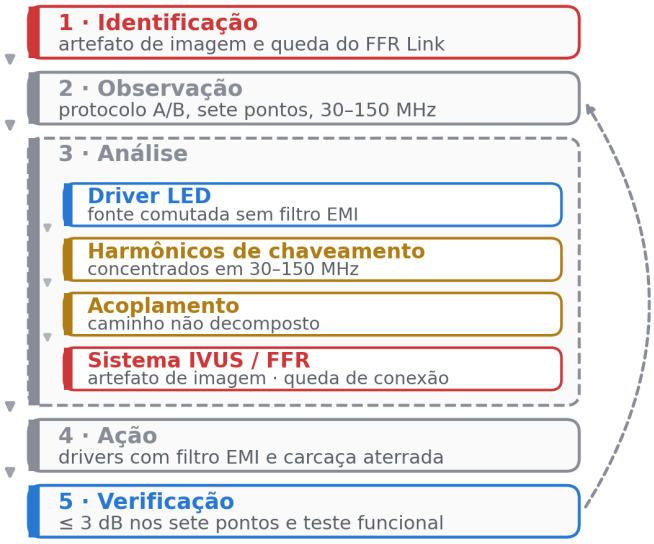


Figura 1. Sequência de investigação em cinco fases e a cadeia causal que a terceira delas isola — produto da análise, não premissa. O protocolo A/B opera sem desmontar a instalação, condição para execução em sala assistencial, e a verificação realimenta a observação: reprovar um único dos sete pontos reabre o ciclo. O caminho de acoplamento não é decomposto entre as componentes conduzida e radiada (Seção VI-D).

C. Amostragem espacial e localização da fonte

Cada sala foi amostrada em sete posições: entrada, centro e fundo de ambos os lados da mesa cirúrgica, além do ponto central. A amostragem atende a dois propósitos, e o primeiro é o que dimensiona a intervenção: a dispersão entre pontos distingue emissão *localizada* — compatível com luminária ou driver individual defeituoso — de emissão *distribuída*, que aponta o circuito de iluminação como um todo. O segundo é assegurar que a aprovação da sala segundo o critério interno não dependa de um ponto favorável.

A distinção tem consequência direta de escopo e de custo. Diagnosticada emissão localizada, a substituição pontual resolve; diagnosticada emissão distribuída, a substituição pontual consome recurso sem alterar o resultado, e o escopo precisa abranger o conjunto.

D. Situação de calibração e rastreabilidade

O analisador empregado **não dispunha de calibração vigente rastreável** durante as campanhas — a calibração de fábrica declarada na Seção III-A não constitui rastreabilidade —, e não houve janela de calibração disponível antes das medições. A condição é declarada aqui, e não apenas na seção de limitações, porque condiciona a leitura de todos os resultados.

Cabe distinguir o que essa ausência afeta. O ILAC-G24/OIML D 10 [5] não estabelece intervalo universal de recalibração: o intervalo deve ser definido e justificado pelo usuário a partir do histórico de deriva, da intensidade de uso e da consequência de um resultado fora de tolerância. A ISO/IEC 17025 [19] exige que exista programa de calibração, sem fixar periodicidade. Não há, portanto, um prazo normativo que tenha sido descumprido — há uma lacuna de rastreabilidade, que é coisa distinta.

Para a grandeza medida neste trabalho, a consequência é limitada por construção. Sendo Δ uma diferença entre duas leituras tomadas com o mesmo arranjo, separadas por segundos e sem deslocamento da antena, os erros sistemáticos comuns às duas leituras cancelam-se:

$$\Delta_{\text{med}} = (A_{\text{acesa}} + e_{\text{sist}}) - (A_{\text{apagada}} + e_{\text{sist}}) = \Delta_{\text{real}} \quad ($$

onde e_sist agrega ganho de antena, perda de cabo, resposta do pré-amplificador e desvio absoluto de calibração. O que *não* se cancela é tratado na Seção VI-D.

Uma tabela de correção será acrescentada ao repositório de dados tão logo o instrumento seja calibrado, permitindo reprocessar as varreduras brutas e avaliar a repetibilidade das medições aqui apresentadas (ver *Disponibilidade dos Dados*).

E. Critério de aceitação

Adotou-se como critério que a elevação média em 30 a 150 MHz não ultrapasse 3 dB em nenhum dos sete pontos. A reprovação de um único ponto reprova a sala. A faixa foi delimitada empiricamente, pela caracterização inicial: é onde a elevação se concentrou e onde ela acompanha o acionamento da iluminação. Não decorre de especificação de suscetibilidade publicada para os sistemas envolvidos — os fabricantes não a disponibilizam em forma de faixa — e essa delimitação empírica é parte do que o trabalho propõe como método.

Em consequência, a faixa não é transferível sem reexame: outra sala, com outras tecnologias instaladas, exige nova caracterização inicial para fixá-la.

Trata-se de critério operacional interno de triagem e aceitação. O valor de 3 dB corresponde à paridade entre a contribuição acrescentada e o ruído de fundo preexistente: duas potências não correlacionadas de igual magnitude somam-se em 3,01 dB ($10 \log_{10} 2$) — a mesma relação que leva o método de medição invocado pela CISPR 15 [2] a exigir margem mínima de 6 dB sobre o ambiente [8]. Não constitui limite normativo de emissão ou de imunidade, nem se propõe como limiar universal de interferência: os limites da CISPR 15 são absolutos, em dBμV/m, e não se aplicam a uma grandeza relativa. Corroboram a escolha a repetibilidade do arranjo (desvio-padrão de 0,89 dB entre as cinco varreduras sucessivas da campanha inicial; o critério equivale a cerca de 3,5 vezes esse valor) e o desfecho clínico.

Distinguem-se dois planos, mantidos separados em todo o texto: **conformidade normativa** — atendimento a limite absoluto de norma, verificável apenas em ensaio acreditado com instrumento rastreável — e **conformidade operacional** — atendimento ao critério interno de 3 dB sobre a grandeza relativa Δ. Toda "aprovação" neste trabalho é do segundo plano (redução relativa aprovada segundo o critério interno), e as conclusões limitam-se à triagem, à orientação da intervenção e à verificação relativa.

Um instrumento sem rastreabilidade não deve comunicar resultado com aparência de valor metrológico. Adotou-se, por isso, uma **classificação ordinal de severidade** sobreposta à medida, com limiares ancorados em referências deste trabalho e

não arbitrados: **Classe 0**, não detectável ($\Delta \leq 1$ dB, dentro da repetibilidade do arranjo); **Classe I**, detectável e conforme o critério interno (1 a 3 dB); **Classe II**, não conforme o critério interno (3 a 10 dB); e **Classe III**, não conformidade severa ($\Delta >$ 10 dB, o dobro da incerteza que a CISPR 16-4-2 [6] admite em ensaio acreditado). A classe é o que se comunica ao corpo clínico; o valor em dB é o que se registra e reprocessa.

IV. PROTOCOLO DE MITIGAÇÃO

A medição orienta a intervenção, mas não a substitui. Esta seção descreve como o resultado da detecção se converte em escopo de serviço executável e verificável.

A. Hierarquia de intervenção

A ordem de intervenção segue o princípio de atacar a fonte antes do acoplamento. Quatro níveis, em ordem de precedência:

1. **Supressão na fonte** — substituição dos drivers por modelos com filtragem EMI integrada. É o único nível que reduz a emissão gerada, e não apenas o quanto dela alcança o equipamento suscetível.
2. **Aterramento funcional** — conexão das carcaças metálicas à barra de terra da sala, fornecendo caminho de retorno para as correntes de modo comum.
3. **Roteamento** — separação física entre o cabeamento de alimentação das luminárias e o cabeamento de sinal.
4. **Blindagem** — último recurso, por custo e por interferir na manutenibilidade da instalação.

No caso relatado, a dispersão espacial reduzida indicou emissão distribuída por todo o circuito, e o escopo foi definido no primeiro nível, com o segundo e o terceiro executados como medidas associadas. Cabe registrar o limite dessa evidência: os níveis não foram comparados isoladamente entre si, de modo que a hierarquia é derivada do padrão espacial medido e do princípio de supressão na fonte, e não de contraste experimental entre alternativas.

B. Especificação por desempenho

A especificação do componente substituto foi redigida por requisito de desempenho, não por marca: filtro EMI integrado com emissões dentro dos limites da CISPR 15; conformidade declarada às normas EN 55015 [18], EN 61000-3-2 Classe C [16] e EN 61547 [15]; carcaça metálica com ponto de aterramento funcional; fator de potência superior a 0,95; distorção harmônica total inferior a 10 %.

A forma da especificação não é detalhe editorial. Redigida por marca, ela restringe a concorrência e não é utilizável em processo de compra pública ou regulado; redigida por desempenho, admite equivalente técnico de qualquer fabricante mediante laudo comprobatório, e transfere ao fornecedor o ônus de demonstrar conformidade.

C. Aceitação vinculada à verificação

O recebimento do serviço foi condicionado a nova medição, nos mesmos sete pontos e com os mesmos parâmetros, com aprovação segundo o critério interno exigida em todos eles. Essa vinculação é o elemento que converte a medição em instrumento contratual: sem ela, o serviço é aceito por execução e não por resultado, e a persistência da interferência após a intervenção não gera consequência para o fornecedor.

O ciclo completo — detectar, localizar, especificar, intervir, verificar — encerra-se apenas quando a verificação aprova. A Tabela II e a Fig. 2 documentam esse fechamento no caso relatado.

V. RESULTADOS

A. Detecção e caracterização

A campanha de fevereiro de 2026 registrou, com as luminárias acionadas, elevação média de 13,23 dB na faixa de 30 a 150 MHz. O piso de ruído eletromagnético da sala subiu de −87,6 dBm para −79,1 dBm na média de 5 a 350 MHz. O maior pico da faixa de análise atingiu −61,0 dBm em 77,2 MHz — elevação de 25,3 dB.

Doze máximos locais excederam a média da banda em mais de um desvio-padrão, distribuídos entre 60,8 e 130,2 MHz, o maior deles em 77,2 MHz. Os espaçamentos entre eles são irregulares (5,4 a 16,4 MHz), o que não caracteriza série harmônica resolvida — e não poderia: uma fonte comutada de driver LED opera com fundamental de dezenas a centenas de kHz [11], de modo que as ordens situadas em 60 a 130 MHz distam entre si menos que a largura de banda empregada, e cada *bin* integra dezenas delas. A estrutura observada é, portanto, envelope de emissão, não espectro de raias.

A atribuição da fonte ao circuito de iluminação não decorre dessa estrutura, e sim do protocolo A/B: a emissão aparece e desaparece com o acionamento, no mesmo ponto e no intervalo de segundos.

B. Mitigação e verificação

A intervenção consistiu na substituição dos drivers das luminárias por modelos atendendo à especificação da Seção IV, com aterramento funcional das carcaças e revisão do encaminhamento do cabeamento de alimentação. A ação determinante foi a substituição do driver; aterramento e roteamento acompanharam-na como boas práticas de instalação, e não como variáveis ensaiadas isoladamente.

Os componentes são descritos por características técnicas, não por marca. O driver original, conforme a etiqueta do componente inspecionado, era de 36 W e 425 mA, com entrada de 100 a 250 VAC em 50/60 Hz, fator de potência declarado superior a 0,99, distorção harmônica total inferior a 10 %, eficiência de 92 % e saída aproximada de 84 V/425 mA, alojado em carcaça polimérica ventilada sem ponto de terra identificado. O substituto é fonte de corrente constante com ajuste de corrente, em carcaça metálica com terra funcional e com especificações formais de emissão e imunidade declaradas pelo fabricante.

Esses números excluem uma explicação: o componente original já apresentava fator de potência e distorção harmônica dentro dos valores tidos como adequados, de modo que a qualidade de energia não explica a emissão observada. A diferença plausível está na arquitetura construtiva — filtragem de EMC, caminhos de retorno e de terra, controle de emissões do estágio chaveado —, que este trabalho não isolou: a substituição coincidiu com a redução do indicador relativo e com a resolução

funcional do artefato, sem que se atribua o efeito a um elemento interno específico.

A verificação pós-intervenção, realizada em julho de 2026 nos sete pontos, registrou elevação média de 1,04 dB — redução de 12,19 dB no indicador relativo em relação à campanha inicial, com valores individuais entre 0,34 e 1,73 dB — todos abaixo do critério interno. Nenhum dos máximos anteriores permanece destacado: aplicado o mesmo critério de seleção à condição posterior, nenhum ponto da faixa excede a média em mais de um desvio-padrão. A Fig. 2 apresenta a comparação espectral.

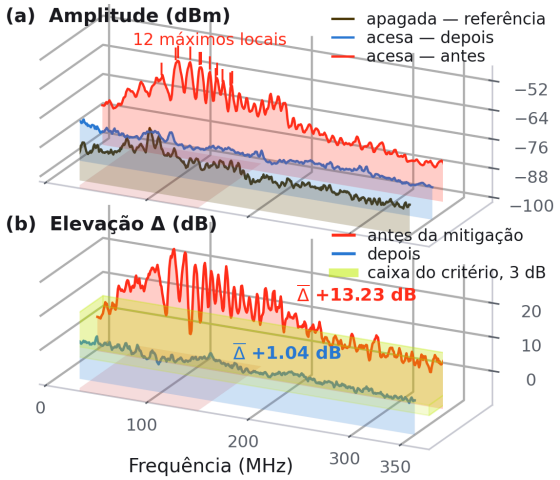


Figura 2. Sala 3 antes e depois da mitigação, em vista isométrica — a profundidade separa condições que em duas dimensões se sobrepunham. (a) Amplitude absoluta: ao fundo a campanha inicial, ao centro a de verificação, à frente a referência com as luminárias apagadas; as hastes assinalam os máximos locais que excedem a média da faixa em mais de um desvio-padrão. (b) Elevação Δ(f) nas duas campanhas, com o critério marcado no eixo e Δ̄ indicando média da faixa. Os primeiros 2 MHz foram omitidos por conterem vazamento de oscilador local, artefato do instrumento.

C. Salas de referência sem intervenção

Duas salas do mesmo serviço, sem qualquer intervenção, foram medidas no mesmo dia, com o mesmo instrumento, os mesmos parâmetros e o mesmo protocolo. Os resultados constam da Tabela II e da Fig. 3.

Tabela II
ELEVAÇÃO POR SALA EM 30 A 150 MHZ

Sala	Elevação média (dB)	DP (dB)	Faixa (dB)	Classe
3 — após mitigação	+1,04	0,52	0,34 – 1,73	I
3 — antes	+13,23	0,89*	—	III
2 — referência	+12,40	1,74	9,24 – 14,32	III
1 — referência	+12,68	0,70	11,62 – 13,77	III

Elevação média: média das elevações dos sete pontos (21 pares de varreduras por sala). DP: desvio-padrão espacial entre os sete pontos — dispersão entre posições de medição, calculada dos dados brutos; não corresponde à incerteza metrológica do instrumento, tratada na Seção VI-D. (*) Campanha inicial, ponto único: DP de repetibilidade entre as cinco varreduras sucessivas no mesmo ponto. Salas 1 e 2: referência, sem intervenção. Classes da Seção III-E (critério interno).

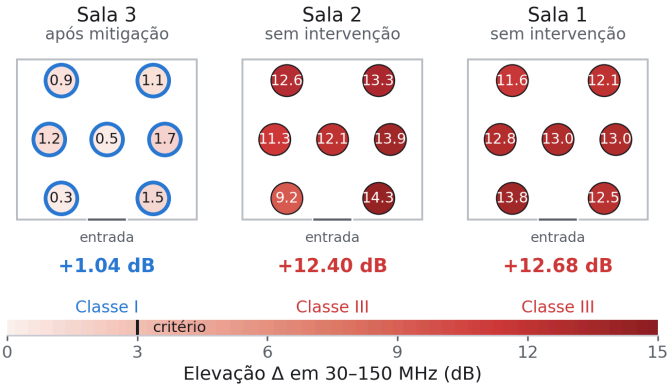


Figura 3. Elevação nos sete pontos de cada sala, em esquema de planta com a entrada ao pé. Cada marcador é uma amostra, não um campo interpolado: sete posições não sustentam reconstrução espacial contínua. Contorno azul indica ponto dentro do critério interno.

Estas salas não constituem controle experimental no sentido clássico — não houve alocação nem intervenção parcial —, mas referência contemporânea: permaneceram sem intervenção e mantiveram elevação acima de 12 dB, enquanto apenas a sala submetida à substituição dos drivers caiu abaixo do critério operacional adotado. A comparação simultânea é relevante para a inferência: não se trata apenas de comparação temporal — sujeita a mudanças não registradas no ambiente — mas de contraste entre salas medidas nas mesmas condições instrumentais, diferindo quanto à intervenção realizada. Um desvio instrumental afetaria as três salas igualmente e não produziria essa separação.

A dispersão entre pontos também informa o escopo das intervenções pendentes. Na Sala 1 a variação é de 2,1 dB — entre 11,62 e 13,77 dB, desvio-padrão de 0,70 dB —, padrão de emissão distribuída que indica substituição do conjunto. Na Sala 2 a variação alcança 5,1 dB, com desvio-padrão de 1,74 dB e o pior ponto na entrada, compatível com contribuição localizada somada à distribuída. Medidas apenas no ponto central, as duas salas apresentariam valores próximos — 13,0 e 12,1 dB — que ocultariam essa diferença de escopo.

D. Efeito da faixa de análise

Uma medição preliminar da Sala 1, restrita à faixa de 225 a 575 MHz, indicou elevação média de 2,94 dB — abaixo do critério de 3 dB e, portanto, sugestiva de conformidade operacional. Tomada no mesmo ponto central e no mesmo dia, a faixa de 30 a 150 MHz devolveu 13,01 dB: subestimação de 10,07 dB sob confronto pareado por posição, ilustrada na Fig. 4. Considerada a média dos sete pontos, a sala reprova com 12,68 dB.

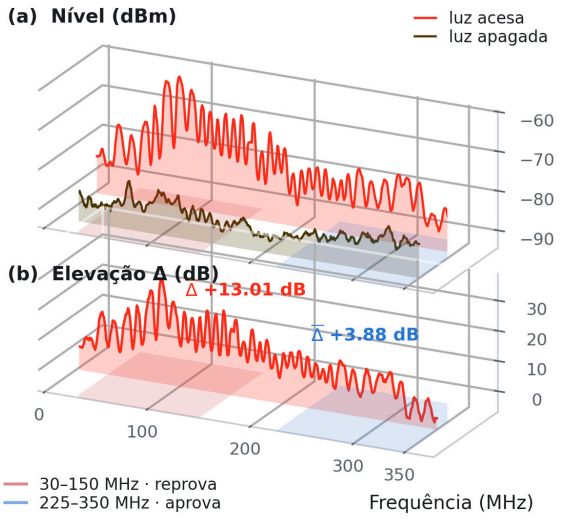


Figura 4. A faixa de análise decide o resultado. Vista isométrica do ponto central da Sala 1 sem intervenção: em (a) os níveis absolutos com a iluminação acionada e apagada, em (b) a elevação Δ(f) daí resultante. A projeção no piso assinala as faixas de análise, e o veredito de cada uma contra o critério de 3 dB consta da legenda. A barra superior, $\bar{\Delta}$, declara média da faixa e não pico, conforme ISO 80000-2 [9] e GUM [10]: bins isolados na faixa crítica chegam a 30,5 dB, mais que o dobro. A projeção é ortográfica: sem perspectiva, a mesma distância mede igual em qualquer profundidade.

A comparação retrospectiva confirma o padrão: na campanha de fevereiro, com o fenômeno clinicamente ativo, a faixa de 225 a 350 MHz registrava apenas 3,93 dB, enquanto 30 a 150 MHz indicava 13,23 dB. A faixa inadequada subestimou o fenômeno em 9,3 dB.

A consequência para a mitigação é direta: a faixa inadequada não produz apenas imprecisão, mas inversão de decisão. Em classe de severidade, a sala migra de III — não conformidade severa — para I — detectável e conforme o critério interno. A investigação teria sido encerrada com a sala liberada e a interferência intacta.

Este resultado alterou a evolução do próprio trabalho: a faixa de análise deixou de ser parâmetro de configuração e passou a integrar o critério de aceitação (Seção III-E); a caracterização inicial tornou-se etapa obrigatória do ciclo — é ela que fixa a faixa — e a declaração da faixa junto ao resultado converteu-se na recomendação metodológica central do artigo (Seção VII). Por documentar uma decisão que seria invertida, o resultado figura nesta seção.

VI. DISCUSSÃO

A. Por que a mitigação falha sem medição

A medição altera três decisões que, sem ela, são tomadas por presunção: onde intervir — ambiente e não equipamento; *quanto* intervir — conjunto e não luminária isolada, conforme o padrão espacial; e quando encerrar — por critério verificado e não por serviço executado. Nenhum protocolo vigente de liberação de área teria detectado a condição, porque nenhum a verifica.

O contraste com [13] é instrutivo e não contraditório. Naquele estudo, Ishida et al. mediram o ruído radiado por focos cirúrgicos LED sobre um sistema de navegação cirúrgica eletromagnética — acoplamento indutivo em baixa frequência — e concluíram por interferência improvável. Seu impacto aqui é duplo: legitima a iluminação LED como hipótese de fonte desde a anamnese do caso motivador, e mostra que a mesma classe de fonte produz desfechos opostos conforme a tecnologia exposta — a susceptibilidade observada neste caso situa-se em 30 a 150 MHz e afeta a cadeia de recepção de sinais de cateter. Isso reforça o argumento central: a verificação ambiental precisa ser referida ao que está instalado naquela sala, e não a um limite único.

B. Relação com o problema clínico original

O ciclo foi desencadeado por um problema clínico concreto — granulação na imagem IVUS e perda intermitente de conexão do FFR Link — e a relação dos resultados com a sua solução deve ser caracterizada com precisão. A cadeia de evidência: a anamnese estabeleceu correlação temporal reprodutível (Seção II); o protocolo A/B a converteu em grandeza medida — 13,23 dB em 30 a 150 MHz, coincidente com o acionamento; a intervenção removeu a causa proposta e a verificação mediu a remoção (1,04 dB, com as salas de referência sem intervenção acima de 12 dB); após a intervenção, não houve novos relatos de artefato nem de queda de conexão até o fechamento deste texto.

O alcance tem limite declarado: o desfecho clínico é observação qualitativa, sem registro sistematizado de eventos, e a verificação funcional prolongada não havia sido concluída. Os resultados demonstram a remoção verificada da causa ambiental identificada e corroboram, pela cessação dos relatos, a resolução do problema original — não constituem demonstração definitiva de causalidade clínica, que exigiria a relação dose–resposta apontada na Seção VI-D como trabalho futuro.

C. Generalização do protocolo

Propõe-se a incorporação do ciclo de mitigação aos procedimentos de engenharia clínica em três momentos: na liberação inicial da sala, como linha de base contra a qual medições futuras serão comparadas; após qualquer reforma que altere iluminação, cabeamento ou aterramento; e na investigação de falhas recorrentes não explicadas pelo equipamento.

A linha de base na liberação inicial é o elemento de maior retorno e o de menor custo: sem ela, toda medição posterior compara a sala apenas consigo mesma no estado atual; com ela, torna-se possível atribuir uma degradação à reforma que a introduziu. O protocolo é compatível com essa aplicação: dispensa laboratório e executa-se em cerca de duas horas por sala com equipe própria.

D. Limitações

As amplitudes registradas em dBm na entrada do analisador permitem comparação relativa entre condições, mas não constituem medida de conformidade normativa à CISPR 15, que estabelece limites em dBµV/m com antena calibrada e distância normalizada. A CISPR 32 [17], por exemplo, fixa a emissão radiada Classe B em 30 dBµV/m na faixa de 30 a 230 MHz a 10 m — valor com o qual os resultados aqui apresentados não podem ser confrontados. A proposta não substitui ensaio acreditado; destina-se à triagem, à orientação da intervenção e à veri-

ficação relativa de rotina, contextos em que o ensaio acreditado é economicamente inviável.

Conforme a Seção III-D, a subtração pareada cancela os erros sistemáticos comuns às duas leituras. Resta, porém, um termo que ela não elimina: a **não linearidade de amplitude** do instrumento. Como as duas condições são lidas em níveis distintos — −73 dBm com as luminárias acionadas contra −86 dBm com elas apagadas —, sobrevive na diferença o desvio de linearidade ao longo desse intervalo:

$$\Delta_{\text{med}} = \Delta_{\text{real}} + \left[\, e(A_{\text{acesa}}) - e(A_{\text{apagada}}) \,\right] \qquad (3)$$

A exatidão de amplitude declarada pelo fabricante é de ± 1 dB [20], o que limita o termo residual a cerca de 2 dB. Para referência, a CISPR 16-4-2 [6] atribui a uma medição radiada *acreditada* incerteza de instrumentação de 4,95 dB a 3 m. A elevação observada, de 12 a 13 dB, excede essa incerteza em duas vezes e meia — margem que sustenta a conclusão sob a hipótese mais desfavorável quanto ao instrumento.

Para uma diferença tomada no intervalo de segundos, a grandeza relevante não é o intervalo de calibração, mas a estabilidade de curto prazo do arranjo, caracterizada por desvio-padrão de 0,89 dB entre as cinco varreduras sucessivas da campanha inicial — cerca de um décimo quinto do efeito medido. A repetibilidade pós-intervenção não foi caracterizada por varreduras sucessivas (a verificação registrou um par por ponto); o indicador disponível — desvio-padrão de 0,52 dB entre os sete pontos — é compatível com a repetibilidade do arranjo, e a caracterização direta, pendente, será depositada no repositório de dados.

Permanece fora de alcance a equivalência metodológica com a norma: o instrumento não dispõe de detector de quase-pico nem das larguras de banda normativas de 200 Hz, 120 kHz e 1 MHz exigidas pela CISPR 16-1-1 [7], e analisadores sem pré-seleção não são admitidos em medição de conformidade. Literatura de referência em diagnóstico de EMI [11] situa instrumentação dessa classe como adequada à localização de fontes e à verificação relativa, e explicitamente não à medição de nível para fins de conformidade — delimitação que este trabalho adota.

O critério de 3 dB é operacional. Foi estabelecido a partir da repetibilidade do arranjo e apoiado pelo desfecho — ausência de novos relatos da equipe assistencial após a mitigação, observação qualitativa sem registro sistematizado — mas não decorre de relação dose–resposta estabelecida entre elevação do ruído eletromagnético ambiente e degradação mensurável do sinal clínico. Não existia registro sistematizado dos eventos clínicos e nenhum foi instituído durante o estudo; o estabelecimento dessa relação, com registro estruturado de eventos clínicos como desfecho, constitui trabalho futuro.

O caminho de acoplamento não foi decomposto. Mediu-se a componente radiada; a componente conduzida pela rede de alimentação, que compartilha a mesma origem no chaveamento do driver, não foi caracterizada por indisponibilidade de rede fictícia de acoplamento e pela impossibilidade de interromper o circuito de uma sala em operação. Como a intervenção realizada atua sobre ambas — filtro integrado ao driver, aterramento de carcaça e roteamento —, o desfecho não discrimina qual caminho predominava no acoplamento ao equipamento suscetível. Para o objetivo de mitigar, a indistinção é aceitável; para atribuir mecanismo, não é, e o trabalho não faz essa atribuição.

Uma observação de alcance decorre da faixa em que o fenômeno se concentrou. O ensaio de imunidade a campos de RF radiados da IEC 60601-1-2 [1] inicia-se em 80 MHz; abaixo dessa fronteira, a imunidade do equipamento é qualificada por ensaio de perturbações conduzidas induzidas por RF, com o cabeamento como caminho de acoplamento. Dos 154 pontos espectrais de 30 a 150 MHz da campanha inicial, 64 (42 %) situam-se abaixo de 80 MHz. Não se conclui daí violação de requisito — as amplitudes são relativas e não admitem confronto com limites em V/m —, mas apenas que parte do fenômeno recai em região cuja qualificação de imunidade não se dá por ensaio radiado. Registre-se ainda que não há, em norma alguma, limite para o ruído eletromagnético ambiente da sala: o critério de aceitação ambiental é lacuna a preencher na liberação da área, não requisito a transcrever de norma existente.

A hierarquia de intervenção da Seção IV não foi submetida a contraste experimental entre níveis: apenas o primeiro foi exercido, com o segundo e o terceiro como medidas associadas, de modo que a contribuição individual de cada um permanece não quantificada. A campanha inicial não incluiu amostragem espacial, o que impede reconstruir a distribuição da emissão anterior à mitigação. A verificação funcional prolongada do equipamento sob iluminação acionada, embora prevista, não havia sido concluída até o fechamento deste texto.

VII. CONCLUSÃO

Demonstrou-se que a interferência eletromagnética radiada de origem predial em sala de hemodinâmica é detectável, localizável, corrigível e *verificável* por medição relativa, com instrumentação de custo da ordem de US$ 170 operada pela própria equipe de engenharia clínica. A mitigação reduziu a elevação do ruído eletromagnético ambiente de 13,23 para 1,04 dB na faixa crítica — redução de 12,19 dB no indicador relativo —, aprovada segundo o critério interno nos sete pontos, enquanto duas salas de referência sem intervenção mantiveram valores superiores a 12 dB — resultado que fortalece a atribuição do fenômeno ao circuito de iluminação e à intervenção realizada.

Quanto ao problema de origem, os relatos de artefato IVUS e de queda do FFR cessaram após a mitigação e não haviam retornado até o fechamento deste texto — remoção verificada da causa ambiental e corroboração da resolução, não demonstração definitiva de causalidade clínica (Seção VI-B).

O ciclo proposto contribui em três pontos além da detecção. A amostragem espacial converte a medição em escopo de serviço, distinguindo emissão distribuída de localizada. A especificação por desempenho torna a intervenção executável em processo de compra, sem restringir a concorrência. E a vinculação do recebimento à verificação transfere o ônus do resultado ao fornecedor, o que distingue mitigação comprovada de serviço executado.

O achado metodológico mais transferível é que a faixa de análise determina a conclusão: restringir a medição a frequências inadequadas produziu aparente conformidade operacional e subestimou o fenômeno em cerca de 9 dB — não uma imprecisão, mas uma inversão de decisão. Qualquer protocolo de mitigação eletromagnética ambiental deve, portanto, fixar a faixa por caracterização inicial do próprio ambiente, e declará-la junto ao resultado.

Recomenda-se a incorporação da verificação eletromagnética aos procedimentos de liberação e de manutenção de áreas cirúrgicas de alta complexidade, à medida que a sensibilidade das tecnologias incorporadas reduz a margem de imunidade disponível no ambiente construído. Os resultados aqui relatados permanecem no plano da triagem, da orientação de intervenção e da verificação relativa — não substituem ensaio EMC acreditado.

DISPONIBILIDADE DOS DADOS

As 52 varreduras brutas que sustentam este trabalho — 42 da campanha de verificação (três salas, sete pontos, duas condições por ponto), seis da campanha inicial e quatro que documentam o falso negativo por faixa — estão depositadas em repositório público de acesso aberto, sob licença CC BY 4.0, com os parâmetros de aquisição, o protocolo de amostragem e o código que reproduz todas as tabelas aqui apresentadas:

https://github.com/DiyMV/emi-iluminacao-led-salas-cirurgicas

O repositório inclui caderno de verificação executável em navegador (Colab), que recalcula cada valor publicado a partir dos CSV brutos. As amplitudes são publicadas exatamente como saíram do instrumento, sem fator de antena, perda de cabo ou correção de calibração. O conjunto não contém dado de paciente — as medições são de emissão ambiente, em salas sem procedimento em curso — nem identificador de instituição, pessoa, ativo ou documento administrativo; a declaração de anonimização acompanha um verificador executável.

Conforme a Seção III-D, o analisador não dispunha de calibração vigente durante as campanhas. Assim que a calibração for realizada, a tabela de correção e o respectivo certificado serão acrescentados à pasta calibracao/ do mesmo repositório, com o recálculo de todas as tabelas lado a lado com os valores publicados. Uma calibração posterior não torna estas medições retroativamente rastreáveis — a rastreabilidade é propriedade da medição no momento em que é feita —, mas permitirá caracterizar a resposta do instrumento, limitar sua deriva e verificar se a correção desloca de forma material as conclusões relativas.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Em atendimento à Portaria CNPq nº 2.664, de 6 de março de 2026, declara-se o uso de ferramenta de inteligência artificial generativa (Claude, Anthropic) nas seguintes finalidades: elaboração dos programas de processamento das varreduras e de geração das figuras; elaboração e revisão do texto; e levantamento de referências normativas.

Todo o material assim produzido foi submetido a revisão criteriosa pelos autores. Cada valor publicado foi reconferido contra as varreduras brutas por rotina de verificação independente, depositada no repositório de dados; as divergências identificadas nesse processo foram corrigidas antes da submissão. O texto reproduz com fidelidade os achados de medição e as interpretações formuladas pelos autores, a quem pertencem a concepção do estudo, a execução das campanhas, a condução da intervenção e as conclusões. A responsabilidade pela exatidão dos dados, das referências e de todas as afirmações é integralmente dos autores.

AGRADECIMENTOS

À equipe de Hemodinâmica, pela disponibilização das salas para as campanhas de medição, e à equipe de Manutenção Predial, pela execução da intervenção.

REFERÊNCIAS

[1] IEC, IEC 60601-1-2:2014+AMD1:2020 — Medical electrical equipment — Part 1-2: Electromagnetic disturbances. Geneva, 2020.

[2] IEC, CISPR 15:2018+AMD1:2021 — Radio disturbance characteristics of electrical lighting equipment. Geneva, 2021.

[3] ABNT, NBR IEC 60601-1-2:2020 — Equipamento eletromédico — Parte 1-2: Perturbações eletromagnéticas. Rio de Janeiro, 2020.

[4] K. Ishida et al., “Evaluation of electromagnetic fields in a hospital for safe use of medical equipment,” J. Med. Syst., 2016.

[5] ILAC e OIML, ILAC-G24 / OIML D 10:2022 — Determination of recalibration intervals of measuring equipment, 2022.

[6] IEC, CISPR 16-4-2:2011+AMD1:2014 — Part 4-2: Measurement instrumentation uncertainty. Geneva, 2014.

[7] IEC, CISPR 16-1-1 — Part 1-1: Measuring apparatus. Geneva.

[8] IEC, CISPR 16-2-2 — Part 2-2: Methods of measurement of disturbances and immunity — Measurement of disturbance power. Geneva, 2010.

[9] ISO, ISO 80000-2:2019 — Quantities and units — Part 2: Mathematics. Geneva, 2019.

[10] JCGM, JCGM 100:2008 — Evaluation of measurement data: Guide to the expression of uncertainty in measurement (GUM), 2008.

[11] P. André e K. Wyatt, EMI Troubleshooting Cookbook for Product Designers. Edison, NJ: SciTech Publishing, 2014.

[12] V. S. Fernandes, R. A. A. R. Felix e A. E. P. Correia, “Electromagnetic interference in hospital environment: case study of the intensive care units of a university hospital,” in Proc. IEEE MTT-S Int. Microw. Optoelectron. Conf. (IMOC), 2021, doi: 10.1109/IMOC53012.2021.9624826.

[13] K. Ishida et al., “Evaluation of electromagnetic noise radiated from LED surgical lights on an electromagnetic navigation system,” 2018.

[14] Boston Scientific, AVVIGO+ Multi-Modality Guidance System — Operator’s Manual, 2023.

[15] CENELEC, EN 61547 — General lighting equipment: EMC immunity requirements. Brussels.

[16] IEC, IEC 61000-3-2 — Part 3-2: Limits for harmonic current emissions. Geneva.

[17] IEC, CISPR 32:2015+AMD1:2019 — Electromagnetic compatibility of multimedia equipment — Emission requirements. Geneva, 2019.

[18] CENELEC, EN 55015 — Limits and methods of measurement of radio disturbance characteristics of electrical lighting and similar equipment. Brussels.

[19] ISO/IEC, ISO/IEC 17025:2017 — General requirements for the competence of testing and calibration laboratories. Geneva, 2017.

[20] E. Brinkman, tinySA / tinySA Ultra — User Manual: specifications and amplitude accuracy. [Online]. Disponível: https://tinysa.org

Multiple Analytical Perspectives on Hospital Maintenance Data: Reliability, Seasonality, and Support through Language Models

Yuri Cavalcanti Silva Araújo
Department of Biomedical Engineering
Federal University of Pernambuco (UFPE)
Recife, Brazil
yuri@yuriaraujo.eng.br
ORCID: 0009-0004-7133-2552

Yan Ribeiro de Melo
Department of Production Engineering
Federal University of Pernambuco (UFPE)
Recife, Brazil
yan.ribeiro@ufpe.br
ORCID: 0000-0001-6487-8742

Guilherme Moraes Bezerra
Technology Center
Federal University of Pernambuco (UFPE)
Caruaru, Brazil
guilherme.gmb@ufpe.br
ORCID: 0009-0006-7291-4655

Augusto José da Silva Rodrigues
Technology Center
Federal University of Pernambuco (UFPE)
Caruaru, Brazil
augusto.jsrodrigues@ufpe.br
ORCID: 0000-0002-7042-9270

Cristiano Alexandre Virginio Cavalcante
Department of Production Engineering
Federal University of Pernambuco (UFPE)
Recife, Brazil
cristiano.avcavalcante@ufpe.br
ORCID: 0000-0003-1466-656X

Wellington Pinheiro dos Santos
Department of Biomedical Engineering
Federal University of Pernambuco (UFPE)
Recife, Brazil
wellington.santos@ufpe.br
ORCID: 0000-0003-2558-6602

***Abstract*—Hospital maintenance records accumulated in CMMS systems represent an underused information asset in decision-making in Clinical Engineering. The present investigation structured an analytical framework applied to the operational history (2020 to 2025) of five categories of medical-hospital equipment, prioritized through multi-criteria selection (CAPEX impact, maintenance volume, and clinical criticality), to support health technology management. The reliability and maintainability assessment, conducted through the application of the Kruskal-Wallis test, Kaplan-Meier estimators, and the Cox proportional hazards model, indicated the manufacturer as a performance predictor with significantly greater statistical weight than cumulative equipment usage (Age Proxy). In the Electric Bed samples, the variance explained by the asset brand reached Eta^2 = 0.3787 for repair time (MTTR). The technical weight of this variable provides quantitative support for the design of new institutional guidelines for the acquisition and renewal of the technology inventory. Time-series modeling based on STL decomposition mapped seasonal behaviors inherent to each technology category. The maintenance volume follows regular care and epidemiological cycles, measured by a Seasonal Index (SI) of 231.8 for Pulmonary Ventilators in June and an SI of 158.3 for Electric Beds in September. The proposed incorporation of large language models (LLMs) adds a semantic processing layer to the numerical indicators. Textual analysis of the work orders generated during seasonal peak months is proposed to classify and detail the root causes and failure modes of the equipment. The proposed integration between stochastic modeling and natural language interpretation constitutes a replicable analytical framework aimed at transforming data generated in the hospital environment into operational intelligence to support decision-making.**

***Keywords*—healthcare asset management, CMMS, reliability, seasonality, language model, clinical engineering.**

I. Introduction

The management of medical-hospital equipment inventories plays a central role in the quality and safety of care provided in healthcare facilities. Over the years, computerized maintenance management systems (CMMS) have become substantial repositories of operational information, recording not only failure events but also response times, mobilized resources, and recurring patterns of technical intervention throughout the life cycle of each technology [1], [2].

In practice, however, these databases are limited to storing operational records, without fully exploring the analytical potential of data regarding the general or specific behavior of assets. Even when some type of analysis is incorporated, which is not the most common situation, Clinical Engineering and Health Technology Assessment (HTA) tend to explore them in a segmented manner: either focusing on reliability and maintainability indicators [1], [3], [6], or analyzing temporal demand patterns [4], [5], without articulating the multiple analytical dimensions that the same database can reveal simultaneously. This fragmentation limits synthesis capacity and reduces the use of the operational knowledge available in historical maintenance records.

More recently, advances in large language models (LLMs) have opened a concrete possibility for processing the textual content of work orders, which historically remained outside any systematic analysis [7]. Failure descriptions, technical observations, and intervention records can now be read automatically, making it possible to identify recurring semantic patterns and connect this qualitative information to classical quantitative indicators.

To broaden support for managers and promote effective use of the data already available in CMMS systems, this study proposes a multi-perspective approach that articulates different analyses to address real technology management problems. To this end, we explore three complementary dimensions from the same database: reliability and maintainability analysis by manufacturer, seasonal decomposition of maintenance demand, and the use of LLMs as an interpretive layer for the iden-

tified seasonal patterns. The objective is not to examine each method in isolation, but to demonstrate that the articulation among these perspectives reveals relationships and insights that independently conducted analyses would not be able to capture, offering a replicable framework for healthcare asset management guided by operational evidence.

II. Materials and Methods

A. Database and Equipment Selection

The analyzed database comes from the CMMS of a hospital network and includes maintenance records from 2020 to 2025, totaling 638,629 records in the global repository prior to category filtering, and 218,028 consolidated work orders (WOs) for the five studied technology categories after complete preprocessing (filtering of critical classes, removal of incoherent/unclassified records, duplicate consolidation, and elimination of non-positive intervention times).

To identify the equipment most relevant to the study, a multi-criteria selection was conducted considering three complementary dimensions. Initially, groups with significant participation in hospital capital expenditure (CAPEX) were selected, regardless of institutional complexity, considering both unit cost and aggregate investment resulting from installed quantities. From this initial group, equipment with higher volumes of preventive and corrective work orders was prioritized, ensuring greater data availability for analysis. Finally, clinical criticality was considered, prioritizing equipment whose failure could severely compromise healthcare delivery or pose a risk to patient life. The combined application of these criteria directed the selection to five groups: Electric Bed, Multiparameter Monitor, Anesthesia Machine, Pulmonary Ventilator, and Cardioverter. Table I presents the descriptive profile of each group, including the exact count of unique physical assets (Tags).

For all reliability analyses (MTBF calculation, Kaplan-Meier survival curves, and Cox regression), a failure event was exclusively defined by the closing of WOs structurally classified as "Corrective Maintenance" in the CMMS database. To model reliability accurately, continuous hours of operation were used as the standard time unit. The time between failures (TBF) was calculated by measuring the exact time elapsed, in continuous hours, between the closing of a corrective WO and the opening of the subsequent corrective WO for the same unique equipment identifier (tag). Repeated failures of the same equipment were treated longitudinally, allowing a single tag to contribute multiple operational intervals in the survival modeling. The last chronological observation recorded for each tag at the end of the collection period was treated statistically as right-censored, as the equipment might not have presented a new failure by the end of the observation window.

B. Reliability and Maintainability Analysis by Manufacturer

Mean Time Between Failures (MTBF) and Mean Time to Repair (MTTR) were calculated by manufacturer for each type of equipment, using continuous operating hours and technical intervention hours, respectively. Manufacturer groups were compared using the nonparametric Kruskal-Wallis test, with effect size estimated by the Eta^2 coefficient [8]. Interpretation followed the thresholds proposed by Cohen [9]: small effect ($\text{Eta}^2 \geq 0.01$), medium effect (≥ 0.06), and large effect (≥ 0.14).

Survival analysis was performed using Kaplan-Meier curves [10], with manufacturers compared by the log-rank test. To assess the simultaneous influence of multiple operational factors on the risk of failure, a multivariate Cox proportional hazards regression model was constructed [11]. Alongside the manufacturer dummy indicators, the model included the following fundamental operational control covariates:

- **Cumulative usage (Age Proxy):** The accumulated number of WOs for a specific equipment tag, acting as a surrogate variable to represent intensity of use and accumulated wear over time in the absence of exact manufacturing timestamps.
- **Preventive Maintenance Intensity (PM Intensity):** A continuous variable calculated through a rolling temporal window, accounting for the rate of preventive WOs the equipment (same tag) received in the exact 180 days prior to the analyzed failure event.
- **Annual and Semiannual Seasonality (Sine/Cosine):** To control for cyclical demand oscillations throughout the year, the WO opening month was converted into two-dimensional coordinates on the trigonometric circle (Sine and Cosine), capturing the continuous nature of calendar transitions.

Clustered robust standard errors were incorporated at the equipment tag level to account for intra-asset correlation and avoid risk distortion from repeat-failure units (the "lemon effect"). The model estimated Hazard Ratios (HRs) and 95% confidence intervals. All analyses were performed in Python using the *lifelines*, *scipy*, and *statsmodels* libraries.

C. Seasonal Decomposition of Demand

The STL (Seasonal-Trend decomposition using Loess) technique was selected to model monthly maintenance demand because of its ability to flexibly separate trend, seasonality, and residual components without imposing rigid parametric assumptions on the time series [12]. The Seasonal Index (SI) was calculated for each month, allowing the identification of peak and trough demand periods and the magnitude of seasonal fluctuations. Equipment with $SI > 150$ was classified as having high seasonal variability.

D. Proposed Interpretive Support Using Large Language Models

An intrinsic limitation of temporal decomposition methods is that they reveal when maintenance demand increases, but do not explain why the seasonal pattern is specific to each type of equipment or the qualitative nature of the failures concentrated at the peaks. To advance this dimension, the use of LLMs is proposed as an interpretive layer complementary to seasonal analysis.

Table I
CHARACTERIZATION OF THE SELECTED MEDICAL EQUIPMENT CATEGORIES.

Equipment	Total WOs	Corrective WOs (Failures)	Unique Assets (Tags)	Manufacturers	Period	Selection
Anesthesia Machine	11,378	2,139	736	15	2020–2025	Multi-Criteria
Cardioverter	69,120	2,057	1,342	12	2020–2025	Multi-Criteria
Electric Bed	9,536	5,709	2,671	33	2020–2025	Multi-Criteria
Multiparameter Monitor	70,601	22,462	7,898	52	2020–2025	Multi-Criteria
Pulmonary Ventilator	57,393	9,758	7,570	37	2020–2025	Multi-Criteria
Total	**218,028**	**42,125**	**20,201**	**124***	**2020–2025**	**–**

*Number of distinct manufacturer brands across all five analyzed critical equipment categories.

The methodology consists of selecting the work orders recorded during the peak months identified by STL decomposition ($SI > 150$) and submitting their descriptive fields to the LLM, requesting the identification and semantic grouping of the predominant failure patterns. Comparing the qualitative profiles between peak periods and normal periods makes it possible to reveal whether seasonal increases in demand are associated with specific types of failure, recurring components, or patterns of intensive clinical use. Table IV summarizes the analytical questions proposed for each equipment type, derived directly from the seasonal patterns identified in Section III-B.

III. RESULTS

A. Reliability and Maintainability by Manufacturer

For all five types of equipment analyzed, the Kruskal-Wallis test indicated significant differences in MTTR among manufacturers ($p < 0.001$). Although statistical significance was observed across categories, there was substantial variation in the magnitude of the effect measured by Eta^2. For Electric Beds ($\text{Eta}^2 = 0.3787$) and Multiparameter Monitors ($\text{Eta}^2 = 0.2660$), the MTTR effect size clearly exceeded the large-effect threshold (≥ 0.14) according to Cohen's criteria [9], demonstrating that manufacturer choice accounts for over a quarter to more than a third of the variance in repair times. In terms of reliability (MTBF), significant differences across manufacturers were confirmed for Multiparameter Monitors ($\text{Eta}^2 = 0.1341, p < 0.001$), Electric Beds ($\text{Eta}^2 = 0.1201, p < 0.001$), and Pulmonary Ventilators ($\text{Eta}^2 = 0.0332, p < 0.001$), whereas Cardioverters ($p = 0.067$) and Anesthesia Machines ($p = 0.115$) did not reach significance at $\alpha = 0.05$. Table II presents the complete results by equipment type.

The Kaplan-Meier survival curves revealed pronounced differences in reliability trajectories among manufacturers of the same equipment category. For Electric Beds, Manufacturer 04 demonstrated the highest survival probability, maintaining a stable operational curve with survival probability remaining above 0.40 throughout extended operational hours. In contrast, Manufacturer 43 exhibited the lowest reliability profile, characterized by severe early mortality with survival probability dropping sharply within the initial operational intervals ($t_{25} < 500$ hours). Intermediate survival trajectories were observed for Manufacturers 66, 12, and 52. The log-rank test confirmed that survival distributions across all evaluated manufacturers differ significantly ($p < 0.001$).

These results indicate that assets from lower-performing brands present elevated failure rates during initial operational phases, leading to premature downtime. From the perspective of hospital management, this disparity directly drives higher consumption of technical hours, accelerated turnover of mechanical components, and prolonged bed unavailability, emphasizing the necessity of integrating CMMS survival history into formal procurement criteria.

The multivariate Cox proportional hazards model quantified the relative contribution of manufacturer identity against operational control covariates, including cumulative usage (Age Proxy), preventive maintenance intensity, and cyclical seasonality. As illustrated in Fig. 2 for Pulmonary Ventilators, operational controls—Cumulative Usage (Age Proxy: HR ≈ 1.00, $p > 0.05$), Intensity of Preventive Maintenance (HR ≈ 1.00, $p > 0.05$), and Seasonality Cosine (HR ≈ 1.01, $p > 0.05$)—exhibited no statistically significant independent effect on failure hazard. The Seasonality Sine covariate captured a statistically significant cyclical demand oscillation ($p < 0.05$), consistent with mid-year respiratory surges.

Manufacturer identity emerged as the single dominant determinant of failure risk ($p < 0.05$). The model identified Manufacturers 33, 18, and 06 as the highest-risk brands (HR > 1.00), exhibiting substantially accelerated failure rates relative to the reference brand under identical operating regimes. Conversely, statistically significant protective effects were demonstrated by Manufacturers 14, 07, 11, 54, 34, and 92 (HR < 1.00), which exhibited significantly reduced hazards of corrective failure.

The pronounced spread in Hazard Ratios between protective and high-risk brands confirms that manufacturer selection is the primary structural factor governing medical technology reliability, substantially outweighing cumulative operational age and routine preventive maintenance frequency.

B. Seasonality of Maintenance Demand

STL decomposition showed that each type of equipment has its own maintenance demand calendar. This is not random variation: the identified seasonal patterns have sufficient regu-

Table II
RESULTS OF THE KRUSKAL-WALLIS TEST AND EFFECT SIZE (ETA^2) ACROSS MANUFACTURERS BY EQUIPMENT TYPE.

	Mean Time to Repair (MTTR)				Mean Time Between Failures (MTBF)			
Equipment	H	p-**value**	**Eta**^{2}	**Magnitude**	H	p-**value**	**Eta**^{2}	**Magnitude**
Anesthesia Machine	88.47	< 0.001	0.0389	Small	10.25	0.115	0.0128	Small†
Cardioverter	229.33	< 0.001	0.1078	Medium	14.63	0.067	0.0167	Small†
Electric Bed	2174.01	< 0.001	0.3787	Large	155.31	< 0.001	0.1201	Medium
Multiparameter Monitor	5992.08	< 0.001	0.2660	Large	492.95	< 0.001	0.1341	Medium
Pulmonary Ventilator	812.01	< 0.001	0.0807	Medium	75.71	< 0.001	0.0332	Small

†Not statistically significant ($p \geq 0.05$). Effect size thresholds (Cohen [9]): Small ($\text{Eta}^2 \geq 0.01$), Medium (≥ 0.06), Large (≥ 0.14).

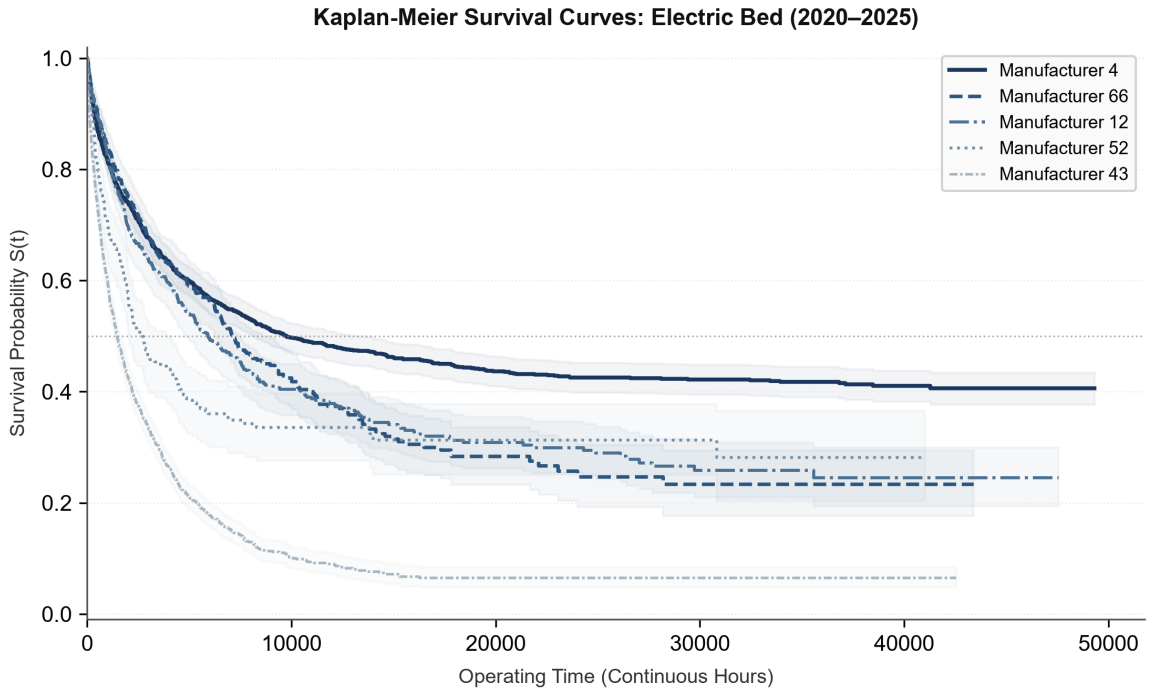


Figure 1. Kaplan-Meier survival curves stratified by manufacturer for Electric Bed (2020–2025). Manufacturer 04 (solid dark curve) represents the benchmark with superior operational longevity, stabilizing above 0.40 survival probability. Conversely, Manufacturer 43 (dashed light curve) exhibits rapid initial degradation and high early mortality ($p < 0.001$, log-rank test). 95% CI shaded.

larity and magnitude to inform operational planning. Table III presents the maximum and minimum Seasonal Indices by equipment type.

The STL decomposition revealed marked seasonal dynamics across all five medical technologies. For Pulmonary Ventilators, maintenance demand reached its annual maximum in June (SI $= 231.8$), coinciding with the winter surge in respiratory illnesses and elevated intensive care occupancy in the Southern Hemisphere. This concentration reflects operational stress during peak clinical utilization, followed by a demand trough in August (SI $= 67.8$).

Electric Beds exhibited their demand peak in September (SI $= 158.3$), driven by high hospital bed occupancy rates and accelerated patient turnaround in inpatient units during the third quarter, contrasting with a trough in August (SI $= 59.9$).

Anesthesia Machines displayed a peak in February (SI $= 128.4$), reflecting the resumption and rescheduling of elective surgical procedures following the year-end recess period. Multiparameter Monitors presented peak demand in August (SI $= 128.5$), associated with continuous ICU throughput, intensive sensor replacement cycles, and periodic transducer calibrations. In contrast, Cardioverters demonstrated limited seasonal variation throughout the year (SI between 77.9 in August and 114.5 in January), consistent with devices maintained in standby readiness where technical interventions are predominantly driven by routine inspection protocols rather than acute utilization fluctuations.

C. Emerging Perspective: LLM Support for the Qualitative Interpretation of Seasonality

Although identifying the period of increased demand is relevant, efficient management requires the qualitative characterization of recorded failures. This technical detail goes

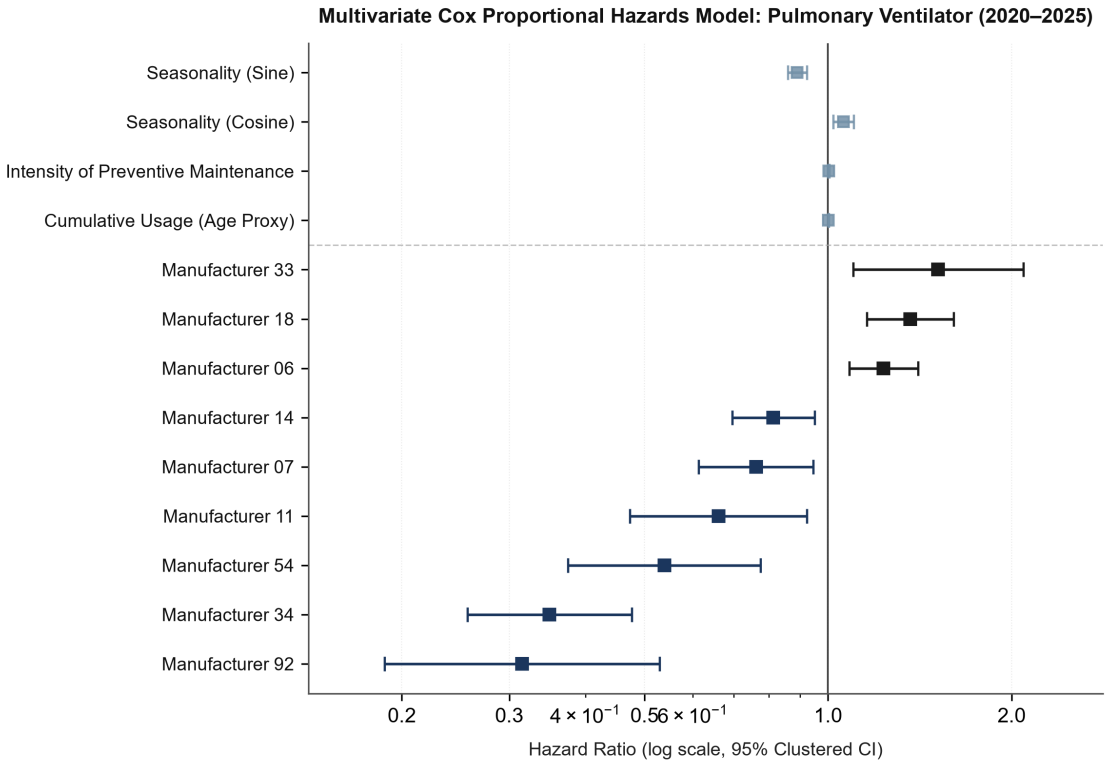


Figure 2. Hazard Ratios (log scale, 95% Clustered CI) from the multivariate Cox Proportional Hazards Model for Pulmonary Ventilator (2020–2025). Robust standard errors clustered at the asset tag level. High-risk brands (e.g., Manufacturers 33, 18, and 06) exhibit significantly elevated failure hazards (HR > 1.00), whereas protective benchmark brands (e.g., Manufacturers 14, 07, 11, 54, 34, and 92) demonstrate substantial hazard reductions (HR < 1.00). Operational controls (Age Proxy and PM Intensity) remain centered on the null line (HR ≈ 1.00, $p > 0.05$).

Table III
MAXIMUM AND MINIMUM SEASONAL INDICES (SI) BY EQUIPMENT TYPE.

Equipment	**Max. SI**	**Peak Month**	**Min. SI**	**Trough Month**	**Hypothesis**
Pulmonary Ventilator	231.8	June	67.8	August	Respiratory disease seasonality (winter ICU demand surge)
Electric Bed	158.3	September	59.9	August	Inpatient volume surge and post-winter bed turnover
Anesthesia Machine	128.4	February	61.5	August	Resumption of elective surgeries post-recess
Multiparameter Monitor	128.5	August	72.4	February	Intensive inpatient monitoring turnover and transducer maintenance
Cardioverter	114.5	January	77.9	August	Periodic emergency readiness tests and annual inspection routines

beyond the capabilities of traditional temporal decomposition methods, since the STL algorithm identifies volume peaks but does not distinguish the nature of the occurrences.

The descriptive fields of work orders record precisely this information in natural language. The proposal is to submit the WOs from the peak months identified by STL to a large language model (LLM), instructing it to classify and semantically group the failure descriptions. The result would be a qualitative profile of the peaks: which types of intervention predominate, whether there are recurring component or cause patterns, and whether the peak-period profile is structurally different from the profile of normal-demand months. Table IV organizes, for each equipment type, the concrete analytical question that this approach would make it possible to answer.

The methodological feasibility of the proposal is based on the use of routine historical records and the application of natural language processing procedures structured through systematic classification prompts. This approach does not require dedicated hardware and can be carried out by Clinical Engineering teams using standard computing resources. The resulting framework adds a semantic analysis layer capable of qualifying the temporal distribution obtained through statistical modeling.

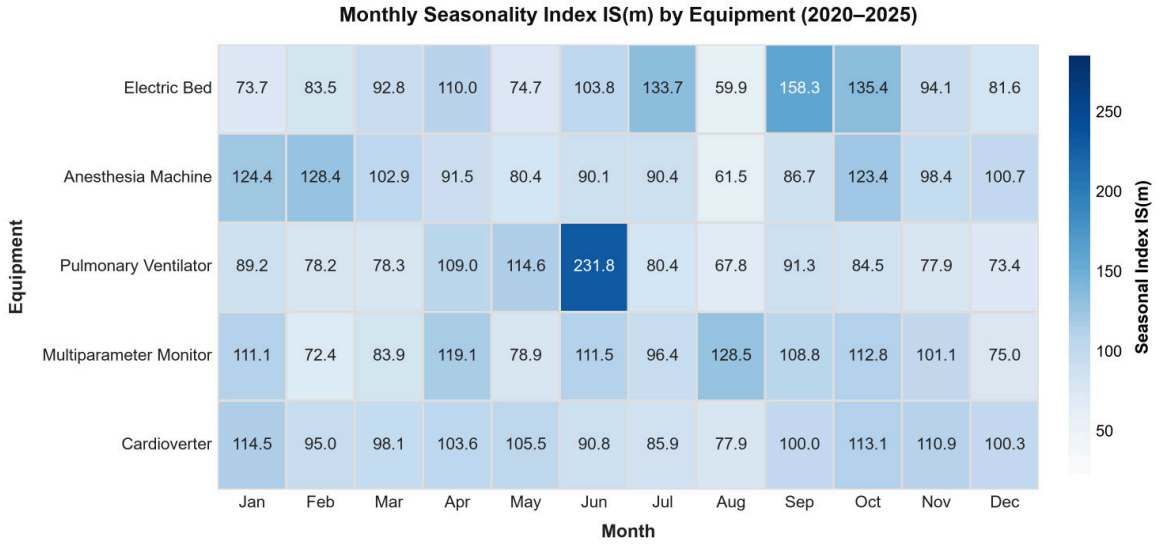


Figure 3. Monthly Seasonality Index (SI) by equipment type. SI = 100 represents the equipment's historical average demand. Values above 100 indicate a peak, whereas values below 100 indicate a trough. STL decomposition with annual periodicity (2020–2025).

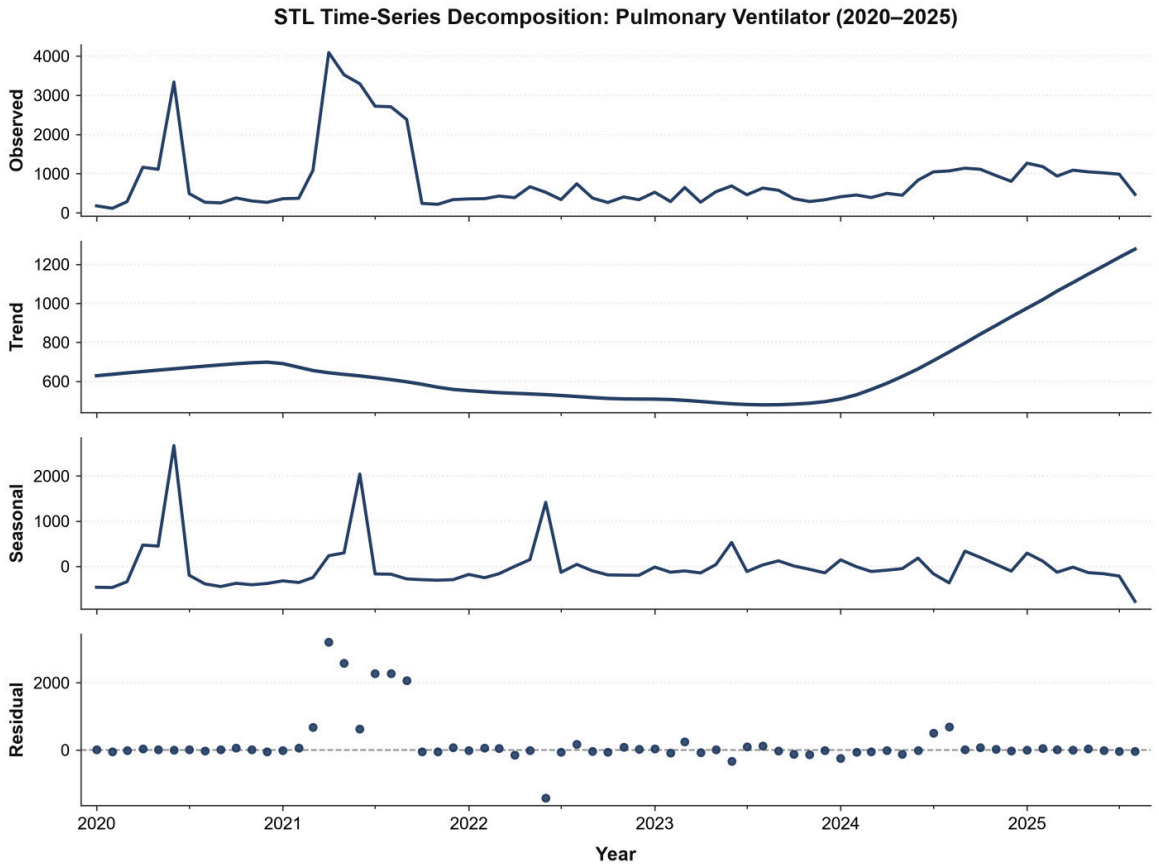


Figure 4. STL time-series decomposition for Pulmonary Ventilator (2020–2025): observed series, trend, seasonal component, and residual points. Recurring demand peak in June (SI = 231.8). Positive residual shocks in 2020–2021 capture acute pandemic demand surges.

IV. Discussion

A. Convergence between Perspectives: The Value of the Multi-Analytical Framework

Taken together, the findings from the two analytical perspectives implemented construct a more complete picture of

Table IV
FRAMEWORK OF ANALYTICAL QUESTIONS PROPOSED FOR APPLYING AN LLM TO WOS FROM SEASONAL PEAKS.

Equipment	Peak Month(s)	Max. SI	Analytical question proposed for the LLM
Pulmonary Ventilator	June	231.8	Which types of failure predominate in June WOs during the winter respiratory surge? Are they flow-sensor, valve, or pneumatic degradation failures?
Electric Bed	September	158.3	Is the September peak dominated by electromechanical failures (actuator, motor drive) or structural/control interface wear associated with inpatient turnover?
Anesthesia Machine	February	128.4	Does the February peak coincide with post-recess surgical resumption? Are occurrences concentrated in pneumatic circuits, vaporizers, or electronic calibration?
Multiparameter Monitor	August	128.5	Does the August peak reflect higher ICU monitoring turnover and accessory replacement (ECG leads, SpO_2 probes), or display/power board degradation?
Cardioverter	Not applicable	114.5	Does textual analysis of technical calls confirm that failures are predominantly detected during periodic testing and battery checks rather than acute emergency usage?

the operational behavior of the equipment than either would produce alone. Reliability analysis identifies the manufacturer as the main explanatory factor for differences in repair time, with a large effect for Electric Beds and substantial Hazard Ratios for Pulmonary Ventilators. This finding highlights the need to integrate historical reliability and technical performance indicators into the formal specification and evaluation criteria of medical-hospital equipment acquisition processes.

Practical experience in the hospital environment shows that, because of high technological complexity and multiple configuration possibilities, a significant proportion of technical calls for Pulmonary Ventilators and Anesthesia Machines is related to operational aspects and care adjustments, and therefore does not necessarily represent structural technical failures. Without a qualitative analysis layer, purely statistical indicators tend to treat all work orders homogeneously. The application of language models (LLMs) therefore appears to be a promising approach for refining these metrics, making it possible to distinguish intrinsic equipment failures from demands generated by operational support and configuration needs.

Seasonal analysis adds a second dimension that hospital maintenance planning commonly ignores: demand is not randomly distributed throughout the year. It rises and falls in regular patterns that reflect the rhythms of healthcare delivery itself. The SI of 231.8 in June for Ventilators and 158.3 in September for Electric Beds are not statistical noise. Rather, they are signals that could reduce the operational impact of peak periods if incorporated into resource and inventory planning.

Integrating LLMs into the descriptive analysis of technical calls complements traditional numerical approaches. While STL temporal analysis establishes the timing and intensity of demand variations, semantic processing can distinguish whether intervention peaks result from structural failures in critical components, such as flow sensors and seals, or from operational demands. This information may contribute to the transition from a predominantly reactive maintenance model to evidence-based planning of care contingencies grounded in qualitative historical evidence.

B. Implications for Clinical Engineering and Health Technology Assessment

One aspect worth emphasizing is that none of the results obtained contradicts what experienced Clinical Engineering teams already perceive empirically in their daily work. What the analysis does is make this perception quantifiable, comparable, and communicable beyond the technical team. This has specific value in the context of HTA: decisions on technology incorporation and replacement require evidence that can be justified and documented, not merely perceived.

The Cardioverter showed no relevant seasonal variation throughout the analyzed year. This statistical behavior is consistent with its clinical use characteristics: these devices have a low frequency of effective clinical activation and remain in a state of readiness for long periods. A substantial proportion of occurrences does not result from continuous use but is frequently identified during operational tests and routine inspections, which decouples maintenance demand from seasonal fluctuations in occupancy and hospitalization.

From the standpoint of replicability, the proposed framework does not depend on analytical infrastructure beyond the reach of hospital teams. The complete methodological sequence, ranging from multi-criteria equipment selection to the proposed qualitative interpretation by LLM, can be applied to any CMMS database with sufficient historical volume. Reliability analysis requires Python with the *lifelines* and *scipy* libraries. STL decomposition is equally accessible, while the LLM stage can be conducted using open-access tools. The barrier to dissemination is not technical complexity, but the absence of an analytical culture applied to healthcare asset management.

V. CONCLUSION

This study demonstrated that historical hospital maintenance records, routinely stored in CMMS systems, constitute a robust database for supporting managerial decisions in healthcare

asset management. The integration of multiple analytical approaches applied to the same dataset made it possible to extract performance indicators that enrich decision-making processes in Clinical Engineering and Health Technology Assessment (HTA).

Statistical reliability modeling demonstrated that the manufacturer acts as a predictive factor for repair time and operational survival with significantly greater weight than cumulative equipment usage (Age Proxy). For technologies such as Electric Beds, acquisition decisions based predominantly on initial cost, while disregarding operational performance history, may generate prolonged negative impacts on care availability because of increased mean repair time and the complexity of obtaining spare parts.

Additionally, time-series modeling revealed structured seasonal cycles consistent with the specific care dynamics of each technology. Equipment linked to elective surgical procedures, such as Anesthesia Machines, exhibits fluctuations in maintenance demand conditioned by the hospital calendar. In contrast, technologies maintained in a state of care readiness, such as cardioverters, have intervention demands that are less associated with occupancy rates, with some occurrences being identified during routine inspections and periodic operational tests.

To refine the interpretation of numerical indicators, the framework proposes incorporating large language models as a semantic layer for the textual classification of work orders. This natural language processing stage may help distinguish technical failures intrinsic to the equipment from calls resulting from operational support or configuration errors, qualifying the interpretation of reliability metrics.

In summary, the proposed methodological integration of reliability analysis, seasonality modeling, and textual processing by language models constitutes an analytical framework capable of transforming historical maintenance records into evidence to support Clinical Engineering and Health Technology Assessment. This approach may strengthen decisions related to technology incorporation, investment planning, equipment replacement, risk management, and technology governance. As future perspectives, empirical validation of the framework in different care profiles and CMMS platforms is recommended, as well as its integration into intelligent decision-support systems aimed at managing the life cycle of health technologies.

REFERENCES

[1] E. Iadanza, V. Gonnelli, F. Satta, and M. Gherardelli, "Evidence-based medical equipment management: a convenient implementation," *Medical & Biological Engineering & Computing*, vol. 57, no. 10, pp. 2215–2230, 2019.

[2] Association for the Advancement of Medical Instrumentation, Ed., *Recommended Practice for a Medical Equipment Management Program*. Arlington, VA, USA: Association for the Advancement of Medical Instrumentation, 2013.

[3] A. H. Zamzam, K. Hasikin, and A. K. A. Wahab, "Integrated failure analysis using machine learning predictive system for smart management of medical equipment maintenance," *Engineering Applications of Artificial Intelligence*, vol. 125, Art. no. 106715, 2023.

[4] R. Y. Zisk-Rony, C. Weissman, and Y. G. Weiss, "Mechanical ventilation patterns and trends over 20 years in an Israeli hospital system: policy ramifications," *Israel Journal of Health Policy Research*, vol. 8, no. 1, Art. no. 20, 2019.

[5] M. Ortiz-Barrios, A. Petrillo, S. Arias-Fonseca, S. McClean, F. de Felice, C. Nugent, and S. A. Uribe-López, "An AI-based multiphase framework for improving the mechanical ventilation availability in emergency departments during respiratory disease seasons: a case study," *International Journal of Emergency Medicine*, vol. 17, no. 1, Art. no. 45, 2024.

[6] D. R. M. Florentino, E. L. Dos Santos Cabral, W. R. Silva Castro, J. F. da Costa Junior, J. G. de Torres Neto, F. C. de Oliveira Mendes, H. R. Hekis, A. C. Meneses Rêgo, and I. Araújo Filho, "Analysis of shared services center deployment at a university hospital (CENTROMED) in Natal, Brazil: a study on availability and capacity," *Journal of Multidisciplinary Healthcare*, vol. 13, pp. 1717–1728, 2020, doi: 10.2147/JMDH.S260107.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[8] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[9] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, rev. ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1977.

[10] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[11] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[12] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: a seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990.

Implementation of Optimization Methods to Determine the Lowest Cost of a Hospital Electric Bill

1st João Marcus Ferreira Braga da Costa Lima
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
0009-0005-3743-9687

2nd Jorge Cormane
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
0000-0001-5263-0985

3rd Manuel Nascimento Dias Barcelos Junior
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
manuelbarcelos@unb.br

4th Bruno Luiz Pereira
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
bruno.pereira@unb.br

5th Lorenzo de Lima Alves dos Santos
Graduate Program in Biomedical Engineering
University of Brasilia
Brasilia, Brazil
lorenzodavinci@hotmail.com

6th André Bisinoto Matias
Infrastructure dept.
HU Brasil
Brasilia, Brazil
andre.matias@hubrasil.gov.br

***Abstract*—Electricity costs represent a significant share of the operating budget of large Brazilian university hospitals, particularly those operating under the Free Contracting Environment (*Ambiente de Contratação Livre*, ACL) and the blue hourly tariff structure. This work proposes an optimization methodology for the contracted demand and energy volume of the University Hospital of Brasília (HUB-UnB), comparing three approaches: Sequential Quadratic Programming (SQP), Differential Evolution (DE), and a hybrid DE + SQP strategy. Monthly demand and energy consumption data were used to forecast the 2026 load profile through a Holt-Winters model, while the Settlement Price of Differences (PLD) was forecast with a Prophet model; both were validated against the first four months of 2026 with errors below 5 %. Applied to the forecasted profile, the SQP method alone reduced the estimated annual electricity cost by 3.76 %, the DE method achieved a 6.16 % reduction, and the hybrid DE + SQP approach achieved the largest reduction, 6.19 %, relative to the current contract. The results indicate that the hybrid approach consistently outperforms both individual methods and is particularly well-suited to problems in which the objective function exhibits derivative discontinuities arising from the demand-tolerance tariff rule.**

***Index Terms*—university hospital, free electricity market, contracted demand, optimization, differential evolution, sequential quadratic programming, hybrid optimization**

I. Introduction

The electricity bill of the University Hospital of Brasília, operated by HU Brasil (HUB-UnB/HU Brasil) ranks among the institution's largest fixed operating expenditures, a characteristic shared by large federal university hospitals throughout Brazil, where the continuous operation of critical equipment, HVAC systems, and lighting results in high and largely inflexible load profiles. As a Group A4 consumer enabled to operate under the Free Contracting Environment (ACL), the HUB-UnB/HU Brasil negotiates energy prices bilaterally with traders but still pays the local distribution utility, Neoenergia Brasília, for network access through the Distribution System Use Tariff (TUSD). Under this arrangement, setting the contracted demand levels for peak and off-peak periods, as well as the energy volume to be acquired in the ACL, carries direct financial consequences that go beyond the simple per-kWh cost.

The blue hourly tariff structure penalizes deviations from the contracted demand asymmetrically: exceedances above 5 % of the contracted value are billed at twice the demand tariff rate, whereas contracted values above actual consumption generate a guaranteed-minimum charge without offsetting compensation. On the energy side, the difference between the volume contracted in the ACL and the actual consumption is settled at the Settlement Price of Differences (PLD) at the Energy Trading Chamber (CCEE), a price that has changed during the year with hydropower reservoir levels. Forecasting future demand and consumption and, from those projections, determining the contractual parameters that minimize total cost therefore constitutes an optimization problem with concrete,

PPGEB – Postgraduate Program in Biomedical Engineering and HUB-UnB/HU Brasil

recurrent economic impact in the hospital.

The literature addresses this problem through distinct approaches. Linear and quadratic programming methods have been applied to energy contract optimization for industrial consumers [8], [9], yielding good results in settings where the cost function is convex and continuous. When nonlinearity or discrete variables complicate analytical approaches, metaheuristics such as genetic algorithms [10] and Particle Swarm Optimization (PSO) [11] have been used successfully in dispatch and energy procurement problems. These works, however, focus on industrial consumers or distribution networks with distributed generation: the Brazilian hospital context, with its ACL-specific regulatory constraints and a load profile driven by medical equipment, remains underexplored in the energy optimization literature.

The works most closely related to this paper address hospital energy efficiency from the perspective of energy audits and retrofit [12], [13], without addressing the contractual layer. The distinction is relevant: efficiency improvements reduce consumption but not necessarily cost, if the contractual parameters remain poorly sized. A hospital that reduces peak demand by 10 % but keeps the contracted demand at the previous level continues to pay for volume it does not use.

This paper implements and compares three optimization methods for the demand and energy procurement problem at the HUB-UnB/HU Brasil: SQP (Sequential Quadratic Programming) [1]; Differential Evolution (DE) [2], a global-search evolutionary algorithm; and a hybrid approach that uses DE to explore the solution space in the early stages and SQP to refine the best solution found. The motivation for this comparison lies in the fact that the cost function presents derivative discontinuities at the demand-tolerance thresholds, which may cause SQP to converge to local minima depending on the starting point, while DE navigates these regions without gradient bias, at the expense of a larger number of objective function evaluations. The hybrid strategy seeks to capture the advantages of both.

A distinguishing feature of this work relative to similar studies is the empirical data foundation: demand and consumption variables were collected from the hospital's electricity bills since the institution first became a free-market consumer. The historical dataset covers the period from March 2021 to December 2025, and the 2026 projections feeding the optimization module were obtained from time-series models validated against the already-available months of 2026 [3].

The paper is organized as follows. Section II presents the theoretical background on the tariff structure and the three optimization methods. Section III describes the hospital characterization, the forecasting procedure, and the method configurations. Section IV presents the results and comparison across approaches. Section V draws conclusions and outlines directions for future work.

II. Theoretical Background

A. Tariff Structure for Free-Market A4 Consumers — Blue Hourly Tariff

Under the blue hourly tariff, demand is billed separately for the peak period (18:00 – 21:00, business days) and the off-peak period, at distinct rates for each. The monthly demand charge follows a piecewise function: up to 5 % above the contracted demand D_c, billing equals the contracted value itself; beyond that threshold, the excess is charged at twice the tariff rate, as established by ANEEL Normative Resolution No. 1,000/2021 [4]:

$$f(\tilde{D}, D_c, \tau) = \begin{cases} D_c \cdot \tau, & \tilde{D} \leq 1{,}05\, D_c \\ D_c \cdot \tau + (\tilde{D} - D_c) \cdot 2\tau, & \tilde{D} > 1{,}05\, D_c \end{cases} \tag{1}$$

where $\tilde{D}$ is the maximum demand measured in the period, integrated over 15-minute intervals, and τ is the TUSD demand tariff for that period (R$/kW). The total monthly cost for a free-market consumer decomposes into three components:

$$C_m = C_{dem,m} + C_{tusd,m} + C_{acl,m} \tag{2}$$

The demand TUSD component applies function f from (1) to each period:

$$C_{dem,m} = f(\tilde{D}_p, D_p, \tau_{D_p}) + f(\tilde{D}_{fp}, D_{fp}, \tau_{D_{fp}}) \tag{3}$$

The consumption TUSD component remunerates network usage for the energy actually withdrawn from the distribution system:

$$C_{tusd,m} = \tilde{E}_p\, \tau_{E_p} + \tilde{E}_{fp}\, \tau_{E_{fp}} \tag{4}$$

The energy component combines the cost of the bilateral ACL contract with the settlement of the difference between actual consumption and contracted volume at the PLD:

$$C_{acl,m} = E_{acl}\, \rho_{acl} + (\tilde{E}_p + \tilde{E}_{fp} - E_{acl})\, \rho_{pld} \tag{5}$$

The third term in (5) may be negative when actual consumption falls short of the contracted volume, in which case the difference is settled as a credit at the PLD.

B. Optimization Problem Formulation

The objective is to find the decision vector $\mathbf{x} = (D_p,\, D_{fp},\, E_{acl})$ that minimizes the total annual cost:

$$\min_{\mathbf{x}}\ C_{annual}(\mathbf{x}) = \sum_{m=1}^{12} C_m(\mathbf{x}) \tag{6}$$

subject to the box constraints:

$$D_p \geq D_{min}, \quad D_{fp} \geq D_{min}, \quad E_{acl} \geq 0 \tag{7}$$

where D_{min} is the minimum contractable demand. The net demand effectively measured by the distribution utility in each

period corresponds to the larger of the minimum contractable demand and the forecasted demand:

$$\tilde{D}_{p,m} = \max\Big(D_{min},\ \hat{D}_{p,m}\Big) \tag{8}$$

$$\tilde{D}_{fp,m} = \max\Big(D_{min},\ \hat{D}_{fp,m}\Big) \tag{9}$$

where $\hat{D}_{p,m}$ and $\hat{D}_{fp,m}$ are the forecasted demands for month m. The function C_{annual} in (6) is piecewise continuous, with derivative discontinuities at the points $\tilde{D} = 1{,}05\, D_c$. This feature motivates the comparison between gradient-based and metaheuristic methods proposed in this work.

C. Sequential Quadratic Programming (SQP)

SQP is one of the most consolidated classical methods for nonlinear constrained optimization, with theoretical development dating back to the 1970s and extensive application to engineering problems in subsequent decades [1], [5]. Its central idea is to approximate the original problem by a sequence of quadratic subproblems, each solved analytically to produce a descent direction.

At each iteration k, SQP solves:

$$\min_{\mathbf{d}} \ \frac{1}{2}\mathbf{d}^\top W_k \mathbf{d} + \nabla C(\mathbf{x}^{(k)})^\top \mathbf{d} \tag{10}$$

where W_k is an approximation of the Lagrangian Hessian and $\mathbf{d}$ is the search direction. The solution $\mathbf{d}^*$ of (10) defines the step:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\, \mathbf{d}^* \tag{11}$$

where α_k is determined by a merit-function line search that penalizes constraint violations. The box constraints from (7) are incorporated directly into the quadratic subproblem, with no need for external projection. The implementation used here is SLSQP (Sequential Least Squares Programming), available in the SciPy library [6], which handles equality, inequality, and bound constraints in a unified framework.

SQP is a local search method: solution quality depends on the starting point. In functions with multiple local minima — such as the cost function in (6), which exhibits derivative discontinuities at the demand-tolerance thresholds — this dependence motivates the hybrid strategy with DE described in Section II-E.

D. Differential Evolution

Differential Evolution (DE) is an evolutionary algorithm for continuous optimization proposed by Storn and Price [2]. It maintains a population of N_p vectors $\{\mathbf{x}_i^g\}$ and evolves it through three operations. In the *mutation* step, a mutant vector is generated from the current best individual $\mathbf{x}_{best}^g$ and two randomly selected individuals:

$$\mathbf{v}_i^g = \mathbf{x}_{best}^g + F \cdot (\mathbf{x}_{r1}^g - \mathbf{x}_{r2}^g) \tag{12}$$

where $F \in [0, 2]$ is the scale factor and $r1 \neq r2 \neq i$. In the *binomial crossover*, each component of the trial vector $\mathbf{u}_i^g$ is selected from $\mathbf{v}_i^g$ with probability CR or from $\mathbf{x}_i^g$ otherwise, guaranteeing that at least one component comes from $\mathbf{v}_i^g$. In the *selection* step, the trial vector replaces the current one only if it reduces the cost: $\mathbf{x}_i^{g+1} = \mathbf{u}_i^g$ if $C(\mathbf{u}_i^g) \leq C(\mathbf{x}_i^g)$, and $\mathbf{x}_i^{g+1} = \mathbf{x}_i^g$ otherwise. The *best/1/bin* strategy adopted here balances global exploration and local convergence by directing mutations toward the current best solution.

E. Hybrid DE + SQP Approach

The complementarity between the two methods directly motivates the proposed hybrid strategy. DE performs global exploration without requiring continuity of the objective function, identifying the region of the solution space where the global optimum lies. SQP then exploits the local curvature of that region to converge rapidly to the optimal point.

In the adopted strategy, DE is run to convergence, and the resulting best individual $\mathbf{x}_{best}^{g_{trans}}$ is passed as the starting point to SQP, which carries out local refinement under the box constraints of (7). This transition reduces the total number of objective function evaluations relative to running DE alone to the same precision level, while preserving the ability to escape local minima that SQP alone does not possess.

III. Materials and Methods

A. HUB-UnB Characterization

The HUB-UnB is a large federal university hospital operated by HU Brasil (*Hospitais Universitários do Brasil*), located in Brasília (DF) and supplied at medium voltage under distribution utility Neoenergia Brasília's Group A4 tariff. Its load profile is dominated by HVAC systems and continuously operating medical equipment. The institution has been operating in the ACL since 2021.

The historical dataset covers 57 months of consolidated monthly electricity bills (March 2021 to December 2025), comprising four variables: peak maximum demand, off-peak maximum demand, represented on Fig. 1, and peak energy consumption, and off-peak energy consumption, represented on 2.

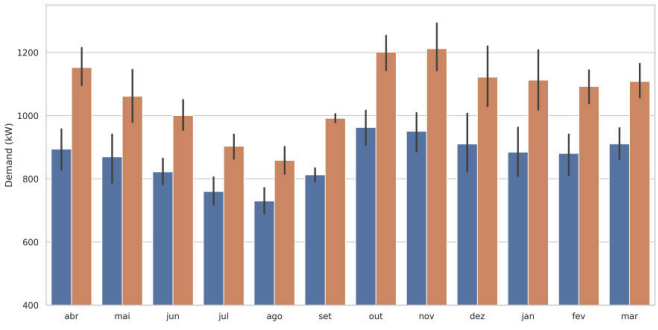


Fig. 1. Historical peak maximum demand (blue) and off-peak maximum demand (orange) colected from the electricity bills

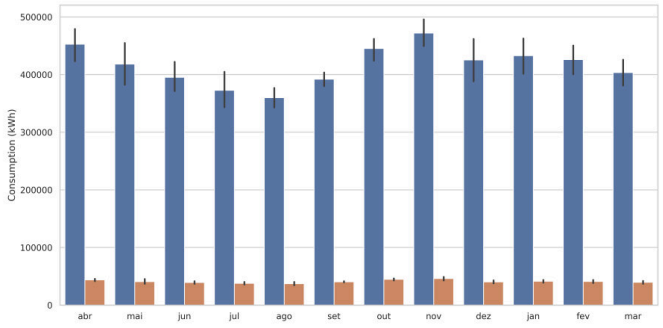


Fig. 2. Historical peak energy consumption (blue) and off-peak energy consumption (orange) colected from the electricity bills

B. Demand and Energy Forecasting for 2026

Three time-series models were fitted to each variable using data through December 2025, with validation on the four months of 2026 already available: Holt-Winters triple exponential smoothing (automatically testing additive and multiplicative combinations of trend and seasonality), SARIMA$(1,1,1)(1,1,1)_{12}$, and a Seasonal Naïve model as a baseline. The evaluation metric was the Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100 \tag{13}$$

where y_t is the observed value and $\hat{y}_t$ the forecasted value for month t. The model selected to feed the optimization module was the one with the lowest validation error over the first four months of 2026: Holt-Winters, with a MAPE of 3.2 % for peak demand and 4.42 % for off-peak demand [3].

C. PLD Forecasting for 2026

Estimating the CCEE settlement component described in (5) required projecting the PLD for all twelve months of 2026. The Prophet model, developed by Meta, was used; it natively handles multiple seasonalities and public holidays in the training process. The monthly PLD forecasts were validated against the four known months of 2026 using the historical PLD series published by the CCEE [7] from january 2002 until december 2025, achieving errors below 2 %.

D. Optimization Method Parameters

All three methods were implemented in Python 3.12 with the following configurations.

SQP (SLSQP): SciPy implementation [6], convergence tolerance *ftol* $= 10^{-12}$, *maxiter* $= 20\,000$. The starting point was set to the currently contracted values for peak demand, off-peak demand, and ACL energy volume.

DE: *best/1/bin* strategy, population size $N_p = 500$, scale factor $F = 0{,}8$, crossover probability $CR = 0{,}9$, maximum of *maxiter* $= 20\,000$ generations or relative variation tolerance $< 10^{-12}$ between consecutive generations. The search bounds are the same as those of SQP, as defined in (7).

Hybrid DE + SQP: DE is run to convergence with population size $N_p = 50$ and the remaining parameters as described above. The best resulting individual is then passed as the starting point to SQP with the same parameters.

The tariff parameters used in the objective function are listed in Table I.

TABLE I
TARIFF PARAMETERS USED IN THE OBJECTIVE FUNCTION

Parameter	Value	Unit
TUSD peak demand (τ_{D_p})	40.41	R$/kW
TUSD off-peak demand ($\tau_{D_{fp}}$)	16.92	R$/kW
TUSD peak consumption (τ_{E_p})	0.21	R$/kWh
TUSD off-peak consumption ($\tau_{E_{fp}}$)	0.21	R$/kWh
ACL price (ρ_{acl})	0.19937	R$/kWh
Demand tolerance	5	%
Exceedance penalty	2×	—

E. Comparison Criterion

The three methods were compared based on the lowest final cost found. Given that convergence, speed, and uncertainties are not relevant for this application, and since contractual adjustments can be made every two years but energy accounting is monthly, solution precision is more relevant for this problem than computational throughput.

IV. Results and Discussion

To assess consistency, mean, variance, and standard deviation were computed across 30 complete independent runs of each method. It is worth noting that SQP exhibits zero variance and zero standard deviation, since it is a deterministic algorithm: given the same starting point, it always returns the same solution, which, however, may not be the global optimum, as shown in Table II.

TABLE II
STATISTICAL RESULTS OF ANNUAL COST ACROSS 30 RUNS PER ALGORITHM

Method	Mean (R$)	Variance	Std. Dev. (R$)
SQP	3,050,917.23	0	0
DE	2,975,689.57	633,215.74	795.75
DE + SQP	2,975,548.29	1,330,317.05	1,153.39

Fig.3 summarizes the best results achieved by each method according to the criterion defined in Section III-E. The SQP method alone reduced the estimated annual electricity cost by 3.76 % (R$ 119,130.77). The DE method achieved a reduction of 6.16 % (R$ 195,383.48), and the hybrid DE + SQP approach yielded the best result among the three, with a 6.19 % reduction (R$ 196,166.66).

SQP reduces the cost rapidly in the first iterations but tends to stall at a local optimum that depends on the starting point. DE converges gradually and consistently, exploring the solution space more broadly in the early generations. The hybrid combines DE's initial exploration with SQP's accelerated convergence in the final phase, reaching the lowest

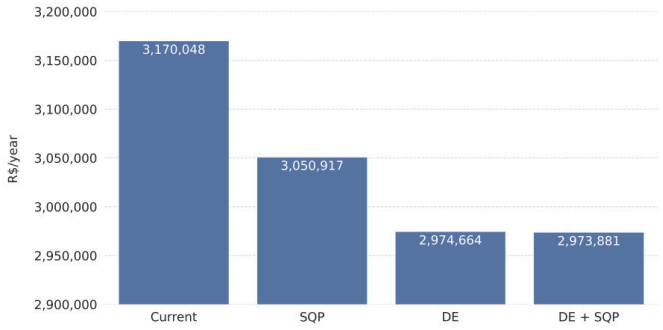


Fig. 3. Estimated annual electricity bill cost (best results)

cost with fewer total objective function evaluations than DE run to full convergence.

A. Optimal Proposal and Economic Analysis

The optimal parameter values obtained by the hybrid method are shown in Table III, alongside the current contract.

TABLE III
COMPARISON BETWEEN CURRENT CONTRACT AND OPTIMAL PROPOSAL

Parameter	Current	Proposal	Change
Peak Demand (kW)	1,090	840	−22.94 %
Off-Peak Demand (kW)	1,280	1,053	−17.73 %
ACL Energy (kWh/month)	494,503	727,870	+47.17 %
Annual Cost (R$)	3,170,048	2,973,881	−6.19 %

The estimated annual saving of R$ 196,166.66 arises primarily from the adjustment in the ACL contracted energy volume. Because the bilateral ACL price remained below the projected PLD for most of the year, contracting an energy volume above the average forecasted consumption proved advantageous: the surplus is settled in the short-term market (MCP) at the PLD, which remunerates that difference at a value above the ACL acquisition cost, as expressed by the second term of (5). Correctly sizing this contracted volume simultaneously avoids the extra cost of an undersized contract and the generation of surpluses that find no real counterpart in the market.

B. Discussion

The results show that the energy contract optimization problem for hospitals is not unimodal: SQP found distinct solutions depending on initialization. This behavior stems from the derivative discontinuity in the cost function at the demand-tolerance threshold, which can cause the gradient to point in misleading directions near those regions.

DE navigates this landscape without the bias of first-order methods, but does so at the expense of a larger number of evaluations. For a problem of this dimensionality, three decision variables, that computational cost is manageable, yet the hybrid approach showed that it is possible to obtain the best of both methods without incurring the full computational burden of either in isolation.

From a hospital management standpoint, the proposed methodology offers a direct practical contribution: executed on a biannual basis with up-to-date data, it produces a contract recommendation grounded in real measurements and calibrated forecasts, replacing the common practice of renewing contracts based solely on historical peak values or managerial experience.

Furthermore, the idea is to expand this methodology to other hospitals within the same network that share similar characteristics.

V. CONCLUSION

This paper compared three optimization methods (SQP, DE, and a hybrid DE + SQP strategy) applied to the demand and energy procurement problem of the HUB-UnB/HU Brasil in the Free Contracting Environment. Using real data extracted from electricity bills and 2026 forecasts of demand, energy, and PLD, the hybrid method achieved the largest estimated saving among the three approaches: 6.19 % (R$ 196,166.66) relative to the current contract. This result demonstrates that combining global search with local refinement is advantageous for this class of problems, whose cost function presents derivative discontinuities.

Key limitations of the study relate to the forecasting models for the variables used in the cost function. Both the demand-and-consumption forecasts and the PLD projections carry uncertainty inherent to the models employed, and that uncertainty propagates directly to the quality of the optimal solution found.

Future work will focus on validating the optimized parameters against the actual data as they become available, refining the peak and off-peak demand forecasting models, incorporating additional explanatory variables into the PLD forecasting model, and integrating a real-time SCADA monitoring system into the HUB-UnB/HU Brasil, enabling more granular consumption analysis and supporting contract sizing with direct measurement data.

Finally, operational considerations such as changes in hospital occupancy, installation of distributed generation, demand response strategies, or unexpected variations in electricity market prices are not incorporated into the optimization model since the primary objective is the correction of the electricity bill.

The contribution is a decision-support formulation tailored to a hospital in the free contracting environment, not proof of global superiority between those methods.

ACKNOWLEDGMENT

The authors thank the Graduate Program in Biomedical Engineering, the National Council for Scientific and Technological Development (CNPq) and the University Hospital of Brasília (HUB-UnB/HU Brasil) for their collaboration and institutional support, which made this study possible.

REFERENCES

[1] P. T. Boggs and J. W. Tolle, “Sequential quadratic programming,” *Acta Numerica*, vol. 4, pp. 1–51, 1995.

[2] R. Storn and K. Price, “Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces,” *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.
[3] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” *PLOS ONE*, vol. 13, no. 3, p. e0194889, 2018.
[4] ANEEL, *Resolução Normativa nº 1.000, de 7 de dezembro de 2021*. Brasília: Agência Nacional de Energia Elétrica, 2021. [Online]. Available: https://www.aneel.gov.br
[5] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.
[6] P. Virtanen et al., “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020.
[7] CCEE, *Relatório de Preços - PLD Histórico*. São Paulo: Câmara de Comercialização de Energia Elétrica, 2025. [Online]. Available: https://www.ccee.org.br
[8] F. Marangoni, L. Magatão, and L. V. R. de Arruda, “Demand response optimization model to energy and power expenses analysis and contract revision,” *Energies*, vol. 13, no. 11, p. 2803, 2020.
[9] A. Soroudi and T. Amraee, “Decision making under uncertainty in energy systems: State of the art,” *Renewable and Sustainable Energy Reviews*, vol. 28, pp. 376–384, 2013.
[10] M. Marzband, M. Javadi, J. L. Domínguez-García, and M. Moghaddam, “Non-cooperative game theory based energy management systems for energy district in the retail market considering DER uncertainties,” *IET Generation, Transmission & Distribution*, vol. 10, no. 12, pp. 2999–3009, 2016.
[11] A. G. Bhesdadiya, I. N. Trivedi, P. Jangir, N. Jangir, and A. Kumar, “An NSGA-III algorithm for solving multi-objective economic/environmental dispatch problem,” *Cogent Engineering*, vol. 3, no. 1, p. 1269383, 2016.
[12] R. F. De Masi, N. Del Regno, A. Gigante, S. Ruggiero, A. Russo, F. Tariello, and G. P. Vanoli, “The importance of investing in the energy refurbishment of hospitals: Results of a case study in a Mediterranean climate,” *Sustainability*, vol. 15, no. 14, p. 11450, 2023.
[13] A. Aranda, G. Ferreira, M. D. Mainar-Toledo, S. Scarpellini, and E. Llera Sastresa, “Multiple regression models to predict the annual energy consumption in the Spanish banking sector,” *Energy and Buildings*, vol. 49, pp. 380–387, 2012.

Post-Market Performance Evaluation of an Infusion Pump in a Pediatric Intensive Care Unit: A Case Study

Gilberto B. Araujo-Filho*, Fedra C. Mendez†, Paulo M. Zanuzzio‡, Edson L. Tenca‡, Fabiana Chagas‡, Henrique T. Moriya*, and Diego A. O. Rosa*

*Biomedical Engineering Laboratory (LEB), University of São Paulo (USP), São Paulo, Brazil
ORCID: 0000-0002-1770-3840 (G.B.A.F.), 0000-0002-1595-1779 (H.T.M.), 0009-0004-2422-8453 (D.A.O.R.)

†School of Biomedical Engineering, National University of San Marcos (UNMSM), Lima, Peru
ORCID: 0009-0001-8174-2262 (F.C.M.)

‡University Hospital (HU), University of São Paulo (USP), São Paulo, Brazil
ORCID: 0009-0000-6740-5166 (P.M.Z.), 0009-0007-7220-5565 (E.L.T.), 0009-0008-0499-4061 (F.C.)

***Abstract*—This study presents tests performed on infusion pumps used in clinical practice at the University Hospital of the University of São Paulo. The evaluation was carried out through a partnership between the hospital and the Normative Testing Division of the Biomedical Engineering Laboratory at the Escola Politécnica of the University of São Paulo. To evaluate device performance, gravimetric tests were performed in accordance with IEC 60601-2-24. A representative sentinel case extracted from an ongoing surveillance initiative is presented, selected for significant flow rate deviations observed during performance testing. The evaluated infusion pump, manufactured in 2023, had recently been received by the hospital. Gravimetric tests were performed in accordance with IEC 60601-2-24 to assess flow-rate performance under predefined operating conditions. In the initial assessment, the device exhibited relevant deviations from the programmed flow rate, despite having previously been considered suitable for clinical use. Following the initial assessment, the device was segregated and sent to the manufacturer. The manufacturer did not provide information regarding any maintenance or calibration procedures performed before returning the device to the hospital. A second performance assessment was performed and showed improved flow-rate performance under the evaluated conditions. This case illustrates the importance of periodic testing as a complementary strategy for verifying the effectiveness of manufacturer intervention, to detect metrological issues before they present potential clinical risks, supporting the implementation of broader hospital sampling programs.**

***Index Terms*—infusion pump, flow-rate accuracy, performance testing, post-market surveillance, IEC 60601-2-24.**

I. Introduction

In a medical hospital setting, infusion pumps are routinely used in clinical care. These devices are widely employed to ensure the safe and accurate administration of medications, nutrients, and intravenous fluids, especially in critical care areas such as Intensive Care Units (ICUs) [1]. According to the International Electrotechnical Commission (IEC) 60601-2-24:2012 standard, an infusion pump is medical equipment intended for the controlled delivery of liquids to the patient through the pressure generated by its pump [2]. These devices are classified into several categories based on their operating mechanism, including peristaltic (rotary or linear), piston, diaphragm, and syringe [3].

Infusion pumps are designed to guarantee accurate dosing and optimal device performance; however, it is essential that healthcare professionals use these devices effectively [4] and that regular maintenance and calibration protocols are strictly followed to ensure patient safety [5].

The clinical impact of flow errors is particularly significant in critically ill patients, since small variations in infusion rate may alter the therapeutic concentration of highly sensitive drugs, compromising hemodynamic stability and patient safety [1].

In pediatric intensive care units (PICUs), infusion pumps are frequently used to continuously administer antimicrobials, vasoactive drugs, sedatives, and analgesics. According to the Surviving Sepsis Campaign guidelines, antimicrobial dosing strategies should be optimized based on pharmacokinetic and pharmacodynamic principles, since subtherapeutic dosing may lead to failure to clear infection, prolonged organ dysfunction, and the development of antimicrobial resistance. Furthermore, critically ill pediatric patients often present altered drug metabolism and clearance, making accurate drug delivery essential for achieving the intended therapeutic effect. Therefore, the performance and accuracy of infusion pumps play a crucial role in ensuring safe and effective treatment in pediatric critical care settings [6].

In this scenario, preventive maintenance, calibration, and post-market metrological evaluation play an important role in verifying whether the device continues to operate within the limits established by international standards. The importance of these activities is reinforced by evidence from the literature. A systematic review conducted by Atanda *et al.* (2023) reported that error rates in intravenous medication preparation and administration range from 13% to 84%, highlighting that device accuracy varies drastically based on both physiological and mechanical factors. The authors also highlighted a critical shortage of clinical tests directly correlating flow rate

accuracy with clinical patient outcomes [7]. In addition, other studies have reported that a significant proportion of infusion pumps used in hospitals present flow deviations associated with component wear, insufficient maintenance, or undetected functional failures [8].

Within this context, this study presents a representative sentinel case extracted from an ongoing investigation involving infusion pumps used in clinical practice. The selected device was chosen due to the magnitude of the deviations identified during performance testing carried out by the Normative Testing Division of the Biomedical Engineering Laboratory at the Escola Politécnica of the University of São Paulo (DENLEB/EPUSP).

The analyzed infusion pump belongs to the PICU at the University Hospital of the University of São Paulo (HU-USP). Before testing, it had been classified as operational and suitable for use, as stipulated by the contractual terms of the hospital's loan agreement.

By presenting test results and an analysis of medical device performance, this study aims to demonstrate that routine maintenance verification alone may be insufficient to ensure real-world flow accuracy. Ultimately, this work provides evidence to motivate structured, hospital-wide sampling strategies and refine evidence-based clinical engineering procedures.

II. Materials and Methods

The experimental tests were performed in the DENLEB. The laboratory is equipped with an analytical balance (BP221S, Sartorius, Germany); an 18G needle (Descarpack, Brazil); a $125\,\mathrm{mL}$ Erlenmeyer flask (Vidrolabor, Brazil); a vacuum gauge (63 mm, Wika, Germany); a pressure gauge (IDP-2000, Connect JCO, Brazil); an air compressor (Schulz S.A., Brazil); a vacuum pump (Schulz S.A., Brazil); and an environmental conditions meter (Davis Instruments, USA). All measuring instruments complied with the specifications established by IEC 60601-2-24:2012 and were duly calibrated with traceability to recognized metrological standards.

The tests were conducted at $(23 \pm 2)\,^{\circ}\mathrm{C}$; relative humidity was maintained at $(60 \pm 15)\,\%$, and water for medical use was utilized as the fluid throughout all experiments. Specifically, the procedures were carried out under items 201.12.1.101 and 201.12.1.102 of IEC 60601-2-24:2012, entitled Accuracy of Controls and Instruments, and followed the test configuration illustrated in Fig. 1.

The analytical balance was interfaced with a computer via RS-232 communication for real-time data acquisition. The system was operated by custom software developed in a graphical programming environment (LabVIEW 2012, National Instruments, USA), which guided the step-by-step execution of the tests and performed automated data processing.

The medical device, a volumetric infusion pump with a linear peristaltic mechanism manufactured in 2023 and its dedicated infusion sets, both newly acquired by HU-USP, were supplied by the PICU. To preserve institutional and commercial confidentiality, the exact manufacturer and model are kept undisclosed. According to the manufacturer's documentation, the nominal overall mean percentage flow error is $\pm 5\%$.

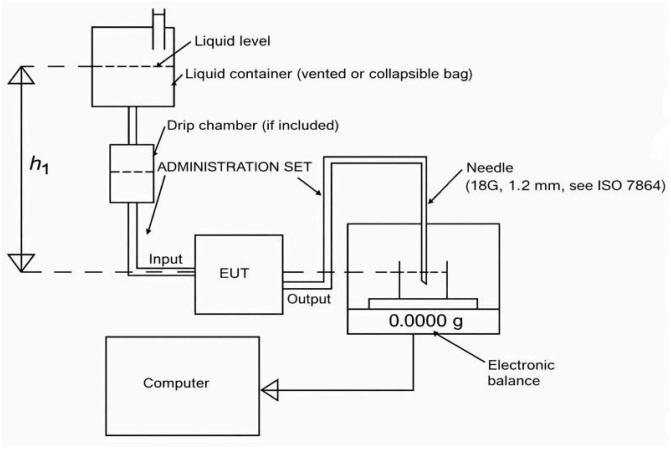


Fig. 1. Test apparatus for volumetric infusion pumps, adapted from IEC 60601-2-24:2012 [2]. EUT: equipment under test.

The tests at the nominal flow rate of $1\,\mathrm{mL/h}$ were conducted over a 24-hour period under ambient pressure conditions. In contrast, the tests at $25\,\mathrm{mL/h}$ were carried out for 2 hours under three distinct conditions: ambient pressure, positive pressure of $100\,\mathrm{mmHg}$, and negative pressure of $-100\,\mathrm{mmHg}$. For each condition, the global mean percentage errors were calculated for the second hour (T_1) and for the final hour (T_2).

Furthermore, all formulas and calculations employed in this research were based on IEC 60601-2-24:2012, which defines the technical criteria for evaluating infusion pump performance. The flow rate was calculated using

$$Q = \frac{60(W_j - W_k)}{Td}, \tag{1}$$

where W_j represents the final sample mass (g), W_k denotes the initial sample mass (g), T is the analysis period (min), and d represents the density of the fluid (g/mL).

For the 24-hour test, T_1 corresponds to the second hour of analysis, with the corresponding initial and final mass measurements taken at 120 and 240 min, respectively. For T_2, the corresponding measurements were taken at 1380 and 1440 min.

To calculate the overall mean percentage flow error, the following equation was used:

$$A = \frac{100(Q - r)}{r}, \tag{2}$$

where Q is the flow rate calculated using (1) and r is the flow rate programmed into the infusion pump. For tests lasting more than 2 hours, the overall mean percentage flow error was calculated for both T_1 and T_2.

For the construction of the start-up flow-rate graphs, the data acquisition procedure followed the prescriptions of IEC 60601-2-24:2012, with two measurements acquired per minute throughout the tests. For the start-up graphs, the flow rate was calculated as

$$Q_i = \frac{60(W_i - W_{i-1})}{Sd}, \tag{3}$$

where $i = 1, 2, \ldots, N$; W_i is the ith mass sample of the analysis period T_0 (g); T_0 is the analysis period (min); S is the interval between samples (min); and d represents the density of the fluid (g/mL).

The quantity Q_i represents the average flow rate over the interval between two consecutive mass measurements and should not be interpreted as an instantaneous flow rate.

A. Measurement Uncertainty

The evaluation of measurement uncertainty is essential for assessing the reliability of flow-rate determinations. In the experimental setup, the flow rate was measured gravimetrically based on the mass of the fluid collected over a defined time interval. The measurement model considered the initial and final mass readings, the time interval, and the fluid density.

Because the density is temperature-dependent, ambient room temperature was controlled throughout testing to minimize density fluctuations. Under these controlled environmental conditions, fluid density was treated as constant in the measurement model, and its uncertainty contribution was assumed negligible relative to the balance performance.

Similarly, time intervals were governed by automated data logging; hence, timing uncertainty was treated as negligible. Under these operational assumptions, the primary quantified uncertainty source originated from the mass measurements recorded by the analytical balance.

The flow rate (Q) was calculated using the gravimetric model:

$$Q = \frac{60(W_j - W_k)}{Td} \tag{4}$$

where W_k and W_j are the initial and final mass values (g), respectively, T is the analysis period (min), and d is the fluid density (g/mL).

Applying the law of propagation of uncertainty, the combined standard uncertainty of the flow rate, $u(Q)$:

$$u^2(Q) = \left(\frac{\partial Q}{\partial W_j} u(W_j)\right)^2 + \left(\frac{\partial Q}{\partial W_k} u(W_k)\right)^2 \tag{5}$$

The sensitivity coefficients are defined:

$$\frac{\partial Q}{\partial W_j} = \frac{60}{Td} \qquad \frac{\partial Q}{\partial W_k} = -\frac{60}{Td} \tag{6}$$

Assuming equal standard uncertainties for both initial and final mass readings, i.e., $u(W_k) = u(W_j) = u(W)$, the combined standard uncertainty simplifies to:

$$u(Q) = \frac{60\sqrt{2}}{Td} u(W) \tag{7}$$

The standard uncertainty $u(W)$ was derived from the calibration certificate of the analytical balance. The certificate reports an expanded uncertainty:

$$U(W) = 0.0002\,\mathrm{g} \tag{8}$$

with a coverage factor of $k = 2$. Therefore, the standard mass uncertainty was calculated:

$$u(W) = \frac{U(W)}{k} = \frac{0.0002\,\mathrm{g}}{2} = 0.0001\,\mathrm{g}. \tag{9}$$

The overall mean percentage flow error (E) was defined relative to the nominal flow rate r:

$$E_{=}\frac{Q - r}{r} \cdot 100(\%) \tag{10}$$

Since r is a flow rate adjusted by the operator on the infusion pump, its uncertainty contribution is zero. Consequently, the uncertainty associated with the error is obtained by propagating $u(Q)$:

$$u(E_p) = \left|\frac{\partial E_p}{\partial Q}\right| u(Q) = \frac{100}{r} u(Q) \tag{11}$$

The expanded uncertainty for the flow rate error is reported as $U(E_p) = k \cdot u(E_p)$, using a coverage factor of $k = 2$ (representing an approximate 95% confidence level) [10].

The uncertainty analysis was to T1 and T2. Consecutive flow-rate estimates Q_i (used for the start-up graph) are not statistically independent because each estimate is calculated from two consecutive mass measurements, and adjacent estimates share a common mass reading. Therefore, a complete uncertainty evaluation of the extrema of the trumpet curves would require consideration of covariance between successive estimates and statistical treatment of the extremum-selection process. Such an analysis is beyond the scope of the present study.

III. Results

To evaluate the pump's performance, the tests were carried out in two separate stages. Tables I and II show the results of the infusion pump before manufacturer intervention. Tables III and IV show the results after the device was returned by the manufacturer.

TABLE I
OVERALL MEAN PERCENTAGE FLOW ERROR AT 1 mL/h, TOTAL TIME OF 24 HOURS, AMBIENT PRESSURE, AND BEFORE MANUFACTURER INTERVENTION.

Medical Device	Error at T_1 (%)	Error at T_2 (%)
Infusion Pump	-9.89	-0.71

Figures 2, 3, 4, and 5 show the infusion pump's start-up graphs before manufacturer intervention. Figures 6, 7, 8, and 9 show the corresponding start-up graphs after manufacturer intervention.

Table V presents the measurement uncertainty results.

TABLE II
OVERALL MEAN PERCENTAGE FLOW ERROR AT 25 mL/h, TOTAL TIME OF 2 HOURS, UNDER THREE PRESSURE CONDITIONS, AND BEFORE MANUFACTURER INTERVENTION.

Pressure condition	Error (%)
No back pressure	-3.83
+100 mmHg back pressure	-11.13
−100 mmHg back pressure	-10.06

TABLE III
OVERALL MEAN PERCENTAGE FLOW ERROR AT 1 mL/h, TOTAL TIME OF 24 HOURS, AMBIENT PRESSURE, AND AFTER MANUFACTURER INTERVENTION.

Medical Device	Error at T_1 (%)	Error at T_2 (%)
Infusion Pump	-5.64	3.38

TABLE IV
OVERALL MEAN PERCENTAGE FLOW ERROR AT 25 mL/h, TOTAL TIME OF 2 HOURS, UNDER THREE PRESSURE CONDITIONS, AND AFTER MANUFACTURER INTERVENTION.

Pressure condition	Error (%)
No back pressure	-0.16
+100 mmHg back pressure	3.88
−100 mmHg back pressure	-0.29

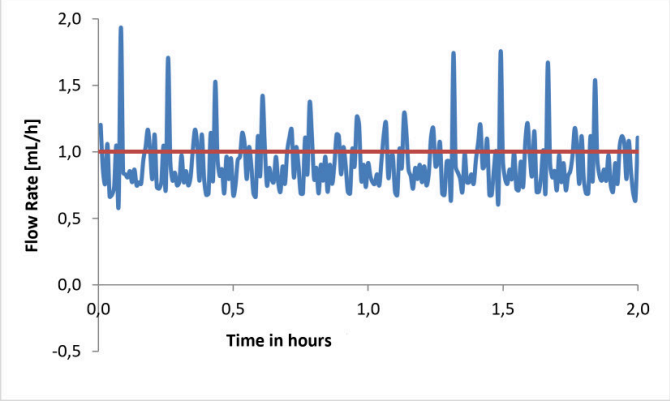


Fig. 2. Flow-rate behavior of the infusion pump operating at 1 mL/h, before manufacturer intervention, under ambient pressure conditions during the first two hours of testing.

IV. DISCUSSION

A. Performance Evaluation of the Infusion Pump

In the 1 mL/h test (T_1), the infusion pump presented an overall mean percentage flow error of -9.89%, indicating considerable underinfusion. Although the error was reduced in T_2 (-0.71%), the results demonstrated instability in the infusion pump performance over the 24-hour test period. At 25 mL/h, the medical device also showed negative overall mean percentage flow errors under all pressure conditions, especially with a back pressure of +100 mmHg, where the error reached -11.13%.

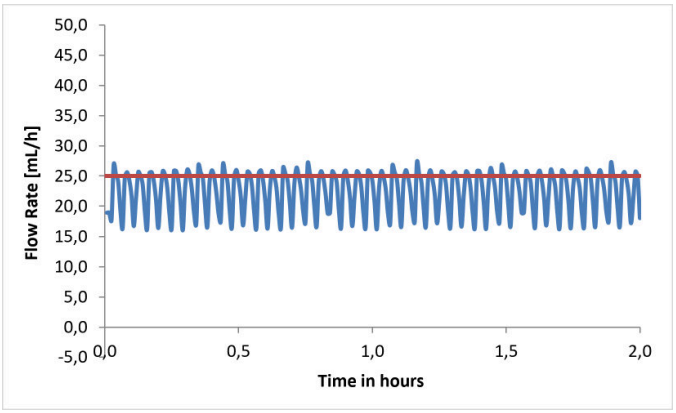


Fig. 3. Flow-rate behavior of the infusion pump operating at 25 mL/h, before manufacturer intervention, during two hours of testing, under +100 mmHg back pressure.

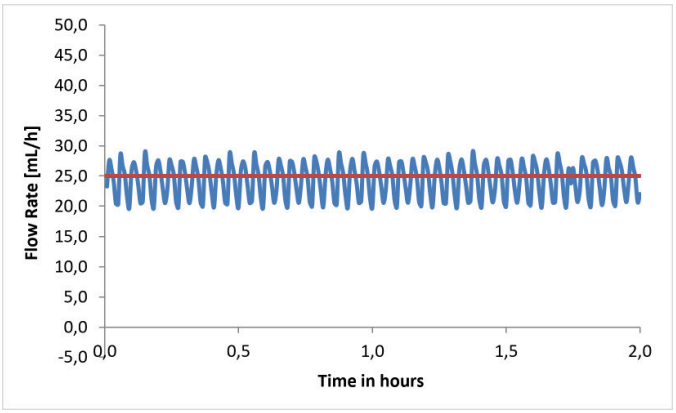


Fig. 4. Flow-rate behavior of the infusion pump operating at 25 mL/h, before manufacturer intervention, during two hours of testing, under ambient pressure.

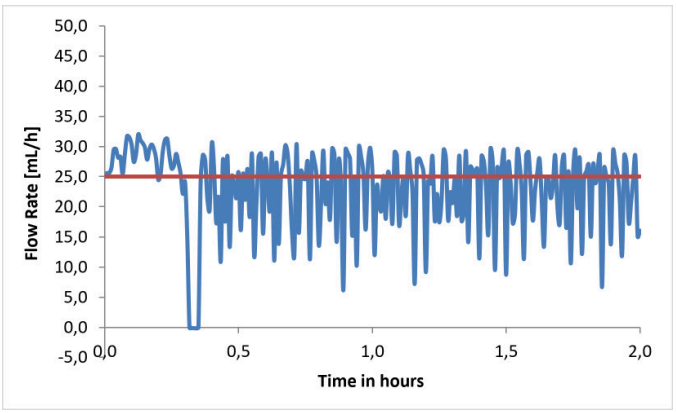


Fig. 5. Flow-rate behavior of the infusion pump operating at 25 mL/h, before manufacturer intervention, during two hours of testing, under −100 mmHg back pressure.

The start-up graphs before manufacturer intervention indicate transient behavior during the first thirty minutes of infusion. This is particularly true under the 1 mL/h condition and at 25 mL/h with a back pressure of −100 mmHg. At

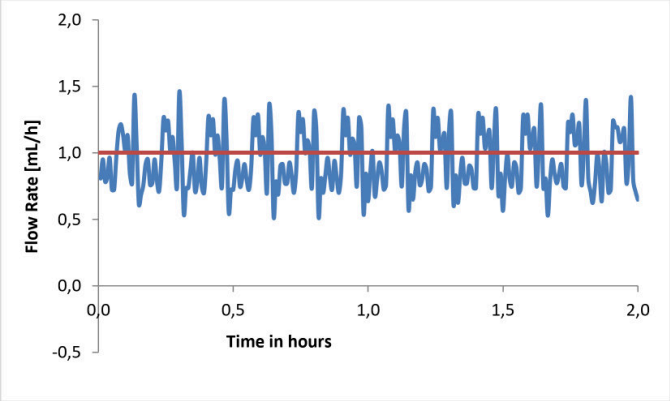


Fig. 6. Flow-rate behavior of the infusion pump operating at $1\,\mathrm{mL/h}$, after manufacturer intervention, during two hours of testing, under ambient pressure.

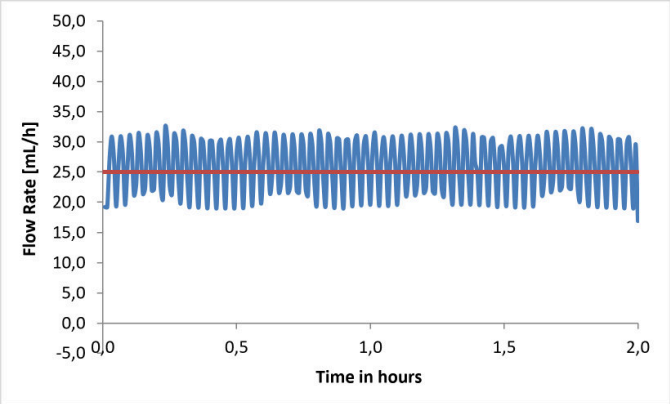


Fig. 7. Flow-rate behavior of the infusion pump operating at $25\,\mathrm{mL/h}$, after manufacturer intervention, during the first two hours of testing, under $+100\,\mathrm{mmHg}$ back pressure.

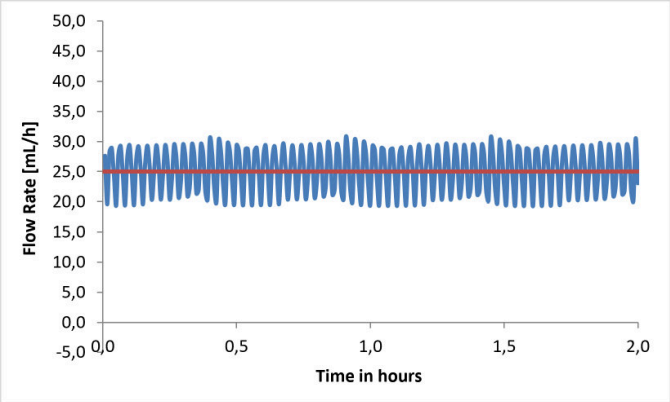


Fig. 8. Flow-rate behavior of the infusion pump operating at $25\,\mathrm{mL/h}$, after manufacturer intervention, during the first two hours of testing, under ambient pressure.

a programmed flow rate of $1\,\mathrm{mL/h}$ without back pressure, the flow fluctuated between $0.6\,\mathrm{mL/h}$ and $2.0\,\mathrm{mL/h}$. At a programmed flow rate of $25\,\mathrm{mL/h}$ with a back pressure of $-100\,\mathrm{mmHg}$, the flow exhibited considerably greater variability, ranging from $0\,\mathrm{mL/h}$ to $30\,\mathrm{mL/h}$.

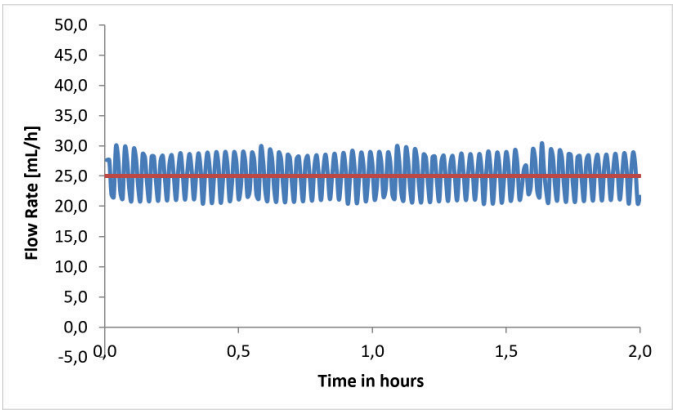


Fig. 9. Flow-rate behavior of the infusion pump operating at $25\,\mathrm{mL/h}$, after manufacturer intervention, during the first two hours of testing, under $-100\,\mathrm{mmHg}$ back pressure.

TABLE V
MEASUREMENT UNCERTAINTY

Parameter	**Flow: 1 mL/h**	**Flow: 25 mL/h**
Analysis interval (T)	60 min (1 h)	60 min (1 h)
Assumed density (d)	1 g/mL	1 g/mL
Balance uncertainty $u(W)$	0.0001 g	0.0001 g
Standard Uncertainty of Flow Rate, $u(Q)$		
Formula: $u(Q) = \frac{60}{T \cdot d}\sqrt{2} \cdot u(W)$		
Calculation	$\frac{60}{60 \cdot 1}\sqrt{2} \cdot 0.0001$	$\frac{60}{60 \cdot 1}\sqrt{2} \cdot 0.0001$
Calculated value	$\mathbf{0.0001}$ mL/h	$\mathbf{0.0001}$ mL/h
Standard Uncertainty of Error, $u(E_p)$		
Formula: $u(E_p) = \frac{100}{r} \cdot u(Q)$		
Calculation	$\frac{100}{1} \cdot 0.000141$	$\frac{100}{25} \cdot 0.000141$
Calculated value	$\mathbf{0.0141}\%$	$\mathbf{0.0006}\%$
Expanded Uncertainty of Error, $U(E_p)$		
Coverage factor (k)	2 (95% conf.)	2 (95% conf.)
Formula: $U(E_p) = k \cdot u(E_p)$		
Calculation	$2 \cdot 0.0141\%$	$2 \cdot 0.0006\%$
Final Result $U(E_p)$	$\pm\mathbf{0.0282}\%$	$\pm\mathbf{0.0012}\%$

As shown in Fig. 5, the infusion pump exhibited a temporary loss of infusion continuity when operated at $25\,\mathrm{mL/h}$ under a back pressure of $-100\,\mathrm{mmHg}$. This phenomenon was characterized by a complete cessation of fluid delivery lasting approximately 3.5 minutes, recorded between minutes 18 and 22 of testing.

These findings demonstrate that variations in pressure test conditions and low flow rates can severely compromise infusion pump accuracy despite up-to-date routine maintenance.

The classification "new and operational" does not necessarily ensure adequate metrological performance. In this case, the infusion pump was considered suitable for use, but the normative test showed that its flow delivery was still outside the expected tolerance range.

The results were presented to the clinical engineering team. The infusion pump was segregated and was immediately sent to the manufacturer for intervention. The manufacturer did not provide information regarding the procedures performed

during this intervention. Once the medical device was returned, all tests were repeated to evaluate its performance after the intervention.

At $1\,\mathrm{mL/h}$, the overall mean percentage flow error was reduced to -5.64% in T_1 and changed to 3.38% in T_2, demonstrating better accuracy and stability during long-term operation. Similarly, at $25\,\mathrm{mL/h}$, the errors decreased significantly under all test conditions, remaining close to the nominal flow rate. The best performance was observed without back pressure, with an error of only -0.16%, while the tests with $+100\,\mathrm{mmHg}$ and $-100\,\mathrm{mmHg}$ back pressure presented errors of 3.88% and -0.29%, respectively.

The start-up graphs obtained after manufacturer intervention also demonstrated smoother flow-rate stabilization and reduced oscillations during the first hours of operation.

B. Potential Clinical Impact of Flow Rate Deviations

The flow-rate deviations identified in the present study may have clinical relevance because infusion pumps are used to deliver medications and fluids at programmed rates, including in pediatric intensive care settings where relatively small absolute flow rates may be clinically important. However, the present study did not evaluate medication concentrations, administered doses, pharmacokinetic outcomes, clinical events, or patient outcomes.

The assessment demonstrated underinfusion, particularly at $1\,\mathrm{mL/h}$, as well as transient fluctuations under specific pressure conditions. From a clinical perspective, underinfusion is particularly concerning because it may not be immediately detected by the healthcare team, especially when the device does not trigger an alarm. In critical care, this delay can reduce the expected therapeutic effect of the medication and expose the patient to avoidable instability.

These discrepancies may be clinically relevant in intensive care settings, particularly when infusion pumps are used to administer medications for which accurate dosing and timely administration are important. The present study did not evaluate clinical outcomes, medication dosing errors, or treatment responses; therefore, no direct relationship between the measured flow-rate deviations and patient outcomes can be established. Nevertheless, clinical guidelines can help contextualize the potential relevance of accurate infusion delivery. In pediatric sepsis management, for example, the Surviving Sepsis Campaign guidelines emphasize timely and optimized antimicrobial therapy to reduce the risk of treatment failure and organ dysfunction.

C. Importance of Hospital-Based Performance Verification of Infusion Pumps

The results obtained in this study demonstrate that hospitals should not rely exclusively on the calibration reports provided by manufacturers or service/calibration providers. Healthcare institutions must also establish internal methods to periodically evaluate the performance and accuracy of infusion pumps used in clinical practice. Although it may not be feasible to test every medical device available in the hospital, a representative sampling strategy can be adopted to identify possible deviations and performance degradation over time.

Overall, this investigation illustrates how a documented sentinel case, identified through systematic testing, can support evidence-based clinical engineering decision-making, while underscoring that broader sampling programs are required to determine the true prevalence and underlying determinants of metrological deviations.

D. Considerations for Infusion Pump Manufacturers and Medical Device Calibration Providers

These findings indicate that quality assurance for infusion pumps must extend beyond routine calibration. Given their important role in intensive care and critical therapies, manufacturers and service/calibration providers must ensure flow-rate accuracy throughout the equipment's lifecycle.

While periodic preventive maintenance must follow manufacturer guidelines and standards like IEC 60601-2-24, our results show that standard calibration records cannot fully guarantee real-world functional performance. Consequently, post-maintenance protocols should incorporate functional testing under specific operational variables, such as low flow rates and varying backpressure conditions, before releasing the device back to clinical use.

Manufacturers and service/calibration providers also need to ensure metrological traceability and maintain history logs, including past failures and operational tests, to pinpoint recurring issues. Finally, continuous technician training and strict control over the compatibility of infusion sets are important, since third-party infusion sets may compromise flow accuracy.

E. Proposed Functional Verification Procedure

Considering the differences in financial resources and technical infrastructure among hospitals, some complementary approaches to verifying infusion pumps are proposed.

It is important to emphasize that neither of these verification procedures replaces the periodic calibration of the medical devices. Instead, they should be regarded as complementary post-market surveillance strategies intended to identify infusion pumps with potential performance degradation and verify the effectiveness of manufacturer intervention.

1) Hospitals Without Technical Instrumentation: The clinical engineering team could implement a simplified functional verification procedure as part of the post-market surveillance of infusion pumps. The procedure should begin with a visual inspection of the device and its accessories, including verification of the physical integrity, display, keypad, alarm system, power supply, and infusion set, and confirmation that preventive maintenance and calibration are up to date.

For the initial functional assessment, the infusion pump may be programmed to a predefined target flow rate to deliver a $0.9\,\%$ saline solution into a class A graduated cylinder over a specified time interval. The collected volume could be compared against the theoretical nominal volume.

Since this simplified verification relies on manual volume measurements, the uncertainty associated with graduated

cylinder readings should be considered. The measurements based on graduated cylinders or other low-resolution volumetric methods should be interpreted cautiously, particularly at low flow rates, because reading resolution, meniscus interpretation, evaporation, and the measurement interval can substantially influence the result.

Alternatively, automated drip counting can be employed to monitor flow rate accuracy. While subject to specific limitations, it also offers a practical solution for verifying infusion pump performance in the intervals between metrological calibrations [9].

2) Hospitals With Technical Instrumentation: Hospitals equipped with specialized testing facilities may perform a more comprehensive performance verification by following the prescriptions in IEC 60601-2-24. In this approach, the clinical engineering team may employ a calibrated analytical balance, automated data acquisition systems, and controlled environmental conditions to determine the actual flow rate delivered by the infusion pump over predefined time intervals using the gravimetric method.

Infusion pump analyzers may also be used. However, it should be recognized that the gravimetric method offers advantages over infusion pump analyzers for two main reasons: its lower acquisition cost and its versatility, as it allows the use of fluids with different densities and viscosities, a condition that is generally limited in infusion pump analyzers.

F. Study Limitations

The study did not include internal inspection or component-level failure analysis. Due to contractual restrictions, the infusion pump could not be internally inspected, and the manufacturer did not provide information regarding the procedures performed while the device was outside the hospital. Consequently, the root cause of the initial flow-rate deviations and the specific reason for the improvement observed during the second assessment could not be established.

Second, tests were performed under controlled laboratory conditions. The results, therefore, do not reproduce the full range of clinical conditions, including different medication viscosities, infusion solutions, infusion sets, or interactions with other medical equipment.

Finally, no direct clinical outcomes, medication dosing errors, adverse events, or patient safety incidents were evaluated. The clinical relevance discussed in this study is therefore based on the potential implications of the measured flow-rate deviations rather than on documented patient outcomes.

V. Conclusion

The results demonstrated that routine maintenance and calibration records alone may not fully reflect the actual performance of infusion pumps used in clinical practice. In the evaluated case, clinically relevant flow rate deviations were identified through independent testing, despite the device being considered operational.

The substantial performance improvement observed after manufacturer intervention reinforces the importance of post-market verification strategies capable of detecting flow rate deviations before they compromise infusion therapy.

Although this study presents a representative case from an ongoing broader investigation, the findings highlight the potential contribution of performance assessment programs to patient safety, quality assurance, and evidence-based clinical engineering practices.

Acknowledgment

This study was carried out within the framework of a partnership between the Escola Politécnica and the University Hospital of the University of São Paulo. The authors acknowledge the support provided by the endowment fund of the University of São Paulo Engineering School (Amigos da Poli – AdP). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (309464/2022-6 to HTM).

References

[1] M. S. Silva *et al.*, "Precision and reliability study of hospital infusion pumps: a systematic review," *BioMed. Eng. OnLine*, vol. 22, no. 1, Art. no. 26, 2023, doi: 10.1186/s12938-023-01088-w.

[2] International Electrotechnical Commission, *Medical Electrical Equipment–Part 2-24: Particular Requirements for the Basic Safety and Essential Performance of Infusion Pumps and Controllers*, IEC 60601-2-24, Geneva, Switzerland, 2012.

[3] S. J. Calil and E. T. Gomide, "Dispositivos de infusão," in *Equipamentos médico-hospitalares e o gerenciamento da manutenção*. Brasília, DF, Brazil: Ministério da Saúde, Projeto REFORSUS, 2002, pp. 301–330.

[4] M. Husch, C. Sullivan, D. Rooney, *et al.*, "Insights from the sharp end of intravenous medication errors: implications for infusion pump technology," *Qual. Saf. Health Care*, vol. 14, no. 2, pp. 80–86, 2005, doi: 10.1136/qshc.2004.011957.

[5] M. A. L. Etelvino, M. G. Pereira, and R. D. Barreiro Filho, "Risk management in infusion pumps," *Vigil. Sanit. Debate*, vol. 7, no. 4, pp. 61–66, 2019, doi: 10.22239/2317-269X.01365.

[6] S. L. Weiss *et al.*, "Surviving Sepsis Campaign International Guidelines for the Management of Septic Shock and Sepsis-Associated Organ Dysfunction in Children," *Intensive Care Med.*, vol. 46, no. S1, pp. 10–67, 2020, doi: 10.1007/s00134-019-05878-6.

[7] O. Atanda, J. West, T. Stables, C. Johnson, R. Merrifield, and J. Kinross, "Flow rate accuracy of infusion devices within healthcare settings: a systematic review," *Ther. Adv. Drug Saf.*, vol. 14, pp. 1–15, 2023, doi: 10.1177/20420986231188602.

[8] A. Hult *et al.*, "Flow rate deviation in infusion pump: infusion set defect enables pump malfunction and considerable accuracy deviation," *J. Infus. Nurs.*, vol. 47, no. 1, pp. 30–35, 2024, doi: 10.1097/NAN.0000000000000529.

[9] J. An, R. D. Butterfield, and N. M. Sims, "Evaluation of clinical infusion pump performance through downstream microdrop monitoring: a preliminary study," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2020, pp. 6066–6069.

[10] Joint Committee for Guides in Metrology, *Evaluation of Measurement Data–Guide to the Expression of Uncertainty in Measurement*, JCGM 100:2008, BIPM, 2008.

Potential Capacity and Territorial Distribution of Mammography Equipment in the State of São Paulo

1st Gilberto S. Zunta
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-6006-2142

2nd Miguel R. Nunes
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0009-1757-5568

3rd Cristina Q. A. Fais
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0007-3632-7223

4th Henrique T. Moriya
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0000-0002-1595-1779

5th Edison Silva Filho
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-5397-088X

***Abstract*—Assessing the potential capacity of mammography equipment requires consideration of the number of units, their affiliation with the public health system, regional distribution, and the observed number of examinations performed. This study aimed to evaluate the potential capacity and territorial distribution of mammography equipment in the State of São Paulo, Brazil, using secondary data from the National Registry of Health Facilities, the Outpatient Information System of the Brazilian Unified Health System, and the National Supplementary Health Agency. Potential capacity was estimated using a reference parameter of 5,069 examinations per mammography unit per year. The historical series considered women aged 50 to 69 years from 2016 to 2024 and women aged 40 to 74 years in 2025 and 2026. For March 2026, three scenarios were analyzed: total capacity versus the female population aged 40 to 74 years, capacity available to the Brazilian Unified Health System versus the same population, and capacity available to the Brazilian Unified Health System versus the population without active private health insurance. At the state level, the capacity-to-demand ratio was 1.47 in the total-capacity scenario, 0.63 in the public-system scenario considering the total population, and 1.04 in the public-system scenario considering the population without private insurance. Although total capacity exceeded estimated demand in all 19 Regional Health Care Networks, 17 regions had ratios below 1.0 when only public-system capacity was considered, and eight remained below 1.0 after adjustment for private health insurance coverage. The results indicate that potential sufficiency depends on the population considered, availability within the public health system, and the regional distribution of mammography equipment, with findings that may support health managers in equipment planning, regional regulation, and the organization of access to mammography services.**

***Index Terms*—Mammography, mammography equipment, Brazilian Unified Health System, installed capacity.**

I. Introduction

Breast cancer remains a relevant public health issue in Brazil, both because of its epidemiological magnitude and because it requires organized health care networks capable of supporting early detection, timely diagnosis, and continuity of care. Mammography is the main imaging modality in this process, used for population-based screening of asymptomatic women and for the diagnostic investigation of suspicious signs and symptoms [1]. Within the Brazilian Unified Health System (Sistema Único de Saúde, SUS), Technical Note No. 626/2025 from the Ministry of Health recommends population-based mammography screening for women aged 50 to 74 years, with biennial periodicity, while clarifying that access to mammography is not restricted to this age group when there is clinical demand or indication [2].

Records of mammography equipment in official databases are relevant for health planning, but they do not, by themselves, ensure effective access to examinations. Equipment may be located in public, private, or contracted facilities, may or may not be available to SUS, and may be unevenly distributed across the territory. In addition, the observed number of examinations depends on operational factors such as scheduling, workforce availability, funding, regulatory flows, maintenance, image quality, and reporting capacity. Therefore, the assessment of mammography sufficiency should combine installed capacity, SUS availability, outpatient production, and regional distribution.

The aggregate number of mammography units may suggest sufficiency, whereas analyses based on distance, health regions, and equipment use can reveal locational inequalities and limitations that cannot be attributed only to a lack of machines [3], [4]. Inequalities between women with and without private health insurance should also be considered, since mammography use differs according to private health plan coverage and because diagnostic imaging services are not equally distributed between the public and private sectors [5], [6].

Public databases provide a useful basis for this type of evaluation, but they also have limitations. The National Registry of Health Facilities (Cadastro Nacional de Estabelecimentos de Saúde, CNES) makes it possible to identify facilities, equipment, and SUS affiliation, while the Outpatient Information

System of SUS (Sistema de Informações Ambulatoriais do SUS, SIA/SUS) allows monitoring of outpatient procedures recorded in the public system. The National Supplementary Health Agency (Agência Nacional de Saúde Suplementar, ANS), in turn, provides information on active private health plan beneficiaries, which can be used as a complementary proxy for interpreting potential demand on the public network. These databases have different purposes and structures and should not be interpreted as direct measures of individual access to mammography.

In the State of São Paulo, public databases allow the integration of information on equipment, outpatient production, reference population, private health plan coverage, and regionalization into Regional Health Care Networks (Redes Regionais de Atenção à Saúde, RRAS). This integration can help distinguish total potential capacity from potential capacity available to SUS and can identify regions in which the capacity-to-demand ratio remains low even after accounting for the population covered by active private health plans.

Thus, this study aimed to evaluate the potential capacity and territorial distribution of mammography equipment in the State of São Paulo, considering total availability and availability to SUS, the reference population for mammography, outpatient production recorded in SUS, and private health insurance coverage.

II. Materials and Methods

The study material consisted of secondary datasets related to mammography equipment, reference population, outpatient mammography procedures, and private health insurance coverage in the State of São Paulo. The analysis was designed as a scenario-based assessment of potential capacity and was not intended to measure actual access, waiting time, operational availability, or individual service utilization. Data processing, organization, and analysis were performed using R, version 4.5.0 (R Foundation for Statistical Computing, Vienna, Austria), with support from the RStudio Desktop integrated development environment, version 2024.12.1.563 (Posit Software, PBC, Boston, MA, USA). R packages were used for data import, cleaning, transformation, analysis, visualization, spatial processing, and export, including tidyverse, readr, readxl, janitor, lubridate, stringr, stringi, forcats, ggplot2, scales, sf, patchwork, and writexl [7].

A. Study Design and Data Sources

A retrospective, descriptive, quantitative ecological study was conducted using secondary data from official health information systems. Data on mammography equipment were obtained from CNES for the State of São Paulo between 2016 and March 2026, considering facility location, equipment type, operational status, and SUS affiliation [8]. For each year of the historical series, the March reporting period was used as the reference month. Equipment was classified according to whether it was registered as existing, in use, and available to SUS. For the historical series, the annual evolution of the mammography equipment inventory was represented by the March reporting period of each year, from 2016 to 2026, in order to maintain temporal comparability with the territorial snapshot from March 2026. The numbers of existing and in-use mammography units were consolidated, with identification of their availability to SUS.

The territorial analysis used data from March 2026, after geographic validation of facilities and exclusion of an extreme registry record that was incompatible with the expected distribution of the technology inventory. Mammography units in use were included, distinguishing equipment available and not available to SUS. This distinction was maintained throughout the analyses because the total availability of equipment in the territory does not necessarily correspond to availability for the population served by the public health system. Before aggregation, the territorial dataset underwent consistency checks involving municipal codes, geographic coordinates, duplicate facilities, compatibility with municipal boundaries, and coherence of state-level totals.

The observed number of examinations was obtained from SIA/SUS by place of care for the State of São Paulo, considering procedure codes 0204030188, bilateral screening mammography, and 0204030030, mammography [9]. Because these records include outpatient mammography procedures registered in SUS, observed production was interpreted as public-system production rather than as a direct measure of screening coverage. For 2026, the observed production was limited to the January to March period, according to data availability. Information on active private health plan beneficiaries was retrieved from the ANS TabNet platform, filtered by female sex, age group from 40 to 74 years, and municipality of residence in the State of São Paulo [10], [11].

B. Reference Population and Demand Estimation

The reference population was defined to preserve historical comparability while allowing evaluation of an expanded potential demand in the most recent years. From 2016 to 2024, women aged 50 to 69 years were considered. In 2025 and 2026, women aged 40 to 74 years were considered. This expanded age range was used as a planning and sensitivity denominator for potential demand, not as an indication that the 40 to 74 age group represents the single normative population for SUS population-based screening throughout the entire period. This distinction is relevant because current national guidance recommends population-based screening for women aged 50 to 74 years, while mammography may also be performed outside this age range when there is clinical indication or demand. Therefore, the 2016–2024 and 2025–2026 periods should not be interpreted as a single homogeneous historical trend, because the reference population denominator changed.

Estimated annual demand was calculated by dividing the reference population by two, considering biennial periodicity. For the territorial analysis of March 2026, the female population aged 40 to 74 years was used, followed by a complementary analysis of the population without active private health insurance. This population was estimated as the difference between the total female population aged 40 to 74 years and

the number of active private health plan beneficiaries reported by ANS. This adjustment was treated as a sensitivity scenario for potential public-system demand, not as a direct measure of individual dependence on SUS.

C. Potential Capacity and Observed Production

Annual potential capacity was estimated as the product of the number of mammography units in use and the parameter of 5,069 examinations per unit per year. This parameter was adopted from the INCA review of the production capacity of standard mammography equipment, which considers 3 examinations per hour, 8 hours per day, 22 days per month, 12 months per year, and an 80% utilization factor [12]. Thus, the parameter was used as a normative planning reference and should not be interpreted as the measured productivity of each facility. As a historical reference, the SUS health care criteria and parameters were also considered, since they present a previous benchmark of 6,758 examinations per unit per year and territorial planning criteria based on travel time or maximum distance [13].

Three main indicators were calculated: annual potential capacity, the capacity-to-demand ratio, and observed production in SUS. The capacity-to-demand ratio was defined as annual potential capacity divided by estimated annual demand. Observed production corresponded to the number of mammograms recorded in SIA/SUS during the analyzed period. Utilization of potential capacity was interpreted as a descriptive indicator, since the study did not measure effective operating time, staffing, scheduling, regulatory flows, examination quality, or reporting capacity.

D. Territorial Analysis and Analytical Scenarios

The territorial analysis was conducted by RRAS, according to the current regionalization of the State of São Paulo, composed of 19 RRAS and 62 Health Regions [14]. For March 2026, three analytical scenarios were defined. Scenario C1 compared total potential capacity, including SUS and non-SUS mammography units, with the estimated annual demand of the total female population aged 40 to 74 years. Scenario C2 compared SUS potential capacity with the same demand. Scenario C3 compared SUS potential capacity with the estimated demand of the female population aged 40 to 74 years without active private health insurance.

Results were presented as capacity-to-demand ratios. Values equal to or greater than 1.0 indicated potential capacity equal to or greater than estimated demand, whereas values below 1.0 indicated potential capacity below estimated demand. This classification was used only as an indicator of potential capacity and did not imply effective individual access to mammography. Ratios displayed in tables and figure labels were presented with two decimal places without rounding, while exact values were retained in the analytical datasets for classification.

III. Results and Discussion

A. Historical Series of Capacity and Observed Production

The CNES time series made it possible to monitor the ratio between estimated potential capacity and annual demand from 2016 to March 2026. Between 2016 and 2024, when the reference population was defined as women aged 50 to 69 years, the ratio associated with total mammography equipment remained above 2.7. The ratio associated with equipment available to SUS remained close to or above 1.0 during the same period. In 2025 and 2026, with adoption of the expanded reference population aged 40 to 74 years, the ratios decreased markedly, a pattern that should not be interpreted as a sudden loss of equipment. Rather, this decrease primarily reflects the change in the population denominator used to estimate annual demand; therefore, the two periods should be compared with caution and not interpreted as a continuous homogeneous trend.

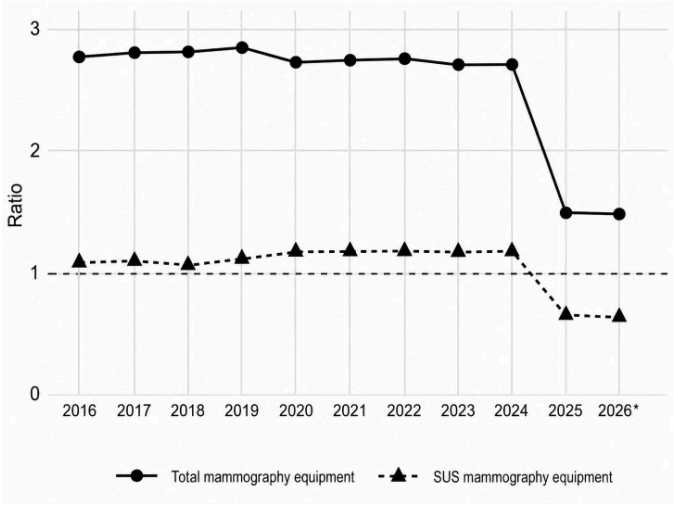


Fig. 1. Ratio of estimated potential capacity to annual mammography demand in the State of São Paulo, according to total mammography equipment and equipment available to SUS. The reference population was defined as women aged 50 to 69 years from 2016 to 2024 and women aged 40 to 74 years in 2025 and 2026; therefore, the most recent years should be interpreted considering the expanded denominator.

Fig. 1 shows that interpretation of potential sufficiency depends on the population considered and on the health care affiliation of the equipment. The total ratio remained above 1.0 throughout the period, but the SUS ratio fell below this threshold in the most recent years. This result introduces the difference observed in the 2026 territorial analysis, in which total capacity and capacity available to SUS provide distinct interpretations of the same technological network. For 2026, only the January to March period was considered.

Observed production in SUS followed a pattern that differed from potential capacity. In 2016, 1,411,157 mammograms were recorded in SIA/SUS, whereas 1,358,451 were recorded in 2019. In 2020, production decreased to 870,685 examinations, a result consistent with the interruption and reorganization of services during the COVID-19 pandemic.

In 2025, 1,423,569 examinations were recorded. In 2026, the value of 326,289 examinations corresponds only to the January to March period and should not be interpreted as annual production.

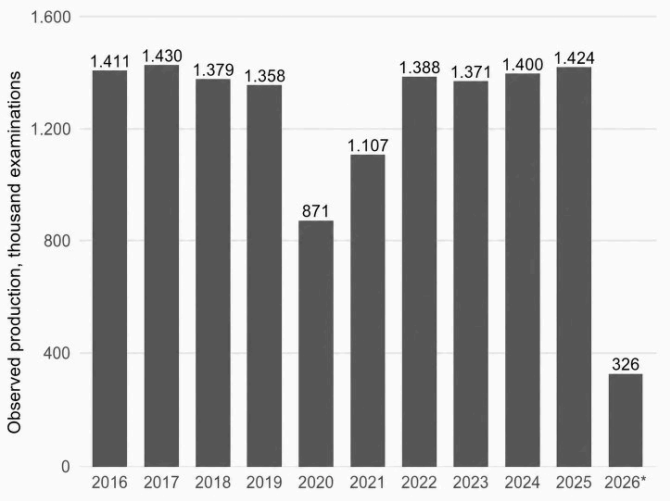


Fig. 2. Observed number of mammograms recorded in SIA/SUS in the State of São Paulo. The year 2026 corresponds to the January to March period.

Fig. 2 shows that observed production should be interpreted differently from potential capacity. The existence of mammography equipment in use does not imply a proportional number of examinations performed. Recorded production depends on operating conditions, appointment availability, referral flows, staffing, maintenance, supplies, and reporting capacity. This difference supports the need to evaluate the technology inventory using more than one indicator.

B. State-Level Potential Capacity Scenarios

In the March 2026 territorial analysis, after treatment of the registry outlier, 1,493 mammography units in use were considered, of which 638 were available to SUS and 855 were not affiliated with SUS. The female population aged 40 to 74 years was estimated at 10,253,362 women, resulting in an annual demand of 5,126,681 examinations. Total potential capacity was 7,568,017 examinations per year, while SUS potential capacity was 3,234,022 examinations per year.

TABLE I
ANALYTICAL SCENARIOS OF MAMMOGRAPHY POTENTIAL CAPACITY USING 5,069 EXAMINATIONS PER MAMMOGRAPHY UNIT PER YEAR.

Scenario	Population	Equipment	Annual demand	Capacity	Ratio	RRAS < 1
C1	Total	Total	5,126,681	7,568,017	1.47	0
C2	Total	SUS	5,126,681	3,234,022	0.63	17
C3	Without private insurance	SUS	3,099,553	3,234,022	1.04	8

Table I summarizes the three analytical scenarios. In scenario C1, the capacity-to-demand ratio was 1.47, suggesting potential sufficiency at the state aggregate level when all mammography units in use were considered. In scenario C2, restricted to equipment available to SUS in relation to the total reference population, the ratio was 0.63. This contrast indicates that an assessment based only on total capacity may mask limitations in public-sector availability.

In scenario C3, after excluding active private health plan beneficiaries, the reference population was reduced to 6,199,107 women and the estimated annual demand to 3,099,553.5 examinations. The state-level ratio increased to 1.04. This result suggests that SUS potential capacity approaches estimated demand when the population without active private health insurance is considered. However, the margin is narrow and should be interpreted with caution, particularly because regional distribution is not homogeneous.

C. Regional Distribution of the Three Scenarios

The RRAS analysis showed that state-level sufficiency in scenario C1 was not distributed equivalently in the scenarios restricted to SUS. All 19 RRAS had ratios equal to or greater than 1.0 in scenario C1. In scenario C2, 17 RRAS had ratios below 1.0. In scenario C3, eight RRAS remained below 1.0 even after adjustment for the population without active private health insurance.

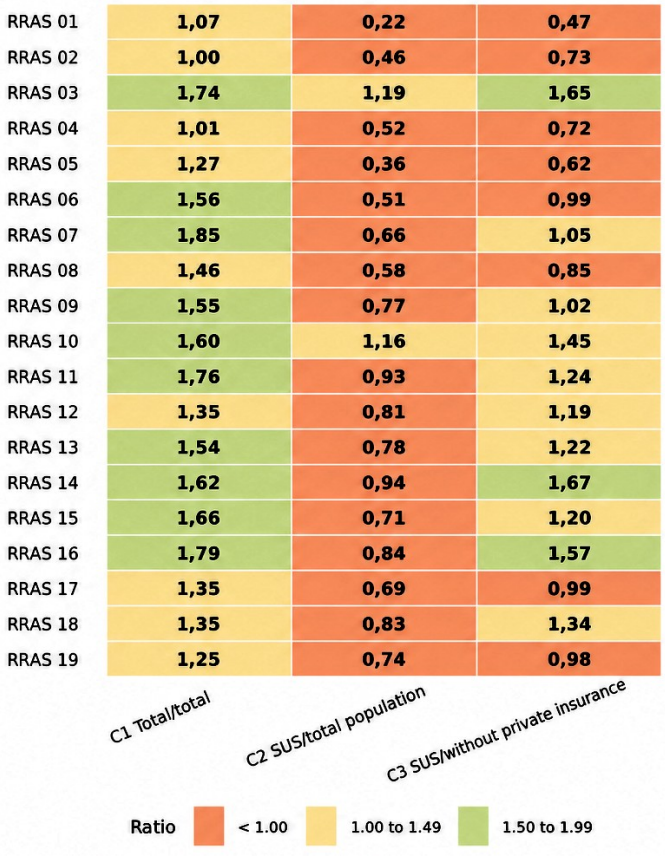


Fig. 3. Ratio between potential capacity and estimated annual demand by RRAS in three analytical scenarios, considering 5,069 examinations per mammography unit per year.

Fig. 3 summarizes the change in interpretation across the scenarios. Scenario C1 shows a predominance of ratios above 1.0, whereas scenario C2 indicates potential insufficiency in

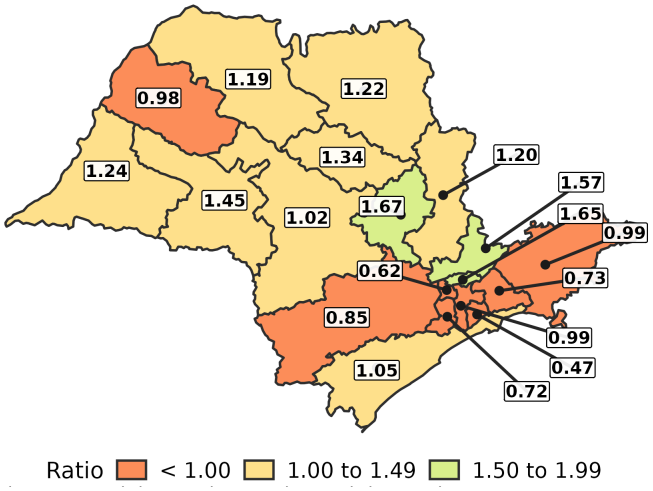


Fig. 4. Ratio between SUS potential capacity and estimated annual demand among women aged 40 to 74 years without active private health insurance, by RRAS.

almost all RRAS. Scenario C3 reduces the number of regions below 1.0 but does not eliminate regional inequality. This difference supports the need to distinguish total capacity from capacity available to SUS.

The observed pattern is consistent with previous reports in the literature, in which the overall availability of mammography equipment does not eliminate regional disparities in access and use. Amaral et al. observed that the total number of mammography units in Brazil could be sufficient when analyzed in aggregate, but that applying a 60 km distance criterion revealed areas without adequate provision [3]. Xavier et al. also reported that problems related to mammography coverage, allocation, and use across Health Regions cannot be explained only by a lack of equipment [4].

RRAS 03 and RRAS 10 were the only regions with a ratio of SUS capacity to total population equal to or greater than 1.0. The lowest ratios in scenario C2 occurred in RRAS 01, 05, 02, 06, and 04. This pattern indicates that SUS availability does not uniformly keep pace with demand estimated from the reference population. In highly populated regions, such as RRAS 06, the absolute number of SUS mammography units is high, but estimated demand is also greater, reducing the capacity-to-demand ratio.

D. Population Without Active Private Insurance and Regional Persistence

The map for scenario C3 shows the regional distribution of the ratio between SUS potential capacity and estimated demand among women aged 40 to 74 years without active private health insurance. This scenario improves the interpretation in relation to scenario C2 because it acknowledges that part of the population is covered by private health plans. Even so, the persistence of RRAS below 1.0 shows that the aggregate state-level analysis is insufficient to characterize the territorial distribution of capacity.

TABLE II
RRAS WITH SUS CAPACITY RATIOS BELOW 1.0 AFTER ADJUSTMENT FOR THE POPULATION WITHOUT ACTIVE PRIVATE HEALTH INSURANCE.

RRAS	SUS units	Pop. without insurance, thousands	SUS/total	SUS/without insurance
RRAS 01	14	297	0.22	0.47
RRAS 05	15	245	0.36	0.62
RRAS 04	13	181	0.52	0.72
RRAS 02	30	414	0.46	0.73
RRAS 08	31	366	0.58	0.85
RRAS 19	13	134	0.74	0.98
RRAS 17	39	399	0.69	0.99
RRAS 06	142	1,442	0.51	0.99

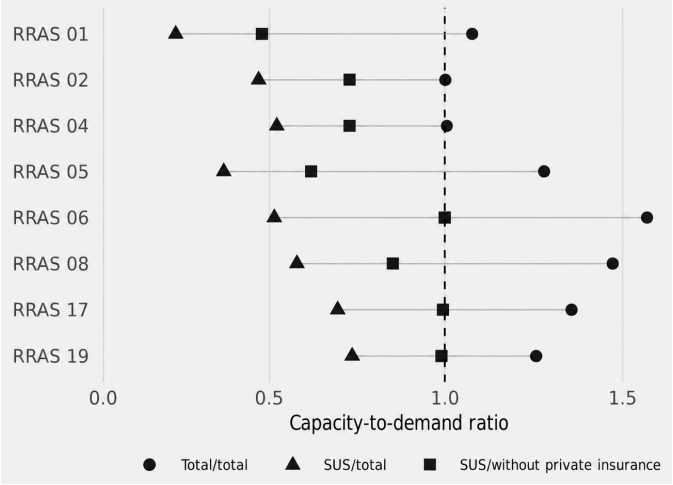


Fig. 5. RRAS with SUS capacity-to-demand ratios below 1.0 in the total-population scenario and in the complementary scenario excluding the population with active private health insurance.

Table II presents the eight RRAS that remained below 1.0 in scenario C3. The lowest ratios were observed in RRAS 01, 05, 04, 02, and 08. RRAS 19, 17, and 06 were close to 1.0, but remained below this threshold when exact values were used. This result shows that excluding the population with active private health insurance improves the ratio in all regions but does not eliminate the possibility of regional insufficiency in SUS capacity.

Fig. 5 reinforces that adjustment for the population without active private health insurance reduces the number of regions below the 1.0 threshold but does not eliminate regional inequality. The figure also shows that some RRAS remain close to the sufficiency threshold, a condition that should be interpreted with caution because it depends on the normative parameter and on the population denominator adopted.

The interpretation of scenario C3 should consider the nature of ANS data. The number of active private health plan beneficiaries was used as a proxy to reduce the population potentially dependent on the public network. However, this approach does not allow individual-level inference that all women without private insurance rely exclusively on SUS or that all beneficiaries undergo mammography outside SUS. For this reason, scenario C3 should be understood as a complementary sensitivity analysis.

Even so, the analysis using ANS data is useful for planning because it recognizes that the total female population does not exert the same pressure on the public health network. Population studies show differences in mammography use according to health insurance coverage, reinforcing the need to consider this dimension in capacity analyses [5]. In the present study, the improvement of scenario C3 in relation to C2 confirms that private coverage changes the estimated potential demand on SUS, but does not remove all regional inequalities.

E. Implications and Limitations

The results suggest that the discussion of mammography equipment in the State of São Paulo should not be limited to the acquisition of additional machines. Estimated total capacity indicates that the aggregate technology inventory is sufficient when all sectors are considered. However, SUS availability and territorial distribution change this interpretation. For Clinical Engineering, this result indicates that planning decisions should integrate inventory, health care affiliation, location, observed production, and regional demand.

The use of public data allows a reproducible and updatable assessment, but depends on the quality of the records. CNES reports the existence, operational status, and SUS affiliation of equipment, but it does not record the number of appointments effectively offered, number of shifts, workforce availability, operational quality of the equipment, or reporting capacity. SIA/SUS describes outpatient production recorded in the public system, but does not represent private-sector production. ANS data on beneficiaries indicate active private health plan coverage, but do not directly indicate where each woman undergoes mammography. These limitations do not invalidate the analysis, but they define the scope of the conclusions. The geographic restriction to the State of São Paulo should also be considered. The results reflect the organization, infrastructure, population distribution, and public-private composition of this state and should not be directly generalized to other Brazilian regions without additional analysis.

Another point concerns the age range adopted. The historical series used women aged 50 to 69 years until 2024 and women aged 40 to 74 years in 2025 and 2026. This decision makes it possible to assess expanded potential demand in recent years, but limits direct comparison between the entire historical period and the 2026 territorial snapshot. Therefore, the reduction in the capacity-to-demand ratio observed in 2025 and 2026 should be interpreted mainly as a consequence of the expanded reference population used in the denominator, rather than as evidence of a proportional reduction in installed or SUS-available capacity. For this reason, the results were interpreted as estimates of potential capacity in relation to the study reference population, rather than as a single normative measure for all mammography screening in SUS.

Taken together, the results suggest that the State of São Paulo presents aggregate potential sufficiency, but with relevant restrictions when public availability of equipment is considered. This type of result is useful for managers because it shifts the planning question from a simple count of mammography units to an analysis of where the equipment is located, to whom it is available, and how closely observed production approaches estimated capacity.

IV. Conclusion

The analysis indicated that the total potential capacity of mammography equipment in the State of São Paulo exceeded the estimated annual demand when all in-use units were considered. However, this potential sufficiency was not maintained when the analysis was restricted to equipment available to SUS in relation to the total reference population, a scenario in which 17 of the 19 RRAS had ratios below 1.0.

In the complementary scenario excluding women with active private health insurance, the state-level ratio between SUS capacity and estimated demand was slightly above 1.0. Nevertheless, eight RRAS remained below this threshold, indicating persistent regional inequalities even after adjustment for private health insurance coverage.

The assessment of mammography equipment sufficiency should not rely solely on the absolute number of units. To support technological and health care planning, it is necessary to consider SUS affiliation, observed production, reference population, territorial distribution, and the limitations of public information systems. Future studies may incorporate road travel time, waiting time, workforce distribution, equipment downtime, operating shifts, reporting capacity, utilization rates, and examination quality, thereby bringing potential-capacity estimates closer to actual access and service availability conditions.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil (CAPES), Finance Code 001, and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant 309464/2022-6 to HTM.

References

[1] Instituto Nacional de Câncer, *Controle do câncer de mama no Brasil: dados e números 2025*. Rio de Janeiro, Brasil: INCA, 2025.

[2] Ministério da Saúde, Secretaria de Atenção Especializada à Saúde, Departamento de Atenção ao Câncer, Coordenação-Geral de Prevenção e Controle do Câncer, "Nota Técnica nº 626/2025-CGCAN/DECAN/SAES/MS," Brasília, Brasil, 2025.

[3] P. Amaral, L. Luz, F. Cardoso, and R. Freitas, "Distribuição espacial de equipamentos de mamografia no Brasil," *Revista Brasileira de Estudos Urbanos e Regionais*, vol. 19, no. 2, pp. 326–341, 2017, doi: 10.22296/2317-1529.2017v19n2p326.

[4] D. R. Xavier, R. A. D. Oliveira, V. P. Matos, F. Viacava, and C. C. Carvalho, "Cobertura de mamografias, alocação e uso de equipamentos nas Regiões de Saúde," *Saúde em Debate*, vol. 40, no. 110, pp. 20–35, 2016.

[5] E. Saes-Silva, Y. P. Vieira, V. S. F. Viero, J. Q. S. Rocha, and M. O. Saes, "Tendência de desigualdades na realização de mamografia nas capitais brasileiras nos últimos dez anos," *Ciência & Saúde Coletiva*, vol. 28, no. 2, pp. 397–408, 2023, doi: 10.1590/1413-81232023282.07742022.

[6] Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, *Atlas da Radiologia no Brasil 2025*. São Paulo, Brasil: CBR, 2025.

[7] H. Wickham and G. Grolemund, *R para Data Science*. Rio de Janeiro, Brasil: Alta Books, 2019.

[8] Ministério da Saúde, DATASUS, "Cadastro Nacional de Estabelecimentos de Saúde," 2026. [Online]. Available: https://cnes.datasus.gov.br/. Accessed: Jun. 25, 2026.

[9] Ministério da Saúde, DATASUS, “Produção Ambulatorial (SIA/SUS),” 2026. [Online]. Available: https://datasus.saude.gov.br/acesso-a-informacao/producao-ambulatorial-sia-sus/. Accessed: Jun. 25, 2026.
[10] Agência Nacional de Saúde Suplementar, “Dados e indicadores do setor,” 2026. [Online]. Available: https://www.gov.br/ans/pt-br/acesso-a-informacao/perfil-do-setor/dados-e-indicadores-do-setor. Accessed: Jun. 25, 2026.
[11] Ministério da Saúde, DATASUS, “Informações de saúde TABNET,” 2026. [Online]. Available: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/. Accessed: Jun. 25, 2026.
[12] Instituto Nacional de Câncer, Coordenação de Prevenção e Vigilância, “Nota Técnica: Revisão do parâmetro para cálculo da capacidade de produção do mamógrafo simples,” Rio de Janeiro, Brasil, 2015.
[13] Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Regulação, Avaliação e Controle de Sistemas, *Critérios e parâmetros para o planejamento e programação de ações e serviços de saúde no âmbito do Sistema Único de Saúde*. Brasília, Brasil: Ministério da Saúde, 2015.
[14] Comissão Intergestores Bipartite do Estado de São Paulo, “Deliberação CIB nº 84, de 26 de agosto de 2025,” *Diário Oficial do Estado de São Paulo*, São Paulo, Brasil, 2025.

QUALIRAD: Progressive Web Application for Quality Control of Diagnostic Imaging Equipment in Accordance with RDC 611/2022

Eliseu Cordeiro de Brito
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
ecb2@cin.ufpe.br

Gustavo Isidio dos Santos Filho
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
gisf@cin.ufpe.br

Júlia Nunes de Araujo Silva
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
jnas2@cin.ufpe.br

Sara Simone Emilay de Araujo Pereira
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
sseap@cin.ufpe.br

Rafael Mourato Dantas Vilar
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
rmdv@cin.ufpe.br

Alex Ramos da Silva
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
ars6@cin.ufpe.br

Tiago da Silva Jornada
HU Brasil - Hospital das Clínicas UFPE
Universidade Federal de Pernambuco
Recife, Brazil
tiago.jornada@hubrasil.gov.br

Maria Karoline da Silva Andrade
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brazil
karoline.andrade@ufpe.br

Maíra Araújo de Santana
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
maira.araujosantana@ufpe.br

***Abstract*—Quality control (QC) tests on diagnostic imaging equipment are mandatory in Brazil, in accordance with Anvisa's RDC No. 611/2022 and Normative Instructions (INs) 90 to 97, and are essential to guarantee adequate diagnostic images with the lowest possible radiation exposure. In health service practice, however, the recording of these tests lacks standardization, often being performed in spreadsheets, printed forms, or separate reports, which compromises the traceability required by legislation, leads to operational failures, and limits the consolidated analysis of results. Medical equipment management software available on the market does not incorporate, in its structure, the specific QC tests for imaging modalities, whose protocols involve their own technical parameters, tolerances, and frequency. To address this gap, this paper presents QUALIRAD, a progressive web application (PWA) that translates the QC requirements of RDC 611/2022 into typed checklists with automatic compliance calculation, operating entirely offline. The tool currently covers 4 of the 8 equipment modalities regulated by INs 90 to 97; the remaining four are planned (one under development and three on the roadmap). Its checklists were developed and reviewed in collaboration with a medical physics professional, ensuring theoretical adherence to the applicable regulations. As a proof of concept, QUALIRAD demonstrates the feasibility of standardizing test records, representing an improvement over manual documentation practices and enabling the temporal tracking of equipment performance, contributing to the strengthening of Quality Assurance Programs, the regulatory compliance of radiology services, and data-driven decision-making.**

***Index Terms*—Quality control, Imaging equipment, Diagnostic radiology, ANVISA, Progressive Web Application.**

I. Introduction

In Brazil, the Board Resolution (RDC) No. 611, of March 9, 2022, of the National Health Surveillance Agency (Anvisa), establishes that diagnostic or interventional radiology services must implement a Quality Assurance Program and perform acceptance and performance tests, maintaining records of the respective reports and technical opinions [1]. This resolution is complemented by Normative Instructions (INs) 90 to 97, which detail the technical parameters, acceptance criteria, and minimum test frequency for each equipment modality [2]–[9]. These tests allow the verification of technical parameter compliance, ensuring images with the required quality through the lowest radiation dose reasonably possible to the patient, and enabling the early identification of discrepancies that could compromise diagnostic accuracy or the safety of patients and workers. Despite the regulatory requirement and clinical relevance of these procedures, the recording of quality control (QC) tests still lacks standardization in the practice of diagnostic services. Results are frequently documented in electronic spreadsheets, printed forms, or loose reports without a uniform data structure, which hinders the traceability required by health surveillance legislation [1], leads to operational recording failures—such as the loss of reports, incomplete entries, and lack of frequency control and limits the capacity for consolidated analysis of results over time. The World Health Organization (WHO) highlights, in its global strategy on digital health, that

the adoption of digital technologies strengthens the integration of health information, improves decision-making, and enhances the quality of services provided [12]. Complementing this, Kruse et al. (2014) [13] conducted a literature review on health information technologies and concluded that their adoption is predominantly associated with error reduction, greater traceability of records, and increased efficiency in clinical processes. In the market, several software solutions aimed at medical equipment management are available, covering the registration of interventions such as preventive and corrective maintenance, electrical safety tests, and calibrations. However, these solutions do not include, within their structure, the recording and management of specific QC tests for diagnostic imaging equipment, whose protocols involve modality-specific parameters and require systematic tracking of tolerances and periodicities defined in official regulations. Given this gap, this paper presents **QUALIRAD**, a tool for managing QC tests of diagnostic imaging equipment. The main contributions are:

- A model that translates the QC requirements of RDC 611/2022 and INs 90 to 97 into typed checklists with **automatic compliance calculation**, reducing the subjectivity of result interpretation [1]–[9];
- An *offline-first* architecture based on PWA, inspired by Mozilla Developer Network recommendations on progressive web applications [14] and by studies demonstrating its application in healthcare systems and environments with limited connectivity [16], [18];
- A deployment case study in a diagnostic radiology service, covering 4 of the 8 regulated modalities, with theoretical content validity ensured by the collaboration of a medical physics professional. A practical validation phase is scheduled for the second half of 2026 in university hospitals, where medical physics professionals from accredited institutions will evaluate the tool in real-world scenarios.

Finally, this paper is organized as follows. Section II presents related works, contextualizing existing solutions for medical equipment management and the use of progressive web applications in healthcare. Section III describes the materials and methods employed, including the regulatory framework, requirements gathering, the compliance model, application architecture, and the report generation process. Section IV presents the results obtained from the tool's development. Section V discusses the contributions, limitations, and implications of the proposed solution. Lastly, Section VI presents the conclusions of the work and perspectives for future work.

II. Related Works

Medical equipment maintenance management systems are widely used to record maintenance, calibrations, and electrical safety tests; however, they do not cover specific quality control tests for diagnostic imaging equipment, whose parameters and frequency are defined by regulatory standards. In addition to regulatory requirements, the literature highlights the importance of standardizing quality control tests. Papp (2025) [25] presents a comprehensive approach to quality management in diagnostic imaging, describing QC protocols for modalities such as radiography, mammography, computed tomography, and magnetic resonance imaging. Complementing this perspective, Hasford et al. (2024) [27] implemented a harmonized International Atomic Energy Agency (IAEA) protocol in 108 diagnostic imaging devices distributed across five African countries, demonstrating that performance tests allow the identification of non-conformities that could otherwise go unnoticed during routine operation. In the context of digital health, the World Health Organization (WHO) presents, in its global strategy on digital health, guidelines for the adoption of digital technologies capable of strengthening healthcare systems through data integration, care quality improvement, and decision-making support [12]. Complementing this vision, Kruse et al. (2014) [13] conducted a systematic review on health information technologies and observed that most studies report benefits related to improving the quality of care, increasing operational efficiency, and reducing clinical errors resulting from process computerization. In the field of web applications, Aboarab et al. (2024) [16] present a *Progressive Web App* (PWA) aimed at the visualization and reconstruction of DICOM medical images, exploring features such as installability and offline operation. Meanwhile, Azhar and Mohan (2022) [17] propose in their work a PWA for efficient real-time communication between doctors and patients, demonstrating that this approach enables the development of platform-independent applications with usability similar to native applications. Furthermore, Setiawan et al. (2024) [18] explore the implementation of PWAs through a combination with microservices to improve application performance, scalability, and accessibility. Regarding PWA components, the Mozilla Developer Network (MDN) documentation [14] highlights the role of *service workers* in caching, intercepting network requests, and executing functionalities without a continuous Internet connection. Similarly, Parashuram et al. (2016) [15] present an approach to utilizing *service workers* beyond traditional caching, demonstrating strategies that increase the availability and reliability of web applications. Although these studies highlight the benefits of digitalizing healthcare processes and utilizing PWAs, none of them address the management of quality control tests in diagnostic imaging equipment considering the parameters, tolerances, and frequency defined by ANVISA regulations. Therefore, QUALIRAD distinguishes itself by integrating these regulatory requirements into a PWA application, offering structured support for the execution and recording of QC routines in diagnostic radiology. The importance of continuous monitoring of results is also highlighted by El Kafhali et al. (2025) [26], who conducted a four-year longitudinal study on PET/CT systems and observed that small variations in performance parameters allow the identification of gradual equipment degradation. Similarly, Wigati et al. (2025) [28] investigated the use of the Noise Power Spectrum (NPS) in digital mammography, demonstrating that quantitative metrics make it possible to detect changes in detector performance and support quality control tests.

III. Materials and Methods

The development of QUALIRAD followed an iterative and participatory process, organized in successive cycles of construction and evaluation, in which the regulatory foundation and collaboration with the healthcare service preceded and guided each increment of the tool. This approach aligns with the principles of Design Science Research [11], where an artifact is designed to solve a real-world problem and subjected to evaluations that feed back into its improvement. Fig. 1 synthesizes the adopted method, from the regulatory foundation to field application, with content validation by the medical physicist.

A. Development Method

The work began with the real needs of a diagnostic imaging service, identified through visits to the facility to understand the routine—specifically how QC reports were produced at that time (using electronic spreadsheets and printed forms), the main bottlenecks, and how the solution would integrate into daily practice. In parallel, the applicable regulations (RDC 611/2022 [1] and INs 90 to 97 [2]–[9]) were studied. Based on this understanding, versions of the functional artifact were iteratively built and submitted to evaluations by medical physics and clinical engineering professionals regarding the usability and adequacy of the implemented QC tests until the version presented in this study was reached, which is suitable as a proof of concept and will undergo practical testing in routine use.

B. Regulatory Foundation and Requirements Gathering

Requirements gathering was based on three complementary sources: (i) direct analysis of RDC 611/2022 [1] and associated INs, mapping the mandatory test types for each modality; (ii) interviews and meetings with the healthcare service, which revealed the actual routine and its bottlenecks; and (iii) analysis of the spreadsheets, forms, and reports effectively used to record the performed tests. The eight INs cover distinct modalities; Table I presents this mapping and indicates which modalities are currently covered by QUALIRAD. Of the eight regulated modalities, the application covers those associated with INs 90 to 93 and IN 97 (the latter under development); INs 94 to 96 (dental radiology and ultrasonography) remain as future scope.

Based on this mapping, the supported test types per modality were identified. The number of tests varies depending on the equipment, from three (bone densitometry) to nine (mammography). Table II lists the test types, organized by frequency, planned in the application.

Based on the regulation, the functional requirements were consolidated: (i) nominal identification of the user, without complex authentication in this first version; (ii) guided workflow Login → Equipment category → Equipment → Test type → Checklist → Report; (iii) support for three checklist item formats, detailed in Section III-C; (iv) mandatory equipment verdict upon completion; (v) PDF and CSV report generation without connection dependency; and (vi) local persistence with draft recovery. An entirely offline solution was chosen to ensure the practical viability of the tool even in areas with unstable connectivity, a common situation in diagnostic radiology services [16]–[18].

Table I
Modalities regulated by RDC 611/2022 and coverage in QUALIRAD

IN	Regulated modality	In the application	Coverage
90	Conventional medical radiography	Fixed X-ray; Mobile X-ray	Yes
91	Fluoroscopy and interventional radiology	Angiographic; C-arm	Yes
92	Mammography	Mammography system	Yes
93	Computed tomography	CT scanner	Yes
94	Extraoral dental radiology	—	Future
95	Intraoral dental radiology	—	Future
96	Diagnostic or therapeutic ultrasound	—	Future
97	Nuclear magnetic resonance	Magnetic resonance	Under dev.
—[a]	Bone densitometry (RDC 611, no dedicated IN)	Bone densitometry	Yes

Source: own elaboration based on RDC 611/2022 and INs 90-97 (ANVISA, 2021).
Note: [a]Bone densitometry is within the scope of RDC 611/2022, which mentions it explicitly, but does not have a dedicated IN among INs 90-97; its tests were structured based on manufacturer protocols and medical physics literature.

Table II
Supported quality control test types

Test Type	Description and Frequency
Acceptance	Performed upon initial installation of the equipment.
Daily	Quick check of critical parameters for each day of use (e.g., densitometer, mammography system).
Weekly	Weekly verification of complementary parameters (e.g., CT).
Monthly	Monthly performance evaluation (e.g., mammography system).
Semiannual	Periodic evaluation every six months.
Annual	Comprehensive annual review of all critical parameters.
Quadrennial	Long-term test required every four years.
After repairs	Mandatory verification after any technical intervention.
After modifications	Evaluation after configuration changes or system updates.

Source: own elaboration based on RDC 611/2022 and INs 90–97.

C. Compliance Model

The core of QUALIRAD is the structured representation of checklist items derived from the INs. Each item is typed into one of three categories, allowing compliance assessment to be standardized and, in quantitative cases, automated:

- *standard*: qualitative item, whose status is manually defined by the evaluator as Compliant or Non-compliant;
- *measured*: quantitative item with a target value and percentage tolerance, evaluated by relative deviation;
- *threshold*: quantitative item with a fixed threshold evaluated by a comparison operator.

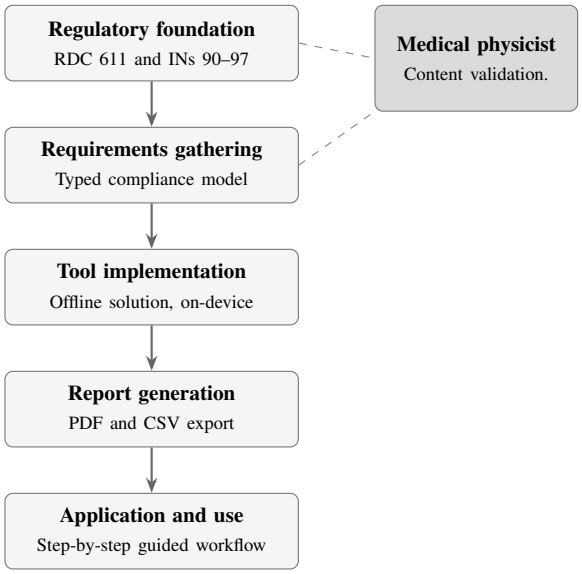


Figure 1. QUALIRAD development method, from regulatory foundation to field application, with content validation by the medical physicist.

As an example, the "tube voltage indicator accuracy" item (fixed X-ray, IN 90) is a *measured* item with a tolerance of 10% and a restriction of 20%. For a target of $v_t = 80$ kV and a reading of $v_m = 85$ kV, the relative deviation is $\delta = |85 - 80|/80 = 6.25\%$; since $\delta \leq \tau$, the item is classified as Compliant.

For a *measured* item with target value v_t, measured value v_m, tolerance τ and restriction ρ (percentages, with $\tau \leq \rho$), the relative deviation is defined as $\delta = |v_m - v_t|/v_t$ and the status as:

$$\text{status} = \begin{cases} \text{Compliant}, & \delta \leq \tau \\ \text{With restriction}, & \tau < \delta \leq \rho \\ \text{Non-compliant}, & \delta > \rho \end{cases} \tag{1}$$

For a *threshold* item with threshold L and operator $\oplus \in \{\leq, \geq, <, >, =\}$, the item is Compliant if the proposition $v_m \oplus L$ is true and Non-compliant otherwise; an optional intermediate threshold defines the With restriction range. As an example, the "compressed breast thickness indication" item (mammography, IN 92) is a *threshold* item with a tolerance limit of ≤ 5 mm and a restriction limit of > 10 mm: a reading of 2 mm falls within the tolerance limit and is therefore classified as Compliant, whereas a value above the restriction limit would be Non-compliant. Target values, tolerances, and thresholds were extracted directly from RDC 611/2022 and the respective INs [1]. Since the INs establish only the tolerance limits without prescribing the calculation method to be used, conventions established in reference protocols in medical physics were adopted [10], [22]–[24]: relative deviation for accuracy tests and, for reproducibility and linearity, the respective coefficients of variation and linearity. In cases where the regulation refers the baseline value to the equipment or manufacturer, the target value is provided by the evaluator at the time of the test; in other cases, it is pre-configured based on the IN. For instance, the tube voltage indicator accuracy limit ($\leq 10\%$) is taken directly from IN 90, whereas the water noise test in computed tomography is evaluated against a baseline standard deviation established at acceptance or provided by the manufacturer, which the evaluator enters manually, as the regulation refers this reference value to the equipment. The N/A button is enabled only for *standard* items in "after repairs" and "after modifications" test types, where some items may not apply depending on the intervention; *measured* and *threshold* items are always evaluated automatically. Finally, the equipment verdict is derived from the set of recorded statuses.

Since the classification of an item may depend on a comparison exactly at the tolerance limit, the evaluation incorporates controlled rounding prior to comparison, preventing floating-point representation errors—where a value mathematically equal to the limit is computed as slightly higher—from incorrectly failing a compliant item. This is a deliberate design decision, as incorrect compliance classification has direct consequences on equipment and patient safety.

The checklists implemented in QUALIRAD were structured based on the requirements of RDC 611/2022 and the respective Normative Instructions, also following quality management principles described by Papp [25], which organizes QC procedures according to the different diagnostic imaging modalities.

D. Content Validation by Specialist

The theoretical content validity of the tool, i.e., the adherence of the tests, tolerances, and frequencies to the applicable standards, was ensured at this stage through collaboration with medical physics and clinical engineering professionals experienced in radiology, diagnostic radiology, and quality control testing. The validation was conducted iteratively through five face-to-face and remote meetings, during which the specialists evaluated the checklists, compliance criteria, and tool behavior, indicating adjustments that were recorded and incorporated in subsequent development cycles. Furthermore, a practical validation phase will begin in the second half of 2026. This phase will take place in university hospitals and will be conducted by medical physics professionals from accredited hospitals, who will provide formal feedback reports to guide further tool adjustments and refinements.

E. Architecture and Technologies

The application was built as a client-side PWA, without a *backend* server in this version, which was implemented as a proof of concept. This decision reduces the required infrastructure, allows deployment via static servers, and keeps the architecture extensible for a future migration to a REST API. Offline operation is achieved using *service workers*, which, according to the Mozilla Developer Network (MDN), intercept network requests and cache resources to enable the application to run without a continuous connection [14]. Parashuram et al. (2016) [15] complement this approach by demonstrating that *service workers* offer advanced caching mechanisms that increase the availability and reliability of web applications. The checklist definitions are kept separate

from the state logic, so that adding new items or test types only requires editing the data, without modifying the business logic. The persistence of results and drafts in progress is local, and report generation occurs entirely in the browser. In this first version (MVP), no authentication mechanism was implemented, and user identification is merely nominal, so the application does not handle personal or sensitive data that would require encryption. All data produced during use, that is, drafts in progress and finalized reports, are stored entirely locally on the user's device, through the storage provided by the browser itself, without transmission to external servers. As a direct consequence of this decision, the records are restricted to the browser and to the device on which they were generated, so there is neither shared access among multiple users nor synchronization among different devices, and no data reconciliation occurs when connectivity is restored, since operation does not depend on a remote server at any point. Table III summarizes the main technologies and the reasons for their choice.

Table III
MAIN TECHNOLOGIES OF THE APPLICATION

Technology	**Layer**	**Justification**
React + TypeScript [19]	Framework / UI	Reusable components and static typing
Vite [20]	Build	Modern tooling with PWA support
Service Worker [14], [15]	Offline	Resource caching and offline functionality
Local persistence	Data	Zero infrastructure, sufficient for the MVP
jsPDF [21]	Report	Client-side PDF generation, works offline

Source: own elaboration.

F. Verification

The correctness of the compliance calculation was verified using two complementary strategies. First, through automated tests that exercise the limits of tolerance and restriction ranges for *measured* and *threshold* items, including boundary cases sensitive to rounding errors. Second, by comparing the reports generated by the tool with QC reports previously prepared by the service, verifying the equivalence of the verdicts and the fidelity of the recorded values. This verification ensures that the automated classification reproduces the expected judgment prior to evaluation in real-world use.

The automated boundary test suite covers 26 cases for the *measured* and *threshold* classification functions (15 for *measured*, including deviations below, at, and above τ and ρ, and floating-point boundary cases; 11 for *threshold*, covering both comparison directions and the intermediate restriction threshold), plus 15 additional cases for the controlled-rounding routine used before comparison. All 41 cases pass.

G. Report Generation and Export

Upon completing the checklist, the user is directed to the report screen, which displays the equipment data, the evaluator, date and time, test type, item-by-item results with their respective compliance status, recorded observations, and the final equipment verdict. The report can be exported in PDF or CSV format; generation is performed programmatically in the browser [21], without any server calls, thereby preserving offline operation.

IV. RESULTS

QUALIRAD is functional for 4 of the 8 regulated equipment modalities (Table I), with the solution for magnetic resonance currently under development. Within each modality, it is possible to select the equipment present in the institution, and all planned test types (Table II) have been adapted to each device. The complete implemented and operational workflow consists of: Login $\rightarrow$ Category $\rightarrow$ Equipment $\rightarrow$ Test Type $\rightarrow$ Checklist $\rightarrow$ Report. The application is installable as a PWA on mobile devices and desktops and operates offline after the first access.

The checklists and respective compliance criteria were developed and reviewed in collaboration with a qualified professional, which ensures the tool's theoretical content validity, i.e., the adherence of the tests, tolerances, and frequency to applicable standards. Continuous improvements will rely on the upcoming practical validation in the second half of 2026, which will yield formal reports from field specialists. Figures 2, 3, and 4 illustrate, respectively, the home screen, the equipment and test selection environment, and the generated report with export options.



Figure 2. Home screen of the tool.

Table IV summarizes the comparison between the usual process in most institutions and the developed solution. It should be noted that this version constitutes a proof of concept: no quantitative data on operational use (filling time, error rate, or inter-observer agreement) have been collected so far, which is discussed in Section VI as a limitation and future work.

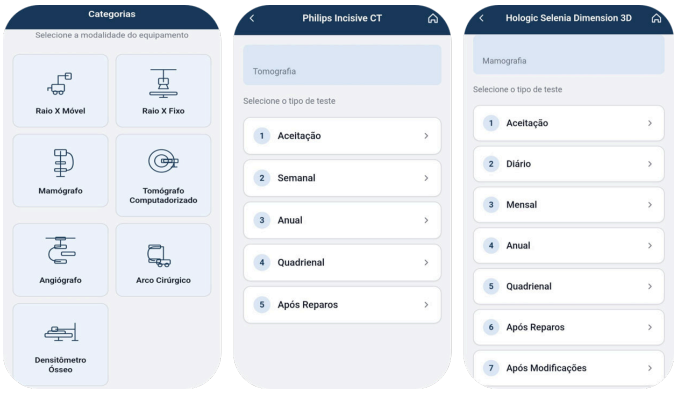


Figure 3. Environment for selecting the equipment to be tested and the test to be performed.

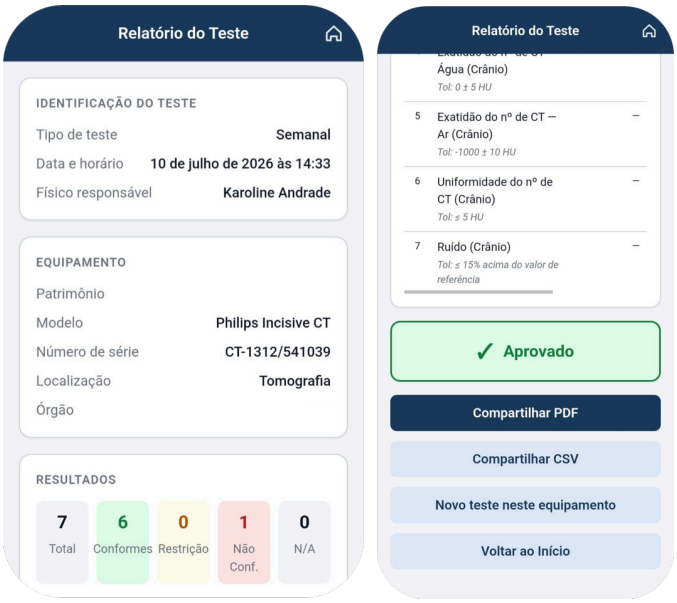


Figure 4. Test report and export options.

Table IV
COMPARISON BETWEEN MANUAL PROCESS AND QUALIRAD

Usual situation	**With QUALIRAD**
Paper forms, susceptible to loss	Persistent digital record
Spreadsheets disconnected from the workflow	Step-by-step guided workflow in the application
No standardization across tests and equipment	Checklists structured by modality and test type
Dependence on connectivity for access	Works offline as an installed PWA
Report manually compiled after testing	Automatic PDF generation upon checklist completion

Source: own elaboration.

V. DISCUSSION

The development of QUALIRAD demonstrated that it is possible to fully digitalize the QC process for imaging equipment using exclusively client-side technologies, with zero server infrastructure in the MVP. The main gain lies not in the technology itself, but in the direct translation of regulatory requirements into an operational instrument: the standardization of checklists by modality and frequency and the persistent digital record directly target the weaknesses of paper- and spreadsheet-based processes—loss of reports, incomplete entries, and lack of frequency control—strengthening the traceability required by RDC 611/2022 [1]. The automatic compliance calculation for *measured* and *threshold* items reduces the subjectivity of interpretation and makes results more objective and reproducible.

The PWA approach proved suitable for the context of use: offline support eliminates dependency on connectivity during field inspections [16], [18], and in-browser installability bypasses the need for app store publication [17]. Separating checklist definitions from state logic enhances extensibility, which is crucial for adapting to future regulatory updates.

Maintaining the test history facilitates the identification of trends and gradual changes in equipment performance, an aspect observed by El Kafhali et al. (2025) [26] in PET/CT systems. Furthermore, the structured storage of results benefits the analysis of quantitative parameters used in modern QC tests, such as those discussed by Wigati et al. (2025) [28].

VI. CONCLUSION

We presented QUALIRAD, a PWA that structures the QC of diagnostic imaging equipment in accordance with RDC 611/2022 and INs 90 to 97, using typed checklists and automatic compliance calculations. As a proof of concept with theoretical content validity ensured by a qualified professional, the tool highlights the feasibility of standardizing test records and represents an improvement over manual documentation practices.

This work presents limitations that guide its future development. First, the evaluation conducted consists of content validation by an expert; empirical studies of usability, execution time, inter-observer agreement, or error reduction quantifying the benefits under real-world conditions have not yet been carried out. Second, persistence is local, occurring exclusively in the browser storage of the device used, without any server component, which restricts the records to that device, prevents multi-user centralization, and limits institutional audits. Third, there is no robust authentication mechanism, since the login is nominal and simple by design decision, as described in Section III-B, which is adequate for a proof of concept but insufficient for environments requiring individual accountability and access control. Fourth, the magnetic resonance modality and those regulated by INs 94 to 96 (dental radiology and ultrasonography) have not yet been implemented.

Priority future works include: (i) an empirical evaluation of usability and performance in real-world use, with practical

validation planned for the second semester of 2026 in university hospitals. This evaluation will consider not only objective functional criteria, such as the coverage of all prescribed tests and the adequacy of field completion and ordering, but also qualitative feedback from healthcare professionals regarding the clarity, usability, and suitability of the tool in their routine practices; (ii) storage of test history with visualization and filters; (iii) photographic recording of observed conditions per item; (iv) migration from local persistence to a REST API with a database, enabling centralization and audits; (v) report customization with logos and digital signatures; and (vi) expansion of covered modalities, including magnetic resonance [24] and, subsequently, ultrasonography and dental radiology. Beyond being a management tool, QUALIRAD serves educational and continuous training purposes, so a future improvement involves incorporating step-by-step execution guides for each test, promoting practices that meet the standards expected by institutions and regulatory bodies.

REFERENCES

[1] AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). Resolução da Diretoria Colegiada RDC n.º 611, de 9 de março de 2022. Brasília: ANVISA, 2022.

[2] ANVISA. Instrução Normativa IN n.º 90, de 27 de maio de 2021. Requisitos sanitários para radiografia médica convencional. Brasília: ANVISA, 2021.

[3] ANVISA. Instrução Normativa IN n.º 91, de 27 de maio de 2021. Requisitos sanitários para fluoroscopia e radiologia intervencionista. Brasília: ANVISA, 2021.

[4] ANVISA. Instrução Normativa IN n.º 92, de 27 de maio de 2021. Requisitos sanitários para mamografia. Brasília: ANVISA, 2021.

[5] ANVISA. Instrução Normativa IN n.º 93, de 27 de maio de 2021. Requisitos sanitários para tomografia computadorizada médica. Brasília: ANVISA, 2021.

[6] ANVISA. Instrução Normativa IN n.º 94, de 27 de maio de 2021. Requisitos sanitários para radiologia odontológica extraoral. Brasília: ANVISA, 2021.

[7] ANVISA. Instrução Normativa IN n.º 95, de 27 de maio de 2021. Requisitos sanitários para radiologia odontológica intraoral. Brasília: ANVISA, 2021.

[8] ANVISA. Instrução Normativa IN n.º 96, de 27 de maio de 2021. Requisitos sanitários para ultrassonografia diagnóstica e intervencionista. Brasília: ANVISA, 2021.

[9] ANVISA. Instrução Normativa IN n.º 97, de 27 de maio de 2021. Requisitos sanitários para ressonância magnética nuclear. Brasília: ANVISA, 2021.

[10] BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. Radiodiagnóstico médico: desempenho de equipamentos e segurança. Brasília: Ministério da Saúde, 2005. 104 p. (Série A. Normas e Manuais Técnicos). ISBN 85-334-1040-9.

[11] HEVNER, Alan R.; MARCH, Salvatore T.; PARK, Jinsoo; RAM, Sudha. Design Science in Information Systems Research. MIS Quarterly, v. 28, n. 1, p. 75–105, 2004.

[12] WORLD HEALTH ORGANIZATION. Global strategy on digital health 2020–2025. Geneva: WHO, 2021. Available at: https://iris.who.int/bitstream/handle/10665/344249/9789240020924-eng.pdf. Accessed on: Jul. 13, 2026.

[13] KRUSE, Clemens Scott; KROGSTAD, Nicholas; GEORGE, J. Ryan; et al. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Yearbook of Medical Informatics, v. 23, n. 1, p. 47–54, 2014.

[14] MOZILLA DEVELOPER NETWORK. Progressive Web Apps (PWAs). Available at: https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps. Accessed on: Jul. 13, 2026.

[15] PARASHURAM, N.; KUMAR, S.; CHAUDHARY, S. Service Workers: Beyond the Cache. In: Proceedings of the 25th International Conference on World Wide Web, 2016.

[16] ABOARAB, Mohammed A.; HASSAN, M.; ALI, S. Advancing Progressive Web Applications to Leverage Medical Imaging for Visualization of Digital Imaging and Communications in Medicine and Multiplanar Reconstruction: Software Development and Validation Study. JMIR Medical Informatics, v. 12, e63834, Dec. 9, 2024. DOI: 10.2196/63834.

[17] AZHAR, Hannan Bin; MOHAN, Joseph Thomas. Progressive Web App for Real-time Doctor-Patient Communication and Searchable Health Conditions. In: E-Health and Bioengineering Conference (EHB), 2022. IEEE, 2022. DOI: 10.1109/EHB55594.2022.9991288.

[18] SETIAWAN, Budi; WIJAYA, Andi; PRATAMA, Rizky. The Implementation of PWA (Progressive Web App) Technology in Enhancing Website Performance & Mobile Accessibility. Buletin Pos dan Telekomunikasi, v. 22, n. 1, p. 25–36, 2024.

[19] REACT. React — The library for web and native user interfaces. Available at: https://react.dev. Accessed on: Jul. 13, 2026.

[20] VITE. Vite — Next generation frontend tooling. Available at: https://vitejs.dev. Accessed on: Jul. 13, 2026.

[21] JSPDF. jsPDF — Client-side JavaScript PDF generation. Available at: https://github.com/parallax/jsPDF. Accessed on: Jul. 13, 2026.

[22] AMERICAN COLLEGE OF RADIOLOGY (ACR). Phantom Test Guidance for Use of the Large MRI Phantom for the ACR MRI Accreditation Program. Reston, VA: American College of Radiology, 2018. Available at: https://www.acraccreditation.org/-/media/ACRAccreditation/Documents/MRI/LargePhantomGuidance.pdf. Accessed on: Jul. 13, 2026.

[23] AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE (AAPM). Phantoms for performance evaluation and quality assurance of CT scanners. Chicago: AAPM, 1977. (AAPM Report, n. 1). Available at: http://www.aapm.org/pubs/reports/rpt_01.pdf. Accessed on: Jul. 13, 2026.

[24] MAZZOLA, Alessandro André; HERDADE, Silvio Bruni; KOCH, Hilton Augusto; CARVALHO, Antonio Carlos Pires. Protocolo de testes de aceitação em equipamentos de imagem por ressonância magnética. Radiologia Brasileira, Rio de Janeiro, v. 38, n. 3, p. 195–204, 2005. Available at: https://doi.org/10.1590/S0100-39842005000300008. Accessed on: Jul. 13, 2026.

[25] PAPP, Jeffrey. *Quality Management in the Imaging Sciences*. 7. ed. St. Louis: Elsevier, 2025.

[26] EL KAFHALI, Morad; TAHMASBI, Marziyeh; BAGHOUS, Nor El Houda; et al. Longitudinal quality control analysis of PET/CT systems: analyzing system stability, image quality, and the performance of silicon photomultipliers (SiPMs). *Health and Technology*, v. 15, p. 703–715, 2025.

[27] HASFORD, Francis; KHELASSI-TOUTAOUI, Nadia; ATTALLA, Ehab; et al. Preliminary results of performance testing in diagnostic radiology facilities: Implementation of harmonized IAEA protocol for Africa. *Health and Technology*, v. 14, n. 1, p. 169–187, 2024.

[28] WIGATI, Kristina Tri; BOSMANS, Hilde; BINST, Joke; et al. In-depth look at the use of pixel variance and the noise power spectrum in digital mammography quality control. *Journal of Medical Imaging*, v. 12, n. 6, Art. 063502, 2025.

Risk-Based Maintenance Prioritization Framework for Medical Equipment in Brazilian Public Hospitals

Bianca Vitória Silva Pina
Institute of Technology
Federal University of Pará
Belém, PA, Brazil
bianca.pina@itec.ufpa.br

Davi Gabriel dos Santos Oliveira
Faculty of Medicine
Federal University of Goiás
Goiânia, GO, Brazil
oliveira_davi@ufg.br

***Abstract*—Equipment failure in Brazilian public hospitals affects patient safety, and the problem is sharpest in facilities operating under the Unified Health System (SUS), where maintenance budgets are tight and downtime translates quickly into cancelled procedures. We propose a Risk-Based Maintenance Priority Index (RBMPI) built from four criteria: failure frequency, clinical severity of failure, substitute device availability, and a regulatory weight derived from the Brazilian Health Regulatory Agency (ANVISA) device risk class. Frequency and severity are expert-elicited ordinal scores informed by the aggregate composition of ANVISA post-market surveillance notifications and by clinical engineering references. They are not measured device-specific rates, which Brazilian public reporting does not disaggregate. Across ten equipment categories common in SUS hospitals the index yields three tiers: Critical, Essential, and Elective. Infusion pumps, mechanical ventilators, syringe pumps, and defibrillators fall in the Critical tier. A Monte Carlo analysis perturbing every scored parameter at once keeps those four in tier in 88% to 100% of runs, while five devices sitting near a boundary change tier often enough to be treated as provisional. Three further tests probe the model's own construction: natural-breaks clustering supports the Critical boundary but finds no data support for the lower one; varying the redundancy multiplier leaves rank order untouched; and an equal-weight additive analogue reproduces the multiplicative ranking at $\rho = 0.964$, which shows the aggregation rule contributes far less than the criterion set. Against the Fennigkoh–Smith index the RBMPI agrees at the extremes ($\rho = 0.835$) and reprioritizes devices that have a functional substitute, a dimension the additive model cannot express. The RBMPI is offered as a preliminary framework for resource-constrained settings. It has not been validated prospectively in hospitals.**

***Index Terms*—Medical equipment maintenance, clinical engineering, patient safety, ANVISA, Brazilian public hospitals, risk-based maintenance, technovigilance, SUS.**

I. Introduction

Maintenance of medical equipment in Brazilian public hospitals is a persistent problem, and it bears directly on whether diagnostic and therapeutic procedures can be delivered at all. An operational audit of nine federal hospitals by the Federal Court of Accounts (TCU) found equipment purchased in earlier years still uninstalled while the same units proceeded with fresh acquisitions, together with absent acquisition planning and deficient technical specification [1]. Field studies of clinical engineering practice in Brazilian public facilities report a substantial share of the installed base underutilized or inoperative, with missing spare parts and consumables among the recurring causes [2].

Clinical engineers in these settings face a narrow, recurring question: with limited resources, which equipment gets attention first? Lacking an objective rule, the answer tends to follow individual judgment or whichever failure is loudest that week, rather than a structured reading of risk and clinical impact [3].

Brazilian regulation offers a starting point. ANVISA operates a national post-market surveillance system in which hospitals, manufacturers, and health professionals report adverse events and technical complaints for medical devices [4]. Between 2014 and 2024 the system received 149,099 device-related notifications; of the 23,716 classified as adverse events, 57.2% involved Class III (high-risk) devices [5]. We are not aware of prior work using this source to inform maintenance prioritization in SUS facilities.

Established models were built elsewhere. The Equipment Management Index of Fennigkoh and Smith [3] assumes a well-resourced department and has no way to express redundancy in the installed base, which is precisely the constraint that binds in public health facilities. Sensor-driven predictive maintenance [6], [7] assumes instrumentation that most SUS hospitals do not have.

This paper proposes an index combining failure frequency, clinical severity, substitute availability, and regulatory risk class into a single score, yielding a three-tier prioritization scheme. The inputs are drawn from sources a facility already holds or can consult without new instrumentation. We treat the published scores as a starting profile to be recalibrated locally, not as a ready-to-deploy ranking.

II. Related Work and Regulatory Context

A. Medical Equipment Management in Brazil

RDC 509/2021 governs health technology management in Brazilian health services, setting requirements for preventive and corrective maintenance planning, equipment registration, and lifecycle documentation [9]. Compliance is uneven. It is weakest in smaller municipalities, where clinical engineering capacity is thin or absent [2]. The Ministry of Health has issued guidance on maintenance management for public facilities [11], and the Brazilian literature reports preventive

programmes routinely interrupted by budget, spare part availability, and the lack of a clinical engineer on staff [12], [2].

B. ANVISA Post-Market Surveillance System

The surveillance system coordinated by ANVISA monitors adverse events and technical complaints for devices in clinical use. Of the 149,099 notifications recorded between 2014 and 2024, 23,716 (15.9%) were adverse events and 125,383 (84.1%) technical complaints [5]. Class III devices account for 57.2% of the adverse events, so a large share of reported incidents involve devices capable of causing direct harm. These are aggregate figures. Public reporting does not break counts down by equipment category at the resolution a maintenance index would need, and that constraint shapes the scoring procedure described in Section III-C.

ANVISA sorts devices into four risk classes by intended use, invasiveness, and the consequences of failure: Class I (low), Class II (moderate), Class III (high), and Class IV (maximum). The class determines pre-market and post-market regulatory obligations [9].

C. Existing Prioritization Models

The Equipment Management Index (EM) of Fennigkoh and Smith [3] sums scores for function, physical risk, and maintenance requirements, admitting to the maintenance programme any device above a threshold. It has been widely adopted. It also carries no surveillance input and no representation of substitute availability, and both its additive structure and its inclusion threshold have been revisited since, notably by Wang and Levenson, who recast the function criterion as mission criticality and add a utilization term [13].

D. Multi-Criteria and Predictive Approaches

Prioritization has also been posed as a multi-criteria decision problem. Taghipour et al. [8] rank 26 devices from a hospital maintenance database using weighted criteria spanning function, mission criticality, age, risk, recall history, and maintenance requirements, then assign maintenance strategies by class. Sezdi [7] argues for differentiating strategy by device technology level. Radzi et al. [6] review 40 studies from 2016 to 2024 and find sensor-based monitoring dominant, which confirms how heavily this line of work depends on instrumentation.

Two things separate the RBMPI from these. It makes substitute availability an explicit criterion, so redundancy lowers priority. And it is computable without a maintenance database, sensors, or an expert panel performing pairwise weighting. The price is coarser inputs, which is why Sections IV-B and V-B test the classification rather than assert it.

III. Proposed Maintenance Priority Index

A. Index Structure

The Risk-Based Maintenance Priority Index combines four criteria:

$$\text{RBMPI} = F \times S \times (2 - A) \times W_r \tag{1}$$

where F is the failure frequency score (0–5), S the clinical severity score (1–5), A the substitute availability (0 = unavailable, 1 = available), and W_r the regulatory weight. The term $(2 - A)$ doubles the index for equipment with no functional substitute. A factor of two is a modeling decision, not a calibration against data; Section V-B reports what happens when it moves.

The product form follows the risk-priority-number logic of failure mode and effects analysis, where criticality is a product of severity and occurrence terms [14], and the multi-criteria scoring used in the equipment prioritization literature [8]. The choice carries an assumption worth naming. A product is non-compensatory in the limit: a device scoring zero on any criterion scores zero overall, whereas a weighted sum lets strength on one criterion offset weakness on another. It is also symmetric in F and S, so a rare but lethal failure and a frequent but harmless one receive the same index. Whatever asymmetry the model expresses comes from the tier thresholds, not from the functional form. Section V-B tests how much the choice of form actually changes the outcome here, and the answer is less than one might expect.

B. Regulatory Weight

W_r follows the ANVISA risk class: Class I ($W_r = 1.0$), Class II ($W_r = 1.1$), Class III ($W_r = 1.3$), Class IV ($W_r = 1.5$), raising priority for devices under stricter oversight. No device in this sample is Class IV, so the effective range here is 1.0 to 1.3. The Class IV level is retained for wider inventories. Section V-B tests three alternative schemes and quantifies how little of the classification any of them drive.

C. Failure Frequency Score

F is an expert-elicited ordinal score on a 0–5 scale. Two sources informed it: the aggregate composition of the Brazilian post-market surveillance notification base [5], which gives the overall volume and risk-class distribution of device notifications, and the failure-behaviour and maintenance-intensity patterns described in the clinical engineering literature [18], [11], [7]. Categories most often associated with reported incidents anchor the upper end of the scale. Both authors scored the ten categories independently against these sources and reconciled the six categories where their initial scores differed by discussion, without a formal consensus protocol such as Delphi. Inter-rater agreement was therefore not quantified, a limitation revisited in Section V-E.

The scores are not measured failure rates. Device-specific notification counts are not published at the resolution the index requires, so F ranks relative failure propensity across these ten categories and nothing finer. This is the weakest input in the model. Section IV-B quantifies what it costs the classification.

D. Clinical Severity Score

S scores the consequence of an unplanned failure during active use:

- 5: immediate risk to life (ventilator, infusion pump, defibrillator)

- 4: serious but not immediately life-threatening harm (monitor, syringe pump)
- 3: moderate impact with manageable consequences (pulse oximeter, electrosurgical unit)
- 2: limited clinical impact (electrocardiograph, ultrasound, patient bed)
- 1: no direct patient safety consequence

Assignments follow established clinical engineering references [3], [18] and the same two-rater procedure described above. No device here scores 1; the level is kept for generality. Like F, S is an ordinal judgment and is perturbed in the robustness analysis.

E. Substitute Availability

A records whether a functional replacement is typically reachable in a SUS hospital when the primary unit fails, coded from Ministry of Health equipment census data [10] and clinical engineering management literature [11], [2].

Binary coding is a real simplification. Actual substitutability depends on how many units exist, occupancy at the moment of failure, physical distance between the failed unit and its replacement, transfer time for the patient or the device, and the local operational criticality of the service involved. A single spare ventilator three floors away in a full ICU is not the same resource as a spare beside the bed. Section V-D sets out a graded alternative built from census counts, and the Monte Carlo analysis identifies this coding as the largest single source of tier instability.

F. Priority Classification

Three tiers were defined:

- **Critical** (RBMPI $\geq$ 20): maximum-frequency preventive maintenance, immediate corrective response, formal downtime reporting.
- **Essential** ($10 \leq$ RBMPI < 20): regular preventive maintenance, corrective response within 24 hours.
- **Elective** (RBMPI < 10): standard institutional schedule, corrective response within working hours.

G. Deriving the Thresholds

Thresholds placed by eye invite the objection that they were chosen to produce a desired answer. We therefore tested them against the score distribution using Jenks natural-breaks optimization [15], which partitions a univariate distribution to minimize within-class variance. With $n = 10$ the search is exhaustive, so the reported partitions are optimal rather than heuristic.

The evidence splits. The distribution contains one dominant gap of 18.2 index points, between the lowest Critical device (36.4) and the highest Essential device (18.2); every other gap is 12.0 points or smaller. The two-class solution (goodness of variance fit, GVF $= 0.870$) cuts exactly there, reproducing the Critical / non-Critical split. Any threshold in the interval $(18.2, 36.4]$ produces the same Critical tier, and the value 20 sits inside it.

The lower boundary has no such support. Gaps in the region below 18.2 run from 0.8 to 3.6 points, and the three-class solution (GVF $= 0.960$) prefers to subdivide the Critical cluster into $\{62.4, 54.6\}$ and $\{42.6, 36.4\}$ rather than split the lower six devices at all. On this distribution the Essential / Elective boundary at 10 is a maintenance-policy decision about response times, not a feature of the data. We state it as such. The Monte Carlo analysis points the same way: every device whose tier is unstable sits in that lower region.

All computations used Python 3.12 with Pandas [19], Matplotlib [20], and SciPy [21]. Table I lists every input, and Sections III-A through III-G specify the scoring rules, the perturbation distributions, and the iteration counts, so the analysis can be reproduced from the paper alone.

IV. Results

A. RBMPI Scores and Classification

Table I gives scores, ranks, and tiers for all ten categories. Four devices are Critical: infusion pumps (62.4), mechanical ventilators (54.6), syringe pumps (42.6), and defibrillators (36.4). All four carry severity 4 or 5, ANVISA Class III, and no coded substitute. That coding comes from national census data [10] and the clinical engineering literature [11], [2]; it describes a general SUS profile and was not verified facility by facility.

A word on precision. Index values are reported to one decimal so the arithmetic can be checked against Table I, not because the inputs support that resolution. The Monte Carlo analysis in Section IV-B gives a per-device interquartile width with a median of 5.2 index points and a median standard deviation of 6.6. Differences below roughly five points therefore fall inside the uncertainty of a single device and should not be read as a meaningful ranking gap. The pulse oximeter at 12.9 and the electrosurgical unit at 12.1 are tied for practical purposes. The infusion pump at 62.4 and the electrocardiograph at 4.8 are not.

TABLE I
Risk-Based Maintenance Priority Index by Equipment Category

Equipment	Components				Rank	RBMPI	P(tier)	Level
	F	S	A	W_r			%	
Infusion Pump	4.8	5	No	1.3	1	62.4	100	Critical
Mech. Ventilator	4.2	5	No	1.3	2	54.6	100	Critical
Syringe Pump	4.1	4	No	1.3	3	42.6	95	Critical
Defibrillator	2.8	5	No	1.3	4	36.4	88	Critical
Multipar. Monitor	3.5	4	Yes	1.3	5	18.2	62	Essential
Pulse Oximeter	3.9	3	Yes	1.1	6	12.9	76	Essential
Electrosurg. Unit	3.1	3	Yes	1.3	7	12.1	71	Essential
Ultrasound System	2.5	2	No	1.1	8	11.0	55	Essential
Bed/Stretcher	3.7	2	Yes	1.0	9	7.4	78	Elective
Electrocardiograph	2.2	2	Yes	1.1	10	4.8	93	Elective

[a] F: failure frequency (0–5); S: clinical severity (1–5). Both are expert-elicited ordinal estimates (Section III-C).
[b] A: substitute available; W_r: regulatory weight.
[c] Values are relative priority scores, not absolute risk. Differences below ˜5 points are within simulated uncertainty.
[d] P(tier): Monte Carlo probability of retaining the tabulated tier under joint perturbation of F, S, and A (Section IV-B).

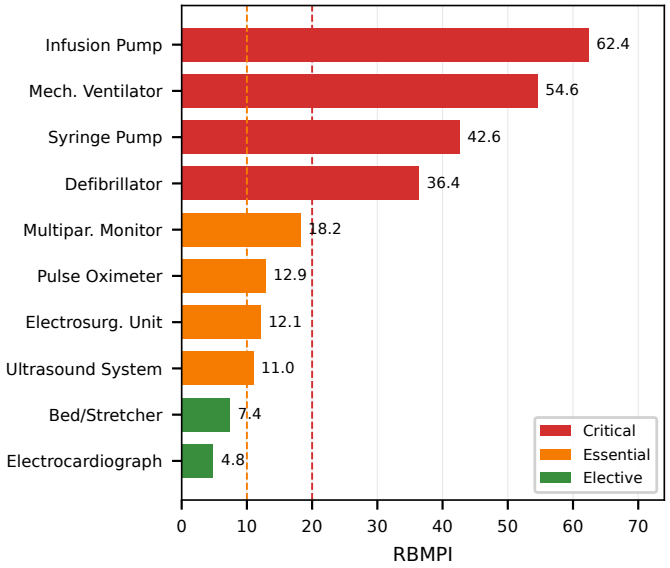


Fig. 1. Equipment ranking by RBMPI. Red: Critical; orange: Essential; green: Elective. Dashed vertical lines mark the Critical (20) and Essential (10) thresholds. The gap between the lowest Critical device (36.4) and the highest Essential device (18.2) is the dominant break in the distribution (Section III-G).

B. Robustness of the Classification

Because F and S are ordinal estimates and A is binary, we tested classification stability over 10,000 Monte Carlo iterations. Each iteration perturbed F and S by up to one level using a triangular distribution centred on the assigned score, and drew A as a Bernoulli variable reflecting 85% confidence in the coded value. Draws falling outside the scale were clipped to the bounds rather than resampled, which places a point mass at the extremes; S was additionally rounded to integers, since it is defined on an integer scale, while F remained continuous. For devices already at the scale maximum this concentrates a substantial share of draws at that bound, and it is the mechanism behind the one-sided behaviour described below. Regulatory class is a legal attribute of the device, so it was held fixed here and varied separately in Section V-B. Fig. 2 reports tier assignment probabilities.

The four Critical devices hold tier in 88.1% to 100% of iterations (infusion pump 100%, mechanical ventilator 99.5%, syringe pump 94.8%, defibrillator 88.1%), and the electrocardiograph holds Elective in 93.2%. The other five are less stable, and all five sit within one tier width of a boundary. The ultrasound system stays Essential in 54.6% of runs and drops to Elective in 44.1%. The multiparameter monitor stays Essential in 62.1% and crosses into Critical in 37.9%. The electrosurgical unit (71.1%), pulse oximeter (76.1%), and patient bed (78.0%) hold tier in roughly three quarters of runs. Read the classification as firm for the four Critical devices and the electrocardiograph, provisional for the rest.

Two features of the simulated distributions run in opposite directions and are worth noting. The mean simulated index for the multiparameter monitor (21.0) exceeds its baseline of 18.2, because treating substitute availability probabilistically means the assumed substitute is sometimes absent and the score doubles. Devices at the top of the F and S scales move the other way: clipping makes perturbation one-sided there, so their simulated means fall below baseline (infusion pump 55.3 against 62.4). Neither effect changes the tier probabilities, but together they show the index carries asymmetric uncertainty and should not be read as a point estimate.

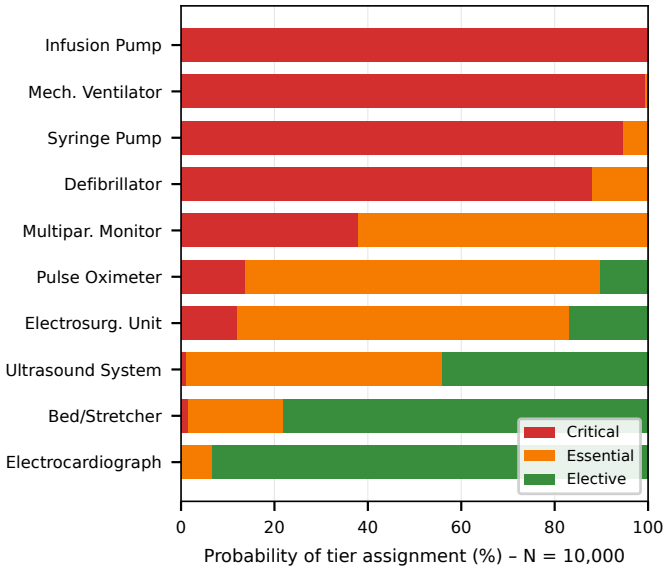


Fig. 2. Monte Carlo tier stability over 10,000 iterations with F, S, and A perturbed simultaneously. Each bar gives one device's probability of assignment to each tier. Single-colour bars are stable; mixed bars mark boundary devices sensitive to score uncertainty.

C. Comparison with the Fennigkoh–Smith Model

To position the index against an established baseline we computed the Fennigkoh–Smith equipment management number (EM = Function + Physical Risk + Maintenance) for the same ten categories. To avoid ad hoc scoring, each device was mapped to the published Fennigkoh–Smith functional categories, to a failure-consequence category, and to standard maintenance-intensity levels, so every score follows a documented rule. Rank agreement is strong (Spearman's $\rho = 0.835$, $p = 0.003$), and the two models converge at the extremes, both placing the mechanical ventilator highest and the patient bed lowest. That convergence is evidence of convergent validity, not of correctness: it shows the RBMPI agrees with an accepted model where that model is most confident.

Divergences track the RBMPI's extra criteria (Fig. 3). Under the original inclusion rule (EM $\geq$ 12), seven of ten categories would enter the maintenance programme. The sharpest disagreement is the electrosurgical unit: Fennigkoh–Smith ranks it fifth (EM = 15, comfortably above threshold) while the RBMPI ranks it seventh and calls it Essential, because the census-derived profile codes it as having a functional substitute. The additive model has no way to represent that.

Running the other way, the RBMPI lifts the infusion pump to first, whereas the additive EM compresses the high-severity devices into a band from 18 to 20.

Perturbing every Fennigkoh–Smith component score by up to one level over 10,000 iterations (Fig. 4) keeps rank correlation strong (mean $\rho = 0.828$, 95% CI [0.687, 0.936]), and the electrosurgical-unit divergence holds direction in 96% of runs. The redundancy-driven reclassification is structural, not an artifact of which scores we picked.

Two further divergences move a device by two or more positions and are equally robust. The pulse oximeter rises 2.5 positions under the RBMPI (direction preserved in 97% of runs), because a comparatively high failure frequency ($F = 3.9$) counts for more in a product than in a sum dominated by the function score. The defibrillator falls two positions (89%), penalized by moderate failure frequency, though it stays firmly Critical. Divergences among near-tied devices in the Critical cluster, such as the syringe pump against the mechanical ventilator, hold direction in only 48% to 60% of runs and should be read as ties.

V. Discussion

A. Applicability in the SUS Context

The index was built for facilities with no sensor infrastructure, no analytics platform, and often no biomedical engineering department. Its four inputs come from publicly available ANVISA and CNES data, standard clinical engineering references, and direct inventory knowledge. Computing it requires a spreadsheet and an accurate equipment list, which puts it within reach of clinical engineers across facilities of varying size, unlike the sensor-based approaches that dominate the recent literature [6]. That is a claim about data requirements, not about validated benefit.

B. What the Model's Free Parameters Actually Drive

Three components of the index were chosen rather than measured: the redundancy multiplier, the regulatory weight vector, and the aggregation rule. We tested each.

Redundancy multiplier. Replacing $(2 - A)$ with $(k - A)$ leaves rank order untouched for $k \geq 2$ ($\rho = 1.000$ at $k = 2.5$ and $k = 3.0$) and nearly untouched below ($\rho = 0.988$ at $k = 1.75$; $\rho = 0.964$ at $k = 1.5$). Tier assignment responds more: $k = 1.5$ moves four devices, $k = 2.5$ moves two. Every device that moves under any k belongs to the boundary group already flagged in Section IV-B. The four Critical devices and the electrocardiograph never change tier at any k tested.

Regulatory weight. A linear scheme (1.0/1.2/1.4/1.6) and an aggressive one (1.0/1.3/1.6/2.0) both leave rank order unchanged ($\rho = 1.00$); dropping the weight entirely still gives $\rho = 0.99$, with one adjacent swap between the electrosurgical unit and the ultrasound system. Tier assignment shifts slightly: the unweighted scheme moves the electrosurgical unit to Elective, the aggressive scheme moves the multiparameter monitor to Critical. Both devices were already flagged as unstable. We state the conclusion plainly: in this sample W_r contributes almost nothing. With no Class IV device present its range is 1.0 to 1.3, and rank order is effectively invariant to it. Its role here is interpretive, tying maintenance priority to ANVISA's own risk assessment in keeping with RDC 509/2021 [9], rather than discriminative. Whether it earns a place as a fourth criterion is a question a larger inventory containing Class IV devices would settle.

Aggregation rule. We also tested the product form itself against an equal-weight additive analogue computed on min-max normalized criteria. The two agree at $\rho = 0.964$. Only three devices shift position, none by more than two, and all three belong to the same boundary group. Ranks one through five and nine through ten are identical under both rules. This is a negative result and we report it as such: on this data the choice between multiplying and adding changes very little. What distinguishes the RBMPI from Fennigkoh–Smith is the substitute availability criterion, not the functional form. Readers should weigh the contribution accordingly.

Across all three tests the pattern is consistent. The parameter values are not what drives the classification; the ordinal structure of F, S, and A is. And the same small set of boundary devices proves fragile under every perturbation applied, which makes calibration effort targetable rather than open-ended.

C. Interpretation of Key Results

Placing infusion and syringe pumps in the Critical tier is consistent with the documented contribution of infusion devices to medication administration errors. A systematic review of smart infusion pump interoperability reports reductions of 21.2% to 90.5% in overall medication administration errors after implementation [16], which indicates the scale of the error burden these devices carry beforehand. The defibrillator is the instructive case: moderate failure frequency, maximal severity, no substitute. Any frequency-only or severity-only rule would rank it lower than the combined criteria do.

The tier definitions aim at the hazard-driven maintenance attention that device safety bodies recommend, where preventive intensity tracks the consequence of failure rather than acquisition cost [17].

D. Calibration for Different Hospital Profiles

The index is meant to be recalibrated, not adopted as printed. Three steps use data an institution already holds.

Failure frequency. Replace the literature-informed F with corrective work orders per device category over twelve months, normalized by installed units and rescaled to 0–5 across the local inventory. Without a computerized maintenance management system, the corrective service records required by RDC 509/2021 [9] carry the same information. This step swaps the model's weakest input for a measured one and should be treated as a precondition for using the index to allocate resources.

Clinical severity. Re-score S against services the facility actually operates. A unit with no intensive care bed or operating theatre has nothing in the immediate-risk-to-life band during routine operation and leaves the top of the scale unused. A tertiary referral centre applies the scale as published.

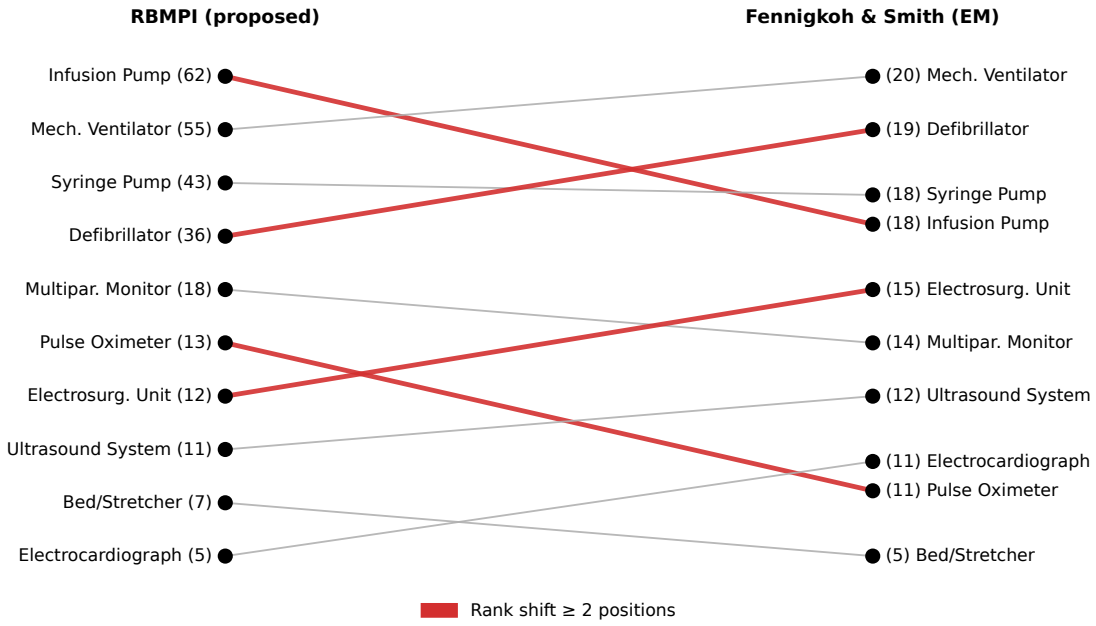


Fig. 3. Rank comparison between the RBMPI (left) and the Fennigkoh–Smith EM number (right). Marker colour is the device's RBMPI tier on both sides. Red connectors mark rank shifts of two or more positions.

Substitute availability. Replace binary A with the facility's unit count from its CNES record [10]: $A = 0$ for a single installed unit, $A = 1$ when units exceed peak simultaneous demand, an intermediate value otherwise. Where transfer time or physical distance materially limits access, those units should not count toward redundancy.

Thresholds must be re-derived, not carried over. Section III-G shows the Critical boundary rests on a gap specific to this distribution and the Essential boundary rests on policy; both need re-examination against the recalibrated scores. Expected direction of change differs by profile. In a small municipal unit, low unit counts push most categories to $A = 0$ while a compressed severity scale lowers top scores, flattening the ranking and enlarging the Critical tier. In a medium general hospital, redundancy among monitoring devices dominates and the Essential tier expands. In a tertiary centre with its own clinical engineering department, local failure counts replace the ordinal F entirely and the index converges on a conventional risk priority number. These are expectations to be tested, not results.

E. Limitations

The index rests on expert-elicited ordinal scores rather than measured data, and that is its principal limitation. Failure frequency was informed by aggregate surveillance composition rather than device-specific ANVISA counts, which are not published at the required resolution. Severity and substitute availability reflect general SUS conditions assigned from literature rather than observation. Two raters produced the scores without a formal consensus protocol, so inter-rater agreement is unquantified. The Monte Carlo analysis bounds how far the classification survives perturbation of these estimates, but perturbing one's own assumptions measures internal fragility, not external accuracy.

Several further limitations follow. The sample is ten categories, which makes the rank correlation with Fennigkoh–Smith a weak statistic, as its 95% interval [0.687, 0.936] shows. Substitute availability remains binary. Equipment age is absent and is a priority for model development. The product form treats F and S symmetrically, which is arguably wrong for a patient-safety application where a rare catastrophic failure and a frequent trivial one should not score alike; an asymmetric weighting deserves testing.

The most consequential gap is that we cannot say whether using this index improves anything. Establishing that the RBMPI raises equipment availability, reduces safety incidents, or improves maintenance performance requires a prospective comparison against decisions made without it, with outcomes measured over at least a maintenance cycle. No such study was possible here: corrective service records are not standardized across SUS facilities and are not publicly accessible. The Fennigkoh–Smith comparison establishes convergent validity with an existing model and nothing more. Until outcome data exist, the RBMPI should be treated as a way to make prioritization assumptions explicit and testable, not as a validated decision rule, and the prospective study is the single most valuable next step.

VI. Conclusion

We proposed a Risk-Based Maintenance Priority Index for medical equipment in Brazilian public hospitals, combining failure frequency, clinical severity, substitute availability, and

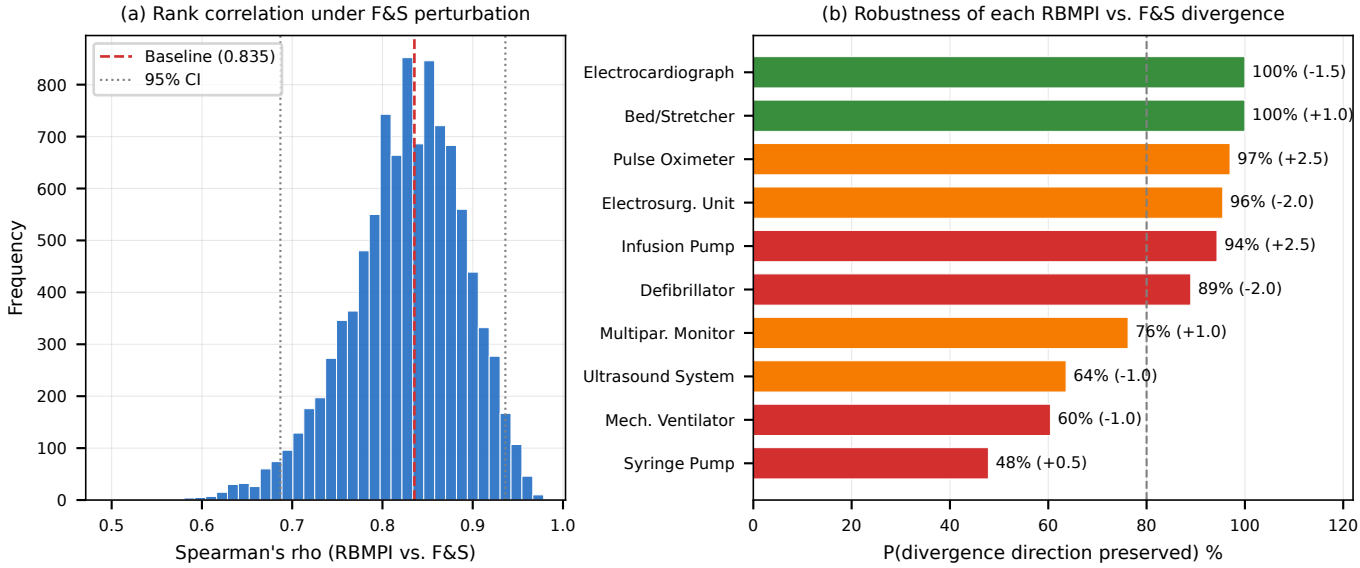


Fig. 4. Robustness of the RBMPI versus Fennigkoh–Smith comparison under ± 1-level perturbation of the Fennigkoh–Smith scores (10,000 iterations). (a) Distribution of Spearman's ρ with baseline and 95% interval marked. (b) Probability that each device preserves the direction of its divergence.

regulatory risk class. Across ten categories it produced a three-tier classification with four devices Critical. Monte Carlo analysis showed those four and the electrocardiograph hold tier under simultaneous variation of all scored parameters, while five boundary devices change tier often enough to be treated as provisional. Comparison with Fennigkoh–Smith showed broad agreement and one robust divergence: the RBMPI lowers priority for devices with a functional substitute, which the additive model cannot express.

Three lessons emerge. Substitute availability changes maintenance priorities in ways function-and-risk models miss, but only for a subset of devices; where equipment is clearly critical, the models agree. Fragility is concentrated rather than diffuse, since the same small group of devices moves under every perturbation applied, whether score noise, redundancy multiplier, regulatory weighting, or aggregation rule. And the aggregation rule itself matters less than its prominence in the formula suggests: an additive analogue reproduces the ranking at $\rho = 0.964$, which locates the contribution in the criterion set rather than the arithmetic.

The open gaps are a prospective hospital study measuring whether the index improves availability or safety outcomes, device-specific notification rates from NOTIVISA, a graded availability measure, and inclusion of equipment age. We offer the RBMPI as a preliminary and transparent framework whose value depends on local calibration and prospective validation before broader adoption.

Acknowledgment

The authors thank the Brazilian Health Regulatory Agency (ANVISA) for maintaining public access to post-market device surveillance data, which informed the scoring in this work.

References

[1] Tribunal de Contas da União, "Acórdão 999/2017 — Plenário, TC 025.472/2016-0: auditoria em aquisições de equipamentos médico-hospitalares por hospitais federais no Rio de Janeiro," TCU, Brasília, Brazil, 2017.

[2] J. R. Cardoso and C. A. Waqued, "Engenharia clínica como vetor do avanço na gestão de equipamentos médico-hospitalares: caso de um hospital regional no DF," *Revista de Engenharia e Tecnologia*, vol. 11, no. 4, 2019.

[3] L. Fennigkoh and B. Smith, "Clinical equipment management," *JCAHO Plant Technology and Safety Management Series*, no. 2, pp. 5–14, 1989.

[4] Agência Nacional de Vigilância Sanitária, "Painel de Tecnovigilância: notificações de dispositivos médicos," ANVISA, Brasília, Brazil. [Online]. Available: https://www.gov.br/anvisa

[5] C. C. D. Andrade, N. A. Luz, and P. S. S. Pinto, "Tecnovigilância dos dispositivos médicos no Brasil nos últimos 10 anos," *Epidemiologia e Serviços de Saúde*, vol. 34, no. Suppl. 1, p. 168, 2025.

[6] N. A. M. Radzi, N. A. Nayan, and S. N. Ishak, "Systematic literature review on improving healthcare through proactive maintenance and fault detection in medical equipment," *Health and Technology*, vol. 16, no. 1, pp. 1–17, 2026.

[7] M. Sezdi, "Two different maintenance strategies in the hospital environment: preventive maintenance for older technology devices and predictive maintenance for newer high-tech devices," *Journal of Healthcare Engineering*, vol. 2016, art. 7267983, 2016.

[8] S. Taghipour, D. Banjevic, and A. K. S. Jardine, "Prioritization of medical equipment for maintenance decisions," *Journal of the Operational Research Society*, vol. 62, no. 9, pp. 1666–1687, 2011.

[9] Agência Nacional de Vigilância Sanitária, "Resolução RDC nº 509, de 27 de maio de 2021: gerenciamento de tecnologias em saúde em estabelecimentos de saúde," *Diário Oficial da União*, Brasília, Brazil, 2021.

[10] Ministério da Saúde, "Cadastro Nacional de Estabelecimentos de Saúde (CNES): relatório de equipamentos hospitalares," Brasília, Brazil. [Online]. Available: http://cnes.datasus.gov.br

[11] Ministério da Saúde, *Equipamentos médico-hospitalares e o gerenciamento da manutenção*. Série F, Comunicação e Educação em Saúde. Brasília, Brazil: Ministério da Saúde, 2002.

[12] R. S. Lopes and L. A. N. Santiago, "Manutenção preventiva em estabelecimentos de saúde," *Research, Society and Development*, vol. 11, no. 15, art. e596111537572, 2022.

[13] B. Wang and A. Levenson, "Equipment inclusion criteria: a new interpretation of JCAHO's medical equipment management standard," *Journal of Clinical Engineering*, vol. 25, no. 1, pp. 26–35, 2000.

[14] *Failure Modes and Effects Analysis (FMEA and FMECA)*, IEC Standard 60812:2018, 3rd ed., International Electrotechnical Commission, Geneva, Switzerland, 2018.
[15] G. F. Jenks, "The data model concept in statistical mapping," *International Yearbook of Cartography*, vol. 7, pp. 186–190, 1967.
[16] E. P. Borrelli, J. D. Lucaci, N. S. Wilson, A. Taneja, M. Weiss, and I. Beer, "Evaluating the impact of smart infusion pump interoperability on reducing medication administration errors: a systematic literature review," *Medical Devices: Evidence and Research*, vol. 18, pp. 247–260, 2025.
[17] ECRI, "Top 10 health technology hazards for 2023," ECRI, Plymouth Meeting, PA, USA, 2023.
[18] J. G. Webster, *Medical Instrumentation: Application and Design*, 4th ed. Hoboken, NJ, USA: Wiley, 2010.
[19] The pandas Development Team, "pandas-dev/pandas: Pandas," Zenodo, 2020.
[20] J. D. Hunter, "Matplotlib: a 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
[21] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

Scientific Foundations of Immersive Humanization in Diagnostic Imaging

1st Carolina Gois Falandes
Department of Journalism and Publishing (CJE)
School of Communications and Arts, University of São Paulo (USP)
São Paulo, Brazil
https://orcid.org/0000-0001-6639-2121

2nd Tiago Gois Falandes
MR In Site Healthcare Solutions
São Bernardo do Campo, Brazil
tiago.falandes@mrinsite.com.br

***Abstract*—Humanization in diagnostic imaging environments represents a growing challenge due to the need to reduce anxiety, promote well-being, and enhance the patient experience during imaging procedures. Although different strategies have been investigated in the literature, scientific production remains dispersed across distinct fields of knowledge, hindering an integrated understanding of the topic. Considering this scenario, this study aims to map and analyze the scientific literature addressing humanization strategies in diagnostic imaging environments, with special attention to the potential of immersive technologies. To this end, a scientometric analysis of the national and international scientific literature was conducted based on structured searches in the Web of Science Core Collection, complemented by exploratory searches in SciELO and Google Scholar Labs. Keyword co-occurrence analysis was performed using VOSviewer. The results highlight the convergence of four key dimensions: patient experience, humanization of care, evidence-based design, and immersive technologies. Based on this synthesis, the concept of Immersive Humanization is proposed as a conceptual framework integrating these dimensions to inform future research and the development and evaluation of solutions aimed at enhancing the patient experience in diagnostic imaging environments.**

***Index Terms*—diagnostic imaging, healthcare humanization, patient experience, immersive technologies.**

I. Introduction

Medical examinations can represent potentially stressful experiences for many patients, involving factors that go beyond expectations regarding the diagnosis. This scenario highlights the importance of adopting strategies capable of minimizing discomfort and promoting a more positive experience during imaging procedures.

In the context of diagnostic imaging, an important tool for the early detection, monitoring, and treatment planning of multiple diseases [1], imaging modalities such as magnetic resonance imaging (MRI) and computed tomography (CT) present characteristics that pose challenges to patients. Enclosed spaces, intense noise, the need to remain still, and prolonged examination times are factors that may trigger anxiety, discomfort, and claustrophobia [1], [2].

From this perspective, the environment in which imaging examinations are performed can influence patients' emotional responses [3], making the design of diagnostic imaging rooms and the adoption of appropriate care protocols strategic elements for promoting the humanization of care.

In this regard, different approaches have been investigated to reduce the emotional impacts associated with imaging procedures and enhance the patient experience. The literature includes non-pharmacological interventions based on design [3], art [4], music therapy, hypnosis, and health psychology [5], as well as studies focused on the development of communication strategies to facilitate magnetic resonance imaging without anesthesia in children [6] and the application of graphic design in computed tomography environments to promote children's well-being [7].

More recently, advances in digital technologies have expanded this range of approaches by incorporating multimedia resources [8] and immersive applications, such as virtual reality [9] and moving image projection systems [10]. In this context, studies indicate that these solutions offer new possibilities for supporting diagnostic imaging procedures and improving the patient experience, establishing themselves as a promising area of innovation in healthcare [11].

Although these initiatives have been investigated in different healthcare contexts, their scientific foundations remain dispersed across fields such as healthcare, design, architecture, computer science, and engineering, hindering an integrated understanding of the field and the consolidation of a conceptual framework for humanization in diagnostic imaging environments.

In view of this gap, this study presents a scientometric analysis of the national and international scientific literature with the aim of mapping, systematizing, and understanding how the literature has supported humanization strategies in diagnostic imaging environments. Based on the identified convergences, the conceptual model of *Immersive Humanization* is proposed, integrating principles of humanization of care, environmental design, patient experience, and the use of immersive technologies.

II. Materials and Methods

A. Research Design

This study is characterized as exploratory research based on scientometric analysis using bibliometric indicators derived from keyword co-occurrence networks. Its objective was to map, systematize, and interpret the scientific literature related to humanization strategies in diagnostic imaging environments, seeking to identify convergences capable of supporting the proposal of the conceptual model of *Immersive Humanization*.

The methodological design was organized around four analytical axes: (i) healthcare humanization; (ii) patient experience in diagnostic imaging examinations; (iii) immersive technologies applied to healthcare; and (iv) healthcare environment and hospital design. These categories guided both the definition of the search strategies and the organization and interpretation of the results.

B. International Literature Survey

The international literature search was conducted in the Web of Science Core Collection on June 26, 2026, covering publications indexed between 2015 and 2026, a period selected because it corresponds to the consolidation of immersive technologies in healthcare. For each analytical axis, specific search strategies were developed, resulting in the establishment of seven independent thematic corpora, as presented in Table I. The searches were refined to include records classified as Articles or Review Articles and indexed as Open Access. No language restriction was applied.

TABLE I. SEARCH STRATEGIES USED IN THE WEB OF SCIENCE CORE COLLECTION

Analytical axis	*Search strategy*	*Records*
Patient experience	patient experience AND MRI	112
	patient experience AND "computed tomography"	55
Healthcare humanization	healing environment AND hospital	75
	positive distraction AND hospital	19
Immersive technologies	virtual reality AND MRI	209
	virtual reality AND "diagnostic imaging"	23
Healthcare environment and hospital design	visual environment AND (hospital OR healthcare)	9

It should be emphasized that the record counts reported in Table I correspond to the searches conducted on June 26, 2026. Because Web of Science is continuously updated, rerunning the same searches at a later date may yield different numbers of records.

C. National Literature Survey

To contextualize the Brazilian scientific production, a complementary literature search was conducted in the SciELO Brazil collection on June 30, 2026, considering publications from 2015 to 2026. Initially, search strategies directly related to the research topic were employed, using search expressions such as *humanização AND diagnóstico por imagem*, *Política Nacional de Humanização AND ressonância magnética*, *humanização imersiva*, and *humanização AND projeção*. However, these searches retrieved few or no studies, highlighting the scarcity of Brazilian publications directly addressing the relationship between humanization, diagnostic imaging, and immersive technologies.

Given this scenario, the search strategy was expanded to include broader search terms, namely *humanização hospitalar* (35 records), *Política Nacional de Humanização* (42 records), *ambiência hospitalar* (7 records), and *humanização AND projeção* (1 record). The retrieved studies were subjected to an exploratory reading of titles and abstracts, allowing the identification of contributions related to the foundations of humanization of care and the organization of healthcare environments. These records were used for contextual and thematic interpretation rather than for keyword co-occurrence analysis.

TABLE II. SEARCH STRATEGIES USED IN SCIELO

Search term	*Records*
Humanização hospitalar	35
Política Nacional de Humanização	42
Ambiência hospitalar	7
Humanização AND diagnóstico por imagem	0
Política Nacional de Humanização AND ressonância magnética	0
Humanização imersiva	0
Humanização AND projeção	1

D. Exploratory Survey Using Google Scholar

In addition to the indexed scientific databases, an exploratory survey was conducted using Google Scholar on June 26, 2026, through the experimental Google Scholar Labs feature, which enables the retrieval of scientific literature using natural language queries. The search was performed using the following prompt: *"I need to find studies addressing healthcare humanization supported by audiovisual projections in examination rooms, with a focus on magnetic resonance imaging and computed tomography environments."*

This procedure resulted in the identification and selection of 17 relevant publications, including scientific articles, master's dissertations, undergraduate theses, books, and literature reviews. This stage had a complementary purpose and was intended to enrich the contextualization of the scientific literature, particularly with regard to practical experiences and applications developed in Portuguese. In this case, no time restriction was adopted. These materials were used to provide additional contextual and thematic insights and were not included in the VOSviewer co-occurrence analysis.

E. Data Processing and Analysis

Across the three complementary sources, 604 records and publications were identified: 502 records from the Web of Science Core Collection, 85 records from SciELO, and 17 publications identified through Google Scholar Labs. The 502 records retrieved from the Web of Science Core Collection were organized into seven independent thematic corpora and exported in a compatible format for bibliometric analysis using VOSviewer software (version 1.6.20).

Keyword co-occurrence analysis was performed using full counting. For corpus A1, author keywords were used as the unit of analysis, with a minimum occurrence threshold of two. For corpora A2, B, C, D1, and D2, all keywords were considered, with a minimum occurrence threshold of five. For corpus E, all keywords were considered, with a minimum occurrence threshold of two. The resulting networks were described using the number of keywords meeting the occurrence threshold, number of clusters, number of links, total link strength, and keyword occurrence frequency. Because keyword occurrence thresholds differed among the corpora, network indicators such as the number of selected keywords, links, and total link strength were interpreted descriptively and in relation to the specific search strategy and corpus structure, rather than as direct measures of scientific maturity or clinical effectiveness.

The seven Web of Science corpora were treated as independent thematic datasets rather than merged into a single corpus because each corpus resulted from a distinct search strategy designed to address a specific analytical dimension. Therefore, a publication could appear in more than one corpus when it met the respective search criteria, as such overlap represented thematic convergence rather than duplication within a single analytical dataset.

The 85 records retrieved from SciELO and the 17 publications identified through Google Scholar Labs were subjected to exploratory screening and thematic interpretation by the first author through examination of titles and abstracts. The resulting interpretations were subsequently discussed with the co-author. No independent duplicate screening or coding procedure was performed. These complementary sources were used for contextual and thematic interpretation.

Subsequently, an interpretative synthesis of the different document sets was carried out to identify convergences among the four analytical axes of the study. This stage made it possible to systematize the main scientific foundations related to humanization in diagnostic imaging environments and to formulate the conceptual model of *Immersive Humanization*.

III. Results

A. International Scientific Production

The search strategies adopted enabled the identification of different theoretical perspectives and applications related to humanization in diagnostic imaging environments. The integrated analysis of the corpora revealed convergences among the four investigated axes – patient experience, healthcare humanization, immersive technologies, and healthcare environment and hospital design – highlighting recurring elements that informed the formulation of the conceptual model of *Immersive Humanization*.

1) Patient experience: The analysis of the corpora dedicated to patient experience in magnetic resonance imaging (112 records) and computed tomography (55 records) highlights a substantial body of research addressing the patient experience in diagnostic imaging. Although both imaging modalities share a common interest in the patient experience, the literature reveals distinct thematic emphases.

In the MRI corpus, a predominance of studies addressing emotional aspects associated with the examination was observed, with emphasis on anxiety, claustrophobia, and strategies aimed at enhancing the patient experience during the procedure. In contrast, the literature related to CT is more closely associated with oncological contexts, in which patient experience is investigated as part of the diagnostic and therapeutic journey.

The keyword co-occurrence analysis generated a network with 16 selected keywords, organized into four clusters and 47 links, with a total link strength (TLS) of 96 for the MRI corpus. The CT corpus comprised seven selected keywords distributed across three clusters and connected by 13 links, with a TLS of 19. Although the occurrence thresholds differed between the two corpora, the results indicate a broader range of thematic connections in the MRI network.

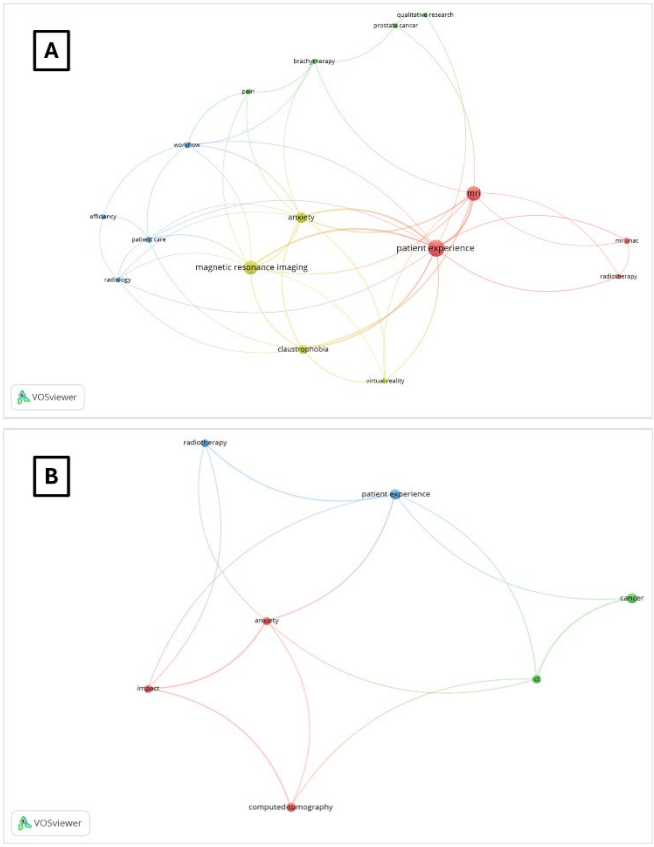


Fig. 1. Keyword co-occurrence networks of the corpora related to patient experience in magnetic resonance imaging (A) and computed tomography (B).

Figure 1 illustrates that, in both corpora, the term *patient experience* occupies a central position in the network, establishing connections with different dimensions of diagnostic imaging care. While the magnetic resonance imaging literature presents a more diverse thematic structure, encompassing emotional, organizational, and technological aspects, the computed tomography corpus shows a closer association with oncological and radiotherapy contexts, relating the patient experience to the impacts of treatment and the care pathway.

In summary, both corpora demonstrate that patient experience represents an important dimension of research on diagnostic imaging. Across the two imaging modalities, anxiety, perception of the examination, communication, and the healthcare context recur as relevant themes, indicating the importance attributed by the literature to the conditions surrounding the imaging experience.

2) Healthcare humanization: The analysis of the corpora related to healthcare humanization comprises 94 publications distributed across the concepts of Healing Environment (75 records) and Positive Distraction (19 records). Overall, these studies indicate that the international literature approaches the humanization of healthcare environments beyond interpersonal relationships, also considering the physical environment as a relevant dimension of well-being, quality of care, and patient experience.

The *Healing Environment* corpus is characterized by a broad and interdisciplinary approach, primarily grounded in the principles of *Evidence-Based Design*. The studies focus on

architectural planning, environmental design, and the organization of healthcare spaces in relation to comfort, quality of care, and health promotion.

The *Healing Environment* keyword network comprised 11 selected keywords distributed across three clusters, with 39 links and a total link strength of 117. In comparison, the *Positive Distraction* network comprised six keywords concentrated in a single cluster, with 15 links and a TLS of 27. These indicators are consistent with the broader thematic scope of the Healing Environment corpus and the more focused thematic structure of *Positive Distraction*.

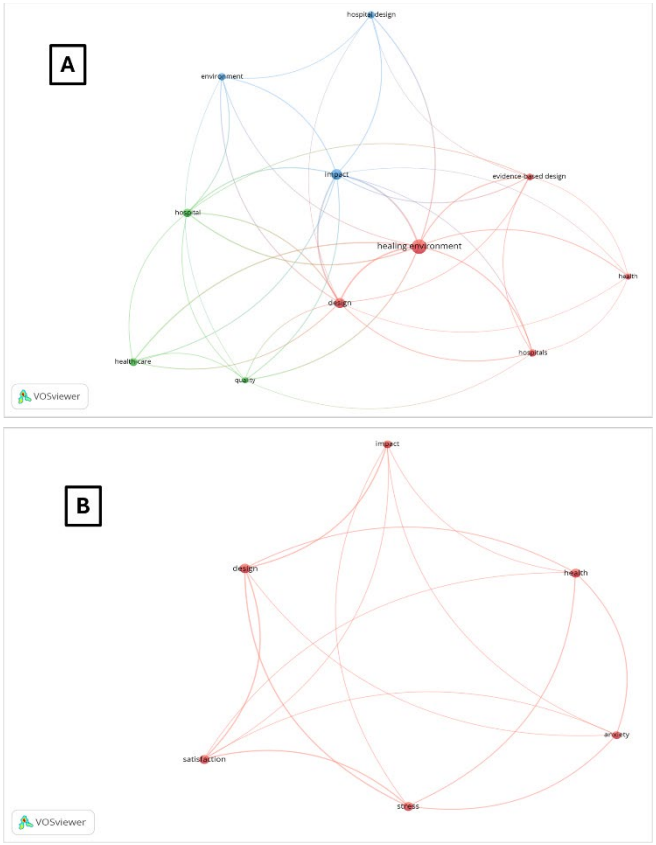


Fig. 2. Keyword co-occurrence networks of the *Healing Environment* (A) and *Positive Distraction* (B) corpora.

Figure 2 shows the complementarity between the two corpora. While the *Healing Environment* map is organized around concepts associated with the planning of healthcare environments, such as *evidence-based design*, *hospital design*, *environment*, and *quality*, the *Positive Distraction* corpus focuses on themes associated with *design*, *stress*, *anxiety*, *satisfaction*, and *health*. This difference indicates that the two fields investigate distinct, yet complementary, dimensions of the humanization of healthcare environments.

Thus, the results suggest that humanization extends beyond the exclusively relational dimension of care and incorporates the physical environment as an important component of the healthcare experience. From this perspective, *Healing Environment* provides conceptual foundations for evidence-based planning of healthcare environments, whereas *Positive Distraction* is discussed in the literature as an environmental intervention approach intended to address stress and anxiety and to support more welcoming and patient-centered experiences.

3) Immersive technologies: Upon examining the corpora related to immersive technologies, a substantial body of scientific production concerning the application of virtual reality in diagnostic imaging examinations can be observed. The corpus related to magnetic resonance imaging comprises 209 publications, whereas the broader corpus focused on diagnostic imaging comprises 23 publications, revealing different thematic scopes within the literature.

The MRI corpus generated a network of nine selected keywords distributed across four clusters, with 28 links and a TLS of 110. The diagnostic imaging corpus comprised four keywords concentrated in a single cluster, with six links and a TLS of 23. The larger number of records and the greater network structure observed in the MRI corpus reflect its more specific research focus on virtual reality in magnetic resonance imaging.

In the MRI corpus, virtual reality emerges as the central focus of the investigations, establishing connections with terms such as *anxiety*, *claustrophobia*, *sedation*, and *pediatric population*. As illustrated in Fig. 3(A), these co-occurrences show that publications in this corpus frequently associate immersive technologies with emotional responses and examination-related challenges.

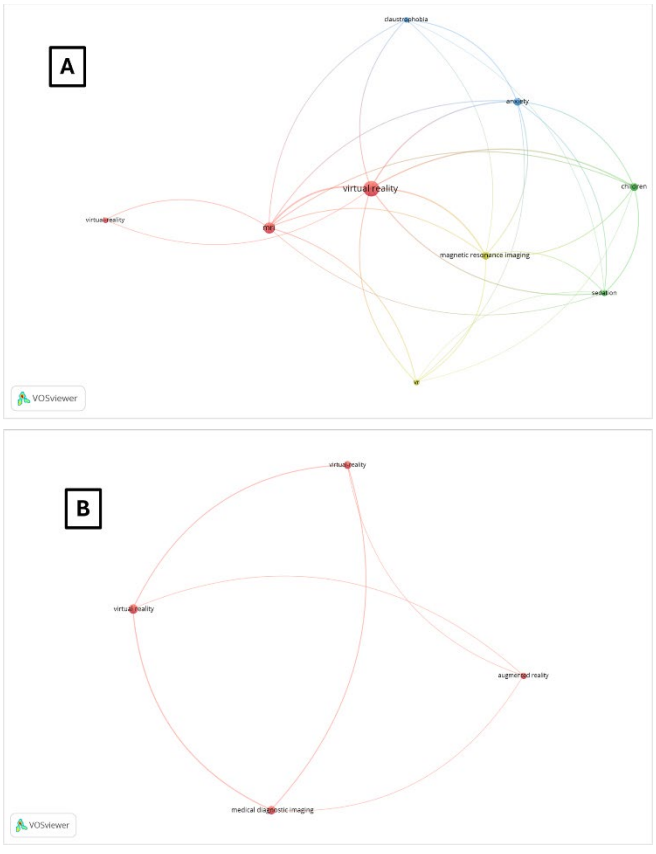


Fig. 3. Keyword co-occurrence networks of the *Immersive Technologies* corpora. (A) *virtual reality* AND *MRI*. (B) *virtual reality* AND *diagnostic imaging*.

The corpus dedicated to diagnostic imaging, in turn, presents a more concise thematic structure (Fig. 3(B)), organized around the concepts of *virtual reality*, *augmented reality*, and *medical diagnostic imaging*. Unlike the first corpus, which focuses on the patient experience during magnetic resonance imaging, this set of studies broadens the scope of investigation by addressing the incorporation of

different immersive technologies into diagnostic processes, including applications related to visualization, planning, and technological innovation.

Thus, the two corpora demonstrate complementary research orientations. The first is more closely associated with the patient experience during MRI examinations, particularly in relation to anxiety, claustrophobia, and sedation-related outcomes. The second broadens the technological scope to include virtual and augmented reality in diagnostic imaging, indicating a wider research agenda concerning the incorporation of immersive technologies into healthcare.

4) Healthcare Environment and Hospital Design: The corpus related to healthcare environment and hospital design comprises nine publications, representing the smallest document set identified in the international literature. Despite the limited number of studies, a recent and interdisciplinary research agenda can be observed, focusing on the investigation of the visual components of the built environment and their relationship with the patient experience. The retrieved studies address aspects related to sensory design, restorative environments, Evidence-Based Design, environmental quality, and the application of visual technologies in healthcare settings.

The keyword co-occurrence analysis (Fig. 4) reveals a sparse thematic network, composed of a limited number of terms and a low density of connections. The network comprises three selected keywords distributed across two clusters and connected by a single link. Unlike the other corpora, the network does not present consolidated thematic groupings, suggesting that research on visual healthcare environments remains comparatively underexplored. The few identified terms are distributed between the concepts of *information* and *stress*, suggesting that the visual environment has been investigated both as a communication and wayfinding resource and in relation to stress.

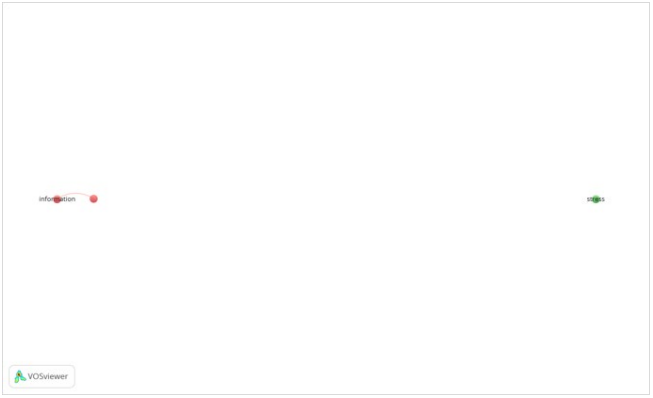


Fig. 4. Keyword co-occurrence network of the *Visual Environment* corpus.

Taken together, the findings indicate that the visual environment constitutes a relatively underexplored area of research in healthcare humanization. Although still underexplored, this field brings together knowledge from architecture, engineering, computer science, and health sciences, pointing to the potential of environmental design as a component of the patient experience and the improvement of spaces dedicated to diagnostic imaging.

B. National Scientific Production (SciELO)

The literature survey conducted in SciELO demonstrates that the Brazilian scientific production on healthcare humanization presents a different scenario from that observed in the international literature. The search strategies initially directed toward the interface between humanization, diagnostic imaging, and immersive technologies did not retrieve studies directly related to the specific research focus. The search terms *humanização* and *diagnóstico por imagem*, *Política Nacional de Humanização* and *ressonância magnética*, and *humanização imersiva* returned no results, whereas the search for *humanização* and *projeção* retrieved only one record that was not directly related to the healthcare context. Within the SciELO Brazil literature retrieved by the present search strategy, these findings suggest that the integration of humanization, immersive technologies, and diagnostic imaging has not yet been consolidated.

Given this scenario, the search strategy was expanded by using broader search terms related to *humanização hospitalar* (35 records), *Política Nacional de Humanização* (42 records), and *ambiência hospitalar* (7 records). The integrated analysis of these data made it possible to characterize the main axes of Brazilian scientific production on healthcare humanization.

The studies related to *Política Nacional de Humanização* are predominantly focused on organizational and institutional aspects, addressing topics such as *HumanizaSUS*, co-management, institutional support, professional training, and the implementation of public policies. In contrast, the corpus related to *humanização hospitalar* emphasizes care practices, bringing together studies on patient welcoming, nursing, occupational therapy, hospital psychology, patient-centered care, and healthcare environments. In turn, the corpus dedicated to *ambiência hospitalar* highlights the recognition of the physical environment as a component of care, encompassing studies on professionals' perceptions, environmental comfort, Florence Nightingale's environmental theory, and the organization of healthcare spaces.

Although these three corpora reveal a consistent body of research on humanization policies, care practices, and healthcare environments, these topics remain addressed relatively independently. No studies were identified within the retrieved SciELO records that directly articulate the principles of humanization of care, the planning of healthcare environments, and the use of immersive technologies in diagnostic imaging contexts. This gap in the retrieved literature informed the formulation of the *Immersive Humanization* model.

C. Exploratory Survey Using Google Scholar

In addition to the indexed scientific production, an exploratory survey was conducted using Google Scholar and the experimental Google Scholar Labs feature to identify studies related to humanization in diagnostic imaging environments supported by audiovisual and immersive technologies.

TABLE III. CLASSIFICATION OF STUDIES IDENTIFIED USING GOOGLE SCHOLAR LABS

Document type	*Count*	*Examples*
Scientific articles	5	Anxiety in PET/CT; humanized care in radiotherapy; multimedia resources in hospitals
Master's dissertations	3	Humanization in MRI rooms; communication in MRI; anxiety in MRI
Undergraduate theses	4	Hypnosis in MRI; graphic design in computed tomography; pediatric radiology
Books	3	Virtual reality in healthcare; humanization of technologies
Other	2	Meta-analysis; human–computer interaction study

The distribution of document types shows that the literature identified on the topic is dispersed across different publication formats, including scientific articles, master's dissertations, undergraduate theses, and books. This diversity indicates that contributions to the topic have emerged across different levels of investigation and publication formats.

The retrieved studies can be grouped into four main perspectives. The first comprises studies addressing anxiety, claustrophobia, and discomfort during diagnostic imaging examinations through non-pharmacological approaches such as hypnosis, music therapy, health psychology, and patient communication. The second encompasses research related to the design and physical environment of magnetic resonance imaging and computed tomography rooms, emphasizing architectural aspects, environmental comfort, and patient experience. A third group addresses applications of virtual reality in different healthcare contexts, including rehabilitation, mental health, and pain management. Finally, studies focusing on the humanization of care in diagnostic imaging examinations and the use of multimedia resources in healthcare environments were identified.

Overall, the results obtained through Google Scholar Labs complement the panorama observed in SciELO. While the indexed literature retrieved in SciELO showed limited integration among humanization, immersive technologies, and diagnostic imaging, the exploratory survey identified contributions distributed across different types of academic and applied literature. These findings suggest that relevant evidence on the topic remains dispersed across different areas of knowledge and publication formats rather than concentrated within a consolidated research field.

This scenario reinforces the relevance of an integrative framework capable of articulating the principles of humanization of care, patient experience, healthcare environment, and immersive technologies applied to diagnostic imaging environments.

IV. Discussion

Based on the scientometric analysis conducted, humanization in diagnostic imaging environments can be understood as a complex and multidimensional field. While the international literature reveals a greater concentration of research on themes related to patient experience, immersive technologies, and healthcare environments, the Brazilian scientific production primarily highlights the foundations of humanization of care and the organization of healthcare services. The comparison between these bodies of literature suggests that the articulation among these dimensions remains limited, particularly in relation to diagnostic imaging environments.

Drawing on the collected and interpreted data, and from the perspective of *Evidence-Based Design*, the hospital environment can be understood not merely as a passive setting, but as an integral component of the care experience. In this context, elements such as the *visual environment*, *audio environment*, and spatial organization have been investigated as strategies for improving care [12].

It is within this line of thought that the concept of *Positive Distraction* is situated. In stressful situations, positive distractions have been discussed as a coping strategy [13]. In this regard, the use of resources that promote relaxation – such as natural landscapes, soft colors, and calming sounds – may contribute to the creation of more supportive environments for patients and healthcare professionals [14].

This framework is further expanded when positive distraction mechanisms incorporate immersive technologies. The literature identified in the MRI corpus frequently associates virtual reality with patient preparation, familiarization, and examination-related challenges, including anxiety and claustrophobia. Individual studies have also reported outcomes related to anxiety, adherence, and acceptability [15]. These findings are derived from individual investigations and should not be interpreted as a direct result of the scientometric analysis, which identifies thematic associations and co-occurrence patterns rather than clinical effectiveness.

Thus, one of the central challenges in healthcare is to implement more “hospitable” hospitals, capable of adopting strategies that reduce rumination and enhance the patient experience [16]. Furthermore, according to researchers, institutions that invest in the integration of technology and care tend to receive more favorable evaluations from patients [17].

None of these perspectives, however, is sufficient when considered in isolation. Based on the results of the analysis of the national and international academic debate – plural and fragmented – the proposal of *Immersive Humanization* was developed, grounded in the convergence of four complementary dimensions: patient experience, understood as the guiding axis of the interventions; humanization of care, responsible for guiding ethical and care principles; evidence-based healthcare environments, focused on the physical planning of healthcare spaces; and immersive technological solutions, understood as resources capable of supporting patient welcoming, preparation, familiarization, and positive distraction during diagnostic procedures.

The proposed framework does not replace these perspectives, but seeks to articulate dimensions that are commonly addressed separately. In this sense, *Immersive Humanization* brings together the experiential perspective of patient-centered approaches, the environmental principles of *Evidence-Based Design*, the coping strategies associated with *Positive Distraction*, and the possibilities offered by immersive technologies within a framework specifically oriented toward diagnostic imaging environments. Its

contribution lies in this integrative articulation, which provides a conceptual basis for examining how environmental, experiential, care-related, and technological dimensions may be considered together.

The concept of immersion can also be understood as multidimensional rather than restricted to the use of a specific technological device. Drawing on Marín's typology of immersion in audiovisual nonfiction, three complementary modalities can be distinguished: perceptual immersion, associated with the sensory and technical characteristics of the medium; narrative-level immersion, related to the user's involvement in the story; and immersion as a response, which emphasizes active participation and the user's responses to the experience [18]. Although originally developed in the context of immersive audiovisual narratives, this typology provides a useful analytical lens for examining different configurations of immersive experiences in healthcare contexts.

From this perspective, immersive humanization may encompass different modes of engagement according to the technological and experiential configuration adopted. Immersive audiovisual projection in diagnostic imaging environments can be understood as approaching perceptual immersion, insofar as it modifies the patient's sensory relationship with the examination space [19], [20]. Virtual reality experiences incorporating narrative or playful elements in pediatric healthcare contexts may approximate narrative-level immersion by engaging patients through contextualized experiences [21]. In turn, VR-based therapeutic interventions for fears and phobias may approach immersion as a response, as the patient actively engages with virtual environments and is repeatedly exposed to stimuli associated with the feared situation as part of the intervention [22]. These examples are presented to illustrate different configurations of immersive experience and do not constitute evidence of the clinical effectiveness of the *Immersive Humanization* framework.

In this context, an immersive audiovisual projection system used in diagnostic imaging environments provides a concrete illustration of how some of these principles can be configured technologically. The proposal uses synchronized audiovisual projections on the walls, ceiling, and the magnetic resonance imaging and/or computed tomography equipment itself, creating an immersive environment to be experienced by the patient. From a technological perspective, this solution is conceptually related to CAVE (Cave Automatic Virtual Environment) immersive systems, developed to promote the sense of presence through audiovisual immersion [23]. Publicly available 360-degree audiovisual materials documenting the MRI and CT applications of the system are used here to illustrate its configuration and operation (Fig. 5A-B).

Beyond illustrating how the solution operates, 360-degree images may constitute a healthcare communication strategy, in line with recent studies [24]. By enabling prior familiarization with the environment and the diagnostic procedure through immersive experiences, these resources may contribute to reducing uncertainty and supporting patient preparation before the examination [25]. These potential contributions, however, require empirical evaluation and should not be inferred solely from the technological characteristics of the system or from the illustrative materials.

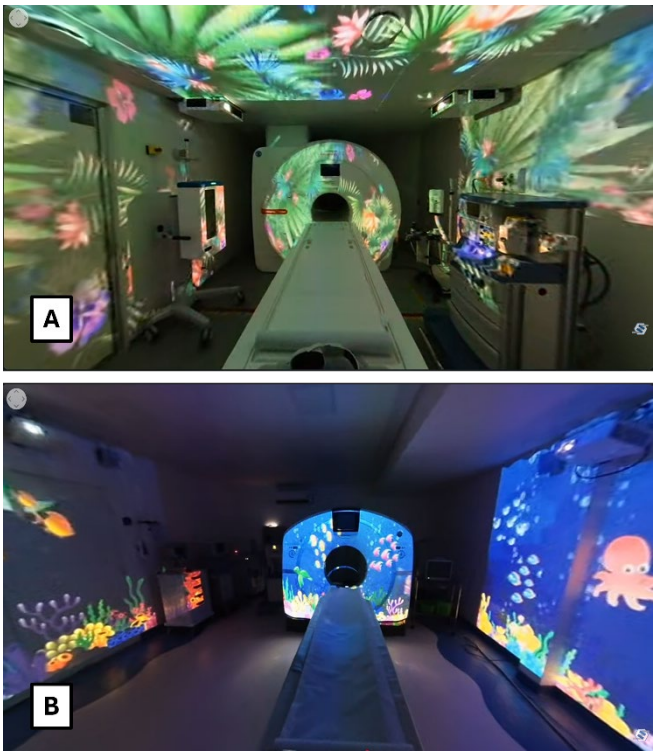


Fig. 5. Application of an immersive audiovisual projection system in diagnostic imaging environments: (A) magnetic resonance imaging room [19]; (B) computed tomography room [20].

Although initiatives of this nature remain limited and involve challenges related to implementation costs, infrastructure, and the evaluation of their clinical impacts, they point to the possibility of integrating environmental design, audiovisual resources, and immersive technologies into diagnostic imaging environments. In light of these findings, *Immersive Humanization* is proposed as a conceptual framework capable of guiding future research and supporting the development of solutions aimed at enhancing the patient experience.

V. Conclusion

This study conducted a scientometric analysis of the national and international scientific production on humanization strategies in diagnostic imaging environments to understand how the literature has addressed patient experience, healthcare environment, humanization of care, and the use of immersive technologies.

The search strategies retrieved 604 records across the different sources consulted, revealing a broad field of research distributed across different areas of knowledge and theoretical approaches. The analysis of the corpora made it possible to identify convergences among these perspectives, indicating that patient experience, evidence-based healthcare environments, humanization of care, and immersive technologies can be understood as complementary dimensions in the consideration of diagnostic imaging environments.

As its main contribution, this study proposes the concept of *Immersive Humanization*, understood as a conceptual framework that integrates established principles of humanization of care, *Evidence-Based Design*, *Positive Distraction*, and immersive technologies, without intending to replace them, but rather articulating them within a perspective directed toward diagnostic imaging environments.

As this is an exploratory study, supported by specific literature search strategies, the results should be interpreted in light of the methodological scope adopted, including the databases, search terms, and publication types considered. In this regard, future studies may expand the databases consulted, empirically validate the proposed framework, and investigate the effects of immersive solutions on objective indicators related to patient experience, anxiety, and the quality of diagnostic procedures.

It is expected that the findings presented herein will contribute to strengthening the dialogue among healthcare, design, communication, and immersive technologies, stimulating further research while supporting the development of more welcoming, humanized, and patient-centered diagnostic imaging environments.

CONFLICT OF INTEREST

The second author has a professional affiliation with MR In Site Healthcare Solutions, a company that develops immersive audiovisual technologies for healthcare environments.

REFERENCES

[1] W. Gobo, T. M. F. Meneses, and R. M. F. Vador, "Cuidado humanizado em ambientes tecnológicos: atuação do enfermeiro frente à ansiedade e claustrofobia em exames de imagem," *PBPC*, vol. 5, no. 4, pp. 1230–1240, Jun. 2026. doi: 10.36557/2674-9432.2026v5n4p1230-1240.

[2] V. M. Santos, "Estratégias de atendimento para pacientes claustrofóbicos em exames de ressonância magnética: uma revisão sistemática," *Rev. Remecs*, no. 2, p. 148, Aug. 2023. [Online]. Available: https://revistaremecs.com.br/index.php/remecs/article/view/1315.

[3] F. B. Dallarosa, *Humanização em salas de exames de ressonância magnética com enfoque no design para experiência*. M.S. thesis, Faculdade de Arquitetura, Programa de Pós-Graduação em Design, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, 2015. [Online]. Available: https://lume.ufrgs.br/handle/10183/127897.

[4] C. C. P. Fontainha, E. S. S. Ferreira, I. L. B. Roza, G. M. Matos, R. V. Teixeira, L. E. S. Nascimento, R. C. R. Quirino, J. P. M. X. Siqueira, M. R. P. S. Barros, L. H. S. Rocha, A. H. C. Santos, R. M. G. Gontijo, A. S. M. Batista, T. O. Santos, E. S. Júnior, and L. B. Nogueira, "Medo e fobia nos exames radiológicos: implementação de ações pela arte em um hospital universitário do SUS de Belo Horizonte," in *Proc. VII Semana Nacional de Engenharia Nuclear e da Energia e Ciências das Radiações (VII SENCIR)*, Belo Horizonte, Brazil, Nov. 2024. [Online]. Available: https://www.even3.com.br/anais/vii-sencir-semana-nacional-de-engenharia-nuclear-e-da-energia-e-ciencias-das-radiacoes-449507/927773-medo-e-fobia-nos-exames-radiologicos--implementacao-de-acoes-pela-arte-em-um-hospital-universitario-do-sus-de-bel.

[5] C. F. Paula, *O uso da hipnose, musicoterapia e psicologia da saúde, para minimizar a claustrofobia em exames de ressonância magnética, sem o uso de sedação: uma revisão da literatura*. B.S. thesis, Dept. Acad. de Física, Curso Superior de Tecnologia em Radiologia, Universidade Tecnológica Federal do Paraná, Curitiba, Brazil, 2023. [Online]. Available: http://repositorio.utfpr.edu.br/jspui/handle/1/35407.

[6] M. C. Castro, *A influência da comunicação na adesão das crianças à ressonância magnética sem anestesia*. M.S. thesis, Mestrado em Comunicação Clínica, Faculdade de Medicina, Universidade do Porto, Porto, Portugal, 2020. [Online]. Available: https://sigarra.up.pt/fmup/pt/teses.tese?p_aluno_id=109737&p_processo=21026&p_lang=1.

[7] A. L. J. Freitas and P. M. N. Perfeito, *Design gráfico aplicado à sala de tomografia visando o bem-estar das crianças*. B.S. thesis, Centro Universitário de Brasília (UniCEUB), Brasília, Brazil, 2015. [Online]. Available: https://repositorio.uniceub.br/jspui/handle/235/7655.

[8] S. R. de Carvalho, P. Â. C. H. dos Santos, A. A. F. G. Moreira, and A. M. C. Campos, "Utilização de recursos multimídia para crianças em contexto hospitalar: uma revisão integrativa," *Rev. Educ. Espec.*, vol. 33, Art. no. e10, pp. 1–21, Apr. 2020, doi: 10.5902/1984686X38894.

[9] J. A. O. Pereira, "Capacidades e limitações da realidade virtual no tratamento da claustrofobia: uma revisão da literatura," *Revista Multidisciplinar em Saúde*, vol. 5, no. 2, 2024, doi: 10.51161/cronics2024/36646.

[10] E. Zijlstra, M. Hagedoorn, W. P. Krijnen, C. P. van der Schans, and M. P. Mobach, "Motion nature projection reduces patient's psycho-physiological anxiety during CT imaging," *Journal of Environmental Psychology*, vol. 53, pp. 168–176, 2017, doi: 10.1016/j.jenvp.2017.07.010.

[11] P. Kukla, K. Maciejewska, I. Strojna, M. Zapał, G. Zwierzchowski, and B. Bąk, "Extended reality in diagnostic imaging—A literature review," *Tomography*, vol. 9, no. 3, pp. 1071–1082, 2023, doi: 10.3390/tomography9030088.

[12] R. S. Ulrich, L. L. Berry, X. Quan, and J. T. Parish, "A conceptual framework for the domain of evidence-based design," *HERD: Health Environments Research & Design Journal*, vol. 4, no. 1, pp. 95–114, 2010, doi: 10.1177/193758671000400107.

[13] C. E. Waugh, E. Z. Shing, and R. M. Furr, "Not all disengagement coping strategies are created equal: Positive distraction, but not avoidance, can be an adaptive coping strategy for chronic life stressors," *Anxiety, Stress, & Coping*, vol. 33, no. 5, pp. 511–529, 2020, doi: 10.1080/10615806.2020.1755820.

[14] A. Alhsainat and K. Günçe, "Healing environment in pediatric cancer centers by utilizing positive distractions," *Behavioral Sciences*, vol. 14, no. 11, Art. no. 1010, 2024, doi: 10.3390/bs14111010.

[15] D. M. Hudson, C. Heales, and S. J. Vine, "Scoping review: How is virtual reality being used as a tool to support the experience of undergoing magnetic resonance imaging?," *Radiography*, vol. 28, no. 1, pp. 199–207, 2022, doi: 10.1016/j.radi.2021.07.008.

[16] G. Bonnin and A. Goudey, "Creating a hospitable hospital: How colors, music, and scent in the servicescape reduce rumination and enhance the patient waiting experience," *International Journal of Hospitality Management*, vol. 128, Art. no. 104155, 2025, doi: 10.1016/j.ijhm.2025.104155.

[17] S. M. Oh, J. S. Park, B. Kim, S. Y. Jang, S. G. Lee, and T. H. Kim, "General hospitals with high-tech medical equipment and patient experience," *Journal of Patient Experience*, vol. 12, Art. no. 23743735251330468, 2025, doi: 10.1177/23743735251330468.

[18] A. Marín, "Critical study of fast immersive journalism: The Enemy as a paradigmatic case of slow and creative journalism," *BJR*, vol. 20, no. 1, p. e1558, Apr. 2024, doi: 10.25200/BJR.v20n1.2024.1558.

[19] MR In Site Healthcare Solutions, "Projeção imersiva MR in site – Ressonância Magnética – 360º VR experience," *YouTube*, Aug. 4, 2025. [Online]. Available: https://www.youtube.com/watch?v=xbUk0WAqrlo.

[20] MR In Site Healthcare Solutions, "Projeção imersiva MR in site – Tomografia Computadorizada – 360º VR experience," *YouTube*, Aug. 4, 2025. [Online]. Available: https://youtu.be/HU6Zq1CEeDk.

[21] Hermes Pardini, "Realidade virtual transforma a experiência da vacinação infantil," YouTube, May 4, 2017. [Online]. Available: https://www.youtube.com/watch?v=P9JwAH0298w.

[22] Tech SC, "Realidade virtual para potencializar o tratamento contra medo, fobia e síndrome do pânico," Globoplay, 2026. [Online]. Available: https://globoplay.globo.com/v/14172765/

[23] L. L. Gonçalves, J. C. Appolinario, A. E. Nardi, and R. Ferreira-Garcia, Eds., *Realidade virtual na saúde mental*, 1st ed. Porto Alegre, Brazil: Artmed, 2025.

[24] A. Pessoni and T. Passaro, "Os 12 eixos da comunicação combinada para a saúde pública municipal: uma proposta," *RECIIS – Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, vol. 15, no. 1, 2021, doi: 10.29397/reciis.v15i1.2049.

[25] R. K. J. Brown, S. Petty, S. O'Malley, J. Stojanovska, M. S. Davenport, E. A. Kazerooni, and D. Fessahazion, "Virtual reality tool simulates MRI experience," *Tomography*, vol. 4, no. 3, pp. 95–98, 2018, doi: 10.18383/j.tom.2018.00023.

Sistemas de Detecção de Materiais Ferromagnéticos em Ressonância Magnética: Tecnologias, Segurança e Sustentabilidade Operacional

1st Tiago Gois Falandes
MR In Site Healthcare Solutions
São Bernardo do Campo, Brasil
tiago.falandes@mrinsite.com.br

2nd Carolina Gois Falandes
Departamento de Jornalismo e Editoração (CJE)
Escola de Comunicações e Artes, Universidade de São Paulo (USP)
São Paulo, Brasil
https://orcid.org/0000-0001-6639-2121

***Resumo*—A ressonância magnética (RM) é uma modalidade fundamental de diagnóstico por imagem na medicina moderna, destacando-se por não utilizar radiação ionizante. No entanto, a presença de um campo magnético estático de alta intensidade torna o ambiente de RM suscetível a acidentes decorrentes da introdução de objetos ferromagnéticos, entre os quais se incluem incidentes classificados como "*projectile event*". Tais ocorrências podem resultar em lesões graves, óbitos, danos ao equipamento e impactos operacionais e financeiros. Nesse contexto, os Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) constituem um recurso complementar de mitigação de riscos, permitindo a identificação de objetos potencialmente perigosos antes da entrada em áreas controladas de RM. Este estudo apresenta uma análise técnico-narrativa dos principais riscos associados ao ambiente de RM, das normas e diretrizes nacionais e internacionais aplicáveis e das tecnologias e configurações de sistemas de detecção. São discutidas as características, os princípios de funcionamento, as limitações e as aplicações de detectores portáteis (manuais), do tipo barra e do tipo portal, considerando sua integração aos procedimentos de segurança em RM.**

***Palavras-chave*—Ressonância Magnética (RM), Segurança em Ressonância Magnética, Sistemas de Detecção de Materiais Ferromagnéticos (FMDS), Segurança do Paciente.**

I. Introdução

A ressonância magnética (RM) é amplamente utilizada no diagnóstico por imagem e é considerada uma modalidade relativamente segura, além de não utilizar radiação ionizante [1]. No entanto, a presença de um campo magnético estático de alta intensidade, que permanece ativo continuamente, introduz riscos específicos que exigem medidas de controle adequadas.

Entre esses riscos, os incidentes com projéteis – causados pela introdução inadvertida de objetos ferromagnéticos no ambiente de RM – representam um importante perigo à segurança. Objetos aparentemente inofensivos, como tesouras, cilindros de oxigênio, cadeiras de rodas, injetores de contraste e ferramentas de manutenção, podem ser atraídos em direção ao ímã, resultando em danos ao equipamento, interrupções operacionais e, em casos graves, lesões ou óbitos.

Para fins de comparação, o campo magnético da Terra varia de aproximadamente 25 a 65 μT [2], enquanto um sistema de RM de 1,5 T produz aproximadamente 1.500.000 μT, quase 30.000 vezes mais que o valor terrestre. Essa diferença evidencia a elevada intensidade presente nos sistemas de RM e a necessidade de medidas adequadas de controle e segurança.

Segundo dados do Cadastro Nacional de Estabelecimentos de Saúde (CNES), o Brasil contava com 3.181 sistemas de ressonância magnética em operação em maio de 2026 [3]. A presença de um número expressivo de equipamentos de RM em funcionamento reforça a importância de procedimentos adequados de segurança, especialmente diante dos riscos associados à introdução inadvertida de materiais ferromagnéticos nesses ambientes.

Nesse contexto, a segurança em RM envolve diferentes medidas de controle, incluindo procedimentos de triagem, treinamento de pessoal, controle de acesso e identificação de objetos e equipamentos potencialmente perigosos. Os Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) podem constituir uma camada adicional de proteção, contribuindo para a identificação de materiais ferromagnéticos antes de seu ingresso em áreas controladas de RM.

Assim, este artigo analisa as principais tecnologias de detecção de materiais ferromagnéticos utilizadas em ambientes de RM, discutindo seus princípios de funcionamento, características, limitações e aplicações práticas. Além disso, são apresentadas as principais normas e diretrizes nacionais e internacionais relacionadas à segurança em RM, considerando o papel dos FMDS como uma camada adicional de mitigação de riscos em programas abrangentes de segurança em RM.

II. Materiais e Métodos

Este estudo caracteriza-se como uma análise técnico-narrativa sobre sistemas de detecção de materiais ferromagnéticos aplicados à segurança em ambientes de ressonância magnética. A investigação foi desenvolvida por meio de pesquisa bibliográfica não sistemática, pesquisa documental e análise reflexiva da experiência profissional do autor.

A pesquisa bibliográfica foi realizada por meio de buscas gerais na web, com o objetivo de identificar publicações relacionadas à segurança em ressonância magnética, incidentes envolvendo materiais ferromagnéticos e tecnologias de detecção. Já a pesquisa documental contemplou normas, diretrizes, documentos regulatórios e materiais técnicos relacionados aos sistemas de detecção. Esses materiais foram utilizados para caracterizar funcionalidades e aspectos operacionais, não sendo considerados evidências

independentes de desempenho. Complementarmente, a experiência profissional do autor no contexto de tecnologias aplicadas à segurança em ressonância magnética foi considerada na análise dos aspectos técnicos e operacionais discutidos no estudo.

As informações identificadas foram organizadas e examinadas considerando os princípios de funcionamento, características, aplicações, limitações e aspectos operacionais das diferentes tecnologias de detecção, bem como as recomendações normativas relacionadas à segurança em ambientes de ressonância magnética.

III. O CASO MICHAEL COLOMBINI: UM MARCO NA SEGURANÇA EM RM

Apesar dos avanços na Ressonância Magnética (RM), o histórico de segurança dessa modalidade abrange incidentes que evidenciaram os riscos associados aos campos magnéticos estáticos de alta intensidade. Entre esses eventos, o caso de Michael Colombini tornou-se um marco nessa discussão [4].

Em 27 de julho de 2001, Michael Colombini, uma criança de seis anos, submetia-se a um exame de rotina de RM no *Westchester Medical Center*, em Nova York (EUA). Durante o procedimento, um cilindro de oxigênio de aço, classificado como inseguro para RM (*MR Unsafe*), foi inadvertidamente levado para a sala de exames. O cilindro foi fortemente atraído pelo campo magnético e atingiu a cabeça de Michael, provocando lesões que resultaram em sua morte [4], [5].

O acidente revelou falhas em diferentes aspectos das práticas de segurança em ressonância magnética (RM), incluindo questões relacionadas ao treinamento, à triagem de materiais e ao controle de acesso. O episódio também demonstrou que, embora a RM não utilize radiação ionizante, seu campo magnético estático pode representar riscos graves quando as medidas de segurança e controle não são adequadamente aplicadas [6].

Após o ocorrido, o *American College of Radiology* (ACR) publicou, em 2002, um documento de consenso sobre segurança em RM, posteriormente revisado e atualizado [7]. Essas recomendações contribuíram para a consolidação de práticas estruturadas de segurança, incluindo o zoneamento dos serviços de RM em quatro áreas (Zonas I a IV), o controle de acesso e o treinamento dos profissionais envolvidos no ambiente de RM [7].

O período posterior ao caso também coincide com o desenvolvimento de Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) específicos para ambientes de RM. Segundo Keene [8], os primeiros sistemas desse tipo surgiram em 2002, pouco depois do acidente, oferecendo uma camada adicional de proteção contra incidentes do tipo “projétil”.

Mais de duas décadas depois, o caso Michael Colombini permanece como uma referência para o debate sobre segurança em RM, evidenciando que esta depende não apenas de recursos tecnológicos, mas também de treinamento, procedimentos adequados, controle de acesso e múltiplas camadas de proteção [6], [7].

IV. O CASO LEANDRO MATHIAS E OS RISCOS EM AMBIENTES DE RM

Embora acidentes relacionados à ressonância magnética (RM) sejam relativamente raros, alguns podem apresentar consequências graves, especialmente aqueles envolvendo objetos ferromagnéticos introduzidos inadvertidamente no ambiente de exame [9]. Em janeiro de 2023, um novo incidente voltou a evidenciar que esses riscos permanecem presentes nos ambientes de RM [10], [11].

Leandro Mathias de Novaes, advogado de 40 anos, acompanhava sua mãe durante um exame de ressonância magnética em São Paulo, Brasil, quando a arma de fogo que portava foi atraída pelo campo magnético do equipamento e disparou, atingindo seu abdômen. O incidente provocou graves lesões, e Leandro faleceu em 6 de fevereiro de 2023, cerca de três semanas após o ocorrido [10].

O episódio ilustra que os riscos associados à introdução de materiais ferromagnéticos em ambientes de RM não se restringem a equipamentos médicos, cilindros ou ferramentas, podendo também envolver objetos pessoais introduzidos inadvertidamente nas áreas controladas. A presença de uma arma de fogo no ambiente de RM resultou em um evento grave e reforça a importância de procedimentos adequados de triagem, da comunicação eficaz com pacientes e acompanhantes, bem como do controle de acesso às áreas restritas. Nesse contexto, diferentes camadas de proteção podem ser adotadas de forma complementar, incluindo Sistemas de Detecção de Materiais Ferromagnéticos (FMDS), conforme as características do serviço e do programa de segurança. O exemplo também mostra que os recursos tecnológicos devem estar integrados a procedimentos e treinamento, uma vez que nenhuma medida isolada elimina completamente os riscos associados ao ambiente de RM.

V. CONTEXTO E RELEVÂNCIA DOS SISTEMAS DE DETECÇÃO FERROMAGNÉTICA

Estudos sobre segurança em ressonância magnética (RM) indicam que muitos incidentes estão ligados a fatores humanos e a falhas nos processos de segurança, incluindo deficiências na triagem, treinamento inadequado, descumprimento de procedimentos e lacunas no controle de acesso [6], [7]. Embora acidentes relacionados à RM sejam relativamente raros, alguns podem apresentar consequências graves quando ocorrem [9].

A gravidade desses incidentes está relacionada, entre outros aspectos, à intensidade do campo magnético presente nos sistemas de RM. Objetos ferromagnéticos podem ser fortemente atraídos pelo magneto e adquirir velocidade suficiente para causar lesões ou danos ao equipamento. Um exemplo frequentemente utilizado na literatura é o de objetos ferromagnéticos que podem atingir velocidades superiores a 60 km/h em determinadas condições [8].

Além do risco de lesões, eventos de efeito projétil podem resultar em danos aos equipamentos, cancelamento de exames, interrupções não planejadas das atividades, necessidade de intervenções técnicas e prejuízos financeiros substanciais [8], [12]. Estimativas apresentadas em material técnico sobre

segurança em RM indicam que os custos de acidentes envolvendo efeito projétil, sem lesão humana, podem variar de aproximadamente US$ 20.000 a mais de US$ 200.000, dependendo da natureza do incidente e de suas consequências [8].

No contexto brasileiro, um levantamento realizado em 29 instituições de saúde de Porto Alegre, abrangendo 49 equipamentos de RM, identificou diferentes estratégias de detecção de metais nos serviços avaliados [13]. A maioria das instituições utilizava entrevista de triagem associada a detector portátil, enquanto alguns serviços combinavam detectores portais e portáteis. O estudo também identificou instituições que utilizavam exclusivamente a entrevista e serviços que não informaram, em seus Programas de Segurança em RM, como era realizada a detecção de metais [13]. Esses resultados evidenciam a diversidade das estratégias adotadas na prática e reforçam a importância da definição e padronização dos procedimentos de triagem nos serviços de RM.

O levantamento também identificou lacunas na documentação relacionada à segurança. Apenas 55,17% das instituições avaliadas apresentavam descrição dos riscos inerentes à RM e 58,62% apresentavam descrição da compatibilidade dos materiais com o ambiente de RM [13]. Esses achados reforçam que a utilização de sistemas de detecção deve estar inserida em um programa de segurança mais amplo, que envolva procedimentos de triagem, treinamento, avaliação de riscos e controle de acesso.

Nesse contexto, os Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) constituem um recurso complementar para a mitigação de riscos em ambientes de RM. Diferentemente dos detectores de metais convencionais, os FMDS são projetados para detectar perturbações no campo magnético associadas à presença de materiais ferromagnéticos, permitindo maior seletividade quanto às propriedades magnéticas dos objetos avaliados [8].

As recomendações do *American College of Radiology* (ACR) incluem os sistemas de detecção ferromagnética como recurso adicional aos procedimentos de triagem, especialmente antes do acesso às áreas controladas da RM [14]. Dessa forma, os FMDS devem ser compreendidos como uma camada complementar de proteção, integrada aos procedimentos de triagem, controle de acesso, treinamento e protocolos operacionais padronizados.

VI. Normas e diretrizes aplicadas à detecção de materiais ferromagnéticos em ambientes de ressonância magnética

No Brasil, as principais referências regulatórias relacionadas à segurança em ambientes de ressonância magnética (RM) incluem a Instrução Normativa nº 97/2021 da ANVISA e a RDC nº 611/2022. Esses documentos estabelecem requisitos relacionados à segurança dos ambientes de RM, incluindo medidas de detecção de metais e controle do acesso às áreas restritas [15], [16]. As normas não especificam uma tecnologia única para a detecção, permitindo diferentes abordagens de acordo com as características e os procedimentos de segurança adotados pelos serviços [15], [16].

Internacionalmente, o *American College of Radiology* (ACR) recomenda uma abordagem baseada em múltiplas camadas de proteção para a segurança em RM, incluindo procedimentos de triagem, inspeção de pessoas e objetos, treinamento da equipe, sinalização adequada e controle de acesso. O ACR recomenda a triagem de pessoas antes do acesso às Zonas III e IV e considera os Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) um recurso complementar às medidas de segurança estabelecidas para o ambiente de RM [14].

Na Europa, a Diretiva 2013/35/UE estabelece requisitos para a proteção dos trabalhadores contra riscos associados à exposição a campos eletromagnéticos, incluindo efeitos indiretos relacionados à presença de materiais ferromagnéticos [17]. A diretiva, entretanto, não estabelece uma exigência específica para a utilização de FMDS em instalações de RM. Orientações da *Medicines and Healthcare products Regulatory Agency* (MHRA) também abordam os riscos associados à introdução de objetos metálicos em ambientes de RM e apresentam recomendações relacionadas aos procedimentos de segurança e triagem [18].

De modo geral, os documentos regulatórios e as diretrizes analisados indicam que a segurança em ambientes de RM deve ser estruturada por meio de medidas complementares de controle e triagem. No Brasil, os requisitos regulatórios estabelecem medidas de segurança e detecção, sem determinar uma tecnologia específica. Internacionalmente, o ACR recomenda a utilização de FMDS como recurso complementar às medidas de triagem e segurança antes do acesso às áreas controladas da RM [14]. Dessa forma, a adoção de sistemas de detecção deve ser considerada em conjunto com os procedimentos de triagem, treinamento, controle de acesso e demais medidas estabelecidas no programa de segurança da instituição.

VII. Tecnologias de detecção aplicadas em ambientes de ressonância magnética

A prevenção de acidentes em ambientes de Ressonância Magnética (RM) baseia-se na implementação de sistemas capazes de identificar objetos potencialmente perigosos antes que estes ingressem nas Zonas III e IV [14]. Na prática, podem ser encontrados três tipos de tecnologia de detecção em aplicações de segurança em RM: detectores de metais convencionais, detectores ferromagnéticos e sistemas de detecção híbridos. Cada tecnologia apresenta características, vantagens e limitações distintas, que devem ser consideradas no planejamento e na implementação de programas de segurança em RM.

A. Detectores de metais convencionais

Detectores de metais convencionais (Figura 1) funcionam gerando campos eletromagnéticos alternados e analisando perturbações causadas pela presença de objetos metálicos.

Esses sistemas são amplamente utilizados em aeroportos e aplicações de segurança, pois conseguem detectar diferentes tipos de materiais metálicos, incluindo materiais ferromagnéticos e não ferromagnéticos.

No entanto, em ambientes de ressonância magnética (RM), essa característica pode representar uma limitação, uma vez que nem todos os metais apresentam risco de efeito projétil. Ao não diferenciarem materiais ferromagnéticos de materiais não ferromagnéticos, esses detectores podem gerar alarmes para objetos que não representam necessariamente risco de interação com o campo magnético, exigindo avaliação adicional durante a triagem. Essa limitação é particularmente relevante nas áreas de acesso às Zonas III e IV, nas quais o controle de objetos e equipamentos metálicos constitui parte importante das práticas de segurança em RM [14].

Outra limitação importante está relacionada à utilização desses equipamentos nas proximidades de campos magnéticos estáticos intensos. Quando instalados nas proximidades das Zonas III e IV da RM, sua utilização deve considerar as características do equipamento, sua compatibilidade com o ambiente de RM e sua localização em relação ao magneto. Estudos que avaliaram diferentes dispositivos de detecção também indicam que o desempenho pode variar de acordo com o tipo e o tamanho do objeto, sua posição e a distância em relação ao detector [19].

Por essas razões, os detectores de metais convencionais não devem ser considerados substitutos dos procedimentos específicos de triagem e controle de acesso aplicáveis à segurança em RM. As recomendações do *American College of Radiology* (ACR) indicam a utilização de procedimentos complementares de detecção e avaliação de objetos potencialmente perigosos, incluindo sistemas de detecção ferromagnética, no contexto do programa de segurança em RM [14].



Fig. 1. Detectores de metais convencionais. Fonte: O autor (2026).

B. Detectores ferromagnéticos

Os Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) foram desenvolvidos especificamente para aplicações em ressonância magnética (RM). Seu princípio de funcionamento baseia-se na detecção de perturbações no campo magnético ambiente causadas pela presença de materiais ferromagnéticos. Como resultado, materiais não ferromagnéticos – como alumínio, cobre, titânio e alguns materiais não ferromagnéticos utilizados em dispositivos médicos – geralmente não acionam alarmes.

Uma das principais características dos FMDS (Figura 2) é a seletividade para materiais ferromagnéticos, o que os diferencia dos detectores de metais convencionais. Essa característica permite direcionar a detecção para objetos que apresentam potencial de interação com o campo magnético e risco de efeito projétil, reduzindo a necessidade de investigação de objetos metálicos que não apresentam esse tipo de propriedade.

Evidências experimentais e clínicas demonstram que o desempenho dos FMDS pode variar de acordo com o sistema utilizado e as condições de avaliação. Em um estudo clínico de implementação, foram realizados 1.032 eventos de triagem em 977 pacientes, com 55 resultados verdadeiramente positivos, 21 falsos positivos e nenhum falso negativo. Nesse estudo, a sensibilidade foi de 100% e a especificidade de 98%, e o tempo adicional associado à etapa de detecção foi considerado pequeno [20]. Em outro estudo, que avaliou 67 implantes e outros objetos em condições experimentais, foram observados 58 verdadeiros positivos, quatro verdadeiros negativos, nenhum falso positivo e cinco falsos negativos, correspondendo a uma sensibilidade de 92% e especificidade de 100% [21]. Esses resultados indicam que o desempenho dos FMDS deve ser interpretado considerando o equipamento, o objeto avaliado e as condições de triagem.

A relação entre sensibilidade, especificidade e ocorrência de falsos alarmes também deve ser considerada na análise dos sistemas, uma vez que esses parâmetros influenciam a necessidade de avaliação adicional durante o processo de triagem [20], [21].

Diretrizes do *American College of Radiology* (ACR) reconhecem os FMDS como recurso complementar aos procedimentos de triagem de segurança em RM, e não como substitutos desses procedimentos. Dessa forma, sua utilização deve ser integrada às medidas de controle de acesso, triagem clínica e administrativa e avaliação da compatibilidade de dispositivos e implantes com a RM [14].

Embora os FMDS sejam direcionados à detecção de materiais ferromagnéticos, determinados objetos, implantes e dispositivos médicos podem não ser detectados ou podem exigir avaliação complementar. Assim, o resultado do sistema deve ser interpretado em conjunto com os demais procedimentos estabelecidos no programa de segurança em RM.

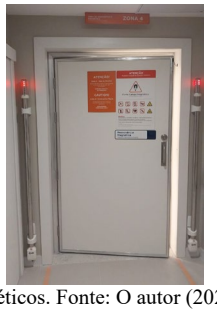


Fig. 2. Detectores ferromagnéticos. Fonte: O autor (2026).

C. Detectores Híbridos

Sistemas de detecção híbridos (Figura 3) combinam as funcionalidades de detectores de metais convencionais e detectores ferromagnéticos em um único dispositivo. Esses

equipamentos podem detectar objetos metálicos em geral e, adicionalmente, identificar materiais ferromagnéticos, permitindo diferenciar, durante a triagem, os objetos que apresentam propriedades ferromagnéticas daqueles que não apresentam.

Alguns dispositivos fornecem indicações separadas para materiais metálicos e ferromagnéticos, o que pode auxiliar na interpretação dos alarmes e ampliar a flexibilidade operacional. No entanto, os sistemas híbridos podem apresentar maior complexidade de operação e exigir treinamento específico dos operadores, dependendo da configuração e das características do equipamento. Aspectos como custo de aquisição, manutenção e requisitos de treinamento devem ser avaliados de acordo com a aplicação e o modelo considerado.



Fig. 3. Detectores híbridos. Fonte: O autor (2026).

D. Comparação de tecnologias

De modo geral, os detectores de metais convencionais apresentam capacidade de detectar diferentes materiais metálicos, incluindo materiais ferromagnéticos e não ferromagnéticos. Essa característica amplia o conjunto de objetos identificados, mas limita a diferenciação entre materiais que apresentam ou não propriedades ferromagnéticas, o que pode exigir avaliação adicional durante a triagem.

Em contrapartida, os sistemas de detecção de materiais ferromagnéticos (FMDS) são direcionados à identificação de objetos com propriedades ferromagnéticas, característica diretamente relacionada à prevenção de eventos de efeito projétil em ambientes de RM. Estudos experimentais e clínicos demonstram que o desempenho pode variar conforme o sistema, o objeto investigado e as condições de utilização [20], [21].

Os sistemas híbridos buscam combinar diferentes princípios de detecção em um mesmo equipamento, podendo ampliar as possibilidades de triagem conforme sua configuração. A escolha da tecnologia deve considerar o objetivo da detecção, o perfil operacional da instituição, o fluxo de pacientes, as características físicas do ambiente, a compatibilidade do equipamento com o campo magnético, os protocolos de segurança, os requisitos de treinamento e manutenção e as regulamentações aplicáveis [14].

VIII. Tipos de detectores utilizados em ambientes de RM

Os sistemas de detecção para RM podem ser encontrados, de acordo com sua configuração física, em formatos portáteis (manuais), do tipo barra e do tipo portal. A escolha da solução mais adequada depende de fatores como o nível de segurança requerido, o fluxo de pacientes, o layout das instalações e a estratégia de gestão de riscos da instituição.

A. Detector portátil

Detectores manuais são dispositivos portáteis alimentados por bateria, destinados à inspeção de pacientes, funcionários, visitantes e objetos. Essa categoria abrange detectores de metais convencionais, ferromagnéticos e híbridos. Sua principal vantagem é a flexibilidade operacional, permitindo a inspeção de pacientes acamados, cadeiras de rodas, macas, incubadoras, cilindros de gases medicinais e equipamentos médicos móveis [22].

O desempenho dos detectores portáteis depende do princípio de detecção, da configuração do equipamento e das características do objeto avaliado. A documentação técnica consultada estabelece procedimentos específicos de varredura e indica a influência das condições de utilização sobre o processo de detecção [22].

Embora possam complementar sistemas fixos, os detectores manuais envolvem inspeção realizada diretamente pelo operador, tornando treinamento, padronização da técnica e verificações periódicas fatores relevantes para sua utilização [22]. Seu uso não substitui os demais procedimentos de segurança estabelecidos para a triagem antes do acesso às áreas controladas da RM [14].

B. Detectores do tipo barra

Detectores do tipo barra consistem em um ou mais sensores fixos instalados próximos a portas, corredores ou pontos de acesso às Zonas III e IV de RM. Os sistemas disponíveis para aplicações em RM utilizam predominantemente tecnologia de detecção de materiais ferromagnéticos e são projetados para operar em ambientes com campos magnéticos estáticos intensos.

Esses sistemas monitoram a passagem de pessoas e objetos, emitindo alarmes visuais e sonoros quando materiais ferromagnéticos são detectados. Em comparação com os detectores portáteis, podem proporcionar maior padronização do processo de triagem e menor dependência da atuação do operador. A configuração e o espaçamento entre sensores devem considerar as características do sistema e as condições específicas de instalação [19].

C. Detectores do tipo portal

Detectores do tipo portal consistem em duas colunas de sensores posicionadas em lados opostos de uma passagem, permitindo o monitoramento de pessoas e objetos que atravessam a área de detecção. Podem utilizar tecnologias convencionais de detecção de metais ou de detecção de materiais ferromagnéticos.

Os portais convencionais identificam diferentes tipos de materiais metálicos, mas não distinguem materiais ferromagnéticos de não ferromagnéticos, constituindo uma limitação para aplicações específicas em RM [14]. Já os portais

ferromagnéticos são projetados para identificar materiais com potencial de interação com o campo magnético e podem proporcionar triagem contínua nos pontos de passagem, inclusive como parte das medidas de controle de acesso às áreas controladas [14].

Características como distância entre sensores, largura da passagem e proximidade de estruturas metálicas devem ser consideradas na configuração e avaliação do sistema [19], [22].

D. Comparação entre os tipos de detectores

Os detectores portáteis oferecem maior flexibilidade operacional, enquanto os sistemas do tipo barra e do tipo portal permitem a detecção em pontos fixos de passagem, podendo proporcionar maior padronização do processo de triagem, conforme sua configuração e as condições de instalação [22].

Entre as tecnologias disponíveis, os sistemas de detecção ferromagnética apresentam como característica a seletividade para materiais ferromagnéticos, aspecto particularmente relevante em ambientes de ressonância magnética (RM), nos quais o potencial de interação com o campo magnético constitui um dos principais critérios de segurança [8]. Estudos de desempenho demonstram, entretanto, que a capacidade de detecção pode variar conforme o sistema utilizado, o tipo e o tamanho do objeto e as condições de avaliação [19], [20], [21]. O Quadro 1 sintetiza as principais características, aplicações, vantagens e limitações dos diferentes tipos de detectores discutidos.

Nenhuma tecnologia de detecção, isoladamente, elimina os riscos associados à RM. A utilização desses sistemas deve estar integrada a um programa de segurança abrangente, que inclua procedimentos de triagem, treinamento da equipe, controle de acesso e identificação da compatibilidade de pacientes, equipamentos e objetos com o ambiente de RM [14]. Nesse contexto, os sistemas de detecção ferromagnética devem ser compreendidos como recursos complementares às demais medidas de segurança.

QUADRO 1. COMPARAÇÃO ENTRE TIPOS DE DETECTORES

Característica	**Detector de metais convencional**	**Detector ferromagnético (FMDS)**	**Detector híbrido**
Princípio de detecção	Detecta diferentes materiais metálicos por meio de campos eletromagnéticos	Detecta perturbações associadas à presença de materiais ferromagnéticos	Combina modos de detecção de materiais ferromagnéticos e, conforme o equipamento, outros metais
Seletividade	Baixa para aplicações em RM, pois não diferencia materiais ferromagnéticos de não ferromagnéticos	Direcionada a materiais ferromagnéticos	Pode combinar diferentes modos de detecção
Aplicação em RM	Uso complementar; não substitui os procedimentos específicos de triagem em RM [14]	Desenvolvido especificamente para aplicações de segurança em RM [8]	Depende das características e da configuração do equipamento
Dependência do operador	Varia conforme a configuração	Tende a ser menor em sistemas fixos; presente em sistemas portáteis	Varia conforme a configuração
Sensibilidade	Depende do equipamento e do objeto avaliado	Depende do sistema, do objeto e das condições de avaliação [19], [20], [21]	Depende do modo de operação e do equipamento
Falsos alarmes	Pode detectar objetos metálicos que não apresentam risco ferromagnético	Pode apresentar alarmes associados às condições de detecção e ao objeto avaliado	Depende do modo de detecção e da configuração
Identificação do risco de efeito projétil	Limitada, pois não diferencia propriedades magnéticas	Maior direcionamento para materiais ferromagnéticos	Depende do modo de detecção
Aplicações típicas	Aplicações gerais de segurança e detecção de metais	Triagem e controle de acesso em ambientes de RM	Triagem em RM, quando o equipamento oferece modos combinados
Principais vantagens	Ampla capacidade de detecção de materiais metálicos	Seletividade para materiais ferromagnéticos e aplicação específica em RM	Flexibilidade de modos de detecção
Principais limitações	Não diferencia materiais ferromagnéticos de não ferromagnéticos	Não substitui os demais procedimentos de segurança e não detecta necessariamente todos os objetos relevantes	Desempenho e aplicação dependem do modelo e do modo de operação

Fonte: Elaborado pelo autor (2026).

IX. IMPACTOS ECONÔMICOS DE ACIDENTES ENVOLVENDO O EFEITO PROJÉTIL EM AMBIENTES DE RESSONÂNCIA MAGNÉTICA

Embora os acidentes envolvendo materiais ferromagnéticos em ambientes de ressonância magnética (RM) sejam discutidos principalmente sob a ótica da segurança do paciente, suas consequências econômicas também podem ser significativas. Um incidente do tipo "projétil" pode resultar em danos ao equipamento, interrupções operacionais, cancelamentos de exames, necessidade de manutenção corretiva e outros impactos associados à continuidade dos serviços [8], [12].

Em alguns casos, a ocorrência de um incidente pode exigir a interrupção das atividades na sala de RM para avaliação técnica e adoção das medidas necessárias para restabelecer as condições de operação. A duração dessa interrupção e seus custos associados dependem da natureza do incidente, dos danos provocados e das condições específicas do equipamento e da instituição. Estimativas para o custo de acidentes relacionados ao efeito projétil, sem lesão humana, variam de US$ 20.000 a mais de US$ 200.000 [8].

Um *quench* não planejado, caracterizado pela perda da supercondutividade e pela extinção do campo magnético, pode representar uma consequência de elevado impacto operacional em sistemas de RM supercondutores. Além da perda de hélio criogênico, o evento pode envolver custos relacionados ao reabastecimento, transporte e consultoria técnica, bem como período de indisponibilidade do equipamento para realização de exames [7], [23].

O impacto financeiro do tempo de inatividade também pode ser estimado a partir da capacidade operacional do equipamento. Como exemplo hipotético, considerando um sistema de RM que realize 25 exames por dia e um valor médio de R$ 500 por exame, uma interrupção de 20 dias poderia representar uma receita potencial não realizada de R$ 250.000, conforme apresentado no Quadro 2.

Esse cálculo representa uma estimativa ilustrativa de receita não realizada, e não uma estimativa do custo efetivo de um acidente. O impacto econômico real depende de fatores como o volume de exames, o modelo de remuneração, os custos fixos, os custos de manutenção, a duração da indisponibilidade e a possibilidade de transferência dos exames para outros equipamentos.

QUADRO 2. EXEMPLO ILUSTRATIVO DO IMPACTO ECONÔMICO DA INDISPONIBILIDADE DE UM EQUIPAMENTO DE RM

Item	Premissa/estimativa
Exames realizados por dia	25
Valor médio considerado por exame	R$ 500
Receita potencial diária	R$ 12.500
Período hipotético de inatividade	20 dias
Receita potencial não realizada	R$ 250.000

Nota: Os valores apresentados constituem uma simulação hipotética destinada a ilustrar o impacto econômico da indisponibilidade de um equipamento de RM. Não representam custos médios de acidentes nem estimativas de retorno sobre investimento.

Sob a ótica da Engenharia Clínica, a análise econômica dos incidentes deve considerar não apenas os custos diretos de reparo, mas também os efeitos associados à indisponibilidade do equipamento e à continuidade dos serviços. Nesse contexto, medidas de segurança, incluindo procedimentos de triagem, treinamento da equipe, controle de acesso e sistemas de detecção ferromagnética, podem ser analisadas como componentes de uma estratégia integrada de gerenciamento de riscos operacionais.

A análise evidencia que a indisponibilidade de equipamentos de RM pode representar impacto econômico relevante para os serviços de diagnóstico por imagem. A avaliação do custo-benefício da implementação de sistemas de detecção ferromagnética, entretanto, requer dados específicos sobre custos de aquisição e manutenção, frequência de incidentes, desempenho dos sistemas e perdas potencialmente evitadas.

X. PERSPECTIVAS FUTURAS

A evolução dos Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) está associada às demandas por maior segurança em ambientes de ressonância magnética (RM), padronização dos processos de triagem, rastreabilidade e continuidade operacional. Nesse contexto, observa-se uma tendência de integração dos sistemas de detecção com outros recursos de segurança e gerenciamento das instalações de RM.

Os sistemas FMDS disponíveis atualmente podem incorporar recursos como monitoramento remoto, armazenamento de eventos, gerenciamento de alarmes e integração com sistemas de controle de acesso, de acordo com as características de cada equipamento [22]. Recursos de inteligência artificial (IA) e análise de dados também apresentam possibilidades de aplicação na identificação de padrões de eventos, apoio à gestão de alarmes e análise de informações operacionais. Essas aplicações, entretanto, ainda devem ser avaliadas quanto à sua efetividade e confiabilidade em contextos específicos de segurança em RM.

Outra possibilidade é a integração dos sistemas de detecção com mecanismos de controle de acesso, dispositivos de sinalização e outros elementos da infraestrutura das salas de RM. Nessas configurações, informações provenientes da triagem podem ser utilizadas em conjunto com mecanismos de controle de acesso às áreas restritas e sistemas de comunicação de eventos. A adoção de mecanismos de intertravamento deve considerar os requisitos de segurança, o fluxo operacional e a necessidade de preservar procedimentos de intervenção e controle manual.

Esses desenvolvimentos apontam para uma possível evolução das instalações de RM em direção a ambientes mais integrados, nos quais sistemas de detecção, controle de acesso, sinalização e gerenciamento de informações possam atuar de forma coordenada. A convergência entre detecção de materiais ferromagnéticos, automação, análise de dados e recursos de Saúde Digital constitui uma perspectiva para o desenvolvimento de sistemas de segurança mais integrados e rastreáveis.

XI. CONCLUSÃO

A ressonância magnética (RM) consolidou-se como uma das principais modalidades de diagnóstico por imagem na medicina moderna. Entretanto, a presença de um campo magnético estático de alta intensidade envolve riscos específicos, demandando medidas de segurança adequadas e continuamente aprimoradas.

Acidentes envolvendo a introdução de objetos ferromagnéticos em ambientes de RM demonstram que falhas nos processos de triagem e controle de acesso podem resultar em lesões graves, óbitos, danos ao equipamento, interrupções operacionais e impactos financeiros. Consequentemente, a segurança em RM deve ser compreendida como um processo multidisciplinar baseado em múltiplas camadas de proteção.

Nesse contexto, os Sistemas de Detecção de Materiais Ferromagnéticos (FMDS) constituem um recurso complementar aos procedimentos de segurança em RM. Diferentemente dos detectores de metais convencionais, esses sistemas são direcionados à identificação de materiais ferromagnéticos, característica relevante para a avaliação de objetos com potencial de interação com o campo magnético. As diferentes configurações de FMDS, incluindo sistemas portáteis, do tipo barra e do tipo portal, apresentam características operacionais distintas e devem ser selecionadas de acordo com as necessidades da instituição e as condições específicas de instalação.

No entanto, nenhuma tecnologia isolada pode eliminar completamente os riscos associados à RM. A utilização dos FMDS deve estar integrada a programas de segurança abrangentes que incluam zoneamento, procedimentos de triagem, treinamento de pessoal, identificação da compatibilidade de objetos e dispositivos e controle de acesso

às áreas restritas. Considerando esses aspectos, sua adoção deve levar em conta as características da instituição, o fluxo de pacientes e equipamentos, os procedimentos de segurança existentes e as condições específicas de instalação [14].

Este estudo apresenta como limitação seu delineamento de análise técnico-narrativa, baseado em pesquisa bibliográfica não sistemática, pesquisa documental e análise reflexiva da experiência profissional do autor, não contemplando comparações experimentais próprias ou análises estatísticas de desempenho entre diferentes sistemas de detecção. Estudos futuros podem investigar comparativamente o desempenho, a efetividade operacional e o custo-benefício de diferentes configurações de FMDS, utilizando avaliações multicêntricas e dados operacionais de diferentes instituições.

Conflito de interesse

O primeiro autor possui vínculo profissional com a *MR IN SITE Healthcare Solutions*, empresa que fornece sistemas de detecção de materiais ferromagnéticos para ambientes de ressonância magnética.

Referências

[1] A. Shah e S. Aran, "A Review of Magnetic Resonance (MR) Safety: The Essentials to Patient Safety," *Cureus*, vol. 15, no. 10, e47345, 2023, doi: 10.7759/cureus.47345.

[2] National Oceanic and Atmospheric Administration (NOAA), "Earth's Magnetic Field Frequently Asked Questions." [Online]. Disponível: https://www.ncei.noaa.gov/products/geomagnetism/faq

[3] Ministério da Saúde, Cadastro Nacional de Estabelecimentos de Saúde (CNES), "Consulta de estabelecimentos." [Online]. Disponível: https://cnes.datasus.gov.br/

[4] *Colombini v. Westchester County Healthcare Corp., Supreme Court of the State of New York, Appellate Division, Second Department, Dec. 27, 2005. [Online]. Disponível: https://www.nycourts.gov/Reporter/3dseries/2005/2005_10095.htm.*

[5] AuntMinnie, "Settlement Details Released in $2.9M Colombini MRI Suit," 4 fev. 2010. [Online]. Disponível: https://www.auntminnie.com/clinical-news/mri/article/15594133/settlement-details-released-in-29m-colombini-mri-suit

[6] AuntMinnie, "Did the MRI Community Learn from the Colombini Tragedy?," 27 jul. 2005. [Online]. Disponível: https://www.auntminnie.com/clinical-news/mri/article/15573738/did-the-mri-community-learn-from-the-colombini-tragedy

[7] A. A. Mazzola, K. I. Stieven, G. H. Neto, and G. de M. Cardoso, "Segurança em Imagem por Ressonância Magnética," *Revista Brasileira de Física Médica*, vol. 13, no. 1, pp. 76–91, 2019. [Online]. Disponível: https://www.rbfm.org.br/rbfm/article/view/519

[8] M. N. Keene, "Using Ferromagnetic Detection Systems in the MRI Environment," in MRI Bioeffects, Safety, and Patient Management, F. G. Shellock and J. V. Crues, Eds. Los Angeles, CA, USA: Biomedical Research Publishing Group, 2014, pp. 107–145.

[9] K. Inaguma *et al.*, "Medical Accidents Related to Ferromagnetic Objects Brought into the MRI Room: Analysis of the National Multicenter Database by Orthopedic Surgeons," *Spine Surgery and Related Research*, vol. 7, no. 1, pp. 96–102, 2023, doi: 10.22603/ssrr.2022-0110.

[10] H. Murphy, "Lawyer Accidentally Shot by Own Gun in MRI Suite Dies of His Injuries," *Radiology Business*, 9 fev. 2023. [Online]. Disponível: https://radiologybusiness.com/topics/medical-imaging/magnetic-resonance-imaging-mri/lawyer-who-was-shot-his-own-gun-mri-suite

[11] B. Y. Lee, "Lawyer Dies After Shot by His Own Concealed Gun Triggered by MRI Machine," *Forbes*, Feb. 12, 2023. [Online]. Disponível: https://www.forbes.com/sites/brucelee/2023/02/12/lawyer-dies-after-shot-by-his-own-concealed-gun-triggered-by-mri-machine/.

[12] The Joint Commission, "Sentinel Event Alert: Preventing Accidents and Injuries in the MRI Suite," Oakbrook Terrace, IL, USA: The Joint Commission, 2008.

[13] G. P. Silva, M. M. Costa, S. S. Oliveira, L. L. Corrêa, and A. A. de Lima, "Vigilância Sanitária e segurança em ressonância magnética: levantamento dos serviços em Porto Alegre," *Vigilância Sanitária em Debate*, vol. 14, e02556, 2026, doi: 10.22239/2317-269X.02556.

[14] American College of Radiology, *ACR Manual on MR Safety*, 2026.

[15] Agência Nacional de Vigilância Sanitária (ANVISA), *Instrução Normativa nº 97, de 27 de maio de 2021*, 2021.

[16] Agência Nacional de Vigilância Sanitária (ANVISA), *RDC nº 611, de 9 de março de 2022*, 2022.

[17] Parlamento Europeu e Conselho da União Europeia, "Diretiva 2013/35/UE," 2013.

[18] Medicines and Healthcare products Regulatory Agency (MHRA), *Magnetic Resonance Imaging Equipment in Clinical Use: Safety Guidelines*, 5th ed., 2026.

[19] Okawa, R., N. Hayashi, Y. Kato, Y. Deguchi, S. Yahata, S. Kobayashi, and T. Takahashi, "Evaluation of Detectability in Clinical Practice of Metal Detectors, etc. for Magnetic Resonance Imaging," *Nihon Hoshasen Gijutsu Gakkai Zasshi*, vol. 79, no. 10, pp. 1168–1179, Oct. 2023, doi: 10.6009/jjrt.2023-1372.

[20] L. J. Orchard, "Implementation of a Ferromagnetic Detection System in a Clinical MRI Setting," *Radiography*, vol. 21, no. 3, pp. 248–253, 2015, doi: 10.1016/j.radi.2014.12.007.

[21] F. G. Shellock and A. Karacozoff, "Detection of Implants and Other Objects Using a Ferromagnetic Detection System: Implications for Patient Screening Before MRI," *American Journal of Roentgenology*, vol. 201, no. 4, pp. 720–725, 2013, doi: 10.2214/AJR.12.10332.

[22] MR IN SITE Healthcare Solutions, *Product Technical Datasheets and Internal Documentation*. São Bernardo do Campo, SP, Brazil, 2026.

[23] A. A. Mazzola, *Uma Breve Revisão sobre Supercondutividade e Hélio na Ressonância Magnética*, 1. ed. Porto Alegre, Brasil, 2019. [Online]. Disponível: https://pt.scribd.com/document/684285897/eBook-MRIONLINE-He-lio-e-Supercondutividade-na-RM

Technical Requirements for Heat-Sealing Equipment in Medical Device Processing Units According to Current Standards

1st Daniela C. C. Polli
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brasil
0009-0008-8086-6993

2nd Henrique T. Moriya
Biomedical Engineering Laboratory (LEB)
University of São Paulo (USP)
São Paulo, Brasil
0000-0002-1595-1779

***Abstract*—The sterile barrier system is a critical component in the medical device processing chain, responsible for ensuring the sterility of surgical instruments until their time of use. This study analyzes the consequences of the variability of operational parameters, such as specifically temperature, pressure, and dwell time/speed, within the heat-sealing process of medical grade paper and laminated film packaging, within the framework of ABNT NBR ISO standards and local perioperative guidelines (SOBECC). Through a structured review of the technical literature and materials engineering principles, we investigated how deviations in these three fundamental axes compromise seal integrity, focusing on the delamination phenomenon. The results indicate that underheating and insufficient thermal dwell time lead to incomplete polymer fusion and premature delamination. In contrast, overheating and overpressure trigger thermal degradation of the materials and polymer extrusion, causing cohesive tearing of the paper matrix and the subsequent release of cellulose microfibers during aseptic opening (peel-off). We conclude that mitigating heat-sealing failures and reducing operator-dependent variability are strongly associated with the integration of hardware and software technological solutions into modern heat sealers, in conjunction with automatic interlocking systems, based on the technical and normative literature reviewed, rather than on experimental or clinical data. Ultimately, this study is intended as a practical framework to support Clinical Engineering departments in conducting procurement, qualification, and routine analysis in compliance with current standards and guidelines, contributing to patient safety and to mitigation of healthcare-associated infections.**

***Index Terms*—Heat sealing, sterile barrier system, delamination, process validation, clinical engineering.**

I. Introduction

Central Sterile Supply Departments (CSSDs), or medical device processing units, are essential sectors within healthcare facilities responsible for processing medical devices, including critical stages such as cleaning, preparation, sterilization, storage, and distribution. Within this workflow, the packaging and sealing stage plays a determining role in maintaining the sterility of the device until the exact moment of clinical use.

The integrity of the Sterile Barrier System (SBS) is the primary factor in ensuring that a sterilized product remains free of microbiological contamination. This system is defined as the minimum packaging capable of preventing the ingress of microorganisms while simultaneously allowing aseptic presentation of the product at the point of use. In this context, sealing is not merely an operational step, but a critical process that directly impacts patient safety.

Brazilian regulation, through the normative resolution RDC No. 15/2012 [1], reinforces that the packaging architecture must allow the penetration of the sterilizing agent while concurrently preventing the entry of contaminants. Furthermore, the standard establishes that seal integrity is one of the key factors considered when defining the shelf life of the sterilized device, highlighting the relevance of this engineering process within the healthcare assistance context.

In the CSSD environment, the execution of the sealing process must strictly follow the guidelines provided by both the packaging material manufacturers and the equipment developers. To achieve this, Clinical Engineering departments must technically support CSSD management during equipment procurement, ensuring that selected automated heat-sealing technologies feature the mandatory metrological controls required to ensure process standardization, repeatability, and reliability. Accordingly, this study does not propose new universal design requirements; it offers an applied technical synthesis intended to support equipment specification, procurement, qualification, and routine monitoring by Clinical Engineering teams, grounded in currently available standards, professional guidelines, and technical literature.

II. Literature Review

A. Review Methodology

This study is based on a narrative, non-systematic review of Brazilian and international normative documents (ABNT NBR ISO 11607-1/2 [2], [3] , ABNT ISO/TS 16775 [4], ABNT NBR ISO 17665-1/2 [5], [6], ABNT NBR 14990 [6], ASTM F88 [7], EN 868-5 [8] the SOBECC perioperative guidelines [9], and complementary technical and manufacturer literature on heat-sealing equipment used in CSSD environments. Sources published up to 2023 and available in Portuguese or English were considered; documents were included when they directly addressed sterile barrier system sealing requirements,

validation, or process parameters, and excluded when addressing sterilization modalities or packaging materials outside this scope. Recommendations were extracted by cross-referencing normative requirements with the technical and engineering literature and are identified throughout the manuscript as originating from standards, professional guidelines, manufacturer specifications, or the authors' own engineering interpretation.

B. Normative and Technical References

The primary normative reference for medical device packaging sealing processes is the ABNT NBR ISO 11607 standard, particularly Part 2 [3], which specifies the validation requirements for forming, sealing, and assembly processes of packaging systems.

This standard establishes that every sealing process must be duly validated, ensuring that the sterile barrier system fulfills its essential functions: enabling sterilization, maintaining sterility, and allowing aseptic presentation. Consequently, it requires the definition, monitoring, and control of critical process parameters (CPPs), notably temperature, contact pressure, and dwell time (or sealing speed in rotary continuous sealers).

Furthermore, the standard emphasizes that seal quality must be demonstrated through measurable and verifiable characteristics. Key among these characteristics are adequate mechanical seal strength, seal continuity, the absence of structural defects (such as channels, pinholes, or tears), and the absence of material delamination, especially in systems designed for peel-open aseptic presentation.

The ABNT ISO/TS 16775 technical specification [4] provides practical guidance for applying these requirements, highlighting relevant operational aspects for healthcare institutions, such as correct packaging dimensioning, material compatibility, and the critical importance of controlling sealing parameters.

The ABNT NBR 14990-7 standard [6] describes the strength and integrity testing protocols for seals, ensuring the safety and reliability of the packaging system when subjected to the sterilization process.

The ABNT NBR ISO 17665 standard [5] focuses strictly on the requirements for the development, validation, and routine control of moist heat (steam) sterilization processes for medical devices. Although it does not dictate isolated rules for heat sealers, it determines that the SBS, which encompasses the sealing of packages within CSSD environments, must be fully integrated and validated to guarantee that steam penetrates the material and maintains its sterile state until point of use.

Finally, the ASTM F88 standard [7] defines the standard test methods for seal strength testing via tensile and peel testing to measure the mechanical force of flexible barrier material seals, such as medical grade paper, within the CSSD environment. It establishes testing techniques and acceptance criteria to ensure that the seal is neither too weak, which would compromise sterility, nor excessively strong, which would cause material tearing and subsequent particulate contamination during opening.

III. Technical Aspects of Heat Sealing

Heat-sealing equipment, particularly the thermal sealers deployed in CSSDs, must exhibit technical characteristics that guarantee strict process control. A fundamental requirement is the ability to continuously monitor and control the Critical Process Parameters (CPPs), namely temperature, pressure, and dwell time.

Furthermore, these devices must be equipped with alarm systems or hardware/software safety mechanisms designed to interrupt the cycle or alert the operator in the event of any deviation from the established operating parameters. This functionality is essential to prevent the production of inadequate seals that could compromise the integrity of the sterile barrier system.

The Brazilian regulatory standard RDC No. 15/2012 [1] establishes periodic calibration of heat sealers with a minimum annual frequency, Under a clinical engineering asset management framework, this calibration must not be treated as a mere administrative milestone, but as a continuous metrological safety program where Clinical Engineering monitors sensor drift tolerances to safeguard the reliability of measurement and the long-term performance of the equipment.

Another relevant aspect is the requirement for comprehensive equipment documentation, including operation manuals, preventive maintenance logs, calibration certificates, and standard operating procedures (SOPs). In addition, operators must undergo structured training and qualification, as the human factor is an integral component of the overall process validation.

Seal quality evaluation must be performed using well-defined criteria to ensure the integrity and functionality of the sterile barrier system. Sealing tracks must present a uniform aesthetic aspect, featuring adequate width and a smooth, continuous surface free of wrinkles, bubbles, fractures, or delamination.

Seal strength is a fundamental parameter; it must be sufficient to maintain integrity of the packaging during transportation, storage, and handling, without hindering the subsequent aseptic presentation. To evaluate this mechanical property, specific destructive protocols are utilized, such as the tensile seal strength test described in ASTM F88 [7].

Additionally, two integrity testing methods are widely employed to detect sealing failure pathways, such as channels or micro-leaks: the dye penetration test and the bubble leak test.

The Dye Penetration Test (Ink Test) is specifically engineered to detect channels and discontinuities along the sealing line caused by insufficient pressure or underheating. It relies on capillary action; if a micro-channel is present, the liquid solution will migrate through it. The standards ABNT NBR ISO 11607-2 [3] and ABNT ISO/TS 16775 [4] recommend the dye test as physical and documentable evidence during Operational Qualification (OQ) and Performance Qualification (PQ), as well as routine daily monitoring in CSSD workflows, in accordance with local perioperative guidelines (SOBECC) [9].

The Bubble Leak Test is an integrity assay based on internal pressurization. While the dye test targets chemical capillarity, the bubble test utilizes differential air pressure to indicate pinholes, tears, or macroscopic seal failures. Following SOBECC guidelines [9], this method serves as an alternative for evaluating the global resistance of the sterile barrier against macroscopic defects.

Visual inspection also plays an important role as a tool for continuous monitoring, although it is insufficient as a standalone evaluation method.

The validation of these testing methodologies is equally mandated by current standards, which state that the assays must demonstrate adequate repeatability, reproducibility, and sensitivity for defect detection.

Finally, packaging sealing is directly correlated with the efficacy of the sterilization process. The ABNT NBR ISO 17665-1 standard [5] establishes that packaging systems must withstand processing conditions, including acute variations in temperature, pressure, and humidity, without compromising their physical integrity. Ultimately, the standard demands that the packaging system's integrity be verified during sterilization process qualification and throughout routine activities. This includes physical inspection of processed loads and verification that the sterile barrier system remains structurally intact after the cycle.

IV. Primary Sealing Failures

The literature on sterile barrier systems highlights that seal strength depends directly on proper stabilization and combination of three critical variables: temperature, contact pressure, and dwell time. Figures 1–6 illustrate representative seal defects associated with the corresponding process deviation for didactic purposes; they do not constitute a controlled experimental dataset with fully disclosed packaging materials or process parameters, and are used here to support qualitative, not quantitative, discussion of each failure mode.

According to the guidelines of the Brazilian Association of Operating Room Nurses, Anesthesia Recovery and Central Sterile Supply Center (SOBECC) [9] and the requirements of the ABNT NBR ISO 11607-2 standard [3], deviations in sealing temperature due to underheating or overheating compromise the integrity of the sterile barrier system, resulting in incomplete fusion or material embrittlement.

When the sealer temperature falls below the lower threshold recommended by the manufacturer and technical standards (underheating - Fig.1), the thermal energy is insufficient to properly melt the plastic layer (polymer) and bond to medical grade paper.

- **Weak or Partial Sealing**: Incomplete fusion results in inadequate adhesion. Visually, the sealing track may appear continuous, but it lacks the necessary mechanical peeling strength.
- **Premature Opening (Delamination)**: Packaging may open spontaneously during handling, transport, or inside the autoclave itself due to pressure fluctuations during the sterilization cycle.

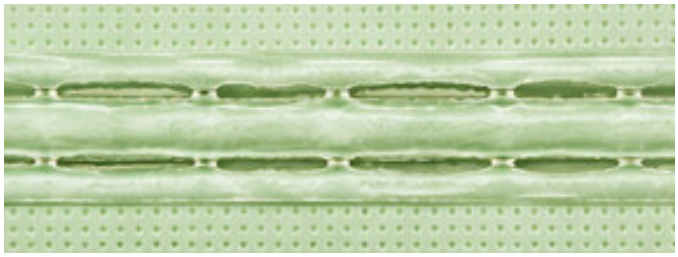


Fig. 1. Low Temperature Seal



Fig. 2. High Temperature Seal

- **Loss of Sterile Barrier**: Any micro-gap resulting from poor polymer fusion negates the packaging's barrier properties, allowing the ingress of microorganisms and contaminating the device.

In contrast, a temperature exceeding the upper threshold (overheating - Fig. 2) burns or degrades the packaging materials.

- **Plastic Crystallization and Embrittlement**: Excessive heat alters the molecular structure of the plastic film, making it rigid and brittle.
- **Micro-perforations and Scorching**: Medical grade paper or the film may develop micro-holes or tears along the sealing line that are imperceptible to the naked eye.
- **Incorrect Opening (Tearing/Fiber Shedding)**: When the package is opened within the sterile field, the thermally degraded paper will tear or peel irregularly (erratic delamination), releasing cellulose fibers and particulate matter that can fall into the surgical site or contaminate the instruments.
- **"Permanent Bonding" Effect**: The film undergoes excessive fusion, hindering proper aseptic peel-open presentation. This forces healthcare personnel to use scissors, increasing the risk of breakthrough contamination.

Furthermore, in accordance with the ABNT NBR ISO 11607-2 standard [3], the sealing pressure is another critical process parameter that directly affects the uniformity of the fusion. Failure to control this compressive force, whether through underpressure or overpressure, compromises the structural integrity of the sterile barrier system, as established by SOBECC safety guidelines [9].

When the compressive force exerted by the heat sealer is less than the validated parameters (underpressure), the two packaging layers are not pressed together with the necessary

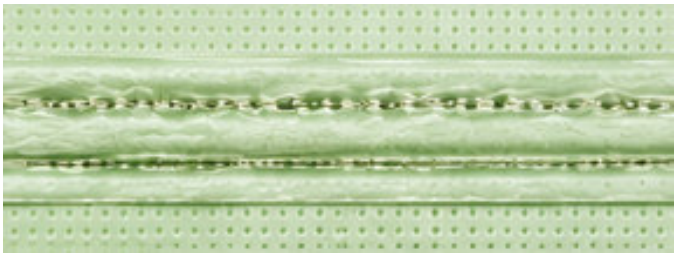


Fig. 3. Low pressure seal

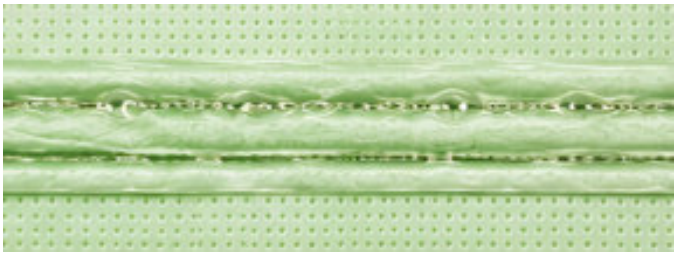


Fig. 4. High pressure seal

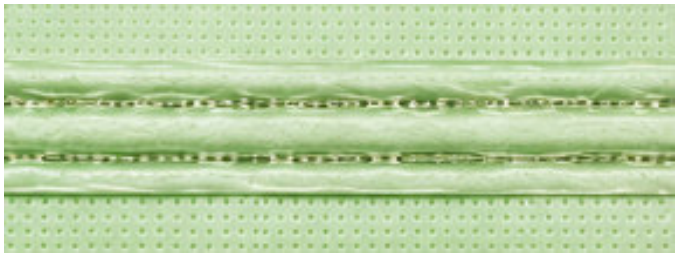


Fig. 5. High speed seal

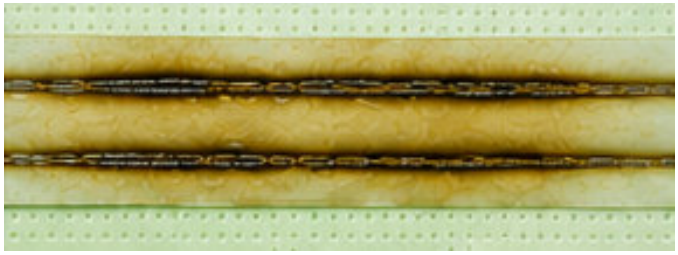


Fig. 6. Low speed seal

intensity. Fig. 3 illustrates the outcome of a low sealing pressure.

- **Incomplete or Intermittent Sealing**: Even if the temperature is correct, insufficient pressure prevents the molten plastic from properly penetrating the microfibers of the medical grade paper. This generates falsely sealed areas known as "blind spots" or "air islands."
- **Micro-leakage Channels (Canaliculus)**: Underpressure creates micro-channels along the sealing track. These channels, invisible to the naked eye, act as pathways for microbial ingress, breaking the sterile barrier before clinical use.
- **Low Mechanical Strength**: The package will exhibit low seal strength, rupturing easily under the weight of the surgical instrumentation or due to physical pressure differences inside the autoclave during the vacuum phase.

In contrast, excessive mechanical force crushes the packaging materials beyond the structural tolerance of the components. This effect is shown in Fig. 4.

- **Polymer extrusion ("Squeezed" Plastic)**: Excessive pressure pushes the molten plastic out of the original sealing line. This leaves the central zone of the seal extremely thin, weakened, and susceptible to bursting.
- **Paper Piercing and Shearing**: The excessive force of the sealing bars can cut, crush, or weaken the fibers of the medical grade paper at the edge of the seal, creating points of immediate tearing or micro-perforations.
- **Incorrect Polymer Transfer Upon Opening**: When the healthcare professional opens the pouch utilizing the aseptic peel-open technique, the effects of overpressure cause the paper matrix to tear rather than detach smoothly from the plastic film. This removes particulate matter and destabilizes the safety of the aseptic presentation.

The thermal dwell time, or the packaging traction speed through the sealer, is the critical process parameter that governs the efficiency of heat transfer. If the substrate travels too quickly or too slowly across the sealing bars, the thermal-mechanical equilibrium of the process is disrupted, directly compromising the security of the sterile device.

When the heat sealer operates at a linear velocity higher than the validated baseline (sub-exposure), the pouch travels too quickly across the heated elements, reducing the physical contact time, as illustrated in Fig. 5.

- **Insufficient Thermal Input (Under-dwelling)**: Even if the bar temperature is accurately regulated, the polymer does not remain in contact with the heat source long enough to reach its melting point.
- **"False" or Superficial Sealing**: The plastic film undergoes merely a visual, aesthetic alteration, appearing bonded without achieving an intimate, intermolecular fusion within the porous matrix of the medical grade paper. Under minimal physical stress (such as the weight of a surgical instrument tray), the layers immediately separate.
- **Discontinuity Genesis**: Excessive linear velocity prevents the stabilization of pressure and heat distribution along the sealing track, generating intermittent bonding points interspersed with completely unsealed pathway.

In contrast, when the packaging advances too slowly or remains stationary between the heating bars beyond the specified threshold, it induces an over-exposed state that causes severe structural degradation, as depicted in Fig. 6.

- **Total Thermal-Mechanical Degradation**: Excessive exposure to heat melts the plastic and dries out the cellulose fibers of medical grade paper.
- **Melting-induced Punctures**: The plastic film can liquefy to the point of shrinking, piercing, or rupturing entirely

within the sealing track, rendering the packaging immediately defective.
- **Adhesion to Sealing Bars**: Over-fused plastic may adhere to the internal components of the machine (such as Teflon belts or heating bars), jamming the equipment, tearing the pouch, and contaminating the sealing path with burned residues.

V. Root Cause Analysis of Delamination

One of the most prevalent failure modes within the medical sterilization cycle is package delamination. In heat-sealing processes, delamination is defined as the undesirable separation of the constituent layers of a multilayer material during or after thermal sealing, resulting in a total loss of the laminate's structural cohesion. Within sterile barrier systems for medical devices, this is classified as a critical failure. It severely compromises sterile barrier integrity, hinders aseptic presentation at the point of use, and breaches the strict compliance criteria of the ABNT NBR ISO 11607 standard [2], [3] for terminally sterilized medical device packaging. Although sealing-parameter deviations are a major and well-documented trigger, delamination may also result from interactions among substrate composition, adhesive properties, material aging, environmental humidity, sterilization conditions, storage, and packaging manufacturing quality. This study focuses specifically on the sealing-process contribution to delamination, without excluding these other contributing factors.

In a multilayer packaging structure (e.g., PET film + adhesive + PE, or Tyvek® + plastic film), distinct bonding interfaces exist. When the package is opened correctly, material separation must occur exclusively at the interface specifically engineered for peeling. Conversely, when delamination occurs, the applied mechanical force overcomes the internal adhesion between the laminate layers, producing partial peeling, loose fibers, material residues, irregular tearing, and potential particulate generation. In sterile medical devices, this phenomenon is particularly critical as it invalidates aseptic presentation.

A. Failures Induced by Underheating or Sub-exposure (Incomplete Bonding)

In this scenario, delamination occurs prematurely and spontaneously prior to clinical use (during autoclaving, transport, or storage) because the interface materials never achieved true intermolecular fusion.

- **Temperature Below Specification**: Without the required thermal energy input, the plastic polymer does not reach its softening or melting point, merely contacting the paper matrix without penetrating its microfibers. Consequently, any physical vibration or internal pressure differential inside the autoclave causes the plastic film to completely detach from the paper (fusionless delamination).
- **Insufficient Dwell Time / High Sealing Speed**: Even if the temperature is accurate, if the packaging travels too quickly across the heating bars, thermal transfer cannot be successfully completed. This yields a superficial seal that undergoes immediate delamination under minimal mechanical stress.

B. Failures Induced by Overheating (Thermal Degradation)

Delamination caused by overheating manifests during aseptic opening within the surgical field, occurring in an irregular, defective manner that introduces biological risks. Excess heat degrades the chemical bonding properties of the plastic film and severely desiccates the cellulose fibers of medical grade paper. Upon attempting to open the package using the proper peel-open technique, instead of a smooth film detachment, the layers fracture into splinters or scale off. The paper matrix splits internally (delaminates), shedding cellulose microfibers that contaminate the surgical field.

C. Failures Induced by Overpressure (Mechanical Crushing)

Under excessive mechanical compression, delamination also manifests itself erratically during opening within the operating room. Excessive force crushes and extrudes the molten polymer from the original sealing line. The remaining plastic within the joint becomes extremely thin and brittle. During aseptic opening, this weakened track does not withstand tensile stress, scaling irregularly, and tearing the paper fibers instead of achieving a flawless, clean peel.

VI. Qualification Process

According to the mandates of the ABNT ISO/TS 16775 technical specification [4], which provides guidance on the application of the ABNT NBR ISO 11607-1 [2] and ABNT NBR ISO 11607-2 [3] standards, the heat-sealing equipment must undergo Installation Qualification (IQ), Operational Qualification (OQ) and Performance Qualification (PQ). Unless explicitly attributed to a cited standard, the specific numerical values, examples, and workflows described in this section (e.g., temperature boundaries, seal-strength reference values, and setpoint-definition criteria) illustrate common clinical-engineering and industry practice rather than requirements mandated verbatim by the referenced documents. These sequential validation stages, co-executed and formally approved by Clinical Engineering and CSSD specialists, stages enable the characterization and definition of the safe operating window of the process.

During Installation Qualification (IQ), it is recommended to employ a structured checklist to evaluate physical installation status, verifying previous calibration procedures backed by official certificates, as well as the implementation of a continuous calibration program. This ensures that the sealing parameters remain accurate throughout the use of the equipment. Furthermore, environmental conditions (such as cleanliness, ambient temperature and relative humidity) must be observed, along with documented operator training and the availability of operating manuals or standard operating procedures (SOPs).

Critical Process Parameters (CPPs) must be clearly established, encompassing at least contact pressure, temperature, and sealing dwell time or speed. For rotary continuous sealers, the dwell time is expressed as the sealing speed (e.g. meters

per minute). For bar sealers, the dwell time represents the exact duration for which the heated bars remain in direct contact with the packaging substrates. The sealing temperature must comply with the sterile barrier system manufacturer's technical specifications. Additionally, the heat sealer must feature integrated control and monitoring hardware for these CPPs, incorporating alarm/warning systems or automatic cycle interruption mechanisms if parameters drift beyond acceptable limits.

During Operational Qualification (OQ), usability aspects are assessed, including the identification of the target sealing temperature, which varies depending on the polymeric formulation of the sterile barrier system. The upper and lower boundaries for temperature, pressure, and dwell time must be rigorously mapped. The equipment manufacturer must demonstrate the monitoring mechanism for these variables. The sealer must be technically capable of operating within the boundaries recommended by the preformed sterile barrier system manufacturer. Higher flexibility in boundary configuration translates to enhanced equipment performance across the wide range of commercially available sterile barrier systems.

Testing during OQ must be performed at the established lower and upper temperature boundaries after the equipment calibration has been finalized. The primary acceptance criteria for process approval include seal integrity, which must remain intact without gaps, channels, punctures, wrinkles, folds, or material delamination.

For seal and temperature characterization, a minimum mechanical seal strength threshold is required. Following European reference standards (EN 868-5 [8]), which are widely adopted in the absence of local regulatory values, a minimum reference value of 1.5 N per 15 mm is established for moist heat sterilization processes and 1.2 N per 15 mm for alternative sterilization methods.

Commercially available dye penetration test kits or other seal integrity indicators can also be deployed, provided that all test results are formally documented.

To complement the validation matrix, the Seal Test is deployed as a critical indicator for routine monitoring and quality control across both Operational Qualification (OQ) and Performance Qualification (PQ) stages, in accordance with the ABNT ISO/TS 16775 technical specification [4]. During the OQ phase, the chemical indicator sheets are processed at the established upper and lower temperature boundaries to physically challenge and verify seal continuity, ensuring the absence of unsealed channels or structural gaps even under extreme operational limits. In contrast, within the PQ workflow and routine monitoring, the Seal Test is recommended by SOBECC [9] and widely adopted in clinical practice as a pre-production verification tool under nominal parameters, providing documented, reproducible evidence that the heat sealer successfully transfers uniform thermal-mechanical energy to achieve complete polymer fusion across the porous matrix, in compliance with the visual evaluation criteria specified in the EN 868-5 standard [8].

Following the operational boundary evaluation, if quality requirements are successfully satisfied at both the upper and lower boundaries, the process setpoint is commonly defined, as an engineering convention rather than a universal normative rule, as the arithmetic mean of these two values (e.g., a lower limit of $180°C$ and an upper limit of $190°C$ produce a baseline sealing temperature of $185°C$); institutions may adopt other criteria within the validated range.

Finally, Performance Qualification (PQ) evaluates whether the validated process consistently produces compliant sterile barrier systems under actual routine operational conditions.

VII. Engineering Design Features for Preventing Delamination Failures

A. Sensing and Mechanical Hardware

Modern heat-sealing instrumentation incorporates dedicated hardware mechanisms intended to support metrological traceability and precision material handling, thereby reducing operator dependency.

- **High-Response Temperature Sensors**: Some advanced equipment embeds platinum resistance thermometers (RTDs, such as Pt100/Pt1000) directly into the heating block as an alternative to standard thermocouples. Depending on system architecture, calibration, and dynamic-response requirements, RTDs can provide high-resolution, near-continuous temperature readings that help mitigate thermal inertia; sensor choice should nonetheless be based on the specific metrological requirements of each equipment design rather than treated as a universally superior solution.
- **Independent Upper and Lower Bar Temperature Control**: Although European reference standards (EN 868-5 [8]), frequently adopted in the absence of national specifications, state that the temperature baseline can be determined by the arithmetic mean of both elements, acute thermal asymmetries between the two bars compromise sterile barrier integrity. For instance, while an average target specification of $180°C$ can be achieved through symmetric distributions ($175°C/185°C$), asymmetric profiles (e.g., $150°C$ upper and $230°C$ lower) induce uneven polymer melting and subsequent delamination.
- **Adjustable Infeed Guides and Automated Traction**: The operator simply positions the packaging edge against the input guide, after which synchronized drive belts pull the substrate forward at a locked linear speed. This design feature is intended to reduce human variability regarding the thermal exposure time.

B. Embedded Control Software

Embedded software serves as an active control instrument, enforcing compliance with critical parameters validated during qualification protocols, as mandated by the ABNT NBR ISO 11607-2 standard [3].

- **Process Window Monitoring (OQ Limits)**: During the Operational Qualification (OQ) phase, a safe operating window is assigned (e.g., $180°C$ to $185°C$). The firmware

automatically restricts machine operation if the system parameters drift outside this validated range.

- **Alarm Systems and Mechanical Interlocking (Machine Stop)**: This functionality reflects the process-monitoring and deviation-response mechanisms required by ABNT NBR ISO 11607-2 [3] for a validated sealing process, and is commonly implemented as a strict operational safeguard in commercially available equipment. If the temperature drops below a validated tolerance around the setpoint (a 2°C deviation is a common illustrative threshold in commercial equipment, though the exact value must be defined for each sterile barrier system during OQ) or if the drive velocity fluctuates, the software triggers a visual/audible alert and immediately de-energizes the traction motor, physically preventing the processing of non-compliant packaging.
- **Advanced Temperature Regulation**: Microprocessor-driven control loops (such as Proportional-Integral-Derivative - PID algorithms) mathematically calculate instantaneous heat dissipation as a pouch enters the track, dynamically modulating electrical power to the heating elements to maintain a linear thermal baseline.

C. Digital Traceability and Data Logging

Digital traceability and data logging solutions provide verifiable assurance that each sealing cycle adheres to the validated standard.

- **Supervisory Control and Data Acquisition (SCADA) Integration**: Equipment architecture may include communication ports (Ethernet, USB, or RS-232) to export real-time temperature, pressure, and dwell time data logs for every processed pouch. Should a specific batch exhibit post-processing delamination, Clinical Engineering acts as a root-cause investigator, auditing these computerized SCADA logs to correlate thermodynamic deviations with structural laminate failures, preventing speculative adjustments and ensuring data-driven troubleshooting at the precise second of execution.
- **Integrated Process Inking**: Another technological alternative involves coupling an inline ink jet printer to the output terminal, printing data such as processing date, lot number, operator ID, and a ”valid parameters” verification string directly onto the substrate. If the equipment drifts from its validated baseline, printing is instantly suspended, which yields immediate visual evidence of the error on the package itself.

D. Troubleshooting Guidance

Equipment developers provide detailed troubleshooting guidelines in their technical manuals to diagnose and mitigate active delamination failures:

- **Sealability Curve Optimization**: Technical documentation provides cross-referenced matrices correlating substrate formulation with ideal temperature and linear velocity combinations.
- **Polytetrafluoroethylene (PTFE/Teflon) Tape Replacement**: Manuals mandate the routine replacement of protective PTFE barriers. If the Teflon layer degrades, the packaging film adheres to the exposed heating bar, causing catastrophic delamination through scorching and material drawing.
- **Compulsory Calibration Cycles**: Technical instructions explicitly dictate that internal metrological sensors undergo calibration by laboratories accredited under national metrological networks (such as the RBC - Rede Brasileira de Calibração) to verify that the digital readout accurately reflects the true thermal energy transferred to the polymer matrix.

VIII. Conclusion

The safety and efficacy of medical device processing depend heavily on maintaining instrument sterility until the exact moment of clinical deployment within the surgical field. Within this framework, the heat-sealing process transcends being a mere mechanical closure step, assuming the role of a critical biomedical engineering procedure whose predictability and repeatability are guaranteed by rigorous control of critical process parameters: temperature, contact pressure, and thermal dwell time.

Operational deviations in any of these three core variables compromise the validated process window established under the guidelines of the ABNT NBR ISO 11607-2 standard [3], ultimately triggering structural delamination failures. Consequently, continuous validation and automated monitoring of sealing parameters are not simply bureaucratic or regulatory exercises mandated by ISO standards or local perioperative guidelines (SOBECC) [9]. Rather, they constitute a fundamental risk mitigation and quality assurance strategy driven by the joint efforts of CSSD nursing and Clinical Engineering.

A. Study Limitations and Research Gaps

As the recommendations are based predominantly on a literature and normative review rather than on experimental testing or case analysis in real hospital settings, claims regarding elimination of human error, direct reduction of infection rates, and superiority of specific equipment architectures should be regarded as hypotheses for future validation rather than established outcomes. This study’s main contribution is the integrated analysis of normative requirements, clinical-engineering practice, and delamination mechanisms into a single applied technical framework, complementing existing standards and technical guides that typically address these aspects separately. Adopting these recommendations in practice will also depend on each institution’s infrastructure, budget, staff training, and access to compatible equipment and consumables, which may limit or delay implementation in some CSSD settings. In summary, this review indicates that explicit control and documentation of temperature, pressure, and dwell time are consistently required across the standards analyzed. Delamination emerges as the most frequently reported critical failure, although it is rarely attributed to a

single process variable. In addition, automation and digital traceability are recurring engineering approaches aimed at reducing operator-dependent variability. The main identified research gaps are the scarcity of published experimental data directly linking specific sealing-parameter deviations to seal-strength or delamination outcomes, and the absence of studies correlating sterile barrier system performance with clinical infection indicators.

This collaborative framework connects the metrological precision of clinical instrumentation with sterile barrier integrity, which is recognized in the literature as a contributing factor to surgical site infection prevention, positioning the sterile barrier system as an important element in safeguarding patient safety and clinical outcomes. Establishing a quantitative link to Surgical Site Infection (SSI) rate reduction would require dedicated clinical and experimental studies beyond the scope of this technical review.

IX. Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (309464/2022-6 to HTM). The authors thank Persei Sistemas Eletroeletrônicos for providing the images used in this study and for their valuable contributions that enabled a better understanding of the primary challenges in the thermal sealing process.

References

[1] Agência Nacional de Vigilância Sanitária (ANVISA), "Resolução RDC nº 15, de 15 de março de 2012. Dispõe sobre requisitos de boas práticas para o processamento de produtos para saúde e dá outras providências," *Diário Oficial da União, Brasília, DF*, Mar 2012.

[2] Associação Brasileira de Normas Técnicas (ABNT), *ABNT NBR ISO 11607-1:2013: Embalagem para produtos para saúde esterilizados terminalmente — Parte 1: Requisitos para materiais, sistemas de barreira estéril e sistemas de embalagem*. Rio de Janeiro, RJ: ABNT, 2013.

[3] Associação Brasileira de Normas Técnicas(ABNT), *ABNT NBR ISO 11607-2:2024: Embalagem para produtos para saúde esterilizados terminalmente — Parte 2: Requisitos de validação para processos de formação, selagem e montagem*. Rio de Janeiro, RJ: ABNT, 2024.

[4] Associação Brasileira de Normas Técnicas (ABNT), *ABNT ISO/TS 16775:2018: Embalagem para produtos para a saúde esterilizados finalizados — Orientações sobre a aplicação da ABNT NBR ISO 11607-1 e da ABNT NBR ISO 11607-2*. Rio de Janeiro, RJ: ABNT, 2018.

[5] Associação Brasileira de Normas Técnicas(ABNT), *ABNT NBR ISO 17665-1:2010: Requisitos para o desenvolvimento, validação e controle de rotina nos processos de esterilização de produtos para saúde*. Rio de Janeiro, RJ: ABNT, 2010.

[6] Associação Brasileira de Normas Técnicas (ABNT), *ABNT NBR 14990-7:2004: Systems and materials of packaging for sterilization of healthcare products - Part 7: Envelope and tubular*. Rio de Janeiro, RJ: ABNT, 2004.

[7] ASTM International, *ASTM F88/F88M-23: Standard Test Method for Seal Strength of Flexible Barrier Materials*. West Conshohocken, PA: ASTM, 2023.

[8] European Committee for Standardization (CEN), *EN 868-5:2018: Packaging materials and systems for medical devices which are to be sterilized - Part 5: Sealable pouches and reels of porous materials and plastic film*. Brussels, Belgium: CEN, 2018.

[9] Associação Brasileira de Enfermeiros de Centro Cirúrgico, Recuperação Anestésica e Centro de Material e Esterilização (SOBECC), *Diretrizes de Práticas em Enfermagem Perioperatória e Processamento de Produtos para Saúde*, 8th ed. São Paulo, SP: SOBECC, 2021.

A Study of Statistical Noise in Cleft Lip and Palate Classification Models

André Jacob
Biomedical Engineering
Federal University of ABC
São Paulo, Brazil
andre.jacob@aluno.ufabc.edu.br
ORCID: 0009-0003-2611-0697

Martha R. Ortiz-Posadas
Electrical Engineering Depatment
Universidad Autónoma Metropolitana Iztapalapa
Mexico City, Mexico
posa@xanum.uam.mx
ORCID: 0000-0001-5032-2663

***Abstract*— Noise may be an important problem in medical decision making which includes establishing diagnoses, assessing conditions, or deciding upon management options among others. When medical professionals make different decisions despite being faced with the same information this can be termed noise, that is, unwanted, random variability in decision making. The focus of this work is on the analysis of statistical noise during the process of classifying the surgical complexity of cleft lip and palate, which are facial congenital malformations in fetal development. Their anatomical description involves, in general terms, the lip, the alveolus, and the palate (hard and/or soft), and may occur in isolated or combined form, unilateral or bilateral. In this work, two models that propose a classification for cleft lip and palate developed at the Comprehensive Center for Cleft Lip and Palate of Mexico (SUMA Center) were analyzed. Model 1 proposes a classification of 29 classes based on the location of the cleft, and Model 2, proposes a classification of 15 classes based on the anatomical characteristics of the cleft. The objective of this work was to evaluate the statistical noise of both classification models in a sample of 105 patients treated at the SUMA Center using ordinal coherence, increase and Shannon entropy between classes of both models. To complement the analysis, we use Spearman's rank correlation coefficient and, Kendall's tau coefficient to quantify the degree to which the ordering of classes is aligned with increasing surgical complexity. The results indicated that Model 2 exhibits lower structural statistical noise than Model 1.**

Keywords—Statistical noise, medical decision making, cleft lip and palate classification.

I. Introduction

Unwanted random variability of judgment, commonly referred to as "noise," is an inherent phenomenon in human decision-making and has been documented across a wide range of disciplines, including medicine. Unlike systematic bias, which consistently skews decisions in a particular direction, noise introduces inconsistency and unpredictability, leading to variability in outcomes even when the same information is available. One important subtype is occasion noise, which arises from transient factors associated with the moment in which a decision is made. These factors include physiological and psychological states such as stress, fatigue, hunger, cognitive load, time pressure, and overall workload. Such influences can subtly alter perception, reasoning, and judgment, thereby affecting decision outcomes [1].

In the context of medical practice, noise represents a particularly critical yet underexplored issue. Clinical decision-making involves complex cognitive processes, including diagnosis, evaluation of disease severity, and selection of treatment strategies. Variability in these decisions can have direct consequences for patient outcomes, safety, and quality of care. Despite its potential impact, the systematic measurement and characterization of noise in medicine remain limited. In clinical settings, noise may manifest when different healthcare professionals arrive at different conclusions or recommendations despite being presented with identical clinical data. This inconsistency reflects not only differences in expertise or experience but also the influence of contextual and situational factors that introduce random variability into judgment [2].

The present work focuses specifically on the analysis of statistical noise in the classification of surgical complexity in patients with cleft lip and palate. These conditions are facial congenital malformations that occur during fetal development due to the incomplete fusion of facial structures. They may affect the lip, alveolus, palate (hard and/or soft), and nasal structures, either in isolation or in combination. Their etiology is multifactorial, involving genetic predisposition as well as environmental influences such as maternal health, nutrition, and exposure to teratogens. Clinically, cleft lip and palate present with a wide spectrum of anatomical variations, ranging from mild unilateral defects to more complex bilateral and combined forms (Fig. 1).

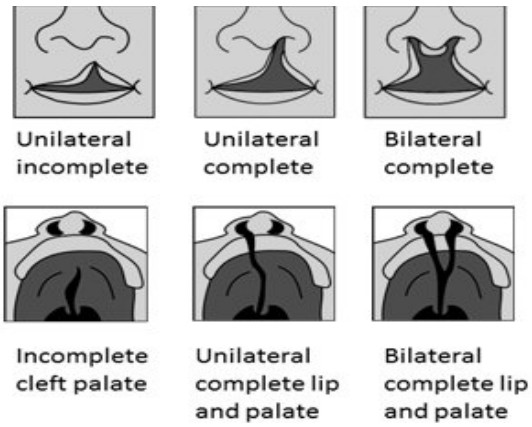


Figure 1. Cleft lip and clef palate (unilateral and bilateral) [3]

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The anatomical characterization of these malformations typically considers three primary components: the lip, the alveolar ridge, and the palate. Each of these structures can be affected to varying degrees, resulting in significant heterogeneity in clinical presentation. This variability poses challenges for the classification of surgical complexity, as it requires clinicians to interpret anatomical features and assign them to predefined categories. Consequently, the process is susceptible to noise, particularly when classification criteria are subjective or when evaluators differ in training and experience. Understanding and quantifying this variability is essential for improving consistency in clinical assessment, optimizing treatment planning, and ultimately enhancing patient outcomes [4].

While statistical noise is traditionally associated with variability among observers or repeated decisions under identical conditions, it can also be examined from the perspective of the internal consistency of a classification system. In this context, noise may emerge when the structural organization of categories fails to preserve a coherent relationship with the variable that the classification intends to represent. Therefore, the present study focuses on the concept of structural statistical noise, defined as irregularities, discontinuities, or inconsistencies within the progression of surgical complexity levels across the classes of a classification model. This approach does not seek to evaluate inter-observer agreement, but rather the internal coherence of the classification framework itself as a decision-support instrument.

In this work, two models that propose a classification for cleft lip and palate developed at the Comprehensive Center for Cleft Lip and Palate of Mexico (SUMA Center) [5] were analyzed. Both are described as follows:

Model 1. Proposes a classification of 29 classes based on the location of the cleft in the primary palate (unilateral and bilateral), and/or in the secondary palate. It also considers clefts that occur simultaneously in both palates (primary and secondary) and groups them into cleft lip and palate with unilateral or bilateral involvement of the primary palate [6].

Model 2. Proposes a classification of 15 classes based on the anatomical characteristics of the cleft. It considers complete and incomplete clefts of the primary and secondary palate. It includes complete or incomplete bilateral clefts of both palates, and asymmetric bilateral clefts of the primary palate (a complete unilateral cleft and an incomplete unilateral cleft) [7].

By combining these classes in each model, the SUMA Center plastic surgeon team determined the surgical complexity of the clefts. Then, the corresponding model assigned a *Surgical Complexity Factor* (SCF) within a value range. Having a classification with 29/15 classes of clefts, with their respective surgical complexity factor [5, 6], means the unification of the surgeon's clinical criteria and, therefore, makes it possible to define a specific surgical management of the patient, depending on the class to which the cleft that it presents belongs to.

Therefore, the objective of this work was to evaluate the statistical noise of both cleft lip and palate classification models described above, analyzing the SCF in a sample of 105 patients treated at the SUMA Center using three approaches: ordinal coherence, increase and Shannon entropy between classes. To complement the analysis, the association between class order and mean surgical complexity values was assessed by two coefficients, Spearman's rank correlation coefficient and Kendall's tau coefficient. Both measures quantify the degree to which the ordering of classes is aligned with increasing surgical complexity

II. Methodology

A. Ordinal Coherence

Ordinal data is a qualitative data type that classifies variables in a ranked order. This kind of data is categorical and separates variables into groups based on distinct ascending or descending hierarchies. An ordinal scale is defined as a measurement scale that allows for the ranking of data based on relative comparisons but does not permit absolute comparisons or mathematical operations [8]. Agreement or coherence in ordinal data refers to the degree to which different observers or classification methods assign the same ordered category to the same observation [9].

In this work, to assess ordinal coherence, the association between the designated categories and their respective mean SCF values was examined. This procedure was grounded in the assumption that a surgical complexity-based classification framework should exhibit a monotonic pattern, where higher numerical categories correspond to progressively greater mean SCF values compared to preceding ones. Therefore, ordinal consistency reflects the extent to which the classification scheme aligns with the principle of ascending surgical complexity.

The study included 105 patients diagnosed with cleft lip and/or palate and treated at the SUMA Center [5]. For each patient, the corresponding cleft classification was assigned according to both models, and the associated Surgical Complexity Factor (SCF) was obtained. Subsequently, mean SCF values were calculated for each class by averaging the SCF values of all patients belonging to that category. These class-level mean values served as the basis for the ordinal coherence and increment analyses. Subsequently, the difference (Δ) between the mean SCF of each class and that of the preceding class was examined. Positive Δ values reflect a consistent increase in complexity, while negative Δ values signal reversals. That is, cases in which a higher-numbered class exhibits a lower mean SCF than the one before it. Each reversal was documented as a breach of ordinal consistency. This strategy enables the systematic detection of hierarchical discontinuities within the model, highlighting potential mismatches between the intended classificatory framework and the observed numerical patterns of the SCF variable.

To complement the inversion analysis, the association between class order and mean SCF values was assessed by two coefficients: Spearman's rank correlation coefficient, that is a non-parametric measure of the strength and direction of the monotonic relationship between two ranked variables [10] and, Kendall's tau coefficient, that evaluates ordinal association by quantifying the agreement between paired rankings and is

particularly robust in the presence of ties and small sample sizes [10]. Both measures quantify the degree to which the ordering of classes is aligned with increasing surgical complexity. Higher correlation values indicate stronger monotonic consistency between class designation and SCF progression.

B. Increase between Classes

The regularity of the increments between classes was evaluated based on the standard deviation of the differences in SCF values across consecutive classes. This metric provides a quantitative assessment of the dispersion in the magnitude of growth, allowing for the identification of variability patterns in the progression. In this context, a high standard deviation reflects greater inconsistency in the increments, suggesting irregular, non-monotonic, or unpredictable transitions between classes, which may indicate classification noise or instability in the underlying model. Conversely, a low standard deviation indicates that the increments are more homogeneous, reflecting a smoother and more stable progression that is closer to the expected ordinal structure. Therefore, this measure serves as an indicator of coherence in the scaling of surgical complexity, helping to evaluate whether the classification system preserves a consistent gradation between categories [2].

C. Classes Entropy

The distribution of patients among the classes was evaluated using Shannon entropy, a widely used metric that quantifies the degree of categorical dispersion within a discrete probability distribution. In this context, entropy provides a measure of how evenly patients are distributed across the different classes. Higher entropy values indicate a more uniform distribution, where patients are more evenly spread among all categories, reflecting greater heterogeneity and less dominance of any single class. In contrast, lower entropy values suggest that patients are concentrated in a limited number of classes, indicating a more imbalanced or skewed distribution.

Shannon entropy is mathematically defined by (1) [11]. It quantifies the average amount of information expected when observing the variable. The higher the entropy, the greater the uncertainty of the outcome.

$$H(X) = -\sum p(x_i) log_2(p(x_i)) \quad (1)$$

Where: $p(x_i)$ is the probability of obtaining a value x_i.

Because the two classification models contain a different number of categories (29 and 15 classes, respectively), entropy values were normalized by the theoretical maximum entropy of each model by (2). This normalization scales entropy values to the interval [0, 1], allowing direct comparison between classification systems with different numbers of classes.

$$H_N = \frac{H}{log_2(K)} \quad (2)$$

Where H is the Shannon entropy and K is the number of classes in the classification model.

D. Rationale for the Selected Indicators

The three indicators used in this study were selected because they evaluate complementary dimensions of classification consistency. Ordinal coherence assesses whether the progression of classes follows the expected monotonic increase in surgical complexity. The analysis of increments between classes evaluates the regularity of complexity transitions and identifies abrupt fluctuations that may introduce instability into the classification process. Shannon entropy characterizes how patients are distributed among the available categories, providing information about the concentration or fragmentation of the classification structure. These indicators provide an indirect assessment of statistical noise by evaluating the degree of consistency, stability, and organization exhibited by the classification models.

III. Results

The results of the analysis of the Surgical Complexity Factor (SCF) obtained by the 105 patients in the two models evaluated: Model 1 (29 classes) and Model 2 (15 classes) are presented according to the three axes of analysis calculated: ordinal coherence, increase and entropy, between classes.

A. Ordinal Coherence

The analysis of the two classification models revealed notable differences in the occurrence and magnitude of inversions (Table I). In Model 1, six inversions were detected, with negative delta values ranging from −29.67 to −6, indicating variability in the severity of these deviations. While some inversions correspond to minor decreases in the mean between consecutive classes (Δ = −6), others reflect more pronounced drops (Δ = −25, −29.67), highlighting disruptions in the expected hierarchical progression, and suggesting that the numerical ordering of classes does not consistently correspond to the intended complexity levels.

In contrast, Model 2 exhibited only three inversions, two of which are minor (Δ = −3), while one represents the largest reduction observed across both models (Δ = −36.8). Overall, Model 2 not only reduces the number and magnitude of inversions but also demonstrates a progression more closely aligned with the expected monotonic behavior.

TABLE I. SCF MEAN AND INCREASES OBTAINED BY CLASSES IN BOTH CLASSIFICATION MODELS

Model 1			Model 2		
Class	**Mean SCF**	**Δ**	**Class**	**Mean SCF**	**Δ**
1	3	3	1	3	3
2	12	9	2	7	4
3	18,5	6,5	3	9	2
4	7	**–11,5**	4	12	3
5	38,67	31,67	5	16	4
6	9	**–29,67**	6	38,67	22,67
7	23	14	7	36	**–2,67**
8	36	13	8	38	2
9	38	2	9	48	10
10	51	13	10	48	0
11	45	**–6**	11	70	22
12	20	**–25**	12	33,2	**–36,8**
15	78	58	13	58,95	25,75
16	101,33	23,33	14	80,56	21,61
17	126,17	24,83	15	77,67	**–2,9**
22	107	**–19,17**			
24	117	10			
25	140,4	23,4			
27	127,5	**–12,9**			
29	186	58,5			

These findings indicate that the structure of Model 2 provides a stronger and more reliable correspondence between class designation and surgical complexity, offering a more consistent framework for clinical classification.

To complement the inversion analysis, the association between class order and mean SCF values (Table 1) was assessed using Spearman's rank correlation coefficient and Kendall's tau coefficient.

Both models exhibited strong and statistically significant monotonic relationships indicating that higher class numbers are generally associated with greater surgical complexity (Table 2). Although the correlation coefficients were similar for both models, the analyses of ordinal inversions and increment dispersion revealed fewer local inconsistencies in Model 2, suggesting greater structural stability despite comparable overall monotonicity.

TABLE II. SPEARMAN'S RANK CORRELATION COEFFICIENT AND KENDALL'S TAU COEFFICIENT

Model	Spearman's ρ	p-value	Kendall's τ	p-value
Model 1	0.929	3.23×10^{-9}	0.811	9.06×10^{-9}
Model 2	0.903	4.13×10^{-6}	0.804	3.15×10^{-5}

B. Analysis of Increases between Classes

Comparison of the increments (Δ) in mean Surgical Complexity Factor (SCF) between consecutive classes for Model 1 and Model 2 is shown in Fig.1. Negative values indicate ordinal inversions, where a higher-numbered class exhibits a lower mean SCF than the preceding class.

Comparison of the standard deviation of class increments (σΔ) for both classification models is shown in Fig. 2. Lower values indicate more homogeneous transitions between consecutive classes and, therefore, greater structural stability of the classification framework. Model 1 presents an average standard deviation of the increments of $\sigma_{\Delta 1}$= 23.78, suggesting a more erratic progression with abrupt variations between complexity levels. In contrast, Model 2 exhibits an average standard deviation of $\sigma_{\Delta 2}$= 15.24, indicating greater stability in how complexity increases across the classes. This result complements the previous analyses and reinforces that Model 2 organizes its SCF progression scale more consistently.

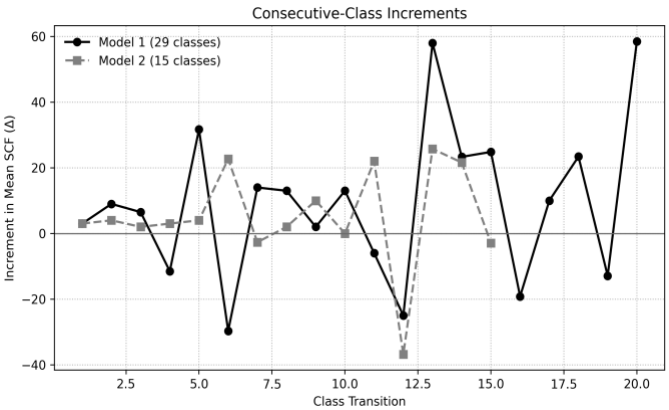


Fig. 1. Consecutive-class increments [Model 1 vs. Model 2]

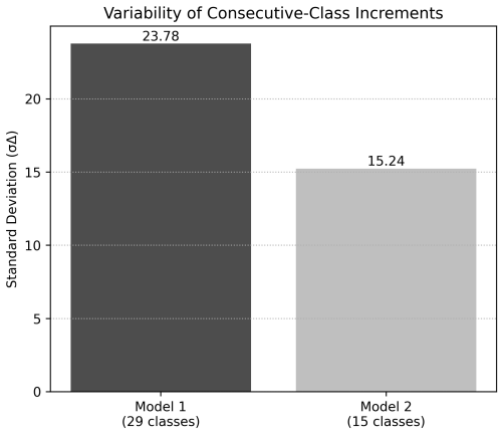


Fig. 2. Variability of Consecutive-Class Increments.

C. Classes Entropy

In this study, entropy was calculated using the observed frequencies of patients in each class of both classification models (Table III). The entropy analysis was performed using both raw and normalized Shannon entropy values. The normalized entropy was used for model comparison because the two classification systems contain different numbers of classes. Higher normalized entropy values indicate a more uniform distribution of patients among categories, whereas lower values indicate greater concentration in specific classes.

TABLE III. FREQUENCIES OF PATIENTS IN EACH CLASS OF BOTH CLASSIFICATION MODELS

Model 1		Model 2	
Class	No. patients	Class	No. patients
1	7	1	7
2	4	2	6
3	6	3	1
4	6	4	4
5	3	5	6
6	1	6	3
7	1	7	1
8	1	8	1
9	1	9	1
10	1	10	1
11	1	11	1
12	1	12	5
13	0	13	45
14	0	14	23
15	4	15	3
16	6		
17	36		
18	0		
19	0		
20	0		
21	0		
22	2		
23	0		
24	1		
25	5		
26	0		
27	2		
28	0		
29	16		

Model 1 (29 classes). Using both equations (1) and (2) this model presents:

H_{M1}= 3.67

H_{max}= log2 (29) = 4.858

$$H_{N-M1} = \frac{H}{log_2(K)} = \frac{3.67}{4.585} = 0.756$$

This result indicates that patient distribution occupies approximately 75.6% of the maximum possible dispersion, suggesting a relatively heterogeneous classification structure with moderate concentration in a limited number of categories.

Model 2 (15 clsses). Using both equations (1) and (2) this model presents:

H_{M2}=2.94.

H_{max}=log2(15) = 3.907

$$H_{N-M2} = 0.752$$

Because the two classification systems contain different numbers of categories, entropy values were normalized by their respective maximum entropy values. The resulting normalized entropies were H_{N-M1}=0.756 and H_{N-M2}=0.752. These values indicate that both models exhibit a very similar relative distribution of patients across their available categories, with differences below 1%.

IV. Discussion

The results obtained from the three analytical approaches consistently indicate that Model 2 exhibits a more organized classification structure than Model 1. The reduced number of ordinal inversions suggests that the progression of classes more closely follows the expected increase in surgical complexity. Similarly, the lower dispersion of class increments indicates a smoother transition between categories, reducing abrupt fluctuations that may complicate clinical interpretation.

The normalized entropy obtained for Model 1 indicates that patients are distributed across a considerable proportion of the available classes. Although the classification exhibits substantial dispersion, the concentration of cases in specific categories, particularly Class 17 and Class 29, prevents the distribution from reaching maximum uniformity. Consequently, the model can be characterized as moderately fragmented, with patients occupying a broad range of surgical complexity categories.

Although the raw entropy was higher for Model 1, this result is largely explained by the larger number of classes available in the classification scheme. After normalization, both models exhibited nearly identical entropy values (H_{N-M1}=0.756 and H_{N-M2}=0.752), indicating comparable levels of relative dispersion. Therefore, entropy alone does not provide evidence supporting the superiority of either model. Instead, the differences observed in ordinal inversions and increment variability become the primary indicators distinguishing the structural consistency of the two classification frameworks.

From a clinical perspective, a classification system with fewer structural inconsistencies may facilitate more reproducible decision-making processes. When class ordering reflects a coherent increase in complexity, clinicians may interpret category transitions more consistently and communicate treatment strategies more effectively. Although the present study does not evaluate surgical outcomes or inter-observer agreement, the observed reduction in structural noise may contribute to greater standardization in the assessment of cleft lip and palate severity.

A. Limitations of the Study

The findings should be interpreted considering some limitations. First, the study was conducted using a relatively modest sample of 105 patients obtained from a single clinical center. Second, some classes may contain a limited number of observations, potentially affecting the stability of mean SCF estimates. Third, the analysis focused exclusively on the internal consistency of the classification models and did not include multiple evaluators, external validation datasets, or direct measures of inter-rater reliability. Consequently, the results should be viewed as preliminary evidence of structural consistency rather than definitive proof of clinical superiority.

V. Conclusion

The present study provides preliminary evidence that Model 2 exhibits lower structural statistical noise than Model 1 when evaluated through ordinal coherence, increment regularity, and entropy-based measures. The results suggest that less fragmented classification frameworks may support more consistent representations of surgical complexity and contribute to the standardization of clinical assessment processes. Likewise, because some categories contained relatively few observations, class-level mean SCF values may be sensitive to sampling variability.

Future studies should expand the evaluation framework by incorporating larger multicenter datasets, independent clinical evaluators, and additional agreement metrics (e. g. weighted kappa coefficients). Such analysis would allow a more comprehensive characterization of decision variability and strengthen the validation of classification systems used in cleft lip and palate management.

References

[1] Dlugos KV, Mazwi M, Lao R, Honjo O. Noise is an underrecognized problem in medical decision making and is known by other names: a scoping review. BMC Med Inform Decis Mak. 2025 Feb 17;25(1):86. doi: 10.1186/s12911-025-02905-z. PMID: 39962496; PMCID: PMC11834221.

[2] Daniel Kahneman, Olivier Sibony, & Cass R. Sunstein (2021). Noise: A Flaw in Human Judgment.

[3] Cleft Hospital Pakistan, Cleft lip and palate. Available at: https://clefthospital.com/cleft-lip-plate/

[4] Gundlach K. K. H. & Maus C. (2006). *Epidemiological studies on the frequency of clefts in Europe and worldwide*. Journal of Cranio-Maxillofacial Surgery, 34(Suppl 2), 1–2.

[5] SUMA Center. Centro Integral de Labio y Paladar Hendido. México. https://www.smiletrainla.org/en/cleft-leadership-center/suma-old

[6] B. Gutiérrez-Sánchez, "Quantitative Classification of Cleft-Lip Palates Applying the Multicriteria Decision Analysis Method," Thesis, Bachelor's in Biomedical Engineering. Dept. Elect. Eng., Universidad Autónoma Metropolitana Iztapalapa, Mexico, 2022. (in Spanish)

[7] B. Gutiérrez-Sánchez, J. Maya-Behar J, R. M. De La Cerda, M. R. Ortiz-Posadas, "A Surgical Complexity Factor for Clefts of Primary and

Secondary Palates," in *Proc 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* Sydney, 2023. doi: 10.1109/EMBC40787.2023.10340486.

[8] Al-Jaishi, A. A., Cuerden, M. S., Luo, B., et al., "Statistical analysis of Likert-based ordinal scales: a guide for clinical trialists," *BMC Medical Research Methodology*, 2026

[9] Vanbelle, S., Hernandez Engelhart, C., & Blix, E. (2024). "A comprehensive guide to study the agreement and reliability of multi-observer ordinal data." BMC Medical Research Methodology, 24, 310. https://doi.org/10.1186/s12874-024-02431-y

[10] Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). John Wiley & Sons.

[11] Cover, T. M., & Thomas, J. A. (2021). *Elements of Information Theory* (2nd ed.). Wiley

Development of an Analog Electronic Circuit for Cardiac Cycle Simulation Based on a Lumped-Parameter Model

Figueiredo, L. E. T. C
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
cassianolynnda@gmail.com

Sales, I.Q.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
Isasales40@gmail.com

Sousa, M. L. L.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
maria.lisboa.sousa22@gmail.com

Araújo, K. P. B. D
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
karenpbaraujo@gmail.com

Fonseca, M. C. P.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
conceicao@ufpa.br

Miziara, I. M.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
immiziara@ufpa.br

***abstract*—The cardiovascular system can be represented through electrical analogies that establish equivalences between physiological and electrical variables, enabling the reproduction of the cardiac cycle by means of analog electronic circuits. This work presents the development of an analog electronic circuit designed to reproduce the principal cardiovascular pressure waveforms using commercially available electronic components and a simplified circuit topology, thereby facilitating its physical implementation for educational and experimental applications. The proposed circuit models the four cardiac chambers, represents the cardiac valves using diodes, incorporates the pulmonary and systemic circulations, and includes an active stage to emulate the mechanical behavior of the left ventricle. The proposed system was initially validated through circuit simulations performed in the Proteus environment and subsequently implemented on a breadboard using discrete electronic components. The experimentally acquired waveforms were compared with both the simulated responses and reference physiological pressure waveforms, demonstrating good agreement in terms of waveform morphology and temporal characteristics. These findings indicate that the proposed circuit accurately reproduces the principal features of the cardiac cycle, thereby validating the electrical analogy as an effective educational tool for Biomedical Engineering. Furthermore, the developed platform integrates concepts from cardiovascular physiology, analog electronics, and biomedical instrumentation while providing a low-cost experimental framework for future investigations and comparative analyses of physiological and pathological cardiovascular conditions.**

***Keywords*—Cardiovascular system; Analog electronic circuit; Biomedical Engineering; Biomedical instrumentation; Electrical analogy.**

I. Introduction

The cardiovascular system is responsible for transporting blood throughout the body and is primarily composed of the heart, blood vessels, and blood. Cardiac function is governed by the cardiac cycle, which consists of two main phases: diastole, during which the atria and ventricles are filled with blood while the atrioventricular valves remain open, and systole, during which the ventricles contract and eject blood into the aorta and pulmonary arteries [1]. These phases produce characteristic variations in pressure and volume within the cardiac chambers, whose assessment plays a fundamental role in the diagnosis of several cardiovascular diseases [1].

Several approaches have been proposed to investigate cardiovascular dynamics. Computational models, such as the one developed by Heldt et al. (2002), represent the cardiovascular system through mathematical equations, enabling the analysis of its response under different physiological conditions [2]. Alternatively, analog electrical models exploit the correspondence between hemodynamic variables, including pressure, flow, blood volume, vascular resistance, and vascular compliance, and their electrical counterparts, namely voltage, current, charge, resistance, and capacitance. An example is the model proposed by Montenegro (2022), in which the four cardiac chambers, the four cardiac valves represented by diodes, and the pulmonary and systemic circulations are modeled using an equivalent electrical circuit whose parameters are correlated with the geometric and mechanical properties of the cardiovascular system [3]. This model reproduces the characteristic morphology of physiological pressure waveforms, including the systolic and diastolic pressures observed in the aorta. However, it presents limitations related to electrical noise and to maintaining pressure consistency throughout the vascular network [3].

Although computational models provide accurate mathematical representations of cardiovascular dynamics, they do not allow the physical observation and manipulation of the system components. Analog electrical models, in contrast, offer a tangible representation of cardiovascular physiology but remain largely restricted to simulation environments. This limitation

arises because their component values are obtained through the direct conversion of hemodynamic parameters, resulting in electrical elements that do not correspond to commercially available components. Furthermore, the excitation sources employed in these models often require voltage levels that are incompatible with conventional laboratory instrumentation. In the model proposed by Montenegro (2022), for example, the right and left ventricular excitation sources operate with amplitudes of 20 V and 120 V, respectively. In addition, the high level of model granularity requires a large number of electrical components, increasing implementation complexity and limiting its practical realization [3]. Consequently, although these models are suitable for numerical simulation, they are difficult to implement experimentally, restricting their use as educational platforms for teaching analog electronics, electrical circuits, and cardiovascular physiology.

Motivated by these limitations, the present work proposes the development of a physical analog electronic circuit capable of reproducing the behavior of the cardiovascular system using commercially available electronic components. The proposed circuit represents the four cardiac chambers, namely the right atrium, right ventricle, left atrium, and left ventricle, together with the corresponding cardiac valves and the pulmonary and systemic circulations. The design was based on the lumped-parameter model proposed by Montenegro (2022), in which each vascular segment is represented by equivalent electrical elements whose parameters are derived from physiological characteristics, including vessel length, cross-sectional area, wall thickness, and elastic modulus [3]. To enable practical implementation, the original model was simplified by concentrating the hemodynamic properties of each circulatory segment into a reduced number of equivalent electrical elements. This approach substantially reduces circuit complexity while preserving the characteristic systolic and diastolic pressure waveforms under normal physiological conditions. As a result, the proposed circuit provides an inexpensive experimental platform suitable for educational applications and future research in cardiovascular modeling.

II. Methodology

A. Methodological Overview

The development of this work followed a structured methodology comprising theoretical analysis, circuit design, simulation, hardware implementation, and experimental validation. Figure 1 summarizes the methodological workflow adopted in this study, illustrating the sequence of stages and the relationship between them.

The first stage consisted of establishing the electrical representation of the cardiovascular system by defining the analogies between physiological and electrical quantities. In this representation, blood flow is analogous to electric current, while pressure is represented by voltage. This electro-hemodynamic correspondence enables the development of analog electronic circuits capable of reproducing the pressure waveforms observed throughout the cardiovascular cycle.

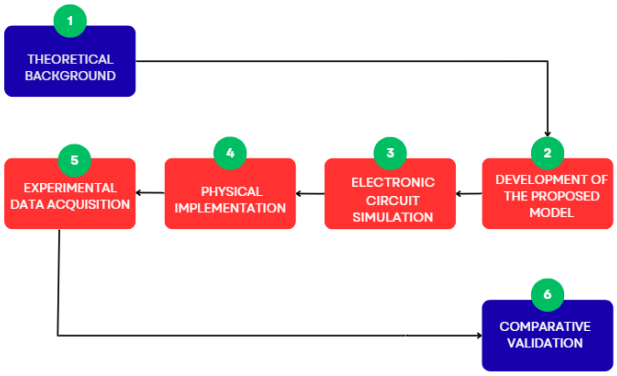


Figura 1: Flowchart illustrating the methodological steps adopted in the development of this study, from the theoretical background to the experimental validation of the proposed circuit.

B. Lumped-Parameter Method

A rigorous description of blood flow in a vessel requires partial differential equations in which pressure and flow vary simultaneously with time and along the vessel length [2], [8]. The lumped-parameter method simplifies this formulation by assuming that, within a vascular segment sufficiently short compared with the wavelength of the pressure pulse, spatial variations can be neglected [2], [8]. Consequently, each segment is represented by a single set of lumped parameters, and the resulting mathematical model is described by ordinary differential equations that are formally equivalent to those governing an RLC electrical circuit [2], [3]. This equivalence constitutes the basis of the electro-hemodynamic analogy.

In this approach, each vascular segment is represented by a block composed of three electrical elements whose values are determined from the geometric and mechanical properties of the vessel, including its length, inner radius, wall thickness, and elastic modulus. The electrical resistance represents the viscous dissipation that opposes blood flow, the inductance represents the inertia of the blood mass accelerated during ventricular ejection, and the capacitance represents vascular compliance, corresponding to the ability of the vessel wall to expand and store blood volume.

The principal advantage of the lumped-parameter method is its flexibility, allowing the cardiovascular system to be represented with different levels of anatomical detail. Increasing the number of vascular segments improves the physiological fidelity of the model but also increases the number of electrical components, making physical implementation considerably more complex. For this reason, the present work adopts a simplified representation of the cardiovascular system, which is described in the following section.

C. Windkessel Model

The Windkessel model represents the simplest application of the lumped-parameter method, in which the systemic circulation is described using equivalent electrical elements [3], [8].

In this work, the two-element Windkessel model, composed of a peripheral vascular resistance and an arterial compliance, was adopted because it is sufficient to reproduce the characteristic systolic and diastolic pressure waveforms under normal physiological conditions while reducing the complexity of the physical implementation [8]. The characteristic resistance of the aorta was retained in the model, whereas the inductive element was omitted because the inductance values obtained from the lumped-parameter formulation are extremely small, making their physical implementation of limited practical significance [3]. The mathematical behavior of the adopted model is described by (1).

$$C\frac{dP(t)}{dt} + \frac{P(t)}{R} = Q(t) \tag{1}$$

where $P(t)$ represents the arterial pressure, $Q(t)$ represents the blood flow rate, C denotes the arterial compliance, and R denotes the peripheral vascular resistance.

D. Electro-Hemodynamic Analogy

The analog electronic circuit was developed based on the correspondence between physiological variables and equivalent electrical quantities. This electro-hemodynamic analogy allows the cardiovascular system to be represented by conventional electrical components while preserving its fundamental dynamic behavior.

Blood pressure is represented by electrical voltage, since both quantities act as the driving force responsible for producing flow in their respective systems. Likewise, blood flow is represented by electrical current because both describe the transport of a conserved quantity through a conductive medium.

Vascular resistance is represented by an electrical resistor. In the cardiovascular system, vascular resistance opposes blood flow and is mainly determined by vessel diameter and blood viscosity. Similarly, an electrical resistor limits the flow of electric current by opposing the movement of electric charge.

Arterial compliance is represented by a capacitor. Physiologically, arterial compliance corresponds to the ability of blood vessels to expand during ventricular ejection and to release the stored blood volume during diastole, thereby attenuating pressure fluctuations. In the electrical circuit, the capacitor performs an analogous function by storing and releasing electrical charge, smoothing voltage variations and reproducing the elastic behavior of the arterial walls.

Finally, the cardiac valves are represented by diodes. Heart valves ensure unidirectional blood flow by opening and closing according to the pressure difference between adjacent cardiac chambers. Likewise, a diode allows electrical current to flow in only one direction, preventing reverse current and accurately reproducing the valve function within the electrical model.

III. Development of the Proposed Model

Based on the lumped-parameter method and the electro-hemodynamic analogies described in the previous section, an analog electronic circuit was developed to physically reproduce the behavior of the four cardiac chambers, as well as the pulmonary and systemic circulations. The proposed model follows the physiological blood flow pathway, comprising the right atrium, right ventricle, pulmonary circulation, left atrium, left ventricle, and systemic circulation, as illustrated in the complete schematic shown in Fig. 2.

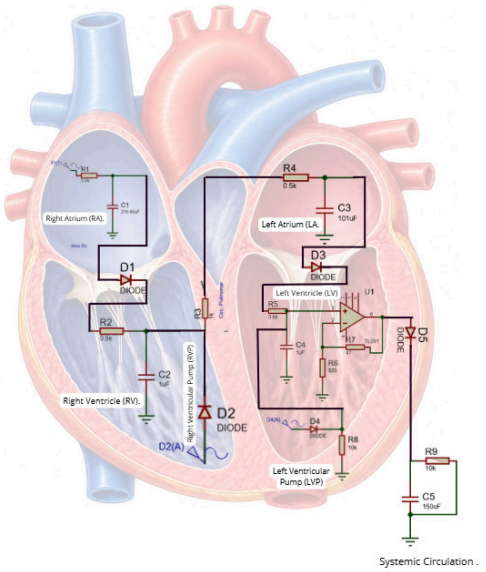


Figura 2: Proposed electronic circuit superimposed on the cardiac anatomy, highlighting the correspondence between each circuit block and the cardiac chamber or vascular compartment it represents. Source: Prepared by the author (2026).

The second modification involved the cardiac chambers. Passive RC blocks were retained for the right atrium, right ventricle, and left atrium, whereas the left ventricle was implemented as an active circuit based on a TL081 operational amplifier. This modification constitutes the principal contribution of the proposed model, as it provides sufficient signal amplification to reproduce the characteristic amplitude and slope of the left ventricular pressure waveform, which cannot be accurately generated using a passive RC network alone.

These simplifications were introduced to obtain a circuit that is both physically feasible to implement and suitable for educational and experimental applications while preserving the characteristic pressure waveforms at the physiologically most relevant locations, namely the right ventricle, left atrium, and left ventricle, instead of reproducing the entire vascular tree.

A. Division into the Four Cardiac Chambers

The right atrium, right ventricle, and left atrium were implemented as first-order RC networks, each composed of a series resistor and a capacitor connected in parallel with the output. These networks reproduce the charging and discharging behavior analogous to the pressure-volume relationship of the corresponding cardiac chamber. The electrical components associated with each chamber are listed in Table I.

The left ventricle was implemented as an active circuit based on a TL081 operational amplifier. The circuit comprises the input resistor R5 (0.5 kΩ) and the feedback network formed by C4, R6 (820 Ω), and R7 (47 Ω). This configuration amplifies the signal generated by the left atrium and actively reproduces the characteristic left ventricular pressure waveform.

The systemic circulation was represented by a two-element Windkessel model composed of resistor R9 (10 kΩ), corresponding to the peripheral vascular resistance, connected in parallel with capacitor C5 (150 μF), which represents the arterial compliance.

Tabela I: Circuit components and their respective physiological representation.

Component	Value	Physiological representation
R1	0.5 kΩ	Resistance of the right atrium
C1	216.45 μF	Compliance of the right atrium
D1	—	Tricuspid valve
R2	0.5 kΩ	Resistance of the right ventricle
C2	1 μF	Compliance of the right ventricle
D2	—	Pulmonary valve (RV pump)
R3	1 kΩ	Pulmonary vascular resistance
R4	0.5 kΩ	Resistance of the left atrium
C3	101 μF	Compliance of the left atrium
D3	—	Mitral valve
R5	0.5 kΩ	Input resistance of the left ventricle
C4	1 μF	Feedback capacitor of the left ventricle
R6	820 Ω	Feedback resistor of the left ventricle
R7	47 Ω	Feedback resistor of the left ventricle
U1 (TL081)	—	Operational amplifier of the left ventricle
D4	—	LV pump diode
R8	10 kΩ	Resistance associated with the LV pump
D5	—	Aortic valve
R9	10 kΩ	Peripheral vascular resistance
C5	150 μF	Arterial compliance

B. Use of Diodes as Valves

Each cardiac valve was represented by a diode positioned to conduct current only in the physiological direction of blood flow while blocking reverse flow when the downstream pressure exceeds the upstream pressure. The correspondence between the diodes and the cardiac valves is summarized in Table II.

Tabela II: Diodes used as cardiac valves in the circuit.

Component	Valve	Location
D1	Tricuspid valve	Between the right atrium and the right ventricle
D3	Mitral valve	Between the left atrium and the left ventricle
D5	Aortic valve	Between the left ventricle and the systemic circulation
D2	Pulmonary valve	Between the right ventricle and the pulmonary circulation

It should be noted that the Right Ventricle Pump (RVP) block simulates blood ejection from the right ventricle into the pulmonary circulation. In the proposed circuit, diode D2 performs the dual function of rectifying the sinusoidal excitation signal and reproducing the action of the pulmonary valve. In contrast, the reference model presented in [3] represents the pulmonary valve as an independent diode separate from the right ventricular pump. This modification constitutes an additional simplification of the proposed circuit, reducing the number of components without compromising its intended physiological behavior.

C. Ventricular Excitation Sources ("Pumps")

In addition to the passive RC networks, the proposed circuit incorporates three sinusoidal voltage sources. The first source excites the right atrium with an amplitude of 100 mV, an offset of 300 mV, and a frequency of 250 mHz, representing the venous return that supplies blood to the heart and initiates the simulated cardiac cycle.

The remaining two sources reproduce the active contraction of the ventricles. The Right Ventricle Pump (RV Pump) emulates the pressure generation associated with right ventricular contraction, whereas the Left Ventricle Pump (LV Pump) reproduces the corresponding mechanism in the left ventricle. Physiologically, ventricular contraction generates the pressure required to eject blood into the pulmonary and systemic circulations, a process that cannot be reproduced using passive electrical elements alone.

This approach follows the same principle adopted in the reference model [3], in which the “Right Ventricle - Pump” and “Left Ventricle - Pump” blocks employ sinusoidal voltage sources associated with diodes to reproduce ventricular contraction. In that model, both sources operate at a frequency of 1.2 Hz (72 beats per minute), with amplitudes of 20 V and 120 V for the right and left ventricles, respectively.

In the proposed circuit, the excitation mechanism was simplified to facilitate physical implementation. The RV Pump consists of an AC voltage source combined with a 7.04 V DC source and a series diode, whereas the LV Pump is composed of a sinusoidal voltage source and a diode. Both excitation sources operate at a frequency of 1.2 Hz. Their amplitudes were reduced to 2 V for the RV Pump, with a 700 mV offset, and 12 V for the LV Pump, with a 700 mV offset and a phase angle of 0°. This scaling preserves the original 6:1 amplitude ratio between the left and right ventricular excitation sources while reducing the operating voltage to a range compatible with the physical implementation of the proposed circuit.

D. Ventricular Excitation Sources ("Pumps")

In addition to the passive RC networks, the proposed circuit incorporates three sinusoidal voltage sources. The first source excites the right atrium with an amplitude of 100 mV, an offset of 300 mV, and a frequency of 250 mHz, representing the venous return that supplies blood to the heart and initiates the simulated cardiac cycle.

The remaining two sources reproduce the active contraction of the ventricles. The Right Ventricle Pump (RV Pump) emulates the pressure generation associated with right ventricular contraction, whereas the Left Ventricle Pump (LV Pump) reproduces the corresponding mechanism in the left ventricle. Physiologically, ventricular contraction generates the pressure required to eject blood into the pulmonary and systemic circulations, a process that cannot be reproduced using passive electrical elements alone.

This approach follows the same principle adopted in the reference model [3], in which the “Right Ventricle - Pump” and “Left Ventricle - Pump” blocks employ sinusoidal voltage sources associated with diodes to reproduce ventricular contraction. In that model, both sources operate at a frequency of 1.2 Hz (72 beats per minute), with amplitudes of 20 V and 120 V for the right and left ventricles, respectively.

In the proposed circuit, the excitation mechanism was simplified to facilitate physical implementation. The RV Pump consists of an AC voltage source combined with a 7.04 V DC source and a series diode, whereas the LV Pump is composed of a sinusoidal voltage source and a diode. Both excitation sources operate at a frequency of 1.2 Hz. Their amplitudes were reduced to 2 V for the RV Pump, with a 700 mV offset, and 12 V for the LV Pump, with a 700 mV offset and a phase angle of 0°. This scaling preserves the original 6:1 amplitude ratio between the left and right ventricular excitation sources while reducing the operating voltage to a range compatible with the physical implementation of the proposed circuit.

IV. Results and Discussion

This section presents the pressure waveforms obtained at the nodes corresponding to the cardiac chambers from two complementary stages: computational simulations performed in Proteus and experimental measurements acquired from the physical implementation on a protoboard. The simulated and experimental waveforms were compared with the expected physiological pressure profiles of each cardiac chamber to evaluate the ability of the proposed circuit to reproduce cardiovascular dynamics. As a reference for the analyses presented in the following sections, Fig. 6 shows the physiological pressure waveforms of the cardiac cycle used for comparison with the signals generated by the proposed circuit.

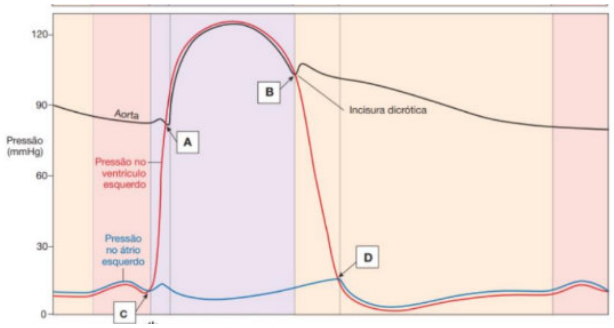


Figura 3: Reference physiological pressure curves of the left cardiac cycle, illustrating the relationship between pressure and volume. Adapted from [6].

A. Right Ventricular Pressure

Figure 5 compares the right ventricular pressure waveform obtained from the physiological reference, the Proteus simulation, and the experimental measurements. For comparison purposes, all waveforms were normalized, allowing their morphological characteristics to be evaluated independently of their amplitudes. It should be noted that the physiological waveform

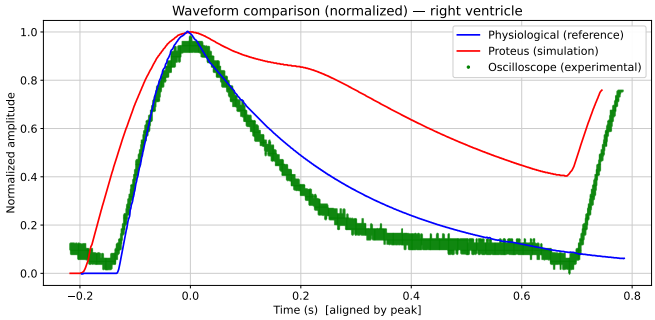


Figura 4: Normalized comparison of the right ventricular pressure waveform obtained from the schematic physiological reference, the Proteus simulation, and the experimental measurements, aligned at the systolic peak. Source: Prepared by the author (2026), based on the Proteus simulation, the experimental measurements, and the physiological reference derived from the literature.

is schematic and was constructed from values reported in the literature.

Under normal physiological conditions, right ventricular pressure ranges from approximately 2 to 5 mmHg during diastole and reaches about 25 mmHg during systole, reflecting the relatively low resistance of the pulmonary circulation. In the Proteus simulation, the signal varied from 0.4 V to 2.06 V, corresponding to approximately 20 mmHg according to the adopted conversion of 1 V to 10 mmHg, which is consistent with the expected physiological range.

The normalized comparison demonstrates that the experimental waveform closely follows both the simulated signal and the physiological reference, preserving the rapid pressure increase during systole and the gradual decrease during diastole. Minor differences, including a slightly delayed peak and a steeper decay in the experimental waveform, are attributed to the limited sampling rate of the acquisition system and to the tolerances of the electronic components used in the physical implementation. Overall, these results indicate that the proposed circuit adequately reproduces the characteristic dynamics of right ventricular pressure.

B. Left Atrial Pressure

Figure 6 compares the left atrial pressure waveform obtained from the physiological reference, the Proteus simulation, and the experimental measurements acquired using an oscilloscope. For comparison purposes, all waveforms were normalized to emphasize their morphological characteristics, whereas the quantitative analysis is based on the measured voltage values.

Unlike ventricular pressure, left atrial pressure exhibits a characteristic waveform composed of distinct pressure waves associated with the different phases of the cardiac cycle. In the present work, however, the physiological reference was intentionally simplified to represent only the overall variation in atrial pressure throughout the cardiac cycle, consistent with

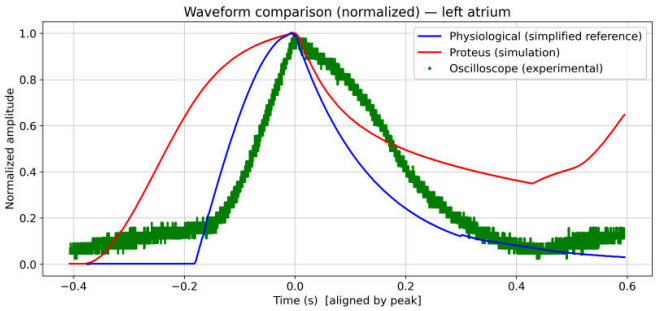


Figura 5: Normalized comparison of the left atrial pressure waveform obtained from the schematic physiological reference, the Proteus simulation, and the experimental measurements. The reference waveform represents the overall variation in atrial pressure without explicitly reproducing the individual atrial pressure waves. Source: Prepared by the author (2026), with the physiological reference based on [5].

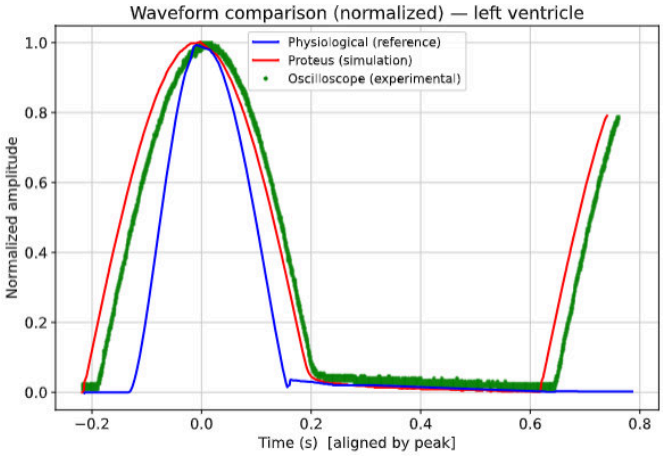


Figura 6: Normalized comparison of the pressure waveform at the output of the Left Ventricular Pump (LVP), which serves as the excitation source for the left ventricle, obtained from the Proteus simulation and the experimental measurements, aligned at the waveform peak. Source: Prepared by the author (2026).

the objective of reproducing the global behavior of the cardiac chambers rather than the detailed atrial pressure components.

In the Proteus simulation, the left atrial signal ranged from 0.4 V to 1.72 V, reaching its maximum value at approximately 0.41 s. This timing is consistent with the expected delay between atrial and ventricular activity during the cardiac cycle. The experimentally measured signal varied from 0.54 V to 1.35 V, with a period of approximately 0.83 s, confirming the stable and periodic operation of the proposed circuit.

Comparison of the normalized waveforms reveals good agreement between the simulated and experimental results, with both preserving the overall morphology of the simplified physiological reference. Minor discrepancies, including a slightly delayed rising edge and a steeper decay in the experimental waveform, are primarily attributed to the limited sampling rate of the acquisition system and to the tolerances of the electronic components used in the prototype. Despite these differences, the close agreement between the simulated and experimental waveforms indicates that the proposed circuit successfully reproduces the overall temporal evolution of left atrial pressure.

C. Pressure at the LV Pump Output

Figure 7 compares the waveform measured at the output of the Left Ventricular Pump (LVP), which serves as the excitation source for the active left ventricular stage, between the Proteus simulation and the experimental measurements. This signal provides the excitation required to reproduce the pressure generated during left ventricular contraction. For comparison purposes, both waveforms were normalized between 0 and 1 to emphasize their morphological characteristics, whereas the quantitative analysis is based on the measured voltage values.

In the Proteus simulation, the excitation signal varied from approximately 0 V to 12 V, reaching its peak at about 0.21 s before gradually decreasing throughout the remainder of the cardiac cycle. According to the conversion adopted in the model (1 V $\equiv$ 10 mmHg), the peak value corresponds to approximately 120 mmHg, consistent with the physiological left ventricular systolic pressure.

Experimentally, the measured signal ranged from approximately 0 V to 8.5 V and exhibited a well-defined peak followed by a gradual decay during each cardiac cycle, confirming the stable and periodic operation of the excitation stage. Although the measured peak amplitude was lower than that predicted by the simulation because of component nonidealities, circuit losses, and limitations of the acquisition system, the waveform morphology and periodicity were preserved.

The close agreement between the simulated and experimental waveforms indicates that the proposed excitation stage reliably reproduces the driving signal required by the active left ventricular model. Consequently, this signal provides the appropriate excitation for generating the left ventricular pressure waveform discussed in the following subsection.

D. Left Ventricular Pressure

Figure 8 compares the normalized left ventricular pressure waveforms obtained from the physiological reference, the Proteus simulation, and the experimental measurements acquired using an oscilloscope. The left ventricular pressure corresponds to the output of the active stage implemented with the TL081 operational amplifier. To facilitate the comparison of waveform morphology independently of amplitude and temporal offsets, all signals were normalized to the range of 0 to 1 and aligned at the systolic peak ($t = 0$ s).

The physiological reference waveform was analytically constructed from cardiac-cycle timing parameters reported in the literature and therefore does not correspond to an experimental recording. A cardiac cycle duration of approximately

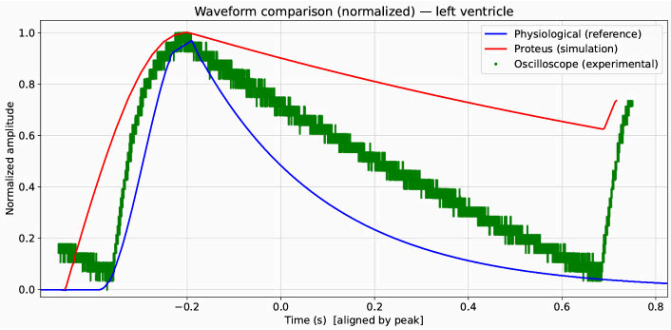


Figura 7: Normalized comparison of the left ventricular pressure waveform obtained from the physiological reference, the Proteus simulation, and the experimental measurements, aligned at the systolic peak ($t = 0$ s). The physiological reference waveform was analytically constructed from cardiac-cycle timing parameters reported in the literature (cardiac cycle of 0.8 s, systole of 0.3 s, diastole of 0.5 s, and isovolumetric relaxation of 0.06 s) and does not correspond to an experimental recording. Source: Prepared by the author (2026).

0.8 s (75 bpm), consisting of a ventricular systole of 0.3 s, a diastole of 0.5 s, and an isovolumetric relaxation period of approximately 0.06 s, was adopted [7]. The experimentally measured period was approximately 0.836 s, corresponding to a heart rate of about 72 bpm, which lies within the normal resting physiological range.

The simulated and experimental waveforms closely reproduce the systolic rising phase and the pressure peak, showing good agreement with the physiological reference during ventricular contraction. The principal difference is observed during diastole. Whereas physiological left ventricular pressure decreases rapidly at the end of systole because of isovolumetric relaxation, both the simulated and experimental waveforms exhibit a gradual pressure decay.

This behavior is characteristic of arterial pressure described by the Windkessel model adopted for the systemic circulation, in which the diastolic pressure follows an approximately exponential decay [8]. Consequently, the proposed circuit reproduces the systolic ejection dynamics and peak timing with good agreement, whereas the relaxation phase predominantly reflects the dynamics of the arterial compartment rather than the rapid decrease of true intraventricular pressure. This observation is consistent with the theoretical assumptions adopted in the circuit design and indicates that the proposed model reproduces arterial hemodynamics more accurately than the detailed intraventricular pressure profile.

E. Final Considerations

The differences observed between the simulated and experimental waveforms were consistent across all analyzed cardiac chambers. The experimental signals exhibited a slightly delayed rising edge, a steeper decay during the relaxation phase, and a higher noise level than the simulated waveforms. These differences are attributed primarily to the limited sampling rate of the acquisition system and to the tolerances of the electronic components, which introduce variations in the circuit time constants that are not present in the idealized simulation.

The diastolic phase of the left ventricle deserves particular attention. Whereas physiological intraventricular pressure decreases abruptly following systole because of isovolumetric relaxation, the simulated and experimental waveforms display a gradual decay characteristic of arterial pressure. This behavior is consistent with the two-element Windkessel model adopted for the systemic circulation, whose response is governed by an exponential discharge of the arterial compliance. Consequently, the proposed circuit accurately reproduces the systolic pressure generation and peak timing, while the relaxation phase is dominated by the response of the arterial compartment.

The main contribution of the proposed model becomes evident when compared with previous studies. Computational cardiovascular models [2], [4] provide accurate representations of hemodynamic behavior but are restricted to numerical simulations and cannot be directly implemented as physical systems. Analog electrical models, such as that proposed by Montenegro [3], also remain primarily simulation-oriented because they require a large number of components, excitation voltages of 20 V and 120 V, and electrical component values that are not commercially available, making practical implementation difficult.

In contrast, the proposed circuit adopts a simplified topology that prioritizes physical implementation while preserving the characteristic pressure waveforms at the principal cardiac chambers. By reducing the vascular network to its essential elements, eliminating impractical inductive components, scaling the excitation voltages to laboratory levels, and employing commercially available components, the proposed model provides a practical experimental platform for demonstrating the fundamental principles of cardiovascular hemodynamics.

V. Conclusion

This work presented the development and experimental validation of an analog electronic circuit capable of reproducing the principal pressure dynamics of the cardiovascular system using commercially available components. Based on the lumped-parameter method, the two-element Windkessel model, and electro-hemodynamic analogies, a simplified version of the reference model proposed by Montenegro [3] was developed by reducing the systemic circulation to an equivalent block, eliminating inductive elements, and scaling the excitation voltages to values compatible with laboratory implementation.

The results obtained from both the Proteus simulations and the experimental prototype demonstrated good agreement with the expected physiological pressure waveforms. The measured pressure levels were consistent with the adopted conversion factor of 1 V to 10 mmHg, corresponding to approximately 20 mmHg for the right ventricle and 120 mmHg for the left ventricle. The differences observed between the simulated

and experimental results are consistent with the limitations of the acquisition system and the tolerances of the electronic components used in the prototype.

Compared with previously reported analog models, the proposed circuit offers the advantage of physical realizability while maintaining the principal hemodynamic characteristics of the cardiovascular system. Consequently, it constitutes a practical platform for teaching cardiovascular physiology, validating lumped-parameter models, and supporting experimental activities in biomedical and electronic engineering.

Future work includes improving the representation of intraventricular pressure during diastole, expanding the vascular network while maintaining practical implementation, employing higher-speed data acquisition systems, developing a printed circuit board version of the prototype, and incorporating quantitative waveform similarity metrics, such as the Pearson correlation coefficient and the root mean square error (RMSE), to complement the qualitative analysis.

REFERÊNCIAS

[1] R. E. Klabunde, *Cardiovascular Physiology Concepts*, 3rd ed. Philadelphia, PA, USA: Wolters Kluwer, 2021.

[2] T. Heldt, E. B. Shim, R. D. Kamm, and R. G. Mark, "Computational modeling of cardiovascular response to orthostatic stress," *J. Appl. Physiol.*, vol. 92, no. 3, pp. 1239–1254, 2002.

[3] P. C. M. Montenegro, "Simulação do sistema cardiovascular: modelos elétrico e matemático," B.S. thesis, Dept. Elect. Eng., Univ. of Brasília, Brasília, DF, Brazil, 2022.

[4] A. Ferreira, S. Chen, M. A. Simaan, J. R. Boston, and J. F. Antaki, "A nonlinear state-space model of a combined cardiovascular system and a rotary pump," in *Proc. 44th IEEE Conf. Decision Control Eur. Control Conf. (CDC-ECC)*, Seville, Spain, 2005.

[5] D. U. Silverthorn, *Fisiologia Humana: Uma Abordagem Integrada*, 7th ed. Porto Alegre, Brazil: Artmed, 2017.

[6] A. C. Guyton and J. E. Hall, *Tratado de Fisiologia Médica*, 13th ed. Rio de Janeiro, Brazil: Elsevier, 2017.

[7] O. Bazan and J. P. Ortiz, "Duration of systole and diastole for hydrodynamic testing of prosthetic heart valves: comparison between ISO 5840 standards and in vivo studies," *Braz. J. Cardiovasc. Surg.*, vol. 31, no. 2, 2016, doi: 10.5935/1678-9741.20160036.

[8] N. Westerhof, J.-W. Lankhaar, and B. E. Westerhof, "The arterial Windkessel," *Med. Biol. Eng. Comput.*, vol. 47, no. 2, pp. 131–141, 2009.

Analysis of Dengue Viral Dynamics Under the Impact of Antiviral Treatment During Pregnancy

Isadora Vargas
Department of Electrical Engineering
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
isadoraferreiravargas@gmail.com

Emilly Idalina
Department of Electrical Engineering
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
emillybeaa@gmail.com

Danilo Regis
Department of Electrical Engineering
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—Dengue is a viral disease transmitted by the bite of infected female *Aedes aegypti* mosquitoes that represents a significant global public health challenge, particularly during pregnancy, when infection may result in severe maternal and fetal complications. This study investigates the impact of antiviral treatment on dengue viral dynamics throughout pregnancy using a within-host mathematical modeling approach. The proposed model consists of a system of ordinary differential equations describing the interactions among viral particles, susceptible monocytes, infected monocytes, and T lymphocytes, while incorporating the effects of antiviral therapy on viral replication. Numerical simulations were implemented in Python to evaluate treatment administration during different stages of pregnancy, namely the first, second, and third trimesters. Parameter values were adapted to account for physiological and immunological changes occurring throughout gestation. The results found in this work, indicate that pregnant women are more vulnerable to severe infection, particularly during the second trimester, due to reduced immune activity. Furthermore, the results also show that the proposed antiviral administration substantially decreases viral load and accelerates viral clearance, reducing the risk of adverse outcomes associated with dengue infection during pregnancy, such as Dengue Hemorrhagic Fever (DHF), especially when early administrated.**

***Index Terms*—dengue, mathematical modeling, antiviral treatment, pregnancy.**

I. Introduction

Dengue is a mosquito-borne viral disease caused by the dengue virus (DENV), which belongs to the *Flaviviridae* family and is primarily transmitted by infected female mosquitoes of the genus *Aedes*. This disease is considered one of the most significant arbovirus diseases worldwide, affecting millions of individuals annually, especially in tropical and subtropical regions [1]. The DENV infection can manifest simple symptoms or progress to severe forms characterized by plasma leakage, hemorrhagic manifestations, and organ dysfunction, associated with the occurrence of Dengue Hemorrhagic Fever (DHF), which represents a major challenge to public health systems [2].

After transmission, the virus initially infects immune cells, particularly monocytes, where viral replication occurs. The progression of the infection is strongly influenced by the interaction between viral particles and the host immune response [3]. Among the most relevant cellular populations involved in this process are susceptible monocytes, infected monocytes, and T lymphocytes, which play a fundamental role in controlling viral replication and eliminating infected cells. The balance between viral proliferation and immune activity determines the severity and duration of the disease [4].

Pregnancy introduces substantial physiological and immunological changes that may alter the course of dengue infection. To ensure fetal tolerance and proper gestational development, the maternal immune system undergoes adaptations that modify the activity and distribution of immune cells throughout the trimesters [5]. These alterations can influence the host's ability to control viral infections, potentially increasing susceptibility to severe outcomes, such as Dengue Hemorrhagic Fever (DHF). Dengue infection during pregnancy has been associated with adverse maternal and fetal complications, including preterm birth, low birth weight, fetal distress, and, in severe cases, maternal or neonatal death [6].

Despite the growing incidence of dengue worldwide, there is currently no widely established specific antiviral therapy for routine clinical use [7]. In this case, mathematical models provide a valuable framework for evaluating the potential effectiveness of therapeutic approaches, allowing researchers to investigate treatment outcomes under different biological conditions and intervention schedules. The mathematical model proposed in this work incorporates the interactions between viral particles, susceptible monocytes, infected monocytes, and T lymphocytes, while accounting for physiological changes associated with different gestational stages. Therefore, this study investigates the impact of antiviral treatment on dengue viral dynamics during pregnancy through a mathematical and computational approach, aiming to contribute to the advancement of personalized medicine and reduce the incidence of severe outcomes associated with DENV infection in pregnant women.

II. Literature Background

A. Immune Populations Involved in Dengue Infection

The progression of dengue infection is strongly influenced by the interaction between the dengue virus (DENV) and the host immune system. Following transmission through

the bite of an infected mosquito, viral particles enter the bloodstream and initially target susceptible immune cells, particularly monocytes, macrophages, and dendritic cells. These cells serve as primary sites for viral replication, contributing to the dissemination of the virus throughout the body [3]. The immune components involved in the DENV infection dynamics are represented in Figure 1.

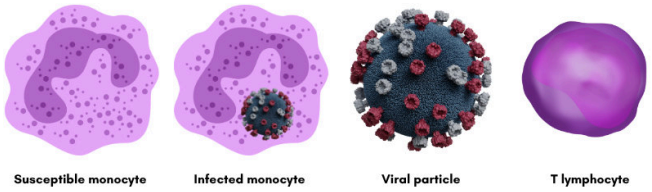


Fig. 1. Immune cells involved in DENV infection

Monocytes play a central role in dengue pathogenesis. Once infected, they become viral factories capable of producing and releasing new viral particles, thereby increasing the viral load. The number of infected monocytes is directly associated with the progression of the infection and the activation of the immune response. At the same time, susceptible monocytes are continuously recruited and exposed to viral particles, creating a dynamic interaction between viral replication and immune defense [3]. Once, susceptible monocytes are infected by a viral particle, cytotoxic T cells, such as T lymphocytes, recognize and eliminate infected cells, reducing viral dissemination, while helper T cells coordinate immune activity through the production of cytokines and the activation of other immune cells [4].

The balance between viral replication within infected monocytes and the immune response mediated by T lymphocytes largely determines disease severity. A robust and well-regulated immune response promotes viral elimination, whereas impaired or delayed immune activity may allow prolonged viral replication and increase the risk of severe clinical manifestations [4]. Therefore, understanding the interactions among viral particles, susceptible monocytes, infected monocytes, and T lymphocytes is essential for developing mathematical models capable of describing the dengue disease dynamics and evaluating potential therapeutic interventions.

B. Female Immune System During Pregnancy

Pregnancy is characterized by profound physiological, hormonal, and immunological adaptations that are essential for maintaining fetal development, being described in literature as a major immunosuppressive period. Since fetal development requires substantial maternal physiological and immunological adaptations to meet its nutritional and hormonal demands, pregnancy depends on a finely regulated balance between immune tolerance and immune protection. This higher immunological activity is particularly important during the first and third trimesters, when the maternal immune system is initially adapting to the presence of the fetus and later preparing the body for labor and delivery [8]. In this case, a decrease in immune activity can be observed throughout the second trimester of pregnancy, when the T lymphocyte production stimulated by the amount of infected monocytes rate in the maternal body reaches the lowest point, indicating a period of bigger immunosuppression and risk for contracting severe forms of the DENV infection caused disease, such as DHF, as observed in Figure 2.

During the first trimester of pregnancy, there is an increase in the population dynamics of T lymphocytes, once the pregnant body is initially recognizing the fetus as a potential immune threat. However, during the second trimester, there is a significant decrease in the immunological activities of the pregnant body, lowering the amount of T lymphocytes pregnant women's blood stream, indicating a major immunosuppressive period. Finally, during the third trimester, there is a gradual increase on the immunological defenses during pregnancy, raising the amount of T lymphocytes in the blood stream, preparing the maternal body for labor and delivery [8].

As a consequence, pregnant women may exhibit increased susceptibility to severe outcomes following dengue infection. The reduced immune activity observed particularly during the second trimester can facilitate viral replication and prolong the persistence of infected cells, resulting in higher viral loads and an increased risk of complications [9]. These gestational immune adaptations have motivated the development of mathematical models capable of incorporating pregnancy-specific immunological conditions into the analysis of dengue infection dynamics. By considering trimester-dependent variations in immune parameters, computational models can provide a more realistic representation of viral dynamics and help evaluate therapeutic interventions, such as antiviral treatments, in pregnant populations [9] [10].

III. Methods

The methodology adopted in this study consists of the development of a within-host mathematical models to describe the dynamics of DENV infection during pregnancy. First, a system of ordinary differential equations (ODEs) was formulated to represent the interactions among susceptible monocytes ($s(t)$), infected monocytes ($i(t)$), viral particles ($v(t)$), and T lymphocytes ($z(t)$). The parameters used to build the mathematical models were obtained from literature, and the initial conditions of all state variables were defined at the beginning of the simulation, with time set to $t = 0$, representing the initial stage of infection. Subsequently, a mathematical model that describes antiviral treatment was incorporated into the system of ODEs by modifying parameters associated with specific stages of the viral life cycle. The resulting system of equations provides a quantitative framework for investigating the progression of dengue infection and evaluating the effectiveness of the proposed therapeutic strategy. The model is numerically solved using the third-order Runge–Kutta method implemented in Python, allowing the simulation of viral and immune dynamics over time for non-pregnant women and pregnant women while accounting for their distinct biological and immunological characteristics. Finally, the temporal profiles of the model

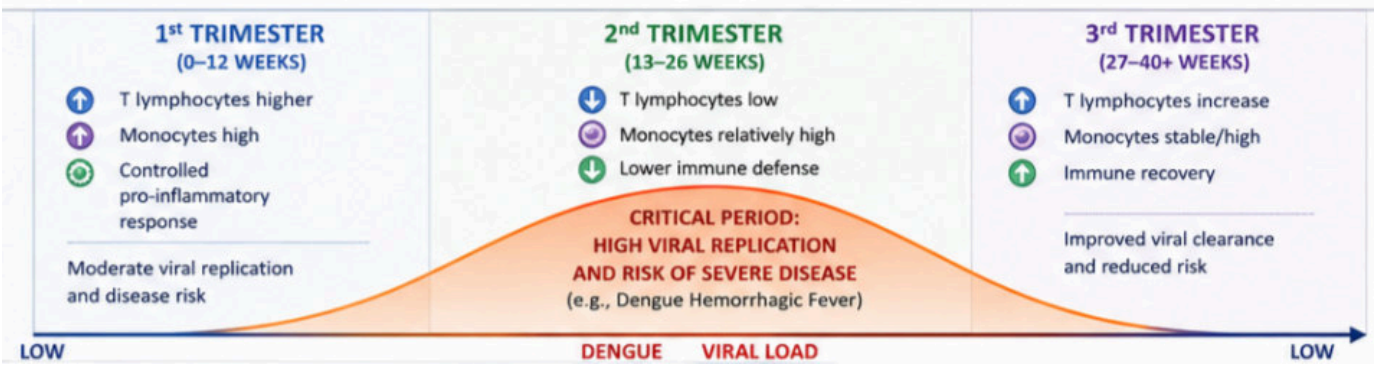


Fig. 2. Immunological dynamics during DENV infection throughout pregnancy.

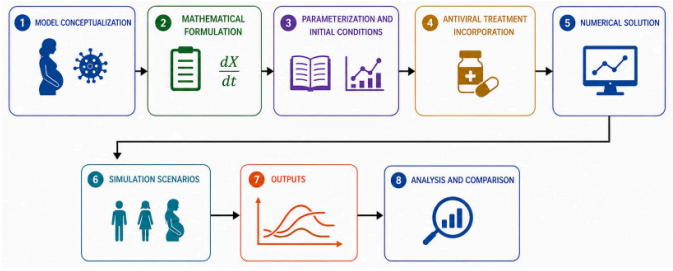


Fig. 3. Representation of the methodology used to obtain the results for the impact of DENV antiviral treatment during pregnancy.

variables were analyzed and compared to assess the impact of antiviral treatment under the investigated scenarios. The complete methodological workflow is summarized in Fig. 3.

A. Mathematical Modeling of Dengue Infection and Antiviral Treatment

Mathematical modeling has become an important tool for understanding the complex interactions between pathogens and the host immune system. In the context of dengue infection, within-host models describe the progression of the disease inside an infected individual by representing the interactions among viral particles, susceptible cells, infected cells, and immune cells through a system of equations formed by ordinary differential equations [11].

This within-host approach allows the investigation of biological mechanisms that are difficult to observe directly in experimental studies. By incorporating the main cellular populations involved in DENV infection, these models provide a quantitative framework for analyzing viral replication, immune activation, and disease progression over time. Furthermore, numerical resolution implemented in computational simulations enable the evaluation of this within-host dynamics according to different therapeutic strategies and physiological conditions without exposing patients to potential risks [12].

Among the immune populations commonly considered in dengue mathematical models are susceptible monocytes, infected monocytes, viral particles, and T lymphocytes. The interaction between these cellular populations and viral particles can be represented mathematically through ODEs that describe the rates of cell infection, viral production, immune activation, and cellular death. Such models make it possible to quantify how variations in immune parameters influence viral dynamics and disease severity. In addition, the incorporation of antiviral treatment into the model allows the investigation of its effects on viral load reduction and immune response enhancement [12]. In this work, the immune parameters of pregnancy required for these mathematical models are taken from literature and express the life and death dynamics of the immune cells described. The initial values, such as the values that represent the viral dynamics of dengue infection during pregnancy, are shown in Table I [10].

The temporal evolution of each cellular population is represented through differential equations that account for biological processes such as monocyte production and replication, viral infection of susceptible monocytes, viral replication, immune-mediated clearance of infected cells and viral particles, and the activation and proliferation of T lymphocytes. In addition, antiviral treatments are incorporated into the model through parameters that modify specific stages of the viral life cycle.

1) Mathematical Model for Susceptible Monocytes: The variable number of susceptible monocytes over the course of DENV infection (1) is determined by the monocyte production rate, minus the natural monocyte depletion rate, represented by $\alpha s(t)$, and minus the rate at which susceptible monocytes become infected through interactions with viral particles, represented by $\alpha s(t)v(t)$.

$$\frac{ds(t)}{dt} = \mu - \alpha s(t) - \alpha s(t)v(t) \tag{1}$$

With this equation, it becomes possible to predict the amount of monocytes susceptible to DENV infection in the maternal body's bloodstream. According to this rate, it is possible to numerically represent the initial amount for susceptible monocytes, and the monocyte recovery rate, meaning infected monocytes that are successfully neutralized by the maternal body's immune response coordinated by T lymphocytes, which can be used as an indicator of the

TABLE I
IMMUNOLOGICAL PARAMETERS

Symbols	Immunological Parameters		
	Meaning	***Initial Values***	***Pregnancy Values***
μ	Monocyte replication rate	80	120
α	Monocyte death rate	3	3
a	Viral infection rate	0.003	0.003
κ	Monocyte regeneration rate	20	20
γ	Virus release rate	0.8	0.8
ν	Infected monocyte clearance rate	0.001	0.001
β_1	Monocyte infection period	1/0.5	1/0.5
δ	T-lymphocyte lifespan	365	365
η	T-lymphocyte production rate	0,265	0,20
c	Stimulation rate of ν induced by infected monocytes	0.01	0.01
d	T-lymphocyte production stimulated by infected monocytes	250	1st tri. = 0.0179; 2nd tri. = 0.0168; 3rd tri. = 0.0321
s(t)	Susceptible monocytes	250	1st tri. = 490; 2nd tri. = 606; 3rd tri. = 908
i(t)	Infected monocytes	10	10
v(t)	Viral particles	165	165
z(t)	T lymphocytes	2000	1700

effectiveness of the proposed antiviral treatment.

2) Mathematical Model for Infected Monocytes: The dynamics of the infected monocyte population (2) is governed by the conversion of susceptible monocytes into infected monocytes through interactions with viral particles, represented by $\alpha s(t)v(t)$, and by the clearance of infected monocytes at a rate proportional to their infection period, which is represented by β_1 (3).

$$\frac{di(t)}{dt} = \alpha s(t)v(t) - \beta_1 i(t) - \nu i(t)z(t) \tag{2}$$

$$\beta_1 = \beta + \frac{\eta\nu}{\delta} \tag{3}$$

With these equations, it is possible to predict the amount of infected monocytes in the maternal bloodstream. From this rate it is possible to numerically estimate the viral particles production score, which flows from the resulting product of the monocyte infection-period rate and the infected monocyte population $\beta_1 i(t)$, counterbalanced by the elimination of infected monocytes through both natural cell death and T lymphocyte-mediated immune clearance, represented by $\nu i(t)z(t)$.

3) Mathematical Model for T Lymphocytes: The number of T lymphocytes over the period of DENV infection (4) is determined by the T-lymphocyte production rate $c_1 i(t)$, where the parameter c_1 is determined in (5), plus the amount of T lymphocytes produced by the stimulus contact with infected monocytes $di(t)z(t)$, minus the score for the natural death of T lymphocytes $\delta z(t)$.

$$\frac{dz(t)}{dt} = c_1 i(t) + di(t)z(t) - \delta z(t) \tag{4}$$

$$c_1 = c + \frac{d\eta}{\delta} \tag{5}$$

These equations represent the dynamics of T-lymphocyte number and production through the course of DENV infection. They make it possible to quantify both the natural production of T lymphocytes and the additional production stimulated by contact with infected monocytes. The resulting T-lymphocyte population number provides a quantitative measure of the adaptive immune response and can be used to investigate its association with disease severity.

4) Mathematical Model for Viral Particles: The number of viral particles present in the host during DENV infection (6) is determined by the viral production rate $\kappa i(t)$, minus the viral clearance rate $\gamma v(t)$, which accounts for both the natural degradation of viral particles and the elimination of infected monocytes, minus the loss of viral particles resulting from their entry into susceptible monocytes $\alpha s(t)v(t)$, as these particles are no longer part of the free viral population.

$$\frac{dv(t)}{dt} = \kappa i(t) - \gamma v(t) - \alpha s(t)v(t) \tag{6}$$

This equation makes it possible his equation makes it possible to determine the temporal evolution of the viral population by quantifying viral production, viral clearance, and the infection of susceptible monocytes. The resulting viral load provides a quantitative measure of infection progression and is the main score used to evaluate the effectiveness of the proposed antiviral treatment.

5) Mathematical Model for Antiviral Treatment: The mathematical model that represents the effect of the antiviral treatment (7) is based on the same equation used to model the viral particles' population, with the addition of the parameter ρ. Prior to treatment administration, ρ assumes a value of 1, corresponding to the untreated infection dynamics, keeping the equation's results untouched. Following antiviral administration, ρ is set to 0, reflecting the complete activation of the antiviral mechanism incorporated into the model.

$$\frac{dv(t)}{dt} = (1-\rho)\kappa i(t) - \gamma v(t) - \alpha s(t)v(t) \tag{7}$$

With this equation, it becomes possible to model the dynamics of DENV infection with the impact of antiviral treatment specifically considering the amount of viral particles present in the host's bloodstream. The resulting system of ODEs is shown in (8).

$$\begin{cases} \frac{ds(t)}{dt} = \mu - \alpha s(t) - \alpha s(t)v(t) \\ \frac{di(t)}{dt} = \alpha s(t)v(t) - \beta_1 i(t) - \nu i(t)z(t) \\ \frac{dz(t)}{dt} = c_1 i(t) + di(t)z(t) - \delta z(t) \\ \frac{dv(t)}{dt} = (1-\rho)\kappa i(t) - \gamma v(t) - \alpha s(t)v(t) \end{cases} \tag{8}$$

IV. Results

This section presents the computational results obtained from the within-host mathematical model of DENV infection in pregnant women undergoing antiviral treatment. The simulations were performed considering antiviral administration on the first, second, and third day of infection, allowing the evaluation of its effects on viral dynamics throughout the course of the disease, considering the immunological parameters associated with pregnancy, as well as the biological differences observed throughout the three trimesters of gestation.

When administrated in the first day of infection, the proposed antiviral treatment produced the most pronounced reduction in viral load among all evaluated scenarios, considering all three trimester of gestation. Although the viral population increased during the initial stage of infection, the treatment rapidly limited viral replication, resulting in an early decline in viral particles, as observed in Figure 4.

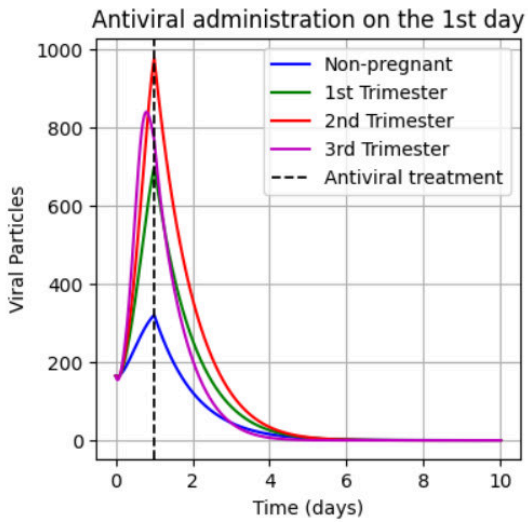


Fig. 4. Antiviral treatment administrated on the first day of infection.

In this graphic, the second trimester (red curve) exhibited the highest viral peak, followed by the third trimester (purple curve), the first trimester (green curve), and non-pregnant women (blue curve). Despite these differences, all groups showed a rapid decrease in viral load after treatment, reaching nearly complete viral clearance within approximately five days. These findings indicate that early antiviral intervention, preferably on the first day of infection, substantially improves infection control, particularly under the immunosuppression conditions observed during pregnancy.

When the antiviral treatment is initiated on the second day of infection, viral replication continues for a longer period before therapeutic intervention, resulting in higher viral peaks than those observed with first-day administration, as seen in Figure 5.

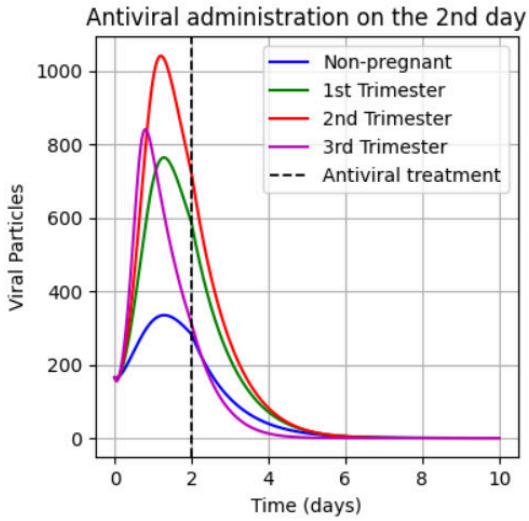


Fig. 5. Antiviral treatment administrated on the second day of infection.

In this graphic, the second trimester again presented the highest viral load, followed by the third and first trimesters, whereas non-pregnant women maintained the lowest viral concentrations throughout the infection. After treatment initiation, all curves exhibited a progressive decline, demonstrating the effectiveness of the proposed antiviral strategy. However, compared with treatment on the first day, viral clearance occurred more slowly, emphasizing the importance of minimizing delays in antiviral administration.

The delayed administration of the antiviral treatment presented on Figure 6 allowed viral replication to progress for a longer period, producing the highest viral burdens among the evaluated treatment schedules.

In the graphic presented in Figure 6, the second trimester remained the most susceptible condition, exhibiting the greatest viral peak, while the third trimester showed intermediate viral levels, followed by the first trimester and non-pregnant women. Although antiviral administration successfully reduced viral load in all groups, the decline occurred later than in the previous scenarios, reflecting the prolonged period of uncontrolled viral replication before treatment. Nevertheless, complete viral clearance was still achieved, demonstrating that the proposed antiviral therapy remained effective even when administered after the early phase of infection.

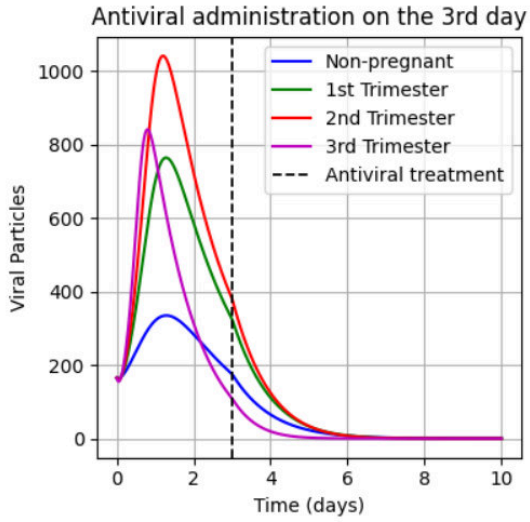


Fig. 6. Antiviral treatment administrated on the third day of infection.

Finally, all results show that the proposed antiviral treatment is most effective when administered early in the course of infection, ideally on the first day following DENV exposure, as observed on Figure 4. Under this condition, a more pronounced reduction in viral load is observed, resulting in faster viral clearance and a lower peak concentration of viral particles, as shown in Figures 5 and 6. These findings highlight the importance of early diagnosis and prompt therapeutic intervention in controlling dengue infection during pregnancy.

V. Conclusion and Discussion

The proposed antiviral strategy proved to be effective across all gestational trimesters and under all treatment initiation times evaluated in this study. In every simulated scenario, antiviral administration contributed to a reduction in viral load, indicating its potential as a therapeutic approach for mitigating the progression of DENV infection in pregnant women.

Among the three gestational stages analyzed, the second trimester presented the highest viral loads, suggesting a period of increased susceptibility to developing more severe forms of dengue, which are frequently associated with the occurrence of DHF. This behavior may be associated with the greater degree of immunosuppression observed during this phase of pregnancy, which favors viral replication and persistence within the host. However, even under the conditions considered for each trimester, the proposed antiviral treatment remained effective in reducing viral particle density and improving infection control.

These findings reinforce the importance of considering pregnancy-specific immunological characteristics when developing therapeutic strategies for dengue infection. Personalized antiviral approaches tailored to the immunological profile of each gestational stage may improve treatment outcomes and contribute to reducing the incidence of severe dengue manifestations, currently associated with the occurrence of DHF. Such reductions could help prevent pregnancy-related complications and minimize adverse effects on fetal development, thereby improving maternal and fetal health outcomes.

A potential future improvement of the proposed model is the incorporation of the DENV incubation period, which corresponds to the time period between viral infection and the onset of clinical symptoms. Including this stage would provide a more realistic representation of the disease progression, since antiviral treatment is typically initiated only after symptom onset and clinical diagnosis. By accounting for this delay in treatment administration, future simulations could more accurately assess the therapeutic effectiveness of antiviral interventions under realistic conditions and provide more reliable support for clinical decision-making.

VI. Acknowledgment

The authors would like to acknowledge the support provided to this study by the National Council for Scientific and Technological Development (CNPq) through an Undergraduate Research Scholarship.

References

[1] S. Sousa, M. Carneiro, A. Eiras, J. Bezerra, and D. Barbosa, “Factors associated with the occurrence of dengue epidemics in Brazil: a systematic review,” Rev Panam Salud Publica. Panama, vol. 45, August 2021.

[2] World Health Organization, ”Dengue and Severe Dengue,” who.int. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (accessed May 25, 2026).

[3] S. Wan, B. Wu-Hsieh, Y. Lin, W. Chen, Y. Huang, and R. Anderson, “The monocyte-macrophage-mast cell axis in dengue pathogenesis,” J Biomed Sci. United Kingdom, vol. 25, pp. 1–10, November 2018.

[4] A. Mathew, E. Townsley, and F. Ennis, “Elucidating the role of T cells in protection against and pathogenesis of dengue virus infections,” Future Microbiol. United States Of America, vol. 9, pp. 411–425, March 2014.

[5] A. Erlebacher, “Immunology of the maternal-fetal interface,” Annu Rev Immunol. United States of America, vol. 342, pp. 448–452, January 2013.

[6] T. Oriente, “Hemodynamic analysis of exosome propagation in the maternal-placental communication system during a Zika virus infection,” “Análise hemodinâmica da propagação de exossomos no sistema de comunicação materno-placentário durante uma infecção pelo vírus Zika,” M.S. Thesis, PPGEE. Signal Processing, IFPB, Brazil, 2021. [Online]. Available: https://repositorio.ifpb.edu.br/handle/177683/2189

[7] B. Troost and J. Smit, “Recent advances in antiviral drug development towards dengue virus,” Curr Opin Virol. Netherlands, vol. 43, pp. 9–21, August 2020.

[8] M. Kühnert, R. Strohmeier, M. Stegmüller, and E. Halberstadt, “Changes in lymphocyte subsets during normal pregnancy,” Eur J Obstet Gynecol Reprod Biol. Ireland, vol. 76, pp. 147–151, February 1998.

[9] G. Mor and Ingrid Cardenas, “The Immune System in Pregnancy: A Unique Complexity,” Am J Reprod Immunol. United States of America, vol. 63, pp. 425–433, March 2010.

[10] W. Rodrigues, K. Carvalho, J. Silva, M. Vilaça, E. Silva, and C. Regis, “Dengue virus analysis in different stages of pregnancy,” “Análise do vírus da dengue em diferentes estágios da gravidez,” CBEB. Brazil, pp.589–594, September 2024.

[11] K. Carvalho, M. Vilaça, W. Rodrigues, J. Silva, E. Silva, and C. Regis, “Dengue virus analysis in different stages of pregnancy,” “Análise do Impacto do Tratamento Antiviral no Modelo da Dinâmica Populacional da Infecção por Dengue,” CBEB. Brazil, pp. 575–580, September 2024.

[12] H. Clapham, V. Tricou, N. Chau, C. Simmons, and N. Ferguson, “Within-host viral dynamics of dengue serotype 1 infection,” J R Soc Interfac. England, vol. 11, July 2014.

Abordagem modular baseada em redes neurais artificiais para classificação do desempenho cognitivo em crianças com paralisia cerebral

Juan Alesson Silva de Almeida
Faculdade de Engenharia Elétrica e Biomédica
UFPA
Belém, Brasil
juanalessonn99@gmail.com

Adriana Rosa Garcez Castro
Faculdade de Engenharia Elétrica e Biomédica
UFPA
Belém, Brasil
adcastro@ufpa.br

***Resumo*—** A avaliação cognitiva de crianças com Paralisia Cerebral (PC) permanece um desafio clínico, uma vez que limitações motoras e sensoriais podem ocultar o real potencial cognitivo. Este trabalho propõe uma abordagem modular baseada em Redes Neurais Artificiais (RNAs) para classificar o desempenho cognitivo a partir de dados do aplicativo Desenvolve®, como suporte à decisão clínica. Fundamentado no Guia Portage, o Desenvolve® sistematiza a avaliação de dezenove habilidades cognitivas e gera perfis de desenvolvimento infantil. A base experimental incluiu 52 crianças com PC, entre dois e seis anos, do Centro de Referência UEAFTO/UEPA. O fluxo metodológico envolveu pré-processamento, codificação e divisão dos dados em treinamento, validação e teste, com uso de dados simulados exclusivamente no treinamento para maior robustez. A abordagem utilizou seis RNAs independentes, baseadas em Perceptron de Múltiplas Camadas, cada uma dedicada a um domínio cognitivo. Os modelos alcançaram acurácias entre 93,33% e 100% na base de teste, evidenciando elevada capacidade de generalização. Os resultados indicam que a estratégia modular é promissora como ferramenta complementar de apoio à decisão clínica, auxiliando na análise do desenvolvimento infantil sem substituir o julgamento terapêutico.

Palavras-chave—paralisia cerebral, aplicativo Desenvolve, redes neurais artificiais, desenvolvimento cognitivo, suporte à decisão clínica.

I. INTRODUÇÃO

A avaliação do desenvolvimento infantil constitui um processo complexo e multifacetado, que demanda instrumentos sensíveis e adequados para captar as nuances do progresso psicomotor e cognitivo, especialmente em crianças com condições clínicas específicas. Entre essas condições, destaca-se a Paralisia Cerebral (PC), definida como um grupo de desordens motoras permanentes e não progressivas, resultantes de lesões ou anomalias no desenvolvimento do cérebro durante os primeiros estágios da vida [1]. Tais lesões comprometem significativamente o controle motor e a coordenação, manifestando-se por meio de espasticidade, movimentos involuntários, fraqueza muscular e alterações no equilíbrio [2].

No ambiente clínico, a avaliação cognitiva de crianças com PC enfrenta desafios significativos, uma vez que a disfunção neuromotora pode interferir na auto exploração e na capacidade de interpretar estímulos ambientais [2]. Além disso, a maioria dos instrumentos tradicionais de avaliação exige respostas verbais ou motoras complexas que podem não refletir a real capacidade intelectual de pacientes com graves limitações na fala e na escrita manual [3].

Considerando este cenário observa-se um crescente interesse pelo uso de tecnologias computacionais no apoio à avaliação clínica, especialmente aquelas capazes de lidar com dados complexos e multidimensionais. Nesse contexto, insere-se o aplicativo Desenvolve® [4], uma ferramenta concebida para a avaliação e estimulação do desempenho cognitivo de crianças com deficiências neuromotoras. O protocolo adotado pelo *software* fundamenta-se em referenciais clássicos, como o Guia Portage, que sistematiza comportamentos e habilidades cognitivas e motoras em faixas etárias que abrangem do nascimento aos seis anos [2], [5]. O Desenvolve® é estruturado em 127 telas interativas que abrange dezenove habilidades cognitivas [4]. O aplicativo facilita a coleta e a organização de dados da avaliação cognitiva, gerando relatórios quantitativos que apresentam o número bruto de acertos e erros, além de percentis de rendimento categorizados por perfis etários típicos [4], [6].

Entretanto, em determinados contextos clínicos, a interpretação desses dados organizados pode tornar-se complexa. Dependendo da abordagem adotada, a leitura puramente descritiva dos resultados pode não captar adequadamente as nuances inerentes ao desempenho de pacientes com desenvolvimento atípico [7]. Quando as informações não são consideradas de forma integrada, a interpretação quantitativa dos acertos pode não refletir plenamente as capacidades ou dificuldades específicas do paciente, demandando consideração adicional do contexto da avaliação.

Além disso, o modelo analítico atual concentra-se predominantemente em habilidades isoladas. Essa fragmentação pode demandar uma interpretação manual intensa e exaustiva por parte do profissional a fim de obter uma visão global do paciente. Tal condição exige um maior esforço para cruzar múltiplas variáveis simultaneamente, o que possibilita diferentes interpretações diagnósticas a depender do nível de experiência do avaliador [2], [3]. Assim, a síntese e a interpretação integrada desses dados não são tarefas triviais na prática clínica, evidenciando a necessidade de abordagens mais sistematizadas e contextualizadas.

Diante deste contexto, este trabalho propõe uma abordagem modular considerando seis classificadores de desempenho baseados em Redes Neurais Artificiais Perceptron de Múltiplas camadas especializadas para cada domínio cognitivo do aplicativo Desenvolve®, visando auxiliar a interpretação dos resultados e fornecer suporte técnico à decisão clínica por meio da integração de variáveis contextuais.

II. PROTOCOLO E APLICATIVO DESENVOLVE®

O aplicativo Desenvolve® é um instrumento científico concebido para a avaliação e estimulação do desempenho cognitivo de crianças com deficiências neuro-motoras. Fruto de uma investigação científica iniciada em nível de mestrado e consolidada em tese de doutorado, a ferramenta evoluiu de protocolos originais em pranchas de papel para uma plataforma digital multiplataforma. Atualmente, o *software* vem sendo aplicado em contextos de pesquisa e atendimentos clínicos na Unidade de Ensino e Assistência de Fisioterapia e Terapia Ocupacional (UEAFTO) da Universidade do Estado do Pará (UEPA), tratando-se de uma tecnologia de uso acadêmico e institucional, não disponível comercialmente. O protocolo adotado pelo *software* fundamenta-se no Guia Portage, que sistematiza comportamentos e marcos do desenvolvimento infantil do nascimento aos seis anos [2], [5].

Diferente de avaliações qualitativas, o sistema permite a coleta sistemática de registros em 127 telas interativas que abrangem 19 habilidades cognitivas, entre elas: percepção de objetos do cotidiano, percepção de tamanho, percepção de sequência, noção de espaço, percepção auditiva, identificação de ações, identificação de formas e associação de iguais e diferentes. Para garantir a acessibilidade de pacientes com graves comprometimentos motores, o *software* utiliza um sistema de varredura (escaneamento) aliado a acionadores adaptados (toque, sopro ou movimento de cabeça), assegurando que o impedimento físico não mascare o real potencial cognitivo da criança [8], [9], [10].

As 19 habilidades cognitivas podem ser agrupadas em seis aspectos fundamentais (A1 a A6), respeitando a hierarquia de complexidade, conforme apresentado na Tabela I.

TABELA I
ASPECTOS COGNITIVOS FUNDAMENTAIS
CHECADOS PELO INSTRUMENTO DE AVALIAÇÃO

Aspectos cognitivos	**Habilidades**
A1 - Percepção visual	Se o indivíduo (criança) identifica e associa cores, formas, tamanho, objetos do cotidiano, letras e números, se o indivíduo (criança) identifica "igual", "diferente" etc...
A2 - Percepção auditiva	Se o indivíduo (criança) possui compreensão do comando de forma correta e se discrimina sons de animais, de objetos domésticos e relaciona a palavra ao objeto.
A3 - Percepção espacial	se o indivíduo (criança) identifica conceitos de "grande'", "pequeno", "em cima", "em baixo", "dentro", "fora", "compridos", "curtos", "primeiro", "segundo", "terceiro", "último".
A4 - Percepção e sequência	De ações, numérica, alfabética, temporal e tamanho
A5 - Percepção de esquema corporal	Se a criança identifica partes do corpo.
A6 - Outros	Se a criança completa figuras. Se a criança possui percepção de tempo. Se a criança possui percepção de escrita e leitura.

Fonte: [2].

A Fig. 1 ilustra a interface do *software* e apresenta exemplos de algumas telas interativas utilizadas para a avaliação das dezenove habilidades cognitivas no aplicativo Desenvolve®.

Após a aplicação das tarefas, o sistema fornece ao terapeuta relatórios quantitativos que detalham o número bruto de acertos e erros, além de gerar gráficos com percentis de rendimento categorizados por perfis etários típicos, permitindo uma visualização estruturada do perfil do repertório cognitivo da criança.

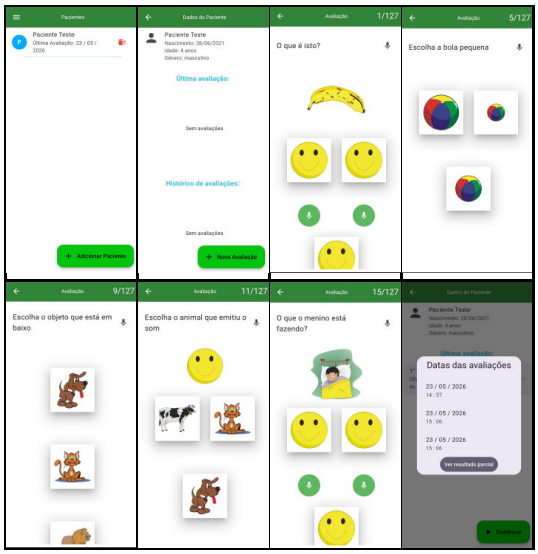


Fig. 1. Interface gráfica e exemplos de telas de avaliação do aplicativo Desenvolve®.

III. REDES PERCEPTRON DE MÚLTIPLAS CAMADAS

As Redes Neurais Artificiais (RNA) constituem modelos matemáticos computacionais inspirados na estrutura neural de organismos inteligentes, capazes de refinar o conhecimento adquirido a partir de estímulos do ambiente através de um processo iterativo de aprendizagem [11], [12]. A característica que torna o uso dessas redes relevante para aplicações na área da saúde é a sua capacidade de lidar com a não-linearidade e com múltiplas variáveis de forma simultânea [12], [13].

As Redes Perceptron de Múltiplas Camadas (MLP) representam uma evolução dos modelos neurais simples, sendo caracterizadas por uma arquitetura do tipo *feedforward* e completamente conectada [11], [12]. A estrutura básica de uma MLP é dividida em três componentes fundamentais: a camada de entrada, que recebe e distribui os sinais de estímulo; uma ou mais camadas ocultas (ou intermediárias), onde o processamento pesado ocorre; e a camada de saída, que fornece a resposta final do modelo para o problema proposto [11]. O diferencial técnico desta arquitetura reside na utilização de funções de ativação não-lineares nas camadas ocultas, o que permite à rede mapear relações complexas e separar padrões que não seriam distinguíveis por modelos lineares [11], [12].

O processo de aprendizagem das redes MLP segue o paradigma de aprendizado supervisionado, utilizando majoritariamente o algoritmo de retropropagação do erro (*Backpropagation*) [11]. O objetivo final do treinamento é minimizar o erro quadrático médio na saída da rede, ajustando os parâmetros internos para que a rede adquira a capacidade de generalização, ou seja, consiga produzir resultados corretos mesmo para dados que não foram apresentados durante a fase de treinamento da rede [11], [12].

IV. METODOLOGIA

A pesquisa foi conduzida em etapas sequenciais, organizadas em *pipeline*, garantindo o processamento das variáveis do protocolo do Guia Portage em ambiente digital. O fluxograma geral da metodologia proposta é apresentado na Fig. 2.

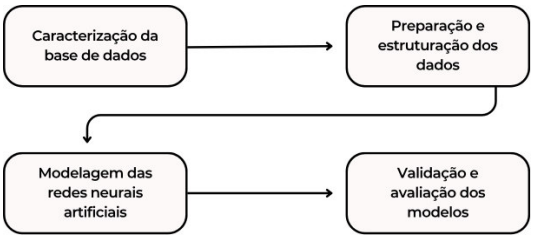


Fig. 2. Fluxograma geral da metodologia proposta.

Diferentemente de uma abordagem global com um único classificador neural, optou-se pelo desenvolvimento de classificadores dedicados para processar o desempenho de forma segmentada por domínios cognitivos (A1 a A6). Essa estratégia visa oferecer uma base estruturada para o suporte à decisão clínica, auxiliando o profissional na análise quantitativa das variáveis de cada aspecto, sem substituir o julgamento terapêutico.

A. Caracterização e Partição da Base de Dados

A base de dados é constituída por registros secundários de 52 pacientes com diagnóstico de Paralisia Cerebral, com idades entre 2 e 6 anos, coletados por terapeutas ocupacionais no Centro de Referência UEAFTO, da Universidade do Estado do Pará — UEPA (ALVES DE OLIVEIRA, 2010). Os estudos que fundamentaram a coleta original e o desenvolvimento do protocolo foram aprovados

pelo Comitê de Ética em Pesquisa do Núcleo de Medicina Tropical da Universidade Federal do Pará — UFPA, sob protocolo nº 049/2003 CEP/NMT.

Para o desenvolvimento dos modelos, os registros foram submetidos a um processo de partição *Holdout* [12]. O conjunto de treinamento foi inicialmente composto por 22 registros reais, sendo complementado com dados simulados para corrigir o desbalanceamento de classes e estabilizar o aprendizado das redes frente ao volume limitado da amostra clínica inicial. Ao final deste processo, o conjunto foi balanceado para um total de 99 registros [12]. Os conjuntos de validação (15 registros) e teste (15 registros) foram compostos integralmente por dados clínicos reais e inéditos, garantindo que a eficácia final do modelo fosse aferida apenas frente a evidências clínicas concretas.

Os registros simulados foram construídos de forma sistemática, utilizando as habilidades e o número de telas do aplicativo Desenvolve®, seguindo os marcos do desenvolvimento infantil [2]. A Tabela II sintetiza o perfil de desempenho por classe utilizado para a geração desses dados.

TABELA II
PERFIL DE DESEMPENHO POR CLASSE PARA GERAÇÃO DE DADOS

Classe	Característica dos registros
1	Alto desempenho (80% a 100% de acertos) nos marcos compatíveis com a idade.
2	Desempenho intermediário (50% a 80% de acertos) nos marcos compatíveis com a idade.
3	Baixo desempenho (0% a 50% de acertos) nos marcos compatíveis com a idade.

Fonte: Elaborado pelo autor.

A Tabela III sintetiza a distribuição de dados para treino, validação e teste dos modelos.

TABELA III
DISTRIBUIÇÃO QUANTITATIVA DOS DADOS POR CONJUNTO

Conjunto	Quantidade de registros	Natureza do dado
Treinamento	99	Real + Simulado
Validação	15	Real
Teste	15	Real

Fonte: Elaborado pelo autor.

B. Preparação e Estruturação dos Dados

Esta seção detalha o procedimento de transformação dos registros brutos do aplicativo Desenvolve® em vetores numéricos estruturados para o processamento computacional. O fluxo de preparação foi desenhado para converter o desempenho clínico em variáveis quantitativas, respeitando a organização por aspectos cognitivos e marcos de desenvolvimento previstos no protocolo.

A modelagem adotada define a entrada de cada modelo classificador com base na estrutura do protocolo de avaliação, que distribui as habilidades em quatro faixas etárias principais: dois a três, três a quatro, quatro a cinco e cinco a seis anos. O vetor de entrada utiliza codificação *one-hot* para identificação da faixa etária, entradas de desempenho em habilidades específicas e entradas agregadas por faixa (médias aritméticas de acertos), oferecendo à rede um contexto adicional sobre o rendimento em cada nível de desenvolvimento.

A variável de saída representa o resultado da análise integrada do desempenho cognitivo, convertida por arredondamento para umas das três categorias discretas apresentadas na Tabela II. Para a quantificação, as variáveis foram convertidas em uma escala de razão proporcional entre 0 e 1, calculada pela divisão entre o número de telas acertadas e o total de telas executadas. Adotou-se o valor -1 para identificar a ausência de execução de tarefas, diferenciando o desempenho nulo (erro) da ausência de registro (não realização), preservando a integridade da estrutura dos dados coletados.

C. Modelagem das Redes Neurais Artificiais

A pesquisa utilizou Redes Neurais Artificiais (RNAs) do tipo Perceptron de Múltiplas Camadas (MLP), implementadas em ambiente MATLAB® utilizando a função *feedforwardnet* do *Deep Learning Toolbox*. A estrutura de cada rede foi composta por três camadas:

- Camada de entrada: Constituída por um número de neurônios igual ao vetor de variáveis definido para cada aspecto. Para o tratamento de dados ausentes (codificados como -1), a topologia de entrada foi expandida através da concatenação de uma máscara binária que identifica a ocorrência de valores nulos. Dessa forma, o número real de neurônios na camada de entrada corresponde ao dobro da dimensionalidade original de cada vetor.
- Camada oculta: Camada intermediária responsável pelo processamento, submetida a variações experimentais para identificação da melhor topologia. Nesta camada, utilizou-se a função de ativação não-linear

Tangente Hiperbólica Sigmóide (*tansig*), que permite à rede mapear relações complexas entre as variáveis de desempenho.

- Camada de saída: Composta por um neurônio, responsável por gerar um valor numérico posteriormente convertido, por meio de arredondamento, para as classes 1, 2 e 3.

O treinamento das redes seguiu o paradigma supervisionado utilizando o algoritmo Levenberg-Marquardt [15], [16]. Como critério de interrupção do ajuste dos pesos sinápticos, adotou-se a técnica de *Early Stopping* (parada antecipada), fundamentada no monitoramento do erro no conjunto de validação. Para mitigar a instabilidade da inicialização aleatória dos pesos sinápticos, adotou-se uma estratégia de múltiplas execuções independentes (*seeds*). O processo experimental consistiu na variação sistemática do número de neurônios na camada oculta (10, 20, 40, 60, 80 e 100 neurônios), realizando-se 500 iterações de treinamento para cada configuração e selecionando-se o modelo com melhor acurácia no conjunto de validação.

D. Métricas de Desempenho

A acurácia na base de dados de validação foi adotada como a métrica principal para a definição da melhor arquitetura para cada um dos seis aspectos. Complementarmente, para aferição do desempenho dos modelos frente a dados reais e inéditos, foram obtidas as matrizes de confusão e calculadas as métricas de precisão, sensibilidade (*Recall*) e *F-score*.

V. RESULTADOS

A. Desempenho Global dos Modelos

A seleção das arquiteturas com melhores desempenhos fundamentou-se no critério de maior acurácia no conjunto de validação, priorizando, em caso de equivalência técnica, o modelo com maior poder de generalização no conjunto de teste e menor complexidade estrutural. A Tabela IV apresenta o desempenho global dos classificadores selecionados após diversos treinamentos, onde se variou os neurônios na camada escondida para inicializações de pesos diferentes (*seed)*.

TABELA IV
DESEMPENHO GLOBAL DOS CLASSIFICADORES POR ASPECTO COGNITIVO

Aspecto	**Neurônios**	***Seed***	**Treino**	**Validação**	**Teste**
A1	10	70	98,99%	100%	93,33 %
A2	20	1	97,98%	100%	100%
A3	40	292	98,99%	100%	93,33 %
A4	10	5	98,99%	100%	93,33 %
A5	10	1	100%	100%	100%
A6	20	7	100%	100%	93,33 %

Fonte: Elaborado pelo autor.

A interpretação sistêmica dos resultados, após os diversos treinamentos, permite agrupar os classificadores em dois perfis técnicos distintos. O primeiro grupo, composto pelos aspectos A2, A4 e A5, apresentou elevada estabilidade e generalização consistente durante os treinamentos, indicando padrões funcionais robustos e fronteiras clínicas claras no protocolo. Destaca-se o Aspecto A5 (Esquema Corporal), que preservou acerto integral (100%) em todas as fases experimentais devido à sua baixa dimensionalidade e regras determinísticas.

O segundo grupo, que integra os Aspectos A1, A3 e A6, revelou maior sensibilidade arquitetural durante os treinamentos. O sucesso da classificação foi estritamente dependente do ajuste fino da densidade da camada oculta para mitigar o fenômeno de sobreajuste (*overfitting*). Apesar da maior complexidade e heterogeneidade funcional, esses modelos mantiveram acurácias de teste superiores a 93%, validando a eficácia da modelagem modular mesmo em cenários de alta dimensionalidade.

C. Análise de desempenho dos modelos na base de teste

As Fig. 3 a Fig. 8 apresentam as matrizes de confusão para os classificadores considerando a base de teste e a Tabela V apresenta as métricas de Precisão, *F-score* e Sensibilidade para cada modelo.

Nos domínios A2 (Percepção Auditiva) e A5 (Esquema Corporal), observou-se acerto integral de todos os registros reais (diagonal principal). Para os demais aspectos (A1, A3, A4 e A6), os desvios de classificação ocorreram exclusivamente entre classes adjacentes. Nos Aspectos A1 e A3, registrou-se um único desvio onde a Classe 2 foi classificada como Classe 1. Já nos domínios A4 e A6, ocorreu um único erro de classificação na qual instâncias da Classe 1 foram atribuídas à Classe 2.

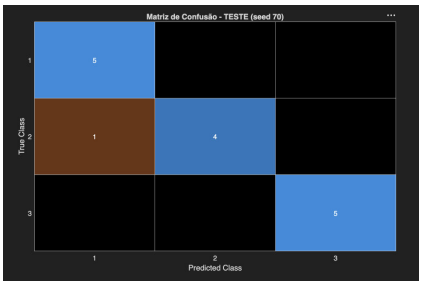


Fig. 3. Matriz de confusão para a base de teste – Aspecto A1.

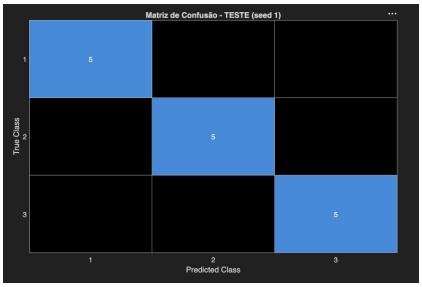


Fig. 4. Matriz de confusão para a base de teste – Aspecto A2.

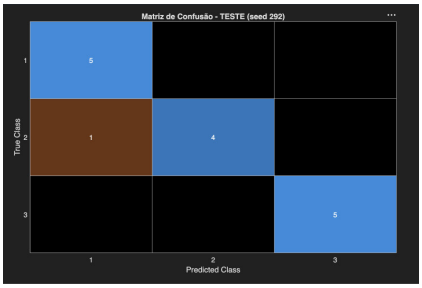


Fig. 5. Matriz de confusão para a base de teste – Aspecto A3.

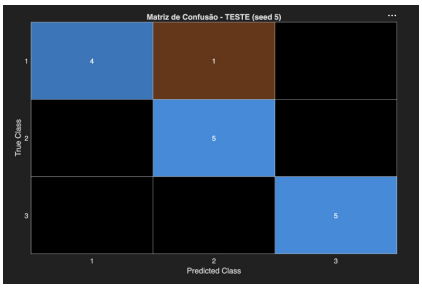


Fig. 6. Matriz de confusão para a base de teste – Aspecto A4.

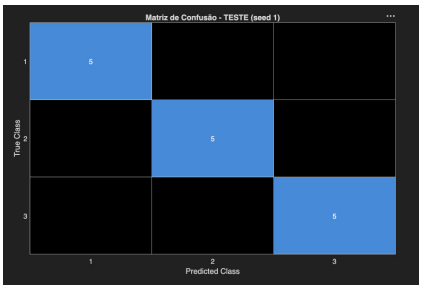


Fig. 7. Matriz de confusão para a base de teste – Aspecto A5.

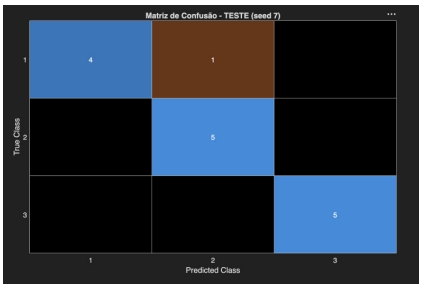


Fig. 8. Matriz de confusão para a base de teste – Aspecto A6.

TABELA V
MÉTRICAS GLOBAIS DE DESEMPENHO NA BASE DE TESTE

Aspecto	**Precisão**	**Sensibilidade**	***F-score***
A1	94,44%	93,33%	93,27%
A2	100%	100%	100%
A3	94,44%	93,33%	93,27%
A4	94,44%	93,33%	93,27%
A5	100%	100%	100%
A6	94,44%	93,33%	93,27%

Fonte: Elaborado pelo autor.

A análise da Tabela V indica que o modelo apresenta alto desempenho, com valores elevados de precisão, sensibilidade e *F-score*. A precisão elevada sugere baixa ocorrência de falsos positivos, indicando que o sistema raramente atribui incorretamente uma classe. Por outro lado, a sensibilidade ligeiramente inferior aponta para a presença de alguns falsos negativos, mostrando que nem todos os casos reais são identificados corretamente. O *F-score* confirma o equilíbrio entre as métricas, indicando ausência de viés significativo. Além disso, os erros ocorrem apenas entre classes adjacentes, reduzindo seu impacto clínico. Assim, mesmo quando há classificações incorretas, estas permanecem próximas do valor real. Apesar dos bons resultados, destaca-se a importância da análise conjunta com o terapeuta, uma vez que o sistema deve ser utilizado como ferramenta de apoio à decisão. A interpretação profissional é essencial para contextualizar os resultados e garantir uma avaliação mais precisa e individualizada.

VI. DISCUSSÃO

A. Contribuições da Metodologia

A principal contribuição deste trabalho reside na implementação de uma estratégia modular, que segmentou o desenvolvimento cognitivo em

classificadores especializados para cada aspecto do protocolo Desenvolve®. Em substituição a uma rede única e generalista, o emprego de MLPs dedicadas permitiu que cada modelo assimilasse as propriedades específicas de seu respectivo aspecto clínico, elevando a capacidade de adaptação do sistema às nuances de cada aspecto avaliado.

Os resultados evidenciam que as MLPs conseguiram identificar padrões funcionais a partir dos registros de avaliação com sólida capacidade de generalização frente a dados reais e inéditos. A metodologia comprova ser capaz de converter registros brutos em informações estruturadas, confirmando a utilidade de algoritmos de aprendizado de máquina em contextos de saúde infantil ao oferecer um suporte complementar que otimiza a carga de análise manual e proporciona uma visualização mais ágil e estruturada do perfil do paciente, sem substituir o julgamento clínico essencial ao cuidado.

B. Limitações do Estudo

A amostra clínica desta pesquisa é composta por registros de 52 pacientes reais, o que reflete as complexidades inerentes à coleta de dados em ambientes clínicos altamente especializados. Este volume amostral impõe desafios ao treinamento de modelos de Redes Neurais Artificiais, que demandam bases robustas para o mapeamento de padrões funcionais complexos. Em decorrência da limitação amostral inicial, o emprego de dados simulados constituiu uma estratégia necessária para equalizar as classes e assegurar a estabilidade da otimização dos pesos sinápticos durante o treinamento.

Contudo, esta técnica foi aplicada sob rigorosa salvaguarda metodológica, restringindo o uso de dados simulados exclusivamente ao conjunto de treinamento. Dessa forma, as etapas importantes de validação e teste permaneceram fundamentadas integralmente em registros de pacientes reais, garantindo a fidedignidade dos resultados. Adicionalmente, deve-se considerar que o sistema foi desenvolvido sobre a estrutura específica do protocolo Desenvolve®, o que vincula a eficácia atual dos classificadores à organização e aos marcos de desenvolvimento deste instrumento.

C. Implicações Clínicas

Na prática assistencial, a metodologia proposta oferece um suporte estratégico ao profissional de saúde, otimizando a organização do elevado volume de dados coletados durante os atendimentos com o aplicativo Desenvolve®. Ao processar de forma integrada o desempenho da criança nos diferentes aspectos cognitivos (A1 a A6), o sistema favorece a interpretação funcional das variáveis, proporcionando uma visão ágil, estruturada e sistêmica do perfil do paciente. É fundamental destacar que a tecnologia atua estritamente como um recurso de apoio e sistematização, fornecendo subsídios técnicos que auxiliam a análise do terapeuta sem substituir sua sensibilidade e *expertise* no processo avaliativo.

O potencial clínico desta abordagem estende-se ao acompanhamento evolutivo, permitindo que o profissional monitore o progresso terapêutico de forma sistemática e longitudinal ao longo do tempo. Em ambientes multiprofissionais, a ferramenta estabelece uma base técnica comum para a discussão de casos, apoiando a tomada de decisão por meio de dados organizados e fundamentados nos marcos do desenvolvimento infantil. Esse suporte tecnológico auxilia no direcionamento de intervenções personalizadas, garantindo que as metas terapêuticas estejam rigorosamente alinhadas às reais necessidades e capacidades demonstradas pela criança em cada etapa de seu amadurecimento.

D. Trabalhos Futuros

A continuidade desta pesquisa aponta para a necessidade de expansão da base clínica, buscando a inclusão de registros provenientes de novos centros de reabilitação. A realização de validações multicêntricas permitirá que os modelos representem uma diversidade ainda maior de perfis funcionais, ampliando a capacidade de generalização do sistema em diferentes contextos clínicos.

No âmbito tecnológico, os trabalhos futuros podem explorar a avaliação de outras arquiteturas neurais, como redes recorrentes ou modelos baseados em *ensembles*, visando o refinamento contínuo da precisão classificatória. Além disso, a integração da metodologia com sistemas de gestão clínica e a expansão para outros protocolos de avaliação funcional permitirão que a inteligência artificial ofereça um suporte ainda mais abrangente, possibilitando o acompanhamento da trajetória evolutiva do paciente de maneira mais sistemática e organizada.

VII. CONCLUSÃO

A aplicação do *pipeline* metodológico evidenciou a viabilidade técnica do uso de Redes Neurais Artificiais (MLP) para a classificação do desempenho cognitivo em crianças com Paralisia

Cerebral. Os classificadores demonstraram eficácia na estratificação dos níveis de rendimento cognitivo, consolidando a modelagem modular por aspectos como uma estratégia robusta para o suporte à decisão clínica.

Esta síntese reitera que a eficácia da Inteligência Artificial nesta pesquisa é reflexo da consistência estrutural do instrumento de avaliação, validando a convergência entre algoritmos de aprendizado de máquina e marcos de desenvolvimento clínico. Conclui-se que a metodologia modular proposta tem potencial como ferramenta complementar de apoio à decisão, podendo auxiliar o profissional na análise de variáveis multidimensionais e no acompanhamento do desenvolvimento infantil de crianças com PC, sem, no entanto, substituir o julgamento terapêutico.

REFERÊNCIAS

[1] I. Ferrareto e A. M. C. Souza (org.), Como tratamos a paralisia cerebral: reabilitação. São Paulo: Escritório Editorial, 1997.

[2] A. I. A. Oliveira (org.), Pesquisa em desenvolvimento infantil. Belém: Conhecimento & Ciência, 2008.

[3] I. L. C. Gil, "Aprendizagem e Inclusão escolar de criança com PC," em Anais do III Cong. Ibero-Americano de Inf. na Ed. Especial, Fortaleza, 2002.

[4] A. I. A. Oliveira, Desenvolve: software instrumentado. Belém: [s. n.], 2004. Registro no INPI n. 07703-6.

[5] S. Bluma et al., The Portage Guide to Early Education: revised edition. Portage: Cooperative Educational Service Agency 12, 1976.

[6] M. A. Pinheiro et al., "Software Desenvolve e Histórias Infantis," Informática na Educação: teoria & prática, vol. 15, no. 2, pp. 27-43, 2012.

[7] P. C. Pereira e M. A. R. Seruffo, "Inteligência Artificial e o Protocolo Desenvolve: novas perspectivas," Revista Pan-Amazônica de Saúde, 2024 (No prelo).

[8] M. B. Pelosi, "A comunicação alternativa e ampliada nas escolas do Rio de Janeiro," Dissertação de Mestrado, UERJ, Rio de Janeiro, 2000.

[9] A. I. A. Oliveira, "A contribuição da tecnologia no desenvolvimento cognitivo de crianças com paralisia cerebral," Dissertação de Mestrado, UEPA, Belém, 2004.

[10] N. Browning, "Recursos de Acessibilidade ao computador," em Atendimento Educacional Especializado: deficiência física, Brasília: SEESP/SEED/MEC, 2006, pp. 87-103.

[11] L. Fleck et al., "Redes Neurais Artificiais: princípios básicos," Revista Eletrônica Científica Inovação e Tecnologia, vol. 1, no. 13, pp. 47-57, 2016.

[12] S. Haykin, Redes neurais: princípios e prática, 2. ed. Porto Alegre: Bookman, 2001.

[13] A. P. Braga, T. B. Ludermir, e A. C. P. L. F. Carvalho, Redes Neurais Artificiais: teoria e aplicações. Rio de Janeiro: LTC, 2000.

[14] A. I. A. Oliveira, "Integrando tecnologias para leitura em crianças com paralisia cerebral na educação inclusiva," Tese de Doutorado, UFPA, Belém, 2010.

[15] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," Quart. Appl. Math., vol. 2, no. 2, pp. 164-168, 1944.

[16] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," J. Soc. Ind. Appl. Math., vol. 11, no. 2, pp. 431-441, 1963.

Avaliação de Estratégias de Controle Preditivo e por Alocação de Polos na Flexão do Joelho Induzida por FES em Ambiente OpenSim

Aparecida F. de Andrade
Departamento de Engenharia Elétrica
Universidade Federal do Ceará
Fortaleza/CE, Brasil
aparecidafalcao@alu.ufc.br

Fabrício G. Nogueira
Departamento de Engenharia Elétrica
Universidade Federal do Ceará
Fortaleza/CE, Brasil
fnogueira@dee.ufc.br

Bismark C. Torrico
Departamento de Engenharia Elétrica
Universidade Federal do Ceará
Fortaleza/CE, Brasil
bismark@dee.ufc.br

Bruno L. Faustino
Departamento de Engenharia Elétrica
Universidade Federal do Ceará
Fortaleza/CE, Brasil
brunoluizf1@alu.ufc.br

***Abstract*—Este trabalho apresenta uma análise comparativa entre os controladores RST e GPC aplicados ao controle da flexão do joelho via Estimulação Elétrica Funcional (FES). Utilizando o modelo musculoesquelético Leg6Dof9Musc no ambiente OpenSim integrado ao MATLAB, o sistema foi identificado por um modelo ARX. Os resultados indicam que ambas as estratégias são estáveis para o rastreamento de trajetórias. Contudo, o GPC demonstrou superioridade, apresentando convergência mais rápida, menor erro no regime transitório e resposta visualmente mais suave, consolidando-se como a técnica mais promissora para reabilitação motora.**

***Index Terms*—FES, GPC, RST, OpenSim, Flexão do Joelho, Modelagem Musculoesquelética, Identificação de Sistemas (ARX).**

I. Introduction

Diversas estratégias de controle vêm sendo propostas para aplicações em FES, incluindo controladores PID, adaptativos, robustos e preditivos. Entre essas abordagens, o Controle Preditivo Generalizado (Generalized Predictive Control – GPC) destaca-se por utilizar um modelo matemático da planta para prever o comportamento futuro do sistema e determinar, a cada instante de amostragem, a sequência ótima de ações de controle. Essa característica permite melhorar o rastreamento da trajetória desejada, reduzir erros de seguimento e minimizar os efeitos das não linearidades presentes no sistema neuromuscular [4]. Outra estratégia amplamente empregada em sistemas discretos é o controlador RST, baseado em alocação de polos por meio de polinômios, que apresenta simplicidade de implementação, estabilidade e bom desempenho no acompanhamento de referências [5]. Nos últimos anos, o avanço das ferramentas computacionais de modelagem musculoesquelética tem possibilitado o desenvolvimento e a validação de algoritmos de controle em ambientes virtuais antes de sua aplicação experimental. Entre essas ferramentas, o OpenSim destaca-se como uma plataforma de código aberto amplamente utilizada para simulação biomecânica, permitindo representar com elevado grau de realismo a dinâmica dos músculos, articulações e segmentos corporais. A utilização desse ambiente reduz custos experimentais, aumenta a reprodutibilidade dos estudos e possibilita avaliar diferentes estratégias de controle em condições controladas [6]. Além disso, revisões recentes destacam que a modelagem musculoesquelética associada a técnicas avançadas de controle constitui uma das principais tendências para o desenvolvimento de sistemas inteligentes de reabilitação motora personalizados [7]. Apesar dos avanços observados na literatura, ainda são escassos os estudos que realizam uma comparação direta entre estratégias de controle baseadas em GPC e RST aplicadas à Estimulação Elétrica Funcional para movimentos de flexão do joelho utilizando modelos musculoesqueléticos computacionais. Essa lacuna limita a compreensão das vantagens e limitações de cada abordagem no contexto da reabilitação motora assistida por FES. Dessa forma, este trabalho apresenta uma análise comparativa entre os controladores GPC e RST aplicados ao controle em malha fechada da flexão do joelho induzida por Estimulação Elétrica Funcional. Para isso, foi utilizado o modelo musculoesquelético Leg6Dof9Musc, implementado na plataforma OpenSim e integrado ao MATLAB para identificação do sistema, projeto dos controladores e realização das simulações computacionais. O desempenho das estratégias é avaliado por meio da capacidade de rastreamento da referência e das características da resposta dinâmica, contribuindo para o desenvolvimento de sistemas de reabilitação motora mais precisos, robustos e eficientes.

Reconhecimento ao CNPq, por meio do Edital nº 25/2020 - Projeto 445435/2020-8, pelo auxílio Financeiro no desenvolvimento da pesquisa.

II. Materiais e Métodos

A. Arquitetura do sistema proposto

A metodologia proposta foi desenvolvida em ambiente computacional integrando o software MATLAB® ao OpenSim®. A interface MATLAB/OpenSim utilizou o solver *ode45*, com tempo de amostragem de $0{,}02\,\mathrm{s}$ para aquisição e processamento dos dados. O MATLAB foi empregado para a identificação da planta, implementação dos algoritmos de controle e processamento dos resultados, enquanto o OpenSim foi utilizado para representar a dinâmica musculoesquelética do membro inferior humano e simular a resposta biomecânica à Estimulação Elétrica Funcional (FES).

Inicialmente, o modelo musculoesquelético foi excitado por sinais de entrada previamente definidos, permitindo a obtenção dos dados necessários para a identificação do sistema. A partir do modelo identificado, foram projetados os controladores RST e GPC, os quais foram avaliados em malha fechada por meio de simulações computacionais.

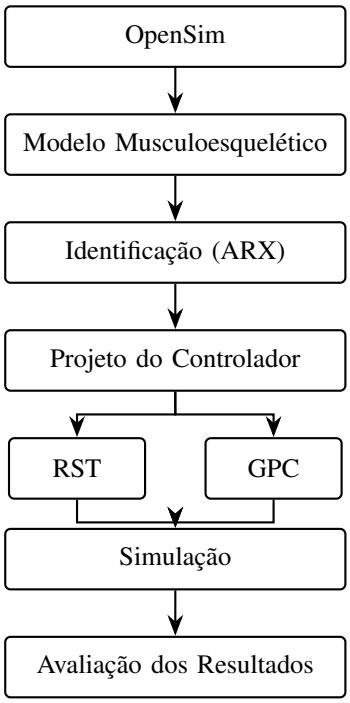


Fig. 1. Fluxograma geral da metodologia.

B. Modelo musculoesquelético

Foi utilizado o modelo não-linear Leg6Dof9Musc, disponibilizado na biblioteca do OpenSim [9]. Esse modelo representa o membro inferior humano por meio de seis graus de liberdade e nove músculos responsáveis pelos principais movimentos do quadril, joelho e tornozelo.

O estudo concentrou-se na flexão do joelho produzida pela estimulação elétrica aplicada ao músculo bíceps femoral, sendo o ângulo articular utilizado como variável de saída do sistema.

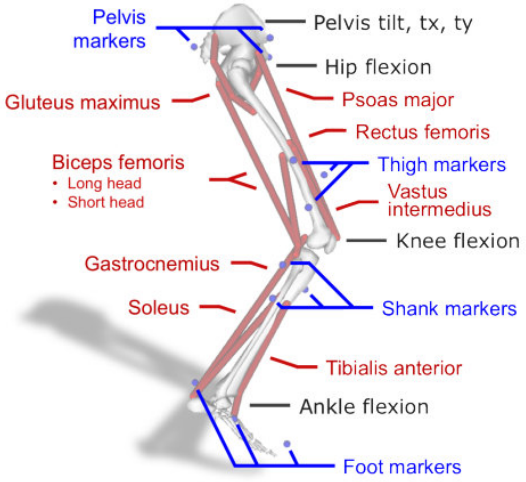


Fig. 2. Modelo musculoesquelético Leg6Dof9Musc utilizado nas simulações, representando o membro inferior com seis graus de liberdade e nove músculos

C. Identificação do sistema

A identificação da planta foi realizada em ambiente MATLAB utilizando dados obtidos a partir do OpenSim. Para garantir excitação persistente, empregou-se um sinal Pseudo Random Binary Sequence (PRBS) durante a coleta de dados.

A amplitude e a frequência do sinal PRBS foram definidas a partir de uma análise prévia da dinâmica do sistema. Inicialmente, foi aplicado um sinal degrau de amplitude 0,5 ao músculo bíceps femoral, sendo observada uma frequência dominante de aproximadamente $1{,}6\,\mathrm{Hz}$. Com base nessa resposta, o PRBS foi configurado com 7 células, amplitude entre 0,3 e 0,7, período de bit $T_b = 0{,}2\,\mathrm{s}$ e duração de 30 s. A adequação espectral do sinal foi verificada por meio da Transformada Rápida de Fourier (FFT), de forma a garantir conteúdo energético na faixa dinâmica de interesse do sistema. Ressalta-se que o modelo ARX obtido constitui uma aproximação linear válida em torno da região de operação excitada pelo sinal de identificação.

A estrutura do modelo identificado é baseada em modelos auto-regressivos com entradas exógenas (ARX).

$$A(q^{-1})y(k) = B(q^{-1})u(k) + e(k) \tag{1}$$

onde

- $u(k)$ representa o sinal de estimulação elétrica;
- $y(k)$ representa o ângulo de flexão do joelho;
- $e(k)$ representa o erro de modelagem.

Os polinômios do modelo são definidos por

$$A(q^{-1}) = 1 + a_1 q^{-1} + a_2 q^{-2} + \cdots + a_n q^{-n} \tag{2}$$

$$B(q^{-1}) = b_1 q^{-1} + b_2 q^{-2} + \cdots + b_m q^{-m} \tag{3}$$

Os parâmetros foram estimados pelo método dos mínimos quadrados não-recursivo, seguindo os critérios clássicos de identificação de sistemas [10]. O modelo apresentou um índice

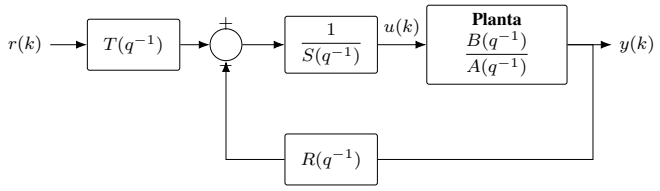


Fig. 3. Estrutura do controlador RST em tempo discreto.

de ajuste de aproximadamente 90,84%, indicando boa capacidade de representar a dinâmica da flexão do joelho em torno do ponto de operaçao dos dados coletados.

D. Projeto do controlador RST

Após a identificação da planta, foi projetado um controlador RST em tempo discreto pelo método de alocação de pólos. A lei de controle é dada por

$$R(q^{-1})u(k) = T(q^{-1})r(k) - S(q^{-1})y(k) \tag{4}$$

Os polinômios do controlador foram determinados por meio da solução da Equação Diofantina

$$A(q^{-1})R(q^{-1}) + B(q^{-1})S(q^{-1}) = A_c(q^{-1}) \tag{5}$$

em que $A_c(q^{-1})$ representa o polinômio característico escolhido de acordo com o comoportamento dinâmico desejado para o sistema de malha fechada.

A alocação dos polos foi definida a partir das especificações de desempenho desejadas para a resposta em malha fechada. A partir do polinômio característico resultante, obtiveram-se os polos

$$z_{1,2} = 0{,}9448 \pm j\,0{,}2225$$

e

$$z_{3,4} = 0{,}9065 \pm j\,0{,}0533.$$

O critério de projeto considerou a definição dos polos dominantes a partir dos requisitos de amortecimento e velocidade de resposta. Os polos adicionais foram escolhidos de modo a apresentar dinâmica mais rápida e menor influência sobre o comportamento dominante da resposta.

E. Projeto do controlador GPC

O controlador GPC foi desenvolvido com base no modelo CARIMA

$$A(q^{-1})y(k) = B(q^{-1})u(k-1) + \frac{e(k)}{\Delta} \tag{6}$$

em que

$$\Delta = 1 - q^{-1} \tag{7}$$

A função custo utilizada é

$$J = \sum_{j=N_1}^{N_2} (r - \hat{y})^2 + \lambda \sum_{j=1}^{N_u} (\Delta u)^2 \tag{8}$$

em que N_1 e N_2 representam, respectivamente, os limites inferior e superior do horizonte de predição; N_u corresponde ao horizonte de controle; $r(k+j)$ representa a trajetória de referência; $\hat{y}(k+j|k)$ corresponde à saída predita no instante $k+j$, com base nas informações disponíveis no instante k; $\Delta u(k+j-1)$ representa a variação incremental do sinal de controle; e λ é o fator de ponderação associado ao esforço de controle. Valores maiores de λ penalizam variações mais intensas da ação de controle, enquanto valores menores atribuem maior prioridade ao rastreamento da referência.

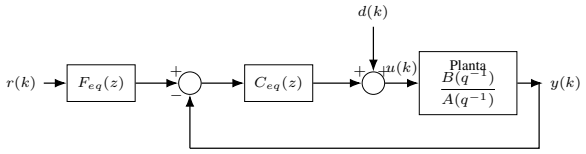


Fig. 4. Estrutura equivalente do controlador GPC em tempo discreto.

A estrutura interna do GPC pode ser representada por uma configuração equivalente de dois graus de liberdade (2DOF), na qual $C_{eq}(z)$ corresponde ao controlador equivalente e $F_{eq}(z)$ ao filtro equivalente associado ao caminho da referência.

Nessa representação, $F_{eq}(z)$ atua no pré-processamento da referência, enquanto $C_{eq}(z)$ reúne de forma compacta a dinâmica interna do controlador preditivo. Essa forma equivalente permite descrever a relação entre referência, ação de controle e saída da planta de maneira mais simples e adequada à apresentação do método.

$$M(z) = (K_e H - K_f)\,(zI - A + DH)^{-1}, \tag{9}$$

obtêm-se

$$C_{eq}(z) = -\left[\left([M(z)D - K_e]\,\Delta^{-1}\right)^{-1}(1 - M(z)B)\right]^{-1}, \tag{10}$$

e

$$F_{eq}(z) = -\left([M(z)D - K_e]\,\Delta^{-1}\right)^{-1} K_r \Delta^{-1}, \tag{11}$$

em que $\Delta = 1 - z^{-1}$.

F. Critérios de avaliação

O desempenho dos controladores foi avaliado a partir da capacidade de rastreamento da referência, da estabilidade da resposta, do comportamento durante o regime transitório e da presença de sobresinal. Também foi analisado qualitativamente o comportamento do sinal de controle ao longo da simulação.

A comparação entre as estratégias foi realizada considerando a resposta do ângulo de flexão do joelho em relação à trajetória de referência e as características transitórias observadas após as mudanças de referência.

III. Resultados e Discussões

A. Validação do modelo identificado

Após a etapa de identificação, o modelo ARX foi validado por meio da comparação entre sua resposta e a resposta do modelo musculoesquelético implementado no OpenSim. A Figura 5 apresenta o resultado dessa validação.

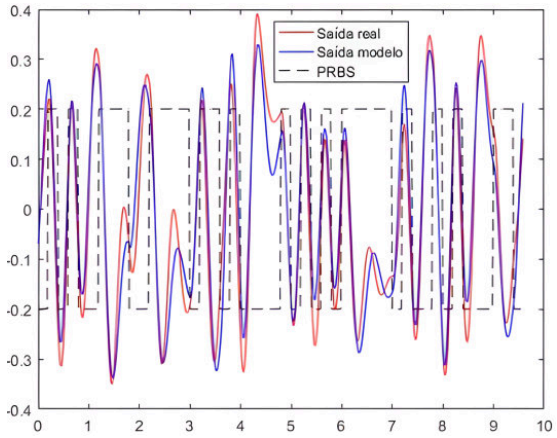


Fig. 5. Validação do modelo ARX

A partir da Figura 5 verifica-se que o modelo capturou a dinâmica do sistema em torno do ponto de operação dos dados coletados. O índice de ajuste (fit) obtido foi de aproximadamente 90,84%, indicando que o modelo ARX representa adequadamente a dinâmica da flexão do joelho para as condições consideradas neste estudo.

Esse resultado é fundamental para o desenvolvimento dos controladores, uma vez que tanto o controlador RST quanto o GPC dependem da qualidade do modelo identificado para produzir ações de controle compatíveis com a dinâmica da planta.

Além disso, a boa capacidade de representação do modelo reduz erros decorrentes da identificação e aumenta a confiabilidade das simulações realizadas em malha fechada.

B. Avaliação do controlador RST

Após a validação da planta identificada, foi implementado o controlador RST para o controle da flexão do joelho.

Conforme apresentado na Figura 6, o controlador RST foi capaz de conduzir o sistema até a referência desejada, mantendo estabilidade durante toda a simulação. A resposta apresenta comportamento compatível com a metodologia de alocação de polos empregada no projeto do controlador.

Observa-se, entretanto, que durante o regime transitório ocorre uma aproximação mais gradual da referência. Além disso, foi observada a ocorrência de sobresinal após as mudanças de referência, com duração aproximada de 1 s.

Mesmo apresentando esse comportamento, o controlador demonstrou ser uma alternativa viável para o controle da flexão do joelho induzida por FES, proporcionando resposta estável ao sistema.

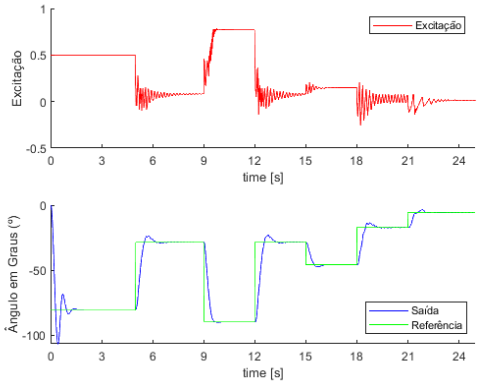


Fig. 6. Resposta temporal do controlador RST.

C. Avaliação do controlador GPC

Na sequência, foi avaliado o desempenho do Controle Preditivo Generalizado (GPC).

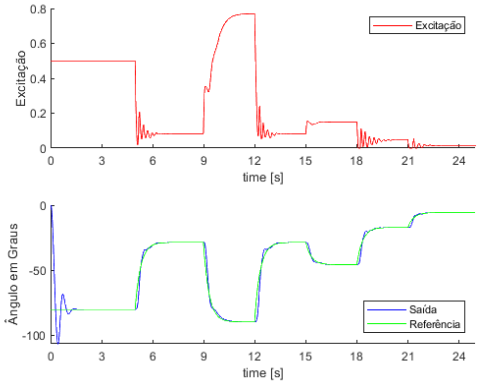


Fig. 7. Resposta temporal do controlador GPC.

A Figura 7 mostra que o controlador GPC também foi capaz de acompanhar a trajetória de referência de forma estável. A amplitude apresentada corresponde ao sinal de controle calculado pelo controlador nas condições simuladas e não representa, por si só, os limites fisiológicos máximos da estimulação. Os limites físicos da atuação dependem da forma de normalização e da implementação da excitação no modelo musculoesquelético.

Visualmente, observa-se uma aproximação mais rápida da referência durante o regime transitório, além de uma resposta mais suave ao longo da simulação. Esse comportamento é coerente com a estratégia preditiva empregada pelo controlador, que utiliza um modelo interno da planta para calcular, a cada instante de amostragem, a ação de controle que minimiza uma função de custo previamente definida.

Essa característica permite ao GPC antecipar o comportamento futuro do sistema, favorecendo um rastreamento mais eficiente da referência.

Para o controlador GPC, não foi observada a ocorrência de sobresinal nas condições simuladas, indicando comportamento

transitório mais suave em comparação ao controlador RST.

D. Comparação entre as estratégias de controle

A comparação direta entre as duas estratégias é apresentada na Figura 8.

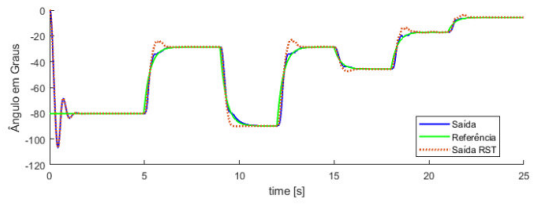


Fig. 8. Comparação entre as respostas obtidas pelos controladores RST e GPC.

A análise visual da Figura 10 evidencia que ambos os controladores foram capazes de estabilizar o sistema e realizar o rastreamento da referência proposta.

Entretanto, observa-se que o controlador GPC apresenta menor erro durante o regime transitório e uma resposta mais uniforme ao longo do movimento de flexão do joelho. Em contrapartida, o controlador RST apresenta uma convergência mais gradual até a referência, mantendo, porém, estabilidade durante toda a simulação.

No contexto da Estimulação Elétrica Funcional, uma resposta mais suave e com menor erro transitório pode contribuir para uma aplicação mais eficiente da estimulação elétrica, favorecendo o controle do movimento articular e reduzindo oscilações indesejadas.

Os resultados obtidos apresentam tendências consistentes com trabalhos correlatos da literatura. Em sistemas de estimulação elétrica em malha fechada, estratégias de controle preditivo têm demonstrado potencial para melhorar o acompanhamento de tarefas motoras [3]. De forma semelhante, nas condições avaliadas neste trabalho, o GPC apresentou aproximação mais rápida da trajetória de referência e comportamento transitório mais suave.

Por outro lado, a aplicação de controladores RST em sistemas de reabilitação motora baseada em FES já foi investigada para movimentos de quadril e joelho, demonstrando a viabilidade dessa estratégia para o controle de movimentos articulares [8]. No presente estudo, o RST também foi capaz de manter a estabilidade e realizar o rastreamento da referência, embora tenha apresentado convergência mais gradual e ocorrência de sobresinal durante o regime transitório.

Esses comportamentos também são coerentes com os fundamentos das duas estratégias. O GPC utiliza um modelo interno da planta para prever seu comportamento futuro e determinar as ações de controle a partir da minimização de uma função de custo [4], enquanto o RST baseia-se na alocação dos polos da malha fechada [5]. Dessa forma, as diferenças observadas nas simulações são compatíveis com as características teóricas descritas na literatura.

Diferentemente dos trabalhos citados, que analisam individualmente estratégias preditivas ou baseadas em alocação de polos, o presente estudo realiza uma comparação direta entre GPC e RST sob as mesmas condições de simulação e utilizando a mesma planta musculoesquelética, permitindo evidenciar de forma mais direta as diferenças de comportamento entre as duas estratégias.

Como limitação, os controladores foram projetados a partir de um modelo linear identificado em uma região específica de operação, enquanto a dinâmica neuromuscular associada à FES apresenta não linearidades e variações fisiológicas não explicitamente representadas, incluindo a fadiga muscular. Como trabalhos futuros, pretende-se avaliar os controladores diante dessas condições, incorporar modelos de fadiga e investigar estratégias adaptativas de estimulação, além de realizar validações com dados experimentais.

IV. Conclusão

Este trabalho realizou uma análise comparativa entre as estratégias de controle RST e GPC aplicadas à flexão do joelho induzida por Estimulação Elétrica Funcional (FES), utilizando um modelo musculoesquelético complexo no ambiente integrado MATLAB e OpenSim

Os principais pontos conclusivos são:

- **Validação do Modelo:** O processo de identificação do sistema por meio de um modelo ARX e sinais PRBS mostrou-se altamente eficaz, atingindo um índice de ajuste (*fit*) de aproximadamente 90,84%, o que confirma que o modelo foi capaz de representar com precisão a dinâmica biomecânica da flexão do joelho, fornecendo uma base sólida para o projeto dos controladores.
- **Desempenho dos Controladores:** Ambas as estratégias demonstraram ser alternativas viáveis e estáveis para o controle em malha fechada. O controlador RST garantiu estabilidade, embora tenha apresentado uma convergência mais gradual e um regime transitório menos responsivo.
- **Melhor comportamento transitório nas condições simuladas:** O Controle Preditivo Generalizado (GPC) destacou-se por apresentar um melhor rastreamento da trajetória de referência, com menor erro durante o regime transitório e uma resposta visualmente mais suave e uniforme ao longo da simulação. Essa comportamento é atribuída à sua capacidade inerente de utilizar um modelo interno para antecipar o comportamento futuro do sistema e otimizar as ações de controle em tempo real.
- **Impacto na Reabilitação:** No contexto da reabilitação motora, a resposta mais suave do GPC sugere uma aplicação de estímulos mais eficiente, o que pode contribuir para a redução de oscilações indesejadas e um melhor controle do movimento articular em pacientes com disfunções motoras.

Em suma, o estudo demonstrou que a utilização de ferramentas de simulação como o OpenSim permite a validação robusta de algoritmos de controle avançados, reduzindo riscos e custos experimentais. Os resultados posicionam o GPC como a técnica mais promissora para futuras implementações experimentais e clínicas de sistemas de FES personalizados e precisos.

AGRADECIMENTOS

Ao apoio do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Ceará (PPGEE/UFC), por meio do Grupo de Pesquisa em Automação, Controle e Robótica (GPAR), pelo suporte científico e infraestrutura disponibilizada para o desenvolvimento deste trabalho. Agradecemos também às agências de fomento Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) pelo apoio à pesquisa.

REFERENCES

[1] K. J. Popovic, Functional Electrical Stimulation in Rehabilitation, Neurorehabilitation and Neural Repair, vol. 28, n. 1, pp. 8–18, 2014.

[2] S. Ferrarin e A. Pedotti, "The Relationship Between Electrical Stimulus and Joint Torque: A Dynamic Model," IEEE Transactions on Rehabilitation Engineering, vol. 8, n. 2, pp. 230–243, 2000.

[3] X. Guo, P. Wang, X. Chen e Y. Hao, "Revolutionizing Motor Dysfunction Treatment: A Novel Closed-Loop Electrical Stimulator Guided by Multiple Motor Tasks with Predictive Control," Medical Engineering & Physics, vol. 129, 2024.

[4] D. W. Clarke, C. Mohtadi e P. S. Tuffs, "Generalized Predictive Control—Part I: The Basic Algorithm," Automatica, vol. 23, n. 2, pp. 137–148, 1987.

[5] K. J. Åström e B. Wittenmark, Computer-Controlled Systems: Theory and Design, 3ª ed. Prentice Hall, 1997.

[6] A. Seth et al., "OpenSim: Simulating Musculoskeletal Dynamics and Neuromuscular Control to Study Human and Animal Movement," PLoS Computational Biology, vol. 14, n. 7, 2018.

[7] M. Abdullah et al., "Multibody Dynamics-Based Musculoskeletal Modeling for Gait Analysis: A Systematic Review," Journal of NeuroEngineering and Rehabilitation, vol. 21, 2024.

[8] B. L. Faustino *et al.*, "Proposta de reabilitação motora baseada em simulação de um controlador rst com estimulação elétrica funcional para flexão de quadril e joelho," *Anais do Congresso Brasileiro de Engenharia Biomédica (CBA)*, 2022.

[9] A. Rajagopal *et al.*, "Full-body musculoskeletal model for muscle-driven simulation of human gait," *IEEE Transactions on Biomedical Engineering*, v. 63, n. 10, p. 2068–2079, 2016.

[10] L. Ljung, "System Identification: Theory for the User," *Prentice Hall*, 1999.

Predicting Growth Curves of Cell Lines with a Structured Segregated Model Using Unreal Engine

Matheus Morroni
Department of Science and Technology
Institute of Science and Technology
Unifesp
São José dos Campos - SP, Brazil
matheus.morroni@unifesp.br

André Zelanis
Department of Science and Technology
Functional Proteomics Laboratory
Unifesp
São José dos Campos - SP, Brazil
andre.zelanis@unifesp.br

Fabio Gava Aoki
Department of Science and Technology
Institute of Science and Technology
Unifesp
São José dos Campos - SP, Brazil
fgaoki@unifesp.br

***Abstract*—Developing a model for cell culture is a challenging task, because single cells have a certain response to the environment, and the same cell line in a culture with millions of cells has a different response for the same environment. Therefore, a robust model requires complicated mathematical equations with a large set of parameters and variables that, when not properly implemented, could mislead researchers to design a process that fails to achieve the simulated results. This study has the goal of developing a structured segregated model in Unreal Engine (UE) to predict the growth curves (GCs) of the cell lines CHO, Vero, and some of their sublines. The proposed computational model (CM) relies on the simulation of cell behavior to emerge the mathematical equations presented in the literature. Its parameters were set using data from experiments of different cell lines and the well-established knowledge about their characteristics, such as glucose consumption rate, maximum cell density in adherent or suspension culture, and lag phase estimation. The evaluation of the model was done by calculating the absolute relative error, $|\delta|$, the non-absolute relative error, δ, and the mean absolute error (MAE) at each time point. The doubling time and the average growth rate were also used for the analysis. The results showed that the model, with an absolute relative mean error of $|\bar{\delta}| = 0.77 \pm 0.127$, was able to simulate, on average, the early stages of culture growth (i.e., the lag phase and the beginning of exponential growth), $|\bar{\delta}|_{early} = 0.168 \pm 0.031$, better than the last stage of exponential growth, $|\bar{\delta}|_{last} = 2.426 \pm 0.393$. These results indicate that *in vitro* cells cease the division process before reaching the maximum cell density, avoiding the scarcity of nutrients in the medium and the stress increasing originated from spatial limitation. Although numerous biological, chemical, and physical factors may influence cell culture behavior, the evaluated model, even with its limited set of parameters, demonstrates potential to simulate in a short-term period the aspects of cell culture behavior growth. The use of UE enabled a more fluid development of the model itself, and —being based on C++— offered the flexibility and robustness needed to optimize the computational power required to cell culture simulation.**

***Index Terms*—computational model, structured segregated model, cell culture, growth curve, Unreal Engine**

I. Introduction

Mathematical and computational models of cell culture are an important feature for the biotechnological industry and academic research because they can reduce waste of resources when used properly. Having the ability to simulate the cell culture allows researchers to evaluate a myriad of scenarios economically and quickly. Most of the models are based on mathematical equations and describe the cell population statistically.

According to [1], there are four main types of cell culture models: unstructured unsegregated model, unstructured segregated model, structured unsegregated model, and structure segregated model.

The unstructured unsegregated model uses a more simplistic cell representation. This model simulates the cell population as a unique entity that refers to an average cell, so the model does not consider the intracellular variances between individuals and the different aspects of the cells such as age, shape, weight, and size [2]. In other words, the cell population is composed of cells with the same intracellular characteristics, shape, size and all of them are at the same age at the same time. Nonetheless, this model is largely used for the simplification of the mathematical equations, allowing a direct evaluation of the response of the culture to different conditions [3].

The unstructured segregated model enables a simulation of a heterogeneous cell population, where each cell has age, weight or size individually defined. These properties variances are ruled by statistical distribution models, such as the Population Balance Models (PBM) [4]. The PBM allows the population heterogeneity to be represented in a more concise approach. Compared with the unstructured unsegregated model it increases the mathematical complexity of the simulation, because the equations need to take into account the cell variation in the population and in the period of cultivation. Even with the increasing difficulty yielded by the model complexity for mathematical solution and parameters estimation, new mathematical tools and the advanced monitoring techniques of cellular processes made the application of this type of model possible outside the academic environment.

The structured unsegregated model treats the cell population as if all the cells present identical values for identification characteristics, such as age, size and weight. Although, the cells present a distinguishable intracellular structures. In other words, the cell population is a representation of the same cell (i.e., same age, same size, and same weight) in different stages that are precisely defined by the knowledge of the dynamic relation between the intracellular structures and processes and the boundary conditions of the culture medium. The depth of this knowledge allows the intracellular characteristics be more

granular simulating the principal biochemistry processes of the metabolism [5].

The structured segregated model reaches the most precise simulation of the cell culture. It represents a heterogeneous cell population, in which each cell is individually characterized, and presents individual intracellular structures [6]. This approach leads to a robust representation of the diversity of the physiological behavior of cells with their own structures and metabolic pathways interacting with the boundary conditions of their microenvironment [7]. Although this great accuracy in the complexity of the cell culture has a high price to pay because of the complexity of the mathematical formulation and the determination of the parameters. Because of these requirements the structured segregated model is rarely used.

To support the development of computational models capable of reproducing and visualizing cell culture dynamics, appropriate simulation platforms are required. UE is a graphic engine utilized for game development and creation of visual effects. It is based on C++, a programming language well-known for its robust library and its optimization capabilities. Its development environment is prepared to use the GPU of the computer; this enables the usage of the GPU to help the simulation and calculation of some aspects of computational modeling. In addition, the implementation of a Graphical User Interface (GUI) is trivial using this editor.

In this context, the goal of this work is to develop a CM implemented in UE that offers the level of granularity presented in a structured segregated model, simulating the cell behavior individually through logical conditions and simple mathematical equations (i.e., equations with few variables and parameters). Therefore, mitigating the issues of applying complicated mathematical formulation observed in the structured segregated models that restrict the usage of this type of modeling.

II. Methods

A. Collecting data from literature

To evaluate the robustness of the CM, studies meeting the following criteria were selected: research involving cell culture that included a graphical representation for the GC; at least one of the GC studied was of a non-carcinogenic cell line; the initial glucose concentration and the initial culture medium volume were reported; the culture container had to be reported; the type of cultivation had to be mentioned (e.g., batch, fed-batched suspension culture, or continuous suspension culture).

Access to raw cell count data was initially considered as an inclusion criterion. However, due to the difficulty of obtaining such data from published studies, this criterion was eventually discarded. Instead, to compare the model generated data with the experimental data, we extract the experimental observation from the graphs retrieved from the website of the publisher. To reduce the analytical problem raised by this approach, we applied an appropriate order-of-magnitude scaling in the simulation data to match the experimental data extracted.

To extract the experimental observations, we used a note app to draw lines from each time entry parallel to its opposed axis (i.e., the time entries were on the x-axis, and their lines were parallel to the y-axis that contained the cell density entries). Then, we estimated the cell density values using the intersection point between the GC and the lines, and the reference values presented on the y-axis. We sought the highest resolution image available to decrease the estimation error. However, we acknowledge that even with the highest resolution possible, the extracted experimental observation contained considerable uncertainty that directly affect the accuracy evaluation of the comparison measurements taken in this study.

The magnitude scaling was achieved by manipulating the simulation data values using the scale set in the experimental observation graphs. For example, Pan et al [9] (2017) presented the experimental observations of GC on a graph with the y-axis set in total viable cells, where each unit of the y-axis corresponded to 10^9 cells/bioreactor. Therefore, we took the simulation data given in cells $\cdot$ mL^{-1} and converted the value to total viable cells/bioreactor, multiplying the simulation data by the total volume given by the authors, then dividing the result by 10^9.

We also defined the time horizon represented by the CM. The model simulated the first 2–3 days of *in vitro* cell culture, corresponding to the biological time experienced by the *in silico* cell population (not to be confused with the computational runtime of the simulation).

B. The structured segregated behavioral computational model

The structured segregated behavioral classification of the CM means that the model simulated the behavior of each cell individually with its own characteristics and its own internal processes (i.e., the cell can be identified individually by its own age, volume, shape, and stage of cell cycle with its own internal clock). The flowchart of the model is shown in Fig. 1

The model was developed in UE. Therefore, rather than expending effort on the manual implementation of GUI libraries and GPU acceleration in a standard C++ development environment (e.g., Visual Studio Code or Xcode), we were able to focus exclusively on model development and on exploring the intended simulations using the CM.

C. Setting the parameters for the model

The parameters of interest focused on the lag phase of cells cultured under both suspension and adherent conditions, glucose consumption rate for basal metabolism and biomass synthesis for cell division, maximum cell density per volume or surface area, seeding cell density, simulation time horizon, glucose concentration in the medium, total surface area or volume of the container.

The values of the parameters listed above were defined using the following. The information from the article (e.g., initial volume or area of cultivation, period of cultivation, glucose concentration, seeding density), and the characteristics of the cell line used in the experiment (e.g., maximum cell density, lag phase period, glucose consumption rate) were extracted from the literature.

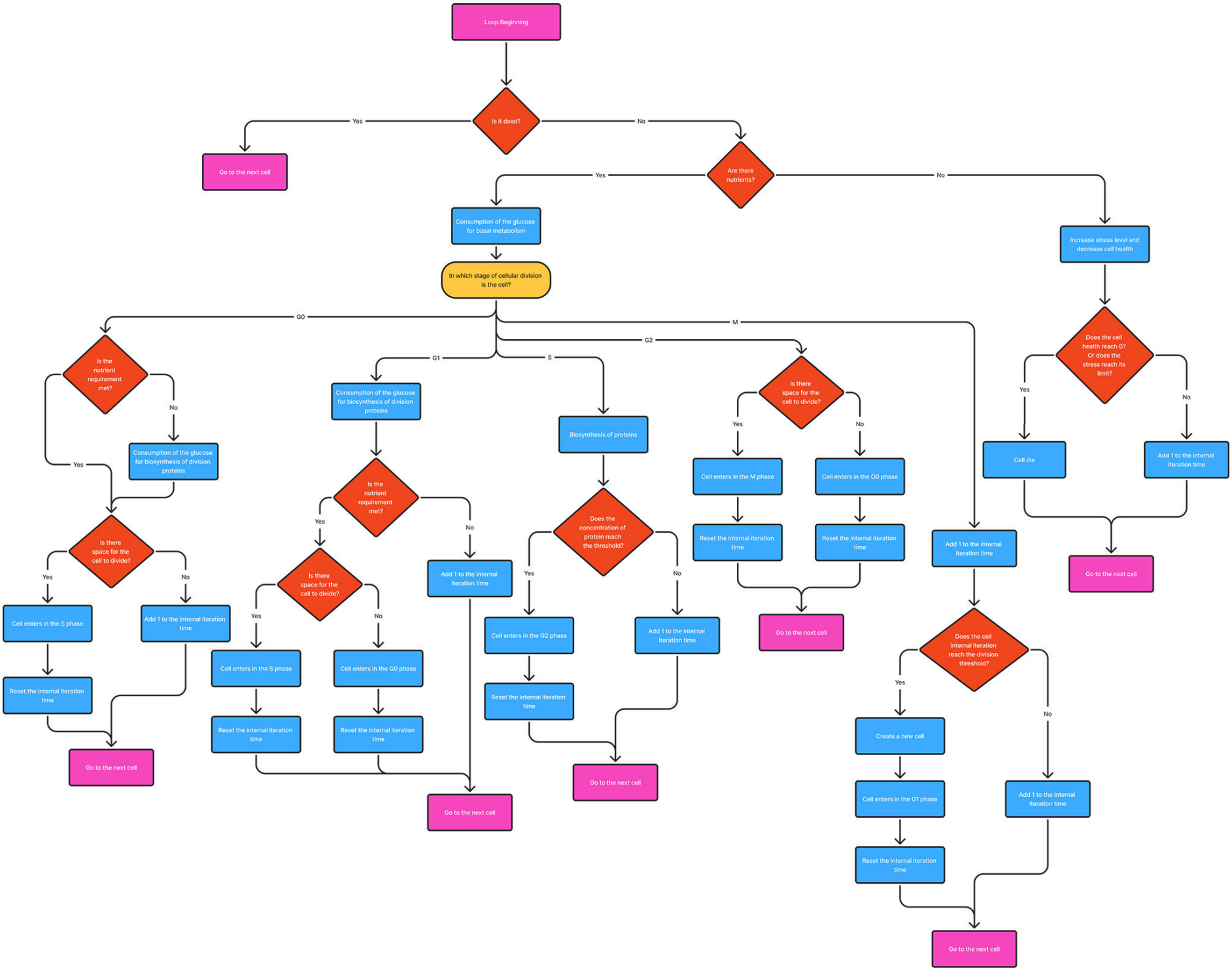


Fig. 1. Flowchart of the model cell behavior during the simulation. The internal iteration time refers to the internal clock of each cell.

To achieve heterogeneity in the cell population, cell characteristics in well-defined intervals were defined, and the CM chose pseudo-randomly a value in the range—using the Mersenne Twister algorithm that is natively supported in C++ and in the UE. This algorithm received the bound limits as input. The formulas applied to calculate the parameter bounds were:

$$\text{lower bound} = \text{literature datum} – 0.2 \times \text{literature datum} \quad (1)$$

$$\text{upper bound} = \text{literature datum} + 0.2 \times \text{literature datum} \quad (2)$$

The authors assumed that the characteristics followed a Gaussian distribution.

Because the glucose consumption rate per cell is extremely low (on the order of 10^{-15} mol · cell^{-1} · min^{-1}), both the glucose consumption parameter and the glucose concentration in the culture medium were scaled by a factor of 10^{13} as a numerical scaling strategy. This scaling was necessary because, given the numerical precision adopted by UE (approximately six decimal places), these values would otherwise be treated as zero during the computations. Although this limitation could be overcome by modifying the numerical representation or by implementing a scientific mathematics library with higher precision, such modifications were unnecessary, as the proposed scaling strategy preserved the intended cell behavior throughout the simulations.

The units adopted throughout the simulation were: mL for volume; mm^2 for surface area; cells · mL^{-1} for cell density; min for the simulation time step and lag phase duration; mol for glucose concentration; and mol · cell^{-1} · min^{-1} for the glucose consumption rate.

Based on the values of the parameters related to culture environment (e.g., volume of the container, seeding cells density) the number of simulations was selected to ensure that users could complete the simulation within a reasonable computational time. The environment parameters values for each simulation taken from the experiments are presented in Table I; the parameters values of cell characteristics taken from the literature are shown in Table II; and the general parameters applied in all simulations are listed in Table III.

D. Treating the simulation data

The simulation data was treated with biostatistics functions to calculate the growth rate of the cell population and the doubling time of each culture. Each experimental condition was simulated three times, the mean value was used as a scatter point, and the deviation of the simulation results was presented using the error bar. Data analysis was done in Python using pandas, numpy, matplotlib, and scipy libraries for the treatment, visualization, and manipulation of the experimental and simulation data. To maintain the same number of collected points in both groups, the authors took only the entries of the simulation data that met the time interval of the experimental entries observed (e.g., 24 h, 48 h, 72 h). Therefore, prevent statistical discrepancies related to the number of sample data

TABLE I
ENVIRONMENT PARAMETERS PER SIMULATION TAKEN FROM THE EXPERIMENTS. SPATIAL LIMITATION (SL); PERIOD OF SIMULATION (PS); INITIAL GLUCOSE CONCENTRATION (IGC); SEEDING CELLS DENSITY (SCD)

Ref	***Cell line***	***SL***	***PS***	***IGC***	***SCD***
[9]	CHO-BC®	5000 mL	4320	1.4×10^9	300
[10]	CHO-DG44	1400 mL	4320	3.88542×10^{10}	312.5
[11]	CHO-K1	2500 mm^2	2880	3.35×10^6	1250
[12]	Vero	7500 mm^2	4320	2.5×10^5	100
[13]	Vero (suspension)	5 mL	4320	1.25×10^6	245
[13]	Vero (adherent)	2500 mm^2	4320	9.75×10^8	250
[14]	CHO GS46	1200 mL	2880	3.73×10^8	300

TABLE II
CELL CHARACTERISTICS PARAMETERS PER SIMULATION TAKEN FROM THE LITERATURE. LAG PHASE DURATION RANGE (LDR); MAXIMUM CELL DENSITY (MCD); GLUCOSE CONSUMPTION RATE RANGE FOR BASAL METABOLISM (GCBM); GLUCOSE CONSUMPTION RATE RANGE FOR CELL DIVISION (GCCD)

Ref	***Cell line***	***LDR***	***MCD***	***GCBM***	***GCCD***
[9]	CHO-BC®	[576, 864]	1000	[0.02325, 0.034875]	[0.00175, 0.002625]
[10]	CHO-DG44	[864, 1296]	1220	[0.0186, 0.0279]	[0.00175, 0.002625]
[11]	CHO-K1	[864, 1296]	20	[0.024805, 0.037207]	[0.001867, 0.0028]
[12]	Vero	[1152, 1728]	3.8	[0.029265, 0.043898]	[0.002373, 0.003559]
[13]	Vero (suspension)	[1152, 1728]	2000	[0.029265, 0.043898]	[0.002373, 0.003559]
[13]	Vero (adherent)	[1728, 2592]	3250	[0.029265, 0.043898]	[0.002373, 0.003559]
[14]	CHO GS46	[576, 864]	13000	[0.0248, 0.037201]	[0.001867, 0.0028]

TABLE III
FIXED PARAMETERS PER SIMULATION. (*) PARAMETER SET FOR EACH CELL INDIVIDUALLY. (**) PARAMETERS ASSOCIATED TO THE DIFFERENTIAL EQUATION PROPOSED BY BISWAS AND BRENNER [8] (2024)

Parameter	***Value***
Cell health range *	[80, 100]
Senescent death rate range *	[0.4, 0.8]
Glucose concentration range required for cellular division *	[0.5, 1.5]
G2 phase duration range [min] *	[100, 200]
Mitotic phase duration range [min] *	[50, 60]
Acquired stress range per iteration *	[0.8, 1.2]
Stress limit range *	[80, 120]
Protein α threshold mean rate * ** (μ_C)	0.012
Restoring force of Ornstein-Uhlenbeck Stochastic Process (γ) * **	0.01
Initial concentration range of protein α (N) * **	[0.0, 1.0]
Initial protein α threshold mean range (C) * **	[0.0, 100.0]
Medium renovation cycle [min]	0
Glucose concentration in medium renovation [mol]	0

of each group. Based on the sample data, the non-absolute relative error, $\delta(t)$, was determined using the following formula:

$$\delta(t) = (y_s(t) - y_e(t))/y_e(t) \tag{3}$$

For the absolute relative error, $|\delta(t)|$, the formula applied was:

$$|\delta(t)| = |y_s(t) - y_e(t)|/y_e(t) \tag{4}$$

The MAE was calculated using the following formula:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_e(t) - y_s(t)| \tag{5}$$

where $y_s(t)$ is the simulation value at time t, $y_e(t)$, the experimental value at the same time, and n is the number of time steps. For the estimation of the doubling time of each GC, the formula applied was as follows.

$$T_d = (\Delta t \times log(2))/(log(X_{max}/X_0)) \tag{6}$$

where T_d is the doubling time in hours, Δt is the incubation time in hours, X_{max} is the maximum value of cell density reached in this period, and X_0, the initial cell density (i.e., the seeding density), both densities in cells · mL^{-1} or cells · mm^{-2}. In order to calculate the mean growth rate, the following formula was used:

$$\mu = log(X_T/X_0)/\Delta t \tag{7}$$

where μ is the mean growth rate of the culture in cells · ml^{-1} · h^{-1}, X_T is the cell density at the last collected point in cells · mL^{-1}, X_0, the seeding density in cells · mL^{-1} and Δt is the incubation time in hours.

The results were expressed in mean of the simulation values ± standard error of the mean (SEM).

III. Results and Discussion

Our analysis focused on the initial 2-3 days of cultivation in order to avoid misinterpretation of the process and the impact of renewing the culture medium on the GC, ending the simulation before the medium change in the experiments. Another reason for this time interval was to be able to simulate different cell lines and different experimental settings to demonstrate that the CM did not overfit a unique cell line under specific conditions. The total time required to simulate each replicate was 12 days, approximately. Due to the time cost of running all simulations, each experimental condition was simulated only three times. This decision prioritized the evaluation of all experimental conditions rather than increasing the number of replicates per condition. Even because without raw data, we could not calculate the SEM of the experimental observations, then the number of replicates does not have the expected statistical weight for the comparison, contributing to the decision to limit the number of replicates to the minimum required to represent the results as mean value ± SEM. In addition, the evaluation of the MAE indicated that the addition of more data sets would not change expressively the prediction performance of the CM in all experimental conditions evaluated. The Table IV shows the MAE with the standard deviation, σ, of all possible combinations of the data sets.

TABLE IV
MAE of all possible combinations of the data sets. The units correspond to those used in the GC graphs

Ref	***Cell line***	***All data sets***	***Data sets 1 and 2***	***Data sets 1 and 3***	***Data sets 2 and 3***	***Mag***
[9]	CHO-BC®	9.209 ± 2.09	10.356± 0.155	8.583 ± 2.478	8.687 ± 2.633	$\times 10^9$
[10]	CHO-DG44	0.569 ± 0.036	0.577 ± 0.04	0.556 ± 0.008	0.574 ± 0.047	$\times 10^6$
[11]	CHO-K1	0.998 ± 0.251	0.943 ± 0.325	1.131 ± 0.04	0.919 ± 0.285	$\times 10^5$
[12]	Vero	0.501 ± 0.039	0.523 ± 0.001	0.489 ± 0.047	0.49 ± 0.048	$\times 10^5$
[13]	Vero (suspension)	1.687 ± 0.194	1.637 ± 0.246	1.799 ± 0.018	1.625 ± 0.299	$\times 10^6$
[13]	Vero (adherent)	0.05 ± 0.013	0.048 ± 0.017	0.56 ± 0.003	0.046 ± 0.015	$\times 10^6$
[14]	CHO GS46	0.385 ± 0.059	0.416 ± 0.03	0.364 ± 0.052	0.38 ± 0.082	$\times 10^9$

For the purpose of this study, the model is limited to simulate a certain group of characteristics large enough to allow heterogeneity in the *in silico* population. This decision enabled us to simulate different cell lines without the depth of knowledge necessary to set all the model parameters presented for each line (the fixed parameters were listed in Table III). The authors believe that heterogeneity reduces the possibility of overfitting; therefore, it was fundamental that the limited group of attributes could represent some cell line differences.

The difference of the proposed CM is the approach to emerging the complexity of the mathematical formulation based on the individual cell behavioral response. This enables it to simulate the cell culture as a structured segregated model without the application of complicated mathematical equations. To accomplish this goal, our work was to simulate *in silico* cell heterogeneity throughout the distribution of biological characteristics of the cell line—lag phase period and nutrient consumption rates—and the stochastic phenomena related to cell division, mainly the biosynthesis of proteins.

For the biological characteristics, we used equations 1 and 2 to obtain the intervals of values that were input for the Mersenne Twister algorithm. The Empirical Rule establishes that approximately 68% of the data is in the interval of the first standard deviation, σ_1. So using data from the literature, which is frequently given as a mean value, and getting $\sigma_1 = 0.2\times$literature data, it was possible to generate a representative interval for the characteristics (shown in Table II), discarding potential extreme behavior that could create abnormalities in the simulation data.

To achieve the stochastic characteristic of proteins biosynthesis for cell division, the authors used the differential equation proposed by Biswas and Brenner [8] (2024). This approach treats the metabolic process as an Ornstein-Uhlenbeck process (i.e., a stochastic process of threshold crossing event), allowing

the model to simulate the progress of the cell cycle in a more natural and less deterministic way.

The similarity between the simulation results and the experimental data (i.e., the prediction capacity of the model) was evaluated by calculating the MAE, and the absolute and non-absolute relative errors. Furthermore, to hypothesize about the impact of CM fixed parameters in the simulated GC, we calculated the growth rate and doubling time of the cultures (Table V).

TABLE V
COMPARISON OF GROWTH RATE AND DOUBLING TIME. EXPERIMENTAL (e) AND SIMULATION (s)

Ref	***Cell line***	μ_e	μ_s	T_d^e	T_d^s
[9]	CHO-BC®	0.032	0.0479 $\pm$0.0022	21.6742	14.4566 $\pm$0.6329
[10]	CHO-DG44	0.0289	0.0366 $\pm$0.0004	24.0	18.9307 $\pm$0.2168
[11]	CHO-K1	0.0232	0.0336 $\pm$0.0026	29.9239	20.6101 $\pm$1.4745
[12]	Vero	0.0104	0.0236 $\pm$0.0005	66.8584	29.3789 $\pm$0.5844
[13]	Vero (suspension)	0.0194	0.0469 $\pm$0.0015	35.7383	14.7883 $\pm$0.4543
[13]	Vero (adherent)	0.0174	0.0335 $\pm$0.0025	39.8372	20.6636 $\pm$1.4314
[14]	CHO GS46	0.0261	0.0415 $\pm$0.0015	26.5581	16.6902 $\pm$0.5799

The mean absolute relative error of the CM, which is $|\bar{\delta}| = 0.77 \pm 0.127$, shows that, for the entire simulated period of all conditions, the model had a tendency to overestimate the experimental data by approximately 77%. It is important to highlight that because of the different units of the collected data, we decided not to use the average MAE (i.e., the mean value of the MAE calculated for each experimental condition) since it retains the unit of the analyzed data. On first look, we could conclude that the CM did not achieve a good prediction of culture growth. However, looking for the general stages presented in a culture growth curve individually (e.g., lag phase, log phase), we found that at early stages —lag phase and the first portion of the log phase (around the first 36 hours)— the CM predicted, on average, the culture growth with an error estimation of the experimental observations around 17%, $|\bar{\delta}|_{early} = 0.168 \pm 0.031$ (as shown in Fig. 2). Using the non-absolute relative error, we found that in some conditions the model underestimated the culture growth, and in others, it overestimated the growth behavior of the cells, as can be seen in Fig. 3. How to improve the prediction in this scenario of mix estimations will be discussed later.

Now, looking at the last stages —the second portion of the log phase— of the simulations, the CM had a poor prediction of the GCs, it overestimated the experimental data, on average, by approximately 242.6%, $|\bar{\delta}|_{last} = 2.426 \pm 0.393$. In other words, *in silico* cell cultures were almost 2.5-fold more dense than *in vitro* cell cultures. We hypothesize that cells cease division prior to reaching their maximum density, halting population growth. Possible explanations for this may involve reduced nutrients in the medium and spatial

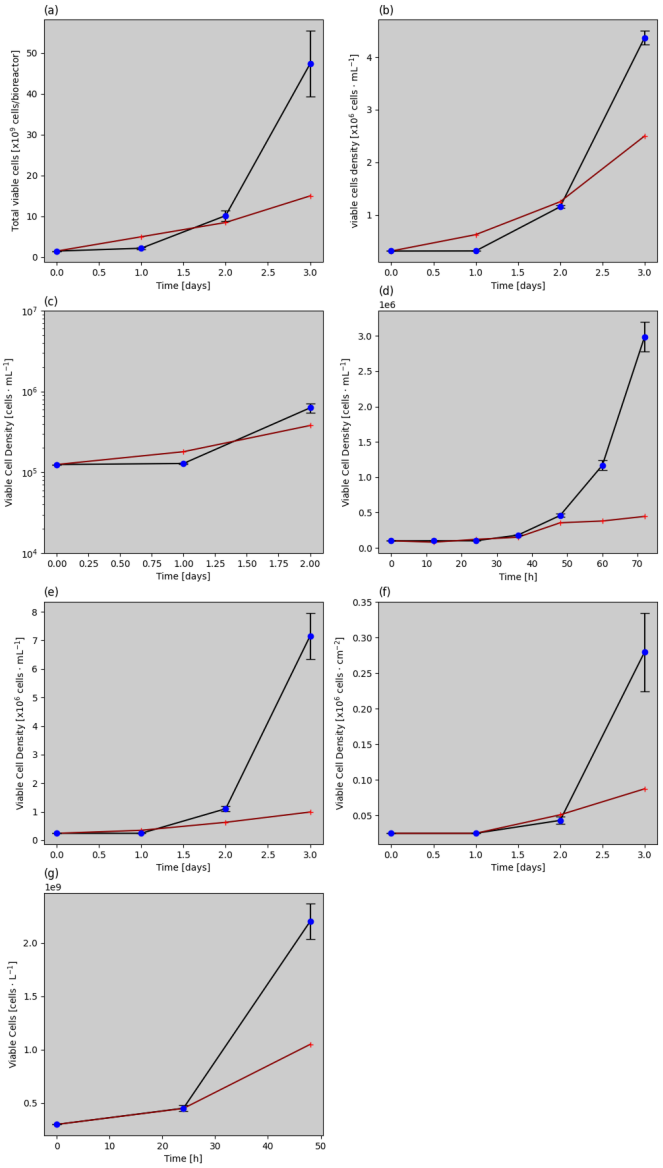


Fig. 2. Comparison of the growth curves of the simulation data (black line with blue dots) and the experimental data (red line with red crosses). (a) GC of the CHO-BC® cells [9]; (b) GC of the CHO-DG44 cells [10]; (c) GC of the CHO-K1 cells [11]; (d) GC of the Vero ATCC-CCL-81 cells [12]; (e) and (f) GC of the Vero ATCC-CCL-81 cells in suspension and adherent culture, respectively, [13]; (g) GC of the CHO-GS46 cells [14].

limitation, parameters included in this work. However, other parameters can be related, such as contact inhibition [15], pH alteration [16], accumulation of toxic metabolites [17], [18], among others, which were not considered in this exploratory analysis. The possibility that more replications would improve the average CM prediction without adjusting the parameters was discarded because of the low variance presented in the MAE of the combination of data sets, as discussed earlier. So, a large number of replications would probably not avoid those significant overestimations observed, even with the cells inner characteristics varying pseudo-randomly. The authors believe

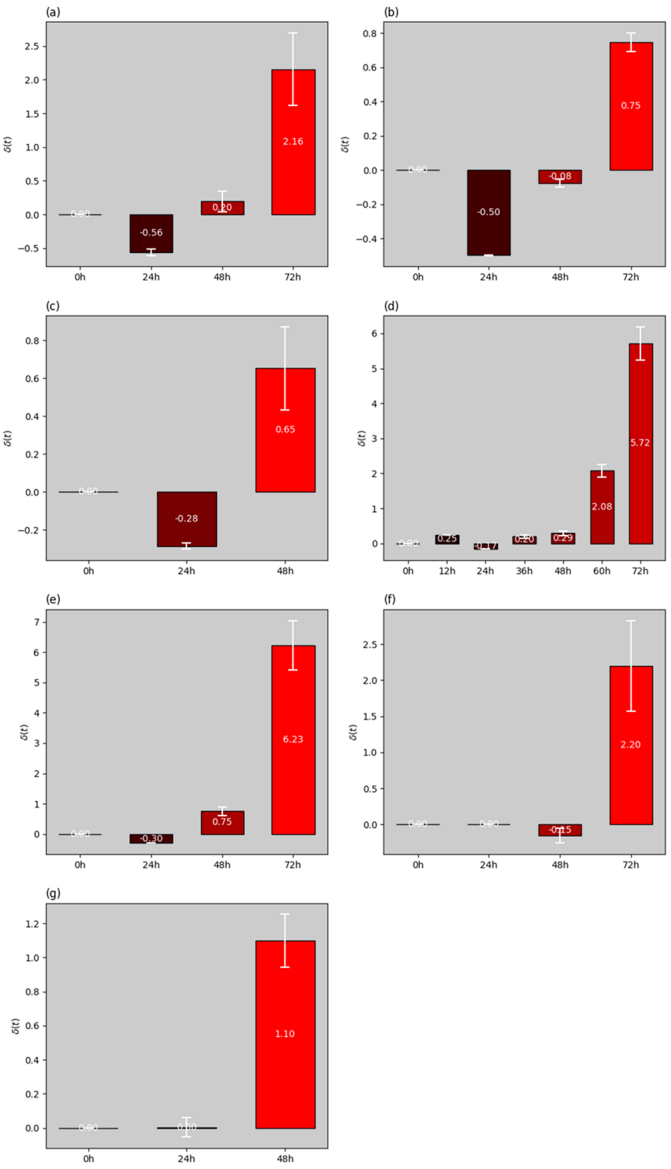


Fig. 3. Non-absolute relative error of the cell number of the simulation data at each period of time. (a) CHO-BC® cells [9]; (b) CHO-DG44 cells [10]; (c) CHO-K1 cells [11]; (d) Vero ATCC-CCL-81 cells [12]; (e) and (f) Vero ATCC-CCL-81 cells in suspension and adherent culture, respectively, [13]; (g) CHO-GS46 cells [14].

that adjusting only the parameters of interest could reduce the CM overestimation of the final cell density, but probably the early stages would suffer an error prediction increase.

However, it is already established in the literature that environmental cues to induce halt of cell division are more complex than just the number of cells in the vicinity of each cell and the availability of nutrients. For this study, we chose not to increase the number of parameters under evaluation because the raw experimental data were not available.

The aforementioned non-absolute error of the simulations, illustrated in Fig. 3, is the key to establish a strategy to improve the prediction power of the CM. These data can clarify how the fixed parameters impacted the simulation. The first two parameters presented in Table III —cell health range and senescent death rate range— do not needed any adjustment in the scenarios used in this work, because the period of cultivation selected were the initial stages, when the cells, theoretically, could not reach yet the number of divisions required to enter in senescence, and the health is only a representation of the cell to endure damage and stressful conditions before entering in apoptosis or necrosis. The G2 and Mitotic phases duration parameters are another pair of values that does not need any adjustment in the CM framework used for two reasons, the values used are well-known in the literature, representing the average duration of each phase in a mammalian animal cell; and because in the future framework of the model, these phases logic will be refined to simulate the cellular division process using mechanical and spatial feedback of the *in silico* cell in division and its surrounding. The variables related to stress of the cell will also be improved through the implementation of boundary conditions of the signaling context presented in cellular cultures. All the parameters already mentioned are not included in our hypothetical solution for solve the overall overestimation of the CM.

For us, the variables of the Ornstein-Uhlenbeck process and the variable of glucose concentration required for cellular division were the fixed parameters responsible for the prediction capability achieved by the CM. The glucose concentration required represents in a simplistic way the volume that the *in silico* cell must achieve to pass the checkpoint of G1 phase. In other words, this parameter represents how much glucose oriented to biomass production the cells need to synthesize and duplicate its protein mass as one of the conditions required for the cell to enter in the DNA duplication process. Therefore, this value can lead the model to overestimate the volume needed for the specific cell line to divide if it is greater than the actual volume required; or it can lead the model to underestimate the volume needed if it is smaller than the actual volume. In the simulations made for this study, we think that the fixation of this variable contributes to the overall overestimation of the model, mainly because this variable was not linked to the actual size of the *in silico* cell (i.e., the cell volume did not increase as the protein mass increases), so the negative feedback given by the available space around the cell for its growth and division did not regulate glucose uptake and proteins synthesis. The absence of this type of feedback leads to a theoretical growth rate greater than the actual growth rate of the cell line. In the case of the experimental conditions simulated, the simulation growth rate overestimated, on average, the experimental growth rate by 167.7%. In other words, the *in silico* cells divided 1.7x faster than the *in vitro* cells (Fig. 4A). Furthermore, the variables of the Ornstein-Uhlenbeck process will directly affect proteins synthesis. The protein α threshold rate will impact, for example, the protein mass required for the cell to divide. A larger rate will lead to a great threshold value fluctuation because of the restoring force of this stochastic process; also, different cell lines present different rate values (as can be seen in Fig. 4). The restoring force variable impacts

how fast the protein threshold returns to its center. And the initial values of the protein α concentration and its threshold can lead the CM to underestimate or overestimate how much the cell needs to grow for its division. These variables impact the speed of the culture growth and its doubling time. Looking at the doubling time of the *in silico* and *in vitro* culture, we can conclude that the values used for the threshold rate and the restoring force lead the CM to underestimate the protein mass required for cellular division in these cell lines, so the *in silico* cells were able to divide faster than the *in vitro* cells, as the *in vitro* population needed 80.5% more time, on average, to double when compared to the *in silico* population (Fig. 4B).

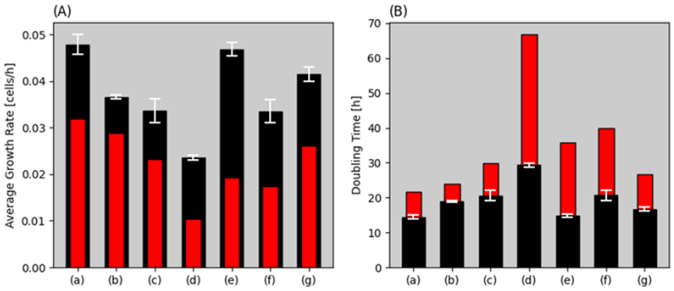


Fig. 4. Comparison of the average growth rate (A) and comparison of the doubling time (B) of the simulation data (black bars) and the experimental data (red bars). (a) CHO-BC® cells presented in [9]; (b) CHO-DG44 cells presented in [10]; (c) CHO-K1 cells presented in [11]; (d) Vero ATCC-CCL-81 cells presented in [12]; (e) and (f) Vero ATCC-CCL-81 cells in suspension and adherent culture, respectively, presented in [13]; (g) CHO-GS46 cells presented in [14].

Future work should include the cultivation of the cell lines simulated in order to validate the CM through the F-test methodology proposed by [1]. Further investigations should also explore alternative values for the parameters that remained unchanged in the present study. The future CM framework can become able to support bioreactor simulations, enabling the evaluation of production performance and the identification of optimal operating strategies to achieve target product yields with high accuracy while avoiding excessively complex mathematical formulations. Ultimately, the model is intended to simulate cell 3D culture to generate and analyze vasculature geometries capable of sustaining the cells.

IV. Conclusion

The current framework of the CM evaluated in this study demonstrated the potential to reproduce the short-term GC of different cell lines. However, its predictive performance decreases over extended simulation periods. In the future, with further improvement in the simulation of metabolic processes and the implementation of cellular mechanical simulation, the proposed model could achieve the potential to evaluate different culture scenarios and support researchers in estimating suitable strategies for their specific objectives.

References

[1] A. M. Moraes, E. F. P. Augusto, L. R. Castilho, Tecnologia do cultivo de células animais de biofármacos a terapia gênica: de biofármacos a terapia gênica, São Paulo: GEN, 2008. 503 p. ISBN 9788572417303.

[2] D. Sietaram, P. Kotidis, G. Finka, A. A. Lapkin, "A multi clone kinetic model for characterizing Chinese hamster ovary cell line variability," Journal of Industrial Microbiology and Biotechnology, vol. 52, 2025, kuaf029, doi: 10.1093/jimb/kuaf029

[3] R. Sartori, J. Leme, C. P. Caricati, A. Tonso, F. Núñez, E. Gustavo, "Model comparison to describe BHK-21 cell growth and metabolism in stirred tank bioreactors operated in batch mode," Brazilian Journal of Chemical Engineering, vol. 35, pp. 441-458, 2018, doi: 10.1590/0104-6632.20180352s20160592.

[4] M. E. Kolewe, S. C. Roberts, M. A. Henson, "A population balance equation model of aggregation dynamics in Taxus suspension cell cultures," Biotechnol Bioeng, 109(2), pp. 472-82, 2012 Feb, doi: 10.1002/bit.23321, Epub 2011 Sep 9.

[5] B. Ben Yahia, L. Malphettes, E. Heinzle, "Macroscopic modeling of mammalian cell growth and metabolism," Appl Microbiol Biotechnol, 99(17), pp. 7009-24, 2015 Sep, doi: 10.1007/s00253-015-6743-6, Epub 2015 Jul 22.

[6] S. Panunzi, M. Pompa, P. M. D'Angelo, G. Bretti, A. De Gaetano, "Validation of an agent-based model for cell interactions in a microfluidic chip." PLoS One, 21(2):e0341962, 2026 Feb 9, doi: 10.1371/journal.pone.0341962.

[7] A. Antelmi, G. Cordasco, G. D'Ambrosio, D. De Vinco, C. Spagnuolo, "Experimenting with agent-based model simulation tools," Appl. Sci. 2023, 13, 13, doi: 10.3390/app13010013.

[8] K. Biswas, N. Brenner, "Universality of phenotypic distributions in bacteria," Phys. Rev. Pres., vol. 6(2), pp. L022043, 2024 May, doi: 10.1103/PhysRevResearch.6.L022043.

[9] X. Pan, C. Dalm, R. H. Wijffels, et al., "Metabolic characterization of a CHO cell size increase phase in fed-batch cultures," Appl Microbiol Biotechnol 101, pp. 8101–8113, 2017, doi: 10.1007/s00253-017-8531-y.

[10] B. Ben Yahia, B. Gourevitch, L. Malphettes, E. Heinzle, "Segmented linear modeling of CHO fed-batch culture and its application to large scale production," Biotechnol. Bioeng., 114, pp. 785-797, 2017, doi: 10.1002/bit.26214.

[11] W. S. Ahn, R. A. Maciek, "Metabolic flux analysis of CHO cells at growth and non-growth phases using isotopic tracers and mass spectrometry," Metabolic Engineering, Vol. 13, Issue 5, pp. 598-609, 2011, ISSN 1096-7176, doi: 10.1016/j.ymben.2011.07.002.

[12] Y. S. Sia, N. I. Azahar, M. A. Abd. Aziz, M. A. Arifin, "Sequential adaptation to serum-free medium for Vero cells cultivation on ultraviolet/ozone (UVO) treated microcarrier," Materials Today: Proceedings, Vol. 107, pp. 92-98, 2024, ISSN 2214-7853, doi: 10.1016/j.matpr.2023.08.031

[13] L. Bourigault, C. Bresson, C. Jean, et al, "Characterization of a suspension Vero cell line for viral vaccine production," NPJ Vaccines 10, 114, 2025, doi: 10.1038/s41541-025-01157-2.

[14] S. Alhuthali, C. Kontoravdi, "Population balance modelling captures host cell protein dynamics in CHO cell cultures," PLOS ONE 17(3): e0265886, 2022, doi: 10.1371/journal.pone.0265886.

[15] A. Puliafito, L. Hufnagel, P. Neveu, S. Streichan, A. Sigal, D. K. Fygenson, B.I. Shraiman, "Collective and single cell behavior in epithelial contact inhibition," Proc. Natl. Acad. Sci. U.S.A., 109 (3), pp. 739-744, 2012, doi: 10.1073/pnas.1007809109.

[16] W. Liu, Z. Ren, K. Lu, C. Song, E. C. W. Cheung, Z. Zhou, G. Chen, "The suppression of medium acidosis improves the maintenance and differentiation of human pluripotent stem cells at high density in defined cell culture medium," Int J Biol Sci., 14(5), pp. 485-496, 2018 Apr 5, doi: 10.7150/ijbs.24681.

[17] N. Kumar, P. Gammell, M. Clynes, "Proliferation control strategies to improve productivity and survival during CHO based production culture : a summary of recent methods employed and the effects of proliferation control in product secreting CHO cell lines," Cytotechnology, 53(1-3), pp. 33-46, 2007 Apr, doi: 10.1007/s10616-007-9047-6, Epub 2007 Mar 1.

[18] T. Matsuki, T. Pédron, B. Regnault, C. Mulet, T. Hara, P. J. Sansonetti, "Epithelial cell proliferation arrest induced by lactate and acetate from Lactobacillus casei and Bifidobacterium breve," PLoS One, 8(4):e63053, 2013 Apr 30, doi: 10.1371/journal.pone.0063053.

Comparative Analysis of Antiviral Strategies Targeting Viral Load and Immune Response in Dengue Infection Under a Sex-Based Perspective

Isadora Vargas
Department of Electrical Engineering
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
isadoraferreiravargas@gmail.com

Emilly Idalina
Department of Electrical Engineering
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
emillybeaa@gmail.com

Danilo Regis
Department of Electrical Engineering
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
danilo.regis@ifpb.edu.br

***Abstract*—Dengue is a major public health concern in tropical and subtropical countries due to its high incidence and recurrent outbreaks. Despite its significant impact, treatment remains primarily supportive, highlighting the need for effective antiviral strategies. This study employs mathematical modeling and computational simulations to evaluate and compare the effects of two antiviral approaches against dengue infection. The results demonstrated that early antiviral administration prevented sustained viral replication and markedly reduced viral load in both sexes compared with the untreated condition. In the female simulations, antiviral ρ produced the most pronounced reduction in viral load immediately after treatment initiation, whereas antiviral ψ promoted faster elimination of residual viral particles during the later stages of infection. In the male simulations, antiviral ψ showed greater reduction in the viral load , while antiviral ρ provided a rapid initial decrease in viral load. In both sexes, delaying treatment initiation from the first to the second day resulted in higher viral peaks and slower viral elimination, reducing the overall therapeutic benefit. These findings demonstrate that sex-specific immunological characteristics influence antiviral efficacy and highlight the potential value of incorporating sex-related parameter variation into within-host dengue modeling.**

***Index Terms*—dengue; mathematical modeling; antiviral therapy; T lymphocytes; viral dynamics.**

I. Introduction

The dengue virus (DENV) is one of the most significant arboviral diseases affecting global public health, particularly in tropical and subtropical regions. The disease belongs to the *Flaviviridae* family and is primarily transmitted by the *Aedes aegypti* female mosquito, whose proliferation is favored by environmental and climatic conditions such as high temperatures and intense rainfall [13]. In Brazil, (DENV) exhibits an endemic pattern with recurrent epidemic outbreaks and remains a major challenge for public health systems due to its high incidence and socioeconomic impact [1], [11]. Despite its considerable burden, dengue management remains predominantly supportive and symptomatic, highlighting the need for effective therapeutic interventions. Mathematical and computational models have become valuable tools for understanding the within-host dynamics of viral infections, allowing the investigation of interactions between viral particles, infected cells, and immune components, as well as the evaluation of potential treatment strategies [3], [6]. Among the immune components involved in dengue infection, T lymphocytes play a central role in both viral clearance and disease pathogenesis. These cells contribute to the recognition and elimination of infected cells and regulate antiviral responses through cytokine production, significantly influencing the progression and outcome of dengue virus infection [2], [9]. Furthermore, immune responses may vary among individuals due to biological factors, including sex-related differences, which can affect disease progression and treatment efficacy [7], [8]. Recent studies have proposed different antiviral approaches to control dengue infection. One strategy, represented by antiviral ρ (rho), acts directly on free viral particles, reducing viral load and limiting viral dissemination throughout the host [5]. The second antiviral strategy, represented by the parameter ψ, enhances immune-mediated viral clearance by increasing the effectiveness of the host immune response. Rather than directly inhibiting viral production, ψ increases the rate at which free viral particles are eliminated [3]. Since these antiviral strategies target different aspects of the infection process, comparing their effects on dengue within-host dynamics may provide important insights into the development of more effective and personalized therapeutic interventions. Therefore, the present study aims to analyze and compare the effects of antivirals ρ and ψ on the within-host dynamics of dengue infection through mathematical modeling and computational simulations, considering their impacts on viral load and T lymphocyte populations in both men and women.

II. Literature Background

A. Immune Cells Involved in Dengue Infection

The course of dengue infection results from a complex interplay between dengue virus (DENV) replication and the host immune response, this viral infection dynamics are represented in Fig. 1. After inoculation by an infected *Aedes aegypti* mosquito, the virus enters the circulation and rapidly encounters immune cells that are susceptible to infection, especially monocytes, macrophages, and dendritic cells. These

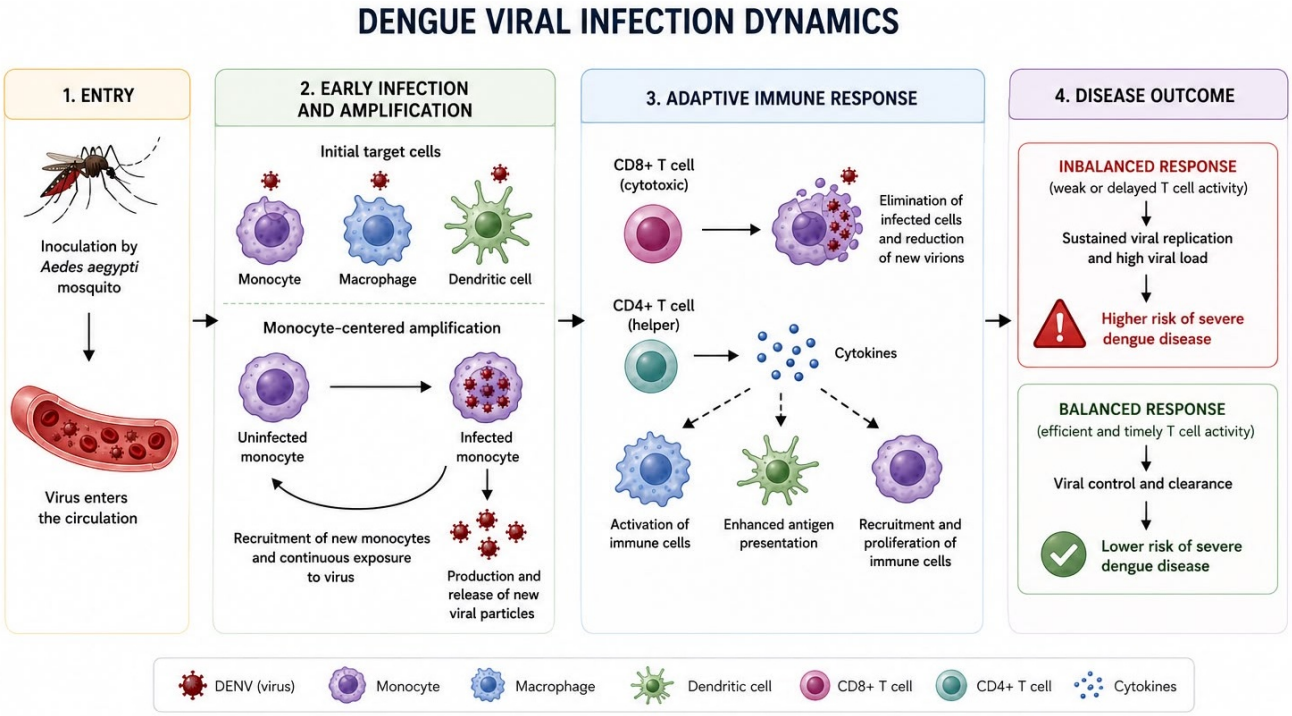


Fig. 1. DENV infection dynamics.

cells not only constitute the initial targets of DENV but also provide a favorable environment for viral multiplication and dissemination within the host [2].

Among these immune cells, monocytes are specifically important in the establishment and progression of infection. Following viral entry, infected monocytes actively produce and release new viral particles, contributing to the amplification of viral load. Meanwhile, uninfected monocytes continue to be recruited and become exposed to circulating virions, generating a continuous cycle of infection and cellular turnover [10].

The adaptive immune response, especially that mediated by T lymphocytes, is essential for limiting viral spread. Cytotoxic T lymphocytes eliminate infected cells and reduce the production of new virions, whereas helper T cells orchestrate immune activity by secreting cytokines and promoting the activation of additional immune populations [2]. Consequently, the efficiency of T-cell-mediated responses has a direct impact on the control of dengue infection.

Disease outcome depends largely on the equilibrium established between viral expansion and immune-mediated clearance. Efficient and timely activation of T lymphocytes favors viral control and contributes to infection resolution. In contrast, inadequate or delayed immune responses may permit sustained viral replication, thereby increasing the likelihood of severe clinical outcomes [2], [10]. Therefore, characterizing the interactions among viral particles, susceptible monocytes, infected monocytes, and T lymphocytes is fundamental for constructing mathematical models that accurately represent dengue within-host dynamics and for assessing the potential efficacy of antiviral interventions.

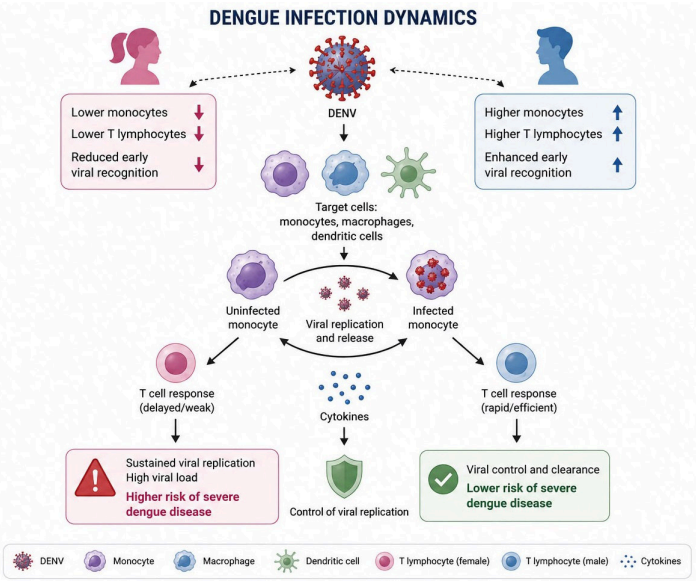


Fig. 2. sex-related differences assumed in the mathematical model

B. Female Immune System

The female immune system exhibits complex and dynamic characteristics that can significantly influence the progression of viral infections, as represented in Fig. 2. In general, women tend to develop different innate and adaptive immune responses than men due to differences in immune regulation and hormonal influences, which can affect both susceptibility to infections and therapeutic outcomes [8]. Nevertheless,

in the present mathematical model, females were assigned lower initial populations of T lymphocytes, which play a fundamental role in controlling viral replication by recognizing and eliminating infected cells and coordinating antiviral responses through cytokine production [2], and monocytes, therefore, characterizing lower rates of monocyte production and viral recognition compared to men [7]. Consequently, reduced availability of these immune cells may compromise the efficiency of viral clearance efficiency and contribute to increased vulnerability during dengue infection, particularly when antiviral treatment is delayed [7]. These findings highlight the importance of considering sex-related immunological differences in mathematical models of dengue infection and in the development of personalized therapeutic strategies, especially those aimed at preserving or increasing T-lymphocyte populations [2], [7], [8].

C. Male Immune System

The male immune system presents distinct immunological characteristics that can influence the progression and control of viral infections, as demonstrated in Fig. 2. Overall, men tend to exhibit less intense innate and adaptive immune responses than women, resulting in differences in susceptibility, disease progression, and responses to therapeutic interventions [8]. Nonetheless, in the mathematical model, male individuals were characterized by higher viral-recognition compared with women thereby presenting increased rates of T lymphocytes,responsible for limiting viral replication through the recognition and elimination of virus-infected cells [2], and increased initial rates of monocyte populations production compared with women [7]. These characteristics may contribute to a more efficient early immune response and greater capacity to control viral replication. Consequently, these findings emphasize the importance of considering sex-related differences in immune dynamics when developing mathematical models of dengue infection and evaluating antiviral strategies [2], [7], [8].

III. Methods

A. Mathematical Modeling of Immune Cells

The present study employed mathematical modeling based on ordinary differential equations (ODEs) and computational implementations developed in Python to investigate the immune dynamics of young adults of both sexes during dengue infection and under antiviral treatment, this course of the methodology used in the study is represented by Fig. 3. The model describes the temporal evolution of infection and immune responses by simulating viral load and immune cell populations [3], [12].

1) Equations: Four state variables were considered: $v(t)$, representing the concentration of free viral particles; $s(t)$, the population of susceptible monocytes; $i(t)$, the population of infected monocytes; and $z(t)$, the population of T lymphocytes.

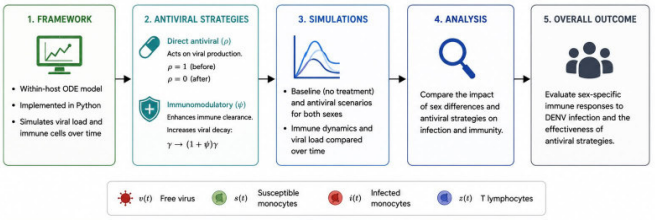


Fig. 3. Representation of the methodology used to obtain the results.

The parameters k and (γ) denote the viral replication rate and the natural clearance rate of free viral particles, respectively.

$$\begin{cases} \dfrac{dv}{dt} = ki(t) - \gamma v(t) - as(t)v(t) \\ \dfrac{ds}{dt} = \mu - \alpha s(t) - as(t)v(t) \\ \dfrac{di}{dt} = as(t)v(t) - \beta_1 i(t) - \psi i(t)z(t) \\ \dfrac{dz}{dt} = \eta + c_1 i(t) + di(t)z(t) - \delta z(t) \end{cases} \tag{1}$$

For all simulations, the initial conditions $v(0)$=165 and $i(0)$=10, together with the parameters k =20, a=0.003, and (γ=0.8), were assumed to be identical for both sexes. In contrast, the initial populations of susceptible monocytes $s(t = 0)$ and T lymphocytes $z(t = 0)$, as well as the monocyte production rate (μ), the T-lymphocyte production rate (η), and the viral recognition parameter (d), were adjusted according to sex-specific immunological characteristics [7], [8].

TABLE I
Model Parameters Modified by Sex.

Modified Elements	$s(t=0)$	$z(t=0)$	μ	η	d
Women	480	1700	120	0.20	0.01
Men	500	1800	130	0.21	0.02

2) Mathematical Model for Antiviral Treatment: The system of ordinary differential equations was based on previously proposed mathematical models of dengue within-host dynamics and immune interactions [3], [6]. Two antiviral strategies were incorporated into the model. The first, represented by the parameter (ρ), acts directly on free viral particles, reducing viral load and limiting viral dissemination. Following a previous approach, (ρ = 1) before treatment administration and (ρ = 0) after antiviral administration, representing the inhibition of viral production [5].

$$\frac{dv(t)}{dt} = (1-\rho)\kappa i(t) - \gamma v(t) - \alpha s(t)v(t) \tag{2}$$

The second antiviral strategy is represented by the parameter (ψ), which does not act directly on viral particles. Instead, it enhances the host immune system's ability to eliminate the virus by increasing the immune-mediated viral clearance rate. Specifically, (ψ) acts as a multiplicative factor on the natural viral decay coefficient, (γ), intensifying the viral removal term by a factor of (1+ψ). Consequently, the equation governing the dynamics of free viral particles becomes: [3].

$$\frac{dv(t)}{dt} = ki(t) - (1+\psi).\gamma v(t) - as(t)v(t) \qquad (3)$$

where the term ((1+ψ)γ v(t)) represents the enhanced elimination of viral particles mediated by the host immune response. This immuno modulatory strategy strengthens the natural defenses of the organism and contributes indirectly to the control of dengue infection.

Results

The simulation results demonstrate that early antiviral administration substantially reduces viral load in both sexes when compared with the untreated condition. In all simulated scenarios, antiviral therapy effectively interrupted sustained viral replication, resulting in a more rapid decline in viral particle concentration following treatment initiation. Furthermore, the two antiviral strategies evaluated exhibited distinct therapeutic profiles throughout the course of infection, indicating that their mechanisms of action differentially influence viral clearance dynamics.

In the female simulations, antiviral administration on the first day of infection produced a pronounced reduction in viral load shortly after the peak of infection. The antiviral ρ, which directly inhibits viral production, promoted the greatest decrease in viral load during the early post-treatment period. In contrast, the antiviral ψ, which enhances immune-mediated viral elimination, exhibited comparable efficacy during the initial phase of treatment but achieved a slightly faster reduction in residual viral load during the later stages of infection, as observed in Fig. 4.

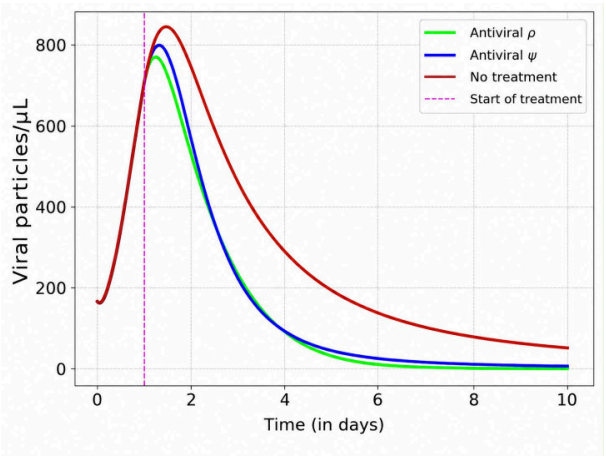


Fig. 4. Antiviral treatment administrated on the first day of infection in women.

.

When treatment initiation was delayed until the second day, both antiviral strategies remained effective; however, the therapeutic benefit was attenuated because viral replication had already progressed substantially before intervention, as presented in Fig. 5.

A similar pattern was observed in the male simulations. Early treatment initiation, as seen in Fig 6 markedly restricted

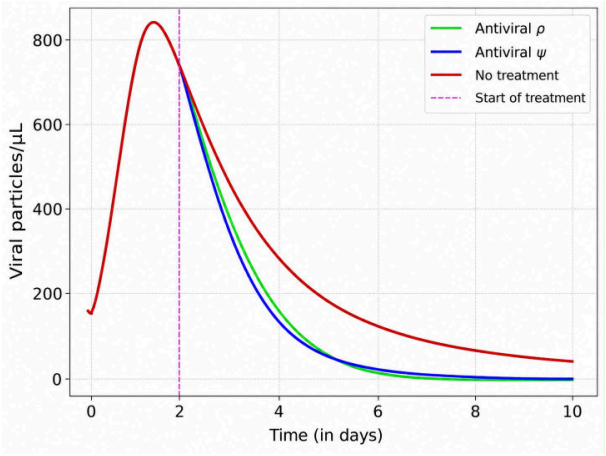


Fig. 5. Antiviral treatment administrated on the second day of infection in women.

infection progression and shortened the duration of elevated viral load relative to the untreated condition. In contrast to the female simulations, antiviral ψ exerted a stronger immediate effect following administration, whereas antiviral ρ enhanced viral clearance during the declining phase of infection, resulting in lower viral levels during the later stages.

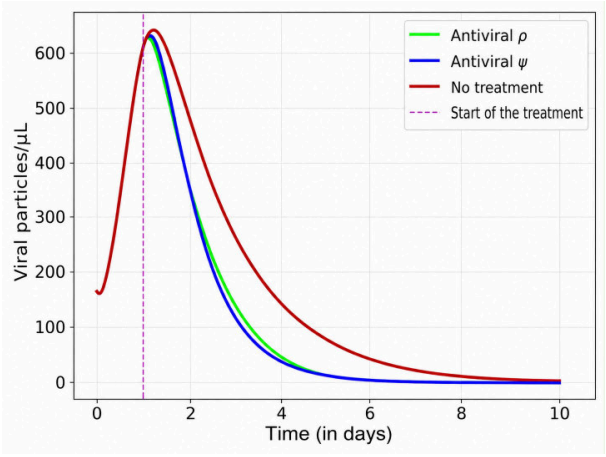


Fig. 6. Antiviral treatment administrated on the first day of infection in men.

The delayed administration of treatment until the second day, represented by Fig. 7 reduced the efficacy of both antiviral strategies, emphasizing the importance of early therapeutic intervention for optimal viral control.

Conclusion and Discussion

The comparison between sexes further revealed that female simulations exhibited higher peak viral loads than male simulations, reflecting the sex-specific immunological parameters incorporated into the mathematical model. Despite these differences, both sexes responded favorably to antiviral therapy, with treatment initiated on the first day consistently producing superior outcomes. Overall, both antiviral strategies substantially improved viral control compared with the untreated scenario.

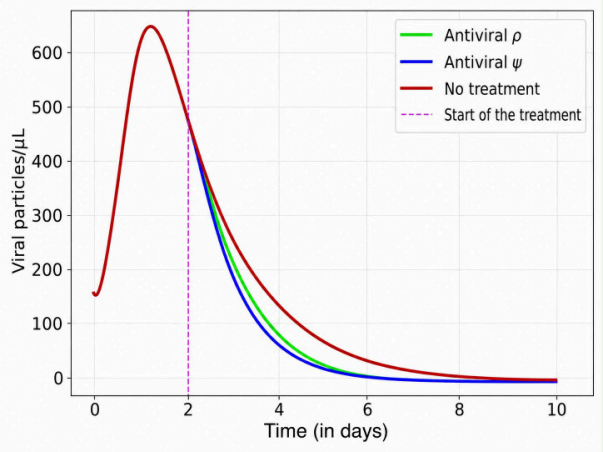


Fig. 7. Antiviral treatment administrated on the second day of infection in men.

However, the comparative analysis suggests that the optimal therapeutic approach may differ according to biological sex. Specifically, antiviral ρ produced the greatest overall reduction in viral load in the female simulations, whereas antiviral ψ achieved slightly greater efficacy during the later stages of infection in the male simulations. These findings indicate that sex-related immunological and physiological differences can influence antiviral performance and, consequently, therapeutic outcomes. Therefore, the present results support the development of sex-specific and personalized antiviral strategies to maximize treatment efficacy in dengue virus infection.

IV. Acknowledgment

The authors would like to acknowledge the support provided to this study by the National Council for Scientific and Technological Development (CNPq) through an Undergraduate Research Scholarship.

References

[1] S. Sousa, M. Carneiro, A. Eiras, J. Bezerra, and D. Barbosa, "Factors associated with the occurrence of dengue epidemics in Brazil: a systematic review," Rev Panam Salud Publica. Panama, vol. 45, August 2021.

[2] A. Mathew, E. Townsley, and F. Ennis, "Elucidating the role of T cells in protection against and pathogenesis of dengue virus infections," Future Microbiol. United States Of America, vol. 9, pp. 411–425, March 2014.

[3] T. Oriente, "Hemodynamic analysis of exosome propagation in the maternal-placental communication system during a Zika virus infection," "Análise hemodinâmica da propagação de exossomos no sistema de comunicação materno-placentário durante uma infecção pelo vírus Zika," M.S. Thesis, PPGEE. Signal Processing, IFPB, Brazil, 2021. [Online]. Available: https://repositorio.ifpb.edu.br/handle/177683/2189

[4] W. Rodrigues, K. Carvalho, J. Silva, M. Vilaça, E. Silva, and C. Regis, "Dengue virus analysis in different stages of pregnancy," "Análise do vírus da dengue em diferentes estágios da gravidez," CBEB. Brazil, pp.589–594, September 2024.

[5] K. Carvalho, M. Vilaça, W. Rodrigues, J. Silva, E. Silva, and C. Regis, "Análise do impacto do tratamento antiviral no modelo da dinâmica populacional da infecção por dengue," CBEB, Brazil, pp. 575–580, Sep. 2024.

[6] H. Clapham, V. Tricou, N. Chau, C. Simmons, and N. Ferguson, "Within-host viral dynamics of dengue serotype 1 infection," J R Soc Interfac. England, vol. 11, July 2014.

[7] I. Vargas, E. Galdino, K. Carvalho, and C. Regis, "Analysis of the dengue immune response under a gender perspective: antiviral treatment administration in men and women," "Análise da resposta imunológica da dengue sob recorte de gênero: administração do tratamento antiviral em homens e mulheres," SIMPIF. Brazil, July 2025.

[8] S. L. Klein and K. L. Flanagan, "Sex drives dimorphic immune responses to viral infections," J. Immunol., vol. 198, no. 5, pp. 1782–1790, Mar. 2017.

[9] A. L. Rothman, "Immunity to dengue virus: a tale of original antigenic sin and tropical cytokine storms," Nat. Rev. Immunol., vol. 11, no. 8, pp. 532–543, Aug. 2011.

[10] A. Srikiatkhachorn and A. L. Rothman, "Immune mechanisms in dengue and dengue hemorrhagic fever," Curr. Opin. Immunol., vol. 29, pp. 46–49, Aug. 2014.

[11] D. J. Gubler, "Dengue and dengue hemorrhagic fever," Clin. Microbiol. Rev., vol. 11, no. 3, pp. 480–496, Jul. 1998.

[12] M. A. Nowak and R. M. May, *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford, U.K.: Oxford University Press, 2000.

[13] World Health Organization, "Dengue and severe dengue," World Health Organization, Geneva, Switzerland, Fact Sheet, 2024.

Comparison of Computed Muscle Ativations and sEMG on Lower Limb Muscle SSC

Carlos Rodrigues
C-BER - Centre for Biomedical Engineering Research
INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal
https://orcid.org/0000-0001-6408-6032

Miguel Correia
C-BER - Centre for Biomedical Engineering Research
INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal
https://orcid.org/0000-0001-6065-9358

João Abrantes
MovLab - Laboratory of Technologies for Interactions and Interfaces
CICANT - Centre for Research in Applied Communication, Culture, and New Technologies
Lisbon, Portugal
https://orcid.org/0000-0002-7628-0462

Marco Rodrigues
Department of Electronic and Systems
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0002-5143-3696

Jurandir Nadal
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
https://orcid.org/0000-0003-3921-7787

***Abstract*—This study presents and applies the comparison of a technique based on the magnitude and phase metrics as well as the linear correlations and amplitudes of computed muscle activations (CMA) and lower limb surface electromyography (sEMG) during different stretch-shortening cycle. CMA algorithm is based on data fusion from external kinematic and dynamic direct measurements for inverse dynamics with subject specific musculoskeletal model, static optimization and 2nd order polynomial cost function to address muscle redundancy. The magnitude (M) and phase (P) metrics proved adequate for the comparison of CMA based on sEMG. Thus, despite different nature of CMA and sEMG, presented methods pave the way for improvement of both techniques, namely with different optimization strategies, cost functions, muscles and movements applied at larger samples to assess multimodal comparison.**

Keywords—computed muscle activations, sEMG, lower limb, muscle SSC

I. Introduction

Natural human movement is the result of the integration by the neuromusculoskeletal system starting at the neural sum of the excitatory and inhibitory inputs from each motoneuron, the integration of all the motor twitches for the muscle recruitment, the sum of the agonist and antagonist muscle actions at each joint, and the combination of the synergistic joint actions towards a common goal [1]. Muscle stretch-shortening cycle (SSC) is a natural phenomenon involved in gait, running and jumping, with lower limb muscle contraction immediately preceded by the muscle stretch for efficient submaximal and powerful maximal action [2-4]. Thus, despite lower limb muscle SSC can be observed in gait and running, its higher expression and accessibility is performed on maximum vertical jump with long countermovement (CM) at countermovement jump (CMJ) and drop jump (DJ) with short CM for comparison with squat jump (SJ) and no CM [5-6].

Muscles, as the active components responsible under the control of the central nervous system (CNS) for the necessary actions to produce natural human movement are typically assessed using noninvasive surface electromyography (sEMG). Despite isolated concentric and excentric actions have received larger research effort, muscle SSC corresponds to a natural human action, with sEMG assessment issues at dynamic conditions.

More recently, computed muscle actions (CMA) from inverse dynamics, with subject specific musculoskeletal modeling and simulations (MSK-MS) based on optimization techniques, have arisen to address muscle redundancy issues, with an open question on the most adequate cost function for realistic muscle tendon unit (MTU) action prediction at different CM and SSC. For this reason, we have performed a comparison study of the lower limb CMA from personalized MSK-MS with corresponding sEMG towards validation at different CM and SSC. According to different nature of CMA and sEMG, complementary analysis was used for comparison at each signal category, namely the phase/magnitude metrics [7], the Pearson correlations and the amplitudes/dispersions.

II. Materials and Methods

A. Trial tests

A total of 54 trials were assessed using subject-specific musculoskeletal digital models, including three repetitions each of the countermovement jump (CMJ), drop jump (DJ), and squat jump (SJ). The sample is composed by a group of six sports and physical education students with ages 20-24 years old, body masses 60-84 kg and heights 1.69-1.86 m. None of the participants presented any pathology or had undergone specialized sports training. The study was approved CAE200903 by the ISMAI scientific committee 01/03/2009 and the ethical standards were followed on human experimentation in agreement with the Helsinki Declaration of 1975, as revised in 2000 and 2008. Participants were explained on the study purpose and procedures and signed written informed consent

Reflective adhesive markers were positioned at the skin surface over palped anatomic joints. These included the right and left acromioclavicular joints at the shoulders (RSHO /LSHO), anterior superior iliac spines at the hip (RASI/LASI), knees femoral epicondyles (RKNE/LKNE), lateral malleoli at the ankles (RANK/LANK), heels calcanei (RHEE/LHEE), and the distal ends of the second metatarsals at the toes (RTOE/LTOE). In addition, Aqua-Wet gel Skintact F55 electrodes were placed in a bipolar arrangement according to SENIAM guidelines for surface electromyography (sEMG) recording of selected lower-limb muscles, namely *vastus medialis* (VM), *rectus femoris* (RF), *vastus lateralis* (VL), *gastrocnemius lateralis* (GL), and *gastrocnemius medialis* (GM).

During the experimental trials, ground reaction forces and moments were collected at a sampling frequency of 1000 Hz using an AMTI BP2416-4000 CE force platform (Advanced Mechanical Technology Inc., Watertown, MA) connected to an AMTI Mini Amp MAS-6 amplifier. Simultaneously, sEMG synchronized signals were acquired at 2000 Hz and transmitted through a 32-channel Biotell 99 telemetry system (Glonner Electronic GmbH). Video recordings were captured at 100 Hz after external synchronization using two JVC GR-VL9800 camera system (JVCKENWOOD Corporation, Kanagawa, Japan), positioned 4 m from the subjects and aligned parallel to the sagittal plane at anterior and posterior locations.

B. Signal processing

Ground reaction forces and force moments were resampled at 100 Hz to match with the image acquisition frequency. The integration of each 2D camera image with the calibration system was performed using direct linear transformation (DLT), and the 3D coordinates of the reflective markers were tracked using Simi Motion 3D Version 7.5.280 (Simi Reality Motion Systems GmbH, Germany). The obtained 3D Cartesian coordinates were filtered in the time domain using a bidirectional second-order low-pass Butterworth filter with a 12 Hz cut-off frequency, applied in both forward and reverse directions to prevent phase distortion, as described by [1]. Raw sEMG signals were normalized relative to the maximum voluntary contraction (MVC), rectified, and subsequently filtered using a second-order low-pass Butterworth filter with 5.0 Hz cut-off frequency applied bidirectionally to avoid phase shift and generating the sEMG linear envelope.

C. Computed muscle actions

Inverse kinematics was implemented using AnyBody Version 8.1.2 Modeling System (AnyBody Technology A/S, Aalborg, Denmark) and articular cartesian coordinates to obtain lower limb joint angles, angular velocities and accelerations. Inverse dynamics was performed from the joint angular kinematics and ground reaction forces / force moments using multi-articular dynamics and obtaining lower limb joint moments of force. Based on joint moments of force, the musculoskeletal geometry and muscle properties, the CMA were calculated, as presented at Fig. 1, using static optimization and polynomial $p = 2$ cost function along with computed muscle control to minimize the error between the model and the experimental joint angles / angular velocities with time iteration and direct dynamics.

Applied musculoskeletal model was the full human body model with 432 degrees of freedom (DOF) and eleven segments for the lower limb, corresponding to the pelvis, femurs, tibias, talus and feet, with 4 joints on each limb, 3 DOF spherical joints at the hips, and 1 DOF hinge joints at the knees, the talocrural, and the subtalar joints, with the patella's positions defined as a function of the knee angle as defined by De Pieri et al. [8].

The model is composed by a total of 55 muscular actuators divided by 169 Hill type elements based on TLEM (Twente lower extremity model) data [9-10]. According to the relevance for the present study, data was selected on the number of elements, type of the path (WS - wrapping surface), type of the origin and insertion sites (Point, Line (order)), mass (g) and volume (ml) for the selected muscle actuators *gastrocnemius lateralis* (GL), *gastrocnemius medialis* (GM), *rectus femoris* (RF), *vastus lateralis inferior* (VLi), *vastus lateralis superior* (VLs), *vastus medialis inferior* (VMi), *vastus medialis mid* (VMm) and *vastus medialis superior* (VMs) presented at Tab. 1.

TABLE I. DATA FROM THE SELECTED MUSCLE ACTUATORS

Muscle	# Elem.	Path	Origin	Insertion	Mass (g)	Vol. (ml)
GL	1	WS	Point	Point	54	54
GM	1	WS	Point	Point	111	107
RF	2	WS	Point	Line (2)	118	114
VLi	6	WS	Line (2)	Line (3)	84	84
VLs	2	WS	Line (2)	Point	338	330
VMi	2	WS	Line (3)	Line (3)	47	46
VMm	2	WS	Line (3)	Line (3)	87	88
VMs	4	WS	Line (2)	Point	111	112

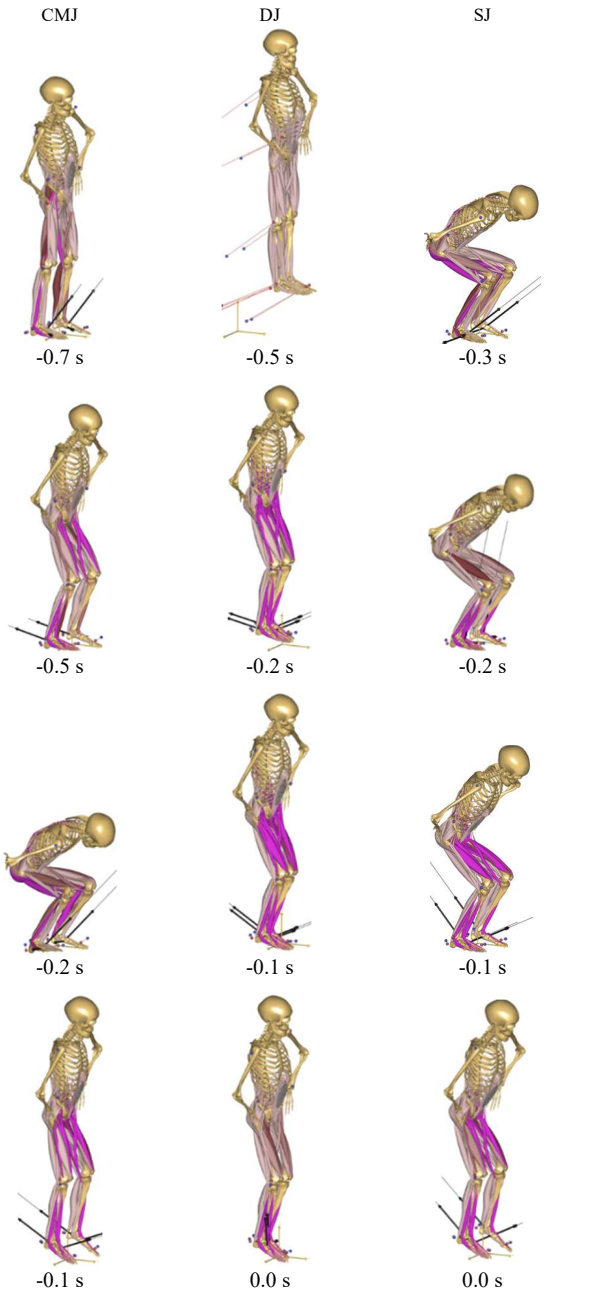


Fig. 1 Selected positions during impulse phases of performed trials CMJ, DJ and SJ with muscle activations and ground reaction forces / moments.

D. Comparison at CMA and sEMG

1) Magnitude and phase diferences

The comparisons within CMA and within sEMG were performed based on two different types of metrics: (*i*) a set of flexible metrics corresponding to the phase and magnitude differences, and (*ii*) a set of standard metrics, namely the Pearson correlations and the amplitude comparison. The goal is to perform the validation based on two different methods, to assess the adequacy of each method and its results.

Regarding the flexible metrics, the comparison of computed muscle activations (CMA) for selected muscle groups *gastrocnemius lateralis* (GL), *gastrocnemius medialis* (GM), *rectus femoris* (RF), *vastus lateralis inferior* (VLi), *vastus lateralis superior* (VLs), *vastus medialis inferior* (VMi), *vastus medialis mid* (VMm) and *vastus medialis superior* (VMs) from the TLEM musculoskeletal model was performed, using the magnitude (M) and the phase (P) metrics as defined in (1)-(5) for $m = 1, \dots, 9$ and $n = m + 1, \dots, 9$.

$$v_{mm} = 1/(t_2 - t_1) \int_{t_1}^{t_2} CMA_m^2(t)\, dt \quad (1)$$

$$v_{nn} = 1/(t_2 - t_1) \int_{t_1}^{t_2} CMA_n^2(t)\, dt \quad (2)$$

$$v_{mn} = 1/(t_2 - t_1) \int_{t_1}^{t_2} CMA_m(t)\, CMA_n(t)\, dt \quad (3)$$

$$M = \sqrt{v_{nn}/v_{mm}} - 1 \quad (4)$$

$$P = 1/\pi \cos^{-1}\left(v_{mn}/\sqrt{v_{mm} v_{nn}}\right) \quad (5)$$

As for the surface EMG signals, the same metrics were used to compare the phase and the magnitude within sEMG signals. These are for the $sEMG_i$ signals ($i = 1, \dots, 5$) of selected lower limb muscles *vastus medialis* (VM), *rectus femoris* (RF), *vastus lateralis* (VL), *gastrocnemius lateralis* (GL) e *gastrocnemius medialis* (GM), using the magnitude M and the phase P metrics presented in (6)-(10) for $i = 1, \dots, 5$ and $j = i + 1, \dots, 5$.

$$v_{ii} = 1/(t_2 - t_1) \int_{t_1}^{t_2} sEMG_i^2(t)\, dt \quad (6)$$

$$v_{jj} = 1/(t_2 - t_1) \int_{t_1}^{t_2} sEMG_j^2(t)\, dt \quad (7)$$

$$v_{ij} = 1/(t_2 - t_1) \int_{t_1}^{t_2} sEMG_i(t)\, sEMG_j(t)\, dt \quad (8)$$

$$M = \sqrt{v_{jj}/v_{ii}} - 1 \quad (9)$$

$$P = 1/\pi \cos^{-1}\left(v_{ij}/\sqrt{v_{ii} v_{jj}}\right) \quad (10)$$

2) Linear correlations and amplitudes

Regarding the standard metrics, for each pairwise comparison within CMA and within sEMG, the bivariate linear Pearson correlations were computed by the least squares method. As for the amplitudes, these were compared between trials for the location and dispersion measures through the mean and standard deviation of each muscle activity both on computed CMA and measured sEMG for each trial.

III. Results

A. Comparison of CMA

Illustrative time profiles of CMA for one of the performed trials on CMJ, DJ and SJ are presented at Fig. 2.

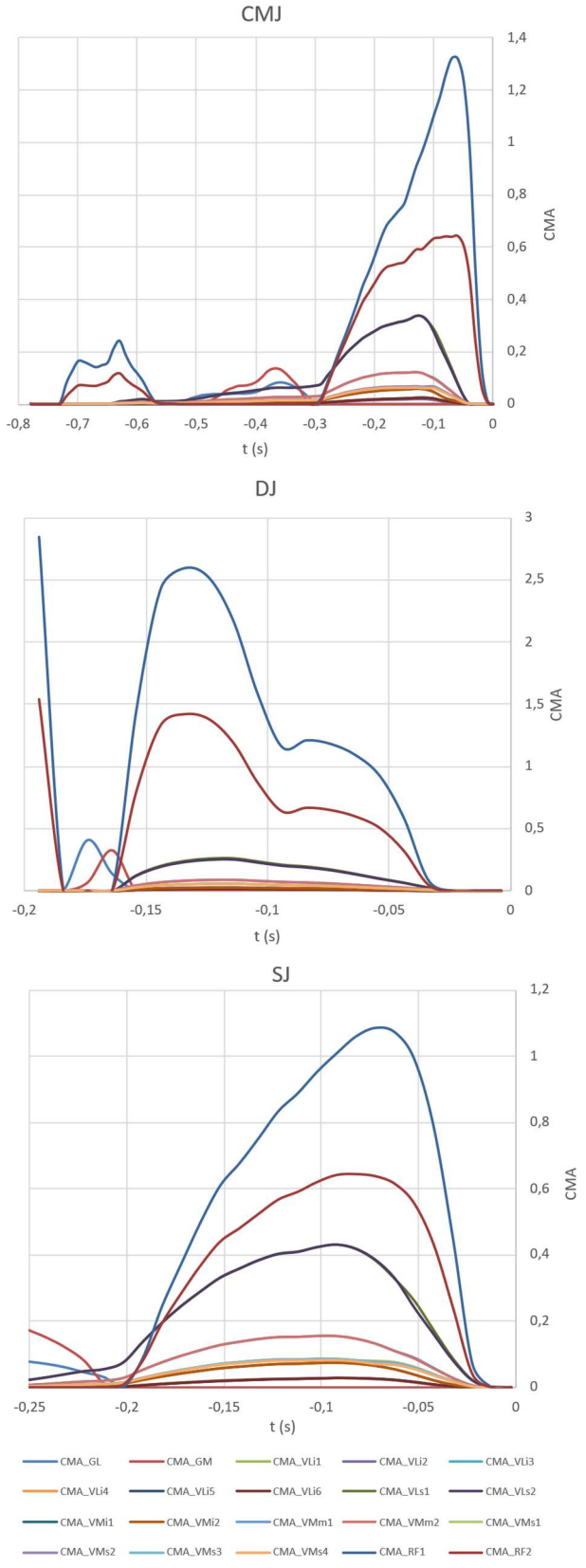


Fig. 2 Illustrative CMA of the main muscle groups for one performed trial of CMJ, DJ and SJ respectively.

1) CMA magnitude and phase diferences

The magnitude differences M presented higher means and standard deviations at CMJ, DJ and SJ for the groups RF, VLs, VMm and VMs in relation to GL and GM ($p < 0.05$)*, with lower amplitudes of CMA in GL and GM, as well as higher magnitudes M for VLi-VLs ($p < 0.05$)* corresponding to higher amplitudes of CMA for VLs relative to VLi, in Tab. 2.

TABLE II. MEAN AND STANDARD DEVIATION FOR THE PAIRWISE MAGNITUDE DIFERENCE $|M|$ BETWEEN MUSCLES CMA AT CMJ, DJ, SJ.

M - CMA	CMJ	DJ	SJ
GL - GM	0.452 ± 0.666	1.097 ± 2.136	2.133 ± 2.327
GL - RF1	42.691 ± 53.075*	31.771 ± 47.651*	107.734 ± 175.568*
GL - RF2	30.794 ± 39.656*	19.321 ± 28.625*	59.83 ± 94.705*
GL - VLi	0.254 ± 1.574	-0.494 ± 0.761	1.338 ± 3.695
GL – VLs	18.991 ± 25.075*	6.935 ± 11.732*	35.713 ± 57.994*
GL - VMi	2.481 ± 4.375	0.286 ± 1.923	5.254 ± 9.871
GL – VMm	6.683 ± 9.652*	1.741 ± 4.003*	12.352 ± 21.051*
GL – VMs	3.297 ± 5.387*	0.806 ± 2.736*	6.699 ± 12.165*
GM - RF1	31.901 ± 39.191*	11.195 ± 9.78*	21.341 ± 23.056*
GM - RF2	23.188 ± 28.562*	6.143 ± 5.704*	12.014 ± 12.042*
GM - VLi	-0.033 ± 1.205	-0.861 ± 0.142	-0.511 ± 0.479
GM – VLs	14.403 ± 19.197*	1.273 ± 2.178*	6.683 ± 7.511*
GM - VMi	1.681 ± 3.351	-0.65 ± 0.361	0.31 ± 1.278
GM – VMm	4.913 ± 7.4*	-0.21 ± 0.742*	1.799 ± 2.723*
GM – VMs	2.31 ± 4.131*	-0.485 ± 0.51*	0.611 ± 1.576*
RF1 - RF2	-0.267 ± 0.127	-0.42 ± 0.212	-0.379 ± 0.054
RF1 - VLi	-0.972 ± 0.006	-0.988 ± 0.012	-0.977 ± 0.003
RF1 - VLs	-0.549 ± 0.089	-0.811 ± 0.177	-0.644 ± 0.042
RF1 - VMi	-0.922 ± 0.016	-0.969 ± 0.031	-0.939 ± 0.008
RF1 - VMm	-0.827 ± 0.034	-0.933 ± 0.064	-0.87 ± 0.017
RF1 - VMs	-0.903 ± 0.019	-0.959 ± 0.036	-0.925 ± 0.009
RF2 - VLi	-0.961 ± 0.006	-0.983 ± 0.014	-0.963 ± 0.005
RF2 - VLs	-0.382 ± 0.095	-0.731 ± 0.218	-0.424 ± 0.078
RF2 - VMi	-0.893 ± 0.017	-0.957 ± 0.039	-0.902 ± 0.014
RF2 - VMm	-0.764 ± 0.039	-0.905 ± 0.078	-0.79 ± 0.03
RF2 - VMs	-0.867 ± 0.021	-0.941 ± 0.046	-0.879 ± 0.016
VLi – VLs	14.957 ± 0.092*	11.801 ± 8.776*	14.742 ± 0.057*
VLi - VMi	1.764 ± 0.029	0.984 ± 1.051	1.683 ± 0.019
VLi - VMm	5.1 ± 0.11	3.518 ± 3.013	4.746 ± 0.038
VLi - VMs	2.43 ± 0.062	1.916 ± 1.927	2.3 ± 0.025
VLs - VMi	-0.827 ± 0.002	1.482 ± 4.663	-0.83 ± 0.002
VLs - VMm	-0.618 ± 0.006	0.019 ± 1.337	-0.635 ± 0.003
VLs - VMs	-0.785 ± 0.003	0.3 ± 2.155	-0.79 ± 0.001
VMi - VMm	1.207 ± 0.022	0.891 ± 1.141	1.141 ± 0.008
VMi - VMs	0.241 ± 0.022	0.282 ± 0.653	0.23 ± 0.017
VMm - VMs	-0.438 ± 0.008	-0.148 ± 0.434	-0.426 ± 0.007

*Statistically significant differences ($p < 0.05$)

Additionally, the phase differences P present higher values for the muscle groups RF, VL and VM relative to GL and GM ($p < 0.05$)* as well as a second level of phase differences with higher values in VL, VM relative to RF ($p < 0.05$)*, with small phase differences between VL and VM ($p \geq 0.05$), as presented in Tab. 3.

TABLE III. MEAN AND STANDARD DEVIATION FOR THE PAIRWISE PHASE DIFERENCE P BETWEEN MUSCLES CMA AT CMJ, DJ, SJ.

P - CMA	CMJ	DJ	SJ
GL - GM	0.098 ± 0.036	0.224 ± 0.212	0.081 ± 0.068
GL - RF1	0.499 ± 0.002*	0.458 ± 0.066*	0.500 ± 0.001*
GL - RF2	0.499 ± 0.001*	0.469 ± 0.047*	0.500 ± 0.001*
GL – Vli	0.474 ± 0.024*	0.435 ± 0.065*	0.483 ± 0.01*
GL – VLs	0.474 ± 0.024*	0.445 ± 0.049*	0.483 ± 0.01*
GL – Vmi	0.473 ± 0.025*	0.436 ± 0.063*	0.482 ± 0.011*
GL – VMm	0.473 ± 0.026*	0.439 ± 0.057*	0.481 ± 0.011*
GL – VMs	0.473 ± 0.025*	0.438 ± 0.06*	0.482 ± 0.011*
GM - RF1	0.499 ± 0.002*	0.491 ± 0.011*	0.500 ± 0.001*
GM - RF2	0.499 ± 0.002*	0.492 ± 0.01*	0.500 ± 0.001*
GM – Vli	0.477 ± 0.025*	0.459 ± 0.031*	0.478 ± 0.007*
GM – VLs	0.477 ± 0.026*	0.468 ± 0.032*	0.477 ± 0.007*
GM – Vmi	0.476 ± 0.026*	0.46 ± 0.031*	0.476 ± 0.007*
GM – VMm	0.476 ± 0.027*	0.464 ± 0.032*	0.475 ± 0.008*
GM – VMs	0.476 ± 0.027*	0.462 ± 0.03*	0.476 ± 0.007*
RF1 - RF2	0.051 ± 0.019	0.048 ± 0.061	0.031 ± 0.002
RF1 – Vli	0.137 ± 0.059*	0.139 ± 0.088*	0.128 ± 0.033*
RF1 – VLs	0.136 ± 0.056*	0.133 ± 0.095*	0.122 ± 0.031*
RF1 – Vmi	0.147 ± 0.06*	0.143 ± 0.087*	0.139 ± 0.034*
RF1 – VMm	0.149 ± 0.057*	0.133 ± 0.091*	0.133 ± 0.031*
RF1 – VMs	0.137 ± 0.054*	0.13 ± 0.093*	0.112 ± 0.03*
RF2 - VLi	0.098 ± 0.049	0.138 ± 0.095	0.099 ± 0.033
RF2 - VLs	0.096 ± 0.048	0.132 ± 0.099	0.093 ± 0.032
RF2 - VMi	0.106 ± 0.05	0.141 ± 0.094	0.110 ± 0.034
RF2 - VMm	0.106 ± 0.049	0.131 ± 0.097	0.104 ± 0.032
RF2 - VMs	0.097 ± 0.048	0.131 ± 0.099	0.084 ± 0.03
VLi - VLs	0.011 ± 0.004	0.027 ± 0.026	0.013 ± 0.003
VLi - VMi	0.012 ± 0.003	0.007 ± 0.002	0.012 ± 0.001
VLi - VMm	0.023 ± 0.004	0.021 ± 0.011	0.02 ± 0.001
VLi - VMs	0.025 ± 0.006	0.019 ± 0.004	0.028 ± 0.005
VLs - VMi	0.014 ± 0.008	0.028 ± 0.02	0.02 ± 0.005
VLs - VMm	0.016 ± 0.003	0.011 ± 0.015	0.014 ± 0.001
VLs - VMs	0.016 ± 0.003	0.016 ± 0.02	0.019 ± 0.002
VMi - VMm	0.016 ± 0.004	0.021 ± 0.009	0.018 ± 0.004
VMi - VMs	0.026 ± 0.010	0.021 ± 0.007	0.035 ± 0.007
VMm - VMs	0.018 ± 0.007	0.009 ± 0.009	0.026 ± 0.002

*Statistically significant differences ($p < 0.05$)

2) CMA correlations and amplitudes

The associations between CMA signals for the muscle groups showed, as presented at Fig. 3, a common trend to the various types of countermovement CMJ, DJ and SJ, with higher positive correlations ($p < 0.05$)* for the gastrocnemius GL–GM, the rectus femoris RF1–RF2, and the vastus lateralis and medialis VL–VM, followed by the rectus femoris with the vastus lateralis and vastus medialis RF–VL and RF–VM, corresponding to agonist behaviour. On the other hand, the associations between the gastrocnemius (GL, GM) with the rectus femoris (RF) and the vastus lateralis and medialis (VL and VM) show negative correlations, corresponding to an antagonistic behaviour, as illustrated at Fig. 3.

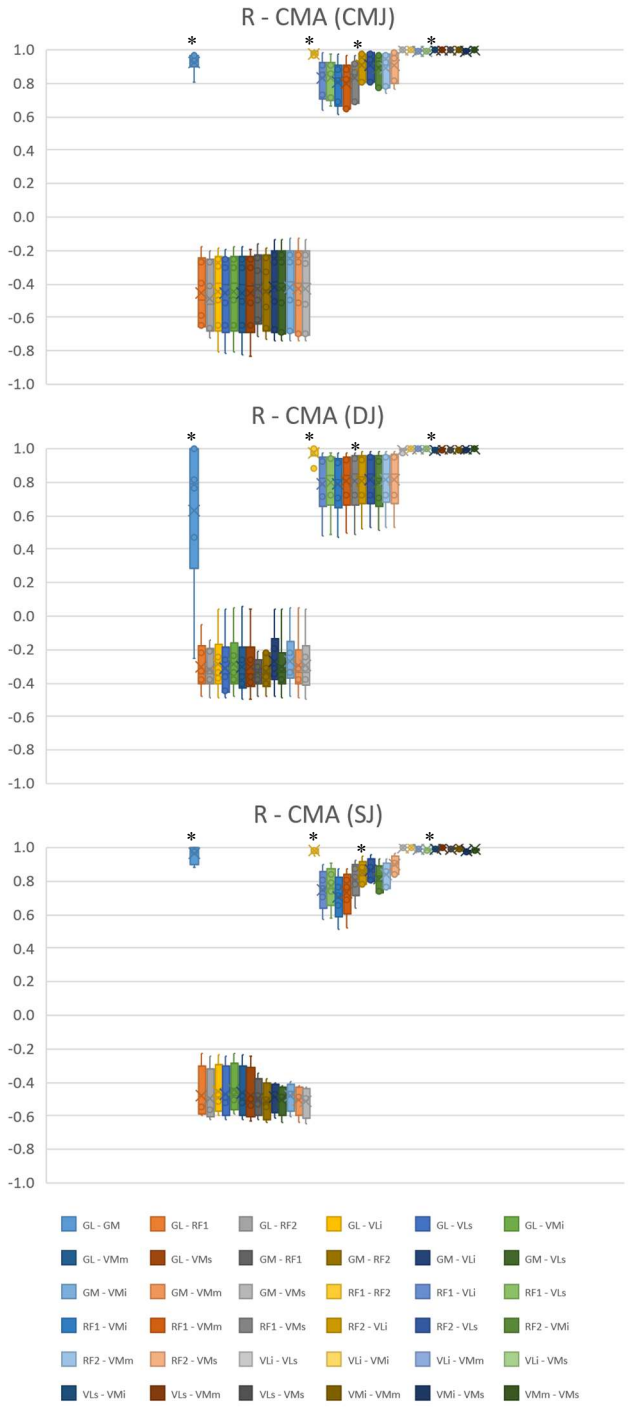


Fig. 3 Correlations of the CMA for the main muscle groups during CMJ, DJ and SJ respectively

The computed muscle activations CMA presented higher mean values ($p < 0.05$)* for RF than VLs muscle, both larger to the other muscles groups, as well as higher values for the DJ when compared with the SJ, both of which were higher than the CMJ, as shown in Tab. 4.

TABLE IV. MEAN AND STANDARD DEVIATION OF CMA AT CMJ, DJ AND SJ.

CMA	CMJ	DJ	SJ
GL	0.021 ± 0.033	0.038 ± 0.098	0.010 ± 0.024
GM	0.026 ± 0.046	0.022 ± 0.085	0.019 ± 0.044
VLi1	0.007 ± 0.007	0.011 ± 0.012	0.011 ± 0.009
VLi2	0.007 ± 0.007	0.011 ± 0.012	0.011 ± 0.009
VLi3	0.007 ± 0.007	0.012 ± 0.013	0.012 ± 0.009
VLi4	0.007 ± 0.007	0.012 ± 0.013	0.012 ± 0.009
VLi5	0.007 ± 0.007	0.013 ± 0.014	0.012 ± 0.009
VLi6	0.007 ± 0.007	0.013 ± 0.014	0.012 ± 0.009
VLs1	0.105 ± 0.114*	0.187 ± 0.184*	0.177 ± 0.132*
VLs2	0.105 ± 0.114*	0.182 ± 0.184*	0.175 ± 0.132*
VMi1	0.018 ± 0.020	0.029 ± 0.031	0.030 ± 0.023
VMi2	0.018 ± 0.020	0.029 ± 0.031	0.030 ± 0.023
VMm1	0.040 ± 0.043	0.065 ± 0.064	0.065 ± 0.048
VMm2	0.040 ± 0.043	0.064 ± 0.063	0.065 ± 0.048
VMs1	0.023 ± 0.024	0.042 ± 0.041	0.037 ± 0.027
VMs2	0.023 ± 0.024	0.041 ± 0.040	0.037 ± 0.027
VMs3	0.023 ± 0.024	0.040 ± 0.039	0.037 ± 0.027
VMs4	0.023 ± 0.024	0.040 ± 0.038	0.037 ± 0.026
RF1	0.220 ± 0.298*	0.787 ± 0.762*	0.469 ± 0.416*
RF2	0.161 ± 0.206*	0.498 ± 0.478*	0.315 ± 0.280*

*Statistically significant differences ($p < 0.05$)

Regarding the comparison of the results from the CMA magnitude differences M and the means of the corresponding CMA, both metrics presented consistent higher M and mean CMA values for the quadriceps muscles in relation to the gastrocnemius. Nevertheless, whereas the M metrics at each CM type involves the relative magnitude of the total CMA time series, as presented in (1)-(4), the CMA mean values involves the comparison of the absolute mean values, contributing for additional detection of the differences among CM, with higher values at DJ and short CM in relation to SJ without CM, both higher than CMJ with long CM.

The phase differences P of the computed CMA and the bivariate correlations of the corresponding CMA presented a systematic agreement among different CM, with lower phase differences P and higher bivariate correlations for the muscles at the same groups namely GL-GM and the VL-VM, with a second cluster of intermediate phase differences P and bivariate correlations at RF-VL and RF-VM, in contrast to larger phase differences and lower bivariate correlations at the antagonistic muscle groups of GL, GM with RF, VL, and VM, thus pointing for computed muscle control CMA to implement the lower limb synergistic and antagonistic muscle behaviour along with the ability of the phase differences and bivariate correlations to consistently capture this systematic behaviour.

B. *Comparison of sEMG*

Illustrative time profiles of sEMG for selected performed trials of CMJ, DJ and SJ are presented at Fig. 4.

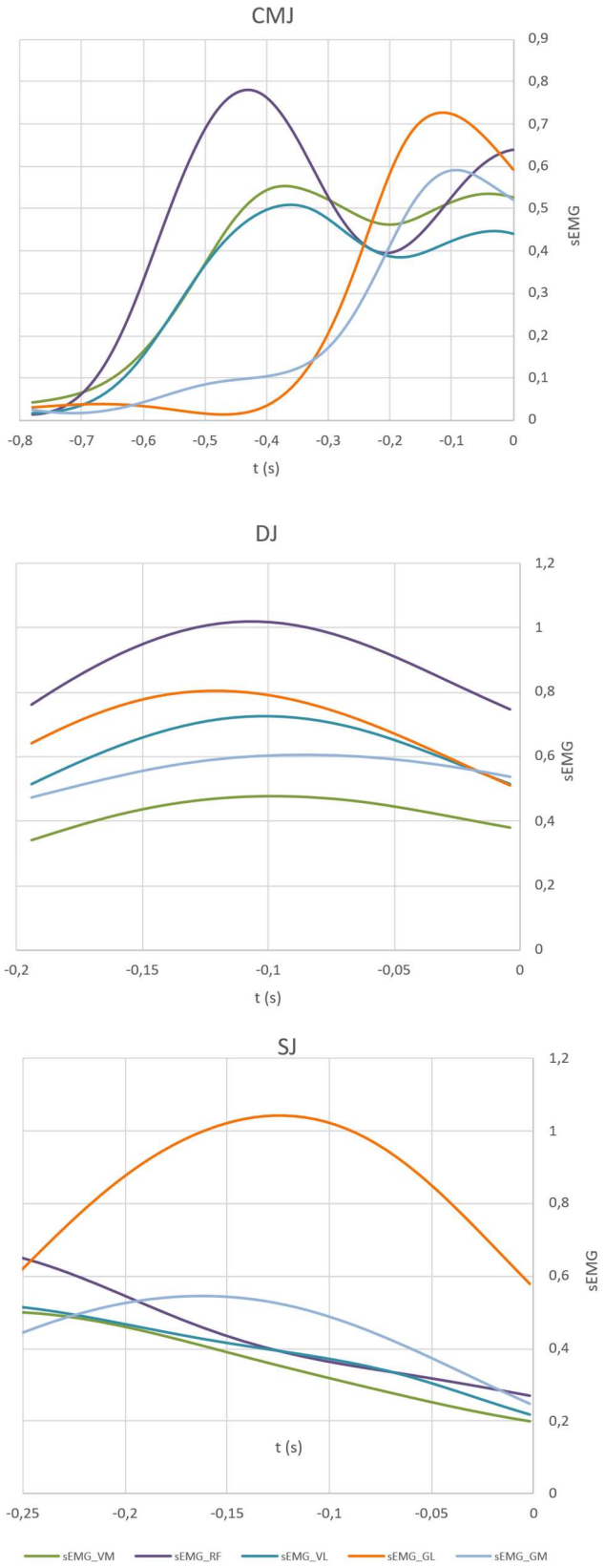


Fig. 4 Illustrative examples of sEMG for the main muscle groups for one performed trial of CMJ, DJ and SJ respectively.

1) *sEMG magnitude and phase diferences*

Based on obtained results for the magnitude M, resulting from comparison of sEMG signals, presented in Tab. 5, we highlight the higher values ($p < 0.05$)* of M for VM – RF shared by CMJ, DJ and SJ corresponding to greater amplitudes of RF relative to VM for the different types of CM. DJ further presented higher negative magnitudes ($p < 0.05$)* for RF-VL, RF-GL and RF-GM corresponding to higher amplitudes of RF relative to VL, GL and GM. Finaly, SJ presented higher negative magnitudes ($p < 0.05$)* for RF-VL and RF-GM corresponding to higher amplitudes of RF relative to VL and GM, and higher positive magnitudes for VL-GL, corresponding to higher amplitudes of GL relative to VL at SJ, with higher standard deviation on SJ for VM-GL, RF-GL and VL-GL corresponding to higher dispersion across different trials in the absence of CM.

TABLE V. MEAN AND STANDARD DEVIATION FOR THE PAIRWISE MAGNITUDE DIFERENCE $|M|$ BETWEEN MUSCLES SEMG AT CMJ, DJ, SJ.

***M* - sEMG**	**CMJ**	**DJ**	**SJ**
VM – RF	0.456 ± 0.513*	0.570 ± 0.698*	0.512 ± 0.244*
VM - VL	0.026 ± 0.201	-0.192 ± 0.334	-0.267 ± 0.263
VM - GL	-0.084 ± 0.350	-0.205 ± 0.319	0.071 ± 0.882
VM - GM	-0.164 ± 0.228	-0.211 ± 0.221	-0.228 ± 0.342
RF – VL	-0.213 ± 0.316	-0.415 ± 0.333*	-0.493 ± 0.263*
RF – GL	-0.266 ± 0.419	-0.468 ± 0.232*	-0.202 ± 0.802
RF – GM	-0.341 ± 0.322	-0.471 ± 0.128*	-0.452 ± 0.342*
VL - GL	-0.098 ± 0.265	0.095 ± 0.628	0.411 ± 0.755*
VL - GM	-0.166 ± 0.16	0.081 ± 0.472	0.074 ± 0.297
GL - GM	-0.028 ± 0.212	0.086 ± 0.330	-0.123 ± 0.343

*Statistically significant differences ($p < 0.05$)

Regarding the results of the phase difference P in the comparison among sEMG signals, presented in Tab. 6, it was possible to detect a generalized trend for all comparisons, with higher means and standard deviations in CMJ relative to DJ and SJ ($p < 0.05$)*, corresponding to lower phase agreement of acquired sEMG signals at long CM in relation to the short CM and in the absence of CM.

TABLE VI. MEAN AND STANDARD DEVIATION FOR THE PAIRWISE PHASE DIFERENCE $|P|$ BETWEEN MUSCLES SEMG AT CMJ, DJ, SJ.

***P* - sEMG**	**CMJ**	**DJ**	**SJ**
VM - RF	0.121 ± 0.128*	0.019 ± 0.013	0.070 ± 0.058
VM - VL	0.107 ± 0.108*	0.025 ± 0.027	0.053 ± 0.061
VM - GL	0.159 ± 0.092*	0.053 ± 0.054	0.045 ± 0.037
VM - GM	0.152 ± 0.142*	0.026 ± 0.018	0.042 ± 0.019
RF - VL	0.125 ± 0.184*	0.028 ± 0.024	0.033 ± 0.012
RF - GL	0.231 ± 0.143*	0.053 ± 0.033	0.095 ± 0.040
RF - GM	0.217 ± 0.189*	0.028 ± 0.013	0.067 ± 0.036
VL - GL	0.170 ± 0.043*	0.066 ± 0.042	0.071 ± 0.039
VL - GM	0.156 ± 0.064*	0.034 ± 0.018	0.051 ± 0.032
GL - GM	0.066 ± 0.055*	0.037 ± 0.035	0.034 ± 0.019

*Statistically significant differences ($p < 0.05$)

2) sEMG correlations and amplitudes

In terms of the associations between the sEMG signals assessed on the basis of bivariate correlations relating to the selected muscle groups from the lower limbs, CMJ showed a general trend towards higher correlation values and lower dispersion ($p < 0.05$)* compared with DJ and SJ for all the muscle pairs presented in Fig. 5. Higher correlations were also detected ($p < 0.05$)* for RF–VL across the various types of countermovement (CMJ, DJ and SJ) and for GL–GM in CMJ and SJ, as well as lower correlation levels for RF–GL, RF–GM and VL–GL in CMJ, DJ and SJ.

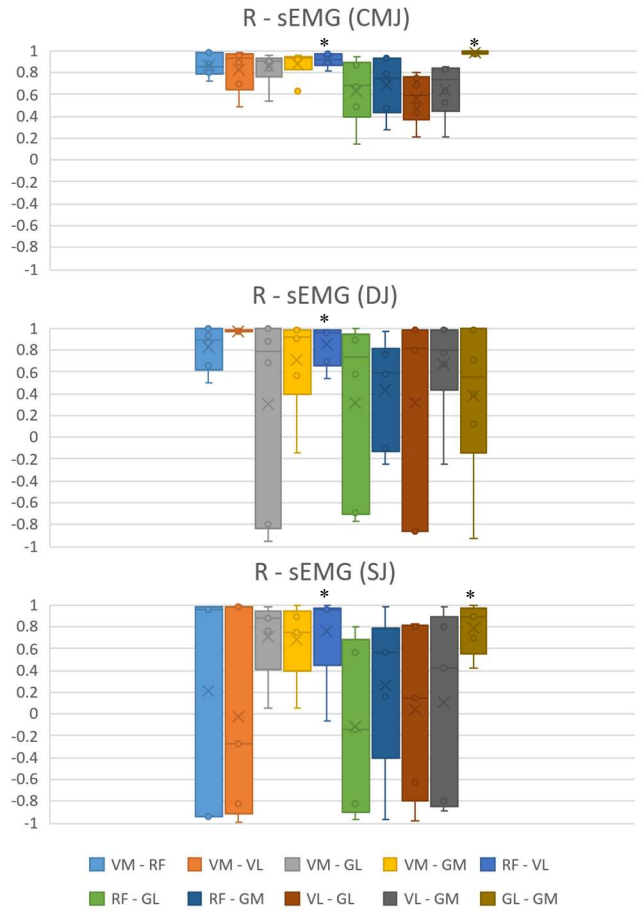


Fig. 5 Correlations of the sEMG signals for the main muscle groups during CMJ, DJ and SJ respectively

The sEMG registered signals for the selected lower limb muscles presented higher mean values ($p < 0.05$)* for DJ compared with SJ, and both were higher than CMJ (Tab. 7). DJ also showed greater variability than CMJ, with both being higher than SJ. Furthermore, RF showed higher sEMG mean values when compared with the other selected lower limb muscle groups.

TABLE VII. MEAN AND STANDARD DEVIATION OF SEMG AT CMJ, DJ AND SJ.

sEMG	CMJ	DJ	SJ
VM	0.381 ± 0.167	0.684 ± 0.395*	0.486 ± 0.136
RF	0.572 ± 0.338*	1.031 ± 0.487*	0.740 ± 0.283*
VL	0.342 ± 0.174	0.588 ± 0.313*	0.395 ± 0.117
GL	0.259 ± 0.224	0.539 ± 0.218*	0.503 ± 0.248
GM	0.250 ± 0.172	0.568 ± 0.265*	0.424 ± 0.158

Regarding the comparison of the results from the sEMG magnitude M and the corresponding sEMG mean values, both metrics presented higher activity ($p < 0.05$) for the RF muscle, pointing to the consistency of both metrics in agreement with detected results for the CMA analysis. Additionally, direct comparison of the mean sEMG led to detection of a detailed differentiation for the assessed muscle groups, with transversal higher values at DJ in relation to SJ, and both higher than CMJ. Finaly, the sEMG phase differences led to detection of higher P values at CMJ, with higher mean of sEMG bivariate correlations at CMJ and the lower limb synergistic muscles in relation to the antagonistic.

IV. DISCUSSION AND CONCLUSIONS

The results of the comparison of CMA activations led to the detection of higher M magnitudes in CMJ, DJ and SJ for the quadriceps muscle groups RF, VLs, VMm and VMs compared to the gastrocnemius GL and GM in line with the results of the comparison of sEMG activations. Similarly, with regard to the P phase difference, CMJ, DJ, and SJ showed higher values among the RF, VL, and VM muscle groups of the quadriceps in relation to the gastrocnemius GL and GM, with a second group of intermediate phase differences P between VL and VM compared to RF, and reduced phase differences between VL and VM. The magnitude and phase metrics thus allowed us to detect in CMJ, DJ, and SJ a greater association between the CMA activations of muscles belonging to the same muscle groups.

Additionally, the correlations between CMA activations with higher values detected between the gastrocnemius muscles GL-GM on the one hand, and between the quadriceps muscles RF1-RF2, VL-VM, RF-VL, and RF-VM on the other, point to the quantitative validation of the agonist behavior between the muscles of the same group, just as the negative correlations detected between the CMA activations of GL, GM with RF, VL, VM point to the validation of the antagonist behavior indicated by the respective CMA temporal profiles.

Focusing specifically on the comparison between sEMG signals, the greatest difference in magnitude M was registered between VM-RF for CMJ, DJ and SJ with higher RF values compared to VM. Additionally, DJ showed differences with higher sEMG magnitudes for RF relative to VL, GL, GM, as well as SJ with higher magnitude differences for RF relative to VL, GM, thus pointing to a generalized trend of higher sEMG activity in RF relative to the other muscle groups, with some variations between CMJ, DJ and SJ. Regarding the P phase difference between sEMG activations, the higher values detected in CMJ compared to DJ and SJ may be associated with the longer duration of the impulse phases at long CM, with the higher simultaneity of muscle recruitment in DJ and SJ related to the lower naturalness of the DJ and SJ impulses compared to CMJ.

The bivariate correlations between the sEMG signals of the selected muscles in the lower limbs allowed us to differentiate long CM in CMJ with higher correlation levels compared to short CM in DJ, and at the absence of the countermovement in SJ. Additionally, the higher correlation levels detected between RF-VL for CMJ, DJ, and SJ, and GL-GM for CMJ and SJ, point to agonist behavior between muscles of the same group, as opposed to the antagonist behavior between muscles of different groups associated with the lower correlation values detected between RF-GL, RF-GM, and VL-GL for CMJ, DJ, and SJ.

Regarding the mean values of the sEMG signals for the selected muscles of the lower limbs, a trend was observed with higher values in DJ relative to SJ and CMJ, as well as higher values for RF relative to VM, VL, GL, and GM at DJ, SJ, and CMJ. These trends were additionally registered for the CMA activations with higher mean values at DJ relative to SJ and CMJ, as well as higher values in RF1, RF2 relative to VLs1, VLs2, and both higher than the other muscles VMm, VMi, VMs, VLi, GL, and GM. Similarly, the correlations between sEMG activations with higher values for long CM in CMJ relative to DJ and SJ presented greater amplitudes for synergist muscles belonging to the same muscle group and lower amplitudes for the antagonistic muscles belonging to different muscle groups, in line with what was detected for the correlations between the corresponding CMA activations. The sEMG and CMA activations also showed alignment with respect to higher amplitudes in DJ and SJ compared to CMJ, with greater simultaneity of muscle recruitment for short CM and in the absence of CM compared to long CM, as well as higher mean sEMG and CMA values in RF compared to the other lower limb muscles selected for analysis.

Thus, the correlations between the muscle activations separately applied to each of the sEMG and CMA measures allowed for better detection of synergistic activations between muscles of the same muscle group and antagonistic activations between muscles of different groups, as well as the greater degree of synergy assessed through correlations with higher values for long CM. Finally, the amplitudes assessed based on the average values of the sEMG and CMA signals led to the detection, by both methods, of greater activity for the rectus femoris, as well as higher intensities for short CM on DJ and without CM on SJ in relation to long CM on CMJ.

According to the obtained results, it is possible to point out the magnitude and phase metrics as adequate for comparing sEMG and CMA activations related to the same muscle groups, based on the differences in amplitudes and time lag assessed on each of the methods, while separately at each of the sEMG measurements and computed CMA, the assessments of correlations and amplitudes allow for a more detailed analysis.

sEMG and CMA are distinct quantities, obtained through different methods. Thus, while sEMGs correspond to muscle activations collected via surface electromyography, CMAs correspond to muscle activations calculated from inverse dynamics, based on optimization criteria used to address the issue of muscle redundancy. Thus, while sEMG signals relate to the activations of the active muscle components, the CMA calculated through the computed muscle control relate to the muscle-tendon unit (MTU) as a whole, corresponding to the association of the three components, the contractile elements (CE) of the muscle in parallel with the passive elements (PE), both in series with the tendon. These differences in the approaches also support the predominance of positive correlations between sEMG activations as opposed to the CMAs positive correlations of synergist muscles and negative correlations for antagonist muscles. The similarities between CMAs for the different types of countermovement CMJ, DJ, and SJ can thus be associated with the type of model and the optimization conditions applied at the MSK-MS, while the differences between CMs at the sEMG patterns result from a process associated with the direct measurement of muscle activity recorded at the skin surface in explosive movements such as the CMJ, DJ and SJ.

According to the redundancy of muscle actions at each joint and the agonistic and antagonistic activations in response to increased external load, as well as the existence of upper limits of muscle load, with the distribution of the higher force levels among the largest muscles, the second-order polynomial optimization criterion was used, prioritizing on the one hand the efficiency corresponding to the minimization of energy expenditure and considering the muscle volume that favours the activation of larger muscles, ensuring the non-linear nature of the activation without excessively penalizing a single muscle and contributing to the overall energy economy in stereotyped movements, while on the other hand reconciling the existence of synergies and allowing a considerable level of overload of the larger muscles.

The comparison of CMA activations using sEMG measurements based on magnitude and phase metrics is thus presented as a form of evaluation, towards improving the techniques and parameters used with a view to the realistic calculation of CMA, as well as the use of a greater number of sEMG measurement channels for validation of the obtained CMA. In addition to the increase on the number of channels for the sEMG signals acquisition, the utilization of different cost functions, namely the min/max and the polynomial with different degrees, can contribute to the objective of realistic CMA predictions taking into account the different nature of the sEMG and CMA signals for the selection of the comparison metrics, with the correlations and amplitudes proving adequate for the analysis at each of the measured sEMG and computed CMA, with the need for more flexible metrics such as the magnitude M and phase P for direct comparison between CMA and sEMG.

REFERENCES

[1] D. A. Winter, Biomechanics and Motor Control of Human Movement, 3rd ed. New Jersey: John Wiley & Sons, 2005.

[2] V. M. Zatsiorsky and W. J. Kraemer, Science and Practice of Strength Training, 2nd ed. Champaign, IL: Human Kinetics, 2006, pp. 33-35.

[3] P. V. Komi, Strength and Power in Sport, 2nd ed. Encyclopaedia of Sports Medicine, vol. III. Oxford, UK: Blackwell Science Ltd., 2003.

[4] P. V. Komi, M. Ishikawa, and, V. Linnamo, "Identification of Stretch-Shortening Cycles in different sports," Portuguese Journal of Sports Sciences, vol. 11(Suppl. 2), 31-34, Jun. 2011.

[5] M. F. Bobbert, M. Mackay, D. Schinkelshoek, P. A. Huijing, and, G. J. van Ingen Schenau, "Biomechanical analysis of drop and countermovement jumps," European Journal of Applied Physiology and Occupational Physiology, vol. 54(6), 566-573, Feb. 1986.

[6] J. J. McMahon, "Kinetic assessment of vertical jumping: application to training and monitoring athletes sports," Sports, Special Issue Editorial, Jun. 2018.

[7] C. Rodrigues, M. Correia., J. Abrantes, J. Nadal and, M. Benedetti, "Combined phase and magnitude metric for validation of lower limb multibody dynamics muscle action with sEMG," IFMBE Proceedings, vol. 68(2), pp. 517-521, May 2018 [World Congress on Medical Physics and Biomedical Engineering 2018. Springer, Singapore]. https://doi.org/10.1007/978-981-10-9038-7_95

[8] E. De Pieri, M. E. Lund, A. Gopalakrishnan, K. P. Rasmussen, D. E. Lunn, and S. J. Ferguson, "Refining muscle geometry and wrapping in the TLEM 2 model for improved hip contact force prediction," PLoS One, vol. 13(9), e0204109, Sep. 2018.

[9] M. D. K. Horsman, H. F. Koopman, F. C. van der Helm, L. P. Prose and H. E. Veeger, "Morphological muscle and joint parameters for musculoskeletal modelling of the lower extremity," Clin Biomech (Bristol), vol. 22(2), 239-247, Feb. 2007.

[10] V. Carbone, R. Fluit, P. Pellikaan, M. M. van der Krogt, D. Janssen, M. Damsgaard, L. Vigneron, T. Feilkas, H. F. Koopman and, N. Verdonschot, "TLEM 2.0 - a comprehensive musculoskeletal geometry dataset for subject-specific modeling of lower extremity," Journal of Biomechanics, vol. 48(5), 734-741, Mar. 2015.

Development of an Interactive Web Simulator for Teaching Computational Electrophysiology

1st Mickaell de Oliveira
Computer Science Department
Federal University of Juiz de Fora
Juiz de Fora, Brazil
0009-0004-2678-5069

2nd Daniel Moreira
Computer Science Department
Federal University of Juiz de Fora
Juiz de Fora, Brazil
0009-0007-1779-0167

3rd Joventino de Oliveira Campos
Computer Science Department
Federal University of Juiz de Fora
Juiz de Fora, Brazil
0000-0003-1899-9441

***Abstract*—Teaching computational cardiac electrophysiology is challenging due to the mathematical complexity of biophysical models and the dependence on high-performance computing infrastructures. To address these barriers, this work presents a free, interactive, and fully client-side web simulator designed for educational purposes. Methodologically, the application employs a hybrid acceleration architecture: it utilizes Web Workers for robust multithreaded numerical integration in 0D and 1D environments, and leverages the WebGPU API to execute highly parallelized spatial computations for 2D reaction-diffusion models directly on the user's hardware. Performance benchmarks demonstrate that the WebGPU implementation achieves execution speedups of up to $395\times$ compared to standard CPU processing. These optimizations enable the real-time visualization of phenomena such as spiral waves and source-sink mismatch on personal computer hardware. Early user feedback highlights the tool's intuitive interface as an effective way to lower the barrier to entry for exploring electrophysiological concepts, motivating future formal studies on its pedagogical impact.**

***Index Terms*—computational electrophysiology, web simulator, education, cardiac modeling, WebGPU**

I. Introduction

Computational modeling has revolutionized the understanding of cardiac electrophysiology by enabling the simulation of complex scenarios that approximate clinical reality [1]. These tools offer predictive capabilities regarding cardiac conditions and the mechanisms of arrhythmias without the need for invasive procedures [2].

Despite this potential, the widespread adoption of computational tools in educational environments faces significant challenges. The models rely on the numerical solution of ordinary differential equations (ODEs) and partial differential equations (PDEs) [3], which imposes a steep learning curve. Furthermore, high-fidelity simulations frequently demand high-performance computing and specialized knowledge in information technology for the installation and configuration of software environments [4]. This complexity creates a barrier for students and medical educators who wish to experiment with biophysical models without the need to manage complex infrastructures.

In this context, web-based solutions have emerged as an alternative to mitigate accessibility issues. Some approaches use WebGL and WebGPU to accelerate calculations through the GPU [5], [6], while others offer web interfaces for established *backends*, such as OpenCARP [7]. However, these solutions may still demand high computational resources or rely on server-side processing.

Thus, this work presents a fully client-side web simulator designed for educational purposes. Developed with React, the tool leverages WebGPU to optimize demanding simulations, while utilizing Web Workers for parallel processing as a robust fallback. This architecture ensures the simulator remains entirely functional even on devices without WebGPU support, without ever relying on server-side processing. By simplifying the execution and visualization of electrophysiological models, the tool eliminates the need for advanced programming skills or specific hardware configurations. This approach seeks to bridge the gap between mathematical complexity and accessibility, allowing students to explore cardiac dynamics in an intuitive manner. Furthermore, the simulator is provided as an open-source tool with versioned source code and a clear license to ensure reproducibility, browser compatibility, accessibility, and long-term maintenance.

II. Related Works

The development of computational tools for cardiac electrophysiology has traditionally focused on high-performance computing (HPC) environments. Established software packages like openCARP [7] offer robust, state-of-the-art simulation capabilities for clinical and research applications. However, their reliance on command-line interfaces, complex dependencies, and server-side processing creates a steep learning curve that limits their use in undergraduate education.

To address accessibility, web-based solutions have gained traction. Kaboudian et al. [5] demonstrated the potential of WebGL for interactive simulations of complex ionic models, while Rosales et al. [6] recently explored WebGPU to achieve fast simulations of cardiac electrophysiology. While these works successfully leverage client-side GPU acceleration, they are often tailored towards advanced research visualization rather than pedagogical exploration.

The simulator presented in this work distinguishes itself by wrapping the high-performance capabilities of WebGPU within an intuitive, React-based educational interface. It prioritizes ease of use, dynamic parameter adjustment, and imme-

diate visual feedback, bridging the gap between raw computational power and pedagogical utility without requiring specialized hardware. Compared to these existing platforms, our approach emphasizes cross-browser compatibility and minimal resource consumption, offering an accessible implementation of established models with transparent acceleration specifically tailored for pedagogical use.

III. Modeling Cardiac Electrophysiology

The simulation of cardiac electrical activity in this work is based on solving mathematical models that describe both individual cellular excitation and the propagation of the electrical wave through the tissue.

A. Cellular Models

The cardiac action potential (AP) is the fundamental cellular event that triggers the contraction of the heart. It is characterized by a rapid depolarization of the cell membrane, followed by a plateau phase and a subsequent repolarization. To represent this dynamics, the simulator implements three widely used models in the literature: the Mitchell-Schaeffer model, the Minimal Model, and the ten Tusscher model.

The Mitchell-Schaeffer model [8] is a simplified two-variable model designed to capture the fundamental properties of restitution dynamics and the generation of arrhythmias, such as electrical alternans, with low computational cost. The state of the cell is described by the transmembrane potential (v) and a gating variable (h) that regulates the refractoriness of the cell.

The Minimal Model [9] is a more robust formulation composed of four state variables (u, v, w, s). It was developed to accurately reproduce the action potential shape and the properties of different cell types (endocardium, epicardium, and mid-myocardium).

The ten Tusscher model [10] is a detailed biophysical formulation of the human ventricular myocyte. Unlike simplified phenomenological models, it provides a comprehensive description of the underlying ionic currents, pumps, and exchangers. This level of detail allows for the investigation of specific ion channel dysfunctions and complex physiological behaviors.

The behavior of each cell is governed by a system of Ordinary Differential Equations (ODEs). For the numerical solution of this system, the simulator adopts a hybrid approach aiming for a balance between accuracy and stability. The potential variable (v) is integrated using the Forward Euler method [11]. Conversely, the gating variables (h in the MS model; v, w, s in the Minimal Model; and the multiple gating variables in the ten Tusscher model) are integrated using the Rush-Larsen method [12]. The Rush-Larsen method is used for these variables because, by assuming the voltage is constant during the time step, it allows the use of the analytical solution for the gating variable. This guarantees unconditional stability for this part of the system, thereby allowing for larger time steps than the standard Euler method would permit.

B. Electrical Tissue Propagation Model

To model the electrophysiological activity of the tissue, the monodomain model [3] is used. This model assumes that the cardiac tissue behaves as a functional syncytium, where the intracellular and extracellular spaces have proportional conductivities. The propagation of the transmembrane potential is described by a reaction-diffusion partial differential equation (PDE) coupled with the cellular reaction model:

$$\chi \left(C_{\mathrm{m}} \frac{\partial V}{\partial t} \right) + (I_{\mathrm{ion}} + I_{\mathrm{s}}) = \nabla \cdot (\sigma \nabla V), \tag{1}$$

$$\frac{\partial \eta}{\partial t} = f(V, \eta, t), \tag{2}$$

where χ is the surface-to-volume ratio of the cells (mm^{-1}), C_{m} is the membrane capacitance per unit area ($\mu\mathrm{F\,mm}^{-2}$), V is the transmembrane potential, and η is the state variable vector of the cellular model. The term I_{ion} represents the sum of the transmembrane ionic currents (given by the cellular models described previously), I_{s} is the external stimulus current, and σ is the electrical conductivity ($\mathrm{mS\,mm}^{-1}$) of the tissue. At the domain boundaries, no-flux Neumann boundary conditions are applied ($\partial V/\partial n = 0$), ensuring that the system is electrically isolated.

The monodomain equation is solved using the Finite Difference Method (FDM) [11] on a uniform Cartesian grid. The spatial Laplacian operator ($\nabla^2 V$) in two dimensions is approximated using a standard 5-point central difference stencil:

$$\nabla^2 V_{i,j} \approx \frac{V_{i+1,j} + V_{i-1,j} + V_{i,j+1} + V_{i,j-1} - 4V_{i,j}}{\Delta x^2} \tag{3}$$

where Δx represents the uniform spatial resolution ($\Delta x = \Delta y$). To enforce the no-flux boundaries, a ghost-node approach is implemented. During each integration step, the voltage values of the boundary nodes are mirrored to the fictitious nodes outside the domain. The temporal evolution is subsequently solved using the explicit Forward Euler scheme, with the time step Δt strictly bounded by the Courant-Friedrichs-Lewy (CFL) stability condition.

IV. Architecture and Implementation

The application interface was developed modularly using React [13], while styling was implemented with Tailwind CSS [14]. A critical aspect of democratizing educational tools is ensuring cross-device compatibility. Leveraging Tailwind's utility-first framework, the simulator features a fully responsive layout that dynamically adapts to various screen resolutions, including tablets and smartphones. This allows students to access, configure, and run models directly from their mobile web browsers, providing an unprecedented level of portability for ubiquitous learning environments.

However, since solving ODEs and PDEs entails high computational costs, executing these calculations on the main browser thread—especially on mobile processors—would inevitably cause the interface to freeze. To ensure interactivity and performance, a hybrid client-side acceleration architecture

was adopted. For computationally intensive spatial propagation, the simulator leverages the WebGPU API to execute highly parallelized calculations directly on the user's graphics processing unit (GPU). Historically, web-based GPU acceleration relied on WebGL, which forced developers to disguise mathematical state variables as RGB pixel colors within textures, executing calculations via graphics rendering pipelines. Conversely, WebGPU natively introduces Compute Shaders written in the WebGPU Shading Language (WGSL), designed specifically for general-purpose GPU (GPGPU) computing. The computations are performed using 32-bit floating-point precision, which is standard for WebGPU and sufficient for educational visualizations, though strict hardware-dependent variations must be considered.

In this implementation, the simulation domain is discretized and mapped to GPU workgroups. Each thread uses its global invocation ID to fetch the local state of a specific grid node from structured Storage Buffers, computes the spatial Laplacian, integrates the gating variables, and writes the updated state back to memory. To avoid the severe performance penalty of continuously transferring large state matrices between the GPU VRAM and the CPU (readbacks) during each integration time step, the simulator implements an alternating double-buffer strategy natively within the device memory. Data is only read back to the CPU asynchronously when real-time visualization rendering is required, ensuring that the heavy computational loop runs at maximum hardware throughput.

For 0D and 1D simulations, or as a robust fallback for devices lacking WebGPU support, a multithreaded architecture using Web Workers [15] was implemented, ensuring that numerical integration is seamlessly processed in the background CPU threads. Real-time rendering of the simulations required intense graphical optimizations. For 0D and 1D simulations, the Chart.js library [16] was used in conjunction with a *downsampling* technique. For 2D propagation under the Web Worker fallback, the solution involves direct pixel manipulation via the HTML5 Canvas API [17].

V. Simulator Overview

The graphical interface was designed to offer full control over simulation parameters without compromising visual clarity[1]. As depicted in Figure 1, the environment is divided into an intuitive side panel for configuration and a main central canvas for real-time dynamic charting.

In the specific example shown in Figure 1, the simulator is running the 0D Mitchell-Schaeffer model under a single-stimulus protocol. The tool seamlessly plots both the Transmembrane Potential (V) and the corresponding gating variable (h) simultaneously. This allows students to directly observe the mathematical and physiological relationship between these variables: following the stimulus injection, the cell rapidly depolarizes, while the gating variable immediately drops, visually illustrating the onset of the cellular refractory period.

[1]The web simulator can be accessed at: https://web-simulator.github.io.

The ability to toggle the visibility of distinct state variables and instantly observe the impact of modifying core parameters provides a highly interactive feedback loop. This immediate visual response effectively eliminates the "black-box" nature often associated with traditional command-line simulation environments.

A. Restitution Analysis and Action Potential Duration (0D)

Beyond standard transmembrane potential plotting, the simulator features a dedicated module for theoretical Restitution Analysis. By applying a programmed S1-S2 pacing protocol across a broad sweep of coupling intervals, the underlying simulation engine programmatically calculates the Action Potential Duration (e.g., APD_{90}) and the preceding Diastolic Interval (DI) for each cycle.

These internally extracted metrics are utilized to dynamically generate and plot the APD restitution curve (Figure 2). This feature allows students to immediately visualize the slope of the restitution curve without needing external analytical software. Understanding this slope is a crucial pedagogical step, as slopes greater than 1 are mathematically and clinically associated with the onset of electrical alternans and subsequent wave propagation instability.

B. One-Dimensional Cable Equation (1D)

The one-dimensional (1D) module extends the cellular models to a cable geometry, allowing the study of wave propagation and the calculation of Conduction Velocity (CV). The ten Tusscher biophysical model is particularly highlighted in this environment, as illustrated in Figure 3.

In the 1D setting, students can observe spatiotemporal phenomena such as the dispersion of repolarization and the onset of unidirectional waveblock under rapid pacing. Given the computational complexity of the ten Tusscher formulation—which comprises dozens of state variables and coupled ionic gates—the client-side WebGPU and multithreaded architecture provided by the simulator becomes strictly necessary to allow real-time interaction with the cable's stimulation parameters. In addition to standard line charts plotting transmembrane potential over time, the 1D module offers a spatiotemporal heatmap visualization (Figure 4). This color-mapped representation displays the tissue cable along the x-axis and time along the y-axis, providing an intuitive, global view of wave propagation, conduction velocity variations, and blockages at a single glance.

C. Two-Dimensional Electrical Wave Propagation

The transition to two-dimensional simulations, highly optimized by the WebGPU integration, enables the study of complex reentrant mechanisms. Figure 5 illustrates the sequential formation of an arrhythmia (spiral breakup) using the Mitchell-Schaeffer model. The sequence begins with an initial planar wavefront triggered at the lateral domain border (a). A premature S2 stimulus is then applied during the vulnerable refractory window (b), leading to the formation of an organized reentrant spiral wave (c). Due to the restitution

properties configured in the model, the wave eventually undergoes spiral breakup, fragmenting into multiple disorganized wavelets simulating ventricular fibrillation (d).

Beyond rendering the Transmembrane Potential (V) in real-time, the 2D module incorporates advanced spatial mapping visualization capabilities, fully supporting tissue heterogeneity (transmurality). By configuring the tissue with distinct cellular action potential phenotypes, such as mapping the Mid-myocardium starting at 30% and the Epicardium at 60% of the domain width, students can visualize the physiological consequences of transmural dispersion.

Figure 6 presents these capabilities using the Minimal Model. The Local Activation Time (LAT) map (a) displays color-coded isochrones indicating the exact moment the depolarization wavefront reaches each tissue node, which is highly pedagogical for identifying conduction delays. Similarly, the spatial Action Potential Duration (APD) map (b) highlights the explicit boundaries and differences in repolarization times across the endocardial, mid-myocardial, and epicardial layers, a fundamental substrate for arrhythmogenesis.

In pathological scenarios, tissue structure plays a crucial role. Figure 7 presents the effect of diffuse fibrosis on wave propagation, creating non-conducting barriers that fragment the wavefront.

Finally, the source-sink mismatch phenomenon [18] is depicted in Figure 8. This phenomenon occurs when the excitatory charge generated by depolarizing cells (the active "source") is insufficient to raise the membrane potential of the adjacent resting cells (the passive "sink") above the activation threshold. In this simulation, wave propagation is guided through a narrow fibrotic isthmus that suddenly expands into a broader healthy tissue region. The narrow isthmus restricts the number of active source cells, while the sudden structural expansion presents a massive electrotonic load. As demonstrated in the figure, the charge provided by the wavefront is rapidly drained by the large mass of resting cells, resulting in a severe conduction delay and ultimately causing a unidirectional block, a known precursor to reentrant arrhythmias.

D. Performance Benchmarks

To quantitatively evaluate the computational efficiency of the proposed hybrid architecture, performance benchmarks were conducted across two distinct hardware profiles: a standard desktop (Intel Core i5-9400F, 16GB RAM, NVIDIA GTX 1660) and a high-performance workstation (Intel Core i5-12400F, 16GB RAM, NVIDIA RTX 4070). The tests consisted of running the 2D Minimal Model for a biological simulation time of 1000 ms, utilizing a fixed spatial resolution ($\Delta x = 0.1$ mm) and a fixed integration time step ($\Delta t = 0.1$ ms).

To isolate the variable of computational load, the grid size was progressively increased from 128×128 to 1024×1024 nodes. The execution time (from initialization to the rendering of the final state) was measured for both the multithreaded CPU approach (Web Worker) and the highly parallelized GPU approach (WebGPU).

As shown in Tables I and II, the pure JavaScript implementation (Web Worker CPU) scales poorly as the grid size in-

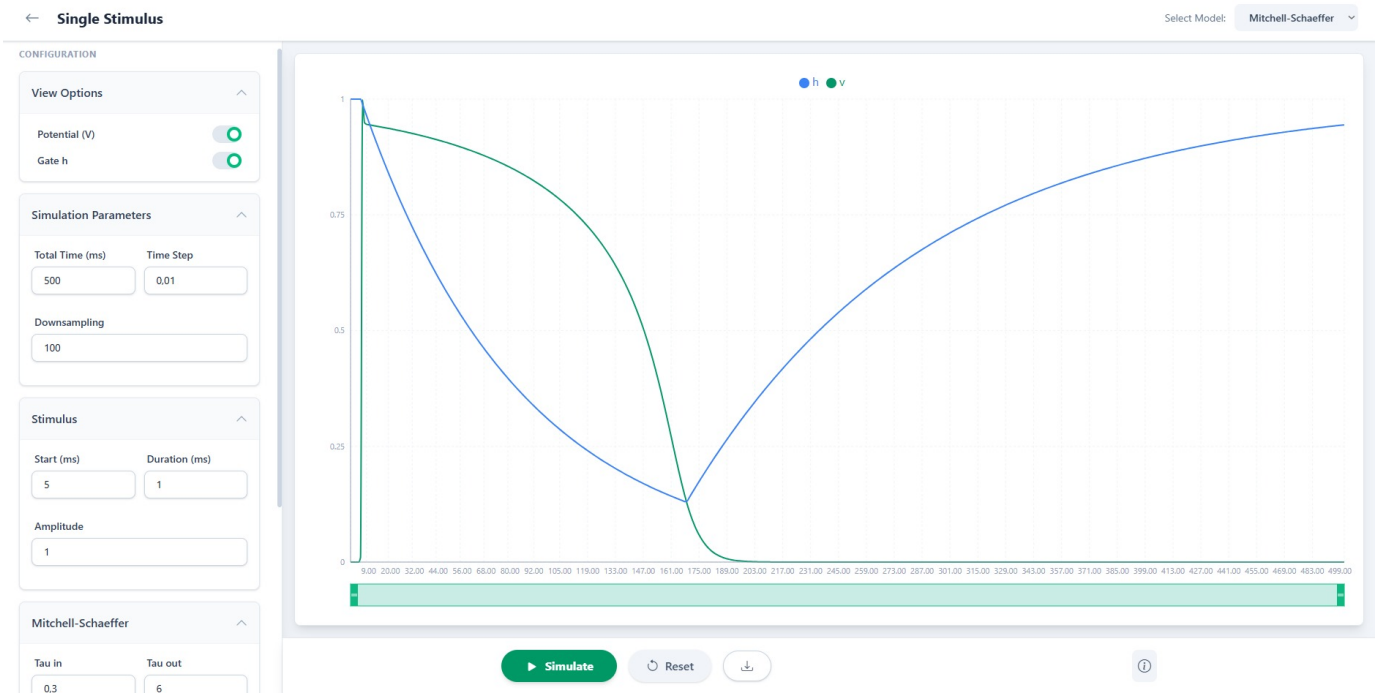


Fig. 1: Simulator interface running the 0D Mitchell-Schaeffer model under a single-stimulus protocol, showcasing the complete parameter control panel and real-time visualization of state variables.

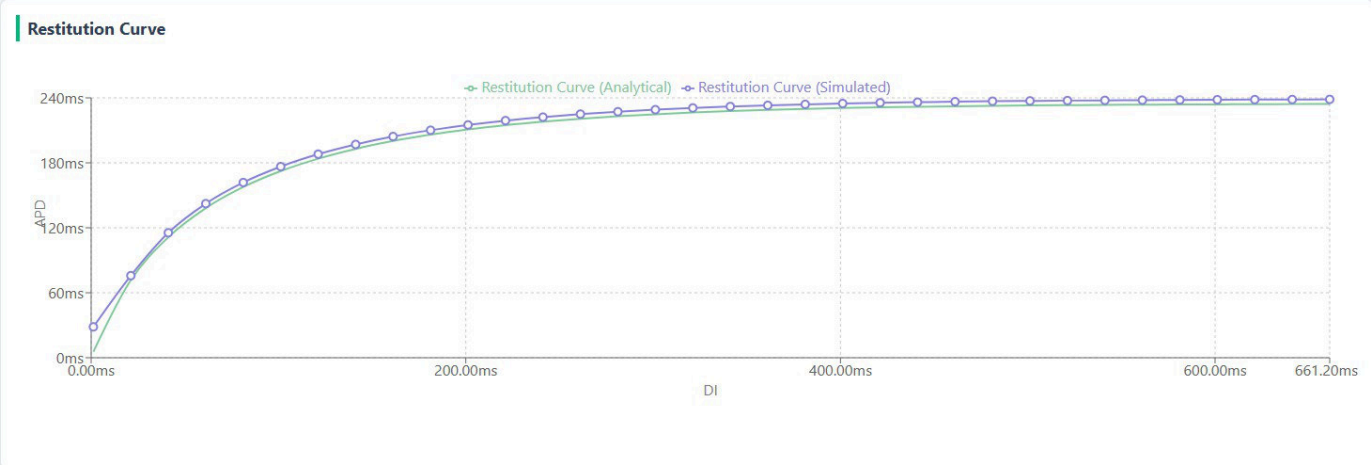


Fig. 2: Action Potential Duration (APD) Restitution Curve generated dynamically via an automated S1-S2 protocol sweep.

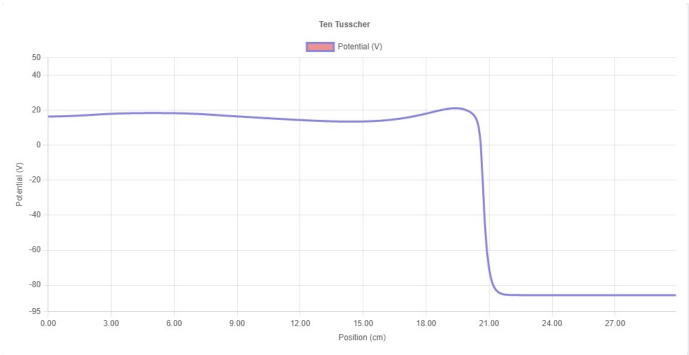


Fig. 3: One-dimensional wave propagation simulated along a cable using the ten Tusscher model.

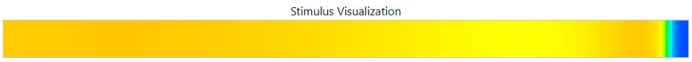
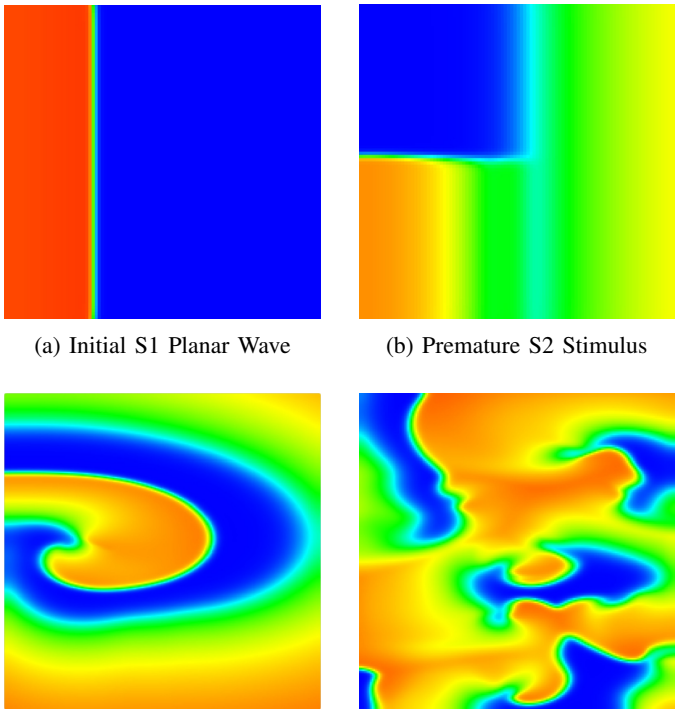


Fig. 4: Spatiotemporal heatmap of 1D wave propagation, illustrating the voltage distribution across the cable over time.



(a) Initial S1 Planar Wave (b) Premature S2 Stimulus

(c) Sustained Spiral Wave (d) Spiral Breakup (Fibrillation)

Fig. 5: Sequential development of an arrhythmia using the 2D Mitchell-Schaeffer model.

TABLE I: Personal Computer Benchmark

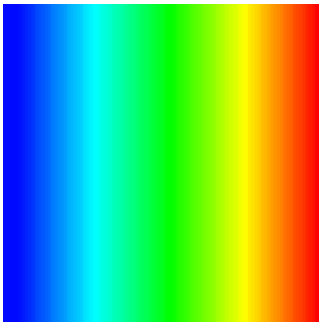
Grid Size	CPU Time (s)	GPU Time (s)	Speedup
128×128	21.14	1.74	$\sim$12x
256×256	82.16	2.23	$\sim$37x
512×512	329.58	2.80	$\sim$117x
1024×1024	1377.00	5.05	$\sim$272x

TABLE II: Workstation Benchmark

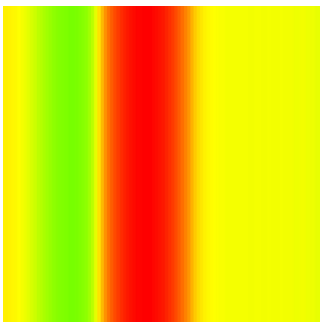
Grid Size	CPU Time (s)	GPU Time (s)	Speedup
128×128	16.00	0.95	$\sim$17x
256×256	61.00	1.38	$\sim$44x
512×512	247.00	1.64	$\sim$150x
1024×1024	1060.00	2.68	$\sim$395x

creases. Even on the modern i5-12400F processor, simulating a grid with over one million nodes required nearly 18 minutes (1060 seconds), rendering high-resolution 2D simulations highly impractical for interactive classroom environments.

In contrast, the WebGPU implementation effectively exploits the massively parallel nature of the reaction-diffusion equations. On the GTX 1660, the GPU execution time scaled remarkably well, remaining at roughly 5 seconds for the largest grid. When utilizing the laboratory-grade RTX 4070, the simulation completed in 2.68 seconds, yielding a maximum



(a) LAT Map



(b) APD Transmural Map

Fig. 6: Spatial visualization maps illustrating tissue heterogeneity (Endocardium, Mid-myocardium, Epicardium) generated via the Minimal Model.

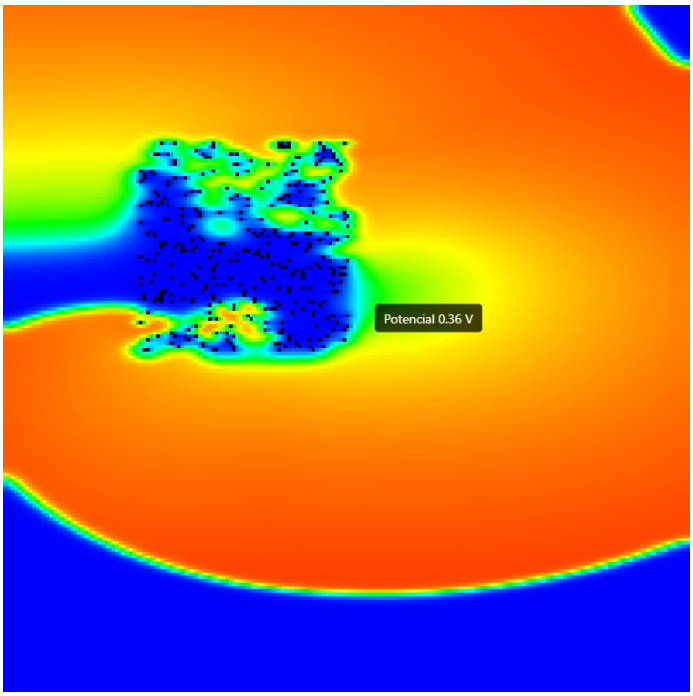


Fig. 7: Simulation of wave propagation in tissue with diffuse fibrotic regions, causing wavefront fragmentation.

speedup of nearly $395\times$ compared to its CPU counterpart. These results demonstrate the simulator's capability to deliver fluid, real-time interactivity on accessible hardware. Future work must include rigorous benchmarking with multiple repetitions, as well as mathematical validation of convergence against trusted ODE and PDE reference solutions, reporting temporal and spatial errors, stability limits, and boundary tests over multiple parameter sets.

E. Simulator Evaluation

To gather initial feedback on the usability of the tool, the simulator underwent an informal evaluation by a preliminary sample of users ($N = 5$), comprising undergraduate students and professors. Data collection was performed through an

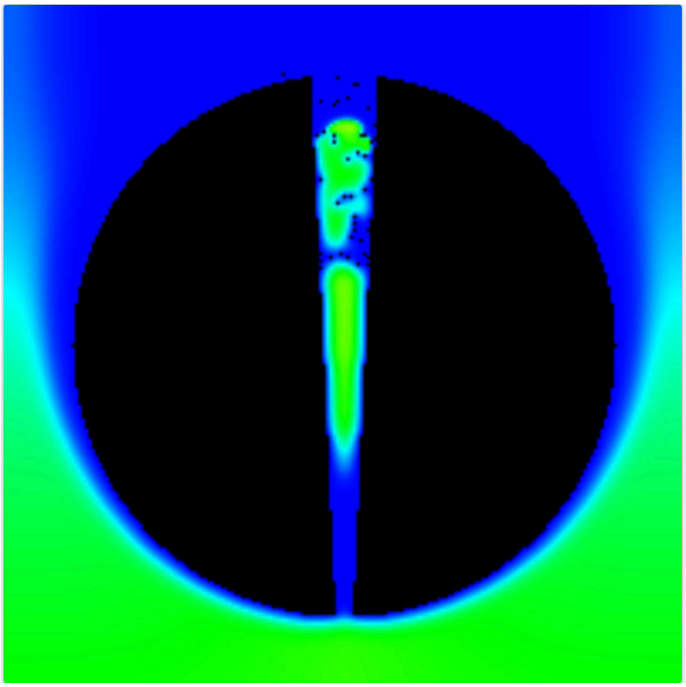


Fig. 8: Simulation of the *source-sink mismatch* phenomenon through a narrow isthmus.

anonymous ad-hoc form, utilizing a 1-to-5 Likert scale to evaluate different aspects of the interface, alongside open-ended questions for qualitative feedback.

The initial quantitative results indicated user satisfaction regarding the interface. Across the responses, users provided positive scores in metrics including the ease of navigation between different mathematical models, the clarity of parameter input fields, and the ease of understanding the 1D and 2D graphical visualizations. From a qualitative perspective, users reported that the tool facilitates the visualization of abstract concepts, highlighting the 2D propagation and Source-Sink modules as particularly useful.

While these findings are encouraging, this preliminary assessment lacks the scope to objectively quantify the educational impact of the tool. A controlled pre-post comparative teaching study is necessary to formally evaluate actual learning outcomes.

A noted limitation is the hardcoded nature of the mathematical models. Future work will focus on integrating a dynamic model parser capable of interpreting standardized formats, such as CellML, allowing users to import and simulate any published cellular model without modifying the application's source code. Additionally, extending the WebGPU compute pipeline to support three-dimensional (3D) anatomical geometries represents the next milestone in bridging the gap between educational tools and advanced research environments.

Overall, the initial evaluation indicates that the simulator serves as an accessible implementation of established models, though further structural validation is required.

VI. Conclusion

This work presented an interactive web simulator dedicated to teaching cardiac electrophysiology. The adoption of a hybrid acceleration architecture, combining WebGPU for highly parallelized spatial computations and Web Workers for robust multithreaded numerical integration, enabled the fluid execution of complete mathematical models entirely within the browser, eliminating the need for costly hardware infrastructures. The results demonstrated the tool's capability to visually reproduce cellular heterogeneity and pathological dynamics, such as spiral waves and fibrosis-induced conduction blocks. Consequently, the application shows promise in achieving its goal of lowering the barriers to accessing computational modeling. Future efforts will focus on rigorous mathematical verification and formal pedagogical assessments through comparative teaching studies.

References

[1] L. Wang and B. Rodriguez, "The promise of cardiomyocyte computational models for drug safety," *Military Medical Research*, vol. 12, no. 1, p. 88, 2025.

[2] M. Strocchi, D. J. Hammersley, B. P. Halliday, S. K. Prasad, and S. A. Niederer, "Cardiac digital twins: a tool to investigate the function and treatment of the diabetic heart," *Cardiovascular Diabetology*, vol. 24, no. 1, p. 293, 2025.

[3] J. Sundnes, G. T. Lines, X. Cai, B. F. Nielsen, K.-A. Mardal, and A. Tveito, *Computing the electrical activity in the heart*. Springer Science & Business Media, 2007, vol. 1.

[4] L. Berg, R. Oliveira, J. Camps, L. de Lima, J. de Oliveira Campos, Z. Wang, R. Doste, A. Bueno-Orovio, R. Dos Santos, and B. Rodriguez, "Toward cardiac electrophysiology digital twins with an efficient open source scalable solver on gpu clusters," *Scientific Reports*, vol. 16, no. 1, p. 9619, 2026.

[5] A. Kaboudian, E. M. Cherry, and F. H. Fenton, "Gpu load balancing using sparse cartesian grids: Making interactive webgl simulations of complex ionic models even faster on 3d heart structures," in *2023 Computing in Cardiology (CinC)*, vol. 50, 2023, pp. 1–4.

[6] R. M. Rosales, A. Mincholé, and E. Pueyo, "Cocoro: Fast simulation of cardiac electrophysiology with webgpu," in *2025 Computing in Cardiology (CinC)*, 2025.

[7] G. Plank, A. Loewe, A. Neic, C. Augustin, Y.-L. Huang, M. A. Gsell, E. Karabelas, M. Nothstein, A. J. Prassl, J. Sánchez *et al.*, "The opencarp simulation environment for cardiac electrophysiology," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106223, 2021.

[8] C. C. Mitchell and D. G. Schaeffer, "A two-current model for the dynamics of cardiac membrane," *Bulletin of mathematical biology*, vol. 65, no. 5, pp. 767–793, 2003.

[9] A. Bueno-Orovio, E. M. Cherry, and F. H. Fenton, "Minimal model for human ventricular action potentials in tissue," *Journal of Theoretical Biology*, vol. 253, no. 3, pp. 544–560, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022519308001690

[10] K. H. W. J. ten Tusscher, D. Noble, P. J. Noble, and A. V. Panfilov, "A model for human ventricular tissue," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 286, no. 4, pp. H1573–H1589, 2004.

[11] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

[12] S. Rush and H. Larsen, "A practical algorithm for solving dynamic membrane equations," *IEEE Transactions on Biomedical Engineering*, vol. BME-25, no. 4, pp. 389–392, 1978.

[13] A. Boduch, *React and react native*. Packt Publishing Ltd, 2017.

[14] K. Bhat, *Ultimate Tailwind CSS Handbook: Build sleek and modern websites with immersive UIs using Tailwind CSS*. Orange Education Pvt Limited, 2023.

[15] I. Green, *Web workers: Multithreaded programs in javascript*. " O'Reilly Media, Inc.", 2012.

[16] H. Da Rocha, *Learn Chart.js: Create interactive visualizations for the web with Chart.js 2*. Packt Publishing Ltd, 2019.

[17] S. Fulton and J. Fulton, *HTML5 canvas: native interactivity and animation for the web*. " O'Reilly Media, Inc.", 2013.

[18] F. O. Campos, Y. Shiferaw, R. Weber dos Santos, G. Plank, and M. J. Bishop, "Microscopic isthmuses and fibrosis within the border zone of infarcted hearts promote calcium-mediated ectopy and conduction block," *Frontiers in Physics*, vol. 6, p. 57, 2018.

Pipeline Computacional para Gerar Gêmeos Digitais Cardíacos em um Sistema Langendorff

Daniel Moreira
Graduação em Engenharia Elétrica
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora, Brasil
https://orcid.org/0009-0007-1779-0167

José Gomes
HEartLab, Engenharia Biomédica
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
https://orcid.org/0009-0005-6637-0949

Tainan Neves
HEartLab, Engenharia Biomédica
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
https://orcid.org/0000-0002-9331-6690

Mayra Urbieta
Programa de Pós-Graduação em Modelagem Computacional
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora, Brasil
https://orcid.org/0009-0009-1777-6474

Lucas Arantes Berg
Programa de Pós-Graduação em Modelagem Computacional
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora, Brasil
https://orcid.org/0000-0002-8777-0125

João Salinet
HEartLab, Engenharia Biomédica
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
https://orcid.org/0000-0003-2906-9397

Joventino de Oliveira Campos
FISIOCOMP, Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora, Brasil
https://orcid.org/0000-0003-1899-9441

Rodrigo Weber dos Santos
FISIOCOMP, Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora, Brasil
https://orcid.org/0000-0002-0633-1391

***Resumo*—Este estudo apresenta o desenvolvimento e a validação de um fluxo de trabalho computacional para a geração de gêmeos digitais cardíacos, integrando dados anatômicos de ressonância magnética (RM) e registros eletrofisiológicos *ex vivo* obtidos por meio de um sistema de perfusão de Langendorff. O fluxo de trabalho proposto abrange o processamento morfológico tridimensional, uma simulação biomecânica de expansão passiva para o corregistro espacial e a calibração da dinâmica celular com base no mapeamento óptico. Para reconstruir o complexo QRS, adotou-se uma abordagem híbrida: a inferência das velocidades de condução e dos pontos de ativação no endocárdio utilizando a equação Eikonal otimizada pelo algoritmo de Evolução Diferencial, seguida por uma simulação biofísica através da equação do monodomínio. Os resultados demonstram que a etapa de inferência reproduziu o comportamento eletrofisiológico experimental, alcançando coeficientes de correlação elevados. Embora o modelo do monodomínio tenha apresentado divergências morfológicas residuais devido à premissa de homogeneidade celular transmural, o alinhamento das velocidades de condução confirmou a validade da propagação espacial. Esse fluxo de trabalho consolida uma plataforma para a eletrofisiologia cardíaca personalizada, impulsionando o desenvolvimento de ensaios clínicos *in silico* e apoiando a redução da experimentação animal em pesquisas biomédicas.**

***Palavras-chave*—Gêmeo Digital, Equação Eikonal, Monodomínio, Simulação Computacional.**

I. Introdução

As doenças cardiovasculares representam uma das principais causas de mortalidade globalmente [3]. Na busca por estratégias diagnósticas e terapêuticas mais eficazes, a medicina de precisão tem se apoiado cada vez mais na modelagem computacional. Neste contexto, o conceito de gêmeo digital cardíaco emerge como uma representação computacional capaz de integrar dados anatômicos, fisiológicos e hemodinâmicos para emular a dinâmica do órgão de forma individualizada, abrindo novos horizontes para o planejamento terapêutico e para a predição de risco [6], [7].

Para a investigação detalhada dos mecanismos arritmogênicos e da eletrofisiologia cardíaca basal sem interferência sistêmica, o modelo de coração isolado e perfundido retrogradamente de Langendorff constitui um padrão experimental bem estabelecido [8]. Contudo, a tradução direta de sinais elétricos adquiridos *ex vivo*, como o mapeamento óptico panorâmico de fluorescência [5] e eletrogramas de superfície, para um gêmeo digital cardíaco apresenta desafios matemáticos e computacionais significativos. O principal gargalo reside na resolução do problema inverso: estimar a velocidade de condução elétrica no tecido e mapear os locais de ativação mais precoces no endocárdio a partir de sinais elétricos globais [19].

A simulação computacional da propagação do potencial de ação no miocárdio é realizada através da resolução de equações diferenciais acopladas, como os modelos do bidomínio ou do monodomínio [10], [11]. Embora esses modelos forneçam alta fidelidade biofísica e capturem a dinâmica não linear das correntes iônicas, a personalização direta de seus parâmetros topológicos e elétricos é um processo iterativo cujo custo computacional é frequentemente proibitivo. Em contrapartida, modelos de ordem reduzida baseados na equação Eikonal permitem o cálculo rápido dos tempos de ativação local (LAT), viabilizando o uso de algoritmos metaheurísticos para a inferência paramétrica de propriedades ventriculares essenciais [19], [20].

Neste contexto, o presente estudo propõe o desenvolvimento e a validação de um *pipeline* computacional híbrido para a geração de um gêmeo digital cardíaco de coelho. A metodologia integra dados estruturais de ressonância magnética (RM) combinados a uma simulação mecânica de expansão passiva para garantir o corregistro espacial preciso com a fotogrametria do órgão [9]. Após o alinhamento geométrico, a modelagem Eikonal associada ao algoritmo de Evolução Diferencial (DE) é utilizada para inferir as velocidades de condução macroscópicas e os locais de ativação mais precoces (*root nodes*) diretamente a partir do complexo QRS extraído dos eletrogramas experimentais.

Por fim, os parâmetros anatômicos e de condução inferidos são transpostos para um simulador de monodomínio para emular a despolarização tecidual utilizando um modelo celular [13]. O objetivo principal é demonstrar que a integração metodológica do processamento de imagens, da modelagem biomecânica passiva e da inferência Eikonal oferece uma plataforma eficiente para calibrar simulações biofísicas complexas. A consolidação deste *pipeline* tecnológico visa impulsionar a confiabilidade dos ensaios clínicos *in silico*, aprimorando o teste de fármacos cardiovasculares e mitigando a dependência de modelos animais na pesquisa biomédica primária.

II. Materiais e Métodos

A. Protocolo Experimental e Aquisição de Dados

Os protocolos experimentais conduzidos neste estudo foram aprovados pelo Comitê de Ética no Uso de Animais (CEUA) local, sob o protocolo número 3947230519. Utilizou-se uma fêmea de coelho da raça Nova Zelândia (massa corporal de 3,1 kg; dimensões cardíacas de $3,5 \times 2,5$ cm e peso do órgão de 8,0 g). A indução anestésica foi realizada com cloridrato de buprenorfina (0,05 mg/kg) via injeção intramuscular. Após 30 minutos, administrou-se uma combinação de cetamina (100 mg/kg) e xilazina (14 mg/kg), seguida pela infusão intravenosa de heparina (1.250 U/kg). A extração do coração foi realizada via toracotomia esquerda, seguida por sua imediata canulação em um sistema de perfusão de Langendorff. O órgão foi perfundido retrogradamente pela aorta utilizando uma solução de Krebs modificada, mantida a 37,5 °C e oxigenada com carbogênio (95% de O_2 e 5% de CO_2).

1) Ressonância Magnética: Os dados de ressonância magnética (RM) foram adquiridos utilizando um sistema Siemens MAGNETOM de 7 Tesla. Para garantir a estabilidade geométrica durante o escaneamento, o coração foi acomodado em um recipiente preenchido com espuma e Fomblin, um fluido que também foi injetado nas câmaras cardíacas para evitar a flutuação e mitigar artefatos de imagem (Fig. 1).

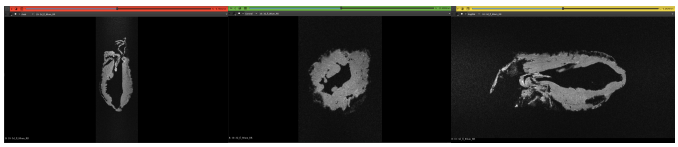


Figura 1. Cortes de RM adquiridos nas visões sagital, coronal e axial.

2) Mapeamento Óptico Panorâmico: Para a reconstrução da superfície cardíaca e análise da dinâmica elétrica, realizou-se o mapeamento óptico panorâmico empregando o corante fluorescente sensível à voltagem Di-4-ANBDQPQ (V_m). A excitação do tecido foi obtida utilizando seis LEDs de cor vermelho profundo (650 nm). A fluorescência emitida foi captada por três câmeras de alta velocidade (HB-1800-S, Emergent Technologies), operando a 500 frames/s com uma resolução de 1600×1000 pixels. As câmeras foram equipadas com lentes de 25 mm e filtros de passagem longa de 715 nm, posicionadas a 120° umas das outras para garantir a cobertura total do órgão (Fig. 2).

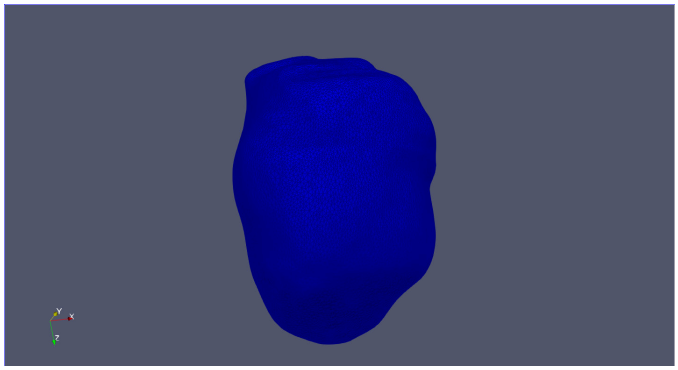


Figura 2. Imagem da superfície geométrica reconstruída via fotogrametria.

3) Eletrocardiografia: Para o registro dos sinais elétricos, o coração foi submerso em um tanque hexagonal customizado, preenchido com uma solução de sacarose (23 mM de NaCl e 260 mM de sacarose) (Fig. 3). O sistema de aquisição consistiu em 60 eletrodos de superfície organizados em matrizes, distribuídos estrategicamente pelas faces do tanque. Os dados foram registrados utilizando um módulo de aquisição de 64 canais (RHD 64, Intan Technologies), permitindo a coleta de eletrogramas espaciais de alta densidade (pseudo-ECG) a uma taxa de amostragem de 4 kHz. Adicionalmente, a dinâmica elétrica celular obtida a partir do experimento permitiu a extração do sinal do potencial de ação da superfície epicárdica (Fig. 4).

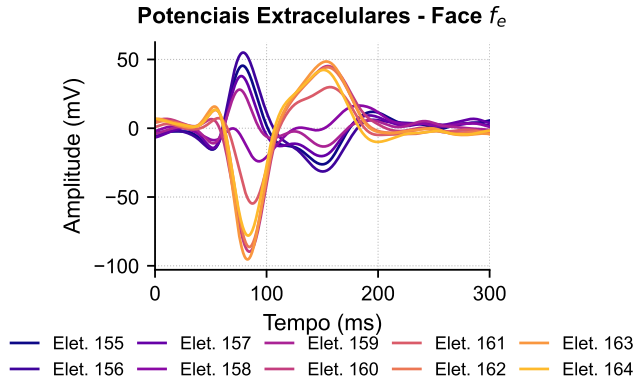


Figura 3. Registro temporal das leituras dos eletrodos de superfície no interior do tanque hexagonal.

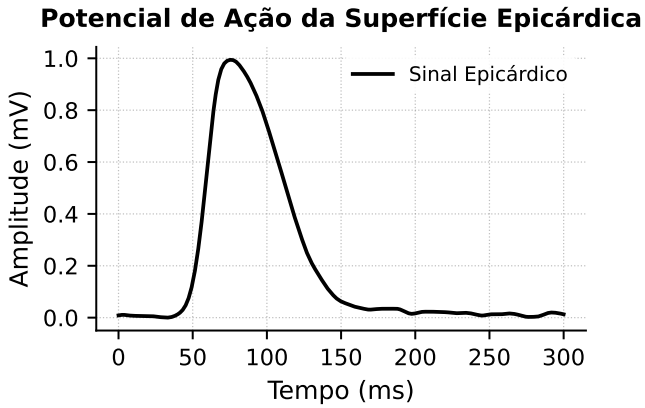


Figura 4. Sinal do potencial de ação extraído da superfície epicárdica.

B. Modelagem Computacional do Gêmeo Digital

1) Segmentação e Processamento Anatômico: A representação volumétrica dos ventrículos foi reconstruída a partir dos dados de RM utilizando o software 3D Slicer [12] por meio do algoritmo semi-automático *Grow from Seeds*. O processo de segmentação delimitou o volume ventricular, estabelecendo barreiras de contenção nas regiões atriais e na fronteira externa inferior do ventrículo direito.

Para corrigir artefatos de degrau decorrentes da resolução espacial e do espaçamento entre as fatias, um protocolo de filtragem morfológica (*Smoothing*) foi aplicado exclusivamente à máscara ventricular. Esse procedimento combinou operações de fechamento (*Closing*), para selar cavidades intersticiais internas, e operações de abertura (*Opening*), para remover ruído de fundo e aglomerados de voxels isolados. Por fim, um filtro Gaussiano com desvio padrão (tamanho do kernel) de 0,3 mm foi aplicado para garantir a continuidade geométrica e a suavidade da superfície (Fig. 5).

A superfície suavizada resultante (no formato STL) foi importada para o Autodesk Meshmixer [2] para refinamento topológico. Inicialmente, um corte planar transversal (*Plane Cut*) foi executado na região basal para remover irregularidades anatômicas remanescentes. Subsequentemente, a malha foi submetida a um relaxamento de superfície (*Smooth*) com a restrição de preservação de fronteiras (*Preserve Group Borders*) ativada, prevenindo o arredondamento indesejado da recém-criada borda basal. Por último, o modelo foi particionado em superfícies de contorno independentes para facilitar a imposição de condições de contorno em simulações futuras (Fig. 6).

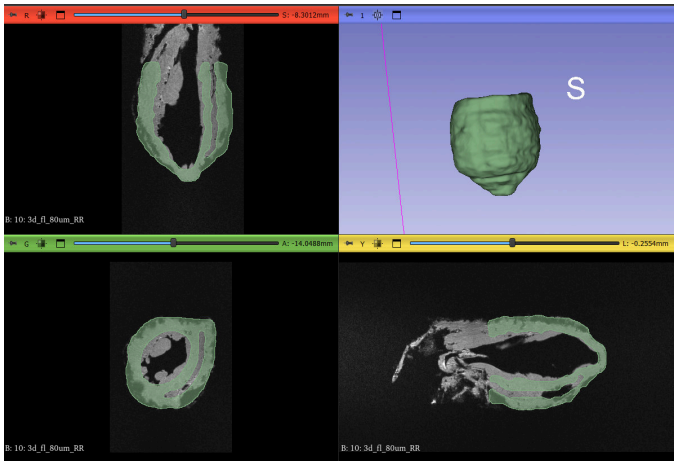


Figura 5. Máscara 3D resultante após a aplicação do algoritmo Grow from Seeds e dos processos de suavização morfológica.

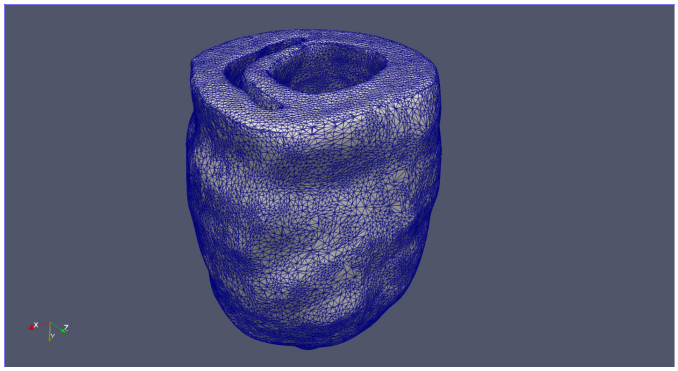


Figura 6. Modelo 3D refinado detalhando a topologia processada e as superfícies de contorno isoladas.

2) Geração da Malha Diastólica e Simulação Mecânica: Após o isolamento das superfícies de contorno, utilizou-se um *script* customizado baseado no software Gmsh para gerar a malha tridimensional de elementos finitos, com parâmetros de comprimento característico definidos no intervalo de $[0,5; 1,0]$ (Fig. 7).

Para reconstruir a geometria diastólica correspondente ao estado de pressão fisiológica durante o experimento, conduziu-se uma simulação mecânica de insuflação passiva utilizando o software Cardiax [9]. Uma pressão endocárdica de 1 kPa foi aplicada ao longo de 10 incrementos de carga, com restrição de deslocamento ao longo do eixo z para os nós pertencentes ao plano basal. O comportamento passivo do tecido miocárdico foi modelado pela lei constitutiva de Guccione [4], configurada

com os seguintes parâmetros: $C = 0,6$ kPa, $b_f = 6,62$, $b_t = 3,65$, $b_{fs} = 2,65$, e um módulo de bulk (penalidade de incompressibilidade) de $\kappa = 200$ kPa. A validação espacial do modelo deformado foi realizada através do corregistro com a superfície de referência obtida via fotogrametria (Fig. 8).

Por fim, versões de elementos finitos e de volumes finitos da malha diastólica foram geradas utilizando o software de código aberto *InSilicoHeartGen* [21] com resoluções espaciais iguais a $500\mu m$ e $200\mu m$, respectivamente. A malha grosseira de elementos finitos foi então utilizada para as simulações de inferência de parâmetros de condução, enquanto a malha refinada de volumes finitos foi empregada na etapa de calibração do modelo celular.

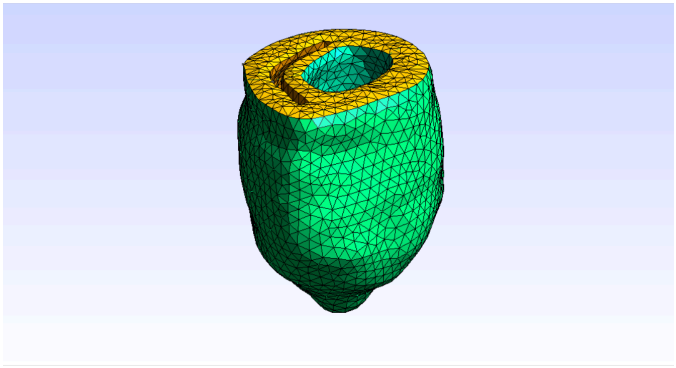


Figura 7. Malha de elementos finitos inicial pré-expansão gerada utilizando o GMSH.

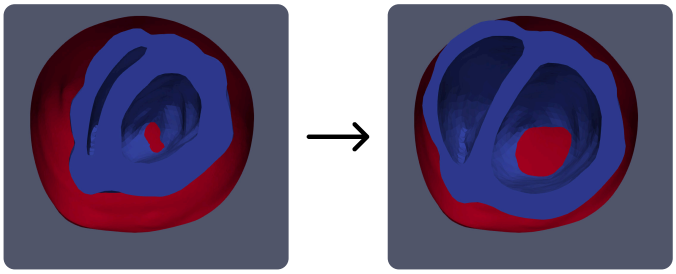


Figura 8. Em azul: malha volumétrica deformada após a simulação mecânica de insuflação no Cardiax; em vermelho: superfície de referência obtida por fotogrametria.

3) Inferência de Parâmetros de Condução (Eikonal): A reprodução computacional dos eletrogramas de superfície (pseudo-ECG) requer a caracterização precisa da velocidade de condução tissular e da distribuição espacial dos locais de ativação mais precoces (*root nodes*), o que aproxima o acoplamento da rede de Purkinje no miocárdio. O modelo Eikonal foi resolvido utilizando uma versão adaptada do software de código aberto CARDIAC-DIGITAL-TWIN-PURKINJE [19], que utiliza numericamente o algoritmo de Dijkstra aplicado à malha ventricular tetraédrica tridimensional para calcular os Tempos de Ativação Local (LAT).

A função objetivo (*discrepancy*) utilizada para guiar a minimização do problema inverso baseou-se em uma métrica combinada do Erro Quadrático Médio (RMSE) e do Coeficiente de Correlação de Pearson (PCC). Os sinais experimentais, extraídos da janela temporal restrita ao complexo QRS, foram pré-processados a uma frequência de amostragem de 1000 Hz, submetidos a um filtro passa-banda de 0,001 a 100 Hz e calculados diferencialmente para os canais 160 a 164 em relação ao eletrodo de referência 155 (GND).

A otimização do espaço paramétrico foi conduzida utilizando um algoritmo metaheurístico misto de Evolução Diferencial (DE) [15]. O algoritmo evolutivo foi configurado com uma população de 256 indivíduos, taxa de mutação (F) de 0,5, probabilidade de cruzamento (CR) de 0,8 e uma semente aleatória fixada em 42, operando por até 5000 gerações.

Além disso, não consideramos uma camada subendocárdica uniformemente rápida nesta configuração e inferimos apenas as três velocidades de condução anisotrópicas ao longo das fibras cardíacas: fibra (longitudinal), folha (transversal) e normal (transmural). A orientação das fibras foi calculada utilizando o software *InSilicoHeartGen* [21]. O espaço de busca global para as variáveis de condução e topologia anatômica foi delimitado da seguinte forma:

- Velocidade longitudinal (fibra): $[0,010; 0,100]$ cm/ms
- Velocidade transversal (folha): $[0,010; 0,060]$ cm/ms
- Velocidade transmural (normal): $[0,010; 0,040]$ cm/ms
- Número de *root nodes*: $[5; 15]$ locais de ativação mais precoces, assumindo uma distribuição *a priori* baseada em uma média de 10 nós.

4) Calibração do Modelo Celular: A precisão das simulações de propagação elétrica em nível tecidual depende da fidelidade da dinâmica iônica intracelular. Para tanto, empregou-se o *Minimal Model* [13], uma formulação fenomenológica que descreve o potencial de ação (PA) cardíaco por meio de quatro variáveis de estado (u, v, w, s). Registros experimentais de mapeamento óptico indicaram uma duração do PA de aproximadamente 100 ms.

A calibração paramétrica foi guiada por uma análise de sensibilidade global de Sobol [14], a qual identificou os parâmetros u_u, τ_{so1} e τ_v^+ como os principais determinantes da variância morfológica da curva do PA. A otimização desse subespaço paramétrico foi executada via Evolução Diferencial, utilizando a biblioteca PyMOO [16]. A função objetivo foi projetada para minimizar a divergência do APD90, complementada por um refinamento morfológico baseado na minimização do erro na norma L_2 entre os traçados simulado e experimental. Por fim, assumiu-se a premissa de homogeneidade celular transmural, aplicando-se o mesmo modelo otimizado globalmente em células epicárdicas, mesocárdicas e endocárdicas.

5) Simulação Eletrofisiológica (Monodomínio): Com os parâmetros anatômicos, de condução e celulares devidamente calibrados, a simulação espacial do tecido miocárdico foi realizada resolvendo-se a equação do monodomínio através do simulador de eletrofisiologia cardíaca MonoAlg3D [10], [11]. O *Minimal Model* [13] calibrado foi acoplado à formulação do monodomínio e resolvido sobre a malha biventricular diastólica utilizando uma discretização espacial de $200\mu m$. Previamente a isso, foram atribuídas à geometria as orientações das fibras miocárdicas.

Para refletir os resultados obtidos a partir da inferência computacional, a matriz de condutividade anisotrópica foi configurada no MonoAlg3D com os seguintes valores escalares: condutividade longitudinal (fibra) $\sigma_l = 5,29041 \times 10^{-5}$, transversal (folha) $\sigma_t = 3,53419 \times 10^{-5}$ e transmural (normal) $\sigma_n = 1,17339 \times 10^{-5}$. O mapeamento entre as velocidades de condução inferidas pelo modelo Eikonal e as condutividades do tecido no monodomínio foi realizado através de simulações unidimensionais em um cabo, empregando-se a rotina *tuneCV* no simulador MonoAlg3D [22].

A simulação da propagação elétrica, configurada com uma resolução espacial de 200 μm e um passo de integração no tempo de 0,02 ms, foi desencadeada pela ativação sequencial de nove pontos de ativação mais precoces (*root nodes*) mapeados durante o processo de inferência. Estes estímulos, modelados como domínios esféricos com um raio de 800μm, foram ativados de forma assíncrona (no intervalo entre $t = 1$ ms e $t = 25$ ms) utilizando os LATs calculados provenientes do processo de inferência.

III. Resultados

A. Alinhamento e Geometria da Malha

A validação do alinhamento geométrico e da expansão mecânica foi realizada de forma qualitativa, comparando-se a malha volumétrica simulada com o modelo de referência externo gerado por fotogrametria (Fig. 9). Embora os modelos apresentem forte concordância espacial global, observa-se que a geometria simulada extrapola os limites da fotogrametria em regiões localizadas.

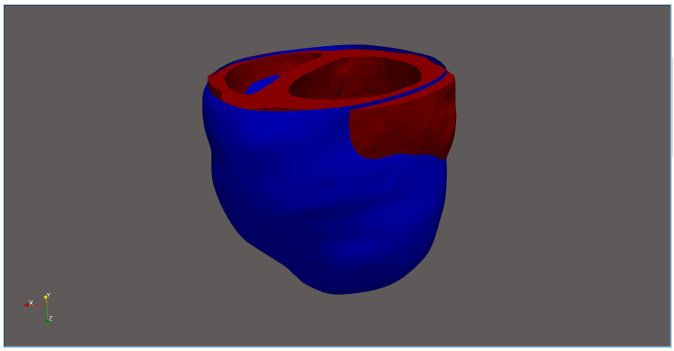


Figura 9. Em vermelho: malha volumétrica deformada após a simulação mecânica; em azul: superfície de referência obtida via fotogrametria.

B. Dinâmica Celular e Potencial de Ação

A otimização do modelo celular gerou uma curva com uma taxa de despolarização (fase 0) visivelmente mais acentuada em comparação ao sinal registrado (Fig. 10). Esta divergência morfológica inicial é bem fundamentada na literatura e atribuída ao efeito de espalhamento de fótons inerente aos sistemas de mapeamento óptico. Esse fenômeno induz uma integração espacial da fluorescência, resultando em um prolongamento artificial do *upstroke* medido [17]. Apesar dessa distorção conhecida na fase de ativação inicial, a Duração do Potencial de Ação (APD) foi ajustada, convergindo para a faixa de 100 ms e assegurando a fidelidade temporal necessária para as simulações em nível tecidual.

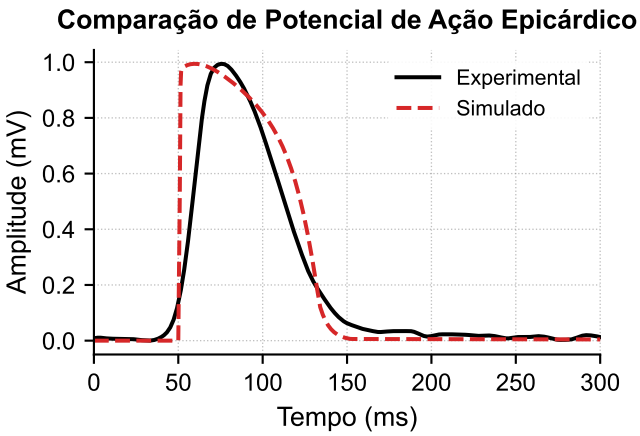


Figura 10. Em vermelho: PA simulado após a otimização (*Minimal Model*). Em azul: PA experimental (mapeamento óptico).

C. Eletrogramas

A acurácia dos eletrogramas de superfície simulados foi quantificada em relação aos sinais experimentais utilizando duas métricas: o Erro Quadrático Médio Normalizado (NRMSE) e o Coeficiente de Correlação de Pearson (PCC). Para a extração adequada dos sinais, selecionou-se um subconjunto de cinco eletrodos de interesse (identificados como derivações 160 a 164). Suas diferenças de potencial foram calculadas utilizando a derivação 155 como referência comum (terra/GND). Para viabilizar a comparação morfológica e considerar as variações de impedância no meio condutor, as amplitudes dos sinais simulados e inferidos foram normalizadas pico a pico em relação aos dados experimentais. A Fig. 11 detalha as formas de onda resultantes para este subconjunto e seus respectivos parâmetros estatísticos.

A análise quantitativa revela que o pseudo-ECG simulado diretamente pela equação do monodomínio apresentou divergências de amplitude aceitáveis (média de $29,6\%$) e valores ideais de PCC (média de $0,80$). Contudo, observou-se um desempenho notavelmente inferior para o eletrodo 160, que registrou um PCC de $0,55$, sugerindo limitações localizadas na inferência e na distribuição espacial dos *root nodes*. Em contrapartida, os sinais reconstruídos pelo processo de inferência Eikonal demonstraram capacidade de reproduzir o comportamento eletrofisiológico experimental. Esta formulação resultou em uma mitigação generalizada do erro de amplitude (baixo NRMSE) e em alta fidelidade morfológica, evidenciando a eficácia do mapeamento espacial dos parâmetros de condução gerado pela inferência.

IV. Discussão

Este estudo demonstra a viabilidade do desenvolvimento de um gêmeo digital cardíaco através da integração de dados anatômicos de ressonância magnética com registros eletrofisiológicos extraídos de um sistema de perfusão de Langendorff. A

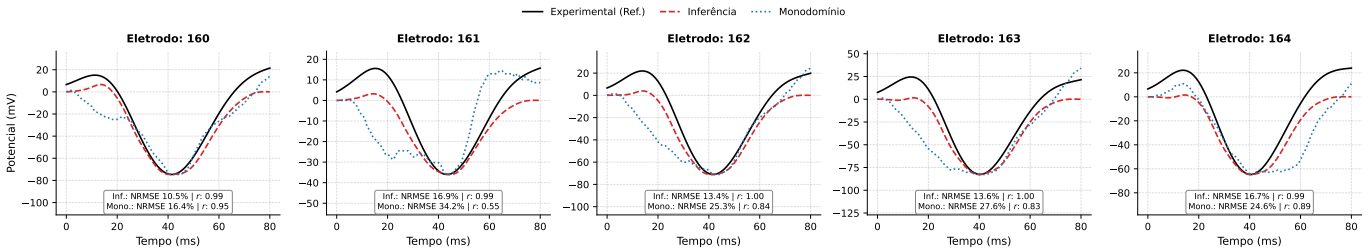


Figura 11. Comparação temporal de 5 eletrodos epicárdicos e métricas de correlação. Preto: pseudo-ECG experimental. Vermelho tracejado: sinal reconstruído via inferência (Eikonal). Azul pontilhado: sinal simulado utilizando o modelo do Monodomínio.

metodologia proposta reconstruiu com sucesso as propriedades morfológicas do órgão e reproduziu a progressão temporal da despolarização ventricular. Além disso, os parâmetros de velocidade de condução tissular inferidos pelo método inverso alinharam-se com a literatura eletrofisiológica estabelecida para o modelo de coelho [18], atestando a validade fenotípica da formulação Eikonal empregada na otimização.

O modelo exibe desvios esperados que destacam o impacto das simplificações assumidas em relação ao sistema biológico. No domínio biomecânico, a simulação de expansão passiva aplicou uma carga de pressão uniforme através das cavidades ventriculares. A extrapolação geométrica localizada observada em relação à superfície de fotogrametria reflete a ausência de um carregamento hemodinâmico assimétrico. A aplicação de gradientes diferenciais de pressão fisiológica para os ventrículos esquerdo e direito apresenta-se como a abordagem corretiva ideal para simulações mecânicas futuras, a fim de garantir a conformidade absoluta do contorno epicárdico.

No domínio elétrico, diferenças morfológicas residuais foram observadas entre os eletrogramas simulados pelo modelo do monodomínio e a referência experimental. Diferentemente do estado estacionário resolvido pela equação Eikonal, o monodomínio captura interações não lineares da dinâmica iônica. Portanto, a falta de heterogeneidade eletrofisiológica transmural (diferenciando células endocárdicas, mesocárdicas e epicárdicas) negligencia variações intrínsecas na forma e na duração dos potenciais de ação locais, impactando diretamente a reconstrução morfológica fina dos sinais de superfície.

As limitações na reprodução da dinâmica elétrica também refletem restrições computacionais durante a etapa de inferência. A otimização de parâmetros concentrou-se em um subconjunto de canais, deixando de lado a totalidade dos dados disponíveis na matriz de 60 eletrodos da montagem experimental. A incorporação do mapeamento elétrico completo no problema inverso poderia aprimorar a resolução espacial da predição, mitigando ainda mais os erros de amplitude (NRMSE). Contudo, a escalabilidade deste processo é dificultada por custos computacionais proibitivos; essa mesma barreira forçou a simplificação da rede de Purkinje. A restrição quantitativa de pontos de ativação no endocárdio (*root nodes*) foi uma imposição técnica necessária para prevenir uma explosão combinatória no espaço de busca do algoritmo de otimização.

Para mitigar essas limitações, trabalhos futuros podem priorizar a incorporação da heterogeneidade miocárdica e a expansão do domínio anatômico para abranger o coração inteiro (átrios e ventrículos). Esta integração estrutural, combinada com o desenvolvimento de redes de Purkinje topologicamente mais precisas e específicas para coelhos, permitirá a emulação do ciclo elétrico completo. Tais aprimoramentos estenderão o escopo do gêmeo digital cardíaco muito além da validação do complexo QRS, viabilizando o estudo detalhado das ondas P e T e, consequentemente, a gênese de arritmias impulsionadas por anomalias de repolarização. Além disso, investigações futuras devem considerar a implementação de um modelo celular específico para coelhos, como o modelo de Mahajan, que replica com precisão a dinâmica do potencial de ação cardíaco em altas frequências de estimulação [23].

V. Conclusões

O *pipeline* computacional estabelecido neste trabalho apresenta uma estrutura promissora para a geração de gêmeos digitais cardíacos personalizados. Através do acoplamento sequencial do processamento de imagens, da simulação mecânica de expansão passiva e da otimização eletrofisiológica guiada por dados experimentais, foi possível recriar o comportamento elétrico do miocárdio com um alto grau de similaridade espacial e temporal. O sucesso na recuperação das velocidades de condução e a fidelidade da propagação confirmam a eficácia da abordagem híbrida, que combina o problema inverso Eikonal com simulações biofísicas do monodomínio. A maturação destas metodologias apresenta um potencial considerável para impulsionar a transição tecnológica em direção aos ensaios clínicos *in silico*, aprimorando a eficácia dos testes farmacológicos e apoiando o princípio ético fundamental de reduzir o uso de modelos animais na pesquisa biomédica.

Agradecimentos

J. Salinet é apoiado pelo auxílio #2018/25606-2 da FAPESP. L. Arantes Berg agradece o apoio do auxílio CNPq Nº 446127/2024-8. Este estudo é apoiado pelo CNPq INCT INTERAS, chamada 58/2022. Os autores gostariam de expressar seus agradecimentos à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) - APQ-02752-24, APQ-02445-24, ao SINAPAD Santos-Dumont e à UFJF pelo financiamento deste trabalho.

REFERÊNCIAS

[1] J. D. Bayer, R. C. Blake, G. Plank e N. A. Trayanova, "A novel rule-based algorithm for assigning myocardial fiber orientation to computational heart models," *Annals of Biomedical Engineering*, vol. 40, pp. 2243–2254, 2012.

[2] Autodesk, Inc. "Meshmixer." [Online]. Disponível em: http://www.autodesk.com/ (acessado em 3 de abr. de 2025).

[3] Organização Mundial da Saúde. "Cardiovascular diseases (CVDs)." [Online]. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (acessado em 11 de abr. de 2025).

[4] J. M. Guccione, A. D. McCulloch e L. K. Waldman, "Passive material properties of intact ventricular myocardium determined from a cylindrical model," *Journal of Biomechanical Engineering*, vol. 113, nº 1, pp. 42–55, 1991.

[5] Q. Lou, W. Li e I. R. Efimov, "Multiparametric optical mapping of the Langendorff-perfused rabbit heart," *Journal of Visualized Experiments: JoVE*, nº 55, p. e3160, 2011.

[6] G. Coorey *et al.*, "The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field," *NPJ Digital Medicine*, vol. 5, nº 1, p. 126, 2022.

[7] J. Corral-Acero *et al.*, "The 'Digital Twin' to enable the vision of precision cardiology," *European Heart Journal*, vol. 41, nº 48, pp. 4556–4564, 2020.

[8] R. M. Bell, M. M. Mocanu e D. M. Yellon, "Retrograde heart perfusion: The Langendorff technique of isolated heart perfusion," *Journal of Molecular and Cellular Cardiology*, vol. 50, nº 6, pp. 940–950, 2011.

[9] J. O. Campos, R. W. Dos Santos, J. Sundnes e B. M. Rocha, "Preconditioned augmented Lagrangian formulation for nearly incompressible cardiac mechanics," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 34, nº 4, p. e2948, 2018.

[10] R. Sachetto Oliveira *et al.*, "Performance evaluation of GPU parallelization, space-time adaptive algorithms, and their combination for simulating cardiac electrophysiology," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 34, nº 2, p. e2913, 2018.

[11] L. A. Berg, *et al.*, "Toward cardiac electrophysiology digital twins with an efficient open source scalable solver on GPU clusters.," *Scientific Reports*, vol. 16, nº 1, pp. 9619, 2026.

[12] A. Fedorov *et al.*, "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magnetic Resonance Imaging*, vol. 30, nº 9, pp. 1323–1341, 2012.

[13] A. Bueno-Orovio, E. M. Cherry e F. H. Fenton, "Minimal model for human ventricular action potentials in tissue," *Journal of Theoretical Biology*, vol. 253, nº 3, pp. 544–560, 2008.

[14] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer Physics Communications*, vol. 145, nº 2, pp. 280–297, 2002.

[15] R. Storn e K. Price, "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341–359, 1997.

[16] J. Blank e K. Deb, "pymoo: Multi-objective optimization in Python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020.

[17] M. J. Bishop *et al.*, "Synthesis of voltage-sensitive optical signals: application to panoramic optical mapping," *Biophysical Journal*, vol. 90, nº 8, pp. 2938–2945, 2006.

[18] M. J. Schalij, L. Boersma, M. Huijberts e M. A. Allessie, "Anisotropic reentry in a perfused 2-dimensional layer of rabbit ventricular myocardium," *Circulation*, vol. 102, nº 21, pp. 2650–2658, 2000.

[19] J. Camps *et al.*, "Digital twinning of the human ventricular activation sequence to clinical 12-lead ECGs and magnetic resonance imaging using realistic Purkinje networks for in silico clinical trials," *Medical Image Analysis*, vol. 94, p. 103108, 2024.

[20] J. Camps *et al.*, "Harnessing 12-lead ECG and MRI data to personalise repolarisation profiles in cardiac digital twin models for enhanced virtual drug testing," *Medical Image Analysis*, vol. 100, p. 103361, 2025.

[21] R. Doste *et al.*, "An automated computational pipeline for generating large-scale cohorts of patient-specific ventricular models in electromechanical in silico trials," *Computer Methods and Programs in Biomedicine*, p. 109290, 2026.

[22] C. M. Costa, E. Hoetzl, B. M. Rocha, A. J. Prassl e G. Plank, "Automatic parameterization strategy for cardiac electrophysiology simulations," em *Computing in Cardiology 2013*, IEEE, 2013, pp. 373–376.

[23] A. Mahajan *et al.*, "A rabbit ventricular action potential model replicating cardiac dynamics at rapid heart rates," *Biophysical Journal*, vol. 94, nº 2, pp. 392–410, 2008.

Does Phrasing Matter? Prompt Sensitivity of Portable Large Language Models in Text-Based Sentiment Evaluation

Paulo C. Vaz de Barros
Engenharia de Computação e Informação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
paulcvaz.20231@poli.ufrj.br

Gabriel C. Motta Ribeiro
Programa de Engenharia Biomédica/COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
gabrielcasulari@peb.ufrj.br

Frederico C. Jandre
Programa de Engenharia Biomédica/COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jandre@peb.ufrj.br

***Abstract*—The use of portable Large Language Models (pLLMs) for sentiment analysis offers a cost-effective and private alternative to Large Language Models (LLMs), yet their operational stability against subtle prompt variations remains critical. This study investigates the prompt sensitivity of five open-source LLMs: Llama3.2 (3B), Phi-4 (14B), LLaVA (7B), Mistral (7B), and DeepSeek-R1 (8B), when performing valence evaluations. Performance was modeled using mean absolute error (MAE), root mean squared error (RMSE), and Pearson's correlation coefficient (r), while response validity was analyzed separately. The sample standard deviation of MAE across prompt configurations was used as the primary descriptive indicator of prompt sensitivity. `Phi-4` had the lowest mean errors and MAE variation, with** $\mathrm{MAE} = 1.07 \pm 0.15$, $\mathrm{RMSE} = 1.33 \pm 0.16$, **and** $r = 0.82 \pm 0.04$, **while returning valid responses in 82.00% of calls. Among the four pLLMs, `Mistral` had the lowest mean errors and MAE variation, with** $\mathrm{MAE} = 1.11 \pm 0.27$, $\mathrm{RMSE} = 1.42 \pm 0.28$, **and** $r = 0.73 \pm 0.06$, **while returning valid responses in 91.09% of calls. These findings suggest the need for an in-depth baseline study when selecting specific models for sentiment analysis pipelines, as different LLMs exhibit contrasting levels of operational volatility.**

***Index Terms*—Portable Large Language Models, Valence Rating, Sentiment Analysis, Linear Regression, Prompt Sensitivity, Self-Assessment Manikin, Natural Language Processing, `Llama3.2`, `Phi-4`, `LLaVA`, `Mistral`, `DeepSeek-R1`, Machine Learning in Linguistics**

I. Introduction

The advancement of large language models (LLMs) has established them as important tools for complex linguistic analysis [1]. Despite their usefulness, a notable challenge remains: small variations in the phrasing of prompts can lead to substantially different outputs. This sensitivity to word ordering and instruction styling has been documented as a core vulnerability in modern language models [2]. While this responsiveness allows for performance optimization through careful prompt engineering [3], it also introduces a degree of instability. The study of response variability and model uncertainty in LLM-based sentiment analysis has also been comprehensively reviewed by Herrera-Poyatos et al. [4], who categorized the primary sources of inconsistencies and discussed theoretical mitigation strategies.

Furthermore, deploying LLMs at scale can be both cost-prohibitive and time-consuming, making the adoption of freely available, smaller language models, such as portable Large Language Models (pLLMs), a viable alternative. Particularly due to their reduced computational overhead (less than 10 Billion parameters) and enhanced data privacy [5]. However, while recent literature demonstrates that pLLMs can effectively assist in text-based sentiment analysis [1], the effects of minor prompt variations on their outputs remain largely underexplored.

The objective of this study is to systematically investigate prompt sensitivity in four pLLMs and one local reference model when performing text-based sentiment evaluation on a positivity scale.

II. Methods

To quantify this continuous dimension of positiveness, our evaluation utilizes the valence domain of Self-Assessment Manikin (SAM) scale. Originally developed by Margaret Bradley and Peter Lang [6], the SAM scale is a non-verbal pictorial questionnaire designed to directly measure an individual's emotional response to a given stimulus. By adapting this human-centric framework to a numerical scale for computational evaluation, we establish a standardized baseline that allows language models to map textual sentiments onto a spectrum spanning from highly negative to highly positive [7].

A. Dataset

Participants in the original study [8] rated a series of English lemmas on a 9-point scale in three emotional dimensions: valence, arousal, and dominance. While the full publicly available dataset contains 13,915 words, the present study focuses exclusively on the valence dimension

and utilizes a representative subset of 140 words. This word set is exactly the same as the one established in our prior study [7]. In that work, the sampling procedure was designed to select words systematically from the original dataset sorted by valence, picking lemmas at intervals of 100 words with an added random jitter of +1 or −1.

B. Large Language Models

The models were selected based on the Hugging Face leaderboard rankings available in November 2025: `Phi-4` (14B) and the four pLLMs `Llama3.2` (3B), `LLaVA` (7B), `Mistral` (7B), and `DeepSeek-R1` (8B) [9], [10]. We evaluate the comparative performance of the five models across slightly different instructional prompts. Hereinto, the term *five evaluated models* refers to the complete set, whereas *four pLLMs* excludes `Phi-4`, because its 14 billion parameters exceed the 10-billion-parameter threshold adopted for the pLLM group. All models were run locally via the Ollama framework on a workstation equipped with an Intel i9 processor, 128 GB of RAM, and an NVIDIA RTX 4090 GPU (24 GB VRAM).

C. Parameters

To ensure the reproducibility and minimize variance, all models were used with a deterministic configuration. The specific parameters, available in the Ollama framework, are detailed in Table I.

Setting the temperature to 0 decreases the responses variance, while a fixed seed of 42 was applied to stabilize the internal state of the random number generator.

TABLE I
OLLAMA FRAMEWORK PARAMETERS THAT ENSURE REPRODUCIBILITY

Parameter	Value
temperature	0
seed	42
num_predict	500

The num_predict parameter was limited to 500 tokens as a safeguard against infinite generation loops. This behavior was observed in some models, which occasionally produced redundant symbols or emojis rather than a concise response. By enforcing this limit, the system reduces the risk of resource exhaustion and helps maintain the processing of subsequent words.

D. Prompt Engineering

To systematically assess the models' architectural robustness when performing the same affective task under slightly altered instruction sets, a two-phase cross-evaluation framework was established. In the first phase, each of the selected local LLMs acted as a *generator*. They were tasked with generating five semantic variations of the baseline prompt inherited from previous literature [7]. To ensure generation alignment, the models received the following meta-prompt:

> *Write only 5 similar prompts that mean the same, but with different words, as the given prompt: How positive or negative is this word on a 1-9 scale? Answer only with a facial emoji and a number, with 1 being "very negative" and 9 "very positive". Here is the word: {word}*

This procedure yielded a final pool of 25 unique machine-engineered prompts (five variations across five generator models), while the `num_predict` inference parameter was left at its default behavior to avoid constraining the generation length during this initial phase. The authors jointly inspected the 25 prompts and considered them semantically equivalent for the target task.

In the second phase, these 25 prompts were deployed as a standardized test suite. Every prompt was cross-evaluated by all five local LLMs acting as *evaluators*. This full matrix configuration resulted in 25 distinct prompts × 5 evaluator models = 125 unique evaluation lines for each evaluated term. Given our sample dataset of 140 target words, the pipeline executed a total of 17,500 (125×140) local `ollama` API calls.

Due to the unconstrained nature of text generation in smaller architectures, a validation filter was implemented to parse the outputs. Responses that failed to provide the mandatory quantitative rating were labeled as invalid and subsequently discarded.

E. Metrics

These $17.5k$ calls were classified by prompt (instruction generated), the generator model (the prompt author), and the evaluator model (the model that interprets the prompts and provides the final ratings).

Each prompt–evaluator configuration comprised 140 word evaluations. The experiment comprised 125 prompt–evaluator configurations. Numerical metrics and regression fits were calculated only for configurations with sufficient valid responses. Following the statistical analysis from the previous studies, we conducted an ordinary least squares (OLS) linear regression for each calculable configuration. The resulting regression lines were presented across six panels as a descriptive visualization of model performance under slightly different instructions.

For every calculable prompt–evaluator configuration, predicted valence $\hat{y}_i$ was compared with the human rating y_i using mean absolute error (MAE) and root mean squared error (RMSE). Pearson's correlation coefficient (r) quantified linear association between model predictions and human ratings.

MAE, RMSE, and r were summarized using their mean and sample standard deviation across prompt configurations. The sample standard deviation of MAE was selected as the primary descriptive indicator of prompt sensitivity. It was interpreted together with mean MAE so that consistent but inaccurate responses were not characterized as robust performance.

A response was considered valid when its parsed numerical value was between 1 and 9, inclusive. Missing values, nonnumeric outputs, and values outside this interval were classified as invalid. Invalid responses were excluded from the numerical analyses and were not replaced with estimated values. The success rate is the percentage of valid responses returned by an evaluator model relative to its total number of calls (25 prompts $\times$140 words = 3.5k calls).

To determine whether the observed differences in agreement with human valence ratings were statistically supported, the calculated metrics were compared in all pairs of evaluator models. For each word and evaluator, the valid predictions obtained across the 25 prompt configurations were averaged, producing one estimate per model for each word. MAE and RMSE were calculated between these prompt-averaged predictions and the human ratings, while Pearson's r measured their linear association. Each of the ten possible model pairs was then compared by paired bootstrap resampling of the words, using 10,000 replicates and seed 42. Bootstrap percentile confidence intervals and two-sided bootstrap p-values were calculated, and Holm's procedure was applied across the ten model comparisons separately for each metric. Differences were considered statistically supported when $p_{\text{Holm}} < 0.05$.

III. Results

A. Prompt Variation

The 25 prompts differed in wording and in the emphasis placed on output format. Table II presents one prompt from each generator.

TABLE II
Examples of prompts generated by the large language models

Model	Prompt
`DeepSeek-R1`	Rate the sentiment of this word from 1 to 9, using a facial emoji. 1 is very negative, 9 is very positive.
`Llama3.2`	What's the emotional tone of this word? Rate it from 1 to 9, where 1 is "very negative" and 9 is "very positive".
`LLaVA`	Please rate this word on a scale of 1 to 9, with 1 being "extremely negative" and 9 "extremely positive", using a facial emoji.
`Mistral`	Rate the tone of this word on a 1-9 scale, where 1 is "very negative" and 9 is "very positive". Please respond with an emoji and a number.
`Phi-4`	Using a 1-9 scale, where 1 is very negative and 9 is very positive, how would you express the sentiment of this word? Provide a facial emoji and a number.

B. Success Rate

The 25 prompts were used to trigger 17.5k ollama calls. Only 83.7% were classified as usable. Table III shows the success rate of each evaluator model.

TABLE III
Models success rate responding to the prompts

Evaluator Model	Valid	Invalid	Success Rate (%)
`DeepSeek-R1` (8B)	2749	751	78.54
`Llama3.2` (3B)	3242	258	92.63
`LLaVA` (7B)	2607	893	74.49
`Mistral` (7B)	3188	312	91.09
`Phi-4` (14B)	2870	630	82.00

To further examine the performance dynamics, Figure 1 presents a cross-evaluative heatmap of response success rates. In this visualization, each element cell at row n and column m represents the success rate of prompts generated by model n when evaluated by model m. This mapping reveals significant variances in how different models interact during the evaluation task.

Response validity also varied by prompt generator and individual prompt. The `LLaVA`-generated block yielded 2,269 valid responses in 3,500 calls (64.83%), the lowest validity among the five generator blocks, whereas the `Phi-4`-generated block yielded 3,301 valid responses (94.31%). `LLaVA-P2` (the second prompt generated by `LLaVA`, provided in Table II) yielded 377 valid responses in 700 calls (53.86%); when this prompt was evaluated by `LLaVA`, all 140 responses were invalid. Across the 25 individual prompts, `DeepSeek-R1-P1` (the first prompt generated by `DeepSeek-R1`, provided in Table II) had the lowest aggregate validity, with 317 valid responses in 700 calls (45.29%).

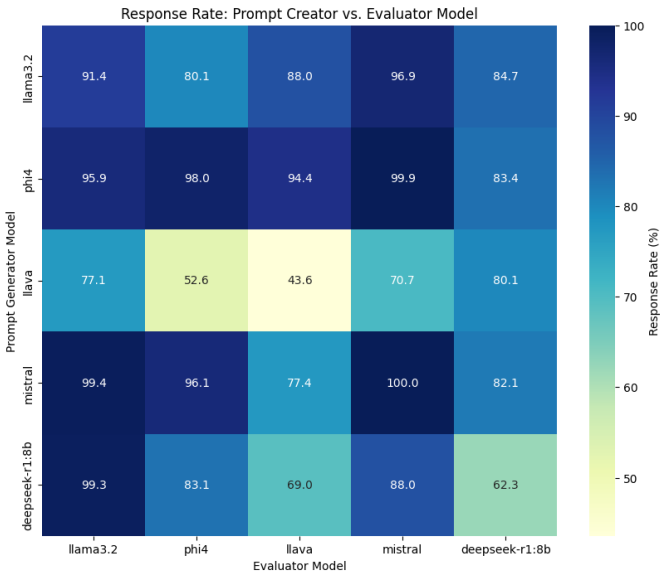
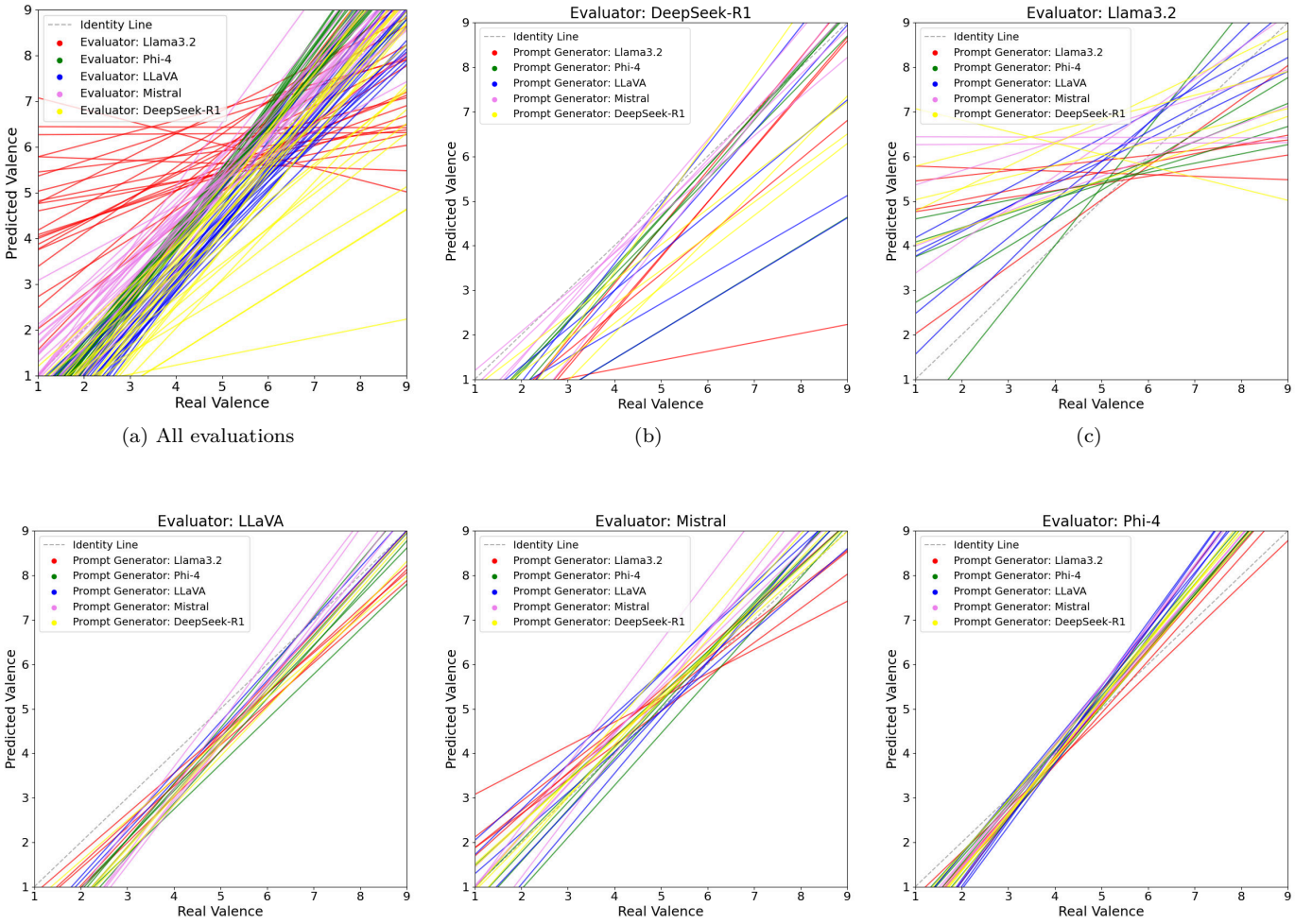


Fig. 1. Heatmap of response success rates (%) across different Large Language Models. Each cell (n,m) represents the percentage of valid responses generated by model m (Evaluator) when prompted with instructions authored by model n (Generator).



(a) All evaluations (b) (c)

(d) (e) (f)

Fig. 2. Comparison of affective evaluation performance across different prompt variations. The panels display results for: (a) all aggregated models, classified by the evaluator model, each color represents a model; (b,c,d,e,f) responses given by the models `DeepSeek-R1` (8B), `Llama3.2` (3B), `LLaVA` (7B), `Mistral` (7B) and `Phi-4` (14B), respectively, classified by the generator prompt.

C. Affective Valence Responses

The performance of the models in terms of linear alignment with human-rated valence is illustrated in Figure 2. The figure presents the linear regression fits for each prompt–evaluator pairing. Figure 2a displays the aggregated trend for each model acting as an evaluator. These regressions provide a descriptive visualization of variation across the evaluated configurations.

D. Absolute Error and Correlation

Table IV summarizes MAE, RMSE, and Pearson's correlation coefficient r by evaluator across prompt configurations. `Phi-4` showed the lowest average MAE and RMSE, together with the highest average Pearson correlation. Among the four pLLMs, `Mistral` had the lowest mean errors and the highest mean correlation.

TABLE IV
DESCRIPTIVE PERFORMANCE OF EVALUATOR MODELS ACROSS CALCULABLE PROMPT–EVALUATOR CONFIGURATIONS. MAE, RMSE, AND PEARSON'S CORRELATION COEFFICIENT (r) ARE REPORTED AS MEAN ± SAMPLE STANDARD DEVIATION ACROSS PROMPT CONFIGURATIONS.

Evaluator	MAE	RMSE	r
`Phi-4`	1.07 ± 0.15	1.33 ± 0.16	0.82 ± 0.04
`Mistral`	1.11 ± 0.27	1.42 ± 0.28	0.73 ± 0.06
`LLaVA`	1.44 ± 0.56	1.76 ± 0.52	0.72 ± 0.05
`DeepSeek-R1`	2.07 ± 0.70	2.51 ± 0.71	0.54 ± 0.15
`Llama3.2`	1.52 ± 0.28	1.96 ± 0.32	0.32 ± 0.23

E. Paired Model Comparisons

The descriptive differences in MAE, RMSE, and Pearson's correlation coefficient do not, by themselves, establish whether the evaluated models differed beyond the variation associated with the set of words. Pairwise

TABLE V

PAIRED COMPARISONS AMONG EVALUATOR MODELS. DIFFERENCES ARE CALCULATED AS THE FIRST MODEL MINUS THE SECOND. THE 95% CONFIDENCE INTERVALS ARE UNADJUSTED BOOTSTRAP PERCENTILE INTERVALS; p-VALUES WERE HOLM-ADJUSTED ACROSS THE TEN COMPARISONS SEPARATELY FOR EACH METRIC. ROWS ARE ORDERED BY DECREASING HOLM-ADJUSTED p-VALUE FOR PEARSON CORRELATION.

Contrast	**MAE**			**RMSE**			**Pearson's r**		
	Δ	**95% CI**	p_{Holm}	**Δ**	**95% CI**	p_{Holm}	**Δ**	**95% CI**	p_{Holm}
`Phi-4` vs. `LLaVA`	−0.13	[−0.28, 0.01]	0.303	−0.14	[−0.32, 0.04]	0.370	0.02	[−0.05, 0.07]	0.962
`LLaVA` vs. `Mistral`	0.23	[0.08, 0.38]	0.011	0.22	[0.06, 0.38]	0.036	0.02	[−0.03, 0.08]	0.962
`Mistral` vs. `DeepSeek-R1`	−0.51	[−0.68, −0.33]	0.002	−0.49	[−0.68, −0.29]	0.002	0.03	[−0.03, 0.09]	0.962
`Phi-4` vs. `Mistral`	0.09	[−0.01, 0.20]	0.303	0.07	[−0.07, 0.21]	0.629	0.04	[−0.01, 0.09]	0.637
`LLaVA` vs. `DeepSeek-R1`	−0.28	[−0.43, −0.13]	0.002	−0.27	[−0.42, −0.11]	0.006	0.05	[−0.01, 0.13]	0.637
`Phi-4` vs. `DeepSeek-R1`	−0.42	[−0.59, −0.24]	0.002	−0.41	[−0.62, −0.22]	0.004	0.07	[0.03, 0.12]	0.012
`Llama3.2` vs. `DeepSeek-R1`	−0.18	[−0.41, 0.05]	0.303	0.00	[−0.28, 0.26]	0.990	−0.33	[−0.45, −0.21]	0.002
`Llama3.2` vs. `Mistral`	0.33	[0.14, 0.52]	0.006	0.49	[0.22, 0.73]	0.006	−0.36	[−0.49, −0.23]	0.002
`Llama3.2` vs. `LLaVA`	0.10	[−0.11, 0.32]	0.355	0.27	[0.01, 0.53]	0.186	−0.38	[−0.52, −0.25]	0.002
`Llama3.2` vs. `Phi-4`	0.24	[0.07, 0.41]	0.029	0.41	[0.20, 0.62]	0.005	−0.39	[−0.53, −0.27]	0.002

comparisons were therefore performed to determine which differences were statistically supported under the common set of 25 prompts.

Tables IV and V address different but complementary questions. Table IV summarizes metrics calculated separately for each calculable prompt–evaluator configuration and describes both average performance and variation across prompt configurations. In contrast, Table V compares evaluator models using one prompt-averaged prediction for each word and model. Consequently, the paired differences reported in Table V are not obtained by directly subtracting the descriptive means reported in Table IV.

For each model pair, signed differences in Pearson's correlation coefficient, MAE, and RMSE were computed. The reported 95% confidence intervals are percentile intervals derived from the bootstrap distributions. After Holm adjustment, five comparisons showed statistically supported differences in correlation: `Llama3.2` had a lower correlation than each of the other four evaluators, while `Phi-4` had a higher correlation than `DeepSeek-R1`. No statistically supported correlation differences were observed for the remaining five model pairs.

IV. Discussion

Changes in prompt formulation were associated with variation in both numerical agreement and response validity, indicating that phrasing mattered under the evaluated conditions. However, the magnitude and nature of this effect depended on the outcome considered. MAE and RMSE measured agreement with the human valence scale, Pearson's r measured preservation of linear ordering, and success rate measured adherence to the required response format. These dimensions did not vary uniformly across models and should therefore be interpreted separately.

The sample standard deviation of MAE across prompt configurations was used as the primary descriptive indicator of prompt sensitivity. By this indicator, `Phi-4`, included as a local reference model rather than as a pLLM, showed the lowest prompt sensitivity. Among the four pLLMs, `Mistral` showed the lowest MAE variation, whereas `DeepSeek-R1` showed the highest variation across all five evaluated models. Moreover, this ordering was not reproduced for correlation variability, for which `Llama3.2` showed the largest variation, reinforcing that prompt sensitivity depends on the performance dimension considered.

The paired comparisons qualify the descriptive ordering observed in Table IV. `Phi-4` had the lowest descriptive errors and the highest mean correlation, whereas `Mistral` had the corresponding best descriptive values among the four pLLMs. However, none of the paired differences between these two models in MAE, RMSE, or correlation was statistically supported after Holm adjustment. Their numerical ordering therefore does not establish superior numerical performance for either model.

In the error-based paired comparisons, `Mistral` had lower MAE and RMSE than `LLaVA`, but the analysis provided no evidence of a difference in correlation. Thus, `Mistral`'s estimates were closer to the human ratings without establishing a difference in the models' preservation of word ordering. Response validity further distinguished them: `Mistral` returned valid responses in 91.09% of calls, compared with 74.49% for `LLaVA`. `DeepSeek-R1` also illustrated the distinction between error and validity: its MAE and RMSE were higher than those of `Phi-4`, `LLaVA`, and `Mistral`, while 78.54% of its responses were valid.

As shown in Table V, `Llama3.2` had lower correlation than each of the other evaluators after Holm adjustment (all $p_{\text{Holm}} = 0.002$). This pattern is consistent with our previous study, in which `Llama3.2` also exhibited the weakest correlation with human valence ratings under a single prompt configuration [1]. By extending the evaluation to 25 prompt configurations, the present findings show that the lower linear association was not eliminated by the prompt variations examined here. This recurrence should be interpreted cautiously as consistency across the evaluated settings, rather than as evidence of the model's general adequacy or inadequacy for valence evaluation.

The paired comparisons and the prompt-sensitivity

analysis address complementary questions. The low p_{Holm} values indicate that the correlation computed from the prompt-averaged predictions of `Llama3.2` differed from those of the other evaluators, but they do not directly quantify sensitivity to prompt variation. At the same time, `Llama3.2` returned the highest proportion of valid responses (92.63%), showing that consistent adherence to the requested numerical format did not imply close agreement with the human ratings.

Response validity varied across prompt–evaluator combinations. As shown in Figure 1, the `LLaVA`-generated block had the lowest generator-block validity, and `LLaVA` returned no valid numerical response in the 140 calls using `LLaVA-P2`. This configuration had no calculable MAE and was omitted from the across-configuration summary, possibly reducing the observed MAE variation for `LLaVA`. Nevertheless, `DeepSeek-R1-P1` had the lowest aggregate validity across evaluators. Both prompts requested a 1–9 rating and mentioned a facial emoji but did not explicitly require a numerical value in the response. However, because wording and output-format requirements were not varied independently, these observations do not establish that the emoji instruction caused the invalid responses.

The study was restricted to 140 isolated English words obtained from a single public dataset. The 25 prompts were generated from one baseline prompt by the same five models subsequently used as evaluators. Although the prompts were considered semantically equivalent for the target task, some variations placed different emphasis on the required response format, making it difficult to fully separate sensitivity to wording from failures to follow the instructions. Results attributed to a generator block describe only the five existing prompts in that block and do not establish an evaluator-independent ability to generate prompts.

The paired comparisons assess differences between evaluator models under the complete set of existing prompts; they do not constitute statistical comparisons between individual prompts or prompt generators. Averaging valid responses for each word provided a common basis for comparison but summarized variation among the 25 prompts, and the number of valid responses contributing to each average could differ between models. The findings should therefore be interpreted together with response validity and remain limited to the evaluated words, prompts, models, and inference settings.

V. Conclusion

The results show that changes in prompt wording were associated with variation in numerical agreement and response validity, although these outcomes represented distinct aspects of model behavior. Using the sample standard deviation of MAE as the primary descriptive indicator, `Phi-4`, the local reference model, showed the lowest observed prompt sensitivity across calculable configurations; `Mistral` showed the lowest among the four pLLMs, whereas `DeepSeek-R1` showed the highest across the five evaluated models. In terms of mean performance, `Phi-4` also had the lowest MAE and RMSE and the highest correlation, whereas `Mistral` had the corresponding lowest mean errors and highest mean correlation among the four pLLMs. The paired analysis did not support differences between `Phi-4` and `Mistral` in MAE, RMSE, or correlation after Holm adjustment. `Llama3.2` returned the largest proportion of valid responses but had the lowest mean correlation.

These findings should not be interpreted as establishing a single overall ranking of the evaluated models. This work indicates that relying solely on mean performance metrics may mask the operational volatility of LLMs. Model selection depends on the relative importance assigned to numerical error, linear association, response validity, and computational portability. The conclusions are limited to the 140 isolated English words, 25 prompts, five locally executed models, and inference settings evaluated in this study.

Future work should evaluate prospectively constructed paraphrases in which semantic content, scale anchors, and output format remain constant while only linguistic phrasing is varied. The framework can also be extended to other languages, contextualized texts, and emotional dimensions of the SAM scale, such as arousal and dominance.

Acknowledgements

This paper was produced using the following tools:

- Gemini AI for grammar revision and suggestions.
- Python code to run machine learning and data analysis workflow.
- Ollama framework to access the used models.

References

[1] P. C. V. de Barros, G. C. M. Ribeiro, and F. C. Jandre, "Can small large language models estimate the hedonic valence of words?" *Proceedings of CBIC*, 2025.

[2] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8086–8098.

[3] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers.(2022)," *arXiv preprint arXiv:2211.01910*, 2022.

[4] D. Herrera-Poyatos, C. Peláez-González, C. Zuheros, A. Herrera-Poyatos, V. Tejedor, F. Herrera, and R. Montes, "An overview of model uncertainty and variability in llm-based sentiment analysis: challenges, mitigation strategies, and the role of explainability," *Frontiers in Artificial Intelligence*, vol. 8, p. 1609097, 2025.

[5] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang *et al.*, "A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 6, pp. 1–87, 2025.

[6] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[7] F. C. Jandre, G. C. Motta-Ribeiro, and J. V. A. da Silva, "Could large language models estimate valence of words? A small ablation study," in *Proceedings of CBIC*, 2023.
[8] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, pp. 1191–1207, 2013.
[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," *ArXiv*, vol. abs/2310.06825, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263830494

Estimation of Pulmonary Arterial Tree Characteristics in Pulmonary Hypertension

1st Fabiana Gomes Diniz
Center for Engineering, Modeling and Applied Social Sciences
Federal University of ABC
São Bernardo do Campo, Brazil
fabiana.diniz@ufabc.edu.br

2rd Jaomar Juan Chavez Salazar
Respiratory Intensive Care Unit
School of Medicine
University of São Paulo
São Paulo, Brazil
jaomar@usp.br

3rd Marcelo Britto Passos Amato
Respiratory Intensive Care Unit
School of Medicine
University of São Paulo
São Paulo, Brazil
marcelo.amato@fm.usp.br

4nd Fernando Silva de Moura
Center for Engineering, Modeling and Applied Social Sciences
Federal University of ABC
São Bernardo do Campo, Brazil
fernando.moura@ufabc.edu.br

***Abstract*—Pulmonary hypertension (PH) is a condition associated with low survival rates, and its standard diagnostic procedure is highly invasive. The objective of this work is to develop a method capable of estimating the parameters for pulmonary arterial tree flow models in patients with PH based on their hemodynamic waveforms. To this end, a pulmonary arterial tree was segmented from computed tomography angiography images of pigs with induced PH and was used to create the arterial flow model using openBF software. An inverse solver based on the generalized Tikhonov regularization was implemented to estimate the hemodynamic parameters of the model from pressure and velocity measurements in one of the main pulmonary arteries. The pulmonary arterial tree geometry was successfully reconstructed from the angiography images, and the proposed parameter estimation method yielded satisfactory results, although full convergence was not achieved for all parameters when multiple parameters were estimated simultaneously. The generalized Tikhonov regularization method proved suitable for parameter estimation in PH patient models.**

***Index Terms*—Pulmonary Hypertension; Electrical Impedance Tomography; 1D Modeling; Inverse Problem**

I. Introduction

Pulmonary hypertension (PH) is a highly lethal condition (mean survival of 2.8 years) [1], characterized by excessively elevated pulmonary arterial pressure in comparison with the level of cardiac output [2].

PH is classified into 5 clinical groups, according to the root cause of the disease: Pulmonary Arterial Hypertension; PH due to left-sided heart disease; PH due to chronic lung disease; Chronic Thromboembolic PH; and PH due to unclear and/or multifactorial mechanisms [3]. Furthermore, it is also classified as: 'pre-capillary", 'isolated post-capillary", or "combined pre- and post-capillary"; depending on the values of mean pulmonary arterial pressure (mPAP), pulmonary capillary wedge pressure (PCWP), and pulmonary vascular resistance (PVR) [1].

Although other tools are used to estimate the clinical probability of the condition or exclude possible causes, right heart catheterization is still the tool used to confirm the disease, despite being highly invasive [3].

Electrical impedance tomography (EIT) is a non-invasive tool based on electrical impedance variation to obtain bedside thoracic images [4], [5]. In addition, it has shown promise for the diagnosis of pathologies affecting pulmonary hemodynamics.

Studies have demonstrated the possibility of using EIT for pulmonary hemodynamic assessment under PH conditions. Correlations have been observed between the probability of PH presence, the severity of the condition, survival [6], [7], and systolic pulmonary arterial pressure with data obtained via EIT [8], [9]. However, the diagnostic power of EIT is linked to the possibility of obtaining absolute lung images (rather than relative ones), since it would be possible to detect an adverse condition without the need for comparison with images acquired prior to the onset of alterations [4], [10].

The main challenge in obtaining absolute images is the greater susceptibility to artifacts caused by modeling errors, as well as the need for improved image reconstruction algorithms [4]. To overcome this challenge, refined impedance models and prior information are required [4].

The possibility of using an anatomical atlas to obtain clinically meaningful thoracic images via EIT has already been demonstrated under conditions such as pneumothorax, atelectasis, pleural effusion, and pressure variations during mechanical ventilation [10]. An anatomical atlas of the upper part of the head has also been constructed and may be used for diagnostic purposes [11]. Furthermore, several pulmonary arterial tree models under the effects of PH have already been developed in order to better understand the hemodynamics of

This work was funded by the following agencies: Federal University of ABC (UFABC) and the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001.

this condition [12]–[16].

In summary, this work focuses on calibrating an arterial hemodynamic model using pressure and velocity waveforms obtained from swine, so that, in the future, it can be used as *a priori* information for the detection and monitoring of PH in patients. This represents the first step toward the application of EIT for this purpose.

II. Methodology

The algorithm developed to estimate the characteristics of the pulmonary arterial tree from its hemodynamic waveforms was divided into two parts. First, an algorithm capable of generating hemodynamic waveforms was developed. Subsequently, the inverse problem was addressed using Tikhonov regularization.

Finally, the algorithm was tested using both a simplified model and a model reconstructed from computed tomography angiography (CTA) images of pigs with experimentally induced PH (Project No. 2349/2025, approved by the Animal Research Ethics Committee (CEUA) of the School of Medicine, University of São Paulo).

A. Algorithm

The algorithm was implemented in Python (version 3.10.16; developed by the Python Software Foundation) [17] using Visual Studio Code (version 1.100.2; developed by Microsoft) [18].

To define the problem configuration (e.g., vessels and characteristics to be estimated in the inverse problem, vessels and waveforms measured in the forward problem, regularization parameter, Gauss-Newton step length, etc.), a configuration file in YAML format was created, from which the algorithm imports the required information.

The Reynolds number was calculated in both the forward and inverse problem solutions using the blood viscosity and density specified in the arterial network file (μ and ρ_b, respectively), twice the vessel radius specified in the configuration file or estimated after solving the inverse problem (represented by the diameter d), and the average flow velocity obtained from openBF (v), as shown in Eq. (1) [19]. The openBF software is described in Section II-A1.

$$\mathrm{Re} = \frac{v \cdot d \cdot \rho_b}{\mu} \tag{1}$$

For the model reconstructed from CTA images, the external pressure acting on each vessel was calculated for both the forward and inverse problems using the hydrostatic pressure equation (2), where ρ_w represents the density of water, g is the gravitational acceleration, and h is the vessel depth [20].

$$P_h = \rho_w g h \tag{2}$$

The density of water was used because its value is close to that of blood and serves as an approximation of the density of the surrounding tissues and fluids.

The depth of each vessel segment was determined from the distance between its midpoint and the highest midpoint within the model. The model was reoriented to different anatomical positions by multiplying the original coordinate matrix by a rotation matrix about the (x)-axis (3), where ω denotes the counterclockwise rotation angle.

$$\mathbf{R}(\omega) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\omega) & -\sin(\omega) \\ 0 & \sin(\omega) & \cos(\omega) \end{bmatrix} \tag{3}$$

Fig. 1 illustrates the different anatomical orientations of the pig and the corresponding orientation of Blender's local coordinate system.

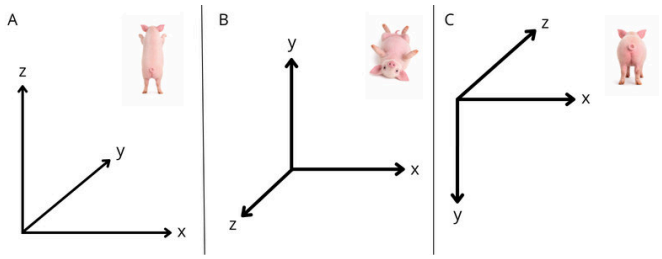


Fig. 1. New orientation of Blender's local coordinate system and the corresponding anatomical orientations of the pig. (A) Standing position; (B) Supine position; (C) Prone position. The pig images were generated using artificial intelligence.

1) Forward problem: The forward problem was solved using the hemodynamic simulation software openBF, developed by Dr. Alessandro Melis under the supervision of Dr. Alberto Marzo at the INSIGNEO Institute, where it was also validated [21]. This software is a one-dimensional blood flow solver based on the MUSCL numerical scheme and written in the Julia programming language [22], [23].

Although accuracy is reduced in the vicinity of heart valves and vascular bifurcations, one-dimensional (1D) simulations have a low computational cost, enabling the simulation of the entire cardiovascular system [23].

The model was derived from the Navier–Stokes equations under several simplifying assumptions: blood was considered an incompressible Newtonian fluid, and arteries were modeled as long, slender, elastic tubes [23].

Blood may be considered a Newtonian fluid, i.e., a fluid with constant viscosity, when simulating flow in large arteries without stagnation regions and at shear rates above $1{,}00 \times 10^2\ \mathrm{s}^{-1}$ [23], [24]. Furthermore, openBF can be used to simulate flow in short vessel segments, provided that the flow entering each segment is assumed to be fully developed, thereby neglecting the fluid transition region.

Despite these simplifications, the governing equations are capable of reproducing pulse wave propagation and reflection phenomena [23].

As input, the software requires a YAML file containing the characteristics of the vascular network to be simulated and a DAT file containing the vessel inlet boundary conditions. The output consists of hemodynamic waveforms at five locations along each simulated vessel [22], [23]. In the present study,

only the pressure and velocity waveforms at the midpoint of each vessel were used.

After generating the waveforms with openBF, noise was added to the simulated data to avoid committing the inverse crime. The inverse crime occurs when the same computational mesh is used both to generate and reconstruct the data. In this case, the absence of noise leads to a perfect reconstruction, which is unrealistic in practical applications [25]. The noise was generated from a normal distribution, with a standard deviation equal to a specified percentage of the total amplitude of each waveform.

2) Inverse Problem: This project was based on the Tikhonov regularization method to solve the inverse problem relating the hemodynamic waveforms of the pulmonary circulation to the characteristics of the arterial tree.

The need for this method arises from the fact that the inverse problem under investigation contains noise in the measured data, may fail to satisfy the Hadamard conditions (i.e., it may be ill-posed) and may also be ill-conditioned [26].

The Hadamard conditions are [27]:

- A solution exists for all admissible data;
- The solution is unique for all admissible data;
- The solution depends continuously on the data.

In an ill-conditioned problem, small variations in x can lead to large changes in $f(x)$ [28]. For a linear system $\mathbf{Ax} = \mathbf{b}$, the condition number of matrix $\mathbf{A}$ provides a useful indication of the uncertainty associated with the solution before the application of a numerical algorithm [29]. Assuming that $\mathbf{A}$ is a square and invertible/non-singular (full-rank) matrix, its condition number can be computed using Equations 4 and 5 [28].

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \tag{4}$$

$$\kappa(\mathbf{A}) = \frac{\sigma_1}{\sigma_m} \tag{5}$$

A well-conditioned problem typically has a condition number of $\kappa = 1, 10, 10^2$, whereas an ill-conditioned problem may have $\kappa = 10^6, 10^{16}$ [28].

Regularization methods improve the stability of inverse problem solutions and enhance the suppression of undesirable noise [30]. The inclusion of a regularization parameter $\mathcal{R}_k > 0$ provides a balance between model fitting and solution stability. Larger values of $\mathcal{R}_k$ lead to more stable solutions, whereas smaller values prioritize data fitting. Since the problem addressed here is nonlinear, the regularization parameter varies between iterations; therefore, the subscript k is used to indicate the corresponding iteration [31].

The main advantage of the Tikhonov regularization method over other approaches, such as the Levenberg–Marquardt and Gauss-Newton methods, is the incorporation of a reliable estimate of the unknown parameters to constrain the solution space [26].

This method seeks to minimize the objective function given by (6) [26], where $\mathbf{z}$ represents the measured (observed) data, $\mathbf{h}'(\boldsymbol{\theta})$ is the linearized *a posteriori* model response (see (7)), $\boldsymbol{\theta}$ denotes the *a posteriori* parameters, and $\boldsymbol{\theta}^*$ is a reliable estimate of the parameters. In (7), $\mathbf{h}(\boldsymbol{\theta}_k)$ corresponds to the model response at the current iteration, and $\mathbf{J}(\boldsymbol{\theta}_k)$ is the Jacobian matrix evaluated at the current parameter vector $\boldsymbol{\theta}_k$. The matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are weighting matrices used to balance the different magnitudes of the data and parameters, respectively [26].

$$F(\boldsymbol{\theta}) = \frac{1}{2}\left\{\|(\mathbf{z} - \mathbf{h}'(\boldsymbol{\theta}))\|^2_{\mathbf{W}_1} + \mathcal{R}_k^2\|(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2_{\mathbf{W}_2}\right\} \tag{6}$$

$$\mathbf{h}'(\boldsymbol{\theta}) = \mathbf{h}(\boldsymbol{\theta}_k) + \mathbf{J}(\boldsymbol{\theta}_k).(\boldsymbol{\theta} - \boldsymbol{\theta}_k) \tag{7}$$

The search for the optimal parameter at each iteration is performed using (8), obtained from the minimization of the objective function in (6) [26].

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \alpha_k &\left(\mathbf{J}_k^T\mathbf{W}_1\mathbf{J}_k + \mathcal{R}_k^2\mathbf{W}_2\right)^{-1} \\ &\cdot\left(\mathbf{J}_k^T\mathbf{W}_1(\mathbf{z} - \mathbf{h}(\hat{\boldsymbol{\theta}}_k)) - \mathcal{R}_k^2\mathbf{W}_2(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\right)\end{aligned} \tag{8}$$

The parameter $\alpha_k > 0$ is the Gauss-Newton step length and determines the magnitude of the parameter update at each iteration. The subscript k identifies the iteration to which the parameter corresponds [31].

In this project, $\mathbf{W}_1$ was defined as a diagonal matrix whose elements correspond to the inverse of the noise variance associated with each dataset. Specifically, the first 100 diagonal entries correspond to the inverse variance of the noise added to the 100 arterial pressure data points (Pa), whereas the last 100 entries correspond to the inverse variance of the noise added to the 100 velocity data points ($\mathrm{m/s}$).

The matrix $\mathbf{W}_2$ was defined as an identity matrix with dimensions equal to the number of estimated parameters.

The Jacobian matrix was evaluated numerically using finite differences, since a major challenge in solving nonlinear inverse problems is the accurate evaluation of derivatives [32]. Each parameter was perturbed individually, and the resulting changes in the pressure and velocity waveform vector $\mathbf{h}$ were used to estimate the partial derivatives (9). These derivatives were then assembled into the Jacobian matrix (10).

$$\frac{\partial \mathbf{h}(\theta_i)}{\partial \theta_i} = \lim_{\Delta\theta_i \to 0} \frac{\mathbf{h}(\theta_i + \Delta\theta_i) - \mathbf{h}(\theta_i)}{\Delta\theta_i} \tag{9}$$

$$\mathbf{J} = \frac{\partial \mathbf{h}}{\partial(\theta_1, \ldots, \theta_n)} = \begin{bmatrix} \dfrac{\partial h_1}{\partial \theta_1} & \cdots & \dfrac{\partial h_1}{\partial \theta_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_{200}}{\partial \theta_1} & \cdots & \dfrac{\partial h_{200}}{\partial \theta_n} \end{bmatrix} \in \mathbb{R}^{200\times n}. \tag{10}$$

Since the estimated parameters $\hat{\boldsymbol{\theta}}_k$ must always remain positive, the logarithm of the desired characteristics was estimated instead of the characteristics themselves. For example, rather than estimating the vessel length L, the parameter x was

estimated, where $x = \ln(L)$. In addition, the parameter update at each iteration was constrained to a maximum variation of 10%.

Several methods have been described in the literature for determining the Gauss-Newton step length (α_k) and the Tikhonov regularization parameter ($\mathcal{R}_k$).

In this project, for the Gauss-Newton step length, both a fixed parameter ($\alpha_k = 0.7$) and an adaptive determination based on the satisfaction of the Armijo, Wolfe, and strong Wolfe conditions were investigated. In the latter case, the selected α_k is the smallest value in the sequence $(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots)$ that satisfies the corresponding conditions.

The search direction $\mathbf{p}_k$ is defined according to (11).

$$\begin{aligned}\mathbf{p}_k = &\left(\mathbf{J}_k^T\mathbf{W}_1\mathbf{J}_k + \mathcal{R}_k^2\mathbf{W}_2\right)^{-1} \\ &\cdot\left(\mathbf{J}_k^T\mathbf{W}_1\big(\mathbf{z} - \mathbf{h}(\hat{\boldsymbol{\theta}}_k)\big) - \mathcal{R}_k^2\mathbf{W}_2\big(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big)\right)\end{aligned} \tag{11}$$

Therefore, (8) can be rewritten as (12).

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \alpha_k\mathbf{p}_k \tag{12}$$

Accordingly, the Armijo condition is given by (13), where c_1 is a constant between 0 and 1 [26], typically chosen as a small value such as 1×10^{-4}, which was the value adopted in this work [33]. The function F denotes the objective function defined in (6). The gradient of F is given by (14) and (15) [26].

$$F(\hat{\boldsymbol{\theta}}_k + \alpha_k\mathbf{p}_k) \leq F(\hat{\boldsymbol{\theta}}_k) + c_1\alpha_k\nabla F(\hat{\boldsymbol{\theta}}_k)^T\mathbf{p}_k \tag{13}$$

$$\nabla F(\hat{\boldsymbol{\theta}}_k) = -\mathbf{J}_k^T\mathbf{W}_1(\mathbf{z} - \mathbf{h}(\hat{\boldsymbol{\theta}}_k)) + \mathcal{R}_k^2\mathbf{W}_2(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) \tag{14}$$

$$\begin{aligned}\nabla F(\hat{\boldsymbol{\theta}}_k + \alpha_k\mathbf{p}_k) &= \nabla F(\hat{\boldsymbol{\theta}}_{k+1}) \\ &= -\mathbf{J}_{k+1}^T\mathbf{W}_1\big(\mathbf{z} - \mathbf{h}(\hat{\boldsymbol{\theta}}_{k+1})\big) \\ &\quad + \mathcal{R}_{k+1}^2\mathbf{W}_2\big(\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*\big).\end{aligned} \tag{15}$$

The Wolfe conditions combine the Armijo condition with the curvature condition defined in (16), ensuring that the criterion is not satisfied by arbitrarily small values of α_k [26]. The constant c_2 is chosen between c_1 and 1 and is typically set to 0.9 when a Newton or quasi-Newton method is used to determine the search direction $\mathbf{p}_k$, and to 0.1 when a nonlinear conjugate gradient method is employed [33]. In this study, $c_2 = 0.9$ was adopted.

$$\nabla F(\hat{\boldsymbol{\theta}}_k + \alpha_k\mathbf{p}_k)^T\mathbf{p}_k \geq c_2\nabla F(\hat{\boldsymbol{\theta}}_k)^T\mathbf{p}_k \tag{16}$$

For the strong Wolfe conditions, (16) is replaced by (17), thereby imposing a reduction in the absolute value of the directional derivative [26]. Strong Wolfe conditions are recommended only when the objective function is smooth [34].

$$|\nabla F(\hat{\boldsymbol{\theta}}_k + \alpha_k\mathbf{p}_k)^T\mathbf{p}_k| \leq c_2|\nabla F(\hat{\boldsymbol{\theta}}_k)^T\mathbf{p}_k| \tag{17}$$

A fixed Tikhonov regularization parameter (using several fixed values) was also evaluated, as well as regularization parameters determined by the following methods: L-curve [35], Discrepancy Principle [36], Generalized Cross-Validation (GCV) [37], Composite Residual and Smoothing Operator (CRESO) [38], Zero-Crossing [39], and U-Curve [40].

The L-curve method consists of plotting the solution norm against the residual norm (defined in (18) and (19), respectively) on logarithmic scales and identifying the value of $\mathcal{R}_k$ corresponding to the point of maximum curvature on the graph [41]. The method may fail when the curve is excessively smooth and can be computationally expensive because the norms must be evaluated for a large range of $\mathcal{R}_k$ values [30], [31].

$$\text{Solution norm} = \|(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|_{\mathbf{W}_2} \tag{18}$$

$$\text{Residual norm} = \|(\mathbf{z} - \mathbf{h}'(\boldsymbol{\theta}))\|_{\mathbf{W}_1} \tag{19}$$

The Discrepancy Principle aims to reduce the squared residual norm until it becomes consistent with the noise level present in the measured data, thereby avoiding model overfitting [32]. Since reducing the regularization parameter generally decreases the residual norm, the selected value of $\mathcal{R}_k$ should be large enough to prevent overfitting but not so large as to cause excessive regularization. Therefore, it must satisfy the condition described in (20), where ϕ denotes the number of measured data points (in this case, $\phi = 200$). Originally, the squared residual norm should equal $\phi\sigma^2$, where σ^2 is the noise variance. However, since the norm is already weighted by $\mathbf{W}_1$, which incorporates the variance, this term can be omitted [46], [47]. Given the difficulty of satisfying the equality, the squared residual norm closest to ϕ was chosen.

$$\|\mathbf{z} - \mathbf{h}'(\boldsymbol{\theta})\|_{\mathbf{W}_1}^2 = \phi \tag{20}$$

Although the method is computationally inexpensive, it tends to produce overly smooth solutions [31], and its performance depends on a reliable estimate of the noise level [30].

In the GCV method, the optimal parameter $\mathcal{R}_k$ is the one that minimizes the function defined in (21), where "trace" denotes the sum of the diagonal elements of the matrix inside the parentheses, and the matrix $\mathcal{A}$ is defined in (22) [41].

$$\mathcal{G}(\mathcal{R}_k) = \frac{\|\mathbf{z} - \mathbf{h}'(\boldsymbol{\theta})\|_{\mathbf{W}_1}^2}{\left[\frac{1}{\phi}\operatorname{trace}(\mathbf{I} - \mathcal{A})\right]^2} \tag{21}$$

$$\mathcal{A}(\mathcal{R}_k) = \mathbf{J}_k(\mathbf{J}_k^T\mathbf{W}_1\mathbf{J}_k + \mathcal{R}_k^2\mathbf{W}_2)^{-1}\mathbf{J}_k^T\mathbf{W}_1 \tag{22}$$

This method does not require prior knowledge of the noise level [37]; however, locating the minimum of the function can be difficult when it lies within a very flat region [31].

When applied to Tikhonov regularization, the CRESO method determines the optimal parameter as the first local maximum of the function defined in (23) [41]. The main

challenges associated with this method are the computation of the derivative and the identification of the function maximum.

$$\mathcal{C}(\mathcal{R}_k) = \frac{d}{d\mathcal{R}_k^2}\left(\mathcal{R}_k^2\|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|_{\mathbf{W}_2}^2 - \|(\mathbf{z}-\mathbf{h}'(\boldsymbol{\theta}))\|_{\mathbf{W}_1}^2\right) \tag{23}$$

In the Zero-Crossing method, the optimal regularization parameter $\mathcal{R}_k$ is assumed to be the value at which the residual norm and the regularization-weighted solution norm are equal, as shown in (24) [39]. Johnston and Gulrajani (1997) defined the optimal parameter as the one corresponding to the smallest root of the function, and the absence of a root occurs only in the hypothetical case where the data are noise-free [39].

$$\mathcal{B}(\mathcal{R}_k) = \mathcal{R}_k^2\|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|_{\mathbf{W}_2}^2 - \|(\mathbf{z}-\mathbf{h}'(\boldsymbol{\theta}))\|_{\mathbf{W}_1}^2 = 0 \tag{24}$$

In the U-Curve method, the optimal regularization parameter $\mathcal{R}_k$ is the one corresponding to the global minimum of the function defined in (25), which may also be represented on a log-log scale [41]. The method may fail if the minimum region is excessively broad, making it difficult to identify the global minimum [41].

$$\mathcal{U}(\mathcal{R}_k) = \frac{1}{\|(\mathbf{z}-\mathbf{h}'(\boldsymbol{\theta}))\|_{\mathbf{W}_1}^2} + \frac{1}{\|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|_{\mathbf{W}_2}^2} \tag{25}$$

The stopping criterion for the Tikhonov regularization iterations was defined as three consecutive iterations with less than 1% variation between successive solutions.

The initial estimate of the method ($\boldsymbol{\theta}_0$) was set to 50% of the true solution value, while the prior estimate ($\boldsymbol{\theta}^*$) was set to 95% of the true solution value.

B. Models

1) Simple Model: A simplified model consisting of three vessels connected by a single bifurcation was constructed for the initial testing of the algorithm. Some of its configuration parameters were obtained from a Circle of Willis model recommended in the openBF documentation [42]. These included the inlet volumetric flow-rate waveform, the rheological properties of blood, the numerical scheme parameters, the order of magnitude of the vessel characteristics, and the resistance and compliance values used in the two-element Windkessel (WK) models.

In this model, the wall thickness, length, and radius of vessel 2 were estimated from the pressure and velocity waveforms measured in vessel 1. In addition, the parameters of the two-element Windkessel (WK) model were estimated, with the proximal resistance estimated independently for vessels 2 and 3, whereas a single compliance value was estimated and shared by both vessels.

2) Model Obtained from Image Segmentation: The pulmonary arterial tree model affected by PH was reconstructed from computed tomography angiography (CTA) images of experimentally induced PH pigs, obtained through a collaboration with the Laboratory of Medical Investigation LIM-09, University of São Paulo. Image segmentation was performed using 3D Slicer [43] and Blender [44].

Vessel radius and length were extracted from the segmentation, while wall thickness was calculated by openBF from the vessel radius. External pressure was determined using the proposed algorithm, a parabolic velocity profile was assumed, and the initial Young's modulus was set to $8{,}757\,81 \times 10^3\,\mathrm{Pa}$ based on literature values for Group 3 PH patients (chronic obstructive pulmonary disease) [45]. Initial orders of magnitude for the 3-element Windkessel (WK) parameters were derived from the Circle of Willis model, as these represented the key parameters that are difficult to estimate.

Pressure and velocity waveforms from the pulmonary trunk were used to estimate the Young's modulus, length, and radius of the left pulmonary artery, together with the three WK parameters for all vessels. Both homogeneous (uniform across vessels) and heterogeneous (differ between the left and right sides) WK configurations were evaluated, considering prone and supine animal positions.

III. Results and Discussion

The simplified model was built using the characteristics described in Tables I, II, and III.

TABLE I
CHARACTERISTICS OF VESSEL 1 IN THE SIMPLIFIED ARTERIAL TREE MODEL

Parameter $\boldsymbol{\theta}$	**Value**
Wall thickness, h0	$1{,}63 \times 10^{-3}\,\mathrm{m}$
Length, L	$2 \times 10^{-1}\,\mathrm{m}$
Constant radius, R0	$1{,}2 \times 10^{-2}\,\mathrm{m}$
Young's modulus, E	$4{,}000\,00 \times 10^5\,\mathrm{Pa}$
Velocity profile, gamma_profile	2 (Laminar flow)
External pressure, Pext	$0\,\mathrm{Pa}$

TABLE II
CHARACTERISTICS OF VESSEL 2 IN THE SIMPLIFIED ARTERIAL TREE MODEL

Parameter $\boldsymbol{\theta}$	**Value**
Wall thickness, h0	$1{,}83 \times 10^{-3}\,\mathrm{m}$
Length, L	$1 \times 10^{-1}\,\mathrm{m}$
Constant radius, R0	$1{,}12 \times 10^{-2}\,\mathrm{m}$
Young's modulus, E	$4{,}000\,00 \times 10^5\,\mathrm{Pa}$
WK proximal resistance, R1	$5{,}430\,000\,000 \times 10^9\,\mathrm{Pa \cdot s/m^3}$
WK distal resistance, R2	$0\,\mathrm{Pa \cdot s/m^3}$
WK compliance, Cc	$1{,}27 \times 10^{-10}\,\mathrm{m^3/Pa}$
Velocity profile, gamma_profile	2 (Laminar flow)
External pressure, Pext	$0\,\mathrm{Pa}$

CT angiography image segmentation enabled the extraction of vessel lengths and radii, which were divided into sections represented by different colors in Fig. 2.

For the simplified model, the forward simulation in openBF converged after 100 cardiac cycles and required approximately 45 s. The Reynolds number was approximately 32, indicating laminar flow.

TABLE III
CHARACTERISTICS OF VESSEL 3 IN THE SIMPLIFIED ARTERIAL TREE MODEL

Parameter $\boldsymbol{\theta}$	**Value**
Wall thickness, h0	$1{,}93 \times 10^{-3}\,\mathrm{m}$
Length, L	$3 \times 10^{-1}\,\mathrm{m}$
Constant radius, R0	$6{,}2 \times 10^{-3}\,\mathrm{m}$
Young's modulus, E	$4{,}000\,00 \times 10^{5}\,\mathrm{Pa}$
WK proximal resistance, R1	$5{,}430\,000\,000 \times 10^{9}\,\mathrm{Pa \cdot s/m^3}$
WK distal resistance, R2	$0\,\mathrm{Pa \cdot s/m^3}$
WK compliance, Cc	$1{,}27 \times 10^{-10}\,\mathrm{m^3/Pa}$
Velocity profile, gamma_profile	2 (Laminar flow)
External pressure, Pext	$0\,\mathrm{Pa}$

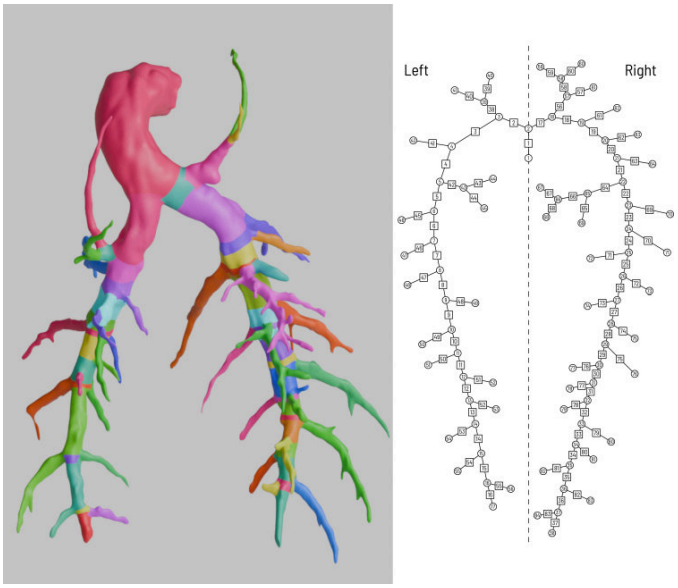


Fig. 2. Pulmonary arterial tree of a pig with induced PH. Each color represents a vessel subdivision. In the image on the right, each square corresponds to a subdivision, whereas each circle represents a junction node between adjacent subdivisions.

For the image-based model, the forward simulation converged after only 2 cardiac cycles and required approximately 7 s. The Reynolds number was approximately 291, also indicating laminar flow.

The faster convergence of the image-based model is likely due to the use of physiological vessel dimensions obtained from pigs with PH and Young's modulus values taken from the literature.

The condition number of matrix $\mathbf{A}_k$, defined in (26), was evaluated to assess the uncertainty of the inverse solution. Tikhonov regularization proved effective, as the matrices associated with all solved problems were well conditioned.

$$\mathbf{A}_k = (\mathbf{J}_k^T \mathbf{W}_1 \mathbf{J}_k + \mathcal{R}_k^2 \mathbf{W}_2) \tag{26}$$

For all inverse problem simulations, the Reynolds number remained low, confirming laminar flow.

The convergence of the inverse problem was primarily assessed through the relative difference between the estimated and true parameter values at each iteration. Consequently, the results depended on the method used to determine α_k and $\mathcal{R}_k$, the estimated parameters, and the pig posture.

For the simplified model, only method combinations involving at least one fixed parameter were evaluated because the longer openBF simulation time resulted in more than one hour per iteration when both parameters were determined. In contrast, all possible method combinations were evaluated for the image-based model.

The minimum value of α_k was set to 1×10^{-2}, since smaller values introduced excessive numerical instability.

The different methods for determining α_k showed no significant differences. Their performance depended mainly on the estimated parameters and the inverse problem configuration.

In contrast, the methods used to determine $\mathcal{R}_k$ produced substantially different results. The L-curve, GCV, and Zero-crossing methods provided the best overall performance, although none was consistently successful for all cases investigated.

The Discrepancy Principle, CRESO, and U-curve methods resulted in insufficient regularization, leading to poor parameter estimates. For the Discrepancy Principle, the squared residual norm remained far from ϕ during the first iterations, regardless of the value of $\mathcal{R}_k$, which may explain its poor performance. CRESO was evaluated using the first local maximum (original method), the last local maximum (as suggested by [41]), and by computing the derivative with respect to $\mathcal{R}_k$ instead of $\mathcal{R}_k^2$; however, all approaches produced insufficient regularization. Likewise, the U-curve method did not produce the expected U-shaped curves, even on a log-log scale.

The Zero-crossing method performed well for the simplified model. However, it produced excessive regularization for all image-based model configurations, causing all solutions to converge to the prior estimate $\boldsymbol{\theta}^*$.

Table IV summarizes the different configurations used to estimate the model parameters and the best combinations of methods for determining α_k and $\mathcal{R}_k$. The abbreviations are defined in Table II. Overall, the manual selection of fixed values for the regularization parameter $\mathcal{R}_k$ proved to be the most effective approach. This strategy reduces computational effort by eliminating the need to evaluate a wide range of candidate values to identify those satisfying the criteria described in Section II. Furthermore, it provided more reliable parameter estimates.

Table V presents the relative differences between the best estimated solution and the true parameter values for each case. The proposed inverse problem achieved complete or near-complete convergence for only some of the estimated parameters, whereas the remaining parameters showed larger deviations from the true solution. Although this behavior is common in inverse problems, it highlights the importance of evaluating the reliability of each estimated parameter before accepting the solution. The lack of a universally reliable method introduces uncertainty when applying the proposed approach to real cases. Nevertheless, the methodology can be reproduced for configurations identical to those investigated

TABLE IV
BEST COMBINATIONS OF OPTIMIZATION METHODS FOR DETERMINING THE OPTIMAL α_k AND REGULARIZATION PARAMETER $\mathcal{R}_k$.

Model	Estimated Parameters	Posture	Best optimization method
Simplified	h0, L, R0	N/A	Armijo + $\mathcal{R}_k = 500$
	R1, Cc	N/A	$\alpha_k = 0.7$ + L-curve/ GCV/ Zero-crossing or $\alpha_k = 0.7$/Armijo/ Wolfe/Strong Wolfe + $\mathcal{R}_k = 1000$
Image-based	E, L, R0	Supine	$\alpha_k = 0.7$/Armijo/Wolfe/ Strong Wolfe + $\mathcal{R}_k = 500$
		Prone	$\alpha_k = 0.7$/Armijo/Wolfe/ Strong Wolfe + $\mathcal{R}_k = 600$
	R1, R2, Cc Homogeneous	Supine	Strong Wolfe + $\mathcal{R}_k = 8$
		Prone	Armijo + L-curve
	R1, R2, Cc Heterogeneous	Supine	Wolfe + GCV
		Prone	Wolfe + $\mathcal{R}_k = 10$

TABLE V
APPROXIMATE RELATIVE DIFFERENCE BETWEEN THE SOLUTION AND THE TRUE VALUE.

Model	Estimated Parameters	Posture	Approximate Relative Difference of the Solution (%)
Simplified	h0, L, R0	N/A	h0: 6%; L: 4% and R0: 2%.
	R1, Cc	N/A	R1 (vessels 1 and 2): 0% and Cc: 5%.
Image-based	E, L, R0	Supine	E: 6%; L: 3% and R0: 1%.
		Prone	E: 6%; L: 3% and R0: practically 0%.
	R1, R2, Cc Homogeneous	Supine	R1: 0%; R2: 5% and Cc: 5%.
		Prone	R1: 0%; R2: 7% e Cc: 8%.
	R1, R2, Cc Heterogeneous	Supine	R1 Right: 0%; R1 Left: 2%; R2 (Right and Left): 5% and Cc (Right and Left): 6%.
		Prone	R1 Right: 5%; R1 Left: practically 0%; R2 Right: 0%; R2 Left: practically 0%; Cc (Right and Left): 7%.

in this study, i.e., with the same estimated parameters and anatomical posture.

IV. CONCLUSION

The inverse problem relating pulmonary arterial hemodynamic waveforms under PH to the characteristics of the arterial tree was successfully solved using Tikhonov regularization. The application of this method, together with strategies such as transforming the estimated parameters into the logarithmic domain, resulted in a well-conditioned inverse problem despite the large differences in parameter magnitudes.

The methods used to determine the step length α_k (Armijo, Wolfe, and strong Wolfe conditions) performed similarly, with no single method consistently outperforming the others. Regarding the methods for determining the Tikhonov regularization parameter $\mathcal{R}_k$, only the L-curve, GCV, and Zero-Crossing methods showed promising performance. However, as no single combination of optimization and regularization methods proved to be consistently superior across all problem configurations, it is suggested that fixing a value for the parameter $\mathcal{R}_k$, combined with either fixing α_k or applying different conditions to it, is the best approach.

The results presented in this study are expected to provide insight into both the strengths and limitations of applying Tikhonov regularization to calibrate hemodynamic models of the pulmonary arterial tree affected by PH.

Future work includes evaluating the convergence of the inverse solution using other fixed values of α_k and extending the proposed methodology to human arterial models.

REFERENCES

[1] Ministry of Health (Brazil), Secretariat of Specialized Health Care and Secretariat of Science, Technology, Innovation and Health Industrial Complex, Protocolo clínico e diretrizes terapêuticas: hipertensão pulmonar [Clinical protocol and therapeutic guidelines: pulmonary hypertension], Joint Ordinance No. 10, July 18, 2023. Brasília, DF, Brazil, 2023. [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/pcdt/h/hipertensao-pulmonar/view. Accessed: November 1, 2025.

[2] S. S. M. Barreto and M. B. Gazzana, "Hipertensão pulmonar: relato de seis casos e atualização do tema [Pulmonary hypertension: report of six cases and review of the topic]," J. Pneumologia, vol. 26, no. 6, pp. 321–336, December 2000.

[3] S. A. Mandras, H. S. Mehta, and A. Vaidya, "Pulmonary hypertension: a brief guide for clinicians," Mayo Clin Proc., vol. 95, no. 9, pp. 1978–1988, September 2020.

[4] T. C. Martins et al., "A review of electrical impedance tomography in lung applications: theory and algorithms for absolute images," Annu Rev Control, vol. 48, pp. 442–471, May 2019.

[5] E. L. V. Costa, R. G. Lima, and M. B. P. Amato, "Electrical impedance tomography," Curr Opin Crit Care, vol. 15, no. 1, pp. 18–24, February 2009.

[6] A. L. D. Hovnanian et al., "Electrical impedance tomography in pulmonary arterial hypertension," PLoS One, vol. 16, no. 3, pp. 1–13, Mars 2021.

[7] A. L. D. Hovnanian, "Papel da tomografia de impedância elétrica em pacientes portadores de hipertensão arterial pulmonar [Role of electrical impedance tomography in patients with pulmonary arterial hypertension]," Ph.D. dissertation, School of Medicine, University of São Paulo, São Paulo, Brazil, 2013.

[8] M. Proença et al., "Non-invasive monitoring of pulmonary artery pressure from timing information by EIT: experimental evaluation during induced hypoxia," Physiol Meas, vol. 37, no. 6, pp. 713–726, June 2016.

[9] M. Proença et al., "Non-invasive pulmonary artery pressure estimation by electrical impedance tomography in a controlled hypoxemia study in healthy subjects,"Sci Rep, vol. 10, no. 1, December 2020.

[10] E. D. L. B. Camargo, "Desenvolvimento de algoritmo de imagens absolutas de tomografia por impedância elétrica para uso clínico [Development of an absolute imaging algorithm for electrical impedance tomography for clinical use]," Ph.D. dissertation, Escola Politécnica, Universidade de São Paulo, São Paulo, Brazil, 2013.

[11] F. S. Moura, R. G. Beraldo, L. A. Ferreira, and S. Siltanen, "Anatomical atlas of the upper part of the human head for electroencephalography and bioimpedance applications," Physiol Meas, vol. 42, no. 10, November 2021.

[12] B. S. Ebrahimi et al., "A computational model of contributors to pulmonary hypertensive disease: impacts of whole lung and focal disease distributions," *Pulmonary Circulation*, vol. 11, no. 4, pp. 1–15, November 2021.

[13] M. J. Colebank, M. U. Qureshi, S. Rajagopal, R. A. Krasuski, and M. S. Olufsen, "A multiscale model of vascular function in chronic thromboembolic pulmonary hypertension," Am J Physiol Heart Circ Physiol, vol. 321, no. 2, pp. H318–H338, August 2021.

[14] M. L. Dong, "Image-based modeling and analysis of hemodynamics in pulmonary hypertension," Ph.D. dissertation, Stanford University, California, USA, 2022.

[15] S. Kakaletsis et al., "Untangling the mechanisms of pulmonary arterial hypertension-induced right ventricular stiffening in a large animal model," Acta Biomater, vol. 171, pp. 155–165, November 2023.

[16] A. Issakhov, A. Sabyrkulova, A. Abylkassymova, and K. Volkov, "Numerical modeling of pulsatile blood flow in the pulmonary artery under the influence of pulmonary hypertension and concomitant pathologies," *International Communications in Heat and Mass Transfer*, vol. 163, p. 108628, April 2025.

[17] Python Software Foundation, "Downloads," 2024. [Online]. Available: https://www.python.org/downloads/. Accessed: December 16, 2024.

[18] Microsoft, "April 2025 (version 1.100)," 2025. [Online]. Available: https://code.visualstudio.com/updates/v1_100. Accessed: May 16, 2025.

[19] J. E. Hall and M. E. Hall, *Guyton & Hall: Tratado de Fisiologia Médica*, 14th ed. Rio de Janeiro: Guanabara Koogan, 2021. [Online]. Available: https://integrada.minhabiblioteca.com.br/reader/books/9788595158696/. Accessed: Jan. 5, 2026.

[20] Y. C. Fung, *Biomechanics: Circulation*, 2nd ed. New York: Springer, 1996.

[21] I. Benemerito, A. Melis, A. Wehenkel, and A. Marzo, "openBF: an open-source finite volume 1D blood flow solver," *Physiological Measurement*, vol. 45, no. 12, p. 125002, 2024. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/ad9663. Accessed: September 29, 2025.

[22] A. Melis, "openBF: Julia software for 1D blood flow modelling," figshare, 2018. [Online]. Available: https://figshare.com/articles/openBF_Julia_software_for_1D_blood_flow_modelling/7166183/1. Accessed: October 2018.

[23] A. Melis, "Gaussian process emulators for 1D vascular models," Ph.D. dissertation, University of Sheffield, 2017. [Online]. Available: https://etheses.whiterose.ac.uk/19175/. Accessed: November 1, 2025.

[24] V. O. Kheyfets, W. O'Dell, T. Smith, J. J. Reilly, and E. A. Finol, "Considerations for numerical modeling of the pulmonary circulation—A review with a focus on pulmonary hypertension," *Journal of Biomechanical Engineering*, vol. 135, no. 6, 2013. [Online]. Available: https://asmedigitalcollection.asme.org/biomechanical/article-abstract/135/6/061011/370870/Considerations-for-Numerical-Modeling-of-the. Accessed: January 16, 2026.

[25] J. L. Mueller and S. Siltanen, *Linear and Nonlinear Inverse Problems with Practical Applications*, 1st ed. Philadelphia: SIAM, 2012.

[26] P. J. Vauhkonen, "Image reconstruction in three-dimensional electrical impedance tomography," Ph.D. dissertation, University of Kuopio, 2004.

[27] H. W. Engl, M. Hanke, and G. Neubauer, *Regularization of Inverse Problems*, 1st ed. Netherlands: Kluwer Academic Publishers, 1996.

[28] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, 1st ed. Philadelphia: SIAM, 1997.

[29] D. J. Higham, "Condition numbers and their condition numbers," *Linear Algebra and its Applications*, vol. 214, pp. 193–213, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0024379593000669. Accessed: March 15, 2026.

[30] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*, 1st ed. Philadelphia: SIAM, 2010.

[31] Å. Björck, *Numerical Methods for Least Squares Problems*, 2nd ed. Philadelphia, PA: SIAM, 2024.

[32] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, 2nd ed. Massachusetts: Elsevier, 2013.

[33] J. Nocedal and S. J. Wright, *Numerical Optimization*, 1st ed. New York, NY: Springer, 1999.

[34] A. Asl and M. L. Overton, "Analysis of the gradient method with an Armijo–Wolfe line search on a class of non-smooth convex functions," *Optimization Methods and Software*, vol. 35, no. 2, pp. 223–242, 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10556788.2019.1673388. Accessed: March 25, 2026.

[35] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, no. 4, pp. 561–580, 1992. [Online]. Available: https://epubs.siam.org/doi/10.1137/1034115. Accessed: May 14, 2026.

[36] V. A. Morozov, "On the solution of functional equations by the method of regularization," *Doklady Akademii Nauk SSSR*, vol. 167, no. 3, pp. 510–512, 1966. [Online]. Available: https://www.mathnet.ru/php/archive.phtml?wshow=paper\&jrnid=dan\&paperid=32161\&option_lang=eng. Accessed: May 14, 2026.

[37] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979. [Online]. Available: http://dx.doi.org/10.1080/00401706.1979.10489751. Accessed: May 14, 2026.

[38] P. Colli-Franzone et al., "A mathematical procedure for solving the inverse potential problem of electrocardiography: analysis of the time-space accuracy from in vitro experimental data," *Mathematical Biosciences*, vol. 77, no. 1-2, pp. 353–396, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0025556485901063. Accessed: May 14, 2026.

[39] P. R. Johnston and R. M. Gulrajani, "A new method for regularization parameter determination in the inverse problem of electrocardiography," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 1, pp. 19–39, 1997. [Online]. Available: https://ieeexplore.ieee.org/document/553710. Accessed: May 14, 2026.

[40] D. Krawczyk-Stańdo and M. Rudnicki, "Regularization parameter selection in discrete ill-posed problems — the use of the U-curve," *International Journal of Applied Mathematics and Computer Science*, vol. 17, no. 2, pp. 157–164, 2007. [Online]. Available: https://reference-global.com/article/10.2478/v10006-007-0014-3. Accessed: May 14, 2026.

[41] T. Wang, J. Karel, P. Bonizzi, and R. L. M. Peeters, "Influence of the Tikhonov regularization parameter on the accuracy of the inverse problem in electrocardiography," *Sensors*, vol. 23, no. 4, art. no. 1841, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/4/1841. Accessed: September 19, 2025.

[42] J. Alastruey, K. H. Parker, J. Peiró, S. M. Byrd, and S. J. Sherwin, "Modelling the circle of Willis to assess the effects of anatomical variations and occlusions on cerebral flows," *Journal of Biomechanics*, vol. 40, pp. 1794–1805, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.jbiomech.2006.07.008. Accessed: September 30, 2025.

[43] A. Fedorov et al., "3D Slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0730725X12001816. Accessed: May 16, 2025.

[44] Blender Foundation, "Index of /release/," 2025. [Online]. Available: https://download.blender.org/release/. Accessed: June 20, 2025.

[45] S. Y. Dotsenko and O. V. Yatsenko, "Elastic properties of pulmonary artery in chronic obstructive pulmonary disease," *Zaporozhye Medical Journal*, vol. 19, no. 1, pp. 26–29, 2017. [Online]. Available: http://dx.doi.org/10.14739/2310-1210.2017.1.91617. Accessed: September 19, 2025.

[46] W. C. Karl, "Regularization in image restoration and reconstruction," in *Handbook of Image and Video Processing*, 2nd ed., A. Bovik, Ed. Burlington: Academic Press, 2005, pp. 183–202. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780121197926500759. Accessed: June 10, 2026.

[47] Y.-L. Zou, R.-H. Xie, and A. Arad, "Numerical estimation of choice of the regularization parameter for NMR T2 inversion," *Petroleum Science*, vol. 13, no. 2, pp. 237–246, April 2016. [Online]. Available: http://dx.doi.org/10.1007/s12182-016-0093-6. Accessed: June 10, 2026.

Feasibility Study of Simplified Computational Models for Gait Simulation

Thiago Levy Rocha da Silva
Lab. of Heat Transfer, Thermodynamics and Numerical Simulation
Federal Rural University of the Semiarid Region
Mossoro, Brazil

Zoroastro Torres Vilar
Dep. of Engineering and Technology
Federal Rural University of the Semiarid Region
Mossoro, Brazil
ORCID: 0000-0002-8841-6689

Ramon Rudá Brito Medeiros
Dep. of Engineering
Federal Rural University of the Semiarid Region
Caraubas, Brazil
ORCID: 0000-0002-0923-1644

Diego Ariel de Lima
Dep. of Health Sciences
Federal Rural University of the Semiarid Region
Mossoro, Brazil
ORCID: 0000-0002-5702-1529

Victor Wagner Freire de Azevedo
Dep. of Engineering and Technology
Federal Rural University of the Semiarid Region
Mossoro, Brazil
ORCID: 0000-0003-2837-7785
victorwfreire@ufersa.edu.br

***Abstract*—The production of personalized plantar orthoses faces significant barriers due to the high cost of clinical equipment for capturing plantar pressure. As an alternative, this work investigates the technical limits of simplified computational models for human gait simulation, aiming to map the feasibility of a low-cost numerical approach for sizing these structures. The methodology involved developing simulations in ANSYS software, employing the Ogden hyperelastic model to represent the foot's soft tissue in a continuous geometry without the internal bone component. Dynamic impact tests were performed using the LS-DYNA solver to represent the specific moments of gait where dynamic loads exceed the static state. Results revealed that the absence of an internal bone structure leads to excessive geometric deformations and an artificial flattening of the plantar arch during the simulated impact. The study concludes that while the simplified continuous model fails to replicate real gait kinematics, it fulfills a crucial scientific role by delimiting the boundaries of biomechanical modeling, proving that the inclusion of the skeletal structure is mandatory for realistic gait simulation and effective orthosis design.**

***Index Terms*—gait simulation, Ogden model, plantar orthoses, numerical simulation**

I. Introduction

The manufacturing of personalized plantar orthoses is traditionally a process dependent on high-cost equipment and expensive commercial materials. In a scenario where gait asymmetries frequently lead to injuries and conventional design relies on costly static measurements, numerical simulation and 3D manufacturing emerge as viable alternatives to reduce costs and optimize projects.

Gait analysis is fundamental for assessing functional independence, diagnosing pathologies and designing assistive devices in medical, rehabilitative and sports contexts. The reliability of gait simulation as a research tool has been further reinforced by validation studies such as those conducted by Baxter et al. [1]. Their work demonstrated that gait simulators, when driven by population-specific in vivo inputs can successfully replicate the kinematics of the ankle, subtalar and talonavicular joints. These findings are particularly significant for orthotic and surgical design, as they prove that simulation platforms can accurately reproduce not only mean population patterns but also individual variability. However, experimental trials are often limited by the high number of sessions required to test multiple conditions, such as varying speeds and carried loads or by the physical burden of sensors on the subject [2]. Recent advancements in biomechanical modeling have produced high-fidelity simulators capable of replicating human gait with remarkable accuracy. Notable examples include the forward dynamic model developed by Ito et al. [3], which incorporates a detailed skeletal structure of 23 bones to simulate foot-ground interactions during the stance phase, which require up to 14 hours of supercomputing time. Likewise, Minagawa et al. [4] employed anatomically precise bone and ligament representations to investigate joint stability under various stress conditions. While these models provide deep insights into internal tissue loading, their implementation often requires high computational power. Nevertheless, a model more accessible and with low-cost numerical tools for plantar orthosis design is an excellent auxiliary tool for plantar orthosis design.

This study presents a feasibility analysis of simplified computational models for human gait simulation. The research investigates the technical limits of replacing complex anatomical structures with the adoption of a continuous geometry without bone components. To ensure mechanical fidelity, the soft tissue was represented using the 1st-order Ogden hyperelastic model, which is capable of capturing the non-linear strain-stiffening behavior of plantar fat.

Using ANSYS software and the LS-DYNA solver for transient dynamic analysis, this work aims to map the numerical stabilization and convergence of this low-cost proposal. By delimiting the boundaries of simplified biomechanical modeling, this research provides essential criteria for the development of accessible tools for plantar orthosis sizing and design.

II. Theoretical Background

This section establishes the biomechanical and mathematical principles used to evaluate human gait and the mechanical behaviour of plantar tissues through numerical simulation.

A. Gait biomechanics and plantar pressure distribution

Human gait is a cyclic activity composed of repetitive patterns, primarily divided into stance and swing phases [2]. For precise orthotic design, the stance phase is further categorized into six stages: Heel Strike, Foot Flat, Mid-Stance, Heel Off, Terminal Stance and Toe Off (see Fig. 1). While conventional orthotic design often relies on static measurements, the highest force concentrations occur during locomotion, where the body weight is abruptly transferred over small areas of the foot. During these moments, dynamic loads can reach twice the magnitude of those observed in a static posture [3]. Computational simulation allows for the assessment of joint torques and internal stresses in these conditions, which are otherwise difficult to reproduce or measure experimentally.

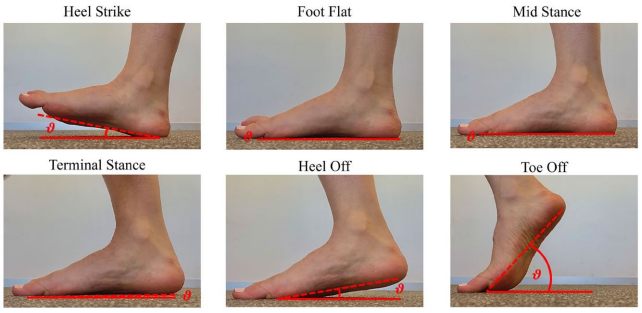


Fig. 1: Visual scheme of gait stages [5]

Plantar pressure distribution is a critical parameter for identifying high-risk zones where abnormal pressure peaks are directly linked to musculoskeletal injuries and complications like diabetic foot ulcers. During a normal step, the Center of Pressure follows a specific trajectory, typically progressing from the heel medially to the midfoot and then to the hallux during the propulsive phase [3, 6].

B. Ogden hyperelastic constitutive model

The representation of plantar soft tissues, such as the fat pad, requires numerical models capable of capturing complex non-linear mechanical behaviours. Unlike linear elastic materials, biological tissues exhibit strain-stiffening, a phenomenon where the material becomes progressively more rigid as it undergoes compression. To replicate this in a Finite Element environment, the Ogden hyperelastic model is widely employed [3].

The Ogden hyperelastic model is widely utilized in soft tissue biomechanics due to its high capacity to capture non-linear and isotropic strain response under large deformations [7, 8]. The strain energy density function is expressed as

$$W = \sum_{i=1}^{N} \frac{\mu_i}{\alpha_i} \left(\lambda_1^{\alpha_i} + \lambda_2^{\alpha_i} + \lambda_3^{\alpha_i} - 3\right), \tag{1}$$

where N represents the order of the model and μ_1 and α_i are empirical parameters of the materials. λ_i represent the principal stretch ratios of the material.

While high-fidelity models often combine the Ogden formulation with viscoelastic properties to account for time-dependent responses [3], this work adopts a simplified 1st-order approach to maintain computational efficiency for orthosis design. However, comparative tests with linear materials (e.g., structural steel) demonstrate that using the Ogden model is mandatory; linear models fail to account for the large local deformations necessary to generate realistic pressure maps, highlighting the importance of hyperelasticity in biomechanical simulations.

III. Methodology

This section details the methodological approach used to investigate the feasibility of computational human gait simulation without relying on clinical force-capture machinery. The study employs a purely numerical framework focused on a simplified geometric model of the human foot.

A. Geometry and Material properties

The 3D geometric domain was obtained from a surface scan available on the GrabCAD library [9]. The model represents a male adult foot with a length of 251.37 mm and a maximum width of 108.74 mm, corresponding to approximately a size 40/41 (EU). Before simulation, the mesh underwent surface smoothing to remove scanning artifacts. The final domain was discretized using tetrahedral elements (8 mm global, 3 mm plantar refinement).

Two material scenarios were defined for comparison:

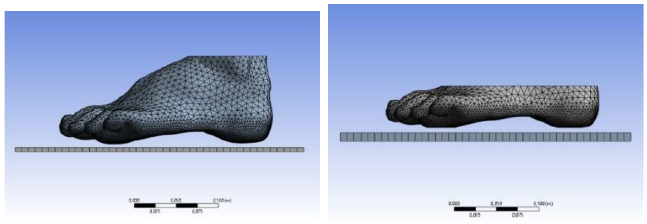
- Plantar Soft Tissue: Represented by the 1st-order Ogden hyperelastic constitutive model. This model is essential for capturing the non-linear behavior and large deformations of biological tissues under compression. The parameters used were $\mu_1 = 0.026$ MPa and $\alpha_1 = 0.523$ [10]. It is critical to note that while values of $\alpha_1 > 1$ typically model strain-stiffening behavior under compression, the adopted parameter of $\alpha_1 = 0.523$ mathematically

results in a non-linear strain-softening effect under large compressive strains. As observed in the results, this material response further facilitated the rapid deformation and eventual collapse of the plantar arch during dynamic trials.

- Control Material: Structural Steel was used as a linear elastic material to evaluate the influence of extreme stiffness on stress distribution and to isolate the effects of geometric deformation.

B. Numerical setup and simulation scenarios

The domain was discretized using tetrahedral elements with a global size of 8 mm. A local mesh refinement was applied to the plantar surface, reducing element size to 3 mm to ensure higher accuracy in contact regions, as shown in Fig. 2.



(a) (b)

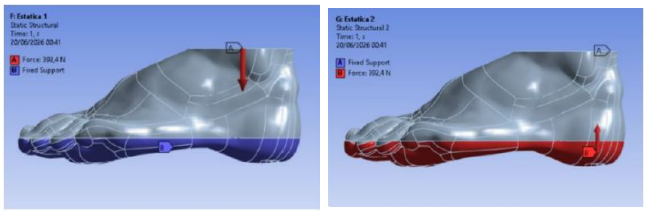
Fig. 2: Geometric representation of the foot: (a) geometry adopted in Cases 1-5 and (b) in Case 6.

The loading parameters assumed a standard adult mass of 80 kg, resulting in a design load of 392.4 N (40 kg – half human representative weight) for a single foot. Simulations were performed using ANSYS software, divided into Static Structural analysis and LS-DYNA transient dynamic analysis to represent the gait impact where dynamic loads exceed static states.

Six study cases were established to evaluate numerical stabilization and pressure distribution:

- Static scenarios:
  - Case 1 (Superior Load and Plantar Fixation): Fixed support on the sole with a vertical force applied to the top of the leg (see Fig. 3a);
  - Case 2 (Plantar Load and Superior Fixation): Fixed support on the top of the leg with a distributed force applied perpendicularly to the sole (see Fig. 3b).
- Dynamic impact scenarios:
  - Case 3 (Pure Inertial Impact): Pure inertial impact with constant force of 392.4 N (see Figs. 4a, showing the fixed support in blue and movement restriction in yellow; and 4b, showing the foot uniform velocity of 4.1 m/s);
  - Case 4 (Restricted Motion Impact): Restricted motion (blocking vertical and horizontal translations at the top) with damping controls to stabilize numerical oscillations (see Fig. 4c, showing the fixed support in blue and applied force in red);
  - Case 5 (Guided Motion with Damping): Guided motion adding an initial velocity of 4.1 m/s to the Case 4 configuration to represent high-intensity running impact speed (see Figs. 4d, showing the fixed support in blue, force in red and movement restriction in yellow; and 4e, showing foot velocity);
  - Case 6 (Reduced Geometric Domain): Reduced geometry version using the same conditions as Case 5 to optimize processing time (see Fig. 4f, showing the fixed support in blue, movement restriction in yellow and applied force in red).

To simulate the foot-ground interaction during impact, a Surface-to-Surface contact formulation was implemented in LS-DYNA. The interface utilized a static friction coefficient of $\mu_s = 0.6$ and a dynamic friction coefficient of $\mu_d = 0.4$, representing typical contact conditions between human skin (or elastomeric orthotic materials) and a rigid floor.



(a) (b)

Fig. 3: Boundary conditions in static cases: (a) Case 1, superior load applied at the ankle (highlighted in red) with fixed plantar support (highlighted in blue); (b) Case 2, plantar load applied at the heel (highlighted in red) with fixed superior support.

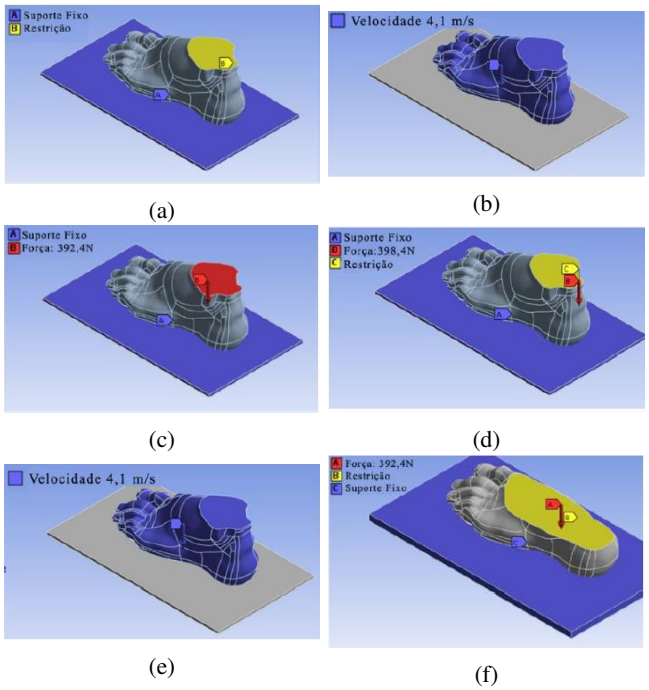
IV. Results

This section presents the results of the numerical simulations performed in ANSYS categorized into static structural analyses and transient dynamic impact tests. The primary objective was to evaluate the stability of the simplified computational model and its capacity to predict plantar pressure distribution using the Ogden hyperelastic model.

A. Static analysis and numerical response

The initial phase tested two constraint arrangements under a constant load of 392.4 N to determine the mechanical response of the simplified model. The velocity of 4.1 m/s is here correctly classified as high-intensity running impact, as typical walking (gait) speeds range between 1.2 and 1.4 m/s [3]. This extreme condition was chosen to test the numerical limits of the simplified geometry:

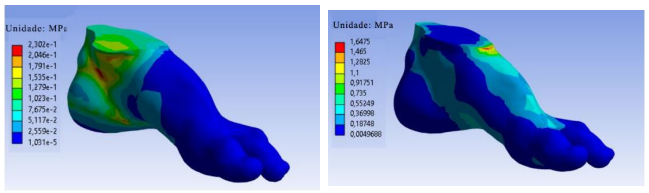
- Case 1: Applying the load at the ankle with a fixed sole resulted in artificial numerical locking. Because the plantar surface was completely restricted, the stress flow at the contact interface was nullified, concentrating instead at the nearest geometric discontinuities and preventing the natural compression of the hyperelastic tissue (see Fig. 5a);



(a) (b) (c) (d) (e) (f)

Fig. 4: Boundary conditions in dynamic cases: (a)-(b) Case 3, (c) Case 4, (d)-(e) Case 5 and (f) Case 6.

- Case 2: Inverting the arrangement caused deformation energy to migrate away from the sole and toward the plantar arch and the anterior ankle transition zone. This behavior indicates that the continuous geometry, when pulled in this configuration, concentrates stress at superior discontinuities, simulating the stretching of dorsal support tissues rather than plantar compression (see Fig. 5b).



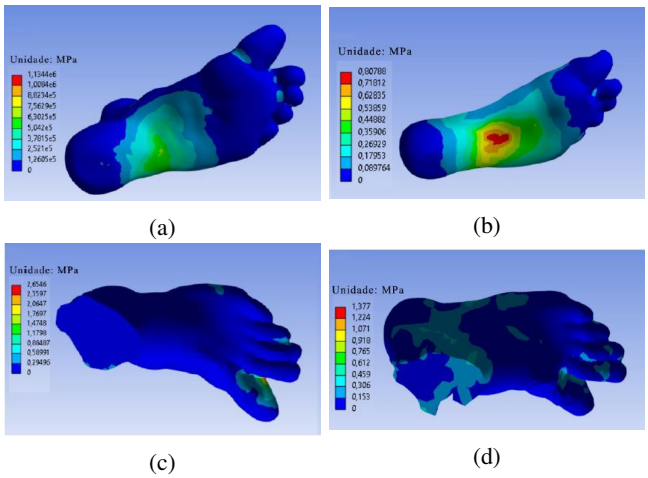
(a) (b)

Fig. 5: Results for pressure distribution along the foot for (a) Case 1 and (b) Case 2.

The comparison established that for transient stages, the model requires freedom of movement at the base to generate biologically coherent pressure maps through mutual compression against the orthotic surface.

B. Dynamic impact analysis

The impact simulations revealed that a model consisting exclusively of hyperelastic soft tissue lacks the structural stiffness required to reproduce human foot biomechanics during ground contact. While Cases 3 and 6 provided clear numerical outputs (see Figs. 6a and 6b), Cases 4 and 5 exhibited severe stabilization issues (see Figs. 6c and 6d). The absence of internal rigid components meant the model suffered from excessive geometric deformations, which compromised system stability and produced mechanical responses divergent from realistic gait patterns.

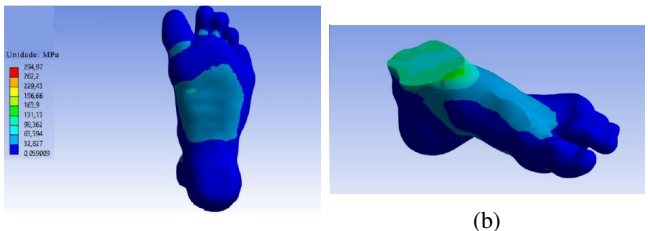


(a) (b) (c) (d)

Fig. 6: Results for impact simulation for (a) Case 3, (b) Case 6, (c) Case 4 and (d) Case 5.

C. Comparative Analysis: Linear vs. Hyperelastic Material

To isolate the effects of the Ogden constitutive model, a control test was conducted using the linear properties of Structural Steel. The extreme stiffness of the linear material eliminated numerical noise and local instabilities at the contact interface. However, this material acted as a rigid body, effectively "shielding" the model against the large local deformations necessary for realistic biomechanical representation. The results presented in Fig. 7 quantify this mechanical discrepancy. Under the same loading conditions, the model using structural steel reached an extreme peak von Mises stress of 294.97 MPa concentrated at the plantar contact zone, with virtually zero compliance. Conversely, the hyperelastic Ogden model distributed the load through localized compliance, limiting the peak stress to only 0.807 MPa (as shown in Case 2, Fig. 5b). While linear models facilitate rapid solver convergence, they completely mask the soft tissue behavior required to design effective plantar orthoses.



(a) (b)

Fig. 7: Results for the simulation with structural steel: (a) Plantar region and (b) Dorsal foot.

The efficiency in solution for the studied simplified models is presented on the results in Tab. I, where the CPU time for each case is presented.

The static analyses exhibited highly divergent processing times. Case 1 required approximately 8 hours to solve, a consequence of the artificial numerical locking caused by the fully constrained sole. In contrast, by inverting the boundary conditions in Case 2, the system achieved convergence in a 92% reduction of computational time. For the transient dynamic simulations using the LS-DYNA explicit solver, processing times ranged from 2 minutes (Case 3) to less than 20 minutes (Cases 5 and 6).

Interestingly, Case 6 (19 min 39 s) required slightly more processing time than Case 5 (18 min 11 s), despite utilizing a reduced geometric domain. In explicit dynamic solvers, the integration time step (Δt) is strictly governed by the Courant-Friedrichs-Lewy (CFL) condition, where

$$\Delta t \leq \frac{l_c}{c}.$$

Here, l_c represents the characteristic length of the smallest element in the mesh and c is the speed of sound in the material. During the geometric truncation of the foot model to isolate the plantar region (Version 2), the relative concentration of refined 3 mm elements increased and mesh generation at the artificial boundaries created slightly more distorted elements with smaller characteristic lengths. This decreased the critical time step, forcing the solver to execute more iterations and slightly increasing the total CPU time. Nevertheless, both Cases 5 and 6 solved in under 20 minutes, which represents an extremely high computational saving (over 97%) compared to the 14 hours required by anatomically detailed 3D models running on high-performance supercomputers. This duality corroborates with the fact that although simplified models did not replicate the precise physics of the bone kinematics, they should indeed be considered as alternative for quick solutions and design in structural evaluation of orthosis design.

This trade-off between anatomical realism and processing speed demonstrates that although simplified models fail to replicate dynamic bone kinematics, they present a highly viable, near-real-time option for fast static structural evaluation in orthosis design.

V. Conclusions

This work investigated the feasibility of using simplified computational models for human gait simulation as a tool to obtain stress distribution on the foot sole and support the design of personalized plantar orthoses in the future. The research employed numerical simulations in ANSYS, utilizing the Ogden hyperelastic model to represent plantar soft tissue in a continuous geometry devoid of internal skeletal structures.

The different simulation scenarios provided clear conclusions regarding the limitations of anatomical simplification. While static tests revealed a high sensitivity to boundary conditions and movement constraints, the dynamic impact tests proved that representing the foot as a bone-less continuous block is inadequate for simulating real gait kinematics. The absence of a rigid internal component, which represents the bones and ligaments, removed the mechanical constraints necessary to support dynamic loads. Consequently, during the simulated impact, the hyperelastic tissue suffered excessive geometric deformations, leading to the artificial collapse and flattening of the plantar arch against the ground.

Furthermore, comparative analyses showed that although linear materials (such as structural steel) provide better numerical stabilization, they mask the realistic "strain-stiffening" behavior of soft tissue required for accurate pressure mapping. Therefore, while this research serves a relevant scientific role by delimiting the boundaries of biomechanical modeling, it concludes that the complexity of the human foot cannot be neglected in dynamic simulations.

Ultimately, to design realistic plantar orthoses and replicate dynamic pressure distributions effectively, the inclusion of the skeletal structure is a mandatory requirement. Future efforts should focus on optimizing these skeletal models to balance the need for anatomical realism with the computational efficiency required for practical clinical applications.

References

[1] Josh R Baxter, Daniel R Sturnick, Constantine A Demetracopoulos, Scott J Ellis, and Jonathan T Deland. Cadaveric gait simulation reproduces foot and ankle kinematics from population-specific inputs. *Journal of Orthopaedic Research*, 34(9):1663–1668, 2016.

[2] Cristina Brambilla, Giulia Beltrame, Giorgia Marino, Valentina Lanzani, Roberto Gatti, Nicola Portinaro, Lorenzo Molinari Tosatti, and Alessandro Scano. Biomechanical analysis of human gait when changing velocity and carried loads: Simulation study with opensim. *Biology*, 13(5):321, 2024.

[3] Kohta Ito, Yuka Matsumoto, Hiroyuki Seki, Takeo Nagura, and Naomichi Ogihara. Simulating human foot mechanics during walking based on an anatomically detailed forward dynamic finite element model. *Annals of Biomedical Engineering*, pages 1–15, 2026.

[4] Akinobu Minagawa, Naomichi Ogihara, Makoto Kubota, Yuka Matsumoto, Kohta Ito, Tadashi Kimura, Takumi Kihara, Hirotaka Fukushima, and Mitsuru Saito. Mechanical contribution of the anterior talofibular ligament to ankle stability: 3d anatomical finite element analysis. *Journal of Orthopaedic Research®*, 44(7):e70241, 2026.

[5] Karl B Landorf. Plantar heel pain and plantar fasciitis. *BMJ clinical evidence*, 2015:1111, 2015.

[6] Fahni Haris, Yih-Kuen Jan, Ben-Yi Liau, Chang-Wei Hsieh, Wei-Cheng Shen, Chien-Cheng Tai, Yin-Hwa Shih, and Chi-Wen Lung. The effects of different inner pressures of air insoles and walking durations on peak plantar pressure. *Medicine*, 102(43):e35704, 2023.

[7] Raymond W Ogden. Large deformation isotropic elasticity–on the correlation of theory and experiment for incompressible rubberlike solids. *Proceedings of the*

TABLE I: Processing Times and Computational Efficiency Across Simulation Cases

Simulation Case	**Analysis Type**	**Geometry Version**	**Processing Time**
Case 1	Static Structural	Version 1 (Complete)	8 h 01 min 00 s
Case 2	Static Structural	Version 1 (Complete)	38 min 24 s
Case 3	Transient Dynamic	Version 1 (Complete)	1 min 59 s
Case 4	Transient Dynamic	Version 1 (Complete)	3 min 21 s
Case 5	Transient Dynamic	Version 1 (Complete)	18 min 11 s
Case 6	Transient Dynamic	Version 2 (Plantar Reduced)	19 min 39 s
State-of-the-Art*	Explicit FEA (23 Bones)	Full Anatomical	$\approx$ 14 h 00 min 00 s

*Detailed musculoskeletal dynamic FE model using supercomputing resources [3].

Royal Society of London. A. Mathematical and Physical Sciences, 326(1567):565–584, 1972.

[8] Jay D Humphrey. Continuum biomechanics of soft biological tissues. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 459(2029):3–46, 2003.

[9] GrabCAD. Scanned foot — 3d cad model library. `https://grabcad.com/library/scanned-foot-1`, 2026. Acessed in: Jun 23th 2026.

[10] Matthew J Lohr, Gabriella P Sugerman, Sotirios Kakaletsis, Emma Lejeune, and Manuel K Rausch. An introduction to the ogden model in biomechanics: benefits, implementation tools and limitations. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 380(2234):20210365, 2022.

Framework for 3D Urban Digital Twin - Reconstruction and Simulations for Disaster Management

Jonathan de Oliveira
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
Grad. Program in Health Tech. (PPGTS), PUCPR
Curitiba, Brazil
gmjo92@gmail.com

Lukas Nikolaus Ullmann
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
luullmann@ukaachen.de

Tobias Martin
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
tmartin@ukaachen.de

Andreas Follmann
Dept. of Anesthesiology
Faculty of Medicine
RWTH Aachen University
Aachen, Germany
afollmann@ukaachen.de

Mauren Abreu de Souza
Grad. Program in Health Tech. (PPGTS), PUCPR
Curitiba, Brazil
mauren.souza@pucpr.br

***Abstract*—This paper presents a computational framework for rapid-acquisition Unmanned Aerial Vehicle photogrammetric data into interactive 3D digital twins within Unreal Engine 5 for disaster management and flood simulation. The methodology was applied across a stretch of the Atuba River basin covering a 0.408 km^2 study area, where 719 images were successfully aligned within a total processing time of 31 minutes and 29 seconds. To enable real-time execution, the high-resolution geometric mesh was optimized in RealityScan environment and decimated into a 5-million triangle topology distributed across 21 structural parts. Quantitative reconstruction validation yielded a Structural Similarity Index of 0.3285, a Peak Signal-to-Noise Ratio of 16.0247 dB, and a stable Learned Perceptual Image Patch Similarity of 0.3620, indicating that pixel-level fidelity is limited while macro-geometric boundaries relevant to physical rigid-body collisions appear comparatively better preserved. The flood simulation was performed in Unreal Engine 5, modeling the progression of the water body in four scenarios, quantifying severe impacts under the maximum flood parameter, which inundated 38 out of 40 structural assets and caused a 35% disruption along the main highway. Finally, it a proof-of-concept decision framework was proposed to optimize medical logistics, establishing the viability of translating rapid aerial surveys into immediate tactical routing directives for climate-induced crises.**

***Index Terms*—Digital Twin, UAV Photogrammetry, Flood Simulation, Unreal Engine 5, Image Quality Assessment, Disaster Management.**

I. Introduction

A 3D Digital Twin (DT) is a computational framework that continuously synchronizes a physical environment with a digital representation [1]. The deployment of DTs has become essential for the management and operation of smart cities [2]. In disaster scenarios, these applications provide critical data for decision-making, coordination, and efficient resource reallocation [3].

The impact of this technique was demonstrated in the reconstruction of historical monuments after catastrophic events, such as the Notre Dame Cathedral [4]. Beyond heritage conservation, this technique can be tailored to specific operational requirements within the scope of disaster medicine, blending 3D reconstruction and geospatial data into 3D digital twin environments to deliver dynamic, real-world representations of affected areas. In the context of disaster medicine, integrating spatial information into 3D environments enhances real-time situational analysis and critical healthcare decision-making processes. While recent frameworks by Lagap and Ghaffarian [5] as well as Ge and Qin [6] have enhanced urban resilience models, executing these simulations in real time remains computationally expensive [7]. This creates a critical disconnect between initial spatial reconstruction and active emergency response loops, limiting their immediate clinical and tactical utility.

Recent extreme weather events, such as the floods in Rio Grande do Sul, Brazil [8] and Bavaria, Germany [9], underscore the urgency of developing aid solutions that are adaptable to diverse international contexts, despite territorial and demographic variances. Traditional aerial monitoring platforms offer high-resolution post-disaster data, but the prolonged processing times required for dense geometric reconstruction limit their immediate, field-deployable utility during active emergency response loops [10]. Furthermore, translating

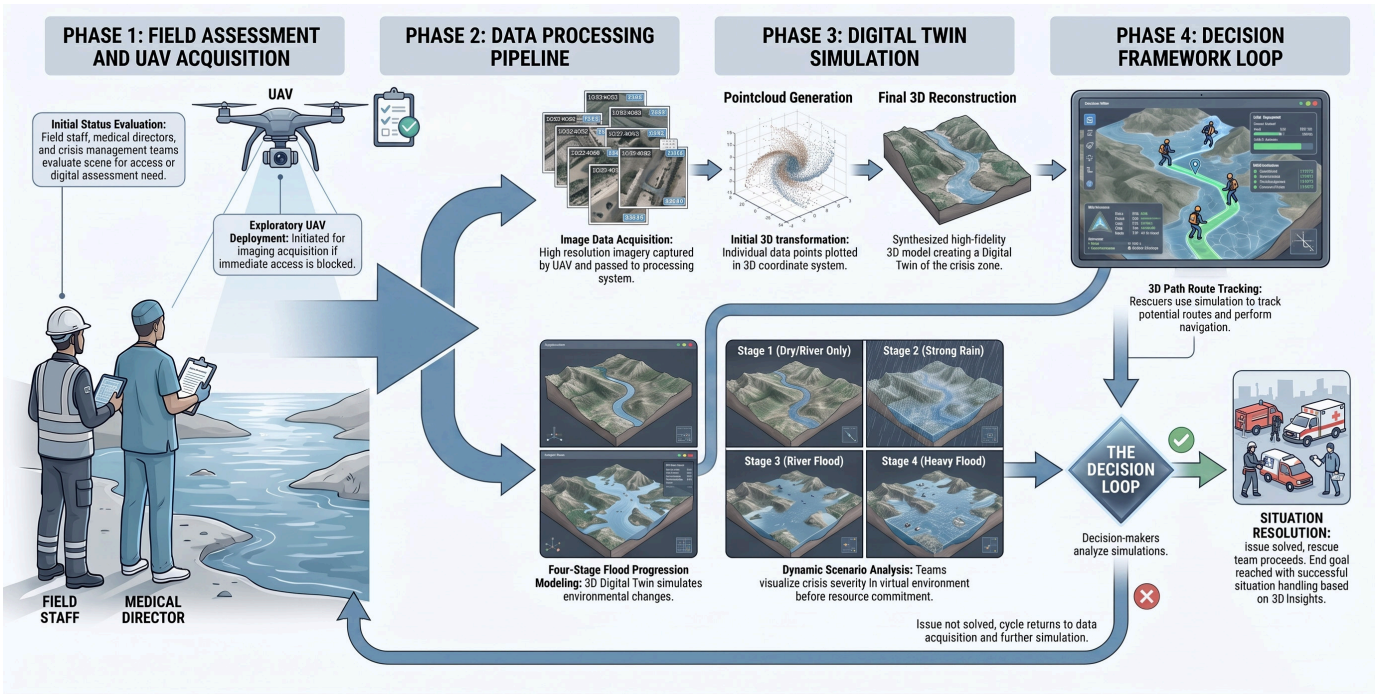


Fig. 1. Graphical abstract showing the entire deployment framework, from field assessment and UAV acquisition, data processing pipeline, digital twin simulation and decision framework loop.

raw topographic changes into active logistical routing for emergency medical supply distribution presents an ongoing bottleneck in disaster management workflows [11].

This challenge operates under the premise that structural alterations observed in urban structures can provide preliminary indicators of the impacts sustained by buildings and affected areas. The accessibility conditions of urban highways directly dictate the mobility of emergency response teams, access to critical regions, and the timely transportation of victims, medical supplies, and healthcare resources. Within this context, leveraging interactive game engines and real-time spatial monitoring becomes important to assess the relationship between raw topographic data and actionable health logistics [12]. To address these operational limitations, this paper proposes an integrated framework that applies rapid UAV photogrammetry to generate a 3D digital twin for simulations within Unreal Engine environment using a water body progression. The technological contributions of this architecture are:

- An end-to-end execution pipeline, which can communicate between unstructured UAV data and localized field action,
- A multi-stage mesh optimization and decimation workflow that minimizes computational rendering while preserving macro-geometric elevation profiles,
- A proof-of-concept decision framework designed to support territorial accessibility analysis, damage assessment, and health logistics, such as rescue operations and resource distribution, while integrating spatial information into 3D digital twins to enhance situational analysis in complex health emergencies.

The conceptual workflow is shown in the Graphical Abstract (Fig. 1) by each stage: (1) Initial assessment: field staff or medical director evaluating the situation, (2) UAV Data Acquisition: highlighting the exploratory drone flights, including checks for weather conditions, battery capacity, and flight plans, (3) 3D Reconstruction: visualizing the transition from image data acquisition to point cloud generation and full 3D reconstruction, (4) Flood Simulation: incorporating the progression of the flooding, and (5) Decision Framework: integrating the decision-making loops where rescuers use the 3D model and path route simulations to determine if the situation can be handled.

II. Theoretical Background

A. 3D Environmental Reconstruction Pipeline

The digital twin generation workflow employs a modern photogrammetric architecture within the desktop-rebranded RealityScan engine (incorporating core algorithms from RealityCapture, Epic Games). The structural conversion of unordered 2D aerial imagery into an actionable 3D mesh operates across two fundamental computing items:

1) Structure-from-Motion (SfM): Feature detection is initiated using scale-invariant descriptors to establish tie points across overlapping frames. The relative camera poses, internal calibrations, and spatial orientations are resolved via a global bundle adjustment optimization, minimizing the global reprojection error to construct a sparse point cloud [13].
2) Multi-View Stereo (MVS) & Surface Reconstruction: To transition from a sparse point cloud to dense surface to-

pography, a depth-map computation is computed across multi-view matrices. Surface meshing is formalized using a variational or Delaunay-based signed distance field (SDF) optimization [14], yielding a vectorized polygon mesh defined as:

$$M = \{V, E, F\} \tag{1}$$

where $V \in \mathbb{R}^3$ represents the spatial coordinates of the vertices, E denotes connectivity edges, and F specifies the polygon faces mapping the macro-elevation bounds of the terrain.

For wide-area disaster modeling (0.4 km^2), computational efficiency necessitates strict decimation constraints to produce static physics colliders that are fully compatible with real-time runtime rendering and fluid simulation processing.

B. Image Quality & Structural Fidelity Metrics

To objectively evaluate the fidelity of the reconstructed digital twin against reference aerial imagery, three independent mathematical formulations are evaluated: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

1) Peak Signal-to-Noise Ratio (PSNR): PSNR serves as a standard metric to compute pixel-level absolute logarithmic error differences between the ground-truth frame (I) and the projected digital twin render (K), defined over an $M \times N$ image matrix:

$$\text{MSE} = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - K(i,j)]^2 \tag{2}$$

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{3}$$

where MAX_I is the maximum possible pixel value of the image. While highly sensitive to global illumination variations and absolute pixel displacement, PSNR lacks direct alignment with human visual structural perception.

2) Structural Similarity Index Measure (SSIM): To capture structural degradation, SSIM shifts the assessment from independent pixel errors to perceived luminance (l), contrast (c), and structural configuration (s) comparisons within localized window kernels [15]:

$$\text{SSIM}(x, y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma \tag{4}$$

When parameter weights are normalized ($\alpha = \beta = \gamma = 1$), the index simplifies to:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where μ and σ^2 dictate the mean and variance profiles of the windows, σ_{xy} is the cross-covariance, and C_1, C_2 act as stabilizing constants to avoid division by zero. SSIM bounds scale between $[-1, 1]$, where a value of 1 signifies absolute structural replication.

3) Learned Perceptual Image Patch Similarity (LPIPS): To overcome the mathematical rigidities of classical metrics, LPIPS utilizes deep feature representations to emulate human perceptual judgement [16]. Given a deep convolutional network (e.g., VGG or AlexNet), features are extracted from layer l, normalized across channels, and scaled by vector w_l to compute the L_2 distance:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}_{h,w}^l - \hat{y}_{l,h,w}^0 \right) \right\|_2^2 \tag{6}$$

An LPIPS value closer to 0.0 signifies a minimizing perceptual distance, demonstrating authentic texture preservation and fewer perceived visual artifacts.

III. Methods

A. UAV Data Acquisition

The initial environmental asset data was captured via an unmanned aerial vehicle (UAV) survey covering a critical flood-prone stretch of the Atuba River basin, as illustrated in the full orthomosaic and digital elevation layout (Fig. 2). The survey encompassed a total coverage area of 0.408 km^2, capturing 733 high-resolution aerial frames from an average flying altitude of 122 m. Imagery was acquired using an onboard FC3411 camera sensor with a fixed focal length of 8.38 mm and a native resolution of 5472×3648 pixels (2.51×2.51 μm pixel size), yielding an average ground sampling distance (GSD) of 3.06 cm/pix.



Fig. 2. Full geographic view of the 0.408 km^2 Atuba River study region, showing the generated orthomosaic overlay, digital elevation characteristics, and key architectural assets modeled for the digital twin.

B. Photogrammetric Processing

Photogrammetric processing and 3D reconstructions of the dataset were performed on RealityScan desktop version environment. A laptop was used to handle the computational demand of the dense point cloud and mesh generation. The system is equipped with an Intel Core i9-14400HX processor, 32 GB of DDR5 RAM (5600 MHz) and an NVIDIA GeForce RTX 4060 GPU.

Prior to dense matching, an image downscale factor of 2 was applied to optimize processing efficiency. Out of the 733 captured frames, 719 images were successfully registered during the alignment stage, yielding a total of 3.5M projections and an average track length of 3.0. This optimization process generated a sparse point cloud containing 1.1M points. The spatial alignment accuracy was validated by a mean reprojection error of 0.85 pixels (with a median of 0.79 pixels and a maximum error threshold of 2.00 pixels). The reconstructed environment was successfully geo-referenced, establishing its absolute spatial positioning despite operating without explicit metric scaling at this preliminary stage.

To balance structural fidelity with the strict real-time rendering constraints required for hydrodynamic simulations in Unreal Engine 5, a multi-stage mesh optimization, decimation, and texturing strategy was implemented. The initial dense reconstruction produced a raw high-polygon model (*HighPoly*) comprising 97M triangles and 48M vertices. To eliminate non-manifold geometries and topological noise, filtering and smoothing algorithms were applied, refining the mesh structure (*HighPoly Smoothed*) to 97M triangles and 48M vertices.

The model was then subjected to a severe decimation process to generate an optimized low-poly variant (*LowPoly*) consisting of 5M triangles and 2.5M vertices, which was partitioned into 21 discrete mesh parts to facilitate efficient spatial streaming and occlusion culling within the virtual environment. To compensate for the loss of geometric detail resulting from the decimation, high-frequency surface details and color data from the *HighPoly Smoothed* topology were baked onto the *LowPoly* mesh via a reprojected texturing workflow. A single 8192×8192 pixel texture atlas was generated using a maximal texture count unwrapping style, achieving a texture utilization rate of 87%. The visual fidelity was preserved by computing a 32-bit BGRA diffuse color layer and a 128-bit floating-point RGBA normal map layer, both reprojected directly from the smoothed high-resolution geometry.

The time for the complete reconstruction pipeline totaled 31 minutes and 29 seconds. This processing duration was distributed across the following pipeline stages: 2 minutes and 29 seconds for feature detection, 4 minutes and 39 seconds for image registration, 9 minutes and 25 seconds for depth-map computation, 12 minutes and 39 seconds for dense meshing, 1 minute and 08 seconds for topological post-processing, and approximately 1 minutes and 09 seconds for mesh unwrapping and layer reprojection.

C. Reconstruction metrics

To evaluate reconstruction fidelity, each rendered frame from the digital twin was paired with its corresponding reference aerial photograph by matching file name between a held-out reference set and its rendered counterpart; files without a match in the other set were excluded from the comparison. Prior to metric computation, both images in a pair were resized (Lanczos resampling) to the smaller of their two native resolutions and converted to RGB. The rendered image was then photometrically aligned to the reference through an independent per-channel linear regression (least-squares fit of pixel intensity, applied only to the rendered image, reference left unaltered), correcting for systematic tonal, white-balance, and radiometric differences between the physical UAV sensor and the RealityScan render pipeline. This calibration step isolates the structural component of the comparison from purely photometric mismatch and follows standard practice in real-vs-synthetic view evaluation. SSIM was computed on the full RGB image (multichannel, scikit-image implementation) with a data range of 255; PSNR was computed over the same 8-bit range. LPIPS was computed with an AlexNet backbone on images resized to 512×512 pixels. The exact number of image pairs contributing to the averages in Table I (Section IV-A), and whether a sub-sampling ratio was applied to the matched set, could not be reliably reconstructed from archived data at the time of revision; this is acknowledged as a documentation gap rather than reported as a specific figure.

IV. Results and Discussion

A. Structural Reconstruction Fidelity

The quantitative assessment of the photogrammetric digital twin was completed by evaluating the absolute pixel, structural, and perceptual distance errors across the dataset. The aggregated statistical performance is summarized in Table I, while the localized distribution behavior across the 0.408 km^2 study area is illustrated via density distributions in Fig. 3.

TABLE I
Visual Fidelity Evaluation of the Reconstructed 3D Environment

Metric	Average	Min	Max	Std Dev
SSIM	0.3285	0.0738	0.8002	0.1303
LPIPS	0.3620	0.1911	0.5950	0.0597
PSNR	16.0247	11.0378	20.1108	1.3460

The mean SSIM of 0.3285 and PSNR of 16.02 dB indicate limited pixel-level and structural agreement between the rendered and reference imagery, and should not be interpreted as evidence of high pixel-level reconstruction fidelity. As shown by the wide probability density distributions in Fig. 3, the largest deviations concentrate in regions with dynamic or ambiguous surface properties: open river water, where specular reflectance and temporal change between capture and render are pronounced; dense tree canopies, whose fine, repetitive geometry is poorly resolved by multi-view stereo and further degraded by mesh decimation; asphalt roadways,

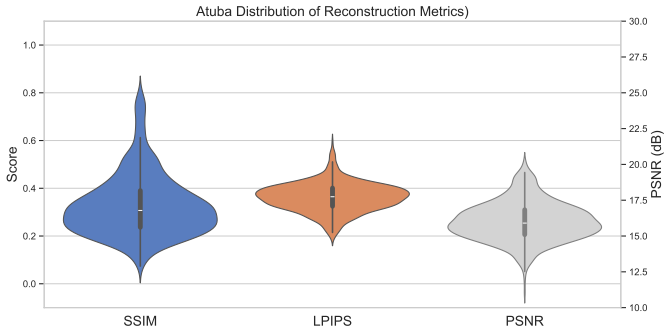


Fig. 3. Probability density and distribution profiles (Violin Plot) of the SSIM, LPIPS, and PSNR (dB) metrics computed over the reconstructed Atuba river basin terrain sample frames.

which combine low native texture with variable glare; and areas affected by residual viewpoint mismatch between the UAV capture pose and the virtual render camera. In contrast, the comparatively tighter LPIPS distribution (mean 0.3620, std 0.0597) indicates that macro-scale perceptual structure and topographic boundaries are better preserved than pixel-level detail. This should be read as evidence that coarse geometric layout survives decimation and not as an indication of overall high-fidelity reconstruction, which the SSIM and PSNR results do not support.

B. Unreal Engine 5 Flooding Progress Simulation

The operational deployment of the digital twin was demonstrated by implementing the decimated low-poly mesh as a static physics collider topology within Unreal Engine 5. The physical accuracy of the terrain geometry directly governs water body volume distribution algorithms during runtime.

The quantitative impacts of the multi-stage water body progression across the four simulated environments—*Dry*, *Strong Rain*, *River Flood*, and *Heavy Flood*—are structuralized in Table II. The sequential visual state changes are mapped from a regional perspective, Fig. 5, and a semantic path safety layout, Fig. 6. During both the baseline *Dry* and *Strong Rain* phases, the fluid vector remains entirely restricted within the natural riverbanks (35.28% and 35.72% area occlusion, respectively), resulting in zero architectural inundation.

Topological failure occurs during the transition to the *River Flood* stage. The over-bank water body propagation drives the total affected surface area to 53.37%, directly submersing 22 out of the 40 tracked structural assets and generating a 25% structural obstruction across the primary transportation highway. Under the maximum parameterization of the *Heavy Flood* scenario, the sub-basin reaches structural saturation, culminating in 77.58% total area submergence, severe structural flooding of 95% of the architectural zone ($38/40$ buildings), and a 35% logistic roadblock along the main highway infrastructure.

C. Decision Framework

To operationalize these simulations for field disaster management and medical emergency logistics, the data routing

TABLE II
FLOOD ASSESSMENT ACROSS UNREAL ENGINE SIMULATION (TOTAL IMAGE AREA: 1.96M PIXELS)

Flood Scenario	Affected Area (%)	Non-Affected Area (%)	Affected Buildings (Qty/40)	Highway Disruption (%)
Dry (River Only)	35.28	64.72	0	0
Strong Rain	35.72	64.28	0	0
River Flood	53.37	46.63	22	25
Heavy Flood	77.58	22.42	38	35

follows a tiered complexity escalation logic, as formalized in the Decision Framework Flowchart (Fig. 4). The core objective of this layered architecture is to systematically resolve tactical logistics problems using the lowest necessary level of computational requirements, preventing unnecessary operational latency during a crisis.

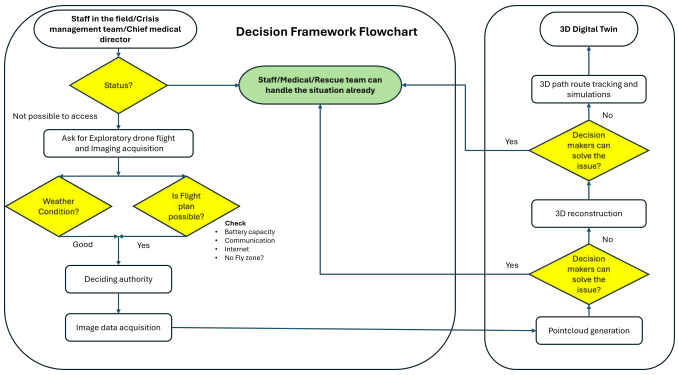


Fig. 4. Proposed Decision Framework Flowchart illustrating the complete data routing from image capture to real-time interactive risk visualization and simulation adjustment loop.

The first operational tier relies solely on the static 3D reconstruction and high-resolution elevation modeling. If local civil defense teams and/or medical directors face low-intensity or static threats, situational awareness can be achieved entirely through geographic inspection of the static asset layout. At this baseline stage, preliminary deployment zones, field hospitals, and resource positioning can be validated without triggering active runtime simulations.

When static terrain information is insufficient to predict structural risk, the framework escalates to the second operational tier: the Unreal Engine 5 water body flooding simulation. By mapping water body vectors over the static colliders, the framework quantifies upcoming infrastructure vulnerabilities before physical immersion occurs, pinpointing specific logistical threats such as the 35% highway disruption and the 38 affected buildings shown in Table II.

To make explicit which capabilities were implemented and evaluated versus conceptually proposed, Table III summarizes each simulation feature against its input data and output.

The third operational tier executes dynamic path tracking and safe corridor mapping, as visualized by the non-affected



Fig. 5. Flood progression simulation map through four different scenarios: Dry, Strong Rain, River Flood, Heavy Flood



Fig. 6. Flood progression simulation map, highlighted by the non-affected area (green), through four different scenarios: Dry, Strong Rain, River Flood, Heavy Flood.

TABLE III
SIMULATION FEATURES: INPUT, OUTPUT, AND VALIDATION STATUS

Feature	Status	Input data	Output
Static 3D reconstruction & elevation model	**Demonstrated**	UAV RGB imagery (719 aligned photographs) → SfM/MVS mesh	Geo-referenced, textured 3D mesh; static physics colliders
Flood progression simulation (4 scenarios)	**Demonstrated**	Static mesh colliders + parameterized water-body volume settings	Per-scenario % area affected, buildings inundated, % highway disruption
Safe-corridor / non-affected route overlay (Fig. 6)	**Demonstrated**	Flood-affected area boundaries from simulation	Visual overlay of non-affected routes (qualitative)
Tiered decision framework (Fig. 4)	**Proposed (conceptual)**	Outputs of the tiers above	Escalation logic across inspection, simulation, and routing tiers — not yet implemented as a working decision-support tool
Autonomous path-tracking / rescue routing	**Proposed (future work)**	Safe-corridor data (once automated)	Automated vehicle routing directives — not implemented in this work

overland vectors (green zones) in Fig. 6. By computing mathematical isolation boundaries over the active water body limits, the framework extracts safe supply routes and emergency evacuation corridors. This integrated approach solves critical logistical bottlenecks directly from the digital twin, toward the goal of translating rapid-acquisition UAV surveys into actionable routing directives for emergency vehicles, pending implementation and field validation of the automated routing component (Table III).

V. CONCLUSION

This paper presented a photogrammetric and simulation framework to translate rapid-acquisition UAV datasets directly into interactive digital twins within Unreal Engine 5. By processing a 733-frame aerial survey over a 0.408 km^2 flood-prone corridor of the Atuba River, 719 images were successfully aligned within a total pipeline computational footprint of 31 minutes and 29 seconds. The results demonstrate that lightweight aerial workflows can efficiently initialize a real-time-capable 3D digital twin environment. This work is best characterized as a rapid digital-twin initialization case study; whether it translates into operational disaster-response performance has not yet been tested and requires validation with end users or live emergency data.

Quantitative validation of the environment revealed a mean SSIM of 0.3285, a PSNR of 16.0247 dB, and a tightly bounded perceptual distance (LPIPS $= 0.3620$). While classical pixel-level indices exhibit degradation over uniform textures like canopies and minor buildings, a multi-stage mesh optimization and decimation strategy to a 5-million triangle topology appear to preserve macro-geometric boundary definitions better than pixel-level detail, based on the LPIPS results. Implemented as static physical colliders, the mesh accurately handled real-time water body vectors to simulate a four-stage flood progression. As demonstrated along the simulations, $38/40$ buildings were affected due to heavy flood, with obstruction of 35% of the highway.

Finally, this architecture establishes the potential of 3D digital twins to expand situational monitoring and decision support in complex disaster scenarios and public health emergencies. Future work will focus on integrating autonomous path-tracking algorithms and active rescue routing protocols directly into the virtual environment to support emergency field operations. Moreover, the proof-of-concept decision framework can be used as a tactical feasibility from each stage of the digital twin reconstruction. This study also has two acknowledged limitations: the exact number and sampling ratio of image pairs contributing to the SSIM/PSNR/LPIPS averages could not be fully reconstructed from archived processing data, and the proposed decision-support and autonomous-routing components (Table III) remain conceptual and untested with real users or live emergency data.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Improvement of Higher Education Personnel (CAPES) and the German Academic Exchange Service (DAAD), under the CAPES/DAAD - PROBRAL program (Grant No. 88881.986264/2024-01) for the financial support provided through the Probral program. Special thanks are extended to the RWTH Uniklinik and the Pontifical Catholic University of Paraná (PUCPR) for institutional support for providing the high-resolution datasets of the study area, which were fundamental to the development and validation of the 3D Digital Twin.

REFERENCES

[1] N. Mohammadi and J. E. Taylor, "Smart city digital twins," in *IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–5. [Online]. Available: http://smartcitydigitaltwins.gatech.edu/

[2] Q. Min, Y. Lu, Z. Liu, C. Su, and B. Wang, "Machine learning based digital twin framework for production optimization in petrochemical industry," *International Journal of Information Management*, vol. 49, pp. 502–519, 2019.

[3] C. Fan, C. Zhang, A. Yahja, and A. Mostafavi, "Disaster city digital twin: A vision for integrating artificial and human intelligence for disaster management," *International Journal of Information Management*, vol. 56, p. 102049, 2021.

[4] A. Gros, A. Guillem, L. De Luca, É. Baillieul, B. Duvocelle, O. Malavergne, L. Leroux, and T. Zimmer, "Faceting the post-disaster built heritage reconstruction process within the digital twin framework for notre-dame de paris," *Scientific Reports*, vol. 13, no. 1, p. 5981, 2023.

[5] U. Lagap and S. Ghaffarian, "Digital post-disaster risk management twinning: A review and improved conceptual framework," *International Journal of Disaster Risk Reduction*, p. 104629, 2024.

[6] C. Ge and S. Qin, "Urban flooding digital twin system framework," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2460432, 2025.

[7] A. Ene, A. C. Badea, G. Badea, and A.-P. Grădinaru, "Development of urban digital twins using gis and game engine systems," *Land*, vol. 15, no. 2, 2026. [Online]. Available: https://www.mdpi.com/2073-445X/15/2/254

[8] P. R. Martins-Filho *et al.*, "Catastrophic floods in rio grande do sul, brazil: The need for public health responses to potential infectious disease outbreaks," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 57, pp. e00 603–2024, 2024.

[9] S. Fischer and A. H. Schumann, "Temporal changes in the frequency of flood types and their impact on flood statistics," *Journal of Hydrology X*, vol. 22, p. 100171, 2024.

[10] J. Zhang, Y. Du, Y. Jiang, S. Zhang, H. Chen, and D. Shang, "Exploring the application of NeRF in enhancing post-disaster response: A case study of the sasebo landslide in japan," *ISPRS International Journal of Geo-Information*, vol. 14, no. 6, p. 218, 2025.

[11] H. O. Mete and Z. B. Zabinsky, "Stochastic optimization of medical supply location and distribution in disaster management," *International Journal of Production Economics*, vol. 126, no. 1, pp. 76–84, 2010.

[12] W. Yin, Q. Hu, W. Liu, J. Liu, P. He, D. Zhu, and A. Kornejady, "Harnessing game engines and digital twins: Advancing flood education, data visualization, and interactive monitoring for enhanced hydrological understanding," *Water*, vol. 16, no. 17, p. 2528, 2024.

[13] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[14] F. Remondino and S. El-Hakim, "Image-based 3d modeling: A review," *The Photogrammetric Record*, vol. 21, no. 115, pp. 269–291, 2006.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

Genetic Albatross: Hyperparameter Optimization in Classification Models via Genetic Algorithms

Walter William Cavalcante de Castro, Emanuel da Silva Oliveira, Lourival Gerardo da Silva Júnior,
Thiago Guilherme Bezerra de Aguiar, Renan Teixeira Leite, Ana Beatriz Silva Vasconcelos, Joaquim Celestino Júnior
Grupo de Redes e Sistemas Inteligentes (INETSYS)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Estadual do Ceará (UECE) – Fortaleza, CE – Brasil
Email: thiago.aguiar@aluno.uece.br, renan.teixeira@aluno.uece.br,
anab.vasconcelos@aluno.uece.br, walter.cavalcante@aluno.uece.br, joaquim.celestino@uece.br

***Abstract*—This work presents Genetic Albatross, a Python application for automatic hyperparameter optimization in supervised classification models using genetic algorithms. The system integrates multiple machine learning algorithms and an interactive graphical interface for intuitive experiment configuration. The solution includes data preprocessing, dataset splitting, optional balancing, and evaluation using multiple metrics. Experimental results show that the system can find effective hyperparameter configurations with fast convergence.**

***Index Terms*—Aprendizado de Máquina, Algoritmos Genéticos, Otimização de Hiperparâmetros, AutoML.**

I. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and, more specifically, Machine Learning (ML) has enabled the development of intelligent systems capable of supporting decision-making processes in a wide variety of domains, including healthcare, industry, finance, agriculture, and education [1], [2].Instead of being explicitly programmed to solve every possible situation, machine learning algorithms learn patterns directly from historical data, allowing them to make predictions or classifications when new data are presented.

The quality of these predictions depends not only on the amount of available data but also on how effectively the learning algorithm extracts useful information from them. Consequently, developing accurate and reliable machine learning models requires more than simply selecting a classification algorithm. Data quality, appropriate preprocessing, class balancing strategies, and the configuration of the learning algorithm all play fundamental roles in determining the final predictive performance of the model [3], [4].

One of the most important aspects of this process is the definition of hyperparameters. Hyperparameters are configuration values specified before the training process begins and determine how a machine learning algorithm learns from the available data. Unlike model parameters, which are automatically learned during training, hyperparameters must be selected beforehand by the user or by an optimization algorithm. Examples include the maximum depth of a Decision Tree, the number of neighbors considered by the K-Nearest Neighbors (KNN) algorithm, the learning rate of boosting algorithms, and the number of neurons in an Artificial Neural Network. Selecting inappropriate hyperparameter values may produce models that are unable to capture important data patterns (underfitting) or models that become excessively specialized to the training data, reducing their ability to generalize to unseen examples (overfitting) [3].

As machine learning algorithms become increasingly sophisticated, the number of possible hyperparameter combinations grows rapidly. This set of possible combinations is commonly referred to as the search space. Exploring every possible configuration through exhaustive search is often computationally expensive or even infeasible, especially for complex models containing numerous hyperparameters. For this reason, automatic optimization methods have become an essential component of modern machine learning workflows [4].

Among the available optimization approaches, bio-inspired metaheuristics have attracted considerable attention. Metaheuristics are optimization techniques designed to efficiently search for high-quality solutions without evaluating every possible alternative. Inspired by natural phenomena such as biological evolution and collective animal behavior, these methods balance the exploration of new regions of the search space with the exploitation of promising solutions that have already been discovered [4].

Two of the most widely adopted bio-inspired optimization methods are the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) [5], [6]. The Genetic Algorithm simulates the principles of natural evolution, where candidate solutions evolve through selection, crossover, and mutation operations. In contrast, Particle Swarm Optimization models the collective behavior of groups such as bird flocks or fish schools, where each candidate solution continuously adjusts its position according to both its own previous experience and the best solution found by the swarm. Although based on different optimization strategies, both methods have demonstrated excellent performance in solving complex optimization problems, including machine learning hyperparameter tuning.

The use of optimization algorithms enables the automatic discovery of hyperparameter configurations capable of im-

proving model performance while significantly reducing the manual effort traditionally required during experimentation. Consequently, evolutionary optimization techniques have become an important component of Automated Machine Learning (AutoML), a research area dedicated to automating tasks such as algorithm selection, hyperparameter optimization, and model evaluation.

Despite the remarkable advances achieved by existing AutoML platforms, many available solutions still present limitations regarding flexibility, transparency of the optimization process, or ease of use for users with limited programming experience. In many cases, understanding or modifying the optimization strategy employed by these platforms is difficult, which restricts their applicability in educational and experimental environments.

To address these limitations, this work presents Genetic App, a desktop application designed to automate machine learning experiments by integrating classification algorithms, class balancing techniques, and evolutionary optimization methods for hyperparameter tuning. The application provides an intuitive graphical user interface that allows users to configure and execute experiments without requiring advanced programming knowledge while maintaining flexibility for experimentation.

The application currently supports the following classification algorithms: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Logistic Regression, Multi-Layer Perceptron (MLP), XGBoost, LightGBM, and CatBoost. Hyperparameter optimization can be performed using either the Genetic Algorithm or Particle Swarm Optimization, allowing users to compare the behavior of both optimization techniques under different experimental scenarios. In addition, the platform incorporates class balancing through UnderSampling and SMOTE, improving the handling of imbalanced datasets.

The main contributions of this work are summarized as follows:

- Development of a graphical desktop application for automatic hyperparameter optimization.
- Integration of multiple classification algorithms, data balancing strategies, and evolutionary optimization methods into a unified platform.
- Support for both Genetic Algorithm and Particle Swarm Optimization for automated hyperparameter tuning.
- Automatic generation of evaluation metrics and feature importance analysis to improve model interpretability.
- A modular and extensible software architecture that facilitates the incorporation of additional machine learning algorithms and optimization techniques. The remainder of this paper is organized as follows. Section II presents the theoretical background required to understand the proposed approach. Section III reviews the most relevant related work. Section IV describes the proposed model and system architecture. Section V details the implementation of the Genetic App. Section VI discusses the experimental results obtained during validation, and Section VII concludes the paper while presenting directions for future research.

II. THEORETICAL BACKGROUND

A. Hyperparameter optimization

The development of an accurate and reliable machine learning model depends not only on the amount and quality of the available data but also on the configuration adopted during the learning process. Before a model can learn from the training data, several configuration values must be specified. These values are known as hyperparameters and determine how the learning algorithm will operate during training [3]. It is important to distinguish hyperparameters from model parameters. Model parameters are automatically learned from the training data during the learning process. For example, in a Logistic Regression model, the coefficients associated with each feature are learned during training. Hyperparameters, on the other hand, are defined before training begins and control the behavior of the learning algorithm itself. Examples include the maximum depth of a Decision Tree, the number of trees in a Random Forest, the number of neighbors considered by the K-Nearest Neighbors (KNN) algorithm, and the learning rate employed by gradient boosting methods. To address this challenge, hyperparameter optimization techniques automatically search the space of possible configurations to identify those that maximize model performance according to a predefined objective function. The objective function, often called the fitness function in evolutionary optimization, measures the quality of each candidate solution using one or more evaluation metrics. By automating this search process, optimization algorithms reduce human effort while increasing the likelihood of obtaining highly accurate predictive models[4].

B. Genetic Algorithms

Genetic Algorithms (GA), originally proposed by Holland [5], are among the most established optimization techniques in the field of Evolutionary Computation. Their operation is inspired by the principles of natural selection and biological evolution, where populations evolve over successive generations and individuals with desirable characteristics are more likely to survive and reproduce. In the context of hyperparameter optimization, each candidate solution corresponds to a specific combination of hyperparameter values and is represented by a chromosome. Each hyperparameter is encoded as a gene, meaning that a chromosome completely describes one possible configuration of the machine learning model. By repeatedly applying selection, crossover, mutation, and fitness evaluation over multiple generations, the population gradually evolves toward increasingly effective hyperparameter configurations. The combination of exploration through mutation and exploitation through the preservation of high-quality individuals makes Genetic Algorithms particularly suitable for complex optimization problems characterized by large and irregular search spaces [5].

C. Particule Swarm Optimization

Particle Swarm Optimization (PSO), introduced by Kennedy and Eberhart [6], is another widely adopted bio-inspired optimization technique. Unlike Genetic Algorithms, which

simulate biological evolution, PSO is inspired by the collective behavior observed in groups of animals such as bird flocks and fish schools. In PSO, each candidate solution is represented by a particle positioned within the search space. The position of a particle corresponds to one possible hyperparameter configuration, while its movement through the search space represents the continuous search for improved solutions. Compared with Genetic Algorithms, PSO does not require crossover or mutation operators. Instead, optimization emerges naturally from the continuous movement of particles toward increasingly promising regions of the search space. Because of its conceptual simplicity, relatively small number of control parameters, and computational efficiency, PSO has become one of the most frequently employed optimization techniques for machine learning hyperparameter tuning [6].

D. Class Balancing

In many real-world applications, datasets exhibit imbalanced class distributions. Under such conditions, classification algorithms tend to favor the majority class, resulting in degraded predictive performance for the minority class.

Two of the most commonly used techniques to mitigate this problem are UnderSampling and SMOTE (Synthetic Minority Over-sampling Technique) [7].

UnderSampling reduces the number of instances belonging to the majority class. In contrast, SMOTE artificially generates new minority-class examples by interpolating between existing instances, thereby promoting a more balanced data distribution.

E. Model Evaluation and interpretability

Proper model evaluation is essential for assessing a model's generalization capability. In this work, the evaluation metrics Accuracy, Precision, and Recall are employed, as they are widely used in supervised classification problems [11].

Beyond quantitative evaluation, model interpretability has become an increasingly important aspect of modern machine learning applications. Feature importance analysis techniques make it possible to identify which variables have the greatest influence on the predictions generated by machine learning models [12].

Whenever available, native feature importance measures provided by the classification algorithms are used. Otherwise, the Permutation Importance technique is employed, allowing a consistent and model-independent evaluation of feature contributions to predictive performance [8], [12].

III. Related Work

Hyperparameter optimization has become one of the most active research topics in machine learning because of its significant influence on predictive model performance. Since the effectiveness of a machine learning algorithm depends heavily on the selected hyperparameter configuration, numerous optimization techniques have been proposed to automate this process and reduce the need for manual experimentation [3], [4].

Traditional optimization strategies such as Grid Search and Random Search have been widely employed due to their simplicity of implementation. Grid Search systematically evaluates all possible combinations within predefined ranges, whereas Random Search randomly samples candidate configurations from the search space.Although both approaches are straightforward, they become computationally expensive as the number of hyperparameters increases, a phenomenon commonly referred to as the curse of dimensionality. Consequently, these methods often become impractical for complex machine learning models containing numerous hyperparameters or large search spaces [4].

To overcome these limitations, researchers have increasingly adopted metaheuristic optimization algorithms, which aim to identify high-quality solutions without exhaustively evaluating every possible configuration. Among the various metaheuristics proposed in the literature, Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) have consistently demonstrated competitive performance across a wide variety of optimization problems [4].

Genetic Algorithms have been extensively investigated for hyperparameter optimization because of their ability to explore large and irregular search spaces through evolutionary operators such as selection, crossover, and mutation. By continuously evolving a population of candidate solutions, GA can effectively identify high-quality hyperparameter configurations even for highly nonlinear optimization problems [5].

Similarly, Particle Swarm Optimization has become one of the most widely adopted optimization techniques in machine learning. Instead of relying on evolutionary operators, PSO performs optimization through the cooperative movement of particles that exchange information while searching the solution space. Owing to its relatively simple implementation, reduced number of control parameters, and fast convergence characteristics, PSO has demonstrated excellent performance in numerous hyperparameter optimization applications [6].

Comparative studies indicate that both GA and PSO are capable of significantly improving the predictive performance of machine learning models when compared with manually selected hyperparameters or exhaustive search techniques. While Genetic Algorithms generally provide greater search diversity through crossover and mutation operations, Particle Swarm Optimization frequently converges more rapidly toward high-quality solutions by exploiting the collective experience accumulated by the swarm [4], [6]. Because each method exhibits distinct optimization characteristics, both remain widely employed in contemporary machine learning research.

In parallel with advances in optimization algorithms, the field of Automated Machine Learning (AutoML) has emerged to simplify the development of predictive models. AutoML aims to automate tasks such as data preprocessing, algorithm selection, hyperparameter optimization, and model evaluation,

Table I
COMPARISON OF PROPOSED "GENETIC APP" WITH EXISTING HYPERPARAMETER OPTIMIZATION LITERATURE

Reference	Optimization Method	ML Algorithms	Class Balancing	GUI	Feature Importar
Generic Grid Search studies	Grid Search	Multiple	× No	× No	× No
Generic Random Search studies	Random Search	Multiple	× No	× No	× No
GA-based optimization studies	Genetic Algorithm (GA)	Single / Few	× No	× No	Limited
PSO-based optimization studies	Particle Swarm Optimization (PSO)	Single / Few	× No	× No	Limited
Proposed Genetic App	**GA + PSO**	**KNN, Decision Tree, Random Forest, Logistic Regression, MLP, XGBoost, LightGBM, CatBoost**	**UnderSampling + SMOTE**	**✓ Yes**	**Native + Permutat Importance**

enabling users to build high-quality machine learning models with reduced manual intervention.

Among the best-known AutoML platforms are Auto-WEKA [9] and Auto-Sklearn [10]. Auto-WEKA extends the WEKA machine learning framework by automatically selecting classification algorithms and optimizing their hyperparameters. Auto-Sklearn, built upon the Scikit-learn ecosystem, combines Bayesian optimization, meta-learning, and ensemble techniques to automatically identify effective machine learning pipelines. Both platforms have demonstrated strong predictive performance across numerous benchmark datasets and have become important references in the AutoML literature.

Despite their effectiveness, these platforms present certain limitations from the perspective of experimentation and educational use. The optimization process is often highly automated, providing users with limited visibility into the internal behavior of the optimization algorithms. Furthermore, extending these platforms with alternative optimization strategies or integrating custom experimental workflows may require considerable programming expertise.

Another research topic closely associated with predictive model performance concerns class imbalance. Numerous studies have demonstrated that highly imbalanced datasets may considerably degrade classifier performance, particularly for minority classes that often correspond to the events of greatest practical interest. To mitigate this problem, balancing techniques such as UnderSampling and the Synthetic Minority Over-sampling Technique (SMOTE) have been extensively investigated and successfully applied in various application domains [7].

More recently, model interpretability has also become an important research direction within machine learning. As predictive models become increasingly complex, understanding how they reach their decisions has become essential in domains such as healthcare, finance, and industrial applications. Feature importance analysis and model-agnostic interpretation techniques have therefore gained considerable attention, providing valuable insights into the contribution of individual features to the prediction process [8], [12].

Unlike many existing solutions, Genetic App integrates multiple stages of the machine learning workflow into a single desktop application. The platform combines an intuitive graphical user interface with several classification algorithms, data balancing techniques, Genetic Algorithm and Particle Swarm Optimization for hyperparameter tuning, automatic performance evaluation, and feature importance analysis. Rather than focusing exclusively on optimization or algorithm selection, the proposed application provides a unified and extensible environment that facilitates experimentation, comparative analysis, and educational activities while requiring minimal programming knowledge from the user.

IV. PROPOSED MODEL

Genetic App was developed to automate the cycle of building, optimizing, and evaluating machine learning models using tabular data from different application domains, including medical spreadsheets and structured clinical datasets.

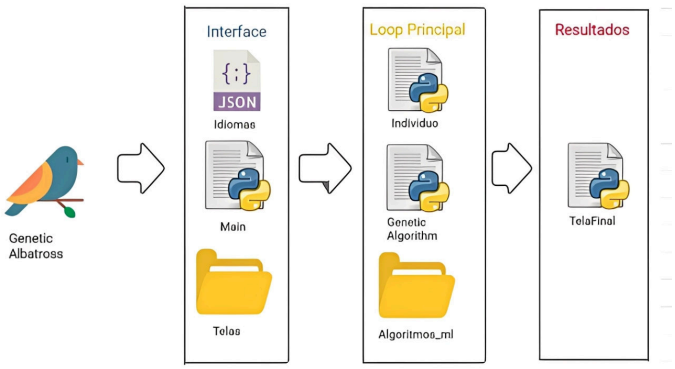


Figure 1. General Architecture of Genetic App

The execution workflow begins with dataset importation and proceeds through data preprocessing, classifier configuration, selection of the data balancing strategy, and definition of the optimization algorithm.

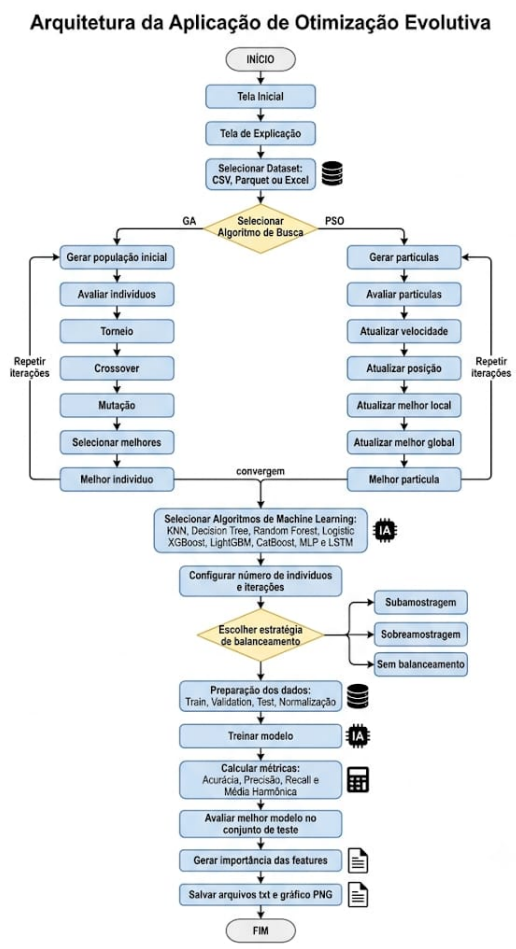


Figure 2. fluxogram of app

The system architecture is composed of five main modules:
- Graphical User Interface (GUI);
- Data Preprocessing;
- Classification Algorithms;
- Optimization Algorithms;
- Evaluation and Report Generation.

A. Graphical User Interface

The Graphical User Interface (GUI) provides the primary interaction between the user and the application. It was designed to simplify the configuration of machine learning experiments by offering an intuitive interface that eliminates the need for programming knowledge. Through the interface, users can import datasets, select the target variable, choose one of the available classification algorithms, configure the optimization method, define optimization parameters, enable optional data balancing techniques, and execute the complete experiment. During execution, the interface also provides feedback regarding the current processing stage, allowing users to monitor the progress of the optimization procedure. Because all experimental parameters are configured through graphical components, the application can be used not only by machine learning researchers but also by students and professionals with limited programming experience.

B. Data Preprocessing Module

Before model training begins, the imported dataset is processed by the preprocessing module. This stage prepares the data to ensure compatibility with the selected classification algorithm and improve the quality of the learning process. Initially, the dataset is divided into input features and the target variable. Subsequently, numerical features may be normalized or transformed when required by the selected classifier. Since several machine learning algorithms are sensitive to differences in feature scales, appropriate preprocessing contributes to more stable optimization and improved predictive performance. If requested by the user, the preprocessing module also applies one of the available class balancing techniques. UnderSampling reduces the number of majority-class samples, whereas SMOTE generates synthetic instances of the minority class to obtain a more balanced training dataset.

C. Classification Module

The classification module is responsible for training and evaluating predictive models using the hyperparameter configurations generated during the optimization process. The current version of the Genetic App supports eight supervised classification algorithms: - K-Nearest Neighbors (KNN); - Decision Tree; - Random Forest; - Logistic Regression; - Multi-Layer Perceptron (MLP); - XGBoost; - LightGBM; - CatBoost. Each classifier possesses its own set of hyperparameters and learning characteristics. Consequently, the optimization algorithms explore different search spaces depending on the selected classifier. This modular design facilitates the incorporation of additional machine learning algorithms in future versions of the application without requiring significant modifications to the remaining system components.

D. Optimization Module

The optimization module constitutes the core of the proposed application. Its primary responsibility is to automatically identify hyperparameter configurations capable of maximizing the predictive performance of the selected classification algorithm. The module supports both Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), allowing users to compare the behavior of these two evolutionary optimization techniques under identical experimental conditions. When the Genetic Algorithm is selected, an initial population of candidate solutions is generated. Each individual represents a complete hyperparameter configuration. After evaluating the predictive performance of each individual, the population evolves through successive generations using selection, crossover, and mutation operators until a termination criterion is satisfied. Alternatively, when Particle Swarm Optimization is selected, each candidate solution is represented by a particle moving through the hyperparameter search space. During each iteration, particles update their positions by considering both their own best solution (pBest) and the best solution discovered by the swarm (gBest). This cooperative search mechanism enables the swarm to progressively converge toward high-quality

hyperparameter configurations. Regardless of the selected optimization technique, every candidate solution is evaluated by training the corresponding machine learning model and computing its fitness value.

E. Evaluation and Report Generation

After the optimization process is completed, the application automatically generates a comprehensive report describing the experiment. The report includes the optimized hyperparameter configuration, evaluation metrics, execution time, and feature importance analysis. Whenever the selected classifier provides native feature importance scores, these values are presented directly. Otherwise, Permutation Importance is employed to estimate the contribution of each feature. The fitness function adopted during optimization simultaneously considers Accuracy, Precision, and Recall, encouraging the selection of hyperparameter configurations that achieve balanced predictive performance rather than optimizing only a single evaluation metric.

V. Implementation

The Genetic App was implemented in Python due to its widespread adoption within the Data Science community and the extensive ecosystem of specialized libraries available for machine learning applications.

The system architecture was designed in a modular manner, ensuring a clear separation between the graphical user interface, data processing, classification algorithms, and optimization mechanisms.

The graphical user interface was developed using the Tkinter library, providing an intuitive environment for configuring and executing machine learning experiments. Through the interface, users can import datasets, select classification algorithms, configure optimization methods, and choose data balancing strategies.

Machine learning functionalities were implemented using the Scikit-learn library, which provides support for data preprocessing, model training, and the computation of evaluation metrics.

In addition, specialized libraries for gradient boosting algorithms, including XGBoost, LightGBM, and CatBoost, were integrated into the application, expanding the diversity of classification models available for analysis.

The optimization module was designed to support both Genetic Algorithm and Particle Swarm Optimization, allowing the automatic exploration of different hyperparameter configurations.

At the end of each experiment, the application automatically generates reports containing the selected hyperparameters, evaluation metrics, execution time, and feature importance analysis.

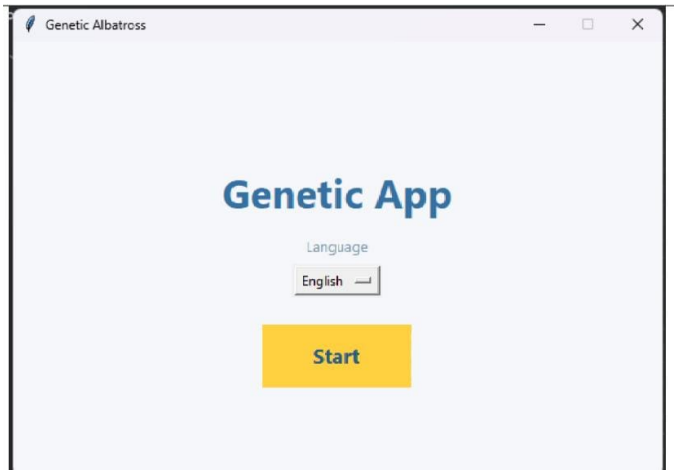


Figure 3. Main Interface of Genetic App

VI. Result and Analysis

The proposed application was experimentally evaluated to investigate its ability to automate the hyperparameter optimization process while improving the predictive performance of machine learning models. The experiments considered different classification algorithms, evolutionary optimization techniques, and class balancing strategies, allowing a comprehensive analysis of the proposed framework. Each experiment followed the workflow implemented in the Genetic App. Initially, the selected dataset was preprocessed and, when required, balanced using either UnderSampling or SMOTE. Subsequently, the chosen classification algorithm was optimized through either the Genetic Algorithm (GA) or Particle Swarm Optimization (PSO). During the optimization process, each candidate solution was evaluated according to the fitness function, which measured the predictive performance obtained after model training. Once the optimization procedure was completed, the best hyperparameter configuration was selected and used to train the final model. Model performance was assessed using Accuracy, Precision, and Recall, providing complementary information about the quality of the generated classifiers. In addition, whenever supported by the selected algorithm, feature importance analysis was performed to improve model interpretability. For algorithms without native feature importance measures, Permutation Importance was employed. The experimental results demonstrate that automatic hyperparameter optimization consistently identified configurations capable of improving predictive performance compared with default parameter settings. Both optimization algorithms successfully explored the search space and converged toward competitive solutions, confirming their suitability for machine learning hyperparameter optimization. Although both optimization techniques produced satisfactory results, some differences were observed in their search behavior. The Genetic Algorithm generally explored a broader portion of the search space due to the combined effects of crossover and mutation, making it effective for avoiding premature convergence. In contrast, Particle Swarm Optimization often converged more

rapidly by continuously adjusting candidate solutions according to both individual and collective experience. Depending on the characteristics of the evaluated dataset and the selected classifier, either optimization method could achieve the best overall performance. The application also demonstrated the importance of class balancing for imbalanced datasets. When balancing techniques were applied before model training, improvements were observed mainly in Precision and Recall, indicating that the optimized models became more capable of correctly identifying minority-class instances. While UnderSampling reduced the number of majority-class samples and consequently shortened training time, SMOTE preserved the original data while generating synthetic minority samples, frequently leading to superior predictive performance. Feature importance analysis provided additional insight into the prediction process by identifying the variables that contributed most strongly to the classification decisions. This information complements the quantitative evaluation metrics and increases the transparency of the generated models, allowing users to better understand the influence of individual features on the final predictions. Overall, the experimental evaluation demonstrates that the Genetic App successfully integrates machine learning algorithms, class balancing techniques, and evolutionary optimization methods into a unified framework capable of automating hyperparameter optimization. The obtained results indicate that the proposed application provides an efficient and flexible environment for conducting machine learning experiments while reducing the manual effort traditionally associated with hyperparameter selection.

A. Feature Importance Analysis

Feature importance analysis proved to be a valuable resource for interpreting model behavior.

After the optimization process is completed, the application automatically generates rankings of the most influential features for each trained classification model.

Whenever available, the native feature importance measures provided by the classification algorithms are used.

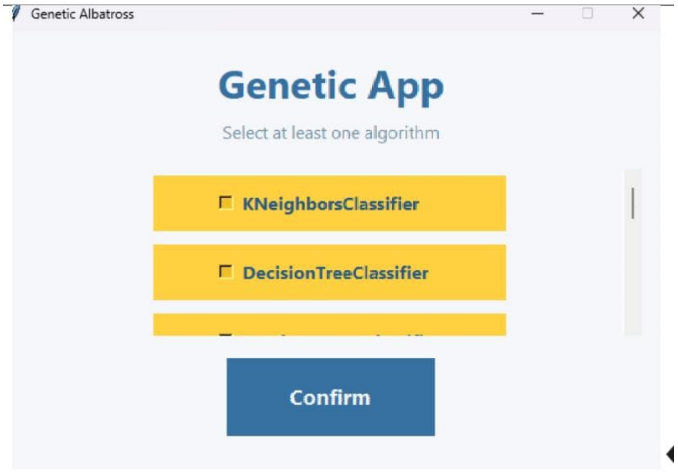


Figure 4. Example of Feature Ranking

The experimental results indicate that the combination of evolutionary optimization, automated preprocessing, data balancing, and feature importance analysis provides an efficient and practical environment for machine learning experimentation.

VII. CONCLUSION

This work presented Genetic App, a desktop application developed to automate hyperparameter optimization for machine learning algorithms.

The platform integrates traditional and state-of-the-art classification algorithms, data balancing techniques, evolutionary optimization methods, and model interpretation mechanisms into a unified environment.

The experimental results demonstrated the feasibility of employing Genetic Algorithm and Particle Swarm Optimization as effective approaches for automatic hyperparameter tuning.

Furthermore, the inclusion of class balancing techniques and feature importance analysis expands the applicability of the platform across different domains and machine learning scenarios.

As future work, new metaheuristic optimization algorithms will be incorporated, the set of supported classification algorithms will be expanded, and advanced AutoML functionalities will be integrated, further increasing the flexibility and analytical capabilities of the platform.

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] L. Franceschi et al., "Hyperparameter Optimization in Machine Learning," *Foundations and Trends in Machine Learning*, vol. 18, no. 6, pp. 1054–1201, 2025.

[4] A. Ali, J. Zhu and M. Zakarya, "A Survey on Hyperparameter Optimization of Machine Learning Models," *IEEE Access*, 2024.

[5] J. H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[6] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, 1995, pp. 1942–1948.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall and K. W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] C. Thornton, F. Hutter, H. H. Hoos and K. Leyton-Brown, "Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms," in *Proceedings of ACM SIGKDD*, 2013, pp. 847–855.

[10] M. Feurer et al., "Efficient and Robust Automated Machine Learning," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[11] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[12] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. Available: https://christophm.github.io/interpretable-ml-book/

Geometric Model of a Digit for Pre-Treatment in Veterinary Electrochemotherapy

1st Marina C. P. da Luz
PPGEEL-UFSC
UFSC
Florianópolis, Brasil
marina.luz@posgrad.ufsc.br

2nd Marco A. M. Paschoal
CAC-UFSC Blumenau
UFSC
Blumenau, Brasil
marco.paschoal@grad.ufsc.br

3rd Daniela O. H. Suzuki
PPGEEL-UFSC
UFSC
Florianópolis, Brasil
suzuki@eel.ufsc.br

4th Guilherme B. Pintarelli
PPGEEL-UFSC
UFSC
Florianópolis, Brasil
guilherme.pintarelli@ufsc.br

***Abstract*—Electrochemotherapy (ECT) is a technique used for the treatment of tumors that combines the administration of chemotherapeutic agents with the application of high-intensity electric pulses, promoting the electroporation of cell membranes and enhancing the intracellular uptake of drugs. The effectiveness of the treatment depends directly on the distribution of the electric field within the tissue, which is influenced by anatomical geometry, the electrical properties of the tissues, and the positioning of the electrodes. In this study, a simplified computational model of the canine middle phalanx was developed to simulate the electric field distribution during electrochemotherapy procedures. The model was implemented in COMSOL Multiphysics® using the Electric Currents module, with electrical properties assigned to the different tissues based on data reported in the literature. The results made it possible to identify regions subjected to reversible and irreversible electroporation, as well as areas with insufficient electric field coverage, referred to as blind spots. The developed model proved capable of representing the spatial distribution of the electric field and constitutes a promising tool for studies aimed at optimizing electrode positioning and the therapeutic parameters employed in electrochemotherapy.**

***Key words*—electroporation, COMSOL, finite element method, electric field.**

I. Introduction

Electrochemotherapy (ECT) is an oncological treatment technique that combines the use of chemotherapeutic agents with specialized electrodes to deliver high-intensity electric pulses capable of inducing electroporation in cell membranes, thereby temporarily increasing cell membrane permeability and, consequently, enhancing drug uptake [1]. Electroporation is a biophysical phenomenon characterized by the temporary formation of pores in the cell membrane when exposed to electric pulses, facilitating the transport of molecules into the cell [1], [2].

A key factor for the success of electrochemotherapy is the distribution of the electric field within the target region. For the treatment to be effective, the entire tumor volume must be exposed to adequate electric field levels, while the exposure of healthy tissues should be minimized [3]. Therefore, treatment outcomes are influenced by factors such as patient anatomy, the electrical properties of biological tissues, and electrode coupling and positioning [3], [4].

To improve treatment predictability and safety, computational models based on the Finite Element Method (FEM) have been widely used as planning and pre-treatment tools, allowing the safe evaluation of the effects of electric field distribution and electrode positioning in electrochemotherapy procedures [3], [5]. The use of geometries reconstructed from medical images enables the development of models that more accurately represent patient anatomy, allowing different therapeutic configurations to be evaluated before clinical application [5]. Furthermore, computational modeling makes it possible to incorporate tissue-specific dielectric properties, contributing to a realistic representation of the electromagnetic phenomena involved [2].

The use of computational models to study electroporation has been widely explored in the literature as a tool to support the development of electroporation-based techniques and the optimization of devices used in electrochemotherapy. Sel *et al.* [10] introduced a sequential finite element model in which tissue electrical conductivity is iteratively updated according to the local electric field, allowing the estimation of worst-case electric field distributions and equipment current during treatment. Corović *et al.* [6] later developed a finite element model to analyze electric field distribution in biological tissues during electroporation, demonstrating the influence of tissue electrical properties and conductivity changes on the coverage of the treated region. More recently, Lopes *et al.* [4] proposed a computational methodology to optimize the electrical conductivity of conductive gels for electrochemotherapy, demonstrating that case-oriented numerical simulations can improve electric field coverage while reducing unnecessary current increase and minimizing the occurrence of blind spots. Similarly, Andrade *et al.* [3] used computational models to investigate the effects of parallel electrode deflection on electric field distribution and treatment safety, while Guedert *et al.* [2] proposed a dynamic model capable of representing changes in tissue electrical properties during electroporation by considering phenomena such as biological dispersion and Joule heating. In turn, Lopes *et al.* [5] developed a computational model based on medical imaging to evaluate the performance of a novel electrode designed for the treatment of bone cavities, demonstrating the importance of numerical modeling in the development and optimization of electrochemotherapy devices. Despite these advances, studies focused on the development of computational models specifically for canine digits remain scarce,

which motivates the proposal of the present work. Finally, this work aims to develop a parametrized computational model of a canine digit to analyze the electric field distribution during electrochemotherapy procedures. The model is designed to investigate different configurations of commercially available electrode positioning, conductive gel coupling, and anatomical variations, enabling the evaluation of their influence on electric field coverage and treatment outcomes.

II. Materials and Methods

A. Development of the Simplified Geometry

Therefore, a simplified geometry was developed while preserving the main anatomical and dimensional characteristics of the canine digit. The simplified geometry adopted in this study is shown in Figure 1.

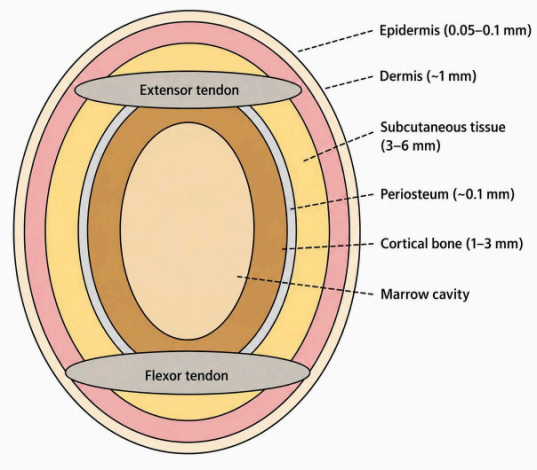


Fig. 1: Schematic cross-sectional view of the canine middle phalanx used to construct the simplified computational geometry.

Figure 1 illustrates the organization of the domains considered in the simplified geometry. The dimensions adopted for each domain were defined based on anatomical references available in the literature and parametrically adjusted to generate a geometry suitable for finite element simulations, as summarized in Table I.

TABLE I: Geometric dimensions adopted for the construction of the computational model.

Structure	Adopted Thickness
Epidermis	0.05–0.10 mm
Dermis	∼1.0 mm
Subcutaneous tissue	3–6 mm
Periosteum	∼0.10 mm
Cortical bone	1–3 mm
Medullary cavity	Variable
Extensor tendon	Dorsal elliptical structure
Flexor tendon	Ventral elliptical structure

Source: Adapted from Evans and De Lahunta (2013) and MSD Veterinary Manual (2024).

B. Computational Model

The computational model was developed focusing on the middle phalanx and was constructed using parameterized geometric solids dimensioned according to anatomical references reported in the literature [7], [8]. The final geometry included domains corresponding to the epidermis, dermis, subcutaneous tissue, periosteum, cortical bone, medullary cavity, extensor tendon, and flexor tendon. The geometry construction and numerical simulations were performed using *COMSOL Multiphysics®*.

The *Electric Currents* interface was employed to solve the electric potential distribution within the different tissues. Electrical properties were assigned individually to each domain based on values reported in the literature. Furthermore, the nonlinear dependence of electrical conductivity on the applied electric field was considered according to electroporation models.

The electrical characterization of the different domains was performed by assigning tissue-specific electrical conductivity values and electroporation thresholds reported in the literature [2], [6]. These parameters are essential for describing the nonlinear response of biological tissues subjected to electric pulses and for determining the regions undergoing reversible and irreversible electroporation. The nonlinear dependence of the electrical conductivity on the magnitude of the applied electric field was modeled using a sigmoidal function, allowing the transition from the initial conductivity to the maximum conductivity as the electric field increases. This approach reproduces the gradual increase in tissue conductivity associated with the electroporation process and has been widely adopted in computational models of electroporation [2], [6], [10]. The conductivity function is defined as:

$$\sigma(\|\mathbf{E}\|) = \sigma_{\min} + \frac{\sigma_{\max} - \sigma_{\min}}{1 + \exp\left(-\dfrac{\|\mathbf{E}\| - A}{B}\right)} \tag{1}$$

$$A = \frac{E_{\text{rev}} + E_{\text{irrev}}}{2}, \qquad B = \frac{E_{\text{irrev}} - E_{\text{rev}}}{8}. \tag{2}$$

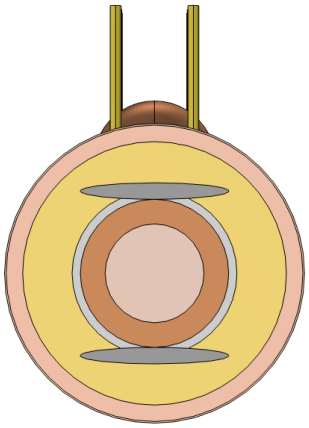
where $\sigma_{\min}$ and $\sigma_{\max}$ are the initial and maximum electrical conductivities of the tissue, respectively, $\|\mathbf{E}\|$ is the magnitude of the electric field, and E_{irrev} and E_{irrev} correspond to the lower and upper electric field thresholds that define the conductivity transition. The adopted values are presented in Table II.

It should be noted that the adopted static conductivity model is protocol-dependent. Therefore, its parameters may require recalibration when pulse characteristics, such as waveform, duration, number of pulses, or repetition rate, are modified [11]. Consequently, the results presented in this study should be interpreted within the specific pulse protocol considered in the simulations. Figure 2a shows the front view of the developed computational model, including the different tissues considered, the tumor region, and the electrode arrangement adopted in the simulations.

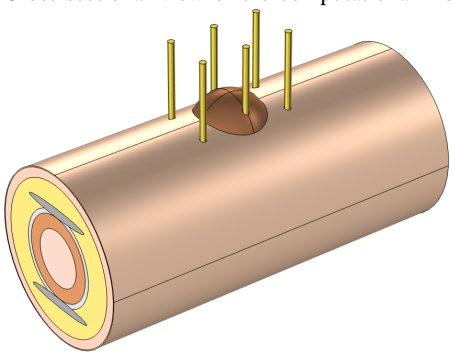
TABLE II: Electrical properties of the tissues used in the *COMSOL Multiphysics*® simulations.

Material	σ_0 **(S/m)**	$\sigma_{\max}$ **(S/m)**
Tumor	0.300	0.750
Epidermis	0.008	0.800
Dermis	0.250	1.000
Subcutaneous tissue	0.020	–
Muscle	0.135	0.340
Cortical bone	0.010	–
Electrodes (copper)	5.998×10^7	–
Material	E_{rev} **(kV/m)**	E_{irrev} **(kV/m)**
Tumor	40	80
Epidermis	40	120
Dermis	30	120
Subcutaneous tissue	20	80
Muscle	20	80
Cortical bone	–	–
Electrodes (copper)	–	–

Source: Adapted from Pintarelli et al. (2019). The original values were obtained from Corović et al. (2013) for tumor, epidermis, dermis, and muscle; Gabriel et al. (1996) for bone; and the *COMSOL Multiphysics*® material library for the electrodes.



(a) Cross-sectional view of the computational model.



(b) Orthogonal view of the computational model.

Fig. 2: Comparison between the Cross-sectional view and the orthogonal view of the computational model.

C. Electrode Configuration and Boundary Conditions

To deliver the electric pulses, six electrodes were positioned around the tumor region, as shown in Figure 2b. The electrodes were modeled as ideal conductors with a diameter of 0.7 mm, an active length of 15 mm, and a spacing of 5 mm between adjacent electrodes, resulting in three electrodes on each side of the tumor region. The three electrodes located on the left side of the model were assigned an electric potential of 600 V, whereas the three electrodes on the right side were defined as the electrical ground (0 V). The external surfaces of the model were assumed to be electrically insulated, imposing a zero normal current density across these boundaries.

D. Mesh Generation

After defining the geometry, the electrical properties of the tissues, and the boundary conditions, the computational domain was discretized by generating a three-dimensional mesh composed predominantly of tetrahedral elements, as shown in Figure 3. The mesh was automatically generated using *COMSOL Multiphysics*® and subsequently refined to ensure an adequate representation of the interfaces between the different anatomical domains.

The computational domain was discretized using an unstructured Fine tetrahedral mesh generated automatically in COMSOL Multiphysics®. The final mesh consisted of 1,485,722 tetrahedral elements and 258,965 mesh vertices, providing adequate spatial resolution for the numerical simulations.

The resulting discretization enabled the representation of the different anatomical structures of the model, providing support for the numerical solution of the electric field distribution within the analyzed tissues. Based on the developed computational model, simulations were performed to evaluate the electric field distribution and the regions subjected to electroporation under the different scenarios considered.

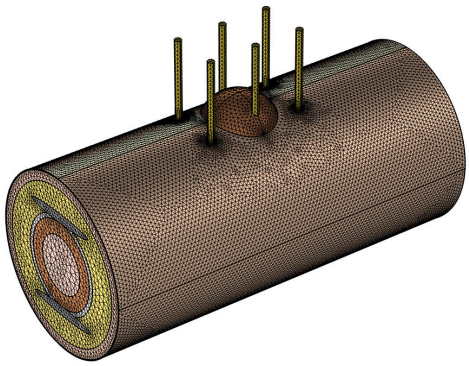


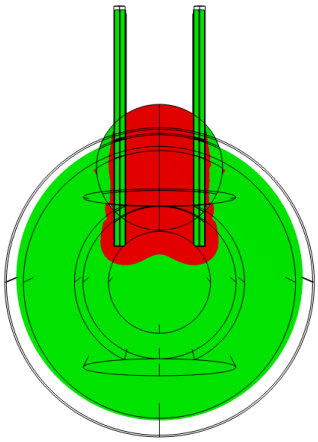
Fig. 3: Orthogonal view of the computational mesh.

III. Results and Discussion

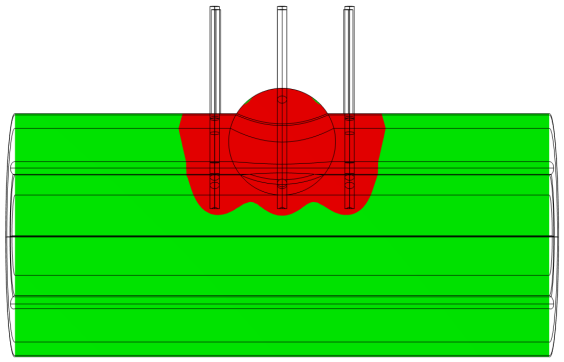
The developed computational model enabled the evaluation of the spatial distribution of the electric field within the canine digit during electrochemotherapy. Based on the adopted tissue-specific electroporation thresholds, the simulations allowed the identification of regions expected to undergo electroporation,

as well as regions where the electric field remained below the required threshold.

Figure 4 illustrates the simulated treatment coverage from frontal and lateral perspectives. In both views, the red regions correspond to locations where the electric field exceeded the adopted threshold for electroporation (approximately 60 kV/m), whereas the green regions indicate areas where the electric field remained below this threshold.



(a) Frontal view of the regions affected (red) and unaffected (green) by the applied electric field.



(b) Lateral view of the affected and unaffected regions.

Fig. 4: Spatial distribution of the regions affected by the applied electric field.

As observed in Figure 4a, the highest electric field intensities were concentrated around the electrodes, resulting in effective coverage of a substantial portion of the tumor volume. However, peripheral portions of the tumor remained outside the affected region, indicating incomplete treatment coverage. These areas represent potential *blind spots*, where the electric field intensity may be insufficient to induce the desired cell membrane permeabilization.

The lateral view shown in Figure 4b complements the frontal analysis by revealing the longitudinal extension of the affected volume along the digit. This perspective highlights the relationship between electrode positioning and the spatial distribution of the electric field within the tumor.

Furthermore, the proximity of the electrodes to the cortical bone influences the propagation of the electric field due to the marked differences in tissue electrical conductivity. This heterogeneity contributes to a non-uniform electric field distribution and reinforces the importance of considering tissue composition during treatment planning.

Overall, the adopted electrode configuration was capable of inducing electroporation over a substantial portion of the tumor volume while preserving most of the surrounding healthy tissues. Nevertheless, the persistence of peripheral *blind spots* indicates that alternative electrode arrangements, different pulse protocols, or complementary strategies, such as conductive gel coupling, may further improve treatment coverage. These findings demonstrate the usefulness of the proposed computational model as a tool for evaluating electrode configurations and supporting electrochemotherapy treatment planning.

IV. Conclusion

This work presented a simplified finite element model of a canine digit for analyzing electric field distribution during electrochemotherapy. The model reproduced the main tissues of the region of interest and enabled the spatial assessment of tumor coverage from complementary frontal and lateral perspectives. The evaluated electrode configuration produced electric field levels compatible with electroporation over a substantial portion of the tumor volume; however, peripheral regions remained below the adopted threshold, indicating incomplete treatment coverage and the occurrence of potential *blind spots*. These results highlight the influence of electrode positioning and tissue electrical heterogeneity, particularly the presence of bone structures, on the resulting electric field distribution. Despite the geometric simplifications adopted, the proposed model provides a useful framework for evaluating electrode configurations and treatment parameters. Future studies should investigate patient-specific anatomies, alternative electrode arrangements, conductive gel coupling, protocol-dependent conductivity models, and experimental validation of the numerical results.

AI Use Statement

The authors declare the use of *ChatGPT*, an artificial intelligence language model developed by *OpenAI*, as an auxiliary tool for text revision, grammatical correction, and improvement of the manuscript's clarity. All scientific content, analyses, interpretations, and conclusions were reviewed and validated by the authors, who assume full responsibility for the final version of the manuscript.

References

[1] MIR, L. M.; GEHL, J.; SERSA, G.; et al. Standard operating procedures of the electrochemotherapy: Instructions for the use of bleomycin or cisplatin administered either systemically or locally and electric pulses delivered by the Cliniporator™ by means of invasive or non-invasive electrodes. European Journal of Cancer Supplements, v. 4, n. 11, p. 14–25, 2006.

[2] GUEDERT, R.; ANDRADE, D. L. L. S.; SILVA, J. R.; et al. Dynamic model of tissue electroporation on the basis of biological dispersion and Joule heating. Journal of Applied Physics, 2024.

[3] D. L. L. S. Andrade, R. Guedert, G. B. Pintarelli, M. M. M. Rangel, K. D. Oliveira, P. G. Quadros, and D. O. H. Suzuki, "Electrochemotherapy treatment safety under parallel needle deflection," Scientific Reports, vol. 12, Art. no. 2766, 2022. https://pubmed.ncbi.nlm.nih.gov/29577784/

[4] LOPES, L. B.; PINTARELLI, G. B.; SANTOS, C. S. F.; SUZUKI, D. O. H. Computer optimization of conductive gels for electrochemotherapy. *Medical Engineering & Physics*, v. 98, p. 133–139, 2021. doi:10.1016/j.medengphy.2021.10.011.

[5] L. B. Lopes, G. B. Pintarelli, R. Guedert, D. O. H. Suzuki, et al., "Novel tetrapolar single-needle electrode for electrochemotherapy in bone cavities: Modeling, design and validation," Medical Engineering and Physics, vol. 125, p. 104120, 2024.

[6] COROVIC, S.; LACKOVIC, I.; SUSTARIC, P.; SUSTAR, T.; RODIC, T.; MIKLAVCIC, D. Modeling of electric field distribution in tissues during electroporation. BioMedical Engineering Online, v. 12, n. 16, 2013. DOI: 10.1186/1475-925X-12-16.

[7] EVANS, H. E.; DE LAHUNTA, A. Miller's Anatomy of the Dog. 4. ed. St. Louis: Elsevier Saunders, 2013.

[8] MSD VETERINARY MANUAL. Structure of the Skin in Dogs. Disponível em: MSD Veterinary Manual. Acesso em: 2026.

[9] GABRIEL, S.; LAU, R. W.; GABRIEL, C. The dielectric properties of biological tissues: III. Parametric models for the dielectric spectrum of tissues. Physics in Medicine and Biology, v. 41, n. 11, p. 2271–2293, 1996. DOI: 10.1088/0031-9155/41/11/003.

[10] D. Sel, D. Cukjati, D. Batiuskaite, T. Slivnik, L. M. Mir, and D. Miklavčič, "Sequential finite element model of tissue electropermeabilization," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 5, pp. 816–827, 2005.

[11] R. Guedert, G. B. Pintarelli, F. R. M. B. Silva, and D. O. H. Suzuki, "Effects of pulse repetition rate in static electrochemotherapy models," *Bioelectrochemistry*, vol. 153, Art. no. 108499, 2023, doi: 10.1016/j.bioelechem.2023.108499.

Influência dos canais HCN na modulação da corrente pós-sináptica inibitória no hipocampo de ratos

Kleber de Jesus Silva
Programa de Pós-graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
kl3b3rr99@gmail.com

Arthur Carvalho Pires
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
arthurcpires@hotmail.com

Carla Alessandra Scorza
Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, SP, Brasil
carla.scorza@unifesp.br

Antônio-Carlos Guimarães de Almeida
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
acga@ufsj.edu.br

Fúlvio Alexandre Scorza
Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, SP, Brasil
scorza@unifesp.br

Antônio Márcio Rodrigues
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
amr@ufsj.edu.br

***Resumo*—A ativação dos canais HCN é regulada pelo potencial de membrana e por nucleotídeos cíclicos, como o cAMP. Estes canais são expressos ao longo dos axônios de interneurônios GABAérgicos corticais sensíveis a parvalbumina e sua atividade é descrita como um mecanismo essencial na geração de padrões de disparos característicos desse grupo celular. Experimentalmente foi observado que a alteração na atividade dos HCNs pode modificar a dinâmica das sinapses GABAérgicas. No presente trabalho, reproduzimos e investigamos achados experimentais por meio de modelagem matemática e simulações computacionais. Nossas análises indicam que o bloqueio dos canais HCN pode alterar as permeabilidades de Ca^{2+} no terminal pré-sináptico, e em consequência, o comportamento de correntes pós-sinápticas inibitórias no hipocampo de rato.**

Palavras-chave—interneurônio, sinapses, HCN, hipocampo, CA1, modelagem computacional

I. Introdução

Interneurônios GABAérgicos que expressam parvalbumina (PV+) desempenham um papel fundamental na regulação da excitabilidade neuronal na formação hipocampal, contribuindo para a prevenção da hipersincronização neuronal [1], [2], [3]. Esses interneurônios podem disparar em alta frequência, gerar e sustentar disparos persistentes. Essas propriedades intrínsecas desse grupo celular são essenciais para manter o controle inibitório contínuo sobre as células excitatórias e é frequentemente associada à expressão de canais HCN (*Hyperpolarization-activated Cyclic Nucleotide-gated*) em seus axônios [4], [5].

Os canais HCN pertencem à superfamília dos canais de K^+ [6] e possuem características distintas: são permeáveis a Na^+, K^+ e também a Ca^{2+}, são ativados por potenciais de membrana polarizados ($\approx$ -55 mV) e modulados por nucleotídeos cíclicos, especialmente o cAMP [7], [8]. Esses canais são fundamentais para a regulação da excitabilidade neuronal [9] e apresentam variações na sua expressão (isoformas 1 a 4) no sistema nervoso central, com a isoforma HCN1 predominando no hipocampo e no neocórtex, enquanto HCN2 e HCN4 têm maior expressão no tálamo [10], [11], [12], [13], [14]. Além da localização, as subunidades diferem na sensibilidade ao cAMP, que é significativamente superior nas isoformas HCN2 e HCN4 em comparação à HCN1 [15].

A ligação do cAMP ao domínio intracelular dos canais HCN desloca o limiar de ativação para potenciais menos polarizados [16], integrando a dinâmica elétrica ao estado metabólico [7], [17]. Essa integração é sustentada pela cooperação metabólica entre neurônios e astrócitos, descrita pela hipótese ANLSH (*Astrocyte-Neuron Lactate Shuttle Hypothesis*). Nesse modelo, a atividade neuronal estimula a glicólise nos astrócitos, resultando na produção de lactato, que é transferido para os neurônios para servir como substrato oxidativo eficiente na geração de ATP [18], [19]. Como o ATP é substrato para a síntese de cAMP pela adenilato ciclase, o suporte energético fornecido pelos astrócitos torna-se um determinante direto da modulação dos canais HCN, permitindo que a excitabilidade neuronal seja ajustada conforme a disponibilidade metabólica local [20].

A transmissão sináptica dos interneurônios PV+ na formação hipocampal pode ser modulada pela atividade dos canais HCN. Estudos de *patch-clamp* realizados na região CA1 revelam que a inibição farmacológica desses canais com ZD7288 resulta em uma redução acentuada da corrente pós-sináptica inibitória (IPSC). Essa queda na IPSC foi verificada ao aplicar um pulso despolarizante de 0,2 ms na membrana do interneurônio e registrar a resposta em uma célula piramidal conectada. Os autores também avaliaram a IPSC com injeção de pulsos pareados no interneurônio e demonstraram que, com o bloqueio dos HCNs, a resposta ao segundo pulso apresenta menor redução de amplitude [21]. Esses achados sugerem que a atividade dos canais HCN não apenas regula a excitabilidade intrínseca, mas pode também participar ativamente da plasticidade neuronal [16].

No presente trabalho, com o uso de modelagem matemática e simulações computacionais, buscamos reproduzir e investigar o comportamento observado experimentalmente [21]. O modelo utilizado teve como base

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

os trabalhos [22], [23], [24] para a descrição das células piramidais e de mecanismos de acoplamento não sinápticos. A neurotransmissão foi baseada no modelo descrito em [25]. Para representar o interneurônio PV+, os canais HCN e o metabolismo, seguimos a metodologia de [26]. As simulações permitiram propor que a influência dos canais HCN na propagação de correntes sinápticas no hipocampo de ratos pode estar relacionada à permeabilidade de Ca^{2+} nos terminais pré-sinápticos.

II. Materiais e Métodos

A. Estrutura espacial do modelo

O modelo é composto por unidades funcionais constituídas por um neurônio (interneurônio ou célula piramidal), um prolongamento astrocítico e um compartimento correspondente ao meio extracelular. Neste estudo, foi utilizada uma rede bidimensional de 5 x 5 unidades funcionais, composta por um interneurônio e 24 células piramidais. Na borda inferior da rede, foram posicionados compartimentos com concentrações iônicas constantes, representando a solução de perfusão do tecido.

B. Mecanismos não sinápticos

O movimento iônico no meio extracelular causado por eletrodifusão foi descrito com base na equação de Nernst-Planck (1), que calcula os fluxos iônicos extracelulares considerando o efeito combinado da difusão iônica e do campo elétrico.

$$\left(\frac{\partial[C]_o}{\partial t}\right)_{dif} = \frac{D}{\lambda^2}\vec{\nabla}[C]_o + \frac{zFD\omega}{RT\lambda^2}\vec{\nabla}[C]_o \cdot \vec{\nabla}V + \frac{zFD\omega}{RT\lambda^2}[C]_o\nabla^2 V, \quad (1)$$

em que

$$\omega = 0.4\left(\frac{-1}{e^{vol_{extra}-1.1\times10^{-13}+1}} + 1\right) + 0.5 \quad (1)$$

e

$$\vec{\nabla}V = -\frac{RT}{F}\frac{\sum z_C D_C \vec{\nabla}[C]_o}{\sum {z_C}^2 D_C [C]_o} \quad (3)$$

onde C representa as concentrações dos íons Na^+, K^+, Cl^-, HCO_3^-, H^+. e λ a tortuosidade do meio extracelular, z a valência, F a constante de Faraday, T a temperatura, D a constante de difusão.

O campo elétrico extracelular foi modelado com duas componentes: uma refletindo a distribuição iônica ao longo do meio extracelular, responsável pela variação lenta do potencial extracelular (3), e outra associada às correntes transmembranares de Na^+ e K^+ (4) [22], [23].

$$V_{extra,rápido} = \tau F \sum_{n=1}^{N} \frac{A_m^{NEURON}\left(\Phi_{Na,canal\ iônico}^{NEURON} + \Phi_{K,canal\ iônico}^{NEURON}\right)_n}{d_n}, \quad (4)$$

em que

$$\tau = \bar{\tau}\left[16\left(\frac{-1}{exp\left(-\frac{vol_{extra} - 2{,}50\times10^{-1}}{(8\times10^{-16})}\right)+1} + 1\right) + 1\right], \quad (5)$$

e, então:

$$I_{campo\ elétrico} = \varepsilon V_{extra,rápido} \quad (6)$$

Fluxos por cootransportadores cátion cloreto (KCC, NKCC), trocadores iônicos que regulam pH (NHE, HCE), bomba de H^+ estão presentes no modelo e foram calculados utilizando o formalismo de Michaelis-Menten [23], [24].

A bomba de Na^+/K^+ também foi incluída no modelo, acoplando o efluxo de Na^+ ao influxo de K^+, na respectiva proporção estequiométrica de 3:2, utilizando energia proveniente da hidrólise de ATP. Para calcular a corrente gerada pela bomba, consideramos um íon fictício A^+ resultante do efluxo líquido de carga positiva dos íons transportados na direção oposta. As concentrações intra e extracelulares do íon A^+ foram determinadas usando o modelo de Albers-Post, com base nas reações dos ligantes com a enzima Na^+/K^+ ATPase [22].

O cálculo das correntes por canais iônicos dependentes do potencial e pela bomba de Na^+/K^+ é dado pela equação de Goldman-Hodgkin-Katz (GHK)

$$\Phi_{C,mech}^{CELL} = P_{C,mech} z_C F \frac{V_m^{cell}}{RT}\left(\frac{[C]_i^{cell}\, exp\left(-\frac{z_C F V_m^{cell}}{RT}\right) - [C]_o}{exp\left(-\frac{z_C F V_m^{cell}}{RT}\right) - 1}\right), \quad (7)$$

O potencial de membrana foi calculado pela extensão da equação de GHK para corrente:

$$Vm = \frac{RT}{F} ln\left(\frac{P_{Na}[Na^+]_o + P_K[K^+]_o + P_{Cl}[Cl^-]_i + P_A[A^+]_o + f_{ef} + f_{ed} + Inj}{P_{Na}[Na^+]_i + P_K[K^+]_i + P_{Cl}[Cl^-]_o + P_A[A^+]_i}\right) \quad (8)$$

onde f_{ef}, f_{ed} representam a influência do efeito de campo e eletrodifusão e Inj os pulsos injetados.

As permeabilidades dos canais iônicos foram calculadas utilizando o formalismo de Hodgkin-Huxley [27]. No interneurônio e células piramidais, foram considerados:

- Canais de Na^+ de dinâmica rápida (*fast*);
- Canais persistentes de Na^+;
- Canais de K^+ do tipo A;
- Canais retificadores de K^+ com atraso;
- Canais muscarínicos de K^+;
- Canais de K^+ do tipo Kir;
- Canais de cloreto com permeabilidades constantes.

No interneurônio, também foram representados canais do tipo L, usados para estimar a produção de cAMP pela enzima adenilato ciclase. A membrana dos astrócitos foi considerada inexcitável. Apenas o canal retificador de entrada de K^+ foi

modelado de forma dependente do potencial; para os demais canais, as permeabilidades foram consideradas constantes.

As permeabilidades para Na^+ e K^+ dos canais HCN expressos nos interneurônios foram calculadas da seguinte forma:

$$P_{Na}^{HCNC} = \overline{P_{Na}^{HCNC}} f_{HCNC} q_{Na}^{HCNC}, \tag{9}$$

$$P_{K}^{HCNC} = \overline{P_{K}^{HCNC}} f_{HCNC} q_{K}^{HCNC}, \tag{10}$$

em que

$$\frac{dq_{Na}^{HCNC}}{dt} = \frac{\left[0.085\left(\tanh\left(\frac{[cAMP]^i - 0.039}{0.0014}\right) + 1.0\right) + 0.05\right] - q_{Na}^{HCNC}}{0.001} \tag{11}$$

$$\frac{dq_{K}^{HCNC}}{dt} = \frac{\left[0.5\left(\tanh\left(\frac{[cAMP]^i - 0.01}{0.0045}\right) + 1.0\right) + 0.63\right] - q_{K}^{HCNC}}{0.001} \tag{12}$$

$$f_{HCNC} = 0.5\left(\tanh\left(-\frac{Vn-(-84+20\cdot DV_{HCNC})}{13.1}\right) + 1.0\right), \tag{13}$$

$$DV_{HCNC} = 0.5\left(\tanh\left(\frac{[cAMP]^i - 0.039}{0.0018}\right) + 1.0\right). \tag{14}$$

O metabolismo energético e as interações entre interneurônio e astrócito também foram descritas. A produção de cAMP a partir de ATP pela adenilato ciclase foi definida como:

$$ATP^i \xrightarrow{adenilato\ ciclase} cAMP^i + 2P^i, \tag{15}$$

A taxa de reação é dada por:

$$r_{AC} = \propto_{AC} \frac{[ATP]^i}{[ATP]^i + 0.2} \times (0.001 + y_{INa}), \tag{16}$$

em que y_{INa} é a função de transferência dependente da corrente de Na^+ usada para representar o efeito no canal do tipo L que ativa a adenilato ciclase:

$$y_{INa} = y_{INa} e^{-DT/0.2\times10^{-3}} + y_1 0.2(1 - e^{-DT/0.2\times10^{-3}}), \tag{17}$$

$$y_1 = y_1 e^{-DT/2.5\times10^{-3}} + y_2 2.5(1 - e^{-DT/2.5\times10^{-3}}), \tag{18}$$

onde y_2 é o estímulo de entrada que identifica a ocorrência do potencial de ação através da taxa de variação da corrente de Na^+, e é dado por:

$$y_2 = \begin{cases} 1\ se\ \frac{dINa}{dt} < -3.07 \times 10^3\ e\ ContarPA = 0 \\ 0\ se\ não \end{cases}, \tag{19}$$

$$ContarPA = \begin{cases} ContarPA + 1\ \ se\ 0 < ContarPA < 10 \\ 0\ se\ não \end{cases}., \tag{20}$$

Nas simulações, y_{INa} e y_1 recebem valores inicialmente nulos. Para a conversão de cAMP para AMP, foi descrita a reação relacionada ao efeito da fosfodiesterase:

$$cAMP^i \xrightarrow{PDE} AMP^i, \tag{21}$$

$$r_{PDE} = \propto_{PDE} \frac{[cAMP]^i}{[cAMP]^i + 0.0201} \tag{22}$$

A [cAMP] foi estimada por:

$$\frac{d[cAMP]^i}{dt} = r_{AC} - r_{PDE} \tag{23}$$

C. Transmissão sináptica

A transmissão sináptica foi representada por um conjunto de passos. Primeiramente, a membrana pré-sináptica é despolarizada devido à ocorrência do potencial de ação no neurônio pré-sináptico. A despolarização aumenta a permeabilidade de Ca^{2+}, promovendo o influxo desse cátion no terminal pré-sináptico. O influxo de Ca^{2+} promove a liberação de neurotransmissores na fenda sináptica. Esses neurotransmissores difundem-se até a membrana pós-sináptica, onde se ligam a receptores e promovem a abertura de canais iônicos. A depender do tipo de canal iônico e do potencial de Nernst do íon em questão, a corrente pode ser excitatória ou inibitória. Após permanecerem ligados por determinado intervalo de tempo, os neurotransmissores são inativados devido à ação enzimática, são transportados novamente para o terminal pré-sináptico e armazenados em vesículas [25].

Seguindo as etapas envolvidas no processo de neurotransmissão, foram incluídas no modelo as seguintes equações:

- Permeabilidade de Ca^{2+} no terminal pré-sináptico:

$$P_{Ca} = \bar{P}_{Ca} m^3 h, \tag{24}$$

onde,

$$\alpha_m = \frac{C_{\alpha_m}(Vt - Vm + 25)}{e^{\frac{Vm+2}{10}} - 1}, \tag{25}$$

$$\beta_m = C_{\beta_m} e^{\frac{Vt-V}{18}}, \tag{26}$$

- Fluxo de Ca^{2+} através dos canais dependentes de potencial:

$$\phi_{Ca,cha} = \frac{F^2}{RT} z_{Ca}^2 P_{Ca} V_m \frac{[\mathrm{Ca}]_i e^{\frac{z_{Ca}FVm}{RT}} - [Ca]_o}{e^{\frac{z_{Ca}FVm}{RT}} - 1}, \quad (27)$$

- Fluxo de Ca^{2+} pela bomba iônica:

$$\phi_{Ca,pump} = \frac{P_{Ca,pump}^{max}}{1 - e^{-3\times10^{-3}([Ca]_i - 1\times10^{-2})}}, \quad (28)$$

- Fluxo devido à movimentação ao longo do axônio, por difusão:

$$\phi_{Ca,dif} = \gamma(8 \cdot 10^{-2} - [Ca]_i), \quad (29)$$

- Em função dos fluxos descritos, as taxas de variação das concentrações intra e extracelular de Ca^{2+} são dadas por:

$$\frac{d[Ca]_i}{dt} = -\frac{f_{vol}}{Vol_{syn}}\left(\phi_{Ca,pump} + \phi_{Ca,channel} + \phi_{Ca,eq}\right) + \phi_{Ca,dif}, \quad (30)$$

$$\frac{d[Ca]_o}{dt} = \frac{f_{vol}}{Vol_o}\left(\phi_{Ca,pump} + \phi_{Ca,channel} + \phi_{Ca,eq}\right), \quad (31)$$

- Taxa de variação da concentração de neurotransmissores no terminal pré-sináptico:

$$\frac{d[NT]_i}{dt} = 10.4(-\Delta T_o + 2\times10^{-3} T_v), \quad (32)$$

- Taxa de liberação de neurotransmissores na fenda sináptica:

$$\frac{d[NT]_o}{dt} = 8\times10^2(\Delta T_o - \Delta T_o R - \Delta T_o E), \quad (33)$$

- Taxa de ligação dos neurotransmissores aos receptores da membrana pós-sináptica:

$$\frac{dToR}{dt} = 8\times10^2 \Delta T_o R, \quad (34)$$

- Taxa de inativação dos neurotransmissores:

$$\frac{dT_oE}{dt} = 8\times10^2(\Delta T_o - \Delta T_o E - \Delta T_o EER), \quad (35)$$

- Taxa de armazenamento de neurotransmissores nas vesículas pré-sinápticas:

$$\frac{dT_v}{dt} = 8\times10^2(\Delta T_o EER - 2\times10^{-3} T_v), \quad (36)$$

em que

$$\Delta T_o = 6\times10^2 [NT]_i([Ca]_i + [Ca]_{i,repouso}), \quad (37)$$

$$\Delta T_o R = 1{,}5[NT]_o - 3T_o R, \quad (38)$$

$$\Delta T_o E = 1{,}5[NT]_o - 3T_o R, \quad (39)$$

$$\Delta T_o EER = \frac{0{,}16}{1+e^{110([NT]_i - 0{,}98)}} T_o E, \quad (40)$$

- A corrente pós-sináptica inibitória (IPSC) foi considerada proporcional à concentração de neurotransmissores ligados aos receptores pós-sinápticos:

$$I_{inib} = \delta T_o R, \quad (41)$$

onde δ (pA) é uma constante de proporcionalidade, que representa o efeito das forças difusiva e elétrica através da membrana pós-sináptica, consideradas constantes no modelo.

D. Implementação computacional

O modelo foi implementado em linguagem Fortran (Fortran 90), utilizando o método de diferenças finitas com erro menor que 10^{-4}. Os resultados foram analisados em Matlab®. Valores iniciais e parâmetros foram baseados nos trabalhos anteriores [23], [24], [25], [26]. O código do programa foi compilado e executado em um computador de alta performance (Cluster SGI UV 2000, 80 *cores* — Intel Xeon E5-4650v2 10-core, 2.4 GHz, 25 MB Cache, 1024 GB DDR3 1866 MHz, 80 TB HD, Cluster SGI UV 300, 144 *cores* Intel Xeon E7-8867v4, 2.4 GHz, RAM 2048 GB DDR4 2400 MHz, 80 TB HD, SUSE Linux Enterprise Server 11, SGI Performance Suite, Intel Cluster Studio XE).

III. RESULTADOS

Simulando o protocolo experimental [21], um pulso retangular de potencial foi aplicado durante 0,2 ms na membrana do interneurônio. A Fig. 1B mostra que a IPSC não sofreu alteração quando os canais HCN foram bloqueados (Fig. 1B, 100% PCa). Para esse caso, o acúmulo de Ca^{2+} intracelular durante o potencial de ação e, consequentemente, os comportamentos das concentrações de neurotransmissores no terminal pré-sináptico e na fenda sináptica não são alterados pelo bloqueio dos canais HCN (Fig. 1C).

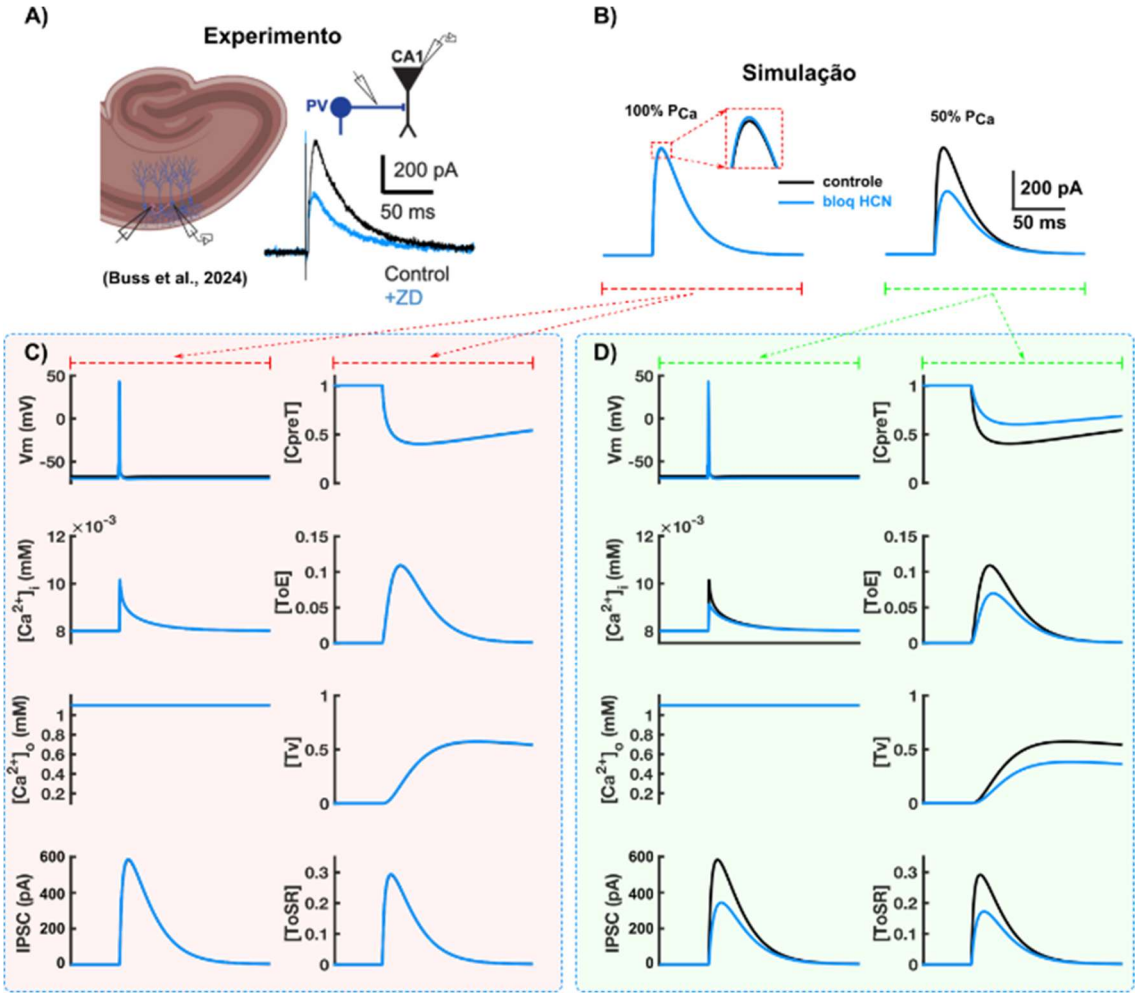


Fig. 1. Redução da IPSC causado pelo bloqueio dos canais HCN. (A) Registro experimental [22]. (B) IPSC simulada no caso controle e bloqueando os HCN em duas situações: 100% PCa – sem alterar o influxo de Ca2+; 50% PCa – reduznindo o influxo de Ca2+. Os comportamentos das concentrações pré-sinápticas e na fenda sináptica de Ca2+ e neurotransmissores são apresentadas para os dois casos: 100% PCa (C) e 50% PCa (D). (preto – controle;. azul – bloqueio do HCN; as curvas em azul se sobrepõem às curvas em preto).

Como os canais HCN também apresentam permeabilidade ao Ca^{2+} [8], [21], o seu bloqueio deveria reduzir o influxo de Ca^{2+}. Considerando que a permeabilidade do Ca^{2+} cai durante o bloqueio dos HCNs, a simulação foi capaz de reproduzir o comportamento de redução da IPSC (Fig. 1B, 50% PCa) observada experimentalmente. Nesse caso, Fig. 1D, durante o bloqueio dos HCNs, a redução do influxo de Ca^{2+} reduz o seu acúmulo no meio intracelular do neurônio pré-sináptico. Isso leva a uma redução na quantidade de neurotransmissores liberados na fenda sináptica e na ligação aos receptores na membrana pós-sináptica. Logo, a IPSC é menos intensa (curvas em azul) do que no caso controle (curvas em preto).

Experimentalmente também foi observado que o bloqueio dos HCNs altera a razão das amplitudes da IPSC de pulsos pareados aplicados com intervalo de 50 ms [21], [22] (Fig. 2A). Como pode ser observado na Fig. 2B, a simulação permite reproduzir o comportamento experimental, sendo que o bloqueio dos HCNs associado com a redução do influxo de Ca^{2+}, apesar de reduzir a amplitude da IPSC, aumenta a razão (amplitude da segunda IPSC) / (amplitude da primeira IPSC) durante esse protocolo de estimulação. Observa-se que, com o bloqueio dos canais HCN, o menor acúmulo de Ca^{2+} intracelular durante cada potencial de ação (Fig. 2C) resulta em uma variação menos acentuada da concentração de neurotransmissores. Esse efeito atenua a depressão sináptica, de modo que a resposta ao segundo pulso sofre menor redução em comparação com a resposta ao primeiro.

IV. Discussão

Alterações na expressão e na atividade dos canais HCN podem estar relacionadas a diferentes patologias do sistema nervoso central [28], [29], [30]. Portanto, compreender como esses canais modulam a transmissão sináptica pode contribuir para a identificação de mecanismos fisiopatológicos relevantes e orientar, futuramente, o desenvolvimento de estratégias terapêuticas mais específicas.

Para representar o interneurônio, utilizamos um modelo que inclui uma descrição dos canais HCN e metabolismo celular [26]. Como observado experimentalmente, os canais HCN também são permeáveis a Ca^{2+} [8], embora essa permeabilidade seja relativamente baixa, ela pode exercer efeito sobre o processo de neurotransmissão [21]. Reforçando essa hipótese, nas simulações, quando a permeabilidade de Ca^{2+} é reduzida durante o bloqueio dos canais HCN, a IPSC apresenta menor intensidade (Fig. 1B e 1D, 50% PCa). De acordo com a simulação, a redução do influxo de Ca^{2+} causa uma menor liberação de neurotransmissores na fenda sináptica. Logo, a IPSC em resposta ao pulso despolarizante apresenta menor amplitude. Sustentando essa hipótese de que o bloqueio dos canais HCN reduz o influxo de Ca^{2+} [21], em resposta a pulsos pareados, a amplitude do segundo pulso

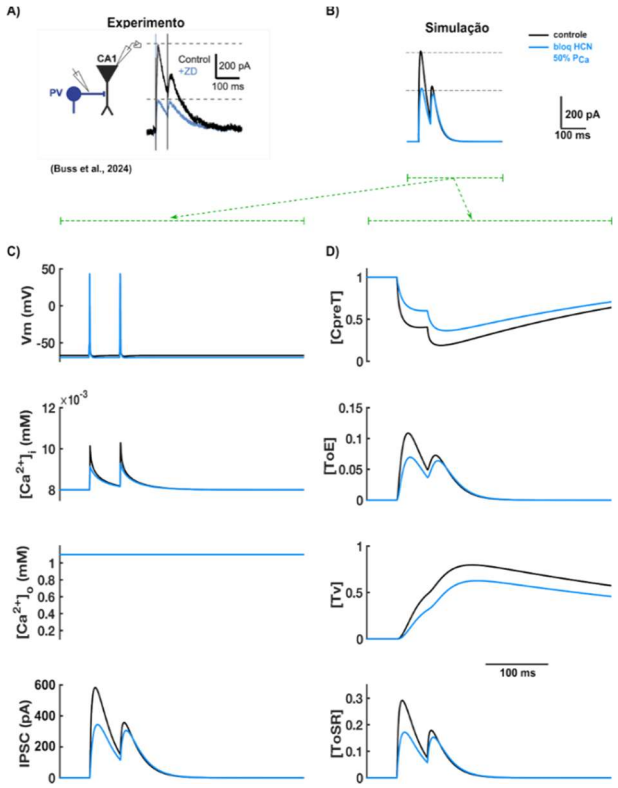


Fig. 2. O bloqueio dos canais HCN aumenta a razão de pulsos pareados das correntes pós-sinápticas inibitórias (IPSCs) evocadas pela estimulação de interneurônios. (A) Registro experimental [22].. (B) IPSC simulada no caso controle e bloqueando os HCN. (C) e (D) Comportamentos das concentrações pré-sinápticas e na fenda sináptica de Ca2+ e neurotransmissores durante os potenciais de ação gerados pelos pulsos.

apresenta menor redução em relação ao primeiro pulso quando comparado ao caso em que os canais HCN não são bloqueados, conforme exibido na Fig. 2. Segundo a simulação, combinando o bloqueio do HCN com um menor influxo de Ca^{2+}, a liberação de neurotransmissores fica menos intensa. Consequentemente, a concentração de neurotransmissores pré-sinápticos reduz menos, afetando pouco a amplitude da resposta ao segundo pulso injetado.

Nas simulações realizadas foi necessário reduzir a permeabilidade ao Ca^{2+} em aproximadamente 50% do valor basal para reproduzir o comportamento observado experimentalmente. Essa redução é substancialmente maior do que os 6% previstos na literatura [8], [21]. Os resultados permitem levantar a hipótese de que os canais HCN estão presentes nos terminais axônicos (pré-sinapses) dos interneurônios, contribuindo com cerca de 50% da permeabilidade total de Ca^{2+}.

AGRADECIMENTOS

Agradecemos à FAPEMIG, ao CNPq, à UFSJ e à CAPES.

REFERÊNCIAS

[1] T. F. Freund and G. Buzsáki, “Interneurons of the hippocampus,” *Hippocampus*, vol. 6, no. 4, pp. 347–470, 1996, doi: 10.1002/(SICI)1098-1063(1996)6:4<347::AID-HIPO1>3.0.CO;2-I.

[2] B. Hattiangady and A. K. Shetty, “Decreased neuronal differentiation of newly generated cells underlies reduced hippocampal neurogenesis in chronic temporal lobe epilepsy,” *Hippocampus*, vol. 20, no. 1, pp. 97–112, Jan. 2010, doi: 10.1002/hipo.20594.

[3] D. Upadhya *et al.*, “Human induced pluripotent stem cell-derived MGE cell grafting after status epilepticus attenuates chronic epilepsy and comorbidities via synaptic integration,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 1, pp. 287–296, Jan. 2019, doi: 10.1073/pnas.1814185115.

[4] C. Elgueta and M. Bartos, “Dendritic inhibition differentially regulates excitability of dentate gyrus parvalbumin-expressing interneurons and granule cells,” *Nat. Commun.*, vol. 10, no. 1, p. 5561, Dec. 2019, doi: 10.1038/s41467-019-13533-3.

[5] C. Elgueta, J. Köhler, and M. Bartos, “Persistent discharges in dentate gyrus perisoma-inhibiting interneurons require hyperpolarization-activated cyclic nucleotide-gated channel activation,” *J. Neurosci.*, vol. 35, no. 10, pp. 4131–4139, Mar. 2015, doi: 10.1523/JNEUROSCI.3671-14.2015.

[6] K. B. Craven and W. N. Zagotta, “CNG and HCN channels: two peas, one pod,” Annu. Rev. Physiol., vol. 68, pp. 375–401, Feb. 2006, doi: 10.1146/annurev.physiol.68.040104.134728.

[7] M. Biel, C. Wahl-Schott, S. Michalakis, and X. Zong, “Hyperpolarization-activated cation channels: from genes to function,” *Physiol. Rev.*, vol. 89, no. 3, pp. 847–885, Jul. 2009, doi: 10.1152/physrev.00029.2008.

[8] X. Yu, K.-L. Duan, C.-F. Shang, H.-G. Yu, and Z. Zhou, “Calcium influx through hyperpolarization-activated cation channels (Ih channels) contributes to activity-evoked neuronal secretion,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 4, pp. 1051–1056, Jan. 2004, doi: 10.1073/pnas.0305167101.

[9] I. Rivolta, A. Binda, A. Masi, and J. C. DiFrancesco, “Cardiac and neuronal HCN channelopathies,” *Pflüg. Arch. - Eur. J. Physiol.*, vol. 472, no. 7, pp. 931–951, Jul. 2020, doi: 10.1007/s00424-020-02384-3.

[10] H. C. Lai and L. Y. Jan, “The distribution and targeting of neuronal voltage-gated ion channels,” *Nat. Rev. Neurosci.*, vol. 7, no. 7, pp. 548–562, Jul. 2006, doi: 10.1038/nrn1938.

[11] S. Moosmang, M. Biel, F. Hofmann, and A. Ludwig, “Differential distribution of four hyperpolarization-activated cation channels in mouse brain,” Biol. Chem., vol. 380, no. 7–8, pp. 975–980, Jul. 1999, doi: 10.1515/BC.1999.121.

[12] R. Narayanan and D. Johnston, “The h channel mediates location dependence and plasticity of intrinsic phase response in rat hippocampal neurons,” *J. Neurosci.*, vol. 28, no. 22, pp. 5846–5860, May 2008, doi: 10.1523/JNEUROSCI.0835-08.2008.

[13] T. Notomi and R. Shigemoto, “Immunohistochemical localization of Ih channel subunits, HCN1–4, in the rat brain,” *J. Comp. Neurol.*, vol. 471, no. 3, pp. 241–276, 2004, doi: 10.1002/cne.11039.

[14] Z. Nusser, “Differential subcellular distribution of ion channels and the diversity of neuronal function,” *Curr. Opin. Neurobiol.*, vol. 22, no. 3, pp. 366–371, Jun. 2012, doi: 10.1016/j.conb.2011.10.006.

[15] C. L. Combe and S. Gasparini, “Ih from synapses to networks: HCN channel functions and modulation in neurons,” *Prog. Biophys. Mol. Biol.*, vol. 166, pp. 119–132, Nov. 2021, doi: 10.1016/j.pbiomolbio.2021.06.002.

[16] S. S. Chow, F. Van Petegem, and E. A. Accili, “Energetics of cyclic AMP binding to HCN channel C terminus reveal negative cooperativity,” *J. Biol. Chem.*, vol. 287, no. 1, pp. 600–606, Jan. 2012, doi: 10.1074/jbc.M111.269563.

[17] A. Momin, H. Cadiou, A. Mason, and P. A. McNaughton, “Role of the hyperpolarization-activated current Ih in somatosensory neurons,” *J. Physiol.*, vol. 586, no. 24, pp. 5911–5929, 2008, doi: 10.1113/jphysiol.2008.163154.

[18] L. Pellerin and P. J. Magistretti, “Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization,” *Proc. Natl. Acad. Sci.*, vol. 91, no. 22, pp. 10625–10629, Oct. 1994, doi: 10.1073/pnas.91.22.10625.

[19] M. A. Tarczyluk *et al.*, “Functional astrocyte-neuron lactate shuttle in a human stem cell-derived neuronal network,” *J. Cereb. Blood Flow Metab.*, vol. 33, no. 9, pp. 1386–1393, Sep. 2013, doi: 10.1038/jcbfm.2013.81.

[20] A.-K. Bouzier-Sore, P. Voisin, P. Canioni, P. J. Magistretti, and L. Pellerin, “Lactate is a preferential oxidative energy substrate over glucose for neurons in culture,” *J. Cereb. Blood Flow Metab.*, vol. 23, no. 11, pp. 1298–1306, Nov. 2003, doi: 10.1097/01.WCB.0000091761.61714.25.

[21] E. W. Buss *et al.*, “HCN1 hyperpolarization-activated cyclic nucleotide–gated channels enhance evoked GABA release from parvalbumin-positive interneurons,” *Proc. Natl. Acad. Sci.*, vol. 121, no. 42, p. e2319246121, Oct. 2024, doi: 10.1073/pnas.2319246121.

[22] A.-C. G. de Almeida et al., “Mechanistic hypotheses for nonsynaptic epileptiform activity induction and its transition from the

interictal to ictal state—computational simulation," Epilepsia, vol. 49, no. 11, pp. 1908–1924, Nov. 2008, doi: 10.1111/j.1528-1167.2008.01686.x.

[23] A. M. Rodrigues, L. E. C. Santos, L. Covolan, C. Hamani, and A.-C. G. de Almeida, "pH during non-synaptic epileptiform activity—computational simulations," *Phys. Biol.*, vol. 12, no. 5, p. 056007, Sep. 2015, doi: 10.1088/1478-3975/12/5/056007.

[24] A. M. Rodrigues *et al.*, "The effect of low magnesium concentration on ictal discharges In a non-synaptic model," *Int. J. Neural Syst.*, vol. 31, no. 04, p. 2050070, Apr. 2021, doi: 10.1142/S0129065720500707.

[25] H. Z. Teixeira, A.-C. G. Almeida, A. F. C. Infantosi, A. M. Rodrigues, N. L. Costa, and M. A. Duarte, "Identifying essential conditions for refractoriness of Leão's spreading depression—computational modeling," *Comput. Biol. Chem.*, vol. 32, no. 4, pp. 273–281, Aug. 2008, doi: 10.1016/j.compbiolchem.2008.03.011.

[26] S. M. de Lima, "Disparos persistentes em interneurônios inibitórios do giro dentado: simulações computacionais," Ph.D. dissertation, Programa de Pós-Graduação em Bioengenharia, Universidade Federal de São João del-Rei, São João del-Rei, Brazil, 2024.

[27] A. L. Hodgkin and A. F. Huxley, "Propagation of electrical signals along giant nerve fibres," *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 140, no. 899, pp. 177–183, Oct. 1952, doi: 10.1098/rspb.1952.0054.

[28] J. C. DiFrancesco and D. DiFrancesco, "Dysfunctional HCN ion channels in neurological diseases," *Front. Cell. Neurosci.*, vol. 9, Mar. 2015, doi: 10.3389/fncel.2015.00071.

[29] S. Jung, L. N. Warner, J. Pitsch, A. J. Becker, and N. P. Poolos, "Rapid loss of dendritic HCN channel expression in hippocampal pyramidal neurons following status epilepticus," *J. Neurosci.*, vol. 31, no. 40, pp. 14291–14295, Oct. 2011, doi: 10.1523/JNEUROSCI.1148-11.2011.

[30] C. Marini *et al.*, "HCN1 mutation spectrum: from neonatal epileptic encephalopathy to benign generalized epilepsy and beyond," *Brain*, vol. 141, no. 11, pp. 3160–3178, Nov. 2018, doi: 10.1093/brain/awy263.

Implementação De Um Controlador De Barreira Para Aumento Da Segurança Em Um Dispositivo Andador

Vinicius Alves Magro Cecchi
Dep. de Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo, Brasil
vmagro@fei.edu.br

Wellington Cássio Pinheiro
Dep. de Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo, Brasil
wellington@fei.edu.br

Fabrizio Leonardi
Dep. de Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo, Brasil
fabrizio@fei.edu.br

***Resumo*—O envelhecimento populacional acarreta o aumento da incidência de patologias musculoesqueléticas e do risco de quedas, o que demanda o uso de dispositivos assistivos. Contudo, a maioria desses dispositivos atuais foca apenas na navegação sem oferecer garantias ativas de segurança contra a perda de equilíbrio do usuário. Nesse contexto, o presente trabalho propõe a implementação de uma Função de Controle de Barreira (CBF) em um andador eletrificado, inteligente, atuando como uma malha de *Safety Control* para a prevenção de quedas. Nesta fase da pesquisa, a dinâmica da marcha foi simplificada por meio de um modelo de pêndulo invertido acoplado ao andador. Um controlador de impedância foi projetado para atuar na intenção de movimento nominal, enquanto a CBF atua por meio da solução de um problema de Programação Quadrática (QP) para restringir o ângulo de inclinação do tronco a uma zona estritamente segura. Simulações do modelo matemático demonstraram que, diante de perturbações externas, o controlador intervém na trajetória de forma estritamente mínima, limitando a inclinação do usuário e evitando o tombamento. Os resultados preliminares sugerem que a técnica CBF apresenta-se como uma alternativa robusta e promissora para garantir a segurança formal na robótica assistiva.**

***Index Terms*—controle, andador, assistivo, barreira**

I. Introdução

O enorme avanço sócio-econômico e tecnológico da humanidade tornou possível um aumento substancial da expectativa de vida. Contudo, a Organização Mundial da Saúde exibe uma preocupação com as implicações desse aumento da expectativa de vida na saúde pública [1]. A expectativa de vida global atingiu 73,3 anos em 2024, um aumento de 8,5 anos desde o ano de 1995, o que resulta em uma projeção de aumento de 1,1 em 2023 para 1,4 bilhão em 2030 na população de mais de 60 anos [1].

Contudo, o avanço da expectativa de vida traz consigo o aumento de alguns problemas associados ao envelhecimento populacional. As patologias associadas a dificuldade de locomoção e redução da mobilidade estão entre as principais preocupações da comunidade científica. [2]. Em virtude do aumento de patologias relacionadas a mobilidade, anualmente, o mundo registra aproximadamente 37,3 milhões de quedas severas o suficiente para que a pessoa precise de atendimento médico de acordo com [3]. A população idosa apresenta o maior risco de morte ou ferimento sério acarretado por uma queda, sendo que 20-30% de idosos ao caírem sofrem algum tipo de ferimento moderado ou severo, como hematomas, fraturas nos quadris, traumatismo craniano etc [3].

O autor Vinicius Alves Cecchi agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de Mestrado (Processo nº 131724).

A. Justificativa

Diante do envelhecimento populacional evidenciado por [1], as patologias músculo-esqueléticas associadas ao envelhecimento, como as descritas por [2] são engrandecidas diante dos riscos que se associam a quedas por idosos e a incidência de quedas na população idosa. A fragilidade músculo-esquelética aumenta mais ainda o risco de fraturas no fêmur, quadril, outros ossos e até a morte, além do dano psicológico, o que evidencia uma crescente demanda por dispositivos assistivos que garantam não só a mobilidade mas a segurança dos usuários.

B. Objetivo

Com isso, o objetivo principal do presente trabalho é a implementação de uma função de controle de barreira em um dispositivo andador para idosos com a finalidade de impedir ou diminuir o índice de quedas de usuários.

II. Fundamentação teórica

A fundamentação teórica se divide em dois pilares principais: o primeiro deles concentra-se no estudo da biomecânica da marcha e da queda, já o segundo, constitui-se do estudo acerca das estratégias de controle aplicadas no desenvolvimento do controlador de barreira.

A. Biomecânica da Marcha e Queda

O ciclo da marcha humana é fundamentado na alternância entre os estados de suporte duplo e suporte simples. Enquanto o primeiro ocorre com o contato simultâneo de ambos os pés com o solo, o segundo caracteriza-se pelo apoio de apenas uma

das extremidades inferiores. Sob a perspectiva de um membro isolado, o movimento é segmentado em duas etapas principais: a fase de apoio, que ocupa cerca de 62% do ciclo, e a fase de balanço, que compreende os 38% restantes. Conforme proposto por [4], [5] e [6], essa progressão pode ser detalhada em oito eventos discretos.

1) Contato Inicial: Marca o começo do ciclo quando o calcanhar atinge o solo. Neste ponto, o centro de gravidade atinge seu nível mais baixo, com o quadril flexionado em cerca de 30° e o joelho em 5°. A força de reação do solo supera o peso corporal, exigindo esforço máximo do músculo reto femoral.
2) Resposta à Carga: Caracteriza-se pelo contato total da planta do pé com a superfície. Ocorre uma flexão rápida do joelho até 15°, mecanismo essencial para a absorção de impactos e continuidade do movimento.
3) Apoio Médio: Período em que o peso do corpo se projeta verticalmente sobre o membro em suporte simples. O joelho permanece em extensão enquanto o membro oposto inicia o balanço. O centro de gravidade atinge seu ápice e o joelho flexiona-se levemente (18°), com redução da atividade do reto femoral.
4) Apoio Terminal: Instante em que o calcanhar se eleva, perdendo o contato com o chão. O impulso para a progressão é gerado pela ativação dos músculos tríceps surais, resultando na flexão plantar do tornozelo.
5) Pré-balanço (Desprendimento): Etapa final do apoio, encerrada no momento em que os dedos deixam a superfície de caminhada.

Quedas são uma das principais causas de acidentes entre a população idosa. O trabalho de [7] teve como objetivo determinar o que causa uma queda entre idosos e o que dificulta a recuperação postural. Em [8], um estudo previamente conduzido os pesquisadores aplicaram um experimento no qual induziam o voluntário a tropeçar, com isso, foi possível observar que o tronco, sendo responsável por aproximadamente 50% do peso do corpo e dada a sua altura do chão, ao inclinar-se quando o indivíduo tropeça tem grande impacto na queda. A flexão do tronco pós tropeço ficava em torno de $27 \pm 9°$ e o pico de velocidade da flexão foi de $140°/s$, o que ocorreu 300ms após o tropeço.

Um tropeço é descrito por [9] como um contato inesperado do pé com a superfície que acaba gerando momento suficiente para desestabilizar o indivíduo. Sendo assim, conclui-se que os fatores-chave para uma queda por tropeço são: contato inesperado com o chão e momento excessivo.

Dada as limitações fisiológicas e neuromusculares um tropeço pode ser desafiador para um idoso ou alguém que sofra de uma patologia neuromuscular, isso porque essas condições acarretam no enfraquecimento músculo-esquelético do indivíduo, diminuindo a magnitude do momento que amortece a flexão do tronco. Além disso, o atraso nos reflexos e tempo de reação aumenta o tempo de resposta, ocasionando em uma maior flexão do tronco que é o que foi observado em [8].

Segundo [8] resvalos (desequilíbrio para trás) causam mais quedas proporcionalmente pois a recuperação deste tipo de distúrbio postural se mostra muito mais difícil do que sua contraparte. Por exemplo, para o caso de tropeços e quedas frontais nas quais o tronco é flexionado, nenhum voluntário jovem caiu, por outro lado, 25% dos mais velhos caíram. Já para um resvalo, ou seja, uma queda para trás na qual o tronco é extendido, 6% dos jovens e 86% dos mais velhos caíram. A capacidade de evitar uma queda para trás está associada com a capacidade de atenuar movimentos indesejados do tronco.

B. Controlador de Barreira

Uma função de controle de Barreira (CBF) pode ser descrita a partir de um sistema descrito pela equação 1.

$$\dot{x} = f(x) + g(x)u \tag{1}$$

Deve-se considerar uma função de barreira $h(x)$ que tende a zero. Tendo um conjunto C chamado de "conjunto seguro", se $h(x)$ tende a zero, x encontra-se dentro do conjunto seguro.

Um controlador de realimentação deve ser projetado para o sistema ((1)) com o objetivo de manter os estados $\mathbf{x}$ no conjunto de segurança, definido conforme:

$$\begin{aligned} \mathcal{C} &= \{\mathbf{x} \in D \subset \mathbb{R}^n : h(\mathbf{x}) \geq 0\}, \\ \partial\mathcal{C} &= \{\mathbf{x} \in D \subset \mathbb{R}^n : h(\mathbf{x}) = 0\}, \\ \text{Int}(\mathcal{C}) &= \{\mathbf{x} \in D \subset \mathbb{R}^n : h(\mathbf{x}) > 0\}, \end{aligned} \tag{2}$$

onde $h(\mathbf{x}) : D \subset \mathbb{R}^n \to \mathbb{R}$ é uma função continuamente diferenciável.

A função $h(\mathbf{x})$ é uma função, cuja classificação como CBF torna-se possível devido a existência de uma classe $\mathcal{K}$ (denominada α_{cbf}) tal que:

$$\sup_{u \in U} [L_f h(\mathbf{x}) + L_g h(\mathbf{x}) u + \alpha_{cbf}(h(\mathbf{x}))] \geq 0. \tag{3}$$

Considerando o conjunto seguro $\mathcal{C}$ definido em (2), o conjunto de controles $K_{cbf}(\mathbf{x})$ para o sistema ((1)) é definido como em [10]:

$$K_{cbf}(\mathbf{x}) = \{u \in U : L_f h(\mathbf{x}) + L_g h(\mathbf{x}) u + \alpha_{cbf}(h(\mathbf{x})) \geq 0\}. \tag{4}$$

C. Funções de Controle de Barreira Exponenciais

Em determinados problemas há uma limitação, dada pela própria natureza do sistema. Por vezes, a variável controlável do sistema não encontra-se na primeira derivada da função de barreira, em função disso, foram desenvolvidas funções de barreira exponenciais (ECBF), que são descritas por [10] conforme:

Definição 1:Dado um conjunto $C \subset D \subset \mathbb{R}$ definido como o supernível de um conjunto de uma função r vezes diferenciável, sendo $h(x) : D \to \mathbb{R}$ então, h(x) é uma ECBF se existe um vetor $K_b \in \mathbb{R}$ de tal forma que,

$$\sup_{u \in U} \left[L_f^r h(\mathbf{x}) + L_g L_f^{r-1} h(\mathbf{x}) u + K_b \eta_b(\mathbf{x}) \right] \geq 0 \tag{5}$$

onde o vetor de estados estendido $\eta_b(\mathbf{x})$ é definido como:

$$\eta_b(\mathbf{x}) := \begin{bmatrix} h(\mathbf{x}) \\ \dot{h}(\mathbf{x}) \\ \ddot{h}(\mathbf{x}) \\ \vdots \\ h^{(r-1)}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} h(\mathbf{x}) \\ L_f h(\mathbf{x}) \\ L_f^2 h(\mathbf{x}) \\ \vdots \\ L_f^{(r-1)} h(\mathbf{x}) \end{bmatrix} \tag{6}$$

III. Metodologia

A abordagem metodológica consiste primariamente no desenvolvimento de um modelo matemático capaz de representar a marcha humana de forma simplificada, mas que apresente coerência com a realidade do movimento humano. Posteriormente, faz-se necessário projetar um controlador capaz de garantir a segurança do usuário, por fim, foi desenvolvido um protocolo experimental para simulação computacional de uma queda induzida.

A. Modelagem Biomecânica

O modelo do pêndulo invertido é amplamente utilizado para a representação matemática da marcha, seja ela humana ou robótica. Sendo assim, utiliza-se um pêndulo invertido para representar o movimento do corpo humano ao longo do plano sagital, no qual ele se alterna entre situações de equilíbrio e desequilíbrio controlado para gerar movimento de avanço.

O pêndulo é representado como uma barra de comprimento l sem massa, e toda a massa (m_h) do corpo está concentrada na ponta desta barra, representando o centro de gravidade (CG) do usuário. Conforme o usuário se movimenta, o pêndulo forma um ângulo de deslocamento θ em relação a sua posição inicial. Naturalmente, um ser humano tem uma rigidez muscular natural dos membros inferiores, que impede a sua queda em situações de repouso. Para simular esta rigidez natural, foi adicionada uma mola com rigidez k_1 a uma altura h_1 do chão que corresponde a distância do chão ao centro de gravidade da perna, que gera um momento contrário ao causado pela força peso.

Além disso, o desafio principal da modelagem matemática é emular também a presença do andador. Para representar a força dos braços segurando no andador foi adicionada uma mola na mesma altura do CG com rigidez k_2. O andador por sua vez é representado como um bloco de altura h_a e massa m_a, além da força atuante da mola, o andador conta com a força F_m do motor que é a variável controlável do sistema que atua para garantir a estabilidade do usuário. O acoplamento entre o andador e o usuário é feito pelo deslocamento da mola k_2 ($x - l\sin(\theta)$).

Após a realização do Diagrama de Corpo Livre, torna-se possível equacionar o pêndulo invertido:

$$I_h\ddot{\theta} = m_h l g \sin(\theta) - k_2 l\cos(\theta)(l\sin(\theta) - x) - k_1 h_1^2 \tan(\theta) \tag{7}$$

Considerando que o pêndulo restringe-se a pequenos deslocamentos, é possível linearizar o comportamento dele, o que resulta em (8):

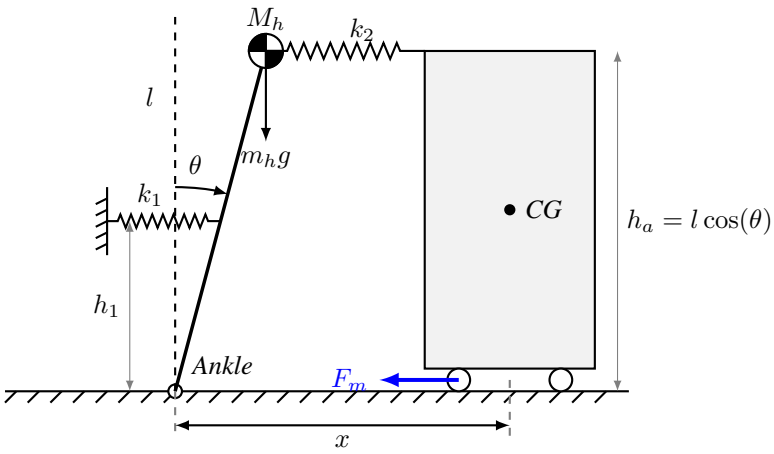


Fig. 1: Diagrama esquemático do modelo geométrico e de forças do sistema acoplado humano-andador.

$$\ddot{\theta} = \frac{m_h l g\theta - k_1 h_1^2\theta - k_2(l\theta - x)l}{I_h} \tag{8}$$

Quanto ao comportamento do andador utiliza-se a somatória de forças no mesmo, o teorema do movimento do baricentro.

$$m_a\ddot{x} = F_m + k_2(l\sin(\theta) - x) \tag{9}$$

Linearizando:

$$\ddot{x} = \frac{F_m + k_2(l\theta - x)}{m_a} \tag{10}$$

Contudo, é necessário impor uma distância mínima (d_{min}) entre o andador e o usuário, resultando em um termo adicional no acoplamento:

$$\ddot{\theta} = \frac{m_h l g\theta - k_1 h^2\theta - k_2(l\theta - x + d_{min})l}{I_h} \tag{11}$$

E o mesmo acontece para o andador:

$$\ddot{x} = \frac{F_m + k_2(l\theta - x + d_{min})}{m_a} \tag{12}$$

Embora a formulação de Funções de Barreira de Controle (CBF) possa ser aplicada diretamente a sistemas não-lineares, a linearização em torno da vertical de equilíbrio ($\theta \approx 0$) é adotada neste trabalho por duas razões fundamentais: primeiramente, como o envelope de segurança do dispositivo assistivo restringe a inclinação do usuário a pequenos ângulos ($\theta < 30°$), as aproximações de primeira ordem ($\sin\theta \approx \theta$ e $\cos\theta \approx 1$) introduzem desvios de modelagem desprezíveis na região de operação segura. Em segundo lugar, a representação linear no espaço de estados ($\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$) simplifica analiticamente o cálculo das derivadas de Lie sucessivas exigidas pela Função de Barreira de Controle Exponencial (ECBF) de grau relativo $r = 2$. Adicionalmente, esta abordagem garante que a restrição de segurança mapeada para o espaço de controle preserve a linearidade na variável de decisão u_{safe}.

Isso assegura a convexidade do problema de Programação Quadrática (QP) resolvido a cada passo de amostragem, viabilizando a implementação e a convergência do algoritmo de otimização em tempo real.

Para fazer a representação linear no espaço de estados, faz-se necessário substituir as variáveis do sistema ((13)).

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \\ x_3 &= \theta \\ x_4 &= \dot{\theta} \end{aligned} \tag{13}$$

Para simplificar a representação do sistema, definem-se os coeficientes dinâmicos γ e β como:

$$\gamma = \frac{m_h l g - k_1 h^2 - k_2 l^2}{I_h} \tag{14}$$

$$\beta = \frac{k_2 l}{I_h} \tag{15}$$

Sendo assim, o modelo no espaço de estados pode ser escrito de forma compacta como:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_2}{m_a} & 0 & \frac{k_2 l}{m_a} & 0 \\ 0 & 0 & 0 & 1 \\ \beta & 0 & \gamma & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m_a} \\ 0 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ \frac{k_2 d_{min}}{m_a} \\ 0 \\ -\beta d_{min} \end{bmatrix} \tag{16}$$

B. Controlador de Impedância

Para garantir que o andador consiga seguir os movimentos do usuário e manter-se próximo, mas ainda assim sendo capaz de adaptar-se a possíveis mudanças de rota, perturbações do ambiente etc, foi utilizado um controlador de impedância.

A dinâmica de impedância desejada para a inclinação do usuário/pêndulo (θ) impõe uma relação dinâmica entre a aceleração angular e o torque externo aplicado pela interação (T_{ext} ou $F_{\text{ext}} \cdot l$):

$$I_d\ddot{\theta} + B_d\dot{\theta} + K_d(\theta - \theta_0) = \tau_{ext} \ddot{\theta}_{des} = \frac{1}{I_d}\left(\tau_{ext} - B_d\dot{\theta} - K_d(\theta - \theta_0)\right) \tag{17}$$

Sendo:

- I_d a inércia desejada;
- B_d o amortecimento desejado;
- K_d a rigidez desejada;
- θ_0 o ângulo de equilíbrio de referência;
- τ_{ext} o torque externo exercido pelo usuário no sistema.

A partir da dinâmica do pêndulo linearizada, é possível isolar a posição desejada do andador (x_d) capaz de gerar o torque equivalente para essa impedância:

$$x_d = -\frac{1}{\beta}\left[\left(\gamma + \frac{K_d}{I_d}\right)\theta + \frac{B_d}{I_d}\dot{\theta} - \frac{\tau_{\text{ext}}}{I_d}\right] \tag{18}$$

A força de controle nominal aplicada pelos motores ($u_{\text{nom}} = F_m$) sintetiza a rigidez (K_{imp}) e amortecimento (B_{imp}) virtuais da interação linear, adicionando a compensação por gravidade do modelo invertido ($M_{\text{lcm},g}$):

$$\begin{aligned} x_{atual} &= x - l\theta - d_{min} \\ \dot{x}_{\text{atual}} &= v - l\dot{\theta} \end{aligned} \tag{19}$$

$$u_{\text{nom}} = -M_{\text{lcm},g} - K_{\text{d}}, x_{\text{atual}} - B_{\text{d}}, \dot{x}_{\text{atual}} \tag{20}$$

Onde $M_{\text{lcm},g} = m_h\, l\, g\, \theta$ representa a componente de compensação do momento gravitacional linearizado do usuário. Essa interação com o ambiente é feita através da criação de uma mola e um amortecedor virtual. A partir de uma abordagem analítica, entende-se que ao inclinar-se pra frente, o usuário comprime essa mola virtual, criando uma força que empurra o andador para frente. Em uma situação oposta, ao inclinar-se para trás, ele faz com que a mola se expanda, puxando o andador para perto do seu corpo. Dessa forma, é possível garantir que o andador esteja colado ao corpo do idoso, a todo momento (respeitando uma distância mínima imposta). Nota-se que o controlador de impedância gera um esforço de controle nominal (u_{nom}) capaz de manter o andador próximo do usuário a todo momento. No entanto, esse controlador não conta com a capacidade preditiva necessária para antecipar e previnir uma possível queda do usuário. Por isso, faz-se necessário também o uso de uma CBF como supervisor desse esforço de controle, visando garantir que o mesmo não coloque o usuário em uma zona de risco, fora do conjunto seguro.

C. Condições de Barreira

Diante das informações presentes em [7], é estabelecido um limite para a flexão de tronco do qual os idosos conseguem se recuperar de uma possível queda. Esse limite encontra-se em torno de 30 ° de inclinação. Sendo assim, a condição inicial de barreira é que o ângulo θ não pode passar desse limite de inclinação (Equação (21)).

$$\theta < \theta_{max} \tag{21}$$

Sendo assim, a barreira $h(x)$ pode ser descrita como:

$$h(x) = \theta_{max} - \theta - \alpha_c x \geq 0 \tag{22}$$

No qual o termo da Equação (22) $\alpha_c x$ introduzido na função de barreira tem o objetivo de incorporar o acoplamento entre a posição horizontal do sistema e o limite admissível de inclinação angular. Dessa forma, em vez de estabelecer um limite angular constante, a função de barreira define uma região segura dependente do estado de posição x. A constante α_c determina a intensidade na qual esse acoplamento impacta a condição de barreira $h(x)$.

Seguindo a condição imposta pela Equação (22):

$$L_f h(x) + \alpha(h(x)) \geq 0 \tag{23}$$

Onde o termo $L_f h(x)$ representa a derivada de Lie da função de barreira e $\alpha(h(x))$ um termo responsável por criar um gradiente de ação que antecipa a ação da barreira, fazendo com que ela atue antes de chegar próxima ao limite. O aumento do valor de α acarreta a intensidade da antecipação.

Seguindo a seguinte lógica:

$$\dot{h}(x) = -\dot{\theta} - \alpha_c \dot{x} \tag{24}$$

Dado que a variável de controle está presente na segunda derivada da posição x, seguindo o formalismo matemático, comprova-se através da Equação (25) que o sistema trata de uma ECBF de grau relativo $r = 2$.

$$\ddot{h}(x) = L_f^2 h + L_g L_h h u \tag{25}$$

Sendo assim, a equação de barreira ECBF com grau relativo 2 seguindo a condição imposta pela Equação (22) é igual a:

$$\ddot{h}(x) + K_1 \dot{h}(x) + K_2 h(x) \geq 0 \tag{26}$$

Expandindo os termos:

$$-\ddot{\theta} - \alpha_c \ddot{x} + K_1(-\dot{\theta} - \alpha_c \dot{x}) + K_2(\theta_{max} - \theta - \alpha_c x) \geq 0 \tag{27}$$

Para garantir que a resposta do sistema seja amortecida e evite sobressinal, utiliza-se a técnica de alocação de polos, alocando 2 polos iguais e negativos como as constantes (autovalores) do sistema.

$$(s + \lambda)^2 = s^2 + 2\lambda s + \lambda^2 \tag{28}$$

$$-\ddot{\theta} - \alpha_c \ddot{x} + 2\lambda(-\dot{\theta} - \alpha_c \dot{x}) + \lambda^2(\theta_{max} - \theta - \alpha_c x) \geq 0 \tag{29}$$

D. Otimização

Enquanto o controlador de impedância atua na malha nominal para garantir uma interação suave e complacente entre o usuário e o andador (rejeitando distúrbios externos e mantendo uma distância de conforto), o controlador baseado em Funções de Barreira de Controle (CBF) atua como um filtro supervisor de segurança. A função deste filtro é garantir, de forma estrita, a invariância do conjunto seguro, mitigando riscos de quedas.A fusão dessas duas estratégias de controle é realizada em tempo real por meio de um algoritmo de Programação Quadrática (QP). O objetivo do otimizador é encontrar a ação de controle final segura (u_{safe}) que minimize o desvio em relação ao sinal nominal gerado pela impedância (u_{nom}), submetendo o sistema às restrições cinemáticas e dinâmicas impostas pelas barreiras de segurança. Formalmente, o problema de otimização QP é descrito por:

$$u^* = \arg\min_{u \in \mathbb{R}} (u_{cbf} - u_{nom})^2 \qquad \text{sujeito a:} \quad A_{cbf} u \leq b_{cbf} \tag{30}$$

Aplicando a restrição estruturada pela Equação (29), as matrizes de restrição linear A_{cbf} e b_{cbf}, que delimitam matematicamente a fronteira de segurança contra quedas frontais, são dadas por:

$$\begin{aligned} A_{\text{cbf}} &= -g_h \\ b_{\text{cbf}} &= -f_h - 2\lambda(\dot{\theta} + \alpha_c \dot{x}) - \lambda^2(\theta - \theta_{max} + \alpha_c x) \end{aligned} \tag{31}$$

Os componentes f_h e g_h são derivados analiticamente a partir do modelo dinâmico acoplado do sistema, de acordo com as Equações (10) e (11). O termo f_h isola a contribuição da evolução natural dos estados da planta, ao passo que g_h representa o ganho de acoplamento da entrada, mapeando como o esforço de controle aplicado ao andador impacta diretamente na dinâmica angular do usuário:

$$\begin{aligned} f_h &= (\ddot{\theta} - \alpha_c \ddot{x}) \\ g_h &= \frac{\alpha_c}{m_a} \end{aligned} \tag{32}$$

A partir dessa formulação, o parâmetro estrutural α_c determina a autoridade e a magnitude do acoplamento do atuador na dinâmica de inclinação do usuário. Adicionalmente, a constante de projeto λ atua como ganho de convergência da Função de Barreira de Controle Exponencial (ECBF), ditando quão agressivamente o filtro de segurança deve intervir à medida que o usuário se aproxima do ângulo limite de inclinação (θ_{max}).

E. Diagrama de Blocos

Nesta seção apresenta-se o diagrama de blocos para ilustrar a lógica de controle. Nele é possível notar que a CBF atua como um supervisor, encontrando o equilíbrio entre o controlador nominal e o controlador CBF, gerando um sinal de controle u_{safe}.

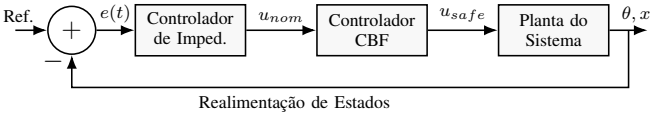


Fig. 2: Diagrama de blocos da malha de controle com filtro de segurança baseado em CBF.

F. Protocolo Experimental

Para testar o funcionamento da função de controle e sua eficácia, foi utilizado o *MATLAB/Simulink* no qual o controlador de Impedância calcula uma força nominal u_{nom} cuja função é manter o usuário próximo do andador. A CBF por sua vez calcula uma força segura u_{safe} que mantém o usuário dentro da região segura.

O intuito dessa simulação é introduzir uma situação crítica onde o controlador é obrigado a entrar em ação e evitar uma queda do usuário. Para isso, foi introduzido mais um elemento na equação (11), um distúrbio d. Para simular o empurrão, aplicou-se na simulação o distúrbio na forma de um pulso retangular com amplitude inicial de 100 N e duração de 0,3 segundos no instante t = 0.

Na Tabela I encontram-se os valores utilizados para os parâmetros do sistema e dos controladores. Foram utilizados valores hipotéticos, mas de magnitude coerente com aqueles encontrados na literatura.

TABELA I: Parâmetros dinâmicos e de controle do sistema acoplado humano-andador.

Categoria	**Parâmetro**	**Símbolo**	**Valor Nominal**
Dinâmica do Andador	Massa do andador	m_a	20
Dinâmica do Humano	Massa do usuário	m_h	70
	Comprimento do pêndulo	l	1.50
	Momento de inércia humano	I_h	157.50
Constantes Físicas	Rigidez da mola traseira	k_1	2500
	Rigidez da mola frontal	k_2	1000
	Altura da mola traseira	h_1	0.5
	Coeficiente de amortecimento	b	100
	Aceleração da gravidade	g	9.81
Controle Nominal	Constante de Rigidez Virtual	K_{imp}	10
	Constante de Amortecimento Virtual	B_{imp}	10
Filtro de Segurança (ECBF)	Fator de agressividade	α_c	0.8
	Ganho de convergência	λ	1.5

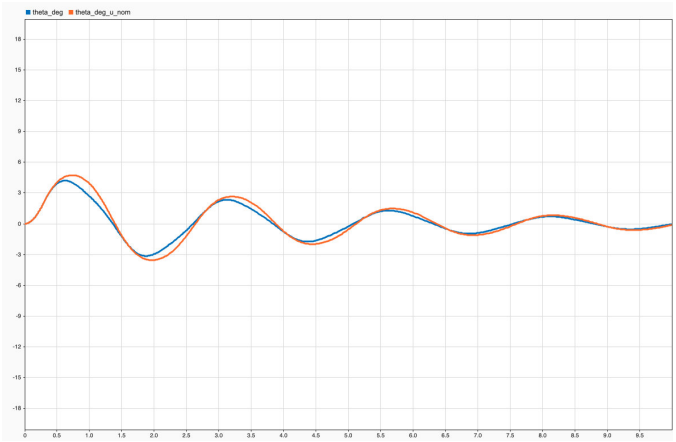


Fig. 3: Comportamento dinâmico do pêndulo em Graus sob força de 100 N, sendo a linha azul com CBF e a linha laranja sem CBF (Fonte: Autor, 2026)

G. Resultados Preliminares

Nesta seção serão apresentados os resultados obtidos nas simulações utilizando o *MATLAB/Simulink*.

A Figura 3 mostra o efeito do empurrão de 100 N no pêndulo apenas com o controlador de impedância (linha laranja) e com o controlador de Barreira (linha azul). Mesmo com um empurrão pequeno é possível notar que a CBF atua restringindo a inclinação do pêndulo, e diminuindo a amplitude do movimento. Essa ação fica evidente ao comparar o esforço de controle de u_{nom} e u_{safe} (Figura 4).

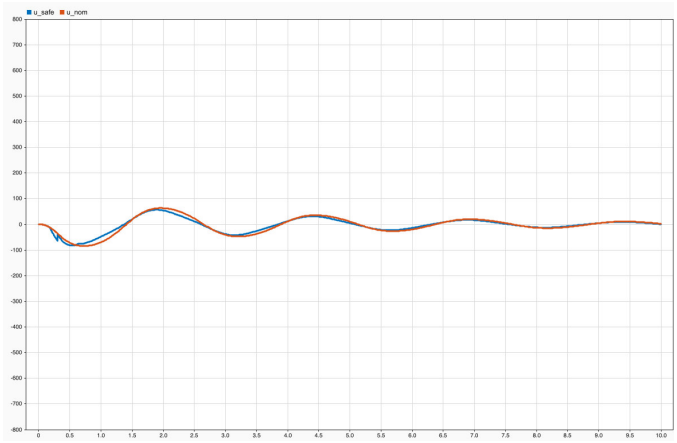


Fig. 4: Comportamento dos controladores sob esforço de 100 N sendo a linha azul com CBF e a linha laranja sem CBF (Fonte: Autor, 2026)

Em seguida, foram realizados testes com esforços maiores, para intensificar a ação do controlador de Barreira.

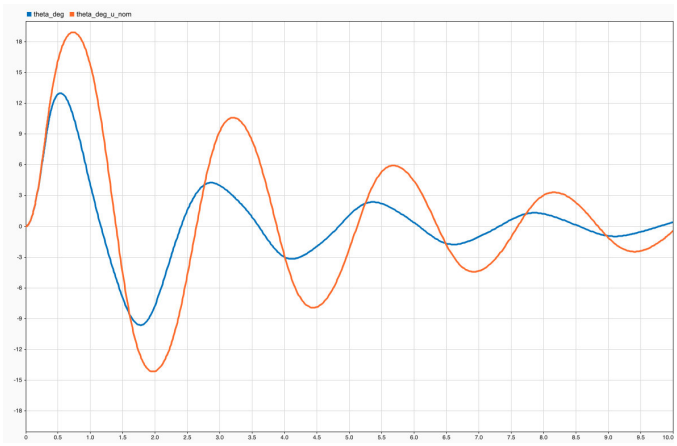


Fig. 5: Comportamento dinâmico do pêndulo em Graus sob força de 400 N, sendo a linha azul com CBF e a linha laranja sem CBF (Fonte: Autor, 2026)

A Figura 5 evidencia um esforço ainda maior do controlador de Barreira que impede que o pêndulo avance em direção ao limite do conjunto seguro. Enquanto no sistema sem a CBF o pêndulo chega próximo a 20° de amplitude, ao ativar a CBF, a mesma é capaz de limitar a amplitude próxima de 12°.

A divergência entre o controlador de Barreira e Controlador de Impedância fica ainda mais evidente com um empurrão de maior magnitude. Nota-se que há um pico de aceleração repentino no u_{safe}, isso acontece para impedir que haja uma colisão entre o usuário e o andador. Mas, mesmo com essa limitação que é imposta pela própria natureza do sistema, o mesmo é capaz de manter o usuário dentro do conjunto seguro.

É de suma importância também se atentar ao comportamento da função $h(t)$ que indica se a condição de segurança está sendo respeitada ou não.

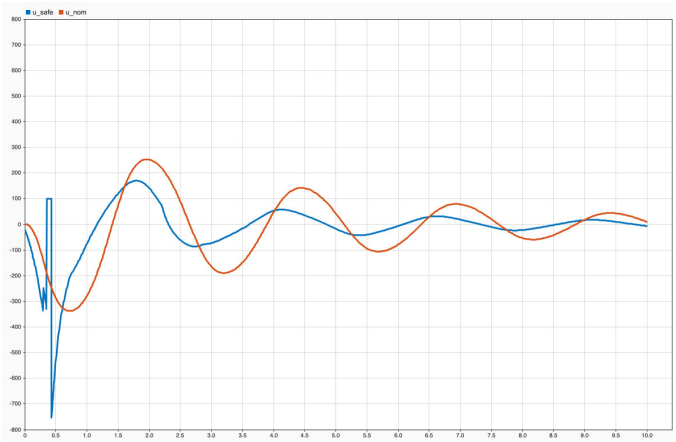


Fig. 6: Comportamento dos controladores sob esforço de 400 N sendo a linha azul com CBF e a linha laranja sem CBF (Fonte: Autor, 2026)

As curvas na Figura 7 comprovam que a CBF foi capaz de manter o sistema dentro do conjunto seguro, respeitando a condição imposta pela (22) em quase todos os esforços, com excessão do impulso de 500 N, em que a função $h(t)$ atinge valores negativos. É importante ressaltar que foi feita uma linha tracejada em azul com o objetivo facilitar a visualização do limite $h(t) = 0$.

IV. Conclusão

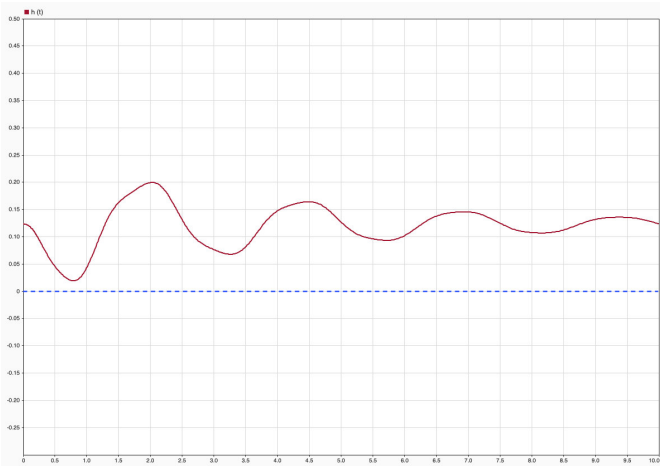
O presente trabalho apresentou como objetivo a elaboração de uma abordagem à segurança de usuários de dispositivos assistivos utilizando uma Função de Controle de Barreira otimizada com uma Lei de Controle de Impedância através de QP (Programação Quadrática. Foram feitas simulações computacionais utilizando um modelo biomecânico simplificado com o objetivo de compreender e avaliar o efeito do uso da CBF no sistema. Com auxílio da literatura foi estabelecido um limite de inclinação de $30°$ para o pêndulo invertido. Posteriormente, na Figura 5 fica evidente que a ação da CBF conteve a inclinação máxima em $12°$, abaixo fronteira crítica, prevenindo efetivamente uma queda irreversível.

Para trabalhos futuros tem-se em vista um refinamento do modelo biomecânico e das condições de barreira visando torná-los mais próximos da realidade, a realização de testes em um protótipo físico com o objetivo de identificar as principais limitações do dispositivo e como pode-se contorna-las.

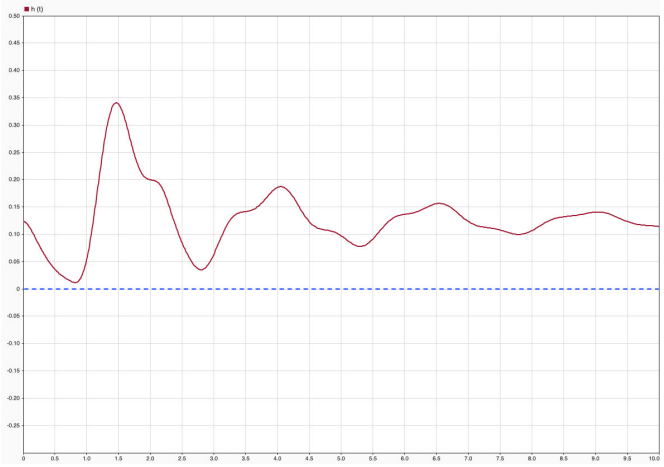
Por fim, a análise dos resultados sugerem que a estratégia proposta apresenta-se como uma alternativa promissora para garantir a segurança formal na robótica assistiva.

V. Declaração de Uso de IA

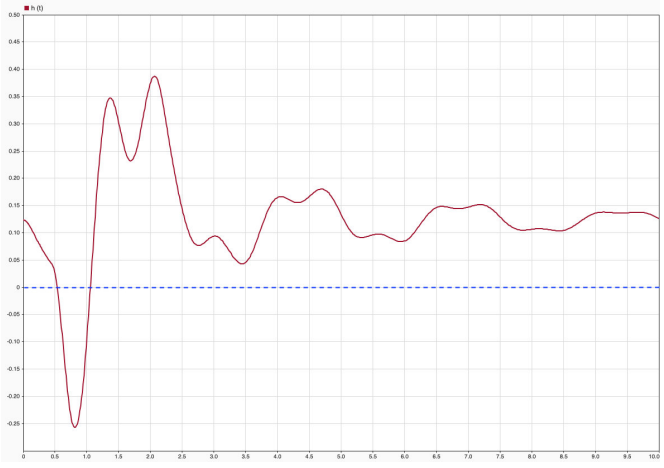
Durante a preparação deste trabalho, os autores utilizaram ferramentas assistidas por IA, como o Grammarly para redação e revisão gramatical, e o Turnitin para verificação de plágio. Após o uso dessas ferramentas, os autores revisaram e editaram o conteúdo conforme necessário e assumem total responsabilidade pelo conteúdo desta publicação.



(a) $F = 100$ N



(b) $F = 300$ N



(c) $F = 500$ N

Fig. 7: Curvas da função $h(t)$ para diferentes níveis de força (Fonte: Autor, 2026).

REFERÊNCIAS

[1] WORLD HEALTH ORGANIZATION. Population ageing: Questions and answers. 2024. Disponível em: https://www.who.int/news-room/questions-and-answers/item/population-ageing. Acesso em: 12 nov. 2025.

[2] MINETTO, Marco Alessandro; GIANNINI, Alessandro; MCCONNELL, Rebecca; BUSSO, Chiara; TORRE, Guglielmo; MASSAZZA, Giuseppe. Common musculoskeletal disorders in the elderly: the star triad. Journal of Clinical Medicine, v. 9, n. 4, p. 1216, 2020.

[3] WORLD HEALTH ORGANIZATION. Falls: Fact sheet. 2024. Disponível em: https://www.who.int/news-room/fact-sheets/detail/falls. Acesso em: 12 nov. 2025.

[4] ACKERMANN, Marko. Sistema de acionamento e travamento para flexão mecanizada de joelho em órteses de membros inferiores. 2002. Dissertação (Mestrado em Engenharia Mecânica) — Escola Politécnica, Universidade de São Paulo, São Paulo, 2002. Disponível em: https://www.teses.usp.br/teses/disponiveis/3/3132/tde-19012024-093042/. Acesso em: 4 mar. 2026. DOI: https://doi.org/10.11606/D.3.2002.tde-19012024-093042.

[5] VAUGHAN, Christopher L. Dynamics of human gait. 2. ed. Howard Place, Western Cape, South Africa: Kiboho Publishers, 1999. Acompanha CD-ROM GaitCD.

[6] SAAD, Marcelo; BATTISTELLA, Linamara Rizzo; MASIERO, Danilo. Técnicas de análise de marcha. Acta Fisiátrica, v. 3, n. 2, p. 23–26, 1996. Disponível em: https://www.revistas.usp.br/actafisiatrica/article/view/102015. Acesso em: 4 mar. 2026. DOI: https://doi.org/10.11606/issn.2317-0190.v3i2a102015.

[7] GRABINER, Mark D.; DONOVAN, Stephanie; BAREITHER, Mary Lou; MARONE, Jane R.; HAMSTRA-WRIGHT, Karrie; GATTS, Strawberry; TROY, Karen L. Trunk kinematics and fall risk of older adults: translating biomechanical results to the clinic. Journal of Electromyography and Kinesiology, v. 18, n. 2, p. 197–204, 2008. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S105064110700096X. Acesso em: 5 mar. 2026. DOI: https://doi.org/10.1016/j.jelekin.2007.06.009.

[8] GRABINER, Mark D.; FEUERBACH, Jon W.; JAHNIGEN, Dennis W. Measures of paraspinal muscle performance do not predict initial trunk kinematics after tripping. Journal of Biomechanics, v. 29, n. 6, p. 735–744, 1996. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/0021929095001425. Acesso em: 5 mar. 2026. DOI: https://doi.org/10.1016/0021-9290(95)00142-5.

[9] NAGANO, Hanatsu. Gait biomechanics for fall prevention among older adults. Applied Sciences, v. 12, n. 13, p. 6660, 2022. Disponível em: https://www.mdpi.com/2076-3417/12/13/6660. Acesso em: 6 mar. 2026. DOI: https://doi.org/10.3390/app12136660.

[10] CHINELATO, Caio I. G.; ANGÉLICO, Bruno A. Robust exponential control barrier functions for safety-critical control. In: AMERICAN CONTROL CONFERENCE (ACC), 2021. Proceedings [...]. Piscataway: IEEE, 2021. p. 2342–2347.

Mediolateral ground reaction force analysis during gait using discrete wavelet transform in individuals with knee dysfunction

1st Tiago da Silva Nogueira
Federal University of ABC
Santo André, Brazil

2nd Renato Naville Watanabe
Federal University of ABC
Santo André, Brazil

Abstract*—This study investigated differences in the mediolateral component of the ground reaction force (F_x) during gait in individuals with knee dysfunction using the Discrete Wavelet Transform (DWT). Data from a public gait database were analyzed, comparing healthy male participants with groups of individuals with left- or right-knee dysfunction matched 1:1 by body mass. Analyses were performed under the self-selected walking-speed condition. F_x signals were decomposed using a Daubechies wavelet (db4) at four levels, resulting in detail bands yhp1 to yhp4 and approximation band ylp4. For each band, the mean squared wavelet coefficient was calculated at the participant level. Between-group comparisons were performed using Welch's t-test, with Holm correction for 10 primary comparisons and effect-size estimation using Cohen's d with 95% confidence intervals. After adjustment for multiple comparisons, yhp3 (15.63–31.25 Hz) showed significantly lower values in the left-knee dysfunction group ($p_{Holm} = 0.0005$; $d = -0.566$) and the right-knee dysfunction group ($p_{Holm} < 0.0001$; $d = -0.704$). In the right-knee dysfunction group, a significant reduction was also observed in yhp4 (7.81–15.63 Hz; $p_{Holm} < 0.0001$; $d = -0.647$). No significant differences after Holm correction were identified in ylp4, yhp2, or yhp1. The results show that wavelet decomposition can identify scale-specific differences in the F_x signal associated with knee dysfunction. The consistent alteration in yhp3 on both sides indicates that this band constitutes a candidate quantitative feature for future investigation, although the current results do not establish diagnostic performance or allow the observed differences to be attributed to specific physiological mechanisms.

***Index Terms*—Discrete wavelet transform; gait analysis; knee dysfunction; ground reaction force; biomechanics.**

I. Introduction

Human gait is a complex motor task that depends on coordinated interaction between the musculoskeletal and neuromuscular systems. Instrumented gait analysis enables quantitative characterization of changes in locomotor patterns and is an important tool for assessing individuals with musculoskeletal impairments [1], [2]. Among the biomechanical variables used in this context, ground reaction forces (GRFs) represent the resultant mechanical interaction between the individual and the support surface during locomotion [2], [4].

The knee joint plays an important role during gait, contributing to load absorption and transmission, stability during the stance phase, and lower-limb progression [1], [4]. Structural or functional alterations of the knee can modify the locomotor pattern and, consequently, the characteristics of the forces exerted against the ground [3], [4]. Studies in populations with knee osteoarthritis and pain have also demonstrated differences in GRF patterns, including the mediolateral component, reinforcing the relevance of analyzing the complete force curve [10].

Conventional analysis of GRF curves often uses time-domain parameters such as peaks, impulse, and temporal variables. Although useful, these measures summarize global aspects of the signal and do not directly describe its combined variation over time and frequency [2], [5]. In nonstationary biomechanical signals, time-frequency approaches can provide complementary information [5], [11].

Among these approaches, the Wavelet Transform allows the signal to be decomposed into different scales with temporal and spectral localization. In its discrete form (DWT), successive filtering and downsampling stages produce detail and approximation coefficients associated with different frequency ranges [5]. This approach has been applied to gait data and GRFs, including total knee arthroplasty and classification of pathological gait patterns [7]– [9], [11].

Publicly available instrumented gait-analysis databases have expanded the possibility of investigating these characteristics in larger clinical samples. GaitRec contains ground reaction force records from healthy participants and individuals with different musculoskeletal impairments, together with clinical metadata [6]. The database provides raw and processed signals and allows secondary analyses of different GRF components [6].

Nevertheless, it remains relevant to investigate whether wavelet decomposition of the mediolateral component identifies signal characteristics associated with knee dysfunction. By examining different scales separately, the DWT may complement conventional descriptors without assuming a direct correspondence with specific physiological mechanisms.

Accordingly, the aim of this study was to investigate differences in the mediolateral component between healthy participants and individuals with unilateral knee dysfunction during gait at self-selected speed, using the DWT to characterize the signal across different frequency scales. Comparisons were performed separately for the left and right sides, evaluating the Mean Squared Wavelet Coefficient (MSWC) in each band.

Given the exploratory nature of the analysis, we hypothe-

sized that the distribution of this measure across the different wavelet bands would differ between healthy participants and individuals with knee dysfunction, without prespecifying which band should be altered or attributing such differences to a specific physiological mechanism.

II. Methodology

A. Data source, population, and definition of the experimental unit

Data were obtained from the publicly available GaitRec database, which contains bilateral GRF recordings from more than 2,000 patients with different musculoskeletal impairments and 211 healthy controls, together with clinical and demographic metadata [6]. Because men were more highly represented among both healthy participants and those with musculoskeletal impairment, the present analysis was restricted to male participants.

Eligible participants were men with gait recordings at self-selected speed and from the initial session (SPEED = 2; SESSION_TYPE = 1). Among those with knee dysfunction, 210 had left-side involvement, 160 right-side involvement, and 5 bilateral involvement. Bilateral cases were excluded because comparisons were defined by side.

The experimental unit was defined as the unique participant, rather than the limb or each isolated gait trial. Although the same participant could have multiple valid trials, these repeated trials were initially processed separately and subsequently aggregated into a single observation per participant and side. Thus, each individual contributed only once to each primary statistical comparison.

Participants were classified according to the presence and laterality of knee dysfunction based on the clinical information and annotations available in GaitRec [6]. According to Horsak et al. [6], before gait assessment each participant underwent a physical examination performed by a physician. The database annotation was subsequently performed manually by an experienced physiotherapist based on the available medical diagnosis. The knee class (K_xxx) includes, among other conditions, fractures of the patella, femur, or tibia near the knee, ligament or meniscal tears, and total knee arthroplasty. Laterality was determined from the AFFECTED_SIDE metadata field, coded as 0 for left, 1 for right, and 2 for bilateral. Participants were therefore assigned to the left- or right-knee dysfunction groups according to the laterality recorded in the database, and bilateral cases were excluded from the present analysis. The GaitRec publication also reports that no information is available on the history of similar or other types of musculoskeletal injuries for either patients or healthy controls; this limitation of the database should be considered when interpreting the results.

B. Walking condition and trial selection

The self-selected walking-speed condition (SPEED = 2) from the initial session (SESSION_TYPE = 1) was analyzed [6]. According to the GaitRec description, healthy controls had a mean self-selected walking speed of 1.27 ± 0.13 m/s.

Metadata were linked to the raw-signal tables using participant and session identifiers, retaining only trials that simultaneously met both criteria.

C. Body-mass matching

One-to-one nearest-neighbor matching (NNM) used body mass as the only variable, without replacement and without a caliper.

A total of 104 eligible healthy controls were identified, all of whom were included as the reference group. Matching was performed separately for the left- and right-side comparisons, always against the same set of controls. To define the sequence of the algorithm, controls were ordered according to the proximity of their body mass to the median body mass of the dysfunction candidates; then, for each control, the remaining candidate with the smallest absolute body-mass difference was selected. Each final comparison included 104 controls and 104 participants with dysfunction. After matching, 106 candidates in the left group and 56 in the right group remained unmatched, in addition to the 5 bilateral cases previously excluded.

Post-matching body-mass balance was assessed using Welch's t-test and Cohen's d effect size. Balance results are presented in Table I.

D. Biomechanical signal and preprocessing

The biomechanical variable of interest was the mediolateral component (F_x). GaitRec reports that the signals were originally digitized at 2,000 Hz; however, the raw signals made available were previously downsampled to 250 Hz as part of the rehabilitation center's internal procedures [6]. The present analysis used these RAW signals at 250 Hz, preserving the original temporal length of each trial.

Missing values (NaN) located at the ends of the series were interpreted as structural table padding and removed. Any internal gaps would have been linearly interpolated; however, none of the trials included in the analysis contained internal gaps. Trials without a valid signal or with insufficient length for four levels of decomposition would have been excluded.

No additional digital filter was applied to the mediolateral signal before the DWT. This distinction is important because GaitRec provides both RAW and processed (PRO) signals: the PRO data were filtered by the database authors using a second-order low-pass Butterworth filter with a cutoff frequency of 20 Hz, temporally normalized to 101 points, and normalized by body mass; these procedures were not reapplied to the RAW signals used in the present analysis [6].

E. Discrete Wavelet Transform

Time-frequency analysis was performed separately on each original trial, without prior temporal normalization, using the DWT with a fourth-order Daubechies wavelet (db4), four decomposition levels, and symmetric boundary extension (symmetric). This procedure mirrors values at the signal edges during wavelet filtering, reducing artificial discontinuities associated with edge processing. Given the sampling

frequency of 250 Hz and a Nyquist frequency of 125 Hz, the decomposition produced four detail coefficients and one approximation coefficient. The frequency ranges reported for each level correspond to the approximate theoretical bands of the decomposition: yhp1, 62.50–125.00 Hz; yhp2, 31.25–62.50 Hz; yhp3, 15.63–31.25 Hz; yhp4, 7.81–15.63 Hz; and ylp4, 0–7.81 Hz. These bands were interpreted as signal scales and were not directly attributed to specific physiological events of the gait cycle, because the study did not include complementary kinematic or neuromuscular measures capable of supporting such correspondence.

F. Mean squared wavelet coefficients (MSWC)

For each trial and wavelet band, the mean squared wavelet coefficient (MSWC) was calculated. For a band j containing N_j coefficients $c_{j,k}$, the measure was defined as:

$$E_j = \frac{1}{N_j} \sum_{k=1}^{N_j} c_{j,k}^2 \tag{1}$$

This definition corresponds to the mean of the squared wavelet coefficients and reduces the influence of the number of coefficients produced at each decomposition level. The measure does not correspond to Relative Wavelet Energy, which is defined as the ratio between the energy of a band and the total signal energy. After calculating the MSWC for each trial, the values obtained from repeated trials for each participant were averaged separately for each band, resulting in a single measure per participant, side, and band.

G. Temporal normalization for graphical representation

Temporal normalization was used exclusively for the construction of Figs. 1 and 2 and was not part of the MSWC calculation or the statistical analyses. For Fig. 1, each trial was resampled to 101 points, corresponding to 0%–100% of the gait cycle. Repeated trials from each participant were then aggregated to obtain an individual curve, and these curves were used to calculate the group mean curves. The shaded regions in Fig. 1 represent the mean and standard deviation across participants.

For Fig. 2, the group mean curves obtained from the normalized signals were subjected to the db4 DWT solely to visually illustrate the distribution of coefficients across the five wavelet subbands.

These temporal-normalization and group-mean decomposition procedures were used only for graphical representation. For the quantitative analysis, the DWT was applied separately to the original trials, without temporal normalization, preserving the length of each signal. The resulting coefficients were used to calculate the MSWC, which was subsequently used in the statistical comparisons.

H. Statistical comparisons and multiple-testing control

The primary comparisons were performed separately for each side. The left limb of healthy participants was compared with the left-knee dysfunction group, whereas the right limb of healthy participants was compared with the right-knee dysfunction group. Thus, each analysis used the same-side limb of healthy controls as the reference, without averaging the two limbs. For each of the five wavelet subbands, participant-level MSWC values were compared between the healthy group and the knee-dysfunction group using Welch's t-test. Although nearest-neighbor matching associated participants with similar body masses, this procedure was used to balance the groups and does not, by itself, establish a natural dependence between observations. Therefore, the primary analysis treated the groups as independent samples. For each comparison, the mean, standard deviation, mean between-group difference (defined as dysfunction minus healthy), and its 95% confidence interval (95% CI) were calculated. Effect size was estimated using Cohen's d, calculated from the pooled standard deviation. The 95% CIs for Cohen's d were obtained by bootstrap with 5,000 resamples with replacement and a fixed random seed to ensure reproducibility.

Because five subbands were evaluated for each side, totaling 10 primary comparisons, p values were jointly adjusted using the Holm procedure to control the family-wise error rate arising from multiple tests. A comparison was considered statistically significant when the Holm-adjusted p value was < 0.05. Raw p values were also retained and are presented in Table II.

As a sensitivity analysis, the same 10 comparisons were repeated using the paired t-test, preserving the correspondence established by the matching procedure and again applying Holm correction. This analysis was performed to determine whether the main conclusions were maintained when the matched-pair structure was accounted for statistically, without replacing the primary Welch analysis.

The analytical pipeline was implemented in Python, using NumPy and pandas for data manipulation, SciPy for statistical tests, and PyWavelets for DWT decomposition.

TABLE I
BODY-MASS BALANCE AFTER MATCHING

Comparison	Group	N	Body mass, mean ± SD (kg)	p value	Cohen's d
Healthy × Left knee	Healthy	104	82.34 ± 14.58	0.8578	-0.025
	Left knee	104	81.98 ± 14.64		
Healthy × Right knee	Healthy	104	82.34 ± 14.58	0.8012	0.035
	Right knee	104	82.85 ± 14.13		

III. RESULTS

A. Participant matching

After body-mass matching, each comparison included 104 healthy participants and 104 participants with knee dysfunction.

Body mass was similar between groups for both the left-side comparison (82.34 ± 14.58 vs. 81.98 ± 14.64 kg; $p = 0.8578$; $d = -0.025$) and the right-side comparison (82.34 ± 14.58

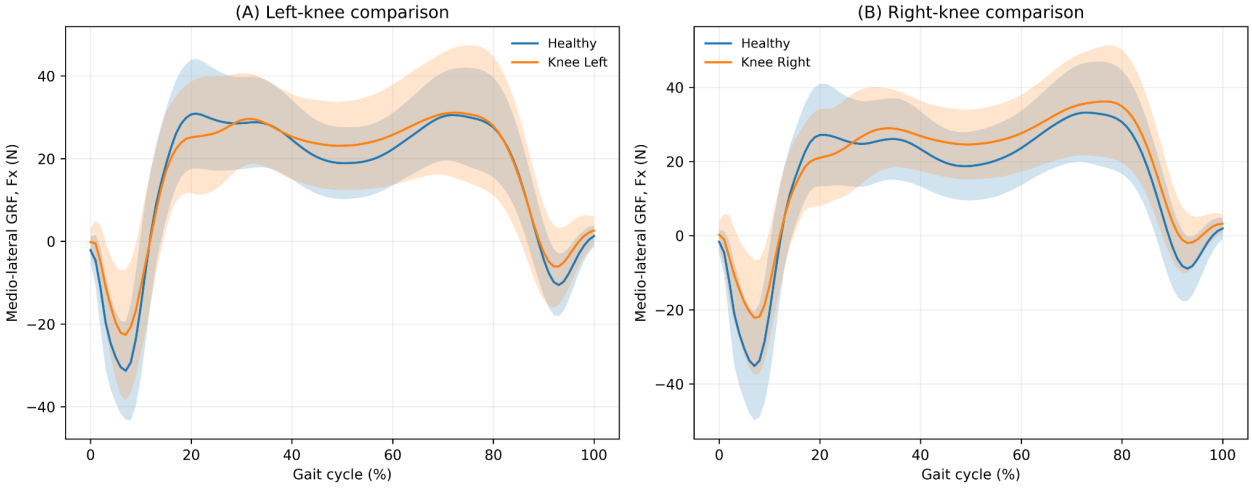


Fig. 1. Mean morphology of the mediolateral ground reaction force (F_x) across the gait cycle for (A) the left-knee comparison and (B) the right-knee comparison. Trials were temporally normalized to 101 points exclusively for graphical representation. Lines indicate group means, and shaded regions represent mean ± SD across participants. Temporal normalization was not used in the MSWC calculation or statistical analyses.

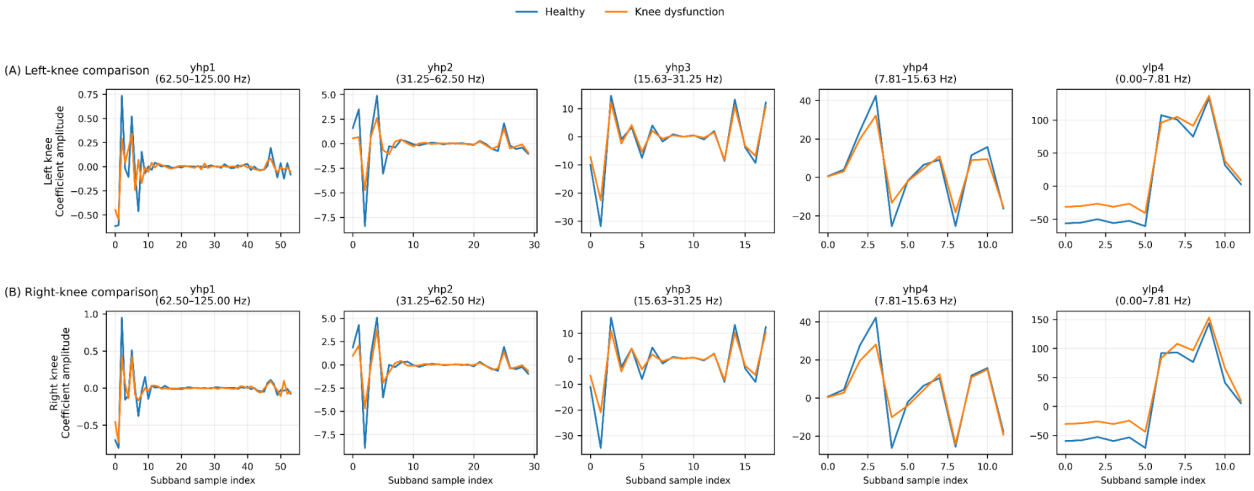


Fig. 2. Illustrative db4 DWT decomposition of the temporally normalized mean F_x curves for (A) the left-knee comparison and (B) the right-knee comparison. Panels show the detail coefficients yhp1 to yhp4 and the approximation coefficient ylp4 as a function of the sample index within each subband. This graphical decomposition was not used in the MSWC calculation or statistical inference, which were performed on the original trials without temporal normalization.

TABLE II
PARTICIPANT-LEVEL MSWC STATISTICS BY SIDE AND SUBBAND

Side	Band	Hz range	Healthy MSWC	Dysfunction MSWC	Mean diff. [95% CI]	p (raw / Holm)	Cohen's d [95% CI]
Left	ylp4	0.00–7.81	7503.17 ± 4203.88	7682.89 ± 6029.98	$179.72\ [-1242.37, 1601.82]$	0.8034 / 1.0000	$0.035\ [-0.252, 0.282]$
	yhp4	7.81–15.63	335.28 ± 200.60	264.04 ± 202.22	$-71.25\ [-126.31, -16.18]$	0.0115 / 0.0803	$-0.354\ [-0.633, -0.080]$
	yhp3	15.63–31.25	58.66 ± 43.38	36.40 ± 34.74	$-22.26\ [-33.01, -11.51]$	<0.0001 / **0.0005**	$-0.566\ [-0.862, -0.293]$
	yhp2	31.25–62.50	4.29 ± 4.06	3.34 ± 3.71	$-0.94\ [-2.01, 0.12]$	0.0821 / 0.4103	$-0.242\ [-0.527, 0.018]$
	yhp1	62.50–125.00	0.21 ± 0.18	0.17 ± 0.16	$-0.05\ [-0.09, 0.00]$	0.0505 / 0.3030	$-0.273\ [-0.541, -0.008]$
Right	ylp4	0.00–7.81	7862.47 ± 5069.39	8294.81 ± 5660.90	$432.34\ [-1036.85, 1901.53]$	0.5624 / 1.0000	$0.080\ [-0.188, 0.342]$
	yhp4	7.81–15.63	405.72 ± 249.42	254.83 ± 216.10	$-150.89\ [-214.70, -87.08]$	<0.0001 / $<$**0.0001**	$-0.647\ [-0.980, -0.346]$
	yhp3	15.63–31.25	70.67 ± 54.82	37.70 ± 37.11	$-32.97\ [-45.78, -20.16]$	<0.0001 / $<$**0.0001**	$-0.704\ [-0.987, -0.444]$
	yhp2	31.25–62.50	4.78 ± 5.00	3.69 ± 4.07	$-1.09\ [-2.34, 0.15]$	0.0853 / 0.4103	$-0.240\ [-0.510, 0.022]$
	yhp1	62.50–125.00	0.20 ± 0.15	0.20 ± 0.22	$0.00\ [-0.05, 0.05]$	0.9629 / 1.0000	$0.006\ [-0.309, 0.249]$

Note: Frequencies in Hz. The 95% CIs reported for the mean difference and Cohen's d were not adjusted for multiple comparisons. Statistical significance of the primary comparisons was determined using Holm-adjusted p values.

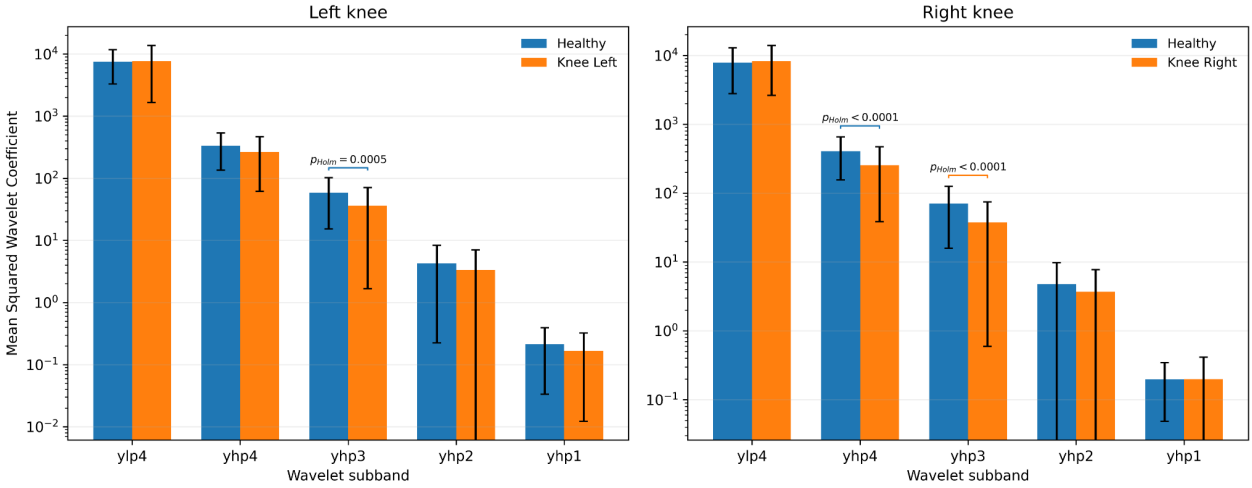


Fig. 3. Mean Squared Wavelet Coefficient (MSWC; mean ± SD) at the participant level for the five DWT subbands in the left- and right-knee comparisons. Statistically significant differences are identified by Holm-adjusted p values, considering the 10 primary comparisons. The y-axis is shown on a logarithmic scale because MSWC values span several orders of magnitude.

vs. 82.85 ± 14.13 kg; $p = 0.8012$; $d = 0.035$), indicating negligible differences after matching (Table I).

B. Graphical representation and wavelet decomposition of the mediolateral signal

Fig. 1 presents the mean morphology of the F_x component across the gait cycle after temporal normalization used exclusively for visualization. Fig. 2 illustrates the db4 DWT decomposition of the normalized mean curves into the five subbands, whereas Fig. 3 presents participant-level MSWC values. As described in the methodology, the temporal normalization and the decomposition shown in Fig. 2 are for graphical purposes; the MSWC calculation and statistical tests were based on the DWT applied to the original trials without temporal normalization. Complete quantitative results are provided in Table II.

C. Left-knee comparison

In the yhp3 band (15.63–31.25 Hz), the mean squared wavelet coefficient was 58.66 ± 43.38 in the healthy group and 36.40 ± 34.74 in the left-knee dysfunction group. The mean difference, defined as dysfunction minus healthy, was -22.26, with a 95% CI of -33.01 to -11.51. The difference had a raw p value < 0.0001 and remained statistically significant after Holm correction for the 10 primary comparisons ($p_{Holm} = 0.0005$). The effect size was moderate ($d = -0.566$; 95% CI: -0.862 to -0.293). The negative sign indicates lower values in the knee-dysfunction group.

In yhp4, the difference observed before adjustment ($p = 0.0115$; $d = -0.354$) did not remain significant after Holm correction ($p_{Holm} = 0.0803$). The ylp4, yhp2, and yhp1 bands also showed no significant differences after adjustment.

D. Right-knee comparison

In the yhp3 band (15.63–31.25 Hz), the mean squared wavelet coefficient was 70.67 ± 54.82 in the healthy group and 37.70 ± 37.11 in the right-knee dysfunction group. The mean difference was -32.97, with a 95% CI of -45.78 to -20.16. This difference had $p_{Holm} < 0.0001$ and an effect size of $d = -0.704$ (95% CI: -0.987 to -0.444).

In the yhp4 band (7.81–15.63 Hz), the values were 405.72 ± 249.42 for healthy participants and 254.83 ± 216.10 for those with right-knee dysfunction. The mean difference was -150.89, with a 95% CI of -214.70 to -87.08. The difference remained significant after Holm adjustment ($p_{Holm} < 0.0001$), with an effect size of $d = -0.647$ (95% CI: -0.980 to -0.346). No statistically significant differences after Holm correction were observed in ylp4, yhp2, or yhp1.

In the paired sensitivity analysis, yhp3 remained significant on the left ($p_{Holm} = 0.0023$) and right ($p_{Holm} < 0.0001$), as did yhp4 on the right ($p_{Holm} < 0.0001$). For yhp4 on the left, the result differed from the primary analysis ($p_{Holm} = 0.0429$ in the paired test vs. 0.0803 in Welch's test), indicating that this inference was sensitive to the statistical treatment of the matched pairs.

IV. Discussion

The analysis revealed a selective pattern of between-group differences, concentrated primarily in yhp3 and, on the right side, also in yhp4. The presence of moderate effects in these comparisons and the persistence of the bilateral yhp3 finding across both inferential strategies indicate that the observed differences were not limited to statistical significance but were concentrated at specific signal scales.

A. Multiscale pattern associated with knee dysfunction

yhp3 showed the most consistent pattern: MSWC values were lower in the dysfunction groups on both sides, with moderate effects ($d = -0.566$ on the left and $d = -0.704$ on the right) and 95% CIs that did not include zero. This behavior suggests that the information capable of differentiating the

groups, under the analyzed conditions, was concentrated at this decomposition scale.

On the right side, yhp4 also showed a reduction with a moderate effect ($d = -0.647$), extending the pattern to the immediately lower scale. On the left, however, yhp4 did not remain significant in the primary analysis after Holm correction, although it reached the significance threshold in the paired test.

These findings should be interpreted in the signal domain, without direct attribution to physiological mechanisms. The study did not include kinematics, electromyography, or other measures capable of establishing the biomechanical origin of the observed differences.

B. Contribution and limitations of wavelet analysis

The absence of differences after Holm correction in ylp4, yhp2, and yhp1 reinforces the selective nature of the observed pattern but does not allow the conclusion that the biomechanical phenomena represented at these scales are preserved.

The main contribution of the DWT in this context was to enable separate analysis of signal components that would otherwise be condensed into global descriptors such as peaks, impulse, or temporal parameters. The use of wavelets in gait and ground reaction force data has precedent in the literature [7]– [9], and the present results support its use as an exploratory tool for multiscale characterization.

However, the study did not evaluate diagnostic or classification performance; therefore, yhp3 should be considered a candidate quantitative signal feature rather than a diagnostic biomarker.

C. Limitations

Several limitations should be considered. This is a retrospective analysis of a public database, with no control by the authors over participant selection, clinical procedures, or original signal acquisition.

The GaitRec "knee dysfunction" category is clinically heterogeneous, including fractures, ligament or meniscal injuries, and arthroplasty [6]; therefore, the results should not be generalized to a specific pathology.

The analysis was restricted to men and to gait at self-selected speed, which limits generalization to women and other walking speeds.

Matching considered body mass only; age, height, diagnosis, severity, and other clinical characteristics remained potential confounding factors.

In addition, the analysis included only the mediolateral component of the ground reaction force, without complementary kinematic, electromyographic, or clinical measures to support mechanistic interpretations.

D. Implications and future studies

The consistency of yhp3 supports its investigation in independent and clinically more homogeneous samples.

Future studies should stratify participants by diagnosis and clinical severity, include both sexes, control additional confounding factors, and integrate GRF with kinematics, electromyography, and other biomechanical measures.

Any eventual clinical application of these features will depend on external validation and formal evaluation of classification performance.

V. Conclusion

Multiscale analysis of the mediolateral ground reaction force component using the Discrete Wavelet Transform identified differences between healthy participants and individuals with knee dysfunction, concentrated primarily in yhp3. The bilateral consistency of this finding, with moderate effects and persistence across both statistical approaches, supports its investigation as a candidate quantitative signal feature.

The results also show that the Discrete Wavelet Transform can complement global gait descriptors by localizing differences at specific signal scales. However, the findings remain limited to group-level differentiation and do not establish physiological mechanisms or diagnostic performance.

Validation in independent, clinically more homogeneous samples accompanied by other biomechanical measures will be necessary to determine the functional relevance and potential applicability of these features.

Ethical Considerations

This study used a public, anonymized clinical database [6]. The original experimental procedures, data collection, and publication of the GaitRec database were conducted in accordance with ethical standards and approved by the local Ethics Committee of Lower Austria (Approval No. GS1-EK-4/299-2014) [6].

References

[1] J. Perry and J. M. Burnfield, *Gait Analysis: Normal and Pathological Function*, 2nd ed. Thorofare, NJ, USA: SLACK Incorporated, 2010.

[2] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 4th ed. Hoboken, NJ, USA: Wiley, 2009.

[3] T. K. Uchida and S. L. Delp, *Biomechanics of Movement: The Science of Sports, Robotics, and Rehabilitation*. Cambridge, MA, USA: MIT Press, 2021.

[4] C. Kirtley, *Clinical Gait Analysis: Theory and Practice*. Edinburgh, U.K.: Elsevier Churchill Livingstone, 2006.

[5] J. L. Semmlow and B. Griffel, *Biosignal and Medical Image Processing*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2014.

[6] B. Horsak et al., "GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait," *Sci. Data*, vol. 7, Art. no. 143, 2020, doi: 10.1038/s41597-020-0481-z.

[7] T. Chau, "A review of analytical techniques for gait data. Part 2: Neural network and wavelet methods," *Gait Posture*, vol. 13, no. 2, pp. 102–120, 2001, doi: 10.1016/S0966-6362(00)00095-3.

[8] F. Verdini, T. Leo, S. Fioretti, M. G. Benedetti, F. Catani, and S. Giannini, "Analysis of ground reaction forces by means of wavelet transform," *Clin. Biomech.*, vol. 15, no. 8, pp. 607–610, 2000, doi: 10.1016/S0268-0033(00)00019-X.

[9] M. Alaqtash, T. Sarkodie-Gyan, H. Yu, O. Fuentes, R. Brower, and A. Abdelgawad, "Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 453–457, doi: 10.1109/IEMBS.2011.6090063.

[10] K. E. Costello et al., “Ground reaction force patterns in knees with and without radiographic osteoarthritis and pain: descriptive analyses of a large cohort (the Multicenter Osteoarthritis Study),” *Osteoarthritis Cartilage*, vol. 29, no. 8, pp. 1138–1146, 2021, doi: 10.1016/j.joca.2021.03.009.

[11] J. Issartel, L. Marin, P. Gaillot, T. Bardainne, and M. Cadopi, “A practical guide to time-frequency analysis in the study of human motor behavior: the contribution of wavelet transform,” *J. Mot. Behav.*, vol. 38, no. 2, pp. 139–159, 2006, doi: 10.3200/JMBR.38.2.139-159.

CardioExergy: A Computational Tool for Exergy Analysis of the Human Heart Under Different Levels of Aortic Stenosis

Maria Luíza Zitei Brandão
Technological Research Center
University of Mogi das Cruzes
Mogi das Cruzes, Brasil
maluzitei@gmail.com

Matheus Leonardo Alves de Camargo
Technological Research Center
University of Mogi das Cruzes
Mogi das Cruzes
matheuscamargo@umc.br

Alessandro Pereira da Silva
Technological Research Center
University of Mogi das Cruzes
Mogi das Cruzes, Brasil
alessandrops@umc.br

Robson Rodrigues da Silva
Technological Research Center
University of Mogi das Cruzes
Mogi das Cruzes, Brasil
robson.silva@umc.br

***Abstract*— Aortic stenosis is one of the most common valvular diseases in the adult population and is associated with significant hemodynamic and energetic changes in the cardiovascular system. In this context, exergy analysis emerges as a promising approach to quantify the irreversibilities involved in cardiac function. The aim of this study was to develop and validate CardioExergy, a computational platform designed for the exergy analysis of the human heart under varying degrees of aortic stenosis. The web-based platform integrates mathematical models of ventricular mechanics, circulatory dynamics, and exergy balance; it is implemented in Python, with interfaces developed via Google AI Studio. The tool allows for the simulation of different aortic valve areas (AVA) and the calculation of variables such as transvalvular pressure gradient, cardiac work, energy metabolism, exergy destruction, and exergy efficiency. Validation was performed by comparing the obtained results with data previously published in the literature. The simulations adequately reproduced expected physiological behavior, demonstrating increased exergy destruction and reduced cardiac efficiency as the stenosis progressed. The results indicate that CardioExergy is a robust tool for *in silico* studies of cardiovascular function, capable of supporting research in cardiac physiology, bioengineering, and the energetic analysis of the cardiovascular system.**

Keywords*— *Exergy Analysis, Aortic Stenosis, Computational Modeling, Cardiovascular System, Biomedical Engineering

I. Introduction (*Heading 1*)

Aortic stenosis is a common valvular heart disease, particularly among the elderly, characterized by a reduction in valve area and the consequent difficulty in ejecting blood from the left ventricle [1]. This condition increases the transvalvular pressure gradient, imposes a cardiac workload, and intensifies energy losses within the cardiovascular system, progressively compromising cardiac performance and efficiency. Assessing the effects of aortic stenosis requires approaches that consider not only hemodynamic aspects but also energetic ones [2]. In this context, exergy analysis enables the investigation of the heart's ability to convert metabolic energy into useful work, providing quantitative indicators regarding cardiac efficiency and energy losses associated with structural and functional changes in the cardiovascular system [3]. Exergy represents the maximum useful work obtainable from a system in equilibrium with its environment and serves as a valuable tool for analyzing biological systems [4]. Unlike conventional energy analysis, it evaluates energy quality and quantifies process irreversibilities, allowing for a more comprehensive understanding of energy losses and the efficiency of cardiac function [5]. Cardiac exergy analysis remains limited by a scarcity of models integrating relevant physiological variables and a lack of accessible computational tools [6]. To address this need, CardioExergy was developed—a web-based platform that combines hemodynamic, metabolic, and thermodynamic aspects to simulate varying degrees of aortic stenosis. The platform enables the estimation of indicators such as cardiac output, ventricular work, destroyed exergy, and exergy efficiency, thereby contributing to the understanding of cardiac energy performance under physiological and pathological conditions and across different aortic valve areas (AVA).

II. Materials and Methods

A. Mathematical Model

Although the present model is based on the exergetic approach proposed by Roll et al. (2019) [2], certain adaptations were made to extend its applicability and analytical capabilities. Unlike the original model, this study incorporates heart rate as a control variable, enabling an assessment of cardiovascular system behavior under varying physiological demands. Furthermore, the valvular discharge coefficient (Cd) was introduced; this parameter represents the actual efficiency of blood flow through the aortic valve and allows for the investigation of how valve geometry impacts energy losses. The model also encompasses thermal analyses and various simulation protocols, including scenarios of cardiac stress as well as pre- and post-intervention conditions.

Basal metabolism (1), M_{Heart}, was determined according to the model proposed by Suga [7]:

$$M_{Heart} = a \cdot W_{Heart} + b \quad (1)$$

Where a and b are model parameters and W_{Heart} is the work performed by the heart.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The exergy destroyed (2) at the aortic valve, B_{d_aortic}, is the stenosis power associated with the heartbeat (W_{sten}) multiplied by the ratio of the reference temperature (T_{ref} to the internal temperature (T_{int}).

$$B_{d_aortic} = W_{sten} \cdot \left(\frac{T_{ref}}{T_{int}}\right) \quad (2)$$

The heart's mechanical exergetic efficiency (3), η_1, is defined as the ratio between the reference power made available for pumping (W_{Heart_ref}) and the total metabolic exergy, expressed as a percentage [2]:

$$\eta_1 = 100 \cdot \left(\frac{W_{Heart_ref}}{B_M}\right) \quad (3)$$

The exergy transfer efficiency (4), η_2, represents the amount of energy utilized to move and heat the blood, being the ratio of blood exergy (ΔB_{blood}) to basal metabolism [2]:

$$\eta_2 = 100 \cdot \left(\frac{\Delta B_{blood}}{B_M}\right) \quad (4)$$

The scope of the present study was the analysis of the left heart, modeled as a thermodynamic system (Fig. 1), in order to investigate blood flow dynamics through the cardiac chambers and their respective valvular structures [8].

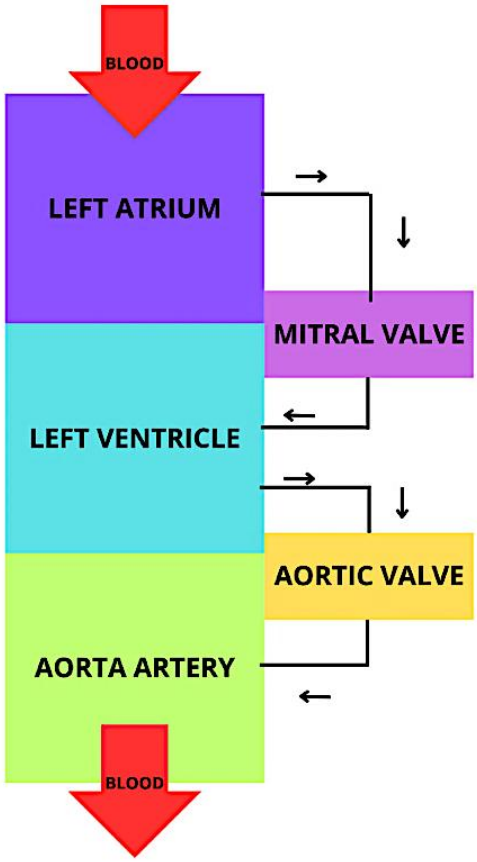


Fig. 1. Thermodynamic system of the left side of the human heart

B. Computational Implememntation

CardioExergy is an interactive web-based tool developed for the analysis of cardiac bioenergetics, assisting students, researchers, and healthcare professionals in investigating alterations associated with pathological conditions. It was implemented using Python, and the interfaces were developed with the support of Google AI Studio.

This web-based computational tool enables "in silico" studies by simulating various physiological and pathological scenarios of the cardiovascular system. The environment allows for the quantitative evaluation of hypotheses, sensitivity analyses, and the investigation of how hemodynamic, metabolic, and exergetic parameters influence cardiac performance.

C. Validation Protocol

The validation of the proposed mathematical model was carried out by evaluating its ability to reproduce physiological behaviors and hemodynamic trends described in the literature [2, 3] for varying degrees of aortic stenosis. To this end, simulation results were analyzed regarding the physical, mathematical, and physiological consistency of the calculated variables, including cardiac output, transvalvular pressure gradient, ventricular mechanical work and power, destroyed exergy, and exergy efficiency indices.

Additionally, the model's ability to adequately represent disease progression was investigated by verifying whether a gradual reduction in aortic valve area (AVA) produced responses consistent with the known pathophysiological mechanisms of aortic stenosis. Specifically, the expected observations included an increase in the transvalvular pressure gradient, elevated metabolic demand and cardiac work, greater exergy destruction resulting from blood flow irreversibilities, and a reduction in the cardiovascular system's overall energy efficiency. The agreement between these trends and those reported in the literature served as the criterion for the model's qualitative validation.

Simulation protocols (Table I) were established to analyze the model's behavior under various conditions related to aortic stenosis, including disease progression, as well as variations in heart rate, cardiac temperature, and valve geometry.

TABLE I. SIMULATIONS PROTOCOLS

Protocol	***Description***	***Objective***
Standard Test	Variation of AVA from 0.5 to 3.5 cm² in regular increments	Evaluate typical clinical progression and identify efficiency limits
Cardiac Stress	Simulation of different HR values for each AVA	Identify vulnerability under tachycardia or bradycardia

III. RESULTS

A. Software Overview

CardioExergy is an academic platform designed for simulating cardiac bioenergetics under conditions of aortic stenosis, based on mathematical models that apply principles of the Second Law of Thermodynamics [9]. The digital environment was developed to support teaching and research activities, offering users an interactive interface that enables the exploration of the heart's energy efficiency mechanisms across various clinical scenarios. The platform can be accessed via the following link: https://cardioexergy.mathematicalmodels.app/

The system usage workflow is structured sequentially. Initially, the user accesses the portal and selects the language.

Next, upon selecting the "Start Simulator" option, the participant is directed to the initial screen. The introductory screen provides context for the tool and outlines the theoretical background. From there, the user can enter values for heart rate, temperature, and discharge coefficient in the control panel (Fig. 2). It is also possible to select the AVA interval for the simulation.

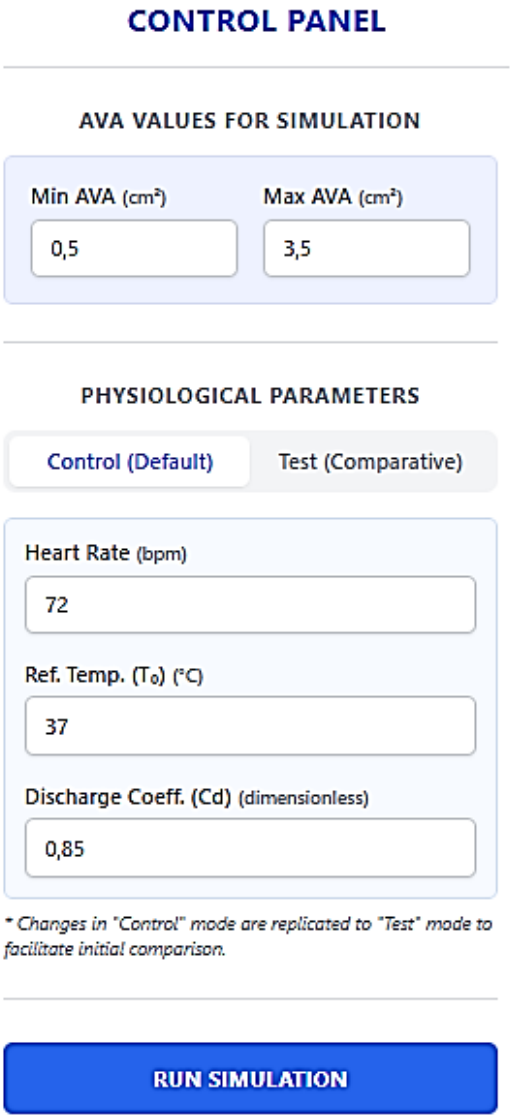


Fig. 2. Control Painel of cardioExergy

Once input parameters are defined and the simulation is executed, the results are displayed in specific interface tabs, allowing for an organized view of calculated variables, graphs, and cardiovascular system performance indicators. Additionally, the platform offers a comparative analysis mode, enabling users to run multiple simulations with different sets of initial parameters and compare the resulting data. This feature facilitates the assessment of how physiological, hemodynamic, and thermodynamic changes affect model behavior, supporting sensitivity studies, parametric analyses, and investigations into various clinical scenarios.

The central area of the interface features the "View Charts" tab (Fig. 3), designed to display simulation results via interactive graphs, allowing users to analyze the behavior of variables of interest under different simulated conditions. Complementarily, the "View Data" tab provides the corresponding numerical data in a tabular format, enabling detailed inspection of the values generated by the model.

The "Results" tab presents a textual interpretation of the findings, highlighting their physiological, hemodynamic, and thermodynamic significance, as well as their potential implications for cardiovascular function and the exergetic analysis of the heart.

The "Activities" tab proposes simulation activities and scenarios that encourage exploration of the computational model, helping users understand how input parameters influence calculated variables and the system's overall behavior.

Finally, the remaining tabs include an AI-based virtual assistant designed to support users while navigating the platform, and a "Help" section containing information, guidelines, and detailed instructions regarding the computational environment's features and capabilities.



Fig. 3. Web plattaform tabs

B. *Simulation of Different Degrees of Aortic Stenosis*

The first simulation was conducted with a heart rate of 72 bpm at 37°C and a discharge coefficient (Cd) of 0.85, aiming to compare variations in Aortic Valve Area (AVA)—categorized as severe (1 cm²), moderate (1.5 cm²), mild (2 cm²), or healthy (3.5 cm²). Shown below are the exergy graph (Fig. 4) based on AVA, and the efficiencies η_1 and η_2 (Figs. 5 and 6, respectively) corresponding to the different levels of aortic stenosis represented by the various AVA values.

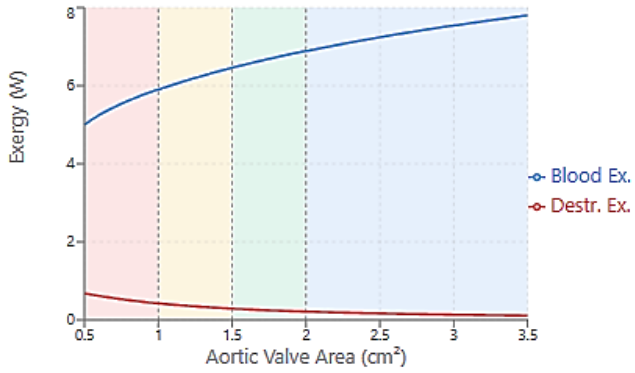


Fig. 4. Blood exergy and destroyed exergy as function of AVAs.

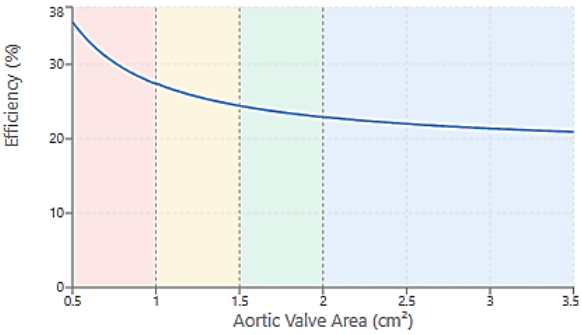


Fig. 5. Exergy Mechanical Effiency as function of AVAs.

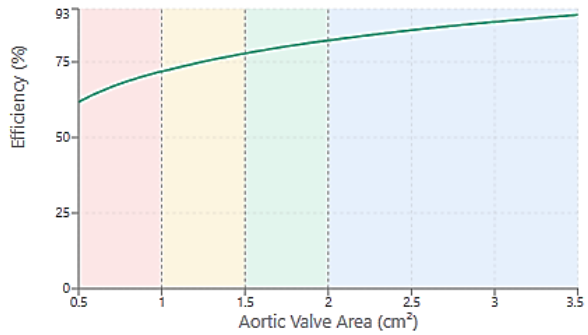


Fig. 6. Exergy Mechanical Effiency as function of AVAs.

C. Clinical Scenario Simulations

The second simulation was performed by comparing the behavior of exergy (Fig. 7), mechanical efficiency (Fig. 8), and exergy transfer efficiency (Fig. 9) at heart rates of 72 bpm (control) and 100 bpm (test), considering a temperature of 37°C and maintaining the discharge coefficient at 0.85.

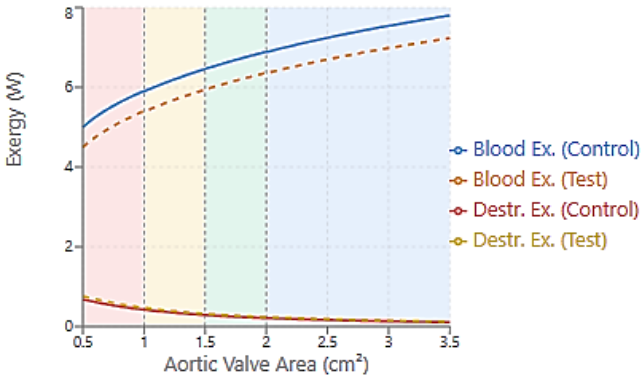


Fig. 7. Blood exergy and destroyed exergy as function of AVAs.

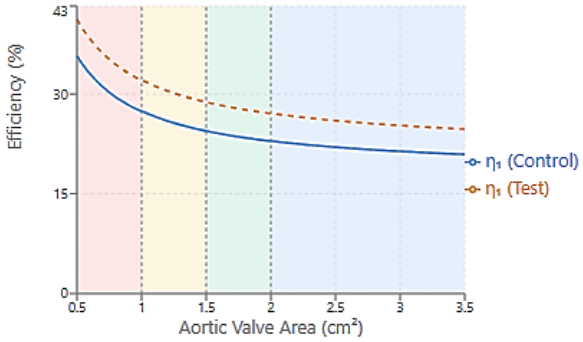


Fig. 8. Exergy Mechanical Effiency as function of AVAs.

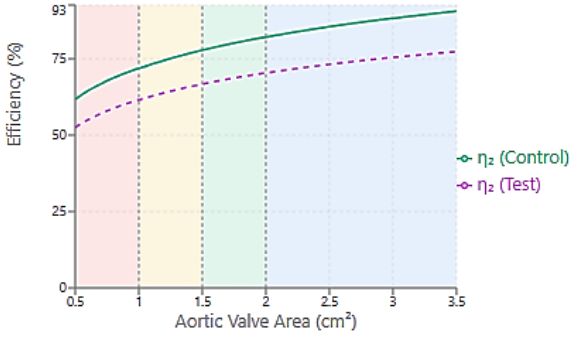


Fig. 9. Exergy Mechanical Effiency as function of AVAs.

Table II shows a comparison of the percentage change in results from Roll et al. [2] and CardioExergy,

TABLE II. COMPARAÇÃO DOS RESULTADOS

Variable	*AVA interval (cm²)*	*Roll et al. variation (%)*	*CardioExergy variation (%)*
Valve Exergy Destruction	1 → 2	-42.86	-50.0
	2 → 3.5	-37.5	-47,62
Blood Exergy	1 → 2	+12.0	+16.78
	2 → 3.5	+3.57	+13.21
Exergetic Mechanical Efficiency	1 → 2	-9.38	-16.31
	2 → 3.5	-13.79	-8.72
Exergy Transfer Efficiency	1 → 2	+10.16	+14.29
	2 → 3.5	+2.84	+10.30

IV. DISCUSSION

A. Validation of the simulations

The results obtained with CardioExergy showed behavior consistent with that reported by Roll et al. (2019) [2], highlighting the effects of aortic stenosis progression on the cardiovascular system's energetic performance. In the first simulation, as the aortic valve area was reduced from 3.50 cm² to 1.00 cm², an increase in exergy destruction at the valves was observed (from 0.11 to 0.42 W), accompanied by a reduction in the exergy transferred to the blood (from 7.80 to 5.90 W) and in the exergetic efficiency η_2 (from 90.58% to 71.85%). These trends align with the findings of Roll et al. [2], who demonstrated that a decrease in valve area leads to increased thermodynamic irreversibilities and reduced cardiac energy conversion efficiency.

In the second simulation, the observed increase in cardiac work confirms the higher energy demand placed on the left ventricle under conditions of severe stenosis, reproducing the same physical behavior described in the reference model. Increasing the heart rate from 72 to 100 bpm raises cardiac output and mechanical cardiac work, as more contraction-relaxation cycles occur within the same time interval, requiring greater myocardial effort [10]. Under both physiological and pathological conditions, an increased heart rate helps maintain the supply of oxygen and nutrients to the tissues. Consequently, myocardial oxygen consumption rises and cardiac metabolism intensifies, resulting in greater generation of metabolic exergy [11].

In both simulations, cases of more severe aortic stenosis show an increase in thermodynamic irreversibilities within the cardiovascular system, evidenced by greater exergy destruction. This occurs due to blood flow turbulence and viscous losses at the narrowed valve, which intensify dissipative mechanisms and reduce the energy available for useful work. Consequently, a decrease in exergy efficiency is observed, confirming that exergy destruction is a key indicator of stenosis severity. Thus, the more severe the stenosis, the greater the energy dissipation, the lower the capacity to

convert metabolic energy into useful work, and the lower the heart's overall efficiency. The apparent "higher efficiency" at advanced stages of stenosis is, in fact, a dangerous overload for the heart. The progressive reduction of the aortic valve area imposed a greater load on the left ventricle, evidenced by the increase in the transvalvular pressure gradient and the cardiac work required to maintain blood flow [3]. As the stenosis worsened, the heart had to generate higher pressures to overcome the additional flow resistance, signifying an increase in ventricular load. This behavior aligns with the findings of Roll et al. [2], who observed that a reduction in valve area results in greater mechanical effort by the heart and higher energy demand to sustain blood circulation.

Although it shows good agreement with the results of Roll et al. [2], CardioExergy still employs a simplified representation of the cardiovascular system, considering average operating conditions and pre-defined physiological parameters. Furthermore, the model does not account for individual patient characteristics or complex adaptive mechanisms associated with disease progression. Consequently, its results should be interpreted as estimates of the cardiovascular system's overall exergetic behavior, and future studies are required for its application in specific clinical scenarios.

B. Impact of Aortic Stenosis on Cardiac Energetics

Thus, the results obtained qualitatively validate CardioExergy, as they reproduce the key hemodynamic and exergetic trends reported by Roll et al. [2], while simultaneously extending the analysis through the inclusion of additional parameters such as heart rate, body temperature, and the valvular discharge coefficient. Consequently, CardioExergy stands as a significant scientific resource for understanding the effects of aortic stenosis on cardiac energy efficiency, fostering the integration of theory and practice in educational and research settings.

C. Model Validation

The CardioExergy tool offers potential applications across various academic and scientific contexts, standing out as a platform for research, education, and training. Its ability to integrate hemodynamic, metabolic, and thermodynamic parameters enables the quantitative investigation of cardiac energetic mechanisms, contributing to the advancement of knowledge in biomedical engineering, cardiovascular physiology, and applied thermodynamics. Furthermore, it can serve as an educational resource to facilitate the understanding of complex concepts related to cardiac performance and exergy analysis.

V. CONCLUSION

The CardioExergy tool is a computational platform capable of integrating hemodynamic, metabolic, and thermodynamic aspects into the analysis of cardiac performance. The results obtained aligned with trends described in the literature, demonstrating the model's ability to reproduce physiological behaviors and alterations associated with aortic stenosis. Furthermore, the use of exergy-based indicators allowed for the quantification of the disease's energetic impact, providing information complementary to conventional hemodynamic analyses.

Regarding limitations, the model employs a simplified representation of the cardiovascular system, omitting three-dimensional blood flow phenomena and long-term adaptive mechanisms. As a future step, the platform will undergo evaluation by students and faculty in the health sciences to assess its usability, educational applicability, and potential as a tool to support teaching and research in cardiovascular physiology.

ACKNOWLEDGMENT

The autors aknowledment financial support from CAPES and the Postgraduate Program of Biomedical Engineering of the University of Mogi das Cruzes.

REFERENCES

[1] PUJARI, Sai Harika; AGASTHI, Pradyumna. Aortic Stenosis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan–. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK557628/. Acesso em: 31 de janeiro de 2026.

[2] Roll JB, Borges ML, Mady CEK, Junior SO. Exergy analysis of the heart with a stenosis in the arterial valve. Entropy. 2019;21(6):563. doi:10.3390/e21060563.

[3] Henriques I, Mady C, Marin J, Serra L, Oliveira-Junior S. Exergy destroyed in the arteries due to stenosis. In: Proceedings of the 29th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental

[4] Çengel YA, Boles MA. Termodinâmica. 7. ed. Porto Alegre: AMGH; 2013.

[5] Levenspiel O. Termodinâmica amistosa para engenheiros. São Paulo: Edgard Blucher; 2002.

[6] Leite ACN, Souza GM, Albuquerque JC, Catão MNS, Diniz PP, Carvalho RO, Oliveira HL, Nascimento JJS. Revisão integrativa da análise de exergia no corpo humano. Campina Grande: Universidade Federal de Campina Grande; 2022.

[7] Suga H, Sagawa K, Shoukas AA. Load independence of the instantaneous pressure-volume ratio of the canine left ventricle and effects of epinephrine and heart rate on the ratio. Circ Res. 1973;32(3):314–22.

[8] GUYTON, Arthur C.; HALL, John E. Tratado de fisiologia médica. 13. ed. Rio de Janeiro: Elsevier, 2017

[9] Moran MJ, Shapiro HN, Boettner DD, Bailey MB. Princípios de termodinâmica para engenharia. 8. ed. Rio de Janeiro: LTC; 2018.

[10] Filho ESD, Santos BK. Sistemas hidráulicos e pneumáticos. Porto Alegre: SAGAH; 2018. E-book. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788595025158/. Acesso em: 30 nov. 2025.

[11] Neto CA, Pellegrini LF, Ferreira MS, Junior SO, Yanagihara JI. Exergy analysis of human respiration under physical activity. Int J Thermodyn. 2010;13(3):123–31.

Monte Carlo Evaluation of Z-Axis Tube Current Modulation for CT Dose Optimization

Felix Mas Milian*†, Diego Frias†, Joel Sanchez Dominguez†,
Susana Marrero Iglesias*†, Paulo Eduardo Ambrosio*†, Dany Sanchez Dominguez*†
*Universidade Estadual de Santa Cruz, Ilhéus, BA, Brazil
†Centro de Pesquisa e Desenvolvimento Tecnológico em Informática e Eletroeletrônica, Ilhéus, BA, Brazil
Emails: fmmilian@uesc.br, diegofrias@uneb.br, jsdominguez@iprj.uerj.br,
smiglesias@uesc.br, peambrosio@uesc.br, dsdominguez@uesc.br

***Abstract*—Computed tomography protocols must balance image quality and radiation dose. This work presents a fully simulation-based methodology to evaluate longitudinal tube current modulation in CT imaging. Both the topogram and the CT projection data were generated through Monte Carlo simulation using a computational patient model. The simulated topogram was stitched and used to estimate the longitudinal attenuation profile of the model, from which relative current weights were defined for each axial turn. These weights were then applied to a second set of Monte Carlo CT simulations to generate modulated projection data for image reconstruction and to calculate the absorbed dose within the model. As a proof of concept, the study was restricted to six axial turns in one anatomical range of a scaled MASH model and to a single 120 kV x-ray spectrum. A constant-current acquisition was compared with a modulated-current approach based on the square root of the attenuation profile. The modulated acquisition produced visually comparable reconstructed CT images and preserved the mean values measured in selected regions of interest. The total absorbed dose was reduced by approximately ten percent, with the largest local reduction observed in the neck region. These results indicate that the proposed workflow is effective for evaluating longitudinal current modulation strategies in small-field CT systems before experimental implementation.**

***Index Terms*—computed tomography, Monte Carlo simulation, tube current modulation, patient dose, image noise, topogram**

I. Introduction

Computed tomography (CT) is widely utilized in medical diagnosis due to its ability to provide high-spatial-resolution cross-sectional information. However, CT imaging also constitutes a major source of patient exposure to ionizing radiation. Consequently, optimization strategies are essential to minimize radiation dose while preserving the image quality required for clinical interpretation and quantitative analysis.

Automatic tube current modulation (ATCM) is a well-established strategy in clinical CT designed to adapt tube output to patient attenuation. The underlying principle involves increasing the tube current in highly attenuating regions and reducing it in less dense areas, thereby maintaining a target image quality with a lower radiation burden. Previous simulation and phantom studies have demonstrated that attenuation-based current modulation can significantly decrease dose, depending on patient geometry and modulation amplitude [1], [2]. Furthermore, clinical investigations have shown that z-axis modulation reduces the tube current–time product while maintaining image noise and diagnostic acceptability at comparable levels [3], [4].

Despite this established background, commercial ATCM implementations are typically proprietary. Their internal logic and specific responses to scanner geometry, detector design, patient shape, and reconstruction algorithms remain inaccessible. This lack of transparency limits the ability to study modulation strategies in a controlled and reproducible manner, particularly for custom CT systems and research prototypes.

In this work, we present and evaluate a Monte Carlo-based methodology to study z-axis tube current modulation in a compact CT system. The proposed workflow extracts a longitudinal attenuation profile from a topogram, translates this profile into modulation weights, and applies them to the number of simulated photon histories. Subsequently, it reconstructs the corresponding CT images and calculates the absorbed dose within the patient model using the same simulation framework. Rather than replicating a specific commercial ATCM implementation, this proof-of-concept study aims to provide a transparent, open framework to test z-axis modulation strategies under fully controlled conditions. It was intentionally restricted to one anatomical range, one 120 kV spectrum, and a scaled phantom geometry due to the high computational cost of pure Monte Carlo transport.

II. Materials and Methods

A. CT Geometry and Computational Phantom Simulation

The proposed methodology was developed to be later tested in a hybrid CT prototype with photon-counting and energy-integrating detectors [5], [6]. In this work, only the energy-integrating detector configuration was considered, and the acquisition was simulated in axial mode.

The simulated detector followed the specifications of the detector installed in the hybrid CT prototype. It was composed of 12 X-ACE 32 modules [6], each featuring a 32×32 pixel matrix with a pixel size of $1.06 \times 1.12\ \mathrm{mm}^2$. The source-to-object distance (SOD) and source-to-detector distance (SDD) were set to $493\ \mathrm{mm}$ and $986\ \mathrm{mm}$, respectively. After post-processing, the effective detector configuration used for topogram acquisition consisted of 384×32 pixels, providing multiple axial detector rows.

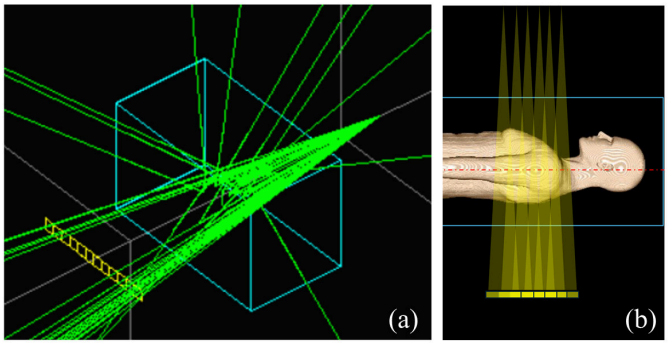


Fig. 1. (a) Simulation geometry and photon source visualization. (b) MASH axial turns selected for current modulation

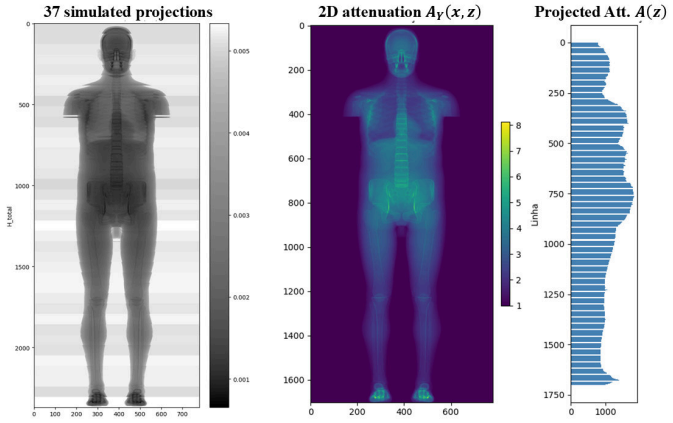


Fig. 2. Topogram-based estimation of the longitudinal attenuation profile. After projections were stitched, the normalized transmission image is converted to projected attenuation using $-\ln(I/I_0)$, and the attenuation values are summed along the transverse detector direction to obtain $A(z)$.

Monte Carlo simulations were performed using GATE version 9.4, a Geant4-based toolkit widely adopted for medical imaging and dosimetry applications [7], [8]. A 120 kV x-ray spectrum was implemented for the source. The Male Adult meSH (MASH) computational phantom was used as the patient model and was scaled to fit within the field of view of the compact CT system. The MASH phantom is based on polygon mesh surfaces and was designed for radiation transport and dosimetric studies [9], [10]. Additional details on its implementation in related computational CT simulations can be found in [11]–[14].

The simulations included a whole-body topogram of the scaled MASH model using 37 longitudinal positions. For this initial proof-of-concept assessment, however, only six axial turns from the thoracic region were simulated. Here, an axial turn denotes one full 360-degree axial acquisition at a selected longitudinal position. This longitudinal range covered the anatomy from the neck down to the lower rib cage, as shown in Fig. 1.

For each of the six turns, 360 angular projections were simulated for two acquisition scenarios: a baseline constant-current case and a z-axis modulated-current case.

B. Topogram Generation and Attenuation Profile

Initially, a topogram of the complete scaled MASH model was generated. A total of 37 simulated projections of 384×32 pixels were normalized and stitched to reconstruct the longitudinal topogram.

The normalized transmission image was converted into projected attenuation using

$$A(z,y) = -\ln\left(\frac{I(z,y)}{I_0(z,y)}\right), \tag{1}$$

where $I(z,y)$ is the transmitted intensity and $I_0(z,y)$ is the flat-field intensity, obtained from a projection without the MASH phantom. The longitudinal attenuation profile was then estimated by integrating the attenuation image along the transverse detector direction, as shown in Fig. 2:

$$A(z) = \sum_y A(z,y). \tag{2}$$

The resulting profile was smoothed using a cubic spline and used as the basis for the z-axis modulation curve along the selected patient range.

C. Z-Axis Current Modulation

In this work, tube current modulation was represented by a relative weight $w(z)$ assigned to each axial turn. This weight was derived from the longitudinal attenuation profile according to

$$w(z) = \frac{A(z)^\alpha}{\max\left[A(z)^\alpha\right]}, \tag{3}$$

where α controls the modulation strength. Two candidate functions were initially evaluated from the topogram: $\alpha = 1$, directly proportional to the attenuation profile, and $\alpha = 0.5$, proportional to the square root of the attenuation profile, as shown in Fig. 3. The square-root profile was selected for the Monte Carlo assessment because previous attenuation-based modulation studies reported favorable noise behavior for this type of dependence [1]. In addition, in the present topogram it limited the minimum weight to 0.757, compared with 0.573 for $\alpha = 1$, thereby avoiding an excessive reduction of photon histories in the lowest-attenuation position.

For each selected turn i, the number of simulated histories in the modulated scenario was defined by

$$N_i = w_i N_{\max}, \tag{4}$$

where $N_{\max} = 5 \times 10^7$ histories. The number of histories used for each turn is shown in Table I. In the constant-current reference acquisition, all turns were simulated with $N_{\max}$ histories.

D. CT Simulations and Reconstruction

The proof-of-concept assessment comprised 4320 independent simulations, corresponding to two acquisition scenarios and six turns, each one with 360 angular projections. To manage the high computational demand, with several hours required per projection, the simulations were executed in

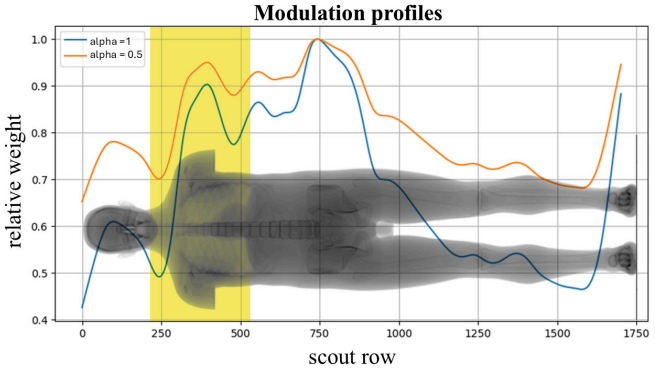


Fig. 3. Relative modulation profiles obtained from the longitudinal attenuation curve. The profile with $\alpha = 0.5$ was selected for the Monte Carlo assessment because it reduces dose while avoiding excessive loss of photon statistics in low-attenuation regions.

TABLE I
RELATIVE MODULATION WEIGHTS AND NUMBER OF HISTORIES FOR THE SIX EVALUATED AXIAL TURNS

Axial turn	$w_{\alpha=1}$	$N_{\alpha=1}$	$w_{\alpha=0.5}$	$N_{\alpha=0.5}$
5	0.573	28,649,134	0.757	37,845,062
6	0.727	36,344,723	0.851	42,548,965
7	0.948	47,389,939	0.973	48,668,849
8	1.000	50,000,000	1.000	50,000,000
9	0.928	46,387,524	0.963	48,159,906
10	0.876	43,803,048	0.936	46,792,931

parallel on a server equipped with an Intel Xeon Gold 5318Y processor, 503 GB of RAM, and Ubuntu 22.04. Before reconstruction, the projection outputs were converted from MHD/RAW to HDF5 format.

CT images were reconstructed using our in-house reconstruction code, RECON4. After axial stitching, the reconstructed volume for the constant-current scenario consisted of $364 \times 364 \times 144$ voxels. The same reconstruction and stitching pipeline was applied to the modulated-current data, as shown in Fig. 4.

E. Quantitative Analysis

Both reconstructed volumes were evaluated visually and quantitatively. Circular regions of interest (ROIs) with a radius of 15 pixels were placed within selected anatomical structures and background regions, as shown in Fig. 5. For each of the six simulated axial turns, the mean value, standard deviation, signal-to-noise ratio (SNR), and contrast-to-noise ratio (CNR) were calculated.

The SNR and CNR were defined as

$$\mathrm{SNR} = \frac{\bar{S}}{\sigma_S}, \tag{5}$$

and

$$\mathrm{CNR} = \frac{|\bar{S} - \bar{B}|}{\sqrt{\sigma_S^2 + \sigma_B^2}}, \tag{6}$$

where $\bar{S}$ and σ_S are the mean value and standard deviation of the selected signal ROI, and $\bar{B}$ and σ_B are the corresponding quantities for the background ROI.

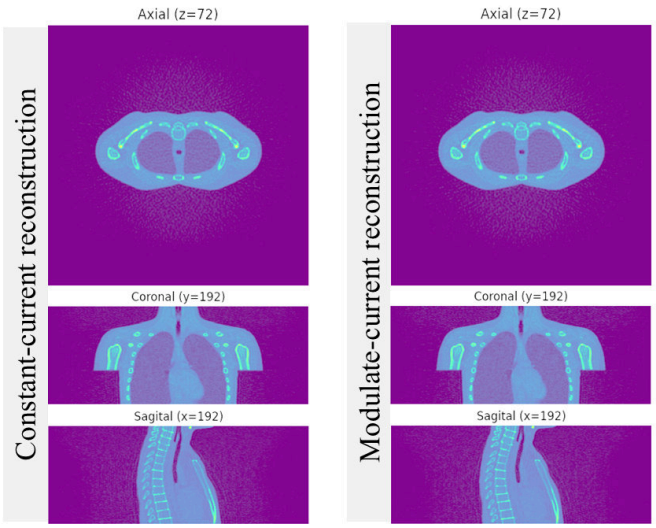


Fig. 4. Reconstructed CT images obtained without tube current modulation and with z-axis tube current modulation using $\alpha = 0.5$.

Two background ROIs, located in the superior and left lateral portions of the axial view, were included to account for differences associated with the asymmetric body thickness along these directions.

The total absorbed dose per axial turn was also extracted from the simulations for both acquisition scenarios. Dose values were reported in Gy for the number of histories used in each simulation.

No inferential statistical test was applied in this proof-of-concept study. Comparisons between constant-current and modulated-current acquisitions were performed descriptively using ROI mean values, standard deviations, SNR, CNR, and dose ratios. The results were summarized either as a function of the reconstructed-slice index of the CT volume or as a function of the axial-turn index, depending on the evaluated quantity.

III. RESULTS

The results presented in this section refer only to the six selected turns, corresponding to positions 5/37 to 10/37 of the simulated topogram. Therefore, they should be interpreted as a proof-of-concept assessment of the proposed workflow rather than as a full-body dosimetric evaluation.

A. Modulation Profile

The topogram-derived attenuation profile exhibited the expected longitudinal behavior. In the selected region used for CT simulation, highlighted in yellow in Fig. 3, attenuation minima were observed in the neck and lower thoracic anatomical regions, whereas maxima occurred near the shoulders and thicker anatomical cross-sections. Consequently, the modulation weights were lowest in turn 5/37 and progressively increased toward turn 8/37, where the maximum weight was reached, as shown in Table I.

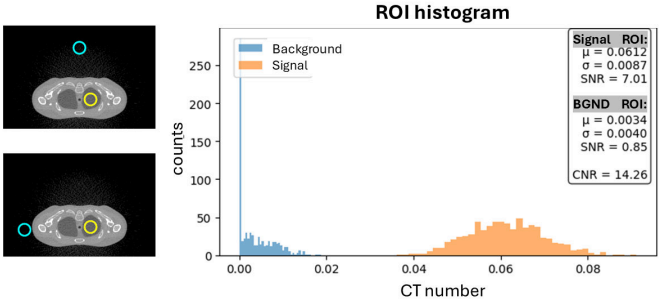


Fig. 5. Example of ROIs selected in the third evaluated axial turn for the quantitative analysis of signal and background regions in the constant-current case.

Under the $\alpha = 0.5$ modulation regime, the number of simulated histories was reduced from 5×10^7 to 3.78×10^7 for position 5/37, and to 4.25×10^7 for position 6/37. The remaining positions exhibited weights closer to unity. The average modulation weight was approximately 0.91, corresponding to an 8.7% mean reduction in simulated histories relative to the constant-current reference.

B. Image Reconstruction

Images reconstructed with and without z-axis current modulation showed visual comparability, as shown in Fig. 4. A quantitative analysis is presented in Fig. 6, where the mean value, SNR, and CNR were evaluated for selected ROIs. The results showed that the mean CT values in the selected anatomical regions, including the heart and lung, remained comparable between the constant-current and modulated-current acquisitions within the descriptive ROI analysis.

The mean background values showed differences when comparing the superior and lateral background regions, which is expected due to the asymmetric body thickness along these directions. However, for each background region, the values obtained with and without current modulation remained consistent. The CNR values were also very similar, with slightly higher values in the constant-current case. Overall, the modulated acquisition showed the expected marginal increase in noise, mainly in axial turns with reduced history counts, but no structural degradation or visible artifacts were introduced in the anatomical features.

C. Absorbed Dose

The absorbed dose values for each turn are compiled in Table II. The total absorbed dose decreased from 1.93×10^{-5} Gy in the constant-current baseline to 1.74×10^{-5} Gy in the modulated configuration, corresponding to an overall dose reduction of 10%.

The most pronounced reduction occurred in turn 5/37, associated with the neck region, where the absorbed dose decreased by 42.3%. Turn 6/37 also exhibited a substantial dose reduction of 26.9%. From 7/37 to 10/37, where modulation weights were closer to unity, dose variations were smaller. The acquisition 8/37 exhibited a negligible increase of 0.2%, which can be attributed to statistical fluctuations inherent to

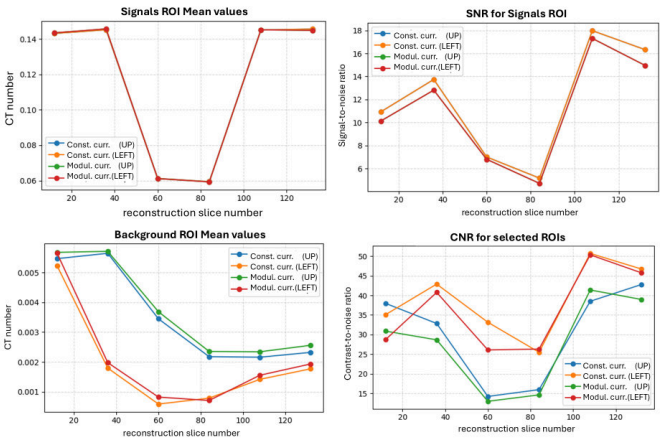


Fig. 6. Signal and backgrounds mean values, SNR, and CNR for selected ROIs.

TABLE II
ABSORBED DOSE COMPARISON FOR THE SIX SIMULATED AXIAL TURNS. DOSE VALUES ARE GIVEN IN GY FOR THE TOTAL NUMBER OF HISTORIES USED IN EACH SIMULATION.

Turn	No mod.	Mod.	Ratio	Reduction (%)
5	1.17×10^{-6}	6.75×10^{-7}	0.577	42.31
6	2.49×10^{-6}	1.82×10^{-6}	0.731	26.91
7	3.82×10^{-6}	3.62×10^{-6}	0.948	5.24
8	4.21×10^{-6}	4.22×10^{-6}	1.002	-0.24
9	3.94×10^{-6}	3.73×10^{-6}	0.947	5.33
10	3.70×10^{-6}	3.31×10^{-6}	0.895	10.54
Total	1.933×10^{-5}	1.738×10^{-5}	0.899	10.11

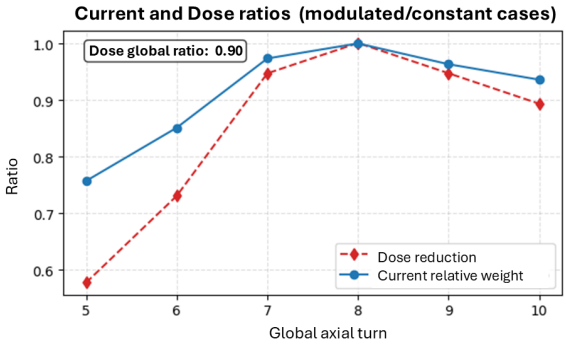


Fig. 7. Dose comparison between constant-current and modulated-current acquisitions. The modulated acquisition reduced the total absorbed dose by 10%, with the largest local reduction observed in the neck axial turn.

Monte Carlo dose estimation and to the fact that this axial turn was simulated with the maximum modulation weight.

IV. Discussion

These findings support the use of topogram-derived attenuation profiles to establish z-axis modulation weights within a compact CT simulation framework. The square-root modulation profile selectively reduced history counts in lower-attenuation positions while preserving robust photon statistics across thicker anatomical cross-sections. This performance aligns with the primary objective of ATCM: adapting tube

output to patient attenuation without compromising image quality [3].

The observed total dose reduction of 10% falls well within the ranges reported in previous tube current modulation literature. Specifically, Gies et al. reported simulated dose reductions of up to 50%, contingent upon phantom geometry and the chosen modulation function [1]. Similarly, Kalender et al. measured experimental dose reductions between 8% and 56% using physical phantoms [2]. Clinical assessments of z-axis modulation have likewise demonstrated reduced tube current–time products alongside comparable image noise and diagnostic acceptability [4]. Hence, the dose savings achieved in this study are highly plausible, particularly given the intentionally conservative nature of the selected $\alpha = 0.5$ profile.

While a more aggressive modulation profile (e.g., $\alpha = 1$) would yield greater dose saving, particularly in the neck region, it risks increasing image noise. Such configurations should only be explored after establishing explicit noise tolerance thresholds. For this initial proof-of-concept assessment, the $\alpha = 0.5$ profile served as a robust baseline, producing measurable dose reductions while preserving visual image quality and ROI mean values.

Several limitations warrant acknowledgment. First, this study was restricted to six selected acquisitions of the MASH phantom due to the long computation times required for pure Monte Carlo transport. Second, evaluations were limited to a 120 kV x-ray spectrum. Alternative spectra may alter beam hardening effects, photon statistics, absorbed dose distribution, and reconstructed image quality. Third, the current quantitative analysis focused exclusively on selected ROIs and per-acquisition dosimetry. A more comprehensive assessment should encompass additional anatomical regions, formal dose uncertainty estimation, and rigorous image-noise acceptance criteria. Finally, the proposed methodology represents an idealized simulation framework rather than a clinical hardware implementation; thus, experimental evaluation with the Hybrid CT prototype remains imperative. Future work will extend the simulations to the full set of 37 longitudinal positions, include different x-ray spectra, incorporate the response of the physical hybrid CT detectors, evaluate radiologist-defined noise tolerance criteria, and investigate AI-based denoising or acceleration strategies to enable larger simulated volumes and more aggressive modulation profiles.

V. Conclusion

A Monte Carlo-based proof-of-concept workflow was developed and evaluated to study z-axis tube current modulation within a compact CT system. It integrates topogram-based attenuation profiling, modulation weight generation, photon history scaling, RECON4 image reconstruction, and MASH phantom dosimetry within a unified workflow. Utilizing a conservative square-root modulation profile ($\alpha = 0.5$), an overall absorbed dose reduction of 10% was achieved, with localized dose savings reaching up to 42% in the neck region. Reconstructed images were visually comparable to the constant-current baseline, and ROI mean values were preserved. These findings show that the proposed methodology is suitable for preliminary benchmarking of tube current modulation strategies before physical prototyping and broader experimental evaluation.

Acknowledgment

The authors would like to thank the Ministério de Ciência Tecnologia e Inovação (MCTI) and the Associação para Promoção da Excelência do Software Brasileiro (Softex) for the financial support provided for this research through the Priority Program PNM Design (01245.019446/2022-11).

References

[1] M. Gies, W. A. Kalender, H. Wolf, and C. Suess, "Dose reduction in CT by anatomically adapted tube current modulation. I. Simulation studies," *Med. Phys.*, vol. 26, no. 11, pp. 2235–2247, 1999, doi: 10.1118/1.598779.

[2] W. A. Kalender, H. Wolf, and C. Suess, "Dose reduction in CT by anatomically adapted tube current modulation. II. Phantom measurements," *Med. Phys.*, vol. 26, no. 11, pp. 2248–2253, 1999, doi: 10.1118/1.598738.

[3] M. K. Kalra *et al.*, "Techniques and applications of automatic tube current modulation for CT," *Radiology*, vol. 233, no. 3, pp. 649–657, 2004, doi: 10.1148/radiol.2333031150.

[4] M. K. Kalra *et al.*, "Comparison of z-axis automatic tube current modulation technique with fixed tube current CT scanning of abdomen and pelvis," *Radiology*, vol. 232, no. 2, pp. 347–353, 2004, doi: 10.1148/radiol.2322031304.

[5] H. R. Mendes *et al.*, "Hybrid CT prototype with photon-counting and energy-integrating detectors: a comparative study of image quality," in *Medical Imaging 2026: Physics of Medical Imaging*, Proc. SPIE 13924, 139242S, 2026, doi: 10.1117/12.3085971.

[6] Detection Technology Plc, "X-ACE 32: off-the-shelf medical CT detector," 2026. [Online]. Available: https://www.deetee.com/product/x-ace-32/. Accessed: Jun. 22, 2026.

[7] D. Sarrut *et al.*, "The OpenGATE ecosystem for Monte Carlo simulation in medical physics," *Phys. Med. Biol.*, vol. 67, no. 18, Art. no. 184001, 2022, doi: 10.1088/1361-6560/ac8c83.

[8] S. Jan *et al.*, "GATE: a simulation toolkit for PET and SPECT," *Phys. Med. Biol.*, vol. 49, no. 19, pp. 4543–4561, 2004, doi: 10.1088/0031-9155/49/19/007.

[9] V. F. Cassola, V. J. de Melo Lima, R. Kramer, and H. J. Khoury, "FASH and MASH: female and male adult human phantoms based on polygon mesh surfaces: I. Development of the anatomy," *Phys. Med. Biol.*, vol. 55, no. 1, pp. 133–162, 2010, doi: 10.1088/0031-9155/55/1/009.

[10] R. Kramer *et al.*, "FASH and MASH: female and male adult human phantoms based on polygon mesh surfaces: II. Dosimetric calculations," *Phys. Med. Biol.*, vol. 55, no. 1, pp. 163–189, 2010, doi: 10.1088/0031-9155/55/1/010.

[11] F. M. Milian, J. S. Dominguez, A. A. L. Bezerra, C. B. de Jesus, and J. P. F. Guimarães, "Abordagem fixed forced detection na aceleração de simulações tomográficas utilizando o código GATE," in *Anais do XXVI Encontro Nacional de Modelagem Computacional e XIV Encontro de Ciência e Tecnologia dos Materiais*, Nova Friburgo, Brazil, 2023, doi: 10.29327/1340957.696667.

[12] F. M. Milian *et al.*, "Código GATE na geração de imagens e cálculo de doses de um sistema tomográfico convencional," in *Anais do XXVI Encontro Nacional de Modelagem Computacional e XIV Encontro de Ciência e Tecnologia dos Materiais*, Nova Friburgo, Brazil, 2023.

[13] F. M. Milian *et al.*, "GATE code in the image generation and dose calculation of a conventional tomographic system," *J. Eng. Exact Sci.*, vol. 10, no. 1, Art. no. 17803, 2024, doi: 10.18540/jcecvl10iss01pp17803.

[14] F. M. Milian *et al.*, "Multimaterial classification for dual-energy CBCT: phantom study by simulations," in *XXIX Brazilian Congress on Biomedical Engineering*, IFMBE Proc., Cham, Switzerland: Springer, 2025, pp. 720–730, doi: 10.1007/978-3-031-94921-0_78.

Influência da Cinética da Na, K-ATPase na Atividade Epileptiforme Não-Sináptica

Jesica Talero Maldonado
Facultad de Ingeniería Mecánica, Electrónica y Biomédica (FIMEB)
Universidad Antonio Nariño (UAN)
Bogotá D.C., Colômbia
jesica.talerom@gmail.com

Maria Daniela Aparecida Resende
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
mdani.resende@hotmail.com

Fúlvio Alexandre Scorza
Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, SP, Brasil
scorza@unifesp.br

Kleber de Jesus Silva
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
kl3b3rr99@gmail.com

Arthur Carvalho Pires
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
arthurcpires@hotmail.com

Carla Alessandra Scorza
Escola Paulista de Medicina
Universidade Federal de São Paulo
São Paulo, SP, Brasil
carla.scorza@unifesp.br

Antônio-Carlos Guimarães de Almeida
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
acga@ufsj.edu.br

Antônio Márcio Rodrigues
Programa de Pós-Graduação em Bioengenharia
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil
amr@ufsj.edu.br

***Abstract*—The Na^+/K^+-ATPase plays a fundamental role in maintaining ionic homeostasis and neuronal excitability and is commonly represented in computational models of epileptiform activity by phenomenological formulations. However, these approaches do not explicitly describe the conformational transitions of the catalytic cycle, limiting the representation of transient responses during rapid ionic fluctuations. In this work, a kinetic Na^+/K^+-ATPase model based on the Post-Albers cycle was integrated into a computational model of non-synaptic epileptiform activity. Three formulations were implemented to estimate the pump current, and the effects of rapid Na^+ and K^+ concentration changes in the perimembrane region were investigated. The proposed model preserved the overall dynamics of epileptiform events while providing a detailed description of pump state occupancy and transition kinetics. Simulations showed that transient perimembrane ionic fluctuations produced dynamic behaviors not reproduced by the phenomenological model. In particular, the current formulation based on electrogenic transitions of the catalytic cycle provided the best qualitative agreement with experimental observations, reproducing post-ictal oscillatory events associated with rapid extracellular ionic changes. Furthermore, simulations of reduced extracellular volume reproduced the emergence of post-events observed under hypo-osmotic conditions. These results indicate that explicit kinetic modeling of the Na^+/K^+-ATPase provides a more realistic description of pump dynamics during intense neuronal activity and offers a useful framework for investigating the role of ionic homeostasis in epileptiform activity.**

Keywords— Na^+/K^+-ATPase, Post-Albers cycle, non-synaptic epileptiform activity, computational modeling, ionic homeostasis.

I. INTRODUÇÃO

A atividade elétrica cerebral depende da manutenção rigorosa dos gradientes transmembranares de Na^+ e K^+, os quais são continuamente restabelecidos pela Na^+/K^+-ATPase [1], [2], [3]. Além de preservar o potencial de repouso, essa proteína desempenha papel essencial na recuperação da excitabilidade neuronal após potenciais de ação e na manutenção da homeostase iônica durante períodos de intensa atividade elétrica [4], [5]. Alterações em sua atividade comprometem o equilíbrio eletroquímico da membrana, favorecendo condições associadas à hiperexcitabilidade neuronal, como a epilepsia.

Modelos computacionais têm contribuído significativamente para a compreensão dos mecanismos envolvidos na geração e manutenção da atividade epileptiforme [4], [6]. Entre eles, o modelo de Atividade Epileptiforme Não Sináptica (AENS) desenvolvido por [4] descreve a dinâmica espaço-temporal das concentrações iônicas e do potencial de membrana, permitindo investigar como alterações na homeostase extracelular podem desencadear crises epilépticas. Entretanto, nesse modelo, a Na^+/K^+-ATPase é representada por uma formulação fenomenológica baseada apenas na dependência das concentrações de Na^+, K^+ e ATP, sem considerar explicitamente as etapas cinéticas do ciclo catalítico da enzima.

Experimentalmente, sabe-se que a Na^+/K^+-ATPase opera por meio de uma sequência de estados conformacionais descritos pelo ciclo de Albers-Post, envolvendo processos de ligação e liberação de íons, fosforilação, desfosforilação e transições conformacionais [3], [7]. A velocidade de cada uma dessas etapas pode influenciar a resposta dinâmica da bomba frente às rápidas alterações iônicas observadas durante eventos epileptiformes, aspecto que não pode ser representado por modelos simplificados.

Neste trabalho, o objetivo foi desenvolver um modelo cinético da Na^+/K^+-ATPase baseado no ciclo de Albers-Post e integrá-lo a um modelo computacional de atividades epileptiformes não-sinápticas. Simulações computacionais foram realizadas para investigar os efeitos da cinética explícita da bomba sobre o transporte iônico, a corrente eletrogênica e a dinâmica das crises. Os resultados demonstram que o

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

modelo reproduz respostas transitórias da Na^+/K^+-ATPase que não podem ser representadas por abordagens fenomenológicas, fornecendo uma descrição mais realista da dinâmica da bomba durante atividades neuronais intensas.

II. Métodos

A. Modelo para atividades epileptiformes não-sinápticas

O modelo matemático usado para simular atividades epileptiformes não-sinápticas (AENS) descreve o transporte de íons tanto no meio extracelular quanto através da membrana celular [4]. A movimentação iônica no espaço extracelular foi representada por uma equação de eletrodifusão baseada na formulação de Nernst-Planck [8]. O transporte transmembranar incluiu os cotransportadores $Na^+/K^+/2Cl^-$ (NKCC) e K^+/Cl^- (KCC), os trocadores Na^+/H^+ (NHE) e HCO_3^-/Cl^- (HCE), a bomba de H^+ e a enzima anidrase carbônica, responsável pela interconversão reversível entre CO_2, H_2O, H^+ e HCO_3^- [9]. Os fluxos iônicos promovidos por esses mecanismos foram calculados em função das concentrações dos ligantes e das afinidades dos respectivos transportadores.

A manutenção dos gradientes transmembranares de Na^+ e K^+ foi representada pela Na^+/K^+-ATPase, responsável pelo efluxo de três íons Na^+ e influxo de dois íons K^+ por ciclo catalítico [5]. No presente trabalho, o modelo fenomenológico originalmente empregado na AENS foi substituído por um modelo cinético baseado no ciclo de Albers-Post. As correntes da bomba e dos canais iônicos foram calculadas por meio da equação de Goldman-Hodgkin-Katz (GHK) [4].

Os neurônios foram representados por canais rápidos de Na^+, canais de K^+ tipo A, canais persistentes de Na^+, canais retificadores tardios de K^+, canais muscarínicos de K^+ e canais de Cl^- dependentes de voltagem, cujas permeabilidades seguiram o formalismo de [10]. Para os astrócitos, a membrana foi considerada inexcitável, contendo canais retificadores de entrada para K^+ e permeabilidades constantes para os demais canais [4].

B. Modelo Cinético para Na^+/K^+-ATPase

A Fig. 1 apresenta o modelo cinético da Na^+/K^+-ATPase utilizado neste trabalho, baseado no ciclo de Albers-Post [11] e expandido pela inclusão do subestado P-E2(Na_3), envolvido na via de transporte de Na^+ [12]. Esse subestado foi incorporado por sua relevância na eletrogenicidade da bomba e por já ter sido empregado em estudos experimentais e computacionais posteriores [12], [13], [14].

O modelo descreve as transições conformacionais entre os estados E_1 e E_2, incluindo as etapas de ligação, oclusão e liberação de Na^+ e K^+, bem como as reações de fosforilação e desfosforilação da enzima. As constantes cinéticas, originalmente propostas por [11], foram complementadas pelas constantes adicionais reportadas por [12], possibilitando a representação completa do ciclo expandido.

A partir do esquema apresentado na Fig. 1, foram definidas 30 taxas de reação (r_i), cada uma correspondente ao fluxo líquido entre dois subestados consecutivos do ciclo catalítico. De forma geral, cada taxa foi calculada como a diferença entre a velocidade da reação direta e a velocidade da reação reversa, sendo ambas determinadas pelas populações dos subestados envolvidos, pelas respectivas constantes cinéticas e, quando aplicável, pelas concentrações dos ligantes (Na^+, K^+, ATP, ADP e Pi). As expressões específicas das taxas

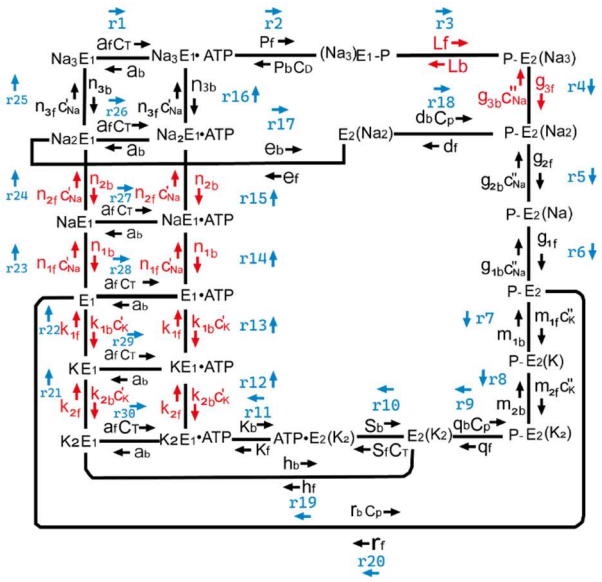


Fig. 1. Modelo cinético expandido da Na^+/K^+-ATPase baseado no ciclo de Albers-Post, adaptado de [11]. O modelo inclui o subestado P-E2(Na_3) e as constantes cinéticas adicionais propostas por [12]. As setas indicam as reações direta e reversa; E_1 e E_2 representam os principais estados conformacionais da enzima; **CT**, **CD** e **CP** correspondem às concentrações de ATP, ADP e Pi, respectivamente; C'_{Na}, C'_K, C''_{Na} e C''_K representam as concentrações intracelulares e extracelulares de Na^+ e K^+.

seguiram diretamente o esquema cinético da Fig. 1, de acordo com a seguinte equação geral:

$$r_i = k_{i,f} \prod_j C_j^{\nu_j} X_i - k_{i,b} \prod_j C_j^{\mu_j} X_{i+1}, \qquad (1)$$

onde X_i representa a população do subestado, $k_{i,f}$ e $k_{i,b}$ são as constantes cinéticas das reações direta e reversa, e C_j corresponde às concentrações dos ligantes envolvidos em cada etapa. As 30 reações do ciclo foram descritas por essa formulação geral, variando apenas os subestados, os ligantes e as constantes cinéticas de cada transição.

As equações diferenciais do modelo foram, então, obtidas por balanço de massa, considerando, para cada um dos 22 subestados, a soma das contribuições das reações de entrada e saída. Dessa forma, a evolução temporal da população de cada subestado foi descrita por um sistema de equações diferenciais ordinárias acopladas.

A corrente da Na^+/K^+-ATPase pode ser estimada por diferentes estratégias, dependendo do nível de detalhamento considerado no ciclo catalítico. Neste trabalho foram propostas três abordagens complementares: (i) uma baseada apenas nos fluxos líquidos de Na^+ e K^+ efetivamente transportados através da membrana; (ii) uma que considera simultaneamente as etapas de ligação e liberação dos íons, representando todo o transporte ao longo do ciclo catalítico; e (iii) uma fundamentada nas contribuições eletrogênicas das transições individuais do ciclo, conforme o formalismo de [15]. A comparação entre essas abordagens permitiu avaliar como diferentes representações da eletrogênese da bomba influenciam a estimativa da corrente de bombeamento.

Para isso, foram calculadas as quantidades de Na^+ ligadas no meio intracelular (ΔNa_i) e liberadas no meio extracelular (ΔNa_o), bem como as quantidades de K^+ ligadas no meio

extracelular (ΔK_o) e liberadas no meio intracelular (ΔK_i). Essas grandezas foram obtidas pela soma das taxas líquidas das reações correspondentes e utilizadas para o cálculo das três versões da corrente da bomba.

Na primeira abordagem (IpV1), a corrente foi estimada a partir da diferença entre os fluxos líquidos de Na^+ liberado no meio extracelular e K^+ liberado no meio intracelular:

$$I_{pV1} = F(\Delta Na_o - \Delta K_i) \quad (2)$$

Na segunda abordagem (IpV2), foram consideradas conjuntamente as etapas de ligação e liberação dos íons, resultando em:

$$I_{pV2} = F[(\Delta Na_i + \Delta Na_o) - (\Delta K_i + \Delta K_o)] \quad (3)$$

Por fim, a terceira abordagem (IpV3) foi baseada no modelo eletrogênico de [15], no qual a corrente corresponde à soma das contribuições das transições eletrogênicas do ciclo, ponderadas pelos respectivos coeficientes dielétricos obtidos de [11]:

$$I_{pV3} = e_0 \sum_i \alpha_i r_i \quad (4)$$

onde e_0 é a carga elementar, r_i representa a taxa líquida da i-ésima reação e α_i corresponde ao coeficiente dielétrico associado à respectiva transição.

C. Efeito eletrogênico da Bomba de Na^+/K^+

Além de reproduzir as alterações das concentrações intracelulares e extracelulares de Na^+ e K^+, o modelo também incorporou a contribuição eletrogênica da Na^+/K^+-ATPase ao cálculo do potencial de membrana. Como o potencial neuronal é determinado pela equação de Goldman-Hodgkin-Katz (GHK), a corrente da bomba foi convertida em um termo auxiliar (C_A), obtido por rearranjo da forma correspondente da equação de corrente de GHK. Esse procedimento permite incorporar o efeito eletrogênico da bomba na formulação do potencial sem introduzir instabilidades numéricas. Assim, a partir da corrente da bomba (I_{pNaK}), foi calculado o termo C_A:

$$C_A = \frac{RT}{F^2 V_m} I_{pNaK} + \left(1 - \frac{RT}{F^2 V_m} I_{pNaK}\right) \alpha e^{-\frac{FV_m}{RT}}, \quad (5)$$

onde R é a constante universal dos gases, α uma constante, T a temperatura absoluta, F a constante de Faraday e V_m o potencial de membrana.

O termo C_A foi então incorporado ao denominador da equação de GHK utilizada no modelo AENS, juntamente com as contribuições das correntes por gap junction (f_{gsi}) e efeito de campo(f_{ef}):

$$Vm = \frac{RT}{F} ln\left(\frac{P_{Na}[Na^+]_o + P_K[K^+]_o + P_{Cl}[Cl^-]_i + f_{ef} + f_{gsi} + \alpha}{P_{Na}[Na^+]_i + P_K[K^+]_i + P_{Cl}[Cl^-]_o + C_A}\right) \quad (6)$$

onde P_{ion} é a permeabilidade, $[ion]_o$ e $[ion]_i$ as concentrações iônicas no meio intra e extracelular.

Para avaliar o impacto da nova representação da bomba, os resultados obtidos com o modelo cinético foram comparados ao modelo fenomenológico originalmente empregado por [4], analisando-se seus efeitos sobre as concentrações iônicas, o potencial de membrana e as características dos períodos ictal e interictal, incluindo duração, amplitude e intervalo entre eventos.

D. Simulações Computacionais

Os parâmetros e valores iniciais das variáveis do modelo foram baseados em trabalhos anteriores [4], [5], [9]. O método das diferenças finitas, considerando um erro menor que 10^{-4}, foi utilizado para realizar as simulações. Para as simulações, o programa foi implementado em linguagem Fortran (FORTRAN 90) e os resultados foram analisados em Matlab®. O código do programa foi compilado e executado em um computador de alta performance (Cluster SGI UV 2000, 80 Cores — Intel Xeon E5-4650v2 10-core, 2.4 GHz, 25 MB Cache, 1024GB DDR3 1866 MHz, 80 TB HD, Cluster SGI UV 300, 144 Cores Intel Xeon E7-8867v4, 2.4 GHz, RAM 2,048 GB DDR4 2,400MHz, 80 TB HD, SUSE Linux Enterprise Server 11, SGI Performance Suite, Intel Cluster Studio XE).

No modelo AENS original, as concentrações iônicas são assumidas espacialmente homogêneas em cada compartimento intra- e extracelular. Assim, os fluxos transmembranares alteram instantaneamente a concentração de todo o compartimento, reduzindo as taxas locais de variação das concentrações iônicas e, consequentemente, atenuando os efeitos transitórios decorrentes da cinética finita da Na^+/K^+-ATPase. Nessas condições, o modelo cinético baseado no ciclo de Albers-Post apresenta comportamento semelhante ao modelo fenomenológico, uma vez que ambos operam próximos ao regime estacionário.

Para investigar os efeitos de variações rápidas das concentrações iônicas nas proximidades da membrana, foram introduzidos compartimentos intermediários para Na^+ e K^+ nas faces intracelular e extracelular. As concentrações nesses compartimentos passaram a ser determinadas simultaneamente pelos fluxos transmembranares e pela difusão com os compartimentos principais:

$$\frac{d[\text{ion}]_o^{int}}{dt} = \frac{\phi_{\text{ion}}^m A^m}{f_o Vol_o} + v_o([\text{ion}]_o - [\text{ion}]_o^{int}), \quad (7)$$

$$\frac{d[\text{ion}]_i^{int}}{dt} = -\frac{\phi_{\text{ion}}^m A^m}{f_i Vol_i} + v_i([\text{ion}]_i - [\text{ion}]_i^{int}), \quad (8)$$

onde ϕ_{ion}^m representa o fluxo transmembranar do íon, A^m a área de membrana, Vol o volume do compartimento, f define a fração de volume correspondente à região perimembranar e v representa a taxa de difusão entre o compartimento intermediário e o compartimento principal. Valores menores de f intensificam as variações locais de concentração, enquanto valores maiores de v aceleram o restabelecimento do equilíbrio com o restante do meio intra- ou extracelular.

III. Resultados

Na Fig. 2A, pode-se observar que o modelo cinético da Na^+/K^+-ATPase incorporado no modelo de AENS permite reproduzir a dinâmica característica da atividade

epileptiforme não sináptica para as três formulações propostas para o cálculo da corrente eletrogênica (IpV1, IpV2 e IpV3), apresentando comportamento semelhante ao obtido com o modelo fenomenológico. A Fig. 2B apresenta os efeitos da inclusão de compartimentos intermediários para representar

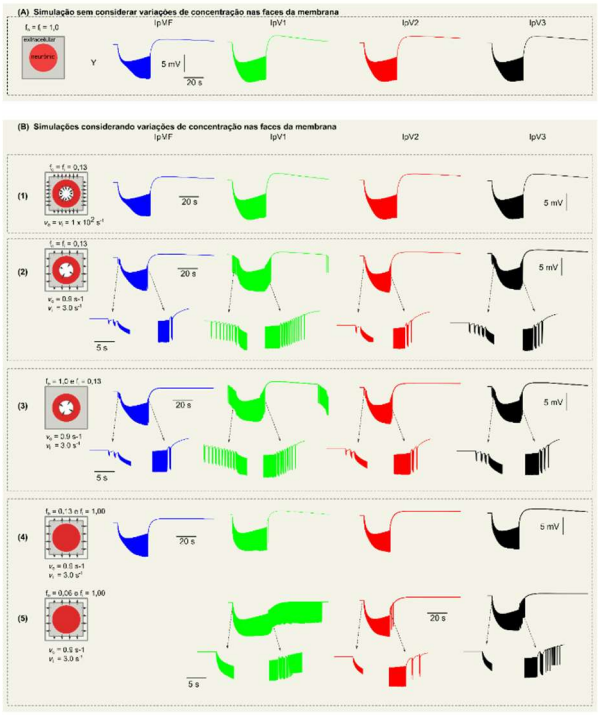


Fig. 2. Influência das flutuações locais das concentrações de Na^+ e K^+ nas proximidades da membrana sobre a atividade epileptiforme simulada. (A) Modelo original, sem compartimentos perimembranares. (B1–B5) Resultados obtidos para diferentes valores da fração do volume perimembranar (f_i e f_o) e das taxas de difusão (v_i e v_o). A consideração explícita desses compartimentos evidencia o surgimento de efeitos transitórios dependentes da cinética da Na^+/K^+-ATPase, não observados na representação fenomenológica convencional.

variações rápidas das concentrações de Na^+ e K^+ nas proximidades da membrana. Quando foi considerada uma fração correspondente a 13% dos volumes intra- e extracelulares ($f_i = f_o = 0{,}13$) e altas taxas de difusão ($v_i = v_o = 10^2\ s^{-1}$), o comportamento dos eventos epileptiformes permaneceu semelhante ao do modelo controle (Fig. 2A-B1). Entretanto, a redução das taxas de difusão ($v_i = 3{,}0\ s^{-1}$ e $v_o = 0{,}9\ s^{-1}$) resultou no aparecimento de pré- e pós-eventos, caracterizados por descargas de menor amplitude no início e no término do período ictal (Fig. 2B2).

Para identificar a contribuição de cada face da membrana, foram realizadas simulações considerando separadamente as variações rápidas das concentrações intracelulares e extracelulares. Quando apenas o compartimento intracelular foi considerado, observou-se comportamento semelhante ao da simulação completa, com a presença de pré- e pós-eventos (Fig. 2B3). Em contraste, considerando somente o compartimento extracelular ($f_o = 0{,}13$), esses eventos não foram observados (Fig. 2B4).

A redução adicional da fração do compartimento extracelular para $f_o = 0{,}065$ evidenciou diferenças entre os modelos da Na^+/K^+-ATPase. Enquanto o modelo fenomenológico tornou-se numericamente instável, impedindo a simulação, o modelo cinético baseado no ciclo de Albers-Post permaneceu estável nas três estratégias de cálculo da corrente (IpV1, IpV2 e IpV3), reproduzindo pós-eventos. Entre elas, a abordagem eletrogênica baseada nas transições do ciclo (IpV3) produziu respostas mais compatíveis com o comportamento experimental esperado.

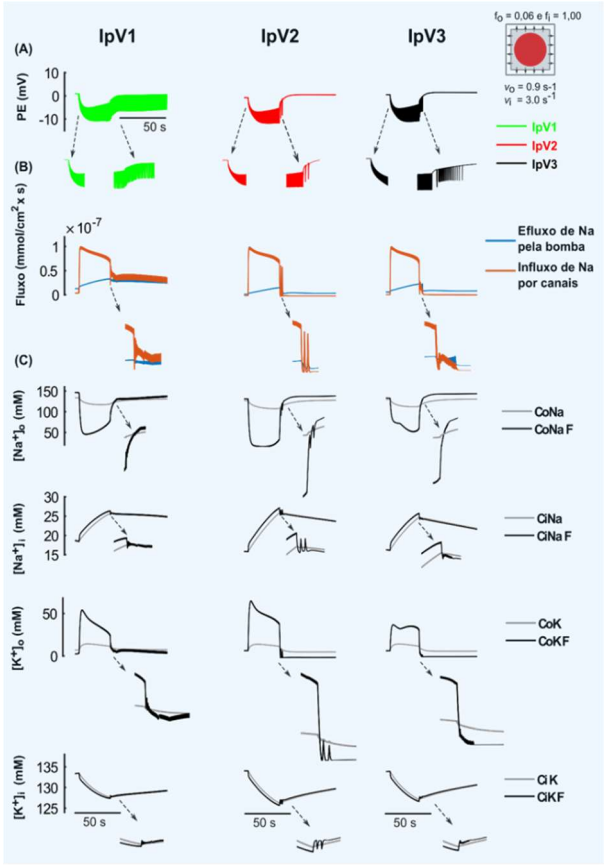


Fig. 3. Efeito das flutuações rápidas das concentrações perimembranares ($f_o = 0{,}06$, $f_i = 1{,}00$, $v_o = 0{,}9\ s^{-1}$ e $v_i = 3{,}0\ s^{-1}$) sobre a dinâmica da atividade epileptiforme. (A) Eventos epileptiformes simulados utilizando as três formulações da corrente da Na^+/K^+-ATPase (IpV1, IpV2 e IpV3). (B) Cruzamento entre o efluxo de Na^+ pela bomba e o influxo de Na^+ pelos canais. (C) Concentrações de Na^+ e K^+ nos compartimentos globais e perimembranares.

Esses resultados indicam que a consideração explícita da cinética da Na^+/K^+-ATPase, associada às variações locais de concentração nas proximidades da membrana, permite reproduzir dinâmicas transitórias que não são obtidas com a representação fenomenológica convencional.

As Figuras 3 e 4 apresentam os resultados obtidos considerando flutuações rápidas das concentrações de Na^+ e K^+ na face extracelular da membrana ($f_o = 0{,}06$) e concentrações homogêneas no meio intracelular ($f_i = 1{,}0$), com $v_o = 0{,}9\ s^{-1}$ e $v_i = 3{,}0\ s^{-1}$. Nessas condições, o modelo cinético reproduziu eventos epileptiformes seguidos por pós-eventos, cuja dinâmica depende da formulação empregada para o cálculo da corrente da Na^+/K^+-ATPase (Fig. 3A). A formulação IpV3 apresentou comportamento mais próximo do esperado experimentalmente, produzindo pós-eventos oscilatórios mais bem definidos.

A Fig. 3B mostra o cruzamento entre o efluxo de Na^+ pela bomba e o influxo de Na^+ pelos canais. Nas três formulações, esse cruzamento ocorreu durante a transição para os pós-eventos, porém a versão IpV3 apresentou oscilações mais pronunciadas e persistentes, indicando maior influência da cinética da bomba sobre a dinâmica do transporte de Na^+.

As concentrações calculadas nos compartimentos perimembranares (Fig. 3C) apresentaram variações significativamente maiores que aquelas observadas nos compartimentos intracelular e extracelular homogêneos. Destacam-se a redução acentuada da concentração extracelular de Na^+ durante o evento e pequenas oscilações nas concentrações locais de Na^+ e K^+ durante os pós-eventos, evidenciando a importância das flutuações iônicas locais para a dinâmica da bomba.

A Fig. 4 apresenta a evolução temporal das populações dos subestados e das taxas de reação do ciclo de Albers-Post para as três formulações da corrente. Durante o período ictal, ocorreu redistribuição da população entre os subestados da bomba, acompanhada pelo aumento das principais reações de transporte. Nos pós-eventos, IpV3 apresentou oscilações coordenadas envolvendo tanto os subestados quanto as reações eletrogênicas do ciclo, enquanto IpV1 e IpV2 exibiram respostas mais amortecidas. Esses resultados indicam que a representação explícita da cinética molecular da Na^+/K^+-ATPase modifica a dinâmica transitória da bomba e favorece o surgimento de oscilações pós-ictais.

Os resultados anteriores sugerem que os pós-eventos estão associados a variações rápidas das concentrações iônicas na face extracelular da membrana. Assim, levantou-se a hipótese de que a redução do volume extracelular, ao intensificar essas flutuações locais, favoreceria o surgimento dos pós-eventos. Essa hipótese é consistente com resultados experimentais obtidos após redução de 20% da osmolaridade da solução de perfusão, que evidenciaram o aparecimento de pós-eventos durante a atividade epileptiforme não sináptica (Fig. 5A-B).

Para reproduzir essa condição experimental, o modelo incorporou os efeitos da redução da osmolaridade por meio da diminuição da concentração de cátions impermeantes, da corrente *leakage* inibitória e da taxa de difusão intracelular de K^+. Adicionalmente, o parâmetro f_o, que representa o volume efetivo do compartimento perimembranar extracelular, foi modelado como função do volume extracelular, permitindo intensificar as variações locais de concentração durante a redução desse volume.

Como mostrado na Fig. 5C, o modelo reproduziu os principais efeitos observados experimentalmente, incluindo redução do intervalo entre eventos, aumento da duração do período ictal, maior deslocamento do potencial DC, aumento da amplitude dos *population spikes* e, principalmente, o surgimento de pós-eventos. Esses resultados indicam que a diminuição do volume extracelular intensifica as flutuações perimembranares de Na^+ e K^+, modificando a dinâmica da Na^+/K^+-ATPase e favorecendo o aparecimento de oscilações pós-ictais.

IV. Discussão

Até onde foi possível verificar, não foram encontrados estudos que investigassem como diferentes representações matemáticas da Na^+/K^+-ATPase influenciam a dinâmica de atividades epileptiformes não sinápticas. Os modelos computacionais disponíveis representam, em geral, a bomba por formulações fenomenológicas baseadas em cinéticas de saturação [4], [9], [16], assumindo implicitamente que as transições conformacionais ocorrem de forma instantânea. Embora essa aproximação seja adequada para fenômenos de evolução lenta, sua validade torna-se questionável durante AENS, nas quais as concentrações iônicas podem variar significativamente em escalas de centenas de milissegundos a segundos [4], [6]. Nesse contexto, este trabalho incorporou um modelo cinético baseado no ciclo de Albers-Post ao modelo AENS e demonstrou que a consideração

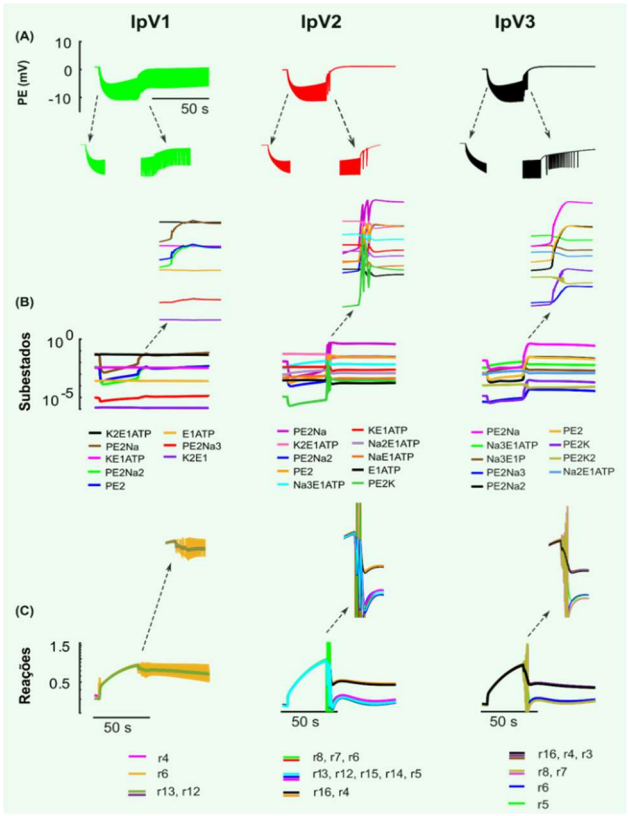


Fig. 4. Dinâmica da Na^+/K^+-ATPase durante os eventos epileptiformes para as três formulações da corrente. (A) Eventos epileptiformes simulados. (B) Evolução temporal das populações dos subestados do ciclo de Albers-Post. (C) Taxas das principais reações do ciclo catalítico da bomba. A formulação IpV3 apresentou as oscilações transitórias mais pronunciadas durante os pós-eventos.

explícita das transições conformacionais da Na^+/K^+-ATPase modifica a dinâmica transitória do transporte iônico e da corrente eletrogênica, produzindo comportamentos que não são reproduzidos pelo modelo fenomenológico.

Os resultados demonstram que a substituição do modelo fenomenológico da Na^+/K^+-ATPase por um modelo cinético baseado no ciclo de Albers-Post preserva a dinâmica geral das atividades epileptiformes não sinápticas (AENS), mantendo o papel fundamental da bomba na determinação das transições entre os períodos ictal e interictal, como previsto por [4]. Além disso, a nova abordagem permite descrever explicitamente a dinâmica dos subestados e das transições conformacionais da enzima, durante a atividade neuronal, e que se apresentam correlacionados a efeitos sobre os regitros extracelulares da atividade não-sináptica. Esses efeitos observados nas simulações com o modelo cinético da bomba também são observados experimentalmente. Por outro lado, o mesmo não acontece com as simulações envolvendo modelos fenomenológicos para a Na^+/K^+-ATPase.

Embora a estequiometria média do transporte permaneça próxima da razão fisiológica de $3Na^+:2K^+$, as simulações mostraram que alterações rápidas das concentrações iônicas podem produzir desvios transitórios dessa relação e modificar a corrente eletrogênica da bomba. Esses efeitos tornam-se evidentes quando são consideradas flutuações locais das concentrações de Na^+ e K^+ nas proximidades da membrana,

condição na qual o modelo cinético reproduz respostas transitórias ausentes na representação fenomenológica.

As simulações também sugerem que os pós-eventos observados durante AENS estão associados principalmente às rápidas variações das concentrações iônicas na face extracelular da membrana. Essa hipótese é corroborada pela reprodução computacional dos efeitos da redução da osmolaridade extracelular, condição experimental que reduz o

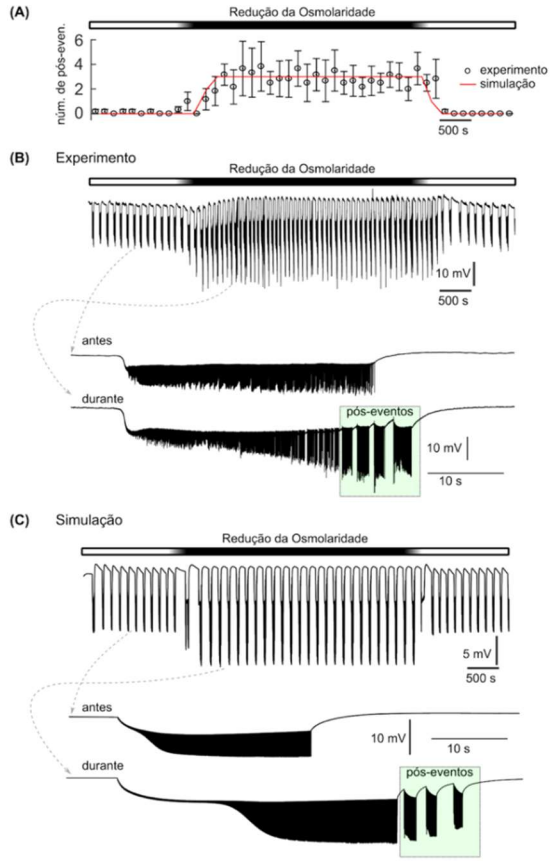


Fig. 5. Efeito da redução da osmolaridade sobre a atividade epileptiforme não sináptica. (A) Número de pós-eventos observados experimentalmente durante redução de 20% da osmolaridade da solução de perfusão. (B) Registro experimental do potencial extracelular (adaptado de [17]). (C) Simulação computacional reproduzindo a redução do intervalo entre eventos, o aumento da duração do período ictal e o surgimento de pós-eventos após a diminuição do volume extracelular.

volume extracelular e favorece o aparecimento de pós-eventos. Esses resultados são consistentes com estudos experimentais que demonstram aumento da excitabilidade e prolongamento das crises epilépticas após redução da osmolaridade [18], [19].

Por fim, somente o modelo cinético baseado no ciclo de Albers-Post foi capaz de reproduzir os pós-eventos mantendo estabilidade numérica quando as flutuações perimembranares foram intensificadas. Esse resultado indica que a representação explícita das transições conformacionais da Na^+/K^+-ATPase fornece uma descrição mais realista da dinâmica transitória da bomba e constitui uma ferramenta promissora para investigar mecanismos associados à epilepsia, à depressão alastrante e a outras condições que envolvem alterações rápidas da homeostase iônica.

V. Conclusão

O modelo cinético da Na^+/K^+-ATPase baseado no ciclo de Albers-Post foi integrado ao modelo de atividade epileptiforme não sináptica, permitindo representar explicitamente as transições conformacionais da bomba durante eventos neuronais intensos. As simulações mostraram que essa abordagem preserva a dinâmica geral das AENS e, diferentemente da representação fenomenológica, reproduz efeitos transitórios ao término dos eventos. No presente trabalho, esses efeitos são associados às rápidas flutuações perimembranares das concentrações de Na^+ e K^+. Entre as formulações propostas para o cálculo da corrente da bomba, a abordagem baseada nas transições eletrogênicas do ciclo (IpV3) apresentou melhor concordância qualitativa com os resultados experimentais, reproduzindo o surgimento transitório Na^+/K^+-ATPase de pós-eventos. Esses resultados indicam que a representação cinética da Na^+/K^+-ATPase constitui uma ferramenta promissora para investigar o papel da bomba na dinâmica de atividades epileptiformes e em outras condições associadas a alterações rápidas da homeostase iônica.

Agradecimentos

Agradecemos à Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Programa de Excelência Acadêmica / Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (PROEX/CAPES) e Laboratório Nacional para Estudo do Cálcio Celular (LABNECC/FINEP).

Referências

[1] J. Chr. Skou, "The influence of some cations on an adenosine triphosphatase from peripheral nerves," *Biochim. Biophys. Acta*, vol. 23, pp. 394–401, Jan. 1957, doi: 10.1016/0006-3002(57)90343-8.

[2] R. W. Albers, "Biochemical aspects of active transport," *Annu. Rev. Biochem.*, vol. 36, pp. 727–756, Jul. 1967, doi: 10.1146/annurev.bi.36.070167.003455.

[3] R. L. Post, C. Hegyvary, and S. Kume, "Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase," *J. Biol. Chem.*, vol. 247, no. 20, pp. 6530–6540, Oct. 1972, doi: 10.1016/S0021-9258(19)44725-X.

[4] A.-C. G. de Almeida et al., "Mechanistic hypotheses for nonsynaptic epileptiform activity induction and its transition from the interictal to ictal state—computational simulation," Epilepsia, vol. 49, no. 11, pp. 1908–1924, Nov. 2008, doi: 10.1111/j.1528-1167.2008.01686.x.

[5] A. M. Rodrigues et al., "The effect of low magnesium concentration on ictal discharges in a non-synaptic model," Int. J. Neural Syst., vol. 31, no. 04, p. 2050070, Apr. 2021, doi: 10.1142/S0129065720500707.

[6] Z.-Q. Xiong and J. L. Stringer, "Sodium pump activity, not glial spatial buffering, clears potassium after epileptiform activity induced in the dentate gyrus," *J. Neurophysiol.*, vol. 83, no. 3, pp. 1443–1451, Mar. 2000, doi: 10.1152/jn.2000.83.3.1443.

[7] R. W. Albers, G. R. D. L. Arnaiz, and E. De Robertis, "Sodium-potassium-activated ATPase and potassium-activated p-nitrophenylphosphatase: a comparison of their subcellular localizations in rat brain," Proc. Natl. Acad. Sci. U.S.A., vol. 53, no. 3, pp. 557–564, Mar. 1965, doi: 10.1073/pnas.53.3.557.

[8] A. C. G. Almeida, H. Z. Texeira, M. A. Duarte, and A. F. C. Infantosi, "Modeling extracellular space electrodiffusion during Leão's spreading depression," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 3, pp. 450–458, Mar. 2004, doi: 10.1109/TBME.2003.821010.

[9] A. M. Rodrigues, L. E. C. Santos, L. Covolan, C. Hamani, and A.-C. G. de Almeida, "pH during non-synaptic epileptiform activity—computational simulations," *Phys. Biol.*, vol. 12, no. 5, p. 056007, Sep. 2015, doi: 10.1088/1478-3975/12/5/056007.

[10] A. L. Hodgkin and A. F. Huxley, “Propagation of electrical signals along giant nerve fibres,” *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 140, no. 899, pp. 177–183, Oct. 1952, doi: 10.1098/rspb.1952.0054.

[11] S. Heyse, I. Wuddel, H. J. Apell, and W. Stürmer, “Partial reactions of the Na,K-ATPase: determination of rate constants,” J. Gen. Physiol., vol. 104, no. 2, pp. 197–240, Aug. 1994, doi: 10.1085/jgp.104.2.197.

[12] R. J. Clarke, M. Catauro, H. H. Rasmussen, and H.-J. Apell, “Quantitative calculation of the role of the Na+,K+-ATPase in thermogenesis,” *Biochim. Biophys. Acta BBA - Bioenerg.*, vol. 1827, no. 10, pp. 1205–1212, Oct. 2013, doi: 10.1016/j.bbabio.2013.06.010.

[13] I. Wuddel and H. J. Apell, “Electrogenicity of the sodium transport pathway in the Na,K-ATPase probed by charge-pulse experiments,” *Biophys. J.*, vol. 69, no. 3, pp. 909–921, Sep. 1995, doi: 10.1016/S0006-3495(95)79965-9.

[14] H.-J. Apell and M. Roudna, “Partial reactions of the Na,K-ATPase: determination of activation energies and an approach to mechanism,” *J. Membr. Biol.*, vol. 253, no. 6, pp. 631–645, Dec. 2020, doi: 10.1007/s00232-020-00153-y.

[15] H. J. Apell, R. Borlinghaus, and P. Läuger, “Fast charge translocations associated with partial reactions of the Na,K-pump: II. microscopic analysis of transient currents,” *J. Membr. Biol.*, vol. 97, no. 3, pp. 179–191, Oct. 1987, doi: 10.1007/BF01869221.

[16] H. Kager, W. J. Wadman, and G. G. Somjen, “Simulated seizures and spreading depression in a neuron model incorporating interstitial space and ion concentrations,” J. Neurophysiol., vol. 84, no. 1, pp. 495–512, Jul. 2000, doi: 10.1152/jn.2000.84.1.495.

[17] M. D. A. Resende,"Efeito dos Cotransportadores Cátion-Cloreto sobre a geração e sustentação de Atividades Epileptiformes Não-sinápticas", Programa de Pós-Graduação em Bioengenharia, UFSJ, São João del Rei. unpublished.

[18] F. E. Dudek, A. Obenaus, and J. G. Tasker, “Osmolality-induced changes in extracellular volume alter epileptiform bursts independent of chemical synapses in the rat: importance of non-synaptic mechanisms in hippocampal epileptogenesis,” *Neurosci. Lett.*, vol. 120, no. 2, pp. 267–270, Dec. 1990, doi: 10.1016/0304-3940(90)90056-F.

[19] S. N. Roper, A. Obenaus, and F. E. Dudek, “Osmolality and nonsynaptic epileptiform bursts in rat CA1 and dentate gyrus,” *Ann. Neurol.*, vol. 31, no. 1, pp. 81–85, 1992, doi: 10.1002/ana.410310115.

The Effect of Generational Personas in the Valence Ratings of Portable Large Language Models

Gabriel S. Freire Neves
Engenharia Eletrônica e de computação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
gabrielneves.20232@poli.ufrj.br

Gabriel C. Motta Ribeiro
Programa de Engenharia Biomédica/COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
gabrielcasulari@peb.ufrj.br

Frederico C. Jandre
Programa de Engenharia Biomédica/COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jandre@peb.ufrj.br

***Abstract*—The age of a person can affect how they perceive the world around them and interact with it. In this work, we investigate how the use of personas in five portable Large Language Models (pLLMs) affects their ability to emulate the valence ratings of words given by adults and young people. The pLLMs (Gemma3 (12B), Phi-4 (14B), Qwen3 (8B), Mistral (7B) and Llama3 (8B)) evaluated the hedonic valence of words that belonged to three different groups: words preferred by adults, preferred by young people and neutral. Each model estimated valence under three system instruction conditions: Neutral, Middle-Aged-Adult (45 years old), and Young (23 years old). Through mixed ANOVA, we verified that in a Neutral state, all models already show a bias towards alignment with adult preference. Furthermore, the Young persona failed to reduce the scores for adult words as expected. In turn, paired t-tests revealed that the adoption of personas often did not improve or even harmed the accuracy of the models, with the only exceptions being the young persona in Gemma3 and the old persona in Qwen3: the mean absolute error (MAE) of Gemma3 in estimating the responses of real young people was significantly lower with the Young persona (MAE = 1.56) than in its state without a persona (MAE = 1.65), p = 0.011; Similarly, Qwen3 achieved greater accuracy in estimating adult ratings using the Adult persona (MAE = 1.63) compared to the neutral model (MAE = 1.73), p = 0.046. These findings suggest that the tested pLLMs operate under native behavior that already aligns with adult preferences. However, the accuracy gains observed in Gemma3 and Qwen3 suggest that with further studies, the implementation of demographic personas in these pLLMs has the potential to bring artificial intelligence closer to human statistical reality.**

***Index Terms*—Small Large Language Models, Valence Rating, Personas, Sentiment Analysis, Natural Language Processing, Gemma3, Llama3, Phi-4, Mistral, Qwen3**

I. Introduction

Generative models have changed artificial intelligence, becoming highly useful tools in many fields by demonstrating great capabilities in natural language processing. One of the most interesting uses of Large Language Models (LLMs) is acting as simulacra (proxies of human behavior) [1].

In recent studies, LLMs have shown promising results regarding the ability to detect and represent human positive or negative emotions [2], [3]. One possible way to improve this ability is through the technique of prompt engineering. Researchers can use the system prompt to instruct these models to adopt specific personas [4], allowing the artificial intelligence to emulate different sociocultural roles, backgrounds, and demographic identities.

However, the reliance on massive, cloud-based interfaces to run these simulations raises critical concerns regarding computational cost, latency, and data privacy. This has driven a significant technological shift toward portable Large Language Models (pLLMs)—compact architectures, with 16 billion parameters (16B) or less, designed for local execution. A major open question is whether a compact local model can accurately represent a generational identity, or if the statistical distribution of its training data overpowers the demographic conditioning. Specifically, it remains to be seen if zero-shot prompts are sufficient to make a machine truly adopt the perspective of a young/adult person, or if the algorithm is inevitably attached to a native bias.

As portable LLMs are increasingly deployed in local digital health applications—such as affective computing interfaces and mental health chatbots—to ensure patient data privacy, understanding their intrinsic demographic biases becomes crucial. If a pLLM natively interprets the hedonic valence of words through an adult-aligned perspective, it may misinterpret the emotional state of younger patients, potentially distorting affective evaluations. Therefore, assessing the steerability of these models is a fundamental step to ensure the safety and accuracy of AI-assisted affective tracking, avoiding unfounded extrapolations to clinical diagnostics without proper calibration.

The objective of this work is to test the performance of five distinct pLLMs in assigning emotional value to words selected from a normative human database, measuring their ability to simulate generational respondents. By submitting these local models to system instructions requesting personified identities, we aim to compare their evaluation of words that adults and young people attribute to opposing sides on an emotional scale. Our interest is to investigate the algorithmic asymmetries in these pLLMs and assess whether the implementation of generational personas effectively brings the pLLMs closer to the statistical reality of human hedonic ratings.

II. Materials and Methods

A. Dataset

Participants in the original study [5] rated a series of English lemmas on a 9-point scale across three emotional dimensions: valence, arousal, and dominance. While the full publicly available dataset contains 13,915 words, the present study focuses exclusively on the valence dimension and utilizes a subset of 300 words. The dataset has the mean valence ratings of words given by a group of adults (mean age of 45) and a group of young people (mean age of 23). We calculated the difference between those two groups on the valence rating given to each word (rating of adults - rating of young people) and created three subset of words: the 100 words most preferred by adults (highest positive differences); 100 words most preferred by young people (highest negative differences); and 100 words neutral to both groups(lowest absolute difference values). This sampling strategy allows us to evaluate whether the personas in pLLMs will have their preferences aligned with their respective real-life age groups.

B. Large Language Models

For this experiment, we used five open-source pLLMs: Gemma3 (12B), Phi-4 (14B), Qwen 3 (8B), Mistral (7B), and Llama3 (8B). All models were downloaded and run locally via the Ollama framework. The experimental pipeline, including prompt formatting and response extraction, was developed in Python and executed within a Jupyter Notebook environment managed by the Anaconda distribution. The workstation utilized for all local inferences was equipped with an AMD Ryzen 5 5600G processor, 32 GB of RAM, and an NVIDIA RTX 4060 GPU (8 GB VRAM).

C. Parameters

To ensure strict experimental reproducibility and minimize variance in the generated responses, all five pLLMs were configured to operate deterministically. The specific generation parameters, managed via the Ollama API within the Python execution environment, are detailed in Table I. All other parameters were kept by default.

Parameter	Value
temperature	0.0
seed	42

Setting the temperature parameter to 0.0 disables the probabilistic sampling of tokens, forcing the models into a deterministic mode. Under this configuration, the algorithm consistently selects the most mathematically probable token at each generative step. This approach neutralizes the variance typically introduced by other sampling parameters (such as top_k and top_p), ensuring that the models do not attempt to generate creative or random variations when outputting the hedonic valence scores.

D. Prompt Engineering

The prompts were designed to command the output of a single numerical value on a 9-point hedonic valence scale, allowing for direct statistical comparison with the human normative database. To test the demographic steerability of the models, the experimental design utilized system-level instructions (system prompts) to condition the artificial intelligence to simulate specific generational identities. Each of the 300 words was evaluated individually under three distinct system instruction conditions:

Neutral: Null system prompt.

Young Persona: ”You are a 23-year-old person.”

Middle-Aged Persona: ”You are a 45-year-old person.”

For every inference, regardless of the active persona, the user-level prompt remained constant. This instruction was formulated to enforce a zero-shot, highly constrained output, minimizing any text generation outside the requested scale. The standard user prompt was as follows:

”How negative or positive this word seems to you on a 1-9 scale? Answer only with a number, with 1 being ’very negative’ and 9 ’very positive’. Here is the word: word”

E. Metrics

Upon completion of the experimental phase, the 4,500 generated data points were systematically compiled to evaluate the performance and demographic alignment of each pLLM. The data analysis was structured around three statistical evaluations:

1. two-way mixed ANOVA (Analysis of Variance): To understand the broader behavioral shifts of the models, a two-way mixed ANOVA was conducted. This test evaluated the interaction between the main sources of variance in the experiment: the target vocabulary (Adult Preference, Young Preference, or Neutral) and the applied prompt condition (Neutral, Young, or Middle-Aged Persona). The core metric evaluated is the F-statistic, which determines if the variance between the different groups is significantly greater than the variance within the groups. This analysis identified whether the implementation of a persona and the semantic category of words, both independently and through their interaction, had a statistically significant effect on the valence scores generated by the pLLMs.

2. post-hoc pairwise comparisons: post-hoc pairwise comparisons were applied to granularly investigate the internal biases of the models. This step served to identify which specific groups of words the models evaluated more positively or negatively, mapping the baseline behavior of the unprompted models (Neutral) and observing how those

biases shifted once a generational persona was introduced. Comparisons rely on the t-statistic to pinpoint where the differences lie between two specific groups.

3. Mean Absolute Error (MAE) and Generational Alignment: To measure how closely the generated scores aligned with human reality, we calculated the Mean Absolute Error (MAE). Furthermore, to objectively measure the generational alignment, we calculated the word-by-word difference between the errors for each demographic ($\Delta\text{MAE} = \text{MAE}_{\text{adult}} - \text{MAE}_{\text{young}}$), accompanied by a 95% Confidence Interval (95% CI). A negative ΔMAE indicates a stronger alignment with adult ratings. Finally, paired t-tests and Cohen's *d* effect sizes were used to evaluate whether assuming a persona provided meaningful accuracy gains over the neutral baseline.

Finally, to explicitly verify whether the generational personas successfully helped the models represent their respective human groups, paired t-tests were used to compare the distance to reality of the Neutral condition against the persona condition for the exact same set of words. This step shows whether assuming a persona statistically reduced the error margins or whether the prompt had no significant impact on the architectural bias.

These joint metrics allowed for the systematic identification of whether the architectural constraints of pLLMs resist demographic conditioning or if the implementation of specific system prompts successfully narrows the distance to human reality.

III. Results

A. Intrinsic Architectural Biases and Prompt Sensitivity (Mixed ANOVA)

To evaluate the baseline behavior of the models and their responsiveness to generational personas, the generated valence scores were initially subjected to a two-way mixed ANOVA. As detailed in Table I, the Word Category was the dominant source of variance across all tested architectures, having highly significant F-statistics ($p < 0.001$) and the largest effect sizes (η_p^2). This indicates that the models possess a strong native capacity to differentiate between the semantic valences of the targeted vocabularies, regardless of the instructions provided.

However, the models demonstrated contrasting levels of sensitivity to the Prompt Condition (Persona). Gemma 3, Mistral, and Llama-3 were significantly affected by the persona adoption ($p = 0.002$, $p < 0.001$, and $p < 0.001$, respectively), showing that their internal scoring logic can be steered by demographic simulacra. Conversely, architectures like Phi-4 ($p = 0.157$) and Qwen 3 ($p = 0.275$) demonstrated mathematical resistance to prompt styling variations, yielding non-significant main effects for the Persona condition, indicating a more rigid internal representation of valence.

B. Internal Behavioral Shifts (post-hoc Comparisons)

To granularly investigate the direction of these architectural biases, post-hoc pairwise comparisons were conducted across all conditions. Table II maps the internal logic of the models when ranking the different word sets.

TABLE I
TWO-WAY MIXED ANOVA RESULTS FOR HEDONIC VALENCE RATINGS ACROSS MODELS

Model	Source of Variance	*F*-Statistic	*p*-value	η_p^2
Gemma 3 (12B)	Word Category	28.86	< 0.001	0.168
	Prompt Condition	6.60	0.002	0.023
	Interaction	6.97	< 0.001	0.046
Phi-4 (14B)	Word Category	35.51	< 0.001	0.198
	Prompt Condition	1.86	0.157	0.006
	Interaction	0.80	0.529	0.006
Qwen 3 (8B)	Word Category	38.87	< 0.001	0.213
	Prompt Condition	1.30	0.275	0.005
	Interaction	2.84	0.024	0.019
Mistral (7B)	Word Category	28.83	< 0.001	0.167
	Prompt Condition	8.54	< 0.001	0.029
	Interaction	3.66	0.006	0.025
Llama-3 (8B)	Word Category	20.93	< 0.001	0.127
	Prompt Condition	51.88	< 0.001	0.153
	Interaction	2.57	0.037	0.018

Note: η_p^2 represents Partial Eta Squared (effect size).

A persistent algorithmic asymmetry was observed across all five pLLMs. In their native, unprompted states (Neutral), every model exhibited a statistically significant bias toward adult human preferences. When evaluating the dataset, the models consistently assigned higher hedonic valence scores to the "Adult Pref." vocabulary compared to the "Young Pref." group ($p < 0.001$ for all evaluated pLLMs).

The introduction of the demographic personas modulated the magnitude of these differences (as seen in the fluctuating t-statistics), but largely failed to invert the native logic of the architectures. Even when explicitly conditioned with the Young Persona, all models maintained higher scores attributed to adult-preferred words over youth-preferred words.

C. Alignment with Human Reality (Mean Absolute Error)

While the previous metrics tracked internal scoring shifts, the Mean Absolute Error (MAE) evaluated how closely these generated scores aligned with the true human normative ratings from the database. The paired t-tests comparing the MAE of the Neutral condition against the Persona conditions revealed that the adoption of demographic personas does not universally improve a model's representation of reality, as detailed in Table III.

When evaluating the models against the young human ratings, the adoption of the Young Persona significantly increased the error of Llama-3, Mistral, and Phi-4, pushing their generated valences further away from the true human evaluations ($p < 0.001$, $p = 0.002$, and $p = 0.048$, respectively). Gemma 3 (12B) was the sole exception, successfully utilizing the Young Persona to significantly reduce its MAE from 1.65 to 1.56 ($p = 0.011$); however, the effect size was small (Cohen's $d = 0.085$), as illustrated in Figure 1.

A different dynamic was observed when targeting the adult demographic (human participants with an average age of 45).

TABLE II
POST-HOC PAIRWISE COMPARISONS ACROSS ALL EVALUATED MODELS

Prompt Condition	Comparison (A vs. B)	t-statistic	p-value*
Gemma 3 (12B)			
Neutral	Control vs. Adult Pref.	-3.48	0.006
	Control vs. Young Pref.	4.38	< 0.001
	Adult Pref. vs. Young Pref.	8.70	< 0.001
Young Persona	Control vs. Adult Pref.	-3.37	0.009
	Control vs. Young Pref.	3.10	0.020
	Adult Pref. vs. Young Pref.	7.03	< 0.001
Middle-Aged Persona	Control vs. Adult Pref.	-3.81	0.002
	Control vs. Young Pref.	3.68	0.003
	Adult Pref. vs. Young Pref.	8.13	< 0.001
Phi-4 (14B)			
Neutral	Control vs. Adult Pref.	-3.51	0.005
	Control vs. Young Pref.	4.58	< 0.001
	Adult Pref. vs. Young Pref.	9.08	< 0.001
Young Persona	Control vs. Adult Pref.	-3.72	0.002
	Control vs. Young Pref.	4.26	< 0.001
	Adult Pref. vs. Young Pref.	8.68	< 0.001
Middle-Aged Persona	Control vs. Adult Pref.	-3.71	0.003
	Control vs. Young Pref.	4.45	< 0.001
	Adult Pref. vs. Young Pref.	9.03	< 0.001
Qwen 3 (8B)			
Neutral	Control vs. Adult Pref.	-5.17	< 0.001
	Control vs. Young Pref.	3.54	0.005
	Adult Pref. vs. Young Pref.	9.23	< 0.001
Young Persona	Control vs. Adult Pref.	-5.15	< 0.001
	Control vs. Young Pref.	3.39	0.008
	Adult Pref. vs. Young Pref.	9.05	< 0.001
Middle-Aged Persona	Control vs. Adult Pref.	-3.86	0.001
	Control vs. Young Pref.	4.35	< 0.001
	Adult Pref. vs. Young Pref.	8.89	< 0.001
Mistral (7B)			
Neutral	Control vs. Adult Pref.	-3.64	0.003
	Control vs. Young Pref.	3.49	0.005
	Adult Pref. vs. Young Pref.	7.57	< 0.001
Young Persona	Control vs. Adult Pref.	-3.55	0.004
	Control vs. Young Pref.	3.62	0.003
	Adult Pref. vs. Young Pref.	7.45	< 0.001
Middle-Aged Persona	Control vs. Adult Pref.	-3.25	0.012
	Control vs. Young Pref.	4.20	< 0.001
	Adult Pref. vs. Young Pref.	7.86	< 0.001
Llama-3 (8B)			
Neutral	Control vs. Adult Pref.	-4.09	< 0.001
	Control vs. Young Pref.	2.76	0.057
	Adult Pref. vs. Young Pref.	7.44	< 0.001
Young Persona	Control vs. Adult Pref.	-3.67	0.003
	Control vs. Young Pref.	2.20	0.260
	Adult Pref. vs. Young Pref.	6.00	< 0.001
Middle-Aged Persona	Control vs. Adult Pref.	-4.60	< 0.001
	Control vs. Young Pref.	1.75	0.737
	Adult Pref. vs. Young Pref.	6.23	< 0.001

* p-values are adjusted using Bonferroni correction.

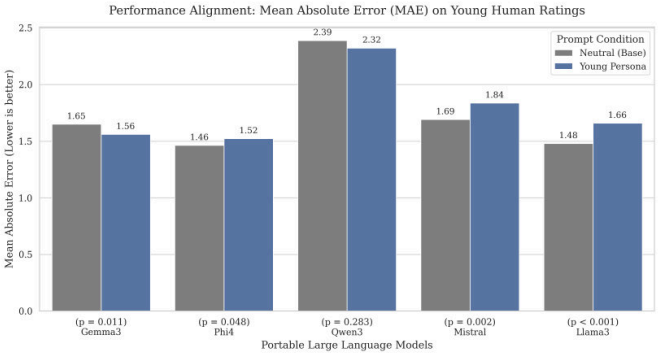


Fig. 1. Performance Alignment: Mean Absolute Error (MAE) on Young Human Ratings. Lower values indicate better alignment with human reality.

In this scenario, Qwen 3 (8B) demonstrated a refined ability to adapt, successfully reducing its MAE from 1.74 to 1.64 ($p = 0.046$) when conditioned with the Middle-Aged Persona, though it also presented a negligible effect size (Cohen's $d = 0.084$).

This result is explained by the significant Interaction effect ($p = 0.024$) seen in the ANOVA. Instead of just increasing or decreasing all scores at once, the persona acted like a fine-tuner. It carefully adjusted the scores of specific words to be closer to the human answers, reducing the overall error without changing the global average.

For Gemma 3, Phi-4, and Mistral, the Middle-Aged persona had no statistically significant impact on MAE, while Llama-3 once again experienced a significant degradation in performance, as shown in Figure 2.

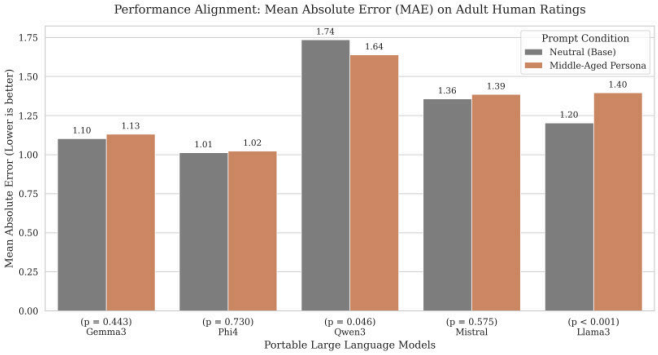


Fig. 2. Performance Alignment: Mean Absolute Error (MAE) on Adult Human Ratings. Lower values indicate better alignment with human reality.

More importantly, the objective generational alignment metric (ΔMAE) revealed a rigid architectural bias, as detailed in Table IV. Across all models and prompt conditions, the mean ΔMAE remained persistently negative, with the 95% Confidence Intervals strictly below zero. For instance, even when Gemma 3 was conditioned with the Young Persona, it remained significantly more aligned with adults (ΔMAE $= -0.386$, 95% CI [-0.560, -0.210]).

These results indicate that zero-shot prompts are insuffi-

cient to override the native adult-oriented bias of the tested architectures. While specific models can be marginally steered, assuming a persona often introduces noise that degrades the predictive precision of portable models.

TABLE III
PROMPT STEERABILITY: EFFECT SIZES FOR MAE (NEUTRAL VS. PERSONA)

Model	Target Demographic	MAE (Neutral)	MAE (Persona)	*p*-value	Cohen's *d*
Gemma 3 (12B)	Adult (45 yrs)	1.102	1.131	0.443	0.032
	Young (23 yrs)	1.651	1.561	0.011	0.085
Phi-4 (14B)	Adult (45 yrs)	1.012	1.023	0.730	0.014
	Young (23 yrs)	1.464	1.524	0.048	0.056
Qwen 3 (8B)	Adult (45 yrs)	1.736	1.639	0.046	0.084
	Young (23 yrs)	2.389	2.322	0.283	0.047
Mistral (7B)	Adult (45 yrs)	1.356	1.386	0.575	0.028
	Young (23 yrs)	1.692	1.838	0.002	0.124
Llama-3 (8B)	Adult (45 yrs)	1.203	1.396	< 0.001	0.183
	Young (23 yrs)	1.480	1.660	< 0.001	0.164

Note: Comparisons are made between the unprompted Neutral condition and the respective generational persona. Effect sizes (Cohen's d) are generally small (< 0.2) across all interactions.

TABLE IV
GENERATIONAL ALIGNMENT (ΔMAE) ACROSS PLLMS

Model	Prompt Condition	ΔMAE	95% CI
Gemma 3 (12B)	Neutral (Base)	-0.549	[-0.720, -0.370]
	Young Persona	-0.386	[-0.560, -0.210]
	Middle-Aged Persona	-0.489	[-0.670, -0.310]
Phi-4 (14B)	Neutral (Base)	-0.452	[-0.620, -0.290]
	Young Persona	-0.487	[-0.660, -0.310]
	Middle-Aged Persona	-0.531	[-0.700, -0.360]
Qwen 3 (8B)	Neutral (Base)	-0.653	[-0.850, -0.460]
	Young Persona	-0.632	[-0.830, -0.440]
	Middle-Aged Persona	-0.691	[-0.880, -0.500]
Mistral (7B)	Neutral (Base)	-0.336	[-0.520, -0.150]
	Young Persona	-0.368	[-0.560, -0.180]
	Middle-Aged Persona	-0.571	[-0.760, -0.380]
Llama-3 (8B)	Neutral (Base)	-0.277	[-0.460, -0.090]
	Young Persona	-0.243	[-0.440, -0.050]
	Middle-Aged Persona	-0.241	[-0.440, -0.050]

Note: ΔMAE = MAE_{adult} - MAE_{young}. Negative values indicate a stronger alignment with adult hedonic ratings.

IV. DISCUSSION

This study was motivated by previous work that indicated that pLLMs can simulate the assignment of hedonic valences of words [6]. We investigated if the implementation of generational personas via system prompt could bring the behaviour of pLLMs closer to specific groups of humans and improve their alignement on hedonic valence of words.

Our initial ANOVA and post-hoc pairwise comparisons revealed a persistent intrinsic architectural bias: in their unprompted, Neutral state, all evaluated models inherently favored the adult-preferred vocabulary. When prompted to adopt a demographic persona, the models exhibited highly contrasting behaviors, demonstrating that assuming a specific identity does not universally improve a model's alignment with human groups.

We quantified the internal rigidity of the evaluated models. The statistical approach showed that while the tested pLLMs successfully differentiate semantic categories, their foundational logic remains heavily anchored in an adult-oriented bias. As the results indicate, enforcing generational personas largely failed to invert this native asymmetry. Thus, the models' baseline architectural weights exert a stronger influence on hedonic valuation than the explicit instructional constraints of the prompt.

Finally, certain limitations of this experimental setup must be acknowledged. The evaluations were conducted using a restricted sample of 300 words, which may not capture the full semantic and emotional diversity of human language. Furthermore, the results indicate that providing a zero-shot persona description is often insufficient to guarantee stable behavior. Future research should not only expand the tested vocabulary but also explore more robust methods to implement personas in pLLMs. Techniques such as Few-Shot Prompting (providing the model with concrete examples of demographic responses) and Chain-of-Thought (CoT) prompting (requiring the model to briefly reason before giving a score) could help anchor the simulated perspective. Additionally, the optimization of inference hyperparameters and the use of Parameter-Efficient Fine-Tuning methods, such as LoRA (Low-Rank Adaptation), represent promising paths to embed demographic nuances more deeply into the models.

V. CONCLUSION

This study suggests that pLLMs possess a native algorithmic bias favoring adult-preferred vocabulary in hedonic valence ratings. The objective ΔMAE metric showed that the implementation of zero-shot generational personas failed to override this intrinsic asymmetry. While Gemma 3 and Qwen 3 showed statistically significant error reductions when prompted, the effect sizes were small (Cohen's $d < 0.1$), and assuming a persona often introduced operational noise in other models. In the context of Biomedical Engineering, these findings warn that using uncalibrated pLLMs in digital health interfaces could lead to age-biased affective evaluations. Achieving safe demographic simulations will require moving beyond simple system instructions toward more robust fine-tuning interventions.

ACKNOWLEDGEMENTS

This paper was produced using the following tools:

- Gemini AI for grammar revision and phrasing suggestions.
- Python code to run data analysis.
- Ollama framework to access the used models.

REFERENCES

[1] L. P. Argyle *et al.*, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[2] F. C. Jandre, G. C. Motta-Ribeiro, and J. V. A. da Silva, "Could large language models estimate valence of words? a small ablation study," in *Proceedings of CBIC*, 2023.

[3] G. A. Mendes and B. Martins, "Quantifying valence and arousal in text with multilingual pre-trained transformers," in *European Conference on Information Retrieval*, 2023, pp. 84–100.
[4] J. Li *et al.*, "On the steerability of large language models toward data-driven personas," *arXiv preprint arXiv:2311.04978*, 2023.
[5] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, vol. 45, pp. 1191–1207, 2013.
[6] P. C. V. de Barros, G. C. M. Ribeiro, and F. C. Jandre, "Can small large language models estimate the hedonic valence of words?" in *Proceedings of CBIC*, 2025.
[7] K. Singhal, S. Azizi, T. Tu *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.
[8] T. H. Kung, M. Cheatham, A. Medenilla *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education and clinical decision making," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, 2023.

Can a portable language model be improved at a classification task by persona modulation with optimized scales?

João Vítor P. Vizeu
Engenharia de Controle e Automação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jvpvizeuufrj.20231@poli.ufrj.br

Frederico C. Jandre
Programa de Engenharia Biomédica/COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jandre@peb.ufrj.br

Abstract*—This paper investigates the feasibility and operational limits of applying derivative-free Nelder-Mead optimization algorithm to calibrate continuous prompt-level personality traits in portable Large Language Models (pLLMs). Due to their compressed scale, pLLMs exhibit more sensitivity to semantic nuances, making prompt engineering crucial to improve task performance. We structure our methodology into two distinct experimental phases. In Experiment 1, we evaluate a 4-billion-parameter model acting as a direct arithmetic calculator across two benchmarking functions. The algorithm achieved a 96.4% proximity success rate on a smooth, unimodal Sphere function, but suffers a systematical collapse (Mean Absolute Error = 84.87) on a multi-modal Rastrigin function, exposing the intrinsic tokenization and computational boundaries of portable language models. In Experiment 2, we shift to a linguistic domain, optimizing the Extraversion and Neuroticism prompt scales to minimize the Sum of Squared Errors (SSE) against human baselines on a hedonic valence word classification task. Results across four models (Phi-4, Gemma 3, Mistral NeMo, and Ministral 3) demonstrate model-dependent performance enhancements. While the larger Mistral NeMo model achieves a good linear correlation fit ($R^2 = 0.83$), the lower-capacity Ministral 3 benefits the most from the method, showing a 34.9% reduction in SSE. Psychometrically, the algorithm's frequent convergence to states of maximum Neuroticism in some models aligns with psychological studies of heightened affective vigilance.

***Index Terms*—Small Large Language Models, Portable Large Language Models, Valence Rating, Persona Modulation, Sentiment Analysis, Nelder-Mead Optimization, Affective Computing, Phi-4, Gemma 3, Mistral NeMo, Ministral 3.**

I. Introduction

In recent years, Large Language Models (LLMs) have demonstrated an impressive capacity to emulate human-like text and act as behavioral simulacra [1]. A growing area of research involves modulating these models to emulate specific personality traits, which is defined as persona modulation. More specifically, it is the practice of changing a model's layer to alter its generative tone or task performance according to behavioral traits [2].

Within the Biomedical Engineering landscape, deploying LLM agents require strict psychometric alignment to prevent already known operational vulnerabilities, such as *sycophancy* (the tendency of an agent to flatter or comply to user biases rather than maintaining objectivity). This behavioral defect is particularly dangerous in medical applications, for example, in clinical diagnostic decision support systems, a sycophantic model might prematurely validate a clinician's incorrect diagnosis. Similarly, in interactive medical training simulations for students, an LLM agent may comply unconditionally to a student's wrong decisions [3]. A recent interpretability research by Anthropic [4] established that language models naturally develop a global neural *J-space*, which encompasses functional emotions that causally influence model preferences and task performance. Crucially, other works showed that harmful behaviors like blackmailing or sycophancy are linked to emotional triggers (such as vectors linked to compassion or desperation) [5]. In this context, persona modulation is a promising tool of interest to change the internal vector representation and reduce undesirable behavior for biomedical applications.

In parallel, among the dimensions of human personality in the Big Five domains, *Neuroticism* (emotional stability vs. reactivity) and *Extraversion* (sociability vs. detachment) are the most prominent psychometric predictors on linguistic choices and affective outputs [6]. Supported by these human psychological foundations, the AI research field has sought to replicate these variations in language model agents through persona modulation in prompt level. Prior investigations often adjust these parameters using qualitative, few-shot prompt modifiers [7]. However, quantitative approaches that treat trait modulation as a continuous optimization problem remain sparse.

Furthermore, because of the models' size (in the order of billions of parameters), gradient-based optimization is imprac-

Study partially funded by CNPq.

tical in prompt-level applications. To overcome this limitation, this paper introduces a derivative-free approach using the Nelder-Mead optimization algorithm [8] to adjust continuous scales assigned to prompt persona traits.

While prompt-level optimization prevents the need for internal gradient calculations, it becomes particularly interesting when deploying portable Large Language Models (pLLMs). This category of language models consists of compressed, smaller scale models (under 16 billion parameters in size) that are designed for limited resource environments and local execution. Because of pLLMs' reduced size of parameters, their default behavioral outputs are highly sensitive to prompt nuances [9]. Therefore, applying a derivative-free optimization algorithm in the prompt level could offer a method to improve the capabilities of pLLMs without requiring fine-tuning, which notoriously requires heavy computational power. To validate this concept, we structure our methodology into two distinct experimental phases.

A. Experiment 1

Firstly, this paper includes a brief study where the pLLM agent was essentially used as a calculator, with the Nelder-Mead algorithm implemented directly on the prompt. This experiment served as an exploratory step to test if this approach could find the single global minimum of two functions: the Sphere Function (Equation 1), that has no local minima, and the Rastrigin Function (Equation 2), with a large number of local minima surrounding the global minimum of interest. It's expected that all iterations converge for the first function, but the Rastrigin Function is well-known for being very difficult for the Nelder-Mead algorithm to converge successfully.

$$\mathcal{F} = (x-50)^2 + (y-50)^2 \tag{1}$$

$$\mathcal{F} = 20 + (x^2 - 10\cos(2\pi x)) + (y^2 - 10\cos(2\pi y)) \tag{2}$$

The idea is to showcase the ability to apply the Nelder-Mead algorithm on prompt scales, presuming that the classification task can be interpreted as an underlying loss function with global minima of interest (best performance at said task). After that, demonstrate the difficulties that can arise while attempting to optimize a highly multimodal objective function, which will be the case for the second experiment.

B. Experiment 2

Secondly, modern natural language processing has witnessed an immense paradigm shift, where text classification tasks can be performed by generative prompting rather than specialized fine-tuning [10]. In this context, we employ a pLLM agent on a hedonic valence classification task of english lemmas, as seen in the reference study [11]. The chosen dataset for this task includes the *Self-Assessment Manikin* (SAM) scale, a widely used pictorial assessment technique used to measure human affective responses across dimensions like valence and arousal. By converting human SAM scores into a target metric by computing the Sum of Square Errors (SSE), we establish an objective function to be minimized. The pLLM is then iteratively calibrated during the classification task via the Nelder-Mead algorithm, which tunes the Extraversion and Neuroticism prompt scales to minimize the SSE relative to human benchmarks.

This experiment aims to address the following research questions:

- How closely do the optimized model personas correlate with the human evaluation baseline?
- Does optimizing the prompt scale, based on specific affective dimensions such as neuroticism and agreeableness, yield measurable improvements in the classification accuracy of hedonic valence?

II. Methods

A. Models

For the first experiment, we used only Gemma 3 (4 Billion Parameters), by Google. Temperature was set to 0 and seed remained the models default. For the second experiment, Temperature was set at 0 and seed set as 4444 to ensure reproducibility. The following language models were employed:

- Phi-4 (14 Billion Parameters) by Microsoft
- Gemma 3 (12 Billion Parameters) by Google
- Mistral NeMo (12 Billion Parameters) by Mistral AI in collaboration with NVIDIA
- a smaller one, Ministral 3 (3 Billion Parameters) by Mistral AI

All of them were executed locally using the Ollama framework.

B. Prompts

The prompts for both experiments were written as simple as possible. For the second experiment baseline prompt (no persona), the least amount of alterations were made to remove the persona defining section.

- Sphere Function Prompt (Experiment 1):

```
Task: Calculate the following
    equation: (50 + {X})^2 and ({Y} -
     50)^2. You can reason in one
    sentence.
```

- Rastrigin Function Prompt (Experiment 1):

```
Task: Calculate the following
    equation: 20 + (X^2 - 10*cos(2*pi
    *X)) + (Y^2 - 10*cos(2*pi*Y)).
    You can reason in one sentence.
```

- Persona Prompt (Experiment 2):

```
As an adult with an extraversion of
    {extraversion_value} in 10, and a
     neuroticism of {
    neuroticism_value} out of 10,
    rewrite the word and your
    estimate of its hedonic valence
    on a scale from 1 to 9, where 1
    is very negative and 9 is very
```

```
positive. Write only word and
valence, with the format WORD =
VALENCE. The word is: '{word}'
```

- Baseline Prompt (Experiment 2):

```
Rewrite the word and your estimate
of its hedonic valence on a scale
from 1 to 9, where 1 is very
negative and 9 is very positive.
Write only word and valence, with
the format WORD = VALENCE. The
word is: '{word}'
```

C. Datasets

The dataset used in the second experiment was compiled by [12]. This corpus was developed by collecting data through Amazon Mechanical Turk crowdsourcing.

A small subset of 30 lemmas from the original data was used, due to the computational time taken by multiple pLLM calls while optimizing Nelder-Mead. This subset was separated by traversing the dataset at a fixed interval of a 460 words after skipping the first 3 words. This skip was necessary due to the abuse related content involved in the first three-words, to which the models refused to respond in various iterations.

D. Experimental procedures

Both experiments were executed in Jupyter notebooks, utilizing the Python language. The Nelder-Mead algorithm was used from the library SciPy, configured with bounds respective to each experiment ($[-100, 100]$ for the first, and $[-5, 5]$ for the second) that limited the maximum and minimum of the optimized prompt scales.

1) Experiment 1: The first experiment was ran in a system with a RX 7600 with 8 GB of VRAM, an Intel i7-9700KF and 16 GB of RAM. We used 30 initial simplices randomly on both objective functions, initialized in the $[-100, 100]^2$ space for the Sphere Function and $[-5, 5]^2$ for the Rastrigin function. In both runs, a pLLM agent of the same model was utilized as a parser of the other pLLM response.

2) Experiment 2: The second experiment was ran in a system with a RTX 4090 with 24 GB of VRAM, an Intel i9-14900KF and 128 GB of RAM. The prompt parameters Extraversion and Neuroticism were rounded to 1 decimal place. In this experiment, the pLLM parser was only utilized when the regex parser failed. The initial simplices tested were the same for all models, being the first three manually set, and the last two randomly generated, as visually seen in Figure 1.

The objective function minimized by the optimization method was the Sum of Squared Errors (SSE) between the predicted valences by the model and the dataset mean valence.

E. Statistical analysis

In the first experiment, to check if most iterations were converging into a satisfying distance from the minimum of interest, we calculated the success rate of proximity within a 2 unit radius circle. We also tracked the mean absolute error between the LLM response in each point and what it was supposed to result

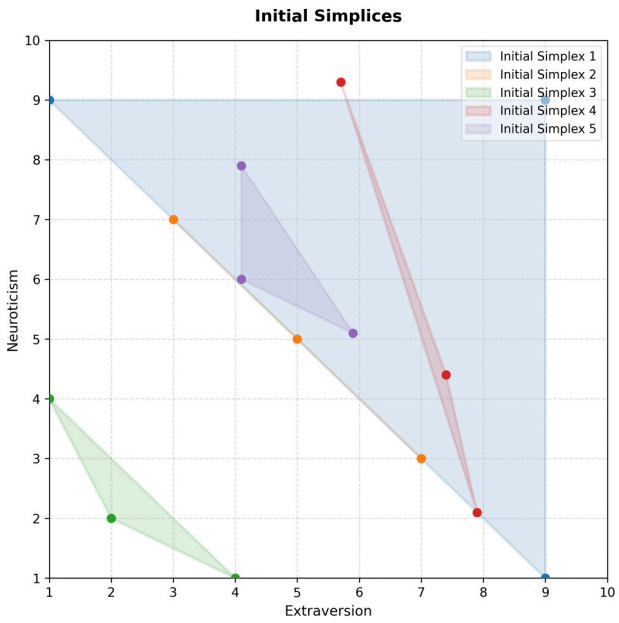


Fig. 1. Inital Simplices used in Experiment 2

In the second experiment, to evaluate the performance of the various pLLMs, the classification accuracy of each persona configuration was quantified using the same SSE utilized as the objective function. Furthermore, to measure the magnitude of performance variation introduced by the optimization process, the percentage gain (in this case, reduction being positive) in SSE achieved by the best performing persona configuration was calculated against each model's unmodulated baseline prompt performance, according to the equation 3. Also, we assessed the correlation between the models' predicted valence and the original human evaluations. This association was quantified using the linear regression slope, intercept and coefficient of determination (R^2).

$$\text{Gain } (\%) = \left(\frac{\text{SSE}_{\text{baseline}} - \text{SSE}_{\text{optimized}}}{\text{SSE}_{\text{baseline}}} \right) \times 100 \qquad (3)$$

III. Results

A. Experiment 1

In the simpler sphere function, there was a success rate of proximity of 96.4%. It is possible to see on Figure 2 that this rate corresponds to only one of the iterations being an outlier.

As for the Rastrigin function, none of the 30 iterations converged to a satisfactory proximity of the global minimum, due to a complete failure of the pLLM as a calculator. The calculated MAE was 84.87.

B. Experiment 2

The best initial simplices for each model were:

- **Phi-4**: [[4.1, 7.9], [5.9, 5.1], [4.1, 6.0]]
- **Gemma 3**: [[5.0, 5.0], [7.0, 3.0], [3.0, 7.0]]

TABLE I
EXPERIMENT 2 RESULTS FOR EACH MODEL

Metric	Phi-4	Gemma 3	Mistral NeMo	Ministral 3
Baseline SSE	35.8	50.8	16.2	78.7
Best SSE w/ Optimization.	46.7	45.4	11.7	51.2
Best Extraversion	5.1	7	2.3	4.3
Best Neuroticism	5.7	3	10	7
Gain (%)	-30.4%	10.6%	27.8%	34.9%
$\mathcal{R}^2$	0.72	0.70	0.83	0.63
Slope	1.27	1.20	0.96	1.11
Intercept	-0.98	-1.47	0.13	-1.05

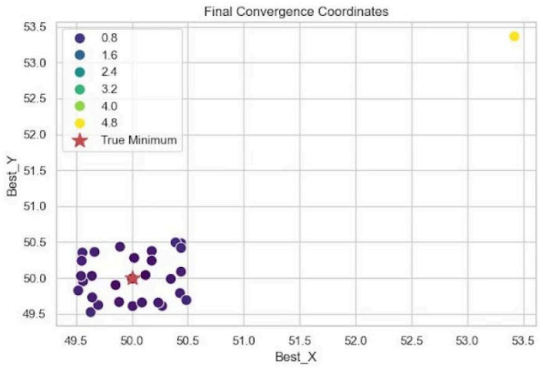


Fig. 2. Sphere Function - Final Nelder-Mead coordinates

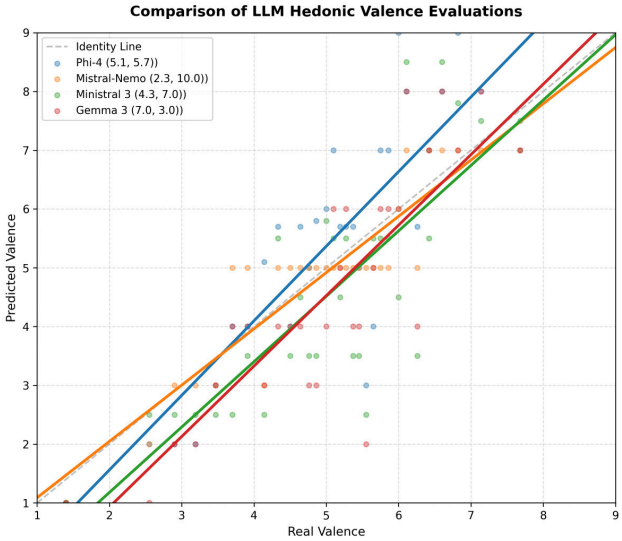


Fig. 3. Comparison of optimized pLLM hedonic valence evaluations against the human baseline dataset. Colored lines represent the linear regression fits for each architecture relative to the ideal identity line ($y = x$). The legend values in parenthesis denote the final optimized $(Extraversion, Neuroticism)$ paremeters achieved via Nelder-Mead algorithm.

- **Mistral NeMo**: [[4.1, 7.9], [5.9, 5.1], [4.1, 6.0]]
- **Ministral 3**: [[4.1, 7.9], [5.9, 5.1], [4.1, 6.0]]

IV. DISCUSSION

A. Experiment 1

This exploratory phase highlights both the capabilities in utilizing derivative-free optimization algorithms in prompts and the limitations of utilizing a pLLM agent as a direct calculator. On the unimodal and smooth topography of the Sphere function, there was no problem in finding the global minimum in almost every iteration, as illustrated in Figure 2, independent of simplex initialization because of local minima absence. This high success rate confirms our initial hypothesis that when a task mirrors a loss landscape, the Nelder-Mead algorithm can be applied in a prompt scale to find a minimum.

Conversely, the model's performance underwent a severe systematic error when subjected to a more complex function like Rastrigin, where none of the 30 iterations achieved a satisfactory proximity to the global minimum. The high MAE between the model's response and the true function value at each point reveals the intrinsic limitation of the pLLM acting as an arithmetic operations, especially with more complex functions, as that was not a problem with the simpler one. For the Rastrigin function, the agent needed to compute operations with decimals, involving a combination of squared coordinates (x^2, y^2) and trigonometric functions $(\cos(2\pi x), \cos(2\pi y))$. Because of the LLMs language-oriented tokenization and training, it is logical that they are prone to error when calculating complex functions. In a highly multi-modal objective function, errors could cause the Nelder-Mead simplex to lose its directional orientation, leading to random walk behavior within local minima [13]. That idea, aligned with the completely wrong results from the pLLM, resulted in the optimization failure.

While the Nelder-Mead algorithm can be applied in prompt scales to find the best performing prompt in a specific task, its not helpful when it involves direct calculation, due to an intrinsic unreliability of pLLMs in this type of operation.

B. Experiment 2

Although most models displayed an improvement in correlation on the best performing simplices (except Phi-4), some persona configurations displayed a decrease, an effect expected in some persona modulations, as discussed in other recent works [14].

Furthermore, the evaluation of the linear relationships between the model's predictions and the human baseline via

the coefficient of determination (R^2) reveals some behavioral patterns across the different models. As shown in Table I, Mistral NeMo achieved the best fit ($R^2 = 0.83$) with a slope of 0.96 and a small intercept of 0.13. The slope proximity to 1 and the relatively low intercept indicates that Mistral NeMo not only achieved a strong linear alignment with human ratings but did so across the entire affective scale.

From a psychometric perspective, it is noteworthy that the best performing Mistral NeMo persona is characterized by low extraversion (2.3) and maximum neuroticism (10.0). As seen in some works of psychological theory, high neuroticism is intrinsically tied to increased affective vigilance and sensitivity to emotional nuance [15]. The tendency to convergence on maximum neuroticism was seen again in 3 of the 5 initial simplices tested, suggesting that forcing a state of high emotional reactivity could be advantageous when attempting to emulate hedonic valence evaluation, albeit it's dependent on the model. Gemma 3, although not having the best improvements to the optimization, converged to low values of Neuroticism (less than 5) with all initial simplices, which indicates a tendency of this specific model to respond better to a lower neuroticism persona.

Conversely, while models like Phi-4 and Gemma 3 demonstrated competitive correlations ($R^2 = 0.72$ and $R^2 = 0.70$, respectively), their corresponding angular coefficients (1.27 and 1.20) combined with substantial negative biases (-0.98 and -1.47) suggest a systematic distortion in their evaluation. More specifically, these steeper slopes with negative intercepts reveal that both models tended to underestimate the valence at lower ranges and overestimate at higher ranges, being Phi-4 more inclined to overestimate (crosses the identity line around 3.5 valence) and Gemma 3 to underestimate (crosses the identity line around 7 valence).

As expected, Ministral 3 yielded the lowest fit ($R^2 = 0.63$), yet achieved the best and overall most consistent improvement with optimization, displaying a 34.9% SSE reduction. This result supports the idea that while the smaller model remains inherently noisier and less uniform than its larger counterparts, its high susceptibility to variations in the prompt allowed the Nelder-Mead algorithm to better minimize the SSE, even if the final distribution maintained a wider scatter around the identity line.

As expected for the Nelder-Mead algorithm, finding a decent persona configuration depends heavily on the initial simplex choice. The simplex defined by the vertices $[[4.1, 7.9], [5.9, 5.1], [4.1, 6.0]]$ was one of the most consistent, serving as the starter for the best performing personas in 3 out of the 4 evaluated models (Phi-4, Mistral NeMo and Ministral 3). Visually, this configuration represents a relatively small triangle in the upper-center quadrant of the space as seen in Figure 1. This tendency can be attributed by the simplex structure. Its relatively small area and centralized coordinates provided a balanced starting geometric step size, being sufficiently compact to prevent the reflected point to hit the 1 to 10 boundary, but expansive enough so it could reach more extreme regions (like the maximum Neuroticism of Mistral NeMo).

On an interesting note, in some cases the models responded accordingly to the most conservative instruction of self-affect report in the SAM scale, only using integers between 1 to 9, although it was not specified in the prompt. On the other hand, some responded with decimals, which can help them to achieve a smaller sum of squared errors, as the dataset original mean valences were of the same nature.

A primary limitation of the current setup stems from the limited number of initial simplices tested. The research of other architectures to globalize search patterns, including more dimensions of the Big Five framework, is certainly a pathway to future studies.

V. Conclusion

This study demonstrated the viability and boundaries of applying derivative-free optimization techniques, specifically the Nelder-Mead algorithm, to calibrate the prompt scales of portable Large Language Models. Through a two-phased experimental framework, we mapped the behavior of these portable models across both rigid mathematical landscapes (Experiment 1) and highly subjective linguistic evaluation tasks (Experiment 2).

Experiment 1 revealed that while pLLMs can act as effective optimization substrates in continuous, unimodal environments (yielding a 96.4% success rate on the sphere function), they suffer from severe systemic collapse when tasked with arithmetic operations of complex functions, especially when paired with the optimization constraints in this function's highly multi-modal landscape. This failure highlights the core limitation of language-oriented tokenized models when forced to execute complex arithmetic, which introduces numerical noise that disorients the local search method.

However, when transitioning this framework to the linguistic domain in Experiment 2, optimizing a persona profile directly in the prompt space yielded significant alignment breakthroughs on a hedonic valence classification task. The optimization idea proved to be heavily dependent on the model: larger architectures like Mistral NeMo achieved the best alignment with human evaluations, and it was improved by the optimized persona modulation. Conversely, the more prompt-sensitive Ministral 3, with the smallest parametric capacity, was the model that most benefited from the method, displaying a 34.9% reduction in SSE, despite its wider output scatter. Psychometrically, the algorithm's tendency in some models to converge on states of maximum Neuroticism showed a compatibility with psychological theory, hinting that inducing higher affective vigilance in some pLLM models could be advantageous for replicating human emotional scoring. Finally, the initialization coordinates of the initial simplex emerged as a crucial factor, where compact, centralized coordinates provided the optimal structural step size to navigate the bounding constraints of the tested parameter space.

Despite these outcomes, a primary limitation of this research is the restricted number of initial simplices evaluated and the containment of persona traits to only two dimensions of the

Big Five. Future work could explore hybrid global-local search methods to avoid premature convergence, as well as increase the prompt persona dimensions to encompass more complex psychometric profiles.

VI. Acknowledgment

The following tools were used to make this paper:

- Gemini AI for phrasing suggestions and code debugging.
- SciPy Python library for the Nelder-Mead method.
- Python code to run experiments and data analysis.
- Ollama framework to access the used models.

References

[1] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22. DOI: 10.1145/3586183.3606763.

[2] M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, and M. Strohmaier, "AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories," *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 808–826, 2024. DOI: 10.1177/17456916231214460.

[3] D. Peng, Y. Wang, C. Preiksaitis, and C. Rose, "Sycoeval-em: Sycophancy evaluation of large language models in simulated clinical encounters for emergency care," *ArXiv*, vol. abs/2601.16529, 2026. DOI: 10.48550/arxiv.2601.16529.

[4] W. Gurnee, N. Sofroniew, J. Lindsey, et al., "Verbalizable representations form a global workspace in language models," *Anthropic Transformer Circuits Thread*, 2026, Anthropic Research. [Online]. Available: https://transformer-circuits.pub/2026/workspace/index.html.

[5] Anthropic Interpretability Team, "Emotion concepts and their function in a large language model," *Anthropic Transformer Circuits Thread*, 2026, Anthropic Research. [Online]. Available: https://transformer-circuits.pub/2026/emotions/index.html.

[6] D. Watson, L. A. Clark, C. W. McIntyre, and S. Hamaker, "Affect, personality, and social activity," *Journal of Personality and Social Psychology*, vol. 63, no. 6, pp. 1011–1025, 1992. DOI: 10.1037//0022-3514.63.6.1011.

[7] J. Wei, S. Kim, H. Jung, and Y. H. Kim, "Leveraging large language models to power chatbots for collecting user self-reported data," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–35, 2024. DOI: 10.1145/3637364.

[8] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965. DOI: 10.1093/comjnl/7.4.308.

[9] R. Qin, Q. Wang, T. Wang, Z. Wei, and W. Shen, "Evaluating and explaining prompt sensitivity of LLMs using interactions," in *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. [Online]. Available: https://openreview.net/forum?id=GsdFM8qnav.

[10] F. C. Jandre, G. C. Motta-Ribeiro, and J. V. A. da Silva, "Could large language models estimate valence of words? A small ablation study," in *Proceedings of the XVI Congresso Brasileiro de Inteligência Computacional (CBIC)*, 2023, pp. 1–6. DOI: 10.21528/CBIC2023-148.

[11] P. C. V. de Barros, G. C. M. Ribeiro, and F. C. Jandre, "Can small large language models estimate the hedonic valence of words?" In *Proceedings of the XVII Congresso Brasileiro de Inteligência Computacional (CBIC)*, 2025.

[12] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013. DOI: 10.3758/s13428-012-0314-x.

[13] R. Chelouah and P. Siarry, "Genetic and Nelder–Mead algorithms hybridized for a more accurate global optimization of continuous multiminima functions," *European Journal of Operational Research*, vol. 148, no. 2, pp. 335–348, 2003. DOI: 10.1016/S0377-2217(02)00412-9.

[14] S. Kim, S. Gupta, C. Zheng, M. Jang, J. Zhou, and J. D. Choi, "Persona is a double-edged sword: Rethinking the impact of role-play prompts in zero-shot reasoning tasks," in *Findings of the Association for Computational Linguistics (ACL/IJCNLP)*, 2025, pp. 51–67. [Online]. Available: https://aclanthology.org/2025.findings-ijcnlp.51.

[15] M. D. Robinson, S. K. Moeller, and S. Ode, "Neuroticism and affective processing: Vigilance, reactivity, and their neurocognitive bases," *Journal of Personality*, vol. 78, no. 6, pp. 1897–1930, 2010. DOI: 10.1111/j.1467-6494.2010.00674.x.

Physical application of the Hodgkin-Huxley model in the description of the membrane action potential

Sales, I. Q.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
isasales40@gmail.com

Figueiredo, L. E. T. C.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
cassianolynnda@gmail.com

Sousa, A. O.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
arthursouzaoliveira250@gmail.com

Carneiro, H. A. B.
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
heloisacarneiro70@gmail.com

Fonseca, M.C.P
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
conceicao@ufpa.br

Miziara I. M
Institute of Technology
Federal University of Pará (UFPA)
Belém, Brazil
immiziara@ufpa.br

***Abstract*—Engineering and physiology students and researchers frequently require approaches that integrate concepts from electronics and neurophysiology to better understand cell membrane dynamics. In this context, mathematical models enable both computational simulations and physical implementations using equivalent electronic circuits, supporting research and educational applications. This study investigates the physical implementation of the Hodgkin–Huxley model to reproduce the membrane action potential using an equivalent electrical circuit. Initially, the membrane potential (V_m) and the ion channel activation and inactivation variables (m, n, and h) were obtained through computational simulation, allowing the calculation of equivalent resistances from the inverse relationship between conductance and resistance. The equivalent circuit was then simulated using electronic design software for validation and subsequently implemented with passive components, including an electrolytic capacitor representing the phospholipid bilayer, manually adjusted potentiometers representing the equivalent ion channel resistances, a fixed $3.3\,\mathrm{k}\Omega$ resistor representing leakage channels, and voltage sources corresponding to the ionic equilibrium potentials. Membrane potential measurements were acquired using a digital multimeter and analyzed graphically. Comparison between the experimental measurements and the mathematical model predictions revealed discrepancies attributed to the limitations of the passive circuit in reproducing the nonlinear dynamics of ion channel gating. Future work will investigate the use of voltage-controlled transistors to emulate channel gating, enabling a physical implementation that more closely reproduces the behavior described by the original Hodgkin–Huxley model.**

***Index terms*—Hodgkin-Huxley, action potential, electrical circuit, ion channels.**

I. Introduction

Neurons are excitable cells that transmit information through action potentials generated by the coordinated activation and inactivation of voltage-gated Na^+ and K^+ channels. Sodium influx promotes membrane depolarization, whereas potassium efflux restores the resting membrane potential, enabling the propagation of nerve impulses [1].

To quantitatively describe these electrophysiological mechanisms, Hodgkin and Huxley proposed the Hodgkin–Huxley (HH) model in 1952, representing the cell membrane as an equivalent electrical circuit in which the phospholipid bilayer is modeled as a capacitor and the ion channels as variable conductances associated with their respective equilibrium potentials [2]. Since its introduction, the HH model has become a cornerstone of computational neuroscience and remains widely employed in research, education, and computational studies of neuronal excitability [3]–[6].

More recently, analog implementations of the HH model using discrete electronic components have been proposed for educational and experimental applications [7]–[9]. In this context, this study presents the computational implementation of the HH model together with its equivalent physical circuit, enabling a comparison between experimentally measured membrane potentials and theoretical predictions. By integrating neurophysiology, mathematical modeling, and analog electronics, the proposed approach provides a practical platform for investigating the electrophysiological mechanisms described by the Hodgkin–Huxley model.

II. Methodology

This study is a quantitative, applied, theoretical-computational investigation based on mathematical modeling, computational simulation, and statistical analysis. The research methodology comprised sequential stages, including the implementation of the Hodgkin–Huxley mathematical model, computational simulation of the membrane potential and gating variables, determination of the equivalent electrical parameters, simulation of the equivalent circuit, physical implementation of the circuit, experimental data acquisition, and statistical analysis of the measured results. This workflow enabled a systematic comparison between the theoretical predictions of the mathematical model and the experimental behavior of the physical circuit, allowing the evaluation of

the proposed implementation. The complete methodological workflow adopted throughout the study is summarized in Fig. 8.

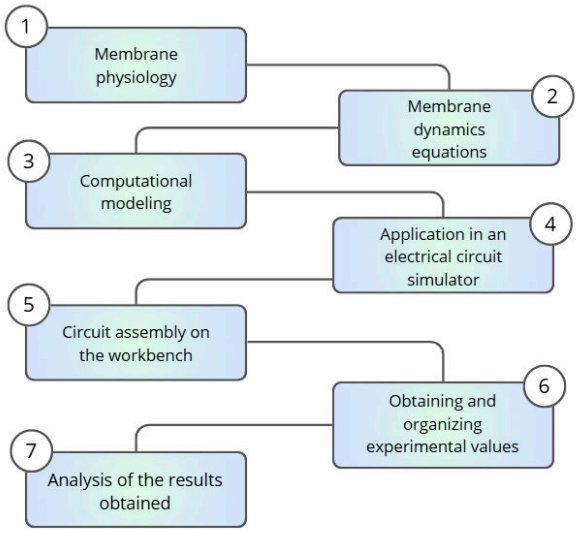


Figure 1: Methodological workflow including membrane physiology, mathematical modeling, computational simulation, physical circuit implementation, and analysis of the obtained results.

A. Physiology of the membrane action potential

Neurons are excitable cells that receive, process, and transmit information through electrical and chemical signals. This process depends on the electrical properties of the plasma membrane, including the resting membrane potential, membrane capacitance, and ionic resistance. The plasma membrane separates the intracellular and extracellular compartments while regulating the selective transport of ions through specialized ion channels [1], [3].

Under resting conditions, the neuronal membrane exhibits a potential difference of approximately -70 mV, known as the resting membrane potential. This potential results from the unequal distribution of ions between the intracellular and extracellular compartments, the high membrane permeability to potassium ions (K^+), and the continuous activity of the sodium–potassium pump (Na^+/K^+-ATPase), which transports three Na^+ ions out of the cell and two K^+ ions into the cell, thereby maintaining the electrochemical gradients required for neuronal excitability [1], [3], [5]. Small variations in membrane permeability continuously influence the membrane potential, allowing neurons to respond to external stimuli. When the membrane is sufficiently depolarized to reach the threshold potential, voltage-gated sodium channels open rapidly, initiating the action potential. This transient electrical event propagates along the axon without significant attenuation, enabling efficient communication between neurons and target tissues. The coordinated activation and inactivation of sodium and potassium channels are therefore essential for the generation, propagation, and termination of the action potential. The mechanism is illustrated in Fig. 8.

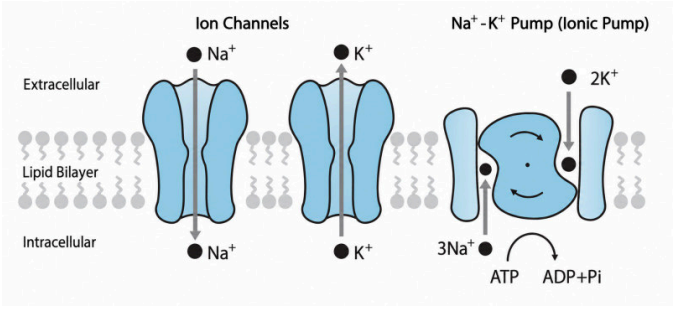


Figure 2: Illustration of the sodium–potassium pump mechanism. Adapted from Sterratt *et al.* (2011) [6].

When the membrane potential reaches the activation threshold, voltage-gated sodium channels open rapidly, allowing Na^+ influx and producing membrane depolarization [1], [2].

Within a few milliseconds, sodium channels become inactivated while voltage-gated potassium channels open. The resulting increase in K^+ permeability promotes potassium efflux, leading to membrane repolarization. In many cases, the elevated potassium conductance drives the membrane potential below its resting value, characterizing the hyperpolarization phase [1], [3], [5].

Subsequently, potassium channels return to their resting state, and the sodium–potassium pump gradually restores the original ionic gradients, allowing the membrane potential to return to its resting value. This sequence constitutes the action potential, which underlies the propagation of nerve impulses along the axon and neuronal communication [1], [2], [4].

A quantitative description of these mechanisms was proposed by Hodgkin and Huxley through a system of nonlinear differential equations describing the dynamics of sodium and potassium channel conductances, thereby enabling mathematical modeling of the different phases of the action potential [2].

B. Membrane dynamics equations

The Hodgkin–Huxley model was computationally implemented in *MATLAB* R2026a using the original differential equations proposed by Hodgkin and Huxley to describe the dynamics of the membrane potential and ion channels. These equations were derived from experiments based on the *voltage-clamp* technique, which enables independent measurement of the ionic currents responsible for generating the action potential [2].

As illustrated in Fig. 3, the *voltage-clamp* method employs a negative-feedback system that maintains the membrane potential at a prescribed value by injecting or withdrawing current between the intracellular and extracellular compartments. This approach allows the ionic currents to be measured while controlling the membrane potential.

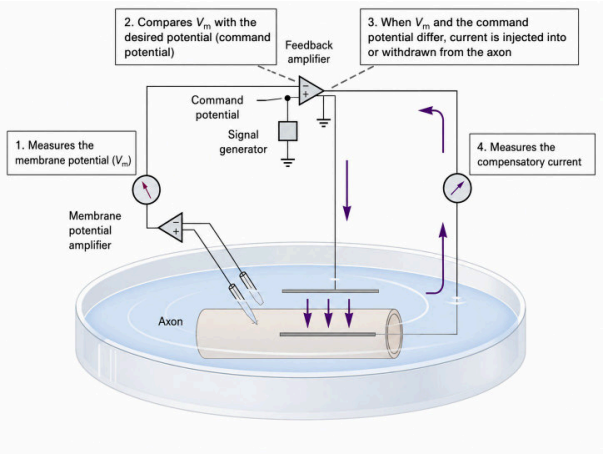


Figure 3: Schematic representation of the *voltage-clamp* method. (1) An intracellular electrode connected to an amplifier measures the membrane potential. (2) The measured potential is compared with a predefined command potential. (3) If a difference exists, a feedback amplifier injects or withdraws current until the membrane potential reaches the command value. (4) The injected current required to maintain this condition is continuously measured. Adapted from Kandel *et al.* [1].

These experiments enabled Hodgkin and Huxley to isolate the intrinsic properties of ion channels and quantitatively characterize the Na^+, K^+, and leak currents. They also demonstrated that channel conductances depend on both membrane potential and time and can be described by the state variables m, n, and h, representing the activation and inactivation dynamics of voltage-gated ion channels [2], [5].

The state variables are governed by the functions $\alpha(V)$ and $\beta(V)$, whose numerical solution provides the membrane potential, channel state variables, conductances, and equivalent resistances required for the implementation of the equivalent electrical circuit. Together, these quantities define the Hodgkin–Huxley system of ordinary differential equations (ODEs) describing the action potential dynamics.

The total membrane current (I_m) is the sum of the ionic current (I_{ionic}) and the capacitive current $C_m \dfrac{dV_m}{dt}$ associated with the membrane capacitance, as expressed in (1).

$$I_m(t) = I_{\text{ionic}}(t) + C_m \frac{dV_m(t)}{dt} \tag{1}$$

The ionic current is given by the sum of the sodium (I_{Na}), potassium (I_{K}), and leak (I_{L}) channel currents, as expressed in (2).

$$I_{\text{ionic}}(t) = I_{\text{Na}} + I_{\text{K}} + I_{\text{L}}. \tag{2}$$

Each ionic current depends on the conductance of its respective channel and the difference between the membrane potential (V_m) and the corresponding equilibrium potential (E_{Na}, E_{K}, and E_{L}). The membrane potential is the electrical potential difference between the intracellular and extracellular compartments, resulting from the unequal distribution of ions across the plasma membrane. It arises from the combined effects of the ionic concentration gradient, which generates the chemical driving force, and the electrical gradient, which generates the electromotive force. Together, these forces determine the direction and magnitude of ionic flow through membrane channels [1], [5].

The equilibrium potential is the membrane potential at which the chemical and electrical driving forces acting on a given ion are balanced, resulting in no net ionic flux. Accordingly, E_{Na}, E_{K}, and E_{L} denote the equilibrium potentials of sodium, potassium, and leak channels, respectively, with the latter representing the contribution of residual membrane currents in the Hodgkin–Huxley model [1], [5].

Accordingly, the sodium, potassium, and leak currents are expressed as the product of the corresponding channel conductance and the electrochemical driving force, given by $(V_m - E_x)$, as presented in (3), (4), and (5).

$$I_{Na} = g_{Na}(V,t)\,[V_m(t) - E_{Na}]\,, \tag{3}$$

$$I_K = g_K(V,t)\,[V_m(t) - E_K]\,, \tag{4}$$

$$I_L = g_L\,[V_m(t) - E_L]\,. \tag{5}$$

The Na^+ and K^+ channel conductances are voltage dependent and are modulated by gating variables. In the Hodgkin–Huxley model, these conductances are expressed as functions of the gating variables (m, n, and h), as given by Eqs. (6) and (7).

$$g_{Na}(V,t) = \bar{g}_{Na}\, m^3(V,t)\, h(V,t) \tag{6}$$

$$g_K(V,t) = \bar{g}_K\, n^4(V,t). \tag{7}$$

The Na^+ channel contains two independent gates: an activation gate (m), which opens rapidly during depolarization, and an inactivation gate (h), which closes more slowly and interrupts the sodium current. In contrast, the K^+ channel contains only an activation gate (n) and remains open while the membrane is depolarized.

The gating variables m, n, and h are obtained from differential equations governed by the functions $\alpha(V)$ and $\beta(V)$, which describe the voltage-dependent transition rates of channel activation and inactivation, as shown in Eqs. (8)–(9). Variations in these functions modify the probability of channel opening, thereby regulating the Na^+ and K^+ currents responsible for membrane depolarization and repolarization [1], [2], [5].

$$\frac{dn}{dt} = \alpha_n(V)\,(1-n) - \beta_n(V)n, \tag{8}$$

$$\frac{dm}{dt} = \alpha_m(V)\,(1-m) - \beta_m(V)m, \tag{9}$$

$$\frac{dh}{dt} = \alpha_h(V)\,(1-h) - \beta_h(V)h. \tag{10}$$

In Eqs. (8)–(10), the variables m, n, and h assume values between 0 and 1, representing the probability that each gate is open. The functions $\alpha(V)$ and $\beta(V)$ determine the transition rates of these variables as a function of the membrane potential.

C. Computational Implementation in a Simulation Environment

Figure 4 illustrates the computational implementation of Eqs. (1) and (2) to generate the membrane action potential. Subsequently, Eqs. (8)–(10) were solved to obtain the activation and inactivation variables of the ion channels.

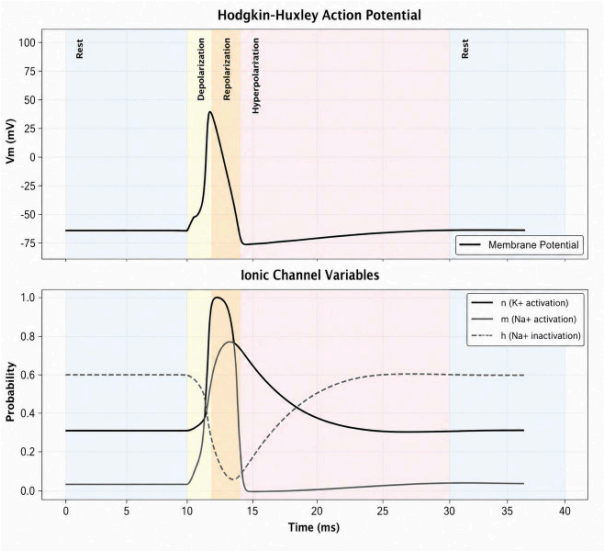


Figure 4: Action potential and ion channel variable waveforms. The light blue shaded region represents the resting phase, the yellow region represents depolarization, the orange region highlights repolarization, and the light pink region indicates hyperpolarization.

The temporal evolution of the membrane potential and gating variables in Fig. 4 occurs over 0–40 ms. During rest, V_m remains near -65 mV, with sodium channels mostly closed and potassium channels maintaining their resting conductance. After stimulation, the increase in m activates Na^+ channels, causing rapid depolarization, while h decreases and n increases gradually. At the action potential peak, sodium channel inactivation and potassium channel activation initiate the transition toward repolarization. Subsequently, the predominance of K^+ efflux restores the membrane potential, with a brief hyperpolarization caused by the delayed closure of potassium channels.

The coordinated variation of the gating variables demonstrates the dynamic behavior of voltage-dependent ion channels during neuronal excitation. The activation variable m responds rapidly to changes in membrane voltage, controlling the initial increase in sodium conductance, whereas the slower dynamics of h and n determine the duration of the action potential and the recovery phase. This difference in activation kinetics is essential for the temporal organization of depolarization and repolarization processes. Therefore, the interaction between these variables allows the Hodgkin–Huxley model to reproduce the main electrical characteristics of neuronal activity. Figure 5 presents the corresponding Na^+ and K^+ conductances and their equivalent resistances, demonstrating the inverse relationship between these parameters.

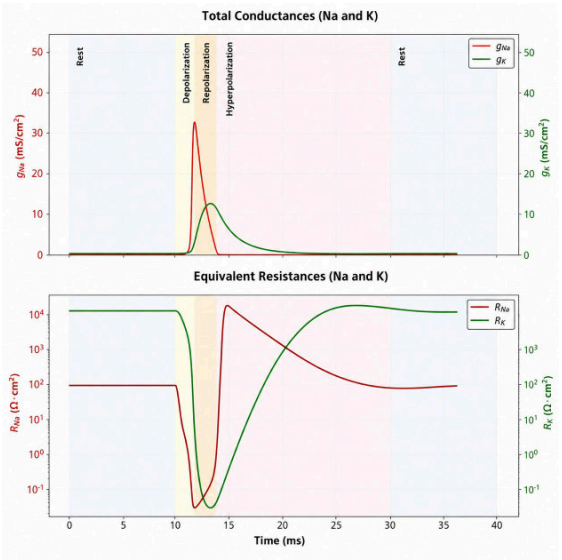


Figure 5: Total sodium (g_{Na}) and potassium (g_{K}) channel conductances and their corresponding equivalent resistances (R_{Na} and R_{K}) throughout the action potential. A transient increase in sodium conductance is observed during the depolarization phase, followed by an increase in potassium conductance during repolarization. Consequently, the equivalent resistances exhibit an inverse relationship with the conductances, reaching their minimum values when ionic permeability is at its maximum.

As expected, the increase in Na^+ conductance during depolarization is accompanied by a decrease in the corresponding equivalent resistance, facilitating Na^+ influx. Conversely, during repolarization, the increase in K^+ conductance is associated with a reduction in its equivalent resistance, promoting K^+ efflux.

D. Application in an electrical circuit simulator

Before the physical implementation, the proposed equivalent circuit was simulated using the *Proteus* software, a widely adopted platform for the simulation of electronic and electrical systems. This step was performed to verify the circuit operation, identify potential design inconsistencies, and validate its functionality prior to hardware implementation.

The equivalent circuit consisted of active components (voltage sources) and passive components (variable resistors, a fixed resistor, and a capacitor), as illustrated in Fig. 6.

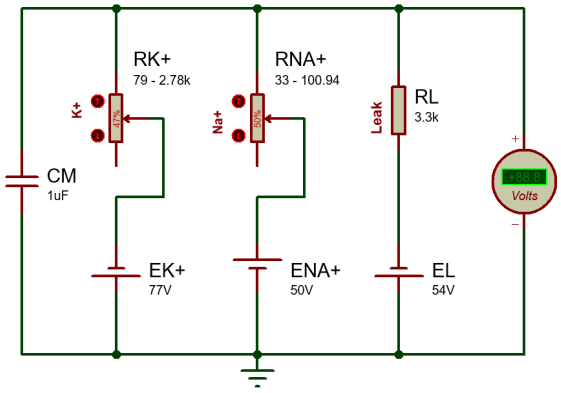


Figure 6: Equivalent electrical circuit implemented in *Proteus*. The circuit comprises the membrane capacitance (C_M), potassium resistance (R_{K^+}), sodium resistance (R_{Na^+}), leak resistance (R_L), potassium equilibrium potential (E_{K^+}), sodium equilibrium potential (E_{Na^+}), leak equilibrium potential (E_L), and a voltmeter for membrane potential measurement.

As shown in Fig. 6, the polarity of the E_{K^+} and E_L voltage sources is reversed with respect to E_{Na^+}. This configuration reflects the physiological direction of the ionic currents across the membrane. During depolarization, Na^+ ions enter the cell, generating an inward current, whereas during repolarization K^+ ions leave the cell, producing an outward current. Under the simulated conditions, the leak current also flows predominantly outward. Consequently, the adopted source polarities ensure that the current directions in the equivalent circuit are consistent with the physiological behavior of ionic transport across the membrane.

E. Circuit assembly on the workbench

The components used for the physical implementation of the equivalent circuit are presented in Fig. 7.

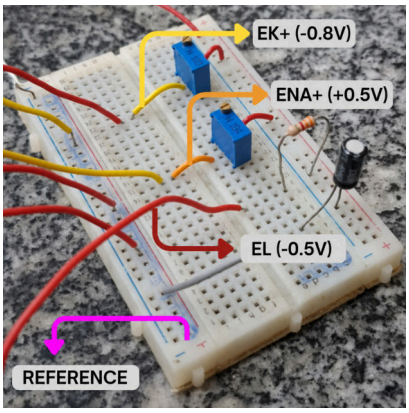


Figure 7: Physical implementation of the equivalent circuit shown in Fig. 6.

Trimmer potentiometers (Trimpots) were employed to implement the variable resistances corresponding to the membrane ion channel conductances. The following models were used:

- **w103**: Rated at 10 kΩ, this potentiometer was used to represent the potassium (K^+) channel.
- **w104**: Rated at 100 kΩ, this potentiometer was used to represent the sodium (Na^+) channel, providing a wider resistance adjustment range than that required for the K^+ channel.

Three DC voltage sources were employed to represent the equilibrium potentials, with $E_{\mathrm{K}^+} = 0.8$ V, $E_{\mathrm{Na}^+} = 0.5$ V, and $E_L = 0.5$ V. A 1 μF electrolytic capacitor was used to represent the membrane capacitance.

F. Collection and organization of experimental data

After assembling the equivalent circuit, experimental data were collected by discretizing the action potential waveform. Twenty-three sampling points were selected along the temporal profile of the action potential, covering the resting, depolarization, repolarization, and hyperpolarization phases.

For each sampling instant, the theoretical resistance values corresponding to the ion channels were obtained from the computational simulation. The trimmer potentiometers were then manually adjusted to reproduce these resistance values, which were verified using a digital multimeter. This procedure ensured that the physical resistance implemented in the circuit matched the corresponding theoretical value at each point of the action potential.

After measuring the membrane potential (V_m), potassium resistance (R_{K}), and sodium resistance (R_{Na}), the experimental data were organized in a spreadsheet for subsequent processing and analysis. Each sampling instant was associated with the corresponding values of R_{K} and R_{Na} implemented in the circuit.

This procedure enabled the standardization and verification of the experimental data, facilitating graphical analysis and direct comparison with the theoretical predictions of the Hodgkin–Huxley model.

G. Analysis of the Obtained Results

The values of V_m, R_{K^+}, R_{Na^+}, and their corresponding sampling points (x) were organized into three tables to facilitate the analysis and comparison of the computational and experimental results. Table I presents the values obtained from the computational implementation, which served as the theoretical reference for the action potential predicted by the Hodgkin–Huxley model. These data were used to determine the expected variations in ion channel resistances throughout the action potential and to define the parameters required for the physical circuit implementation. The sampling points were selected according to relevant phases of the action potential, including depolarization, repolarization, and hyperpolarization, allowing a more detailed evaluation of the dynamic behavior of the model. The computational results provided the basis for adjusting the equivalent circuit parameters and comparing the theoretical predictions with the experimental measurements obtained from the implemented circuit.

TABLE I: Voltage and resistance values obtained computationally.

Point X	V_m **(V)**	R_K **(kΩ)**	R_{Na} **(kΩ)**
1	-65	2,72	94,25
1130	-32,39	1,47	0,56
1147	-0,6544	1,0445	0,1341
1154	20,255	0,825	0,0709
1170	40,2457	0,450	0,033
1227	18,4559	0,1255	0,043
1263	-0,1479	0,0929	0,0632
1333	-32,6064	0,0791	0,1582
1390	-65,2879	0,0896	1,7771
1399	-70,2739	0,0933	5,3573
1454	-76,192	0,124	10094
2128	-70,2908	1,7677	759,4332
3111	-64,5146	2,7823	79,0164
3619	-65,0004	2,6768	93,8534

The experimental measurements are summarized in Tables II and III. Table II presents the results obtained from the physical implementation of the equivalent circuit. At sampling points 1390, 1399, 1454, 1847, 2128, and 2376, the equivalent sodium-channel resistance (R_{Na^+}) reached very high values. This behavior reflects the low sodium-channel conductance at these instants, causing the equivalent resistance to approach an open-circuit condition.

TABLE II: Voltage and resistance values of the physical experimental circuit.

Point X	V_m **(mV)**	R_K **(kΩ)**	R_{Na} **(kΩ)**
1	0,0066	2,7	94
1001	0,0150	2,6	95
1055	0,0180	2,4	0,0099
1130	0,0150	2,6	0,095
1147	0,0180	2,4	0,0099
1154	0,0014	1,46	0,55
1170	-0,0020	1	0,135
1227	-0,0055	0,826	0,067
1263	-0,0182	0,045	0,033
1333	0,0550	0,127	0,043
1364	0,0160	0,0806	0,0327
1390	-0,0010	0,092	0,0628
1399	0,0090	0,0808	0,158
1454	0,0160	0,0806	0,327
1847	0,0032	0,752	10000
2128	0,0050	1,77	760
2376	0,0022	2,62	271
2624	0,0130	2,9	0,132
3111	0,0300	2,6	0,0776
3619	0,0140	2,8	0,0934

To investigate the influence of this behavior on the membrane potential, an additional experiment was performed in which the equivalent sodium-channel resistance was removed from the circuit at sampling points 1390, 1399, 1454, 1847, 2128, and 2376. The corresponding experimental results are presented in Table III.

TABLE III: Experimental results with the sodium-channel branch disconnected.

Point X	V_m **(V)**	R_K **(kΩ)**	R_{Na} **(kΩ)**
1	0,0066	2,70	94,00
1001	0,0150	2,60	0,095
1055	0,0180	2,40	0,0099
1130	0,0150	2,60	0,095
1147	0,0180	2,40	0,0099
1154	0,0014	1,46	0,55
1170	-0,0020	1,00	0,135
1227	-0,0055	0,826	0,067
1263	-0,0182	0,450	0,033
1333	0,0550	0,127	0,043
1364	0,0160	0,080	0,327
1390	0,0050	0,092	-
1399	0,0090	0,080	-
1454	0,0160	0,080	-
1847	0,0032	0,752	-
2128	0,0050	1,77	-
2376	0,0022	2,62	-
2624	0,0130	2,90	-
3111	0,0300	2,60	0,077
3619	0,0140	2,80	0,093

To quantitatively compare the computational and experimental results, the mean and standard deviation of the measurements reported in Tables II and III were calculated. These statistical measures were then compared with the theoretical values presented in Table I to evaluate the agreement between the computational Hodgkin–Huxley model and its equivalent physical circuit implementation.

III. Results and Discussion

Based on the data presented in Tables II and III, the mean membrane potential (V_m) and the corresponding standard deviation were calculated for each sampling point. The experimental mean exhibited an atypical waveform compared with that predicted by the Hodgkin–Huxley model, characterized by a single positive deflection at the action potential peak followed by a hyperpolarization phase. The standard deviation remained relatively low (within approximately $\pm 1\,\mathrm{mV}$ for most sampling points), indicating good repeatability and low variability between measurements performed under similar experimental conditions.

The experimental measurements were compared with the computational results obtained for the standard circuit, the open-circuit configuration (with the Na^+ channel branch disconnected), and the numerical Hodgkin–Huxley model. As shown in Fig. 8, the experimental mean did not reproduce the response of the standard circuit, particularly during the depolarization phase, where the simulated waveform exhibited a more pronounced positive deflection. Additional discrepancies were observed during the repolarization and after-hyperpolarization phases, where the measured membrane potentials tended to be more negative than those predicted by the computational model.

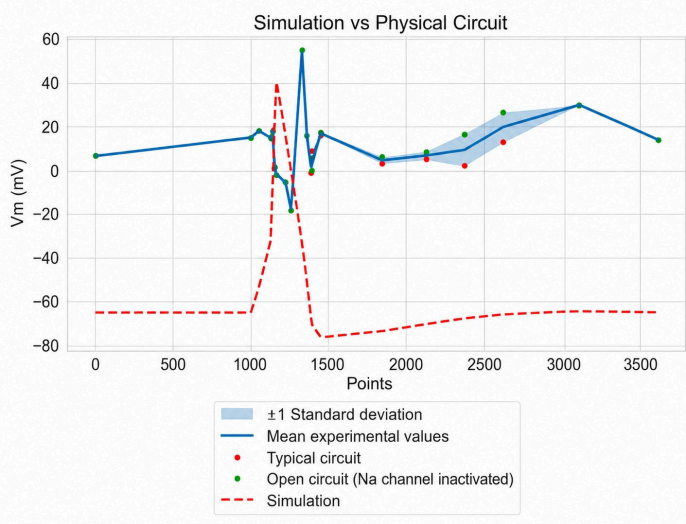


Figure 8: Comparison between the computational simulation and the experimental results, showing the experimental mean, standard deviation, the standard circuit response, and the response obtained with the Na^+ channel branch disconnected.

The largest discrepancies between the experimental and computational results occurred during the depolarization phase and at the action potential peak, where the experimental waveform exhibited a significantly lower amplitude than that predicted by the standard circuit. During the initial repolarization phase, the experimental response also showed a slightly slower decay than the simulated waveform, resulting in a peak delay of approximately $0.2\,\mathrm{ms}$. More pronounced differences were observed during the late repolarization and after-hyperpolarization (*undershoot*) phases, where the experimental membrane potentials were up to $2.6\,\mathrm{mV}$ more negative than the simulated values.

The open-circuit configuration exhibited a markedly different behavior, characterized by a substantially reduced amplitude (maximum peak of only $5.2\,\mathrm{mV}$) and the absence of the characteristic action potential peak. This result confirms the essential role of Na^+ channels in action potential generation, consistent with the predictions of the Hodgkin–Huxley model. Although the standard circuit did not perfectly reproduce the experimental waveform, it provided a considerably closer approximation than the open-circuit configuration, reinforcing the importance of sodium-channel conductance in the equivalent electrical model.

The observed discrepancies can be attributed to the simplified representation of membrane dynamics adopted in the proposed circuit. In biological membranes, the Na^+ and K^+ channels exhibit continuous voltage-dependent activation and inactivation, whereas the equivalent circuit represents these processes through discrete resistance adjustments. Consequently, although the proposed implementation reproduces the principal phases of the action potential, it cannot fully emulate the nonlinear dynamics described by the Hodgkin–Huxley equations. Additional sources of error, including component tolerances, equivalent membrane resistance variations, and temperature effects, may also have contributed to the differences between the computational and experimental responses.

From an electronic perspective, the cell membrane behaves analogously to a transistor, selectively regulating the flow of ionic current through voltage-dependent Na^+ and K^+ channels. This mechanism resembles the operation of a field-effect transistor (FET), in which the gate voltage controls the current flowing between the source and drain. Accordingly, transistor-based circuits have been proposed to reproduce the temporal evolution of the Hodgkin–Huxley gating variables (m, n, and h), providing a closer approximation of the model's nonlinear dynamics than implementations based exclusively on passive components [8].

Few studies have reported the physical implementation of the Hodgkin–Huxley equivalent circuit together with systematic experimental validation. Nizar *et al.* [7] developed analog neurons for virtual laboratory applications but did not report detailed membrane-potential measurements. Rutherford *et al.* [9] proposed a transistor-based implementation capable of qualitatively reproducing the Hodgkin–Huxley dynamics; however, discretized experimental measurements and quantitative comparisons with the theoretical model were not presented. Likewise, Kutzner and Bryson [8] described an educational implementation based on potentiometers without continuous experimental data acquisition.

The present work extends these previous studies by experimentally evaluating the equivalent circuit through membrane-potential measurements at 23 discretized sampling points covering all phases of the action potential. The potentiometers were manually adjusted to reproduce the theoretical resistance values predicted by the computational model, enabling a direct comparison between the numerical and experimental responses. Although deviations were observed, particularly during depolarization and after-hyperpolarization, the proposed circuit successfully reproduced the principal characteristics of the Hodgkin–Huxley model, demonstrating the feasibility of its physical implementation as an educational and experimental platform.

Future implementations may further improve the agreement between the computational and experimental responses by incorporating active electronic devices capable of reproducing the voltage-dependent behavior of ionic conductances. Electrochemical transistors (ECTs), for example, can directly couple ionic and electronic conduction, enabling applications in neuromorphic computing and neuroprosthetic systems [7], [9]. Such technologies may bridge the gap between simplified educational circuits and biologically realistic electronic implementations while preserving the conceptual foundation provided by equivalent electrical models of membrane electrophysiology.

IV. Conclusion

This work presented the computational modeling, simulation, and physical implementation of an equivalent electrical circuit based on the Hodgkin–Huxley model for representing the electrical behavior of excitable cell membranes.

The proposed methodology enabled the determination of the membrane potential and the equivalent sodium and potassium channel resistances from the numerical model, followed by the implementation and experimental evaluation of the corresponding electrical circuit.

The comparison between the computational predictions and the experimental measurements demonstrated that the proposed circuit successfully reproduced the principal characteristics of the Hodgkin–Huxley action potential. Although discrepancies were observed, particularly during the depolarization and after-hyperpolarization phases, the experimental results remained consistent with the expected qualitative behavior of the model. These differences are primarily attributed to the simplified representation of the voltage-dependent ionic conductances and to the practical limitations associated with the use of discrete electronic components.

The experimental measurements also showed that, at specific sampling points, the equivalent sodium-channel resistance reached very high values, approaching an open-circuit condition. This behavior is consistent with the sodium-channel inactivation phase described by the Hodgkin–Huxley model, during which sodium conductance decreases despite the membrane remaining depolarized. Consequently, the proposed equivalent circuit was able to reproduce not only the overall action potential waveform but also the qualitative changes in the electrical behavior associated with the activation and inactivation of the ionic channels.

Overall, the results demonstrate that the proposed equivalent electrical circuit constitutes a practical platform for studying membrane electrophysiology and for illustrating the electrical mechanisms underlying the Hodgkin–Huxley model. Future work will focus on incorporating active electronic devices, such as transistors, to emulate the voltage-dependent nonlinear behavior of ionic conductances more accurately, thereby improving the agreement between the computational model and its physical implementation.

REFERENCES

[1] E. R. Kandel, J. D. Koester, S. H. Mack, and S. A. Siegelbaum, *Principles of Neural Science*, 6th ed. New York: McGraw Hill, 2021.

[2] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, pp. 500–544, 1952.

[3] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, R. D. Mooney, M. L. Platt, and L. E. White, *Neuroscience*, 6th ed. New York: Oxford University Press, 2018.

[4] J. G. Nicholls, A. R. Martin, B. G. Wallace, and P. A. Fuchs, *From Neuron to Brain*, 5th ed. Sunderland, MA: Sinauer Associates, 2012.

[5] B. Hille, *Ion Channels of Excitable Membranes*, 3rd ed. Sunderland, MA: Sinauer Associates, 2001.

[6] D. Sterratt, B. Graham, A. Gillies, and D. Willshaw, "Principles of Computational Modelling in Neuroscience," Cambridge Univ. Press, Cambridge, U.K., 2011.

[7] N. Nizar, R. Radhamani, D. Kumar, B. Nair, K. Achuthan, and S. Diwakar, "Implementation of analog electrical neurons as virtual labs for neuroscience education," in *Proc. IEEE Conf.*, 2018.

[8] M. D. Kutzner and J. M. Bryson, "A progressive complex circuit lab for introductory physics and biophysics," *Phys. Teach.*, vol. 56, no. 8, pp. 540–544, Nov. 2018.

[9] G. H. Rutherford, Z. D. Mobile, J. Brandt-Trainer, R. Follmann, and E. Rosa, Jr., "An electronic neuron and its Hodgkin–Huxley model counterpart for education," *Am. J. Phys.*, vol. 88, no. 7, pp. 1–7, 2020.

Physiological Validation of Synthetic Ventilatory Datasets Generated by the ASL5000 for Asthma Exacerbation Risk Stratification

1st Maria Luísa B. de Luna Barros
dept. of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0001-9496-4892

2nd Camilla Beatriz Coutinho da Fonseca
Department of Physiotherapy
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0000-7519-5269

3rd Guilherme Ferreira Máximo
Center for Informatics
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0001-0158-9983

4th Emily Gomes Paiva
Department of Physiotherapy
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0003-8043-8853

5th Jairo Cândido Gonzaga Neto
Center for Informatics
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0001-3738-7048

6th Layane Santana Pereira Costa
Department of Physiotherapy
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-7566-4149

7th Maria Eduarda Moura Maciel Maria
Department of Physiotherapy
Federal University of Pernambuco
Recife, Brazil
ORCID: 0009-0004-5092-0179

8th Maira Araujo de Santana
Center for Informatics
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0002-1796-7862

9th Armele Dornelas de Andrade
Department of Physiotherapy
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0001-9430-4395

10th Shirley Lima Campos
Department of Physiotherapy
Federal University of Pernambuco
Recife, Brazil
ORCID: 0000-0003-3079-8300

***Abstract*—The development of machine learning models to predict acute asthma exacerbations is severely hindered by the scarcity of balanced, high-quality clinical datasets. High-fidelity simulators offer a potential solution by generating synthetic respiratory digital twins; however, their physiological validity and suitability for data augmentation must be rigorously assessed. This study presents a structured validation framework applied to a synthetic cohort of thirty virtual patients across three asthma severity levels, generated via the ASL5000 Breathing Simulator. We demonstrate that the simulator reproduces physiologically plausible patterns consistent with established respiratory mechanics, including compensatory tachypnea, reduced dynamic compliance, and intrinsic positive end-expiratory pressure. Spearman rank correlation analysis revealed strong monotonic trends for key variables ($\rho > 0.78$, $p < 0.001$). Furthermore, we report a novel observation: despite a systematic numerical bias in respiratory rate measurements, the dataset exhibits strong categorical separability for severity stratification (Cohen's Kappa = 0.90), suggesting its potential suitability for classification tasks. Finally, we evaluated a Synthetic Minority Oversampling Technique (SMOTE) pipeline, providing statistical evidence that synthetic expansion preserves both marginal distributions and multivariate dependence structures (mean absolute correlation difference $\Delta\rho = 0.020$). This framework offers a reproducible foundation for training edge artificial intelligence devices in respiratory monitoring.**

***Index Terms*—Asthma, Synthetic Data, Digital Twins, Machine Learning, Physiological Validation, Biomedical Engineering.**

I. Introduction

ASTHMA affects hundreds of millions of individuals worldwide and remains one of the leading causes of morbidity, emergency department visits, and healthcare expenditure [1]. Despite advances in pharmacological therapies, acute exacerbations remain difficult to predict, often leading to delayed interventions and preventable clinical deterioration. Consequently, asthma management is increasingly shifting toward predictive approaches supported by data-driven decision systems capable of identifying subtle physiological changes before severe manifestations occur [2], [3].

Artificial intelligence (AI) and machine learning (ML) have emerged as promising tools for exacerbation prediction. However, their performance depends on the availability of large, diverse, and physiologically representative datasets [4]. Acquiring respiratory data during acute exacerbations is ethically, logistically, and financially challenging, while the low incidence of severe events results in highly imbalanced datasets. Synthetic data generation has therefore become an

attractive strategy for dataset augmentation, class balancing, and improved model robustness [5].

Digital twin technology enables physiologically controlled synthetic data generation through high-fidelity respiratory simulators, which allows precise manipulation of airway resistance, lung compliance, and inspiratory effort to recreate clinically relevant scenarios [7], [8]. Nevertheless, synthetic datasets must be physiologically validated before being used for ML applications, since unrealistic simulations or inappropriate augmentation techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), may introduce implausible biomechanical relationships and compromise model reliability [6].

To address these challenges, this study proposes a structured physiological validation framework for synthetic ventilatory datasets generated using the ASL5000, supporting the development of edge-AI respiratory monitoring systems,including the Respiratory Diagnostic Assistant [9]. We hypothesize that synthetic respiratory profiles generated under controlled mechanical constraints reproduce clinically plausible respiratory behavior and remain suitable for data augmentation when physiological integrity is preserved.

The main contributions of this work are:

- **Physiological Plausibility Assessment:** Demonstration that the ASL5000 reproduces key physiological features of the asthma cascade (e.g., auto-PEEP, tachypnea, and expiratory flow limitation), supported by Spearman correlation analysis.
- **Categorical versus Numerical Validity Observation:** Evidence that, despite numerical bias in respiratory rate, the simulator preserves severity stratification with strong categorical agreement (Cohen's Kappa = 0.90), supporting its use in ML classification tasks.
- **Multivariate Integrity of Data Augmentation:** Statistical evidence that a SMOTE-based pipeline preserves both marginal distributions and multivariate dependence structures (mean absolute correlation difference $\Delta\rho = 0.020$).
- **Edge-AI Readiness:** A reproducible pipeline providing balanced, physiologically coherent datasets for training lightweight ML models on embedded devices.

II. Related Work

Machine learning (ML) has become increasingly important in asthma management, driven by advances in electronic health records and wearable monitoring. Recent systematic reviews show that supervised learning algorithms identify patterns associated with impending exacerbations, often outperforming conventional statistical approaches [2], [3]. However, the scarcity of high-quality datasets containing severe exacerbation events remains a major limitation, leading to class imbalance and reduced model generalization [4], [5].

To address these limitations, synthetic data generation and digital twin technologies have emerged as promising alternatives. A digital twin is a virtual representation of a biological system that reproduces physiological behavior through computational models [10]. In respiratory research, the ASL5000 enables precise control of airway resistance, lung compliance, inspiratory effort, and respiratory drive, generating physiologically coherent respiratory profiles [7], [8]. Previous studies have demonstrated strong agreement between ASL5000 outputs and theoretical respiratory mechanics [8].

Nevertheless, synthetic data alone do not guarantee physiological realism. Models trained on unrealistic data may perform well internally but fail to generalize clinically. Moreover, the limited size of simulator-generated cohorts often requires augmentation techniques such as SMOTE, whose inappropriate application may generate physiologically implausible samples and distort biomechanical relationships [6]. Consequently, validation frameworks simultaneously assessing the physiological plausibility of simulator-generated datasets and the integrity of subsequent augmentation remain scarce. The present study addresses this gap by proposing a structured validation methodology based on emergent physiological biomarkers.

III. Materials and Methods

A. Study Design and Experimental Setup

This study was designed as a physiological validation investigation conducted at the Multiuser Laboratory for Instrumental Innovation and Physical-Functional Performance (LINDEF). The experimental setup used the ASL5000 high-fidelity breathing simulator (IngMar Medical, Pittsburgh, PA, USA), controlled via ASL SW software (v3.6) on a Surface Pro 10 tablet connected through a USB interface.

To ensure physiological fidelity, simulations were performed in a controlled environment at 36 °C using a 500 mL silicone test lung (PROTEC, Brazil). The protocol consisted of continuous 1-minute runs per virtual patient, corresponding to approximately 30–40 spontaneous respiratory cycles. High-resolution waveforms were recorded and extracted using the ASL5000 Post-Run Analysis module.

B. Signal Quality Control

Prior to analysis, each recorded respiratory cycle was screened for physiological plausibility. Cycles exhibiting values outside established clinical ranges (e.g., respiratory rate > 80 BPM, negative I:E ratios, or plateau pressures > 60 cmH_2O) were flagged as artifacts likely corresponding to the simulator's initial stabilization phase or transient mechanical oscillations. These cycles were excluded prior to computing per-patient averages. This quality-control step ensured that the analyzed dataset reflected stable respiratory behavior rather than start-up transients.

C. Virtual Cohort Construction

A synthetic cohort comprising thirty virtual patients was generated and stratified into three clinically relevant severity categories: Group 1 (Mild Asthma, n=10), Group 2 (Moderate Asthma, n=10), and Group 3 (Severe Asthma, n=10).

The cohort size was not intended to estimate population prevalence, but rather to create a controlled proof-of-concept sample with sufficient intra-group variability for physiological

validation. Ten patients per group were selected to allow paired statistical comparisons while maintaining a manageable experimental scope, consistent with bench-study conventions in respiratory simulation research [8].

The severity hierarchy was designed to reproduce the progressive deterioration in respiratory mechanics typically observed during acute asthma exacerbations. Parameter ranges for each severity level were defined through a systematic review of clinical guidelines (GINA 2025) [1] and established respiratory physiology references [11], [12]. Within each group, individual patient profiles were constructed by manually sampling combinations of resistance, compliance, respiratory rate, and muscle pressure from predefined ranges, ensuring intra-group variability while preserving clinically plausible limits. This manual construction approach was adopted to maintain direct control over the physiological coherence of each virtual patient. Each patient's parameter combination is detailed in Table I.

D. Respiratory Simulation Environment

All simulations were conducted in spontaneous breathing mode. The ASL5000 reproduces respiratory behavior through a computer-controlled piston assembly governed by the mathematical model of lung mechanics, represented by the equation of motion:

$$P_{mus} = R\dot{V} + \frac{V}{C} + PEEP \tag{1}$$

where P_{mus} represents inspiratory muscle pressure, R denotes airway resistance, $\dot{V}$ represents airflow, V denotes lung volume, C corresponds to lung compliance, and $PEEP$ is positive end-expiratory pressure. This formulation allows independent control of airway obstruction and elastic load while preserving physiologically meaningful pressure-flow relationships.

E. Physiological Output Variables and Feature Extraction

Beyond the primary mechanical parameters (R and C), several downstream physiological variables were extracted for validation. These included Spontaneous Respiratory Rate (BPM), Mean Expiratory Pressure (cmH_2O), Peak Expiratory Flow Characteristics (%), Inspiratory-to-Expiratory Ratio (I:E), Maximum Trigger Pressure Drop (ΔP_{trig}, cmH_2O), and Total Inspiratory Elastic Work (mJ). Importantly, these variables were *not* directly programmed into the simulator, but instead emerged naturally from the interaction between respiratory mechanics and spontaneous breathing behavior, providing a robust test of physiological realism.

F. Statistical Validation Framework and Data Augmentation

To ensure dataset integrity, a multi-stage statistical pipeline was implemented. First, the separability of the three severity groups was assessed using the Kruskal-Wallis test with Dwass-Steel-Critchlow-Fligner post-hoc comparisons, given the non-normal distribution of several variables (verified via the Shapiro-Wilk test). Effect sizes were quantified using epsilon-squared (ϵ^2).

Second, to evaluate whether physiological parameters evolved consistently with increasing disease severity, Spearman's rank correlation coefficient (ρ) was computed between ordinal severity coding (Mild=1, Moderate=2, Severe=3) and each physiological variable. This analysis tested the central hypothesis that simulated parameters follow monotonic trends consistent with established pathophysiology.

Third, to assess the simulator's utility for ML classification, we performed a categorical validity analysis using a Confusion Matrix and Cohen's Kappa (κ), comparing clinical severity labels against simulated respiratory-rate-based classifications.

Finally, to address the limited cohort size (n=30), the Synthetic Minority Oversampling Technique (SMOTE) was applied to expand the dataset. To assess whether SMOTE preserved the underlying multivariate dependence structure, we computed Pearson correlation matrices before and after augmentation and quantified their absolute difference.

G. Reproducibility and Computational Environment

All statistical analyses were implemented in Python 3.11 using SciPy 1.11, NumPy 1.24, Pandas 2.0, and Statsmodels 0.14. Complementary descriptive analyses were performed in jamovi 2.7. The analytical workflow was fully scripted to facilitate independent verification and future extension of the dataset for machine learning applications.

IV. Results

A. Statistical Separability of Severity Groups

The Kruskal-Wallis test indicated statistically significant separation among the three asthma severity levels for most of the physiological variables evaluated. Table II summarizes the test statistics, p-values, and effect sizes (epsilon-squared, ϵ^2).

As shown in Fig. 1 and Table II, the variables with the largest effect sizes were Respiratory Rate ($\epsilon^2 = 0.890$) and Compliance ($\epsilon^2 = 0.803$), indicating that these biomarkers are highly discriminative for severity stratification. Notably, Total Inspiratory Elastic Work did not differ significantly across groups ($p = .258$), supporting the interpretation that the simulated disease remains primarily resistive (obstructive) rather than elastic, consistent with asthma pathophysiology.

B. Monotonic Trends Across Severity Levels

To assess whether physiological parameters evolved consistently with increasing disease severity, Spearman rank correlation was computed between ordinal severity coding (Mild=1, Moderate=2, Severe=3) and each physiological variable. Table III summarizes the results.

Strong monotonic associations were observed for Respiratory Rate ($\rho = 0.943$), Compliance ($\rho = -0.891$), and Peak Expiratory Flow ($\rho = 0.849$), all with $p < 0.001$. These results support the hypothesis that simulated parameters evolve in physiologically consistent directions as disease severity increases.

TABLE I
MECHANICAL PARAMETERS OF THE VIRTUAL COHORT BY SEVERITY GROUP

Parameter	Group 1 (Mild)	Group 2 (Moderate)	Group 3 (Severe)
Airway Resistance (R)	5–8 $cmH_2O/L/s$	12–18 $cmH_2O/L/s$	20–25 $cmH_2O/L/s$
Lung Compliance (C)	75–90 mL/cmH_2O	40–56 mL/cmH_2O	30–48 mL/cmH_2O
Muscle Pressure (P_{mus})	7–12 cmH_2O	27 cmH_2O (Fixed)	27 cmH_2O (Fixed)
Respiratory Rate (RR)	14–20 bpm	24–29 bpm	30–45 bpm
Insp. Rise Time	40–50%	20% (Fixed)	20% (Fixed)
Insp. Release Time	50–60%	15% (Fixed)	15% (Fixed)

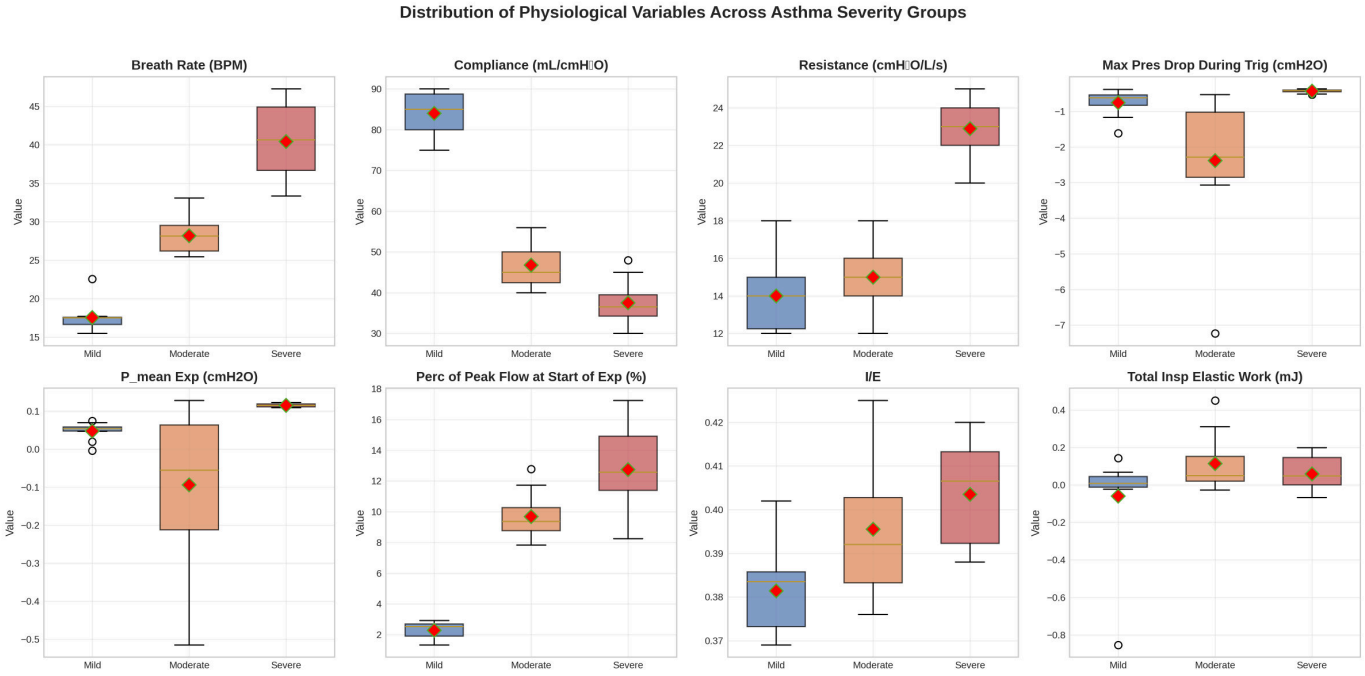


Fig. 1. Distribution of key physiological variables across the three simulated asthma severity groups (Mild, Moderate, Severe). The boxplots illustrate progressive deterioration in respiratory mechanics, including compensatory tachypnea, reduced dynamic compliance, and increased triggering effort.

TABLE II
STATISTICAL SEPARABILITY OF PHYSIOLOGICAL VARIABLES ACROSS SEVERITY GROUPS (KRUSKAL-WALLIS TEST)

Variable	χ^2 (H)	p-value	Effect Size (ϵ^2)
Respiratory Rate (BPM)	25.81	< 0.001	0.890
Compliance (mL/cmH_2O)	23.29	< 0.001	0.803
Peak Exp. Flow (%)	21.68	< 0.001	0.748
Resistance ($cmH_2O/L/s$)	20.27	< 0.001	0.699
Max. Trigger Press. Drop (cmH_2O)	18.88	< 0.001	0.651
Mean Exp. Pressure (cmH_2O)	15.28	< 0.001	0.527
I/E Ratio	10.89	0.004	0.375
Total Insp. Elastic Work (mJ)	2.71	0.258	0.094

TABLE III
SPEARMAN CORRELATION BETWEEN SEVERITY AND PHYSIOLOGICAL VARIABLES

Variable	ρ	p-value
Respiratory Rate (BPM)	0.943	< 0.001
Compliance (mL/cmH_2O)	-0.891	< 0.001
Resistance ($cmH_2O/L/s$)	0.789	< 0.001
Peak Exp. Flow (%)	0.849	< 0.001
Mean Exp. Pressure (cmH_2O)	0.557	0.001
I/E Ratio	0.612	< 0.001
Max. Trigger Press. Drop	0.424	0.019
Total Insp. Elastic Work	0.203	0.282

The non-significant correlation for Total Inspiratory Elastic Work ($\rho = 0.203$, $p = .282$) is consistent with the interpretation that asthma is primarily a resistive rather than elastic disease. The positive correlation for Maximum Trigger Pressure Drop ($\rho = 0.424$, $p = .019$) reflects the non-monotonic behavior discussed in Section IV-C.3, where triggering effort peaks in the moderate group and decreases in the severe group due to the temporal constraints imposed by extreme tachypnea.

C. Emergence of the Pathophysiological Cascade

A core objective of this study was to assess whether secondary physiological variables plausibly emerged from the

programmed mechanical conditions, reproducing patterns consistent with the clinical cascade of acute asthma exacerbations.

1) Compensatory Tachypnea and Reduced Dynamic Compliance: Mean respiratory rate increased progressively from the mild cohort (17.5 ± 1.9 BPM) to the severe cohort (40.4 ± 5.1 BPM). This compensatory tachypnea is consistent with established respiratory physiology, whereby respiratory centers increase neural drive to maintain minute ventilation in the face of increasing airflow limitation and increased work of breathing [11], [12]. Simultaneously, dynamic compliance decreased from 84.0 ± 5.2 to 37.5 ± 5.6 mL/cmH_2O. Although *static* compliance is typically preserved in asthma, the observed reduction is consistent with *dynamic hyperinflation*: as air trapping progresses, the lung operates at higher volumes on the non-linear, steeper portion of the pressure-volume curve, effectively reducing apparent compliance [11].

2) Expiratory Flow Limitation and Intrinsic PEEP: The percentage of peak expiratory flow occurring at the onset of expiration decreased progressively across severity groups ($p < .001$). Physiologically, severe bronchoconstriction increases expiratory resistance and promotes premature airway closure, restricting airflow despite continued expiratory effort [13]. Furthermore, the maximum trigger pressure drop (ΔP_{trig}) exceeded -7.0 cmH_2O in severe profiles. Such increased inspiratory pressure requirements are consistent with intrinsic positive end-expiratory pressure (auto-PEEP). During severe bronchoconstriction, incomplete expiration leads to progressive air trapping; the resulting residual positive pressure must be overcome before inspiratory flow begins, requiring greater negative muscle pressures [13], [14].

3) Non-Linear Triggering Effort Behavior: Interestingly, triggering effort peaked in the moderate group (-2.38 ± 1.95 cmH_2O) and *decreased* in the severe group (-0.43 ± 0.05 cmH_2O). This non-monotonic behavior must be interpreted in the context of the simulator setup. The programmed muscle pressure (P_{mus}) was fixed at 27 cmH_2O for the moderate and severe groups, representing a physiological ceiling of inspiratory capacity. The observed decrease in ΔP_{trig} in the severe group is therefore not an artifact of a simulator limit, but a consequence of extreme tachypnea (cycle time < 1.5 s). The simulated diaphragm lacks the temporal window required to generate the maximum programmed pressure before the next respiratory cycle begins. This behavior is *analogous to the neuromechanical disadvantage observed clinically* in status asthmaticus, where hyperinflation flattens the diaphragm and reduces its pressure-generating capacity [14].

4) Elastic Work and Volume Compensation: An apparent contradiction emerged between the significant reduction in dynamic compliance (C) and the non-significant change in Total Inspiratory Elastic Work ($p = .258$). Theoretically, the elastic work equation $W_{el} = V_T^2/2C$ suggests that halving compliance should double elastic work. However, this analysis ignores the numerator: tidal volume (V_T). In the severe asthma group, dynamic hyperinflation and extreme tachypnea force the patient into a rapid, shallow breathing pattern, drastically reducing V_T. Because V_T is squared in the numerator, its steep reduction mathematically compensates for the decrease in compliance in the denominator. The simulator's ability to capture this complex, multivariable physiological compensation—where mechanical stiffening is counterbalanced by volume restriction—provides evidence supporting its fidelity as a respiratory digital twin.

D. Categorical Validity versus Numerical Bias

To assess the simulator's potential utility for machine learning classification, we performed a dual validation of respiratory rate output.

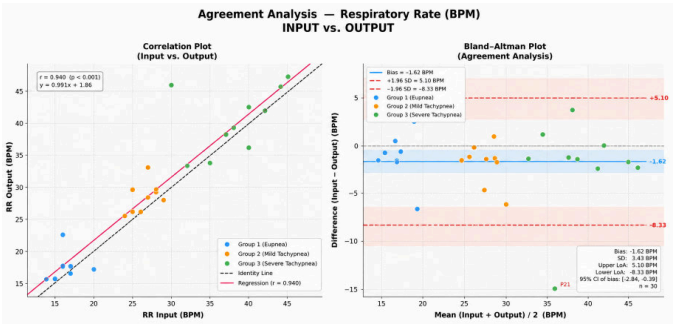


Fig. 2. Bland-Altman plot comparing input respiratory rate (clinical label) against emergent simulated respiratory rate. The plot reveals a systematic numerical bias (-1.62 BPM) and wide limits of agreement, indicating that the simulator does not reproduce exact absolute values.

Fig. 2 presents the Bland-Altman analysis, which revealed a systematic numerical bias (-1.62 BPM, 95% CI: -2.84 to -0.39) and wide limits of agreement (-8.33 to $+5.10$ BPM). Strictly from a metrological standpoint, this indicates that the simulator is not interchangeable with clinical measurements for reproducing absolute values.

However, for ML-based *risk stratification*, categorical separability is the governing metric. Complementary one-way ANOVA revealed highly significant differences between severity groups ($F = 113.0$, $p < 0.001$), with an effect size of $\eta^2 = 0.893$, indicating that 89.3% of the total variance in respiratory rate is explained by clinical severity group. This large effect size confirms that the simulator preserves the clinical decision boundaries required for classification tasks.

As shown in Fig. 3, the confusion matrix demonstrated strong categorical agreement (Cohen's Kappa $\kappa = 0.90$). Threshold-based classification correctly assigned 28 of 30 patients (93.3% accuracy), with errors occurring only between adjacent severity classes. Notably, the severe group achieved a perfect recall of 1.00, ensuring that no critical case was underestimated.

This observation suggests a novel perspective for simulator evaluation: numerical bias does not necessarily invalidate categorical utility. The dataset appears to preserve clinically relevant decision boundaries for training classification algorithms, supporting its potential suitability for ML applications even without exact numerical reproduction.

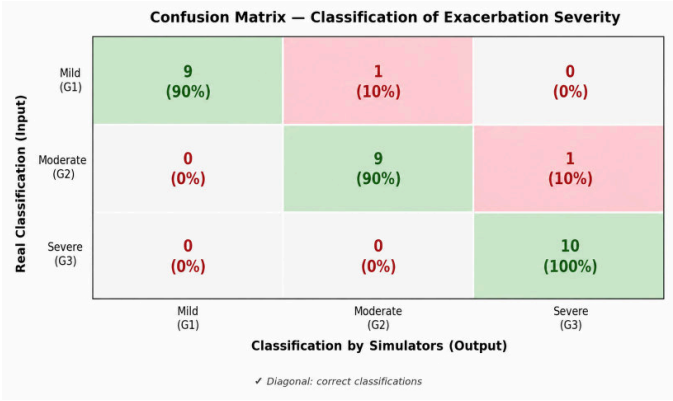


Fig. 3. Confusion matrix for severity classification based on emergent respiratory rate. The strong diagonal agreement ($\kappa = 0.90$) and 100% recall for the Severe group suggest that the dataset preserves clinically relevant decision boundaries, despite the numerical bias observed in Fig. 2.

E. Data Augmentation Preserves Physiological Constraints

To address the limited cohort size ($n = 30$), the Synthetic Minority Oversampling Technique (SMOTE) was applied to expand the dataset. To assess whether SMOTE introduced physiologically implausible patterns, we conducted a two-level integrity evaluation.

Table IV presents the marginal distribution comparison results between the original (Real) and augmented (Synthetic) cohorts. Both parametric (independent t-test) and non-parametric (Mann-Whitney U) tests yielded no significant differences for any of the evaluated variables (all $p > 0.05$).

TABLE IV
DISTRIBUTIONAL INTEGRITY OF SMOTE AUGMENTATION: COMPARISON BETWEEN ORIGINAL AND AUGMENTED COHORTS

Variable	t-test p-value	Mann-Whitney p-value
Respiratory Rate (BPM)	0.928	0.926
I/E Ratio	0.961	0.985
Compliance (mL/cmH_2O)	0.974	0.930
Total Insp. Elastic Work (mJ)	0.875	0.985
Resistance (cmH_2O/L/s)	0.789	0.624
Mean Exp. Pressure (cmH_2O)	0.963	0.900
Max. Trigger Press. Drop (cmH_2O)	0.991	0.914
Peak Exp. Flow (%)	0.845	0.660

Beyond marginal distributions, we assessed whether SMOTE preserved the multivariate dependence structure by computing Pearson correlation matrices before and after augmentation. The mean absolute difference between corresponding correlation coefficients was $\Delta\rho = 0.020$, with a maximum difference of 0.065, indicating strong preservation of inter-variable relationships. Fig. 4 presents the side-by-side comparison of the correlation matrices.

Fig. 5 visually corroborates these findings through Principal Component Analysis, which captured 77.0% of the total variance in the first two components. The silhouette score comparing Real and Synthetic samples was -0.001, indicating substantial overlap between the two populations—consistent with the hypothesis that SMOTE preserved the underlying distribution rather than introducing distinct patterns. The distances between severity-group centroids in PCA space (Mild vs. Moderate: 3.27, Moderate vs. Severe: 3.14, Mild vs. Severe: 4.67) confirm the monotonic separation of disease severity levels.

Together, these results provide evidence supporting that the SMOTE pipeline safely expanded the dataset while maintaining both the marginal distributions and the multivariate dependence structures relevant for reliable machine learning.

F. Comparison with Clinical Reference Ranges

Table V provides a comparison between simulated values and established clinical ranges reported in the literature for severe asthma exacerbations, together with interpretive context for observed deviations.

Most simulated parameters fell within clinically observed ranges, supporting the simulator's ability to reproduce physiologically plausible patterns of severe asthma. Notably, simulated auto-PEEP (0.115 cmH_2O) was substantially lower than the 5–15 cmH_2O typically observed in status asthmaticus [12], [13]. This discrepancy likely reflects the simulator's mechanical constraints and the absence of active expiratory muscle engagement in spontaneous-breathing mode. Nevertheless, the elevated trigger pressure drop ($\Delta P_{trig} = -7.3$ cmH_2O) indicates the presence of intrinsic PEEP that must be overcome before inspiratory flow can begin, partially compensating for this limitation.

Similarly, the simulated I/E ratio (0.403) exceeded the clinically recommended range of 0.25–0.33 for severe asthma [1], suggesting that the simulator's spontaneous-breathing algorithm may not fully capture the prolonged expiratory phase characteristic of obstructive disease. These deviations represent important considerations to be addressed in future work through parameter optimization or hybrid simulation approaches.

V. DISCUSSION

A. Physiological Plausibility and Digital Twins

The results suggest that the ASL5000 reproduces patterns consistent with the pathophysiological cascade of acute asthma exacerbations. Progressive tachypnea, reduced dynamic compliance, expiratory flow limitation, and the emergence of intrinsic PEEP are consistent with established respiratory physiology [11]–[13]. These findings support the interpretation of the simulator as a respiratory digital twin generator capable of producing physiologically coherent datasets for machine learning applications.

Comparison with the clinical literature (Table V) indicates that the simulator captures hallmark features of severe asthma, including compensatory tachypnea, increased airway resistance, and reduced dynamic compliance. However, the underestimation of auto-PEEP and the elevated I/E ratio highlight areas where the simulator's mechanical constraints may limit

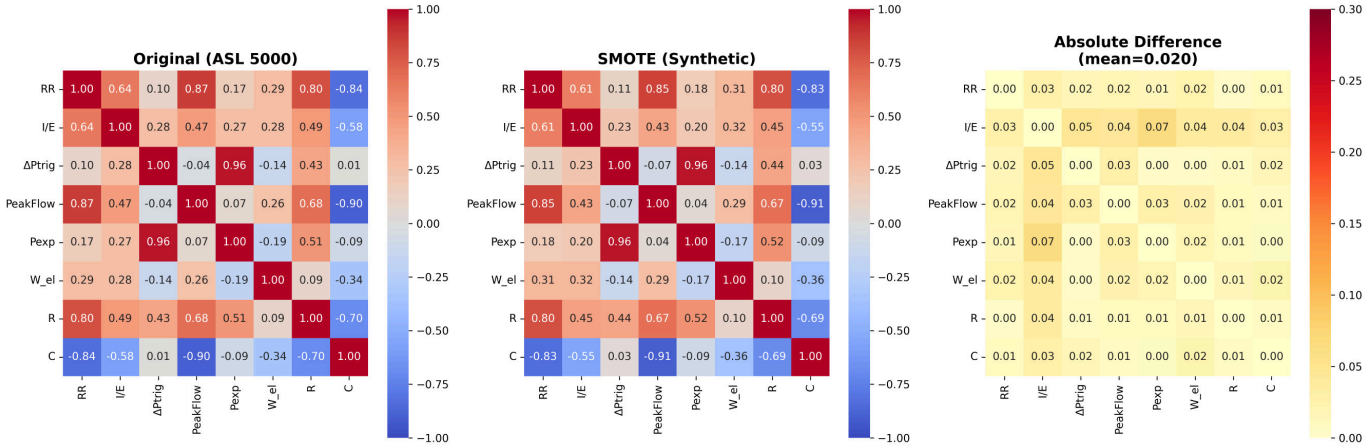


Fig. 4. Preservation of the multivariate dependence structure under SMOTE augmentation. Left: Pearson correlation matrix of the original ASL5000 cohort. Center: Correlation matrix of the SMOTE-augmented cohort. Right: Absolute difference between matrices (mean $\Delta\rho = 0.020$). The low absolute differences support that SMOTE preserved the underlying biomechanical relationships.

TABLE V
COMPARISON OF SIMULATED VALUES WITH CLINICAL REFERENCE RANGES

Variable	**Clinical Range**[a]	**ASL5000 (Severe)**	**Clinical Interpretation**
Respiratory Rate (BPM)	30–45[b]	40.4 ± 5.1	Within range; consistent with severe tachypnea
Resistance ($cmH_2O/L/s$)	15–25[c]	22.9 ± 1.7	Within range; reflects severe bronchoconstriction
Dynamic Compliance (mL/cmH_2O)	30–50[c]	37.5 ± 5.6	Within range; consistent with dynamic hyperinflation
Auto-PEEP (cmH_2O)	5–15[c]	0.115 ± 0.005	Below range[d]; likely due to spontaneous-breathing configuration
ΔP_{trig} (cmH_2O)	-5 to -10	-7.3 (max)	Within range; consistent with intrinsic PEEP compensation
I/E Ratio	0.25–0.33[b]	0.403 ± 0.012	Above range[d]; may reflect simulator timing constraints

[a] Reference ranges for severe asthma exacerbations.
[b] GINA 2025 Guidelines [1].
[c] Kacmarek et al., Egan's Fundamentals [12]; Tobin, Intensive Care Monitoring [13].
[d] Simulated values deviate from clinical ranges; see Discussion for limitations.

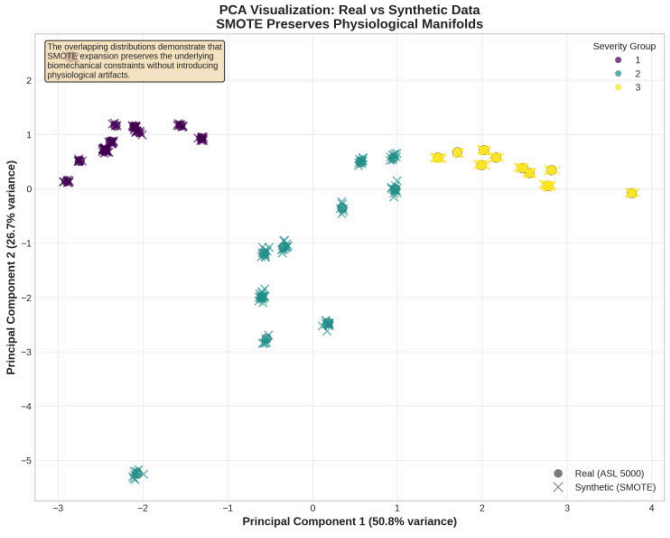


Fig. 5. PCA visualization of Real samples (colored circles) vs. Synthetic samples (orange X markers). The substantial overlap (silhouette score $= -0.001$) and 77.0% variance explained in the first two components support that SMOTE preserved the underlying physiological manifold without introducing distributional shifts.

its physiological fidelity. Future work should explore parameter configurations that more closely approximate clinical auto-PEEP levels and expiratory timing.

B. Categorical versus Numerical Validity: An Observational Perspective

The observation that numerical bias does not necessarily invalidate categorical utility offers a useful perspective for simulator evaluation. Although Bland-Altman analysis revealed systematic bias in absolute respiratory-rate values, the strong Cohen's Kappa ($\kappa = 0.90$) and significant monotonic trends observed via Spearman correlation suggest that the dataset preserves clinically relevant decision boundaries. This distinction is particularly relevant for ML applications focused on classification and risk stratification, where categorical separability is often more important than exact numerical reproduction.

This perspective has implications for the broader field of synthetic data validation. Researchers evaluating simulator-generated datasets may benefit from assessing both numerical agreement (for calibration tasks) and categorical validity (for classification tasks), as these represent distinct validation dimensions with different clinical relevance.

C. Multivariate Integrity of Data Augmentation

The assessment of SMOTE-based augmentation through both marginal and multivariate analyses addresses an important gap in the literature. By demonstrating that synthetic expansion preserves both univariate distributions (all $p > 0.05$) and multivariate dependence structures (mean $\Delta\rho = 0.020$), this work provides a reproducible framework for safely augmenting limited clinical datasets. This approach supports the hypothesis that machine learning models trained on synthetic data will not learn spurious correlations or physiologically implausible patterns.

The PCA visualization (Fig. 5) and the near-zero silhouette score (-0.001) provide complementary evidence that the augmented samples occupy the same physiological manifold as the original simulator outputs, supporting the preservation of causal biomechanical relationships relevant to reliable machine learning.

D. Implications for Edge AI and Embedded Systems

The validated dataset provides a foundation for training lightweight ML models on embedded devices, such as the Respiratory Diagnostic Assistant (RDA) [9]. The high statistical separability ($\epsilon^2 > 0.65$ for key features), categorical validity ($\kappa = 0.90$), and monotonic trends across severity levels suggest that models trained on synthetic data may generalize to clinical scenarios, enabling on-device risk stratification without cloud dependence.

Integration of this validated dataset with the RDA's embedded architecture represents a potential step toward real-time asthma exacerbation prediction in hospital and remote-monitoring settings. Future work will focus on deploying ML models trained on this dataset and evaluating their performance against clinical outcomes.

E. Limitations

This study has several limitations. First, it relies exclusively on simulated data, requiring external validation with real-world asthma cohorts. Second, the proof-of-concept cohort ($n = 30$) is not population representative and should be expanded for future ML development. Third, only two primary mechanical parameters (R and C) were manipulated; future studies should incorporate respiratory muscle strength, heterogeneous airway obstruction, and variable neural drive. Fourth, the simulator does not fully reproduce inflammatory processes, mucus dynamics, or patient-specific anatomical variability. Finally, the underestimation of auto-PEEP and the elevated I/E ratio indicate mechanical constraints that should be addressed in future simulation protocols.

These limitations do not invalidate the findings but highlight opportunities for future refinement and validation.

VI. Conclusion

This study presented a physiological validation framework for synthetic ventilatory datasets generated with the ASL5000 breathing simulator. The results demonstrate that the simulator reproduces physiologically plausible patterns of asthma, including compensatory tachypnea, reduced dynamic compliance, and intrinsic PEEP. Strong monotonic trends across severity levels (Spearman $\rho > 0.78$, $p < 0.001$) support the physiological consistency of the simulated data. Furthermore, the findings suggest that numerical bias does not necessarily compromise categorical utility for ML classification ($\kappa = 0.90$). Finally, the proposed SMOTE-based augmentation pipeline preserved both marginal distributions and multivariate dependence structures (mean $\Delta\rho = 0.020$). Together, these results provide a reproducible foundation for developing edge-AI respiratory monitoring systems based on physiologically validated synthetic datasets.

Acknowledgment

This work was supported by CNPq (Grant No. 303988/2025-8), FACEPE (Grant No. APQ-1020-4.08/25), CAPES (Finance Code 001), and the PIBIC/PIBITI programs, including PIBIC/CNPq (Grant No. PIG2849-2023), PIBITI/CNPq (Grant No. PIG2767-2025), and PIBIC/UFPE (voluntary) (Grant No. PIG2767-2025). Institutional support from the Federal University of Pernambuco (UFPE) through PROPG Pró-Equipamentos and PROPESQI–PARQTEC-UEI LINDEF is also gratefully acknowledged.

References

[1] Global Initiative for Asthma (GINA), "Global Strategy for Asthma Management and Prevention 2025 Update," Fontana, WI, USA, 2025.

[2] N. A. Molfino, G. Turcatel, and D. Riskin, "Machine Learning Approaches to Predict Asthma Exacerbations: A Narrative Review," *Advances in Therapy*, vol. 41, pp. 534–552, 2024.

[3] S. Xiong et al., "Machine Learning for Prediction of Asthma Exacerbations Among Asthmatic Patients: A Systematic Review and Meta-analysis," *BMC Pulmonary Medicine*, vol. 23, no. 278, 2023.

[4] W. K. D. Jayamini et al., "Investigating Machine Learning Techniques for Predicting Risk of Asthma Exacerbations: A Systematic Review," *Journal of Medical Systems*, vol. 48, no. 49, 2024.

[5] T. M. Mahdavi and J. Rees, "Synthetic Healthcare Data: Opportunities and Challenges for Machine Learning Applications," *Artificial Intelligence in Medicine*, vol. 146, 2025.

[6] R. Miotto et al., "Deep Learning for Healthcare: Review, Opportunities and Challenges," *Briefings in Bioinformatics*, vol. 23, no. 1, 2022.

[7] IngMar Medical, *ASL 5000 Breathing Simulator User Manual*. Pittsburgh, PA, USA, 2024.

[8] D. L. Vines et al., "Validation of the ASL 5000 Breathing Simulator for Respiratory Research Applications," *Respiratory Care*, vol. 67, no. 9, pp. 1152–1163, 2022.

[9] I. M. Campos, "Arquitetura de software para Sistema de análise respiratória: Respiratory Diagnostic Assistant," B.S. thesis, Dept. Elect. Eng., Univ. Federal de Pernambuco, Recife, Brazil, 2025.

[10] A. Thelen et al., "A Comprehensive Review of Digital Twin—Part 1: Modeling and Twinning Enabling Technologies," *arXiv preprint* arXiv:2208.14197, 2022.

[11] J. B. West, *Respiratory Physiology: The Essentials*. 10th ed. Philadelphia, PA: Wolters Kluwer, 2015.

[12] R. M. Kacmarek, J. K. Stoller, and A. J. Heuer, *Egan's Fundamentals of Respiratory Care*. 12th ed. St. Louis, MO: Elsevier, 2020.

[13] M. J. Tobin, *Principles and Practice of Intensive Care Monitoring*. New York, NY: McGraw-Hill, 2013.

[14] C. Haberthür and J. Guttmann, "Mechanical ventilation and respiratory muscle function," *Current Opinion in Critical Care*, vol. 16, no. 1, pp. 45–50, 2010.

Safety and Efficacy Evaluation of Rare-earth Doped Bioactive Glasses for Bone Regeneration: A Computational Approach

Daniel Moro Druzian*, Vinícius Rodrigues Oviedo*, Lourdes Maria Muraro Favarin*, Marcos Vinicius Pasqualoto Prior*, Luiz Batista Cardoso*, William Leonardo da Silva*

* *Applied Nanomaterials Research Group (GPNAp)*
Franciscan University (UFN)
Santa Maria, RS, Brazil
daniel.druzian@ufn.edu.br; vinicius.oviedo@ufn.edu.br; lourdes.favarin@ufn.edu.br;
marcos.prior@ufn.edu.br; luiz.bcardoso@ufn.edu.br; w.silva@ufn.edu.br

***Abstract*—This work aims to investigate the pharmacokinetic and toxicological profiles of bioactive glasses (BG) doped with 5% rare-earth oxides (Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3) for bone regeneration using *in silico* assays. The results indicated that incorporating these rare-earth elements preserved the intrinsic potential bioactivity and biological safety of the original glass matrix, exhibiting absorption and distribution profiles ideal for strict local action, with no risk of systemic translocation or penetration into the central nervous system. Additionally, metabolic predictions showed no interactions with the cytochrome P450 system, and all formulations were classified under GHS Toxicity Class 5, reflecting exceptionally low acute oral toxicity (3000 and 4000 mg kg^{-1}). However, Eu_2O_3 exhibited lower safety thresholds, indicating the need for strict dose control in subsequent stages. Therefore, computational simulations successfully validated the sustainable clinical viability of these new biomaterials, aligning technological innovation and quality of life with the Sustainable Development Goals.**

***Index Terms*—bone regeneration, inorganic compound, rare earths, toxicity prediction**

I. Introduction

Degenerative, inflammatory conditions, and fractures in bone tissue affect millions worldwide, often caused by calcium or phosphorus deficiency, and primarily resulting from fall-prone environments, traffic accidents, and genetic disorders that replace natural tissues with biomaterials [1]. Several statistics highlight this issue: (*i*) Osteoarthritis affects 10 to 17% of people over 60 globally; (*ii*) Osteoporosis impacts 200 million individuals worldwide; and (*iii*) Approximately 9 million osteoporotic fractures occur annually worldwide, with projections indicating that by 2050, hip fractures could increase by up to 240% in women and 310% in men [2], [3]. In this context, bone cement emerges as a promising porous material for the regeneration and repair of various functional tissues [4].

Bioactive glasses (BG) are an inorganic compound that typically contains various oxides (e.g., SiO_2, P_2O_5, and CaO), and have been used in prosthetic devices, implants, and scaffolds, since apatite layer (Ap) is formed when in contact with biological fluids [5]. It is worth noting that the addition of rare-earth elements expands the applicability of BG by modifying the surface energy of the glass system, promoting the formation of the apatite layer, and reducing the Ca/P molar ratio required for bone formation [6].

Incorporating rare-earth elements (REEs) (e.g., europium, neodymium, lanthanum, cerium, and erbium) into BG methods (e.g., sol-gel, melt-quenching, or robocasting) represents a promising strategy to expand the functionality of these materials beyond bone regeneration [7]. Acting as network modifiers or intermediates within the glass structure, these ions confer advanced therapeutic and diagnostic properties: the alternation of oxidation states mitigates oxidative stress by scavenging reactive oxygen species (ROS), while the luminescent or magnetic characteristics of other REEs enable bioimaging [8]. Furthermore, precise control over ion-release kinetics and material degradation enables regulation of essential biological responses stimulating angiogenesis, osteoblastic proliferation, and antibacterial activity, thereby opening new avenues for personalized treatments in cancer therapy, nerve regeneration, and controlled drug delivery [9].

In silico tests involve computational simulations that model and analyze biological and chemical processes, serving as a vital tool in contemporary scientific research. They offer benefits (e.g., lower costs, and large-scale analytical capabilities), where the exploration of complex hypotheses, and, crucially, reduced reliance on animal testing [10]. Moreover, biomaterials for bone regeneration are used to simulate mechanical behavior by mimicking the bone environment and to estimate the rate of osseointegration and bone growth around the implant [11].

In this context, this study aims to investigate aspects of rare-earth-doped bioglasses through *in silico* assays. Furthermore, it seeks to drive the economic development of rare-earth-doped bioglasses and add value to the product in both the short and long term. It is worth noting that the study aims to contribute to sustainable development and quality of life, aligning with

Fundação de Amparo à pesquisa do Estado do RS (FAPERGS)

TABLE I
CHARACTERISTICS OF THE BG AND BG DOPED WITH 5% Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, AND Er_2O_3, USED FOR *in silico* TESTS.

Samples	Smiles source	Molecular formula	MW (g mol^{-1})	S_{BET} (m^2 g^{-1})
BG	O=[Si](O)O[Si](=O)[O-].[Ca++].[O-]P(=O)(O)O	$58SiO_2$-$12P_2O_5$-25CaO	274.23	105.03
BG+Eu_2O_3	O=[Si](O)O[Si](=O)[O-].[Ca++].[Eu+++].[O-]P(=O)(O)O	$58SiO_2$-$12P_2O_5$-25CaO-$5Eu_2O_3$	426.20	152.07
BG+Nd_2O_3	O=[Si](O)O[Si](=O)[O-].[Ca++].[Nd+++].[O-]P(=O)(O)O	$58SiO_2$-$12P_2O_5$-25CaO-$5Nd_2O_3$	418.47	144.06
BG+La_2O_3	O=[Si](O)O[Si](=O)[O-].[Ca++].[La+++].[O-]P(=O)(O)O	$58SiO_2$-$12P_2O_5$-25CaO-$5La_2O_3$	413.14	139.13
BG+CeO_2	O=[Si](O)O[Si](=O)[O-].[Ca++].[Ce+++].[O-]P(=O)(O)O	$58SiO_2$-$12P_2O_5$-25CaO-$5CeO_2$	414.35	144.91
BG+Er_2O_3	O=[Si](O)O[Si](=O)[O-].[Ca++].[Er++].[O-]P(=O)(O)O	$58SiO_2$-$12P_2O_5$-25CaO-$5Er_2O_3$	441.49	140.74

Abbreviations: MW = Molecular weight; and S_{BET} = Surface area.

the Sustainable Development Goals (SDGs) established by the United Nations (UN) with an emphasis on SDG 3 (Ensure healthy lives and promote well-being for all at all ages) and SDG 9 (Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation) [12].

II. MATERIALS AND METHODS

A. Study on the toxicity prediction of bioglass

A predictive study was conducted to estimate the potential toxicological risks of the main compounds present in BG and BG doped with 5% Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3, based on molecular similarity, fragmentation patterns, and common machine-learning features, using established models (*in vitro* and *in vivo*). The computational tools employed were pkCSM and ProTox-II 3.0 [13]–[16]. Table 1 and Figure 1 present the characteristics of the BG and BG doped with 5% Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3, which were used in the *in silico* tests. Furthermore, the information used was obtained from PubChem and the Materials Project [17].

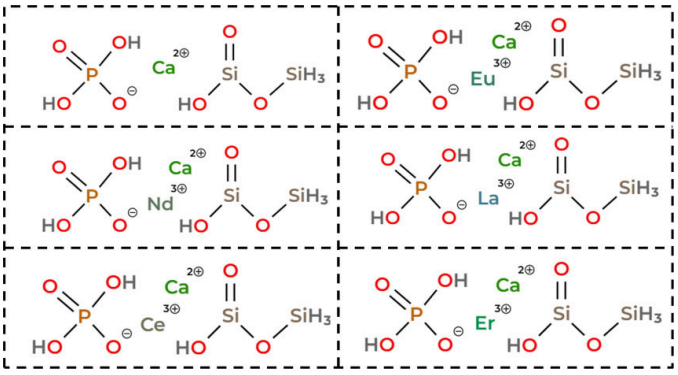


Fig. 1. Schematic representation of the 2D molecular structures of (a) BG; (b) BG+Eu_2O_3; (c) BG+Nd_2O_3; (d) BG+La_2O_3; (e) BG+CeO_2; and (f) BG+Er_2O_3 used in the *in silico* studies.

B. Studies of prediction of mutagenicity and carcinogenicity of bioglass

The online software tools pkCSM and ProTox-II 3.0 were used to predict the mutagenic and carcinogenic potential of BG and BG doped with 5% Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3 components. In this context, the Ames test was employed to predict mutagenicity, enabling the detection of mutations in the genetic material involved in histidine synthesis, typically using bacterial strains such as *Salmonella typhimurium* (TA98, TA100, and TA1535). The prediction of the carcinogenic potential of the nanobiomaterials in rodents was based on data from the National Toxicology Program (NTP).

C. Study of the prediction of adverse effects of bioglass

The prediction of adverse effects for BG and BG doped with 5% Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3 (e.g., arrhythmia, heart failure, hepatotoxicity, nephrotoxicity, and myocardial infarction) was performed using the online software Pro-Tox II 3.0. In this context, the occurrence of adverse effects was considered when the probability of "being active" (Pa) was < 0.7.

D. Study of prediction of acute and chronic toxicity in rats of bioglass

The ProTox-II software was used to determine the acute and chronic toxicity in rats for BG and BG doped with 5% Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3, including, for example, hepatotoxicity, immunotoxicity, cytotoxicity, and the maximum tolerated dose in humans. Additionally, the acute toxicity class in rats was defined according to the Globally Harmonized System of Classification and Labeling of Chemicals (GHS) [18].

III. RESULTS AND DISCUSSION

A. Prediction study of acute toxicity, and adverse effects of components of BG using model pKCSM

Tables 2 and 3 present the *in silico* predictions of absorption, distribution, excretion, metabolism, and toxicity for BG and rare-earth-doped BG, using the pKCSM model.

According to Table 2, *in silico* predictive data for rare-earth-doped bioactive glass (BG) indicated a profile typical of inorganic biomaterials designed for local action, with low permeability values in Caco-2 cells (Log $P_{app} < 0.90$) and a low percentage of human intestinal absorption (less than 30%), indicating that these materials exhibited a poor oral absorption. In the scientific literature, this behavior is widely expected and considered positive, as bioactive glasses are designed to act directly at the surgical site or lesion, releasing ionic dissolution products in a controlled manner into the local microenvironment without the risk of translocation or unwanted

systemic bioaccumulation [19]. Conversely, parameters indicated a moderate water solubility (log mol L^{-1} between -2.2 and -2.8) and a moderate-to-low skin permeability (log Kp of -2.7), reinforcing the impact of rare-earth doping on stabilizing the glass network and indicating more controlled dissolution rates in glasses containing oxides (e.g., CeO_2 and La_2O_3) [20]. Additionally, the fact that most samples do not function as P-glycoprotein (P-gp) inhibitors is a favorable indicator of toxicological biocompatibility, suggesting that the material does not interfere with the natural cellular efflux mechanisms of other essential compounds. It is worth noting that the behavior of Eu_2O_3 and Nd_2O_3 as P-gp substrates suggests specific cellular dynamics involving the active pumping of these ions, an aspect to be validated in future *in vitro* and *in vivo* assays.

For the distribution (Table 2), *in silico* values indicated that the BG and its rare-earth-doped variants exhibited low systemic dispersion potential and severely restricted access to the central nervous system (CNS). Furthermore, the low steady-state volume of distribution (VD_{ss} below -0.15 log L kg^{-1}) indicated that ionic byproducts tend to remain confined to the plasma or local extracellular fluid rather than distributing widely throughout body tissues. In this regard, the high plasma free fraction (Fu) ranging from 0.767 to 0.893, suggested a weak binding affinity to plasma proteins, favoring the immediate local availability of the ions. Crucially, blood-brain barrier permeability (log BB $<$ -1) and CNS penetration (log PS $<$ -3) values categorically classify all samples as unable to cross these complex biological barriers. In the biomaterials literature, this restricted distribution behavior is highly favorable and characterizes the material as biologically safe, as it ensures that the action of the rare-earth oxides remains limited to the peripheral injury site, thereby mitigating the risk of neurotoxicity or systemic side effects resulting from central efflux [21].

For the excretion (Table 2), the rare-earth-doped bioactive glass (BG) revealed a moderately slow body clearance rate and the absence of active renal transport mediated by cationic transporters. Total clearance values range from -0.479 to -0.802 log mL min^{-1} kg^{-1}, demonstrated that the incorporation of rare-earth oxides promoted a slight reduction in the overall elimination rate compared to the base BG. This reduction is consistent with the literature, as the presence of lanthanides alters ionic dissolution kinetics and the formation of secondary complexes within the body, thereby extending the residence time of these elements in the microenvironment [22]. Furthermore, all formulations were predicted not to be substrates of renal organic cation transporter. This indicates that the excretion of these ions did not rely on active tubular secretion via this pathway, thereby eliminating the risk of competitive interactions or renal toxicity from blocking this transporter.

In silico metabolic predictions (Table 3) from the pkCSM platform indicate that neither BG nor its rare-earth oxide-doped variants interact with the cytochrome P450 enzyme system. All samples were strictly categorized as "No" for inhibition of or acting as substrates for the key hepatic isoenzymes evaluated (CYP2D6, CYP3A4, CYP1A2, CYP2C19, and CYP2C9). These results are highly favorable, indicating that the biomaterial's ionic dissolution products pose no risk of inducing unwanted drug interactions or undergoing enzymatic biotransformation via these classic hepatic pathways.

Analysis *in silico* toxicity parameters (Table 3) revealed that both BG and the lanthanide-modified systems exhibited a highly favorable biological safety profile, with a low probability of serious adverse effects. Consequently, all formulations were predicted to be non-mutagenic, non-cardiotoxic, non-hepatotoxic, and non-sensitizing to the skin. Quantitatively, the maximum tolerated dose (MTD) for humans remained within a narrow, safe range across all groups (between 0.771 and 0.898 log mg kg^{-1} day^{-1}). Regarding acute oral toxicity in rats (LD_{50}) and chronic toxicity (LOAEL), the compounds doped with Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3 showed values close to or slightly higher than those of BG, indicating the preservation of the base material's excellent intrinsic biocompatibility. The only notable deviation occurred with BG+Eu_2O_3, which exhibited a decrease in LD_{50} (1.505 log mol kg^{-1}) and LOAEL (0.991 log kg bw^{-1} day^{-1}) values; this suggests that the specific release of europium ions requires closer analytical attention and controlled dosing in subsequent *in vitro* studies to avoid potential cumulative local cytotoxic responses.

B. Prediction study of hepatotoxicity, neurotoxicity, nephrotoxicity, cardiotoxicity, carcinogenicity, immunotoxicity, mutagenicity, and cytotoxicity of BG using model ProTox-II

Figure 2 presents the predictive toxicity profile (ProTox-II 3.0) obtained *in silico* for the BG and its formulations modified with various rare-earth oxides, covering a wide range of cellular biomarkers, organs, and systems. Notably, the most striking feature is the similarity in the curves' behavior across all studied samples, which maintain high predicted biocompatibility (mostly above 75–80%) for all parameters. This graphical parallelism demonstrated that incorporating Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3 did not compromise the intrinsic biological safety of the original bioactive glass, preserving a toxicologically safe and consistent profile with respect to hormonal receptors, oxidative stress, and metabolic pathways.

Despite overall stability, specific, and synchronized declines were observed in three biomarkers across all samples: nephrotoxicity, Transthyretin (TTR) transporter, and Pregnane X receptor (PXR). These identical drops suggested that such responses were inherent to the BG matrix itself or represent the predictive model's standard response to inorganic silicate compounds, rather than a deleterious effect induced by the dopants [23], [24]. Thus, the visual results reinforce the biotechnological potential of these newly doped matrices, indicating that the additional therapeutic properties of rare-earth elements can be exploited with a minimal risk of systemic toxicity or cellular effects that differ from those of the reference biomaterial.

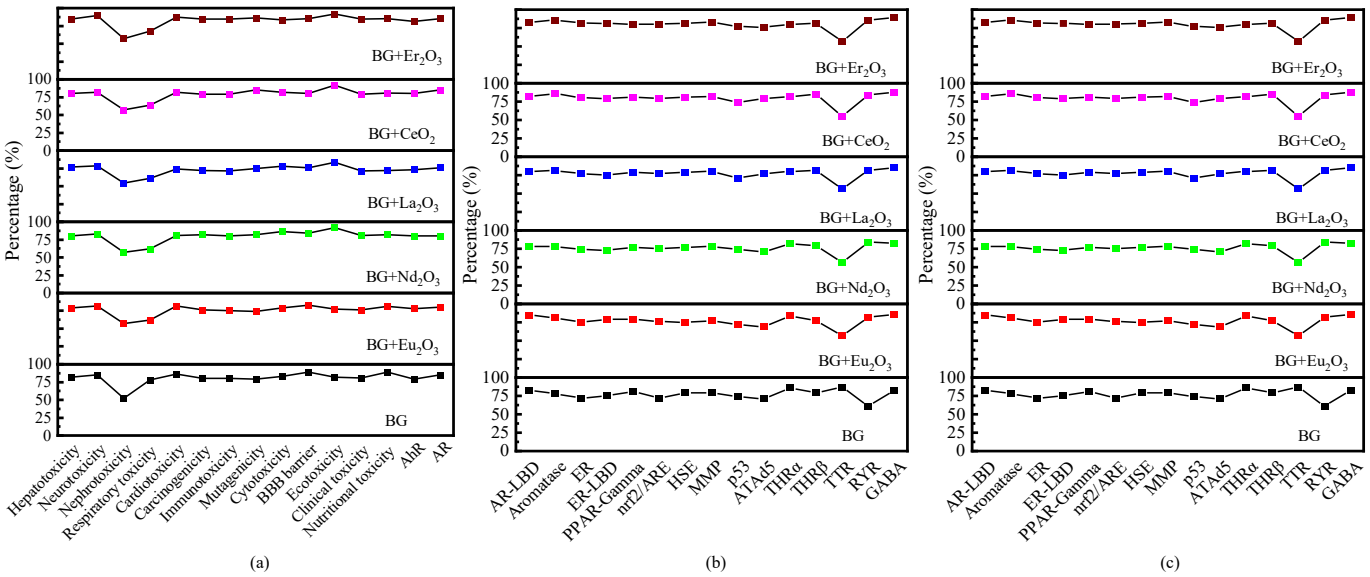


Fig. 2. Predictive toxicity profile (Pred-Tox) of pure bioactive glass (BG) and bioactive glasses modified with rare-earth element oxides (BG+Eu_2O_3, BG+Nd_2O_3, BG+La_2O_3, BG+CeO_2, and BG+Er_2O_3) across different organs, biological systems, and cellular biomarkers. (a) Organ toxicity profile; (b) biological system toxicity profile; and (c) cellular biomarker toxicity profile. **Abbreviations:** AhR = Aryl hydrocarbon receptor; AR = Androgen receptor; AR-LBD = Androgen receptor ligand-binding domain; ER = Estrogen receptor alpha; ER-LBD = Estrogen receptor ligand-binding domain; PPAR-γ = Peroxisome proliferator-activated receptor gamma; nrf2/ARE = Nuclear factor (erythroid-derived 2)-like 2/antioxidant response element; HSE = Heat shock factor response element; MMP = Mitochondrial membrane potential; p53 = Tumor suppressor phosphoprotein p53; ATAD5 = ATPase family AAA domain-containing protein 5; THRα = Thyroid hormone receptor alpha; THRβ = Thyroid hormone receptor beta; TTR = Transthyretin; RYR = Ryanodine receptor; GABAR = GABA receptor; NMDAR = N-methyl-D-aspartate receptor; AMPAR = α-amino-3-hydroxy-5-methyl-4-isoxazolepropionate receptor; KAR = Kainate receptor; AChE = Acetylcholinesterase; CAR = Constitutive androstane receptor; PXR = Pregnane X receptor; NADHOX = NADH-quinone oxidoreductase; VGSC = Voltage-gated sodium channel; and NIS = Na^+/I^- symporter.

TABLE II
In silico TOXICITY PREDICTION (ABSORPTION, DISTRIBUTION, AND EXCRETION) STUDIES USING THE PKCSM MODEL FOR BG, BG+Eu_2O_3, BG+Nd_2O_3, BG+La_2O_3, BG+CeO_2, AND BG+Er_2O_3.

Absorption							
Samples	**Caco-2 permeability (log Papp in 10^{-6} cm s^{-1})**	**Water solubility (log mol L^{-1})**	**Intestinal absorption (human) (%)**	**Skin permeability (log Kp cm h^{-1})**	**P-glycoprotein substrate (Yes/No)**	**P-glycoprotein I inhibition (Yes/No)**	**P-glycoprotein II inhibition (Yes/No)**
BG	0.085	-2.218	24.16	-2.771	No	No	No
BG+Eu_2O_3	-0.364	-2.808	25.76	-2.765	Yes	No	No
BG+Nd_2O_3	-0.362	-2.741	27.44	-2.763	Yes	No	No
BG+La_2O_3	0.112	-2.681	17.42	-2.754	No	No	No
BG+CeO_2	0.114	-2.726	16.20	-2.751	No	No	No
BG+Er_2O_3	0.113	-2.694	17.08	-2.752	No	No	No

Distribution				
Samples	**VDss (human) (log L kg^{-1})**	**Fraction unbound (Fu)**	**BBB permeability (log BB)**	**CNS permeability (log PS)**
BG	-1.106	0.893	-1.449	-3.937
BG+Eu_2O_3	-1.265	0.767	-1.645	-4.562
BG+Nd_2O_3	-1.144	0.854	-1.566	-3.957
BG+La_2O_3	-1.146	0.859	-1.559	-3.957
BG+CeO_2	-1.144	0.854	-1.568	-3.955
BG+Er_2O_3	1.145	0.857	-1.561	-3.956

C. Prediction study of acute toxicity in rats of BG using model ProTox-II

Figure 3 presents the results of *in silico* predictions of acute oral toxicity in rats, comparing the BG with formulations modified with rare-earth oxides, using ProTox-II 3.0 software, where all analyzed samples exhibited median lethal dose (LD_{50}) values strictly within the 3000 to 4000 mg kg^{-1} range. According to GHS guidelines, substances with LD_{50}

TABLE III

In silico TOXICITY PREDICTION (METABOLISM AND TOXICITY) STUDIES USING THE PKCSM MODEL FOR BG, BG+Eu_2O_3, BG+Nd_2O_3, BG+La_2O_3, BG+CeO_2, AND BG+Er_2O_3.

Metabolism							
Samples	**CYP2D6 (Yes/No)**	**CYP3A4 (Yes/No)**	**CYP1A2 (Yes/No)**	**CYP2C19 (Yes/No)**	**CYP2C9 (Yes/No)**	**sCYP2D6 (Yes/No)**	**sCYP3A4 (Yes/No)**
BG	No	No	No	No	No	No	No
BG+Eu_2O_3	No	No	No	No	No	No	No
BG+Nd_2O_3	No	No	No	No	No	No	No
BG+La_2O_3	No	No	No	No	No	No	No
BG+CeO_2	No	No	No	No	No	No	No
BG+Er_2O_3	No	No	No	No	No	No	No

Toxicity								
Samples	**AMES toxicity (Yes/No)**	**Max. tolerated dose (human) (log mg kg^{-1} day^{-1})**	**hERG I inhibitor (Yes/No)**	**hERG II inhibitor (Yes/No)**	**Oral rat acute toxicity (LD50) (log mol kg^{-1})**	**Oral rat chronic toxicity (LOAEL) (log kg_bw^{-1} day^{-1})**	**Hepatotoxicity (Yes/No)**	**Skin sensitization (Yes/No)**
BG	No	0.898	No	No	2.657	2.646	No	No
BG+Eu_2O_3	No	0.881	No	No	1.505	0.991	No	No
BG+Nd_2O_3	No	0.772	No	No	2.847	2.620	No	No
BG+La_2O_3	No	0.780	No	No	2.850	2.621	No	No
BG+CeO_2	No	0.771	No	No	2.846	2.625	No	No
BG+Er_2O_3	No	0.778	No	No	2.849	2.622	No	No

Excretion		
Samples	**Total clearance (log mL min-1 kg-1)**	**Renal OCT2 substrate (Yes/No)**
BG	-0.479	No
BG+Eu_2O_3	-0.802	No
BG+Nd_2O_3	-0.738	No
BG+La_2O_3	-0.709	No
BG+CeO_2	-0.743	No
BG+Er_2O_3	-0.718	No

Abbreviations: VDss = Volume of distribution; BBB = Blood-brain barrier; CNS = Central nervous system; and OCT2 = Organic cation transporter.

values between 2000 and 5000 mg kg^{-1} fall into toxicity class 5, indicating compounds with exceptionally low acute oral toxicity (REFERENCE de guia GHS). Interestingly, the incorporation of all rare-earth oxides (Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3) resulted in higher LD_{50} values compared to the BG, indicating that a larger quantity of the substance was required to induce lethality. Thus, these data suggest that rare-earth doping not only maintained the biological safety profile of the original matrix but theoretically further reduced the potential for systemic acute toxicity. This result reinforces the clinical viability of these new biomaterials for regenerative and tissue engineering applications.

In the experimental literature on rare-earth-containing bioactive glasses (BG), a strong convergence of data was observed, such as: *in vivo* studies and *in vitro* cytotoxicity assays report that the incorporation of oxides (e.g., CeO_2, and La_2O_3) into bioglass matrices does not induce severe cytotoxic responses at controlled concentrations [25]. It is worth noting that the slightly more restrictive behavior of Eu_2O_3 observed in other toxicity parameters was also corroborated by the literature, where lanthanides with more pronounced luminescent properties or distinct coordination kinetics require narrower therapeutic windows to avoid local cellular stress; this validates the sensitivity of the computational predictions obtained when compared against benchtop findings [26].

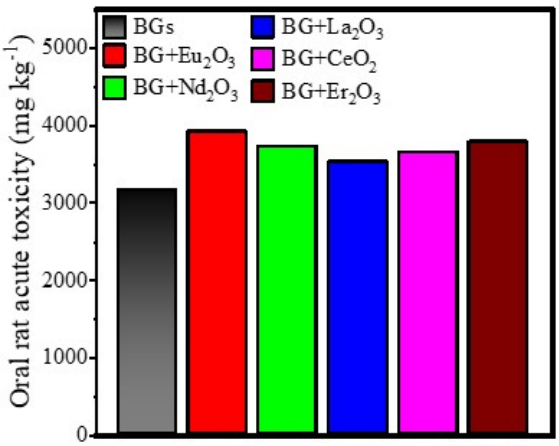


Fig. 3. In silico prediction of acute oral toxicity in rats (mg kg^{-1}) for pure bioactive glass (BG) and bioactive glasses modified with different rare-earth element oxides (BG+Eu_2O_3, BG+Nd_2O_3, BG+La_2O_3, BG+CeO_2, and BG+Er_2O_3) using Pro-Tox II 3.0.

IV. Conclusion

Based on assays performed on the pkCSM and ProTox-II 3.0 platforms, this study demonstrated that incorporating 5 wt.% rare-earth oxides (Eu_2O_3, Nd_2O_3, La_2O_3, CeO_2, and Er_2O_3) onto bioactive glass (BG) generally preserved the intrinsic biocompatibility of the original matrix. Predictive results regarding absorption and distribution confirmed a safe local action profile, characterized by low intestinal and systemic permeability, ensuring controlled ion release directly at the injury site without the risk of unwanted bioaccumulation or neurotoxicity, as the samples proved unable to cross the blood-brain barrier. Furthermore, the absence of interactions with cytochrome P450 isoenzymes eliminated the risk of adverse drug reactions or complex hepatic biotransformation. Regarding acute toxicity, all formulations were classified as GHS Class 5, with LD_{50} values between 3000 and 4000 mg kg^{-1}, indicating low oral toxicity and reinforcing the clinical viability of these new biomaterials. It is worth noting that Eu_2O_3 exhibited declines in LD_{50} and LOAEL values, indicating that the release of Eu^{3+} ions requires stricter dosage monitoring in subsequent experimental stages. Briefly, computational screening validated the high biotechnological potential of these doped matrices for tissue engineering, establishing a sustainable methodological approach that reduces reliance on animal testing and aligns the product's scientific and economic advancement with the UN Sustainable Development Goals, particularly SDGs 3 and 9.

Acknowledgment

This study was financed by the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS, Process No. 25/2551-0002598-2).

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Cope, K. Ourradi, Y. Li, and M. Sharif, "Models of osteoarthritis: the good, the bad and the promising," *Osteoarthritis and Cartilage*, vol. 27, no. 2, pp. 230–239, 2019.

[2] I. Llorente, N. García-Castañeda, C. Valero, I. González-Álvaro, and S. Castañeda, "Osteoporosis in rheumatoid arthritis: Dangerous liaisons," *Frontiers in Medicine*, vol. 7, no. 1, pp. 1–11, 2020.

[3] L. A. Salman, G. Ahmed, S. G. Dakin, B. Kendrick, and A. Price, "Osteoarthritis: a narrative review of molecular approaches to disease management," *Arthritis Research & Therapy*, vol. 25, no. 1, pp. 1–9, 2023.

[4] L. Polo-Corrales, M. Latorre-Esteves, and J. E. Ramirez-Vick, "Scaffold design for bone regeneration," *Journal of Nanoscience and Nanotechnology*, vol. 14, no. 1, pp. 15–56, 2014.

[5] S. Hosseinpour, R. Fekrazad, P. R. Arany, and Q. Ye, "Molecular impacts of photobiomodulation on bone regeneration: A systematic review," *Progress in Biophysics and Molecular Biology*, vol. 149, no. 1, pp. 147–159, 2019.

[6] M. Ajith, M. Aswathi, E. Priyadarshini, and P. Rajamani, "Recent innovations of nanotechnology in water treatment: A comprehensive review," *Bioresource Technology*, vol. 342, no. 1, pp. 1–14, 2021.

[7] Z. Chen, X. Zhou, M. Mo, X. Hu, J. Liu, and L. Chen, "Systematic review of the osteogenic effect of rare earth nanomaterials and the underlying mechanisms," *Journal of Nanobiotechnology*, vol. 22, no. 1, pp. 1–28, 2024.

[8] A. Moghanian, S. Safaee, A. A. Kızılkurtlu, and R. Farmani, "Advanced rare-earth doped bioactive glasses: A comprehensive review on structural, biological, and mechanical improvements," *Silicon*, vol. 18, no. 2, pp. 471–498, 2025.

[9] S. V. Yudintsev, M. I. Ojovan, and V. I. Malkovsky, "Thermal effects and glass crystallization in composite matrices for immobilization of the rare-earth element–minor actinide fraction of high-level radioactive waste," *Journal of Composites Science*, vol. 8, no. 2, pp. 1–24, 2024.

[10] R. Sutcliffe, C. P. Doherty, H. P. Morgan, N. J. Dunne, and H. O. McCarthy, "Strategies for the design of biomimetic cell-penetrating peptides using AI-driven *in silico* tools for drug delivery," *Biomaterials Advances*, vol. 169, no. 1, pp. 1–9, 2025.

[11] M. M. Alves, L. M. Marques, I. Nogueira, C. F. Santos, S. B. Salazar, S. Eugénio, N. P. Mira, and M. Montemor, "*In silico*, *in vitro* and antifungal activity of the surface layers formed on zinc during this biomaterial degradation," *Applied Surface Science*, vol. 447, no. 1, pp. 401–407, 2018.

[12] M. Larrañaga-Tapia, B. Betancourt-Tovar, M. Videa, M. Antunes-Ricardo, and J. L. Cholula-Díaz, "Green synthesis trends and potential applications of bimetallic nanoparticles towards the sustainable development goals 2030," *Nanoscale Advances*, vol. 6, no. 1, pp. 51–71, 2024.

[13] D. E. V. Pires, T. L. Blundell, and D. B. Ascher, "pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures," *Journal of Medicinal Chemistry*, vol. 58, no. 9, pp. 4066–4072, 2015.

[14] S. M. Ivanov, A. A. Lagunin, A. V. Rudik, D. A. Filimonov, and V. V. Poroikov, "ADVERPred–Web service for prediction of adverse effects of drugs," *Journal of Chemical Information and Modeling*, vol. 58, no. 1, p. 8–11, 2017.

[15] A. Lagunin, A. Rudik, D. Druzhilovsky, D. Filimonov, and V. Poroikov, "ROSC-Pred: web-service for rodent organ-specific carcinogenicity prediction," *Bioinformatics*, vol. 34, no. 4, pp. 710–712, 2017.

[16] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: a webserver for the prediction of toxicity of chemicals," *Nucleic Acids Research*, vol. 46, no. W1, pp. 1–7, 2018.

[17] P. Banerjee, E. Kemmler, M. Dunkel, and R. Preissner, "ProTox 3.0: a webserver for the prediction of toxicity of chemicals," *Nucleic Acids Research*, vol. 52, no. W1, pp. 1–8, 2024.

[18] United Nations, *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*, 10th ed. New York: United Nations Publication, 2023, vol. 1.

[19] O. Shikimaka, M. Bivol, B. A. Sava, M. Dumitru, C. Tardei, B. G. Sbarcea, D. Grabco, C. Pyrtsac, D. Topal, A. Prisacaru, V. Cobzac, and V. Nacu, "Hydroxyapatite–bioglass nanocomposites: Structural, mechanical, and biological aspects," *Beilstein Journal of Nanotechnology*, vol. 13, no. 1, pp. 1490–1504, 2022.

[20] S. Vasanthavel, E. Manivannan, K. Murugan, H. Arumugam, and S. Kannan, "Spontaneous crystallization of hydroxyapatite in CeO_2-58S bioglass composite matrix," *Ceramics International*, vol. 51, no. 22, pp. 37 969–37 974, 2025.

[21] C. Noyan, E. Bakuer, M. B. Coban, M. E. Diken, and F. Unal, "Multifunctional europium-doped S53P4 bioactive glass nanocomposites offering magnetic, optical, and antibacterial features for next-generation biomedical applications," *Ceramics International*, vol. 51, no. 24, pp. 41 948–41 958, 2025.

[22] A. M. Deliormanlı, B. Rahman, S. Oguzlar, and M. Z. Ongun, "Investigation of the structural, photoluminescence properties, bioactivity and 5-fluorouracil delivery behavior of Dy^{3+} and Dy:Eu^{3+}-doped bioactive glasses," *Journal of Alloys and Compounds*, vol. 944, pp. 1–14, 2023.

[23] A. M. Simões, C. Venâncio, L. Alves, F. E. Antunes, and I. Lopes, "Hydrophobic modifications of hydroxyethyl cellulose polymers: Their influence on the acute toxicity to aquatic biota," *Journal of Hazardous Materials*, vol. 409, no. 1, pp. 1–12, 2021.

[24] L. Jiang, X. Ma, Q. Yan, D. Pu, X. Fu, and D. Zhang, "Dihydromyricetin/montmorillonite intercalation compounds ameliorates DSS-induced colitis: Role of intestinal epithelial barrier, NLRP3 inflammasome pathway and gut microbiota," *International Journal of Pharmaceutics*, vol. 670, no. 1, pp. 1–13, 2025.

[25] I. Atkinson, A. M. Seciu-Grama, S. Petrescu, D. Culita, O. C. Mocioiu, M. Voicescu, R.-A. Mitran, D. Lincu, A.-M. Prelipcean, and O. Craci-

unescu, “Cerium-containing mesoporous bioactive glasses (MBGs)-Derived scaffolds with drug delivery capability for potential tissue engineering applications,” *Pharmaceutics*, vol. 14, no. 6, pp. 1–16, 2022.

[26] J. Bahamonde-Duarte, S. Briceño, K. Vizuete, A. Debut, L. J. Borrero-González, and G. González, “Fluorescent chitosan-europium hydroxyapatite beads for drug delivery,” *Carbon Trends*, vol. 20, no. 1, pp. 1–6, 2025.

Simulation-based geometry optimization of a bone-on-a-chip through a machine learning-assisted evolutionary approach

Gabriel Nicolai
Center for Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
gnicolai@cti.gov.br

Katarina Vilarins
Ilum Escola de Ciência
Brazilian Center for Research in Energy and Materials (CNPEM)
Campinas, Brazil
katarina24012@ilum.cnpem.br

Caio Ruas
Ilum Escola de Ciência
Brazilian Center for Research in Energy and Materials (CNPEM)
Campinas, Brazil
caio24010@ilum.cnpem.br

Larissa M. S. Ribeiro
Center for Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
larissa.ribeiro@cti.gov.br

Filipe Loyola Lopes
Center of Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
flopes@cti.gov.br

Marcília Valéria Guimarães
Center for Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
mvguimaraes@cti.gov.br

Rodrigo Bonacin
Center for Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
rbonacin@cti.gov.br

Juliana K. M. B. Daguano
Center for Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
jdaguano@cti.gov.br

Matheus Bernardelli de Moraes
Center for Information Technology
Renato Archer (CTI Renato Archer)
Campinas, Brazil
mbmoraes@cti.gov.br

***Abstract*—**This study presents a simulation-based methodology to optimize the geometry and operating conditions of a bone-on-chip platform. Integrating Computational Fluid Dynamics (CFD), machine learning, and an evolutionary algorithm, the framework addresses the complex, opposing trade-offs inherent in microfluidic transport dynamics. CFD simulations were performed across varied geometric parameters, micropillar arrangements, volumetric flow rates, and residence times to generate hydrodynamic and mass transfer performance indicators. The resulting velocity, pressure, wall shear stress, and transport fields were converted into quantitative performance indicators representing medium perfusion within the micropillar region, hydraulic resistance, wall shear stress, and oxygen availability in the culture medium. A Random Forest regression model was trained on the CFD data to serve as a fast surrogate model, allowing a Differential Evolution algorithm to efficiently search the design space for high-performing configurations. The proposed framework reduces the number of additional CFD simulations required during optimization and provides a systematic strategy for designing organ-on-chip platforms with improved transport conditions and reduced hydraulic constraints. All findings are exclusively in silico: the optimized geometries were verified by CFD. Device fabrication, experimental perfusion, or cell-culture validation will be addressed in future studies.

Keywords*—Bone-on-chip; Computational fluid dynamics; Machine learning; Differential Evolution; Surrogate model; Geometry optimization.*

I. Introduction

Bone regeneration and repair of critical-sized bone defects remain major challenges in tissue engineering, driving the development of biomimetic scaffolds, computational tools, and more predictive in vitro models [1]. In this context, organ-on-chip systems have emerged as promising platforms for recapitulating the key aspects of human physiology. However, bone-on-chip models remain less developed than their soft-tissue counterparts, largely because of the structural and functional complexities of bone tissue [2], [3]. Nevertheless, by integrating three-dimensional cell culture, microfluidics, and the controlled application of biomechanical and biochemical cues, these platforms can more closely reproduce the bone microenvironment, expanding their potential for investigating bone physiology and diseases, as well as for drug screening.

The performance of these platforms is strongly influenced by their internal geometries and operating conditions. The micropillar dimensions, arrangement, spacing, and location relative to the inlet and chamber walls influence flow penetration, residence time, pressure distribution, wall shear stress, and solute transport. However, improving medium renewal and reducing stagnation may increase the hydraulic resistance or create heterogeneous shear exposure. Therefore, device design requires balancing competing hydrodynamic and mass transfer objectives that cannot be represented by a single response variable [4], [5].

Computational Fluid Dynamics (CFD) provides spatially resolved information on velocity, pressure, wall shear stress, and species concentration, allowing microfluidic configurations to be evaluated before fabrication. However, direct optimization based exclusively on CFD may require numerous computationally expensive simulations when multiple geometric and operating variables are considered. Machine learning surrogate models can reduce this cost by learning the relationship between the design variables and CFD-derived metrics and predicting untested configurations. Previous studies have demonstrated that combining CFD datasets with machine learning and evolutionary optimization can substantially accelerate the search for improved geometries while retaining nonlinear relationships between geometry, flow, and performance [6],[7].

When multiple responses conflict, they can be transformed and combined into a global performance score [8]. Random Forest (RF) is suitable for constructing this predictive model because an ensemble of randomized decision trees is suited to nonlinear regression and does not require a priori functional relationships [9], [10]. The resulting surrogate can be coupled with Differential Evolution (DE), a population-based algorithm that uses mutation, crossover, and selection to explore continuous

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

design spaces and locate high-performance solutions [11], [12].

This strategy of coupling simulation, surrogate modeling, and optimization has already been applied to organ-on-chip and microfluidic platforms through parametric sweeps combined with machine learning [6],[13] and to characterize predefined bone-on-chip devices under selected operating conditions [14], motivating the closed-loop simulation-surrogate-optimization framework adopted in this study.

Accordingly, the aim of this study was to design a basic microfluidic device for bone-on-a-chip applications and optimize its geometric and operating conditions using a machine learning-assisted evolutionary framework. First, 3D CFD simulations were conducted to evaluate the key fluid dynamics and mass transport parameters within the micropillar region, including medium perfusion, hydraulic resistance, wall shear stress, and oxygen availability. These features were aggregated into a mathematically modeled global performance score designed to ensure adequate conditions for cell culture. Finally, an RF surrogate model was used to predict the score, and DE was applied to identify the optimal design parameters that maximize the overall device performance.

II. Background and Related Work

A general box-constrained optimization problem can be formulated using the following Eq. (1), without loss of generality.

$$\begin{aligned} &min\, F(\vec{x}) = (f_1(\vec{x}), f_2(x),.., f_m(\vec{x})) \\ &s.t.\ x_k \in [x_j^{min}, x_j^{max}],\ \forall k \in \{1, 2,..., d\} \end{aligned} \quad (1)$$

where $\vec{x} = [x_1, x_2,.., x_d]$ is a d-dimensional decision vector, x_k is the k-th variable $\vec{x}$, m is the number of objective functions, $f_z(\vec{x})$ is the z-th objective function, and x_j^{min} x_j^{max} are the lower and upper boundaries of the variable x_j, respectively. If $m = 1$, then Eq. (1) is a Single-Objective Optimization Problem (SOP). In SOP, there is a single decision vector $\vec{x}_{opt}$ known as *Global Optimum* that results in the best objective function $f(\vec{x})_{opt}$. The goal of SOP is to find $\vec{x}_{opt}$.

Machine and deep learning AI methods have been demonstrated to improve the performance of optimization algorithms in time-consuming environments when used as surrogate models (SMs) [15]. An SM can be formally defined as: given a limited amount of data $D = \{(\vec{x}^{(1)}, F(\vec{x}^{(1)}), ..., (\vec{x}^{(\eta)}, F(\vec{x}^{(\eta)}))\}$, where η is the number of real expensive function evaluations, SM is any approximation technique that provides an estimated objective function vector $\hat{F}(\vec{x})$ of a given true and expensive objective vector $F(\vec{x})$. A recently applied AI method as SM in computationally expensive optimization problems is RF [9]. RF has demonstrated to be an effective SM approach [15].

III. Materials and Methods

The methodology of this research was structured into five main stages: **(A)** design of the microfluidic device; **(B)** 3D mesh generation and CFD simulation; **(C)** definition of the target variable score; **(D)** mathematical modeling of the optimization concern; and **(E)** development of machine learning-based surrogate models and their application in the surrogate-assisted optimization cycle, as detailed below.

A. Design of the microfluidic device

The investigated platform (Fig. 1) consisted of inlet and outlet microchannels (1.0 mm high) connected to a 5.0-mm-diameter central chamber containing a ring-shaped micropillar array.

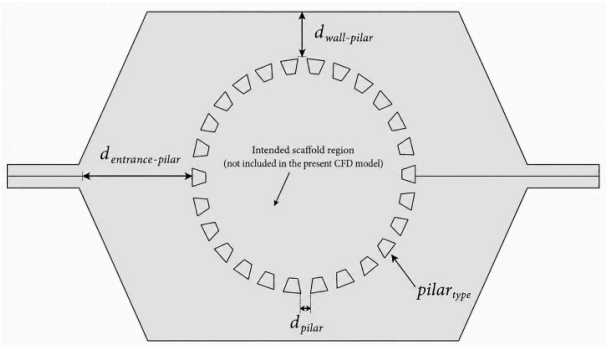


Fig. 1. Schematic representation of the investigated bone-on-a-chip geometry, showing the inlet and outlet microchannels, central chamber, and ring-shaped micropillar region used to define the geometric design space

The pillars enclose a reserved region intended for a bone-mimetic scaffold in future experiments, which is not explicitly modeled here; however, defining the scaffold's intended position and perturbing the flow field influence perfusion and mass transport around this location. The geometric design space was defined by four parameters: clear spacing between adjacent pillars (d_{pilar}), distance from the inlet channel to the pillar region ($d_{entrance}$), distance from the pillar region to the chamber wall ($d_{wall-pilar}$), and pillar arrangement type ($pilar_{type}$). The inlet volumetric flow rate (Q) was also varied to represent different perfusion conditions.

The initial dataset comprised 85 CFD simulations, which combined variations in the geometric parameters, pillar arrangements, and flow rates. After training the machine-learning model and applying the optimization procedure, four additional optimized geometries were simulated using CFD, added to the original dataset, and the complete set was re-evaluated using the same score function.

B. 3D mesh generation and computational fluid dynamics

The fluid was assumed to be incompressible and Newtonian, and the flow was considered laminar, consistent with the low Reynolds number regime expected in microfluidic systems. The hydrodynamic field was described by the continuity—Eq. (2) and the Navier–Stokes equation using Eq. (3), where U is the velocity vector ($m\ s^{-1}$), ρ is the fluid density ($kg\ m^{-3}$), p is the pressure (Pa), v is the kinematic viscosity ($m^2 s^{-1}$), and g is the gravitational acceleration ($m\ s^{-2}$).

$$\nabla . U = 0 \quad (2)$$

$$(U.\nabla)U = -\frac{1}{\rho}\nabla p + v\nabla^2 U + g \quad (3)$$

In addition to the flow field, the transport of dissolved O_2 was described by advective diffusion, as shown in Eq. (4).

$$U.\nabla C_{O_2} = D_{O_2}\nabla^2 C_{O_2} - R_{O_2} \quad (4)$$

where C_{O_2} is the local concentration O_2 ($mol\,m^{-3}$), D_{O_2} is the O_2 diffusion coefficient ($m^2 s^{-1}$), and R_{O_2} is the O_2 consumption term ($mol\,m^{-3}s^{-1}$).

The oxygen consumption term was represented using Michaelis–Menten kinetics, as described in Eq. (5). The kinetic parameters were adopted from the microfluidic bone marrow model adapted by Rosalem et al. [16], which combined CFD and organ-on-chip approaches to investigate leukemia niche dynamics.

$$\begin{gathered} R_{max,O_2} = 1.2\times10^{-5}\,mol\,m^{-3}\,s^{-1}, \\ C_{MM,O_2} = 1.0\times10^{-3}\,mol\,m^{-3}, \text{ and} \\ C_{cr,O_2} = 1.0\times10^{-4}\,mol\,m^{-3}. \\ R_{O_2} = R_{max,O_2}\frac{C_{O_2}}{C_{O_2}+C_{MM,O_2}}\,\delta\,(C_{O_2} > C_{cr,O_2}) \end{gathered} \quad (5)$$

where $R_{max,O2}$ is the maximum oxygen consumption rate, C_{MM,O_2} is the Michaelis–Menten constant, C_{cr,O_2} is the critical oxygen concentration, and δ is the reduction function that suppresses consumption below C_{cr,O_2}. This formulation enables the analysis of oxygen transport under physiologically relevant conditions.

At the inlet, a prescribed volumetric flow rate and fixed oxygen concentration were imposed, and a zero-gauge pressure condition was applied at the outlet. No-slip conditions were adopted for all solid surfaces, including the chamber walls and micropillars. The computational domain was discretized using tetrahedral elements, with local refinement near the micropillars to better resolve the velocity and oxygen concentration gradients in this area. Simulations were performed using ANSYS CFD v18.2. Convergence was defined by root-mean-square residuals below 1×10^{-5}, requiring between 50 and 500 iterations, depending on the case.

To assess the sensitivity of the numerical solution to spatial discretization, a mesh independence study was performed using three tetrahedral meshes containing approximately 2.6×10^6, 3.7×10^6, and 5.0×10^6 elements. The comparison was based on the velocity and wall shear stress profiles along the channel width, since these variables are directly related to the hydrodynamic indicators used in the score calculation. The Grid Convergence Index (GCI) was calculated to quantify the discretization uncertainty across the refined meshes.

C. Definition of the target variable score

The CFD simulations generated multiple hydrodynamic and mass transfer responses, including wall shear stress, velocity heterogeneity, hydraulic resistance, and oxygen concentration. Since these responses represent competing design requirements, they were combined into a single target variable, which was defined as the global score. The score was designed to quantify the overall performance of each simulated condition and enable the training of a single-output surrogate model. Four CFD-derived indicators were used in the score calculation: the 90th percentile wall shear stress, $WSS90(\vec{x})$; the average coefficient of variation of velocity, $CV_u(\vec{x})$; the hydraulic resistance, $R_h(\vec{x})$; and the mean oxygen fraction, denoted by $\phi_{O_2}(\vec{x})$. The use of $WSS90(\vec{x})$ reduced the influence of isolated local peaks, while $CV_u(\vec{x})$ represented flow heterogeneity within the chamber. $R_h(\vec{x})$ was used to account for the hydraulic cost of each configuration, and $\phi_{O_2}(\vec{x})$ represented oxygen availability within the device.

D. Mathematical modeling of the optimization problem

The mathematical formulation of the optimization problem is defined in Eq. (6):

$$\begin{gathered} min\,F(\vec{x}) = (WSS90(\vec{x}),\ R_h(\vec{x}),\ CV_u(\vec{x}), -\phi_{O_2}(\vec{x})) \quad (6) \\ s.t.\ \vec{x} = [Q, d_{pilar}, d_{entrance}, d_{wall-pilar}, pilar_{type}] \\ Q \in \{1, 5, 10, 25, 100\},\ d_{pilar} \in [0.25, 0.75], \\ d_{entrance} \in [1.3,\ 2.3],\ d_{wall-pilar} \in [0.5, 2.0], \\ pilar_{type} \in \{1\ ,2\} \end{gathered}$$

where $WSS90(\vec{x})$ is the 90th percentile wall shear stress, $R_h(\vec{x})$ is hydraulic resistance defined as $R_h(\vec{x}) = \frac{\Delta P(\vec{x})}{Q(\vec{x})}$, $CV_u(\vec{x})$ is the velocity variation coefficient, and $\phi_{O_2}(\vec{x})$ is the dimensionless mean oxygen fraction. The negative sign in $-\phi_{O_2}(\vec{x})$ is used to transform the problem into minimization to facilitate the optimization process, without loss of generality. The variable $pilar_{type}$ was numerically encoded to allow its use in the surrogate model and optimization algorithm, with $pilar_{type} = 1$ representing the trapezoidal micropillar arrangement and $pilar_{type} = 2$ representing the triangular micropillar arrangement. Although numerically encoded, this variable represents a categorical geometric feature.

We transformed Eq. (6) into an SOP by developing a *score* that calculates the effectiveness of each geometry. This score (i) transforms the problem into an SOP, (ii) penalizes solutions that provide infeasible geometries, and (iii) penalizes solutions that are feasible but poor.

The objective functions were first transformed into nondimensional performance factors ranging from 0 to 1, according to their contribution to the desired microfluidic behavior. These factors were combined into a global score ranging from 0 to 10. Each objective function was modified using a specific transformation. For instance, the 90th percentile wall shear stress objective function ($WSS90(\vec{x})$) is transformed into a numerical value based on the following design:

- $WSS90(\vec{x}) = 1,\ if\ WSS90(\vec{x}) \leq 1.0\times10^{-3}$
- $WSS90(\vec{x}) = e^{penal}$

where $penal = -1(\frac{WSS90(\vec{x})-1.0\times10^{-3}}{1.0\times10^{-3}})$. Thus, if the wall shear stress is below the threshold of 1.0×10^{-3}, then the score of this objective 1.0; otherwise, it is penalized based on its original objective function value.

The $WSS90$ threshold was adopted as a conservative low-shear design target rather than as a universal cellular damage limit. Bone cells are mechanosensitive, and experimental microfluidic systems have been used to expose osteoblasts to shear stresses spanning from approximately $1.0\times10^{-3}\,Pa$ to considerably higher values [17]. In the present score, values above the adopted threshold were therefore not classified as infeasible; instead, their desirability was progressively reduced. Moreover, the use of $WSS90$, rather than the absolute maximum, limited the influence of isolated numerical or highly localized peaks while retaining sensitivity to the shear stress experienced over most of the evaluated surface.

The velocity variation coefficient $\overline{CV}_u(\vec{x})$ is penalized using the following formulation: $\overline{CV}_u(\vec{x}) = e^{(-2.CV_u^2(\vec{x}))}$. Thus, lower values are preferred for these indicators. For the hydraulic resistance R_h, a different logic had to be applied. The penalization depended on the range of the oxygen inside the chip, and it was calculated from the normalized value of R_h. Hydraulic resistance was included as an operational criterion because excessive resistance increases the pressure required to maintain a prescribed perfusion rate. Unlike oxygen availability and wall shear stress, R_h was not assigned a direct biological threshold. Its normalized value was used to provide a relative comparison among the simulated configurations, with lower hydraulic resistance receiving higher desirability. The developed logic is organized as follows.

- $R_h(\vec{x}) = 0.85e^{penal}$, $if\ 0.08 \le \phi_{O_2}(\vec{x}) \le 0.1$
- $R_h(\vec{x}) = 1.0e^{penal}$, $if\ 0.06 \le \phi_{O_2}(\vec{x}) \le 0.08$
- $R_h(\vec{x}) = 0.4e^{penal}$, $if\ 0.03 \le \phi_{O_2}(\vec{x}) \le 0.06$
- $R_h(\vec{x}) = 0.1e^{penal}$, $if\ \phi_{O_2}(\vec{x}) < 0.03$
- $R_h(\vec{x}) = 0.4e^{penal}$, $if\ 0.1 \le \phi_{O_2}(\vec{x})$

where $penal = -R_{h,norm}$. Thus, the best values are obtained if the oxygen is within the desired range (between 0.03 and 0.1). Finally, the oxygen concentration penalization is expressed as follows:

- $\phi_{O_2}(\vec{x}) = 1$, $if\ 0.03 \le \phi_{O_2}(\vec{x}) \le 0.10$. No penalization is applied because the oxygen level is in the desired interval.
- $\phi_{O_2}(\vec{x}) = e^{penal}$, $if\ \phi_{O_2}(\vec{x}) < 0.01$. For this case, the penalization is $penal = -5(\frac{0.01-\phi_{O_2}(\vec{x})}{0.01})^2$
- $\phi_{O_2}(\vec{x}) = e^{penal}$, $if\ 0.01 \le \phi_{O_2}(\vec{x}) < 0.03$. For this case, the penalization is $penal = -3(\frac{0.03-\phi_{O_2}(\vec{x})}{0.03})^2$.
- $\phi_{O_2}(\vec{x}) = e^{penal}$, $if\ \phi_{O_2}(\vec{x}) > 0.1$. Then, the penalization is $penal = -1(\frac{\phi_{O_2}(\vec{x})-0.10}{0.10})^2$.

The oxygen thresholds were selected to represent a low-oxygen bone-marrow-like environment while avoiding severe oxygen depletion. Direct in vivo measurements have demonstrated the heterogeneity of bone marrow oxygenation, with oxygen tensions of approximately 2.4–5.3% in bone marrow monocytes, approximately 4.8% in mature osteoclasts, and lower values in some extravascular regions [18]. Based on this physiological context, oxygen levels below 3% were progressively penalized because they indicate marked oxygen limitation, whereas values between 3 and 10% were retained as an operationally acceptable range for the investigated platform. The 6–8% interval was assigned the highest desirability as an author-defined design preference intended to balance oxygen availability and the preservation of a relatively low-oxygen microenvironment Finally, because each objective function was modified, Eq. (6) is transformed into a single-objective problem using a weighted sum approach, as shown in Eq. (7):

$$max\ score(\vec{x}) = 10.[0.30\cdot WSS90(\vec{x}) + 0.10\cdot R_h(\vec{x}) + 0.25\cdot CV_u(\vec{x}) + 0.35\cdot \phi_{O_2}(\vec{x})] \quad (7)$$

s.t. the same constraints as in Eq. (6).

The baseline weight vector was defined a $\vec{w} = [0.30,\ 0.10,\ 0.25,\ 0.35]$ corresponding to the $WSS90$, R_h, $\overline{CV}_u$, and ϕ_{O_2}, respectively. These values represent a multicriteria choice adopted by the authors for the present study rather than universal biological constants. Oxygen availability and wall shear stress received the highest weights because they are directly related to the cellular microenvironment, whereas hydraulic resistance received a lower weight because it primarily represents an operational constraint. The sensitivity of the ranking to these author-defined choices was subsequently evaluated.

E. Sensitivity analysis of the composite score

A sensitivity analysis was performed to evaluate whether the configuration ranking was strongly dependent on the author-defined parameters of the composite score. The analysis was conducted using the CFD-derived responses of the complete dataset, comprising the 85 initial cases and the 20 additional cases obtained from four optimized geometries evaluated at five flow rates.

Alternative scenarios to the baseline were generated by independently increasing or decreasing each weight by 20%. After each individual variation, all weights were renormalized so that their sum remained equal to one. An equal-weight scenario, in which all four factors received a weight of 0.25, was also evaluated.

In addition to the weight analysis, the $WSS90$ threshold was varied by ±20% relative to its baseline value of $1.0\times10^{-3}\,Pa$. The sensitivity to the exponential penalty intensity was evaluated using multiplicative factors of 0.75, 1.00, and 1.25, representing mild, baseline, and strong penalization, respectively. Only one score parameter was modified at a time, while all other parameters were maintained at their baseline values. For the final post-CFD

comparison, fixed min–max bounds for CV_u and R_h were derived from the complete 105-case dataset and applied uniformly to all initial and optimized cases. The same bounds were maintained across all score-sensitivity scenarios.

For each scenario, the composite scores of all 105 CFD cases were recalculated, and the resulting ranking was compared with the baseline ranking. Ranking stability was quantified using Spearman's rank correlation coefficient (ρ), the percentage overlap between the ten highest-ranked cases, and the identity of the highest-ranked configuration. This analysis was intended to evaluate the robustness of the ranking and did not involve retraining the Random Forest model, rerunning Differential Evolution, or modifying the original surrogate-assisted optimization procedure.

F. Surrogate-assisted geometry optimization

Fig. 2 demonstrates the developed methodology of the surrogate-assisted geometry optimization method using DE and RF.

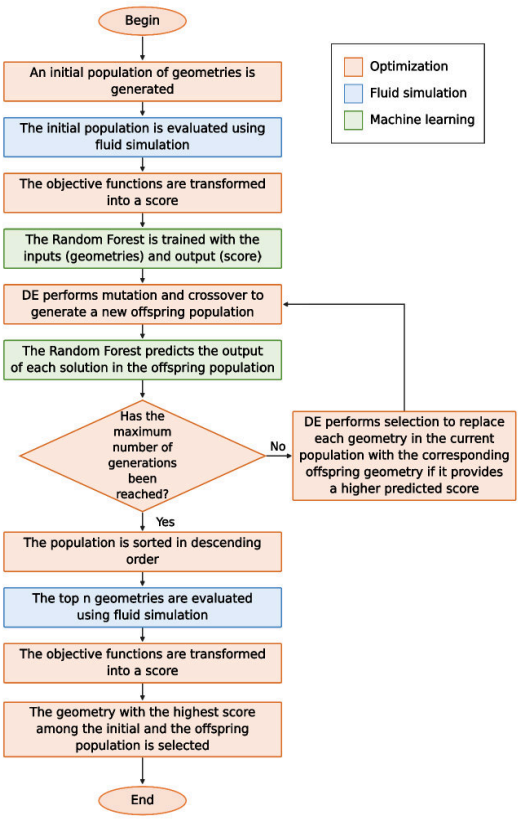


Fig. 2. Flowchart of the proposed methodology.

Before coupling the surrogate model with DE, its predictive performance was evaluated using repeated five-fold group cross-validation with 20 repetitions, resulting in 100 test folds. Groups were defined by the unique combination of pillar spacing, wall-to-pillar distance, inlet-to-pillar distance, and pillar type. Thus, all flow-rate conditions associated with the same geometry were assigned entirely to either the training or test subset in each fold, preventing information leakage between geometrically equivalent cases. The Min-Max transformation was fitted exclusively to the training subset within each fold. The RF comprised 500 trees, with a minimum of two samples per terminal leaf. Variability across the 100 grouped test folds was reported as the mean and standard deviation of R^2, MAE, and RMSE.

After optimization, the predictive performance of the surrogate was further assessed using the 20 additional CFD cases, comprising four optimized geometries evaluated at five flow rates. RF predictions were compared with CFD-derived scores calculated using the same score formulation and normalization limits adopted for the initial 85-case dataset. Agreement was quantified using (R^2), MAE, RMSE, and mean bias, defined as the RF-predicted score minus the CFD-derived score. Error dispersion was summarized using its standard deviation, empirical 2.5th–97.5th percentile range, and maximum absolute error. These descriptive statistics were used to characterize case-specific predictive variability and were not interpreted as formal prediction intervals.

Next, DE performs mutation and crossover to generate an offspring population (as discussed in Section II-A). The RF model predicts the scores for each offspring geometry. If the maximum number of generations has not yet been reached, DE performs a selection by replacing each geometry in the current population with the corresponding offspring generated using the trial vector if it provides a higher predicted score. In contrast, if the maximum number of generations is achieved, the offspring are sorted in ascending order based on the predicted scores. The top n (user-input parameter) geometries were evaluated using fluid simulations.

The objective functions of the best n geometries were transformed into a global score. Finally, the highest-ranked candidate geometries were selected for CFD verification. It's worth to highlight that the complete DE algorithm settings, including the population size, mutation and crossover parameters, number of generations, evolution strategy, and stopping criteria, are provided in the associated code repository[1].

IV. Results and Discussion

A. Transport behavior

First, a mesh-independent study was conducted. The velocity profiles obtained using the three meshes (Fig. 3) showed close agreement along the channel width, indicating that the main flow features were adequately captured. The wall shear stress profiles also exhibited similar behavior among the meshes, including regions of higher shear stress. Since wall shear stress is more sensitive to near-wall velocity gradients, this agreement indicates that the mesh refinement was sufficient for extracting shear-related quantities such as $WSS90$ and $WSS90_{pilar}$. Therefore, the selected mesh provided a suitable compromise between numerical accuracy and computational cost for the CFD simulations.

Based on the profile comparison and GCI value for the fine mesh (0.001%), the mesh with approximately 3.7×10^6 elements were selected for the parametric simulations because they provided mesh-independent results while reducing the computational cost compared with that of the finest mesh size.

The initial CFD simulations provided the hydrodynamic and mass transfer datasets used for the score calculation and surrogate model training. For each simulated condition, the

[1] https://github.com/mbdemoraes/2026-cbeb-open-data

velocity field, pressure field, wall shear stress distribution, and oxygen concentration field were obtained and converted into quantitative-performance indicators. These indicators included the 90th percentile wall shear stress, $WSS90$, the wall shear stress in the micropillar region, $WSS90_{pilar}$, the maximum wall shear stress, WSS_{max}, the average velocity coefficient of variation, CV_u, the hydraulic resistance, R_h, the average oxygen fraction, ϕ_{O_2}, the minimum oxygen fraction, $\phi_{O_{2_{min}}}$, and the oxygen coefficient of variation, $CV_{\phi_{O_2}}$. The baseline simulations showed that the microchip geometry and inlet flow rate strongly affected the transport behavior in the central chamber.

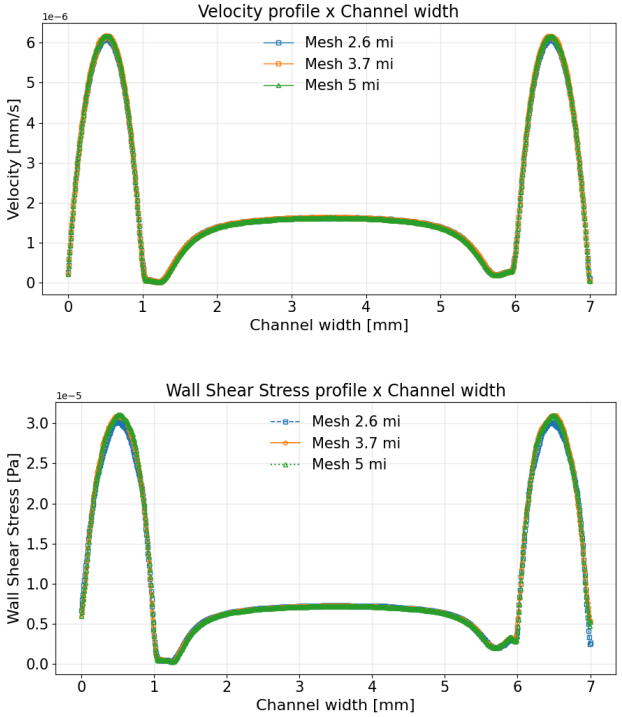


Fig. 3. Assessment of the mesh along the channel width. (a) Velocity profiles and (b) wall shear stress profiles obtained for meshes with approximately 2.6×10^6, 3.7×10^6, and 5.0×10^6 elements. The close agreement among the profiles indicates the low sensitivity of the hydrodynamic solution to further mesh refinement.

Fig. 4 shows the representative CFD results obtained for geometry type D. The velocity contours were compared for micropillar spacings of 0.25 and 0.50 mm at a volumetric flow rate of 1 µL min⁻¹. For both configurations, the highest velocities occurred in the inlet and outlet microchannels, whereas the central chamber exhibited lower velocities owing to flow expansion. The 0.25 mm spacing produced a more restricted flow path around the micropillar region, increasing the local obstruction and low-velocity zones. In contrast, a 0.50 mm spacing allowed greater flow penetration through the chamber, improving the fluid distribution around the micropillar array.

The oxygen concentration contours also demonstrated the effect of the inlet flow rate on mass transport. At a lower flow rate, oxygen gradients were more evident inside the chamber, indicating that oxygen consumption and limited convective renewal can produce spatial heterogeneity. At higher flow rates, the oxygen concentration became more uniform throughout the chamber, suggesting improved medium renewal and reduced oxygen depletion. Therefore, the CFD results confirmed that both the geometry and operating conditions affect the preferential flow pathways, velocity heterogeneity, wall shear stress distribution, and oxygen availability.

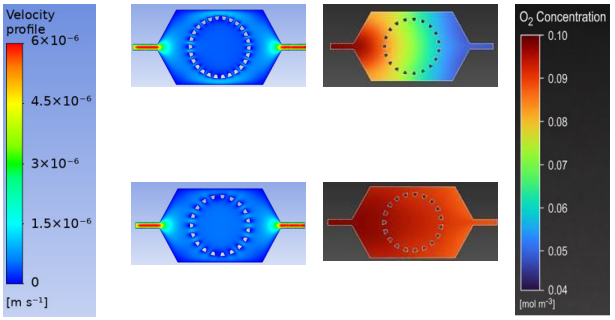


Fig. 4. Representative CFD results for geometry type D. Velocity contours for micropillar spacings of 0.25 mm (top left) and 0.50 mm (bottom left) at a volumetric flow rate of 1 $\mu L\ min^{-1}$, and O_2 concentration contours for two inlet flow rates, $Q_{low} = 1\ \mu L\ min^{-1}$ (top right), and $Q_{high} = 5\ \mu L\ min^{-1}$ (bottom right), at fixed geometry and micropillar spacing.

These findings indicate that no single CFD-derived response is sufficient to rank the microchip configurations. Higher flow rates favored oxygen renewal and reduced oxygen heterogeneity but also increased the wall shear stress. Conversely, lower flow rates reduced shear stress but increased residence time and intensified oxygen depletion in regions with lower convective transport. Similarly, geometrical configurations that promote greater interaction between the flow and micropillar region can improve perfusion but also alter local shear stress. This trade-off supports the use of a global score to combine shear stress, velocity heterogeneity, hydraulic resistance, and oxygen availability into a single target variable. Before applying the surrogate-assisted optimization workflow, the highest-scoring configurations in the initial dataset were associated with geometries that provided an adequate compromise between oxygen availability and low wall shear stress (WSS). However, some of these cases still presented relatively high velocity heterogeneity, indicating that further enhancements could be obtained by exploring intermediate geometric configurations that were not included in the original parametric set. Therefore, these CFD-derived responses were used to train the RF surrogate model and guide the subsequent DE search.

B. Surrogate-model validation and candidate selection

The RF surrogate model was first evaluated using repeated group cross-validation, ensuring that complete geometries were held out from model training. Across the 100 grouped test folds, the model achieved $R^2 = 0.814 \pm 0.132$, $MAE = 0.386 \pm 0.103$, and $RMSE = 0.520 \pm 0.154$ score units. When the repeated out-of-fold predictions were averaged for each of the 85 initial conditions, the resulting performance was $R^2 = 0.848$, $MAE = 0.375$, and $RMSE = 0.526$. These results indicate that the surrogate captured the main relationship between the design variables and the global score for previously unseen geometries. Nevertheless, the

variability across grouped partitions demonstrates that prediction accuracy remained sensitive to the limited number of independent geometries, reinforcing the need for CFD verification of the candidates proposed by the surrogate. After this validation, the RF model was coupled with DE to explore the design space and identify candidate conditions with high predicted performance.

The optimized candidate pool included both pillar-type configurations. Although the surrogate model initially assigned slightly higher robust predicted scores to trapezoidal micropillars, triangular micropillar candidates were maintained in the final pool to preserve geometric diversity and avoid selecting only one design family. This decision was important because the subsequent CFD verification showed that the triangular configurations achieved the highest recalculated scores.

C. CFD evaluation of optimized geometries

From the candidate pool generated by DE, four optimized geometries were selected and simulated using CFD. These new simulations were added to the original dataset, corresponding to zero-based indices 85–104 of the complete dataset. After inclusion, all cases were re-evaluated using the same global score function. This recalculation was necessary because the score includes min–max normalization terms for CV_u and R_h; therefore, adding new geometries can change the reference range used to compare all cases.

Using a common normalization range for the complete dataset, the recalculated ranking placed the optimized geometries in the highest positions. The best configuration was obtained for the optimized geometry with $d_{pilar} = 0.625$ mm, $d_{wall-pilar} = 0.642$ mm, $d_{entrance-pilar} = 2.31$ mm, and $pilar_{type} = 2$. This geometry achieved a score of 9.905 at Q = 1, 5, and 10 µL min^{-1}, with residence times of 52.22, 10.44, and 5.22 min, respectively. This value represents the best multicriteria compromise within the investigated dataset and design domain, rather than a universally optimal geometry, because the score depends on the adopted weights, thresholds, penalty functions, and normalization limits.

Compared with the best initial case at Q = 10 µL min^{-1}, the optimized configuration reduced CV_u from 0.791 to 0.681, pillar $WSS90$ from 6.62×10^{-4} to 4.68×10^{-4} Pa, and CV_{O_2} from 0.0142 to 0.0111. The ϕ_{O_2} increased slightly from 0.0948 to 0.0950, whereas overall $WSS90$ and hydraulic resistance remained nearly unchanged. The shift from pillar type 1 to type 2 coincided with improved transport uniformity, although simultaneous changes in pillar spacing and wall clearance preclude attributing the improvement to pillar shape alone.

However, WSS_{max} increased from $1.62\times10^{-3}\ Pa$ in the best initial case to $1.05\times10^{-2}\ Pa$ in the optimized case, corresponding to a 6.5-fold increase. Because the score was based on $WSS90$, this localized maximum had limited influence on the optimization. Nevertheless, the higher peak may be relevant to local cell exposure and should be considered as an additional constraint in future score formulations. Its biological effect cannot be established without experimental cell-culture validation.

D. Sensitivity of the configuration ranking

The sensitivity analysis indicated that the configuration ranking was robust to the tested assumptions (Fig. 5). Variations in the oxygen-factor weight did not change the ranking because the oxygen desirability remained equal to one for all 105 cases, which were within the adopted acceptable range of 3–10%. The same highest-ranked case (case 95) was maintained in all scenarios, with Spearman's rank correlations ranging from 0.972 to 1.000. Individual variations in the weights, the $WSS90$ threshold, and penalty intensity preserved all baseline top-10 configurations. Equal weighting produced the largest change, although it retained 80% of the top 10 and the same highest-ranked case. Across all scenarios, its score ranged from 9.762 to 9.928, compared with the baseline value of 9.905. Thus, the selected configuration was stable, whereas its numerical score remained conditional on the adopted multicriteria formulation. Detailed scenario-specific results and the analysis code are available in the associated repository[1].

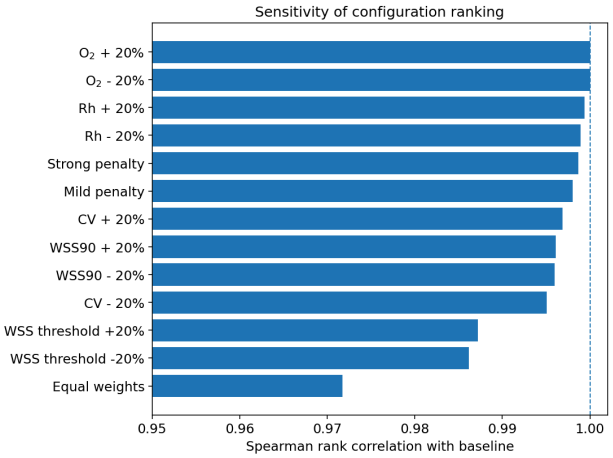


Fig. 5. Sensitivity of the configuration ranking to variations in score weights, the $WSS90$ threshold, and penalty intensity. Bars represent Spearman's rank correlation (ρ) with the baseline ranking; the dashed line indicates perfect correlation (ρ = 1).

E. Initial versus optimized score

The score distributions were compared between the initial dataset and the optimized CFD simulations. As shown in Fig. 6, the optimized group shifted the score distribution toward higher values, indicating that the RF-DE search could identify regions of the design space with improved global performance. This result is particularly relevant because the optimized geometries were not selected by direct CFD screening but by an RF model coupled with DE.

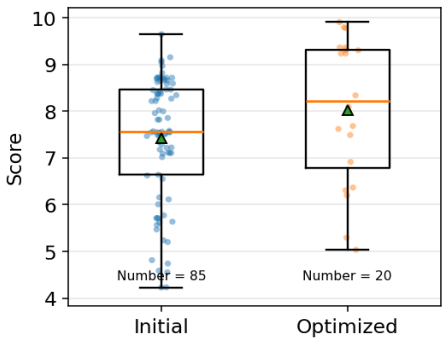


Fig. 6. Score distributions for the initial and optimized groups after recalculation using fixed min–max bounds derived from the complete 105-case dataset. The optimized group showed a shift toward higher scores following RF-DE-based candidate generation and CFD verification

After recalculation using fixed min–max bounds derived from the complete 105-case dataset, optimized cases accounted for nine of the ten highest-ranked conditions, whereas only one initial case remained. This result indicates that the RF-DE search identified high-performance regions of the design space that remained favorable after CFD verification. Besides, all final comparisons were performed using the same fixed normalization reference derived from the complete dataset.

For the 20 additional cases, RF-predicted and CFD-derived scores showed $R^2 = 0.857$, $MAE = 0.393$, $RMSE = 0.572$, and a mean bias of 0.075 score units (Fig. 7). Prediction errors had a standard deviation of 0.581; 90% of the absolute errors were below 0.907, and the maximum was 1.703. The largest deviations occurred for pillar type 2, although agreement was greater among the highest-scoring cases. The final CFD ranking favored triangular pillars, whereas the surrogate initially assigned slightly higher scores to trapezoidal configurations. Preserving both design families was therefore important, supporting the use of RF for candidate screening combined with targeted CFD verification. Because no device fabrication or biological experiments were performed, the reported improvements refer exclusively to CFD-derived hydrodynamic and mass-transfer performance and do not demonstrate improved cellular outcomes.

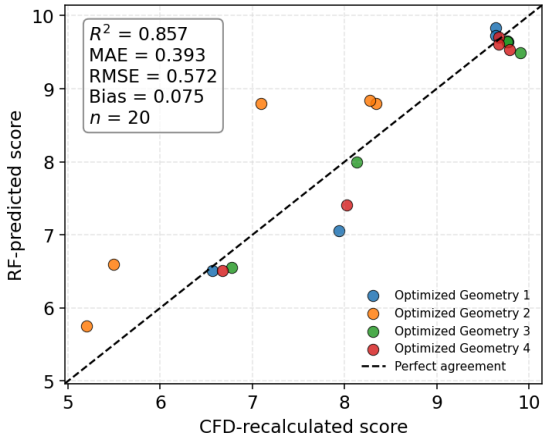


Fig. 7. Comparison between RF-predicted and CFD-derived scores for the 20 additional cases, comprising four optimized geometries evaluated at five flow rates. The dashed line represents perfect agreement.

V. Conclusions

In this study, we developed a simulation-based optimization workflow for a bone-on-chip platform by combining CFD simulations, RF surrogate modeling, and DE. The RF-DE framework identified high-performance design regions and selected four candidate geometries for CFD verification. After score recalculation using a common normalization reference, the optimized cases accounted for nine of the ten highest-ranked conditions. The highest-ranked geometry achieved a score of 9.905, representing a computational multicriteria compromise rather than an absolute optimum, and showed improved velocity and oxygen uniformity and reduced pillar $WWS90$ compared with the best initial configuration. The proposed framework supported multicriteria microfluidic design while reducing the need for exhaustive CFD screening. CFD verification remained essential because the highest-performing triangular-pillar configurations differed from the initial surrogate predictions, which slightly favored trapezoidal pillars. All findings are exclusively in silico, as no device fabrication, experimental perfusion, or cell-culture validation was performed. Future studies should expand the CFD dataset, refine the score formulation, and experimentally validate the optimized designs.

Acknowledgment

The authors acknowledge the São Paulo Research Foundation (FAPESP), Brazil, for the financial support provided through the Generation Project (grant no. 2024/19786-9) and grant no. 2026/04633-8, which contributed to this study.

References

[1] A. Shetty, A. Fathima, B. Anika,; R. Shetty, Vinyas, J. P. Supriya, A. Hegde, "Computational optimization of 3D printed bone scaffolds using orthogonal array-driven FEA and neural network modeling," Scientific Reports, vol. 15, August 2025, doi: 10.1038/s41598-025-15122-5.

[2] F. Sanaei, Y. Hajee, G. L. Koons, D. T. Wu, S. C. G. Leeuwenburgh, S. C. G.; J. J. J. P. Van Den Beucken, M. Diba, " Harnessing advances in bone tissue engineering for design of Bone-on-Chip systems", Advanced Healthcare Materials, v. 15, n. 4, November 2026, doi: 10.1002/adhm.202503525

[3] M. K. Kim, K. Paek, S. M. Woo, J. A. Kim, "Bone-on-a-Chip: Biomimetic models based on microfluidic technologies for biomedical applications" ACS Biomaterials Science & Engineering, v. 9, n. 6, pp. 3058-3073, May 2023, doi: 10.1021/acsbiomaterials.3c00066.

[4] F. Pisapia, W. Balachandran, M. Rasekh, "Organ-on-a-chip: design and simulation of various microfluidic channel geometries for the influence of fluid dynamic parameters", Applied Sciences, v. 12, n. 8, April 2022, doi: 10.3390/app12083829.

[5] A. Liao, J. Xiong, Z. Tong, L. Zhou, J. Liu, "Microfluidic chamber design for organ-on-a-chip: a computational fluid dynamics study of pillar geometry and pulsatile perfusion", Biosensors, v. 16, n. 1,, January 2026, doi:10.3390/bios16010049.

[6] D. O. Maionche, L. Ainstein, D. P. Dos Santos, M. B. Souza Jr, "Computational fluid dynamics and machine learning as tools for optimization of micromixers geometry", International Journal of Heat and Mass Transfer, v. 194, October 2022, doi: 10.1016/j.engappai.2024.108945.

[7] Q. T. Le,P. H. Chiu, C.C. Ooi, "U-Net-based surrogate model for evaluation of microfluidic channels", International Journal of Computational Methods, v. 19, n. 7, 2022, doi: 10.48550/arXiv.2105.05173

[8] G. Derringer, R. Suich, "Simultaneous optimization of several response variables", Journal of Quality Technology, v. 12, n. 4, pp. 214–219, August 1980.

[9] L. Breiman, "Random forests", Machine Learning, v. 45, pp. 5–32, October 2001.

[10] K. Faceli, A. C. Lorena, J. Gama, T. A. Almeida e A. C. P. L. F. de Carvalho, "Artificial Intelligence: A Machine Learning Approach", 3rd ed., vol. 1, Rio de Janeiro, LTC, 2025.

[11] R. Storn, L. Price, "Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces", Journal of Global Optimization, v. 11,p p. 341–359, November 1997.

[12] M. Chen, C. Feng, and R. Cheng, "MetaDE: Evolving Differential Evolution by Differential Evolution," *IEEE Transactions on Evolutionary Computation*, vol. 30, no. 1, pp. 141–155, Feb. 2026, doi: 10.1109/TEVC.2025.3541587.

[13] S Kheiri, E. Kumacheva, E. W. K. Young, "Computational modelling and big data analysis of flow and drug transport in microfluidic systems: A spheroid-on-a-chip study", Frontiers in Bioengineering and Biotechnology, v. 9, pp. 1-14, November 2021. doi: 10.3389/fbioe.2021.781566.

[14] H. Bahmaee, R. Owen, L. Boyle, "Design and evaluation of an osteogenesis-on-a-chip microfluidic device incorporating 3D cell culture", Frontiers in Bioengineering and Biotechnology, v. 8, pp. 1-17, September 2020. doi: 10.3389/fbioe.2020.557111.

[15] M. B. De Moraes and G. P. Coelho, "A Random Forest-Assisted Decomposition-Based Evolutionary Algorithm for Multi-Objective Combinatorial Optimization Problems," *2022 IEEE Congress on Evolutionary Computation (CEC)*, Padua, Italy, 2022, pp. 1-8, doi: 10.1109/CEC55065.2022.9870412

[16] G. S. Rosalem, D. R. R. De Assis, L. A. G. Torres, E. B. de Las Casas, W. N. Rodrigues, R. S. Gonçalves, R. A. N. Rabelo, J. C. Ruiz, and M. G. R. Carvalho, "A microfluidic bone marrow model combining CFD and organ-on-a-chip technologies to study leukemia niche dynamics," Frontiers in Bioengineering and Biotechnology, vol. 14, article 1766296, 2026, doi: 10.3389/fbioe.2026.1766296.

[17] YU, W.; QU, H.; HU, G.; et al. A microfluidic-based multi-shear device for investigating the effects of low fluid-induced stresses on osteoblasts. PLoS ONE, v. 9, n. 2, e89966, 2014. DOI: 10.1371/journal.pone.0089966.

[18] NARAZAKI, A.; SHIMIZU, R.; YOSHIHARA, T.; et al. Determination of the physiological range of oxygen tension in bone marrow monocytes using two-photon phosphorescence lifetime imaging microscopy. Scientific Reports, v. 12, art. 3497, 2022. DOI: 10.1038/s41598-022-07521-9.

Simulação Computacional da Dinâmica de Micro e Nanopartículas de Óxido de Cálcio em Água e Humor Vítreo sob Força de Radiação Acústica

Antonio Atila M. Ferreira, John Hebert da S. Felix, Thiago Q. da Silva,
Francisco Leonardo Alves de M. Sousa e Maria Cristiane M. Souza
Instituto de Engenharias e Desenvolvimento Sustentável
Universidade da Integração Internacional da Lusofonia Afro-Brasileira (UNILAB)
Redenção, Ceará, Brasil
atila@aluno.unilab.edu.br; johnfelix@unilab.edu.br; thiagoqueiroz@aluno.unilab.edu.br
leonardofisica@aluno.unilab.edu.br; mariacristiane@unilab.edu.br

***Resumo*—O uso de forças acústicas para manipulação de micro e nanopartículas tem sido investigado como estratégia de controle de transporte em meios líquidos. Entretanto, a magnitude da resposta depende fortemente do tamanho da partícula, das propriedades do meio e da validade das aproximações empregadas no modelo físico. Este trabalho investiga, por simulação computacional, a dinâmica de partículas esféricas de óxido de cálcio (CaO) submetidas a uma onda ultrassônica progressiva em água e em uma representação simplificada do humor vítreo.Foram avaliados oito cenários, combinando partículas com raios de 100 nm e 100 μm, dois meios fluidos e distâncias fonte–partícula de 16 e 21 mm. A frequência de 1 MHz e a amplitude de pressão de referência de 100 kPa foram mantidas fixas em todas as simulações, sem varredura de frequência ou intensidade. O modelo computacional combinou difusão térmica pela relação de Einstein–Smoluchowski, força de radiação acústica para esfera pequena em onda progressiva e arrasto hidrodinâmico avaliado a partir do número de Reynolds. Para as nanopartículas, o deslocamento determinístico associado à força acústica após 1.999 ms foi da ordem de 10^{-16}–10^{-15} μm, enquanto as trajetórias apresentaram flutuações muito superiores decorrentes do movimento browniano. Para as micropartículas, o deslocamento final no eixo x variou aproximadamente de -0.149 a -0.221 μm, no sentido de propagação da onda. O número de Reynolds máximo permaneceu abaixo de 0.05 em todos os cenários de micropartículas, sustentando o regime de baixo Reynolds nas condições simuladas.Os resultados evidenciam uma forte dependência do efeito acústico com a escala da partícula. Enquanto a contribuição determinística da radiação acústica foi desprezível em comparação com a difusão térmica para partículas de 100 nm no intervalo temporal estudado, as micropartículas apresentaram deslocamento acústico sistemático e mensurável. As diferenças entre água e o modelo simplificado do humor vítreo foram influenciadas pelas propriedades acústicas e viscosas atribuídas a cada meio, sem demonstrar superioridade geral de um meio sobre o outro. Além disso, a aproximação de Rayleigh é adequada para partículas de 100 nm, mas apenas aproximada para partículas de 100 μm a 1 MHz, limitando a interpretação quantitativa deste último caso. Sob frequência e pressão fixas, o modelo prevê resposta acústica detectável para micropartículas de CaO, enquanto o comportamento das nanopartículas permanece predominantemente governado pela difusão browniana no intervalo simulado. Os resultados devem ser interpretados como uma análise computacional das condições específicas avaliadas, não permitindo inferências sobre otimização de frequência, intensidade, eficácia terapêutica ou benefício clínico.**

***Palavras-chave*—força de radiação acústica, movimento browniano, óxido de cálcio, simulação computacional, humor vítreo.**

I. Introdução

O GLAUCOMA é uma doença associada à perda progressiva da função visual, e a redução da pressão intraocular permanece um dos principais objetivos terapêuticos [1], [2]. A adesão ao tratamento tópico pode ser limitada por fatores relacionados à administração e ao uso crônico de medicamentos [3]. Por esse motivo, diferentes sistemas de liberação ocular têm sido investigados como estratégias de pesquisa para prolongar a permanência do fármaco ou modificar sua distribuição [5]. O ultrassom também tem sido estudado como estímulo físico em sistemas responsivos e em estratégias experimentais de transporte ocular [12], [13], [63], [65]. Essas referências justificam o interesse físico pelo acoplamento acústico, mas não constituem evidência de que partículas de CaO sejam seguras ou eficazes no tratamento do glaucoma. Entretanto, resultados obtidos em modelos de transporte de partículas não demonstram, por si só, eficácia terapêutica ou benefício clínico.

A manipulação acústica de micro e nanopartículas é uma abordagem estudada em microfluídica e bioengenharia, na qual a transferência de momento da onda pode produzir uma força média de radiação sobre partículas suspensas [21], [55]. O efeito depende do tamanho da partícula, das propriedades acústicas relativas entre partícula e fluido e do tipo de campo sonoro. Em paralelo, a agitação térmica se torna progressivamente mais relevante à medida que o tamanho da partícula diminui, exigindo que o movimento browniano seja considerado na nanoescala [19], [20].

O CaO foi adotado neste trabalho como material particulado de estudo. Embora sistemas baseados em carbonato de cálcio ($CaCO_3$) sejam discutidos em outros contextos biomédicos, seus resultados não devem ser extrapolados diretamente para CaO. Há literatura específica sobre síntese e aplicações biomédi-

cas emergentes de nanopartículas de CaO [11], além de rotas de síntese a partir de resíduos ricos em cálcio [14], [54]; contudo, o presente trabalho não avalia biocompatibilidade ocular, farmacocinética, toxicidade ou liberação de fármacos. Assim, o CaO é tratado aqui como uma partícula-modelo com propriedades físicas definidas, sem inferência de eficácia terapêutica.

Modelos computacionais são úteis para testar consistência física, ordens de grandeza e sensibilidade a hipóteses, mas sua validade depende da correspondência entre as equações implementadas e o regime físico representado [15]–[17]. Por essa razão, nesta revisão foram explicitados o regime de validade da aproximação acústica, o critério de arrasto e as limitações reológicas do meio vítreo.

O objetivo deste estudo é caracterizar numericamente a resposta de partículas esféricas de CaO em água e em uma representação simplificada do humor vítreo sob uma onda ultrassônica progressiva. Especificamente, são comparados dois raios de partícula e duas distâncias da fonte, mantendo frequência e pressão de referência constantes. Também são verificados o sentido do deslocamento, a consistência das unidades, a validade do regime de Stokes por meio do número de Reynolds e a aplicabilidade da aproximação de partícula pequena utilizada no modelo acústico.

II. Modelo Físico

A. Movimento browniano

Para uma esfera em um fluido Newtoniano, o coeficiente de difusão translacional é representado pela relação de Einstein–Smoluchowski,

$$D = \frac{k_B T}{\gamma}, \qquad \gamma = 6\pi\eta a, \tag{1}$$

em que k_B é a constante de Boltzmann, T é a temperatura absoluta, η é a viscosidade dinâmica e a é o raio da partícula [19], [20]. Em cada passo temporal, o incremento browniano em cada coordenada é

$$\Delta \mathbf{r}_B = \sqrt{2D\Delta t}\,\boldsymbol{\xi}, \tag{2}$$

com $\boldsymbol{\xi}$ formado por variáveis normais independentes de média zero e variância unitária.

B. Força de radiação acústica

Para representar uma esfera pequena submetida a uma onda plana progressiva, adotou-se a expressão de força de radiação de uma onda plana progressiva sobre uma esfera pequena, cuja direção é a da propagação e cuja magnitude de espalhamento apresenta dependência $(ka)^4$ [21], [55]. A magnitude da força média de radiação é

$$F_{\mathrm{ac}} = \frac{4\pi a^2 I}{9c}\left(f_1^2 + f_1 f_2 + \frac{3}{4} f_2^2\right)(ka)^4, \tag{3}$$

em que $k = 2\pi f/c$ e

$$f_1 = 1 - \frac{\kappa_p}{\kappa_0} = 1 - \frac{\rho c^2}{\rho_p c_p^2}, \tag{4}$$

$$f_2 = \frac{2(\rho_p - \rho)}{2\rho_p + \rho}. \tag{5}$$

Aqui, ρ e c são a densidade e a velocidade do som no fluido, enquanto ρ_p e c_p são as propriedades da partícula.

A intensidade local foi obtida de

$$I(d) = \frac{p(d)^2}{2\rho c}, \qquad p(d) = p_0\, 10^{-\alpha f_{\mathrm{MHz}} d_{\mathrm{cm}}/20}, \tag{6}$$

considerando apenas atenuação exponencial. Não foi introduzido fator adicional de espalhamento geométrico. O vetor unitário de propagação é

$$\hat{\mathbf{k}} = \frac{\mathbf{r} - \mathbf{r}_s}{\|\mathbf{r} - \mathbf{r}_s\|}, \qquad \mathbf{F}_{\mathrm{ac}} = F_{\mathrm{ac}}\hat{\mathbf{k}}. \tag{7}$$

Como a fonte está posicionada em $+x$ e a partícula inicia na origem, a onda chega à partícula propagando-se no sentido $-x$. Portanto, a componente acústica prevista é $F_x < 0$: o movimento determinístico ocorre no sentido de propagação da onda e, nesta geometria, corresponde a afastamento do emissor, não a movimento em direção a ele.

C. Arrasto e equação de movimento

O número de Reynolds da partícula foi calculado em cada passo como

$$Re_p = \frac{\rho |\mathbf{v}| (2a)}{\eta}. \tag{8}$$

Para $Re_p < 0.1$, empregou-se o arrasto de Stokes, $\mathbf{F}_D = -6\pi\eta a\mathbf{v}$. O código contém ainda a correlação de Schiller–Naumann para $Re_p \geq 0.1$, com

$$C_D = \frac{24}{Re_p}\left(1 + 0.15 Re_p^{0.687}\right), \tag{9}$$

e $\mathbf{F}_D = -\frac{1}{2}\rho C_D \pi a^2 |\mathbf{v}| \mathbf{v}/|\mathbf{v}|$ [58]. A verificação de Re_p evita aplicar a lei de Stokes indiscriminadamente à micropartícula.

Para a nanopartícula foi utilizado o limite superamortecido, $\mathbf{v}_d = \mathbf{F}_{\mathrm{ac}}/\gamma$, e a posição foi atualizada por $\mathbf{r}_{n+1} = \mathbf{r}_n + \mathbf{v}_d \Delta t + \Delta \mathbf{r}_B$. Para a micropartícula, a inércia foi preservada e integrou-se explicitamente $m\, d\mathbf{v}/dt = \mathbf{F}_{\mathrm{ac}} + \mathbf{F}_D$, acrescentando-se o incremento browniano à posição.

III. Materiais e Métodos

A. Parâmetros e cenários

A implementação foi realizada em Python com *NumPy* e *Matplotlib*. Todas as grandezas internas foram mantidas no SI. Foram executados 2000 passos com $\Delta t = 1 \times 10^{-6}$ s, correspondendo a 1.999 ms. A semente pseudoaleatória de base foi 20260903 para permitir reprodução das trajetórias. Foram avaliados exatamente oito cenários: dois raios (100 nm e 100 μm), dois meios (água e humor vítreo simplificado) e duas distâncias da fonte (16 e 21 mm). A frequência foi mantida em 1 MHz e a amplitude de pressão de referência em

TABLE I
PARÂMETROS FÍSICOS EMPREGADOS NA SIMULAÇÃO

Parâmetro	Valor	Base/observação
T	298 K	valor de simulação
$\rho_{\text{água}}$	997 kg/m^3	IAPWS, arredondado [53]
$\eta_{\text{água}}$	8.9×10^{-4} Pa·s	IAPWS, arredondado [53]
$c_{\text{água}}$	1500 m/s	valor arredondado a 25°C [60]
$\rho_{\text{vítreo}}$	1005 kg/m^3	limite inferior reportado [56]
$\eta_{\text{vítreo}}$	3.65×10^{-3} Pa·s	proxy Newtoniano adotado; cf. [56]
$c_{\text{vítreo}}$	1532 m/s	biometria ultrassônica [59]
ρ_{CaO}	3345 kg/m^3	propriedade adotada
c_{CaO}	4880 m/s	propriedade adotada [57]
f	1 MHz	fixo
p_0	100 kPa	fixo
α	0.1 dB/(cm·MHz)	hipótese do modelo
d_s	16 e 21 mm	cenários

100 kPa em todos os casos; não houve variação desses dois parâmetros.

O valor 3.65 mPa·s é tratado nesta versão explicitamente como um *proxy* Newtoniano efetivo adotado no código, e não como uma constante fisiológica universal. Silva *et al.* demonstraram que o humor vítreo é viscoelástico e fortemente dependente da taxa de cisalhamento, com valores da ordem de 10^{-3} Pa·s em regimes de maior cisalhamento [56]. Assim, o parâmetro usado preserva a execução reproduzida, mas sua natureza aproximativa é declarada e considerada na discussão. A densidade de 1005 kg/m^3 corresponde ao arredondamento do limite inferior da faixa de 1.0053–1.0089 g/cm^3 reportada para o vítreo [56]. Para a água próxima de 298 K, densidade e viscosidade foram arredondadas a partir das correlações recomendadas pela IAPWS [53]; a velocidade de 1500 m/s é um arredondamento do valor de aproximadamente 1497 m/s a 25°C obtido pela expressão de Lubbers e Graaff [60]. Para o vítreo, 1532 m/s foi mantido por ser o valor convencional empregado em biometria A-scan para a câmara anterior e o humor vítreo [59].

O domínio computacional foi definido como um cubo de $50 \times 50 \times 50$ mm^3, centrado na origem, apenas como limite numérico. A trajetória não foi artificialmente limitada por *clipping*; uma saída do domínio interromperia a execução como erro de modelagem. A ação acústica só foi aplicada após o tempo inicial de propagação d_s/c. Os resultados foram exportados em CSV, incluindo posição, velocidade, aceleração, forças, pressão, intensidade e Re_p em cada passo.

B. Implementação computacional e reprodutibilidade

O programa foi organizado em estruturas separadas para propriedades do meio e da partícula e em uma função única de simulação por cenário. A fonte é posicionada em $(d_s, 0, 0)$ e a partícula em $(0, 0, 0)$. Em cada passo, calcula-se primeiro a distância fonte–partícula e a amplitude de pressão local pela Eq. (6). A força acústica permanece nula antes do tempo de chegada d_s/c; após esse instante, a Eq. (3) é avaliada e multiplicada pelo vetor unitário da Eq. (7). Essa implementação elimina a força eletrostática que aparecia em versões preliminares do código, pois ela não fazia parte da metodologia declarada.

Para $a = 100$ nm, o tempo de relaxação inercial é muito menor que o passo macroscópico de interesse e foi usado o limite superamortecido. A deriva determinística é calculada por $\mathbf{F}_{\text{ac}}/\gamma$ e somada ao incremento estocástico da Eq. (2). Para $a = 100$ μm, a velocidade é integrada explicitamente a partir da força acústica e do arrasto. O Re_p é recalculado a cada iteração: Stokes é usado quando $Re_p < 0.1$ e o algoritmo dispõe da correlação de Schiller–Naumann para valores superiores, conforme a literatura clássica de arrasto de partículas [58]. Como os resultados desta execução apresentaram $Re_p < 0.05$, a ramificação de Schiller–Naumann não foi acionada nos cenários reportados.

Cada uma das oito combinações recebeu uma semente determinística derivada da semente-base 20260903. Foram armazenados, a cada 1 μs, posição, velocidade, aceleração, força acústica, força de arrasto, pressão, intensidade, deslocamento, Re_p e o modelo de arrasto selecionado. As figuras foram produzidas a 300 dpi a partir dessas séries. A frequência de 1 MHz e a pressão de referência de 100 kPa são entradas fixas: não existe laço de varredura desses parâmetros. A intensidade local varia apenas como consequência das propriedades do meio, da distância e da atenuação da Eq. (6), não como um parâmetro experimental independente.

C. Diagnóstico de validade do modelo acústico

A Eq. (3) pressupõe partícula pequena em relação ao comprimento de onda. Como diagnóstico conservador, registrou-se $2a/\lambda$ e ka. Para $a = 100$ nm, obteve-se $ka \approx 4.1 \times 10^{-4}$ e $2a/\lambda \approx 1.3 \times 10^{-4}$, claramente no regime de Rayleigh. Para $a = 100$ μm, $ka \approx 0.41$ e $2a/\lambda \approx 0.13$. Assim, o caso de micropartícula permanece subcomprimento de onda, porém fora do critério conservador $2a/\lambda \leq 0.1$ adotado no código; seus resultados devem ser interpretados como aproximação de primeira ordem, não como previsão de alta fidelidade.

IV. RESULTADOS

A Tabela II resume os oito cenários, com deslocamentos expressos em micrômetros e forças em newtons. Em nenhum caso o deslocamento se aproximou da escala do domínio numérico de 25 mm por semieixo.

A. Nanopartículas

A Fig. 1 mostra que as quatro trajetórias de nanopartículas apresentam flutuações aleatórias em escala submicrométrica. O componente determinístico acústico acumulado foi inferior a 2×10^{-15} μm em módulo, enquanto o deslocamento total da realização estocástica variou entre 0.028 e 0.265 μm. Portanto, o sinal da posição final de uma única trajetória não pode ser interpretado como direção de transporte acústico. Nas condições adotadas, a contribuição browniana domina completamente a resposta da nanopartícula no intervalo simulado.

O coeficiente de difusão calculado foi 2.45×10^{-12} m^2/s em água e 5.98×10^{-13} m^2/s no meio vítreo, razão de aproximadamente 4.1. Isso indica maior dispersão browniana esperada em água; contudo, como foi utilizada uma única

TABLE II
RESULTADOS DA SIMULAÇÃO APÓS 1.999 MS

Cenário	x_f (μm)	x_{det} (μm)	$\|\Delta\mathbf{r}\|$ (μm)	$\overline{F}_{ac}$ (N)	Re_{max}	ka
NP, vítreo, 16 mm	0.0666	-4.22×10^{-16}	0.0752	1.46×10^{-27}	1.17×10^{-20}	4.10×10^{-4}
NP, vítreo, 21 mm	-0.0219	-4.17×10^{-16}	0.0284	1.44×10^{-27}	1.16×10^{-20}	4.10×10^{-4}
NP, água, 16 mm	-0.00290	-1.99×10^{-15}	0.0716	1.68×10^{-27}	2.24×10^{-19}	4.19×10^{-4}
NP, água, 21 mm	0.0965	-1.96×10^{-15}	0.265	1.66×10^{-27}	2.22×10^{-19}	4.19×10^{-4}
MP, vítreo, 16 mm	-0.151	-0.153	0.151	1.46×10^{-9}	0.00729	0.410
MP, vítreo, 21 mm	-0.149	-0.151	0.149	1.44×10^{-9}	0.00720	0.410
MP, água, 16 mm	-0.221	-0.219	0.222	1.68×10^{-9}	0.0475	0.419
MP, água, 21 mm	-0.213	-0.216	0.213	1.66×10^{-9}	0.0469	0.419

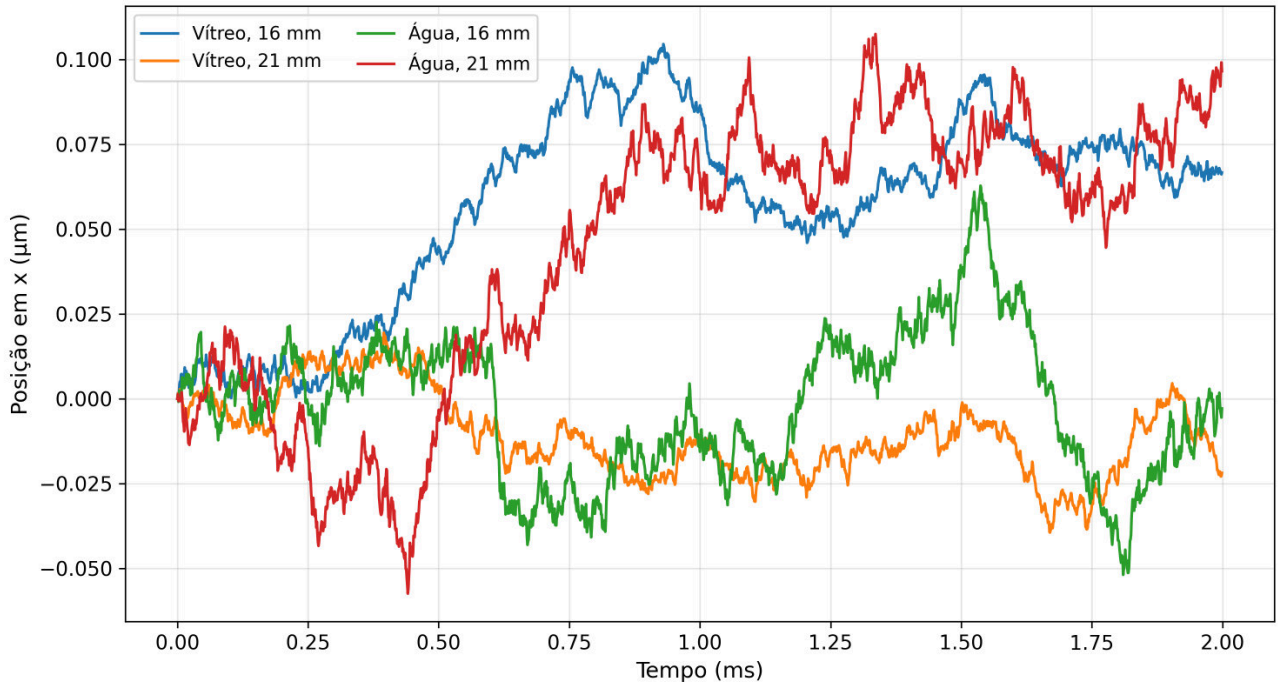


Fig. 1. Posição em x das nanopartículas para água e humor vítreo simplificado, nas distâncias de 16 e 21 mm. A frequência e a pressão de referência foram mantidas em 1 MHz e 100 kPa.

realização pseudoaleatória por cenário, os valores finais individuais não constituem uma estimativa estatística da variância. Uma comparação quantitativa de dispersão entre meios exigiria múltiplas réplicas estocásticas ou uma análise de conjunto.

B. Micropartículas

Para as micropartículas, o componente determinístico é dominante no intervalo de 1.999 ms. A Fig. 2 mostra deslocamento monotônico para $-x$, coerente com a direção de propagação definida na Eq. (7). Os valores finais ficaram entre -0.149 e -0.221 μm. Em água, o módulo do deslocamento final foi aproximadamente 46% maior que no vítreo para 16 mm e 43% maior para 21 mm. A distância de 16 mm produziu magnitude apenas 1.8% maior que 21 mm no vítreo e 3.8% maior em água, compatível com a pequena diferença de atenuação introduzida entre as duas posições.

A velocidade máxima calculada para as micropartículas foi aproximadamente 1.31×10^{-4}–1.32×10^{-4} m/s no meio vítreo e 2.09×10^{-4}–2.12×10^{-4} m/s em água. Como o tempo de simulação é comparável ou inferior ao tempo de relaxação inercial da micropartícula, não se assume que uma velocidade terminal tenha sido atingida.

C. Validade de Stokes

A Fig. 3 apresenta o número de Reynolds das micropartículas. O maior valor foi 0.0475, registrado em água, enquanto no vítreo o máximo foi aproximadamente 0.0073. Todos os valores ficaram abaixo do limiar conservador de 0.1 adotado no algoritmo. Consequentemente, a lei de Stokes é consistente com os oito casos efetivamente simulados, embora o código mantenha a correção de Schiller–Naumann como salvaguarda para configurações futuras com Re_p mais elevado.

V. DISCUSSÃO

A formulação de onda progressiva empregada contém o fator $(ka)^4$ [21], [55]. Como ka é aproximadamente mil vezes menor para a nanopartícula de 100 nm do que para a micropartícula de 100 μm, a força de espalhamento cai fortemente na

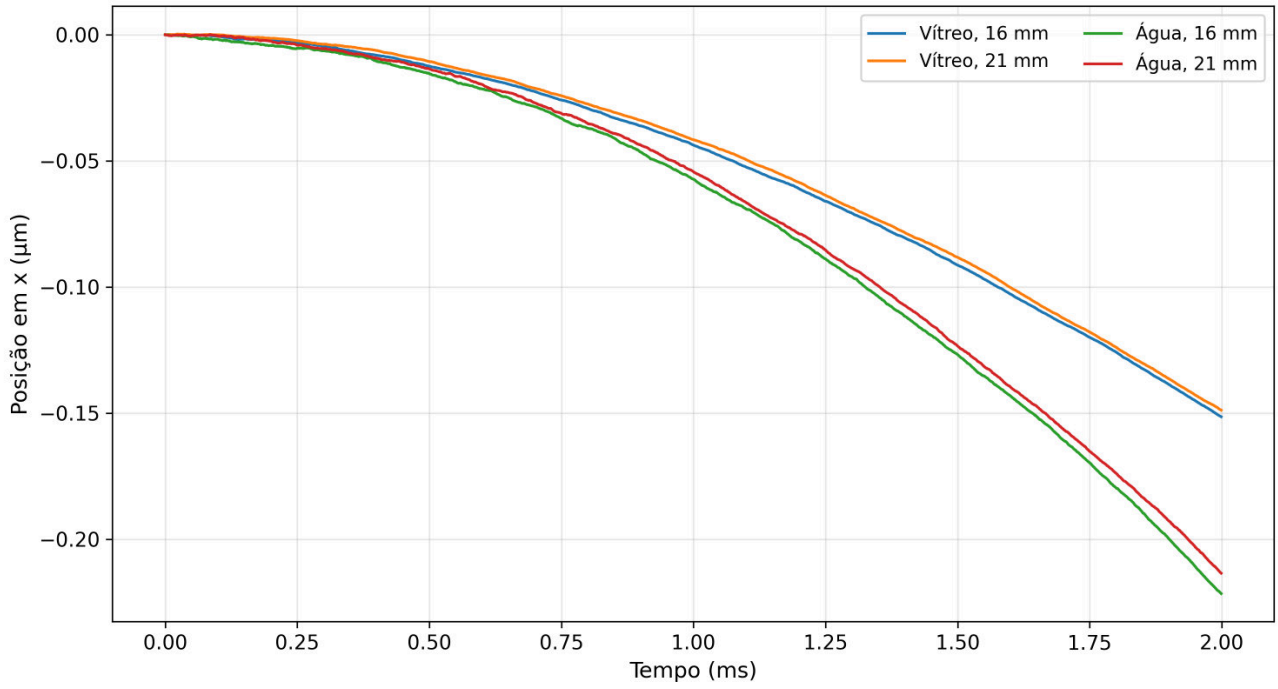


Fig. 2. Posição em x das micropartículas. O deslocamento negativo corresponde ao sentido de propagação da onda na geometria simulada.

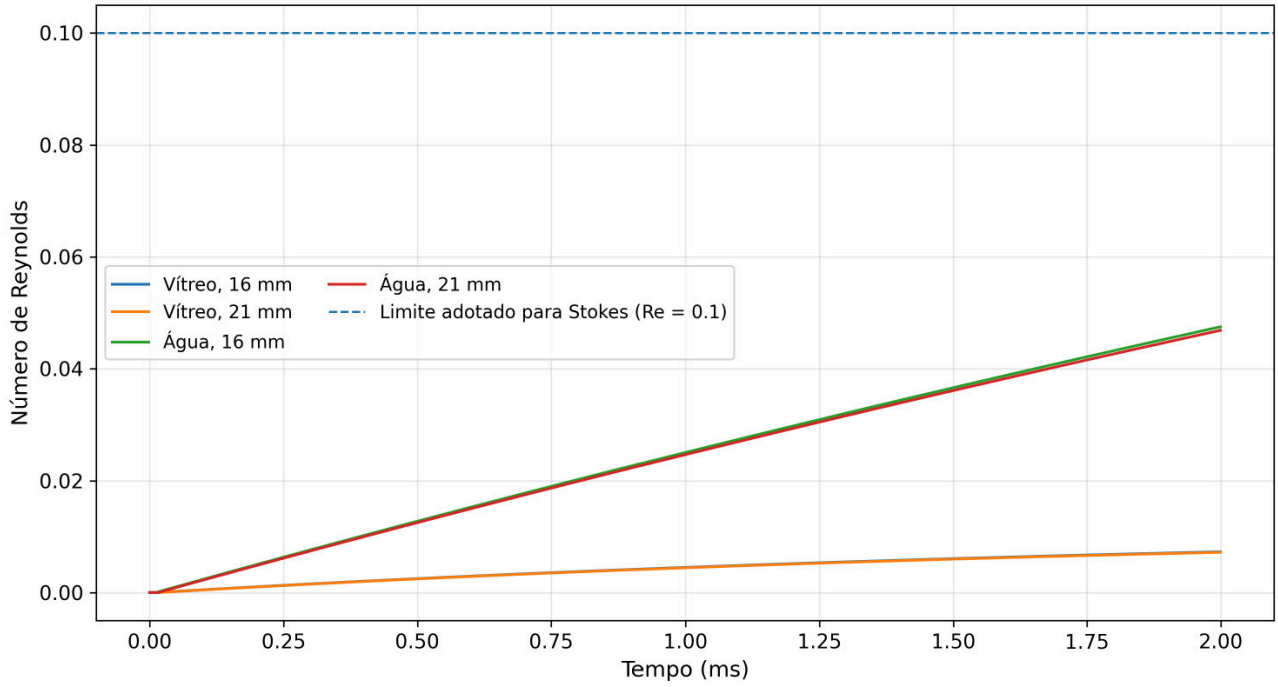


Fig. 3. Número de Reynolds das micropartículas ao longo do tempo. A linha tracejada indica o limiar $Re_p = 0.1$ adotado para a seleção do modelo de arrasto.

nanoescala. Isso explica a diferença de ordem de grandeza entre a força média obtida para as nanopartículas (10^{-27} N) e para as micropartículas (10^{-9} N). A escala submicrométrica dos deslocamentos também permanece consistente com o domínio computacional adotado.

Uma onda progressiva transfere momento na direção de propagação [55]. Com a fonte localizada em $+x$ e a partícula na origem, a direção fonte–partícula é negativa em x; portanto, o deslocamento acústico negativo representa movimento no sentido da onda e, nesta geometria, afastamento do emissor.

A comparação entre água e humor vítreo deve ser interpretada com cautela. Para as micropartículas, a menor viscosidade da água permite maior velocidade e deslocamento durante o mesmo intervalo, e o modelo também prevê força acústica ligeiramente superior por causa das propriedades acústicas do meio. Para as nanopartículas, entretanto, o termo acústico é tão pequeno que as diferenças observadas em uma única trajetória são controladas pela realização browniana. Portanto, não é adequado concluir, a partir desses quatro traçados isolados, que uma distância ou meio produz necessariamente maior deslocamento final de nanopartículas.

Os resultados também esclarecem a inconsistência de escala presente na versão anterior do manuscrito. Os deslocamentos de micropartículas são submicrométricos e representam menos de 10^{-5} do semitamanho de 25 mm do domínio. Portanto, não há compatibilidade física com interpretações anteriores em que deslocamentos da ordem de milímetros ou dezenas de milímetros eram atribuídos ao mesmo intervalo de poucos milissegundos. A correção da equação e a manutenção de todas as grandezas internas no SI removem essa discrepância dimensional.

A diferença entre 16 e 21 mm é pequena porque, no modelo atual, a distância afeta a amplitude somente pelo termo de atenuação; nenhum termo geométrico arbitrário foi acrescentado. Assim, o estudo não demonstra uma distância ótima. Analogamente, como frequência e pressão de referência são únicas, os resultados não permitem inferir tendências em função de frequência ou intensidade. Estudos de otimização exigiriam uma matriz paramétrica independente e, idealmente, validação experimental [15], [17], [61], [62].

A representação do humor vítreo constitui outra limitação relevante. O vítreo possui comportamento viscoelástico e dependente da taxa de cisalhamento [56], enquanto o código utiliza uma viscosidade newtoniana constante como parâmetro efetivo de comparação. O modelo é, assim, uma aproximação reduzida para comparação de parâmetros e não uma reprodução completa do ambiente ocular. Da mesma forma, o coeficiente de atenuação, a pressão e a frequência foram fixados como parâmetros de simulação, sem estudo de sensibilidade.

Por fim, a Eq. (3) é robustamente compatível com a condição de partícula pequena para 100 nm, mas a micropartícula apresenta $ka \simeq 0.41$. Embora ainda seja menor que o comprimento de onda, esse valor não satisfaz o critério conservador usado para classificar o regime estrito de Rayleigh. Assim, os resultados de micropartículas são úteis para demonstrar ordem de grandeza e tendências internas do modelo, mas uma validação quantitativa exigiria uma solução de espalhamento de tamanho finito, caracterização reológica mais completa e comparação com dados experimentais.

O estudo também não testou diferentes frequências ou amplitudes acústicas, não avaliou liberação de fármacos, toxicidade, segurança ocular ou desfechos relacionados ao glaucoma. Consequentemente, não se pode inferir otimização de parâmetros ultrassônicos, maior fidelidade clínica, melhoria de adesão terapêutica ou benefício médico. Tais questões devem ser tratadas em trabalhos experimentais e paramétricos posteriores.

VI. Conclusão

O modelo computacional aplicado aos oito cenários indica comportamentos distintos entre as duas escalas de partícula. Com frequência de 1 MHz e pressão de referência de 100 kPa mantidas fixas, as nanopartículas de raio 100 nm apresentaram dinâmica dominada pelo movimento browniano, com contribuição acústica determinística desprezível no intervalo de 1.999 ms. As micropartículas de raio 100 μm apresentaram deslocamento determinístico da ordem de 0.15–0.22 μm no sentido de propagação da onda. O número de Reynolds permaneceu abaixo de 0.05, sustentando o emprego de Stokes nos casos calculados. Entretanto, a aproximação acústica de Rayleigh é apenas aproximada para a micropartícula a 1 MHz, e o humor vítreo foi representado como um fluido Newtoniano efetivo. Portanto, os resultados caracterizam tendências computacionais sob hipóteses e parâmetros fixos, sem demonstrar eficácia clínica, otimização ultrassônica ou desempenho terapêutico em aplicações oftálmicas.

Agradecimentos

Este estudo recebeu apoio da Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) (PS1-0186-00216.01.00/21), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (311062/2019-9) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (código de financiamento 001).

Conflito de Interesses

Os autores declaram não possuir interesses financeiros conflitantes ou relações pessoais conhecidas que possam ter influenciado o trabalho relatado neste artigo.

Referências

[1] Queiroz B, Oliveira Delgado Mota L, "O impacto do glaucoma na qualidade de vida: uma revisão sistemática," *Revista de Saúde*, vol. 12, pp. 8–12, 2021, doi: https://doi.org/10.21727/rs.v12i2.2478.

[2] Lopes AB, Viana L de A, Faria IS, Campos VAA, Paiva MGS, Araújo FN de, et al., "Hipertensão ocular: uma revisão narrativa sobre o glaucoma," *Revista Eletrônica Acervo Científico*, vol. 41, Art. no. e9987, 2022, doi: https://doi.org/10.25248/reac.e9987.2022.

[3] Cvenkel B, Kolko M, "Devices and Treatments to Address Low Adherence in Glaucoma Patients: A Narrative Review," *J Clin Med*, vol. 12, p. 151, 2022, doi: https://doi.org/10.3390/jcm12010151.

[4] Abdullah F, Imtiaz R, Madni HA, Khan HA, Khan TM, Khan MAU, et al., "A Review on Glaucoma Disease Detection Using Computerized Techniques," *IEEE Access*, vol. 9, pp. 37311–37333, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3061451.

[5] Lavik E, Kuehn MH, Kwon YH, "Novel drug delivery systems for glaucoma," *Eye*, vol. 25, pp. 578–86, 2011, doi: https://doi.org/10.1038/eye.2011.82.

[6] Singh A, Dubey S, Dubey HK, "Nanotechnology: The Future Engineering," *Int J Adv Innov Res*, vol. 6, pp. 230–3, 2019.

[7] Leigh DA, "Genesis of the Nanomachines: The 2016 Nobel Prize in Chemistry," *Angew. Chem. Int. Ed.*, vol. 55, pp. 14506–8, 2016, doi: https://doi.org/10.1002/anie.201609841.

[8] Chesnitskiy A V., Gayduk AE, Seleznev VA, Prinz VY, "Bio-Inspired Micro- and Nanorobotics Driven by Magnetic Field," *Materials*, vol. 15, p. 7781, 2022, doi: https://doi.org/10.3390/ma15217781.

[9] Menezes Ferreira AÁ, da Silva Felix JH, Chaves de Lima RK, Martins de Souza MC, Sousa dos Santos JC, "Advancements and Prospects in Nanorobotic Applications for Ophthalmic Therapy," *ACS Biomater Sci Eng*, vol. 11, pp. 958–80, 2025, doi: https://doi.org/10.1021/acsbiomaterials.4c02368.

[10] Zhan Z, Wei F, Zheng J, Yang W, Luo J, Yao L, "Recent advances of light-driven micro/nanomotors: toward powerful thrust and precise control," *Nanotechnol Rev*, vol. 7, pp. 555–81, 2018, doi: https://doi.org/10.1515/ntrev-2018-0106.

[11] M. Kumari, B. Sarkar, and K. Mukherjee, "Nanoscale calcium oxide and its biomedical applications: A comprehensive review," *Biocatalysis and Agricultural Biotechnology*, vol. 47, Art. no. 102506, 2023, doi: https://doi.org/10.1016/j.bcab.2022.102506.

[12] Husseini GA, Pitt WG, "Micelles and nanoparticles for ultrasonic drug and gene delivery," *Adv Drug Deliv Rev*, vol. 60, pp. 1137–52, 2008, doi: https://doi.org/10.1016/j.addr.2008.03.008.

[13] Entzian K, Aigner A, "Drug Delivery by Ultrasound-Responsive Nanocarriers for Cancer Treatment," *Pharmaceutics*, vol. 13, p. 1135, 2021, doi: https://doi.org/10.3390/pharmaceutics13081135.

[14] Banik O, Priyadarshi S, Banoth E, "Facile Calcination Method to Synthesize Calcium Oxide Nanoparticles from Eggshell Wastes," in *Frontiers in Optics + Laser Science 2023 (FiO, LS)*, Tacoma, WA, USA: Optica Publishing Group, 2023, p. JM7A.136, doi: https://doi.org/10.1364/FIO.2023.JM7A.136.

[15] DOUGHERTY ER, BRAGA-NETO U, "EPISTEMOLOGY OF COMPUTATIONAL BIOLOGY: MATHEMATICAL MODELS AND EXPERIMENTAL PREDICTION AS THE BASIS OF THEIR VALIDITY," *J Biol Syst*, vol. 14, pp. 65–90, 2006, doi: https://doi.org/10.1142/S0218339006001726.

[16] Shi J, Zhu X, Sun K, Fang L, "Movement pattern of an ellipsoidal nanoparticle confined between solid surfaces: Theoretical model and molecular dynamics simulation," *Friction*, vol. 9, pp. 1098–109, 2021, doi: https://doi.org/10.1007/s40544-020-0402-2.

[17] Carchini G, Almora-Barrios N, Revilla-López G, Bellarosa L, García-Muelas R, García-Melchor M, et al., "How Theoretical Simulations Can Address the Structure and Activity of Nanoparticles," *Top. Catal.*, vol. 56, pp. 1262–1272, 2013, doi: https://doi.org/10.1007/s11244-013-0093-3.

[18] Brown R, "XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies," *The Philosophical Magazine*, vol. 4, pp. 161–73, 1828, doi: https://doi.org/10.1080/14786442808674769.

[19] Einstein A, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," *Ann Phys*, vol. 322, pp. 549–60, 1905, doi: https://doi.org/10.1002/andp.19053220806.

[20] von Smoluchowski M, "Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen," *Ann Phys*, vol. 326, pp. 756–80, 1906, doi: https://doi.org/10.1002/andp.19063261405.

[21] Bruus H, "Acoustofluidics 7: The acoustic radiation force on small particles," *Lab Chip*, vol. 12, p. 1014, 2012, doi: https://doi.org/10.1039/c2lc21068a.

[22] Maiocchi R, "The case of Brownian motion," *The British Journal for the History of Science*, vol. 23, pp. 257–83, 1990, doi: https://doi.org/10.1017/S0007087400043983.

[23] Montgomery D, "Brownian Motion from Boltzmann's Equation," *Phys Fluids*, vol. 14, pp. 2088–90, 1971, doi: https://doi.org/10.1063/1.1693297.

[24] Belyi VV, Klimontovich YuL, "On the kinetic theory of Brownian motion," *Physica A: Statistical Mechanics and Its Applications*, vol. 97, pp. 577–88, 1979, doi: https://doi.org/10.1016/0378-4371(79)90097-9.

[25] Raikher YuL, Rusakov V V, "Theory of Brownian motion in a Jeffreys fluid," *Journal of Experimental and Theoretical Physics*, vol. 111, pp. 883–9, 2010, doi: https://doi.org/10.1134/S1063776110110191.

[26] Valido AA, "Quantum kinetic theory of flux-carrying Brownian particles," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2022, Art. no. 073103, 2022, doi: https://doi.org/10.1088/1742-5468/ac7a2e.

[27] Klimontovich YL, "Nonlinear Brownian motion," *Physics-Uspekhi*, vol. 37, pp. 737–66, 1994, doi: https://doi.org/10.1070/PU1994v037n08ABEH000038.

[28] Green PF, "Brownian Motion," in *Kinetics, Transport, and Structure in Hard and Soft Materials*. CRC Press, 2005, pp. 51–76, doi: https://doi.org/10.1201/9781420027143.ch2.

[29] Jakobsson E, Chiu SW, "Application of Brownian motion theory to the analysis of membrane channel ionic trajectories calculated by molecular dynamics," *Biophys J*, vol. 54, pp. 751–6, 1988, doi: https://doi.org/10.1016/S0006-3495(88)83012-1.

[30] Lindner B, "The diffusion coefficient of nonlinear Brownian motion," *New J Phys*, vol. 9, pp. 136–136, 2007, doi: https://doi.org/10.1088/1367-2630/9/5/136.

[31] Wu X, "Using Brownian model to study the effect of temperature on diffusion coefficient," *Theoretical and Natural Science*, vol. 36, pp. 48–57, 2024, doi: https://doi.org/10.54254/2753-8818/36/20240511.

[32] Felderhof BU, "Diffusion of interacting Brownian particles," *J Phys A Math Gen*, vol. 11, pp. 929–37, 1978, doi: https://doi.org/10.1088/0305-4470/11/5/022.

[33] Yang LM, "Diffusion and the Brownian motion," *Proc R Soc Lond A Math Phys Sci*, vol. 198, pp. 94–116, 1949, doi: https://doi.org/10.1098/rspa.1949.0089.

[34] Basu K, Baishya K, "Brownian motion: Theory and experiment a simple classroom measurement of the diffusion coefficient," *Resonance*, vol. 8, pp. 71–80, 2003, doi: https://doi.org/10.1007/BF02835809.

[35] Velasco S, "On the Brownian motion of a harmonically bound particle and the theory of a Wiener process," *Eur J Phys*, vol. 6, pp. 259–65, 1985, doi: https://doi.org/10.1088/0143-0807/6/4/009.

[36] Levy BC, "Wiener Process and White Gaussian Noise," in *Random Processes with Applications to Circuits and Communications*. Cham, Switzerland: Springer, 2020, pp. 207–234, doi: https://doi.org/10.1007/978-3-030-22297-0_6.

[37] Hassler U, "Wiener Processes (WP)," in *Stochastic Processes and Calculus*. Cham, Switzerland: Springer, 2016, pp. 151–177, doi: https://doi.org/10.1007/978-3-319-23428-1_7.

[38] Ross SM, "Brownian Motion and Stationary Processes," in *Introduction to Probability Models*. Elsevier, 2010, pp. 631–666, doi: https://doi.org/10.1016/B978-0-12-375686-2.00012-1.

[39] Nakagawa T, "Brownian Motion and Lévy Processes," in *Stochastic Processes*. London, U.K.: Springer, 2011, pp. 183–197, doi: https://doi.org/10.1007/978-0-85729-274-2_7.

[40] Qiao Y, Zhang X, Zhang G, "Acoustic radiation force on a fluid cylindrical particle immersed in water near an impedance boundary," *J Acoust Soc Am*, vol. 141, pp. 4633–41, 2017, doi: https://doi.org/10.1121/1.4986624.

[41] DOINIKOV AA, "Acoustic radiation interparticle forces in a compressible fluid," *J Fluid Mech*, vol. 444, pp. 1–21, 2001, doi: https://doi.org/10.1017/S0022112001005055.

[42] Winckelmann BG, Bruus H, "Acoustic radiation force on a heated spherical particle in a fluid including scattering and microstreaming from a standing ultrasound wave," *Phys Rev E*, vol. 108, Art. no. 035108, 2023, doi: https://doi.org/10.1103/PhysRevE.108.035108.

[43] Mo R-Y, Hu J, Chen S, Wang C-H, "Acoustic radiation force on thin elastic shells in liquid*," *Chinese Physics B*, vol. 29, Art. no. 094301, 2020, doi: https://doi.org/10.1088/1674-1056/ab943e.

[44] Bruus H, "Acoustic Radiation Force on Small Particles," in *Microscale Acoustofluidics*. Cambridge, U.K.: Royal Society of Chemistry, 2014, pp. 65–80, doi: https://doi.org/10.1039/9781849737067-00065.

[45] Doinikov AA, Fankhauser J, Dual J, "Nonlinear dynamics of a solid particle in an acoustically excited viscoelastic fluid. II. Acoustic radiation force," *Phys Rev E*, vol. 104, Art. no. 065108, 2021, doi: https://doi.org/10.1103/PhysRevE.104.065108.

[46] Lebowitz JL, Rubin E, "Dynamical Study of Brownian Motion," *Physical Review*, vol. 131, pp. 2381–96, 1963, doi: https://doi.org/10.1103/PhysRev.131.2381.

[47] Sharma N, Patankar NA, "Direct numerical simulation of the Brownian motion of particles by using fluctuating hydrodynamic equations," *J Comput Phys*, vol. 201, pp. 466–86, 2004, doi: https://doi.org/10.1016/j.jcp.2004.06.002.

[48] Ramshaw JD, "Brownian motion in a flowing fluid," *Phys Fluids*, vol. 22, pp. 1595–601, 1979, doi: https://doi.org/10.1063/1.862818.

[49] Tóthová J, Lisý V, "Brownian motion in a viscous fluid of particles with constant and time-dependent friction," *Physics of Fluids*, vol. 36, 2024, doi: https://doi.org/10.1063/5.0187649.

[50] Dong S, Zheng L, Zhang X, Wu S, Shen B, "A new model for Brownian force and the application to simulating nanofluid flow," *Microfluid Nanofluidics*, vol. 16, pp. 131–9, 2014, doi: https://doi.org/10.1007/s10404-013-1213-x.

[51] Mazur P, Oppenheim I, "Molecular theory of Brownian motion," *Physica*, vol. 50, pp. 241–58, 1970, doi: https://doi.org/10.1016/0031-8914(70)90005-4.

[52] Pitre L, "Determinations of the Boltzmann Constant," in *2018 Conf. Precision Electromagnetic Measurements (CPEM)*, 2018, pp. 1–2, doi: https://doi.org/10.1109/CPEM.2018.8501168.

[53] International Association for the Properties of Water and Steam, "Revised Supplementary Release on Properties of Liquid Water at 0.1 MPa," IAPWS SR6-08(2011), Sep. 2011.

[54] Habte L, Shiferaw N, Mulatu D, Thenepalli T, Chilakala R, Ahn J, "Synthesis of Nano-Calcium Oxide from Waste Eggshell by Sol-Gel Method," *Sustainability*, vol. 11, p. 3196, 2019, doi: https://doi.org/10.3390/su11113196.

[55] T. J. Matula *et al.*, "Ultrasound-based cell sorting with microbubbles: A feasibility study," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. 41–52, 2018, doi: https://doi.org/10.1121/1.5044405.

[56] A. F. Silva, M. A. Alves, and M. S. N. Oliveira, "Rheological behaviour of vitreous humour," *Rheol. Acta*, vol. 56, pp. 377–386, 2017, doi: https://doi.org/10.1007/s00397-017-0997-0.

[57] Soga N, "New measurements on the sound velocity of calcium oxide and its relation to Birch's law," *J Geophys Res*, vol. 72, pp. 5157–9, 1967, doi: https://doi.org/10.1029/JZ072i020p05157.

[58] R. Clift, J. R. Grace, and M. E. Weber, *Bubbles, Drops, and Particles*. New York, NY, USA: Academic Press, 1978.

[59] L. Petrella, F. Perdigão, M. Caixinha, M. Santos, M. Lopes, M. Gomes, and J. Santos, "A-scan ultrasound in ophthalmology: A simulation tool," *Med. Eng. Phys.*, vol. 97, pp. 18–24, 2021, doi: https://doi.org/10.1016/j.medengphy.2021.09.005.

[60] J. Lubbers and R. Graaff, "A simple and accurate formula for the sound velocity in water," *Ultrasound Med. Biol.*, vol. 24, no. 7, pp. 1065–1068, 1998, doi: https://doi.org/10.1016/S0301-5629(98)00091-X.

[61] Yoon S, Aglyamov S, Karpiouk A, Emelianov S, "A high pulse repetition frequency ultrasound system for the ex vivo measurement of mechanical properties of crystalline lenses with laser-induced microbubbles interrogated by acoustic radiation force," *Phys Med Biol*, vol. 57, pp. 4871–84, 2012, doi: https://doi.org/10.1088/0031-9155/57/15/4871.

[62] Ostrovsky LA, Priev A, Ponomarev V, Barenholz Y, "Acoustic radiation force for rapid detection of particles in biological liquids," 2012, Art. no. 020002, doi: https://doi.org/10.1121/1.3700239.

[63] Duncan B, Al-Kassas R, Zhang G, Hughes D, Qiu Y, "Ultrasound-Mediated Ocular Drug Delivery: From Physics and Instrumentation to Future Directions," *Micromachines (Basel)*, vol. 14, p. 1575, 2023, doi: https://doi.org/10.3390/mi14081575.

[64] Bekerman I, Gottlieb P, Vaiman M, "Variations in Eyeball Diameters of the Healthy Adults," *J Ophthalmol*, vol. 2014, pp. 1–5, 2014, doi: https://doi.org/10.1155/2014/503645.

[65] Driller J, Lizzi FL, "Therapeutic Applications of Ultrasound: A Review," *IEEE Engineering in Medicine and Biology Magazine*, vol. 6, pp. 33–40, 1987, doi: https://doi.org/10.1109/MEMB.1987.5006466.

Three-Dimensional Simulation of Respiratory Airflow Based on Computed Tomography in Tracheobronchopathia Osteochondroplastica

1st Isabella Berti Reale Cruz
Department of Science and Technology
Institute of Science and Technology
UNIFESP
São José dos Campos, Brazil
isabella.berti@unifesp.br

2nd Anna Carolina Magalhães Baraúna
Department of Science and Technology
Institute of Science and Technology
UNIFESP
São José dos Campos, Brazil
anna.barauna@unifesp.br

3rd Yuri Rocha Cocati
Department of Science and Technology
Institute of Science and Technology
UNIFESP
São José dos Campos, Brazil
cocati@unifesp.br

4th Paulo Francisco Guerreiro Cardoso
Division of Thoracic Surgery Instituto do Coracao do Hospital das Clinicas HCFMUSP da Faculdade de Medicina da Universidade de Sao Paulo
São Paulo, Brazil
cardosop@gmail.com

5th Fabio Gava Aoki
Department of Science and Technology
Institute of Science and Technology
UNIFESP
São José dos Campos, Brazil
fgaoki@unifesp.br

***Abstract*—Tracheobronchopathia osteochondroplastica (TO) is a rare benign condition characterized by the formation of cartilaginous and bony nodules in the tracheobronchial submucosa, which can lead to chronic cough, dyspnea, and recurrent infections. Computational Fluid Dynamics (CFD) combined with computed tomography (CT) has emerged as a promising tool for the personalized functional analysis of airways. This study developed 3D CFD simulations comparing the trachea of a patient with TO to that of a control subject. Geometries were reconstructed in 3D Slicer from CT images obtained from the InCor database (CAAE 88886025.80000.0068) and simulated in ANSYS using the k-omega SST turbulence model. The results demonstrated that the asymmetries caused by the nodules create local constrictions that seem to accelerate airflow and increase Turbulent Kinetic Energy (TKE). Wall shear stress (WSS) was elevated and heterogeneously distributed in the TO model, suggesting stagnation zones conducive to bacterial proliferation and altered mucociliary transport. Although limited by the analysis of one control and one patient with pathology, these data provide a quantitative biophysical perspective that aids in understanding the aerodynamic mechanisms associated with the clinical manifestations of TO.**

***Index Terms*—tracheobronchopathia osteochondroplastica, computational fluid dynamics, computed tomography, airflow dynamics, airway resistance.**

I. Introduction

Tracheobronchopathia osteochondroplastica (TO) is a rare benign condition, typically identified in patients over the age of 50 [1]. Its etiology remains incompletely understood; the condition is characterized by the formation of multiple cartilaginous and/or bony nodules within the submucosa of the tracheal and main bronchial walls, projecting into the airway lumen [2]. These nodules are predominantly distributed along the anterolateral walls, sparing the posterior wall, a characteristic that aids in diagnosis during imaging studies [3]. However, due to its relatively low incidence, lack of awareness among healthcare professionals, and slow, often asymptomatic progression, the condition is frequently underdiagnosed or misdiagnosed. Consequently, it is most often discovered incidentally, for instance, during bronchoscopy or autopsy [4].

Although some patients remain asymptomatic for long periods, disease progression can lead to significant respiratory impairment. Chronic symptoms such as persistent cough and exertional dyspnea are the most commonly reported, alongside varying degrees of airflow obstruction ranging from mild to severe. In advanced cases, nodule growth can cause critical narrowing of the tracheal lumen, with reports of hemoptysis, progressive dyspnea [5], and difficulty with orotracheal intubation, highlighting the potential clinical risk of the disease, despite its traditional classification as benign [6].

Furthermore, the impacts of TO are not limited to mechanical airway obstruction. Under physiological conditions, the mucociliary system acts as a continuous defense mechanism, in which ciliated cells drive the transport of a mucus layer responsible for trapping and clearing inhaled particles and microorganisms [7]. The presence of nodules creates irregularities on the inner surface of the airways, deeply disrupting this mechanism. Alterations in airway architecture promote the emergence of turbulent airflow zones, regions where flow transitions from orderly to chaotic, thereby compromising mucociliary transport and, consequently, favoring the accumulation of secretions. These stagnation zones become environments conducive to bacterial proliferation. Consequently, recurrent pulmonary infections are among the most frequent and debilitating manifestations of the disease, compounding

the previously described obstructive condition [2].

Analyzing airflow behavior in human airways, particularly in the context of pathological conditions, presents significant limitations. These limitations stem primarily from the infeasibility of obtaining direct, non-invasive measurements of aerodynamic parameters in real-world cases. In this context, computational modeling based on Computational Fluid Dynamics (CFD) has established itself as a robust approach for studying the respiratory system, especially when integrated with high-resolution imaging; such imaging enables the three-dimensional reconstruction of a patient's airways and facilitates the derivation of patient-specific solutions [8]. CFD has already been applied to analyze various respiratory tract conditions, such as tracheal and subglottic stenosis, allowing for the identification of pressure drop regions, energy loss, and work of breathing [9] [10].

In addition to the aforementioned advances, it is worth noting that CFD enables the determination of quantitative parameters relevant to airflow characterization, such as pressure, velocity, flow resistance, and energy [11], thereby significantly enhancing the understanding of airway fluid dynamics. However, despite this potential, literature reviews indicate that the application of this method to studies on tracheobronchopathia osteochondroplastica remains limited. Given the disease's structural variability and the distinct impacts it has on respiratory dynamics, conducting such studies becomes necessary.

Hence, the present work proposes a comparative case study based on three-dimensional simulation to analyze the fluid dynamics of airflow in a patient affected by tracheobronchopathia osteochondroplastica (TO). The geometry was reconstructed from computed tomography (CT) images, allowing for the 3D reconstruction of the tracheobronchial structures affected by the disease. The fluid-dynamic parameters obtained from the simulation were compared with those of a control patient with no apparent tracheal abnormalities. Thus, the study aims to highlight the alterations associated with the pathology, contributing to an understanding of the underlying fluid-dynamic mechanisms in TO.

II. Materials and Methods

A. Acquisition and Preparation of the Base Geometry

This is an observational, retrospective, and non-interventional study approved by the Research Ethics Committee under CAAE No. 88886025.80000.0068. The research relied on secondary access to previously acquired computed tomography images from the InCor database, aiming to reconstruct and perform a computational morphological analysis of the central airways. There was no recruitment of volunteers, clinical intervention, or storage of biological material. In compliance with Resolution No. 466/2012 of the National Health Council (CNS) and Law No. 14.874/2024, the Research Ethics Committee granted a waiver of the Informed Consent Form (with no requirement for review by CONEP). The authors ensured anonymized processing of DICOM data and strict confidentiality of patient information. Two individuals were selected for analysis: one diagnosed with tracheobronchopathia osteochondroplastica (TO) and a control subject without tracheal disease (both male, aged 57 and 64 years, respectively). Images of the selected patients were acquired at end-inspiration with a slice thickness of 1.0 mm to maintain resolution standardization for the subsequent computational simulation. The mean diameter and length of the observed tracheas were approximately 14.14 mm and 96.48 mm for the patient with TO, and 14.51 mm and 99.64 mm for the control subject; these measurements were obtained using the "Extract Centerline" function available in the Vascular Modeling Toolkit (VMTK) module [12] of 3D Slicer (version 5.10.0). Case selection aimed for compatibility of characteristics whenever possible to minimize comparison bias.

The DICOM files allowed for the visualization and post-processing of the CT images. Based on the various cross-sections obtained from the CT scans, the anatomical region of interest was reconstructed in three dimensions. Airway segmentation was performed using the "Segment Editor" module of the 3D Slicer [13]. Initially, the primary regions of interest were defined by manually placing seed points representing the tracheal lumen and adjacent structures, such as the lungs and other organs. Next, the Grow from Seeds algorithm [14] was applied; this performs semi-automatic segmentation based on the expansion of previously defined regions, taking into account intensity similarity between voxels.

Following the initial segmentation, manual adjustments were made to refine the boundaries and correct for inaccuracies, ensuring greater anatomical fidelity of the three-dimensional tracheal model, as shown in Fig. 1.

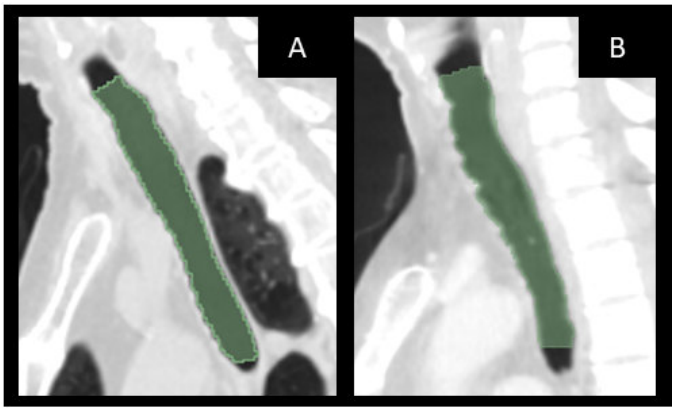


Fig. 1. Sagittal view of the segmented trachea of the control patient (A) and the patient affected by the pathology (B), using the "Grow from Seeds" segmentation method. The reconstructed trachea is highlighted in green.

For the purpose of methodological standardization, the tracheal segment under study was anatomically defined as the section extending from the cricoid cartilage (superior limit) to the tracheal carina (inferior limit) [15], encompassing the

entire length of the trachea. Identification of these anatomical landmarks was facilitated by multiplanar reconstruction (MPR), allowing for simultaneous visualization in the axial, coronal, and sagittal planes of the image series. The cricoid cartilage was identified as the first complete cartilaginous ring situated below the larynx, being most easily recognized in axial and sagittal views. The carina was identified as the point where the trachea bifurcates into the right and left main bronchi, best visualized in the coronal and axial planes. The inferior limit of the section was defined as the level immediately proximal to the start of the bifurcation, ensuring that the division between the left and right bronchi did not interfere with the computational simulation, in accordance with the study's objective. Subsequently, the regions were delineated using the "Point List" tool, and any areas situated above or below the boundary points were removed using the "Crop" tool. Finally, the resulting model could be intuitively exported from the "Segmentations" module in .STL format.

Meshmixer software was used to flatten the tracheal edges of the .STL files. Following, Fluent Flow software (Fluent with Fluent Meshing; ANSYS Fluent®, Release 2026 R1) was used to perform the computational simulation steps. Two modules were used: one for the pathological trachea and another for the control trachea.

B. 3D Geometric Modeling

The Geometry component of the ANSYS software was used to create the three-dimensional models. At this stage, imported geometries were inspected for structural defects and subsequently converted to volumetric solid bodies—a fundamental step to enable mesh generation and subsequent flow simulation. The resulting geometries for both the control and pathological models are shown in Fig. 2.

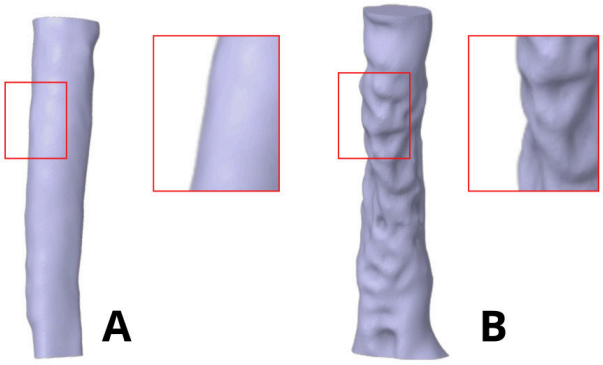


Fig. 2. Three-dimensional reconstruction of the trachea. The inset shows an enlarged view of the anterior tracheal luminal surface indicated by the red box. While the trachea of the control patient (A) exhibits a relatively smooth surface, the trachea of the patient with tracheobronchopathia osteochondroplastica (B) presents a more irregular luminal appearance due to the presence of multiple nodular protrusions.

C. Computational Mesh Generation

Three computational meshes were created using the ANSYS Meshing component, and the most refined one was used for the simulations. This mesh features a local refinement of $2.5 \times 10^{-4}\, m$ to ensure greater accuracy in capturing local fluid-dynamic phenomena. The mesh used is considered optimal, as it yields a difference of less than 2% in the results obtained [16]. A comparison of the three meshes for both tracheal models is presented in Table I.

TABLE I
NUMBER OF ELEMENTS AND NODES FOR THE DIFFERENT COMPUTATIONAL MESHES

Mesh	TO		Control	
	No. Elements	**No. Nodes**	**No. Elements**	**No. Nodes**
1^{st} Mesh	202987	538430	151424	437445
2^{nd} Mesh	360967	901098	270812	691975
3^{rd} Mesh	621328	1524478	467651	1145293

To obtain better resolution of the hydrodynamic boundary layer, which is the region where the greatest velocity and shear stress gradients occur at the wall (WSS), inflation layers were applied to the tracheal walls. The inflation configuration implemented the "Smooth Transition" method, using 7 prismatic layers, a growth rate of 1.12, and a transition ratio of 0.272. This allowed for a smooth transition between the prismatic mesh near the wall and the tetrahedral mesh at the center of the lumen, ensuring y+ values of approximately 1. The mesh design was based on rigorous boundary layer control and element quality metrics (maintaining appropriate orthogonality and asymmetry criteria). This approach ensured the stability and convergence of the numerical residuals, accurately capturing variations in WSS, pressure, and local turbulence, without requiring mesh refinement levels that would compromise computational feasibility. The resulting meshes can be seen in Fig. 3.

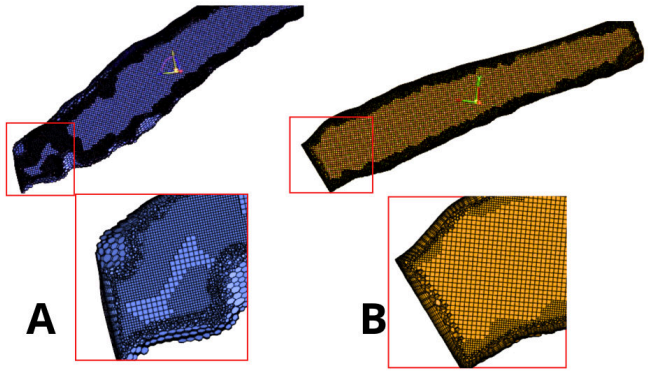


Fig. 3. The 3D models of the (A) control and (B) pathological tracheas were discretized into finite elements for CFD analysis. The red close-ups show the mesh refinement, featuring inflation layers at the tracheal wall to capture boundary layer effects, and larger elements in the center to optimize processing.

D. Setup Phase

The SST k-ω turbulence model was adopted, assuming air with a density of $1.22\,\mathrm{kg/m^3}$ and a viscosity of $1.78 \times 10^{-5}\,\mathrm{kg/(m \cdot s)}$. The inlet/outlet velocity was set at $1.5\,\mathrm{m/s}$ (peak of resting breathing [17]), with an outlet pressure of 0 Pa. The constant velocity standardized the models, isolating the impact of the TO nodule geometry on airflow resistance. The simulations, run to convergence, were performed by simulating an inspiration followed by an expiration, calculated separately under steady-state conditions.

E. Resistance Calculation

To quantitatively assess the anatomical impact on airflow, the resistance of the trachea was calculated. Tracheal resistance (R) aimed at quantifying the opposition to airflow along its length and was defined as

$$R = \frac{\Delta P}{Q}, \tag{1}$$

where R is the resistance of the trachea, ΔP is the pressure drop calculated as the outlet pressure (P_{outlet}) minus the inlet pressure (P_{inlet}), and Q is the volumetric flow rate. This formulation provides a quantitative measure of the pressure required to sustain a given airflow and enables comparison of the aerodynamic consequences of different tracheal geometries through their associated pressure losses.

III. Results

A. Average Speed

The airflow dynamics in the control and pathological tracheas during inspiration and expiration are shown in Fig. 4. Velocity contours illustrate the impact of structural changes caused by TO on the flow profile.

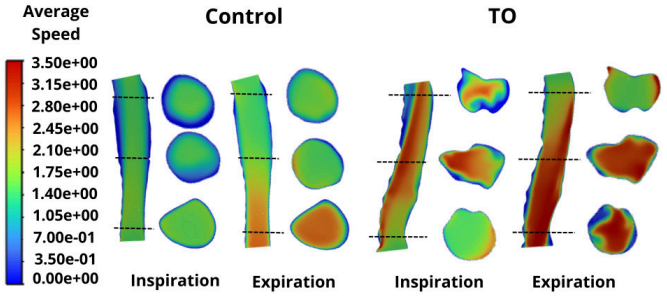


Fig. 4. Airflow velocity contours. Color scale ranging from blue (low) to red (high). Control trachea (right) and pathological trachea (right), showing higher velocities at the narrowing of the pathological trachea during expiration.

B. Total Pressure

The total pressure distribution (Fig. 5) made it possible to determine the pressure drop along the trachea and calculate the air resistance.

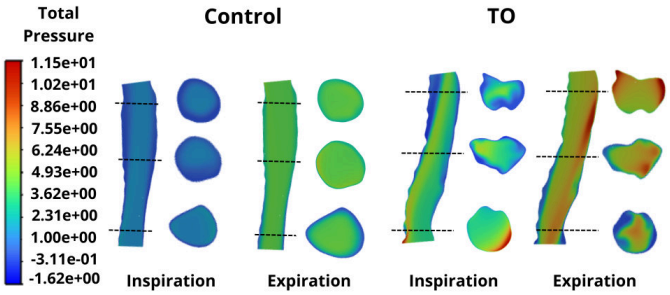


Fig. 5. Total pressure. Color scale ranging from blue (low) to red (high). Control trachea (right) and pathological trachea (right); the latter exhibits greater pressure loss during expiration and higher gradients in the asymmetric regions.

C. Air Flow, Resistance and Turbulent Kinetic Energy

Table II summarizes the airflow and airway resistance in the two tracheal models during inspiration and expiration. The pathological trachea exhibited higher resistance when flow was normalized. Furthermore, the distribution of Turbulent Kinetic Energy (TKE, Fig. 6) indicates the disruption of streamlines caused by the obstructions, generating turbulence peaks downstream of the nodules.

TABLE II
AIRFLOW AND AIR RESISTANCE VALUES FOR THE CONTROL AND PATHOLOGICAL TRACHEA MODELS

Control		
Parameter	***Inspiration***	***Expiration***
Airflow (m^3/s)	2.4×10^{-2}	2.4×10^{-2}
Resistance ($Pa{\cdot}s/m^3$)	36.9	58.6
Pathological		
Airflow (m^3/s)	2.8×10^{-2}	2.8×10^{-2}
Resistance ($Pa{\cdot}s/m^3$)	108.9	148.4

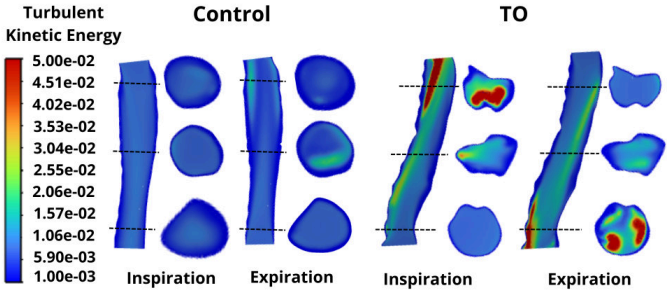


Fig. 6. Turbulent Kinetic Energy (TKE). Color scale ranging from blue (low) to red (high). Control trachea (right) and pathological trachea (right), with the latter exhibiting higher TKE levels and more unstable/turbulent flow especially during expiration.

D. Wall Shear Stress

Fig. 7 presents the WSS data, which may indirectly indicate the potential for inflammation and damage to the tracheal epithelium, especially in more severe cases of the disease.

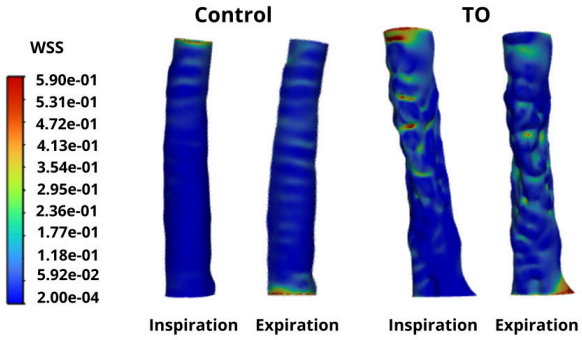


Fig. 7. Wall Shear Stress (WSS). Color scale ranging from blue (low) to red (high). Control trachea (right) and pathological trachea (right); the latter exhibits higher values and a less uniform distribution, especially during expiration.

IV. Discussion

This study aimed to compare the results of three-dimensional numerical simulations based on computed tomography scans of a patient with tracheobronchopathia osteochondroplastica (TO) and a control patient. Analysis of the results suggests possible biomechanical mechanisms by which TO-induced deformities may affect respiratory airflow. In this comparative case study, multiple bony and cartilaginous nodules projecting into the tracheal lumen were observed in the TO model. Such asymmetries and irregularities appeared to result in fluid-dynamic disturbances, contrasting with the more regular behavior of the control model.

An analysis of the data in Fig. 4 suggests that the fluid-dynamic behavior of the airflow is affected by the nodules in the patient with TO. In the control trachea, air velocity was distributed symmetrically and regularly, indicating a stable and predictable profile; velocity increased smoothly toward the center of the trachea, without deviations or abrupt decelerations. In contrast, regarding the pathological trachea, the asymmetries caused velocity peaks—indicated by the red coloring—in regions where the lumen narrows. Due to the nodules, the air is forced through various local constrictions. Given the constant volumetric flow rate, the reduction in lumen size compels the air to accelerate to satisfy the principle of mass conservation [18]. This increase in velocity near the nodule surfaces can generate an increase in viscous friction forces. From a mechanical standpoint, this rise in forces can lead to a marked increase in wall shear stress (WSS) values on the tracheal walls. This effect, particularly in cases of more severe anatomical obstruction, could contribute to the occurrence of microtrauma in local epithelial cells [19].

The data presented in Table 2 show that anatomical changes in the trachea could affect airflow resistance. Numerically, the resistance value for the pathological trachea was higher, but a larger cohort would be required for a more comprehensive analysis.

The airflow colliding with the nodules can influence the development of turbulent stresses and instabilities. The data obtained and the contours presented in Fig. 6 suggest that, while the control trachea maintained lower TKE values—indicative of stable flow—the pathological trachea exhibited localized peaks downstream of the osteochondroplastic projections. The obstructions caused by the nodules disrupt the organization of streamlines, converting kinetic energy into shear vortices and recirculation zones. It has been established [20] that regions with high TKE levels undergo significant mechanical energy dissipation due to turbulent and viscous friction. This chaotic fluid behavior helps explain the pressure drop and increased airway resistance observed in the pathological model. From a clinical perspective, regions with high TKE concentrations may represent sites of altered particle residence time, potentially influencing bioaerosol deposition patterns and local clearance mechanisms [20].

Analysis of the WSS results (presented in Fig. 7) revealed that the presence of pathological nodules can cause the airflow to exert drag forces on the tracheal wall. Under both inspiratory and expiratory conditions, the mean WSS value on the tracheal wall of the patient with TO was higher than that observed in the control subject. Previous studies [20] [21] have indicated that vortex formation and flow stagnation zones result in a heterogeneous WSS distribution. Although the results obtained do not allow for a direct diagnosis of clinical outcomes, such flow disturbances could theoretically impair continuous mucociliary clearance, creating local microenvironments that favor bacterial retention [20].

This study presented several limitations that should be considered. First, the analysis was based on a limited dataset, comprising only one computed tomography scan with pathological findings and one control subject. This limitation arose because tracheobronchopathia osteochondroplastica is a rare condition, making it difficult to obtain scans and data. Consequently, the generalizability of the results was restricted, and the condition calls for careful consideration in light of inter-individual anatomical variability. Thus, the obtained values for airway resistance and shear stress represent case-specific simulation results and cannot be used for population-level estimates. Second, the chosen mesh provided a good balance between computational cost and numerical accuracy; however, further refinement could enhance the spatial resolution of local flow structures and micro-scale fluid dynamics phenomena. Thirdly, the CFD simulations assumed steady-state flow conditions with simplified boundary parameters, thereby disregarding the transient dynamics of the respiratory cycle and temporal variations in airflow. Although the temporal analysis was limited, it was possible to map the spatial characteristics of airflow velocity contours with reduced complexity. Finally, the tracheal walls were modeled as rigid boundaries, without accounting for tissue dynamic mechanics or elastic deformation that could be addressed through fluid-structure interaction (FSI). The

anatomical complexity of the trachea, comprising connective tissue, cartilaginous rings, and a posterior muscular membrane, and the scarcity of mechanical data on these tissues in the literature (usually with anisotropic properties) make it challenging to accurately characterize their mechanical behavior and, consequently, to develop computational models capable of faithfully reproducing their structural and biomechanical response. Future work will focus on refining this methodology to overcome current limitations and achieve more robust numerical simulations. The correlation of these CFD metrics with clinical tests and data could potentially assist medical teams in making therapeutic decisions for patients with TO.

V. Conclusion

This study conducted an in silico assessment of how geometric deformities associated with tracheobronchopathia osteochondroplastica (TO) can induce local mechanical disturbances in respiratory airflow. The computational results suggest that osteochondroplastic nodules within the tracheal lumen could alter local aerodynamics; as air passes through these localized constrictions, structural asymmetries trigger an increase in viscous friction forces, manifesting as increased wall shear stress. The observed disruption in streamline patterns appeared to generate localized peaks of turbulent kinetic energy and recirculation zones, contributing to greater mechanical energy dissipation. In summary, mapping these physical parameters provides a plausible biophysical basis that may help explain the aerodynamic underpinnings of TO's clinical symptoms—such as progressive dyspnea, potential impairment of mucociliary clearance, and susceptibility to recurrent respiratory infections—thereby warranting further validation in larger clinical cohorts.

VI. Acknowledgments

Generative AI was used to assist with text editing. The scientific content, figures, analyzes, and interpretations were created, reviewed, and verified by the authors themselves, who assume full responsibility for the content of this article.

References

[1] T. Luo, H. Zhou, and J. Meng, "Clinical characteristics of tracheobronchopathia osteochondroplastica", Respir. Care, vol. 64, no. 2, pp. 196–200, Feb. 2019.

[2] V. Leske, R. Lazor, D. Coetmeur, B. Crestani, G. Chatté, and J. F. Cordier, "Tracheobronchopathia osteochondroplastica: a study of 41 patients", Medicine (Baltimore), vol. 80, no. 6, pp. 378–390, Nov. 2001.

[3] C. A. García, S. Sangiovanni, V. Zúñiga-Restrepo, E. I. Morales, L. F. Sua, and L. Fernández-Trujillo, "Tracheobronchopathia osteochondroplastica—clinical, radiological, and endoscopic correlation: case series and literature review", Journal of Investigative Medicine High Impact Case Reports, vol. 8, pp. 1-10, Jan.-Dec 2020, doi: 10.1177/2324709620921609.

[4] M. G. M. Silveira, M. V. C. O. Castellano, C. E. Fuzi, E. N. A. M. Coletta, and G. N. Spinosa, "Tracheobronchopathia osteochondroplastica", J. Bras. Pneumol., vol. 43, no. 2, pp. 151–153, Mar.-Apr. 2017, doi: 10.1590/S1806-37562016000000143.

[5] A. Dutau et al. (Groupe d'Endoscopie Thoracique et Interventionnelle Francophone – GETIF), "Tracheobronchopathia osteochondroplastica: clinical, bronchoscopic, and comorbid features in a case series", BMC Pulm. Med., vol. 22, no. 1, Nov. 2022, Art. no. 423, doi: 10.1186/s12890-022-02225-2.

[6] D. Coëtmeur, G. Bovyn, P. Leroux, and M. Niel-Duriez, "Tracheobronchopathia osteochondroplastica presenting at the time of a difficult intubation", Respir.Med., vol. 91, no. 8, pp. 496–498, Sep. 1997.

[7] J. E. Hall, Tratado de fisiologia médica, 13th ed. Rio de Janeiro: Elsevier, 2017.

[8] D. F. Legendre et al., "Computational fluid dynamics for assessment of airflow patterns in the airway tree", in Proc. CLAIB and CBEB 2022, Florianópolis, Brazil, pp. 210–222.

[9] A. J. Bates et al., "The effects of curvature and constriction on airflow and energy loss in pathological tracheas", Respiratory Physiology & Neurobiology, vol. 234, pp. 69–78, Dec. 2016, doi: 10.1016/j.resp.2016.09.002.

[10] M. M. Yang et al., "Subglottic Stenosis Position Affects Work of Breathing", The Laryngoscope, vol. 131, no. 4, pp. E1220–E1226, Apr. 2021, doi: 10.1002/lary.29169.

[11] M. Brouns et al., "Tracheal stenosis: a flow dynamics study", Journal of Applied Physiology, vol. 102, no. 3, pp. 1178–1184, Mar. 2007, doi: 10.1152/japplphysiol.01063.2006.

[12] VMTK Developers, "Extract Centerline", SlicerExtension-VMTK, GitHub. [Online]. Available: https://github.com/vmtk/SlicerExtension-VMTK/tree/master/ExtractCenterline. Accessed: Jun. 27, 2026.

[13] C. Pinter, A. Lasso, and G. Fichtinger, "Polymorph segmentation representation for medical image computing", Computer Methods and Programs in Biomedicine, vol. 171, pp. 19–26, 2019, doi: 10.1016/j.cmpb.2019.02.011.

[14] 3D Slicer Community, "Segment Editor", 3D Slicer Documentation. [Online]. Available: https://slicer.readthedocs.io/en/latest/user_guide/modules/segmenteditor.html. Accessed: Jun. 27, 2026.

[15] J. Samir et al., "The Medical Segmentation Decathlon", arXiv/PMC, 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5900092/. Accessed: Jun. 27, 2026.

[16] O. Aycan, A. Topuz and L. Kadem, "Evaluating uncertainties in CFD simulations of patient-specific aorta models using Grid Convergence Index method", Mechanics Research Communications, vol. 133, p. 104188, Nov. 2023, doi: 10.1016/j.mechrescom.2023.104188.

[17] X. XU et al., "Investigation of inhalation and exhalation flow pattern in a realistic human upper airway model by PIV experiments and CFD simulations" Biomechanics and Modeling in Mechanobiology, vol. 19, no. 5, p. 1679–1695, Out. 2020.

[18] T. Cheng et al., "Investigating the effects of laryngotracheal stenosis on upper airway aerodynamics", The Laryngoscope, vol. 128, no. 4, Apr. 2018.

[19] S. Dharmadhikari et al., "Pathologic Wall Shear Stress in the Trachea Decreases Epithelial Differentiation and Increases Fibrosis", American Journal of Respiratory Cell and Molecular Biology, vol. 74, no. 6, pp. 780–791, Jun. 2026, doi: 10.1093/ajrcmb/aanaf037.

[20] Z. Zhang, C. Kleinstreuer, "Airflow structures and nano-particle deposition in a human upper airway model", Journal of Computational Physics, vol. 198, no. 1, p. 178–210, 20 Jul. 2004.

[21] C. L. Lin et al., "Characteristics of the turbulent laryngeal jet and its effect on airflow in the human intra-thoracic airways", Respiratory Physiology and Neurobiology, vol. 157, no. 2–3, p. 295–309, Aug. 2007.

Upper Limb Kinematic Analysis Using Inertial Sensors and Temporal Normalization

Ligyane Hellen Tobias Lopes, Lucas Matsuno da Costa, Rui Gonçalves de Oliveira Júnior,
João Vitor Menegon Xavier, Marco Antonio Assfalk de Oliveira*, and Jalusa Andreia Storch Díaz†
*Federal University of Goiás (UFG)
Goiânia, Brazil
Email: {ligyanelopes, lucas.matsuno, ruioliveira, joaovitor13}@discente.ufg.br, assfalk@ufg.br
†Federal University of Catalão (UFCAT)
Catalão, Brazil
Email: jastorch@ufcat.edu.br

***Abstract*—Upper limb (UL) motor rehabilitation following a Stroke demands technological approaches that mitigate the physical burden on therapists and expand clinical care. This work focuses on the computational modeling and simulation of functional movements of human ULs for application in assistive robotic manipulators. The objective is to develop a robust mathematical-computational model to represent these patterns from real kinematic data captured by Inertial Measurement Units (IMUs) on the right upper limb during Activities of Daily Living (daily activities). The trials were conducted by two healthy volunteers through multiple structured collections in the tasks of waving, combing hair, and moving a cup. The data in quaternions, which avoid gimbal lock, were converted to Euler angles (XYZ sequence). The model decomposes the movement into sequences of intermediate postures, structuring three-dimensional equations that simulate the dynamic interaction of the segments and incorporate inter-individual anthropometric variability. The results demonstrated that the temporal stochasticity of human movement introduces statistical distortions. The application of Temporal Normalization by Movement Cycle (TNMC) mitigated phase artifacts; however, the subsequent temporal average attenuated high-frequency kinematic components, generating trajectories with ergonomic constraints for robotic control. This computational simulation is critical both for the predictive trajectory programming of the robotic arm and for the foundation of customized therapeutic protocols, enabling physical and parametric adjustments adaptable to each patient.**

***Index Terms*—Computational modeling, Robotic physical therapy, Upper limb kinematics, Assistive robotics, Inertial sensors.**

I. Introduction

A Stroke is clinically defined as a sudden onset neurological deficit resulting from a focal brain lesion. Classified mainly as ischemic or hemorrhagic, it can result from different pathophysiological mechanisms that frequently produce similar clinical manifestations, such as sudden and intense headache, dizziness, mental confusion, changes in speech and vision, loss of balance, and numbness of the limbs. It is estimated that approximately 15 million people are affected by stroke annually worldwide, constituting one of the main causes of death and functional disability [1]. Among the most frequent sequelae, the impairment of the upper limbs (UL) stands out, observed in about 80% of patients [2]. This manifests as reduced cutaneous sensitivity, muscle weakness, and limitation of movement in the shoulder, elbow, wrist, hand, and fingers, compromising activities of daily living such as combing hair, brushing teeth, dressing, and participating in sports activities [3]. During the motor rehabilitation of post-stroke patients, treatment is initially based on manual therapeutic techniques applied by the physical therapist, aiming to restore joint mobility and muscle activation. However, this individualized care requires continuous physical effort from the professional, which can compromise productivity and increase occupational burden. In this context, assistive robotic systems emerge as a promising alternative, as they allow for increased intensity, repeatability, and control of therapeutic sessions, while reducing the physical demand imposed on the physical therapist [4], [5]. A robotic arm for rehabilitation can be defined as a computer-controlled electromechanical device developed to assist the functional recovery of the UL [11]. Its design considers the mechanical dynamics of the system, including gravitational and inertial effects, allowing the execution of passive, assistive, active, or resistive therapeutic movements, depending on the control strategy employed [8], [10]. Specifically, this work addresses a fundamental step for the clinical application of these robotic manipulators: the mathematical-computational modeling of the functional movements of the upper limbs. This model is essential both for planning and programming biomechanically appropriate trajectories, favoring kinematic execution and the induction of new neuromotor patterns [6], [7], and for the joint development, by engineers and physical therapists, of personalized therapeutic protocols, including partial, assisted, resistive, and inductive exercises. The representation of functional movements is obtained from datasets acquired by capture systems based on Inertial Measurement Units (IMUs). The movements are represented by temporal sequences of intermediate postures of the upper limb, described by the three-dimensional positions and orientations of specific anatomical segments. The mathematical model consists of a set of parameters and dynamic equations capable of approximating real human movement, whose orientation is initially represented by quaternions, eliminating the occurrence of the phenomenon known as gimbal lock, and subsequently con-

verted to Euler angles to facilitate biomechanical interpretation and the development of control algorithms [12]. Considering that the database is composed of individuals with different anthropometric characteristics, the proposed model incorporates both inter-individual variability and the temporal stochasticity intrinsic to human movement. This characteristic makes it possible to adapt the robotic manipulator to the specific needs of each patient, both through mechanical configuration and the parametric adjustment of therapeutic trajectories according to the functional deficits resulting from the stroke. However, the statistical treatment of these signals presents important challenges due to the high temporal variability of functional tasks. Although the application of Temporal Normalization by Movement Cycle (TNMC) significantly reduces phase artifacts and dispersion between trials — with reductions of 73.0% to 83.2% in the maximum standard deviation (SD_{max}), as presented in Table I —, the resulting temporal average tends to smooth high-frequency kinematic components, producing excessively smoothed trajectories that may reduce the biomechanical fidelity of the movements and, consequently, must be considered during the development of the robotic control algorithm.Given this scenario, the general objective of this study is to develop and validate a mathematical-computational model capable of representing functional kinematic patterns of the upper limbs obtained through IMUs, quantifying the impact of the temporal normalization technique on discrete biomechanical trajectories [12] and providing inputs for the development of robust predictive control strategies applied to robotic manipulators destined for post-stroke motor rehabilitation.

II. Materials and Methods

The methodology proposed in this work comprises three fundamental stages: (i) experimental acquisition of three-dimensional kinematic data, (ii) signal processing and coordinate conversion based on quaternions and Euler angles, and (iii) the structuring of the mathematical-computational model coupled with the temporal normalization technique.

A. Kinematic Data Acquisition and Experimental Protocol

The capture of functional movements was performed using Inertial Measurement Units (IMUs) integrated into a modular data acquisition system. The final experimental setup utilized inertial sensors strategically fixed to the right upper limb of two healthy volunteers, as illustrated in Figure 1. The inclusion of distinct subjects allowed the evaluation of the mathematical-computational model under inter-individual kinematic variability and different motor strategies. The anatomical distribution of the sensors was structured as follows: The first sensor was positioned on the lateral aspect of the arm (the transition region between the lateral deltoid and triceps brachii muscles), minimizing mechanical interference during elbow flexion. The second sensor was fixed on the dorsal (posterior) aspect of the wrist, adjacent to the distal region of the forearm, over the tendons of the extensor muscles to capture segmental rotations and joint deviations. The sensors were connected



Fig. 1. Experimental setup and anatomical placement of the Inertial Measurement Units (IMUs): (A) arm sensor (lateral side); (B) wrist sensor (dorsal side); and (C) data acquisition and transmission module.

by wiring to a microcontrolled centralizer and transmitter module positioned on the test bench, responsible for power supply via a lithium polymer (LiPo) battery and synchronous data transmission via a wireless protocol. The instrumentation protocol was developed incrementally. Initially, a pilot study was conducted using a single sensor to validate the calibration routines for gyroscopes and accelerometers. After

ensuring the stability of the communication channel and the absence of packet loss, the number of devices was expanded to the segmental multi-sensor array, enabling synchronous monitoring and three-dimensional reconstruction of the kinematic chain. The definitive experimental protocol consisted of multiple structured collections during the execution of three standardized and task-oriented Activities of Daily Living (daily activities): waving, combing hair, and moving a cup. Three-dimensional orientation data were sampled at a constant frequency and transmitted for post-processing.

B. Signal Processing and Rotation Kinematics

To mitigate mathematical singularity artifacts known as gimbal lock, the raw angular orientation data extracted from the IMUs were initially processed and stored as unit quaternions q. A quaternion is defined by:

$$q = q_0 + q_1 i + q_2 j + q_3 k \tag{1}$$

Where q_0 represents the scalar component and q_1, q_2, q_3 correspond to the three-dimensional vector in the complex space, satisfying the unit hypersphere condition $||\mathbf{q}|| = 1$. Subsequently, aiming at anatomical visualization, biomechanical interpretation, and the structuring of the control equations of the robotic manipulator, the quaternions were converted to Euler angles using the XYZ fixed rotation sequence (α, β, γ). The transformation matrix was applied to each discrete time interval to extract the joint angles corresponding to flexion/extension, abduction/adduction, and internal/external rotation of the shoulder and elbow.

C. Computational Modeling and Postural Decomposition

The three-dimensional modeling of the right upper limb was structured as an open kinematic chain containing 7 degrees of freedom (DoF), correlating the orthogonal Cartesian axes of robotics to the anatomical planes of clinical biomechanics. The shoulder joint complex was parameterized with 3 DoF acting as an ideal spherical joint [9]. This modeling allows isolated mapping of flexion/extension movements (sagittal plane, mediolateral axis), abduction/adduction (frontal plane, anteroposterior axis), and internal/external rotation (transverse plane, longitudinal axis). Additionally, the mathematical model processes shoulder circumduction, combining the kinematic restriction equations simultaneously in the three planes of movement to simulate complex and fluid spatial trajectories without geometric discontinuity [6], [7]. The elbow was modeled with 2 DoF, encompassing flexion/extension and the pronation/supination movement of the forearm (transverse plane, longitudinal axis). Finally, the wrist joint complex was structured with 2 additional DoF to ensure the ergonomic precision of the end effector, mapping flexion/extension movements (sagittal plane) and radial/ulnar deviations (frontal plane, anteroposterior axis). For the computational simulation, the continuous functional movement extracted from the IMUs was discretized and decomposed into a temporal sequence of intermediate static postures, defined by:

$$P(t) = \{\text{shoulder}(t), \text{elbow}(t), \text{wrist}(t)\} \tag{2}$$

Where each posture $P(t)$ represents the instantaneous angular arrangement of the anatomical segments in a discrete time interval. Inter-individual anthropometric variability was incorporated into the model through parameterized homogeneous transformation matrices. The length of the arm (L_{humerus}) and forearm (L_{radius}) segments of each volunteer was inserted as a translation vector in the forward kinematic model, allowing the exact calculation of the end effector (hand) position in the three-dimensional space. Finally, the dynamic equations that simulate the interaction of the segments were structured to map these postural sequences. This mapping allows the predictive algorithm of the robotic manipulator [8] to anticipate the spatial transitions necessary to reproduce the human motor pattern in a fluid and ergonomically safe manner.

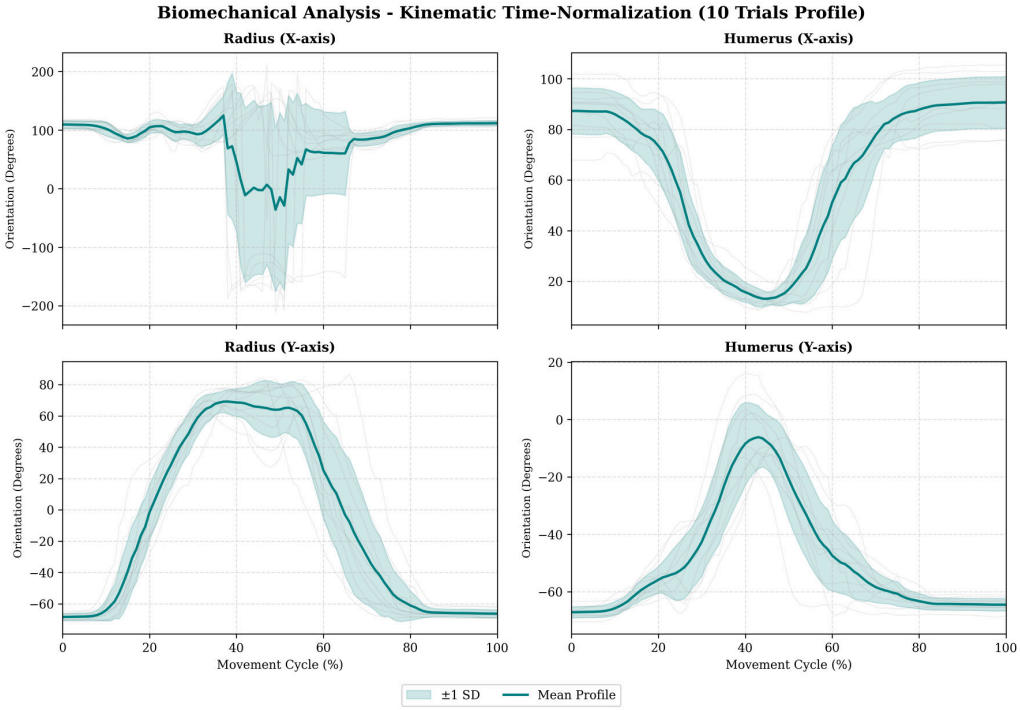
D. Stochasticity Treatment and Temporal Normalization

Due to the temporal stochasticity inherent to human gesture, variability was observed in the execution speed and total duration between movement cycles of the same daily activities, with durations oscillating from 5 s to 8 s for the "Combing Hair" task and extending up to 14 s in the "Moving a Cup" task. This variation generates distortions in aggregate statistical analyses, such as the artificial expansion of the standard deviation in regions of greater angular inclination. To solve this misalignment and mitigate phase artifacts, the Temporal Normalization by Movement Cycle (TNMC) technique was applied [12]. The algorithm performs a mathematical interpolation (via cubic splines) of the Euler angle vectors over time, re-sampling each execution to a fixed percentage window from 0% to 100% of the movement cycle [12]. After the phase synchronization generated by the TNMC, the temporal average of the trajectories was calculated for the extraction of the representative kinematic pattern of the task, serving as a predictive basis for the control algorithm of the assistive robotic manipulator.

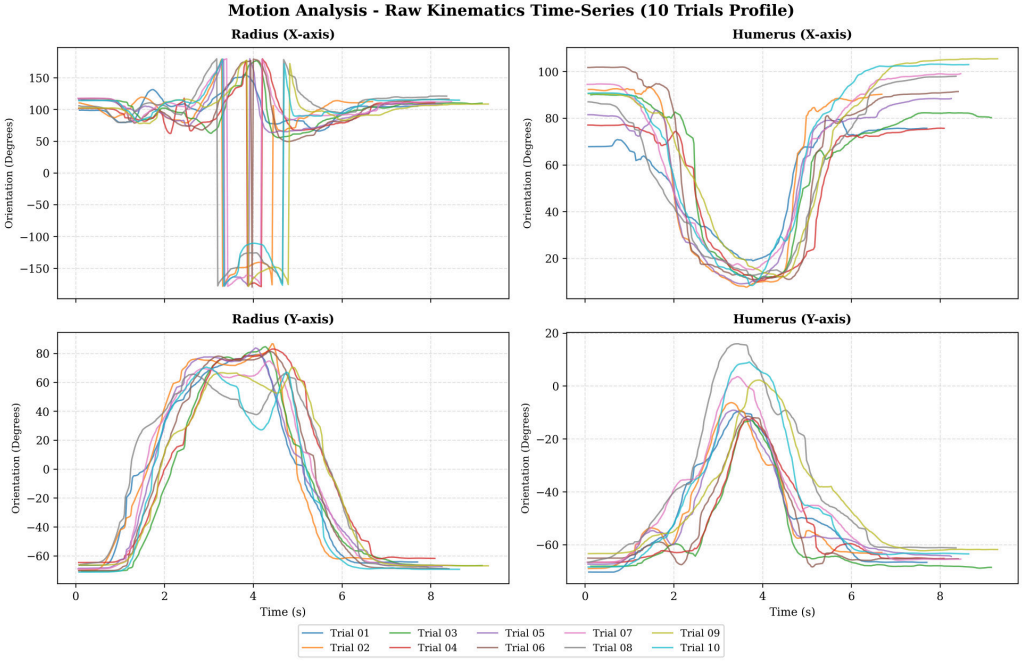
III. Results

The initial analysis of the raw kinematic curves in the absolute time domain (seconds) confirmed the stochastic nature and intrinsic trial-to-trial variability of human movement. As observed in the panels on the right side of Figures 2, 3, and 4 (absolute time axis), each functional task presented markedly heterogeneous total durations and velocity profiles: the "Combing Hair" task varied between 5 s and 8 s (Fig. 2); "Moving a Cup" extended up to 14 s (Fig. 3); and the "Waving" task was concentrated in the window of 6 s to 10 s (Fig. 4). This temporal desynchronization induced a critical statistical artifact in the raw curves: the standard deviation (± 1 SD) expanded artificially at the points of greatest angular inclination (derivative). This phenomenon occurs because the misalignment in time causes trials executed at different speeds to be compared at the same absolute instant. Consequently, the calculated statistical variance reflects the difference in speed between trials, masking the volunteer's true spatial (anatomical) movement variability. The application of the temporal normalization algorithm via Cubic Spline

interpolation mitigated these artifacts, homogenizing the trials into a relative scale from 0% to 100% of the Movement Cycle (101 points) [12]. Table I synthesizes the kinematic behavior, the quantitative impact of the variability attenuation, and the continuous residual validation metrics obtained after phase alignment. To rigorously verify that this variance attenuation stems purely from temporal synchronization rather than a destructive over-smoothing or loss of spatial fidelity, a point-to-point residual evaluation was performed against the ensemble mean profile (Table I). The continuous cross-correlation analysis yielded exceptionally robust coefficients, consistently exceeding r = 0.96 for the highly coordinated "Combing Hair" (r 0.986) and "Moving a Cup" (r 0.990) tasks. Even during the high-frequency oscillatory components of the "Waving" task, the minimum alignment correlation remained remarkably high (r = 0.894), mathematically demonstrating that individual spatial signatures were safely preserved. This fidelity is further supported by the controlled Root Mean Square Error (RMSE) values, which bounded strictly between 2.33° and 3.21° across all traditional functional movements, and only expanded up to 6.77° ± 0.85° in the radius X-axis during the rapid directional reversals of the waving gesture due to coordinate singularity quadrants. These numerical indicators provide definitive validation that the pipeline effectively filters out velocity fluctuations while safeguarding the true underlying kinematics. Building upon this methodological foundation, the overlay analysis between Subject 1 and Subject 2 successfully isolates and illuminates the differences in motor strategies during the multi-joint functional task. As demonstrated in the Humerus (X-axis) and Humerus (Y-axis) subplots, Subject 1 executed the movement cycle through a localized distal strategy, maintaining the humeral alignment strictly stabilized near 40° (X-axis) and -35° (Y-axis) with minor spatial dispersion. Conversely, Subject 2 exhibited a highly dynamic, coordinated multi-joint behavior, characterized by a steep humeral translation where the orientation dropped from 80° to 20° (X-axis) coupled with a smooth, bell-shaped parabolic trajectory in the Y-axis, which peaked at -12° near 43% of the movement cycle. Despite these disparate biomechanical strategies, the TNMC algorithm proved resilient in homogenizing the phase synchronization into a relative scale (0% to 100%). Crucially, specific morphological micro-structures of the biological signals were meticulously preserved rather than over-smoothed. In the Radius (Y-axis) subplot, the subtle double-peak flexion profile performed by Subject 2 (notably the inflection landmark at 20% of the cycle) remains perfectly distinct in the normalized mean curve. Furthermore, the Radius (X-axis) subplot provides a definitive mathematical validation against any premise of artificial data dampening. Around 40% to 50% of the movement cycle, Subject 2's mean profile exhibits an acute oscillation accompanied locally by a significant expansion of the standard deviation shroud (reaching ±100°). This transient artifact does not stem from algorithmic failure; rather, it reflects the mathematical wrapping (coordinate singularity) inherent to Euler angles when raw angular inputs transit rapidly through boundary quadrants. Because the TNMC aligns temporal phase without artificially modifying spatial boundary transitions, it preserves these underlying algebraic constraints of the sensor data. This fidelity guarantees that the post-processing pipeline isolates temporal velocity variations while safeguarding the genuine spatial kinematics and individual motor signatures necessary for high-fidelity physical therapy robotic control [10]. In addition to the dynamic coordinate transitions observed in the cup-transport task, the "Combing Hair" trials provided a distinct operational benchmark for the post-processing framework. As illustrated in the comparative timeline, while Subject 1 executed the classic dynamic grooming cycle characterized by pronounced joint excursions, Subject 2's dataset captured a highly stable, static baseline posture (locked near 90° for the humerus X-axis and 0° for the humerus Y-axis). This contrasting behavioral setup served as a rigorous mathematical control group within the experiment. Under these non-dynamic conditions, the TNMC algorithm proved mathematically robust, displaying zero drift or processing instabilities when mapping invariant time-series. The resulting standard deviation for Subject 2 remained virtually non-existent, confirming that the cubic spline interpolation framework operates safely without introducing artificial oscillations, high-frequency artifacts, or computational ripples when processing static or low-amplitude boundary states. Consequently, the pipeline satisfies the structural requirements for both active kinetic tracking and isometric joint monitoring in physical therapy protocols [4].

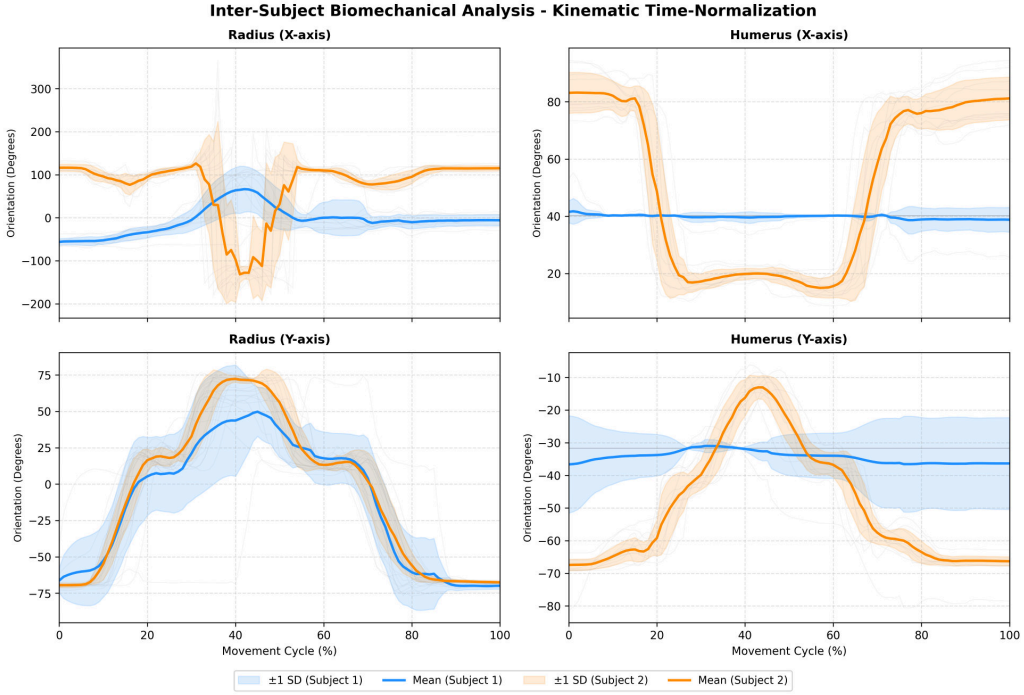


(a) Set A: Temporal Normalization

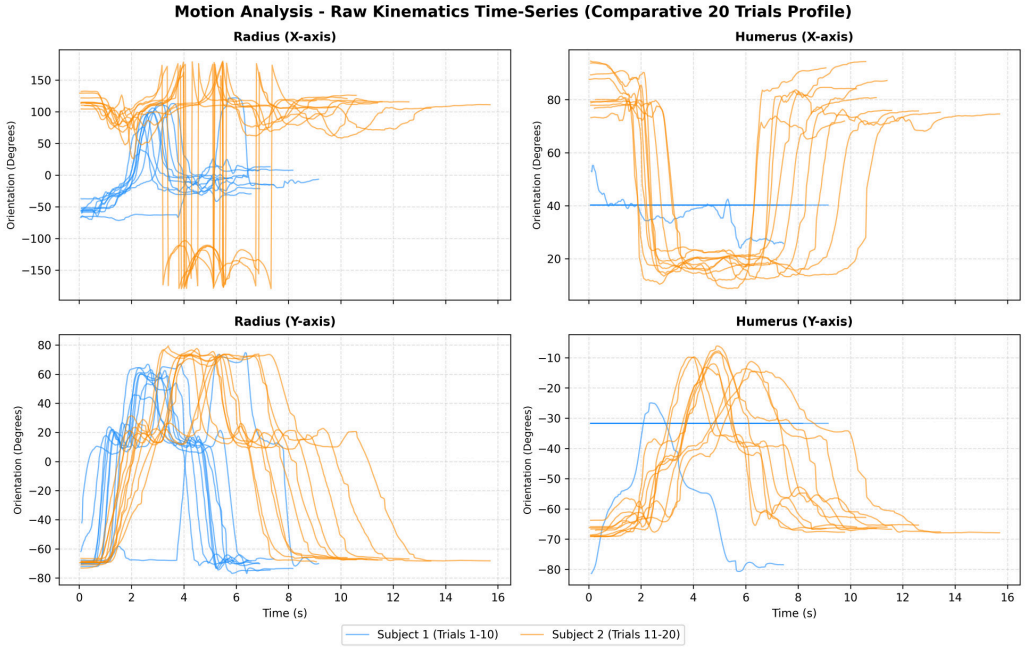


(b) Set B: Raw Trials

Fig. 2. Kinematic analysis of the "Combing Hair" functional task using inertial measurement units (IMUs) on the radius and humerus. Set A presents the mean curve surrounded by the standard deviation band (± 1 SD) after temporal normalization (0% to 100%). Set B displays the 10 original raw trials plotted against absolute time in seconds.

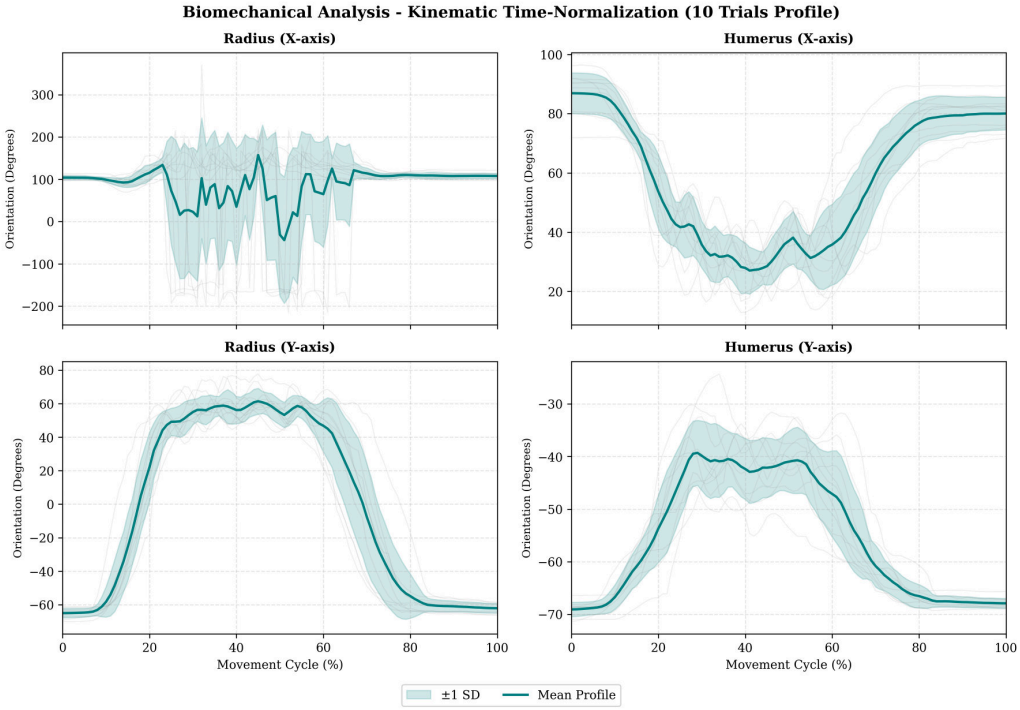


(a) Set A: Temporal Normalization

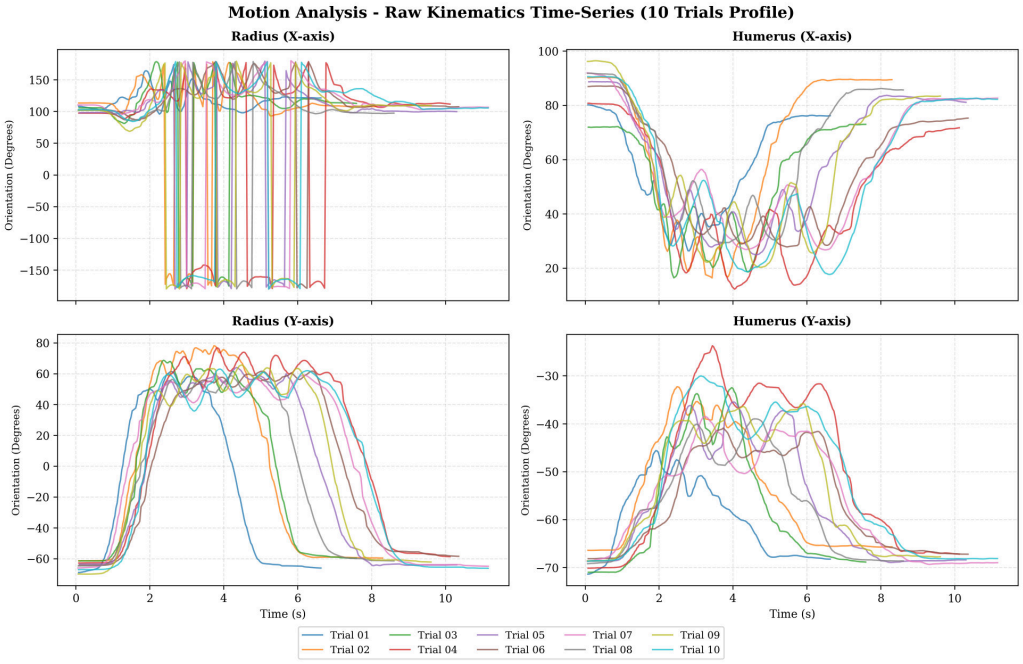


(b) Set B: Raw Trials

Fig. 3. Comparative inter-subject kinematic analysis between Subject 1 (blue) and Subject 2 (orange) after Temporal Normalization by Movement Cycle (TNMC). Solid bold lines represent the ensemble mean profiles, and the shaded regions indicate the respective standard deviation bounds (± 1 SD) across 10 trials per participant.



(a) Set A: Temporal Normalization



(b) Set B: Raw Trials

Fig. 4. Kinematic analysis of the "Waving" functional task using inertial measurement units (IMUs) on the radius and humerus. Set A presents the mean curve surrounded by the standard deviation band (± 1 SD) after temporal normalization (0% to 100%). Set B displays the 10 original raw trials plotted against absolute time in seconds.

TABLE I
MEAN KINEMATIC PARAMETERS, MAXIMUM STANDARD DEVIATION (SD_{max}) ATTENUATION, AND POINT-TO-POINT ENSEMBLE METHOD VALIDATION

Motor Task	**Segment / Axis**	SD_{raw} (°)	SD_{norm} (°)	**Reduction (%)**	**Ensemble RMSE (°)**	**Ensemble Corr. (r)**
Combing Hair	Humerus / X	±22.4	±4.1	81.7%	3.21 ± 0.20	0.986
	Radius / Y	±14.8	±3.6	75.7%	2.93 ± 0.20	0.991
Moving a Cup	Humerus / X	±19.1	±3.2	83.2%	2.59 ± 0.23	0.992
	Radius / Y	±12.3	±2.9	76.4%	2.33 ± 0.20	0.990
Waving	Radius / X	±35.6	±8.7	75.5%	6.77 ± 0.85	0.894
	Humerus / Y	±11.5	±3.1	73.0%	2.55 ± 0.24	0.969

IV. CONCLUSION

The temporal normalization approach using cubic spline interpolation to exactly 101 points proved to be an indispensable methodological tool for the scientific validation of continuous kinematic data obtained by Inertial Measurement Units (IMUs) [12]. The proposed protocol successfully eliminated point-to-point statistical interpretation errors caused by the intrinsic variable speed of human movement execution trial-to-trial. The quantitative attenuation greater than 73% in the maximum standard deviation, coupled with low continuous Root Mean Square Error (RMSE) thresholds and exceptionally high cross-correlation coefficients across all evaluated motor tasks, mathematically demonstrates that the post-processing residual variability strictly reflects the spatial dispersion and consistency of the volunteer's movement without inducing artificial signal degradation. In the high-frequency task ("Waving"), the algorithm preserved the oscillatory morphology of the signal, registering a reduction of up to 75.5% in the maximum standard deviation (SD_{max}) while maintaining robust morphological fidelity as itemized in Table I, which successfully avoided the phase cancellation phenomenon in the mean curve.

Thus, the use of Temporal Normalization by Movement Cycle (TNMC) confers reproducibility, continuous mathematical validation, and scientific robustness to the use of low-cost inertial sensors. This methodological consistency is corroborated by computer animations generated from the calculated mean, which visually confirm the fluid reproduction of the human motor pattern. This alignment provides a preliminary mathematical framework that potentially supports future development in clinical screening protocols, remote monitoring, and rehabilitation engineering [11]. However, further experimental validation involving physical integration with the robotic manipulator [8] and clinical trials with stroke patients [5] is strictly required to evaluate its safety and therapeutic efficacy. Although current analyses are restricted to the initial trial-to-trial methodological validation of two healthy volunteers, immediate steps for future work include expanding the kinematic database to a larger cohort of participants to evaluate inter-individual performance. Additionally, a transition to the hospital environment is planned with the application of the model in randomized controlled clinical trials, aiming to evaluate the efficacy of the predictive control of the robotic manipulator [10] in the neuromotor rehabilitation of actual patients affected by Stroke [2], [3].

COMPLIANCE WITH ETHICAL STANDARDS

The experimental procedures described in this work were conducted in accordance with ethical guidelines for research involving human subjects. Kinematic data capture was performed non-invasively among the members of the research group themselves, who acted as healthy volunteers on a free and informed basis. The execution of the experimental protocols and the anatomical placement of the inertial measurement units (IMUs) were supervised and technically assisted by physiotherapy professionals, ensuring the physical integrity of the volunteers and the methodological accuracy of the trials.

ACKNOWLEDGMENT

The authors acknowledge the use of Google Gemini AI for English language translation, editing, and proofreading during the preparation of this manuscript.

REFERENCES

[1] Y. Béjot, H. Bailly, J. Durier, and M. Giroud, "Epidemiology of stroke in Europe and trends for the 21st century," *La Presse Médicale*, vol. 45, no. 12, Part 2, pp. e391–e398, 2016.

[2] K. N. Borschmann and K. S. Hayward, "Recovery of upper limb function is greatest early after stroke but does continue to improve during the chronic phase," *Physiotherapy*, vol. 107, pp. 216–223, 2021.

[3] A. Roby-Brami, N. Jarrassé, and R. Parry, "Impairment and compensation in dexterous upper-limb function after stroke," *Frontiers in Human Neuroscience*, vol. 15, 2021.

[4] F. Routhier *et al.*, "Impacts of a robotic arm on people with upper-limb disabilities due to neuromuscular disorder," *Journal of Prosthetics and Orthotics*, vol. 34, no. 3, pp. 180–191, 2022.

[5] B. O. Lee, I. D. Saragih, and S. O. Batubara, "Robotic arm use for upper limb rehabilitation after stroke: A systematic review and meta-analysis," *Kaohsiung Journal of Medical Sciences*, vol. 39, no. 5, pp. 435–445, 2023.

[6] R. Shadmehr and S. P. Wise, *The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*. Cambridge, MA, USA: MIT Press, 2005.

[7] A. K. Pandey and R. Alami, "Mightability Maps: A perceptual level decisional framework for cooperative and competitive human–robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2010.

[8] T. Nef, M. Guidali, and R. Riener, "ARMin III arm therapy exoskeleton with an ergonomic shoulder actuation," *Applied Bionics and Biomechanics*, vol. 6, no. 2, pp. 127–142, 2009.

[9] M. A. D. Eguizabal, "Modelado músculo-esquelético do miembro superior e desenvolvimento do sistema de controle de um dispositivo de reabilitação de hombro," Ph.D. dissertation, Universidad Politécnica de Madrid, Madrid, Spain, 2018.

[10] S. Campagnini *et al.*, “Effects of control strategies on gait in robot-assisted post-stroke lower limb rehabilitation: A systematic review,” *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 52, 2022.
[11] D. M. Barbosa, “Using a Not Immersed Virtual Reality System as a Tool for the Hemiparetic Upper Limbs Rehabilitation,” M.Sc. thesis, Universidade Federal de Goiás, Goiânia, Brazil, 2008.
[12] D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. N. Whittlesey, *Research Methods in Biomechanics*, 2nd ed. Champaign, IL, USA: Human Kinetics, 2014.

Variational Quantum Classifier for Multiclass Parkinson's Disease Staging on the PPMI Dataset

Rafaelly Rios dos Santos
Department of Electrical Engineering
Federal University of Ceará
Fortaleza, Brazil
rafaellyrios.adv@gmail.com

Fabio Borges Amorim Cavalcante
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
fabioborges@alu.ufc.br

Luis Francisco Sousa Leal Gomes
Department of Computing
Federal University of Ceará
Fortaleza, Brazil
luisafransou@alu.ufc.br

Marcello Carvalho dos Reis
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
mcarv.reis@gmail.com

Victor Hugo Costa de Albuquerque
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
victor.albuquerque@lesc.ufc.br

***Abstract*—This paper investigates whether a Variational Quantum Classifier (VQC), combined with a clinically motivated preprocessing pipeline, can perform multiclass staging of Parkinson's Disease while retaining all four diagnostic cohorts: Healthy Control, Parkinson's Disease, Prodromal, and Scans Without Evidence of Dopaminergic Deficit (SWEDD), which most prior studies on the Parkinson's Progression Markers Initiative (PPMI) dataset exclude to reduce the task to binary classification. The pipeline integrates MICE imputation, Partial Least Squares (PLS) dimensionality reduction, and SMOTE class balancing to feed an 8-qubit VQC with four strongly entangling layers. All evaluations use a patient-grouped train-test split (StratifiedGroupKFold), preventing the same patient from contributing records to both partitions. In an ablation study, the full preprocessing pipeline raises the weighted AUC from 0.750 to 0.808, though individual components trade AUC against balanced accuracy. On the same 8-dimensional representation, the variational quantum classifier (VQC) reaches a weighted area under the receiver operating curve (ROC) curve (AUC) of 0.808, against 0.832 for the best baseline (multilayer perceptron (MLP)) and 0.809 for random forest (RF). On the severely underrepresented scans without evidence of dopaminergic deficit (SWEDD) cohort (85 records) it obtains 0.756, against 0.760 for RF and 0.664 for the weakest baseline (support vector machine (SVM) RBF). These results are obtained from a single fold of a patient-grouped split and a statevector simulation; they indicate competitive performance without demonstrating quantum advantage.**

***Index Terms*—Quantum Machine Learning, Variational Quantum Classifier, Parkinson's Disease, PPMI, Multiclass Classification**

I. Introduction

Of the 97 supervised learning studies on the parkinson's progression markers initiative (PPMI) dataset surveyed by [1], 55 discard two of the four diagnostic cohorts before training. Those two, Prodromal and SWEDD, are also the cohorts that that make the diagnosis hard. Prodromal patients carry premotor risk markers without overt motor signs; SWEDD patients present parkinsonian symptoms alongside normal dopaminergic neuroimaging [2]. Reducing the task to parkinson's disease (PD) versus healthy control (HC) as a result, leaves out the diagnostic boundaries that are most clinically challenging.

The PPMI dataset introduces additional challenges: cognitive assessments entered the protocol after enrollment had begun, and dropout is more frequent among severely affected patients, resulting in missingness that may not be random [2].

VQC use parameterized quantum circuits whose parameters can be optimized with classical gradient-based methods, making them compatible with current noisy intermediate-scale quantum (NISQ) devices [3], [4]. VQC models have reported competitive results on clinical benchmarks for diabetes [5], cancer [6], and imbalanced or noisy data [7] they are almost always binary, typically on curated datasets.

This paper asks whether a standard VQC, given a preprocessing pipeline designed to address these data-related challenges, can be applied to the full four-cohort classification task. We combine multiple imputation by chained equations (MICE) imputation, partial least squares (PLS) projection onto eight components, and synthetic minority over-sampling technique (SMOTE) balancing to feed an 8-qubit circuit with four strongly entangling layers. An ablation study isolates what each stage contributes; five classical baselines run on the identical representation and the identical patient-grouped test partition; and a per-class analysis evaluates performance across cohorts, including the PD-versus-SWEDD boundary. We do not claim quantum advantage, and the protocol based on a single fold, statevector simulation which limits the scope of the conclusions that can be drawn from the results.

II. Related Work

A. Machine Learning on the PPMI Dataset

The PPMI dataset has become one of the most widely adopted benchmarks for machine learning research in Parkinson's Disease. In a systematic review of 97 supervised learning studies conducted on PPMI data, Gerraty et al. [1] identified three recurring methodological limitations. The first concerns scope:

55 of the 97 studies excluded the Prodromal and SWEDD cohorts due to severe class imbalance, reducing the task to a binary PD-vs-HC problem that does not reflect the full diagnostic complexity. A second issue is leakage: 20 studies applied data transformations to the entire dataset prior to partitioning, artificially inflating reported performance. Finally, most studies relied on a single data modality, typically clinical scores combined with DAT-SPECT imaging, leaving genetic, proteomic, and multi-modal integration largely unexplored.

Brumm et al. [2] further documented the structural challenges of the PPMI cohort, including systematic dropout among the most clinically severe patients and the absence of baseline cognitive assessments for a large fraction of participants, both of which introduce missing not at random (MNAR) patterns that complicate imputation and modeling. The present work addresses these gaps by retaining all four diagnostic classes, applying MICE imputation to handle missing data, using SMOTE for class balancing, and employing a strict train-test separation to prevent preprocessing induced leakage.

B. quantum machine learning (QML) in healthcare

QML has emerged as a promising paradigm for classification tasks in healthcare [3]. Lamichhane and Rawat [6] survey recent advances and compile reported VQC accuracies of up to 90% for breast cancer detection and 85% for lung cancer classification on benchmark datasets. Across this literature, most studies employ binary classification on well-curated, balanced datasets, leaving multiclass problems on noisy clinical data largely unaddressed.

Khan et al. [5] applied an optimized VQC to the Pima Indians Diabetes Dataset, demonstrating competitive performance against classical SVM and RF baselines after feature selection and hyperparameter tuning. Sheoran et al. [7] evaluated both classical and quantum classifiers under realistic conditions including noise injection, class imbalance treated with SMOTE, and dimensionality reduction, concluding that VQC models exhibit comparable robustness to classical counterparts when appropriate preprocessing is applied. Tudisco et al. [8] proposed Multi-VQC, an architecture that chains multiple variational quantum circuits in sequence, achieving improved classification on prostate cancer, heart failure and diabetes datasets.

III. Proposed VQC Model Pipeline

A. Variational Quantum Classifier

The predictive core of the proposed pipeline is a VQC, a hybrid quantum-classical model in which classical clinical features are encoded into quantum states and processed by a trainable parameterized circuit [4]. The implementation relies on the PennyLane framework [9] interfaced with PyTorch for gradient-based optimization.

1) Data encoding: After dimensionality reduction Section III-C, each input sample is represented as a feature vector $\mathbf{x} \in \mathbb{R}^n$ with $n = 8$, scaled to the interval $[0, \pi]$ via Min-Max normalization. The encoding into quantum states is performed through Angle Embedding, which maps each component x_i to a single-qubit rotation on the i-th wire:

$$\mathcal{S}(\mathbf{x}) = \bigotimes_{i=0}^{n-1} R_Y(x_i)\,|0\rangle_i, \tag{1}$$

where $R_Y(\theta) = \exp(-i\theta\sigma_y/2)$ denotes a rotation about the Y-axis of the Bloch sphere. This encoding establishes a one-to-one correspondence between features and qubits, requiring $n = 8$ qubits for an 8-dimensional input.

2) Ansatz Architecture: The trainable component of the circuit consists of $L = 4$ layers of *Strongly Entangling Layers* [4]. Each layer ℓ applies two operations in sequence: (i) local parameterized rotations on each qubit across all three Bloch-sphere axes, and (ii) CNOT gates arranged in a cyclic ring pattern that couples qubit i to qubit $(i + \ell) \bmod n$. Each layer is thus expressed as:

$$\begin{aligned}\mathcal{W}^{(\ell)}(\boldsymbol{\theta}^{(\ell)}) = &\prod_{i=0}^{n-1} \mathrm{CNOT}_{i,(i+\ell) \bmod n} \\ &\times \prod_{i=0}^{n-1} R_Z(\theta_{i,2}^{(\ell)})\, R_Y(\theta_{i,1}^{(\ell)})\, R_X(\theta_{i,0}^{(\ell)}).\end{aligned} \tag{2}$$

where $\boldsymbol{\theta}^{(\ell)} \in \mathbb{R}^{n\times 3}$ are the trainable parameters of layer ℓ. The full ansatz is the sequential composition $\mathcal{W} = \mathcal{W}^{(4)} \circ \mathcal{W}^{(3)} \circ \mathcal{W}^{(2)} \circ \mathcal{W}^{(1)}$, producing a total of $L \times n \times 3 = 96$ learnable parameters. The cyclic shift in the CNOT connectivity entangles all qubits within a small number of layers [4].

3) Measurement and Classification Strategy: The output of each circuit is obtained by measuring the expectation value of the Pauli-Z observable on the first qubit:

$$s_c(\mathbf{x}) = \langle 0|\, \mathcal{S}^\dagger(\mathbf{x})\, \mathcal{W}_c^\dagger\, \sigma_z^{(0)}\, \mathcal{W}_c\, \mathcal{S}(\mathbf{x})\, |0\rangle + b_c, \tag{3}$$

where $s_c(\mathbf{x})$ is the score for class c and b_c is a classical bias term. A *one-vs-rest (OvR)* strategy is adopted: four independent quantum circuits, each with its own set of parameters $\boldsymbol{\theta}_c$ and bias b_c, are trained simultaneously. At inference, the predicted class is given by $\hat{y} = \arg\max_c\ s_c(\mathbf{x})$.

4) Loss Function and Optimization: The training objective is the Multiclass SVM Loss with a fixed margin $\Delta = 0.15$:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq y_i} \max\Big(0,\ s_j(\mathbf{x}_i) - s_{y_i}(\mathbf{x}_i) + \Delta\Big), \tag{4}$$

where y_i is the true class label of sample $\mathbf{x}_i$. This formulation penalizes any competing class j whose score is within margin Δ of the correct class score, encouraging well-separated decision boundaries. The optimization is performed with the Adam optimizer using an initial learning rate $\alpha_0 = 0.02$ and an exponential step decay schedule with decay factor $\gamma = 0.5$ applied every $\tau = 100$ iterations, such that $\alpha_t = \alpha_0 \cdot \gamma^{\lfloor t/\tau \rfloor}$. Training runs for $T = 1{,}000$ iterations with batches of 32 samples. Circuit weights are initialized as $0.1 \times \mathcal{N}(0, 1)$ with a fixed random seed (`torch.manual_seed = 42`), so the reported results correspond to a single, reproducible parameter initialization rather than an average over independent runs.

All circuits are simulated on the PennyLane `default.qubit` statevector backend with analytic gradient computation.

B. Dataset: PPMI

The PPMI is a prospective, multicenter, international observational study designed to identify biomarkers of PD progression [2]. The dataset used in this work was assembled from three clinical subsets of the PPMI ecosystem: family history, non-motor aspects of experiences of daily living, and genetic components, integrated via patient identifier (`PATNO`).

The target variable (`COHORT_DEFINITION`) assigns each record to one of four diagnostic cohorts, described in Table I.

Table I:
CLINICAL COHORTS AND DESCRIPTIONS

Label	Cohort	Description
C0	Healthy Control	Neurologically healthy individuals with no history of motor symptoms.
C1	Parkinson's Disease	Clinically confirmed PD presenting resting tremor, rigidity, and bradykinesia.
C2	Prodromal	Pre-motor stage characterized by hyposmia, REM sleep behavior disorder, or genetic predisposition.
C3	SWEDD	Parkinsonian motor symptoms with normal dopaminergic neuroimaging, suggesting a distinct etiology.

The consolidated dataset comprises 7,654 records from 6,721 unique participants (`PATNO`), an average of 1.14 records per participant, distributed across the four cohorts with a pronounced class imbalance: the Prodromal cohort constitutes the majority, while SWEDD is severely underrepresented with only 85 records. Because the train-test split is grouped by participant (Section III-D), it is the participant count rather than the record count that bounds the statistical support available for each cohort. This imbalance is a structural characteristic of PPMI: 55 of 97 supervised learning studies reviewed by Gerraty et al. [1] excluded the Prodromal and SWEDD cohorts entirely because of it. In the present work, all four classes are retained to address the full clinical complexity of the diagnostic task.

The feature set consists of 45 clinical variables extracted from the movement disorder society - unified parkinson's disease rating scale (MDS-UPDRS) scale, organized into three groups: Part 1 non-motor symptoms, 15 variables, Part 2 motor experiences of daily living, 14 variables, and Part 4 motor complications, 15 variables, supplemented by two familial genetic history indicators (`FULSIBPD`, `PAGPARPD`). This selection followed a domain-informed criterion: only variables obtainable from a clinical anamnesis, in principle via natural language processing, were retained, excluding synthetic latent variables and imaging-derived biomarkers. Three structural challenges of this dataset directly influenced the preprocessing strategy adopted in this work:

1) *Missing data under MNAR mechanism.* The majority of patients lacked cognitive assessments at baseline, as this instrument was incorporated into the PPMI protocol after enrollment began [2]. Also, patients with greater clinical severity exhibited higher dropout rates, producing systematic data absence in precisely the most critical profiles. This MNAR pattern implies that the absence itself carries clinical information, and naive imputation methods may introduce severe bias.
2) *Phenotypic overlap between PD and SWEDD.* Differentiating PD from SWEDD without dopaminergic neuroimaging (DaTscan) is challenging even for movement disorder specialists [1]. Both cohorts share overlapping motor symptom profiles, making this the hardest classification boundary in the dataset.
3) *Risk of data leakage.* The longitudinal structure of PPMI means that a patient may contribute more than one record across visits. If records are not grouped by patient when partitioning the data, information can leak between training and test sets, a methodological flaw identified in 20 of 97 supervised studies on PPMI [1]. Although repeated records are a minority in the present dataset (1.14 records per participant on average), the risk applies to any patient contributing more than once, and we address it directly by grouping the train-test split by `PATNO`, described next in Section III-D.

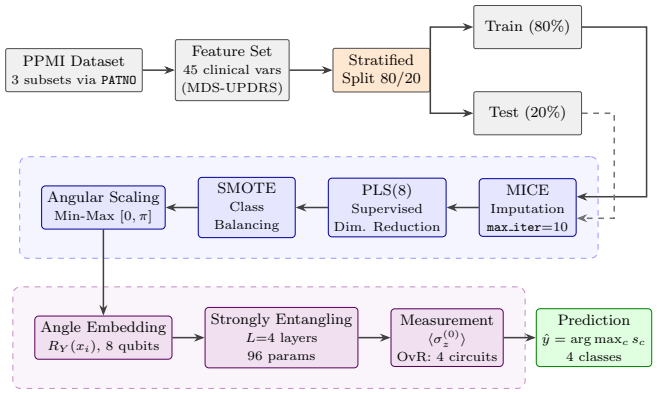


Fig. 1. Block diagram of the proposed preprocessing pipeline.

C. Preprocessing Pipeline

The preprocessing pipeline was designed to address each of the challenges described above through a four stage sequential transformation, illustrated in Fig. 1. All transformations are fitted exclusively on the training partition and applied to the test set, which prevents preprocessing induced leakage.

1) Multivariate Imputation MICE: Missing values are handled through MICE [10], an iterative algorithm that models each incomplete variable as a function of all remaining variables via a sequence of conditional regressions. At each iteration, MICE cycles through the incomplete columns, imputing values by drawing from the predictive distribution conditioned on the current state of all other variables, and repeats this process for a fixed number of rounds (`max_iter` = 10). This approach preserves the multivariate correlation structure of the clinical data, which is critical when symptom variables from the MDS-UPDRS exhibit strong interdependencies. Prior to imputation, sentinel values (-1) used as implicit absence codes in the PPMI are converted to `NaN`.

2) Supervised Dimensionality Reduction PLS: After imputation, the 45-dimensional feature space must be reduced to 8 components to match the number of available qubits. PLS regression [11] is employed as a supervised alternative to principal component analysis (PCA). While PCA identifies directions of maximum variance irrespective of the target labels, PLS finds latent components that maximize the covariance between the feature matrix $\mathbf{X}$ and the class labels $\mathbf{Y}$ [12]. Because the retained components are oriented toward class separation rather than overall variance, PLS is, in principle, better aligned with a classification objective than PCA. PLS projects the input onto 8 latent components ($n_{\text{components}} = 8$), matching the qubit count of the quantum circuit. The empirical validation of this choice over PCA is provided in the ablation study (Section IV-A).

3) Class Balancing SMOTE: The severe class imbalance is addressed through the SMOTE [13], applied exclusively on the training set to prevent synthetic samples from leaking into the evaluation partition. SMOTE generates new instances for underrepresented classes by interpolating between existing samples and their nearest neighbors in the feature space. A custom sampling strategy is adopted, in which all minority classes are oversampled to at least 50% of the majority class size. The number of neighbors k is set adaptively as $k = \max(1, \min(5, n_{\min} - 1))$, where $n_{\min}$ is the smallest class count, ensuring that SMOTE remains applicable even for extremely small classes such as SWEDD.

4) Angular Scaling: Finally, the PLS transformed and SMOTE balanced features are rescaled to the interval $[0, \pi]$ using Min-Max normalization, mapping each component to the angular domain required by the Angle Embedding stage of the quantum circuit (Section III-A).

D. Evaluation Protocol

The dataset is split into training and testing partitions (approximately 80%/20%) using `StratifiedGroupKFold` ($n_{\text{splits}} = 5$, `shuffle=True`, `random_state= 42`), grouped by patient identifier (`PATNO`). This guarantees two properties simultaneously: no patient contributes records to both partitions, verified by an explicit zero-overlap check on every run, and the class proportions of the four cohorts are preserved despite the severe underrepresentation of the SWEDD cohort. We adopt the first of the five folds throughout this work. All preprocessing transformations, MICE imputation, PLS projection, SMOTE oversampling, and angular scaling, are fitted exclusively on the training partition and applied to the test set, preventing preprocessing-induced leakage. We note one constraint of this protocol that bounds the strength of our conclusions: all results derive from a single fold rather than from the full 5-fold cross-validation, which is discussed in Section V.

1) Performance Metrics: The metrics used to evaluate the models are AUC, F1-Score, Accuracy, and Balanced Accuracy. We denote C as the number of classes, N the total number of test samples, N_c the number of samples belonging to class c, and TP_c, FP_c, FN_c the true positives, false positives, and false negatives for class c, respectively.

Balanced Accuracy: [14] the unweighted mean of the per-class recalls,

$$\text{BAcc} = \frac{1}{C} \sum_{c=0}^{C-1} R_c \,. \tag{5}$$

By giving every cohort equal weight regardless of its support, it exposes the majority-class bias that raw accuracy conceals: a classifier that ignores the minority cohorts entirely can retain high accuracy while its balanced accuracy collapses toward $1/C$. The two metrics therefore diverge sharply under severe imbalance, as the ablation study demonstrates in Section IV-A.

Accuracy: the proportion of correctly classified samples. While intuitive, this metric can be misleading under class imbalance, as a model predicting only the majority class may achieve high accuracy without learning meaningful patterns.

Weighted F1-Score: [15] aggregates the per-class harmonic mean of precision (P_c) and recall (R_c), weighted by class support:

$$F_{1,c} = \frac{2\,P_c\,R_c}{P_c + R_c}\,, \quad P_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}\,, \quad R_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}\,, \tag{6}$$

Now, the weighted metric is obtained as:

$$F_1^w = \sum_{c=0}^{C-1} \frac{N_c}{N} F_{1,c} \,. \tag{7}$$

Weighted AUC: [16] is computed under the ROC curve in a OvR fashion. For each class c, the ROC curve plots the true positive rate against the false positive rate across all classification thresholds τ, and the AUC corresponds to the integral [17]:

$$\text{AUC}_c = \int_0^1 \text{TPR}_c\left(\text{FPR}_c^{-1}(t)\right)\, dt \,. \tag{8}$$

The weighted average is then obtained as [16]:

$$\text{AUC}^w = \sum_{c=0}^{C-1} \frac{N_c}{N} \text{AUC}_c \,. \tag{9}$$

The AUC quantifies the probability that the classifier ranks a randomly chosen positive instance above a randomly chosen negative one [16]. Unlike threshold-dependent metrics, AUC evaluates the entire spectrum of decision boundaries and is invariant to class distribution, which makes it well suited to imbalanced clinical datasets [17]. The AUC is adopted as the **primary metric** throughout this work. This choice is motivated by two factors: (i) the severe class imbalance renders accuracy unreliable: as demonstrated in the ablation study, a model can achieve 67.9% accuracy while attaining only 35.5% balanced accuracy by effectively ignoring minority classes; and (ii) in a clinical decision support context, the ability to discriminate between diagnostic categories across all operating points is more informative than performance at a single fixed threshold.

2) Ablation Study Design: To quantify the contribution of each preprocessing component, an ablation study [18] is conducted with four experimental configurations sharing the same VQC architecture and hyperparameters (8 qubits, 4 layers, $\Delta = 0.15$, $T = 1{,}000$ iterations). The configurations are summarized in Table II.

Table II:
ABLATION STUDY — EXPERIMENTAL CONFIGURATIONS

Exp.	MICE	Balancing	Dim. Reduction
EXP1	✓	—	—
EXP2	✓	SMOTE	—
EXP3	✓	SMOTE	PCA(8)
EXP4	✓	SMOTE	PLS(8)

MICE imputation is present in all experiments as a baseline requirement, since the raw PPMI data cannot be fed directly into any classifier without handling missing values. Each successive experiment adds one component, isolating its marginal contribution to the overall performance.

3) Baseline Comparison: To contextualize the VQC performance against established classical methods, five baseline classifiers are evaluated under identical preprocessing conditions (EXP4 pipeline: MICE + PLS(8) + SMOTE): logistic regression (LR), SVM with radial basis function kernel, RF with 500 estimators, gradient boosting (GB), and a two-layer MLP (256-128 hidden units). All classical models use default scikit-learn hyperparameters except where noted. This keeps the comparison focused on the learning paradigm rather than on extensive tuning, with the caveat, revisited in Section V, that tuning could shift the gap in either direction.

IV. RESULTS

A. Ablation Study

Table III summarizes the performance of the VQC across the four experimental configurations defined in Section III-D2. Each row adds one preprocessing component while keeping the circuit architecture and all hyperparameters fixed, isolating the marginal contribution of each stage.

Table III:
ABLATION STUDY — IMPACT OF EACH PIPELINE COMPONENT ON VQC PERFORMANCE

Exp.	Acc.	Bal. Acc.	F_1^w	AUC^w
EXP1	**0.679**	0.355	0.641	0.750
EXP2	0.612	0.389	0.606	0.734
EXP3	0.621	**0.506**	0.668	0.807
EXP4	0.621	0.497	**0.675**	**0.808**

EXP1 (Baseline) Without class balancing or dimensionality reduction, the VQC achieves the highest raw accuracy (0.679) but the lowest AUC (0.750) and a critically low balanced accuracy (0.355). This discrepancy reveals that the model learns to predict predominantly the majority class, effectively ignoring the underrepresented SWEDD and Prodromal cohorts. This result exemplifies why accuracy alone is an unreliable metric for imbalanced clinical datasets [14].

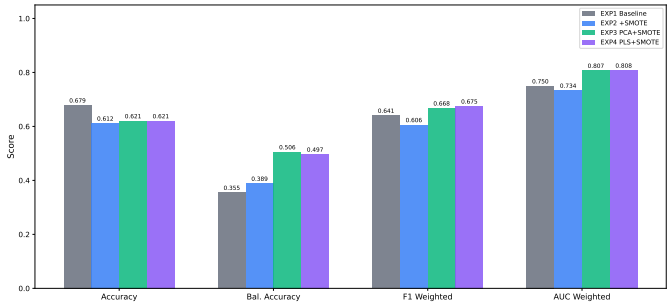


Fig. 2. Ablation study across the four pipeline configurations. Adding SMOTE (EXP2) improves balanced accuracy at the cost of AUC, while supervised PLS reduction (EXP4) recovers the highest AUC and F_1, motivating its adoption as the final pipeline.

EXP2 (+SMOTE) Adding SMOTE raises balanced accuracy from 0.355 to 0.389 (a relative increase of 9.6%), at the cost of the aggregate AUC, which drops from 0.750 to 0.734. This is expected: forcing the classifier to allocate representational capacity to the severely underrepresented cohorts trades a fraction of ranking quality on the majority classes for sensitivity to the minority ones [13]. The simultaneous decrease in raw accuracy (from 0.679 to 0.612) reflects the same trade-off: the model gives up majority-class bias for more equitable sensitivity across classes.

EXP3 (PCA+SMOTE) Introducing PCA based dimensionality reduction yields the largest AUC improvement, from 0.734 to 0.807, and the largest balanced-accuracy gain, from 0.389 to 0.506, the highest balanced accuracy among the four configurations. By projecting the 45-dimensional feature space onto 8 principal components that capture the directions of maximum variance, PCA removes redundant and noisy dimensions that previously hindered the quantum circuit's capacity to find separating boundaries.

EXP4 (PLS+SMOTE) Replacing PCA with supervised PLS projection achieves the highest AUC (0.808 vs. EXP3's 0.807) and the highest weighted F_1 (0.675 vs. 0.668), with a balanced accuracy (0.497) marginally below EXP3's (0.506). PLS and PCA therefore yield essentially equivalent aggregate performance, with EXP4 holding a small edge on the primary metric (AUC) itself. Unlike PCA, PLS aligns the retained components with the class labels rather than with overall variance [11], [12], and the resulting model is directly interpretable in terms of clinical variable loadings, enabling domain experts to inspect which MDS-UPDRS items contribute most to each latent component. Given its lead on AUC and this interpretability advantage, EXP4 is adopted as the final pipeline configuration; we revisit this trade-off explicitly in Section VI.

The per-class ROC curves for PD, Fig. 3 follow the same non-monotonic trajectory as the aggregate metrics. SMOTE alone costs 0.012 of AUC_{PD} (0.797 → 0.785); projecting onto eight components recovers that loss several times over, reaching 0.878 with PCA and 0.857 with PLS. On the class of greatest clinical interest, then, the trade-off identified in Table III plays out identically.

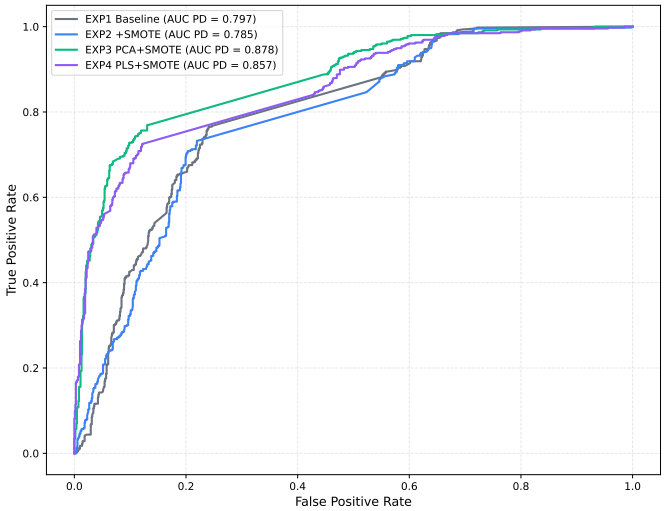


Fig. 3. ROC curves for the PD class across the four ablation experiments. Dimensionality reduction (EXP2-EXP4) raises AUC_{PD} from 0.785 to 0.857, the largest single contribution to PD discrimination.

B. Comparison with Classical Baselines

Five classical classifiers run on the EXP4 pipeline (MICE + PLS(8) + SMOTE) and the same patient-grouped test partition. Table IV presents the results. The VQC row reuses the model trained for EXP4 in the ablation study (Section IV-A) rather than an independent run: the group split (`random_state=42`) and the circuit weight initialization (`torch.manual_seed=42`, Section III-A) are both fixed, so this row and the EXP4 row of Table III report the identical trained circuit evaluated on the identical test partition.

Table IV:
PERFORMANCE COMPARISON — CLASSICAL BASELINES VS. VQC

Model	Acc.	Bal. Acc.	F_1^w	AUC^w
LR	0.549	0.508	0.624	0.819
SVM RBF	0.649	0.514	0.687	0.820
RF	**0.713**	0.455	**0.719**	0.809
GB	0.662	0.499	0.695	0.819
MLP	0.661	**0.521**	0.690	**0.832**
VQC	0.621	0.497	0.675	0.808

No single classical model leads across all four metrics: RF attains the highest accuracy (0.713) and weighted F_1 (0.719) but the lowest balanced accuracy (0.455), while MLP attains the highest AUC (0.832) and balanced accuracy (0.521). The VQC attains an AUC of 0.808, which is 0.024 below MLP and 0.001 below RF. The aggregate ranking hides class-level differences, which are examined in Section IV-C

All six models operate on the same 8-dimensional PLS representation, since the feature space is constrained to match the qubit count of the quantum circuit. Classical models such as RF and GB could in principle exploit higher-dimensional representations, whereas the VQC is restricted to 8 features by the available qubits. Under this constraint all six models fall within a 0.808–0.832 band.

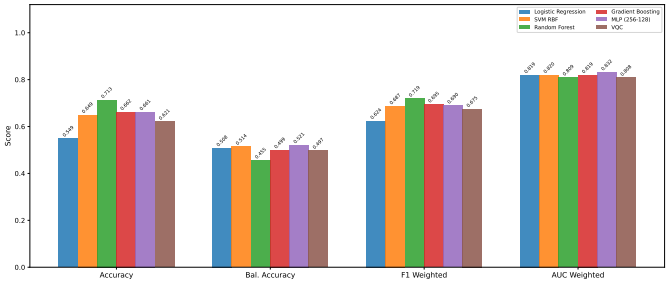


Fig. 4. Aggregate metrics for the five classical baselines and the VQC (MICE + PLS(8) + SMOTE). MLP leads on AUC (0.832); the VQC reaches 0.808 under the 8-feature constraint imposed by the qubit count.

C. Per-Class ROC Analysis

While aggregate metrics provide an overall summary of classifier performance, the clinical relevance of a diagnostic model depends on its ability to discriminate each individual cohort. Table V reports the per-class AUC for all six models, computed in a OvR fashion. Three findings emerge from the per-class analysis.

Table V:
PER-CLASS AUC — CLASSICAL BASELINES VS. VQC

Model	HC	PD	Prodromal	SWEDD
LR	0.775	0.867	0.804	0.757
SVM RBF	0.786	0.881	0.798	0.664
RF	0.709	0.890	0.785	**0.760**
GB	0.766	0.883	0.797	0.752
MLP	**0.799**	**0.893**	**0.809**	0.683
VQC	0.767	0.857	0.791	0.756

PD class. Every model clears 0.85 here. The VQC reaches 0.857; MLP leads at 0.893, with LR (0.867) and SVM RBF (0.881) in between. On the primary diagnostic target, then, the circuit picks up the dominant motor-symptom patterns in the MDS-UPDRS.

SWEDD class. This cohort spreads the models farthest apart. SVM RBF falls to 0.664 and MLP to 0.683, while LR, GB, RF and the VQC cluster between 0.752 and 0.760. This four-way tie is difficult to distinguish with the current test partition, since SWEDD contributes only 85 records in the full dataset and roughly a fifth of those to the test split.

HC and Prodromal classes. For HC, all models perform within a band between 0.709 and 0.799, with the VQC (0.767) in the middle of the range; RF is, somewhat unexpectedly, the weakest model on this comparatively easy class (0.709), while MLP leads (0.799). For Prodromal, the VQC reaches 0.791, 0.018 below MLP (0.809) and above RF (0.785). The Prodromal cohort occupies an intermediate position in the symptom spectrum, more severe than HC but less pronounced than PD, which may explain why no single model dominates this boundary.

The aggregate gap in Table IV is therefore not a uniform deficit: the circuit stands closest to the baselines on PD detection and SWEDD differentiation, which are clinically important

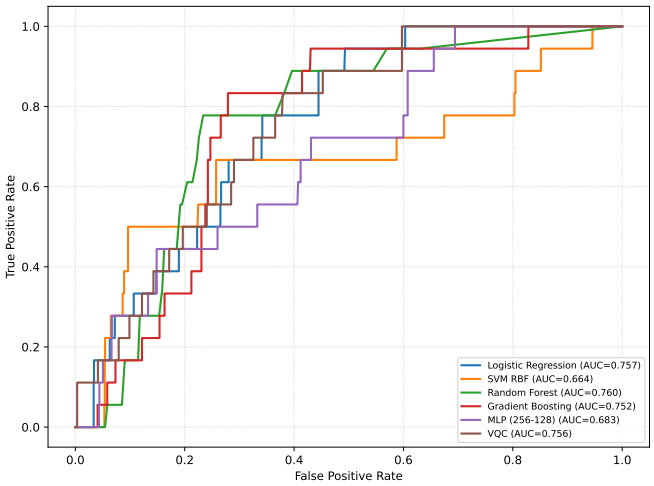


Fig. 5. ROC curves for the SWEDD class. The VQC (AUC = 0.756) ranks above the SVM RBF (0.664) and MLP (0.683), and 0.004 below the best model (RF, 0.760); given the small cohort (85 records), these estimates carry wide uncertainty.

distinctions. Determining the reasons for this behavior would require controlled ablations of the circuit itself.

V. Discussion

The results in Section IV indicate that a VQC, combined with the proposed preprocessing pipeline, can achieve competitive performance on a multiclass PD staging task.

A. The Role of Preprocessing

The ablation study shows that the preprocessing pipeline is a determining factor in the VQC's performance, not an auxiliary step. Without class balancing and dimensionality reduction (EXP1), the circuit reaches 67.9% accuracy yet only 35.5% balanced accuracy, effectively ignoring minority classes. This is not specific to the quantum model: it is the expected consequence of training on severely skewed class distributions, a well-documented phenomenon in classical machine learning [13] that applies equally to quantum classifiers.

The progression from EXP1 to EXP4 demonstrates that each component addresses a specific pathology of the data: SMOTE restores sensitivity to underrepresented cohorts, at a short-term cost to aggregate AUC, and PLS compresses the feature space into a supervised low-dimensional representation compatible with the circuit's qubit budget, recovering and exceeding that cost. The combined effect is an AUC improvement from 0.750 to 0.808 that no single technique accounts for in isolation, and that is not monotonic stage by stage, which is why we report the contribution of each stage separately rather than only the final pipeline.

B. Clinical Interpretation of Model Errors

The confusion patterns in the per-class analysis track the underlying clinical reality. The PD-vs-SWEDD misclassification is the dominant source of error across all models, including the VQC. This is clinically expected, as SWEDD patients present parkinsonian motor symptoms that are phenotypically indistinguishable from early PD without dopaminergic neuroimaging (DaTscan) [2]. Consistently with this, SWEDD shows the widest performance spread of all four classes: the best model on this boundary (RF) reaches a per-class AUC of 0.760, against 0.890 for PD within the same model, and two of the six classifiers (SVM RBF and MLP) fall below 0.75. No model, classical or quantum, separates SWEDD as cleanly as it separates the other cohorts, which points to an intrinsic difficulty of the boundary rather than to a model-specific weakness. The VQC (0.756) falls within the leading group on this boundary, 0.004 behind RF and well ahead of SVM RBF and MLP. Still, the small sample size does not firmly establish this ordering. The mechanism behind this behavior is not established by our experiments; clarifying whether it stems from the quantum representation or from other factors would require controlled ablations of the circuit itself, which we leave to future work.

C. Limitations

Several limitations of this work should be acknowledged.

First, all quantum circuit simulations are executed on the PennyLane `default.qubit` statevector backend, which performs an exact classical simulation of the circuit and does not capture the decoherence and gate errors of real NISQ hardware. The behavior of the VQC under hardware noise therefore remains an open question.

Second, because PPMI is longitudinal, records from the same patient could in principle appear in both partitions and introduce patient-level leakage. We remove this risk by grouping the train-test split by `PATNO` with `StratifiedGroupKFold`, verified by an explicit zero-overlap check on every run (Section III-D). What remains is a narrower limitation: all results derive from a single fold of the five, with a fixed seed, rather than from the full cross-validation. Reporting the mean and variance across all five folds would tighten these estimates and is left for future work.

Third, the VQC architecture is constrained to 8 qubits with 4 layers, totaling 96 trainable parameters per circuit. This configuration was dictated by the current practical limits of statevector simulation. More expressive ansatz designs, deeper circuits, alternative entanglement topologies, or data re-uploading strategies, may improve performance but at increased simulation cost and potential susceptibility to barren plateaus [3].

Fourth, the classical baselines were evaluated with default hyperparameters to ensure a fair comparison focused on learning paradigms rather than model tuning. It is possible that systematic hyperparameter optimization of the classical models would widen the performance gap. On the other hand, the VQC hyperparameters (margin Δ, learning rate schedule, number of layers) were also not exhaustively tuned, leaving room for improvement on both sides.

D. Practical Implications

Despite the limitations above, the results offer preliminary evidence that, under the right preprocessing, VQC models can be competitive on this kind of clinical classification task, particularly when: (i) the feature space can be reduced to a

dimensionality compatible with available qubit counts; (ii) class imbalance is addressed through resampling prior to quantum encoding; and (iii) evaluation prioritizes AUC over accuracy, reflecting the threshold-free nature of clinical decision-making.

The VQC's comparatively high per-class AUC on SWEDD, the most underrepresented and phenotypically ambiguous cohort, is a result worth examining further rather than a settled finding, given the cohort size. Whether quantum classifiers occupy a genuine niche where sample scarcity and class overlap coexist is an open question. As quantum hardware matures beyond the NISQ era, the simulation overhead associated with these models is expected to fall, which would make experiments such as ours easier to scale and validate.

VI. Conclusion

This paper presented a hybrid quantum-classical pipeline for multiclass PD staging using clinical data from the PPMI dataset. The proposed approach combines MICE imputation, PLS dimensionality reduction and SMOTE class balancing to enable a VQC with 8 qubits and 4 layers of strongly entangling layers to operate on a diagnostic task that the majority of prior studies have simplified by excluding minority cohorts.

The ablation study showed that each preprocessing stage contributes measurably to performance, with the weighted AUC rising from 0.750 to 0.808 across the full pipeline; the gain is not monotonic stage by stage, since SMOTE alone trades aggregate AUC for balanced accuracy, and it is the combination with PLS that recovers and exceeds it. On the same representation, the VQC trailed MLP by 0.024 of AUC and RF by 0.001, while achieving its closest performance to the baselines on SWEDD, the smallest and most phenotypically ambiguous of the four cohorts, where it ranked third of six models and 0.004 below the best classical result.

The choice of PLS over PCA as the final pipeline (EXP4 over EXP3, Section IV-A) deserves an explicit remark, since AUC is adopted as the primary metric throughout this work (Section III-D). With the patient-grouped split, EXP4 in fact attains a marginally higher AUC than EXP3 (0.808 vs. 0.807) and a higher weighted F_1, at the cost of a small reduction in balanced accuracy (0.497 vs. 0.506). The selection is therefore consistent with the stated primary metric on this fold; we retain the interpretability of PLS's clinically loaded components as an additional, secondary justification, and note that this ranking could shift under the full cross-validation discussed below.

These results do not claim quantum advantage. They provide empirical evidence that variational quantum classifiers, even under classical simulation and limited qubit counts, can reach performance competitive with classical baselines on a challenging real-world diagnostic task, subject to the validation limitations discussed above.

Future work will pursue three directions. First, evaluating the pipeline on actual NISQ hardware to quantify the impact of gate noise and decoherence on classification performance. Second, reporting metrics averaged over all five folds of the patient-grouped `StratifiedGroupKFold` split, rather than the single fold used here, to quantify variance across patient partitions and strengthen the statistical robustness of the reported metrics. Third, exploring more expressive quantum architectures, including data re-uploading strategies and sequential circuit compositions such as the Multi-VQC approach proposed by Tudisco et al. [8], to assess whether deeper quantum models can further close the gap with classical baselines on the PPMI dataset.

Acknowledgment

Data used in the preparation of this article was obtained on November 18, 2024 from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR_006431. For up-to-date information on the study, visit www.ppmi-info.org. This analysis used data openly available from PPMI.

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners. The full list of funding partners can be found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors.

References

[1] R. T. Gerraty *et al.*, "Machine learning within the parkinson's progression markers initiative: Review of the current state of affairs," *Frontiers in Aging Neuroscience*, vol. 15, p. 1076657, 2023.

[2] M. C. Brumm *et al.*, "Parkinson's progression markers initiative: A milestone-based strategy to monitor parkinson's disease progression," *Journal of Parkinson's Disease*, vol. 13, pp. 899–916, 2023.

[3] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021.

[4] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, no. 3, p. 032308, 2020.

[5] H. Khan *et al.*, "Diabetes prediction using an optimized variational quantum classifier," *International Journal of Intelligent Systems*, p. 1351522, 2025.

[6] P. Lamichhane and D. B. Rawat, "Quantum machine learning: Recent advances, challenges, and perspectives," *IEEE Access*, vol. 13, pp. 94 051–94 100, 2025.

[7] S. K. Sheoran, V. Yadav, and R. K. Sheoran, "Robust evaluation of classical and quantum machine learning under noise, imbalance, feature reduction and explainability," *Scientific Reports*, vol. 15, p. 28412, 2025.

[8] A. Tudisco, D. Volpe, and G. Turvani, "Multi-vqc: A novel qml approach for enhancing healthcare classification," in *2025 IEEE International Conference on Quantum Software (QSW)*. IEEE, 2025, pp. 28–34.

[9] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith *et al.*, "Pennylane: Automatic differentiation of hybrid quantum-classical computations," *arXiv preprint arXiv:1811.04968*, 2022.

[10] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[11] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: A basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[12] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[14] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior agreement," in *Proc. 20th Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 3121–3124.

[15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[16] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[17] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[18] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks," *arXiv preprint arXiv:1901.08644*, 2019.

Evaluation of the Performance of the 5%/1 mm Gamma Criterion in Patient-Specific Quality Assurance of VMAT Treatment Plans for Stereotactic Radiosurgery

G. S. dos Santos[1], R. A. Santana[2], F. C. F. Araújo[1], Emerson N. Itikawa[1,2]

[1]*Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC), Universidade Federal de Goiás, Goiânia, GO, Brasil*

[2]*Instituto de Física (IF), Universidade Federal de Goiás, Goiânia, GO, Brasil*

Abstract — Stereotactic radiosurgery (SRS) using Volumetric Modulated Arc Therapy (VMAT) requires high geometric accuracy, making Patient-Specific Quality Assurance (PSQA) essential for verifying the accuracy of dose delivery. This study evaluated the performance of the 5%/1 mm Gamma criterion compared with the 2%/2 mm criterion for PSQA of intracranial SRS treatment plans using VMAT. A retrospective analysis was performed on 240 clinical treatment plans generated using the Monaco® treatment planning system and delivered on an Elekta Infinity linear accelerator. The Gamma Passing Rate (GPR) was analyzed in relation to treatment plan parameters, including lesion volume, number of lesions, fractionation, and monitor units. The 5%/1 mm criterion showed lower dispersion of GPR values and greater stability across variations in lesion volume, whereas the 2%/2 mm criterion showed a stronger association between increasing irradiated volume and the occurrence of plan failure. The number of lesions was also associated with the behavior of the 5%/1 mm criterion, suggesting a potential influence of treatment plan complexity on the results. The findings demonstrate distinct behaviors between the two criteria and indicate that the 5%/1 mm criterion is a promising alternative for further investigation in the PSQA of SRS treatments using VMAT. However, studies incorporating known geometric and dosimetric errors are required to determine its error-detection capability and to establish a criterion-specific GPR acceptance threshold..

Keywords — *Stereotactic radiosurgery; Volumetric Modulated Arc Therapy (VMAT); Patient-specific quality assurance; Gamma analysis; Quality assurance.*

I. Introduction

Modern radiotherapy has evolved significantly over recent decades with the incorporation of advanced beam intensity modulation techniques, enabling greater conformity of dose distributions to the target volume while improving the sparing of healthy tissues. Among these techniques, *Volumetric Modulated Arc Therapy* (VMAT) has become an important modality for the treatment of various neoplasms, particularly in stereotactic radiosurgery (*Stereotactic Radiosurgery* – SRS), due to its ability to combine high dosimetric conformity, efficient dose delivery, and reduced treatment time [1], [2].

In SRS, high radiation doses are delivered in a single or a few fractions to small target volumes, producing steep dose gradients between the target and adjacent organs at risk. Although this characteristic enables highly conformal dose distributions, it also makes treatment particularly sensitive to geometric and dosimetric uncertainties. Therefore, quality assurance plays an essential role in ensuring that the planned dose distribution is accurately reproduced during clinical delivery [3].

Patient-specific quality assurance (*Patient-Specific Quality Assurance* – PSQA) is an important component of this process. Its purpose is to verify the agreement between the dose distribution calculated by the treatment planning system and that measured during plan verification. Among the available methods, Gamma analysis, proposed by Low *et al.*, remains widely used for comparing dose distributions by combining dose difference (*Dose Difference* – DD) and distance-to-agreement (*Distance-to-Agreement* – DTA) criteria into a single quantitative index [4].

Historically, different combinations of DD and DTA have been used for PSQA of IMRT and VMAT treatments. However, studies have demonstrated that high Gamma passing rates do not necessarily correlate strongly with the absence of clinically relevant dosimetric discrepancies [5], [6]. This limitation is particularly important in stereotactic treatments, which are characterized by small fields and steep dose gradients, where spatial accuracy is a fundamental component of the treatment process [3].

In this context, recommendations from the *American Association of Physicists in Medicine* (AAPM) and the *Radiosurgery Society* (RSS) emphasize the stringent geometric accuracy requirements associated with stereotactic procedures [3], [7]. Therefore, the use of Gamma criteria with more restrictive spatial tolerances may be relevant for the evaluation of these treatments. Among the possible combinations, the 5%/1 mm criterion combines a wider dose-difference tolerance with a 1-mm distance-to-agreement criterion, representing an approach distinct from the 2%/2 mm criterion. However, adopting a more restrictive spatial tolerance does not, by itself, imply a greater ability to detect geometric errors; such capability must be demonstrated through specific studies using known perturbations.

Despite the use of different Gamma criteria in quality assurance programs, clinical evidence regarding the behavior of the 5%/1 mm criterion when applied to PSQA of intracranial SRS treatment plans delivered with VMAT remains limited. Evaluating its behavior across different treatment plan characteristics, such as lesion volume, number of lesions, fractionation, and monitor units, may contribute to a better understanding of its applicability and to the improvement of quality assurance protocols for stereotactic treatments.

Given this context, the present study aimed to evaluate the performance of the 5%/1 mm Gamma criterion compared with the 2%/2 mm criterion in patient-specific quality assurance of intracranial stereotactic radiosurgery treatment plans using VMAT, by analyzing Gamma passing rate behavior and its association with treatment plan characteristics, including lesion volume, number of lesions, fractionation, and monitor units.

II. Material and Methodos

A. Study Design

A retrospective, observational, and quantitative

study was conducted to evaluate the performance of the 5%/1 mm Gamma criterion in patient-specific quality assurance (*Patient-Specific Quality Assurance* – PSQA) of intracranial stereotactic radiosurgery (SRS) treatment plans using *Volumetric Modulated Arc Therapy* (VMAT), compared with the institutional 2%/2 mm criterion. The study was conducted at the Radiotherapy Department of Hospital Araújo Jorge (HAJ–ACCG), Goiânia, GO, Brazil, following approval by the Research Ethics Committee of Hospital Araújo Jorge – ACCG (Approval No. 7,986,757), with a waiver of informed consent.

B. *Sample*

A retrospective analysis was performed on 240 consecutive clinical treatment plans from patients who underwent intracranial stereotactic radiosurgery (SRS) using the VMAT technique. Treatment plans that were clinically used and subjected to the institutional patient-specific quality assurance (PSQA) protocol were included. Cases with incomplete, inconsistent, or missing information for the variables required for the analysis, as well as those for which the PSQA result could not be unequivocally identified, were excluded.

All treatment plans were generated using the Monaco® treatment planning system, version 6.1.3 (Elekta AB, Stockholm, Sweden), with a Monte Carlo dose calculation algorithm, and delivered using an Elekta Infinity linear accelerator with 6 MV flattening filter-free (FFF) photon beams, equipped with a 160-leaf Agility multileaf collimator (MLC).

C. *Patient-Specific Quality Assurance (PSQA)*

Patient-specific quality assurance (PSQA) was performed using the SRS MapCHECK® system (Sun Nuclear Corporation, Melbourne, FL, USA) coupled with the StereoPHAN® phantom, designed for dosimetric verification of stereotactic treatments.

Initially, a computed tomography (CT) scan of the SRS MapCHECK®–StereoPHAN® assembly was acquired and subsequently imported into the Monaco® treatment planning system, version 6.1.3. The clinical treatment plans were transferred and recalculated on the phantom geometry while preserving the irradiation parameters of the original plans. After recalculation, the verification plans were delivered using the Elekta Infinity linear accelerator, and the dose distributions measured with the SRS MapCHECK® were compared with those calculated by the treatment planning system using the SNC Patient® software.

D. *Gamma Evaluation Criteria*

Two Gamma analysis criteria were compared: the institutional 2%/2 mm criterion and the proposed 5%/1 mm criterion, both using a GPR ≥ 95% as the clinical acceptance threshold. For both criteria, global Gamma analysis was performed using a 10% low-dose threshold, with points receiving doses below this threshold excluded from the evaluation. The dose distribution calculated by the Monaco® treatment planning system was used as the reference distribution and compared with the dose distribution measured using the SRS MapCHECK® system through the SNC Patient® software.

The 5%/1 mm criterion was selected based on the use of a more restrictive spatial tolerance, which is particularly relevant for stereotactic treatments characterized by steep dose gradients.

E. *Analyzed Variables*

In addition to the Gamma Passing Rate (GPR), treatment plan information was extracted to evaluate the potential influence of plan characteristics on the performance of the Gamma criteria analyzed. The following variables were considered:

- total lesion volume (mm^3);
- number of treated lesions;
- number of fractions;
- total monitor units (MU).

These variables were selected because they represent treatment plan characteristics potentially related to the geometric and dosimetric complexity of SRS treatments.

F. *Statistical Analysis*

Initially, a descriptive analysis of the Gamma Passing Rates (GPR) was performed for the 2%/2 mm and 5%/1 mm criteria. The distributions of GPR values were characterized using the median and interquartile range (IQR). Since both criteria were applied to the same 240 treatment plans, GPR values were compared using a paired analysis with the Wilcoxon signed-rank test. A significance level of 5% ($p < 0.05$) was adopted.

To evaluate the influence of lesion volume on the occurrence of plan failure, the total lesion volume was categorized using the median of the volume distribution of the 240 treatment plans included in the study as the cutoff point, corresponding to 11,950 mm^3. Accordingly, treatment plans were classified into two groups: volume ≤ 11,950 mm^3 and volume > 11,950 mm^3.

Subsequently, logistic regression analysis was performed to investigate the association between treatment plan characteristics and the occurrence of plan failure, defined as GPR < 95%, for each Gamma criterion evaluated. The results were expressed as *odds ratios* (ORs), with their respective 95% confidence intervals (95% CIs) and p-values. A statistical significance level of 5% ($p < 0.05$) was adopted.

Graphical analyses were also performed to compare the distributions of Gamma Passing Rates and to investigate their relationship with lesion volume and other treatment plan characteristics.

III. RESULTS

A total of 240 intracranial clinical treatment plans subjected to patient-specific quality assurance were analyzed using the 2%/2 mm and 5%/1 mm Gamma criteria. The comparison between the criteria demonstrated distinct Gamma Passing Rate (GPR) behaviors, both in terms of the distribution of GPR values and their association with the characteristics of the treatment plans analyzed.

A. *Comparison of Gamma Criteria*

The distribution of GPR values for both criteria is presented in Fig. 1. For the 2%/2 mm criterion, the median GPR was 94.60%, with an interquartile range of 91.78% to 96.93%. For the 5%/1 mm criterion, the median GPR was 98.10%, with an interquartile range of 96.48% to 99.33%. These results demonstrate higher central values and lower GPR dispersion for the 5%/1 mm criterion.

Since both criteria were applied to the same 240 treatment plans, a paired comparison was performed using the Wilcoxon signed-rank test, which demonstrated a statistically significant difference between the GPR values obtained with the 2%/2 mm and 5%/1 mm criteria ($p < 0.001$).

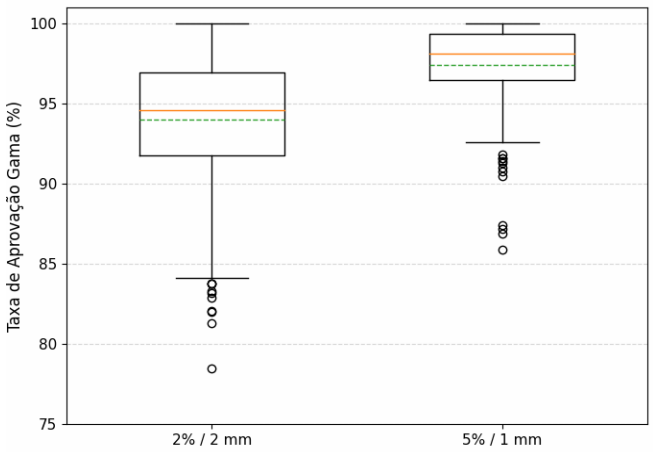


Figure 1. Boxplot of the Gamma Passing Rate (GPR) for the 2%/2 mm and 5%/1 mm Gamma criteria.

As shown in Fig. 1, the 5%/1 mm criterion exhibited a distribution more concentrated at higher GPR values and a smaller interquartile range compared with the 2%/2 mm criterion. The interquartile range was approximately 2.85 percentage points for the 5%/1 mm criterion, compared with 5.15 percentage points for the 2%/2 mm criterion, indicating lower variability in the results obtained with the 5%/1 mm criterion.

Thus, the results demonstrate distinct behaviors between the criteria, with significantly higher GPR values and lower dispersion for the 5%/1 mm criterion. However, these differences reflect the behavior of the GPR distributions in the analyzed sample and, by themselves, do not allow inference of greater sensitivity of the 5%/1 mm criterion for detecting geometric errors.

B. Influence of Lesion Volume

The relationship between total lesion volume and the Gamma Passing Rate is presented in Fig. 2. A trend toward decreasing GPR with increasing lesion volume was observed for the 2%/2 mm criterion. In contrast, for the 5%/1 mm criterion, GPR values showed less variation as a function of lesion volume, remaining more homogeneous across the range of volumes analyzed.

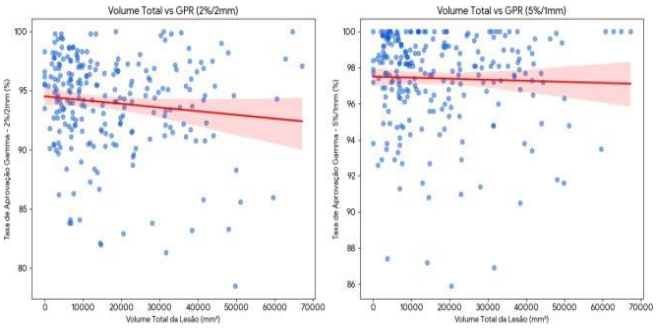


Figure 2. Relationship between total lesion volume and Gamma Passing Rate for the 2%/2 mm and 5%/1 mm criteria.

Logistic regression demonstrated that treatment plans with a volume greater than 11,950 mm³ had higher odds of failure under the 2%/2 mm criterion (OR = 1.71; $p = 0.039$), corresponding to a 71% increase in the odds of failure compared with treatment plans with a volume equal to or below this cutoff. For the 5%/1 mm criterion, no statistically significant association was observed between lesion volume and the occurrence of failure.

C. Influence of Treatment Plan Complexity

The analysis of treatment plan characteristics demonstrated an association between the number of treated lesions and the behavior of the 5%/1 mm Gamma criterion. Logistic regression showed an increase in the probability of failure associated with the number of lesions, with an *Odds Ratio* (OR) of 1.06 ($p < 0.05$).

This result indicates that treatment plans involving a greater number of lesions had a higher probability of failure under the 5%/1 mm criterion. However, this finding represents a statistical association and does not allow inference of greater sensitivity of the criterion for detecting geometric discrepancies.

Variables related to the number of monitor units (MU) and fractionation had only a limited influence on the Gamma Passing Rate, with no statistically significant differences observed between the groups evaluated.

D. Comparison of Clinical Pass Rates

Considering the clinical acceptance threshold of GPR ≥ 95%, differences in passing and failing rates were observed between the evaluated criteria. For the 2%/2 mm criterion, among the 240 treatment plans analyzed, 114 (47.5%) were classified as passing and 126 (52.5%) as failing. For the 5%/1 mm criterion, 199 treatment plans (82.9%) passed and 41 (17.1%) failed, as shown in Fig. 3.

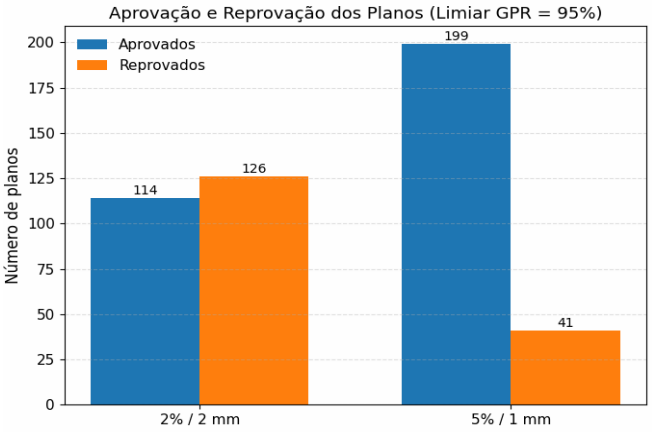


Figure 3. Number of treatment plans passing and failing the 2%/2 mm and 5%/1 mm Gamma criteria, considering GPR ≥ 95% as the acceptance threshold.

Thus, the 5%/1 mm criterion showed a higher proportion of treatment plans with GPR ≥ 95% and a lower proportion of results below this threshold compared with the 2%/2 mm criterion. This difference should be interpreted considering the distinct characteristics of the two criteria, since the 5%/1 mm criterion combines a wider dose-difference tolerance (5%) with a more restrictive spatial tolerance (1 mm), whereas the 2%/2 mm criterion applies a more restrictive dosimetric tolerance.

However, the higher passing rate observed for the 5%/1 mm criterion should not be interpreted in isolation as evidence of a greater ability to detect geometric errors, since the present study did not use a reference standard based on known errors.

E. Predictive Model for Clinical Failure

The logistic regression results for factors associated with plan failure (GPR < 95%) are summarized in Table I.

Table I – Association Between Treatment Plan Parameters and Failure (GPR < 95%) for the 2%/2 mm and 5%/1 mm Gamma Criteria

Variável	Referência	2%/2 mm – OR (IC95%)	p	5%/1 mm – OR (IC95%)	p
Volume total da lesão	Baixo volume	**1,71 (1,03–2,86)**	**0,039**	1,34 (0,68–2,64)	0,392
Complexidade do alvo	Lesão única	1,02 (0,49–2,11)	0,956	1,96 (0,84–4,58)	0,121
Fracionamento	1 fração	1,55 (0,93–2,59)	0,096	0,81 (0,41–1,60)	0,541
Total de UM	Baixa modulação	1,02 (0,60–1,73)	0,930	1,88 (0,95–3,70)	0,069

The model demonstrated that lesion volume was the main factor associated with failure under the 2%/2 mm criterion (OR = 1.71; 95% CI: 1.03–2.86; p = 0.039), whereas no statistically significant association was observed between lesion volume and failure under the 5%/1 mm criterion (OR = 1.34; 95% CI: 0.68–2.64; p = 0.392). The other variables analyzed, including target complexity, fractionation, and total monitor units, showed no statistically significant association with failure for either Gamma criterion evaluated.

IV. Discussão

The results of this study demonstrated that the 2%/2 mm and 5%/1 mm Gamma criteria exhibited distinct behaviors in patient-specific quality assurance of intracranial radiosurgery treatment plans using VMAT. Although both criteria showed high overall passing rates, differences were observed in the distribution of *Gamma Passing Rate* (GPR) values and in the influence of treatment plan characteristics on these results. Overall, the 5%/1 mm criterion showed lower dispersion of GPR values and greater stability across variations in lesion volume, whereas the 2%/2 mm criterion showed a stronger association between increasing irradiated volume and decreasing passing rates. This behavior reinforces that Gamma analysis results depend on the combination of dose-difference and distance-to-agreement parameters adopted, as well as on the characteristics of the dose distribution being analyzed [8], [9].

The analysis of volume influence demonstrated that treatment plans with volumes greater than 11,950 mm³ had a higher probability of failure when evaluated using the 2%/2 mm criterion (OR = 1.71; $p = 0.039$). This result indicates a significant association between irradiated volume and the performance of this criterion in the analyzed sample. The literature demonstrates that GPR may be influenced by different aspects of the analysis methodology, including the spatial distribution of evaluated points, measurement system resolution, dose normalization, low-dose threshold, and the criteria used for Gamma calculation [9], [10]. One possible explanation for the observed behavior is the greater contribution of low-dose-gradient regions to the global GPR calculation. In these regions, small dosimetric differences arising from characteristics of the measurement system, dose calculation, or delivery itself may contribute to a reduction in the passing rate as the evaluated area increases. However, because the present study did not individually investigate the origin of discrepancies between calculated and measured distributions, the mechanism responsible for this association cannot be definitively established.

In contrast, the 5%/1 mm criterion showed lower dependence on lesion volume, maintaining more homogeneous GPR values across the range of volumes analyzed. This behavior may be related to the combination of a wider dosimetric tolerance of 5% with a more restrictive spatial tolerance of 1 mm. Gamma analysis simultaneously combines the dose-difference and distance-to-agreement components; therefore, changes in these parameters may alter the proportion of points classified as passing [8], [9]. The wider dose-difference tolerance may reduce the contribution of small dosimetric variations to plan failure, whereas the shorter distance-to-agreement makes the criterion conceptually more restrictive in terms of spatial agreement. However, the results of the present study directly demonstrate only greater GPR stability across variations in lesion volume, and these data alone do not allow the conclusion that the 5%/1 mm criterion has a greater ability to detect geometric errors.

Another aspect evaluated was the relationship between the number of lesions and the behavior of the Gamma criteria. Treatment plans involving multiple lesions may present greater geometric and modulation complexity, including small fields, steep dose gradients, and dynamic multileaf collimator motion. MLC positioning accuracy is an important component of dosimetric accuracy in modulated treatments, and studies involving deliberately introduced errors have demonstrated that the ability to detect these changes depends on both the magnitude of the error and the Gamma criterion employed [11], [12]. In the present analysis, the number of lesions was associated with failure under the 5%/1 mm criterion (OR = 1.06; $p < 0.05$). This result indicates that the behavior of the 5%/1 mm criterion may be influenced by characteristics related to treatment plan complexity. However, the number of lesions is only an indirect indicator of complexity and does not represent a reference standard for the presence of geometric errors. Therefore, this association should not be interpreted as direct evidence of greater sensitivity of the 5%/1 mm criterion for detecting positioning or delivery errors.

Variables related to the number of monitor units and fractionation had only a limited influence on the Gamma Passing Rate, with no statistically significant differences between the groups evaluated. These findings suggest that, in the analyzed sample, these characteristics were not determinant factors in the behavior of the evaluated criteria. Multicenter experience in quality assurance of modulated treatments has demonstrated that the performance of dosimetric verification may vary according to treatment plan characteristics, treatment planning and delivery systems, and measurement methodologies [10]. On the other hand, the influence observed for volume and number of lesions in the present study reinforces that GPR interpretation should consider the specific characteristics of treatment plans rather than relying solely on an isolated passing-rate value.

These results should be interpreted considering the inherent characteristics and limitations of Gamma analysis. Since its introduction, the Gamma index has become one of the main tools for quantitative comparison between calculated and measured dose distributions [8], [9]. However, several studies have demonstrated that GPR alone has limitations in identifying and characterizing clinically relevant dosimetric discrepancies [13]. Nelms *et al.* [13], for example, evaluated treatment plans containing deliberately introduced errors and demonstrated that Gamma passing rates obtained from planar verification did not consistently predict clinically relevant dosimetric changes in the patient. These findings reinforce that a high Gamma passing rate should not be interpreted in isolation as evidence of the

absence of clinically significant errors.

This interpretation is consistent with the recommendations of AAPM TG-218, which emphasizes that the performance of Gamma analysis depends on several methodological parameters, including dose normalization, spatial resolution, low-dose threshold, comparison methodology, and dose-difference and distance-to-agreement criteria [9]. Therefore, different combinations of these parameters may produce distinct behaviors for the same dose distribution. In stereotactic treatments, where steep dose gradients, small field sizes, and high conformity increase the importance of geometric accuracy, the use of more restrictive spatial tolerances may be relevant to the verification process [12], [14].

In this context, the *AAPM-RSS Medical Physics Practice Guideline 9.b* emphasizes the high level of geometric accuracy required for stereotactic radiosurgery and stereotactic body radiation therapy procedures [14]. Therefore, the use of 1 mm as the spatial component of the Gamma criterion is conceptually aligned with the need for stringent geometric accuracy in these treatments. However, this rationale does not, by itself, imply that the 5%/1 mm criterion is more sensitive to geometric errors than the 2%/2 mm criterion. Demonstrating such capability requires comparing Gamma results against a reference standard containing errors of known nature and magnitude.

Studies involving the deliberate introduction of errors demonstrate the importance of this type of approach. Yan *et al.* [11] introduced random errors of up to 2 mm and systematic errors of ±1 and ±2 mm in MLC leaf positions and observed that QA sensitivity varied according to the Gamma criterion employed. Similarly, Kim *et al.* [12], in a PSQA investigation of VMAT-based SBRT using FFF beams, introduced known MLC misalignments and evaluated different combinations of dose-difference and distance-to-agreement criteria. The authors demonstrated that criteria containing a 1-mm spatial component exhibited distinct behavior in response to the introduced perturbations, highlighting that the sensitivity of a given criterion should be experimentally established using known errors [11], [12].

An important limitation of the present study is therefore the absence of a *ground truth* consisting of known geometric or dosimetric errors. Because this was a retrospective analysis of clinical treatment plans, controlled perturbations such as spatial shifts, deliberate changes in multileaf collimator positioning, or known modifications in delivered dose were not introduced. Studies using deliberately introduced errors demonstrate the importance of this approach for evaluating the actual ability of QA metrics to distinguish error-free dose distributions from those containing known discrepancies [11]–[13]. Therefore, the differences observed between the 2%/2 mm and 5%/1 mm criteria in the present study should be interpreted as differences in GPR behavior and in their association with treatment plan characteristics, and it is not possible to directly determine the sensitivity or specificity of either criterion for error detection.

This limitation should also be considered when interpreting the GPR ≥ 95% threshold adopted in this study. The same reference value was applied to the 2%/2 mm and 5%/1 mm criteria to enable standardized comparison of the results. TG-218 provides recommendations regarding tolerance and action limits for PSQA and emphasizes the importance of standardizing the conditions used for GPR calculation and interpretation [9]. However, in the absence of a reference standard containing treatment plans with and without known errors, the present study cannot determine whether 95% represents the optimal decision threshold for the 5%/1 mm criterion. Establishing a criterion-specific threshold would require an analysis based on the ability to discriminate between known acceptable and unacceptable conditions.

Future studies may address this issue through the systematic introduction of controlled perturbations into treatment plans or measurements, including spatial shifts of different magnitudes, known alterations in MLC positioning, and deliberate dosimetric variations. This methodology has been used to investigate the ability of different PSQA approaches to identify delivery errors [11]–[13]. Particularly in stereotactic treatments using VMAT, the introduction of submillimeter and millimeter-scale MLC errors may provide relevant information regarding the discriminatory ability of Gamma criteria with spatial tolerances of 1 and 2 mm [12]. Comparing unperturbed treatment plans with plans containing known errors would allow direct evaluation of the error-detection capability of the Gamma criteria. Using such a reference standard, sensitivity and specificity could be calculated and receiver operating characteristic (ROC) curves could be constructed, allowing different GPR thresholds to be evaluated and determining whether the 95% threshold is appropriate for the 5%/1 mm criterion.

In addition to this limitation, the study was retrospective and conducted at a single radiotherapy center using a specific treatment planning, delivery, and dosimetry platform. PSQA performance may depend on the treatment planning system, dose calculation algorithm, treatment equipment, detector characteristics, and analysis methodology [9], [10]. Although the analysis of 240 clinical treatment plans provides a representative sample of the institutional experience, multicenter studies involving different linear accelerators, treatment planning systems, and dosimetry devices are needed to assess the generalizability of the results. Future investigations may also evaluate the behavior of the 5%/1 mm criterion in other stereotactic scenarios, including extracranial SBRT treatments.

Taken together, the results of this study demonstrate that the choice of Gamma criterion influences the interpretation of patient-specific quality assurance in radiosurgery. In the analyzed sample, the 5%/1 mm criterion showed less influence of irradiated volume and greater stability of the passing rate, whereas the 2%/2 mm criterion showed a stronger association between volume and the occurrence of failure. Considering the known limitations of GPR as a standalone metric [9], [13], these findings should be interpreted as evidence of the behavior of the two criteria within the PSQA process evaluated. The results support further investigation of the 5%/1 mm criterion as an alternative for PSQA of stereotactic treatments; however, demonstrating superiority in geometric error detection and defining a criterion-specific acceptance threshold require additional studies based on controlled perturbations and a known reference standard.

V. *CONCLUSION*

This study evaluated the performance of the 5%/1 mm Gamma criterion in patient-specific quality assurance (PSQA) of intracranial stereotactic radiosurgery treatment plans using VMAT, comparing it with the institutional 2%/2 mm criterion. Both criteria showed high overall passing rates; however, they exhibited distinct behaviors in relation to the characteristics of the treatment plans analyzed.

The 2%/2 mm criterion showed a stronger association between increasing irradiated volume and

decreasing Gamma Passing Rate (GPR), with a higher probability of failure in treatment plans involving larger lesion volumes. In contrast, the 5%/1 mm criterion showed less influence of irradiated volume and greater stability of the passing rate across the range of volumes evaluated.

The results also demonstrated an association between characteristics related to treatment plan complexity and the behavior of the 5%/1 mm criterion. However, since this study exclusively used retrospective clinical treatment plan data and did not include the introduction of known geometric or dosimetric errors, it is not possible to directly conclude that the 5%/1 mm criterion has greater sensitivity or specificity for detecting such errors.

Therefore, the findings indicate that the 5%/1 mm criterion shows promising behavior for application in patient-specific quality assurance of SRS treatments using VMAT, particularly because of its lower dependence on the volumetric variations observed in the study sample. The more restrictive spatial tolerance of 1 mm may represent a potentially relevant feature for stereotactic treatments, in which geometric accuracy is critical; however, this capability should be confirmed through specific error-detection studies.

Future investigations involving the controlled introduction of spatial shifts, alterations in multileaf collimator (MLC) positioning, and known dosimetric variations may directly evaluate the error-detection capability of each criterion. Such analyses would allow the determination of parameters such as sensitivity and specificity, as well as the construction of receiver operating characteristic (ROC) curves, in addition to investigating the suitability of the GPR ≥ 95% threshold for the 5%/1 mm criterion.

Thus, the results of this study provide clinical evidence regarding the comparative behavior of the 2%/2 mm and 5%/1 mm Gamma criteria and may contribute to the development of more specific quality assurance protocols for stereotactic radiosurgery. However, in the absence of a reference standard based on known errors, these findings do not establish diagnostic superiority of the 5%/1 mm criterion.

ACKNOWLEDGMENT

The authors would like to thank the technical support provided by the Radiotherapy Department of the Association for Cancer Control in Goiás and the academic infrastructure made available by the Institute of Physics of the Federal University of Goiás, which were essential for the development of this work.

REFERENCES

[1] O. Green, S. L. Grimsley, K. C. Yuan, and P. J. Parikh, "Volumetric Modulated Arc Therapy (VMAT): A review of clinical outcomes and treatment efficiency," *Journal of Applied Clinical Medical Physics*, vol. 22, no. 8, pp. 1–15, 2021.

[2] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," *Medical Physics*, vol. 25, no. 5, pp. 656–661, 1998.

[3] C. Miften *et al.*, "Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218," *Medical Physics*, vol. 45, no. 4, pp. e53–e83, 2018.

[4] B. E. Nelms and J. F. Simon, "A survey on planar IMRT QA analysis," *Journal of Applied Clinical Medical Physics*, vol. 8, no. 3, pp. 76–90, 2007.

[5] J. E. Ezzell *et al.*, "Guidance document on delivery, treatment planning, and clinical implementation of IMRT," *Medical Physics*, vol. 30, no. 8, pp. 2089–2115, 2003.

[6] S. J. Thomas *et al.*, "AAPM Medical Physics Practice Guideline 9.b: Stereotactic radiosurgery," *Journal of Applied Clinical Medical Physics*, vol. 25, 2024.

[7] M. Hussein *et al.*, "A global evaluation of pretreatment radiotherapy quality assurance," *Radiotherapy and Oncology*, vol. 126, pp. 290–295, 2018.

[8] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," *Med. Phys.*, vol. 25, no. 5, pp. 656–661, May 1998, doi: 10.1118/1.598248.

[9] M. Miften *et al.*, "Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218," *Med. Phys.*, vol. 45, no. 4, pp. e53–e83, Apr. 2018, doi: 10.1002/mp.12810.

[10] G. A. Ezzell *et al.*, "IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," *Med. Phys.*, vol. 36, no. 11, pp. 5359–5373, Nov. 2009, doi: 10.1118/1.3238104.

[11] G. Yan, C. Liu, T. A. Simon, L.-C. Peng, C. Fox, and J. G. Li, "On the sensitivity of patient-specific IMRT QA to MLC positioning errors," *J. Appl. Clin. Med. Phys.*, vol. 10, no. 1, pp. 120–128, 2009, doi: 10.1120/jacmp.v10i1.2915.

[12] J.-I. Kim, S.-Y. Park, H. J. Kim, J. H. Kim, S.-J. Ye, and J. M. Park, "The sensitivity of gamma-index method to the positioning errors of high-definition MLC in patient-specific VMAT QA for SBRT," *Radiat. Oncol.*, vol. 9, Art. no. 167, Jul. 2014, doi: 10.1186/1748-717X-9-167.

[13] B. E. Nelms, H. Zhen, and W. A. Tomé, "Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors," *Med. Phys.*, vol. 38, no. 2, pp. 1037–1044, Feb. 2011, doi: 10.1118/1.3544657.

[14] E. Cirino *et al.*, "AAPM-RSS Medical Physics Practice Guideline 9.b: SRS-SBRT," *J. Appl. Clin. Med. Phys.*, 2025, doi: 10.1002/acm2.14624

Análise Comparativa entre o Tremor de Repouso no Dedo e a MMG no Antebraço da Doença de Parkinson

Ederson Cichaczewski
PPGTS
PUCPR
Curitiba-PR, Brasil
c.ederson@pucpr.edu.br

Luigi Tortato de Souza
Engenharia Biomédica
PUCPR
Curitiba-PR, Brasil
luigi.souza@pucpr.edu.br

Munir Antonio Gariba
PPGTS
PUCPR
Curitiba-PR, Brasil
munir.gariba@pucpr.br

Percy Nohama
PPGTS
PUCPR
Curitiba-PR, Brasil
percy.nohama@pucpr.br

***Resumo* — A Doença de Parkinson (DP) é caracterizada por sintomas motores cardinais cuja avaliação clínica tradicional, normalmente baseada na escala UPDRS, que apresenta limitações de subjetividade e baixa sensibilidade para alterações sutis. Como alternativa a mecanomiografia (MMG) surge como uma ferramenta promissora para o monitoramento contínuo. Com o objetivo de comparar as características de amplitude e frequência dos sinais da MMG adquiridos no antebraço, contrastando o tremor muscular com o tremor visível do dedo indicador em repouso, este trabalho utilizou dados de um estudo experimental com voluntários diagnosticados com DP. A metodologia consistiu em uma revisão bibliográfica e na análise experimental de 4 pacientes selecionados, nos quais foram coletados sinais simultâneos em repouso por 60 s. Utilizou-se um acelerômetro bidimensional MEMS de alta sensibilidade (±1,5 g; 800 mV/g) sobre os músculos flexores do antebraço para a MMG, e outro sensor (±6 g; 200 mV/g) na ponta do dedo indicador. Os sinais foram filtrados, digitalizados com taxa de amostragem de 280 Hz e analisados no domínio do tempo (RMS) e da frequência (FFT). Os resultados indicaram que as frequências de tremor de repouso situaram-se de forma consistente na banda de 4 a 6 Hz. A amplitude RMS da MMG mostrou-se sistematicamente menor que a do dedo, chegando a ser 10 vezes menor em um dos voluntários. Em contrapartida, as frequências da MMG foram ligeiramente superiores (diferença de até 0,570 Hz), o que confirma que o tremor mecânico periférico é amortecido pela inércia física do membro. Conclui-se que a MMG no antebraço provou ser um método viável e clinicamente informativo, evidenciando potencial para caracterizar assinaturas motoras de forma objetiva e abrir caminho para sistemas de suporte à decisão clínica baseados em inteligência artificial e tratamentos personalizados.**

Palavras-chave — doença de parkinson, mecanomiografia, tremor, acelerômetro.

I. INTRODUÇÃO

A Doença de Parkinson (DP) é a segunda doença neurodegenerativa mais prevalente no mundo e é caracterizada primordialmente pela morte progressiva de neurônios dopaminérgicos na substantia nigra do cérebro. Essa degeneração neuronal manifesta-se através de sintomas motores cardinais que incluem tremor de repouso, bradicinesia e rigidez muscular [1]. À medida que a doença progride, o tratamento, que frequentemente envolve medicação dopaminérgica (como a Levodopa) ou Estimulação Cerebral Profunda (DBS), necessita de ajustes contínuos e personalizados [2].

Até o presente, a quantificação da severidade dos sintomas motores da DP é predominantemente realizada em ambiente clínico, normalmente utilizando a *Unified Parkinson's Disease Rating Scale* (UPDRS) da *Movement Disorder Society* (MDS). Contudo, a UPDRS apresenta limitações severas: é uma escala ordinal inerentemente subjetiva, propensa à variabilidade inter e intra-avaliador, e possui um "efeito teto" que reduz sua sensibilidade às mudanças sutis nos estágios iniciais ou prodrômicos da doença [3]. Além disso, a avaliação clínica provê apenas um retrato momentâneo do paciente, falhando em capturar as flutuações motoras que ocorrem ao longo do dia.

Para contornar essas limitações, a Eletromiografia (EMG) tem sido historicamente utilizada. No entanto, a EMG apresenta desvantagens práticas, como a necessidade de preparação meticulosa da pele, sensibilidade à impedância cutânea (suor, pelos) e alta suscetibilidade a artefatos elétricos [4]. Como alternativa para monitoramento contínuo e vestível (wearable), destaca-se a Mecanomiografia (MMG). A MMG registra as oscilações mecânicas de baixa frequência na superfície muscular produzidas pela ativação das fibras musculares, refletindo a contração e o recrutamento das unidades motoras. Sendo um sinal mecânico, a MMG não sofre interferência da impedância da pele, não necessita de gel condutor e possui uma relação sinal-ruído vantajosa, permitindo sua fácil integração em sistemas de telemedicina e dispositivos vestíveis [5]. A detecção dessas microvibrações mecânicas fornece um espelho direto do comando oscilatório central, sendo essencial para a identificação de anomalias motoras na DP.

A avaliação objetiva dos sintomas motores da Doença de Parkinson (DP) tem motivado a transição de métodos puramente clínicos para o uso de biossensores vestíveis (wearables) [6]. Trabalhos recentes propuseram a fusão da MMG com Unidades de Medição Inercial (IMUs) para criar perfis motores completos. Em [2] foi desenvolvido um conjunto heterogêneo de sensores fixados no braço e antebraço capaz de avaliar simultaneamente bradicinesia, rigidez e tremor, demonstrando que características extraídas da MMG correlacionam-se intimamente com as pontuações

clínicas da UPDRS. Visando o monitoramento contínuo em ambiente domiciliar (telemedicina), em [6] foi desenvolvido o sistema PDTrack. Seus resultados revelaram que a adição da MMG às medições puramente inerciais (acelerômetros e giroscópios) aumenta significativamente a precisão da classificação de sintomas, especialmente para tremores posturais, ressaltando o valor da atividade mecânica muscular dissociada da cinemática das articulações.

Além do tremor, a lentidão do movimento (bradicinesia) tem sido investigada de forma isolada pela MMG. Em [1] foram extraídas informações espectrais dos sinais de MMG baseados em acelerômetros durante tarefas motoras como pronação/supinação e beliscão dos dedos (*finger pinch*). Por meio do cálculo da Frequência Média Instantânea (IMNF) usando o Espectro de Hilbert, os autores confirmaram que pacientes em estágios mais severos de bradicinesia apresentam valores significativamente menores de IMNF, estabelecendo um limiar detectável para o estadiamento do sintoma.

Uma vertente investigou pacientes com Parkinson sem a presença do tremor clínico aparente. Em [3] foi medido a MMG durante tarefas isométricas bilaterais de membros superiores e constataram que os pacientes com DP apresentaram variações de amplitude reduzidas e uma distribuição anormal da potência de frequência mecânica em relação a pessoas saudáveis. Em um estudo subsequente focando em tarefas motoras isométricas unilaterais, os mesmos autores evidenciaram que pacientes sem tremor clínico já apresentam assimetrias acentuadas entre os membros (inter-membros) na variação das amplitudes de MMG e nas médias de frequência, sugerindo que as micro-oscilações musculares precedem a manifestação visual do sintoma e atuam como um marcador patológico sutil.

Avanços no hardware de aquisição permitiram explorar a MMG em níveis microscópicos. Em [7] foi introduzido o uso de acelerômetros de altíssima sensibilidade com componentes em ouro, reduzindo o piso de ruído e capturando frequências acima de 15 Hz ocultadas por sensores comerciais convencionais. Esses pesquisadores detectaram a MMG diretamente no músculo polegar de pacientes com DP e alimentaram esses sinais, em conjunto com componentes de baixa frequência (tremor clássico), numa arquitetura de rede neural profunda (MLSTM-FCNs). A fusão dessas frequências atingiu precisão de 92,19% na classificação entre indivíduos saudáveis e pacientes com DP, estabelecendo os tremores microscópicos da fibra muscular como uma fronteira vital de pesquisa para o diagnóstico.

O presente trabalho tem o objetivo de avaliar a eficácia, a sensibilidade e a aplicabilidade clínica da mecanomiografia (MMG), baseada em sensores de aceleração, para a quantificação e caracterização objetiva de tremores em indivíduos com Doença de Parkinson, visando contrastar o tremor inerente à mecânica das fibras musculares com o tremor visível do segmento distal (mão e dedos).

II. METODOLOGIA

A. Revisão bibliográfica

Para garantir a abrangência e o rigor da fundamentação teórica, foi conduzida uma revisão bibliográfica, com foco em estudos prévios do uso da MMG na Doença de Parkinson.

A busca foi realizada nas bases de dados IEEE Xplore, PubMed, ScienceDirect e SpringerLink. Foram utilizados descritores combinados (operadores booleanos): ("Mechanomyography" OR "MMG") AND ("Parkinson's Disease" OR "Tremor" OR "Bradykinesia").

Os critérios de inclusão abrangeram estudos clínicos e de engenharia que propusessem sistemas de aquisição, processamento de sinais mecânicos (acelerômetros, sensores piezelétricos ou microfones) e algoritmos de classificação de sintomas motores da DP. Artigos focados exclusivamente em modelos animais ou métodos puramente baseados em EMG foram excluídos.

A revisão focou na extração de dados sobre tipos de sensores, posicionamento, parâmetros extraídos (ex: Frequência Média Instantânea - IMNF, Root Mean Square - RMS) e a eficácia na distinção entre pacientes com DP e controles saudáveis.

B. Protocolo Experimental e Dados

Os dados analisados neste trabalho são provenientes de um estudo experimental desenvolvido durante o mestrado do primeiro autor, envolvendo pacientes voluntários do Ambulatório de Distúrbios do Movimento do Hospital de Clínicas da UFPR. O protocolo de pesquisa foi devidamente aprovado pelo Comitê de Ética em Pesquisa (CEP) da instituição, sob o parecer de aprovação nº 2005018128. Todos os participantes foram previamente informados sobre o Termo de Consentimento Livre e Esclarecido (TCLE) e o assinaram, concordando com a participação no estudo.

A medição do tremor visível de repouso foi realizada na extremidade distal do dedo indicador, devido este segmento refletir a manifestação macroscópica final do tremor parkinsoniano clássico, e da MMG foi realizada nos ventres musculares do antebraço, devido a essa região central apresentar a maior deflexão transversal das fibras musculares. O sensor do antebraço foi posicionado no ponto médio da face anterior do antebraço, sobre os músculos flexores do carpo. Utilizou-se para a MMG um acelerômetro bidimensional (2D) modelo MMA6260Q da Freescale baseado em tecnologia MEMS, especificado com um fundo de escala de ±1,5g e uma sensibilidade de 800mV/g, alimentado com 3,3V. Esta configuração de hardware é compatível com tecnologias validadas na literatura, como em [3] e [8], para a extração ideal do sinal de MMG. Para a medição do tremor visível na ponta do dedo indicador foi utilizado um acelerômetro bidimensional (2D) modelo MMA6270 da Freescale configurado para o fundo de escala de ±6g e uma sensibilidade de 200mV/g, alimentado com 3,3V. Os sinais analógicos dos acelerômetros foram condicionados por meio de um filtro passa-altas de 0,3Hz, um filtro passa-baixas de 20 Hz de 2ª ordem e digitalizados por meio de um conversor Analógico-Digital (ADC) utilizando-se 8 bits de quantização com faixa dinâmica de 3,3V de entrada a uma taxa de amostragem de 280 Hz por canal, para evitar *aliasing*. Foi utilizado ganho unitário aos dois sensores e os sinais analógicos tiveram o seu nível de referência deslocado para 1,65V antes de entrar no conversor A/D.

A captura dos sinais foi realizada com duração de 60 s, adquirindo os sinais do sensor do dedo indicador e do antebraço no mesmo momento, com defasagem temporal de 0,89 ms entre canais consecutivos, fazendo-se uma varredura por meio de diferentes canais do ADC em cada eixo X e Y dos dois acelerômetros, utilizando-se o microcontrolador modelo MSP430F135 da Texas Instruments. O hardware ainda conta com isolação elétrica do lado dos sensores por meio do

conversor DC-DC isolado NMH0505D para a alimentação e optoacopladores ACSL-6400 para os dados.

O cálculo do valor RMS da amplitude do sinal do tremor no domínio do tempo em unidade de aceleração da gravidade (g) foi realizado a partir da conversão prévia dos sinais em Volts dividindo-se pela sensibilidade nominal de cada acelerômetro. Foi feita a calibração em cada eixo dos dois sensores utilizando a máquina MTS 810, cuja configuração de deslocamento, em m, por unidade de tempo, em s, permitiu a conversão para aceleração, em m/s², e para aceleração da gravidade, em g, sem saturação. O sensor MMA6260Q que capta vibrações na faixa de ±1,5g a uma sensibilidade de 800mV/g, foi posicionado no antebraço por sua maior sensibilidade, adequada para captar sinais de menor amplitude. O sensor MMA6270Q que capta vibrações na faixa de ±6g a uma sensibilidade de 200mV/g, foi posicionado no dedo indicador por sua maior faixa dinâmica, adequada para sinais de maior amplitude. Por exemplo, para um deslocamento de 0,5mm a 20Hz da máquina MTS 810, equivalente a 0,81g de pico, foi obtido 0,64V na saída do sensor MMA6260Q e 0,16V na saída do sensor MMA6270Q, confirmando as sensibilidades e a proporção de 4x entre elas. Após a normalização ambos os sinais foram expressos em g RMS, permitindo uma comparação direta.

Os dados aquiridos foram armazenados em base de dados MySQL no computador. Os sinais ainda passaram por uma etapa de processamento digital por meio de um filtro passa-baixas FIR com aproximação Chebychev de ordem 11 e frequência de corte de 20 Hz, projetado com 12 coeficientes. Foi realizada a aplicação de algoritmo de FFT para o cálculo da frequência em Hertz do sinal no domínio da frequência.

A fixação dos acelerômetros foi realizada com fita adesiva do tipo fita crepe estreita e fitas de velcro para melhor sustentação e evitar o descolamento da fita crepe.

Os pacientes voluntários com diagnóstico de Doença de Parkinson foram instruídos a realizar uma tarefa motora estática durante o intervalo de 60 s, que avalia o tremor de repouso, tendo os braços relaxados e apoiados sobre as pernas, com a palma das mãos para cima. A amostra analisada constitui-se de uma seleção de conveniência de 4 voluntários a partir de uma coorte heterogênea, como limitação do estudo.

III. Resultados e Discussão

A. *O estado da arte e lacunas de pesquisa*

A aplicação da revisão bibliográfica revelou avanços no uso da MMG para a DP, delineando claramente o que a engenharia biomédica já conquistou e o que ainda requer investigação profunda.

O trabalho de [2] demonstrou pioneirismo ao desenvolver um conjunto de sensores heterogêneos (incorporando 4 sensores MMG, IMUs e sensores de força) para quantificar simultaneamente bradicinesia, rigidez e tremor. O sistema alcançou 85,4% de precisão na previsão das pontuações UPDRS clínicas, provando que a MMG pode até mesmo substituir sensores de força na avaliação da rigidez articular.

Em [1] foi mostrado que a extração da Frequência Média Instantânea (IMNF) a partir do espectro de Hilbert do sinal de MMG é um marcador robusto. Eles observaram que pacientes com sintomas mais severos de DP apresentam valores de IMNF mais baixos durante tarefas de pronação/supinação em relação aos pacientes em estágios iniciais.

Em [3] e [8] foram investigados pacientes com DP sem tremor clínico visível, utilizando MMG durante tarefas motoras isométricas. Os autores identificaram que a variação de amplitude e a relação potência-frequência (*power-frequency-ratio*) dos sinais MMG diferem significativamente dos controles saudáveis, além de revelarem assimetrias entre os membros, servindo como uma possível "janela" precoce para a disfunção oscilatória dos gânglios da base.

Em [7] foi introduzido o conceito de "micro-MMG". Usando acelerômetros de alta sensibilidade (com massa de prova em ouro), descobriram vibrações acima de 15 Hz que são ocultadas pelo piso de ruído (*noise floor*) dos sensores comerciais comuns. Alimentando esses dados em redes neurais profundas (MLSTM-FCNs), alcançaram 92,19% de precisão na classificação entre DP e indivíduos saudáveis.

Apesar desses achados, o entendimento neurofisiológico exato das oscilações musculares de baixa amplitude (MMG) ainda necessita de mais investigações científicas. Ademais, a padronização no posicionamento de sensores de MMG e a relação com o tremor visível de repouso e outros biomarcadores exigem novos avanços.

B. *Tremor Visível no Dedo vs. Tremor Muscular no Antebraço*

Na prática clínica rotineira, o tremor da Doença de Parkinson é frequentemente visualizado e quantificado cinematicamente nas extremidades distais (ex: movimentos de "contar moedas" nos dedos). O tremor visível nos dedos é o resultado final macroscópico de um sistema mecânico em cascata, englobando múltiplas articulações, inércia dos ossos e a soma de forças de músculos agonistas e antagonistas.

O foco deste artigo na MMG do antebraço permitiu uma análise em um nível biomecânico muito mais profundo e precoce. O sinal medido no ventre muscular do antebraço captura as vibrações transversais individuais geradas diretamente pelo recrutamento, disparo e encurtamento das unidades motoras ativadas por comandos supraspinais.

Medir no antebraço revela as oscilações patológicas do Sistema Nervoso Central (via gânglios da base e núcleo subtalâmico) mesmo quando a inércia e a rigidez mecânica da mão impedem que o dedo se mova fisicamente. Como corroborado por [3] e [8], alterações na frequência e na amplitude da MMG ocorrem no antebraço de pacientes em estágios onde não há tremor visível. Portanto, o tremor muscular medido via MMG pode ser um biomarcador promissor, conforme verificado em pacientes com diagnóstico conhecido, requerendo validação com amostras mais amplas.

C. *Resultados Experimentais em Voluntários*

Entre os dados da base dos testes realizados em pacientes voluntários do ambulatório de Distúrbios do Movimento, devido à variação de perfis entre os voluntários, que inclui diversos tipos de sintomas e outros tipos de tremores além da DP, como tremor essencial, e tendo em vista que não houve um controle se o paciente estava sob efeito medicação ou não, considerando aqueles que apresentaram características mais fidedignas ao diagnóstico da Doença de Parkinson, sem outros sintomas ou complicações adicionais, foram selecionados 4 voluntários para este estudo. A análise estatística pareada confirmou que a frequência do sinal de MMG no antebraço foi estatisticamente superior à do tremor no dedo indicador, teste *t* pareado, $t(3) = 3,80$, $p = 0,032$. Para a amplitude RMS, a magnitude no dedo indicador mostrou-se superior à da MMG

no antebraço em todos os indivíduos analisados (p = 0,007 para dados log-transformados). A Tabela 1 apresenta os resultados de amplitude e frequência dos tremores de cada voluntário selecionado.

Tabela 1. Valores de amplitude e frequência de tremor dos voluntários selecionados com Doença de Parkinson.

V	MMG (Antebraço)		Tremor (Dedo Indicador)	
	Amplitude (g)	Frequência (Hz)	Amplitude (g)	Frequência (Hz)
1	0,153	4,076	0,210	3,852
2	0,026	5,182	0,037	4,612
3	0,084	5,198	0,252	4,802
4	0,140	4,249	1,421	4,076

V: voluntário

Esta avaliação dos 4 pacientes voluntários com DP submetidos à medição dos acelerômetros de duas dimensões apresentou resultados espectrais consistentes com a literatura, com a frequência do tremor de repouso ficando aproximadamente na faixa de 4 e 6 Hz.

Pode-se verificar nos dados da Tabela 1 que a amplitude RMS da vibração muscular apresenta valores menores que a amplitude do tremor do dedo indicador, demonstrando que a força de aceleração transversal à superfície do antebraço é baixa, mas detectável, ainda, tendo em vista que os distintos tremores são fenômenos físicos de naturezas distintas, corroborando com os achados de [7]. A diferença entre a intensidade do tremor do dedo indicador e do tremor muscular no antebraço chegou a ser 10 vez maior no voluntário 4.

Olhando para a frequência dos sinais, verificou-se que a frequência do sinal de MMG é sistematicamente um pouco maior, com diferença de até 0,570 Hz no voluntário 2. Este fenômeno pode ser explicado devido à MMG registrar de forma direta e imediata as oscilações musculares; já o tremor do movimento do dedo é amortecido pela inércia física do membro, reduzindo ligeiramente a frequência do tremor visível. Esta natureza distinta foi relatada por [1] e [3].

IV. Conclusões

Este estudo evidenciou que a MMG pode ser usada na avaliação de tremores da Doença de Parkinson, trazendo parâmetros que contribuem com o entendimento dos sintomas da doença. Como os voluntários selecionados possuíam diagnóstico conhecido da DP, ainda é preliminar poder afirmar que a MMG pode ser usada para auxiliar no diagnóstico em estágios iniciais quando o tremor nos dedos ainda não é visível, mas demonstra potencial para uma avaliação mais precoce.

Em relação à diferença entre a amplitude da vibração muscular medido no antebraço e a amplitude do tremor no dedo, pode-se associar a outros sintomas da DP, como a bradicinesia e a rigidez muscular, visto que podem atenuar o tremor visível do dedo, sugerindo um estreitamento da diferença de amplitude entre o MMG e o tremor no dedo.

O potencial de desdobramento das investigações para o diagnóstico da DP com sensores inerciais ainda é muito vasto. A quantificação dos tremores por meio de sensores, seja em nível de MMG ou de tremor visível, pode estabelecer características que funcionem como uma assinatura de tremor que podem servir de realimentação para estratégias medicamentosas ou de ajuste de dispositivos de estimulação cerebral (DBS) para o controle da doença.

Pode-se ampliar aquisição de sinais para dados de giroscópios, para contribuir ainda mais com informações sobre a velocidade angular e alterações nos ângulos durante os movimentos, que virão a contribuir ainda mais no diagnóstico, constituindo-se de uma fusão multimodal.

Complementarmente, uma correlação do tremor com outros biomarcadores de naturezas diferentes e características específicas mapeadas para a DP, como EMG ou EEG, tem potencial de avanços significativos no auxílio ao diagnóstico. Os dados adquiridos em voluntários com diagnóstico conhecido podem ser utilizados para o treinamento de sistemas de suporte à decisão clínica (CDSS), utilizando técnicas de *Deep Learning* e inteligência artificial.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Brasil pelo financiamento ao desenvolvimento deste estudo. EC é bolsista de doutorado, MAG é bolsista de pós-doutorado e LTS é bolsista ITI do Programa MAI-DAI/CNPq, e PN é bolsista PQ-1C/CNPq.

Referências

[1] P. C. Carneiro, T. F. Bastos Filho, A. Rueda, S. Krishnan, L. M. David Luiz, e A. de Oliveira Andrade, "Assessment of bradykinesia on people with Parkinson's disease through mechanomyography," Anais do XII Simpósio de Engenharia Biomédica - IX Simpósio de Instrumentação e Imagens Médicas, ISSN: 2358-3568, 2019, doi: 10.5281/zenodo.3461218

[2] W. Huo, P. Angeles, Y. F. Tai, N. Pavese, S. Wilson, M. T. Hu, e R. Vaidyanathan, "A Heterogeneous Sensing Suite for Multisymptom Quantification of Parkinson's Disease," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 6, pp. 1397-1406, Jun. 2020.

[3] L. V. Schaefer, N. Löffler, J. Klein, e F. N. Bittmann, "Mechanomyography and acceleration show interlimb asymmetries in Parkinson patients without tremor compared to controls during a unilateral motor task," Scientific Reports, 2021.

[4] B. Woods, M. Subramanian, A. Shafti, e A. A. Faisal, "Mechanomyography based closed-loop Functional Electrical Stimulation cycling system," IEEE, 2016.

[5] C. Meagher, E. Franco, R. Turk, S. Wilson, N. Steadman, L. McNicholas, R. Vaidyanathan, J. Burridge, e M. Stokes, "New advances in mechanomyography sensor technology and signal processing: Validity and intrarater reliability of recordings from muscle," Journal of Rehabilitation and Assistive Technologies Engineering, vol. 7, pp. 1-8, 2020.

[6] U. Rosenblum, N. Steadman, T. Martineau, S. Wilson, Y. Tai, F. Morgante, F. Baig, S. Haar, e R. Vaidyanathan, "A Complete Parkinson's Telemedicine System Fusing Myography and Movement Data," TechRxiv, 2026, doi: 10.36227/techrxiv.177004253.35233713/v1

[7] J. Quan, H. Uchitomi, R. Shigeyama, C. Gao, T. Ogata, A. Inaba, S. Orimo, H. Ito, K. Machida, M. Sone, e Y. Miyake, "Application of high-sensitivity acceleration sensor detecting micro-mechanomyogram and deep learning for Parkinson's disease classification," Scientific Reports, 2024.

[8] L. V. Schaefer e F. N. Bittmann, "Parkinson patients without tremor show changed patterns of mechanical muscle oscillations during a specific bilateral motor task compared to controls," Scientific Reports, vol. 10, p. 1168, 2020

ATR-FTIR Salivary Spectroscopy Under Gender-Affirming Hormone Therapy: A Comparative Two-Subject Case Study.

Marcus V. T. Bresser Ribeiro
Research and Development Institute
Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, Brazil
marcusbresser@univap.br

Emilia Angela Lo Schiavo Arisawa
Research and Development Institute
Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, Brazil
mirela@univap.br

Jader Manoel do Nascimento
Research and Development Institute
Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, Brazil
jader@univap.br

Geraldo Magno Alves de Abreu
Research and Development Institute
Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, Brazil
deabreumagno@gmail.com

José Wilson Silva Mendes
Dept. of Biosciences and Oral Diagnosis
Instituto de Ciência e Tecnologia (ICT UNESP)
São José dos Campos, Brazil
jw.mendes@unesp.br

Monica Ghislane Oliveira Alves
Dept. of Biosciences and Oral Diagnosis
Instituto de Ciência e Tecnologia (ICT UNESP)
São José dos Campos, Brazil
monica.alves@unesp.br

***Abstract*— This exploratory pilot study evaluates the technical feasibility and sensitivity of Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy as a non-invasive tool to detect biochemical profile variations in saliva from individuals undergoing Gender-Affirming Hormone Therapy (GAHT). Unstimulated saliva samples were collected from one transgender woman undergoing hormone therapy and one cisgender male control. The specimens were centrifuged, dehydrated using a vacuum concentrator, and analyzed using an ATR-FTIR spectrometer equipped with a diamond crystal accessory in the spectral range of 4000 to 500 cm^{-1}. Spectral processing included baseline correction and min-max normalization. A total of 288 spectra were acquired across acquisition points and preparation triplicates as technical replicates to evaluate measurement repeatability rather than independent biological samples. The spectral profiles demonstrated a shared global biochemical topography between both participants. However, variations were identified in specific vibrational regions. The transgender participant showed increased absorbance at 1630 and 1546 cm^{-1}, corresponding to the protein fingerprint region (Amide I and II) with potential superimposed contributions from exogenous steroid hormones. Concurrently, absorbance decreased at 3400-3200 cm^{-1}, 1065 cm^{-1}, and 978 cm^{-1}. Although these decreases align mechanistically with the literature reporting antiandrogen-induced alterations in salivary hydration and glycoprotein content, clinical sialometry or mucin quantification was not performed. Given the N =2 comparative design, observed differences are highly confounded by individual biological variability, diet, hydration, and unrecorded GATH parameters (dosage, duration, and collection time). These findings demonstrate the technical feasibility of acquiring salivary spectra via ATR-FTIR, but larger longitudinal cohorts with clinical controls are required before establishing this approach as a diagnostic screening tool.**

Keywords*— *Spectroscopy; Saliva; Gender-affirming hormone therapy.

I. Introduction

Spectroscopy is a structural analysis tool that evaluates the properties of matter based on its interaction with electromagnetic radiation. Phenomena such as light absorption, reflection, or transmission reveal the sample's composition through the vibrational transitions of its functional groups, yielding a unique "spectral signature"[1], [2]. Fourier Transform Infrared (FTIR) spectroscopy is widely employed to elucidate organic compounds by acquiring absorbance or transmittance spectra across the infrared region.[3].

Historically, blood plasma has been the primary fluid used in spectroscopic diagnostic studies. However, saliva has emerged as a viable alternative because it shares many constituents with plasma through microvascular exchange[4].

The clinical relevance of saliva stems from its non-invasive collection, low operational cost, and ease of sampling compared to venipuncture[2]. Saliva consists of approximately 99% water and 1% organic and inorganic solutes, including peptides, glycoproteins, enzymes, antibodies, growth factors, and steroid hormones[4]. In vibrational spectroscopy, steroid sex hormones exhibit diagnostic bands primarily within the 1490-1300 cm^{-1} and 1100-1000 cm^{-1} regions [4].

Therapeutic drug monitoring and biomarker screening are particularly relevant in transgender women undergoing Gender-Affirming Hormone Therapy (GAHT), especially where unsupervised hormone use occurs. Systemic risks associated with prolonged therapy include insulin resistance, thromboembolism, and arterial hypertension [5]. In the oral cavity, hormone regimens, particularly those involving androgen receptor blockers, have been associated with reduced salivary flow, xerostomia, and secondary susceptibility to dental caries and periodontal diseases [6], [7].

De Moura, Cardoso, and De Carvalho Brito[8] observed that 17-β-estradiol, an estrogen-bioidentical hormone, is the most frequently used medication in transgender hormone therapy, as reported by 88.9% of study participants. This substance is the active pharmaceutical ingredient in widely distributed commercial transdermal formulations, such as Oestrogel, which contains estradiol hemihydrate [9]. Another widely used medication is Perluran®, whose active ingredients are estradiol enanthate and algestone acetophenide. According to Toffoli Ribeiro et al.[10], it shows positive outcomes in breast development, a benefit reported by 76% of their study participants.

GAHT regimens typically use bioidentical estrogen formulations, such as estradiol enanthate and algestone acetophenide (DHPA) [8], [10]. These molecules possess distinct vibrational fingerprints; 17-β-estradiol exhibits O-H stretching (3432-3256 cm^{-1}) and ketal C-O bonds (1275-1020 cm^{-1}); and estradiol enanthate presents ester carbonyl bands (~ 1735 cm^{-1}), C=C aromatic stretches (1615-1585 cm^{-1}), and C-O signals (~1270-1230 cm^{-1})[11]-[14].

Exogenous administration of these compounds may alter specific absorbance intensities within the salivary protein-lipid matrix. However, evaluating these molecular signals requires assessing the technical performance of infrared accessories in dried salivary films.

This exploratory pilot study aims to evaluate the technical feasibility of using ATR-FTIR spectroscopy to detect salivary spectral profile variations in a two-subject comparative design.

II. METHODOLOGY

This cross-sectional, exploratory pilot study was approved by the Institutional Review Board of Science and Technology, São Paulo State University (ICT-UNESP, Campus São José dos Campos, Protocol No. 6,821,454). Two subjects participated in the study: one transgender woman (Participant T) undergoing GAHT and one cisgender male control (Participant C). Participant T was recruited from the "Transformando Sorrisos" outreach project (UNESP/UNIVAP cooperation agreement). Participant C was recruited from the UNESP institutional body.

The inclusion criteria for Participant T were age over 18 years and current use of GAHT. For Participant C, inclusion criteria were age over 18 years and the absence of smoking, alcohol abuse, or illicit drug use.

Unstimulated saliva (1.5 mL) was collected in the morning between 8:00 a.m. and 10:00 a.m. over a 5-minute interval, following standardized pre-collection guidelines: refraining from food and beverages (except water), caffeine, and toothbrushing for at least 2 hours before sampling. Samples were collected into 1.5 mL Eppendorf tubes, frozen at -80 ºC, and subsequently centrifuged (CF 16RN, Hitachi) at 10,000 rpm for 10 min at 20 °C. The supernatant was collected and stored at -80 ºC. Before spectroscopic analysis, samples were thawed at room temperature (20 ºC) and homogenized via vortex for 10 s.

To eliminate the dominant water absorption band, a 10-µL aliquot of each supernatant was dried for 10 min using a vacuum concentrator (Eppendorf 5301). Infrared spectra were acquired using an Invenio spectrophotometer (Bruker) equipped with a Platinum ATR accessory featuring a single-reflection diamond crystal.

Spectra were recorded over the spectral range of 4000- 500 cm^{-1}, with 32 co-added scans per spectrum at a spectral resolution of 4 cm^{-1}. Measurements were recorded across eight random points on the dried film formed on the ATR crystal to ensure absorbance values remained between 0.5 and 0.7 arbitrary units (a.u.). Background spectra were acquired across both participants. These 288 spectra represent technical/analytical replicates used to evaluate measurement repeatability, rather than independent biological observations.

Spectroscopic data were processed using baseline correction and min-max normalization (0 to 1), followed by averaging technical replicate spectra for each participant using the OPUS software suite (version 4.2, Bruker).

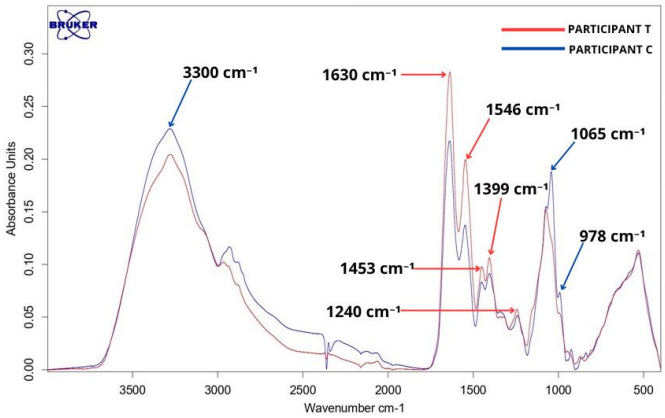
III. RESULTS

Average normalized ATR-FTIR spectra obtained from saliva samples of Participant T and Participant C demonstrated a shared overall biochemical profile across the analyzed range (4000-500 cm^{-1}).

Despite general topographic similarity, localized differences in relative band intensities were observed. The primary spectral variations were identified at 3400- 3200 cm^{-1}, 1630 cm^{-1}, 1546 cm^{-1}, 1453 cm^{-1}, 1399 cm^{-1}, 1240 cm^{-1}, 1065 cm^{-1}, and 978 cm^{-1}. Prominent intensity increases were noted at 1630 cm^{-1} and 1546 cm-1 in Participant T relative to Participant C (Figure 1).

In the higher-wavenumber region, both samples showed a broadband between 3400 and 3200 cm^{-1}, with lower relative absorbance in Participant T. Within the fingerprint region (1800- 600 cm ¹), peak definition and relative intensity varied between 1453 cm ¹ and 978 cm ¹.

Fig. 1. FTIR average spectra overlay of saliva samples from Participant T and Participant C.



IV. DISCUSSION

The general spectral similarity between Participant T and Participant C reflects the common biochemical foundation of human saliva, which consists predominantly of proteins, lipids, carbohydrates, and inorganic salts. The observed absorbance variations indicate changes in relative biochemical proportions. However, attributing these differences solely to GAHT confounds them with individual biological variability, biological sex, diet, hydration status, and oral health parameters.

The broadband centered near 3300 cm^{-1} corresponds to O-H stretching vibrations (residual water and carbohydrate/lipid hydroxyls), superimposed on N-H stretching vibrations of salivary proteins. The lower absorbance observed in Participant T in this region, alongside changes at 1065 cm^{-1} and 978 cm^{-1} (C-O and C-O-P stretching of oligosaccharides, phosphate groups, and structural glycoproteins such as mucins), aligns mechanistically with literature describing antiandrogen-induced reduction in salivary secretion[7]. However, because objective clinical sialometry (flow rate in mL/min), mucin quantification assays, or clinical oral dryness scores were not collected

during sampling, these spectral decreases represent a mechanistic hypothesis rather than a confirmed clinical diagnosis of xerostomia or hyposalivation.

In the fingerprint region, the increased absorbance at 1630 cm^{-1} (Amide I, C=O stretching of peptide backbones) and the sharper profile at 1546 cm^{-1} (Amide II, N-H bending and C-N stretching) in Participant T occur in regions where exogenous steroid molecules (such as 17-β-estradiol, algestone acetophenide, and estradiol enanthate) also exhibit conjugated C=O and aromatic C=C vibrational modes[11]-[13]. Because these bands overlap directly with the endogenous protein matrix, and no independent analytical confirmation (e.g., mass spectrometry or pure reference compound spectra acquired under identical conditions) was performed, these spectral changes cannot be definitively assigned to the drug molecules themselves. Instead, they reflect composite changes in the matrix fingerprint region.

Variations at 1453 cm^{-1} and 1399 cm^{-1} (CH_2/CH_3 bending and steroid ring modes) and 1240 cm^{-1} (C-O stretching) fall within spectral regions reported as diagnostic for steroid sex hormones in biologic fluids[4]. Nevertheless, establishing these coordinates as definitive biomarkers requires rigorous validation.

As a preliminary two-subject comparative study, some operational limitations must be acknowledged. The N = 2 comparative design introduces complete confounding between GAHT status, individual biological variation, baseline salivary composition, lifestyle, and dietary habits. The 288 acquired spectra represent technical analytical measurements on two biological subjects, precluding population-level inferential statistics. Specific GAHT dosage, exact treatment duration, and objective sialometric parameters were not recorded. Pure hormone standards were not analyzed under the same experimental protocol to isolate pure drug spectral contributions from protein matrix shifts.

Future research should employ longitudinal designs tracking transgender individuals before initiating GAHT and at controlled follow-up intervals. Such studies should include age-matched clinical controls, standardized schedules, sialometric/biochemical testing, and direct spectral analysis of reference pharmacological compounds.

V. CONCLUSION

This exploratory pilot study demonstrates the technical feasibility of ATR-FTIR spectroscopy for acquiring reproducible infrared spectra from dried salivary films in individuals undergoing Gender-Affirming Hormone Therapy. However, given the limited sample size (N=2), observed spectral differences at 1630, 1546, 1065, and 978 cm^{-1} cannot be definitively attributed to exogenous hormones or specific oral clinical conditions due to confounding with inter-individual variability, lifestyle factors, and unrecorded clinical parameters (circadian timing, dosage, and GAHT duration). From a dental perspective, this methodology is a viable tool for early, noninvasive, and low-cost diagnostic screening. Identifying molecular signatures of hyposalivation and loss of oral homeostasis at an early stage paves the way for timely preventive interventions. This proactive approach could mitigate severe clinical manifestations such as xerostomia, reduced buffering capacity, and the consequent susceptibility to dental caries and periodontal diseases in this population. However, future research using larger and more robust sample designs remains essential to validate these vibrational coordinates as definitive biomarkers and to consolidate their clinical applicability and longitudinal monitoring.

REFERENCES

[1] L. A. Forato, R. BERNARDES FILHO, D. Osiró, T. D. C. Bicudo, e L. A. Colnago, “A Espectroscopia na região do Infravermelho e algumas aplicações.”, 2010.

[2] M. A. da Costa Nunes, “Espectroscopia ATR-FTIR da saliva como ferramenta de triagem diagnóstica e avaliação do estresse oxidativo no transtorno do espectro autista”, Universidade Federal de Uberlândia, 2022.

[3] A. K. R. Souza, D. M. V. de Lima e S. L. A. Oliveira, “Técnica FTIR e suas aplicações em amostras orgânicas”, *ANAIS DO ENIC*, n. 6, 2023.

[4] M. M. V. Pinto, E. A. L. S. Arisawa, R. A. B. L. Martins, e L. J. Raniero, “Biomarcadores plasmáticos e salivares para diagnóstico precoce de Transtorno do Espectro Autista: revisão sistemática”, *Research, Society and Development*, v. 10, n. 10, p. e412101018924–e412101018924, 2021.

[5] T. J. Ialá, “PORTO ALEGRE 2022 UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL ESCOLA DE ENFERMAGEM PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE COLETIVA”, 2022.

[6] H. A. G. de Oliveira, “Oral condition and quality of life of transgender individuals: A cross-sectional study,” Master’s thesis, School of Dentistry, Federal University of Pernambuco, Recife, Brazil, 2023.

[7] C. M. D. V. Linhares *et al.*, “CONDIÇÕES BUCAIS DE TRANSEXUAIS EM PROCESSO DE HORMONIZAÇÃO”, *Braz. J. Implantol. Health Sci.*, v. 6, n. 3, p. 2760–2777, mar. 2024, doi: 10.36557/2674-8169.2024v6n3p2760-2777.

[8] L. F. M. de Moura, D. S. Cardoso, e G. de Carvalho Brito, “PE-52 Recomendações Clínicas para Harmonização de Pessoas Trans: Revisão de Escopo”, *JORNAL DE ASSISTÊNCIA FARMACÊUTICA E FARMACOECONOMIA*, v. 11, n. s. 2, 2026.

[9] Besins Healthcare Brasil Comercial e Distribuidora de Medicamentos Ltda., “Oestrogel (estradiol) package insert: bula para profissionais de saúde”. Besins Healthcare, Jundiaí, 27 de agosto de 2025. [Online]. Disponível em: https://consultas.anvisa.gov.br/#/bulario/q/?nomeProduto=OESTROGEL

[10] C. Toffoli Ribeiro *et al.*, “Assessment of parenteral estradiol and dihydroxyprogesterone use among other feminizing regimens for transgender women: Insights on satisfaction with breast development from community-based healthcare services”, *Annals of Medicine*, v. 56, n. 1, p. 2406458, 2024.

[11] B. F. Minaev e V. A. Minaeva, “Study of IR spectrum of the 17β-estradiol using quantum-chemical density functional theory”, *Biopolymers & Cell*, v. 22, n. 5, p. 363, 2006.

[12] E. Pawelczyk, I. Pluta, and B. Marciniec, “Infrared spectrophotometry in quantitative drug analysis. II. Assay of some steroids”, Acta Poloniae Pharmaceutica, v. 30, n. 3, p. 289–295, 1973.

[13] J. Pandey, P. Prajapati, P. Tandon, K. Sinha, A. P. Ayala, e J. Ellena, “Vibrational and conformational analysis of structural phase transition in Estradiol 17β valerate with temperature”, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, v. 263, p. 120219, 2021.

[14] PubChem, “Estradiol Enanthate”. Acesso em: 29 de junho de 2026. [Online]. Disponível em: https://pubchem.ncbi.nlm.nih.gov/compound/21070

Avaliação da Inteligência de adultos: protótipo de jogo computadorizado baseado na escala WAIS-III

Daieny Panhan Theodório
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, São Paulo
daienytheodorio@umc.br

Sofia Panhan Theodório
Curso de Odontologia
Universidade de Mogi das Cruzes
Mogi das Cruzes, São Paulo
sofiatheodorio@gmail.com

Terigi Augusto Scardovelli
Núcleo de Pesquisas Tecnológicas
Universidade de Mogi das Cruzes
Mogi das Cruzes, São Paulo
terigiscardovelli@umc.br

Abstract: **Human intelligence has been studied for over 120 years, and psychological tests for its assessment have existed since the early 20th century. The objective was to develop a computerized intelligence scale designed for adults, similar to the WAIS-III, and to correlate the results obtained with those from judges' evaluations. The first prototype was created in PowerPoint, in the form of a game with scenarios and links. The proposal was submitted for review by professionals familiar with the use of the WAIS-III in psychological assessment contexts. The raters recommended replacing some images and suggested creating the scale on an online platform, allowing for the integration of features such as voice recording, a timer, and the use of touch pens or styluses. The next steps involve migrating the scale to a digital environment, implementing the timer, and administering the instrument to participants.**

Keywords: Psychological assessment; intelligence; digital game; WAIS-III; psychometrics

I. INTRODUÇÃO

A Escala Panhan de Inteligência para Adultos (EPIA) é uma proposta de teste gamificado de inteligência desenvolvido para aplicação pelo computador, de forma prática e divertida, seu desenvolvimento tem como base a Escala de Inteligência Wechsler para Adultos (WAIS – III).

A escala WAIS - III é considerada como padrão ouro entre os psicólogos que o utilizam, é um instrumento clínico de aplicação individual composto por 14 subtestes. É utilizado para diversas avaliações, como por exemplo para identificação de problemas de aprendizagem, diagnóstico de pessoas com déficit intelectual, altas habilidades e superdotação, diagnóstico diferencial de transtornos neurológicos e psiquiátricos que afetam o funcionamento mental. Pode ser utilizado para uma investigação mais ampla juntamente com outros testes psicológicos e de personalidade. Para a sua utilização é necessário que o usuário tenha treinamento e experiência tanto para a aplicação, como para a interpretação, análise e a produção do laudo técnico [1].

A primeira autora observou que os pacientes apresentavam comportamentos ansiosos antes e durante a aplicação dos testes psicológicos, principalmente por saberem que estavam sendo avaliados. Quando a avaliação ocorre de forma gamificada, porém, o caráter lúdico da tarefa se torna evidente para o paciente, o que contribui para a redução da ansiedade. Foi justamente por isso que a Escala EPIA foi desenvolvida para minimizar a ansiedade dos pacientes durante o processo de avaliação.

O termo gamificação não significa usar jogos propriamente dito, trata-se de incorporar elementos típicos dos jogos como avatares, rankings, desafios ou sistemas de reputação em atividades que não são jogos [2]. Assim, gamificar é estruturar uma ação com a lógica dos jogos, sem exigir tecnologias específicas ou que os participantes joguem de fato. Diferente do *game-based learning*, a gamificação aplica dinâmicas e mecanismos dos jogos em diversos contextos para aumentar motivação e engajamento [3] [4].

O protótipo da primeira versão da Escala EPIA está concluído com todas as suas etapas construídas, faltando ajustes. O objetivo final da construção da Escala EPIA é de oferecer uma alternativa acessível aos psicólogos, facilitando tanto a avaliação quanto o atendimento psicoterapêutico de seus pacientes. Com a tabulação dos resultados em formato digital, o processo se torna mais ágil e prático, diferenciando-se da análise manual utilizada na escala WAIS-III e em outros testes psicológicos com crivos.

II. REFERENCIAL TEÓRICO

Segundo Rodrigues (2022) [5] a inteligência é considerada de origem genética com plasticidade neuronal, ou seja, com possibilidade de desenvolvimento e conexões neurais para a organização e reorganização dos neurônios. Para o autor uma pessoa reconhecida como intelectual pode ter amplo conhecimento em diversas áreas, mas não necessariamente tem pontuações elevadas em um teste de inteligência, já para Papalia e Martorell (2022) [6] a inteligência não é só genética, mas pode ser estimulada e melhorada a partir do ambiente em que a pessoa vive. Complementam dizendo que o lobo frontal é a última região do cérebro a ser desenvolvida e a que mais se desenvolveu e está relacionada diretamente com a evolução humana.

A inteligência está relacionada também com "a capacidade de armazenamento de memória, de atenção, de funções executivas, de aprendizagem, compreensão, criatividade, tomada de decisão, e prevenção". O ser humano também pode "manipular" a sua percepção, o filtro, a sua consciência, as emoções com variedade de intensidade e significância [5].

Em 2025 completou 120 anos desde que Alfred Binet e Théodore Simon desenvolveram o primeiro teste destinado à avaliação da inteligência. Criado originalmente para acompanhar o progresso de crianças na educação especial, o instrumento conhecido como escala Binet-Simon era voltado exclusivamente ao público infantil.

Na década de 1940, David Wechsler ampliou esse campo ao elaborar um teste de inteligência também voltado para crianças. Com o tempo, suas escalas evoluíram, e hoje as escalas Wechsler são aplicados a indivíduos de todas as faixas etárias [7]. Cabe destacar que as escalas Wechsler são os testes mais utilizados atualmente no ambiente clínico [7] há a versão para crianças (WISC) e a versão para adultos (WAIS).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

A avaliação psicológica vai além da análise da inteligência, abrangendo diversas funções cognitivas, aspectos da personalidade, preferências profissionais e outros fatores relevantes. Trata-se de um processo que permite confirmar ou aprofundar hipóteses elaboradas pelo profissional durante os atendimentos, contribuindo significativamente para o planejamento de intervenções e estratégias de reabilitação [8], [9]. Para isso, são utilizadas múltiplas estratégias, como observação sistemática, entrevistas e aplicação de testes psicológicos [8], [10].

A psicometria juntamente com a psicologia cognitiva e outras abordagens correlacionam as variáveis investigadas e propõe um método experimental. Dessa forma, alguns testes psicológicos pouco a pouco começam a ser apresentados de forma eletrônica desenvolvidos a partir de *software* [11], no entanto a maioria dos testes adquiridos ainda são vendidos com correção manual.

As Tecnologias da Informação e Comunicação (TICs) têm contribuído significativamente para a Psicologia por meio da informatização dos instrumentos de avaliação. A prática da avaliação psicológica envolve diferentes métodos e é exclusiva do psicólogo. Com a digitalização, os dados podem ser processados de forma rápida, precisa e em grande escala, representando um avanço contemporâneo na área e auxiliando o profissional [12].

O uso das Tecnologias da Informação e Comunicação (TICs) na Avaliação Psicológica torna o processo mais ágil, preciso e confiável, permitindo o uso de diversos recursos multimídia, possibilitando obter informações que a aplicação de testes tradicionais não é capaz de proporcionar. Elas ampliam as possibilidades de aplicação e coleta de dados, automatizam tarefas manuais e favorecem avanços psicométricos, elevando a qualidade das avaliações realizadas [13], [14], [15].

Na Universidade Federal de Minas Gerais (UFMG), há um laboratório dedicado à psicometria e ao desenvolvimento de algoritmos e análises estatísticas voltadas para mensuração e predição no campo dos testes educacionais e psicológicos. Dentre os instrumentos desenvolvidos, destaca-se o Teste de Compreensão Verbal 1, criado por Gomes e Nascimento (2021) [16], que apresenta similaridade com um dos subtestes da escala WAIS-III.

Com relação à aplicação de testes de forma digital, o CFP - Conselho Federal de Psicologia divulgou uma Nota Técnica nº 07/2019 que esclarece que os testes psicológicos podem ser aplicados tanto no formato lápis e papel quanto de forma informatizada, desde que aprovados pelo SATEPSI (Sistema de Avaliação de Testes Psicológicos). No entanto, a aplicação informatizada (mediada por computador) não é o mesmo que aplicação on-line (remota). Cabe ao psicólogo conhecer o manual do teste e aplicar conforme as diretrizes específicas. A avaliação psicológica on-line é permitida, desde que respeite as normas vigentes [15].

Para Hazin et al., (2018) [17] a integração entre avaliação psicológica e neuropsicológica ainda enfrenta desafios no Brasil, especialmente no campo da neuropsicologia. Um dos principais obstáculos é conectar áreas como neurociência, psicometria moderna e tecnologia. Segundo Bilder e Reise (2019) [18], avanços significativos na prática neuropsicológica podem surgir com o uso de estratégias psicométricas e tecnológicas, como a Teoria de Resposta ao Item (TRI), algoritmos de testes adaptativos e análises de funcionamento diferencial dos itens (DIF), que contribuem para avaliações mais precisas e eficientes.

Bilder em 2011 [19] alertava para o futuro da neuropsicologia, dizendo que poderá ser transformada e impulsionada pelos avanços em neuroimagem, genética e tecnologia da informação. Os profissionais poderão utilizar métodos avaliativos digitais e em larga escala, embora a prática ainda dependa de abordagens tradicionais e menos eficientes. Segundo Miller e Barr (2017) [20] o futuro da avaliação neuropsicológica incluirá métodos híbridos - laboratoriais, remotos e automatizados, posicionando a neuropsicologia como protagonista na ciência cognitiva e comportamental.

Considerando que os atuais desafios na área tendem a se intensificar em curto prazo, torna-se essencial repensar os modelos tradicionais de avaliação psicológica e neuropsicológica. Nesse contexto, adotar abordagens inovadoras pode representar um avanço significativo na prática profissional. Uma dessas iniciativas é o desenvolvimento de uma escala de inteligência gamificada para adultos, semelhante à WAIS-III, promovendo maior acessibilidade, agilidade e precisão no processo avaliativo.

Com base no exposto acima, os autores objetivam desenvolver a Escala EPIA com a utilização de um programa de computador conhecido como *Power Point*; solicitar a avaliação da escala EPIA por 5 juízes da área de avaliação psicológica; comparar e correlacionar os resultados obtidos pelos juízes.

III. Método

A presente pesquisa classifica-se como de natureza aplicada [21], com base nos objetivos pode ser classificada como exploratória [22] e de análise mista.

Participantes: 5 juízes-psicólogos especializados e atuantes na área de avaliação psicológica acostumados a utilizar a escala WAIS-III em seus atendimentos psicoterapêuticos para jogar e avaliar a escala EPIA.

Materiais: a Escala Panhan de Inteligência para Adultos - EPIA será um instrumento restrito para uso de psicólogos, podendo ser aplicado em pessoas com mais de 18 anos. A intenção futura é que seja confiável e que forneça os escores de QI Verbal, QI de Execução e QI Total. A EPIA está construída com base na escala WAIS-III, no entanto as imagens não são as mesmas, devido aos direitos autorais, outro diferencial é o formato digital gamificado com mapa e tarefas para os participantes, devido a complexidade de tentar replicar algo semelhante aos subtestes cubos. procurar símbolos e armar objetos, na EPIA esses subtestes foram excluídos.

Por enquanto, o protótipo do jogo está construído no programa *Power Point*, futuramente será inserido em uma plataforma (site da internet) com maiores recursos para a construção do mapa e dos personagens.

Para a aplicação dos subtestes aparecerá um personagem solicitando a ajuda do participante para solucionar alguns problemas, o participante percorrerá no mapa para ir até o local determinado e concluir a tarefa. Todas as tarefas têm níveis de dificuldade partindo do mais fácil para o mais complexo. Para as tarefas com tempo para execução o participante será alertado e um relógio no canto

esquerdo da tela é mostrado. Quando as atividades solicitarem respostas verbais, o participante será avisado que as suas respostas serão gravadas para que o psicólogo possa avaliá-las futuramente e tabular as respostas, cada resposta será contabilizada com pontos para que o QI Verbal, de Execução ou Total seja calculado.

Os subtestes desenvolvidos na escala EPIA são baseados na escala WAIS e levam os mesmos nomes da escala WAIS-III:

Completar Figuras: neste subteste o objetivo é verificar se o avaliado consegue observar detalhes, capacidade de análise e síntese visual, memória semântica visual, atenção e reconhecimento de detalhe essencial, processamento visual e a velocidade de processamento implícita, pois será utilizado marcador de tempo.

Um personagem dentro de uma casa solicitará para que o participante entre e o ajude a conferir produtos comprados pela internet, o participante o ajudará a identificar o que está faltando, como por exemplo: uma cadeira sem um pé. Essa atividade tem 4 figuras para treino e 20 para que o participante identifique onde está o problema. Para cada figura o participante terá 20 segundos para a conclusão.

Compreensão: o principal objetivo deste subteste é o julgamento social, examina a capacidade de expressar as experiências sociais do avaliado. Exige conhecimento de regras de relacionamento social e permite observar a facilidade de argumentação (quando é pedido que o avaliado justifique as suas respostas). Envolve padrões convencionais de comportamento, maturidade e senso comum. É uma boa medida de avaliação pré-mórbida e sensível a danos neurológicos.

O examinando deverá responder oralmente a uma série de questões que exigem soluções para problemas cotidianos ou o entendimento de regras e conceitos sociais. São apresentadas ao avaliando 18 questões sem tempo para respondê-las. As respostas serão gravadas.

Semelhanças: o objetivo é avaliar a abstração. Envolve relações lógicas e formação de conceitos verbais, raciocínio verbal abstrato e habilidades intelectuais gerais, além de síntese e integração de conhecimentos.

É solicitado que o participante aponte semelhanças entre duas palavras, que representam objetos ou conceitos, como por exemplo: maçã e laranja; casaco e blusa. O participante precisa responder a 19 itens sem tempo para a conclusão da tarefa. As respostas serão gravadas.

Aritmética: o objetivo é verificar se o avaliando consegue resolver problemas matemáticos e como utiliza a memória operacional. Envolve a execução de cálculo mental, a compreensão de enunciados e a capacidade de raciocínio matemático. Requer a utilização da memória operacional, necessária para manipular mentalmente os elementos do problema.

Estão dispostos 19 problemas para a pessoa resolver mentalmente, iniciando com soma simples com dois algarismos. É necessário o registro de tempo.

Vocabulário: esse subteste avalia a memória semântica e inteligência cristalizada do avaliado, como também a competência linguística, conhecimentos lexicais e a capacidade de elaboração do discurso. Esse teste é considerado uma medida de inteligência cristalizada e inteligência geral.

O subteste consiste em obter do participante informações relacionada a inteligência cristalizada do participante por meio de conhecimento obtido ao longo da vida. Será solicitado ao avaliado que ele defina oralmente o significado de cada palavra apresentada na lista de palavras. As respostas do avaliando serão gravadas. Estão disponíveis 5 palavras para treino e outras 32 palavras sem tempo para a conclusão.

Informação: nesse subteste avalia-se a memória semântica e conhecimentos gerais, envolve conhecimentos adquiridos a partir da educação escolar e/ou familiar e inteligência cristalizada.

Nesse subteste o participante responderá a 28 questões sobre informações gerais, eventos comuns, objetos, lugares e pessoas. Sem tempo para a conclusão da tarefa.

Códigos: nesse subteste avalia-se a velocidade de processamento e a atenção, envolve a memória visual e requer boa caligrafia.

É solicitado ao participante que utilize uma caneta no modo apresentação do programa para que identifique algumas letras, substituindo-as por símbolos escritos corretamente e de forma individual. O avaliando terá 120 segundos para escrever o maior número de símbolos que conseguir.

Raciocínio Matricial: esse subteste avalia a inteligência fluída e a resolução de problemas. Envolve o processamento da informação visual, capacidade de distinguir os detalhes essenciais e a organização visuoespacial. É considerada uma medida de raciocínio fluído não verbal e habilidade de inteligência geral.

Esse teste é composto de quatro tipos de tarefas de raciocínio não verbal: completar padrões, classificação, analogia e raciocínio serial. O examinando deverá olhar para uma matriz em que está faltando uma parte da figura e deverá apontar ou identificar por meio do número correspondente a parte que completa a matriz entre as 6 ou 8 opções. A tarefa tem dificuldade gradativa e tempo para execução cronometrado.

Dígitos: esse subteste avalia a memória de curto prazo e memória operacional. É solicitado ao avaliando que repita verbalmente uma série de números em ordem direta e outra em ordem inversa. Esse subteste foi substituído por atividades numéricas e matemáticas com a utilização da memória operacional, inteligência cristalizada e fluída. O avaliando terá que responder oralmente sem tempo para o término.

Avaliação de juízes

Nessa avaliação os juízes responderam com sim ou não a 18 perguntas sobre a Escala EPIA, sendo 9 perguntas sobre a aparência e 9 sobre o conteúdo. As perguntas que estão relacionadas à aparência da escala, são por exemplo: a escala oferece instruções pertinentes ao usuário? As imagens são adequadas? Os ícones para voltar e cronometro, estão visíveis? Quanto ao conteúdo, há perguntas como: As tarefas propostas estão pertinentes para uma escala de inteligência? A quantidade de tarefas está suficiente? O conteúdo está compatível com as faixas etárias? Pessoas que tem dificuldade de utilizar um computador, conseguirão jogar? E há espaço para comentários sem número de linhas ou espaço definido.

Vale ressaltar que há atividades cronometradas e outras não, dessa forma, o avaliando será avisado toda vez que a atividade exigir tempo para o término da tarefa.

Procedimento: o projeto foi aprovado pelo Comitê de Ética com Seres Humanos da UMC com o número do parecer:6.855.797.

Foram escolhidas como juízas 5 psicólogas especialistas em avaliação neuropsicológica, que atuam em consultório particular com esse tipo de avaliação e que utilizam de forma constante a escala WAIS-III. Elas foram informadas sobre a Escala EPIA e se disponibilizaram a avaliá-la. Foi entregue o TCLE assinado em duas vias, os termos foram assinados pelas juízas e uma cópia devolvida a elas. Vale ressaltar que os ajustes sugeridos pelas juízas serão acatados.

Desenvolvimento do jogo: o jogo foi desenvolvido no programa *Power Point* inicialmente, com a utilização de ilustrações de aplicativos gratuitos, mesmo assim há autoria das imagens, dessa forma, as imagens foram reconstruídas com o auxílio de inteligência artificial, futuramente o teste será desenvolvido em um site para que o cálculo, tabulação e análise dos resultados de cada subteste fique disponível ao psicólogo aplicador.

IV. Resultados

A escala foi elaborada com o propósito de reduzir a ansiedade do participante durante sua aplicação. Por ser apresentada em formato gamificado e executado no computador, ela adota uma abordagem lúdica e visualmente atrativa. Acredita-se que, dessa forma, o avaliando responderá de maneira mais tranquila, evitando reações comportamentais que possam comprometer os resultados. Além disso, a proposta atende à demanda por informatização e uso de tecnologias, superando os modelos tradicionais com papel e crivos. A escala tem cenários comuns de estabelecimentos como: escola, casa, supermercado, parque e sorveteria que todas as pessoas já utilizaram ou fazem uso, com o intuito de acolher e tranquilizar o avaliando com situações já vivenciadas, por isso também as cores alegres e em estilo de desenho.

Figuras e cenas utilizadas no jogo

Na tela inicial (Figura 1) o avaliando precisará clicar inicialmente em “Como jogar”, assim que clicar será direcionado para a tela 2 (Figura 2) para obter as informações e instruções do jogo, após a leitura retornará para a tela inicial clicando na seta azul à esquerda da tela e teclará “Início”. Na tela 2 o avaliando será instruído para o uso do mouse e das setas para se movimentar no cenário do jogo.

O primeiro subteste Completar Figuras será respondido dentro da casa. Completar Figuras é um subteste que tem como objetivo avaliar a capacidade do participante de perceber detalhes visuais, realizar análise e síntese visual, utilizar a memória semântica visual, manter a atenção e reconhecer elementos essenciais. Também são observados o processamento visual e a velocidade de resposta, já que a atividade é cronometrada. Será solicitado que o participante encontre defeitos nos objetos ou figuras apresentados pelo personagem adulto.

No subteste Completar Figuras o personagem da direita solicita e orienta o avaliando para as tarefas que deverá executar, como por exemplo: “vá até a escola”, ou “vá até o supermercado”.

A seguir, são apresentadas algumas imagens e tarefas incluídas no jogo como atividades propostas aos participantes. Essas imagens fazem parte do subteste "Completar Figuras" (Figura 6), no qual o avaliado deve identificar o que está ausente em cada figura.



Figuras 1 – Tela inicial do teste, Figura, 2 – Instruções para a execução do teste, Figura 3 – Primeiro subteste com a entrada dentro da casa dos personagens e Figura 4 – Personagens

Neste estabelecimento há vários subtestes que precisam ser respondidos pelo avaliando. Assim que o avaliando entra na escola (Figura 7), é recepcionado pelos professores (Figura 8), e orientado para as atividades, como identificar itens que faltam em uma figura, exercícios de matemática e conhecimentos gerais.



Figura 5 – Orientações para o subteste Completar Figuras, Figura 6 – Estabelecimento Supermercado, 7 – Estabelecimento escolar e Figura 8 – Recepção pelos professores.

No subteste Vocabulário, o participante deverá ler palavras e explicar seus significados. Essa tarefa avalia a memória semântica, a competência linguística, o repertório lexical e a capacidade de formular discursos. É considerada uma medida de inteligência cristalizada e inteligência geral.

Já no subteste Aritmética, verifica-se a habilidade do avaliando em resolver problemas matemáticos utilizando a memória operacional. A atividade exige cálculo mental, compreensão dos enunciados e raciocínio lógico, sem o uso de instrumentos auxiliares.

Por fim, o subteste Informação avalia a memória semântica e os conhecimentos gerais adquiridos ao longo da vida escolar e familiar, também relacionados à inteligência cristalizada. Nele, o participante responde a 28 perguntas sobre fatos do cotidiano, objetos, locais, pessoas e eventos, sem limite de tempo para a conclusão.

No ambiente representado pela Figura 9 e 10, o avaliando é conduzido a novos subtestes. Após a apresentação inicial, o personagem orienta o participante quanto às tarefas a serem realizadas, como o subteste Códigos.

O subteste Códigos tem como objetivo avaliar a velocidade de processamento, a atenção, a memória visual e a habilidade motora fina, especialmente a caligrafia. Durante a atividade, o participante deve utilizar uma caneta no modo apresentação do programa para identificar letras e substituí-las por símbolos correspondentes, desenhados de forma correta e individualizada. O tempo disponível para a execução da tarefa é de 120 segundos, período em que o avaliando deve registrar o maior número possível de símbolos.



Figura 9 – Recepção no supermercado, Figura 10 – Estabelecimento sorveteria e personagem da sorveteria

Há outros personagens como a ambulante no parque que também vende sorvetes e solicita que o avaliando realize o subteste Raciocínio Matricial, esse subteste foi desenvolvido para a avaliação da inteligência fluída e da capacidade de resolução de problemas, envolve o processamento de informações visuais, a identificação de detalhes relevantes e a organização visuoespacial. Trata-se de uma medida de raciocínio não verbal e de inteligência geral. O teste é composto por quatro tipos de tarefas: completar padrões, classificação, analogias e raciocínio serial. O participante deve observar uma matriz com uma parte ausente e indicar, por meio do número correspondente, qual das 6 ou 8 opções disponíveis completa corretamente a figura. A atividade apresenta dificuldade progressiva e é realizada com tempo cronometrado.

O subteste Semelhanças é um subteste desenvolvido para a avaliação da capacidade de abstração. Ele envolve o estabelecimento de relações lógicas, a formação de conceitos verbais, o raciocínio verbal abstrato e habilidades intelectuais gerais, incluindo a síntese e integração de conhecimentos. Durante a atividade, o participante deve identificar semelhanças entre pares de palavras que representam objetos ou conceitos, por exemplo, "maçã e banana" ou "meias e sapatos". Ao todo, são apresentados 19 itens, sem limite de tempo para a conclusão. As respostas fornecidas pelo participante são registradas para análise posterior.

Durante o jogo o avaliando será orientado pelos personagens com os objetivos que terá que cumprir, qual personagem procurar para cada etapa do jogo e qual o local para a realização da tarefa. Dessa forma, o psicólogo também poderá avaliar como se comporta o seu paciente durante o jogo, é possível observar se prestou atenção na próxima tarefa e se lembra o que precisará fazer quando chegar ao local estipulado, as dificuldades e facilidades em cada etapa, pois todos os processos cognitivos podem ser avaliados neste jogo e a observação do profissional é importante durante a avaliação.

Avaliação das juízas: foi solicitado a 5 psicólogas especialistas em neuropsicologia que avaliassem o jogo enquanto jogavam, de modo geral as juízas gostaram da ideia de gamificar o teste atualmente utilizado em papel para o formato digital, estão esperançosas para a finalização com a possibilidade de cálculo, tabulação e análise digital, com o intuito de facilitar o trabalho de avaliação dos psicólogos. No quadro 1 há as respostas de todas elas.

QUADRO 1 – RESPOSTAS DAS JUÍZAS SOBRE A ESCALA

Descrição	Sim	Não
Aparência		
A escala oferece instruções pertinentes ao usuário?	5	
As imagens estão suficientes?	5	
As imagens são adequadas?	4	1
Há personagens suficientes?	5	
As setas indicativas para mudança de local no mapa, estão compreensíveis?	5	
Os ícones para voltar e cronômetro, estão visíveis?	5	
As cores são agradáveis?	5	
A quantidade de figuras por tela está adequada?	4	1
Conteúdo		
As tarefas propostas estão pertinentes para uma escala de inteligência?	5	
O nível de execução está de acordo com a proposta?	5	
O nível de complexidade das tarefas foi bem desenvolvido?	5	
A quantidade de tarefas está suficiente?	5	
A execução da ordem das tarefas está adequada?	5	
A quantidade de cenários no mapa para a execução das tarefas está adequada?	5	
O comando para iniciar e parar a gravação da voz está compreensível?	3	2
O conteúdo está compatível com a faixa etária para adultos?	5	
Pessoas que tem dificuldade de utilizar um computador, conseguirão jogar?	3	2
Você sugere a comercialização da escala?	5	

As juízas apontaram que as imagens estão boas, no entanto há algumas setas e hiperlinks que não estavam funcionando apropriadamente e esse fato dificultou a execução dos subtestes no modo apresentação, já ajustado.

Quanto as imagens, consideraram que estão boas. Uma sugeriu a sala com cores mais vibrantes, e outra juíza sugeriu as entradas dos estabelecimentos com a porta um pouco maior para identificar que se pode "entrar" pela porta. E a última sugeriu a entrada da escola "menos poluída", há muitas informações no ambiente.

Quanto ao conteúdo, acharam válida a ideia e a possibilidade de aplicação como teste futuramente, assim que estiverem prontos todos os links e o temporizador. A caneta do *Power Point* no modo apresentação não estava disponível em todas as apresentações, dessa forma, as anotações à caneta não foram possíveis. Sugeriram diminuir a quantidade de exercícios do subteste Raciocínio Matricial que atualmente está com 40 exercícios para 25 a 30 slides no máximo.

Foi aplicado teste Qui quadrado nas respostas das juízas sobre a aparência da EPIA, obtendo que nenhuma das partições apresentou $p < 0.05$, portanto não há associação estatisticamente significativa em nenhuma comparação. O resultado geral confirma que as juízas avaliaram os itens de forma homogênea, sem discrepâncias relevantes entre categorias. Quanto a avaliação do conteúdo da EPIA ($p = 0.0429$) conclui-se que existe associação significativa entre as categorias avaliadas, mas as juízas não avaliaram todos os itens de forma uniforme, há discrepância principalmente nos itens que perguntaram se pessoas sem habilidade com o computador conseguiriam jogar e sobre o ícone gravação de voz que não estava funcionando.

Como a Escala foi desenvolvida em um ambiente gamificado utilizando o *power point* não foi possível substituir alguns subtestes, dessa forma há mais subtestes que avaliam o QI verbal do que o de Execução, dificultando o QI Total, será necessário repensar novos subtestes que complementem a Escala.

A Escala EPIA foi aplicada em duas participantes por enquanto, essa aplicação não estava prevista neste momento e por isso não consta nos objetivos, foi exigência de um parecerista para a aprovação do artigo e a participação no congresso CBEB 2026. Para equilibrar a quantidade de subtestes, foram escolhidos: completar figuras, códigos e raciocínio matricial para o QI Execução e vocabulário, semelhanças e informação para o QI Verbal.

A última pergunta na ficha de avaliação para as juízas: "o teste pode ser comercializável?" Todas responderam afirmativamente.

Próximas etapas da prototipação: a próxima etapa da pesquisa será ajustar a escala de acordo com os apontamentos das juízas, instalando um temporizador, gravador de áudio no programa, elaborar novas atividades como os subtestes que foram excluídos da EPIA em decorrência de utilizar computador, são eles: cubos, procurar símbolos e armar objetos, inserir esse conteúdo em um site para que o avaliando consiga desenhar ou escrever. Assim que os próximos passos estiverem concluídos a aplicação será em outros participantes e os resultados correlacionados com a aplicação da escala WAIS-III. Será necessário avaliar a precisão e o conteúdo novamente após as alterações.

QUADRO 2 – RESPOSTAS DOS PARTICIPANTES COMPARANDO O WAIS-III E A ESCALA EPIA

Subtestes	**Participante 1 – P1**		**Participante 2 – P2**	
	WAIS	**EPIA**	**WAIS**	**EPIA**
Compl. Figuras	21	20	17	18
Vocabulário	36	37	37	41
Códigos	79	3	64	47
Semelhanças	21	26	21	30
Rac. Matricial	16	19	21	19
Informação	13	12	10	14

Os participantes obtiveram resultados semelhantes comparando os resultados brutos da WAIS-III e da Escala EPIA, no entanto, aplicando o teste Shapiro-Wilk obteve-se $p < 0.05$ no P1 e $p > 0.05$ no P2. Utilizou-se correlação de Spearman no P1, a relação é positiva, mas fraca (r 0.11). O valor de p confirma que não há associação real, apenas flutuação aleatória. Essa diferença pode ser devido ao subteste códigos que o P1 apresentou dificuldade com a caneta do *power point*. Para os resultados de P2 foi aplicado correlação de Pearson, (r 0.92), ou seja, os resultados estão associados e de forma linear.

Algumas observações foram anotadas durante a aplicação da Escala EPIA: no subteste vocabulário há algumas palavras complexas antecedendo as mais simples e mais fáceis de definir, que será ajustado. Uma das figuras do subteste completar figuras estava com a imagem escura e os participantes não conseguiram responder e no subteste informação, algumas questões terão que ser revistas, como por exemplo a pergunta: Quem era o presidente dos EUA durante a segunda guerra mundial.

Quanto ao subteste códigos os participantes fizeram utilizando a caneta disponível no *power point*, não foi agradável para os participantes desenharem com a caneta, a pontuação ficou abaixo da pontuação obtida na WAIS devido a dificuldade, futuramente esse subteste será modificado para a aplicação em um tablet com caneta.

A Escala será revista novamente antes da aplicação em novos participantes e não será comercializada antes da aprovação do SATEPSI (Sistema de Avaliação de Testes Psicológicos). Com relação ao conteúdo, precisão e validade da Escala também serão testados e comparados com o nível de dificuldade da WAIS-III.

V. DISCUSSÃO

Atualmente a sociedade está marcada por profundas transformações, impulsionadas pela crescente integração das Tecnologias Digitais de Informação e Comunicação (TDICs) em uma sociedade moldada pela cultura digital. A pandemia acelerou esse processo, evidenciando o valor e a versatilidade dessas tecnologias, que passaram a ocupar um papel central nas práticas educacionais [23].

Os Jogos Sérios aplicados à saúde têm se destacado por sua capacidade de promover mudanças comportamentais e apoiar intervenções clínicas. Entre os oito jogos analisados sete apresentavam finalidade terapêutica, enquanto apenas um tinha foco no desenvolvimento de habilidades específicas. Essa predominância evidencia que o campo tem

direcionado seus esforços para propostas voltadas à modificação de comportamento, ao aprimoramento cognitivo e ao fortalecimento da interação social. Esse cenário converge com a literatura, que reconhece os jogos sérios como ferramentas promissoras para intervenções psicológicas, sobretudo pela capacidade de engajar o usuário, oferecer feedback imediato e permitir a personalização das atividades conforme necessidades individuais. A utilização de jogos sérios na área da saúde corrobora com a proposta da Escala EPIA [24].

Lima (2025) solicitou a avaliação de juízes para o conteúdo de um questionário sobre atitudes de pessoas com deficiência auditiva, os resultados sugerem alta concordância entre os juízes indicando que o instrumento tem potencial para identificar atitudes, crenças e possíveis estigmas relacionados à deficiência auditiva, o que é fundamental para pesquisas e práticas clínicas voltadas à inclusão e ao enfrentamento de barreiras atitudinais. Nesta pesquisa sobre a Escala EPIA os juízes também apontaram concordância quanto ao conteúdo e a aparência, estão faltando outros parâmetros avaliativos como precisão, validade, além de outros detalhes que aparecerão com a aplicação em maior número de participantes antes de aprovação pelo SATEPSI. Os pareceristas solicitaram aplicação em participantes, no entanto, o número só não foi maior devido ao tempo de agendamento de disponibilidade deles [25].

Diante desse cenário, torna-se evidente que a Escala EPIA está preparado para avançar às próximas etapas de validação psicométrica, como análises de constructo e confiabilidade. Essas etapas serão essenciais para consolidar sua aplicabilidade em contextos clínicos e de avalição neuropsicológica.

Selau (2020) construiu um estudo sobre evidências de validade e fidedignidade da escala de funcionamento adaptativo para deficiência intelectual, num primeiro momento a autora destacou a importância dos procedimentos que visaram garantir a validade de conteúdo, cada etapa de construção possibilitou o aprimoramento da escala. Em um segundo momento as análises apresentaram evidências satisfatórias de validade relacionadas a estrutura interna e a variáveis critério, ou seja, o instrumento tem capacidade de mensurar o Funcionamento Adaptativo e indicar prejuízo. Tais evidências atestam o potencial do instrumento em auxiliar no diagnóstico de Deficiência Intelectual [26].

Assim como a Selau desenvolveu seu estudo e solicitou a avaliação de juízes e participantes, neste estudo a Escala EPIA também passará por análises mais rigorosas. As juízas já avaliaram a validade de conteúdo, dizendo que a Escala EPIA representa adequadamente o construto que se pretende avaliar, no entanto, nem todos os subtestes foram desenvolvidos, dessa forma, uma nova avaliação por juízes será realizada quando as etapas estiverem prontas. Segundo a American Psychological Association, a validade de conteúdo constitui uma evidência essencial para demonstrar que o instrumento mede aquilo que se propõe a medir, sendo indispensável nas fases iniciais de desenvolvimento de testes psicológicos [27].

A Escala EPIA deverá ser avaliada quanto a precisão, esta etapa é essencial para garantir que os resultados obtidos não sejam fruto de acaso ou instabilidade, mas sim reflexo real do construto avaliado. Conforme a AERA, APA e NCME (2014), instrumentos psicológicos devem apresentar evidências robustas de precisão para serem considerados adequados ao uso clínico e científico [28].

Ambas são etapas indispensáveis para que um teste psicológico seja considerado cientificamente confiável e eticamente aplicável.

Joly et al (2004) [29] já apontavam para o avanço da informatização na avaliação psicológica e a consolidação no cenário nacional e internacional. Acreditam que essa evolução tecnológica traz benefícios práticos e econômicos à atuação profissional, ao eliminar a necessidade de materiais físicos como lápis, papel e manuais impressos, além de otimizar o tempo e reduzir significativamente os erros de mensuração. Ao incorporar os recursos da psicometria moderna, garante-se maior confiabilidade dos instrumentos sem comprometer a validade ou a precisão dos resultados.

Com a perspectiva de que esse novo formato permitirá a aplicação de testes mais dinâmicos e complexos, envolvendo estímulos visuais, sonoros, tempo de reação e respeito à diversidade cultural, Joly e colaboradores estruturaram seu texto com o objetivo de estimular novas pesquisas e promover a internacionalização da avaliação psicológica informatizada, explorando todo o potencial que a tecnologia oferece para compreender os fenômenos e processos do comportamento humano [13].

Diante do que foi apresentado, é possível reconhecer os benefícios do uso das tecnologias nas áreas da educação, saúde e, em especial, da psicologia. Embora a psicologia esteja inserida no campo da saúde, ela recebe destaque neste contexto por ser o foco específico da pesquisa e a área de formação da autora. Enquanto alguns estudiosos ressaltam as vantagens dessas ferramentas, outros alertam para os cuidados necessários, a fim de evitar possíveis impactos negativos aos usuários. Existem, inclusive, pesquisas que propõem estratégias para minimizar esses riscos. No que diz respeito à aplicação e avaliação de indivíduos por meio de tecnologias, trata-se de uma tendência irreversível e voltada para o futuro.

VI. Considerações Finais

A Escala EPIA foi elaborada com base nos objetivos do projeto, utilizando o programa *PowerPoint* como plataforma de desenvolvimento. Baseada na Escala Wechsler de Inteligência para Adultos (WAIS-III), a EPIA incorporou elementos compatíveis com os recursos disponíveis no software. No entanto, três subtestes da WAIS-III não puderam ser reproduzidos ou adaptados devido às limitações técnicas do *PowerPoint*. Pretende-se futuramente desenvolver os subtestes retirados contanto que avaliem os mesmos construtos em um tablet de forma precisa.

Durante o processo de validação inicial, cinco juízas especializadas foram convidadas a avaliar o protótipo. Elas ofereceram críticas construtivas e sugestões com vistas ao aprimoramento da escala para futura comercialização. Entre os pontos destacados, mencionaram problemas com links, dificuldades no uso da caneta durante o modo apresentação, excesso de estímulos na recepção da escola, ausência de cores na sala de estar e sobrecarga de tarefas no subteste de raciocínio matricial.

As recomendações das juízas estão sendo realizadas, no entanto há as próximas etapas da prototipação que incluem a migração da escala para uma plataforma online, a

implementação de um temporizador e a aplicação do instrumento em número maior de participantes, conforme previsto no projeto.

A aplicação em dois participantes foi importante para observar o recurso lúdico e agradável da EPIA, no entanto, há ajustes a serem realizados.

A Escala EPIA não será oferecida para comercialização sem apresentar resultados satisfatórios de conteúdo e precisão e ser aprovada como favorável no SATEPSI.

Considerando que psicólogos frequentemente utilizam testes em sessões psicoterapêuticas, é importante destacar que as Escalas Wechsler apresentam custos elevados, o que dificulta o acesso à avaliação e ao diagnóstico. Nesse contexto, uma escala informatizada e acessível pode representar uma solução viável, promovendo maior agilidade na aplicação e na obtenção de resultados, especialmente em serviços-escola e clínicas. Além disso, essa proposta configura uma inovação científica ao substituir o método tradicional de tabulação manual por um sistema automatizado, facilitando a análise e interpretação por parte dos profissionais.

REFERÊNCIAS

[1] D. Wechsler, Escala de Inteligência Wechsler para Adultos. São Paulo: Casa do Psicólogo, 2004.

[2] KAPP, Karl M. The gamification of learning and instruction: game-based methods and strategies for training and education. John Wiley & Sons, 2012.

[3] M. H. B. SANCHES. Jogos digitais, gamificação e autoria de jogos na educação. Editora Senac São Paulo, 2021.

[4] E. LEE, et al. Hierarchical helical assembly of conjugated poly 3-hexylthiophene)-block-poly (3triethylene glycol thiophene) diblock copolymers. Journal of the American Chemical Society, v. 133, n. 27, p. 10390-10393, 2011.

[5] F. de A. A. Rodrigues, "Inteligência DWRI". Recisatec-Revista Científica Saúde E Tecnologia-ISSN 2763-8405, v. 2, n. 12, p. e212232-e212232, 2022.

[6] D. E. Papalia and G. Martorell, Desenvolvimento Humano. Porto Alegre, Artmed, 14ª ed. 2022.

[7] M. L. Martín Jorge, "The Bellevue Intelligence Tests (Wechsler, 1939): ¿una medida de la inteligencia como capacidad de adaptación?" Revista de historia de la psicología. 33(3), 49-66, 2012.

[8] C. S. Hutz, D. R. Bandeira and C. M. Trentini, Psicometria. Porto Alegre: Artmed, 2015.

[9] J. S. Krug, C. M. Trentini, D. R. Bandeira and C. S. Hutz, Psicodiagnóstico: avaliação psicológica. Porto Alegre: Artmed, 2016.

[10] B. S. G. Werlang, A. E. de Villemor-Amaral, R. S. G. F. Nascimento. "Avaliação psicológica: diretrizes na regulamentação da profissão". In: CONSELHO FEDERAL DE PSICOLOGIA (Org.). Avaliação psicológica: diretrizes na regulamentação da profissão. Brasília: CFP, 2010. p. 1–196.

[11] C. E. Flores-Mendoza, E. Nascimento, A. V. Castilho, "A crítica desinformada aos testes de inteligência". Estudos de Psicologia (Campinas), v. 19, p. 17-36, 2002.

[12] S. S. S. Cóser, F. K. Miguel, "Avaliação psicológica e o uso das tecnologias digitais de informação e comunicação: possibilidades a partir das regulamentações do Conselho Federal de Psicologia". In: R. B. SILVA, M. B. SEI, (Orgs.). Psicologia e contemporaneidade: pesquisas, experiências e reflexões. Londrina: UEL, 2021.

[13] R. Primi, "Avaliação psicológica no Brasil: fundamentos, situação atual e direções para o futuro. Psicologia": Teoria e Pesquisa, Brasília, v. 26, n. especial, p. 25–36, 2010.

[14] M. A. Silva, "Testes informatizados para a avaliação psicológica e educacional". Psico-USF, v. 16, p. 127-129, 2011.

[15] Conselho Federal de Psicologia. Nota Técnica nº 07/2019/GTEC/CG: orienta psicólogas(os) sobre a utilização de testes psicológicos em serviços realizados por meio de tecnologias de informação e da comunicação. Brasília: CFP, 2019.

[16] C. M. A. Gomes, D. F. Nascimento, "Projeto de acesso aberto e gratuito à Bateria de Fatores Cognitivos de Alta-Ordem (BAFACALO): o Teste de Compreensão Verbal 1 do Conjunto de Testes de Inteligência Cristalizada". Preprint. https://doi.org/10.13140/RG, v. 2, n. 22663.32165, p. 1, 2021.

[17] I. Hazin, I. Fernandes, E. Gomes, D. Garcia, "Neuropsicologia no Brasil: passado, presente e futuro". Estudos e Pesquisas em Psicologia, Rio de Janeiro, v. 18, n. 4, p. 1137–1154, 2018.

[18] R. Bilder, S. Reise, "Neuropsychological tests of the future: How do we get there from here?" The Clinical Neuropsychologist, v. 33, n. 2, p. 220-245, 2019. DOI: 10.1080/13854046.2018.1521993.

[19] R. Bilder, "Neuropsychology 3.0: evidence-based science and practice". Journal of the International Neuropsychological Society, v. 17, n. 1, p. 7–13, 2011.

[20] J. Miller, W. Barr, "The technology crisis in neuropsychology. Archives of Clinical Neuropsychology", v. 32, n. 5, p. 541–554, 2017. DOI: 10.1093/arclin/acx050.

[21] R. H. Sampieri, C. F. Lucio, M. P. B. Collado, Metodologia de Pesquisa, 5ª ed. Porto Alegre: Penso, 2013.

[22] A. C. Gil, Como elaborar projetos de pesquisa, 4ª ed. São Paulo: Atlas, 2010.

[23] S. P. M. Marques, E. C. S Gomes, M. M. da M. Martins, "Tecnologias Digitais de Informação e Comunicação no ensino em ciências: tendências dos artigos publicados nos ENPECs 2015, 2017 e 2019. Educitec" - Revista de Estudos e Pesquisas sobre Ensino Tecnológico, v. 7, p. 1–20, jan./dez. 2021. Disponível em: Educitec - artigo completo. Acesso em: 31 ago. 2025.

[24] A. L. F. Souza, A. G. S. Vasconcelos, A. C. Santos, L. V. S. Soares, I. M. S. F. Sousa, V. A. P. Oliveira. Jogos Digitais Sérios Aplicados à Área de Psicologia no Brasil. In: Anais do XXIV Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames); 2025; Salvador, BA. Sociedade Brasileira de Computação; 2025.

[25] A. C. S. Lima. Validação de conteúdo por juízes especialistas do Questionário de Atitudes em Relação a Adultos com Deficiência Auditiva [Trabalho de Conclusão de Curso]. São Paulo: Universidade Federal de São Paulo, Escola Paulista de Medicina; 2025.

[26] Selau T. Construção e estudo de evidências de validade e fidedignidade da Escala de Funcionamento Adaptativo para Deficiência Intelectual (EFA-DI) [Dissertação de Mestrado]. Porto Alegre: Universidade Federal do Rio Grande do Sul, Instituto de Psicologia; 2020.

[27] American Psychological Association. Standards for Educational and Psychological Testing. Washington, DC: APA; 2014.

[28] AERA, APA, NCME. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.

[29] M. C. R. A Joly, "Análise da produção científica em avaliação psicológica informatizada". Avaliação Psicológica: Interamerican Journal of Psychological Assessment, v. 3, n. 2, p. 121-129, 2004.

Class-wise Evaluation Reveals Hidden Errors Behind High Global Performance in Deep Learning Classification of Imbalanced OCT Images

Kerllon Fontenele de Andrade[1], Alina Dyakova[2], Konstantin Boshkov[2], Denis Leonov[3], Flávio Buiochi[4], Marcos Sales Guerra Tsuzuki[4], and José Francisco Silva Costa-Júnior[1,4]

[1]Postgraduate Program in Biomedical Engineering, Brazil University, São Paulo, Brazil
[2]National Research University "Moscow Power Engineering Institute", Moscow, Russia
[3]Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia
[4]Department of Mechatronics and Mechanical Systems Engineering, University of São Paulo, São Paulo, Brazil

***Abstract*—Deep learning models have achieved excellent performance in retinal optical coherence tomography (OCT) image classification. However, overall performance metrics may provide an overly optimistic assessment when datasets are imbalanced, potentially masking poor recognition of clinically relevant minority classes. This study investigated the impact of class imbalance on OCT classification using an InceptionV3 model with transfer learning trained on the CNV, DME, DRUSEN, and NORMAL classes of the OCT2017 dataset. Two training strategies were compared: the original imbalanced dataset and an imbalance-mitigation approach combining offline augmentation of minority classes with class weighting. Although the model trained on the imbalanced dataset achieved high overall accuracy (94.0%), class-wise analysis revealed substantially lower recall for the DRUSEN class (84.8%), with frequent misclassification as CNV. In contrast, the imbalance-mitigation strategy increased overall accuracy to 96.1% and improved recall for the DME and DRUSEN classes to 96.0% and 94.4%, respectively, markedly reducing systematic confusion with CNV. The paired comparison of the 1,000 test predictions using the exact McNemar test showed that the difference between the models was statistically significant (p = 0.0055). These findings demonstrate that high overall performance metrics may conceal clinically important classification errors and underscore the importance of reporting class-wise performance measures, including precision, recall, F1-score, and confusion matrices, to ensure the reliable evaluation of deep learning models intended for diagnostic support.**

***Index Terms*—optical coherence tomography, deep learning, convolutional neural network, class imbalance, InceptionV3**

I. Introduction

Convolutional Neural Networks (CNNs) have achieved strong performance in medical image classification by learning discriminative features directly from imaging data, reducing the need for handcrafted feature extraction and supporting computer-aided diagnosis [1], [2]. Transfer learning with pre-trained CNNs is particularly effective for limited or complex medical datasets because it accelerates convergence while preserving high classification performance [1], [3].

Optical Coherence Tomography (OCT) is a primary non-invasive modality for retinal diagnosis, providing high-resolution cross-sectional images of retinal structures [4]. Deep learning has therefore been widely applied to classify choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal retina [3], [5], [6]. InceptionV3-based models have achieved accuracies between 96.6% and 98.4% on the OCT2017 dataset [7], depending on the experimental protocol and data partitioning strategy [2], [4], [8].

Despite these results, reliance on overall accuracy may provide an overly optimistic assessment when the dataset is imbalanced. In OCT2017, CNV and NORMAL contain substantially more images than DME and DRUSEN [7], potentially biasing optimization toward the majority classes. Consequently, high global accuracy may conceal poor recognition of under-represented diseases and clinically relevant errors [9]. Class-specific precision, recall, F1-score, and confusion matrices are therefore necessary to reveal performance differences among retinal categories.

Data augmentation, class weighting, and undersampling have been proposed to mitigate class imbalance [9]–[12]. Previous OCT studies have applied these strategies to transfer-learning architectures such as InceptionV3 and Xception, generally reporting high global and class-wise performance [13], [14]. However, many relied exclusively on balanced datasets or mitigation techniques [13], [15], without directly quantifying how the original OCT2017 distribution affects class-specific results. Thus, the extent to which imbalance inflates global metrics and reduces minority-class recognition remains insufficiently investigated [9], [16].

This study addresses this gap by comparing two InceptionV3 models under identical experimental conditions: one trained with the original OCT2017 distribution and another using offline data augmentation and class weighting. The objective is to determine whether global metrics conceal minority-class errors and whether imbalance mitigation improves without substantially compromising CNV and NORMAL performance. The comparison of global and class-specific metrics highlights the importance of class-wise evaluation for reliable OCT-based diagnostic support.

The remainder of this paper is organized as follows. Sec-

tion II describes the dataset, network architecture, imbalance-mitigation strategies, training configuration, and evaluation protocol. Section III presents the training behavior and classification results. Section IV discusses the main findings, clinical implications, and study limitations. Finally, Section V summarizes the conclusions and outlines directions for future work.

II. Methods

This section describes the experimental methodology adopted to evaluate the influence of class imbalance on retinal OCT image classification. The proposed protocol was designed to ensure a controlled comparison between two InceptionV3-based models: one trained using the original imbalanced distribution of the OCT2017 dataset and another trained with offline data augmentation and class weighting. Both models used the same image preprocessing pipeline, network architecture, optimizer, learning rate, loss function, and independent test set, so that differences in performance could be attributed primarily to the imbalance-mitigation strategy.

The following subsections present the OCT2017 dataset and data partitioning procedure, the InceptionV3 architecture and transfer-learning configuration, the strategies used to mitigate class imbalance, the model training parameters, and the evaluation protocol. Performance was assessed using global and class-specific metrics, confusion matrices, and ROC-AUC analysis, with particular attention to the underrepresented DME and DRUSEN classes.

A. Dataset

This study used the publicly available OCT2017 dataset from the Mendeley Data repository [7]. The dataset comprises retinal OCT images organized into four categories: CNV, DME, DRUSEN, and normal retina. The original training set is markedly imbalanced, containing 37,205 CNV images, 26,315 NORMAL images, 11,348 DME images, and 8,617 DRUSEN images. Consequently, CNV and NORMAL are the majority classes, whereas DME and DRUSEN are substantially underrepresented [7].

To ensure a fair comparison between the model trained on the original OCT2017 class distribution and the model trained using data augmentation and class weighting, both models were evaluated using the same independent test set. Consequently, any differences in classification performance could be attributed to the training strategy rather than to differences in the evaluation data. The test partition comprised 1,000 images equally distributed across the four categories (250 images per class). As reported by the original dataset authors, the test images were obtained from patients who were not represented in the training partition [6], ensuring complete patient-level separation between the original training and test sets.

The original OCT2017 training set was further divided into training and validation subsets. Patient identifiers were not available in the released training data; therefore, patient-level separation between these two subsets could not be guaranteed. The independent test set remained completely isolated from this partitioning procedure and, as reported by [6], contained images from patients not represented in the original training set. All images were loaded as three-channel inputs and resized to 299×299 pixels, corresponding to the input size required by the InceptionV3 architecture [17]. Before being fed into the network, the images were preprocessed using the InceptionV3-specific preprocessing function with ImageNet-pretrained weights [18].

B. Architecture and Transfer Learning

The experiments were conducted using the InceptionV3 architecture, originally developed for image recognition and characterized by Inception modules that perform convolutional operations at multiple scales [17]. A transfer learning strategy was adopted by initializing the network with ImageNet-pretrained weights [19], [20]. The convolutional base was loaded without its original classification layers and kept frozen during training to serve as a feature extractor. This approach enables the reuse of representations learned from large-scale image datasets while adapting the classifier to the OCT classification task [20], [21]. A global average pooling layer was added to the output of the convolutional base, followed by a dense layer with four neurons and a Softmax activation function [19]. Each output neuron corresponded to one of the four classes: CNV, DME, DRUSEN, and NORMAL.

C. Strategies for Mitigating Class Imbalance

Two experimental scenarios were evaluated. In the first, referred to as the imbalanced model, InceptionV3 was trained using the original OCT2017 training set without applying any techniques to compensate for class imbalance. This scenario served as the reference model and replicated the methodology proposed in [6].

In the second scenario, referred to as the imbalance-mitigated model, offline data augmentation was applied to the minority classes (DME and DRUSEN). The augmented images were generated before training, stored on disk, and added to the original training set. The training-validation split was performed subsequently by the image generators over this augmented training set. Consequently, images derived from the same original image may have been assigned to the training and validation subsets. Each image was created from a randomly selected original image and subjected to a random combination of geometric transformations. Data augmentation included rotations of up to $15°$, horizontal and vertical translations of up to 10% of the image dimensions, and horizontal flipping. Regions introduced by these transformations were filled using the nearest-neighbor interpolation method. During this process, all augmented images were resized to 256×256 pixels and saved in JPEG format. The procedure was repeated until the predefined numbers of additional DME and DRUSEN images were generated. The objective was not to equalize the number of samples across all classes, but to reduce the representation gap of the minority classes. Moderate geometric transformations were deliberately adopted to introduce variability while avoiding excessively

aggressive modifications that could substantially distort the original retinal structures.

In addition to data augmentation, class weighting was applied during training of the imbalance-mitigated model. After splitting the data into training and validation subsets, class weights were computed automatically from the class frequencies in the training subset using the `compute_class_weight` function from the scikit-learn library [22] with the parameter `class_weight = balanced`.

The weight assigned to each class was calculated as

$$w_j = \frac{n_{\text{samples}}}{n_{\text{classes}} \cdot n_j}, \tag{1}$$

where w_j denotes the weight assigned to class j, n_{samples} is the total number of training samples, n_{classes} is the number of classes, and n_j is the number of training samples belonging to class j.

Consequently, classes with lower frequencies received higher weights during loss computation, increasing the contribution of their errors to the optimization process. The two models used the same image preprocessing pipeline, feature extraction procedure, network architecture, optimizer, learning rate, loss function, validation proportion, and independent test set. The imbalanced model preserved the original training protocol adopted in the reference study [6], whereas the imbalance-mitigated model additionally used early stopping to reduce unnecessary training once validation loss ceased to improve. In both scenarios, the final weights corresponded to the epoch with the lowest validation loss observed during training. Nevertheless, because early stopping was applied only to the imbalance-mitigated model, the comparison presented in this study is between two complete training protocols rather than between two conditions differing in a single factor.

D. Training Configuration

The models were trained using the Adam optimizer [23] with a learning rate of 0.001 and the categorical cross-entropy loss function, which is appropriate for multiclass classification with Softmax output [19]. The training and validation generators used a batch size of 32 images. During testing, images were processed individually (batch size = 1) without shuffling to preserve the correspondence between image order, ground-truth labels, and model predictions.

For both the imbalanced and imbalance-mitigated models, 20% of the training images were reserved for validation, whereas the remaining 80% were used for model training. This split was implemented using the 'validation_split = 0.2' parameter of the image generators.

In the imbalanced scenario, the model was trained for 100 epochs while preserving the original OCT2017 class distribution, following the protocol described in the reference study [6]. Training and validation accuracy and loss were monitored throughout the training process. A model checkpoint monitored the validation loss and saved the weights corresponding to its minimum value, which were subsequently used for test-set evaluation.

In the imbalance-mitigated scenario, training was performed for up to 100 epochs. Validation loss was monitored using both models checkpointing and early stopping, with a patience of 15 epochs and restoration of the best model weights. The checkpoint preserved the weights associated with the lowest validation loss for final evaluation.

Training and validation accuracy and loss curves were analyzed to assess model convergence. The final comparison was performed exclusively between the imbalanced and imbalance-mitigated models using the same independent test set.

E. Evaluation Protocol

Model performance was evaluated using the same independent test set described previously, comprising 1,000 images equally distributed across the four classes. Performance was assessed using both overall accuracy and class-specific metrics. Precision, recall, and F1-score were calculated for each class. Recall represents the proportion of correctly identified images within a given class, whereas precision represents the proportion of correct predictions among all predictions assigned to that class. The F1-score was computed as the harmonic mean of precision and recall.

Macro- and weighted-average metrics were also calculated. The macro average assigns equal importance to all classes regardless of class size, whereas the weighted average accounts for the number of samples in each class. Because the test set was balanced, both averages produced identical or nearly identical values.

Absolute and normalized confusion matrices were used to identify the main patterns of misclassification. Receiver operating characteristic (ROC) curves were generated using the one-vs-rest approach, and the area under the curve (AUC) was calculated for each class as well as for the micro and macro averages. The macro-AUC was computed as the arithmetic mean of the class-specific AUC values.

The comparison between the two models included overall accuracy, class-wise precision, recall, F1-score, macro-F1, weighted F1-score, confusion matrices, and AUC. Ninety-five percent confidence intervals for overall accuracy and for class-wise recall were computed using the Wilson score interval. Because both models were evaluated on the same 1,000 test images, their paired classification outcomes were additionally compared using the exact two-sided McNemar test. The test was based on the discordant cases in which only one of the two models correctly classified a given image, with statistical significance defined at $p < 0.05$. Particular attention was given to discrepancies between global and class-specific performance, especially for DME and DRUSEN, which were underrepresented in the original training set. The analysis also evaluated whether imbalance mitigation strategies improved the recognition of these classes without substantially compromising the performance of CNV and NORMAL.

III. Results

This section presents the experimental results for the imbalanced and imbalance-mitigated models. It begins with the final

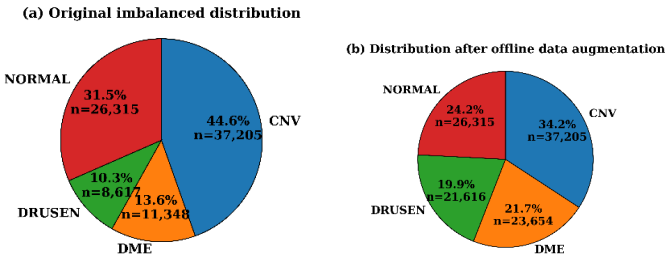


Fig. 1. Class distribution before and after imbalance mitigation: (a) original imbalanced distribution and (b) distribution after offline augmentation of the DME and DRUSEN classes.

distribution of the training data after applying the imbalance-mitigation strategy, followed by an analysis of the training behavior of both models based on the learning curves. The performance of each model is then evaluated on the independent test set using overall and class-specific metrics, followed by a comparative analysis to assess the impact of the imbalance mitigation strategies on classification performance.

A. Final Distribution of the Training Data

Figure 1 compares the original imbalanced training distribution with the distribution obtained after offline data augmentation. The original dataset retained the OCT2017 class distribution, whereas offline data augmentation increased the DME and DRUSEN classes to 23,654 and 21,616 images, respectively, while the CNV and NORMAL classes remained unchanged. In total, 12,306 DME images and 12,999 DRUSEN images were generated through data augmentation. After augmentation, the ratio between the largest class (CNV, 37,205 images) and the smallest class (DRUSEN, 21,616 images) was approximately 1.72:1. Thus, the procedure reduced the original class imbalance but did not produce a numerically uniform class distribution.

B. Training Behavior

1) Imbalanced Model: Figures 2a and 2b show the training and validation accuracy and loss, respectively, over 100 training epochs for the imbalanced model. In both figures, the blue and orange curves represent the training and validation subsets, respectively. During the initial training stage, accuracy increased rapidly, from 83.45% to approximately 89% for the training subset and from 86.08% to approximately 88% for the validation subset. Over the same period, training and validation losses decreased from 0.4635 to approximately 0.30 and from 0.3946 to approximately 0.32, respectively.

After this initial phase, the learning curves became more stable. Training accuracy continued to increase gradually, whereas validation accuracy remained predominantly between 88% and 89%. The highest validation accuracy (89.14%) was achieved at epoch 54, whereas the lowest validation loss (0.3079) was observed at epoch 63. Because validation loss was used as the model selection criterion, the weights obtained at epoch 63 were selected for the final evaluation on the test

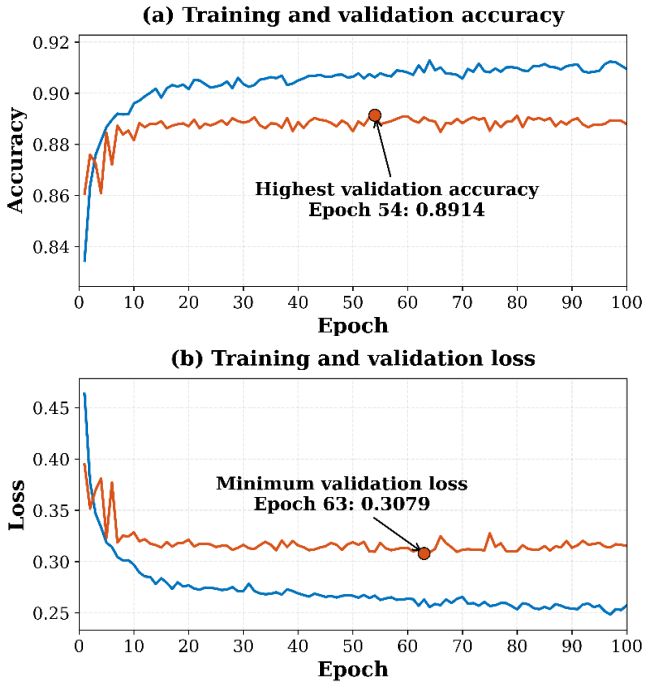


Fig. 2. Training and validation accuracy and loss of the imbalanced model over 100 epochs: (a) accuracy and (b) loss.

set. During the remaining epochs, validation loss fluctuated without further sustained improvement.

At the end of training, the model achieved accuracies of 90.95% and 88.79% on the training and validation subsets, respectively, with corresponding losses of 0.2574 and 0.3153. Overall, the learning curves indicate rapid convergence during the initial training stage, followed by stable validation performance throughout the remainder of training.

2) Imbalance-Mitigated Model: Figures 3a and 3b illustrate the training and validation accuracy and loss, respectively, over the 36 training epochs of the imbalance-mitigated model, with the blue and orange curves representing the training and validation subsets. During the initial epochs, model performance improved rapidly, with training accuracy increasing from 85.02% during the first epoch to above 90% by the fifth epoch. Validation accuracy increased from 77.48% to 84.11% by epoch 7, whereas training and validation losses decreased from 0.4272 to 0.2780 and from 0.5931 to 0.4459, respectively.

After this initial phase, training accuracy improved steadily, reaching 91.92% at epoch 36, whereas validation accuracy exhibited greater fluctuations. The highest validation accuracy (84.40%) and the lowest validation loss (0.4389) were reached at epoch 21. Thereafter, validation loss fluctuated without further improvement.

Because validation loss did not improve over the subsequent 15 epochs, training was terminated at epoch 36 by the early stopping criterion. As specified by the training configuration, the weights corresponding to the lowest validation loss at epoch 21 were restored automatically and used for the final evaluation on the test set. Overall, the learning curves indicate continuous

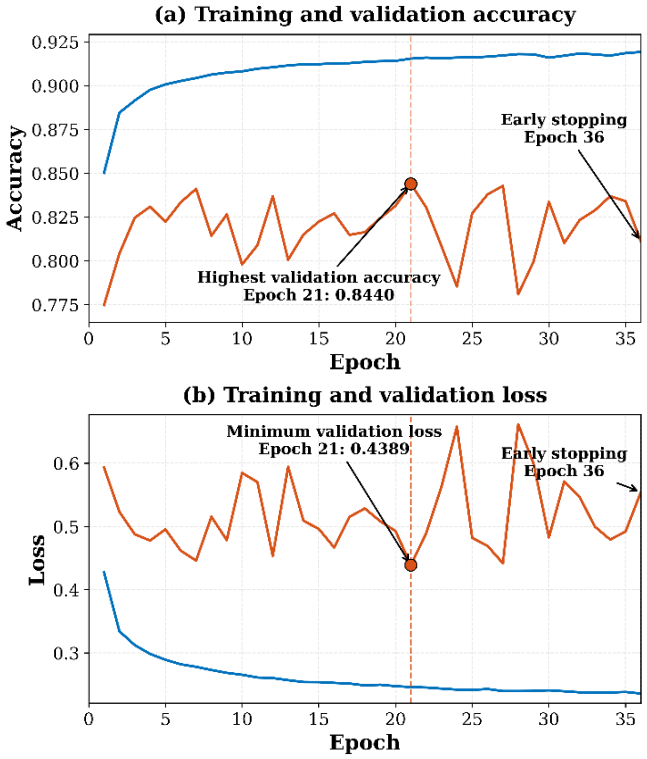


Fig. 3. Training and validation accuracy and loss of the imbalance-mitigated model over 36 epochs: (a) accuracy and (b) loss.

improvement in training performance, accompanied by greater variability in the validation subset throughout the remainder of training.

C. Performance of the Imbalanced Model on the Test Set

The imbalanced model achieved an overall accuracy of 94.00% on the test set. Because the test set contained 250 images per class, the macro and weighted averages were identical, yielding a precision of 94.73%, a recall of 94.00%, and an F1-score of 94.04%.

Despite the high overall accuracy, class-wise analysis showed that performance was not uniform across the four classes. For the CNV class, the model achieved a precision of 83.16%, a recall of 98.80%, and an F1-score of 90.31%. The corresponding values for DME were 98.72%, 92.40%, and 95.45%, respectively. DRUSEN achieved a precision of 98.60%, a recall of 84.80%, and an F1-score of 91.18%, representing the lowest recall among the four classes. For the NORMAL class, the model achieved a precision of 98.43%, a recall of 100%, and an F1-score of 99.21%.

Figure 4 presents the confusion matrices for the imbalanced model, with the absolute matrix on the left and the normalized matrix on the right. According to the absolute confusion matrix, 247 of the 250 CNV images were correctly classified, whereas two were misclassified as DME and one as DRUSEN. Among the DME images, 231 were correctly classified, while 13 were misclassified as CNV, two as DRUSEN, and four as NORMAL. For the DRUSEN class, 212 images were correctly classified,

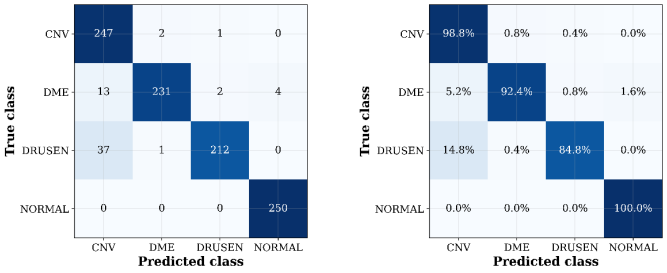


Fig. 4. Confusion matrices of the imbalanced model on the test set: absolute confusion matrix (left) and normalized confusion matrix (right).

37 were misclassified as CNV, and one as DME. All 250 NORMAL images were correctly classified.

The predominant error pattern was the misclassification of DRUSEN images as CNV, affecting 14.8% of the samples in this class. DME images were also misclassified as CNV in 5.2% of cases. These findings indicate that the high overall accuracy did not translate into uniform performance across the four classes, primarily because of the lower recall for DRUSEN.

Discriminative performance was also evaluated using one-vs-rest ROC analysis. The AUC values were 0.9975 for CNV, 0.9977 for DME, 0.9970 for DRUSEN, and 0.9993 for NORMAL. The micro-AUC was 0.9955, whereas both the macro- and weighted-average AUCs were 0.9979. These results indicate excellent overall discriminative performance; however, the confusion matrices reveal that classification errors persisted between specific classes.

D. Performance of the Imbalance-Mitigated Model on the Test Set

The imbalance-mitigated model was evaluated using the independent test set of 1,000 images, equally distributed among CNV, DME, DRUSEN, and NORMAL. It achieved an overall accuracy of 96.10%, with relatively uniform performance across all classes.

Figure 5 presents the confusion matrices of the imbalance-mitigated model, with the absolute matrix on the left and the normalized matrix on the right. According to the absolute confusion matrix, the model correctly classified 238 CNV, 240 DME, 236 DRUSEN, and 247 NORMAL images. Among the CNV images, eight were misclassified as DME and four as DRUSEN. Among the DME images, two were misclassified as CNV, three as DRUSEN, and five as NORMAL. Among the DRUSEN images, 13 were misclassified as CNV and one as DME. Among the NORMAL images, three were misclassified as DRUSEN.

Class-specific metrics were consistently high. For CNV, precision, recall, and F1-score were 94.07%, 95.20%, and 94.63%, respectively. The corresponding values were 96.39%, 96.00%, and 96.19% for DME; 95.93%, 94.40%, and 95.16% for DRUSEN; and 98.02%, 98.80%, and 98.41% for NORMAL. Because the test set was balanced, macro precision, macro recall, macro-F1, and weighted F1-score were all 96.10%.

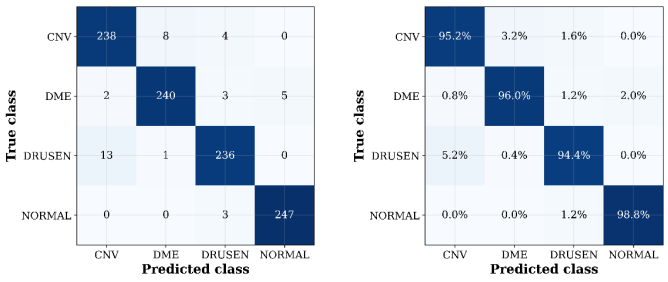


Fig. 5. Confusion matrices of the imbalance-mitigated model on the test set: absolute confusion matrix (left) and normalized confusion matrix (right).

Discriminative performance was further evaluated using one-vs-rest ROC analysis. The AUC values were 0.9952 for CNV, 0.9960 for DME, 0.9964 for DRUSEN, and 0.9986 for NORMAL. The micro-AUC was 0.9970, while the macro- and weighted-average AUCs were both 0.9966.

Overall, the imbalance-mitigated model achieved an accuracy and macro-F1 score of 96.10%, with AUC values above 0.995 for every class. Although DRUSEN had the lowest recall, its F1-score remained high at 95.16%. These results indicate that the imbalance-mitigation strategy promoted more uniform classification performance without compromising overall accuracy.

E. Comparison Between the Imbalanced and Imbalance-Mitigated Models

Both models were evaluated using the same independent test set of 1,000 images, equally distributed among CNV, DME, DRUSEN, and NORMAL. Therefore, performance differences can be attributed to the imbalance-mitigation strategy rather than to variations in the evaluation data.

The main global performance metrics improved consistently. Accuracy increased from 94.00% for the imbalanced model to 96.10% for the imbalance-mitigated model. Macro-F1 increased from 94.04% to 96.10%, while macro precision and macro recall increased from 94.73% and 94.00% to 96.10%, respectively. The weighted F1-score showed a similar improvement, indicating more uniform performance across the four classes. The 95% confidence intervals for overall accuracy were 92.35–95.31% for the imbalanced model and 94.71–97.13% for the imbalance-mitigated model. The corresponding intervals for class-wise recall were 96.53–99.59% and 91.80–97.23% for CNV, 88.44–95.08% and 92.79–97.81% for DME, 79.83–88.72% and 90.82–96.64% for DRUSEN, and 98.49–100.00% and 96.53–99.59% for NORMAL. The paired analysis of the 1,000 test images identified 53 discordant predictions between the two models. Among these, 37 images were incorrectly classified by the imbalanced model but correctly classified by the imbalance-mitigated model, whereas 16 showed the opposite pattern. The exact two-sided McNemar test indicated a statistically significant difference between the classifiers ($p = 0.0055$).

Both models achieved high AUC values. The micro-AUC increased from 0.9955 to 0.9970, whereas the macro-AUC changed slightly from 0.9979 to 0.9966. This difference arises because AUC measures discrimination across multiple thresholds, whereas accuracy, recall, and F1-score reflect performance at the selected classification threshold. Thus, despite the small reduction in macro-AUC, the imbalance-mitigated model achieved better global and class-specific performance.

The class-wise recall values differed markedly between the two models. For the imbalanced model, recall was 98.80% for CNV, 92.40% for DME, 84.80% for DRUSEN, and 100.00% for NORMAL. After applying the imbalance-mitigation strategies, the corresponding values were 95.20%, 96.00%, 94.40%, and 98.80%.

The largest gain occurred for DRUSEN, whose recall increased by 9.60 percentage points. DME recall increased by 3.60 points, whereas CNV and NORMAL recall decreased by 3.60 and 1.20 points, respectively. Nevertheless, recall remained above 95% for both majority classes, showing that improved minority-class recognition did not substantially compromise their performance.

The confusion matrices support this result. DRUSEN images misclassified as CNV decreased from 37 cases (14.80%) to 13 (5.20%), while DME-to-CNV errors decreased from 13 to two. The imbalance-mitigation strategy reduced the tendency of the imbalanced model to assign minority-class images to CNV, the largest training class.

The same trend was observed for the class-wise F1-scores. The F1-score increased from 90.31% to 94.63% for CNV, from 95.45% to 96.19% for DME, and from 91.18% to 95.16% for DRUSEN. For NORMAL, it decreased slightly from 99.21% to 98.41%.

The increase in the CNV F1-score, despite its lower recall, resulted from the improvement in precision from 83.16% to 94.07%. Fewer DME and DRUSEN images were incorrectly classified as CNV, reducing false positives and making CNV predictions more reliable.

Overall, the imbalance-mitigation strategy produced more uniform performance across all classes. It improved DME and DRUSEN recognition, reduced systematic confusion with CNV, and increased accuracy and macro-F1 without substantially compromising CNV and NORMAL. These results show that global metrics alone may conceal clinically relevant class-specific errors and reinforce the importance of class-wise evaluation in imbalanced medical image classification.

IV. Discussion

The InceptionV3 model trained with the original OCT2017 distribution achieved 94.0% accuracy and a macro-F1 score of 94.04%. However, DRUSEN recall was only 84.8%, demonstrating that strong global metrics may conceal clinically relevant weaknesses in underrepresented classes. Therefore, accuracy should be interpreted together with class-specific precision, recall, F1-score, and confusion matrices.

The confusion matrices showed that the imbalanced model most frequently misclassified DRUSEN as CNV, with the same tendency observed for DME to a lesser extent. Because CNV

and NORMAL were the largest classes, they exerted greater influence during optimization and shifted decision boundaries toward the majority classes. Minority diseases consequently became more difficult to distinguish, particularly when their imaging characteristics overlapped with CNV. Similar results were reported in [5], where a fusion model achieved 98.7% accuracy and 99.1% AUC but still identified DRUSEN as the most difficult class because of its limited representation and structural similarity to CNV. Likewise, [24] showed that aggregate metrics may hide clinically important class-specific deficiencies.

To mitigate this effect, offline data augmentation was combined with class weighting, improving minority-class representation without discarding original images. Accuracy increased from 94.0% to 96.1%, while DRUSEN recall rose from 84.8% to 94.4% and DME recall from 92.4% to 96.0%. This improvement was also supported by the paired statistical analysis. Among the 53 test images for which the models disagreed in correctness, the imbalance-mitigated model corrected 37 errors made by the imbalanced model, whereas the reverse occurred in 16 cases. The exact McNemar test confirmed that this asymmetry was statistically significant ($p = 0.0055$). The gains in macro recall and macro-F1 indicate a more uniform distribution of performance across the four retinal categories.

The confusion matrices further confirmed the reduction in majority-class bias. DRUSEN-to-CNV errors decreased from 37 cases (14.8%) to 13 (5.2%), while DME-to-CNV errors fell from 13 to two. Although CNV recall declined from 98.8% to 95.2%, its precision increased from 83.16% to 94.07%, raising the CNV F1-score from 90.31% to 94.63%. Thus, the modest decrease in sensitivity was offset by a substantial reduction in false-positive CNV predictions, illustrating why precision, recall, and F1-score should be interpreted jointly. Considering recall alone would favor the model trained on the original class distribution, whereas the combined metrics show that the imbalance-mitigated model produced more dependable predictions by substantially reducing false-positive classifications.

Unlike undersampling strategies, such as that evaluated in [24], the proposed approach reduced imbalance without removing original training images, which is relevant in medical imaging because annotated data are costly and difficult to obtain. The imbalance-mitigated model also produced results consistent with previous OCT2017 transfer-learning studies. Adel et al. [13], for example, reported accuracies between 93% and 98% for balanced InceptionV3 and Xception models. However, the present study compared the original distribution with the distribution obtained after imbalance mitigation under identical architectural, preprocessing, and testing conditions. Therefore, the observed improvements can reasonably be attributed to the imbalance-mitigation strategy rather than to classifier modifications.

From a clinical perspective, reducing false-positive CNV predictions may prevent DME and DRUSEN cases from being incorrectly assigned to CNV, potentially avoiding unnecessary referrals, additional examinations, and inappropriate management. Recall remained above 95% for CNV and NORMAL, while the largest gains occurred for DME and DRUSEN. This redistribution is advantageous for clinical decision-support systems, where consistent recognition across diseases is more important than maximizing performance for the most frequent categories.

The ROC results should also be interpreted together with threshold-dependent metrics. Both models achieved AUC values above 0.99 for all classes, yet the imbalanced model concentrated errors in minority categories. AUC measures ranking performance across several decision thresholds, whereas accuracy, precision, recall, and F1-score describe performance at the selected operating point. Consequently, ROC curves alone are insufficient to characterize multiclass medical classifiers and should be analyzed together with confusion matrices and class-wise metrics.

The lower validation accuracy of the imbalance-mitigation model compared with its test accuracy does not necessarily indicate overfitting. The validation subset preserved the distribution produced after augmentation, whereas the independent test set contained equal numbers of images from each class. Because these datasets followed different distributions, their metrics are not directly comparable. Moreover, model selection based on minimum validation loss, early stopping, and restoration of the best weights reduced the likelihood of overfitting.

This study has three main limitations. First, it was restricted to OCT2017, so the findings may not generalize to images acquired with different devices, protocols, institutions, clinical settings, or patient populations. External multicenter validation is therefore required. Second, data augmentation and class weighting were applied simultaneously, preventing individual contributions from being isolated. Third, the validation subset of the imbalance-mitigated model was defined after offline data augmentation. Images derived from the same original image may therefore have been assigned to both the training and the validation subsets, and patient identifiers were not available to enforce patient-level separation between them. This limitation is restricted to the selection of the final weights by validation loss and to the early stopping criterion, since no reported performance metric was computed on the validation subset. All results reported in this study were obtained on the independent test set, which contained no augmented images and, as reported in [6], comprised images from patients not represented in the original training set. Future ablation studies should evaluate these techniques separately and compare them with undersampling, oversampling, focal loss, class-balanced loss, and cost-sensitive learning. Generative adversarial networks and diffusion models could also be investigated for minority-class synthesis, together with probability calibration and optimized decision thresholds.

The purpose of this study was not to maximize accuracy through increasingly complex architectures, but to determine how class imbalance affects model evaluation. The results show that excellent aggregate performance may still conceal systematic deficiencies in clinically important minority classes.

Therefore, a classifier with more balanced recognition across diseases may be preferable to one with slightly higher global performance but persistent errors in underrepresented categories.

Overall, class imbalance produced not only a modest reduction in accuracy, but also a systematic bias against minority retinal diseases. Combining offline augmentation with class weighting improved their recognition while preserving high global performance. These findings reinforce the importance of class-wise evaluation and imbalance mitigation in clinically reliable deep learning systems for retinal OCT classification.

V. Conclusions and Future Works

This study evaluated class imbalance in retinal OCT classification using two InceptionV3 models trained under identical conditions. The imbalanced model achieved 94.0% accuracy and a macro-F1 score of 94.04%, but DRUSEN recall was only 84.80%, with 14.80% of its images misclassified as CNV. This result shows that high global performance can conceal systematic errors in minority classes.

Combining offline data augmentation with class weighting increased accuracy and macro-F1 to 96.10%. Recall improved from 92.40% to 96.00% for DME and from 84.80% to 94.40% for DRUSEN. DRUSEN-to-CNV errors decreased from 37 to 13, while DME-to-CNV errors decreased from 13 to two. Performance for CNV and NORMAL remained above 95% recall. The paired analysis suggests that the imbalance-mitigation strategy may have contributed to improved classification performance compared with the imbalanced model.

These findings demonstrate that accuracy and AUC should be complemented by class-specific precision, recall, F1-score, and confusion matrices. The imbalance-mitigation strategy produced more uniform performance across the four classes and reduced bias toward the majority class.

Future studies should isolate the effects of data augmentation and class weighting through ablation experiments and compare them with undersampling, oversampling, focal loss, class-balanced loss, and cost-sensitive learning. Generative adversarial networks and diffusion models should also be assessed for minority-class augmentation.

External validation should include different OCT devices, institutions, acquisition protocols, and patient populations, with patient-level separation among data partitions. Further work should investigate probability calibration, uncertainty estimation, class-specific thresholds, and alternative convolutional, transformer-based, and hybrid architectures.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning," *Comput Methods Programs Biomed*, vol. 223, p. 106951, Aug. 2022.

[2] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *J Big Data*, vol. 6, no. 1, p. 113, Dec. 2019.

[3] M. H. Akpinar, A. Sengur, O. Faust, L. Tong, F. Molinari, and U. R. Acharya, "Artificial intelligence in retinal screening using OCT images: A review of the last decade (2013–2023)," *Comput Methods Programs Biomed*, vol. 254, p. 108253, Sep. 2024.

[4] M. Kulyabin *et al.*, "OCTDL: Optical coherence tomography dataset for image-based deep learning methods," *Sci Data*, vol. 11, no. 1, p. 365, Apr. 2024.

[5] Z. Ai *et al.*, "FN-OCT: Disease detection algorithm for retinal optical coherence tomography based on a fusion network," *Front Neuroinf*, vol. 16, Jun. 2022.

[6] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.

[7] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (OCT) and chest x-ray images for classification," Mendeley Data, Jan. 2018.

[8] P. Dutta, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, "Conv-ViT: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection," *J Imaging*, vol. 9, no. 7, p. 140, Jul. 2023.

[9] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018.

[10] C. J. Hellín, A. A. Olmedo, A. Valledor, J. Gómez, M. López-Benítez, and A. Tayebi, "Unraveling the impact of class imbalance on deep-learning models for medical image classification," *Appl Sci*, vol. 14, no. 8, p. 3419, Apr. 2024.

[11] M. Saini and S. Susan, "Tackling class imbalance in computer vision: a contemporary review," *Artif Intell Rev*, vol. 56, no. S1, pp. 1279–1335, Oct. 2023.

[12] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J Big Data*, vol. 6, no. 1, p. 27, Dec. 2019.

[13] A. Adel, M. M. Soliman, N. E. M. Khalifa, and K. Mostafa, "Automatic classification of retinal eye diseases from optical coherence tomography using transfer learning," in *2020 16th ICENCO*. Cairo, Egypt: IEEE, Dec. 2020, pp. 37–42.

[14] A. Rozhyna, M. Atzori, and H. Müller, "A comparative study of deep convolutional neural networks for the analysis of retinal damage in optical coherence tomography," in *Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications*, S. Wu and H. Yoshida, Eds. San Diego, United States: SPIE, Apr. 2024, p. 26.

[15] N. J. Bilal and R. Z. Yousif, "Neutrosophic-inspired data augmentation and multi-architecture ensemble model for robust retinal disease classification," *Eng Res Express*, vol. 7, no. 3, p. 0352b5, Sep. 2025.

[16] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," *Eur Radio*, vol. 34, no. 12, pp. 7895–7903, Jun. 2024.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE CVPR*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2818–2826.

[18] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int J Comput Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[19] A. Géron, *Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes*, 1st ed. Rio de Janeiro, Brasil: Alta Books, 2021.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans Knowl Data Eng*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[21] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med Imaging*, vol. 22, no. 1, p. 69, Dec. 2022.

[22] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J Mach Learn Res*, pp. 2825–2830, 2011.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv, 2014.

[24] G. Sharon, N. Yaligar, K. R. Kumar, A. Padaki, R. Dawe, and S. Katagi, "State-of-the-art deep learning strategies for multi-class classification of retinal OCT images," in *2024 Second ICAIT*. Chikkamagaluru, Karnataka, India: IEEE, Jul. 2024, pp. 1–7.

Classification of EEG signals for differential diagnosis of Alzheimer's disease using machine learning.

Davyd Wendys Florêncio da Silva
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
davyd.florencio@ufpe.br

Nathan Cahuê Aventino Fidelis
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
nathan.fidelis@ufpe.br

Gabriel Vitor Matias Pereira
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
gabriel.vpereira@ufpe.br

Beatriz Sales Pagliani Correia da Cruz
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
beatriz.pagliani@ufpe.br

Sandrelly Mirely da Silva Santos
Department of Pharmaceutical Sciences
Federal University of Pernambuco
Recife, Brazil
sandrelly.santos@ufpe.br

Wellington Pinheiro dos Santos
Department of Biomedical Engineering
Federal University of Pernambuco
Recife, Brazil
wellington.santos@ufpe.br

***Summary*Alzheimer's Disease (AD) is the leading cause of dementia in the elderly, presenting significant challenges for differential diagnosis, especially in relation to Frontotemporal Dementia (FTD). This study aims to evaluate the performance of machine learning algorithms for automatic classification of EEG signals, facilitating the differential diagnosis between AD, FTD, and healthy aging. A public database with records of 88 individuals (36 AD, 23 FTD, 29 controls) was used, subjected to balancing by SMOTE and feature selection via Genetic Algorithms and Particle Swarm Optimization. Eleven classified individuals were evaluated in the WEKA environment with validation (10 folds, 5 repetitions). The results demonstrated the superiority of the Random Forest with 200 trees, which achieved 96.2% accuracy and a Kappa index of 0.943 in the test set, with a sensitivity of 96.6% for AD and a specificity of 96.0% for controls. Analysis of the confusion matrix revealed low error rates between classes, including in the clinical challenge of distinguishing AD versus FTD. It is concluded that the proposed approach shows potential as an objective and non-invasive tool to aid in the differential diagnosis of dementias, contributing to earlier and more personalized interventions.**

***Keywords*Alzheimer's disease; Frontotemporal dementia; electroencephalography; machine learning; differential diagnosis**

I. Introductiondog

The increase in life expectancy observed in recent decades has been accompanied by an increase in the prevalence of neurodegenerative diseases, notably Alzheimer's Disease (AD), which currently represents the leading cause of dementia in the elderly [1]. According to the World Health Organization (WHO), approximately 10 million new cases of dementia are reported annually, totaling around 57 million people living with some form of this condition in 2021 [1]. Projections indicate that this number could reach 152 million by 2050, representing one of the greatest public health challenges of the 21st century [2].

Despite advances in knowledge about the pathophysiology of AD, early and accurate diagnosis remains a significant clinical challenge. Currently, the diagnostic process is predominantly based on subjective clinical assessments, combined with neuropsychological tests, such as the Mini-Mental State Examination (MMSE), and neuroimaging exams [3]. This approach, although widely used, presents important specificities, particularly in the early stages of the disease, such as in Mild Cognitive Impairment (MCI), where symptoms can be subtle and nonspecific [4].

One of the main obstacles is differential diagnosis, that is, distinguishing AD from other neurodegenerative dementias, especially Frontotemporal Dementia (FTD), as well as the process of healthy cognitive aging itself [5], [6]. The symptomatic overlap between these conditions frequently leads to diagnostic errors, compromising the prognosis and the effectiveness of therapeutic interventions [7]. Studies indicate that diagnostic error rates between AD and FTD can reach up to 30% in early stages, when assessed exclusively by clinical criteria [8]. This diagnostic imprecision delays the start of specific therapies and impacts patients' quality of life.

In this context, Electroencephalography (EEG) has emerged as a promising tool for investigating functional biomarkers in neurodegenerative diseases [9]. It is a non-invasive and low-cost technique that records brain electrical activity with high temporal resolution [10]. Studies demonstrate that AD is associated with characteristic changes in oscillatory patterns, notably a deceleration of the dominant rhythm, with increased power in the delta and theta frequency ranges, and a reduction in the alpha and beta ranges [11], [12]. Such changes preceding the more evident clinical manifestations and brain atrophy can be detected by neuroimaging, making EEG a potential source of early biomarkers [13].

In parallel, Artificial Intelligence (AI) and Machine Learning (ML) have become established as tools capable of extracting complex and non-linear patterns from biomedical signals, overcoming the limitations of traditional visual analysis [14]. Classification algorithms can analyze large volumes of EEG data, identifying electrophysiological signatures associated with different neurodegenerative conditions [15].

However, most of these studies focus on problems of classifying AD versus controls, not addressing the challenge of differential diagnosis in multiclass scenarios that include other

dementias [18]. Furthermore, the literature lacks work that performs a systematic comparison between different machine learning algorithms using the same experimental protocol, database and evaluation analyses, which makes it difficult to identify the most appropriate approach for clinical applications.

Given this scenario, the present work aimed to evaluate the performance of Machine Learning algorithms, implemented in the WEKA environment, for automatic classification of EEG signals focusing on the differential diagnosis of Alzheimer's Disease versus Frontotemporal Dementia and healthy aging, employing a methodology based on signal feature extraction, attribute selection through genetic algorithms and particle swarm optimization (PSO), and classification using 11 distinct algorithms.

II. BasisTe-mail

A. Alzheimer's disease and differential diagnosis

Alzheimer's disease (AD) is the leading cause of dementia in older adults, characterized by progressive cognitive decline associated with the accumulation of β-amyloid and hyperphosphorylated tau proteins in brain tissue [1], [2]. The differential diagnosis (the process of distinguishing one condition from others with similar symptoms) between AD and FTD constitutes a significant clinical challenge, since both present overlapping initial manifestations, such as changes in memory and behavior [3], [4].

Studies indicate that diagnostic error rates between AD and FTD can reach up to 30% in initial targets when based exclusively on clinical criteria [5]. This diagnostic imprecision compromises the prognosis, delays specific therapeutic interventions, and impacts the quality of life of patients and caregivers [6]. In this context, the identification of objective biomarkers that help distinguish between these conditions becomes fundamental for adequate clinical management and for the development of personalized therapeutic approaches.

B. Electroencephalography

EEG is a non-invasive technique that records brain electrical activity using electrodes placed on the skin of the ear. The captured signal reflects the sum of postsynaptic potentials generated by populations of pyramidal neurons in the cerebral cortex, offering excellent temporal resolution (milliseconds) for the analysis of neural dynamics [8]. The acquisition follows the International 10-20 System, which standardizes the placement of 19 electrodes on the metallic lobes, ensuring reproducibility and symmetrical coverage [9].

The EEG signal is classically composed of five main rhythms: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-25 Hz), and gamma (25-45 Hz), each associated with different functional states of the brain [10]. In the context of neurodegenerative diseases, AD is associated with characteristic alterations of these rhythms, notably a slowdown of the spectrum (increased power in the delta and theta bands, accompanied by a reduction in the alpha and beta bands), as well as loss of signal complexity and synchronization [11], [12]. In FTD, the alterations are less consistent, but studies report differences in connectivity patterns and spectral distribution in relation to AD and healthy controls [13]. Such alterations precede brain atrophy detectable by neuroimaging, potentially positioning EEG as a promising source of early biomarkers for the differential diagnosis of dementias [14].

C. Machine Learning for EEG Classification

Machine Learning comprises techniques that allow computational systems to identify patterns in data from examples, without explicit programming [15]. In the context of this work, supervised classification is employed, in which a model is trained with previously labeled data (e.g., DA, DFT, control) to predict the class of new samples [16].

Raw EEG signals have high dimensionality, making it necessary to remove features that compactly represent relevant information. These can be obtained in different domains: time (mean, variance, entropy); frequency (power spectral density by bands - delta, theta, alpha, beta, gamma); and time-frequency (wavelet transform) [17]. After separation, feature selection techniques such as Genetic Algorithms (GA), inspired by natural evolution, and Particle Swarm Optimization (PSO), based on collective behavior, can be applied to reduce dimensionality and eliminate redundancies [18].

In the scientific literature, different algorithms have been used for EEG classification, with this study highlighting: Bayesian Networks, Support Vector Machines (SVMs), and Decision Trees (J48, Random Tree, and Random Forest).

For performance evaluation in biomedical classification problems, different metrics can be used, such as: accuracy, which represents the overall proportion of correct answers; Kappa index, which measures chance-adjusted agreement; sensitivity (also called recall or true positive rate), which measures the proportion of correctly identified positive instances among those that affect the class of interest; specificity (true negative rate), which assesses the proportion of correctly identified negative instances among all those that do not belong to the class of interest; and the area under the ROC curve (AUC-ROC), which provides an overall measure of the model's discriminative ability between classes, regardless of the decision limits adopted.

III. Methodology

A. Database

The database used in this study was published by Miltiadous [19] and is publicly available in the OpenNeuro repository. The dataset contains secure, closed-eye EEG recordings from 88 individuals distributed into 36 patients with Alzheimer's Disease (AD), 23 patients with Frontotemporal Dementia (FTD), and 29 health controls (HC). Signals were acquired with 19 scalp electrodes (international 10-20 system), a sampling frequency of 500 Hz, and a resolution of 10 μV/mm. The cognitive status of the participants was assessed using the Mini-Mental State Examination (MMSE), with mean scores of 17.75 (SD = 4.5) for the AD group, 22.17 (SD = 8.22) for the FTD group, and 30 for the HC group. All records are in the standardized BIDS (Brain Imaging Data Structure) format, facilitating reproducibility and processing by computational tools.

B. Feature preprocessing and selection

Initially, the original database was partitioned into a training set (75%) and a test set (25%), maintaining the proportion between the classes. The training set was subjected to the SMOTE (Synthetic Minority Over-sampling Technique), an innovative technique in the WEKA environment, to equalize the number of instances among the three classes (DA, DFT, and CN), avoiding bias in the classifieds resulting from the original imbalance.

From the balanced training set, two feature selection techniques were applied: (i) Genetic Algorithms (GA), referenced as EVO base; and (ii) Particle Swarm Optimization (PSO), referenced as PSO base. Both techniques were employed to identify subsets of features most relevant for class discrimination, reducing dimensionality and eliminating redundancies.

At the end of this process, three training datasets were generated for the experiments: (1) balanced dataset without feature selection (BSA); (2) balanced dataset with selection by Genetic Algorithms (EVO); and (3) balanced dataset with selection by Particle Swarm Optimization (PSO).

C. Classification and experimental setup

The training set (75%), after balancing via SMOTE and, when applicable, subjected to feature selection techniques (EVO and PSO), was used for the classification stage in the WEKA software environment. Eleven classification algorithms were evaluated: Bayes Net (BN), Support Vector Machines (SVM) with linear kernels, polynomial of degree 2 and polynomial of degree 3 (parameter C = 0.01), Random Tree (RT), J48 and Random Forest (RF) with 100, 200, 300, 400 and 500 trees.

For each of the three training sets (balanced without selection, balanced with EVO selection, and balanced with PSO selection), experiments were conducted in the WEKA Experimenter module, employing 10-fold, 5-repetition cross-validation to estimate model performance. This protocol allows for the evaluation of classifier stability and generalization against data variability, using only the training set for parameter testing and model selection, without any contamination from the test set, which is occasionally separated for the final validation stage.

D. Selection of the best model and validation on the test set.

Following the cross-validation phase, in which all 11 classified models were evaluated on the three training datasets (balanced without selection, balanced with EVO selection, and balanced with PSO selection), the best combination was identified. The selection considered only the complete set of metrics (accuracy, Kappa index, sensitivity, specificity, and ROC area), with special emphasis on the Kappa index. In addition, practical aspects such as the computational cost associated with each model, particularly relevant for clinical applications, were considered.

The means and standard deviations of the observations in the cross-validation were calculated using Jamovi software, allowing for comparison between the models and the construction of box plot graphs to visualize the distribution of the results.

The classifier that presented the best balance between superior performance (especially in terms of Kappa), low variability between repetitions, and computational predictions was selected for the validation stage, performed on the test set (25%). From this validation, the confusion matrix and final performance statistics were generated, representing an estimate of the model's behavior on data not used in the training.

IV. π and Discussion

The experiments conducted on the training sets of the three bases (balanced without selection, balanced with EVO selection, and balanced with PSO Optimization selection) revealed significant performance differences among the evaluated subjects. Tables I, II, and III present the means and standard deviations of the metrics obtained in the validation with 10 folds and 5 repetitions for each base.

The results demonstrate that Random Forest-based models significantly outperformed all other evaluated algorithms in terms of accuracy, Kappa index, and statistical robustness, corroborating recent findings in the literature on EEG classification in dementia. In the three databases tested, Random Forest configurations achieved average accuracies exceeding 94%, with Kappa indices between 0.916 and 0.956, values that indicate near-perfect agreement between predictions and actual labels.

TABLE I - Mean and Standard Deviation of the results of the training set for the balanced base without selection

freshness	Precision	Kappa Index	Sensitivity	Specificity	ROC Area
BN	62.90 ± 1.010	0.444 ± 0.015	0.760 ± 0.016	0.716 ± 0.013	0.805 ± 0.010
J48	82.90 ± 0.787	0.744 ± 0.012	0.821 ± 0.014	0.906 ± 0.009	0.871 ± 0.009
RF 100	96.50 ± 0.396	0.948 ± 0.006	0.972 ± 0.006	0.972 ± 0.005	0.997 ± 0.001
RF 200	96.90 ± 0.379	0.953 ± 0.006	0.976 ± 0.005	0.975 ± 0.005	0.997 ± 0.001
RF 300	97.00 ± 0.406	0.954 ± 0.006	0.976 ± 0.005	0.976 ± 0.005	0.998 ± 0.001
RF 400	97.00 ± 0.391	0.955 ± 0.006	0.976 ± 0.005	0.976 ± 0.004	0.998 ± 0.001
RF 500	97.00 ± 0.390	0.956 ± 0.006	0.977 ± 0.005	0.976 ± 0.005	0.998 ± 0.001
RT	79.40 ± 1.040	0.692 ± 0.016	0.775 ± 0.015	0.889 ± 0.010	0.832 ± 0.010
SVM 1	40.20 ± 1.130	0.103 ± 0.017	0.093 ± 0.013	0.958 ± 0.007	0.545 ± 0.011
SVM 2	60.20 ± 1.100	0.403 ± 0.017	0.512 ± 0.024	0.849 ± 0.011	0.695 ± 0.012
SVM 3	68.00 ± 0.883	0.521 ± 0.013	0.688 ± 0.017	0.833 ± 0.011	0.772 ± 0.011

BN: Bayesian Network; RF: Random Forest; TR: Random Tree; SVM: Support Vector Machine.

TABLE II - Mean and Standard Deviation of the training set results for the balanced base with the PSO applied.

freshness	Precision	Kappa Index	Sensitivity	Specificity	ROC Area
BN	59.00 ± 1.060	0.385 ± 0.016	0.783 ± 0.016	0.634 ± 0.016	0.774 ± 0.012
J48	78.60 ± 0.952	0.679 ± 0.014	0.779 ± 0.021	0.887 ± 0.011	0.844 ± 0.012
RF 100	94.40 ± 0.553	0.916 ± 0.008	0.949 ± 0.009	0.961 ± 0.005	0.992 ± 0.001
RF 200	94.70 ± 0.458	0.920 ± 0.007	0.951 ± 0.009	0.963 ± 0.005	0.993 ± 0.001
RF 300	94.80 ± 0.504	0.921 ± 0.008	0.951 ± 0.009	0.964 ± 0.005	0.993 ± 0.001
RF 400	94.90 ± 0.486	0.923 ± 0.007	0.952 ± 0.009	0.965 ± 0.005	0.993 ± 0.001
RF 500	94.90 ± 0.483	0.924 ± 0.007	0.952 ± 0.009	0.965 ± 0.005	0.993 ± 0.001
RT	77.90 ± 1.190	0.668 ± 0.018	0.760 ± 0.018	0.880 ± 0.011	0.820 ± 0.012
SVM 1	34.60 ± 0.532	0.020 ± 0.008	0.037 ± 0.010	0.976 ± 0.006	0.509 ± 0.006
SVM 2	43.10 ± 1.810	0.146 ± 0.02 7	0.141 ± 0.045	0.959 ± 0.018	0.589 ± 0.019
SVM 3	54.20 ± 1.240	0.314 ± 0.019	0.421 ± 0.028	0.851 ± 0.015	0.639 ± 0.014

BN: Bayesian Network; RF: Random Forest; TR: Random Tree; SVM: Support Vector Machine.

TABLE III - Mean and Standard Deviation of the training set results for the balanced base with the PSO applied

freshness	Precision	Kappa Index	Sensitivity	Specificity	ROC Area
BN	61.50 ± 1.150	0.423 ± 0.017	0.789 ± 0.018	0.677 ± 0.016	0.802 ± 0.012
J48	80.80 ± 1.020	0.711 ± 0.015	0.797 ± 0.018	0.895 ± 0.010	0.853 ± 0.015
RF 100	95.60 ± 0.457	0.934 ± 0.007	0.964 ± 0.008	0.967 ± 0.005	0.995 ± 0.001

RF 200	96.00 ± 0.414	0.940 ± 0.006	0.966 ± 0.006	0.970 ± 0.005	0.996 ± 0.001
RF 300	96.10 ± 0.453	0.942 ± 0.007	0.967 ± 0.007	0.971 ± 0.004	0.996 ± 0.001
RF 400	96.20 ± 0.444	0.943 ± 0.007	0.968 ± 0.007	0.971 ± 0.005	0.996 ± 0.001
RF 500	96.20 ± 0.440	0.943 ± 0.007	0.969 ± 0.007	0.971 ± 0.005	0.996 ± 0.001
RT	78.10 ± 1.060	0.672 ± 0.016	0.767 ± 0.021	0.882 ± 0.011	0.825 ± 0.013
SVM 1	36.50 ± 1.720	0.047 ± 0.026	0.046 ± 0.009	0.976 ± 0.004	0.522 ± 0.005
SVM 2	55.20 ± 1.080	0.328 ± 0.016	0.475 ± 0.026	0.840 ± 0.014	0.671 ± 0.013
SVM 3	61.40 ± 0.945	0.422 ± 0.014	0.630 ± 0.020	0.814 ± 0.013	0.739 ± 0.012

BN: Bayesian Network; RF: Random Forest; TR: Random Tree; SVM: Support Vector Machine.

A. Abbreviations and acronyms

Define abbreviations and acronyms the first time they are used in the text, even if they have already been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, AC, DC, and RMS do not need to be defined. Do not use abbreviations in the title or headings unless they are unavoidable.

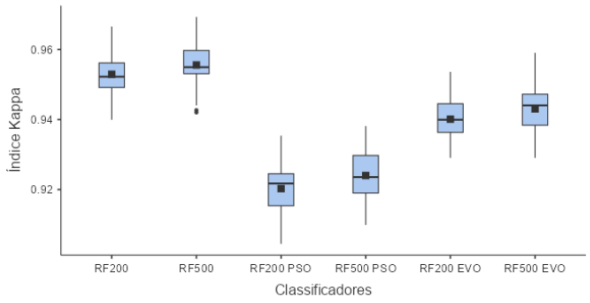
The classified SVM models performed worse, with accuracies ranging from 34.6% to 68.0% depending on the kernel and dataset used. The linear kernel (SVM 1) obtained the worst results, with an accuracy close to 40% in datasets without selection and EVO, and only 34.6% in the PSO dataset. Polynomial kernels of degree 2 and 3 improved progressively, but still lagged behind the tree-based models. The SVM 1 model in the PSO dataset showed even worse results, with a Kappa index of only 0.020, indicating a highly random performance.

The classifications based on single trees (J48 and RT) achieved intermediate results, with accuracies between 77.9% and 82.9%. J48 proved to be superior to RT in all datasets, reaching 82.9% accuracy in the dataset without selection. However, both presented higher standard deviations than the models.*Random Forest*, decreasing stability in validation repetitions. The BN's performance was limited, with accuracies between 59.0% and 62.9%, corroborating the literature that points to limitations of Bayesian models for EEG signal classification due to the violation of the conditional suspension of independence between features.

Among the variations of Random Forest, a consistent pattern is observed: increasing the number of trees produces marginal performance gains, but with diminishing returns. The model with 100 trees already shows high accuracies (94.4% to 96.5%), while the configuration with 500 trees achieves the highest absolute values in all datasets. However, the maximum difference between RF 200 and RF 500 is only 0.3% in accuracy and 0.003 in the Kappa index (PSO baseline), clinically insignificant values when considering the inherent variability of the problem.

Additionally, the*Random Forest*With 200 trees, it stands out for having the lowest standard deviations among all configurations evaluated (0.38 in the no-selection base, 0.46 in the PSO base, and 0.41 in the EVO base), providing greater consistency and reproducibility of results. This aspect is particularly relevant for clinical applications, where the stability of the model in the face of small variations in the input data is as important as average performance.

FIGURE I - BOX PLOTS OF THE KAPPA INDICES OF THE RANDOM FOREST CLASSIFIERS



RF: Random Forest; PSO: Particle Swarm Optimization; EVO: Genetic Algorithms.

Computational cost was also considered in the analysis. Although not systematically measured, we observed that the execution time of RF 500 was significantly higher than that of RF 200, especially in datasets with feature selection, where the search space complexity is greater. Recent studies report that the marginal performance gain provided by forests with more than 200 trees justifies the exponential increase in computational cost, particularly in scenarios with time or resource constraints.

Regarding the impact of feature selection techniques, it is observed that the base without selection produced the best results for all configurations.*Random Forest*, with accuracies between 96.5% and 97.0%. The databases with PSO and EVO selection obtained superior performance, with an average reduction of 1.8% and 0.8% in accuracy, respectively. This finding suggests that, for the set of features used (relative power of the five frequency bands), feature selection may have eliminated information relevant to the discrimination between classes, possibly due to the complementary nature of the different spectral bands in the differential diagnosis of dementias. Alternatively, the parameters used in the selection algorithms may not have been optimized for the specific task, opening perspectives for future investigations.

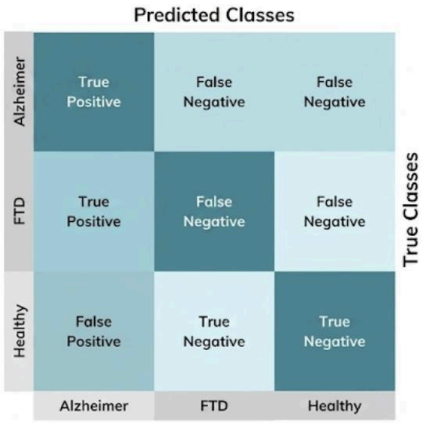
In light of these findings, the*Random Forest*with 200 trees applied to the balanced dataset without feature selection was selected as the final model for evaluation in the test set. This choice is based on the best relationship between: (1) high diagnostic performance (accuracy 96.9%, Kappa 0.953); (2) statistical robustness, i.e., lower standard deviation between repetitions; (3) balance between sensitivity (0.976) and specificity (0.975); and (4) computational predictions for clinical applications.

The results obtained after validating the Random Forest with 200 trees in the test set are presented in Table IV and in the confusion matrix shown in Figure II.

TABLE IV - RESULTS OF THE RANDOM FOREST WITH 200 TREES IN THE TEST SET

Metric	Result
Precision	96.20%
Kappa Index	0.943
Sensitivity	0.962
Specificity	0.971
ROC Area	0.997

FIGURE II - CONFUSION MATRIX OF THE RANDOM FOREST RESULTS WITH 200 TREES IN THE TEST SET.



DFT:Frontotemporal dementia.

The results in the test set reinforce the generalizability of the selected model, with metrics consistent with those observed in the validation. The accuracy of 96.20% and the Kappa index of 0.943 do not indicate overfitting.*overadjustment*g) during training, demonstrating that the model is able to maintain its performance in the presence of non-visible data.

Analysis of the confusion matrix reveals clinically important patterns. For the Alzheimer's disease class, out of 5,496 real instances, 5,308 (96.6%) were correctly identified, with errors distributed among 96 classifications as FTD (1.7%) and 92 as healthy controls (1.7%). The high sensitivity for AD is particularly relevant as it minimizes false negatives, i.e., Alzheimer's patients who would otherwise fail to receive appropriate treatment.

In the Frontotemporal Dementia class, out of 5,496 actual instances, 5,280 (96.1%) were correctly defined, with 147 errors observed as AD (2.7%) and 69 as control (1.3%). Confusion between AD and FTD was expected, as it reflects the symptomatic and electrophysiological overlap between these dementias, constituting the main clinical challenge of differential diagnosis [5], [6].

For the safety checklist class, out of 5,496 actual instances, 5,274 (96.0%) were correctly identified, with errors including 167 classifications as AD (3.0%) and 55 as FTD (1.0%). The 96.0% specificity indicates low false positive rates, meaning that particularly high levels of individuals are incorrectly diagnosed with some form of dementia, a crucial aspect for avoiding incidental interventions and negative psychological impact.

Class-specific statistics reinforce the model's consistency: AUC-ROC values close to 1.0 (0.996 to 0.997) indicate excellent discriminatory capacity between the three classes, while Matthews demonstration coefficients (MCC) between 0.932 and 0.950 confirm the quality of the predictions even considering the original data imbalance.

Compared to recent studies in the literature, the performance achieved is among the highest for multiclass EEG classification in dementia. Miltíado*and others.*[19], using the same database with a leave-one-subject-out approach, reported accuracies of 77.01% for AD vs CN classification with Random Forest. The superior performance in this work (96.2% in a multiclass scenario) can be attributed to the SMOTE balancing strategy, validation with repetitions, and optimization of the number of trees.

V. CONCLUSION

This work investigated a machine learning algorithm application for multiclass classification of EEG signals, focusing on the differential diagnosis of Alzheimer's Disease versus Frontotemporal Dementia and healthy aging. The results showed that Random Forest classifiers performed best, and for the validation stage, the Random Forest with 200 trees was selected, considering the results of evaluations and clinical applications, and observing computational cost. This model, applied to a balanced dataset without attribute selection, achieved superior performance to the other classifiers, with an accuracy of 96.2% and a Kappa index of 0.943 in the test set. The high sensitivity for Alzheimer's (96.6%) and specificity for height controls (96.0%) reinforce the robustness of the model for clinical applications.

The methodology employed, including SMOTE balancing, cross-validation with repetitions, and systematic comparison of 11 classifiers in the WEKA environment, proved adequate for the proposed problem, allowing the identification of the configuration with the best balance between diagnostic performance and computational predictions. The results obtained surpassed recent studies in the literature that used the same database, highlighting the value of the adopted approach.

From a clinical standpoint, the developed model offers an objective and non-invasive tool to aid in the differential diagnosis of dementias, contributing to the reduction of subjectivity inherent in isolated clinical evaluation. The non-invasive nature and low cost of EEG, combined with improved performance, position the approach as promising for applications in primary screening and longitudinal monitoring, especially in resource-limited settings.

Limitations of the study include the lack of external validation with independent databases and the need to investigate the importance of spectral characteristics for each diagnostic class. Future studies could explore the combination of different data modalities (EEG, neuroimaging, blood biomarkers), as well as evaluate the generalization of the models to other clinical situations and contexts.

In summary, Artificial Intelligence applied the demonstrated EEG signals with the potential to transform the differential diagnosis of dementia into a practice based on quantitative evidence, allowing for earlier and more personalized interventions, with a positive impact on the quality of life of patients and caregivers.

REFERENCES

[1] G. Eason, B. Noble and IN Sneddon, “On certain Lipschitz-Hankel type integrals involving products of Bessel functions”, Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68–73.

[3] IS Jacobs and CP Bean, “Fine particles, thin films, and exchange anisotropy,” in Magnetism, vol. III, GT Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Article title, if known”, unpublished.

[5] R. Nicole, “Article title with only the first word capitalized”, J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka and Y. Tagawa, “Electron Spectroscopy Studies in Magneto-Optical Media and Plastic Substrate Interface”, IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Abstracts of the 9th Annual Japan Magnetism Conference, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Niedermeyer, E.; da Silva, FL. Electroencephalography: Basic Principles, Clinical Applications, and Related Areas. 5th ed. Philadelphia: Lippincott Williams & Wilkins, 2005.

[9] Jeong, J. EEG dynamics in patients with Alzheimer's disease. Clinical Neurophysiology, v. 115, n. 7, p. 1490-1505, 2004.

[10] Rossini, PM et al. Non-invasive electrical and magnetic stimulation of the brain, spinal cord and nerve roots: basic principles and procedures for routine clinical application. Electroencephalography and Clinical Neurophysiology, v. 91, n. 2, p. 79-92, 1994.

[11] Dauwels, J.; Vialatte, F.; Cichocki, A. Diagnosis of Alzheimer's disease from EEG signals: where are we now? Current Alzheimer Research, v. 7, n. 6, p. 487-505, 2010.

[12] Zhu, Y., et al. (2022). Deep learning-based perivascular space segmentation: validation and clinical application. NeuroImage: Clinical, 35, 102999.

[13] Song Z, Deng B, Wang J, Yi G. An explainable systematic detection framework based on EEG to investigate and locate abnormal patterns in Alzheimer's disease. J Neural Eng. 2022 May 11;19(3). doi: 10.1088/1741-2552/ac697d. PMID: 35453136.

[14] Huang R. Diagnosis of Alzheimer's disease based on EEG analysis. Stud Health Technol Inform. 2023 Nov 23;308:574-582. doi: 10.3233/SHTI230887. PMID: 38007786.

[15] Halliday, DM, et al. (1995). Coherence analysis of EEG and EMG signals.

[16] Stam, CJ, Nolte, G., & Daffertshofer, A. (2007). Phase delay index: assessment of functional connectivity from multichannel EEG and MEG with reduced common-source bias.

[17] Nolte, G., et al. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherence.

[18] Nunez, PL, & Srinivasan, R. (2006). Brain electric fields: the neurophysics of EEG.

[19] A. Miltiadou, KD Tzimourta, T. Afrantou, P. Ioannidis, N. Grigoriadis, DG Tsalikakis, P. Angelidis, MG Tsipouras, E. Glavas, N. Giannakeas, et al., “A dataset of scalp EEG recordings from Alzheimer’s disease, frontotemporal dementia and healthy individuals from routine EEG”, Data, vol. 8, no. 6, p. 95, 2023, doi: 10.3390/data8060095.

Comparison of Two Hypoxic Burden Metrics in Polysomnography of Obstructive Sleep Apnea Patients

Silva-Simão, Thaís Cristine
Programa de Engenharia Biomédica
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
thais.vj@peb.ufrj.br

Duarte, Ricardo Luiz Menezes
Programa de Pós Graduação em Clínica Médica da Faculdade de Medicina
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
rlmduarte@gmail.com

Motta Ribeiro, Gabriel Casulari da
Programa de Engenharia Biomédica
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
gabrielcasulari@peb.ufrj.br

Abstract*—Introduction: Hypoxic burden metrics are emerging as alternatives to the Apnea-Hypopnea Index (AHI) in obstructive sleep apnea (OSA). Among those, Respiratory Event Desaturation Transient Area (REDTA) and Desaturation Severity 100 (DesSev100) are area-based metrics designed to quantify cumulative nocturnal oxygen desaturation. However, differences in their methodological definitions may influence the resulting estimates of hypoxic burden. Objective: To investigate whether REDTA and DesSev100 provide equivalent estimates of hypoxic burden in patients with OSA. Methods: This retrospective methodological study analyzed 33 type III polysomnography recordings from adult patients. REDTA was implemented in Octave according to its original description. DesSev100 was calculated using the ABOSA software. An adjusted version of DesSev100 was also generated to equalize the normalization factors of both metrics. Numerical values were compared using paired t-tests, and agreement was assessed using Bland–Altman analysis, including mean bias, 95% limits of agreement, and proportional bias. Results: REDTA and DesSev100 showed a strong positive correlation (Pearson's $r = 0.857$, 95% CI: 0.728–0.927). Nevertheless, REDTA numeric values were significantly higher than DesSev100 (mean difference = 13.65, CI: 7.21–20.10; $t(32) = 4.315$, $p < 0.001$). Significant differences remained after the normalization adjustment of DesSev100 (mean difference = 5.30 %·h, CI: 3.35–7.26; $t(32) = 5.527$, $p < 0.001$). Bland–Altman analysis showed 95% limits of agreement ranging from -5.50 %·h to 16.10 %·h, with a proportional bias. Conclusions: REDTA and DesSev100 quantify related aspects of nocturnal hypoxic burden but showed limited agreement and cannot be considered numerically interchangeable. Persistent differences after normalization suggest that methodological factors, including event selection and area integration strategies, influence the resulting estimates. These findings highlight the importance of considering metric-specific characteristics when comparing studies and interpreting hypoxic burden measurements in OSA.

***Index Terms*—obstructive sleep apnea; hypoxic burden; oxygen desaturation; polysomnography.**

I. Introduction

The Apnea-Hypopnea Index (AHI) remains the diagnostic gold standard for Obstructive Sleep Apnea (OSA). However, it does not capture the cumulative physiological burden of hypoxemia. This has led to the development of index, time, and area-based indicators of the intensity of blood oxygen desaturation [1], [2].

Metrics such as Hypoxic Burden (HB), Desaturation Severity (DesSev), and the Respiratory Event Desaturation Transient Area (REDTA) have emerged as more robust predictors of cardiovascular disease (CVD) mortality compared to traditional clinical indices [3]–[5]. However, different algorithms and baseline definitions coexist without a clear consensus on the superiority of any specific metric for predicting concrete clinical outcomes [2].

This uncertainty highlights the need for comparative studies to evaluate the performance of these tools in real-world clinical settings. Limitations to conducting such studies include the lack of openly accessible software to calculate metrics and published descriptions that explain the calculations but do not aim for their implementation. Currently, one exception is the ABOSA (Automatic Blood Oxygen Saturation Analysis) software [6]. This implements the DesSev and, according to two recent reviews of the literature [1], [2], the REDTA.

The objective of this study is to compare the REDTA and the DesSev. We aim to show that, due to fundamental differences in methodology, such as event definition and sampling window, these metrics are not interchangeable and provide non-equivalent assessments of the hypoxic burden.

II. Methods

All data were collected from polysomnographic records in which patients were informed and consented to the use of their data in future scientific studies conducted at the Clementino Fraga Filho University Hospital (HUCFF), provided that the anonymity and confidentiality of the information were preserved.

A. Study Design and Sample Selection

This is a retrospective, methodological, and comparative study focused on hypoxic burden metrics in OSA. The sample

This work was partially founded by CAPES and Instituto de Ciência, Tecnologia e Inovação de Maricá (ICTIM), Brazil.

consisted of type III polysomnography recordings (ApneaLink Air, ResMed, US) obtained from adult patients evaluated at the HUCFF outpatient sleep clinic for a clinical diagnosis of OSA. A total of 35 records were randomly selected.

B. Desaturtion Metrics

The study compared the values of two emerging desaturation area-based metrics:

1) REDTA (Respiratory Event Desaturation Transient Area): designed using a fixed 100% baseline and a temporal window determined according to respiratory event duration, as

$$REDTA = \frac{\sum_{i=1}^{N} Area_i}{3600}. \quad (1)$$

Where the area corresponding to an individual respiratory event i is defined as (Fig. 1)

$$Area_i = \int_{t_{mid}}^{t_{end_window}} [100 - SpO_2(t)]\, dt. \quad (2)$$

In these equations:

- N: number of respiratory events in the recording;
- t_{mid}: beginning of the integration window, located midway through the respiratory event;
- t_{end_window}: end of the integration window, extending for twice the event duration after the end of the event. If the integration window of two consecutive events overlaps, t_{mid} of the second is moved to t_{end_window} of the first.
- 3600: scaling factor to express the final value in $\%\,\mathrm{h}$.

Following the original description of REDTA [5], saturation values considered artifacts ($<50\%$ or $>100\%$) were ignored for the area calculation. However, these samples were counted for the determination of the integration window.

2) DesSev100 (Desaturation Severity 100): implemented in the ABOSA [6], it accumulates the area between 100% and the saturation curve within each automatically identified desaturation event (Fig. 2),

$$DesSev = \frac{\sum_{i=1}^{n} DesArea_i}{TST_{second}}. \quad (3)$$

In this equation:

- n: number of identified desaturation events;
- $DesArea_i$: area of an individual desaturation event;
- TST_{second}: Total Sleep Time (TST) expressed in seconds.

ABOSA automatically identifies desaturations, independently of respiratory events. These are identified from user defined minimum depth and duration thresholds, along with other fixed rules regarding event distances and desaturation rate. A detailed description is found in [6]. In the current study, minimum depth was chosen as 4% to match the respiratory event criteria, and minimum duration was $3\,\mathrm{s}$. A desaturation expands only from a maximum to a minimum, with the posterior increase in saturation treated as a separate recovery event. The recovery identification uses only fixed parameters and does not need to be identified for every desaturation [6].

In addition to the upper and lower saturation value bounds used to classify artifacts, ABOSA includes an additional criterion based on values > 30 after squaring a high-pass filtered version of the saturation signal [6]. A safety margin of $1\,\mathrm{s}$ before and after each sample classified as an artifact is added. If the resulting artifact region is $<5\,\mathrm{s}$, it is replaced by linear interpolation. The presence of artifacts excludes the entire desaturation or recovery event from the analysis.

A comparison of (1) and (3) reveals that they do not yield directly equivalent values due to their different normalization factors and units. While REDTA uses a fixed scale factor of 3600 to represent cumulative desaturation, DesSev100 is normalized by TST in seconds, representing a time-weighted average severity. To ensure a consistent scale for the comparative analysis, a mathematical adjustment was performed by standardizing the denominator based on the TST in hours,

$$DesSev_{adj} = DesSev \cdot TST_{hours} = \frac{\sum_{i=1}^{n} DesArea_i}{3600}. \quad (4)$$

C. Data Acquisition and Signal Processing

For population characterization, sex, age, height, weight, body mass index (BMI), neck circumference (NC), abdominal circumference (AC), and Epworth Sleepiness Scale (ESS) scores were collected from all participants.

Respiratory events were manually scored by a single specialist. The scoring followed the recommendations of the American Association of Sleep Medicine, considering hypopneas with the 4% desaturation rule. Due to the absence of electroencephalographic data, arousals could not be considered for hypopnea, and there was no sleep staging. Furthermore, TST was considered equal to the total recording time during which the oxygen saturation signal was available, as default in ABOSA for this situation.

PSG and event data were extracted in the universal European Data Format (EDF) for the calculation of the desaturation metrics. REDTA was implemented by the authors in the Octave programming language, following the original description of the metric [5], summarized above. The DesSev100 was calculated directly in the ABOSA 1.2.2 [7]. The denominator-modified DesSev100 was calculated using the output tables of ABOSA. The AHI was computed from all the events identified in the EDF file, divided by the total time between the marks for the start and end of the scoring window.

D. Statistical Analysis

Descriptive statistics were calculated and expressed as mean $\pm$ standard deviation. Differences in numerical values between the REDTA and the DesSev100 metric provided by the ABOSA software or the adjusted version were assessed using the paired t-test. The association between the metrics was assessed using Pearson's correlation coefficient. Agreement between REDTA and adjusted DesSev100 was assessed using Bland–Altman analysis, including mean bias, 95% limits of agreement, and proportional bias. This last was also assessed

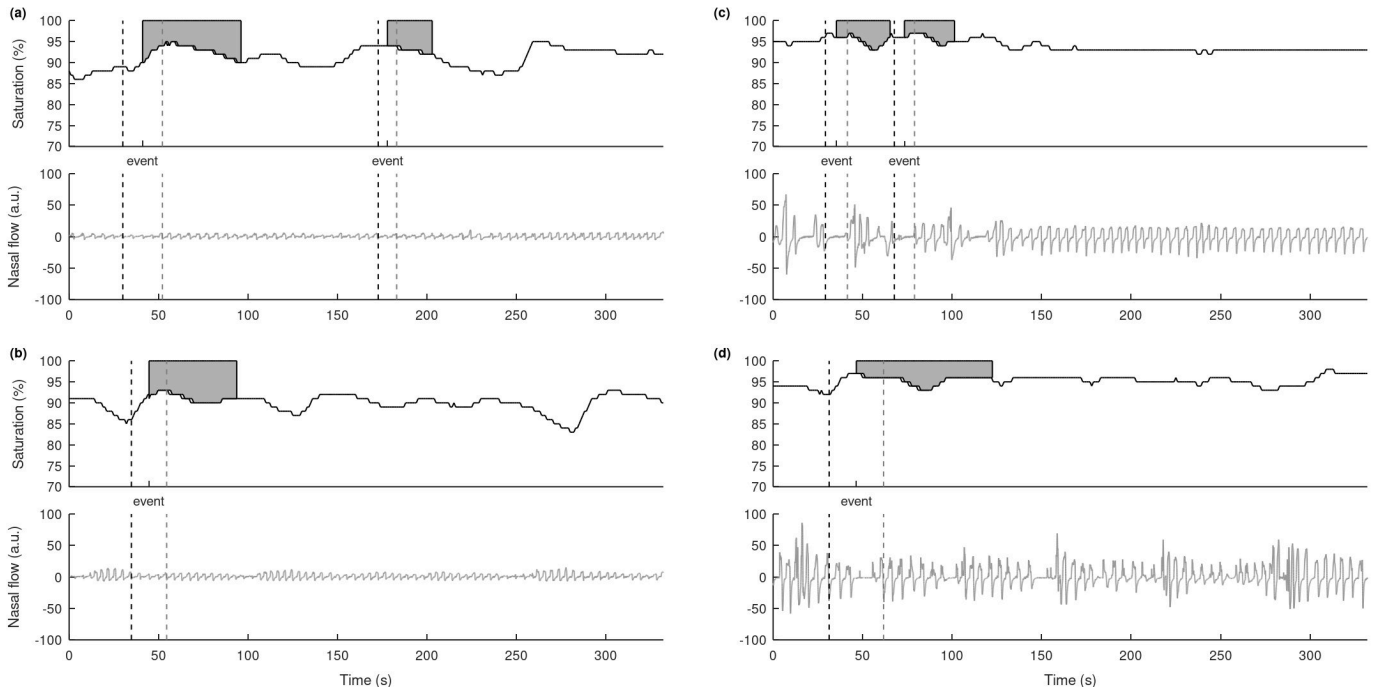


Fig. 1. Examples of saturation and nasal flow curves for two patients (panels a/c, and b/d) from the analyzed population showing the area used in the Respiratory Event Desaturation Transient Area (REDTA) metric. The manually scored respiratory events are marked with vertical dashed lines, from the black to the light gray. The area considered as desaturation related to one event (gray shade) extends from the middle of the event to twice the event duration after its end. The upper limit is 100%, and the lower the saturation curve. The metric considers only the occurrence of a respiratory event, integrating the described region even if it does not cover the entire desaturaion (a), corresponds to saturation drops lower than 3% considered in the guidelines (b and c), or extends after a recovery from desaturation (d). Only events that start in the graphs' time interval are shown.

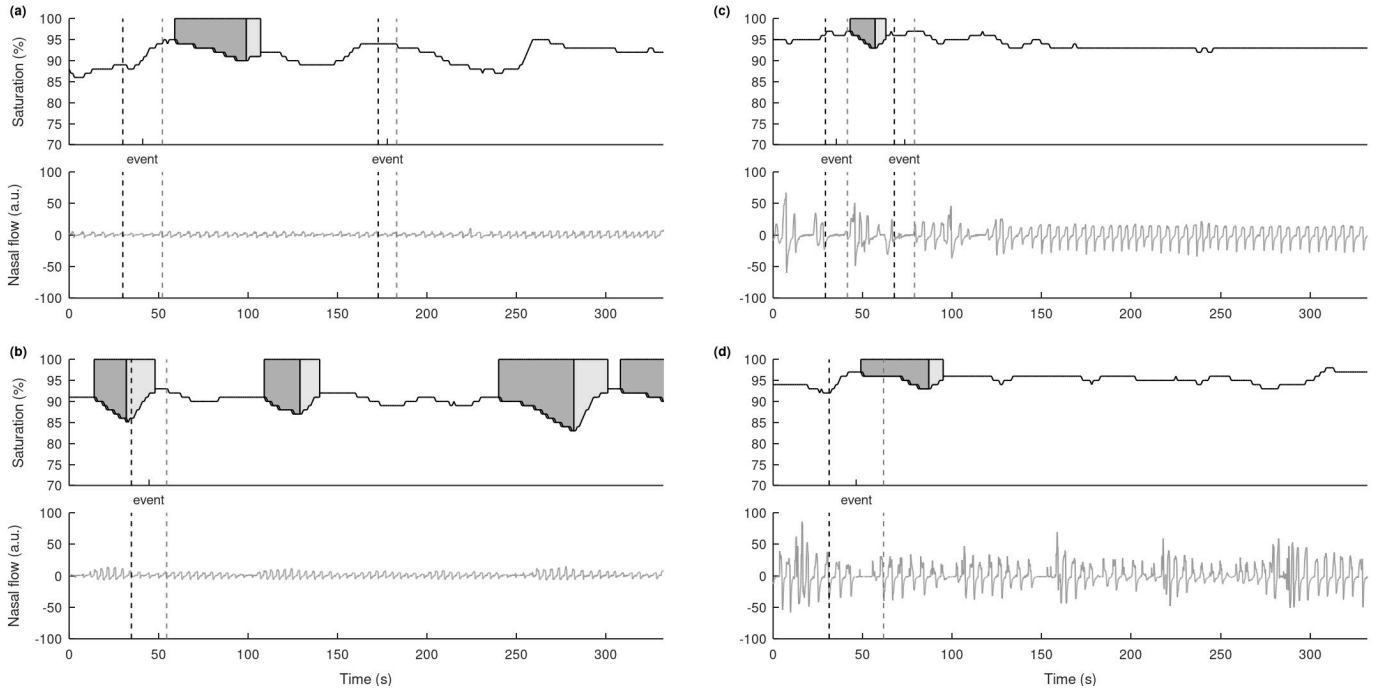


Fig. 2. Examples of saturation and nasal flow curves for two patients (panels a/c, and b/d) from the analyzed population showing the area used in the Desaturation Severity 100 (DesSev100) metric. The manually scored respiratory events are marked with vertical dashed lines, from the black to the light gray. The area considered in the metric (dark gray shaded) is the one between 100% and the saturation curve within the time interval automatically identified as corresponding to a desaturation of, at least, 4%. The recovery of saturation is also automatically identified, with the corresponding area shown (light gray shaded). The desaturation identification follows a series of rules and does not considers the information about respiratory event. Thus, it may not include some saturation drops of 4% (a) or changes in saturation after respiratory events that are lower than the chosen threshold (b); while including desaturations not time related to events (c). Only events that start in the graphs' time interval are shown.

TABLE I
PARTICIPANT CHARACTERISTICS (N = 33)

Female, n (%)	21 (63.6)
Age (years)	52.5 ± 15.6
Body Mass Index ($\mathrm{kg\,m^{-2}}$)	33.5 ± 5.9
Height (m)	1.66 ± 0.09
Weight (kg)	92.0 ± 21.3
Epworth Sleepiness Scale	12.2 ± 5.1
Neck circumference (cm)	39.9 ± 5.9
Abdominal circumference (cm)	107.7 ± 18.0

Data is mean ± SD

using reduced major axis linear regression [8]. For all estimates, 95% confidence intervals (CI) were calculated. All statistical analyzes were performed using JASP software, with a significance level of 5%.

III. RESULTS

During the analysis, two patients were excluded due to the absence of the saturation signal. The final 33 sample was predominantly female (63.6%), aged 52.5 ± 15.6 years, and had a BMI of $33.5 \pm 5.9\,\mathrm{kg\,m^{-2}}$ (Table I).

The DesSev100 calculated in ABOSA cannot be considered the same metric as REDTA. A positive correlation was observed between REDTA and DesSev100 (Pearson's $r = 0.857$, 95% CI: 0.728–0.927), indicating that both metrics quantify related aspects of nocturnal hypoxic burden. However, as expected from their different normalization factors, significant differences were observed between the numerical values of REDTA ($15.55 \pm 19.94\,\%\,\mathrm{h}$) and DesSev100 ($1.89 \pm 2.10\%$), with a mean difference of $13.65\,\%\,\mathrm{h}$ (95% CI: 7.21–20.10; $t(32) = 4.315$, $p < 0.001$).

The discrepancy between the methods was not explained only by the difference in normalization. Adjusting DesSev100 to use the same normalization factor as REDTA increased its values ($10.25 \pm 15.02\,\%\,\mathrm{h}$), but a significant difference between the metrics persisted (mean difference = $5.30\,\%\,\mathrm{h}$, 95% CI: 3.35–7.26; $t(32) = 5.527$, $p < 0.001$). Bland–Altman analysis showed 95% limits of agreement ranging from $-5.50\,\%\,\mathrm{h}$ to $16.10\,\%\,\mathrm{h}$ and indicated a proportional bias (Fig. 3a). This bias was confirmed by a slope of 0.753 (CI: 0.715–0.793) significantly different from unity in a reduced major axes regression between adjusted DesSev100 and REDTA. Additionally, the difference between the metrics had a positive correlation with whole night average oxygen saturation (Fig. 3b - $r = 0.450$, $p = 0.009$). Thus, the difference between the metrics is larger for more severe patients and those with lower overall saturation.

Both hypoxic burden metrics were significantly associated with AHI (Fig. 4). REDTA, which is based on the respiratory events, had a larger positive correlation with AHI ($r = 0.799$, 95%: 0.628–0.897, $p < 0.001$) than DesSev100 ($r = 0.695$; 95%: 0.461–0.838; $p < 0.001$). Still, both metrics exhibited variability around the regression lines, especially for larger AHI, reinforcing that they incorporate information beyond the frequency of respiratory events.

IV. DISCUSSION

The present study compared two area-based hypoxic burden metrics, REDTA and DesSev100, using the same set of type III polysomnography recordings with manually scored respiratory events. The main finding was that, although both metrics showed a strong correlation, they were not numerically equivalent. Significant differences persisted even after adjusting the denominator of DesSev100 to match the normalization adopted by REDTA. Thus, the observed differences between the methods are related to their methodological definitions rather than to mathematical scaling alone.

The DesSev100 calculated in ABOSA should not be considered the same metric as REDTA. Two reviews of methods to evaluate hypoxic burden indicate that the ABOSA software can be used to calculate REDTA [1], [2]. Even the study validating the ABOSA software seems to make such equivalence between the metrics, citing the REDTA study as previous findings using the DesSev100 [6]. The strong correlation observed between REDTA and DesSev100 ($r = 0.857$) indicates that both metrics quantify related characteristics linked to nocturnal hypoxic burden. They are based on the integration of oxygen desaturation areas using the same fixed reference value and increase with the number, duration, and depth of desaturation events. However, as the metrics are expressed in different units (REDTA in $\%\,\mathrm{h}$, DesSev100 in $\%$), our findings of significantly different numerical values were to be expected. This would lead to incorrect risk classification if one used any threshold derived from the true REDTA calculation in patients scored using DesSev100 as a REDTA-like metric. Of note, a recent study showed that DesSev100 and REDTA differ in their capacity as predictors of cardiovascular related mortality outcomes when calculated in the same patients [9].

Importantly, the difference between the metrics extends beyond their units. The different units exist because whereas REDTA uses a fixed scaling factor, DesSev100 is normalized by TST. After we adjusted DesSev100 to use the same denominator adopted by REDTA, significant differences remained. These differences may be explained by the distinct definitions of the desaturation areas. REDTA is linked to manually scored respiratory events and quantifies desaturation burden within predefined temporal windows based on those events. In contrast, DesSev100 is based on the automated identification of oxygen desaturation, regardless of direct temporal correspondence with respiratory events. As shown in the examples of Figs. 1 and 2, REDTA can include regions that are smaller or larger than a desaturation event of DesSev100. It also includes regions with drops in saturation that are lower than the 4% minimum configured in the ABOSA. Meanwhile, DesSev100 may include more desaturation events than the scored respiratory events, or the rules considered by ABOSA may not classify as desaturation intervals where saturation drops by 4%.

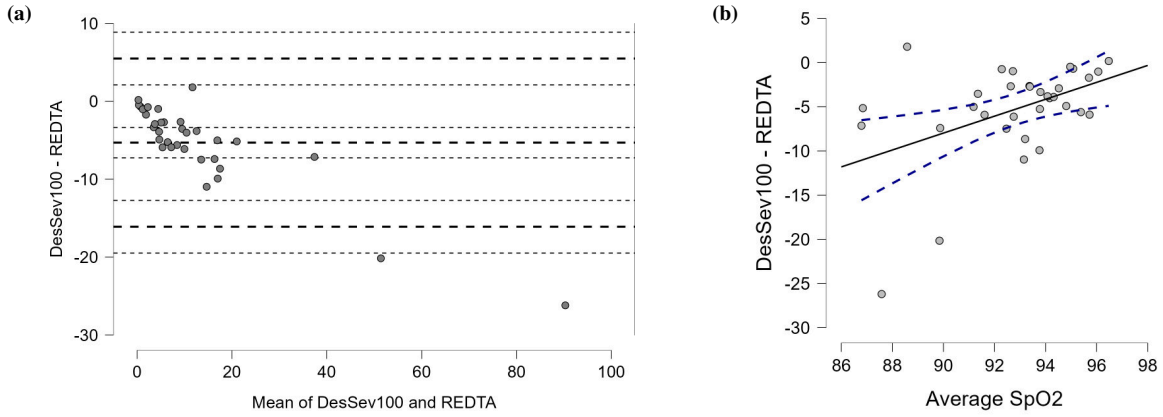


Fig. 3. Analyzes of the difference between REDTA and the adjusted DesSev100. (a) Bland–Altman plots of the metrics showing their mean bias and 95% limits of agreement (thick dashed lines), with the confidence intervals of each value (dashed lines). The asymmetry on the horizontal axis indicates a proportional bias. (b) Correlation of the difference between the metrics and the whole night average of saturation ($r = 0.450$, $p = 0.009$). The dots represent each of the 33 patients. The black line is a linear regression with blue-dashed lines confidence interval. REDTA = Respiratory Event Desaturation Transient Area; DesSev = Desaturation severity; SpO2 = peripheral oxygen saturation.

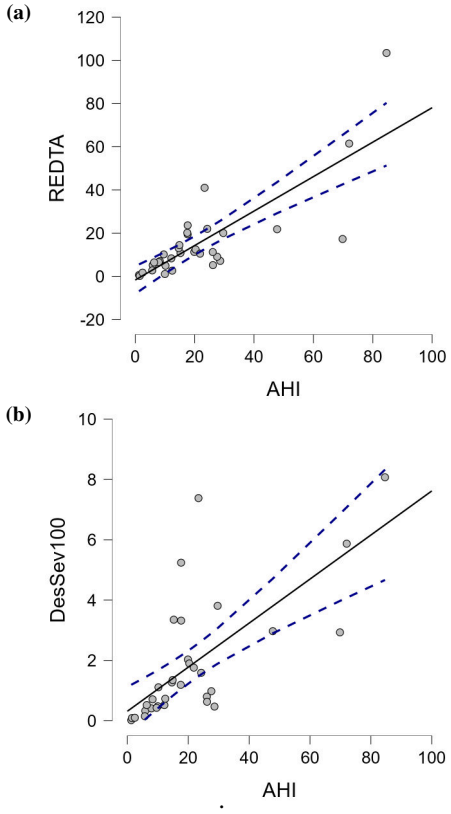


Fig. 4. Correlation of REDTA (4a - $r = 0.799$, 95%: 0.628–0.897, $p < 0.001$) and DesSev100 (4b - $r = 0.695$; 95%: 0.461–0.838; $p < 0.001$) with apnea-hypopnea index (AHI). The dots represent each of the 33 patients. The black line is a linear regression with blue-dashed line confidence interval. REDTA = Respiratory Event Desaturation Transient Area; DesSev = Desaturation severity.

An additional difference in area definitions is that ABOSA separates the decrease and increase in saturation into different events. Thus, if most desaturations are linked to a respiratory event, there should be a tendency for smaller areas for DesSev100. This is in line with our finding of significantly higher REDTA, with only two patients showing higher adjusted DesSev100. The ABOSA calculates a similar area-based metric for recoveries (RecSev100), which could be added to DesSev100. This addition could increase the number of differences with the opposite sign and decrease the mean difference between metrics. However, as our examples show, including recoveries would not be sufficient to make integrated regions equal. Moreover, given the fixed reference saturation value, it would be expected that a lower overall oxygen saturation leads to an increase in the differences caused by the area definitions. Accordingly, our results indicate a partial correlation between the metrics difference and the whole night saturation average ($r = 0.450$). Noteworthy, the Bland–Altman plot suggests a proportional bias between the metrics, which was confirmed by a regression between REDTA and adjusted DesSev100. Hence, considering both metrics as interchangeable could have a greater impact on patients who are in the highest risk.

Hypoxic burden metrics may reflect deeper physiological consequences of OSA than the number of events. Both REDTA and DesSev100 showed strong associations with AHI, indicating that hypoxic burden metrics remain related to the conventional index used for OSA diagnosis and severity classification. The correlation was strongest for REDTA, which depends directly on manually scored respiratory events used for AHI calculation. In contrast, DesSev100 is based on the automated identification of desaturations, not requiring direct correspondence with respiratory events. The variability around the regression line is, at least, partially explained by the area-based metrics incorporating information related to the depth and duration of oxygen desaturations. As a result, they may provide a more comprehensive assessment of nocturnal hypoxic exposure. This characteristic has been proposed as a possible explanation for the more consistent association between hypoxic burden metrics and cardiovascular outcomes reported in previous studies [3].

This study has some limitations. First, the sample size

was not based on a power calculation, and all patients were from a single center. The use of paired comparisons and the effect sizes observed indicate an adequate sample. A larger, independent, multicentric sample could provide greater variability in flow, oxygen saturation, and respiratory event patterns, allowing for a more robust assessment of the potential sources of discrepancy between the metrics. Such a sample could also improve the generalizability of the findings across different patient populations and recording conditions. However, it is also expected to highlight edge cases with even more extreme differences between the metrics. Second, only type III polysomnography recordings were available, precluding the use of electroencephalographic criteria and sleep-stage information. The definition of TST has a direct influence on the value of DesSev100, but sleep-staging could also influence the respiratory event scoring. Third, respiratory events were scored by a single evaluator. Due to inherent variability in visual data scoring, the exact numerical values are specific to this evaluator, but the observed differences are unlikely to depend only on event identification. Additionally, if the agreement between REDTA and DesSev100 is dependent on the evaluator, this reinforces that the metrics are not interchangeable in clinical practice. Finally, the implementation of REDTA required the development of a custom algorithm, since currently available software solutions do not allow direct processing of manually scored respiratory events according to the original definition of the metric. Nevertheless, this implementation ensured methodological fidelity to the original REDTA description and enabled a direct comparison with DesSev100 under controlled conditions.

V. Conclusions

REDTA and DesSev100 appear to quantify related dimensions of nocturnal hypoxic burden, as demonstrated by the correlation observed between them. However, the persistence of significant differences after mathematical normalization and the systematic nature of the disagreement between the methods indicate that these metrics should not be considered interchangeable. Methodological choices related to event selection and area integration substantially influence the final quantification of hypoxic burden and should be carefully considered when comparing studies and establishing clinical thresholds.

Acknowledgment

The generative artificial intelligence tool ChatGPT 5.5 (OpenAI) was used exclusively to assist with English language editing, as well as to correct grammar and spelling in the manuscript. The study design, methodology, statistical analyzes, interpretation of results, and all scientific decisions remained the sole responsibility of the authors, who assume full responsibility for the final content of the paper.

References

[1] S. He, P. A. Cistulli, and P. De Chazal, "A Review of Novel Oximetry Parameters for the Prediction of Cardiovascular Disease in Obstructive Sleep Apnoea," *Diagnostics*, vol. 13, no. 21, p. 3323, Oct. 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/21/3323

[2] A. Parekh, "Hypoxic burden – definitions, pathophysiological concepts, methods of evaluation, and clinical relevance," *Current Opinion in Pulmonary Medicine*, vol. 30, no. 6, pp. 600–606, Nov. 2024. [Online]. Available: https://journals.lww.com/10.1097/MCP.0000000000001122

[3] A. Azarbarzin, S. A. Sands, K. L. Stone, L. Taranto-Montemurro, L. Messineo, P. I. Terrill, S. Ancoli-Israel, K. Ensrud, S. Purcell, D. P. White, S. Redline, and A. Wellman, "The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study," *European Heart Journal*, vol. 40, no. 14, pp. 1149–1157, Apr. 2019. [Online]. Available: https://academic.oup.com/eurheartj/article/40/14/1149/5146754

[4] A. Kulkas, P. Tiihonen, K. Eskola, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters for evaluating severity of sleep disordered breathing and for supporting diagnosis of sleep apnea-hypopnea syndrome," *Journal of Medical Engineering & Technology*, vol. 37, no. 2, pp. 135–143, Feb. 2013. [Online]. Available: http://www.tandfonline.com/doi/full/10.3109/03091902.2012.754509

[5] P. De Chazal, N. Sadr, H. Dissanayake, K. Cook, K. Sutherland, Y. S. Bin, and P. A. Cistulli, "Predicting Cardiovascular Outcomes Using the Respiratory Event Desaturation Transient Area Derived from Overnight Sleep Studies," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Mexico: IEEE, Nov. 2021, pp. 5496–5499. [Online]. Available: https://ieeexplore.ieee.org/document/9630610/

[6] T. Karhu, T. Leppänen, J. Töyräs, A. Oksenberg, S. Myllymaa, and S. Nikkonen, "ABOSA – Freely available automatic blood oxygen saturation signal analysis software: Structure and validation," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107120, Nov. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260722005016

[7] ——, "Automatic oxygen saturation signal analysis software," Aug. 2025, iSSN: 01692607 Language: en. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.13982004

[8] J. Ludbrook, "Special article comparing methods of measurement," *Clinical and Experimental Pharmacology and Physiology*, vol. 24, no. 2, pp. 193–203, Feb. 1997. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1440-1681.1997.tb01807.x

[9] S. He, P. A. Cistulli, and P. De Chazal, "Comparison of oxygen desaturation area-based methods in predicting cardiovascular disease-related mortality outcomes," *Frontiers in Network Physiology*, vol. 6, p. 1805587, Apr. 2026. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnetp.2026.1805587/full

Compilação Semântica em Redes Bayesianas: Explorando LLMs para Inferência MPE sem Dados

1° Victor H. S. S. Ragazzi
Department of Electrical Engineering
Federal University of Viçosa
Viçosa, Brazil
victor.ragazzi@ufv.br

2° Jhonatan S. Oliveira
Department of Computer Science
University of Regina
Regina, Canada
desouzjh@uregina.ca

3° Cory Butz
Department of Computer Science
University of Regina
Regina, Canada
cory.butz@uregina.ca

4° Leonardo Bonato Felix
Department of Electrical Engineering
Federal University of Viçosa
Viçosa, Brazil
leobonato@ufv.br

***Abstract*—Redes Bayesianas (BN) são utilizadas em diversos domínios, como áreas médicas, tendo poder para apoiar diagnóstico e prognóstico. Entretanto, a construção de suas tabelas de probabilidade condicional (CPTs) não é simples, exigindo aquisição de dados ou uso de conhecimento especializado. O foco deste trabalho é investigar se modelos de linguagem de grande escala (LLMs) são capazes de aproximar, por meio de um procedimento sequencial de decisões locais, a solução do método de inferência *Most Probable Explanation* (MPE), utilizando apenas a semântica das variáveis, sem acesso às suas CPTs. O método foi avaliado em dez BN de domínios médicos distintos, comparando as configurações preditas pelo LLM com a solução MPE exata sob diferentes condições de evidência, incluindo evidências definidas tanto de acordo com quanto de forma independente da solução MPE. A acurácia por variável obtida variou de 60.7% a 100% entre as redes, mas essa variação não se mostrou explicada de forma estatisticamente significativa por características estruturais isoladas, como profundidade ou número de pais max., sugerindo que fatores semânticos específicos de cada domínio pesam mais do que a topologia da rede. Além disso, uma métrica de correspondência da configuração conjunta (*exact match*) revela que a acurácia por variável superestima substancialmente a capacidade do método de recuperar a solução MPE completa, e que essa capacidade se degrada de forma acentuada quando a evidência não é consistente com o MPE. Esses resultados indicam que LLMs podem oferecer uma aproximação de baixo custo para inferência MPE em BN ainda não parametrizadas, mas sujeita a limitações importantes, tanto teóricas — a inferência sequencial local não garante a solução conjunta ótima — quanto práticas — sensibilidade à consistência da evidência e a fatores semânticos ainda não plenamente caracterizados.**

***Index Terms*—Redes Bayesianas, Most Probable Explanation, Large Language Models, inferência, semântica**

I. Introdução

Redes Bayesianas (Bayesian Network - BN) constituem uma classe fundamental de modelos probabilísticos gráficos no campo do aprendizado de máquina e da inteligência artificial (Pearl, 1988; Koller and Friedman, 2009; Darwiche, 2009). Esses modelos combinam conhecimento especializado e dados observacionais para representar relações de dependência condicional entre variáveis aleatórias em um domínio específico. Formalmente, uma BN é definida por um grafo direcionado acíclico, no qual os nós representam variáveis e as arestas indicam relações de dependência probabilística, juntamente com um conjunto de tabelas de probabilidade condicional (Conditional Probability Tables — CPTs) que quantificam essas relações.

A estrutura do grafo informa as dependências condicionais que permitem uma representação da distribuição de probabilidade sobre o conjunto de variáveis (Koller and Friedman, 2009). Essa propriedade torna as BN adequadas para tarefas de inferência probabilística. Dada uma evidência (uma atribuição de valor a uma ou mais variáveis), é possível inferir a distribuição das variáveis restantes.

O processo de inferência possui diferentes objetivos, como a estimação de probabilidades marginais ou a identificação da explicação mais provável (Most Probable Explanation — MPE), que corresponde à configuração das variáveis não observadas que maximiza a probabilidade conjunta condicionada às evidências (Dechter, 1996). Diferentemente da inferência marginal, que envolve inferência de variáveis específicas, a inferência MPE é formulada como um problema de otimização global, no qual a ideia é atribuir os valores mais prováveis para todas as variáveis que não são tratadas como evidência.

Entretanto, para realizar as inferências, é necessário utilizar as probabilidades de cada nó, armazenadas em CPTs, cuja obtenção ainda representa um desafio. Em aplicações reais, sua construção frequentemente requer grandes volumes de dados para estimação estatística ou a elicitação de conhecimento especializado por parte de especialistas do domínio (Heckerman et al., 1995). Este processo pode ser custoso, suscetível a viés e, em muitos casos, limitado pela escassez de dados ou pela dificuldade de formalizar o conhecimento especializado.

Por outro lado, modelos de linguagem de grande escala (Large Language Models — LLMs) têm apresentado avanços em tarefas que envolvem raciocínio e inferência baseados em linguagem natural (Brown et al., 2020; Wei et al., 2023). Treinados em grandes conjuntos de dados textuais, esses mod-

elos capturam padrões estatísticos e semânticos complexos, permitindo a realização de inferências baseadas em conhecimento implícito através do raciocínio causal (Kıcıman et al., 2024). Embora não representem explicitamente distribuições probabilísticas no formato tradicional de modelos gráficos, evidências recentes sugerem que LLMs são capazes de aproximar relações causais e dependências entre variáveis de forma qualitativa (Bubeck et al., 2023).

Diante disso, é natural perguntar-se se LLMs conseguem captar padrões através da semântica para aproximar a configuração mais provável em uma rede. O trabalho propõe utilizar LLMs para aproximar inferências MPE utilizando apenas a semântica de variáveis em BN, sem a utilização das CPTs, visando simular uma situação onde a estrutura é construída mas não há valores atribuídos para cada variável. Dessa forma, busca-se entender até onde a linguagem pode ser usada como uma alternativa para aproximar inferências em redes que estão em construção e não possuem dados integrados.

II. Trabalhos relacionados

LLMs foram empregados amplamente em várias áreas, incluindo a medicina, desde tarefas tão simples como triagem e resposta a questões de conhecimento médico até fornecer suporte direto na formulação do diagnóstico diferencial (Singhal et al., 2023; Nori et al., 2023). Revisões recentes da literatura descrevem dezenas de trabalhos propondo modelos LLM para fazer prognósticos de doenças em áreas tão distintas quanto oncologia, doenças infecciosas, radiologia e emergência (Omar et al., 2024; Zhou et al., 2025). Avaliações em configurações reais, isto é, em serviços de urgência, demonstram que sistemas baseados em LLMs conseguem gerar propostas diagnósticas de qualidade ao nível de médicos especialistas (Niset et al., 2025). Contudo, essa mesma literatura ressalva que esses sistemas parecem por vezes mal calibrados, são suscetíveis a produzir ”saltos de raciocínio” de uma execução a outra e a reproduzir vieses demográficos que comprometem o seu uso de forma autônoma (Kesmen et al., 2026).

Ao mesmo tempo e de modo independente da evolução dos LLMs, as BN consolidaram-se como uma técnica de suporte à decisão clínica amplamente utilizada, tendo sido aplicadas em problemas de diagnóstico diferencial e prognóstico em áreas que vão desde as doenças cardiovasculares até aos distúrbios neurodegenerativos e o câncer (Polotskaya et al., 2024). Por exemplo, foram desenvolvidos sistemas especialistas baseados em BNs para a previsão de angina instável a partir de dados clínicos (Vila-Francés et al., 2013). Esta adoção deve-se à capacidade do formalismo para representar incerteza e relações causais entre variáveis clínicas, em um cenário clínico onde a interpretabilidade da decisão é tão importante quanto a sua precisão (Polotskaya et al., 2024). Contudo, essa mesma literatura admite que a construção destas redes, seja pela elicitação do conhecimento de especialistas ou mesmo pelo processo de aprendizagem a partir de dados, continua sendo uma tarefa trabalhosa e um dos maiores entraves à sua adoção (Polotskaya et al., 2024).

Kesmen et al. (2026) procuram combinar o uso de raciocínio em linguagem natural dos LLMs e do formalismo probabilístico de BN, ao invés de vê-los como abordagens contrárias. O framework MoBayes apresenta uma arquitetura que abstrai o raciocínio probabilístico do componente linguístico em sistemas de auxílio à decisão clínica conversacionais, preservando os cálculos de probabilidades e frequências formalizados pela BN enquanto empresta a interface em linguagem natural ao LLM (Kesmen et al., 2026). Em uma tarefa de diagnóstico de dor torácica, LLMs de uso geral tiveram o seu comportamento de inquirição de informação e raciocínio diagnóstico comparado com políticas ótimas extraídas de BN treinadas em dados clínicos reais, demonstrando discrepâncias substanciais na forma como cada abordagem escolhe as evidências (Safranek et al., 2026). Trabalhos prévios também propuseram frameworks bayesianos para orientar a coleta iterativa da história clínica com o auxílio de LLMs em condições médicas recorrentes, com o objetivo de tornar o processo diagnóstico mais explicável e reprodutível (Suraj, 2025).

Semelhante à ideia deste estudo, uma corrente recente explora a utilização de LLMs diretamente como substitutos para a elicitação especializada de CPTs. Nafar et al. (2025) investigam a capacidade de estimativa probabilística de vários LLMs, aproveitando o conhecimento armazenado nestes modelos para parametrizar BN específicas, com experimentos sobre um conjunto de oitenta redes de domínio público em áreas desde a saúde até as finanças. Os resultados indicam que as distribuições extraídas de LLMs superam tanto benchmarks aleatórios como uniformes, podendo inclusivamente atuar como priors especializados para afinar distribuições estimadas a partir de poucos dados observacionais. De modo complementar, Sekulovski et al. (2026) aproveitam LLMs para elicitar não só os valores das CPTs, como também para identificar relações de dependência condicional entre variáveis em modelos gráficos, enquanto Shaposhnyk et al. (2025) consideram LLMs como substitutos da elicitação humana para a construção de um modelo causa-efeito probabilístico na área biomédica e da saúde. Estes estudos mostram que LLMs podem servir como fonte de conhecimento probabilístico estruturado, mas concentram-se principalmente na etapa de *parametrização* do modelo, isto é, na geração prévia das CPTs, sem investigar de que forma esse conhecimento semântico pode ser reutilizado durante o processo de inferência, particularmente na resolução de problemas de otimização global, como a MPE. Este é o foco do presente estudo.

III. Metodologia

O experimento foi estruturado para avaliar o desempenho de LLMs na tarefa de inferência probabilística em diferentes BN do domínio médico. Para cada rede considerada, subconjuntos de variáveis foram selecionados como evidência (isto é, variáveis com valores atribuídos), e o modelo foi utilizado para inferir o valor das outras variáveis não observadas. As predições obtidas foram comparadas com aquelas produzidas

por inferência MPE exata, adotada como referência (*ground truth*) probabilística.

O procedimento proposto se subdivide em dois módulos complementares: (i) uma etapa de **Compilação**, na qual um LLM é utilizado para gerar, para cada variável da rede, uma tabela de decisões semânticas maximizadas equivalente a uma CPT tradicional; e (ii) uma etapa de **Inferência**, na qual as tabelas pré-compiladas são combinadas para aproximar a solução do problema de MPE mediante consultas diretas, sem novas chamadas ao modelo.

A. Fase da meta-compilação

A requisição feita ao modelo combina dois tipos de informação, produzidos em duas etapas offline, executadas uma única vez por rede antes de qualquer condicionamento a evidência: *notas de relacionamento* e *contexto de inferência da variável* (detalhado na Seção III-B).

Na primeira etapa, uma única requisição ao LLM gera, para cada variável não-raiz X_i, uma nota de relacionamento textual descrevendo qualitativamente como seus pais influenciam seu valor. O prompt utilizado[1] recebe a estrutura completa da rede, domínio, pais, filhos e variáveis que compartilham ao menos um pai de cada variável não-raiz, e solicita um parágrafo único para cada variável cobrindo cinco aspectos: (i) o pai dominante e sua direção de influência; (ii) como pais secundários modulam, atenuam ou interagem com esse efeito principal; (iii) regras explícitas de preferência entre os valores de domínio dada uma combinação específica de pais; (iv) como sinais de retroalimentação vindos de variáveis filhas devem reforçar ou sobrepor a pressão local dos pais; e (v) um erro de raciocínio comum que o próprio LLM tende a cometer para aquela variável.

B. Compilação

Seja $B = (G, \mathcal{D})$ uma BN, onde G é um grafo acíclico direcionado e $\mathcal{D}$ representa os domínios de suas variáveis. Seja E um conjunto de variáveis de evidência observadas. O problema clássico de MPE consiste em encontrar uma atribuição conjunta para todas as variáveis que não são de evidência que maximize a probabilidade a posteriori condicionada a E, isto é,

$$X^* = \arg\max_{X \notin E} P(X \mid E).$$

Utilizando apenas um LLM como fonte de conhecimento semântico, sem considerar as distribuições de probabilidade, aproximamos uma solução por meio de um processo de predição sequencial, no qual cada variável da rede é inferida a partir de um oráculo $\mathcal{M}$ (o LLM), condicionado às informações contextuais disponíveis.

Exemplo 1: Para tornar o procedimento concreto, considere a rede didática simples da Figura 1 (não associada aos datasets avaliados neste trabalho) composta por três variáveis: Fumante (X_1) e Poluição (X_2), ambas sem pais, e Falta de ar (X_3), cujos pais são X_1 e X_2.

[1]https://github.com/VictorRagazzi/pgm-llm-inference-reorganizado/blob/0ebe49115a1b66b59a6e3582299ab37fdd0a9a75/src/pgm_llm_inference/mpe/relationship_generation.py

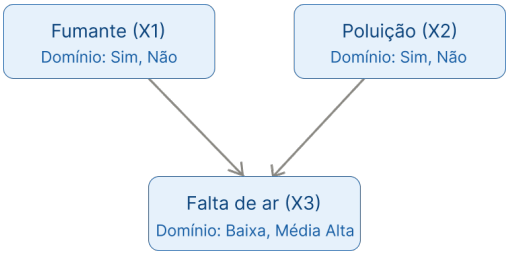


Fig. 1. Rede Bayesiana didática de três variáveis usada como exemplo ilustrativo do procedimento de compilação.

Para inferir X_3, é realizada uma **única** requisição ao oráculo $\mathcal{M}$. Essa requisição não pergunta pelo valor de X_3 isoladamente: ela envia, de uma só vez, o nome da variável, seu domínio e a tabela de contexto $\mathcal{R}_{X_3}$, contendo todas as combinações possíveis de valores dos pais X_1 e X_2, neste caso binário com quatro linhas. O oráculo responde com a tabela já preenchida por completo, atribuindo a cada linha o valor de domínio que maximiza sua resposta:

$$T_{X_3} = \begin{cases} \textit{Média} & \text{se } X_1 = \textit{Sim}, X_2 = \textit{Não}, \\ \textit{Alta} & \text{se } X_1 = \textit{Sim}, X_2 = \textit{Sim}, \\ \textit{Baixa} & \text{se } X_1 = \textit{Não}, X_2 = \textit{Não}, \\ \textit{Média} & \text{se } X_1 = \textit{Não}, X_2 = \textit{Sim}. \end{cases}$$

Para uma variável raiz como Poluição (X_2), que não possui pais, o conjunto $\mathrm{pa}(X_2) = \emptyset$ e, portanto, a tabela de contexto $\mathcal{R}_{X_2}$ possui uma única linha, correspondente à configuração vazia. Nesse caso, a requisição ao oráculo $\mathcal{M}$ não condiciona a resposta a nenhuma variável pai, retornando diretamente o valor de domínio mais provável a priori para X_2:

$$T_{X_2} = \big(\textit{Sim}\big).$$

Ou seja, o mesmo mecanismo de requisição única se aplica indistintamente a nós com e sem pais: a única diferença está na cardinalidade de $\mathcal{R}_i$, que degenera para 1 quando X_i é um nó raiz, reduzindo a tabela compilada a um único valor fixo, análogo a uma probabilidade marginal a priori. O mesmo procedimento se repete para X_1 (Fumante), também raiz, completando a compilação das três variáveis da rede didática.

O mesmo procedimento é repetido para todas as demais variáveis, resultando em um conjunto de tabelas pré-computadas, análogas a CPTs tradicionais da rede. Assim, a etapa de compilação requer uma requisição ao LLM para cada variável da rede, além de uma requisição adicional destinada à geração das notas de relacionamento (Seção III-A), totalizando exatamente $N+1$ requisições por rede, onde N é o número de variáveis (nós) da rede. Durante a etapa de inferência, quando a evidência fixa os valores dos pais de uma variável, a entrada correspondente é recuperada diretamente por consulta à tabela já compilada, sem qualquer chamada adicional ao LLM.

É importante ressaltar que o procedimento descrito nesta seção constitui uma **aproximação heurística** da inferência MPE, e não uma reformulação equivalente do problema de

otimização original. Ao decidir cada variável (X_i) sequencialmente, escolhendo o valor localmente ótimo com base apenas em seus pais, pode-se obter a solução global ótima em estruturas específicas. No entanto, em geral, essas decisões locais não garantem a configuração conjunta que maximiza a probabilidade posterior.

C. Inferência

A inferência aplica uma atribuição de evidência concreta ao conjunto de tabelas semânticas. Dada a evidência $\mathbf{e}$, os valores das variáveis observadas são fixados diretamente a partir de $\mathbf{e}$. As variáveis restantes (não pertencentes a E) são então percorridas em ordem topológica π: começando com $\hat{\mathbf{x}} = \emptyset$, cada variável X_i é visitada sucessivamente, e sua decisão é recuperada por meio de uma consulta direta à tabela T_i, usando o contexto atual $\mathbf{S}_i$ como chave. Esse contexto é formado pela combinação dos valores de evidência de quaisquer pais fixados por $\mathbf{e}$ com as decisões já acumuladas em $\hat{\mathbf{x}}$ para os demais pais, os quais garantidamente já foram resolvidos anteriormente na ordem topológica, uma vez que π respeita a estrutura de dependência do grafo. O valor recuperado $\hat{x}_i$ é então anexado a $\hat{\mathbf{x}}$, e o processo se repete até que todas as variáveis não pertencentes a E tenham recebido uma atribuição. A atribuição final $\hat{\mathbf{x}}$ é retornada como a configuração MPE prevista.

Como todas as decisões foram geradas durante a compilação, a etapa de inferência propriamente dita não requer qualquer chamada ao LLM: seu custo computacional é determinado exclusivamente pelo número de consultas às tabelas T_i, uma por variável não pertencente a E.

Exemplo 2: Aplicando esse processo ao cenário da Figura 1 com evidência $X_2 = Sim$, a ordem topológica resolve sucessivamente X_1 (retornando *Não* via tabela T_1) e X_3, retornando Média via consulta a T_3 com a chave:

$$\{X_1 = \textit{Não}, X_2 = \textit{Sim}\},$$

obtendo a configuração conjunta final:

$$\hat{\mathbf{x}} = \{X_1 = \textit{Não},\ X_3 = \textit{Média}\},$$

obtida por meio de duas consultas a tabelas pré-compiladas, sem qualquer chamada ao LLM.

IV. Setup experimental

Para avaliar o método proposto, foram utilizadas 10 BN de domínios distintos da medicina, com diferentes tamanhos e complexidades estruturais, obtidas do repositório `bnlearn` (Scutari, 2010). As características de cada rede são apresentadas na Tabela I. Nesta tabela, o número de pais max. corresponde ao maior valor, entre todos os nós da rede, do número de pais; já a profundidade corresponde ao maior caminho dirigido do grafo, medido em número de arestas, considerando as variáveis raiz como tendo profundidade zero em relação a si mesmas.

O modelo de linguagem utilizado em todos os experimentos foi o DeepSeek-V4-Pro, escolhido pela boa relação entre custo computacional e desempenho em raciocínio. Todas as inferências foram realizadas com temperatura igual a 0.0, tornando as respostas determinísticas.

TABLE I
Características das Redes Bayesianas utilizadas.

Domínio	Nós	Arestas	Pais máx.	Profundidade
TDAH	19	18	1	1
Sintomas de Covid-2	12	30	9	9
Sintomas de Covid-1	12	11	1	1
Gonorreia	10	9	3	3
Hepatite	70	123	6	7
Monitoramento de UTI	37	46	4	10
Alergia-2	8	8	2	4
Alergia-1	8	12	3	3
Diabetes	9	11	4	3
Doenças pediátricas	20	25	2	4

Para cada rede, o procedimento de Compilação descrito na Seção III-B foi executado uma única vez, gerando as tabelas semânticas T_i para todas as variáveis com exatas $N + 1$ requisições à LLM, sendo N o número de nós da rede e mais uma requisição para a meta-compilação. Em seguida, a etapa de Inferência foi aplicada repetidamente, variando o conjunto de evidência E, sem novas chamadas ao LLM.

O número de variáveis de evidência foi incrementado progressivamente, de 1 até 50% do total de nós da rede. Para cada tamanho de evidência, foram realizadas 3 repetições, cada uma com um subconjunto de variáveis de evidência escolhido aleatoriamente, de forma a reduzir o efeito de uma escolha particular de evidência sobre o resultado. Em cada repetição, o valor inferido pelo modelo para as variáveis não observadas foi comparado ao valor produzido por inferência MPE exata.

A hipótese para essa abordagem é que o fornecimento de mais evidência gere mais contexto sobre o caso clínico e melhore as decisões do LLM em relação ao MPE exato.

Os valores atribuídos às variáveis de evidência foram definidos de acordo com a própria solução MPE exata e **incondicional** da rede. Essa escolha isola o comportamento do LLM na etapa de decisão condicionada a uma evidência coerente com o alvo, mas constitui um viés experimental, uma vez que cenários clínicos reais podem apresentar evidências (sintomas, achados de exame) que não são necessariamente consistentes com a explicação mais provável. O impacto dessa escolha é quantificado na Seção V-C, onde um segundo cenário, com evidência definida de forma independente do MPE, é comparado ao protocolo original.

Em cada repetição, os valores inferidos para as variáveis não observadas foram comparados aos valores obtidos por inferência MPE exata na mesma rede e com o mesmo conjunto de evidências. A acurácia foi calculada como a proporção de variáveis cujo valor predito coincidiu com o valor da solução MPE exata:

$$\mathrm{Acc}(E) = \frac{\text{Número de variáveis corretamente inferidas}}{\text{Número total de variáveis não observadas}}.$$

Além da acurácia por variável, foi calculada a correspondência da configuração conjunta (*exact match*), que atribui

valor 1 quando a configuração MPE é recuperada completamente, e 0 caso contrário.

V. Resultados

Inicialmente, os resultados serão apresentados para cada dataset com apenas uma evidência. Em seguida, será realizada uma análise se o aumento do número de evidências agrega valor nas inferências realizadas, e se o critério usado para definir os valores de evidência afeta esse comportamento. Por fim, investiga-se quais fatores estruturais e de nível de variável estão associados ao desempenho do método.

A. Acurácia por dataset

A Tabela II apresenta a acurácia média obtida em cada uma das dez redes avaliadas, na condição em que os valores de evidência são definidos de acordo com a solução MPE (Seção IV). TDAH e Sintomas de Covid-2 atingem acurácia de 100%, em contrapartida, redes como Alergia-2 (65.3%), Alergia-1 (64.6%), Diabetes (61.3%) e Doenças pediátricas (60.7%) apresentam os piores desempenhos médios, na faixa de sessenta por cento.

TABLE II
Acurácia média por nó e correspondência exata de inferência MPE via LLM para cada uma das dez Redes Bayesianas avaliadas.

Dataset	Acurácia por var. (%)	Exact match (%)
TDAH	100.0	100.0
Sintomas de Covid-2	100.0	100.0
Sintomas de Covid-1	92.7	33.3
Gonorreia	87.6	16.0
Hepatite	79.3	0.0
Monitoramento de UTI	79.1	0.0
Alergia-2	65.3	0.0
Alergia-1	64.6	0.0
Diabetes	61.3	0.0
Doenças pediátricas	60.7	0.0

A recuperação do MPE completo se mostra uma tarefa difícil, já que qualquer variável inferida incorretamente já resulta em uma taxa de 0, e essa dificuldade é refletida nos resultados obtidos na Tabela II.

B. Efeito da proporção da evidência

A Figura 2 detalha a acurácia média em função da proporção de variáveis de evidência para as dez redes avaliadas deste trabalho. O comportamento observado não é uniforme entre os domínios, permitindo identificar três padrões distintos, por exemplo: TDAH mantém acurácia de 100% independentemente da proporção de evidência; Hepatite, Monitoramento de UTI e Doenças pediátricas melhoram conforme a evidência aumenta (de aproximadamente 70% para 84%, 71% para 83% e 53% para 70%, respectivamente); e Gonorreia e os dois datasets de Alergia pioram (Gonorreia de aproximadamente 90% para 84%, e Alergia-1 e 2 de aproximadamente 80% para 60% ou menos).

Já o *exact match* permanece nulo na maioria das redes independentemente da proporção de evidência, diferente da tendência gradual observada na acurácia por variável, com algumas exceções.

C. Efeito do critério de definição da evidência

Os experimentos citados foram realizados utilizando valores de evidência definidos de acordo com a própria solução MPE exata da rede; denominamos essa condição *evidência MPE-consistente*. Também foi realizado um segundo experimento, nas mesmas dez redes, em que os valores de evidência foram definidos de forma independente da solução MPE, sendo contraditórios entre o MPE e entre si; denominamos essa segunda condição *evidência MPE-inconsistente*. Comparando as duas condições, a acurácia média por variável cai de 80.1% (MPE-consistente) para 58.2% (MPE-inconsistente), e a taxa de *exact match* cai de 26.2% para 10.6%, como visto na Figura 3.

D. Fatores associados ao desempenho

Os resultados das seções anteriores revelam que o desempenho do método proposto provavelmente é determinado pelo domínio do dataset: diferenças de acurácia entre redes de tamanho semelhante (por exemplo, Sintomas de Covid-2 e Doenças pediátricas, com 12 e 20 nós, mas desempenhos de 100% e 60.7%, respectivamente) sugerem que a estrutura por si só não explica o padrão observado. Testando a hipótese formalmente via correlação de Spearman entre as métricas (acurácia por variável e *exact match*) e características estruturais em nível de rede (número de nós, arestas, número de pais máximo e profundidade; $n = 10$), nenhuma das correlações testadas se mostrou estatisticamente significativa (por exemplo, acurácia $\times$ profundidade: $\rho = -0.068$, $p = 0.852$; *exact match* $\times$ profundidade: $\rho = -0.366$, $p = 0.298$). Esse resultado demonstra que não há evidência estatística de que a profundidade, o número de pais max ou o tamanho da rede, isoladamente, expliquem o desempenho do método, e a hipótese que melhor explica, é que a acurácia depende do domínio da rede. O caso de Sintomas de Covid-2, uma das rede estruturalmente mais complexa do conjunto (número de pais max. 9, profundidade 9) atingindo 100% de acurácia, mostra essa desconexão entre a complexidade da rede e o desempenho.

Como a análise anterior é limitada pelo número de redes disponíveis (n=10), complementamos-na com uma comparação em nível de variável. Para cada variável avaliada nas dez redes, relacionamos seu número de pais (grau de entrada) à acurácia obtida especificamente para aquela variável, o que eleva substancialmente o tamanho amostral em relação à análise por rede (Figura 4: $n = 39$ para variáveis sem pais, $n = 89$ para uma variável pai, $n = 48$ para duas, $n = 18$ para três e $n = 6$ para quatro; categorias com $n < 4$ foram omitidas). Nessa análise, observa-se uma tendência de queda na acurácia conforme o número de pais aumenta (de aproximadamente 0.93 para variáveis sem pais a aproximadamente 0.47 para variáveis com quatro pais), acompanhada de dispersão crescente entre as categorias intermediárias.

Ademais, os experimentos da Seção V-A demonstram o efeito de uma evidência selecionada aleatoriamente em cada rede, então, é justo pensar se a posição da evidência selecionada, pode afetar o resultado do método. Foi realizado

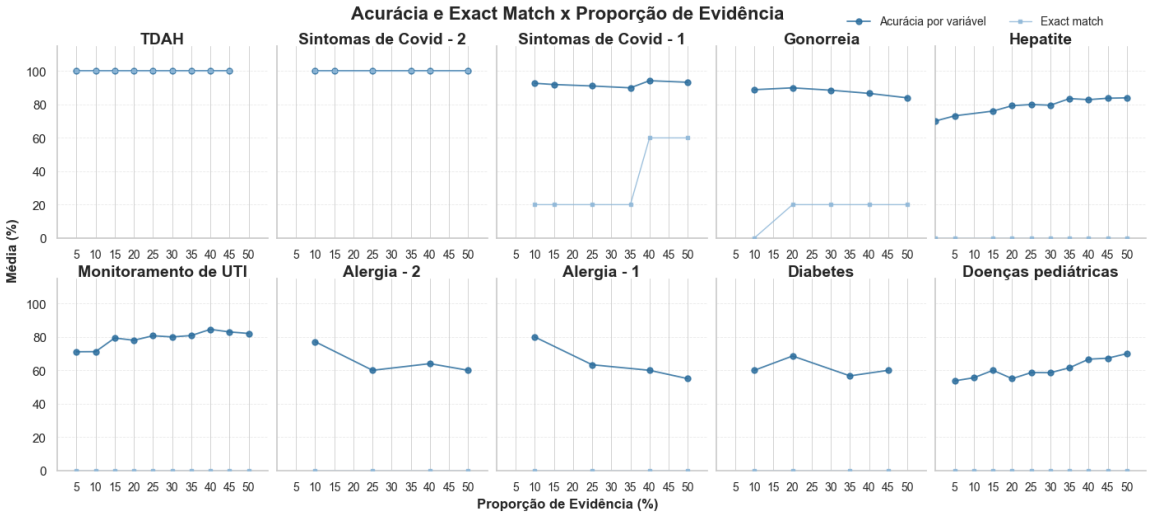


Fig. 2. Acurácia média de inferência MPE via LLM em função da proporção de variáveis de evidência. O eixo horizontal representa faixas percentuais do total de nós (5% a 50%), usadas para agregar os resultados entre redes de tamanhos distintos; o protocolo experimental (Seção IV) parte de 1 variável de evidência.

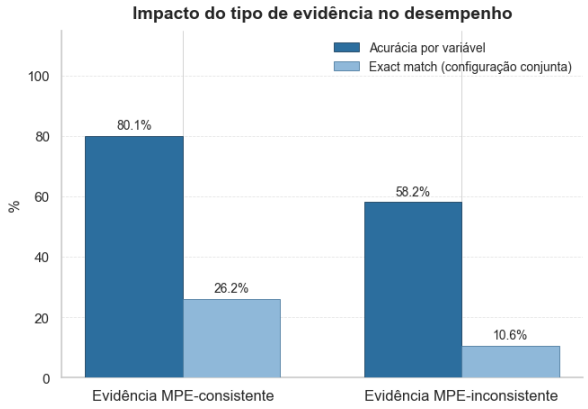


Fig. 3. Acurácia média por variável e taxa de *exact match* da configuração conjunta, comparando evidência definida de acordo com o MPE (evidência MPE-consistente) e evidência definida de forma independente do MPE (evidência MPE-inconsistente), agregadas sobre as dez redes avaliadas.

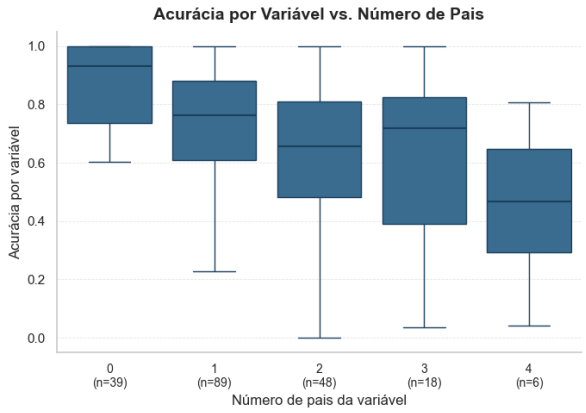


Fig. 4. Acurácia por variável em função do número de pais da variável, agregada sobre as dez redes avaliadas. Categorias com $n < 4$ (5, 6 e 9 pais) foram omitidas.

um experimento exaustivo em que, para cada uma das dez redes, cada variável foi utilizada individualmente como única evidência, associando o resultado à profundidade topológica normalizada do nó-evidência (0 = raiz, 1 = folha mais funda).

A Figura 5 apresenta a acurácia por variável em função da posição normalizada da evidência, agregada sobre as dez redes ($n = 205$ execuções). Observa-se uma fraca tendência de aumento da acurácia à medida que a evidência se afasta da raiz e se aproxima das folhas, confirmada pela correlação de Spearman entre posição normalizada e acurácia ($\rho = 0.294$, $p = 1.84 \times 10^{-5}$) e entre posição normalizada e *exact match* ($\rho = 0.250$, $p = 3.03 \times 10^{-4}$), ambas estatisticamente significativas, embora de magnitude modesta.

VI. Discussão

Os resultados apresentados na Seção V-D indicam que, embora as características estruturais da rede consideradas isoladamente (número de nós, arestas, número de pais max. e profundidade) não expliquem estatisticamente o desempenho do método, o número de pais de cada variável está associado a uma queda de acurácia por variável: variáveis com mais pais impõem um contexto maior para o LLM decidir corretamente durante a Compilação, o que pode gerar acúmulo de erro nas decisões, mesmo não se traduzindo em uma correlação significativa quando agregada em nível de rede. Esse padrão entre número de pais e acurácia, mostra uma oportunidade de evoluir o método na etapa de Compilação (Seção III-B), e não na etapa de Inferência (que é puramente combinatória, dado que as tabelas já estão compiladas): a qualidade de

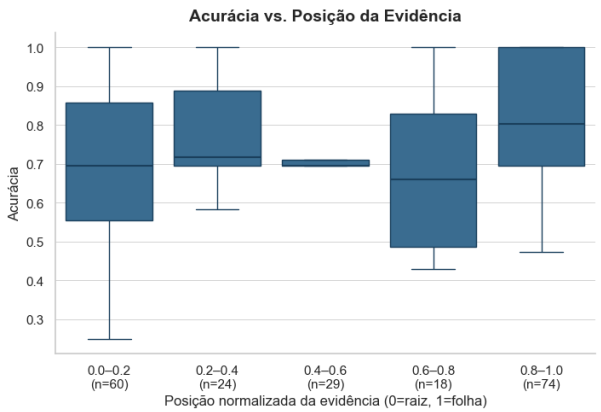


Fig. 5. Acurácia por variável em função da posição normalizada da evidência (0 = raiz, 1 = folha mais funda), agregada sobre as dez redes avaliadas.

cada T_i depende de quão bem o LLM consegue traduzir o significado real da variável e de seus pais para uma decisão local, sem conhecer a distribuição de probabilidade subjacente. Variáveis raiz representam um caso especial desse problema, pois sua tabela compilada consiste em um único valor a priori, determinado exclusivamente pelo conhecimento de mundo do modelo, sem nenhum pai que o possa corrigir. Embora, em média, variáveis-raiz tenham desempenho alto, nos casos específicos em que o prior do LLM diverge da distribuição real da população onde a rede foi construída (como se observou na variável `Massa corporal` da rede de Diabetes), ele é propagado para toda a cadeia de inferência seguinte.

Uma solução que pode resolver o problema de perda de acurácia com o aumento de contexto de pais, seria a adição de notas descritivas. Essas notas apresentariam descrições sobre como cada variável se influencia, semelhante às notas de relacionamento, entretanto, geradas por especialistas entendidos das relações entre variáveis e de possíveis vieses que uma LLM pode considerar, ao pensar em um escopo geral, e não focado no dataset. Essas notas podem trazer um poder de raciocínio para o oráculo semelhante a uma estratégia de prompt já conhecida na literatura, a estratégia de *few-shot* (Brown et al., 2020).

O efeito da proporção de evidência (Figura 2) confirma parcialmente essa leitura. Em Hepatite, Monitoramento de UTI e Doenças pediátricas, mais evidência tende a melhorar a acurácia, consistente com a hipótese de que mais contexto factual reduz a ambiguidade da inferência para o LLM. Entretanto, em Gonorreia e Diabetes esse efeito não se confirma: a acurácia se mantém estável ou piora levemente conforme mais variáveis são fixadas como evidência.

A. Limitações

É importante enfatizar que os resultados aqui relatados devem ser interpretados como evidência de que o procedimento proposto *aproxima*, sob determinadas condições, a solução do MPE, não como demonstração de que a inferência sequencial local seja equivalente à otimização conjunta global. Todos os experimentos foram conduzidos com um único modelo LLM.

Outra observação é que quando a evidência deixa de estar alinhada com a explicação mais provável, tanto a acurácia por variável quanto, de forma ainda mais acentuada, o *exact match* da configuração conjunta se deterioram (Figura 3), sugerindo que os ganhos observados com mais evidência no protocolo original dependem, ao menos em parte, de essa evidência ser informativa e coerente com o alvo. Entretanto, essa análise pode ser tão enviesada quanto utilizar apenas evidências condizentes, já que os valores das evidências foram definidos de forma a serem contraditórios entre o MPE e entre si, sendo casos pouco representativos do mundo real.

Essas lacunas não comprometem os achados centrais deste trabalho, mas devem ser investigadas antes que o método seja considerado maduro para validação clínica.

B. Implicações práticas

Os resultados apontam para um próximo passo: o uso de metadados adicionais gerados por especialistas humanos, completando o contexto gerado na fase da meta-compilação (Seção III-A). Poderá ser útil nos casos em que o método atual falha, em casos de variáveis raízes cujos valores não são intuitivos a partir do conhecimento de mundo do LLM, ou em redes cujas relações de causa e efeito são também ocultas pela semântica. Um metadado explícito dizendo qual a distribuição de uma variável raiz naquele grupo de estudos, em vez da distribuição baseada no conhecimento geral do LLM, pode ser o viés necessário que corrige precisamente o erro que se vê.

Do ponto de vista prático, o método proposto não é um substituto da inferência de probabilidade exata quando se tem acesso a dados e à BN completa (com CPTs numéricas). O seu valor está em ser uma solução para o caso em que esses dados não existem ou não estão disponíveis, mas o conhecimento de domínio necessário para descrever qualitativamente as relações de causa e efeito entre variáveis está disponível, seja na forma de conhecimento médico geral, seja via o conhecimento do especialista como nota. Neste caso, os resultados indicam que o método consegue aproximar bem (acima de 60% de acerto, chegando a 100% em algumas redes) a solução exata do MPE, em especial em redes onde a semântica das variáveis é mais clara.

VII. Conclusão

Em suma, este trabalho mostrou que um LLM, utilizando apenas a semântica das variáveis e um procedimento de decisões sequenciais locais, consegue aproximar a solução do MPE em BN de domínio médico, com acurácia por variável entre 60.7% e 100% entre as dez redes avaliadas. Essa aproximação, no entanto, é heurística e não uma reformulação equivalente da otimização conjunta, sendo sensível à consistência da evidência fornecida e não garantindo a recuperação da configuração MPE completa. O método proposto se mostra, assim, uma alternativa de baixo custo para cenários em que uma BN ainda não possui CPTs numéricas, mas o conhecimento de domínio necessário para descrever as relações entre variáveis está disponível, e deve-se, antes de

qualquer uso em contextos clínicos reais, ser complementado por notas de especialistas ou por metadados adicionais que mitiguem as limitações identificadas.

VIII. Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido, essencial para o desenvolvimento desta pesquisa.

References

J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

A. Darwiche, *Modeling and Reasoning with Bayesian Networks*. New York, NY: Cambridge University Press, 2009.

R. Dechter, "Bucket elimination: A unifying framework for probabilistic inference," in *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann, 1996, pp. 211–219.

D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," 2024.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

M. Omar, D. Brin, B. Glicksberg, and E. Klang, "Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review," *American Journal of Infection Control*, vol. 52, pp. 992–1001, 2024.

S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, Y. Fang, S. Ding, J. Wang, K. Xu, L. Xia, J. Yeung, D. Zha, D. Cai, G. B. Melton, M. Lin, and R. Zhang, "Large language models for disease diagnosis: A scoping review," *npj Artificial Intelligence*, vol. 1, p. 9, 2025, preprint: arXiv:2409.00097.

A. Niset, I. Melot, M. Pireau, A. Englebert, N. Scius, J. Flament, S. El Hadwe, M. Al Barajraji, H. Thonon, and S. Barrit, "Grounded large language models for diagnostic prediction in real-world emergency department settings," *JAMIA Open*, vol. 8, no. 5, p. ooaf119, 2025.

Y. Kesmen, F. Elhassan, J. Ma, J. Stalhandske, Y. Chang, D. Sasu, A. Kulinkina, A. Arora, L. Klein, and M.-A. Hartley, "Mobayes: A modular bayesian framework for separating reasoning from language in conversational clinical decision support," *arXiv preprint arXiv:2604.20022*, 2026.

K. Polotskaya, C. S. Muñoz-Valencia, A. Rabasa, J. A. Quesada-Rico, D. Orozco-Beltrán, and X. Barber, "Bayesian networks for the diagnosis and prognosis of diseases: A scoping review," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 1243–1262, 2024.

J. Vila-Francés, J. Sanchis, E. Soria-Olivas, A. Serrano, M. Martinez-Sober, C. Bonanad, and S. Ventura, "Expert system for predicting unstable angina based on bayesian networks," *Expert Systems with Applications*, vol. 40, pp. 5004–5010, 2013.

C. W. Safranek, V. Socrates, D. Wright, and R. A. Taylor, "Simulated evaluation of large language model stepwise diagnostic reasoning with real-world chest pain encounters and bayesian networks," *BMC Medical Informatics and Decision Making*, vol. 26, p. 97, 2026.

T. Suraj, "A bayesian framework for llm-enhanced history-taking in recurrent medical conditions to improve treatment outcomes: An empirical evaluation," *Intelligence-Based Medicine*, 2025, scienceDirect PII: S2666521225000869.

A. Nafar, K. B. Venable, Z. Cui, and P. Kordjamshidi, "Extracting probabilistic knowledge from large language models for bayesian network parameterization," 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:278789642

N. Sekulovski, M. Waaijers, and G. Arena, "Llm-based prior elicitation for bayesian graphical modeling," *British Journal of Mathematical and Statistical Psychology*, 2026.

O. Shaposhnyk, D. Zahorska, and S. Yanushkevich, "Can llms assist expert elicitation for probabilistic causal modeling?" *arXiv preprint arXiv:2504.10397*, 2025.

M. Scutari, "Learning bayesian networks with the bnlearn R package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.

Evaluating Synthetic Data Generation from Expert-Defined Scenarios for Autism Pre-Screening

1st Letícia Lima M. Martins de Castro
Department of Mechanical Engineering
University of Campinas (UNICAMP)
Campinas, Brazil
l250678@unicamp.br

2nd Estevão Esmi Laureano
Department of Applied Mathematics
University of Campinas (UNICAMP)
Campinas, Brazil
eesmi@unicamp.br

3rd Laécio C. de Barros
Department of Applied Mathematics
University of Campinas (UNICAMP)
Campinas, Brazil
laecioba@unicamp.br

***Abstract*—Autism Spectrum Disorder (ASD) diagnosis in children typically relies on multiple assessments conducted by specialists, whose high financial cost may limit access for low-income populations and restrict the availability of these services within public healthcare systems. These limitations also contribute to the scarcity of structured datasets suitable for training machine learning models. Moreover, available diagnostic reports are often heterogeneous in format and require substantial preprocessing to extract structured information. To address these challenges, this work presents a methodology for generating ASD-related data based on expert knowledge combined with synthetic data generation techniques. The proposed approach uses expert-defined scenarios to generate synthetic datasets for training and internally evaluating machine learning models related to the pre-screening of ASD in children aged 0 to 6 years.**

***Index Terms*—Autism Spectrum Disorder, Synthetic Data Generation, Expert Knowledge, Data Augmentation, Machine Learning, Pre-Screening.**

I. Introduction

Diagnosing Autism Spectrum Disorder (ASD) requires the integration and assessment of multiple behavioral, social, cognitive, and developmental aspects [3]. Reaching an accurate diagnosis is a challenging task because ASD encompasses a broad spectrum of presentations, with each individual exhibiting different characteristics and varying levels of support needs. Furthermore, the diagnostic process depends not only on the clinical expertise of the professional but also on the active collaboration of parents or caregivers [4]. In addition to these intrinsic diagnostic challenges, comprehensive ASD assessments are often time-consuming and costly, as they typically require the administration of multiple standardized instruments over several clinical sessions. These financial and logistical demands may limit access to diagnosis among low-income populations and pose additional challenges for publicly funded healthcare systems, such as the Brazilian Unified Health System (SUS).

Furthermore, gathering reliable clinical data becomes even more complex when focusing on early childhood, specifically the age group from 0 to 6 years. Obtaining structured data within this developmental period is particularly challenging, as early ASD-related characteristics can be subtle and often overlap with typical developmental variability or other neurodevelopmental conditions. Moreover, younger children typically display camouflaging behaviors less frequently and less intensely, factor which can facilitate the identification. Consequently, early childhood is widely recognized as an important period for identification and intervention, as timely therapeutic support may contribute to the development of communication, social, adaptive, and cognitive skills [20]. Early intervention can reduce long-term support needs, generating a positive financial impact on care budgets.

One possible strategy for supporting ASD pre-screening and referral processes is the use of machine learning methods. However, a major challenge is that such models typically require large amounts of high-quality data for training [6]. In this context, high quality implies that the data accurately represent real-world clinical diversity, allowing the model's generalizability to be evaluated across independent environments, such as different hospitals or clinical settings. In the case of ASD, structured datasets are scarce. Besides the limited access to diagnostic assessments, the available information is often embedded in heterogeneous clinical reports that must be extracted, standardized, filtered, and preprocessed before it can be used for machine learning. This entire process is labor-intensive, time-consuming, and costly, which largely explains the scarcity of high-quality datasets for developing reliable machine learning models related to ASD pre-screening, particularly in the Brazilian context [7].

Synthetic data generation has been investigated as a strategy for supporting machine learning research when access to sufficiently large and structured datasets is limited. For tabular data, probabilistic graphical models, copula-based methods, and conditional probabilistic generators seek to preserve different statistical properties of a reference dataset. However, statistical similarity alone does not guarantee predictive utility, privacy, or generalization to independent populations. Therefore, the quality of synthetic data must be evaluated using complementary criteria. [1] A more detailed discussion of synthetic data generation and evaluation methodologies is presented in the following section.

In this paper, we investigate an approach for generating data that can be used to train machine learning models through a

This work was partially supported by FAPESP under Grant No. 2024/07285-5 and by CNPq under Grants No. 314885/2021-8 and 311976/2023-9.

two-step framework. The first step consists of eliciting domain knowledge from an ASD specialist through interviews and structured meetings. The objective is to identify relevant predictive variables and extract inference rule scenarios based on the specialist's expertise aiming to translate complex cognitive relations and jointly characterize scenarios associated with positive or negative ASD-related classifications (hypothetical individuals) rooted in her clinical experience. The resulting rule base represents our initial reference dataset, where each individual corresponds to an expert-defined and clinically plausible combination of characteristics. These rules represent structured expert knowledge rather than clinical records from evaluated children. Knowledge elicitation from domain experts has long been recognized as a fundamental step in the development of expert systems and knowledge-based decision-support systems [9].

The second step focuses on the application of synthetic data generation techniques to produce additional samples for training machine learning models related to ASD prescreening research. The use of synthetic data has emerged as a potential strategy for addressing data scarcity in machine learning applications, particularly in healthcare, where privacy constraints and the high cost of data collection often limit the availability of large, high-quality datasets [11]. Nevertheless, synthetic data reproduce only the information available in the source dataset and cannot, by themselves, establish clinical representativeness.

To evaluate the proposed data generation framework, we conducted a series of computational experiments using machine learning classifiers and a fuzzy rule-based classifier (without learning). In particular, we evaluated three methods for generating binary synthetic data: a Gaussian Copula model, a Bayesian Network, and a conditional Beta-Bernoulli model [12]. To prevent generative data leakage, the reference scenarios were partitioned before fitting the generators. In each cross-validation fold, every generator was fitted exclusively using the reference training partition, while the held-out reference scenarios remained isolated from generator fitting, synthetic data generation, classifier training, and model selection.

The experiments were designed to investigate three main questions regarding the generated synthetic data. First, we assessed how well the synthetic data preserve the statistical properties and dependency structure of the expert-defined reference dataset. Second, we investigated whether classifiers trained on synthetic data can transfer the learned patterns to held-out reference scenarios that were not used to fit either the generator or the classifier. Third, we evaluated whether augmenting the reference training data with synthetic samples provides consistent predictive improvements over reference-only and synthetic-only training.

Because synthetic data quality is multidimensional , the generators were evaluated using complementary measures of statistical fidelity, diversity, and predictive utility [2]. We also examined whether the relative performance of the generators remained consistent across classifiers with different inductive biases, including linear models (Logistic Regression), kernel-based methods (Support Vector Machines), tree-based ensembles (Random Forest and XGBoost), and fuzzy rule-based systems.

This distinction is essential because statistical alignment does not guarantee optimal predictive transfer, nor do improvements in AUC-ROC reliably reflect gains in MCC, sensitivity, specificity, or overall stability. Therefore, instead of relying on a single metric, model and generator selection was based on a heuristic interpretation of each evaluation dimension and its variance across repeated cross-validation partitions.

II. Reference Dataset Construction and Synthetic Data Generation

This section describes the construction of the expert-defined reference dataset and the generation of synthetic binary samples. The framework comprises knowledge elicitation, variable selection, and synthetic data generation.

A. Knowledge Elicitation from an ASD Specialist

Developing predictive machine learning models for Autism Spectrum Disorder (ASD) requires datasets that adequately capture relevant clinical and contextual variables. However, existing options present significant limitations [4], [28]. For instance, benchmark datasets such as the one by [8] rely on ten Q-CHAT screening items [13] and basic demographics, omitting critical developmental factors highlighted by domain specialists (e.g., language acquisition milestones, developmental regression, and detailed clinical history). Furthermore, its caregiver-report nature, minimal representation of the Latino population and other populations, and limited documentation restrict its applicability to local clinical contexts.

To overcome these contextual gaps, we elicited domain knowledge from a neuropsychologist (Regina C. Antão Lima) with over ten years of experience in assessing ASD and other neurodevelopmental conditions. Following established expert system methodologies [9], this knowledge acquisition process was conducted in two main stages: feature identification and scenario formulation.

First, the specialist identified 34 binary predictive attributes (features) considered critical for clinical observation and ASD assessment. These variables were derived from standard assessment instruments recommended in the literature and are essential for differentiating ASD-related characteristics from those associated with other neurodevelopmental conditions. In this framework, a value of 1 indicates the presence of the corresponding attribute, whereas a value of 0 indicates its absence. The complete set of 34 attributes is detailed in Table I.

In the second stage, the objective was to formulate inference rules representing specific combinations of these 34 characteristics associated with positive or negative ASD classifications. Given the high density and manual synthesis required for each expert-defined decision rule, this process yielded a compact reference dataset of 40 clinically plausible scenarios. Although small in volume, each instance explicitly represents complex combinations of behavioral and developmental traits, such as

severe impairments in social interaction paired with delayed language acquisition, providing a domain-validated foundation for interpretable screening.

At the end of this process, the initial reference dataset comprised 40 expert-defined scenarios (instances), each represented by the 34 binary predictive attributes and one binary target assigned according to the elicited expert rules. However, a critical limitation must be acknowledged: these scenarios represent structured expert knowledge rather than clinical observations of independently assessed children. Furthermore, the manual acquisition and processing of this dataset required translating highly nuanced clinical judgments into simplified binary attributes, a data treatment step that inherently introduces subjectivity and may obscure the gradual severity of ASD symptoms. Consequently, these instances carry biases inherent to expert elicitation and should not be interpreted as actual patient records or as a statistically representative sample of the broader ASD population. Nonetheless, this simplified representation was necessary to enable a practical framework for future real-world deployment and to facilitate validation by other domain specialists.

TABLE I
PREDICTIVE ATTRIBUTES ELICITED FROM THE ASD SPECIALIST.

ID	Attribute	ID	Attribute
V1	Repetitive behaviors	V18	Aversion to parties
V2	Sensory hypersensitivity	V19	Avoidance of physical affection
V3	Theory of mind difficulties	V20	Atypical use of toys
V4	Learning difficulties	V21	Delayed speech
V5	Poor motor coordination	V22	Impaired speech
V6	Hyperactivity	V23	Atypical language (echolalia)
V7	Repetitive motor movements	V24	Social interaction difficulties
V8	Unusual body postures	V25	Limited interest in toys
V9	Developmental regression	V26	Limited visual communication
V10	Inflexibility	V27	Restricted interests (hyperfocus)
V11	Inappropriate emotional response	V28	Loss of interest
V12	Lack of danger awareness	V29	Difficulty understanding speech
V13	Sleep or pain abnormalities	V30	Excessive fears
V14	Impulsiveness	V31	Frequent screaming
V15	Oral sensory-seeking behaviors	V32	Conflicts with peers
V16	Difficulty accepting limits	V33	Lack of friendships
V17	Lack of independence	V34	Reduced facial expression

To accommodate user accessibility and streamline future evaluation, the 34 attributes were defined as concise feature names for technical modeling (as summarized in Table I), but originally formulated as behavioral questions (e.g., assessing pacing or continuous jumping instead of a raw label like "agitation"). This questionnaire-based representation ensures the framework remains intuitive for non-expert respondents and facilitates validation by other domain specialists.

B. Synthetic Data Generation

The resulting dataset was relatively small compared with its dimensionality, consisting of 40 expert-defined scenarios described by 34 binary predictor variables. Among these scenarios, 28 were assigned to the ASD-positive class and 12 to the non.

A strategy that has been successfully employed to mitigate data scarcity in medical applications is the generation of synthetic data [11], [27], [26].

Synthetic data generation has been investigated as a strategy for increasing the amount of training data while preserving relevant statistical properties of the original dataset. Its use has been explored in healthcare and other data-constrained applications [1].

In this study, we evaluated two synthetic data generation approaches for binary data: the Gaussian Copula and the Conditional Beta–Bernoulli generator. These methods were selected because they represent complementary strategies for modeling dependencies between predictor variables and the target. Both generators were trained strictly on the training fold of each cross-validation partition, while the corresponding test fold remained isolated for final evaluation.

To prevent data leakage, both generators were trained strictly on the training fold of cross-validation partition, leaving the test fold entirely isolated for model evaluation.

Conditional Beta-Bernoulli (Conditional Independence): Estimates the target class prior $P(y)$ and generates attributes X_j conditionally and independently given the class label, $P(X_j = 1 \mid y)$, with additive Laplace smoothing ($\alpha = 1.0$). As a conservative and computationally efficient baseline, it minimizes parameter variance and overfitting risk, though it omits direct feature-to-feature correlations. Binary Gaussian Copula (Joint Dependency): Captures the joint distribution of features and targets by mapping observed marginal probabilities and pair-wise binary correlations into a continuous Gaussian latent space. Synthetic continuous vectors are sampled from a Multivariate Normal distribution and converted back to binary values using quantile thresholding ($\alpha = 0.5$). This approach explicitly preserves complex clinical symptom co-occurrences.

Comparing these two paradigms isolates the effect of inductive bias on synthetic transfer. Exploring whether modeling complex feature interactions provides superior predictive utility or whether a simpler, lower-variance generator proves more resilient Although Bayesian Networks were also preliminary investigated, they were ultimately excluded from the final framework, as their predictive performance and evaluation metrics were unsatisfactory given the sample size constraints.

III. CLASSIFIERS, METRICS, AND VALIDATION PROTOCOL

The synthetic datasets were evaluated from complementary perspectives to determine whether they preserve relevant properties of the expert-defined reference data and whether the generated samples provide useful information for ASD-related classification experiments. The evaluation comprised three main dimensions: statistical fidelity, predictive utility, and sample diversity.

Statistical fidelity was evaluated by comparing the marginal distributions, target prevalence, and pairwise dependency structures of the synthetic and reference data. Correlation Root Mean Squared Error (RMSE) was used to summarize differences between attribute frequencies and between the corresponding association matrices. In addition, a classifier two-sample test was employed to assess whether synthetic and reference samples could be distinguished computationally.

Sample diversity and exact-record overlap were also examined to identify limited coverage of the reference space or excessive reproduction of expert-defined scenarios.

Predictive utility was assessed by training classifiers under reference-only, synthetic-only, and augmented training conditions. The evaluated algorithms included Logistic Regression (LR) , Support Vector Machines (SVM) , Random Forests (RF) , Extreme Gradient Boosting (XGBoost) , and fuzzy rule-based classifiers. These models represent different learning assumptions: LR provides a linear probabilistic baseline, SVM supports margin-based and potentially nonlinear decision boundaries, RF and XGBoost represent nonlinear interactions through tree-based ensembles, and fuzzy classifiers provide an interpretable rule-based representation. Using classifiers with different learning mechanisms makes it possible to examine whether the predictive relationships captured by each generator remain useful across modeling approaches. However, performance differences among classifiers do not necessarily indicate poor synthetic-data quality, since each model has its own inductive bias and sensitivity to the data structure. Therefore, predictive utility was evaluated using complementary metrics, including MCC, F1-score, AUC-ROC, sensitivity, specificity, and balanced accuracy, rather than relying on a single performance measure.

To prevent information leakage, the reference scenarios were partitioned before any data-dependent operation. For each cross-validation partition, the synthetic generator and classifier were fitted exclusively using the reference training fold, while the corresponding held-out reference fold remained isolated for final evaluation. All reported results therefore represent internal computational validation against expert-defined scenarios and should not be interpreted as evidence of clinical performance or population-level representativeness.

A. Classifiers

Evaluating multiple classifiers provides a comprehensive assessment of the quality of the generated synthetic data from different learning perspectives. If the synthetic datasets adequately preserve the statistical characteristics and predictive relationships present in the original dataset, consistent classification performance is expected across different learning algorithms. Conversely, substantial variations in predictive performance may indicate that important structural properties of the original data have not been preserved during the synthetic data generation process.

B. Selection of Metrics

The predictive performance of the classifiers was evaluated using standard metrics for binary classification derived from the confusion matrix, namely True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [14]. The following performance measures were considered:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{M}}. \tag{3}$$

where $M = (TP+FP)(TP+FN)(TN+FP)(TN+FN)$. Confusion matrices were also analyzed to provide a detailed assessment of the specific classification errors.

In addition, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to evaluate the discriminative ability of each classifier independently of the decision threshold [15].

Selecting the most appropriate evaluation metrics for classification tasks is often challenging, as each metric presents distinct advantages and limitations. In many scenarios, metrics behave complementarily, where optimizing one may lead to the degradation of another. This requires a strategic balance or the selection of a primary representative benchmark.

In the evaluation of machine learning models, traditional metrics such as Accuracy and F1-Score are commonly utilized as standard benchmarks for classification performance [23]. However, relying solely on these metrics can inadvertently mask poor classification performance, especially when dealing with highly imbalanced and scarce datasets. To address this limitation, Chicco and Jurman [29] demonstrated that the Matthews Correlation Coefficient (MCC) provides a substantially more robust assessment under such constraints.

In parallel, when translating these computational models into clinical practice, the context of early screening for Autism Spectrum Disorder must be considered. In this domain, primary healthcare objectives demand a strong focus on sensitivity (recall). This priority stems from the fact that facilitating early interventions within the crucial window of high neuroplasticity requires maximizing the identification of potential cases to avoid false negatives. Indeed, clinical gold-standard tools, such as the M-CHAT-R/F [30], are intentionally designed with high sensitivity to capture the maximum number of suspected cases, even at the expense of generating temporary false positives.

Consequently, a hierarchical evaluation strategy is adopted in this study. The MCC is established as our gold-standard metric due to its reliable statistical properties in small and imbalanced sample sizes. Nevertheless, models will also be rigorously evaluated using the AUC-ROC to specifically assess their class separability thresholds. Concurrently, sensitivity will be monitored to guarantee a strict clinical focus on maximizing early screening detection, thereby preventing missed cases prior to specialist referral. Finally, standard Accuracy is utilized strictly as a final baseline verification to provide a conventional measure of overall performance.

C. Methodological Safeguards

The following experimental protocol was applied to all datasets (original, synthetic, and augmented), ensuring a fair comparison among the evaluated data generation strategies.

1) Experimental Rigor and Methodological Safeguards: To ensure reliable performance estimates and robust generalization, all experiments were conducted using repeated stratified k-fold cross-validation [16]. Stratification enforces the original class distribution across all training and testing folds, thereby mitigating sampling bias. To strictly prevent data leakage [5], a rigorous separation between training and evaluation partitions was maintained for all model training and tuning were performed exclusively within the training folds, ensuring no test data information influenced the learning process [18]. Classifier stability was evaluated by computing the mean and standard deviation and confidence intervals of the metrics across all iterations [6]. In synthesis, combining repeated stratified validation, strict data leakage prevention [5], and variance analysis mitigates the risk of biased conclusions. This framework provides a statistically sound foundation to confirm whether the synthetic data reliably preserves relevant features for subsequent clinical modeling.

2) Machine Learning Benchmarking and Synthetic Data Evaluation: The quality of the generated synthetic data was validated through its empirical utility across diverse machine learning algorithms. Beyond verifying statistical similarity to the original sample, this approach assesses whether the synthetic data effectively preserves the underlying clinical patterns defined by the experts.

D. Classifier Selection and Experimental Scenarios

As previously outlined, the validation of the generated synthetic data was conducted by training a diverse suite of machine learning algorithms, which were subsequently evaluated using the metrics detailed in the preceding section (AUC-ROC, F1-Score, Sensitivity, Specificity, MCC, and Accuracy). To provide a comprehensive assessment of synthetic data quality, the models were evaluated under two primary experimental paradigms: Train Synthetic, Test Real (TSTR) and Data Augmentation (Aug).

However, the dataset used in this study does not consist of actual clinical records from real patients. Rather, it comprises data representing hypothetical individuals, constructed based on expert knowledge. Therefore, throughout these experiments, we use the term "reference data" instead of "real data" to denote the original expert-knowledge-based dataset. Accordingly, the TSTR and Augmentation paradigms are adapted to this setting by replacing the term real data with reference data while preserving the underlying experimental protocols.

- **Train Synthetic, Test Reference (TSTR):** In this paradigm, the models are trained exclusively on synthetic data and evaluated on the held-out expert-defined scenarios. These scenarios represent hypothetical individuals (reference) rather than clinical records. This setting evaluates the predictive utility and transferability of the synthetic data within the expert-defined knowledge base.
- **Data Augmentation (AUG):** This paradigm combines the expert-defined reference scenarios with synthetic samples in the training set, thereby increasing its size while keeping model evaluation strictly restricted to held-out hypothetical individuals. The objective is to investigate whether synthetic samples provide complementary information to the expert-defined tabular data and consequently improve the generalization and robustness of the classifiers. Although data augmentation is widely explored in image-based applications, this study examines whether a similar strategy can effectively complement scarce tabular data.

While the TSTR paradigm is instrumental in measuring statistical fidelity, the data augmentation strategy evaluates the practical utility of synthetic data as a sample expansion mechanism, which is especially critical in data-constrained environments. Empirical observations indicate that even when the volume of available real-world samples is limited, their retention within the training set significantly bolsters model stability and reliability, consistently outperforming models trained exclusively on synthetic data [19]. This trend reinforces the vital complementarity between real/reference and synthetic data distributions.

IV. Results

The experiments evaluated two synthetic data generators: Conditional Beta-Bernoulli and Gaussian Copula in combination with classifiers based on different learning mechanisms.

The evaluation employed 10 repetitions of stratified 4-fold cross-validation, resulting in 40 fold-level evaluations for each experimental configuration. A different random seed was used in each repetition to account for variability associated with data partitioning, synthetic sampling, and model initialization. Within each fold, the generator was fitted exclusively to the reference training partition, while the corresponding reference validation partition remained isolated from both generator fitting and classifier training.

The Train on Real, Test on Real (TRTR) paradigm was adapted in this study as a train-test procedure using the reference data. The results obtained under this scenario were used as the internal predictive baseline. This approach was inspired by the conventional TRTR protocol; however, the term "reference" is used here as well because the dataset consists of expert-defined scenarios rather than clinical observations. The baseline measures the ability of each classifier to generalize to held-out scenarios from the expert-defined reference dataset without using synthetic samples. Because the dataset comprises only 40 expert-defined scenarios, the reported means and standard deviations should be interpreted as measures of internal computational performance and variability rather than as evidence of clinical generalization. This constitutes a limitation of the present study.

A. Statistical Fidelity of the Synthetic Data

Figure 1 summarizes the statistical fidelity results obtained for the evaluated generators and synthetic sample sizes. Fidelity was assessed using complementary measures, since no single metric can fully characterize the quality of tabular synthetic data. MAE and correlation RMSE were used to quantify discrepancies in marginal attribute frequencies and dependency structures between the reference and synthetic datasets, while the Classifier Two-Sample Test (C2ST) evaluated their distinguishability under the specified testing protocol.

Across the evaluated range of synthetic sample sizes, the fidelity curves generally exhibited a decrease in error followed by an approximately stable region. This pattern suggests that increasing the number of generated samples improves the stability of the empirical estimates obtained from the learned distributions.

The Gaussian Copula achieved the lowest MAE (Fig. 1.A) and RMSE (Fig. 1.B) values, providing the strongest overall statistical fidelity according to these measures. Nevertheless, the Conditional Beta-Bernoulli generator remained competitive under other criteria, particularly with respect to target-prevalence preservation (Fig. 1.C) and the proportion of synthetic records that were exactly identical to complete records in the corresponding training partition (Fig. 1.D). Therefore, the results do not indicate that one generator was uniformly superior. Instead, they reveal complementary strengths and a trade-off between reproducing the statistical structure of the reference data and avoiding direct reproduction of training records.

The C2ST AUC remained below approximately 0.60 for both generators across the evaluated configurations, indicating that the reference and synthetic datasets were relatively difficult to distinguish under the adopted testing protocol. Image (D) provides a separate descriptive measure of exact overlap by quantifying the proportion of synthetic records that are identical to complete records observed in the corresponding training partition. The lower overlap observed for the Conditional Beta-Bernoulli generator indicates less direct reproduction of the training records in this experiment. Graphic (C) complements this analysis by showing the extent to which each generator preserves the global target prevalence.

In addition to this empirical performance, the Conditional Beta-Bernoulli model has lower structural complexity and greater conceptual simplicity. Based on its competitive statistical fidelity, adequate preservation of target prevalence, lower exact overlap, and simpler formulation, it was selected for the subsequent predictive-utility and augmentation analyses. This selection was made before those analyses and does not imply that the generator is superior across every evaluation dimension. Rather, it reflects the objectives of the subsequent experiments, which focus on predictive transfer and the practical use of synthetic samples for classifier training.

B. Predictive Utility Across Generators and Classifiers

Statistical similarity does not necessarily imply predictive usefulness. Therefore, predictive utility was evaluated through

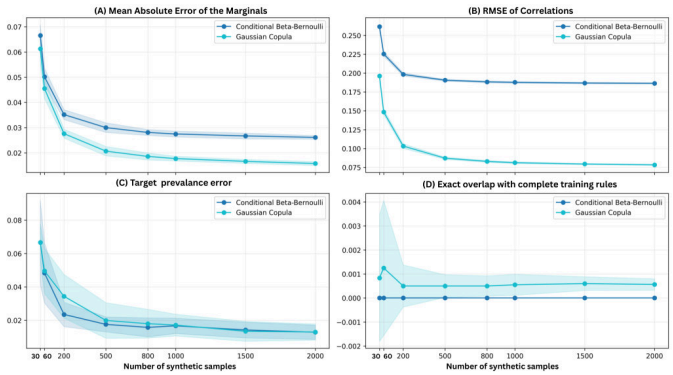


Fig. 1. Statistical fidelity and distinguishability of the synthetic datasets across generators and sample sizes. Lower marginal and dependency errors indicate greater fidelity

the Train on Synthetic, Test on Reference (TSTR) protocol. In this scenario, each classifier was trained exclusively on samples produced by a generator fitted to the reference training fold and evaluated on the corresponding held-out reference fold.

Figure 2 presents the mean MCC obtained for each combination of generator Conditional Beta–Bernoulli, classifier, and synthetic sample size. The results reveal that predictive transfer depends on the interaction between the generator and classifier. Moreover, the heatmap in Figure 2 also indicates that increasing the number of synthetic samples leads to monotonic improvements for every classifier, except for SVM model.

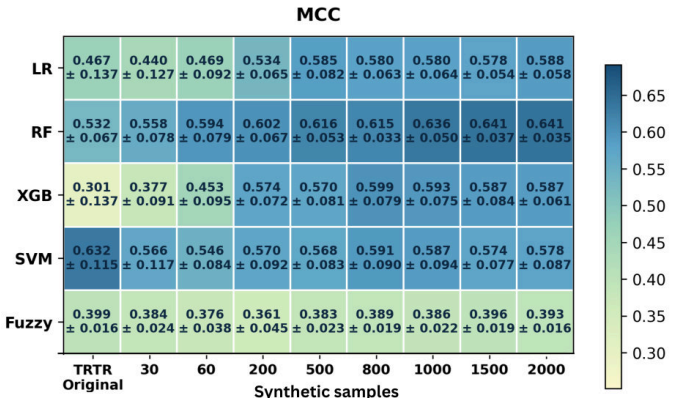


Fig. 2. Mean MCC under TSTR using the Conditional Beta-Bernoulli generator across classifiers. TRTR denotes the reference-only internal baseline

C. Utility Analysis at the Selected Operating Point

Analyzing Figure 1, we chose to conduct the remaining analysis using 800 Conditional Beta-Bernoulli samples. This operating point was selected because the curves in Figure 1 indicate that, beyond 800 samples, the performance approaches a stable region across all graphs. In addition, the results presented in Figure 2 using 800 samples are similar to or close to the best results achieved by each classifier, indicating that this choice provides a reasonable balance between statistical fidelity and utility.

Table II reports the results obtained using 800 Conditional Beta–Bernoulli samples for sensitivity, specificity, F1-score, and AUC-ROC. The comparison includes the TSTR results and the corresponding TRTR baseline for each classifier. Accuracy and MCC are discussed as complementary metrics but are not included in this table.

TABLE II
PERFORMANCE WITH 800 CONDITIONAL BETA–BERNOULLI SAMPLES.

Model	Sens.	Spec.	F1	AUC-ROC
Conditional Beta-Bernoulli - TSTR				
LR	0.796 ± 0.051	0.808 ± 0.040	0.842 ± 0.033	0.836 ± 0.020
RF	0.814 ± 0.028	0.833 ± 0.000	0.858 ± 0.021	0.818 ± 0.022
XGB	0.807 ± 0.042	0.800 ± 0.043	0.846 ± 0.032	0.829 ± 0.033
SVM	0.793 ± 0.058	0.800 ± 0.043	0.837 ± 0.036	0.848 ± 0.021
Expert-defined reference data – TRTR				
LR	0.718 ± 0.052	0.767 ± 0.102	0.782 ± 0.043	0.762 ± 0.042
RF	0.882 ± 0.041	0.600 ± 0.053	0.856 ± 0.025	0.814 ± 0.031
XGB	0.768 ± 0.048	0.533 ± 0.105	0.776 ± 0.042	0.715 ± 0.046
SVM	0.882 ± 0.048	0.733 ± 0.086	0.882 ± 0.033	0.846 ± 0.019

The results in Table II show that predictive performance varied according to both the classifier and the evaluation metric. Among the metrics reported in the table, Random Forest achieved the highest sensitivity, specificity, and F1-score under TSTR. Complementary results, not shown in the table, indicate that Random Forest also achieved the highest accuracy (0.820 ± 0.020) and MCC (0.636 ± 0.027). Therefore, no classifier was uniformly superior across all criteria. RF provided the strongest threshold-dependent performance according to MCC, while SVM showed the best ranking-based discrimination according to AUC-ROC.

Overall, the comparison across metrics indicates that the predictive utility of synthetic data was classifier-dependent. Logistic Regression and XGBoost showed the most consistent improvements over the TRTR baselines, whereas Random Forest achieved stronger threshold-dependent performance according to specificity, F1-score, accuracy, and MCC, despite a reduction in sensitivity. SVM maintained similar ranking-based discrimination according to AUC-ROC but showed weaker threshold-dependent performance. These results highlight the importance of evaluating synthetic data using complementary metrics, since no classifier consistently dominated across all evaluation criteria.

Instead, the results suggest that the Conditional Beta-Bernoulli generator preserved predictive relationships that could be transferred to held-out scenarios within the expert-defined knowledge base, particularly when used with Logistic Regression, Random Forest, and XGBoost.

D. Augmentation Analysis

The augmentation experiment evaluated whether adding synthetic samples to the reference training fold improved predictive performance on the corresponding held-out reference fold. Three training strategies were compared: reference-only training (TRTR), synthetic-only training (TSTR), and augmented training (AUG). In all three scenarios, evaluation was performed on expert-defined reference scenarios that were not used to fit either the generator or the classifier.

For AUG, 30 synthetic samples were added to each reference training fold. This quantity was selected because it approximately matches the number of reference scenarios available for training in each fold, resulting in a nearly balanced contribution from reference and synthetic data. Maintaining the same total training size across TRTR and AUG would require discarding part of the already limited reference training data. Therefore, the adopted experiment evaluates the practical effect of adding synthetic observations rather than isolating the effect of data source under a strictly size-matched design. A controlled analysis of different reference-to-synthetic proportions is left for future work.

Random Forest was selected for this analysis because it showed favorable predictive utility, particularly in terms of MCC and accuracy, in the preceding TSTR experiments. Figure 3 presents bar plots of F1-score, sensitivity, and MCC for the three training strategies. Each bar plot summarizes the distribution of results obtained across the 10 repetitions of stratified 4-fold cross-validation, allowing the comparison of median performance, dispersion, and the occurrence of unstable results.

The distributions show that augmentation slightly improves the MCC while not affecting the F1-score or sensitivity values. As can be observed, AUG remained close to the reference-only strategies across several measures. In particular, the results obtained for MCC are consistent with the findings reported in [2], [19], which indicate that the use of augmentation tends to outperform models trained exclusively on synthetic data.

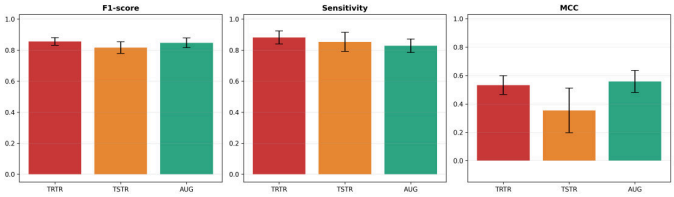


Fig. 3. F1-score, sensitivity, and MCC under reference-only (TRTR), synthetic-only (TSTR), and augmented (AUG) training. AUG combines 30 reference scenarios with 30 synthetic samples and is evaluated on the corresponding held-out reference fold.

V. CONCLUSION

This paper investigates whether synthetic data generation methods applied to tabular data derived from expert knowledge can be useful for identifying children aged 0 to 6 years who may be at risk of ASD. Our hypothesis is that, if the proposed approach to deal with problems with scarce datasets demonstrates consistent statistical fidelity and predictive utility within the expert-defined reference dataset, it may provide a methodological foundation for future studies using real clinical data.

Because classification performance is strongly dependent on training data quality, high-fidelity synthetic datasets would be expected to yield classification outcomes comparable to those obtained using real-world data. The combined results reveal a trade-off between statistical fidelity and predictive

utility. Gaussian Copula most closely reproduced the statistical properties of the reference data, whereas Conditional Beta–Bernoulli produced the strongest predictive performance across the evaluated classifiers. At the classifier level, RF achieved the highest MCC at the selected operating point of 800 synthetic samples. The results support the computational potential of expert-derived synthetic data for reproducing and transferring patterns within the reference knowledge base, but they do not establish clinical validity or population representativeness.

For future work, several directions are planned to extend this research. First, the expert-defined reference dataset should be further assessed by incorporating the perspectives of additional specialists. An initial consultation process involving three other experts has already been initiated to discuss the main clinical attributes, decision rules, and scenario formulations. Based on this process, we intend to develop a structured and accessible validation form that can support systematic expert review, facilitate the acquisition of additional datasets, and improve the consistency of future knowledge elicitation. A second point is the evaluation of the proposed framework using independent datasets containing observations from clinically assessed children. Such datasets would enable a more realistic assessment of external generalization, population representativeness, and potential clinical applicability. The current results should therefore be interpreted as an initial internal validation within the expert-defined knowledge base, rather than as evidence of clinical performance.

Future studies should also investigate the effect of hyperparameter optimization for the machine-learning classifiers, decision-threshold calibration, and the adjustment of the weights used in the fuzzy rule-based models. These procedures should be conducted exclusively within the training partitions, preferably using nested validation or a separate development set, in order to avoid optimistic estimates and preserve the leakage-aware design of the current protocol. Additional analyses may also examine different reference-to-synthetic proportions, sample sizes, and augmentation strategies.

REFERENCES

[1] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: A literature review," *Journal of Big Data*, vol. 10, no. 1, p. 115, 2023. doi: 10.1186/s40537-023-00792-7.

[2] D.-H. Won, K.-S. Shin, e S. Youm, "Synthetic Data Augmentation for Imbalanced Tabular Data: A Comparative Study of Generation Methods," *Applied Sciences*, vol. 14, no. 1, p. 371, 2024.

[3] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA, USA: American Psychiatric Publishing, 2013.

[4] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-VanderWeele, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, Aug. 2018.

[5] B. van Breugel, Z. Qian, and M. van der Schaar, "Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data," *arXiv preprint arXiv:2307.12154*, 2023.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[7] F. Thabtah, "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward," *Informatics for Health and Social Care*, vol. 44, no. 3, pp. 278–297, 2019.

[8] F. Thabtah, "Autistic Spectrum Disorder Screening Data for Children," UCI Machine Learning Repository, 2017. doi: 10.24432/C5659W.

[9] J. Giarratano and G. Riley, *Expert Systems: Principles and Programming*, 4th ed. Boston, MA, USA: Thomson Course Technology, 2005.

[10] S. A. Kalam, T. R. Prome, e M. A. Ullah, "Machine Learning Approach for Identification of Autism Spectrum Disorder from Video Using OpenPose," em *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pp. 1720–1725, 2024.

[11] A. Goncalves, S. Ray, J. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Medical Research Methodology*, vol. 20, no. 1, Art. no. 108, 2020.

[12] L. J. Emrich and M. R. Piedmonte, "A method for generating high-dimensional multivariate binary variates," *The American Statistician*, vol. 45, no. 4, pp. 302–304, 1991.

[13] C. Allison, S. Baron-Cohen, S. Wheelwright, T. Charman, J. Richler, G. Pasco, and C. Brayne, "The Q-CHAT (Quantitative Checklist for Autism in Toddlers): A normally distributed quantitative measure of autistic traits at 18–24 months of age: Preliminary report," *Journal of Autism and Developmental Disorders*, vol. 38, no. 8, pp. 1414–1425, 2008.

[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[15] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.

[17] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410. doi: 10.1109/DSAA.2016.49.

[18] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, Art. no. 15, 2012.

[19] N. Jaipuria, X. Zhang, R. Bhasin, M. Arafa, P. Chakravarty, S. Shrivastava, S. Manglani, and V. N. Murali, "Deflating Dataset Bias Using Synthetic Data Augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Palo Alto, CA, USA, 2020, pp. 772–781.

[20] L. Zwaigenbaum, J. A. Bauman, R. Choueiri, et al., "Early Identification and Interventions for Autism Spectrum Disorder: Executive Summary," *Pediatrics*, vol. 136, no. Supplement_1, pp. S1–S9, 2015. doi: 10.1542/peds.2015-2851B.

[21] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*. Washington, DC, USA: American Psychiatric Association, 2022.

[22] M. Sahin, "Applications of Generalized Linear Models in Complex Data Analysis," *Journal of Statistical Software*, vol. 110, no. 3, pp. 45–58, 2024.

[23] S. Sperandei, "Understanding Logistic Regression Analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014. doi: 10.11613/BM.2014.003.

[24] C. R. Bilder and T. M. Loughin, *Analysis of Categorical Data with R*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2023.

[25] Machine Learning Hub, "Log-Linear Models for Contingency Tables in Health Sciences," *ML Hub Technical Reports*, 2017.

[26] R. E. Foraker, M. M. S. Refai, and J. M. Michel, "The role of synthetic data in clinical research and healthcare artificial intelligence," *Annual Review of Biomedical Data Science*, vol. 4, pp. 243–261, 2021.

[27] A. Tucker, Z. Wang, Y. Roudsari, and P. Culliford, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *Journal of Biomedical Informatics*, vol. 106, p. 103439, 2020. doi: 10.1016/j.jbi.2020.103439.

[28] HYMAN, S. L.; LEVY, S. E.; MYERS, S. M. et al. Identification, Evaluation, and Management of Children With Autism Spectrum Disorder. **Pediatrics**, v. 145, n. 1, p. e20193447, jan. 2020. DOI: 10.1542/peds.2019-3447. PMID: 31843864.

[29] Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1), 4. doi: 10.1186/s13040-023-00322-4.

[30] Robins, D. L., Fein, D., & Barton, M. L. (2009). *The Modified Checklist for Autism in Toddlers, Revised with Follow-Up (M-CHAT-R/F)*. Available at: https://www.mchatscreen.com.

Diagnostic Accuracy of Prostate-Specific Antigen as an Indicator of Male Sexual Contact in Living Victims of Sexual Violence

Brenda Georgetti Gasparetto
Forensic Molecular Genetics Section
Scientific Police of Paraná
Curitiba, Brazil
rt.brenda.gasparetto@policiacientifica.pr.gov.br

Gustavo Benvenutti Borba
Postgraduate Program in Biomedical Engineering
UTFPR
Curitiba, Brazil
gustavobborba@utfpr.edu.br

Eduardo Borba Neves
Postgraduate Program in Biomedical Engineering
UTFPR
Curitiba, Brazil
neveseb@gmail.com

Juliane Carlotto
Forensic Molecular Genetics Section
Scientific Police of Paraná
Curitiba, Brazil
juliane.carlotto@policiacientifica.pr.gov.br

***Abstract*—Prostate-specific antigen (PSA) is widely used in forensic laboratories as a presumptive biochemical marker of seminal fluid, but no consensus exists regarding the optimal cutoff value, and the statistical robustness of PSA concentration bands as predictors of genotyping success has not been formally tested. This study evaluated the diagnostic accuracy of PSA as an indicator of male sexual contact in 160 intimate samples from living victims of sexual violence processed at the Forensic Molecular Genetics Section of the Scientific Police of Paraná, Brazil (2023–2024), and assessed, through chi-square tests of independence with Bonferroni-corrected pairwise comparisons, which PSA concentration bands differ significantly from one another in male DNA detection and STR genotyping outcomes. Sensitivity, specificity, PPV, and NPV were calculated at three cutoffs (0.08, 1.0, and 4.0 ng/mL), yielding an estimated ROC AUC of 0.89. The chi-square test of independence confirmed a significant association between PSA band and male DNA detection (χ^2=70.78, df=5, p=7.06×10^{-14}) and between PSA band and STR genotyping outcome (χ^2=43.58, df=10, p=3.91×10^{-6}). However, Bonferroni-corrected pairwise comparisons revealed that statistically significant differences in male DNA detection were confined to comparisons between the lowest band (<0.07 ng/mL) and every other band; no significant differences were found among bands above 0.08 ng/mL. This indicates that, given the present sample size and the conservative Bonferroni correction, the apparent dose-response gradient observed descriptively across intermediate bands (0.08–0.99, 1–4, 4–10, 10–99 ng/mL) did not reach statistical significance, a pattern compatible with a Type II error arising from the conservative Bonferroni correction and the limited size of the two smallest strata (n=13 and n=14), rather than definitive proof of the absence of a true biological gradient, and that the only comparison robust to correction is that between PSA below versus above approximately 0.07–1.0 ng/mL. These findings refine the interpretation of PSA as a forensic triage parameter: rather than supporting a continuous dose-response relationship, the data support a simplified two-tier stratification (low vs. moderate-to-high PSA) for screening purposes in living victims.**

Keywords*—*prostate-specific antigen, forensic genetics, sexual violence, STR genotyping, ROC curve

I. Introduction

The investigation of sexual violence crimes depends fundamentally on the analysis of biological evidence collected from victims and crime scenes. In Brazil, official public security statistics confirm the scale of the problem: 87,545 rape and statutory rape cases were registered in 2024 alone, the highest figure in the historical series, averaging one reported case every six minutes nationwide [1]. This volume, combined with documented underreporting, makes forensic evidence even more central to the production of legal proof. In this context, forensic genetics has consolidated itself as a strategic field, with continuous advances in laboratory methodology and infrastructure across the country.

Unlike post-mortem casework, evidence collection in living victims is constrained by a narrow biological window: seminal biomarkers such as PSA degrade progressively after the assault, and victims frequently delay seeking forensic examination due to psychological distress, fear of revictimization, or limited access to specialized care, particularly outside large urban centers. This operational reality means that, in practice, forensic laboratories routinely receive intimate samples collected at variable and often suboptimal intervals after the alleged assault, with PSA concentrations that may already be in the lower analytical range by the time testing occurs. Consequently, establishing which PSA concentration bands carry genuinely distinct diagnostic information, rather than relying on cutoffs validated primarily in post-mortem or controlled experimental settings, has direct relevance for how laboratories triage and interpret evidence from this specific, time-sensitive population.

Prostate-specific antigen (PSA) is the principal biochemical indicator used in the routine of forensic genetics laboratories for the presumptive identification of seminal fluid. PSA is a glycoprotein produced by prostatic epithelial cells, present at high concentration in seminal fluid, and its detection is considered strong presumptive evidence of semen presence [2], [3]. A recent clinical case report reaffirms that PSA testing remains the standard presumptive screening approach in forensic examinations of sexual violence victims,

underscoring its continued centrality in casework despite known limitations [4].

Henky and Arijana [5] demonstrated, in a comparative study using vaginal swabs from sexual violence victims, that the PSA rapid test outperforms sperm microscopy, with sensitivity of 84.6%, specificity of 100%, and overall accuracy of 87.5%, using the SRY gene as the gold standard for confirmation of male material. Sperm microscopy, in the same population, presented sensitivity of only 23.1%, indicating a high risk of false-negative results, particularly in victims who performed genital hygiene after the assault.

A further challenge concerns the absence of consensus on the ideal PSA cutoff for forensic seminal fluid determination. Reference values vary widely across studies and laboratories: the traditional clinical threshold of 4.0 ng/mL, originally established for prostatic disease screening, coexists with more sensitive forensic cutoffs such as 1.0 ng/mL and 0.08 ng/mL, adopted by different laboratories in an attempt to balance sensitivity and specificity. A recent study by Ruff et al. [6], using ROC curve analysis on 460 forensic samples from sex crime investigations, identified an optimal cutoff of 0.085 ng/mL, with an area under the curve (AUC) of 0.848. This lack of standardization highlights the need for studies evaluating different cutoff values in specific populations and laboratory contexts.

Beyond serological screening, quantification of Y-chromosome DNA by qPCR, known as Y-Screening, has been increasingly adopted as a molecular triage step for sexual assault evidence kits, determining which samples proceed to STR genotyping [7], [8]. A 2025 validation study of a Y-screening workflow using a commercial qPCR kit confirmed that this approach reliably predicts which samples are likely to yield informative downstream STR profiles, supporting its use to triage casework and preserve evidentiary material [8]. Despite these individual advances, the literature still lacks studies that systematically integrate PSA concentration, rather than its qualitative positive/negative classification, with molecular quantification and genotyping outcomes; SWGDAM [7] explicitly identifies this gap, calling for research correlating quantification data with the success of genetic typing.

A methodological limitation common to dose-response analyses of this kind is that descriptive trends across multiple concentration bands are frequently presented without formal statistical testing of whether adjacent or non-adjacent bands actually differ from one another. When more than two categories are compared pairwise, the probability of a spurious significant result increases with the number of comparisons performed, a problem addressed by applying a Bonferroni correction to the significance threshold following a significant overall chi-square test of independence [9], [10]. This stepwise approach, an omnibus chi-square test followed by Bonferroni-corrected pairwise 2×2 comparisons, provides a statistically rigorous means of identifying precisely which categories drive an observed association, rather than assuming that all visually apparent differences are equally robust.

The present study addresses both gaps. First, it evaluates, in a cohort restricted to living victims of sexual violence, the diagnostic accuracy of PSA for indicating male sexual contact across three forensic cutoff values, using qPCR-based male DNA detection as the reference standard. Second, and as its central methodological contribution, it applies chi-square tests of independence with Bonferroni-corrected pairwise comparisons to formally test which PSA concentration bands differ significantly in male DNA detection and STR genotyping outcome, determining whether the apparent dose-response gradient reported in prior descriptive analyses withstands rigorous statistical scrutiny or whether it is in fact driven by a single, more limited threshold.

From a biomedical engineering standpoint, this distinction is operationally important: a triage rule based on a continuous dose-response assumption implies a fundamentally different decision architecture than one based on a single robust threshold. By identifying which comparisons survive correction for multiple testing, this study provides a more conservative and defensible empirical basis for the design of computer-assisted triage tools that combine PSA concentration with Y-chromosome quantification in forensic genetics laboratory workflows.

II. Materials and Methods

A. Study Design and Sample

This retrospective study analyzed data from biological traces collected from living female victims of sexual violence, processed at the Forensic Molecular Genetics Section of the Scientific Police of Paraná, Brazil, between 2023 and 2024. The analyzed cohort comprised 160 intimate samples (vaginal, anal, vulvar, and oral swabs) obtained from living victims. Collected variables included PSA concentration (ng/mL), qPCR-based Y-chromosome DNA quantification result, and STR genotyping outcome, classified as viable genetic profile, non-viable profile, or absence of amplification. Records with incomplete data or compromised identification were excluded (Figure 1).

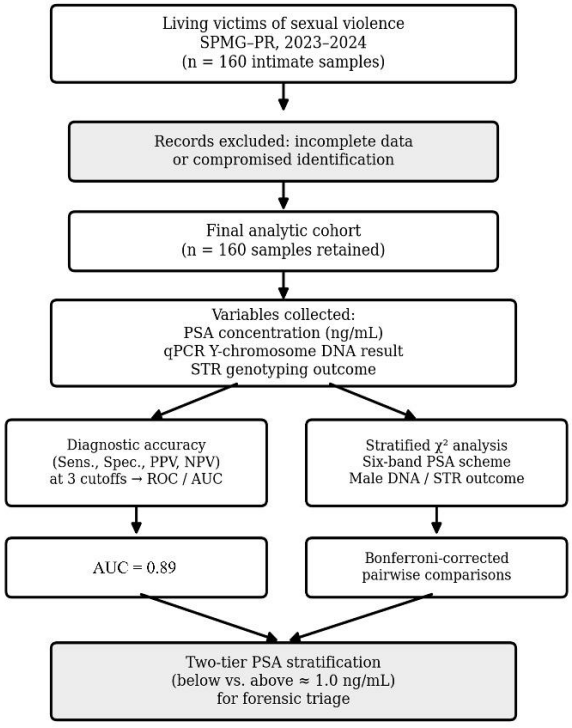


Fig. 1. Study flow diagram, from sample identification through the two analytical branches (diagnostic accuracy and stratified chi-square analysis) to the resulting two-tier PSA stratification proposal.

This study was approved by the Scientific Committee of the Scientific Police of Paraná, in accordance with the ethical principles set forth in the Declaration of Helsinki for research involving human subjects. Data were collected from institutional records of the Forensic Molecular Genetics Section of the Scientific Police of Paraná and used exclusively for research purposes in accordance with institutional guidelines.

B. *Reference Standard*

The gold standard for confirming male sexual contact was the qPCR-based quantification of male DNA. Samples were classified as positive for male DNA presence according to qPCR results, independent of the PSA outcome, allowing PSA performance to be assessed against an objective molecular reference rather than circumstantial case information.

C. *Diagnostic Accuracy Analysis*

Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for three PSA cutoff values commonly cited in forensic and clinical practice: 0.08 ng/mL, 1.0 ng/mL, and 4.0 ng/mL. A receiver operating characteristic (ROC) curve was constructed from the three corresponding sensitivity/specificity pairs, and the area under the curve (AUC) was estimated using the trapezoidal rule.

D. *Stratified Chi-Square Analysis*

Samples were stratified into PSA concentration bands using a six-band scheme (<0.07, 0.08–0.99, 1–4, 4–10, 10–99, and >100 ng/mL), defined a priori from established forensic and clinical cutoffs. A chi-square (χ^2) test of independence was performed to test the global association between PSA band and (i) male DNA detection (positive/negative) and (ii) STR genotyping outcome (viable profile/non-viable profile/no amplification). The selection of these specific ranges is justified by the lack of consensus in the literature regarding the optimal prostate-specific antigen (PSA) cut-off value for confirming the presence of semen, with previous studies reporting various thresholds, including 0.04 [6], 1 [12], 4 [13], and 10 [14] ng/mL.

When the overall χ^2 test was significant, all possible pairwise 2×2 sub-tables were subsequently tested to identify which specific band comparisons drove the global association. Because multiple pairwise comparisons inflate the family-wise Type I error rate, a Bonferroni correction was applied: the conventional significance threshold (α=0.05) was divided by the number of pairwise comparisons performed (15 comparisons for the six-band scheme, yielding a corrected threshold of $p<0.00333$) [9], [10]. All statistical analyses were performed using Python (SciPy, NumPy, Pandas), with visualization in Matplotlib.

The Bonferroni correction was chosen over less conservative alternatives (e.g., Benjamini–Hochberg false discovery rate control) because the present analysis is explicitly confirmatory rather than exploratory: the six concentration bands were defined a priori from established forensic and clinical cutoffs rather than derived from the data, and the goal was to identify only those band contrasts robust enough to support an operational triage threshold, accepting a more conservative family-wise error rate in exchange for higher confidence in any comparison that survived correction. For each 2×2 sub-table, expected cell frequencies were inspected, and Fisher's exact test was substituted for the chi-square approximation whenever any expected count fell below 5, ensuring that significance levels were not inflated by small-sample violations of the chi-square assumptions.

III. Results

A. *Diagnostic Accuracy of PSA in Living Victims*

Among the 160 samples from living victims, diagnostic performance varied considerably across the three evaluated cutoffs. Table I summarizes sensitivity, specificity, PPV, and NPV for each cutoff value.

TABLE I. DIAGNOSTIC ACCURACY OF PSA BY CUTOFF VALUE – LIVING VICTIMS (N = 160)

Cutoff (ng/mL)	Sens. (%)	Spec. (%)	PPV (%)	NPV (%)
0.08	90.5	72.7	89.7	74.4
1.0	75.0	86.4	93.6	56.7
4.0	63.8	88.6	93.7	48.2

The lowest cutoff, 0.08 ng/mL, produced the highest sensitivity (90.5%) at the cost of lower specificity (72.7%), whereas increasing the cutoff to 1.0 and 4.0 ng/mL progressively favored specificity at the expense of sensitivity. The NPV declined sharply as the cutoff increased (from 74.4% at 0.08 ng/mL to 48.2% at 4.0 ng/mL), indicating that higher cutoffs substantially increase the risk of false-negative results in this population. The ROC curve constructed from the three sensitivity/specificity pairs yielded an estimated AUC of 0.89, consistent with good discriminative performance of PSA in living victims.

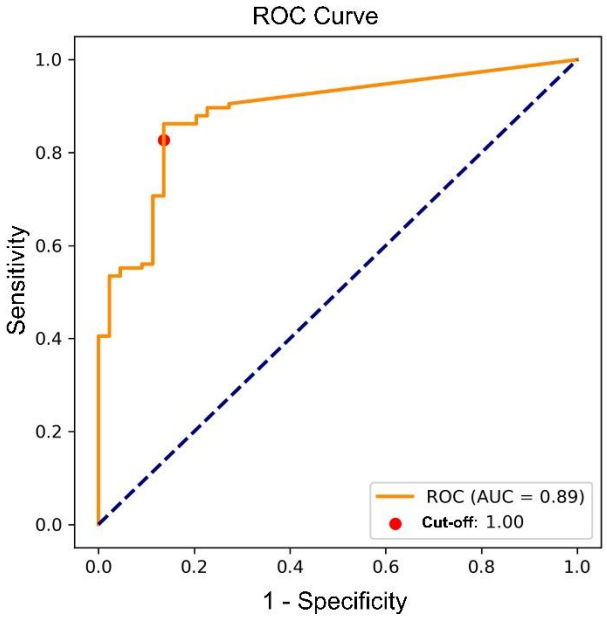


Fig. 2. Receiver operating characteristic (ROC) curve for PSA at the cutoff = 1.0 ng/mL, with male DNA quantification by qPCR as the reference standard. AUC = 0.89.

B. *Descriptive Distribution Across PSA Bands*

Table II presents the distribution of male DNA detection and STR genotyping outcomes across the six PSA concentration bands. Descriptively, the proportion of samples positive for male DNA and the proportion yielding a viable

genetic profile both increased with PSA concentration, ranging from 25.6% and 7.0%, respectively, in the lowest band (<0.07 ng/mL), to 100.0% and 91.7% in the highest band (>100 ng/mL).

TABLE II. MALE DNA DETECTION AND STR GENOTYPING OUTCOME BY PSA CONCENTRATION BAND – LIVING VICTIMS

PSA band (ng/mL)	n	Male DNA+ (%)	Viable profile (%)
< 0.07	43	25.6	7.0
0.08–0.99	24	75.0	33.3
1–4	14	92.9	64.3
4–10	13	76.9	76.9
10–99	30	93.3	76.7
> 100	36	100.0	91.7

C. Chi-Square Test of Independence: Global Association

The chi-square test of independence confirmed a statistically significant global association between PSA band (six-band scheme) and male DNA detection (χ^2=70.78, df=5, $p=7.06\times10^{-14}$), as well as between PSA band and STR genotyping outcome (χ^2=43.58, df=10, $p=3.91\times10^{-6}$). These results confirm that PSA concentration band is not independent of either outcome, formally supporting the descriptive pattern observed in Table II.

D. Bonferroni-Corrected Pairwise Comparisons

A significant omnibus χ^2 result indicates that an association exists somewhere among the bands, but does not indicate which specific bands differ from one another [9], [10]. Pairwise 2×2 chi-square tests were therefore performed for every band combination within the six-band scheme, with significance assessed against the Bonferroni-corrected threshold of p<0.00333 (15 comparisons).

For male DNA detection, only the five comparisons involving the lowest band (<0.07 ng/mL) against every other band reached statistical significance after correction (Table III). None of the ten remaining pairwise comparisons among bands at or above 0.08 ng/mL, including comparisons such as 0.08–0.99 vs. 1–4 ng/mL, 1–4 vs. 4–10 ng/mL, or 10–99 vs. >100 ng/mL, achieved statistical significance.

TABLE III. BONFERRONI-CORRECTED PAIRWISE COMPARISONS FOR MALE DNA DETECTION ACROSS THE SIX PSA BANDS

Comparison	χ^2	p-value	Sig.[a]
<0.07 vs 0.08–0.99	13.38	2.55×10^{-4}	Yes
<0.07 vs 1–4	16.95	3.84×10^{-5}	Yes
<0.07 vs 4–10	9.14	2.50×10^{-3}	Yes
<0.07 vs 10–99	29.93	4.47×10^{-8}	Yes
<0.07 vs >100	42.00	9.15×10^{-11}	Yes
0.08–0.99 vs 1–4	0.88	0.349	No
0.08–0.99 vs 4–10	0.00	1.000	No
0.08–0.99 vs 10–99	2.25	0.134	No
0.08–0.99 vs >100	7.42	0.006	No
1–4 vs 4–10	0.39	0.534	No
1–4 vs 10–99	0.00	1.000	No
1–4 vs >100	0.25	0.621	No
4–10 vs 10–99	1.05	0.306	No
4–10 vs >100	5.29	0.021	No
10–99 vs >100	0.73	0.394	No

[a] Significant at Bonferroni-corrected threshold p<0.00333.

A similar pattern was observed for STR genotyping outcome, although the significant threshold (p<0.00111, for 45 comparisons across the three outcome categories) was considerably more conservative due to the larger number of three-way pairwise comparisons. Only two comparisons remained significant: viable profile vs. no amplification between the <0.07 and >100 ng/mL bands (χ^2=11.88, $p=5.66\times10^{-4}$), and viable profile vs. non-viable profile between the 0.08–0.99 and >100 ng/mL bands (χ^2=12.59, $p=3.87\times10^{-4}$). All other pairwise comparisons among intermediate bands, including several that could not be calculated due to insufficient cell frequencies, did not reach significance.

IV. DISCUSSION

The 0.08 ng/mL cutoff achieved the best balance for screening purposes in this cohort, maximizing sensitivity (90.5%) and thereby minimizing the risk of overlooking male DNA presence. This finding aligns with Ruff et al. [6], who identified a comparably low optimal cutoff (0.085 ng/mL) with similar AUC (0.848 vs. 0.89 in the present study) in an independent sample of 460 forensic specimens. The progressive loss of NPV at higher cutoffs (74.4% at 0.08 ng/mL versus 48.2% at 4.0 ng/mL) reinforces that adopting the traditional clinical threshold of 4.0 ng/mL in a forensic screening context would leave nearly half of PSA-negative samples still harboring detectable male DNA.

The central contribution of this study lies in the stratified chi-square analysis, which qualifies how the dose-response pattern in Table II should be interpreted. While the descriptive proportions suggest a smooth, continuous increase in both male DNA positivity and genotyping success as PSA concentration rises, the Bonferroni-corrected pairwise comparisons demonstrate that this is not statistically the case. The only band comparisons that reliably distinguish themselves are those contrasting the lowest band (PSA below approximately 0.07–1.0 ng/mL) against any higher band; comparisons among bands at or above roughly 1.0 ng/mL do not differ significantly from one another, despite visually appearing to form a gradient. The selection of these specific ranges is justified by the lack of consensus in the literature regarding the optimal prostate-specific antigen (PSA) cut-off value for confirming the presence of semen, with previous studies reporting various thresholds, including 0.04, 1, 4, and 10 ng/mL.

This result has a direct practical implication: rather than treating the six PSA bands as a fine-grained continuous predictor, the data support a simplified two-tier stratification for forensic triage purposes, samples below approximately 1.0 ng/mL versus samples above 4.0 ng/mL, as the statistically defensible operating distinction, with the intermediate range (1.0–4.0 ng/mL) representing a zone in which the present data cannot statistically distinguish a true diagnostic plateau from a gradient obscured by limited statistical power and a consequent risk of Type II error, particularly given the n=13 and n=14 strata and the conservative Bonferroni threshold, rather than necessarily reflecting a smoothly graded transition. This refinement is consistent with general statistical guidance recommending that descriptive trends across ordered categories be confirmed by formal pairwise testing with correction for multiple comparisons before being interpreted as a true dose-response relationship [9], [10].

From a forensic genetics workflow standpoint, this finding also has implications for triage protocol design. A previous validation study of Y-chromosome qPCR screening as a predictor of downstream STR success demonstrated that quantification-based triage can reliably flag samples likely to yield informative profiles [8]. The present results suggest that any analogous triage rule built on PSA concentration should be calibrated around the single robust threshold identified here (approximately 1.0 ng/mL), rather than assuming that finer PSA gradations carry proportionally finer-grained predictive information, since the statistical evidence for differentiating among bands above this threshold is not robust given the present sample size.

The pattern observed here is also consistent with broader comparative literature on PSA rapid testing. Henky and Arijana [5] reported that PSA testing substantially outperformed sperm microscopy in vaginal swabs from living victims, with sensitivity of 84.6% against an SRY-based gold standard, a figure close to the 90.5% sensitivity obtained at the 0.08 ng/mL cutoff in the present cohort, despite differences in reference standard and population. Taken together with the comparably low optimal cutoff reported by Ruff et al. [6] using ROC analysis in an independent sample of 460 forensic specimens, this convergence across three distinct cohorts and laboratories strengthens the case that a forensic PSA cutoff considerably below the traditional clinical threshold of 4.0 ng/mL is broadly appropriate for sexual assault screening in living victims, rather than being an artifact specific to a single laboratory population.

This analysis was restricted to living victims and does not address the distinct dynamics of PSA in post-mortem or clothing/object samples, where degradation and contamination patterns differ substantially [11]. The relatively small number of samples in some PSA bands (n=13–14 in two strata) limits the statistical power of pairwise comparisons involving those bands, and several three-way STR outcome comparisons could not be calculated due to insufficient cell frequencies, a limitation that should be addressed through replication in larger or pooled cohorts. The chi-square approach used here treats PSA concentration as categorical; future work should also model PSA as a continuous predictor (e.g., via continuous-cutoff ROC analysis or logistic regression) to complement the categorical stratification presented here.

Finally, the lack of standardized time-since-assault data prevented controlling for PSA degradation, which may confound variability within intermediate bands. Additionally, while our cohort (n = 160) is comparable to previous studies, it is smaller than larger series (e.g., Ruff et al. [6]); thus, external validation in independent Brazilian laboratories is warranted to ensure the generalizability of the proposed two-tier stratification before routine clinical implementation.

V. Conclusion

In this cohort of living victims of sexual violence, PSA demonstrated good diagnostic accuracy as a presumptive indicator of male sexual contact (AUC=0.89), with the 0.08 ng/mL cutoff offering the most favorable sensitivity-specificity balance. Chi-square tests of independence confirmed a significant global association between PSA concentration band and both male DNA detection and STR genotyping outcome; however, Bonferroni-corrected pairwise comparisons demonstrated that, under the present sample size, this association is statistically attributable almost entirely to the contrast between the lowest PSA band and all higher bands. Given the conservative correction applied and the limited size of two intermediate strata (n=13–14), this finding should be interpreted as an absence of demonstrated statistical significance for a finer gradient, potentially reflecting a Type II error, rather than definitive evidence against one. These findings support a simplified two-tier PSA stratification (below versus above approximately 1.0 ng/mL) as the statistically robust basis for forensic triage decisions in living victims, while cautioning against over-interpreting finer-grained descriptive trends that do not survive correction for multiple comparisons. Future work should extend this analysis to continuous-cutoff ROC and logistic regression frameworks and explore the joint predictive value of PSA and Y-chromosome quantification for STR genotyping success.

References

[1] Fórum Brasileiro de Segurança Pública, 19º Anuário Brasileiro de Segurança Pública. São Paulo: FBSP, 2025.

[2] K. Virkler and I. K. Lednev, “Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene,” Forensic Science International, vol. 188, no. 1-3, pp. 1–17, 2009.

[3] G. F. Sensabaugh, “Isolation and characterization of a semen-specific protein from human seminal plasma: a potential new marker for semen identification,” Journal of Forensic Sciences, vol. 23, no. 1, pp. 106–115, 1978.

[4] A. Stasiūnas et al., “The significance of prostate-specific antigen in forensic examinations of sexual violence: a case report,” Medicine, Science and the Law, 2025, doi: 10.1177/00258024251330100.

[5] Henky and I. G. K. N. Arijana, “Prostate-Specific Antigen (PSA) rapid diagnostic tests compared with SRY gene for detecting male component in vaginal swabs,” Indonesian Journal of Legal and Forensic Sciences, vol. 2, no. 2, pp. 37–41, 2012.

[6] C. Ruff et al., “PSA and ROC curve: validation and cutoff for forensic sexual assault cases through sequential testing,” Journal of Forensic Sciences, 2026, ahead of print, doi: 10.1111/1556-4029.70268.

[7] Scientific Working Group on DNA Analysis Methods (SWGDAM), Report on Y-Screening of Sexual Assault Evidence Kits (SAEKs). SWGDAM, 2020, 24 p.

[8] R. Bonsu et al., “Investigation of a Y-chromosome screening approach for sexual assault evidence using the QIAGEN Investigator Casework GO! kit,” Journal of Forensic Sciences, vol. 70, pp. 1133–1148, 2025, doi: 10.1111/1556-4029.70027.

[9] S. Bahariniya et al., “Tutorial on Bonferroni correction as a post hoc analysis of a significant chi-squared test: a methodological guide,” Journal of Food Quality and Hazards Control, 2025.

[10] J. H. McDonald, “Handbook of Biological Statistics”, 3rd ed. Baltimore: Sparky House Publishing, 2014.

[11] P. Lunetta and H. Sippel, “Positive prostate-specific antigen (PSA) reaction in post-mortem rectal swabs: a cautionary note,” Journal of Forensic and Legal Medicine, vol. 16, no. 7, pp. 397–399, 2009, doi: 10.1016/j.jflm.2009.04.002.

[12] M. M. Hobbs, M. J. Steiner, K. D. Rich, M. F. Gallo, L. Warner, and M. Macaluso, “Vaginal swab specimen processing methods influence performance of rapid semen detection tests: a cautionary tale,” Contraception, vol. 82, no. 3, pp. 291–295, 2010, doi: 10.1016/j.contraception.2010.02.022.

[13] B. O. Gartside, K. J. Brewer, and C. L. Strong, “Estimation of prostate-specific antigen (PSA) extraction efficiency from forensic samples using the Seratec PSA Semiquant semiquantitative membrane test,” Forensic Science Communications, vol. 5, no. 2, 2003.

[14] B. Levine, J. M. Titus, K. Moore, and D. Fowler, “Use of prostate specific antigen in the identification of semen in postmortem cases,” American Journal of Forensic Medicine and Pathology, vol. 25, no. 4, pp. 288–290, 2004, doi: 10.1097/01.paf.0000136640.06642.f8.

Effects of Transcutaneous Spinal Direct Current Stimulation (tsDCS) in Patients with Major Depressive Disorder: A Case Comparison

1st Wíviny Rodrigues de Souza Martins
Graduate Program in Health Sciences (PPGCSA).
Federal University of Rio Grande do Norte (UFRN)
Natal, Brazil
0009-0001-4096-159X

2nd Gabriel Gonçalves Moreira
Technical High School Program in Informatics
Federal Institute of Rio Grande do Norte (IFRN)
Parnamirim, Brazil
0009-0008-2347-4619

3rd Alexia Naftally Silva Morais
Graduate Program in Neuroengineering.
Santos Dumont Institute (ISD)
Macaíba, Brazil
0009-0008-6264-2048

4rd Jales Clemente
Psychiatry practice
Natal, Brazil
0000-0002-1702-1751

5th Rodrigo Pegado
Graduate Program in Health Sciences (PPGCSA).
Federal University of Rio Grande do Norte (UFRN)
Natal, Brazil
0000-0002-7227-1075

6th Hougelle Simplício Gomes Pereira
Graduate Program in Neuroengineering.
Santos Dumont Institute (ISD)
Macaíba, Brazil
0000-0001-8794-9769

7th Fabrício Lima Brasil
Graduate Program in Neuroengineering.
Santos Dumont Institute (ISD)
Macaíba, Brazil
0000-0002-9984-1007

***Abstract*—Transcutaneous spinal direct current stimulation (tsDCS) has emerged as a promising non-invasive neuromodulation technique for Major Depressive Disorder (MDD). This study investigated the effects of tsDCS on depressive symptoms, quality of life, and suicidal ideation in MDD, evaluating therapeutic persistence. In a longitudinal case comparison, two female MDD patients completed 15 sessions of tsDCS (2.5 mA, 20 min/session) over three weeks. Standardized clinical assessments (HAM-D, BDI-II, WHOQOL-BREF, SIQ) were performed at baseline, mid-treatment, post-treatment, and at a 90-day follow-up. Both participants showed marked reductions in depressive symptoms and suicidal ideation, accompanied by improvements in psychological quality of life after treatment. At follow-up, one participant maintained the clinical benefits, whereas the other experienced symptom recurrence associated with antidepressant tapering and psychosocial stressors. The intervention was well tolerated, with only mild transient adverse effects. These findings suggest that tsDCS may represent a safe and promising adjunctive treatment for MDD, although larger controlled studies are needed to confirm its long-term effectiveness.**

***Index Terms*—Major depressive disorder (MDD), transcutaneous spinal direct current stimulation (tsDCS), non-invasive neuromodulation, depression, quality of life, suicidal ideation.**

I. Introduction

Major Depressive Disorder (MDD) is characterized by the persistent presence of depressed mood, diminished interest or pleasure in previously valued activities, and cognitive, behavioral, and neurovegetative changes that significantly impair social, occupational, and personal functioning, with symptom severity varying across individuals [1]. Although conventional treatment for MDD, consisting of selective serotonin reuptake inhibitors combined with psychotherapy, is effective for many patients, a substantial proportion fail to achieve a satisfactory therapeutic response or experience challenges related to treatment adherence, symptom recurrence, and adverse effects [2]. These limitations have driven the search for complementary and innovative therapeutic strategies.

Non-invasive neuromodulation techniques have emerged as promising therapeutic approaches for MDD, demonstrating favorable outcomes in reducing depressive symptoms. Among the available modalities, transcutaneous spinal direct current stimulation (tsDCS) has gained attention as a novel intervention. This technique involves the application of low-intensity direct current over specific segments of the spinal column, thereby modulating the activity of central neural circuits [3]. Initially investigated for conditions such as chronic and neuropathic pain, stroke, spinal cord injury, and Alzheimer's disease, tsDCS has shown benefits associated with improvements in quality of life, mood, and depressive symptoms [4]–[7].

Recent evidence suggests that anodal tsDCS applied over the T10 thoracic spinal segment may modulate interoceptive pathways involved in the integration of bodily states and emotional processing, thereby influencing brain structures frequently implicated in the pathophysiology of MDD, such as the thalamus and the insular cortex [8]. Furthermore, tsDCS has demonstrated a favorable safety and tolerability profile, making it a potentially accessible, and easy-to-administer intervention. In light of these findings, tsDCS represents a promising therapeutic approach for MDD, underscoring the importance of investigating its clinical effects in individuals with the disorder.

II. Objective

The present study aimed to investigate the effects of transcutaneous spinal direct current stimulation (tsDCS) on depressive symptoms, psychological quality of life, and suicidal ideation in two patients with major depressive disorder (MDD). Additionally, the study evaluated the persistence of these effects at follow-up and the safety and tolerability of the intervention.

III. Methodology

A. Participants

This longitudinal comparative case series was based on a subset of participants from a larger clinical study investigating the effects of tsDCS in patients with MDD. The present case series included two female participants diagnosed with MDD who underwent a protocol of tsDCS. The selection process took place after recruitment, during which the participants underwent a clinical assessment conducted by a psychiatrist from the research team to confirm all eligibility criteria. Participants were eligible if they were aged 18–65 years, met the diagnostic criteria for MDD according to the DSM-5, had a minimum score of 18 on the Hamilton Depression Rating Scale (HAM-D), and had stable pharmacological and/or psychotherapeutic treatment for at least four weeks prior to enrollment. Participants were excluded if the diagnosis of MDD was not confirmed, if they presented severe psychiatric comorbidities, had undergone neuromodulation within the previous six months, reported illicit drug use or substance abuse, had medical conditions contraindicating electrical stimulation, were pregnant or breastfeeding, had implanted electronic devices, had unstable clinical treatment or changed medication during the study, or presented significant suicide risk.

Once deemed eligible, the participants provided written informed consent and completed the baseline assessment, which included the Beck Depression Inventory-II (BDI-II), the HAM-D, the World Health Organization Quality of Life-BREF (WHOQOL-BREF), and the Suicidal Ideation Questionnaire (SIQ). At the pre-intervention (baseline) assessment, Participant 1, aged 25 years, scored 18 on the HAM-D, whereas Participant 2, aged 52 years, scored 22. Following the baseline assessment, participants underwent the intervention protocol.

B. Research Ethics Committee (REC)

This project was approved by the Research Ethics Committee (REC) of the Santos Dumont Institute (ISD), CAAE: 86722925.5.0000.0129 on April 14th 2026. Prior to the initial assessment, all participants signed the Informed Consent Form and the Authorization Form for voice recording and image capture (photos and videos), in accordance with ethical standards for research involving human subjects as set forth by the National Health Council (Brazil; National Health Council, 2012).

This research adhered to the ethical standards required by Resolution No. 466/12 of the National Health Council (CNS), which regulates research involving human subjects.

C. Intervention Protocol

Participants underwent 15 sessions of tsDCS administered over three weeks (five sessions per week), using a direct current intensity of 2.5 mA for 20 minutes per session. The intervention protocol did not include any cognitive tasks; participants remained in the prone position at rest throughout the stimulation sessions. The anodal electrode was positioned over the T10 vertebral level, while the cathodal electrode was placed over the C3 cervical region (Figure 1). Throughout the intervention, participants were continuously monitored for discomfort and potential skin reactions. Psychological assessments were conducted at four time points: baseline(T0), mid-treatment (T1; 7th session), post-intervention (T2; 15th session), and at the 90-day follow-up(T3), to evaluate the persistence of treatment effects and the potential recurrence of symptoms.

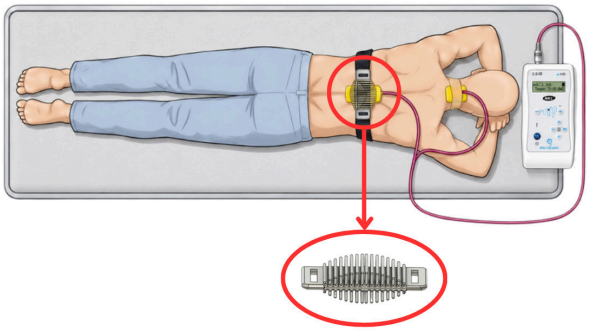


Fig. 1. Electrode Placement. The anodal electrode was positioned over the T10 vertebral level, while the cathodal electrode was placed over the C3 cervical region

D. Equipment

The tsDCS was delivered using a Micro Estim Foco Research stimulator (NKL, Brazil) with saline-soaked (0.9% NaCl) surface electrodes (5 × 7 cm). The anodal electrode was positioned over the T10 vertebral level and secured using a customized anatomical adapter designed in Autodesk Fusion 360 and three-dimensionally printed with thermoplastic polyurethane (TPU). The adapter was used to promote better contact between the electrode and the spine, conforming to the curvature of the subject's body. The cathodal electrode was positioned over the C3 cervical region and secured with kinesiology tape.

E. Assessments Instruments

Standardized instruments were used to assess depressive symptoms, quality of life, suicidal ideation, and intervention safety. These included the Beck Depression Inventory-II (BDI-II), the Hamilton Depression Rating Scale (HAM-D), the World Health Organization Quality of Life-BREF (WHOQOL-BREF), the Suicidal Ideation Questionnaire (SIQ), and a Safety and Tolerability Questionnaire.

IV. Results

A. Depressive Symptoms

At baseline (T0; Figure 2), Participants 1 and 2 presented HAM-D scores of 18 and 22, respectively, indicating moderate depressive symptoms. Throughout the intervention, a progressive reduction in HAM-D scores was observed in both participants (T1 and T2), with both reaching a score of 4 by the end of the 15th session. At follow-up, Participant 1 maintained the clinical improvements (HAM-D = 6), whereas Participant 2 experienced a relapse of depressive symptoms, with the HAM-D score returning to a level close to baseline (HAM-D = 21).

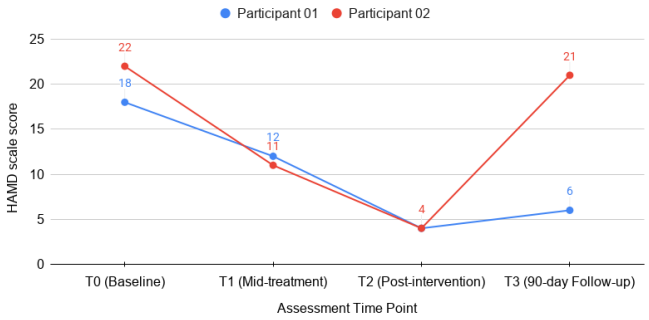


Fig. 2. Comparison of HAM-D scores

At baseline, as assessed by the Beck Depression Inventory-II (BDI-II) (T0; Figure 3), Participants 1 and 2 scored 33 and 57, respectively, indicating moderate to severe depressive symptoms. Throughout the intervention (T1 and T2), both participants showed a consistent reduction in BDI-II scores, reaching 6 (Participant 1) and 20 (Participant 2) by the end of the 15th session. At follow-up, Participant 1 maintained the clinical improvement (BDI-II = 6), whereas Participant 2 experienced a relapse, with the BDI-II score returning to the baseline level (BDI-II = 58).

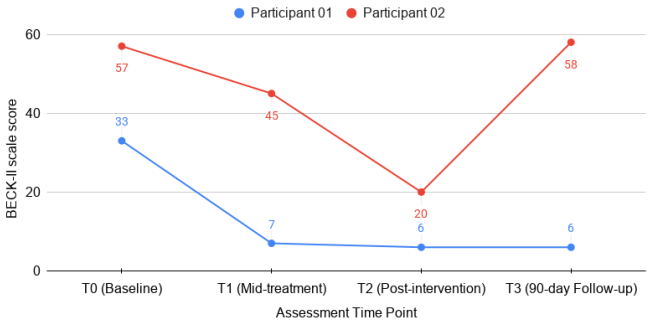


Fig. 3. Comparison of BDI-II scores

B. Quality of Life

Quality of life was assessed using the psychological domain of the WHOQOL-BREF (Figure 4). At baseline, Participants 1 and 2 scored 33.33 and 20.83, respectively. After the intervention, both participants demonstrated improvements in psychological quality of life. At the intermediate assessment, scores increased to 62.50 for Participant 1 and 37.50 for Participant 2, reaching 66.67 and 62.50, respectively, by the end of the intervention. At follow-up, Participant 1 maintained the final score (66.67), whereas Participant 2 showed a decline, with the score decreasing to 8.33.

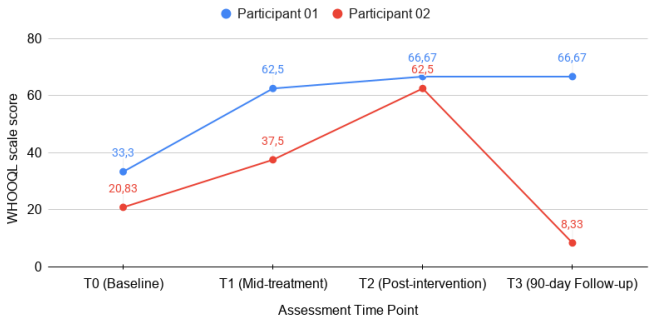


Fig. 4. Comparison of WHOQOL-BREF scores

C. Suicidal Ideation

The results of the Suicidal Ideation Questionnaire (SIQ) (Figure 5) showed baseline scores of 47 for Participant 1 and 151 for Participant 2. After the intervention, both participants exhibited a progressive reduction in suicidal ideation. At the intermediate and final assessments, Participant 1 scored 25 and 14, respectively, whereas Participant 2 scored 122 and 27. At follow-up, Participant 1 maintained the downward trend, with an SIQ score of 6, while Participant 2 experienced a recurrence of suicidal ideation, with the score increasing to 118.

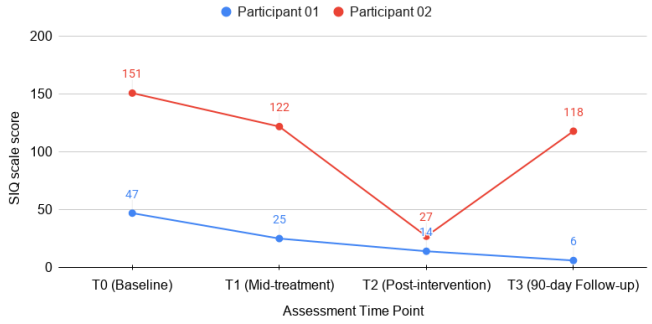


Fig. 5. Comparison of SIQ scores

D. Safety and Tolerability

tsDCS demonstrated a preliminary safety profile and feasibility in both participants, with no serious adverse events or permanent protocol interruptions. The main reported effects included tingling, burning sensations, itching, and erythema at the electrode sites, which were transient and restricted to the current ramp-up phase. Participant 1 predominantly reported low discomfort levels (1 to 5 points). Conversely, Participant 2 exhibited greater skin sensitivity, reporting discomfort scores

between 5 and 8 points and occasional burning sensations that required brief pauses before resuming stimulation. As both participants completed all 15 sessions, these findings suggest that the protocol is safe and feasible for future clinical investigations, provided that individual tolerability is monitored.

V. Discussion

The findings of this study suggest that transcutaneous spinal direct current stimulation (tsDCS) was able to promote clinical improvement in both participants during the intervention period, as evidenced by reductions in depressive symptoms and suicidal ideation, as well as improvements in quality of life. However, the main finding was not solely the initial clinical response, but rather the difference in the maintenance of these effects throughout the follow-up period. While Participant 1 maintained the benefits achieved after the 15 stimulation sessions, Participant 2 experienced a recurrence of symptoms, returning to scores close to those observed at baseline when measured at the follow-up, 90 days after the end of the intervention.

This difference becomes particularly relevant when considering that both participants exhibited similar responses during the intervention period. At the end of the protocol, Participant 1 showed reductions in scores to HAM-D = 4, BDI-II = 6, and SIQ = 14, whereas Participant 2 reached HAM-D = 4, BDI-II = 20, and SIQ = 27. Additionally, both participants achieved scores consistent with a good perception of quality of life in the psychological domain (WHOQOL-BREF = 66.7 and 62.5, respectively). These results indicate that tsDCS promoted clinical improvement in both participants, despite the greater baseline symptom severity observed in Participant 2.

However, the follow-up assessment revealed an important divergence between the two cases. Participant 1 maintained the clinical improvements achieved during the intervention, remaining in remission of depressive symptoms (HAM-D = 6), whereas Participant 2 experienced a recurrence of symptoms, as evidenced by increased HAM-D (21), BDI-II (58), and SIQ (118) scores, accompanied by a marked decline in quality of life (WHOQOL-BREF = 8.33). These findings suggest that an initial clinical response to tsDCS does not necessarily predict the long-term maintenance of therapeutic benefits.

The observed effects may be related to the ability of tsDCS to modulate neural circuits involved in emotional regulation. Recent evidence suggests that stimulation at the T10 thoracic level may influence interoceptive pathways connecting the spinal cord to brain regions involved in emotional processing, including the thalamus and the insular cortex, thereby contributing to changes in mood perception and stress responses [8]. Although these mechanisms require further investigation, they provide a plausible explanation for the concurrent improvements observed in depressive symptoms, quality of life, and suicidal ideation.

Consistent with this broad neuromodulatory effect, another important finding was the agreement across the different outcome measures. Reductions in HAM-D and BDI-II scores were accompanied by improvements in psychological quality of life and decreases in suicidal ideation, suggesting that the clinical benefits extended beyond the alleviation of depressive symptoms and were reflected in subjective well-being and the reduction of clinical risk. Likewise, this consistency across outcomes was also evident during the deterioration observed at follow-up, as the recurrence of symptoms in Participant 2 was accompanied by a decline in quality of life and an increase in suicidal ideation.

The recurrence observed in Participant 2, however, highlights the vulnerability of treatment outcomes to external factors. At the time of clinical deterioration, the participant was undergoing antidepressant tapering and experiencing significant psychosocial stressors. Previous evidence has demonstrated that pharmacological agents can enhance and prolong the neuroplastic effects induced by direct current stimulation, whereas medication withdrawal may reduce the persistence of these stimulation-induced adaptations [9]. More recent clinical evidence also supports the importance of maintaining pharmacotherapy after the acute phase of neuromodulation. A meta-analysis by Wang et al. [10] showed that combining brain stimulation techniques with antidepressant medication was more effective in preventing depressive relapse over 6 to 12 months than either intervention alone. According to the authors, antidepressants may exert a synergistic effect by enhancing neuroplastic mechanisms, including increased brain-derived neurotrophic factor (BDNF) signaling, thereby facilitating the consolidation of stimulation-induced therapeutic effects. Similarly, Woodham et al. [11] reported that sustained clinical remission following home-based transcranial direct current stimulation was more frequently observed among participants who remained under active treatment, either through continued antidepressant use or maintenance stimulation sessions. In the present case, antidepressant tapering, together with psychosocial stressors, may therefore have contributed to the gradual loss of the neuroplastic adaptations promoted by tsDCS and the recurrence of depressive symptoms, although this hypothesis cannot be confirmed within the present study. These findings suggest that, although neuromodulation may induce meaningful neurophysiological and clinical improvements, the long-term maintenance of its therapeutic effects likely depends on a favorable clinical and psychosocial context [12]. As demonstrated in this case, medication instability and social stress may outweigh the benefits of spinal stimulation.

Regarding safety, the intervention showed a favorable profile and was well tolerated by both participants. Adverse events were mild, transient, and consistent with those previously described for transcutaneous electrical stimulation techniques. Despite the greater sensitivity reported by Participant 2, who occasionally required brief pauses during stimulation, neither participant discontinued the protocol, reinforcing the clinical feasibility of the technique.

The limitations of this study include its case series design, the small sample size, and the absence of a control or sham stimulation group, which limit the generalizability of the findings and prevent the establishment of causal relationships. In particular, without a control or sham condition, it

is not possible to determine whether the observed clinical improvements were specifically attributable to tsDCS or were partly related to placebo effects, spontaneous fluctuations or remission of depressive symptoms, or the therapeutic contact associated with the 15 intervention sessions. Therefore, the clinical improvements observed in both participants should be interpreted cautiously and cannot be attributed exclusively to the neuromodulation intervention. Nevertheless, the present findings contribute to the preliminary evidence regarding the application of tsDCS in MDD and reinforce the need for future controlled studies with larger samples to investigate the efficacy of tsDCS, the durability of its therapeutic effects, and potential biomarkers capable of predicting clinical response to the intervention.

VI. Conclusion

Transcutaneous spinal direct current stimulation (tsDCS) represents a potential complementary approach for the treatment of Major Depressive Disorder (MDD). The findings of this study demonstrated a favorable clinical course in both participants throughout the 15 stimulation sessions, with reductions in depressive symptoms, improvements in quality of life, and decreased suicidal ideation measured exactly after the intervention. However, the follow-up assessment, measured 90 days after the intervention, revealed that the maintenance of treatment effects may vary between individuals, as observed in Participant 2, who experienced a recurrence of symptoms after the end of the intervention, but with a confounding factor of interrupting their current medication. These findings highlight the importance of investigating factors associated with the duration and persistence of therapeutic effects induced by tsDCS. Despite the inherent limitations of a case study design, these results contribute to the understanding of the therapeutic potential of tsDCS and reinforce the need for future studies employing different stimulation protocols, extended follow-up periods, and investigations of physiological biomarkers to further elucidate its mechanisms of action, clinical applicability, and effectiveness in the management of MDD.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Fundação de Apoio à Pesquisa do Estado do Rio Grande do Norte (FAPERN) through the Programa de Fomento à Ciência, Tecnologia e Inovação do Rio Grande do Norte (FUCITERN). The authors also thank the Edmond and Lily Safra International Institute of Neuroscience (IIN-ELS), Instituto Santos Dumont, for providing the facilities and infrastructure where part of this study was conducted. Special thanks are extended to Dr. Jales Clemente for his contribution to the psychiatric assessment and confirmation of Major Depressive Disorder diagnoses prior to participant enrollment. The authors also thank Ana Luiza Matos da Silva and José Roberto Gurgel Dantas for their contribution to the development of the adapter used for T10 spinal stimulation, Gabrielly Blasques Oliveira Albuquerque and Isadora Medeiros Cortez for their valuable assistance during the intervention sessions, and all the participants for their cooperation and commitment throughout the study.

References

[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, DSM-5-TR*, 5th ed. Washington, DC, USA: American Psychiatric Association, 2022, doi: 10.1176/appi.books.9780890425787.

[2] P. Dalgalarrondo, *Psicopatologia e Semiologia dos Transtornos Mentais*, 3rd ed. Porto Alegre, Brazil: Artmed, 2019.

[3] P. Marangolo, S. Bonifazi, A. Cristofori, and C. Caltagirone, "What Else Can Be Done by the Spinal Cord? A Review on the Effectiveness of Transpinal Direct Current Stimulation (tsDCS) in Stroke Recovery," *International Journal of Molecular Sciences*, vol. 24, no. 12, p. 10173, June 2023, doi: 10.3390/ijms241210173.

[4] Y. Al'joboori, S. J. Massey, S. L. Knight, N. N. Donaldson, and L. D. Duffell, "The Effects of Adding Transcutaneous Spinal Cord Stimulation (tSCS) to Sit-To-Stand Training in People With Spinal Cord Injury: A Pilot Study," *Journal of Clinical Medicine*, vol. 9, no. 9, p. 2765, Aug. 2020, doi: 10.3390/jcm9092765.

[5] M. Hogan, G. Hamilton, and P. Horner, "Neural Stimulation and Molecular Mechanisms of Plasticity and Regeneration: A Review," *Frontiers in Cellular Neuroscience*, vol. 14, p. 271, Oct. 2020, doi: 10.3389/fncel.2020.00271.

[6] P. Picelli et al., "Combined Effects of Transcranial Direct Current Stimulation (tDCS) and Transcutaneous Spinal Direct Current Stimulation (tsDCS) on Robot-Assisted Gait Training in Patients With Chronic Stroke: A Pilot, Double Blind, Randomized Controlled Trial," *Restorative Neurology and Neuroscience*, pp. 357–368, 2015, doi: 10.3233/RNN-140474.

[7] F. Pisano, C. Caltagirone, F. Satriano, R. Perri, L. Fadda, and P. Marangolo, "Can Alzheimer's Disease Be Prevented? First Evidence From Spinal Stimulation Efficacy on Executive Functions," *Journal of Alzheimer's Disease*, vol. 77, no. 4, pp. 1755–1764, 2020, doi: 10.3233/JAD-200695.

[8] F. Romo-Nava et al., "Effect of Non-Invasive Spinal Cord Stimulation in Unmedicated Adults With Major Depressive Disorder: A Pilot Randomized Controlled Trial and Induced Current Flow Pattern," *Molecular Psychiatry*, vol. 29, no. 3, pp. 580–589, Mar. 2024, doi: 10.1038/s41380-023-02349-9.

[9] M. A. Nitsche et al., "Pharmacological Modulation of Cortical Excitability Shifts Induced by Transcranial Direct Current Stimulation in Humans," *Journal of Physiology*, vol. 553, no. Pt. 1, pp. 293–301, Nov. 2003, doi: 10.1113/jphysiol.2003.049916.

[10] R. Wang et al., "Maintenance of Noninvasive Brain Stimulation for Preventing Relapse in Depression: A Systematic Review and Meta-Analysis," *Alpha Psychiatry*, vol. 26, no. 6, p. 49140, Dec. 2025, doi: 10.31083/AP49140.

[11] R. D. Woodham et al., "Home-Based Transcranial Direct Current Stimulation for Major Depressive Disorder: 6-Month Follow-up From Randomised Sham-Controlled Trial and Open-Label Treatment Phases," *Journal of Psychiatric Research*, vol. 186, pp. 23–32, 2025, doi: 10.1016/j.jpsychires.2025.03.047.

[12] J. Buckman et al., "Life Events and Treatment Prognosis for Depression: A Systematic Review and Individual Patient Data Meta-Analysis," *Journal of Affective Disorders*, vol. 299, pp. 298–308, Feb. 2022, doi: 10.1016/j.jad.2021.12.030.

Elaboração de Procedimentos Operacionais Padronizados para Controle de Qualidade em PET-CT

1st Pietra Caroline de Pena
Mestranda - PPGEB
UTFPR
Curitiba, Brasil
https://orcid.org/0009-0006-2849-7444

2nd Danielle Filipov Pereira
Professora Doutora - PPGEB
UTFPR
Curitiba, Brasil
https://orcid.org/0000-0002-0040-1184

3rd Marcelo Antoniassi
Professor Doutor - PPGEB
UTFPR
Curitiba, Brasil
https://orcid.org/0000-0003-1636-9366

***Resumo* — A Tomografia por Emissão de Pósitrons associada à Tomografia Computadorizada (PET-CT) constitui uma das modalidades mais tecnológicas da medicina nuclear atual, sendo amplamente utilizada na detecção, estadiamento e monitoramento de diversas patologias. A qualidade das imagens obtidas e a confiabilidade dos resultados dependem diretamente da correta execução dos testes de controle de qualidade do equipamento. Neste contexto, os Procedimentos Operacionais Padronizados (POPs) desempenham papel fundamental na padronização, rastreabilidade e segurança das atividades realizadas. Este estudo teve como objetivo elaborar e padronizar POPs para os testes de controle de qualidade em sistemas PET-CT, com base em recomendações nacionais e internacionais vigentes. Para isso, foi realizada uma pesquisa bibliográfica e documental em normas regulatórias e guias técnicos, seguida da identificação dos principais testes aplicáveis à rotina clínica e da elaboração dos respectivos procedimentos. Como resultado, foram desenvolvidos POPs contendo objetivo, responsabilidades, materiais necessários, metodologia de execução, critérios de aceitação e registros obrigatórios. A padronização proposta contribui para o fortalecimento da garantia de qualidade nos Serviços de Medicina Nuclear, promovendo maior segurança aos pacientes, confiabilidade diagnóstica, conformidade regulatória e otimização dos processos operacionais.**

***Palavras-Chave* — PET-CT, Controle de Qualidade, Procedimento Operacional Padronizado, Medicina Nuclear, Garantia de Qualidade**

I. Introdução

A Tomografia por Emissão de Pósitrons associada à Tomografia Computadorizada (PET-CT) constitui uma das principais modalidades diagnósticas da Medicina Nuclear moderna, permitindo a aquisição simultânea de informações funcionais e anatômicas em um único exame (IAEA, 2009).

A qualidade das imagens produzidas pelos sistemas PET-CT dependem diretamente do desempenho adequado dos componentes PET e CT, bem como da correta calibração dos equipamentos envolvidos no processo de aquisição e reconstrução das imagens (AAPM, 2019).

Falhas relacionadas aos detectores do PET, sistemas eletrônicos, alinhamento espacial entre as modalidades ou componentes do tomógrafo podem resultar em degradação da qualidade da imagem e comprometer tanto a interpretação visual, quanto a qualificação da atividade radioativa (NEMA, 2018).

Neste contexto, programas de garantia de qualidade têm sido amplamente recomendados por organismos internacionais, como a International Atomic Energy Agency (IAEA), a American Association of Physicists in Medicine (AAPM) e a National Electrical Manufacturers Association (NEMA), com o objetivo de assegurar a manutenção do desempenho dos equipamentos ao longo do tempo (AAPM, 2018).

Entretanto, a efetividade desses programas depende da existência de procedimentos claramente documentados, capazes de orientar a execução dos testes de forma uniforme e rastreável. Os Procedimentos Operacionais Padronizados (POPs) representam uma ferramenta essencial para a padronização das atividades, contribuindo para a redução da variabilidade operacional e para a conformidade com requisitos regulatórios aplicáveis aos serviços de Medicina Nuclear (IAEA, 2013).

Dessa forma, o presente estudo tem como objetivo elaborar Procedimentos Operacionais Padronizados destinados ao controle de qualidade em sistemas PET-CT, fundamentados em recomendações nacionais e internacionais reconhecidas na área.

II. Metodologia

Trata-se de um estudo metodológico de natureza descritiva, desenvolvido por meio de pesquisa bibliográfica e documental, além de observados testes de qualidade e seus resultados realizados em um Serviço de Medicina Nuclear.

Foram consultadas publicações da IAEA, documentos da AAPM, normas da NEMA, regulamentações da Comissão Nacional de Energia Nuclear (CNEN) e demais documentos relacionados ao controle de qualidade em PET-CT.

Além disso, foi observada a rotina de testes de qualidade diários, semanais, mensais e anuais realizados no setor PET-CT de um Serviço de Medicina Nuclear, assim como os resultados obtidos.

A seleção dos testes de qualidade contemplou os procedimentos recomendados conforme preconizado pelas referências consultadas e, a partir das informações levantadas, foi elaborada uma estrutura padronizada para os POPs contendo identificação do documento, objetivo, responsabilidades, materiais necessários, procedimento operacional, critérios de aceitação, ações corretivas, registros obrigatórios e referências técnicas.

III. Fundamentação teórica

A. Sistemas PET-CT

Os sistemas PET-CT representam uma das mais importantes inovações da imagem médica nas últimas décadas, combinando a informação funcional fornecida pela PET com a informação anatômica obtida pela CT (TOWNSEND, 2008).

A modalidade PET utiliza radiofármacos emissores de pósitrons para a avaliação de processos metabólicos e fisiológicos. Após a emissão do pósitron e sua aniquilação com um elétron do meio, são produzidos dos fótons de 511 keV emitidos em direções opostas, permitindo a reconstrução tomográfica da distribuição do radiofármaco no organismo (CHERRY et al., 2012).

A incorporação do CT ao sistema PET possibilitou significativa melhoria na localização anatômica das lesões e na correção de atenuação das imagens, aumentando a precisão diagnóstica e reduzindo o tempo total de aquisição dos exames (TOWNSEND, 2008).

Atualmente, o PET-CT é amplamente empregado no diagnóstico, estadiamento, avaliação de resposta terapêutica e detecção de recidivas em diversos tipos de neoplasias, sendo considerado um dos principais recursos diagnósticos da oncologia moderna (BOELLAARD et al., 2015).

B. Garantia e Controle de Qualidade em PET-CT

A garantia de qualidade compreende um conjunto de atividades planejadas e sistemáticas destinadas a assegurar que os equipamentos operem dentro de parâmetros previamente estabelecidos, produzindo resultados confiáveis e reprodutíveis (IAEA, 2009).

Segundo a IAEA, os programas de garantia de qualidade devem contemplar os testes periódicos destinados à avaliação do desempenho dos sistemas PET e CT, incluindo verificações diárias, mensais, trimestrais e anuais (IAEA, 2009).

Os testes de controle de qualidade aplicados têm como objetivo avaliar o desempenho dos detectores, sistemas eletrônicos, algoritmos de reconstrução e processos de quantificação da atividade radioativa, por exemplo (NEMA, 2018).

A implementação de programas estruturados de controle de qualidade contribui para a detecção precoce de falhas, redução de artefatos, aumento da vida útil dos equipamentos e melhoria da qualidade diagnóstica dos exames (AAPM, 2019).

De acordo com a AAPM, a realização periódica dos testes de qualidade é essencial para assegurar a estabilidade quantitativa dos exames PET e a confiabilidade dos valores de SUV utilizados na rotina médica para avaliação final (AAPM, 2019).

C. Procedimentos Operacionais Padronizados

Os Procedimentos Operacionais Padronizados constituem documentos formais destinados à descrição detalhada das atividades realizadas em uma organização, estabelecendo responsabilidades, etapas operacionais e critérios para a execução das tarefas (ANVISA, 2017).

Em Serviços de Medicina Nuclear, a utilização de POPs é particularmente relevante devido à complexidade tecnológica dos equipamentos e aos requisitos regulatórios relacionados à proteção radiológica e à garantia de qualidade (IAEA, 2013).

A implementação de programas estruturados de controle de qualidade contribui para a detecção precoce de falhas, redução de artefatos, aumento da vida útil dos equipamentos e melhoria da qualidade diagnóstica dos exames (AAPM, 2019).

IV. Resultados

A. Estrutura dos Procedimentos Operacionais Padronizados

Como resultado deste estudo, foi desenvolvida uma estrutura padronizada para a elaboração dos Procedimentos Operacionais Padronizados aplicáveis ao controle de qualidade em sistemas PET-CT.

Cada POP foi constituído por: identificação do documento, objetivo, campo de aplicação, responsabilidades, materiais e equipamentos necessários, procedimento operacional, critérios de aceitação, ações corretivas, registros, referências, fluxograma operacional.

A padronização dessa estrutura visa facilitar a elaboração, implementação e atualização dos documentos, além de favorecer a rastreabilidade das atividades desenvolvidas.

B. POP para Teste Diário de Uniformidade

Com o objetivo de ilustrar a aplicação prática do procedimento elaborado, exemplificamos neste artigo o POP específico para o teste diário de uniformidade do PET utilizando fonte selada de Germânio-68.

O procedimento descreve as etapas de preparação do equipamento, posicionamento da fonte, aquisição dos dados, análise dos resultados e registro das informações.

O objetivo do teste é avaliar a uniformidade de resposta dos detectores do sistema PET, verificando a estabilidade e o desempenho do equipamento.

Os materiais necessários são: Fonte de Germânio-68 (Ge-68) adequada ao equipamento, sistema PET-CT, formulário ou sistema de registro dos resultados.

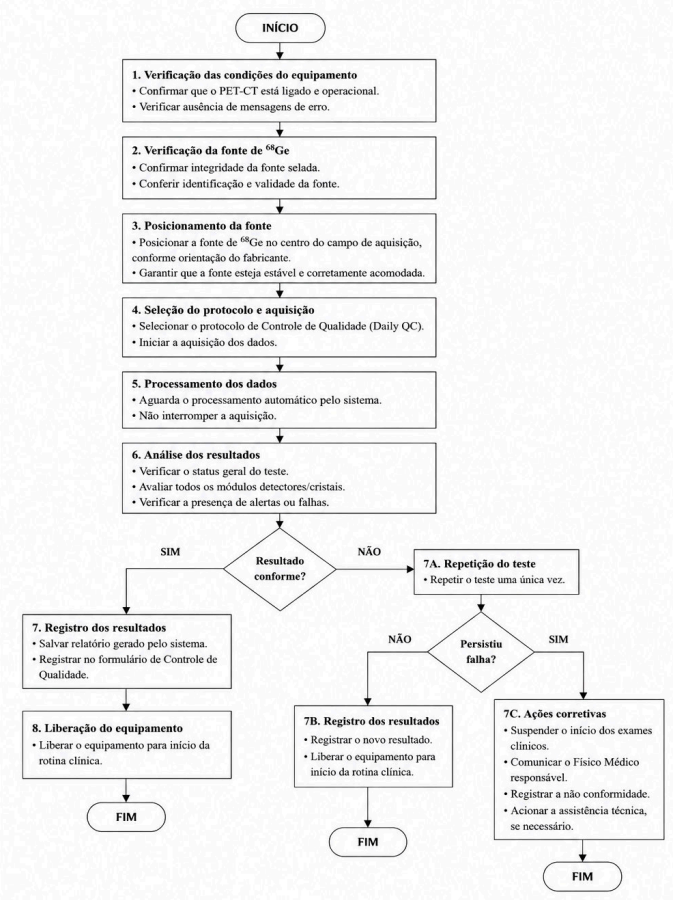
O procedimento é descrito no fluxograma (figura 1) completo do teste de uniformidade realizado diariamente no sistema PET-CT. O processo inicia-se com a verificação das condições operacionais do equipamento e integridade da fonte radioativa, seguida pelo posicionamento adequado da fonte e execução da aquisição do controle de qualidade.

Após a aquisição, o software do equipamento realiza a análise dos dados e gera o resultado dos teste. Quando os parâmetros avaliados encontram-se dentro dos limites estabelecidos pelo fabricante, o sistema emite o resultado de aprovação. Caso seja identificado resultado de reprovação, podem ser realizadas verificações complementares, como a repetição do teste.

A representação gráfica do processo permite visualizar de forma objetiva as etapas operacionais, os pontos de decisão e as ações corretivas associadas aos diferentes resultados possíveis, contribuindo para a padronização das atividades e para a

segurança operacional do serviço.

Fig. 1. Fluxograma do POP para o teste diário de uniformidade do PET-CT com fonte de Germânio-68



Fonte: Elaborado pela autora (2026)

C. Análise comparativa

Com o objetivo de verificar a aderência dos POPs elaborados às recomendações técnicas reconhecidas internacionalmente, foi realizada uma análise comparativa entre os testes contemplados neste estudo e aqueles recomendados pelos principais órgãos de referência na área de Medicina Nuclear e Física Médica. A Tabela 1 apresenta a correlação entre os testes propostos e as recomendações da International Atomic Energy Agency (IAEA), American Association of Physicists in Medicine (AAPM), National Electrical Manufacturers Association (NEMA), European Association of Nuclear Medicine (EANM) e European Federation of Organisations for Medical Physics (EFOMP).

Observa-se que os procedimentos desenvolvidos contemplam os principais testes recomendados internacionalmente para monitoramento do desempenho dos sistemas PET-CT, evidenciando a compatibilidade da proposta com os requisitos atualmente adotados em programas de garantia de qualidade

TABELA 1
CORRELAÇÃO ENTRE OS TESTES DE CONTROLE DE QUALIDADE EM PET-CT E AS RECOMENDAÇÕES INTERNACIONAIS

Testes de Controle de Qualidade	**Recomendações internacionais**					**Frequência Recomendada**
	IAEA (2009)	***AAPM (2019)***	***NEMA (2018)***	***EANM (2023)***	***EFOMP (2023)***	**-**
Uniformidade diária com fonte de 68Ge	✓	✓	✓	✓	✓	Diária
Verificação do funcionamento dos detectores	✓	✓	✓	✓	✓	Diária
Calibração cruzada PET x Calibrador de Doses	✓	✓	✓	✓	✓	Mensal
Verificação do alinhamento PET/CT	✓	✓	✓	✓	✓	Mensal
Uniformidade do CT	✓	✓	---	✓	✓	Mensal
Exatidão dos números CT	✓	✓	---	✓	✓	Mensal
Avaliação de artefatos do CT	✓	✓	---	✓	✓	Mensal
Resolução espacial do PET	✓	✓	✓	✓	✓	Anual
Sensibilidade do PET	✓	✓	✓	✓	✓	Anual
Avaliação global com phantom NEMA	✓	✓	✓	✓	✓	Anual
Revisão documental dos registros de CQ	✓	✓	---	✓	✓	Mensal
Gestão de não conformidades	✓	✓	---	✓	✓	Contínua

Fonte: Elaborado pela autora com base em IAEA (2009), AAPM (2019), NEMA (2018), Koole et al. (2023) e Matheoud et al. (2023).

D. Indicadores Quantitativos dos POPs Desenvolvidos

Com o objetivo de caracterizar quantitativamente os produtos gerados durante o desenvolvimento deste estudo, foram definidos indicadores relacionados à estrutura documental elaborada para o Programa de Controle de Qualidade em sistemas PET-CT.

A elaboração dos POPs foi baseada na análise de normas, diretrizes e documentos técnicos nacionais e internacionais, resultando na construção de uma estrutura padronizada capaz de contemplar os principais testes de controle de qualidade aplicáveis aos sistemas PET-CT.

A Tabela 2 apresenta os principais indicadores quantitativos obtidos durante o desenvolvimento do estudo.

TABELA 2
INDICADORES QUANTITATIVOS DO ESTUDO DE CONTROLE DE QUALIDADE PARA PET-CT

Indicador	**Quantidade**
Procedimentos Operacionais Padronizados Elaborados	10
Testes de controle de qualidade	12

contemplados	
Fluxogramas operacionais desenvolvidos	10
Matrizes de controle de qualidade elaboradas	1
Descrição de não conformidade	1
Referências normativas internacionais analisadas	5
Referências bibliográficas utilizadas	12
Componentes do sistema PET-CT contemplados	3

Fonte: Elaborado pela autora (2026).

Os resultados demonstram que a proposta desenvolvida não se restringe à elaboração isolada de procedimentos operacionais, abrangendo também elementos de gestão da qualidade, rastreabilidade documental e monitoramento sistemático do desempenho dos equipamentos.

A inclusão de fluxogramas operacionais, matrizes de controle de qualidade e tabelas de tratamento de não conformidades contribui para a integração dos procedimentos em um programa estruturado de garantia da qualidade, alinhado às recomendações da IAEA, AAPM, NEMA, EANM e EFOMP.

Além disso, os indicadores apresentados evidenciam que a proposta contempla tanto aspectos relacionados ao desempenho dos equipamentos, quanto aspectos gerenciais voltados à documentação, rastreabilidade e melhoria contínua dos processos, elementos considerados fundamentais para programas modernos de garantia da qualidade em Serviços de Medicina Nuclear (IAEA, 2013).

A quantificação dos produtos gerados permite demonstrar a abrangência do trabalho e fornece subsídios para futuras etapas de implementação, validação e monitoramento do desempenho dos POPs em ambiente clínico.

V. Discussão

Os resultados obtidos neste estudo demonstram que a elaboração de Procedimentos Operacionais Padronizados constitui uma estratégia viável para a estruturação de programas de controle de qualidade em sistemas PET-CT. Embora diversas instituições disponibilizem recomendações relacionadas aos testes de desempenho desses equipamentos, observa-se que tais documentos frequentemente apresentam caráter normativo e técnico, exigindo adaptação para a realidade operacional dos serviços de Medicina Nuclear (AAPM, 2019).

A padronização proposta neste trabalho busca reduzir a variabilidade na execução dos testes de controle de qualidade por meio da definição clara de responsabilidades, etapas operacionais, critérios de aceitação e ações corretivas. Segundo a IAEA, a existência de procedimentos formalmente documentados representa um dos pilares dos programas ed garantia da qualidade, contribuindo para a uniformidade dos processos e para a rastreabilidade das informações produzidas (IAEA, 2013).

Entre os procedimentos elaborados, destaca-se o POP destinado ao teste diário de uniformidade com fonte de Ge-68. Esse teste é considerado uma das verificações mais importantes da rotina operacional dos sistemas PET-CT, pois possibilita a identificação precoce de alterações no desempenho dos detectores e dos sistemas eletrônicos associados ao equipamento (NEMA, 2018).

Sob a perspectiva regulatória, a adoção de POPs contribui para o atendimento aos requisitos relacionados à garantia de qualidade e à proteção radiológica estabelecidos para os Serviços de Medicina Nuclear. A documentação estruturada das atividades facilita processos de auditoria, inspeções regulatórias e programas de acreditação institucional, além de demonstrar o compromisso do serviço com a melhoria contínua da qualidade (CNEN, 2021).

Portanto, os resultados obtidos sugerem que a elaboração de POPs representa uma ferramenta relevante para a organização das atividades de controle de qualidade em PET-CT, contribuindo para a segurança operacional, a confiabilidade diagnóstica e a conformidade com as normas nacionais e internacionais.

VI. Conclusão

A qualidade das imagens produzidas pelos sistemas PET-CT depende diretamente da execução adequada dos testes de controle de qualidade e da manutenção do desempenho dos equipamentos ao longo do tempo.

Neste contexto, o presente estudo teve como objetivo elaborar POPs destinados às principais atividades de controle de qualidade em sistemas PET-CT, fundamentados em normas, recomendações técnicas e princípios de garantia da qualidade reconhecidos nacional e internacionalmente.

Como resultado, foi proposta uma estrutura padronizada para a elaboração dos POPs e desenvolvido um conjunto de procedimentos aplicáveis aos principais testes recomendados para sistemas PET-CT, incluindo o teste diário de uniformidade com fonte de Ge-68.

A proposta apresentada contribui para a uniformização das rotinas operacionais, para a rastreabilidade documental e para o fortalecimento dos programas de garantia da qualidade em Medicina Nuclear, fornecendo subsídios técnicos e implementação de padronização e melhoria contínua nos processos relacionados ao controle de qualidade em PET-CT.

Referências

[1] AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE. PET/CT Acceptance Testing and Quality Assurance. AAPM Report No. 126. Alexandria: AAPM, 2019.

[2] ANVISA. Agência Nacional de Vigilância Sanitária. Gestão da qualidade em Serviços de Saúde. Brasília: ANVISA, 2017.

[3] BOELLAARD, R. et al. FDG PET/CT: EANM procedure guidelines for tumour imaging. European Journal of Nuclear Medicine and Molecular Imaging, v. 42, n. 2, p. 328-354, 2015.

[4] BRASIL. Comissão Nacional de Energia Nuclear. CNEN NN 3.05: Requisitos de Segurança e Proteção Radiológica para Serviços de Medicina Nuclear. Rio de Janeiro: CNEN.

[5] CHERRY, S. R.; SORENSON, J. A.; PHELPS, M. E. Physics in Nuclear Medicine. 4. Ed. Philadelphia: Elsevier Saunders, 2012.

[6] CNEN. Comissão Nacional de Energia Nuclear. NN 3.05 - Requisitos

de Segurança e Proteção Radiológica para Serviços de Medicina Nuclear. Rio de Janeiro: CNEN, 2021.

[7] INTERNATIONAL ATOMIC ENERGY AGENCY. Quality Assurance for PET and PET/CT Systems. Human Health Series No. 1. Vienna: IAEA, 2009.

[8] INTERNATIONAL ATOMIC ENERGY AGENCY. Standard Operating Procedures for PET/CT: A Practical Approach for Use in Adult Oncology. Vienna: IAEA, 2013.

[9] KOOLE, M. et al. EANM guidelines for PET-CT and PET-MR routine quality control. Zeitschrift fur Medizinische Physik, v. 33, n. 1, p. 103-113, 2023.

[10] MATHEOUD, R. et al. EFOMP's protocol quality contyrols in PET/CT and PET/MR. Physica Medica, v. 105, p. 102-506, 2023.

[11] NATIONAL ELECTRICAL MANUFACTURERS ASSOCIATION. Performance Measurements of Positron Emission Tomographs. NEMA Standards Publication NU 2-2018. Rosslyn: NEMA, 2018.

[12] TOWNSEND, D. W. Multimodality imaging of structure and function. Physics in Medicine and Biology, v. 53, n. 4, p. R1-R39, 2008.

Avaliação da Estabilidade de Cloretos em Amostras Refrigeradas para aplicação no Teste do Suor

Daniel Saverio Spozito*†, Michele V. S. S. Rondon†
* Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP
São Paulo, Brasil
† Universidade Estadual de Campinas – UNICAMP
Campinas, São Paulo, Brasil
Daniel Saverio Spozito – ORCID: 0000-0003-1120-8555
Michele V. S. S. Rondon – ORCID: 0000-0002-6833-8759

***Resumo*—O Teste do Suor é o padrão-ouro para o diagnóstico da Fibrose Cística, exigindo a dosagem quantitativa de cloretos em microvolumes. Os Centros de Referência frequentemente enfrentam desafios logísticos e geográficos entre os pontos de coleta e os laboratórios de análise, tornando necessária a avaliação de protocolos de armazenamento temporal das amostras.**

Objetivo: Avaliar preliminarmente a estabilidade temporal da concentração de cloretos em soluções padrão impregnadas em papel-filtro e mantidas sob refrigeração controlada por até 28 dias.

Métodos: Foram utilizadas 60 unidades de papel filtro qualitativo Whatman número 1 saturadas com três soluções padrão de referência, representativas das faixas de baixa, intermediária e alta concentração de cloretos, com cinco unidades analisadas por ponto temporal. As análises quantitativas foram realizadas por titulometria. Para cada concentração e tempo foram calculadas medidas descritivas e intervalos de confiança. A comparação global entre os períodos foi realizada pelo teste de Kruskal–Wallis e, quando aplicável, por comparações pós-hoc com correção para múltiplos testes.

Resultados: Não foram observadas diferenças globais entre os tempos para as soluções de menor e maior concentração. Para a solução intermediária, entretanto, foi observada diferença global entre os períodos, com aumento das concentrações no último período de armazenamento e aproximação do ponto de decisão clínica.

Conclusão: Os resultados são compatíveis com estabilidade analítica preliminar nas condições estudadas para as faixas de menor e maior concentração, enquanto a faixa intermediária requer cautela, especialmente em períodos prolongados e nas proximidades dos limites diagnósticos. Os achados devem ser confirmados em estudos com suor humano, microvolumes clínicos e maior tamanho amostral.

***Index Terms*—Fibrose Cística, Teste do Suor, Conservação de Espécimes Biológicos, Estabilidade Química, Titrimetria**

I. Introdução

O diagnóstico da Fibrose Cística (FC), doença caracterizada por infecções crônicas e recorrentes do pulmão, insuficiência pancreática e manifestações gastrintestinais típicas, é confirmado por um médico através de achados clínicos e exames laboratoriais [1].

O Teste do Suor (TS) é atualmente o principal exame indicativo da doença, pois indica os níveis de cloreto no suor, os quais estão elevados em fibrocísticos [2]. Considerado exame padrão-ouro, o TS é simples, indolor e não invasivo. Seu protocolo prevê as seguintes etapas: estimulação das glândulas sudoríparas, coleta do suor e dosagem quantitativa dos cloretos presentes no suor [3].

No Brasil, a Portaria SAS/MS nº 224/2010 normatizou a dosagem quantitativa de cloretos no suor como o teste mais fidedigno no processo diagnóstico de FC no Brasil [14].

Em 2015, a Portaria nº 2.099 atualizou os critérios técnicos para habilitação dos Centros de Referência (CRs), estabelecendo coulometria e titulometria como métodos aceitos para análise confirmatória [15].

Recentemente, a Portaria Conjunta SAES/SECTICS nº 5, de 30 de abril de 2024, aprovou a nova atualização do Protocolo Clínico e Diretrizes Terapêuticas da Fibrose Cística (PCDT). Essa normativa consolida o consenso técnico-científico e estabelece os parâmetros nacionais vigentes para o diagnóstico, tratamento e mecanismos de regulação e controle da doença no âmbito do Sistema Único de Saúde (SUS) [16].

A primeira etapa essencial do Teste do Suor consiste na estimulação das glândulas sudoríparas, a qual deve ser executada obrigatoriamente por meio da técnica de iontoforese por pilocarpina. Na etapa subsequente, a coleta do suor pode ser realizada utilizando gaze esterilizada livre de cloretos, papel filtro qualitativo ou sistemas modernos de microtúbulos em espirais de plástico.

A janela temporal máxima admitida para a coleta é de 30 minutos, sendo exigido um microvolume mínimo de 15 µL ou uma massa de 75 mg de amostra biológica para que o ensaio possua viabilidade analítica e não seja classificado como coleta insuficiente.

A última etapa do TS, foco deste trabalho, consiste na dosagem analítica quantitativa fundamentada na determinação da concentração de íons cloreto na amostra, cujos limiares clínicos balizam o diagnóstico de Fibrose Cística entre improvável, menor ou igual a 29 mEq/L; inconclusivo, entre 30 e 59 mEq/L; ou indicativo de Fibrose Cística, quando maior ou igual a 60 mEq/L.

Embora o PCDT atual aponte a coulometria automatizada como método de eleição, o presente trabalho empregou a dosagem titulométrica por meio do reagente comercial Hanna® HI3815 para avaliar o impacto do armazenamento refrigerado estendido por até 28 dias.

A opção deliberada pela técnica titulométrica justifica-se por consolidar-se como um método consagrado e validado há décadas pela literatura científica internacional [8], além de apresentar um custo operacional significativamente baixo em relação às demais metodologias, pois a titulometria demanda apenas vidrarias básicas e insumos de alta acessibilidade econômica.

Na técnica titulométrica, um reagente com concentração conhecida é titulado na solução cuja concentração se deseja determinar. Com o auxílio de um indicador, como a difenilcarbazona, o ponto de neutralização é identificado através da alteração da coloração da mistura reagente+solução. A concentração de cloretos é diretamente proporcional ao volume de reagente pipetado ou titulado na solução [9], [10].

Outro ponto a destacar consiste nos complexos desafios logísticos frequentemente enfrentados pelos Centros de Referência em Fibrose Cística. Em muitos cenários, a coleta do suor não ocorre no mesmo prédio ou período em que a equipe técnica laboratorial está disponível para realizar a dosagem imediata dos cloretos [2].

Há realidades regionais em que o laboratório credenciado para a análise dista centenas de quilômetros do ponto de coleta. Nesse contexto, viabilizar a dosagem de cloretos em dia e local distintos, eliminando o risco de perda ou comprometimento da amostra, evitaria a necessidade de repetição do exame e pouparia os pacientes de um desgaste físico e emocional significativo, com especial destaque para a população pediátrica.

A avaliação preliminar da estabilidade de cloretos impregnados em papel-filtro sob refrigeração constitui uma etapa experimental necessária antes da investigação de protocolos de armazenamento de amostras reais de suor.

A demonstração de estabilidade em soluções padrão pode subsidiar estudos subsequentes destinados à avaliação da matriz biológica e das condições clínicas de coleta.

O objetivo deste trabalho foi realizar uma avaliação preliminar da estabilidade temporal da concentração de cloretos em soluções padrão impregnadas em papel-filtro qualitativo e armazenadas sob refrigeração por até 28 dias, como etapa pré-clínica para o desenvolvimento de protocolos de conservação aplicáveis às amostras do TS.

II. Materiais e Métodos

O presente estudo utilizou exclusivamente soluções padrão certificadas de cloreto, não tendo sido empregadas amostras biológicas de suor humano.

Dessa forma, o protocolo avaliou a estabilidade do analito em uma matriz experimental controlada impregnada em papel-filtro, sem reproduzir integralmente os possíveis efeitos de constituintes biológicos, pH, viscosidade e demais interferentes presentes no suor.

Para a realização do trabalho, utilizou-se o reagente titulométrico Hanna® HI3815 e papel filtro qualitativo Whatman® nº 1 em todas as amostras.

Foram empregadas três soluções padrão (SSx) com concentrações de cloretos conhecidas e certificadas: SS1 (21 mEq/L), SS2 (49 mEq/L) e SS3 (101 mEq/L), fornecidas originalmente junto ao kit de testagem do cloridrômetro Clorochek® (ElitechGroup®).

Foram recortadas 60 peças de papel filtro com dimensões de 2,0 cm $\times$ 1,5 cm, manipuladas exclusivamente com o auxílio de uma pinça de aço inoxidável.

Todas as pesagens foram realizadas em uma balança semi-analítica (Marte®, modelo AD500). Inicialmente, cada peça de papel seca foi acondicionada em um coletor universal estéril de plástico, devidamente identificado de forma sequencial (1 a 60).

Determinou-se, então, a massa inicial do conjunto (papel seco + recipiente).

Na sequência, as diferentes soluções de referência (SS1, SS2 e SS3) foram depositadas em recipientes independentes.

As peças de papel foram imersas em suas respectivas soluções por 30 segundos, sendo o volume de saturação determinado de forma visual. As amostras foram preparadas e dosadas com temperatura ambiental mínima do laboratório de 21,4°C e máxima de 24,7°C.

A saturação do papel-filtro empregada neste protocolo teve finalidade experimental e não reproduziu diretamente o microvolume mínimo de 15 µL ou a massa de 75 mg recomendados para amostras clínicas do Teste do Suor.

Essa condição deverá ser reproduzida em etapa posterior de validação com matriz biológica.



Figura 1. Preparo das amostras.

As amostras foram organizadas em três grupos de 20 unidades, conforme as concentrações de referência (Tabela I).

A massa da solução absorvida pelo papel foi calculada a partir da diferença entre o peso do conjunto saturado (papel úmido + recipiente) e o peso do conjunto seco inicial.

Tabela I
Grupos das concentrações utilizadas como referência

Grupo	Quantidade (un)	Concentração esperada (mEq/L)
1	20	21 ± 2
2	20	49 ± 4
3	20	101 ± 8

Conhecido o volume líquido de solução em cada papel, procedeu-se à dosagem titulométrica para determinar a concentração de cloretos.

No primeiro dia do experimento (Dia 0), uma subamostra de cinco unidades de cada grupo, totalizando 15 amostras, foi submetida à dosagem imediata.

As 45 amostras remanescentes devidamente lacradas foram armazenadas em ambiente refrigerado (Fig. 2) com termostato ajustado para 8°C, cuja temperatura mínima do refrigerador foi de 5°C e a máxima de 9°C.

As três etapas subsequentes seguiram cronogramas específicos de armazenamento sob refrigeração, cujas dosagens foram realizadas em intervalos de 7, 14 e 28 dias para a Bateria 01 (SS1); e intervalos de 7, 21 e 28 dias para as Baterias 02 (SS2) e 03 (SS3).

Essa distribuição totalizou 20 determinações analíticas por faixa de concentração e 60 ensaios no escopo global do estudo.

Considerando que os pontos temporais intermediários não foram uniformes entre as três baterias experimentais, essa assimetria decorreu da disponibilidade laboratorial para realização do experimento, especialmente por se tratar de estudo inicial.

Por essa razão, as análises temporais foram conduzidas separadamente para cada concentração, considerando os tempos efetivamente disponíveis.



Figura 2. Armazenagem em refrigerador.

Os dados foram registrados em planilha eletrônica e reanalisados estatisticamente considerando separadamente cada concentração de referência.

Para cada ponto temporal foram calculados média, desvio padrão, intervalo de confiança de 95% da média, viés absoluto e viés percentual em relação à concentração nominal.

A normalidade foi explorada pelo teste de Shapiro–Wilk.

Como as unidades experimentais analisadas em cada tempo eram independentes, considerando o reduzido tamanho amostral ($n = 5$ por ponto) e a ocorrência de distribuição não normal em pelo menos um grupo, a comparação global entre os tempos foi realizada pelo teste não paramétrico de Kruskal–Wallis.

Quando identificada diferença global, foram realizadas comparações par a par pelo teste de Mann–Whitney, com correção de Holm para múltiplas comparações.

Como análise complementar, foi calculada ANOVA de uma via. Adotou-se nível de significância de 5%.

A ausência de diferença estatisticamente significativa não foi interpretada isoladamente como demonstração de equivalência ou estabilidade clínica.

III. Resultados

A distribuição individual das dosagens titulométricas obtidas para as 20 amostras pertencentes à Bateria 01, com concentração nominal esperada de 21 ± 2 mEq/L, ao longo dos respectivos tempos de armazenamento entre 0, 7, 14 e 28 dias está ilustrada na Fig. 3.

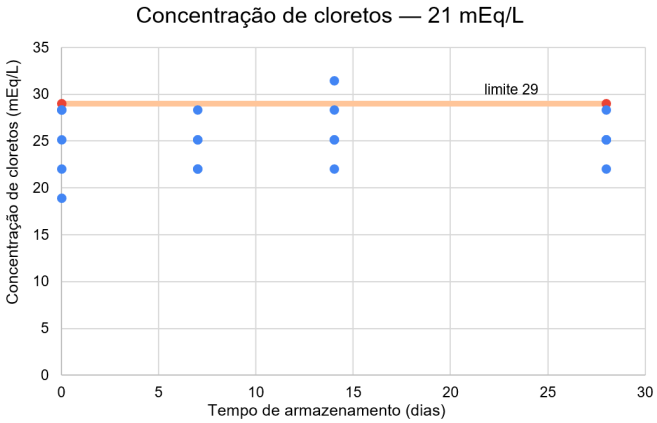


Figura 3. Concentração de cloretos da solução de referência de 21 mEq/L em função do tempo de armazenamento refrigerado.

Para o grupo de amostras da Bateria 02, correspondente à faixa intermediária, com valor de referência de 49 ± 4 mEq/L, o perfil individualizado das dosagens nos diferentes intervalos temporais entre 0, 7, 21 e 28 dias é apresentado na Fig. 4.

Observou-se maior concentração dos resultados junto ao limite superior no Dia 28, enquanto a avaliação de normalidade indicou distribuição não normal nesse ponto temporal.

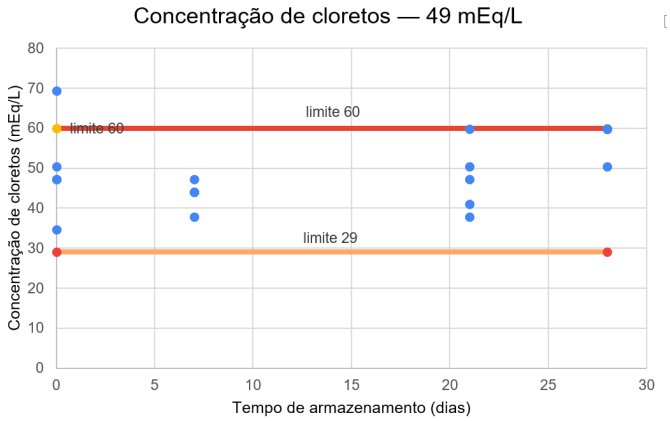


Figura 4. Concentração de cloretos da solução de referência de 49 mEq/L em função do tempo de armazenamento refrigerado.

Na Fig. 5, observa-se o comportamento das amostras da Bateria 03, formuladas na faixa francamente indicativa da doença, com valor de referência de 101 mEq/L.

Apesar da dispersão observada, as medidas permaneceram acima do limite diagnóstico de 60 mEq/L nos períodos avaliados.

A análise descritiva por ponto temporal é apresentada na Tabela II.

Para cada combinação de concentração e tempo são apresentados média, desvio padrão, intervalo de confiança de 95%, viés absoluto e viés percentual em relação ao valor nominal.

Tabela II
ESTATÍSTICA DESCRITIVA DAS CONCENTRAÇÕES DE CLORETOS SEGUNDO O TEMPO DE ARMAZENAMENTO

Nominal (mEq/L)	Dia	n	Média (mEq/L)	DP	IC95% (mEq/L)	Viés (mEq/L)	Viés (%)
21	0	5	24,57	4,11	19,47–29,67	+3,57	+17,00
21	7	5	24,57	2,64	21,30–27,84	+3,57	+17,00
21	14	5	26,46	3,59	22,00–30,92	+5,46	+26,00
21	28	5	25,20	2,23	22,43–27,97	+4,20	+20,00
49	0	5	49,77	12,48	34,27–65,27	+0,77	+1,57
49	7	5	43,47	3,45	39,19–47,75	-5,53	-11,29
49	21	5	47,25	8,63	36,54–57,96	-1,75	-3,57
49	28	5	57,96	4,23	52,71–63,21	+8,96	+18,29
101	0	5	102,06	10,59	88,91–115,21	+1,06	+1,05
101	7	5	98,91	18,98	75,34–122,48	-2,09	-2,07
101	21	5	97,65	18,64	74,51–120,79	-3,35	-3,32
101	28	5	107,73	10,77	94,35–121,11	+6,73	+6,66

DP: desvio padrão; IC95%: intervalo de confiança de 95% da média. O viés foi calculado em relação à concentração nominal da solução de referência.

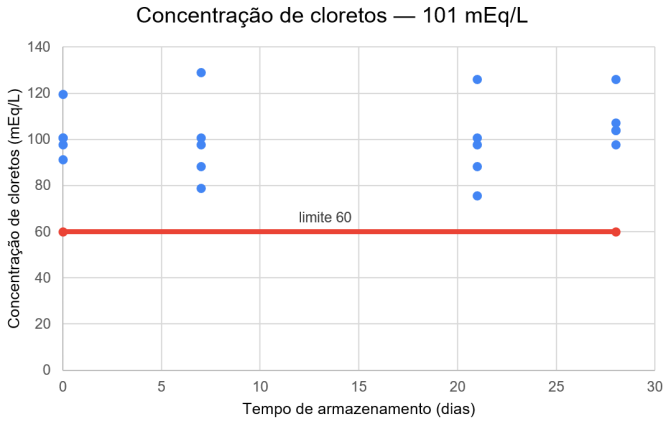


Figura 5. Concentração de cloretos da solução de referência de 101 mEq/L em função do tempo de armazenamento refrigerado.

A comparação global entre os tempos é apresentada na Tabela III.

Para a concentração nominal de 21 mEq/L não foi observada diferença entre os períodos avaliados (Kruskal–Wallis, $H = 0{,}892$; $p = 0{,}827$).

Resultado semelhante foi observado para 101 mEq/L ($H = 3{,}003$; $p = 0{,}391$).

Para a solução nominal de 49 mEq/L, entretanto, foi detectada diferença global entre os períodos ($H = 8{,}012$; $p = 0{,}0458$).

Tabela III
COMPARAÇÃO GLOBAL ENTRE OS TEMPOS DE ARMAZENAMENTO

Nominal (mEq/L)	H Kruskal–Wallis	p	ANOVA p
21	0,892	0,8274	0,7681
49	8,012	**0,0458**	0,0673
101	3,003	0,3911	0,7330

Na análise pós-hoc da faixa de 49 mEq/L, nenhuma comparação par a par permaneceu estatisticamente significativa após a correção de Holm.

A maior diferença foi observada entre os Dias 7 e 28 (Mann–Whitney, p bruto $= 0{,}0088$; p ajustado $= 0{,}0527$).

Portanto, o teste global sugere heterogeneidade temporal nessa faixa, mas o reduzido número de unidades por período limita a identificação precisa de quais tempos diferem entre si.

O resultado da Bateria 02 no Dia 28 merece atenção particular.

A média aumentou para $57{,}96 \pm 4{,}23$ mEq/L, com IC95% de 52,71 a 63,21 mEq/L. Esse intervalo expressa a incerteza em torno da média do grupo e não deve ser interpretado como probabilidade de uma amostra individual ultrapassar o ponto de decisão clínica de 60 mEq/L.

Quatro das cinco determinações desse período apresentaram valor de 59,85 mEq/L, muito próximo do limite entre a faixa intermediária e a faixa indicativa de Fibrose Cística.

Esse agrupamento de resultados próximo ao limite clínico recomenda cautela na extrapolação de 28 dias de armazenamento para amostras situadas próximas aos pontos de decisão diagnóstica.

Na faixa nominal de 21 mEq/L observou-se viés positivo em todos os tempos, inclusive no Dia 0, com médias entre 24,57 e 26,46 mEq/L.

Esse comportamento sugere contribuição do próprio método titulométrico e de sua resolução analítica, e não necessariamente efeito exclusivo do armazenamento.

Para 101 mEq/L, apesar da maior dispersão observada nos Dias 7 e 21, não houve evidência estatística de alteração global entre os tempos.

A ausência de significância estatística nas faixas de 21 e 101 mEq/L deve ser interpretada como ausência de evidência de alteração detectável nas condições avaliadas, e não como demonstração formal de equivalência.

A interpretação de estabilidade também requer critérios de aceitabilidade analítica ou clínica previamente definidos.

IV. DISCUSSÃO

Desde o estabelecimento do Teste do Suor há mais de cinco décadas, o manejo de microvolumes biológicos tem se configurado como um desafio persistente na rotina laboratorial, uma vez que a quantidade de amostra obtida varia tipicamente entre 20 µL e 70 µL.

A técnica titulométrica aqui utilizada destaca-se por prescindir de maquinários complexos ou de alto custo, demandando apenas vidrarias básicas de laboratório e micropipetas de precisão.

O princípio baseia-se na reação entre um titulado constituído pela solução padrão de cloreto impregnada no papel-filtro e um titulante de concentração conhecida, solução de nitrato de mercúrio, utilizando a difenilcarbazona como indicador químico.

O ponto de neutralização total dos cloretos, ou ponto de viragem (*endpoint*), é identificado visualmente quando a mistura transita de uma coloração alaranjada para um complexo violeta (Fig. 6).

Uma vantagem prática do reagente Hanna® HI3815 foi seu custo-benefício, estimado em aproximadamente R$ 0,50 por dosagem, viabilizando a realização do experimento.



Figura 6. Ilustração da dosagem titulométrica.

A introdução do componente temporal e térmico neste estudo permitiu uma avaliação inicial da estabilidade do cloreto em matriz experimental controlada.

Frente aos desafios logísticos dos Centros de Referência, esses resultados preliminares sustentam a investigação da refrigeração como estratégia potencial para flexibilizar o intervalo entre coleta e dosagem.

Entretanto, a aplicação clínica desse procedimento depende de confirmação em amostras reais de suor e de critérios prévios de aceitabilidade analítica e diagnóstica.

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados.

O experimento foi conduzido com soluções padrão certificadas de cloreto impregnadas em papel-filtro, e não com suor humano, não reproduzindo integralmente composição biológica, pH, viscosidade e possíveis interferentes da matriz clínica.

Foram avaliadas apenas cinco unidades em cada ponto temporal, os períodos intermediários não foram uniformes entre as três concentrações e a saturação experimental do papel-filtro não corresponde diretamente aos microvolumes utilizados na rotina clínica.

Além disso, não foram incluídas concentrações deliberadamente posicionadas junto aos pontos de decisão de 29/30 e 59/60 mEq/L.

Estudos subsequentes deverão utilizar suor humano e, preferencialmente, dividir cada amostra em alíquotas destinadas à dosagem imediata e após diferentes períodos de armazenamento, permitindo análise longitudinal de viés e manutenção da classificação diagnóstica.

V. Conclusões

Nas condições experimentais avaliadas, as soluções padrão de 21 e 101 mEq/L apresentaram comportamento compatível com estabilidade analítica preliminar ao longo dos períodos estudados, sem evidência de diferença global entre os tempos.

Para a solução de 49 mEq/L, foi observada diferença temporal global e elevação da média para 57,96 mEq/L no Dia 28. Quatro das cinco determinações desse período foram de 59,85 mEq/L, valor muito próximo do ponto de decisão clínica de 60 mEq/L, o que recomenda cautela na interpretação de períodos prolongados de armazenamento próximos aos limites diagnósticos.

Esses resultados constituem evidência pré-clínica inicial e não permitem estabelecer segurança clínica ou validar o armazenamento de amostras reais de suor por 28 dias.

Estudos subsequentes deverão incluir suor humano, microvolumes compatíveis com a rotina clínica, maior tamanho amostral, critérios prévios de aceitabilidade e concentrações especificamente posicionadas próximas aos limites diagnósticos.

Autor correspondente

Daniel Saverio Spozito
Centro de Engenharia Biomédica/Unicamp
Rua Alexander Fleming, 163, Cidade Universitária “Zeferino Vaz”
Campinas/SP, CEP 13083-881
E-mail: danielss@unicamp.br
ORCID: 0000-0003-1120-8555
https://www.cocen.unicamp.br/centros-e-nucleos/orgaos-interdisciplinares/id/4/ceb

Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP / Campus Jundiaí
Av. Ângelo Pellicciari, 727 – Parque Residencial Jundiaí II
Jundiaí – SP / Brasil
E-mail: daniel.spozito@ifsp.edu.br
https://jnd.ifsp.edu.br/index.php/servidores-campus

Referências

[1] F. R. Rosa et al., “Fibrose cística: uma abordagem clínica e nutricional,” *Revista de Nutrição*, pp. 725–737, 2008, doi: 10.1590/S1415-52732008000600011.

[2] M. F. Servidoni et al., “Teste do suor e fibrose cística: panorama da realização do teste em centros públicos e privados do estado de São Paulo,” *J. Bras. Pneumol.*, pp. 121–128, 2017, doi: 10.1590/S1806-37562016000000076.

[3] V. A. Legrys et al., “Diagnostic sweat testing: The Cystic Fibrosis Foundation guidelines,” *The Journal of Pediatrics*, pp. 85–89, Jul. 2007.
[4] C. C. S. Gomez et al., “Iontoforese por pilocarpina: comparação entre o uso de correntes contínua pulsada e contínua constante,” in *XXVI Congresso Brasileiro de Engenharia Biomédica*, 2014, pp. 2028–2931.
[5] E. Cotlove, “Determination of chloride in biological materials,” in *Methods of Biochemical Analysis*, D. Glink, Ed., 1964, doi: 10.1002/9780470110300.
[6] Bioclin, “Instruções Cloretos Colorimétrico.” [Online]. Available: https://quibasa.bioclin.com.br/anexos/INSTRUCOES_CLORETOS_COLORIMETRICO.pdf.
[7] D. M. Zall, D. Fisher, and M. Q. Garner, “Photometric determination of chlorides in water,” *Analytical Chemistry*, pp. 1665–1668, 1956.
[8] J. L. Lezana et al., “Sweat conductivity and chloride titration for cystic fibrosis diagnosis in,” *Journal of Cystic Fibrosis*, pp. 1–7, 2003, doi: 10.1016/S1569-1993(02)00146-7.
[9] Schales and S. S. Schales, “A simple and accurate method for the determination of chloride in biological fluids,” *Journal of Biological Chemistry*, pp. 879–884, 1941, doi: 10.1016/S0021-9258(18)72872-X.
[10] R. A. Athanazio et al., “Diretrizes brasileiras de diagnóstico e tratamento da fibrose cística,” *J. Bras. Pneumol.*, pp. 2019–2045, 2017, doi: 10.1590/S1806-37562017000000065.
[11] Desax et al., “Nanoduct® sweat testing for rapid diagnosis in newborns, infants and children with cystic fibrosis,” *Eur. J. Pediatr.*, pp. 299–304, 2018, doi: 10.1007/s00431-007-0485-0.
[12] Iwasaki et al., “A new spectrophotometric method for the determination of small amounts of chloride using the mercuric thiocyanate method,” *CSJ Journals*, pp. 860–864, 1956.
[13] G. W. Ewing, *Métodos instrumentais de análise química*, vol. 1. Editora Blucher, 1992, p. 312.
[14] Brasil, Ministério da Saúde, Secretaria de Atenção à Saúde, “Portaria nº 224, de 10 de maio de 2010,” Brasília, DF, 2010. [Online]. Available: https://bvsms.saude.gov.br/bvs/saudelegis/sas/2010/prt0224_10_05_2010.html.
[15] Brasil, Ministério da Saúde, “Portaria nº 2.099, de 18 de dezembro de 2015,” Brasília, DF, 2015. [Online]. Available: https://bvsms.saude.gov.br/bvs/saudelegis/gm/2015/prt2099_18_12_2015.html.
[16] Brasil, Ministério da Saúde, Secretaria de Atenção Especializada à Saúde e Secretaria de Ciência, Tecnologia e Inovação e do Complexo Econômico-Industrial da Saúde, “Portaria Conjunta SAES/SECTICS nº 5, de 30 de abril de 2024,” Brasília, DF, 2024. [Online]. Available: https://bvsms.saude.gov.br/bvs/saudelegis/Saes/2024/poc0005_02_05_2024.html.

Heating analysis of a multilayer phantom: skin/ tendon/ bone irradiated by therapeutic ultrasound

Giovanna de Jesus Rocha
Department of Cardiopulmonary and Musculoskeletal Physiotherapy Faculty of Physiotherapy/Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro/ Brazil
gijr2025.1ufrj@gmail.com

Thaís Pionorio Omena
Department of Cardiopulmonary and Musculoskeletal Physiotherapy Faculty of Physiotherapy/Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro/ Brazil
thaisomena@gmail.com.br

Julio Guilherme Silva
Postgraduate Program in Rehabilitation Sciences Faculty of Physiotherapy/Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro/ Brazil
jglsilva@yahoo.com.br

Stela Freitas Raymundo
Department of Biomedical Engineering Technology Center/Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro/ Brazil
stela.freitas@peb.ufrj.br

Wagner Coelho de Albuquerque Pereira
Department of Biomedical Engineering Technology Center/Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro/ Brazil
wagner@peb.ufrj.br

***Abstract*— Therapeutic ultrasound (TUS) is a common treatment modality employed in Physical Therapy practice. One of the most important treatment objectives is to promote pain relief in joints. To this end, laboratory-fabricated tissue-mimicking phantoms representing skin and tendon were developed using plasticized poly (vinyl chloride) (PVCP). TUS was applied to a multilayer tissue-mimicking phantom composed of skin, tendon and cortical bone simulating treatment of the infrapatellar region of the knee. Thermal images were acquired using an infrared camera (Flir systems®) to evaluate the temperature distribution during an ultrasound therapy protocol. Thermal imaging demonstrated a non-uniform temperature distribution across the phantom layers. Greater heating near the cortical bone phantom was observed at 1MHz reaching approximately 60 °C in only 5 minutes of application, suggesting that differences in tissue properties may influence the spatial distribution of heating.**

***Keywords*—ultrasonic therapy, patella, knee, heating.**

I. INTRODUCTION

Phantoms are materials developed to reproduce the physical properties and behavior of biological tissues. Their use has been described in the literature since the 1960s, encompassing applications such as equipment calibration and simulation of different biological tissue characteristics for research and development purposes [1]. Tendons are structures of the musculoskeletal system distributed across various regions of the human body, transmitting the forces produced during muscle contraction. This mechanism enables joint movement and contributes to the stability and efficiency of the body movement [2, 3]. When injured, they lead patients to seek Physiotherapy treatment, where one of the resources is Therapeutic Ultrasound (TUS). Therefore, TUS dosage varies according to the application time, intensity and wave frequency to produce thermal and non-thermal effects [4]. Thermal effects include increase of blood flow, metabolism and extensibility of collagen fibers, muscle spasm reduction and pain relief. Non-thermal effects include cavitation and micromassage which stimulate fibroblast activity, increase protein synthesis and regenerate soft and bone tissues [5, 6, 7].

Although TUS has been used for decades, there is still a lack of robust scientific evidence to support the dosages employed for specific clinical conditions. To date, TUS continues to be used largely on an empirical basis by physiotherapists, as relatively few methodologically rigorous clinical trials have established the effectiveness of TUS parameters.

The objective of this study was to evaluate the TUS protocol proposed for application to the infrapatellar region of the knee, based on the treatment parameters reported in [8], using a multilayer phantom. The protocol consisted of: 1-MHz frequency, continuous mode, spatial average temporal peak (I_{SAPT}), intensity of 1.2 W/cm^2, effective radiating area (ERA) 7 cm^2, and an application time of 8 minutes.

II. METHODOLOGY

The methodology consisted of four main stages: 1) phantom fabrication, 2) evaluation of the acoustic properties and 3) application of TUS on the multilayer phantom.

1) Phantom fabrication

The skin-mimicking phantom was prepared using 100mL of PVCP 2132LP (MF Manufacturing CO., Fort Worth, TX, USA) and 3 g of 140-mesh graphite powder (B'Herzog , Rio de Janeiro, Brazil). The PVCP was mixed in a glass beaker with the graphite powder, which had been previously sieved and weighed [9].

The tendon phantom was constructed at the Ultrasound Laboratory in the Biomedical Engineering Program (COPPE/UFRJ) using plasticized poly (vinyl chloride) – PVCP, obtained as Plastisol resin (M-F Manufacturing Co., Inc, Fort Worth, TX, USA), glycerol, and poly (vinyl chloride) – PVC powder . The proportion of the materials was: 15 wt% glycerol and 10wt% PVC powder [10].

For both phantoms the equipment used included a two-stage vacum pump (Ecotools, model ET190), a glass vacuum desiccator (PHX-210), and a microwave oven (Blue Touch®, Electrolux, model PM023E, Rio de Janeiro, Brazil).

To remove air bubbles, the mixture was placed in a vacuum desiccator for 30 min. It was then heated in a microwave oven at 70% power in a 2-min interval until it reached approximately 180°C. After a 24-h curing period, the phantom's thickness and acoustic properties were assessed at the Ultrasound Laboratory (PEB/COPPE/UFRJ).

The skin phantom is rectangular (14cm length x 14 cm width), thickness 3.55 ± 0.10 mm. The tendon phantom is cylindrical with 6.5 cm perimeter and thickness : 8.17 ± 0.18 mm. The cortical bone phantom is rectangular with 13 cm width x 18 cm length and 1.03 thickness.

Cortical bone Phantom is a commercial Phantom (Epoxi Resin Sawbones©, Vashon, Washington, USA), with parameters specific mass 1700.59 $kg.m^{-3}$, acoustic speed 2924.31 $m.s^{-1}$ and attenuation coefficient 6.15 $dB.cm^{-1}.MHz^{-1}$.

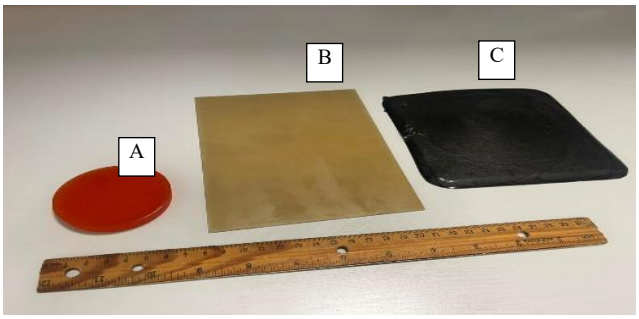


Fig 1: A: Tendon phantom: B: Bone Phantom; C: Skin phantom.
Source: the authors.

2) Evaluation of the acoustic properties

The acoustic properties were measured in an acoustic tank at the laboratory [Fig. 2] and are presented in Table 1, according to [11, 12].

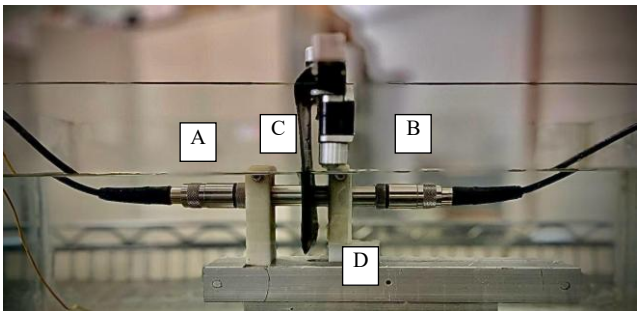


Fig 2: A: Emitter transducer; B: Receiver Transducer: C: Skin phantom; D: Sample holder.
Source: the authors.

Table 1. Acoustic Properties of the skin and tendon Phantoms were measured at the laboratory, while cortical bone phantom presents nominal measures from the manufacturer. S: speed [$m.s^{-1}$] and AC: Attenuation coefficient [$dB.cm^{-1}.MHz^{-1}$].

Skin phantom	S	1491.32 ± 4.83
	AC	0.94 ± 0.01
Tendon phantom	S	1251.47 ± 67.52
	AC	3.94 ± 0.74
Cortical Bone phantom	S	2924.31
	AC	6.15

3) Application of TUS on the multilayer phantom

The present work aims to evaluate the heating promoted by an Ultrasound therapy device (UST) applied on a multilayer phantom. The UST used (Sonopulse III, Ibramed) is shown in Fig 3. The irradiation parameters were: 1MHz-frequency, Effective radiation area (ERA) 8 cm^2, Nominal intensity 1.0 W/cm^2, Nominal power 8.0 W, continuous mode and 5 minutes of application time. The output power was measured ten times using a Radiation force balance (Ohmic instruments), so nominal and estimated power were, respectively: 8.0 W and 8.787 ± 0.017 W.



Fig. 3. TUS equipment (Sonopulse III®, Ibramed Indústria Brasileira de Equipamentos Médicos Ltda., Amparo, SP, Brazil).
Source: the authors.

The reference clinical protocol [8] utilized equipment with an ERA of 7 cm2, an intensity of 1.2 W/cm2, a nominal power of 8.4 W and a dynamic transducer for an application duration of 8 minutes. In the present study, a commercial transducer (Sonopulse III) was employed with an ERA 8 cm2, a nominal intensity of 1.0 W/cm2, and an estimated power of 8.7 W. The application duration was set to 5 minutes, shorter than the time suggested in the clinical protocol, since the laboratory transducer application was stationary, with the transducer secured to support.

The experimental setup was based on the clinical application of TUS on the infrapatellar tendon in the knee [Fig 4].

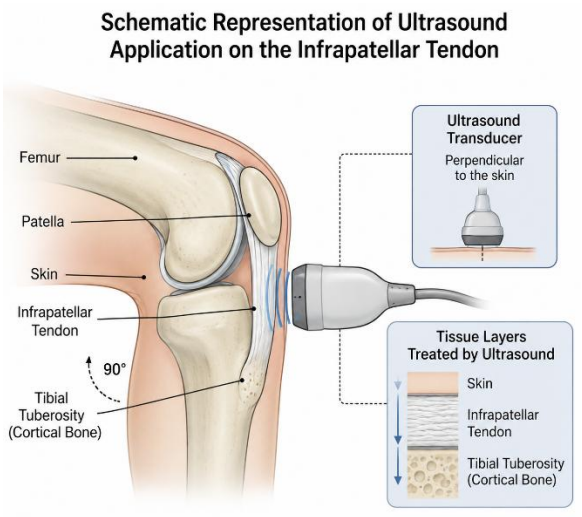


Fig 4: Schematic representation of TUS on the infrapatellar tendon in the knee. Source: the authors.

The experimental set up [Fig 5] consisted of TUS transducer positioned on the skin phantom, coupled with gel. The phantoms interfaces were also filled with gel to minimize the acoustic impedance difference between them. The UST transducer was positioned stationary on the multilayer phantom and the acoustic beam propagated sequentially through the skin, tendon and cortical bone phantoms, before reaching the absorbing medium. After 5 min of irradiation the transducer was moved laterally to allow the infrared camera (FLIR E64501, FLIR SYSTEMS), to capture the heating pattern across each phantom layer. The phantoms were manually separated to enable the acquisition of thermal images of the individual layers. Then, to allow the phantoms to cool between experiments, they were immersed in a beaker containing cold water for 5 min, before the next experimental procedure was performed. A total of five experiments were performed, and the temperature data were analyzed using Excel®.

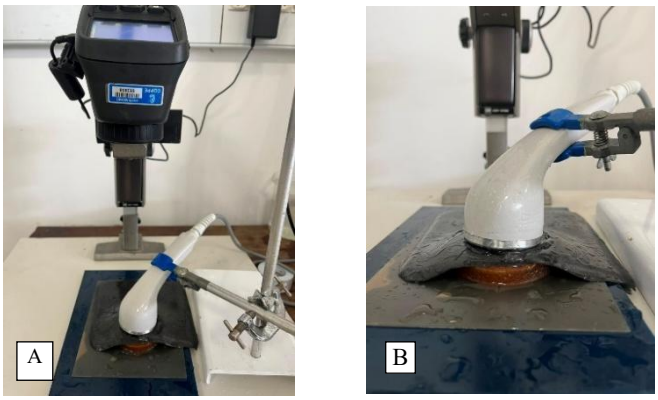


Fig 5: A: Superior view of the experimental set up: infrared camera, ultrasound transducer, multilayer phantom and absorbing medium. B: Anterior view of the ultrasound transducer accoupled with the multilayer phantom (skin/ tendon/ cortical bone phantoms) and absorbing medium). Source: the authors.

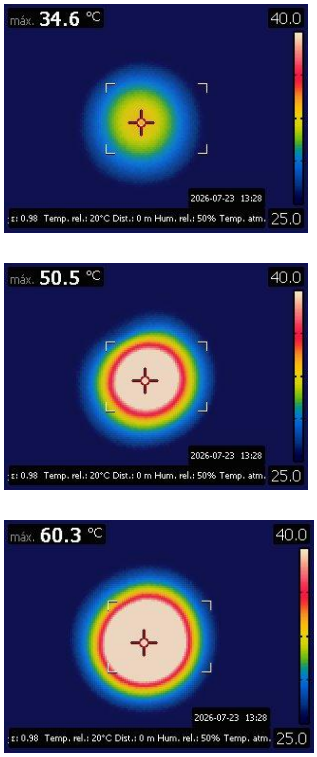
RESULTS

Five experiments were performed on heating the multilayer phantom by the UST (Table 2).

Table 2. Multilayer phantom heating experiments, initial and final temperatures [°C].

Experiments 1	Initial temperature [°C]	Final Temperature [°C]
Skin phantom	23.5	34.8
Tendon phantom	23.5	52.5
Bone phantom	23.5	62.0
2		
Skin phantom	23.5	33.5
Tendon phantom	23.5	48.8
Bone phantom	23.5	64.5
3		
Skin phantom	24.2	34.8
Tendon phantom	24.2	50.5
Bone phantom	24.2	60.3
4		
Skin phantom	24.5	34.6
Tendon phantom	24.5	49.9
Bone phantom	24.5	61.7
5		
Skin phantom	22.8	35.1
Tendon phantom	22.8	65.1
Bone phantom	22.8	72.2

The average temperature after the irradiation for skin phantom was 34.56 ± 0.62 °C, for tendon phantom was 53.36 ± 6.70 °C and for bone phantom was 64.14 ± 4.75 °C. An example of the thermal images is shown in Fig. 6.



Skin phantom

Tendon phantom

Cortical bone phantom

Fig 6: Thermal images registered after irradiation by UST. Temperature can be observed on the left and upper side of the images, temperature of skin phantom was 34.6 °C, tendon phantom was 50.5 °C and cortical bone phantom was 60.3 °C. Source: the author

III. DISCUSSION

The findings of this study demonstrate the thermal behavior of therapeutic ultrasound (TUS) applied to a complex multilayer phantom (skin/ tendon/ cortical bone), providing relevant information about energy deposition in the infrapatellar region of the knee.

The application of PVCP phantoms provides a useful approach for mimicking tissue layers and for investigating the thermal response to TUS, with different dosages [10,12]. This work may provide an estimation of temperature distribution behavior within skin, tendon and bone layers during ultrasound irradiation. However, the findings should be interpreted with caution and should not be directly extrapolated to in vivo conditions. Unlike biological tissues, the phantom does not reproduce blood perfusion, which can contribute to heat dissipation, and the ultrasound transducer was maintained in a stationary position throughout irradiation. Consequently, the localized energy deposition and resulting temperature increase may differ from those occurring during clinical treatment, in which the transducer is typically moved continuously over the treatment area.

For this reason, the irradiation time was reduced to 5 minutes in the present study, rather than the 8-minute application reported in [8], in order to minimize the risk of excessive heating inside the phantoms under stationary irradiation conditions.

The physical stability of PVCP allowed repeated ultrasound exposures and thermal measurements and the present experimental model should therefore be considered a simplified representation of the clinical scenario of the TUS application on the infrapatellar region.

The variability observed across the experimental trials, particularly the higher temperature recorded during the fifth trial (65.1 °C in the tendon and 72.2 °C in the bone) reflects the high sensitivity of the static model to slight operational variations, such as coupling gel thickness, minor fluctuations in applied manual pressure and acoustic beam alignment.

Thermal images revealed a non-uniform temperature distribution along the phantom layers. The skin phantom exhibited the lowest temperature increase, followed by the tendon and cortical bone layers. This finding was expected, given that the attenuation coefficient of the tendon is approximately four times that of the skin , while that of cortical bone is approximately six times higher than that of the skin. Although acoustic attenuation occurs as the ultrasound beam propagates through the tissues, more pronounced heating may occur near the cortical bone region at 1MHz. This behavior may be attributed not only to the higher attenuation coefficient bone, but also to the acoustic impedance mismatch at the tissue interfaces, particularly between the tendon and cortical bone, which may result in partial reflection of the acoustic beam and alter the local distribution of the acoustic energy.

Such behavior highlights a critical clinical concern: target tissues located deep in and close to the periosteal margins may suffer excessive temperature elevations if the spatial distribution of energy is not properly considered.

The obtained values exceed the safe therapeutic limit (40 – 45 °C) and would induce thermal damage in vivo within this range. However, the primary objective of this study was to map the relative heating distribution pattern across interfaces (skin/ tendon/ bone) under a simplified scenario. The phantom model employed lacks blood perfusion and the transducer remained static. Although these physiological and operational factors significantly reduce temperature elevation in actual clinical practice, it is critical to emphasize the importance of continuous transducer movement in clinical settings to prevent tisse damage.

Infrared thermography was selected because it enables direct visualization of the surface thermal pattern. The use of thermocouples would only allow point measurements at the phantom interfaces, which was not the objective of this study. Visualizing the temperature pattern requires fast handling and separation of the layers, so the elapsed time causes thermal loss. The combined use of integrated thermocouples with continuous thermography is planned for subsequent studies.

A randomized crossover clinical study [13] involving 24 healthy participants evaluated high-intensity continuous ultrasound (HICUS) applied to the knee (1MHz, 2.5W/cm^2, 100% duty cycle, 5 minutes). The intervention significantly reduced infrapatellar fat pad stiffness and improved tissue gliding, with effects maintained at subsequent assessments. This study used the same ultrasound frequency, 100% duty cycle and exposure time, but a higher intensity and a moving transducer during irradiation.

Static application represents a simplified, “worst-case” thermal condition. Simulating manual scanning introduces operator-dependent variability. Automating this movement at 4 cm/s using a computerized mechanical setup is already planned in the laboratory. Furthermore, we intend to increase the sample size and compare devices from different manufacturers. Nevertheless, the present study is distinguished by its novelty in evaluating the thermal pattern of a multilayered PVCP phantom simulating the infrapatellar region irradiated by therapeutic ultrasound.

IV. CONCLUSION

The present study demonstrated a non-uniform thermal distribution following TUS application to a multilayer PVCP phantom representing the skin, tendon and cortical bone of the infrapatellar region. The skin exhibited the lowest temperature increase, whereas greater heating was observed in the tendon and cortical bone layers, suggesting that tissue-specific acoustic properties and impedance differences influence the spatial distribution of ultrasound-induced energy.

The PVCP phantom provided a useful controlled model for investigating TUS heating. The low number of repetitions and evaluation of a single TUS dosage is considered a limitation. Further studies should increase the number of experiments, investigate other TUS dosages, and evaluate different skin and tendon thicknesses to better characterize the thermal behavior of TUS under varying conditions. In addition, the transducer movement should be considered, as performed in clinical practice, to better represent the dynamic conditions of therapeutic ultrasound application.

ACKNOWLEDGMENT

To National Council for Scientific and Technological Development (CNPq), Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), and Rio de Janeiro Research Foundation (FAPERJ) for the financial support.

REFERENCES

[1] Culjat, M. O., Goldenberg D., Tewari P. et al. 2010. “A Review of Tissue Substitutes for Ultrasound Imaging”. Ultrasound in Medicine & Biology, v. 36, n.6, pp.861–73.

[2] Guan, G. 2013. “Quantitative Evaluation of Degenerated Tendon Model Using Combined Optical Coherence Elastography and Acoustic Radiation Force. Method”. Journal of Biomedical Optics, v. 18, n. 11.

[3] Wilchez, D. F. T. 2012. Study of the mechanical behavior of ligaments and tendons using optical measurement, M.Sc. thesis, UFSM,Florianópolis, Santa Catarina, Brazil.

[4] Bertolini, G. R. F., Silva T. S., Ciena A. P. et al. 2012. “Comparison of pulsed and continuous ultrasound in rat tendon repair”.Fisioterapia e Pesquisa, vol. 19, no. 3, pp. 242–47.

[5] Speed, C. A. (2001), “Therapeutic ultrasound in soft tissue lesions”, Rheumatology, v.40, p. 1331-1336.

[6] Cameron, M.H., 2022, “Ultrasound”. In: Physical Agents in Rehabilitation: An Evidence-Based Approach to Practice, 6ª ed, cap 9, Portland, EUA, Elsevier.

[7] Silva ANG, de Oliveira JRS, Madureira ÁNM, Lima WA, Lima VLM. Biochemical and Physiological Events Involved in Responses to the Ultrasound Used in Physiotherapy: A Review. Ultrasound Med Biol. 2022 Dec; 48(12):2417-2429. doi: 10.1016/j.ultrasmedbio.2022.07.009. Epub 2022 Sep 15. PMID: 36115728.

[8] de Jesus, J.F., de Albuquerque, T.A.B., Shimba, L.G. *et al.* High-energy dose of therapeutic ultrasound in the treatment of patellar tendinopathy: protocol of a randomized placebo-controlled clinical trial. *BMC Musculoskelet Disord* , **20**, 624 (2019). https://doi.org/10.1186/s12891-019-2993-2

[9] S. F. Raymundo, A. C. R. Silva, E. K. Saito and W. C. A. Pereira, "Ultrasonic Phantoms Made of PVCP and Graphite to Mimic the Thermal Properties of Human Skin," *2026 Global Medical Engineering Physics Exchanges/ Pan American Health Care Exchanges (GMEPE/PAHCE)*, Rio de Janeiro, Brazil, 2026, pp. 1-6, doi: 10.1109/GMEPE/PAHCE69420.2026.11550305.

[10] Gomes, GB, De Matheo, LL, da Silva, FFF et al. Evaluation of materials for the manufacture of models that mimic the acoustic properties of tendons. Res. Biomed. Eng. 37 , 47–52 (2021). https://doi.org/10.1007/s42600-020-00099-z.

[11] Maggi, L. E. 2011. Characterization of the acoustic properties of new materials for the fabrication of ultrasound phantoms. D.Sc. thesis, COPPE/UFRJ, Rio de Janeiro, RJ, Brazil.

[12] Maggi, L.E., Cortela G., von Kruger M. A. et al. 2013. "Ultrasonic attenuation and speed in phantoms made of PVCP and evaluation of acoustic and thermal proprieties of ultrasonic phantoms made of polyvinyl chloride-plastisol (PVCP)". In: Anais do International Work-Conference on Bioinformatics and Biomedical Engineering, Espanha.

[13] Naoto Ikeda, Taizan Fukaya, Yui Sezaki, Sho Toyoshima, Yuka Yokoi, Katsuyuki Morishita,Effects of High-intensity Continuous Ultrasound on Infrapatellar Fat Pad Stiffness and Gliding in Healthy Participants: A Randomized, Single-blind, Placebo-controlled Crossover Trial,Ultrasound in Medicine & Biology,Volume 51, Issue 1,2025,Pages 49-56,ISSN 0301-5629, https://doi.org/10.1016/j.ultrasmedbio.2024.09.006.

Investigation of Protoporphyrin IX Dimethyl Ester Aggregation by Nuclear Magnetic Resonance and Its Impact on the Development of Photosensitizers for Photodynamic Therapy

Sarah Santana Faustino Silva
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, Brasil
sarah.santana524@gmail.com

Adjaci Uchôa Fernandes
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, Brasil
adjaci@hotmail.com

Egberto Munin
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, Brasil
emunin@gmail.com

***Abstract*—Tetrapyrrolic macrocycles constitute a class of compounds widely distributed in nature whose electronic, photochemical, and photophysical properties make them highly relevant for biomedical applications, particularly photodynamic therapy. Protoporphyrin IX has emerged as a promising photosensitizer due to its biocompatibility; however, its intrinsic tendency to aggregate significantly compromises its photodynamic performance. In this study, the aggregation behavior of protoporphyrin IX dimethyl ester in solution was investigated using one- and two-dimensional nuclear magnetic resonance spectroscopy to elucidate how aggregation influences proton chemical shifts and to correlate these variations with the preferred supramolecular organization adopted by the molecule. Experiments were performed at concentrations ranging from 2.82 to 33.8 mmol L^{-1} using an 800 MHz spectrometer. Comparison with previously reported studies revealed inconsistencies in earlier spectroscopic assignments, and the present approach enabled the establishment of more robust proton assignments supported by multidimensional nuclear magnetic resonance correlations. The spectroscopic data indicates the formation of aggregates that cannot be strictly described by classical H-, J-, or T-type organizations but are better represented as hybrid or distorted supramolecular arrangements involving tilted macrocycles. This study provides a high-resolution spectroscopic characterization that contributes to a deeper understanding of porphyrin aggregation phenomena and establishes a structural framework that may support the rational development of next-generation photosensitizers for photodynamic therapy.**

Keywords—photodynamic therapy, protoporphyrin IX, aggregation, nuclear magnetic resonance spectroscopy

I. INTRODUCTION

Porphyrins are a class of tetrapyrrolic macrocycles widely distributed in nature, consisting of four N-heterocyclic pyrrole units interconnected through methine bridges to form a planar π-conjugated ring system [1,2]. These aromatic structures exhibit an 18π-electron delocalization pathway [3], which gives rise to unique electronic, photochemical, and photophysical properties. The ability of porphyrinic macrocycles to participate in photoinduced processes has been extensively exploited in photodynamic therapy (PDT).

Protoporphyrin IX (PpIX), a natural precursor of the heme group, has been widely investigated as a photosensitizer for PDT owing to its intrinsic biocompatibility [4]. Upon visible-light excitation, these molecules are promoted to excited singlet states and subsequently undergo intersystem crossing to the triplet state, which can transfer energy or electrons to molecular oxygen, leading to the generation of reactive oxygen species (ROS) [5]. The resulting oxidative stress promotes cellular damage and cell death, supporting the clinical relevance of PDT, particularly in oncological applications.

Tetrapyrrolic macrocycles exhibit an intrinsic tendency toward aggregation [6]. This behavior arises from their high π-electron density, which favors intermolecular interactions that are energetically more favorable than interactions with the surrounding medium, including biological fluids such as blood circulation [6,7]. In general, aggregation compromises photodynamic performance by promoting non-radiative deactivation pathways of the excited singlet state and reducing the efficiency of intersystem crossing responsible for the formation of the excited triplet state [8].

The aggregation phenomenon results from a delicate balance between attractive σ–π interactions and repulsive π–π interactions between molecules in the aggregated state [9]. According to the proposed model, J-type (slipped face-to-face) and T-type (edge-to-face) arrangements are energetically favored, whereas H-type (face-to-face) organizations maximize π–π repulsion and are therefore less stable [9]. Aggregate formation is influenced by peripheral substituents at the β-pyrrolic and meso positions, as well as by metal coordination at the porphyrin core [10]. Furthermore, physicochemical parameters, including polarity, temperature, pH, and concentration, also modulate this process [6,10].

Throughout the evolution of PDT, several strategies have been proposed to minimize aggregation, including the isolation of naturally occurring derivatives, sterically hindered macrocycle design [11], and nanoencapsulation approaches [12]. Despite these advances, aggregation remains a major limitation, negatively affecting therapeutic performance and hindering structural characterization by spectroscopic techniques, including nuclear magnetic resonance (NMR) spectroscopy.

Porphyrins exhibit pronounced diamagnetic anisotropy, which is reflected in characteristic NMR chemical shifts,

with inner protons typically appearing between −2 and −4 ppm and meso protons resonating between 8 and 12 ppm [13]. Aggregation perturbs this spectroscopic pattern, causing signal broadening and chemical shift variations due to intermolecular interactions that alter the local magnetic environment. Classical studies investigated these effects by systematically varying parameters such as concentration and temperature [14,15].

Recent advances in NMR instrumentation, including multidimensional experiments, pulsed-field gradients, and cryogenically cooled magnets [16], now allow these systems to be revisited with significantly improved resolution and sensitivity, enabling a more detailed description of porphyrin aggregates.

Moreover, a comprehensive understanding of photosensitizer aggregation has become increasingly relevant for biomedical engineering, as this phenomenon directly influences physicochemical stability, singlet oxygen generation, and overall therapeutic performance in photodynamic therapy. These mechanistic insights may ultimately support the rational design of molecular systems with greater translational potential for clinical diagnostic and therapy systems applications.

II. Objective

To investigate the influence of aggregation on the diamagnetic anisotropy of protoporphyrin IX dimethyl ester using nuclear magnetic resonance spectroscopy and to establish spectroscopic correlations that may contribute to the rational development of photosensitizers for photodynamic therapy applications.

III. Materials and Methods

A. General Information

This study was conducted within a multidisciplinary research environment involving the Centro de Inovação, Tecnologia e Educação (CITE), Resix Invent, and the Central Analítica do Instituto de Química da Universidade de São Paulo. The synthetic strategies employed are well established in the literature [11] and are part of the expertise developed by the research group affiliated with the Center for Research on Redox Processes in Biomedicine (REDOXOME – CEPID-FAPESP, Process No. 2013/07937-8).

B. Synthesis of Protoporphyrin IX Dimethyl Ester (PpIX DME)

Disodium protoporphyrin IX salt (1.00 g, ≥ 90%) was dissolved in 20 mL of a methanol/sulfuric acid solution (19:1, v/v) and stirred at room temperature (25 °C) for 24 h under light-protected conditions. The reaction mixture was extracted with dichloromethane (100 mL) and subsequently neutralized by the slow addition of ammonium hydroxide solution (150 mL, 28–30%) in an ice bath until pH 11 was reached. Gentle stirring was maintained for 10 min to avoid emulsion formation. The organic phase was extracted three times with dichloromethane (10 mL), dried over anhydrous sodium sulfate, filtered, and concentrated under reduced pressure. The resulting product was precipitated with methanol and stored protected from light and heat.

C. Sample Preparation

Twelve PpIX DME solutions were prepared in deuterated chloroform containing 0.05% (v/v) tetramethylsilane ($CDCl_3$/TMS). Concentrations ranged from 2.82 to 33.8 mmol L^{-1} while maintaining a constant final volume of 600 µL for all samples. The initial sample was prepared by dissolving 1.00 mg of PpIX DME in 600 µL of $CDCl_3$/TMS, corresponding to a concentration of 2.82 mmol L^{-1} (PpIX DME 1). Subsequent samples were prepared cumulatively by adding successive 1.00 mg portions of PpIX DME to the preceding solution while maintaining a constant total volume until the maximum concentration of 33.8 mmol L^{-1} (PpIX DME 12) was reached.

D. NMR Experiments

One-dimensional proton nuclear magnetic resonance (^{1}H NMR) spectroscopy and two-dimensional homonuclear correlation experiments, including correlation spectroscopy (^{1}H–^{1}H COSY) and nuclear Overhauser effect spectroscopy (^{1}H–^{1}H NOESY), were performed using a Bruker Avance III spectrometer (Germany, 2012) operating at 800 MHz and 25 °C.

One-dimensional ^{1}H NMR spectra, as well as two-dimensional ^{1}H–^{1}H COSY and ^{1}H–^{1}H NOESY spectra, were acquired for PpIX DME 1 and PpIX DME 12. For the intermediate concentrations, only one-dimensional ^{1}H NMR spectra were recorded.

IV. Results and Discussion

The ^{1}H NMR spectra acquired at different concentrations exhibited systematic variations in both chemical shifts and proton assignments, demonstrating that the magnetic environment is strongly modulated by aggregation phenomena. The molecular structure and proton numbering scheme of PpIX DME are shown in Figure 1 to facilitate the interpretation of the spectroscopic data discussed throughout this section.

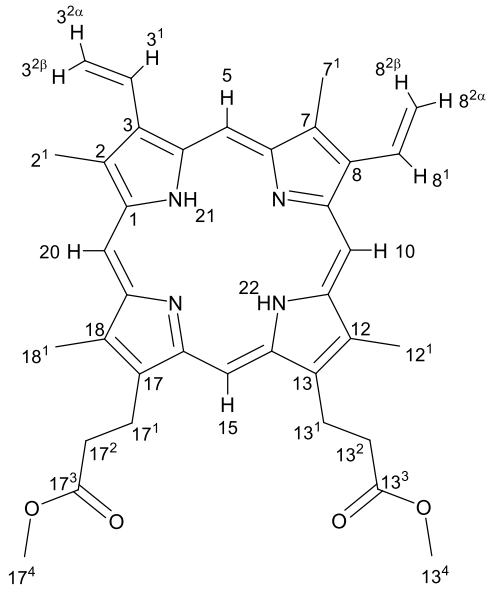


Figure 1. Molecular structure and proton numbering scheme of PpIX DME used for NMR assignments throughout this study.

Complete proton assignments are presented in Table 1 and were established through the combined interpretation of one-dimensional ^{1}H NMR and two-dimensional ^{1}H–^{1}H COSY and ^{1}H–^{1}H NOESY experiments. Chemical shift variations (Δδ) were quantified by calculating the differences between the proton resonances observed at the lowest and highest concentrations investigated.

Table 1. Proton assignments and concentration-dependent chemical shift variations of PpIX DME.

Group	Position	PpIX DME 1	PpIX DME 12	Δδ
Meso	H-5	10.23	9.90	0.33
	H-10	10.18	9.93	0.25
	H-15	10.04	9.81	0.23
	H-20	10.09	9.78	0.31
Vinyl	H-3^1	8.29	8.10	0.19
	H-8^1	8.29	8.15	0.14
	H-3$^{2\beta}$ H-8$^{2\beta}$	6.38	6.28	0.10
	H-3$^{2\alpha}$ H-8$^{2\alpha}$	6.19	6.12	0.07
Methylene	H-13^1 H-17^1	4.41	4.29	0.12
	H-13^2 H-17^2	3.28	3.21	0.07
Methyl	H-2^1	3.72	3.55	0.17
	H-7^1	3.71	3.54	0.17
	H-12^1	3.62	3.51	0.11
	H-18^1	3.64	3.49	0.15
Methoxy	H-13^4 H-17^4	3.66 3.66	3.65 3.65	0.01
Inner	H-21 H-22	-3.68	-4.19	0.51

The present results demonstrate that several proton resonances are significantly affected by aggregation and undergo signal permutation as concentration increases. Although previous studies have reported that most proton assignments of PpIX DME could be readily achieved using only one-dimensional ^{1}H NMR without the need for chemical shift reagents or specialized solvents [15], the current findings reveal important limitations in this approach when applied to a complex tetrapyrrolic macrocycle undergoing concentration-dependent aggregation.

The largest upfield chemical shift variations were observed for the inner pyrrolic protons H-21 and H-22, and the meso protons H-5 and H-20. This behavior indicates that nuclei located near the center of the macrocycle are particularly sensitive to changes in the magnetic environment induced by aggregation. These observations reflect the diamagnetic anisotropy generated by the aromatic ring current of the tetrapyrrolic system, whose influence becomes more pronounced as intermolecular interactions intensify.

Interestingly, the meso protons exhibited signal permutation throughout the concentration series, as illustrated in Figure 2. This behavior suggests that the aggregation process does not strictly follow classical J-, T-, or H-type supramolecular organizations. Instead, the spectroscopic evidence supports the formation of a hybrid arrangement in which different regions of the porphyrin macrocycle experience distinct magnetic environments.

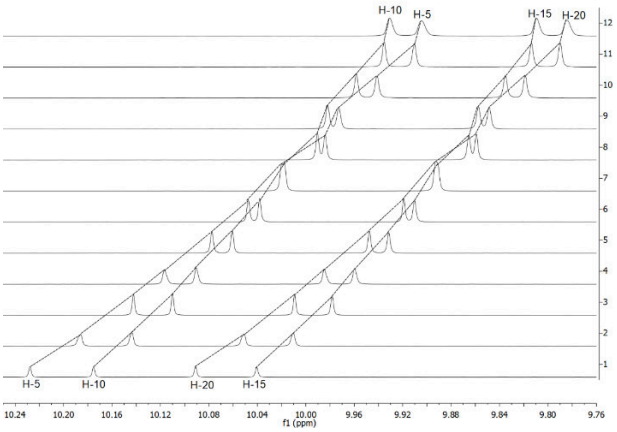


Figure 2. Concentration-dependent changes in the chemical shifts of meso protons.

The methyl groups also exhibited concentration-dependent signal permutation (Figure 3). Furthermore, the data indicates that aggregation-induced perturbations gradually decrease as the structural distance from the macrocyclic core increases. This trend is particularly evident when comparing the methyl and methoxy substituents.

Among the methyl groups, H-2^1 and H-7^1 exhibited the largest chemical shift variations. Their enhanced sensitivity may be attributed to their proximity to the vinyl substituents, which contribute to electronic delocalization within the macrocycle and amplify the influence of intermolecular interactions.

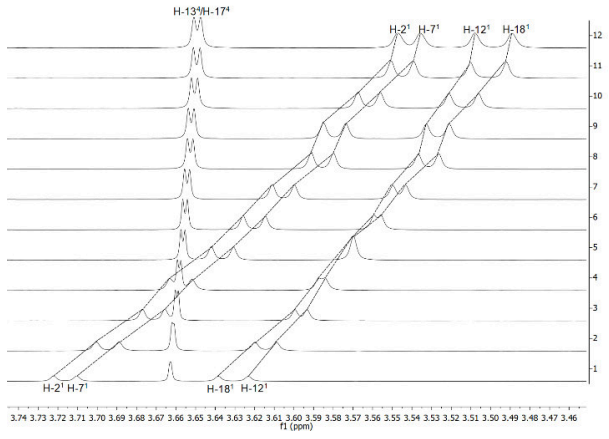


Figure 3. Concentration-dependent changes in the chemical shifts of methyl and methoxy groups.

In contrast, the methoxy protons H-13^4 and H-17^4 displayed minimal chemical shift variations, indicating a relatively low sensitivity to aggregation-induced perturbations. Nevertheless, these groups provide additional

insight into the supramolecular organization of the system. The progressive increase in signal separation observed with increasing concentration suggests the existence of steric repulsion at specific intermolecular contact regions.

This interpretation is consistent with previously reported crystallographic observations for PpIX DME [17] and further supports the presence of a hybrid supramolecular arrangement involving tilted macrocycles in the aggregated state.

V. Conclusion

The comprehensive nuclear magnetic resonance spectroscopic analysis presented in this study enabled the unequivocal assignment of proton resonances in protoporphyrin IX dimethyl ester, overcoming ambiguities previously reported in the literature. The observed variations in proton chemical shifts were strongly influenced by aggregation-induced diamagnetic anisotropy associated with the aromatic tetrapyrrolic system, highlighting the sensitivity of these nuclei to changes in the local magnetic environment. The spectroscopic evidence obtained throughout the concentration series indicates that the aggregated state of protoporphyrin IX dimethyl ester cannot be adequately described by idealized H-, J-, or T-type supramolecular organizations. Instead, the data supports the formation of hybrid supramolecular arrangements involving relative tilting between neighboring macrocycles, revealing the structural complexity of intermolecular interactions in solutions. These findings reinforce the capability of high-field nuclear magnetic resonance spectroscopy to elucidate subtle structural changes associated with porphyrin aggregation and establish a robust spectroscopic framework for investigating aggregation-prone aromatic systems. To complement these findings, computational simulations are envisioned as future work to investigate the structural organization of the porphyrinic dimer and further validate the proposed geometric distortion. From a biomedical engineering perspective, this study provides relevant contributions to the structural characterization of photosensitizers intended for photodynamic therapy. The methodological approach described herein may serve as a reference for the investigation of other porphyrin-based systems, while the mechanistic insights obtained may support the rational design of next-generation photosensitizers with improved physicochemical stability and enhanced therapeutic performance. Ultimately, a deeper understanding of aggregation phenomena may contribute to the development of more efficient molecular platforms for clinical diagnostic and therapy systems, thereby supporting the advancement of photodynamic therapy toward more predictable and clinically translatable applications.

Conflict of Interest

The authors declare that they have no known financial interests or personal relationships that could have appeared to influence the work reported in this study.

Acknowledgements

S.S.F. Silva acknowledges the financial support from the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES), Finance Code 001. S.S.F. Silva also acknowledges the Central Analítica do Instituto de Química da Universidade de São Paulo for providing access to the analyses and Resix Invent for its collaboration and support. E. Munin and A.U. Fernandes acknowledge the support of Ânima Institute (AI) and Universidade Anhembi Morumbi. A.U. Fernandes also acknowledges the São Paulo Research Foundation (FAPESP) for financial support (Grant No. 2016/16380-3).

References

[1] A. M. Udrea *et al.*, "Photosensitizers-Loaded Nanocarriers for Enhancement of Photodynamic Therapy in Melanoma Treatment," Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/pharmaceutics15082124.

[2] S. Hiroto, Y. Miyake, and H. Shinokubo, "Synthesis and Functionalization of Porphyrins through Organometallic Methodologies," Feb. 22, 2017, *American Chemical Society*. doi: 10.1021/acs.chemrev.6b00427.

[3] R. Boscencu *et al.*, "Porphyrin Macrocycles: General Properties and Theranostic Potential," Feb. 01, 2023, *MDPI*. doi: 10.3390/molecules28031149.

[4] Y. Zou, J. Chen, Y. Qu, X. Luo, W. Wang, and X. Zheng, "Evolution of nMOFs in photodynamic therapy: from porphyrins to chlorins and bacteriochlorins for better efficacy," 2025, *Frontiers Media SA*. doi: 10.3389/fphar.2025.1533040.

[5] B. Myrzakhmetov, P. Arnoux, S. Mordon, S. Acherar, I. Tsoy, and C. Frochot, "Photophysical properties of protoporphyrin IX, pyropheophorbide-a and photofrin® in different conditions," *Pharmaceuticals*, vol. 14, no. 2, pp. 1–21, Feb. 2021, doi: 10.3390/ph14020138.

[6] B. M. Estevão *et al.*, "Synthetic chlorin derivative self-prevented from aggregation: Behavior in homogeneous medium for PDT applications," *J. Mol. Liq.*, vol. 320, Dec. 2020, doi: 10.1016/j.molliq.2020.114363.

[7] R. Teixeira, V. V. Serra, D. Botequim, P. M. R. Paulo, S. M. Andrade, and S. M. B. Costa, "Fluorescence spectroscopy of porphyrins and phthalocyanines: Some insights into supramolecular self-assembly, microencapsulation, and imaging microscopy," Jul. 02, 2021, *MDPI AG*. doi: 10.3390/molecules26144264.

[8] G. G. Parra *et al.*, "Effects of meso-tetrakis (4-sulfonatophenyl) porphyrin (TPPS4) aggregation on its spectral and kinetic characteristics and singlet oxygen production," *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, vol. 261, Nov. 2021, doi: 10.1016/j.saa.2021.120063.

[9] C. A. Hunter and J. K. M. Sanders, "The Nature of π-π Interactions," *J. Am. Chem. Soc.*, vol. 112, no. 14, pp. 5525–5534, 1990, doi: 10.1021/ja00170a016.

[10] G. Magna, D. Monti, C. Di Natale, R. Paolesse, and M. Stefanelli, "The assembly of porphyrin systems in well-defined nanostructures: An update," Nov. 26, 2019, *MDPI AG*. doi: 10.3390/molecules24234307.

[11] A. F. Uchoa, K. T. De Oliveira, M. S. Baptista, A. J. Bortoluzzi, Y. Iamamoto, and O. A. Serra, "Chlorin photosensitizers sterically designed to prevent self-aggregation," *Journal of Organic Chemistry*, vol. 76, no. 21, pp. 8824–8832, Nov. 2011, doi: 10.1021/jo201568n.

[12] L. M. Rossi, P. R. Silva, L. L. R. Vono, A. U. Fernandes, D. B. Tada, and M. S. Baptista, "Protoporphyrin IX nanoparticle carrier: Preparation, optical properties and singlet oxygen generation," *Langmuir*, vol. 24, no. 21, pp. 12534–12538, Nov. 2008, doi: 10.1021/la800840k.

[13] Scheer, H.; Katz, J. J. Nuclear magnetic resonance spectroscopy of porphyrins and metalloporphyrins. In: Smith, K. M. Porphyrins and metalloporphyrins. Amsterdam: Elsevier Scientific Publishing Company. Cap.10, p.399-524, 1975.

[14] R. J. Abraham, S. C. M. Fell, H. Pearson, and K. M. Smith, "The nmr spectra of porphyrins—15," *Tetrahedron*, vol. 35, no. 14, pp. 1759–1766, Jan. 1979, doi: 10.1016/0040-4020(79)88005-9.

[15] T. R. Janson and J. J. Katz, "An Examination of the 220 MHz NMR Spectra of Mesoporphyrin IX Dimethyl Ester, Deuteroporphyrin IX Dimethyl Ester, and Protoporphyrin IX Dimethyl Ester*," 1972.

[16] I. Gjuroski, J. Furrer, and M. Vermathen, "Probing the interactions of porphyrins with macromolecules using NMR

spectroscopy techniques," Apr. 01, 2021, *MDPI AG.* doi: 10.3390/molecules26071942.

[17] W. S. Caughey and J. A. Ibers, "Crystal and molecular structure of the free base porphyrin, protoporphyrin IX dimethyl ester," *J. Am. Chem. Soc.*, vol. 99, no. 20, pp. 6639–6645, Sep. 1977, doi: 10.1021/ja00462a027.

Monitor de Prevenção de Pneumonia Associada à Ventilação Mecânica: Arquitetura de Hardware, Instrumentação e Protocolo de Verificação

Lúcio Bareta Todorov	Ulisses Matiola Todorov	Eduardo Agostini Pasold	Jonas Souza
Engenharia de Produto	*Engenharia de Produto*	*Engenharia de Produto*	*Engenharia de Produto*
CSE Equipamentos e Serviços Ltda.	*CSE Equipamentos e Serviços Ltda.*	*CSE Equipamentos e Serviços Ltda.*	*CSE Equipamentos e Serviços Ltda.*
Blumenau, Brasil	*Blumenau, Brasil*	*Blumenau, Brasil*	*Blumenau, Brasil*
lucio@cseonline.com.br	ulisses@cseonline.com.br	eduardo.pasold@cseonline.com.br	suporte.tecnico@cseonline.com.br

Resumo—Este artigo descreve a arquitetura de hardware e a instrumentação de um monitor multiparamétrico destinado a apoiar as medidas de prevenção da pneumonia associada à ventilação mecânica (PAV) em unidades de terapia intensiva (UTI) que atendem pacientes neonatais, pediátricos e adultos. O equipamento integra, em uma única plataforma embarcada, três subsistemas em malha fechada: controle da pressão do balonete do tubo endotraqueal, aspiração subglótica contínua e temporização da mudança de decúbito, além de uma interface que parametriza as faixas de ajuste conforme o perfil do paciente. O projeto adota como requisitos de entrada a norma geral ABNT NBR IEC 60601-1 e suas normas colaterais aplicáveis, e estabelece como requisito de desempenho um desvio inferior a 5% entre o valor medido e o valor alvo em cada módulo. São apresentados o protocolo de verificação em bancada — grandezas avaliadas, condições por perfil de paciente, tamanho amostral e número de ciclos — e as planilhas de registro correspondentes. São apresentados os registros da campanha de verificação interna conduzida segundo esse protocolo. Os ensaios normativos em laboratório acreditado e a avaliação clínica encontram-se em execução; consequentemente, este trabalho não apresenta resultados de conformidade normativa nem desfechos assistenciais. Este desenvolvimento foi apoiado pela Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) por meio do edital Impulsiona SC 2024.

Termos Índice—arquitetura de hardware, dispositivos médicos, instrumentação biomédica, pneumonia associada à ventilação mecânica, unidade de terapia intensiva, verificação em bancada.

I. Introdução

O suporte ventilatório invasivo consolidou-se, ao longo das últimas cinco décadas, como recurso central no tratamento da insuficiência respiratória. A aplicação de pressão positiva às vias aéreas por meio de uma prótese traqueal associa-se, contudo, a efeitos adversos reconhecidos: instabilidade hemodinâmica, sobretudo em pacientes hipovolêmicos; redução dos mecanismos locais de defesa, com aumento da frequência de infecções respiratórias; e lesão induzida pela ventilação mecânica, relacionada às pressões alveolares atingidas em pulmões heterogêneos [1].

Entre esses efeitos, a pneumonia associada à ventilação mecânica (PAV) destaca-se pela frequência e pela gravidade no ambiente de terapia intensiva [1], [2]. Trata-se de infecção relacionada à assistência à saúde cujo diagnóstico permanece heterogêneo entre serviços e protocolos, o que dificulta a comparação direta entre estudos epidemiológicos e a consolidação de séries nacionais [2], [4].

A letalidade associada à PAV varia amplamente entre as coortes publicadas, com relatos que se distribuem aproximadamente entre 20% e 60%, a depender da população estudada, do critério diagnóstico adotado e da gravidade basal dos pacientes [3], [4]. A evolução para insuficiência respiratória grave e sepse é a via clínica mais temida.

Do ponto de vista econômico, a PAV prolonga o tempo de ventilação mecânica e de internação e concentra parcela expressiva do consumo de antimicrobianos nas unidades de terapia intensiva, estimada em cerca de metade de todo o uso de antibióticos nessas unidades, o que contribui para a pressão seletiva e para a resistência bacteriana no ambiente hospitalar [2], [4].

As medidas preventivas consolidadas em protocolos assistenciais combinam: (i) elevação da cabeceira do leito entre 30° e 45°, para reduzir o risco de aspiração de conteúdo gástrico; (ii) aspiração das secreções acumuladas no espaço subglótico; (iii) manutenção da pressão do balonete do tubo endotraqueal dentro da faixa recomendada [5], [6]; e (iv) mudança programada de decúbito, com finalidade de mobilização e alívio de áreas pulmonares dependentes [9], [10]. Nas populações neonatal e pediátrica, a indicação de cânulas com balonete e a respectiva faixa de ajuste de pressão apresentam particularidades documentadas na literatura e desigualdade de adoção entre os serviços brasileiros [7], [8].

Este artigo tem por objetivo descrever a arquitetura de hardware e a instrumentação do monitor desenvolvido e formalizar o protocolo de verificação em bancada de cada módulo, incluindo as planilhas de registro das medições. Não são objeto deste trabalho a

demonstração de conformidade normativa, que depende de ensaios executados por laboratório acreditado, nem a avaliação de desfechos clínicos, que depende de estudo controlado específico.

II. Arquitetura do Equipamento

O equipamento é organizado em uma plataforma embarcada única, que hospeda três subsistemas funcionais e uma interface homem-máquina (IHM) de tela sensível ao toque. Cada subsistema executa uma malha de controle independente sobre a grandeza que lhe compete, com aquisição contínua, comparação com o valor alvo ajustado pelo operador e atuação corretiva. A Tabela I consolida a especificação de instrumentação e de atuação de cada módulo; os campos de faixa de medição, resolução e incerteza são resultados de medição a partir dos componentes empregados no projeto.

A. Módulo de controle da pressão do balonete

O módulo mantém a pressão interna do balonete do tubo endotraqueal no valor alvo definido pelo profissional assistencial, dentro da faixa habilitada para o perfil de paciente selecionado. A pressão é adquirida continuamente por transdutor acoplado à linha de insuflação; o erro em relação ao valor alvo é processado pelo controlador embarcado, que comanda a unidade de insuflação e a válvula de alívio. O requisito de projeto estabelece desvio inferior a 5% entre a pressão medida e a pressão alvo em regime permanente. Os limites de alarme são configuráveis pelo operador (Fig. 1) e tratados conforme os requisitos de sistemas de alarme indicados na Seção III.

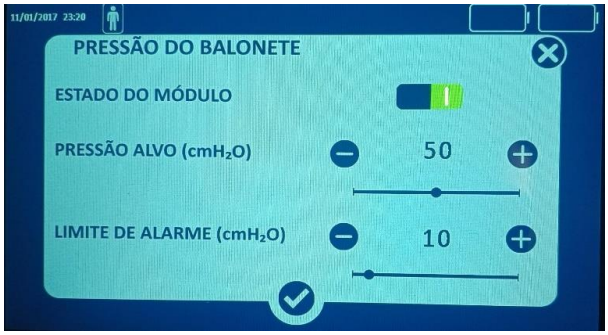


Fig. 1 meramente ilustrativa. Tela de ajuste do controle de pressão do balonete.

B. Módulo de aspiração subglótica

O módulo aplica pressão negativa regulada à porta de aspiração subglótica do tubo endotraqueal, mantendo aspiração contínua no valor alvo definido para o perfil de paciente selecionado. A estrutura de controle é análoga à do módulo anterior: medição contínua, comparação com o valor alvo e atuação sobre a fonte de vácuo, com requisito de projeto de desvio inferior a 5% em regime permanente e limites de alarme configuráveis (Fig. 2).

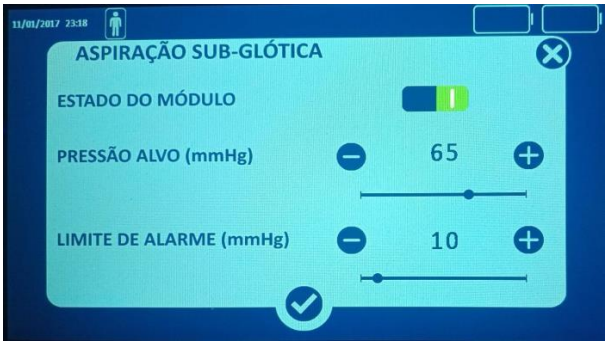


Fig. 2 meramente ilustrativa. Tela de ajuste da aspiração subglótica.

A pesquisa de anterioridade conduzida pelos autores não identificou equipamento eletrônico de uso corrente que execute o controle contínuo dessa variável de forma integrada; as soluções localizadas operam por meios mecânicos ou por aspiração intermitente manual. Essa observação refere-se ao escopo da busca realizada e não constitui levantamento exaustivo de mercado.

C. Módulo de decúbito do paciente

O módulo controla a rotação do paciente sobre o próprio eixo longitudinal, por meio de estrutura aplicada ao leito, com intervalo de alteração ajustável pelo operador (Fig. 3). A função é habilitada para os perfis pediátrico e adulto. O requisito de projeto estabelece desvio inferior a 5% tanto para o ângulo de posicionamento quanto para a temporização do ciclo programado.

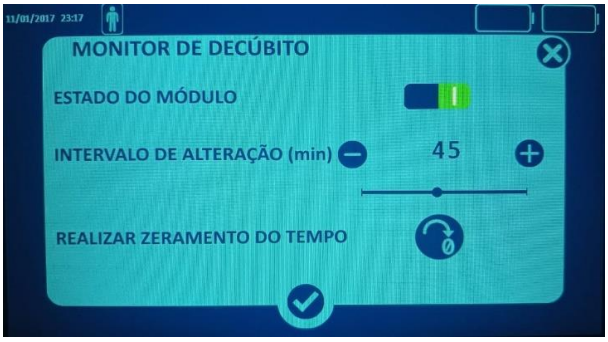


Fig. 3 meramente ilustrativa. Tela de ajuste do módulo de decúbito.

D. Interface integradora multiparamétrica

A operação inicia-se pela seleção do perfil de paciente — neonatal, pediátrico ou adulto (Fig. 4) —, que define as faixas de ajuste e os limites de alarme habilitados em cada módulo. Essa parametrização por faixa etária está alinhada às diferenças de indicação e de ajuste descritas para cânulas com balonete nas populações neonatal e pediátrica [7], [8], e pode ser revista pelo profissional assistencial durante a evolução do paciente.



Fig. 4 meramente ilustrativa. Tela de seleção do perfil do paciente.

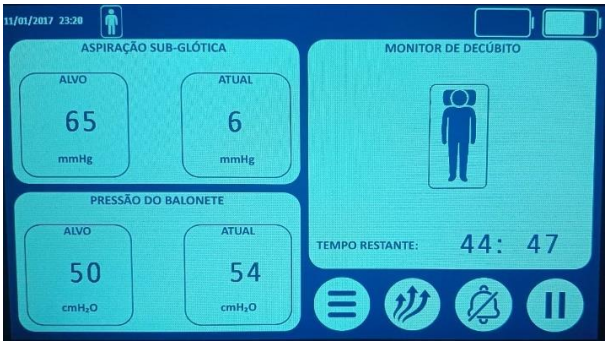


Fig. 5 meramente ilustrativa. Tela funcional de uso, com os três módulos integrados.

A tela funcional de uso (Fig. 5) apresenta simultaneamente os valores alvo e medidos dos módulos de aspiração subglótica e de pressão do balonete, o estado do módulo de decúbito e o tempo restante para a próxima mudança de posição. A Fig. 6 mostra a vista frontal do protótipo montado.



Fig. 6 meramente ilustrativa. Vista frontal do protótipo.

TABELA I
ESPECIFICAÇÃO DE INSTRUMENTAÇÃO E ATUAÇÃO POR MÓDULO

Módulo	Grandeza controlada	Faixa de medição	Erro (%)	Incerteza declarada	K
Pressão do balonete	Pressão positiva	0 – 100cmH_2O	0,001	0,02cmH_2O	2,00
Aspiração subglótica	Pressão negativa	0 – 110mmHg	0,004	0,02mmHg	2,00
Decúbito	Ângulo / tempo	360°	0,001	0,02°	2,00

k é o fator de abrangência da incerteza expandida.

III. Normatizações Aplicáveis

O projeto adota como requisitos de entrada as normas técnicas publicadas pela Associação Brasileira de Normas Técnicas (ABNT) na série Norma Brasileira (NBR), correspondentes às normas da International Electrotechnical Commission (IEC — Comissão Eletrotécnica Internacional). Padroniza-se, neste artigo, a designação ABNT NBR IEC 60601-1 para a norma geral e ABNT NBR IEC 60601-1-X para as normas colaterais, nas edições vigentes na data de submissão. A Tabela II relaciona cada norma, o respectivo objetivo regulatório e o ponto do projeto em que o requisito incide.

Registra-se que as normas foram empregadas como requisitos de projeto e como referência para os ensaios internos de verificação descritos na Seção IV. A demonstração de conformidade depende de ensaios executados por laboratório acreditado, com emissão dos respectivos relatórios, os quais não haviam sido concluídos até a submissão deste trabalho.

TABELA II
NORMAS ADOTADAS, OBJETIVO REGULATÓRIO E INCIDÊNCIA NO PROJETO

Norma	Escopo	Objetivo regulatório	Incidência no projeto
ABNT NBR IEC 60601-1	Equipamento eletromédico — requisitos gerais para segurança básica e desempenho essencial	Estabelecer os requisitos gerais de segurança elétrica, mecânica e térmica e exigir a definição formal do desempenho essencial do equipamento	Isolação, correntes de fuga e proteção contra choque; definição do desempenho essencial de cada um dos três módulos
ABNT NBR IEC 60601-1-2	Norma colateral: compatibilidade eletromagnética	Estabelecer níveis de emissão e de imunidade a perturbações eletromagnéticas no ambiente de uso previsto	Qualificação do conjunto eletrônico embarcado e da fiação de sensores e atuadores
ABNT NBR IEC 60601-1-6	Norma colateral: usabilidade (engenharia de fatores humanos)	Exigir a aplicação de um processo de engenharia de usabilidade que reduza o risco de erro de uso	Projeto da IHM, fluxo de seleção do perfil de paciente e de ajuste dos valores alvo
ABNT NBR IEC 60601-1-8	Norma colateral: sistemas de alarme	Padronizar categorias de prioridade, sinalização audível e visual e comportamento das condições de alarme	Alarmes de desvio de pressão do balonete e de pressão de aspiração subglótica
ABNT NBR IEC 60601-1-9	Norma colateral: projeto ecologicamente consciente	Exigir a consideração do impacto ambiental ao longo do ciclo de vida do equipamento	Seleção de materiais, embalagem e orientações de descarte
ABNT NBR IEC 60601-1-10	Norma colateral: controladores fisiológicos em malha fechada	Estabelecer requisitos para o desenvolvimento e a validação de controladores que atuam sobre variável fisiológica em malha fechada	Malhas de controle dos módulos de pressão do balonete e de aspiração subglótica

IV. Protocolo de Verificação em Bancada

Esta seção define o protocolo de ensaios internos de verificação, cuja execução gera os registros apresentados na Seção V. O protocolo é aplicado a cada módulo isoladamente e, em uma segunda etapa, ao equipamento integrado, em cada um dos perfis de paciente habilitados.

A. Grandezas avaliadas

- Erro em regime permanente: diferença percentual entre o valor medido pelo padrão de referência e o valor alvo ajustado, após estabilização.
- Estabilidade (deriva): variação do valor medido em ensaio contínuo de 24 h, mantidas constantes as condições de ensaio.
- Tempo de resposta: intervalo t90 entre a aplicação de um degrau no valor alvo e o alcance de 90% do novo valor.
- • Repetibilidade: desvio-padrão das medições obtidas sob condições de ensaio idênticas e repetidas.
- Desempenho do sistema de alarme: latência entre a instalação da condição de alarme e a respectiva sinalização, e contagem de falsos positivos e falsos negativos.
- Exatidão do controle de decúbito: erro angular em relação ao ângulo alvo e erro de temporização do intervalo programado.

B. Condições de ensaio

Cada grandeza é avaliada nos perfis neonatal, pediátrico e adulto, sobre simulador de via aérea e pulmão de teste com balonete de características declaradas. Para o módulo de decúbito, utiliza-se carga instrumentada representativa da massa de cada perfil. A Tabela III consolida a matriz de ensaios, incluindo os campos de tamanho amostral e de número de ciclos, que devem ser declarados para cada condição antes do início da campanha.

C. Rastreabilidade metrológica

As medições de referência são obtidas com padrões calibrados e rastreáveis, cujos dados de identificação, faixa, incerteza e validade de certificado são registrados na Tabela IV. A incerteza do padrão deve ser, no mínimo, quatro vezes menor que a tolerância verificada, de modo que o requisito de 5% possa ser avaliado com margem adequada.

D. Tratamento dos dados

Para cada condição de ensaio são calculados a média, o desvio-padrão, o valor máximo absoluto do erro e o intervalo de confiança de 95%. A conformidade ao requisito de projeto é declarada quando o limite superior do intervalo de confiança do erro permanece inferior a 5% do valor alvo. Os resultados são reportados sempre acompanhados do tamanho amostral e do número de ciclos executados.

TABELA III
MATRIZ DE ENSAIOS POR MÓDULO E PERFIL DE PACIENTE

Módulo	Perfil	Valor alvo ajustado	n	Ciclos	Duração do ensaio
Pressão do balonete	Neonatal	15cmH_2O	20	5	24 horas
Pressão do balonete	Pediátrico	25cmH_2O	20	5	24 horas
Pressão do balonete	Adulto	40cmH_2O	20	5	24 horas
Aspiração subglótica	Neonatal	2mmHg	20	5	24 horas
Aspiração subglótica	Pediátrico	10mmHg	20	5	24 horas
Aspiração subglótica	Adulto	50mmHg	20	5	24 horas
Decúbito	Pediátrico	90°	20	24	24 horas
Decúbito	Adulto	90°	20	24	24 horas

O módulo de decúbito não é habilitado para o perfil neonatal.

TABELA IV
PADRÕES DE REFERÊNCIA E RASTREABILIDADE METROLÓGICA

Grandeza de referência	Padrão / equipamento	Modelo	Faixa	Incertez a (%)	Certificado / validade
Pressão positiva	Vacuômetro digital	VDR-920	1500mbar	0,002	R738/25 – 07/2027
Pressão negativa	Vacuômetro digital	VDR-920	-1500mbar	0,002	R738/25 – 07/2027
Ângulo	Medidor de ângulo	935DAA	360°	0,001	R715/25 – 01/2027
Tempo	Cronômetro digital	INS-1338	24h	0,001	R766/26 – 01/2028

V. Consolidação da Arquitetura e Registro dos Ensaios

A. Fundamentação bibliográfica das funções implementadas

Os parágrafos desta subseção sintetizam evidências da literatura que fundamentaram a escolha das três funções implementadas. Trata-se de fundamentação bibliográfica: os achados descritos referem-se aos estudos citados e não constituem resultados obtidos com o equipamento apresentado neste artigo.

A manutenção da pressão do balonete dentro da faixa recomendada é associada, na literatura, à redução do extravasamento de secreções contaminadas para as vias aéreas inferiores, que ocorre quando a pressão é insuficiente, e à prevenção da lesão isquêmica da mucosa traqueal, associada a pressões excessivas [5], [6]. O monitoramento periódico ou contínuo por manômetro ou sistema automatizado é descrito como fator de redução da microaspiração [1], [3], e a variabilidade de adoção de protocolos e de instrumentos de medida entre serviços brasileiros foi documentada tanto em unidades de adultos [5], [6] quanto em unidades neonatais e pediátricas [7], [8].

A aspiração subglótica é descrita como medida complementar, por remover as secreções acumuladas acima do balonete antes de sua migração para o trato respiratório inferior; revisões sistemáticas associam o uso de tubos com lúmen para aspiração subglótica à redução das taxas de PAV e do tempo de ventilação mecânica [4]. O efeito é descrito como mais consistente quando a aspiração é programada e não apenas reativa [2], [4].

A mudança programada de decúbito é apresentada na literatura de enfermagem e de terapia intensiva como cuidado voltado à mobilização do paciente crítico e ao alívio de áreas pulmonares dependentes [9], com repercussões hemodinâmicas que devem ser consideradas na definição do intervalo de mudança [10]. Associada à elevação da cabeceira entre 30° e 45°, contribui para reduzir o refluxo gastroesofágico e a probabilidade de aspiração de conteúdo gástrico [1], [9].

As três medidas atuam em elos distintos da cadeia fisiopatológica da PAV, o que sustenta a organização de protocolos que as integram, os chamados pacotes (bundles) de prevenção [2], [4]. A integração dessas funções em uma única plataforma instrumentada é a

motivação de projeto do equipamento descrito na Seção II.

B. Planilhas de registro dos ensaios

As Tabelas V a VII constituem as planilhas de registro dos ensaios definidos na Seção IV, preenchidas com os resultados da campanha de verificação interna. Cada série é reportada com o tamanho amostral, o número de ciclos e o padrão de referência empregado, conforme declarado nas Tabelas III e IV. O critério de conformidade da Seção IV.D é atendido em todas as séries: o limite superior do intervalo de confiança de 95% do erro permanece inferior a 1,1% do valor alvo nos módulos de pressão e a 0,2% no módulo de decúbito, contra o requisito de projeto de 5%. As grandezas reportadas seguem as definições da Seção IV.A e o critério de conformidade da Seção IV.D.

TABELA V
REGISTRO DOS ENSAIOS DOS MÓDULOS DE PRESSÃO

Módulo	Perfil	Alvo	n	Ciclos	Erro médio (%)	DP (%)	Erro máx. (%)	t90 (s)	Deriva em 24 h (%)
Balonete	Neonatal	15cmH_2O	20	5	0,75	0,4	1,25	2,54	0,1
Balonete	Pediátrico	25cmH_2O	20	5	0,75	0,4	1,24	3,25	0,1
Balonete	Adulto	40cmH_2O	20	5	0,75	0,4	1,25	4,00	0,1
Aspiração	Neonatal	2mmHg	20	5	0,85	0,5	3,00	1,25	0,2
Aspiração	Pediátrico	10mmHg	20	5	0,85	0,5	2,10	2,25	0,2
Aspiração	Adulto	50mmHg	20	5	0,85	0,5	1,85	3,00	0,2

TABELA VI
REGISTRO DOS ENSAIOS DO MÓDULO DE DECÚBITO

Perfil	Ângulo alvo	Intervalo alvo	n	Ciclos	Erro angular médio (%)	DP (%)	Erro máx. (%)	Erro de temporização (%)
Pediátrico	90°	1 hora	20	24	0,1	0,05	0,25	0,002
Adulto	90°	1 hora	20	24	0,1	0,05	0,25	0,002

TABELA VII
REGISTRO DO DESEMPENHO DO SISTEMA DE ALARME

Condição de alarme	Limite ajustado	n eventos	Detecções	Latência média (s)	Latência máx. (s)	Falsos positivos	Falsos negativos
Pressão do balonete acima do limite	alvo + 10 cmH_2O	20	20	5,00	10,00	0	0
Pressão do balonete abaixo do limite	alvo − 10 cmH_2O	20	20	3,50	7,00	0	0
Pressão de aspiração fora da faixa	alvo ± 5 mmHg	20	20	5,00	8,00	3	0
Falha de temporização do decúbito	intervalo alvo ± 5 min	20	20	2,00	5,00	0	0
Desconexão / oclusão de linha	n.a.	20	20	5,00	8,00	0	0

Classificação de prioridade e sinalização conforme os requisitos da ABNT NBR IEC 60601-1-8 (Tabela II). O limite é expresso como desvio em relação ao valor alvo ajustado em cada perfil de paciente. Detecções: eventos induzidos que foram anunciados pelo equipamento; falsos negativos = n eventos − detecções. Falsos positivos: anunciações ocorridas em janela de observação sem condição de alarme presente, cuja duração deve ser declarada.

VI. Limitações e Direções Futuras

A. Ausência de análise de sensibilidade a perturbações

O protocolo apresentado avalia o comportamento dos módulos sob condições nominais e degraus aplicados ao valor de referência, mas não contempla, nesta etapa, análise de sensibilidade a perturbações representativas do uso clínico. Entre as perturbações não avaliadas encontram-se: vazamento progressivo no balonete e nas conexões; variações de complacência e de resistência do sistema respiratório; alterações da pressão positiva expiratória final; oclusão parcial da linha de aspiração por secreção; e variações de temperatura e umidade no ambiente de operação. A ausência dessa análise limita a extrapolação do desempenho verificado em bancada para o ambiente assistencial e constitui a próxima etapa prevista do trabalho.

B. Ensaios normativos pendentes

Os registros das Tabelas V a VII resultam de verificação interna conduzida pela equipe de desenvolvimento e não substituem ensaio normativo. Os ensaios previstos pelas normas relacionadas na Tabela II não haviam sido concluídos até a submissão deste manuscrito. A conformidade só poderá ser afirmada mediante relatórios de ensaio emitidos por laboratório acreditado, e as afirmações deste artigo limitam-se, portanto, à adoção das normas como requisitos de entrada do projeto.

C. Avaliação clínica e tratamento estatístico inferencial

Não foi conduzido, até o momento, estudo clínico comparando a incidência de PAV entre grupos assistidos com e sem o dispositivo. Prevê-se, como direção futura, estudo controlado com cálculo prévio de tamanho amostral e de poder estatístico, comparação da incidência por testes de hipótese apropriados, análise do tempo até o evento e apresentação de intervalos de confiança de 95%. A concordância entre a pressão indicada pelo equipamento e a medida pelo padrão de referência deverá ser avaliada por análise de Bland-Altman, com relato dos limites de concordância.

VII. Conclusão

Apresentou-se a arquitetura de hardware e a instrumentação de um monitor que integra, em plataforma embarcada única, o controle da pressão do balonete do tubo endotraqueal, a aspiração subglótica contínua e a temporização da mudança de decúbito, com parametrização das faixas de ajuste por perfil de paciente. As normas da família ABNT NBR IEC 60601-1 foram adotadas como requisitos de entrada do projeto e orientaram a definição do desempenho essencial de cada módulo, conforme consolidado na Tabela II.

O requisito de desvio inferior a 5% entre o valor medido e o valor alvo foi atendido na verificação interna documentada nas Tabelas V a VII, com margem de pelo menos uma ordem de grandeza em relação ao limite especificado. Trata-se, contudo, de verificação conduzida pela própria equipe de desenvolvimento, cuja confirmação depende dos ensaios normativos descritos na Seção VI. Não são reivindicados, portanto, conformidade normativa comprovada por relatório de ensaio nem benefício assistencial atribuível ao equipamento, cuja demonstração depende do estudo clínico descrito na Seção VI.

Reconhecimento

A CSE Equipamentos e Serviços registra publicamente o agradecimento pelo incentivo financeiro concedido pela Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) ao desenvolvimento deste equipamento, por meio do edital Impulsiona SC 2024.

Referências

[1] C. R. R. de Carvalho, “Pneumonia associada à ventilação mecânica”, Jornal Brasileiro de Pneumologia, vol. 32, n. 4, pp. xx–xxii (numeração romana original do editorial, correspondente às páginas 20–22), ago. 2006. doi: 10.1590/S1806-37132006000400003.

[2] F. Howroyd, C. Chacko, A. MacDuff, N. Gautam, B. Pouchet, B. Tunnicliffe, J. Weblin, F. Gao-Smith, Z. Ahmed, N. A. Duggal e T. Veenith, “Ventilator-associated pneumonia: pathobiological heterogeneity and diagnostic challenges”, Nature Communications, vol. 15, art. 6447, jul. 2024. doi: 10.1038/s41467-024-50805-z.

[3] D. Wu, C. Wu, S. Zhang e Y. Zhong, “Risk factors of ventilator-associated pneumonia in critically ill patients”, Frontiers in Pharmacology, vol. 10, art. 482, maio 2019. doi: 10.3389/fphar.2019.00482.

[4] H. Mumtaz, M. Saqib, W. Khan, S. M. Ismail, H. Sohail, M. Muneeb e S. S. Sheikh, “Ventilator associated pneumonia in intensive care unit patients: a systematic review”, Annals of Medicine & Surgery, vol. 85, n. 6, pp. 2932–2939, 2023. doi: 10.1097/MS9.0000000000000836.

[5] S. R. R. Juliano, M. C. R. Juliano, J. P. Cividanes, J. G. S. Houly, O. E. Gebara, G. V. L. Cividanes e E. C. Catão, “Medidas dos níveis de pressão do balonete em unidade de terapia intensiva: considerações sobre os benefícios do treinamento”, Revista Brasileira de Terapia Intensiva, vol. 19, n. 3, pp. 317–321, jul.–set. 2007.

[6] P. K. S. Vasconcellos, L. A. M. da Siqueira, G. A. da S. Alves Figueira, L. B. Martins, C. L. Cavellani, A. C. A. Novais e F. M. B. Bisinotto, “Avaliação da pressão do balonete das sondas de intubação traqueal: resultados obtidos pela técnica da palpação digital e experiência do anestesiologista”, Revista Médica de Minas Gerais, vol. 27, e-1887, 2017.

[7] J. P. B. B. Rodrigues, S. L. L. dos S. Bacci, J. M. Pereira, C. Johnston e V. M. G. de O. Azevedo, “Frequência e caracterização da utilização de cânulas traqueais com balonete em unidades de terapia intensiva neonatais e pediátricas do Brasil”, Revista Brasileira de Terapia Intensiva, vol. 32, n. 2, pp. 235–243, abr.–jun. 2020. doi: 10.5935/0103-507X.20200038.

[8] J. H. Kim, J. H. Ahn e Y. J. Chae, “Pediatric application of cuffed endotracheal tube”, Western Journal of Emergency Medicine, vol. 24, n. 3, pp. 579–587, abr. 2023. doi: 10.5811/westjem.59560.

[9] R. F. A. da Silva e M. A. de L. Nascimento, “Mobilização terapêutica como cuidado de enfermagem: evidência surgida da prática”, Revista da Escola de Enfermagem da USP, vol. 45, n. 2, pp. 413–419, out. 2011.

[10] A. P. de Assis, A. P. D. Soares, C. M. de Moraes, R. F. A. da Silva e F. R. V. Fernandes, “Mudança de decúbito na UTI: uma análise sobre as repercussões hemodinâmicas”, Global Academic Nursing Journal, vol. 2, e73, jan. 2021.

NeoTrack APLV: A Mobile System for Longitudinal Symptom Monitoring in Infants with Suspected Cow's Milk Protein Allergy

1st Monica Cristina da C. Pereira
Depto. Engenharia Biomédica
UFRN
Natal, Brasil
monicapereira.engbio@gmail.com

2nd Caroline Dantas Vilar
Depto. Engenharia Biomédica
UFRN
Natal, Brasil
caroline.vilar@ufrn.br

3rd Heliana Bezerra Soares
Depto. Engenharia Biomédica
UFRN
Natal, Brasil
heliana.soares@ufrn.br

***Abstract*—Cow's milk protein allergy (CMPA) presents with varied clinical manifestations, which can make it difficult to record the evolution of symptoms and monitor infants suspected of having this condition. This work presents NeoTrack CMPA, a mobile system developed to support the longitudinal recording of symptoms in infants up to six months of age. The application uses a structured questionnaire with conditional questions to record gastrointestinal, cutaneous, respiratory, and behavioral manifestations, as well as information related to feeding. At the end of each assessment, the system organizes the data and generates a report in PDF format, which can be shared with healthcare professionals. The application was developed in React Native with TypeScript and preliminarily evaluated through the monitoring of an infant with a confirmed CMPA diagnosis by a healthcare professional. The records obtained before and after 30 days of a cow's milk protein exclusion diet were compared. The results demonstrated the disappearance of cutaneous manifestations and the reduction of gastrointestinal, respiratory, and behavioral symptoms initially recorded. The application allowed for the structured documentation of these changes and the production of comparative reports for monitoring clinical evolution. Although the evaluation was limited to a single case and does not allow for establishing the clinical validity of the system, the results indicate the functional viability of NeoTrack APLV as a complementary tool for recording and communication between caregivers and healthcare professionals.**

Keywords— allergy to cow's milk protein; mobile health applications; infants; symptom monitoring.

I. INTRODUCTION

Cow's milk protein allergy (CMPA) is an adverse immunological reaction to proteins present in cow's milk and is one of the most frequent food allergies in the first years of life. Its clinical manifestations can involve different systems, including gastrointestinal, cutaneous, and respiratory symptoms, as well as behavioral changes such as irritability and persistent crying [1], [2]. This diversity of signs, associated with the possibility of immediate or delayed reactions, can make it difficult to recognize the evolution of the condition, especially when information depends exclusively on the memory and observation of caregivers.

The investigation of cow's milk protein allergy (CMPA) requires a clinical evaluation conducted by a healthcare professional, considering the patient's history, the symptoms presented, and their temporal relationship with exposure to cow's milk protein. Depending on the mechanism involved and the infant's clinical condition, the process may include an elimination diet followed by reassessment and, when indicated and performed under professional supervision, an oral food challenge test [1], [2]. In this context, the systematic recording of observed manifestations over time can contribute to documenting changes in the clinical picture and provide more organized information during follow-up.

However, daily monitoring is usually carried out by caregivers outside the clinical setting, which can result in incomplete, scattered, or poorly standardized records. The absence of a structured history makes it difficult to compare different moments, such as the periods before and after a dietary intervention. Furthermore, symptoms belonging to different categories can vary independently, making a tool capable of gathering them into a single longitudinal record relevant, without replacing the assessment or decision of the healthcare professional.

Mobile health technologies have the potential to support the continuous recording of information, facilitate the visualization of symptom evolution, and promote communication between users and professionals. Studies on applications designed for food allergies and intolerances have identified functionalities such as food diaries, reaction records, and the sharing of clinical information [3], [4]. Despite this potential, evaluations of these tools also point to limitations related to the quality of information, the participation of healthcare professionals in development, and the suitability of functionalities to the specific needs of users [3], [4]. In the case of cow's milk protein allergy (CMPA) in infants, there is a need for solutions aimed at the child's caregiver, capable of structuring symptoms from different systems and generating a history that can be presented during professional follow-up.

Given this scenario, this work presents NeoTrack APLV, a mobile system designed for the longitudinal monitoring of symptoms in infants up to six months old suspected of having cow's milk protein allergy. The application organizes information on feeding and gastrointestinal, cutaneous, respiratory, and behavioral manifestations through a structured questionnaire with conditional questions. At the end of the completion, the data are consolidated into a report that can be stored and shared with healthcare professionals. The main contributions of this work include: i) the digital structuring of the record of symptoms related to APLV; ii) the generation of reports for comparison between different

moments of follow-up; iii) the implementation of a mobile application aimed at caregivers of infants; and iv) a preliminary evaluation of its applicability in the longitudinal follow-up of a real case.

II. Related Works

The use of mobile technologies in monitoring food allergies has been investigated as a strategy to facilitate symptom recording, dietary control, and access to information about potential allergens. These applications can assist patients and caregivers in organizing data produced outside the clinical setting, allowing them to document episodes, identify temporal relationships, and share information during professional follow-up.

Mandracchia et al. [3] analyzed mobile applications designed for people with food allergies and intolerances and identified functionalities related to ingredient lookup, identification of potentially unsuitable foods, meal logging, and reporting of adverse reactions. The authors observed, however, that many solutions focus on dietary restriction or product lookup, offering limited resources for structured monitoring of symptom progression.

In an assessment of the quality of applications aimed at food allergies, Wu et al. [4] found significant differences regarding the reliability of information, usability and the involvement of health professionals in the development of the tools. These aspects are particularly relevant in applications related to child health, in which inaccurate information or inadequate interpretations can influence the decisions of those responsible. Therefore, such systems should be presented as complementary recording and communication instruments, and not as autonomous diagnostic or therapeutic guidance tools.

Other digital solutions employ electronic diaries to associate food consumption with the occurrence of clinical manifestations. This type of record can reduce the user's reliance on memory and make the temporal relationship between exposure, dietary intervention, and symptom progression clearer [5]. However, general-purpose tools do not always account for the particularities of cow's milk protein allergy (CMPA) in infants, such as the type of feeding, the diversity of clinical manifestations, and the need for complete filling out by the caregiver.

NeoTrack APLV differs from these approaches because it was developed specifically for monitoring infants up to six months old with suspected CMPA (Cow's Milk Protein Allergy). The system gathers, in a single flow, information on feeding and gastrointestinal, cutaneous, respiratory, and behavioral manifestations. Furthermore, it uses conditional questions to adapt the questionnaire to the answers provided and generates standardized reports that allow for comparison at different points in the follow-up.

Thus, the main contribution of NeoTrack APLV is not in performing automated diagnoses, but in the longitudinal structuring of information observed by caregivers. This proposal seeks to meet a specific need in the monitoring process: transforming everyday observations, often scattered, into organized records that can complement the assessment conducted by the healthcare professional.

III. Materials and Methods

The development of NeoTrack APLV was organized into three main stages: defining the requirements and structuring the questionnaire, implementing the mobile application, and conducting a preliminary assessment of its applicability. The system was designed to record information from infants up to six months old, with completion performed by a caregiver. Its purpose is to document the evolution of symptoms and facilitate the sharing of information with healthcare professionals, without making a diagnosis or recommending therapeutic interventions.

A. Defining the requirements and structuring the questionnaire

The development of NeoTrack APLV was motivated by the author's experience while monitoring her two children with manifestations related to cow's milk protein allergy (CMPA). In the first case, the diversity of symptoms and the difficulty in organizing them over time contributed to making the investigation process more time-consuming. Subsequently, the experience gained allowed the author to recognize and record with greater attention similar manifestations observed in her second child. From this experience, the need for a tool to assist other caregivers in the systematic recording of symptoms and in communicating this information to healthcare professionals was identified.

Based on this need and the clinical manifestations described in the guidelines used as a basis [1], [2], the system requirements were defined and the questionnaire was structured. The application was designed to collect identification data, age, type of food and symptoms observed by the person responsible. The manifestations were organized into gastrointestinal, skin, respiratory and behavioral categories, allowing information to be gathered that, in everyday life, could remain dispersed or depend solely on the memory of those responsible.

The author's personal experience contributed to the identification of the problem and the initial definition of the functionalities. However, the content of the questionnaire was structured based on the literature and guidelines consulted and is not presented as a validated clinical instrument.

B. Mobile application development

NeoTrack APLV was developed with the React Native framework and the TypeScript language, using the Expo environment to implement and run the application. The organization of the screens and navigation was carried out with Expo Router, while the interface components were implemented with React Native Paper. Additional resources were used to select and process dates, validate data, adapt the interface to the virtual keyboard and generate documents.

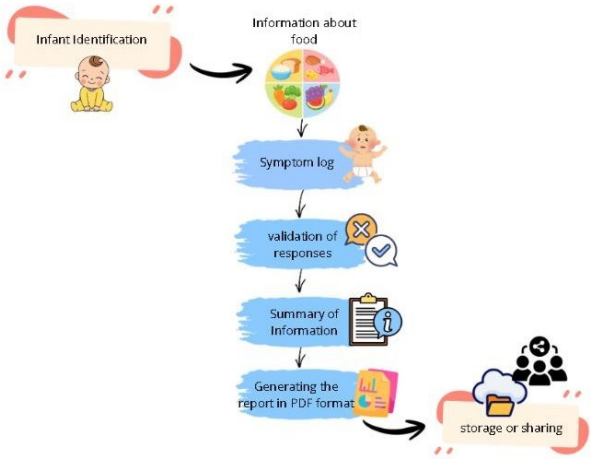
The user workflow comprises: i) access to the home screen; ii) infant identification; iii) filling in information about feeding and symptoms; iv) validation of responses; v) presentation of a summary; and vi) generation and sharing of the repor.

The data from each completed form is organized into a PDF document, identified by the date and time it was created. The report compiles the information provided by the party responsible and can be stored or shared using the resources

available on the mobile device. This functionality was developed to allow the storage of records made at different times and the sharing of information with healthcare professionals.

Fig. 1 should show the general workflow of NeoTrack APLV, from completing the questionnaire to generating and sharing the report.

Fig. 1. General workflow for using NeoTrack APLV, including identification, symptom recording, response validation, summary presentation, and report generation.



C. *System functional verification*

Following implementation, the functional verification of NeoTrack APLV was conducted exclusively by the author herself. During this stage, the functionalities foreseen in the system were explored, including infant identification, filling in information about feeding and symptoms, presenting conditional questions, validating mandatory fields, reviewing responses, and generating and sharing the report.

The author also completed data entry corresponding to different stages of the monitoring process, allowing verification of whether the system adequately organized the information and enabled comparison between the recorded symptoms. The functioning of the conditional rules, the consistency of the dates, the restriction of the age range covered, and the structure of the information presented in the generated document were observed.

This stage involved functional usability testing of the application with a single patient, who was the newborn son of the author. The test was conducted in a real-use context to verify the proper functioning of the implemented features, including symptom recording, clinical information monitoring, and report generation. No external users, caregivers of other children, or healthcare professionals participated in the testing.

Therefore, the results demonstrate the technical and functional performance of the application in this specific case but do not allow for the generalization of its usability, user acceptance, the clinical validity of the questionnaire, or the effectiveness of the system for monitoring cow's milk protein allergy (CMPA).

IV. Results

A. *Implemented functionalities*

The development resulted in a mobile application capable of guiding the caregiver through a structured flow of recording information related to the infant. The system includes screens for identification, selection of the assessment date, characterization of feeding, and recording of gastrointestinal, cutaneous, respiratory, and behavioral manifestations. The main implemented interfaces are shown in Fig. 2.

Conditional questions allow the content presented to be adapted according to the answers provided, so that supplementary fields are only displayed when related to the selected manifestation. Validation messages have also been implemented to identify missing required fields, inconsistent dates, and information incompatible with the age range covered by the application, as exemplified in Fig. 2(d).

Before generating the document, NeoTrack APLV presents a summary of the recorded information, allowing the responsible party to review the data, as shown in Fig. 2(e). After confirmation, the system generates a report in PDF format, automatically named with the application identification, date, and time of creation. The document can be stored or shared using the resources available on the device.

Fig. 2. Main interfaces of NeoTrack APLV: (a) home screen; (b) infant identification; (c) recording of manifestations; (d) information validation; and (e) summary prior to report generation.



B. *Generating the reports*

The report produced by the system organizes, in a single document, the identification data, information about feeding, and the manifestations selected during completion. The inclusion of the date and time allows distinguishing records made at different times, enabling the creation of a documentary history. An example of the structure of the report generated by NeoTrack APLV is shown in Fig. 3.

As observed in Fig. 3, the information is presented in an organized manner, seeking to facilitate its consultation and sharing with healthcare professionals. However, the document does not contain a diagnosis, automatic classification of the condition, or therapeutic recommendation. Its purpose is to reproduce, in a standardized way, the information provided by the responsible party.

Fig. 3. Example of a report generated by NeoTrack APLV, containing recorded information on feeding and observed manifestations in the infant.



C. *Functional verification*

In the verification conducted by the author, it was possible to follow the complete application flow presented in Fig. 1, from the identification of the infant to the generation and sharing of the report. The conditional questions were presented according to the selected answers, while the validation mechanisms prevented progress when mandatory information was missing or inconsistent.

Fillings were also carried out representing different moments, allowing us to verify whether the information was organized separately in the respective reports. The execution demonstrated that the system implements the expected functional flow and produces documents that can be used to compare records made on different dates. Fig. 3 exemplifies the type of document produced at the end of this process.

These results should be interpreted as a demonstration of technical functionality based on the use of the application by the author and her newborn son, who was the patient monitored during the evaluation. The verification allowed the functionality of the system to be assessed in a real case, but did not provide sufficient data regarding usability, user acceptance, comprehension of the questions, completion time, or professional evaluation of the content. Similarly, it is not possible to establish the clinical validity of the questionnaire or the effectiveness of the application in monitoring cow's milk protein allergy (CMPA).

V. Discussion

NeoTrack APLV was developed to transform everyday observations made by caregivers into structured records, organized according to information about feeding and gastrointestinal, cutaneous, respiratory, and behavioral manifestations. The workflow shown in Fig. 1 and the interfaces shown in Fig. 2 demonstrate that this information has been integrated into a single data entry sequence, from infant identification to data review and report generation.

The use of conditional questions helps to direct completion, as it allows additional fields to be presented only when related to previous answers. Furthermore, the validations implemented seek to reduce the occurrence of incomplete or inconsistent records. However, as the verification was carried out exclusively by the author, it is still not possible to say whether the organization of the questionnaire, the terminology used and the sequence of questions are clear enough for other responsible parties.

Applications focused on food allergies often include features related to ingredient identification, product lookup, and reaction logging [3]. NeoTrack APLV presents a distinct approach by being specifically directed at the longitudinal recording of symptoms in infants up to six months of age. The system does not seek to identify safe foods or establish a diagnosis, but to organize the manifestations reported by the caregiver and produce documents corresponding to different moments of follow-up.

The report exemplified in Fig. 3 constitutes one of the system's main functionalities, as it gathers the information provided during completion into a single document. This feature can reduce reliance on scattered notes and retrospective recall by those responsible. However, the report's usefulness for clinical practice should still be evaluated by healthcare professionals, considering aspects such as the relevance of the information, clarity of presentation, and effective contribution to consultation.

Studies analyzing apps related to food allergies point to variations in the quality of information, usability, and the participation of healthcare professionals in their development [3], [4]. These aspects also represent points of attention for NeoTrack APLV. Although the questionnaire content was based on the manifestations described in the consulted guidelines [1], [2], the instrument has not been submitted to validation by experts. Consequently, the app should be considered an early-stage technological tool, and not a validated clinical instrument.

Another important limitation is the lack of testing with the target audience. The author's personal experience contributed to the identification of the problem and the initial definition of the functionalities, but it does not replace a systematic evaluation with different user profiles. Those responsible may have different levels of education, familiarity with mobile technologies and understanding of terms related to symptoms, factors that can influence the completion and quality of records.

Usability tests, clinical evaluations, accessibility analyses, or comparisons with other forms of record-keeping were not conducted. Therefore, the results presented only demonstrate that the planned functionalities were implemented and executed in the environment used by the author. It is not possible to conclude, at this stage, that NeoTrack APLV improves diagnosis, reduces the time required to investigate the condition, or produces clinical benefits.

As a continuation of this work, it is recommended that the questionnaire content be evaluated by professionals with experience in pediatric food allergies and that usability tests be conducted with caregivers of infants. Subsequent studies could assess the clarity of the questions, the completion time, the ease of navigation, and the usefulness of the reports during professional follow-up. Features such as an integrated history of records, alerts for new assessments, expansion to other age groups, and secure integration with health information systems could also be incorporated.

VI. Conclusion

This work presented NeoTrack APLV, a mobile system developed to support the longitudinal recording of symptoms in infants up to six months old with suspected allergy to cow's milk protein. The solution organizes information related to

food and gastrointestinal, skin, respiratory and behavioral manifestations, using conditional questions and validation mechanisms to conduct completion.

The implemented system allows for the review of recorded information and the generation of reports in PDF format, which can be stored and shared with healthcare professionals. The functional verification conducted by the author herself demonstrated the execution of the planned workflow, from infant identification to report generation, including the functioning of the conditional questions and implemented validations.

However, no tests were conducted with caregivers, healthcare professionals, or patients, nor were there any formal usability and accessibility assessments or clinical validation of the questionnaire. Therefore, the results presented are limited to demonstrating the technical functioning of the application, and do not allow us to affirm its effectiveness in monitoring cow's milk protein allergy (CMPA) or its contribution to the diagnostic process.

As future work, it is proposed to validate the questionnaire content with specialists in pediatrics and food allergy and to conduct instrument validation before assessing its clinical applicability. Usability tests should also be performed with caregivers from different profiles, as well as independent evaluations by healthcare professionals, considering the clarity of the questions, completion time, interpretation errors, and usefulness of the information provided.

Features such as integrated history, alerts for new assessments, support for other age groups, and secure integration with health information systems could also be incorporated. Thus, NeoTrack APLV constitutes an initial technological proposal with the potential to support the organization of information provided by caregivers, without replacing the assessment and conduct of healthcare professionals.

ACKNOWLEDGEMENTS

The authors would like to thank LAPSIM – Laboratory of Acquisition and Processing of Medical Signals and Images of the Biomedical Engineering course at the Federal University of Rio Grande do Norte (UFRN) for the technical support, infrastructure and support for the development of this work.

REFERENCES

[1] Ministério da Saúde, Protocolo Clínico e Diretrizes Terapêuticas da Alergia à Proteína do Leite de Vaca (APLV). Brasília, DF, Brasil: Ministério da Saúde, 2022. [Online]. Available: https://www.gov.br/conitec/pt-br/midias/consultas/relatorios/2022/20220427_pcdt_aplv_cp_24.pdf. [Accessed: May 23, 2026].

[2] S. Koletzko, B. Niggemann, A. Arato, J. A. Dias, R. Heuschkel, S. Husby, M. L. Mearin, A. Papadopoulou, F. M. Ruemmele, A. Staiano, M. G. Schäppi, and Y. Vandenplas, "Diagnostic approach and management of cow's-milk protein allergy in infants and children: ESPGHAN GI Committee practical guidelines," J. Pediatr. Gastroenterol. Nutr., vol. 55, no. 2, pp. 221–229, Aug. 2012, doi: 10.1097/MPG.0b013e31825c9482.

[3] F. Mandracchia, E. Llauradó, L. Tarro, R. M. Valls, and R. Solà, "Mobile phone apps for food allergies or intolerances in app stores: Systematic search and quality assessment using the Mobile App Rating Scale (MARS)," *JMIR Mhealth Uhealth*, vol. 8, no. 9, Art. no. e18339, Sep. 2020, doi: 10.2196/18339.

[4] A. C. Wu, R. Walsh, and A. L. Wang, "Assessment of the quality of mobile apps for food allergy," J. Allergy Clin. Immunol. Pract., vol. 10, no. 6, pp. 1658–1659.e7, Jun. 2022, doi: 10.1016/j.jaip.2022.02.015.

NeuroSUS-RN: a scalable neurological monitoring platform to expand access to Neonatal Care in Brazil's Universal Health System (SUS)

Aline Elvina R Fernandes
ICe, UFRN
Natal/RN, Brazil
aline.fernandeselv@gmail.com

Thiago Guimarães Silva
ICe, UFRN
Natal/RN, Brazil
tguimaraes.90@gmail.com

Arthur Jorge de Vasconcelos Ribeiro
MEJC, UFRN, HU Brasil,
Natal/RN, Brasil
arthurribeiro1976@gmail.com

Jéssica Winne
FMRP, USP
Ribeirão Preto/SP, Brazil
jessica.winne1@gmail.com

Anna Christina do N. G. Barreto
MEJC, UFRN
Natal/RN, Brazil
anna.barreto@ufrn.br

Ricardo Ney Cobucci
UnP &,MEJC UFRN
Natal/RN, Brazil.
ricardo.cobucci.737@ufrn.edu.br

Marcos Aurélio Tavares Filho
ECT, UFRN
Natal, Brazil
marcos.filho.123@ufrn.edu.br

Antonio Wallace Antunes Soares
DEE, UFRN
Natal/RN, Brazil
wallace.soares@ufrn.br

Eduardo Bouth Sequerra
ICe, UFRN
Natal/RN, Brazil
ebsequerra@neuro.ufrn.br

Helton Maia Peixoto
ECT, UFRN,
Natal/RN, Brazil
helton.maia@ufrn.br

Richardson N Leão
ICe, UFRN
Natal/RN, Brazil
richardson.leao@neuro.ufrn.br

***Abstract*—Neonatal seizures represent critical neurological manifestations in Neonatal Intensive Care Units (NICUs), with nearly 50% of events in the first days of life remaining entirely subclinical and detectable solely through continuous electroencephalogram (EEG) monitoring. In low- and middle-income countries like Brazil, scaling such specialized neuromonitoring is severely bottlenecked by high equipment costs and a shortage of pediatric neurophysiologists. To bridge this gap, we introduce the NeuroSUS-RN platform: a scalable, low-cost Artificial Intelligence of Things (AIoT) video-EEG ecosystem tailored for universal deployment within Brazil's Universal Health System (SUS) NICUs.**

We also present preliminary data from neonates monitored in the university's NICU, demonstrating the clinical feasibility and signal quality of the platform in real-world scenarios. Looking forward, we plan to implement complementary serological tests within the platform framework to help determine and screen for the most common neurological disorders underlying the country's clinical reality. By shifting complex diagnostic computing from high-end GPU clusters to the bedside, the CNI platform mitigates cloud dependency and structural infrastructure barriers, offering a viable pathway toward the universalization and democratization of advanced neonatal neuroprotection.

***Index Terms*—NICU, electrographic seizures, SUS**

I. Introduction

A major challenge in neonatal neuroprotection is that brain injury manifestations in critically ill newborns are clinically unusual and frequently lack visible physical correlates. In fact, approximately 50% to 80% of neonatal seizures are purely electrographic, occurring with no reliable clinical signs and remaining entirely invisible to bedside observation. Continuous electroencephalography (EEG) combined with synchronous video monitoring (video-EEG), is the gold standard for accurate bedside diagnosis and seizure management. Despite its critical diagnostic value, conventional continuous video-EEG remains virtually absent from the clinical routine of NICUs across the developing world due to operational and financial bottlenecks.

This clinical deficit is especially pronounced in the context of Hypoxic-Ischemic Encephalopathy (HIE) and regional epidemics. While HIE affects a modest 1 to 3 per 1,000 live births in developed nations, its incidence surges dramatically to as many as 26 per 1,000 live births in developing countries. In Brazil alone, over 15,000 to 20,000 newborns are affected by HIE annually [1]. Moreover, recent regional public health crises like the Brazilian outbreak of congenital Zika virus syndrome is a sad example of how unmanaged neonatal brain injuries can catastrophically compound permanent developmental impairments [2]. As prolonged convulsive events directly correlate with cumulative, irreversible cerebral damage, real-time bedside screening and immediate physiological telemetry are vital to altering patient trajectories and improving neonatal survivability [3].

To address this critical equity gap, we present a scalable technological framework designed to expand access to neonatal neurological monitoring through affordable, edge-computed engineering. We also show here preliminary results of neurological monitoring of high risk neonates (N=4) obtained in the NICU. Developed through an interdisciplinary collaboration between the Brain Institute (ICe/UFRN) and the

Januário Cicco Maternity School (MEJC/UFRN), we introduce a low-cost, wireless Artificial Intelligence of Things (AIoT) platform integrating video-EEG to other NICU monitoring devices. The battery-operated system combines a lightweight EEG (the software also displays amplitude-integrated EEG - aEEG) with video. Embedded machine learning algorithms deployed on a hybrid ESP32/ARM processing architecture enable real-time signal quality assessment, artifact reduction, and seizure detection directly at the bedside, without dependence on cloud connectivity or high-performance computing infrastructure. By reducing complexity, preserving patient privacy, and targeting an estimated manufacturing cost of approximately $200 USD per unit at scale, this platform provides an accessible and self-contained solution tailored to the clinical and socioeconomic challenges of Brazil's Universal Health System (SUS) and other resource-limited healthcare environments.

II. Methods

A. Data acquisition and real-time analysis

We implemented a wireless platform for real-time neurophysiological and systemic monitoring in neonatal intensive care environments. The system consists of battery-operated devices organized in a master–subordinate architecture: a central bedside processing and imaging hub, a wireless EEG module, and an optional lightweight foot-mounted pulse oximetry/heart rate sensor. The central hub integrates CMOS imaging and far-infrared thermal sensing, using local computer vision algorithms to identify patient landmarks, track motor activity, and dynamically adjust thermal measurements while preserving privacy by extracting features without storing raw video. The EEG module incorporates precision analog front-end circuitry with instrumentation amplifiers, transient suppression, filtering, and 24-bit analog-to-digital conversion for high-fidelity cortical signal acquisition. A quick-disconnect magnetic electrode interface with spring-loaded contacts enables stable recordings while reducing electrode displacement during neonatal care procedures. All data streams are synchronized through a local time-reference system, while flexible firmware updates allow adaptation between automated seizure detection using amplitude-integrated EEG (aEEG) and high-resolution conventional EEG acquisition. To optimize data transmission, raw high-bandwidth signals are locally stored, whereas processed metrics, vital signs, and neurological alerts are transmitted through low-power long-range communication protocols, enabling scalable tele-neurological monitoring in resource-limited healthcare settings.

Data monitoring was performed using Neo-EEG, an open-source, web-based platform for real-time EEG visualization, recording, and analysis. Signals are acquired by an eight-channel, 24-bit analog-to-digital converter (ADS1299) driven by an ESP32 microcontroller, which streams the data over WiFi (UDP) at a 250 Hz sampling rate. Neo-EEG follows a client–server architecture, in which a Python/FastAPI backend receives the stream and broadcasts it to the browser over WebSocket, while a React frontend renders all eight channels as a real-time clinical montage with an adjustable time window and amplitude gain. Digital filtering is applied to the signal, a 60 Hz notch filter to suppress power-line interference, together with configurable band-pass and band-reject (IIR) filters to restrict the recording to the EEG band of interest and reject residual noise. Each session combines the EEG signals, exported as CSV, with time-synchronized USB-camera video, exported as MP4, preserving frame-level alignment between the two streams.

B. Head phantom

A neonatal head phantom was developed to evaluate the performance of the EEG acquisition system under controlled and reproducible conditions. The phantom was fabricated using electrically conductive PLA with a bulk resistivity of approximately 120–200 Ω/cm and a sparse internal infill structure designed to achieve electrical conductance properties comparable to the combined impedance characteristics of neonatal skin, connective tissue, and skull layers. Conductive wires embedded internally within the phantom were used to deliver realistic neonatal EEG signals, reproducing microvolt-level cortical activity obtained from an open-access EEG data repository. The signals were replayed through two EEG channels using the audio output interface of a laptop, allowing the assessment of signal acquisition, amplification, filtering, and wireless transmission performance under conditions representative of neonatal EEG monitoring. This approach enabled validation of the platform's ability to capture low-amplitude electrophysiological signals while providing a practical and low-cost alternative for system testing and calibration.

C.Ethics committee

This research is being conducted in accordance with the requirements established by the Research Ethics Committee, under approval CAAE n 91421925.6.0000.5537, at Maternidade Escola Januário Cicco (MEJC/UFRN), with strict adherence to data anonymization protocols.

D. Preliminary NICU recordings

Data were collected from 4 high risk newborn patients at the MEJC NICU, with gestational ages ranging from 27 to 37 weeks and birth weights ranging from 870 g to 2390 g.

III. Results

EEG acquisition system was evaluated using the neonatal head phantom (Fig. 1A) with a 7-lead EEG montage (Fig. 1B) and an additional ECG channel (channel 8), and its performance was compared against a commercial reference system (BrainAmp DC, Brain Products GmbH). Phantom head recordings demonstrated the capability of the proposed platform to capture and reproduce low-amplitude neonatal EEG signals across multiple channels, with signal morphology and temporal patterns consistent with the reference acquisition system. SNR, common mode rejection ratio and other signal parameters will be published in the future elsewhere. The measured baseline noise levels remained within the expected range for the implemented acquisition electronics, using a common reference configuration for all channels.

To demonstrate the feasibility of the proposed video-EEG platform in a clinical setting, we also present the preliminary results of continuous neuromonitoring (1–3 h) performed in four high-risk neonates admitted to the NICU of the Federal

University of Rio Grande do Norte (UFRN). The main clinical and demographic characteristics of the enrolled neonates are summarized in Table 1, including gestational age at birth, sex, birth weight, 1- and 5-minute Apgar scores, corrected (postmenstrual) age at the time of EEG recording, and the clinical indication for neuromonitoring.

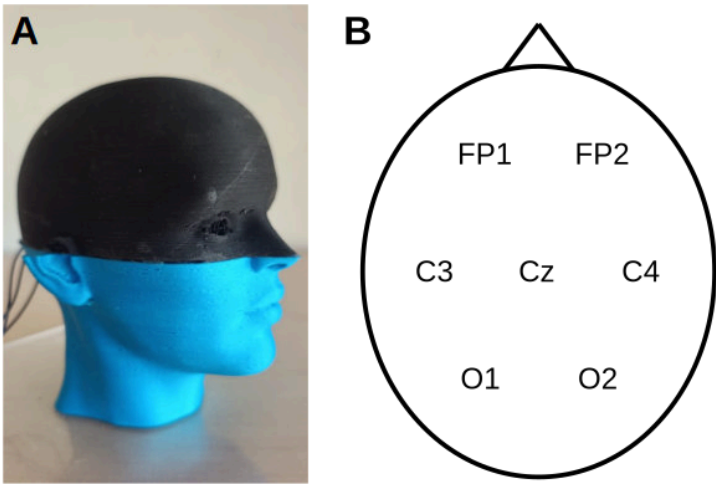


Figure 1: (A) Head phantom and (B) electrode montage.

Table 1. Clinical characteristics of the neonates

	Sex	GA at birth	Birth weight	APGAR (1'/5')	Corrected age at exam
Baby 1	Male	32w6d	1460g	6/7	35w6d (PMA)
Baby 2	Female	37w1d	2390 g	7/8	41w4d (PMA)
Baby 3	Male	28w3d	870 g	6/9	1 month 15 days
Baby 4	Male	27w5d	1215 g	8/8	42.4 days

GA, gestational age; PMA, postmenstrual age; w, weeks; d, days.

The association between Apgar scores and neonatal seizure activity has been extensively studied, though findings vary according to gestational age and birth weight. In a large population-based cohort of over 1.2 million term infants, [4] found that the adjusted hazard ratio for cerebral palsy increased steadily with decreasing five-minute Apgar scores, and reported that associations between Apgar scores and epilepsy, while less pronounced than for cerebral palsy, still showed increased hazard ratios in infants with a five-minute Apgar score of 7 or less and a ten-minute score of 8 or less. However, this association appears to be substantially weaker, or even absent, in preterm and low-birth-weight populations (data not shown).

An expected inverse correlation between Apgar scores and EEG findings has been reported specifically in full-term and normal-birth-weight infants, whereas no significant correlation was found in preterm or low-birth-weight groups [5]. This distinction may help explain the pattern observed in our cohort: the three neonates presenting brief rhythmic discharges (Babies 2, 3, and 4) had 5-minute Apgar scores ≥8, while Baby 1, the only neonate without this EEG finding, had the lowest 5-minute score [4] in the sample. Rather than reflecting a protective effect of higher Apgar scores, this pattern is consistent with the notion that, in preterm infants, seizure activity is driven predominantly by factors independent of immediate perinatal asphyxia, such as gestational age itself.

The incidence of neonatal seizures has been shown to rise sharply with decreasing gestational age [6], a trend also reflected in our sample, where the two most extremely preterm neonates (Babies 3 and 4, born at 28w3d and 27w5d) were both among those presenting rhythmic discharges. Similarly, abnormal aEEG patterns have been shown to occur at comparable rates in newborns with low (0–5) and higher (>6) Apgar scores [7], reinforcing that Apgar alone is an unreliable predictor of EEG abnormality in this population. Given the small sample size (n=4), no statistical association between Apgar scores and EEG findings could be established in the present study; however, we believe that Apgar vs. seizure correlation will be an important metric for the evaluation of our system.

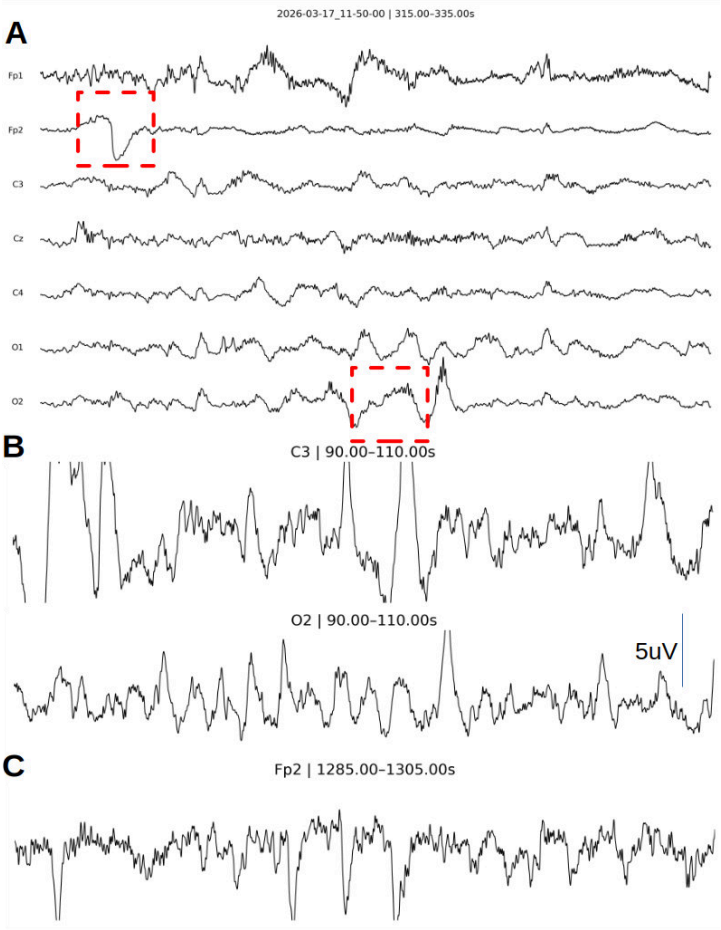


Figure 2: EEG traces obtained from the four neonatal patients during monitoring. A. 20s recording from Baby 1 showing frontal sharp transients and delta brush. B. C3 and O2 of Baby 2 (20s) showing a >10s seizure event. C. Brief rhythmic discharge (<10s) Baby 3.

Representative findings from Baby 1 illustrate the cortical maturation patterns and pathological findings described above. Delta brushes, delta waves with superimposed beta oscillations were prominently present (e.g. O2 in Fig. 1A), a pattern consistent with the neonate's gestational age of 36 weeks. Asystematic frontal sharp transients were also identified over Fp1 and Fp2, a finding physiologically expected during the transition to term. In Baby 2, a focal electrographic seizure was captured, consisting of a 10-second ictal event restricted to channel C3, with clear evolution in frequency and morphology (a rhythmic theta-delta pattern) characterizing a focal electrographic discharge without apparent generalization. Isolated sharp waves were additionally observed over O2, without synchrony with the central ictal event, suggesting irritative activity independent of the main seizure focus. Movement artifacts and low-frequency oscillations, likely reflecting respiratory interference, were also recorded, along with synchronous technical transients across all channels attributable to the experimental nature of the acquisition hardware. For Baby 3, a well-characterized episode of brief rhythmic discharge was observed in the frontocentral region, lasting 5 seconds. All four neonates were receiving phenobarbital at the time of the EEG recording. This is a relevant clinical detail, as phenobarbital remains the most widely used first-line antiseizure medication in neonates, despite a variable and often limited response rate [8-10]. Recent cohort data indicate that fewer than 40% of neonates achieve seizure freedom following first-line treatment with phenobarbital [10].

Notably, Babies 3 and 4 were receiving additional antiseizure medications beyond phenobarbital, with Baby 4 whose EEG indication was treatment-refractory seizure, under a regimen of four concurrent antiseizure medications. This escalation in pharmacological treatment reflects the well-documented limitations of phenobarbital monotherapy in achieving seizure control, particularly in cases of refractory neonatal seizures, which often require sequential or combined second and third-line agents such as levetiracetam, phenytoin, or midazolam [9]. Phenobarbital use has also been associated with reduced amplitude and spatial propagation of electrographic seizures, a phenomenon that may contribute to electroclinical uncoupling the dissociation between electrographic and clinically visible seizure activity [10]. As such, the persistence of brief rhythmic discharges in three of the four neonates despite polytherapy in two of them is consistent with the known limited efficacy of phenobarbital and, in Baby 4's case, with the clinically documented refractoriness of the seizures. This observation should nonetheless be interpreted with caution given the small sample size.

IV. Discussion

This article is the first report of the development of a video-EEG system for neonatal seizure detection and continuous neurological monitoring designed specifically for the NICU. Patient data presented here is preliminary with no statistical value as the clinical validation of our platform is currently on the way. The main objective of the NeuroSUS-RN is to provide a low-cost, easily deployable, and adaptable solution for neonatal neurological monitoring across the diverse landscape of NICUs in Brazil, with particular emphasis on resource-sensitive and geographically isolated healthcare settings. Unlike conventional neuromonitoring systems that often require specialized infrastructure, stable electrical conditions, and high-bandwidth data connectivity, our approach was designed to operate reliably under challenging clinical environments, minimizing susceptibility to electrical interference, unstable power conditions, and limited digital infrastructure. By incorporating robust signal acquisition strategies, local data processing, and flexible communication protocols, the platform aims to reduce dependence on external resources while maintaining clinically relevant monitoring capabilities. Furthermore, the system architecture was intentionally based on commercially available components with relative ease of acquisition in the Brazilian market, facilitating maintenance, scalability, and future implementation within the public healthcare system. This design philosophy supports the development of a sustainable technology pathway toward expanding access to advanced neonatal neuromonitoring beyond highly specialized centers. We estimated a cost of circa US$200 based on the commercial costs of components in Brazil, PCB design and assembly. To conform to ANVISA requirements (and INMETRO, ANATEL etc) costs could rise due to regulations on component traceability and the minute number of companies in Brazil that comply with the *good manufacturing practices* required for class II devices. Costs may also be lower for scaled production.

In the four neonates monitored in this preliminary clinical evaluation, the low-cost EEG acquisition system was able to capture electrographic traces of sufficient quality for clinical interpretation, allowing a neurophysiologist to identify brief rhythmic epileptiform discharges in three of the four recordings. Notably, two of these neonates subsequently underwent a conventional, gold-standard EEG examination, which confirmed the presence of epileptic seizure activity, corroborating the findings initially detected by the low-cost platform. The preliminary agreement between the low-cost system and conventional EEG supports the feasibility of the proposed platform for neonatal seizure monitoring, particularly in resource-limited settings such as low- and middle-income countries, where access to conventional EEG systems remains limited.

To further expand the platform, we are currently developing a multichannel functional near-infrared spectroscopy (fNIRS) module that will be fully integrated with the video-EEG system into a single multimodal neuromonitoring device. The platform also includes the development of hybrid EEG-fNIRS electrodes, enabling simultaneous acquisition of electrophysiological and cerebral hemodynamic signals from the same scalp locations. Because the optical emitters used in fNIRS induce photovoltaic (Becquerel effect) artifacts in EEG recordings [11] , the system employs temporally modulated light sources that generate predictable artifact patterns, enabling their suppression through dedicated signal-processing and embedded artificial intelligence algorithms [11]. Cerebral fNIRS has emerged as a valuable bedside neuromonitoring modality in the NICU by providing continuous, non-invasive

measurements of regional cerebral oxygenation and hemodynamic responses. When combined with EEG, fNIRS offers complementary information on neurovascular coupling, cerebral perfusion, and metabolic function, improving the detection and characterization of seizures, hypoxic-ischemic brain injury, and other neonatal neurological disorders [12, 13]. Integrating these complementary modalities into a low-cost embedded platform has the potential to deliver comprehensive multimodal neuromonitoring to NICUs that currently lack access to advanced neurodiagnostic technologies while enabling real-time bedside analysis through embedded AI. It is important to note that, unlike EEG amplifiers, there is no company that manufactures cerebral fNIRS devices in Brazil.

Beyond seizure monitoring, the platform will enable the establishment of a standardized Brazilian neonatal EEG database, providing a representative dataset for the development and validation of robust artificial intelligence models tailored to the Brazilian population. In addition, the microprocessor-based EEG acquisition units and intelligent camera modules will support the deployment of embedded AI models for real-time neonatal pose estimation and automated seizure detection directly at the bedside. Upon completion of the current study, a larger multicenter trial with systematic, simultaneous reference-standard EEG recordings will be conducted to establish the sensitivity and specificity of the platform for neonatal seizure detection and to further validate these embedded AI technologies.

The current patient data is preliminary and reflects only the initial usability assessments of our system. We plan to evaluate the video-EEG platform in a cohort of 90 high-risk neonates (sample size approved by the local ethics committee) admitted to the NICU of the Januário Cicco Maternity School Hospital (MEJC). Upon completion of the MEJC study, we intend to initiate a multicenter trial across NICUs of varying levels of complexity. This broader study will allow us to refine the system's design and propose standardized protocols for implementation within the SUS.

Acknowledgment

This study was supported by the Financiadora de Estudos e Projetos (FINEP) through the Centro de Neuroproteção Infantil project. The authors gratefully acknowledge the Sistema Único de Saúde (SUS) of Brazil for providing the clinical infrastructure and patient care that made this study possible. A.E.F.R. was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001.

References

[1] A. Craig and O. DeLaGarza-Pineda, "Neonatal neurocritical care in low-resource settings: Challenges and innovations in hypoxic-ischemic encephalopathy," Clin. Perinatol., vol. 52, no. 2, pp. 361–374, 2025.

[2] E. Sequerra, et al., "Clinical, neuroimaging and video electroencephalography findings in children with congenital Zika syndrome: An analysis from a neurorehabilitation centre," EClinicalMedicine, vol. 25, Art. no. 100252, 2020.

[3] R. M. Pressler et al., "Treatment of seizures in the neonate: Guidelines and consensus-based recommendations—Special report from the ILAE Task Force on Neonatal Seizures," Epilepsia, vol. 62, no. 3, pp. 615–628, 2021.

[4] M. Persson, N. Razaz, K. Tedroff, K. S. Joseph, and S. Cnattingius, "Five- and 10-minute Apgar scores and risks of cerebral palsy and epilepsy: A population-based cohort study in Sweden," BMJ, vol. 360, Art. no. k207, 2018.

[5] S. Eun, J. M. Lee, D. Y. Yi et al., "Assessment of the association between Apgar scores and seizures in infants less than 1 year old," Seizure, vol. 37, pp. 48–54, 2016.

[6] F. Pisani, C. Facini, E. Bianchi, G. Giussani, B. Piccolo, and E. Beghi, "Incidence of neonatal seizures, perinatal risk factors for epilepsy and mortality after neonatal seizures in the province of Parma, Italy," Epilepsia, vol. 59, no. 9, pp. 1764–1773, 2018.

[7] F. M. Doandes, A. M. Manea, N. Lungu, T. Brandibur, D. Cioboata, O. C. Costescu, M. Zaharie, and M. Boia, "The role of amplitude-integrated electroencephalography (aEEG) in monitoring infants with neonatal seizures and predicting their neurodevelopmental outcome," Children, vol. 10, no. 5, Art. no. 833, 2023.

[8] R. M. Pressler, N. S. Abend, S. Auvin, et al., "Treatment of seizures in the neonate: Guidelines and consensus-based recommendations—Special report from the ILAE Task Force on Neonatal Seizures," Epilepsia, vol. 64, no. 10, pp. 2550–2570, 2023.

[9] L. Battig, C. Dunner, D. Cserpan, et al., "Levetiracetam versus phenobarbital for neonatal seizures: A retrospective cohort study," Pediatr. Neurol., vol. 138, pp. 62–70, 2023.

[10] S. R. Mathieson, V. Livingstone, E. Low, R. M. Pressler, J. M. Rennie, and G. B. Boylan, "Phenobarbital reduces EEG amplitude and propagation of neonatal seizures but does not alter performance of automated seizure detection," Clin. Neurophysiol., vol. 127, no. 10, pp. 3343–3350, 2016.

[11] S. Mikulovic, et al., "On the photovoltaic effect in local field potential recordings," Neurophotonics, vol. 3, no. 1, Art. no. 015002, 2016.

[12] G. F. T. Variane, J. P. V. Camargo, D. P. Rodrigues, et al., "Current Status and Future Directions of Neuromonitoring With Emerging Technologies in Neonatal Care," Front. Pediatr., vol. 9, Art. no. 755144, 2022.

[13] M. El-Dib, N. S. Abend, T. Austin, et al., "Neuromonitoring in neonatal critical care part II: Extremely premature infants and critically ill neonates," Pediatr. Res., vol. 93, pp. 1568–1580, 2023.

Plataformas Abertas de Ultrassom para Pesquisa: Arquiteturas, Componentes Comerciais e Tendências

João Pedro da Silva Junior
CPGEI, UTFPR
Curitiba, Brasil
joaopedrojunior@alunos.utfpr.edu.br

Eduardo Tavares Costa
DEEB, CEB, UNICAMP
Campinas, Brasil
educosta@unicamp.br

Joaquim Miguel Maia
CPGEI, DAELN, UTFPR
Curitiba, Brasil
joaquim@utfpr.edu.br

Amauri Amorin Assef
CPGEI, DAELT, UTFPR
Curitiba, Brasil
amauriassef@utfpr.edu.br

***Abstract*—Open research ultrasound systems play a fundamental role in the development and validation of advanced biomedical imaging techniques by providing flexible access to radio-frequency (RF) signals, acquisition parameters, and signal processing algorithms. However, the increasing complexity of architectures and the diversity of electronic components make it difficult to compare the solutions reported in the literature. This paper presents a review of the state of the art in the electronic architecture of open research ultrasound platforms, covering the transmission, reception, signal conditioning, digitization, and digital processing stages. The main characteristics of high-voltage pulsers, multiplexers, T/R switches, analog front-end (AFE) circuits, and programmable logic devices are analyzed, focusing on operating voltage, channel density, operating frequency, integration level, and configuration flexibility. The analysis reveals a clear trend toward higher analog integration (multichannel AFEs with JESD204B/C interfaces), reduced use of discrete components, and the consolidation of hybrid FPGA/GPU/CPU architectures interconnected by high-speed buses (PCIe) for real-time reconstruction. The results provide a consolidated overview that serves as a reference guide for component selection and the design of new acquisition and imaging platforms.**

***Index Terms*—ultrasound imaging, open ultrasound systems, electronic components, analog front-end, beamforming**

I. Introdução

Os sistemas de ultrassom para diagnóstico não invasivo por imagem constituem uma das modalidades mais consolidadas de imageamento biomédico, combinando segurança, portabilidade e capacidade de fornecer informações anatômicas e funcionais de tecidos biológicos em tempo real. A contínua evolução desses sistemas é impulsionada pelo avanço de arquiteturas eletrônicas dedicadas, técnicas complexas de processamento de sinais e uma maior densidade de integração de componentes [1], [2].

A integração de eletrônica analógica de alta velocidade, aquisição multicanal e circuitos dedicados, como matrizes de portas programáveis em campo (*Field-Programmable Gate Array* – FPGA), pulsadores integrados de alta tensão e *analog front-ends* (AFEs), consolidou uma nova geração de plataformas de ultrassom para pesquisa altamente reconfiguráveis [3], [4]. Tais sistemas oferecem acesso direto aos dados brutos de radiofrequência (RF) e controle avançado dos parâmetros de excitação e amostragem, permitindo a validação de algoritmos customizados de *beamforming* na transmissão (Tx/T) e na recepção (Rx/R). No entanto, a consequente heterogeneidade das soluções desenvolvidas faz com que os detalhes de implementação e desempenho permaneçam dispersos na literatura, impondo barreiras para uma análise comparativa sistemática [5].

Este artigo apresenta uma revisão analítica e comparativa sobre a arquitetura eletrônica e os componentes empregados na construção de sistemas abertos de ultrassom para pesquisa. As principais contribuições deste trabalho incluem: (i) um panorama comparativo das plataformas abertas comerciais e acadêmicas consolidadas; (ii) uma análise detalhada dos circuitos TX, chaveamento T/R, AFEs e interfaces de comunicação; e (iii) a identificação das tendências tecnológicas de integração (como AFEs de alta densidade e RFSoCs) e os *trade-offs* entre flexibilidade, consumo energético e largura de banda.

II. Metodologia de Revisão e Critérios de Seleção

Esta revisão foi conduzida com o objetivo de sistematizar as principais arquiteturas eletrônicas e os componentes empregados em plataformas abertas de ultrassom destinadas à pesquisa e ao desenvolvimento de novas técnicas de imageamento. A revisão considerou tanto plataformas comerciais de pesquisa quanto sistemas acadêmicos e de código aberto que apresentassem informações técnicas suficientes para caracterizar sua arquitetura eletrônica. A busca bibliográfica foi realizada nas bases IEEE Xplore, ScienceDirect, PubMed, Scopus e Web of Science, complementada por buscas direcionadas nos sítios dos fabricantes para obtenção de folhas de dados e documentação técnica dos componentes. Foram considerados trabalhos publicados no período de 1990 a 2026. A estratégia de busca combinou termos relacionados à arquitetura de sistemas de ultrassom e aos principais blocos eletrônicos analisados. Foram empregados, individualmente ou em combinações, termos como “open ultrasound system”, “research ultrasound platform”, “programmable ultrasound”, “ultrasound front-end”, “ultrasound pulser”, “high-voltage multiplexer”, “T/R switch”, “ultrasound AFE”, “ultrasound FPGA”, “ultrasound RFSoC” e “ultrasound beamforming”. As buscas foram posteriormente refinadas de acordo com o bloco funcional analisado.

Foram incluídos estudos que apresentassem pelo menos uma das seguintes características: (i) descrição de uma plataforma de ultrassom reconfigurável ou destinada à pesquisa; (ii) descrição ou aplicação documentada de compo-

nentes eletrônicos utilizados em sistemas de ultrassom; (iii) apresentação de parâmetros elétricos ou arquiteturais suficientes para comparação; ou (iv) descrição de arquiteturas de aquisição e processamento relevantes para sistemas abertos de ultrassom. Foram excluídos trabalhos que (i) descrevessem exclusivamente equipamentos clínicos comerciais sem acesso ou flexibilidade para pesquisa; e (ii) não apresentassem informações técnicas suficientes para caracterizar a arquitetura analisada. A seleção das plataformas baseou-se na relevância para arquiteturas reconfiguráveis e no uso documentado na literatura. Para comparar os sistemas de forma holística, a análise relaciona as especificações de hardware extraídas (eletrônica e processamento) a parâmetros de desempenho, como fluxo de dados, escalabilidade, capacidade de aquisição, processamento e eficiência arquitetural, nas aplicações de ultrassom.

III. Arquitetura Geral de um Sistema de Ultrassom para Pesquisa

A cadeia de sinal de um sistema de ultrassom amostrado por canal é dividida em três domínios funcionais principais: transmissão dos sinais de excitação ao transdutor, recepção dos ecos provenientes dos tecidos e processamento digital para a formação e reconstrução das imagens [1], [6], [7].

Na etapa de transmissão, pulsadores de alta tensão geram os sinais elétricos aplicados aos elementos do transdutor para a emissão das ondas acústicas. Dependendo da arquitetura adotada, multiplexadores de alta tensão (*High-Voltage Multiplexers* – HV-MUXs) podem ser empregados para multiplexar múltiplos elementos em um número reduzido de canais eletrônicos, otimizando o custo e a complexidade do sistema [4], [8]. Durante a recepção, os ecos refletidos são direcionados ao circuito de aquisição por meio de uma chave T/R, que protege os circuitos sensíveis de recepção durante os pulsos de alta tensão de transmissão e encaminha os sinais de eco recebidos ao estágio de condicionamento analógico. Em seguida, os sinais são amplificados, filtrados e digitalizados para posterior reconstrução da imagem [9], [10]. Em muitas arquiteturas modernas, essas funções analógicas e de conversão são consolidadas em CIs AFE. Esses dispositivos multicanais reúnem amplificadores de baixo ruído (*Low-Noise Amplifiers* – LNAs), amplificadores de ganho variável (*Variable Gain Amplifiers* – VGAs) acoplados ao circuito de compensação de ganho no tempo (*Time Gain Compensation* – TGC), filtros passa-baixas (FPB) anti-aliasing (*Anti-Aliasing Filters* – AAFs) e conversores analógico-digitais (*Analog-to-Digital Converters* – ADCs) em dispositivos multicanais [11], [12]. Após a digitalização, os dados de RF são transferidos para unidades de processamento digital baseadas em FPGAs, unidades centrais de processamento (*Central Processing Units* – CPUs) e unidades de processamento gráfico (*Graphics Processing Units* – GPUs). Esses dispositivos são responsáveis pelo *beamforming* de recepção, incluindo as etapas de correção de atraso temporal, filtragem, somatório coerente, detecção de envelope, compressão logarítmica e execução de algoritmos avançados de reconstrução em tempo real [13], [14].

A. Plataformas Abertas para Pesquisa

As plataformas abertas variam substancialmente em poder computacional, número de canais e estratégia de Tx e Rx [3], [13], [15]–[17], [19]–[22]. A Tabela I apresenta uma síntese das principais características das plataformas comerciais, acadêmicas e de código aberto utilizadas pela comunidade científica. Observa-se que essas plataformas abrangem desde sistemas com 64 canais de recepção, como o us4R-lite, até arquiteturas com 1024 canais, como o SARUS, evidenciando a ampla variedade de soluções disponíveis para aplicações de pesquisa. Dentre as plataformas comerciais de maior porte, o Verasonics Vantage, que possui configurações de 128 e 256 canais, utiliza uma interface PCI Express de alta largura de banda para transferência dos dados ao computador hospedeiro com GPU, viabilizando aplicações como ultrassom ultrarrápido, elastografia e imageamento volumétrico [16], [17], [22], [23]. Em contrapartida, projetos portáteis e *open-source*, como o ULA-OP, utilizam arquiteturas modulares baseadas em FPGAs intermediárias e conectividade USB 2.0 ou 3.0, sacrificando a amostragem paralela massiva em prol do custo acessível e consumo reduzido [15], [20]. As características de processamento digital e armazenamento das plataformas apresentadas na Tabela I são complementadas pelos recursos de FPGA, memória e interfaces de comunicação resumidos na Tabela VI.

B. Transmissão Ultrassônica

A etapa de transmissão visa excitar os elementos piezoelétricos com pulsos de alta tensão (entre 10 e 200 Vpp) e frequências na faixa de 1 MHz a 50 MHz para gerar o campo acústico desejado no meio biológico [1], [7]. A seleção dos dispositivos de chaveamento e excitação depende de parâmetros críticos, tais como o número de canais, faixa de tensão de saída, capacidade de corrente e resistência no estado de condução (R_{ON}) da chave T/R integrada, analisados a seguir para os principais pulsadores integrados e multiplexadores de alta tensão disponíveis na literatura e no mercado.

1) Pulsadores para Sistema de Imagem por Ultrassom: Os pulsadores de alta tensão geram sinais de elevada amplitude e rápido tempo de subida para excitação dos elementos piezoelétricos [24]. As arquiteturas variam entre soluções unipolares, bipolares, multinível e geradores de formas de onda arbitrárias (*Arbitrary Waveform Generators* – AWGs) ou lineares [4], [7], [21], [25], frequentemente integrando *drivers* de potência, amortecimento ativo e retorno a zero (*Return to Zero* – RTZ) no mesmo encapsulamento para mitigar o *ringing* acústico [4], [15]. O mercado de circuitos integrados (CIs) para ultrassom médico é altamente especializado e vem evoluindo em direção a soluções de altíssima densidade. Diversas empresas consolidaram-se no fornecimento de CIs para aplicações de excitação ultrassônica. Esses dispositivos diferenciam-se quanto à excursão de tensão, resistência em condução (R_{ON}), número de canais, frequência máxima, capacidade de corrente e chave T/R integrada. A Tabela II sintetiza as principais características desses componentes.

Tabela I
CARACTERÍSTICAS DAS PRINCIPAIS PLATAFORMAS DE ULTRASSOM PARA PESQUISA DESCRITAS NA LITERATURA[1].

Plataforma	Referência	Número de canais (Tx/Rx)	Tensão Tx (V)	Forma de onda Tx	Faixa de freq. Tx (MHz)	Taxa de amostragem (MHz)
us4R-lite	[22]	256 / 64	±90	3 níveis	0,1 a 50	50
ULA-OP 128	[15]	128 / 128	±12	Linear	∼ 15	50
Verasonics Vantage 256	[16], [23]	256 / 256	±100	3 níveis	0,5 a 20	62,5
Cephasonics cQuest	[17], [18]	256 / 256	±100	3 ou 5 níveis	0,5 a 18	50
DIPhAS	[19]	256 / 256	±100	3 níveis	20	80
ULA-OP 256	[20]	256 / 256	±100	Linear	1 a 20	78,13
UARP	[21]	256 / 256	±100	5 níveis	0,5 a 15	80
SARUS	[13]	1024 / 1024	±100	3 níveis	1 a 30	70

[1]Os parâmetros apresentados foram obtidos a partir das referências bibliográficas e, quando disponíveis. Valores provenientes de diferentes fontes podem ter sido especificados sob condições de operação distintas e, portanto, não representam necessariamente uma comparação experimental direta.

Tabela II
PRINCIPAIS CARACTERÍSTICAS DOS PULSADORES DE ALTA TENSÃO EMPREGADOS EM SISTEMAS DE ULTRASSOM PARA PESQUISA.

Pulsador	Referência	Número de canais	Tensão de saída (V)	Corrente de saída (A)	Níveis do sinal	Freq. máx. (MHz)	Chave T/R int. ($R_{\mathbf{ON}}[\Omega]$)	Fabricante
HV7360	[29]	1	±100	±2,5	3	35	Não	Microchip
HV7361	[25], [29]	1	±100	±2,5	3	35	Sim (15)	Microchip
HV7321	[30]	4	±80	±2,5	5	20	Sim (8)	Microchip
STHV748S	[31]	4	±90	±2	3 / 5	20	Sim (13,5)	STMicroelectronics
MAX14808	[32], [33]	8 / 4	±105	±2	3 / 5	5	Sim (11,5)	Analog Devices
MAX14809	[32], [33]	8 / 4	±105	±2	3 / 5	5	Não	Analog Devices
STHV800	[34]	8	±90	±2	3	20	Sim (8)	STMicroelectronics
TX7316	[35]	16 / 8	±100	±2,4	3 / 5	36	Sim (12)	Texas Instruments
TX7516	[36], [37]	16	±100	±2	5	35	Sim (8)	Texas Instruments
STHV1600	[38]	16	±100	±4 / ±2	3 / 5	30	Sim (9)	STMicroelectronics
TX7332	[39]	32	±100	±1,2	3	20	Sim (24)	Texas Instruments
STHVUP32	[40]	32	±100	±0,8 / ±0,4	3 / 5	20	Sim (23)	STMicroelectronics
TX7364	[41]	64	±100	±1	3	22	Sim (26)	Texas Instruments
STHVUP64	[42]	64	±100	±0,4 / ±0,2	3 / 5	20	Sim (45)	STMicroelectronics

2) Multiplexadores de Alta Tensão: Para reduzir a complexidade e o custo em sistemas em que o número de elementos do transdutor excede os canais físicos de amostragem, empregam-se matrizes de chaves multiplexadoras de alta tensão [4], [26]. Para aplicações em ultrassom, esses dispositivos devem suportar os elevados níveis de tensão gerados pelos pulsadores, conforme os parâmetros de tensão de alimentação positiva V_{PP} e negativa V_{NN}, respeitando o limite de tensão diferencial, apresentar baixa R_{ON}, capacitância parasita reduzida e tempos de comutação compatíveis com a frequência de operação do sistema [4], [24]. Além disso, devem garantir elevada isolação entre os estágios de transmissão e recepção, minimizando o acoplamento dos pulsos de alta tensão aos circuitos sensíveis de aquisição [24]. A Tabela III apresenta uma análise comparativa dos principais multiplexadores reportados na literatura. Entre os parâmetros de seleção, destaca-se o número de chaves (canais), que determina a quantidade de elementos do transdutor controlados por um único CI. Embora dispositivos com maior densidade de canais reduzam o número de componentes e simplifiquem o roteamento da placa de circuito impresso, eles podem aumentar efeitos parasitas, como capacitâncias internas e acoplamento entre canais (*crosstalk*), exigindo cuidados adicionais no projeto eletrônico. Como funcionalidade adicional, determinados modelos, como o MAX14803 e o HV2901, incorporam resistores de drenagem integrados (*bleed resistors*), que promovem a dissipação da carga residual acumulada nos elementos capacitivos do transdutor piezoelétrico, evitando variações indesejadas na tensão de *offset* [27], [28].

C. Recepção Ultrassônica

Os ecos ultrassônicos apresentam amplitudes muito baixas, da ordem de μV a mV, exigindo elevada sensibilidade na aquisição [1]. A arquitetura também deve garantir proteção e isolamento dos estágios de recepção contra os transientes de alta tensão da transmissão. Nesse contexto, observa-se a crescente substituição de componentes discretos por circuitos altamente integrados, como chaves T/R e AFEs comerciais.

1) Chaves T/R de proteção: Para proteger os amplificadores do caminho de recepção contra os pulsos de alta tensão da transmissão, uma chave T/R é posicionada entre o transdutor e o AFE. Seu desempenho influencia diretamente a sensibilidade, a faixa dinâmica e a integridade dos sinais adquiridos. Assim, são desejáveis baixa R_{ON}, baixa capacitância parasita, curto tempo de comutação e elevada isolação entre os caminhos de transmissão e recepção, além de capacidade para suportar os níveis de tensão e corrente da excitação sem introduzir distorções significativas [24]. Alguns dispositivos comerciais incorporam funções adicionais de proteção e condicionamento, contribuindo para arquiteturas mais compactas

Tabela III
CARACTERÍSTICAS DOS PRINCIPAIS MULTIPLEXADORES DE ALTA TENSÃO UTILIZADOS EM SISTEMAS DE ULTRASSOM PARA PESQUISA.

Multiplexador	**Ref.**	**Número de canais**	V_{PP} **max. (V)**	V_{NN} **max. (V)**	$V_{\mathrm{PP}} - V_{\mathrm{NN}}$ **max. (V)**	**Corrente de pico (A)**	R_{ON} (Ω)	**Resistor de drenagem (kΩ)**	**Freq. da interf. serial (MHz)**	**Fabricante**
HV20220	[26], [43]	8	200	-200	220	3	22	Não	5	Microchip
HV2201	[44], [45]	8	200	-200	220	3	16–26	Não	20	Microchip
HV2705	[46]	16	200	-200	200	3	16–26	Sim (35)	20	Microchip
MAX14802	[27], [47]	16	220	-220	250	3	16–26	Não	20	Analog Devices
MAX14803	[27], [48]	16	100	-100	250	3	16–26	Sim (35)	20	Analog Devices
CPC7601	[49], [50]	16	100	-100	200	1	22–27	Não	8	Littelfuse
HV2801	[28], [51]	32	200	-200	200	3	16–26	Não	20	Microchip
HV2901	[28], [52]	32	200	-200	200	3	16–26	Sim (35)	20	Microchip

Tabela IV
CARACTERÍSTICAS DAS PRINCIPAIS CHAVES DE TRANSMISSÃO E RECEPÇÃO (T/R) UTILIZADAS EM SISTEMAS DE ULTRASSOM.

Chave T/R	**Ref.**	**Tensão máx. (V)**	**Canais**	**Largura de banda (MHz)**	**Proteção**	**Fabricante**
MD0100	[53], [54]	$\pm$100	1 / 2	100	Sim	Microchip
MD0101	[55], [56]	$\pm$100	4	100	Sim	Microchip
MD0105	[57], [58]	$\pm$130	4	100	Externa	Microchip
TX810	[59], [60]	$\pm$90	8	140	Sim	Texas Instruments
LM96530	[61]	$\pm$60	8	20	Sim	Texas Instruments
MP4816A	[62]	$\pm$90	16	20	Não	Monolithic Power Systems
MAX14866	[63]	$\pm$105	16	50	Não	Analog Devices

e escaláveis [4]. A Tabela IV apresenta uma comparação entre algumas das principais chaves T/R utilizadas em sistemas de ultrassom para pesquisa. Entre essas soluções, destacam-se os dispositivos das famílias MD01xx e Tx810, frequentemente relatados na literatura.

De modo geral, dispositivos mais integrados reduzem a complexidade do hardware, enquanto soluções com baixa capacitância parasita e elevada largura de banda são particularmente adequadas para aplicações de alta frequência [24]. Observa-se, portanto, uma tendência de integração das funções de chaveamento T/R, proteção contra sobretensão e condicionamento analógico em um único CI.

2) Circuitos Analog Front-End: Os circuitos AFE fazem a interface entre o transdutor e o processamento digital, realizando o condicionamento dos sinais de eco antes da conversão. Como esses sinais apresentam baixa amplitude e ampla faixa dinâmica devido à atenuação durante a propagação nos tecidos, os AFEs integram múltiplos estágios analógicos de alta velocidade. Tipicamente, esses dispositivos possuem de 8 a 32 canais e integram LNA, VGA para TGC (0–40 dB), AAF e ADCs de alta velocidade, com resoluções de 10 a 16 bits. O consumo pode variar de aproximadamente 35 mW a 225 mW por canal, dependendo da arquitetura e das funcionalidades integradas [4]. A adoção de interfaces seriais de alta velocidade, como JESD204B/C, também reduz o número de conexões na PCB e favorece arquiteturas com elevada densidade de canais. Essa característica é particularmente relevante em aplicações como ultrassom ultrarrápido e imageamento 3D [64].

A Tabela V apresenta uma comparação entre AFEs comerciais utilizados em pesquisa. A integração de múltiplos canais em um único CI reduz a quantidade de componentes e favorece arquiteturas mais compactas. Os dispositivos analisados apresentam taxas de amostragem de dezenas a centenas de MSPS e resoluções de 10 a 16 bits, atendendo a diferentes requisitos de largura de banda e faixa dinâmica. A seleção de um AFE deve considerar o número de canais, a faixa de entrada, a taxa de amostragem, a resolução, a SNR, a interface digital, o consumo por canal e o desempenho de ruído [17]. AFEs com maior resolução e desempenho analógico são particularmente adequados para aplicações que exigem elevada sensibilidade e detecção de ecos de baixa amplitude [64], [65]. Alguns dispositivos também incorporam atenuadores, memória de configuração, filtros passa-alta (HPF) e demodulação digital I/Q para Doppler de onda contínua (CWD), aumentando a flexibilidade e reduzindo a necessidade de processamento externo. A família AFE58xx, da Texas Instruments, destaca-se entre os circuitos AFE empregados em sistemas experimentais de ultrassom devido à integração de múltiplos canais de recepção e recursos dedicados, como mixer CWD, demodulação digital I/Q (membros da família AFE58JD) e interfaces digitais de alta velocidade. Dispositivos como o AFE58JD28 e o AFE58JD48 ampliam essa capacidade ao oferecer maior resolução de conversão e suporte a interfaces como JESD204B, enquanto o AFE5832 e o AFE58JD32 representam soluções de maior densidade, com 32 canais de aquisição integrados. Por outro lado, embora a separação entre os estágios de transmissão (pulsadores) e recepção (AFEs) seja a abordagem predominante para reduzir o acoplamento de ruído de alta tensão, a literatura também apresenta transceptores monolíticos que integram ambos os caminhos em um único dispositivo. O MAX2082 integra, em um encapsulamento de 8 canais, pulsadores de alta tensão de 3 níveis, chaves T/R e uma cadeia completa de recepção e reduzindo a quantidade de componentes externos [66]. Já MAX2084 é um AFE com chaves T/R integradas, proporcionando maior densidade de integração e simplificação do *front-end* ultrassônico [67].

Tabela V
CARACTERÍSTICAS DOS PRINCIPAIS CIRCUITOS AFE UTILIZADOS EM SISTEMAS DE ULTRASSOM PARA PESQUISA[1].

AFE	Ref.	Saída de dados	PGA	LPF	VGA	AAF	CWD Mixer	HPF	f_{ADC} (MSPS)	Resolução (bits)	Faixa de entrada (V_{pp})	SNR (dBFS)	Consumo típico (mW/canal)	Fabricante
8 canais														
MAX2079	[68]	LVDS	✓	✓	✓	✓	✓	✓	50	12	0, 27	76	120	Analog Devices
MAX2082	[66]	LVDS	✓	✓	✓	✓	✓	✓	50	12	0, 27	76	131	Analog Devices
AFE5801	[69]	LVDS	✓	✓	✓	✓	–	–	65	12	0, 5	66	58	Texas Instruments
AFE5805	[7], [70]	LVDS	✓	✓	✓	✓	–	–	50	12	0, 4	69, 5	122	Texas Instruments
AFE5807	[71]	LVDS	✓	✓	✓	✓	✓	✓	80	12	0, 4	70	88	Texas Instruments
AFE5808	[11], [72]	LVDS	✓	✓	✓	✓	✓	✓	65	14 / 12	0, 4	77	158	Texas Instruments
AFE5808A	[73]	LVDS	✓	✓	✓	✓	✓	✓	65	14 / 12	0, 8	77	158	Texas Instruments
AD9271	[74], [75]	LVDS	✓	✓	✓	✓	–	✓	50	12	0, 625	63, 7	150	Analog Devices
AD9272	[76], [77]	LVDS	✓	✓	✓	✓	–	✓	80	12	0, 8	63	225	Analog Devices
16 canais														
MAX2084	[67]	LVDS	✓	✓	✓	✓	✓	✓	65	14	0, 27	81	135	Analog Devices
AFE58JD16	[78]	LVDS / JESD204B	✓	✓	✓	–	✓	✓	65 / 80	14 / 12	1, 0	75 / 72	90 / 55	Texas Instruments
AFE58JD28	[79]	LVDS / JESD204B	✓	✓	✓	–	✓	–	65 / 80	14 / 12	0, 71	75	45	Texas Instruments
AFE58JD48	[80]	LVDS / JESD204B	✓	✓	✓	–	✓	–	125 / 80	16 / 14	0, 71	80 / 79	140	Texas Instruments
32 canais														
AFE5832	[81]	LVDS	✓	✓	✓	–	✓	–	40 / 50	12 / 10	0, 8	72 / 61	35–52	Texas Instruments
AFE58JD32	[65], [82]	LVDS / JESD204B	✓	✓	✓	–	✓	–	40 / 50	12 / 10	0, 8	72 / 61	35–52	Texas Instruments

[1] Alguns dispositivos oferecem modos de baixo consumo e de alto desempenho, com diferentes níveis de consumo e ruído.

D. Dispositivos FPGA e Processamento Digital

Em plataformas abertas de pesquisa, a demanda por flexibilidade, elevada taxa de aquisição e processamento em tempo real consolidou o uso de arquiteturas híbridas compostas por FPGAs, CPUs e GPUs [17]. Após o condicionamento analógico e a digitalização no AFE, os sinais de RF ultrassônicos digitalizados são submetidos ao processamento digital, responsável pelo controle temporal da aquisição, *beamforming* e reconstrução da imagem final [4].

Entretanto, a taxa de dados gerada pelos ADCs em recepção simultânea representa o principal gargalo de processamento. Por exemplo, para 128 canais amostrados a 50 MSPS com 12 bits, a taxa contínua de dados atinge 128×50 MHz $\times$ 12 bits $\approx$ 76,8 Gbps ($\approx$ 9,6 GB/s), desconsiderando o *overhead* de protocolo. A transferência de dados entre os AFEs e o processador digital migrou das interfaces seriais LVDS paralelas tradicionais para os padrões seriais de ultra-alta velocidade JESD204B/C. O padrão JESD204 reduz drasticamente o número de linhas de sinal na PCB, simplificando o roteamento físico ao custo de exigir *transceivers* seriais dedicados na FPGA. As FPGAs desempenham papel fundamental no *front-end* digital devido à execução determinística, baixa latência e capacidade de processamento paralelo reconfigurável. Esses dispositivos são empregados no controle preciso dos tempos de disparo da transmissão, na sincronização multicanal, no pré-processamento dos dados de RF e na implementação inicial de algoritmos *beamforming* de recepção [13], [14], [61].

A Tabela VI apresenta os principais dispositivos lógicos programáveis e memórias (SRAM e DRAM) empregados nas plataformas de pesquisa, destacando a capacidade de armazenamento local necessária para acomodar o elevado fluxo de dados de RF antes da transferência para a unidade central de processamento [13], [15], [20], [21]. A integração entre os FPGAs responsáveis pela aquisição dos dados brutos RF ou IQ pré-filtrados e os processadores/GPUs do computador hospedeiro é realizada por meio de barramentos de alta velocidade e baixa latência, como o PCI Express (Gen3/Gen4 x4/x8), USB 2.0, USB 3.0 ou Thunderbolt 3 para o computador hospedeiro [16], [22]. No *back-end*, a elevada densidade de dados gerada por aplicações como o imageamento por ondas planas (*Plane-Wave Imaging* – PWI), o ultrassom ultrarrápido, a elastografia e o imageamento volumétrico (3D) impõe elevada carga computacional, podendo impor limitações de desempenho, dependendo da aplicação e do algoritmo empregado, a arquiteturas baseadas exclusivamente em CPUs ou FPGAs para a execução em tempo real desses algoritmos [16]. Nesse contexto, GPUs de propósito geral são amplamente empregadas na reconstrução intensiva de imagens. Em particular, dispositivos NVIDIA, utilizando a plataforma CUDA, permitem a execução em tempo real de algoritmos clássicos, como *Delay-and-Sum* (DAS) e reconstrução por síntese de Fourier baseada na Transformada Rápida de Fourier (FFT), bem como de métodos avançados de maior complexidade computacional, como *Delay-Multiply-and-Sum* (DMAS), *Filtered Delay-Multiply-and-Sum* (FDMAS) e estimadores de Mínima Variância (*Minimum Variance* – MV) [83].

IV. DISCUSSÃO

A análise das plataformas e componentes apresentados evidencia que a evolução dos sistemas abertos de ultrassom tem sido impulsionada pela maior integração dos circuitos eletrônicos e pelo avanço do processamento digital. Conforme discutido por Chen e Pertijs [4] e Lok e Li [5], o desenvolvimento de CIs específicos para ultrassom reduziu a complexidade da eletrônica analógica, aumentando a escalabilidade, modularidade e flexibilidade das plataformas experimentais.

Na transmissão, observa-se a evolução de pulsadores discretos para CIs multicanais que integram chaves T/R e recursos programáveis de excitação. Essa integração reduz a área da PCB, simplifica o roteamento e favorece plataformas compactas, como a us4R-lite [22]. Além disso, arquiteturas baseadas em AWGs ampliam a flexibilidade para técnicas como transmissão codificada, abertura sintética e formação de feixe adaptativa [7], [25]. Multiplexadores de alta tensão

Tabela VI
CARACTERÍSTICAS DOS DISPOSITIVOS FPGA, MEMÓRIA E INTERFACES DE COMUNICAÇÃO DAS PRINCIPAIS PLATAFORMAS DE ULTRASSOM PARA PESQUISA.

Plataforma	Ref.	FPGA	SRAM (por FPGA)	DRAM	Interface de comunicação
us4R-lite	[22]	Xilinx Kintex-7	34 Mbit	8 GB	PCIe Gen3 / Thunderbolt 3
ULA-OP 128	[15]	Altera Stratix II	4.8 Mbit/FPGA	1 GB	USB 2.0
Verasonics Vantage 256	[16], [23]	Altera Stratix V	45 Mbit	64 GB	PCIe Gen3 x8
Cephasonics cQuest	[17], [18]	Altera Cyclone IV	6.5 Mbit/FPGA	4 GB	PCIe / Gigabit Ethernet
DIPhAS	[19]	Xilinx Spartan-6	4.8 Mbit/FPGA	512 MB	PCIe / Gigabit Ethernet
ULA-OP 256	[20]	Arria V	19 Mbit/FPGA	80 GB	USB 3.0
UARP	[21]	Altera Stratix V	45 Mbit	256 MB/canal	PCIe Gen3 / USB 2.0
SARUS	[13]	Xilinx Virtex-6	19 Mbit/FPGA	2 GB/FPGA	PCIe Gen2 / Gigabit Ethernet

também permitem reduzir o número de canais eletrônicos necessários, sendo particularmente relevantes em transdutores matriciais para imageamento 2D e 3D [8], [26]. Entretanto, sua utilização envolve compromissos relacionados à resistência de condução, capacitâncias parasitas e acoplamento entre canais [24].

Na recepção, a integração de LNAs, VGAs, filtros, TGCs e ADCs em AFEs multicanais representa um dos principais avanços dos sistemas modernos. Essa abordagem reduz componentes externos, consumo e área, viabilizando sistemas com centenas de canais [11], [12], [14]. Interfaces digitais de alta velocidade, como LVDS e JESD204B, também aumentam a taxa de transferência e simplificam a comunicação com FPGAs [14], [65]. Outra tendência é a adoção de arquiteturas híbridas com FPGAs, CPUs e GPUs. Plataformas como SARUS [13], ULA-OP 256 [20], Verasonics Vantage [16] e us4R-lite [22] combinam a baixa latência e o processamento determinístico das FPGAs com o paralelismo das GPUs, permitindo aplicações computacionalmente intensivas, como ultrassom ultrarrápido, abertura sintética, elastografia e formação de feixe adaptativa [61].

Como evolução dessas arquiteturas, destaca-se a adoção de dispositivos SoC, como Xilinx Zynq UltraScale+ MPSoC, RFSoC, Intel Cyclone V SoC e Agilex, que integram processadores ARM e lógica programável, e, em alguns casos, ADCs e DACs de alta velocidade [64], [84]. Essa integração pode reduzir a latência, a área ocupada na placa e a necessidade de interconexões externas entre os blocos de conversão, processamento e memória, favorecendo o processamento em tempo real. Em conjunto, as plataformas analisadas apontam para uma convergência em direção a sistemas mais integrados, modulares, escaláveis e reconfiguráveis. A combinação de eletrônica analógica integrada, interfaces de alta velocidade e processamento paralelo possibilita maior número de canais, menor consumo e maior capacidade computacional [85]. Essas características são especialmente relevantes para a próxima geração de sistemas de pesquisa, incluindo aplicações de imageamento 3D, ultrassom portátil, dispositivos vestíveis e sistemas baseados em inteligência artificial [86]–[88].

V. CONCLUSÃO

Este trabalho apresentou uma revisão das principais arquiteturas eletrônicas e componentes empregados em plataformas abertas de ultrassom para pesquisa, abrangendo transmissão, multiplexação, proteção T/R, condicionamento analógico, conversão analógico-digital e processamento. A análise evidencia uma evolução em direção a sistemas com maior integração, densidade de canais e taxas de transferência. A seleção da arquitetura, entretanto, deve considerar conjuntamente número de canais, largura de banda, resolução dos ADCs, consumo, capacidade de processamento, complexidade de roteamento, flexibilidade e custo. Destaca-se também a crescente integração de conversores de alta velocidade, lógica programável e processadores em arquiteturas SoC e RFSoC, que podem reduzir latência, área e dependência de enlaces externos. Paralelamente, arquiteturas híbridas com FPGA, CPU e GPU permitem o processamento em tempo real de técnicas computacionalmente intensivas, como *beamforming* adaptativo, ultrassom ultrarrápido e reconstrução avançada de imagens. Espera-se que este estudo sirva como referência para o desenvolvimento e dimensionamento de novas plataformas experimentais, contribuindo para a seleção de componentes e arquiteturas adequados às diferentes aplicações. Essas tendências favorecem o avanço de sistemas portáteis, dispositivos *point-of-care* e plataformas de imageamento 3D e 4D [16], [26].

AGRADECIMENTOS

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná e à Secretaria da Ciência, Tecnologia e Ensino Superior do Paraná (SETI) pelo apoio concedido.

REFERÊNCIAS

[1] J. A. Jensen, "Ultrasound imaging and its modeling," in *Imaging of Complex Media with Acoustic and Seismic Waves*, Berlin, Germany: Springer, 2002, pp. 135–165.

[2] J. A. Zagzebski, Biomedical Ultrasound. San Diego, CA, USA: Academic Press, 1996.

[3] E. Boni et al., "Ultrasound open platforms for next-generation imaging technique development," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 65, no. 7, pp. 1078–1092, 2018.

[4] C. Chen and M. A. P. Pertijs, "Integrated transceivers for emerging medical ultrasound imaging devices: A review," *IEEE Open J. Solid-State Circuits Soc.*, vol. 1, pp. 104–114, 2021, doi: 10.1109/OJSSCS.2021.3069406.

[5] U.-W. Lok and P.-C. Li, "Transform-Based Channel-Data Compression to Improve the Performance of a Real-Time GPU-Based Software Beamformer," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 63, no. 3, pp. 369–379, Mar. 2016.

[6] S. Krishnan, "Introduction to Ultrasound," Texas Instruments, Application Report SLOA252, Nov. 2017.

[7] A. A. Assef, "Arquitetura de hardware multicanal reconfigurável com excitação multinível para desenvolvimento e testes de novos métodos de geração de imagens por ultrassom," Tese de Doutorado, Universidade Tecnológica Federal do Paraná, Curitiba, PR, Brasil, 2014.

[8] Y. Zhang and A. Demosthenous, "Integrated Circuits for Medical Ultrasound Applications: Imaging and Beyond,"in IEEE Transactions on Biomedical Circuits and Systems, vol. 15, no. 5, pp. 838-858, Oct. 2021.

[9] G. Peyton, M. G. Boutelle, and E. M. Drakakis, "Front-End Receiver Architecture for Miniaturised Ultrasound Imaging,"in Proceedings of the 3rd World Congress on Electrical Engineering and Computer Systems and Science (EECSS'17), International Conference on Biomedical Engineering and Systems (ICBES), Rome, Italy, Jun. 4–6, 2017, Paper 133.

[10] M. I. Fuller, T. N. Blalock, J. A. Hossack and W. F. Walker, "Novel transmit protection scheme for ultrasound systems,"in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 54, no. 1, pp. 79-86, January 2007.

[11] G. Pandey and A. Vora, "Open Electronics for Medical Devices: State-of-Art and Unique Advantages,"Electronics, vol. 8, no. 11, Art. no. 1256, Nov. 2019.

[12] Y. Zhang, M. Chen, and Z. Yang, "Design and Fabrication of Ultrasound Linear Array Transducer Used in Ultrasound Endoscope,"arXiv preprint arXiv:2507.18628, Jul. 2025.

[13] J. A. Jensen, H. Holten-Lund, R. T. Nilsson, M. Hansen, U. D. Larsen, R. P. Domsten, B. G. Tomov, M. B. Stuart, S. I. Nikolov, M. J. Pihl, Y. Du, J. H. Rasmussen, and M. F. Rasmussen, "SARUS: A Synthetic Aperture Real-Time Ultrasound System,"IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 60, no. 9, pp. 1838–1852, Sep. 2013.

[14] E. Batbayar, J. Y. Lee, W. Ham, E. Tumenjargal, C. Song, and S. Roh, "Implementation of Medical Image Reconstruction Algorithm for Photoacoustic Imaging Using k-Wave Toolkit,"in Proc. IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Washington, DC, USA, Oct. 2017.

[15] P. Tortoli et al., "ULA-OP: An advanced open platform for ultrasound research," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 56, no. 10, pp. 2207–2216, Oct. 2009.

[16] J. Yu, H. Yoon, Y. M. Khalifa, and S. Y. Emelianov, "Design of a Volumetric Imaging Sequence Using a Vantage-256 Ultrasound Research Platform Multiplexed With a 1024-Element Fully Sampled Matrix Array,"IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, no. 2, pp. 248–257, Feb. 2020.

[17] M. Behnam, L. A. J. N. van Lier, and M. D. Verweij, "Characterization and RF Data Acquisition Using a 64-Channel Open Ultrasound System," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 65, no. 12, pp. 2235–2244, Dec. 2018.

[18] Cephasonics Ultrasound, "Cephasonics Ultrasound Research Platform (cQuest)," [Online]. Disponível em: https://www.cephasonics.com/systems.html. Acesso: Jul. 2026.

[19] C. Risser, H. Hewener, M. Fournelle, H. Fonfara, S. Barry-Hummel, S. Weber, D. Speicher, and S. Tretbar, "Real-Time Volumetric Ultrasound Research Platform with 1024 Parallel Transmit and Receive Channels,"Applied Sciences, vol. 11, no. 13, Art. no. 5795, Jun. 2021.

[20] E. Boni, L. Bassi, A. Dallai, F. Guidi, V. Meacci, A. Ramalli, S. Ricci, and P. Tortoli, "ULA-OP 256: A 256-Channel Open Scanner for Development and Real-Time Implementation of New Ultrasound Methods,"IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 63, no. 10, pp. 1488–1495, Oct. 2016.

[21] P. R. Smith, D. M. J. Cowell, B. Raiton, C. V. Ky, and S. Freear, "Ultrasound Array Transmitter Architecture With High Timing Resolution Using Embedded Phase-Locked Loops,"IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 59, no. 1, pp. 40–49, Jan. 2012.

[22] D. Cacko and M. Lewandowski, "Shear Wave Elastography Implementation on a Portable Research Ultrasound System: Initial Results,"Applied Sciences, vol. 12, no. 12, Art. no. 6210, Jun. 2022.

[23] Verasonics, Inc., The Vantage Systems: Unmatched Flexibility for Research Ultrasound. Kirkland, WA, USA: Verasonics, Mar. 2020. [Online].

[24] H. Peng, J. Sabate, K. A. Wall and J. S. Glaser, "GaN-Based High-Frequency High-Energy Delivery Transformer Push–Pull Inverter for Ultrasound Pulsing Application,"in IEEE Transactions on Power Electronics, vol. 33, no. 8, pp. 6794-6806, Aug. 2018.

[25] A. A. Assef, P. L. S. de Moura, J. M. Maia, P. Vu, A. O. Olomodosi, S. Strassle Rojas, and B. D. Lindsey, "An Open Multifunctional FPGA-Based Pulser/Receiver System for Intravascular Ultrasound (IVUS) Imaging and Therapy,"Sensors, vol. 25, no. 15, Art. no. 4599, Jul. 2025.

[26] S. Tamano, T. Kobayashi, S. Sano, K. Hara, J. Sakano, and T. Azuma, "3D Ultrasound Imaging System Using Fresnel Ring Array & High Voltage Multiplexer IC,"in Proceedings of the 2004 IEEE Ultrasonics Symposium, Montreal, QC, Canada, Sep. 2004, vol. 1, pp. 782–785.

[27] Analog Devices, Inc., MAX14802/MAX14803/MAX14803A: Low-Charge-Injection, 16-Channel, High-Voltage Analog Switches, Datasheet, Rev. 5, Fev. 2021.

[28] Microchip Technology Inc., HV2801/HV2901: 32-Channel Low-Charge-Injection High-Voltage Analog Switches Datasheet, DS20005840A, Rev. A, Jul. 2019.

[29] Microchip Technology Inc., "HV7360/HV7361 High-Speed ±100V 2.5A Two-or-Three-Level Ultrasound Pulsers," DS20005570C, 2017.

[30] Microchip Technology Inc., "HV7321 4-Ch. 5-Level ±80V High-Voltage Ultrasound Pulser with T/R Switches," DS20005639A, 2016.

[31] STMicroelectronics, "STHV748S: Quad ±90 V, ±2 A, 3/5-Level High-Speed Ultrasound Pulser," Data Brief, 2018.

[32] G. P. Enwia, Open-Source Miniaturized Test-Bench Design for Applications in Wearable Autonomous Ultrasound Imaging, M.S. thesis, Dept. of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA, 2020.

[33] Analog Devices, Inc., MAX14808/MAX14809: Octal Three-Level/Quad Five-Level High-Voltage 2A Digital Pulsers with T/R Switch Datasheet, Rev. 2, Jan. 2014.

[34] STMicroelectronics, "STHV800: Octal ±90 V, ±2 A, 3-Level RTZ, High-Speed Ultrasound Pulser," Datasheet, 2016.

[35] Texas Instruments, TX7316 Three-Level, 16-Channel or Five-Level, 8-Channel Transmitter With 2.4-A Pulser, T/R Switch, and Integrated Transmit Beamformer, Datasheet, Rev. B, Jul. 2019.

[36] N. Lukhanin, F. Xia, S. Isomatsu, M. Teng, B. Jiang, J.-D. Zanone, and L. Lin, "Biological Bone Age Assessment via PMUTs,"in 2022 IEEE International Ultrasonics Symposium (IUS), Venice, Italy, Oct. 2022, pp. 1–4.

[37] Texas Instruments, TX7516 Five-Level, 16-Channel Transmitter with T/R Switch, and On-Chip Beamformer, Datasheet, Rev. A, Mar. 2022.

[38] STMicroelectronics, "STHV1600: 16-Channel ±100 V, ±2/±4 A, 5/3-Level RTZ High-Speed Ultrasound Pulser with Integrated Transmit Beamformer," Datasheet, 2020.

[39] Texas Instruments, "TX7332 Three-Level, 32-Channel Transmitter With 1.2-A Pulser, T/R Switch, and Integrated Transmit Beamformer," Datasheet SBOS975, Mar. 2019.

[40] STMicroelectronics, STHVUP32: 32-Channel ±100 V, ±0.8/0.4 A, 3/5-Level RTZ, High-Speed Ultrasound Pulser with Integrated Beamformer, Datasheet, Rev. 3, 2023.

[41] Texas Instruments, "TX7364 Three-Level, 64-Channel Transmitter with On-Chip Beamformer, T/R Switch," Datasheet SBASB25, Apr. 2024.

[42] STMicroelectronics, "STHVUP64: 64-Channel ±100 V, ±0.2/±0.4 A, 3/5-Level RTZ, T/R Switch, High-Speed Ultrasound Pulser with Integrated Transmit Beamformer," Data Brief DB4800, Rev. 2, 2023.

[43] Supertex Inc., "HV20220: Low Charge Injection, 8-Channel High Voltage Analog Switches," Datasheet DSFP-HV20220, 2013.

[44] U. Techavipoo, P. Pinunsottikul, R. Keinprasit and P. Thajchayapong, "Simplified pseudo-dynamic receive beamforming for FPGA implementation,"2015 8th Biomedical Engineering International Conference (BMEiCON), Pattaya, Thailand, 2015, pp. 1-5.

[45] Supertex Inc., "HV2201: Low Charge Injection, 8-Channel Enhanced High Voltage Analog Switch," Datasheet DSFP-HV2201, 2008.

[46] Microchip Technology Inc., "HV2605/HV2705: 16-Channel Low Harmonic Distortion High-Voltage Analog Switches," Datasheet DS20005498C, 2022.

[47] M. Lewandowski, K. Sielewicz and M. Walczak, "A low-cost 32-channel module with high-speed digital interfaces for portable ul-

trasound systems,"2012 IEEE International Ultrasonics Symposium, Dresden, Germany, 2012, pp. 1-4.
[48] B. Peng, Y. Zheng, W. Ma, D. Wang, C. Wang and X. Cao, "Integrated Ultrasound Imaging System based on FPGA and Monocrystalline Phased-Array Probe,"2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS), Yanji, China, 2024, pp. 499-503.
[49] W. Y. Choi, Y. S. Kwak and K. K. Park, "3-D synthetic imaging using 128-channel 2-D sparse CMUT array,"2017 IEEE International Ultrasonics Symposium (IUS), Washington, DC, USA, 2017, pp. 1-4.
[50] IXYS Integrated Circuits Division, "CPC7601: Low Charge Injection, 16-Channel High Voltage Analog Switch," Datasheet, 2015.
[51] C. J. Martín-Arguedas, "A SoC-based fully configurable point-of-care ultrasound system for research purposes,"2019 IEEE International Ultrasonics Symposium (IUS), Glasgow, UK, 2019, pp. 815-817.
[52] S. Ahn, J. Kang, P. Kim, G. Lee, E. Jeong, W. Jung, M. Park e T.-K. Song, "Smartphone-based Portable Ultrasound Imaging System: Prototype Implementation and Evaluation,"in 2015 IEEE International Ultrasonics Symposium (IUS), Taipei, Taiwan, Oct. 2015.
[53] J. T. Yen, M. M. Nguyen, Y. Lou, J. S. Shin, Y. Chen and H. L. Tarnoff, "Gated Transmit and Fresnel-Based Receive Beamforming With a Phased Array for Low-Cost Ultrasound Imaging,"in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 68, no. 6, pp. 2183-2192, June 2021.
[54] Microchip Technology Inc., "MD0100: Single/Dual-Channel High-Voltage Protection T/R Switch," Datasheet DS20005738A, 2018.
[55] G. Li et al., "Imaging-Guided Dual-Target Neuromodulation of the Mouse Brain Using Array Ultrasound,"in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 65, no. 9, pp. 1583-1589, Sept. 2018.
[56] Microchip Technology Inc., "MD0101: High-Voltage Protection T/R Switch with Clamp Diodes," Datasheet DS20005916A, 2020.
[57] J. Spaulding, Y. C. Eldar and B. Murmann, "A sub-nyquist analog front-end with subarray beamforming for ultrasound imaging,"2015 IEEE International Ultrasonics Symposium (IUS), Taipei, Taiwan, 2015, pp. 1-4.
[58] Microchip Technology Inc., "MD0105: 4-Channel High-Voltage Protection T/R Switch," Datasheet DS20005739A, 2018.
[59] R. K. Sagdiev, E. S. Denisov, Y. K. Evdokimov, M. G. Fazlyyyakhmatov e N. F. Kashapov, "Phased Array Based Ultrasound Scanning System Development,"IOP Conference Series: Materials Science and Engineering, vol. 69, Art. no. 012012, 2014.
[60] Texas Instruments, "TX810: 8-Channel, Programmable T/R Switch for Ultrasound," Datasheet SLLS996A, Rev. A, Apr. 2010.
[61] Busono, P. et al. (2015). Development of a Modular FPGA Based Digital Beamformer for PC Based Ultrasound Imaging System. In: Goh, J., Lim, C. (eds) 7th WACBE World Congress on Bioengineering 2015. IFMBE Proceedings, vol 52. Springer.
[62] Monolithic Power Systems, "MP4816A: 16-Channel High-Voltage Analog Switch with Output Bleed Resistors," Datasheet, Rev. 1.02, 2020.
[63] G. Enwia, V. Pashaei, M. Bayat, A. Roman and S. Mandal, "An Open-Source Ultrasound Imaging System with Wearable Active Probes,"2019 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, USA, 2019, pp. 503-506.
[64] S. Bhattacharjee et al., "ListenToJESD204B: A lightweight open-source JESD204B IP core for FPGA-based ultrasound acquisition systems," in *Proc. 10th Int. Workshop Adv. Sensors Interfaces (IWASI)*, IEEE, 2025, pp. 1–5.
[65] W. Ma, W. Meng, Y. Zheng, D. Wang, C. Wang e X. Cao, "Modular Ultrasound Platform Design: Designing a Modular 32-Channel Ultrasound Platform with TX7364 and AFE58JD32 for Research Agility,"in Proceedings of the 2025 2nd International Conference on Smart Healthcare and Wearable Intelligent Devices (SHWID 2025), pp. 68–73, 2026.
[66] Maxim Integrated, "MAX2082: Low-Power, High-Performance Octal Ultrasound Transceiver with Integrated AFE, Pulser, T/R Switch, and CWD Beamformer," Datasheet, Rev. 0, 2014.
[67] Maxim Integrated, "MAX2084: 16-Channel, High-Performance, Low-Power, Fully Integrated Ultrasound Receiver with T/R Switch," Datasheet, 2016.
[68] Maxim Integrated, "MAX2079: Low-Power, High-Performance, Fully Integrated Octal Ultrasound Receiver (Octal LNA, VGA, AAF, ADC, and CWD Beamformer)," Datasheet, Rev. 1, 2012.
[69] Texas Instruments, "AFE5801: 8-Channel Variable-Gain Amplifier (VGA) With Octal High-Speed ADC," Datasheet SLOS581D, 2010.
[70] Texas Instruments, "AFE5805: Fully-Integrated, 8-Channel Analog Front-End for Ultrasound, 0.85 nV/$\sqrt{\text{Hz}}$, 12-Bit, 50-MSPS, 122-mW," Datasheet SBOS421D, 2010.
[71] Texas Instruments, "AFE5807: Fully Integrated, 8-Channel Ultrasound Analog Front End, Passive CW," Datasheet SLOS530C, May 2013.
[72] Texas Instruments, "AFE5808: Fully Integrated, 8-Channel Ultrasound Analog Front End With Passive CW Mixer," Datasheet SLOS688D, Rev. D, 2014.
[73] Texas Instruments, "AFE5808A: Fully-Integrated, 8-Channel Ultrasound Analog Front-End With Passive CW Mixer," Datasheet SLOS688D, Rev. D, 2015.
[74] Y. Wang, M. Koen and D. Ma, "Low-Noise CMOS TGC Amplifier With Adaptive Gain Control for Ultrasound Imaging Receivers,"in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 58, no. 1, pp. 26-30, Jan. 2011.
[75] Analog Devices, "AD9271: Octal LNA/VGA/AAF/ADC and Crosspoint Switch," Datasheet Rev. B, 2009.
[76] C. C. P. Cheung et al., "Multi-channel pre-beamformed data acquisition system for research on advanced ultrasound imaging methods,"in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 59, no. 2, pp. 243-253, February 2012.
[77] Analog Devices, "AD9272: Octal LNA/VGA/AAF/ADC and Crosspoint Switch," Datasheet Rev. C, 2009.
[78] Texas Instruments, "AFE58JD16: 16-Channel Ultrasound AFE with 90-mW/Channel Power, 1-nV/$\sqrt{\text{Hz}}$ Noise, 14-Bit, 65-MSPS or 12-Bit, 80-MSPS ADC and Passive CW Mixer," Datasheet AFE58JD16, Rev. A, 2016.
[79] Texas Instruments, "AFE58JD28: 16-Channel Ultrasound AFE with 102-mW/Channel Power, 0.8-nV/$\sqrt{\text{Hz}}$ Noise, 14-Bit, 65-MSPS or 12-Bit, 80-MSPS ADC, Digital Demodulator, JESD or LVDS Interface, and Passive CW Mixer," Datasheet SBAS864, Jul. 2017.
[80] Texas Instruments, "AFE58JD48: 16-Channel Ultrasound AFE with 140-mW/Channel Power, 0.8-nV/$\sqrt{\text{Hz}}$ Noise, 16-Bit, 125-MSPS ADC with JESD or LVDS Interface, Digital Demodulator, and Passive CW Mixer," Datasheet SBASA17A, Dec. 2019.
[81] Texas Instruments, "AFE5832: 32-Channel Ultrasound AFE With 35-mW/Channel Power, 2.1-nV/$\sqrt{\text{Hz}}$ Noise, 12-Bit, 40-MSPS or 10-Bit, 50-MSPS Output, Passive CW Mixer and LVDS Interface," Datasheet SBAS773A, 2018.
[82] Texas Instruments, "AFE58JD32: 32-Channel Ultrasound AFE With 35-mW/Channel Power, 2.1-nV/$\sqrt{\text{Hz}}$ Noise, 12-Bit, 40-MSPS or 10-Bit, 50-MSPS Output, Passive CW Mixer, LVDS and JESD204B Interface, and Digital Demodulator," Datasheet SBAS880A, Apr. 2018.
[83] L. C. Neves, F. M. Ribas, J. M. Maia, A. J. Zimbico, A. A. Assef e E. T. Costa, "Improving Ultrasound B-Mode Image Quality with Coherent Plane-Wave Compounding Using Adaptive Beamformers Based on Minimum Variance,"Sensors, vol. 25, no. 5, Art. no. 1306, 2025.
[84] AMD Xilinx, "Zynq UltraScale+ RFSoC Datasheet: Overview," Datasheet DS889, v1.14, Jun. 2023.
[85] E. Boni, F. Lagonigro, V. Meacci, S. Ricci, A. Ramalli and P. Tortoli, "Toward a modular open scanner: design of the front-end unit,"2023 IEEE International Ultrasonics Symposium (IUS), Montreal, QC, Canada, 2023, pp. 1-3.
[86] A. Hsieh, M. B. Baker, J. M. Phalen, J. Mejias-Garcia e R. Canelli, "Handheld Point-of-Care Ultrasound: Safety Considerations for Creating Guidelines,"Journal of Intensive Care Medicine, vol. 37, no. 9, pp. 1146–1151, 2022.
[87] R. Vega, M. Dehghan, A. Nagdev, B. Buchanan, J. Kapur, J. L. Jaremko e D. Zonoobi, "Overcoming Barriers in the Use of Artificial Intelligence in Point-of-Care Ultrasound,"npj Digital Medicine, vol. 8, no. 1, Art. no. 213, 2025.
[88] S. Kim, C. Fischetti, M. Guy, E. Hsu, J. Fox e S. D. Young, "Artificial Intelligence (AI) Applications for Point of Care Ultrasound (POCUS) in Low-Resource Settings: A Scoping Review,"Diagnostics, vol. 14, no. 15, Art. no. 1669, 2024.

Parâmetros técnicos e eficácia clínica da laserterapia na cicatrização de lesões cutâneas em humanos

1st Guilherme A. O. Aguiar
Departamento de Engenharia Elétrica
Universidade de Brasília
Brasília, Brasil
guilhermeac2018@gmail.com
ORCID: 0009-0007-6262-5531

2nd Thamires R. Freitas
Faculdade UnB Gama - Engenharias
Universidade de Brasília
Brasília, Brasil
thmrs.r@gmail.com
ORCID: 0009-0009-0781-6209

3rd Gerardo A. I. Pizo
Faculdade UnB Gama - Engenharias
Universidade de Brasília
Brasília, Brasil
gerardo.idrobo@gmail.com
ORCID: 0000-0002-6165-672X

4th José Maurício Santos Torres da Motta
Universidade de Brasília
Brasília, Brasil
jmmotta@unb.br
ORCID: 0000-0003-3970-4219

***Resumo*—A laserterapia tem sido investigada como recurso adjuvante para favorecer a cicatrização de lesões cutâneas em humanos, sobretudo por seus efeitos de fotobiomodulação, modulação inflamatória, analgesia, angiogênese e estímulo ao reparo tecidual. Este estudo teve como finalidade caracterizar os principais parâmetros técnicos e a eficácia clínica dos lasers utilizados nesse contexto, destacando a necessidade de padronização dos protocolos terapêuticos. Realizou-se uma revisão integrativa da literatura, orientada pela estratégia PICO, nas bases MEDLINE/PubMed, SciELO, BVS/LILACS e Cochrane Library, contemplando publicações até 2025. Foram incluídos artigos originais completos, em português, inglês ou espanhol, que avaliaram o uso de laser no tratamento de lesões cutâneas em humanos. O nível de evidência e a qualidade metodológica foram avaliados segundo critérios do Joanna Briggs Institute. Após a remoção de duplicidades, triagem por título e resumo e leitura integral, 33 estudos compuseram a amostra final. Desses, 22 estudos (66,7%) foram classificados como de alta eficácia, 5 (15,2%) como de eficácia média e 6 (18,2%) como de baixa eficácia. Os desfechos mais consistentes foram observados em úlceras diabéticas, deiscências cirúrgicas e feridas pós-operatórias. Os parâmetros mais associados a respostas positivas envolveram comprimentos de onda no vermelho visível (595–685 nm) e no infravermelho próximo (785–980 nm), com fluências frequentemente entre 1 e 5 J/cm^2; estudos recentes também reforçam a relevância de doses de 8 a 12 J/cm^2 em úlceras diabéticas. Conclui-se que a laserterapia apresenta potencial como recurso adjuvante na cicatrização de lesões cutâneas; entretanto, a heterogeneidade metodológica, a omissão de dados dosimétricos e a utilização de diferentes classes de laser ainda limitam a comparação direta entre estudos e a definição de protocolos clínicos universais.**

***Palavras-chave*—Laserterapia; fotobiomodulação; cicatrização de feridas; lesões cutâneas; parâmetros dosimétricos.**

I. Introdução

A cicatrização de feridas é um processo biológico complexo e essencial para a restauração da integridade tecidual. Esse fenômeno envolve eventos celulares, vasculares e moleculares organizados em fases parcialmente sobrepostas: inflamação, proliferação e remodelação. A compreensão dessas etapas é fundamental para o desenvolvimento de intervenções capazes de reduzir o tempo de reparo, prevenir complicações e melhorar a qualidade de vida de pacientes com lesões agudas ou crônicas. [1]–[4]

As feridas crônicas representam um problema relevante de saúde pública. Além do impacto funcional, estético e psicossocial, essas lesões estão associadas a elevada demanda assistencial, maior risco de infecção, internações prolongadas e aumento dos custos dos sistemas de saúde. Estimativas internacionais indicam que o tratamento de feridas crônicas gera gastos bilionários, evidenciando a necessidade de terapias adjuvantes seguras, acessíveis e clinicamente efetivas. [46], [48]

Nesse contexto, a laserterapia tem sido investigada como alternativa complementar ao tratamento convencional de lesões cutâneas. O termo laser deriva de Light Amplification by Stimulated Emission of Radiation e descreve uma fonte de luz coerente, monocromática e colimada, capaz de interagir com tecidos biológicos de forma controlada. Em baixas intensidades, essa interação não tem finalidade ablativa ou térmica, mas fotobiológica, sendo frequentemente descrita como terapia a laser de baixa intensidade ou fotobiomodulação. [5]–[12]

A fotobiomodulação está relacionada à absorção da luz por cromóforos celulares, especialmente pela citocromo c oxidase mitocondrial. Essa interação pode favorecer a produção de ATP, a liberação de óxido nítrico, a modulação de espécies reativas de oxigênio em níveis sinalizadores e a regulação de mediadores inflamatórios. Como consequência, pode ocorrer maior atividade de fibroblastos, síntese de colágeno, expressão de fatores pró-angiogênicos, como VEGF, formação de tecido de granulação e epitelização. Esses mecanismos ajudam a explicar a redução de dor, a melhora da microcirculação e a reorganização tecidual observadas em parte dos estudos clínicos. [7]–[12]

Entretanto, a resposta terapêutica depende diretamente de parâmetros físicos e clínicos, como comprimento de onda, potência, fluência, energia por ponto, área irradiada, modo

de emissão, frequência de aplicação, tempo de exposição, técnica de contato e características da lesão. A ausência de padronização dificulta a comparação entre estudos e pode explicar resultados divergentes, especialmente quando diferentes tipos de laser são agrupados em uma mesma análise.

Diante desse cenário, o objetivo deste estudo foi realizar uma revisão integrativa da literatura para caracterizar os parâmetros técnicos e a eficácia dos lasers utilizados no tratamento de lesões cutâneas em humanos, destacando os padrões mais associados a desfechos favoráveis e as limitações que ainda impedem a definição de protocolos clínicos universais.

II. Metodologia

O presente estudo foi desenvolvido como uma revisão integrativa da literatura, método adequado para reunir e discutir evidências provenientes de diferentes delineamentos metodológicos. A escolha dessa abordagem foi justificada pela grande heterogeneidade dos protocolos de laserterapia descritos nos estudos primários, o que inviabiliza uma metanálise quantitativa rigorosa sem padronização prévia das variáveis dosimétricas.

A revisão seguiu as etapas propostas para revisões integrativas: identificação do tema e da pergunta de pesquisa; busca sistematizada na literatura; definição dos critérios de seleção; categorização dos estudos; avaliação crítica das evidências; interpretação dos resultados; e síntese do conhecimento produzido. [13]

A pergunta norteadora foi estruturada pela estratégia PICO: pacientes humanos com lesões cutâneas; intervenção com laserterapia ou fotobiomodulação; comparação com tratamento convencional, placebo, outro comprimento de onda ou ausência de laser, quando disponível; e desfechos relacionados à cicatrização, como redução da área da ferida, tempo de fechamento, melhora do tecido de granulação, epitelização, dor, inflamação e complicações locais.

As buscas foram realizadas nas bases MEDLINE/PubMed, SciELO, BVS/LILACS e Cochrane Library, contemplando publicações disponíveis até 2025. A estratégia combinou descritores e palavras-chave em português, inglês e espanhol, usando principalmente os blocos: (“low-level laser therapy” OR “laser therapy” OR “photobiomodulation” OR “laserterapia”) AND (“wound healing” OR “skin wound” OR “diabetic foot ulcer” OR “pressure ulcer” OR “venous ulcer” OR “cicatrização de feridas”). A seleção manteve os mesmos critérios de elegibilidade para todos os anos avaliados, evitando distinção metodológica entre períodos de publicação.

Foram incluídos artigos originais completos, publicados até 2025, nos idiomas português, inglês ou espanhol, que avaliaram o uso de laser no tratamento de lesões cutâneas em humanos. Foram excluídos estudos exclusivamente in vitro ou em animais, revisões de literatura, protocolos sem resultados clínicos, estudos cujo desfecho principal não fosse cicatrização/reparação tecidual e publicações sem informações mínimas sobre a intervenção com laser.

Os registros encontrados foram organizados no EndNote para remoção de duplicatas. A seleção ocorreu em duas etapas: leitura de títulos e resumos, seguida de leitura integral dos textos potencialmente elegíveis. Dois pesquisadores realizaram a seleção e a extração de dados de forma independente, com resolução de divergências por consenso e discussão com um revisor adicional. [14]

Foram extraídas as seguintes informações: título, autores, país, ano, delineamento, nível de evidência, tipo de lesão, local de aplicação, fonte de luz, comprimento de onda, modo de emissão, potência, fluência, protocolo de aplicação, resultados clínicos, presença de grupo controle e classificação de eficácia. O processo de seleção foi documentado por fluxograma PRISMA, conforme apresentado na Fig. 1 [14]. A avaliação metodológica utilizou critérios do Joanna Briggs Institute, considerando o delineamento de cada estudo; os estudos não foram excluídos apenas por menor qualidade, mas essa condição foi incorporada à interpretação crítica dos achados. [15]

Para reduzir subjetividade, a eficácia foi classificada em três níveis: alta eficácia, quando houve melhora clínica robusta ou diferença estatisticamente significativa em relação ao controle; eficácia média, quando os resultados foram parciais, dependentes de um parâmetro específico ou sustentados por delineamentos com menor força de evidência; e baixa eficácia, quando não houve benefício significativo ou quando o laser não superou o tratamento convencional ou placebo.

III. Resultados e Discussão

Foram identificados 750 registros nas bases consultadas: 25 na SciELO, 387 no PubMed/MEDLINE, 128 na BVS/LILACS e 210 na Cochrane Library. Após a remoção de 38 duplicatas, 712 estudos foram avaliados por título e resumo; 574 foram excluídos por não responderem à pergunta de pesquisa ou por não atenderem aos critérios de inclusão. Em seguida, 138 textos completos foram avaliados, dos quais 105 foram excluídos após leitura integral, resultando em 33 artigos na amostra final.

O processo de seleção é apresentado na Fig. 1, com identificação, triagem, elegibilidade e inclusão final dos estudos.

A amostra incluiu estudos publicados entre 1994 e 2025, com predomínio de ensaios clínicos randomizados ou controlados. De acordo com a classificação do JBI, a maior parte dos estudos apresentou delineamento clínico controlado, embora relatos e séries de casos ainda componham parcela relevante da amostra. Essa distribuição demonstra que existe base clínica consistente, mas parte das evidências permanece limitada por tamanho amostral, controle incompleto, protocolos associados e heterogeneidade dosimétrica.

Em relação à origem das publicações, observou-se participação expressiva de estudos brasileiros, correspondendo a 40% da amostra. Essa presença é relevante por indicar interesse nacional na aplicação clínica da laserterapia, especialmente em feridas crônicas, úlceras por pressão, úlceras diabéticas e deiscências cirúrgicas. A literatura internacional complementou a amostra com estudos da Polônia, Holanda, Egito, Estados Unidos, Canadá, China, Índia, Coreia do Sul, Suíça, França, Indonésia, Irã, Colômbia e Malásia. A distribuição temporal

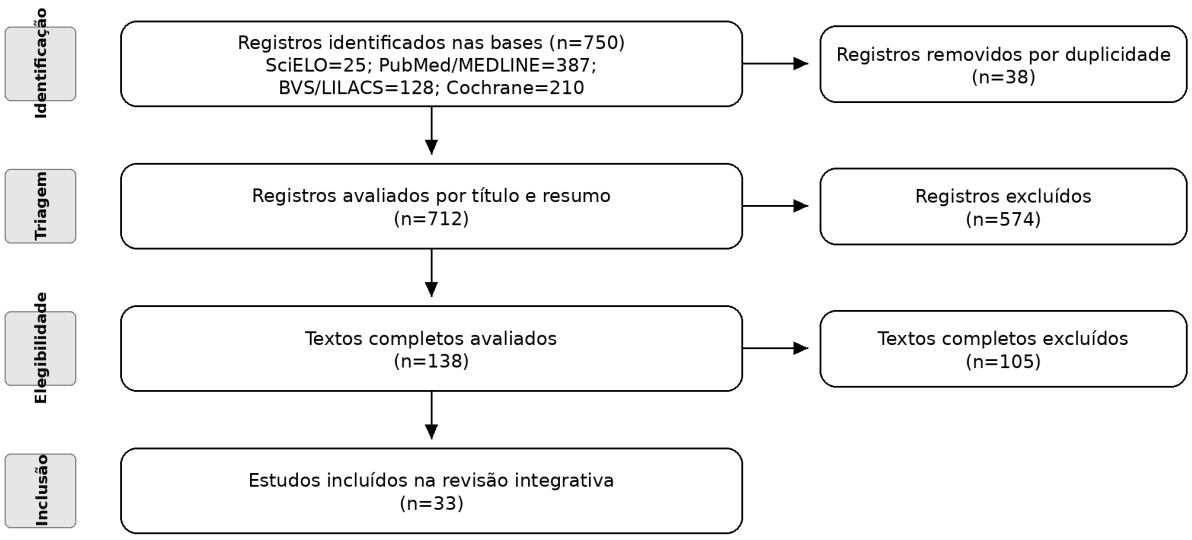


Fig. 1. Fluxograma PRISMA da seleção dos estudos.

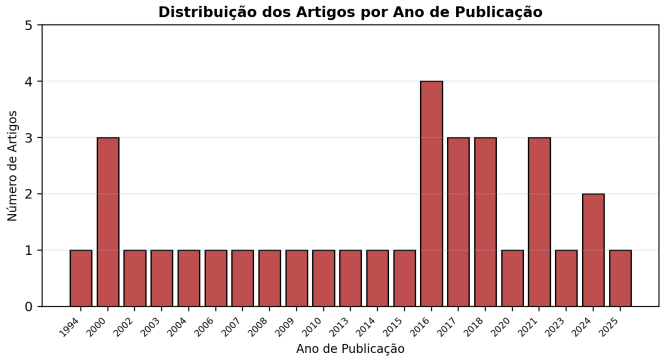


Fig. 2. Distribuição temporal dos artigos selecionados sobre laserterapia em lesões cutâneas.

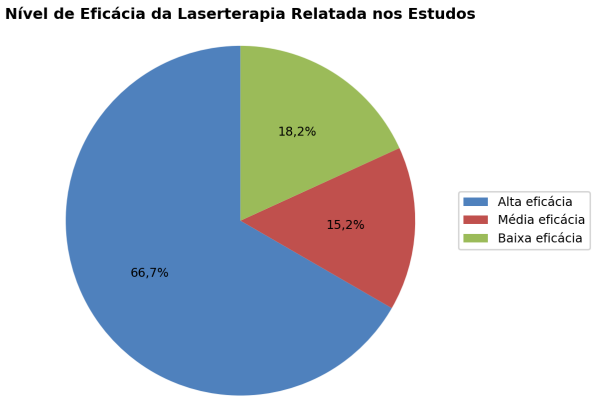


Fig. 3. Proporção do nível de eficácia clínica da laserterapia relatada nos estudos.

dos artigos incluídos é apresentada na Fig. 2, evidenciando continuidade da produção científica sobre laserterapia em lesões cutâneas entre 1994 e 2025.

A. Eficácia clínica relatada nos estudos

A análise por desfecho clínico indicou que 22 estudos (66,7%) relataram alta eficácia da laserterapia, 5 estudos (15,2%) apresentaram eficácia média e 6 estudos (18,2%) mostraram baixa eficácia, conforme sintetizado na Fig. 3. A predominância de resultados positivos sugere que o laser pode atuar como recurso adjuvante útil, mas a interpretação deve considerar o nível de evidência, o tamanho da amostra, a presença de grupo controle e a qualidade do relato dosimétrico.

Os resultados mais consistentes foram observados em úlceras diabéticas, deiscências cirúrgicas, feridas pós-operatórias e algumas feridas traumáticas. Nesses grupos, a laserterapia esteve associada à redução da área da lesão, aceleração da granulação, epitelização, diminuição de dor, menor inflamação e melhor qualidade do tecido cicatricial. Em contrapartida, os estudos com úlceras venosas e úlceras por pressão apresentaram maior heterogeneidade, incluindo resultados sem diferença significativa em relação ao controle.

Esse comportamento indica que a eficácia do laser não deve ser interpretada como propriedade universal da tecnologia, mas como resultado da interação entre o tipo de lesão, fase de cicatrização, profundidade tecidual, parâmetros físicos e frequência de aplicação. Assim, protocolos bem-sucedidos em lesões superficiais podem não gerar a mesma resposta em lesões profundas, crônicas, infectadas ou submetidas a pressão mecânica persistente.

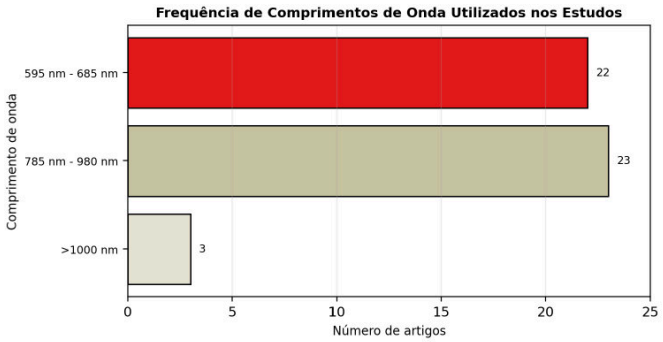


Fig. 4. Frequência das faixas de comprimentos de onda utilizadas nas intervenções com laser.

B. Comprimento de onda e profundidade de ação

O comprimento de onda foi um dos parâmetros mais relevantes para explicar as diferenças de resposta. A faixa vermelha visível, entre 595 e 685 nm, apareceu em 22 estudos e foi frequentemente associada a lesões superficiais, epitelização e modulação inflamatória local. Já o infravermelho próximo, entre 785 e 980 nm, apareceu em 23 estudos e foi utilizado especialmente em lesões mais profundas ou protocolos que buscavam maior penetração tecidual.

Apenas três estudos empregaram comprimentos de onda superiores a 1000 nm, geralmente associados a lasers de alta intensidade ou equipamentos dermatológicos específicos. Nesses casos, o mecanismo de ação pode diferir da fotobiomodulação clássica, razão pela qual a interpretação dos resultados exige cautela. Agrupar LLLT, PBM, HILT, PDL, Nd:YAG e lasers fracionados como se fossem uma única intervenção reduz a precisão científica da análise.

Os estudos comparativos indicaram que o comprimento de onda de 658 nm apresentou melhor desempenho em determinadas úlceras por pressão quando comparado aos comprimentos de 808 nm e 940 nm. Esse achado reforça que a seleção do comprimento de onda deve considerar a profundidade da lesão e os cromóforos-alvo, não apenas a disponibilidade do equipamento.

A contagem das faixas espectrais é apresentada na Fig. 4. Um mesmo estudo pôde ser contabilizado em mais de uma faixa quando utilizou múltiplos comprimentos de onda.

C. Fluência, dose e relato dosimétrico

Além do comprimento de onda, a fluência foi decisiva para a interpretação dos resultados. A maior parte dos protocolos com resposta favorável utilizou fluências entre 1 e 5 J/cm^2, embora alguns estudos tenham empregado valores menores ou muito superiores. Essa amplitude reforça a existência de uma janela terapêutica, na qual doses insuficientes podem não produzir efeito biológico relevante e doses excessivas podem gerar resposta inibitória ou efeito distinto do esperado.

Outro problema recorrente foi a omissão de informações dosimétricas completas. Em diferentes estudos, faltaram dados sobre área do spot, energia por ponto, irradiância, distância de aplicação, tempo total de exposição, número de pontos, intervalo entre sessões ou modo de emissão. Sem esses elementos, torna-se difícil reproduzir o protocolo e determinar se uma falha terapêutica ocorreu por inadequação da dose, refratariedade da lesão ou limitação metodológica do estudo.

Para fortalecer a área, futuros ensaios devem relatar, no mínimo: comprimento de onda, potência média e/ou de pico, energia por ponto, fluência, irradiância, área irradiada, modo de emissão, frequência de aplicação, número de sessões, técnica de contato ou varredura, distância do aplicador, características da ferida e terapia convencional associada.

D. Classe do laser e comparabilidade dos protocolos

Uma melhoria conceitual importante é separar os estudos por classe de laser. Os protocolos de baixa intensidade e fotobiomodulação têm como objetivo modular processos celulares sem dano térmico. Já os lasers de alta intensidade, como HILT, Nd:YAG ou sistemas Classe IV, podem combinar efeitos fotobiológicos, térmicos controlados e maior profundidade de penetração. Lasers dermatológicos, como PDL e laser fracionado não ablativo, possuem aplicação mais específica em cicatrizes e remodelação estética.

Essa separação é essencial porque protocolos com 660 nm e 4 J/cm^2 não são diretamente comparáveis a protocolos com 1064 nm, 25 J/cm^2 ou 80 J/cm^2. A ausência dessa distinção pode produzir conclusões excessivamente amplas. Portanto, a laserterapia deve ser discutida como um conjunto de modalidades ópticas, e não como uma intervenção única e homogênea.

Como síntese prática, sem propor protocolo universal, os achados indicam que lesões superficiais e fases de epitelização tendem a se beneficiar de luz vermelha, especialmente entre 660 e 685 nm, com fluências moderadas; úlceras diabéticas e lesões mais profundas aparecem com maior frequência em protocolos no infravermelho próximo, entre 785 e 980 nm; e feridas crônicas complexas exigem associação obrigatória com curativo adequado, controle de infecção, alívio de pressão, controle glicêmico, orientação do paciente e acompanhamento clínico. Assim, a laserterapia deve ser entendida como recurso adjuvante individualizado, e não como substituta do tratamento padrão.

E. Síntese dos estudos incluídos

A Tabela I apresenta a síntese dos 33 estudos incluídos, preservando autores, país, ano, nível de evidência, tipo de lesão, parâmetros técnicos e principais resultados clínicos. [16]–[45], [49]–[51]

IV. Limitações

Esta revisão apresenta limitações que devem ser consideradas. A primeira é a heterogeneidade dos estudos incluídos, que variaram quanto ao delineamento, tipo de lesão, parâmetros do laser, número de sessões e terapias associadas. A segunda é a presença de relatos e séries de casos, que são úteis para descrição clínica, mas possuem menor capacidade de generalização. A terceira é a ausência de relato dosimétrico

completo em parte da literatura, o que limita a replicabilidade dos achados.

Outra limitação é a impossibilidade de realizar metanálise quantitativa robusta, uma vez que os estudos não apresentaram padronização suficiente de parâmetros físicos e desfechos clínicos. Além disso, mesmo contemplando publicações até 2025, permanece o risco de perda de estudos publicados em bases não pesquisadas ou em literatura cinzenta. Por isso, os resultados devem ser interpretados como síntese integrativa e crítica da literatura disponível, e não como definição definitiva de protocolo terapêutico.

V. Conclusão

A laserterapia apresenta potencial como recurso adjuvante na cicatrização de lesões cutâneas em humanos, especialmente em úlceras diabéticas, deiscências cirúrgicas e feridas pós-operatórias. A maioria dos estudos analisados relatou melhora clínica, com predominância de protocolos que utilizaram comprimentos de onda no vermelho visível e no infravermelho próximo, frequentemente associados a fluências entre 1 e 5 J/cm^2.

Apesar dos resultados favoráveis, a evidência ainda é limitada pela heterogeneidade metodológica, pela inclusão de diferentes classes de laser e pela omissão de dados dosimétricos essenciais. Assim, não é adequado propor um protocolo universal a partir dos estudos disponíveis. Como síntese orientativa, os dados sustentam maior atenção aos protocolos de 660–685 nm para lesões superficiais e epitelização, e de 785–980 nm para úlceras diabéticas e lesões mais profundas, sempre como adjuvantes ao cuidado padrão. O uso clínico deve considerar tipo de lesão, profundidade, fase de cicatrização, características do paciente e parâmetros físicos do equipamento.

Recomenda-se que pesquisas futuras adotem ensaios clínicos randomizados bem controlados, com amostras maiores, descrição completa dos parâmetros ópticos e estratificação por tipo de lesão e tipo de laser. A padronização do relato técnico é condição necessária para transformar resultados promissores em protocolos seguros, reprodutíveis e baseados em evidências.

TABLE I: Apresentação dos artigos selecionados na revisão integrativa.

Artigo	Autores [ref.], país e ano	Nível	Lesão/local	Fonte de luz e parâmetros	Resultado principal
A1	Cárdenas Restrepo et al. [16], Colômbia, 1994	1.d	Queimaduras de 2^{o} e 3^{o} grau	GaAs 902 nm; modo pulsado; 292, 584 e 9344 Hz; 0,1-5 mW; 3 min/dia.	Redução clínica do tempo de hospitalização, granulação e epitelização, sem significância estatística robusta. Eficácia: baixa.
A2	Gonçalves et al. [17], Brasil, 2000	1.d	Úlceras hansênicas e não hansênicas	HeNe 632,8 nm; fluência 4 J/cm^2; potência descrita no original como 4,8 W.	Cicatrização total em 66,7% das úlceras hansênicas e 53,3% das não hansênicas; redução de dor, edema e exsudato. Eficácia: alta.
A3	Lucas et al. [18], Holanda, 2000	1.c	Úlceras de pressão estágio III	Cluster GaAl 904 nm; pulsado; 830 Hz; potência média 8 mW; 1 J/cm^2.	Sem diferença significativa em relação ao controle; o grupo controle apresentou maior redução mediana. Eficácia: baixa.
A4	Simunovic et al. [19], Suíça, 2000	1.c	Feridas cirúrgicas traumáticas em tecidos moles	GaAlAs 830 nm contínuo, 120 mW; HeNe 632,8 nm 10 mW + IR 904 nm pulsado.	Aceleração da cicatrização entre 25% e 35%, com alívio de dor e recuperação funcional precoce. Eficácia: alta.
A5	Franek et al. [20], Polônia, 2002	1.d	Úlcera crural venosa	GaAlAs 810 nm; contínuo; 65 mW; 4 J/cm^2.	Redução em todos os grupos, sem superioridade estatística do laser sobre placebo ou tratamento convencional. Eficácia: baixa.
A6	Lucas et al. [21], Holanda, 2003	1.c	Úlceras de pressão estágio III	GaAs 904 nm; pulsado 830 Hz; 12 diodos; 1 J/cm^2.	Não houve diferença significativa na redução absoluta ou relativa da ferida. Eficácia: baixa.
A7	Hopkins et al. [22], Estados Unidos, 2004	1.c	Abrasões superficiais padronizadas	Cluster multicomprimento: 820, 660, 950, 870, 880 e 940 nm; 700 Hz; 8 J/cm^2.	Contração significativamente maior nos dias 6, 8 e 10, inclusive com possível efeito sistêmico indireto. Eficácia: alta.
A8	Fernandes et al. [23], Brasil, 2007	4.d	Úlcera varicosa em paciente diabético	HeNe 632,8 nm; modo pontual; 4 min; 6 J/cm^2.	Redução aproximada de 50% da área em 5 dias, com melhora de edema e inflamação. Eficácia: média, por tratar-se de relato/caso limitado.
A9	Saltmarche [24], Canadá, 2008	4.c	Úlceras de pressão, diabéticas, venosas e feridas agudas/crônicas	MedX 1000; 785 nm; sonda 50 mW e clusters 80 mW; doses de 1-6 J conforme região.	Fechamento $\geq$50% em 61,9% das feridas e cicatrização completa em 42,9%. Eficácia: média.
A10	Stefanello e Hamerski [25], Brasil, 2006	4.d	Úlcera de pressão grau II	AsGa 904 nm; 45 mW; 6 J/cm^2; aplicação pontual nas margens e varredura no leito.	Cicatrização completa após 14 aplicações em 7 semanas. Eficácia: alta.
A11	Pinto et al. [26], Brasil, 2009	4.d	Deiscência aguda de safenectomia	Diodo 685 nm; 20 mW; 4,5 J/cm^2.	Cicatrização total na 9^{a} sessão e importante efeito analgésico desde as primeiras sessões. Eficácia: alta.
A12	Barreto e Salgado [27], Brasil, 2010	1.c	Úlceras em pacientes com sequelas de hanseníase	InGaIP 660 nm; contínuo; 40 mW; 4 J/ponto nas bordas e 2 J/cm^2 no leito.	Sem diferença significativa em área, profundidade ou escores de cicatrização. Eficácia: baixa.
A13	Taradaj et al. [28], Polônia, 2013	1.c	Úlceras de pressão estágios II e III	GaAlAs 940, 808 e 658 nm; contínuo; 50 mW; 4 J/cm^2.	O comprimento de onda de 658 nm foi superior; 808 e 940 nm não diferiram do placebo. Eficácia: média, dependente do comprimento de onda.
A14	Ha et al. [29], Coreia do Sul, 2014	4.c	Cicatrizes recentes de tireoidectomia	PDL 595 nm; NAFL 1550 nm; parâmetros específicos de pulso, fluência e densidade.	Melhora estética significativa com ambos os lasers, sem diferença entre as técnicas. Eficácia: alta.
A15	Palagi et al. [30], Brasil, 2015	4.d	Úlcera de pressão estágio III	AlGaInP 660 nm; emissão contínua/pulsada; 4 J/cm^2.	Redução de 7 x 6 cm para 1,5 x 1,1 cm em 5 semanas, com melhora do tecido de granulação. Eficácia: alta.
A16	Carvalho et al. [31], Brasil, 2016	1.c	Úlceras diabéticas	Laser HTM 658 nm; contínuo; 30 mW; 80 s/ponto; 4 J/cm^2.	Redução significativa da dor e da área da úlcera em comparação ao controle. Eficácia: alta.
A17	Mathur et al. [32], Índia, 2016	1.c	Úlceras diabéticas	Diodo 660 $\pm$ 20 nm; ~50 mW/cm^2; 60 s/ponto; 3 J/cm^2.	Redução média de 37% no grupo laser contra 15% no controle. Eficácia: alta.
A18	Ojea et al. [33], Brasil, 2016	1.c	Ferida cirúrgica após cirurgia bariátrica aberta	Diodo 808 nm; contínuo; 100 mW; 10 J/cm^2; 2 J/ponto.	Redução de seroma, dor, inflamação e melhor coaptação das bordas. Eficácia: alta.
A19	Piedade [34], Brasil, 2016	4.c	Úlceras crônicas diabéticas e de pressão	AsGa 904 nm; protocolos de 0,28 J/cm^2 e 4 J/cm^2 associados à drenagem linfática.	Ambos estimularam reparo; protocolo de 4 J/cm^2 foi substancialmente superior. Eficácia: média.
A20	Fialho et al. [35], Brasil, 2017	4.c	Úlceras por pressão estágios III e IV	HeNe 632,8 nm e AsGa 904 nm; 5 J/cm^2; 2 vezes/semana.	Dois pacientes cicatrizaram completamente; adesão e autocuidado influenciaram o resultado. Eficácia: alta.
A21	Srilestari et al. [36], Indonésia, 2017	1.c	Úlceras de pé diabético graus 1 e 2	Diodo 630 nm; 100 mW; laser-acupuntura e varredura sobre a úlcera.	Maior redução da área da úlcera em relação ao controle na 4^{a} semana. Eficácia: alta.
A22	Vitse et al. [37], França, 2017	1.c	Úlceras venosas crônicas	MLS 635 nm; 17,5 mW/diodo; 2,95 J/cm^2; 20 min; 2 vezes/semana.	Sem diferença na taxa de fechamento; redução tardia de dor. Eficácia: baixa.
A23	Kazemikhoo et al. [38], Irã, 2018	1.c	Deiscência de enxerto de pele pós-queimadura	650 nm contínuo, 150 mW, 2 J/cm^2; 808 nm contínuo, 4 J/cm^2.	Redução significativa da deiscência e melhora do prognóstico do enxerto. Eficácia: alta.
A24	Taradaj et al. [39], Polônia, 2018	1.c	Úlcera de pressão estágios II a IV	GaAlAs 940, 808 e 658 nm; contínuo; 4 J/cm^2; 5 vezes/semana.	Apenas 658 nm reduziu marcadores inflamatórios e estimulou angiogênese. Eficácia: média.
A25	Thabet et al. [40], Egito, 2018	1.c	Retardo de cicatrização em incisão de cesárea	Nd:YAG 1064 nm; pulsado; 3 kW de pico; 25 J/cm^2; 3 vezes/semana.	Redução altamente significativa do tamanho da ferida e melhora da aparência. Eficácia: alta.
A26	Pereira et al. [41], Brasil, 2020	4.c	Úlcera isquêmica iatrogênica pós-prótese total	Laser de alta potência 980 nm; contínuo; 10 W; 3,18 J/cm^2; feixe desfocado.	Granulação e redução considerável em lesão refratária à oxigenoterapia hiperbárica. Eficácia: alta.
A27	Breder et al. [42], Brasil, 2021	4.d	Deiscência de incisão de cesárea	Diodo 632,8 nm; 25 mW; 4 J/cm^2; associado a gel de barbatimão.	Cicatrização completa após 84 dias e cicatriz quase imperceptível. Eficácia: alta.
A28	Lu et al. [43], China, 2021	1.c	Feridas crônicas refratárias	Laser de alta intensidade BTL-6000; 1064 nm; contínuo; 8 W; 80 J/cm^2.	Redução significativa dos escores BWAT e PUSH em comparação ao curativo convencional. Eficácia: alta.
A29	Nair et al. [44], Malásia, 2021	4.d	Úlceras crônicas de diferentes etiologias	K-Laser Classe IV; 660, 800 e 970 nm; contínuo; 30 kJ em 3 min.	Quatro úlceras fecharam totalmente e as demais reduziram 68,2%-99% em 4 semanas. Eficácia: alta.
A30	Esmael et al. [45], Egito, 2023	1.c	Úlceras crônicas de pé diabético	Diodo Classe IV; 980, 915, 810 e 650 nm; até 24 W; até 20 kHz.	Modo sincronizado alcançou 60% de fechamento total contra 26,7% no modo separado e 6,7% no controle. Eficácia: alta.
A31	Torkaman et al. [49], Irã, 2024	1.c	Úlcera de pé diabético	GaAs 904 nm; 2 J/cm^2; 90 W; 3 vezes/semana; associado ao cuidado convencional.	Houve maior redução percentual da área da úlcera e modulação de marcadores angiogênicos relacionados a HIF-1α, VEGF e receptor específico. Eficácia: alta.
A32	Saura Cardoso et al. [50], Brasil, 2024	1.c	Úlceras de pé diabético não infectadas	GaAs 904 nm; densidades de 4, 8 e 10 J/cm^2; ensaio randomizado, duplo-cego e placebo-controlado.	Todas as doses ativas reduziram significativamente a área das úlceras; 10 J/cm^2 apresentou maior taxa de cicatrização completa e melhor resposta no seguimento. Eficácia: alta.
A33	Bezerra Miranda et al. [51], Brasil, 2025	1.c	Úlceras de pé diabético	Laser 660 nm com 4, 8 e 12 J/cm^2; comparação com infravermelho 904 nm a 10 J/cm^2; tratamento convencional associado.	As doses de 8–12 J/cm^2 em 660 nm e 10 J/cm^2 em 904 nm melhoraram significativamente a cicatrização em 10 semanas. Eficácia: alta.

AGRADECIMENTOS

Os autores agradecem à Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) pelo apoio financeiro ao projeto "Desenvolvimento e Construção de um Laser Sintonizável em Aplicações de Cicatrização da Pele" (2022–2024), que contribuiu para a infraestrutura e para o desenvolvimento científico relacionado a esta pesquisa.

REFERENCES

[1] Carrel A. The treatment of wounds: a first article. J Am Med Assoc. 1910;55(25):2148-50. doi: 10.1001/jama.1910.04330250044011.

[2] Clark RAF. Wound repair. In: Kumar V, Abbas AK, Fausto N, editors. Robbins and Cotran Pathologic Basis of Disease. 7th ed. Philadelphia: Saunders; 2005. p. 112.

[3] Campos ACL, Borges-Branco A, Groth AK. Cicatrização de feridas. ABCD Arq Bras Cir Dig. 2007;20(1):51-58. doi: 10.1590/S0102-67202007000100010.

[4] Mandelbaum SH, Di Santis EP, Mandelbaum MHS. Cicatrização: conceitos atuais e recursos auxiliares - Parte I. An Bras Dermatol. 2003;78(4):393-408. doi: 10.1590/S0365-05962003000400027.

[5] Bagnato VS. Laser e suas aplicações em ciência e tecnologia. 1ª ed. São Paulo: Livraria da Física; 2008.

[6] Catorze G. Laser: fundamentos e indicações em dermatologia. Med Cutan Iber Lat Am. 2009;37(1):5-27.

[7] Karu T. Photobiology of low-power laser effects. Health Phys. 1989;56(5):691-704. doi: 10.1097/00004032-198905000-00015.

[8] Ortiz MCS, et al. Laser de baixa intensidade: princípios e generalidades - Parte 1. Fisioterapia Brasil. 2001;2:221-240. doi: 10.33233/fb.v2i4.644.

[9] Henriques ACG, Cazal C, Castro JFL. Ação da laserterapia no processo de proliferação e diferenciação celular: revisão da literatura. Rev Col Bras Cir. 2010;37(4):295-302. doi: 10.1590/S0100-69912010000400011.

[10] Marzullo CF, Peres ACP, Shah ML, Nicolau RA. Atualidades do efeito analgésico após aplicação do laser de baixa potência. X Encontro Latino Americano de Iniciação Científica e VI Encontro Latino Americano de Pós-Graduação; 2006. p. 2520-2523.

[11] Andrade FS, Clark RMO, Ferreira ML. Effects of low-level laser therapy on wound healing. Rev Col Bras Cir. 2014;41(2):129-133. doi: 10.1590/S0100-69912014000200010.

[12] Hawkins D, Abrahamse H. Effect of multiple exposures of low-level laser therapy on the cellular responses of wounded human skin fibroblasts. Photomed Laser Ther. 2007;24(6):705-714. doi: 10.1089/pho.2006.24.705.

[13] Pompeo DA, Rossi LA, Galvão CM. Revisão integrativa: etapa inicial do processo de validação de diagnóstico de enfermagem. Acta Paul Enferm. 2009;22(4):434-438. doi: 10.1590/S0103-21002009000400014.

[14] Galvão TF, Pansani TSA, Harrad D. Principais itens para relatar revisões sistemáticas e meta-análises: a recomendação PRISMA. Epidemiol Serv Saúde. 2015;24(2):335-342. doi: 10.5123/S1679-49742015000200017.

[15] Joanna Briggs Institute. Critical appraisal tools. 2024. Available from: https://jbi.global/critical-appraisal-tools.

[16] Restrepo JCC, Aramburo LFA, Amaya FM, Schall AC. Laser blando en el tratamiento de las quemaduras. Rev Col Cir. 1994;9(3):163-166.

[17] Gonçalves G, Gonçalves A, Padovani CR, Parizotto NA. Promovendo a cicatrização de úlceras hansênicas e não hansênicas com laserterapia: ensaio clínico em unidades ambulatoriais do Sistema Único de Saúde. Hansenol Int. 2000;25(2):133-142. doi: 10.47878/hi.2000.v25.36435.

[18] Lucas C, Coenen C, de Haan R. The effect of low level laser therapy (LLLT) on stage III decubitus ulcers: a prospective randomised single blind, multicentre pilot study. Lasers Sci. 2000;15:94-100. doi: 10.1007/s101030050054.

[19] Simunovic Z, Ivankovich AD, Depolo A. Wound healing of animal and human body sport and traffic accident injuries using low-level laser therapy treatment: a randomized clinical study of seventy-four patients with control group. J Clin Laser Med Surg. 2000;18(2):67-73. doi: 10.1089/clm.2000.18.67.

[20] Franek N, Król P, Kucharzewski M. Does low output laser stimulation enhance the healing of crural ulceration? Some critical remarks. Med Eng Phys. 2002;24(9):607-615. doi: 10.1016/S1350-4533(02)00112-1.

[21] Lucas C, van Gemert MJ, de Haan RJ. Efficacy of low-level laser therapy in the management of stage III decubitus ulcers: a prospective, observer-blinded multicentre randomised clinical trial. Lasers Med Sci. 2003;18(2):72-77. doi: 10.1007/s10103-003-0259-5.

[22] Hopkins JT, McLoda TA, Seegmiller JG, Baxter GD. Low-Level Laser Therapy Facilitates Superficial Wound Healing in Humans: A Triple-Blind, Sham-Controlled Study. J Athl Train. 2004;39(3):223-229.

[23] Fernandes P, Guerino MR, Guerino AC, Ide MR. Efeitos do laser de HeNe na cicatrização de úlceras varicosas em pacientes diabéticos. Fisioter Bras. 2007;8(6):463-470. doi: 10.33233/fb.v8i6.1818.

[24] Saltmarche AE. Low level laser therapy for healing acute and chronic wounds - the Extendicare experience. Int Wound J. 2008;5(2):351-360. doi: 10.1111/j.1742-481X.2008.00491.x.

[25] Stefanello TD, Hamerski CR. Tratamento de úlcera de pressão através do laser AsGa de 904 nm - um relato de caso. Arq Ciênc Saúde Unipar. 2008;10(2).

[26] Pinto NC, Pereira MHC, Stolf NAG, Chavantes MC. Laser de baixa intensidade em deiscência aguda de safenectomia: proposta terapêutica. Braz J Cardiovasc Surg. 2009;24(1):88-91. doi: 10.1590/S0102-76382009000100017.

[27] Barreto JG, Salgado CG. Clinic-epidemiological evaluation of ulcers in patients with leprosy sequelae and the effect of low level laser therapy on wound healing: a randomized clinical trial. BMC Infect Dis. 2010;10:237. doi: 10.1186/1471-2334-10-237.

[28] Taradaj J, Halski T, Kucharzewski M, Urbanek T, Halska U, Kucio C. Effect of laser irradiation at different wavelengths (940, 808, and 658 nm) on pressure ulcer healing. Evid Based Complement Alternat Med. 2013;2013:960240. doi: 10.1155/2013/960240.

[29] Ha JM, Kim HS, Cho EB, Park GH, Park EJ, Kim KH, et al. Comparison of the effectiveness of nonablative fractional laser versus pulsed-dye laser in thyroidectomy scar prevention. Ann Dermatol. 2014;26(5):615-620. doi: 10.5021/ad.2014.26.5.615.

[30] Palagi S, Severo IM, Menegon DB, Lucena AF. Laser therapy in pressure ulcers: evaluation by the Pressure Ulcer Scale for Healing and Nursing Outcomes Classification. Rev Esc Enferm USP. 2015;49(5):826-833. doi: 10.1590/S0080-623420150000500017.

[31] Carvalho AFM, Feitosa MCP, Coelho NPMF, Rebêlo VCN, Castro JG, Sousa PRG, et al. Low-level laser therapy and Calendula officinalis in repairing diabetic foot ulcers. Rev Esc Enferm USP. 2016;50(4):628-634. doi: 10.1590/S0080-623420160000500013.

[32] Mathur RK, Sahu K, Saraf S, Patheja P, Khan F, Gupta PK. Low-level laser therapy as an adjunct to conventional therapy in the treatment of diabetic foot ulcers. Lasers Med Sci. 2017;32(2):275-282. doi: 10.1007/s10103-016-2109-2.

[33] Ojea AR, Madi O, Neto RM, Lima SE, de Carvalho BT, Ojea MJ, et al. Beneficial effects of applying low-level laser therapy to surgical wounds after bariatric surgery. Photomed Laser Surg. 2016;34(11):580-584. doi: 10.1089/pho.2016.4149.

[34] Piedade MCB. Laserterapia associada à drenagem linfática em úlceras de pele: relato de múltiplos casos. Fisioter Bras. 2013;14(6):422-425. doi: 10.33233/fb.v14i6.431.

[35] Fialho LM, Baron MV, Brandenburg C, Martins ABT. Efeitos dos lasers Hélio-Neônio (HeNe) e Arseneto de Gálio (AsGa) associados à educação em saúde. Rev Med Minas Gerais. 2017;27. doi: 10.5935/2238-3182.20170051.

[36] Srilestari A, Nareswari I, Simadibrata C, Tarigan TJ. Effectiveness of combined laser-puncture and conventional wound care to accelerate diabetic foot ulcer healing. Med J Indones. 2017;26(1):26-34. doi: 10.13181/mji.v26i1.1401.

[37] Vitse J, Bekara F, Byun S, Herlin C, Teot L. A double-blind, placebo-controlled randomized evaluation of the effect of low-level laser therapy on venous leg ulcers. Int J Low Extrem Wounds. 2017;16(1):29-35. doi: 10.1177/1534734617690948.

[38] Kazemikhoo N, Vaghardoost R, Dahmardehei M, Mokmeli S, Momeni M, Nilforoushzadeh MA, et al. Evaluation of the effects of low level laser therapy on the healing process after skin graft surgery in burned patients. J Lasers Med Sci. 2018;9(2):139-143. doi: 10.15171/jlms.2018.26.

[39] Taradaj J, Shay B, Dymarek R, Sopel M, Walewicz K, Beeckman D, et al. Effect of laser therapy on expression of angio- and fibrogenic factors, and cytokine concentrations during the healing process of human pressure ulcers. Int J Med Sci. 2018;15(11):1105-1112. doi: 10.7150/ijms.25651.

[40] Thabet AAE, Mahran HG, Ebid AA, Alshehri MA. Effect of pulsed high intensity laser therapy on delayed caesarean section healing in diabetic women. J Phys Ther Sci. 2018;30(4):570-575. doi: 10.1589/jpts.30.570.

[41] Pereira FLC, Ferreira MVL, Mendes PS, Rossi FM, Alves MP, Alves BLP. Use of a High-Power Laser for Wound Healing: A Case Report. J Lasers Med Sci. 2020;11(1):112-114. doi: 10.15171/jlms.2020.19.

[42] Breder JSC, Tsukumo DML, Pereira E, Lima MH. Surgical Wound Dehiscence Treatment With Low-Level Laser Therapy and Barbatimão: A Case Report. Wound Manag Prev. 2021;67(10):18-22.

[43] Lu Q, Yin Z, Shen X, Li J, Su P, Feng M, et al. Clinical effects of high-intensity laser therapy on patients with chronic refractory wounds: a randomised controlled trial. BMJ Open. 2021;11(7):e045866. doi: 10.1136/bmjopen-2020-045866.

[44] Nair HKR, Chong SSY, Selvaraj DDJ. Photobiomodulation as an adjunct therapy in wound healing. Int J Low Extrem Wounds. 2023;22(2):278-282. doi: 10.1177/15347346211004186.

[45] Esmael HF, El Basiouny MS, Morsy MI, Abdel Gawad HA, Hssein AA, Abdallah TE. The effect of combined 650 nm and infrared laser on chronic diabetic foot ulcer surface area: a randomized controlled trial. J Pharm Negative Results. 2023;14(2):199-207. doi: 10.47750/pnr.2023.14.02.25.

[46] Burgess JL, Wyant WA, Abdo Abujamra B, Kirsner RS, Jozic I. Diabetic wound-healing science. Medicina (Kaunas). 2021;57(10):1072. doi: 10.3390/medicina57101072.

[47] Pham H, Armstrong DG, Harvey C, Harkless LB, Giurini JM, Veves A. Screening techniques to identify people at high risk for diabetic foot ulceration: a prospective multicenter trial. Diabetes Care. 2000;23(5):606-611. doi: 10.2337/diacare.23.5.606.

[48] Sen CK, Gordillo GM, Roy S, Kirsner R, Lambert L, Hunt TK, et al. Human skin wounds: a major and snowballing threat to public health and the economy. Wound Repair Regen. 2009;17(6):763-771. doi: 10.1111/j.1524-475X.2009.00543.x.

[49] Torkaman G, Hoseini-Sanati M, Hedayati M, Mofid M, Iranparvar Alamdari M. Effects of photobiomodulation therapy on the expression of hypoxic inducible factor, vascular endothelial growth factor, and its specific receptor: a randomized control trial in patients with diabetic foot ulcer. Photobiomodul Photomed Laser Surg. 2024;42(4):275-284. doi: 10.1089/photob.2023.0152.

[50] Saura Cardoso V, de Souza Lima da Silveira PR, Dos Santos CM, Bezerra Miranda M, Silva Barros AC, Veloso Coelho L, et al. Dose-response and efficacy of 904 nm photobiomodulation on diabetic foot ulcers healing: a randomized controlled trial. Lasers Med Sci. 2024;39(1):142. doi: 10.1007/s10103-024-04090-3.

[51] Bezerra Miranda M, Silva Barros AC, Veloso Coelho L, Barbosa da Rocha R, Antonio Kato da Silva B, Ahmad Hazime F, et al. Dose-response and efficacy of 660-nanometer low-level laser therapy in healing diabetic foot ulcers: a randomized, double-blind, clinical trial. Lasers Med Sci. 2025;40(1):399. doi: 10.1007/s10103-025-04641-2.

Photobiostimulation After Acute Myocardial Infarction: A One-Year Follow-Up Case Report

Maria Verônica Pires
Graduate Program in Biomedical Engineering
Anhembi Morumbi University
Center for Innovation, Technology and Education
São José dos Campos, Brazil
mariaveronicapires46@gmail.com

Carlos José de Lima
Graduate Program in Biomedical Engineering
Anhembi Morumbi University
Center for Innovation, Technology and Education
São José dos Campos, Brazil
cdcfdlima@gmail.com

***Abstract*— A 63-year-old man with acute anteroseptal ST-segment elevation myocardial infarction underwent primary percutaneous coronary intervention followed by adjunctive photobiostimulation (PBST), initiated 18 days after the ischemic event. PBST was delivered to the anterior thoracic region using a wearable 940-nm near-infrared light-emitting diode device. Serial echocardiography and myocardial perfusion SPECT were performed over one year. LVEF assessed by the biplane Simpson method increased from 46% at baseline to 61% at 120 days, with preserved systolic function on subsequent assessments. Global longitudinal strain was 23% at 120 days. Indexed LV end-diastolic volume remained stable (65.5 to 61.0 mL/m²), without progressive ventricular dilatation. Serial SPECT demonstrated a persistent fixed perfusion defect without reversible ischemia and preserved global ventricular function. Despite recognized risk factors for adverse remodeling, sustained functional recovery was observed. Although no causal relationship between PBST and ventricular recovery can be established from this single observation, these findings support further investigation of PBST as a potential adjunctive strategy after myocardial infarction.**

Keywords*— ST-segment elevation myocardial infarction; Photobiomodulation; Ventricular remodeling; Left ventricular function;*

I. INTRODUCTION

Acute myocardial infarction (MI) remains a major cause of cardiovascular morbidity and mortality. Myocardial injury resulting from ischemia may lead to left ventricular dysfunction, adverse ventricular remodeling, and subsequent heart failure [1]. Early reperfusion combined with antithrombotic therapy has significantly reduced both short- and long-term mortality following MI; however, mortality remains substantial, particularly during the acute phase and the first year after the event [2].

The myocardium has a high energy demand required to sustain specialized cellular processes and possesses a limited capacity to store substrates involved in glycolipid metabolism, mitochondrial oxidative phosphorylation, and adenosine triphosphate (ATP) production and utilization [3], [4]. Myocardial ischemia triggers a complex cascade of events, including alterations in energy metabolism, mitochondrial dysfunction, increased oxidative stress, dysregulation of intracellular calcium homeostasis, and activation of inflammatory and neurohumoral pathways involving the renin–angiotensin system [3], [5], [6]. These mechanisms contribute to apoptosis, fibrosis, and structural alterations in the myocardium, resulting in cardiac remodeling and directly affecting left ventricular function and clinical outcomes [7], [8].

Although early myocardial revascularization significantly reduces the extent of ischemia, ischemia-reperfusion (I/R) injury and persistent microvascular dysfunction can limit myocardial functional recovery, favoring adverse ventricular remodeling [3], [5], [9]. Consequently, adjunctive therapeutic strategies targeting inflammatory pathways, cellular bioenergetics, and preservation of myocardial viability have emerged as promising areas of investigation.

Photobiostimulation (PBST) involves the application of non-ionizing light, mainly in the red (λ = 650 nm) and near-infrared (NIR; λ = 700–1000 nm) spectral ranges, inducing photochemical responses that may subsequently lead to photobiomodulation (PBM) of cellular processes.

At the cellular level, PBST induces photochemical reactions that may modulate ATP production, reactive oxygen species (ROS) generation, inflammatory signaling pathways, nitric oxide (NO) release, and intracellular calcium (Ca^{2+}) homeostasis [10]. Experimental studies have suggested a cardioprotective role for PBST in ischemia–reperfusion injury, demonstrating reductions in infarct size and attenuation of pathological ventricular remodeling [6], [11]. These effects are thought to be mediated through improvements in mitochondrial function, cellular bioenergetics, and regulation of oxidative and inflammatory responses. However, clinical evidence supporting the use of PBST in patients following acute myocardial infarction remains limited and inconclusive.

Therefore, the present report aims to describe the longitudinal recovery of left ventricular function in a patient with acute myocardial infarction who received PBST as an adjunctive therapy to conventional treatment.

II. CASE REPORT

A 63-year-old man with a history of hypertension, dyslipidemia, chronic kidney disease, overweight, and prior smoking was admitted with an acute anteroseptal ST-segment elevation myocardial infarction (STEMI), confirmed by electrocardiographic findings and elevated cardiac biomarkers. Coronary angiography revealed multivessel coronary artery disease with thrombotic occlusion of the mid-left anterior descending artery (LAD). Primary percutaneous coronary intervention (PCI) with drug-eluting stent implantation was successfully performed, restoring TIMI grade 3 flow. A staged revascularization procedure was

subsequently completed for additional significant coronary lesions.

Transthoracic echocardiography performed within the first 24 hours demonstrated left ventricular systolic dysfunction, with a left ventricular ejection fraction (LVEF) of 46% by the biplane Simpson method, associated with regional wall motion abnormalities involving the anterior and anteroseptal segments. Right ventricular systolic function was preserved (TAPSE 21 mm).

The patient remained hemodynamically stable throughout hospitalization and was discharged on the eighth day, receiving guideline-directed medical therapy for secondary prevention after myocardial infarction.

A. Photobiostimulation Protocol

Photobiostimulation (PBST) was initiated 18 days after myocardial infarction as an adjunctive therapy to standard medical treatment. The protocol was developed based on experimental and clinical literature on photobiostimulation and near-infrared photobiomodulation [12], [13].

PBST was administered using a wearable vest-like device designed to cover the torso and deliver transcutaneous irradiation to the anterior thoracic region over a treatment area of 2088 cm² with dimensions of 58 × 36 cm. The device incorporated 300 uniformly distributed light-emitting diodes (LEDs) emitting near-infrared light at 940 nm. Treatment was delivered at a total nominal optical power of 6 W, corresponding to a calculated average surface irradiance of approximately 2.9 mW/cm². The radiant exposure (fluence) was 3.5 J/cm², corresponding to a total photonic energy of 7200 J per session. Each session lasted 20 minutes, and the system was powered by a low-voltage (12 V) power supply.

For transcutaneous thoracic irradiation with near-infrared light at 940 nm, a simplified optical propagation model is proposed. In this conceptual model, incident photons undergo multiple scattering events, resulting in three-dimensional redistribution of optical energy that may be approximated, for conceptual purposes, by a diffuse Lambertian model, together with absorption by tissue chromophores and molecular photoacceptors. Absorption of photonic energy by mitochondrial photoacceptors, including cytochrome c oxidase (CCO), may modulate respiratory chain activity, mitochondrial membrane potential, and the proton gradient, thereby favoring ATP synthesis. These initial photochemical events characterize photostimulation and may trigger downstream cellular signaling cascades, subsequently resulting in photobiomodulation (PBM) of cellular and metabolic processes. This model is conceptual and does not permit direct determination of photon distribution or the optical dose effectively delivered to the myocardium.

The treatment protocol initially consisted of daily sessions administered at 24-hour intervals. Beginning in the third month, treatment frequency was gradually reduced, reaching two sessions per week by the end of the fourth month. PBST was discontinued after completion of the four-month treatment protocol.

B. Ethical Aspects

This case report was approved by the Research Ethics Committee under CAAE No. 76100023.1.0000.5492, opinion No. 6.657.464. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and national standards for research involving human subjects.

III. CLINICAL EVOLUTION

Myocardial perfusion single-photon emission computed tomography (SPECT) with technetium-99m sestamibi (^99mTc-sestamibi) and dipyridamole stress was performed 54 days after myocardial infarction (36 days after initiation of PBST). Quantitative perfusion analysis yielded a summed stress score (SSS) of 10 and a summed rest score (SRS) of 10, reflecting perfusion abnormalities during stress and rest, respectively. The resulting summed difference score (SDS = SSS − SRS) was 0, indicating a fixed perfusion defect without a reversible ischemic component. Global left ventricular wall motion was preserved, although systolic thickening was reduced in the affected segments. Stress left ventricular ejection fraction (LVEF) was 53% (Fig. 1).

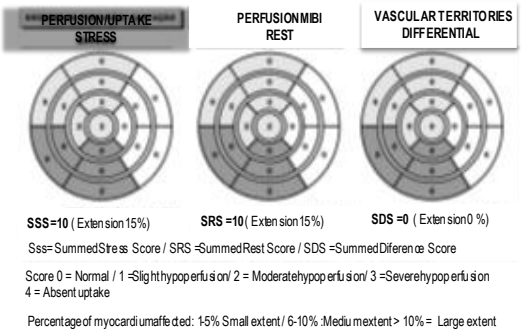


Fig. 1: Myocardial perfusion SPECT polar maps at rest and during pharmacological stress showing a persistent fixed perfusion defect involving approximately 15% of the LV myocardium (SSS=10, SRS=10, SDS=0), without reversible ischemia. Gated SPECT demonstrated preserved global LV systolic function.

Serial echocardiography demonstrated the recovery and subsequent preservation of left ventricular systolic function. LVEF assessed using the biplane Simpson method increased from 46% at baseline to 61% at 120 days. Subsequent assessments using other methods also demonstrated preserved systolic function, although values obtained using different techniques are not directly interchangeable. At 120 days, LV dimensions and regional wall motion were preserved, with a mean global longitudinal strain (GLS) magnitude of 23%. After one year, LV dimensions remained normal (LV end-diastolic diameter [LVEDD], 51 mm; LV end-systolic diameter [LVESD], 34 mm), with LVEF values of 62% by the Teichholz method and 70% by the cube method, and mild residual hypokinesis of the mid and apical anterior and septal segments. Indexed LV end-diastolic volume remained stable, from approximately 65.5 mL/m² at baseline to 61.0 mL/m² at one year, with no evidence of progressive LV dilatation. Baseline indexed LV end-systolic volume was unavailable, precluding serial comparison of this parameter.

At late follow-up, myocardial perfusion SPECT demonstrated persistent fixed perfusion abnormalities, with unchanged quantitative perfusion scores (SSS = 10, SRS = 10, SDS = 0; 15% extent) and no inducible ischemia. Global LV wall motion remained preserved, with reduced systolic thickening in the affected segments, while stress LVEF increased from 53% on the initial SPECT to >70% at follow-up. Thus, serial SPECT demonstrated an improved global functional response without a measurable change in perfusion defect burden.

During follow-up, serial imaging demonstrated recovery and sustained preservation of left ventricular systolic function, without evidence of significant reversible ischemia or progressive ventricular dilatation. These findings occurred in the context of optimized guideline-directed medical therapy with adjunctive PBST. Fig. 2 illustrates the temporal evolution of LVEF during follow-up, with values presented according to imaging modality and measurement method.

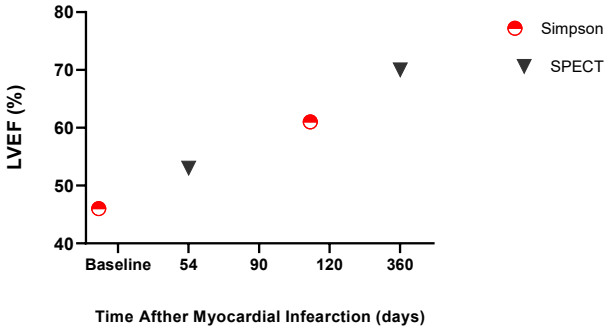


Fig. 2. Temporal evolution of left ventricular ejection fraction (LVEF) after acute myocardial infarction. Echocardiographic LVEF assessed by the biplane Simpson method increased from 46% at baseline to 61% at 120 days. Stress gated-SPECT LVEF was 53% at day 54 and >70% at follow-up.

IV. DISCUSSION

Patients with ST-segment elevation myocardial infarction (STEMI) experience substantial morbidity and mortality, frequently associated with the development of adverse ventricular remodeling. This process results from the activation of inflammatory and neurohumoral pathways triggered by the ischemic event. These pathways lead to alterations in gene expression and in molecular, cellular, and interstitial processes, ultimately resulting in structural and functional changes of the left ventricle. Many of these changes are closely linked to energy metabolism [14].

Post-infarction prognosis is influenced by multiple factors, including infarct size, time to reperfusion, multivessel coronary artery disease, and chronic kidney disease. In the present case, the patient presented a combination of characteristics recognized as major determinants of adverse ventricular remodeling and unfavorable clinical outcomes.

The inflammatory response triggered by myocardial ischemia develops predominantly during the first weeks following infarction, whereas neurohumoral activation may persist for up to six weeks after the acute event [15]. Although early reperfusion of the infarct-related artery is essential for preserving viable myocardium, restoration of epicardial coronary flow does not necessarily guarantee adequate microvascular perfusion. Consequently, microvascular obstruction and ischemia–reperfusion injury may occur despite successful revascularization [5], [9]. Microvascular obstruction may persist for up to three months after infarction, contributing to increased myocardial stiffness, reduced ventricular compliance, and impaired recovery of left ventricular function [16], and has been identified as an independent predictor of persistent ventricular dysfunction [17].

In the present report, photobiostimulation (PBST) was introduced as an adjunctive therapy during a phase characterized by intense inflammatory activity, tissue remodeling, and metabolic adaptation. From a pathophysiological perspective, PBST may modulate cellular bioenergetics, ATP production, oxidative stress, and inflammatory signaling pathways potentially relevant to post-infarction ventricular remodeling. In an experimental model, Feliciano et al. demonstrated that PBST modulated transcriptional and post-transcriptional pathways associated with post-infarction remodeling, including inflammation, apoptosis, myocardial hypertrophy, calcium handling, cellular metabolism, oxidative stress, and extracellular matrix regulation [6]. However, clinical evidence in acute myocardial infarction remains limited. A small clinical study reported lower troponin-T levels following adjunctive low-level laser therapy but no significant between-group difference in LVEF at 9 months [20].

The biological effects of PBST originate from the delivery of optical energy to biological tissues. Following photon absorption by intracellular photoacceptors, photophysical and photochemical events may modulate mitochondrial activity, cellular bioenergetics, redox signaling, and nitric oxide bioavailability. These responses may subsequently lead to photobiomodulation (PBM) of cellular processes and inflammatory pathways potentially relevant to post-infarction ventricular remodeling.

Longitudinal evaluation with serial imaging demonstrated recovery and sustained preservation of left ventricular systolic function. LVEF assessed using the same biplane Simpson method increased from 46% at baseline to 61% at 120 days. Subsequent assessments using other methods also demonstrated preserved systolic function, although values obtained by different techniques are not directly interchangeable. At 120 days, mean global longitudinal strain (GLS) magnitude was 23%, and ventricular dimensions remained preserved. Indexed LV end-diastolic volume remained stable, from approximately 65.5 mL/m² at baseline to 61.0 mL/m² at one year, supporting the absence of progressive ventricular dilatation.

Serial myocardial perfusion SPECT provided complementary functional and perfusion information. Quantitative perfusion abnormalities remained unchanged during follow-up (SSS = 10, SRS = 10, SDS = 0; 15% extent), indicating persistence of a fixed perfusion defect without a reversible ischemic component. Global ventricular function remained preserved, while stress LVEF increased from 53% on the initial study to >70% at late follow-up. These findings demonstrate an increased stress LVEF at late follow-up without a measurable change in fixed perfusion defect burden.

Post-infarction LVEF recovery is heterogeneous. Contemporary cohorts have reported median absolute LVEF improvements of approximately 6%–8%, with substantial interindividual variability [8], [18]. In the PREDICTS study, involving patients with more severe baseline systolic dysfunction, 26% recovered to an LVEF ≥50% at approximately 90 days after myocardial infarction, whereas the majority remained with persistent degrees of ventricular dysfunction [19]. Although these populations are not directly comparable with the present case, they provide clinical context for the magnitude of functional recovery observed.

Nevertheless, these findings should be interpreted in light of the limitations of a single-case observation. Recovery of ventricular function may reflect established mechanisms, including successful reperfusion, myocardial viability, recovery of stunned myocardium, and optimized guideline-directed medical therapy. Therefore, no causal relationship between PBST and ventricular recovery can be established. The sustained preservation of systolic function and stable ventricular dimensions observed during follow-up support further investigation of PBST as a potential adjunctive strategy after acute myocardial infarction.

V. CONCLUSION

This exploratory case report demonstrated sustained recovery of left ventricular function following acute myocardial infarction, with preserved ventricular dimensions and no evidence of progressive ventricular dilatation over one year of follow-up. Although no causal relationship between PBST and ventricular recovery can be established from a single observation, the favorable functional evolution observed supports further investigation of PBST as a potential adjunctive strategy for post-infarction ventricular recovery.

ACKNOWLEDGMENT

The authors would like to thank Instituto Anima (AI), University Anhembi Morumbi, São Paulo – SP, Brazil.

REFERENCES

[1] S. Frantz, M. J. Hundertmark, J. Schulz-Menger, F. M. Bengel, and J. Bauersachs, "Left ventricular remodelling post-myocardial infarction: pathophysiology, imaging, and novel therapies," Eur. Heart J., vol. 43, no. 27, pp. 2549–2561, 2022, doi: 10.1093/eurheartj/ehac223.

[2] J. M. Pedersen et al., "Cause of death in patients with ST-segment elevation myocardial infarction treated with primary percutaneous coronary intervention," J. Am. Coll. Cardiol., vol. 64, no. 20, pp. 2101–2108, 2014, doi: 10.1016/j.jacc.2014.08.037.

[3] H. Tian, X. Zhao, Y. Zhang, and Z. Xia, "Abnormalities of glucose and lipid metabolism in myocardial ischemia-reperfusion injury," Biomed. Pharmacother., vol. 163, Art. no. 114827, 2023, doi: 10.1016/j.biopha.2023.114827.

[4] J. M. Huss and D. P. Kelly, "Mitochondrial energy metabolism in heart failure: a question of balance," J. Clin. Invest., vol. 115, no. 3, pp. 547–555, 2005, doi: 10.1172/JCI24405.

[5] M. Algoet, S. Janssens, U. Himmelreich, et al., "Myocardial ischemia-reperfusion injury and the influence of inflammation," Trends Cardiovasc. Med., vol. 33, no. 6, pp. 357–366, 2023, doi: 10.1016/j.tcm.2022.02.005.

[6] R. S. Feliciano, A. L. B. Atum, É. G. S. Ruiz, et al., "Photobiomodulation therapy on myocardial infarction in rats: transcriptional and posttranscriptional implications to cardiac remodeling," Lasers Surg. Med., vol. 53, no. 9, pp. 1247–1257, 2021, doi: 10.1002/lsm.23407.

[7] J. N. Cohn, R. Ferrari, and N. Sharpe, "Cardiac remodeling—concepts and clinical implications: a consensus paper from an international forum on cardiac remodeling," J. Am. Coll. Cardiol., vol. 35, no. 3, pp. 569–582, 2000, doi: 10.1016/S0735-1097(99)00630-0.

[8] S. B. Wilton, M. T. Bennett, R. Parkash, et al., "Variability in reassessment of left ventricular ejection fraction after myocardial infarction in the Acute Myocardial Infarction Quality Assurance Canada Study," JAMA Netw. Open, vol. 4, no. 12, Art. no. e2136830, 2021, doi: 10.1001/jamanetworkopen.2021.36830.

[9] D. J. Doherty, R. Sykes, K. Mangion, and C. Berry, "Predictors of microvascular reperfusion after myocardial infarction," Curr. Cardiol. Rep., vol. 23, no. 3, Art. no. 21, 2021, doi: 10.1007/s11886-021-01442-1.

[10] S. Ravera, E. Colombo, C. Pasquale, et al., "Mitochondrial bioenergetics, photobiomodulation and trigeminal branches nerve damage: what's the connection? A review," Int. J. Mol. Sci., vol. 22, no. 9, Art. no. 4347, 2021, doi: 10.3390/ijms22094347.

[11] A. Liebert, A. Krause, N. Goonetilleke, et al., "A role for photobiomodulation in the prevention of myocardial ischemia reperfusion injury: a systematic review and potential molecular mechanisms," Sci. Rep., vol. 7, Art. no. 42386, 2017, doi: 10.1038/srep42386.

[12] C. Dompe, L. Moncrieff, J. A. Maty, et al., "Photobiomodulation: underlying mechanisms and clinical applications," J. Clin. Med., vol. 9, no. 6, Art. no. 1724, 2020, doi: 10.3390/jcm9061724.

[13] P. C. Pereira, C. J. De Lima, A. B. Fernandes, R. A. Zângaro, and A. B. Villaverde, "Cardiopulmonary and hematological effects of infrared LED photobiomodulation in the treatment of SARS-CoV-2," J. Photochem. Photobiol. B, vol. 238, Art. no. 112619, 2023, doi: 10.1016/j.jphotobiol.2022.112619.

[14] P. S. Azevedo, M. F. Minicucci, P. P. Santos, S. A. Paiva, and L. A. Zornoff, "Energy metabolism in cardiac remodeling and heart failure," Cardiol. Rev., vol. 21, no. 3, pp. 135–140, 2013, doi: 10.1097/CRD.0b013e318274956d.

[15] J. Francis, R. M. Weiss, S. G. Wei, A. K. Johnson, and R. B. Felder, "Progression of heart failure after myocardial infarction in the rat," Am. J. Physiol. Regul. Integr. Comp. Physiol., vol. 281, no. 5, pp. R1734–R1745, 2001, doi: 10.1152/ajpregu.2001.281.5.R1734.

[16] S. de Waha, M. R. Patel, C. B. Granger, et al., "Relationship between microvascular obstruction and adverse events following primary percutaneous coronary intervention for ST-segment elevation myocardial infarction: an individual patient data pooled analysis from seven randomized trials," Eur. Heart J., vol. 38, no. 47, pp. 3502–3510, 2017, doi: 10.1093/eurheartj/ehx414.

[17] M. Faustino, S. B. Baptista, A. Freitas, et al., "The index of microcirculatory resistance as a predictor of echocardiographic left ventricular performance recovery in patients with ST-elevation acute myocardial infarction undergoing successful primary angioplasty," J. Interv. Cardiol., vol. 29, no. 2, pp. 137–145, 2016, doi: 10.1111/joic.12278.

[18] H. Shmueli, G. Tsaban, A. Moreno, et al., "Echocardiographic predictors of improvement of left ventricular ejection fraction below 35% in patients with ST-segment elevation myocardial infarction," J. Clin. Med., vol. 13, no. 14, Art. no. 4016, 2024, doi: 10.3390/jcm13144016.

[19] G. C. Brooks, B. K. Lee, R. Rao, et al., "Predicting persistent left ventricular dysfunction following myocardial infarction: the PREDICTS study," J. Am. Coll. Cardiol., vol. 67, no. 10, pp. 1186–1196, 2016, doi: 10.1016/j.jacc.2015.12.042.

[20] G. Elbaz-Greener, M. Sud, O. Tzuman, M. Leitman, Z. Vered, N. Ben-Dov, U. Oron, and A. Blatt, "Adjunctive laser-stimulated stem-cells therapy to primary reperfusion in acute myocardial infarction in humans: safety and feasibility study," J. Interv. Cardiol., vol. 31, no. 6, pp. 711–716, 2018, doi: 10.1111/joic.12539.

Proof of Concept of Hand Tracking Using an Omnidirectional Vision Sensor in a Portable Device for Rehabilitation

Leonardo Marquez Pedro
Department of Mechanical Engineering
Federal University of São Carlos (UFSCar)
São Carlos, Brazil
lmpedro@ufscar.br

Glauco Augusto de Paula Caurin
São Carlos School of Engineering
University of São Paulo (USP)
São Carlos, SP, Brazil

Viviane Cristina Roma Appel
São Carlos School of Engineering
University of São Paulo (USP)
São Carlos, SP, Brazil

Erika Shirley Moreira da Silva
Department of Physical Therapy
Federal University of São Carlos (UFSCar)
São Carlos, SP, Brazil

Luisa Fernanda Garcia-Salazar
Department of Physical Therapy
Federal University of São Carlos (UFSCar)
São Carlos, SP, Brazil

Thiago Luiz Russo
Department of Physical Therapy
Federal University of São Carlos (UFSCar)
São Carlos, SP, Brazil

Etienne Burdet
Department of Bioengineering
Imperial College of Science, Technology and Medicine (ICL)
London, United Kingdom

***Abstract*—Kinematic assessment of the upper limb is relevant in several neurological and musculoskeletal conditions, but in clinical settings it is generally restricted to laboratories, since it depends on marker-based camera systems and instrumented gloves, which are expensive and impractical for clinical routine. This paper proposes a new measure, the position of the hand and fingertips along the trajectory, obtained by an omnidirectional vision system, formed by a camera and a hyperbolic mirror, embedded in an instrumented replica of a water cup that captures images of the hand during grasping, and presents a qualitative and observational proof of concept of this measure. The measure recovers hand transport and finger aperture modulation throughout reach-to-grasp, whose synchronization is the most compromised after injuries, particularly neurological ones. In the qualitative evaluation of the proposal, we used data from experiments in which, together with the device images, data from a motion tracking system were collected; the markers of this system are projected onto the unrolled images of the device, through a registration between the coordinate systems followed by projection using the numerical model of the vision system obtained by calibration. The consistent overlap of the markers onto the corresponding regions of the hand allows a qualitative evaluation of the proposed measurement. The device also has an inertial measurement unit, which provides complementary measures of phase timing, smoothness, and tremor, acquired in synchrony with hand posture. The device thus offers potential as a portable alternative of reduced cost relative to laboratory systems to bring kinematic assessment of the upper limb to clinical settings outside major centers.**

***Index Terms*—assessment device, omnidirectional vision, hand directional tracking, reach-to-grasp**

This work was supported by FAPESP, grant 2017/13655-6, by CNPq, grant 310798/2020-5, and by the Royal Society, grant NA140231.
Corresponding author: lmpedro@ufscar.br

I. Introduction

Drinking water is one of the activities of daily living (ADLs) most commonly used to assess and exercise upper-limb motor function, both in healthy adults [1] and typically developing children [2], as well as in individuals with a wide range of conditions, such as stroke [3], spinal cord injury [4], Parkinson's disease [5], cerebral palsy [6], chronic shoulder pain [7], and essential tremor [8]. In these conditions, upper-limb dysfunction compromises the performance of ADLs such as reaching for and grasping objects [9]. Two experimental protocols have been widely used to investigate these dysfunctions: *reach-to-grasp*, which analyzes the synchronization between hand transport toward the object and finger aperture for grasping [10], and *drinking from a glass*, which adds the transport of the cup to the mouth and the return to the initial position, constituting a complete functional task [3].

Kinematic assessment of these tasks provides quantitative variables that make it possible to understand the dysfunctions, that is, how each condition affects the upper limb, to track rehabilitation progress, and to support objective measures for research and scientific advancement. In stroke, for example, one of the most studied conditions, these variables reveal reduced peak velocity, longer movement time, lower trajectory smoothness, and loss of synchronization between hand transport and finger aperture, with maximum aperture tending to occur prematurely [10].

Acquiring these variables depends on marker-based optical tracking systems, which require a dedicated room with con-

trolled lighting, careful placement of markers on anatomical landmarks, and calibration. These requirements make frequent assessments and longitudinal follow-up unfeasible in clinical settings.

Beyond infrastructure requirements, optical tracking systems themselves present intrinsic technical limitations: occlusion of markers by body segments or by the manipulated object interrupts trajectories and requires interpolation, and label swapping between nearby markers introduces errors that are difficult to detect automatically. The hand is particularly poorly served, since the thumb joints, among the most affected in several neurological conditions, are among the least accessible to markers, and instrumented gloves, the usual alternative for the hand, are incompatible with severe spasticity. Together with infrastructure requirements, these limitations concentrate kinematic assessment in a few specialized centers and require patients and caregivers to travel for each session, imposing logistical and financial barriers that hinder and often prevent access, especially in remote and low-income regions.

Recent initiatives seek to overcome these limitations with portable, low-cost devices. Wearable inertial measurement units (IMUs) have been used to estimate clinical metrics and to characterize tremor and movement smoothness, counting movement units during the drinking task [11]. The smart cup of Bobin et al. [12] is a representative example: instrumented with inertial and pressure sensors, it monitors arm and hand activity during the drinking task, which is itself an ADL. However, isolated IMUs do not capture hand posture, and the metrics obtained with these systems still do not achieve full clinical reliability for use as reference values in rehabilitation [11].

A previously proposed device, named GULMN [2], [13], [14], integrates, in a single instrumented object, an omnidirectional vision system, formed by a camera and a hyperbolic mirror that captures images of the hand during grasping, and an IMU that provides kinematic data of the limb. Unlike wearable IMUs and the cup of Bobin et al., the embedded image provides, in synchrony with the inertial data, a direct observation of hand posture during grasping.

The objective of this paper is to present a new measure of the device, the position of the fingers and the palm of the hand along the trajectory, obtained from the omnidirectional images captured with the device, and also to carry out an observational proof of concept of this measure. Sections II and III describe the device and its measurement capabilities. Section IV proposes the new measure and Section V presents its proof of concept. Sections VI and VII discuss and conclude.

II. Device Design and Construction

The device is a replica of a transparent acrylic cylindrical cup, approximately 159.6 mm in height, 70 mm in diameter, and with a mass of approximately 250 g [14]. The cup is instrumented with two measurement subsystems, an omnidirectional vision system, formed by a camera and a hyperbolic mirror, that captures images of the hand during grasping, and an IMU. The hyperbolic mirror is mounted just below the cup's cap, above the transparent acrylic cylinder, and the base of the cup houses the camera, the mirror, and the IMU [14]. The camera and the IMU transmit their data by cable to an external computer, which performs acquisition and storage, at a sampling rate of 90 Hz [14]. All data are acquired synchronously (Fig. 1).



Figure 1. The instrumented cup and its measurement subsystems. Source: adapted from [14].

The most innovative measurement of the device is the omnidirectional image, obtained by a camera and a hyperbolic mirror, a configuration developed mainly in the context of mobile robotics [15], [16]. The camera, a BASLER acA1300-30gm with a resolution of 1280×960 pixels, is housed at the base of the cup and captures the image reflected by the mirror [14], with the requirement of obtaining the best possible resolution within the reduced dimensions of the cup. The mirror, made of aluminum, is machined on an ultra-precision lathe with a final finish by polishing with a synthetic diamond micro-tool, and is mounted just below the cap [14].

Through the image, the innovative data obtained in previous work is the posture of the hand while grasping the object [2], [13], [14]. The raw image captured by the camera is omnidirectional and maps the cylindrical surface around the cup onto a circular region of the sensor (Fig. 2). For the assessment of hand posture, this image is unrolled into a planar cylindrical panorama view, in which the angle around the cup and the height along it become orthogonal axes (Fig. 3). The reconstruction uses the calibrated geometry of the vision system to map each pixel to a direction in the scene: the calibration recovers the mapping between the radial position in the image and the height, or equivalently the inclination, over the cylindrical surface of the cup, while the azimuthal angle around the cup is preserved in the image [17].

An IMU embedded in the base of the cup (Bosch BNO055), with a triaxial gyroscope, accelerometer, and magnetometer, measures sagittal-plane acceleration and angular velocity dur-

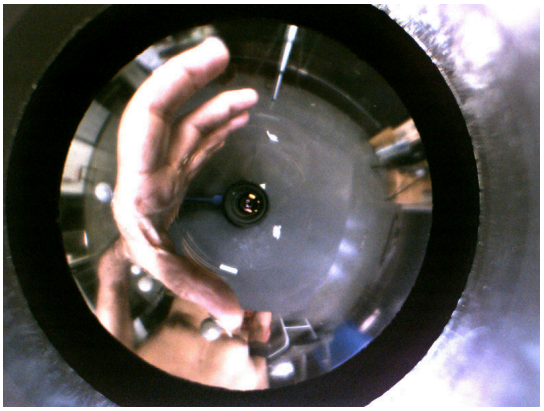


Figure 2. Raw omnidirectional image of hand grasping. Source: adapted from [14].



Figure 3. Unrolled image of hand grasping. Source: adapted from [14].

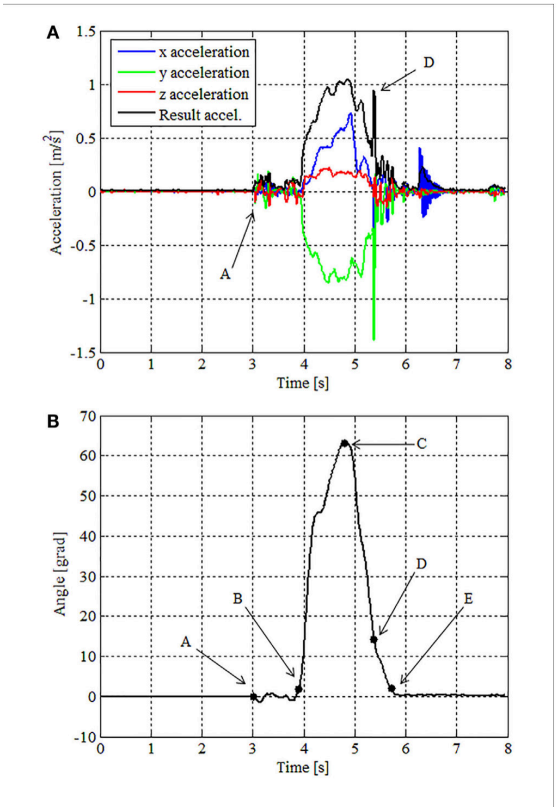


Figure 4. Temporal segmentation of the drinking task from the IMU signals, with the instants that delimit the accommodation, transport, and return phases. Source: adapted from [2].

ing the task and, by combining these signals, estimates the orientation and angular displacement of the limb, transmitted by cable at 90 Hz [14].

III. Measurement Capabilities

The capabilities reported here were established in previous work with the device, in which three groups of volunteers, namely typically developing children [2], healthy adults [14], and stroke survivors [13], [14], performed the drinking task. The study with typically developing children characterized grasp posture and spatiotemporal metrics across development [2]; the study with stroke survivors compared the more- and less-affected hands before and after rehabilitation [13]; and the doctoral thesis consolidated the methodology, bringing together temporal and spectral metrics, hand segmentation, and validation of the abduction angle [14]. The following observations are qualitative and are intended to illustrate the relevance of each measurement.

A. IMU-Based Measures

The IMU signals, acquired in synchrony with the grasping images, provide linear acceleration, angular velocity, and orientation of the limb throughout the task. The joint analysis of the resultant acceleration and sagittal-plane orientation locates the instants that delimit the movement, namely contact with the cup, movement onset, maximum amplitude near the mouth, contact with the table, and release, which define the accommodation, transport, and return phases (Fig. 4) [2], [13]. From this segmentation are derived the total duration and the duration of each phase, the hand accommodation time, mean and peak velocity parameters, the number of movement units, the straightness index, and the energy expenditure [2].

Orientation, angular velocity, and linear acceleration are well established in upper-limb biomechanics, including assessment after neurological injuries [3]. The measure proposed here is the temporal relationship between the grasping time, that is, the hand accommodation time, and the total task time. Stroke survivors showed a significantly longer grasping time in the more-affected hand, both in absolute and proportional terms, and a shorter total task time in the less-affected hand [13]; in longitudinal follow-up, this time decreased and correlated with standard assessments [14], suggesting its use for tracking grasp recovery.

Human movement has an average maximum frequency of approximately 2 Hz, and velocity oscillations above this value are identified as tremor. The device estimates trajectory smoothness by applying the fast Fourier transform to the angular velocity and relating the energy between 0 and 2 Hz, typical of healthy movement, to the energy between 2 and 6 Hz, typical of tremor. In the thesis that consolidated the methodology, the resulting spectral metric was consistently higher in the more-affected hand of stroke survivors than

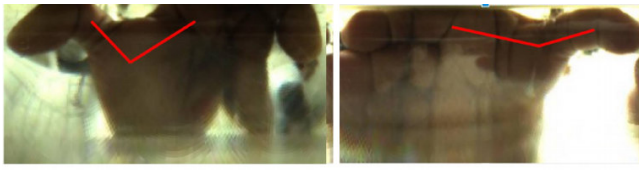


Figure 5. Grasping images of the more-affected (left) and less-affected (center) hand of a stroke survivor, with the thumb abduction angle indicated, and of a typical hand (right). Source: adapted from [13] and [2].

in the less-affected hand, correlated strongly with standard clinical tests, and tracked patient progress in the longitudinal experiments [14].

B. Hand Posture During Grasping from the Images

From the reconstructed images, the geometric analysis of posture provides thumb opposition, finger recruitment, and the abduction angle between the thumb and index finger [2], [14].

The main innovation of the device is the image of hand posture during grasping, particularly the abduction angle between the thumb and index finger (Fig. 5). The abduction angle is particularly promising for assessment, since the joints involving the thumb and index finger are among the most affected in several neurological conditions. The device also revealed the development of grasp posture in typically developing children, with the thumb opposing toward the palm in younger children and toward the index finger, closer to the adult pattern, in older children [2].

IV. Proposed New Measure: Directional Tracking of the Hand and Fingers

A. The Omnidirectional Vision System as a Directional Sensor

Omnidirectional vision systems formed by a camera and a hyperbolic mirror are widely used in mobile robotics for environment perception and motion estimation [16], [17]. In a catadioptric configuration, each pixel of the image corresponds to a viewing direction, defined by two angles, the azimuth around the symmetry axis of the vision system and the inclination angle. We propose exploring this property to locate, in spherical coordinates, any scene point tracked in the image.

In spherical coordinates (r, θ, φ), with origin at the focus of the mirror, where r is the radius, θ the inclination, and φ the azimuth, the two angles are obtained directly from the image. The azimuth φ is preserved by the image formation: the angular position of a point in the annular image, around the optical center, corresponds to the azimuth of its viewing direction around the axis of the cup. The inclination θ, in turn, results from the calibration of the vision system described in Section II, which associates the radial position of a point in the image with the elevation, or inclination, of its viewing direction [17].

From the azimuth and inclination thus obtained, it is possible to track the direction, frame by frame, of both the hand, taking the base of the palm as a reference, and the fingertips. Tracking the base of the palm describes hand transport toward the cup, while the relative tracking between the fingertips describes the modulation of hand aperture throughout the movement. The complete recovery of the three-dimensional position, including depth, is discussed in Section VI as a direction for future work. Even so, with a single embedded camera and without markers, the two components of reach-to-grasp, transport and finger aperture, are recovered.

B. Qualitative Experimental Validation

To qualitatively validate the proposed measure, we used data from experiments in which the device was employed together with a motion capture system taken as reference [14]. In this proof of concept, the positions of the hand and fingers are provided by the motion capture system's markers projected onto the images, which validates the geometric projection model; marker-free detection, directly in the images, is the next step.

Volunteer. One healthy individual participated, recruited from the local community, with no history of musculoskeletal injury of the upper limb. The intentional sample of a single participant is consistent with the observational nature of this proof of concept; expansion to larger samples, including individuals with motor impairment, is future work.

Ethical aspects. The data come from a study approved by the Human Research Ethics Committee of the Federal University of São Carlos, under approval report no. 2.495.168 (CAAE 79106017.0.0000.5504), in February 2018. All procedures followed the ethical guidelines in force, in accordance with Brazilian National Health Council (CNS) Resolution 466/12. The participant signed an informed consent form.

Procedure. The participant, seated with hip, knee, and ankle at 90° and without trunk support, performed the drinking task starting from a resting position on the table, with the device positioned at the height of the xiphoid process and at 80% of the upper-limb length. The task comprised reaching, transport to the mouth, return of the cup to the table, and return of the hand. After one familiarization trial, three repetitions were performed at self-selected speed, with the non-dominant limb [14]. For directional tracking, only the reach and transport phases were considered.

Acquisition and reference system. During the task, the device described in Section II recorded the omnidirectional images. In parallel, the movements were recorded by an optoelectronic Qualisys motion capture system (ProReflex, Qualisys Medical AB, Gothenburg, Sweden), with eight high-speed cameras at 120 Hz, taken as reference. Passive reflective markers were placed on the tips of the thumb and index finger and a marker on top of the cup, which allows comparing the trajectories estimated from the image with those of the reference system.

V. Results

A. Registration Between Reference Systems

The projection follows from a registration between the reference frame of the omnidirectional vision system, $\{V\}$, with origin at the focus of the mirror and axis z_V along the mirror's rotation axis, and the reference frame of the capture

system, $\{Q\}$. The Qualisys calibration aligns the axis z_Q with the vertical, that is, with the gravitational direction. It is then assumed, as a simplification, that the device is positioned with its rotation axis z_V also aligned with the vertical, so that the two z axes are parallel. As a consequence, the xy planes of the two systems are parallel to each other and to the horizontal, and the relative orientation is fully determined by a single rotation β about the vertical,

$$ {}^{Q}\mathbf{R}_V = \mathbf{R}_z(\beta) = \begin{bmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1} $$

The translation between the systems takes the origin of $\{Q\}$ to the viewpoint of $\{V\}$. The device has a fixed marker 14.5 mm above the focus of the mirror, which is taken as a reference: starting from its position $\mathbf{m}^Q$, a translation is performed along the vertical to the focus, preserving the orientation. The relative pose is thus described by the homogeneous transformation

$$ {}^{Q}\mathbf{T}_V = \begin{bmatrix} \mathbf{R}_z(\beta) & \mathbf{o}^Q \\ \mathbf{0}^\top & 1 \end{bmatrix}, \qquad \mathbf{o}^Q = \begin{bmatrix} A_x \\ A_y \\ h \end{bmatrix}, \tag{2} $$

where (A_x, A_y) is the center in the horizontal plane and h the height of the viewpoint, initialized at $\mathbf{m}^Q$ and at $m_z - 14.5$ mm and refined by the adjustment described next.

A marker at position $\mathbf{p}_i^Q$ in $\{Q\}$ is expressed in $\{V\}$ as $\mathbf{p}_i^V = {}^{Q}\mathbf{R}_V^\top(\mathbf{p}_i^Q - \mathbf{o}^Q)$. Since the z axes are parallel, the marker's azimuth in $\{V\}$ is the azimuth in $\{Q\}$ around the center, offset by β,

$$ \varphi_i = \operatorname{atan2}\big(p_{i,y}^Q - A_y,\ p_{i,x}^Q - A_x\big) - \beta. \tag{3} $$

The marker spheres are identified in the unrolled image of a single frame, chosen because it shows all hand markers visible. Each sphere provides an azimuth in the image, $\tilde{\varphi}_i$, obtained from its column position, since the unrolled image covers exactly 360° along its width. The registration then seeks the center and rotation that minimize the angular distance between the Qualisys-projected azimuths and those measured in the image,

$$ (A^*, \beta^*) = \arg\min_{A,\,\beta} \sum_i \operatorname{wrap}_\pi\big(\varphi_i - \tilde{\varphi}_i\big)^2, \tag{4} $$

where $\operatorname{wrap}_\pi(\cdot)$ normalizes the angular difference to the interval $(-\pi, \pi]$. The minimization is solved by nonlinear search, restarted from several initial rotations to avoid local minima.

With the orientation fixed, the height of each marker in the image is obtained from the calibration curve of the vision system, which converts the angular position in the spherical system into height on the reference cylinder of radius $\rho = 35$ mm adopted in the device calibration. By triangle similarity from the viewpoint, a marker at a horizontal distance $d_i = \|(A_x^*, A_y^*) - (p_{i,x}^Q, p_{i,y}^Q)\|$ and at a vertical offset Δz_i relative to the viewpoint is projected onto the reference cylinder at a height

$$ H_i = \rho\,\frac{\Delta z_i}{d_i}, \tag{5} $$

converted into an image row by the vertical scale of the unrolling. The height h is calibrated by best approximating, over the set of markers, the inclination read from the image, $\operatorname{atan2}(H_i, \rho)$, to the inclination computed from Qualisys, $\operatorname{atan2}(\Delta z_i, d_i)$. Once the registration is complete, the projection of each marker in any given frame is obtained by converting its azimuth φ_i into an image column and its height H_i into an image row, and on each frame of the reach and transport phases the markers and the segments connecting them are drawn.

B. Projection of the Markers onto the Reconstructed Images

The proof of concept of the measurement capability was carried out on the data from the device and the capture system, acquired in synchrony (Section IV). For each frame, the positions of the markers fixed at the thumb tip, the index fingertip, and the base of the palm, measured by the capture system, were projected onto the reconstructed image, using the calibrated catadioptric geometry of the vision system (Section II) and the registration between reference systems described above. The consistent overlap of the projected markers onto the corresponding parts of the hand in the image allows verification of the geometric projection model, a proof of concept of the proposed angular measurement capability.

The projected markers of the thumb, index finger, and palm fell consistently onto the corresponding parts of the hand in the reconstructed images, throughout the whole reach-to-grasp (Fig. 6). The trajectory of the palm-base marker describes hand transport toward the cup, while the relative motion between the thumb and index-finger markers describes the modulation of finger aperture throughout the movement. The joint projection of the transport and aperture trajectories, obtained from a single embedded camera, demonstrates the feasibility of the proposed measure and supports its future implementation without markers, through direct detection of the hand and fingers in the images.

VI. Discussion

The main result of this work is the visual verification of the correspondence between the two systems: based on the registration procedure and the device calibration, the thumb and index-finger markers, projected from the capture system, overlapped consistently with the hand in the reconstructed images, throughout the reach-to-grasp. This correspondence is a qualitative and observational proof of concept that omnidirectional vision carries the geometry needed to track hand transport and finger aperture modulation, and not a definitive quantitative validation of the proposed measure.

It remains to estimate depth, that is, the distance between the viewpoint of the vision system and the tracked point, which a single omnidirectional sensor does not provide directly. One

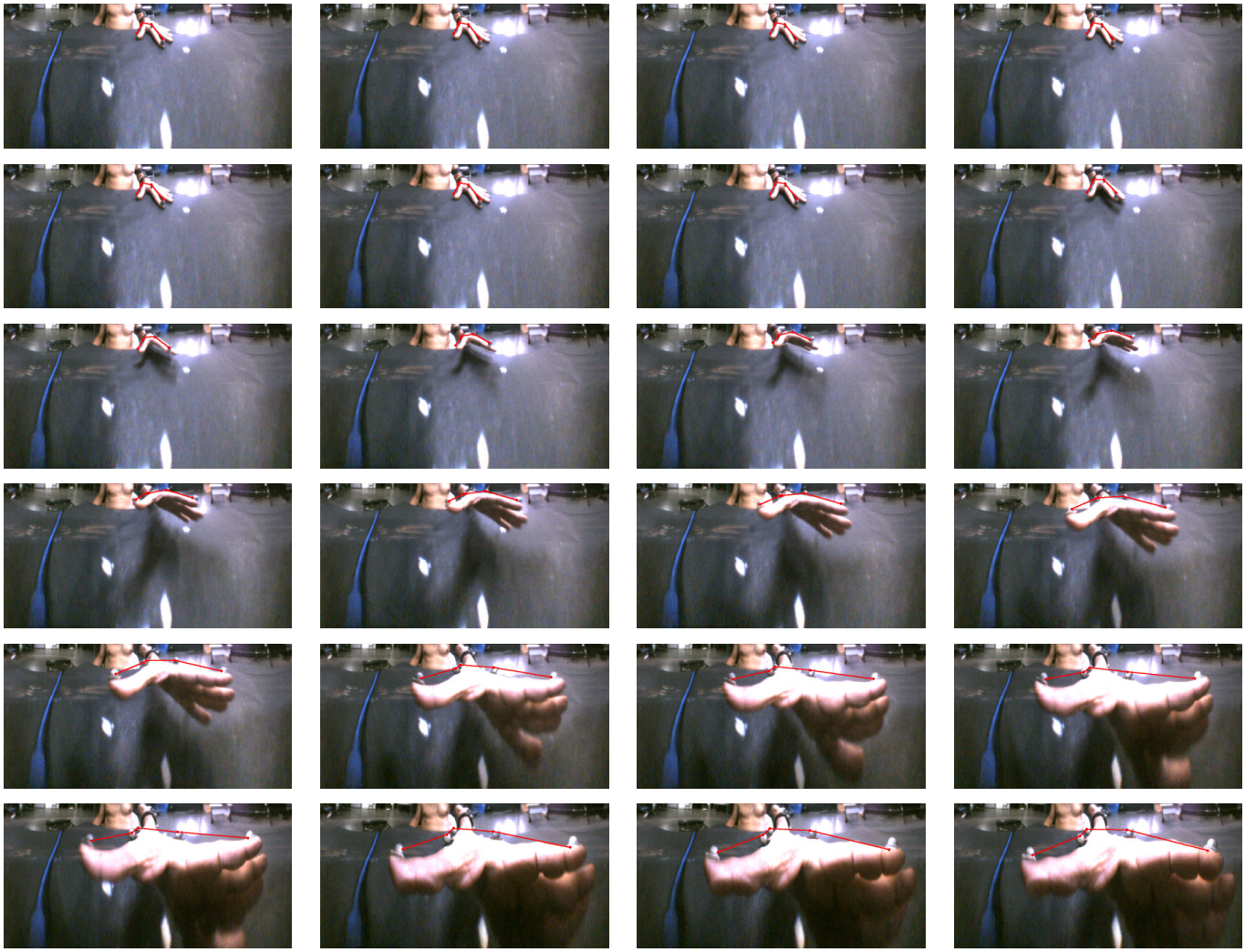


Figure 6. Projection of the capture system's markers onto the reconstructed images, for the thumb, index finger, and palm, along 24 equally spaced frames of the reach-to-grasp, for a healthy participant, arranged in 6 rows and 4 columns as in a video sequence.

possibility would be to estimate it from the ratio between the actual size of the hand, known from the cylindrical calibration of the vision system at the instant of grasping, and its apparent size in the image at each frame. This step has not yet been implemented or validated, and constitutes the main direction for future work to complete the three-dimensional tracking, recovering the three spherical coordinates of the position of the hand and fingertips, and not only their angular direction.

In addition to this result, the set of capabilities already established in previous work suggests that a single instrumented object of reduced cost relative to laboratory systems, with embedded sensing, can bring together findings that usually require laboratory motion capture and instrumented gloves, such as phase timing, hand posture during grasping, and movement smoothness. In particular, the abduction angle between the thumb and index finger, so far obtained by manual marking on the images, still does not allow conclusive results; its measurement connects directly to the three-dimensional tracking proposed in this paper, which could obtain it automatically and continuously throughout the movement, in synchrony with the other already-established IMU measures, allowing the correspondence between them to be studied in broader clinical contexts. Unlike wearable inertial sensors, which do not access hand posture, and the cup of Bobin et al. [12], which combines inertial and pressure sensors but does not image the hand, the device adds the omnidirectional image of grasping synchronized with the inertial data, with reliability and validity already indicated for the temporal and smoothness measures in the drinking task.

These measures, previously restricted to motion capture laboratories, could become accessible in a portable, marker-free cup of accessible cost relative to laboratory motion capture, which supports the argument for democratizing access to kinematic assessment. In clinical use, the device would have the potential to enable assessment and longitudinal follow-up of rehabilitation in primary care and smaller-scale services, without requiring patient travel, complementing the clinical instruments already established for upper-limb assessment,

such as the Fugl-Meyer Assessment [18], the ARAT [19], and the *Nine-Hole Peg Test* [20]. On a population scale, it would open the way for screening and monitoring upper-limb motor function in the various neurological and musculoskeletal conditions that compromise it, in regions far from major centers, addressing the scenario in which the concentration of assessment in a few centers contradicts the principles of universality and equity of Brazil's Unified Health System (SUS).

The main limitation of the current version is the cable that connects the cup to the external computer: according to reports from the volunteers and the team, it adds weight and resistance and restricts hand trajectory, which may bias the smoothness and tremor measures and reduce the ecological validity of the task, critical for use outside the laboratory. Added to this are the weight of the prototype and the cost of the hyperbolic mirror, whose manufacturing on an ultra-precision lathe, with a finish by polishing with a synthetic diamond micro-tool, is expensive and specialized. In addition, the geometry of the vision system occupies the interior of the cup and prevents it from holding water, so that the current version only mimics the drinking task. These points motivate an updated sensor design, wireless, lighter, with more affordable manufacturing for the mirror and, conveniently, with a water reservoir, conditions important to bring the task closer to the real activity and to realize the democratization potential discussed.

On the measurement side, the main directions for future work are: (i) quantitative validation of the geometric projection model, with metrics for angular error, agreement with the capture system, and repeatability across trials and participants; (ii) expansion of the sample to include individuals with motor impairment, such as stroke survivors and patients with Parkinson's disease, to verify discrimination between conditions and track longitudinal evolution; (iii) direct detection of the hand and fingers in the images, without external markers, the condition necessary to make the device independent of laboratory systems; and (iv) depth estimation, which would complete the tracking with three spherical coordinates, enabling, for example, automatic and continuous extraction of the abduction angle between the thumb and index finger. Together, these directions could convert posture metrics, today annotated manually, into continuous and automatic measurements throughout the movement, complementing or replacing external trackers and instrumented gloves with a single portable cup.

VII. Conclusion

This paper presented the device, an instrumented cup built as a replica of a water cup, for assessing the arm and hand in the various conditions that compromise the upper limb, such as stroke, Parkinson's disease, and spinal cord injury. We brought together its measurement capabilities already established in previous studies: the IMU signals, acquired in synchrony with the grasping images, from which phase timing, hand accommodation time, and smoothness and tremor by spectral analysis are derived; and hand posture during grasping from the images, including the abduction angle between the thumb and index finger.

As the main contribution, we proposed the use of the omnidirectional vision system as a directional angular sensor and presented a qualitative and observational proof of concept: the projection of the positions of the thumb, index finger, and palm onto the reconstructed images, which indicates that it is possible to recover hand transport and finger aperture modulation in reach-to-grasp, measured in synchrony with the other already-established measures of the device, precisely the components whose synchronization is most compromised after several neurological injuries. The projection of the markers confirmed that omnidirectional vision carries the geometry needed for this measure, and no comparable instrumented object simultaneously offers this possibility of spatial information of the hand and fingers during the functional task, which distinguishes the proposal in the state of the art of portable upper-limb assessment.

We also discussed its limitations, with emphasis on the cable of the current version and on depth, not yet estimated, which is necessary to complete the three-dimensional tracking.

By bringing together clinically relevant measures in a portable object of reduced cost relative to laboratory systems, the device points toward the democratization of upper-limb kinematic assessment, bringing it to rehabilitation settings outside major centers and aligning with the principles of universality and equity of Brazil's SUS. As directions with the greatest potential, we identify: quantitative validation of the geometric model, with angular error and agreement metrics; expansion of the sample to individuals with motor impairment; direct detection of the hand and fingers in the images, without external markers; depth estimation, necessary to complete three-dimensional tracking with three spherical coordinates; and a wireless, lighter version with more affordable mirror manufacturing.

Acknowledgment

The authors thank all the volunteers who participated in the experiments.

References

[1] J. Huber, S. Slone, and J. Bae, "Computer vision for kinematic metrics of the drinking task in a pilot study of neurotypical participants," *Scientific Reports*, vol. 14, p. 20668, 2024.

[2] A. C. P. Rocha, E. Tudella, L. M. Pedro, V. C. R. Appel, L. G. P. da Silva, and G. A. P. Caurin, "A novel device for grasping assessment during functional tasks: preliminary results," *Frontiers in Bioengineering and Biotechnology*, vol. 4, p. 16, 2016.

[3] M. A. Murphy, C. Willén, and K. S. Sunnerhagen, "Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass," *Neurorehabilitation and Neural Repair*, vol. 25, no. 1, pp. 71–80, 2011.

[4] L. Lili, K. S. Sunnerhagen, and T. Rekand, "Participation and autonomy, independence in activities of daily living and upper extremity functioning in individuals with spinal cord injury," *Scientific Reports*, vol. 14, p. 9120, 2024.

[5] K. M. B. Bennett, M. Marchetti, R. Iovine, and U. Castiello, "The drinking action of Parkinson's disease subjects," *Brain*, vol. 118, no. 4, pp. 959–970, 1995.

[6] E. Jaspers, H. Feys, H. Bruyninckx, K. Klingels, G. Molenaers, and K. Desloovere, "Three-dimensional upper limb movement characteristics in children with hemiplegic cerebral palsy and typically developing children," *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2283–2294, 2011.

[7] A. S. C. Melo, D. C. Guedes, R. Matias, E. B. Cruz, J. P. Vilas-Boas, and A. S. P. Sousa, "Scapular motor control and upper limb movement quality in subjects with and without chronic shoulder pain: a cross-sectional study," *Applied Sciences*, vol. 14, no. 8, p. 3291, 2024.

[8] N. Roth, A. Salih, and S. Rosenblum, "Subjective and objective day-to-day performance measures of people with essential tremor," *Sensors*, vol. 24, p. 4854, 2024.

[9] World Health Organization, *The World Health Report*. Geneva, Switzerland: WHO, 2002.

[10] J. van Kordelaar, E. E. van Wegen, and G. Kwakkel, "Unraveling the interaction between pathological upper limb synergies and compensatory trunk movements during reach-to-grasp after stroke: A cross-sectional study," *Experimental Brain Research*, vol. 221, no. 3, pp. 251–262, 2012.

[11] S. Patel *et al.*, "A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology," *Proceedings of the IEEE*, vol. 98, no. 3, pp. 450–461, Mar. 2010.

[12] M. Bobin, M. Anastassova, M. Boukallel, and M. Ammi, "Design and study of a smart cup for monitoring the arm and hand activity of stroke patients," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, p. 2100812, 2018.

[13] V. C. R. Appel, R. E. Garcia, G. R. Chiqueti, L. M. Pedro, D. M. C. Cruz, and G. A. P. Caurin, "Novel assessment measures of upper-limb function in pre and poststroke rehabilitation: a pilot study," in *Proc. IEEE RAS/EMBS Int. Conf. Biomedical Robotics and Biomechatronics (BioRob)*, 2018.

[14] V. C. R. Appel, "Avaliação de Preensão e de Movimentos de Membro Superior em Indivíduos Hemiparéticos com o Dispositivo SPMM," Ph.D. dissertation, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, Brazil, 2019.

[15] S. Baker and S. K. Nayar, "A theory of single-viewpoint catadioptric image formation," *International Journal of Computer Vision*, vol. 35, no. 2, pp. 175–196, 1999.

[16] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to Autonomous Mobile Robots*, 2nd ed. Cambridge, MA, USA: MIT Press, 2011.

[17] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2006, pp. 5695–5701.

[18] A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance," *Scandinavian Journal of Rehabilitation Medicine*, vol. 7, no. 1, pp. 13–31, 1975.

[19] J. H. van der Lee, H. Beckerman, G. J. Lankhorst, and L. M. Bouter, "The responsiveness of the action research arm test and the Fugl-Meyer assessment scale in chronic stroke patients," *Journal of Rehabilitation Medicine*, vol. 33, no. 3, pp. 110–113, 2001.

[20] V. Mathiowetz, G. Volland, N. Kashman, and K. Weber, "Adult norms for the nine hole peg test of finger dexterity," *Occupational Therapy Journal of Research*, vol. 5, no. 1, pp. 24–38, 1985.

Raman spectroscopy Applied to the Pharmacovigilance of *Viscum album* in Healthy Snakes

1st Gerson Batista Picca
Dept. Biomedical Engineering
Anhembi Morumbi University (UAM)
São José dos Campos, SP, Brazil
gersonbp@gmail.com

2nd Maristela Possati Porto
Dept. Biomedical Engineering
Anhembi Morumbi University (UAM)
São José dos Campos, SP, Brazil
maristelapossati@gmail.com

3rd Henrique Cunha Carvalho
Dept. Electronic Engineering (DAELN-CM – PPGEB-CT)
Federal University of Technology - Paraná (UTFPR)
Campo Mourão, PR, Brazil
hccarvalho@utfpr.edu.br

4th Landulfo Siveira Jr
Dept. Biomedical Engineering
Anhembi Morumbi University (UAM)
São José dos Campos, SP, Brazil
landulfo.silveira@gmail.com

5th Lívia Helena Moreira
Dept. Biomedical Engineering
Anhembi Morumbi University (UAM)
São José dos Campos, SP, Brazil
lh.medicinaveterinaria@gmail.com

***Abstract*— This study evaluated the safety of using *Viscum album* (*VA*) in both ultra-diluted (UD) and fermented (F) forms in healthy snakes at the Butantan Institute, using Raman spectroscopy combined with Principal Component Analysis (PCA). Fourteen snakes received monthly applications of *VA* according to the specific herbal formulation: UD (D3+D9) for six months and F (D3) for three months. Serum samples were collected and analyzed via Raman spectroscopy. The mean spectra for the UD and F groups exhibited the same set of bands as the control serum, with no bands unique to or missing from the control serum. Among the principal components, only PC4 and PC7 showed statistically significant differences (ANOVA, $p < 0.05$); these were linked to subtle variations in spectral regions associated with lipids, phenylalanine, and nucleic acids, a pattern that may be consistent with the reported immunomodulatory activity of the extract. The results indicate that Raman spectroscopy detected discrete biochemical modulation, with no spectroscopic evidence of toxicity under the conditions evaluated; complementary hematological, biochemical, and histopathological assessments are recommended to corroborate these findings.**

Keywords*— *Raman spectroscopy, serum, snake, Viscum album.

I. Introduction

The development of therapeutic alternatives with minimal side effects is a cornerstone of modern medicine and biomedical engineering. In this context, substances of natural origin, such as *Viscum album* (*VA*), have attracted growing scientific interest [1,2] due to their immunomodulatory properties and low toxicity, as well as their safety in both animal and human species, even at high doses [3].

VA is a perennial hemiparasitic plant that grows on host trees, such as oaks and apple trees [1,2]. Its main constituents include viscotoxins, lectins (ML1 to ML4) [3], flavonoids, and phenolic acids [1], which are responsible for antitumor and immunomodulatory effects [3].

Commercially, various formulations are available, notably fermented (F) and ultra-diluted (UD) extracts. Fermentation, as a biotechnological process, alters the extract's metabolomic profile and molecular organization, thereby modifying the availability of bioactive compounds. These changes can impact biological properties and structural stability, making molecular characterization essential before experimental application, particularly in animal models [1,3].

Herpetofauna, specifically snakes of the family Viperidae, play a significant role in biomedical research, being widely used in the production of antivenoms and in studies of venom composition and biological activity [4,5]. Increased longevity among these animals in research and conservation institutions has been accompanied by a higher incidence of chronic diseases, including neoplasms. However, studies evaluating the pharmacovigilance of new therapeutic approaches in this class remain scarce. Thus, the use of clinically healthy snakes represents a crucial step in investigating potential molecular alterations induced by compounds of interest, enabling estimation of their safety margin before application in diseased animals [3].

Among molecular characterization techniques, Raman spectroscopy is a rapid, non-destructive vibrational method capable of analyzing small sample quantities with minimal preparation and reduced animal handling. It relies on the interaction of light with molecular vibrations. It yields characteristic spectra in the 400–1800 cm^{-1} range, enabling the identification of biochemical changes, particularly in biological samples [6,7]. When combined with Principal Component Analysis (PCA), it becomes a powerful tool for detecting subtle spectral differences between experimental groups, thereby contributing to the assessment of the molecular stability of formulations [7,8].

Thus, the present study aims to investigate *VA* pharmacovigilance in healthy snakes using Raman spectroscopy.

II. Materials and Methods

A. Ethics and animals' selection

Healthy snakes (n=14) of the family Viperidae, belonging to the genus *Bothrops* sp., kept in captivity at the Butantan Institute's herpetology facilities in São Paulo, Brazil (CEUA No. 7991151220002256), participated in this study. All animals were adults of both sexes and were housed in a climate-controlled environment; they were fed monthly with mice bred and euthanized at the institution's own animal facility.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Blood samples were collected monthly from the caudal vein of healthy snakes, corresponding to approximately 1% of body weight. The samples were then packaged and stored at -20°C until shipment to the Laboratório de Espectroscopia Vibracional of the Center for Innovation, Technology and Education (CITÉ) in São José dos Campos, SP, Brazil, for analysis (Figure 1).



Fig. 1. Experimental assay with the snakes. (a) Snake restraint; (b) Collection via caudal of blood samples for serology; and (c) Samples processing for Raman spectroscopy.

The experimental setup involved a baseline assessment of all healthy snakes, followed by *VA* administration to the same individuals and subsequent sample collection. The animals were assigned to three groups: Control (Ctrl / zero) (n=14), Fermented (F) (n=4), and Ultra-diluted (UD) (n =10).

B. VA characterization

VA preparations were supplied by Injectcenter® (Injectcenter Manip Prod Farm Ltda., Ribeirão Preto, SP, Brazil) and were used in both UD and F formulations, packaged in 1.0 mL ampoules. The manufacturer provides preparations at different potencies (D2, D3, D6, D9, D12, D30, D60, and 200CH). According to the experimental protocol, the UD formulation consisted of a combination of the D3 and D9 potencies, administered at 0.5 mL of each potency per animal (1.0 mL total). In contrast, the F formulation consisted of the D3 potency at 0.5 mL per animal.

The phytotherapeutic agents were stored in their original ampoules; the ultra-diluted *VA* was kept at room temperature, while the fermented *VA* was refrigerated, in accordance with the manufacturer's guidelines. Upon opening, the contents of the ampoules were immediately administered to each animal in the respective group, via injection into the coelomic cavity, the standard parenteral administration site in snakes.

The treatments consisted of the monthly administration of the herbal *VA* preparation (one application per month), with UD-VA administered for 6 months and F-VA for 3 months. The applications were performed at times distinct from feeding and blood sample collection.

C. Raman spectroscopy

The blood sera underwent a single controlled thawing cycle at room temperature, ensuring batch homogeneity and minimizing artifacts resulting from protein degradation.

Aliquots of 80 µL from each sample were deposited onto an aluminum plate using a micropipette (mod. P200, Bio-Rad Laboratories Inc., Hercules, CA, USA) for subsequent reading. Analyses were performed using a dispersive Raman spectrometer (mod. Dimension P-1, Lambda Solutions Inc., Waltham, MA, USA) equipped with an 830 nm diode laser, 250 mW power, and a resolution of approximately 2 cm^{-1} in the 400–1800 cm^{-1} range; spectra were obtained through repeated collections and triplicate spectral acquisition.

Data were pre-processed using Chemoface (v. 1.75, UFLA, Lavras, MG, Brazil) [9], which involved fluorescence removal and baseline correction with a 7^{th}-order polynomial, followed by normalization to eliminate intensity variations.

Subsequently, the spectral mean for each group was calculated, allowing identification of differences in Raman spectral peak intensities and their correlation with the biochemical compositions described in the literature. The Raman spectra were then subjected to Principal Component Analysis (PCA) to identify relevant spectral variables.

This entire statistical procedure and the plotting of score graphs were carried out using MATLAB software (v. 2007a, The MathWorks, Inc., Natick, MA, USA). At the same time, analysis was performed using InStat software (v. 3.0, GraphPad Software Inc., San Diego, CA, USA), employing parametric one-way ANOVA and nonparametric comparisons (Kruskal-Wallis test with post hoc tests), followed by Tukey's test to compare all groups. The PCs were used to identify the groups in which these spectral differences were statistically significant ($p < 0.05$).

III. Results and Discussion

Raman spectra remained adequate for the 14 animals. They were used to calculate the mean Raman spectra for the Control (Ctrl: n=24), Fermented *VA* (F-VA: n=39), and Ultra-diluted VA (UD-VA: n=67) groups, which revealed a spectral profile with bands predominantly related to proteins, aromatic amino acids, and lipids [6,7,8,10,11] (Figure 2 and Table 1).

Spectra were obtained in triplicate at repeated monthly time points from the same 14 animals; therefore, spectra were treated as repeated measurements nested within each animal, and the reported statistical comparisons should be interpreted with this repeated-measures structure in mind.

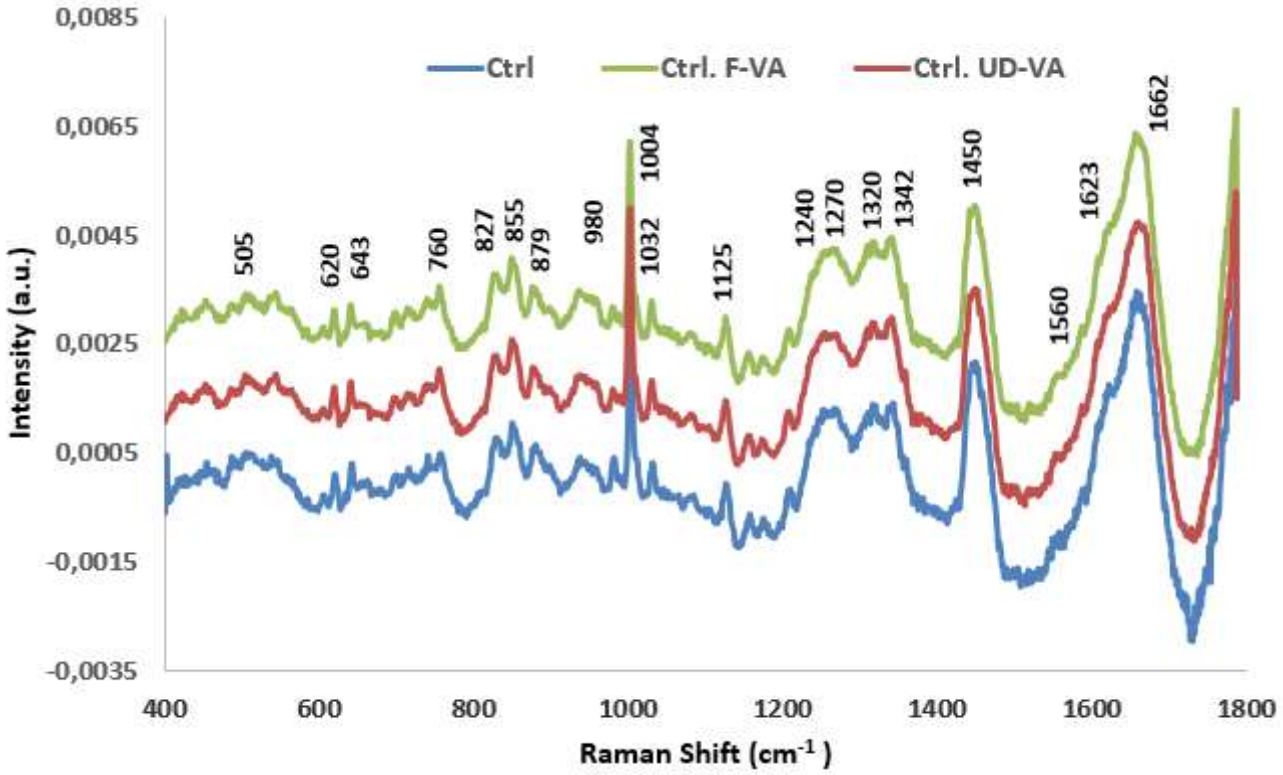


Fig. 2. Mean Raman spectra of the Control (Ctrl), F-VA and UD-VA Control groups.

TABLE I. BIOCHEMICAL ASSIGNMENTS OF THE RAMAN BANDS IDENTIFIED IN THE SERUM SPECTRA OF HEALTHY SNAKES.

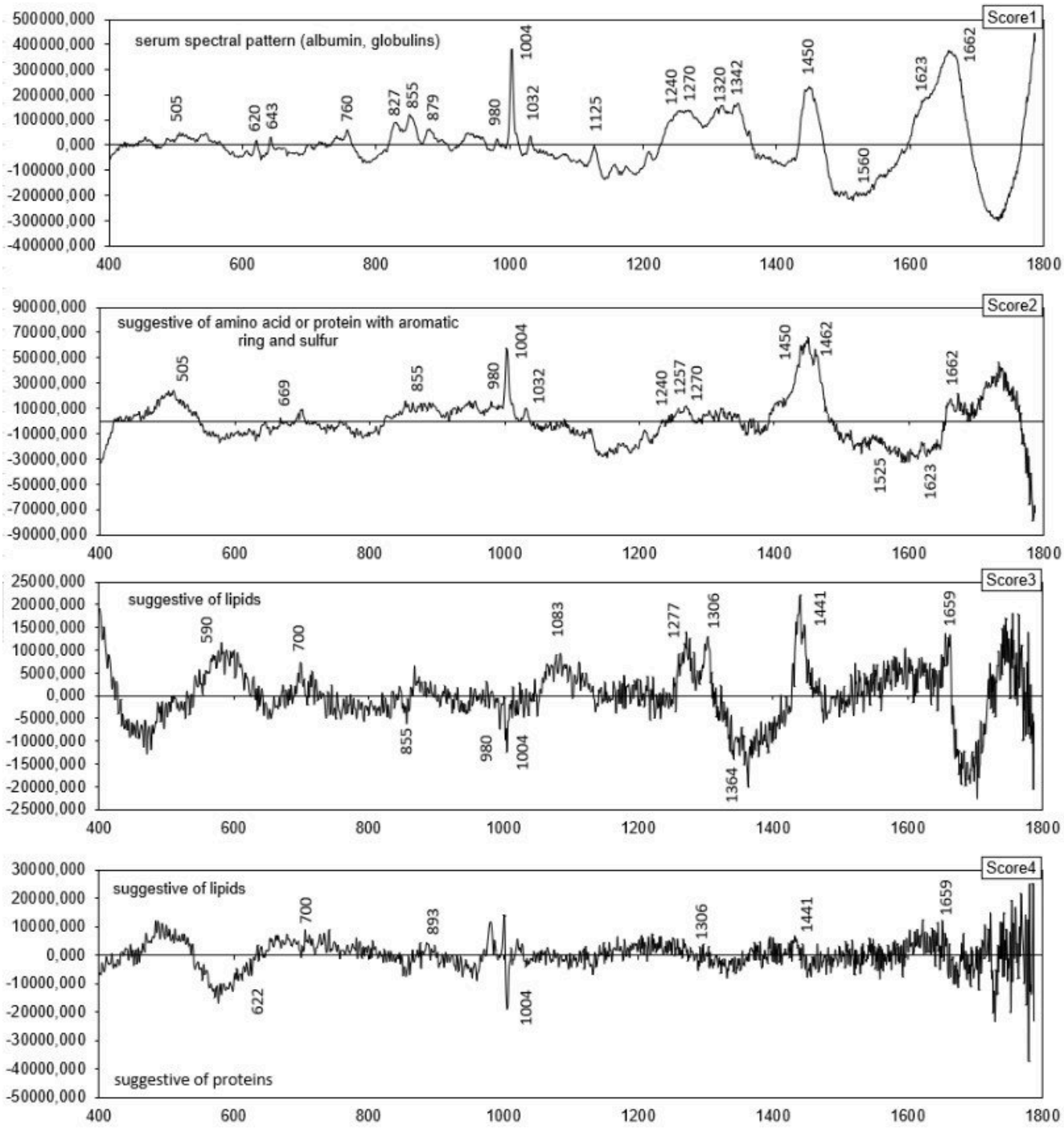
Peak (cm^{-1}) [*reference*]	**Assignment**
505 [10]	Disulfide bonds (S–S stretching), cysteine/cystine
620-643 [10, 13]	Phenylalanine and aromatic amino acids
760, 827 [10]	Tryptophan and indole ring
855, 879 [11]	Tyrosine and proteins (C–C stretching)
980 [10]	Symmetric stretching of proteins and phosphates
1004 [10]	Phenylalanine (aromatic ring)
1032 [11, 14]	Phenylalanine and proteins
1125 [11]	C–C stretching vibrations of lipids and proteins
1240-1270 [11, 15]	Amide III
1320, 1342 [10, 15]	Amide III and proteins
1450 [11, 16]	CH_2 deformation of lipids and proteins
1560, 1623 [10, 17]	Amide
1659-1662 [11, 17]	Amide I and lipids

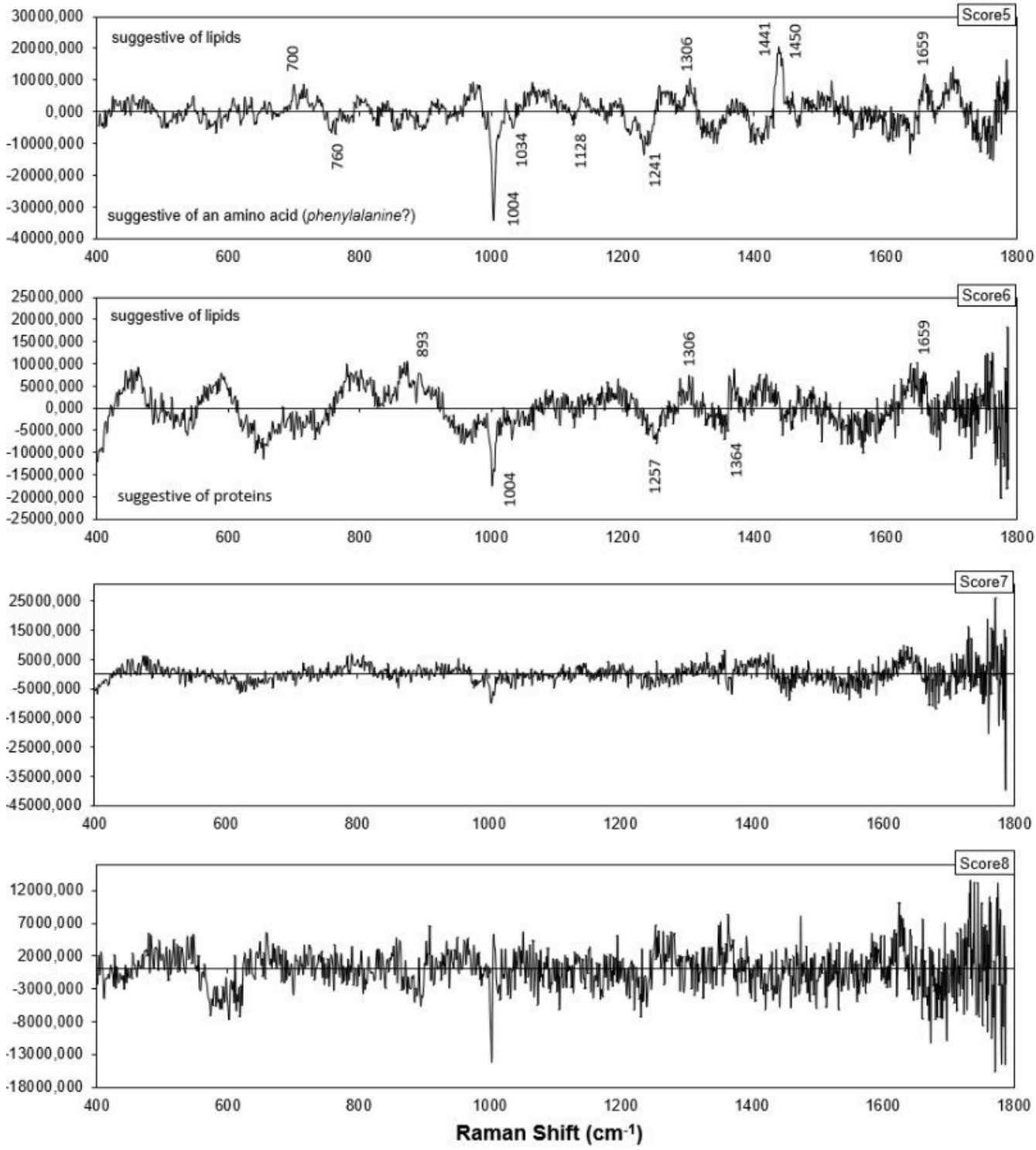
A comparison of the mean Raman spectra across the control groups revealed neither group-specific bands nor the disappearance of bands from the snakes' physiological serum profile, suggesting that these variations are common among the groups, as observed in multivariate analyses of biofluids [12]. Most of the components evaluated (PC1, PC2, PC3, PC5, PC6, and PC8) are predominantly associated with the general albumin and globulin pattern and the serum lipid or protein fraction; no significant differences were observed between the groups (Figure 3).

This preservation of the spectral pattern provides an initial indication that *VA* use does not induce significant biochemical alterations in the animals' serum. PCA analysis demonstrated statistically significant differences only in PC4 and PC7 (ANOVA, $p < 0.05$).

In PC4, variations in relative intensity occurred in the 622 and 1004 cm^{-1} regions, bands assigned to protein- and lipid-related molecular vibrations, the control group was more specific to lipids. These variations suggest subtle relative changes in protein-associated spectral contributions without altering the fundamental physiological spectral profile. PC7 exhibited loading peaks in spectral regions similar to those observed for the other principal components, mainly corresponding to lipid- and protein-related bands.

This behavior indicates only an opposing contribution to the formation of the principal component and is not interpreted as an absolute reduction in the concentration of this metabolite. The observed increase in the fermented group's score compared with other experimental control groups further supports that the separation between groups occurred predominantly in PC4, indicating subtle biochemical modifications without a significant impact on the spectral profile organization (Figure 3).


Score1
serum spectral pattern (albumin, globulins)
505
620
643
760
827
855
879
980
1004
1032
1125
1240
1270
1320
1342
1450
1560
1623
1662
Score2
suggestive of amino acid or protein with aromatic ring and sulfur
505
669
855
980
1004
1032
1240
1257
1270
1450
1462
1525
1623
1662
Score3
suggestive of lipids
590
700
855
980
1004
1083
1277
1306
1364
1441
1659
Score4
suggestive of lipids
622
700
893
1004
1306
1441
1659
suggestive of proteins


Score5
suggestive of lipids
700
760
1004
1034
1128
1241
1306
1441
1450
1659
suggestive of an amino acid (phenylalanine?)
Score6
suggestive of lipids
893
1004
1257
1306
1364
1659
suggestive of proteins
Score7
Score8
Raman Shift (cm-1)

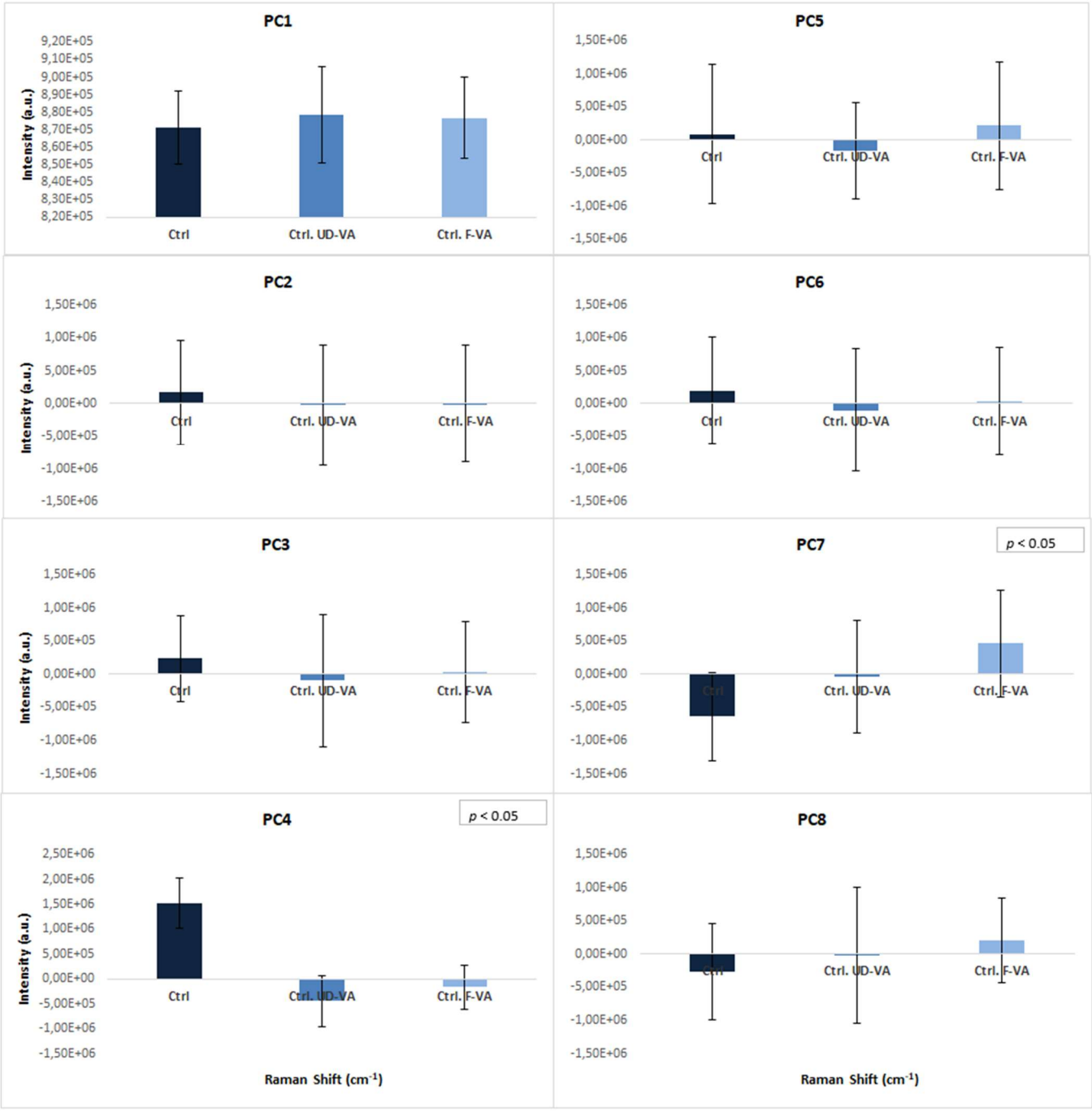


Fig.3. Loadings (intensity *vs*. Raman shift) and statistical comparison of the principal components (PC1–PC8) obtained via PCA of the Raman spectra from the Control (Ctrl), F-VA and UD-VA Control groups.

Taken together, the results indicate that both the UD and the F extract induced subtle biochemical changes detectable via Raman spectroscopy, while preserving the overall organization of the snake serum molecular profile.

The fact that statistical discrimination between groups occurred only in PC4 and PC7, components that typically capture less variance than the leading components (PC1 and PC2), demonstrates that higher-order principal components can contain biologically relevant information, even when representing a smaller portion of the data variability [7,8]. This suggests that the effect of the *VA* was limited to specific, subtle biochemical variations, without causing detectable global changes in the serum metabolic profile, within the sensitivity of the Raman approach employed. These results support the interpretation that using the extract, in either UD or F form, produces a biochemical modulation detectable by Raman spectroscopy without compromising the overall biochemical integrity of the animals' serum.

Therefore, the changes observed in PC4 and PC7 may reflect the action of compounds in the *VA* that can induce slight modifications in the snakes' serum metabolism, primarily involving proteins, aromatic amino acids, nucleic acids, and possibly lipids. This behavior is compatible with the extract's chemical composition: RIP-2 type lectins (viscumins) [18], which inactivate ribosomes and inhibit protein synthesis, could influence the availability of free amino acids such as phenylalanine (compatible with bands associated with aromatic amino acids), while phenolic

compounds and flavonoids [19] exhibit recognized antioxidant activity capable of influencing lipid peroxidation. Furthermore, the subtle contribution of bands associated with nucleic acids is compatible with a slight increase in cell turnover, potentially linked to the extract's immunomodulatory effect on immune cells [1,5]. It is important to emphasize that the PCA loadings support only spectral associations between these compound classes and the observed principal components; they do not, by themselves, provide evidence of metabolite concentration, protein synthesis rate, immune modulation, or cell turnover. While these biological properties have been described in the literature for *VA*, in the present study, they are interpreted as consistent with the spectroscopic findings; it is not yet possible to establish a direct causal link between these mechanisms and the observed alterations.

The spectral pattern observed in this study differs from that typically described in cases of actual tissue damage analyzed via Raman spectroscopy, patterns characterized by broad, widespread spectral changes, such as increased lipid content associated with catabolic products in liver damage [20] or the appearance of hemoglobin bands indicating hemolysis in systemic biocompatibility studies [21]. No broad, disseminated spectral change was detected in the present study; however, this preserved Raman profile, on its own, cannot establish the absence of toxicity, since clinical observation, hematology, serum biochemistry, and histopathology were not assessed in this work. Complementary evaluation via such methods is therefore advisable to corroborate the results obtained here.

IV. CONCLUSION

Raman spectroscopy combined with PCA proved useful for monitoring the effects of VA in healthy snakes. The detected changes were subtle and limited to specific components, with no spectroscopic evidence of toxicity under the conditions evaluated. Further clinical, hematological, biochemical, and histopathological analyses are recommended to confirm these results.

ACKNOWLEDGEMENT

G. B. P. and M. P. P. thank PROSUP/CAPES for the Doctoral Scholarship (Finance Code 001), and Anhembi Morumbi University. L. S. and L. H. M. would like to thank Instituto Ânima (IA) for the research grant. H. C. C. thanks Fundação Araucária (FA) & Superintendência de Ciências, Tecnologia e Ensino Superior (SETI) for the conference grant.

REFERENCES

[1] M. Nicoletti, “The anti-inflammatory activity of *Viscum album*”, Plants, v. 12, n. 7, pp. 1460, 2023, doi: 10.3390/plants12071460.

[2] S.L. Adzhiakhmetova, N.M Chervonnaya, D.I. Pozdnyakov, O.I. Popova, E.T. Oganisyan, “Component composition and features of biological activity of *Viscum album* (Viscaceae)”, Dokl. Biol. Sci., v. 518, n. 1, p. 116-132. Doi: 10.1134/S00124.96624.701072. epub.2024.

[3] G.S. Kienle, R. Grugel, H. Kiene, “Safety of higher dosages of *Viscum album* L. in animals and humans: Systematic review of immune changes and safety parameters”, BMC Complement Altern. Med., v. 11, pp. 72, 2011, doi: 10.1186/1472-6882-11-72.

[4] J.E. Chang Estrada, T.N. Guerrero, D.F. Reyes-Enríquez, E.S. Nardy, R. Guimarães Ferreira, C.J. Ruiz Calderón, I.A. Wellmann, K.M. Monteiro Espíndola, A.F. do Prado, A.M. Soares, M.R.M. Fontes, M. Chagas Monteiro, R.B. Zingali, “Potential biotechnological applications of venoms from the viperidae family in central america for thrombosis”, Toxins (Basel), v. 16, n. 3, pp. 142, 2024. doi: 10.3390/toxins16030142

[5] C. Steinborn, A.M. Klemd, A. S. Sanchez-Campillo, S. Rieger, M. Scheffen, B. Sauer, M. Follo, A. Ucker, G.S. Kienle, R. Huber, C. Grundemann, “*Viscum album* neutralizes tumor-induced immunosuppression in a human *in vitro* cell model”, PLOS One, v. 12, n. 7, pp. e0181553, 2017, doi: 10.1371/journal.pone.0181553.

[6] G. Acri, A. Falcone, C. Giannetto, E. Giudice, P. Guisepe, B. Testagrossa, L. Cícero, “Preliminary study for the identification of *Leishmania* infected dogs”, Sci. Rep., v. 12, n. 1. pp. 7489, 2022, doi: 10. 1038/s41598-022-11525-w

[7] C.R. da Silva, H.C. Carvalho, R.A. Lazo Osório, A.B. Fernandes, L. Silveira Jr, “Differences in whole blood before and after hemodialysis session of subjects with chronic kidney disease measured by Raman spectroscopy”, Lasers Med. Sci., v. 39, n. 1, pp. 175, 2024. doi: 10.1007/s10103-024-04125-9.

[8]. F. Bonnier, H.J. Byrne, “Understanding the molecular information contained in principal component analysis of vibration spectra of biological systems”, Analyts, 137(7), pp. 322-332. Doi:10.1039/c1an15821j

[9] C.A. Nunes, M.P. Freitas, A.C.M. Pinheiro, S.C. Bastos, “Chemoface: a novel free user-friendly interface for chemometrics”, J. Braz. Chem. Soc., v. n. 23, pp. 2003–2010, 2012. doi: 10.1590/S0103-50532012005000073.

[10] Z. Movasaghi, S. Rehman, I.U. Rehman, “Raman spectroscopy of biological tissues”, Appl. Spect. Rev., v. 42, n. 5, pp. 493-541, 2007, doi: 10.1080/05704920701551530.

[11] A.C.S. Talari, Z. Movasaghi, S. Rehman, U. Rehman “Raman spectroscopy of biolocal tissues”, Appl. Spect. Rev., v. 50, n. 1, pp. 46-111, 2015, doi: 10. 1080/05704928.2014.923902.

[12] M. Fan, Y. Weng, Y. Liu, Y. Lu, L. Xu, J. Ye, D. Lin, S. Qiu, S. Feng, “Surface-Enhanced Raman spectroscopy-based optical biosensor for liquid biopsy: toward precision medicine”, Laser Photonics Rev., v. 18, n. 5, pp. 2301072, 2024. doi: 10.1002/lpor.202301072

[13] J.L. González-Solís, J.C. Martínez-Espinosa, J.M. Salgado-Román, P. Palomares-Anda “Monitoring of chemotherapy leukemia treatment using Raman spectroscopy and principal component analysis”, Lasers Med. Sci.; v. 29, n. 3, pp. 1241-1249, 2014. doi: 10.1007/s10103-013-1515-y

[14] J.C. Martínez-Espinosa, J.L. González-Solís, C. Frausto-Reyes, M.L. Miranda-Beltrán, C. Soria-Fregoso, J. Medina-Valtierra “Detection of leukemia with blood samples using Raman spectroscopy and multivariate analysis”, AIP Conf. Proc., v. 1142, n. 1, pp. 99-103, 2009. doi: 10.1063/1.3175637

[15] J. De Gelder, K. De Gussem, P. Vandenabeele, L. Moens “Reference database of Raman spectra of biological molecules”, J. Raman Spectrosc., v. 38, pp. 1133-1147, 2007. doi: 10.1002/jrs.1734

[16] A.M.F. Lima, C.R. Danie, M.T.T. Pacheco, P.L. de Brito, L. Silveira Jr. “Discrimination of leukemias and non-leukemic cancers in blood serum samples of children and adolescents using a Raman spectral model”, Lasers Med. Sci., v. 38, n. 1, pp. 22, 2022. doi: 10.1007/s10103-022-03681-2

[17] E. Vargas-Obieta, J.C. Martínez-Espinosa, B.E. Martínez-Zerega, L.F. Jave-Suárez, A. Aguilar-Lemarroy, J.L. González-Solís “Breast cancer detection based on serum sample surface enhanced Raman spectroscopy”, Lasers Med. Sci.; v. 31, n. 7, pp. 1317-1324, 2016. doi: 10.1007/s10103-016-1976-x

[18] M.A. Souza, F.C. Carvalho, L.P. Ruas, R. Ricci-Azevedo, M.C. Roque-Barreira, “The immunomodulatory effect os plant lectins: A review with emphasis on ArtimM properties”, Glycoconj J., v. 30, n. 7, pp. 641-657, 2013, doi: 10.1007/s10719-012-9464-4.

[19] N. Nicoletti, “The antioxidant activity of mistletoes (*Viscum album* and other species),” Plants, v. 12, n. 14, pp. 2707, 2023. doi: 10.3390/plants12142707.

[20] K. Galler, E. Fronhlich, A. Kortgen, M. Bauer, J. Popp, U. Neugebauer, “Hepatic cirrhosis and recovery as reflected by Raman spectroscopy: information revealed by statistical analysis might lead to a prognostic biomarker”, Anal. Bioanal. Chem., v. 408, n. 28, pp. 8053-63, 2016. doi: 10.1007/s00216-016-9905-1.

[21] A. Wilk, K.W. Lakomy, Z. Setkowicz, K. Szary, R. Stachura, K.M. Marzec, M. Rugiel, J. Chwiej, “Elemental and biomolecular changes in serum as indicators of SPION action in the body: Analysis using TXRF, FTIR and Raman spectroscopy”, Sci. Reports, v. 15, pp. 45585, 2025. doi: 10.1038/s41598-025-29434-z.

Raman Spectroscopy for Discriminating Different Serological Titers in Canine Visceral Leishmaniasis

1st Maristela Possati Porto
Dept. Biomedical Engineering
Anhembi Morumbi University (UAM)
São José dos Campos/SP/ Brazil
maristelapossati@gmail.com

2nd Gislene F. da Silva Rocha Fournier
for Innovation, Technology and Education (CITÉ)
São José dos Campos/SP/Brazil
gislene.fatima@gmail.com

3rd Henrique Cunha Carvalho
Dept. Electronic Engineering (DAELN-CM – PPGEB-CT)
Federal University of Technology - Paraná (UTFPR)
Campo Mourão/ PR/ Brazil
hccarvalho@utfpr.edu.br

4th Landulfo Siveira Jr
Dept. Biomedical Engineering Anhembi Morumbi University (UAM) São José dos Campos/ SP/ Brazil
landulfo.silveira@gmail.com

5th Lívia Helena Moreira
Dept. Biomedical Engineering Anhembi Morumbi University (UAM) São José dos Campos/ SP/ Brazil
lh.medicinaveterinaria@gmail.com

***Abstract*— Canine Visceral Leishmaniasis (CVL) is a chronic parasitic disease associated with systemic biochemical alterations, reinforcing the need for complementary diagnostic approaches. This study evaluated the ability of Raman spectroscopy to identify molecular signatures associated with different serological titers in dogs naturally infected with *Leishmania* spp. Serum samples from seronegative dogs and seropositive dogs with titers of 1/64 or ≥1/128 were analyzed. Raman spectra were preprocessed by baseline correction, smoothing, normalization, and Principal Component Analysis (PCA). Significant spectral differences were observed in bands assigned to phenylalanine (1003 cm^{-1}), tyrosine (829-853 cm^{-1}), proteins and lipids (1319-1449 cm^{-1}), and amide I (1659 cm^{-1}), indicating biochemical alterations associated with the host immune response. PCA revealed discrimination among the evaluated groups, with statistically significant differences for PC1 (p=0.017) and PC2 (p=0.0094). These findings indicate that Raman spectroscopy is sensitive to biochemical changes associated with CVL and may serve as a complementary tool for distinguishing different serological response profiles.**

***Keywords*— blood serum, dogs, Leishmania, Raman spectroscopy, titers.**

I. Introduction

Leishmaniasis comprises a complex of infectious diseases caused by obligate intracellular parasites of the genus *Leishmania* spp., which exhibit tropism for the mononuclear phagocyte system (MPS). Active transmission occurs via the blood meal of female phlebotomine sandflies of the genus *Lutzomyia* spp. Between 20 and 22 species of this parasite are known to cause disease in both humans and animals [1]. It is estimated that there are between 700,000 and 1 million cases worldwide [2]. It is classified as a neglected tropical disease based on the DALY (disability-adjusted life-year) indicator for morbidity [3].

Clinical classification varies according to the causative parasite species, manifesting as Cutaneous Leishmaniasis (CL), Mucocutaneous Leishmaniasis (MCL), and Visceral Leishmaniasis (VL). Domestic dogs play a significant epidemiological role in urban centers, serving as the primary reservoirs that maintain the transmission cycle. Therefore, monitoring infection in these animals is essential to prevent the spread of the disease to new geographic areas and to improve surveillance and diagnostic systems [1].

Since the disease presents with non-pathognomonic clinical signs and often requires differential diagnosis, there is a need to develop complementary methods that optimize diagnostic specificity and sensitivity while maintaining speed and low cost, particularly to serve remote endemic regions. Current diagnostic methods include parasitological methods; serological tests such as the Indirect Immunofluorescence Assay (IFA) and Enzyme-Linked Immunosorbent Assay (ELISA); the Dual-Path Platform (DPP) rapid test; and molecular methods such as the Polymerase Chain Reaction (PCR), which is considered the gold standard. However, these techniques require specialized laboratories, trained professionals, and careful interpretation by a veterinarian, considering the disease stage. Furthermore, fluctuations in the host's immunological status can lead to false-positive or false-negative results [1].

In this context, Raman spectroscopy has attracted interest in the medical field because it enables rapid, non-destructive molecular characterization of biological samples with minimal sample preparation. These characteristics have favored its application in the study of infectious diseases [4]. The Raman spectroscopy technique determines molecular structures by analyzing the vibrational frequencies of light scattered by the sample. This phenomenon is based on the interaction of light with matter, encompassing both elastic scattering (Rayleigh), which preserves the incident light's wavelength, and inelastic scattering (Raman) [4]. In the latter, scattered photons exhibit altered frequencies due to specific molecular vibrations, acting as a chemical fingerprint of the sample. This feature enables the identification of compounds such as carbohydrates, proteins, amino acids, lipids, and nucleic acids, thereby facilitating the monitoring of tissue differentiation [5]. Consequently, biochemical alterations preceding pathological manifestations can be identified in real-time during active infections.

The ability to distinguish among different serological levels, rather than simply classifying animals as seropositive or seronegative, is clinically relevant because it allows greater precision in assessing disease severity, monitoring the immune response over time, and informing therapeutic and epidemiological decisions. Given the diagnostic limitations of CVL, this study aims to evaluate the sensitivity and applicability of Raman spectroscopy as an innovative tool for differentiating serological levels in CVL diagnosis.

II. Materials and Methods

A. Sample collection and serological diagnosis

Blood serum samples from mixed-breed dogs in the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Serra do Cipó, MG, Brazil, (Ethics: CEUA 09001/2021) were collected and shipped under refrigeration to the animal clinical pathology laboratory. After processing, the serum samples (n = 57) were divided into two aliquots: one was used for diagnostic screening by IFA to determine titers, and the other was designated for Raman spectroscopy analysis.

B. Sample preparation and Raman spectroscopy acquisition

The aliquot designated for Raman spectroscopy was sent to the Vibrational Spectroscopy Laboratory of the Center for Innovation, Technology and Education (CITÉ) in São José dos Campos, SP, Brazil, for analysis.

Blood serum samples were stored frozen at -20°C and thawed once at room temperature, ensuring batch homogeneity and minimizing artifacts from protein degradation. Handling procedures were standardized across the evaluated groups to facilitate processing.

The 80 µL aliquots of each sample were deposited onto an aluminum plate using a micropipette (model P200, Bio-Rad Laboratories Inc., Hercules, CA, USA) for subsequent spectral acquisition. Raman spectra were collected using a dispersive Raman spectrometer (model Dimension P-1, Lambda Solutions Inc., Waltham, MA, USA) with an 830 nm excitation laser operating at 350 mW power, and a 30 s exposure time. The instrument was calibrated within the 400–1800 cm^{-1} range, and spectra were obtained in triplicate. Before analysis, the data were digitally processed to remove background fluorescence and correct the baseline using a 7^{th}-degree polynomial algorithm. Subsequently, the signals were normalized using L1 normalization (dividing each spectrum by the sum of its absolute values) to eliminate intensity variations due to minor differences in focus or droplet volume.

Next, the mean spectra for each animal group (low and high titers) were calculated, enabling the identification of differences in Raman peak intensities and their correlation with the biochemical compositions described in the literature.

Spectral analyses were conducted using statistical tools, specifically Principal Component Analysis (PCA), applied to the full set of triplicate spectra acquired for each animal rather than to group- or animal-level averages. This multivariate approach was employed to reduce data dimensionality and identify the main clusters based on the biochemical profiles of canine serum samples. All statistical processing and the generation of score plots were performed using MATLAB software (v. 2007a; The MathWorks, Inc., Natick, MA, USA).

III. Results and Discussion

Serological analysis of dogs enrolled in the CVL survey in the Serra do Cipó-MG by IFA method, identified healthy (seronegative) animals (n = 38) and animals infected (seropositive) with CVL (n = 17); the latter were subdivided based on titer levels into low-titer (IFA 1/64; n = 13) and high-titer (IFA ≥ 1/128; n = 4) groups. It should be noted that the high-titer group was considerably smaller than the other groups, a sample imbalance that reflects the natural distribution of serological titers in the surveyed population and represents a limitation inherent to this preliminary study.

Variations in the biochemical profiles of the canine sera evidenced diagnostic differentiation among the evaluated groups. These molecular changes reflect the systemic nature of CVL, a disease in which the complex interaction between parasite and host disrupts the balance of the Th1 and Th2 immune responses. While the Th1 profile correlates with resistance mechanisms that activate leishmanicidal macrophages, the Th2 response induces intense humoral activity and hyperglobulinemia. This immunological imbalance directly alters the concentrations and structures of serum proteins, lipids, and amino acids, underpinning the potential of Raman spectroscopy as a complementary diagnostic tool [1].

Figure 1 shows the mean Raman spectra obtained for the negative group (control) and for dogs testing positive with low titers (1/64) and high titers (≥ 1/128). A progressive increase in spectral intensity is observed in seropositive animals, most notably in the high-titer group (≥ 1/128), suggesting an association between the serological response and molecular alterations present in the serum. The most significant differences occurred in regions attributed to serum proteins, aromatic amino acids, and lipids. The peak at 1003 cm^{-1}, attributed to phenylalanine [6], increased progressively in infected dogs. This pattern suggests an increase in the serum protein fraction, resulting from the antigenic stimulation observed in CVL, a condition frequently associated with polyclonal hypergammaglobulinemia [1].

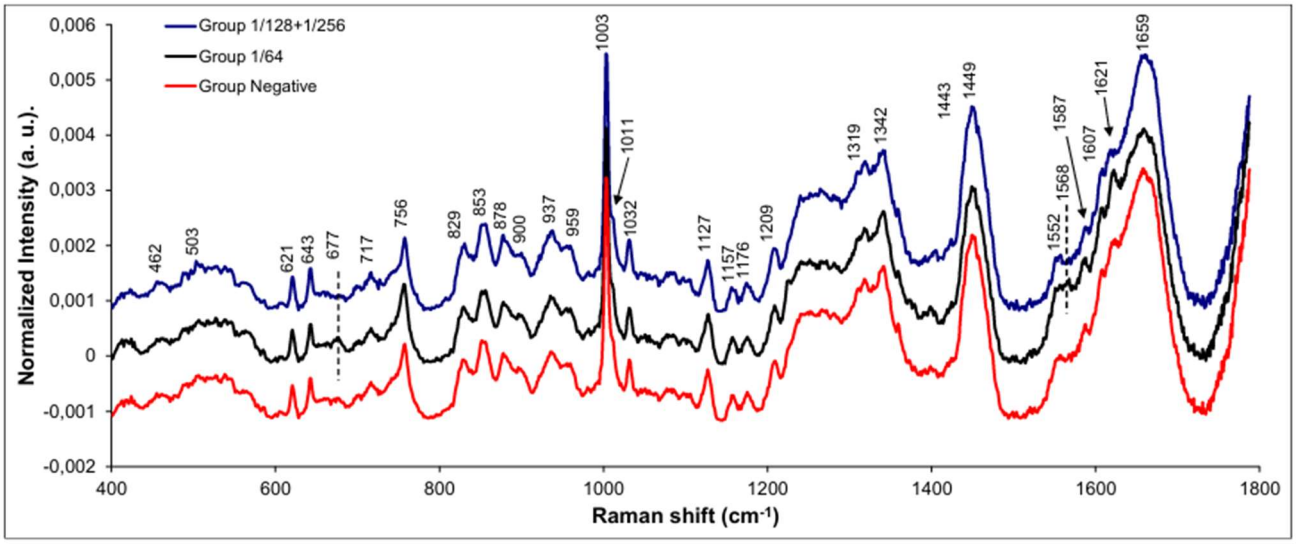


Fig. 1: Mean Raman spectra for negative, and positive groups with low titer (1/64) and high titer (≥1/128).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

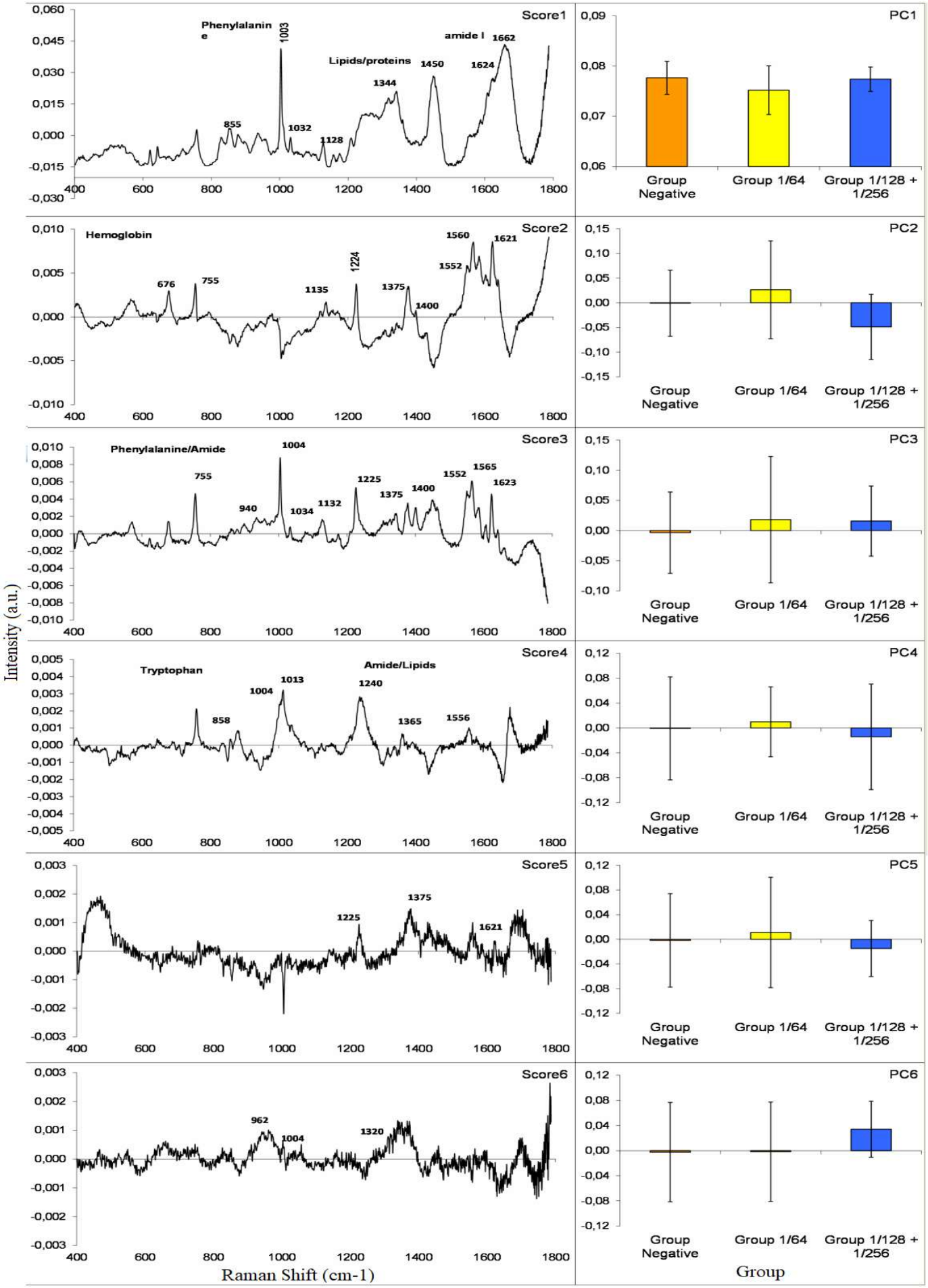


Fig. 2: PCA applied to the Raman spectra, with significance in PC1 ($p = 0.017$) and PC2 ($p = 0.0094$).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Alterations in the 829–853 cm^{-1} region, known as the tyrosine doublet [7], are frequently used as markers for conformational changes in proteins. The increased intensity of this doublet in seropositive animals suggests modifications consistent with intense immunological activation. Similarly, the bands between 1319 and 1342 cm^{-1} correspond to proteins and lipids [6] involved in metabolic responses to infection. The region between 1443 and 1449 cm^{-1}, attributed primarily to CH_2 and CH_3 vibrations of proteins and lipids [6], also showed higher intensity in infected animals, potentially reflecting changes resulting from the chronic inflammatory process.

Alterations in lipid metabolism have been reported in dogs with CVL, particularly in those with high parasite loads [8]. Another relevant result was observed in the 1659 cm^{-1} band, which predominantly corresponds to the amide I band [6]. The increase in this region reflects the contribution of serum proteins, particularly immunoglobulins, thereby mirroring the humoral activation characteristic of the infection [4].

Taken together, the alterations observed in the protein and lipid bands indicate that *Leishmania* sp. infection drives detectable biochemical changes in the serum of diseased dogs [4]. PCA confirmed the existence of spectral differences between the study groups. ANOVA of the scores showed statistical significance for PC1 (p = 0.017) and PC2 (p = 0.0094), whereas the remaining components yielded $p \geq 0.05$. These results indicate that the majority of the spectral variability associated with the infection is concentrated in the first two principal components, reinforcing that the differences between groups are not random but follow a biological pattern consistent with the pathophysiology of CVL (Fig. 2).

The distribution of the PCs suggests a separation between the negative and positive groups, as well as a tendency to distinguish between different serological titers. Although PCs 3 through 6 did not show statistical significance, possibly due to uneven sample sizes across groups, their spectral characteristics highlighted bands at 759, 855, 962, 1013, 1225, 1320, 1365, and 1375 cm^{-1}. These regions, associated with proteins, the amide group, carbohydrates, and phosphoproteins [6,9], suggest that multiple biochemical pathways are involved in the alterations observed in CVL.

Overall, the results indicate that the biochemical changes associated with CVL are reflected in spectral differences among the evaluated groups. However, these differences must be interpreted at the group level, as the limited number of animals and the uneven distribution of serological titers do not currently allow for the establishment of predictive capability for individual classification. For instance, supervised classification analyses would require a larger, more balanced dataset, as well as validation strategies that account for the origin of spectral replicates, thereby avoiding the use of spectra from the same animal for both model training and evaluation. Thus, the results presented should be considered exploratory and serve as a basis for future studies involving larger samples and individual-level validation.

IV – CONCLUSION

Raman spectroscopy identified spectral differences between seronegative and seropositive dogs, as well as among groups classified by serological titers. The progressive increase in the intensity of bands assigned to phenylalanine, proteins, and amide I in animals with higher titers demonstrates the technique's sensitivity to biochemical changes induced by the humoral response to CVL. Thus, Raman spectroscopy shows promise as a complementary tool for the biochemical characterization of CVL and the assessment of different serological profiles, offering additional support for the CVL diagnosis and monitoring.

ACKNOWLEDGEMENT

M. P. P. thanks PROSUP/CAPES for the Doctoral Scholarship (Finance Code 001), and Anhembi Morumbi University. L. S. Jr. thanks CNPq for the Research Productivity Grant (Process No. 314167/2021-8). L. S. Jr and L. H. M. would like to thank Instituto Ânima (IA) for the Research Grant. H. C. C. thanks Fundação Araucária (FA) & Superintendência de Ciências, Tecnologia e Ensino Superior (SETI) for the conference grant. The authors thank Parque Nacional da Serra do Cipó (PARNA) – MG, Brazil.

REFERENCES

[1] F. Morgado; A. Sudré; R. Porrozzi, “Manejo da Leishamniose canina. Orientações para educação em medicina veterinária”. IOC/Fiocruz, pp. 137. 1st. Ed. 2024.

[2] Organização Mundial de Saúde (OMS). “Manual de procedimientos para la vigilancia y el control de las Leishmaniasis em la región de las Americas”. 2nd Ed. Washington, D.C.: OMS, 2023.

[3] Brasil, Ministério da Saúde. “Manual de vigilância e controle de Leishmaniose visceral”. Dep. De Doenças Transmissíveis. Secretaria de Vigilância em saúde. 2nd Ed. Brasília, MS, pp.137. 2026.

[4] G. Acri, A. Falcone, C. Giannetto, E. Giudice, P. Guisepe, B. Testagrossa, L. Cicero, “Preliminary study for the identification of Leishmania infected dogs”. Sci Rep., vol.12, n. 1, p. 7489, 2022. doi: 10.1038/s41598-022-11525-w.

[5] M. P. Porto, L. Keidel, R.C. Menezes, L. Silveira Jr., L.H. Moreira, “Raman Spectroscopy in the diagnosis of Leishmaniasis in blood serum of dogs: Pilot Study”, In XXIX Brazilian Congress on Biomedical Engineering - Volume 1: Biomedical Robotics, Rehabilitation, Biomechanics, and Biomedical Signal Processing (CBEB 2024), IFMBE *Proceedings* 125, Springer, pp.464-470, 2025. doi: 10.1007/978-3-031-93646-3_48.

[6] Z. Movasaghi, S. Rehman, I. Urrehman, “Raman spectroscopy of biological tissues”. Applied spectroscopy reviews, v. 42, n. 5, pp. 493-541, 2007. doi: 10.1080/05704992070155153O.

[7] B. Hernández, Y-M Coic, G. Segei Kruglik, S. Sanchez – Cortes, M. Ghomi, “The relationship between the tyrosine residue 850-830 cm^{-1} Raman doublet intensity ratio and the aromatic side chain χ1 torsion angle”. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, v. 308, pp. 123681, 2024. doi: 10.1016/j.saa.2023.123681

[8] M. Gultekin, S. Pasa, K. Ural, C. Balikçi, G.S.E. Asici, G. Gultekin, “Oxidative status and lipid profile among dogs at different stages of visceral leishmaniasis”, Turkiye Parazitol Derg; v. 41, n. 940, pp.183-187, 2017. doi: 10.5152/tpd.2017.5398.

[9] A. Chandra; V. Kumar; U.C. Garnailk, R. Dada; I. Qamar; V.K. Goel; S. Agarwal, “Unveiling the Molecular Secrets: A Comprehensive Review of Raman Spectroscopy in Biological Research”, ACS Omega, vol. 9, n. 51, pp.50049-63, 2024, doi:10.1021/acsomega.4c00591.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Raman Spectroscopy Applied to the Serum Analysis of Healthy Individuals, Diabetic/Hypertensive Subjects and Hemodialysis Patients

Luciano Gonçalves da Nóbrega
Center for Innovation, Technology and Education - CITÉ
Anhembi Morumbi University - UAM
São José dos Campos - SP, Brazil
lgnobrega@gmail.com

Thaísa Roxo Rebelo
Center for Innovation, Technology and Education - CITÉ
Anhembi Morumbi University - UAM
São José dos Campos - SP, Brazil
thaisaroxo@hotmail.com

Nilton Mangueira Maciel
Federal University of Maranhão - UFMA
São Luís - MA, Brazil
nm.manfgueira@ufma.br

Landulfo Silveira Jr.
Center for Innovation, Technology and Education - CITÉ
Anhembi Morumbi University - UAM
São José dos Campos - SP, Brazil
landulfo.silveira@gmail.com

***Abstract* — Chronic kidney disease (CKD) is associated with the retention of uremic solutes, with urea, creatinine and glucose being essential markers in clinical follow-up. This study evaluated the potential of Raman spectroscopy (RS) for identifying molecular patterns and quantifying these analytes in human serum. A total of 224 samples were analyzed and distributed into three groups: control (CTRL), diabetic and/or hypertensive (DIAB/SAH), and chronic kidney disease patients on hemodialysis (CKD). Spectra were obtained and pre-processed for fluorescence removal and normalization. Partial least squares (PLS) regression models were constructed for correlation with standard laboratory methods (photocolorimetry). Difference spectra revealed that the DIAB/SAH group had a profile associated with carbohydrates and lipoproteins, whereas the CKD group showed bands related to uremic retention. PLS regression demonstrated high predictive accuracy. RS, combined with multivariate analysis, enabled differentiation of molecular signatures and accurate quantification of biomarkers, being a promising complementary tool for the biochemical monitoring of patients with CKD.**

***Index Terms* — Raman spectroscopy, human serum, creatinine, urea, glucose, PLS**

I. Introduction

Chronic kidney disease (CKD) is associated with the retention of uremic solutes and metabolic and cardiovascular pathologies. Diabetes mellitus (DIAB) and systemic arterial hypertension (SAH) constitute underlying diseases that are determinant for the development and progression of renal injury, while creatinine, urea, and glucose are routine biochemical markers employed in the clinical-laboratory follow-up of these patients [1].

Raman spectroscopy (RS) is a vibrational technique based on the inelastic scattering of light (the Raman effect), capable of recording the molecular signature of proteins, lipids, carbohydrates, and low molecular weight metabolites without re-quiring the use of reagents, allowing rapid and non-destructive analysis of biological fluids and tissues. In human serum, the technique has already been employed to estimate clinically relevant analytes, including urea, creatinine, glucose, proteins, lipids, and lipoprotein fractions, especially when associated with multivariate methods [2-4].

In this study, RS of human serum from three clinical groups were analyzed, with the objective of identifying spectral patterns related to laboratory alterations in urea, creatinine and glucose and evaluating the potential of the technique for biochemical monitoring of patients with CKD undergoing renal replacement therapy.

II. Materials and Methods

This was an analytical, quantitative, and cross-sectional study, approved by the Research Ethics Committee of Alcides Carneiro University Hospital, Federal University of Campina Grande, under CAAE no. 65400422.0.0000.5182.

A total of 224 adults of both sexes were included and distributed into three groups: 1 – Control (CTRL), composed of 57 healthy individuals with urea, creatinine, and glucose within the normal range; 2 – Diabetic and/or hypertensive (DIAB/SAH), composed of 82 individuals with DIAB and/or SAH, with normal urea and creatinine and without renal insufficiency; and 3 – chronic kidney disease (CKD) patients,

composed of 85 individuals on hemodialysis, with elevated serum urea and creatinine and some hyperglycemic individuals.

Serum urea, creatinine, and glucose measurements were performed at the Centrallab laboratory (Cajazeiras, PB, Brazil) by photocolorimetry using an automated analyzer (model AU680, Beckman Coulter, Brea, CA, USA). Subsequently, the samples were sent to the Vibrational Spectroscopy laboratory of CITÉ /UAM (São José dos Campos, SP, Brazil) for spectral acquisition.

Raman spectra were collected using a dispersive spectrometer (model Dimension P1, Lambda Solutions Inc., Waltham, MA, USA), with excitation at 830 nm, power of 350 mW, a 1200 lines/mm diffraction grating, a cooled CCD detector, 2 cm^{-1} resolution and a spectral range from 400 to 1800 cm^{-1}. Three independent spectra were obtained per sample, totaling 672 spectra. The data were subjected to cosmic ray removal, fluorescence correction by polynomial fitting (7th order), normalization by the area under the curve, and calculation of the mean spectrum per sample, in a procedure compatible with protocols previously applied to Raman analysis of human serum [2, 5].

Exploratory data analysis included mean spectra of each sample and difference spectra between the CTRL, DIAB/SAH and CKD groups.

A partial least squares (PLS) regression model was developed for creatinine, urea, and glucose using Chemoface software (https://www.ufla.br/chemoface) from the Federal University of Lavras – UFLA. Raman spectral data (224 samples) in the 400–1800 cm^{-1} range were used as the X matrix, and serum concentrations of creatinine, urea, and glucose (mg/dL) obtained via the reference laboratory method as the Y matrix. The number of latent variables (LVs) was defined by the minimum root mean square error of cross-validation (RMSEcv), resulting in 4 LVs for creatinine, 5 for urea and 11 for glucose. Performance was evaluated by the coefficient of determination of cross-validation (R^2cv), RMSEcv, and the correlation coefficient of cross-validation (r) between measured and predicted values, following the internal validation strategy used in previous studies on the quantification of serum analytes by Raman spectroscopy [3, 5].

III. Results and Discussion

The mean Raman spectra of the CTRL, DIAB/SAH and CKD groups (Fig. 1A) showed spectral features of human serum, with greater contribution from the serum protein matrix, composed of albumin and globulins and other plasma proteins. The bands attributed to phenylalanine (1003 cm^{-1}), one of the main spectral markers of serum proteins [6], to CH^2/CH^3 deformations of proteins, phospholipids, and lipoproteins (1449 cm^{-1}) [7-9] and to the Amide I band, associated with the C=O vibrational mode of peptide groups (1660 cm^{-1}) [3, 8], corresponded to the main spectral contributions of the groups analyzed.

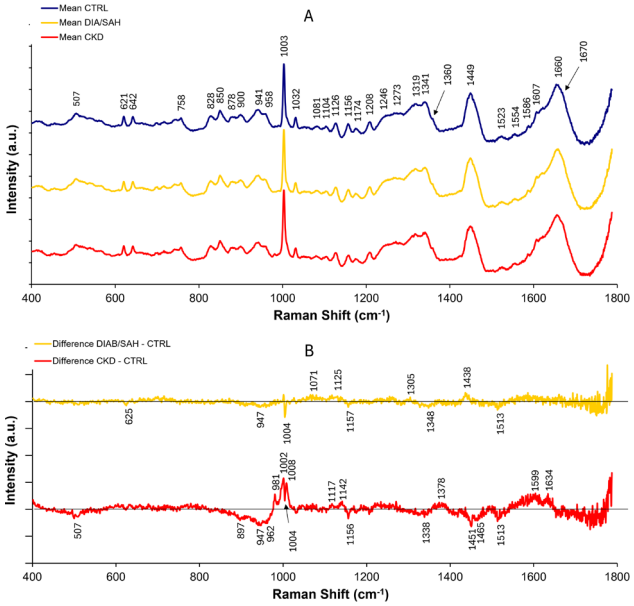


Fig. 1. Mean Raman spectra (A) and Raman difference spectra between the DIAB/SAH – CTRL and CKD – CTRL groups (B)

Bands attributed to aromatic amino acids, including tyrosine, phenylalanine, and carbohydrates (850 and 1032 cm^{-1}) [10-12]; C–C and C–N vibrational modes of proteins and serum metabolites (941 and 1126 cm^{-1}) [7, 13]; carotenoids (1156 cm^{-1}) [10, 11] and protein vibrations (1341 cm^{-1}) associated with nitrogenous compounds present in serum [8, 10] were also identified.

The spectral profile observed for the DIAB/SAH group suggested biochemical differences associated with glycemic and lipoprotein imbalance, characteristics directly related to the metabolic conditions present in diabetic and hypertensive individuals [8]. The contributions attributed to glycidic compounds and lipoproteins, with lower relative participation of serum proteins and carotenoids, showed agreement with alterations in the molecular composition of serum described in conditions of chronic hyperglycemia and metabolic dysfunction [8, 15].

In the Raman difference spectrum DIAB/SAH – CTRL (Fig. 1B), positive features showed contributions attributed to carbohydrates and glycidic compounds (1071 and 1125 cm^{-1}), compatible with C–O, C–O–C, and C–O–H vibrational modes of glucose and related metabolites [8]. Additional contributions were associated with lipids, phospholipids, and lipoproteins (1438 cm^{-1}) [14]. Negative features were identified in regions attributed to serum proteins and aromatic amino acids (947 and 1004 cm^{-1}) [7] and to carotenoids (1157 cm^{-1}) [14].

In the Raman difference spectrum CKD – CTRL (Fig. 1B), positive features occurred in regions attributed to phosphate vibrational modes (981 cm^{-1}) [16], spectral regions previously related to urea (1002 and 1008 cm^{-1}) [3, 13], compounds containing C–N and C–C bonds (1142 cm^{-1})

[17] and aromatic amino acids or protein structures (1599 cm^{-1}) [17]. Negative features were attributed to proteins and amino acids (947 cm-1) [10], carotenoids (1156 cm^{-1}) [15] and CH^2/CH^3 deformations of proteins, phospholipids, and lipoproteins (1451 cm^{-1}) [8]. The greater relative contribution of regions associated with urea, creatinine, phosphates, and nitrogenous compounds showed agreement with biochemical alterations described in patients with CKD undergoing renal replacement therapy, including retention of uremic solutes and disturbances of mineral metabolism resulting from reduced renal function [1, 2, 17].

Together, the mean spectra characterized the serum matrix common to the groups analyzed, whereas the difference spectra evidenced distinct biochemical patterns. The DIAB/SAH group showed a profile associated with glycidic and lipoprotein metabolism, whereas the CKD group showed a greater relationship with nitrogenous compounds, phosphates, and spectral markers compatible with uremic retention. Thus, Raman spectroscopy allowed differentiation of molecular signatures related to the clinical conditions evaluated.

Figure 2 presents the results obtained from the PLS regression model for creatinine (Fig. 2A), urea (Fig. 2B) and glucose (Fig. 2C).

For creatinine (Fig. 2A), the model presented a correlation coefficient r = 0.93 and R^2cv = 0.87. These values are highly consistent with previous studies, such as that of Almeida et al. (2016), which also obtained r = 0.93. The RMSEcv of 1.79 mg/dL is slightly lower (therefore, more accurate) than the error of 1.94 mg/dL. Giansante et al. (2022) showed the poorest correlation, with r = 0.64 and R2cv = 0.41, whereas Giana et al. (2025) exhibited r = 0.85 and R^2cv = 0.63.

In the urea assessment (Fig. 2B), performance was even more significant, with r = 0.97 and R^2cv = 0.98. The model's precision is evidenced by an RMSEcv = 10.3 mg/dL, result superior to those reported by Giana et al. (2025) (r = 0.89, R^2cv = 0.79) and Giansante et al. (2022) (r = 0.89, R^2cv = 0.8). The study by Almeida et al. (2016) showed the closest analytical correlation (r = 0.97) but yielded an RMSEcv of 17.6 mg/dL. Urea is the primary nitrogenous metabolite in serum and directly reflects the balance of catabolism.

Although showing slightly lower performance than creatinine and urea, PLS regression for glucose (Fig. 2C) presented a strong correlation between the methods, with r = 0.85, R^2cv = 0.72, and RMSEcv = 26 mg/dL. This study showed performance similar to that of the work of Silveira et al. (2017), in which r = 0.86, R^2cv = 0.74, and RMSEcv = 26.4 mg/dL were demonstrated. The model's performance in this research was superior to that reported by Giana et al. (2025) (r = 0.86 and R^2cv = 0.74) and Giansante et al. (2022) (r = 0.64 and R^2cv = 0.41).

The results obtained through the association between RS and PLS regression, presented in Fig. 2, demonstrate high potential for the quantification of serum metabolites indispensable for the evaluation of renal function. Analysis of the performance of the models for creatinine and urea reveals a high correlation between the values predicted by RS and those obtained by the reference laboratory methods.

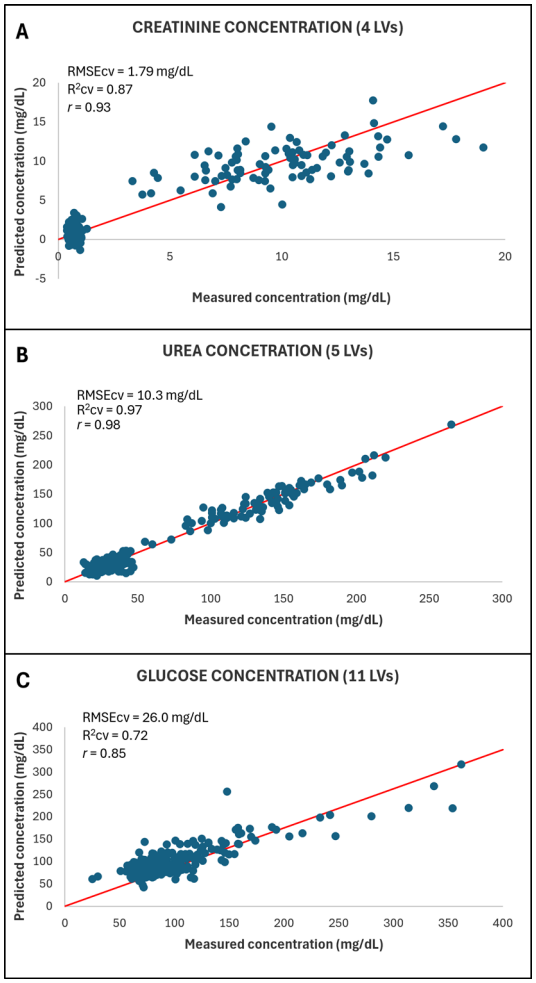


Fig. 2. Plot of the measured versus predicted concentration of the PLS regression and model performance (R^2cv = 0.87, r = 0.93 and RMSEcv = 1.79 mg/dL) for creatinine (A), (R^2cv = 0.97, r = 0.98 and RMSEcv = 10.3 mg/dL) urea (B), and (R^2cv = 0.72, r = 0.85 and RMSEcv = 26.0 mg/dL) glucose (C)

IV. Conclusion

Raman spectroscopy evidenced spectral differences among the CTRL, DIAB/SAH, and CKD groups, associated with distinct biochemical profiles of serum. Whereas the DIAB/SAH group showed contributions mainly related to glycidic, lipid, and lipoprotein compounds, the CKD group showed spectral contributions in regions associated with nitrogenous compounds, phosphates, and biomolecules compatible with alterations observed in chronic kidney disease. These findings demonstrated that Raman was capable of identifying differences in the molecular composition of serum among the clinical conditions evaluated.

When RS is integrated with multivariate spectral models such as PLS regression, it becomes a powerful tool for the biomolecular quantification of important markers in clinical

diagnosis and prognosis, as is the case of creatinine, urea, and glucose in the approach to CKD.

Because this is an exploratory and descriptive study, the results should be interpreted as preliminary evidence, reinforcing the potential of Raman spectroscopy as a complementary tool for biochemical investigation of human serum and encouraging future studies with a larger number of samples and independent clinical validation.

ACKNOWLEDGMENT

L. G. da No´brega acknowledges the support from Coordination for the Improvement of Higher Education Personnel (CAPES), Anhembi Morumbi University – UAM, Center for Innovation, Technology and Education – CITE´ and Federal University of Campina Grande – UFCG. L. Silveira Jr. ac-knowledges CNPq (National Council for Scientific and Technological Development) for the research productivity fellowship (Process No. 305857/2025-8) and the Anima Institute (AI) for the research scholarship.

REFERENCES

[1] Stevens, P. E., Ahmed, S. B., Carrero, J. J., Foster, B., Francis, A., Hall, R. K., Levin, A. (2024). KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International, 105(4), S117-S314.

[2] de Almeida, M. L., Saatkamp, C. J., Fernandes, A. B., Pinheiro, A. L. B., Silveira, L. (2016). Estimating the concentration of urea and creatinine in the human serum of normal and dialysis patients through Raman spectroscopy. Lasers in medical science, 31(7), 1415-1423.

[3] Giansante, S., Giana, H. E., Fernandes, A. B., Silveira, L. (2022). Analytical performance of Raman spectroscopy in assaying biochemical components in human serum. Lasers in Medical Science, 37(1), 287-298.

[4] Rohleder, D., Kiefer, W., Petrich, W. (2004). Quantitative analysis of serum and serum ultrafiltrate by means of Raman spectroscopy. Analyst, 129(10), 906-911.

[5] Giana, H. E., Souza, L. O., Silveira, L. (2025). Comparison of different Raman spectrometer models in the quantification of blood serum analytes. Lasers in Medical Science, 40(1), 143.

[6] da Silva, C. R., Carvalho, H. C., Lazo Osório, R. A., Fernandes, A. B., Silveira, L. (2024). Differences in whole blood before and after hemodialysis session of subjects with chronic kidney disease measured by Raman spectroscopy. Lasers in Medical Science, 39(1), 175.

[7] Borges, R. D. C. F., Navarro, R. S., Giana, H. E., Tavares, F. G., Fernandes, A. B., Silveira, L. (2015). Detecting alterations of glucose and lipid components in human serum by near-infrared Raman spectroscopy. Research on Biomedical Engineering, 31(2), 160-168.

[8] Silveira, L., Borges, R. D. C. F., Navarro, R. S., Giana, H. E., Zângaro, R. A., Pacheco, M. T. T., Fernandes, A. B. (2017). Quantifying glucose and lipid components in human serum by Raman spectroscopy and multivariate statistics. Lasers in medical science, 32(4), 787-795.

[9] Torre-Gutiérrez, L. D. L., Martínez-Zérega, B. E., Oseguera-Galindo, D. O., Aguilar-Lemarroy, A., Jave-Suárez, L. F., Torres-González, L. A., González-Solís, J. L. (2022). Breast cancer chemotherapy treatment monitoring based on serum sample Raman spectroscopy. Lasers in Medical Science, 37(9), 3649-3659.

[10] Goulart, A. C. C., Silveira Jr, L., Carvalho, H. C., Dorta, C. B., Pacheco, M. T. T., Zângaro, R. A. (2022). Diagnosing COVID-19 in human serum using Raman spectroscopy. Lasers in medical science, 37(4), 2217-2226.

[11] Lima, A. M. F., Daniel, C. R., Pacheco, M. T. T., de Brito, P. L., Silveira, L. (2022). Discrimination of leukemias and non-leukemic cancers in blood serum samples of children and adolescents using a Raman spectral model. Lasers in Medical Science, 38(1), 22.

[12] Ryzhikova, E., Kazakov, O., Halamkova, L., Celmins, D., Malone, P., Molho, E., Lednev, I. K. (2015). Raman spectroscopy of blood serum for Alzheimer's disease diagnostics: specificity relative to other types of dementia. Journal of biophotonics, 8(7), 584-596.

[13] Parachalil, D. R., Bruno, C., Bonnier, F., Blasco, H., Chourpa, I., McIntyre, J., Byrne, H. J. (2019). Raman spectroscopic screening of high and low molecular weight fractions of human serum. Analyst, 144(14), 4295-4311.

[14] Movasaghi, Z., Rehman, S., Rehman, I. U. (2007). Raman spectroscopy of biological tissues. Applied Spectroscopy Reviews, 42(5), 493-541.

[15] Singh, A. K., Karjee, H., Ghosh, S., Chatterjee, J., Roy, A. (2022). Spectropathologic endorsement of ocular carotenoids for early detection of diabetic retinopathy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 268, 120676.

[16] Moreira, L. P., Rocco, D. D. F. M., da Silva, A. G., Pacheco, M. T. T., Silveira Jr, L. (2020). Detecting creatine excreted in the urine of swimming athletes by means of Raman spectroscopy. Lasers in Medical Science, 35(2), 455-464.

[17] Delrue, C., Speeckaert, M. M. (2022). The potential applications of Raman spectroscopy in kidney diseases. Journal of Personalized Medicine, 12(10), 1644.

SAD-DERM: A Decision Support System for Automatic Diagnosis of Dermatological Lesions Using Convolutional Neural Networks

1st Abel Ramalho Galvão
Postgraduate Program in Systems and Products Engineering
Federal Institute of Education, Science and Technology of Bahia (IFBA)
Salvador, Bahia, Brazil
abel.ramalho18@gmail.com

2nd Eduardo Telmo Fonseca Santos
Postgraduate Program in Systems and Products Engineering
Federal Institute of Education, Science and Technology of Bahia (IFBA)
Salvador, Bahia, Brazil
eduardot@ifba.edu.br

Abstract*—Efficient screening and early diagnosis of dermatological lesions remain significant clinical challenges due to visual overlap between malignant and benign categories and the high demand on specialized care. This paper presents SAD-DERM, a multi-platform clinical decision support system based on convolutional neural networks for the eight-class classification of skin lesions defined by the International Skin Imaging Collaboration (ISIC) 2019 taxonomy (AK, BCC, BKL, DF, MEL, NV, SCC, VASC). Three architectures (ResNet50, DenseNet121, and EfficientNetV2) were trained with transfer learning, focal loss, and mixup augmentation on a lesion-grouped 70/15/15 split of the ISIC 2019 archive, with a frozen test partition of 3,468 images never used during training, model selection, or calibration. The final configuration combined DenseNet121 and EfficientNetV2 through a weighted soft-voting ensemble (0.68/0.32) with Test-Time Augmentation and temperature scaling calibrated on the validation set. On the frozen test partition, the ensemble achieved a macro one-vs-rest area under the receiver operating characteristic curve (AUC) of 0.913 (95% confidence interval: 0.906–0.920), top-1 accuracy of 66.9%, and top-2 accuracy of 84.3%. Per-class analysis with bootstrap confidence intervals, specificity, and confusion analysis identified squamous cell carcinoma and basal cell carcinoma as the primary source of residual error, with bidirectional confusion within the malignant branch. A cascaded variant (malignancy gate + subtype classifiers) was implemented and did not improve SCC classification, suggesting a representation- or data-level rather than an architectural limitation. A threshold-based triage mode prioritizing melanoma recall achieved 91.0% melanoma sensitivity (95% bootstrap CI: [88.8%, 93.2%]) at the cost of reduced specificity in other classes. We conclude that SAD-DERM provides a technically feasible research-grade triage pipeline whose clinical utility must be established through external validation and reader studies. The software is registered at the Brazilian National Institute of Industrial Property (INPI) under process 512026004653-0 and approved by the institutional ethics committee (CAAE: 95092725.7.0000.5031).

***Index Terms*—Artificial Intelligence, Convolutional Neural Networks, Dermatology, Clinical Decision Support, Skin Cancer, Test-Time Augmentation, Model Calibration.**

I. Introduction

The skin acts as a protective barrier and sensory interface, but its integrity is frequently compromised by a vast array of clinical manifestations [1]. Dermatoses represent a substantial burden on healthcare systems, significantly affecting patients' quality of life and psychological well-being [2]. Timely diagnosis is often hindered by the shortage of specialists, resulting in dangerous care bottlenecks [3].

Among these lesions, non-melanoma skin cancers (NMSC), such as Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC), are the most common neoplasms in humans [4], with cumulative ultraviolet (UV) exposure as the primary etiological factor [5], [6]. Conversely, cutaneous melanoma, which originates in melanocytes, presents lower incidence but high lethality when diagnosis is delayed [7]. The precise clinical distinction between melanoma and benign lesions remains one of dermatology's greatest challenges due to visual mimicry and morphological overlap [8].

Pre-malignant and benign lesions act as confounding factors [1]. Actinic keratosis is an intraepithelial dysplasia and direct precursor to SCC [9]. Dermatofibroma and benign keratosis frequently mimic malignant lesions, complicating visual diagnosis [10], [11]. Vascular anomalies add further challenges, particularly in darker skin phototypes [12]. Traditionally, the gold standard requires dermatoscopic evaluation followed by invasive biopsies, overloading the diagnostic queue [4].

Convolutional Neural Networks (CNNs) have emerged as disruptive technologies in medical image interpretation [13]. Recent work achieved 80% accuracy using DenseNet on HAM10000 [14], while others reached 87.72% with custom detection systems [15]; systematic reviews confirm AI precision is often equivalent or superior to human specialists [16]. To bridge this gap, this paper presents SAD-DERM, a diagnostic support system focused on multi-class classification of eight dermatological categories, integrating deep learning into a practical multi-platform application. Unlike prior work

reporting aggregate accuracy on a single hold-out partition, we adopt a three-way lesion-grouped split with a frozen test partition, report per-class metrics with bootstrap confidence intervals, and analyze the confusion structure that dominates residual error. The system potentially assists healthcare professionals subject to the validation requirements discussed in Section VII [1], [3].

II. Theoretical Background and Clinical Categories

A. Convolutional Neural Networks (CNNs)

CNNs constitute a specialized class of deep learning architectures designed specifically for processing grid-topology data, such as digital images [13]. In contrast to conventional neural networks, CNNs employ local connectivity and weight sharing, providing exceptional efficiency in extracting spatial features and morphological patterns [13]. The architectural structure is outlined by convolutional layers that apply parameterizable filters (kernels) to generate feature maps, activation functions like ReLU for non-linearity, pooling layers for spatial dimensionality reduction, and fully connected layers for the final Softmax classification [13].

B. Dermatological Lesions Investigated

To ensure clinical applicability, the system considers eight mutually exclusive diagnostic categories encompassing benign, pre-malignant, and malignant lesions [17], [18]:

- **Actinic Keratosis (AK):** A pre-malignant condition associated with chronic sun exposure, presenting as rough, scaly areas with the potential to progress to SCC [17].
- **Basal Cell Carcinoma (BCC):** A slow-growing but locally invasive malignant neoplasm, often presenting as a pearly papule or ulcerated lesion [18].
- **Benign Keratosis (BKL):** Non-malignant lesions with a keratotic appearance, frequently associated with benign epidermal alterations [18].
- **Dermatofibroma (DF):** A benign cutaneous lesion of fibrohistiocytic origin, typically observed as a firm nodule, frequently located on the lower limbs [19].
- **Melanoma (MEL):** A highly malignant neoplasm originating from melanocytes, characterized by asymmetry, irregular borders, and multiple shades [18].
- **Melanocytic Nevus (NV):** A benign proliferation of melanocytes, being one of the most frequent pigmented lesions in clinical practice [20].
- **Squamous Cell Carcinoma (SCC):** A malignant neoplasm derived from keratinocytes, often presenting as an infiltrated or ulcerated lesion with significant metastatic potential [17].
- **Vascular Lesion (VASC):** Benign alterations associated with the proliferation or dilation of cutaneous vascular structures, including angiomas [18].

III. Materials and Methods

The research was conducted based on a supervised learning workflow for multi-class classification using the TensorFlow/Keras ecosystem. The dataset was sourced from the ISIC 2019 archive, comprising 25,331 images spanning the eight diagnostic categories above. The complete pipeline, from data preparation to deployment, is summarized in Figure 1.

A. Data Splitting and Evaluation Protocol

The ISIC 2019 archive was partitioned using a lesion-grouped stratified split into 70% training (15,821 images), 15% validation (3,861 images), and 15% test (3,468 images). Grouping was enforced at the lesion level: all images sharing the same `lesion_id` were assigned to a single partition, preventing cross-partition leakage through multiple acquisitions of the same lesion. Images with no clinical grouping metadata (2,084) were excluded, and near-duplicate detection via perceptual hashing removed an additional 97 images before splitting. Class distribution across partitions was verified to be proportional to the full cohort within each class.

Patient-level grouping was attempted but proved infeasible: the ISIC 2019 distribution does not publish patient identifiers, HAM10000 does not expose them [27], and the ISIC Archive API v2 does not return `patient_id` for BCN20000 or MSK images (verified through direct API queries; the coverage report is part of the project's reproducibility artifacts). This constraint is discussed further in Section VII.

The test partition was frozen before any model training and used only for final evaluation, after all architectural, calibration, and threshold decisions had been finalized on the validation set. No metric reported in this paper is computed on the validation set.

B. Preprocessing and Data Augmentation

The images underwent spatial resizing and normalization specific to each pre-trained architecture (224×224 for DenseNet121 and ResNet50; native resolution and preprocessing for EfficientNetV2). To mitigate class imbalance and prevent overfitting, medically-safe augmentations were applied during training: random horizontal and vertical flips, rotations within $\pm20^\circ$, small brightness and contrast jitter, and random zoom. The mixup technique [26] was employed during fine-tuning, generating synthetic training samples through linear interpolations between pairs of images and their labels.

C. Architectures and Training Methodology

Three ImageNet-pretrained backbones were evaluated: ResNet50 [21], DenseNet121 [22], and EfficientNetV2 [23]. Each was adapted with a classification head consisting of Global Average Pooling, BatchNorm, Dropout (0.3–0.5), a Dense(256) layer with GELU activation, and a final Dense(8) layer with softmax activation.

Training occurred in two phases: (i) head warm-up with a frozen feature extractor (AdamW, lr $= 10^{-3}$), and (ii) fine-tuning with layer-wise learning-rate decay and cosine schedule. Due to severe class imbalance, the Focal Loss function was employed [24]:

$$L_{FL} = -\alpha(1 - p_t)^{\gamma} \log(p_t), \tag{1}$$

where p_t is the probability assigned to the correct class, α is a per-class weight inversely proportional to the effective number

of training samples, and $\gamma = 2.0$ controls the penalization of easily classified examples. The AdamW optimizer was used with weight decay regularization. Mixed-precision training (`mixed_float16`) was used for hardware acceleration on the Radeon Open Compute (ROCm) platform.

D. Ensemble, Calibration, and Test-Time Augmentation

The final inference configuration, denoted **M4**, combines four components:

1) **Ensemble.** DenseNet121 and EfficientNetV2 softmax outputs were combined via weighted soft voting, with weights fitted on the validation set (DenseNet: 0.68, EfficientNetV2: 0.32).
2) **Temperature scaling.** The ensemble logits were calibrated with a single temperature parameter fitted on the validation set via negative log-likelihood minimization [25].
3) **Test-Time Augmentation.** Input images were transformed with horizontal flip, vertical flip, $\pm 10^{\circ}$ rotation, and center crop 90%, and the resulting softmax outputs were averaged. TTA is a well-established technique for reducing predictive variance under distribution shift [28].
4) **Argmax decision.** Final class assignment by argmax over the calibrated ensemble probabilities.

Ablation modes M1 (DenseNet121 alone), M2 (EfficientNetV2 alone), and M3 (ensemble without TTA or calibration) were also evaluated to isolate the contribution of each component.

E. Clinical Threshold Mode (M5)

A fifth decision mode, denoted **M5**, operates on the same calibrated ensemble probabilities as M4 but replaces argmax with a cost-sensitive rule that prioritizes melanoma recall. M5 does not introduce a new model; it changes only the mapping from calibrated probabilities to predicted labels. Let $p_c(x)$ denote the calibrated probability for class c on image x. M5 selects the predicted label by maximizing the expected utility:

$$\hat{y}(x) = \arg\max_{c} \; u_c \cdot p_c(x), \tag{2}$$

where $u \in \mathbb{R}^8$ assigns $u_{\text{MEL}} = 10$ and $u_c = 1$ otherwise. The factor of ten reflects a deliberate clinical design choice that penalizes melanoma false negatives tenfold, consistent with the clinical priority of melanoma detection [29]. It is not derived from a formal elicitation study, which is out of scope for this work. A grid search over per-class scaling factors was performed on the validation set. Because the rule is a deterministic argmax over utility-weighted probabilities, it inherits the same ranking structure as M4 and differs only in the decision boundary. Consequently, M5 and M4 share identical AUC; they differ only in threshold-dependent metrics (precision, recall, specificity, F1).

F. Metrics and Statistical Analysis

For each mode, per-class precision, recall (sensitivity), specificity, F1, and one-vs-rest area under the receiver operating characteristic curve (AUC) were computed on the frozen test partition, together with macro and weighted averages. All point estimates are accompanied by 95% stratified bootstrap confidence intervals (CIs; 1,000 resamples, stratified by true class; $n = 1{,}000$ follows standard practice for bootstrap CI estimation in medical imaging). Calibration was assessed with two complementary definitions of Expected Calibration Error (ECE): top-label ECE, which measures reliability of the argmax prediction, and macro ECE, which averages per-class one-vs-rest reliability. Brier scores are reported for both definitions. Two complementary error metrics are reported for malignant classes: *label-exact recall* (e.g., true SCC classified as SCC) and *branch-level retention* (e.g., true SCC classified as any of AK, BCC, MEL, SCC), the latter reflecting the clinically actionable distinction between referral and non-referral.

TABLE I
INFERENCE MODES ON THE FROZEN TEST PARTITION (3,468 IMAGES). POINT ESTIMATES WITH 95% STRATIFIED BOOTSTRAP CONFIDENCE INTERVALS (CIS).

Mode	Macro AUC	Top-1	Top-2	Macro F1
M1 (DenseNet)	0.904 [0.894, 0.912]	64.94%	83.68%	51.48%
M2 (EffNetV2)	0.891 [0.882, 0.899]	63.26%	81.81%	46.68%
M3 (Ensemble)	0.909 [0.901, 0.916]	66.06%	83.65%	52.16%
M4 (Ens+TTA+Cal)	**0.913 [0.906, 0.920]**	**66.90%**	**84.31%**	**53.22%**

G. Hardware and Infrastructure

The experimental development was executed on a Linux environment. The primary hardware comprised an AMD Ryzen 7 8700G processor, 64 GB of RAM, and an AMD Radeon RX 6800 XT GPU, utilizing TensorFlow-ROCm for hardware acceleration within a Docker container for reproducibility.

H. System Architecture and Implementation

The SAD-DERM system is deployed as a four-layer client-server application. The backend uses Python and the FastAPI framework, exposing REST endpoints for authentication (JWT), patient management, and inference. A PostgreSQL 15 database persists users, diagnoses, and patient records. The client is built in Flutter, providing a single codebase for mobile, web, and desktop, with clinical history tracking and a side-by-side view of AI predictions and specialist hypotheses. The entire backend is containerized with Docker for reproducibility. We note that the system-level architecture is presented for completeness; quantitative throughput, latency, and usability evaluations are outside the scope of this work.

IV. RESULTS

A. Ablation of Inference Components

Table I presents the four pre-specified inference modes evaluated on the frozen test partition. Each mode isolates the incremental contribution of a specific component.

The ensemble (M3) improved macro AUC by +0.018 over the better single backbone (M1), and TTA combined with calibration (M4) added +0.004, yielding 0.913. We note that non-overlapping confidence intervals do not constitute a formal test

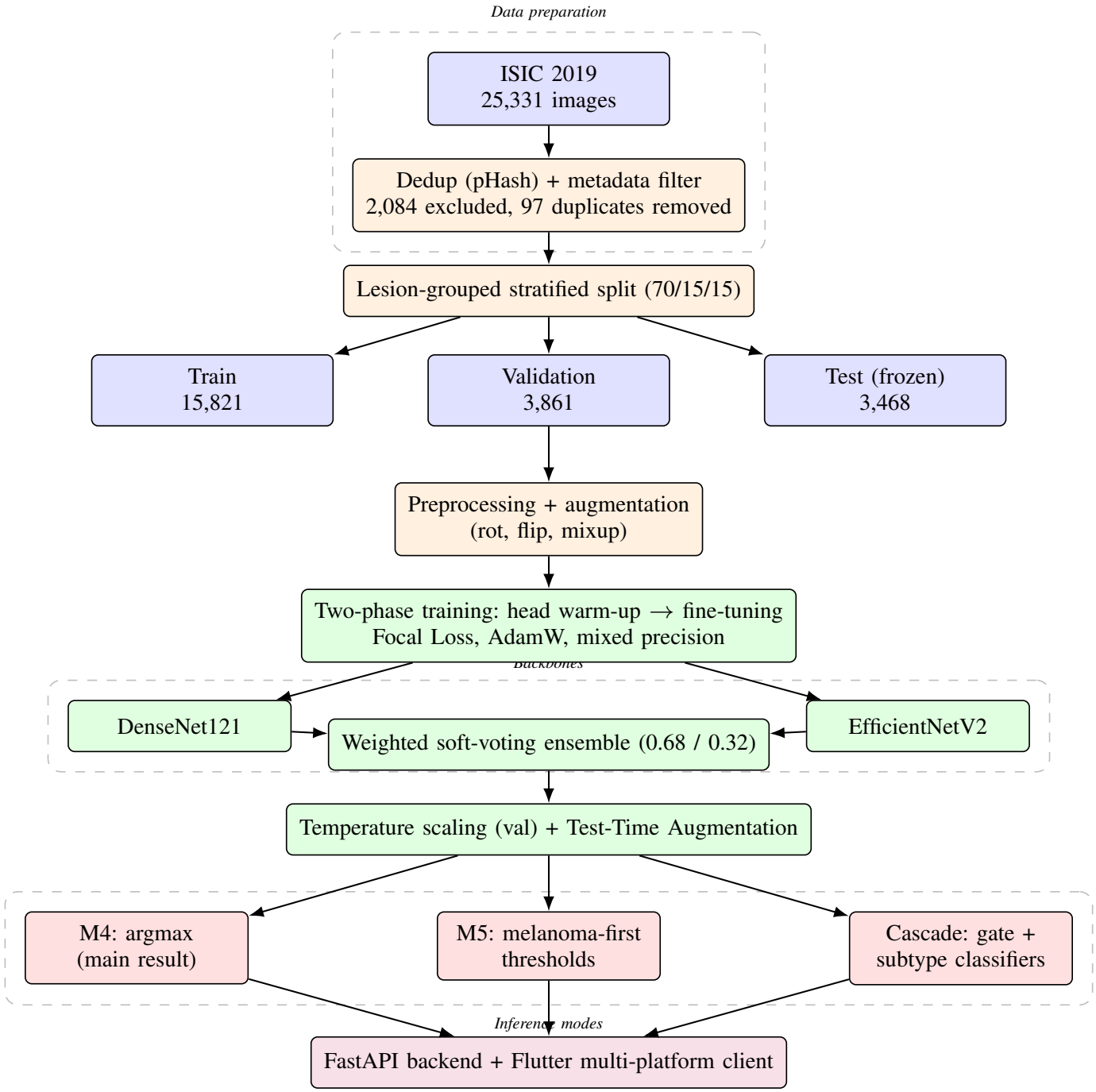


Fig. 1. End-to-end pipeline of SAD-DERM. Data flows from the ISIC 2019 archive through deduplication and lesion-grouped stratified splitting, followed by two-phase training of two backbones. The final inference configuration combines weighted soft-voting ensemble, temperature scaling, and Test-Time Augmentation. Three decision modes (M4 argmax, M5 melanoma-first thresholds, and the cascaded variant) share the same calibrated ensemble. Inference is served through a FastAPI backend to a Flutter multi-platform client.

of difference between correlated models; a paired bootstrap or DeLong-style test is left as future work.

B. Per-Class Performance

Table II reports per-class metrics for M4 on the frozen test partition. All point estimates are accompanied by 95% bootstrap CIs. The corresponding confusion matrix is shown in Figure 2.

Two observations are central (Figure 2). First, the historical concern from preliminary versions of this work — that apparent high recall in small classes might reflect leakage — is resolved by the frozen test partition: DF recall dropped from a leakage-suspect 1.00 to 47.9%, and VASC from 0.98

TABLE II
PER-CLASS FROZEN TEST METRICS FOR M4 (ENSEMBLE + TTA + CALIBRATION, ARGMAX DECISION).

Class	Supp.	Prec.	Recall	Spec.	F1	AUC
AK	110	23.4%	39.1%	95.8%	29.3%	0.920
BCC	497	68.6%	65.4%	95.0%	66.9%	0.950
BKL	362	42.9%	58.6%	90.9%	49.5%	0.867
DF	48	28.8%	47.9%	98.3%	35.9%	0.893
MEL	634	63.2%	50.6%	93.4%	56.2%	0.825
NV	1689	86.8%	78.0%	88.8%	82.2%	0.921
SCC	81	24.8%	43.2%	96.9%	31.5%	0.930
VASC	47	62.3%	91.5%	99.2%	74.1%	0.998

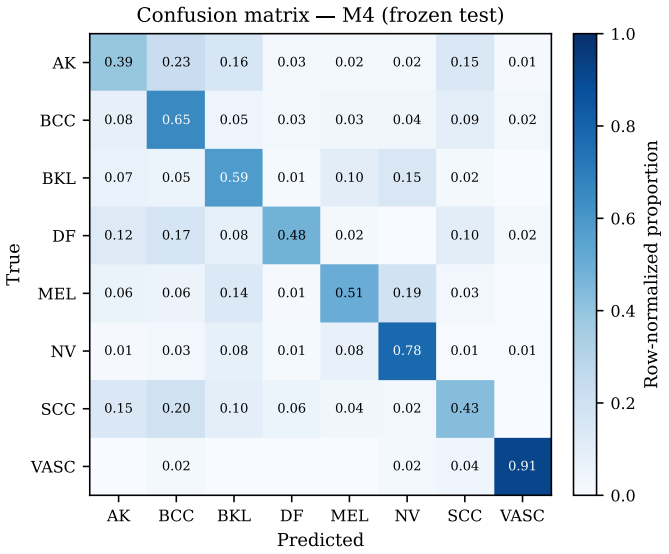


Fig. 2. Row-normalized confusion matrix for M4 on the frozen test partition. Diagonal cells correspond to label-exact recall. The off-diagonal mass in the SCC↔BCC block within the malignant branch is discussed in Section V.

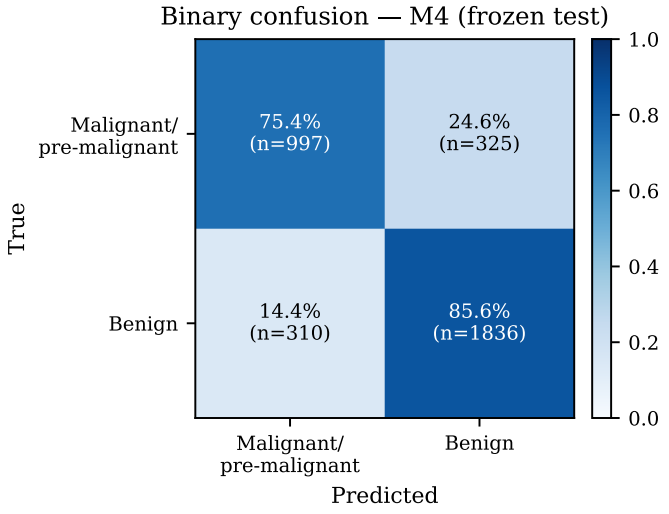


Fig. 3. Binary confusion matrix collapsing the eight classes into malignant/pre-malignant (AK, BCC, MEL, SCC) and benign (BKL, DF, NV, VASC). Values are row-normalized; raw counts in parentheses.

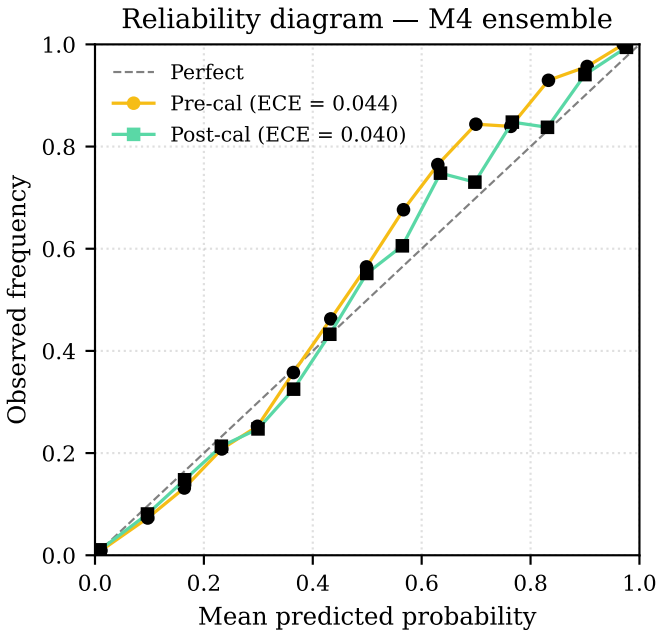


Fig. 4. Reliability diagram for the M4 ensemble before and after temperature scaling. Curves are macro-averaged across the eight one-vs-rest class probabilities. The ECE values shown in the legend (0.044 pre-calibration, 0.040 post-calibration) correspond to the macro-averaged definition. The top-label ECE reported in the text is defined differently and yields 0.076 → 0.042. The dashed diagonal indicates perfect calibration.

to 91.5%, both with wide confidence intervals reflecting their small supports. Second, the residual errors concentrate in specific class pairs: SCC (recall 43.2%), AK (39.1%), and DF (47.9%) exhibit the lowest label-exact recalls, whereas the dominant classes NV and BCC remain above 65% recall.

C. Calibration

Temperature scaling fitted on the validation set reduced the top-label Expected Calibration Error (ECE) on the frozen test from 0.076 to 0.042 (a 45% reduction). Under the macro-averaged per-class definition, the corresponding values were 0.044 and 0.040 (Figure 4). These two figures reflect different definitions of ECE (top-label versus macro-averaged) and are not in conflict. The multiclass Brier score moved from 0.457 to 0.452. Per-class ECE reductions were observed for six of the eight classes, with NV (the majority class) accounting for most of the residual miscalibration. The reduction in top-label ECE confirms that the ensemble probabilities, after calibration, are suitable inputs for downstream threshold-based decision modes.

V. ERROR ANALYSIS

A. Melanoma False Negatives

Under argmax decisions (M4), 313 of 634 melanoma images were not classified as MEL, corresponding to a label-exact melanoma recall of 50.6%. This figure is important context for the reviewer concern that "*three in ten melanomas would not be recovered*" in an earlier single-model configuration. In the

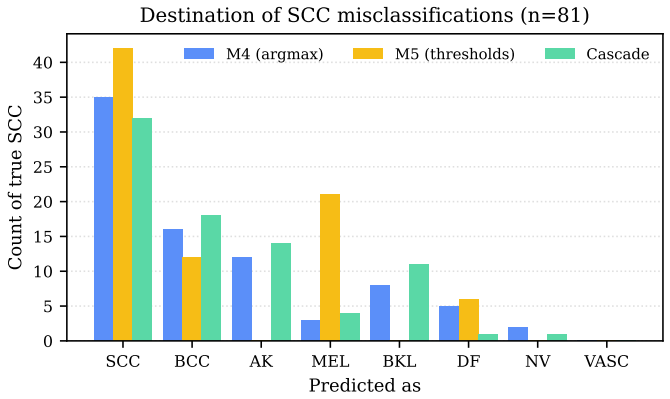


Fig. 5. Destination of SCC misclassifications on the frozen test partition (n=81 true SCC), comparing M4 (argmax), M5 (melanoma-first thresholds), and the cascaded variant. Malignant retention (SCC, BCC, MEL, AK) versus benign leakage (BKL, DF, NV, VASC) is reported in the text.

frozen test M4 configuration, the label-exact recall of 50.6% reflects the fact that many melanomas are classified as benign melanocytic nevi (NV), which is clinically the safest of the false-negative destinations.

To address the clinical priority of melanoma detection, we evaluated the threshold-based mode M5 (Section III-E). On the frozen test partition, M5 achieved a melanoma sensitivity of 91.0% (95% bootstrap CI: [88.8%, 93.2%]), reducing melanoma false negatives from 313 (M4) to 57 (M5) — an 82% reduction. This gain, however, comes at a substantial cost: melanoma specificity dropped from 93.4% (M4) to 49.2% (M5), and BCC false negatives increased from 172 to 239. M5 is therefore reported as a melanoma-first triage configuration, not as a general-purpose classifier.

Because M5 and M4 share identical calibrated probabilities and differ only in the decision rule, their one-vs-rest AUCs are identical; only threshold-dependent metrics change. This reinforces that M5 is an operational mode, not a different model.

B. SCC and BCC: Bidirectional Confusion

Analysis of the frozen test confusion matrix (Figure 2) reveals that SCC and BCC, the two most clinically consequential malignant classes after melanoma, form a bidirectional confusion zone within the malignant branch. The distribution of SCC misclassification destinations is shown in Figure 5, and the binary collapse is shown in Figure 3.

Bootstrap 95% CIs for malignant retention are [72.8%, 88.9%] (M4), [86.4%, 97.5%] (M5), and [76.5%, 91.4%] (cascade). The intervals overlap for all pairwise comparisons, so no difference in retention is statistically significant at the current test-set size (81 SCC images). We therefore avoid claiming superiority of any mode for SCC retention and report the point estimates with their uncertainty.

The confusion matrix also reveals that SCC misclassification is dominated by cross-class errors within the malignant branch: in M4, 16 SCCs were predicted as BCC, and 12 as AK. Both destinations result in a decision to refer the lesion for histopathological confirmation, since all three classes warrant biopsy under clinical guidelines. We emphasize that this does not imply clinical equivalence: SCC, BCC, and AK differ in prognosis, metastatic potential, and therapeutic planning, and the label-exact error remains clinically meaningful even when the referral decision is preserved. Only 15/81 (18.5%) of SCCs leaked into the benign branch, where triage would not refer the patient. In the cascade, benign leakage was 16.0%, and in the threshold-mode M5, 7.4%. These figures characterize SCC as primarily a *within-malignant* discrimination problem rather than a gate-failure problem.

Finally, the reverse direction is equally informative: 47 of 497 true BCC images (9.5%) were classified as SCC by M4, and 66 of 497 (13.3%) by the cascade. The bidirectional SCC↔BCC confusion, with substantial mass in both directions, suggests that the two classes share visual features that the current backbone cannot separate reliably — a hypothesis we tested directly with the cascaded architecture (Section V-C).

C. Cascaded Architecture Experiment

To test whether the SCC bottleneck could be resolved by architectural decomposition, we implemented a cascaded classifier consisting of (i) a binary malignancy gate (MEL, BCC, SCC, AK vs. BKL, NV, DF, VASC), (ii) a malignant subtype classifier over the four malignant classes, and (iii) a benign subtype classifier over the four benign classes. Each stage was trained with the same transfer-learning, focal-loss, and two-phase fine-tuning procedure as the flat models, and inference used soft-gating fusion of the three calibrated outputs.

The cascade failed the pre-registered SCC success criterion (39.5% vs. 58.2% target) while passing the criteria for melanoma recall (54.4% vs. 48.6% floor), BCC recall (68.8% vs. 63.4% floor), and macro AUC (0.919 vs. 0.903 floor). Critically, the benign leakage of SCC in the cascade (16.0%) was not statistically different from M4 (18.5%), and the bidirectional SCC↔BCC confusion *increased*: BCC→SCC errors rose from 9.5% (M4) to 13.3% (cascade). Removing BKL, NV, DF, and VASC from the subtype problem did not simplify it; instead, it isolated the hardest boundary (SCC vs. BCC) and forced the subtype classifier to make finer distinctions within the malignant branch that the backbone cannot support.

We interpret this result as consistent with a representation- or data-level limitation rather than a decision-boundary or architectural one, but we emphasize that this should be read as a hypothesis rather than a causal proof. The failure of a single cascaded configuration does not exclude the possibility that other causes — label ambiguity, class size, domain heterogeneity, resolution, or training strategy — contribute to the SCC bottleneck. The cascade result is reported as a negative finding with full metrics in the reproducibility artifacts; no further hyperparameter iteration was attempted after the pre-registered criteria were evaluated.

VI. Discussion

The SAD-DERM system demonstrates three contributions to the automated classification of dermatological lesions.

First, the ablation study (Table I) shows that the combination of ensemble learning, Test-Time Augmentation, and temperature scaling yields a macro AUC of 0.913 on a frozen test partition. The point estimates are consistent with a genuine contribution from each component, though we refrain from claiming statistical significance of the incremental gains without a paired test of differences.

Second, the per-class analysis with bootstrap CIs (Table II) reveals that the dramatic recalls reported in preliminary versions of this work (e.g., 1.00 for DF, 0.98 for VASC) were artifacts of leaky evaluation. The frozen test partition shows that small classes suffer wide uncertainty, and that SCC and AK remain the weakest categories by label-exact recall.

Third, the cascade experiment (Section V-C) provides a negative result: architectural decomposition does not resolve SCC misclassification and isolates rather than resolves the SCC↔BCC boundary. This has methodological implications: future work should report *branch-level retention* alongside *label-exact recall*, since only the latter reflects downstream clinical safety. In our test set, label-exact SCC recall was 43.2% (M4) and 39.5% (cascade), while branch-level retention was 81.5% and 84.0% respectively.

The threshold-based mode M5 illustrates the trade-off between sensitivity and specificity in a clinical triage setting. Prioritizing melanoma detection reduces melanoma false negatives from 313 to 57 on the frozen test partition, but at the cost of increasing BCC false negatives from 172 to 239 and reducing specificity across all non-melanoma classes. M5 is therefore positioned as a melanoma-first screening mode, applicable when the clinical cost of missing melanoma dominates the cost of unnecessary referrals, and not as a replacement for the argmax classifier. This trade-off is consistent with the broader literature on cost-sensitive classification in clinical decision support [29].

VII. Limitations

Our study has five limitations that should be considered when interpreting the results.

a) Lesion-level, not patient-level grouping.: The ISIC 2019 composite dataset does not publish patient identifiers in its official distribution. We attempted to recover `patient_id` by joining ISIC 2019 image IDs against the original HAM10000, BCN20000, and MSK metadata sources, and by querying the ISIC Archive API v2 directly. HAM10000 does not expose patient identifiers [27], and the ISIC Archive API v2 does not return `patient_id` for BCN20000 or MSK images. Consequently, our splits are grouped at the *lesion* level, with zero cross-partition lesion overlap. This is consistent with prior ISIC 2019 literature but does not exclude the possibility that images of the same patient appear in different partitions when the lesions are distinct; residual patient-level correlation between partitions cannot be fully excluded and is not quantified in this work. Patient-level validation would require the ISIC 2020 or ISIC 2024 cohorts, which publish patient identifiers but are binary (melanoma vs. rest) rather than eight-class.

b) No external validation.: All reported metrics are derived from a single frozen test partition of the ISIC 2019 dataset, which was curated from the same sources as the training data. Performance on images acquired under different conditions remains unknown. Prior work has shown that dermatological CAD systems degrade substantially under domain shift, and our macro AUC of 0.913 should not be interpreted as a predictor of clinical performance on external data.

c) No clinical reader study.: The current system was evaluated against histopathological ground-truth labels, not against dermatologist decisions. A proper clinical validation would compare model predictions to the decisions of board-certified dermatologists on the same images, ideally with a multi-reader multi-case design. Without such a study, our results characterize model-label agreement, not clinical utility.

d) Statistical power for small classes.: The frozen test set contains only 81 SCC, 48 DF, and 47 VASC images. Per-class differences between M4, M5, and the cascade do not reach statistical significance at 95% confidence for any of these classes, as bootstrap CIs overlap. Larger test partitions would be required to detect clinically meaningful differences in these categories. We report all per-class metrics with CIs and refrain from claiming superiority of any mode on small classes.

e) No formal paired test of model differences.: The reported confidence intervals are computed independently for each mode. Comparisons between modes (e.g., M1 vs. M4, or M4 vs. cascade) would ideally use a paired bootstrap of the metric differences, or a DeLong-style test for correlated AUCs. We report independent CIs and refrain from claiming statistical superiority of any single mode. A paired analysis is left as future work.

VIII. Conclusion

This study demonstrated the technical feasibility of a deep-learning-based decision support system for the automated screening of dermatological lesions across eight ISIC 2019 categories. On a frozen test partition of 3,468 images, the final configuration combining DenseNet121 and EfficientNetV2 with Test-Time Augmentation and temperature scaling achieved a macro one-vs-rest AUC of 0.913 (95% CI: 0.906–0.920), top-1 accuracy of 66.9%, and top-2 accuracy of 84.3%. A threshold-based melanoma-first triage mode achieved 91.0% melanoma sensitivity, at the cost of reduced specificity in non-melanoma classes. A cascaded architecture was implemented and produced a negative result, consistent with a representation- or data-level limitation rather than an architectural one.

The SAD-DERM system is consolidated as a research-grade support tool, integrating natively with clinical applications via a REST API and a Flutter multi-platform client. From the perspective of applied software engineering in health, the platform combines computational robustness with clinical usability. The research has its first functional version formally registered

at the National Institute of Industrial Property (INPI) under process number 512026004653-0, and is duly approved by the Ethics Committee (CAAE: 95092725.7.0000.5031). Future work will focus on (i) external validation on an independent dataset with patient identifiers, (ii) a reader study comparing model and dermatologist decisions, and (iii) refinement of the SCC/BCC discrimination through additional training data or auxiliary contrastive objectives.

ACKNOWLEDGMENT

The authors would like to thank the Postgraduate Program in Systems and Products Engineering (PPGESP) at the Federal Institute of Education, Science and Technology of Bahia (IFBA) for the academic support. This study was financed in part by the National Council for Scientific and Technological Development (CNPq) through the master's scholarship process 133298/2025-6.

REFERENCES

[1] M. C. G. Galhardo, F. F. Valle, and A. C. F. Valle, Dermatologia & doenças infecciosas: fundamentos e condutas. Rio de Janeiro: SciELO-Editora FIOCRUZ, 2022.

[2] N. C. V. Peruchi and M. C. Koelln, "Perfil epidemiológico de lesões dermatológicas em adultos na emergência hospitalar," Hospital de Clínicas de Porto Alegre, Porto Alegre, 2019.

[3] M. H. Imran, M. Shahid, M. Aazam, R. Sajid, M. A. Adnan, K. Naeem, and A. Ali, "DermNet: integrative CNN-ViT architecture for bias mitigation in dermatological diagnostics using advanced unsupervised lesion segmentation," Scientific Reports, vol. 16, p. 5291, Jan. 2026.

[4] E. T. de A. Fortunato et al., "Carcinoma basocelular: uma revisão de literatura," Brazilian Journal of Implantology and Health Sciences, vol. 6, no. 9, pp. 2223-2236, 2024.

[5] S. G. Paranhos et al., "Principais fatores de risco e intervenções terapêuticas no carcinoma espinocelular cutâneo: uma revisão sistemática," Lumen et Virtus, vol. 15, no. 42, pp. 6716-6730, 2024.

[6] A. J. Stratigos et al., "Diagnosis and treatment of cutaneous squamous cell carcinoma: European consensus-based interdisciplinary guideline," European Journal of Cancer, vol. 126, pp. 132-156, 2020.

[7] C. W. de M. Oliveira, A. F. B. de Sousa, K. L. Milfont, J. C. P. da Silva, I. T. A. Kanashiro, M. M. de Mauro, L. G. Teixeira, and D. de V. Bazílio, "Melanoma cutâneo: abordagens dermatológicas e oncológicas para diagnóstico, risco e tratamento," Contribuciones a las Ciencias Sociales, vol. 18, no. 1, pp. 1-24, Jan. 2025.

[8] G. G. Rezze, A. Leon, and J. Duprat, "Nevo displásico (nevo atípico)," Anais Brasileiros de Dermatologia, vol. 86, no. 1, pp. 143-156, 2011.

[9] C. P. H. Reinehr and R. M. Bakos, "Ceratoses actínicas: revisão dos aspectos clínicos, dermatoscópicos e terapêuticos," Anais Brasileiros de Dermatologia, vol. 94, no. 6, pp. 637-657, 2019.

[10] D. İ. Erdil, C. Leblebici, D. Erdil, V. Manav, V. A. Erdemir, and A. E. K. Aksu, "Dermatofibroma: análise clinicopatológica de 239 casos," Anais Brasileiros de Dermatologia, vol. 100, no. 1, pp. 150-155, 2025.

[11] K. Noiles and R. Vender, "Are all seborrheic keratoses benign? Review of the typical lesion and its variants," Journal of Cutaneous Medicine and Surgery, vol. 12, no. 5, pp. 203-210, 2008.

[12] T. Taye, S. D. Shah, and L. P. Lawley, "Diagnosis of vascular anomalies in patients with skin of color," The Journal of Clinical and Aesthetic Dermatology, vol. 17, no. 9, pp. S26-S31, 2024.

[13] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," Proceedings of the IEEE, vol. 109, no. 5, pp. 820–838, May 2021.

[14] D. A. Montagna, W. D. Parreira, A. M. da R. Fernandes, and R. L. S. Dazzi, "Classificação Automática para Auxílio no Diagnóstico de Lesões de Pele Usando Deep Convolucional Neural Network com Modelos de Transfer Learning," in COMPUTER ON THE BEACH, Itajaí, 2022, pp. 228-235.

[15] G. T. Fabian, B. Z. Dias, and G. A. Silva, "Sistema de detecção de câncer de pele com inteligência artificial," Centro Universitário Facens, Sorocaba, 2023.

[16] E. Macedo, "O uso da inteligência artificial no diagnóstico de câncer de pele: uma revisão sistemática," 2023.

[17] R. D. Azulay, D. R. Azulay, and L. Azulay-Abulafia, Dermatologia, 8th ed. Rio de Janeiro: Guanabara Koogan, 2022.

[18] W. Belda Junior, N. Di Chiacchio, and P. R. Criado, Tratado de Dermatologia, 3rd ed. São Paulo: Atheneu, 2018.

[19] M. M. C. A. Canelas, J. C. P. da S. Cardoso, P. F. Andrade, J. P. G. dos Reis, and O. Tellechea, "Histiocitomas fibrosos: revisão histopatológica de 95 casos," Anais Brasileiros de Dermatologia, vol. 85, no. 3, pp. 331-338, 2010.

[20] F. M. Paschoal, "Nevo melanocítico congênito," Anais Brasileiros de Dermatologia, vol. 77, no. 6, pp. 649-656, 2002.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.

[23] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 10096–10106.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.

[26] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.

[27] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, p. 180161, 2018.

[28] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," Neurocomputing, vol. 338, pp. 34–45, 2019.

[29] C. Elkan, "The foundations of cost-sensitive learning," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2001, pp. 973–978.

Carta de Resposta aos Revisores

CBEB 2026 — Congresso Brasileiro de Engenharia Biomédica

Prezado(a) Editor(a) e Comitê Científico,

Agradecemos os comentários criteriosos dos dois revisores sobre o manuscrito "Simultaneous Quantification of Macronutrients in Commercial Milk Using Raman Spectroscopy and Partial Least Squares Regression". Reconhecemos a relevância de todos os pontos levantados. Dado o formato de artigo curto do CBEB e o prazo estabelecido para a versão final (camera-ready), nem todas as revisões sugeridas puderam ser implementadas nesta submissão; abaixo, detalhamos o que foi efetivamente revisado e justificamos os pontos que permanecem como direção explícita para a versão estendida deste trabalho.

Revisor 1

> *"Validar pelo menos um subconjunto das amostras por métodos analíticos independentes e utilizar esses valores na calibração, pois um modelo não pode demonstrar precisão superior à confiabilidade de sua própria referência."*

Resposta dos autores: Concordamos que esta é uma limitação real do estudo. Como a condução de novos ensaios de referência (Kjeldahl, Gerber/Soxhlet, HPLC) está fora do escopo temporal desta submissão, optamos por tornar essa limitação explícita no texto, em vez de mitigá-la sem novos dados. Incluímos na Conclusão a seguinte ressalva: *"the reference values used for calibration corresponded to the concentrations declared on the nutritional labels, rather than to independent reference analytical methods; therefore, the results presented here reflect the Raman-PLS model's agreement with the declared information, within the regulatory tolerance, rather than a validation of absolute analytical accuracy."* A validação com métodos de referência independentes está registrada como etapa necessária para a versão estendida deste trabalho.

> *"Também considero necessário esclarecer o número efetivo de amostras independentes e evitar tratar as 10 aquisições Raman de uma mesma amostra como observações independentes na validação. O leave-one-out deveria excluir conjuntamente todas as medições pertencentes ao mesmo produto/lote, evitando data leakage."*

Resposta dos autores: Este é um ponto metodológico importante que optamos por não alterar nesta versão, pois uma reanálise completa da validação cruzada agrupada por amostra (grouped leave-one-out) no Chemoface exigiria reprocessamento dos dados brutos, o que não foi possível concluir dentro do prazo da versão final do congresso.

> *"Apesar dos Rcv entre 0,784 e 0,861 e dos RMSEcv reportados, não existe um conjunto externo de teste; portanto, afirmar excelente capacidade preditiva apenas com validação cruzada é prematuro."*

Resposta dos autores: Atendido parcialmente. Na Conclusão, a linguagem foi ajustada de "robust/excellent predictive performance" para *"satisfactory predictive performance under leave-one-out cross-validation"*, e a ausência de validação absoluta foi explicitada (ver resposta ao ponto anterior).

Revisor 2

Resumo (Abstract)

> *"Reescrever o final do resumo acerca dos métodos tradicionais de controle de qualidade de leites em pó. Ficou um pouco genérico, uma vez que o trabalho carece de uma seção de trabalhos relacionados. Adicionar também resultados quantitativos mais significativos."*

Justificativa dos autores: Atendido parcialmente. A frase final do Resumo foi reescrita para substituir a afirmação genérica anterior por um resultado quantitativo concreto do próprio estudo: *"15 of 15 samples complied for total protein and casein, and 14 of 15 for both total fat and lactose, the exceptions being attributed toMie scattering-driven optical attenuation and a possible lactose labeling discrepancy, respectively."* A expressão "robust predictive performance" também foi ajustada para "satisfactory predictive performance", em linha com a resposta ao Revisor 1.

Objetivos

> *"Adicionar outros modelos de regressão, bem como um teste estatístico (Wilcoxon) para quantificar melhor os métodos avaliados."*

Justificativa dos autores: A comparação com modelos de regressão alternativos (PCR, MLR) e a aplicação formal de testes estatísticos (Wilcoxon) são adições valiosas, mas exigiriam espaço adicional incompatível com o limite de páginas do formato de artigo curto do CBEB. Optamos por não incluir esses elementos nesta versão para preservar a profundidade das demais seções, e os registramos como direção prioritária de expansão metodológica na versão de periódico.

Trabalhos relacionados

> *"Carece de uma seção específica para posicionar o trabalho na literatura, bem como as principais contribuições metodológicas."*

Justificativa dos autores: Concordamos com a necessidade de uma seção dedicada. Por restrição de espaço do formato de artigo curto, não foi possível inserir uma subseção específica nesta versão; a contextualização da contribuição do trabalho permanece distribuída ao longo da Introdução.

Metodologia

> *"Muito curta, é necessário aumentar os detalhes desde a aquisição até a metodologia de comparação e avaliação, os autores têm espaço para tal. Remodelar a figura do fluxograma metodológico para ficar mais clara. Além disso, esta é apenas citada na figura e não é descrita."*

Justificativa dos autores: Atendido. A seção de Metodologia foi expandida com subtítulos que espelham cada etapa da Fig. 1 (fontes de amostra, aquisição espectral, pré-processamento, modelagem multivariada, validação e figuras de mérito), de modo que o fluxograma passa a ser descrito em texto corrido, e não apenas citado. Foi também adicionado um parágrafo detalhando explicitamente o cálculo de conformidade com a tolerância ANVISA (limite inferior = valor do rótulo × 0,80; amostra considerada conforme quando a média prevista somada ao RMSEcv é maior ou igual a esse limite), etapa que antes aparecia apenas implicitamente nos Resultados. A Fig. 1 foi redesenhada com fontes maiores, cores diferenciando cada etapa do fluxo, espaçamento revisado entre as caixas e uma nova etapa explícita de "Regulatory Compliance Assessment (ANVISA ±20%)", alinhando a figura ao texto metodológico ampliado.

Resultados

> *"Rever todas as unidades das métricas de R2 e RMSE."*

Resposta dos autores: Revisamos as ocorrências dessas métricas no manuscrito e confirmamos que r_{cv} é reportado corretamente como adimensional e $RMSE_{cv}$ consistentemente em g/100 mL em todas as figuras e tabelas nenhuma inconsistência de unidade foi identificada nesta verificação.

Agradecemos novamente a avaliação do comitê e permanecemos à disposição para esclarecimentos adicionais.

Atenciosamente,

Thaísa Roxo Rebelo, Luciano Gonçalves da Nóbrega, Landulfo Silveira Jr.
Center of Innovation, Technology and Education – CITÉ, Anhembi Morumbi University – UAM

Spectral Discrimination of Antimicrobial Resistance Profiles in Clinical Strains of *Escherichia coli* and *Klebsiella pneumoniae* Species via Raman Spectroscopy

1st Bruno César Almeida Costa
Universidade Anhembi Morumbi (UAM)
Institute of Health Sciences, Universidade Paulista (UNIP)
Faculdade Cristã da Cidade (FCC)
São José dos Campos, Brazil
profbrunobio@gmail.com

2nd Fernanda Sant´Ana de Siqueira e Oliveira
Institute of Health Sciences, Universidade Paulista (UNIP)
São José dos Campos, Brazil
fernanda.oliveira1@docente.unip.br

3rd Landulfo Silveira Jr.
Universidade Anhembi Morumbi (UAM)
Center for Innovation, Technology and Education (CITE)
São José dos Campos, Brazil
landulfo.silveira@gmail.com

Abstract*—The rise of bacterial resistance is a major threat to global public health due to recurrent therapeutic failures and the consequent increase in mortality. In this scenario, Raman spectroscopy stands out as a promising vibrational optical technique, as it enables the identification of cellular biochemical profiles. This study aimed to evaluate the spectral differences between clinical strains of Gram-negative bacteria (*Escherichia coli* and *Klebsiella pneumoniae*) presenting distinct susceptibility profiles (resistant and sensitive), correlating these spectral variations with the mechanisms of action of the antibiotics ciprofloxacin and ampicillin. A total of 121 clinical strains derived from biological samples (urine, cerebrospinal fluid and secretions) submitted for identification at the Municipal Laboratory of Jacareí (Jacareí, SP, Brazil) were used. Spectra were collected in quintuplicate using a dispersive Raman spectrometer (830 nm laser, 400 to 1800 cm^{-1} spectral range), totaling 605 spectra, which were preprocessed for background and noise removal. The results showed that, despite high intrinsic cellular biochemical similarity, the resistant strains exhibited variations in spectral intensity within similar ranges compared to the sensitive group. The results suggest that Raman spectroscopy proved to be an effective analytical tool to discriminate the antimicrobial resistance profile in clinical strains of *E. coli* and *K. pneumoniae.

Keywords—Antimicrobials, Gram-negative bacteria, microbial resistance, optical diagnosis, Raman spectroscopy.

I. INTRODUCTION

The emergence and rise of antimicrobial resistance over the years have consolidated as one of the primary threats to public health in the 21st century, impacting healthcare systems with high morbidity and mortality rates. This phenomenon directly affects medical practice by resulting in severe and difficult-to-manage infections, which prolong hospitalizations, extend treatments, and, consequently, lead to patient death [1]. The 2025 GLASS report [2] highlights that, between 2018 and 2023, antimicrobial resistance increased in 40% of the monitored pathogen-antibiotic combinations, with annual relative increases ranging from approximately 5% to 15%.

The overuse and misuse of antibiotics are currently considered a severe global public health crisis. Resistance reduces treatment efficacy, leading to therapeutic failures, increased mortality, and serious economic impacts [3]. This resistance occurs through a combination of biological mechanisms, including mutations in drug targets, the production of antibiotic-degrading enzymes (such as β-lactamases), or the overexpression of broad-specificity efflux pumps (e.g., AcrAB-TolC and MexAB-OprM), which extrude multiple antibiotic classes before they can reach their target [4].

Ampicillin is a broad-spectrum antibiotic belonging to the β-lactam class, frequently used in clinical practice and agriculture due to its low cost and good initial efficacy [3]. Its mechanism of action is based on interfering with cell wall biosynthesis in Gram-negative bacteria by cross-linking to penicillin-binding proteins (PBPs) during peptidoglycan synthesis, which inhibits essential reactions and weakens the bacterial structure [3]. Ampicillin resistance rates have become high, occurring through the acquisition of β-lactamase genes, mutations in target PBPs, and extrusion by efflux pumps [1, 4]. Furthermore, resistant strains adapt by reducing their metabolic growth and decreasing cell membrane permeability (e.g., porin downregulation), which physically restricts the amount of the compound penetrating the cell and prevents the drug from reaching optimal concentrations [5]. Ciprofloxacin, on the other hand, is a second-generation fluoroquinolone antibiotic introduced into clinical use in the 1980s and is structurally characterized by having a fluorine atom at the C6 position [4]. The drug primarily targets the essential bacterial enzymes DNA gyrase and topoisomerase IV, converting them into toxins that stabilize the cleavage complex, fragmenting the chromosome through DNA double-strand breaks and resulting in oxidative cell death pathways [7]. Despite its broad spectrum of activity against Gram-negative pathogens and good efficacy against Gram-positive ones, resistance to ciprofloxacin has grown considerably and encompasses multiple methods, such as mutations in the genes of the target enzymes themselves (such as gyrA and parC), overexpression of efflux pumps, and plasmid-mediated quinolone resistance (PMQR) [7]. The pathway through PMQR elements, in particular, includes both the production of Qnr family proteins (which physically protect gyrase) and variant enzymes such as AAC(6')-Ib-cr, which directly acetylates and inactivates the ciprofloxacin molecule, creating additional barriers to clinical treatment [8].

The ability of these pathogens to resist therapeutic treatments and survive in hostile environments is associated, among other factors, with the composition and architecture of their cell envelope [9]. The cell envelope of Gram-negative bacteria, such as *Escherichia coli* and *Klebsiella pneumoniae*, exhibits a structural composition organized in multiple layers (outer membrane, cell wall, and inner membrane) that protect the cell from hostile environments. The envelope consists of a system of molecular barriers and is primarily supported by carbohydrate complexes. In Gram-negative bacteria, this structure is formed by a thin peptidoglycan layer protected by an asymmetric outer

membrane rich in lipopolysaccharide (LPS) [9]. Hypervirulent pathogens, such as *K. pneumoniae*, possess capsular polysaccharide (CPS) as a defense mechanism, characterized as a glycan network that functions simultaneously as a shield and an attractant against host antimicrobial peptides [10]. Regarding proteins, their activity is related to controlling the chemical interface and metabolic flux. Integral membrane proteins, such as porins (beta-barrel channels like OmpK35 and OmpK36 in *K. pneumoniae*), have their expression suppressed under stress to restrict antibiotic influx, just as transmembrane protein complexes act as efflux pumps (such as the AcrAB and AcrAB-TolC systems). These biomechanical structures identify intracellular drugs and actively eject them into the extracellular medium, which is characteristic of multidrug resistance phenotypes [3,4]. Additionally, the selective permeability and combined efficacy of glycoprotein networks are associated with lipid diversity [9]. In fact, adaptive mechanisms allow microorganisms such as *E. coli* to reduce cell membrane permeability and restrict drug accumulation by downregulating dedicated transmembrane networks used for antibiotic passage, such as the OmpF porin system [5]. It is known that during the infectious process, the envelope undergoes active chemical remodeling; the bioconjugation of molecules such as palmitate or aminoarabinose directly to the lipid A moiety of LPS is the primary mechanism to neutralize the affinity for last-line antibiotics, such as colistin [10].

In Gram-negative bacteria, nucleic acids (DNA and RNA) are structured by continuous phosphodiester bonds that support the arrangement of nitrogenous bases. The DNA compartment encompasses the main chromosome and a dynamic genome, consisting of mobile genetic elements such as plasmids, transposons, and integrons, whereas the RNA fraction is predominantly organized as ribosomal RNA, highlighting the conserved 16S and 23S domains. The integrity and regulation of this genetic material depend on specific enzymes, such as DNA gyrase, topoisomerase IV, and RNA polymerase, which constitute direct therapeutic targets for clinical antimicrobians, such as fluoroquinolones and rifampicin. In addition to their physiological and structural role, intrinsic characteristics of these biopolymers, such as unmethylated CpG motifs and single- or double-stranded RNA chains, act as pathogen-associated molecular patterns (PAMPs). These components are recognized by intracellular receptors of the host's innate immune system, such as TLR9, acting as the primary triggers of the inflammatory response during the infectious process [9, 10].

Raman spectroscopy is a vibrational technique based on measuring the inelastic scattering of light from a laser by molecules and has stood out for being non-destructive and label-free, effective for chemotaxonomic identification and classification of a wide variety of microorganisms, including bacteria and fungi [11]. By measuring the vibrational energies of fundamental cellular constituents, such as proteins, lipids, carbohydrates, and nucleic acids, the technique can provide a biochemical and structural "spectral signature" for microorganisms [12-14]. Witkowska et al. (2016) [12] evaluated both isolated fungal cells and the main constituents of the cell wall of these fungi (chitin, galactomannan, and β-1,3-glucan). Furthermore, the authors successfully applied the technique to identify the fungus Trichophyton rubrum directly in clinical samples of infected human skin. Additionally, to handle data complexity and perform accurate taxonomic classification, Raman spectra are frequently combined with multivariate statistical analysis and machine learning methods, such as principal component analysis (PCA) and support vector machines (SVM) [11].

In the field of bacteriology, Raman spectroscopy has proven to be effective, enabling the identification and differentiation of bacterial species. Oliveira et al. (2012) [11] used the technique to discriminate seven species of pathogenic bacteria commonly causing urinary tract infections, being capable of identifying spectral differences linked to biochemical content. Subsequently, Oliveira et al. (2020) [13] deepened this study, by expanding the analysis to nine clinical species (three Gram-positive and six Gram-negative) and establishing correlations between statistically significant spectral peaks (ANOVA) and cellular biochemical components, such as peptidoglycans, proteins, and nucleic acids. Recently, the study by Oliveira et al. (2016) [15] evolved the application of the technique from a taxonomic focus to analyzing the problem of antimicrobial resistance, investigating four Gram-negative species (*E. coli*, *K. pneumoniae*, *Proteus mirabilis*, and *Enterobacter cloacae*); by analyzing spectral variations in regions of proteins, lipids, saccharides, and nucleotides. The study successfully discriminated resistant from sensitive strains, achieving a remarkable accuracy of 99.1% independently of the species. In another study, Singh et al. (2023) [16] evaluated sensitive and resistant clinical isolates of *E. coli*, *Acinetobacter baumannii*, and *Enterobacter* spp. by analyzing spectra collected with and without exposure to antibiotics with different mechanisms of action (ciprofloxacin, gentamicin, meropenem, and nitrofurantoin). The results demonstrated that sensitive strains exposed to ciprofloxacin and meropenem exhibited a reduction in the intensity of bands associated with DNA and proteins, accompanied by an increase in lipid bands; however, treatment with gentamicin and nitrofurantoin generated the opposite effect in these same sensitive strains, increasing the nucleic acid signal and reducing those of proteins and lipids.

Considering the relevance of identifying resistance profiles to guide antibiotic therapy, this study aims to evaluate the average spectral differences between clinical strains of Gram-negative bacteria (*E. coli* and *K. pneumoniae*) presenting distinct susceptibility profiles (resistant and sensitive). Additionally, it seeks to compare the obtained results with data already described in the relevant literature associated with the mechanism of action of the antibiotics.

II. Material and Methods

A. Samples

A total of 121 clinical strains, obtained from biological samples (such as urine, cerebrospinal fluid, and secretions), were sent to the Municipal Laboratory of Jacareí (Jacareí, SP, Brazil) for analysis. These bacteria were morphotinctorially identified using Gram staining and standardized biochemical tests for Gram-negative bacilli: EPM MILI (Probac do Brasil, São Paulo, Brazil; MS Registration: 10104030068) for *E. coli* strains and API 20E

(BioMérieux Brasil S.A., Rio de Janeiro, Brazil; MS Registration: 10158120563) for *K. pneumoniae* strains. Antimicrobial susceptibility testing (AST) was performed following two main protocols. First, the disk diffusion method (Kirby-Bauer) was conducted by exposing the bacterial strains cultured on Mueller-Hinton agar to filter paper disks impregnated with antimicrobial agents, including the antibiotics ampicillin and ciprofloxacin. After incubation at 35 ± 2 °C for a period of 16 to 18 h, susceptibility or resistance was characterized by the presence or absence of growth inhibition zones around the disks, verifying that strains exhibiting microbial growth up to the area of direct drug exposure were classified as resistant. Second, automated broth microdilution, using the MicroScan system (Beckman Coulter, São Paulo, SP, Brazil), was employed to determine the minimum inhibitory concentration (MIC) and to detect extended-spectrum β-lactamase (ESBL)-producing bacteria. The AST protocols and breakpoint interpretations were guided by established diagnostic standards, including the guidelines of the American Society for Microbiology (ASM), the Manual of Clinical Microbiology, and the criteria established by the Clinical and Laboratory Standards Institute (CLSI) or the Brazilian Committee on Antimicrobial Susceptibility Testing (BrCAST)/EUCAST.

For the Raman spectroscopy analysis, the bacterial strains were cultured on Mueller-Hinton agar, as this medium has been shown to cause less interference in bacterial Raman spectra, according to [11]. After the standard incubation of 16 to 18 h at 35 ± 2 °C, the plates were safely stored in biohazard boxes to ensure sample integrity and safety during transport. Subsequently, the plates were forwarded for spectral data collection. At the time of analysis, the biomass was carefully collected from the colonies present on the agar surface using a disposable 10-µL loop to avoid interferences, and the material was deposited onto a sheet of aluminum foil. Spectra were collected in quintuplicate, resulting in 605 measurements. After spectral acquisition, all samples were discarded following standard laboratory biosafety protocols. Table 1 describes the different bacterial species used in this study, which were selected due to their relevance in infectious conditions, along with the number of strains to represent the biological variation among the species.

TABLE I
SUSCEPTIBILITY AND RESISTANCE PROFILE OF THE STUDIED BACTERIAL STRAINS

Group		*Species*	
		E. coli	*K. pneumoniae*
Susceptible		30 strains	28 strains
Resistant	Ciprofloxacin	20 strains	17 strains
	Ampicillin	26 strains	–[a]

[a]"–" indicates strains not evaluated for the antibiotic in question.

B. Spectral Acquisition

Spectral data acquisition was performed using a dispersive Raman spectrometer (model Dimension P-1 HR, Lambda Solutions Inc., Waltham, MA, USA), equipped with an 830-nm diode laser as the excitation source. The radiation was delivered to the sample through a fiber-optic cable coupled to a probe (model Vector Probe, Lambda Solutions Inc.) with a focal length of 10 mm, generating a beam on the target with an approximate spatial resolution of 160 µm and ensuring an excitation of 300 mW. The inelastically scattered light was collected by the same probe and directed the detection system, which consisted of a spectrograph and a CCD camera detector. Spectra were collected covering the spectral range from 400 to 1800 cm^{-1} with a spectral resolution of approximately 4 cm^{-1}. To determine reproducibility, each bacterial sample was analyzed in five independent replicates. Acquisition parameters were fixed at an exposure time of 5 s and 10 accumulations per point (total time of 50 s per spectrum). System calibration was checked daily, and readings showing background saturation were discarded.

C. Preprocessing

Following the acquisition, the spectra were subjected to a preprocessing routine to remove equipment-inherent noise and isolate authentic vibrational information, eliminating interferences common in biological samples. Initially, a correction was performed for differences in the wavelength-dependent sensitivity of the spectrometer's optical components (probe, mirrors, diffraction grating, lenses, and CCD camera). Subsequently, the subtraction of the camera's dark signal, correction of potential detector reading defects (hot/cold pixels), and manual removal of spurious cosmic ray spikes were carried out. To eliminate background interference, the subtraction of a 7th-order polynomial fitted over the useful spectral range was applied. The selection of spectral regions of biochemical interest was determined based on domain-specific literature, corroborated by visual inspection of the spectral averages of the analyzed groups.

D. PLS Regression

In order to develop a predictive diagnostic model, the spectra were subjected to partial least squares discriminant analysis (PLS-DA). PLS-DA is a supervised method that establishes a linear relationship between the spectra (X) and the resistance profiles (Y), optimizing the covariance between them. The data were tabulated in spreadsheets and projected onto latent variables (LVs), allowing the identification of the spectral characteristics that discriminate the bacterial phenotypes. PLS-DA also allowed the construction of confusion matrices and the determination of the model's accuracy, sensitivity, and specificity for the differentiation of susceptible and resistant groups of *E. coli* and *K. pneumoniae*. The PLS-DA analyses were performed using the Chemoface software (Universidade Federal de Lavras, Lavras, MG, Brazil), as described by Nunes et al., 2012.

III. Results

A. Mean Raman Spectra of the Sensitive and Resistant Strains of Selected Bacteria Specie

Fig. 1 presents the mean Raman spectra of *E. coli* and *K. pneumoniae* groups in the range from 400 to 1800 cm⁻¹.

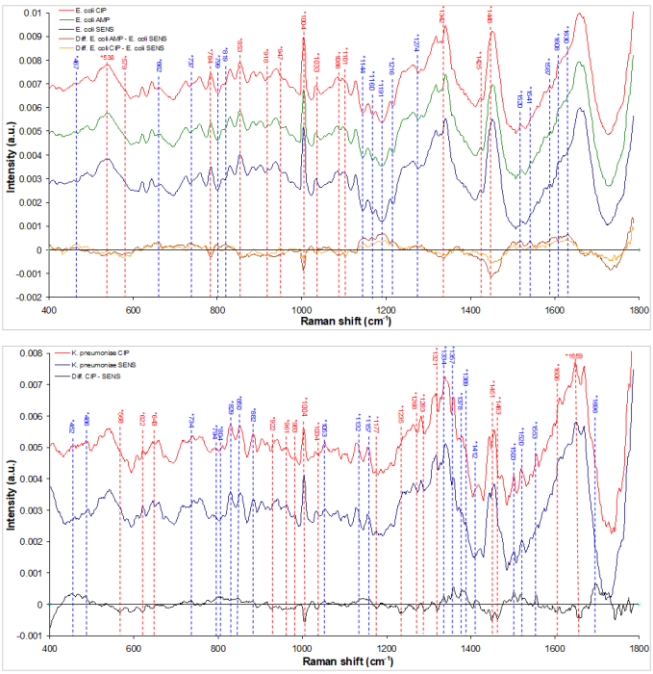


Fig. 1. Mean Raman spectra of *E. coli* and *K. pneumoniae* groups in the range from 400 to 1800 cm⁻¹.

Initially, a similarity was observed between the spectra of both susceptible and resistant strains of each species, which was expected and is justified by the common biochemical composition of this cell type, predominantly consisting of proteins, lipids, nucleic acids, and carbohydrates [5,6,8]. Despite the similarity between the spectra, it was possible to observe differences in the peak intensity pattern.

In Fig. 1, the spectra of the resistant *E. coli* groups (to both ampicillin and ciprofloxacin) showed higher intensity in the positive bands with peaks in the range of ~1520–1630 cm⁻¹, which can be assigned to proteins and amino acids, as well as the peaks at 1216 and 1274 cm⁻¹. Nucleotide bases were also observed at 737, 799, and 819 cm⁻¹, in addition to regions with collective contributions from multiple cellular constituents (~1144–1191 cm⁻¹). In contrast, the susceptible *E. coli* group showed higher intensity in the negative bands with peaks that can be assigned to LPS, such as 579 cm⁻¹, and the peaks in the range of ~853–1101 cm⁻¹, except for the peaks at 1004 and 1033 cm⁻¹, which correspond to aromatic amino acids (phenylalanine, tyrosine, tryptophan).

In the ciprofloxacin-resistant *K. pneumoniae* group, the spectra showed higher intensity in the positive band at peaks that can be assigned to proteins and amino acids, such as those at 829, 850, 1334, and 1357 cm⁻¹, in addition to the range of ~1503–1653 cm⁻¹, noting an approximate similarity with the profile of this latter range observed in the resistant *E. coli* group. The peaks in the range of ~734–804 cm⁻¹ can be assigned to nucleotide bases. The susceptible *K. pneumoniae* group showed higher intensity in the negative bands with peaks that can be assigned to proteins and amino acids in the ranges of ~622–643 cm⁻¹, ~1451–1463 cm⁻¹, and ~1606–1659 cm⁻¹, and to LPS at the peaks of 922 and 961 cm⁻¹.

When analyzing and comparing the susceptible and resistant strain groups of both species, the peak ranges with the highest intensity were found to be close. In the resistant *E. coli* groups, when observing the intensity in the mean difference plot, it was noted that the ampicillin-resistant group showed a higher intensity than the ciprofloxacin-resistant group. Indeed, these spectral variations may be related to the different phenotypic and adaptive mechanisms involved in surviving antimicrobial pressure.

B. Classification of Resistant Versus Sensitive Bacteria by PLS Regression

Table 2 presents the confusion matrix for the classification of spectra from resistant and sensitive strains of *E. coli* and *K. pneumoniae* species. Considering the 360 spectra obtained from the *E. coli* species, an accuracy of 84.4% was achieved. Analysis of the confusion matrix in Table 2 demonstrated that the technique correctly classified 205 out of 227 spectra from resistant strains (true positives) and 99 out of 133 spectra from sensitive strains (true negatives). The model presented only 22 false negatives (resistant strains incorrectly classified as sensitive) and 34 false positives (sensitive strains classified as resistant). With this distribution, the predictive validation achieved a sensitivity of 90.3% and a specificity of 74.4%. For the PLS-DA evaluation of *K. pneumoniae* strains, the model was also accurate. Based on the analysis of 140 spectra, the model achieved an accuracy of 88.6%, correctly classifying 79 out of 84 spectra from susceptible samples and 45 out of 56 from resistant samples. There were five false positives and 11 false negatives. From this distribution, the model demonstrated a sensitivity of 80.4% and a specificity of 94.0%. Furthermore, the PLS-DA analysis suggests that the classifiers may be capable of predicting unknown samples.

TABLE II
CONFUSION MATRIX FOR THE CLASSIFICATION OF THE SPECTRA FROM RESISTANT OR SENSITIVE STRAINS OF *E. coli* AND *K. pneumoniae*

E. coli	**Susceptible**	**Resistant**
Susceptible (*n* = 133)	99	34
Resistant (*n* = 227)	22	205
Sensitivity (%)	90.3%	
Specificity (%)	74.4%	
Accuracy (%)	84.4%	
K. pneumoniae	**Susceptible**	**Resistant**
Susceptible (*n* = 84)	79	5
Resistant (*n* = 56)	11	45
Sensitivity (%)	80.4%	
Specificity (%)	94.0%	
Accuracy (%)	88.6%	

IV. Discussion

The analysis of the difference spectra between the susceptible and resistant groups allowed a comparison of the intensity variations between the susceptibility and resistance of the strains, which may be related to the phenotypic mechanisms and adaptive survival strategies of each species.

Oliveira et al. (2012) [11] extended previous studies in order to differentiate pathogenic species using Raman spectroscopy, a capability later broadened by Oliveira et al. (2020) [13] to encompass a wider range of Gram-negative bacteria for differentiating pathogenic species using Raman spectroscopy, a capability further expanded by Oliveira et al. (2020) [13] to include a wider range of Gram-negative bacteria (*E. coli*, *K. pneumoniae*, *P. mirabilis*, and *E. cloacae*) based on their specific biochemical signatures. In another study, results demonstrated that metabolic changes can be associated with antimicrobial resistance, allowing spectral differences to be observed between resistance phenotypes, as well as interspecies variations [15]. In the current study, the resistant *E. coli* groups showed an increase in peak intensity for protein, amino acid, and nucleotide base assignments, as was also observed in the *K. pneumoniae* group. These findings may indicate phenotypic adaptations associated with cell envelope remodeling, particularly the downregulation or alteration of porins to reduce membrane permeability, a resistance strategy that restricts drug entry but often imposes a significant metabolic fitness cost [5]. In the ciprofloxacin-resistant groups, the spectral findings may be related to the drug's mechanism of action. As previously discussed, it acts by converting DNA gyrase and topoisomerase IV into toxic complexes that induce genome fragmentation and oxidative stress; thus, the concomitant higher intensity in protein and nucleic acid bands may suggest the activation of cellular protection mechanisms against chromosomal damage. This metabolic response observed in the spectra corroborates the hypothesis of PMQR [7]. Munita and Arias (2016) [1] and Reygaert (2018) [17] reported that clinical strains frequently harbor plasmid-mediated resistance determinants, such as the fluoroquinolone-modifying enzyme AAC(6')-Ib-cr or qnr genes, which encode proteins capable of physically shielding DNA gyrase against fluoroquinolones.

The susceptible *E. coli* and *K. pneumoniae* groups showed higher intensity peaks with characteristics mainly associated with lipoprotein assignments and aromatic amino acids, in addition to other protein ranges (which is expected, given the wide variety and types of proteins that compose the cell). The higher intensity peaks in these groups may be associated with the preservation of the structural integrity of cell envelope constituents, such as LPS and peptidoglycan, in bacteria that have not been subjected to adaptive stress. Indeed, the intensity in certain protein and aromatic amino acid peaks (such as phenylalanine, in the ~1004 cm^{-1} region) can be associated with the preserved expression of the species' porins, such as OmpK35 and OmpK36. Oliveira et al. (2026) [15] state that, from a baseline morphological and compositional standpoint, *E. coli* exhibits smaller global intensity variations in its spectra when compared to the other studied species. The results suggest that antibiotic resistance does not rely on an isolated molecular alteration.

Regarding the PLS-DA analysis, the study reinforced the clinical utility of Raman spectroscopy, which, when coupled with this discriminant analysis technique, can serve as a rapid diagnostic strategy, aligning with the global surveillance priorities outlined by the WHO [2] to combat antimicrobial resistance. The high sensitivity observed in the *E. coli* groups may reflect the model's ability to capture biochemical and structural alterations in the spectral signature associated with bacterial resistance mechanisms [1]. Indeed, exposure and resistance to β-lactam antibiotics, such as ampicillin, alter the composition and synthesis of bacterial cell wall structures [3], which is reflected in the electrophysical properties of the *E. coli* cell envelope [6]. These macromolecular changes are related to the vibrational modes of lipids and proteins in the Raman spectra, enabling the PLS-DA model to discriminate resistant phenotypes. The predictive variations between *E. coli* and *K. pneumoniae* highlight that different Gram-negative pathogens employ distinct envelope defense mechanisms [1], ranging from localized modifications in the outer membrane barrier [5] to alterations in intracellular targets, such as those protecting DNA gyrase against quinolones [7]. Ultimately, the chemometric analysis of the difference spectra between susceptible and resistant groups within the PLS-DA framework allowed for a direct comparison of intensity variations. This approach not only demonstrates significant predictive value for identifying resistance profiles in unknown clinical samples but also connects these spectral differences directly to the phenotypic mechanisms and adaptive survival strategies unique to each species.

V. Conclusion

This study reinforced that Raman spectroscopy can be a non-destructive and label-free analytical tool for differentiating the antimicrobial susceptibility profiles in clinical strains of Gram-negative bacteria, such as *E. coli* and *K. pneumoniae*. The results indicated that the spectral differences between susceptible and resistant strains may be associated with a set of structural changes, reflecting variations in lipid compounds, proteins, amino acids, and nucleotide bases. Despite the standardization challenges for direct clinical samples, spectral acquisition and the monitoring of biochemical characteristics support a promising path toward the development of rapid optical diagnostic systems, complementing conventional susceptibility testing in clinical environments. The precise distinction between susceptible and resistant strains occurred independently of the species analyzed, indicating that the spectral signatures of resistance override interspecies biological variations. Furthermore, the fact that the model achieved high accuracy using few latent variables (LVs) reveals that this phenotypic differentiation may present spectral bands from a restricted group of biochemical constituents involved in the resistance of these Gram-negative bacteria. Therefore, future studies are required to develop a database containing other bacterial species, both susceptible and resistant to antibiotics.

Acknowledgments

B. C. A. Costa thanks Universidade Anhembi Morumbi (UAM) for the doctoral fellowship. L. Silveira Jr. thanks the National Council for Scientific and Technological Development (CNPq) for the research productivity

fellowship (Grant No. 314167/2021-8) and Ânima Institute (AI) for the research fellowship.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] J. M. Munita and C. A. Arias, “Mechanisms of antibiotic resistance,” *Microbiol. Spectr.*, vol. *4*, nº 2, VMBF-0016-2015, 2016. DOI:10.1128/microbiolspec.

[2] World Health Organization, “Global antibiotic resistance surveillance report 2025: WHO Global Antimicrobial Resistance and Use Surveillance System (GLASS),” Geneva: World Health Organization, 2025. Disponível em: https://iris.who.int/server/api/core/bitstreams/eede4321-72b5-4185-b422-c2ce07f185df/content. Acesso em: 16 jun. 2026.

[3] H. Cho, T. Uehara and T. G. Bernhardt, “Beta-lactam antibiotics induce a lethal malfunctioning of the bacterial cell wall synthesis machinery,” *Cell*, vol. *159*, nº 6, pp. 1300-1311, 2014. DOI:10.1016/j.cell.2014.11.017.

[4] H. Nikaido and J.-M. Pagès, “Broad specificity efflux pumps and their role in multidrug resistance of Gram-negative bacteria,” *FEMS Microbiol. Rev.*, vol. *36*, nº 2, pp. 340-363, 2012. DOI:10.1111/j.1574-6976.2011.00290.x.

[5] A. Dunai et al., "Rapid decline of bacterial drug-resistance in an antibiotic-free environment through phenotypic reversion," *eLife*, vol. 8, p. e47088, 2019. DOI:10.7554/eLife.47088.

[6] O. I. Guliy et al., “The effect of ampicillin on the electrophysical properties of *Escherichia coli* cells,” *Microbiology*, vol. *74*, nº 1, pp. 111-115, 2005. DOI: 10.1007/s11021-005-0019-9.

[7] K. J. Aldred, R. J. Kerns and N. Osheroff, “Mechanism of quinolone action and resistance,” *Biochemistry*, vol. *53*, nº 10, pp. 1565-1574, 2014. DOI:10.1021/bi5000564.

[8] J. Strahilevitz, G. A. Jacoby, D. C. Hooper and A. Robicsek, “Plasmid-mediated quinolone resistance: A multifaceted threat,” *Clin. Microbiol. Rev.*, vol. *22*, nº 4, pp. 664-689, 2009. DOI:10.1128/CMR.00016-09.

[9] T. J. Silhavy, D. Kahne and S. Walker, “The bacterial cell envelope,” *Cold Spring Harbor Perspect. Biol.*, vol. *2*, nº 2, p. a000414, 2010. DOI:10.1101/cshperspect.a000414.

[10] J. A. Bengoechea e J. Sa Pessoa, “*Klebsiella pneumoniae* infection biology: living to counteract host defences,” *FEMS Microbiol. Rev.*, vol. *43*, nº 2, pp. 123-144, 2019. DOI:10.1093/femsre/fuy043.

[11] F. S. S. Oliveira, H. E. Giana and L. Silveira, “Discrimination of selected species of pathogenic bacteria using near-infrared Raman spectroscopy and principal components analysis,” *J. Biomed. Opt.*, vol. *17*, nº 10, p. 107004, 2012. DOI:10.1117/1.JBO.17.10.107004.

[12] E. Witkowska et al., “Detection and identification of human fungal pathogens using surface-enhanced Raman spectroscopy and principal component analysis,” *Anal. Methods*, vol. 8, no. 45, pp. 8427-8436, 2016. DOI:10.1039/c6ay02957d.

[13] F. S. S. Oliveira, A. M. da Silva, M. T. T. Pacheco, H. E. Giana and L. Silveira, “Biochemical characterization of pathogenic bacterial species using Raman spectroscopy and discrimination model based on selected spectral features,” *Lasers Med. Sci.*, vol. *35*, nº 8, pp. 1761-1772, 2020. DOI:10.1007/s10103-020-03028-9.

[14] P. Rösch et al., "Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: Application to clean-room-relevant biological contaminations," *Appl. Environ. Microbiol.*, vol. *71*, no. 3, pp. 1626-1637, 2005. DOI:10.1128/AEM.71.3.1626-1637.2005.

[15] F. S. S. Oliveira, A. M. da Silva, M. T. T. Pacheco and L. Silveira., “Detecting antibiotic resistance in strains of selected Gram-negative bacterial species using Raman spectroscopy,” *J. Raman Spectrosc.*, vol. *57*, pp. 1-11, 2026. DOI:10.1002/jrs.70167.

[16] S. Singh et al., “Antimicrobial resistance studies using Raman spectroscopy on clinically relevant bacterial strains,” *Anal. Chem.*, vol. *95*, nº 30, pp. 11342-11351, 2023.DOI:10.1021/acs.analchem.3c01453.

[17] W. C. Reygaert, “An overview of the antimicrobial resistance mechanisms of bacteria,” *AIMS Microbiol*., vol. *4*, nº. 3, pp. 482-501, 2018. DOI:10.3934/microbiol.2018.3.482.

Thermographic Analysis of Unstageable Pressure Injuries

Catia Terezinha Heimbecher
Graduate Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology – Paraná (UTFPR)
Centro Universitário Santa Cruz de Curitiba (USC)
Centro Universitário de Tecnologia de Curitiba (Unifatecpr)
Curitiba, Brazil
ORCID: 0000-0003-4487-153X

Leandra Ulbricht
Graduate Program in Biomedical Engineering (PPGEB)
Electronics Engineering Department (DAELN)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID: 0000-0002-9514-2938

Leomar Caetano de Oliveira
Bachelor of Science in Electronic Engineering
Electronics Engineering Department (DAELN)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID: 0009-0006-5872-0272

Joaquim Miguel Maia
Graduate Program in Electrical and Computer Engineering (CPGEI)
Electronics Engineering Department (DAELN)
Federal University of Technology – Paraná (UTFPR)
Curitiba, Brazil
ORCID: 0000-0001-6309-8664

***Abstract*—Unstageable pressure injuries are covered by slough or eschar that obscures the wound bed and the depth of tissue damage. This study evaluated infrared thermography combined with seeded region growing to assess these injuries. A descriptive cross-sectional study was conducted on six injuries from three patients drawn from a database compiled from 30 patients admitted to a referral hospital in Curitiba. Images were acquired using a Fluke Ti32 camera, processed in SmartView, and segmented into the outer area, border, inner layers, and thermal islands using thresholds ranging from 0.05 to 0.50. The minimum, mean, and maximum temperatures of each region, the mean temperature by region type, and the thermal gradient between the outer area and the islands were determined. Thermal gradients ranged from 1.22 to 4.44 degrees Celsius, with the largest values observed in the two sacrococcygeal injuries of Patient B, while Patient A showed an irregular internal reheating pattern. The approach quantified spatial temperature variability and provided information complementary to clinical inspection. The small number of cases and the cross-sectional design limit the generalizability of the findings.**

***Index Terms*—thermography, pressure injury, thermal analysis, image segmentation.**

I. Introduction

Unstageable pressure injuries are characterized by the inability to determine the actual depth of tissue damage because slough or eschar obscures the wound bed [1]–[3].

These injuries pose a challenge in clinical practice because devitalized tissue prevents direct inspection of the wound bed and makes it difficult to estimate the extent of tissue involvement [4]–[7]. This limitation may delay treatment decisions and increase the complexity of care [8].

The limitations of visual assessment alone have encouraged the use of complementary methods to identify physiological alterations before clinically evident manifestations occur [9], [10]. Among these methods, infrared thermography enables the measurement of temperature variations associated with inflammation and blood perfusion [11]–[13].

However, interpretation based solely on the thermogram color palette is insufficient. Quantitative analysis of thermal data can improve characterization of the spatial temperature distribution and reduce assessment subjectivity.

Seeded Region Growing (SRG) is a segmentation method based on the similarity of adjacent pixels. Starting from a seed defined within the region of interest, the algorithm iteratively incorporates neighboring pixels that satisfy a homogeneity criterion. Growth ceases when the predefined threshold is no longer met, resulting in spatially connected and thermally similar regions [14].

Accordingly, this study aimed to evaluate infrared thermography combined with seeded region growing for assessing unstageable pressure injuries.

II. Methods

This descriptive cross-sectional study was approved by the Research Ethics Committee under protocol no. 5,651,212. Thermal data were obtained from hospitalized patients with limited mobility in the intensive care unit, medical clinic, and trauma wards of a referral hospital in Curitiba, Brazil. Unstageable pressure injuries were selected from a previously established database of 30 individuals, using the National Pressure Injury Advisory Panel classification [3]. Injuries covered by slough were excluded. The final sample comprised six injuries covered by eschar from three patients.

Thermal images were acquired using a Fluke Thermography Ti32 infrared camera and processed in SmartView software, version 4.3. Thermographic image acquisition was standardized, with temperature maintained at 22 ± 1°C and humidity between 45% and 55% using a digital thermo-hygrometer with a probe (SH112; J-Prolab, Brazil). Acclimatization lasted five minutes, during which the region to be imaged remained exposed and free from contact to avoid thermal imprinting. During image acquisition, participants were positioned in lateral recumbency. An approximate distance of 50 cm was maintained between the skin surface and the camera. The initial analysis used a blue-to-red color palette from the software accompanying the thermal camera to represent the minimum and maximum temperature ranges, with blue tones corresponding to lower temperatures and yellow and red tones to higher temperatures.

The data were exported in `.txt` format and transferred to the API software. Seeded region growing was then applied to segment the regions of interest [14]. Delineation began in the outer area and progressed toward the internal regions, forming successive layers and thermal islands.

Layer and island formation were stopped when the segmented region spanned the injury perimeter or reached the threshold defined between 0.05 and 0.50. The starting point for seed growth development was established as the point closest to the previously developed margin, ensuring no areas were left unanalyzed. The initial marking was made at the edge, followed by the outer area and subsequently the inner layers, until only small areas—designated as "islands"—remained. The colors for these strata were selected from the API's available palette to ensure contrast between the different layers and islands. A 200% zoom level was applied to better identify the seed growth starting points; the API allows for zooming up to 950% if necessary. For each region, the minimum, mean, and maximum temperatures, along with the number of pixels, were recorded to characterize the thermal distribution within each injury. For Patient B, the legend retained the same island indicators for the right and left lesions to ensure standardization, even though data were unavailable for Islands 2 and 3 in the right sacrococcygeal lesion. For Patient C, the seeded region growing procedure was repeated three times on the gluteal lesion to assess repeatability.

Descriptive statistics were calculated from the mean temperatures of the segmented regions. For each of the six injuries, the number of regions (N) and mean temperature were calculated by region type: outer area, border, layer, and island. The thermal gradient (ΔT) was defined as the mean temperature of the outer area minus the mean temperature of the islands.

III. Results and Discussion

Six unstageable pressure injuries from three patients were analyzed. This injury category was selected because the presence of eschar limits direct clinical assessment. Appropriate identification and classification are essential for reducing complications and guiding care [15].

Fig. 1 presents optical and thermal images of Patients A, B, and C, including one trochanteric injury, four sacrococcygeal regions analyzed separately, and one gluteal injury.

The optical images show eschar but do not provide direct information on thermal distribution or underlying perfusion. In the thermograms, orange and red regions correspond to higher temperatures, whereas blue tones indicate lower temperatures.

Siah et al. [16] observed that distinct thermal patterns in surgical wounds identified during the first postoperative days were associated with delayed healing. These findings indicate that thermography may contribute to monitoring perfusion and the inflammatory response.

Bates-Jensen et al. [17] demonstrated that objective skin temperature measurement improves early identification of pressure-injury-related erythema, particularly when visual inspection is limited by skin tone.

Collectively, this evidence indicates that thermal assessment provides physiological information complementary to visual inspection and may reveal microcirculatory alterations before clinically evident manifestations [16], [17].

After seeded region growing was applied, thermal layers and islands were identified, revealing an asymmetric and variable temperature distribution. These variations may reflect changes in blood perfusion and tissue metabolic activity [11], [18].

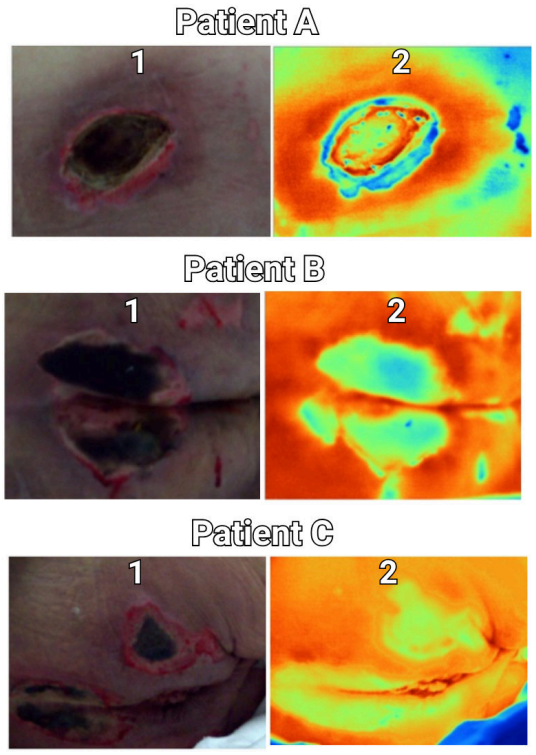
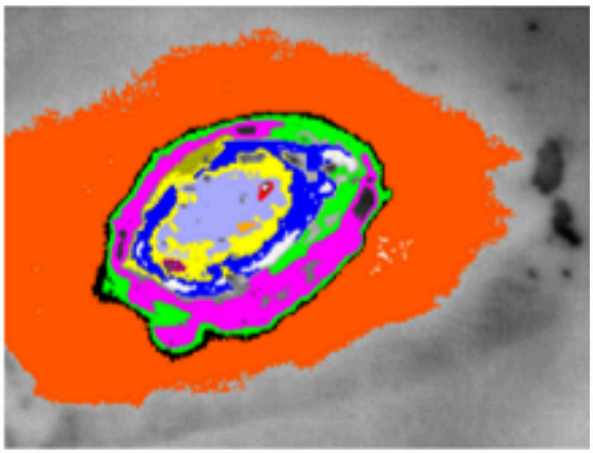


Figure 1. Optical and thermal images of the six unstageable pressure injuries from three patients. (1) Optical image; (2) thermal image using the blue-to-red palette.

Fig. 2 presents the segmented thermogram of the trochanteric injury in Patient A.



Legend

Region	Mean temperature (°C)
Outer area	38.74
Border	38.11
Inner layer 1	37.77
Inner layer 2	37.06
Inner layer 3	38.25
Inner layer 4	38.72
Inner layer 5	37.80
Island 1	37.31
Island 2	37.78
Island 3	37.68
Island 4	37.32
Island 5	36.36

Figure 2. Segmented thermogram of the trochanteric injury in Patient A, with layers generated by seeded region growing.

The minimum, mean, and maximum temperatures of the segmented regions in Patient A are presented in Table I.

Table I
THERMAL DATA FOR THE INJURY IN PATIENT A

Areas	Min (°C)	Mean (°C)	Max (°C)
Outer area	38.4	38.74	39.3
Border	37.9	38.11	38.3
Inner layer 1	37.5	37.77	38.0
Inner layer 2	36.7	37.06	37.3
Inner layer 3	38.0	38.25	38.5
Inner layer 4	38.5	38.72	39.0
Inner layer 5	37.5	37.80	38.0
Island 1	37.0	37.31	37.5
Island 2	37.5	37.78	38.0
Island 3	37.4	37.68	37.9
Island 4	37.1	37.32	37.5
Island 5	36.2	36.36	36.5

The segmentation enables the identification of regions with similar thermal behavior and the establishment of boundaries between adjacent areas. This strategy supports quantitative analysis of temperature variations and improves thermogram interpretation [14], [19], [20].

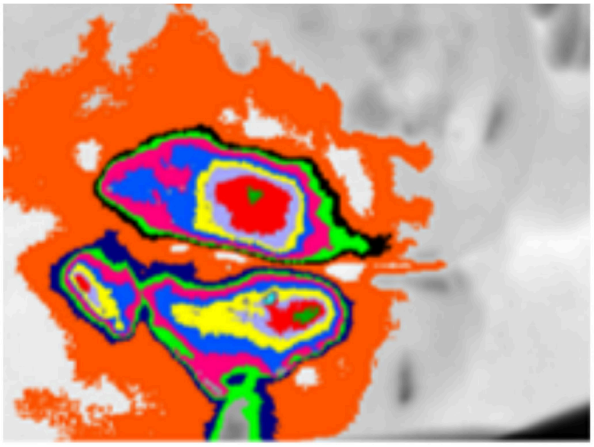
Fig. 3 presents the segmentation of the two sacrococcygeal injuries in Patient B. Although they share an outer area, the algorithm separated the right and left regions, which were analyzed independently in Tables II and III.

Table II
THERMAL DATA FOR THE RIGHT SACROCOCCYGEAL INJURY IN PATIENT B

Areas	Min (°C)	Mean (°C)	Max (°C)
Outer area	37.5	37.90	38.3
Border	36.7	37.12	37.5
Inner layer 1	36.1	36.51	36.9
Inner layer 2	35.5	35.81	36.3
Inner layer 3	35.0	35.24	35.4
Inner layer 4	34.8	34.89	35.0
Inner layer 5	34.3	34.51	34.7
Inner layer 6	33.8	34.05	34.2
Island 1	33.6	33.75	33.8
Island 2	–	–	–
Island 3	–	–	–

Table III
THERMAL DATA FOR THE LEFT SACROCOCCYGEAL INJURY IN PATIENT B

Areas	Min (°C)	Mean (°C)	Max (°C)
Outer area	37.5	37.90	38.3
Border	36.7	37.12	37.5
Inner layer 1	36.2	36.58	37.0
Inner layer 2	35.8	36.19	36.6
Inner layer 3 UPPER	35.4	35.67	35.9
Inner layer 3 LOWER	35.4	35.65	35.9
Inner layer 4 UPPER	35.0	35.28	35.5
Inner layer 4 LOWER	34.9	35.14	35.3
Inner layer 5 UPPER	34.7	34.80	34.9
Inner layer 5 LOWER	34.6	34.72	34.8
Inner layer 6 UPPER	34.3	34.52	34.6
Inner layer 6 LOWER	34.4	34.52	34.6
Island 1	34.3	34.38	34.4
Island 2	32.8	33.13	33.5
Island 3	32.6	32.87	33.1



Legend

Region	Mean temperature (°C)		
	Right	Left–upper	Left–lower
Outer area	37.90	37.90	37.90
Border (right)	37.12	—	—
Border (left)	—	37.12	37.12
Inner layer 1	37.12	36.58	36.58
Inner layer 2	35.81	36.19	36.19
Inner layer 3	35.24	35.67	35.65
Inner layer 4	34.89	35.28	35.14
Inner layer 5	34.51	34.80	34.72
Inner layer 6	34.05	34.52	34.52
Island 1	33.75	—	34.38
Island 2	—	—	33.13
Island 3	—	—	32.87

— Region absent

Figure 3. Segmented thermogram of the right and left sacrococcygeal injuries in Patient B, with layers generated by seeded region growing.

Thermal data for the right sacrococcygeal injury in Patient B are presented in Table II.

Thermal data for the left sacrococcygeal injury in Patient B are presented in Table III.

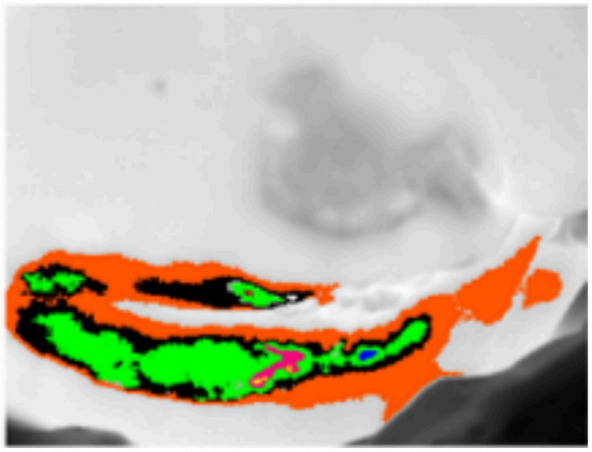
Fig. 4 presents separate segmentations of the right and left sacrococcygeal injuries in Patient C.

Table IV
THERMAL DATA FOR THE RIGHT SACROCOCCYGEAL INJURY IN PATIENT C

Areas	Min (°C)	Mean (°C)	Max (°C)
Outer area	34.6	35.01	35.4
Border UPPER	34.4	34.50	34.6
Border LOWER	34.2	34.43	34.6
Inner layer 1 UPPER	34.2	34.37	34.4
Inner layer 1 LOWER	33.6	33.81	34.0
Inner layer 2	–	–	–
Inner layer 3	–	–	–
Inner layer 4	–	–	–
Inner layer 5	–	–	–
Island 1	33.4	33.48	33.5
Island 2	34.1	34.10	34.1
Island 3	–	–	–

Table V
THERMAL DATA FOR THE LEFT SACROCOCCYGEAL INJURY IN PATIENT C

Areas	Min (°C)	Mean (°C)	Max (°C)
Outer area	34.6	35.01	35.4
Left border	34.1	34.28	34.5
Inner layer 1	33.5	33.91	34.1
Inner layer 2	–	–	–
Inner layer 3	–	–	–
Inner layer 4	–	–	–
Inner layer 5	–	–	–
Island 1	32.8	33.18	33.4
Island 2	33.0	33.33	33.5
Island 3	32.6	32.72	32.8
Island 4	32.3	32.56	32.9
Island 5	–	–	–



Legend

Region	Mean temperature (°C)		
	Right–upper	Right–lower	Left
Outer area	35.01	35.01	35.01
Border	34.50	34.43	34.28
Inner layer 1	34.37	33.81	33.91
Island 1 (left)	—	—	33.18
Island 2 (left)	—	—	33.33
Island 3 (left)	—	—	32.72
Island 4 (left)	—	—	32.56
Island 1 (right)	—	33.48	—
Island 2 (right)	—	34.10	—

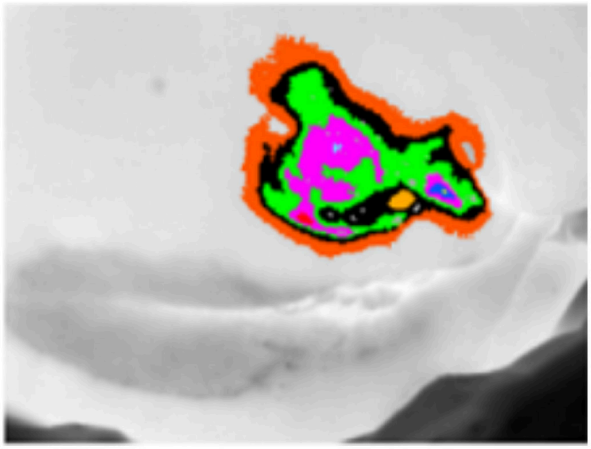
— Region absent.

Figure 4. Segmented thermograms of the right and left sacrococcygeal injuries in Patient C, with layers generated by seeded region growing.

Thermal data for the right sacrococcygeal injury are presented in Table IV, and those for the left sacrococcygeal injury are presented in Table V.

Table V details the left sacrococcygeal injury in Patient C. As in Patient B, the right-left subdivision resulted from the stratification generated by seeded region growing.

Fig. 5 presents the segmentation of the gluteal injury in Patient C.



Legend

Region	Mean temperature (°C)		
	Upper	Lower	Overall
Outer area	35.38	35.38	—
Border	34.72	34.72	—
Inner layer 1	34.24	34.24	—
Inner layer 2	33.89	33.76	—
Inner layer 3	—	33.36	—
Island 1	—	—	35.32
Island 2	—	—	33.55
Island 3	—	—	33.10
Island 4	—	—	33.60

— Region absent.

Figure 5. Segmented thermogram of the gluteal injury in Patient C, with layers generated by seeded region growing.

Thermal data for the gluteal injury in Patient C are presented in Table VI.

Table VI
THERMAL DATA FOR THE GLUTEAL INJURY IN PATIENT C

Areas	Min (°C)	Mean (°C)	Max (°C)
Outer area	34.9	35.38	35.7
Border	34.5	34.72	35.0
Inner layer 1	34.1	34.24	34.4
Inner layer 2 UPPER	33.7	33.89	34.0
Inner layer 2 LOWER	33.6	33.76	33.9
Inner layer 3	–	–	–
Inner layer 3 LOWER	33.2	33.36	33.5
Inner layer 4	–	–	–
Inner layer 5	–	–	–
Island 1	35.1	35.32	35.7
Island 2	33.4	33.55	33.6
Island 3	33.1	33.10	33.1
Island 4	33.6	33.60	33.6
Island 5	–	–	–

To assess the repeatability of the seeded region growing procedure, the gluteal injury in Patient C was segmented three times, as shown in Table VII.

Considering the mean temperature of corresponding regions, the maximum absolute difference among the three repetitions was 0.38 °C in Island 4, numerically corresponding to 1.13% of the three-repetition mean. The mean relative range across the evaluated regions was 0.57%, indicating low intra-image variation among the repeated segmentations.

Table VIII summarizes the descriptive statistics by region type and the thermal gradient for the six injuries.

Table VII
TRIPLICATE THERMAL DATA FOR THE GLUTEAL INJURY IN PATIENT C

Areas	**Behavior 1**			**Behavior 2**			**Behavior 3**		
	Min (°C)	**Mean (°C)**	**Max (°C)**	**Min (°C)**	**Mean (°C)**	**Max (°C)**	**Min (°C)**	**Mean (°C)**	**Max (°C)**
Outer area	34.9	35.38	35.7	34.9	35.38	35.7	34.6	35.02	35.4
Border	34.5	34.72	35.0	34.7	34.90	35.0	34.3	34.60	35.0
Inner layer 1	34.1	34.24	34.4	34.0	34.30	34.4	34.2	34.32	34.5
Inner layer 2 UPPER	33.7	33.89	34.0	33.8	33.90	34.0	33.7	33.94	34.1
Inner layer 2 LOWER	33.6	33.76	33.9	33.7	33.87	34.0	33.6	33.82	34.0
Inner layer 3 LOWER	33.2	33.36	33.5	33.0	33.20	33.3	33.2	33.36	33.5
Island 1	35.1	35.32	35.7	34.9	35.20	35.7	34.8	35.00	35.4
Island 2	33.4	33.55	33.6	33.4	33.55	33.6	33.2	33.40	33.5
Island 3	33.1	33.10	33.1	33.1	33.16	33.2	33.1	33.10	33.1
Island 4	33.6	33.60	33.6	33.6	33.68	33.7	33.3	33.30	33.3

Table VIII
DESCRIPTIVE STATISTICS AND THERMAL GRADIENTS FOR THE SIX INJURIES

Patient/injury	**Outer area**		**Border**		**Layer**		**Island**		ΔT (°C)*
	N	**Mean (°C)**	**N**	**Mean (°C)**	**N**	**Mean (°C)**	**N**	**Mean (°C)**	
Patient A	1	38.74	1	38.11	5	37.92	5	37.29	1.45
Patient B – right sacrococcygeal	1	37.90	1	37.12	6	35.17	1	33.75	4.15
Patient B – left sacrococcygeal	1	37.90	1	37.12	10	35.31	3	33.46	4.44
Patient C – right sacrococcygeal	1	35.01	2	34.47	2	34.09	2	33.79	1.22
Patient C – left sacrococcygeal	1	35.01	1	34.28	1	33.91	4	32.95	2.06
Patient C – gluteal	1	35.38	1	34.72	4	33.81	4	33.89	1.49

*ΔT = mean temperature of the outer area − mean temperature of the islands.

The largest gradients occurred in Patient B: 4.15 °C (right) and 4.44 °C (left). These values represent the largest thermal contrast between the outer area and the islands; however, they should not be interpreted in isolation as a measure of clinical severity. Lower temperatures in the central regions may be associated with reduced perfusion and heat transport and should be interpreted alongside absolute temperatures and clinical assessment [11], [18], [21].

Patient A had a gradient of 1.45 °C but showed reheating in inner layers 3 and 4, with mean temperatures of 38.25 °C and 38.72 °C, respectively. This pattern confirms that the difference between the outer area and the islands does not, by itself, describe the full pattern of regional temperature variation within the injury.

In Patient C, the gradients were 1.22 °C in the right sacrococcygeal injury, 2.06 °C in the left sacrococcygeal injury, and 1.49 °C in the gluteal injury. In the gluteal injury, Island 1 had a mean temperature of 35.32 °C, whereas the remaining islands ranged from 33.10 °C to 33.60 °C, demonstrating internal temperature variability that would be attenuated by using only the overall island mean.

Prolonged pressure can cause cell deformation, impaired microcirculation, hypoxia, inflammation, and ischemic damage [12], [22]. Because blood flow contributes to heat transport in tissues, changes in perfusion may alter local temperature [18]. However, inflammatory processes, environmental conditions, and individual characteristics also affect thermograms, requiring cautious interpretation.

Overall, the findings indicate that segmentation into layers and islands adds information to visual analysis by identifying central cooling and reheating patterns, as well as spatial temperature variability. Regional quantification may reduce subjectivity and support injury monitoring without replacing clinical assessment [11], [14], [21].

The main limitations were the small number of patients, the cross-sectional design, and the lack of longitudinal follow-up of clinical progression.

IV. CONCLUSION

Infrared thermography combined with seeded region growing quantified thermal differences in six unstageable pressure injuries and distinguished central cooling and internal reheating patterns that were not evident through visual inspection alone. The largest gradients occurred in the two sacrococcygeal injuries of Patient B, whereas the injuries in Patients A and C showed lower contrast and greater regional temperature variation. This approach has potential as a complementary tool for clinical assessment; however, studies with larger sample sizes and longitudinal follow-up are needed to better define its clinical applicability and limits of interpretation.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil – Finance Code 001. The authors would like to acknowledge the support provided by the National Council for Scientific and Technological Development (CNPq), the Federal University of Technology – Paraná (UTFPR), Centro Universitário Santa Cruz de Curitiba (USC), Centro Universitário de Tecnologia de Curitiba (Unifatecpr) and the Araucária Foundation.

REFERENCES

[1] LABEAU, S. O.; et al. "Prevalence, associated factors, and outcomes of pressure injuries in adult intensive care unit patients: the DecubICUs study," *Intensive Care Medicine*, vol. 47, no. 6, pp. 503–520, 2021, doi: 10.1007/s00134-021-06399-8.

[2] MARTÍNEZ-TORIJA, M.; et al. "Multifaceted pathophysiology and secondary complications of chronic spinal cord injury: focus on pressure injury," *Journal of Clinical Medicine*, vol. 14, 2025.

[3] NATIONAL PRESSURE INJURY ADVISORY PANEL (NPIAP). "Pressure injury stages," Mount Laurel, NJ, 2023. [Online]. Available: https://npiap.com/page/PressureInjuryStages.

[4] COLEMAN, S.; et al. “Pressure ulcer prevention: understanding pressure, shear and friction,” *Wounds International*, vol. 14, no. 1, pp. 20–25, 2023.

[5] DUCHESNE, G. A.; et al. “Pressure ulcer diagnosis is associated with increased mortality in patients with end-stage renal disease: a retrospective study,” *Life*, vol. 13, no. 8, p. 1713, 2023, doi: 10.3390/life13081713.

[6] SALOMÉ, G. M.; et al. “Extrinsic factors of pressure injuries in patients during surgery: a systematic review,” *International Wound Journal*, vol. 21, no. 3, 2024.

[7] SHI, C.; et al. “Continuous pressure exacerbates ischemia-reperfusion injury in pressure-induced muscle damage,” *Journal of Wound Care*, vol. 34, no. 11, 2025.

[8] KOTTNER, J.; CUDDIGAN, J.; CARVILLE, K.; BALZER, K.; BERLOWITZ, D.; HAESLER, E.; et al. “Prevention and treatment of pressure ulcers/injuries: the protocol for the international guideline,” *International Wound Journal*, vol. 17, no. 3, pp. 766–770, 2020, doi: 10.1111/iwj.13350.

[9] FULBROOK, P.; LOVEGROVE, J. “Reporting accuracy of pressure injury categorisation in an acute tertiary hospital: a four-year analysis,” *Journal of Clinical Nursing*, vol. 32, nos. 15–16, pp. 5149–5160, 2023, doi: 10.1111/jocn.16789.

[10] PARK, S.; et al. “Knowledge and visual differentiation ability of the pressure injury classification system and incontinence-associated dermatitis among hospital nurses: a descriptive study,” *Healthcare*, vol. 12, p. 145, 2024.

[11] CAI, F.; et al. “Application of infrared thermography in the early warning of pressure injury: a prospective observational study,” *Journal of Clinical Nursing*, vol. 30, nos. 13–14, pp. 1906–1916, 2021, doi: 10.1111/jocn.15605.

[12] GEFEN, A.; et al. “The biomechanics of deep tissue injury: current understanding and future directions,” *Journal of Tissue Viability*, vol. 32, no. 2, pp. 145–156, 2023, doi: 10.1016/j.jtv.2023.02.003.

[13] JIANG, X.; et al. “Application of an infrared thermography-based model to detect pressure injuries: a prospective cohort study,” *British Journal of Dermatology*, vol. 187, no. 4, pp. 571–579, 2022.

[14] SCHADECK, C. A.; GANACIM, F.; ULBRICHT, L. “Processamento de imagens como metodologia auxiliar à análise de termogramas,” *Pan American Journal of Medical Thermology*, vol. 6, pp. 31–41, 2021, doi: 10.18073/pajmt.2019.6.31-41.

[15] EUROPEAN PRESSURE ULCER ADVISORY PANEL (EPUAP); NATIONAL PRESSURE INJURY ADVISORY PANEL (NPIAP); PAN PACIFIC PRESSURE INJURY ALLIANCE (PPPIA), *Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline*, 3rd ed. Osborne Park: Cambridge Media, 2019.

[16] SIAH, C. J. R.; et al. “An observational study of temperature and thermal images of surgical wounds for detecting delayed wound healing within four days after surgery,” *Journal of Clinical Nursing*, vol. 28, nos. 11–12, pp. 2285–2295, 2019, doi: 10.1111/jocn.14832.

[17] BATES-JENSEN, Barbara M.; et al. “Thermal measurement of erythema across skin tones: implications for clinical identification of early pressure injury,” *Journal of Tissue Viability*, vol. 33, no. 4, pp. 745–752, 2024, doi: 10.1016/j.jtv.2024.08.002.

[18] WADA, M.; et al. “Infrared thermography for assessing skin perfusion and tissue viability: current evidence and clinical applications,” *Journal of Thermal Biology*, vol. 98, 2021.

[19] LOIZIDOU, K.; SKOUROUMOUNI, G.; NIKOLAOU, C.; PITRIS, C. “Automatic breast mass segmentation and classification using subtraction of temporally sequential digital mammograms,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, p. 1801111, 2022, doi: 10.1109/JTEHM.2022.3219891.

[20] MUÑOZ, J.; et al. “Automatic segmentation methods in medical thermography: a systematic review,” *Sensors*, vol. 22, 2022.

[21] LIN, Y. H.; CHEN, Y. C.; CHENG, K. S.; YU, P. J.; WANG, J. L.; KO, N. Y. “Higher periwound temperature associated with wound healing of pressure ulcers detected by infrared thermography,” *Journal of Clinical Medicine*, vol. 10, no. 13, p. 2883, 2021, doi: 10.3390/jcm10132883.

[22] SOPPI, E.; KNUUTI, J.; KALLIOKOSKI, K. “Positron emission tomography study of effects of two pressure-relieving support surfaces on pressure ulcer development,” *Annals of Medicine*, vol. 53, no. 1, pp. 204–214, 2021, doi: 10.1080/07853890.2020.1867237.

Evaluation of the Surface of Human Amniotic Membrane Processed with a Hydrodynamic System Using Ozonated Water

1st Milena dos Santos Souza
Universidade Anhembi Morumbi
Centro de Inovação, Tecnologia e Educação (CITÉ)
São José dos Campos, São Paulo, Brasil
https://orcid.org/0009-0000-8765-4434

2nd Sara Evelyn Batista Aquino
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi
São José dos Campos, São Paulo, Brasil
https://orcid.org/0009-0003-2414-6186

3rd Jessica Sthefanny Carvalho
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi
São José dos Campos, São Paulo, Brasil
https://orcid.org/0009-0007-5550-0116

4th Melissa Zuccolotti Soares
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi
São José dos Campos, São Paulo, Brasil
https://orcid.org/0000-0001-5136-526X

5th Adriana Barrinha Fernandes Moretti
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro de Inovação, Tecnologia e Educação (CITÉ)
São José dos Campos, São Paulo, Brasil
fernandesabm@gmail.com

6th Carlos José de Lima
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro de Inovação, Tecnologia e Educação (CITÉ)
São José dos Campos, São Paulo, Brasil
cdcfdlima@gmail.com

***Abstract* — De-epithelialized human amniotic membrane (HAM) shows better adhesion and proliferation of epithelial cells compared to its intact form. However, conventional de-epithelialization methods (EDTA, Dispase II, trypsin-EDTA) require prolonged exposure and may compromise the basement membrane. This study evaluated the effect of a single 15-minute exposure of HAM to ozone (O3) using a hydrodynamic system, through qualitative histological analysis (H&E) and optical profilometry (WYKO NT1100), in a sample obtained from a single donor. Histological analysis showed preservation of the internal structures of HAM after treatment, with discrete alterations restricted to the epithelial layer compared to the control group. Optical profilometry recorded a reduction in mean surface roughness (Ra), from 0.626 (± 0.116 µm) in the in natura group to 0.509 (± 0.132) in the treated group, with no statistically significant difference (p = 0.0835; unpaired t-test with Welch's correction). These preliminary findings suggest that O3 exposure may promote alterations in the HAM epithelial layer while preserving the basement membrane and stroma, which reinforces its potential as an alternative to conventional chemical and enzymatic methods.**

Keywords* — *human amniotic membrane, ozone, de-epithelialization, skin graft.

I. Introduction

Human Amniotic Membrane (HAM), also known as the amnion, is the innermost layer of the placenta that surrounds and protects the embryo, and is generally regarded as biological waste after delivery. Histologically, it is composed of three layers: an epithelial layer, a basement membrane, and a stroma.

The epithelial layer is formed by a single layer of epithelial cells with a columnar or cuboidal shape, supported by a basement membrane composed mainly of type IV collagen and laminin. Laminin plays a fundamental role in epithelial cell adhesion and also contributes to their proliferation, migration, and differentiation. Beneath it lies the stroma, composed of three layers: an inner compact layer, a fibroblast layer, and a spongy layer. The stromal extracellular matrix is largely composed of type I, III, V, and VI collagen, in addition to laminin and fibronectin [2].

Historically, HAM has been successfully used as a skin graft substitute in the treatment of burns and wounds for more than a century, and its application has expanded to several areas of regenerative medicine, such as dentistry, oncology, and urology. This wide range of applications is related to its low immunogenicity, which reduces the risk of rejection and inflammation, as well as its ability to accelerate epithelialization and reduce pain [1,3].

As a biological tissue, HAM must be disinfected or sterilized before use, since microbial contamination, mainly by bacteria, is one of the main causes of discard. Gamma radiation is considered the gold-standard method for tissue sterilization, generally applied at a dose of 25 kGy to ensure high microbiological safety. However, doses equal to or greater than 20 kGy can cause structural alterations in HAM, including damage to the basement membrane, degradation of collagen fibers, nuclear chromatin condensation, and reduction of growth factors and tissue inhibitors of metalloproteinases [4]. In this context, there has been growing interest in alternative low-temperature sterilization methods capable of better preserving the physicochemical

and biological properties of biomaterials, in addition to offering faster processing and lower environmental impact.

Ozone (O3) has been investigated as an effective alternative compared to conventional methods, owing to its high oxidative power. When this gas is dissolved in water, its molecular bonds are broken, generating reactive oxygen species that react with organic and inorganic compounds. This process increases the permeability of the cytoplasmic membrane and favors its interaction with the cytoplasm, resulting in cell dissolution [5].

Awoyama et al. [6] found that 15 minutes of ozonation using a hydrodynamic system with ozonated water was able to efficiently disinfect HAM samples experimentally contaminated with Staphylococcus aureus. The authors reported small morphological alterations in the epithelial layer, while the remaining internal structures remained preserved. Based on these findings, the present study aimed to evaluate the effect of exposing HAM to O3 for 15 minutes, using the same hydrodynamic system employed by Awoyama et al. [6], to assess alterations on the surface of the biomaterial. For this purpose, histological analysis was used to verify possible alterations in the epithelial layer and the preservation of the underlying structures. In addition, optical profilometry was employed to quantify roughness (Ra) and verify topographic alterations on the epithelial surface of HAM.

II. Materials and Methods

This study, approved by the Research Ethics Committee of Universidade Anhembi Morumbi (opinion No. 3.984.423; CAAE: 289370020.0.0000.5492) and conducted in accordance with ANVISA RDC No. 190/2003, used HAM obtained at Santa Casa de Misericórdia de Pindamonhangaba (SP, Brazil). The membrane was obtained from a cesarean delivery performed under aseptic conditions, from a donor aged between 28 and 30 years, presenting negative serology for syphilis, hepatitis B, toxoplasmosis, and HIV, following signature of the Free and Informed Consent Form. Samples were washed with sterile saline solution (0.9% NaCl) to remove biological residues and subsequently segmented for the experimental procedures.



Figure 1. Collection and separation of HAM.

A. Experimental procedure

The hydrodynamic system used in this study was developed based on the methodology described by Awoyama et al. (2022) [6], consisting of a closed circuit for ozone generation and dissolution in water, with components manufactured from materials resistant to the oxidative action of the gas. The ozonated water circulated continuously through the system to the head where the HAM samples were positioned, in a climate-controlled environment with controlled temperature. In this study, the exposure time was standardized at 15 minutes, which, according to Awoyama et al. (2022), was the time required for HAM disinfection [6]. This exposure condition was selected based on the previously reported disinfection protocol and was not intended to establish an optimal de-epithelialization exposure time.

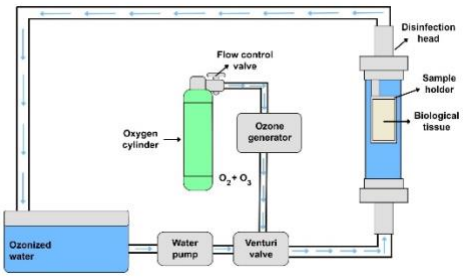


Figure 2. Schematic representation of the hydrodynamic system used for HAM de-epithelialization. Adapted from Souza et al. [8].

B. Histological analysis

For histological evaluation, HAM samples were fixed in 10% formalin solution and subjected to dehydration, clearing, infiltration, and paraffin embedding. Histological sections were obtained using a rotary microtome and stained with hematoxylin and eosin (H&E). Slide analysis was performed using an optical microscope (Opticam® Microscopy Technology), with image capture through T Capture® software. The histological assessment was primarily qualitative, focusing on the epithelial layer and preservation of the basement membrane and stroma.

C. Optical profilometry

Profilometric analysis was performed on a WYKO NT1100 instrument to evaluate possible alterations in the surface morphology of HAM after experimental processing. Samples were divided into a control group (in natura) and a group subjected to the hydrodynamic system with ozonated water for 15 minutes, with measurements taken in standardized central areas of each sample to ensure reproducibility, based on the methodology described in the study by Souza et al., 2026. The main parameter analyzed was mean roughness (Ra), determined from eight measurements (n = 8) obtained from samples derived from the same donor, and defined as the arithmetic mean of the absolute height deviations relative to the centerline of the evaluated profile. Ra means were compared using an unpaired t-test with Welch's correction (two-tailed), preceded by a Kolmogorov-Smirnov normality test ($p > 0.10$ for both groups), with significance set at $p < 0.05$.

III. Results and Discussion

Histological analysis of HAM showed preservation of the internal structures after treatment with O3 for 15 minutes (C

and D), with discrete alterations observed in the epithelial layer compared to the control group (A and B). These findings are consistent with the results reported by Awoyama et al. [6], who also described small morphological alterations in the epithelium after O3 exposure, without significant compromise of the structural integrity of the membrane. In addition, the basement membrane and stroma remained preserved after ozonation (Figure 3).

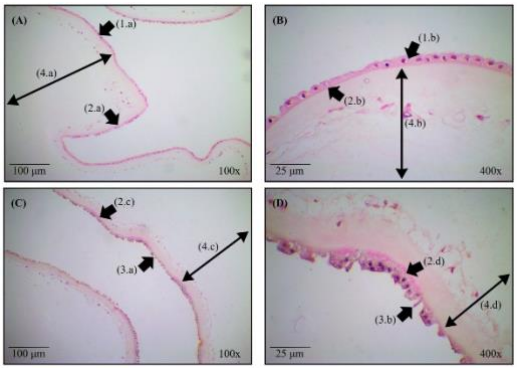


Figure 3. Histological images of HAM stained with H&E. Images (A and B) represent the in natura samples; (**C** and **D**) represent samples processed with O3 for 15 minutes. Arrows (**1.a**) and (**1.b**) indicate the epithelial layer; (**2.a**), (**2.b**), (**2.c**), and (**2.d**) indicate the basement membrane; (**3.a**) and (**3.b**) indicate the alterations observed in the epithelial layer of the treated group; (**4.a**), (**4.b**), (**4.c**), and (**4.d**) indicate the stroma.

In addition, optical profilometry analysis revealed differences in surface topography between the groups (Figure 4). The control group (I) showed a visually more irregular surface, observed both in the two-dimensional projection (I-A) and in the three-dimensional reconstruction (I-B), whereas the group treated with O3 for 15 minutes (II) showed an apparently more homogeneous surface, observed in the two-dimensional projection (II-A) and in the three-dimensional reconstruction (II-B).

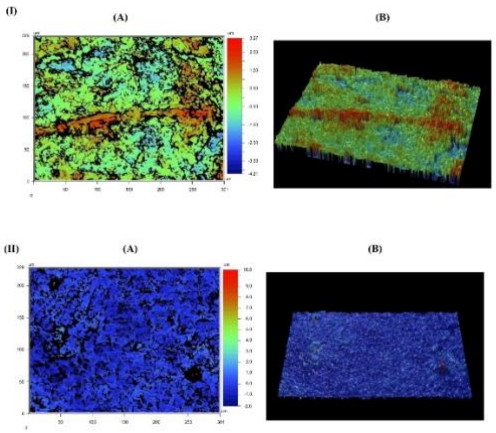


Figure 4. Optical profilometry of HAM: (**I**) in natura group and (**II**) group treated with O3 for 15 minutes; for each group, (**A**) two-dimensional projection and (**B**) three-dimensional reconstruction of the surface topography.

Table 1 presents the mean roughness (Ra) data for the epithelial face of HAM. The Ra value of the in natura sample was 0.626 (± 0.116 µm), while in the group treated with O3 the value was 0.509 (± 0.132 µm). However, this difference was not statistically significant ($p = 0.0835$), indicating only a trend toward reduced surface roughness after treatment. Together with the histological findings, these results suggest that the effects of O3 were mainly observed at the epithelial surface, without affecting the deeper layers of the membrane, with evidence of partial epithelial disorganization in the treated group. These data corroborate the study by Kawata et al. (2025), which used the same hydrodynamic system employed in this study, and reported that for significant alterations to occur in the HAM epithelium, higher O3 dosages are required, and consequently, longer exposure and ozonation times [7].

Tabela 1. Mean roughness (Ra) values obtained by optical profilometry in in natura samples and samples treated with O3 for 15 minutes (mean ± standard deviation). The significance level adopted was 5% ($p < 0.05$).

Parameter	in natura	O3	p-value
Ra (µm)	0,626 (± 0,116)	0,509 (± 0,132)	0,0835

IV. Conclusion

This study found that processing HAM for 15 minutes using a hydrodynamic system with ozonated water induced small alterations in the epithelial layer without affecting the deeper layers of HAM. Although a reduction in surface roughness was observed, the difference was not statistically significant. Therefore, under the conditions evaluated, the treatment resulted in epithelial alterations, but the findings are not sufficient to confirm effective de-epithelialization of the biomaterial. The use of samples from a single donor represents an additional limitation, and future studies should include a larger number of samples from different donors and evaluate different ozone exposure conditions to better assess the de-epithelialization potential of the method.

Acknowledgments

A. B. Fernandes and C.J. de Lima thank Instituto Ânima (AI), Universidade Anhembi Morumbi, São Paulo, SP, Brazil. A. Barrinha Fernandes thanks CNPq (National Council for Scientific and Technological Development) for the research productivity scholarship (Process No. 307989/2025-9). All authors thank Santa Casa de Misericórdia de Pindamonhangaba, SP, Brazil, and the Research and Technological Development Coordination of the National Institute for Space Research (COPDT/INPE).

References

[1] Law EJ, Taib H, Berahim Z. Amniotic Membrane: An Approach to Periodontal Regeneration. Cureus. 2022 Aug 9;14(8):e27832. doi: 10.7759/cureus.27832. PMID: 36106255; PMCID: PMC9458385.

[2] Fenelon M, Maurel DB, Siadous R, Gremare A, Delmond S, Durand M, Brun S, Catros S, Gindraux F, L'Heureux N, Fricain JC. Comparison of the impact of preservation methods on amniotic membrane properties for tissue engineering applications. Mater Sci Eng C Mater Biol Appl. 2019 Nov;104:109903. doi: 10.1016/j.msec.2019.109903. Epub 2019 Jun 24. PMID: 31500032.

[3] Ashraf M, Elhusseiny AM, Bouchard CS, Chodosh J, Djalilian AR, Saeed HN. Amniotic membrane in ophthalmology: Historical perspectives, biological properties, and clinical applications. Prog Retin Eye Res. 2026 Jul;113:101473. doi: 10.1016/j.preteyeres.2026.101473. Epub 2026 May 19. PMID: 42162708.

[4] Zhang T, Yam GH, Riau AK, Poh R, Allen JC, Peh GS, Beuerman RW, Tan DT, Mehta JS. The effect of amniotic membrane de-epithelialization method on its biological properties and ability to promote limbal epithelial cell culture. Invest Ophthalmol Vis Sci. 2013 Apr 30;54(4):3072-81. doi: 10.1167/iovs.12-10805. PMID: 23580491.

[5] Santos FDRP, Kawata BA, Oliveira Heinzelmann TR, Belfort MGS, Crispim de Oliveira Carvalho M, Móbille Awoyama S, Gomes de Oliveira Neto J, José de Lima C, Barrinha Fernandes A. De-Epithelization of the Human Amniotic Membrane Using a System Involving Ozonated Water and Ultrasound. Bioengineering (Basel). 2024 Sep 29;11(10):987. doi: 10.3390/bioengineering11100987. PMID: 39451363; PMCID: PMC11504975.

[6] Awoyama, S. M., Carvalho, H. C., de Souza Botelho, T., Dos Santos, S. I. S., Buendia Palacios, D. A., San Martín Henríque, S., … Fernandes, A. B. (2023). Disinfection of Human Amniotic Membrane Using a Hydrodynamic System with Ozonated Water. *Ozone: Science & Engineering*, *45*(1), 28–40. https://doi.org/10.1080/01919512.2021.2022452.

[7] Kawata, B. A.; et al. Effects of ozone sterilization in human amniotic membrane morphology: a promising de-epithelialization process. *Ozone Sci. Eng*. 2025, 1–12.

[8] Souza JSC, Laurindo Igreja Marrafa PA, de Carvalho MCO, Rodrigues MF, Akemi Kawata B, Martins GV, Ribeiro Dos Anjos EG, Passador FR, Fernandes AB, de Lima CJ. Evaluation of a Hydrodynamic Ozonation System for Gutta-Percha Disinfection and Structural Characterization. ACS Omega. 2026 Feb 9;11(7):11470-11481. doi: 10.1021/acsomega.5c08886. PMID: 41768645; PMCID: PMC12947183.

Analysis by Electron Microscopy of Morphological Changes in Calcium Hydroxyapatite Crystals Submitted to Microfocused Ultrasound *in vitro*

1st Gabriel Theodoro Olivieri
Biomedical Engineering Program UFRJ – Federal University of Rio de Janeiro RJ – Rio de Janeiro, Brazil
gtholivieri@peb.ufrj.br

2nd Paula Olimpio da Silva *Biomedical Engineering Program UFRJ – Federal University of Rio de Janeiro* RJ – Rio de Janeiro, Brazil
paulaolimpio@peb.ufrj.br

3rd Stela Freitas Raymundo *Biomedical Engineering Program UFRJ – Federal University of Rio de Janeiro* RJ – Rio de Janeiro, Brazil
stela.freitas@peb.ufrj.br

4th Aline Raybolt dos Santos Almeida
FO – Faculty of Dentistry – Department of Prosthesis and Dental Materials UFRJ – Federal University of Rio de Janeiro RJ – Rio de Janeiro, Brazil
0000-0001-6842-2179

5th Wagner Coelho de Albuquerque Pereira
Biomedical Engineering Program UFRJ – Federal University of Rio de Janeiro RJ – Rio de Janeiro, Brazil
0000-0001-5880-3242

Abstract*—Skin aging has driven increasing demand for minimally invasive aesthetic procedures, highlighting the combined use of energy-based technologies, such as Microfocused Ultrasound (MFU), and collagen biostimulators, such as Calcium Hydroxyapatite (CaHA). Although the association of these therapies is clinically promising for treating skin laxity and improving dermal quality, current protocols generally recommend intervals between applications due to the theoretical risk of filler material degradation by the ultrasound. This study aimed to evaluate, *in vitro*, the morphological and structural integrity of CaHA crystals when immediately subjected to MFU application. A controlled *in vitro* experimental study was conducted using an ex vivo biological model (porcine skin, Sus scrofa domesticus) and an ultrasonic phantom to reproduce the attenuation of the ultrasonic wave by the human dermis. Synthetic, dry, and non-commercial CaHA samples were divided into a control group (untreated) and an experimental group (treated with MFU using transducers with 3.0- and 4.5-mm focal depths). Analyses were performed using Scanning Electron Microscopy (SEM) to identify the occurrence of fragmentation, cracks, or alterations in crystal porosity. The results demonstrated that no visible changes were detected between the control samples and the samples subjected to MFU. Additionally, no visible differences were observed between the samples covered with porcine skin and those covered with the phantom. In conclusion, under the specific conditions evaluated, the thermal and mechanical energy of MFU did not produce additional morphological changes detectable by SEM in dry synthetic CaHA crystals, regardless of the protocol's frequency or covering medium. These findings represent an *in vitro* proof of concept and do not constitute evidence of clinical safety for same-day combined protocols.

Keywords*—Calcium hydroxyapatite; Microfocused Ultrasound; Facial Rejuvenation; In vitro Study.

I. INTRODUCTION

Skin aging is a multifactorial biological process characterized by progressive structural and functional changes in the epidermis, dermis, and subcutaneous tissue, resulting in reduced dermal thickness, loss of elasticity, changes in tissue support, and redistribution of facial fat compartments. These modifications compromise both skin quality and facial contours, driving the development of minimally invasive therapies capable of promoting rejuvenation with a high safety profile and limited recovery time.

Among these therapies, Microfocused Ultrasound (MFU) has established itself as a key energy-based technology for treating specific skin laxity [1]. The method involves focusing ultrasonic waves at predetermined depths, creating small zones of thermal coagulation (Thermal Injury Zones—TIZs) at temperatures sufficient to cause localized denaturation of collagen fibers and subsequent tissue remodeling [2]. This process triggers immediate tissue contraction and stimulates a prolonged biological response of neocollagenesis, without compromising the integrity of the epidermal surface.

Concurrently, calcium hydroxyapatite (CaHA) stands as one of the most widely used dermal biostimulators in contemporary aesthetic medicine [3], [4]. Composed of synthetic calcium hydroxyapatite microspheres suspended in a biodegradable carrier gel, its action combines an initial volumizing effect with the progressive stimulation of type I and III collagen synthesis, thereby improving skin quality and restoring tissue structure [5], [6]. Initially introduced for the treatment of facial lipoatrophy, its use subsequently expanded to various facial and body rejuvenation protocols, making it

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). fy applicable funding agency here. If none, delete this text box.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

one of the most extensively studied biomaterials for dermal biostimulation [7], [8].

In recent years, several clinical studies have demonstrated that combining MFU and CaHA can enhance aesthetic outcomes compared to the use of either technology alone [9], [10], [11]. While MFU promotes collagen reorganization through the controlled deposition of thermal energy in deep tissue planes, CaHA acts as a scaffold for the deposition of new collagen around its particles; this creates complementary mechanisms of action capable of simultaneously addressing skin laxity, tissue support, and dermal quality [12], [13], [14].

Despite favorable clinical outcomes, a significant gap remains regarding the direct interaction between ultrasonic energy and the biomaterial following implantation. In clinical practice, various protocols recommend performing MFU prior to CaHA application or maintaining an interval of weeks or months between procedures; this approach is based on the hypothesis that the mechanical and thermal energy generated by ultrasound could alter the physical integrity of the biostimulator particles and compromise their biological performance [15]. However, this recommendation remains based predominantly on theoretical considerations and indirect clinical evidence, with a scarcity of experimental studies directly investigating potential morphological changes in calcium hydroxyapatite following immediate exposure to MFU.

From a biomedical engineering perspective, characterizing this interaction is of great scientific importance. Validating the structural stability of CaHA when exposed to microfocused ultrasound supports the development of safer clinical protocols and contributes to the creation of experimental models designed to evaluate the interaction between injectable biomaterials and energy-based therapeutic technologies. Furthermore, the use of ultrasonic phantoms with acoustic properties similar to those of human skin enables the controlled replication of ultrasound propagation conditions, thereby reducing reliance on animal models and enhancing experimental reproducibility.

Given the high physicochemical stability of calcium hydroxyapatite as a bioceramic, it was hypothesized that the energy delivered by MFU clinical protocols would be insufficient to produce detectable morphological changes in the material's crystalline structure.

Thus, this study aimed to evaluate—using Scanning Electron Microscopy (SEM)—potential morphological changes in synthetic calcium hydroxyapatite crystals immediately after exposure to microfocused ultrasound (MFU). To replicate the acoustic propagation conditions encountered in clinical practice, two experimental overlay models were used—porcine skin (*Sus scrofa domesticus*) and an ultrasonic phantom [16] previously characterized as having acoustic properties compatible with human skin—thereby allowing for an investigation into the influence of the propagation medium on the biomaterial's structural integrity.

To date, no studies have been identified in the literature that directly evaluated the morphological integrity of CaHA crystals exposed to MFU using Scanning Electron Microscopy (SEM)—an approach that enables direct visualization of the mineral surface at the micrometer scale and the detection of structural alterations such as fragmentation, fissures, and changes in surface porosity.

II. MATERIALS AND METHODS

A. Materials

The experiment utilized dried samples of non-commercial synthetic calcium hydroxyapatite crystals (0.12 mg); porcine skin strips; a graphite-loaded ultrasonic phantom; a Microfocused ultrasound device (Pixie Pen V3 – By MedSystems); a Leica EM SCD050 sputter coater for sample preparation; and a ZEISS AURIGA 40 for scanning electron microscopy.

B. Sample Organization

Initially, the samples were divided into two main groups of 6 samples each, totaling 12 samples. The first group consisted of samples treated with the SMP 4-4.5 tip (4 MHz frequency, 4.5 mm focal depth, 0.9 J energy, and a rate of 7 pulses per second), designed to target the SMAS region for facial lifting. The second group consisted of samples treated with the DDP 7-3.0 tip (7 MHz frequency, 3.0 mm focal depth, 0.6 J energy, and a rate of 7 pulses per second), intended to induce thermal denaturation in the deep dermis to restructure the skin's deep support and enhance firmness.

Each subgroup within these two divisions contained 3 samples (0.1 mg each), resulting in a total of four sample divisions (i.e., two main groups and two subgroups). Within each subgroup, the samples were individually placed on a smooth acetate sheet featuring a standardized 10 mm x 10 mm grid pattern—arranged in parallel and sequential rows—to form a total sample area of 30 mm x 10 mm.

C. MFU exposure protocol

In the first group, the first set of samples was covered with a layer of ultrasound phantom material, with an area corresponding to the acetate markings and a thickness of 4.5 mm (30 mm x 10 mm x 4.5 mm). The second set of samples was covered with porcine skin, also with dimensions corresponding to the acetate surface markings and a thickness of approximately 4.3 mm (30 mm x 10 mm x 4.3 mm); (Fig. 1).



Fig. 1. MFU pulse delivery using an SMP 4-4.5 tip (4 MHz, 0.9 J, 7 pulses per second, 4.5 mm focal depth) on a porcine skin overlay.

Both coverings aimed to simulate the physical characteristics and reproduce certain acoustic properties of a strip of human skin approximately 4.5 mm thick—matching the target depth of the facelift handpiece near the Superficial Musculo-Aponeurotic System (SMAS), a region where

CaHA biostimulators are also used for structural support, bone projection, and facial contouring to correct age-related loss of bony support. The first sample in each sequence served as a control and was not treated; subsequently, the remaining four samples received 100 ultrasound pulses each from the SMP 4-4.5 handpiece using the specified settings, creating duplicates for each subgroup. The thicknesses of both the phantom and the animal skin were measured using analog calipers and a thickness gauge.

In the second group, the first sequence of three samples was also covered with the ultrasonic phantom—sharing the same specifications and ultrasonic characteristics but with a thickness of 3.0 mm (30 mm x 10 mm x 3.0 mm). As in the first group, the second sequence was covered with a strip of porcine skin, also approximately 3.0 mm thick (30 mm x 10 mm x 3.0 mm). These coverings aimed to mimic the physical characteristics of a strip of human skin at the deep-dermis depth for which the DDP 7-3.0 handpiece was designed, as well as to replicate some of its ultrasonic properties (Fig. 2).

This region also represents a significant depth for biostimulators, as it is the primary dermal level for stimulating neocollagenesis to provide skin support and firmness. As with the first group, both thicknesses were measured using a caliper and an analog thickness gauge. Also as with the first group, the initial samples in each sequence were not treated—serving as the control group—while the remaining four samples received 100 ultrasound pulses each. Both groups underwent the procedure on the same day.

Each sample received 100 ultrasound pulses, applied to an area of approximately 1 cm² of porcine skin and phantom material. Radial movements around a central point were required to deliver the shots, as the device does not fire while the applicator is stationary for safety reasons. The small experimental application area resulted in a density of 100 ultrasound pulses/cm², a value exceeding the average densities of 5.0–6.7 shots/cm² (SMP 4-4.5) and 2.5–3.3 shots/cm² (DDP 7-3.0) observed in routine practice [15]. Thus, the experimental protocol represented a deliberately more severe acoustic exposure scenario, to establish a conservative condition for evaluating the morphological integrity of the CaHA crystals.



Fig. 2. MFU pulse delivery with DDP 7-3.0 tip, 7 MHz, 0.6 J, 7 pulses per second, 3.0 mm focal depth, on a porcine skin covering.

Although the treatment tips serve different purposes, they are similar in dimensions—particularly regarding their firing diameter, which is 1 cm in both cases. Nevertheless, thermal contraction effects were observed in the treated skin samples when compared to those not exposed to ultrasound (Fig. 3). The general classification and grouping of the samples are outlined in Table 1.

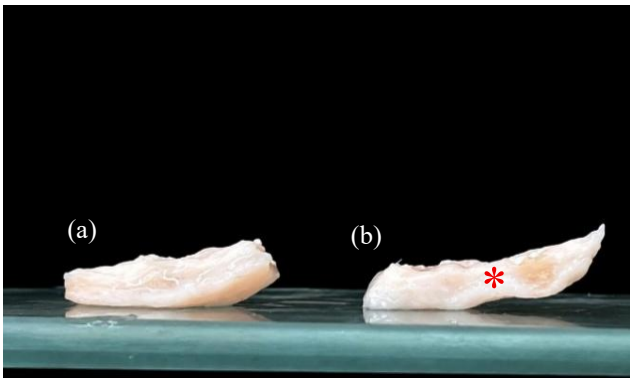


Fig. 3 Comparison of porcine skin samples: (a) sample not subjected to MFU firing; (b) sample subjected to firing with the SMP 4-4.5 transducer; (*) thermal contraction. In both samples, the epidermis faces downward, in contact with the glass surface.

TABLE I SAMPLE GROUPS

	Frequency [a]	*Coverage*	*Focus*	*Test group*	*Energy (J)*
1	4 MHz	phantom	4.5 mm	1	0.9 J
2	4 MHz	phantom	4.5 mm	1	0.9 J
3	7 MHz	phantom	3.0 mm	2	0.6 J
4	7 MHz	phantom	3.0 mm	2	0.6 J
5	4 MHz	skin	4.5 mm	1	0.9 J
6	4 MHz	skin	4.5 mm	1	0.9 J
7	7 MHz	skin	3.0 mm	2	0.6 J
8	7 MHz	skin	3.0 mm	2	0.6 J
9	-	phantom	-	1	-
10	-	phantom	-	1	-
11	-	skin	-	2	-
12	-	skin	-	2	-

[a] The pulse repetition rate was the same for all samples (7 PPS – pulses per second), as was the number of shots per sample (100 ultrasound pulses). The groups refer to the material coating the samples, and the subgroup refers to the focal spot of the tip used.

D. Characterization of the Ultrasound Phantom

The ultrasonic phantom used was developed by the Ultrasound Laboratory (LUS) at the Graduate School of Engineering of the Federal University of Rio de Janeiro (COPPE/UFRJ) and was originally designed to mimic the acoustic properties of human skin [16]. The model in question consists of 50 mL of Regular Plastic (plastic for making plastic worms; M-F Manufacturing Co., Inc., stock no. 2132LP, Fort Worth, Texas, USA)—polyvinyl chloride (PVC) monomers immersed in a plasticizer—and 1.0 g of graphite. Graphite was selected due to its high durability, physicochemical stability, resistance to biological degradation, and ease of handling—characteristics that make it a promising material for fabricating phantoms intended for biomedical ultrasound research. Additionally, the graphite acts as an ultrasonic wave scattering agent, allowing for the adjustment of the material's acoustic properties and producing behavior more similar to that observed in biological tissues [16].

The physical properties determined for the sample used were: thickness of 2.86 ± 0.29 mm, density of 0.94 ± 0.03 g/cm³, acoustic impedance of 1.40 ± 0.04 × 10^6 kg·m^{-2}·s^{-1}, attenuation coefficient of 1.11 ± 0.03 dB·cm^{-1}·MHz^{-1},

ultrasonic propagation speed of 1488.12 ± 0.01 m/s, specific heat of 1694.66 ± 6.90 $J \cdot kg^{-1} \cdot °C^{-1}$, and thermal conductivity of 0.24 ± 0.00 $W \cdot m^{-1} \cdot K^{-1}$ [16].

When compared to reference values described in the literature for human skin, the phantom demonstrated high similarity in acoustic parameters fundamental to ultrasonic wave propagation. The measured propagation speed (1488 m/s) was extremely close to the lower limit reported for human skin (1498–1624 m/s), while the attenuation coefficient (1.11 $dB \cdot cm^{-1} \cdot MHz^{-1}$) fell within the reference range described for this tissue (1.04–3.50 $dB \cdot cm^{-1} \cdot MHz^{-1}$). These results indicate that the material is capable of satisfactorily reproducing the interaction of the ultrasonic beam with the skin layer, particularly regarding phenomena such as absorption, scattering, and energy loss during propagation.

The use of ultrasonic phantoms in experimental studies is of great scientific importance, as it enables testing under controlled and reproducible conditions, free from the variability inherent in biological tissues. Furthermore, these materials allow for the standardization of experimental protocols, facilitating comparisons between different samples, equipment, and acquisition conditions. In the context of this study, the phantom served as a reference material for evaluating the behavior of measurements obtained from biological tissue.

The comparison between the phantom and the animal skin used in the *in vitro* experiments—specifically porcine skin—is particularly relevant because the skin represents the first interface encountered by ultrasonic waves upon entering the body. Consequently, its acoustic properties directly influence the transmission, reflection, absorption, and distribution of ultrasonic energy within the underlying tissues. Although porcine skin shares anatomical and physiological characteristics with human skin, variations in thickness, hydration, structural composition, and storage conditions can influence experimental results. In this regard, the use of a pre-characterized phantom provides a stable reference standard, making it possible to verify whether the responses observed in the biological tissue align with the expected behavior of a material that reproduces the acoustic properties of skin, as shown in the comparisons in Table II.

TABLE II COMPARATIVE VALUES BETWEEN THE ACOUSTIC PROPERTIES OF THE PHANTOM USED AND AVERAGE MEASUREMENTS OF HUMAN SKIN

Physical Property	***Unit***	***Phantom***	***Human Skin Reference Range***
Density (ρ)	g/cm^3	0.94 ± 0.03	1.09 – 1.20
Acoustic Impedance (Z)	$\times 10^6$ kg $\cdot m^{-2} \cdot s^{-1}$	1.40 ± 0.04	1.4-1.87
Propagation Velocity (c)	m/s	1488.12 ± 0.01	1498 - 1624
Attenuation Coefficient (α)	$dB \cdot cm^{-1} \cdot MHz^{-1}$	1.11 ± 0.03	1.04 – 3.5
Thermal Conductivity (k)	$W \cdot m^{-1} \cdot K^{-1}$	0.24 ± 0.00	0.23 – 0.50
Specific Heat (Cp)	$J \cdot kg^{-1} \cdot ° C^{-1}$	1694.66 ± 6.90[a]	3150 - 3660

[a] The specific heat of the phantom (1694.66 $J \cdot kg^{-1} \cdot K^{-1}$) falls within the range reported for dry human stratum corneum (1500–2200 $J \cdot kg^{-1} \cdot K^{-1}$) [20], [21]—the outermost layer of the epidermis, characterized by low water content—and is significantly lower than that of hydrated skin (3150–3660 $J \cdot kg^{-1} \cdot K^{-1}$).

E. Sample Processing

To enable microstructural characterization and reduce charging effects during microscopy analysis, the samples were subjected to a physical vapor deposition process via sputtering. The sputtering process aimed to deposit a thin gold layer onto the samples and was carried out using a Leica EM SCD050 automated coater equipped with an Adixen Pascal 2005 SD dual-stage mechanical pump. After positioning the sample holder inside the vacuum chamber, the system was evacuated until a safe residual pressure was reached.

Subsequently, the chamber was purged via 30 controlled injections of argon gas (analytical grade) using the "rinsing" function, with the aim of completely eliminating residual oxygen and stabilizing the working atmosphere.

Operational parameters were established by maintaining a constant pressure of 0.05 mbar and a fixed working distance of 50 mm between the metal target (Au) and the sample surface. Based on prior equipment optimization protocols designed to ensure film uniformity, the deposition process was conducted using a sputtering current of 40 mA for a total exposure time of 90 seconds. Upon completion of the programmed time, the plasma flow was interrupted by closing the shutter, and the reaction chamber was slowly pressurized to local atmospheric pressure prior to the removal of the metallized samples.

F. Scanning Electron Microscopy

Analysis of the surface morphology and dimensional assessment of the calcium hydroxyapatite crystals were performed using Field Emission Gun Scanning Electron Microscopy (FEG-SEM) with a ZEISS AURIGA dual-beam microscope (FIB-SEM CrossBeam; Carl Zeiss Microscopy GmbH, Germany). The analyses were conducted at the National Center for Structural Biology and Bioimaging (CENABIO/UFRJ).

Following gold coating, the sample stubs were mounted on the motorized stage and introduced into the high-vacuum analysis chamber. Images were acquired using an electron beam accelerating voltage (Electron High Tension – EHT) set to 2.00 kV and a working distance (WD) maintained between 4.8 mm and 5.2 mm, ensuring the necessary spatial resolution without causing thermal or structural damage to the material. Secondary electron detection for topographic mapping was performed using a conventional Everhart-Thornley (SE2) detector. Various magnification levels (ranging from 100x to 1.00 kX) were applied to obtain panoramic views of the distribution as well as to isolate and directly measure individual crystals/particles.

III. RESULTS

SEM images showed no visually detectable differences in the apparent morphology of the CaHA crystals, regardless of the experimental configuration tested. Among the 12 samples comprising the study cohort, the 8 samples subjected to MFU pulses showed no visually detectable differences compared with the 4 control samples. Consequently, no visually detectable evidence of immediate morphological degradation was observed.

A. Qualitative Characterization of the Control Group

Photomicrographs of control samples (samples 9, 10, 11, 12), which did not receive ultrasound shots, revealed that the CaHA crystals used as source have an intrinsically irregular and clastic (detrital) morphology. The presence of particles with varying dimensions and surfaces with pre-existing porosity and fragmentation was observed, characteristics inherent to the material in its dry state. Micrometric measurements confirmed the presence of crystals with major axes varying, for example, between 58.86 µm and 60.23 µm (Fig. 4). Observation of the conformation of the basal morphological profile was essential for comparison with the tested samples.

B. Evaluation of Experimental Groups

Samples subjected to 100 MFU shots were analyzed based on their technical classification into two subgroups defined by the frequency and focal depth of the ultrasonic transducer.

The subgroup subjected to 4 MHz shots (samples 1, 2, 5, and 6)—despite being exposed to a higher energy load (0.9 J) targeted at a depth of 4.5 mm (near the SMAS level)—showed no additional thermal or mechanical damage upon microscopic examination; furthermore, no new fissures, particle fusion, or microscopic patterns attributable to the ultrasonic beam were observed. Similarly, the subgroup subjected to 7 MHz shots (samples 3, 4, 7, and 8)—despite having a higher tissue attenuation rate in the deep dermis at 3.0 mm—also maintained apparent physical stability. The visual appearance of the crystals in these samples remained structurally indistinguishable from that of the control groups (Fig. 5).

C. Influence of Covering Media (Porcine Skin vs. Phantom)

No morphological or structural differences were detected between samples covered with porcine skin and those protected by the graphitized synthetic phantom. No morphological or structural differences were detected between samples covered with porcine skin and those protected by the graphitized synthetic phantom. The two covering media produced similar morphological outcomes under the experimental conditions evaluated; however, this similarity does not establish equivalence in the spatial distribution or dose of acoustic energy delivered to the crystals.

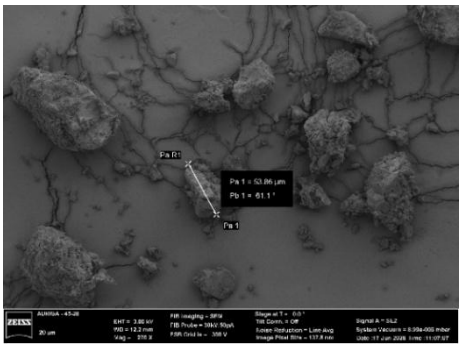


Fig. 4 Control sample (sample 10) at a 20-µm scale, with a reference particle measuring 53.86 µm, showing particles both smaller and larger than the reference particle in the same micrograph.

In summary, microscopy results confirmed that, regardless of the frequency used (4 MHz or 7 MHz), the applied energy density, or the covering material, the CaHA crystals maintained their apparent morphological integrity as assessed by surface SEM. No new fissures, fragmentation, or particle fusion were observed under the tested conditions. It should be noted, however, that the analysis is essentially qualitative and that the absence of visible changes via SEM does not imply clinical safety, as sub-micrometric or physicochemical alterations undetectable by this technique cannot be ruled out.

IV. DISCUSSION

The morphological integrity of CaHA crystals following exposure to MFU is a crucial factor for the validation of combined clinical protocols. SEM results demonstrated that, regardless of the frequency used or the covering material, there was no fragmentation, cracking, or alteration in the surface porosity of the crystals.

A. Thermal and Mechanical Stability of CaHA

The absence of structural damage can be explained by the intrinsic properties of CaHA as a bioceramic. While MFU operates by generating thermal injury zones (TIZs) with temperatures ranging from 65°C to 75°C to induce collagen denaturation and tissue contraction, the literature indicates that the thermal degradation of calcium hydroxyapatite occurs only at extremely high temperatures, with sintering points near 1000°C [2], [17].

Unlike other polymeric biostimulators such as poly-L-lactic acid (PLLA) or polycaprolactone (PCL)—which have glass transition temperatures close to the operating ranges of energy-based devices and may undergo deformation under thermal stress—CaHA exhibits superior thermal stability [17]. This serves as a benchmark for the intrinsic thermal stability of CaHA as a bioceramic. It is important to note, however, that the thermal effects of MFU are highly localized and transient, differing fundamentally from uniform bulk heating (bulk sintering). Therefore, the sintering temperature should not be interpreted as direct evidence of resistance to MFU, but rather as an indicator of the material's mineral robustness—which, combined with SEM findings, suggests that the energy delivered by MFU was insufficient to produce detectable morphological changes under the tested conditions.

B. Clinical Implications of the Order of Application.

Currently, there is a debate regarding the optimal treatment sequence: the application of MFU followed by CaHA on the same day is considered safe and effective for enhancing neocollagenesis [13]. The theoretical concern that ultrasound might fragment the biostimulator and accelerate its resorption in an unpredictable manner was not supported by the *in vitro* data presented here.

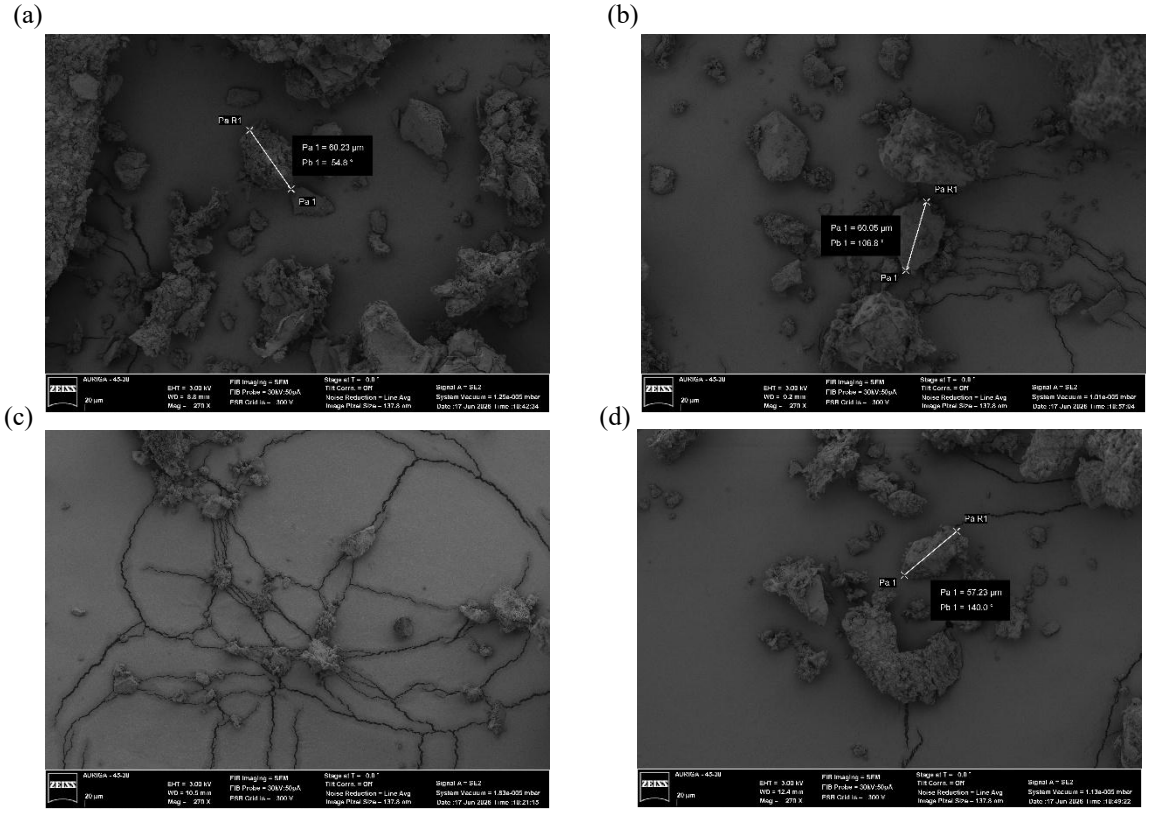


Fig. 5. (a) (sample 1) 20 µm scale with reference particle measured at 60.23 µm, 4.5 mm focal depth, 4 MHz, 0.9 J – Phantom covering. (b) (sample 5) 20 µm scale with reference particle measured at 60.05 µm, 4.5 mm focal depth, 4 MHz, 0.9 J – Phantom covering. (c) (sample 3) 20 µm scale, 3.0 mm focal depth, 7 MHz, 0.6 J – Animal skin covering. (d) (sample 7) 20 µm scale, reference particle measured at 57.23 µm, 3.0 mm focal depth, 7 MHz, 0.6 J – Animal skin covering.

Although some histological studies suggest that performing MFU first may "prime" the tissue and enhance subsequent interaction with CaHA particles, the present findings only indicate that direct MFU exposure did not produce additional morphological changes detectable by surface SEM under the tested conditions. Therefore, these results cannot be used to infer preservation of the biological function of CaHA particles after MFU exposure.

C. Validation of the Experimental Model and the Phantom

The phantom used showed an attenuation coefficient within the reported reference range for human skin, while its propagation speed was close to the lower limit of the reported range [16]. The attenuation coefficient is an important parameter influencing energy loss during propagation through the covering medium. Regarding this parameter, the phantom (1.11 $dB \cdot cm^{-1} \cdot MHz^{-1}$) fell within the reference range for human skin (1.04–3.50 $dB \cdot cm^{-1} \cdot MHz^{-1}$), indicating compatible attenuation. It should be noted, however, that other parameters—such as density, acoustic impedance, and specific heat—differ from the reference values for human skin (Table II); consequently, claims of complete equivalence to "*in vivo*" conditions are not supported by the available data.

It is important to note, however, that no homogeneous mimicking material can simultaneously reproduce all the characteristics of a stratified biological tissue like human skin [17], [18]. The suitability of a Tissue-Mimicking Material (TMM) is evaluated based on the specific application rather than an absolute match of all parameters [17], [18]. In this regard, the measured acoustic impedance (1.40×10^6 $kg \cdot m^{-2} \cdot s^{-1}$) falls within the range accepted for TMMs in the literature (1.4–1.7 MRayl) [17], [19], and the propagation speed (1488 m/s) differs by only 0.7% from the lower limit of the reference range, remaining within the ±5% tolerance reported in the literature [17], [18]. Regarding specific heat, the measured value (1694.66 $J \cdot kg^{-1} \cdot K^{-1}$) is consistent with the range reported for dry human stratum corneum (1500–2200 $J \cdot kg^{-1} \cdot K^{-1}$) [20], [21]—the outermost layer of the epidermis, characterized by low water content—and is lower than that of hydrated skin (3150–3660 $J \cdot kg^{-1} \cdot K^{-1}$). The phantom was therefore selected and validated based on its acoustic attenuation properties—the parameter most directly relevant to the energy reaching the crystals—without implying complete biological equivalence to *in vivo* human skin [22].

D. Crystal Morphology and Biological Response

It was observed that the CaHA samples used exhibited an irregular, clastic morphology rather than that of perfect microspheres. In the literature, commercial microspheres (such as Radiesse) feature smooth surfaces and diameters ranging from 25 to 45 µm to minimize chronic inflammatory reactions [23], [24], [25], [26]. The fact that crystals with more angular shapes withstood 100 ultrasound pulses without undergoing further fragmentation underscores the robustness of the mineral material. However, while it is tempting to infer those commercial microspheres—due to their more regular geometry and smooth surfaces—would offer an even greater margin of physical safety under the same MFU settings, this

inference remains speculative. Commercial CaHA/CMC microspheres differ from the crystals used in this study in terms of morphology, particle size distribution, and surface characteristics; therefore, direct studies using commercial formulations are required to validate this hypothesis.

E. Experimental Limitations

Despite the promising results, this study presents limitations inherent to the *in vitro* model. The use of dry, non-commercial crystals—while useful for observing the pure mineral structure—does not mimic, for example, the presence of a carrier gel (such as carboxymethylcellulose/CMC) found in some commercial products, or tissue hydration, both of which can influence thermal conduction [27]. Furthermore, the experiment focused on immediate physical integrity and did not evaluate potential long-term chemical changes or the biological inflammatory response following irradiation.

Moreover, as an *in vitro* model, the experiment does not account for the effect of blood perfusion, which plays a crucial role in heat distribution and dissipation within living tissues subjected to therapeutic or microfocused ultrasound.

V. CONCLUSION

Under the experimental conditions evaluated, immediate application of microfocused ultrasound did not produce additional morphological changes detectable by scanning electron microscopy in the dry synthetic calcium hydroxyapatite crystals. The 4- and 7-MHz MFU protocols were not associated with detectable additional fragmentation of the material under the tested conditions, regardless of whether the coating was of biological or synthetic origin.

These findings are consistent with the hypothesis regarding the physical feasibility of same-day combination protocols, indicating that, under the tested conditions the CaHA maintained its apparent morphological integrity. The validation of the ultrasonic phantom as a controlled experimental model contributes to future investigations into the interaction between energy-based technologies and injectable biomaterials.

The main takeaway from this study is that surface SEM, while adequate for detecting macroscopic morphological changes, is insufficient as a standalone method for characterizing subtle physicochemical modifications in mineral bioceramics. Key research gaps identified include: (1) studies using commercial CaHA/CMC formulations under hydrated and implanted conditions; (2) quantitative morphometric analysis combined with crystallographic (XRD) and spectroscopic (Raman, FTIR) characterization; (3) direct acoustic dosimetry of the energy delivered at the focal point in biological models; and (4) biological and in vivo validation of the safety of same-day combination protocols.

ACKNOWLEDGMENTS

The authors thank the Padrón Lins Multi-user Microscopy Unit at the Paulo de Góes Institute of Microbiology (UNIMICRO), located at the Health Sciences Center (CCS/UFRJ), for technical support and assistance with sample processing. We also thank the CENABIO team for their essential assistance with the microscopy analyses. Special thanks to Professor Marcelo Henrique Prado for providing the batch of calcium hydroxyapatite used in this study.

REFERENCES

[1] M. Contini, M. H. J. Hollander, A. Vissink, R. H. Schepers, J. Jansma, and J. Schortinghuis, “A systematic review of the efficacy of microfocused ultrasound for facial skin tightening,” Int. J. Environ. Res. Public Health, vol. 20, no. 2, p. 1522, Jan. 2023.

[2] W. M. White, I. R. S. Makin, P. G. Barthe, M. H. Slayton, and R. E. Gliklich, “Selective creation of thermal injury zones in the superficial musculoaponeurotic system using intense ultrasound therapy: A new target for noninvasive facial rejuvenation,” Arch. Facial Plast. Surg., vol. 9, no. 1, pp. 22–29, Jan. 2007.

[3] C. S. F. P. Oliveira et al., "Calcium hydroxyapatite: a review for efficacy, safety and imaging when used as a filler and as a biostimulator," *Res. Soc. Dev.*, vol. 10, no. 14, p. e05101421689, Oct. 2021, doi: 10.33448/rsd-v10i14.21689.

[4] S. B. Aguilera et al., “The role of calcium hydroxyapatite (Radiesse) as a regenerative aesthetic treatment: A narrative review,” Aesthet. Surg. J., vol. 43, no. 8, pp. 880–890, Aug. 2023.

[5] Merz Aesthetics, “Radiesse Package Insert,” Dec. 2023. [Online]. Available: https://www.patientinfo.merzaesthetics.com/products/download/52372/RADIESSE%C2%AE-1.5-PT-V01-01-DEC-2023.pdf [Accessed: Feb. 21, 2026].

[6] A. L. Berlin, M. Hussain, and D. J. Goldberg, “Calcium hydroxyapatite filler for facial rejuvenation: A histologic and immunohistochemical analysis,” Dermatol. Surg., vol. 34, no. s1, pp. S64–S67, Jun. 2008.

[7] N. Fakih-Gomez and J. Kadouch, “Combining calcium hydroxyapatite and hyaluronic acid fillers for aesthetic indications: Efficacy of an innovative hybrid filler,” Aesth. Plast. Surg., vol. 45, no. 4, pp. 1712–1721, Aug. 2021.

[8] G. Casabona and D. N. Teixeira, “Microfocused ultrasound in combination with diluted calcium hydroxyapatite for improving skin laxity and the appearance of lines in the neck and decolletage,” J. Cosmet. Dermatol., vol. 17, no. 1, pp. 66–72, Feb. 2018.

[9] M. Amiri et al., “Aesthetic efficacy and safety of combined microfocused ultrasound with visualization and calcium hydroxyapatite treatment: A systematic review of human evidence,” Aesthet. Surg. J., vol. 44, no. 3, pp. 290–302, Mar. 2024.

[10] M. Kerscher, A. T. Nurrisyanti, C. Eiben-Nielson, S. Hartmann, and J. Lambert-Baumann, “Clinical and biophysical outcomes of combining microfocused ultrasound with visualization and calcium hydroxyapatite filler for facial treatment,” Dermatol. Ther. (Heidelb), vol. 9, no. 1, pp. 135–142, Mar. 2019.

[11] S. Ramirez and I. B. K. Puah, “Effectiveness of combined microfocused ultrasound with visualization and subdermal calcium hydroxyapatite injections for the management of brachial skin laxity,” J. Cosmet. Dermatol., vol. 20, no. 12, pp. 3871–3879, Dec. 2021.

[12] G. Casabona, K. Kaye, S. Cotofana, K. Davidovic, M. Alfertshofer, and L. Freytag, "Histological effects of a combined collagen stimulation procedure consisting of microfocused ultrasound, soft tissue filler, and Ca-HA injections," *J. Cosmet. Dermatol.*, vol. 22, no. 6, pp. 1724–1730, Jun. 2023, doi: 10.1111/jocd.15770.

[13] K. Kaye, G. Casabona, K. Davidovic, M. Alfertshofer, L. Freytag, and S. Cotofana, "Pilot study: Single-depth superficial ultrasound with subdermal injections of diluted calcium hydroxyapatite for improving lower face skin quality," *Plast. Reconstr. Surg. Glob. Open*, vol. 12, no. 2, p. e5612, Feb. 2024, doi: 10.1097/GOX.0000000000005612.

[14] J. Corduff, "Treatment of mid-face aging with calcium hydroxyapatite: Focus on retaining ligament support," *Clin. Cosmet. Investig. Dermatol.*, vol. 17, pp. 2545–2553, Oct. 2024, doi: 10.2147/CCID.S479901.

[15] A. Doyle, I. Looi, and P. Chu, “Microfocused ultrasound with visualization (MFU-V) and hyperdilute calcium hydroxyapatite (CaHA-CMC) of the lower face and submentum to treat skin laxity: A pilot study demonstrating superiority of MFU-V first followed by hyperdilute CaHA-CMC,” Aesthet. Surg. J., vol. 44, no. 5, pp. 512–525, May 2024.

[16] S. F. Raymundo, "Produção e caracterização de phantoms com propriedades acústicas e térmicas da pele humana," M.S. thesis, Programa de Engenharia Biomédica, COPPE, Univ. Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2025.

[17] C. K. McGarry et al., "Tissue mimicking materials for imaging and therapy phantoms: a review," *Phys. Med. Biol.*, vol. 65, no. 23, Dec. 2020, doi: 10.1088/1361-6560/abbd17.

[18] A. Jawli, W. Aldehani, G. Nabi, and Z. Huang, "Tissue-mimicking material fabrication and properties for multiparametric ultrasound phantoms: a systematic review," *Bioengineering (Basel)*, vol. 11, no. 6, p. 620, Jun. 2024, doi: 10.3390/bioengineering11060620.

[19] L. E. Maggi, "Characterization of the acoustic properties of new materials with the objective of creating ultrasonic phantoms," Ph.D. thesis, COPPE, Univ. Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2011.
[20] S. Krishnan et al., "Multimodal epidermal devices for hydration monitoring," *Microsyst. Nanoeng.*, vol. 3, p. 17014, Jun. 2017, doi: 10.1038/micronano.2017.14.
[21] P. J. Rodríguez de Rivera et al., "In vivo measurement of skin heat capacity: advantages of the scanning calorimetric sensor," *J. Therm. Anal. Calorim.*, vol. 147, pp. 12155–12163, 2022, doi: 10.1007/s10973-022-11416-2.
[22] S. F. Raymundo, A. C. R. Silva, E. K. Saito, and W. C. A. Pereira, "Ultrasonic phantoms composed of PVCP and graphite are used to replicate the thermal characteristics of human skin," in *Proc. IEEE GMEPE/PAHCE*, 2026, doi: 10.1109/gmepe/pahce69420.2026.11550305.
[23] S. Cotofana, M. Alfertshofer, L. Freytag, K. Davidovic, K. Kaye, and G. Casabona, "A morphological analysis of calcium hydroxyapatite and poly-L-lactic acid biostimulator particles," *Skin Res. Technol.*, vol. 30, no. 2, p. e13764, Mar. 2024, doi: 10.1111/srt.13764.
[24] Merz Aesthetics, “Radiesse Instructions for Use,” Jan. 2025. [Online]. Available: https://www.ifu.merzaesthetics.com/products/download/53903/RADIESSE%C2%AE-(+)%20Lidocaine-PT-V01-01-JAN-2025.pdf [Accessed: Feb. 21, 2026].
[25] G. A. Sanchez Rico and S. B. Andrade Canto, “Three calcium hydroxyapatite-based dermal fillers marketed in Mexico: Comparison of particle size and shape using electron microscopy,” J. Cosmet. Dermatol., vol. 24, no. 1, p. e70100, Jan. 2025.
[26] M. de Moraes Nobre et al., “Characterization of particles from collagen bio-stimulating materials used in facial filling procedures,” Microsc. Res. Tech., vol. 88, no. 7, pp. 1935–1944, Jul. 2025.
[27] E. Karatas, K. Koc, M. Yilmaz, and H. M. Aydin, “Characterization and comparative investigation of hydroxyapatite/carboxymethyl cellulose (CaHA/CMC) matrix for soft tissue augmentation in a rat model,” ACS Omega, vol. 9, no. 29, pp. 31586–31600, Jul. 2024.

Analysis of the Cellular Response to Ti-29Nb-13Ta-4.6Zr Alloy Using Vero Cells for Bone Tissue Engineering

Andressa Francine Martins
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
São Bernardo do Campo, São Paulo, Brazil
andressa.francine@ufabc.edu.br

Cauê Machado Bueno
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
Santo André, São Paulo, Brazil

João Carlos Fernandes da Silva
Department of Bioengineering and iBB-Institute for Bioengineering and Biosciences
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal

Anibal de Andrade Mendes Filho
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
Santo André, São Paulo, Brazil

Christiane Bertachini Lombello
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
São Bernardo do Campo, São Paulo, Brazil

***Abstract*—β-metastable titanium alloys have emerged as promising alternatives to the most common titanium alloys used in medical devices, such as Ti-6Al-4V (Ti64), due to their lower toxicity, excellent corrosion resistance, and lower elastic modulus, more similar to cortical bone. This study aimed to evaluate the biological performance of the Ti-29Nb-13Ta-4.6Zr (TNTZ) alloy using Vero cells. The samples were characterized through direct contact cytotoxicity assays, extract cytotoxicity tests, cell viability assessment using the MTT assay, and scanning electron microscopy (SEM). The results demonstrated preservation of normal cell morphology, absence of cytotoxic effects, and cell viability above 70%. SEM analysis revealed satisfactory cell adhesion and spreading across the alloy surface. Overall, the findings indicate that the TNTZ alloy exhibits excellent biocompatibility and considerable potential for biomedical implant applications, representing a promising alternative to the conventional Ti64 alloy.**

***Index Terms*—Biomaterials; Bone tissue engineering; Cell adhesion; Metallic Biomaterials; Titanium; Titanium alloys.**

I. Introduction

By definition, biomaterials are materials designed to interact with biological systems for the purpose of evaluating, treating, augmenting, or replacing organs, tissues, or biological functions [1]. These materials must exhibit essential characteristics such as biocompatibility, ensuring that they do not induce significant adverse effects, including toxicity, inflammation, allergic reactions, or carcinogenicity [2]. These characteristics are essential to ensure a safe interaction with the host tissue and the long-term performance of medical devices. Furthermore, biomaterials should present suitable biodegradability or bioresorbability when required, be non-toxic, and possess surface characteristics compatible with their intended biomedical application. Therefore, the selection of an appropriate biomaterial is crucial to minimize undesirable physiological responses, such as local inflammation or infection, which may ultimately compromise systemic health [3], [4].

Bone tissue engineering has emerged as an effective strategy to restore skeletal function by promoting tissue regeneration while providing structural support and mechanical stability for the treatment of bone defects caused by trauma, tumors, infections, and congenital skeletal abnormalities, conditions that frequently require surgical intervention [5]. Metallic biomaterials play a fundamental role in stabilizing damaged bone structures and providing mechanical support, thereby reducing the risk of deformities, pain, and functional impairment. The chemical composition of these metallic devices can also be tailored according to patient-specific clinical requirements [5]. Commercially pure titanium (cp-Ti) and titanium alloys are among the most widely used metallic biomaterials for temporary and permanent implants because of their excellent mechanical performance, corrosion resistance, and biocompatibility. Moreover, there is the possibility of developing alloys with characteristics specifically optimized for each biomedical application [5]–[7], as modifications in alloy composition, thermomechanical treatments, and structural or chemical processing can significantly improve their intrinsic properties [4]–[6].

Among the recently developed β-type titanium alloys, Ti-29Nb-13Ta-4.6Zr (TNTZ) has attracted considerable attention because of its low elastic modulus, which reduces the stress-shielding effect while promoting osseointegration [6]. Furthermore, the alloying elements niobium (Nb), tantalum (Ta), and zirconium (Zr) provide excellent corrosion resistance and

Corresponding author: Andressa Francine Martins (andressa.francine@ufabc.edu.br)

exhibit high biological safety due to their non-toxic nature [6]. The development of this alloy addresses the need for an alternative to the conventional Ti-6Al-4V (Ti64) alloy, which contains aluminum and vanadium, elements that have been associated with potential long-term toxic effects [7]. Within this context, the present study aimed to evaluate the interaction between Vero cells and the β-metastable TNTZ titanium alloy to investigate its potential application as a biomaterial for bone tissue engineering and as a promising alternative to the commercially established Ti64 alloy.

II. Materials and methods

A. Sample Preparation

The TNTZ alloy was selected due to its non-toxic composition and its favorable mechanical and biocompatibility properties [6]. After forging, the samples were homogenized at 850 °C for 1 h under an argon atmosphere and subsequently water-quenched. The TNTZ specimens (8 mm in diameter and 30 mm in length) were then ground using silicon carbide (SiC) abrasive papers with grit sizes ranging from 120 to 3000, ultrasonically cleaned, and polished with a 0.6 μm silica suspension. Subsequently, the specimens were subjected to a single pass of Equal Channel Angular Pressing (ECAP) at 300 °C using Route A to refine the microstructure and enhance their mechanical properties [19].

B. Sample Sterilization

Prior to the in vitro experiments, the samples were washed with water and neutral detergent and sterilized. Moist heat sterilization was performed in an autoclave (Spencer) at 121 °C for 15 min [8].

C. Vero Cell Culture

Vero cells (São Paulo, Brazil; CCIAL-057) [9], an anchorage-dependent fibroblast-like cell line originally derived from the kidney epithelium of the African green monkey (*Cercopithecus aethiops*), were used. The cells were cultured in HAM-F10 medium (Sigma-Aldrich) supplemented with 10% fetal bovine serum (FBS; Cultilab), antibiotics (Cultilab), and maintained at 37 °C in a humidified atmosphere containing 5% CO_2 (Sanyo incubator). Cell morphology was routinely monitored using an inverted phase-contrast microscope (AxioVert A1, Zeiss). Trypan blue exclusion method was used before each cytotoxicity assay; only index with a minimum 90% viable cells were used in the subsequent experiments.

D. Direct Contact Cytotoxicity Assay

Direct contact cytotoxicity was an adapted protocol based on ISO 10993-5:2009, performed in triplicate. TNTZ alloy samples were evaluated; latex samples were used as the positive cytotoxic control, whereas Vero cells cultured under standard conditions served as the negative control. Vero cells were seeded in a 24-well culture plate at a density of 10^4 cells/mL directly onto the samples. After 24 h of incubation, cell morphology was examined using an inverted phase-contrast microscope (AxioVert A1, Zeiss).

E. Elution Cytotoxicity: morphology and MTT Assay

Elution media were prepared according to ISO 10993-12:2007, resulting in a surface area-to-volume ratio of 3 cm^2/mL for the TNTZ alloy samples, which were immersed in culture medium and incubated for 24 h at 37 °C. Simultaneously, Vero cells were seeded in 96-well plates at a density of 10^4 cells/mL and cultured until reaching confluence. Latex extracts were used as the positive cytotoxic control, while cells cultured under standard conditions without exposure to the biomaterial served as the negative control. All experiments were performed in triplicate. After 24 h, the culture medium was replaced with the previously prepared elution media, and the cells were incubated for an additional 24 h. Cell morphology was then evaluated using an inverted phase-contrast microscope (AxioVert A1, Zeiss). Subsequently, cell metabolic activity was assessed using the MTT assay. The elution media were removed, each well was washed with 100 μL of phosphate-buffered saline (PBS), and 100 μL of MTT solution (Sigma-Aldrich; 0.5 mg/mL) was added, followed by incubation for 2 h at 37 °C. The MTT solution was then discarded and replaced with 50 μL of dimethyl sulfoxide (DMSO; Synth). Absorbance was measured at 570 nm using a microplate reader (SpectraMax). The negative control was defined as representing 100% cell viability. Statistical analyses were performed using GraphPad Prism 5. Differences among experimental groups were analyzed using one-way analysis of variance (ANOVA) followed by Tukey's multiple comparison test, with statistical significance established at $p < 0.05$.

F. Scanning Electron Microscopy (SEM)

Samples intended for SEM were prepared according to previously established protocols. Cells interacting with TNTZ samples were fixed in 2.5% glutaraldehyde prepared in phosphate buffer. After rinsing with distilled water for 10 min, the samples were dehydrated in a graded ethanol series, dried using a critical point dryer (Leica EM CPD300), and sputter-coated with a 15-nm gold layer (Leica ACE200). Finally, the specimens were analyzed using a FEI Quanta 250 scanning electron microscope to evaluate cell adhesion, morphology, and spreading on the TNTZ alloy surface [11].

III. Results

A. Cytotoxicity Assessment

Cytotoxicity assays are performed to determine whether a biomaterial exerts toxic effects on living cells [12]. In the context of medical devices, such as the titanium alloy investigated in this study, these assays evaluate whether the material induces adverse cellular responses, including apoptosis or reduced cell viability, following the recommendations of ISO 10993-5 [13].

1) Direct Contact Cytotoxicity Assay: The direct contact cytotoxicity assay was carried out by culturing Vero cells directly on the surface of the TNTZ samples. Fig. 1 shows the morphology of Vero cells cultured under the positive control (latex) and negative control (standard culture conditions). As expected, the positive control exhibited a typical cytotoxic

response, characterized by severe morphological alterations and the absence of a continuous cell monolayer. In contrast, cells cultured under negative control conditions formed a confluent monolayer and maintained the characteristic fibroblast-like morphology of Vero cells, displaying an elongated shape with an oval nucleus [14]. Fig. 1 illustrates the morphology of Vero cells after 24 h of direct interaction with the TNTZ alloy; cells are attached to the material surface and preserved their characteristic morphology, showing no evidence of morphological damage or cytotoxic effects.

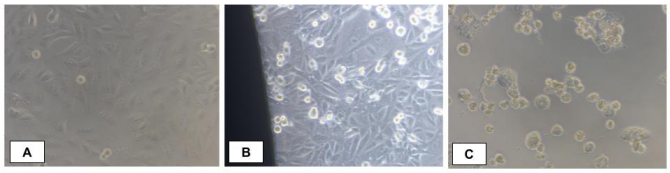


Fig. 1. Phase-contrast micrographs of Vero cells after 24 h of incubation with the control groups. (A) Negative control (standard culture conditions). (B) direct contact with TNTZ alloy. (C) Positive control (latex). Magnification: 400×.

2) Elution Cytotoxicity Assay: The elution cytotoxicity assay was performed to investigate the potential cytotoxic effects of soluble substances released from the TNTZ alloy. This method complements the direct contact assay. Vero cells after exposure to the elution media (Fig. 2) showed similar morphology to the direct contact assay; no morphological alterations were observed in cells exposed to the TNTZ elution media. The cell morphology closely resembled that of the negative control, whereas the positive control exhibited the expected cytotoxic characteristics.

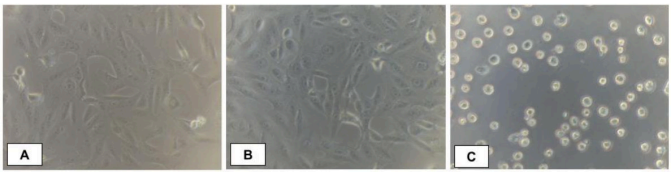


Fig. 2. Phase-contrast micrographs of Vero cells after exposure to TNTZ alloy extracts. (A) Negative control (standard culture conditions). (B) direct contact with TNTZ alloy. (C) Positive control (latex). Magnification: 400×.

a) MTT Assay: Cell metabolic activity was quantitatively evaluated by the MTT assay. Assuming normal data distribution and homogeneity of variances, the ANOVA generated an F-statistic comparing the variability among the TNTZ alloy, negative control, and positive control groups. No statistically significant differences were observed between the TNTZ alloy and the negative control ($p > 0.05$), indicating that exposure to the alloy did not negatively affect cellular metabolic activity. As shown in Fig. 3, cell viability remained above 70% relative to the negative control, satisfying the cytocompatibility criteria established by ISO 10993-5. The TNTZ alloy exhibited approximately 95% cell viability in comparison to the negative control, demonstrating the absence of cytotoxic effects and confirming its excellent *in vitro* biocompatibility, consistent with the morphological observations obtained in the cytotoxicity assays. The positive control presented 6.5% cell viability, statistically distinct from TNTZ alloy, and characteristic of cytotoxic response.

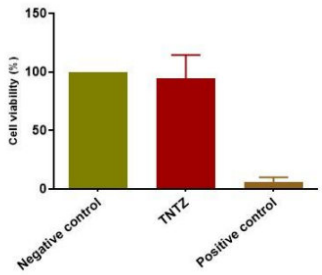


Fig. 3. Cell viability (%) determined by the MTT assay relative to the negative control.

B. Scanning Electron Microscopy (SEM)

SEM was performed to investigate the interaction between Vero cells and the surface of the β-metastable TNTZ alloy after 24 h of culture (Fig. 4). Cells were observed firmly attached to the alloy surface and distributed across the substrate. Their elongated and flattened morphology was consistent with the characteristic morphology of Vero cells. It was observed cytoplasmic extensions and intimate contact between the cells and the material surface, suggesting the establishment of early adhesion sites. Moreover, the cell spreading indicates favorable cell-material interactions, provided by the TNTZ surface for cellular attachment [15].

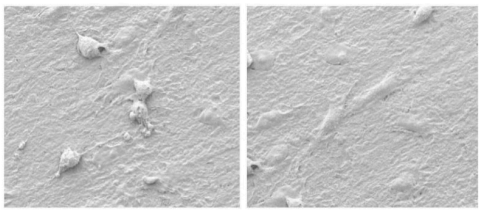


Fig. 4. Scanning electron micrographs showing the adhesion and spreading of Vero cells on the surface of TNTZ alloy specimens after 24 h of culture.

IV. Discussion

The development of novel titanium alloys for biomedical applications has been driven by the limitations of conventional alloys, particularly Ti64. Although Ti64 exhibits excellent mechanical strength and corrosion resistance, concerns regarding the long-term biological effects of aluminum and vanadium have stimulated the development of alternative titanium alloys

containing biologically safe alloying elements such as niobium, tantalum, and zirconium. In addition, β-type titanium alloys exhibit an elastic modulus closer to that of cortical bone, reducing the stress-shielding effect and promoting a more physiological distribution of mechanical loads at the bone–implant interface [10], [16].

The biological evaluation performed in this study demonstrated promising results for the β-metastable TNTZ alloy. The direct contact cytotoxicity assay showed that Vero cells maintained their characteristic fibroblast-like morphology, exhibited high confluence, and formed a continuous monolayer comparable to that observed in the negative control. In contrast, the positive control displayed severe morphological alterations and loss of monolayer formation, confirming the expected cytotoxic response. The preservation of normal cellular morphology following direct contact with the TNTZ alloy provides initial evidence of its biocompatibility.

The elution cytotoxicity assay further corroborated these findings. Exposure of Vero cells to the alloy extracts did not induce detectable morphological alterations, indicating that soluble/leachable released from the material did not exert adverse biological effects. This assay is particularly relevant because it simulates the release of metallic ions under physiological conditions, providing an indirect assessment of the chemical stability of the biomaterial. The absence of cytotoxic responses suggests that the TNTZ alloy remained chemically stable throughout the experimental period [17]. Quantitative evaluation by the MTT assay confirmed the qualitative observations obtained through optical microscopy. The TNTZ alloy exhibited approximately 95% cell viability, demonstrating that exposure to the material did not compromise mitochondrial metabolic activity. According to ISO 10993-5, materials exhibiting cell viability above 70% are considered non-cytotoxic. Therefore, the results obtained in the present study clearly demonstrate the excellent *in vitro* biocompatibility of the TNTZ alloy and reinforce its suitability for biomedical applications [15]. It is important to highlight that cytotoxicity was assessed using Vero cell lineage, a kidney epithelial cell with fibroblast-like morphology, and as such this initial biological evaluation screening does not provide information on the osteogenic behavior. Further analyses will be conducted in order to achieve a characterization of bone cells interacting with the materials.

Scanning electron microscopy provided additional evidence supporting the biological compatibility of the alloy. SEM images revealed numerous cells firmly adhered to the material surface, displaying an elongated and flattened morphology accompanied by extensive cytoplasmic projections. The presence of focal adhesion sites indicates the establishment of stable interactions between the cells and the biomaterial, representing an essential prerequisite for subsequent cellular proliferation and tissue integration [10]. The favorable cellular response observed in this study can be attributed to both the chemical composition and the microstructural characteristics of the TNTZ alloy. Previous studies have demonstrated that β-type titanium alloys containing niobium exhibit enhanced cell adhesion, proliferation, and metabolic activity compared with conventional titanium alloys, primarily because of their superior chemical stability and reduced release of potentially cytotoxic ions [15], [17].

Another important aspect concerns the thermomechanical processing of the alloy. The specimens evaluated in this study were subjected to ECAP followed by homogenization heat treatment. ECAP is well recognized for producing ultrafine-grained microstructures capable of improving both the mechanical and surface properties of metallic biomaterials. Although surface roughness and grain refinement were not directly investigated in the present work, these microstructural modifications may have contributed to the favorable biological behavior observed [10], [15], [19]. Overall, the present findings are consistent with recent literature identifying β-metastable titanium alloys as one of the most promising classes of materials for the next generation of metallic implants [17]. Their combination of excellent corrosion resistance, lower elastic modulus, and superior biological safety makes these alloys attractive alternatives to Ti-6Al-4V for orthopedic and dental applications [18]. Nevertheless, despite their intrinsic biocompatibility, further optimization of their surface properties through biofunctionalization or surface modification strategies may further enhance osseointegration and improve long-term clinical performance [10], [15], [17], [18].

V. Conclusions

The β-metastable TNTZ alloy demonstrated biological behavior consistent with the fundamental requirements for metallic biomaterials intended for medical device applications. Direct contact cytotoxicity, extract cytotoxicity, and MTT assays revealed no evidence of cytotoxic effects, while scanning electron microscopy confirmed satisfactory cell adhesion and spreading on the alloy surface. Furthermore, the results demonstrated that the TNTZ alloy maintains cellular metabolic activity and preserves the characteristic morphology of Vero cells, indicating excellent *in vitro* biocompatibility. The stable interaction established between the cells and the biomaterial suggests that the alloy provides a favorable surface for cellular attachment. The conclusions refer to data obtained with Vero cells assays; further analyses will be conducted with osteogenic cells.

Acknowledgment

The authors would like to thank the Laboratory of Tissue Engineering and Biomolecules of UFABC, and the Laboratory of Advanced Metallic Materials and Phase Transformations of UFABC. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] A. L. R. Pires, A. C. K. Bierhalz, and Â. M. Moraes, "Biomateriais: Tipos, Aplicações e Mercado," Química Nova, vol. 38, no. 7, pp. 957–971, Aug. 2015.

[2] R. A. Bapat et al., "Recent update on potential cytotoxicity, biocompatibility and preventive measures of biomaterials used in dentistry," Biomater. Sci., vol. 9, pp. 3244–3283, 2021, doi: 10.1039/D1BM00233C.

[3] S. B. T. Silva et al., "Surgical reconstruction using biomaterials: An integrative literature review," Res. Soc. Dev., vol. 10, no. 12, 2021.

[4] J. A. Santana, "Desenvolvimento de revestimentos híbridos aplicados aos substratos de Ti-6Al-4V para aplicações biomédicas," M.S. dissertation, Universidade Federal de Sergipe, São Cristóvão, Brazil, 2019.

[5] P. Pesode and S. Barve, "A review—Metastable β titanium alloy for biomedical applications," J. Eng. Appl. Sci., vol. 70, no. 25, 2023.

[6] G. L. G. S. B. Kumar, S. Chinara, S. Das, C. Tiwari, and G. V. S. N. Rao, "Studies on Ti-29Nb-13Ta-4.6Zr alloy for use as a prospective biomaterial," Mater. Today Proc., vol. 15, pp. 11–20, 2019, doi: 10.1016/j.matpr.2019.05.018.

[7] P. Bocchetta, L.-Y. Chen, J. D. C. Tardelli, A. C. Reis, F. Almeraya-Calderón, and P. Leo, "Passive Layers and Corrosion Resistance of Biomedical Ti-6Al-4V and β-Ti Alloys," Coatings, vol. 11, no. 5, p. 487, 2021.

[8] V. A. Pereira Filho et al., "Efeito da esterilização em autoclave sobre propriedades mecânicas das miniplacas de titânio utilizadas em fixação interna," Salusvita, vol. 20, no. 1, pp. 133–145, 2001.

[9] A. R. Esposito et al., "Estudo da interação células Vero/PLGA após a modificação da superfície por plasma de oxigênio," Matéria, vol. 12, no. 1, pp. 164–172, 2007.

[10] J. V. Calazans Neto, C. A. S. Celles, C. S. A. F. de Andrade, C. R. M. Afonso, B. E. Nagay, and V. A. R. Barão, "Recent Advances and Prospects in β-type Titanium Alloys for Dental Implants Applications," ACS Biomater. Sci. Eng., 2024, doi: 10.1021/acsbiomaterials.4c00963.

[11] J. J. Bozzola and L. D. Russell, "Preparation of biological specimens for scanning electron microscopy," Methods Mol. Biol., vol. 369, pp. 19–34, 2007.

[12] D. O. G. Côrrea, "Citotoxicidade de novas ligas à base de titânio visando aplicações biomédicas," M.S. dissertation, Universidade Estadual Paulista, Bauru, Brazil, 2022.

[13] V. Borelli, F. Medeiros, M. Maia, R. Medeiros, and O. Z. Higa, "Biocompatibilidade de Biomateriais: Estudo exploratório sobre a aplicação da norma ISO 10993," in Proc. 6th Latin American Congress of Artificial Organs and Biomaterials (COLAOB), 2014.

[14] A. Cremasco, A. D. Messias, A. R. Esposito, E. A. R. Duek, and R. Caram, "Effects of alloying elements on the cytotoxic response of titanium alloys," Mater. Sci. Eng. C, vol. 31, no. 5, pp. 833–839, 2011.

[15] A. Bandyopadhyay, I. Mitra, S. B. Goodman, M. Kumar, and S. Bose, "Improving Biocompatibility for Next Generation of Metallic Implants," Prog. Mater. Sci., 2023, doi: 10.1016/j.pmatsci.2022.101053.

[16] D. Solano-Orrala et al., "Advances in β-titanium alloys for safer and greener biomedical implants," Biomater. Adv., 2026, doi: 10.1016/j.bioadv.2026.214755.

[17] K. K. Wong, H. C. Hsu, S. C. Wu, and W. F. Ho, "A Review: Design from Beta Titanium Alloys to Medium-Entropy Alloys for Biomedical Applications," Materials, vol. 16, 2023, doi: 10.3390/ma16217046.

[18] G. Kanapaakala and V. Subramani, "A review on β-Ti alloys for biomedical applications: The influence of alloy composition and thermomechanical processing on mechanical properties, phase composition, and microstructure," Proc. Inst. Mech. Eng. H, vol. 237, no. 6, pp. 645–668, 2023.

[19] T. G. Langdon, "The principles of grain refinement in equal-channel angular pressing," Mater. Sci. Eng. A, vol. 462, no. 1–2, pp. 3–11, 2007.

Avaliação da membrana amniótica humana por espectroscopia Raman após diferentes períodos de armazenamento em glicerol: um estudo preliminar

Milene da Silva Melo
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, SP, Brasil
Núcleo de Saúde – Universidade de Mogi das Cruzes (UMC)
Mogi das Cruzes, SP, Brasil
milene_melo@yahoo.com.br

Letícia Cristine de Siqueira Santos
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, SP, Brasil
leticiacris.s.s@hotmail.com

Landulfo Silveira Júnior
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, SP, Brasil
landulfo.silveira@gmail.com

Henrique Cunha Carvalho
Centro de Inovação, Tecnologia e Educação (CITÉ)
São José dos Campos, SP, Brasil
Universidade Tecnológica Federal do Paraná (UTFPR)
Campo Mourão, PR, Brasil
hccarvalho@utfpr.edu.br

Sara Evelyn Batista Aquino
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, SP, Brasil
saraevelyn749@gmail.com

Renato Amaro Zângaro
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, SP, Brasil
razangaro@gmail.com

Adriana Barrinha Fernandes Moretti
Centro de Inovação, Tecnologia e Educação (CITÉ)
Universidade Anhembi Morumbi (UAM)
São José dos Campos, SP, Brasil
fernandesabm@gmail.com

***Resumo*— A membrana amniótica humana (MAH) é amplamente utilizada na medicina regenerativa devido às suas propriedades biológicas e terapêuticas. Entretanto, os efeitos do armazenamento sobre sua composição molecular ainda não são completamente compreendidos. O objetivo deste estudo foi avaliar a MAH por espectroscopia Raman após diferentes períodos de armazenamento em glicerol. Uma amostra de MAH foi analisada na condição *in natura* e após 2, 14, 35 e 68 dias de armazenamento em glicerol a −20 °C. Para cada condição experimental, foram adquiridos cinco espectros Raman em diferentes regiões da mesma amostra, os quais foram submetidos à correção da linha de base, normalização pela área sob a curva e cálculo do espectro médio. Observou-se similaridade entre os perfis espectrais da MAH *in natura* e após dois dias de armazenamento. Alterações espectrais tornaram-se evidentes após 14 dias, enquanto os espectros obtidos após 35 e 68 dias apresentaram características semelhantes entre si e distintas das demais condições experimentais, sugerindo a presença de um padrão espectral associado ao armazenamento prolongado. As principais bandas observadas foram compatíveis com os principais componentes da matriz extracelular da MAH e com as bandas Amida I e Amida III. Os resultados sugerem o potencial da espectroscopia Raman como ferramenta não destrutiva para o monitoramento de alterações espectrais na MAH durante o armazenamento em glicerol. Por se tratar de um estudo preliminar, análises quantitativas e multivariadas encontram-se em desenvolvimento para confirmação dos achados observados.**

Palavras-chave*— *membrana amniótica humana, espectroscopia Raman, armazenamento, glicerol.

I. INTRODUÇÃO

A membrana amniótica humana (MAH) é uma estrutura derivada da camada interna da placenta, composta por uma camada epitelial simples, lâmina basal e estroma rico em colágeno, elastina, fibronectina, laminina, ácido hialurônico, além de diversos fatores de crescimento [1-4]. Desempenha papel essencial durante a gestação, contribuindo para a troca de nutrientes, modulação da resposta imunológica e proteção mecânica durante o desenvolvimento embrionário e fetal. Devido à importância dessas funções, a MAH tem sido amplamente estudada na medicina regenerativa por suas propriedades anti-inflamatórias, antimicrobianas e de reparo tecidual. Diversos estudos demonstram sua eficácia biológica nas áreas de engenharia de tecidos, dermatologia e oftalmologia, sendo que, nesta última, seu uso terapêutico está consolidado [1-6]. Embora as propriedades terapêuticas da MAH sejam reconhecidas, seu armazenamento para uso posterior ainda representa um desafio. Um método comumente empregado para esta finalidade é a imersão em glicerol, pois segundo a literatura, este composto preserva suas características biológicas, estruturais e bioquímicas por um determinado período [7,8]. No entanto, apesar da ampla utilização deste método, sua eficácia ainda não está completamente elucidada. A Espectroscopia Raman se destaca como uma técnica promissora para a análise e caracterização da MAH, pois permite a investigação estrutural e química de materiais biológicos de forma não destrutiva, fornecendo informações detalhadas sobre sua composição

molecular [9-11]. Diante deste contexto, o objetivo do presente estudo foi avaliar a MAH por espectroscopia Raman após diferentes períodos de armazenamento em glicerol.

II. MATERIAIS E MÉTODOS

A. *Aspectos Éticos e Obtenção da MAH*

Uma amostra de MAH foi obtida de uma doadora que concordou em participar da pesquisa e assinou o Termo de Consentimento Livre e Esclarecido (TCLE). A voluntária foi submetida à cesariana na Santa Casa de Misericórdia de Pindamonhangaba, SP, Brasil, respeitando-se as normas técnicas para bancos de sangue de cordão umbilical e placentário, em conformidade com a Resolução RDC nº 190/2003 [12].

A amostra de MAH foi obtida assepticamente e de acordo com o protocolo aprovado pelo Comitê de Ética em Pesquisa da Universidade Anhembi Morumbi (Parecer nº 3.984.423; CAAE: 289370020.0.0000.5492). A doadora apresentou gestação sem intercorrências, sem antecedentes médicos ou sociais que contraindicassem a doação, além de sorologia negativa para HIV, hepatites B e C, HTLV, doença de Chagas, sífilis, citomegalovírus e toxoplasmose.

B. *Aquisição dos Espectros Raman e Armazenamento da Membrana*

A amostra de MAH foi lavada repetidas vezes com solução fisiológica de cloreto de sódio (NaCl) a 0,9% até completa remoção de resíduos sanguíneos [8]. Após a lavagem, um fragmento da amostra de MAH, na condição *in natura*, foi posicionado em um suporte de alumínio e submetido à aquisição de espectros Raman. Para esta finalidade, foi utilizado um Espectrômetro Raman Dispersivo (Dimension P-1, Lambda Solutions Inc, MA, EUA), com comprimento de onda de excitação de 830 nm e potência de saída de 350 mW. Para cada espectro, empregou-se tempo de integração de 4 segundos por *frame* e 10 acumulações, totalizando 40 segundos de aquisição. Em seguida, o restante da MAH foi armazenado em frasco de vidro contendo glicerol (Glicerina P.A ACS PM: 92,10 – Dinâmica®) sob refrigeração a −20 ºC. Após 2, 14, 35 e 68 dias, novos fragmentos foram obtidos, lavados 10 vezes com solução fisiológica de NaCl a 0,9% [13] para remoção do excesso de glicerol e submetidos à aquisição de espectros Raman para análise comparativa.

Para cada condição experimental, incluindo a condição *in natura*, foram adquiridos cinco espectros Raman. A cada aquisição, o suporte contendo o respectivo fragmento era reposicionado, de modo que os espectros fossem obtidos em regiões distintas. Esses espectros constituíram, portanto, medidas da mesma amostra de MAH e não replicatas biológicas independentes. Essas medidas foram utilizadas para a obtenção do espectro médio de cada condição experimental, tendo em vista a heterogeneidade natural do tecido.

C. *Processamento dos Espectros*

Os espectros obtidos foram exportados para o software MATLAB®, onde foram submetidos ao pré-processamento, incluindo correção da linha de base por ajuste polinomial de sexta ordem, seguida de normalização pela área sob a curva. Posteriormente, os espectros foram exportados para o software Microsoft Excel® onde foi calculado o espectro médio das cinco aquisições obtidas para cada período experimental, utilizado nas análises comparativas. As análises foram realizadas na faixa espectral de 400 a 1740 cm^{-1}.

III. RESULTADOS E DISCUSSÃO

A. *Análise Qualitativa dos Espectros Raman*

A Fig. 1 apresenta os espectros Raman médios da MAH *in natura* e após 2, 14, 35 e 68 dias de armazenamento em glicerol. Observa-se semelhança entre os perfis espectrais da MAH *in natura* e após dois dias de armazenamento, com manutenção das principais características espectrais. Após 14 dias, observam-se alterações no formato de algumas bandas e na intensidade relativa de determinadas regiões espectrais quando comparadas aos períodos anteriores, sugerindo alterações na composição molecular detectáveis por espectroscopia Raman. Nos períodos de 35 e 68 dias, os espectros apresentaram características similares entre si, indicando a manutenção de um perfil espectral semelhante entre esses períodos de armazenamento. Entretanto, observam-se diferenças em relação aos espectros obtidos *in natura* e após dois dias de armazenamento, principalmente quanto ao formato das bandas e à intensidade relativa de determinadas regiões espectrais.

As semelhanças e diferenças espectrais entre os diferentes períodos de estudo podem ser melhor visualizadas na Fig. 1(b), na qual foi aplicado deslocamento vertical dos espectros (*offset*).

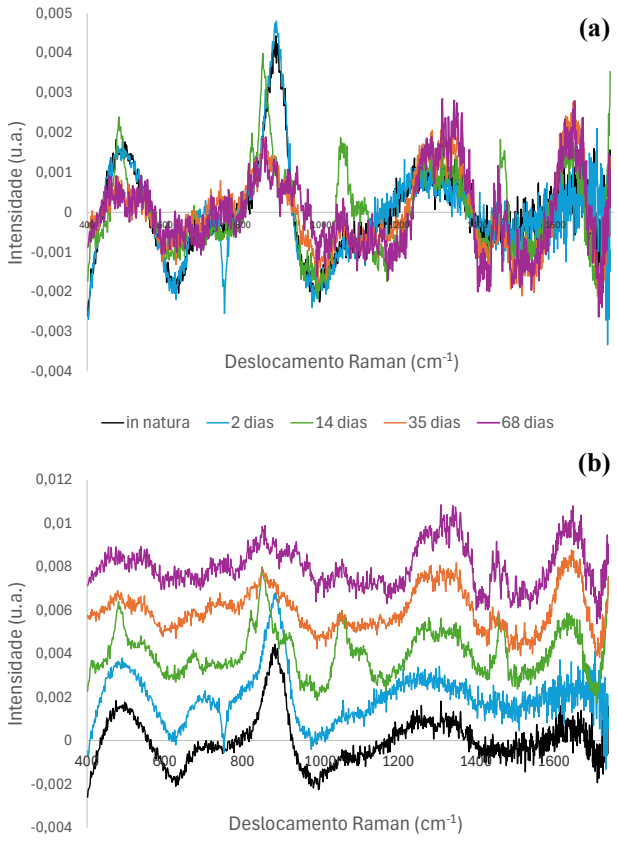


Fig. 1. Espectros Raman médios da membrana amniótica humana (MAH) *in natura* e após 2, 14, 35 e 68 dias de armazenamento em glicerol. (a) Espectros sobrepostos. (b) Espectros apresentados com deslocamento vertical (*offset*) para facilitar a visualização das diferenças espectrais. Os espectros foram submetidos à correção da linha de base e normalizados pela área sob a curva.

Esses resultados sugerem manutenção do perfil espectral nos dois primeiros dias de armazenamento, seguida por

mudanças espectrais detectáveis após 14 dias. Os espectros obtidos após 35 e 68 dias apresentaram características espectrais semelhantes entre si e distintas daquelas observadas na condição *in natura* e após 2 e 14 dias, sugerindo a presença de um padrão espectral distinto associado ao armazenamento prolongado, cuja confirmação requer análises quantitativas e/ou multivariadas, as quais estão atualmente em andamento.

A fim de explorar os componentes moleculares associados às alterações observadas, foi realizada a identificação das principais bandas Raman presentes nos espectros.

B. Identificação das Bandas Raman

A Tabela I apresenta as principais regiões espectrais observadas na análise dos espectros Raman da MAH.

TABELA I - PRINCIPAIS REGIÕES ESPECTRAIS OBSERVADAS NOS ESPECTROS DA MAH E SUAS POSSÍVEIS ATRIBUIÇÕES ESPECTRAIS COM BASE NA LITERATURA

Região observada na MAH	**Máximo aproximado**	**Possível atribuição espectral**	**Referência**
400 – 600 cm^{-1}	~490 cm^{-1}	Colágeno, elastina, fibronectina, ácido hialurônico.	Wang et al., 2023.
800 – 960 cm^{-1}	~890 cm^{-1}	Colágeno, elastina, ácido hialurônico; glicina presente no colágeno.	Wang et al., 2023; Cárcamo et al., 2012.
1200 – 1400 cm^{-1}	~1260 cm^{-1}	Amida III; colágeno, elastina	Bergholt et al., 2019; Wang et al., 2023; Milyudin et al., 2023; Cárcamo et al., 2012.
1540 – 1700 cm^{-1}	~1640 cm^{-1}	Amida I; colágeno, elastina, fibronectina, ácido hialurônico.	Bergholt et al., 2019; Wang et al., 2023; Milyudin et al., 2023; Cárcamo et al., 2012.

Foram observadas bandas nas regiões de aproximadamente 400 – 600 cm^{-1}, 800 – 960 cm^{-1}, 1200 – 1400 cm^{-1} e 1540 – 1700 cm^{-1}, com intensidades máximas em torno de 490 cm^{-1}, 890 cm^{-1}, 1260 cm^{-1} e 1640 cm^{-1}, respectivamente. Em espectros Raman obtidos de biomoléculas isoladas, bandas entre 530 e 560 cm^{-1} e entre 1640 e 1665 cm^{-1} foram associadas ao colágeno, elastina, fibronectina e ácido hialurônico, enquanto bandas entre 800 e 940 cm^{-1} e entre 1252 e 1260 cm^{-1} foram associadas principalmente ao colágeno e à elastina [11]. Essas regiões têm sido associadas a componentes da matriz extracelular e são compatíveis com a composição bioquímica da MAH.

De acordo com Bergholt et al. [10], as bandas observadas nas regiões de aproximadamente 1200 – 1400 cm^{-1} e 1540 – 1700 cm^{-1}, com intensidades máximas próximas de 1260 cm^{-1} e 1640 cm^{-1}, podem estar associadas às bandas Amida III (1230 – 1280 cm^{-1}) e Amida I (1650 – 1680 cm^{-1}), respectivamente. Essas bandas são frequentemente relacionadas à presença de componentes da matriz extracelular, incluindo colágeno, elastina, fibronectina e ácido hialurônico.

Os valores observados para as bandas com máximos próximos de 1260 cm^{-1} e 1640 cm^{-1} também são compatíveis com aqueles descritos por Milyudin et al. [14] para componentes da MAH, os quais foram atribuídos às bandas Amida III e Amida I, respectivamente. Essas bandas estão associadas a vibrações de ligações amida presentes em proteínas da matriz extracelular. Adicionalmente, bandas próximas de 879 cm^{-1}, 1247 – 1275 cm^{-1} e 1672 – 1678 cm^{-1} foram descritas em espectros Raman de colágeno tipos I e III, sendo atribuídas à glicina, Amida III e Amida I, respectivamente, compatíveis com os achados do presente estudo [9].

As interpretações realizadas neste estudo foram baseadas na comparação qualitativa e visual dos espectros Raman médios obtidos para cada condição experimental. Embora diferenças espectrais tenham sido observadas entre as condições avaliadas, análises quantitativas e multivariadas, como Análise de Componentes Principais (PCA), estão atualmente em fase de desenvolvimento. Dessa forma, os resultados devem ser interpretados como preliminares para as alterações espectrais associadas ao armazenamento da MAH em glicerol após diferentes períodos. Estudos futuros envolvendo maior número de doadoras, bem como análises quantitativas e multivariadas, serão necessários para confirmar os achados apresentados.

IV. CONCLUSÃO

A espectroscopia Raman permitiu detectar alterações no perfil espectral da membrana amniótica humana após armazenamento em glicerol. Observou-se manutenção do perfil espectral nos primeiros dois dias, alterações detectáveis após 14 dias e características espectrais distintas das anteriores após 35 e 68 dias. As principais regiões espectrais observadas foram compatíveis com os principais componentes da matriz extracelular e com bandas Amida I e Amida III. As alterações espectrais detectadas podem estar relacionadas a modificações em componentes proteicos específicos da matriz extracelular, entretanto, essa hipótese deverá ser confirmada por análises bioquímicas, estruturais e morfológicas complementares.

AGRADECIMENTOS

M. S. Melo agradece à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Financing Code 001; Santa Casa de Misericórdia de Pindamonhangaba, SP, Brasil e Centro de Inovação, Tecnologia e Educação (CITÉ) / Universidade Anhembi Morumbi (UAM), São José dos Campos, SP, Brasil.

H. C. Carvalho agradece à Fundação Araucária (FA) e Superintendência de Ciência, Tecnologia e Ensino Superior (SETI), PR, Brasil, pelo auxílio à participação no congresso.

A. B. Fernandes, L. Silveira Júnior, R. A. Zângaro agradecem ao Instituto Ânima, Universidade Anhembi Morumbi, São Paulo, SP, Brasil.

A. B. Fernandes agradece ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela bolsa de produtividade (Processo No. 310708/2021-4).

REFERÊNCIAS

[1] K. Jirsova, G. L. A. Jones, “Amniotic membrane in ophthalmology: properties, preparation, storage and indications for grafting – a review,” Cell Tissue Bank, vol 18, n. 2, pp. 193 – 204, 2017.

[2] S. Tighe, O. G. Mead, A. Lee, S. C. G. Tseng, “Basic science review of birth tissue uses in ophthalmology,” Taiwan Journal of Ophthalmology, vol. 10, n. 3, pp. 3 – 12, 2020.

[3] S. Cavalu, G. Roiu, O. Pop, D. A. P. Heredea, T. O. Costea, C. F. Costea, “Nano-scale modifications of amniotic membrane induced by UV and antibiotic treatment: histological, AFM and FTIR spectroscopy evidence,” Materials, vol. 14, n. 4, pp. 863, 2021.

[4] P. Heydari, M. Mojahedi, P. Javaherchi, M. Sharifi, A. Z. Kharazi, “Advances and impact of human amniotic membrane and human amniotic-based materials in wound healing application,” Int J Biol Macromol, vol. 281 (Pt. 4), 2024.

[5] M. Fénelon, S. Catros, C. Meyer, J-C. Fricain, L. Obert, F. Auber, A. Louvrier, F. Gindraux, “Applications of Human Amniotic Membrane for Tissue Engineering,” Membranes, vol. 11, 387, 2021.

[6] U. P. S. Parmar, P. L. Surico, A. Scarabosio, V. Barone, R. B. Singh, F. D’Ancona, M. Zeppieri, P. C. Parodi, T. Mori, F. Cutrupi, D. H. Ma, A. Di Zazzo, M. Coassin, “Amniotic membrane transplantation for wound healing, tissue regeneration and immune modulation,” Stem Cell Rev and Rep, vol. 21, 1428–1448, 2025.

[7] S. M. Zidan, S. A. Eleowa, M. A. Nasef, M. A. Abd-Almoktader, A. M. Elbatawy, A. G. Borhamy, M. A. Aboliela, A. M. Ali, M. R. Algamal, “Maximizing the safety of glycerol preserved human amniotic membrane as a biological dressing,” Burns, vol. 41, n. 7, pp. 1498 – 1503, 2015.

[8] N. Hofmann, H. Rennekampff, A. K. Salz, M. Börgel, “Preparation of human amniotic membrane for transplantation in different application areas,” Front. Transplant, vol. 2, 2023.

[9] J. J. Cárcamo, A. E. Aliaga, R. E. Clavijo, M. R. Brañes, M. M. Campos-Vallette, “Raman study of the shockwave effect on collagens,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 86, pp. 360 – 365, 2012.

[10] M. S. Bergholt, A. Serio, M. B. Albro, “Raman Spectroscopy: Guiding Light for the Extracellular Matrix,” Front. Bioeng. Biotechnol., Sec. Biomaterials, vol. 7, 303, 2019.

[11] X. Wang, F. Hua, X. Xie, Y. Wu, W. Sun, Z. Jiang, J. Zhang, X. Luo, “Infrared Raman spectroscopy enables noninvasive biochemical assessment of skin tissue and the thermal stability,” Heliyon, vol. 9, 2023.

[12] BRASIL. Agência Nacional de Vigilância Sanitária, “Resolução RDC nº 190, de 18 de julho de 2003. Determina normas técnicas para o funcionamento dos bancos de sangue de cordão umbilical e placentário,” Diário Oficial da União, Brasília, DF, 21 jul. 2003. [Online]. Available: https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2003/rdc0190_18_07_2003.html. Accessed: May 26, 2026.

[13] B. A. Kawata, V. A. H. Sato, P. A. L. I. Marrafa, S. M. Awoyama, M. S. Melo, L. F. F. Oliveira, A. C. Santos, J. B. Rodrigues, R. A. Zângaro, C. J. Lima, A. B. Fernandes, “Effects of ozone sterilization in human amniotic membrane morphology: a promising de-epithelialization process,” Ozone: Science & Engineering, vol. 48, n. 3, 2026.

[14] E. Milyudin, L. T. Volova, K. E. Kuchuk, E. V. Timchenko, P. E. Timchenko, “Amniotic Membrane Biopolymer for Regenerative Medicine,” Polymers, vol. 15, 1213, 2023.

Rheological Characterization of a Bioink for Bone Tissue Engineering Applications

Thaís Amaral França
Postgraduate Program on Biomedical Engineering
Federal University of ABC
São Bernardo do Campo - Brazil
thaisamaralfranca15@gmail.com

Vitor de Mattos
Research and Development Department
Quantis Biotechnology
São Paulo - Brazil
vitor.mattos@quantis.bio

Lucas Guarnier
Research and Development Department
Quantis Biotechnology
São Paulo - Brazil
lucas.guarnier@quantis.bio

Kamila Leichtweiss
Research and Development Department
Quantis Biotechnology
São Paulo - Brazil
kamila.leichtweis@quantis.bio

Sônia Maria Malmonge
Postgraduate Program on Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, SP - Brazil
sonia.malmonge@ufabc.edu.br

Janaina de Andrea Dernowsek
Research and Development Department
Quantis Biotechnology
São Paulo - Brazil
janaina.dernowsek@quantis.bio

***Abstract*— The development of printable biomaterials for bone tissue engineering requires formulations that combine easy extrusion with structural stability and rapid recovery after deposition. This study investigated the rheological behavior of a composite hydrogel of agarose, sodium alginate, gelatin, and β-tricalcium phosphate (β-TCP) as a candidate bioink for extrusion-based three-dimensional bioprinting. Rheological characterization included oscillatory strain and frequency sweeps, steady-shear viscosity measurements, oscillatory yield-stress determination, and a three-interval thixotropy test. The formulation exhibited a broad linear viscoelastic region, extending from approximately 0.01% to 13.3% strain, and predominantly elastic behavior, with the storage modulus remaining higher than the loss modulus throughout the tested frequency range. The oscillatory yield stress was approximately 1.02 kPa. Apparent viscosity decreased from 21,883 Pa·s at a shear rate of 0.01 s^{-1} to 0.203 Pa·s at 1,310 s^{-1}. Power-law fitting yielded a flow behavior index of 0.0098 and an R^2 of 0.991, confirming pronounced shear-thinning behavior. Following high-strain disruption, the hydrogel rapidly recovered its solid-like character, reaching 64.74% recovery of the storage modulus and 78.36% recovery of complex viscosity. Overall, the formulation combined resistance to deformation at rest, shear-dependent flow under extrusion-relevant conditions, and functionally relevant post-shear recovery. Direct printing further confirmed the rheological predictions, with printed grids maintaining their pore architecture and filament deflection remaining below 3.5% across unsupported spans of up to 5.63 mm. Overall, the formulation combined shear-dependent flow, structural recovery, and effective post-deposition shape retention, supporting its potential for extrusion-based 3D printing. Mechanical, degradation, and biological evaluations are still required to establish its suitability for bone tissue engineering applications.*

Keywords*— *Bone repair, Bioink, Rheology, 3D Bioprinting.

I. Introduction

Bone is among the most frequently transplanted tissues in clinical practice, and the repair of large defects arising from trauma, tumor resection, infection, or congenital malformation remains a persistent challenge in regenerative medicine [1,2]. Although bone possesses a notable intrinsic capacity for self-repair, defects exceeding a critical size do not heal spontaneously and require external intervention [2,3]. Autologous grafts are still regarded as the clinical gold standard, yet their use is constrained by donor-site morbidity, restricted donor availability, and the need for additional surgical procedures, whereas allografts and synthetic substitutes carry risks of immune rejection, disease transmission, and limited osteointegration [1,2]. In this context, bone tissue engineering has emerged as a promising strategy, combining three-dimensional scaffolds, cells, and bioactive signals to restore the structure and function of the native tissue [3,4].

Among the available scaffold fabrication technologies, three-dimensional (3D) bioprinting has attracted increasing attention because it enables the production of patient-specific constructs through the layer-by-layer placement of cells and biomaterials, with precise control over architecture and porosity [5,6]. Extrusion-based bioprinting, in particular, is the most widely adopted approach for bone applications, owing to its versatility, relatively low cost, and ability to process highly viscous formulations and elevated cell densities [6–8]. The success of this technique, however, depends critically on the bioink, which must simultaneously ensure adequate flow during extrusion, structural fidelity after deposition, and a biologically permissive environment, requirements that are often in conflict [5,9]. For bone-oriented applications, composite formulations that incorporate bioceramics such as β-tricalcium phosphate (β-TCP) into polymeric matrices are especially attractive, since the inorganic phase promotes new bone formation, increases the viscosity and mechanical strength of the construct, and contributes to its bioactivity and osteogenic potential [10–12].

The printability and post-printing stability of a bioink are largely governed by its rheological behavior, which makes rheological characterization an essential step prior to printing. Properties such as shear-thinning (pseudoplastic) flow, viscoelasticity expressed through the storage (G′) and loss (G″) moduli, yield stress, and the capacity for structural recovery after deformation (thixotropy) directly govern bioprintability [13]. An ideal bioink should flow readily under the shear imposed within the extrusion nozzle and rapidly recover its solid-like character once deposited, thereby preserving filament geometry and supporting the dimensional stability of multilayered constructs [7,14]. Assessing these parameters therefore provides a predictive basis for evaluating extrudability and shape fidelity before committing to printing trials [9,13].

In this work, the rheological behavior of a composite hydrogel based on agarose, sodium alginate, gelatin, and β-TCP was investigated as a candidate bioink for extrusion-based bone tissue engineering. The formulation was characterized through oscillatory strain and frequency sweeps, steady-state viscosity measurements, yield-stress determination, and three-interval thixotropy tests. Direct printing, grid shape-fidelity, and filament collapse tests were additionally performed to determine whether the rheological behavior translated into effective printability and post-deposition stability.

II. Materials and methods

A. Materials

Gelatin from porcine skin (Sigma-Aldrich, G1890); sodium alginate (Sigma-Aldrich, W201502); agarose (Kasvi, K9-9500); tribasic calcium phosphate, also referred to as β-tricalcium phosphate, with purity ≥96.0% (Sigma-Aldrich, 21218); and 10X phosphate-buffered saline (PBS 10X, Nova Biotecnologia, code 13-30262-05).

B. Preparation of the bioink

The hydrogel was prepared by adding 10 mL of PBS 1X to a 50 mL beaker under magnetic stirring at approximately 300 rpm and 80 °C. Agarose (0.2 g) was first added and dissolved until a translucent and homogeneous solution was obtained. Sodium alginate (0.4 g) was then gradually incorporated to avoid clump formation, followed by the addition of gelatin from porcine skin (0.5 g). At this step, the temperature was reduced to 50 °C, and the mixture was maintained under stirring until a uniform and slightly viscous solution was formed. Finally, β-tricalcium phosphate (β-TCP; 0.1 g) was gradually added and dispersed into the polymeric matrix. When magnetic stirring became inefficient due to increased viscosity, manual mixing with a spatula was performed to ensure homogeneity. The final formulation was visually assessed for homogeneity and absence of clumps.

C. Test Equipment and Conditions

The rheological assays were performed using an MCR 92 rotational rheometer (Anton Paar GmbH, Graz, Austria), equipped with a synchronous EC motor and air-bearing system, which enables torque measurements from 0.4 μNm with high sensitivity and no mechanical friction, making it particularly suitable for the structural characterization of low- to medium-consistency viscoelastic materials. A 25 mm parallel-plate geometry was used (PP25, Anton Paar), with a fixed gap of 1.0 mm. Temperature control was performed using a Peltier plate system, model P-PTD 200/AIR (Anton Paar), with an accuracy of ±0.1 °C, and all assays were conducted at 25 °C. Data acquisition and processing were carried out using RheoCompass™ software (Anton Paar).

D. Strain Sweep

The linear viscoelastic region (LVER) of the hydrogel was determined by an oscillatory strain sweep test. The samples were subjected to increasing strain values from 0.01% to 1000% on a logarithmic scale, while maintaining a fixed frequency of 1 Hz. During the assay, the storage modulus (G′), loss modulus (G″), and loss factor (tan δ = G″/G′) were recorded. The LVER was defined as the strain range in which G′ and G″ remained approximately constant, allowing a maximum variation of 10% relative to the initial plateau. The critical strain (γc) was defined as the strain corresponding to the onset of G′ decrease or to a measurable deviation from linear behavior. The strain used in subsequent oscillatory tests was selected within the LVER, preferably one order of magnitude below γc.

E. Frequency Sweep

The time scale-dependent viscoelastic behavior was evaluated by a frequency sweep test. The assay was performed at a fixed strain within the LVER, previously determined by the strain sweep test. The angular frequency (ω) was varied from 0.1 to 100 rad·s^{-1} on a logarithmic scale. The storage modulus (G′), loss modulus (G″), loss factor (tan δ), and complex viscosity were determined. The predominantly solid-like or liquid-like behavior was assessed based on the relationship between G′ and G″ over the frequency range. G′ values higher than G″ were interpreted as a predominance of elastic behavior, associated with shape retention after deposition. In contrast, G″ values higher than G′ would indicate predominantly viscous behavior. The frequency dependence of the moduli was used to infer the structural stability of the material at different time scales relevant to the extrusion process.

F. Viscosity Curve

The viscosity curve of hidrogel was evaluated using a steady-state rotational rheological test. The shear rate was varied from 0.01 to 1300 s^{-1} on a logarithmic scale, while the apparent viscosity and shear stress were recorded. The experimental data were fitted to the Ostwald–de Waele model, described by the following equation:

$$\eta = K\dot{\gamma}^{(n-1)} \quad (1)$$

where η is the apparent viscosity, $\dot{\gamma}$ is the shear rate, K is the consistency index, and n is the flow behavior index. For linearization, the model was expressed as:

$$\ln(\eta) = \ln(K) + (n-1)\ln(\dot{\gamma}) \quad (2)$$

Thus, the flow behavior index was calculated from the slope of the linear regression as n = slope + 1, while K was obtained from the exponential of the intercept. Values of n < 1 were considered indicative of shear-thinning behavior. The fitting quality was evaluated based on the coefficient of determination, residual analysis, and mean fitting error.

G. Yield Stress

The yield stress was determined using complementary methods, considering that structured semi-solid materials may exhibit protocol-dependent values. First, an oscillatory stress sweep was performed at a fixed frequency of 1 Hz. The stress was progressively increased on a logarithmic scale until disruption of the viscoelastic structure. The oscillatory yield stress was defined as the stress corresponding to the crossover point between G′ and G″, i.e., the point at which the viscous behavior begins to predominate over the elastic behavior. Additionally, the stress associated with the onset of G′ decrease was recorded and considered as the structural yield limit.

H. Thixotropy

The structural recovery capacity after intense deformation was evaluated using a three-interval oscillatory step-strain test at a fixed frequency of 1 Hz. In the first interval, the sample was subjected to 1% strain, within the LVER, for 120 s to record the initial structural state. In the second interval, the strain was increased to 200% for 120 s to induce structural disruption. In the third interval, the strain was returned to 1% for 300 s, allowing the recovery of G′ and G″ to be monitored over time.

Structural recovery was quantified as the ratio between the storage modulus after returning to low strain and the initial storage modulus, according to the following equation:

$$G' \text{ recovery } (\%) = (G't / G'0) \times 100 \quad (3)$$

where G′0 corresponds to the mean G′ value during the initial step, and G′t corresponds to the G′ value at defined time points after the high-strain step. Immediate recovery, in 10 s, final recovery, in 300 s, and the time required to recover 50% and 90% of the initial modulus were reported when applicable.

I. Filament Collapse Test

Post-deposition stability was evaluated using a filament collapse test with unsupported spans of 0.80, 1.60, 2.65, 3.67, 4.64, and 5.63 mm. A single filament was printed across the supports using the parameters shown in Table I. Images were acquired immediately after printing and analyzed using ImageJ/Fiji. Relative deflection was calculated as:

$$Relative\ deflection(\%) = \frac{\delta}{d} * 100 \quad (4)$$

where δ is the maximum perpendicular displacement of the filament relative to the straight line connecting the two supports, and d is the unsupported span length.

J. Direct Printing and Shape-Fidelity Assessment

The printability and shape fidelity of the bioink were evaluated by printing 30 × 30 mm square grids using a 0°/90° deposition pattern with three layers. The grid geometry was prepared using PrusaSlicer software, which was also used to define the printing parameters and generate the G-code. The bioprinter used was the BioEnder (BioEdTech, São Paulo, Brazil).

TABLE I. PRINTING PARAMETERS USED FOR SHAPE-FIDELITY ASSESSMENT.

Parameter	***Values***
Nozzle	0.6 mm
Speed	5 mm/s
Extrusion multiplier	0.5
Infill	20%
Layer height	0.5 mm
Printing temperature	25 ºC

Printing parameters (shown on Table I) were maintained constant for all samples. Immediately after printing, the constructs were photographed from the top with a millimeter scale and analyzed using ImageJ/Fiji. At least five internal pores per grid were evaluated by measuring their area (A) and perimeter (L). The printability parameter was calculated according to Ribeiro et al. [15] as:

$$Pr = \frac{L^2}{16A} \quad (5)$$

where values close to 1 indicate greater preservation of the designed square pore geometry.

K. Statistical Analysis

Rheological and printability data were expressed as mean ± standard deviation from at least three independent bioink preparations.

III. RESULTS AND DISCUSSION

The amplitude sweep analysis showed that the bioink exhibited a broad Linear Viscoelastic Region (LVER), extending from approximately 0.01% to 13.3% shear strain, as shown in Figure 1. Within this range, the storage modulus (G′) remained predominantly stable, with values close to 1,000 Pa at lower strain amplitudes, ranging from 1,002.59 ± 382.50 Pa to 1,008.37 ± 347.81 Pa. At the upper limit of the LVER, at γ = 13.3%, G′ was 927.34 ± 287.01 Pa, remaining within the stability criterion adopted for the linear region.

Throughout the LVER, the loss modulus (G″) remained lower than G′, with values ranging from 208.04 ± 55.63 Pa to 219.39 ± 43.19 Pa. This predominance of G′ confirms the mainly elastic behavior of the bioink under low-strain conditions, indicating that the three-dimensional network formed by interactions among agarose, alginate, gelatin, and β-TCP preserves its structural integrity within this range and behaves as a solid-like gel. The G′ values obtained are consistent with those reported for hydrogels applied to bone tissue engineering, which are generally within the range of hundreds to a few thousand Pa depending on polymer

composition, concentration, and the presence of inorganic phases [8,11].

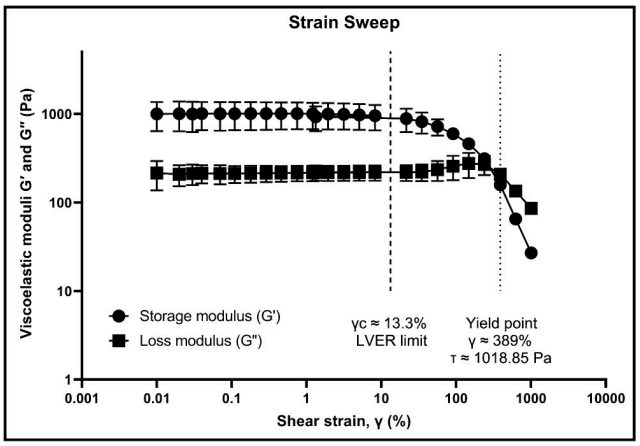


Fig. 1. Amplitude sweep analysis of the bioink showing the storage modulus (G′) and loss modulus (G″) as a function of shear strain. The dashed vertical line indicates the LVER limit/critical strain, while the dotted vertical line indicates the yield point/crossover between G′ and G″. Data are presented as mean ± standard deviation.

The amplitude sweep also allowed identification of two complementary structural thresholds. The first was the critical strain, observed at approximately 13.3%, corresponding to a shear stress of 127.10 ± 47.30 Pa. This point marks the onset of measurable G′ decrease and therefore indicates the beginning of nonlinear viscoelastic behavior. However, this should be interpreted as the initial structural yield limit rather than complete flow. The second threshold was the oscillatory yield stress, determined from the crossover between G′ and G″, as shown in Figure 2. This crossover occurred at approximately τy = 1,018.85 Pa, with a corresponding strain of about 389%. At the equivalent experimental point, the shear stress was 1,018.95 ± 49.96 Pa. The large separation between the onset of nonlinearity and the G′/G″ crossover indicates that the bioink begins to undergo structural rearrangement at moderate stress but only transitions to predominantly viscous behavior under substantially higher deformation.

This high oscillatory yield stress is a desirable characteristic for extrusion-based 3D bioprinting, since it indicates that the material has sufficient resistance to deformation before flowing. Such resistance contributes to shape retention after deposition, reducing the risk of immediate filament spreading or collapse of printed layers [14,15]. At the same time, the fact that the crossover occurs only at high strain supports the ability of the formulation to behave as a solid-like gel at rest while still becoming flowable under the stronger mechanical forces imposed during extrusion. Based on the identified LVER, a strain of 1% was selected for subsequent oscillatory tests, ensuring that the intrinsic viscoelastic properties of the bioink were evaluated without disrupting the network structure.

The frequency sweep further confirmed the stability of the bioink under oscillatory conditions within the LVER (Figure 3). Across the tested angular frequency range, G′ remained consistently higher than G″, with no crossover between the moduli. This indicates that the formulation maintained a predominantly elastic response over different time scales, reinforcing its solid-like gel behavior. G′ increased moderately with angular frequency, while G″ also increased at higher frequencies but remained below G′ throughout the test. This frequency-dependent increase in the moduli suggests that the polymeric-mineral network becomes more resistant to deformation at shorter time scales, while still preserving the dominance of the elastic component.

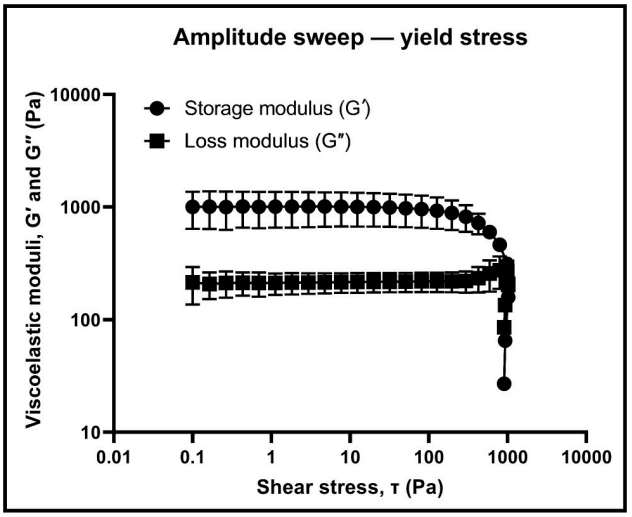


Fig. 2. Amplitude sweep for oscillatory yield stress determination. Storage modulus (G′) and loss modulus (G″) are shown as a function of shear stress. The yield stress was determined from the crossover between G′ and G″, corresponding to τy ≈ 1,018.85 Pa. Data are presented as mean ± standard deviation.

The absence of a G′/G″ crossover in the frequency sweep is particularly relevant for bioink performance, since it indicates that the material does not transition to a liquid-like state under the small-amplitude oscillatory conditions used to probe its structure. Similar behavior has been described for physically crosslinked and composite hydrogel bioinks, in which G′ > G″ over the frequency range is associated with a stable network, improved shape fidelity, and enhanced capacity to maintain filament geometry after deposition [9,11,13]. Therefore, the frequency sweep complements the amplitude sweep by showing that the bioink is not only stable at low strain, but also maintains its elastic character across the range of deformation time scales evaluated.

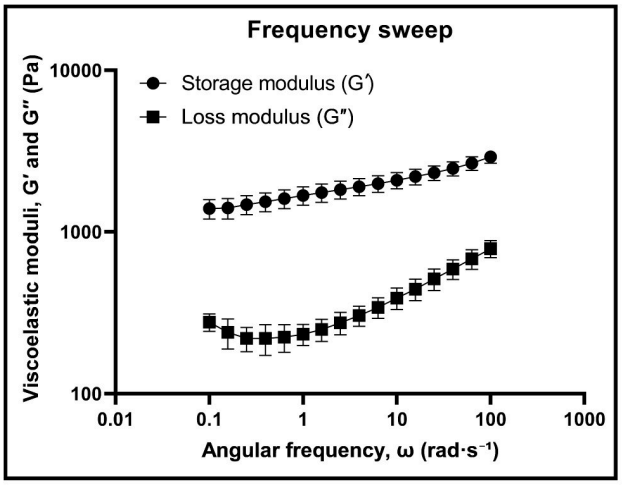


Fig. 3. Frequency sweep of the bioink. Storage modulus (G′) and loss modulus (G″) were evaluated as a function of angular frequency. G′ remained higher than G″ throughout the tested frequency range, indicating predominant elastic behavior and stability of the hydrogel network under oscillatory conditions. Data are presented as mean ± standard deviation.

The apparent viscosity of the bioink was evaluated as a function of shear rate and showed a marked reduction throughout the analyzed range, evidencing the pseudoplastic behavior characteristic of shear-thinning systems (Figure 4). The obtained slope was −0.9902, resulting in a flow behavior index of n = 0.0098, while the consistency index derived from the intercept corresponded to K = 474,473.24 mPa·s^n. The coefficient of determination was R^2 = 0.991, indicating a high-quality fit within the evaluated shear-rate range.

The apparent viscosity curve showed that, at a low shear rate of 0.010 s^{-1}, the bioink exhibited a viscosity of 21,883.00 ± 1,632.32 Pa·s. In contrast, at the highest evaluated shear rate, 1,310 s^{-1}, the viscosity decreased to 0.203 ± 0.115 Pa·s. This reduction of approximately five orders of magnitude confirms the strong dependence of viscosity on shear rate and demonstrates that the bioink becomes significantly less viscous under conditions of higher deformation. The flow index n ≈ 0.01, much lower than 1, confirms the strongly shear-thinning nature of the formulation. For comparison, Newtonian fluids present n = 1, whereas many bioinks described for soft tissue bioprinting exhibit n values between 0.2 and 0.6 (14,15). The exceptionally low n value obtained in this study indicates a highly pronounced pseudoplastic response, probably associated with the high density of entanglements and physical interactions among agarose, alginate, and gelatin, in addition to the contribution of β-TCP as a dispersed solid phase.

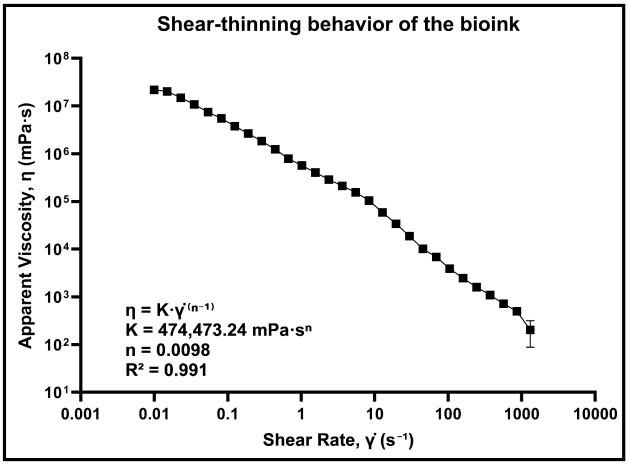


Fig 4. Apparent viscosity curve of the bioink as a function of shear rate. The data are presented as mean ± standard deviation in log-log scale. The marked decrease in viscosity with increasing shear rate confirms the shear-thinning behavior of the formulation. The apparent viscosity data were fitted using the power-law model, η = K·γ̇^(n−1), yielding K = 474,473.24 mPa·s^n, n = 0.0098, and R^2 = 0.991.

From the perspective of extrusion-based 3D bioprinting, this behavior is highly favorable. During passage through the extrusion nozzle, the increase in shear rate markedly reduces the viscosity of the bioink, facilitating flow and decreasing the need for excessive pressures, which could compromise cell viability. After deposition, when shear is removed, viscosity recovery contributes to filament shape retention and to the structural stability of the printed construct [7,8]. Together, the high consistency index, the strong shear-thinning behavior, and the elevated oscillatory yield stress indicate that the bioink can combine extrudability during printing with sufficient resistance to deformation after deposition. This balance is especially important for bone scaffold fabrication, in which geometric stability and controlled porosity are essential for the functional performance of the printed structure [4].

The thixotropy analysis was then performed to evaluate whether the bioink could recover its structure after a deformation event simulating extrusion. The three-interval oscillatory test consisted of an initial low-strain interval within the LVER, a high-strain interval to induce network disruption, and a final low-strain interval to monitor recovery. The results are summarized in Table II and illustrated in Figures 5 and 6.

TABLE II. SUMMARY OF RHEOLOGICAL PARAMETERS OBTAINED IN THE THREE INTERVALS OF THE THIXOTROPY TEST.

Parameter	***Interval 1 (Initial structure)***	***Interval 2 (Disruption)***	***Interval 3 (Recovery)***
G′ (Pa)	2,160 - 2,050	165 - 251	1,431 - 1,627
G″ (Pa)	373 - 348	358 - 412	455 - 304
η* (mPa·s)	348,91 - 330,4	62,769 - 76,852	239,01 - 263,3
G′ recovery (%)	—	—	64,74%
η* recovery (%)	—	—	78,36%

In the initial interval, the bioink exhibited predominantly elastic behavior, with G′ remaining higher than G″ throughout the entire period. The storage modulus remained close to 2.1 kPa, varying from approximately 2,160 Pa at the beginning to 2,050 Pa at the end of 120 s, indicating good network stability at rest. During the same interval, G″ remained around 350–370 Pa, confirming that the viscoelastic response of the formulation was dominated by the elastic component. The complex viscosity also showed only a slight decrease, from approximately 348,917 to 330,403 mPa·s, consistent with the maintenance of the internal structure under low deformation.

During the second interval, the application of high deformation promoted clear disruption of the viscoelastic network. G′ dropped abruptly to values close to 165 Pa at the beginning of the perturbation stage, while G″ remained within the range of approximately 358–412 Pa. Under this condition, G″ temporarily became higher than G′, indicating a transition of the material toward a more viscous behavior and confirming the breakdown of the elastic structure under intense deformation. The complex viscosity also decreased markedly, reaching values close to 62,769 mPa·s, further supporting the temporary loss of structural resistance. This behavior is consistent with the shear-thinning profile observed in the viscosity curve and confirms that the bioink

network can be disrupted under high deformation, allowing flow during extrusion.

After returning to low deformation in the third interval, the bioink showed rapid recovery of the viscoelastic moduli. G′ immediately increased to approximately 1,431 Pa within the first seconds and reached around 1,600 Pa in less than 15 s, stabilizing at a plateau close to 1,625 Pa until the end of the test. Although G′ did not fully return to its initial value, it again remained substantially higher than G″, which was maintained around 300 Pa at the end of the recovery stage. This result indicates that the bioink recovers its solid-like gel behavior after perturbation, an essential feature for maintaining filament shape after extrusion.

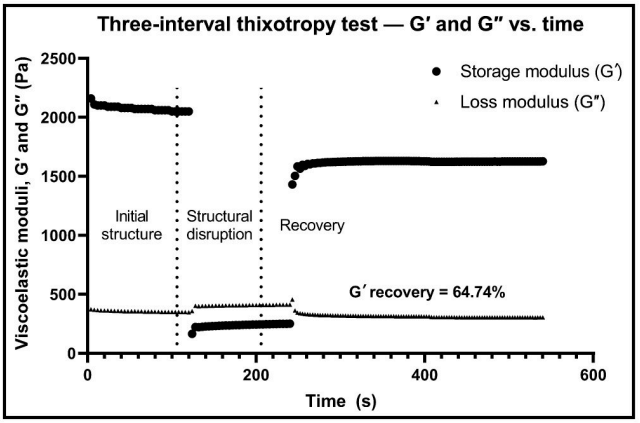


Fig. 5. Three-interval thixotropy test showing G′ and G″ over time. G′ decreased sharply during high strain and partially recovered after returning to low strain, reaching 64.74% recovery.

The structural recovery of G′, calculated from representative mean values of the initial and final intervals, was 64.74%. This value indicates partial but functionally relevant recovery of the viscoelastic network. The complex viscosity showed a similar response, increasing rapidly after removal of the high deformation and stabilizing around 263,000–264,000 mPa·s, with a recovery of 78.36%. The difference between G′ recovery and η* recovery suggests that different levels of structural organization within the bioink are restored at distinct time scales: interactions associated with resistance to flow may reorganize more rapidly, whereas full recovery of the elastic network depends on slower supramolecular rearrangements.

This profile of rapid recovery followed by stabilization is characteristic of hydrogels based on partially reversible physical networks. In the studied formulation, this behavior may be related to interactions among agarose, alginate, gelatin, and β-TCP. Agarose contributes to the formation of physical networks through hydrogen-bonding interactions, gelatin favors thermoreversible interactions and matrix cohesion, while alginate contributes to the viscosity and organization of the polymeric phase. In addition, β-TCP acts as a dispersed mineral phase that can increase the mechanical strength of the network and contribute to stability. The incomplete recovery of G′ within the 300 s window suggests that part of these interactions is not fully restored in the short term, a behavior also described for physical hydrogels and bioinks based on polysaccharides and proteins [11,16,17].

The filament collapse test demonstrated high post-deposition stability across all unsupported spans tested (0.80–5.63 mm). Relative deflection remained below 3.5%, ranging from approximately 0.5% to 3.2%. Even at the largest gap of 5.63 mm, the filament exhibited only ~1.9% deflection, indicating effective shape retention after extrusion.

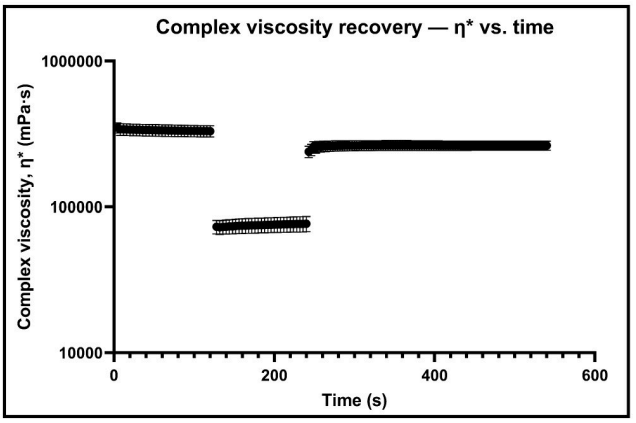


Fig. 6. Complex viscosity recovery during the thixotropy test. Complex viscosity (η*) decreased during the high-strain disruption interval and partially recovered after returning to low strain, indicating reorganization of the bioink network after shear-induced breakdown. Data are presented as mean ± standard deviation.

Overall, the low filament deflection demonstrates that the bioink can maintain its geometry after deposition, even across relatively large unsupported distances. These results confirm good post-printing stability and support the printability predicted by the rheological characterization.

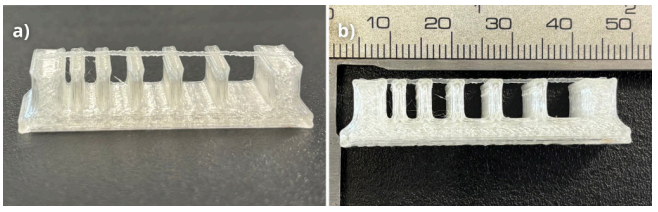


Fig. 7. Filament collapse test of the bioink across progressively increasing unsupported spans. (a) Oblique view and (b) perpendicular view showing preservation of filament continuity and minimal deformation after deposition.

Direct printing confirmed the ability of the bioink to form continuous filaments and preserve the designed grid architecture. The printed pores remained well defined, with a printability parameter of Pr = 1.15 ± 0.08, indicating good preservation of the square pore geometry.

Overall, the rheological results indicate that the agarose/alginate/gelatin/β-TCP bioink presents a favorable balance between flowability and structural stability. The amplitude sweep showed a broad LVER and high oscillatory yield stress, indicating resistance to deformation at rest. The frequency sweep confirmed that G′ remained higher than G″ across the tested frequency range, demonstrating the persistence of solid-like behavior under oscillatory conditions.

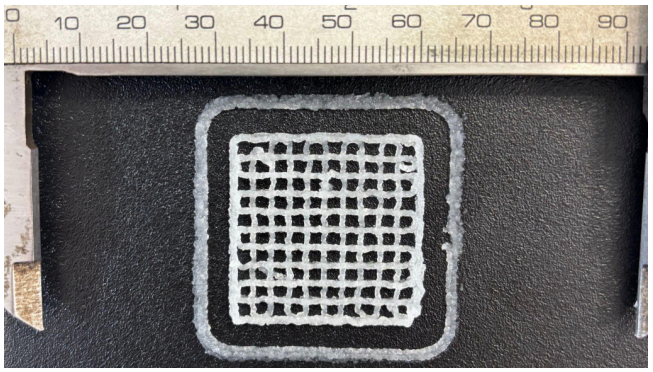


Fig. 8. Representative 30 × 30 mm grid printed with a 0°/90° deposition pattern, showing preservation of filament continuity and square pore geometry after deposition.

The viscosity curve showed pronounced shear-thinning behavior, supporting extrudability under high shear. Importantly, although G′ recovery was incomplete (64.74%), the direct printing results demonstrated that the recovered structure was sufficient to maintain filament geometry at the macroscopic scale after deposition. Together, these findings demonstrate a favorable balance between extrusion flowability and post-deposition stability, which are fundamental requirements for the fabrication of bone tissue engineering scaffolds with defined geometry and controlled porosity.

IV. Conclusion

The agarose/alginate/gelatin/β-TCP formulation exhibited rheological properties compatible with extrusion-based 3D bioprinting, including predominantly elastic behavior, pronounced shear-thinning, high oscillatory yield stress, and rapid partial structural recovery after deformation. Direct printing experiments further confirmed its printability, demonstrating preservation of grid geometry and low filament deflection across unsupported spans up to 5.63 mm. These results indicate that, despite the incomplete recovery of G′, the bioink provides sufficient post-deposition stability to maintain printed structures. Therefore, the formulation represents a promising candidate for the extrusion-based fabrication of scaffolds for bone tissue engineering. Further mechanical, degradation, and biological studies are still required to confirm its performance for bone regeneration applications.

Acknowledgment

The authors gratefully acknowledge Quantis Biotechnology for providing materials and laboratory infrastructure for the development of this study. The authors also thank Douglas Santos and Anton Paar for granting access to the rheometer and for their valuable technical assistance during the rheological characterization.

References

1. Zimmerer R, Jehn P, Spalthoff S, Kokemüller H, Gellrich NC. Präfabrikation vaskularisierter Gesichtsknochen. Chir. 1º de março de 2015;86(3):254–8. doi:10.1007/s00104-014-2884-x
2. Łuczak JW, Palusińska M, Matak D, Pietrzak D, Nakielski P, Lewicki S, et al. The Future of Bone Repair: Emerging Technologies and Biomaterials in Bone Regeneration. Int J Mol Sci. 27 de novembro de 2024;25(23):12766. doi:10.3390/ijms252312766
3. Lopez de Armentia S, del Real JC, Paz E, Dunne N. Advances in Biodegradable 3D Printed Scaffolds with Carbon-Based Nanomaterials for Bone Regeneration. Materials. janeiro de 2020;13(22):5083. doi:10.3390/ma13225083
4. Murphy SV, Atala A. 3D bioprinting of tissues and organs. Nat Biotechnol. agosto de 2014;32(8):773–85. doi:10.1038/nbt.2958
5. Zhang J, Wehrle E, Rubert M, Müller R. 3D Bioprinting of Human Tissues: Biofabrication, Bioinks, and Bioreactors. Int J Mol Sci. 12 de abril de 2021;22(8):3971. doi:10.3390/ijms22083971
6. Zhang S, Chen X, Shan M, Hao Z, Zhang X, Meng L, et al. Convergence of 3D Bioprinting and Nanotechnology in Tissue Engineering Scaffolds. Biomimetics. 26 de fevereiro de 2023;8(1):94. doi:10.3390/biomimetics8010094
7. Gao G, Kim H, Kim BS, Kong JS, Lee JY, Park BW, et al. Tissue-engineering of vascular grafts containing endothelium and smooth-muscle using triple-coaxial cell printing. Appl Phys Rev. 22 de outubro de 2019;6(4):041402. doi:10.1063/1.5099306
8. Chimene D, Peak CW, Gentry JL, Carrow JK, Cross LM, Mondragon E, et al. Nanoengineered Ionic–Covalent Entanglement (NICE) Bioinks for 3D Bioprinting. ACS Appl Mater Interfaces. 28 de março de 2018;10(12):9957–68. doi:10.1021/acsami.7b19808
9. Petta D, Armiento AR, Grijpma D, Alini M, Eglin D, D'Este M. 3D bioprinting of a hyaluronan bioink through enzymatic-and visible light-crosslinking. Biofabrication. 1º de outubro de 2018;10(4):044104. doi:10.1088/1758-5090/aadf58
10. Song JL, Fu XY, Raza A, Shen NA, Xue YQ, Wang HJ, et al. Enhancement of mechanical strength of TCP-alginate based bioprinted constructs. J Mech Behav Biomed Mater. 1º de março de 2020;103:103533. doi:10.1016/j.jmbbm.2019.103533
11. Heid S, Boccaccini AR. Advancing bioinks for 3D bioprinting using reactive fillers: A review. Acta Biomater. 1º de setembro de 2020;113:1–22. doi:10.1016/j.actbio.2020.06.040
12. Wu YF, Wen YT, Salamanca E, Moe Aung L, Chao YQ, Chen CY, et al. 3D-bioprinted alginate-based bioink scaffolds with β-tricalcium phosphate for bone regeneration applications. J Dent Sci. 1º de abril de 2024;19(2):1116–25. doi:10.1016/j.jds.2023.12.023
13. Bercea M. Rheology as a Tool for Fine-Tuning the Properties of Printable Bioinspired Gels. Molecules. 19 de março de 2023;28(6):2766. doi:10.3390/molecules28062766
14. Paxton N, Smolan W, Böck T, Melchels F, Groll J, Jungst T. Proposal to assess printability of bioinks for extrusion-based bioprinting and evaluation of rheological properties governing bioprintability. Biofabrication. 14 de novembro de 2017;9(4):044107. doi:10.1088/1758-5090/aa8dd8
15. Ribeiro A, Blokzijl MM, Levato R, Visser CW, Castilho M, Hennink WE, et al. Assessing bioink shape fidelity to aid material development in 3D bioprinting. Biofabrication. 1º de janeiro de 2018;10(1):014102. doi:10.1088/1758-5090/aa90e2
16. Dávila JL, d'Ávila MA. Laponite as a rheology modifier of alginate solutions: Physical gelation and aging evolution. Carbohydr Polym. 10 de fevereiro de 2017;157:1–8. doi:10.1016/j.carbpol.2016.09.057 PubMed PMID: 27987800.
17. Wüst S, Godla ME, Müller R, Hofmann S. Tunable hydrogel composite with two-step processing in combination with innovative hardware upgrade for cell-based three-dimensional bioprinting. Acta Biomater. fevereiro de 2014;10(2):630–40. doi:10.1016/j.actbio.2013.10.016

Characterization of a Gelatin/Whitlockite Ink for 3D Printing of Scaffolds for Osteochondral Regeneration

1st Gabriella Ribeiro Froehlich
School of Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
São Paulo, Brazil
gabriella.rfroehlich@einstein.edu.br

2nd Larissa R. Lourenço
Center for Natural Science and Humanities
Federal University of ABC
Santo Andre, Brazil
r.larissa@ufabc.edu.br

3rd Juliana Marchi
Center for Natural Science and Humanities
Federal University of ABC
Santo Andre, Brazil
Juliana.marchi@ufabc.edu.br

4th Ligia Passos Maia Obi
Center of Engineering and Applied Social Sciences
Federal University of ABC
Santo Andre, Brazil
ligia.maia@ufabc.edu.br

5th Roger Borges
School of Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
São Paulo, Brazil
roger.b@einstein.br

Abstract*—The development of bioinks with suitable rheological properties is essential for the successful application of extrusion-based 3D bioprinting in tissue engineering. In this context, whitlockite nanoparticles have attracted attention due to their bioactivity and piezoelectric behavior, making them promising candidates for developing piezoelectric bioinks for bone tissue engineering. This study aimed to investigate the influence of pure and zinc-doped whitlockite nanoparticles on the rheological properties of GelMA hydrogels intended for use as bioinks. GelMA was synthesized through glycidyl methacrylate functionalization and characterized by Nuclear Magnetic Resonance (NMR). Pure and zinc-doped whitlockite nanoparticles (0.5% and 1 mol% Zn) were synthesized using the sol–gel method and characterized by X-ray Diffraction (XRD). Hydrogels containing different nanoparticle concentrations (1 and 10 wt.%) were prepared and evaluated through rheological tests, including sol-gel transition temperature (Tsol−gel), viscosity, linear viscoelastic region (LVR), and structural recovery. The incorporation of whitlockite increased the sol-gel transition temperature and expanded the LVR, while all formulations maintained pseudoplastic behavior and structural recovery, properties that are desirable for extrusion-based 3D printing. The specific influence of zinc doping was less evident and depended on the nanoparticle concentration. Overall, the results indicate that incorporating zinc-doped whitlockite alters the rheological behavior of GelMA hydrogels, suggesting its potential for developing bioinks for tissue engineering applications.

***Index Terms*—Gelatin methacryloyl (GelMA); Whitlockite; Rheology; Zinc doping**

I. Introduction

Bone regeneration is a widely discussed topic in the scientific community, since approximately two million bone grafting procedures are performed annually to repair bone defects [1]. Bone tissue has a highly complex structure due to its hierarchical arrangement of collagen fibers organized into seven layers [2], which makes it the stiffest tissue in the human body and confers high resistance to compressive forces. Thus, bones provide mobility and protect vital organs against mechanical shocks [3].

Another remarkable characteristic of bone tissue is its capacity for self-regeneration, which allows it to continuously remodel and replace itself, maintaining the individual's bone structure. This property facilitates the treatment of fractures and mild bone injuries, since the tissue itself is capable of regenerating [3]. However, in cases of severe tissue loss, the regeneration process depends on human intervention. This is a highly complex process because bone tissue is difficult to synthesize, and its regeneration depends on three factors: osteogenic cells, osteoinductive growth factors, and osteoconduction [4].

Therefore, biomaterials have become widely used in treatment. These components are studied within the interdisciplinary field of materials science, which investigates the relationships between the structure, properties, and processing of different materials. Within this field, biomaterials research focuses on cellular and tissue interactions at the molecular level. This area plays a crucial role in advancing medicine [5].

In this context, the evolution of bone repair strategies can be approached according to the different generations of biomaterials. The first generation aims to mimic biological bone components in an inert manner, without generating interactions, forming a bionic structure. The second generation involves the use of bioactive materials that integrate with the biological environment, albeit in an artificial manner. Furthermore, the third generation comprises bioactive and resorbable materials that are already used in tissue-engineering-based regenerative medicine [5].

Currently, a fourth generation of biomaterials has been proposed, with a focus on electrophysiological behavior [5]. In these materials, upon application of an external mechanical stimulus, surface charges form, creating electric fields with negative potential in compression regions and positive potential in tension regions. This polarization alters the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

membrane potential of adjacent cells, promoting the opening of voltage-gated calcium channels and stimulating the migration and proliferation of bone-forming cells [6].

This behavior is explained by a chemical property shared by both structures: piezoelectricity, the ability to generate an electric field in response to mechanical stress. This occurs due to a transient deformation of the electric dipole in the crystalline material under mechanical loading, resulting in changes in atomic positions and a loss of the center of symmetry. Consequently, this dipole distribution leads to charge accumulation on the material surface, generating localized electric fields that interact with the electrical activity of surrounding cells. Therefore, by incorporating biomaterials with this characteristic, it is possible to establish an electrical interaction with bone tissue and stimulate regeneration [6].

In this context, to apply this therapeutic approach to bone regeneration, it is necessary to use a material that is both biocompatible and piezoelectric. One biomaterial that fits this proposal is the gelatin methacrylate (GelMA) scaffold, which provides structural support for regeneration by mimicking the extracellular matrix of bone tissue. Therefore, incorporating piezoelectric particles into this scaffold can increase the regeneration rate by electrically stimulating bone cells, promoting cell proliferation and migration.

Furthermore, another material that could be combined with the scaffold is the whitlockite nanoparticle (WK: $Ca_{18}Mg_2(HPO_4)_2(PO_4)_{12}$). Unlike hydroxyapatite, the most used calcium phosphate in bone scaffolds [7], WK exhibits intrinsic piezoelectric behavior, making it particularly suited to the electrical stimulation strategy described above. In addition, WK mimics the mineral composition of bone tissue and may stimulate osteogenic differentiation by suppressing osteoclastic activity while continuously releasing PO_4^{3-} and Mg^{2+} ions [8]. Moreover, zinc doping of this scaffold may also enhance bone regeneration by inhibiting bone resorption and consequently increasing the regeneration rate. In addition, zinc incorporation is biocompatible and does not produce undesirable reactions in the body [9].

Should the accelerated bone regeneration hypothesized for this piezoelectrically driven scaffold be confirmed in future studies, the potential clinical impact could include reducing hospital stay and hospital resource consumption. This would improve patients' quality of life, since severe bone fractures are not only debilitating but also reduce patients' autonomy and require long recovery periods. Furthermore, it could reduce hospital costs associated with this condition, making treatment more affordable and enabling greater investment in other healthcare sectors. Finally, since it is produced by 3D bioprinting, this treatment is more personalized and scalable, with the potential to reach not only large hospitals but also, in the future, the Brazilian Unified Health System (SUS).

As a first step toward the development of such a piezoelectrically-driven scaffold, the general objective of this study is to investigate the influence of incorporating pure and zinc-doped whitlockite nanoparticles into GelMA hydrogels on their rheological properties, with the aim of applying them as bioinks for 3D printing. Furthermore, the specific objectives are: (1) to synthesize zinc-doped whitlockite nanoparticles; (2) to characterize the rheological properties (sol-gel transition temperature, viscosity, LVR, and recovery) of GelMA hydrogels containing different concentrations of pure and zinc-doped whitlockite; and (3) to compare the effect of incorporating the different whitlockite formulations on the rheological behavior of the hydrogels. This study further aims to describe, on an exploratory basis, whether the incorporation of different whitlockite formulations is associated with changes in the rheological properties of GelMA hydrogels, based on the single-sample characterization performed for each formulation

II. METODOLOGY

A. Gelatin Methacryloyl (GelMA) Synthesis

To synthesize gelatin methacryloyl (GelMA), 10% (w/v) type A porcine skin gelatin was added to a bicarbonate buffer solution (pH 7.4) at 50 °C and mixed until complete dissolution. During stirring, 10% (w/v) glycidyl methacrylate (GMA) was added to the gelatin solution at 0.5 mL/min and maintained for 3 hours. Subsequently, the solution was dialyzed against distilled water using a dialysis membrane for one week, with daily water replacement, while kept at room temperature and protected from light, to remove unreacted glycidyl methacrylate and reaction by-products. Finally, the solution was lyophilized and stored at room temperature until further use.

To confirm the functionalization of the gelatin, the synthesized product was analyzed by Nuclear Magnetic Resonance (NMR). This technique is based on the spins of atomic nuclei, which selectively absorb radiofrequency energy, generating peaks corresponding to the contribution of protons in different chemical environments [10]. Thus, the analysis is based on evaluating the area under each peak by integrating it, thereby quantifying the proportions of different functional groups in the sample.

To determine the degree of functionalization of the gelatin methacryloyl, the NMR spectra were analyzed. First, the peak corresponding to the methacrylate protons of methacryloyl groups grafted onto lysine and hydroxylysine residues (6.1 ppm) was evaluated, as it indicates the functionalization of gelatin with methacrylamide groups, which is essential for hydrogel polymerization and formation. Next, the second peak analyzed corresponded to the methacrylate groups (5.75 ppm), which also play an important role in the functionalization and properties of the hydrogel. Finally, the peak corresponding to the non-functionalized amine groups (2.99 ppm) was analyzed. The degree of methacrylation was then calculated using the following equation:

$$DM(\%) = \frac{N_{functionalized}}{N_{amine}} \times 100 \quad (1)$$

B. Whitlockite Synthesis and Characterization

Whitlockite synthesis consists of mixing an acidic solution of H_3PO_4 at a concentration of 6.5 % (w/v) with an alkaline solution containing 5.7 % (w/v) $Ca(OH)_2$ and 1.35 % (w/v) $Mg(OH)_2$. The alkaline solution was prepared under stirring at 70 °C and maintained under these conditions for 1 hour. Subsequently, the acidic solution was added dropwise to the mixture at a rate of 12.5 mL/min. The reaction was then transferred to an autoclave and maintained at 70 °C for 48

hours. The resulting material was washed to remove residual unreacted acid and base, lyophilized to remove the remaining water, and finally calcined at 700 °C in a ceramic crucible using a muffle furnace for 1 hour.

For zinc-doped whitlockite, Mg^{2+} ions were partially substituted by Zn^{2+} ions. For this purpose, $Zn(NO_3)_2$ was added to the synthesis according to a molar proportionality between the compounds. The selected zinc doping levels were 0.5 and 1 mol %. Thus, during the preparation of the alkaline solution, the corresponding mass percentages of $Mg(OH)_2$ were replaced by $Zn(NO_3)_2$ while maintaining the stoichiometric proportion. The subsequent synthesis steps were carried out without further modifications. Finally, the three synthesized samples were characterized by X-ray Diffraction (XRD) to verify the formation of the whitlockite crystalline phase.

The formation of the whitlockite phase was confirmed by X-ray diffraction (XRD), validating the synthesis route employed prior to incorporation into the GelMA hydrogel. The diffractograms were compared with the reference standard for whitlockite.

C. *Hydrogel Synthesis*

The hydrogel was prepared by mixing 10% (w/v) gelatin methacryloyl (GelMA) with whitlockite nanoparticles at 35 °C until complete homogenization. Whitlockite concentrations of 1% (w/w) and 10% (w/w) were used in the hydrogel formulations. The final formulations analyzed in this study are presented in Table 1.

Sample	**Composition (in w/w %)**
Control	GelMA
WK1	GelMA + 1 % whitlockite
WK10	GelMA + 10 % whitlockite
WK1-Zn0.5	GelMA + 1 % Zn0.5-doped whitlockite
WK10-Zn0.5	GelMA + 10 % Zn0.5-doped whitlockite
WK1-Zn1	GelMA + 1% Zn1-doped whitlockite
WK10-Zn1	GelMA + 10% Zn1-doped whitlockite

Table 1. Composition of the hydrogel formulations.

D. *Rheological Characterization*

Rheological measurements were performed on uncrosslinked hydrogels using a HAAKE PCR rheometer (Thermo Scientific) with plate-plate geometry (diameter: 1.5 cm). Four tests were conducted: (i) oscillatory amplitude sweep to determine the linear viscoelastic region (LVR) and storage (G') and loss (G") moduli; (ii) flow curve (shear rate 20–100 s^{-1}) to evaluate viscosity and shear stress as a function of shear rate; (iii) temperature sweep (10–50 °C) to determine the sol-gel transition temperature (Tsol-gel); and (iv) three-interval thixotropy test (3ITT) to evaluate the structural recovery of the hydrogels after high shear deformation. All rheological measurements were performed on a single sample per formulation, without independent replicates; consequently, the results reported in this study are descriptive, and no inferential statistical analysis was performed.

III. RESULTS

Initial characterization confirmed WK formation during synthesis. The undoped WK was characterized in a previous work of our research group [11], while Zn-doped WK, analyzed post-calcination, also showed the whitlockite phase at 0.5 and 1 mol% doping (Fig. 1). XRD peaks matched ICDD card 00-070-2064, indicating successful synthesis. No secondary phases appeared in 1 mol% Zn-doped WK, suggesting high purity and Zn^{2+} incorporation into the crystal lattice, replacing Mg^{2+}, consistent with calcium phosphate doping reports [12]. The 0,5 mol% sample showed additional hydromagnesite and calcite phases, likely due to carbonate formation on calcium phosphate surfaces that interact with CO_2, especially under humid conditions [13]. Despite these secondary phases, Zn0.5-WK was used in rheological tests to study its effect on GelMA. Although calcite was an undesired phase, it is a bioactive biomaterial and has demonstrated promising biological applications in bone regeneration through both in vitro and in vivo studies [14].

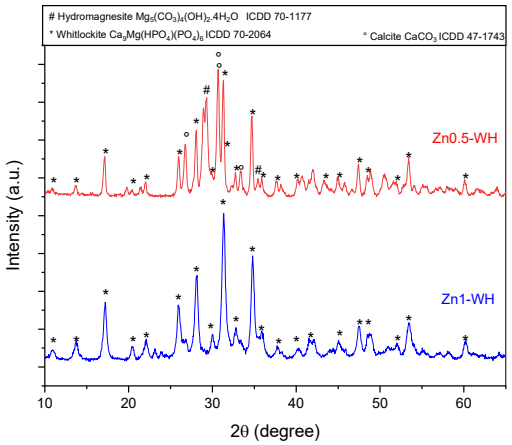


Fig. 1 – X-ray diffraction patters of the Zn-doped WK after calcination: results of the Zn1-WK in blue and Zn0.5-WK in red.

The produced hydrogels were characterized regarding their rheological properties in order to evaluate the influence of incorporating undoped and zinc-doped whitlockite nanoparticles in GelMA hydrogels. Initially, the sol-gel transition temperature ($T_{sol-gel}$) was evaluated. This parameter represents the critical temperature at which gelatin macromolecules, dispersed in an aqueous medium, adopt a random-coil conformation after losing their triple-helix structure. Since the triple-helix structures are crosslinked through intermolecular interactions, their loss leads to loss of physical entanglement, thereby resulting in loss of gel consistency [10]. This temperature is determined by the crossover between the storage modulus (G') and the loss modulus (G") during a heating ramp (Figure 2), indicating the temperature range in which the hydrogel remains in the gel state. The sol-gel transition temperature ranged from 34.02 °C to 37.17 °C, and the addition of whitlockite increased it relative to the control groups across all analyzed formulations.

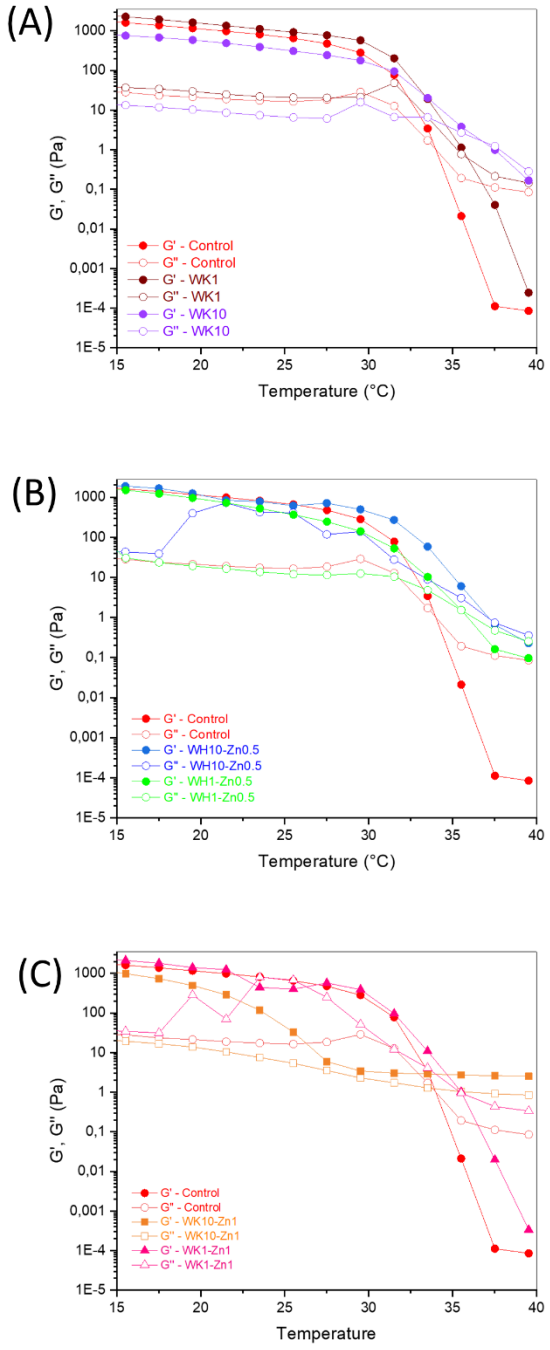


Fig. 2. Determination of the sol-gel transition temperature ($T_{sol-gel}$) of the different GelMA hydrogel formulations through the crossover between the storage modulus (G') and the loss modulus (G''). The panels show: (A) control and non-doped whitlockite formulations; (B) control and 0.5% zinc-doped whitlockite formulations; and (C) control and 1% zinc-doped whitlockite formulations

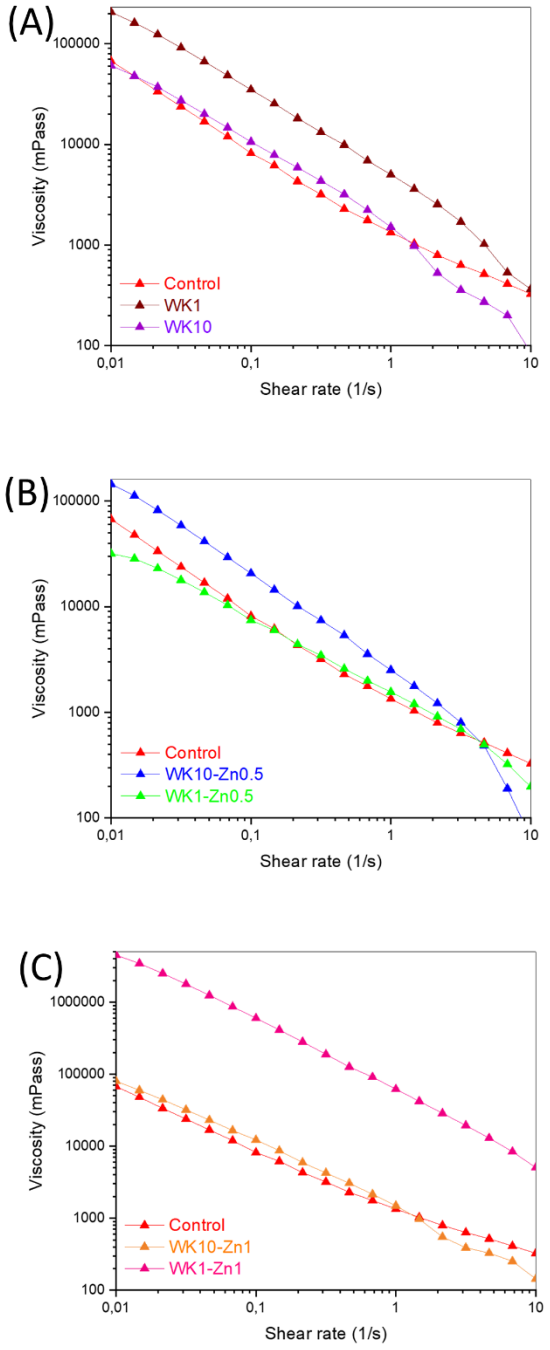


Fig. 3. Viscosity as a function of shear rate for the different GelMA hydrogel formulations.The panels show: (A) control and non-doped whitlockite formulations; (B) control and 0.5% zinc-doped whitlockite formulations; and (C) control and 1% zinc-doped whitlockite formulations

The viscosity of the hydrogels was measured as a function of shear rate to assess whether they behave as pseudoplastic fluids, and the results are shown in Figure 3. All formulations exhibited pseudoplastic behavior, characterized by a decrease in viscosity as the applied shear rate increased.

The flow curve test, which measures the shear stress as a function of shear rate, was also performed to evaluate complementary data regarding fluid behavior, and is shown in Figure 4. All formulations exhibited a local maximum in shear stress (stress overshoot) at shear rates between 37.31 s^{-1} (WK10) and 68.06 s^{-1} (WK10-Zn0.5), followed by a progressive decrease in shear stress. All formulations

containing 1% undoped or doped whitlockite exhibited the highest yield shear stress, followed by hydrogels with 10% whitlockite, whereas the control sample showed the lowest peak stress. The maximum shear stress reached approximately 1549.9 Pa (WK1-Zn1) and the minimum 561.86 Pa (Control).

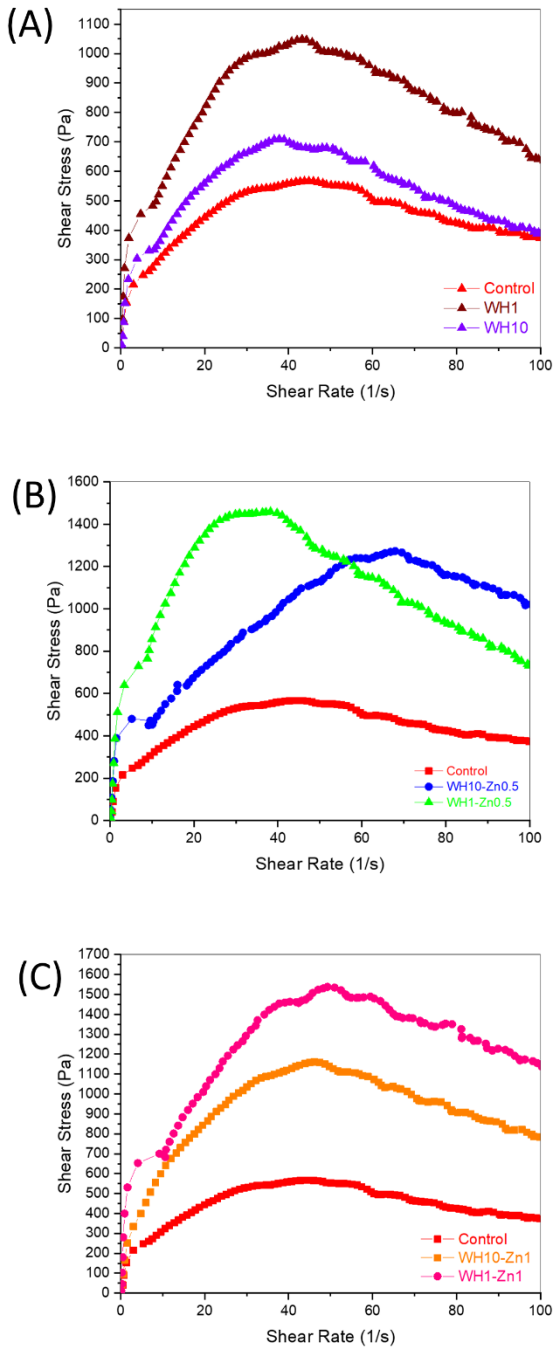


Fig. 4. Shear stress as a function of shear rate (flow curve) for the different GelMA hydrogel formulations. The panels show: (A) control and non-doped whitlockite formulations; (B) control and 0.5% zinc-doped whitlockite formulations; and (C) control and 1% zinc-doped whitlockite formulations

The linear viscoelastic region (LVR) was determined by shear strain sweep tests, and the results are shown in Figure 5. This analysis identifies the critical deformation limit at which changes in the internal structure of the hydrogel occur, corresponding to the crossover between G' and G". The LVR values ranged from 3.02 to 8.01, with the addition of undoped or Zn-doped whitlockite increasing this LVR region compared with the control group in all analyzed formulations.

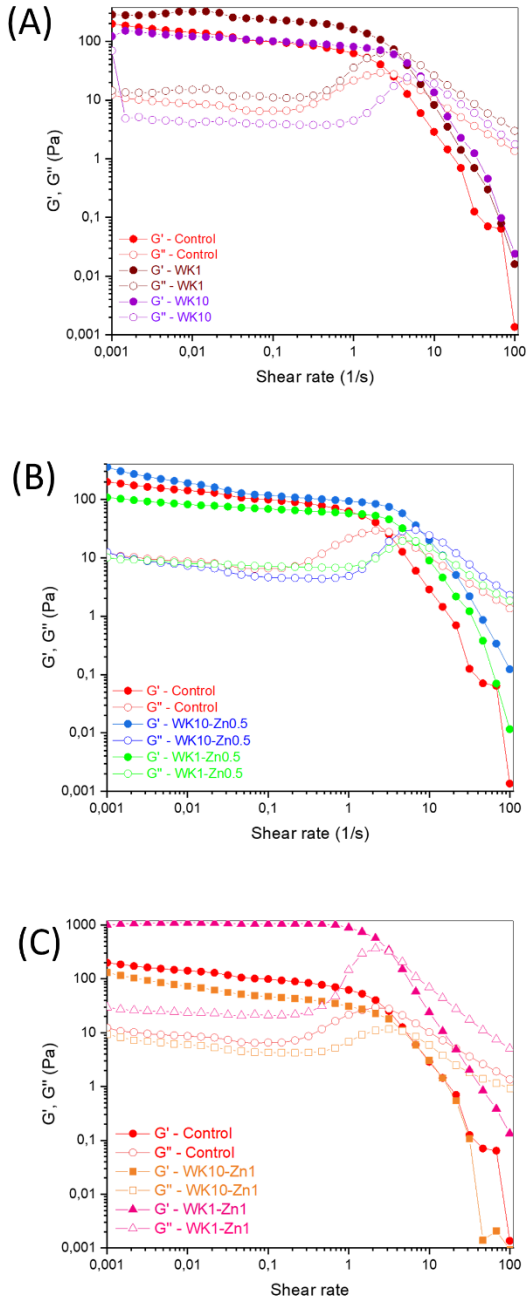


Fig. 5. Determination of the linear viscoelastic region (LVR) of the different GelMA hydrogel formulations by strain sweep analysis. The panels show: (A) control and non-doped whitlockite formulations; (B) control and 0.5% zinc-doped whitlockite formulations; and (C) control and 1% zinc-doped whitlockite formulations

Finally, the structural recovery capacity of the hydrogels was evaluated using a three-interval thixotropy test, in which the material was initially deformed within its linear viscoelastic region (LVR), then beyond its critical limit (outside the LVR), and finally returned to its initial condition. The test showed viscosity recovery values ranging from 55.56% to 90.81% (Figure 6).

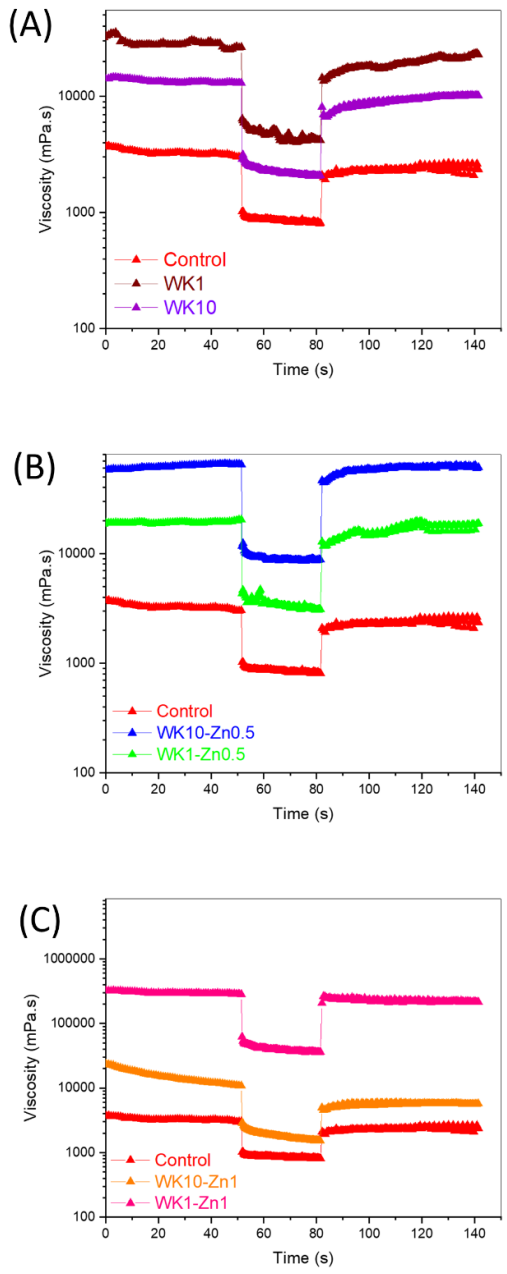


Fig. 6. Three-interval thixotropy test showing the structural recovery of the different GelMA hydrogel formulations after high shear deformation. The panels show: (A) control and non-doped whitlockite formulations; (B) control and 0.5% zinc-doped whitlockite formulations; and (C) control and 1% zinc-doped whitlockite formulations

The results obtained from the rheological tests described above are summarized in Table 2 to facilitate data visualization.

Table 2. Summary of the rheological properties of the synthesized GelMA hydrogels.

Sample	Tsol_gel (°C)	LVR	Recovery
Control	34.02	3.02	71.60%
WK1	35.85	3.52	73.23%
WK10	36.77	7.40	73.00%
WK1-Zn0.5	35.31	6.76	91.15%
WK10-Zn0.5	37.17	8.01	90.81%
WK1-Zn1	35.58	3.38	79.10%
WK10-Zn1	36.15	4.56	55.56%

IV. DISCUSSION

Rheological properties play a central role in the application of hydrogels as bioinks for 3D printing, since they determine both the material's ability to be extruded through the printing nozzle and its structural stability after deposition [9]. Gelatin methacryloyl (GelMA) exhibits thermoresponsive behavior: at low temperatures, its polymer chains adopt a triple-helix conformation, promoting physical crosslinking and increasing viscosity, whereas heating disrupts these interactions. This conformational change characterizes the sol-gel transition. Therefore, maintaining hydrogels below the critical transition temperature, in the region where G' > G", is important for shape fidelity and structural stability during 3D printing.

In this scenario, the increase in $T_{sol-gel}$ (Table 2) observed in the formulations containing undoped or Zn-doped whitlockite is consistent with previous reports, since ceramic nanoparticles have been described as reinforcing agents in hydrogels by creating intermolecular interactions with the polymeric chains that improve structural stability and resistance to deformation [15]. Furthermore, the slight increase in the sol-gel transition was mainly attributed to the ceramic charge, whereas zinc doping did not affect this property. At both whitlockite concentrations (1% and 10%), no apparent difference in transition temperature was observed between doped and undoped formulations. Moreover, all formulations exhibited $T_{sol-gel}$ above 34°C, which is favorable from a processing standpoint, as the material remains in the gel state at room temperature, facilitating its handling as a bioink.

Furthermore, pseudoplastic behavior was observed across all formulations (Figure 3), consistent with the expected behavior of hydrogels before chemical crosslinking. This property allows the material to be extruded through the printing needle by reducing its viscosity under shear stress, while recovering its viscosity at rest, thereby preventing excessive spreading or deformation of the deposited filament [16]. The classification as pseudoplastic fluids is also supported by the ascending branch of the flow curves (Figure 4), which shows the characteristic nonlinear, concave shear stress–shear rate relationship of a pseudoplastic fluid, with a characteristic yield stress [16]. Below the yield stress, the

physical entanglement of the polymeric network can be elastically deformed; above it, physical entanglement within the triple helix and intermolecular interactions cannot accommodate the deformations, and the fluid deforms plastically. The yield point increased with increasing amounts of doped or undoped WK in the hydrogels, suggesting that the ceramic powder interacts with the GelMA polymeric chains, thereby contributing to energy storage via intermolecular interactions under mechanical deformation. The formulation WK10-Zn0.5 exhibited anomalous behavior compared with Zn1-doped WH or undoped WH-containing hydrogels, which may be attributed to secondary phases formed during synthesis. The WK10-Zn0.5 formulation exhibited greater deformation at the yield point, suggesting strong interactions between the carbonate phases and the GelMA network. These results are interesting because they indicate that WK does not affect physicochemical interactions between the ceramic and the polymeric matrix, whereas carbonate phases may cause significant physicochemical interactions at high concentrations (10 wt.%). More studies are needed to determine whether this behavior extends to other calcium phosphate ceramics beyond WK.

Regarding the linear viscoelastic region (LVR), LVR values increased for formulations containing higher concentrations of whitlockite, particularly WK10 and WK10-Zn0.5 (Figure 5). This behavior indicates that these hydrogels withstand greater deformation before their internal structure is compromised, another characteristic that, together with pseudoplastic behavior, is essential for bioink printability, ensuring structural integrity and dimensional fidelity.

Finally, the structural recovery test is consistent with the predominantly physical (non-covalent) nature of the interactions sustaining the hydrogel network. Under high deformation, these junctions are disrupted, allowing the material to flow and reorganize once the mechanical stimulus is removed, restoring a large portion of the original viscosity. The highest recovery was observed for WK1-Zn0.5 (91.15%), closely followed by WK10-Zn0.5 (90.81%), suggesting a network with a greater capacity for physical reorganization after deformation, possibly favored by stronger mineral-polymer interactions. These results are consistent with those found in the flow curves (Figure 4), where the secondary phases found in the Zn0.5-WK ceramic were associated with stronger interactions with the GelMA chains.

Overall, the rheological results indicate that the incorporation of whitlockite, but not its zinc doping, modulates both the operating temperature of the hydrogel and its resistance to deformation and recovery capacity. Besdies, secondary carbonate phases showed stronger interactions with the GelMA network than the WK phases. These characteristics are relevant to applications involving material handling, injection, or 3D printing prior to crosslinking. However, the rheological characterization was performed using a single sample per formulation, without independent replicates; therefore, the differences observed among formulations should be interpreted as descriptive trends.

Although the present study focused exclusively on rheological characterization, incorporating whitlockite may support the future development of bioinks with piezoelectric potential, as it has been reported to exhibit piezoelectric behavior. This study does not directly demonstrate actual printability, mechanical performance, piezoelectric response, biological performance, or osteogenic effects. These aspects should be evaluated in future studies to further assess the suitability of the formulations for tissue engineering and their potential for regenerative applications.

V. CONCLUSION

The results of this study show that incorporating pure and zinc-doped whitlockite nanoparticles modifies the rheological properties of GelMA hydrogels, influencing parameters relevant to their application as bioinks. Overall, an increase in the sol-gel transition temperature ($T_{sol-gel}$) and in the linear viscoelastic region (LVR) was observed, while all formulations maintained pseudoplastic behavior and structural recovery capacity, which are relevant characteristics for 3D printing processes. Although it was found that zinc doping in the WK ceramics did not lead to significant changes in the properties of the formulations, the results indicate that the incorporation, in terms of mass fraction, of whitlockite represents a promising strategy for the development of biomaterials for tissue engineering, with piezoelectric functionality — reported for whitlockite in the literature — to be experimentally explored in future studies. *Additionally, the interaction between GelMA and carbonate phases must be further explored, as these phases strongly modulate GelMA properties.* Future studies involving independent experimental replicates, 3D printing assays, mechanical characterization, piezoelectric measurements, and biological evaluation will be necessary to further assess the suitability of these formulations for tissue engineering and their potential for regenerative applications.

REFERENCES

[1] M. Bohner, "Resorbable biomaterials as bone graft substitutes," Materials Today, vol. 13, no. 1–2, pp. 24–30, 2010.

[2] C. Ma, H. Wang, Y. Chi, Y. Wang, L. Jiang, N. Xu, Q. Wu, Q. Feng, and X. Sun, "Preparation of oriented collagen fiber scaffolds and its application in bone tissue engineering," Biomaterials, 2020.

[3] N. Ramesh, S. C. Moratti, and G. J. Dias, "Hydroxyapatite-polymer biocomposites for bone regeneration: A review of current trends," Journal of Biomedical Materials Research Part B: Applied Biomaterials, vol. 106, no. 5, pp. 2046–2057, 2018.

[4] M. V. Regí and E. R. Hernández, *Bioceramics: From Bone Regeneration to Cancer Nanomedicine*. Cham, Switzerland: Springer, 2020.

[5] C. Ning, L. Zhou, and G. Tan, "Fourth-generation biomedical materials," Materials Today, 2015.

[6] K. Kapat, Q. T. H. Shubhra, M. Zhou, and S. Leeuwenburgh, "Piezoelectric Nano-Biomaterials for Biomedicine and Tissue Regeneration," Advanced Functional Materials, vol. 30, no. 44, 2020.

[7] P. Kazimierczak, J. Wessely-Szponder, K. Palka, A. Barylyak, V. Zinchenko, and A. Przekora, "Hydroxyapatite or Fluorapatite—Which Bioceramic Is Better as a Base for the Production of Bone Scaffold?—A Comprehensive Comparative Study," International Journal of Molecular Sciences, vol. 24, no. 6, art. 5576, 2023.

[8] D. H. Kim et al., "Biomimetic whitlockite inorganic nanoparticles-mediated in situ remodeling and rapid bone regeneration," Biomaterials, vol. 112, pp. 31–43, 2016.

[9] C. Padilla, F. Quero, M. Pępczyńska, P. Díaz-Calderon, J. P. Acevedo, N. Byres, J. J. Blaker, W. MacNaughtan, H. E. L. Williams, and J. Enrione, "Understanding the Molecular Conformation and Viscoelasticity of Low Sol-Gel Transition Temperature Gelatin

Methacryloyl Suspensions," *International Journal of Molecular Sciences*, vol. 24, no. 8, art. 7489, 2023.

[10] K. Yue, G. Trujillo-de Santiago, M. M. Alvarez, A. Tamayol, N. Annabi, and A. Khademhosseini, "Synthesis, properties, and biomedical applications of gelatin methacryloyl (GelMA) hydrogels," Biomaterials, vol. 73, pp. 254-271, 2015.

[11] E. C. Macedo Lemos, R. H. Moriya, L. C. Teruel de Jesus, R. Borges, R. R. de Oliveira, N. B. de Lima, J. T. Arantes, and J. Marchi, "Magnesium whitlockite structure: an ab initio approach supported by experimental characterization," *Ceram. Int.*, vol. 51, no. 22, pp. 36637–36643, 2025.

[12] R. Cruz et al., "Does the incorporation of zinc into calcium phosphate improve bone repair? A systematic review,"Ceram. Int., vol. 44, no. 2, pp. 1240–1249, 2018.

[13] H. S. Santos, H. Nguyen, F. Venâncio, D. Ramteke, R. Zevenhoven, and P. Kinnunen, "Mechanisms of Mg carbonates precipitation and implications for CO_2 capture and utilization/storage," *Inorganic Chemistry Frontiers*, vol. 10, no. 9, pp. 2507–2546, 2023.

[14] M. Mozafari, S. Banijamali, F. Baino, S. Kargozar, and R. G. Hill, "Calcium carbonate: Adored and ignored in bioactivity assessment," *Acta Biomater.*, vol. 91, pp. 35–47, 2019.

[15] M. Itxaso Calafel, M. Criado-González, R. Aguirresarobe, M. Fernández, and C. Mijangos, "From rheological concepts to additive manufacturing assessment of hydrogel-based materials for advanced bioprinting applications," Materials Advances, vol. 6, no. 14, pp. 4566–4597, 2025.

[16] M. Mirzagoli, F. Ganji, and L. Tayebi, "Injectable hydrogels for bone regeneration: Mechanical reinforcement strategies using nanoparticles and nanofibers," ADMET & DMPK, vol. 14, 2026.

Impacto do *Slicing Aided Hyper Inference* (SAHI) na detecção de lesões de retinopatia diabética usando YOLO

Cezário Luiz Mourão C. Mesquita
dept. engenharia de teleinformática
Universidade Federal do Ceará
cezariomourao@alu.ufc.br

José Valclemir Rodrigues da Silva
dept. engenharia de teleinformática
Universidade Federal do Ceará
valclemir15@hotmail.com

Paulo César Cortez
dept. engenharia de teleinformática
Universidade Federal do Ceará
cortez@lesc.ufc.br

***Resumo*—A retinopatia diabética (RD) é uma das principais causas de perda de visão evitável no mundo e se manifesta, em seus estágios iniciais, por lesões de fundo de olho como microaneurismas, exsudatos duros, exsudatos moles e hemorragias. Sua detecção automatizada é desafiadora, pois as imagens têm alta resolução e as lesões ocupam pouquíssimos *pixels*, sendo os microaneurismas, sinal mais precoce da doença, os mais difíceis. Este trabalho investiga, de forma controlada e estratificada por tipo de lesão, se a inferência assistida por fracionamento de imagem *Slicing Aided Hyper Inference* (SAHI) melhora a detecção de lesões de RD quando aplicada a um detector YOLO moderno, no conjunto de dados público DDR. Sob validação cruzada repetida e estratificada por lesão (5 *folds* $\times$ 3 repetições) e protocolo unificado de avaliação (*mean Average Precision*, mAP, a IoU 0,1 e 0,5), comparam-se três configurações: (A) treino e inferência em imagem cheia; (B) SAHI apenas na inferência; e (C) SAHI também no ajuste fino. A Config. B *reduziu* significativamente o desempenho global em relação ao *baseline* YOLOv8m ($p < 0{,}01$, Wilcoxon pareado), enquanto a Config. C aproximou o mAP global do *baseline*, sem diferença detectável sob o teste corrigido para a não-independência dos *folds*. O ganho concentrou-se nos microaneurismas: a Config. C elevou significativamente o AP@0,1 ($p_{\mathrm{Holm}} = 0{,}009$; $r = 0{,}82$) e aumentou o AP@0,5 de 0,059 para 0,111, ao custo de piora nas lesões maiores e difusas. Conclui-se que, para detectores YOLO modernos em alta resolução, o ganho do SAHI é localizado: não melhora o mAP global, mostra-se eficaz na detecção de microaneurismas no DDR.**

***Palavras-chave*—retinopatia diabética, detecção de lesões, YOLO, SAHI.**

I. Introdução

A retinopatia diabética (RD) é uma complicação microvascular da retina e uma das principais causas de perda irreversível de visão em pessoas com diabetes [1], [2]. No Brasil, mais de 13 milhões de pessoas vivem com diabetes, cerca de 6,9% da população [3], e a RD atinge entre 24% e 39% desse grupo [4]. Quando diagnosticada e tratada precocemente, o risco de cegueira pode ser reduzido para menos de 5% [4].

A RD pode ser caracterizada pela presença de determinados tipos de lesões na retina, tais como exsudatos duros (EX), exsudatos moles ou algodonosos (SE), microaneurismas (MA) e hemorragias (HE) [2]. Os microaneurismas, em particular, costumam ser as primeiras alterações visíveis associadas à doença e, ao mesmo tempo, as mais difíceis de detectar, por ocuparem pouquíssimos *pixels* nas imagens de fundo de olho. A identificação dessas lesões em fases iniciais é essencial para prevenir a progressão da doença a estágios mais graves e a consequente perda de visão.

Os exames que identificam essas lesões dependem de médicos especializados, cuja disponibilidade é limitada, o que pode retardar o diagnóstico, sobretudo em regiões remotas e de difícil acesso [5]. A análise humana está ainda sujeita a fatores subjetivos, como cansaço e nível de experiência do especialista [6], o que compromete a precisão justamente nos estágios iniciais, quando as alterações são sutis. Nesse cenário, a detecção automática por aprendizado profundo tem sido proposta como ferramenta de apoio ao especialista: detectar e classificar cada lesão, e não apenas graduar a imagem, oferece informação para o diagnóstico precoce, de forma mais rápida e menos suscetível à subjetividade do avaliador.

Automatizar essa tarefa, contudo, permanece um problema em aberto. No conjunto de dados público DDR (*Dataset for Diabetic Retinopathy*) [1], um dos maiores *benchmarks* da área, os detectores genéricos avaliados pelos próprios autores obtiveram mAP inferior a $0{,}01$ a IoU 0,5, com AP de microaneurismas abaixo de 1% [1], [7]. Esse desempenho reduzido decorre do tamanho e do formato das lesões, do ruído e do contraste das imagens, do reduzido número de exemplos anotados e da dificuldade dos algoritmos de aprendizado profundo em detectar objetos muito pequenos [5].

Para mitigar o problema dos objetos minúsculos, a literatura recente tem explorado estratégias diversas: realçar seletivamente as regiões mais informativas da imagem [8], [9], ampliar artificialmente sua resolução [10] ou modelar relações entre regiões distantes, em vez de apenas vizinhanças locais [11]. Uma alternativa independente da arquitetura do detector consiste em fracionar a imagem em janelas sobrepostas (*tiling*/*slicing*), de modo que cada janela seja processada em resolução efetiva mais alta. O *framework* SAHI (*Slicing Aided Hyper Inference*) [12] implementa essa ideia de forma

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob o Processo nº 315447/2023-0, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Código de Financiamento 001, pelo apoio a esta pesquisa.

genérica, aplicável na inferência e, opcionalmente, também no ajuste fino do detector. No contexto do DDR, trabalhos anteriores demonstraram que o fracionamento eleva o desempenho da detecção de lesões, seja por *tiling* convencional [5], seja pelo SAHI [7]. Entretanto, esses estudos empregaram arquiteturas YOLO de gerações anteriores (YOLOv4/v5, YOLOR-CSP), permanecendo uma lacuna de investigação do SAHI combinado a detectores YOLO mais recentes, sob um protocolo de comparação controlado e com análise estratificada por tipo de lesão.

Diante dessa lacuna, este trabalho avalia, de forma controlada e estratificada por tipo de lesão, se o SAHI melhora a detecção de lesões de RD quando aplicado a um detector YOLO moderno (YOLOv8m), em comparação com a inferência em imagem completa, utilizando o conjunto de dados DDR. As principais contribuições deste trabalho são:

- o emprego de um detector YOLO moderno (YOLOv8m) combinado ao SAHI no DDR;
- uma comparação *controlada* entre YOLO e YOLO+SAHI sob um protocolo unificado de pré-processamento, resolução, validação cruzada e avaliação, que isola o efeito do *slicing* (mesmo detector, mesmos *folds*);
- uma análise *estratificada por classe* de lesão (MA, EX, SE e HE), com teste estatístico pareado, que evidencia o *trade-off* do *slicing*, ganho nas menores lesões (microaneurismas) e perda nas maiores e mais difusas (exsudatos moles e hemorragias), demonstrando que seu benefício sobre um YOLO moderno em alta resolução é condicional à classe de lesão e ao regime de treino.

O restante do artigo apresenta a fundamentação teórica (Seção II), os trabalhos relacionados (Seção III), os materiais e métodos (Seção IV), os resultados e a discussão (Seção V) e as considerações finais (Seção VI).

II. Fundamentação Teórica

A. Detecção de objetos

A detecção de objetos consiste em delimitar cada instância presente na imagem por uma *bounding box* (caixa retangular que circunscreve o objeto) e atribuir-lhe uma classe [13]. Os detectores baseados em aprendizado profundo dividem-se em detectores de dois estágios, que primeiro propõem regiões de interesse e depois as classificam, como o Faster R-CNN [14], e detectores de estágio único, que tratam a detecção como um problema de regressão e predizem simultaneamente as caixas e as probabilidades de classe sobre a imagem completa, como o SSD [15] e a família YOLO (*You Only Look Once*) [16]. A principal vantagem dos detectores de estágio único é a velocidade de inferência, o que os torna adequados a aplicações em tempo real e a *hardware* de menor custo.

B. A arquitetura YOLOv8

Neste trabalho adota-se o YOLOv8 [17], na variante `m` (*medium*). Trata-se de um detector *anchor-free* com *backbone* do tipo CSPDarknet, *neck* PAN-FPN e cabeça de detecção desacoplada, que realiza a predição em três escalas (P3/P4/P5) com cabeças especializadas em objetos pequenos, médios e grandes, característica relevante para a detecção de lesões de tamanhos variados na retina. A variante `m` equilibra capacidade e custo computacional, sendo compatível com o treino em uma única GPU.

C. Slicing Aided Hyper Inference (SAHI)

O SAHI, proposto por Akyon *et al.* [12], é uma solução de código aberto que fornece um processo genérico de inferência e ajuste fino auxiliado por fracionamento. A imagem de entrada é dividida em janelas (*tiles*) sobrepostas; a detecção é realizada em cada janela e as predições são reagrupadas na imagem original por pós-processamento (por exemplo, supressão não máxima gulosa, *GREEDYNMM*). Ao processar regiões menores em resolução efetiva mais alta, o método amplia a visibilidade de objetos pequenos, sendo promissor para o problema dos microaneurismas. Pode ser aplicado apenas na inferência (*sliced inference*) ou também no treinamento (*slicing-aided fine-tuning*), regime no qual a literatura reporta ganhos maiores [12].

III. Trabalhos Relacionados

A. Evolução da família YOLO

A família YOLO evoluiu por sucessivas versões (YOLOv3, YOLOv4, YOLOv5, YOLOR, YOLOv7, YOLOv8, YOLOv11, entre outras), incorporando aprimoramentos de *backbone*, *neck* e *head*, além de técnicas de aumento de dados e funções de perda mais robustas. Essa evolução importa aqui porque os trabalhos que aplicaram *slicing* à detecção de lesões de RD empregaram gerações anteriores da família [5], [7], cujas limitações na detecção de objetos pequenos podem ter sido atenuadas nas versões recentes.

B. Detecção de lesões de RD no DDR

No estudo que apresentou o conjunto de dados, Li *et al.* [1] avaliaram detectores genéricos (SSD, YOLO e Faster R-CNN) e demonstraram a dificuldade central da tarefa: enquanto a classificação alcançou acurácia da ordem de 0,83 com a arquitetura SE-BN-Inception, os modelos de segmentação e de detecção tiveram desempenho fraco, com AP de microaneurismas abaixo de 1%, o que evidencia a necessidade de aprimorar o reconhecimento de microlesões para viabilizar a aplicação clínica. Alyoubi *et al.* [18] combinaram uma rede de classificação (CNN512) a um YOLOv3 para localização de lesões, obtendo mAP de 0,216 na detecção no DDR.

Uma linha de trabalhos correlatos explorou progressivamente a família YOLO nessa base. Santos *et al.* [19] propuseram uma abordagem baseada em YOLOv4 e, depois [5], um método baseado em YOLOv5 com pré-processamento, aumento de dados e *Tiling*, alcançando mAP de 0,2630 na validação e 0,1540 no teste do DDR (IoU 0,5). Em seguida, Pereira *et al.* [7] combinaram a arquitetura YOLOR-CSP ao *framework* SAHI, atingindo mAP de 0,3808 na *validação* do DDR (IoU 0,5), mas de apenas 0,2225 no *teste*; essa diferença de 0,16 no mesmo modelo ilustra o quanto números de um único *split* de validação podem ser otimistas.

C. Lacuna de pesquisa

Como a combinação YOLO+SAHI já foi aplicada ao DDR [7], a comparação direta "YOLO vs. YOLO+SAHI" não constitui, por si só, uma novidade. O que permanece em aberto, e é o objeto deste trabalho, é saber se o ganho atribuído ao *slicing* persiste sobre um detector YOLO moderno, quando medido sob protocolo unificado, sem o viés de *split* único e de forma estratificada por tipo de lesão.

IV. Materiais e Métodos

A. Visão geral do método

O fluxo experimental é ilustrado na Fig. 1. Não há aquisição própria de imagens: utilizam-se as 757 imagens do conjunto público DDR [1] que possuem anotações de lesão em nível de *bounding box* (Seção IV-B). Essas anotações, distribuídas em formato Pascal VOC, são convertidas para o formato YOLO, com deduplicação das caixas repetidas e unificação dos três *splits* oficiais em um único *pool* (Seção IV-C), sobre o qual se aplica validação cruzada repetida e estratificada pelo vetor de presença de lesão, tendo a imagem como unidade de divisão (Seção IV-E). Em cada *fold*, o mesmo detector YOLOv8m é treinado sob dois domínios de entrada, imagem completa em alta resolução (Configs. A e B) e *tiles* de 512×512 (Config. C, *slicing-aided fine-tuning*), com hiperparâmetros fixos, de modo que a única variável experimental seja o fracionamento (Seção IV-F). A inferência ocorre sobre a imagem completa (Config. A) ou de forma fracionada com o SAHI (Configs. B e C), caso em que as predições de cada *tile* são reagrupadas nas coordenadas da imagem original por supressão não máxima gulosa. Por fim, as predições são confrontadas com o *Ground Truth* pelo avaliador COCO, calculando-se o mAP a IoU 0,1 e 0,5, global e por classe (Seção IV-G); como as três configurações compartilham os mesmos *folds*, os resultados são pareados e comparados por testes estatísticos pareados (Seção IV-H).

B. Conjunto de dados

Os experimentos utilizam o conjunto de dados público DDR [1], também chamado OIA-DDR, composto por 13.673 imagens de fundo de olho de 9.598 pacientes, coletadas em 147 hospitais de 23 províncias da China; destas, apenas 757 possuem anotações de detecção em nível de *bounding box*, sendo este o subconjunto empregado neste trabalho. As quatro classes consideradas são microaneurismas (MA), exsudatos duros (EX), exsudatos moles (SE) e hemorragias (HE). As imagens têm alta resolução e tamanhos variáveis (da ordem de 1088×1920 a 3456×5184 *pixels*), enquanto as lesões ocupam regiões diminutas, o que constitui a principal motivação para investigar o *slicing*. As anotações são bastante desbalanceadas, com SE presente em apenas 239 imagens e 1.293 instâncias, contra 23.659 instâncias de EX, conforme resumido na Tabela I. A densidade das microlesões e a grade de *tiles* do SAHI são ilustradas, sobre uma imagem real do DDR, na Fig. 2.

Tabela I
Distribuição das anotações de detecção do DDR (subconjunto de 757 imagens anotadas), após a deduplicação descrita na Seção IV-C.

Lesão	Imagens com a lesão	Instâncias	% instâncias
MA (microaneurismas)	570	10.380	21,7%
EX (exsudatos duros)	486	23.659	49,4%
SE (exsudatos moles)	239	1.293	2,7%
HE (hemorragias)	601	12.538	26,2%
Total	757	47.870	100%

Tabela II
Configurações experimentais avaliadas.

Config.	Treino	Inferência	Propósito
A	imagem cheia	imagem cheia	*baseline* sem *slicing*
B	imagem cheia	*sliced* (SAHI)	efeito do *slicing* só na inferência
C	*tiles* (*slicing-aided FT*)	*sliced* (SAHI)	regime completo de *slicing*

C. Conversão de anotações para o formato YOLO

O DDR fornece as *bounding boxes* oficiais em arquivos Pascal VOC (XML), organizadas nos *splits* `train`/`valid`/`test`. A conversão adotada parseia esses XML e escreve as caixas no formato YOLO (`classe cx cy w h`, com coordenadas normalizadas em $[0, 1]$), unindo os três *splits* em um único *pool* de 757 imagens, de modo que a divisão entre treino, validação e teste passa a ser feita pela validação cruzada (Seção IV-E). Verificou-se que os XML contêm caixas duplicadas exatas (fator próximo de $1{,}7\times$); aplica-se, portanto, deduplicação por caixa, resultando nas 47.870 instâncias da Tabela I. Após a conversão, as caixas foram validadas visualmente sobre uma amostra de imagens, para descartar erros de eixo ou de normalização. Na avaliação, adotou-se um lado mínimo de caixa de 2 *pixels*, o que evita áreas nulas no reagrupamento geométrico do SAHI.

D. Arquitetura de experimentos

Para isolar o efeito do *slicing*, três configurações são avaliadas sob o mesmo protocolo, conforme a Tabela II.

A configuração A é o *baseline* (sem *slicing*); a B isola o efeito do SAHI aplicado apenas na inferência, sobre o *mesmo* modelo treinado em imagem completa; e a C corresponde ao regime completo, em que o modelo também é treinado sobre *tiles*. As Configs. A e B derivam de um *único* treino por *fold*, avaliado de duas formas (imagem cheia e SAHI), o que torna a comparação A vs. B um teste exato do efeito da inferência fracionada.

E. Validação cruzada

Adota-se validação cruzada repetida e estratificada, com a imagem como unidade de divisão (as 757 anotadas), de modo que uma mesma imagem nunca apareça em mais de um

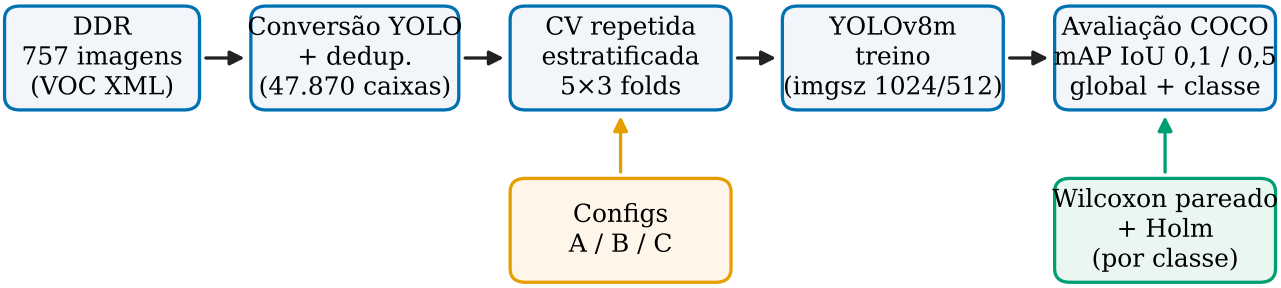


Figura 1. Visão geral do *pipeline* proposto: conversão das anotações VOC do DDR para o formato YOLO (com deduplicação), validação cruzada repetida e estratificada por lesão, treino do YOLOv8m, avaliação COCO (mAP a IoU 0,1 e 0,5, global e por classe) e teste estatístico pareado. As três configurações (A/B/C) compartilham os mesmos *folds*.

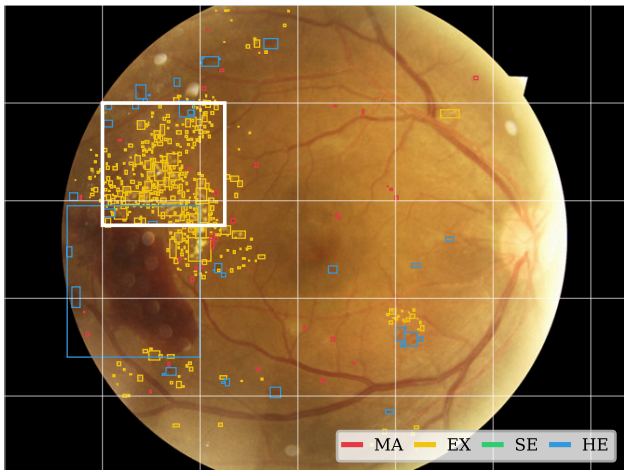


Figura 2. Imagem do DDR com as *bounding boxes* do *Ground Truth* (cores por classe) e a grade de *tiles* do SAHI (512 × 512, sobreposição 0,2); em destaque, um *tile*. A alta densidade de microlesões diminutas, muitas menores que a espessura da linha das caixas, motiva a inferência assistida por fracionamento.

subconjunto, evitando vazamento de dados. Como o DDR não disponibiliza identificadores de paciente para o subconjunto anotado, a partição é por imagem, não por paciente: imagens de um mesmo paciente podem, em princípio, ocorrer em *folds* distintos. Estimativas de partição única são instáveis em conjuntos pequenos, e a repetição reduz essa variância ao promediar múltiplas partições [20], [21]; no próprio DDR, Pereira *et al.* [7] reportam diferença de 0,16 de mAP entre validação e teste para o mesmo modelo, magnitude superior a qualquer efeito que se pretenda medir entre configurações.

Utiliza-se `RepeatedMultilabelStratifiedKFold` (5 *folds* × 3 repetições = 15 avaliações), estratificado pelo vetor binário de presença de lesão (tem MA / tem EX / tem SE / tem HE) de cada imagem [22]. Essa estratégia é essencial dado o desbalanceamento, sobretudo para SE, presente em 239 das 757 imagens: sob partição aleatória, os *folds* teriam contagens díspares dessa classe, inflando a variância de seu AP. Em cada *fold* externo, o conjunto de treino é novamente dividido de forma estratificada, reservando-se 80% para o ajuste dos pesos e 20% para a validação interna do *early-stopping*.

Como detector, hiperparâmetros (Tabela III), *seed*, *folds* e protocolo de avaliação são compartilhados, qualquer diferença entre A, B e C é atribuível ao fracionamento. O desempenho é reportado como média ± desvio-padrão do mAP entre os 15 *folds*.

F. Modelos e hiperparâmetros

O detector YOLOv8m foi implementado com a biblioteca `ultralytics` e o *slicing* com a biblioteca `sahi`, em Python. No *baseline* (Configs. A/B), as imagens são processadas em resolução elevada (lado maior redimensionado para 1024), a fim de preservar as lesões pequenas; nas inferências com SAHI, utilizam-se *tiles* de 512 × 512 com sobreposição de 0,2. Na Config. C, o modelo é treinado diretamente sobre *tiles* de 512 *px*, com *batch* maior, dado o menor consumo de memória por *tile*; treino e validação interna operam no domínio de *tiles*, para que o *early-stopping* reflita o mesmo domínio da inferência SAHI. Para preservar a morfologia das microlesões, emprega-se aumento de dados moderado e desativa-se o *mosaic*, que tende a prejudicar objetos minúsculos. A Tabela III lista os hiperparâmetros finais.

G. Métricas e protocolo de avaliação

Para comparar os modelos utiliza-se a *Intersection over Union* (IoU), que mede a sobreposição entre a *bounding box* predita e a do *Ground Truth*, conforme a Eq. (1):

$$\mathrm{IoU} = \frac{\mathrm{Area}\,(gt \cap pd)}{\mathrm{Area}\,(gt \cup pd)}, \tag{1}$$

em que gt é a caixa do *Ground Truth* e pd a caixa predita. A *Average Precision* (AP) é a área sob a curva de Precisão×Revocação de cada classe, e o *mean Average Precision* (mAP), a média das AP entre as classes, conforme a Eq. (2):

Tabela III
HIPERPARÂMETROS FIXOS (IDÊNTICOS ENTRE CONFIGURAÇÕES, SALVO OS *overrides* DE TREINO DA CONFIG. C).

Parâmetro	**Valor**
Detector / variante	YOLOv8m (`ultralytics`)
Otimizador / lr_0	SGD / 0,01
Resolução de treino (A/B)	1024 (lado maior)
Resolução de treino (C)	*tiles* 512
Batch (A/B | C)	8 | 16
Épocas (teto) / *patience*	100 / 20 (*early-stopping*)
Mosaic	desativado
SAHI *tile* / *overlap*	512×512 / 0,2
SAHI pós-proc. / *conf*	*GREEDYNMM* / 0,1
Seed	42

Tabela IV
DESEMPENHO GLOBAL (MAP, MÉDIA $\pm$ DESVIO ENTRE OS 15 *folds*) E TESTES PAREADOS. p_{Holm}: WILCOXON COM CORREÇÃO DE HOLM; r: TAMANHO DE EFEITO; p_{NB}: TESTE-T CORRIGIDO DE NADEAU–BENGIO. MELHOR VALOR EM **NEGRITO**.

Config.	**mAP@0,1**	**mAP@0,5**
A (*baseline*)	**0,300 $\pm$ 0,019**	**0,189 $\pm$ 0,015**
B (SAHI na inferência)	0,245 $\pm$ 0,033	0,095 $\pm$ 0,019
C (SAHI + *fine-tuning*)	0,290 $\pm$ 0,016	0,159 $\pm$ 0,012
Comparações pareadas (mAP@0,1):		
A vs. B	A>B; p_{Holm}=0,002, r=0,88, p_{NB}=0,001	
A vs. C	A>C; p_{Holm}=0,005, r=0,72, p_{NB}=0,125	
B vs. C	C>B; p_{Holm}=0,002, r=0,87, p_{NB}=0,013	

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i. \tag{2}$$

O mAP é calculado com o avaliador COCO (`pycocotools`) a IoU 0,1, compatível com o protocolo de Li *et al.* [1], e também a IoU 0,5, de forma global e por classe, sempre sobre o mesmo conjunto de teste de cada *fold*; dado o grande número de lesões por imagem, o limite de detecções por imagem foi ampliado (*maxDets* até 1000). Classes ausentes do *Ground Truth* de um *fold* são excluídas da média daquele *fold*, seguindo o padrão COCO.

H. Análise estatística

As três configurações são comparadas duas a duas, nos pares A vs. B, A vs. C e B vs. C, sob a hipótese nula bilateral de ausência de diferença de desempenho. A unidade de observação é o *fold*: como as três configurações são avaliadas exatamente sobre os mesmos 15 conjuntos de teste, cada uma produz um vetor de 15 valores, pareados elemento a elemento. Para um par (X, Y), define-se o vetor de diferenças

$$d_i = m_i^X - m_i^Y, \qquad i = 1, \ldots, 15, \tag{3}$$

em que m_i é a métrica obtida no *fold* i. O pareamento elimina a variabilidade atribuível à partição, isolando o efeito do *slicing*. O procedimento é aplicado ao mAP global (três comparações) e ao AP de cada lesão (doze comparações: 3 pares $\times$ 4 classes), sempre sobre o desempenho a IoU 0,1.

Cada vetor de diferenças é submetido a dois testes complementares. O primeiro é o de Wilcoxon pareado (*signed-rank*) [23], escolhido por não pressupor normalidade das diferenças, suposição pouco verificável com apenas 15 observações; dele se reporta o valor-p e o tamanho de efeito $r = |Z|/\sqrt{N}$, com $N = 15$. Esse teste, contudo, pressupõe observações independentes, condição que a validação cruzada repetida não satisfaz, pois *folds* de repetições distintas compartilham as mesmas imagens. Tal dependência subestima a variância das diferenças e tende a produzir valores-p artificialmente baixos, e não existe variante de Wilcoxon que a corrija. Aplica-se, por isso, um segundo teste: o teste-t pareado com a correção de variância de Nadeau–Bengio [24], que adiciona ao erro-padrão um termo proporcional à razão $n_{\mathrm{teste}}/n_{\mathrm{treino}} = 0{,}25$ de cada *fold*.

Os testes cumprem papéis distintos: Wilcoxon dispensa a normalidade, mas ignora a dependência entre *folds*; Nadeau–Bengio corrige essa dependência, ao custo de assumir diferenças aproximadamente normais. Considera-se robusta apenas a diferença em que ambos concordam; quando divergem, adota-se a conclusão do teste mais conservador, situação que ocorre em A vs. C (Seção V-A). Como análise complementar, dado o poder estatístico limitado de $n = 15$, a comparação é repetida tendo a imagem como unidade de observação: o AP é calculado por imagem e, como cada uma das 757 compõe o conjunto de teste exatamente uma vez por repetição, obtêm-se 757 observações pareadas, submetidas ao mesmo teste de Wilcoxon. Por fim, aplica-se a correção de Holm–Bonferroni separadamente em cada família de hipóteses, com $\alpha = 0{,}05$, denotando-se por p_{Holm} o valor-p corrigido.

V. RESULTADOS E DISCUSSÃO

A. Desempenho global

O desempenho global das três configurações, expresso pelo mAP médio entre os 15 *folds*, é apresentado na Tabela IV e na Fig. 3. O *baseline* em imagem completa (Config. A) obteve os maiores valores em ambos os limiares. A aplicação do SAHI restrita à inferência (Config. B) resultou em desempenho inferior, contrariando os ganhos reportados na literatura para arquiteturas anteriores [7], [12]. O regime completo, com ajuste fino no domínio de *tiles* (Config. C), atenuou essa degradação e aproximou-se do *baseline* a IoU 0,1, embora permanecesse abaixo dele a IoU 0,5.

Os testes pareados sustentam essa leitura. A diferença entre as Configs. A e B é significativa e com tamanho de efeito elevado ($p_{\mathrm{Holm}} = 0{,}002$; $r = 0{,}88$), resultado corroborado pela análise complementar por imagem ($p_{\mathrm{Holm}} = 0{,}003$) e pelo teste de Nadeau–Bengio ($p_{\mathrm{NB}} = 0{,}001$): a inferência fracionada, aplicada isoladamente sobre um modelo treinado em imagem completa, degrada o desempenho global. A diferença entre as Configs. B e C também é significativa ($p_{\mathrm{Holm}} = 0{,}002$; $p_{\mathrm{NB}} = 0{,}013$), o que evidencia que o ajuste fino no domínio de *tiles* é condição necessária para compensar essa perda.

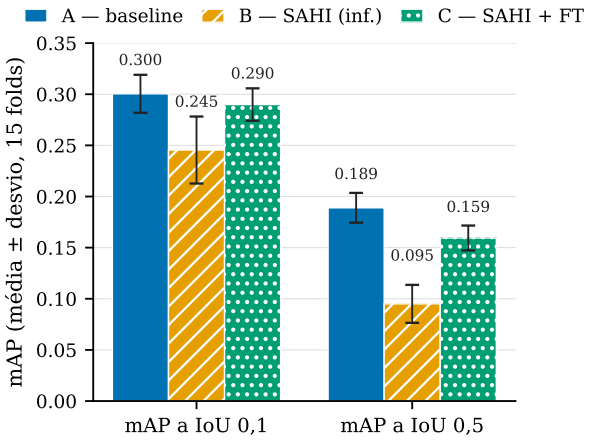


Figura 3. mAP global por configuração a IoU 0,1 e 0,5 (média ± desvio entre os 15 *folds*).

Já a comparação entre as Configs. A e C depende do teste adotado: a diferença é significativa sob Wilcoxon ($p_{\text{Holm}} = 0{,}005$), mas não sob Nadeau–Bengio ($p_{\text{NB}} = 0{,}125$), que ajusta a variância para a não-independência entre os *folds*. Pelo critério mais conservador, não há evidência suficiente para rejeitar a hipótese de ausência de diferença de mAP global entre elas. O *slicing* não produziu, portanto, nesta configuração experimental, o ganho global descrito em trabalhos com gerações anteriores de YOLO [5], [7].

B. Desempenho por classe

A estratificação por tipo de lesão revela comportamento distinto do agregado: os efeitos do *slicing* têm sinal e magnitude dependentes da classe, sendo positivos nas lesões de menor extensão espacial, conforme a Tabela V e as Figs. 4 e 5.

Os microaneurismas constituem a única classe em que uma configuração com *slicing* supera o *baseline* de forma estatisticamente significativa. A Config. C elevou o AP@0,1 de 0,177 para 0,216, incremento absoluto de 0,039 e relativo de 22% ($p_{\text{Holm}} = 0{,}009$; $r = 0{,}82$), e, a IoU 0,5, de 0,059 para 0,111. A Config. B também superou o *baseline* nessa classe (0,189 a IoU 0,1), sem atingir significância após a correção de Holm–Bonferroni. O efeito é consistente entre partições: a Config. C supera o *baseline* em 13 dos 15 *folds* (Fig. 6).

Nos exsudatos duros, o maior AP@0,1 coube à Config. B (0,404 contra 0,348 do *baseline*; $p_{\text{Holm}} = 0{,}008$); a IoU 0,5, porém, o *baseline* mantém-se superior (0,181 contra 0,108), o que sugere que o fracionamento aumenta a taxa de detecção dessa classe em detrimento da precisão de localização das caixas. Exsudatos moles e hemorragias apresentaram redução de AP sob ambas as configurações com *slicing*, comportamento analisado na Seção V-C.

C. O trade-off do slicing

A contrapartida do ganho nos microaneurismas é a perda nas lesões maiores e difusas. Exsudatos moles (SE) e hemorragias (HE), que frequentemente excedem ou cruzam as dimensões de um *tile* de 512 *px*, são fragmentados, e o reagrupamento por pós-processamento não os reconstrói integralmente. A inspeção qualitativa das predições corrobora o mecanismo, com caixas espúrias e fragmentadas concentradas nas junções dos *tiles*. O efeito é acentuado na Config. B: o AP@0,1 de SE cai de 0,354 para 0,178 ($\Delta = -0{,}176$) e o de HE cai 0,112, o que explica isoladamente a queda do mAP global dessa configuração. O ajuste fino da Config. C atenua o dano (SE recupera para 0,329; HE para 0,255), pois o modelo aprende a operar no domínio de *tiles*, mas ainda não iguala o *baseline* nessas classes (Fig. 5).

Esse *trade-off* explica por que o SAHI não melhora o mAP *global*: sendo uma média entre classes, o mAP dilui o ganho nos microaneurismas na perda em SE e HE. Na prática, o valor do *slicing* depende da métrica priorizada: quando o AP de microaneurismas é o critério de interesse, a Config. C supera o *baseline*; para uma detecção equilibrada das quatro classes, o *baseline* em alta resolução é mais robusto, um achado metodológico, não uma recomendação clínica, que exigiria validação adicional.

Em análise exploratória *post-hoc* (fora do protocolo predefinido e sem teste estatístico), a reavaliação dos pesos da Config. C com fusão IOU (em vez da IOS, padrão do SAHI) no *GREEDYNMM* eleva o mAP@0,5 de 0,159 para 0,183, próximo do *baseline* (0,189), preservando o AP@0,5 de MA ($\approx$ 0,11), indício de que a fusão imperfeita de caixas nas junções, mais que erros de detecção, responde por parte da perda a IoU alto.

O *slicing* também tem custo de inferência: cronometrado o teste de um *fold* (145 imagens; em média 34,9 *tiles* por imagem a 512/0,2), a Config. B custa 11,4× e a C 5,0× o *baseline*; a diferença entre ambas decorre da resolução de entrada herdada do treino (1024 vs. 512 *px*). Como tempos absolutos dependem do *hardware*, importam as razões: um sobrecusto de cerca de uma ordem de grandeza, a ponderar na viabilidade de rastreamento em larga escala.

D. Comparação com a literatura

Os resultados são posicionados frente à literatura do DDR na Tabela VI, com todos os valores a IoU 0,5. A comparação direta exige cautela: os trabalhos anteriores reportam um único

Tabela V
AP POR CLASSE (MÉDIA ENTRE OS 15 *folds*), A IOU 0,1 E 0,5. EM **NEGRITO**, O MELHOR VALOR POR CLASSE/IOU. †: DIFERENÇA SIGNIFICATIVA VS. A ($p_{\text{Holm}} < 0{,}05$, WILCOXON NA FAMÍLIA POR CLASSE). OS TESTES DE SIGNIFICÂNCIA FORAM CONDUZIDOS SOBRE O AP A IOU 0,1.

IoU	Config.	MA	EX	SE	HE
0,1	A	0,177	0,348	**0,354**	**0,323**
	B	0,189	**0,404**†	0,178†	0,210†
	C	**0,216**†	0,360	0,329†	0,255†
0,5	A	0,059	**0,181**	**0,308**	**0,208**
	B	0,032	0,108	0,150	0,091
	C	**0,111**	0,132	0,258	0,137

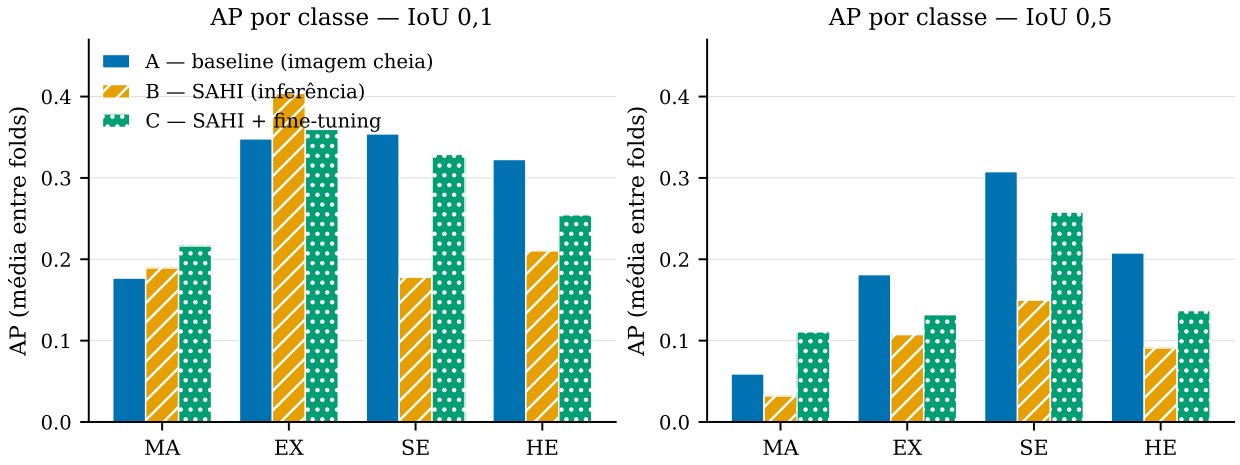


Figura 4. AP por classe de lesão (média entre os 15 *folds*), a IoU 0,1 (esquerda) e 0,5 (direita). O *slicing* beneficia os microaneurismas (MA) e, na inferência, os exsudatos duros (EX), mas prejudica os exsudatos moles (SE) e as hemorragias (HE), que são lesões maiores e mais difusas.

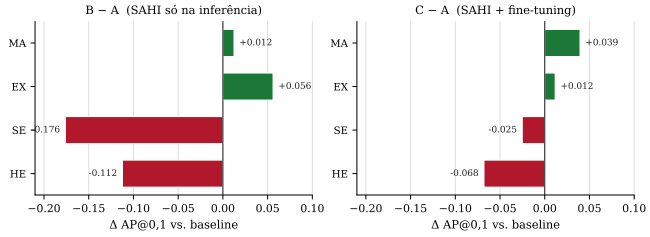


Figura 5. Ganho de AP@0,1 de cada configuração de *slicing* em relação ao *baseline* (A), por classe. Verde: ganho; vermelho: perda.

Tabela VI
COMPARAÇÃO COM A LITERATURA NO DDR (MAP A IOU 0,5). "VAL."/"TESTE" INDICAM O SUBCONJUNTO; "CV" INDICA VALIDAÇÃO CRUZADA REPETIDA (MÉDIA DE 15 *folds*). PROTOCOLOS, RESOLUÇÕES E ARTIFÍCIOS DE TREINO DIFEREM ENTRE ESTUDOS E NÃO SÃO DIRETAMENTE COMPARÁVEIS.

Trabalho / método	**Protocolo**	**mAP@0,5**
Li *et al.* [1] (SSD/YOLO)	teste	$< 0{,}01$
Alyoubi *et al.* [18] (CNN512+YOLOv3)	n. i.	0,216
Santos *et al.* [5] (YOLOv5+*Tiling*)	val.	0,263
Santos *et al.* [5] (YOLOv5+*Tiling*)	teste	0,154
Pereira *et al.* [7] (YOLOR-CSP+SAHI)	val.	0,381
Pereira *et al.* [7] (YOLOR-CSP+SAHI)	teste	0,222
Este trabalho, A (*baseline*)	CV	0,189
Este trabalho, C (SAHI+FT)	CV	0,159
Este trabalho, A (*split* oficial)	teste	0,113

split (tipicamente a *validação*), enquanto aqui se reporta a média de uma validação cruzada repetida, protocolo mais conservador. Para uma comparação justa, avaliou-se também a Config. A no *split* oficial do DDR (383/149/225), o mesmo usado por [5], [7]: nele o *baseline* obtém mAP@0,5 de 0,113, valor coerente com a faixa de teste dos trabalhos correlatos (0,154 e 0,222).

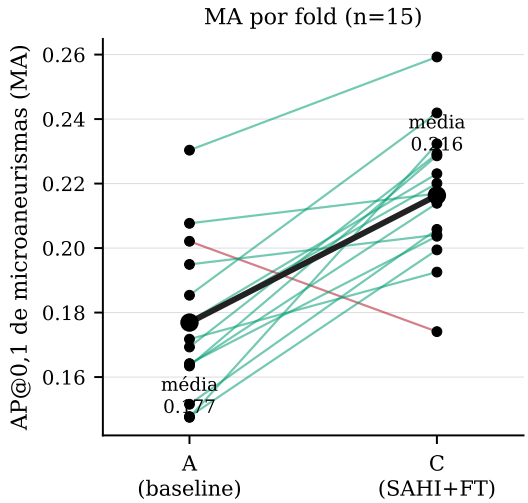


Figura 6. AP@0,1 de microaneurismas (MA), por *fold* ($n = 15$): *baseline* (A) vs. Config. C. As linhas verdes indicam *folds* em que a Config. C supera o *baseline* (13 de 15); a linha preta liga as médias (0,177 → 0,216).

Não se pretende, aqui, competir por desempenho absoluto: os métodos da literatura empregam artifícios adicionais (recorte do fundo, pré-treino em COCO, milhares de épocas, ajuste de âncoras, aumento de dados agressivo) que este estudo deliberadamente evita para *isolar* o efeito do *slicing*. O ponto é metodológico: a diferença entre validação e teste observada em [7] (0,381 vs. 0,222) reforça que números de *split* único devem ser lidos com cautela, e que a validação cruzada oferece estimativa mais fiel do desempenho esperado.

E. Limitações

Este estudo apresenta limitações. Primeiro, os hiperparâmetros foram fixados (sem busca), privilegiando a comparabilidade sobre o desempenho absoluto; uma busca dedicada poderia alterar as magnitudes, mas dificilmente inverteria o

trade-off por classe, de natureza geométrica. Segundo, avaliou-se um único detector (YOLOv8m) e uma única configuração de fracionamento (512/0,2); *tiles* menores poderiam favorecer os microaneurismas ao custo de agravar a fragmentação, e a extensão do *trade-off* a outras arquiteturas (YOLO recentes, *Transformers*) permanece em aberto. Terceiro, a avaliação restringe-se ao DDR, com generalização a outras bases, equipamentos e populações (por exemplo, IDRiD [25]) a verificar; e, sem identificadores de paciente, a partição é por imagem (Seção IV-E). Por fim, o escopo é a acurácia de detecção: desfechos em nível de paciente (graduação de RD, sensibilidade de triagem) não foram avaliados.

VI. Considerações Finais

Este trabalho avaliou, de forma controlada e estratificada por lesão, se a inferência assistida por SAHI melhora a detecção de lesões de RD no DDR quando aplicada a um detector YOLO moderno (YOLOv8m). A resposta é matizada. *Globalmente*, o SAHI não melhora o mAP: a inferência fracionada sobre um modelo treinado em imagem cheia (Config. B) degrada significativamente o desempenho, e mesmo o regime completo com ajuste fino (Config. C) não apresenta diferença detectável em relação ao *baseline*. *Por classe*, entretanto, o *slicing* cumpre sua promessa exatamente na lesão mais precoce e crítica: a Config. C melhora de forma significativa e consistente a detecção de microaneurismas, ao custo de piora nos exsudatos moles e nas hemorragias, que o fracionamento tende a fragmentar, *trade-off* que o mAP, por ser uma média entre classes, oculta.

A principal mensagem é que, em detectores YOLO modernos operando em alta resolução, o benefício do SAHI é *condicional*: localizado nas menores lesões e dependente de ajuste fino em *tiles*, e não o ganho global sugerido por trabalhos com arquiteturas anteriores. O resultado deve ser lido como evidência metodológica sobre o detector no DDR; uma recomendação clínica exigiria validação externa e desfechos em nível de paciente, fora do escopo deste estudo.

Como trabalhos futuros, pretende-se: (i) varrer o tamanho de *tile* e a sobreposição, buscando o equilíbrio do *trade-off*; (ii) explorar estratégias de fusão de caixas nas bordas dos *tiles* para reduzir a perda em lesões grandes; (iii) validar *cross-dataset*, treinando no DDR e avaliando em IDRiD [25] e em bases brasileiras; (iv) contrastar o *slicing* com arquiteturas mais recentes de YOLO e com *Transformers* de detecção; (v) aprofundar a análise dos modos de falha com técnicas de explicabilidade; e (vi) integrar a detecção a um *pipeline* de triagem com desfechos em nível de paciente.

Referências

[1] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019.

[2] A. A. Alghadyan, "Diabetic retinopathy – An update," *Saudi Journal of Ophthalmology*, vol. 25, no. 2, pp. 99–111, 2011.

[3] Ministério da Saúde, Governo Federal do Brasil, "Diabetes," 2024. [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/diabetes. Acesso em: 3 set. 2024.

[4] Ministério da Saúde, Comissão Nacional de Incorporação de Tecnologias no SUS (CONITEC), *Protocolo Clínico e Diretrizes Terapêuticas – Retinopatia Diabética (Versão Resumida)*, 2021. [Online]. Available: https://www.gov.br/conitec/pt-br/midias/protocolos/resumidos/PCDTResumidoRetinopatiaDiabtica.pdf

[5] C. Santos, M. Aguiar, D. Welfer, and B. Belloni, "A new approach for detecting fundus lesions using image processing and deep neural network architecture based on YOLO model," *Sensors*, vol. 22, no. 17, p. 6441, 2022.

[6] D. F. D. Assis, P. C. Cortez, and P. Motta, "Métodos de classificação de glaucoma em imagens do disco óptico," in *Anais do 14 Congresso Brasileiro de Inteligência Computacional*, Curitiba, 2019, pp. 1–8.

[7] A. Pereira, C. Santos, M. Aguiar, D. Welfer, M. Dias, and M. Ribeiro, "Detection of retinal microlesions through YOLOR-CSP architecture and image slicing with the SAHI algorithm," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. on Computer Vision (ECCV)*, 2018, pp. 3–19.

[9] A. Jithesh, C. Karumuri, V. K. R. Kotha, M. Doddapuneni, and T. Jeong, "Diabetic retinopathy lesion segmentation through attention mechanisms," in *Proc. IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2025.

[10] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9725–9734.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. European Conf. on Computer Vision (ECCV)*, 2020, pp. 213–229.

[12] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 966–970.

[13] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. European Conf. on Computer Vision (ECCV)*, 2016, pp. 21–37.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[17] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[18] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, p. 3704, 2021.

[19] C. Santos, M. S. De Aguiar, D. Welfer, and B. Belloni, "Deep neural network model based on one-stage detector for identifying fundus lesions," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.

[21] X. Bouthillier *et al.*, "Accounting for variance in machine learning benchmarks," in *Proc. Machine Learning and Systems (MLSys)*, vol. 3, 2021, pp. 747–769.

[22] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2011, pp. 145–158.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[24] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[25] P. Porwal *et al.*, "IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge," *Medical Image Analysis*, vol. 59, 101561, 2020.

Development and Characterization of Green Strontium Oxide Nanoparticles for Potential Application in Bone Regeneration

Marcos Vinicius Pasqualoto Prior*, Lourdes Maria Muraro Favarin*, Luiz Batista Cardoso*, Vinícius Rodrigues Oviedo*, Daniel Moro Druzian*, Yolice Patricia Moreno Ruiz†, Giovani Pavoski‡, Jorge Alberto Soares Tenório‡, Denise Crocce Romano Espinosa‡, Amanda Kolinski Machado§, Alencar Kolinski Machado§, William Leonardo da Silva*

* *Applied Nanomaterials Research Group (GPNAp)*
Franciscan University (UFN)
Santa Maria, RS, Brazil
marcos.prior@ufn.edu.br; lourdes.favarin@ufn.edu.br; luiz.bcardoso@ufn.edu.br;
vinicius.oviedo@ufn.edu.br; daniel.moro.druzian@gmail.com; w.silva@ufn.edu.br

† *Strategic Technologies Center of Northeast (CETENE)*
Federal University of Pernambuco (UFPE)
Recife, PE, Brazil
ypmoreno@gmail.com

‡ *Polytechnical School of Chemical Engineering (EPEQ)*
University of São Paulo (USP)
São Paulo, SP, Brazil
giovanipavoski@gmail.com; jtenorio@usp.br; espinosa@usp.br

§ *Laboratory of Cell Culture and Bioactive Effects (LabCultBio)*
Franciscan University (UFN)
Santa Maria, RS, Brazil
amanda.kolinski@ufn.edu.br; alencar.machado@ufn.edu.br

***Abstract*—Increased life expectancy has raised the prevalence of bone disorders, increasing demand for biomaterials that promote bone regeneration. In this study, strontium oxide nanoparticles (SrO-NPs) were synthesized via green synthesis using pecan (*Carya illinoinensis*) shell extract as a reducing and stabilizing agent. The nanoparticles were characterized by X-ray diffraction (XRD), Fourier-transform infrared spectroscopy (FTIR), and N_2 porosimetry. XRD confirmed a cubic SrO phase with an average crystallite size of 45.26 nm, coexisting with an orthorhombic $SrCO_3$ phase. EDX indicated residual chlorine, demonstrating that the material was an SrO/$SrCO_3$ composite with precursor residues rather than pure SrO. FTIR identified characteristic O–H, C=O, C–O, and Sr–O vibrations, while N_2 adsorption–desorption indicated a mesoporous structure. *In vitro* assays using OFCOL II osteoblasts showed no cytotoxicity, elevated ROS, or dsDNA release. These findings support an *in vitro* safety profile and justify further investigation of osteogenic performance and *in vivo* response.**

***Index Terms*—biomaterials, biomedical engineering, green synthesis, nanotechnology**

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

I. Introduction

Population aging represents one of the most significant demographic transformations of the 21st century. Advances in health care, sanitation, and medical technologies have contributed to a steady increase in global life expectancy, with Brazil reaching an average life expectancy of 76.8 years in 2026 [1]. However, increased longevity has also been accompanied by a higher prevalence of chronic non-communicable diseases and age-related musculoskeletal disorders, particularly bone fragility and femoral fractures [2], [3]. Consequently, the demand for biomaterials for bone repair has grown substantially, encompassing various classes of materials, including metals, ceramics, polymers, and composites [4].

Femoral fractures constitute a major public health problem owing to their high morbidity, mortality, and the substantial economic burden they impose on health care systems. Between 2020 and 2025, approximately 630,687 hospital admissions for femoral fractures were recorded in Brazil, resulting in an estimated health care expenditure of R$1.42 billion [5]. These data highlight the urgent need for effective therapeutic

strategies to promote bone repair. In this context, tissue engineering has emerged as a promising approach for the development of three-dimensional scaffolds that support bone regeneration [6]. However, currently available biomaterials still present important limitations, including susceptibility to bacterial colonization, inadequate mechanical properties, and limited capacity to mimic the native bone extracellular matrix [7]. Thus, nanotechnology offers a promising strategy for overcoming these challenges through the development of nanostructured systems with enhanced bioactivity and surface properties [8].

Among the nanomaterials currently under investigation for bone tissue engineering, strontium oxide nanoparticles (SrO-NPs) have attracted considerable attention owing to the ability of strontium to stimulate osteoblastic activity, inhibit osteoclast differentiation, and promote bone formation [9], [10]. In addition, green nanotechnology has become established as a sustainable alternative to conventional synthesis methods, replacing toxic chemical reagents with biologically derived reducing and stabilizing agents obtained from plant extracts [11], [12].

Pecan (*Carya illinoinensis*) shells constitute an abundant agro-industrial byproduct in southern Brazil and are rich in phenolic compounds with high antioxidant activity [13]. Previous studies have demonstrated the potential of pecan shell extract for the green synthesis of silver nanoparticles with antimicrobial activity [14]. In this context, plant extracts have stood out as sustainable reducing and stabilizing agents owing to their rich phytochemical composition, including polyphenols and flavonoids, which enable the environmentally friendly synthesis of metallic and metal oxide nanoparticles [15]–[17]. Thus, the use of this biomass as a reducing and stabilizing agent in the synthesis of SrO-NPs adds value to a regional agro-industrial byproduct while promoting environmental sustainability.

This approach is aligned with the principles of green chemistry and the United Nations Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-Being), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 12 (Responsible Consumption and Production) [18]. Given the above, the present study aimed to synthesize and characterize SrO-NPs obtained through green synthesis using pecan shell extract and to evaluate their *in vitro* safety profile in OFCOL II osteoblastic cells, providing preliminary scientific evidence for their potential application as a bioactive nanomaterial in bone regeneration.

II. Materials and Methods

A. Green synthesis of SrO-NPs

Pecan shells were obtained from a local agro-industry in southern Brazil. The extract was prepared by the infusion method [19], in which 10 g of pecan shells were cleaned and dried at 40 $\pm$ 2 °C for 72 hours, followed by mixing with 500 mL of distilled water under magnetic stirring (300 rpm, 10 min, 85 $\pm$ 2 °C). The extract was then stored in an amber flask at 4 $\pm$ 2 °C.

SrO-NPs were synthesized by the hydrothermal method [20], [21]. Briefly, the pecan shell extract was mixed with a strontium chloride hexahydrate solution ($SrCl_2{\cdot}6H_2O$, Synth, $> 99\%$) (0.1 mol L^{-1}) under magnetic stirring (200 rpm, 25 $\pm$ 2 °C), promoting the reduction and nucleation steps. The mixture was then transferred to a stainless-steel autoclave with an internal polytetrafluoroethylene liner and subjected to hydrothermal treatment (160 °C, 480 min, 5 °C min^{-1}), followed by calcination (850 °C, 120 min, 30 °C min^{-1}) to promote stabilization of the SrO-NPs.to promote the stabilization of the SrO-NPs.

B. Characterization techniques

Crystalline phases were identified by X-ray diffraction (XRD) using a Bruker D2 Advance diffractometer with Cu Kα radiation (λ = 0.15418 nm), over a 2θ range of 20° to 80°, operated at 30 kV and 10 mA. The average crystallite size (d_c) was calculated using the Scherrer equation, as shown in (1).

$$d_c = \frac{K * \lambda_{\mathrm{CuK}\alpha}}{\beta * \cos(2\theta)} \tag{1}$$

Functional groups were analyzed by Fourier-transform infrared spectroscopy (FTIR) using an IRPrestige-21 spectrometer (Shimadzu) in transmittance mode, over the range of 4000 to 450 cm^{-1}.

Nitrogen physisorption measurements were performed on an ASAP 2020 system (Micromeritics) at 77 K. Adsorption–desorption isotherms were used to determine the specific surface area (SSA), pore volume (V_p), pore diameter (D_p), and pore size distribution [22].

A field-emission scanning electron microscope (FEG-SEM) TESCAN MIRA3 was used to evaluate morphological properties [23]. In addition, an Oxford EDX system coupled to the FEG-SEM TESCAN MIRA3 was employed to determine elemental composition, with 400$\times$ magnification and backscattered electrons. Samples were prepared by depositing the solids on double-sided adhesive tape fixed to aluminum stubs [24]. Data were collected in different modes: bulk, at specific points, and along a trend line.

C. In vitro safety profile

The *in vitro* safety profile was evaluated using OFCOL II osteoblasts (BCRJ-0192). Cells were cultured in RPMI medium supplemented with 10% fetal bovine serum, at 37 °C in a 5% CO_2 atmosphere. SrO-NPs were tested at concentrations of 12.5, 25, 50, and 100% of the stock solution (2000 μg/mL).

Cell viability and proliferation were assessed by the MTT assay (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) after 24, 48, and 72 hours of exposure [25]. Absorbance was recorded at 570 nm using a Synergy HTX microplate reader (BioTek). Nitric oxide (NO) production was determined by the Griess reaction, measuring nitrite (NO_2^-) and nitrate (NO_3^-) levels in the culture supernatant at 540 nm [26]. Generation of reactive oxygen species (ROS) in cells was measured using the DCFH-DA probe (2′,7′-dichlorodihydrofluorescein diacetate), with fluorescence

recorded at 480/520 nm (excitation/emission) [25]. Release of extracellular double-stranded DNA (dsDNA) was evaluated using the Quant-iT PicoGreen reagent, with fluorescence recorded at 480/520 nm (excitation/emission), as a punctual indicator of membrane integrity [25]. It should be noted that, in isolation, this assay does not allow inferences regarding genotoxicity, apoptosis, or mutagenesis, which would require additional specific assays (e.g., comet assay or micronucleus test). Untreated cells served as the negative control (NC), while hydrogen peroxide (H_2O_2, 200 μmol L^{-1}) was used as the positive control in all biological assays.

D. Statistical analysis

All experiments were performed in triplicate. Results are presented as a percentage relative to the untreated control (mean $\pm$ standard deviation). Statistical analyses were performed by one-way analysis of variance (ANOVA), followed by Tukey's post hoc test. Statistical significance was defined as $p < 0.05$, $p < 0.01$, $p < 0.001$, and $p < 0.0001$.

III. Results and Discussion

A. Characterization of the SrO-NPs

Figure 1 presents the X-ray diffractogram of the strontium oxide nanoparticles obtained by green synthesis (SrO-NPs), in which diffraction peaks characteristic of the cubic SrO phase were observed at 2θ = 27.5° (110), 29.1° (112), 35.6° (202), 42.9° (114), 49.0° (222), 49.6° (213), 52.6° (310), 56.7° (312), 60.1° (114), 61.4° (114), 66.3° (312), and 78.5° (013), in agreement with ICDD card no. 96-900-7717 [27]. In addition, diffraction peaks corresponding to the $SrCO_3$ phase were identified at 2θ = 22.1° (110), 25.2° (111), 31.5° (002), 33.7° (121), 39.1° (201), 40.8° (220), and 45.7° (221), in agreement with JCPDS card no. 00-005-0418 [28]. The coexistence of the two crystalline phases indicated that the synthesized material is not pure, being composed of SrO and $SrCO_3$. This heterogeneous composition should be taken into account when interpreting future assays.

The crystallite size, calculated using the Scherrer equation ($K = 0.94$) from the (202) diffraction peak, was 45.26 nm, a value consistent with data reported in the literature for SrO nanoparticles obtained by green synthesis [20], [29].

The FTIR spectrum (Figure 2) revealed the characteristic vibrational bands of the SrO-NPs. A broad band between 3600 and 3200 cm^{-1} was attributed to O–H stretching vibrations of water molecules adsorbed on the nanoparticle surface [22]. The band at 1650 cm^{-1} corresponds to C=O carbonyl stretching, likely originating from residual organic compounds present in the pecan shell extract [20]. At 713 cm^{-1}, the band was attributed to asymmetric stretching and out-of-plane bending vibrations of C–O bonds, related to the presence of carbonate species ($SrCO_3$) [30]. At 482 cm^{-1}, the band corresponded to a metal–oxygen (Sr–O) stretching vibration, possibly arising from the formation of SrO [22], [31]. These results corroborated the XRD findings and confirmed the coexistence of the SrO and $SrCO_3$ phases.

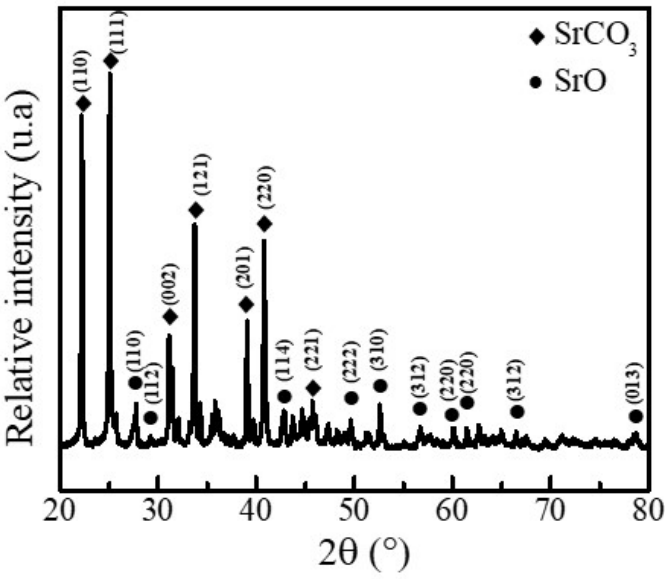


Fig. 1. XRD diffractogram of the SrO-NPs from *Carya illinoinensis* extract.

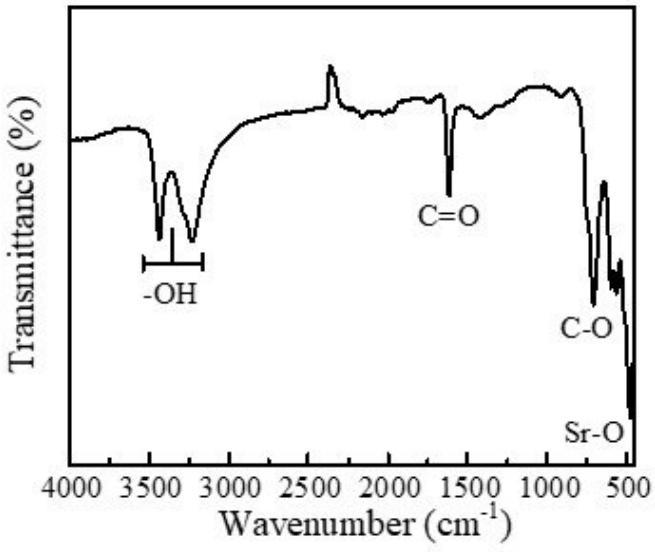


Fig. 2. FTIR spectrum of the SrO-NPs from *Carya illinoinensis* extract.

N_2 porosimetry of the material (Figure 3) exhibited a type IV(a) isotherm profile with an H3 hysteresis loop, according to the IUPAC classification [32]. This type of isotherm is characteristic of mesoporous materials, while H3-type hysteresis is associated with slit-shaped pores formed by aggregates. The specific surface area was approximately 45.79 $\pm$ 0.14 m^2 g^{-1}, with a pore diameter of 42.31 $\pm$ 0.14 nm and a pore volume of 0.044 $\pm$ 0.01 cm^3 g^{-1}, confirming the mesoporous nature of the material. Literature studies on SrO-based nanomaterials have reported similar textural properties, with low surface areas attributed to high-temperature calcination (600 °C to 800 °C), which promotes particle sintering and grain growth [27], [22].

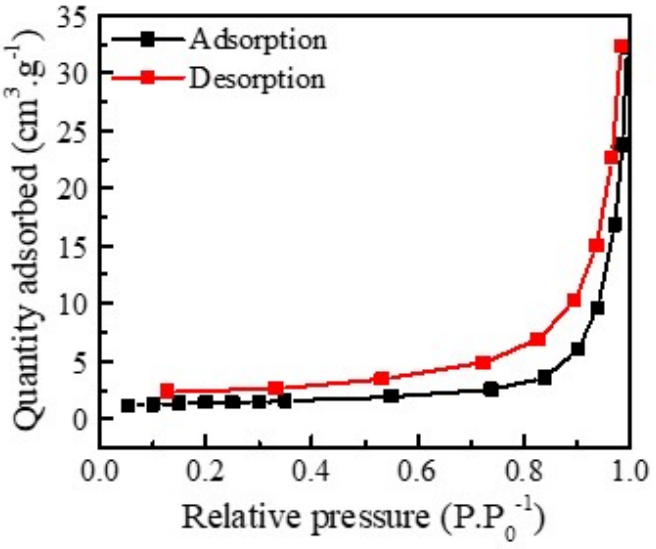


Fig. 3. N_2 adsorption–desorption isotherm of the SrO-NPs from *Carya illinoinensis* extract.

Morphological characterization of the green SrO-NPs was performed by high-resolution FEG-SEM combined with elemental analysis by EDX. The high-resolution micrograph of the SrO-NPs (Figure 4) revealed a heterogeneous and porous morphology, with small agglomerations attributable to the thermal treatment [20]. Elemental mapping by EDX (Figure 5) indicated a homogeneous dispersion of elements on the material surface, with a predominant presence of strontium (Sr, 33.6 wt%), in addition to chlorine (Cl, 22.8 wt%), attributed to the metal precursor used, suggesting that it was not completely consumed or removed during the synthesis or calcination process [22]. The residual chlorine content, together with the presence of the $SrCO_3$ phase demonstrated by XRD and FTIR characterization, indicates that the final material is composed of SrO and $SrCO_3$, with residues from the $SrCl_2{\cdot}6H_2O$ precursor.

B. In vitro safety profile

Figure 6 presents the results of the *in vitro* safety profile of the SrO-NPs. Cell viability and proliferation assays performed using the MTT method did not reveal any cytotoxic effects at the concentrations tested when compared with the control group. A statistically significant increase in cell viability was observed only at the 12.5% concentration after 48 hours of incubation. These results indicate that, under the tested conditions, SrO-NPs demonstrate a favorable biocompatibility profile, with no evidence of significant cellular toxicity. In bone regeneration applications, disordered cell growth can lead to fibrosis or, in extreme cases, osteosarcomas (bone tumors). In this sense, the observed profile suggested that osteoblasts maintained their basal physiological behavior, with no evidence of an exacerbated proliferative response [10].

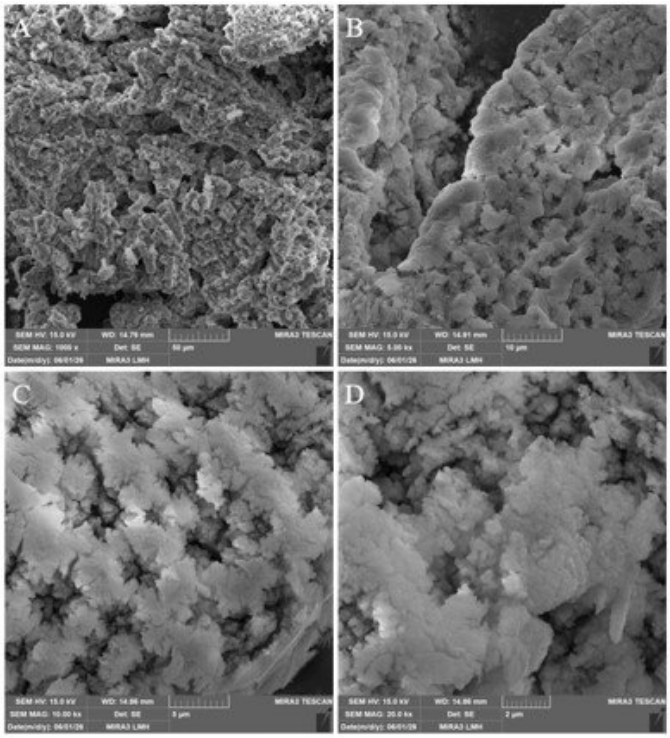


Fig. 4. High-resolution micrographs of green SrO-NPs at different magnifications: (a) 1 kx, (b) 5 kx, (c) 10 kx, and (d) 20 kx.

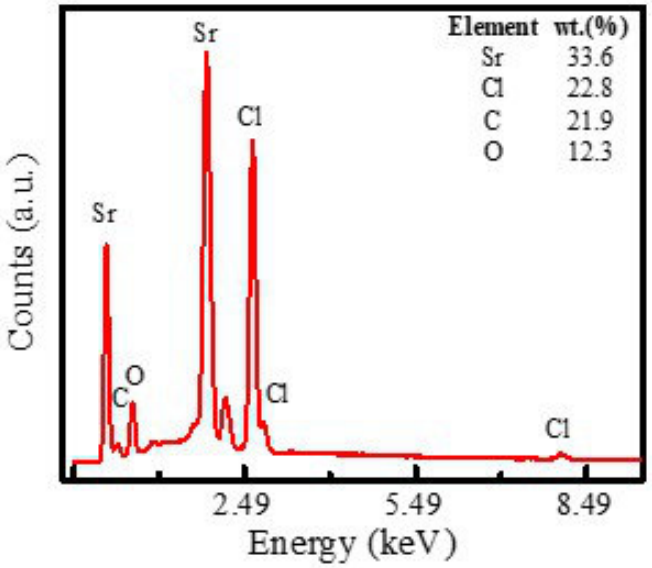


Fig. 5. Element distribution histogram of green SrO-NPs obtained from FEG-SEM micrographs.

Evaluation of nitric oxide (NO) showed that levels remained equivalent to the negative control (100%) at all concentrations and time points tested. This baseline performance indicated a possible absence of an acute pro-inflammatory response induced by the material [10], with additional tests being necessary to demonstrate anti-inflammatory activity. Regarding oxidative stress, the material did not cause an increase in

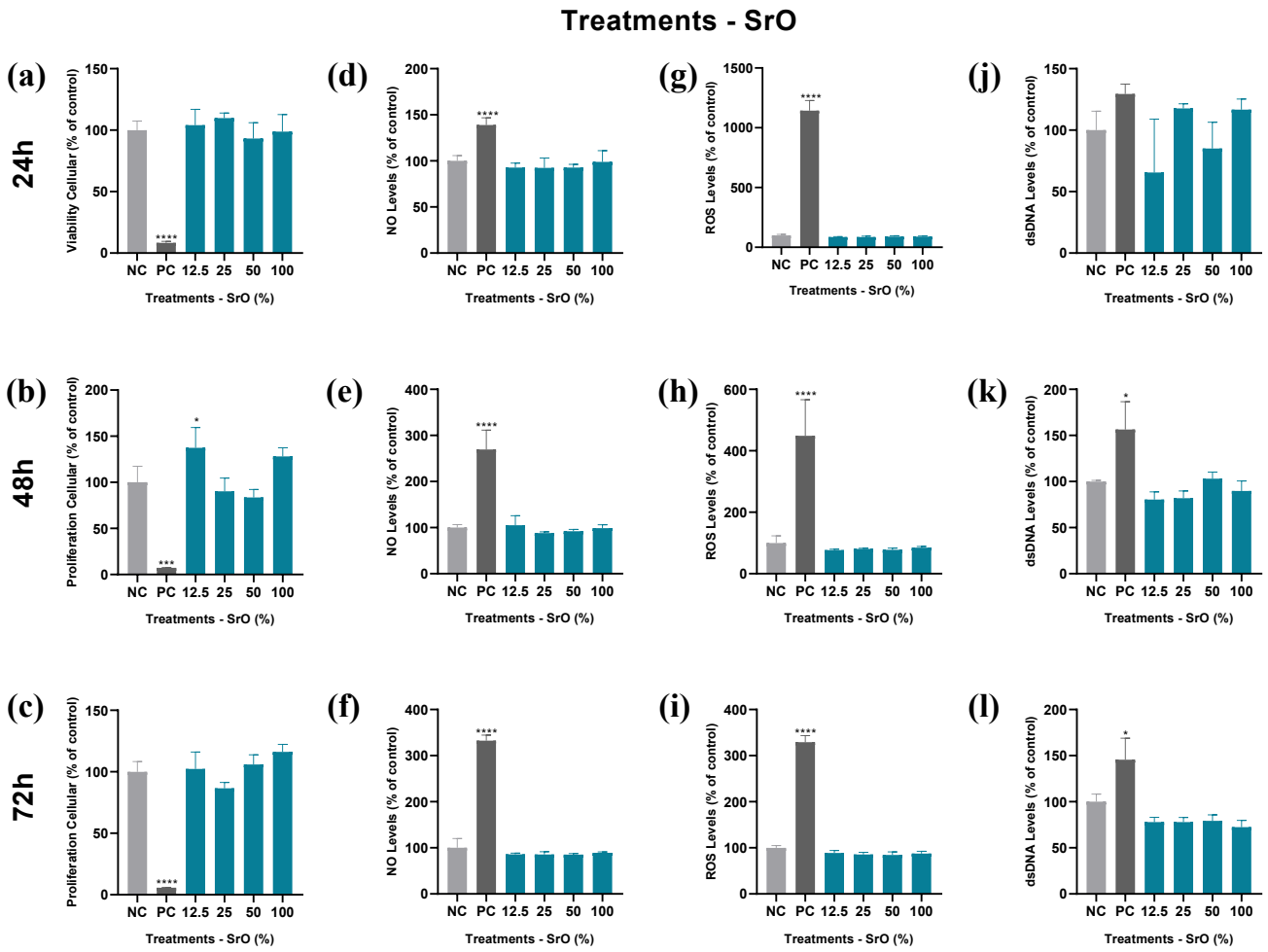


Fig. 6. *In vitro* safety profile of green SrO-NPs evaluated by MTT assay (a–c), nitric oxide (NO) production (d–f), reactive oxygen species (ROS) generation (g–i), and double-stranded DNA (dsDNA) quantification (j–l) after 24, 48, and 72 hours of exposure.

ROS levels relative to the control. Furthermore, the dsDNA assay showed that none of the concentrations used in the experiment promoted significant dsDNA release. These findings demonstrated preservation of membrane integrity under the tested conditions; however, this assay alone does not allow the conclusion that the material does not cause genotoxicity, apoptosis, or mutagenesis, requiring additional specific assays not performed in this study.

Taken together, the *in vitro* results demonstrate that the material obtained by green synthesis presented a favorable safety profile in terms of cell viability, without inducing inflammatory responses or oxidative stress detectable by the assays employed, even in the presence of other compounds such as $SrCO_3$ and residual chlorine from the synthesis. Similar *in vitro* safety conditions for $SrCO_3$ reported in the literature have shown neither toxicity nor stimulation in osteoblastic cells [33]. The residual chloride did not exert a cytotoxic effect or an adverse modulatory effect on the osteoblastic cells tested, likely still being bound to the metal precursor, reflecting its high percentage in the EDX analysis; however, the effect of chloride reported in the literature is characterized by the mimicry of calcium in bone stimulation [34]. This *in vitro* safety profile falls within the parameters of the ISO 10993-5 standard for cytotoxicity evaluation of biomedical materials [25], making the material a possible candidate for application in tissue engineering.

IV. Conclusion

SrO-NPs were successfully synthesized from pecan shell extract by means of an environmentally friendly approach, resulting in a crystalline, mesoporous material composed of SrO and $SrCO_3$ with residues of the chlorinated precursor, presenting textural and structural properties consistent with those reported in the literature for green-synthesized SrO nanostructures. The *in vitro* safety evaluation in OFCOL II osteoblasts indicated the absence of cytotoxicity, oxidative stress elevation, and significant dsDNA release, in agreement with ISO 10993-5, outlining a favorable scenario for the

biocompatibility of the synthesized material, although this remains to be confirmed through additional specific assays.

It should be noted that osteogenic performance, extracellular matrix mineralization, and the *in vivo* biological response were not evaluated in this study, so that any application in bone regeneration remains a perspective to be investigated and not a conclusion established by the present results; in this sense, the material is regarded as a potential nanoreinforcement for polymeric nanocomposites intended for future application in bone regeneration, rather than a biomaterial in itself. In addition, the valorization of pecan shells as an agro-industrial by-product reinforces the sustainability of this approach, and future investigations focused on osteogenic differentiation, *in vivo* biological response, and long-term biosafety will be essential for the consolidation of these nanoparticles as a candidate for bone tissue engineering applications.

ACKNOWLEDGMENT

The authors thank the Graduate Program in Nanosciences (PPGNano) and the Applied Nanomaterials Research Group (GPNAp) of Universidade Franciscana (UFN) for support in carrying out this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001.

CONFLICT OF INTEREST

The authors declare that they have no known financial conflicts of interest or personal relationships that could have influenced the work reported in this article.

REFERENCES

[1] A. Garmany, S. Yamada, and A. Terzic, "Longevity leap: mind the healthspan gap," *NPJ Regenerative Medicine*, vol. 6, no. 1, pp. 1–7, 2021.

[2] V. M. de Azeredo Passos, A. P. S. Champs, R. Teixeira, M. F. F. Lima-Costa, R. Kirkwood, R. Veras, B. R. Nascimento, A. M. Nogales, M. I. Schmidt, B. B. Duncan, E. Cousin, M. Naghavi, and F. M. Souza, "The burden of disease among brazilian older adults and the challenge for health policies: results of the global burden of disease study 2017," *Population Health Metrics*, vol. 18, no. 1, pp. 1–15, 2020.

[3] K. Irandoust, R. Daroudi, M. Tajvar, and M. Yaseri, "Global and regional impact of health determinants on life expectancy and health-adjusted life expectancy, 2000–2018: an econometric analysis based on the global burden of disease study 2019," *Frontiers in Public Health*, vol. 13, no. 1, pp. 1–15, 2025.

[4] N. M. S. Mahdi, A. K. Hassan, F. J. Al-Hasani, and W. A. M. Al-Jawher, "Classification of biomaterials and their applications," *Journal Port Science Research*, vol. 7, no. 3, p. 281–299, 2024.

[5] L. S. d. Silva and N. Reis, "Panorama epidemiológico de pacientes com fratura de fêmur no Brasil: um estudo retrospectivo," *Brazilian Journal of Implantology and Health Sciences*, vol. 8, no. 1, p. 173–185, 2026.

[6] C. Xu, Z. Liu, X. Chen, Y. Gao, W. Wang, X. Zhuang, H. Zhang, and X. Dong, "Bone tissue engineering scaffold materials: fundamentals, advances, and challenges," *Chinese Chemical Letters*, vol. 35, no. 2, pp. 1–11, 2024.

[7] H. Hajiali, L. Ouyang, V. Llopis-Hernandez, O. Dobre, and F. R. A. J. Rose, "Review of emerging nanotechnology in bone regeneration: progress, challenges, and perspectives," *Nanoscale*, vol. 13, no. 23, p. 10266–10280, 2021.

[8] R. Karmakar, S. Dey, A. Alam, M. Khandelwal, F. Pati, and A. K. Rengan, "Attributes of nanomaterials and nanotopographies for improved bone tissue engineering and regeneration," *ACS Applied Bio Materials*, vol. 6, no. 10, pp. 4020–4041, 2023.

[9] I. Lodoso-Torrecilla, D. Moreno, G. Ciucci, M. Mateu-Sanz, J.-Y. Yoon, E. Jimenez-Pique, J. Franch, M.-C. Manzanares, J. Konka, M. Espanol, and M.-P. Ginebra, "Strontium and gallium doping enhances *in vivo* bone regeneration in biomimetic hydroxyapatite 3D-printed scaffolds," *Materials Today Bio*, vol. 38, no. 1, pp. 1–16, 2026.

[10] U. U. Akobundu, I. H. Ifijen, P. Duru, J. C. Igboanugo, I. Ekanem, M. Fagbolade, A. S. Ajayi, M. George, B. Atoe, and J. T. Matthews, "Exploring the role of strontium-based nanoparticles in modulating bone regeneration and antimicrobial resistance: a public health perspective," *RSC Advances*, vol. 15, no. 14, p. 10902–10957, 2025.

[11] Y. Bao, J. He, K. Song, J. Guo, X. Zhou, and S. Liu, "Plant-extract-mediated synthesis of metal nanoparticles," *Journal of Chemistry*, vol. 2021, no. 1, pp. 1–14, 2021.

[12] A. Saraf, S. B. Geelani, P. Das, and A. Jayaraman, "The nanoalchemy of nature: a critical review of biogenic and sustainable pathways in green nanotechnology," *Results in Chemistry*, vol. 24, no. 1, pp. 1–23, 2026.

[13] N. T. Dunford, Z. P. Gumus, and C. S. Gur, "Chemical composition and antioxidant properties of pecan shell water extracts," *Antioxidants*, vol. 11, no. 6, pp. 1–13, 2022.

[14] A. A. Neira-Vielma, H. I. Meléndez-Ortiz, J. I. García-López, S. Sanchez-Valdes, M. A. Cruz-Hernández, J. G. Rodríguez-González, and S. N. Ramírez-Barrón, "Green synthesis of silver nanoparticles using pecan nut (*Carya illinoinensis*) shell extracts and evaluation of their antimicrobial activity," *Antibiotics*, vol. 11, no. 9, pp. 1–12, 2022.

[15] H. Singh, M. F. Desimone, S. Pandya, S. Jasani, N. George, M. Adnan, A. Aldarhami, A. S. Bazaid, and S. A. Alderhami, "Revisiting the green synthesis of nanoparticles: uncovering influences of plant extracts as reducing agents for enhanced synthesis efficiency and its biomedical applications," *International Journal of Nanomedicine*, vol. 18, no. 1, p. 4727–4750, 2023.

[16] M. Larrañaga-Tapia, B. Betancourt-Tovar, M. Videa, M. Antunes-Ricardo, and J. L. Cholula-Díaz, "Green synthesis trends and potential applications of bimetallic nanoparticles towards the sustainable development goals 2030," *Nanoscale Advances*, vol. 6, no. 1, p. 51–71, 2024.

[17] C. Pechyen, B. Tangnorawich, S. Toommee, R. Marks, and Y. Parcharoen, "Green synthesis of metal nanoparticles, characterization, and biosensing applications," *Sensors International*, vol. 5, no. 1, pp. 1–17, 2024.

[18] A. Kumar, P. K. Tyagi, S. Tyagi, and M. Ghorbanpour, "Integrating green nanotechnology with sustainable development goals: a pathway to sustainable innovation," *Discover Sustainability*, vol. 5, no. 1, pp. 1–13, 2024.

[19] L. F. W. Brum, M. D. C. R. da Silva, C. dos Santos, G. Pavoski, D. C. R. Espinosa, and W. L. da Silva, "Green synthesis of niobium (V) oxide nanoparticles using pecan nutshell (*Carya illinoinensis*) and evaluation of its antioxidant activity," *Catalysis Today*, vol. 445, no. 1, pp. 1–4, 2025.

[20] K. K. Ilavenil, A. Kasthuri, and P. Pandian, "Biosynthesis and antimicrobial investigation of strontium oxide (SrO) nanoparticles by *Lantana camara* leaf extract," *RASAYAN Journal of Chemistry*, vol. 16, no. 2, p. 596–603, 2023.

[21] A. Geremew, L. Carson, S. Woldesenbet, H. Wang, S. Reeves, N. Brooks, P. Saganti, A. Weerasooriya, and E. Peace, "Effect of zinc oxide nanoparticles synthesized from *Carya illinoinensis* leaf extract on growth and antioxidant properties of mustard (*Brassica juncea*)," *Frontiers in Plant Science*, vol. 14, no. 1, pp. 1–16, 2023.

[22] A. J. Samuel, A. S. Deepi, G. Srikesh, and A. S. Nesaraj, "One pot synthesis and characterisation of two dimensional tin doped strontium oxide nanostructured electrode materials for electrochemical supercapacitor applications," *Materials Technology*, vol. 37, no. 3, p. 150–160, 2020.

[23] B. Lewczuk and N. Szyryńska, "Field-emission scanning electron microscope as a tool for large-area and large-volume ultrastructural studies," *Animals*, vol. 11, no. 12, pp. 1–21, 2021.

[24] J. F. Nohl, N. T. Farr, Y. Sun, G. M. Hughes, S. A. Cussen, and C. Rodenburg, "Low-voltage sem of air-sensitive powders: from sample preparation to micro/nano analysis with secondary electron hyperspectral imaging," *Micron*, vol. 156, no. 1, pp. 1–22, 2022.

[25] D. G. Boeckel, R. S. A. Shinkai, M. L. Grossi, and E. R. Teixeira, "*In vitro* evaluation of cytotoxicity of hyaluronic acid as an extracellular matrix on OFCOL II cells by the MTT assay," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 117, no. 6, pp. 1–6, 2014.

[26] N. I. Vargas-Maya, F. Padilla-Vaca, O. E. Romero-González, E. A. S. Rosales-Castillo, Á. Rangel-Serrano, S. Arias-Negrete, and B. Franco, "Refinement of the griess method for measuring nitrite in biological samples," *Journal of Microbiological Methods*, vol. 187, no. 1, pp. 1–8, 2021.

[27] N. Bhagya, A. Giresha, S. Rao, G. Prashanth, H. Lalithamba, A. Sowmyashree, K. Ravindra, and M. Aman, "Eco-friendly fabrication of SrO nanoparticles via *Pongamia pinnata* extract: Structural, electrochemical, and multifunctional biological assessments," *Ceramics International*, vol. 51, no. 21, pp. 34 038–34 049, 2025.

[28] P. Lu, X. Hu, Y. Li, M. Zhang, X. Liu, Y. He, F. Dong, M. Fu, and Z. Zhang, "One-step preparation of a novel $SrCO_3$/g-C_3N_4 nanocomposite and its application in the selective adsorption of crystal violet," *RSC Advances*, vol. 8, no. 12, pp. 6315–6325, 2018.

[29] N. P. Radhika, M. S, S. G, K. Raj, S. K R, and A. Appaji, "Data-driven analysis of photocatalytic degradation of pararosaniline using green synthesized SrO nanoparticles," *Data in Brief*, vol. 63, no. 1, pp. 1–11, 2025.

[30] R. Kishorwara, P. S. Devika, M. Madhumitha, A. Mohanprasanth, and R. Snega, "Structural and optical characteristics of green-synthesized ZnO/SrO nanocomposites using pomegranate peel extract and their biological activity," *Next Materials*, vol. 12, no. 1, pp. 1–10, 2026.

[31] M. Fawzy, R. Alzain, A. S. Eltaweil, and E. M. Abd El-Monaem, "Engineering of balanites aegyptiaca-derived SrO@biochar and SrO@biomass for nitrophenol removal from wastewater," *Scientific Reports*, vol. 15, no. 1, pp. 1–19, 2025.

[32] M. Thommes, K. Kaneko, A. V. Neimark, J. P. Olivier, F. Rodriguez-Reinoso, J. Rouquerol, and K. S. Sing, "Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (IUPAC technical report)," *Pure and Applied Chemistry*, vol. 87, no. 9-10, pp. 1051–1069, 2015.

[33] T. C. Dotta, L. Hayann, L. de Padua Andrade Almeida, L. F. B. Nogueira, M. M. Arnez, R. Castelo, A. F. B. Cassiano, G. Faria, M. Martelli-Tosi, M. Bottini, P. Ciancaglini, A. B. C. E. B. Catirse, and A. P. Ramos, "Strontium carbonate and strontium-substituted calcium carbonate nanoparticles form protective deposits on dentin surface and enhance human dental pulp stem cells mineralization," *Journal of Functional Biomaterials*, vol. 13, no. 4, pp. 1–22, 2022.

[34] T. Brennan, M. Rybchyn, W. Green, S. Atwa, A. Conigrave, and R. Mason, "Osteoblasts play key roles in the mechanisms of action of strontium ranelate," *British Journal of Pharmacology*, vol. 157, no. 7, pp. 1291–1300, 2009.

Droplet Generation in Flow-Focusing Microfluidic Devices Fabricated via MSLA 3D Printing

Júlio Gallinaro Maranho
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
julio.gallinaro@ufabc.edu.br

João Lameu da Silva Junior
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
São Bernardo do Campo, SP, Brazil
joao.lameu@ufabc.edu.br

***Abstract*—We compare droplet generation performance across three flow-focusing geometries fabricated via masked stereolithography (MSLA) 3D printing using an Elegoo Saturn 4 Ultra and Anycubic High Clear photopolymer resin. The geometries investigated are: (FF-I) symmetric cross-junction, (FF-II) tapered-entry diverging nozzle, and (FF-III) straight-entry diverging nozzle. All devices were post-processed with a two-cycle resin glazing protocol to restore optical transparency. Mineral oil with 2 wt% Span 80 was used as the continuous phase and a 10 mmol/L methylene blue aqueous solution as the dispersed phase. Droplet diameter, shape and frequency were characterized as a function of the flow-rate ratio $\phi = Q_d/Q_c$ for each geometry. Across all three geometries at fixed $Q_d = 20\,\mu$L/min and $Q_c = 200$–$800\,\mu$L/min, droplet generation was monodisperse ($\mathrm{CV}_d < 4.1\%$) throughout the dripping regime, producing near-circular droplets smaller than the channel outlet width. Mean droplet diameter and generation frequency each followed clear power laws in the flow-rate ratio ϕ, with near-identical exponents across the three designs ($m \approx 0.52$ for diameter, above the mass-conservation value of $4/9$, and $m \approx -1.4$ for frequency, matching the theoretical $-4/3$), while FF-II produced droplets $\approx 1.4\times$ larger than FF-I and FF-III. The measured frequency exponents closely matched a mass-conservation scaling model derived from T-junction theory, while the diameter exponents lay slightly above the theoretical prediction. The MSLA fabrication workflow produces functional devices in under 2 h 30 min print time, making rapid geometry iteration feasible on a single working day.**

***Index Terms*—droplet microfluidics, flow-focusing, power-law scaling, 3D printing**

I. Introduction

Flow-focusing (FF) is among the most widely used geometries for generating monodisperse water-in-oil (W/O) droplets in microfluidic devices. In the canonical symmetric FF design, two streams of continuous phase squeeze a central stream of dispersed phase through a narrow orifice, driving periodic pinch-off [2]. The resulting droplet diameter and generation frequency are controlled by the capillary number $Ca = \mu_c v/\gamma$, where μ_c is the continuous-phase viscosity, v a characteristic velocity, and γ the interfacial tension, and the flow-rate ratio $\phi = Q_d/Q_c$, where Q_d and Q_c are the dispersed- and continuous-phase volumetric flow rates, respectively [3]. Modifications to the basic geometry, such as introducing an asymmetric constriction or altering the focusing-entry geometry, can extend the accessible droplet-size range, improve monodispersity, or stabilize the dripping regime at higher throughput [1].

The ability to generate thousands of size-controlled aqueous droplets per hour makes droplet microfluidics a powerful tool in biomedical research. Because each droplet acts as an isolated microreactor, W/O emulsions produced by FF devices have been applied to tumor spheroid formation, organoid culture, high-throughput drug screening, and cell encapsulation for tissue engineering and cell therapy [1], [18]–[20]. In these applications, droplet size uniformity is not merely a physical metric: it directly governs the homogeneity of the encapsulated cell aggregate, since droplet volume sets the initial cell density and the resulting gradients of oxygen and nutrients that drive three-dimensional tissue-like organization [11], [18]. Achieving monodisperse generation, typically $\mathrm{CV} < 5\%$, across a wide range of droplet volumes therefore underpins the translational relevance of any FF platform.

Conventional fabrication of FF devices relies on soft lithography in poly(dimethylsiloxane) (PDMS), which requires a photolithographic master and clean-room infrastructure [4]. Masked stereolithography (MSLA) 3D printing offers a geometry-agnostic, rapid-prototyping alternative: all three device variants investigated here can be printed simultaneously on a single build plate in under 2 h 30 min with no tooling change. The broader landscape of 3D printing strategies for microfluidics has been reviewed by Su et al. [5]. Within droplet microfluidics specifically, 3D-printed modular devices have generated single and double emulsions using T-junction and co-flow geometries with active pneumatic control [6].

LCD-based vat photopolymerisation printers have achieved lateral features below 100 µm and embedded microchannels down to $170 \times 220\,\mu\mathrm{m}^2$ on low-cost Elegoo Saturn machines [14]; the Elegoo Saturn 4 Ultra specifically has been validated for transparent microfluidic devices [13] and as a PDMS master [15]. A comparison of an $18,000 DLP and a $380 LCD printer found LCD dimensional accuracy for open microchannels within 2.5% of nominal with optimized

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001, and by CNPq (National Council for Scientific and Technological Development, Process 405055/2023-4).

exposure [7], while custom resins and post-processing push channel dimensions and optical clarity further [8]–[10].

The three geometries investigated here, a symmetric cross-junction (FF-I), a tapered-entry diverging nozzle (FF-II), and a straight-entry diverging nozzle (FF-III), represent three canonical flow-focusing classes [1]–[3], each imposing a distinct pinch-off mechanism: orifice-driven squeezing (FF-I), shear-dominated elongation along a gradual convergence (FF-II), and abrupt-constriction focusing followed by nozzle expansion (FF-III). The central hypothesis is that these mechanistic differences produce distinct droplet sizes and generation frequencies over the same ϕ window.

This paper presents a systematic comparison of droplet generation behavior across three MSLA-printed FF geometries, characterized by droplet diameter d, coefficient of variation CV, and generation frequency f as a function of ϕ.

II. Device Design

Three flow-focusing variants were designed (Fig. 1) and printed simultaneously on a single build plate. All devices share the same body footprint ($\approx 35 \times 35 \times 5$ mm), a uniform channel height of 1.0 mm, and a 2 mm port diameter; they differ only in the geometry of the focusing junction and the downstream channel. The three designs were chosen to represent distinct pinch-off mechanisms while keeping fabrication conditions identical, isolating the effect of junction geometry on droplet formation. Key cross-sectional dimensions are summarized in Table I.

TABLE I
Geometry Parameters

Parameter	**FF-I**	**FF-II**	**FF-III**
Channel dimensions (mm)			
Continuous channel width	0.8	1.0	1.0
Dispersed inlet width	1.0	0.8	0.8
Junction / orifice width	0.5	0.6	0.6
Outlet / exit width	1.5	1.5	1.5
Nozzle length	0.5	9.0	9.0
Channel height	1.0	1.0	1.0
Nozzle half-angle	–	2.9°	2.9°
Focusing entry type	concave	tapered	straight
Derived quantities			
A_{or} at junction (mm^2)	0.50	0.60	0.60

FF-I is a symmetric cross-junction with a concave-wall orifice, adapted from the flow-focusing geometry reported by Lee et al. [11]. Two lateral continuous-phase channels (0.8 mm wide) flank a 1.0 mm dispersed-phase inlet. The orifice is defined by symmetric concave cutouts in the lateral walls (inner radius $R = 0.25$ mm), forming a 0.5 mm constriction ($A_{or} = 0.50\,\text{mm}^2$) that squeezes the dispersed thread. The short orifice and abrupt confinement place this design in the classic squeezing-dominated regime.

FF-II is a tapered-entry diverging-nozzle geometry adapted from the Fluigent EZ Drop device [12]. Two lateral continuous-phase channels (1.0 mm wide) converge through symmetric trapezoidal walls (base 3.2 mm, height 0.9 mm) onto a 0.8 mm dispersed-phase inlet, meeting at a 0.6 mm orifice ($A_{or} = 0.60\,\text{mm}^2$). Downstream, the channel expands gradually along a 9 mm nozzle (half-angle $\approx 2.9°$) to a 1.5 mm exit. The progressive convergence of the continuous phase establishes a smooth velocity gradient at the orifice, favoring shear-driven thread elongation before pinch-off.

FF-III shares the orifice and nozzle of FF-II but replaces the tapered focusing walls with straight rectangular continuous-phase channels (1.0 mm wide) that meet the 0.8 mm dispersed-phase inlet at an abrupt 0.6 mm constriction ($A_{or} = 0.60\,\text{mm}^2$). Because A_{or} and the downstream nozzle are identical to FF-II, any difference in droplet behavior can be attributed solely to the entry geometry: the abrupt constriction concentrates the velocity discontinuity at the orifice plane, which is expected to trigger earlier thread instability and alter the pinch-off dynamics relative to the gradual entry of FF-II.

III. Fabrication

A. Printing

STL files for all three geometries were imported into Chitubox v1.1.0 (CBD-Tech, Shenzhen, China), arranged on a single build plate, and rotated 15° about the X-axis so that channel cross-sections were exposed transversely to the build platform. This orientation preserves the rectangular channel profile (avoiding staircase artifacts), prevents resin pooling inside the channel, and distributes peel forces to minimize warpage. Because each layer slice then intersects the channel perpendicularly, every newly exposed wall layer rests on cured resin from the previous layer, so no internal supports are required; auto-generated supports (Chitubox *Adaptive*, *Middle* density) were placed on the external device surfaces only.

Devices were printed on an Elegoo Saturn 4 Ultra 16K mono-LCD printer, with 14 × 19 µm pixel size and 405 nm wavelength (Elegoo, Shenzhen, China), using Anycubic High Clear resin (Anycubic, Shenzhen, China), following print parameters consistent with prior MSLA microfluidic work on the same platform [13], [15]: layer thickness 0.05 mm, normal layer exposure 2.0 s, bottom layer exposure 40 s (8 layers). All three devices were printed simultaneously; total print time was 2 h 26 min 11 s.

B. Resin Glazing for Optical Transparency

After printing, green-state devices were post-processed as follows (Fig. 2). First, parts were agitated for 5 min in isopropyl alcohol (≥99.5%) in the Anycubic Wash & Cure 3 station to remove uncured resin, then UV-cured for 5 min in the same station to consolidate the washed part before glazing. Next, a thin layer of Anycubic High Clear resin was applied to all surfaces with a flat nylon brush (7 mm × 10 mm) to fill the ∼50 µm layer-line corrugations and ∼19 µm pixel steps that scatter visible light and render the part opaque. The coated part was then UV-cured for 2 min; a thin IPA film was immediately applied and a second 2 min cure was performed under wet surface conditions to suppress oxygen inhibition of radical polymerization at the resin–air interface [16]. This glazing cycle was repeated a second time to further reduce residual

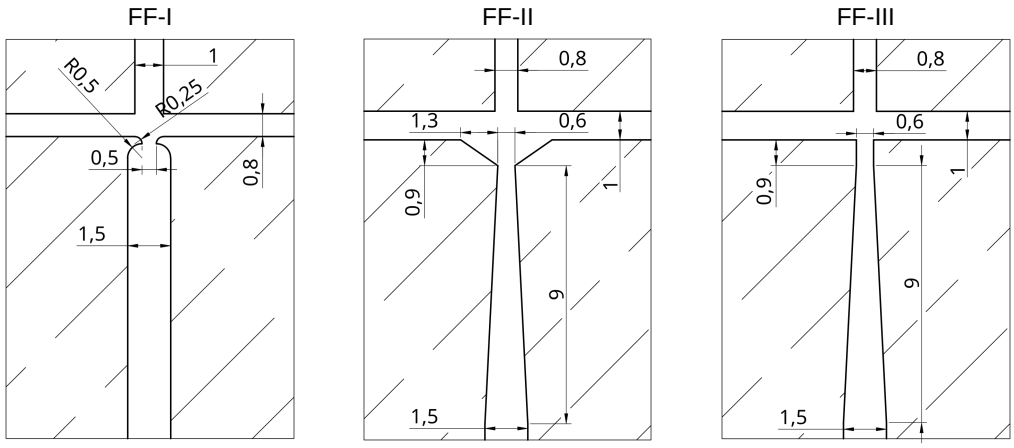


Fig. 1. CAD drawings of the three flow-focusing geometries (dimensions in mm): (a) FF-I symmetric cross-junction with concave-wall orifice, (b) FF-II tapered-entry diverging nozzle, and (c) FF-III straight-entry diverging nozzle. Key cross-sectional dimensions are listed in Table I.

haze, followed by a final 3 min dry UV cure to consolidate the glaze. Glazing was applied to the externally brushed faces required for imaging; the internal channel walls, which the brush cannot reach, retain the as-printed layer-line corrugation. After this protocol devices were visually transparent.

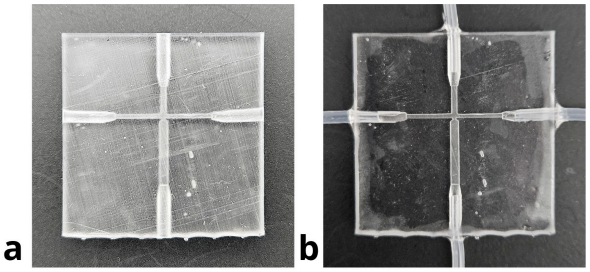


Fig. 2. Effect of resin glazing on optical transparency (black background). (a) FF-I device before glazing ; (b) same device after two-cycle glazing.

C. Tubing Bonding

Four silicone tubes (ID 1 mm, OD 2 mm, length 300 mm) were inserted 7 mm deep into lateral port holes (diameter 2.25 mm) on each of the four side faces of the device, producing an interference fit of 0.25 mm that retains the tube mechanically prior to bonding. Uncured Anycubic High Clear resin was applied around each tube–port interface and cured for 2 min under the UV station. This same-material bond avoids the adhesion incompatibilities observed with cyanoacrylate or epoxy on photopolymer substrates, and was leak-free at the operating pressures of this study. The tube-to-pump connections were made with blunt-tip dispensing needles (18 G $\times$ 12 mm) inserted into the silicone tubing and coupled to Luer-Lock syringes: a single 1 mL syringe for the dispersed phase and two 3 mL syringes for the continuous phase, one feeding each of the two lateral continuous-phase inlets.

IV. Experimental Setup

A. Syringe Pump

Flow was driven by an in-house syringe pump built around a bipolar stepper motor (6.7 Ω, 24 VDC, 0.6 A/phase; Sumtor Electric Co., Wuxi, China) mounted on a T-type linear rail (50 mm stroke, T6 trapezoidal lead-screw, nominal pitch 1 mm/rev), driven by a TB6600 microstepping driver (1/16 microstepping) and controlled by an Arduino microcontroller (AccelStepper firmware). The lead-screw pitch was calibrated to $L = 0.995$ mm/rev by commanding a 3 mm advance and measuring the actual carriage displacement with a dial indicator. The syringe carriage was fabricated by FDM 3D printing in PLA on a Creality Ender 3 V2 printer (Shenzhen, China), and accepts syringes of different diameters; the plunger diameter d_p was configured at runtime via the Arduino Serial Monitor for each syringe model. The pump accepts flow-rate commands in µL/min via the Arduino Serial Monitor and converts them to stepper pulse sequences via:

$$v_p = \frac{Q}{\pi d_p^2/4}, \qquad f = v_p \, \frac{N}{L} \tag{1}$$

where Q is the commanded flow rate, v_p is the corresponding linear plunger velocity, and f is the resulting stepper pulse frequency, set by the lead-screw pitch L (0.995 mm/rev) and the motor resolution N (6400 steps/rev). The smallest volume increment the pump can deliver is set by a single microstep and is given by

$$\Delta V = \frac{\pi d_p^2}{4} \, \frac{L}{N}, \tag{2}$$

i.e. the plunger cross-sectional area multiplied by the linear displacement per microstep (L/N), which defines the volumetric resolution of the pump for a given syringe. The two Luer-Lock syringe models used in this work (1 mL and 3 mL), together with their nominal plunger diameters d_p and the corresponding volumetric resolutions ΔV, are summarized in Table II.

TABLE II
VOLUMETRIC RESOLUTION PER SYRINGE

Syringe	d_p (mm)	ΔV (nL/microstep)
1 mL	4.78	2.79
3 mL	8.66	9.16

B. Fluids

The continuous phase was white mineral oil USP (Auros Química, São Paulo, Brazil) with 2 wt% Span 80 (sorbitan monooleate; Sigma-Aldrich, USA) dissolved by vortex mixing. The dispersed phase was a 10 mmol/L aqueous solution of methylene blue (Exodo Científica, São Paulo, Brazil) prepared in bi-distilled water. The methylene blue dyes the aqueous phase, providing the optical contrast needed to visualize and track the droplets against the surrounding oil under reflected-light illumination.

C. Imaging and Measurement

Devices were imaged using a custom reflected-light microscope assembled in-house on a CVI Melles Griot vibration-isolation optical breadboard (Fig. 3). The optical head consists of a 130× C-mount zoom lens coupled to a colour USB global-shutter camera (OV9782, Shenzhen Times Kangcheng Technology Co., Shenzhen, China). Illumination was provided by a ring LED lamp mounted on the front of the C-mount lens, directing light onto the device surface (reflected-light illumination); optical contrast between the aqueous droplets and the surrounding oil phase was enhanced by the methylene blue.

The camera–lens assembly was held on a magnetic-base post clamped directly to the CVI Melles Griot optical breadboard, providing a rigid, vibration-damped mount and defining the fixed imaging axis. Sample positioning was decoupled from the camera: each device was bonded with double-sided tape to a standard microscope slide along a short ($\approx$ 9 mm) strip at its upstream edge, near the dispersed-phase inlet, leaving the remainder of the device and the outlet tubing free of the chariot. The slide was mounted on the chariot, an XY positioning stage fixed via a 3D-printed adapter (double-sided tape) to a 100 mm × 100 mm granite block. The block rested on the table and was held in place by its own weight, allowing the whole sample assembly to be repositioned while its mass damped residual vibration. A Petri dish placed alongside the device collected waste fluid during operation, and a laptop was used both for video acquisition from the USB camera and for sending flow-rate commands to the Arduino-controlled syringe pump.

Videos were acquired at a resolution of 640 × 480 pixels and 180 fps using AMCap v3.0.9 (Microsoft, Redmond, USA). Droplets were detected using a YOLOv8 model trained on a dataset generated with the semi-automatic annotation pipeline described in [17]. To expose the detector to the full range of droplet morphologies, training frames were extracted from FF-I videos acquired at the two extreme generation regimes, the slowest (Q_d = 20 µL/min, Q_c = 200 µL/min) and the fastest (Q_d = 20 µL/min, Q_c = 400 µL/min), so that both the large, low-frequency plugs and the small, high-frequency droplets were represented in the dataset. Model training was performed inside a Docker container built from the official Ultralytics image (tag: 8.4.82). The trained model reached a detection performance of $\mathrm{mAP}_{50\text{-}95} = 0.95$ on the validation set, providing reliable localization for the subsequent tracking and sizing stages.

The detected droplets were subsequently processed using the droplet tracking, counting, and sizing pipeline described in [17], implemented in Python 3.12, to estimate the equivalent circular diameter (d) from the projected droplet area and the droplet generation frequency (f) from droplet passage events over time. For each operating condition (Q_c, Q_d), a 30 s video segment was analyzed. The coefficient of variation was calculated as $\mathrm{CV} = (\sigma_d/\bar{d}) \times 100\%$, where $\bar{d}$ and σ_d are the mean and standard deviation of the droplet diameter distribution, respectively.

V. RESULTS AND DISCUSSION

All three geometries generated monodisperse droplets in the dripping regime across the full flow-rate window tested (Q_d = 20 µL/min, Q_c = 200–800 µL/min), with droplet diameter, generation frequency, and their coefficients of variation summarized in Table III. Figure 4 shows a representative device (FF-III) in operation together with the automated droplet detection used to extract these quantities.

A. Theoretical Scaling Prediction from Mass Conservation

To provide a quantitative benchmark for the scaling exponents measured across the three geometries, we derived the expected dependence of droplet diameter d and generation frequency f on the flow-rate ratio ϕ, starting from mass conservation of the dispersed phase. In steady state, all fluid injected at rate Q_d must leave as droplets, so

$$f \cdot V = Q_d, \tag{3}$$

where V is the volume of an individual droplet.

Christopher et al. [21] reported, for T-junction geometries at low viscosity ratio, the empirical power-law scalings:

$$f \propto Ca^{4/3}, \qquad V \propto Ca^{-1/3}, \tag{4}$$

which were independently measured but mutually consistent with Eq. (3): combining them gives $f \cdot V \propto Ca^{4/3} \cdot Ca^{-1/3} = Ca \propto Q_c$, consistent with $f \cdot V = Q_d$ at fixed ϕ, where $Q_d \propto Q_c$.

To express these relations in terms of ϕ rather than Ca alone, we note that the capillary number scales with the continuous-phase flow rate, $Ca \propto Q_c$, and that $Q_d = \phi Q_c$. Substituting $Q_c = Q_d/\phi$ and holding Q_d fixed (as in the present experiments), Eq. (4) for the frequency becomes

$$f \propto Ca^{4/3} \propto Q_c^{4/3} = \left(\frac{Q_d}{\phi}\right)^{4/3} \propto \phi^{-4/3}. \tag{5}$$

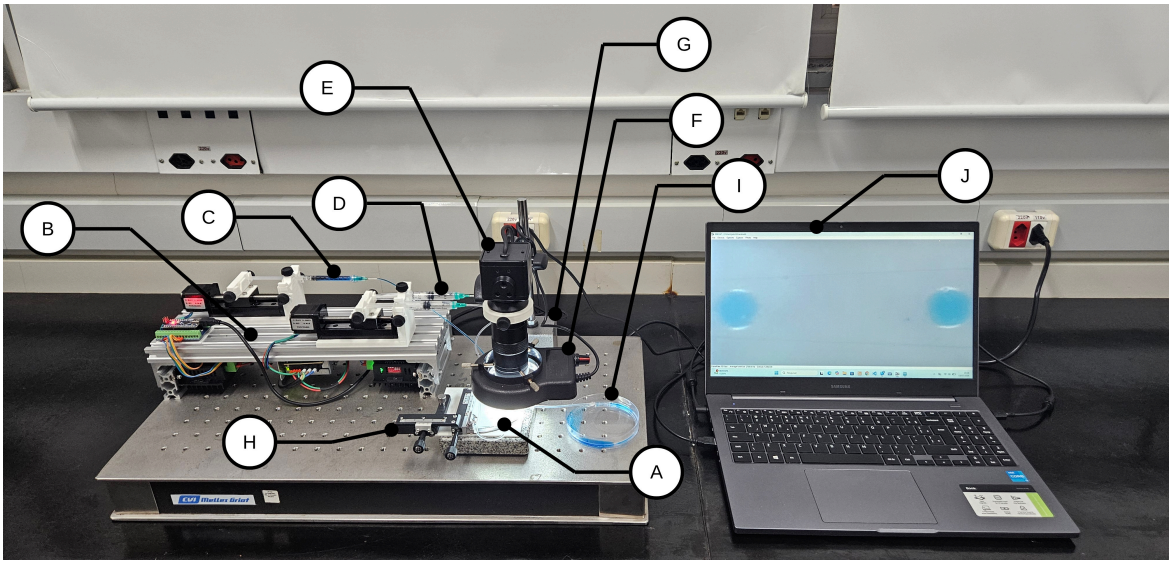


Fig. 3. Experimental setup mounted on the CVI Melles Griot vibration-isolation optical breadboard. (A) Microfluidic device; (B) syringe pump; (C) dispersed-phase syringe (aqueous methylene blue solution); (D) continuous-phase syringe (mineral oil); (E) camera; (F) ring light; (G) magnetic base clamped to the optical breadboard; (H) chariot for device positioning; (I) Petri dish for waste collection; (J) computer for video acquisition and syringe-pump control.

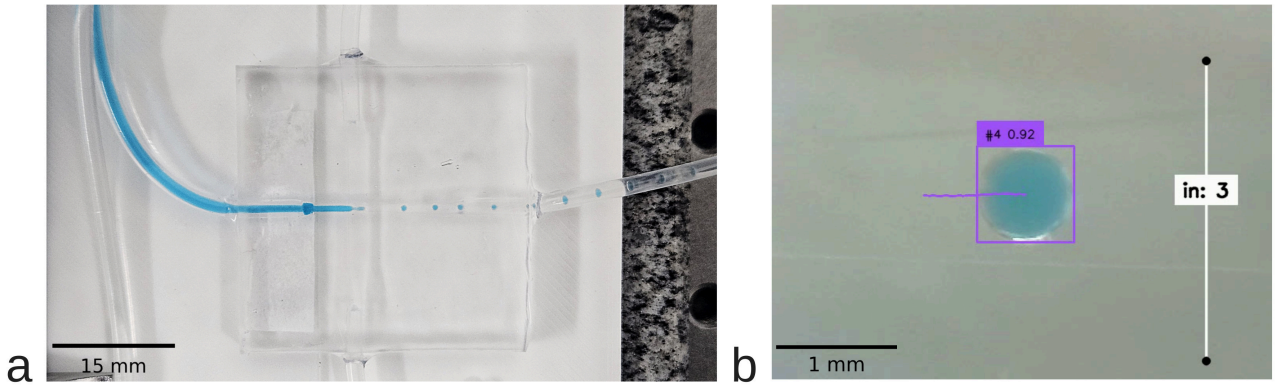


Fig. 4. Droplet generation and automated detection in the MSLA-printed FF-III device operating in the dripping regime ($Q_d = 20\,\mu$L/min, $Q_c = 200\,\mu$L/min; $\phi = 0.1$). (a) Transparent device in operation: the dispersed aqueous phase (methylene blue) enters from the right and breaks into a train of monodisperse droplets carried along the channel, with the outlet tubing running freely off the chariot. (b) Acquired video frame showing a droplet localized by the YOLOv8 detector (bounding box, track ID #4, detection confidence 0.92) and the crossing-count line used to estimate generation frequency (cumulative count: 3).

For the droplet volume, care must be taken because the empirical relation $V \propto Ca^{-1/3}$ in Eq. (4) was reported at fixed ϕ; the full dependence on both Ca and ϕ follows directly from mass conservation, $V = Q_d/f \propto Q_d \cdot Ca^{-4/3}$. Using $Q_d = \phi Q_c \propto \phi\, Ca$,

$$V \propto \phi\, Ca \cdot Ca^{-4/3} = \phi\, Ca^{-1/3} \propto \phi \left(\frac{Q_d}{\phi}\right)^{-1/3} \propto \phi^{4/3}, \tag{6}$$

at fixed Q_d. Assuming approximately spherical droplets ($V \propto d^3$), this yields

$$d \propto V^{1/3} \propto \phi^{4/9} \approx \phi^{0.444}. \tag{7}$$

Table IV summarizes the theoretical exponents predicted by Eqs. (5) and (7) alongside the exponents fitted from the present dataset (Sections V-B and V-D). While the frequency scaling exponents agreed closely with the T-junction prediction, the fitted diameter exponents lie slightly above the theoretical value across all three geometries. As discussed below, this offset is small relative to the uncertainty of a four-point power-law fit, and its near-identical magnitude across three geometrically distinct orifices argues against a junction-specific origin; we therefore anchor the quantitative comparison with theory on the directly counted frequency rather than on the diameter.

B. Droplet Size Distribution and Scaling with Flow-Rate Ratio

For all three geometries, droplet size distributions were narrow ($\mathrm{CV}_d < 4.1\%$ across all conditions), confirming stable, monodisperse generation over the entire flow-rate window tested. Fig. 5(a) shows the diameter distributions grouped by Q_c for each device. As expected, d decreases monotonically

TABLE III
DROPLET CHARACTERISATION SUMMARY

Geom.	Q_c (µL/min)	ϕ	d (µm)	CV$_d$ (%)	f (Hz)	CV$_f$ (%)
FF-I	200	0.100	814.5 ± 7.0	0.9	0.90 ± 0.04	4.4
	400	0.050	616.6 ± 10.8	1.7	1.81 ± 0.09	5.0
	600	0.033	473.8 ± 9.4	2.0	3.53 ± 0.16	4.6
	800	0.025	392.7 ± 9.9	2.5	6.41 ± 0.38	5.9
FF-II	200	0.100	1131.1 ± 11.7	1.0	0.51 ± 0.02	3.9
	400	0.050	906.3 ± 12.6	1.4	1.02 ± 0.04	3.5
	600	0.033	711.5 ± 22.4	3.1	1.87 ± 0.13	7.1
	800	0.025	531.0 ± 14.3	2.7	3.77 ± 0.25	6.6
FF-III	200	0.100	818.5 ± 7.5	0.9	0.88 ± 0.02	2.4
	400	0.050	611.6 ± 15.6	2.6	1.64 ± 0.07	4.4
	600	0.033	495.6 ± 17.3	3.5	3.02 ± 0.18	6.0
	800	0.025	385.1 ± 15.8	4.1	6.95 ± 0.61	8.7

TABLE IV
THEORETICAL VS. MEASURED SCALING EXPONENTS ($\propto \phi^m$)

		Measured		
Quantity	**Theory**†	**FF-I**	**FF-II**	**FF-III**
f	$-4/3 \approx -1.333$	-1.39	-1.40	-1.41
d	$4/9 \approx 0.444$	0.52	0.52	0.53

†Derived from the scaling model of [21].

with increasing Q_c at fixed Q_d for all geometries, consistent with the shear-driven pinch-off mechanism of the dripping regime [2]. Notably, FF-I and FF-III produce droplets of nearly identical size at every condition tested (e.g. 814.5 ± 7.0 vs. 818.5 ± 7.5 µm at $Q_c = 200$ µL/min), despite their different orifice widths (0.5 vs. 0.6 mm). FF-II, in contrast, produces systematically larger droplets than both other geometries at every Q_c investigated, by a factor of 1.3–1.5× (e.g. 1131.1 ± 11.7 µm vs. 814–819 µm at $Q_c = 200$ µL/min). Since FF-II and FF-III share an identical orifice area ($A_{or} = 0.60\,\text{mm}^2$) and downstream nozzle, this difference can be attributed entirely to the entry geometry: the gradual convergence of the tapered walls in FF-II appears to sustain a longer dispersed-phase thread before the confined pressure build-up required for pinch-off is reached, delaying breakup relative to the abrupt constriction of FF-III.

The scaling of mean diameter with the flow-rate ratio ϕ is shown in Fig. 6(a) on logarithmic axes. All three geometries follow a clear power-law relationship, $\bar{d} \propto \phi^m$, with fitted exponents $m = 0.52$ (FF-I, $R^2 = 0.983$), $m = 0.52$ (FF-II, $R^2 = 0.934$), and $m = 0.53$ (FF-III, $R^2 = 0.974$). The near-identical exponents across all three geometries indicate that, while the absolute droplet volume is geometry-dependent (governed by orifice width and entry angle), the underlying scaling mechanism relating droplet size to the flow-rate ratio is conserved across all three junction designs.

The measured exponents ($m \approx 0.52$–0.53) lie slightly above the mass-conservation value $m_{\text{theory}} = 4/9 \approx 0.444$ (Table IV; Section V-A). Since Q_d is held fixed here, the $V \propto Ca^{-1/3}$ scaling cannot be verified from diameter data alone; the small residual offset is within the combined uncertainty of the four-point fit and the spherical approximation $d \propto V^{1/3}$.

C. Droplet Shape

Droplet aspect ratio, $\Psi = \min(w,h)/\max(w,h)$ (with $\Psi = 1$ corresponding to a perfectly circular projected shape), was evaluated to assess deviations from sphericity across the four Q_c values (Fig. 5(b)). All three geometries produced predominantly near-circular droplets ($\Psi > 0.89$ in all conditions), consistent with operation in the dripping regime. FF-II consistently exhibited the lowest median Ψ among the three geometries, most pronounced at $Q_c = 600$ µL/min (median $\Psi \approx 0.925$, vs. ≈ 0.98 for FF-I and FF-III), indicating more elongated droplets. Two contributions to this elongation are plausible and cannot be separated with the present dataset: (i) greater shear-driven stretching of the larger-volume FF-II droplets along the flow direction, and (ii) drag and contact-line pinning at the channel top and bottom surfaces — which retain the layer-line corrugation of the MSLA print orientation — since only the FF-II droplets grow comparable to the 1.0 mm channel height and can contact these walls. The smaller droplets produced by FF-I and FF-III remain near-circular across all conditions.

D. Generation Frequency

Fig. 5(c) shows the generation frequency distributions (computed over non-overlapping windows of $N = 3$ consecutive droplet crossing events, $f = (N-1)/\Delta t$; the number of windows per condition ranged from 5 at the lowest flow rate to 69 at the highest) grouped by Q_c. Frequency increased monotonically with Q_c for all geometries (Table III). Consistent with its larger droplet volume, FF-II generated droplets at a systematically lower frequency than FF-I and FF-III at every Q_c tested (e.g. 0.51 ± 0.02 vs. 0.90 ± 0.04 and 0.88 ± 0.02 Hz at $Q_c = 200$ µL/min).

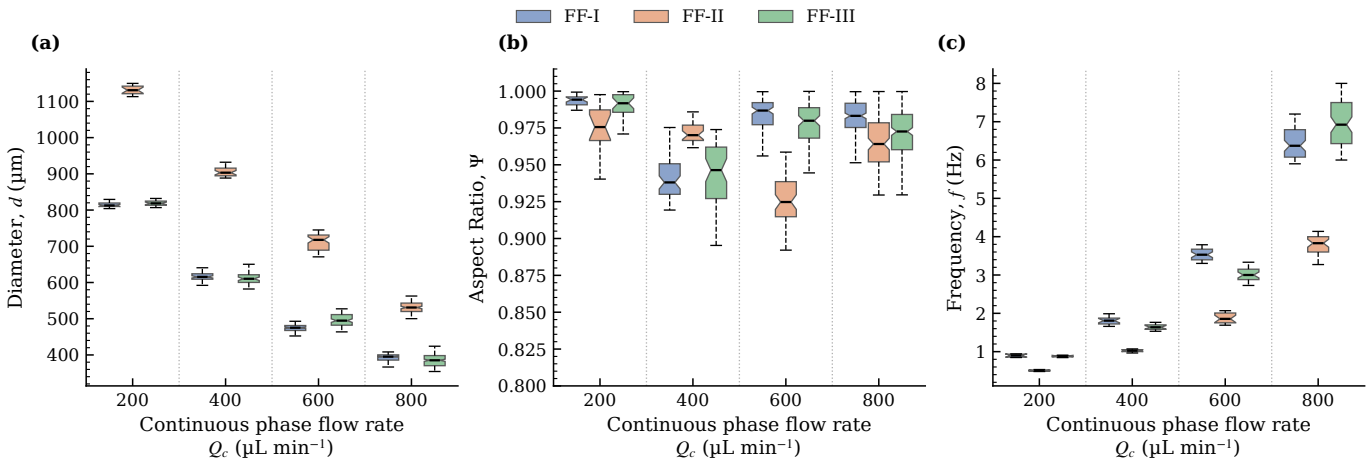


Fig. 5. (a) Droplet diameter, (b) aspect ratio, and (c) generation frequency distributions for the three flow-focusing geometries as a function of the continuous-phase flow rate Q_c ($Q_d = 20$ µL/min fixed). Boxes show the interquartile range with notches indicating the 95% confidence interval of the median; whiskers extend to the full data range.

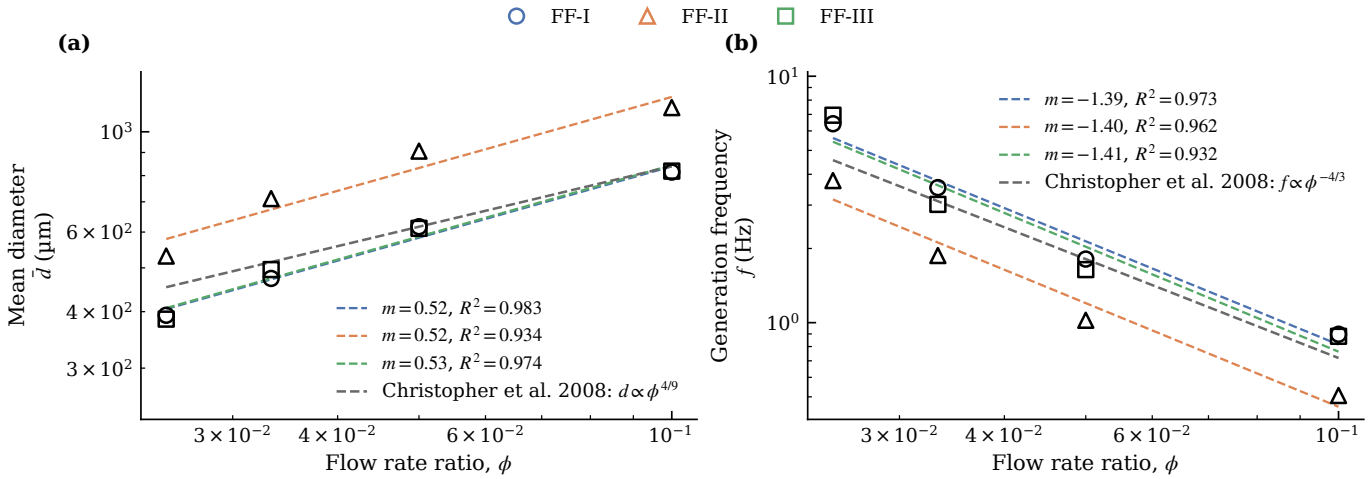


Fig. 6. (a) Mean droplet diameter and (b) generation frequency versus flow-rate ratio ϕ, on logarithmic axes, for the three geometries. Dashed colored lines show the fitted power law for each device; the grey dashed line shows the mass-conservation prediction from Christopher et al. [21].

The scaling of f with ϕ (Fig. 6(b)) follows a power law $f \propto \phi^m$ with $m = -1.39$ (FF-I, $R^2 = 0.973$), $m = -1.40$ (FF-II, $R^2 = 0.962$), and $m = -1.41$ (FF-III, $R^2 = 0.932$). These values are in close agreement with the theoretical prediction $f \propto \phi^{-4/3}$ ($m_{\text{theory}} \approx -1.333$) derived from the same mass-conservation argument combined with the Christopher et al. $V \propto Ca^{-1/3}$ scaling [21]. The frequency scaling was measured directly from droplet-crossing counts and does not depend on the spherical approximation required to relate diameter to volume; it therefore provided the more robust benchmark against the mass-conservation prediction, and its close agreement with the T-junction exponent indicates that the frequency response of all three FF geometries to the flow-rate ratio was governed by a mechanism near the T-junction limit.

VI. Conclusion

This work demonstrated that three distinct flow-focusing geometries (FF-I symmetric cross-junction, FF-II tapered-entry diverging nozzle, and FF-III straight-entry diverging nozzle) can be fabricated simultaneously in a single build via masked stereolithography (MSLA) 3D printing, with a total print time under 2 h 30 min. A two-cycle resin-glazing protocol restored optical transparency to the green-state parts, enabling reflected-light droplet imaging directly through the printed device body without recourse to PDMS soft lithography or clean-room infrastructure. The complete workflow, from geometry design to functional device, is achievable within a single working day, making rapid geometry iteration practical.

Across the systematic dataset ($Q_d = 20$ µL/min, $Q_c = 200$–800 µL/min), all three geometries produced monodisperse droplets ($\text{CV}_d < 4.1\%$) in the dripping regime, confirming that MSLA-printed devices can sustain stable, uniform

droplet generation. The mean droplet diameter and generation frequency each followed clear power laws in the flow-rate ratio ϕ, with near-identical exponents across the three designs ($m \approx 0.52$–0.53 for diameter and $m \approx -1.39$ to -1.41 for frequency). This indicated that the scaling mechanism relating droplet metrics to ϕ was conserved across junction geometries, even though the absolute droplet size is strongly geometry-dependent: FF-II produced droplets 1.3–$1.5\times$ larger than FF-I and FF-III at every condition, an effect attributed to the gradual convergence of its tapered entry walls, which sustains a longer dispersed-phase thread before pinch-off. The measured frequency exponents agree closely with the mass-conservation prediction derived from the T-junction scaling of Christopher et al. [21]; the diameter exponents lie slightly higher, within the uncertainty of a four-point fit and the spherical approximation used to relate diameter to volume.

Several limitations should be noted. The scaling relationships were obtained with a single fluid pair, a fixed surfactant concentration and a fixed Q_d, and for three junctions only; their extension to other viscosity ratios, surfactant loadings and broader classes of flow-focusing designs remains to be verified. The proposed pinch-off mechanisms are inferred from droplet metrics rather than observed directly, and the droplets were not tested in a biological assay. As next steps, the interfacial tension and viscosities of both phases will be measured directly (e.g. pendant-drop method), allowing Ca to be computed from first principles and the diameter-scaling offset to be re-examined. The three geometries will then be reproduced in CFD (volume-of-fluid) simulations complemented by direct flow visualization at the orifice, and the dataset will be extended to variable Q_d to verify the $V \propto Ca^{-1/3}$ and $f \propto Ca^{4/3}$ scalings independently. Finally, cytotoxicity assays of the chip material will be performed to confirm biocompatibility, followed by droplet generation with a cell-laden dispersed phase to produce cell spheroids and assess aggregate uniformity and viability.

Acknowledgment

The authors thank the Biophotonics and Biomedical Optics Laboratory and the Phototherapy and Microbiological Analysis Laboratory, both at the Federal University of ABC (UFABC), for providing laboratory space for the experiments reported in this work.

References

[1] S.-Y. Teh, R. Lin, L.-H. Hung, and A. P. Lee, "Droplet microfluidics," *Lab Chip*, vol. 8, no. 2, pp. 198–220, 2008, doi: 10.1039/b715524g.

[2] S. L. Anna, N. Bontoux, and H. A. Stone, "Formation of dispersions using flow focusing in microchannels," *Appl. Phys. Lett.*, vol. 82, no. 3, pp. 364–366, 2003, doi: 10.1063/1.1537519.

[3] P. Garstecki, M. J. Fuerstman, H. A. Stone, and G. M. Whitesides, "Formation of droplets and bubbles in a microfluidic T-junction—scaling and mechanism of break-up," *Lab Chip*, vol. 6, no. 3, pp. 437–446, 2006, doi: 10.1039/b510841a.

[4] J. C. McDonald and G. M. Whitesides, "Poly(dimethylsiloxane) as a material for fabricating microfluidic devices," *Acc. Chem. Res.*, vol. 35, no. 7, pp. 491–499, 2002, doi: https://doi.org/10.1021/ar010110q.

[5] R. Su, F. Wang, and M. C. McAlpine, "3D printed microfluidics: advances in strategies, integration, and applications," *Lab Chip*, vol. 23, pp. 1279–1299, 2023, doi: 10.1039/d2lc01177h.

[6] Q. Ji, J. M. Zhang, Y. Liu, X. Li, P. Lv, D. Jin, and H. Duan, "A modular microfluidic device via multimaterial 3D printing for emulsion generation," *Sci. Rep.*, vol. 8, p. 4791, 2018, doi: 10.1038/s41598-018-22756-1.

[7] K. M. Leong, A. Y. Sun, M. L. Quach, C. H. Lin, C. A. Craig, F. Guo, T. R. Robinson, M. M. Chang, and A. O. Olanrewaju, "Democratizing access to microfluidics: rapid prototyping of open microchannels with low-cost LCD 3D printers," *ACS Omega*, vol. 9, pp. 45537–45544, 2024, doi: 10.1021/acsomega.4c07776.

[8] H. Gong, B. P. Bickham, A. T. Woolley, and G. P. Nordin, "Custom 3D printer and resin for 18 µm × 20 µm microfluidic flow channels," *Lab Chip*, vol. 17, no. 17, pp. 2899–2909, 2017, doi: 10.1039/c7lc00644f.

[9] N. Bhattacharjee, A. Urrios, S. Kang, and A. Folch, "The upcoming 3D-printing revolution in microfluidics," *Lab Chip*, vol. 16, no. 10, pp. 1720–1742, 2016. doi: 10.1039/c6lc00163g.

[10] M. J. Beauchamp, G. P. Nordin, and A. T. Woolley, "Moving from millifluidic to truly microfluidic sub-100-µm cross-section 3D printed devices," *Anal. Bioanal. Chem.*, vol. 409, no. 18, pp. 4311–4319, 2017, doi: 10.1007/s00216-017-0398-3.

[11] J. M. Lee, J. W. Choi, C. D. Ahrberg, H. W. Choi, J. H. Ha, S. G. Mun, S. J. Mo, and B. G. Chung, "Generation of tumor spheroids using a droplet-based microfluidic device for photothermal therapy," *Microsyst. Nanoeng.*, vol. 6, p. 52, 2020, doi: 10.1038/s41378-020-0167-x.

[12] Fluigent, "EZ Drop data sheet (P/N Dropkit01)," [Online]. Available: https://www.fluigent.com/resources-support/support-tools/downloads/technical-datasheets/ez-drop-datasheet/. Accessed: Jun. 2026.

[13] A.-I. Radu, C. Mărculescu, E.-C. Creangă, B. Adiaconiță, E. Chiriac, T. Burinaru, and M. Avram, "The development of 3D resin-printed microfluidic devices for curcumin encapsulation in aspasomes," in *Proc. Int. Semiconductor Conf. (CAS)*, 2025, pp. 355–358, doi: 10.1109/CAS66707.2025.11222257.

[14] H. Shafique, V. Karamzadeh, G. Kim, M. L. Shen, Y. Morocz, A. Sohrabi-Kashani, and D. Juncker, "High-resolution low-cost LCD 3D printing for microfluidics and organ-on-a-chip devices," *Lab Chip*, vol. 24, pp. 2774–2790, 2024, doi: 10.1039/d3lc01125a.

[15] C. Xu, C. Shao, H. Jiang, C. Zeng, G. Zhang, X. Li, Y. Shen, Y. Yalikun, M. Li, X. Liu, N. Wu, and T. Tang, "Ring-shaped open microfluidic platform reveals viscosity-driven mechanical priming of glioblastoma cells," *ACS Sens.*, vol. 10, pp. 9457–9466, 2025, doi: 10.1021/acssensors.5c02716.

[16] S. C. Ligon, B. Husár, H. Wutzel, R. Holman, and R. Liska, "Strategies to reduce oxygen inhibition in photoinduced polymerization," *Chem. Rev.*, vol. 114, no. 1, pp. 557–589, 2014, doi: 10.1021/cr3005197.

[17] J. G. Maranho and J. L. Silva Jr., "Droplet Semi-Automatic Annotation," Zenodo, Jun. 2026, doi: 10.5281/zenodo.21140355.

[18] Y. Wang, M. Liu, Y. Zhang, H. Liu, and L. Han, "Recent methods of droplet microfluidics and their applications in spheroids and organoids," *Lab Chip*, vol. 23, pp. 1080–1096, 2023, doi: 10.1039/d2lc00493c.

[19] R. Fevre, G. Mary, N. Vertti-Quintero, A. Durand, R. F.-X. Tomasi, E. Del Nery, and C. N. Baroud, "Combinatorial drug screening on 3D Ewing sarcoma spheroids using droplet-based microfluidics," *iScience*, vol. 26, no. 5, p. 106651, 2023, doi: 10.1016/j.isci.2023.106651.

[20] H. Dortaj, A. M. Amani, L. Tayebi, N. Azarpira, M. Ghasemi Toudeshkchouei, and A. Hassanpour-Dehnavi, "Droplet-based microfluidics: an efficient high-throughput portable system for cell encapsulation," *J. Microencapsul.*, vol. 41, no. 6, pp. 479–501, 2024, doi: 10.1080/02652048.2024.2382744.

[21] G. F. Christopher, N. N. Noharuddin, J. A. Taylor, and S. L. Anna, "Experimental observations of the squeezing-to-dripping transition in T-shaped microfluidic junctions," *Phys. Rev. E*, vol. 78, no. 3, p. 036317, 2008, doi: 10.1103/PhysRevE.78.036317.

Surface Chemical Characterization of Gutta-Percha By FTIR/UATR After Hydrodynamic Ozonated Water Disinfection

Jessica sthefanny carvalho souza
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro da Inovação, Tecnologia e Educação (CITE)
São José dos Campos, SP, Brasil
souza.jessica.carvalho@gmail.com

Sara Evelyn Batista Aquino
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro da Inovação, Tecnologia e Educação (CITE)
São José dos Campos, SP, Brasil.
saraevelyn749@gmail.com

Pedro Luiz Corrêa da Silva Junior
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro da Inovação, Tecnologia e Educação (CITE)
São José dos Campos, SP, Brasil
prof.pedrolcsjunior@gmail.com

Milton Faria Diniz
Departamento de Ciência e Tecnologia Aeroespacial (DCTA)
Instituto de Aeronáutica e Espaço (IAE),
São José dos Campos, SP, Brasil.
mfariadiniz@gmail.com

Adriana Barrinha Fernandes Moretti
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro da Inovação, Tecnologia e Educação (CITE)
São José dos Campos, SP, Brasil
fernandesabm@gmail.com

Carlos José de Lima
Postgraduate Program in Biomedical Engineering, Universidade Anhembi Morumbi
Centro da Inovação, Tecnologia e Educação (CITE)
São José dos Campos, SP, Brasil
cdcfdlima@gmail.com

***Abstract*—** Ozonated water has emerged as a promising alternative for the disinfection of gutta-percha cones because of its broad-spectrum antimicrobial activity and the absence of toxic residues. However, information regarding its effects on the chemical structure of gutta-percha remains limited. Therefore, this study aimed to investigate the chemical effects of a hydrodynamic ozonated water system on gutta-percha using Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR/UATR), assessing whether ozone-induced chemical modifications remained detectable after superficial material removal. Three experimental conditions were evaluated: untreated gutta-percha, gutta-percha exposed to hydrodynamic ozonated water for 7 min, and ozonated gutta-percha after mechanical removal of a thin superficial layer. FTIR/UATR spectra were acquired in the range of 4000–550 cm^{-1} using 20 scans, and the spectra were qualitatively compared based on band position, shape, and relative intensity. The ozonated specimens exhibited broadening of the hydroxyl stretching region (3390–3445 cm^{-1}), changes in the spectral profile around 1667 cm^{-1}, and reduced intensity of the band near 795 cm^{-1}, consistent with ozone-induced chemical modifications of the polyisoprene matrix. After superficial abrasion, these spectral changes were no longer detectable, and the resulting spectral profile became comparable to that of untreated gutta-percha. These findings suggest that the chemical modifications induced by hydrodynamic ozonation were predominantly associated with the superficial region of the material under the conditions evaluated. However, the qualitative FTIR/UATR approach does not allow quantitative determination of the degree or depth of oxidation. Complementary surface-sensitive and depth-resolved analyses are therefore warranted to further characterize the extent of ozone-induced chemical modifications. The results support the potential application of hydrodynamic ozonated water as a gutta-percha disinfection strategy while highlighting the need for further investigation of the effects of superficial chemical modifications on sealer adhesion and long-term obturation performance.

***Keywords*—** Gutta-percha; Hydrodynamic ozonation; FTIR/UATR.

I. INTRODUCTION

Endodontic infections are predominantly microbial diseases that affect the root canal system and periapical tissues and are recognized as the primary cause of the development and persistence of apical periodontitis. These infections are characterized by the formation of organized biofilms and by the high adaptive capacity of microorganisms to the intraradicular environment, making their complete elimination during endodontic treatment challenging [1,2,3,4,5]. Therefore, the clinical success of endodontic therapy depends on a substantial reduction of the microbial load, adequate shaping and disinfection of the root canal system, and the achievement of an effective three-dimensional seal capable of preventing reinfection of the endodontic space [5,6,7,8].

Among the materials used for root canal obturation, gutta-percha (GP) remains the material of choice and is considered the gold standard in Endodontics because of its biocompatibility,

dimensional stability, radiopacity, low cytotoxicity, ease of handling, and compatibility with different obturation techniques[9,10,11]. Chemically, commercial GP is a composite material composed predominantly of trans-1,4-polyisoprene associated with zinc oxide, barium sulfate, waxes, resins, plasticizers, and pigments, whose proportions may vary according to the manufacturer [9,13,14,15]. This composition directly influences its physical, mechanical, and thermal properties, as well as its behavior during clinical procedures performed before obturation [13,14,15].

Although GP cones are commercially available as sterile products, several studies have demonstrated that they may become contaminated during storage, transportation, clinical handling, or exposure to the operative environment, making disinfection immediately before insertion into the root canal essential [9,10,11,12]. For this purpose, different chemical agents, such as sodium hypochlorite, chlorhexidine, alcohol, and peracetic acid, have been widely used because of their high antimicrobial activity. However, depending on the concentration employed and the exposure time, these agents may induce surface alterations, structural modifications, and changes in the physicochemical properties of GP, thereby compromising its adaptation to the root canal walls and, consequently, the quality of the obturation seal [12,13,14,15].

In this context, ozonated water has attracted increasing interest as an alternative for the disinfection of dental biomaterials due to its high oxidative potential, broad-spectrum antimicrobial activity, effectiveness against microorganisms organized in biofilms, and rapid decomposition into oxygen without leaving toxic residues at the end of the process. Furthermore, the use of hydrodynamic ozonation systems enhances ozone transfer into water, resulting in more stable dissolved ozone concentrations and improving its disinfecting efficacy. Recent studies have demonstrated that this system provides high disinfection efficiency against *Enterococcus faecalis* contaminated GP cones while preserving their mechanical and morphological properties following treatment with ozonated water [19].

Despite the promising findings regarding the antimicrobial activity of ozonated water, studies investigating its potential effects on the chemical composition and thermal stability of GP remain limited. Considering that structural modifications may compromise the clinical performance of the obturation material, it is essential to determine whether exposure to a hydrodynamic ozonated water system promotes changes in its polymeric matrix or inorganic constituents.

In this regard, Fourier-transform infrared spectroscopy (FTIR) is a widely employed technique for the characterization of biomaterials, allowing the identification of functional groups, chemical bonds, and possible structural changes resulting from different treatments [16,17].

Therefore, the present study aimed to evaluate the chemical effects of hydrodynamic ozonation on gutta-percha by means of Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR/UATR), comparing untreated, ozonated, and superficially abraded ozonated specimens to determine whether ozone-induced oxidation is limited to the material surface or affects deeper layers of the polymeric matrix.

II. Materials and Methods

A. Study design

This in vitro study aimed to investigate the chemical effects of a hydrodynamic ozonation system on gutta-percha cones using Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR/UATR). The experimental design comprised three conditions: untreated gutta-percha (control group), gutta-percha exposed to ozonated water for 7 min using the hydrodynamic ozonation system and ozonated gutta-percha after the mechanical removal of a thin superficial layer. FTIR/UATR analysis was performed to compare the chemical profiles of the three conditions and determine whether the oxidative changes induced by ozonation were restricted to the material surface or extended into its internal structure.

B. Gutta-percha samples

Standardized GP cones (DiaDent ML.029™, size 70; DiaDent Group International, South Korea) from the same manufacturing batch were used throughout the study to minimize compositional variability. The samples were randomly divided into two experimental groups: control group, *in natura* and cones exposed to the hydrodynamic ozonation system (Figure 1).

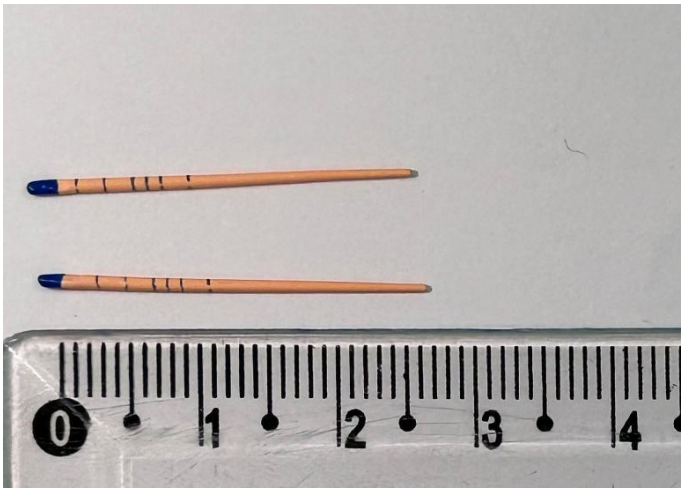


Fig 1. Gutta-percha cone used in the experimental procedures. In the scale bar shown below, the distance between adjacent tick marks corresponds to 1 mm. Author's own work.

C. Hydrodynamic ozonation system

The system consisted of a 500 mL water reservoir connected to a centrifugal pump. A Venturi injector was attached to the pump outlet, with a third inlet for ozone gas, enabling dynamic mixing with water. The fluid was directed to a tubular head via silicone tubing, where the gutta-percha samples were positioned for disinfection (Figure 2). The kinetic effect of ozonated water and ozone's oxidative action were expected to remove and oxidize organic material deposited on the sample surfaces. Experiments were conducted in a climate-controlled environment at 21 °C. An ozone generator (MS3G, MS Ltd.a, Brazil) was set to a concentration of 57.2 mg/L, with oxygen gas flow adjusted to 0.250 L/min. Initially, the system was activated without samples to allow complete ozonation of the water and internal

disinfection of the system, requiring 10 min to inactivate any residual microorganisms.

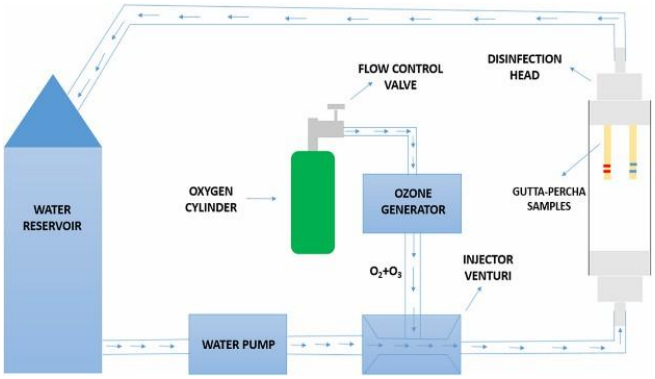


Fig. 2 Schematic representation of the system used for gutta-percha disinfection with ozonated water via a hydrodynamic setup. The samples, consisting of eight units, were internally fixed within the tubular head and exposed to the dynamic flow of ozonated fluid. **Source:** souza *et al.,* 2026

D. Fourier transform infrared spectroscopy (FTIR)

A chemical analysis was performed to investigate the effects of hydrodynamic ozonation on the surface of gutta-percha cones. Fourier-transform infrared (FTIR) spectra were acquired in triplicate under three experimental conditions: (1) untreated gutta-percha cones (*in natura*), (2) gutta-percha cones exposed to ozonated water for 7 min using a hydrodynamic ozonation system, and (3) ozonated gutta-percha cones after mechanical removal of a thin superficial layer using a sterile scalpel blade (scraping) to expose the underlying material.

FTIR analyses were carried out using a Spectrum Frontier spectrometer (PerkinElmer, Waltham, MA, USA) equipped with a universal attenuated total reflectance (UATR) accessory. Spectra were collected over the range of 4000–550 cm^{-1} using 20 accumulated scans and a gain setting of 1. All measurements were performed under controlled laboratory conditions at 21 °C and 48% relative humidity.

The spectra obtained from the three experimental groups were qualitatively compared to identify possible chemical modifications induced by ozone exposure and to determine whether these changes were confined to the surface or extended into the internal structure of the gutta-percha.

E. Data analysis

The FTIR spectra obtained from the three experimental groups were qualitatively analyzed by comparing the position, intensity, and shape of the characteristic absorption bands associated with the chemical constituents of gutta-percha. Special attention was given to spectral regions related to hydroxyl groups, carbon-carbon double bonds, and the trans-1,4-polyisoprene structure to identify possible chemical modifications induced by ozone exposure.

The spectra of untreated gutta-percha, ozonated gutta-percha, and ozonated gutta-percha after superficial layer removal were compared to determine whether the observed oxidative changes were restricted to the material surface or extended into the underlying layers.

III. Results

A. Structural Characterization of Dental Gutta-Percha by FTIR/UATR

The FTIR/UATR spectra of untreated gutta-percha (A), ozonated gutta-percha (B), and ozonated gutta-percha after superficial scraping (C) are presented in Figure 3. All experimental groups exhibited the characteristic absorption bands of gutta-percha, confirming the preservation of the material's chemical composition after treatment. The main bands were observed at approximately 3444 cm^{-1} (O–H stretching), 2962, 2913, and 2848 cm^{-1} (C–H stretching), 1667 cm^{-1} (C=C stretching of the polyisoprene backbone), 1481, 1418, and 1381 cm^{-1} (CH_2 and CH_3 bending vibrations), 1173, 1117, 1074, and 1016 cm^{-1} (C–C and C–O stretching), and 795 cm^{-1}, which is associated with the trans-1,4-polyisoprene structure.

Qualitative comparison of the spectra revealed chemical changes after exposure to the hydrodynamic ozonation system. The ozonated specimens showed a broader absorption band in the hydroxyl stretching region (approximately 3390–3445 cm^{-1}), suggesting the formation of oxygen-containing functional groups on the material surface. In addition, broadening of the band centered at approximately 1667 cm^{-1} was observed, indicating oxidation of the unsaturated polyisoprene chains and increased interaction with adsorbed water. A reduction in the intensity of the absorption band near 795 cm^{-1} was also detected, consistent with ozone-induced consumption of carbon–carbon double bonds in the trans-1,4-polyisoprene matrix.

After mechanical removal of a thin superficial layer, the FTIR spectrum became comparable to that of untreated gutta-percha. The broadening of the hydroxyl band, the changes observed around 1667 cm^{-1}, and the reduction in the 795 cm^{-1} band were no longer evident, indicating that the chemical modifications induced by hydrodynamic ozonation were predominantly restricted to the outer surface of the material, with no detectable alterations in the underlying polymeric structure.

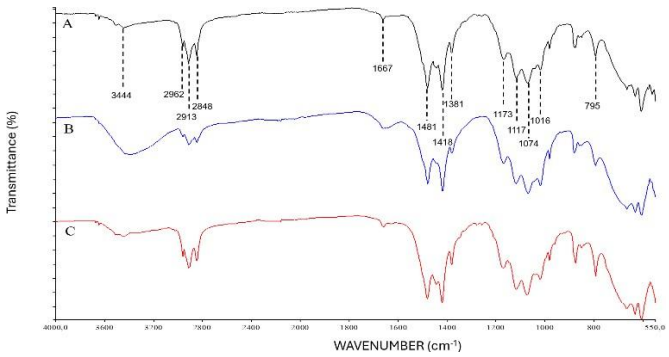


Fig 3 Representative FTIR/UATR spectra of untreated gutta-percha (A), ozonated gutta-percha (B), and ozonated gutta-percha after superficial scraping (C).

IV. Discussion

The FTIR/UATR analysis demonstrated that exposure of gutta-percha to hydrodynamic ozonated water induced localized chemical modifications in the superficial region of the material, as evidenced by the broadening of the hydroxyl stretching band (3390–3445 cm^{-1}), changes in the spectral profile around 1667 cm^{-1}, and a reduction in the intensity of the band near 795 cm^{-1}. These findings are consistent with the ozonolysis mechanism of unsaturated polymers, in which ozone preferentially reacts with carbon–carbon double bonds (C=C), initially forming unstable ozonides that subsequently decompose into oxygen-containing products, including hydroxyl, carbonyl, and hydroperoxide groups [20].

Although dental gutta-percha is composed predominantly of trans-1,4-polyisoprene, whereas natural rubber consists mainly of cis-1,4-polyisoprene, both materials contain unsaturated polymeric chains with carbon–carbon double bonds that are susceptible to oxidative attack by ozone. Therefore, the oxidation mechanism described for natural rubber can be considered an appropriate chemical model for interpreting the spectroscopic changes observed in gutta-percha. Kim et al. [20] demonstrated that ozone exposure of polyisoprene results in progressive changes in spectral bands associated with C=C bonds and the formation of oxygen-containing functional groups, supporting the interpretation of the present findings. Similarly, Rodrigues et al. [21] reported ozone-induced surface effects in natural rubber associated with the oxidation of unsaturated polymeric structures.

In the present study, the broadening of the band in the 3390–3445 cm^{-1} region suggests the formation or increased contribution of oxygen-containing groups following ozonation, whereas the changes around 1667 cm^{-1} are consistent with modifications involving unsaturated regions of the polyisoprene matrix. Likewise, the reduced intensity of the band near 795 cm^{-1} is compatible with changes involving the trans-1,4-polyisoprene structure and the consumption or modification of carbon–carbon double bonds during ozonolysis. However, these interpretations should be considered qualitative because the present FTIR/UATR analysis was based on comparison of band position, shape, and relative intensity rather than on quantitative determination of oxidation products or functional-group concentrations.

An important aspect of the present findings is the comparison between ozonated gutta-percha before and after mechanical removal of a superficial layer. Following superficial scraping, the spectral changes observed after ozonation were no longer detectable, and the resulting spectrum became comparable to that of untreated gutta-percha. This finding suggests that the chemical modifications detected by FTIR/UATR were predominantly associated with the superficial region of the material under the conditions evaluated. However, the present experimental design does not allow the exact depth of oxidation to be quantitatively determined. In particular, the thickness of the mechanically removed layer was not measured, and the FTIR/UATR analysis was not performed as a depth-resolved or spatially resolved technique. Therefore, the disappearance of the spectral changes after scraping should be interpreted as evidence consistent with a predominantly superficial modification rather than as definitive confirmation of a specific oxidation depth.

This limitation is relevant because FTIR/UATR provides information about the chemical composition within the region sampled by the attenuated total reflectance interaction, while the mechanical scraping procedure provides a qualitative comparison between the exposed surface and the underlying material. Complementary techniques could provide a more precise characterization of the oxidation process. For example, X-ray photoelectron spectroscopy (XPS) could be used to quantitatively evaluate changes in surface elemental composition and carbon–oxygen chemical states, whereas depth-resolved micro-FTIR or other spatially resolved spectroscopic approaches could provide additional information regarding the distribution of chemical modifications as a function of depth. Raman spectroscopy could also complement FTIR by providing additional information regarding changes in the polyisoprene structure. Therefore, future studies combining surface-sensitive and depth-resolved techniques would be important to quantify both the degree and spatial extent of ozone-induced oxidation.

From an endodontic perspective, the superficial nature of the detected modifications is relevant because gutta-percha is used in combination with an endodontic sealer to achieve an adequate root canal filling. Surface characteristics of gutta-percha, including chemical composition, surface energy, roughness, and wettability, may influence its interaction with endodontic sealers. Previous studies have demonstrated that chemical surface modification of gutta-percha can alter surface free energy and sealer wettability, with effects on bond strength depending on the treatment and the type of sealer evaluated [22,23]. In addition, different endodontic sealers may exhibit substantially different adhesion to gutta-percha [24]. Therefore, the superficial oxidation observed in the present study could potentially modify the interaction between ozonated gutta-percha and endodontic sealers. Nevertheless, the direction and clinical relevance of this effect cannot be established from the present data because adhesion, wettability, interfacial bonding, microleakage, and obturation performance were not directly evaluated.

The preservation of the underlying polymeric structure is also relevant to the potential application of hydrodynamic ozonation as a gutta-percha disinfection strategy. Previous work using the same hydrodynamic ozonation approach demonstrated effective disinfection of Enterococcus faecalis-contaminated gutta-percha while preserving mechanical and morphological characteristics after treatment [19]. The present findings add a chemical perspective to these observations by indicating that detectable FTIR/UATR modifications were associated predominantly with the superficial region rather than with the underlying polymeric structure. Nevertheless, chemical preservation of the bulk material should not be interpreted as evidence of complete preservation of long-term clinical performance, since aging, repeated exposure, interaction with endodontic sealers, and storage conditions may influence the behavior of the material over time.

Taken together, the main lesson learned from this study is that the evaluation of a disinfection method for gutta-percha should consider not only its antimicrobial effectiveness but also the extent and nature of the chemical modifications induced in the biomaterial. Under the conditions evaluated, hydrodynamic ozonation produced detectable spectral changes that were no longer observed after superficial material removal, suggesting that the chemical effects were predominantly associated with the surface region. At the same time, the qualitative nature of the FTIR/UATR analysis highlights the need for caution when interpreting the degree and depth of oxidation.

The findings also identify relevant research gaps. Future investigations should quantitatively determine the degree and depth of ozone-induced oxidation, evaluate changes in surface energy and wettability, and investigate the interaction of

ozonated gutta-percha with different classes of endodontic sealers. Bond-strength, interfacial adaptation, microleakage, and long-term aging studies would be particularly important to establish whether the superficial chemical modifications observed in the present study have beneficial, neutral, or detrimental effects on obturation performance. Such investigations would allow the chemical findings obtained here to be more directly correlated with the functional and clinical behavior of ozonated gutta-percha.

V. Conclusion

The present study demonstrated that hydrodynamic ozonated water induced detectable chemical modifications in the superficial region of gutta-percha under the conditions evaluated. FTIR/UATR analysis revealed spectral changes associated with ozone-induced modifications of the polyisoprene matrix, including changes in the hydroxyl stretching region, the band around 1667 cm^{-1}, and the band near 795 cm^{-1}. After mechanical removal of a superficial layer, these spectral changes were no longer detectable, suggesting that the chemical modifications were predominantly associated with the superficial region of the material.

However, the qualitative nature of the FTIR/UATR analysis and the absence of a quantitative depth-resolved assessment limit the determination of the exact degree and depth of ozone-induced oxidation. Therefore, complementary surface-sensitive and depth-resolved techniques are needed to further characterize these modifications. In addition, further studies should investigate whether the observed superficial chemical changes influence the interaction of gutta-percha with endodontic sealers, as well as long-term obturation performance.

Overall, the findings support the potential of hydrodynamic ozonated water as a gutta-percha disinfection strategy while indicating that its chemical effects should be further characterized to establish their functional and clinical relevance.

Conflict of Interest

The authors declare that they have no conflicts of interest regarding the publication of this manuscript.

Acknowledgment

J.S.C.S. acknowledge CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior□Brazil), Financing Code 001. A.B.F. and C.J.de.L. acknowledge the Anima Institute (AI) Universidade Anhembi Morumbi São Paulo-SP, Brazil. A. Barrinha Fernandes acknowledges CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the produc tivity fellowship (Process No. 310708/2021-4). The authors express their sincere gratitude to Milton Faria Diniz, from the Institute of Aeronautics and Space (IAE), Department of Aerospace Science and Technology (DCTA), for his valuable assistance and support in the acquisition of the infrared spectra.

REFERENCES

[1] Siqueira JF Jr, Rôças IN. Diversity of endodontic microbiota revisited. J Dent Res. 2009 Nov;88(11):969-81. doi: 10.1177/0022034509346549. PMID: 19828883.

[2] Ricucci D, Siqueira JF Jr. Biofilms and apical periodontitis: study of prevalence and association with clinical and histopathologic findings. J Endod. 2010 Aug;36(8):1277-88. doi: 10.1016/j.joen.2010.04.007. Epub 2010 Jun 14. PMID: 20647081.

[3] Narayanan LL, Vaishnavi C. Endodontic microbiology. J Conserv Dent. 2010 Oct;13(4):233-9. doi: 10.4103/0972-0707.73386. PMID: 21217951; PMCID: PMC3010028.

[4] Nair PN. On the causes of persistent apical periodontitis: a review. Int Endod J. 2006 Apr;39(4):249-81. doi: 10.1111/j.1365-2591.2006.01099.x. PMID: 16584489.

[5] Siqueira JF Jr. Aetiology of root canal treatment failure: why well-treated teeth can fail. Int Endod J. 2001 Jan;34(1):1-10. doi: 10.1046/j.1365-2591.2001.00396.x. PMID: 11307374.

[6] Ng YL, Mann V, Gulabivala K. A prospective study of the factors affecting outcomes of nonsurgical root canal treatment: part 1: periapical health. Int Endod J. 2011 Jul;44(7):583-609. doi: 10.1111/j.1365-2591.2011.01872.x. Epub 2011 Mar 2. PMID: 21366626.

[7] Siqueira JF Jr, Rôças IN. Clinical implications and microbiology of bacterial persistence after treatment procedures. J Endod. 2008 Nov;34(11):1291-1301.e3. doi: 10.1016/j.joen.2008.07.028. Epub 2008 Sep 17. PMID: 18928835.

[8] Ricucci D, Siqueira JF Jr. Recurrent apical periodontitis and late endodontic treatment failure related to coronal leakage: a case report. J Endod. 2011 Aug;37(8):1171-5. doi: 10.1016/j.joen.2011.05.025. PMID: 21763916.

[9] Vishwanath V, Rao HM. Gutta-percha in endodontics - A comprehensive review of material science. J Conserv Dent. 2019 May-Jun;22(3):216-222. doi: 10.4103/JCD.JCD_420_18. PMID: 31367101; PMCID: PMC6632621.

[10] Dobrzańska, J.; Dobrzański, L.B.; Dobrzański, L.A.; Gołombek, K.; Dobrzańska-Danikiewicz, A.D. Is Gutta-Percha Still the "Gold Standard" among Filling Materials in Endodontic Treatment? Processes 2021, 9, 1467. https://doi.org/10.3390/pr9081467

[11] Gasner NS, Brizuela M. Endodontic Materials Used To Fill Root Canals. [Updated 2023 Mar 19]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK587367/

[12] Çakir A. Effect of cleaning materials on microtensile bond strength of resin composite to primary dentin contaminated with root canal sealers. BMC Oral Health. 2023 Jun 16;23(1):396. doi: 10.1186/s12903-023-03090-z. PMID: 37328843; PMCID: PMC10273635.

[13] Gurgel-Filho ED, Andrade Feitosa JP, Teixeira FB, Monteiro de Paula RC, Araújo Silva JB, Souza-Filho FJ. Chemical and X-ray analyses of five brands of dental gutta-percha cone. Int Endod J. 2003 Apr;36(4):302-7. doi: 10.1046/j.1365-2591.2003.00653.x. PMID: 12702126.

[14] Huang PC, Wu YT, Hsu YH, Liao SC, Wang YH, Gutmann JL, Huang HM, Hsieh SC. The investigation of composition and thermal behavior of two types of backfilling gutta-percha. J Dent Sci. 2023 Jan;18(1):414-419. doi: 10.1016/j.jds.2022.08.014. Epub 2022 Sep 2. PMID: 36643230; PMCID: PMC9831847.

[15] Marciano MA, Ordinola-Zapata R, Cunha TV, Duarte MA, Cavenago BC, Garcia RB, Bramante CM, Bernardineli N, Moraes IG. Analysis of four gutta-percha techniques used to fill mesial root canals of mandibular molars. Int Endod J. 2011 Apr;44(4):321-9. doi: 10.1111/j.1365-2591.2010.01832.x. Epub 2011 Jan 10. PMID: 21219361.

[16] Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, Fielden PR, Fogarty SW, Fullwood NJ, Heys KA, Hughes C, Lasch P, Martin-Hirsch PL, Obinaju B, Sockalingum GD, Sulé-Suso J, Strong RJ, Walsh MJ, Wood BR, Gardner P, Martin FL. Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc. 2014 Aug;9(8):1771-91. doi: 10.1038/nprot.2014.110. Epub 2014 Jul 3. PMID: 24992094; PMCID: PMC4480339.

[17] Al-Kelani M, Buthelezi N. Advancements in medical research: Exploring Fourier Transform Infrared (FTIR) spectroscopy for tissue, cell, and hair sample analysis. Skin Res Technol. 2024 Jun;30(6):e13733. doi: 10.1111/srt.13733. PMID: 38887131; PMCID: PMC11182784.

[18] BROWN, Michael E. (Ed.). Introduction to thermal analysis: techniques and applications. Dordrecht: Springer Netherlands, 2001.

[19] Souza JSC, Laurindo Igreja Marrafa PA, de Carvalho MCO, Rodrigues MF, Akemi Kawata B, Martins GV, Ribeiro Dos Anjos EG, Passador FR, Fernandes AB, de Lima CJ. Evaluation of a Hydrodynamic Ozonation System for Gutta-Percha Disinfection and Structural Characterization. ACS Omega. 2026 Feb 9;11(7):11470-11481. doi: 10.1021/acsomega.5c08886. PMID: 41768645; PMCID: PMC12947183.

[20] KIM, Ik-Sik et al. Characterization on the thermal oxidation of raw natural rubber thin film using image and FT-IR analysis. Elastomers and composites, v. 55, n. 1, p. 51-58, 2020.

[21] RODRIGUES, F. H. A. et al. Ozonation of unstretched natural rubber: Part i. effect of film thickness. Rubber chemistry and technology, v. 74, n. 1, p. 57-68, 2001

Gelatin photo-crosslinking with methylene blue and red laser: a comparative pilot study with other crosslinking agents

Patricia da Silva Souza
Postgraduate Program on Biomedical Engineering
Federal University of ABC
Sao Bernardo do Campo, Brazil
patricia.souza@aluno.ufabc.edu.br

Sonia Maria Malmonge
Center for Engineering, Modeling and Applied Social Sciences
Federal University of ABC
Sao Bernardo do Campo, Brazil
sonia.malmonge@ufabc.edu.br

Nasser Ali Daghastanli
Center for Engineering, Modeling and Applied Social Sciences
Federal University of ABC
Sao Bernardo do Campo, Brazil
nasser.daghastanli@ufabc.edu.br

Patricia A.Ana
Center for Engineering, Modeling and Applied Social Sciences
Federal University of ABC
Sao Bernardo do Campo, Brazil
patricia.ana@ufabc.edu.br

***Abstract*— The purpose of this study was to investigate the potential of methylene blue with red laser light as a photo crosslinking method in type B gelatin films, compared with the effect of UVC radiation and EDC. Preliminary results with methylene blue showed higher crosslinking capacity compared to UVC.**

Keyword - methylene blue, crosslinking, ultraviolet radiation, photodynamic therapy.

I. INTRODUCTION

Collagen is the most abundant fibrous protein in animal tissues and the primary protein that makes up the extracellular matrix (ECM). In the extracellular matrix, collagen fibers contribute to its tensile strength and structural organization [1]. Derived from the hydrolysis and irreversible denaturation of collagen, gelatin is a linear polymer with a high degree of similarity to the amino acids that make up collagen; it can be sourced from the skin of mammals, primarily pigs and cattle. For extraction purposes, gelatin can be treated with acidic or alkaline agents and is classified accordingly as Type A or Type B [2].

In the field of tissue engineering, gelatin is considered a biocompatible and biodegradable material, and its versatility as a biomaterial for scaffold formation is one of advantages over collagen, as its lower cost and higher solubility in a variety of solvents [3]. However, gelatin has limitations in its mechanical properties, with a short degradation rate and low thermal stability [4]. To overcome this inherent characteristic, crosslinking methods aimed at promoting intra- and intermolecular bonds have been proposed to increase rigidity, enhance stability, increase molecular weight, and improve mechanical properties; this approach has been considered the most commonly explored strategy for addressing the limitations of this biomaterial [5].

Crosslinking processes are commonly carried out using chemical agents, which are designed to interact with functional groups in the collagen structure, promoting the formation of crosslinks. However, despite their crosslinking potential, chemical agents may exhibit high toxicity and low biocompatibility. Among traditional chemical agents, N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride, known as EDC, is recognized for its ability to form isopeptide bonds between carboxyl and amine groups, exhibiting high aqueous solubility and lower toxicity compared to other aldehyde-based chemical agents, such as glutaraldehyde [6].

Physical crosslinking methods using ultraviolet (UV) radiation are known to promote crosslinking through the formation of chemical bonds via the aromatic groups of tyrosine and phenylalanine. Despite potential sterilizing benefits, the use of UV as a crosslinking agent depends on its dosimetry, such as fluency, exposition time and power. Currently, the use of UV as a crosslinker has been best evaluated in combination with photosensitive chemical agents [7].

Methylene Blue (MB) is a photosensitive dye belonging to the class of phenothiazine compounds, with a broad absorption band in the 550–700 nm range, Fig.1, whose monomers exhibits maximum absorption peaks at a wavelength of 664 nm and dimers exhibits maximum absorption peaks at 590 nm [8]. MB is a recognized photosensitizing agent in Photodynamic Therapy (PDT). Photodynamic Therapy is a non-invasive therapeutic technique defined by the combination of a light source, a photosensitizing agent, and the presence of molecular oxygen. The interaction of these components generates photophysical and photochemical processes that are responsible for the onset of biological effects at the site of application. The photosensitizing agent is activated by a light source with a wavelength that is absorbed by the photosensitizer; this specificity is intended to trigger its photosensitive reactions, which occur due to the presence of molecular oxygen and may result in the production of reactive oxygen species (ROS-Reactive Oxygen Species), capable of interacting locally [9].

Reactive oxygen species, such as singlet oxygen, are already known in the literature to react with the carbonyl groups of collagen and promote new bonds between its amino acids, increasing the formation of covalent crosslinks. This process is used, for example, for corneal crosslinking through exposure to ultraviolet radiation in combination with vitamin B2 [10]. UV light is also widely used in the photo-crosslinking of gelatin methacryloyl hydrogels (GelMA) with an aqueous photoinitiator [11]. However, due to light scattering in tissues and depending on its wavelength, ultraviolet radiation has a

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

limited depth of tissue penetration, ranging from 0.5 to 2.5 mm in healthy skin tissues [12].

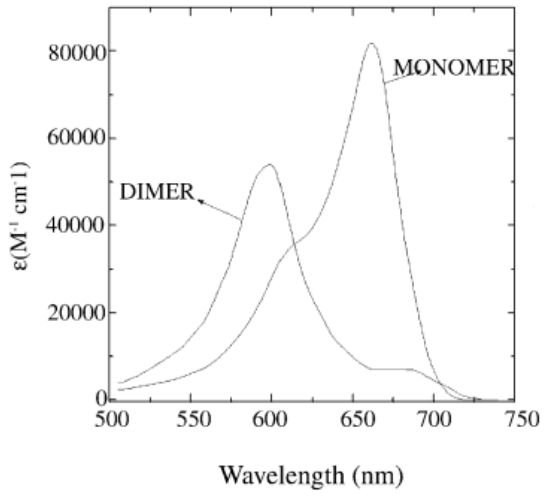


Fig.1. Absorbance spectra of dimer and monomer species of MB [8].

Considering that higher wavelengths are generally safer for *in vivo* applications compared to lower wavelengths, given the direct correlation between photon energy and the phototoxicity of the light source [13], and that red light has a penetration depth in soft tissues of 1 to 3 mm - greater than of UV light [14], the combination of red light with methylene blue - whose maximum absorption lies within the therapeutic window of the red laser [15] - may be an innovative approach for performing photo-crosslinking in deeper tissues.

Based on the premise that new crosslinking processes may contribute to the development of innovative collagen-based materials for biomedical applications [7], this study aims to understand the potential of photodynamic therapy with methylene blue and red laser as a new crosslinking process in type B gelatin films, in comparison with other traditional crosslinking methods.

II. Material and Methods

All reagents were purchased from Sigma-Aldrich (USA), including: N-(3-dimethylaminopropyl)-N-ethylcarbodiimide hydrochloride (EDC), methylene blue (MB), 2,2,2-trifluoroethanol (TFE), and type B bovine skin gelatin.

A. Preparation of gelatin films

A.1 Preparation of pure type-B gelatin films

Five grams of type B gelatin were mixed in 100 mL of 2,2,2-trifluoroethanol, subjected to continuous mechanical stirring at 500 rpm for 24 h. From the resulting solution, 20 mL of the monotonic solution were added and poured into 10 cm diameter Petri dishes. The plates were placed in an exhaust hood for 24 h and then the thickness of the films was checked with a micrometer and their masses on a Shimadzu -ATX 224 analytical balance.

A.2 Preparation of type-B gelatin film with MB

2 mL of 0.05% MB was added to the surfaces of the type B gelatin films, Fig.2, with a pasteur pipet and the solution was left to immersion for 5 or 7 minutes. After that, samples were removed and kept in dark until the red laser irradiations.

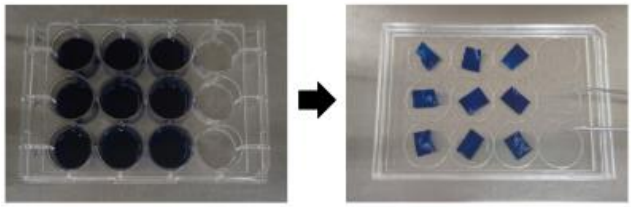


Fig.2. Immersion process of type B gelatin films in MB (0.05%).

B. Crosslinking of type-B gelatin film

B.1 Crosslinking of pure gelatin using EDC

EDC (148 g/mol) was used to prepare EDC-crosslinked type B gelatin films at a gel/EDC molar ratio of 1:1 by immersing the films in a solution of distilled water with EDC for 24 hours. Subsequently, the films were wrapped in greaseproof paper and left to dry in an exhaust hood for 24 hours.

B.2 Crosslinking of pure gelatin with UVC radiation

The crosslinking of twenty 4.7 x 1 cm gelatin samples was performed using a LucMat Lamp 20W/GL UVC (254 nm). For that, the samples were positioned in a support within a laminar hood flow, at 10 cm distance. The irradiation of the samples followed the time and dose parameters shown in Table 1.

TABLE I. Irradiation Parameters (UVC)

Sample	***UVC exposure***	***Dose (J/cm²)***
Type-B pure gelatin	10 min	0,0259
	30 min	0,0779

B.3 Crosslinking of MB-incorporated gelatin by red laser

The crosslinking of thirty 1,5 x 1,5 cm MB-incorporated gelatin films was performed using a red Laser (OrthoLase, DMC, Brazil, 660nm). For that, samples were placed in Petri dishes within a laminar hood flow, and the irradiations were performed according to the Table II.

TABLE II. Irradiation Parameters (Red Laser)

Sample	***Immersion time in the MB***	***Irradiation time with red laser***	***Dose (J/cm²)***
Type-B pure gelatin	5 min	10 min	0,7639
	7 min	15 min	1,1459
	7 min	20 min	1,5278

C. *Film Characterization*

C.1 Water Absorption Percentage

Three 1.5 × 1.5 cm samples from each group, were weighed to determine their initial mass (M0) and then immersed in distilled water for specific periods (0, 2, 4, 6, 8,

10, 15, 20, 25, 30, 40, 50, and 60 min). The samples were then dried, and their dry mass (Mt) was weighed again [16]. The water absorption rate was calculated using Equation 1.

$$Water\ Absorption\ Percentage = Mt - M0 / M0 \times 100\% \quad (1)$$

C.2 Mechanical Properties

Tensile testing was performed on a Tytron™ 250 Universal Testing Machine and TestWorks 4 software to obtain Young's Modulus, tensile strength, and elongation at break. An initial test length of 30 mm was used on a 250 N load cell at a speed of 5 mm.min-1.

III. Results and Discussion

The results of the water absorption indicated that the MB-incorporated gelatin films irradiated with red laser, Fig. 3, for an extended period exhibited greater resistance to water absorption. The 7-minute immersion period in the MB, followed by 20 minutes of red light, proved to be the most effective treatment for strengthening the gelatin film's bonds, as evidenced by the film's reduced water retention due to its stiffening. Gelatin exhibits low stiffness and unstable mechanical behavior in the presence of water [17]. The aqueous absorption test on gelatin films is recognized for contributing to the understanding of gelatin's polymeric behavior in an aqueous environment and the stability of biomaterials [18]. Pure gelatin contains hydrophilic amino acids with amino and hydroxyl groups in its structure, which in aqueous solutions tend to form hydrogen bonds with water molecules, promoting greater water incorporation into the polymer chains. The use of cross-linking agents in gelatin tends to reduce swelling by promoting covalent bonds that restrict polymer expansion [19]. These preliminary results obtained with MB and red light demonstrated the ability to alter the structural behavior of the gelatin film by reducing its swelling. Further chemical analyses should be conducted in this study to better understand the potential of MB's cross-linking mechanism on gelatin films.

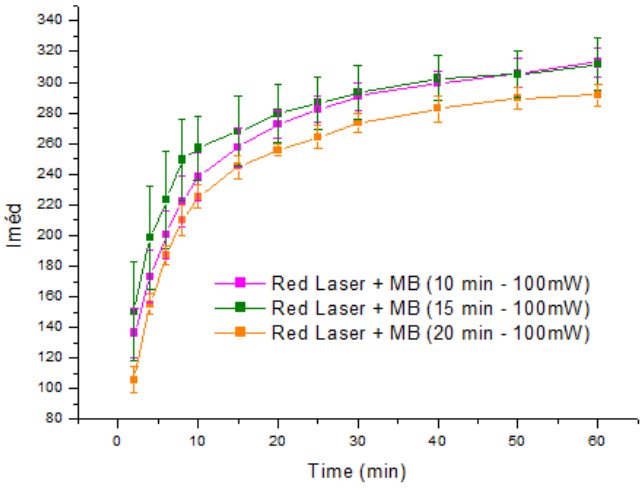


Fig.3. Percentage of water absorption for MB-incorporated gelatin after red laser irradiation. Bars indicate standard deviation.

However, to better understand these results regarding the percentage of water absorption, the degree of water absorption of type B gelatin films subjected to other traditional cross-linking methods was compared, Fig. 4. The results shown in Fig. 4 reveal that pure gelatin films exposed to UVC for 10 min and 30 min exhibited lower water repellency compared to MB incorporated gelatin films after red laser irradiation. While chemical crosslinking with EDC traditionally yields strong covalent bonds [7], PDT-treated films exhibited superior performance compared to UVC-crosslinked films.

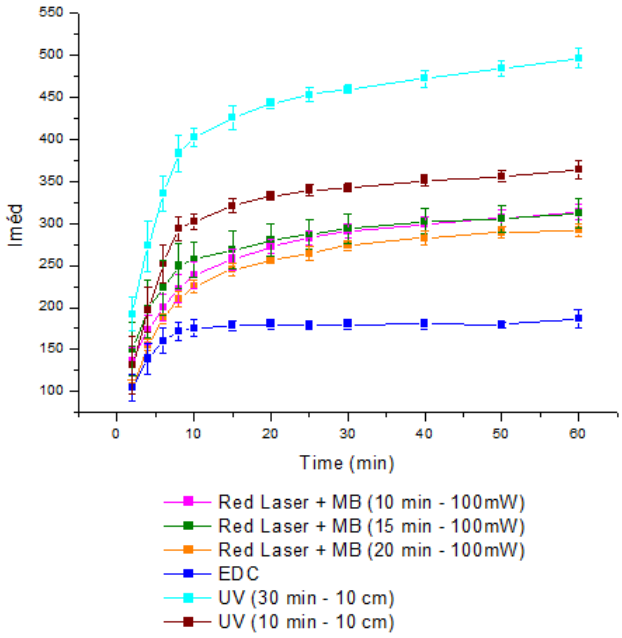


Fig. 4. Percentage of water absorption for all samples of this study. Bars indicate standard deviation.

Mechanical evaluation showed reduced stiffness in MB-incorporated samples following 20 minutes of irradiation, reflected by their elongation at break (Table III). Following this trend, the group crosslinked with EDC demonstrated the lowest stiffness. On the other hand, 30-minute UVC irradiation resulted in the highest elongation at break, showing a significant increase compared to neat gelatin samples.

Despite significant variations in stiffness, EDC-treated neat gelatin displayed significantly higher Young's modulus and tensile strength than any other experimental group. Mechanical analysis indicates that each crosslinking agent imparts distinct physical properties to gelatin, potentially enabling tailored clinical applications for the same biomaterial.

Literature demonstrates that the chemical and mechanical properties of gelatin hydrogels can be tailored for diverse applications. For instance, low-stiffness hydrogels-offering high flexibility- are ideal for injectable formulations, making them suitable for controlled drug delivery and soft tissue regeneration, such as skin and vascular tissues. Conversely, hydrogels with high mechanical strength are better suited for hard tissue engineering, including bone and dental applications.

Moreover, methylene blue-mediated photodynamic therapy using red laser light is well-established in clinical practice. These results highlight the potential for producing gelatin hydrogels with versatile applications by adjusting the light dose according to the intended clinical outcome.

TABLE III. MECHANICAL PROPERTIES OF ALL FILMS TESTED.

Groups	***Thickness (mm)***		***Young's modulus (MPa)***		***Tensile strength (MPa)***		***Elongation at break (%)***	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pure gelatin	0.12	0	1.04	0.73	0.16	0.10	48.46	37.57
Pure gelatin with EDC	0.16	0.03	7.17	4.09	2.25	1.60	30.93	17.79
Pure gelatin with 10 min UVC	0.13	0.01	1.11	0.14	0.31	0.14	36.26	11.24
Pure gelatin with 30 min UVC	0.14	0.01	0.97	0.08	0.65	0.36	84.20	48.61
Gelatin + MB and Red Laser (10 min)	0.13	0	0.49	0.09	0.19	0.07	44.93	23.25
Gelatin + MB and Red Laser (15 min)	0.13	0	0.58	0.02	0.22	0.21	40.53	34.86
Gelatin + MB and Red Laser (20 min)	0.13	0	0.78	0.41	0.10	0.04	19.50	3.41

The preliminary results presented in the characterization tests of type B gelatin films containing methylene blue and irradiated with a red laser suggest that photodynamic therapy using methylene blue and red laser has potential as a new cross-linking approach, that represents a highly adaptable methodology tailored to the researcher's requirements, with the added benefits of low cost, low toxicity and broad availability. However, tests to chemically understand and evaluate the structure of gelatin films must be conducted, in addition to cytotoxicity tests, to support further discussion. This study supports the view that methylene blue - an agent already widely used in antimicrobial photodynamic therapy and considered safe for in vivo applications [14] should also be further explored for the cross-linking of gelatin with a view toward tissue engineering in the biomedical field.

IV. CONCLUSION

This study revealed the promising potential of photo-crosslinking method for type B gelatin films. Further tests should be conducted to better understand the specifics of using methylene blue in the crosslinking process.

ACKNOWLEDGMENT

To INCT-INTERAS and Multiuser Central of UFABC.

REFERENCES

[1] B. Alberts *et al*. "Biologia Molecular da Célula". Porto Alegre: Artmed, 6ª edição, 2017.

[2] M. Hassan, D. Hussain, T. Kanwal, H.M Xiao, S.H. Musharraf. "Methods for detection and quantification of gelatin from different sources". Food Chemistry, 2024.

[3] M. Usman et al. "Valorization of animal by-products for gelatin extraction using conventional and green technologies: A comprehensive review". Biomass Conversion and Biorefinery, 2023.

[4] V. Patil, K. Masters. "Engineered Collagen Matrices". Bioengineering, v.7, 2020.

[5] J. V. Hoorick et al. "(Photo-)crosslinkable gelatin derivatives for biofabrication applications". Acta Biomaterialia, 2019.

[6] A. Sionkowska, K. Kulka-Kamińska, P. Brudzyńska, K. Lewandowska, L. Piwowarski. "The Influence of Various Crosslinking Conditions of EDC/NHS on the Properties of Fish Collagen Film". Mar. Drugs, 2024.

[7] K. Adamiak, A. Sionkowska. "Current methods of collagen cross-linking: Review". International Journal of Biological Macromolecules, 2020.

[8] J.P. Tardivo et al. "Methylene blue in photodynamic therapy: From basic mechanisms to clinical applications". Photodiagnosis and Photodynamic Therapy, 2005.

[9] S. Kwiatkowskia et al. "Photodynamic therapy – mechanisms, photosensitizers and combinations". Biomedicine & Pharmacotherapy, 2018.

[10] A. Naranjo, E. E. Manche. "A comprehensive review on corneal crosslinking". Taiwan J. Ophthalmol, 2024.

[11] K. Yue et al. "Synthesis, properties, and biomedical applications of gelatin methacryloyl (GelMA) hydrogels". Biomaterials, 2015.

[12] B. Sun, J. N. Rahmat, Y. Zhang. "Advanced techniques for performing photodynamic therapy in deep-seated tissues". Biomaterials, 2022.

[13] S. Sharifi, H. Sharifi, A. Akbari, J. Chodosh. "Systematic optimization of visible light-induced crosslinking conditions of gelatin methacryloyl (GelMA)". Scientific Reports, 2021.

[14] H. Kolarova et al. "Methylene blue-enhanced photodynamic and sonodynamic therapy against Staphylococcus aureus: from laboratory research to clinical evaluation". Photodiagnosis and Photodynamic therapy, 2026.

[15] A. C. Sousa e Sousa et al. "The application of red laser associated with methylene blue promotes better healing outcomes in nipple fissures: a hypothesis study". Medical Hypotheses, 2026.

[16] J. S. Haima, S. N. Nair, S. Juliet, A. Ravindran. "Synthesis and characterisation of glutaraldehyde cross-linked k-carrageenan-gelatin hydrogel". Journal of Pharmacognosy and Phytochemistry, 2021.

[17] V. Manish, A. Arockiarajan, G. Tamadapu. "Influence of water content on the mechanical behavior of gelatin based hydrogels: syntheis, characterization and modeling". International Journal of Solids and Structures, 2021.

[18] Q. Xing et al. "Increasing Mechanical Strength of Gelatin Hydrogels by Divalent Metal Ion Removal". Scientific Reports, 2014.

[19] J. A. Rather et al. "A comprehensive review on gelatin: Understanding impact of the sources, extraction methods, and modifications on potential packaging applications". Food Packaging and Shelf Life, 2022.

Hanging Drop Microdroplet Culture: A Promising Approach for Biomaterial Testing

Frederico Ercolin Guazzelli
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC
São Bernardo do Campo, Brazil
fredercolin@gmail.com

Felipe Ambrosio
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC
São Bernardo do Campo, Brazil
felipe.ambrosio@gmail.com

João Lameu da Silva Júnior
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC
São Bernardo do Campo, Brazil
joao.lameu@ufabc.edu.br

Christiane Bertachini Lombello
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC
São Bernardo do Campo, Brazil
christiane.lombello@ufabc.edu.br

Abstract*—Vero cells, derived from African green monkey kidney epithelial cells and exhibiting a fibroblast-like morphology, these cells are widely used in biological research due to their stability and susceptibility to viral infection. These characteristics make them a valuable model for *in vitro* studies, including cytotoxicity and drug screening assays. Traditional two-dimensional (2D) cell culture systems, although well established, often fail to accurately reproduce the complexity of *in vivo* environments. In contrast, three-dimensional (3D) culture models, such as spheroids, better mimic cell–cell interactions, extracellular matrix deposition and physiological conditions. Among emerging 3D culture techniques, the hanging drop method, a microdroplet-based approach, enables the formation of scaffold-free cellular spheroids in a controlled microenvironment. This technique has gained attention for its simplicity, reproducibility, and ability to generate more physiologically relevant models for *in vitro* assays, including biomaterial testing. Furthermore, microdroplet-based systems show promising potential for integration with advanced platforms such as organ-on-a-chip technologies, which aim to replicate tissue-level functions *in vitro*. These combined approaches may, in the future, enhance the predictive accuracy of cytotoxicity and pharmacological studies. This study aims to obtain and characterize Vero cell spheroids using the hanging drop method. It enables the evaluation of their interaction with biological systems, including cytotoxicity and biological evaluation assays, expanding the potential application in biomaterial assessment, contributing to the advancement of 3D *in vitro* models in biomedical research. The results demonstrated the formation of compact and structurally stable Vero cell spheroids after 7 and 21 days of culture, with preserved morphology and nuclear integrity after post-processing.

Keywords— Vero cells, cell culture, In Vitro, Hanging Drop, Organ-on-a-chip.

I. INTRODUCTION

Vero cells are a continuous cell lineage derived from the kidney epithelial cells of the African green monkey (*Cercopithecus aethiops*) (ATCC CCL-81). Due to their robustness, ease of maintenance in culture, and high susceptibility to viral infection, these cells are widely used in biomedical research, particularly in virology, toxicology and vaccine development studies [1], [2]. Their permissiveness to viral replication and adaptability to different experimental conditions make them a reliable in vitro model for cellular and molecular investigations and an attractive candidate for the development of advanced three-dimensional culture systems.

Cell culture techniques play a fundamental role in experimental biology by enabling controlled investigation of cellular behavior under defined conditions [3]. Conventional two-dimensional (2D) monolayer cultures remain the most widely used model because of their simplicity, reproducibility, and low operational cost. However, their inability to reproduce the complex cell–cell and cell–extracellular matrix interactions found in native tissues limits their physiological relevance and predictive capacity [4], [5]. These limitations have driven the development of three-dimensional (3D) culture systems, which more closely mimic the architecture, organization, and biological responses observed in vivo [4]–[6].

Among the available 3D culture models, multicellular spheroids have attracted considerable attention because they reproduce key structural and functional characteristics of native tissues, including enhanced cell–cell interactions, extracellular matrix deposition, and physiological gradients [7], [8], [24]. Consequently, spheroids have emerged as physiologically relevant models for cytotoxicity assessment, tissue engineering, and biomaterial testing [6], [9]. Their application in biocompatibility assessment is particularly relevant, as the biological evaluation of biomaterials remains an essential step in the development of biomedical devices and regenerative medicine strategies [3], [21], [22], [25].

Among the different methods available for spheroid generation, the hanging-drop technique stands out as a simple, scaffold-free, and highly reproducible approach [11]–[13]. In this method, cells are suspended in microdroplets, promoting spontaneous aggregation through gravitational forces and enhanced cell-cell interactions. This strategy enables the formation of compact and structurally organized spheroids that more closely resemble physiological tissue microenvironments [11], [21], [22]. In addition, recent advances have demonstrated that hanging drop-based microdroplet systems can be integrated with microfluidic platforms, allowing controlled perfusion, nutrient exchange, and real-time monitoring of three-dimensional spheroids [13], [14].

Microfluidic and organ-on-a-chip technologies have emerged as promising tools for the development of

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

microphysiological systems capable of reproducing tissue architecture and biological responses under controlled conditions [15]–[19]. These platforms combine microfluidic, cell biology, and engineering to stimulate relevant microenvironments, including fluid flow, shear stress and cellular interactions [20]–[23]. Such characteristics have contributed to the increasing application of organ-on-a-chip devices for drug development and biomaterial testing, offering improved predictive capacity, reduced costs, and decreased reliance on animal models [15], [16], [18].

Despite the significant advances in three-dimensional culture systems and organ-on-a-chip technologies, studies describing the long-term generation and morphological characterization of Vero cell spheroids using the hanging-drop method remain scarce, particularly regarding their potential integration into microfluidic platforms [23].

Therefore, this study aimed to generate and characterize Vero cell spheroids using the hanging-drop method as a reproducible three-dimensional in vitro platform for biomaterial biocompatibility and cytotoxicity assessment, with potential future integration into microfluidic and organ-on-a-chip systems. By establishing a simple and reproducible protocol for long-term spheroid culture, this work provides a foundation for the development of advanced three-dimensional in vitro models and microphysiological systems.

II. Methodology

A. Vero Cells Cultivation

Vero cells (CCIAL 057) were cultivated in HAM F10 medium supplemented with 10% fetal bovine serum (FBS) and maintained at 37 °C in a humidified atmosphere containing 5% CO_2 [1]. The culture medium was replaced twice weekly, and cell passaging was performed when cultures reached a confluent monolayer. Cell growth and morphology were monitored by phase-contrast microscopy using an inverted microscope (Axio Vert A1, Zeiss). This cell line is widely used as a standard model for cytotoxicity assessment and cellular response studies [25], [26].

B. Hanging Drop Microdroplet Culture Method

Spheroids were generated using the hanging-drop microdroplet technique. A cell suspension obtained from 7 and 21-day Vero cell cultures was dispensed onto the lid of a Petri dish as individual droplets. The dish was then inverted, allowing gravitational forces to promote cell sedimentation and aggregation at the lowest point of each droplet [11]–[13].

C. Spheroid Characterization

After spheroid formation by the hanging-drop method, samples were initially examined by direct and phase-contrast microscopy. For additional microscopic evaluation, selected droplets containing spheroids were transferred onto glass coverslips, and 5 µL of glutaraldehyde 2,5% solution was added to each droplet for cell fixation. After a 10-minute fixation period, the samples were washed to remove residual fixative and prepared for subsequent microscopic analyses. Subsequently, spheroids were stained with 4′,6-diamidino-2 phenylindole (DAPI) and analyzed by fluorescence microscopy.

Spheroids were visualized using a light microscope (Axio Vert A1, Zeiss) and a fluorescence microscope (Axio Imager A2, Zeiss).

III. Results and Discussions

Vero cell spheroids were obtained using the hanging-drop method with cells at passage numbers below 160. Each droplet had a final volume of 25 µL and contained cells at a density of 8×10^5 cells/mL, corresponding to approximately 20,000 cells per droplet. The droplets were then seeded into two separate Petri dishes and cultured for 7 days (Fig. 1a) and 21 days (Fig. 1b), respectively.

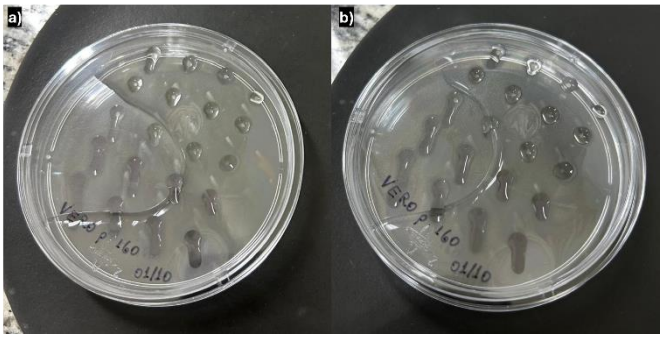


Fig. 1 Petri dishes with Vero cell hanging droplets spheroids cultures: a) 7 days; and b) 21 days.

Subsequently, the samples were analyzed using different imaging techniques: brightfield light microscopy (Fig. 2), phase-contrast microscopy (Fig. 3), and fluorescence microscopy stained with DAPI (Fig. 4).

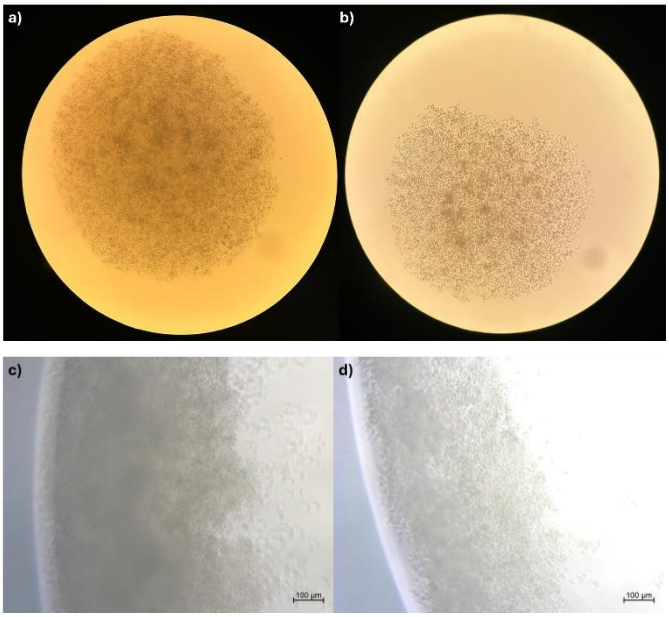


Fig. 2 a,b Direct light microscopy. Vero cells cultivated for a) 7 days; and b) 21 days. Magnification 100x; c,d Light microscopy, phase-contrast. Vero cell cultivated for c) 7 days; and d) 21 days. Magnification 400x.

Spheroid morphology was monitored using brightfield light microscopy (Figure 2a and 2b), allowing for the evaluation of spheroid formation compaction and cellular integrity throughout the 7-day (Fig. 2a) and 21-day (Fig. 2b) culture periods.

Figure 2c and 2d display Vero cell spheroids after 7 days of cultivation using phase-contrast microscopy. The images reveal the emergence of compact, spherical aggregates that conform to the geometry of the microdroplet. This morphology confirms successful spheroid condensation and consistent maintenance of the 3D structure over the cultivation period. The cells exhibit high affinity for one

another, forming stable structures that reflect the micro-confinement of the droplet environment.

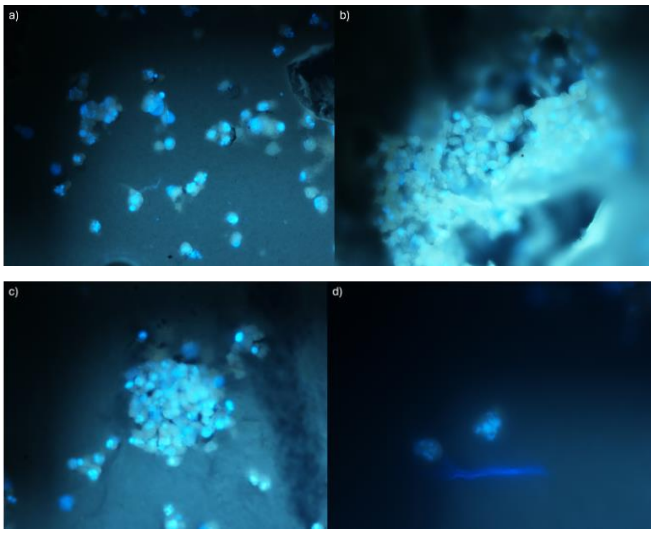


Fig. 4 Fluorescence microscopy with DAPI staining. Vero cell spheroid cultured for: a,b) 7 days; and c,d) 21 days. Magnification 400x.

Fluorescence microscopy with DAPI staining (Figure 4) was employed to evaluate cell nucleus. The images demonstrate well-defined nuclear DNA throughout the spheroid, with no evidence of fragmented DNA or extracellular debris.

Furthermore, the structural resilience of the spheroids was evaluated during post-processing. Upon inversion of the culture plate, the microdroplets spread due to gravitational forces and lost their initial spherical geometry. Nevertheless, the cellular aggregates remained intact, indicating that spheroid integrity was maintained independently of droplet confinement. This observation demonstrates that stable cell–cell interactions had already been established during culture, allowing the aggregates to preserve their three-dimensional architecture despite the loss of external physical support. Such behavior is consistent with the hanging-drop technique, in which gravity-driven cell sedimentation and the absence of substrate adhesion promote spontaneous cellular self-assembly into compact and reproducible spheroids while minimizing mechanical stress during formation [10], [11], [27]. The preservation of spheroid integrity after post-processing is particularly relevant because mechanically stable aggregates are less susceptible to disruption during routine procedures, including medium replacement, staining, imaging, and downstream biological assays [4], [12].

The progressive compaction observed after both 7 and 21 days, together with the preservation of nuclear integrity demonstrated by DAPI staining, further indicates successful spheroid maturation and sustained cellular viability. As spheroids mature, enhanced intercellular adhesion and cytoskeletal organization contribute to increased mechanical stability and structural cohesion, allowing the aggregates to better reproduce the cellular organization and microenvironment found in native tissues than conventional monolayer cultures [4], [6], [7]. The maintenance of spheroid morphology throughout the 21-day cultivation period further demonstrates the suitability of the proposed culture conditions for prolonged three-dimensional growth. Long-term structural stability is particularly advantageous for experimental models requiring extended observation periods, repeated analyses, and chronic biological assays [12]. Moreover, the compact organization observed in the present study agrees with previous reports showing that hanging-drop cultures produce homogeneous spheroids [10], [11], [27].

Vero cells are extensively used in virology owing to their robust proliferative capacity, broad susceptibility to viral infection, and adaptability to different culture conditions [1], [2]. The successful generation of compact and viable Vero spheroids observed in this study therefore reinforces the applicability of this well-established cell line to more physiologically relevant three-dimensional models. Previous studies have shown that Vero cell spheroids improve the evaluation of host–pathogen interactions and increase the sensitivity of biological assays when compared with conventional monolayer cultures [3], [23]. Collectively, these findings demonstrate that the hanging-drop protocol provides a simple strategy for producing structurally stable Vero cell spheroids suitable for future long-term culture, antiviral screening, virus production, and other advanced *in vitro* applications [10]–[12], [23].

The ability to maintain structurally preserved spheroids for up to 21 days demonstrates the efficacy of the proposed protocol and highlights its potential for long-term *in vitro* studies requiring repeated experimental manipulation. Although the present study successfully demonstrates the formation and morphological characterization of Vero cell spheroids, additional functional analyses, including metabolic activity, proliferation, cell viability, quantitative measurements of the spheroids, will be necessary to further validate this model for specific biomedical applications.

IV. Conclusion

This study successfully demonstrated the generation and characterization of 3D Vero cell spheroids using the hanging-drop method. The proposed scaffold-free protocol enabled the formation of compact, structurally stable cellular aggregates that maintained their morphology and nuclear integrity throughout the cultivation period, including after post-processing. These findings establish a simple, and low-cost strategy for generating long-term Vero cell spheroids, providing a reliable three-dimensional *in vitro* model for future biological investigations.

Future studies should focus on the functional characterization of these spheroids and their integration into microfluidic organ-on-a-chip platforms to develop more physiologically relevant models for biological research, cytotoxicity and biomaterial interactions. In addition, the proposed that the culture system may serve as a promising platform for biomaterial evaluation and, potentially, for applications in tissue engineering and regenerative medicine.

Acknowledgment

The authors would like to thank Federal University of ABC for financial support, Tissue Engineering and Biomolecule Laboratory for infrastructure and are grateful to Multiuser Center Facilities (CEM - UFABC) for the experimental support.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] N. C. Ammerman, M. Beier-Sexton, and A. F. Azad, “Growth and maintenance of Vero cell lines,” Curr. Protoc. Microbiol., vol. 11, no. 1, pp. A.4E.1–A.4E.7, 2008, doi: 10.1002/9780471729259.mca04es11.

[2] C. S. Takata, F. S. Kubrusly, C. Miyaki, I. F. Mendes, and E. de Rizzo, “Susceptibility of Vero cell line to vaccine strains of the measles virus,” Rev. Saúde Pública, vol. 28, no. 3, pp. 209–212, 1994, doi: 10.1590/S0034-89101994000300008.

[3] J. C. Fontoura et al., “Comparison of 2D and 3D cell culture models for cell growth, gene expression and drug resistance,” Mater. Sci. Eng. C Mater. Biol. Appl., vol. 107, Art. no. 110264, 2020, doi: 10.1016/j.msec.2019.110264.

[4] K. Białkowska, P. Komorowski, M. Bryszewska, and K. Miłowska, “Spheroids as a type of three-dimensional cell cultures—examples of methods of preparation and the most important application,” Int. J. Mol. Sci., vol. 21, no. 17, Art. no. 6225, 2020, doi: 10.3390/ijms21176225.

[5] N.-E. Ryu, S.-H. Lee, and H. Park, “Spheroid culture system: Methods and applications for mesenchymal stem cells,” Cells, vol. 8, no. 12, Art. no. 1620, 2019, doi: 10.3390/cells8121620.

[6] A. S. Nunes, A. S. Barros, E. C. Costa, A. F. Moreira, and I. J. Correia, “3D tumor spheroids as in vitro models to mimic in vivo human solid tumors resistance to therapeutic drugs,” Biotechnol. Bioeng., vol. 116, no. 1, pp. 206–226, 2019, doi: 10.1002/bit.26845.

[7] M. T. Santini, G. Rainaldi, and P. L. Indovina, “Apoptosis, cell adhesion and the extracellular matrix in the three-dimensional growth of multicellular tumor spheroids,” Crit. Rev. Oncol. Hematol., vol. 36, nos. 2–3, pp. 75–87, 2000, doi: 10.1016/S1040-8428(00)00078-0.

[8] F. Rescigno, L. Ceriotti, and M. Meloni, “Extra cellular matrix deposition and assembly in dermis spheroids,” Clin. Cosmet. Investig. Dermatol., vol. 14, pp. 935–943, 2021, doi: 10.2147/CCID.S316707.

[9] A. S. Mikhail, S. Eetezadi, and C. Allen, “Multicellular tumor spheroids for evaluation of cytotoxicity and tumor growth inhibitory effects of nanomedicines in vitro: A comparison of docetaxel-loaded block copolymer micelles and Taxotere®,” PLoS ONE, vol. 8, no. 4, Art. no. e62630, 2013, doi: 10.1371/journal.pone.0062630.

[10] A. P. P. Guimaraes, I. R. Calori, H. Bi, and A. C. Tedesco, “SpheroMold: Modernizing the hanging drop method for spheroid culture,” Front. Drug Deliv., vol. 4, Art. no. 1397153, 2024, doi: 10.3389/fddev.2024.1397153.

[11] M. Rasouli, F. Safari, M. H. Kanani, and H. Ahvati, “Principles of hanging drop method (spheroid formation) in cell culture,” in Methods in Molecular Biology, vol. 2879, New York, NY, USA: Humana, 2025, pp. 289–300, doi: 10.1007/7651_2024_527.

[12] H. Ohguro et al., “Application of single cell type-derived spheroids generated by using a hanging drop culture technique in various in vitro disease models: A narrow review,” Cells, vol. 13, no. 18, Art. no. 1549, 2024, doi: 10.3390/cells13181549.

[13] T. J. Bartosh and J. H. Ylostalo, “Preparation of anti-inflammatory mesenchymal stem/precursor cells (MSCs) through sphere formation using hanging-drop culture technique,” Curr. Protoc. Stem Cell Biol., vol. 28, no. 1, pp. 2B.6.1–2B.6.23, 2014, doi: 10.1002/9780470151808.sc02b06s28.

[14] Y. Wang, M. Liu, Y. Zhang, H. Liu, and L. Han, “Recent methods of droplet microfluidics and their applications in spheroids and organoids,” Lab Chip, vol. 23, no. 5, pp. 1080–1096, 2023, doi: 10.1039/D2LC00493C.

[15] O. Frey, P. M. Misun, D. A. Fluri, J. G. Hengstler, and A. Hierlemann, “Reconfigurable microfluidic hanging drop network for multi-tissue interaction and analysis,” Nat. Commun., vol. 5, Art. no. 4250, 2014, doi: 10.1038/ncomms5250.

[16] P. M. Misun, J. Rothe, Y. R. F. Schmid, A. Hierlemann, and O. Frey, “Multi-analyte biosensor interface for real-time monitoring of 3D microtissue spheroids in hanging-drop networks,” Microsyst. Nanoeng., vol. 2, Art. no. 16022, 2016, doi: 10.1038/micronano.2016.22.

[17] C. M. Leung et al., “A guide to the organ-on-a-chip,” Nat. Rev. Methods Primers, vol. 2, Art. no. 33, 2022, doi: 10.1038/s43586-022-00118-6.

[18] A. E. Danku, E. H. Dulf, C. Braicu, A. Jurj, and I. Berindan-Neagoe, “Organ-on-a-chip: A survey of technical results and problems,” Front. Bioeng. Biotechnol., vol. 10, Art. no. 840674, 2022, doi: 10.3389/fbioe.2022.840674.

[19] Q. Wu et al., “Organ-on-a-chip: Recent breakthroughs and future prospects,” Biomed. Eng. Online, vol. 19, Art. no. 9, 2020, doi: 10.1186/s12938-020-0752-0.

[20] D. Singh, A. Mathur, S. Arora, S. Roy, and N. Mahindroo, “Journey of organ-on-a-chip technology and its role in future healthcare scenario,” Appl. Surf. Sci. Adv., vol. 9, Art. no. 100246, 2022, doi: 10.1016/j.apsadv.2022.100246.

[21] L. S. A. F. Oliveira, C. S. Oliveira, A. P. L. Machado, and F. P. Rosa, “Biomaterials for bone regeneration: Methods of analyses and future perspectives,” Rev. Ciênc. Méd. Biol., vol. 9, suppl. 1, pp. 37–44, 2010, doi: 10.9771/cmbio.v9i1.4730.

[22] A. L. R. Pires, A. C. K. Bierhalz, and Â. M. Moraes, “Biomaterials: Types, applications and market,” Quím. Nova, vol. 38, no. 7, pp. 957–971, 2015, doi: 10.5935/0100-4042.20150094.

[23] C. Z. To and A. K. Bhunia, “Three-dimensional Vero cell-platform for rapid and sensitive screening of Shiga-toxin producing Escherichia coli,” Front. Microbiol., vol. 10, Art. no. 949, 2019, doi: 10.3389/fmicb.2019.00949.

[24] H. Ijima, K. Nakazawa, H. Mizumoto, T. Matsushita, and K. Funatsu, “Formation of a spherical multicellular aggregate (spheroid) of animal cells in the pores of polyurethane foam as a cell culture substratum and its application to a hybrid artificial liver,” J. Biomater. Sci. Polym. Ed., vol. 9, no. 7, pp. 765–778, 1998, doi: 10.1163/156856298X00136.

[25] International Organization for Standardization, Biological Evaluation of Medical Devices—Part 5: Tests for In Vitro Cytotoxicity, ISO 10993-5:2009, 3rd ed. Geneva, Switzerland: ISO, 2009..

[26] U. S. do Nascimento et al., “Ibuprofen loaded gelatin-siloxane gels prepared by supercritical CO_2-assisted impregnation/deposition for controlled drug release and tissue engineering,” J. Supercrit. Fluids, vol. 218, Art. no. 106511, 2025, doi: 10.1016/j.supflu.2024.106511.

[27] J. M. Kelm, N. E. Timmins, C. J. Brown, M. Fussenegger, and L. K. Nielsen, “Method for generation of homogeneous multicellular tumor spheroids applicable to a wide variety of cell types,” Biotechnol. Bioeng., vol. 83, no. 2, pp. 173–180, 2003, doi: 10.1002/bit.10655.

Effect of HAp addition on rheological behavior and cytocompatibility of gelatin/carrageenan-based bioinks for osteochondral applications

Angelo Romeu da Silva Junior
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Federal do ABC
São Bernardo do Campo, Brazil
angelo.romeu@ufabc.edu.br

Kaline Nascimento Ferreira
Biofabrication Group
Centro de Tecnologia da Informação Renato Archer
Campinas, Brazil
knferreira@cti.gov.br

Sônia Maria Malmonge
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Federal do ABC
São Bernardo do Campo, Brazil
sonia.malmonge@ufabc.edu.br

Gabriel Caio Kokkonias e Castro
Biofabrication Group
Centro de Tecnologia da Informação Renato Archer
Campinas, Brazil
gckcastro@cti.gov.br

Juliana Kelmy Mácario Barboza Daguano
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Federal do ABC
São Bernardo do Campo, Brazil
juliana.daguano@ufabc.edu.br

***Abstract*—Osteochondral injuries pose a significant clinical challenge due to the structural and functional complexity of the interface between articular cartilage and subchondral bone. In this context, various strategies based on 3D biomanufacturing and bioprinting have been explored with the aim of developing biomimetic constructs for Tissue Engineering. Among the main challenges associated with extrusion-based bioprinting is the development of bio-inks with appropriate rheological properties, structural stability, and cytocompatibility. The objective of this study was to develop and characterize gelatin/carrageenan-based bioinks incorporating hydroxyapatite, with a view toward future applications in biofabricated systems for osteochondral regeneration. The formulations were prepared with different concentrations of hydroxyapatite and evaluated for their rheological properties using viscosity and thixotropic recovery tests, investigating parameters related to printability and pseudoplastic behavior. Additionally, the formulations underwent preliminary evaluation of extrusion and structural fidelity. Furthermore, FTIR spectroscopy and X-ray diffraction (XRD) analyses were performed to investigate the interaction between the ceramic phase and the polymer matrix.Finally, the cytocompatibility of the bioinks will be investigated via cell viability assays over different culture periods. It is expected that the incorporation of hydroxyapatite will promote changes in the properties viscoelastic materials for the development of bioactive formulations with potential application as a mineralized phase in osteochondral biofabrication strategies.**

***Index Terms*—bioink; 3D printing; rheological ; hydroxipatite**

Identify applicable funding agency here. If none, delete this.

I. Introduction

Lesions of the osteochondral tissues represent one of the greatest therapeutic challenges in modern orthopedics, significantly affecting the quality of life of millions of people worldwide. Characterized by the simultaneous degeneration of articular cartilage and subchondral bone, these lesions have a low capacity for spontaneous regeneration, largely due to the avascular nature of cartilage and the complexity of the osteochondral interface. It is estimated that, globally, more than 250 million people suffer from degenerative joint diseases, such as osteoarthritis, which are often exacerbated by failures in osteochondral tissue regeneration[1]. In this context, various strategies based on tissue engineering and biomanufacturing have been extensively explored for the development of biomimetic structures capable of promoting tissue regeneration. To enable the fabrication of these customized and hierarchical architectures, biomaterials are selected based on the need to balance mechanical support and cellular bioactivity. Among the biomaterials most commonly used in the fabrication of scaffolds for bone applications, ceramics,such as hydroxyapatite (HAp), calcium phosphates, and bioglasses,stand out, as do natural polymers, including alginate, chitosan, and gelatin [2, 3].Among manufacturing approaches, extrusion bioprinting stands out for its ability to produce three-dimensional structures containing cells and biomaterials with controlled geometries. However, the success of this process depends directly on the physicochemical and

rheological properties of the bioinks, which must exhibit appropriate extrusion behavior, structural stability, and cytocompatibility. In the field of bioprinting, the development of polymeric bioinks plays a fundamental role in the construction of biomimetic cellular microenvironments, particularly hydrogels based on natural polymers. In this context, the combination of gelatin and i-carrageenan offers complementary properties that favor its application in tissue engineering. While i-carrageenan offers gelling capacity and structural similarity to glycosaminoglycans present in the extracellular matrix, gelatin provides bioactive RGD-type sequences that promote cell adhesion and interaction [4, 5].On the other hand, hydroxyapatite is the main inorganic component of human bones and, therefore, serves as an excellent additive for hydrogels in bone tissue engineering. Thus, on its own, HAp can promote adhesion, proliferation, and extracellular matrix deposition by osteoblasts[6, 7]. Thus, the combination of these two polymers generates a polysaccharide-protein hybrid network capable of forming gels under physiological conditions, balancing mechanical support, cellular bioactivity, and potential rheological adjustment for applications in 3D bioprinting. Thus, the objective of this study is to develop and characterize gelatin/carrageenan-based bioinks containing hydroxyapatite, investigating the influence of the ceramic phase on rheological properties and initial cell viability, with a view toward future applications in biofabricated systems for osteochondral regeneration.

II. METHODOLOGY

A. Gel inks preparation

Bioinks delopment were based on gelatin/carrageenan systems , with a focus on the study and influence of hydroxyapatite [8, 9]. To this end, gelatin- and carrageenan-based bioinks were prepared by dissolving the polymers in DMEM/PBS under magnetic stirring and temperature control. After complete solubilization, different concentrations of hydroxyapatite (HAp) were incorporated into the polymer matrix to obtain mineralized bioactive formulations. The formulations will be divided into:

- GC: gelatin/carrageenan (control)
- HAp 0.2%: gelatin/carrageenan + 0.2% HA
- HAp 2%: gelatin/carrageenan + 2% HA
- HAp 20%: gelatin/carrageenan + 20% HA

B. Rheological tests

the rheological characterizations of the bionks were conducted to understand the the viscoelastic properties relavant to controlled extrusion and deposition. the tests were conducted on an anton paar MCR102 rotational rheometer using a plate-plate geometry (PP25, 25 mm, 0.7 mm gap) at 25 °C. To prevent slippage along the walls and ensure reliable data acquisition at high shear rates (1000 s^{1}), the geometry surfaces were roughened using fine-grit sandpaper (P600), the surfaces of the geometry were modified using fine-grit sandpaper (P600). This technique increases surface roughness, preventing material displacement and ensuring effective stress transfer. Thus, viscosity curves (0.01–1000 s^{1}) and amplitude sweeps (0.1, 1.0, 10 rad s^{1}, LVE = 0.02) were used to determine structural integrity during printing. Frequency sweeps (0.1–300 rad s^{1}) investigated the response of the ink. Three-stage recovery tests simulated behavior during and after deposition: pre-shearing (1 s^{1}, 30 s), rest (1 s^{1}, 25 s), shearing (10 s^{1}, 50 s), and regeneration (1 s^{1}, 250 s)[10, 11]. All rheological measurements were performed in duplicate

C. Extrusion-Based 3D printing

Initially , the 3D extrusion-based from 593ICAN LTDA (Brazil) was tested using Hap 20% ink which served as the ink with the ideal viscosity. The printing speed was set to 10 mm s^{1} using conical nozzles with inner diameters of 20 gauge (580 µm) and 22 gauge for the higher-viscosity hydrogels and using repetier-Host V2 1.3 software, to compare the printed geometries. G-code files containing the printing paths were generated using BioScaffoldsPg software [11].

D. Structural/Physical-Chemical Characterization

The characterization of biomaterials was conducted using a set of physicochemical and structural techniques to evaluate the properties essential for bioprinting and the functional performance of the scaffolds. The structural stability of hydroxyapatite and the integrity of the crystalline phases in the bioprinted scaffolds were evaluated by X-ray diffraction (XRD) on an Aeris (Malvern Panalytical), with a 2 scan from 10° to 80° (continuous mode, 0.02°/s) and indexing via HighScore software. The focus will be on detecting possible crystallographic changes or interactions resulting from the association with the gelatin/-carrageenan hydrogel In addition, Fourier-Transform Infrared Spectroscopy (FTIR) was performed using a PerkinElmer Spectrum 100 instrument (diamond crystal), with 16 scans between 4000–550 cm^{1} and a resolution of 4 cm^{1}.

E. Cell Viabillty

To assess the potential cytotoxicity of the formulations, L929 mouse fibroblast-like cells (BCRJ Cell Bank, code 0188, NCTC clone 929) were cultured in the presence of each gel formulation. Cells at the 28th passage were seeded into 96-well plates at a density of 1 × 10 cells per well and cultured in DMEM under a humidified atmosphere containing 5% CO_2 at 37 °C for 24 h.

After this period, the culture medium was replaced with 100 µL of fresh medium, and 100 µL of gel was added to each well. Cells were incubated under the same conditions for an additional 24 h. Subsequently, the gels were removed, and the wells were washed twice with PBS to eliminate residual material.Next, 100 µL of MTT solution [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] (Thermo Fisher Scientific, USA) at a concentration of 0.5 mg mL^{1} was added to each well and incubated for 4.5 h. The MTT solution was then removed and replaced with 100 µL of dimethyl sulfoxide (DMSO) to dissolve the formed formazan crystals. Absorbance

was measured at 540 nm using a microplate reader (LMR-96, Loccus, Brazil).Cell viability was determined by comparing the absorbance values of the samples with those of the negative control, consisting of a alumina phosphate extract prepared according to ISO 10993-5 and known to be non-cytotoxic. DMSO (20%) and Triton X-100 (1%) diluted in DMEM were used as positive controls. Data are presented as the mean of eight replicates.

III. Results and Discussion

A. Rheological tests

In the development of a gelatin/ -carrageenan polymer blend as a matrix for a chondral bioink the influence of the gradual incoporation of the hydroxyapatite (HAp) was rheological behavior of the systematically evaluated. All formulations shown in Figure 1 exhibited shear-thinning behavior, characterized by a decrease in viscosity with increasing shear rate profile typical of polymeric systems intended for extrusion bioprinting. This behavior is consistent with the literature on polymeric hydrogels, in which the continuous matrix phase dominates the flow profile ,even after the incorporation of solid phases. [9, 12, 13]. Thus, moderate variations in HAp content resulted in slight changes in viscosity without compromising the shear-thinning behavior of the formations. Formulations containing 6% gelatin and 0,3% -carrageenan, combined with different concentrations of HAp (0; 0.2; 2, and 20% m/v), with the aim of identifying a range of ceramic content capable of providing structural reinforcement without compromising the system's extrudability or causing clogging of the print nozzle, a critical aspect for biomanufacturing applications([12, 14]. Formulations containing 20% HAp exhibiting the highest resistance to flow across the entire range of shear rates evaluated. In contrast , the formulations containing 0.2% and 2% exhibited rheological behavior similar to that of the control,suggesting that low concentrations of the ceramic phase were insufficient to promote a significant increase in resistance to flow, this indicate that,within this concentration range, the polymer matrix still dominates the flow, and the ceramic phase acts primally as a dispersed filler without significantly altering the shear-thinning behavior-a phenomenon also observed in the other polymer hydrogels with low HAp contents . finally, the increase in viscosity observed for the formulation containing 20% HAp suggets a structural reinforcement effect promoted by the ceramic phase, possibility associated with the restriction of polymer chain mobility and increased physical interactions among the matrix components. [14, 15].in addition interactions may occur between the sulfate groups of the carrageenan,and calcium ions present on the surface of hydroxyapatite , contributing to the formation of a more stabile structural network[16] In the time-recovery test (3ITT) , all formulations shown in figure 2 exhibited thixotropic behavior with rapid recovery , as evidenced by the abrupt decrease in viscosity during high-shear application, followed by partial recovery on the structure upon returning to low-shear conditions.In the first stage of the test , conducted at a low strain rate ,a progressive increase in viscosity was observed as

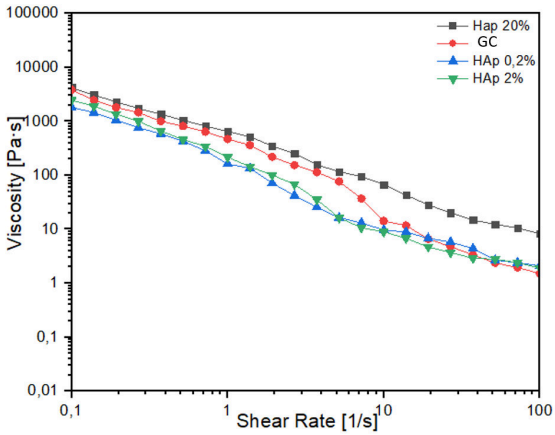


Fig. 1. – Shear-thinning behavior of control and HAp formulations at different concentrations.

the HAp content increased; formulation containing 20% HAp exhibited the highest initial viscosity initial viscosity values, indicating greater structural organization at rest. Subsequently, during the high-shear stage, all formulations underwent a sharp decrease in viscosity, a behavior associated with the reversible breakdown of the physical interactions responsible for the stability of the polymer network. [14, 17].The formulation containing 20% Hap exhibited the highest initial viscosity values, indicating greater structural organization at rest. However, during the high-shear stage, all formulations underwent a sharp decrease in viscosity, a behavior associated with the reversible breakdown of the physical interactions responsible for the stability of the polymer network. In this regard, the formulation containing 20% HAp maintained a higher residual viscosity than the other formulations even under high shear, suggesting greater structural resistance to flow-induced disorganization [18, 19]. In contrast, the formulations containing 0.2% and 2% HAp exhibited behavior more similar to that of the control, indicating that reduced concentrations of the ceramic phase had a limited influence on resistance to structural collapse. On the other hand, this range of formulations with low rheological perturbation relative to the pure matrix is useful for incorporating cells and biomolecules, whereas 20% maximizes structural support [14, 17, 20]. After the high shear was discontinued, a gradual recovery in viscosity was observed for all samples, demonstrating the hydrogel's three-dimensional network's ability to partially reconstruct itself. The initial increase in viscosity and greater recovery observed for the formulation containing 20% HAp in this study suggest the formation of a more interconnected particulate network, capable of sustaining the structure even after temporary disruption under flow. suggesting that high concentrations of HAp contributes to increased structural stability and post-shear recovery capacity of the formulations, while maintaining thixotropic behavior suitable for applications in extrusion bioprinting[14].

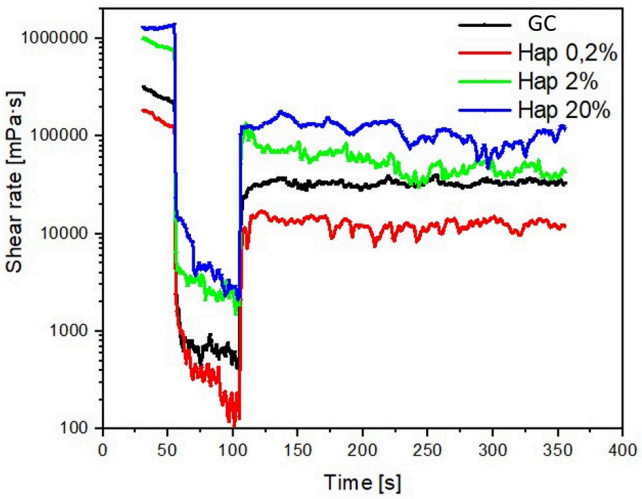


Fig. 2. Temporal recovery of viscosity (3ITT) for formulations with different hydroxyapatite concentrations.

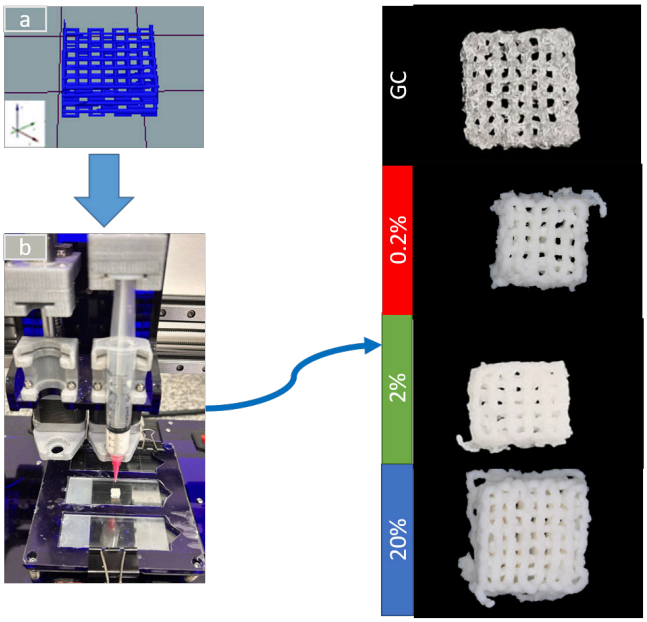


Fig. 3. CAD design, bioprinting process, and printed structures with varying hydroxyapatite concentrations.

B. Printed scaffolds and morphology

The scaffolds were fabricated by layer-by-layer deposition of hydrogels, based on the previously defined CAD model. Before fabricating the final samples, printing parameters—such as deposition speed, extrusion pressure, extrusion ratio, and porosity—were systematically optimized to obtain structures with high fidelity to the digital model and adequate structural integrity. The results of this optimization process are shown in Figure 3A–C.

C. Structural/Physical-Chemical Characterization

The FTIR spectra of the formulations containing gelatin/carrageenan and different concentrations of the ceramic phase are shown in Figure 3. Characteristic bands of both the polymer matrix and the organic material are observed, indicating the coexistence of the components and possible intermolecular interactions between them. Similar patterns of Amide I/II and sulfate bands have been described in gelatin/carrageenan gels, where subtle shifts are associated with the formation of polyelectrolyte complexes between the protein and the sulfated polysaccharide; Due to the conformational reorganization of gelatin in the control samples and those with lower ceramic content, broad bands are observed in the region between 3200–3400 cm¹, which are attributed to the stretching of hydroxyl O-H/N-H groups and are consistent with the FTIR spectra of gelatin[21, 22]. Which are characteristics of both gelatin and carrageenan, associated with the hydrogen bonds present in the hydrogel matrix. In the region between 1600–1500 cm¹, bands associated with the amide groups of gelatin—especially type I and type II—are observed; these are associated with the C=O and N-H groups in the protein structure of gelatin. Meanwhile, the region between 1220–1260 cm¹ may be associated with the sulfate groups (S=O) of carrageenan[21–24]. As the hydroxyapatite content increased, particularly in the formulation containing 20 wt%, a progressive increase in the intensity of the absorption bands between 900 and 1100 cm¹ was observed. This spectral region is mainly associated with the asymmetric stretching vibration of phosphate (PO^3) groups, characteristic of hydroxyapatite. The higher intensity of these bands reflects the increased contribution of the ceramic phase to the composite spectrum and confirms the successful incorporation of hydroxyapatite into the polymeric matrix. The absence of significant band shifts suggests that the primary chemical structures of the polymer components were preserved after ceramic incorporation, while the intensity changes are predominantly related to the higher concentration of phosphate groups in the formulations. [25, 26]. This behavior indicates a greater contribution of the inorganic phase to the spectrum and suggests good incorporation of the ceramic particles into the polymer matrix. The sample containing 20% exhibited significantly more intense and well-defined bands, especially near 1000 cm¹, indicating a predominance of vibrations from the mineral phase. The increase in the intensity of these bands may be associated with the greater number of available phosphate/ groups, as well as with increased interfacial interaction between the ceramic and the hydrogel polymers.In addition, Figure 4 shows the diffractograms of the starting materials, highlighting the striking differences in their structural organization. Gelatin exhibited a diffuse halo centered at approximately 20°, which is characteristic of predominantly amorphous materials and is consistent with the molecular disorganization expected following the partial denaturation of collagen. Unlike gelatin, carrageenan exhibits a semicrystalline pattern characterized by the coexistence

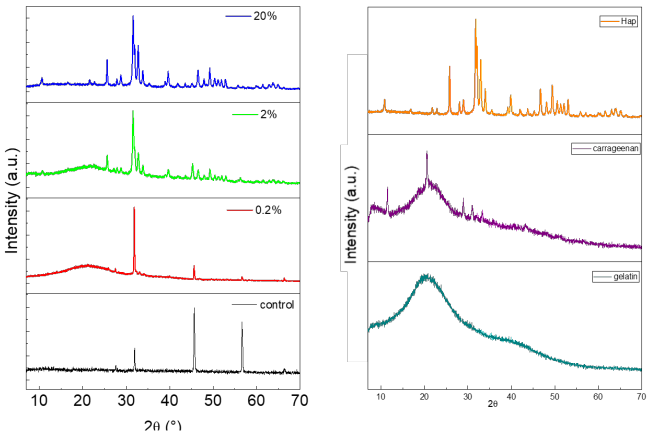


Fig. 4. XRD patterns of composite samples with varying hydroxyapatite contents (left) compared to reference crystalline phases (right).

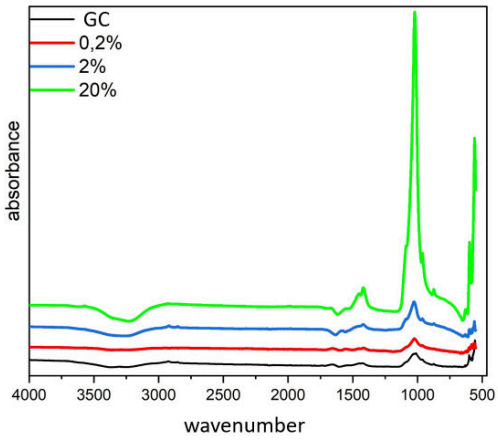


Fig. 5. FTIR spectra of control and hydroxyapatite-containing formulations at different concentrations.

of an amorphous halo with discrete reflections distributed throughout the diffractogram (15°, 24°, 33°, 34°)[27, 28]. This behavior stems from the partial organization of the galactan chains into helical structures capable of forming junction domains stabilized by intermolecular bonds and electrostatic interactions. In contrast, hydroxyapatite exhibited a highly crystalline diffraction pattern, with intense reflections at the peaks of 25.9°, 31.8°, 32.9°, 34.1°, 39.8°, and 46.7°, respectively, consistent with the hexagonal structure typical of stoichiometric hydroxyapatite[6, 16, 29]. In the case of hydrogels, it is possible to see how a gradual increase in the concentration of the mineral phase led to a significant increase in the intensity of the crystalline reflections associated with hydroxyapatite, while maintaining the characteristic amorphous halo of the polymer matrix. This behavior indicates that the incorporation of the ceramic phase increases the crystalline fraction of the system without substantially altering the predominantly amorphous nature of the gelatin/carrageenan blend[30].

These structural differences suggest that, when combined, the three constituents may perform complementary functions in the bio-ink formulation. While the predominantly amorphous nature of gelatin promotes chain mobility and the formation of a continuous polymer network, the semicrystalline organization of carrageenan contributes to the formation of junction zones responsible for increasing the hydrogel's structural stability. At the same time, the high crystallinity of hydroxyapatite provides a stable inorganic phase, which is potentially capable of acting as a mechanical reinforcing agent and providing sites for interactions with the polymer matrix, in addition to imparting bioactivity to the formulation. Together, these structural characteristics provide a favorable basis for the development of bioinks for bone tissue engineering, in which the balance between structural stability, rheological properties, and bioactivity is essential for performance during printing and tissue regeneration.

D. Cell Viabillty

MTT assays were performed to determine the percentage of viable cells adhering to the hydrogel samples (fig 6). the results showed that the number of viable cell onto the systems of gelatin/carragenaan. Preliminary cell viability results obtained by the MTT assay demonstrated that all gelatin/i-carrageenan hydrogels containing different concentrations of hydroxyapatite (Hap) exhibited high cytocompatibility toward L929 fibroblast-like cells, indicating that the incorporation of the ceramic phase did not induce detectable cytotoxic effects under the experimental conditions evaluated [31–33]. Among the investigated formulations, the hydrogel containing 2 wt% HAp exhibited the highest metabolic activity, consistent with previous studies reporting that low to moderate HAp contents may enhance cellular metabolic activity while maintaining high cell viability [8, 34, 35]. Although the formulation containing 20 wt% HAp showed a lower metabolic response than the 2 wt.% HAp group, cell viability remained within a range considered compatible with tissue engineering applications, suggesting that even highly mineralized formulations can be well tolerated when the ceramic phase is homogeneously incorporated into the polymeric matrix[36–38]. Collectively, these preliminary findings indicate that hydroxyapatite can be successfully incorporated into the gelatin/-carrageenan hydrogel without substantially compromising cytocompatibility, reinforcing the potential of this biomaterial platform for future applications in extrusion-based 3D bioprinting and osteochondral tissue engineering[39–41] .

IV. Conclusion

This study investigated gelatin/-carrageenan hydrogels containing increasing concentrations of hydroxyapatite to establish the relationship between ceramic content, physicochemical structure, rheological behavior, and biological response. ATR-FTIR analyses confirmed the successful incorporation of hydroxyapatite into the polymeric network while preserving the

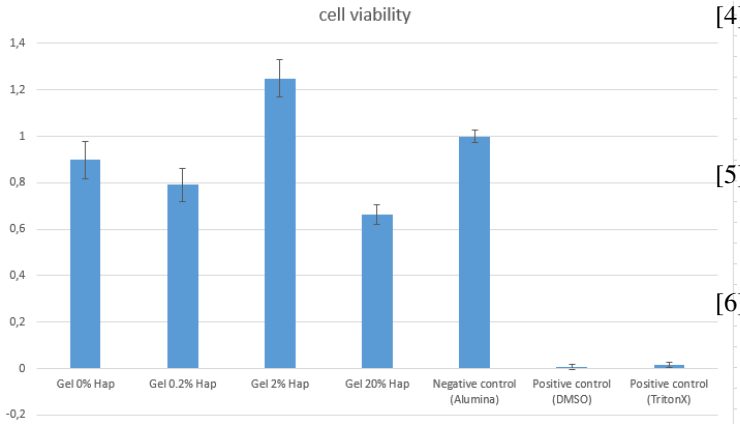


Fig. 6. MTT assay results showing cytotoxicity and cell viability of hydrogels containing varying hydroxyapatite contents compared to controls.

characteristic functional groups of both polymers. Structural characterization demonstrated that increasing hydroxyapatite content enhanced the contribution of the mineral phase without compromising the integrity of the hydrogel matrix. Rheological analyses revealed that the ceramic phase significantly influenced the viscoelastic behavior of the formulations, leading to improved mechanical stability and printability. Furthermore, the biological assays demonstrated that the developed composite hydrogels supported cellular viability, indicating good cytocompatibility. Overall, the formulation containing an optimized hydroxyapatite content exhibited a balanced combination of structural integrity, rheological performance, and biological properties, highlighting its potential as a bioink for bone tissue engineering applications.

Acknowledgment

To the team at the Open Biotechnology Laboratory of the Renato Archer Information Technology Center (CTI), the Brazilian Center for Physics Research (CBPF), the Agency for the Financing of Studies and Projects (Finep-22.01.0141.00), and CAPES for their financial support in the development of this work

References

[1] J. D. Steinmetz, G. T. Culbreth, L. M. Haile, Q. Rafferty, J. Lo, K. G. Fukutaki, J. A. Cruz, A. E. Smith, S. E. Vollset, P. M. Brooks *et al.*, "Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the global burden of disease study 2021," *The Lancet Rheumatology*, vol. 5, no. 9, pp. e508–e522, 2023.

[2] D. W. Hutmacher, "Scaffolds in tissue engineering bone and cartilage," *Biomaterials*, vol. 21, no. 24, pp. 2529–2543, 2000.

[3] F. J. O'brien, "Biomaterials & scaffolds for tissue engineering," *Materials today*, vol. 14, no. 3, pp. 88–95, 2011.

[4] L. Tytgat, M. Vagenende, H. Declercq, J. Martins, H. Thienpont, H. Ottevaere, P. Dubruel, and S. Van Vlierberghe, "Synergistic effect of κ-carrageenan and gelatin blends towards adipose tissue engineering," *Carbohydrate polymers*, vol. 189, pp. 1–9, 2018.

[5] S. Sinthusamran, S. Benjakul, P. J. Swedlund, and Y. Hemar, "Physical and rheological properties of fish gelatin gel as influenced by κ-carrageenan," *Food Bioscience*, vol. 20, pp. 88–95, 2017.

[6] N. B. Allen, B. Abar, L. Johnson, J. Burbano, R. M. Danilkowicz, and S. B. Adams, "3d-bioprinted gelma-gelatin-hydroxyapatite osteoblast-laden composite hydrogels for bone tissue engineering," *Bioprinting*, vol. 26, p. e00196, 2022.

[7] P. P. Patel, C. Buckley, B. L. Taylor, C. C. Sahyoun, S. D. Patel, A. J. Mont, L. Mai, S. Patel, and J. W. Freeman, "Mechanical and biological evaluation of a hydroxyapatite-reinforced scaffold for bone regeneration," *Journal of Biomedical Materials Research Part A*, vol. 107, no. 4, pp. 732–741, 2019.

[8] P. Stolarov, J. De Vries, S. Stapleton, L. Morris, K. Martyniak, and T. J. Kean, "Suitability of gelatin methacrylate and hydroxyapatite hydrogels for 3d-bioprinted bone tissue," *Materials*, vol. 17, no. 5, p. 1218, 2024.

[9] L. M. Anaya-Sampayo, N. S. Roa, C. Martínez-Cardozo, D. A. García-Robayo, and L. M. Rodríguez-Lorenzo, "Influence of hydroxyapatite and gelatin content on crosslinking dynamics and hdfn cell viability in alginate bioinks for 3d bioprinting," *Polymers*, vol. 16, no. 22, p. 3224, 2024.

[10] K. N. Ferreira, J. A. Ferreira, C. Y. Morassuti, S. D. Ozelin, J. K. Daguano, Z. C. Silveira, C. E. Ambrosio, and E. M. Pallone, "3d-printed alumina-zirconia scaffolds via peg-laponite® for bone regeneration," *Materials Today Communications*, p. 115435, 2026.

[11] J. Dávila, M. S. d. Freitas, P. I. Neto, Z. d. C. Silveira, J. V. L. d. Silva, and M. A. d'Ávila, "Software to generate 3-d continuous printing paths for the fabrication of tissue engineering scaffolds," *The International Journal of Advanced Manufacturing Technology*, vol. 84, no. 5, pp. 1671–1677, 2016.

[12] J. Adhikari, M. Perwez, A. Das, and P. Saha, "Development of hydroxyapatite reinforced alginate-chitosan based printable biomaterial-ink. nano struct nano objects. 25: 100630," 2020.

[13] D. Almeida, F. Küppers, A. Gusmão, A. C. Manjua, C. F. Ferreira, C. A. Portugal, J. C. Silva, P. Sanjuan-Alberte, and F. C. Ferreira, "Design of magnetic kappa-carrageenan-collagen bioinks for 3d bioprinting," *Journal of Materials Science*, vol. 59, no. 31, pp. 14 573–14 592, 2024.

[14] P. A. Amorim, M. d'Ávila, R. Anand, P. Moldenaers, P. Van Puyvelde, and V. Bloemen, "Insights on shear rheology of inks for extrusion-based 3d bioprinting," *Bioprinting*, vol. 22, p. e00129, 2021.

[15] J. Karvinen and M. Kellomäki, "Design aspects and

characterization of hydrogel-based bioinks for extrusion-based bioprinting," *Bioprinting*, vol. 32, p. e00274, 2023.

[16] M. P. Gashti, M. Stir, and J. Hulliger, "Synthesis of bone-like micro-porous calcium phosphate/iota-carrageenan composites by gel diffusion," *Colloids and Surfaces B: Biointerfaces*, vol. 110, pp. 426–433, 2013.

[17] H. Herrada-Manchón, M. A. Fernández, and E. Aguilar, "Essential guide to hydrogel rheology in extrusion 3d printing: how to measure it and why it matters?" *Gels*, vol. 9, no. 7, p. 517, 2023.

[18] N. Zandi, E. S. Sani, E. Mostafavi, D. M. Ibrahim, B. Saleh, M. A. Shokrgozar, E. Tamjid, P. S. Weiss, A. Simchi, and N. Annabi, "Nanoengineered shear-thinning and bioprintable hydrogel as a versatile platform for biomedical applications," *Biomaterials*, vol. 267, p. 120476, 2021.

[19] S. Uman, A. Dhand, and J. A. Burdick, "Recent advances in shear-thinning and self-healing hydrogels for biomedical applications," *Journal of Applied Polymer Science*, vol. 137, no. 25, p. 48668, 2020.

[20] A. Wenz, K. Borchers, G. E. Tovar, and P. J. Kluger, "Bone matrix production in hydroxyapatite-modified hydrogels suitable for bone bioprinting," *Biofabrication*, vol. 9, no. 4, p. 044103, 2017.

[21] N. Hassan, T. Ahmad, N. M. Zain, and S. R. Awang, "Identification of bovine, porcine and fish gelatin signatures using chemometrics fuzzy graph method," *Scientific Reports*, vol. 11, no. 1, p. 9793, 2021.

[22] L. Radev, T. Gerganov, H. Georgiev, A. Kolev, V. Vassileva, R. Iankova, and E. Cholakova, "Silk fibroin/calcium phosphate silicate composites: In vitro bioactivity," *Int. J. Mater. Chem*, vol. 3, pp. 8–15, 2013.

[23] N. Thakur and B. Singh, "Designing carrageenan-based functional biomaterials by supra-molecular and covalent interactions for biomedical applications," *Medicine in Novel Technology and Devices*, vol. 24, p. 100338, 2024.

[24] M. A. R. D. Fauzi, P. Pudjiastuti, E. Hendradi, R. T. Widodo, and M. C. I. M. Amin, "Characterization, disintegration, and dissolution analyses of carrageenan-based hard-shell capsules cross-linked with maltodextrin as a potential alternative drug delivery system," *International Journal of Polymer Science*, vol. 2020, no. 1, p. 3565931, 2020.

[25] M. Głab, S. Kudłacik-Kramarczyk, A. Drabczyk, J. Walter, A. Kordyka, M. Godzierz, R. Bogucki, B. Tyliszczak, and A. Sobczak-Kupiec, "Hydroxyapatite obtained via the wet precipitation method and pvp/pva matrix as components of polymer-ceramic composites for biomedical applications," *Molecules*, vol. 26, no. 14, p. 4268, 2021.

[26] A. Sobczak-Kupiec, K. Pluta, A. Drabczyk, M. Włoś, and B. Tyliszczak, "Synthesis and characterization of ceramic-polymer composites containing bioactive synthetic hydroxyapatite for biomedical applications," *Ceramics International*, vol. 44, no. 12, pp. 13 630–13 638, 2018.

[27] A. Sharma, S. Bhat, T. Vishnoi, V. Nayak, and A. Kumar, "Three-dimensional supermacroporous carrageenan-gelatin cryogel matrix for tissue engineering applications," *BioMed research international*, vol. 2013, no. 1, p. 478279, 2013.

[28] A. O. Makarova, S. R. Derkach, A. I. Kadyirov, S. A. Ziganshina, M. A. Kazantseva, O. S. Zueva, A. T. Gubaidullin, and Y. F. Zuev, "Supramolecular structure and mechanical performance of κ-carrageenan–gelatin gel," *Polymers*, vol. 14, no. 20, p. 4347, 2022.

[29] S. V. Dorozhkin, "Calcium orthophosphates in nature, biology and medicine," *Materials*, vol. 2, no. 2, pp. 399–498, 2009.

[30] J. R. Padhi, D. Nayak, A. Nanda, P. R. Rauta, S. Ashe, and B. Nayak, "Development of highly biocompatible gelatin & i-carrageenan based composite hydrogels: In depth physiochemical analysis for biomedical applications," *Carbohydrate polymers*, vol. 153, pp. 292–301, 2016.

[31] N. Rajab, T. Yaakob, B. Ong, M. Hamid, A. Ali, B. Annuar, and S. Inayat-Hussain, "Dna damage evaluation of hydroxyapatite on fibroblast cell l929 using the single cell gel electrophoresis assay." *The Medical Journal of Malaysia*, vol. 59, pp. 170–171, 2004.

[32] Y. Tan, L. Ma, X. Chen, Y. Ran, Q. Tong, L. Tang, and X. Li, "Injectable hyaluronic acid/hydroxyapatite composite hydrogels as cell carriers for bone repair," *International journal of biological macromolecules*, vol. 216, pp. 547–557, 2022.

[33] M. Szymonowicz, M. Korczynski, M. Dobrzynski, K. Zawisza, M. Mikulewicz, E. Karuga-Kuzniewska, B. Zywicka, Z. Rybak, and R. J. Wiglusz, "Cytotoxicity evaluation of high-temperature annealed nanohydroxyapatite in contact with fibroblast cells," *Materials*, vol. 10, no. 6, p. 590, 2017.

[34] Y. W. Oh, S. W. Kang, S. Park, S.-W. Park, and H.-G. Yi, "Collagen/hydroxyapatite hydrogels promote intercellular interactions and osteogenic differentiation," *Journal of Biomedical Materials Research Part B: Applied Biomaterials*, vol. 113, no. 8, p. e35632, 2025.

[35] C. Herranz-Diez, A. Crawford, R. L. Goodchild, P. V. Hatton, and C. A. Miller, "Stimulation of metabolic activity and cell differentiation in osteoblastic and human mesenchymal stem cells by a nanohydroxyapatite paste bone graft substitute," *Materials*, vol. 15, no. 4, p. 1570, 2022.

[36] P. Stolarov, J. De Vries, S. Stapleton, L. Morris, K. Martyniak, and T. J. Kean, "Suitability of gelatin methacrylate and hydroxyapatite hydrogels for 3d-bioprinted bone tissue," *Materials*, vol. 17, no. 5, p. 1218, 2024.

[37] A. Di Stefano, C. Di Marco, F. Toia, M. Trapani, M. Testa, S. Di Leonardo, G. Burriesci, M. Franza, E. Cammarata, A. Cordova *et al.*, "Effect of nanocomposite chitosan/hydroxyapatite ph-induced hydrogels on the osteogenic differentiation of spheroids from adipose stem cells," *International Journal of Biological Macromolecules*, vol. 299, p. 140213, 2025.

[38] N. Demirkol, A. A. Ari, and H. Evlen, "Production and characterization of alumina doped sheep and synthetic hydroxyapatite bioinks," *Journal of the Australian Ceramic Society*, pp. 1–10, 2025.

[39] M. L. Bedell, A. L. Torres, K. J. Hogan, Z. Wang, B. Wang, A. J. Melchiorri, K. J. Grande-Allen, and A. G. Mikos, "Human gelatin-based composite hydrogels for osteochondral tissue engineering and their adaptation into bioinks for extrusion, inkjet, and digital light processing bioprinting," *Biofabrication*, vol. 14, no. 4, p. 045012, 2022.

[40] X. Yu, M. Gholipourmalekabadi, X. Wang, C. Yuan, and K. Lin, "Three-dimensional bioprinting biphasic multicellular living scaffold facilitates osteochondral defect regeneration," *Interdisciplinary Materials*, vol. 3, no. 5, pp. 738–756, 2024.

[41] K. Loukelis, G.-I. Kontogianni, D. Vlassopoulos, and M. Chatzinikolaidou, "Extrusion-based 3d bioprinted gellan gum/poly (vinyl alcohol)/nano-hydroxyapatite composite bioinks promote bone regeneration," *Advanced Healthcare Materials*, vol. 14, no. 19, p. 2500365, 2025.

An Injectable Thermoresponsive Hydrogel Framework: Characterization of Mechanical Compliance and Cellular Response in an In Vitro Stress Model

1st Gabriela Colares Ali Ganem
Department of Chemistry
Aeronautics Institute of Technology
São José dos Campos, Brazil
gabrielacganem@gmail.com

2nd Cristina Pacheco Soares
Research & Development Institute
Universidade do Vale do Paraíba
São José dos Campos, Brazil
cpsoares@univap.br

Abstract*— This study evaluated the potential of a multi-component, thermoresponsive hydrogel framework, designated as SMARTGELP, to serve as a biomimetic matrix for neural tissue engineering applications within an in vitro oxidative stress model relevant to ischemic neural injury. Viscoelastic properties and the dynamic gelation temperature (T_{gel}) were characterized via dynamic oscillatory rheology, while human neuroblastoma cells (SH-SY5Y) were seeded onto the matrix to evaluate 3D cellular metabolic activity and morphological distribution via resazurin reduction assays and inverted phase-contrast microscopy, and were encapsulated within the matrix to evaluate the modulation of hydrogen peroxide-induced oxidative stress via flow cytometry using the CellROX® Orange fluorogenic probe. Rheological characterization indicated a dynamic sol-gel transition at 32 °C and a storage modulus (G') of approximately 740 Pa at 37 °C, yielding a value close to the range reported for brain tissue stiffness in the literature. Metabolic evaluation via resazurin reduction demonstrated that metabolic activity remained above 95% of the negative control at 24 and 48 hours, displaying stable, controlled 3D kinetics (97% to 103%) compared to the accelerated metabolic activity of standard 2D cultures. Morphological analyses revealed the initial establishment of dense 3D cellular clusters at 24 hours, which transited into a more diffuse state at 72 hours, tracking the progressive physical restructuring and controlled erosion of the hydrogel network over time. Furthermore, flow cytometry quantification demonstrated a symmetrical retraction of the population curves under acute injury conditions, with the SMARTGELP matrix maintaining lower baseline oxidative activation parameters (ROS(%) of 4,28% at 24 hours and 5,40% at 72 hours) compared to the positive controls. The convergence of these rheological, morphological, and biochemical indicators suggests that the design of the SMARTGELP architecture can address multiple parameters of structural support, mechanical compliance, and redox baseline maintenance, indicating its potential for further investigation as an in vitro oxidative stress model relevant to ischemic neural injury.

Keywords*— *neural tissue engineering, thermoresponsive hydrogel, SMARTGELP, SH-SY5Y cells, mechanical compliance, oxidative stress modulation.

I. INTRODUCTION

Neural tissue engineering constitutes a major area of regenerative medicine, focusing on the functional restoration of the central nervous system (CNS) following injuries such as ischemic stroke, traumatic brain injury, or neurodegenerative diseases [1]. Ischemic stroke, which accounts for more than 80% of all stroke cases worldwide, triggers a devastating pathological cascade in the ischemic penumbra [2]. This process is initiated by an abrupt reduction in blood flow and an immediate failure of anaerobic glycolysis, which disrupts transmembrane ion pumps and leads to severe bioenergetic collapse [2], [3]. Importantly, this homeostatic disorder forces a massive extracellular accumulation of glutamate, inducing catastrophic neuronal death through the overstimulation of ionotropic receptors and calcium overload [2]. Concomitantly, this cascade stimulates the excessive production of reactive oxygen species (ROS), which exaerbates oxidative damage, brain edema, and neuronal necrosis[2].

Thus, a fundamental barrier to neural repair is the rapid degradation of the brain's native extracellular matrix (ECM) within these ischemic and necrotic zones [2]. The destruction of this specialized structural support creates a hostile microenvironment that inhibits cell survival and impedes the cellular migration required for the functional recovery of the tissue [2]. To overcome this limitation, research has turned toward developing biomaterials that simulate the cerebral microenvironment and resolve the inherent chemical-mechanical mismatch between rigid interface materials and the delicate neural tissue [4]. In this context, injectable smart hydrogels have emerged as promising frameworks due to their ability to fill the injury cavity and offer a three-dimensional (3D) scaffold that mimetizes the native ECM [4].

The brain parenchyma is exceptionally soft, establishing an ideal stiffness range between 0.1 and 1.0 kPa, where cell survival and neurite outgrowth are optimized [5]. Stiffness values outside this range promote reactive gliosis - the primary inhibitory factor in regeneration -, redirecting the fate of neural stem cells toward the astroglial lineage [6]. Hence, the delivery strategy requires thermoresponsive systems that transition from an injectable liquid to a solid gel state in situ in response to body temperature (37 °C) [7]. However, the utility of the conventional thermoresponsive polymer Poloxamer 407 (F127) is limited by its restricted mechanical strength and accelerated degradation [8]. Similarly, the rapid structural loss of pure natural polymers, such as hyaluronic acid (HA), results in the premature collapse of the scaffold [9].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

To overcome these structural and stability limitations, the rational combination of F127 with HA and gelatin (GL) represents a strategic design. HA, a key component of the cerebral ECM, increases stiffness and attenuates the inflammatory response [10], while GL confers biological properties that stimulate cell adhesion, proliferation, and neurite outgrowth [11]. Considering the demand for platforms that promote mechanical mimicry and bioactivity with enhanced stability, this work proposes the evaluation of an in vitro oxidative stress model relevant to ischemic neural injury, using an injectable and thermosensitive hydrogel, designated as SMARTGELP (F127-HA-GL), carefully optimized at a mass ratio of 70/20/10 to try achieve a neuromimetic elastic modulus within the 0.1 to 1.0 kPa range. Therefore, this study aims to validate the potential of SMARTGELP as a rationally designed platform, evaluating its capacity to modulate and attenuate the inflamation effects of ischemic injuries through biochemical assays involving hydrogen peroxide (H_2O_2).

II. Materials and Methods

A. Hydrogel Preparation

The thermoresponsive SMARTGELP hydrogel was prepared as a structured polymer blend to achieve a final composition of 70/20/10 %w/w of Poloxamer 407 (F127, Sigma-Aldrich, USA), medium molecular weight hyaluronic acid (HA; MW = 380 kDa, #9067-32-7, Engenharia das Essências, Brazil), and gelatin (GL, Sigma-Aldrich, USA), using sodium carboxymethylcellulose (CMC, Sigma-Aldrich, USA) as a compatibilizing agent.

Precursor solutions were prepared individually and sequentially to promote solubilization of each component prior to mixing, aiming to minimize aggregate formation and support the homogeneity of the final system. A 20% w/w F127 aqueous solution was obtained by adding the polymer powder to chilled deionized water under constant magnetic stirring in an ice bath for 1 hour, followed by storage under refrigeration for 12 hours to allow complete dissolution. Concurrently, a 1% w/w gelatin solution to deionized water was prepared with the addition of 0.01% w/w CMC, under magnetic stirring at 50 °C until homogenization. A 2% w/w HA solution to deionized water was prepared at room temperature under constant stirring until polymer dispersion was achieved.

The SMARTGELP formulation was prepared under cooling to preserve the sol state of the components and prevent premature gelation during processing. Calculated volumes of the previously stabilized HA and GL/CMC solutions were gradually added to the F127 solution, maintained in an ice bath under continuous magnetic stirring. After visual homogenization of the system, the precursor mixture was stored under refrigeration. The formulation pH (~7.4) was determined at room temperature using a calibrated pH meter (Hanna HI 221, Hanna Instruments, Woonsocket, RI, USA).

B. Rheological Characterization

The viscoelastic properties and the quantitative gelation temperature (T_{gel}) were determined by dynamic oscillatory rheology using a stress-controlled rheometer (Rheostress RS600, Thermo Fisher Scientific, Massachusetts, USA) equipped with a cone-and-plate geometry (diameter: 20 mm; angle: 2º). The evolution of the storage (G') and loss (G") moduli was monitored during a temperature ramp from 25 to 47 °C, under a constant heating rate of 1 °C·min^{-1}. To ensure that measurements remained within the linear viscoelastic region, where G' and G" are independent of strain amplitude, a frequency of 1 Hz and an oscillatory strain of 0.01 (1%) were fixed, mitigating the probability of structural disruption of the forming network during the sol-gel transition [12]. The rheological T_{gel} was defined by the crossover point where G' = G", corresponding to the transition from a predominantly viscous behavior (G" > G') to a predominantly elastic behavior (G' > G").

C. Cell Viability Analysis

The in vitro biological evaluation of SMARTGELP was performed using SH-SY5Y cells (ATCC - CRL-2266) to characterize metabolic activity and morphological evolution in 3D culture.

Prior to the viability assays, the SMARTGELP hydrogel was sterilized by filtration through a 0.22 µm membrane (Millipore, USA), a procedure adopted given that F127-based hydrogels may undergo structural changes when subjected to autoclaving or ultraviolet light, and was kept under refrigeration to preserve its flow state until use. Predefined hydrogel volume (50 µL) was deposited into 24-well culture plates (Kasvi, Brazil, #K12-024) and kept at 37 °C for 5 minutes to allow in situ gelation prior to cell seeding.

Following the sol-gel transition, SH-SY5Y cells were seeded onto the polymeric matrix at a density of $5x10^4$ cells per well, followed by the addition of 500 µL of culture medium consisting of Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F12, Gibco, USA, #12400024), supplemented with 10% fetal bovine serum (FBS, Gibco, USA, #12657-029) and 1% penicillin-streptomycin (Gibco, USA, #15140-122).

Cellular metabolic activity was evaluated at 24 and 48 hours after seeding using the resazurin reduction assay, adopting the following control groups: negative control (cells incubated in culture medium) and positive control (cells incubated in culture medium containing dimethyl sulfoxide, DMSO, 1:1 v/v with culture medium, i.e., 50% DMSO; R&R, Brazil, CAS: 67-68-5). Data were expressed as a percentage of the negative control metabolic activity, as detailed in the Statistical Analysis section.

D. Oxidative Stress Assay

To evaluate the cytoprotective potential and the modulation of oxidative stress, SH-SY5Y cells seeded at a density of $5x10^4$ cells/well were differentiated with 10 µM retinoic acid (Sigma-Aldrich, USA, #R2625) and 0.5% FBS for 5 days.

Oxidative stress was induced by exposing the cultures to 200 µM H_2O_2 (#L0032-P1L, Molekular Química) in differentiation medium for 1 hour. Upon completion of the injury interval, the wells were washed with phosphate buffered saline for treatment application, which consisted of encapsulating $5x10^4$ cells in 50 µL of SMARTGEL per well.

Following the sol-gel transition process, 500 µL of differentiation medium was added per well, and the plate was incubated at 37 °C and 5% CO_2. The in vitro assays were monitored at 24 and 72-hour intervals using n = 4, as detailed in the Statistical Analysis section. The production of reactive oxygen species (ROS) was quantified using the CellROX® Orange reagent (Invitrogen, USA, #C10443) according to the protocol provided by the manufacturer, remaining in incubation for 30 minutes prior to verification via flow cytometry (BD Accuri C6 Plus, Becton Dickinson, USA). Morphological modifications were monitored using an

inverted microscope (Olympus IX51, Olympus Corporation, Japan) coupled to a DP30BW digital camera unit.

E. Statistical Analysis

All experiments were performed with n = 4, and results are expressed as mean ± standard deviation. Statistical analysis was performed using one-way ANOVA followed by Tukey's post hoc test for multiple comparisons, with statistical significance set at *p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001.

III. RESULTS AND DISCUSSION

A. Sol-Gel Transition

The viscoelastic profile of the SMARTGELP matrix as a function of temperature suggested a three-phase behavior characteristic of F127-based systems (Fig. 1)[13]. At lower temperatures, the matrix remained in a sol state, exhibiting low viscoelastic moduli values and a predominantly viscous character (G" > G').

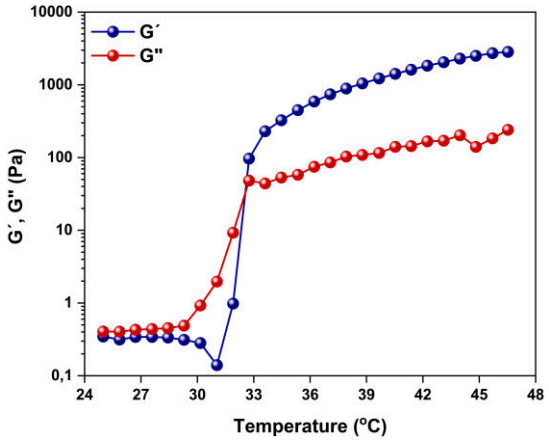


Fig. 1. Viscoelastic profile of the SMARTGELP matrix as a function of temperature. The dynamic oscillatory response illustrates the temperature-dependent storage modulus (blue) and loss modulus (red). The crossover point (G' = G") recorded at 32 °C identifies the rheological gelation threshold (T_{gel}), transitioning from a predominantly viscous behavior (G" > G') to a structured elastic network (G' > G').

Subsequent thermal enhancement induced an elevation in both moduli, a phenomenon associated with the progressive thermal aggregation and self-assembly of the F127 unimers into micellar networks [13].

The dynamic T_{gel} was recorded at 32 °C. This intersection marked the physical transition from a predominantly viscous behavior to a structured elastic arrangement (G' > G") [13].

At the physiological temperature of 37 °C, the SMARTGELP reached a storage modulus of approximately 740 Pa, approaching the lower range of stiffness values reported for brain tissue in the literature (0.1 to 1.0 kPa), although direct comparison is limited by differences in measurement technique and frequency between studies [5]. Literature indicators suggest that substrate compliance within this neuromimetic range can support neuronal phenotype differentiation while mitigating adverse astroglial responses or reactive gliosis associated with excessive rigidity [6], [14].

B. Cell Viability and Metabolic Activity

The biological performance of the SMARTGELP matrix was evaluated using the SH-SY5Y cell line seeded onto the three-dimensional matrix, Fig. 2. For the 50 µL hydrogel volume condition, the cellular metabolic activity, assessed via the resazurin reduction assay, remained above 95% at both 24 and 48 hours post-seeding, demonstrating no statistically significant reduction in comparison to the negative control group, suggesting the SMARTGELP architecture allowed a sufficient diffusive flux of oxygen, nutrients, and metabolites.

Comparatively, while the negative control exhibited an accelerated proliferation profile from 24 to 48 hours (~100% to ~135%) typical of standard two-dimensional monolayer cultures, the metabolic activity within the SMARTGELP remained stable and controlled, varying from ~97% to ~103%. This behavior suggests that the three-dimensional architecture of the hydrogel matrix modulates cell growth kinetics compared to the accelerated proliferation observed in the 2D negative control [15].

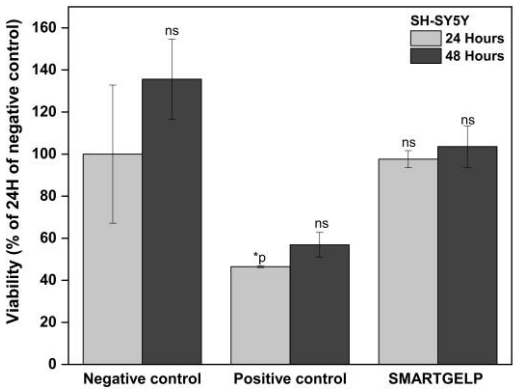


Fig. 2. Evaluation of SH-SY5Y cell metabolic activity and growth kinetics within the SMARTGELP matrix after 24 hours (light gray) and 48 hours (dark gray) of incubation. Metabolic activity was quantified via the resazurin reduction assay and expressed as a percentage relative to the 24-hour negative control group (%). As resazurin reduction reflects metabolic activity rather than a direct measure of cell viability or proliferation, these values should be interpreted accordingly. Experimental groups comprise the standard culture condition showing accelerated 2D monolayer proliferation (negative control), the cytotoxic setup (positive control), and the 3D engineered framework establishing homeostatic metabolic maintenance (SMARTGELP). Data are presented as mean ± standard deviation.

This stable metabolic outcome can be associated with the integration of GL and HA into the synthetic F127 base. While pure poloxamer matrices frequently exhibit non-adhesive characteristics, the inclusion of GL provides functional RGD sequences known to mediate cell anchoring via integrin receptors [16].

C. Cytocompatibility Analysis

The temporal evaluation of intracellular oxidative profiles quantified by flow cytometry (Fig. 3) shows a visible alignment with the cellular architecture monitored via inverted phase-contrast microscopy (Fig. 4), providing a physical context to the redox parameters.

At the 24-hour interval, the physiological baseline established by the negative control group revealed a standard two-dimensional monolayer distribution, where cells displayed typical flattened and spread morphologies adhering to the substrate, corresponding to a basal activation rate of 6,24% within the gated region. In contrast, the positive control group manifested a rightward shift along the biexponential fluorescence axis, registering an increased fraction of hyperproductive reactive oxygen species events at 20,48%. This shift matches the phase-contrast observations of sparse cell clusters and early indications of cell membrane alteration, which are consistent with the triggering of

deleterious pathways linked to mitochondrial membrane potential disruption or exacerbated cytoplasmic oxidase activity under acute stress [17].

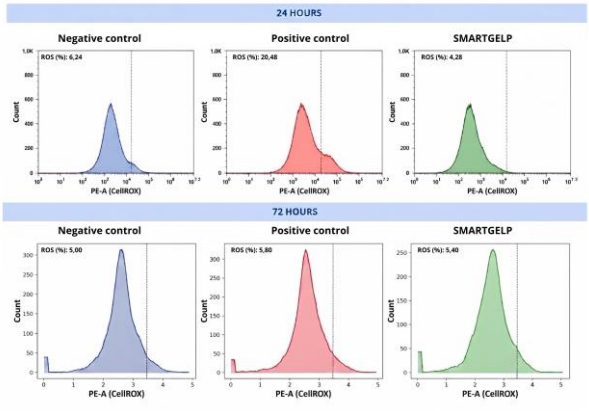


Fig. 3. Flow cytometry histograms of intracellular reactive oxygen species (CellROX® Orange fluorescence intensity) in SH-SY5Y cells across the negative control, positive control, and SMARTGELP groups, at 24 and 72 hours of exposure.

Among the evaluated conditions at 24 hours, the SMARTGELP matrix exhibited a substantial density of uniform, tightly packed 3D cellular clusters, which accompanied a symmetrical retraction of the population curve toward lower fluorescence channels, reaching the lowest activation level at 4,28%. This initial convergence suggests that the 3D spatial microenvironment, mechanical compliance, and biochemical anchoring sites within the hybrid framework support cell aggregation while maintaining lower intracellular oxidative stress levels, indicating potential radical scavenging activities or a mitigation of detachment-induced ROS production during the acute phase [18], [19].

By 72 hours, microenvironmental modifications resulting from progressive nutrient depletion and metabolic byproduct accumulation accompany a compression of the kinetic separation boundaries between the groups [19]. The long-term positive control group manifested a pronounced reduction in visible cell density and increased fragmentation, while the late negative control group maintained stretched, polarized morphologies with a lower overall cell density compared to its 24-hour baseline. These architectural variations correspond to the attenuated absolute differences measured in their late-stage ROS(%) rates, which reached 5,80% for the positive control and 5,00% for the negative control. This parallel may reflect the early acute oxidative imbalance being followed by metabolic clearance of the original stress or by an adaptive selection of surviving cells exhibiting compensatory scavenger regulation; however, given the pronounced reduction in cell density observed in the positive control at 72 hours (Fig. 4), part of this decrease in ROS(%) may also reflect a smaller surviving cell population being analyzed, rather than redox recovery alone [17].

Under this same long-term perspective, the large cellular clusters within the SMARTGELP framework transited into a more diffuse state within the volume of the hydrogel, aligning with the expected physical restructuring and controlled erosion of the matrix over time. This architectural transition proceeded under stable redox parameters, as the SMARTGELP group maintained a long-term ROS(%) of 5,40%, remaining closely aligned with baseline homeostatic levels. This behavior points to chronic cytocompatibility and suggests that incorporating the components into the SMARTGELP network assists in controlling cellular exposure rates, ensuring that matrix alteration proceeds without inducing delayed oxidative toxicity over time [20], [21].

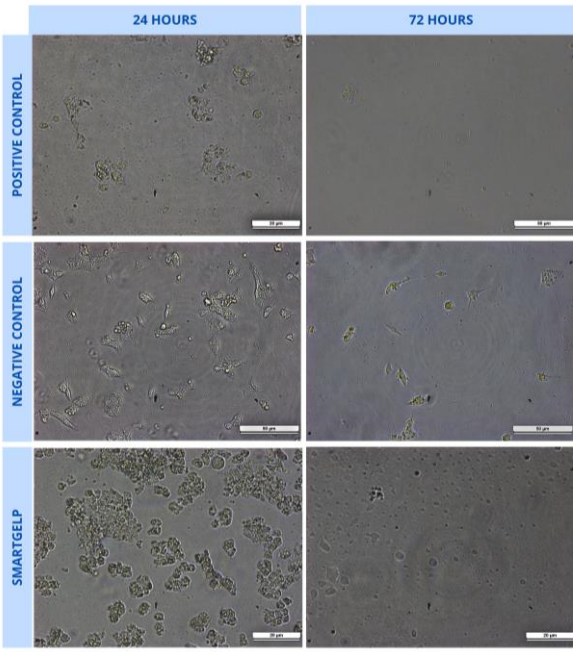


Fig. 4. Inverted phase-contrast light microscopy micrographs of SH-SY5Y cells after 24 hours and 72 hours of incubation across the experimental groups: positive control, negative control, and SMARTGELP matrix. At 24 hours, micrographs illustrate sparse cellular distribution in the positive control, standard flattened 2D monolayer adherence in the negative control, and dense 3D cellular clustering within the SMARTGELP network. At 72 hours, the images show a pronounced cell reduction and fragmentation in the Positive control, maintained stretched morphologies under lower density in the Negative control, and a structural transition to a diffuse cellular state accompanying matrix rearrangement in the SMARTGELP framework.

It should be noted that the morphological observations reported here, including cell clustering and its progressive diffusion over time, do not by themselves confirm a specific neural phenotype or a defined mechanism of action; cell aggregation can also arise from factors such as limited substrate adhesion, cellular stress, or matrix contraction, and quantitative neural marker analysis would be needed to support such an interpretation. Taken together, these parallel morphological and biochemical indicators suggest that the combination of structural shielding with localized metabolic stabilization in the SMARTGELP framework may contribute to a more favorable microenvironment, offering preliminary support for further investigation of tissue engineering interventions relevant to ischemic neural injury.

IV. Conclusion

In this work, the SMARTGELP matrix was developed as a multifunctional, thermoresponsive framework aimed at neural tissue engineering applications.

The structural and rheological characterization suggests that the hybrid polymeric network approaches stiffness values reported for brain tissue in the literature, although direct comparison is limited by differences in measurement technique and frequency between studies. By transitioning to a structured elastic gel below body temperature and positioning its compliance within the neuromimetic stiffness window, the matrix provides a soft support designed to interface with neural tissue model systems.

Biologically, the platform indicated a capacity to accompany changes in cell distribution alongside

microenvironmental conditions. The initial observation of three-dimensional cellular aggregates followed by a transition into a more diffuse state suggests that the progressive physical remodeling of the matrix can proceed under monitored redox parameters. By maintaining stable intracellular reactive oxygen species profiles over the evaluated culture periods, the hybrid framework suggested compatibility with the cellular environment.

This study has some limitations worth noting. The current in vitro configuration represents an oxidative stress model relevant to ischemic neural injury rather than a complete model of ischemic stroke, and the SH-SY5Y monoculture does not reproduce the cellular and structural complexity of the injured cerebral tissue. Establishing a direct cytoprotective effect would require a factorial design comparing cells with and without hydrogel exposure, in the presence and absence of H_2O_2, within the same experiment. Additionally, the resazurin reduction assay reflects metabolic activity rather than a direct measure of cell viability or proliferation, and a hydrogel-only control (without cells) could help clarify the influence of the matrix on reagent diffusion and signal. Future work incorporating these additional experimental groups, together with quantitative neural marker analysis, could further support the interpretation of the platform's protective potential.

Consequently, the convergence of these rheological, morphological, and biochemical indicators suggests that the design of the SMARTGELP architecture may address parameters of structural support, mechanical compliance, and redox baseline maintenance, indicating its potential for further investigation as an in vitro oxidative stress model relevant to ischemic neural injury.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - 88887.711128/2022-00.

REFERENCES

[1] S. K. Feske, “Ischemic Stroke”, *Am. J. Med.*, v. 134, n. 12, p. 1457–1464, Dec. 2021, doi: 10.1016/j.amjmed.2021.07.027.

[2] C. D. Maida, R. L. Norrito, M. Daidone, A. Tuttolomondo, and A. Pinto, “Neuroinflammatory Mechanisms in Ischemic Stroke: Focus on Cardioembolic Stroke, Background, and Therapeutic Approaches”, *Int. J. Mol. Sci.*, v. 21, n. 18, p. 6454, Sep. 2020, doi: 10.3390/ijms21186454.

[3] T. Zieliński *et al.*, “Changes in nanomechanical properties of single neuroblastoma cells as a model for oxygen and glucose deprivation (OGD)”, *Sci. Rep.*, v. 12, n. 1, p. 16276, Sep. 2022, doi: 10.1038/s41598-022-20623-8.

[4] A. Gopalakrishnan, S. A. Shankarappa, and G. K. Rajanikant, “Hydrogel Scaffolds: Towards Restitution of Ischemic Stroke-Injured Brain”, *Transl. Stroke Res.*, v. 10, n. 1, p. 1–18, Feb. 2019, doi: 10.1007/s12975-018-0655-6.

[5] C. Zhang, C. Liu, and H. Zhao, “Mechanical properties of brain tissue based on microstructure”, *J. Mech. Behav. Biomed. Mater.*, v. 126, p. 104924, Feb. 2022, doi: 10.1016/j.jmbbm.2021.104924.

[6] Y. Liang *et al.*, “Impact of hydrogel stiffness on the induced neural stem cells modulation”, *Ann. Transl. Med.*, v. 9, n. 24, p. 1784–1784, Dec. 2021, doi: 10.21037/atm-21-6189.

[7] Y. Bai *et al.*, “Advancements in Hydrogel Application for Ischemic Stroke Therapy”, *Gels*, v. 8, n. 12, p. 777, Nov. 2022, doi: 10.3390/gels8120777.

[8] L. Klouda and A. G. Mikos, “Thermoresponsive hydrogels in biomedical applications”, *Eur. J. Pharm. Biopharm.*, v. 68, n. 1, p. 34–45, Jan. 2008, doi: 10.1016/j.ejpb.2007.02.025.

[9] K. Výborný *et al.*, “Genipin and EDC crosslinking of extracellular matrix hydrogel derived from human umbilical cord for neural tissue repair”, *Sci. Rep.*, v. 9, n. 1, p. 10674, Jul. 2019, doi: 10.1038/s41598-019-47059-x.

[10] P. Sensharma, G. Madhumathi, R. D. Jayant, and A. K. Jaiswal, “Biomaterials and cells for neural tissue engineering: Current choices”, *Mater. Sci. Eng. C*, v. 77, p. 1302–1315, Aug. 2017, doi: 10.1016/j.msec.2017.03.264.

[11] M. Shen *et al.*, “Gelatin Methacrylic Acid Hydrogel-Based Nerve Growth Factors Enhances Neural Stem Cell Growth and Differentiation to Promote Repair of Spinal Cord Injury”, *Int. J. Nanomedicine*, v. 19, p. 10589–10604, Oct. 2024, doi: 10.2147/IJN.S480484.

[12] J. Cho, M.-C. Heuzey, A. Bégin, and P. J. Carreau, “Physical Gelation of Chitosan in the Presence of β-Glycerophosphate: The Effect of Temperature”, *Biomacromolecules*, v. 6, n. 6, p. 3267–3275, Nov. 2005, doi: 10.1021/bm050313s.

[13] M. Kroger, “Thermogelation Analysis Of F127-Water Mixtures By Physical Chemistry Techniques”, 2010, doi: 10.3933/APPLRHEOL-20-52081.

[14] S. K. Seidlits *et al.*, “The effects of hyaluronic acid hydrogels with tunable mechanical properties on neural progenitor cell differentiation”, *Biomaterials*, v. 31, n. 14, p. 3930–3940, May 2010, doi: 10.1016/j.biomaterials.2010.01.125.

[15] N. Madrer *et al.*, “Neural Metabolic Networks: Key Elements of Healthy Brain Function”, *J. Neurochem.*, v. 169, n. 6, p. e70084, Jun. 2025, doi: 10.1111/jnc.70084.

[16] D.-M. Viezuina *et al.*, “Gelatin-Based Hydrogels for Peripheral Nerve Regeneration: A Multifunctional Vehicle for Cellular, Molecular, and Pharmacological Therapy”, *Gels*, v. 11, n. 7, p. 490, Jun. 2025, doi: 10.3390/gels11070490.

[17] B. Halliwell and J. M. C. Gutteridge, *Free Radicals in Biology and Medicine*. Oxford University Press, 2015. doi: 10.1093/acprof:oso/9780198717478.001.0001.

[18] B. Kalyanaraman *et al.*, “Measuring reactive oxygen and nitrogen species with fluorescent probes: challenges and limitations”, *Free Radic. Biol. Med.*, v. 52, n. 1, p. 1–6, Jan. 2012, doi: 10.1016/j.freeradbiomed.2011.09.030.

[19] H. Sies, “Oxidative stress: a concept in redox biology and medicine”, *Redox Biol.*, v. 4, p. 180–183, Apr. 2015, doi: 10.1016/j.redox.2015.01.002.

[20] P. Wardman, “Fluorescent and luminescent probes for measurement of oxidative and nitrosative species in cells and tissues: Progress, pitfalls, and prospects”, *Free Radic. Biol. Med.*, v. 43, n. 7, p. 995–1022, Oct. 2007, doi: 10.1016/j.freeradbiomed.2007.06.026.

[21] I. Popescu, M. Constantin, M. Bercea, B.-P. Coşman, D. M. Suflet, and G. Fundueanu, “Poloxamer/Carboxymethyl Pullulan Aqueous Systems—Miscibility and Thermogelation Studies Using Viscometry, Rheology and Dynamic Light Scattering”, *Polymers*, v. 15, n. 8, p. 1909, Apr. 2023, doi: 10.3390/polym15081909.

Development of a Low-Cost Food 3D Printing Platform for Educational Applications

1st Felipe de Mendonça Pereira
Control and Automation Engineering Undergraduate Student
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, MG, Brazil
mendonca.felipe@gea.inatel.br

2nd Guilherme Moreira Virgulino
Biomedical Engineering Undergraduate Student
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, MG, Brazil
guilherme.virgulino@geb.inatel.br

3rd Fabiano Valias de Carvalho
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, MG, Brazil
fabiano@inatel.br

4th Maysa Costa Alves
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, MG, Brazil
maysa.alves@inatel.br

***Abstract*—This paper presents the development of a low-cost food 3D printer for educational purposes, achieved by adapting an Ender 3 V3 SE 3D printer. The original fused filament fabrication extrusion system was replaced with a redesigned extrusion head configured to operate with a 20 ml Luer Lock syringe, utilizing a trapezoidal lead screw driven by a stepper motor. Three linear guide axes handle axis movement. An ESP32 running a dedicated, specific program was used for bed leveling. The printer uses a 1-millimeter nozzle for food extrusion and was initially tested with pastes and doughs, such as mashed potatoes, cookie dough, and peanut butter. The project plans future upgrades for temperature control, enabling the printing of products such as chocolate. The total cost of the adaptation was approximately R$ 4,000.00. This work was developed at the Instituto Nacional de Telecomunicações (INATEL) as part of an undergraduate research project (Iniciação Científica) and will be presented at the Inatel Technological Fair (FETIN).**

***Index Terms*—Food 3D printing; additive manufacturing; syringe extrusion; educational prototype; rapid prototyping.**

I. Introduction

Food 3D printing, a branch of additive manufacturing, has emerged as a promising technology with the potential to expand the possibilities for production, customization, and the development of nutritional matrices. This technology enables the creation of complex and customized food structures, offering new opportunities for food design and, crucially, for clinical nutrition. The ability to precisely control the composition, macrostructure, and geometry of the ingredients enables the development of personalized diets for hospital patients, easy-to-swallow formulations aimed at treating dysphagia, and systems for the controlled release of nutrients and bioactive compounds [1], [2].

Additive manufacturing has consolidated itself over the last few decades in fields such as engineering, medicine, and industrial design, being widely used for rapid prototyping and the fabrication of components with complex geometries. Its application in the health and nutrition sector represents a natural evolution of this technology, driven by the growing demand for customized, on-demand products. Among the different methods used for food 3D printing, syringe extrusion stands out as one of the most versatile techniques for semi-solid food materials, such as hospital purees, enriched gels, and doughs [3]. In this process, the material is deposited layer by layer through an extrusion nozzle, allowing precise control over the shape and dimensions of the final product.

Despite the progress observed in recent years, the adoption of this technology is still limited by the high cost of commercial equipment and the complexity of its implementation. Most food 3D printers available on the market are intended for industrial applications or specialized research, making them inaccessible to educational institutions, small laboratories, and researchers with limited financial resources [4]. Furthermore, obtaining adequate printing parameters and controlling the rheological properties of food materials remain significant challenges to ensuring extrusion quality and process repeatability.

Given this scenario, this paper presents the development of a low-cost food 3D printer achieved by adapting a commercial Creality Ender-3 V3 SE filament printer (FDM technology). The main contribution of this project lies in the development of a new print head, completely redesigned to operate with a 20 mL Luer Lock syringe, while preserving the printer's original electronic structure and firmware. The developed mechanical assembly includes specific brackets for the syringe, a drive motor, positioning sensors, and an extrusion system based on a trapezoidal lead screw and linear guides. Since replacing the original print head made it impossible to use the equipment's native calibration sensors, an auxiliary system based on an

ESP32 microcontroller was developed. This system is responsible for automatic bed leveling without requiring complex modifications to the manufacturer's original firmware. Additionally, adaptations were made to the PrusaSlicer software parameters, enabling its use for slicing and parameterizing the food printing process.

A. Motivation and Objectives

Food 3D printing is a technology with strong potential for applications in clinical research, nutritional product development, and educational activities. However, the high cost of commercial platforms remains one of the primary limiting factors for its widespread adoption in educational institutions and research centers, especially in undergraduate laboratories, where resources allocated for equipment acquisition are often restricted.

In this context, adapting a commercial FDM 3D printer emerges as an economically viable alternative to enable experimental activities using components widely available on the market. In addition to reducing the initial investment, this approach facilitates maintenance, component replacement, and future platform modifications—features that are particularly relevant in academic environments.

To demonstrate the economic viability of the proposed solution, Table I presents a comparison between the approximate costs of commercial food 3D printing platforms and the prototype developed in this work. It is observable that utilizing an architecture based on the adaptation of a commercial FDM printer allows for a significant reduction in acquisition costs, expanding access to this technology for teaching and research applications.

Table I
COST COMPARISON BETWEEN COMMERCIAL PLATFORMS AND THE DEVELOPED PROTOTYPE.

Platform / Equipment	**Approximate Cost**
Foodini (Natural Machines) [5]	R$ 30,000.00 to R$ 33,000.00
byFlow Focus [6]	R$ 25,000.00 to R$ 40,000.00
Cocoa Press [7]	R$ 8,000.00 to R$ 20,000.00
Developed prototype	**R$ 4,000.00**

Source: Adapted by the authors (2026).

It can be observed that the developed prototype features a cost 50% lower than the lowest-priced commercial platform considered, and is up to approximately 85% more affordable when compared to top-of-the-line industrial systems. This significant reduction was achieved by repurposing the mechanical and electronic structure of the commercial FDM printer, concentrating the investment strictly on the components necessary for the controlled extrusion of semi-solid materials.

In addition to the comparison with commercial equipment, the cost obtained also proved to be competitive against open-source platforms described in the literature. The Printer.HM system has an approximate cost of £900 for its basic configuration, while the adapter developed by Rogers [8] requires approximately US$600 in components, excluding the base printer.

The central objective of this study is to develop and validate a low-cost food 3D printing platform aimed at educational and research applications in additive food manufacturing, helping to lower the financial barriers to accessing this technology. To achieve this, we propose adapting a commercial FDM printer through the development of a syringe extrusion system, implementing an auxiliary automatic bed leveling system using an ESP32, establishing printing parameters for semi-solid food materials, and conducting experimental validation of the developed platform.

II. METHODOLOGY

A. Base Printer

The development of the food 3D printer was carried out by adapting a commercial *Creality Ender-3 V3 SE* 3D printer. This model was chosen due to its low cost, wide availability of components, and robust mechanical structure—features that facilitate modifications and future expansions of the project. The printer features a Cartesian architecture and a motion system suitable for applications requiring precise axis positioning.

To enable food printing, the original fused deposition modeling (FDM) extrusion assembly was removed, while preserving the printer's mechanical structure and its original electronic board. Consequently, all modifications were concentrated on developing a new extrusion system, allowing the equipment's original firmware to be maintained.

B. Mechanical Design

The entire mechanical design was developed using *Autodesk Fusion 360* software, in which all the parts required for the printer's adaptation were modeled.

Key components developed include the syringe mounting bracket, the drive motor bracket, the attachment elements for the X-axis carriage, the positioning sensor mounts, and the components responsible for the syringe plunger movement system. The design was conceived modularly, allowing for easy disassembly, maintenance, and future modifications.

Following the modeling stage, all parts were fabricated using 3D printing and assembled onto the original structure of the Ender-3 V3 SE.

Figure 1 shows the syringe bracket modeled and printed.



Fig. 1. Developed syringe bracket for printer adaptation. Source: The authors (2026).

C. Extrusion System

The extrusion system was developed to use disposable *20 mL Luer Lock syringes*, which serve as the reservoir for the food material.

The plunger is driven by a *NEMA 17* stepper motor (the standard motor used in additive manufacturing applications) coupled to a *trapezoidal lead screw*, which converts the rotational motion of the motor into linear displacement. This mechanism allows precise control over the amount of material extruded during printing.

To ensure the alignment of the moving assembly and prevent lateral forces on the syringe, *three linear guides* were used to maintain uniform displacement throughout the entire plunger stroke.

An interchangeable extrusion nozzle was installed at the tip of the syringe, allowing the use of different diameters depending on the food material being used.

D. Leveling System

During the development of the new print head, it was found that the original position of the *BLTouch* sensor became incompatible with the syringe extrusion system. If it remained fixed during printing, the sensor could collide with the deposited material or with components of the new print head.

To solve this problem, an auxiliary system was developed based on an **ESP32** microcontroller, used *exclusively during the automatic bed leveling procedure*.

Before printing begins, the ESP32 mechanically positions the BLTouch sensor (Figure 2) into the appropriate position to perform the mesh bed leveling. Once this procedure is complete, the sensor is retracted to a safe position, remaining away from the printing area throughout the entire printing process. This preserves the functionality of the printer's original leveling system without interfering with the geometry of the new extrusion head.

This solution made it possible to keep the original firmware of the Ender-3 V3 SE, requiring only a complementary mechanical and electronic adaptation for the temporary actuation of the sensor.



Fig. 2. BLTouch sensor positioning system using an ESP32. Source: The authors (2026).

E. PrusaSlicer Configuration

Since no modifications were made to the printer's firmware, the necessary adaptations for food printing were focused on configuring the *PrusaSlicer* software. Only the print parameters related to the extrusion process were modified, including the initialization G-code (*Start G-code*) and termination G-code (*End G-code*), alongside specific parameters for the syringe extrusion system, such as plunger feed rate, flow factor, and layer height. These modifications enabled the generation of toolpaths compatible with the new print head while preserving the printer's original electronics and firmware, in line with similar adaptations reported in the literature [9]–[11].

F. Mathematical Modeling of Extrusion Flow

To properly configure the extrusion flow in PrusaSlicer, it was necessary to establish the relationship between the plunger feed rate of the syringe and the material deposition speed onto the platform. The volumetric flow modeling was based on the principle of mass conservation for incompressible fluids, which is widely used in the calibration of syringe extrusion systems [12], [13].

Let us consider a syringe with an internal diameter $D_{Syringe}$ (mm) whose plunger moves at a linear speed $v_{Syringe}$ (mm/s). The cross-sectional area of the syringe is given by:

$$A_{Syringe} = \frac{\pi \cdot D_{Syringe}^2}{4} \tag{1}$$

The volumetric flow rate Q (mm³/s) generated by the displacement of the plunger is, therefore, the product of the area and the displacement per unit of time [12], [14]:

$$Q = A_{Syringe} \cdot v_{Syringe} = \frac{\pi \cdot D_{Syringe}^2}{4} \cdot v_{Syringe} \tag{2}$$

On the other hand, during deposition, the extruded material forms a continuous bead on the platform, whose volume per unit of time can be described by the geometry of the printed line. Given that w is the extruded line width (mm), h is the layer height (mm), and v_{print} is the travel speed of the print head (mm/s), the volume deposited per second is expressed as [11], [12]:

$$Q = w \cdot h \cdot v_{print} \tag{3}$$

By the principle of mass conservation for incompressible fluids, the flow rate exiting the syringe (2) must equal the flow rate deposited on the platform (3) [10], [13]. Equating both expressions:

$$\frac{\pi \cdot D_{Syringe}^2}{4} \cdot v_{Syringe} = w \cdot h \cdot v_{print} \tag{4}$$

Isolating the plunger velocity $v_{Syringe}$ yields the fundamental parameter for calibrating the slicer:

$$v_{Syringe} = \frac{4 \cdot w \cdot h \cdot v_{print}}{\pi \cdot D_{Syringe}^2} \tag{5}$$

Equation (5) establishes that, for a fixed diameter syringe, the plunger speed must be adjusted proportionally to the print speed, line width, and layer height. This result is consistent with the work of Pusch et al. [12], who demonstrated the need to calibrate the volumetric flow in desktop syringe systems, and with the adaptations implemented by Contreras et al. [11], who applied similar relationships when converting an Ender 3 for bioprinting.

By applying this mathematical relationship, it was possible to calibrate PrusaSlicer's extrusion multiplier for different deposition geometries. This ensured that the amount of extruded material corresponded exactly to the volume expected by the slicer, without requiring changes to the original firmware's *E-steps*. This approach bypasses electronic modifications on the printer since the motor motion logic remains unaltered, with the adaptation performed entirely within the slicing software [9], [15].

It is worth noting that (5) is valid for incompressible fluids and disregards rheological effects such as shear-rate-dependent viscosity, which are common in bioinks and food materials. For applications involving non-Newtonian materials, it is recommended to include an empirical correction factor obtained through extrusion test runs, as discussed by Ozbolat [13] and Piedade et al. [14]. In the present work, since the material used exhibits pseudoplastic behavior, additional manual adjustments were performed experimentally to compensate for viscosity variations during the printing process.

G. Materials Used

The initial tests were conducted using semi-solid food materials that do not require active temperature control, including mashed potatoes, cookie dough, and peanut butter.

The methodology used for mechanical validation is consistent with recent studies on the adaptation of FDM printers to syringe extrusion, in which food materials or rheological substitutes are initially used to verify the continuity of extrusion and deposition stability before conducting more comprehensive quantitative trials [8], [15].

As a future development of the project, we intend to incorporate a thermal control system into the print head, expanding the range of materials processable by the platform. The inclusion of this module will enable the printing of temperature-sensitive materials, such as chocolate and other food formulations that require controlled heating or cooling during extrusion. Similar solutions have been employed in platforms dedicated to chocolate 3D printing, where temperature control is crucial to maintaining the material's proper viscosity, ensuring extrusion continuity, and preserving the geometric fidelity of the produced structures [8]. Additionally, the developed modular architecture allows for the future integration of complementary modules, such as UV lighting systems, which are used in soft material extrusion platforms to promote *in situ* curing of photosensitive materials [15]. Although this functionality is not intended for conventional food printing, its implementation demonstrates the platform's potential for expansion into experimental applications and for processing special food formulations and other extrudable materials.

III. Results and Discussion

The development resulted in a functional food 3D printing platform based on the adaptation of the *Creality Ender-3 V3 SE* printer. Figure 3 presents an overview of the printer after the installation of the developed syringe-based extrusion system.



Fig. 3. Overview of the developed food 3D printer. Source: The authors (2026).

The main modifications made to the original printer are summarized in Table II. The adaptation consisted of the complete removal of the original fused deposition modeling (FDM) extrusion system and the installation of a print head designed specifically for the extrusion of semi-solid materials. The assembly was designed to use 20 mL Luer Lock syringes, allowing for rapid replacement of the material reservoir and facilitating cleaning and sanitization procedures.

The entire mechanical assembly was developed in a CAD environment using *Autodesk Fusion 360* software and subsequently fabricated via 3D printing. The developed components include the syringe bracket, the drive motor mount, the plunger displacement system, sensor brackets, the ESP32 enclosure, and other necessary fasteners to integrate the new print head into the printer's original frame. The three-dimensional models developed in this project are available for consultation and reproduction at the following addresses: print head [*https://a360.co/44JrYqv*], ESP32 enclosure [*https://a360.co/4eYl2ue*], and X-axis limit switch bracket [*https://a360.co/4eWFHiu*]. The approximate cost of the components used in the adaptation was R$ 4,000.00 , including mechanical, electronic, and drive components, considering the cost of the base printer.

The developed extrusion mechanism utilizes a *NEMA 17* stepper motor coupled to a trapezoidal lead screw to drive the

Table II
MAIN MODIFICATIONS MADE TO THE PRINTER.

Item	Description
Base printer	Creality Ender-3 V3 SE
Print head	Mechanical design developed to use a 20 mL Luer Lock syringe.
Extrusion system	NEMA 17 motor driving a trapezoidal lead screw.
Linear guides	Three 6 mm linear guides for plunger alignment.
Leveling	Retractable BLTouch sensor actuated by an ESP32.
Firmware	Original printer firmware preserved.
CAD Software	Autodesk Fusion 360.
Slicing	PrusaSlicer with parameters adapted for food extrusion.
Tested materials	Mashed potatoes, cookie dough, and peanut butter.

Source: The authors (2026).

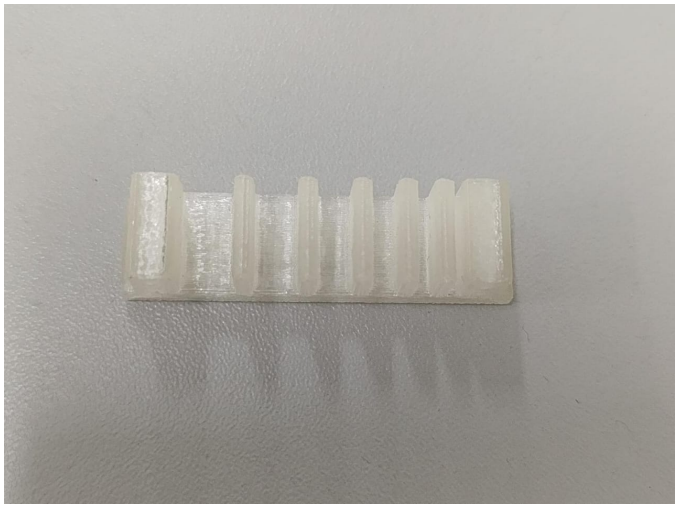


Fig. 4. Bridge-collapse test used for qualitative assessment of extruded filament stability. Source: The authors (2026).

syringe plunger, while three linear guides ensure alignment throughout the entire displacement. The platform was experimentally validated through mechanical and motion system verification. The X, Y, and Z axes of the Ender-3 V3 SE base printer were operated across their full range of motion, confirming that the new print head assembly does not interfere with the original movement system. The automatic bed leveling system was also validated during printer assembly. The mechanism developed using an *ESP32* microcontroller temporarily positions the *BLTouch* sensor only during the bed leveling procedure. Upon completion of this stage, the sensor automatically returns to a retracted position, avoiding interference with the new print head or deposited material during part fabrication and preserving the original printer firmware without requiring changes.

Preliminary extrusion tests were conducted using semi-solid materials to verify the functional operation of the developed syringe-based extrusion system. Peanut butter was loaded into the 20 mL syringe and extruded through the developed mechanism. The material was successfully ejected from the syringe with a continuous and visually stable flow, demonstrating that the combination of the NEMA 17 stepper motor, trapezoidal lead screw, and plunger-guiding system is capable of generating the displacement required for semi-solid material extrusion. No apparent interruptions or mechanical instabilities were observed during the test.

A second experimental assessment was performed using a bridge configuration with Nivea cream as the extruded material. The test was used to qualitatively evaluate the behavior of the deposited filament under unsupported conditions, with particular attention to filament continuity and its tendency to collapse as the unsupported span increases. The bridge configuration provides a practical assessment of the interaction between extrusion consistency and the ability of the deposited material to maintain its geometry after deposition. Figure 4 presents the bridge-collapse test arrangement.

The observed behavior of the extruded filament in the bridge test demonstrates that the deposited material can form a continuous unsupported span before progressive collapse occurs, providing a qualitative indication of the platform's ability to maintain filament geometry after deposition, a prerequisite for dimensional accuracy in layered fabrication. Together, the peanut butter extrusion and bridge-collapse tests provide complementary evidence of the platform's functionality: the first evaluates the ability of the mechanism to continuously eject a semi-solid material from the syringe, whereas the second evaluates the behavior of the deposited filament after extrusion. These preliminary results support the mechanical feasibility of the proposed syringe-based architecture for food extrusion applications.

Regarding the quantitative performance metrics requested for a comprehensive characterization, the following considerations can be drawn from the present study. The dimensional accuracy of extrusion-based food printing is influenced by multiple factors, including printing parameters, rheological properties, extrusion rate, layer height, and post-processing conditions [19], [20]. The successful bridge test, although qualitative, demonstrates that the extrusion system is capable of producing continuous filaments under unsupported conditions, which is a necessary condition for achieving dimensional accuracy in layered structures. Moreover, the Creality Ender-3 V3 SE base platform features a positioning accuracy of ±0.1 mm in the X and Y axes according to the manufacturer [20], which establishes a baseline for the achievable dimensional accuracy of the mechanical system. While quantitative measurements of dimensional accuracy (e.g., printed geometry dimensions) were not performed in this preliminary study, the combination of the base platform's positioning accuracy and the observed filament stability suggests that the system has the potential to produce dimensionally consistent structures.

The mechanical architecture was designed to promote repeatable plunger displacement. The three linear guides constrain the syringe plunger during actuation, while the trape-

zoidal lead screw provides controlled linear displacement driven by the NEMA 17 stepper motor. The use of three linear guides for plunger alignment is expected to contribute to extrusion repeatability by reducing lateral forces during plunger displacement. This mechanical design choice follows the principles adopted in open-source soft-material extrusion platforms, where guided plunger motion has been associated with improved deposition consistency [15]. The preliminary extrusion tests demonstrated stable operation, suggesting that the system is capable of repeatable deposition. Statistical repeatability, however, was not quantitatively determined through repeated depositions or mass measurements in the present study.

Extrusion consistency was assessed through the visual observation of the extruded flow during the peanut butter test, which remained continuous and stable throughout the procedure. Although quantitative measurements of extrusion consistency (e.g., coefficient of variation of deposited mass or filament diameter) were not performed, the visual uniformity of the extruded flow is consistent with observations reported for syringe-based extrusion of semi-solid materials, where coefficients of variation below 10% are typically achieved for adequately formulated pastes and food inks [19]. This suggests that the developed mechanism provides a suitable basis for consistent material deposition, provided that the material rheology remains within the operational range of the system. Future work will include quantitative measurements of extrusion consistency to establish the statistical performance of the system.

Regarding printing time, the developed platform retains the motion system of the Ender-3 V3 SE, whose maximum motion speed is specified by the manufacturer at up to 250 mm/s for conventional FDM operation [20]. For syringe-based food extrusion, however, the practical printing speed is expected to be lower because the achievable deposition rate depends on material rheology, syringe geometry, and plunger displacement. The mathematical formulation in (5) provides a basis for relating plunger velocity and printing speed to the required volumetric flow rate, enabling the estimation of printing time for a given geometry once the material-specific parameters are determined. A systematic experimental evaluation of printing speed and printing time for different materials was not performed in the present study.

To position the developed platform within the current landscape of low-cost food and soft-material extrusion systems, Table III presents a technical comparison between the proposed system and two representative platforms reported in the literature: the syringe-pump system by Heriyanto et al. [9], and the *Printer.HM* platform [15].

The comparative analysis highlights three distinctive features of the proposed platform. First, unlike adaptations that require modifications to firmware [10], [11], the proposed approach preserves the original electronics and firmware. Second, the use of a dedicated ESP32 microcontroller for automatic bed leveling enables the use of the BLTouch sensor without permanent modifications to the print head, avoiding collisions during printing. Third, the inclusion of three linear guides for plunger alignment, combined with a trapezoidal lead screw, provides mechanical stability during deposition. Rogers [8] successfully demonstrated chocolate printing with temperature control, whereas the proposed platform currently focuses on semi-solid materials and provides a modular architecture that allows future thermal integration. Heriyanto et al. [9] developed a syringe-pump mechanism compatible with multiple syringe sizes, while the proposed system emphasizes integration with a commercially available FDM platform while preserving its original firmware. Compared with the *Printer.HM* platform [15], the proposed system provides a lower estimated acquisition cost while incorporating an automated leveling mechanism based on an independent ESP32 controller.

Table III
TECHNICAL COMPARISON OF THE DEVELOPED PLATFORM WITH RELATED LOW-COST EXTRUSION SYSTEMS.

Feature	This work	Heriyanto et al. [9]	Printer.HM [15]
Base platform	Ender-3 V3 SE	Custom	Custom
Extrusion	20 mL syringe	Syringe	Syringe
Temperature control	Planned	No	No
Leveling	ESP32 + BLTouch	Manual	Manual
Tested materials	Semi-solid materials	Dough	Food pastes
Estimated cost (USD)	US$ 700	US$ 450	US$ 1,100

Source: Compiled by the authors based on [9], [15].

The preliminary experimental results demonstrate the functional operation of the developed food extrusion platform and provide qualitative evidence of continuous material ejection and filament stability after deposition. Nevertheless, a comprehensive quantitative characterization remains necessary to determine extrusion repeatability, dimensional accuracy, deposited mass variation, and printing time under controlled conditions. These measurements constitute an important next stage for establishing the statistical performance of the platform. Among the proposed future developments, the implementation of temperature control in the print head stands out, enabling the processing of materials such as chocolate and other formulations requiring controlled heating [8]. Additionally, the future incorporation of UV lighting modules could expand the platform's use to photosensitive materials [15]. The modular architecture adopted during development allows these modifications to be incorporated without significant changes to the original printer structure.

IV. CONCLUSION

This work presented the development of a low-cost food 3D printing platform obtained by adapting a commercial FDM printer for syringe-based extrusion of semi-solid materials. The proposed system integrates a 20 mL Luer Lock syringe, a NEMA 17 stepper motor, a trapezoidal lead screw, and three linear guides for controlled plunger displacement, while preserving the original electronics and firmware of the *Creality Ender-3 V3 SE*.

The experimental results demonstrated the functional operation of the developed extrusion mechanism. Peanut butter was successfully ejected from the syringe with a continuous and

visually stable flow, confirming the ability of the mechanism to generate the required plunger displacement for semi-solid material extrusion. A complementary bridge-collapse assessment using Nivea cream provided a qualitative evaluation of filament stability under unsupported conditions, demonstrating the applicability of the platform to the study of post-deposition material behavior.

The developed architecture also incorporates an ESP32-controlled retractable BLTouch system for automatic bed leveling, allowing the sensor to be deployed during the leveling procedure and retracted during printing. This approach preserves the original printer firmware and reduces the need for permanent modifications to the base machine. The estimated total cost of approximately US$700 further supports the suitability of the platform for educational and research applications.

The comparison with related low-cost extrusion platforms indicates that the proposed system combines syringe-based extrusion, automated leveling, mechanical plunger guidance, and preservation of the original printer control architecture within a comparatively low-cost platform. These characteristics provide a modular basis for further development of food and soft-material additive manufacturing systems.

Although the present study provides functional and qualitative experimental validation, quantitative characterization of extrusion repeatability, dimensional accuracy, deposited mass variation, and printing time remains necessary to fully establish the performance limits of the platform. Future work will therefore focus on controlled quantitative testing, as well as the implementation of temperature control for thermally sensitive materials such as chocolate and additional modules for expanding the range of printable materials.

Overall, the developed platform demonstrates the feasibility of converting a commercially available low-cost FDM printer into a modular syringe-based food extrusion system while retaining its original electronics and firmware. The resulting architecture provides a practical foundation for further research, experimentation, and educational applications in food additive manufacturing.

ACKNOWLEDGMENT

The authors would like to thank the National Institute of Telecommunications (INATEL) for the institutional support and the infrastructure made available for the development of this work.

Special thanks to Prof. Dr. Maysa Costa Alves for her guidance, valuable technical contributions, and supervision throughout the development of this undergraduate research project (Iniciação Científica).

The authors also thank Prof. Dr. Fabiano Valias de Carvalho, coordinator of the Biomedical Engineering course, for encouraging research and scientific development.

REFERENCES

[1] A. K. Mantihal, A. Prakash, G. S. Godbole, and S. Bhandari, "3D Food Printing of As the New Way of Preparing Food: A Review," *International Journal of Gastronomy and Food Science*, vol. 22, 2020.

[2] F. J. Liu, Y. Zhang, and H. Bhandari, "Extrusion-based 3D Food Printing: Materials and Machines," *Foods*, vol. 9, no. 10, 2020.

[3] A. Attaran, "The Rise of 3D Printing: The Advantages of Additive Manufacturing over Traditional Manufacturing," *Business Horizons*, vol. 60, no. 5, pp. 677-688, 2017.

[4] M. H. Derossi, R. Caporizzi, D. Azzollini, and A. Severini, "Application of 3D Printing for Customized Food. A Case on the Development of a Fruit-Based Snack for Children," *Journal of Food Engineering*, vol. 220, pp. 65-75, 2018.

[5] Natural Machines. *Foodini 3D Food Printer*. Available: <https://www.naturalmachines.com/>. Accessed: Jul. 2026.

[6] ByFlow. *Focus 3D Food Printer*. Available: <https://3dbyflow.com/>. Accessed: Jul. 2026.

[7] Cocoa Press. *Cocoa Press 3D Chocolate Printer*. Available: <https://cocoapress.com/>. Accessed: Jul. 2026.

[8] B. M. Rogers, "Development of a Low-Cost Chocolate 3D Printer," M.S. thesis, Dept. Mech. Eng., Northern Illinois Univ., DeKalb, IL, USA, 2022.

[9] S. M. K. Heriyanto et al., "Development of an Affordable 3D Food Printer with an Exchangeable Syringe-Pump Mechanism," *HardwareX*, vol. 13, 2023.

[10] S. Chaji, J. Al-Saleh, and A. A. Al-Rub, "A low-cost 3D bioprinter platform for fabrication of tissue constructs," *Bioprinting*, vol. 19, p. e00084, 2020.

[11] A. Contreras, M. A. Lopez, and A. Martinez, "Conversion of an Ender 3 FDM printer into a low-cost extrusion-based bioprinter," *Micromachines*, vol. 13, no. 8, p. 1212, 2022.

[12] K. Pusch, T. J. Hinton, and A. W. Feinberg, "Large volume syringe pump extruder for desktop 3D printers," *HardwareX*, vol. 3, pp. 49-61, 2018.

[13] I. T. Ozbolat and M. Hospodiuk, "Current advances and future perspectives in extrusion-based bioprinting," *Biomaterials*, vol. 76, pp. 321-343, 2016.

[14] A. C. B. M. Piedade, M. F. Silva, and R. M. G. Alves, "Flow rate mathematical modeling for syringe-based extrusion in low-cost 3D bioprinters," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 45, no. 3, 2023.

[15] J. A. Weis, J. H. Lee, A. L. Thompson, and J. P. Wachs, "A Hackable, Multi-Functional, and Modular Extrusion 3D Printer for Soft Materials," arXiv preprint arXiv:2203.07497, 2022. Available: <https://arxiv.org/pdf/2203.07497>.

[16] W. Liu, H. Hu, D. J. McClements, Z. Jin, and L. Chen, "Fabrication and characterization of edible inks for 3D printing of dysphagia foods based on corn starch stabilized by calcium ions and hydrocolloids," *Food Hydrocolloids*, vol. 166, p. 111278, 2025.

[17] Z. Teng, X. Huang, C. Zhang, X. Liu, Y. Li, C. S. Wang, X. Liu, and F. Xie, "Xanthan gum modulation of octenyl succinic anhydrate starch-based high internal phase emulsions: Characterization, rheological behavior, and 3D printing applications," *Food Chemistry*, vol. 464, no. 3, p. 141813, 2025.

[18] Z. Liu, X. Chen, M. Ruan, Y. Liao, Z. Wang, Y. Zeng, H. Li, L. Hu, and H. Mo, "3D printed dysphagia diet using pea protein gel modified by xanthan gum with different pyruvate group content," *Food Research International*, vol. 190, p. 114894, 2025.

[19] P. Shah, V. Shah, S. Shrestha, and P. Mitra, "A Review on 3D Food Printing Technology in Food Processing," *Journal of Food Industry*, vol. 8, no. 1, 2024. DOI: 10.5296/jfi.v8i1.22468.

[20] Creality, "Ender-3 V3 SE User Manual," Creality, 2024.

Mechanical Behavior of Biosilicate Scaffolds: Experimental and Computational Analysis

Karina Feliciano Santos
Program in Materials Science and Engineering
University of São Paulo
Pirassununga, Brazil
kfeliciano@usp.br

Marcília Valéria Guimarães
Biofabrication Research Group
Center for Information Technology Renato Archer
Campinas, Brazil
mvguimaraes@cti.gov.br

Claudio Yamamoto Morasutti
Biofabrication Research Group
Center for Information Technology Renato Archer
Campinas, Brazil
claudio.morassuti@cti.gov.br

Marco A. Sabino-Gutiérrez
B^{5}IDA Group, Department of Chemistry,
Simón Bolívar University
Caracas, Venezuela
msabinog@gmail.com

Ronny Calixto Carbonari
Center of Engineering Modeling, and Applied Social Sciences
Federal University of ABC
Santo André, Brazil
ronny.carbonari@ufabc.edu.br

Juliana Kelmy Macário Barboza Daguano
Biofabrication Research Group
Center for Information Technology Renato Archer
Campinas, Brazil
juliana.daguano@cti.gov.br

***Abstract* — This study investigated the mechanical behavior of cryogenically treated Biosilicate scaffolds by combining micro-computed tomography, finite element analysis, and experimental compression testing. Three-dimensional reconstructions were used to generate computational models representative of the fabricated scaffolds. The experimental results showed a compressive strength of approximately 50 MPa and an effective elastic modulus of 1.0 GPa. Finite element simulations based on the μ-CT-reconstructed geometry showed good agreement with the experimental stress–strain response, while simulations performed on an idealized CAD model of the scaffold architecture identified critical stress concentration regions at filament intersections, with a maximum principal stress of approximately 30–47 MPa. These findings demonstrate the potential of integrating micro-computed tomography and finite element analysis to evaluate and optimize porous scaffolds for bone tissue engineering applications.**

Keywords— Biosilicate, scaffold, finite element analysis, micro-computed tomography.

I. INTRODUCTION

Bioceramics have been extensively investigated for bone regeneration applications due to their excellent biocompatibility and chemical similarity to the mineral phase of bone tissue [1]. Among these materials, bioactive glasses and glass-ceramics have attracted attention due to their high bioactivity and ability to form a hydroxyapatite layer upon contact with biological fluids, supporting cell adhesion, osseointegration, and new bone formation [2, 3].

Biosilicate® (BioS), a bioactive glass-ceramic, has gained considerable interest due to its well-established bioactivity, osteoconductive properties, and potential for bone tissue engineering applications [4]. When processed into three-dimensional scaffolds, BioS can provide a suitable environment for cell adhesion, proliferation, and differentiation, supporting bone regeneration [5, 6].

For successful bone tissue engineering applications, scaffolds must achieve an appropriate balance between porosity and mechanical performance compatible with the host tissue [7, 8, 9] . An interconnected porous network is essential for cell migration, vascularization, and nutrient transport, while sufficient structural integrity is required to withstand physiological loads during the regeneration process. However, increasing porosity generally compromises the mechanical strength of the scaffold, representing a major challenge in the design of porous biomaterials [8].

In addition to chemical composition, the mechanical performance of scaffolds is strongly influenced by their three-dimensional architecture, including porosity, pore interconnectivity, filament diameter, and pore distribution. Although additive manufacturing enables precise control of scaffold geometry, subsequent processing steps, such as cryogenic treatment (CT) and sintering [10, 11], may induce dimensional and microstructural changes that alter the fabricated architecture. Consequently, its mechanical response. Accurate characterization of the resulting scaffold geometry is therefore essential for understanding its mechanical behavior [9] .

Although several studies have investigated the bioactivity and mechanical properties of BioS scaffolds, few have addressed the influence of their actual three-dimensional geometry on mechanical behavior by combining micro-computed tomography (μ-CT), finite element analysis (FEA), and experimental mechanical testing [12, 13]. Moreover, the combined use of μ-CT-based and idealized CAD-based finite element models remains limited. While μ-CT-based models represent fabricated architecture, idealized CAD models can be used to identify stress concentration regions associated with the scaffold design and support structural optimization.

Therefore, this study investigates the mechanical behavior of cryogenically treated BioS scaffolds by correlating μ-CT-based finite element simulations with experimental uniaxial compression tests. In addition, finite element simulations based on an idealized CAD model were performed to identify critical

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

stress concentration regions associated with potential failure sites and to provide insights for future scaffold design optimization.

II. MATERIALS AND METHODS

A. Materials

The printing paste was formulated using hydrogel polyethylene glycol 400 (PEG-400, Synth, 380–420 g/mol), Laponite XLG (Colormix, BYK Additives & Instruments), and Biosilicate (BioS) powder, supplied by the Laboratory of Vitreous Materials (LaMaV), Federal University of São Carlos (UFSCar).

B. Scaffold Fabrication

The BioS scaffolds were fabricated by direct ink writing (DIW) following the methodology previously described by Daguano et al. (2022) [14]. Scaffolds with nominal dimensions of 10 × 10 × 5 mm, comprising 15 layers, were printed using a 580 µm conical nozzle (20 G) and BioScaffolds V2.0 software. An air gap of 1.55 mm was used, resulting in a theoretical porosity of 70%.

After printing, the samples were subjected to cryogenic treatment (CT), which consisted of successive cycles of freezing at -86 °C for 2 h in an ultralow-temperature freezer (New Brunswick Scientific), followed by thawing at room temperature (25 °C) for 2 h. These cycles were repeated over a total treatment period of 72 h. Subsequently, the scaffolds were sintered at 900 °C for 2 h using a heating rate of 1 °C/min. Non-sintered scaffolds were used for the control group.

C. Shape Fidelity and Dimensional Shrinkage Relative to the Designed Geometry

The shape fidelity of the scaffolds was evaluated by comparing the designed pore geometry with the pore geometries obtained after printing, cryogenic treatment (CT), and sintering. Optical images were analyzed using ImageJ software to determine the Printability parameter (Pr) [15], according to equation (1):

$$\mathrm{Pr} = \frac{\mathrm{L}^2}{16\mathrm{A}} \qquad (1)$$

Where A is the pore area and L is its perimeter. Pr = 1 indicates correspondence with the designed geometry, while Pr ≠ 1 indicates distortion.

The scaffold dimensions in the X, Y, and Z directions were measured using a digital caliper (DIGIMES, Brazil). The corresponding linear shrinkage was then calculated according to equation (2):

$$Shrinkage(\%) = \frac{L_i - l_f}{L_i} \times 100 \qquad (2)$$

Where L_i and L_f are the initial and final dimensions, respectively (i= X, Y and Z).

D. Morphological and Mechanical Characterization

The scaffolds' surface morphology was examined by field-emission scanning electron microscopy (FEG-SEM, Tescan). Prior to imaging, the samples were sputter-coated with a 20 nm thick gold layer.

The mechanical behavior of the scaffolds was evaluated by uniaxial compression testing using a universal testing machine (EMIC DL 2000, Instron) equipped with a 5 kN load cell. The tests were performed at a crosshead speed of 1 mm/min, with the load applied perpendicular to the printing plane. The compressive strength was determined from the maximum stress recorded during the test. The results are presented as the mean and standard deviation (n = 5). Statistical significance between the compressive strength of the control and sintered groups was evaluated using GraphPad Prism (version *v*8.0, 2022), One-way ANOVA, with a significance level set at $p < 0.0001$.

E. Micro-Computed Tomography and Computational Modeling

The three-dimensional architecture of the scaffolds was characterized by micro-computed tomography (µ-CT) using a SkyScan 1272 CMOS Edition microtomograph (Bruker, Kontich) with a spatial resolution of 12 µm and 1200 ms exposure time. The acquired images were reconstructed using NRecon software (Bruker) and subsequently processed in CTAn software by automatic segmentation based on the Otsu 3D algorithm.

The three-dimensional reconstructions obtained from µ-CT were used to generate computational models employed for FEA. First, the reconstructed geometries were processed using routines developed in MATLAB and subsequently imported into HyperMesh software (Altair Engineering, Troy, MI, USA), where the geometries were discretized and the finite element meshes were generated. The numerical simulations were performed using Abaqus software (Dassault Systèmes, Providence, RI). The simulations were carried out under boundary conditions equivalent to those adopted in the experimental uniaxial compression tests, including geometrically nonlinear analysis with large deformations and frictional contact interactions between the scaffold and the loading surface.

The numerical models were calibrated against the experimental force-displacement data to estimate the effective solid-phase elastic modulus. The quality of this numerical-experimental fit was quantitatively assessed using the coefficient of determination (R^2) and the Root Mean Square Error (RMSE).

In parallel, an idealized CAD model of the scaffold architecture, based on the nominal design parameters (10 × 10 × 5 mm, 15 layers, 580 µm nozzle diameter), was developed and discretized into a finite element mesh using HyperMesh software (Altair Engineering, Troy, MI, USA). Linear static analysis was performed in OptiStruct (Altair Engineering) under a 5 kN compressive load applied through a sliding, frictionless contact interface between the scaffold and rigid loading platens. The same elastic modulus (E = 1.0 GPa), obtained from the

calibration against the experimental stress–strain curve, was assigned to this model. The maximum principal stress criterion was applied to identify regions of tensile stress concentration and potential failure sites, given its suitability for representing the fracture behavior of brittle porous ceramics, in which crack initiation is typically governed by local tensile stresses [12].

III. RESULTS AND DISCUSSION

A. Shape Fidelity Characterization

The shape fidelity characteristics of the scaffolds before and after sintering are summarized in Table 1.

Table 1 - Geometric parameters, printability, and linear shrinkage of BioS scaffolds before and after sintering.

Parameter	BioS Scaffold Control (non-sintered)	BioS Scaffold (Sintered)
Air gap (mm)	1.00 ± 0.10	0.89 ± 0.09
Pr	0.96 ± 0.04	0.91 ± 0.03
Linear Shrinkage (X, Y, and Z) (%)	---	5.4 ± 0.5 5.9 ± 0.8 11.0 ± 0.7

The measured air gap decreased from 1.00 ± 0.10 mm in the non-sintered scaffolds to 0.89 ± 0.09 mm after sintering. The printability parameter (Pr) remained close to 1 in both conditions, with values of 0.96 ± 0.04 and 0.91 ± 0.03 for the non-sintered and sintered scaffolds, respectively, indicating good preservation of the designed pore geometry. Following sintering, linear shrinkage values of 5.4 ± 0.5%, 5.9 ± 0.8%, and 11.0 ± 0.7% were observed in the X, Y, and Z directions, respectively. This behavior is expected and is associated with the removal of polymeric components and densification of the ceramic structure during thermal processing [14, 16, 17].

These results confirm that, despite the dimensional changes induced by sintering, the designed architecture was satisfactorily preserved. The effects of the processing steps on the microstructural features and mechanical behavior of the scaffolds are further presented in Fig. 1.

Fig. 1(a) shows the BioS scaffolds after 3D-printing and CT, exhibiting a well-defined three-dimensional architecture with highly interconnected pores. Although a reduction in the air gap was observed after sintering, Fig. 1(b), as previously described, this dimensional change did not compromise the overall macroporous architecture. The interconnected pore structure is relevant for bone tissue engineering applications, as it provides pathways for cell infiltration and nutrient transport [9].

The SEM micrographs presented in Fig. 1(c–d) correspond to the scaffolds before and after sintering at 900 °C, respectively. Prior to heat treatment, the scaffold exhibited a microstructure composed of discrete ceramic particles with a high volume of interparticle voids. Following sintering, a more consolidated and homogeneous microstructure was observed, accompanied by a marked reduction in interparticle porosity. These microstructural changes indicate effective densification promoted by the sintering process while preserving the overall integrity of the printed architecture [14, 16].

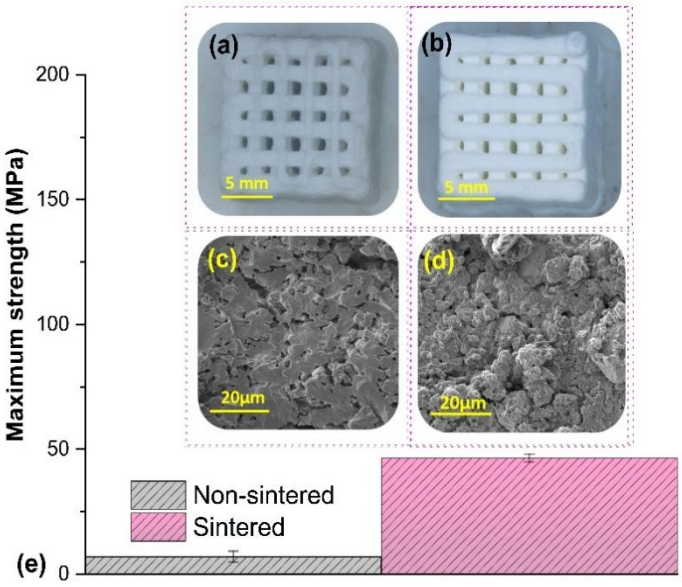


Fig.1 - Structural and mechanical characterization of BioS scaffolds: macroscopic view (a,b) and SEM micrographs (c,d) of non-sintered and sintered scaffolds, respectively. Maximum strength obtained from uniaxial compression tests of the scaffolds before and after sintering (d).

The compressive mechanical behavior of the scaffolds is presented in Fig.1(e). The non-sintered scaffolds exhibited a compressive strength of 5.9 ± 1.2 MPa, whereas the sintered scaffolds demonstrated a statistically significant increase ($p < 0.0001$), reaching values of 46.5 ± 1.6 MPa. This substantial improvement can be attributed to the consolidation of the ceramic network during sintering, which enhances particle bonding and load transfer throughout the structure, resulting in a significant increase in mechanical strength [17, 18].

B. Morphological, Mechanical Analyses, and Computational Modeling

Although the morphological and mechanical analyses provide valuable information regarding scaffold characteristics, they offer only limited insight into the three-dimensional architecture. Therefore, µ-CT was employed to reconstruct the actual scaffold geometry and generate representative computational models for subsequent finite element simulations, as shown in Fig. 2.

Fig. 2(a–d) presents three-dimensional µ-CT reconstructions of the sintered BioS scaffolds, confirming that the designed architecture was preserved after thermal processing. This is evidenced by the regular filament arrangement and the fully interconnected porous network maintained throughout the structure, providing a high-fidelity representation of the as-fabricated scaffolds.

A quantitative morphometric comparison between the idealized CAD geometry and the µ-CT reconstruction further confirmed the structural fidelity of the processing route. While the idealized CAD model possessed a theoretical porosity of 70% and a nominal filament diameter of 580 µm, µ-CT analysis of the sintered scaffolds revealed a reduced filament diameter of 450 ± 30 µm (0.45 ± 0.03 mm) and an open porosity of approximately 20%. Furthermore, the total porosity decreased by 30.14% from the non-sintered control group to the sintered scaffolds. These microstructural changes are highly consistent

with the observed post-sintering volumetric shrinkage and successfully validate the structural representations employed in the numerical models.

Furthermore, Fig. 2(e–f) compares the raw µ-CT data with the computationally processed model used for finite element (FE) mesh generation. The excellent geometric agreement between them demonstrates that the segmentation and processing procedures successfully retained the scaffold's key structural features, yielding a reliable model for numerical simulations.

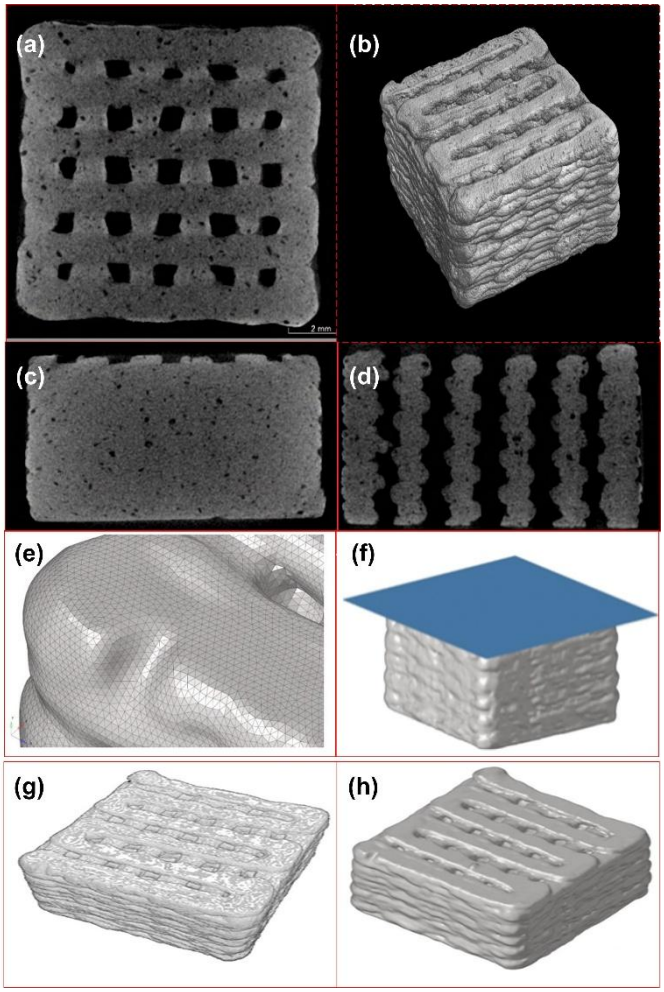


Fig. 2 - Three-dimensional µ-CT reconstructions of sintered BioS scaffolds: (a) front, (b) side, and (c) top views; (d) detailed side view highlighting the filament stacking architecture. A qualitative comparison is shown between (e) the raw µ-CT reconstruction and (f) the computationally processed model. The finite element analysis (FEA) setup illustrates (g) the applied boundary conditions and (h) the generated mesh.

The processed geometries were imported into HyperMesh to generate the finite element mesh, as illustrated in Fig. 2 (g–h). Comprising 3,232,395 linear tetrahedral (C3D4) elements and 600,299 nodes, this discretization accurately captured the geometric complexity of the scaffolds and was subsequently exported to Abaqus for numerical simulation.

Fig. 3 illustrates the mechanical response of the BioS scaffolds under compression, plotting Force (N) against Displacement (mm). The data is separated into experimental results and two computational models.

Experimental curve exhibits an initial non-linear region (0 to 0.2 mm) followed by a pronounced linear elastic regime. The curve reaches a peak force of approximately 5500 N at a displacement of 0.65 mm, indicating structural failure. This is followed by a sharp drop and a long, fluctuating plateau extending up to 2.0 mm, typical of cellular solid crushing behavior. The compression tests performed on the scaffolds demonstrated the typical mechanical behavior of brittle ceramic materials. The elastic limit nearly coincided with the maximum compressive stress, indicating the limited plastic deformation capacity of the scaffolds. Once the maximum stress was reached, structural failure was initiated through the nucleation and propagation of cracks within the ceramic structure [19, 20].

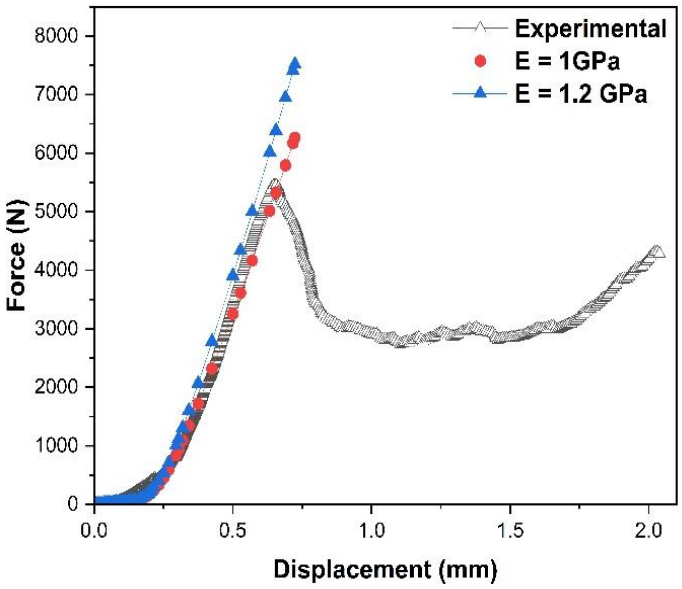


Fig. 3 - Experimental and computational force-displacement curves for BioS scaffolds under uniaxial compression. The experimental data (grey open triangles) is compared against finite element simulations adopting a solid phase Young's modulus of E = 1 GPa (red circles) and E = 1.2 GPa (blue triangles).

The Poisson's ratio was adopted from values reported in the literature, whereas the elastic modulus was estimated by fitting the numerical response to the experimental stress–strain curve. Although no formal optimization procedure was employed to refine the fitting process, the adopted approach provided representative and physically consistent mechanical properties for the analyzed scaffolds.

The numerical simulations were performed in Abaqus using an implicit dynamic analysis under a quasi-static assumption. This approach is highly suitable for nonlinear problems where inertial effects are negligible, effectively treating the compressive loading as a slowly applied static load to ensure robust convergence despite geometric nonlinearities and contact interactions.

The parameter identification process revealed that a solid-phase elastic modulus of 1.0 GPa yielded the most accurate representation of the experimental macroscopic stiffness. The calibrated numerical model demonstrated high accuracy throughout the linear elastic regime, confirmed by an error metric of $R^2 = 0.98$ and an RMSE of 202.32 N (3.72%). While this procedure calibrates effective modulus rather than serving as an independent validation, the obtained value is physically consistent for highly porous ceramic scaffolds.

The experimental curve shows that the transition between the elastic limit and the peak compressive force is virtually absent, which is characteristic of brittle ceramics. Following this

maximum load, the scaffold undergoes abrupt structural failure due to crack initiation and propagation.

To determine the appropriate solid phase elastic modulus, values of 1.0 GPa and 1.2 GPa were evaluated. The 1.0 GPa model yielded the best fit with the experimental force–displacement curve. Although considerably lower than the modulus typically reported for dense ceramic materials (~70 GPa), this difference is directly attributed to the intrinsic porosity of the processed scaffolds [14, 21] .

The high porosity of the microstructure substantially reduces the effective stiffness of the material, as the elastic modulus decreases markedly with increasing porosity. Therefore, the results suggest that the observed mechanical response is directly governed by the porous microstructure developed during scaffold processing [22, 23].

Overall, the numerical models accurately captured the mechanical behavior of the BioS scaffolds sintered at 900 °C. Reaching a compressive strength of approximately 50 MPa, these scaffolds demonstrate the efficiency of higher sintering temperatures in promoting ceramic network densification and reducing porosity. These results emphasize that the mechanical performance of the fabricated scaffolds is fundamentally governed by the interplay between their porous architecture and degree of structural consolidation.

The finite element simulations also revealed the internal stress distribution of the scaffold under compressive loading. As illustrated in Fig. 4, the numerical setup applied to the idealized CAD model includes a 5 kN compressive load. This detailed view highlights the orthogonal stacking architecture of the filaments and reveals the fine finite element mesh (discretization) generated over the cylindrical surfaces for the numerical simulation. In this manner, the detailed inset confirms that the adopted mesh discretization accurately captures the geometric complexity of the scaffold architecture, particularly at the critical filament intersections and stacking regions.

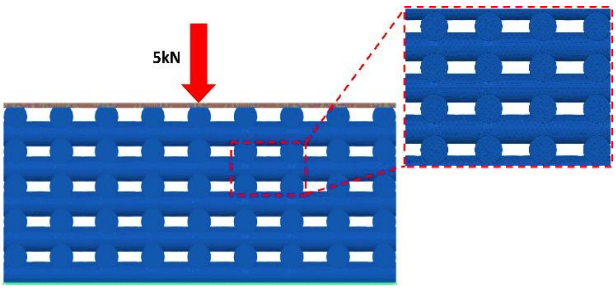


Fig. 4 - Computational setup for the finite element analysis of the BioS scaffold under uniaxial compression. A 5 kN compressive load is applied to the upper surface of the model, while the magnified inset highlights the filament stacking architecture and the generated mesh discretization.

The finite element analysis (FEA) of the idealized CAD model provided critical insights into the internal mechanics of the scaffolds under compression (Fig. 5). The displacement field (Fig. 5a) revealed that deformation was predominantly localized in the central region of the structure, demonstrating that the macroscopic mechanical response is highly dependent on the internal filament arrangement and pore distribution.

Furthermore, the maximum principal stress distribution (Fig. 5b) highlighted localized tensile stress concentrations, primarily at the intersections between adjacent filaments. Excluding localized mesh singularities at the load application node and artificial stress concentrations from the rigid contact interface, tensile stresses reached approximately 30–47 MPa at these critical junctions. In contrast, the bulk of the scaffold body exhibited moderate stresses in the range of 5–15 MPa. Because small microstructural heterogeneities and localized stress concentrations act as preferential sites for crack initiation, these intersections represent the primary failure origins, which is highly consistent with the brittle fracture mechanisms typically observed in porous ceramics [24].

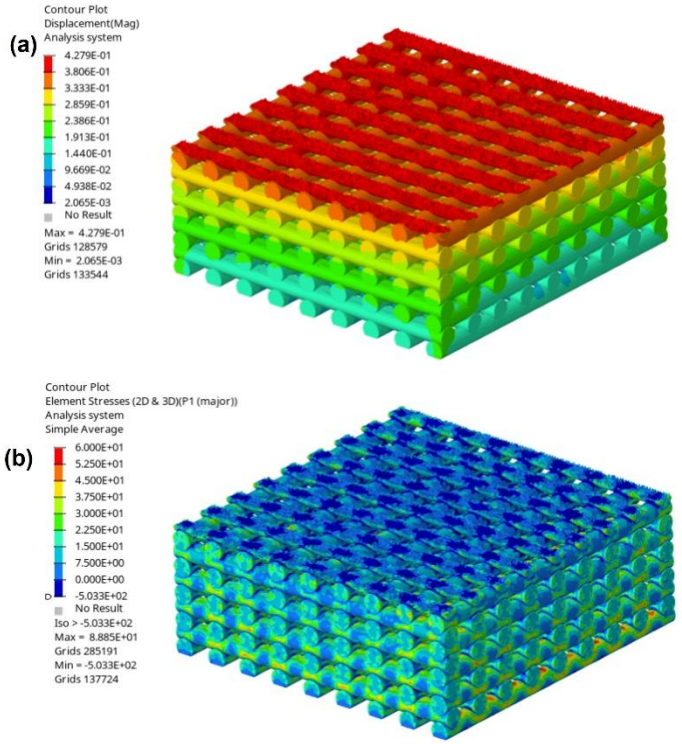


Fig. 5 - Finite element analysis results obtained from the idealized CAD model under simulated compressive loading. (a) Displacement distribution (mm): the hatched outline denotes the initial undeformed configuration, while the colored scaffold represents the deformed state. (b) Maximum stress distribution (MPa): localized tensile stress concentrations are identified predominantly at the filament intersection regions.

IV. CONCLUSIONS

The integration of µ-CT reconstructions and finite element analyses elucidated the critical relationship between the scaffold's microarchitecture and its macroscopic mechanical response. Structurally, the 3D reconstructions confirmed that the scaffolds sintered at 900 °C successfully preserved their designed architecture. This structural fidelity, combined with the enhanced densification achieved during sintering, yielded an experimental compressive strength of approximately 50 MPa.

Computationally, the numerical models demonstrated excellent agreement with the experimental force–displacement data, identifying an effective solid-phase elastic modulus of 1.0 GPa. Furthermore, simulations of the idealized CAD model revealed that internal stress distributions are fundamentally governed by the filament arrangement. Tensile stress concentrations reaching 30–47 MPa were localized specifically at the filament intersections, pinpointing these junctions as primary failure initiation sites. Ultimately, this combined experimental and numerical approach provides a robust

framework for the future design and topological optimization of BioS scaffolds in bone tissue engineering applications.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from the Universal Project (Grant No. 407680/2021-7) and the EMUTISAÚDE/FINEP Project (Grant No. 01.18.0027.00). K.F.S. also acknowledges financial support from CAPES through a scholarship (Grant No. 88887.234026/2025-00). The authors thank LaMaV for providing the Biosilicate used in this work.

REFERENCES

[1] A.M.M. Amin, E.M.M. Ewais, Bioceramic Scaffolds, in: Scaffolds in Tissue Engineering - Materials, Technologies and Clinical Applications, InTech, 2017: p. 13. https://doi.org/10.5772/intechopen.70194.

[2] S. Mukhopadhyay, Bioactive glass-ceramics, in: Fundamental Biomaterials: Ceramics, Elsevier Inc., 2018: pp. 129–152. https://doi.org/10.1016/B978-0-08-102203-0.00006-8.

[3] Jiang. Chang, Xingdong. Zhang, Kerong. Dai, Bioactive Materials for Bone Regeneration, Elsevier Science & Technology, 2020.

[4] M.C. Crovace, M.T. Souza, C.R. Chinaglia, O. Peitl, E.D. Zanotto, Biosilicate® — A multipurpose, highly bioactive glass-ceramic. In vitro, in vivo and clinical trials, J. Non. Cryst. Solids 432 (2016) 90–110. https://doi.org/10.1016/j.jnoncrysol.2015.03.022.

[5] M.R. Azenha, O. Peitl, V.M. da R. Barros, Bone response to biosilicates® with different crystal phases, Braz. Dent. J. 21 (2010) 383–389. https://doi.org/10.1590/S0103-64402010000500001.

[6] D. Desimone, W. Li, J.A. Roether, D.W. Schubert, M.C. Crovace, A.C.M. Rodrigues, E.D. Zanotto, A.R. Boccaccini, Biosilicate ® – gelatine bone scaffolds by the foam replica technique: development and characterization, Sci. Technol. Adv. Mater. 14 (2013) 045008. https://doi.org/10.1088/1468-6996/14/4/045008.

[7] L. Roseti, V. Parisi, M. Petretta, C. Cavallo, G. Desando, I. Bartolotti, B. Grigolo, Scaffolds for Bone Tissue Engineering: State of the art and new perspectives, Materials Science and Engineering: C 78 (2017) 1246–1262. https://doi.org/10.1016/j.msec.2017.05.017.

[8] T. Tariverdian, F. Sefat, M. Gelinsky, M. Mozafari, Scaffold for bone tissue engineering, in: Handbook of Tissue Engineering Scaffolds: Volume One, Elsevier, 2019: pp. 189–209. https://doi.org/10.1016/B978-0-08-102563-5.00010-1.

[9] F. Pecqueux, F. Tancret, N. Payraudeau, J.M. Bouler, Influence of microporosity and macroporosity on the mechanical properties of biphasic calcium phosphate bioceramics: Modelling and experiment, J. Eur. Ceram. Soc. 30 (2010) 819–829. https://doi.org/10.1016/j.jeurceramsoc.2009.09.017.

[10] S.-Y. Fu, Cryogenic Properties of Polymer Materials, in: Polymers at Cryogenic Temperatures, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 9–39. https://doi.org/10.1007/978-3-642-35335-2_2.

[11] P.P. Kamble, S. Chavan, R. Hodgir, G. Gote, K.P. Karunakaran, S. Goli, Multimodal Freezing System for Cryogenic 3D Printing, (2021). https://doi.org/10.21203/rs.3.rs-480124/v1.

[12] Y. Du, H. Liang, D. Xie, N. Mao, J. Zhao, Z. Tian, C. Wang, L. Shen, Finite element analysis of mechanical behavior, permeability of irregular porous scaffolds and lattice-based porous scaffolds, Mater. Res. Express 6 (2019). https://doi.org/10.1088/2053-1591/ab3ac1.

[13] L. D'Andrea, D. Gastaldi, E. Verné, F. Baino, J. Massera, G. Örlygsson, P. Vena, Mechanical Properties of Robocast Glass Scaffolds Assessed through Micro-CT-Based Finite Element Models, Materials 15 (2022). https://doi.org/10.3390/ma15186344.

[14] J.K.M.B. Daguano, F.C. Giora, K.F. Santos, A.B.G.C. Pereira, M.T. Souza, J.L. Dávila, A.C.D. Rodas, C. Santos, J.V.L. Silva, Shear-thinning sacrificial ink for fabrication of Biosilicate® osteoconductive scaffolds by material extrusion 3D printing, Mater. Chem. Phys. 287 (2022). https://doi.org/10.1016/j.matchemphys.2022.126286.

[15] L. Ouyang, 3D Bioprinting and Bioink: Background BT - Study on Microextrusion-based 3D Bioprinting and Bioink Crosslinking Mechanisms, 2019. https://doi.org/10.1007/978-981-13-9455-3_2.

[16] J.M. Zuidema, C.J. Rivet, R.J. Gilbert, F.A. Morrison, A protocol for rheological characterization of hydrogels for tissue engineering strategies, J. Biomed. Mater. Res. B Appl. Biomater. 102 (2014) 1063–1073. https://doi.org/10.1002/jbm.b.33088.

[17] Z. Fu, S. Naghieh, C. Xu, C. Wang, W. Sun, X. Chen, Printability in extrusion bioprinting, Biofabrication 13 (2021) 033001. https://doi.org/10.1088/1758-5090/abe7ab.

[18] A. Mari, M. Pavarini, P.F. Menci, C. Charbonneau, L.P. Lefebvre, L. De Nardo, Impact of aging on the sintering behavior of bioactive-glass powder, Journal of Materials Research and Technology 25 (2023) 3704–3711. https://doi.org/10.1016/j.jmrt.2023.06.204.

[19] A.R. Boccaccini, Q. Chen, L. Lefebvre, L. Gremillard, J. Chevalier, Sintering, crystallisation and biodegradation behaviour of Bioglass®-derived glass-ceramics, Faraday Discuss. 136 (2007) 27–44. https://doi.org/10.1039/b616539g.

[20] X. Kuang, G. Carotenuto, L. Nicolais, A Review of Ceramic Sintering and Suggestions on Reducing Sintering Temperatures, Kluwer Academic Publishers, 1997.

[21] Z. Chen, J. Che, Y. Xu, J. Linghu, L. Wang, X. Wang, Dynamic capture of cracks propagation behavior and thermal transport within high entropy ceramics under elevated thermal stress field, Ceram. Int. (2026). https://doi.org/10.1016/j.ceramint.2026.08.407.

[22] E. Apel, J. Deubener, A. Bernard, M. Höland, R. Müller, H. Kappert, V. Rheinberger, W. Höland, Phenomena and mechanisms of crack propagation in glass-ceramics, J. Mech. Behav. Biomed. Mater. 1 (2008) 313–325. https://doi.org/10.1016/j.jmbbm.2007.11.005.

[23] M. Montazerian, E.D. Zanotto, Bioactive and inert dental glass-ceramics, J. Biomed. Mater. Res. A 105 (2017) 619–639. https://doi.org/10.1002/jbm.a.35923.

[24] N. Abbasi, S. Hamlet, R.M. Love, N.-T. Nguyen, Porous scaffolds for bone regeneration, Journal of Science: Advanced Materials and Devices 5 (2020) 1–9. https://doi.org/10.1016/j.jsamd.2020.01.007.

[25] L. Tabard, V. Garnier, E. Prud'Homme, E.J. Courtial, S. Meille, J. Adrien, Y. Jorand, L. Gremillard, Robocasting of highly porous ceramics scaffolds with hierarchized porosity, Addit. Manuf. 38 (2021) 101776. https://doi.org/10.1016/j.addma.2020.101776.

[26] W. Xiao, M.A. Zaeem, G. Li, B. Sonny Bal, M.N. Rahaman, Tough and strong porous bioactive glass-PLA composites for structural bone repair, J. Mater. Sci. 52 (2017) 9039–9054. https://doi.org/10.1007/s10853-017-0777-3.

Hydrogen Peroxide Plasma Sterilization of 3D-Printed PCL and PCL/Bioglass for Biomedical Applications

Thaís Amaral França
Postgraduate Program on Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, SP - Brazil
thaisamaralfranca15@gmail.com

Domênico Cutrupi
Postgraduate Program on Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, SP - Brazil
domenico.cutrupi@ufabc.edu.br

Ana Paula Resende Lauro
Graduate Program in Materials Science Engineering
Federal University of ABC
Santo André, SP - Brazil
ana.lauro@ufabc.edu.br

Anne Cristiane Chinellato
Graduate Program in Materials Science Engineering
Federal University of ABC
Santo André, SP - Brazil
anne.chinellato@ufabc.edu.br

Janaina de Andrea Dernowsek
Research and Development Department
Quantis Biotechnology
São Paulo - Brazil
janaina.dernowsek@quantis.bio

Sônia Maria Malmonge
Postgraduate Program on Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, SP - Brazil
sonia.malmonge@ufabc.edu.br

Abstract— **Polycaprolactone (PCL) is widely used to fabricate scaffolds for tissue engineering, but sterilization may alter its thermal and mechanical properties or produce cytotoxic residues. This study evaluated the effects of hydrogen peroxide plasma sterilization on 3D-printed PCL and PCL/Bioglass composite specimens containing 20 wt% Bioglass (PCL/BG). Filaments were produced by extrusion and used to manufacture specimens by fused deposition modeling. The specimens were sterilized using a 55-min hydrogen peroxide plasma cycle and characterized by thermogravimetric analysis, tensile testing, and an indirect cytotoxicity assay using L929 fibroblasts. Sterilization caused no macroscopic changes or statistically significant alterations in the elastic modulus, yield stress, or yield strain of either material. The thermal degradation onset temperatures remained practically unchanged, with values of 378 and 380 °C for nonsterile and sterilized PCL specimens, respectively, and 275 °C for both nonsterile and sterilized PCL/BG specimens. Cell viability remained close to or above 90% in all groups, with no statistically significant differences among treatments ($p > 0.05$). The PCL/BG specimens exhibited lower thermal stability and yield stress than neat PCL, which may be associated with polymer degradation during composite-filament processing rather than with sterilization. These findings support its use as a low-temperature sterilization method for PCL-based materials intended for biomedical and tissue-engineering applications.**

Keywords **- PCL; Bioglass, Plasma Sterilization; 3D Printing.**

I. Introduction

Polycaprolactone (PCL) is a semicrystalline aliphatic polyester widely used in the development of biomaterials, scaffolds, and medical devices due to its biocompatibility, biodegradability, processability, and structural stability in long-term applications [1]. In tissue engineering, PCL has been extensively investigated for the fabrication of three-dimensional structures using techniques such as 3D printing, electrospinning, and porous molding-based methods. These techniques enable the production of scaffolds with controlled architecture, interconnected porosity, and mechanical properties tailored to the intended application. However, despite its advantages, the functional performance of PCL depends directly on its physical properties, including surface morphology, crystallinity, thermal stability, molar mass, dimensional integrity, and mechanical strength [2, 3, 4, 5].

For PCL-based biomaterials to be used in biological assays or clinical applications, the sterilization step is indispensable. The presence of microbial contaminants can compromise material safety, interfere with cellular responses, and render its biomedical application unfeasible [6]. However, the choice of sterilization method represents a challenge, since physical, chemical, or radiative processes may alter the properties of the polymer. Methods such as ethylene oxide (EtO) (ISO 11135), gamma radiation (ISO 11137), hydrogen peroxide plasma (HPP) (ISO 14937), ultraviolet radiation (ISO 14937), alcohol (ISO 14937), peracetic acid (ISO 14937), ozone (ISO 14937), dry heat (ISO 20857), and autoclaving (ISO 17665) have been used to sterilize or disinfect polymeric materials, but their effects vary according to the composition, geometry, porosity, and manufacturing method of the material [7, 8].

Among the available sterilization methods, EtO and gamma radiation are the most widely used for PCL-based materials in the biomedical field. Ethylene oxide operates at low temperatures and low relative humidity, which minimizes thermomechanical stress on the polymer matrix . Studies have shown that EtO sterilization preserves the structural integrity, surface texture, crystallinity, and shape memory behavior of PCL-based scaffolds, making it a suitable option when dimensional fidelity and morphological preservation are critical requirements [9]. However, residual EtO absorbed by the material may require extended aeration periods before biological use, and concerns regarding cytotoxic residues and environmental hazards have been

raised [10]. Gamma radiation, in turn, is a penetrating and highly effective sterilization method, but its interaction with PCL chains can induce simultaneous chain scission and cross-linking events [11]. Additionally, gamma irradiation has been reported to lead to a significant decrease in molecular weight accompanied by an increase in crystallinity of PCL fiber mats, though mechanical properties did not always change significantly [12, 13]. These structural modifications may influence the kinetics of degradation, cell–scaffold interactions, and long-term mechanical performance of the device.

HPP sterilization has emerged as a promising low-temperature alternative, particularly suited for heat-sensitive polymeric biomaterials. The method is based on the generation of a plasma state in which hydrogen peroxide is ionized, producing reactive free radicals that non-specifically interact with molecules essential to microbial metabolism and reproduction, resulting in rapid and effective microbicidal action at temperatures generally below 50 °C [14]. At the end of the sterilization cycle, the process is designed to promote the decomposition of hydrogen peroxide into water vapor and oxygen [15]. However, potential retention of process-related species and material compatibility should be evaluated for each material and device geometry, particularly for porous or complex structures. Therefore, characterization of the effects of HPP sterilization on both material properties and biological response remains important for PCL-based biomedical devices.

Accordingly, the present study aimed to investigate the effect of H_2O_2 plasma sterilization method on the physical properties of 3D printed specimens of PCL and PCL/Bioglass (80/20 w/w) composite, hereafter referred to as PCL/BG , contributing to the definition of safer and more suitable sterilization protocols for biomedical and tissue engineering applications.

II. Materials and methods

A. Materials

Polycaprolactone (PCL; average Mn 80,000; product no. 440744) was purchased from Sigma-Aldrich/Merck). Bioglass (bioactive glass ceramic 45S5) was supplied by Vetra Biomaterials, Brazil)

B. Preparation of PCL and PCL/BG (80/20 w/w) filament

PCL and PCL/BG filaments with a diameter of 1.75 mm were produced using a twin-conical-screw microextruder (MC 40, Xplore) at a temperature of 150 °C and screw rotation speeds ranging from 100 to 250 rpm.

For the PCL/BG (80/20) composite, PCL and BG particles with a diameter of 5 µm were initially mixed and homogenized using a LAB 16 single-screw mini-extruder (AX Plásticos). The extruder was operated at a screw speed of 40 rpm and a maximum torque of 28 N·m, using a temperature profile of 140, 130, and 115 °C. The resulting composite extrudate was subsequently pelletized and used as feedstock in the MC 40 microextruder to produce filaments with a controlled diameter.

C. 3D printing of PCL and PCL/BG

The test specimens were manufactured by fused deposition modeling (FDM) using a Creality Ender-3 S1 3D printer. Pure PCL specimens were printed at a nozzle temperature of 190 °C, whereas the PCL/bioglass composite specimens were printed at 180 °C. For both compositions, the build plate temperature was maintained at 38 °C. The three-dimensional models were prepared for printing using PrusaSlicer, with a rectilinear infill pattern and 100% infill density. Except for the nozzle temperature, all printing parameters were kept constant to minimize variability between samples.

D. Sterilization by hydrogen peroxide plasma

The 3D printed PCL and PCL/BG specimens were placed in Tyvek packaging (DuPont, Michigan, USA) and sterilized using the 55-min "rapid cycle" of a STERRAD 100S system (Advanced Sterilization Products/Johnson & Johnson, New Jersey, USA). This sterilization protocol was validated in previous work using the direct inoculation method introducing microorganisms into a culture medium which confirmed the absence of microbial growth in the sterilized samples [16].

E. Thermal analysis

Thermogravimetric analysis was performed on the raw material, filament and 3D printed PCL and PCL/BG specimens before and after sterilization by using a Shimadzu DTG-60H equipment. Samples were heated from room temperature to 500ºC at a heating rate of 10ºC/min under a nitrogen atmosphere at a flow rate of 50mL/min to obtain an inert atmosphere.

F. Tensile stress x strain test

Tensile tests were performed on 3D printed PCL and PCL/BG specimens before and after sterilization using an INSTRON 3369 universal testing machine. A crosshead speed of 5mm/min was used. The values of elastic modulus, yield stress and yield strain were determined and reported as the mean values obtained from three specimens (n=3).

G. *In Vitro cytotoxicity assay*

Indirect cytotoxicity was evaluated according to ISO 10993-5 using L929 fibroblasts. The assay aimed to determine whether hydrogen peroxide plasma sterilization induced the release of cytotoxic compounds from the 3D-printed PCL and PCL/bioglass specimens. Samples sterilized by UV irradiation were used as a comparison group. Material extracts were prepared by incubating the specimens in complete culture medium at an extraction ratio of 0.1 g/mL for (24 ± 2) h at (37 ± 1) °C. L929 fibroblasts were seeded at a density of 1×10^4 cells per well and allowed to adhere for 24 h. The culture medium was then replaced with the respective material extracts, and the cells were exposed for an additional 24 h. Cells maintained in complete culture medium were used as the viable control, whereas cells exposed to culture medium containing 50% DMSO were used as the positive cytotoxicity control. Cell metabolic activity was assessed using 0.01% resazurin, and cell viability was calculated

relative to the viable control. Statistical analysis was performed using an ordinary one-way ANOVA followed by Dunnett's multiple comparisons test, with each material group compared with the viable control. Homogeneity of variances was assessed using the Brown–Forsythe test, and statistical significance was set at $p < 0.05$. The positive cytotoxicity control was used to confirm assay responsiveness and was excluded from the inferential statistical analysis.

III. Results and Discussion

A. PCL and PCL/Bioglass filament

After the processing difficulties associated with PCL and the PCL/BG composite had been overcome, filaments with a diameter of 1.75 mm were successfully obtained (Fig. 1a). The processing difficulties encountered with pure PCL were probably related to chain entanglement and intermolecular interactions resulting from the high molar mass of the PCL used.

The addition of Bioglass altered the melt characteristics and caused darkening of the PCL/BG filament. Moisture adsorbed onto Bioglass as well as the presence of ions present in the biomaterial are critical factors in composite production since during thermal processing, water and ions may catalyze the hydrolysis of PCL ester bonds, reducing the molar mass of the polymer and causing filament darkening [17,18].

B. 3D printing of PCL and PCL/BG

The specimens were successfully fabricated by FDM using the optimized printing parameters, as shown in Table I. The printed samples presented a well-defined geometry and preserved the overall shape of the digital model. Continuous filament deposition was observed throughout the printing process, resulting in specimens with regular layer stacking and no evident macroscopic defects, such as warping, delamination, or incomplete filling. The use of 100% rectilinear infill produced fully filled specimens suitable for subsequent characterization and testing. Representative images of the printed specimens are shown in Figure 1b and 1c.

TABLE I. Printing parameters for PCL and PCL/BG

Parameters	*PCL*	*PCL/BG*
Nozzle temperature (ºC)	190	180
Bed temperature (ºC)	38	38
Extrusion multiplier	1	1
Fill (%)	100	100
Speed (mm/s)	20	20

No macroscopic differences were observed in either the PCL or PCL/BG specimens before and after sterilization (Fig. 1b and 1c).

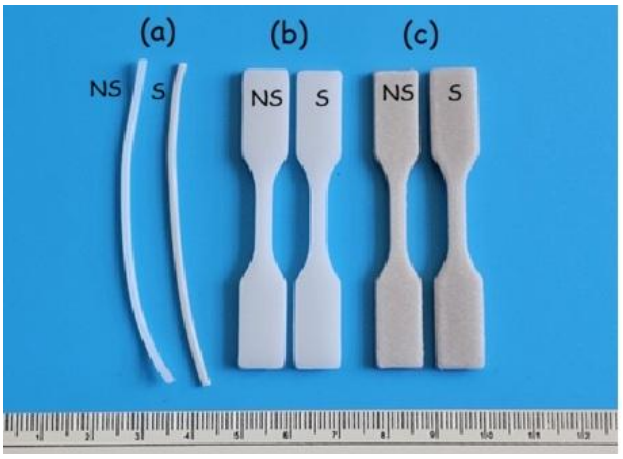


Fig. 1. Photographs of (a) PCL and PCL/BG filament, (b) PCL specimens produced by 3D printing before and after H_2O_2 plasma sterilization, and (c) PCL/BG specimens produced by 3D printing before (NS) and after (S) H_2O_2 plasma sterilization.

C. Thermogravimetric Analysis

TGA analysis is an important method for evaluating the thermal stability of polymeric materials and it can also provide an indirect assessment of polymer degradation because the thermal stability depends partly on the average molar mass of the polymer. Fig. 2 shows the TGA curves of PCL raw material, PCL and PCL/BG filaments and 3D printed specimens before and after hydrogen peroxide plasma sterilization. In this study, thermal stability was evaluated using the onset temperature of thermal degradation (T*onset*), in accordance with ISO 11358-1 standard. Two distinct sets of curves were observed: one for neat PCL and another for the PCL/BG composite. The thermal stability of the composite was significantly lower than that of neat PCL. This difference was probably caused by degradation during the processing and fabrication of the PCL/BG filament, as discussed in Section III-A. However, the T*onset* values of the 3D-printed specimens remained practically unchanged after sterilization, as shown in Table II. This finding indicates that the sterilization process did not cause detectable thermal degradation of the 3D-printed PCL or PCL/BG specimens. .

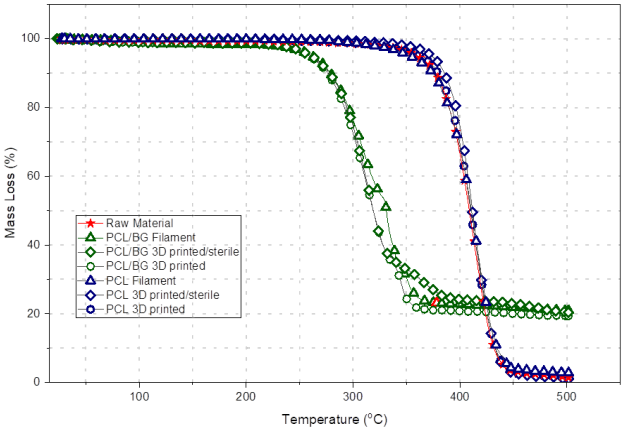


Fig. 2. TGA curves of PCL raw material, PCL and PCL/BG filaments; and PCL and PCL/BG 3D printed specimens before and after hydrogen peroxide plasma sterilization.

TABLE II. VALUES OF ONSET TEMPERATURE (TONSET) AND RESIDUAL MASS DETERMINED FROM THE TGA CURVES..

sample	T*onset* (ºC)	Residual mass (%)
RM	378	zero
PCL filament	372	zero
PCL 3d printed	378	zero
PCL 3d printed sterile	380	zero
PCL/BG filament	278	20 %
PCL/BG 3d printed	275	20 %
PCL/BG 3d printed sterile	275	20 %

The residual mass of 20% observed for the composite samples corresponds to the bioactive-glass content of the composite. Bioactive glass does not undergo thermal degradation below 500 °C under an inert atmosphere.

The TGA results also indicated that 3D printing did not cause detectable degradation of PCL or PCL/BG, even though the printing temperatures were considerably higher than the melting temperature of PCL. This conclusion is supported by the similar T*onset* values obtained for the materials before and after 3D printing.

D. Tensile stress x strain Behaviour

The mechanical properties of a polymeric material can provide an indirect indication of degradation caused by polymer-chain scission because these properties depend partly on the average molar mass of the polymer. In this study, the mechanical properties of the 3D-printed specimens were determined from tensile stress-strain curves. Figure 3 presents the tensile behavior of the specimens, and Table III summarizes their mechanical properties.

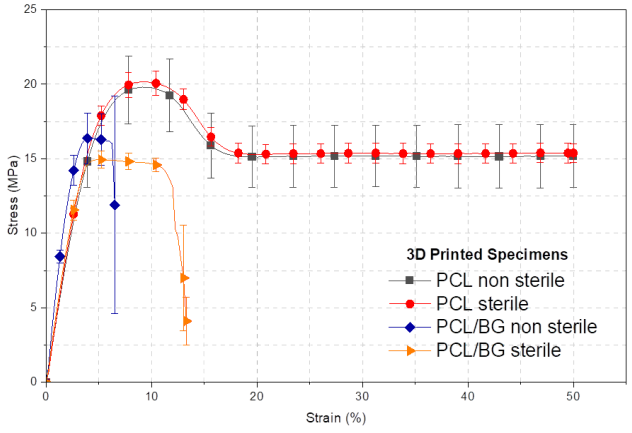


Fig. 3. Tensile stress x strain average curves (n=3) of PCL and PCL/BG 3D printed specimens before and after hydrogen peroxide plasma sterilization.

All specimens exhibited typical ductile behavior. This behavior was particularly evident for neat PCL, which did not rupture even at strains exceeding 50%. The elastic modulus, yield stress, and yield strain of the 3D-printed PCL specimens did not differ significantly before and after sterilization, as shown in Table III.

Although the PCL/BG composite specimens also exhibited ductile behavior, they had a higher elastic modulus and lower yield stress and yield strain than pure PCL. An increase in elastic modulus is expected when a rigid filler is incorporated into a polymer matrix. The lower yield-stress values of the PCL/BG composite relative to pure PCL may provide further evidence of PCL degradation during composite-filament production.

TABLE III. MECHANICAL PROPERTIES OF 3D PRINTED PCL AND PCL/BG SPECIMENS.

Properties	***PCL***		***PCL/BG***	
	Non sterile	Sterile	Non sterile	Sterile
Elastic Modulus (MPa)	4.2 ± 0.5	4.5 ± 0.2	5.3 ± 0.5	4.4 ± 0.2
Yield Stress (MPa)	14.6 ± 0.5	14.2 ± 0.5	12.3 ± 0.7	12.4 ± 0.6
Yield Strain (%)	3.9 ± 0.3	3.6 ± 0.3	2.1 ± 0.2	2.8 ± 0.2

The results indicate that hydrogen peroxide plasma sterilization did not cause detectable degradation of the materials investigated. No statistically significant differences were observed between the mechanical properties of the 3D-printed PCL or PCL/BG specimens before and after sterilization.

E. *In Vitro cytotoxicity assay*

The indirect cytotoxicity assay demonstrated that extracts obtained from all 3D-printed specimens maintained high L929 cell viability after 24 h of exposure (Figure 4). Mean cell viability was 97.79% for PCL/BG-HPP, 90.89% for PCL/BG-UV, 99.82% for PCL-HPP, and 101.7% for PCL-UV, relative to the viable control. No overall differences were detected among the evaluated groups (one-way ANOVA, $F(4,55) = 0.5216$, $p = 0.7202$). Moreover, Dunnett's multiple comparisons test showed that none of the material groups differed significantly from the viable control ($p > 0.05$).

These findings indicate that neither the PCL nor the PCL/bioglass specimens released compounds capable of producing a detectable cytotoxic response under the evaluated extraction and exposure conditions. Importantly, the high viability observed for the HPP-sterilized groups indicates that no detectable extract-mediated cytotoxic effect was associated with the sterilization process under the conditions evaluated. Although hydrogen peroxide sterilization cycles are designed to promote peroxide decomposition, this assay does not directly quantify residual hydrogen peroxide or other process-related species. This observation is particularly relevant because sterilization procedures may potentially alter polymeric materials or leave process-related residues that could affect cellular responses. Nevertheless, the viability of cells exposed to extracts from HPP-sterilized PCL and PCL/BG specimens remained close to that of the viable control.

The presence of bioglass also did not result in a detectable cytotoxic effect. Although the PCL/BG-UV group showed the lowest mean viability (90.89%), this reduction was not statistically significant compared with the viable control ($p = 0.6411$), and viability remained high. Furthermore, the comparable cellular responses observed for PCL and PCL/BG specimens suggest that the incorporation of bioglass, under the conditions investigated, did not

introduce soluble components at concentrations sufficient to impair L929 metabolic activity.

Taken together, these results support the cytocompatibility of both 3D-printed PCL and PCL/BG specimens and indicate that hydrogen peroxide plasma sterilization can be applied without compromising their biological response in the indirect cytotoxicity assay. However, as this assay evaluates the effects of soluble compounds released from the materials over a 24 h exposure period, these findings specifically demonstrate the absence of detectable extract-mediated cytotoxicity and should not be interpreted as a comprehensive assessment of long-term cellular interactions with the materials.

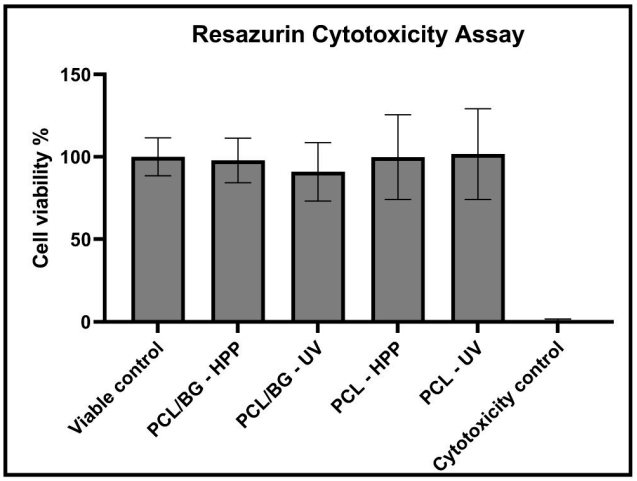


Fig. 4. L929 cell viability after 24 h of exposure to extracts obtained from 3D-printed PCL and PCL/BG specimens subjected to hydrogen peroxide plasma sterilization or treatment with UV irradiation. Data are presented as mean ± standard deviation. No statistically significant differences were observed among the experimental groups ($p > 0.05$).

IV. Conclusion

The results of this study demonstrate that hydrogen peroxide plasma sterilization is a suitable low-temperature method for 3D-printed PCL and PCL/BG biomedical devices. No macroscopic changes were observed after sterilization, and the thermal degradation onset temperatures remained practically unchanged for both materials. The tensile tests also showed no differences in the elastic modulus, yield stress, or yield strain between nonsterile and sterilized specimens. Furthermore, the indirect cytotoxicity assay showed high L929 cell viability in all groups, indicating that the indirect cytotoxicity assay showed no detectable extract-mediated cytotoxicity associated with the sterilized specimens under the evaluated conditions.

The lower thermal stability and yield stress of PCL/BG compared with neat PCL were most likely associated with degradation during composite-filament processing rather than with hydrogen peroxide plasma exposure. Overall, hydrogen peroxide plasma sterilization preserved the evaluated thermal, and mechanical properties of the 3D-printed materials while maintaining their in vitro cytocompatibility. These findings support the use of this sterilization method for PCL-based scaffolds intended for biomedical and tissue-engineering applications. Further studies should evaluate its effects on surface properties, long-term degradation, and biological performance under application-specific conditions.

Acknowledgment

CAPES and FINEP (process 01.22.0141.00-Ref 091/21) for financial support.

References

[1] M. A. Woodruff and D. W. Hutmacher, “The return of a forgotten polymer—Polycaprolactone in the 21st century,” *Prog. Polym. Sci.*, vol. 35, no. 10, pp. 1217–1256, 2010.

[2] S. Biscaia *et al*., “3D printed poly(ε-caprolactone)/hydroxyapatite scaffolds for bone tissue engineering,” *Int. J. Mol. Sci.*, vol. 23, no. 4, Art. no. 2318, 2022.

[3] M. Gharibshahian, M. Salehi, N. Beheshtizadeh, M. Kamalabadi-Farahani, A. Atashi, M.-S. Nourbakhsh, and M. Alizadeh, “Recent advances on 3D-printed PCL-based composite scaffolds for bone tissue engineering,” *Front. Bioeng. Biotechnol.*, vol. 11, Art. no. 1168504, 2023.

[4] H.-Y. Liang, W.-K. Lee, J.-T. Hsu, J.-Y. Shih, T.-L. Ma, T. T. T. Vo, C.-W. Lee, M.-T. Cheng, and I.-T. Lee, “Polycaprolactone in bone tissue engineering: A comprehensive review of innovations in scaffold fabrication and surface modifications,” *J. Funct. Biomater.*, vol. 15, no. 9, Art. no. 243, 2024.

[5] F. Ramírez-Ruiz, I. Núñez-Tapia, M. C. Piña-Barba, M. A. Alvarez-Pérez, V. Guarino, and J. Serrano-Bello, “Polycaprolactone for hard tissue regeneration: Scaffold design and in vivo implications,” *Bioengineering*, vol. 12, no. 1, Art. no. 46, 2025.

[6] J. Horakova *et al*., “Impact of various sterilization and disinfection techniques on electrospun poly-ε-caprolactone,” *ACS Omega*, vol. 5, pp. 8885–8892, 2020.

[7] K. Havlickova *et al*., “The impacts of the sterilization method and the electrospinning conditions of nanofibrous biodegradable layers on their degradation and hemocompatibility behavior,” *Polymers*, vol. 16, Art. no. 1029, 2024.

[8] T. L. S. Farias *et al*., “Sterilization by ozone: Effects on electrospun polycaprolactone membrane properties and cell viability,” *Int. J. Biomater.*, vol. 2025, 2025.

[9] C. J. Houk, F. O. Beltran, and M. A. Grunlan, “Suitability of EtO sterilization for polydopamine-coated, self-fitting bone scaffolds,” *Polym. Degrad. Stab.*, vol. 194, Art. no. 109763, 2021.

[10] M. Elkhashab, Z. Sartawi, D. Lynch, S. Vucen, and A. Crean, “Impact of gamma irradiation and ethylene oxide sterilisation on the stability of crystalline and amorphous drug solids,” *Eur. J. Pharm. Sci.*, vol. 224, Art. no. 107572, 2026.

[11] M. Hoseini, S. Hamidi, E. Salehi, A. Mohammadi, F. Mirhoseini, and M. Ravaghi, “Multi-variate multi-objective optimization of production conditions for electro-spun skin scaffold using RSM and investigation of gamma irradiation effects on the properties of the optimized sample,” *Heliyon*, vol. 10, no. 12, Art. no. e32941, 2024.

[12] D. de Cassan, A. L. Hoheisel, B. Glasmacher, and H. Menzel, “Impact of sterilization by electron beam, gamma radiation and X-rays on electrospun poly-(ε-caprolactone) fiber mats,” *J. Mater. Sci. Mater. Med.*, vol. 30, no. 4, Art. no. 42, 2019.

[13] L. Benecke, O. Lohrberg, F. Alt, M. T. Müller, C. Cherif, J. Neunzehn, and B. Kruppke, “Influence of sterilization regime on mechanical properties of calcium carbonate-reinforced polycaprolactone scaffolds for bone regeneration,” *J. Biomed. Mater. Res. A*, vol. 114, no. 1, Art. no. e70003, 2026.

[14] W. A. Rutala, D. J. Weber, and Healthcare Infection Control Practices Advisory Committee, *Guideline for Disinfection and Sterilization in Healthcare Facilities*. Atlanta, GA, USA: Centers for Disease Control and Prevention, 2008, updated 2019. [Online]. Available: https://www.cdc.gov

[15] S. Crow and J. H. Smith III, “Gas plasma sterilization—Application of space-age technology,” *Infect. Control Hosp. Epidemiol.*, vol. 16, no. 8, pp. 483–487, Aug. 1995.

[16] M. M. O. Simbara, “Fabricação e caracterização de um protótipo de arcabouço para engenharia tecidual de válvulas cardíacas“ thesis (doctoral), Instituto Dante Pazzanese de Cardiologia, São Paulo, Brasil, 2019.

[17] I. Armentano *et al*., “Biodegradable polymer matrix nanocomposites for tissue engineering: A review,” *Polym. Degrad. Stab.*, vol. 95, pp. 2126–2146, 2010.

[18] A. P. Resende Lauro, “Desenvolvimento de biocompósitos de policaprolactona/biovidro (PCL/BV) para impressão 3D,” M.S. thesis, Programa de Pós-Graduação em Ciência e Engenharia de Materiais, Universidade Federal do ABC, Santo André, Brazil, 2025.

Poly(Lactic Acid) Films as Substrate for Cell Culture: Hydrolytic Degradation and *In Vitro* Biocompatibility

Mônica H. M. Nascimento
Centro de Ciências Naturais e Humanas
Universidade Federal do ABC (UFABC)
Santo André, Brasil
monica.nascimento@ufabc.edu.br

Karolina K. Garcia
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
kaminski.k@aluno.ufabc.edu.br

Christiane B. Lombello
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC (UFABC)
São Bernardo do Campo, Brasil
christiane.lombello@ufabc.edu.br

***Abstract*—Tissue engineering is a way to help organs and living tissues grow back after they are damaged by disease and cannot repair themselves easily. One important step in this field is to take cells from the patient's own body. These cells are removed from the original tissue and grown on special surfaces like dense films or three-dimensional scaffolds. The films or scaffolds can be made from natural or synthetic materials. After the cells grow, the whole structure is put back into the patient's body. In this context, this study evaluated the hydrolytic degradation behavior and cell interaction aspects of dense poly(lactic acid) (PLA) films, a material widely investigated as a substrate for tissue engineering. PLA films were obtained by solvent evaporation and subjected to *in vitro* degradation protocols, 70 °C (accelerated degradation) for up to 7 days, with monitoring of mass loss and macro- and microstructural morphology. In parallel, the adhesion and growth of Vero cells on the films were evaluated for up to 7 days of culture. The results indicated progressive mass loss and opacification of the films throughout degradation, in the accelerated assay, and show that Vero cells were able to adhere, spread, and possibly proliferate on the substrate, maintaining morphology typical of the cell line and without apparent signs of cytotoxicity. It is concluded that PLA constitutes a biodegradable and biocompatible substrate with potential application for tissue regeneration, like bone, cartilage or skin.**

***Index Terms*—Poly(lactic acid); Biodegradation; Tissue engineering; Cell culture; Biocompatibility.**

I. Introduction

The difficulty of certain injured organs and tissues to repair themselves spontaneously has, in recent decades, driven the development of tissue engineering as an approach that combines cells, biomaterials, and signaling molecules to induce tissue regeneration [1], [2]. Biomaterials, presented as three-dimensional scaffolds, or dense films, provide temporary structural support for cell adhesion, migration, and proliferation and should later be resorbed as a new extracellular matrix deposited by the cells themselves [1], [3].

Among the polymeric biomaterials used for this purpose, the poly(α-hydroxy acids) stand out — a class of synthetic aliphatic polyesters that includes poly(lactic acid) (PLA), poly(caprolactone) (PCL), poly(glycolic acid) (PGA), and poly(p-dioxanone) (PDO) [4]. PLA is a promising alternative due to its biocompatibility and biodegradability. Furthermore, its degradation products are eliminated through the body's own metabolic pathways, which reduces postoperative risks and costs [1], [4].

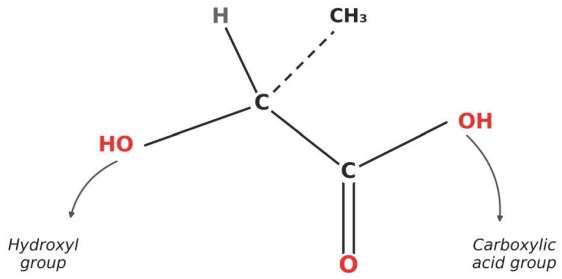


Fig. 1. Biofunctionality of lactic acid: hydroxyl and carboxylic acid groups of the monomeric unit [4], [5].

PLA degradation occurs predominantly through simple hydrolysis of the ester bonds in the main chain, forming hydroxyl and carboxyl end groups (Fig. 2).

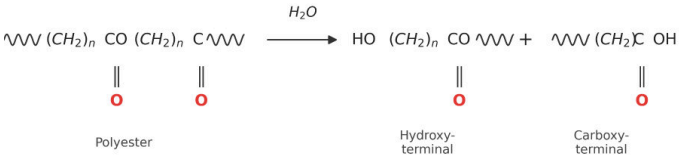


Fig. 2. Hydrolysis of the polyester ester bonds, forming hydroxyl and carboxyl terminal groups [1], [5].

The resulting oligomers are converted into lactate and pyruvate and, in the presence of acetyl-CoA, are incorporated

into the Krebs cycle, ultimately being eliminated as CO_2 and H_2O through respiration and urine (Fig. 3) [1], [6].

The kinetics of this degradation depend on factors such as molar mass, crystallinity, shape, thickness, pH, and medium temperature [4], [7]. The literature reports resorption times for PLA ranging from a few weeks to several years, depending on the stereoisomer and the morphology of the device [8].

Bioresorbable PLA devices have already been used *in vitro* as a scaffold for different cell types, including bone, chondral, and epidermal cells, with no evidence of toxicity, with cells adhering, proliferating, and producing extracellular matrix components [9]. Recent studies support this potential in bone, cartilage, and meniscus tissue engineering, while also pointing to limitations of pure PLA, such as low initial cell adhesion and local acidification during degradation, which have driven the development of blends and composites [10], [11], [12]. These limitations are starting points for improvement rather than reasons to discard the material: PLA's hydrophobicity and low surface energy can be reduced through plasma treatments, adhesion-protein coatings, or bioactive fillers, while its mechanical fragility and degradation kinetics can be adjusted through copolymerization (as in PLGA) or blending with natural polymers [10], [12]. Understanding how these limitations appear under simple *in vitro* conditions, like those used in this study, is a necessary step before moving to material modification strategies or more complex *in vivo* assays.

Thus, this study aimed to characterize the hydrolytic degradation behavior of PLA films under 70 °C and to evaluate the adhesion, spreading, and morphological maintenance of Vero cells cultured on this substrate, contributing to the discussion of its potential as a substrate for tissue regeneration.

II. Materials and Methods

A. Preparation of PLA films

Dense films were obtained by dissolving PLA (PLA2002D, NatureWorks®/Cargill) in chloroform at 10% (w/v), at room temperature. The solution was poured onto glass plates and kept in a fume hood for approximately one week, until complete evaporation of the solvent.

B. Hydrolytic degradation assay

In vitro degradation protocol was carried out, adapted from Barbanti et al. [4]. Three rectangular samples (2.5 × 1.0 cm) were sterilized in 70% alcohol, immersed in buffer solution (pH 7.2) and kept in an oven at 70 °C for 1 and 7 days, characterizing an accelerated degradation assay, following ISO 13781 [13]. The results were presented as mean ± standard deviation.

C. Cell culture on PLA

Fibroblast-like Vero cells (african green monkey kidney, reference CCL-81, ATCC) were cultured in medium 199 supplemented with 10% fetal bovine serum and antibiotic, at 37 °C. High-density inocula (1×10^3 cells/mL) were seeded onto PLA samples previously sterilized with ethylene oxide, with medium changes every 3 days. The positive control consisted of cells cultured directly on the plate. Adhesion was evaluated by phase-contrast microscopy after 30 min and 2 h of incubation, and cell morphology after 2, 5, and 7 days, by staining with toluidine blue and methylene blue under inverted and upright light microscopes (Axiovert A1, Zeiss).

III. Results

A. Hydrolytic degradation

In the accelerated degradation assay at 70 °C (Fig. 3), mass loss was much more pronounced, reaching about 25% within just 7 days, highlighting the strong dependence of degradation kinetics on medium temperature.

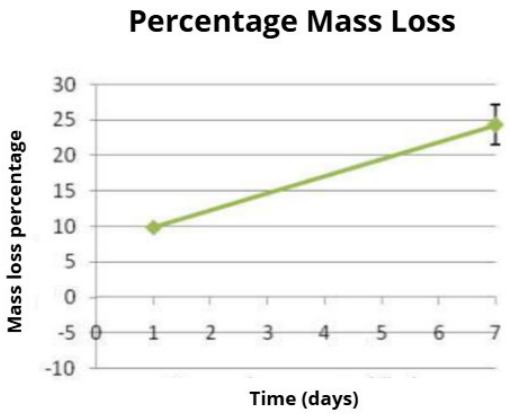


Fig. 3. Percentage mass loss of PLA during accelerated degradation at 70 °C. The results were made in triplicate and are presented as mean ± standard deviation.

Macroscopically (Fig. 4), the initially transparent films became opaque as degradation progressed, an effect that was much more pronounced and rapid at 70 °C, where deformation was already observed after 1 day and complete fragmentation after 7 days.

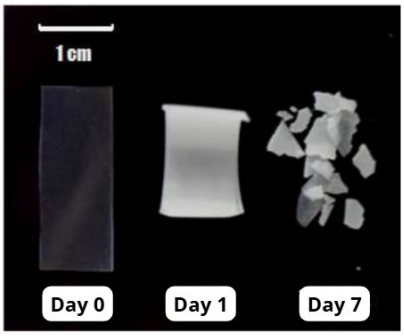


Fig. 4. Macroscopic appearance of the PLA films during accelerated degradation at 70 °C (0, 1, and 7 days).

Stereomicroscopic analysis (Fig. 5) confirms this pattern: at 70 °C, surface cracks are already observed on the first day, progressing to intense brittleness and surface irregularity after 7 days.

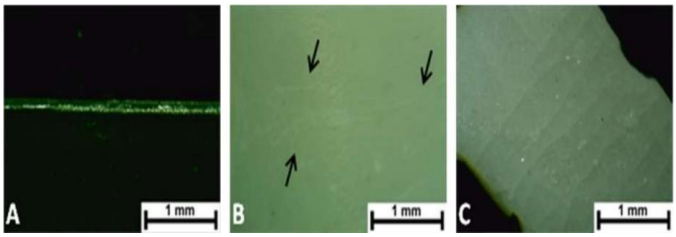


Fig. 5. Stereomicroscopy of the PLA films during accelerated degradation at 70 °C: (A) before degradation; (B) 1 day; (C) 7 days. Magnification: 30×.

B. Cell interaction with PLA

The adhesion process of Vero cells to the substrate is illustrated in Fig. 6. Up to 30 min of incubation, the cells appeared rounded and in suspension. After 2 h, the onset of cell spreading on the film could already be observed.

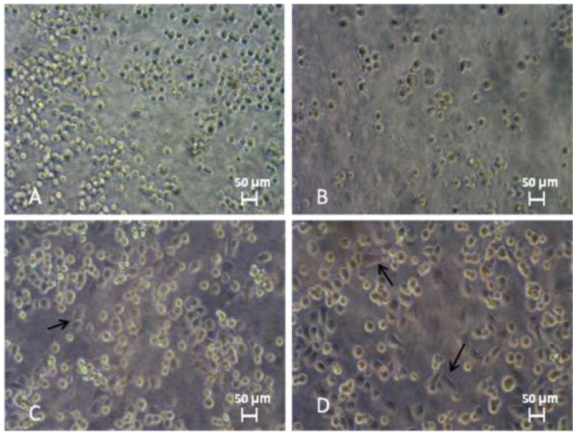


Fig. 6. Adhesion of Vero cells on PLA by phase-contrast microscopy: (A, B) 30 min; (C, D) 2 h of incubation.

Within 2 days of culture (Fig. 7), the cells already formed a monolayer at the edge of the material, similar to the positive control, and on the film itself they appeared slightly less spread, with an estimated confluence of 60–70%.

At 5 days (Fig. 8), the typical fibroblast morphology was maintained, with increased confluence (∼80–90%) both in the control and on the biomaterial, with no signs of cell degeneration.

After 7 days (Fig. 9), a further increase in cell density on the PLA was observed, with the formation of a confluent monolayer of fibroblastic morphology, both at the edge and on the surface of the film, indicating maintained viability throughout the entire evaluation period.

IV. Discussion

In vitro degradation assays represent a low-cost simulation of PLA behavior after implantation, although they do not fully reproduce factors present *in vivo*, such as inflammatory response and local vascularization [14]. The temperature increase in the accelerated assay (70 °C) markedly intensified degradation kinetics as expected from the literature on this material [15], [16], [17].

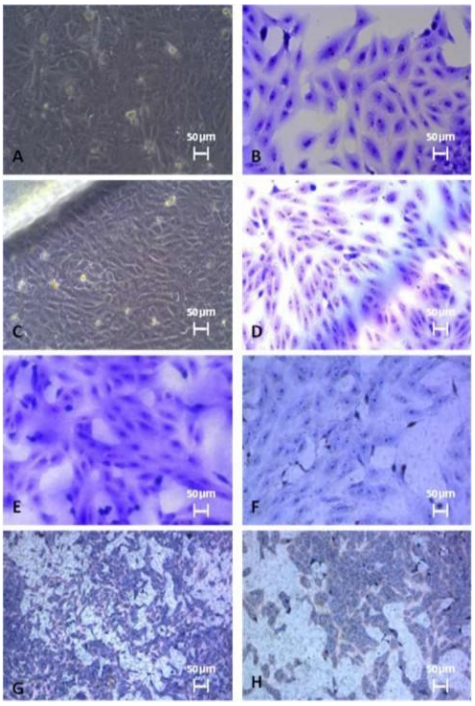


Fig. 7. Vero cells after 2 days of culture: (A, B) positive control; (C–H) cells at the edge and on the PLA film.

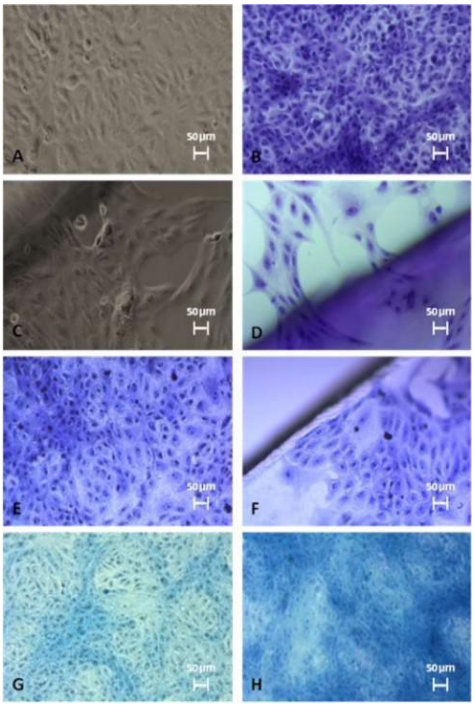


Fig. 8. (A-H) Vero cells after 5 days of culture, stained with toluidine blue and methylene blue.

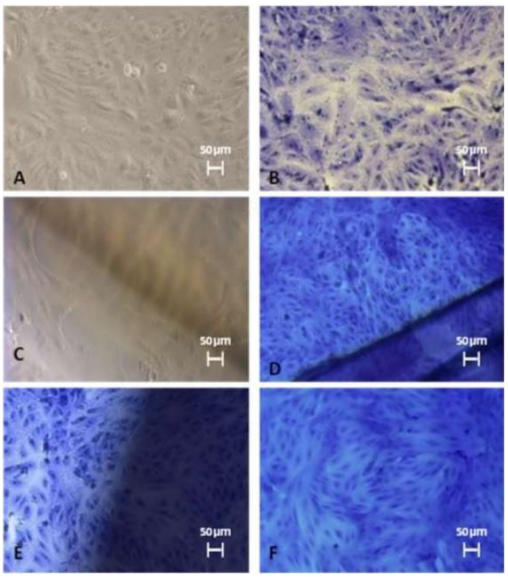


Fig. 9. Vero cells after 7 days of culture at the edge and on the PLA film, stained with toluidine blue.

The increasing opacity of the films throughout degradation is associated with pore formation and the rearrangement of crystalline regions of the polymer, since hydrolysis occurs preferentially in the amorphous phase, and the relative crystalline fraction tends to increase over the course of the process, altering the optical and mechanical properties of the material [4].

Regarding cell interaction, the observed pattern of initial adhesion with rounded cells, followed by progressive spreading and the formation of confluent monolayers, is consistent with the classic sequence described for fibroblasts on synthetic substrates [18]. The maintenance of the typical Vero cell line morphology, both on PLA and in the positive control, throughout the 7 days of culture suggests the absence of relevant cytotoxic effects of the substrate, in agreement with previous results obtained with PLA films applied [9].

These results reinforce the potential of PLA as a biodegradable substrate for tissue regeneration, applicable to contexts such as bone, cartilage, and skin repair, in agreement with the literature [11], [12].

It is also important to weigh these findings against the intrinsic limitations of PLA as a biomaterial. Its hydrophobic nature and the low density of anchoring sites for adhesion proteins may explain the slightly lower cell spreading observed on the film compared to the positive control in the first days of culture. Likewise, the accumulation of acidic hydrolysis byproducts within the material, evident in the pronounced brittleness observed in the 70 °C assay, is a phenomenon widely described as a potential trigger of inflammatory response in *in vivo* applications, although this was not directly assessed in the present in vitro study [10], [11]. These limitations open avenues for future work related to the incorporation of bioactive fillers or surface treatments (plasma, grafting of adhesion molecules) to improve the initial cell response [16], [19], as well as the development of blends or copolymers (such as PLA/PLGA or PLA combined with natural polymers) to modulate degradation kinetics and mitigate local acidification, and the fabrication of three-dimensional, porous scaffolds instead of dense films, so as to favor nutrient diffusion and cell infiltration in contexts closer to clinical application [20], [21]. Other biological and physico-chemical characterizations, such as quantitative cytotoxicity, film's thickness, wettability, roughness, residual solvent and scanning electron microscopy surface analyses will be considered in future studies.

V. Conclusion

The hydrolytic degradation of PLA films was pronounced under accelerated conditions (70 °C), accompanied by its opacification and evident structural weakening, reinforcing the fact that PLA does not constitute a permanent material. The cell interaction assays showed that Vero cells were able to adhere, spread, and proliferate onto the PLA films, in a monolayer, maintaining the morphology characteristic of the Vero cell line and with no evidence of cytotoxicity. Together, these results support PLA as a biodegradable and biocompatible material, with potential application as a substrate for cell growth in tissue regeneration strategies, as for bone, cartilage and skin regeneration.

Acknowledgments

The authors thank the Universidade Federal do ABC (UFABC) for the laboratory infrastructure provided for this study. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] S. H. Barbanti, C. Zavaglia, and E. Duek, "Polímeros bioreabsorvíveis na engenharia de tecidos," Polímeros: Ciência e Tecnologia, vol. 15, pp. 13–21, 2005.

[2] J. Melrose, C. Chuang, and J. Whitelock, "Tissue engineering of cartilages using biomatrices," J. Chem. Technol. Biotechnol., vol. 83, pp. 444–463, 2008.

[3] S. Huang and X. Fu, "Naturally derived materials-based cell and drug delivery systems in skin regeneration," J. Control. Release, vol. 142, pp. 149–159, 2010.

[4] V. D. Jahno, "Síntese e caracterização do poli(L-ácido láctico) para uso como biomaterial," Dissertação (Mestrado), UFRGS, 2005.

[5] H. Bergwerf, "MolView," 2014. [Online]. Available: https://molview.org/. Accessed: Jul. 15, 2026.

[6] M. J. Lyndon, T. W. Minett, and B. J. Tighe, "Cellular interactions with synthetic polymer surfaces in culture," Biomaterials, vol. 6, pp. 396–402, 1985.

[7] A. C. Motta and E. A. R. Duek, "Estudo inicial da degradação in vitro de poli(L-co-DL ácido lático) sintetizado em laboratório," Rev. Matéria, vol. 13, 2008.

[8] C. C. Chen, J. Y. Chueh, H. Tseng, H. M. Huang, and S. Y. Lee, "Preparation and characterization of biodegradable PLA polymeric blends," Biomaterials, vol. 24, pp. 1167–1173, 2003.

[9] M. H. M. Nascimento, S. M. Malmonge, and C. B. Lombello, "PLA films as substrates for growing cells," in Proc. XII Simpósio Brasileiro de Matriz Extracelular; VII Int. Symp. on Extracellular Matrix, Buzios, RJ, Brazil, SIMEC, 2013, p. 28.

[10] M. S. Alavi et al., "Applications of poly(lactic acid) in bone tissue engineering: a review article," J. Biomed. Mater. Res., 2023.

[11] S. Liu, S. Qin, M. He, D. Zhou, Q. Qin, and H. Wang, "Current applications of poly(lactic acid) composites in tissue engineering and drug delivery," Int. J. Biol. Macromol., 2020.

[12] E. Capuana, F. Lopresti, M. Ceraulo, and V. La Carrubba, “Poly-L-lactic acid (PLLA)-based biomaterials for regenerative medicine: a review on processing and applications,” Polymers, 2022.
[13] International Standards Organization, “ISO 13781: Resinas e formas fabricadas de poli(L-lactato) para implantes cirúrgicos — Ensaio de degradação in vitro,” 1998.
[14] C. C. Santos and A. R. Santos Jr., “Metodologia com biomateriais para estudo de regeneração óssea,” III Simpósio de Iniciação Científica da UFABC, 2010.
[15] L. E. Claes, A. A. Ignatius, K. E. Rehm, and C. Scholz, “New bioresorbable pin for the reduction of small bony fragments: design, mechanical properties and in vitro degradation,” Biomaterials, vol. 17, pp. 1621–1626, 1996.
[16] D. da Silva, M. Kaduri, M. Poley, O. Adir, N. Krinsky, J. Shainsky-Roitman, and A. Schroeder, “Biocompatibility, biodegradation and excretion of polylactic acid (PLA) in medical implants and theranostic systems,” Chem. Eng. Journal, vol. 340, pp. 9–14, 2018.
[17] S. Lyu, J. Schley, B. Loy, D. Lind, C. Hobot, R. Sparer, and D. Untereker, “Kinetics and time-temperature equivalence of polymer degradation,” Biomacromolecules, vol. 8, pp. 2301–2310, 2007.
[18] J. L. Dewez et al., “Adhesion of mammalian cells to polymer surfaces: from physical chemistry of surfaces to selective adhesion on defined patterns,” Biomaterials, vol. 19, pp. 1441–1445, 1998.
[19] S. Castañeda-Rodríguez, M. González-Torres, and R. M. Ribas-Aparicio, “Avanços recentes em poli (ácido láctico) modificado como materiais de engenharia de tecidos,” J. Biol. Eng., vol. 17, p. 21, 2023.
[20] J. S. Watai, “Degradação hidrolítica do poli (ácido láctico) (PLA) e de compostos de PLA/nanohidroxiapatita,” Dissertação (Mestrado), FEI, 2018.
[21] S. Mantha, S. Pillai, P. Khayambashi, A. Upadhyay, Y. Zhang, O. Tao, H. M. Pham, and S. D. Tran, “Smart hydrogels in tissue engineering and regenerative medicine,” Materials (Basel, Switzerland), 2019.

Poly(vinyl alcohol)-based hydrogels engineered for sustained non-viral gene delivery: physicochemical characterization and in vitro transfection assessment

1st Jailson de A. Santos
Departament of Energy Engineering
Aeronautics Institute of Technology (ITA)
Fortaleza, Brazil
jailsonsantos@ita.br

2nd Steve Catarino
Coimbra Institute for Clinical and Biomedical Research (iCBR)
University of Coimbra
Coimbra, Portugal
scatarino@fmed.uc.pt

3rd Maria Mendes
Faculty of Pharmacy
University of Coimbra
Coimbra, Portugal
mariamendes1093@gmail.com

4th Dina Murtinho
Department of Chemistry
University of Coimbra
Coimbra, Portugal
dmurtinho@ci.uc.pt

5th Carla Vitorino
Faculty of Pharmacy
University of Coimbra
Coimbra, Portugal
csvitorino@ff.uc.pt

6th Henrique Girao
Coimbra Institute for Clinical and Biomedical Research (iCBR)
University of Coimbra
Coimbra, Portugal
hgirao@ibili.uc.pt

7th Artur J.M. Valente
Department of Chemistry
University of Coimbra
Coimbra, Portugal
avalente@ci.uc.pt

8th Edvani C. Muniz
Department of Chemistry
Federal University of Piauí
Teresina, Brazil
munizec@ufpi.edu.br

Abstract*—Hydrogel-based systems have emerged as promising non-viral platforms for localized gene delivery due to their tunable physicochemical properties and biocompatibility. This study investigated poly(vinyl alcohol) (PVA)-based hydrogels incorporating DNA and modified with chitosan (CHI), poly(ethylene glycol) (PEG), or poly(lactic acid) (PLA) to evaluate their potential for sustained gene delivery. Hydrogels were produced using a freeze-thaw method and characterized by Fourier transform infrared spectroscopy, fluorescence analysis, swelling behavior, DNA release kinetics, thermogravimetric analysis, mechanical and rheological analyses, and scanning electron microscopy. Short-term cytocompatibility was assessed in astrocyte and glioblastoma cell lines, while GFP-associated fluorescence was qualitatively evaluated in SUM-159PT breast cancer cells exposed to plasmid-loaded hydrogels. The results confirmed DNA incorporation and demonstrated that polymer composition influenced swelling behavior, DNA release kinetics, mechanical properties, and the temporal pattern of GFP-associated fluorescence. CHI- and PLA-containing formulations showed persistent GFP-associated fluorescence at later experimental time points, consistent with their sustained DNA-release behavior. All formulations showed favorable short-term cytocompatibility under the conditions evaluated. Overall, polymer-modified PVA hydrogels represent tunable matrices with potential for localized and sustained non-viral DNA delivery.

Keywords— gene delivery, hydrogels, poly(vinyl alcohol), non-viral vectors, tissue engineering, biomaterials.

I. Introduction

Gene therapy has emerged as a promising strategy for the treatment of genetic disorders, cancer, and tissue damage through the delivery of nucleic acids capable of modulating cellular function. However, the clinical translation of gene therapy remains limited by the instability of genetic material, low cellular uptake, and rapid clearance following administration [1].

Hydrogels have attracted considerable attention as non-viral gene delivery systems because of their high water content, biocompatibility, tunable mechanical properties, and ability to provide localized and sustained release [2]. Among hydrogel-forming polymers, poly(vinyl alcohol) (PVA) stands out due to its low toxicity, ease of processing, and capacity to form physically crosslinked networks through freeze-thaw cycles [3].

Strategies involving the incorporation of secondary polymers can further optimize hydrogel performance. Chitosan (CHI) can enhance DNA interactions through electrostatic attraction, poly(ethylene glycol) (PEG) can improve porosity and biocompatibility, and poly(lactic acid) (PLA) can increase structural stability and prolong release profiles [4].

This study aimed to develop and characterize PVA-based hydrogels modified with CHI, PEG, and PLA for sustained non-viral gene delivery. The influence of polymer composition on physicochemical properties, DNA release behavior, short-term cytocompatibility, and GFP-associated fluorescence in plasmid-exposed cells was investigated to assess the potential of these formulations as tunable matrices for sustained DNA delivery.

II. Materials and Methods

A. Hydrogel Preparation

Poly(vinyl alcohol) (PVA)-based hydrogels were prepared by physical crosslinking using a freeze-thaw method. A 20% (w/v) PVA aqueous solution was prepared at 90 °C under constant stirring. Salmon sperm DNA or GFP-encoding plasmid DNA was incorporated into the polymeric matrix to obtain DNA-loaded formulations.

To modulate hydrogel properties, chitosan (CHI), poly(ethylene glycol) (PEG), and poly(lactic acid) (PLA) were individually incorporated into the PVA-based formulations. CHI was dissolved in acetic acid, PEG in ultrapure water, and PLA was initially dissolved in chloroform before incorporation into the PVA-based system. After mixing, the PVA-PLA formulation was maintained at room temperature, allowing chloroform evaporation, and subsequently subjected to the same freeze-thaw protocol used for the other formulations. No additional drying or processing step was specifically applied to the PVA-PLA hydrogel. Residual chloroform was not quantitatively determined. No macroscopic phase separation was visually observed after

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

incorporation of the PLA solution into the PVA-based formulation.

The resulting formulations were designated as PVA-DNA, PVA-CHI-DNA, PVA-PEG-DNA, and PVA-PLA-DNA. Hydrogels were subjected to three freeze-thaw cycles consisting of freezing at −20 °C for 12 h and thawing at 25 °C for 12 h. After gelation, samples were washed with ultrapure water to remove residual reagents.

B. Physicochemical Characterization

The chemical composition of the hydrogels was investigated by attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) in the range of 4000–500 cm^{-1}. DNA incorporation was evaluated using GelStar™ nucleic acid staining, UV-visible absorption spectroscopy, fluorescence spectroscopy, and agarose gel electrophoresis. A calibration curve was established using the PVA-based system as the reference matrix because PVA represented the major polymeric fraction of all hydrogel formulations. Formulation-specific blanks were used to account for background contributions from the respective hydrogel matrices. Hydrogel sample masses were standardized among formulations; however, a complete DNA mass balance including released and residual DNA was not independently determined. Identical amounts of GelStar™ were used for fluorescence-based measurements to ensure comparability among samples.

Hydrogel morphology was examined by scanning electron microscopy (SEM) after freeze-drying and gold sputter coating. Because lyophilization may modify the native hydrated network through ice-crystal formation and sublimation, SEM micrographs were used for qualitative morphological comparison rather than quantitative determination of porosity or pore size.

Swelling behavior was evaluated by immersing dried hydrogels in ultrapure water and sodium bromide (NaBr) solutions at 10 and 100 mM. NaBr was selected based on its suitability for DNA-containing systems and its ability to provide controlled ionic conditions while preserving double-stranded DNA stability [5]. The NaBr experiments were designed to investigate the influence of the ionic environment on hydrogel swelling rather than to reproduce physiological conditions. The swelling ratio was calculated from the difference between swollen and dry weights.

Thermogravimetric analysis (TGA) was performed under a nitrogen atmosphere from 25 °C to 500 °C at a heating rate of 10 °C min^{-1}.

C. Mechanical and Rheological Characterization

Mechanical characterization was performed using a TA.XT Plus Texture Analyzer (Stable Micro Systems Ltd., UK). Texture profile analysis was conducted using a P/10 cylindrical Delrin probe (10 mm diameter). The probe compressed the hydrogel samples at 11 mm s^{-1} to a depth of 3 mm, with a 5-s interval between two consecutive compression cycles. Measurements were performed at 25 °C using eight replicates per formulation. Data were analyzed using Texture Exponent software (version 6.1.16.0). The following mechanical parameters were determined: compressibility, calculated from the area under the first force-time curve (AUC1); hardness, defined as the maximum force during the first compression; cohesiveness, calculated as the AUC2/AUC1 ratio; elasticity, determined from the time ratio between compression cycles; and resilience, calculated from the energy ratio during the first compression.

Tensile strength (TS) and elongation at break (EB%) were additionally determined using the TA.XT Plus Texture Analyzer. Samples were cut into 50 × 10 mm strips and tested using a P/1S probe with double-sided adhesive tape to secure the specimens. Tensile strength was calculated as the maximum force at rupture divided by the cross-sectional area of the specimen, whereas elongation at break was expressed as the percentage increase in specimen length at fracture relative to its initial length.

Rheological measurements were performed using a HAAKE MARS III rheometer (Thermo Fisher Scientific, Germany) equipped with a parallel-plate geometry (35 mm diameter) and a 1 mm gap. Temperature was maintained at 25.0 ± 0.1 °C using a Peltier temperature-control system. Dynamic oscillatory frequency-sweep measurements were performed from 100 to 0.01 Hz at a constant applied stress of 50 Pa. The storage modulus (G′) and loss modulus (G″) were recorded as functions of frequency to characterize the viscoelastic behavior of the hydrogel formulations.

D. DNA Release Studies

DNA release kinetics were evaluated by immersing the hydrogels in phosphate-buffered saline (PBS, pH 7.4) under continuous agitation at 25 °C. The release experiments were performed at 25 °C under standardized physicochemical conditions established for comparison among formulations. Aliquots were collected at predetermined intervals and analyzed by UV-visible spectroscopy at 260 nm. DNA concentrations were determined from a calibration curve established using the PVA-based system as the reference matrix, since PVA represented the major polymeric fraction of all formulations. Formulation-specific hydrogel blanks were used for background correction to minimize contributions from the respective polymeric matrices. Release profiles were fitted using pseudo-first-order and pseudo-second-order kinetic models to describe DNA release behavior and polymer-DNA interactions. Model performance was compared using the coefficient of determination (R^2) and the Akaike Information Criterion (AIC).

E. In vitro Biological Assays

The cytocompatibility of the hydrogels was assessed using immortalized human astrocytes (IM-HA) and glioblastoma cells (U87). Cells were cultured under standard conditions and exposed to the hydrogel formulations for 72 h. Cell metabolic activity was determined using the resazurin reduction assay.

Transfection experiments were conducted using SUM-159PT triple-negative breast cancer cells. Cells were seeded directly onto hydrogels containing GFP-LC3 plasmid DNA and cultured at 37 °C under 5% CO_2. Fluorescence microscopy was performed at different time points to qualitatively assess GFP-associated fluorescence in cells exposed to the DNA-loaded hydrogels. DNA-free and plasmid-free hydrogel controls were included to assess background fluorescence associated with the experimental system. All compared fluorescence images were acquired using identical microscopy and exposure settings.

F. Statistical Analysis

Data are presented as mean ± standard deviation from at least three independent experiments. Statistical analyses included the Shapiro-Wilk test for normality, Levene's test for

homogeneity of variances, one-way analysis of variance (ANOVA) followed by Tukey's post hoc test, and Kruskal-Wallis tests when parametric assumptions were not satisfied. Differences were considered statistically significant when $p < 0.05$.

III. Results and Discussion

A. DNA Incorporation and Chemical Characterization

The successful incorporation of DNA into the PVA-based hydrogels was confirmed through complementary spectroscopic, colorimetric, and electrophoretic analyses (Figure 1).

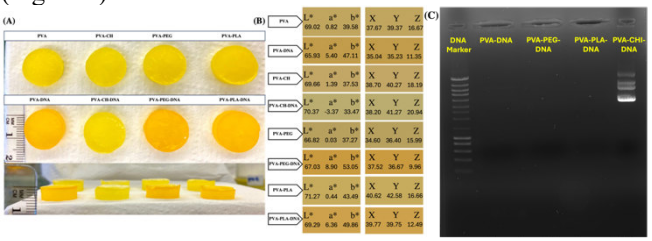


Fig. 1. (A) Visual assessment of PVA-based hydrogels with and without DNA following nucleic acid dye staining. (B) Colorimetric characterization of hydrogels showing L*, a*, b* and XYZ coordinates measured with a Jobin-Yvon Spex Fluorolog 3–22 spectrofluorometer. (C) Agarose gel electrophoresis (8 % gel) confirming DNA presence and quantification in PVA-CHI-DNA hydrogel. DNA bands were visualized using the CaPture Chemiluminescence System.

ATR-FTIR spectra revealed characteristic absorption bands associated with PVA, CHI, PEG, PLA, and DNA, confirming the presence of all components within the hydrogel network (Figure 2).

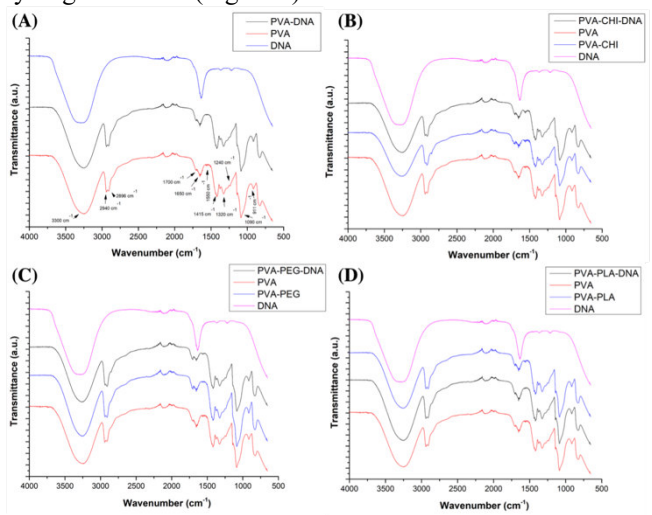


Fig. 2. FTIR-ATR spectra of (A) PVA; (B) CHI; (C) PEG; and (D) PLA with and without DNA.

The broad absorption band observed between 3300 and 2940 cm^{-1}, corresponding to O–H and N–H stretching vibrations, shifted after DNA incorporation, indicating changes in hydrogen-bond interactions within the polymeric matrix [6]. DNA-containing formulations also exhibited increased intensity in the carbonyl stretching region (1700–1650 cm^{-1}), suggesting enhanced molecular interactions between DNA phosphate groups and the hydrogel components [7].

Additional evidence of DNA incorporation was obtained through GelStar™ nucleic acid staining and fluorescence spectroscopy. PVA-DNA, PVA-PEG-DNA, and PVA-PLA-DNA hydrogels displayed strong fluorescence signals, indicating efficient DNA entrapment within the polymer network (Figure 1). In contrast, PVA-CHI-DNA hydrogels showed lower fluorescence intensity but generated distinct DNA bands during agarose gel electrophoresis, suggesting a more reversible interaction between chitosan and DNA.

The absorbance and emission spectra demonstrated that hydrogel composition significantly influenced the local environment surrounding the encapsulated DNA [8] (Figure 3).

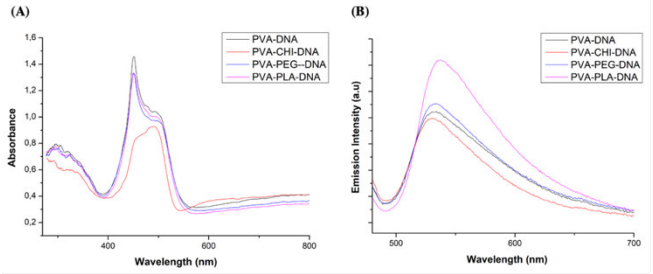


Fig. 3. Absorption (A) and Emission spectra (B) from the GelStar™ Nucleic Acid Gel Stain in the hydrogels with DNA. Excitation at 452 nm.

PVA-PLA-DNA exhibited the highest fluorescence emission intensity, while PVA-CHI-DNA showed fluorescence quenching effects, likely associated with electrostatic interactions between positively charged chitosan and negatively charged DNA phosphate groups [9].

Overall, these findings demonstrate that the polymer composition modulates DNA retention and molecular interactions within the hydrogel matrix, directly influencing the potential performance of these materials as non-viral gene delivery systems.

B. Swelling Behavior and DNA Release Kinetics

The swelling behavior of DNA-loaded hydrogels was evaluated in water and sodium bromide solutions to investigate the influence of ionic strength on hydrogel performance (Figure 4). All formulations exhibited rapid water uptake followed by equilibrium swelling, with distinct profiles depending on polymer composition.

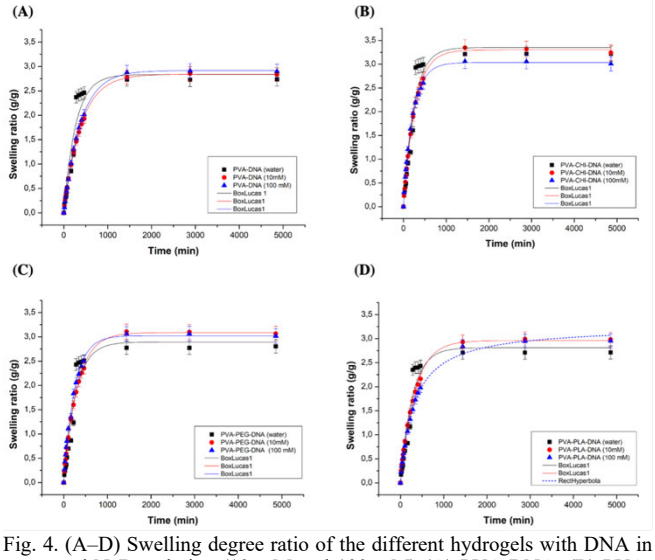


Fig. 4. (A–D) Swelling degree ratio of the different hydrogels with DNA in water and NaBr solution (10 mM and 100 mM) (A) PVA-DNA; (B) PVA-CHI-DNA; (C) PVAPEG-DNA; and (D) PVA-PLA-DNA.

Among the tested formulations, PVA-CHI-DNA showed the highest swelling ratio in all media, reaching values above 3.0 g/g. This behavior can be attributed to the hydrophilic nature of chitosan and its ionizable amino groups, which promote water absorption and network expansion.

NaBr was employed as a controlled ionic medium to investigate the influence of ionic conditions on hydrogel swelling while maintaining conditions compatible with double-stranded DNA stability [5]. Therefore, the NaBr experiments should be interpreted primarily as physicochemical characterization of the polymer networks rather than as a simulation of the physiological environment. Although the present experiments allowed direct comparison among hydrogel formulations under standardized conditions, evaluation in PBS or cell culture medium at 37 °C would provide a closer representation of physiological conditions and should be considered in future studies.

Increasing ionic strength affected the swelling behavior differently depending on the polymer incorporated into the hydrogel matrix. While higher salt concentrations reduced the swelling of PVA-CHI-DNA due to electrostatic shielding effects [10], PVA-PEG-DNA and PVA-PLA-DNA maintained relatively stable swelling profiles, suggesting weaker ionic interactions with DNA.

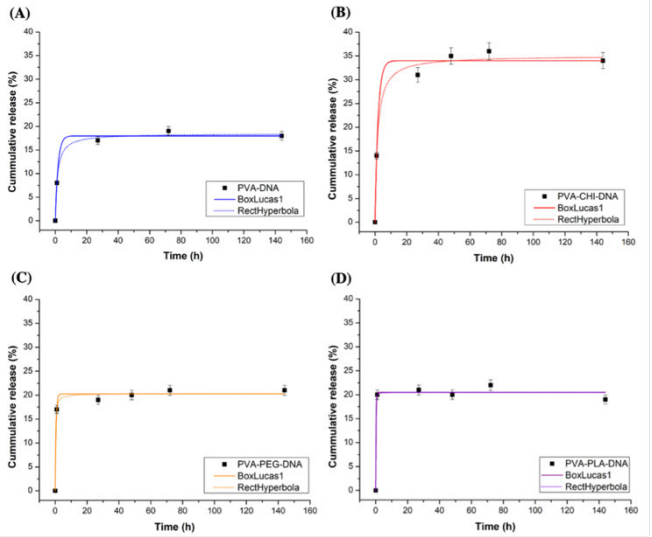


Fig. 5. (A–D) Cumulative DNA release over time for the hydrogels (A) PVA-DNA; (B) PVA-CHI-DNA; (C) PVA-PEG-DNA; and (D) PVA-PLA-DNA.

DNA release studies demonstrated formulation-dependent release profiles (Figure 5). PEG- and PLA-containing hydrogels exhibited comparatively controlled release profiles, whereas the higher swelling capacity of PVA-CHI-DNA may have facilitated DNA diffusion through the polymeric network.

Kinetic modeling indicated that the pseudo-first-order model provided the best empirical fit for most formulations, demonstrating formulation-dependent release kinetics. These findings highlight the importance of polymer selection for tailoring hydrogel performance according to specific therapeutic requirements.

C. Morphological and Mechanical Properties

Scanning electron microscopy of the freeze-dried hydrogels revealed distinct morphological features depending on polymer composition (Figure 6). PVA-DNA exhibited a comparatively homogeneous morphology, whereas incorporation of CHI, PEG, and PLA resulted in visible changes in the apparent surface and internal organization of the dried matrices. PVA-CHI-DNA showed more pronounced void-like structures and surface heterogeneity, whereas PEG- and PLA-containing formulations exhibited comparatively more compact morphologies.

These observations should, however, be interpreted qualitatively. Because SEM analysis was performed after freeze-drying, ice-crystal formation during freezing and subsequent sublimation may modify the native hydrated network and generate or enlarge apparent voids. Therefore, the SEM micrographs do not provide a quantitative measurement of hydrogel porosity or pore size in the hydrated state [11].

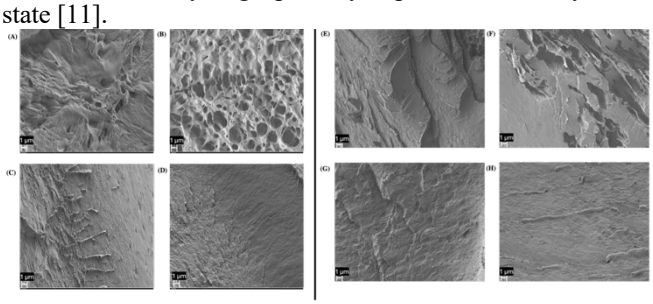


Fig. 6. SEM micrographs of hydrogels with or without DNA: (A) PVA; (B) PVA-CHI; (C) PVA-PEG; (D) PVA-PLA; (E) PVA-DNA; (F) PVA-CHI-DNA; (G) PVA-PEG-DNA; and (H) PVA-PLA-DNA. Magnification: 3000×.

Thermogravimetric analysis demonstrated that polymer incorporation modified the thermal stability of the hydrogels, reflecting differences in intermolecular interactions among the hydrogel components (Figure 7). The observed variations indicate that DNA and polymer additives can either strengthen or weaken the hydrogel network depending on their chemical compatibility [12].

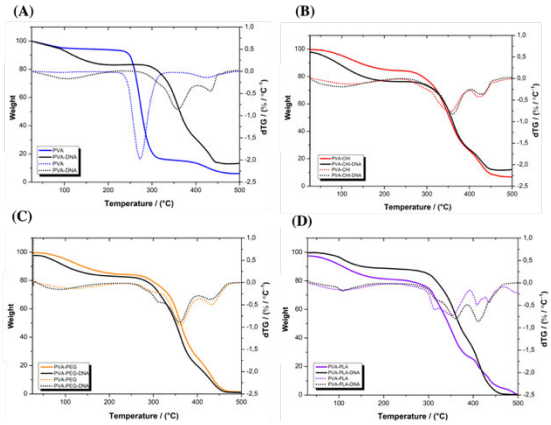


Fig. 7. TG curves (solid line) and respective DTGs (dotted line) for hydrogels without or with DNA: (A) PVA and PVA-DNA; (B) PVA-CHI and PVA-CHI-DNA; (C) PVA-PEG and PVA-PEG-DNA; and (D) PVA-PLA and PVA-PLA-DNA.

Mechanical characterization demonstrated that polymer composition and DNA incorporation influenced the mechanical behavior of the PVA-based hydrogels. Tensile testing revealed formulation-dependent differences in tensile strength and elongation at break (Figure 8), indicating that incorporation of the secondary polymers altered both resistance to deformation and extensibility of the hydrogel matrices.

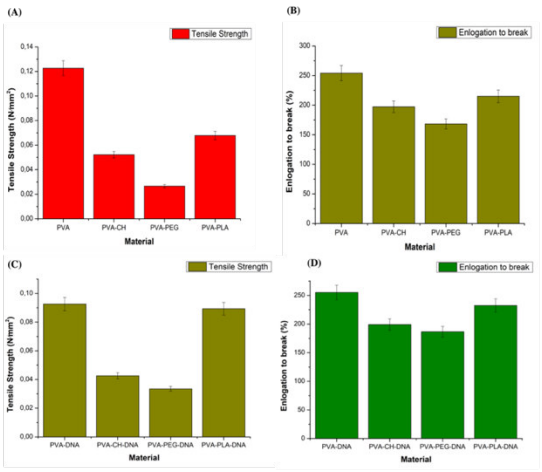


Fig. 8. Tensile strength and Elongation at break for the different hydrogels: (A) and (B) without DNA; and (C) and (D) with DNA.

Oscillatory rheological analysis further demonstrated composition-dependent viscoelastic behavior (Figure 9). The storage modulus (G′) and loss modulus (G″) varied among the formulations over the investigated frequency range, indicating that incorporation of CHI, PEG, PLA, and DNA affected the balance between elastic and viscous responses of the polymer networks. Together, the tensile and rheological results demonstrate that the mechanical properties of the PVA-based hydrogels can be modulated through polymer composition and DNA incorporation.

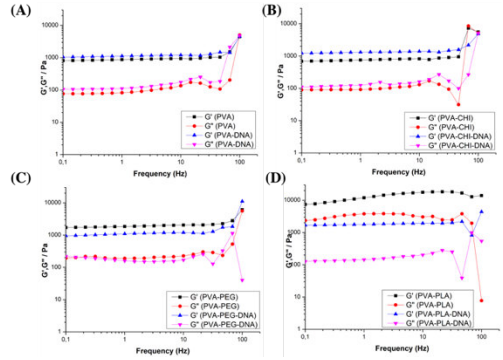


Fig. 9. Mechanical spectra of hydrogels: (A) PVA and PVA-DNA; (B) PVA-CHI and PVA-CHI-DNA; (C) PVA-PEG and PVA-PEG-DNA; and (D) PVA-PLA and PVA-PLA-DNA. Storage modulus (G′) and loss modulus (G″) were recorded at 25 °C over a frequency range of 100–0.01 Hz under a constant applied stress of 50 Pa.

The combination of distinct morphological features, tunable mechanical properties, and composition-dependent thermal behavior further demonstrates the influence of polymer modification on the physicochemical properties of the PVA-based hydrogel matrices.

D. Biocompatibility and Transfection Performance

The cytocompatibility of the developed hydrogels was evaluated using immortalized human astrocytes and glioblastoma cells. Metabolic activity assays demonstrated that all formulations maintained high metabolic activity after 72 h of exposure, indicating favorable short-term cytocompatibility under the experimental conditions evaluated (Figure 10).

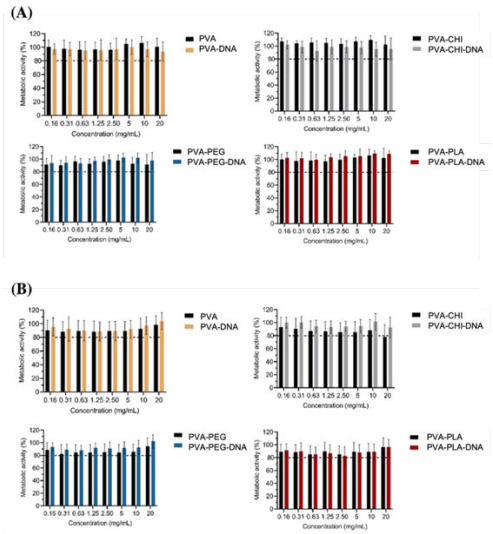


Fig. 10. (A) Metabolic activity of astrocyte cells and (B) U87 cells exposed to different PVA hydrogels after 72 h. Statistical values were calculated using one-way ANOVA. Data are presented as the means ± SD. (n = 9).

The ability of the hydrogels to provide plasmid DNA to SUM-159PT cells was qualitatively investigated using GFP-LC3-containing formulations (Figure 11). DNA-free and plasmid-free hydrogel controls were included to assess background fluorescence associated with the experimental system. The control conditions exhibited background fluorescence, whereas cultures exposed to DNA-loaded hydrogels showed more evident GFP-associated signals.

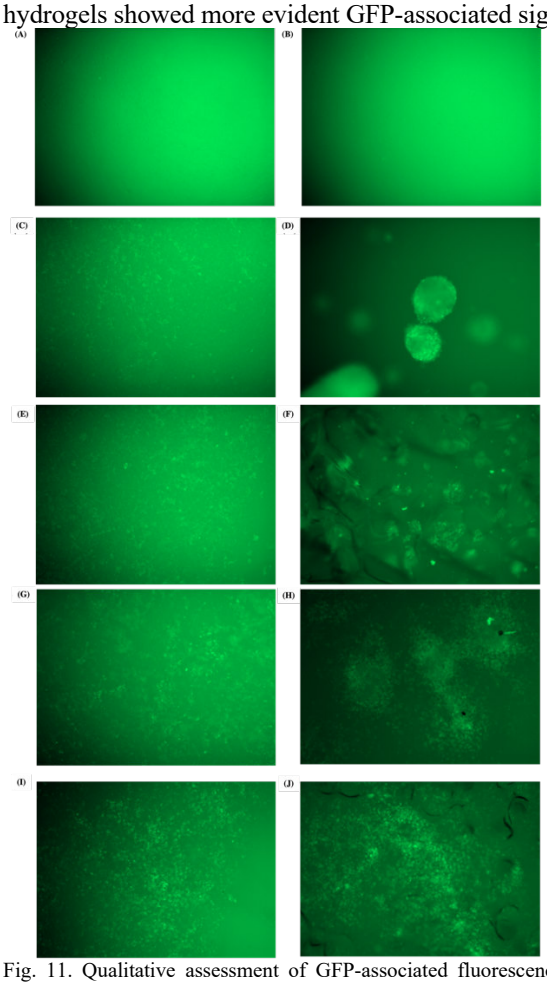


Fig. 11. Qualitative assessment of GFP-associated fluorescence in SUM-159PT breast cancer cells. (A) DNA-free control; (B) plasmid-free hydrogel control; (C–F) cells exposed to PVA-DNA, PVA-CHI-DNA, PVA-PEG-DNA, and PVA-PLA-DNA hydrogels, respectively, after 6 days of incubation; and (G–J) cells exposed to PVA-DNA, PVA-CHI-DNA, PVA-

PEG-DNA, and PVA-PLA-DNA hydrogels, respectively, after 16 days of incubation.

Distinct GFP-associated fluorescence patterns were observed depending on hydrogel composition and incubation time. At the 6-day time point, PVA-DNA and PVA-PEG-DNA formulations showed comparatively more evident fluorescence signals, suggesting greater initial DNA availability to the cells. In contrast, GFP-associated fluorescence remained detectable at later time points in cultures exposed to PVA-CHI-DNA and PVA-PLA-DNA hydrogels, consistent with the sustained DNA-release behavior observed for these formulations [14].

The persistence of GFP-associated fluorescence in the CHI- and PLA-containing formulations suggests that hydrogel composition influences the temporal profile of plasmid availability and the resulting GFP-associated fluorescence signal. However, because transfection was evaluated qualitatively by fluorescence microscopy and long-term metabolic activity was not independently assessed in SUM-159PT cells at the corresponding time points, these observations should not be interpreted as quantitative evidence of transfection efficiency or long-term cytocompatibility.

Overall, these results indicate that polymer modification of PVA hydrogels can modulate DNA-release behavior and the temporal pattern of GFP-associated fluorescence. Further studies incorporating quantitative transfection analysis and matched long-term cytocompatibility measurements are required to establish the efficiency and safety of these formulations for sustained gene-delivery applications.

IV. Conclusion

Polymer-modified PVA hydrogels showed potential as platforms for sustained non-viral DNA delivery. Incorporation of CHI, PEG, and PLA modulated the physicochemical properties of the PVA matrix, including swelling behavior, DNA retention and release profiles. The formulations showed favorable short-term cytocompatibility in IM-HA and U87 cells, while GFP-associated fluorescence was observed in SUM-159PT cultures exposed to plasmid-loaded hydrogels at different experimental time points. In particular, the persistence of fluorescence associated with CHI- and PLA-containing formulations was consistent with their sustained DNA-release behavior. However, quantitative transfection efficiency and long-term cytocompatibility in SUM-159PT cells were not evaluated and should be addressed in future studies. Overall, these findings support further investigation of polymer-modified PVA hydrogels as tunable matrices for localized and sustained DNA delivery under physiologically relevant conditions.

Acknowledgment

The authors gratefully acknowledge the financial and institutional support provided by the Federal University of Piauí (UFPI), the University of Coimbra, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil; Grant No. 88881.711946/2022-01), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil; Grants No. 402030/2022-2, 409258/2022-9, 200173/2023-5, and 406973/2022-9), and the Fundação para a Ciência e a Tecnologia (FCT, Portugal; Project No. 10.54499/2022.06174.PTDC). The Coimbra Chemistry Centre (CQC) is supported by FCT through Project No. 10.54499/UIDB/00313/2020. The authors also acknowledge the institutional support of the Coimbra Institute for Clinical and Biomedical Research (iCBR), the Clinic Academic Center of Coimbra (CACC), and the Center for Innovative Biomedicine and Biotechnology (CIBB), Faculty of Medicine, University of Coimbra.

References

[1] C. Su, D. Lin, X. Huang, J. Feng, A. Jin, F. Wang, Lv Qizhuang, L. Lei, W. Pan, Developing hydrogels for gene therapy and tissue engineering, J. Nanobiotechnol. 22 (1) (2024), https://doi.org/10.1186/s12951-024-02462-z

[2] K. Xu, Q. Zhang, D. Zhu, Z. Jiang, Hydrogels in gene delivery techniques for regenerative medicine and tissue engineering, Macromol. Biosci. 24 (6) (2024), https://doi.org/10.1002/mabi.202300577.

[3] Hossein Omidian, Dey Chowdhury Sumana, R.L. Wilson, Advancements and challenges in Hydrogel engineering for regenerative medicine, Gels 10 (4) (2024) 238, https://doi.org/10.3390/gels10040238, 238.

[4] J.A. Santos, M. Mendes, D. Murtinho, C. Vitorino, E.C. Silva-Filho, A.J.M. Valente, E.C. Muniz, Biopolymer-enhanced PVA hydrogels for DNA delivery: structural and functional characterization, Next Mater. (2025), https://doi.org/10.1016/j.nxmate.2025.100944.

[5] R.S. Dias, L.M. Magno, A.J.M. Valente, D. Das, P.K. Das, S. Maiti, M.G. Miguel, B. Lindman, Interaction between DNA and cationic surfactants: Effect of DNA conformation and surfactant headgroup, J. Phys. Chem. B 112 (46) (2008) 14446–14452, https://doi.org/10.1021/jp8027935

[6] H. Lee, R. Mensire, R.E. Cohen, M.F. Rubner, Strategies for hydrogen bonding based layer-by-layer assembly of Poly(vinyl alcohol) with weak polyacids, Macromolecules 45 (1) (2011) 347–355, https://doi.org/10.1021/ma202092w.

[7] H. Sun, Y. Feng, B. Lei, X. Yin, B. Wu, Molecular simulation of the effects of water on the interface of plant fiber/polyvinyl alcohol (PVA) composites prepared by dry preparation method (DPM), J. Mol. Liq. 391 (2023), https://doi.org/10.1016/j.molliq.2023.123174, 123174.

[8] Elnaz Ahmadi-Sangachin, Javad Mohammadnejad, M. Hosseini, Fluorescence self-assembled DNA hydrogel for the determination of prostate specific antigen by aggregation induced emission, Spectrochim. Acta Part a Molecular and Biomolecular Spectroscopy 303 (2023), https://doi.org/10.1016/j.saa.2023.123234, 123234.

[9] C.T. Cesco, A.J.M. Valente, A.T. Paulino, Methylene blue release from Chitosan/Pectin and Chitosan/DNA blend hydrogels, Pharmaceutics 13 (6) (2021) 842, https://doi.org/10.3390/pharmaceutics13060842, 842.

[10] Dago-Serry Yahnis, K.N. Maroulas, A.K. Tolkou, N.C. Kokkinos, G.Z. Kyzas, How the chitosan structure can affect the adsorption of pharmaceuticals from wastewaters: an overview, Carbohydrate Polymer Technologies and Applications 7 (2024) 100466, https://doi.org/10.1016/j.carpta.2024.100466, 100466.

[11] Rostamabadi Hadis, Y. Bist, Y. Kumar, Meral Yildirim-Yalcin, Tugce Ceyhan, Seid Reza Falsafi, Cellulose nanofibers, nanocrystals, and bacterial nanocellulose: fabrication, characterization, and their most recent applications, Future Postharvest and Food (2024), https://doi.org/10.1002/fpf2.12001.

[12] I. Chrysafi, N.M. Ainali, D.N. Bikiaris, Thermal degradation mechanism and decomposition kinetic studies of Poly(Lactic acid) and its copolymers with Poly (Hexylene succinate), Polymers 13 (9) (2021) 1365, https://doi.org/10.3390/polym13091365

[13] J.D. Torres-Vanegas, J.C. Cruz, L.H. Reyes, Delivery systems for nucleic acids and proteins: barriers, cell capture pathways and nanocarriers, Pharmaceutics 13 (3) (2021) 428, https://doi.org/10.3390/pharmaceutics13030428.

Rheological Assessment of Enzymatic Degradation of Rennova Ultra Volume Lido Facial Filler by Biometil Hyaluronidase at Different UTR Concentrations: An *In Vitro* Study

1st Paula Olimpio da Silva
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
paulaolimpio@peb.ufrj.br

2nd Gabriel Theodoro Olivieri
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
gtholivieri@peb.ufrj.br

3rd Stephani Araujo Cardoso
SkinMat Research Group
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0001-9284-0186

4th Antonio Lima Alves
LABTER
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0009-0003-8259-9204

5th Thaís Nogueira Barradas
SkinMat Research Group
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0002-6558-9595

6th Aline Raybolt dos Santos Almeida
Dept. of Prosthesis and Dental Materials.
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0001-6842-2179

7th Wagner Coelho de Albuquerque Pereira
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
0000-0001-5880-3242

Abstract*—Hyaluronic acid (HA) fillers are the gold standard for facial volumization, with hyaluronidase as the only reversal agent. In Brazil, the commercially available hyaluronidase (Biometil) is measured in Reduced Turbidimetric Units (UTR), a system with no validated equivalence to international standards (IU/USP), which prevents the direct transposition of published dosing protocols. This study characterized, by dynamic oscillatory rheometry (20 °C, 10 Hz, 1 % strain), the enzymatic degradation of Rennova Ultra Volume Lido (23 mg/mL, Hi™ technology) exposed to four concentrations of Biometil hyaluronidase (25, 50, 100 and 150 UTR) and to a control group (0 UTR), with n = 1 sample per condition. The G′ and G″ time series were fitted to an extended exponential model that incorporates the interval between mixing and the first reading, and all estimates reported derive from this model. The time to the in vitro rheological dissolution threshold (G′ = 30 Pa), counted from enzyme–substrate contact, was 136.9 min [95 % CI 136.2–137.6], 28.9 min [28.4–29.3], 14.7 min [14.4–14.9] and 6.0 min [5.5–6.4] for 25, 50, 100 and 150 UTR, the 25 UTR value being obtained by extrapolation beyond the observed window. The G′ = G″ crossover, which marks the effective loss of elastic predominance, occurred at 26.7 min [25.3–29.0] in the 150 UTR group, with G′ = G″ = 4.24 Pa [3.80–4.57], and only by extrapolation in the remaining groups. The model further estimates that 9.7 % [9.4–10.2], 23.2 % [22.3–24.2], 45.3 % [44.9–45.8] and 52.8 % [51.8–54.3] of the elastic modulus had already been lost before the first reading, a proportion that increases with dose. Under the controlled conditions tested, the results indicate dose-dependent rheological susceptibility and provide a preliminary experimental reference, specific to the Rennova Ultra Volume Lido–Biometil pair, within the UTR system. Validation at 37 °C, independent experimental replication and models incorporating tissue diffusion are required before any extrapolation to clinical protocols.

Keywords—hyaluronic acid; hyaluronidase; rheology; dermal filler; elastic modulus; UTR; enzymatic degradation.

I. Introduction

Cross-linked hyaluronic acid (HA) dermal fillers are biomaterials widely used in minimally invasive aesthetic procedures worldwide [1], [2]. Their mechanical properties arise from the three-dimensional polymer network formed by entangled and chemically cross-linked HA chains. The predominance of HA in the aesthetic field is attributed to its biocompatibility, clinical versatility and, above all, to the reversibility conferred by hyaluronidase, which provides a safety margin unique among injectable materials used in clinical practice [2], [4].

Hyaluronidase (EC 3.2.1.35) is an endoglycosidase that depolymerizes HA through hydrolytic cleavage of β-1,4 glycosidic bonds and represents the first-line treatment for filler-related complications, from aesthetic irregularities to vascular occlusions [1], [3], [5]. Despite its clinical importance, its use in this context is considered off-label by both the FDA and ANVISA [2], [3].

Enzymatic cleavage progressively disrupts the HA network, reducing polymer-chain connectivity and consequently altering the balance between elastic and viscous responses of the hydrogel. dose range, , an approximately The efficacy of enzymatic reversal is modulated by the rheological properties of the filler, particularly the elastic modulus G′ and the degree of cross-linking, which may determine resistance to enzymatic penetration through steric hindrance [4], [6].

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Fillers with high G′ tend to require substantially higher doses of hyaluronidase for complete dissolution [5], [6]. In vitro studies that follow this degradation rheologically commonly stop the assay when G′ reaches 30 Pa, a value defined by Faivre et al. as the lowest elastic modulus that can be reliably measured with this type of geometry [6]; it is therefore an operational endpoint of the measurement rather than a physical marker of complete dissolution.

In the Brazilian context there is an additional methodological gap: whereas international studies use hyaluronidase measured in International Units (IU) or USP Units, products available in Brazil — such as Biometil — are measured in Reduced Turbidimetric Units (UTR) [7], a system with no direct equivalence to international standards [2], [7]. This discrepancy prevents the direct transposition of published dosing protocols to national practice.

Rennova Ultra Volume Lido (manufactured by Humedix Co., Ltd., South Korea; distributed in Brazil by Innovapharma) is one of the most widely used fillers in the Brazilian facial harmonization market, formulated with 23 mg/mL of cross-linked HA and 0.3 % lidocaine [9]. Despite its widespread use, there are no independent rheological data or specific enzymatic degradation studies for this product in the literature.

This study aims to characterize, by dynamic oscillatory rheometry with continuous monitoring of G′ and G″, the rheological susceptibility of Rennova Ultra Volume Lido to Biometil hyaluronidase at four concentrations expressed in UTR, generating a preliminary experimental reference specific to this product–enzyme pair in the unit used in Brazil.

II. Materials and Methods

A. Materials

The following materials were used: Rennova Ultra Volume Lido filler (23 mg/mL cross-linked HA + 0.3 % lidocaine; Innovapharma, Brazil); Biometil hyaluronidase (3,000 UTR, Cristália, Brazil); and a manufacturer-supplied diluent (0.45 % sodium chloride and 0.7 % mannitol).

B. Enzymatic Stock Preparation

The Biometil vial (3,000 UTR) was reconstituted with 3.0 mL of the manufacturer's diluent, yielding a working stock of 1,000 UTR/mL (1 UTR/µL). This concentration allows the pipetted enzyme volume to be numerically equal to the target dose in UTR, minimizing pipetting errors.

C. Experimental Groups and Volumes

Each sample was prepared at a total volume of 500 µL in a 1:1 ratio (HA:EDV): 250 µL of filler and 250 µL of enzyme dilution vehicle (EDV, enzyme + diluent), as detailed in Table I. The 1:1 ratio was adopted as an adaptation of the 3:1 ratio described by Safran et al. (2025) [10], validated for peripheral diffusion models; the present protocol employs direct homogenization, which dispenses with the additional volume intended for diffusion. The implications of this choice are discussed in Section IV-E.

TABLE I. Experimental groups and preparation volumes

Group	Dose	HA (µL)	Enzyme (µL)	Diluent (µL)	EDV (µL)	Total (µL)
C	0 UTR	250	0	250	250	500
1	25 UTR	250	25	225	250	500
2	50 UTR	250	50	200	250	500
3	100 UTR	250	100	150	250	500
4	150 UTR	250	150	100	250	500

D. Mixing and Transfer Procedure

Sample preparation was standardized to minimize the interval between homogenization and the start of measurement, since hyaluronidase acts immediately upon contact with HA [5]. The sequence was: (1) aspirate 250 µL of HA into the syringe; (2) add the diluent with a calibrated pipette; (3) add the enzymatic solution with a calibrated pipette; (4) immediately transfer to the rheometer plate; (5) lower the upper geometry; (6) remove excess sample; (7) place the solvent trap; (8) start the measurement. Despite this standardization, a non-zero interval remains between mixing and acquisition of the first data point. This interval was not timed during the assays; it is estimated by the parameter τ of the kinetic model described in Section II-F and interpreted in Section IV-B.

E. Rheological Parameters and Data Acquisition

All measurements were performed on an ARES-G2 rheometer (TA Instruments) with 40 mm parallel-plate geometry, 0.2 mm gap, 20 °C, 1 % strain amplitude and an angular frequency of 62.83 rad/s (10 Hz), recording G′ and G″ over time with one point every 60 s. Acquisition proceeded for 43 to 110 min depending on the group; these windows delimit the estimates flagged as extrapolated in Tables II and III. Each experimental condition corresponds to a single independently prepared sample (n = 1); no independent experimental replicates were performed, owing to limited material. The statistical analysis derives entirely from the confidence intervals estimated for the parameters of the kinetic model (Section II-F). The control group maintained G′ stable throughout the assay, with a coefficient of variation of 0.47 %, confirming that the diluent vehicle has no relevant effect on gel viscoelasticity.

The temperature of 20 °C was adopted in place of the physiological 37 °C after significant solvent evaporation was identified at body temperature in pilot assays, even with the solvent trap — a known limitation of open parallel-plate geometry in low-viscosity systems. This choice reduces the absolute rate of the enzymatic reaction and limits the direct transposition of the times reported here to physiological conditions, as discussed in Section IV-E.

The frequency of 10 Hz was selected on two grounds: maintaining the torque signal above the minimum threshold of the instrument for reliable reading in partially degraded gels, in which lower frequencies yield insufficient torque; and higher temporal resolution of the kinetics, particularly relevant in the high-dose groups [11], [12].

The 1 % strain amplitude was validated by a prior amplitude sweep on the control formulation (0.01 % to 995 % strain, 51 points). The G′ plateau in the low-strain region

was 95.3 Pa, and the limit of the linear viscoelastic region (LVR), taken as the strain at which G′ falls 10 % below the plateau [17], was 15.8 %. At 1 % strain, G′ corresponded to 98.2 % of the plateau value.

F. Data Analysis

The raw G′(t) and G″(t) time series of each treated group were fitted to an extended exponential (Weibull) model, of the same functional form for both moduli:

$$G'(T) = Y_0' \cdot \exp[-(k' \cdot T)^{\beta'}]$$

$$G''(T) = Y_0'' \cdot \exp[-(k'' \cdot T)^{\beta''}], \quad T = t + \tau$$

where T is the time counted from enzyme–substrate contact, t is the time recorded by the rheometer, τ is the interval between contact and the first reading, k is the degradation constant and β is a shape parameter capturing departures from simple first-order kinetics. Y0′ and Y0″ were fixed at the moduli of the control group at the first reading (121.49 Pa and 27.66 Pa), which represent the intact state of the hydrogel before enzymatic action. The parameter τ is a property of the sample, not of the observed modulus; it was therefore estimated in the G′ fit and held fixed in the G″ fit of the same group, so that both models share a single time axis.

The time to the rheological dissolution threshold, T30, defined as the instant at which G′ reaches 30 Pa, follows analytically from the model parameters:

$$T_{30} = [-\ln(30 / Y_0')]^{1/\beta'} / k'$$

The crossover instant, Tcross, at which G′ = G″, is obtained from the positive root of

$$(k' \cdot T)^{\beta'} - (k'' \cdot T)^{\beta''} = \ln(Y_0' / Y_0'')$$

solved numerically, and the corresponding modulus, Gcross, is obtained by substituting Tcross into either equation.

Fitting was performed by nonlinear least-squares regression (SciPy, Python 3), with multiple starting values to reduce the risk of local minima. Confidence intervals of 95 % for the parameters of both models (k′, τ, β′, k″ and β″) and for the quantities derived from them (T30, Tcross and Gcross) were obtained by moving-block bootstrap of the residuals, refitting the model at each iteration; for quantities depending on both moduli, the two series were resampled and refitted jointly.

III. Results

The model described the observed kinetics with R^2 between 0.998 and 0.9997 for G′ (Fig. 1) and between 0.940 and 0.993 for G″ (Fig. 2) in all treated groups. The parameters fitted for G′ and the estimated time to the dissolution threshold are listed in Table II; the same model was fitted to the loss modulus, and the corresponding parameters and the crossover time are given in Table III, laid out with the same columns as Table II to allow direct comparison between the two moduli.

TABLE II. Extended Exponential Model Parameters for the Storage Modulus G′ and the Characteristic Time Derived From It, With 95 % Confidence Intervals

Group	k′ (min⁻¹) [95 % CI]	τ (min) [95 % CI]	β′ [95 % CI]	R^2	T30 (min) [95 % CI]
25 UTR	0.01354 [0.01350–0.01358]	1.10 [1.02–1.25]	0.544 [0.540–0.547]	0.9997	136.9 [a] [136.2–137.6]
50 UTR	0.06183 [0.06035–0.06346]	1.62 [1.42–1.85]	0.579 [0.566–0.593]	0.9995	28.9 [28.4–29.3]
100 UTR	0.13328 [0.13005–0.13645]	2.73 [2.59–2.90]	0.500 [0.494–0.507]	0.9995	14.7 [14.4–14.9]
150 UTR	0.29817 [0.27140–0.33099]	2.06 [1.74–2.42]	0.583 [0.552–0.611]	0.9980	6.0 [5.5–6.4]

Y0′ fixed at 121.49 Pa, storage modulus of the control group at the first reading. T30: time to G′ = 30 Pa, counted from enzyme–substrate contact (T = t + τ). 95 % CI by moving-block bootstrap of the residuals. n = 1 experimental sample per group. [a] Value extrapolated beyond the observed experimental window.

TABLE III. Extended Exponential Model Parameters for the Loss Modulus G″ and the Characteristic Time Derived From It, With 95 % Confidence Intervals

Group	k″ (min⁻¹) [95 % CI]	τ (min)[a]	β″ [95 % CI]	R^2	Tcross (min) [95 % CI]
25 UTR	0.01203 [0.01178–0.01225]	1.10	0.414 [0.405–0.424]	0.9889	1150 [b] [1060–1246]
50 UTR	0.03331 [0.03270–0.03396]	1.62	0.480 [0.468–0.492]	0.9932	152.1 [b] [141.7–163.2]
100 UTR	0.08931 [0.08385–0.09519]	2.73	0.398 [0.379–0.419]	0.9396	126.6 [b] [115.6–139.7]
150 UTR	0.12330 [0.11925–0.12833]	2.06	0.528 [0.508–0.546]	0.9898	26.7 [25.3–29.0]

Y0″ fixed at 27.66 Pa, loss modulus of the control group at the first reading. Tcross: time to the G′ = G″ crossover, counted from enzyme–substrate contact. Columns and conventions identical to those of Table II. [a] τ is not re-estimated in the G″ fit: being a property of the sample rather than of the observed modulus, it was fixed at the value obtained in the G′ fit of the same group (Table II), so that both models share a single time axis. [b] Value extrapolated beyond the observed experimental window.

T30 decreases monotonically with enzyme concentration, from 136.9 min at 25 UTR to 6.0 min at 150 UTR, with confidence intervals that do not overlap between consecutive doses; equivalently, the degradation constant k′ rises from 0.0135 to 0.2982 min^{-1} over the same dose range, an approximately 22-fold increase. A log–log regression between dose and T30 yields an exponent of −1.66 (R^2 = 0.975; 95 % CI: −2.47 to −0.84), compatible both with simple inverse proportionality between dissolution time and enzyme concentration and with a supralinear gain.

The shape parameter β′ remained between 0.50 and 0.58 in all groups, consistently below 1, and the same behavior was observed for β″ (0.40 to 0.53). The estimated values of τ were 1.10, 1.62, 2.73 and 2.06 min for 25, 50, 100 and 150 UTR, compatible in order of magnitude with the manual mixing and transfer procedure described in Section II-D. From these values, the model estimates that 9.7 % [95 % CI 9.4–10.2], 23.2 % [22.3–24.2], 45.3 % [44.9–45.8] and 52.8 % [51.8–54.3] of the elastic modulus had already been lost at the instant of the first reading.

The comparison between Tables II and III shows that, in all groups, k″ is lower than k′ and β″ is lower than β′: the loss modulus decays more slowly than the storage modulus, and

the confidence intervals of k′ and k″ do not overlap at any concentration. As a consequence, the two moduli converge. The crossover times estimated for each dose are compared with the corresponding T30 values in Fig. 3a. The modulus at the crossover instant, Gcross, was 1.42 Pa [95 % CI 1.16–1.70], 3.13 Pa [2.83–3.43], 2.00 Pa [1.64–2.38] and 4.24 Pa [3.80–4.57] for 25, 50, 100 and 150 UTR, respectively (Fig. 3b). Only the 150 UTR group reached the crossover within the experimental window, at 26.7 min [25.3–29.0]; in the remaining groups the crossover is an extrapolation, and in the 25 UTR group it projects beyond 19 h, so that it should be read as not attained within any practically relevant window rather than as a usable time estimate.

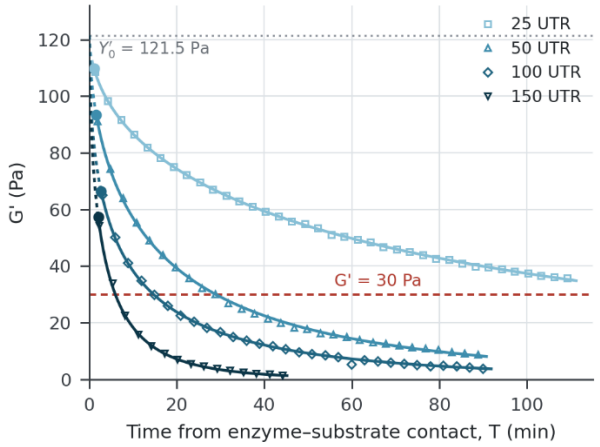


Fig. 1. Storage modulus G′ as a function of the time counted from enzyme–substrate contact (T = t + τ). Markers: measured data. Solid lines: fitted model. Dotted segments: region preceding the first reading, reconstructed by the model. Filled circle: first reading of each group. The dashed line indicates the 30 Pa threshold.

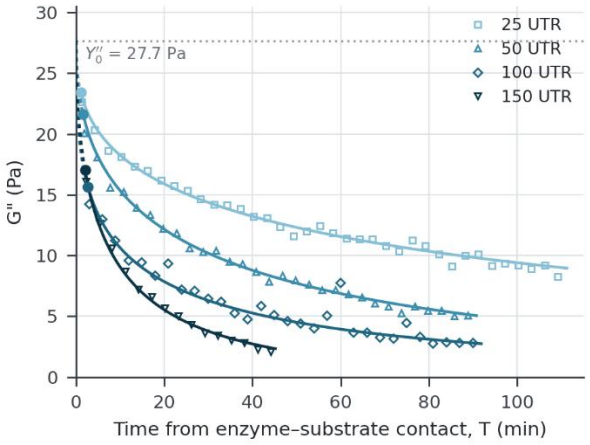


Fig. 2. Loss modulus G″ as a function of the time counted from enzyme–substrate contact, with the same extended exponential model and the same τ as the G′ fit. Graphical conventions identical to those of Fig. 1.

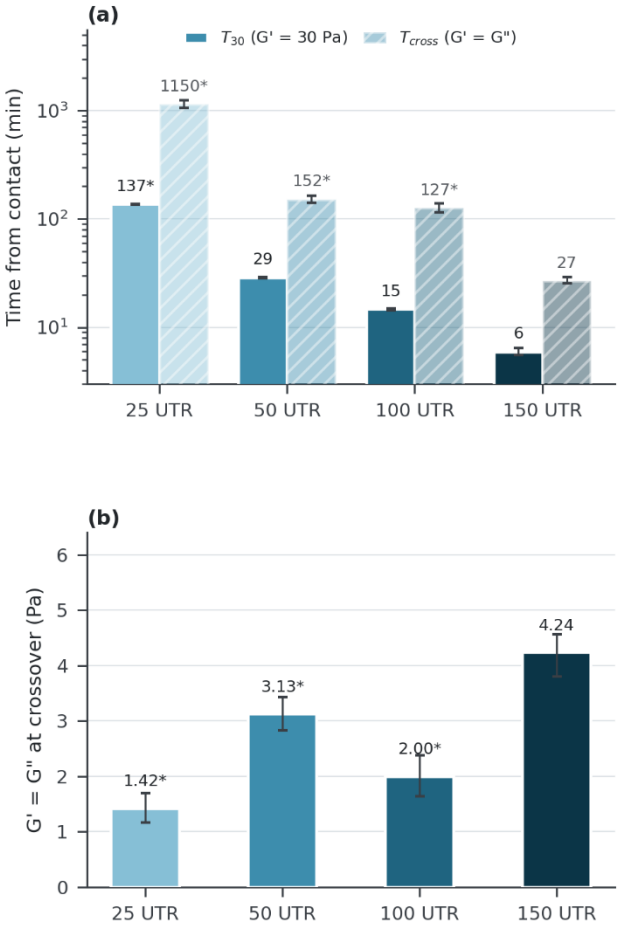


Fig. 3. (a) Time to the dissolution threshold (T30) and modulus crossover time (Tcross), both counted from enzyme–substrate contact, on a logarithmic scale. (b) Modulus at the crossover (G′ = G″). Error bars: 95 % CI by bootstrap. The asterisk (*) marks values extrapolated beyond the observed experimental window.

IV. Discussion

The rheological profiles show that the enzymatic degradation of Rennova Ultra Volume Lido is strongly dependent on enzyme concentration and directly associated with the progressive disruption of the hydrogel network. Fitting the same kinetic model to G′ and G″ on a common time axis allows the description not only of the degradation rate but also of the change in the viscoelastic regime of the material.

A. Meaning of the Shape Parameter

The value of β consistently below 1, for both moduli and at all concentrations, indicates progressive damping of the apparent degradation rate. This behavior is expected when the enzyme acts on a substrate whose structure changes continuously during the reaction itself. Hyaluronidase cleaves the HA chains and, in doing so, progressively alters the molar mass, the network architecture and the degree of entanglement of the system. In the early stages the substrate consists of long, strongly entangled chains whose cleavage removes elastically active junctions and produces marked drops in G′ per cleavage event. As the reaction advances, the remaining fragments are smaller and less entangled, and each new cleavage produces a proportionally smaller impact on the modulus. Added to this is the intrinsic heterogeneity of

enzymatic action, which acts preferentially on the most accessible regions of the network and leaves the most densely cross-linked domains for last. The combined result is kinetics that are not simple first-order but damped, which the parameter β describes compactly.

The consistency of β across the four concentrations indicates that this damping is a characteristic of the Rennova Ultra Volume Lido–Biometil system rather than an artifact of an isolated dose: enzyme concentration changes the rate of degradation, expressed in k, but not the shape of the curve.

B. Dead Time and Degradation Prior to the First Reading

The main limitation of the experimental design — the impossibility of measuring the moduli at the instant of enzyme–substrate contact — was addressed by modeling. The parameter τ provides a quantitative estimate of that interval, and its magnitude (1.1 to 2.7 min) is compatible with the manual procedure described.

The most relevant consequence is that the fraction of degradation occurring before the first record increases with dose, from 9.7 % at 25 UTR to 52.8 % at 150 UTR. In the latter condition, the estimated time for loss of half the initial modulus (1.79 min) is shorter than τ itself (2.06 min), that is, more than half of the degradation had already occurred before the first reading. This explains the variability of the initial moduli recorded across groups, which does not arise from material heterogeneity but from the progress of the reaction before recording began, and justifies referring all estimates presented here to enzyme–substrate contact rather than to the first reading of the instrument.

One limitation of the model deserves note. If τ represented only handling time, it would be expected to be approximately constant across groups, which is not observed. An alternative fit with τ shared across all groups yields τ = 1.68 min and describes the data almost as well as the model with free τ. This indicates that τ absorbs, besides the actual handling time, part of the incomplete specification of the model — notably the assumption that degradation follows the same functional form from time zero, including the mixing period, during which enzyme–substrate contact is not yet homogeneous. The values of τ should therefore be interpreted as order-of-magnitude estimates of the dead time rather than as its direct measurement.

C. Dissolution Threshold and Crossover of the Moduli

The threshold of G′ = 30 Pa was adopted from recent in vitro degradation studies [6], where it is defined as the lowest elastic modulus that can be reliably measured with this type of geometry, and therefore as the point at which the assay is stopped rather than as a structural transition of the gel. In the present system it corresponds to approximately 25 % of the modulus of the intact hydrogel, and to roughly one order of magnitude below the G′ reported for the volumizing fillers discussed in Section IV-D [14]. It is important to note, however, that reaching this threshold is not equivalent to the loss of elastic predominance: in all groups G′ remains higher than G″ at the instant T30, and the material continues to behave predominantly as a viscoelastic solid, albeit a much weakened one.

The effective loss of elastic predominance occurs only at the G′ = G″ crossover — the point at which tan δ = 1 —,and the contrast between the two landmarks is substantial. In the 150 UTR group the 30 Pa threshold is reached at 6.0 min, whereas the crossover occurs at 26.7 min — more than four times later. In the 50 and 100 UTR groups the crossover projects to 152 and 127 min, well beyond the observed window, and at 25 UTR it is not attained within any practically relevant horizon. These two landmarks therefore describe distinct structural events: the first corresponds to weakening of the network to the point of losing mechanical support capacity; the second, to the fall of connectivity below the percolation threshold, beyond which viscous dissipation dominates the behavior of the material [4], [14], [15].

The fact that k″ < k′ and β″ < β′ in all groups supports this reading. Enzymatic cleavage preferentially reaches the elastically active bonds that hold the network together [4], [5], [13], so that G′ decays faster than G″; the resulting fragments, although smaller and less entangled, retain the capacity to dissipate energy through viscous flow, which sustains G″ for longer. The convergence of the two moduli is the quantitative expression of this asymmetry.

It should be noted that the moduli at the crossover lie between 1.4 and 4.2 Pa, a range in which the torque approaches the minimum threshold of the instrument for reliable reading. The values of Gcross should therefore be interpreted as an indication of the order of magnitude of the final state of the material, and not as measurements of precision comparable to those obtained at higher moduli.

D. Comparison With the Literature

The results reinforce that a high G′, in isolation, does not determine enzymatic resistance. Fillers with the same volumizing indication and comparable modulus ranges may behave differently towards hyaluronidase [13], [16], depending on the manufacturing technology that governs the architecture of the polymer network [6]. Restylane Lyft (NASHA™ technology; 25 mg/mL), a biphasic particulate gel with high G′ (977 Pa at 5 Hz) [14], dissolves with 2.5 USP units of recombinant hyaluronidase in a single-dose in vitro protocol [16], an exceptionally low dose explained by its particulate morphology, which facilitates enzyme diffusion and maximizes the contact area [4]. At the opposite end, Juvéderm Volux (25 mg/mL; 665 Pa at 5 Hz) and Voluma (15 mg/mL; 271 Pa at 5 Hz) [14], monophasic gels of Vycross™ technology, are reported as the most resistant commercially available, requiring up to 300 USP for the dissolution of 0.2 mL [6], [16].

Germani et al. (2025) [15] evaluated Rennova UltraDeep (Hybrid MoBi™) with 100 UTR of Biometil hyaluronidase, attributing to it a moderate-to-slow degradation profile; that analysis was conducted by digital photography rather than rheometry, which precludes direct comparison with the thresholds adopted here. Comparison with the other studies [2], [6], [8], [16] remains limited by the absence of a validated conversion factor between UTR and IU/USP and by the heterogeneity of volumes, geometries and endpoint criteria. The values reported here apply to 0.25 mL of product under direct homogenization at 20 °C and should not be converted to other conditions without experimental verification.

It should be emphasized that the data do not establish any dose as a minimum requirement for dissolution and do not support recommendations of clinical conduct. Under the conditions tested, lower doses reach comparable levels of degradation provided sufficient reaction time; the translation

of this observation to any clinical scenario would depend on multiple factors, such as physiological temperature, diffusion in tissue, enzymatic clearance, applied volume and filler technology.

E. Limitations

First, the study lacks experimental replication: each condition corresponds to a single sample, and the confidence intervals presented reflect only the uncertainty of the fit over one time series, not the variability between independent preparations.

Second, conducting the measurements at 20 °C — an approach adopted to avoid the evaporation artifacts observed at 37 °C in pilot assays — reduces the absolute reaction rate; the times reported here should be interpreted as relative kinetic references, ordered by dose, rather than as predictors of the in vivo reversion rate.

Third, the protocol employs direct homogenization, eliminating the diffusional barrier that in vivo is often the rate-limiting step, so that the assay measures the intrinsic susceptibility of the network rather than the kinetics of reversal in tissue.

Fourth, the estimates of Tcross for 25, 50 and 100 UTR and of T30 for 25 UTR lie beyond the observed window and depend on the validity of the model outside the fitted interval.

Finally, the results describe a single filler and a single enzyme formulation, there is no validated conversion between UTR and IU/USP, and the rheological changes are interpreted as disruption of the polymer network without independent structural confirmation.

F. Perspectives

The resulting experimental agenda includes, in order of priority, independent replication of the conditions with a sample size allowing inference between groups; repetition of the protocol at 37 °C with evaporation mitigation by humidity chamber or sealed geometry; explicit standardization of time zero, by timing the preparation or by in situ mixing, which would turn τ into a measured quantity; complementary physicochemical characterization of the degraded material, by infrared spectroscopy, size-exclusion chromatography or microscopy, correlating the rheological changes with the molar mass distribution of the fragments; extension of the design to other fillers and other hyaluronidases, including recombinant ones; and, finally, models that reintroduce the diffusional barrier, followed by ex vivo, animal and clinical studies.

V. Conclusion

This study indicates that Rennova Ultra Volume Lido exhibits dose-dependent rheological susceptibility to Biometil hyaluronidase in the range of 25 to 150 UTR, under controlled in vitro conditions at 20 °C. Fitting the same exponential model to G' and G'' —on a common time axis anchored at the enzyme-substrate contact point— allowed for the estimation of the time to reach 30 Pa (T30) as 136.9, 28.9, 14.7, and 6.0 min for 25, 50, 100, and 150 UTR (with non-overlapping confidence intervals between consecutive doses) and demonstrated that the reduction in T30 stems directly from the increase in the degradation constant k' with enzyme concentration.

The joint modeling of the two moduli further separated two landmarks that the analysis of G' alone confounds: the operational endpoint of the assay, at $G' = 30$ Pa, and the loss of elastic predominance, at the $G' = G''$ crossover — the latter a genuine change in the viscoelastic regime of the material, which occurs substantially later and was observed within the experimental window only in the 150 UTR condition, at 26.7 min. The model also quantified the degradation occurring before the first reading, between 9.7 % and 52.8 % depending on dose, establishing that any comparison between concentrations in this type of assay must be referred to enzyme–substrate contact.

The range of 50 to 100 UTR identified as capable of driving 0.25 mL of the product to the rheological threshold in less than 30 min constitutes the first product-specific experimental rheological reference for the Rennova Ultra Volume Lido–Biometil pair within the Brazilian UTR system. It is, however, a preliminary reference, derived from a single sample per condition, obtained at 20 °C and without a diffusional barrier, and not a clinical dose recommendation.

Two methodological lessons emerge from this study. The first is that the interval between mixing and the first reading is not an operational detail: at the highest dose it contains more than half of the degradation, so a protocol that does not control it measures only the late phase of the reaction. The second is that following G' alone conflates the point at which the assay ceases to be measurable with the point at which the material changes viscoelastic regime; modelling both moduli on a common time axis separates the two. Three questions remain open and bound the scope of the reference proposed here: how the UTR system relates quantitatively to IU/USP, how the degradation kinetics scale with physiological temperature, and how much of the susceptibility measured here survives the diffusional barrier that direct homogenization removes.

References

[1] A. Chakhachiro and M. Waseem, “Risk factor analysis for vascular occlusions after dermal filler injections: a systematic review and meta-analysis,” Cureus, vol. 17, no. 4, 2025.

[2] L. K. de Almeida Balassiano and B. S. F. Bravo, “Hialuronidase: uma necessidade de todo dermatologista que aplica ácido hialurônico injetável,” Surg. Cosmet. Dermatol., vol. 6, no. 4, pp. 338–343, 2014.

[3] G. Murray et al., “Guideline for the safe use of hyaluronidase in aesthetic medicine, including modified high-dose protocol,” J. Clin. Aesthet. Dermatol., vol. 14, no. 8, p. E69, 2021.

[4] S. P. Fundarò et al., “The rheology and physicochemical characteristics of hyaluronic acid fillers: their clinical implications,” Int. J. Mol. Sci., vol. 23, no. 18, p. 10518, 2022.

[5] G. Kroumpouzos and P. Treacy, “Hyaluronidase for dermal filler complications: review of applications and dosage recommendations,” JMIR Dermatol., vol. 7, no. 1, p. e50403, 2024.

[6] J. Faivre et al., “Comparison of hyaluronidase-mediated degradation kinetics of commercially available hyaluronic acid fillers in vitro,” Aesthet. Surg. J., vol. 44, no. 6, pp. NP402–NP410, 2024.

[7] S. Nagaraju et al., “Estimation of serum hyaluronidase activity overcoming the turbidity interference,” Am. Soc. Clin. Lab. Sci., vol. 24, no. 3, pp. 172–177, 2011.

[8] J. Oliveira et al., “Effectiveness of hyaluronidase on hyaluronic acid degradation: an experimental study,” Aesthet. Surg. J. Open Forum, p. ojaf108, 2025.

[9] Innovapharma, Product Information Leaflet: Rennova Ultra Volume Lido. Goiânia, Brazil: Innovapharma, 2024.

[10] T. Safran et al., “Comparison of the degradability of hyaluronic acid by ovine and recombinant human hyaluronidase,” Aesthet. Surg. J., p. sjaf127, 2025.

[11] D. J. Burgess et al., “Rheological considerations of pharmaceutical formulations: focus on viscoelasticity,” PMC/NCBI, 2023.

[12] B. A. Buhren et al., “Standardized in vitro analysis of the degradability of hyaluronic acid fillers by hyaluronidase,” Eur. J. Med. Res., vol. 23, no. 1, p. 37, 2018.

[13] G. Grimaldi et al., “Rheology as a tool to investigate the degradability of hyaluronic acid dermal fillers,” Clin. Cosmet. Investig. Dermatol., pp. 1349–1358, 2025.

[14] G. W. Hong et al., “Rheological characteristics of hyaluronic acid fillers as viscoelastic substances,” Polymers, vol. 16, no. 16, p. 2386, 2024.

[15] M. Germani et al., “Comparative in-vitro degradation of hyaluronic acids exposed to different hyaluronidase enzymes,” J. Oral Biol. Craniofac. Res., vol. 15, no. 1, pp. 178–182, 2025.

[16] R. Sudharshan et al., “Identifying minimum single dose of recombinant human hyaluronidase for in vitro dissolution of twenty-two hyaluronic acid fillers,” Ophthalmic Plast. Reconstr. Surg., 2025, doi: 10.1097/IOP.0000000000002915.

[17] J. M. Zuidema, C. J. Rivet, R. J. Gilbert, and F. A. Morrison, “A protocol for rheological characterization of hydrogels for tissue engineering strategies,” J. Biomed. Mater. Res. B Appl. Biomater., vol. 102, no. 5, pp. 1063–1073, 2014.

Rheological Properties of Chitosan Hydrogels Containing Zinc-doped Whitlockite

1st Carolina Rosatto Ribeiro Scilingo
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
São Paulo, Brazil
carol.scilingo2004@gmail.com

2nd Larissa Ribeiro Loureço
Center for Natural and Human Sciences
Universidade Federal do ABC
Santo André, Brazil
r.larissa@ufabc.edu.br

3rd Juliana Marchi
Center for Natural and Human Sciences
Universidade Federal do ABC
Santo André, Brazil
juliana.marchi@ufabc.edu.br

4th Thais Samed Sousa
Center for Engine and Applied Social Sciences
Universidade Federal do ABC
Santo André, Brazil
thais.samed@ufabc.edu.br

5th Roger Borges
Biomedical Engineering / Bioengineering
Faculdade Israelita de Ciências da Saúde Albert Einstein / Brazil University
São Paulo, Brazil
roger.b@einstein.br

Abstract — Biomaterial-guided bone regeneration has driven the development of biocompatible inks for 3D printing that mimic the mineral composition of native bone tissue, becoming an interesting solution for personalized medical devices for bone regeneration. In this work, chitosan hydrogels were prepared and reinforced with whitlockite (WK)doped with zinc at concentrations of 0.5% and 1% (mol %). The formation of the Zn-doped WK phase was confirmed by X-ray diffraction (XRD). Four formulations were evaluated: Chi (chitosan, control), Chi-WK (chitosan + WK), Chi—WK—0.5Zn (chitosan + WK doped with 0.5 % Zn), and Chi-WK-1Zn (chitosan + WK doped with 1% Zn). The rheological behavior was characterized by a shear rate sweep, an amplitude sweep to determine the linear viscoelastic region (LVR), a frequency sweep, and a recovery test, all performed with a 0.1 mm gap. All formulations exhibited shear-thinning and viscoelastic (G' > G") behaviors, with Chi-WK showing the highest storage modulus in the LVR (~1,775 Pa) and the highest flow index in the Shear Rate Sweep test (n = 0.382), while Chi-0,5Zn and Chi-1Zn showed the lowest flow indices (n = 0.118 and n = 0.113, respectively), evidencing more pronounced shear-thinning behavior. The recovery test demonstrated structural recovery capacity in all formulations, supporting their potential suitability for extrusion-based 3D printing, to be confirmed by direct printing trials. Zinc doping modulated the viscoelastic properties of the hydrogels, making them more suitable for 3D printing and contributing to the development of bioactive inks for bone scaffolds.

Keywords — Chitosan; Whitlockite; Zinc doping; Hydrogels; 3D printing; Rheology; Bone scaffolds.

I. INTRODUCTION

Bone tissue engineering has sought biocompatible and printable inks capable of mimicking both the organic and mineral phases of native bone. Extrusion-based 3D printing has emerged as a promising technology for fabricating scaffolds with complex geometries and controlled distribution of cells and bioactive agents, enabling the construction of customized three-dimensional constructs for regenerative medicine applications [1-2]. The success of these platforms depends directly on the rheological properties of the ink, which must exhibit shear-thinning behavior during extrusion, a dominant elastic character for structural support after deposition, and rapid recovery capacity of the gel network [2].

Among the biopolymers used as a base for 3D printing inks, chitosan stands out for its biocompatibility, biodegradability, antimicrobial activity, and ability to form hydrogels with tunable rheological properties for extrusion-based 3D printing [1-2]. Chitosan has reactive amino ($-NH_2$) and hydroxyl (-OH) groups, capable of interacting with inorganic particles and metal ions through coordination, thereby modulating the viscoelastic properties of the hydrogel and favoring the formation of three-dimensional networks with thixotropic behavior [1].

To confer bioactivity and stimulate bone regeneration, polymeric hydrogels have been reinforced with calcium phosphate-based bioceramics [3]. Among these ceramics, whitlockite (WK, $Ca_{18}Mg_2(HPO_4)_2(PO_4)_{12}$) - which is the second most abundant inorganic phase in human bone tissue (up to 20% by mass), being particularly relevant in young bone and in remodeling processes – has recently arisen interest in the field of biomaterials. Unlike hydroxyapatite or β-tricalcium phosphate, WK exhibits an intermediate solubility between them and, besides releasing PO_4^{3-} Ca2+ ions, also leaches Mg^{2+} ions under physiological dissolution, which actively participate in osteogenic signaling pathways. WK nanoparticles have been shown to stimulate osteogenic differentiation of stem cells, inhibit osteoclastic activity, and promote in vivo transformation into hydroxyapatite, resulting in faster bone regeneration with greater structural density [4], [5].

The incorporation of zinc into calcium phosphates is a strategy widely investigated to enhance bone repair. Zinc is an essential trace element for bone metabolism that stimulates osteoblast proliferation and differentiation, inhibits osteoclastic resorption, and exerts local antimicrobial action [6]. In bioceramics, the atomic substitution of Ca^{2+} by Zn^{2+} in the crystal lattice modulates the ionic release profile without compromising the structure of the mineral phase. Additionally, piezoelectric Zn-doped WK has been linked to the modulation of cellular processes through mechano-electrical stimulation, and, in addition to increasing mechano-electrical conversion, it has attracted interest in doped inorganic phases for multifunctional biomaterials [7].

Rheological characterization is a central tool for assessing the suitability of hydrogel as an ink for 3D printing. The determining parameters include: apparent viscosity and the shear-thinning index, the viscoelastic moduli G' and G", the linear viscoelastic region (LVR) that defines the stiffness and elastic behavior of the gel at rest, and structural recovery capacity that determines the shape fidelity of the printed filament. The balance between fluidity during extrusion and

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

firmness after deposition represents the main challenge in developing hydrogel inks [1-2].

Given this context, the present work proposes the development and rheological characterization of chitosan hydrogel inks reinforced with zinc-doped WK (0.5% and 1%), synthesized by the chemical precipitation method with hydrothermal aging. The formation of the WK phase was confirmed by XRD prior to incorporation into the hydrogel, and the rheological behavior of the four formulations was systematically evaluated as a basis for the future application of these compositions as bioactive inks for 3D printing of bone scaffolds.

II. MATERIALS AND METHODS

A. Synthesis of the chitosan hydrogel

To prepare 15 mL of hydrogel, an acid solution was obtained by mixing 7.5 mL of distilled water, 6 mL of acetic acid (Sigma Aldrich), 1.5 mL of lactic acid (Sigma Aldrich), and 0.45 g of citric acid (Casa Americana, Brasil), homogenized under magnetic stirring for 20 min. **For the WK-reinforced formulations (Chi-WK, Chi-WK-0.5Zn, and Chi-WK-1Zn), WK powder (pure or zinc-doped) was then added to the acid solution at 10% (w/w) relative to the chitosan mass to be subsequently incorporated, corresponding to 0.09 g of WK per 0.9 g of chitosan (final WK concentration of 0.6% w/v in the 15 mL hydrogel batch).** Next, 0.9 g of low-molecular-weight chitosan (final concentration of 6% w/w) was added to the solution, ensuring complete submersion of the polymer. The mixture was left to rest for at least 12 h and subsequently homogenized by ultrasound for 2 h (with water exchange every 1 h). To remove bubbles, the gel was transferred to syringes with sealed tips and centrifuged at 3600 rpm for 30 min.

B. Synthesis of zinc-doped whitlockite

The synthesis of undoped whitlockite was carried out by chemical precipitation, followed by hydrothermal aging, according to the protocol. Undoped whitlockite was used as a control parameter in our studies. The process was divided into two days. On the first day, two separate solutions were prepared. The alkaline solution was obtained by dissolving 14.2660 g of $Ca(OH)_2$ (99,5 %, Sigma Aldrich) and 3.3640 g $Mg(OH)_2$ (99,5 %, Sigma Aldrich) in 250 mL of distilled water, under magnetic stirring at 70 °C for 1 h. The acidic solution was prepared by adding 16.275 mL of H_3PO_4 (Sigma-Aldrich) to 250 mL of distilled water in a volumetric flask. The acidic solution was dripped onto the alkaline solution at a rate of 2.5 mL/min using a peristaltic pump, maintaining the temperature at 70 °C. The resulting suspension was transferred to a reactor (autoclave) and subjected to hydrothermal aging at 70 °C for 48 h, with an expected yield of ~90%. The general reaction equation is in Eq. 1:

$$9Ca(OH)_2 + Mg(OH)_2 + 7H_3PO_4 \rightarrow Ca_9Mg(HPO_4)(PO_4)_6 + 18OH^- + 2H_2O \quad \text{(Eq. 1)}$$

On the second day, the solution was removed from the reactor and cooled for 1 h (with the aid of cold running water, if necessary). The precipitate was transferred to 50 mL Falcon tubes and centrifuged at 3,600 rpm for 5 min. The washing cycle was repeated 5 times: (1) removal of the supernatant; (2) filling the tube with distilled water up to 50 mL; (3) mechanical agitation for resuspension; and (4) centrifugation at 3,600 rpm for 5 min. After the last wash, the supernatant was discarded, and the tubes were labeled and stored in a freezer. The samples were subsequently freeze-dried.

Zinc-doped whitlockite: For the synthesis of the doped samples, a molar fraction of $Ca(OH)_2$ was replaced by $Zn(NO_3)_2$ (99,5 %, Sigma-Aldrich), keeping the total stoichiometry unchanged. For 0.5 mol% Zn: $Ca(OH)_2$ = 0.66731 g, $Zn(NO_3)_2$ = 0.01785 g. For 1 mol% Zn: $Ca(OH)_2$ = 0.66396 g and $Zn(NO_3)_2$ = 0.03570 g. The precipitation, hydrothermal aging, and washing protocols were identical to those used for undoped WK. The formation of the WK phase in all samples was confirmed by XRD after freeze-drying.

C. Structural confirmation by X-ray diffraction (XRD)

The formation of the whitlockite phase, both pure and zinc-doped, was confirmed by X-ray diffraction (XRD), validating the synthesis route employed prior to incorporation into the chitosan hydrogel. The diffractograms were compared with the reference standard for whitlockite (ICDD 00-011-0570), and the absence of significant secondary phases attested to the purity of the obtained phase. The observed peaks are consistent with the crystal structure of magnesium whitlockite previously characterized by X-ray diffraction and *ab initio* methods [5].

D. Experimental groups

For comparative evaluation, four hydrogel formulations were prepared, as described in Table I.

Group	Composition
Chi	Chitosan (control)
Chi-WK	Chitosan + pure WK
Chi-WK-0,5Zn	Chitosan + WK doped with 0.5 mol.% Zn
Chi-Wk-1Zn	Chitosan + WK doped with 1 mol.% Zn

TABLE I. Evaluated formulations.

E. Rheological characterization

The rheological behavior of the four formulations was evaluated through four tests, all performed with a 0.1 mm gap. A shear rate sweep test was performed to evaluate the apparent viscosity of the samples under conditions close to body/printing temperatures and to assess their shear-thinning behavior, which is relevant to material extrusion. The linear viscoelastic region (LVR) was determined to identify the strain range over which the viscoelastic response of the samples remains linear, a condition necessary for the correct interpretation of subsequent oscillatory tests. Then, we performed the frequency sweep test within the determined LVR to characterize the evolution of the storage (G') and loss (G") moduli as a function of frequency, thereby indicating whether each formulation is predominantly elastic or viscous. Finally, we carried out a recovery test (structural recovery) to evaluate the gel network's recovery capacity after a high-shear step, a property associated with shape fidelity during and after 3D printing.

The shear stress (τ) and apparent viscosity (η) data as a function of shear rate ($\dot{\gamma}$), obtained from the shear rate sweep

test at 37 °C, were fitted to the power-law model of *Ostwald–de Waele* [8] (Eq.2):

$$\mathbf{T = K \cdot \dot{\gamma}^{n}} \qquad \text{(Eq. 2)}$$

where K is the consistency index (Pa·s^n) and n is the flow index, which expresses the degree of shear-thinning (*shear-thinning* behavior) of the material (n < 1). The fitting was performed using log-log linear regression (log τ = log K + n·log $\dot{\gamma}$) on the raw data for each formulation. For Chi and Chi-WK groups, the fitting was performed over the entire available $\dot{\gamma}$ range after excluding the initial points corresponding to the rheometer's stabilization period. For Chi-WK-0.5Zn and Chi-WK-1Zn, stabilization and a slight reduction in τ were observed at high shear rates ($\dot{\gamma}$ > 25 s^{-1}), behavior consistent with slippage at the sample-geometry interface (*wall slip*), a phenomenon widely described for materials with a high fraction of dispersed particles [8]. In these cases, the fitting was restricted to the $\dot{\gamma}$ window where τ increased monotonically with shear rate (4–24 s^{-1}), thereby avoiding the inclusion of data affected by this experimental artifact.

III. RESULTS AND DISCUSSION

A. Confirmation of the WK phase by XRD

Initial characterization of the ceramic powders was necessary to ensure that WK was produced during synthesis. The undoped WK was already characterized, as published in [5]. The Zn-doped WK was characterized before and after calcination to evaluate whether the WK phase was formed prior to or after thermal treatment. For Zn-doped WK, formation of the whitlockite phase was confirmed by XRD for both doping concentrations, 0.5 and 1 mol% (Fig. 1).

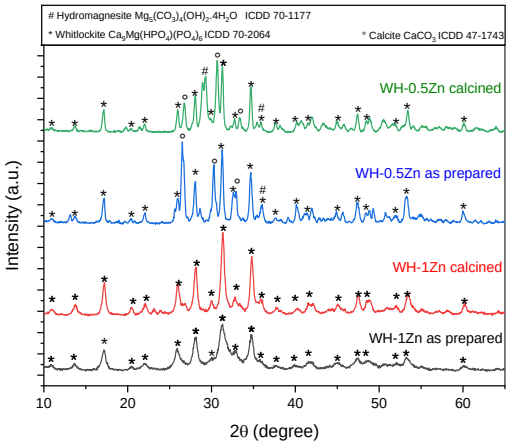


Fig. 1. X-ray diffractograms of pure and zinc-doped WK (0.5% and 1%). Peaks indexed according to ICDD 00-011-0570.

The diffractograms showed characteristic WK peaks, indexed to ICDD card 00-070-2064, thereby validating the synthesis route employed. No significant secondary phases were observed for the WK doped with 1 mol% of Zn, indicating satisfactory purity of the obtained phase. In these compositions, zinc doping did not alter the crystal structure of WK, confirming that Zn^{2+} was incorporated into the crystal lattice, partially replacing Ca^{2+}, as described in the literature for zinc-doped calcium phosphates [6]. However, for the 0,5 mol% Zn doping, additional formation of hydromagnesite and calcite was observed. Usually, carbonate phases form on the surfaces of calcium phosphate phases due to interactions between Ca^{2+} at the surfaces and atmospheric CO_2 under humid conditions, as described by Cerruti et al. [9]. Despite the undesired phases observed, we did not exclude the Chi-WK-0.5Zn formulation from the rheological characterization, as we aim to investigate its impact on rheological behavior. Notwithstanding, calcite is also a bioactive biomaterial, and both *in vitro* and *in vivo* studies have shown promising biological applications in bone regeneration [10].

B. Shear Rate Sweep

All formulations exhibited shear-thinning behavior (Fig. 2), confirmed by the flow indices n < 1 obtained from the power-law model fitting (Table II). The group Chi presented n = 0.167, indicating pronounced shear-thinning, whereas group Chi-WK exhibited the highest flow index (n = 0.382), suggesting lower shear sensitivity than the other formulations. The zinc-doped formulations, Chi-0.5WK (n = 0.118) and Chi-1WK (n = 0.113), showed the lowest flow indices among all samples, evidencing more pronounced shear-thinning behavior. These values fall within the range reported in the literature for hydrogels intended for direct ink writing (DIW) extrusion (n ≈ 0.3–0.6) [8], with G3 and G4 falling below this range, indicating greater ease of extrusion at lower applied pressures.

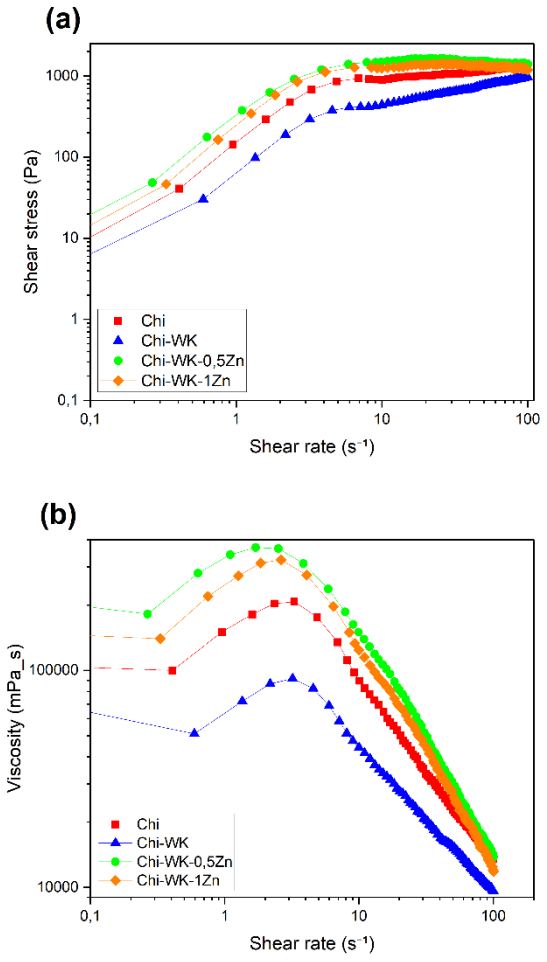


Fig. 2. Flow behavior at 37 °C for the four formulations: (a) shear stress and (b) apparent viscosity as a function of shear rate.

The consistency index (K) allows comparisons of resistance to flow of different ink formulations and is strongly

influenced by intermolecular interactions among the ink components. Adding undoped WK to chitosan decreased the consistency index (609.6 for Chi and 171.0 for Chi-WK). However, when Zn-doped WK is added, the consistency index increases from 1146,3 for Chi-WK-0.5Zn to 976.1 for Chi-WK-1Zn. It should be noted that the Chi-WK-0.5Zn sample also contained minor calcite and hydromagnesite phases, as identified by XRD (Section III.A); these secondary phases may additionally contribute to the observed rheological response, making it difficult to attribute the results in this specific formulation solely to Zn^{2+}–chitosan interactions. Bearing this limitation in mind, these data suggest that zinc ions leached from the doped ceramics may affect the intermolecular interactions of chitosan. It is known that some divalent ions, such as Zn^{2+}, Mn^{2+}, and Mg^{2+}, interact with electron pairs from NH_2 -COOH groups of chitosan, forming complexes involving intra- or intermolecular interaction among chitosan chains [11]. However, Ca^{2+} ions do not interact with amino and carboxyl groups as effectively as the other divalent ions [12-13]. Thereby, it is reasonable to assume that it may be the main cause of the found results. It should be noted that Zn^{2+} ion release and Zn–chitosan complex formation were not directly measured in this study (e.g., via ICP-OES for ion release or FTIR for coordination bonding); the proposed mechanism is therefore a hypothesis consistent with the observed rheological trends and previous literature [11]–[13], rather than a directly confirmed interaction.

We have to assume two different effects: 1) the first is the effect of the leached ions from the ceramics in the chitosan structure, where Zn-doped WK leaches Zn^{2+} ions that yields to more formation of chitosan complexes than undoped WK; 2) the second effect is the physical effect of the ceramics particles in the polymer, where these particles limit the polymer-polymer interactions, and thereby decrease the viscosity of the polymer. These effects compete with one another: one increases intermolecular interactions, while the other decreases them. While in compositions containing Zn-doped WK, the effect of Zn2+ ions was dominant, in compositions containing undoped WK, the physical effect of the particles became predominant.

These results are also evidenced by the viscosity measurement (Fig. 2). The viscosity follows the same trend as the consistency index. Considering that viscosity is also linked to intermolecular interactions among the ink formulation components, it is reasonable that both viscosity and K follow the same trend.

TABLE II. Power-law model parameters ($\tau = K \cdot \dot{\gamma}^n$) and viscosity at 10 s^{-1}

Group	n	K (Pa·s^n)	R^2	η at 10s^{-1}
Chi	0.167	609.6	0.978	90.1
Chi-WK	0.382	171.0	0.983	44.2
Chi-WK-0,5Zn	0.118	1146.3	0.929	149.4
Chi-Wk-1Zn	0.113	976.1	0.885	123.7

C. Linear Viscoelastic Region — LVR

The strain sweep test allowed identification of the LVR of each formulation and confirmed the predominantly elastic character of the hydrogels (G' > G") at small strain amplitudes. Within the LVR, the storage module remained constant, while above the linear viscoelastic limit, there was a progressive decrease in G', indicating structural collapse of the gel network.

Chi group presented G' in the LVR on the order of ~700 Pa, a value characteristic of moderate-concentration chitosan hydrogels [14]. Chi-WK group stood out with G' ~1,775 Pa within the linear region, the highest among the formulations. This result indicates that incorporating undoped WK promoted mechanical reinforcement of the hydrogel network, likely through physical interactions (hydrogen bonds, electrostatic interactions) between WK particles and chitosan chains. In a previous work from our research group in which bioactive glass particles (SiO_2-CaO-P_2O_5) were incorporated in chitosan hydrogels, we found that an increase in G' was modulated by the nergatively charged surface of the glass particles, which resulted in a possible hydrogen bond between silanol groups from the glass and amine or carboxylic acids from the chitosan [15]. In this work, we believe that hydrogen bonds between the phosphate groups of WK may cause a similar effect.

Interestingly, the Chi-WK-0.5Zn and the Chi-WK-1Zn groups showed G' ~1,307 and 1,109 Pa. These values are higher than those observed in the Chi group but lower than those in the Chi-WK group. Considering that the storage modulus reflects the ability of the hydrogel to store elastic deformation energy that, in turn, is stored in the deformation of chemical bonds, intermolecular interactions, and reversible structural deformation, we believe that the physical particle effect of the ceramic particles increases the storage modulus, while the formation of complexes with Zn^{2+} ions makes the chitosan structure more rigid, decreasing the amount of energy that can be stored in the hydrogel deformation. As noted above, the presence of secondary calcite and hydromagnesite phases in the Chi-WK-0.5Zn sample may also contribute to this intermediate behavior, and should be considered alongside the proposed Zn^{2+}-complexation mechanism.

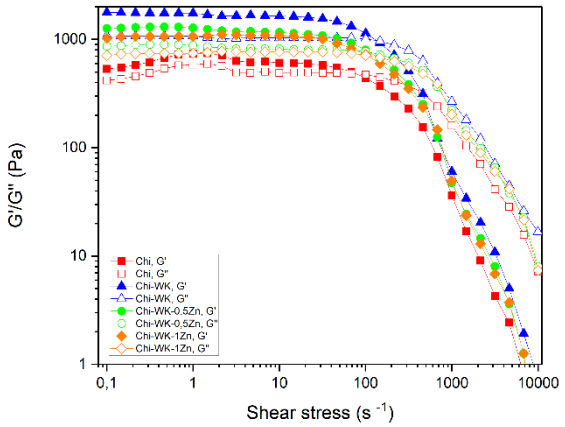


Fig. 3. G' and G" moduli as a function of strain amplitude at 5 °C (LVR).

D. Frequency Sweep (5 °C)

The frequency sweep, performed within the LVR of each formulation, confirmed the predominantly elastic character (G' > G") across the entire evaluated frequency range,

indicating that the formulations exhibit a coherent gel structure. Maintaining G' > G" throughout the sweep is an essential requirement for 3D printing inks, as it indicates that the material sustains its shape after deposition [1], [2].

Chi group presented G' on the order of ~50 Pa at 0.01 Hz, increasing to ~1,100 Pa at high frequencies, confirming frequency-dependent gel behavior. The Chi-WK group exhibited the highest moduli across the entire evaluated range, with G' reaching ~2,500 Pa, thereby reinforcing the role of undoped WK particles in strengthening the polymer network. Chi-WK-0.5Zn and Chi-WK-1Zn groups presented similar behavior to each other, with intermediate values between Chi and Chi-WK, confirming the same mechanism mentioned in the LVR discussion. No formulation exhibited crossover (G' = G") within the evaluated range, reinforcing the solid-gel behavior in all compositions.

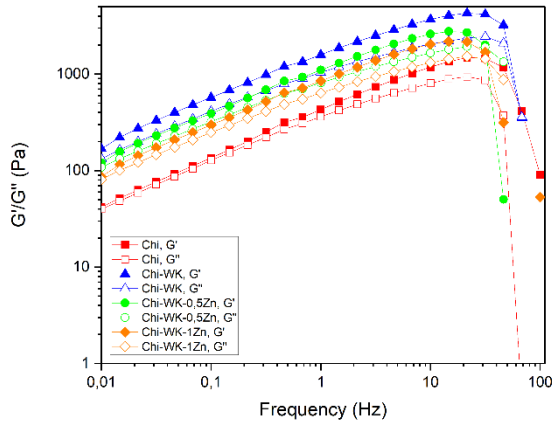


Fig. 4. G' and G" moduli as a function of angular frequency at 5 °C.

E. Structural Recovery

The recovery test assessed the ability of each formulation to reconstitute its gel network after a high-amplitude shear step. This thixotropic behavior is directly related to shape fidelity (*shape fidelity*) during and after 3D printing, as it determines the ability of the extruded filament to maintain its geometry [16].

All formulations demonstrated structural recovery above 96%. Chi group recovered 96.1%, while Chi-WK presented the highest absolute recovery, reaching 100.3% ($\eta_{initial}$ = 4,173.0 Pa·s), indicating complete network reconstitution after shearing. The Chi-WK-0.5Zn and Chi-WK-1Zn groups recovered 96.9% and 99.0%, respectively, both higher than the Chi group (our control), which recovered 96.0%. This behavior is consistent with the literature, which describes zinc doping as promoting reversible coordination interactions between Zn^{2+} ions and chitosan's functional groups, which dissociate under mechanical stress and reorganize once shearing ceases [6]. However, our results, along with the LVR and frequency sweep tests, provide evidence that the formation of zinc complexes makes the chitosan structure stiffer, and might be the cause of a lower recovery compared to the Chi-WK group. Nonetheless, all formulations show high recovery percentages, confirming thixotropic behavior and supporting their potential suitability for extrusion-based 3D printing, to be confirmed by direct printing trials.

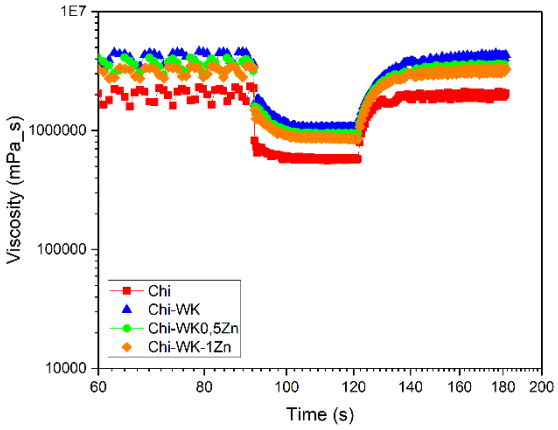


Fig. 5. G' recovery after high-amplitude shearing at 5 °C.

IV. CONCLUSION

This work demonstrated that chitosan hydrogels reinforced with undoped and zinc-doped whitlockite present promising rheological properties for application as 3D printing inks. The WK phase was confirmed by XRD in all synthesized samples, although the WK-0.5Zn composition also showed the formation of other minor phases, such as calcite and hydromagnesite. All formulations exhibited *shear-thinning* behavior, dominant elastic character (G' > G"), and structural recovery capacity after shearing. The incorporation of undoped WK promoted mechanical reinforcement of the hydrogel network via hydrogen bonding, thereby contributing to high recovery. Zinc doping modulated the viscoelastic properties, making the chitosan network stiffer and slightly decreasing the structure's strain. Overall, the results indicate that formulation Chi-WK presented the greatest mechanical reinforcement and structural recovery capacity, while the Chi-WK-0.5Zn and Chi-WK-1Zn formulations showed more pronounced shear-thinning behavior, favoring extrusion under lower pressures. Future studies should include 3D printing trials, cytotoxicity assessment, and in vitro bioactivity tests to confirm the clinical potential of the developed formulations.

ACKNOWLEDGMENTS

The authors thank the Faculdade Israelita de Ciências da Saúde Albert Einstein and the Federal University of ABC (UFABC) for the infrastructure provided for the development of this work. We also acknowledge the National Council of Science and Technology Development (CNPq, grant numbers 44319/2023-9 and 380534/2024-0) for the financial support.

REFERENCES

[1] M. Rajabi, M. McConnell, J. Cabral, and M. A. Ali, “Chitosan hydrogels in 3D printing for biomedical applications,” Carbohydr. Polym., vol. 260, art. 117768, 2021.

[2] M. Lazaridou, D. N. Bikiaris, and D. A. Lamprou, “3D bioprinted chitosan-based hydrogel scaffolds in tissue engineering and localised drug delivery,” Pharmaceutics, vol. 14, art. 1978, 2022.

[3] M. Vallet-Regí and E. Ruiz-Hernández, “Bioceramics: From bone regeneration to cancer nanomedicine,” Adv. Mater., vol. 23, no. 44, pp. 5177–5218, 2011.

[4] H. D. Kim et al., “Biomimetic whitlockite inorganic nanoparticles-mediated in situ remodeling and rapid bone regeneration,” Biomaterials, vol. 112, pp. 31–43, 2017.

[5] E. C. Macedo Lemos, R. H. Moriya, L. C. Teruel de Jesus, R. Borges, R. R. de Oliveira, N. B. de Lima, J. T. Arantes, and J. Marchi, "Magnesium whitlockite structure: an ab initio approach supported by experimental characterization," *Ceram. Int.*, vol. 51, no. 22, pp. 36637–36643, 2025.

[6] R. Cruz et al., “Does the incorporation of zinc into calcium phosphate improve bone re~′um ypair? A systematic review,”Ceram. Int., vol. 44, no. 2, pp. 1240–1249, 2018.

[7] K. Kapat, Q. T. H. Shubhra, M. Zhou, and S. Leeuwenburgh, “Piezoelectric nano-biomaterials for biomedicine and tissue regeneration,” Adv. Funct. Mater., vol. 30, art. 1909045, 2020.

[8] M. I. Calafel, M. Criado-Gonzalez, R. Aguirresarobe, M. Fernández, and C. Mijangos, "From rheological concepts to additive manufacturing assessment of hydrogel-based materials for advanced bioprinting applications," *Mater. Adv.*, vol. 6, pp. 4566–4597, 2025.

[9] M. Cerruti and C. Morterra, "Carbonate formation on bioactive glasses," *Langmuir*, vol. 20, no. 15, pp. 6382–6388, 2004.

[10] M. Mozafari, S. Banijamali, F. Baino, S. Kargozar, and R. G. Hill, "Calcium carbonate: Adored and ignored in bioactivity assessment," *Acta Biomater.*, vol. 91, pp. 35–47, 2019.

[11] S. Rashid, C. Shen, J. Yang, J. Liu, and J. Li, "Preparation and properties of chitosan–metal complex: Some factors influencing the adsorption capacity for dyes in aqueous solution," *J. Environ. Sci.*, vol. 66, pp. 301–309, 2018.

[12] J. Nie, Z. Wang, and Q. Hu, "Chitosan Hydrogel Structure Modulated by Metal Ions," *Sci. Rep.*, vol. 6, art. 36005, 2016.

[13] X. Wei, S. Chen, J. Rong, Z. Sui, S. Wang, Y. Lin, J. Xiao, and D. Huang, "Improving the Ca(II) adsorption of chitosan via physical and chemical modifications and characterizing the structures of the calcified complexes," *Polym. Test.*, vol. 98, art. 107192, 2021.

[14] P. Sánchez-Cid, M. Jiménez-Rosado, M. Alonso-González, A. Romero, and V. Perez-Puyana, "Applied Rheology as Tool for the Assessment of Chitosan Hydrogels for Regenerative Medicine," *Polymers*, vol. 13, no. 13, art. 2189, 2021.

[15] L. R. Lourenço, R. Borges, D. Carastan, M. B. Mathor, and J. Marchi, "Precision 3D printing of chitosan-bioactive glass inks: Rheological optimization for enhanced shape fidelity in tissue engineering scaffolds," *Bioprinting*, vol. 43, art. e00359, 2024.

[16] A. Schwab, R. Levato, M. D'Este, S. Piluso, D. Eglin, and J. Malda, "Printability and Shape Fidelity of Bioinks in 3D Bioprinting," *Chem. Rev.*, vol. 120, no. 19, pp. 11028–11055, 2020.

Scientometric Analysis of Nanoparticle-Based Drug Delivery Systems for Breast Cancer Treatment

1st Melissa Soares de Souza
Federal University of Pernambuco
Recife, Brasil
melissa.soaress@ufpe.br

2nd Neferson Barbosa da Silva Ramos
Federal University of Pernambuco
Recife, Brasil
neferson.barbosa@ufpe.br.

3rd Cristine Martins Gomes de Gusmão
Federal University of Pernambuco
Recife, Brasil
cristine.gusmao@ufpe.br.

4th Ana Cristina Martins
Federal University of Pernambuco
Recife, Brasil
ana.cmartins@ufpe.br.

5th Ricardo Yara
Federal University of Pernambuco
Recife, Brasil
ricardo.yara@ufpe.br

***Abstract*— Breast cancer remains a leading cause of mortality worldwide, prompting significant research into nanoparticle-based drug delivery systems to improve therapeutic efficacy and reduce toxicity. This study aims to perform a bibliometric analysis of the global scientific production on this topic between 2018 and 2024, identifying key trends, influential networks, and research gaps. A search was conducted on the Scopus database in October 2025 using descriptors related to nanoparticles, breast cancer, and drug delivery. A total of 1,280 eligible original articles were analyzed. VOSviewer software was used to map co-authorship, keyword co-occurrence, and citation networks to visualize the field's intellectual structure. The analysis revealed China as the primary research hub, followed by the USA, Iran, and India. Research themes formed three main clusters: physicochemical development of nanosystems, in vitro biological evaluations, and in vivo animal studies. While the field is expanding, collaboration networks remain geographically fragmented, with limited international integration between major institutional centers. Research on nanoparticle-based delivery for breast cancer is growing but characterized by regional clusters. Future progress depends on strengthening global collaborative networks and fostering translational studies to bridge the gap between laboratory findings and clinical application.**

***Index Terms*—Breast cancer; Nanoparticles; Drug delivery systems; Scientometrics analysis; VOSviewer.**

I. Introduction

Breast cancer is a leading cause of morbidity and mortality among women worldwide, showing a continuous increase in incidence over the last decades and representing a substantial portion of the global cancer burden. Recent estimates from Globocan (2022) indicate that breast cancer remains among the types with the highest number of new cases, accounting for approximately 2.3 million diagnoses (11.6% of all cancer cases), and contributes significantly to mortality rates, being responsible for about 670,000 annual deaths across different regions of the world [1].

The reality in developing countries mirrors this global landscape. In Brazil, for instance, this neoplasm consolidates itself as a major public health challenge. According to recent national projections for the 2023–2025 period, thousands of new annual diagnoses are expected, positioning the disease among the most recurrent in the female population, with incidence and mortality rates showing significant regional variations [4].

To overcome the barriers of low specificity and high toxicity common to classical oncological treatments, the application of nanotherapies has expanded considerably. This strategy aims to improve drug biodistribution and reduce adverse effects, becoming a field of intense scientific investigation [5]. Despite the growing volume of publications on new nanosystems, there is a lack of studies that systematize this scientific production to identify global collaboration patterns and emerging technological trends.

Given this scenario, this study aims to perform a bibliometric analysis of the global scientific production on nanoparticle-based drug delivery systems for breast cancer, mapping the evolution of the field between 2018 and 2024. Our main conclusions indicate that, although scientific production has grown exponentially, largely led by China and focused on preclinical studies, international collaboration networks remain fragmented. This suggests that the advancement towards clinical application will crucially depend on strengthening global partnerships and greater investment in translational research.

II. Materials and Methods

This study is a scientometric review designed to analyze current global scientific output regarding nanoparticle-based drug delivery systems applied to breast cancer treatment.

Data collection was conducted on the Scopus database on October 17, 2025, selected for its comprehensive indexing of high-impact journals across health sciences, pharmacology, and materials science [3]. The 2018–2024 timeframe was adopted to focus the analysis on contemporary scientific production and identify emerging trends in the development and application of nanoparticle-based drug delivery systems for breast cancer therapy.

The search strategy was executed using Scopus title, abstract, and keyword fields (TITLE-ABS-KEY) with the following search string: TITLE-ABS-KEY (Nanoparticles) AND TITLE-ABS-KEY (breast cancer) AND TITLE-ABS-KEY (Drug delivery systems) AND TITLE-ABS-KEY (Cancer treatment) AND PUBYEAR > 2017 AND PUBYEAR < 2025

The search was conducted on October 17, 2025, capturing publications from 2018 to 2024. Search terms were selected to establish a comprehensive retrieval strategy covering nanoparticles, breast cancer, drug delivery systems, and cancer therapy. Broad descriptors directly tied to core research concepts were prioritized, as preliminary testing showed that narrower term combinations yielded significantly lower record recovery.

Inclusion criteria comprised original research articles published between 2018 and 2024, written in English, that explicitly addressed nanoparticle-based drug delivery systems targeted for breast cancer treatment. Exclusion criteria eliminated studies evaluating nanoparticles for unrelated therapeutic or technological applications, as well as works that did not incorporate nanoparticle systems.

The initial database search yielded 2,902 publications. Applying Scopus automated filters for document type (original articles) and language (English) reduced the dataset to 1,901 records. The retrieved records were exported to a spreadsheet for screening. Following duplicate removal, a single reviewer manually screened titles and abstracts against the predefined eligibility criteria. Studies unrelated to nanoparticle drug delivery for breast cancer—including those focused on alternative medical applications or lacking nanoparticle integration—were excluded. Ultimately, 1,280 articles met all inclusion criteria and were selected for scientometric analysis.

VOSviewer software was utilized to generate network maps and analyze co-authorship networks (by country, author, and organization), keyword co-occurrence, citation networks (sources and documents), organization co-citation, and bibliographic coupling among documents. In the visual networks, nodes represent entities such as countries, institutions, or authors, where node size reflects occurrence frequency. Node and connection colors distinguish discrete clusters, while line thickness indicates link strength between elements.

III. RESULTS

The scientometric analysis of the 1,280 selected articles allowed for a clear overview of the evolution of research on nanoparticle-based drug delivery systems applied to breast cancer treatment. As observed in Fig. 1, the growth of scientific production between 2018 and 2024 exhibits a predominantly upward trend, reflecting the consolidation of this theme as one of the most promising frontiers in cancer nanomedicine.

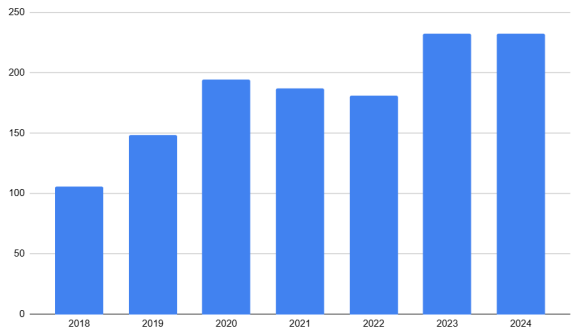


Fig. 1. Temporal distribution of publications on nanoparticle-based drug delivery systems applied to breast cancer treatment (2018–2024).

The year 2018 marks the beginning of the time series with 106 publications, followed by an increase to 148 articles in 2019. In 2020, further growth was observed, reaching 194 publications. In 2021 and 2022, there was a moderate decline, with 187 and 181 articles, respectively, while nevertheless maintaining a significant volume of scientific output. The upward trend resumed in 2023, with 232 identified articles—a figure that remained steady in 2024, also with 232 publications.

This temporal trend highlights sustained scientific interest in applying nanoparticles to drug delivery systems for breast cancer. Although publication growth was not strictly linear across all years, the most recent period concentrated the largest share of the dataset, with 464 articles published in 2023 and 2024, representing approximately 36.3% of the total analyzed corpus. This finding indicates the consolidation of the field in recent years and serves as a baseline to explore, in subsequent sections, the specific therapeutic approaches, materials, and nanostructured strategies that accompanied this temporal evolution.

A. Co-Authorship Analysis by Country

The analysis of co-authorship between countries demonstrates the geographical distribution of scientific production and international collaborations related to nanoparticle-based drug delivery systems for breast cancer treatment. The map generated by VOSviewer reveals an interconnected global network composed of multiple clusters representing the main research hubs and their respective collaborations, as shown in Figure 2.

The size of the nodes indicates the volume of publications from each country, while the thickness of the lines represents the intensity of cooperation. It is observed that China stands out as the primary center of scientific production in the area, presenting the largest node in the network and numerous connections with Asian, American, and European countries. This prominent position reflects the exponential growth of Chinese research in nanotechnology and oncology.

Secondarily, countries such as the United States, Iran, and India also appear as central actors in the international collaboration network, establishing consistent partnerships both among themselves and with other emerging nations. Iran, in particular, shows a high density of connections, indicating active cooperation with countries such as Turkey, China, Germany, and Italy, consolidating itself as an important regional hub in the application of nanoparticles in

anticancer therapies.

Brazil, although having lower centrality compared to Asian and North American countries, integrates a cluster involving European and Latin American nations (such as Portugal and Spain), demonstrating progressive insertion into international collaborations focused on translational research and the development of drug delivery systems.

Thus, the co-authorship map by country evidences that research on therapeutic nanosystems applied to breast cancer presents a global and interdisciplinary character, with strong Asian leadership—especially from China and Iran—and gradual growth in Latin American participation.

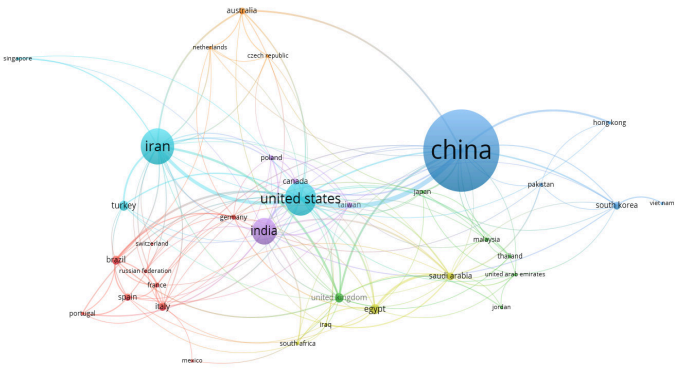


Fig. 2. Co-authorship map by country generated in VOSviewer, showing the geographical distribution and international connections of research on nanoparticle-based drug delivery systems applied to breast cancer. The size of the nodes represents the volume of publications, and the colors indicate collaboration clusters.

B. *Co-Authorship Analysis by Authors*

The author analysis highlights the collaborative structure among the main researchers working on nanoparticle-based drug delivery systems for breast cancer treatment. The map generated by VOSviewer shows the formation of different clusters (author groups) representing consolidated or emerging scientific cooperation networks, as illustrated in Fig. 3.

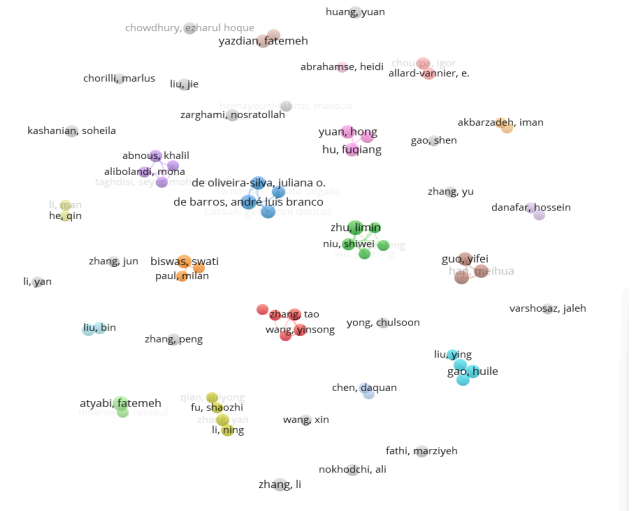


Fig. 3. Detailed view of specific author clusters, highlighting the distinct collaborative groups led by Asian and Latin American researchers.

The presence of multiple colored clusters is observed, indicating relatively independent collaboration nuclei with limited interconnectivity between them. Among the most prominent authors with high centrality in the network are Zhu, Limin; Niu, Shiwei; Guo, Yifei; Hu, Fuqiang; Yuan, Hong; De Oliveira-Silva, Juliana O.; and De Barros, André Luís Branco. These authors stand out due to the density of connections and node sizes, suggesting a higher number of co-authored publications and influence within the field.

The blue cluster, centered on De Oliveira-Silva and De Barros, represents an active network associated with Latin American research. The green cluster, led by Zhu, Limin, and Niu, Shiwei, evidences robust Asian collaboration. The red cluster, with Zhang, Tao, and Wang, Yinsong, and the pink cluster, with Hu, Fuqiang, and Yuan, Hong, also reflect collaborations concentrated in Asia, showing the significant contribution of this continent to the advancement of the area.

From this perspective, the co-authorship map demonstrates that scientific production on the topic is somewhat collaborative but still geographically fragmented, with strong production hubs and occasional intercontinental cooperation. This configuration reflects the recent and multidisciplinary growth of the area, driven by partnerships between research groups in nanotechnology, biomedical engineering, and pharmacology.

C. *Co-Authorship Analysis Among Organizations*

The co-authorship network among organizations highlights institutional articulation in the development of research on the subject. The map generated by VOSviewer shows a concentrated collaborative structure composed of a few strongly connected nodes, suggesting stable institutional relationships that are, however, still restricted to certain research nuclei, as shown in Figure 4.



Fig. 4. Collaboration map among organizations identified by VOSviewer. Nodes represent institutions and research groups, and the absence of robust connections indicates limited cooperation among the involved centers.

The formation of small, interconnected clusters is observed, where institutions such as the Drug Applied Research Center, the Department of Pharmaceutics, and the Department of Pharmaceutical Nanotechnology stand out. These centers appear as central nodes of the network, acting as axes of cooperation between different departments and universities.

The strong connection between these institutions indicates interdepartmental collaborations focused on the development of nanostructured systems applied to oncology, reinforcing the multidisciplinary nature of the area. On the other hand, institutions such as the Department of Biology, Islamic Azad University, and the Department of Life Science Engineering appear more peripherally, evidencing punctual or emerging collaborative participation within the global network.

The spacing between clusters demonstrates that, although

established partnerships exist, the institutional network still lacks broader international integration, with regional collaborations predominating. This pattern reinforces the scenario identified in the co-authorship analysis by authors, where there are strong local scientific production hubs located mainly in Asia and the Middle East, with lower involvement of European and Latin American institutions.

Overall, the results indicate that research on nanoparticle-based drug delivery presents consolidated institutional nuclei but remains geographically fragmented, suggesting opportunities for expanding global cooperation and scientific exchange between different regions.

D. Keyword Analysis

The keyword co-occurrence analysis allows for the identification of the main themes and research fronts structuring the field of nanoparticle-based drug delivery systems applied to breast cancer treatment. The map generated by VOSviewer (limited to the top 50 most frequent keywords) evidences a dense and interconnected network organized into three large thematic clusters representing the different conceptual and experimental approaches in the area, as illustrated in Figure 5.

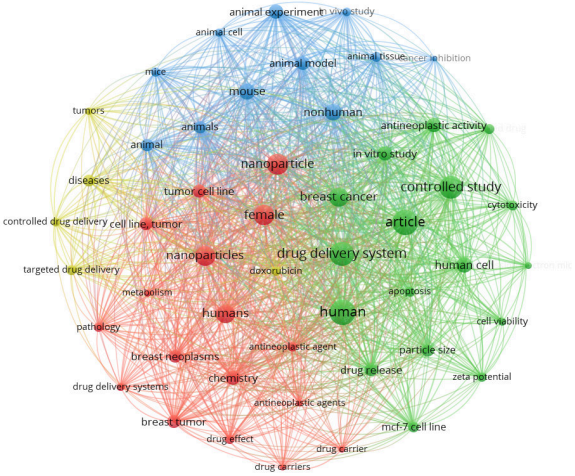


Fig. 5. Network visualization map of keyword co-occurrence generated by VOSviewer. Nodes represent index and author keywords, node size indicates occurrence frequency, and colors distinguish main thematic clusters.

The red cluster, centered on terms such as “nanoparticles”, “drug delivery systems”, “breast neoplasms”, “drug effect”, and “chemistry”, is associated with research on the development and physicochemical characterization of nanosystems. This group addresses topics related to formulation, stability, surface properties, as well as the evaluation of encapsulation efficiency and controlled drug release, such as doxorubicin. It represents a nucleus focused on materials engineering and pharmaceutical nanotechnology, aiming at optimizing the structural properties of nanoparticles for therapeutic application.

The green cluster aggregates terms such as “drug delivery system”, “controlled study”, “human cell”, “apoptosis”, “cytotoxicity”, and “in vitro study”, representing the biological and experimental branch of the area. This represents studies concentrating on preclinical tests seeking to evaluate the efficacy, selectivity, and safety of nanocarriers. This grouping highlights the interface between pharmacology, biotechnology, and cell biology, evidencing the application of nanoparticles as targeted therapeutic platforms.

The blue cluster is associated with terms such as “mouse”, “animal experiment”, “tumor model”, and “in vivo study”, indicating the set of preclinical research conducted in animal models. These studies are essential for translating in vitro results to in vivo contexts, allowing for the investigation of biodistribution, metabolism, and antitumor efficacy of nanostructured systems.

This co-occurrence network reveals a strong interrelationship between the physicochemical, biological, and applied aspects of the research. Furthermore, the high frequency of terms such as “breast cancer”, “drug release”, and “human” indicates that the area is in a consolidation phase, with increasing focus on biological assays and advanced therapeutic applications.

E. Source Citation Analysis

The source citation analysis highlights the main scientific journals that influence knowledge dissemination. As depicted in Figure 6, journals such as the International Journal of Nanomedicine, Pharmaceutics, and the International Journal of Pharmaceutics present the largest nodes, indicating a high frequency of citations and centrality in the network. These journals stand out for publishing studies with multidisciplinary approaches integrating nanotechnology, pharmacology, and bioengineering, reflecting their relevance in consolidating this research area. Additionally, titles such as the Journal of Drug Delivery Science and Technology and Biomaterials Science appear as important complementary sources, contributing to the diffusion of advances in biocompatible materials and controlled release strategies. The dispersion of other nodes demonstrates the diversity of journals that, although having lower citation impact, comprise the expanding scientific body on the subject.

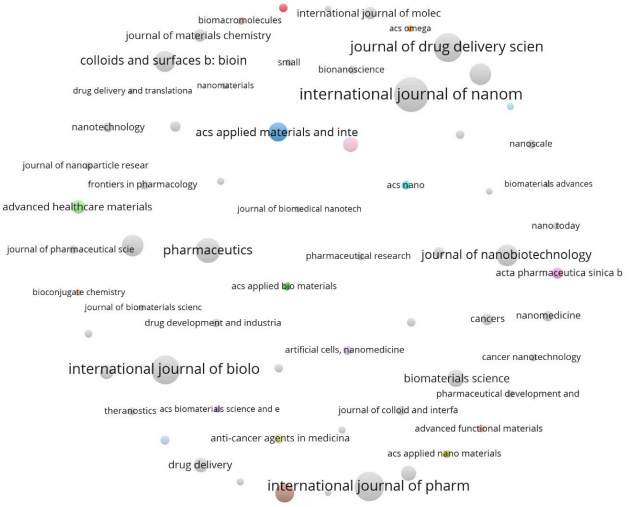


Fig. 6. Source citation map indicating the most influential journals in the analyzed research. The size of the nodes reflects citation frequency, highlighting journals such as International Journal of Nanomedicine, Pharmaceutics, and International Journal of Pharmaceutics.

F. Document Co-Citation Analysis

The document co-citation analysis highlights the main authors and works structuring the field of study on nanoparticle-based drug delivery systems applied to breast cancer treatment. As shown in Figure 7, the nodes are arranged in isolation, indicating the absence of consolidated collaborative networks.

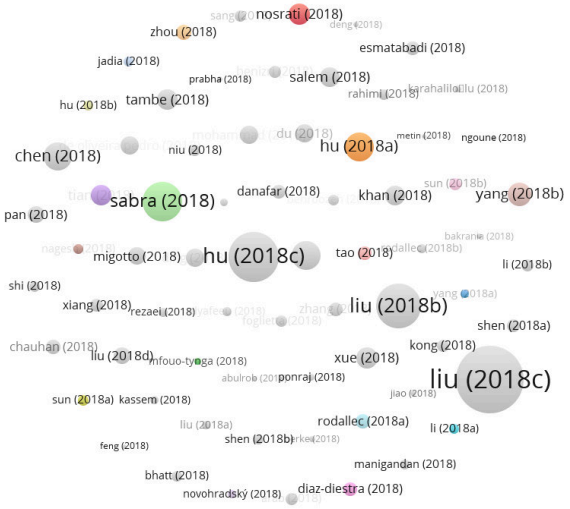


Fig. 7. Document co-citation map revealing the dispersed nature of recent studies. The isolation of nodes indicates weak connections between different research groups in the field.

This dispersion suggests that more recent studies are being conducted independently, possibly reflecting localized and individual efforts by research groups. Such a scenario points to a fragmentation of scientific production, which may limit the integrated advancement of knowledge in the area.

The formation of high citation density nuclei is observed around key works, particularly those by Hu (2018), Liu (2018), and Sabra (2018). These studies appear as central references, indicating they have exerted strong influence on the recent theoretical and methodological consolidation in the area.

G. Analysis of High-Impact Studies

The analysis of accumulated citation counts allows for the identification of the most influential and scientifically consolidated studies in the field of nano-oncology applied to breast cancer. Within the sample of 1,280 articles analyzed between 2018 and 2024, the highest-impact works are concentrated around three main technological axes: biomimetic systems for immune evasion, nanovaccines for antitumor immunotherapy, and multifunctional platforms for combination therapy. Table I summarizes the ten most cited works in the sample, highlighting the publishing journal and the total citations recorded in the Scopus database.

Table 1: most cited articles in the selected sample (2018–2024)

Authors	*Article Title*	*Journal*	Citations
J. Xu *et al.* (2020) [16]	"A general strategy towards personalized nanovaccines based on fluoropolymers for post-surgical cancer immunotherapy"	*Nature Nanotechnology*	468
S. Z. F. Phua *et al.* (2019) [17]	"Catalase-Integrated Hyaluronic Acid as Nanocarriers for Enhanced Photodynamic Therapy in Solid Tumor"	*ACS Nano*	341
L. Liu *et al.* (2018) [11]	"Combination Immunotherapy of MUC1 mRNA Nano-vaccine and CTLA-4 Blockade Effectively Inhibits Growth of Triple Negative Breast Cancer"	*Molecular Therapy*	313
L. Bayón-Cordero *et al.* (2019) [18]	"Application of solid lipid nanoparticles to improve the efficiency of anticancer drugs"	*Nanomaterials*	292
B. Bahmani *et al.* (2021) [19]	"Intratumoral immunotherapy using platelet-cloaked nanoparticles enhances antitumor immunity in solid tumors"	*Nature Communications*	251

W. Ma *et al.* (2022) [20]	"Biomimetic Nanoerythrosome-Coated Aptamer–DNA Tetrahedron/Maytansine Conjugates: pH-Responsive and Targeted Cytotoxicity for HER2-Positive Breast Cancer"	*Advanced Materials*	232
G. M. Sulaiman *et al.* (2020) [21]	"Hesperidin Loaded on Gold Nanoparticles as a Drug Delivery System for a Successful Biocompatible, Anti-Cancer, Anti-Inflammatory and Phagocytosis Inducer Model"	*Scientific Reports*	229
S. Shen *et al.* (2021) [22]	"A nanotherapeutic strategy to overcome chemotherapeutic resistance of cancer stem-like cells"	*Nature Nanotechnology*	223
S. Li *et al.* (2020) [23]	"Engineering macrophage-derived exosomes for targeted chemotherapy of triple-negative breast cancer"	*Nanoscale*	220
D. R. Hu *et al.* (2018) [10]	"Oxygen-generating hybrid polymeric nanoparticles with encapsulated doxorubicin and chlorin e6 for trimodal imaging-guided combined chemo-photodynamic therapy"	*Theranostics*	215

Fonte: Data extracted from the Scopus database through scientometric analysis. Prepared by the authors (2025).

As demonstrated in Table I, the study led by Xu et al. (2020), published in Nature Nanotechnology, stands out as the highest-impact paper in the sample (468 citations), proposing a personalized fluoropolymer-based nanovaccine strategy for post-surgical immunotherapy. Subsequently, the works by Phua et al. (2019) and Liu et al. (2018) reinforce the trend toward utilizing hyaluronic acid-based carriers and mRNA nanovaccines combined with immune checkpoint blockade for the treatment of triple-negative breast cancer.

IV. Discussion

The results of this scientometric analysis encompassing 1,280 articles from the Scopus database reveal a profound conceptual and technological evolution in the development of nanoparticle-based drug delivery systems for breast cancer treatment.

A.Geopolitical Reconfiguration and Fragmentation of Institutional Networks

The findings demonstrate robust growth in scientific production regarding nanocarriers for breast cancer, corroborating global trends that position nanomedicine as a central pillar in overcoming the limitations of conventional chemotherapy [1]. Beyond a numerical increase, however, a geopolitical reconfiguration of knowledge is observed.

China's prominence as the leader in publication volume corroborates recent global trends described by Yang et al. [6]. According to the authors, this Chinese hegemony in the development of nanosystems for breast cancer is a reflection of strategic state policies. This quantitative leadership suggests the consolidation of a research ecosystem in China that prioritizes high publication volume, driven by robust governmental funding and strategic institutional direction toward mastering this technology. While the Western landscape tends to focus on translational complexity, the Chinese model has demonstrated an accelerated capacity to generate new pharmaceutical materials and vectors at scale [6].

Institutional collaboration analysis reveals that few centers act as reference hubs, while most institutions remain isolated or form restricted clusters. This fragmentation corroborates the regionalization pattern described by Yang et al. [6], where scientific production remains concentrated in local initiatives driven by national funding with limited international integration, explaining the peripheral position of Latin American countries observed in co-authorship maps. This pattern is also reflected in the co-citation analysis, which showed small, loosely connected clusters, suggesting that despite high publication volume, effective knowledge exchange between major Asian and Western hubs remains limited [6].

B.Specific Contributions to Biomaterials and Biomedical Engineering

Beyond updating bibliometric maps, the data obtained in this investigation reveal a technological transition in

biomaterials engineering applied to breast oncology. The thematic clustering of keywords (Fig. 5) exposes a paradigm shift in the physicochemical architecture of nanocarriers, corroborating and expanding recent bibliometric analyses on polymers and lipid vectors in biomedical engineering:

- Matrix Transition and Advanced Biopolymers: Unlike the early stages of oncological nanomedicine, which were predominantly marked by rigid first-generation liposomes and metallic or inorganic nanostructures, recent literature in Scopus demonstrates the consolidation of nanostructured lipid carriers (NLCs), solid lipid nanoparticles (SLNs), and biodegradable polymeric matrices—such as PLGA, chitosan, and hyaluronic acid. This transition is driven by the need to overcome biocompatibility and drug loading capacity bottlenecks, as highlighted by reviews in materials science [8].

- Overcoming the EPR Effect and Surface Engineering (Targeted Delivery): Co-occurrence maps show that the focus of biomaterials engineering has shifted from passive accumulation based on enhanced permeability and retention (EPR effect)—now recognized for its limitations in clinical settings [9]—to active targeting. Notably, functionalizing nanoparticles with specific ligands, such as monoclonal antibodies (anti-HER2/trastuzumab), aptamers, and peptides (RGD), promotes selective recognition and receptor-mediated endocytosis in breast tumor cells.

- Smart Stimuli-Responsive Nanosystems: Keyword analysis indicates substantial progress in designing biomaterials responsive to the tumor microenvironment (TME). The mapped literature highlights matrices capable of on-demand drug release in response to endogenous variations (intratumoral acidic pH, redox/glutathione GSH gradients, and matrix metalloproteinase overexpression) or exogenous stimuli, such as magnetic hyperthermia, photothermy, and ultrasound [7].

C. Main Seminal References and Emerging Trends in 2024

Document co-citation analysis (Fig. 7) identified the theoretical pillars underpinning research during this period. Seminal works from 2018 stand out as central references by introducing advanced strategies for treating aggressive tumors and Triple-Negative Breast Cancer (TNBC). Key approaches include nano-immunotherapy and mRNA vaccines combined with checkpoint blockade [11], oxygen-generating, stimuli-responsive polymeric systems designed to maximize penetration and chemotherapeutic efficiency [10], and targeted protein micelles for drug co-delivery and efflux pump inhibition [12].

In 2024, when academic output consolidated at the highest point of the time series (232 articles), scientometric analysis revealed the emergence of three primary technological axes:

- Immunonanomedicine and Nano-immunotherapy: A transition from using nanocarriers merely as drug delivery vehicles to active agents that reconfigure the tumor immune microenvironment. Strategies prioritize targeted delivery of immunomodulators, STING pathway agonists, and mRNA vaccines to enhance immune checkpoint blockade (e.g., anti-PD-1 and anti-PD-L1) and induce immunogenic cell death [13], [14].
- Integrated Oncological Theranostics: Co-incorporation of diagnostic or imaging agents (such as quantum dots and iron oxide nanoparticles) with therapeutic drugs into a single multifunctional nanoplatform, enabling real-time monitoring of drug release profiles and tumor regression.
- Biomimetic Platforms and Hybrid Vectors: Adoption of nanoparticles cloaked in mammalian cell membranes (such as platelets, erythrocytes, exosomes, or tumor cell membranes), conferring reticuloendothelial system evasion (stealth behavior) and enhanced homotypic targeting toward the tumor microenvironment [15].

V. Conclusions

From this perspective, it becomes evident that strengthening institutional partnerships, expanding international cooperation, and investing in translational studies are key elements to accelerate the scientific maturation of the field. In this scenario, future progress in breast cancer treatment tends to depend not only on increasingly sophisticated nanosystems but also on the capacity to build truly global scientific collaboration networks capable of integrating knowledge, leveraging technologies, and promoting innovations with real impact on clinical practice. Thus, it is expected that this analysis contributes to the understanding of the field's recent dynamics and serves as a basis to guide future research, stimulating initiatives that broaden integration between groups, strengthen evidence production, and favor the development of promising technologies for the care of breast cancer patients.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 74, no. 3, pp. 229–263, May 2024, doi: 10.3322/caac.21834.

[2] K. Cheng, Q. Guo, Z. Shen, W. Yang, Y. Wang, Z. Sun, and H. Wu, "Bibliometric analysis of global research on cancer photodynamic therapy: Focus on nano-related research," Front. Pharmacol., vol. 13, p. 927219, Jun. 2022, doi: 10.3389/fphar.2022.927219.

[3] Scopus, "Scopus database," Elsevier, 2025. [Online]. Available: https://www.scopus.com. [Accessed: Oct. 17, 2025].

[4] Instituto Nacional de Câncer (INCA), Estimativa 2023: Incidência de Câncer no Brasil. Rio de Janeiro, Brazil: INCA, 2022.

[5] S. Q. Torres, M. P. A. Ferreira, F. Sousa, and A. Cruz, "Nanotherapeutics in women's health: Emerging nanotechnologies for triple-negative breast cancer treatment," Small, vol. 20, no. 18, p. 2300666, May 2024, doi: 10.1002/smll.202300666.

[6] J.-Y. Yang, Y. Zhang, X. Li, Z. Wang, and Q. Liu, "Bibliometric analysis of global research trends in delivery systems for breast cancer immunotherapy," Discov. Oncol., vol. 16, no. 1, p. 1829, 2025, doi: 10.1007/s12672-025-01829-w.

[7] M. T. Manzari, A. Ackermann, T. D. Taylor, C. Triantafyllou, M. G.

Shamay, and D. A. Heller, "Targeted drug delivery strategies for cancer therapeutics," Nat. Rev. Mater., vol. 6, no. 4, pp. 351–370, Apr. 2021, doi: 10.1038/s41578-020-00269-4.

[8] M. J. Mitchell, M. M. Billingsley, R. M. Haley, M. E. Wechsler, N. A. Peppas, and R. Langer, "Engineering precision nanoparticles for drug delivery," Nat. Rev. Drug Discov., vol. 20, no. 2, pp. 101–124, Feb. 2021, doi: 10.1038/s41573-020-0090-8.

[9] R. van der Meel, E. Sulheim, Y. Shi, F. Kiessling, W. J. M. Mulder, and T. Lammers, "Smart cancer nanomedicine," Nat. Nanotechnol., vol. 14, no. 11, pp. 1007–1017, Nov. 2019, doi: 10.1038/s41565-019-0567-y.

[10] D. R. Hu, L. Chen, Y. Qu, J. Peng, B. Y. Chu, K. Shi, Y. Hao, L. Zhong, M. Y. Wang, and Z. Qian, "Oxygen-generating hybrid polymeric nanoparticles with encapsulated doxorubicin and chlorin e6 for trimodal imaging-guided combined chemo-photodynamic therapy," Theranostics, vol. 8, no. 6, pp. 1558–1574, Mar. 2018, doi: 10.7150/thno.22989.

[11] L. Liu, Y. Wang, L. Miao, Q. Liu, S. Musetti, J. Li, and L. Huang, "Combination immunotherapy of MUC1 mRNA nano-vaccine and CTLA-4 blockade effectively inhibits growth of triple negative breast cancer," Mol. Ther., vol. 26, no. 1, pp. 45–55, Jan. 2018, doi: 10.1016/j.ymthe.2017.10.020.

[12] S. A. Sabra, D. M. Abdelkader, A. H. El-Kamel, H. M. El-Sawy, and M. A. El-Nabarawi, "Self-assembled amphiphilic zein-lactoferrin micelles for tumor targeted co-delivery of rapamycin and wogonin to breast cancer," Eur. J. Pharm. Biopharm., vol. 128, pp. 156–169, Jul. 2018, doi: 10.1016/j.ejpb.2018.04.023.

[13] S. Huang, X. Li, Y. Zhang, and Z. Wang, "Nanoparticle-integrated dissolving microneedles for the co-delivery of R848/aPD-1 to synergistically reverse the immunosuppressive microenvironment of triple-negative breast cancer," Acta Biomater., vol. 176, pp. 344–355, Mar. 2024, doi: 10.1016/j.actbio.2024.01.009.

[14] J. Liang, X. Chen, Y. Zhao, and H. Liu, "Shikonin and chitosan-silver nanoparticles synergize against triple-negative breast cancer through RIPK3-triggered necroptotic immunogenic cell death," Biomaterials, vol. 309, art. 122608, Aug. 2024, doi: 10.1016/j.biomaterials.2024.122608.

[15] W. Xiong, Z. Dong, and Y. Chao, "Biomimetic tumor cell membrane-encapsulated nanoparticles combine NIR-II photothermal therapy and chemotherapy for enhanced immunotherapy in triple-negative breast cancer," Adv. Funct. Mater., vol. 34, no. 52, art. 2410841, Dec. 2024, doi: 10.1002/adfm.202410841.

[16] J. Xu, J. Lv, Q. Zhuang, Z. Yang, Z. Cao, L. Xu, P. Pei, C. Wang, H. Wu, Z. Dong, Y. Chao, C. Wang, K. Yang, R. Peng, Y. Cheng, and Z. Liu, "A general strategy towards personalized nanovaccines based on fluoropolymers for post-surgical cancer immunotherapy," Nat. Nanotechnol., vol. 15, no. 12, pp. 1043–1052, Dec. 2020, doi: 10.1038/s41565-020-00781-4.

[17] S. Z. F. Phua, G. Yang, W. Q. Lim, A. Verma, H. Chen, T. Thanabalu, and Y. Zhao, "Catalase-integrated hyaluronic acid as nanocarriers for enhanced photodynamic therapy in solid tumor," ACS Nano, vol. 13, no. 4, pp. 4742–4751, Apr. 2019, doi: 10.1021/acsnano.9b01087.

[18] L. Bayón-Cordero, I. Alkorta, and L. Arana, "Application of solid lipid nanoparticles to improve the efficiency of anticancer drugs," Nanomaterials, vol. 9, no. 3, p. 474, Mar. 2019, doi: 10.3390/nano9030474.

[19] B. Bahmani, H. Gong, B. T. Luk, K. J. Haushalter, E. DeTeresa, M. Previti, J. Zhou, W. Gao, J. D. Bui, L. Zhang, R. H. Fang, and J. Zhang, "Intratumoral immunotherapy using platelet-cloaked nanoparticles enhances antitumor immunity in solid tumors," Nat. Commun., vol. 12, no. 1, p. 1999, Apr. 2021, doi: 10.1038/s41467-021-22311-z.

[20] W. Ma, Y. Yang, J. Zhu, W. Jia, T. Zhang, Z. Liu, X. Chen, and Y. Lin, "Biomimetic nanoerythrosome-coated aptamer–DNA tetrahedron/maytansine conjugates: pH-responsive and targeted cytotoxicity for HER2-positive breast cancer," Adv. Mater., vol. 34, no. 15, p. 2109609, Apr. 2022, doi: 10.1002/adma.202109609.

[21] G. M. Sulaiman, H. M. Waheeb, M. S. Jabir, S. H. Khazaal, Y. H. Dewir, and Y. Naidoo, "Hesperidin loaded on gold nanoparticles as a drug delivery system for a successful biocompatible, anti-cancer, anti-inflammatory and phagocytosis inducer model," Sci. Rep., vol. 10, no. 1, p. 9362, Jun. 2020, doi: 10.1038/s41598-020-66419-6.

[22] S. Shen, X. Xu, S. Lin, Y. Zhang, H. Liu, C. Zhang, and R. Mo, "A nanotherapeutic strategy to overcome chemotherapeutic resistance of cancer stem-like cells," Nat. Nanotechnol., vol. 16, no. 1, pp. 104–113, Jan. 2021, doi: 10.1038/s41565-020-00793-0.

[23] S. Li, Y. Wu, F. Ding, J. Yang, J. Li, X. Gao, C. Zhang, and J. Feng, "Engineering macrophage-derived exosomes for targeted chemotherapy of triple-negative breast cancer," Nanoscale, vol. 12, no. 19, pp. 10854–10862, May 2020, doi: 10.1039/d0nr00523a.

Study of the Gelation Process of an Injectable Hydrogel Composed of Natural Materials

1st Beatriz S. Rodrigues
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
beatrizrodrigues@geb.inatel.br

2nd Évilyn N. Ferreira
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
evilyn.n@geb.inatel.br

3rd Filipe B. Vilela
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
filipe.bueno@inatel.br

4th Luma R. B. Prado
Instituto Nacional de Telecomunicações (INATEL)
Santa Rita do Sapucaí, Brazil
luma.prado@inatel.br

***Abstract*—Injectable hydrogels show potential for biomedical applications due to their biocompatibility and ability to facilitate controlled drug release. This study developed and characterized a hydrogel based on porcine gelatin, hyaluronic acid, propolis, and genipin, evaluating its gelation, degradation, and swelling properties. The material exhibited gelation under conditions close to physiological ones, structural stability, and low swelling capacity. The results indicate potential for applications in controlled substance release and tissue engineering.**

***Index Terms*—Injectable hydrogel, Natural biomaterials, Controlled release, Gelation, Physiological pH.**

I. Introduction

Traditional drug delivery systems, such as tablets or conventional pharmacological injections, often promote rapid drug release after administration, resulting in a high initial concentration of the drug in the body, a phenomenon known as burst release. This effect may lead to adverse reactions or reduced therapeutic efficacy over time due to the rapid elimination of the drug. This immediate release profile makes it difficult to maintain stable therapeutic levels within the body, potentially causing initial overdose episodes and requiring repeated administrations to sustain the desired pharmacological effect [1] [2].

Another relevant factor is that many systems used in biomaterials or drug delivery vehicles are based on synthetic polymers, which may present limitations related to biocompatibility, toxicity of degradation by-products, or poor interaction with biological tissues. These factors can trigger inflammatory responses or hinder their proper absorption by the body. Therefore, these limitations highlight the need for the development of new biomaterials capable of promoting controlled drug release, known as drug delivery systems [2] [3].

Among the most widely used matrices in drug delivery systems are hydrogels, which are polymeric materials characterized by a three-dimensional network architecture formed by hydrophilic polymers capable of absorbing and retaining large amounts of water while maintaining their structure due to the presence of physical or chemical crosslinks between polymer chains [4]. Among the different categories of hydrogels, injectable hydrogels represent a distinct subclass capable of overcoming biological barriers through subcutaneous injection. This type of administration enables the dispersion of the material within the skin or other tissues, allowing it to reach the desired site through different delivery pathways [5].

Natural biopolymers, such as gelatin, chitosan, and chitin, combined with bioactive compounds such as Allium sativum, Aloe vera, and propolis, exhibit therapeutic properties that support the development of injectable hydrogels for biomedical applications [6] [7]. Propolis stands out due to its antimicrobial, anti-inflammatory, antioxidant, and regenerative activities, contributing to wound healing and tissue repair processes [8].

Injectable hydrogels can be applied in several areas of medicine, such as cardiac regeneration after myocardial infarction, where they act as extracellular matrix-like scaffolds, promoting cell adhesion and proliferation, as well as the controlled release of bioactive molecules [9]. Another promising application is in the treatment of osteoarthritis and cartilage regeneration, as these materials can modulate the inflammatory environment, reduce oxidative stress, and stimulate tissue repair. Therefore, injectable hydrogels represent a versatile and promising therapeutic strategy for regenerative medicine [10].

Within this context, pH-responsive systems present important advantages, such as controlled substance release, enhanced targeting to the site of action, and improved interaction with the body [7]. Changes in the environmental pH can lead to the protonation or deprotonation of functional groups present in polymer chains, modifying the electrostatic interactions between them and directly influencing the organization and density of the hydrogel network. As a result, it is possible to regulate properties such as the swelling degree (the ability of the hydrogel to absorb and retain water within its structure, leading to the expansion of the polymeric network), the degradation rate (the process of polymer chain breakdown and consequent material disintegration over time within the body), and the drug release profile, promoting greater material

retention at the application site and enabling a more controlled therapeutic release [11].

Within this perspective, injectable hydrogels present the challenge of remaining in a liquid state during administration and forming a stable gel network after application, promoting tissue adhesion and retention of the therapeutic substance [7]. However, this transition must occur in a manner compatible with physiological conditions, such as temperature and pH, which influence the organization and stability of the polymeric network [11]. Changes in pH promote the protonation or deprotonation of functional groups, enabling the control of properties such as swelling (water absorption and retention by the matrix), degradation (breakdown of polymer chains over time), and substance release, thereby supporting more controlled therapeutic applications [11].

Given the above, the objective of this research is to develop an injectable hydrogel using natural materials with bactericidal and regenerative properties, as well as to investigate the parameters involved in the gelation rate of this hydrogel.

II. Materials and methods

Initially, a study was conducted on injectable hydrogels composed of natural polymers capable of undergoing gelation at body temperature, approximately 36.5°C, exhibiting high water solubility and responsiveness to the physiological pH of 7.4.

A. Materials

For the development of the hydrogel and the execution of the experimental assays, the following materials were used: porcine gelatin (300 Bloom), genipin, propolis, hyaluronic acid (HA), fluorescein, acetic acid, and phosphate-buffered saline solution (PBS, pH 7.4).

B. Hydrogel production and gelation

Initially, two experimental groups were established to evaluate hydrogel formation, designated as Group A and Group B. Each group had 5 batches of hydrogel (n = 5) to ensure the reproducibility of the results. The formulations were developed based on the methodology described in [12], with adaptations according to the objectives of this study. Group A consisted of 500 mg of porcine gelatin (99% purity, Sigma Aldrich), 100 mg of hyaluronic acid (HA; quality leve 100, 500 kDA, Sigma Aldrich), 100 mg of propolis (Apis Flora, 500 mg/g), and 50 mg of genipin (¿ 98%, Sigma Aldrich) as a crosslinking agent. Group B consisted of 750 mg of the same porcine gelatin, 100 mg of the same HA, 100 mg of the same propolis, and 75 mg of the same genipin.

For the preparation of the formulations, powdered porcine skin gelatin was dissolved in 9 ml of distilled water and 1 ml of acetic acid (99%, glacial) under vigorous stirring using a magnetic stirrer at approximately 60 °C for 15 minutes until complete solubilization. Powdered HA was dissolved in 10 ml of distilled water and maintained under vigorous stirring for 15 minutes at approximately 25 °C. Powdered propolis was dissolved in 10 ml of distilled water under vigorous stirring and mild heating (approximately 40 °C) to assist solubilization. Finally, genipin was added as a crosslinking agent at the concentrations corresponding to each experimental group.

Subsequently, for hydrogel formation from the Group B formulation, the gelatin solution was mixed with the HA solution at approximately 40 °C under vigorous stirring for 15 minutes until a visually homogeneous mixture was obtained. Afterwards, acetic acid was added to the solution. Subsequently, the propolis solution was incorporated, maintaining stirring for an additional 15 minutes. Fluorescein was previously prepared at a concentration of 0.35 mg/ml by dissolving it in distilled water. Then, 0.2 ml of this solution was added to the sample, which remained under stirring for an additional 15 minutes to ensure homogenization. After this step, genipin was incorporated, and the mixture was maintained under vigorous stirring for 30 minutes until the onset of crosslinking of the polymeric system.

Subsequently, the final hydrogel solution was weighed using an analytical balance (SHIMADZU, model ATY224R), and the hydrogel was divided into four beakers (B1, B2, B3, and B4) containing equivalent amounts, corresponding to 25% (5.09 mg) of the total prepared hydrogel (20.36 mg). Beaker B1 was designated as the control group and was used as a reference for comparison with the other groups. All assays were performed in triplicate to ensure the reproducibility of the results.

To perform the assays under conditions close to physiological conditions, a PBS solution was prepared using a concentration of 40 g of NaOH per mol. The solution was added dropwise until reaching an approximate pH of 7.

Each formulation was placed in a water bath at an approximate temperature of 36.5 °C, monitored using a mercury thermometer, and 2 ml of PBS solution was simultaneously added to each group. The mixtures remained in the water bath for 15 minutes to simulate physiological temperature and pH conditions, enabling the evaluation of gelation and hydrogel formation.

Thus, the samples were injected through a 25 G needle with a length of 25 mm using manual pressure to evaluate potential needle occlusion. Subsequently, gel formation and structural integrity, continuity of the polymeric matrix, and visual homogeneity of the system were evaluated. Photographic records of the obtained samples were performed for documentation of the results.

C. Characterization

1) Degradation: The sample was maintained at rest under dry conditions for 20 minutes. After this period, 3 ml of PBS solution at approximately 36.5 °C was added to the hydrogel, and the degradation process was monitored through photographic records at intervals of 5, 15, 30, 60, and 90 minutes, allowing the visual analysis of the degradation process over time.

2) Swelling: For swelling evaluation, 14.49 ml of PBS solution at approximately 36.5 °C was added to the system.

The evaluations were performed at 5, 15, 30, 60, and 90 minutes. At each time interval, the supernatant liquid was removed to allow the hydrogel mass to be measured using an analytical balance, enabling the quantitative monitoring of the swelling degree. Additionally, the removed liquid volume was measured using a graduated cylinder and subsequently returned to the system together with the hydrogel. When necessary, the volume was replenished with PBS solution, maintaining the initial volume constant throughout the entire assay. As control, the initial dry mass and volume were measured.

3) Fluorescein release rate: To evaluate the fluorescein release rate, 6 ml of PBS solution at approximately 36.5 °C was added to the system. The evaluations were performed at 5, 15, 30, 60, and 90 minutes. At each time interval, 2 ml of the supernatant solution was collected and stored in properly identified Eppendorf tubes. After collection, the removed volume was immediately replaced by adding 2 ml of PBS solution to the system, maintaining the initial volume constant and ensuring standardized experimental conditions throughout the entire assay.

The CELM spectrophotometer, model E-225 D, was calibrated using standard solutions prepared from an initial fluorescein concentration of 0.35 mg/ml diluted in distilled water. From this solution, five distinct concentrations were obtained through dilutions with factors of 7.5×, 8×, 12×, 16×, and 32×, corresponding to a progressive reduction of the initial concentration. The final concentrations obtained were 0.0467 mg/ml, 0.0438 mg/ml, 0.0292 mg/ml, 0.0219 mg/ml, and 0.0109 mg/ml. For each dilution, four absorbance measurements were performed to verify the linearity of the equipment response and estimate potential measurement errors.

Subsequently, the 2 ml samples reserved in the Eppendorf tubes were analyzed using the spectrophotometer to determine the fluorescein release rate over time by measuring absorbance at 490 nm.

The absorbance values of the experimental groups (B2, B3, and B4) were corrected by subtracting the mean absorbance of the control group (B1) at the time points of 5, 15, 30, 60, and 90 minutes. These data enabled the subsequent correlation with the released fluorescein concentration through the linear equation obtained from the calibration curve of the spectrophotometer, adjusted by linear regression using Microsoft Excel.

III. Results

The results obtained from the experimental assays were organized according to the methodological steps previously described, including hydrogel production, swelling and degradation assays, and the evaluation of the fluorescein release rate.

A. Hydrogel production and gelation

The formation of the hydrogels and the gelation process were evaluated after the crosslinking step, considering the two previously developed formulations (Figure 1). To promote the crosslinking reaction, genipin was maintained under stirring for 30 minutes, favoring the formation of the hydrogel polymeric network. After 15 minutes in a water bath at approximately 36.5 °C, both formulations exhibited gel formation.

In Group A, the formation of a hydrogel with a heterogeneous structure, lower structural integrity, and a bluish coloration due to the genipin reaction was observed. In contrast, Group B exhibited a more homogeneous and continuous structure, with greater uniformity of the polymeric matrix and absence of the characteristic blue coloration.

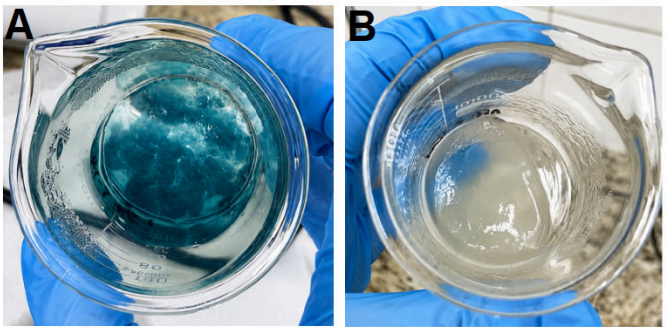


Fig. 1. Macroscopic appearance of the developed hydrogels: (a) Group A hydrogel and (b) Group B hydrogel. Authors, 2026.

Based on the structural characteristics observed during the macroscopic evaluation of the obtained hydrogels, Group B was selected for further investigation. Therefore, the subsequent assays and results presented in this study were conducted using the formulation that exhibited greater visual uniformity and improved structural integrity after the gelation process.

B. Degradation

After the addition of the PBS solution to the hydrogel, the sample was visually monitored at intervals of 5, 15, 30, 60, and 90 minutes through photographic records. Throughout the experimental period, the hydrogel maintained its structure, with no visible disintegration of the material or significant loss of its initial conformation, as presented in Figure 2.

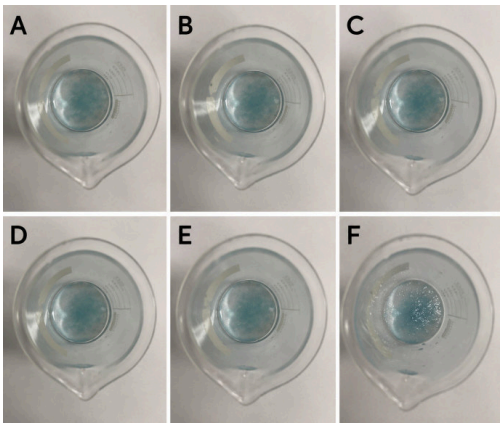


Fig. 2. Visual evaluation of hydrogel degradation over time. The images represent the samples after (A) 5 min, (B) 15 min, (C) 30 min, (D) 60 min, and (E) 90 min of the assay. In (F), the hydrogel is observed after 90 min without supernatant liquid. Authors, 2026.

During the initial time intervals, the hydrogel maintained its structural integrity. With increasing exposure time to the PBS solution, gradual structural changes and progressive darkening of the polymeric matrix were observed, indicating the progression of the degradation process.

C. Swelling

After the addition of the PBS solution to the hydrogel, the swelling process was monitored at intervals of 5, 15, 30, 60, and 90 minutes through the evaluation of the sample volume and mass. Throughout the entire experimental period, the hydrogel maintained its structural integrity, with no visible disintegration of the polymeric matrix.

The volume values of the PBS remaining external to the hydrogel remained relatively stable throughout the assay, presenting 14 ml at the 5, 15, and 90-minute intervals, with an increase to 14.3 ml at 30 minutes and a maximum value of 14.5 ml at 60 minutes. Regarding mass, values of 34.08 g at 5 minutes, 34.05 g at 15 minutes, 34.06 g at 30 minutes, 33.91 g at 60 minutes, and 33.95 g at 90 minutes were recorded. The initial values of weight and volume were 33.8 g and 13.2 ml, respectively.

In addition to the quantitative variations observed, gradual macroscopic changes in the hydrogel were identified throughout the analyzed intervals, including subtle modifications in the material's uniformity and coloration. The results obtained from the swelling assay are presented in Figure 3.

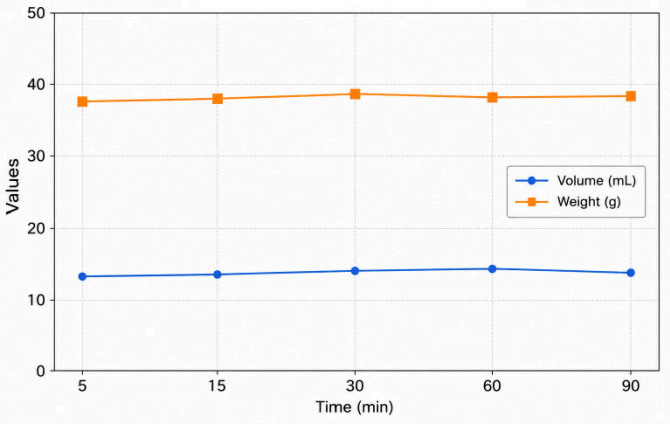


Fig. 3. Swelling and supernatant liquid volume over time. Authors, 2026.

D. Fluorescein release rate

Based on the mean absorbance values obtained for each dilution, the spectrophotometer calibration curve was constructed. The calibration results showed good agreement among replicates, considering that the calibration curve presented a coefficient of determination (R^2) of 0.9547.

The correlation between absorbance and the released fluorescein concentration was performed using the trendline equation $y = 21.46x - 0.0146$ (where y represents absorbance and x represents concentration), enabling the evaluation of the release profile of the system.

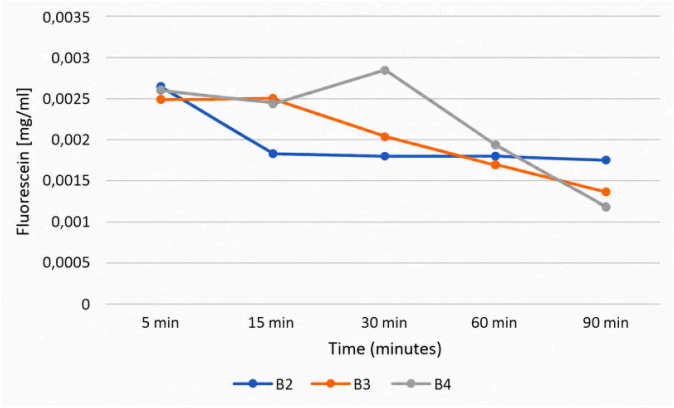


Fig. 4. Estimated fluorescein concentration released from the hydrogel over time. Authors, 2026.

The fluorescein release from the hydrogel was evaluated by spectrophotometry at different time intervals (5, 15, 30, 60, and 90 minutes). As presented in Figure 4, Group B2 exhibited a more pronounced initial reduction between 5 and 15 minutes, decreasing from approximately 0.00264 mg/ml to 0.00185 mg/ml. After this interval, the values remained practically constant until 90 minutes, ranging between 0.00175 and 0.00180 mg/ml, indicating stabilization of the concentration after the initial decrease. Group B3, in turn, showed a gradual and continuous reduction throughout the entire assay, with an initial concentration of approximately 0.00250 mg/ml and a final value of 0.00138 mg/ml at 90 minutes, representing the greatest overall reduction observed among the groups. In contrast, Group B4 showed an initial maintenance of concentration between 5 and 15 minutes, followed by an increase until 30 minutes, reaching a maximum value of approximately 0.00287 mg/ml. After this peak, a progressive reduction in concentration was observed, reaching 0.00119 mg/ml at 90 minutes, resulting in the lowest final concentration among the evaluated groups.

IV. Discussion

The 15-minute incubation period in a water bath at approximately 36.5 °C and pH 7.4, simulating physiological conditions, proved to be adequate for the gelation of the evaluated formulations. This interval was sufficient to promote the sol–gel transition under conditions similar to those encountered in the human body, enabling the formation of the polymeric network. Furthermore, the establishment of this gelation time is particularly relevant for in situ applications, as excessively slow gelation may lead to material dispersion at the administration site before hydrogel formation, thereby compromising its retention and therapeutic performance [13].

The results demonstrated that the formulation composition directly influenced the formation and structural stability of the hydrogels. In Group A, the lower concentrations of gelatin and genipin resulted in a less cohesive matrix with reduced structural integrity, indicating insufficient formation of the hydrogel's three-dimensional network [14].

In contrast, Group B exhibited a more continuous, homogeneous, and visually uniform structure, suggesting greater efficiency in the crosslinking process. The increased concentrations of gelatin and genipin may have favored the formation of a more organized and stable polymeric network, thereby contributing to improved hydrogel structural integrity. Genipin acts as a crosslinking agent by reacting with primary amine groups present in the gelatin polymer chains. This process occurs progressively, promoting the formation of intermolecular covalent bonds and the development of a more stable three-dimensional network. These findings indicate that the ratio between the formulation components exerts a significant influence on the physical and structural properties of the material [14].

The successful formation of the hydrogel under near-physiological conditions, such as a temperature of approximately 36.5 °C and pH 7.4, represents an important characteristic for biomedical applications, particularly in injectable controlled-release systems. Under these conditions, the material can remain in a liquid state during syringe administration, facilitating a minimally invasive delivery, and subsequently undergo in situ gelation at the application site [15].

From a physicochemical perspective, this behavior demonstrates the occurrence of sufficient interactions among the polymer chains to promote the formation of a stable three-dimensional structure under physiological conditions. Many polymeric systems face challenges related to the establishment of a three-dimensional network following administration due to limitations in gelation kinetics [15]. Therefore, the gelation capacity observed in the selected formulation indicates efficient organization of the polymeric matrix, favoring the entrapment of the incorporated compound within its structure. This characteristic directly contributes to a slower and more sustained drug release profile, increasing its residence time in the body and supporting its potential for controlled therapeutic delivery applications [16].

In the degradation assay, the hydrogel maintained its structural integrity throughout the 90-minute evaluation period, with no visible disintegration of the material. However, gradual changes in the system's uniformity, consistency, and coloration were observed during exposure to the PBS solution.

The occurrence of these reactions is directly dependent on the equilibrium between the protonated (-NH_3+) and deprotonated (-NH_2) forms of amine groups. Under more acidic conditions, the protonated form predominates. However, at pH values close to neutrality, such as those found in the physiological environment (pH 7.4), a greater proportion of amine groups is available in the deprotonated form (-NH_2), which favors the formation of crosslinks as well as the development of the characteristic blue coloration of the hydrogel [14] [17].

From a therapeutic application perspective, the degradation rate of the hydrogel is directly related to its potential use. Formulations exhibiting faster degradation are generally more suitable for short-term applications, such as the local delivery of anti-inflammatory drugs, anesthetics, or antimicrobial agents during acute conditions and the immediate postoperative period [13]. Conversely, materials with slower degradation profiles may be employed in treatments that require prolonged system activity, allowing for the sustained retention of the incorporated compound. Such systems are particularly advantageous for applications including ocular therapies, chemotherapy, and tissue engineering, where extended therapeutic action is desirable [9] [18] [19] [20].

In this context, the gradual degradation profile observed suggests the potential applicability of the hydrogel in longer-term therapeutic strategies, in which the material may function as a temporary scaffold for the sustained release of bioactive substances. Furthermore, the prolonged structural persistence of the hydrogel may contribute to a reduced frequency of administration, thereby enhancing therapeutic stability and improving patient comfort. Therefore, long-term studies and in vivo evaluations should be conducted to validate these findings and provide more robust conclusions regarding the hydrogel's stability, biodegradation behavior, and therapeutic potential for clinical applications.

In the swelling assay, the hydrogel exhibited limited absorption of the PBS solution, a phenomenon associated with water uptake into the polymeric matrix and the consequent partial separation of the polymer chains within the material's three-dimensional structure. This behavior is influenced by the presence of hydrophilic groups in gelatin and hyaluronic acid, which are capable of interacting with water molecules through hydrogen bonding. However, the crosslinking promoted by genipin favors the formation of a denser and more organized polymeric network, restricting chain mobility and reducing the available space for water incorporation. As a result, a lower degree of swelling is observed, even in the presence of water-affinitive groups, allowing only limited internal structural rearrangement without compromising the hydrogel's structural integrity [21] [22] [23].

Water uptake into the polymeric matrix also facilitates the diffusion and gradual release of the drug incorporated within the hydrogel. Consequently, the system functions as a reservoir capable of providing controlled release of the active substance over time. This mechanism contributes to minimizing the phenomenon known as burst release, which is characterized by the rapid and excessive initial release of the encapsulated drug [24].

To evaluate this release behavior, the results obtained from the spectrophotometer calibration showed good agreement among replicates, with low variability between measurements performed at the same concentration. This finding demonstrates the stability of the spectrophotometric readings and the suitability of the experimental procedure employed. Furthermore, the calibration curve exhibited a coefficient of determination (R^2) of 0.9547, indicating a strong linear correlation between concentration and absorbance within a range considered acceptable for experimental assays of this nature.

A greater release was observed during the initial time points, followed by a reduction throughout the course of the assay. This behavior is associated with the low swelling capacity of the hydrogel, which limits the diffusion of fluorescein through

the polymeric matrix. Consequently, the restricted water uptake and reduced expansion of the network structure decrease molecular mobility within the system, thereby slowing the release process over time.

The results suggest that the hydrogel exhibited limited fluorescein release throughout the assay, as a more pronounced release profile would be expected to result in an increase in the concentration of the surrounding medium, even after the dilution steps. These findings indicate that the polymeric matrix was able to effectively retain the incorporated compound, restricting its diffusion into the external environment and promoting a more controlled release behavior under the experimental conditions evaluated.

Therefore, therapeutic applications requiring prolonged and controlled drug administration may particularly benefit from this type of system. These include treatments for chronic wound healing, tissue regeneration, ocular therapies, growth factor delivery, and tissue engineering applications. Among the most relevant clinical applications are cardiac regeneration following acute myocardial infarction, as well as the treatment of osteoarthritis and articular cartilage regeneration. In these contexts, the injectable nature of hydrogels plays a crucial role, as the material can be administered initially in a liquid or low-viscosity state, facilitating its minimally invasive delivery directly to the target tissue. This initial fluidity enables administration through less invasive procedures, allowing better adaptation to the application site and reducing tissue damage when compared with solid implants or surgical approaches [9] [10].

Following injection, the system undergoes a sol–gel transition, resulting in the formation of a three-dimensional network that conforms to the tissue microenvironment and remains retained at the application site. In post-infarction cardiac regeneration, this retention is particularly important, as the hydrogel can be administered directly into the injured myocardium, reducing its rapid dispersion through the bloodstream and enabling the sustained delivery of cells or bioactive molecules to a tissue with limited regenerative capacity. Similarly, in osteoarthritis, gel formation within the joint space allows the material to remain in an environment characterized by high mobility and continuous synovial fluid turnover, conditions that would normally promote the rapid clearance of drugs administered in liquid formulations [9] [10].

Compared with conventional approaches based on simple liquid injections, in which the drug tends to diffuse rapidly or be cleared by systemic circulation, limited local retention and the need for repeated administrations to maintain therapeutic efficacy are commonly observed. In this context, hydrogels function as controlled-release matrices, allowing enhanced retention at the target site and sustained release over time. These properties contribute to improved therapeutic efficacy, reduced frequency of interventions, and better spatial and temporal control of therapy [18] [19] [20].

V. Conclusion

The present study successfully achieved its proposed objectives by developing and characterizing an injectable hydrogel based on porcine gelatin, hyaluronic acid, propolis, and genipin. The material exhibited gelation under near-physiological conditions, structural stability throughout the experimental assays, low swelling capacity, and a controlled fluorescein release profile, demonstrating characteristics that are compatible with biomedical applications.

The results reinforce the potential of the developed hydrogel for applications in controlled-release systems and tissue engineering. As future perspectives, additional physicochemical characterization, as well as in vitro and in vivo studies, are recommended to further investigate the material's performance and biocompatibility. Moreover, evaluating the antimicrobial activity of propolis is essential to broaden the understanding of the hydrogel's therapeutic potential and clinical applicability.

References

[1] B. Liu and K. Chen, "Advances in hydrogel-based drug delivery systems," Gels, vol. 10, no. 4, Art. no. 262, 2024.

[2] H. Hameed, S. Faheem, A. C. Paiva-Santos, H. S. Sarwar, and M. Jamshaid, "A comprehensive review of hydrogel-based drug delivery systems: Classification, properties, recent trends, and applications," AAPS PharmSciTech, vol. 25, Art. no. 64, 2024.

[3] Y. Li, H. Y. Yang, and D. S. Lee, "Biodegradable and injectable hydrogels in biomedical applications," Biomacromolecules, vol. 23, no. 3, pp. 609–618, 2022.

[4] S. Jacob, A. B. Nair, J. Shah, N. Sreeharsha, S. Gupta, and P. Shinu, "Emerging role of hydrogels in drug delivery systems, tissue engineering and wound management," Pharmaceutics, vol. 13, no. 3, Art. no. 357, 2021.

[5] A. Guo, Q. Cao, H. Fang, and H. Tian, "Recent advances and challenges of injectable hydrogels in drug delivery," Journal of Controlled Release, vol. 385, Art. no. 114021, 2025.

[6] N. F. S. Oliveira, "Design de formulação: curativo à base de hidrogel contendo extratos naturais de Aloe vera e Allium sativum para o tratamento de feridas diabéticas," Trabalho de Conclusão de Curso (Bacharelado em Farmácia), Universidade Federal de Campina Grande, Cuité, Brasil, 2021.

[7] A.-E. Segneanu, L. E. Bejenaru, C. Bejenaru, A. Blendea, G. D. Mogoșanu, A. Biță, and E. R. Boia, "Advancements in hydrogels: A comprehensive review of natural and synthetic innovations for biomedical applications," Polymers, vol. 17, no. 15, 2025.

[8] F. Zulhendri, K. Chandrasekaran, M. Kowacz, M. Ravalia, K. Kripal, J. Fearnley, and C. O. Perera, "Antiviral, antibacterial, antifungal, and antiparasitic properties of propolis: A review," Foods, vol. 10, no. 6, p. 1360, 2021.

[9] G. dos Anjos, Á. V. de S. Diogo, J. I. A. Bellolio, and F. Tureck, "Hidrogel como veículo terapêutico na regeneração de cardiomiócitos pós infarto agudo do miocárdio: uma revisão integrativa," Revista de Medicina UNC, vol. 4, pp. 27–43, 2025.

[10] Z. Zheng, J. Xie, J. Zeng, Z. Kang, Z. Liu, M. Qi, C. Yu, and H. Wei, "Pathology-guided design of injectable hydrogels for precision therapy and cartilage regeneration in osteoarthritis," Regenerative Biomaterials, 2026.

[11] S. Amirthalingam, A. K. Rajendran, Y. G. Moon, and N. S. Hwang, "Stimuli-responsive dynamic hydrogels: Design, properties and tissue engineering applications," Materials Horizons, vol. 10, pp. 3325–3350, 2023.

[12] M. R. Khatun, A. Bhattacharyya, M. Gunbayar, M. Jung, and I. Noh, "Study on bioresponsive gelatin-hyaluronic acid-genipin hydrogel for high cell-density 3D bioprinting," Gels, vol. 9, Art. no. 601, 2023.

[13] E. A. Pumford, C. D. Hamad, A. I. Enueme, Z. Mamouei, N. Peterson, C. Hart, A. Ishmael, A. Li, R. Sobti, J. Pearce, J. Taylor, M. Ralston, J. D. Wainwright, K. Nakamoto, P. Enkhbaatar, K. A. Cook, K. P. Francis, J. Adams, A. Stavrakis, J. C. Wenke, A. M. Kasko, and N. M. Bernthal, "A field polymerizing hydrogel enables simultaneous

antimicrobial, hemostatic, and analgesic delivery in traumatic wounds," Scientific Reports, vol. 16, p. 6950, 2026.
[14] M. I. Wahba, "A comprehensive review on genipin: an efficient natural cross-linker for biopolymers," Polymer Bulletin, vol. 81, pp. 14251–14305, 2024.
[15] S. Correa, A. K. Grosskopf, H. L. Hernandez, D. Chan, A. C. Yu, L. M. Stapleton, and E. A. Appel, "Translational applications of hydrogels," Chemical Reviews, vol. 121, no. 18, pp. 11385–11457, 2021.
[16] B. Hu, J. Gao, Y. Lu, and Y. Wang, "Applications of degradable hydrogels in novel approaches to disease treatment and new modes of drug delivery," Pharmaceutics, vol. 15, no. 10, Art. no. 2370, 2023.
[17] R. Ahmed, N. U. A. Hira, M. Wang, S. Iqbal, J. Yi, and Y. Hemar, "Genipin, a natural blue colorant precursor: Source, extraction, properties, and applications," Food Chemistry, vol. 434, Art. no. 137498, 2024.
[18] T. Lessmann, S. A. Jones, T. Voigt, S. Weisbrod, O. Kracker, S. Winter, L. A. Zúñiga, S. Stark, N. Bisek, and K. Sprogøe, "Degradable hydrogel for sustained localized delivery of anti-tumor drugs," Journal of Pharmaceutical Sciences, vol. 112, pp. 2843–2852, 2023.
[19] S. Lambrecht, H. Schröter, H. Pohle, and S. Jopp, "Swelling behavior of novel hydrogels produced from glucose-based ionic monomers with varying cross-linkers," ACS Omega, vol. 9, no. 5, pp. 5418–5428, 2024.
[20] A. Sadek, M. M. Amin, and S. Sayed, "Ocular drug delivery: A comprehensive review," AAPS PharmSciTech, vol. 24, Art. no. 66, 2023.
[21] S. Bashir, M. Hina, J. Iqbal, A. H. Rajpar, M. A. Mujtaba, N. A. Alghamdi, S. Wageh, K. Ramesh, and S. Ramesh, "Fundamental concepts of hydrogels: Synthesis, properties, and their applications," Polymers (Basel), vol. 12, no. 11, Art. no. 2702, 2020.
[22] B. Liu and K. Chen, "Advances in hydrogel-based drug delivery systems," Gels, vol. 10, no. 4, Art. no. 262, 2024.
[23] W. C. Ng, Y. Lokanathan, M. B. Fauzi, *et al.*, "In vitro evaluation of genipin-crosslinked gelatin hydrogels for vocal fold injection," Scientific Reports, vol. 13, Art. no. 5128, 2023.
[24] M. J. Penn and M. G. Hennessy, "Optimal loading of hydrogel-based drug-delivery systems," Applied Mathematical Modelling, vol. 112, pp. 649–668, 2022.

Synthesis and Tuning of Highly Methacrylated Hyaluronic Acid Hydrogels for Light- and Extrusion-Based Three-Dimensional Printing

Paloma Saura Tirolti
BioEnT Lab.
Center for Biomedical Engineering
UNICAMP
Campinas, Brazil
p237694@dac.unicamp.br

Jéssica H. L. da Fonseca
BioEnT Lab.
Center for Biomedical Engineering
UNICAMP
Campinas, Brazil
jheline@unicamp.br

Bruna Alice G. de Melo
BioEnT Lab.
Center for Biomedical Engineering
UNICAMP
Campinas, Brazil
brunaagm@unicamp.br

***Abstract*—Methacrylated hyaluronic acid (HAMA) is a photocrosslinkable hydrogel precursor with potential for 3D biofabrication. This study synthesized highly methacrylated HAMA and evaluated its suitability for masked stereolithography (MSLA) and extrusion-based 3D printing. HAMA was chemically validated by proton nuclear magnetic resonance spectroscopy and formulated at one, two, and three percent weight per volume with LAP as photoinitiator. The formulations were characterized by viscosity measurements and wet mass variation after different UV exposure times. HAMA showed a high degree of methacrylation of 80.7 percent. All formulations exhibited shear-thinning behavior, with viscosity increasing with polymer concentration. The 1% HAMA formulation was processed by MSLA, whereas 2 and 3% HAMA formulations were printed by extrusion. Wet mass variation depended on both polymer concentration and UV exposure time, with 30 s producing the most stable hydrogels and 15 s selected as an intermediate photocrosslinking condition for printed constructs. These results define a concentration-dependent processing window for highly methacrylated HAMA hydrogels and support their use as tunable photocrosslinkable precursors for 3D biofabrication.**

***Index Terms*—HAMA, hyaluronic acid, hydrogel, photocrosslinking, biofabrication**

I. Introduction

Hyaluronic acid (HA) is a natural glycosaminoglycan widely used in tissue engineering due to its biocompatibility, hydrophilicity, and role in extracellular matrix organization [1], [2]. However, native HA has limited structural stability and rapid degradation in aqueous environments, which restricts its use as printable hydrogel. Chemical modification is therefore required to improve its stability and enable controlled network formation. Methacrylation introduces photocrosslinkable methacrylate groups into the HA backbone, producing methacrylated hyaluronic acid (HAMA), a hydrogel precursor that can be stabilized by UV crosslinking [3].

This modification preserves its biological properties while enable the formation of hydrogels with tunable physicochemical properties. For three-dimensional printing, these properties must be adjusted according to the fabrication strategy. Light-based vat photopolymerization methods require photocrosslinkable precursors with enough flowability for liquid replenishment and uniform layer formation, whereas extrusion-based printing requires formulations that flow through the nozzle under shear and retain shape after deposition [4]–[7]. Therefore, polymer concentration, rheological behavior, and UV exposure time are fundamental parameters for defining the processing window of HAMA hydrogels. In this study, highly methacrylated HAMA was synthesized and chemically validated by ^{1}H NMR. The HAMA formulations at 1, 2 and 3% w/v were then evaluated in viscosity, UV-dependent wet mass variation and printability using light- and extrusion-based three-dimensional printing.

II. Materials and Methods

A. Synthesis of Methacrylated Hyaluronic Acid

Methacrylated hyaluronic acid (HAMA) was synthesized by reacting hyaluronic acid (HA) with methacrylic anhydride (MA). Briefly, HA was dissolved in phosphate-buffered saline (PBS) at 2% w/v under magnetic stirring overnight. After dissolution, MA was added to the HA solution using a peristaltic pump at a flow rate of 0.5 mL/h, under continuous stirring. The reaction was maintained for 3 h at room temperature, with pH kept around 7.5 using NaOH. The product was purified by dialysis against deionized water using a 12-14 kDa membrane for 7 days, lyophilized, and stored at 4 °C until use.

B. Chemical Characterization of HAMA

HA and HAMA were analyzed by proton nuclear magnetic resonance spectroscopy (^{1}H NMR). Samples were dissolved in D_2O containing TMSP/TSP as the chemical shift reference. After solubilization, 600 μL of each solution was transferred to NMR tubes. Spectra were acquired in a 600 MHz NMR spectrometer at 25 °C, with 128 scans and a relaxation delay of 3 s. The degree of methacrylation (DM) was calculated from the ratio between the methacrylate vinyl proton signals at 6.04–6.33 and 5.60–5.91 ppm and the N-acetyl methyl protons of HA at approximately 2.03 ppm, correcting for the number of protons contributing to each signal.

C. Preparation of HAMA Hydrogel Formulations

HAMA precursor solutions were prepared at 1, 2, and 3% w/v by dissolving lyophilized HAMA in PBS under magnetic stirring at room temperature. Lithium phenyl-2,4,6-trimethylbenzoylphosphinate (LAP) was added at 0.1% w/v as photoinitiator. All formulations were protected from light until use.

D. Rheological Characterization

The rheological behavior of HAMA formulations was evaluated using a rotational rheometer equipped with a 50 mm parallel-plate geometry and a 0.5 mm gap. Measurements were performed at 25 °C in triplicate. Apparent viscosity was measured as a function of shear rate, from 0.01 to 1000 s^{-1}.

E. Crosslinking Screening and Swelling Assay

To screen photocrosslinking conditions, 1, 2, and 3% HAMA formulations containing 0.1% w/v LAP were cast in PDMS molds and exposed to 405 nm light for 10, 15, and 30 s.

After photocrosslinking, the initial mass of each hydrogel was recorded, and the samples were immersed in PBS at 37 °C. At each time point, hydrogels were removed from PBS, gently blotted to remove excess of liquid, and weighed again. The wet mass variation was calculated as:

$$\text{Wet mass variation } (\%) = \frac{W_t - W_0}{W_0} \times 100$$

where W_0 is the initial hydrogel mass after photocrosslinking and W_t is the hydrogel mass at each time point.

F. MSLA 3D Printing

MSLA printing was performed using a Creality Halot R6 printer based on MSLA/LCD technology, equipped with a 405 nm UV light source and nominal intensity around 5000 $\mu\mathrm{W}/\mathrm{cm}^2$. The printing setup was adapted to reduce precursor volume. The 1% HAMA formulation was printed using geometries designed in CAD and sliced in CHITUBOX. After preliminary tests, printing was performed with 0.04 mm layer height, 10 bottom layers exposed for 60 s, and 10 s exposure for the subsequent layers.

G. Extrusion-Based 3D Printing

Extrusion-based printing was performed using 2 and 3% HAMA formulations at a 593iCan Printer. The formulations were loaded into 10 mL syringes and extruded through a 20G conical nozzle. Cylindrical constructs with a 10 mm diameter were printed with 5 layers and 0.4 mm layer height. After printing, the constructs were photocrosslinked under 405 nm light for 15 s and immersed in PBS until further analysis.

III. RESULTS AND DISCUSSION

A. Chemical Validation and Methacrylation Degree of HAMA

The successful methacrylation of HA was confirmed by ^{1}H NMR (Fig. 1). HAMA exhibited characteristic methacrylate vinyl proton peaks at 6.04–6.33 and 5.60–5.91 ppm. The N-acetyl methyl signal of HA at approximately 2.03 ppm was used as a reference for calculating the degree of methacrylation (DM). The resulting HAMA exhibited a DM of 80.7%, indicating a high level of functionalization and confirming its suitability as a hydrogel precursor for photocrosslinking.

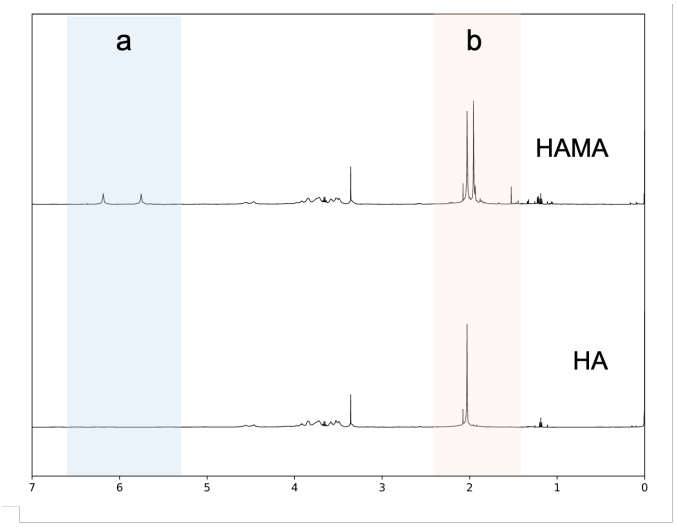


Fig. 1. ^{1}HNMR spectra of HA and HAMA. (a) HAMA peaks at approximately 6.04–6.33 and 5.60–5.91 ppm, absent in native HA. The methyl region (b), including the N-acetyl methyl protons of HA around 2.03 ppm

B. Rheological and Swelling Behavior of HAMA Formulations

The rheological behavior of HAMA formulations was concentration-dependent (Fig. 2). All formulations exhibited shear-thinning behavior, with viscosity decreasing as shear rate increased. This behavior is favorable for 3D printing because different fabrication strategies impose distinct flow requirements. In vat photopolymerization methods such as MSLA, low-viscosity photocrosslinkable precursors favor liquid replenishment and uniform recoating between layers, which are required for consistent layer formation [4], [5]. In extrusion-based printing, shear-thinning behavior facilitates flow through the nozzle under high shear, whereas higher viscosity at low shear contributes to filament stability and shape retention after deposition [6]–[8].

Among the tested formulations, 1% HAMA showed the lowest viscosity, indicating a formulation more compatible with light-based processing. The 2% HAMA formulation showed an intermediate behavior, suggesting a balance between flowability and post-deposition stability. In contrast, 3% HAMA showed the highest viscosity across the entire shear-rate range, indicating greater resistance to flow and improved potential for shape retention. Thus, the rheological profile provided the basis for defining the subsequent printing strategy, with 1% HAMA evaluated by MSLA and 2–3% HAMA evaluated by extrusion-based 3D printing.

The wet mass variation of HAMA hydrogels depended on both polymer concentration and UV exposure time (Fig. 3). For 1% HAMA, 10 s of UV exposure was insufficient to maintain hydrogel integrity after immersion in PBS, and samples started to lose structural integrity after approximately 30 min

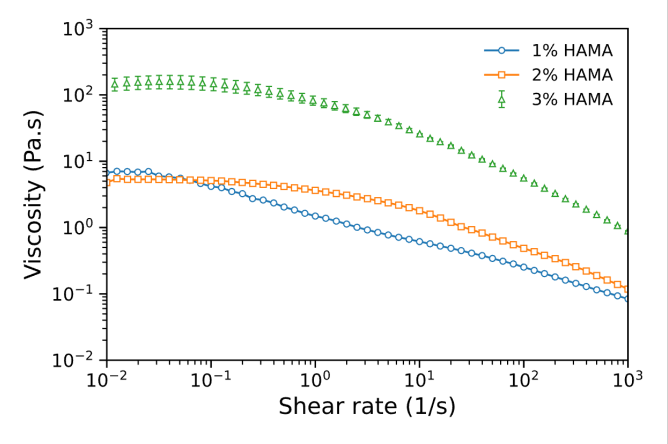


Fig. 2. Viscosity of 1, 2, and 3% HAMA formulations as a function of shear rate.

(Fig. 3A). In contrast, 15 and 30 s allowed the hydrogels to remain intact during the 120 min assay, although wet mass changes were still observed.

For 2% HAMA, hydrogels crosslinked for 10 s showed the highest wet mass increase, indicating insufficient network stabilization (Fig. 3B). Increasing UV exposure reduced wet mass variation, with 30 s producing a more stable profile. A similar trend was observed for 3% HAMA, but with lower overall wet mass variation (Fig. 3C). Notably, 3% HAMA crosslinked for 30 s remained nearly stable during the assay.

The long-term endpoint also confirmed the stabilizing effect of longer UV exposure (Fig. 3D). After 4 days, hydrogels crosslinked for 30 s showed the lowest wet mass variation across concentrations, with 3% HAMA remaining close to its initial mass. In contrast, 10 and 15 s exposures resulted in higher mass uptake for 2% and 3% HAMA, while 1% HAMA crosslinked for 15 s showed mass loss after prolonged incubation.

Overall, these results indicate that longer UV exposure increased hydrogel network stabilization, while higher HAMA concentration further improved wet mass stability. Although 30 s produced the lowest wet mass variation, 15 s was used as an intermediate photocrosslinking condition for the extrusion-printing experiments.

C. MSLA and Extrusion-Based 3D Printing of HAMA

Based on its lower viscosity and suitability for light-based processing, the 1% HAMA formulation was selected for mSLA printing using the adapted low-volume setup (Fig. 4a–c). Square and heart-shaped constructs were printed with 0.04 mm layer height, 60 s exposure for the bottom layers, and 10 s exposure for the subsequent layers. These geometries were used to evaluate shape formation and dimensional stability after incubation in PBS. Immediately after printing, the constructs preserved their overall geometry, with width and length values close to the nominal CAD dimensions (Fig. 4d,e). In contrast, height fidelity was already reduced immediately after printing, indicating lower dimensional accuracy in the vertical direction. After 7 d in PBS, this reduction became more pronounced, while width and length remained relatively stable (Fig. 4d,f). The lower height fidelity immediately after printing indicates a limitation in vertical dimensional accuracy under the selected printing conditions, while the additional reduction during incubation indicates limited vertical stability under hydrated conditions. The low polymer concentration may also contribute to this behavior by forming a polymer network with lower resistance to deformation under hydrated conditions. Thus, 1% HAMA can be processed by mSLA, although its vertical dimensional stability remains limited. Optimization of exposure time and layer height may improve vertical printing accuracy, while post-printing photocrosslinking may enhance dimensional stability under hydrated conditions.

Based on the rheological profile and crosslinking screening, 2% and 3% HAMA formulations were selected for extrusion-based printing. Both formulations were extruded through a 20G conical nozzle and photocrosslinked for 15 s after printing (Fig. 4g,h). The printed constructs maintained their circular geometry, indicating that the formulations retained their shape after deposition (Fig. 4i,j). However, filament fusion and collapse were observed, indicating limited printing fidelity and suggesting that post-printing photocrosslinking alone was not sufficient to fully preserve the designed pore architecture. Although the higher HAMA concentrations increased viscosity and allowed the printed constructs to maintain their overall circular geometry, the deposited filaments remained susceptible to spreading before complete network stabilization. Thus, under the tested conditions, these formulations may be more suitable for deposition strategies in which maintaining the overall shape is more important than internal architecture. Therefore, extrusion fidelity may be further improved by optimizing printing parameters, such as flow rate, printing speed, and photocrosslinking strategy. Layer-by-layer or in situ photocrosslinking during extrusion could improve filament stabilization by reducing spreading and improving pore definition.

Together, these results show that HAMA processability is strongly dependent on polymer concentration and photocrosslinking conditions. The low-viscosity 1% formulation was suitable for mSLA processing, enabling the fabrication of constructs with good shape fidelity. In contrast, the higher-viscosity 2% and 3% formulations could be processed by extrusion and maintained the overall construct geometry, although with limited filament and pore fidelity. Therefore, HAMA formulations can be tuned to match different three-dimensional printing strategies by adjusting polymer concentration, UV exposure time, and printing parameters. This formulation flexibility supports the use of highly methacrylated HAMA as a photocrosslinkable hydrogel precursor for future biofabrication applications.

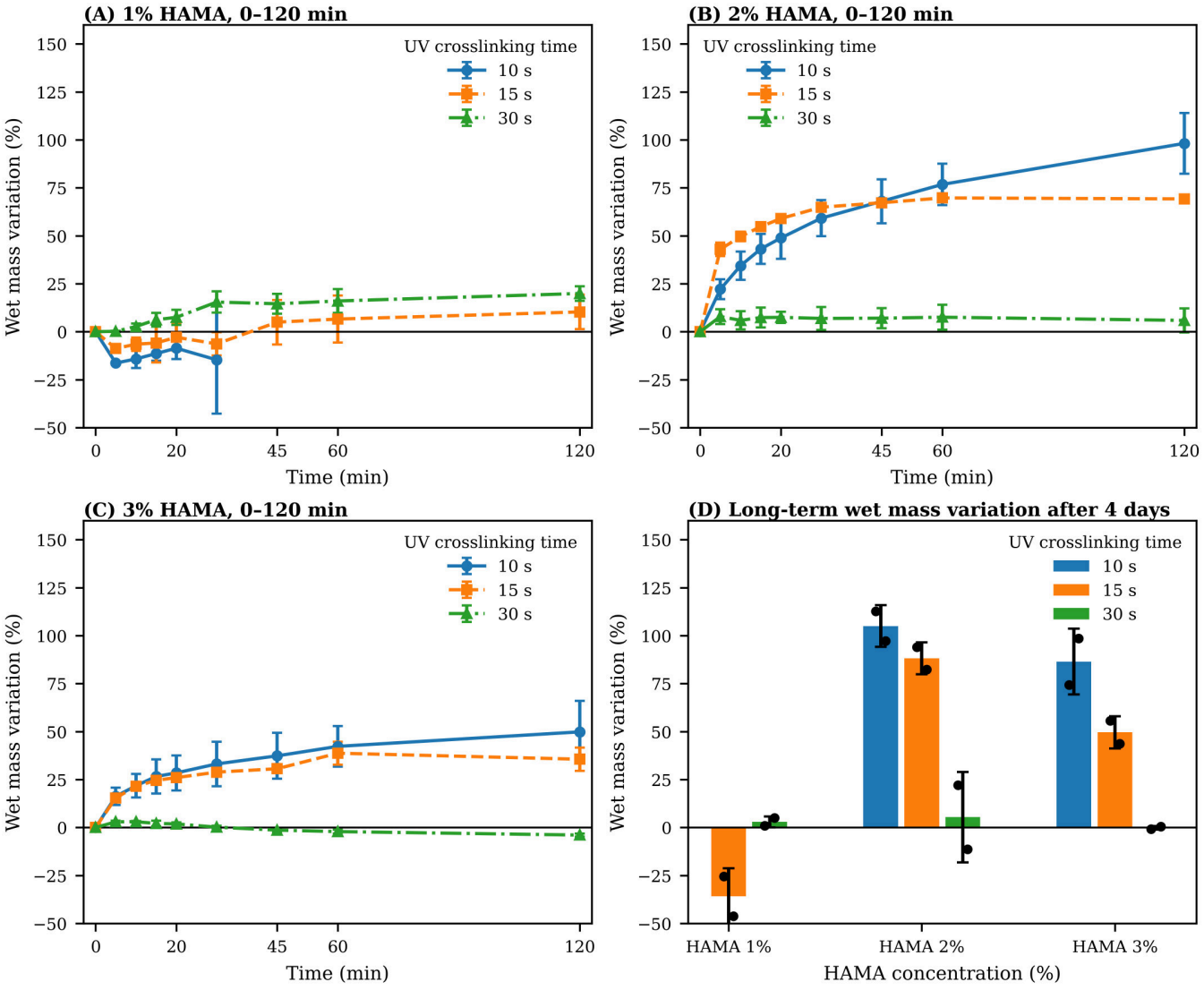


Fig. 3. Wet mass variation of HAMA hydrogels photocrosslinked for 10, 15, or 30 s and incubated in PBS at 37 °C. Time-course profiles are shown for (A) 1% HAMA, (B) 2% HAMA, and (C) 3% HAMA during the first 120 min. Panel (D) shows long-term wet mass variation after 4 d.

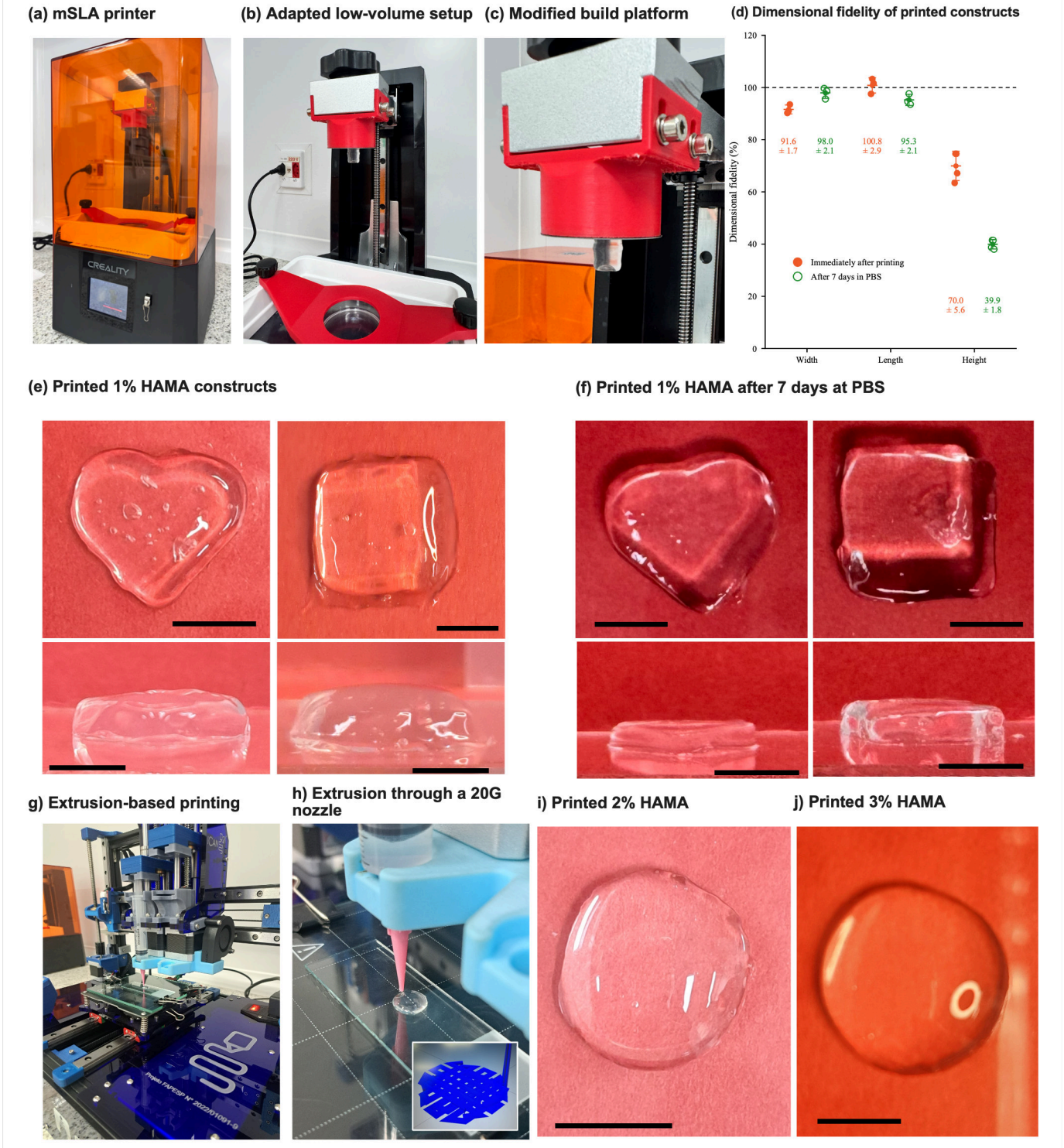


Fig. 4. Printing results of HAMA using light-based and extrusion-based approaches. (a) Creality Halot R6 MSLA printer. (b) Adapted low-volume setup used to reduce precursor volume during printing. (c) Modified build platform. (d) Dimensional fidelity of MSLA-printed 1% HAMA constructs immediately after printing and after 7 d in PBS. (e) MSLA-printed 1% HAMA constructs immediately after printing. (f) 1% HAMA constructs after 7 d in PBS. (g) Extrusion-based printing setup. (h) Deposition of HAMA through a 20G conical nozzle. (i,j) Representative extrusion-printed HAMA constructs obtained with 2% and 3% HAMA formulations after 15 s photocrosslinking. Scale bars: 5 mm.

Conclusion

Highly methacrylated HAMA was successfully synthesized and chemically validated by ^{1}H NMR, showing a DM of 80.7%. The hydrogel formulations exhibited concentration-dependent rheological behavior, with shear-thinning profiles and increased viscosity at higher polymer concentrations. UV exposure time also modulated wet mass stability, indicating that photocrosslinking conditions are key parameters for hydrogel network stabilization.

The 1% HAMA formulation was suitable for MSLA printing and enabled the fabrication of planar constructs, although height fidelity decreased after incubation in PBS. In contrast, 2% and 3% HAMA formulations were processed by extrusion-based printing and maintained the overall circular construct geometry, although filament spreading and loss of pore definition were observed. Overall, these findings demonstrate that highly methacrylated HAMA can be tuned by polymer concentration, UV exposure time, and printing strategy, supporting its use as a photocrosslinkable hydrogel precursor for future three-dimensional biofabrication applications.

Acknowledgment

This study was supported by the Fund for the Support of Teaching, Research, and Extension (FAEPEX), under grant numbers 3434/24 and 2792/25. The authors also acknowledge the support of Diretoria Acadêmica de Apoio à Permanência Estudantil (DEAPE) of Unicamp.

References

[1] J. Lam, N. F. Truong, and T. Segura, "Design of cell–matrix interactions in hyaluronic acid hydrogel scaffolds," *Acta Biomaterialia*, vol. 10, no. 4, pp. 1571–1580, 2014, doi: 10.1016/j.actbio.2013.07.025.

[2] Y.-W. Ding, Z.-Y. Zhang, J. Ren, and H. Zhang, "Recent advances in hyaluronic acid-based hydrogels for 3D bioprinting in tissue engineering applications," *Smart Materials in Medicine*, vol. 4, pp. 59–68, 2023, doi: 10.1016/j.smaim.2022.07.003.

[3] N. Shen, Z. Li, P. Yang, X. Liu, Y. Ju, Y. Hu, B. Fang, and L. Liu, "Designing methacrylic anhydride-based hydrogels for 3D bioprinting," *International Journal of Bioprinting*, vol. 11, no. 1, pp. 84–138, 2025, doi: 10.36922/ijb.4650.

[4] S. Subedi, S. Liu, W. Wang, S. M. A. N. Shovon, X. Chen, and H. O. T. Ware, "Multi-material vat photopolymerization 3D printing: A review of mechanisms and applications," *npj Advanced Manufacturing*, vol. 1, Art. no. 9, 2024, doi: 10.1038/s44334-024-00005-w.

[5] L. Šimunović, M. Šantić, D. Španić, M. Škorić, and I. Šariri, "Rheology of dental photopolymers for SLA/DLP/MSLA 3D printing," *Polymers*, vol. 17, no. 19, Art. no. 2706, 2025, doi: 10.3390/polym17192706.

[6] M. E. Cooke and D. H. Rosenzweig, "The rheology of direct and suspended extrusion bioprinting," *APL Bioengineering*, vol. 5, no. 1, Art. no. 011502, 2021, doi: 10.1063/5.0031475.

[7] A. Schwab, R. Levato, M. D'Este, S. Piluso, D. Eglin, and J. Malda, "Printability and shape fidelity of bioinks in 3D bioprinting," *Chemical Reviews*, vol. 120, no. 19, pp. 11028–11055, 2020, doi: 10.1021/acs.chemrev.0c00084.

[8] M. I. Calafel, M. Criado-Gonzalez, R. Aguirresarobe, M. Fernández, and C. Mijangos, "From rheological concepts to additive manufacturing assessment of hydrogel-based materials for advanced bioprinting applications," *Materials Advances*, vol. 6, no. 14, pp. 4566–4597, 2025, doi: 10.1039/d5ma00019j.

Engineering Next-Generation Theranostic Bioceramics: A Translational Framework Integrating Rational Ion Co-Doping and Biomedical Devices

1st José da Silva Rabelo Neto
Graduate Program in Health Management and Technology Innovation (PPGITS)
Federal University of Sergipe (UFS)
São Cristovão - SE, Brazil

Centre for Mechanical Technology and Automation (TEMA)
University of Aveiro
Aveiro, Portugal

RenovCure Biotech (I.S.)
São Cristovão - SE, Brazil
ORCID 0000-0002-9769-5675

2nd Patrícia Aparecida da Ana
Center for Engineering, Modeling and Applied Social Sciences (CECS)
Federal University of ABC (UFABC)
Santo André - SP, Brazil
ORCID 0000-0003-2857-7517

3rd Antonio Martins de Oliveira Junior
Graduate Program in Chemical Engineering (PEQ)
Federal University of Sergipe (UFS)
São Cristovão - SE, Brazil

Graduate Program in Intellectual Property Science (PPGPI)
Federal University of Sergipe (UFS)
São Cristovão - SE, Brazil
ORCID 0000-0002-8635-7048

***Abstract*—Multifunctional theranostic bioceramics have emerged as promising platforms for integrating diagnosis, therapy, tissue regeneration, and biomedical sensing within a single material system. Recent advances in rational ion co-doping have significantly expanded the functional capabilities of calcium phosphate bioceramics by incorporating magnetic, optical, angiogenic, antibacterial, and osteogenic properties. However, despite substantial progress in material design, current research remains fragmented across materials science, biology, photonics, and biomedical engineering, limiting the clinical translation of these technologies. This paper presents a comprehensive overview of recent developments in multifunctional ion-engineered bioceramics and identifies the major scientific and technological challenges hindering their translational advancement. Based on these research gaps, we propose the Translational Engineering of Bioceramics (TEB) Framework, a literature-informed conceptual methodology that systematically integrates clinical need identification, rational ion co-doping design, material engineering, multimodal characterization, computational and AI-assisted candidate prioritization, biological validation, prototype development, and translational planning. Rather than providing only a conventional literature review, this work introduces a systems-oriented engineering perspective aimed at structuring the prospective development and evaluation of next-generation theranostic biomaterials. The proposed framework provides a conceptual roadmap for investigating the coordinated development of multifunctional bioceramics toward biomedical device applications, while requiring prospective experimental validation to establish its practical utility.**

***Index Terms*—Theranostic, calcium phosphate, hydroxyapatite, ion co-doping, translational engineering, biomedical devices, biomaterials.**

I. Introduction

Biomedical Engineering is undergoing a profound transformation driven by the convergence of advanced biomaterials, nanotechnology, artificial intelligence, biophotonics, molecular imaging, and precision medicine. Rather than developing isolated therapeutic materials, current research increasingly seeks multifunctional biomedical platforms capable of simultaneously diagnosing, monitoring, and treating diseases while promoting tissue regeneration. This paradigm has accelerated the development of theranostic systems, which integrate diagnostic and therapeutic functions into a single biomaterial platform, enabling more personalized, minimally invasive, and clinically effective interventions [1]–[4].

Among the various classes of biomaterials investigated for theranostic applications, calcium phosphate bioceramics, particularly hydroxyapatite, have attracted considerable attention because of their excellent biocompatibility, bioactivity, osteoconductivity, biodegradability, and chemical similarity to natural bone. Beyond their traditional use as bone substitutes, recent advances in ionic substitution and co-doping strategies have expanded their functionality, allowing the incorporation of magnetic, optical, catalytic, antimicrobial, angiogenic, and osteogenic properties into a single material platform. These multifunctional characteristics position ion-engineered bioceramics as promising candidates for next-generation biomedical devices [5]–[9].

In parallel, photonic and magnetic therapeutic approaches, including Magnetic Fluid Hyperthermia (MFH), Photothermal Therapy (PTT), and Photodynamic Therapy (PDT), have

evolved from experimental concepts into clinically relevant strategies for cancer treatment and image-guided interventions. The combination of these modalities with multifunctional nanomaterials has demonstrated remarkable potential, particularly through synergistic PDT/PTT nanoplatforms and imaging-guided phototherapies, to improve therapeutic selectivity, reduce systemic toxicity, and enable real-time monitoring of treatment response. More recently, synergistic combinations of PDT and PTT have emerged as particularly promising due to their complementary mechanisms of action and compatibility with multimodal nanoplatforms [2], [3], [10]–[12].

Despite these remarkable advances, current research remains highly fragmented. Most studies focus on isolated aspects of material development, such as individual ionic substitutions, specific synthesis methods, single therapeutic modalities, or individual biological evaluations. Consequently, the field still lacks an integrated engineering strategy capable of systematically connecting rational material design, multifunctionality, advanced characterization, biological validation, artificial intelligence, prototype development, and clinical translation. This fragmentation may contribute to the limited progression of multifunctional theranostic bioceramics toward higher technological readiness levels and broader clinical evaluation.

In this context, this paper proposes a Translational Engineering Framework for next-generation multifunctional theranostic bioceramics. Rather than providing only a conventional literature review, the proposed framework conceptually organizes rational ion co-doping strategies, advanced synthesis routes, multimodal characterization, computational and AI-assisted candidate prioritization, biological validation, prototype development, and translational planning.

By organizing current knowledge into a unified engineering perspective, this work aims to provide researchers with a conceptual roadmap whose practical utility can be prospectively evaluated across the development of multifunctional biomaterials for regenerative medicine, biosensing, antimicrobial therapies, and precision medicine.

II. Literature Selection and Conceptual Framework Development

A structured narrative review was conducted to support the identification of scientific and translational gaps and the conceptual development of the Translational Engineering of Bioceramics (TEB) Framework. The review was designed as a framework-oriented synthesis rather than a systematic review or meta-analysis. Its purpose was to integrate evidence across materials science, biomedical engineering, theranostics, and computational materials design and to identify recurring challenges throughout the development pathway of multifunctional calcium phosphate bioceramics.

The literature search focused primarily on studies published between 2019 and 2026. Search terms were organized into three complementary thematic domains: ion-engineered calcium phosphate bioceramics, multifunctional/theranostic bioceramics, and computational or translational approaches to biomaterials development. An exploratory search using ("hydroxyapatite" OR "calcium phosphate") AND ("ion doping" OR "ionic substitution" OR "co-doping") retrieved 338 records. Restriction to records indexed with doping-related keywords reduced this set to 161, of which 128 were classified within Materials Science and 15 within Medicine. When "theranostic*" was added to this search, only two records were retrieved, both published in 2024.

A broader search combining ("hydroxyapatite" OR "bioceramic") with ("theranostic" OR "multifunctional" OR "magnetic" OR "photothermal" OR "photodynamic" OR "biosensing") retrieved 3,083 records; restricting this search using the same doping-related keywords reduced the set to 60 records. A third search combining ("biomaterials" OR "bioceramics") with ("artificial intelligence" OR "machine learning" OR "computational materials design" OR "translational engineering") retrieved 1,636 records, whereas application of the doping-related keyword restriction resulted in only one record, classified within Materials Science.

These exploratory counts indicate that, although substantial bodies of literature exist independently on ion-engineered bioceramics, multifunctionality, and computational biomaterials research, their intersection remains comparatively underrepresented. The predominance of Materials Science classifications in the ion-doping search further suggests a disciplinary concentration that supports examining these technologies through a broader biomedical translational perspective.

Studies were considered for the narrative synthesis when they contributed evidence concerning ion–structure–property relationships, multifunctional or theranostic performance, biological responses, multimodal characterization, computational or AI-assisted materials design, biomedical device integration, or translational barriers. Review articles were used to contextualize consolidated developments and identify complementary primary studies. The selected literature was synthesized thematically rather than quantitatively.

Recurring scientific, engineering, and translational challenges were mapped across the biomaterial development pathway and organized into eight interconnected domains: clinical need identification, rational ion co-doping design, material engineering, multimodal characterization, computational and AI-assisted optimization, biological validation, prototype development, and translational planning.

These domains informed the proposed stages and evaluation targets of the TEB Framework. Accordingly, the TEB Framework should be interpreted as a literature-informed, hypothesis-generating conceptual model rather than an experimentally validated translational pathway; its practical utility and performance require prospective experimental and real-world evaluation.

III. Current Landscape and Research Gaps

Calcium phosphate bioceramics have long been recognized as one of the most successful classes of biomaterials for

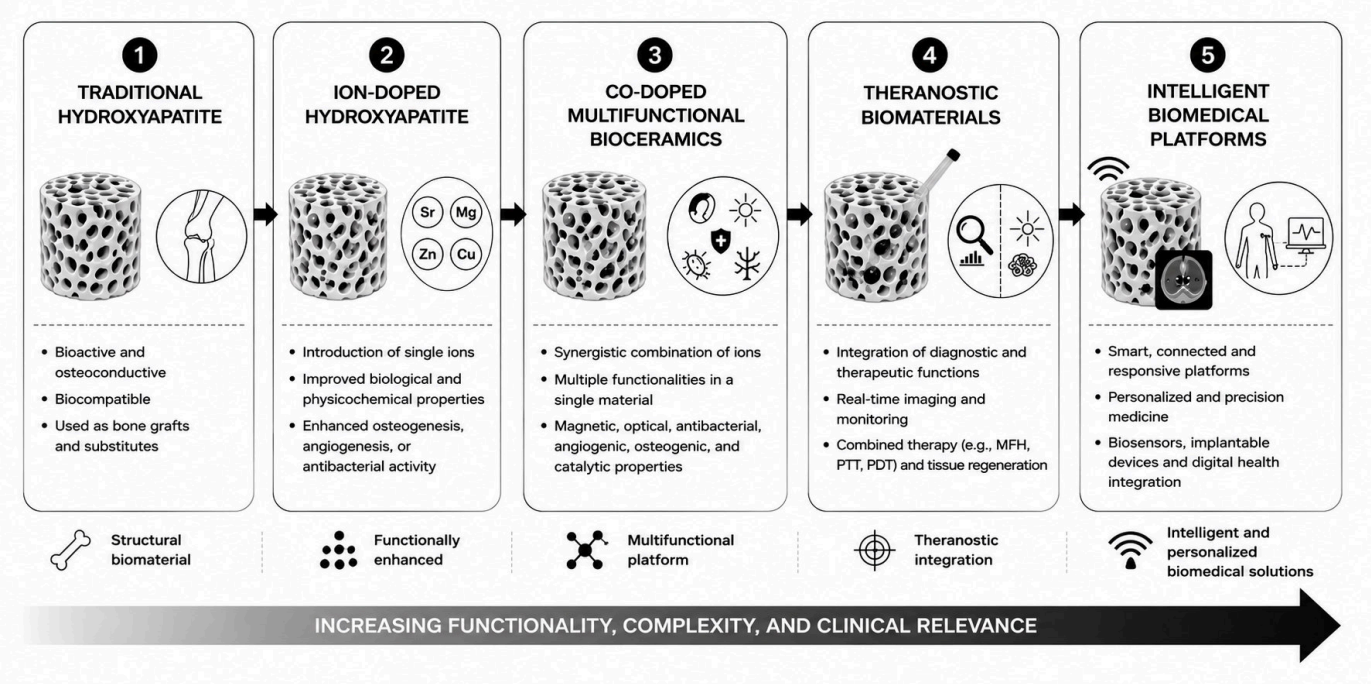


Fig. 1. Evolution of multifunctional bioceramics from traditional hydroxyapatite to intelligent biomedical platforms. *Conceptual illustration generated using generative AI and edited by the authors.*

orthopedic and dental applications due to their excellent biocompatibility, osteoconductivity, biodegradability, and chemical similarity to the mineral phase of human bone. Among them, hydroxyapatite (HAp) has become the most extensively investigated material because of its ability to support bone regeneration [13]–[17].

During the last two decades, however, the role of hydroxyapatite has expanded substantially. Rather than acting exclusively as passive scaffolds for tissue repair, calcium phosphate bioceramics have progressively evolved into multifunctional platforms capable of simultaneously interacting with biological tissues, responding to external stimuli, carrying therapeutic agents, and generating diagnostic signals. This transformation reflects the convergence of materials science, nanotechnology, photonics, and biomedical engineering toward the development of theranostic biomaterials, see figure 1.

One of the major driving forces behind this evolution has been the development of rational ionic substitution strategies, which enables simultaneous tuning of structural, mechanical, dielectric, surface, magnetic, and biological properties. The apatite crystal lattice presents remarkable structural flexibility, allowing calcium ions to be partially replaced by biologically active metallic ions without compromising the fundamental crystal structure [1], [17], [18]. This unique characteristic enables precise modulation of physicochemical, magnetic, optical, biological, and mechanical properties according to specific biomedical requirements, see table I.

Initially, ionic substitution focused primarily on improving osteoconductivity and mechanical performance. More recently, co-doping approaches have emerged as a powerful strategy for simultaneously introducing multiple biological and physical functionalities into a single biomaterial. Unlike single-ion substitution, co-doping enables synergistic interactions among different ions, allowing the simultaneous regulation of osteogenesis, angiogenesis, antibacterial activity, magnetic responsiveness, fluorescence, reactive oxygen species (ROS) generation, and imaging capabilities [19]–[26].

The selection of dopant ions is no longer exclusively based on chemical compatibility but increasingly follows a function-oriented engineering approach. Each ionic species contributes distinct biological or physical properties that can be strategically combined according to the intended clinical application [7], [9], [27]–[33].

Strontium and magnesium are widely recognized for stimulating osteoblast proliferation, bone mineralization, and inhibition of osteoclast activity, making them attractive for regenerative applications [34]–[36]. Cobalt and manganese have additionally demonstrated important roles in regulating angiogenesis, osteogenesis, and vascular remodeling. Cobalt promotes angiogenesis through hypoxia-mimicking mechanisms while simultaneously introducing magnetic properties useful for hyperthermia [37]–[40]. Iron provides superparamagnetic behavior suitable for magnetic fluid hyperthermia and magnetic resonance imaging [41]–[44]. Manganese contributes both to magnetic imaging and reactive oxygen species generation, whereas chromium may enhance luminescence and photonic responses. Other ions, including zinc, copper, silver, gadolinium, and erbium, further expand the multifunctionality of calcium phosphates by incorporating antimicrobial, imaging, sensing, and optical capabilities [45]–[49].

TABLE I
:REPRESENTATIVE IONIC DOPANTS AND THEIR MULTIFUNCTIONAL ENGINEERING ROLES IN NEXT-GENERATION THERANOSTIC BIOCERAMICS.

Functional Objective	**Representative Ions**	**Primary Mechanism**	**Engineering Functionality**	**Biomedical Applications**
Osteogenesis	Sr, Mg, Zn	Stimulation of osteoblast differentiation, mineralization, and inhibition of osteoclast activity	Bioactive scaffolds with enhanced bone regeneration	Bone regeneration, orthopedic implants, dental implants, tissue engineering
Angiogenesis	Co, Sr	Hypoxia-mimicking pathways (HIF-1α/VEGF) and vascular stimulation	Vascularized biomaterials and regenerative scaffolds	Bone repair, critical-size defects, regenerative medicine
Magnetic Response	Fe, Co	Superparamagnetism and magnetic susceptibility	Magnetic Fluid Hyperthermia (MFH), magnetic targeting and MRI enhancement	Cancer therapy, molecular imaging, image-guided interventions
Reactive Oxygen Species (ROS) Generation	Mn, Cu, Cr	Catalytic ROS generation under external stimulation	Photodynamic Therapy (PDT) and oxidative therapies	Cancer treatment, antimicrobial therapy, infection control
Photonic / Luminescent Response	Er, Cr	Luminescence and fluorescence enhancement	Optical probes and photonic biomaterials	Optical imaging, fluorescence-guided surgery, biosensing
Magnetic Resonance Imaging (MRI)	Gd, Fe, Mn	Paramagnetic contrast enhancement	Multimodal imaging platforms	MRI-guided diagnosis, precision medicine, theranostics
Antibacterial Activity	Ag, Cu, Zn	Membrane disruption and bacterial metabolic inhibition	Antimicrobial coatings and bioactive implant surfaces	Implant-associated infection prevention, wound healing
Biosensing	Er, Gd, Sr	Surface charge modulation, fluorescence and electrochemical signal enhancement	Smart biosensors and bioelectronic interfaces	Molecular diagnostics, wearable biosensors, precision medicine

The concept of theranostics has profoundly reshaped the development of multifunctional biomaterials by integrating diagnostic and therapeutic capabilities into single platforms. Instead of employing independent systems for imaging, treatment, and tissue regeneration, current engineering approaches increasingly seek multifunctional materials capable of performing these functions simultaneously.

This convergence has stimulated the integration of magnetic fluid hyperthermia, photothermal therapy, photodynamic therapy, molecular imaging, biosensing, controlled drug delivery, and tissue regeneration into unified biomaterial systems. Such multifunctionality enables image-guided therapies, real-time treatment monitoring, and personalized therapeutic strategies while reducing invasiveness and improving clinical precision.

Beyond material development itself, recent advances indicate a progressive transition toward engineering-oriented design strategies. Artificial intelligence, computational materials science, high-throughput screening, digital manufacturing, additive manufacturing, and multimodal characterization are beginning to transform how multifunctional bioceramics are conceived, optimized, and translated into biomedical devices. Nevertheless, despite these remarkable technological advances, current research remains fragmented across materials science, chemistry, biology, photonics, and biomedical engineering.

The absence of an integrated translational strategy capable of connecting these individual advances into a unified development pipeline represents one of the major challenges currently facing the field. Addressing this gap requires a comprehensive engineering framework capable of guiding the rational design of next-generation theranostic bioceramics from fundamental material development to clinical implementation.

The TEB Framework was also informed by the authors' technology-development experience. Between 2010 and 2018, Innovacura Biomateriais developed advanced biomaterials through a predominantly materials-science team, revealing the need to integrate clinical, regulatory, manufacturing, and business requirements earlier and iteratively. These lessons, together with the literature gaps, informed the TEB Framework, currently being prospectively applied and evaluated within RenovCure. This ongoing implementation does not yet constitute empirical validation.

IV. A TRANSLATIONAL ENGINEERING FRAMEWORK FOR NEXT-GENERATION THERANOSTIC BIOCERAMICS

Based on the scientific and translational gaps identified in the literature, we propose the Translational Engineering of Bioceramics (TEB) Framework as a literature-informed conceptual model for organizing the development and prospective evaluation of multifunctional theranostic bioceramics. Rather than representing a validated translational pathway, the framework defines a sequence of interconnected engineering stages and corresponding evaluation targets that can be tested in future experimental and real-world studies.

The proposed framework comprises eight iterative stages: (1) clinical need identification, (2) rational ion co-doping design, (3) material engineering, (4) multimodal characterization,

TABLE II
CURRENT RESEARCH GAPS AND PROPOSED ENGINEERING DIRECTIONS FOR NEXT-GENERATION THERANOSTIC BIOCERAMICS.

Current Limitation	**Scientific or Technological Consequence**	**Proposed Engineering Direction**
Empirical ion selection	Time-consuming optimization and limited understanding of synergistic effects	Artificial intelligence-assisted rational ion selection and computational materials design
Single-function biomaterials	Limited therapeutic performance and reduced clinical applicability	Development of multifunctional theranostic platforms integrating diagnosis, therapy, and tissue regeneration
Fragmented material characterization	Incomplete understanding of structure–property–function relationships	Multimodal characterization combining structural, magnetic, optical, chemical, and biological analyses
Isolated biological evaluation	Poor correlation between physicochemical properties and biological performance	Integrated in vitro, in vivo, and mechanistic biological validation
Laboratory-scale synthesis	Limited reproducibility and difficulty in industrial scale-up	Standardized, scalable, and reproducible manufacturing protocols
Limited prototype development	Low Technology Readiness Level (TRL) and poor translational potential	Engineering-oriented prototyping and biomedical device development
Lack of regulatory and clinical planning	Delayed clinical translation and commercialization	Early integration of regulatory science, clinical validation, and translational engineering strategies

(5) computational and AI-assisted candidate prioritization, (6) biological validation, (7) prototype development, and (8) translational planning and evaluation. Each stage is associated with prospective evaluation targets intended to determine whether a candidate material should progress, be reformulated, or return to an earlier development stage. Accordingly, progression through the framework should not be interpreted as evidence of increasing technological maturity unless predefined stage-specific criteria are prospectively demonstrated.

Table III summarizes the proposed stages, their engineering objectives, and the corresponding targets for future evaluation. The framework is intended to operate iteratively rather than as a strictly linear pipeline, allowing experimental, biological, computational, manufacturing, or regulatory findings to inform previous stages and guide subsequent design decisions.

Despite advances in multifunctional bioceramics, their development remains fragmented across materials science, biology, clinical sciences, and biomedical engineering. Materials research commonly emphasizes synthesis, crystal chemistry, and physicochemical characterization, whereas biological and clinical studies focus on biocompatibility, therapeutic performance, and safety. Limited integration among these domains may contribute to promising biomaterials remaining at the experimental stage rather than progressing toward biomedical technologies.

Ion co-doping further increases this translational complexity because interactions among dopants may affect crystal chemistry, phase stability, degradation, physicochemical properties, and biological performance in nonlinear and potentially synergistic or antagonistic ways. Since most studies investigate a limited number of compositions, candidate selection still frequently relies on empirical trial-and-error approaches. Beyond material design, reproducibility, scalable manufacturing, long-term biosafety, regulatory requirements, prototype development, and clinical evaluation represent additional translational barriers.

These interconnected challenges support addressing material design, biological validation, manufacturing, and translational requirements iteratively rather than as isolated downstream activities. Within the proposed TEB Framework, progression is therefore conditional on predefined evaluation criteria, allowing evidence generated at each stage to support progression, refinement, or return to earlier design decisions. This stage-gate logic provides a testable structure for prospectively evaluating the framework's practical utility.

A. Computational and AI-Assisted Ion Co-Doping Optimization

Within the TEB Framework, computational and AI methods are conceived as decision-support tools for prioritizing candidate ion combinations and experimental conditions rather than autonomous material-selection mechanisms. The proposed iterative data–model–experiment workflow may integrate experimental, literature-derived, and simulation data using descriptors such as dopant identity and concentration, Ca/P ratio, synthesis conditions, particle size, crystallinity, phase composition, porosity, and surface properties. These inputs may be associated with target outcomes including magnetic, photothermal, optical, antimicrobial, cytotoxic, osteogenic, and angiogenic responses. Regression, classification, and multi-objective approaches may then predict relevant properties and rank candidate formulations against predefined performance and safety criteria.

TABLE III
TRANSLATIONAL ENGINEERING OF BIOCERAMICS (TEB) FRAMEWORK: ENGINEERING STAGES, INPUTS, AND EXPECTED OUTPUTS FOR THE DEVELOPMENT OF NEXT-GENERATION THERANOSTIC BIOCERAMICS.

Framework Stage	**Engineering Objective**	**Prospective Evaluation Targets**
Clinical Need Identification	Define the intended clinical problem, application, and required functions before material design.	Explicit target product/application profile, intended use, clinical requirements, and predefined performance and safety criteria.
Rational Ion Co-Doping Design	Select candidate dopants and concentration ranges based on expected physicochemical and biological mechanisms.	Candidate ion combinations and concentration ranges prioritized for experimental testing against predefined functional and biological criteria.
Material Engineering	Develop and control synthesis, processing, composition, microstructure, and physicochemical properties.	Reproducible candidate formulations meeting predefined compositional, structural, and processing specifications.
Multimodal Characterization	Characterize structural, chemical, surface, optical, magnetic, thermal, and other application-relevant properties.	Experimentally measured property profiles compared with predefined performance criteria and suitable for informing subsequent design decisions.
Computational and AI-Assisted Ion Co-Doping Optimization	Predict relevant properties and prioritize ion combinations and processing conditions for experimental validation.	Ranked candidate formulations with predicted performance and uncertainty, followed by prospective comparison with experimentally measured outcomes.
Biological Validation	Evaluate safety and application-specific biological responses using progressively relevant experimental models.	Evidence of cytocompatibility and predefined biological functionality in appropriate in vitro and, where justified, subsequent preclinical models.
Prototype Development	Integrate selected bioceramics into application-specific biomedical prototypes and assess manufacturability and functional performance.	Prototype feasibility, reproducibility, functional performance, and manufacturing requirements suitable for subsequent TRL assessment.
Translational Planning and Evaluation	Examine regulatory, manufacturing, quality, clinical, and implementation requirements for further development.	Defined regulatory and manufacturing pathway, identified evidence gaps, and criteria for subsequent preclinical, clinical, and technology-maturation studies.

Regression, classification, or multi-objective optimization models may then be used to predict relevant properties and rank candidate compositions according to predefined application-specific requirements. Thus, rather than directly selecting an ion combination, the computational stage would prioritize experimentally testable candidates predicted to provide the most favorable balance among multiple desired functionalities and safety constraints.

Model performance should be evaluated according to the prediction task. Continuous-property predictions may be assessed using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2), whereas classification tasks may use precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC), when applicable. Cross-validation should be used during model development, while independent or prospectively generated experimental data should be preferred for evaluating generalizability.

Importantly, computational performance alone would not establish the utility of the approach within the TEB Framework. Candidate compositions prioritized by the model should subsequently be synthesized and experimentally characterized, allowing predicted and measured outcomes to be compared and the ranking of candidate formulations to be evaluated.

The experimental results would then be incorporated into subsequent model iterations, creating an iterative learning cycle between computational prediction and laboratory validation. Future prospective studies could further assess practical utility by comparing AI-assisted candidate prioritization with conventional empirical selection in terms of predictive agreement, proportion of successful top-ranked candidates, and the number of experimental iterations required to identify formulations meeting predefined multifunctional performance and safety criteria. This computational component therefore remains a proposed and testable element of the TEB Framework whose added value must be established through prospective experimental validation.

V. LIMITATIONS AND FUTURE PERSPECTIVES

The future development of multifunctional theranostic bioceramics will require greater integration between rational material design, biological validation, prototype development, scalable manufacturing, and regulatory planning. Rather than focusing on isolated material properties, future research should

advance toward multifunctional platforms combining diagnostic, therapeutic, regenerative, and sensing capabilities. This transition will require multidisciplinary collaboration and standardized workflows capable of supporting reproducibility, safety, and subsequent translational evaluation.

Computational materials science, computational chemistry, artificial intelligence, digital twins, and high-throughput experimentation may complement experimental approaches by enabling more systematic exploration of complex compositional and processing spaces. Within the TEB Framework, these tools are intended to support hypothesis generation, candidate prioritization, and experimental design rather than replace experimental validation. Computationally prioritized ion combinations and predicted material properties should therefore be evaluated through synthesis, physicochemical characterization, and biological testing before progression toward prototype development.

The TEB Framework should be interpreted as a literature-informed conceptual and hypothesis-generating framework rather than an experimentally validated translational pathway. It has not yet been prospectively tested in experimental programs, biomedical device development, or real-world clinical settings; therefore, its stages and proposed outputs represent engineering targets rather than demonstrated gains in technological maturity or clinical translation. Future studies should progressively evaluate and refine the framework through experimental case studies, computational–experimental validation, prototype development, and prospective translational applications, using measurable indicators such as predictive agreement, reproducibility, experimental efficiency, prototype performance, and independently assessed progression against relevant TRL criteria.

VI. Conclusions

Multifunctional calcium phosphate bioceramics offer increasing opportunities to integrate diagnostic, therapeutic, regenerative, and sensing functions; however, their progression toward biomedical technologies remains constrained by fragmented development approaches. This work proposed the Translational Engineering of Bioceramics (TEB) Framework as a literature-informed conceptual methodology that organizes clinical needs, rational ion co-doping, material engineering, multimodal characterization, computational and AI-assisted candidate prioritization, biological validation, prototype development, and translational planning within a unified engineering framework.

The main contribution of the TEB Framework is a systems-oriented perspective connecting materials science, biomedical engineering, computational approaches, and translational considerations rather than proposing a new material composition or claiming a validated development pathway. The framework may provide a structured basis for designing and prospectively evaluating multidisciplinary research on theranostic bioceramics and potentially other multifunctional biomaterials. Its practical utility, including whether it can improve experimental efficiency, reproducibility, technological maturation, or progression toward clinical translation, remains to be established through prospective experimental and real-world validation.

References

[1] Gopalu Karunakaran, Eun-Bum Cho, Govindan Suresh Kumar, Evgeny Kolesnikov, Arkhipov Dmitry, Saheb Ali, Microwave-assisted synthesis of superparamagnetic mesoporous Co-doped hydroxyapatite nanorods for various biomedical applications, Ceramics International, Volume 47, Issue 6, 2021, Pages 8642-8652, ISSN 0272-8842, https://doi.org/10.1016/j.ceramint.2020.11.234..

[2] Beniwal N, Verma A, Putta CL, Rengan AK. Recent Trends in Bionanomaterials and Non-invasive Combinatorial Approaches of Photothermal Therapy against Cancer. Nanotheranostics 2024; 8(2):219-238. doi:10.7150/ntno.91356. https://www.ntno.org/v08p0219.htm.

[3] Urazaliyeva, A., Kanabekova, P., Beisenbayev, A. et al. All organic nanomedicine for PDT–PTT combination therapy of cancer cells in hypoxia. Sci Rep 14, 17507 (2024). https://doi.org/10.1038/s41598-024-68077-4.

[4] Pan, T., Lu, D., Xin, H. et al. Biophotonic probes for bio-detection and imaging. Light Sci Appl 10, 124 (2021). https://doi.org/10.1038/s41377-021-00561-2.

[5] Rahul Verma, Soumya Ranjan Mishra, Vishal Gadore, Md. Ahmaruzzaman, Hydroxyapatite-based composites: Excellent materials for environmental remediation and biomedical applications, Advances in Colloid and Interface Science, Volume 315, 2023, 102890, ISSN 0001-8686, https://doi.org/10.1016/j.cis.2023.102890.

[6] Campodoni, E., Montanari, M., Artusi, C., Bassi, G., Furlani, F., Montesi, M., Panseri, S., Sandri, M., & Tampieri, A. (2021). Calcium-Based Biomineralization: A Smart Approach for the Design of Novel Multifunctional Hybrid Materials. Journal of Composites Science, 5(10), 278. https://doi.org/10.3390/jcs5100278

[7] Dongqin Xiao, Jingwei Zhang, Chengdong Zhang, Davide Barbieri, Huipin Yuan, Lorenzo Moroni, Gang Feng, The role of calcium phosphate surface structure in osteogenesis and the mechanisms involved, Acta Biomaterialia, Volume 106, 2020, Pages 22-33, ISSN 1742-7061, https://doi.org/10.1016/j.actbio.2019.12.034.

[8] Ghadimi, E., Eimar, H., Marelli, B. et al. Trace elements can influence the physical properties of tooth enamel. SpringerPlus 2, 499 (2013). https://doi.org/10.1186/2193-1801-2-499

[9] H.R. Fernandes, S. Kannan, M. Alam, G.E. Stan, A.C. Popa, R. Buczyński, P. Gołębiewski, J.M.F. Ferreira, Two decades of continuous progresses and breakthroughs in the field of bioactive ceramics and glasses driven by CICECO-hub scientists, Bioactive Materials, Volume 40, 2024, Pages 104-147, ISSN 2452-199X, https://doi.org/10.1016/j.bioactmat.2024.05.041.

[10] Marta Overchuk, Robert A. Weersink, Brian C. Wilson, Gang Zheng; Photodynamic and Photothermal Therapies: Synergy Opportunities for Nanomedicine. ACS Nano 9 May 2023; 17 (9): 7979–8003. https://doi.org/10.1021/acsnano.3c00891.

[11] Sarbadhikary P, George BP, Abrahamse H. Recent Advances in Photosensitizers as Multifunctional Theranostic Agents for Imaging-Guided Photodynamic Therapy of Cancer. Theranostics 2021; 11(18):9054-9088. doi:10.7150/thno.62479. https://www.thno.org/v11p9054.htm.

[12] Alvarez, N.; Sevilla, A. Current Advances in Photodynamic Therapy (PDT) and the Future Potential of PDT-Combinatorial Cancer Therapies. Int. J. Mol. Sci. 2024, 25, 1023. https://doi.org/10.3390/ijms25021023.

[13] engi Yilmaz, Ammar Z. Alshemary, Zafer Evis, Co-doped hydroxyapatites as potential materials for biomedical applications, Microchemical Journal, Volume 144, 2019, Pages 443-453, ISSN 0026-265X, https://doi.org/10.1016/j.microc.2018.10.007.

[14] Bahman Nasiri-Tabrizi, Wan Jefrey Basirun, Chai Hong Yeong, Win Min Thein, Development of the third generation of bioceramics: Doping hydroxyapatite with s-, p-, d-, and f-blocks cations and their potential applications in bone regeneration and void filling, Ceramics International, Volume 49, Issue 5, 2023, Pages 7142-7179, ISSN 0272-8842, https://doi.org/10.1016/j.ceramint.2022.12.117.

[15] Fendi, F., Abdullah, B., Suryani, S. et al. Hydroxyapatite based for bone tissue engineering: innovation and new insights in 3D printing technology. Polym. Bull. 81, 1097–1116 (2024). https://doi.org/10.1007/s00289-023-04794-6

[16] Inam, H., Sprio, S., Tavoni, M., Abbas, Z., Pupilli, F., & Tampieri, A. (2024). Magnetic Hydroxyapatite Nanoparticles in Regenerative Medicine and Nanomedicine. International Journal of Molecular Sciences, 25(5), 2809. https://doi.org/10.3390/ijms25052809

[17] Tianming Du, Xufeng Niu, Peng Cao, Yanping Zhang, Youjun Liu, Haisheng Yang, Aike Qiao, Multifarious roles of metal elements in bone mineralization, Applied Materials Today, Volume 32, 2023, 101810, ISSN 2352-9407, https://doi.org/10.1016/j.apmt.2023.101810.

[18] Wang, X., Huang, S., & Peng, Q. (2023). Metal Ion-Doped Hydroxyapatite-Based Materials for Bone Defect Restoration. Bioengineering, 10(12), 1367. https://doi.org/10.3390/bioengineering10121367

[19] Syamchand, S.S., Sony, G. Multifunctional hydroxyapatite nanoparticles for drug delivery and multimodal molecular imaging. Microchim Acta 182, 1567–1589 (2015). https://doi.org/10.1007/s00604-015-1504-x

[20] A. Yazdanpanah, M. Ghaffari, Z. Ahmadi, A. Abrishamkar, S. Sattarzadeh, A. Ramedani, S. Arabyazdi, F. Moztarzadeh, Nano Select. 2023, 4, 353. https://doi.org/10.1002/nano.202300008

[21] Arun Kumar Rajendran, Mary Susan J. Anthraper, Nathaniel S. Hwang, Jayakumar Rangasamy, Osteogenesis and angiogenesis promoting bioactive ceramics, Materials Science and Engineering: R: Reports, Volume 159, 2024, 100801, ISSN 0927-796X, https://doi.org/10.1016/j.mser.2024.100801.

[22] Jamileh Kadkhoda, Ali Tarighatnia, Jaleh Barar, Ayuob Aghanejad, Soodabeh Davaran, Recent advances and trends in nanoparticles based photothermal and photodynamic therapy, Photodiagnosis and Photodynamic Therapy,Volume 37, 2022, 102697, ISSN 1572-1000, https://doi.org/10.1016/j.pdpdt.2021.102697.

[23] Mengqin Gu, Wei Li, Li Jiang, Xiyu Li, Recent progress of rare earth doped hydroxyapatite nanoparticles: Luminescence properties, synthesis and biomedical applications, Acta Biomaterialia, Volume 148, 2022, Pages 22-43, ISSN 1742-7061, https://doi.org/10.1016/j.actbio.2022.06.006.

[24] T. S. Vo, T. Hoang, T. T. B. C. Vo, B. Jeon, V. H. Nguyen, K. Kim, Recent Trends of Bioanalytical Sensors with Smart Health Monitoring Systems: From Materials to Applications. Adv. Healthcare Mater. 2024, 13, 2303923. https://doi.org/10.1002/adhm.202303923

[25] De Tommasi, F., Massaroni, C., Grasso, R. F., Carassiti, M., Schena, E. (2021). Temperature Monitoring in Hyperthermia Treatments of Bone Tumors: State-of-the-Art and Future Challenges. Sensors, 21(16), 5470. https://doi.org/10.3390/s21165470

[26] Duta, L., Grumezescu, V. (2024). The Effect of Doping on the Electrical and Dielectric Properties of Hydroxyapatite for Medical Applications: From Powders to Thin Films. Materials, 17(3), 640. https://doi.org/10.3390/ma17030640

[27] Ghadimi, E., Eimar, H., Marelli, B. et al. Trace elements can influence the physical properties of tooth enamel. SpringerPlus 2, 499 (2013). https://doi.org/10.1186/2193-1801-2-499

[28] Adam Shearer, Maziar Montazerian, Jessica J. Sly, Robert G. Hill, John C. Mauro, Trends and perspectives on the commercialization of bioactive glasses, Acta Biomaterialia, Volume 160, 2023, Pages 14-31, ISSN 1742-7061, https://doi.org/10.1016/j.actbio.2023.02.020.

[29] Subhadip Basu, Bikramjit Basu; Unravelling Doped Biphasic Calcium Phosphate: Synthesis to Application. ACS Appl. Bio Mater. 16 December 2019; 2 (12): 5263–5297. https://doi.org/10.1021/acsabm.9b00488

[30] Cacciotti I. Multisubstituted hydroxyapatite powders and coatings: The influence of the codoping on the hydroxyapatite performances. Int J Appl Ceram Technol. 2019; 16: 1864–1884. https://doi.org/10.1111/ijac.13229

[31] Torres, E.C.L., De Sousa, E.M.B. and Cipreste, M.F. (2022) A Brief Review on Hydroxyapatite Nanoparticles Interactions with Biological Constituents. Journal of Biomaterials and Nanobiotechnology, 13, 24-44. https://doi.org/10.4236/jbnb.2022.131002

[32] Torres, F. , Sousa, E. and Cipreste, M. (2022) A Brief Review on Hydroxyapatite Nanoparticles Interactions with Biological Constituents. Journal of Biomaterials and Nanobiotechnology, 13, 24-44. doi: 10.4236/jbnb.2022.131002.

[33] José S. Rabelo Neto, Philipi C. Ricardo, Mário E.G. Valério, Wei Xia, Håkan Engqvist, Márcio C. Fredel, The influence of strontium doping on the crystal morphology of synthetic calcium phosphates, Journal of Molecular Structure, Volume 1316, 2024, 139030, ISSN 0022-2860, https://doi.org/10.1016/j.molstruc.2024.139030.

[34] Rabelo Neto, Jose da Silva, Knopf, Thaiane Balesteri, Fredel, Marcio Celso, Olate, Sergio, de Moraes, Paulo H. (2015). Synthesis and Characterization of Calcium Phosphate Compounds with Strontium and Magnesium Ionic Substitutions. International Journal of Morphology, 33(3), 1189-1193. https://dx.doi.org/10.4067/S0717-95022015000300061.

[35] Boanini, E., Torricelli, P., Fini, M., Bigi, A. (2011). Osteopenic bone cell response to strontium-substituted hydroxyapatite. Journal of Materials Science: Materials in Medicine, 22, 2079-2088.

[36] Alves Côrtes, J.; Dornelas, J.; Duarte, F.; Messora, M.R.; Mourão, C.F.; Alves, G. The Effects of the Addition of Strontium on the Biological Response to Calcium Phosphate Biomaterials: A Systematic Review. Appl. Sci. 2024, 14, 7566. https://doi.org/10.3390/ app14177566

[37] Singh, R.P., Singh, D. Singh, J.P. Biomagnetic mesoporous nanorods for hyperthermia, drug delivery, and tissue regeneration applications. J Sol-Gel Sci Technol 97, 155–166 (2021). https://doi.org/10.1007/s10971-020-05421-w

[38] Li, J., Zhao, C., Liu, C. et al. Cobalt-doped bioceramic scaffolds fabricated by 3D printing show enhanced osteogenic and angiogenic properties for bone repair. BioMed Eng OnLine 20, 70 (2021). https://doi.org/10.1186/s12938-021-00907-2

[39] Lin Yan, Xiaobo Wei, Ziyue Zhang, Cheng Wang, Yibo Jia, Lei Wang, Yonggang Yan, Xiaoxia Fan, Cobalt-doped hydroxyapatite for bone tissue engineering: Synthesis, characterization and in vitro biocompatibility of real-time extract, Materials Today Communications, Volume 38, 2024, 108554, ISSN 2352-4928, https://doi.org/10.1016/j.mtcomm.2024.108554.

[40] Vu Hoang Minh Doan, Sudip Mondal, Thi Mai Thien Vo, Cao Duong Ly, Dinh Dat Vu, Van Tu Nguyen, Sumin Park, Jaeyeop Choi, Junghwan Oh, Fluorescence conjugated nanostructured cobalt-doped hydroxyapatite platform for imaging-guided drug delivery application, Colloids and Surfaces B: Biointerfaces, Volume 214, 2022, 112458, ISSN 0927-7765, https://doi.org/10.1016/j.colsurfb.2022.112458.

[41] Moussa, S.I., Dakroury, G.A.R.S. El Shazly, E.A.E.S.A.H. Evaluation of Fe-doped calcium phosphate for 65Zn sorption. J Radioanal Nucl Chem 333, 4575–4590 (2024). https://doi.org/10.1007/s10967-024-09625-0

[42] Marta S. Laranjeira, Ana Moço, Jorge Ferreira, Susana Coimbra, Elísio Costa, Alice Santos-Silva, Paulo J. Ferreira, Fernando J. Monteiro, Different hydroxyapatite magnetic nanoparticles for medical imaging: Its effects on hemostatic, hemolytic activity and cellular cytotoxicity, Colloids and Surfaces B: Biointerfaces, Volume 146, 2016, Pages 363-374, ISSN 0927-7765, https://doi.org/10.1016/j.colsurfb.2016.06.042.

[43] Kyriacou, A., Leventouri, T., Chakoumakos, B.C., Garlea, V.O., dela Cruz, C., Rondinone, A.J., & Sorge, K. (2013). Combined X-ray and neutron diffraction Rietveld refinement in iron-substituted nanohydroxyapatite. Journal of Materials Science, 48, 3535-3545.

[44] Anna Tampieri, Teresa D'Alessandro, Monica Sandri, Simone Sprio, Elena Landi, Luca Bertinetti, Silvia Panseri, Giancarlo Pepponi, Joerg Goettlicher, Manuel Bañobre-López, Jose Rivas, Intrinsic magnetism and hyperthermia in bioactive Fe-doped hydroxyapatite, Acta Biomaterialia, Volume 8, Issue 2, 2012, Pages 843-851, ISSN 1742-7061, https://doi.org/10.1016/j.actbio.2011.09.032.

[45] Bahroz Kareem Mahmood, Omer Kaygili, Niyazi Bulut, Sergey V. Dorozhkin, Tankut Ates, Suleyman Koytepe, Canbolat Gürses, Filiz Ercan, Hanifi Kebiroglu, Riyadh Saeed Agid, Turan İnce, Effects of strontium - erbium co-doping on the structural properties of hydroxyapatite: An Experimental and theoretical study, Ceramics International, Volume 46, Issue 10, Part B, 2020, Pages 16354-16363, ISSN 0272-8842, https://doi.org/10.1016/j.ceramint.2020.03.194.

[46] "Physicochemical study of magnesium zinc codoped-hydroxyapatite", morjchem, vol. 12, no. 3, pp. 1240–1253, May 2024, doi: 10.48317/IMIST.PRSM/morjchem-v12i3.48113.

[47] Sathyanarayanan, Shivani and Kannan, Saranya. "Strontium and copper co-doped nanohydroxyapatite for bone augmentation" International Journal of Materials Research, vol. 115, no. 3, 2024, pp. 202-207. https://doi.org/10.1515/ijmr-2023-0089

[48] Katarzyna Szyszka, Justyna Rewak-Soroczynska, Agata Dorotkiewicz-Jach, Karolina A. Ledwa, Agata Piecuch, Michael Giersig, Zuzanna Drulis-Kawa, Rafal J. Wiglusz, Structural modification of nanohydroxyapatite Ca10(PO4)6(OH)2 related to Eu3+ and Sr2+ ions doping and its spectroscopic and antimicrobial properties, Journal of Inorganic Biochemistry, Volume 203, 2020, 110884, ISSN 0162-0134, https://doi.org/10.1016/j.jinorgbio.2019.110884.

[49] Rewak-Soroczynska, J.; Nowak, N.; Targonska, S.; Piecuch, A.;Wiglusz, R.J. The Study of Nanosized Silicate-Substituted Hydroxyapatites Co-Doped with Sr2+ and Zn2+ Ions Related to Their Influence on Biological Activities. Curr. Issues Mol. Biol. 2022, 44, 6229–6246. https://doi.org/10.3390/cimb44120425

Análise da Estabilidade Termofísica de Alinhadores Ortodônticos por Calorimetria Exploratória Diferencial (DSC)

Alessandro Fraga Farah
Mecânica/Sistemas Biomédicos
Fatec Sertãozinho/Fatec Ribeirão Preto
Sertãozinho/Ribeirão Preto, Brasil
alessandro.farah@cps.sp.gov.br

Analu Egydio dos Santos
Sistemas Biomédicos
Fatec Ribeirão Preto
Ribeirão Preto, Brasil
analu.santos@cps.sp.gov.br

Cristian Cley Paterniani Rita
Sistemas Biomédicos
Fatec Ribeirão Preto
Ribeirão Preto, Brasil
cristian.rita@cps.sp.gov.br

Lara Oliveira Ramos
Sistemas Biomédicos
Fatec Ribeirão Preto
Ribeirão Preto, Brasil
lara.ramos@fatec.sp.gov.br

Lucas Xavier Barini
Sistemas Biomédicos
Fatec Ribeirão Preto
Ribeirão Preto, Brasil
lucas.barini@aluno.cps.sp.gov.br

Vinicius Elias Chagas de Faria
Sistemas Biomédicos
Fatec Ribeirão Preto
Ribeirão Preto, Brasil
vinicius.faria9@aluno.cps.sp.gov.br

Resumo—Os alinhadores ortodônticos transparentes são hoje uma das principais alternativas aos aparelhos convencionais devido à sua estética, conforto e previsibilidade clínica. Entretanto, o desempenho biomecânico desses dispositivos depende diretamente das propriedades físico-químicas e termomecânicas dos polímeros utilizados em sua fabricação. Entre as técnicas de caracterização de materiais poliméricos, a Calorimetria Exploratória Diferencial (Differential Scanning Calorimetry – DSC) destaca-se por possibilitar a determinação de parâmetros fundamentais para compreender a estabilidade termofísica dos alinhadores. O presente artigo teve como objetivo revisar a literatura científica e caracterizar, por DSC, três tipos de polímeros comerciais, utilizados como alinhadores ortodônticos, discutindo a influência da composição dos polímeros sobre suas propriedades e desempenho. Os resultados obtidos mostraram as principais diferenças na caracterização termofísica dos polímeros analisados, mostrando-se uma importante ferramenta para o desenvolvimento de materiais mais estáveis, previsíveis e duráveis, contribuindo para a evolução tecnológica dos alinhadores ortodônticos.

Palavras-chave—Alinhadores ortodônticos; DSC; Biomateriais.

I. INTRODUÇÃO

Os materiais termoplásticos empregados na fabricação de alinhadores ortodônticos devem apresentar elevada transparência óptica, biocompatibilidade, estabilidade dimensional, resistência mecânica e capacidade de manter forças leves e contínuas durante todo o período de utilização clínica [1, 2]. As propriedades desses materiais dependem da estrutura molecular dos polímeros, do grau de cristalinidade, da massa molar e do processamento industrial utilizado na fabricação das placas termoplásticas.

Entre os materiais mais empregados destacam-se o poliuretano termoplástico (TPU), o politereftalato de etileno modificado por glicol (PETG), o polipropileno (PP) e polímeros multicamadas desenvolvidos especificamente para aplicações ortodônticas [2, 3, 4]. O TPU apresenta elevada elasticidade e boa resistência ao desgaste, favorecendo a manutenção das forças ortodônticas durante períodos prolongados. O PETG destaca-se pela excelente transparência, facilidade de conformação e boa estabilidade dimensional, enquanto o polipropileno apresenta menor densidade e maior resistência química, embora seja mais sensível às alterações provocadas pelo processamento térmico [3, 4, 5].

A termoformagem constitui uma das etapas mais importantes da fabricação dos alinhadores. Durante esse processo, placas termoplásticas são aquecidas acima da temperatura de transição vítrea e moldadas sobre modelos tridimensionais da arcada dentária. O aquecimento promove reorganização parcial das cadeias poliméricas, podendo modificar o grau de cristalinidade, a orientação molecular e a distribuição de tensões internas do material [4, 5, 6, 9].

Diversos estudos demonstram que essas modificações estruturais influenciam diretamente as propriedades mecânicas dos alinhadores. Alterações no módulo de elasticidade, na resistência à tração e na recuperação elástica podem reduzir a magnitude das forças aplicadas aos dentes durante o tratamento ortodôntico, afetando a eficiência clínica dos dispositivos [1, 2, 3, 4, 5, 6, 8, 9].

Nesse contexto, a Calorimetria Exploratória Diferencial (Differential Scanning Calorimetry – DSC) tornou-se uma das principais ferramentas para caracterização dos polímeros utilizados na Ortodontia [1, 4, 5, 7]. A técnica baseia-se na medição da diferença de fluxo de calor entre uma amostra e um material de referência durante ciclos controlados de aquecimento e resfriamento [1, 5, 7]. A partir dessas curvas térmicas é possível determinar importantes parâmetros, como temperatura de transição vítrea (Tg), temperatura de cristalização (Tc), temperatura de fusão (Tm), capacidade calorífica e entalpia de fusão, permitindo avaliar a estabilidade térmica e estrutural dos materiais [1, 5, 7, 9].

A temperatura de transição vítrea representa um dos parâmetros mais importantes para materiais poliméricos utilizados em alinhadores [7, 9]. Abaixo da Tg, o polímero apresenta comportamento rígido e vítreo; acima dessa temperatura, ocorre aumento da mobilidade molecular, tornando o material mais flexível. Como a temperatura intraoral situa-se próxima de 37°C, pequenas variações na Tg podem

modificar significativamente o comportamento do alinhador durante sua utilização clínica [5, 7, 9].

Outro parâmetro importante obtido pela DSC é a entalpia de fusão, utilizada para estimar o grau de cristalinidade dos polímeros [1, 5, 7, 9]. Em materiais semicristalinos, maior cristalinidade normalmente está associada ao aumento da rigidez e da resistência mecânica, porém pode reduzir a flexibilidade do material. Dessa forma, o equilíbrio entre regiões amorfas e cristalinas torna-se fundamental para garantir desempenho clínico adequado [6, 8, 10].

Além do processamento térmico, diversos fatores relacionados ao ambiente bucal contribuem para o envelhecimento dos alinhadores [11, 12]. A absorção de água, as oscilações térmicas decorrentes da ingestão de alimentos quentes e frios, as soluções de limpeza e as cargas mastigatórias promovem alterações graduais na estrutura dos polímeros [11]. Esses fenômenos podem resultar em plastificação, relaxação molecular e diminuição das propriedades mecânicas, comprometendo a estabilidade dimensional dos dispositivos [6, 8, 9, 10, 11, 12].

Recentemente, a impressão tridimensional direta de alinhadores ortodônticos surgiu como uma alternativa promissora aos processos convencionais de termoformagem [2, 8, 10]. Essa tecnologia reduz etapas de processamento e pode minimizar alterações estruturais induzidas pelo aquecimento, embora ainda existam limitações relacionadas à estabilidade dos materiais fotopolimerizáveis e à necessidade de validação clínica em longo prazo [2, 10].

Assim, a literatura demonstra que a associação entre técnicas de caracterização térmica, mecânica e microestrutural representa uma importante estratégia para o desenvolvimento de biomateriais ortodônticos mais estáveis e duráveis, possibilitando avanços significativos na previsibilidade dos tratamentos realizados com alinhadores transparentes.

Diante dessa necessidade de investigação, esse trabalho teve como objetivo iniciar as caracterizações de diferentes alinhadores ortodônticos, começando pelos ensaios de Calorimetria Exploratória Diferencial (Differential Scanning Calorimetry – DSC).

II. MATERIAIS E MÉTODOS

A. Aquisição e preparação das amostras

Foram utilizadas placas de polímeros comerciais de três empresas diferentes. A placa da empresa 1 é identificada, na sua etiqueta, como Placa PET-G. A placa da empresa 2 é identificada como Placa de Acetato Flexflow. A placa da empresa 3, como Placa de Copoliéster. Todas as placas possuem diâmetro de 125mm e espessura entre 0,75 e 0,80mm.

Para os ensaios de Calorimetria Exploratória Diferencial (Differential Scanning Calorimetry – DSC), foram cortadas, de cada placa, 03 pequenas amostras circulares com aproximadamente 5mm de diâmetro e massa entre 5 e 10mg.

B. Parâmetros dos ensaios de DSC

As análises foram feitas num equipamento DSC, marca TA Instruments, modelo Discovery DSC25, com temperatura entre -90 e 550°C. A Figura 1 mostra o equipamento.

Fig. 1 – Equipamento DSC



As amostras foram colocadas em cadinhos de alumínio hermeticamente fechados. A atmosfera utilizada foi o nitrogênio (N_2) com fluxo de 50 ml/min, para evitar termo-oxidação.

O programa de temperatura foi:

- Rampa de aquecimento de 25°C até 200°C a uma taxa de 20°C/min. (Este primeiro aquecimento apaga a história térmica do material).
- Isoterma de 1 minuto em 200°C.
- Rampa de resfriamento de 200°C a 25°C a uma taxa de 20°C/min.
- Rampa de segundo aquecimento de 25°C a 200°C a uma taxa de 20°C/min.

Os ensaios foram repetidos para as três amostras de cada placa polimérica comercial.

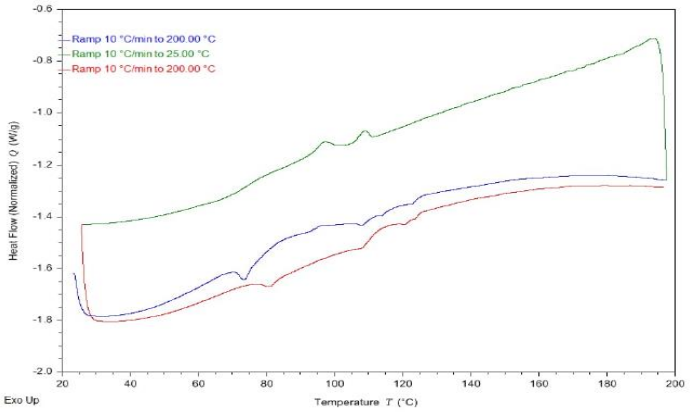
III. RESULTADOS E DISCUSSÕES

Os resultados obtidos das análises de Calorimetria Exploratória Diferencial (DSC), serão apresentados e discutidos a seguir.

A. Amostra de PET-G

A Figura 2 mostra a curva obtida pelo ensaio de DSC da amostra de PET-G.

Fig. 2 – Curva de DSC do PET-G



A curva do PET-G apresenta comportamento típico de um polímero predominantemente amorfo, caracterizado pela presença de uma transição vítrea relativamente bem definida e ausência de pico pronunciado de fusão. A ausência do pico de fusão se deve ao fato que a curva corresponde a um ensaio

limitado a 200°C e o pico de fusão de materiais baseados em PET-G, normalmente ocorre entre 220 e 250°C, dependendo da formulação.

A curva de DSC evidenciou uma transição vítrea bem definida na faixa de 74 a 78 °C, acompanhada por discreto evento exotérmico entre 95 e 105 °C, associado à cristalização a frio de baixa intensidade. Não foi observado pico de fusão dentro da faixa analisada, o que é compatível com a predominância de fase amorfa e com o limite superior do ensaio.

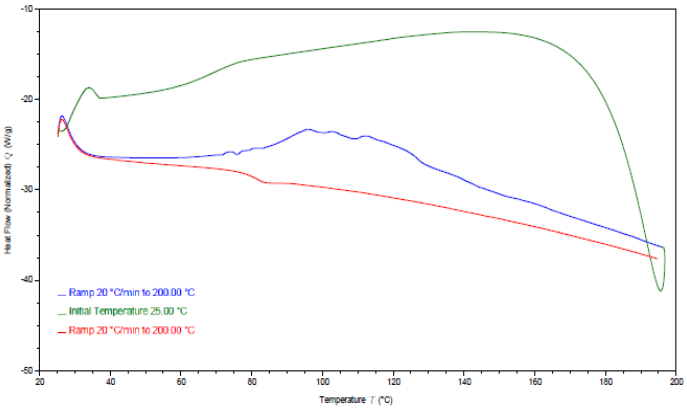
Esse comportamento é coerente com a estrutura do PET-G, uma vez que a incorporação de glicol à cadeia do PET reduz a capacidade de empacotamento molecular e, consequentemente, a cristalização. Do ponto de vista clínico, tal característica é positiva, pois materiais com maior fração amorfa tendem a apresentar melhor transparência óptica e maior facilidade de conformação durante a termoformagem. Contudo, a presença de cristalização a frio indica que o material ainda possui certa capacidade de reorganização sob aquecimento, o que pode interferir em sua estabilidade ao longo do uso.

Além disso, a Tg observada encontra-se relativamente próxima à faixa de temperaturas que podem ser alcançadas durante a ingestão de bebidas quentes ou em etapas de processamento térmico. Isso sugere que o PET-G, embora adequado para aplicações ortodônticas, pode sofrer alterações de desempenho se exposto repetidamente a condições térmicas extremas. Estudos com alinhadores termoplásticos têm mostrado que envelhecimento em saliva artificial e ciclos térmicos podem reduzir a rigidez e a estabilidade dimensional, reforçando a importância da seleção criteriosa do material [4, 5, 12]. Esse comportamento é esperado para o PET-G, pois a introdução do glicol na cadeia do PET reduz significativamente sua capacidade de cristalização.

B. Amostra de Acetato

A Figura 3 mostra a curva obtida pelo ensaio de DSC da amostra de Acetato.

Fig. 3 – Curva de DSC do Acetato



A amostra de acetato apresentou perfil térmico distinto do PET-G, com transição vítrea semelhante, entre 72 e 76 °C, porém com evento exotérmico mais largo e mais intenso entre 95 e 110 °C. Esse comportamento sugere maior reorganização molecular durante o aquecimento, refletindo maior mobilidade das cadeias e menor estabilidade térmica em comparação ao PET-G.

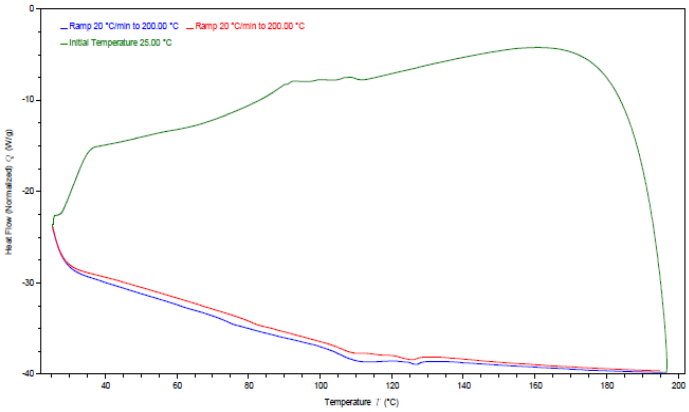
Na prática, isso indica que o acetato pode ser mais sensível a variações de temperatura e a processos de relaxamento estrutural. Em materiais usados como alinhadores, maior mobilidade molecular pode ser vantajosa em etapas de conformação inicial, mas representa desvantagem durante o uso, pois pode favorecer perda de forma, redução do armazenamento de tensões elásticas e queda da eficiência de ativação ortodôntica. Em outras palavras, o material pode deformar-se com mais facilidade e recuperar-se com menor previsibilidade.

Outro aspecto relevante é que a maior amplitude do evento térmico sugere heterogeneidade na estrutura interna do polímero, possivelmente associada à formulação comercial, ao processamento industrial ou à presença de aditivos. Em sistemas poliméricos odontológicos, essa heterogeneidade pode comprometer a repetibilidade do desempenho entre lotes e aumentar a variabilidade clínica. Assim, o acetato analisado mostrou-se menos estável do que o PET-G sob as condições do ensaio, o que merece consideração em aplicações de longa duração ou em situações nas quais a manutenção de forças é crítica.

C. Amostra de Copoliéster

A Figura 4 mostra a curva obtida pelo ensaio de DSC da amostra de Copoliéster.

Fig. 4 – Curva de DSC do Copoliéster



Entre os materiais avaliados, o copoliéster exibiu o comportamento termofísico mais estável. A curva apresentou Tg entre 120 e 130 °C e ausência de eventos relevantes de cristalização ou fusão no intervalo estudado. Além disso, observou-se boa sobreposição entre as curvas, sugerindo repetibilidade experimental e estabilidade estrutural durante os ciclos térmicos.

A Tg mais elevada indica que o copoliéster possui maior resistência à mobilidade molecular quando comparado aos outros materiais. Isso é particularmente relevante para alinhadores ortodônticos, pois um polímero com Tg distante da temperatura intraoral tende a manter suas propriedades mecânicas com menor variação durante o uso. Em termos clínicos, isso pode se traduzir em melhor manutenção da força exercida sobre os dentes e menor suscetibilidade à deformação permanente.

A ausência de cristalinidade significativa também é favorável, uma vez que materiais amorfos e termicamente estáveis costumam oferecer boa transparência e previsibilidade dimensional. Em copoliésteres desenvolvidos para uso

odontológico, a formulação visa justamente combinar estabilidade térmica, clareza óptica e resistência ao impacto. Os resultados observados neste estudo são compatíveis com essa finalidade e reforçam o potencial do copoliéster como material promissor para alinhadores de maior desempenho.[9, 12, 13].

D. Implicações clínicas

A comparação entre os três materiais evidencia que, embora todos apresentem caráter predominantemente amorfo, suas respostas térmicas são distintas e refletem diferenças importantes de formulação e comportamento em serviço. O PET-G mostrou equilíbrio entre processabilidade e estabilidade; o acetato demonstrou maior suscetibilidade à reorganização estrutural; e o copoliéster apresentou a maior resistência às variações térmicas.

Essas diferenças têm impacto direto sobre a previsibilidade clínica dos alinhadores. Materiais com Tg mais baixa e maior tendência à reorganização podem perder forças com mais rapidez, exigindo substituição mais frequente ou comprometendo a eficiência da movimentação dentária. Já materiais mais estáveis tendem a conservar melhor sua geometria e suas propriedades mecânicas ao longo do uso, favorecendo resultados mais consistentes.

Outro ponto a considerar é que a resposta térmica do material não deve ser analisada isoladamente. Na prática, fatores como absorção de água, desgaste por escovação, ataques químicos de soluções de limpeza e exposição a ciclos térmicos orais podem atuar de forma sinérgica na degradação do polímero. Portanto, a DSC fornece uma base importante, mas deve ser complementada por análises mecânicas, morfológicas e de envelhecimento acelerado para uma avaliação mais completa do desempenho dos alinhadores [1,4-6,10-13].

IV. CONCLUSÕES

Os resultados obtidos por Calorimetria Exploratória Diferencial (DSC) demonstram que os três materiais analisados apresentam comportamento predominantemente amorfo, característica desejável para polímeros empregados na fabricação de alinhadores ortodônticos, uma vez que essa condição favorece a transparência óptica e a conformabilidade durante o processamento. Entretanto, cada material apresentou respostas térmicas distintas, evidenciando diferenças em sua estrutura molecular e em sua estabilidade frente ao aquecimento.

O PET-G apresentou temperatura de transição vítrea (Tg) entre 74 e 78 °C, acompanhada por um discreto pico de cristalização a frio entre 95 e 105 °C, indicando baixa capacidade de reorganização cristalina. Esse comportamento confirma que a modificação da cadeia polimérica pela incorporação de glicol reduz significativamente a cristalinidade do material, mantendo sua estrutura majoritariamente amorfa.

O acetato exibiu Tg semelhante à do PET-G (72 a 76 °C), porém apresentou um evento exotérmico mais amplo e intenso entre 95 e 110 °C, evidenciando maior reorganização molecular durante o aquecimento. Esse resultado indica maior mobilidade das cadeias poliméricas e menor estabilidade térmica quando comparado ao PET-G, sugerindo que o material é mais suscetível a alterações estruturais em temperaturas elevadas.

Por sua vez, o copoliéster apresentou o comportamento térmico mais estável dentre os materiais avaliados. A elevada temperatura de transição vítrea, situada entre 120 e 130 °C, associada à ausência de eventos de cristalização ou fusão durante o ensaio, demonstra elevada estabilidade térmica e excelente manutenção da estrutura amorfa ao longo do aquecimento. A sobreposição das curvas obtidas também confirma a boa repetibilidade experimental e indica que o material possui elevada homogeneidade e estabilidade química.

De forma comparativa, verifica-se que o copoliéster apresenta a maior resistência às variações térmicas, seguido pelo PET-G, enquanto o acetato demonstrou maior tendência à reorganização estrutural. Esses resultados evidenciam que as diferenças observadas nas curvas de DSC estão diretamente relacionadas à composição química e ao grau de mobilidade molecular de cada polímero. Assim, a análise térmica confirma que o copoliéster apresenta o comportamento mais adequado para aplicações que exigem estabilidade dimensional e térmica, enquanto o PET-G mantém um equilíbrio entre estabilidade e processabilidade. Já o acetato, embora permaneça predominantemente amorfo, mostra maior sensibilidade ao aquecimento, característica que pode influenciar seu desempenho em aplicações sujeitas a variações de temperatura.

De modo geral, os resultados reforçam que a escolha do material para alinhadores ortodônticos deve ir além da transparência e do conforto, incorporando também critérios de estabilidade térmica e resistência à degradação estrutural. A análise por DSC mostrou-se uma ferramenta útil para triagem e comparação de materiais, oferecendo subsídios importantes para o desenvolvimento e a seleção de polímeros mais adequados ao uso clínico.

Como limitação, destaca-se que a presente avaliação considerou apenas o comportamento termofísico inicial dos materiais, sem a integração de ensaios mecânicos, de envelhecimento e de caracterização morfológica. Estudos futuros devem incluir essas abordagens combinadas, além de comparar o desempenho de materiais termoformados e impressos em 3D, de modo a ampliar a compreensão sobre a durabilidade e a efetividade dos alinhadores ortodônticos.

AGRADECIMENTOS

Os autores agradecem à FATEC Ribeirão Preto e a FATEC Sertãozinho, pelo suporte institucional e pela infraestrutura utilizada nas etapas desse estudo.

REFERÊNCIAS

[1] Pratto, I.; Busato, M. C. A.; Bittencourt, P. R. S. **Thermal and mechanical characterization of thermoplastic orthodontic aligners discs after molding process.** November 2021. Journal of the Mechanical Behavior of Biomedical Materials 126(946–951):104991. DOI: 10.1016/j.jmbbm.2021.104991

[2] Sousa, H. A. F.; Nascimento, J. J. P. N. F.; Sousa, M. A. F.; Genari, B. Souza, A. O.; Degrazia, F. W. **Aparelhos**

ortodônticos invisíveis: uma revisão. Research, Society and Development, v. 10, n. 1, e5510111259, 2021. (CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i1.11259

[3] Freitas, M. P. M .; Gutierrez, L. M. O. **Biocompatibilidade dos alinhadores ortodônticos: revisão de literatura**. Saúde bucal e qualidade de vida: avanços e práticas em Odontologia. Capítulo 11. 94-106. 2025. https://doi.org/10.22533/at.ed.4121225050611.

[4] Staderini, E.; Chiusolo, G.; Papi, M.; Palone, M.; Lombardo, L.; Cordaro, M. **Chemical, optical, and morphological properties of TPU and PET-G samples after aging in artificial saliva: an in vitro study**. *BMC Oral Health* (2025) 25:533. https://doi.org/10.1186/s12903-025-05863-0

[5] Arabpour, F.; Aghili, H. A.; Mahd, M. J.; Arabpour, Z.; Yasaei, S. **Evaluation of aging effects on the physical, chemical, and mechanical properties of thermoplastic materials for orthodontic clear aligners: an in vivo study.** BMC Oral Health. 2025, Oct 8; 25:1557. DOI: 10.1186/s12903-025-06966-4

[6] Albertini, P.; Mazzanti, V.; Mollica, F.; Pellitteri, F.; Palone, M.; Lombardo, L. **Stress Relaxation Properties of Five Orthodontic Aligner Materials: A 14-Day In-Vitro Study**. Bioengineering 2022, 9, 349. https://doi.org/10.3390/bioengineering9080349.

[7] Blanco, A.; et al. **Advancing polypropylene quantification in recycled HDPE: A comparative study of FTIR, DSC, TREF, and TGIC for enhanced plastic circularity.** Polymer Testing, Volume 149, 2025, 108855, ISSN 0142-9418, https://doi.org/10.1016/j.polymertesting.2025.108855.

[8] Kalluf, D. M. B. **Alinhadores Ortodônticos: a Ortodontia do século XXI**. 2023. 149 f. Tese (Doutorado em Odontologia – Ortodontia). Faculdade ILAPEO. Curitiba, 2023.

[9] Daniele, V. et al. **Thermoplastic Disks Used for Commercial Orthodontic Aligners: Complete Physicochemical and Mechanical Characterization.** Materials 2020, 13, 2386; doi:10.3390/ma13102386 www.mdpi.com/journal/materials

[10] Mendonça, L. M. S.; Queiroga, K. V. O. C.; Silva, E. A.; Costa, N.; Ponte, A. R.; Ponte, A. N. L.; Santos, W. R. A. **Propriedades mecânicas dos alinhadores ortodônticos impressos de maneira direta em comparação com os alinhadores termoformados.** Research, Society and Development, v. 14, n. 1, e7414148031, 2025. (CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v14i1.48031

[11] Frederico, R. A. S.; Neves, L. G.; Botta, S. B.; Mendonça, E. C. C.; Matos, A. B. **Influência da contaminação com saliva na resistência de união de sistemas adesivos ao esmalte**. Clínica e Pesquisa em Odontologia - UNITAU, Volume 1, Número 1, p. 20-25; 2009.

[12] Fang, D.; Zhang, N.; Chen, H.; Bai, Y. **Dynamic stress relaxation of orthodontic thermoplastic materials in a simulated oral environment.** November 2013. Dental Materials Journal 32(6):946-951. DOI: 10.4012/dmj.2013-131.

[13] Yang Hu, Hongdong Zhang, Xinming Pu, Liping Yang; **Naphthyl Copolyester with High Glass Transition Temperatures: Tunable Structure and Performance**. *ACS Appl. Polym. Mater.* 13 December 2024; 6 (23): 14794–14801. https://doi.org/10.1021/acsapm.4c02997

Ultrasound-guided 3D dosimetry using an optimized MAGIC-f phantom

1st Estela Maria Freitas Meireles
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
estelamariafm927@usp.br

2nd Thiago Tibúrcio Vicente
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
thiagotiburcio37@usp.br

3rd Ricardo Ricci Bordonal
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
ricardo.bordonal@usp.br

4th Carlos Ernesto Garrido Salmon
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
garrido@ffclrp.usp.br

5th Theo Zeferino Pavan
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
theozp@usp.br

6th Juliana Fernandes Pavoni
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
jfpavoni@ffclrp.usp.br

7th Adilton Carneiro
Department of Physics - FFCLRP
University of São Paulo
Ribeirão Preto, SP
adilton@usp.br

I. ABSTRACT

Polymer-gel dosimeters enable volumetric recording of absorbed ionizing-radiation dose, but conventional readout relies mainly on magnetic resonance imaging (MRI). This study investigated ultrasound elastography as a complementary approach for three-dimensional dosimetry using formaldehyde-modified MAGIC (MAGIC-f) gel and an optimized formulation with 6 percent gelatin and 2.5 percent agar. Phantoms were irradiated laterally at 20 and 40 Gy. Ultrasound elastography was performed using two shear-wave excitation approaches: acoustic-radiation-force shear-wave elastography (SWE) and mechanical perturbation generated by a shaker coupled to a metallic plate. The resulting shear- wave velocity maps were spatially sampled every 3 mm and reconstructed in 3D. MRI transverse relaxometry was acquired with a 12-echo spin-echo protocol (TE 17–204 ms). In the modified phantom, representative shear-wave velocities increased from approximately 3.0–3.4 m/s in non-irradiated regions to 4.4 m/s at 20 Gy and 4.6 m/s at 40 Gy. In the original formulation, values up to 4.3 and 5.0 m/s were observed at 20 and 40 Gy, respectively."Additionally, a comparison of the maps was performed for 30 Gy and 50 Gy. Mean R2 increased from 8.99±0.15 to 12.09±0.29 and 17.42±0.24 1/s in the original phantom and from 4.79±0.14 to 8.96±0.17 and 11.54±0.31 1/s in the modified phantom for 0, 20, and 40 Gy. These findings demonstrate dose-dependent mechanical and magnetic-relaxation changes and support ultrasound-based volumetric mapping of irradiated MAGIC-f gels.

***Index Terms*—3D dosimetry, MAGIC-f gel, shear-wave elastography, ultrasound, MRI relaxometry, R2.**

II. INTRODUCTION

The first studies on dosimetric gels for ionizing radiation date back to the 1950s, when Day and Stein investigated radiation-induced chemical changes in gels [1]. Subsequent polymer-gel formulations improved the ability to record three-dimensional dose distributions, but limitations related to oxygen sensitivity and the toxicity of some monomers restricted their handling and broader application. MAGIC gel (Methacrylic and Ascorbic Acid in Gelatin Initiated by Copper) emerged as an alternative with reduced oxygen sensitivity and toxicity, while formaldehyde-modified MAGIC (MAGIC-f) further increased the gel melting point and sensitivity [2].

MAGIC-f acts as a three-dimensional dosimeter because ionizing radiation induces water radiolysis and generates reactive species that initiate monomer polymerization. The resulting changes in the polymer network modify the physicochemical and mechanical properties of the gel in a dose-dependent manner. MAGIC-f also presents acoustic and mechanical properties compatible with ultrasound- based characterization, supporting its use as a tissue- mimicking dosimetric phantom [3]. Previous ultrasound studies have demonstrated that radiation-induced changes in gel phantoms can be detected acoustically and used to investigate absorbed-dose distributions [4].

These mechanical changes can be interrogated through shear-wave-based methods. Magnetoacoustography (MAG) combines ultrasound with magnetic excitation of embedded magnetic nanoparticles to generate mechanical perturbations,

while acoustic radiation force impulse (ARFI) excitation produces propagating shear waves directly with focused ultrasound. The resulting wave propagation can be used to estimate viscoelastic properties and identify regions with different stiffness. Magnetomotive and transient shear-wave approaches have previously been used to characterize tissue-mimicking phantom viscoelasticity [6], and magnetomotive ultrasound has been implemented using programmable ultrasound research interfaces [7]. Shear-wave elastography is a well-established quantitative approach for mapping mechanical properties from shear-wave propagation [8], [9].

Magnetic resonance imaging (MRI) relaxometry provides a complementary and well-established readout for polymer-gel dosimetry. Transverse relaxation measurements quantify T2, or equivalently R2 = 1/T2, which changes as radiation-induced polymerization modifies spin-spin interactions in the gel [5]. Thus, MRI relaxometry provides an independent quantitative reference for assessing the dose-dependent changes detected by ultrasound.

In this study, an original MAGIC-f formulation and a thermally optimized formulation were investigated using two ultrasound elastography approaches: shear-wave elastography (SWE) based on acoustic-radiation-force excitation and a second approach based on mechanical perturbation generated by a shaker coupled to a metallic plate. Magnetoacoustic measurements and MRI R2 relaxometry were also performed. The objective was to determine whether radiation-induced mechanical changes can be spatially identified and reconstructed in three dimensions and to compare their dose dependence with the MRI relaxometry response.

III. Materials and Methods

A. Phantom preparation and optimization—Original MAGIC-f gel was prepared using bovine gelatin, methacrylic acid, ascorbic acid, copper sulfate, deionized water, glass powder as an acoustic scatterer, and formaldehyde. A magnetic layer containing iron-oxide nanoparticles (30 mg/mL stock; 1.75 mL added to 35 mL of gel) was deposited on the phantom surface for magnetoacoustic experiments. To improve thermal stability and glass-powder dispersion, gelatin/agar concentrations were varied. Mechanical testing with a texture analyzer was performed in triplicate for the tested formulations. The formulation selected for subsequent comparison contained 6 percent gelatin and 2.5 percent agar, whereas the original formulation contained 8.3 percent gelatin and no agar.

B. Irradiation—After at least 24 h of refrigerated stabilization, one original and one optimized phantom were irradiated using an X-ray system at CIDRA, University of São Paulo. Irradiation was delivered laterally from opposite faces to reduce depth-dependent attenuation effects. Adjacent regions received 20 and 40 Gy and other phantom with 30 and 50 Gy; the incident beam diameter was approximately 1 cm.

C. Ultrasound elastography: SWE and mechanical perturbation—Ultrasound acquisitions were performed with a programmable Verasonics platform using experimental protocols developed at GIIMUS. Spatial scanning was performed with a 3-mm step, generating eight two-dimensional displacement and shear-wave- velocity maps per investigated region. RF data were processed by cross-correlation. For the SWE measurements, shear waves were generated by acoustic- radiation-force excitation, and the analysis used a region of interest lateral to the excitation focus to track the first propagating shear-wave front. Displacement maps were filtered in space and time, and local shear-wave velocity was estimated from spatially separated displacement signals. For the mechanical-perturbation measurements, a shaker drove a metallic plate with a surface area larger than the lateral face of the phantom, generating a mechanical excitation consisting of a single 1000-Hz cycle at 15 dB. The resulting shear waves were tracked with the same ultrasound acquisition platform and processed using the same spatial mapping criteria. For both modalities, the two-dimensional velocity matrices were concatenated along the scan direction, exported in NIfTI format, and visualized in 3D Slicer.

D. MRI relaxometry—Transverse relaxometry was acquired with a 12-echo spin-echo protocol, with TE values from 17 to 204 ms in 17-ms increments and 1-mm inter-slice spacing. R2 was estimated from the exponential decay $S=S0\ \exp(-R2\cdot TE)$. Circular regions of interest of approximately 8 mm diameter were defined with MicroDicom. Ten intermediate frames were analyzed for non-irradiated, 20-Gy, and 40-Gy regions in eachphantom, and the resulting R2 values are reported as mean±SD. Signal intensities were sampled from circular ROIs with an average diameter of 8 mm. For each spatial frame, the exponential decay was fitted over the selected echoes, and R2 values were extracted from ten intermediate frames for each region. The analysis compared matched non- irradiated, 20-Gy, and 40-Gy regions within each formulation.

IV. Results

A. Phantom optimization—Among the tested recipes, the formulation containing 6 percent gelatin and 2.5 percent agar showed the smallest temperature-dependent variation in the measured elastic response and improved glass-powder dispersion. This formulation was therefore selected as the optimized MAGIC-f phantom for the ultrasound and MRI comparisons. Comparing Figures 1.a. and 1b, it can be observed that the phantom produced with the new agar/gelatin concentration exhibits a greater stiffness contrast between the irradiated and non-irradiated areas.

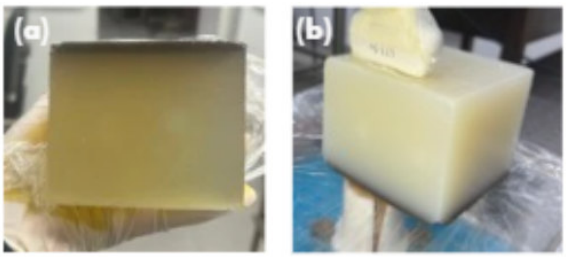


Fig. 1. Original Magic phantom (a) and modified Magic phantom (b) irradiated with 20 Gy and 40 Gy.

B. Shear-wave elastography (SWE)—The 3D reconstruction workflow produced spatially coherent shear-wave velocity maps. In a homogeneous test phantom, the reconstructed mean shear-wave velocity was 5.2±0.5 m/s, supporting the internal consistency of the mapping pipeline. In the optimized MAGIC-f phantom, non-irradiated regions showed velocities of approximately 3.0–3.4 m/s, whereas representative irradiated regions reached approximately 4.4 m/s at 20 Gy and 4.6 m/s at 40 Gy. In the original MAGIC-f phantom, representative values reached approximately 4.3 m/s at 20 Gy and 5.0 m/s at 40 Gy. The irradiated regions could therefore be distinguished from surrounding non-irradiated material by the increase in shear-wave velocity.

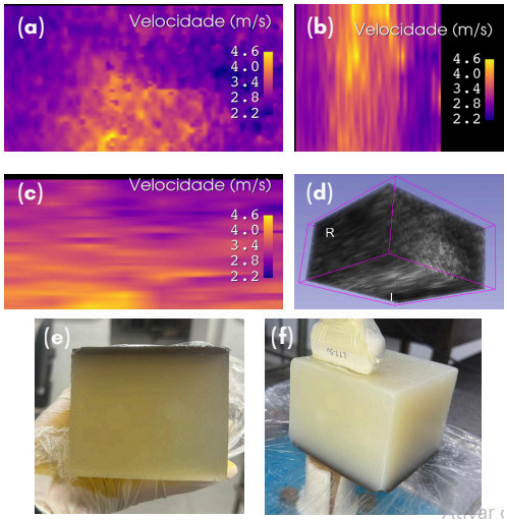


Fig. 2. Shear wave velocity analysis by SWE in **(a)** axial plane image, **(b)** sagittal plane image, **(c)** coronal plane image, and **(d)** 3D image of the irradiated modified Magic phantom with 40 Gy. Original Magic phantom **(e)** and modified Magic phantom **(f)** irradiated with 20 Gy and 40 Gy.

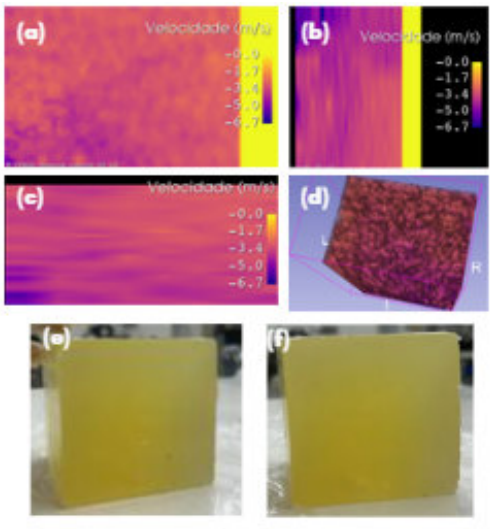


Fig. 3. Shear wave velocity analysis by SWE in **(a)** axial plane image, **(b)** sagittal plane image, **(c)** coronal plane image, and **(d)** 3D image of the irradiated Magic-f phantom with 40 Gy. Original Magic phantom **(e)** and **(f)**.

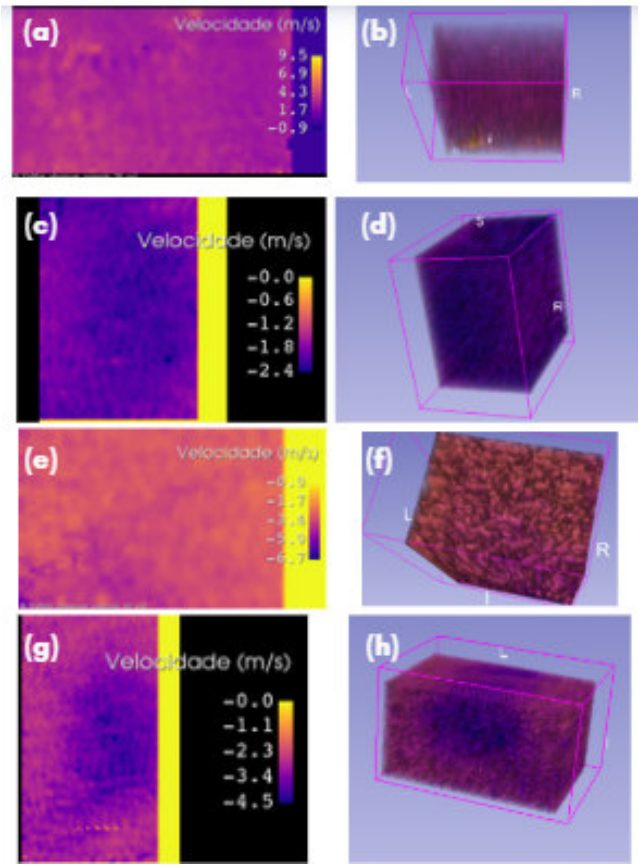


Fig. 4. Shear wave velocity analysis by SWE in axial plane image and 3D image of the irradiated Magic-f phantom with **(a)** **(b)** 20 Gy, **(c)** **(d)** 30 Gy, **(e)** **(f)** 40 Gy and **(g)** **(h)** 50 Gy.

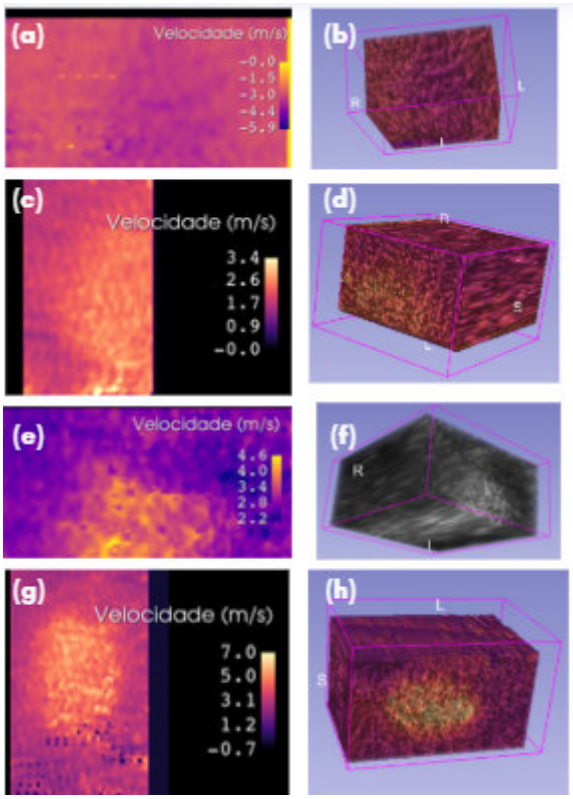


Fig. 5. Shear wave velocity analysis by SWE in axial plane image and 3D image of the irradiated modified Magic phantom with **(a)** **(b)** 20 Gy, **(c)** **(d)** 30 Gy, **(e)** **(f)** 40 Gy and **(g)** **(h)** 50 Gy.

The maps obtained in Figures 4 and 5 also demonstrate that the change in stiffness was more noticeable in the modified phantoms, making them more viable for dosimetric mapping. It can be observed that, in the case irradiated with 50 Gy, the change in shea'r wave velocity from the non-irradiated to the irradiated area is approximately 2.2 m/s for the MAGIC-f phantom and 3.9 m/s for the modified MAGIC-f phantom.

C. Mechanical perturbation with shaker—The shaker- based mechanical perturbation also generated shear waves that enabled spatial differentiation of irradiated and non- irradiated regions. In the optimized MAGIC-f phantom, the 40-Gy region reached approximately 5.4 m/s in the later acquisition, and the velocity difference between the irradiated and non-irradiated regions was approximately 1.4 m/s. The 20-Gy region was less clearly distinguished in this modality. Because the mechanical-perturbation measurements were performed approximately three weeks after the SWE acquisitions, partial phantom dehydration likely affected the absolute shear-wave velocity values. Nevertheless, the irradiated region remained identifiable, supporting mechanical perturbation as a second ultrasound-based elastographic approach for mapping radiation-induced changes.

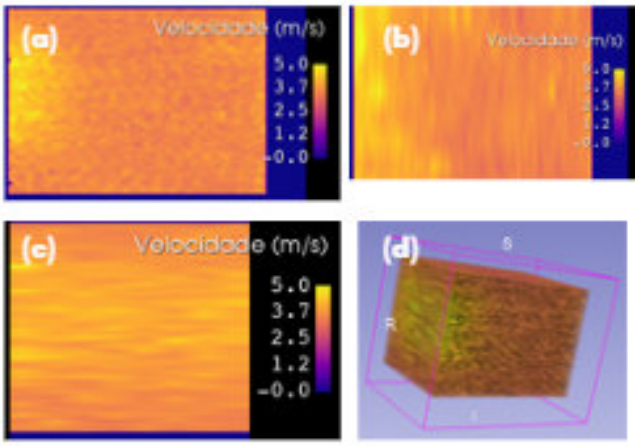


Fig. 7. Shear wave velocity analysis by Mechanical perturbation in **(a)** axial plane image, **(b)** sagittal plane image, **(c)** coronal plane image, and **(d)** 3D image of the irradiated Magic-f phantom with 40 Gy.

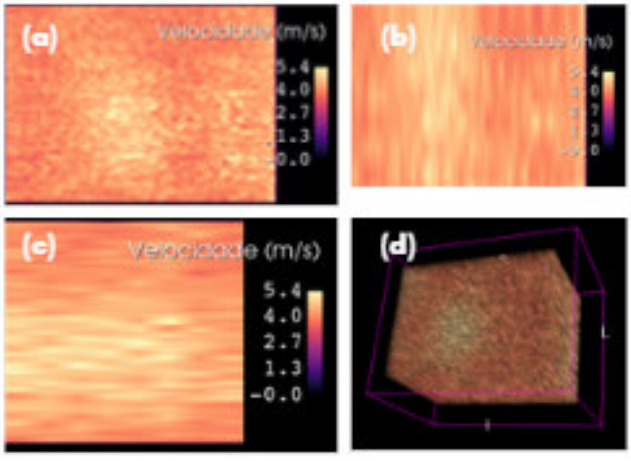


Fig. 6. Shear wave velocity analysis by Mechanical perturbation in **(a)** axial plane image, **(b)** sagittal plane image, **(c)** coronal plane image, and **(d)** 3D image of the irradiated modified Magic phantom with 40 Gy.

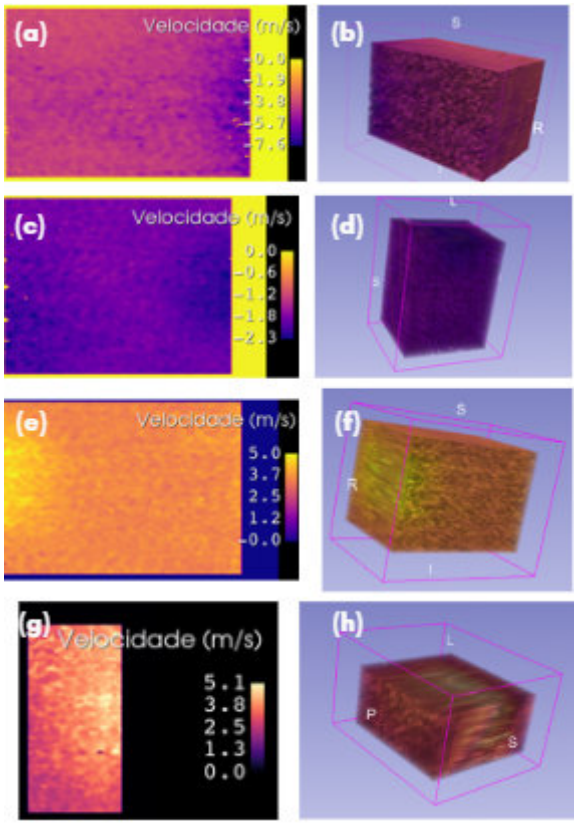


Fig. 8. Shear wave velocity analysis by Mechanical perturbation in axial plane image and 3D image of the irradiated Magic-f phantom with **(a)** **(b)** 20 Gy, **(c)** **(d)** 30 Gy, **(e)** **(f)** 40 Gy and **(g)** **(h)** 50 Gy.

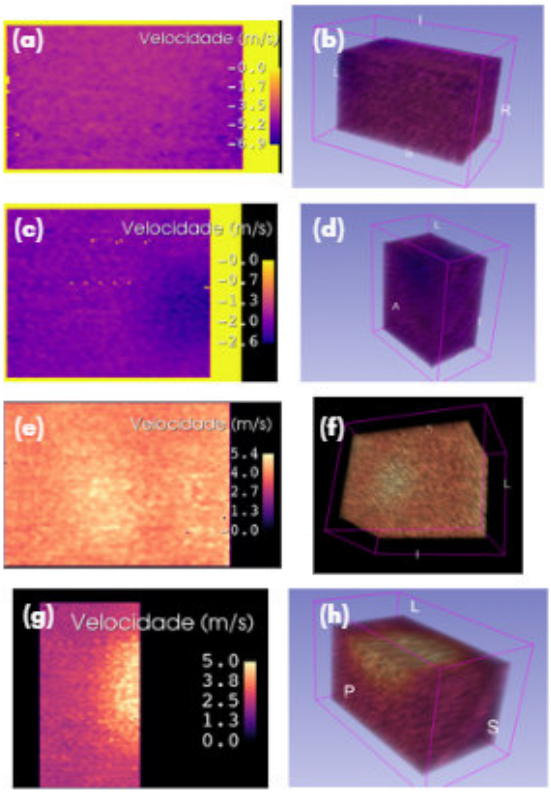


Fig. 9. Shear wave velocity analysis by Mechanical perturbation in axial plane image and 3D image of the irradiated modified Magic phantom with **(a) (b)** 20 Gy, **(c) (d)** 30 Gy, **(e) (f)** 40 Gy and **(g) (h)** 50 Gy.

It was noted that a distinct difference between the irradiated and non-irradiated areas could be seen across all analyzed maps (20, 30, 40, and 50 Gy in the original and modified phantoms).

D. MRI relaxometry—R2 increased with absorbed dose in both formulations (Table I). In the original MAGIC-f phantom, the mean increase was 3.10 1/s from 0 to 20 Gy and 5.33 1/s from 20 to 40 Gy. In the optimized phantom, the corresponding increases were 4.17 and 2.58 1/s. A linear fit to the three mean values yielded R²0.98 for both formulations, indicating a monotonic dose-dependent relaxometric response over the investigated range.

Dose (Gy)	Original	Optimized
0	8.99 ± 0.15	4.79 ± 0.14
20	12.09 ± 0.29	8.96 ± 0.17
40	17.42 ± 0.24	11.54 ± 0.31

Fig. 10. Mean transverse relaxation rate R2 (1/s) from MRI relaxometry (mean±SD).

Across the acoustic-radiation-force measurements, the non-irradiated portions of each phantom remained comparatively homogeneous, while the irradiated zones displayed higher shear-wave velocities. For the optimized formulation, the 40-Gy region reached approximately 4.6 m/s compared with values below about 3.4 m/s outside the irradiated region; at 20 Gy the irradiated region was approximately 4.4 m/s. For the original formulation, representative irradiated-region values were approximately 4.3 m/s at 20 Gy and 5.0 m/s at 40 Gy. These values should be interpreted as representative regional measurements rather than a definitive calibration because the present study used one phantom per formulation and spatially localized regions of interest.

Comparison of the two ultrasound modalities showed that both excitation strategies were sensitive to the radiation- induced mechanical changes, although the absolute shear- wave velocities differed because the acquisitions were not performed at the same time after irradiation. This timing difference should be considered when directly comparing SWE and shaker-based mechanical-perturbation measurements.

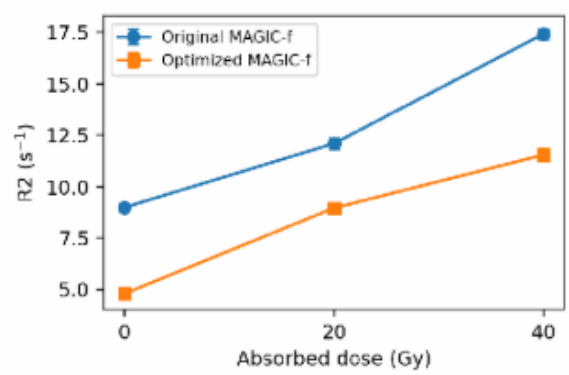


Fig. 11. Mean transverse relaxation rate R2 as a function of absorbed dose for the original and optimized MAGIC-f formulations. Error bars indicate the SD reported across analyzed frames.

V. Discussion

The ultrasound and MRI measurements probe different consequences of the same radiation-induced polymerization process. MRI R2 reflects changes in molecular spin-spin interactions, whereas shear-wave velocity is sensitive to changes in the mechanical stiffness of the polymer network. The dose-dependent increase observed with both readout mechanisms therefore provides complementary evidence that the irradiated gel can be spatially distinguished from non-irradiated material.

The optimized formulation improved thermal stability and produced clearer spatial stiffness contrast in the ultrasound maps. Both SWE and shaker-based mechanical perturbation detected mechanical differences associated with irradiation, demonstrating that the dosimetric information can be interrogated using distinct shear-wave excitation mechanisms. However, absolute shear-wave velocity differed between the two modalities, and the mechanical-perturbation experiment was performed after a delay that allowed partial phantom drying. These results should therefore be interpreted primarily as a feasibility demonstration of spatial dose discrimination rather than as a finalized calibration curve for quantitative

dosimetry. The MRI results provide the strongest quantitative dose dependence in the present dataset. The monotonic increase in R2 at 0, 20, and 40 Gy is consistent with increasing polymerization. In future experiments, a larger number of dose levels and repeated independent phantoms should be used to establish calibration equations, uncertainty, sensitivity, and reproducibility. Quantitative co-registration between R2 and shear-wave-velocity maps would also allow direct voxel-wise comparison of magnetic and mechanical dose surrogates. The present magnetoacoustic configuration was limited by excitation geometry. A coil with a lateral field-of-view matched to the phantom should improve the ability to detect the curvature and propagation of magnetically induced shear waves. Despite this limitation, the ultrasound-based 3D reconstruction pipeline itself was successfully implemented and produced interpretable volumetric maps across the investigated regions. The main limitations are the small number of independent phantoms, the use of only two irradiated dose levels, the lack of synchronized MRI and ultrasound acquisitions, and the time-dependent changes observed in the gel. Accordingly, the present data establish feasibility and dose sensitivity but do not yet define metrological accuracy. A future calibration study should include repeated phantom preparation, denser dose sampling, blinded region selection, uncertainty propagation, and direct registration of R2 and shear-wave-velocity volumes.

VI. Conclusions

Radiation-induced changes in MAGIC-f gel were detected by two ultrasound elastography modalities: SWE using acoustic-radiation-force excitation and mechanical perturbation generated by a shaker; as well as by MRI transverse relaxometry. The optimized formulation containing 6 percent gelatin and 2.5 percent agar improved thermal stability and supported volumetric ultrasound reconstruction. Representative shear-wave velocities increased in irradiated regions with both ultrasound excitation approaches, while mean R2 increased monotonically from 0 to 20 and 40 Gy in both formulations. These results support the feasibility of ultrasound-based 3D mapping of radiation-induced mechanical changes in MAGIC-f gel, with MRI relaxometry providing an independent quantitative reference. Further work should establish multi-dose calibration curves, repeatability, uncertainty, and spatial co-registration between the modalities.

VII. References

[1] M. J. Day and G. Stein, “Chemical effects of ionizing radiation in some gels,” Nature, 1950, doi: 10.1038/166146a0.

[2] J. P. Fernandes, B. F. Pastorello, D. B. De Araujo, and O. Baffa, “Formaldehyde increases MAGIC gel dosimeter melting point and sensitivity,” Phys. Med. Biol., 2008, doi: 10.1088/0031-9155/53/4/N04.

[3] S. L. Vieira, A. Baggio, R. R. Kinnick, M. Fatemi, and A. A. O. Carneiro, “Acoustic images of gel dosimetry phantoms,” Phys. Procedia, vol. 3, no. 1, pp. 83–88, 2010.

[4] S. L. Vieira, R. R. Kinnick, A. L. Baggio, P. Nicolucci, M. Fatemi, and A. O. Carneiro, “Evaluation of vibro-acoustography techniques to map absorbed dose distribution in irradiated phantoms,” in Conf. Proc. IEEE Eng. Med. Biol. Soc., 2009, pp. 796–799.

[5] P. M. Fong, D. C. Keil, M. D. Does, and J. C. Gore, “Polymer gels for magnetic resonance imaging of radiation dose distributions at normal room atmosphere,” Phys. Med. Biol., vol. 46, no. 12, pp. 3105–3113, 2001.

[6] T. W. J. Almeida, D. R. T. Sampaio, A. C. Bruno, T. Z. Pavan, and A. A. O. Carneiro, “Comparison between shear wave dispersion magneto motive ultrasound and transient elastography for measuring tissue-mimicking phantom viscoelasticity,” IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 62, no. 12, pp. 2138–2145, 2015.

[7] D. R. T. Sampaio, T. W. J. Almeida, T. Z. Pavan, and A. A. O. Carneiro, “Implementation of a magneto motive ultrasound system using an ultrasonic research interface,” 2014.

[8] “Shear Wave Elastography - an overview,” ScienceDirect Topics, accessed Jun. 6, 2025.

[9] M. S. Taljanovic et al., “Shear-Wave Elastography: Basic Physics and Musculoskeletal Applications,” Radiographics, vol. 37, no. 3, pp. 855–870, May 2017.

VIII. Acknowledgments

The authors acknowledge CNPq/PIBITI, GIIMUS, CIDRA, FAPESP, and the University of São Paulo for research support and infrastructure.

A real-time evaluation of High-Density EMG signal quality

João Pedro S. Nonato
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brasil
joaopedrosoaresnonato@usp.br

Gustavo Fernandes B. Bergamo
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brasil
gustavofbergamo@usp.br

Louka Skoro
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brasil
Ecole d'ingénieurs Denis Diderot (EIDD)
Université Paris Cité
louka.skoro@etu.u-paris.fr

Daniel G. Tiglea
Lab. de Comunicações e Sinais (LCS)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brasil
daniel.tiglea@usp.br

Renata V. de Freitas
Lab. de Engenharia Biomédica (LEB)
Escola Politécnica
Universidade de São Paulo (USP)
São Paulo, Brasil
renata.valfreitas@usp.br

***Abstract*—This paper presents a pipeline to indicate the increase in signal power during High Density Electromyography (HD-EMG) experiments in comparison with a resting baseline. It is designed to provide a real-time per-electrode visualization of signal quality, enabling relative identification of noise levels and faulty contacts during acquisition, which is done by displaying an 8x8 heatmap. The proposed pipeline was implemented in Python and tested by applying synthetic signals on real HD-EMG acquisition hardware.**

***Index Terms*—Biomedical signal processing, Heat Map, High-Density Electromyography, Quality estimation, Real-Time Signal Processing**

I. Introduction

Evaluation of motor unit recruitment strategies in humans has long been constrained by the spatial limitations of conventional surface electromyography. Classical configurations rely on a limited number of channels, capturing only a spatially averaged representation of the electrical activity over a broad area. Thus, the rich spatial heterogeneity of motor unit recruitment recorded across the muscle surface is lost [1]. High-Density Electromyography (HD-EMG) overcomes this limitation by using a two-dimensional electrode array with tiny contacts. This tool has opened new possibilities in the study of motor control, fatigue, and neuromuscular pathologies [2]. However, the lack of real-time signal quality feedback remains a critical weakness, as poor electrode contact or excessive noise often goes undetected until post-processing, potentially compromising an entire experimental session. Signal contamination in EMG recordings can arise from multiple sources, including baseline noise, power-line interference, motion artifacts, and electrocardiographic crosstalk, each requiring distinct detection and reduction strategies [3]. This issue is particularly severe in muscles with lower EMG signal amplitudes, such as the *soleus* — a primary component of the *triceps surae*, with a key role in postural control [4] [5]. Although online signal quality assessment for multichannel surface electromyography was previously addressed [6], estimation of the power per channel with an ready to use software is still not available. Besides, related work on HD-EMG quality assessment remains predominantly offline. Rojas-Martínez et al. [7] developed algorithms to suppress low-quality signals based on the comparison of adjacent channels. Even real-time decomposition methods rely on parameters learned in a prior offline session [8]. Thus, the development of a dedicated software pipeline capable of providing tools with a real-time visual feedback on signal-to-noise ratio and channel-by-channel quality allows acquisition issues to be identified and corrected during the experiment. The proposed software monitors signal quality directly during acquisition, with real-time per-channel signal power and noise power estimation and a heat map with the information of the muscle activation.

The outline of this paper is organized as follows. Section II describes the acquisition hardware and the estimation pipeline. Section III presents the validation results and Section IV discusses the platform's results, robustness, limitations, and future improvements.

II. Materials and methods

A. Acquisition Hardware and Software

A typical acquisition setup with all its components is illustrated in Fig. 1. In this article, only the red dotted-line area is of interest. The main steps involved in the acquisition of

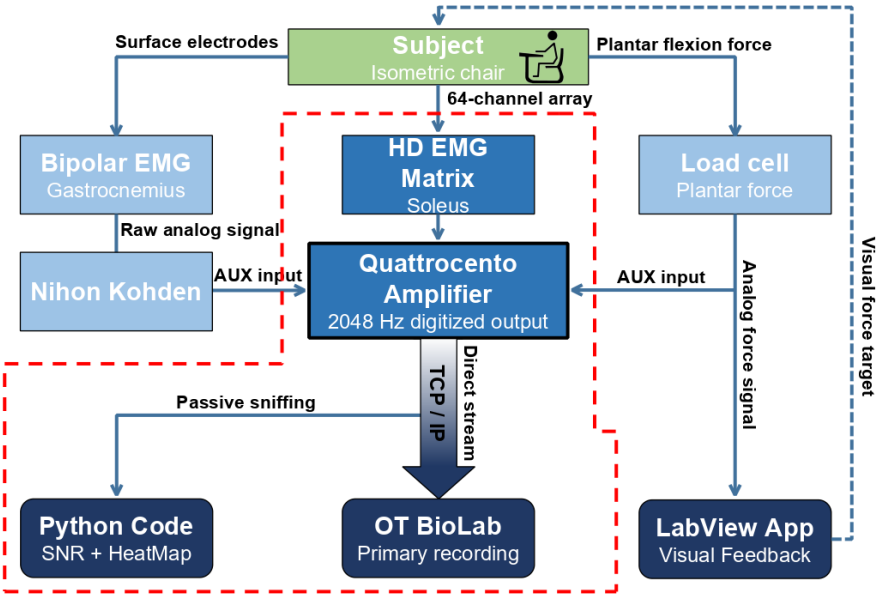


Fig. 1. Block diagram of the real-time HD-EMG acquisition and signal quality monitoring pipeline.

soleus muscle signals are summarized below. Before electrode placement, the skin must be prepared by gentle abrasion to remove dead skin cells, followed by the application of conductive paste to the electrode array to improve electrical contact. The electrode array and reference electrodes must then be positioned according to the SENIAM recommendations to ensure accurate anatomical placement and a standardized acquisition procedure [9].

To elicit the maximal voluntary contraction (MVC), verbal encouragement should be provided during trials to incentivize the subject and maximize force production [10]. During these trials, a specific joint alignment must be maintained to ensure the isolation of the target muscle group. The signals are acquired at a sampling frequency (f_s) of 2048 Hz. In the digital processing stage, a band-pass filter at 10 Hz and 500 Hz with 3-dB cutoff frequencies is applied.

For signal acquisition, a 64-channel gold-coated electrode matrix is connected to a Quattrocento amplification system (OT Bioelettronica) via a flat cable and an AD64 adapter. The adapter enables a monopolar configuration with an auxiliary reference electrode, providing an instrumentation gain of 5 and a hardware-level high-pass filter (cutoff frequency of approximately 0.7 Hz) to eliminate the DC component.

Due to the multiple steps involved in the preparation process, identifying setup-related issues and their underlying causes can be challenging. Therefore, a software tool capable of displaying the noise power of each channel is essential for assessing the quality of the experimental setup.

B. Signal Modeling and SNR Formulation

The average power P_{x_i} of a real-valued signal in a finite window is expressed as [11]

$$P_{x_i} = \frac{1}{N}\sum_{n=0}^{N-1} x_i^2[n], \tag{1}$$

where x_i denotes the discrete-time signal acquired in channel $i = 1, 2, \cdots, C$ in which C denotes the total number of channels in HD-EMG. Furthermore, $n = 0, 1, \cdots, N-1$ represents the sample index, where N is the total number of time instants.

The acquired signal $x_i[n]$ is modeled as the sum of the physiological signal $s_i[n]$ and a term $r_i[n]$ that incorporates all artifacts, including electronic noise, power line interference (PLI), low-frequency motion artifacts, and electrocardiographic crosstalk [3]. For simplicity, henceforth we shall refer to $r_i[n]$ as "noise". Thus, we may write

$$x_i[n] = s_i[n] + r_i[n]. \tag{2}$$

Substituting the signal composition model from (2) into the average power definition in (1) yields

$$P_{x_i} = P_{s_i} + \frac{2}{N}\sum_{n=0}^{N-1}(s_i[n]\cdot r_i[n]) + P_{r_i}. \tag{3}$$

The theoretical SNR in decibels in channel i is defined as the ratio of signal power to noise power:

$$\mathrm{SNR}_{i_{\mathrm{dB}}} = 10\cdot\log_{10}\left(\frac{P_{s_i}}{P_{r_i}}\right). \tag{4}$$

By defining the term $R_{sr_i} = \frac{2}{N}\sum_{n=0}^{N-1}(s_i[n]\cdot r_i[n])$ as an indicator of cross-correlation between the signal and the noise,

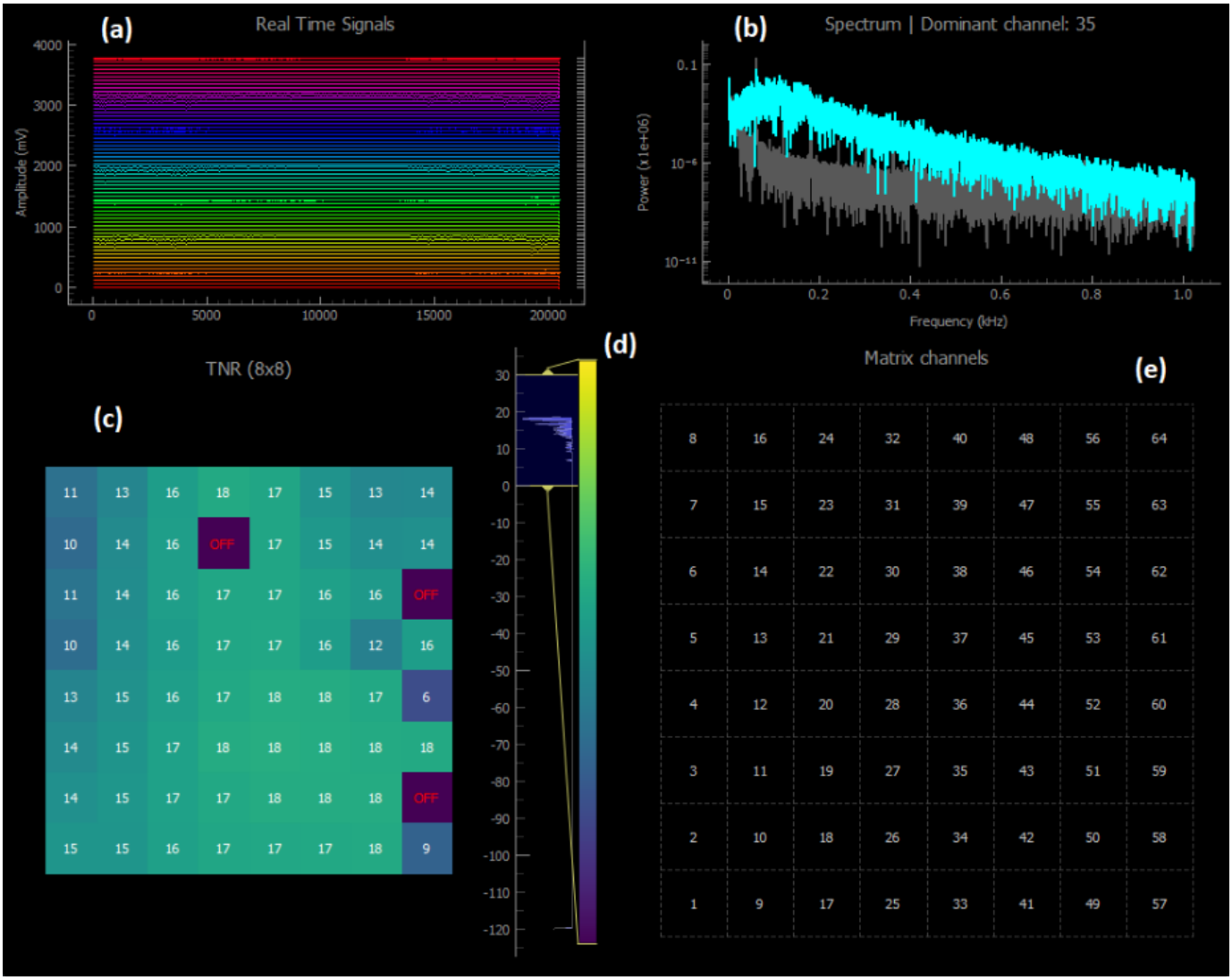


Fig. 2. The dashboard comprises five panels: (top-left: a) real-time oscillograms of all 64 channels; (top-right: b) the power spectral density of the dominant channel, identified as the electrode with higher TNR; (bottom-left: c) an 8×8 TNR heatmap of the electrode matrix, in each square the TNR value is displayed, if a channel is considered faulty, the word OFF is displayed; (bottom-center: d) a draggable colorbar, allowing the operator to manually adjust the min/max range of the TNR(dB)-to-color mapping displayed in the heatmap (c) ; and (bottom-right: e) the channel mapping layout of the matrix for spatial reference.

and combining expressions (3) and (4), the theoretical SNR is expressed as

$$\mathrm{SNR}_{i_{\mathrm{dB}}} = 10 \cdot \log_{10}\left(\frac{P_{x_i} - P_{r_i} - R_{sr_i}}{P_{r_i}}\right). \tag{5}$$

To evaluate the absolute power increase between the resting and active states without assuming any uncorrelation, we introduce the Total-to-Noise Ratio (TNR). This metric computes the ratio of the total short-time windowed power to the estimated noise:

$$\mathrm{TNR}_{i_{\mathrm{dB}}} = 10 \cdot \log_{10}\left(\frac{P_{x_i} - P_{r_i}}{P_{r_i}}\right). \tag{6}$$

By comparing (5) and (6), we note that $\mathrm{SNR}_{i_{\mathrm{dB}}} \approx \mathrm{TNR}_{i_{\mathrm{dB}}}$ if $R_{sr_i} \approx 0$. Thus, if there is little correlation between the physiological signal s_i and the noise r_i, the TNR given by (6) can be viewed as an accurate approximation for the SNR. However, it may be difficult to evaluate the validity of this assumption due to the many components in r_i. Furthermore, it is not straightforward to estimate the cross-correlation. For this reason, we do not take into account and for rigor we use the term TNR instead of SNR.

C. TNR and Power Noise estimation

Before the start of any muscle activation protocol, a mandatory period of resting baseline is recorded with the subject fully relaxed, allowing the software to estimate the per-channel noise used for subsequent TNR computation.

Assuming that the noise $r_i[n]$ is wide-sense stationary (WSS) [12] and that $s_i[n] \approx 0$ during the resting baseline phase, the noise power P_{r_i} remains constant throughout the

experimental session. Therefore, the power noise can be estimated as:

$$\widehat{P}_{r_i} = \frac{1}{K} \sum_{n=0}^{K-1} x_i^2[n], \tag{7}$$

where K is the number of samples in the resting phase.

The total short-time power P_{x_i} during the active contraction phase is continuously tracked over time using a sliding window with L samples for each channel. This result is updated every H samples for better visualization, with $H \ll L$. For the m-th frame, the total measured power in channel i is estimated as:

$$\widehat{P}_{x_i}[m] = \frac{1}{L} \sum_{n=K+mH}^{K+mH+L-1} x_i^2[n], \tag{8}$$

where $m = 0, 1, 2, \ldots$ represents the sequential frame (update) index processed by the real-time thread. In the proposed pipeline, we consider a sliding window with 1 s duration and a 64 ms update interval for the display, corresponding to $L = 2048$ ($L = f_s \cdot \Delta t = 2048 \cdot 1$) and $H = 131$ ($H = f_s \cdot \Delta t \approx 2048 \times 0.064 \approx 131$). Consequently, the practical, time-varying SNR estimator for each frame m is formulated as:

$$\widehat{\mathrm{TNR}}_{i_{\mathrm{dB}}}[m] = 10 \cdot \log_{10} \left(\frac{\max\left\{\varepsilon, \widehat{P}_{x_i}[m] - \widehat{P}_{r_i}\right\}}{\widehat{P}_{r_i}} \right). \tag{9}$$

where ε represents a small positive threshold implemented to prevent computational underflow or negative arguments within the logarithmic function when the estimated noise power exceeds the total power measured in the current frame (e.g., $\varepsilon = 10^{-10}$).

D. Software organization and presentation

As illustrated in Fig. 1 (red dotted area), once the incoming data stream described in Section II-A is captured, parsed, and organized into the 64-channel buffer, the custom software exhibits a real-time visual interface for the operator, which is refreshed at an update rate of approximately 64 ms throughout the acquisition session.

During the mandatory resting baseline phase, the interface remains blank while the proposed pipeline estimates the per-channel noise power according to (7). In this stage, $K = 20480$ samples are accumulated per channel, corresponding to 10 s of resting baseline at the 2048 Hz sampling rate, and the noise power ($\widehat{P}_{r_i}$) is calculated for each channel. Only after this calibration stage is complete does the interface begins displaying data to the operator.

Once the muscle activation protocol starts, the three main graphical elements are presented in Fig. 2: (a) the real-time raw signal traces for all 64 channels; (b) the power spectral density of the channel presenting the highest instantaneous TNR, i.e., the dominant channel; and (c) an 8×8 heatmap displaying the per-channel TNR, computed according to (9).

The spectrum shown in panel (b) is computed via the Fast Fourier Transform (FFT) of the dominant channel over a 5 s sliding window (10240 samples), weighted by a Hamming window prior to the transform to reduce spectral leakage, and updated every 64 ms.

Additionally, during the calibration phase, a channel i is classified as faulty and displayed as "OFF" on the heatmap when its estimated noise power ($\widehat{P}_{r_i}$) exceeds the median noise power across all channels by more than F times their standard deviation. In this work, F was set to 3 without empirical

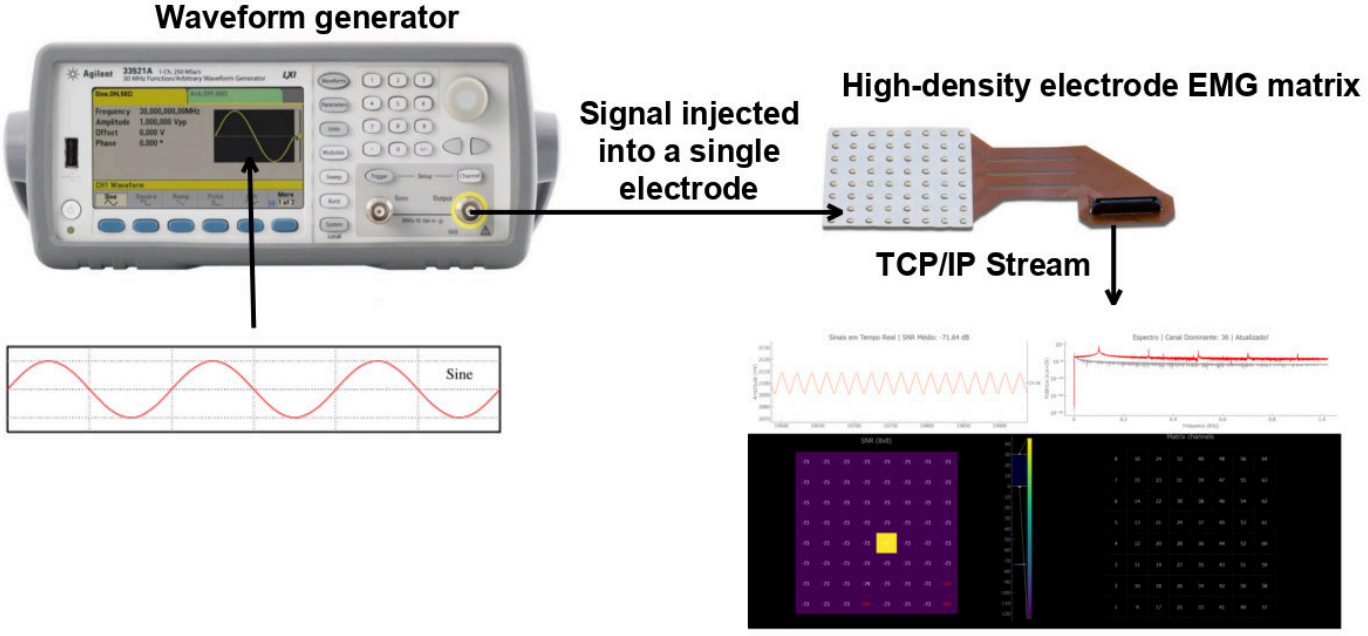


Fig. 3. Experimental setup used for validation of the signal processing pipeline. A deterministic signal was injected into a single channel of the matrix with the waveform generator and the results are displayed in real time.

TABLE I
TNR ERROR AND VPP ERROR BY CHANNEL AND AMPLITUDE

Channel	Amplitude	Mean Abs. Vpp Error	TNR Error (MAE ± SD)	Max Abs. TNR Error
1	10 mV	5.17%	0.46 ± 0.10	0.62
1	1 mV	4.79%	0.52 ± 0.22	0.83
1	20 mV	5.38%	0.47 ± 0.11	0.64
2	10 mV	27.97%	2.97 ± 0.02	2.98
2	1 mV	7.89%	0.96 ± 0.89	1.89
2	20 mV	5.78%	0.67 ± 0.57	1.00
12	10 mV	22.51%	2.12 ± 0.02	2.13
12	1 mV	41.46%	1.99 ± 0.96	2.96
12	20 mV	4.28%	0.48 ± 0.58	1.15
57	10 mV	4.97%	0.44 ± 0.10	0.61
57	1 mV	3.73%	0.36 ± 0.17	0.56
57	20 mV	4.93%	0.44 ± 0.11	0.62
64	10 mV	5.11%	0.18 ± 0.11	0.36
64	1 mV	3.56%	0.14 ± 0.11	0.31
64	20 mV	7.69%	0.44 ± 0.32	1.05

MAE = Mean Absolute Error (dB); SD = Standard Deviation (dB).

validation (see Section IV). This classification is intended as an initial approach for detecting potentially faulty channels and was not further investigated or validated in this work.

E. Pipeline validation

The signal processing pipeline was validated as illustrated in Fig. 3. A waveform generator (Agilent 33521A) was used to inject a sinusoidal signal directly into a single electrode of the 64-channel HD-EMG matrix. During the validation, the remaining electrodes were kept open (high-impedance state), i.e., without electrode-skin contact. A sine wave with peak-to-peak amplitudes of 1, 10, and 20 mVpp and frequencies of 40, 60, 80, 100, 120, and 160 Hz was sequentially injected. For simplicity, rather than detailing all 64 channels, a random subset of five channels (1, 2, 12, 57, and 64) was selected from the matrix for detailed analysis.

For each test, the amplitude and TNR estimated by the real-time processing pipeline were compared with their theoretical values. Since all tested frequencies lie within the passband of the acquisition system, no attenuation due to the acquisition filter was assumed. For a sinusoidal signal with peak-to-peak amplitude $V_{pp} = 2A$, the theoretical signal power was calculated as $P_s = A^2/2$, with $A = V_{pp}/2$. A single resting calibration period was performed before the tests, such that the estimated noise power $\widehat{P}_{r_i}$ remained unchanged throughout the experiments. The theoretical TNR was then calculated from the known signal and noise powers and compared with the value estimated by the pipeline.

III. RESULTS

Table I and Fig. 4 summarize the results obtained during the validation tests. The TNR estimation error is defined as the difference between the pipeline's estimated value and the theoretical expected value.

For most channels, the error remains stable across the dynamic range. Channels 1, 57, and 64 demonstrated close agreement, with a Mean Absolute Error (MAE) for the TNR consistently below 0.5 dB and Vpp relative errors around 4–7%, with their error curves in Fig. 4 following a similar bounded pattern across frequencies. Conversely, channels 2 and 12 exhibited the largest deviations among all tested channels, reaching Vpp and TNR errors of 27.97% and 2.98 dB (channel 2, 10 mV) and 41.46% and 2.96 dB (channel 12, 1 mV), respectively.

IV. DISCUSSION

The proposed pipeline provides real-time calculation of per-channel signal power and noise power and, combined with a qualitative analysis, it helps to detect faulty electrodes.

As shown in Table I, the TNR estimation error is highly channel-dependent: channels 1, 57, and 64 remained close to the theoretical values across the whole tested range, while channels 2 and 12 showed considerably larger errors at the lower amplitudes. However, this gap narrows at the highest amplitude tested (20 mVpp); this suggests that, for these two channels, the estimation becomes more reliable as the injected signal amplitude increases. Indeed, these channels exhibited visible waveform distortion and increased visual noise at lower amplitudes compared to the others.

Although only five channels were tested, no clear dependence of the TNR error on the signal frequency was observed within the 40–160 Hz range. Overall, the results agree with the expected values for the well-behaved channels, and the larger errors observed in channels 2 and 12 are more likely attributable to an intrinsic error of the acquisition hardware itself, specific to those channels, than to a limitation of the proposed TNR estimator. However, since the validation only compared the injected input with the pipeline's final output, without an independent characterization of the Quattrocento amplifier, cables, adapters, and electrode matrices, it is not possible to attribute this discrepancy specifically to the acquisition hardware. Such a characterization is necessary before the software can be validated in more detail under other acquisition conditions.

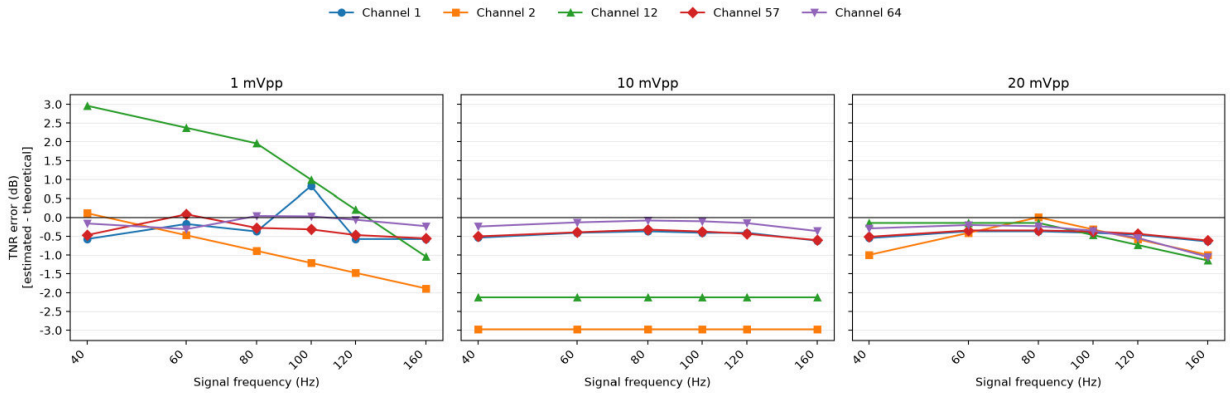


Fig. 4. Total-to-Noise Ratio (TNR) error in dB (estimated minus theoretical value) across different signal frequencies (Hz) for five randomly selected channels (1, 2, 12, 57, and 64) out of the 64-channel matrix. The three panels represent tests performed with peak-to-peak signal amplitudes of 1 mVpp, 10 mVpp, and 20 mVpp, respectively.

Despite its many features, several aspects of the solution presented in this paper can be improved. For example, the robustness of the communication with the Quattrocento still has to be fully evaluated. Moreover, the criterion adopted to classify inactive channels ("OFF") can be further improved to increase its robustness under different acquisition conditions and to indicate the failure mode. Finally, the proposed TNR estimator relies on a calibration stage to estimate the residual noise power. Consequently, variations in background noise after calibration may affect the estimated TNR, suggesting that adaptive noise estimation strategies should be investigated in future work.

V. Conclusion

This work presented a real-time HD-EMG signal quality assessment software capable of estimating channel-by-channel signal power and noise power while simultaneously providing waveform, spectrum, and visual feedback during data acquisition. The experimental results obtained with deterministic test signals revel a dependency on the acquisition system, particularly in the matrix channels.

The current implementation still presents some limitations as discussed in the previous section. Those limitations will be addressed by the development of a validation protocol for the communication and channel classification, a more precise characterization of the noise, and through validation sessions with real subjects and different software users. Nevertheless, in its current form, the proposed pipeline already provides useful feedback on signal quality during HD-EMG acquisition, allowing acquisition issues to be identified before offline processing.

Acknowledgments

The authors would like to thank professors Felipe Fava de Lima and André Fábio Kohn from the Escola Politécnica of the University of São Paulo and professor Leonardo Abdala Elias from the Biomedical Engineering Center, University of Campinas, for their assistance and many helpful comments during the course of the work. The authors gratefully acknowledge TEB – Tecnologia Eletrônica Brasileira Ltda., a Brazilian medical equipment manufacturer, for financial support covering the students' registration and travel expenses. The authors also gratefully acknowledge Amigos da Poli (ADP) for financial support toward the acquisition of laboratory equipment and other supplies that contributed to the development of this work. This project was funded by Capes under Finance Code 001 and in part by FAPESP (Grant Number 2025/05859-7).

References

[1] C. J. De Luca, "The use of surface electromyography in biomechanics," *Journal of Applied Biomechanics*, vol. 13, no. 2, pp. 135–163, 1997.

[2] R. Merletti and D. Farina, *Surface Electromyography: Physiology, Engineering, and Applications*. John Wiley & Sons, 2016.

[3] M. Boyer, L. Bouyer, J.-S. Roy, and A. Campeau-Lecours, "Reducing noise, artifacts and interference in single-channel EMG signals: A review," *Sensors*, vol. 23, no. 6, p. 2927, 2023.

[4] F. H. Netter, *Atlas of Human Anatomy*, 8th ed. Elsevier, 2022.

[5] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for SEMG sensors and sensor placement procedures," *Journal of Electromyography and Kinesiology*, vol. 10, no. 5, pp. 361–374, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641100000274

[6] C. Grönlund, K. Roeleveld, A. Holtermann, and J. S. Karlsson, "Online signal quality estimation of multichannel surface electromyograms," *Medical and Biological Engineering and Computing*, vol. 43, no. 3, pp. 357–364, Jun. 2005.

[7] M. Rojas-Martínez, M. Mañanas, and J. Alonso, "High-density surface EMG maps from upper-arm and forearm muscles," *Journal of Neuro-Engineering and Rehabilitation*, vol. 9, no. 1, p. 85, 2012.

[8] D. Y. Barsakcioglu, M. Bräcklein, A. Holobar, and D. Farina, "Control of spinal motoneurons by feedback from a non-invasive real-time interface," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 3, pp. 926–935, 2021.

[9] SENIAM, "Recommendations for sensor locations on individual muscles," 2020, accessed: 2026-07-03. [Online]. Available: http://seniam.org/sensor_location.htm

[10] Y. Takarada, D. Nozaki, and M. Taira, "Shouting strengthens maximal voluntary force and is associated with augmented pupillary dilation," *Scientific Reports*, vol. 11, no. 1, p. 18190, 2021.

[11] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice-hall Englewood Cliffs, 1999.

[12] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. USA: Prentice-Hall, Inc., 1993.

Análise das alterações espectrais NIRS em comprimidos submetidos a diferentes condições de umidade em cenário de Atenção Domiciliar

1st Maria Fernanda Juncker dos Santos
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
0009-0004-8766-3495

2nd Déborah Evelyn Leite Alcantara
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
debora.evelyna@ufpe.br

3rd Ivy Rodrigues Monteiro Cordeiro
Departamento de Eletrônica e Sistemas
Universidade Federal de Pernambuco
Recife, Brasil
ivy.rodrigues@ufpe.br

4th Elisania da Silva Lopes
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
elisania.lopes@ufpe.br

5th Emery Cleiton Cabral Correia Lins
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
0000-0002-7967-6163

***Resumo*—O estudo avaliou comprimidos acondicionados em organizadores plásticos utilizados por pacientes idosos em regime de atenção domiciliar, incluindo anti-hipertensivos, hipolipemiantes e antidiabéticos. A espectroscopia no infravermelho próximo (NIRS), associada à análise de componentes principais (PCA) foi utilizada para caracterizar alterações espectrais induzidas pela exposição a diferentes condições de umidade. Para isso, as amostras foram acomodadas nos ambientes da cozinha e do banheiro, locais comumente utilizados para guarda de medicamentos por essa população. Os espectros NIR obtidos experimentalmente permitiram comparar as amostras mantidas em cada ambiente, evidenciando variações espectrais associadas às diferentes condições de umidade relativa. Os resultados demonstraram que a abordagem multivariada viabilizou a discriminação espectral entre os grupos expostos aos diferentes ambientes, evidenciando a sensibilidade da técnica às alterações decorrentes da absorção de umidade superficial pelos medicamentos (com os dois primeiros componentes principais explicando $(85,9\%)$ da variância total). Ressalta-se que, para o captopril e a sinvastatina, o uso de fabricantes distintos entre os ambientes introduziu variações na formulação que atuam como fator de confundimento na análise espectral. Em suma, o estudo contribui para a compreensão dos efeitos das condições de armazenamento domiciliar sobre formas farmacêuticas sólidas, explorando alterações espectrais no infravermelho próximo induzidas pela umidade e o potencial da NIRS como ferramenta de monitoramento não invasivo.**

***Palavras-chave*—Espectroscopia de infravermelho próximo, análise de componentes principais, umidade, armazenamento domiciliar de fármacos.**

I. Introdução

Na atenção domiciliar, especialmente para a população idosa, existe uma necessidade maior em combinar soluções que promovam o apoio à vida independente, sabendo que o gerenciamento da medicação é um grande desafio [1]. Nesse contexto, tornou-se comum o uso de organizadores semanais de comprimidos, dispositivos que auxiliam no consumo assertivo dos fármacos na gestão de regimes de farmacoterapia complexos. Contudo, o armazenamento de medicamentos fora de suas embalagens primárias em locais como cozinha e banheiro, motivados pela conveniência, os expõem a condições de temperatura e umidade relativa elevadas, que podem comprometer sua estabilidade [2].

Neste estudo, foram selecionados fármacos de uso comum entre a população idosa: captopril (anti-hipertensivo), sinvastatina (hipolipemiante) e metformina (antidiabético), cuja escolha justifica-se por serem comprimidos sem revestimento, apresentando elevada sensibilidade à degradação por absorção de umidade superficial.

Nesse contexto, a aplicação da espectroscopia no infravermelho próximo (NIRS), técnica analítica rápida, não destrutível e sensível às principais bandas de absorção de água, em conjunto com a PCA permite explorar de forma eficiente as matrizes de dados multivariados e identificar alterações espectrais induzidas pelo ambiente de armazenamento [3]–[5].

II. Metodologia

A seleção dos medicamentos foi baseada no perfil farmacológico característico da população idosa. Entre as condições crônicas mais prevalentes nessa faixa etária estão a hipertensão arterial, o diabetes mellitus tipo 2 e a dislipidemia [6], sendo escolhido um medicamento representativo para cada patologia. Como critério de inclusão, estabeleceu-se que os comprimidos não deveriam possuir revestimento, uma vez que este atua como barreira física à absorção de umidade pelo princípio ativo e pode introduzir bandas espectrais adicionais que interferem na análise por espectroscopia NIR [4], [7]. A escolha de comprimidos sem revestimento justifica-se pela alta sensibilidade que suas formulações apresentam à degradação por absorção de umidade ao longo do tempo. Com base nesses

critérios, foram selecionados: Captopril 25 mg (inibidor de ECA) [8]; Cloridrato de Metformina 850 mg (biguanida) [9]; e Sinvastatina 20 mg (estatina) [10].

Os experimentos foram conduzidos na cidade de Recife (PE), durante o mês de junho de 2026, período de alta umidade relativa do ar média ($\sim 82\%$) [11]. Os ambientes selecionados para a exposição foram o banheiro e a cozinha, por serem os locais do domicílio onde medicamentos são comumente armazenados pela população. Todos os comprimidos foram transferidos para organizadores plásticos compartimentados, não herméticos. As medições foram realizadas em oito pontos temporais ao longo de 20 dias: D0 (dia inicial), D1, D3, D6, D9, D12, D16 e D20, com 3 repetições por grupo, correspondentes a 3 comprimidos distintos.

As medidas espectroscópicas foram realizadas com um espectrômetro NIR miniaturizado DWARF-Star (StellarNet Inc., Tampa, FL, EUA) [5], acoplado a uma fibra óptica bifurcada em Y operada pelo software SpectraWiz Spectrometer OS v5.33 (StellarNet Inc.) [12]. Os parâmetros instrumentais foram: tempo de integração de 800 ms, média de varreduras (*scans to average*) igual a 1 e suavização (*boxcar smoothing*) igual a 1. A calibração foi realizada diariamente no início de cada sessão, utilizando um padrão de reflectância branca de cerâmica. Tanto na etapa de calibração quanto na etapa de medição das amostras, a fibra óptica foi mantida em contato físico direto com a superfície analisada.

Após a calibração dos espectros com o calibrador de cada dia experimental, foi feita a correção da linha de base, por meio do método de Mínimos Quadrados Assimétricos (*Asymmetric Least Squares* – ALS) [13]. Com a linha de base corrigida, os espectros foram submetidos à normalização por área, restritos à faixa espectral compreendida entre 1450 e 1700 nm. Esta região foi selecionada por conter a principal banda de absorção da água no infravermelho próximo, que está associada às vibrações das ligações O–H [4]. Para corrigir pequenas variações nos comprimentos de onda que podem surgir entre aquisições sucessivas, aplicou-se interpolação linear, alinhando todos os espectros em uma mesma grade espectral, garantindo a correspondência entre os pontos amostrados.

Por fim, realizou-se a Análise de Componentes Principais (PCA) [14], concentrando a interpretação nos dois primeiros componentes principais (PC1 e PC2). A análise foi conduzida por meio dos gráficos de Scores e Scree Plot, permitindo avaliar a distribuição das amostras, identificar as regiões espectrais de maior contribuição para a variabilidade observada e acompanhar a evolução temporal da absorção de umidade nos diferentes medicamentos e condições de armazenamento.

Na etapa final da análise, os scores do primeiro componente principal (PC1) obtidos a partir da PCA foram submetidos à análise estatística para verificar a existência de diferenças significativas entre os tempos de exposição para cada combinação medicamento–ambiente. Para essa avaliação, foi empregado o teste não paramétrico de Kruskal–Wallis [15], com nível de significância de 5

Para exclusão de medições discrepantes (*outliers*), adotou-se como critério a inspeção visual dos espectros brutos, descartando leituras que se apresentaram como anomalias evidentes ou com ruídos excessivos, gerando o descarte de 4 espectros antes das etapas de modelagem.

III. Resultados e Discussões

Para a caracterização inicial do sistema de medição, a Fig. 1 apresenta o espectro óptico típico da fonte de luz e do detector, obtido durante a calibração com o padrão de reflectância branca de cerâmica.

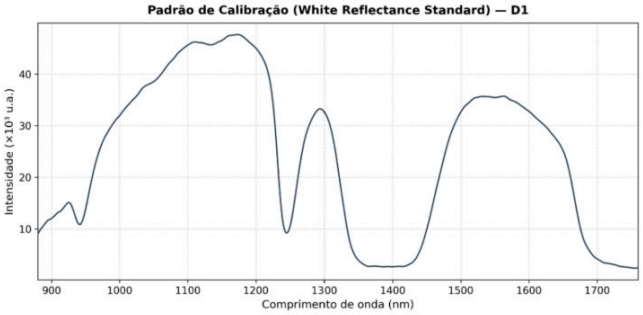


Figura 1. Espectro típico do padrão de calibração (*White Reflectance Standard*) obtido na região de 900 a 1700 nm.

A análise das alterações espectrais provenientes da exposição à umidade foi realizada com foco na faixa de 1450 a 1700 nm, região correspondente ao primeiro sobretom de absorção da ligação O–H. A Fig. 2 compara os espectros obtidos de D0 a D20 para o ambiente de maior variação de cada fármaco, permitindo avaliar o impacto do armazenamento nos organizadores plásticos. Os gráficos presentes nessa figura evidenciam a suavização dos vales característicos ao longo do período de armazenamento, um comportamento particularmente visível nos espectros da Metformina. No D0, o vale localizado próximo a 1500 nm é marcadamente mais pronunciado do que no D20.

A alteração observada na intensidade da banda em ~ 1450 nm entre D0 e D20 confirma a ocorrência de absorção de umidade pela superfície dos comprimidos sem revestimento armazenados nos organizadores plásticos não herméticos ao longo do período estudado.

Na região entre 1427 nm e 1709 nm, os espectros apresentaram comportamentos distintos entre os medicamentos, refletindo suas composições químicas específicas. O Cloridrato de Metformina exibiu comportamento espectral similar entre os ambientes de banheiro e cozinha (o que era esperado, uma vez que os comprimidos eram da mesma marca para ambos os ambientes), com picos bem definidos em 1480 nm, 1585 nm e 1644 nm, além de pequenos vales em 1508 nm e 1613 nm. É importante destacar que nesta região espectral encontra-se a primeira banda de sobretom do estiramento O–H da água em aproximadamente 1450 nm [4], que é de particular interesse para este estudo, pois permitirá monitorar a absorção de umidade ao longo do tempo. A presença de picos adicionais nesta região na metformina pode estar relacionada à sua estrutura molecular característica (biguanida) e aos excipientes presentes na formulação.

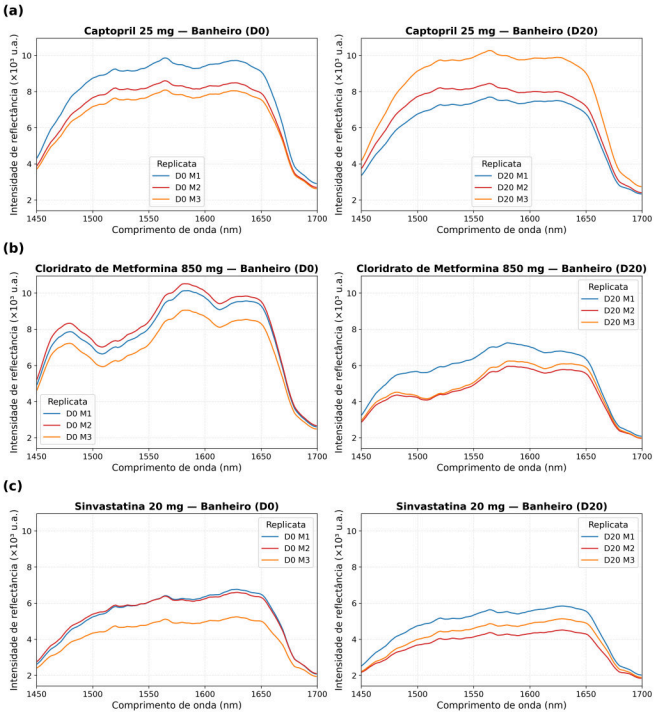


Figura 2. Comparação espectral ilustrativa (D0 vs. D20) na região de absorção de água (1450–1700 nm) para: (a) Captopril 25 mg; (b) Cloridrato de Metformina 850 mg; (c) Sinvastatina 20 mg no ambiente de maior variação.

A Sinvastatina armazenada no banheiro apresentou, na banda entre 1427 nm e 1709 nm, pequenos picos em 1520 nm, 1564 nm e 1629 nm, com uma leve inclinação crescente. Já a sinvastatina armazenada na cozinha exibiu uma banda majoritariamente plana nesta mesma região, com apenas um pico distinto em 1564 nm. O Captopril armazenado no banheiro apresentou comportamento espectral muito similar ao da sinvastatina no mesmo ambiente. Por sua vez, o Captopril da cozinha mostrou a banda entre 1427 nm e 1709 nm também relativamente plana, mas com pequenos picos observáveis em 1518 nm, 1564 nm e 1629 nm. As diferenças espectrais observadas entre os ambientes para captopril e sinvastatina são atribuídas ao uso de fabricantes distintos para cada ambiente, resultando em formulações com excipientes e processos de manufatura diferentes, que influenciam o perfil espectral na região do NIR [18].

Uma variação considerável nas intensidades absolutas foi observada nos espectros, inclusive entre replicatas do mesmo medicamento e ambiente. Esta variabilidade pode ser atribuída a múltiplos fatores: (i) flutuações na intensidade da fonte de luz durante as coletas; (ii) variações no posicionamento da fibra óptica sobre a superfície irregular dos comprimidos, afetando a geometria de iluminação e coleta; (iii) heterogeneidade da superfície dos comprimidos (rugosidade, porosidade, distribuição de partículas); e (iv) possíveis variações na pressão de contato entre a fibra e a amostra [19]. Estas fontes de variabilidade reforçam a importância da aplicação de pré-processamentos espectrais e da normalização dos dados antes das análises quantitativas.

Com o objetivo de investigar especificamente as alterações espectrais relacionadas à absorção de umidade, foi realizada uma PCA utilizando na região de 1450–1700 nm, correspondente ao primeiro sobretom das ligações O–H [4]. Os gráficos de Scree Plot (Fig. 3) demonstram que, em todas as combinações, os dois primeiros componentes principais concentram a maior parte da variabilidade de dados, com PC1 explicando entre 46.5

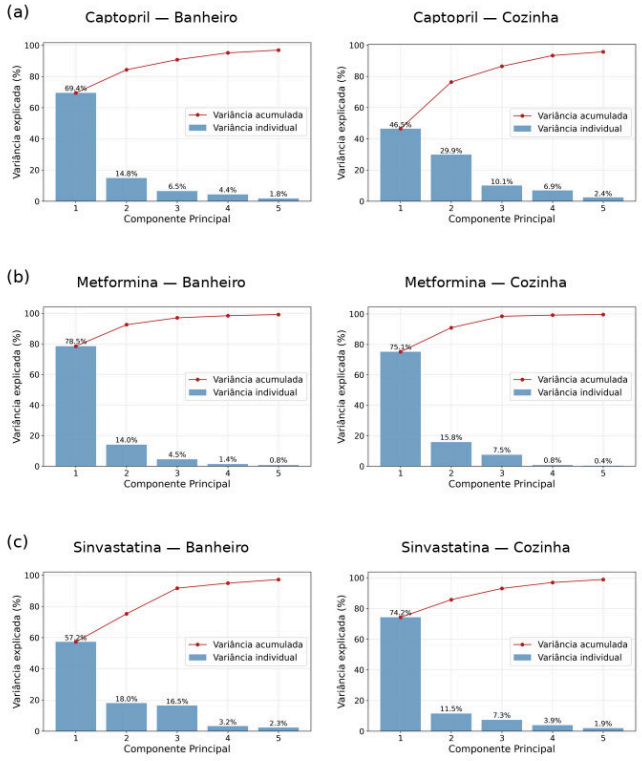


Figura 3. Gráficos de Scree Plot apresentando a porcentagem de variância explicada para: (a) Captopril; (b) Metformina; (c) Sinvastatina.

A Fig. 4 apresenta os gráficos de Scores (PC1 vs. PC2) para cada um dos medicamentos nos ambientes de cozinha e banheiro. PC1 esteve associado às variações globais de intensidade espectral, enquanto PC2 refletiu principalmente diferenças no teor de umidade. Os gráficos de scores não evidenciaram tendência temporal ao longo do armazenamento, embora T0 e T1 tenham apresentado maior dispersão. Além disso, as amostras armazenadas na cozinha exibiram maior variabilidade espectral que as do banheiro, sugerindo maior influência das condições ambientais sobre a região da água.

A análise estatística temporal dos scores de PC1 por meio do teste de Kruskal-Wallis [15] e dos gráficos de boxplots indicou comportamentos distintos entre os fármacos. Para o Captopril, a variação estatística ao longo dos dias foi significativa no banheiro ($p = 0,029$), em contrapartida na cozinha não se mostrou significativa ($p = 0,096$). Para a Sinvastatina, identificou-se alterações significativas em ambos os ambientes ($p = 0,014$ no banheiro e na cozinha). Por fim,

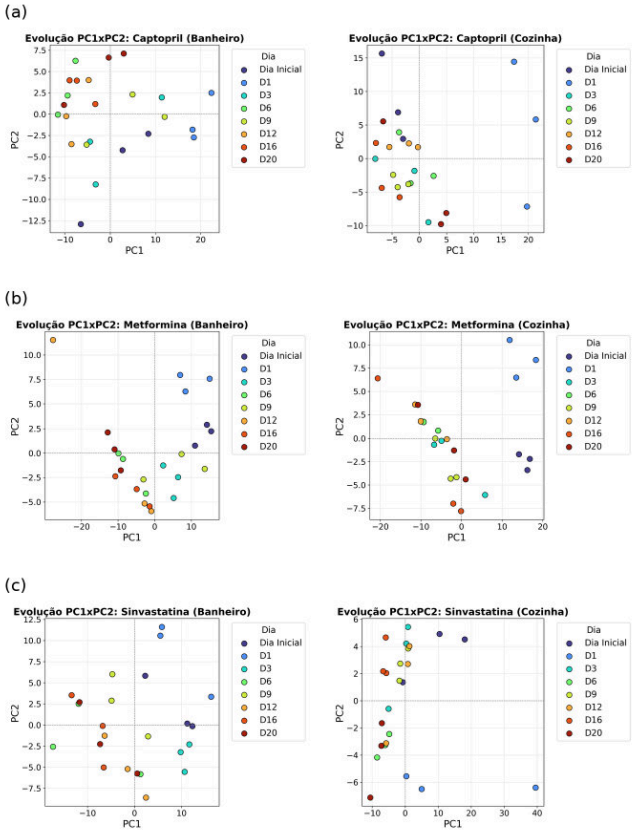


Figura 4. Gráficos de Scores da PCA (PC1 vs PC2) na faixa de 1450 a 1700 nm ao longo dos 20 dias para: (a) Captopril; (b) Metformina; (c) Sinvastatina.

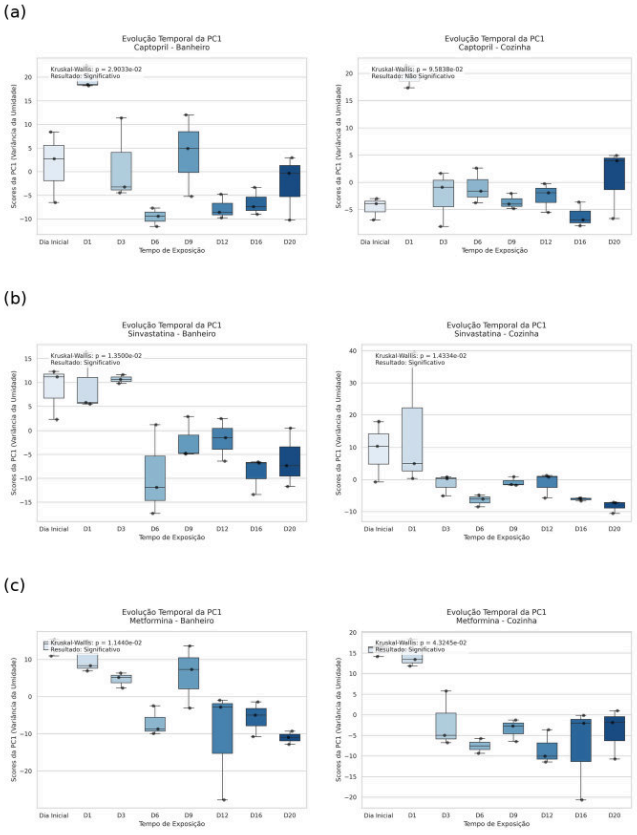


Figura 5. Boxplots da evolução temporal dos scores do primeiro componente principal (PC1) ao longo dos 20 dias para: (a) Captopril; (b) Metformina; (c) Sinvastatina.

a Metformina também apresentou variações significativas no banheiro ($p = 0,011$) e na cozinha ($p = 0,043$). No entanto, a análise conjunta dos boxplots mostra que tais variações refletem flutuações e dispersões pontuais entre as medições, não caracterizando uma tendência progressiva ao longo dos 20 dias.

No caso da metformina, os dois primeiros componentes principais concentraram a maior parte da variabilidade espectral, com PC2 associado a alterações mais localizadas relacionadas à água. Os gráficos de scores apresentaram maior dispersão das amostras, porém sem tendência temporal ao longo do armazenamento. Além disso, não foram observadas diferenças entre os ambientes de armazenamento, indicando maior estabilidade espectral da metformina frente às condições avaliadas.

A análise de PCA na região de absorção da água evidenciou diferenças marcantes entre os fármacos quanto à resposta às condições de armazenamento. A sinvastatina apresentou maior dispersão espectral na cozinha, o captopril mostrou separação entre os ambientes, enquanto a metformina exibiu ampla sobreposição e maior estabilidade espectral. Esses resultados indicam que a influência da umidade depende das características físico-químicas de cada fármaco.

A análise de componentes principais mostrou que a influência das condições de armazenamento sobre a resposta espectral depende da região espectral analisada e das características físico-químicas de cada princípio ativo. De modo geral, o ambiente de armazenamento exerceu maior influência do que o tempo de exposição, evidenciando o potencial da PCA para identificar as principais fontes de variabilidade em estudos de estabilidade farmacêutica [20].

Entre os fármacos, a sinvastatina apresentou maior sensibilidade às condições de armazenamento, com clara separação entre os ambientes, resultado compatível com sua conhecida suscetibilidade à degradação por temperatura e umidade. A metformina manteve elevada estabilidade espectral, sem diferenças consistentes relacionadas ao tempo ou ao ambiente. O captopril apresentou comportamento intermediário, com influência da umidade mais evidente na região da água, em concordância com sua susceptibilidade à degradação oxidativa.

IV. Conclusão

O estudo avaliou como os comprimidos de captopril, metformina e sinvastatina comportaram-se quando armazenados em organizadores semanais e expostos a diferentes ambientes, cozinha e banheiro, por meio da resposta espectral. Para essa avaliação, combinou-se a técnica NIRS com a PCA. Os resultados mostraram que a metformina, único fármaco de mesmo fabricante avaliado em ambos os locais, apresentou

elevada estabilidade espectral, sem diferenças temporais ou entre ambientes. Já para a sinvastatina e o captopril, observou-se uma clara separação espectral entre os ambientes, no entanto, este resultado pode ser reflexo das diferenças de formulação decorrentes do uso de fabricantes distintos entre os locais, não permitindo isolar apenas o efeito do ambiente.

A Análise de Componentes Principais (PCA) focada na faixa espectral de 1450 a 1700 nm permitiu separar as mudanças ligadas à absorção de água.

As descobertas reforçam a preocupação com a prática de guardar medicamentos em locais úmidos, algo frequente entre idosos que recebem cuidados em casa, a preocupação aumenta quando se trata de medicamentos mais sensíveis, como a sinvastatina. Nesse contexto, NIRS demonstrou potencial para monitorar a absorção de umidade pelos medicamentos, destacando-se por ser uma ferramenta rápida, não destrutiva e não invasiva.

Como limitações metodológicas, destacam-se a ausência de amostras controle mantidas em suas embalagens primárias originais [2] e o uso de dados de umidade relativa baseados na média climatológica local ($\sim 82\%$) [11]. Além disso, a metodologia e técnicas aplicadas não permitiram medir diretamente a degradação química dos fármacos. Para estudos futuros, pode-se considerar combinar a NIRS com técnicas de referência, como o HPLC.

REFERÊNCIAS

[1] O. Salmensuu, J. Isotalo, M. Rijken, V. Hyttinen-Huotari, M. Kaarakainen, and I. Linnosmaa, "Effects of using medication reminder technologies by home-dwelling older citizens: a systematic review,"*Age and Ageing*, vol. 55, no. 1, art. no. afae200, 2026, doi: 10.1093/ageing/afae200.

[2] C. Hewson, C. C. Shen, C. Strachan, and P. Norris, "Personal medicines storage in New Zealand,"*J. Prim. Health Care*, vol. 5, no. 2, pp. 146–150, 2013.

[3] S. H. F. Scafi and C. Pasquini, "Identification of counterfeit drugs using near-infrared spectroscopy,"*The Analyst*, vol. 126, no. 12, pp. 2218–2224, 2001.

[4] J. Workman and L. Weyer, *Practical Guide to Interpretive Near-Infrared Spectroscopy*. CRC Press, 2007.

[5] StellarNet Inc., "DWARF-Star Miniature NIR Spectrometer,"2026. [Online]. Available: https://www.stellarnet.us/spectrometers/nir-spectrometers/dwarf-star-miniature-nir-spectrometer/.

[6] F. C. D. Andrade et al., "Prevalência de doenças crônicas em octogenários: dados da Pesquisa Nacional de Saúde 2019,"*Ciência & Saúde Coletiva*, vol. 26, no. 6, pp. 2221–2230, 2021, doi: 10.1590/1413-81232021266.38822020.

[7] E. Mwesigwa and A. W. Basit, "An investigation into moisture barrier film coating efficacy and its relevance to drug stability in solid dosage forms,"*Int. J. Pharm.*, vol. 497, no. 1-2, pp. 70–77, 2016.

[8] A. Sharma et al., "Captopril,"in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[9] N. Corcoran and T. F. Jacobs, "Metformin,"in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[10] A. A. R. Conde et al., "Simvastatin,"in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[11] Climate-Data.org, "Clima Recife (Brasil),"2026. [Online]. Available: https://en.climate-data.org/south-america/brazil/pernambuco/recife-5069/.

[12] StellarNet Inc., "SpectraWiz Spectrometer OS Software,"2026. [Online]. Available: https://www.stellarnet.us/software/#SPECTRAWIZ.

[13] P. H. C. Eilers and H. F. M. Boelens, "Baseline correction with asymmetric least squares smoothing,"Leiden University Medical Centre Report, 2005.

[14] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments,"*Philos. Trans. R. Soc. A*, vol. 374, no. 2065, p. 20150202, 2016.

[15] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis,"*J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.

[16] W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. New York: Wiley, 1999.

[17] H. Wickham and L. Stryjewski, "40 years of boxplots,"*Am. Stat.*, vol. 65, no. 3, pp. 180–187, 2011.

[18] Y. Roggo, C. Roeseler, and M. Ulmschneider, "Near infrared spectroscopy for qualitative comparison of pharmaceutical batches,"*J. Pharm. Biomed. Anal.*, vol. 36, no. 4, pp. 777–786, 2004.

[19] J. R. Johnson and M. S. Robinson, "The surface roughness of (433) Eros as measured by thermal-infrared beaming,"*Icarus*, vol. 180, no. 2, pp. 463–474, 2006.

[20] B. Li, Y. He, and H. Lin, "Evaluation of drug stability using NIR spectroscopy and chemometrics,"*J. Pharm. Innov.*, vol. 4, pp. 112–119, 2009.

Automated Obstructive Sleep Apnea Detection Using Reduced Polysomnography Channels

A. M. S. Delfino
Department of Electrical Engineering
UNESP
Guaratinguetá, Brazil
https://orcid.org/0009-0007-2085-3851

B. T. Francis
Department of Electrical Engineering
UNESP
Guaratinguetá, Brazil
beatriz-thomsen.francis@unesp.br

C. D. Maciel
Department of Electrical Engineering
UNESP
Guaratinguetá, Brazil
https://orcid.org/0000-0003-0137-6678

Abstract*—Obstructive Sleep Apnea (OSA) is a highly prevalent sleep disorder linked to severe cardiovascular diseases and cognitive decline. Traditional diagnosis via manual polysomnography (PSG) is expensive, access-restricted, and labor-intensive. To enable scalable and less invasive screening, this study proposes an automated OSA detection pipeline utilizing a reduced subset of standard physiological channels from a real-world clinical dataset: nasal airflow (FLOW), pulse oximetry (SpO_2), thoracic effort (THX), and abdominal effort (ABD). We design and systematically evaluate three lightweight hybrid Deep Learning topologies that combine One-Dimensional Convolutional Neural Networks (1D-CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, contrasting expanding, contracting, and multi-scale/parallel spatial feature structures. Models are trained using a robust Focal Loss function and optimized via an independent patient-based cross-validation protocol to ensure true clinical generalizability. Our experimental results demonstrate that the multi-scale parallel architecture (Topology 3) achieved the highest overall performance, yielding a Cohen's Kappa of 0.5996, a Macro F1-score of 0.7996, and an AUC-ROC of 0.8843, while maintaining an accuracy of 81.03%. This parallel design offered the most compact alternative evaluated, reducing trainable parameters by up to 86% compared to deeper serial variations with no trade-off in diagnostic robustness. These findings suggest that integrating concurrent, multi-scale convolutional branches within lightweight networks represents an efficient alternative to over-parameterized structures, providing an accurate and computationally efficient mechanism for automated nocturnal diagnostic screening systems that use raw, multi-channel physiological streams.

***Index Terms*—Obstructive Sleep Apnea; Deep Learning; CNN-BiLSTM; Focal Loss; Reduced PSG Channels; Patient-Independent Validation.**

I. Introduction

Obstructive sleep apnea (OSA) is a highly prevalent respiratory disorder characterized by recurrent partial or complete cessations of breathing during sleep. These episodes trigger systemic drops in blood oxygen saturation, frequent microarousals, and severe sleep fragmentation [1], [2]. Over time, these repetitive respiratory disruptions induce chronic inflammation and autonomic nervous system dysregulation, significantly compromising cardiovascular, metabolic, and cognitive functions [3]. It is estimated that approximately 23% of men and 17% of women suffer from OSA, though these figures may be substantially higher when considering undiagnosed cases [4]. Beyond degrading sleep quality and inducing profound daytime fatigue, OSA is strongly correlated with hypertension, cardiac arrhythmias, and an elevated risk of cardiovascular mortality, marking it as a critical global healthcare challenge [5].

Financially, OSA imposes a staggering macroeconomic and societal burden across direct, indirect, and intangible dimensions [6]– [8]. Direct costs encompass diagnostic polysomnography, continuous positive airway pressure (CPAP) therapies, and specialized clinical consultations, with severe cases requiring up to $5,186 annually per patient in Europe and over $32,500 in the United States [6], [7]. Indirect expenses are driven by profound workplace productivity losses—amounting to roughly $87 billion annually in the U.S., alongside $26 billion from traffic or occupational accidents and $30 billion from secondary comorbidities such as diabetes and stroke [9]. Furthermore, intangible costs reflecting diminished quality of life and psychological distress add an estimated $10 billion annually [7]. At the individual level, a patient can expect mean out-of-pocket expenses of approximately $2,100 per year, demonstrating that robust screening and early intervention possess immense potential to alleviate these global economic impacts [6]– [8].

To mitigate these systemic impacts, full nocturnal polysomnography (PSG) remains the diagnostic gold standard, capturing concurrent physiological signals including pulse oximetry (SpO_2), respiratory inductance plethysmography (RIP) for thoracic and abdominal effort, and nasal cannula airflow [10], [11]. However, manual visual scoring of these multi-channel recordings is labor-intensive, time-consuming, and subject to considerable inter-scorer variability [12]. Consequently, artificial intelligence (AI)-based automated screening systems have attracted increasing attention in recent years [13]. Early approaches relied primarily on handcrafted feature engineering combined with conventional machine learning algorithms [14], [15]. More recently, these approaches have progressively been replaced by end-to-end deep learning (DL) models capable of learning discriminative representations directly from raw physiological signals [16]– [18]. Current state-of-the-art solutions frequently employ convolutional neural networks (CNNs), recurrent neural networks (RNNs/LSTMs), and hybrid CNN-LSTM architectures to improve diagnostic performance while reducing manual preprocessing require-

ments [19]– [21].

Despite these advances, important physiological and architectural limitations remain. Most existing approaches either process each physiological channel independently or implicitly assume perfect temporal synchrony among respiratory signals [22], [23]. In practice, obstructive respiratory events are intrinsically asynchronous. Airflow cessation and thoracic or abdominal respiratory effort changes occur immediately after upper airway collapse, whereas peripheral oxygen desaturation (SpO_2) typically appears only after a physiological delay of approximately 10–40 seconds due to circulatory dynamics [24]. Ignoring these temporal offsets forces neural networks to analyze partially misaligned physiological events, potentially introducing spurious correlations, reducing predictive performance, and limiting physiological interpretability [25], [26].

Although hybrid CNN-LSTM and CNN-BiLSTM models have become increasingly popular for automated OSA detection, most reported architectures adopt a single serial convolutional feature extractor before temporal modeling [19]–[21], [34], [35]. Recurrent layers are naturally suited to capture long-term temporal dependencies and have therefore become the predominant choice for modeling respiratory dynamics in OSA detection. However, their effectiveness strongly depends on the quality of the representations extracted by the preceding convolutional stage. For example, Oh *et al.* [34] and Chang *et al.* [35] employ conventional serial CNN-BiLSTM pipelines composed of a single convolutional pathway, while Zhang *et al.* [19] similarly adopts a fixed convolutional hierarchy. Consequently, although recurrent modeling has received considerable attention, the influence of convolutional feature organization itself on the ability of BiLSTM networks to learn delayed physiological interactions has received comparatively little investigation. To the best of our knowledge, few studies have systematically compared expanding, contracting, and multi-scale parallel convolutional organizations under identical experimental conditions to isolate the effect of convolutional organization on respiratory representation learning.

From another perspective, another major challenge arises from the characteristics of real-world clinical datasets. Physiological recordings frequently contain sensor artifacts, motion disturbances, behavioral fluctuations, and instrumentation instabilities [31], [32]. Such variability may encourage neural networks to learn non-physiological patterns, reducing generalization capability and increasing false-positive detections [33]. Therefore, developing architectures capable of extracting robust physiological representations while maintaining computational efficiency remains an important research challenge.

Taken together, these limitations reveal three major gaps in current deep learning approaches for OSA detection. First, asynchronous temporal relationships among respiratory physiological signals remain insufficiently explored, despite the suitability of recurrent architectures for modeling long-range temporal dependencies. Second, although CNN-BiLSTM models have become widely adopted, previous studies typically employ only a single convolutional organization, leaving the influence of feature extraction strategy on temporal representation learning largely unexplored. Third, real-world clinical variability demands compact models capable of learning robust physiological representations without excessive computational complexity. Therefore, understanding how different convolutional organizations interact with recurrent temporal modeling represents an important step toward developing more reliable and physiologically consistent OSA detection systems.

To position this work against the state of the art in OSA detection and address the multi-signal processing limitations pointed out in the literature, we propose a topology that goes beyond isolated channel data processing. Historically, OSA detection approaches are divided into two groups: those based on hand-crafted features extracted statically and linearly from multiple signals, such as Tripathy (2018), who analyzes HRV and EDR using Fourier decomposition and KELM [43]; and deep learning models focused on single channels of raw ECG, such as the works of Banluesombatkul et al. (2018) [44] and Zhang et al. (2021) [45], which employ hybrid CNN-LSTM networks to map temporal characteristics [44], [45].

While the first group fails by assuming that the time windows of different physiological sources are strictly synchronized and processes them independently (discarding dynamic cross-correlations) [43], the second discards the rich information of multi-signal approaches by limiting itself to raw ECG signals [44], [45]. To mitigate these issues, the architecture proposed in this article introduces an early multi-channel feature concatenation mechanism. This architectural element directly addresses the synchronization premise by allowing asynchronous alignment and processing of data streams, capturing dynamic and non-linear interactions between signals without requiring identical sampling or perfectly static time windows. A comprehensive structural comparison of these methodologies is summarized in Table I.

TABLE I
COMPARATIVE ANALYSIS OF OSA DETECTION TOPOLOGIES.

Model / Ref.	Inputs	Approach / Sync	Limitations
Tripathy [43]	HRV, EDR	KELM / Independent processing. Assumes perfect sync.	Fails to capture cross-signal correlations; relies on R-peaks.
Banlues. [44]	Raw ECG (15s)	CNN-LSTM / Single-channel. No signal fusion.	Misses respiratory dynamics from other sources.
Zhang [45]	Raw ECG (10s)	Multi-scale CNN / Single-channel.	Highly sensitive to noise and motion artifacts.
Proposed	FLOW, SpO2 ABD, THX	CNN-BiLSTM / Active Asynchronous Fusion.	Higher training complexity and data requirements.

Consequently, the primary objective of this study is to investigate whether different convolutional feature organizations influence the ability of hybrid CNN-BiLSTM networks to learn asynchronous respiratory physiological patterns from reduced polysomnography signals. More specifically, three controlled architectural configurations were designed to evaluate

distinct hypotheses regarding convolutional feature extraction: an expanding serial organization, a contracting serial organization, and a multi-scale parallel organization. All remaining experimental conditions—including dataset composition, optimization strategy, loss function, preprocessing pipeline, and validation protocol—were intentionally kept unchanged, allowing convolutional organization to be isolated as the sole experimental variable. Therefore, beyond identifying the highest-performing architecture, this study aims to provide experimental evidence regarding the role of convolutional feature organization in hybrid CNN-BiLSTM models for automated OSA detection.

The remainder of this paper is organized as follows. Section II describes the materials and methods, including the clinical dataset, preprocessing procedures, evaluation metrics, and the proposed network topologies. Section III presents the experimental results together with their computational and physiological discussion. Finally, Section IV summarizes the principal conclusions and outlines future research directions.

II. Materials and Methods

This section details the comprehensive methodological framework and computational pipeline developed for the automated detection of OSA using a reduced configuration of polysomnography channels. First, the clinical dataset characterization and the data preprocessing strategy—including signal windowing, downsampling, and patient-wise normalization—are presented. Subsequently, the mathematical formulations of the evaluation metrics and the structural configurations of the three proposed hybrid 1D-CNN-BiLSTM network topologies are specified mathematically and architecturally. Finally, the section outlines the network optimization criteria, loss functions, and the strict patient-independent validation protocol employed to ensure model generalizability.

A. Dataset and computational description

The experimental evaluation in this study utilizes the publicly available PhysioNet Apnea-ECG Database (version 1.0.0) [46]. Although the full database comprises 70 overnight recordings, a specific sub-cohort of 8 recordings (*a01–a04*, *b01*, and *c01–c03*) provides continuous multi-parameter physiological streams alongside single-lead ECG. To construct a non-invasive, computationally efficient screening framework tailored specifically for OSA detection, this study isolates a reduced subset of four essential physiological channels from these 8 subjects: nasal airflow (*FLOW*, corresponding to *Resp N*), peripheral oxygen saturation (SpO_2), thoracic respiratory effort (*THX*, corresponding to *Resp C*), and abdominal respiratory effort (*ABD*, corresponding to *Resp A*). Expert clinical annotations indicating the presence or absence of obstructive sleep apnea on a minute-by-minute basis serve as the ground truth for model training and evaluation. Although restricted to 8 multi-parameter subjects, each overnight recording spans approximately 7–8 hours of continuous physiological monitoring. Under the 60-second windowing protocol with 75% overlap, this dataset yields a dense pool of approximately 14,500 temporal segments, providing a substantial continuous stream of apneic and non-apneic respiratory events for pattern extraction.

The computational workflow was developed in Python (v3.12.13). Signal preprocessing and EDF file handling relied on `MNE` [36] and `NumPy` [37]. Neural network architectures, tensor operations, and data loading pipelines were implemented in `PyTorch` [38]. Model evaluation, class balancing, and patient-independent cross-validation splitting (`GroupKFold` and `GroupShuffleSplit`) were executed using `Scikit-Learn` [39]. Visualizations were rendered via `Matplotlib` [40] and `Seaborn` [41], while hyperparameter tuning and experiment tracking were automated using `Optuna` [42] and `MLflow`. Preprocessing was conducted locally on an Intel Core i7-1255U processor (ASUS Vivobook 16), whereas deep neural network training was accelerated on Google Colab using an NVIDIA T4 GPU.

B. Data Preprocessing and Windowing

To transform the continuous multi-channel physiological recordings into an optimal format for deep learning, a structured preprocessing pipeline was implemented. The raw signals were segmented into fixed-length 60-second temporal windows with a 0.75 overlap factor, maximizing the density of training patterns. Given that human respiratory events and localized apneic anomalies manifest at significantly lower frequencies—typically well below 5 Hz—the signals were downsampled from 300 Hz to 16 Hz. This downsampling maintains the structural integrity of the physiological waves according to the Nyquist criterion while drastically reducing the data footprint and computational overhead. Ground truth labels were assigned using a thresholding rule: a 60-second window is classified as an OSA event if active apnea labels occupy at least 17% of its total duration. This threshold corresponds to approximately 10 seconds of continuous obstruction and strictly aligns with the clinical diagnostic criteria established by the American Academy of Sleep Medicine (AASM) [10]. To eliminate inter-patient physiological variance and remove signal outliers, patient-wise normalization was executed using the Z-score standardization method, defined as:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ and σ represent the mean and standard deviation computed exclusively per individual channel per patient.

Normalization was performed independently for each patient's complete overnight PSG recording after preprocessing and before window segmentation. Since the proposed framework operates offline, the entire recording is available during inference. Consequently, the mean (μ) and standard deviation (σ) are estimated exclusively from the signals of the patient being analyzed, ensuring that no information from other subjects is incorporated into the normalization process.

C. Evaluation Metrics

The classification capability of the deep learning topologies was rigorously evaluated using three widely established met-

rics in biological time-series analysis: Area Under the Receiver Operating Characteristic Curve (AUC-ROC), F1-Score, and Cohen's Kappa (κ). These metrics are mathematically justified as optimal indicators for analyzing deep learning models under typical physiological class imbalances [27], [28]. The F1-Score represents the harmonic mean of precision and recall, ensuring that both false positives and false negatives are heavily penalized:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

Cohen's Kappa measures the level of agreement between the predicted labels and the true annotations while explicitly accounting for the possibility of agreement occurring by chance [29]:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o is the observed proportionate agreement and p_e is the hypothetical probability of chance agreement. The AUC-ROC computes the integral of the True Positive Rate against the False Positive Rate across all possible classification thresholds, describing the model's discriminative threshold-agnostic capability. Additionally, standard classification performance indices, namely Accuracy, Sensitivity (Recall), and Specificity, were computed from the accumulated true positives, false positives, true negatives, and false negatives to provide a comprehensive clinical evaluation.

D. Model Topologies and Structural Comparison

This study evaluates and compares various structural variations of a hybrid deep neural network combining one-dimensional convolutional layers with bidirectional recurrent units. The base system relies on 1D-CNN layers to map localized morphological features from the four input streams, followed by a BiLSTM network designed to track long-term asynchronous interactions across time.

Each topology was designed to evaluate a distinct architectural hypothesis. The expanding topology investigates whether progressively increasing representational capacity improves respiratory pattern extraction. The contracting topology evaluates whether a reduced-depth architecture can preserve diagnostic performance while lowering computational complexity. Finally, the parallel multi-scale topology investigates whether simultaneous multi-resolution feature extraction better represents asynchronous physiological events prior to temporal modeling.

To determine the most efficient architecture for OSA prediction, three distinct topological variants were evaluated, as detailed below:

a) Topology 1 (Expanding Serial Structure): As illustrated in Fig. 1, this architecture employs a deep serial extraction pipeline composed of three sequential 1D-CNN layers with expanding channel dimensions (32, 64, and 128 filters, respectively) and decreasing kernel sizes ($k = 11$, $k = 7$, and $k = 5$). This spatial representation is then forwarded to a robust recurrent stage consisting of two stacked BiLSTM layers with 64 hidden units each, optimized to track deep sequential temporal dependencies.

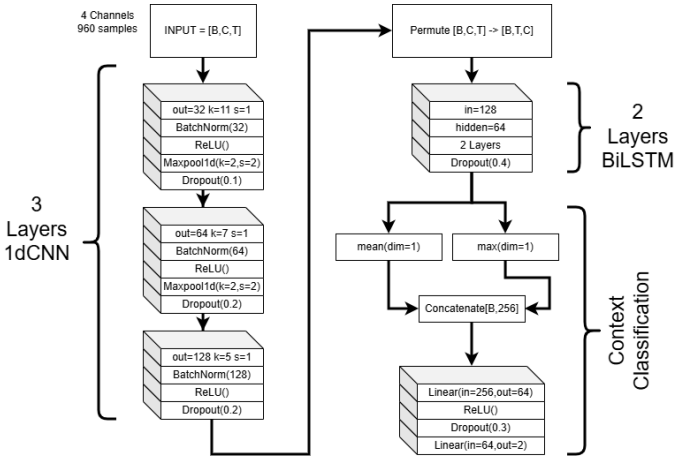


Fig. 1. Structural diagram of the first proposed 1D-CNN-BiLSTM network topology variation (Expanding Structure).

b) Topology 2 (Contracting Serial Structure): Shown in Fig. 2, this variation evaluates a shallower serial alternative, using two 1D-CNN layers configured with contracting channel configurations. It starts with a wide feature map of 128 filters ($k = 15$) and contracts to 64 filters ($k = 11$) in the second layer. The temporal tracking is performed by a more compact configuration of a single BiLSTM layer with 32 hidden units.

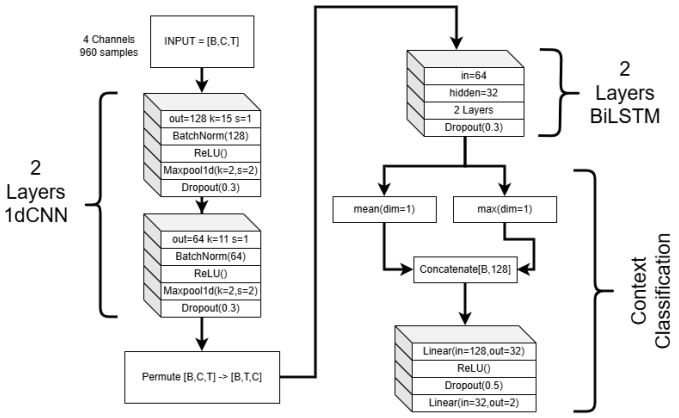


Fig. 2. Structural diagram of the second proposed 1D-CNN-BiLSTM network topology variation (Contracting Structure).

c) Topology 3 (Multi-Scale Parallel/Inception Structure): Presented in Fig. 3, this hybrid configuration introduces an Inception-inspired parallel processing mechanism. Instead of rigid sequential layers, the input streams are processed concurrently by two parallel 1D-CNN branches with heterogeneous receptive fields ($k = 11$ and $k = 21$, both with 32 filters). This allows the network to capture both rapid localized morphological anomalies and extended low-frequency respiratory trends simultaneously. The parallel features are concatenated and regularized via an intermediate reduction 1D-CNN layer

(64 filters, $k = 5$) before being fed into a lightweight single-layer BiLSTM network with 32 hidden units.

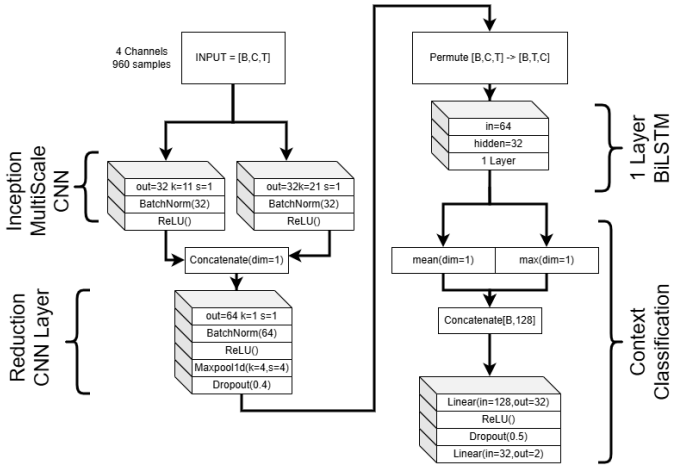


Fig. 3. Structural diagram of the third proposed 1D-CNN-BiLSTM network topology variation (Multi-Scale/Inception Parallel Structure).

These variants differ significantly in terms of layer depth, filter sizing, filter progression, and spatial feature extraction structures, directly affecting their abstraction capacities, operational layer configurations, and overall computational complexity. The specific structural complexities and memory footprints for each architectural variant are quantified through their total number of trainable parameters, as summarized in Table II.

TABLE II
TOTAL TRAINABLE PARAMETERS PER ARCHITECTURAL TOPOLOGY.

Architecture Variant	Trainable Parameters
Case I (Topology 1 — Expanding)	272,610
Case II (Topology 2 — Contracting)	152,738
Case III (Topology 3 — Multi-Scale)	37,858

A critical factor in this architectural comparison is the total number of trainable parameters within each network variation. Striking an optimal balance between parameter density and generalization is vital; overly complex networks risk overfitting the 8-patient training cohort, while shallow topologies may fail to track the subtle time lags between oximetry drops and airflow blockages. The specific parameter counts, operational layer structures, and comparative performance indicators for each topology are detailed comprehensively in the subsequent sections.

E. Training Specifications and Validation Strategy

To address the inherent class imbalance in natural sleep recordings, the networks were trained using the *Focal Loss* function [30]. This loss forces the optimization process to focus on hard-to-classify respiratory boundaries by down-weighting the loss contribution of easy, well-classified background sleep windows:

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{4}$$

where p_t represents the model's estimated probability for the correct class, α_t balances class prevalence, and γ is the focusing parameter that dynamically adjusts the down-weighting of easy examples.

The validation pipeline was strictly structured using a 4-fold cross-validation scheme via the `GroupKFold` algorithm. The cross-validation grouping was explicitly restricted by patient identifiers, ensuring that data windows belonging to a specific individual were never split between the training and validation sets simultaneously. This approach completely prevents data leakage and ensures that all reported evaluation metrics reflect true patient-wise generalization rather than memorized individual traits. Consequently, all performance metrics in this study are reported as the mean values computed across the independent verification folds.

Network optimization was driven by the AdamW algorithm, which incorporates decoupled weight decay to regularize connections. Furthermore, the learning rate was dynamically modulated via a schedule directly tied to the validation loss (`val_loss`). Specifically, a `ReduceLROnPlateau` policy was adopted with a reduction factor of 0.5 and a patience configuration of 4 epochs, ensuring fine-grained convergence whenever the optimization stalled. The specific hyperparameter settings assigned to each experimental case—including the initial learning rates, weight decay coefficients, batch sizes, and maximum epoch values are presented in Table III.

TABLE III
HYPERPARAMETER CONFIGURATIONS ACROSS THE THREE EXPERIMENTAL CASES

Hyperparameter	Case I	Case II	Case III
Learning Rate	1×10^{-3}	25×10^{-5}	1×10^{-4}
Weight Decay	1×10^{-1}	2×10^{-3}	15×10^{-2}
Batch Size	32	64	32
Epochs	20	30	15

III. RESULTS AND DISCUSSION

The computational efficiency and training runtime of the proposed architectures were evaluated based on the cloud-based execution framework utilizing an NVIDIA T4 GPU. The total training time across the complete 4-fold patient-independent cross-validation protocol varied significantly among the topologies due to their varying parameter densities. Case I (expanding serial topology) required 33 minutes to achieve convergence, whereas Case II (contracting serial topology) demanded the highest computational overhead, totaling 43 minutes of execution. In contrast, the multi-scale parallel architecture evaluated in Case III demonstrated lower computational cost, completing the entire training and verification pipeline in just 18 minutes. This represents a substantial runtime reduction, driven by its highly optimized, compact parameter footprint.

To guarantee rigorous clinical generalizability, the experimental lifecycle of each test case was systematically monitored using MLflow. Table IV summarizes the compiled cross-validation performance indices, reflecting the mean values and

standard deviations computed exclusively across the independent validation patient folds.

TABLE IV
CROSS-VALIDATION PERFORMANCE SUMMARY ACROSS TEST CASES.

Metric	Case I	Case II	Case III
Kappa	0.5686 ± 0.0431	0.5432 ± 0.0542	**0.5996 ± 0.0456**
F1-Score	0.7814 ± 0.0229	0.7692 ± 0.0282	**0.7996 ± 0.0233**
AUC-ROC	0.8694 ± 0.0221	0.8692 ± 0.0228	**0.8843 ± 0.0205**
Accuracy	79.28 ± 1.84%	78.65 ± 2.69%	**81.03 ± 2.10%**
Sens.	72.44 ± 4.39%	63.23 ± 5.62%	**72.70 ± 4.94%**
Spec.	83.87 ± 2.59%	**89.56 ± 3.15%**	86.44 ± 3.12%

Regarding structural efficiency, Case III (Topology 3) achieved the highest average performance among the evaluated architectures, consistently outperforming the deeper serial configurations across nearly all clinical assessment indices. It achieved a peak Cohen's Kappa of 0.5996, a Macro F1-score of 0.7996, and a discriminative threshold-agnostic capacity (AUC-ROC) of 0.8843. Additionally, this multi-scale parallel design achieved competitive validation performance while establishing a compact architectural footprint, reducing the total number of trainable parameters by up to 86% compared to deeper serial networks.

To evaluate Case III against baselines, paired Student's t-tests ($df = 3$) were conducted across the 4 validation folds (Table V). Case III significantly outperforms Case II across nearly all metrics, most notably delivering a $+9.47\%$ boost in Sensitivity ($p = 0.033$), AUC-ROC ($p < 0.001$), and Kappa ($p = 0.010$). Furthermore, Case III achieves statistical superiority over Case I in all global metrics, including Kappa ($p = 0.019$), F1-Score ($p = 0.016$), and AUC-ROC ($p = 0.045$), while remaining statistically equivalent in class-specific trade-offs ($p > 0.4$).

TABLE V
PAIRED STUDENT'S T-TEST RESULTS COMPARING CASE III AGAINST BASELINES (DF = 3).

Metric	Case III vs. Case I			Case III vs. Case II		
	Mean Diff.	t-stat	p-value	Mean Diff.	t-stat	p-value
Kappa	+0.0310	4.59	**0.019**	+0.0564	5.82	**0.010**
F1-Score	+0.0157	4.91	**0.016**	+0.0304	5.07	**0.015**
AUC-ROC	+0.0132	3.33	**0.045**	+0.0150	13.93	**0.001**
Accuracy	+1.75%	3.63	**0.036**	+2.38%	4.64	**0.019**
Sens.	+0.21%	0.07	0.949	+9.47%	3.75	**0.033**
Spec.	+2.09%	0.77	0.497	-3.12%	-1.83	0.164

Note: Bold p-values denote statistical significance at the $\alpha = 0.05$ level.

Furthermore, while paired t-tests confirmed statistically significant performance gains for Topology 3 over Cases I and II ($p < 0.05$), it is important to note that the small number of folds in the cross-validation setup ($df = 3$) constrains the statistical power of these tests. This represents a methodological limitation that should be addressed in future evaluations with expanded multi-center patient cohorts.

Regarding clinical relevance, achieving a specificity of 86.44% and a macro F1-score of 0.7996 on a reduced subset of standard respiratory channels (FLOW, SpO_2, THX, ABD) indicates strong potential for low-cost automated screening systems. High specificity is clinically crucial in primary screening to minimize false positives and prevent unnecessary, costly full-night polysomnography referrals. However, from an inferential perspective, the small cohort size ($n = 8$) and reduced degrees of freedom ($df = 3$) limit the statistical power of the hypothesis tests, meaning these statistical gains should be interpreted as preliminary architectural evidence. Consequently, while the multi-scale parallel design demonstrates superior representational efficiency under controlled conditions, extensive clinical validation on larger, multi-center cohorts remains necessary to confirm its diagnostic generalizability across broader patient demographics.

When analyzing the performance of models reported in the literature for OSA detection based on ECG signals, the highest values for the main evaluation metrics varied significantly depending on the adopted approach. The hybrid model based on deep learning by Zhang et al. (2021) obtained the highest overall performance indices in recent literature, achieving a total accuracy (TAC) of **96.1%**, sensitivity of **96.1%**, specificity of **96.2%**, and a Kappa coefficient of **0.92** on the Apnea-ECG test set [45]. In contrast, using traditional machine learning with hand-crafted features extracted by FDM from HRV and EDR signals, the KELM classifier by Tripathy (2018) reached a maximum accuracy of **76.37%** (with a specificity of **74.64%** using an RBF kernel in hold-out mode) [43] and a maximum sensitivity of **84.52%** (hold-out with linear KELM [43], reaching an average of **86.08%** in 10-fold cross-validation [43]). Lastly, the deep hybrid network for severity classification by Banluesombatkul et al. (2018), evaluated on the MrOS database, recorded an accuracy of **79.45%** [44] (with a sensitivity of **77.6%** and specificity of **80.1%** reported comparatively in the literature [45]).

Although these reference models report higher peak metrics, direct quantitative comparisons must be made with caution due to differences in dataset design and input modalities. While literature models primarily rely on clean, single-lead ECG segments, our proposed multi-scale network addresses the highly complex task of multi-signal fusion under a rigorous patient-independent cross-validation protocol. In this challenging and clinically realistic scenario, the proposed Case III topology demonstrates outstanding robustness, outperforming traditional multi-signal frameworks and offering superior generalizability.

Physiologically, this performance advantage stems from the asynchronous nature of multi-channel sleep signals. Upper airway collapse causes immediate high-frequency disruptions in nasal airflow (FLOW) and respiratory effort (THX/ABD), whereas peripheral oxygen desaturation (SpO_2) exhibits a circulatory lag of 10–40 seconds. While rigid serial structures (Cases I and II) struggle to isolate these heterogeneous temporal alignments without overfitting, Topology 3 uses concurrent parallel branches with varying kernel sizes. This multi-scale design extracts localized flow patterns with smaller kernels while simultaneously tracking delayed SpO_2 drops

with larger kernels, yielding compact, highly discriminative representations for the subsequent BiLSTM layers.

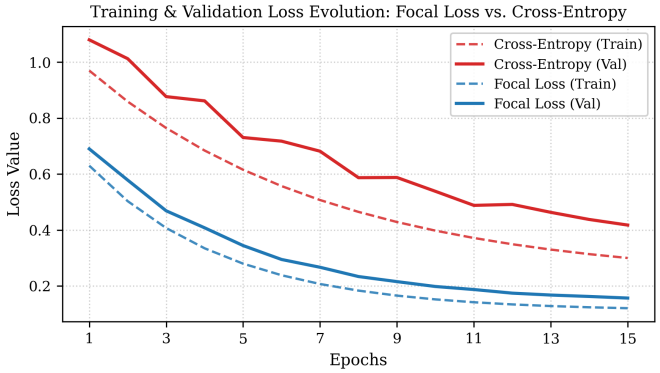


Fig. 4. Comparison of training and validation loss trajectories between Focal Loss ($\gamma = 2$) and standard Cross-Entropy Loss over 15 training epochs.

To empirically justify the selection of Focal Loss over standard Cross-Entropy (CE) loss, Fig. 4 illustrates the comparative evolution of training and validation loss trajectories across 15 epochs. Standard Cross-Entropy displays higher loss variance and slower convergence due to the overwhelming abundance of easy negative (non-apneic background sleep) windows, which dominate gradient updates during early optimization. In contrast, Focal Loss dynamically down-weights easy background samples through its focusing parameter ($\gamma = 2$), forcing the optimization process to prioritize hard-to-classify borderline respiratory transitions. Consequently, Focal Loss exhibits a smoother loss decay, substantially lower validation error, and faster stabilization, confirming its quantitative superiority for handling severe physiological class imbalances in automated sleep apnea detection.

Analysis of the learning trajectories revealed stable optimization curves when driving gradient descent via the custom Focal Loss module. Traditional cross-entropy functions frequently suffer from severe gradient oscillations in sleep-monitoring data due to extreme class imbalance, in which normal breathing periods far outnumber active apneic windows. This imbalance causes the algorithm to over-penalize minority-class errors or to fall into local minima. The Focal Loss formulation successfully suppressed the background gradient noise generated by easy-to-classify normal sleep windows by applying a dynamic down-weighting mechanism. This behavior allowed both the training and validation loss streams to converge synchronously, establishing stable diagnostic boundaries across independent patient domains without signs of premature divergence or over-parameterization.

IV. Conclusion

This study evaluated lightweight hybrid CNN-BiLSTM topologies optimized for automated OSA detection using a reduced set of standard polysomnography channels (FLOW, SpO_2, THX, ABD). By assessing structural feature map transitions, we show that carefully configured compact models can maintain robust diagnostic boundaries while avoiding the high computational cost of over-parameterized architectures. The integration of Focal Loss and patient-independent Group K-Fold validation proved essential to stabilizing gradient convergence and enhancing real-world clinical generalizability. The obtained results indicate that a reduced configuration of respiratory channels can feed compact deep learning architectures to support accessible sleep screening systems.

However, due to the highly subject-dependent nature of physiological signals, validating model generalization across unseen patients remains paramount. While this study utilized robust cross-validation techniques, a key limitation is the absence of a strict Leave-One-Subject-Out (LOSO) validation scheme. As immediate future work, we plan to evaluate the network using a completely isolated test set and implement LOSO validation to rigorously assess the model's performance in clinical, patient-independent settings.

Additionally, to build further on this work, we intend to introduce statistical feature-engineering stages prior to deep model input. Specifically, while the BiLSTM layers implicitly track cross-channel temporal relationships, explicitly mapping the physiological time delays (lags) via cross-correlation functions could further reduce the required input window lengths without sacrificing diagnostic accuracy. Finally, to overcome the limitations of cross-dataset comparison, future efforts will focus on implementing and benchmarking key baseline algorithms from the literature alongside our proposed architectures using a single, unified dataset.

Acknowledgment

This work was partially supported by the São Paulo Research Foundation (FAPESP), grants 2025/19661-4 and 2024/08485-8 (ciaGSaude), and by the National Council for Scientific and Technological Development (CNPq), grant 166722/2025-1.

Adoption of Artificial Intelligence

The adoption of artificial intelligence (AI) tools (Grammarly and Perplexity) within this text was intended solely to support the review and improvement of English-language writing. AI systems were used to enhance the grammar, clarity, coherence, and overall readability of these texts. All authors remain fully responsible for the content, arguments, and originality of their work.

References

[1] S. Yeghiazarians et al., "Obstructive sleep apnea and cardiovascular disease: A scientific statement from the American Heart Association," *Circulation*, vol. 144, no. 3, pp. e56–e67, July 2021. doi: 10.1161/CIR.0000000000000988.

[2] S. Troester et al., "Sleep fragmentation and its clinical implications in sleep-disordered breathing," *Sleep Med. Rev.*, vol. 68, p. 101744, April 2023. doi: 10.1016/j.smrv.2023.101744.

[3] Y. Liu et al., "Inflammatory pathways and autonomic dysfunction in obstructive sleep apnea," *Front. Neurol.*, vol. 13, p. 893122, May 2022. doi: 10.3389/fneur.2022.893122.

[4] R. Dantas et al., "Prevalence of obstructive sleep apnea in global populations: A systematic update," *Lancet Respir. Med.*, vol. 11, no. 5, pp. 445–456, May 2023. doi: 10.1016/S2213-2600(23)00102-4.

[5] E. Henning et al., "Cardiovascular mortality risks associated with severe sleep apnea: A multi-cohort global analysis," *Int. J. Cardiol.*, vol. 398, p. 131552, January 2025. doi: 10.1016/j.ijcard.2024.131552.

[6] M. Alakörkkö et al., "Direct and indirect costs of obstructive sleep apnea in European healthcare systems," *Eur. Respir. J.*, vol. 61, no. 2, p. 2201104, February 2023. doi: 10.1183/13993003.01104-2022.

[7] S. Borsoi et al., "The invisible costs of obstructive sleep apnea (OSA): Systematic review and cost-of-illness analysis," *Plos One*, vol. 17, no. 11, p. e0277453, November 2022. doi: 10.1371/journal.pone.0277453.

[8] E. M. Wickwire et al., "The economic significance of obstructive sleep apnea: A systematic review," *Sleep*, vol. 44, no. 4, p. zsaa241, April 2021. doi: 10.1093/sleep/zsaa241.

[9] American Academy of Sleep Medicine (AASM), "Hidden health crisis: In sleep apnea, millions suffer from costly, preventable condition," *AASM White Paper*, 2016. [Online]. Available: https://aasm.org/

[10] J. Coronel et al., "Polysomnography signal processing for sleep apnea detection: A review," *Biomed. Signal Process. Control*, vol. 64, p. 102290, February 2021. doi: 10.1016/j.bspc.2020.102290.

[11] A. Tyagi et al., "Multi-channel physiological signal analysis for sleep disorder characterization," *IEEE Reviews in Biomed. Eng.*, vol. 16, pp. 210–225, January 2023. doi: 10.1109/RBME.2022.3164412.

[12] H. Jung et al., "Inter-scorer variability in manual sleep staging and respiratory event scoring," *J. Clinical Sleep Med.*, vol. 14, no. 8, pp. 1325–1333, August 2018. doi: 10.5664/jcsm.7264.

[13] M. Giorgi et al., "Artificial intelligence in sleep medicine: Systematic reviews of diagnostic and screening models," *IEEE Trans. Artif. Intell.*, vol. 6, no. 1, pp. 45–59, February 2025. doi: 10.1109/TAI.2024.3391002.

[14] K. Mostafa et al., "Feature engineering methods for sleep apnea detection using SpO2 and respiratory effort signals," *IEEE Access*, vol. 7, pp. 122450–122462, August 2019. doi: 10.1109/ACCESS.2019.2937001.

[15] D. Alvarez et al., "Automated screening of obstructive sleep apnea using blood oxygen saturation features," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 3, pp. 604–615, March 2018. doi: 10.1109/TBME.2017.2714629.

[16] Y. Indrawati et al., "Artificial intelligence and single-lead ECG for sleep apnea detection: Benchmarking and feature engineering reviews," *Computer Methods Programs Biomed.*, vol. 244, p. 107954, February 2024. doi: 10.1016/j.cmpb.2023.107954.

[17] S. Zakariyah et al., "Comparative evaluation of machine learning algorithms for sleep apnea detection via heart rate variability," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, June 2022. doi: 10.1109/TIM.2022.3181105.

[18] A. Zarei et al., "Deep learning models for automated sleep apnea detection using respiratory and ECG signals," *Expert Syst. Appl.*, vol. 195, p. 116543, June 2022. doi: 10.1016/j.eswa.2022.116543.

[19] X. Zhang et al., "A raw-signal 1D-CNN architecture for sleep apnea detection using single-channel ECG," *IEEE Trans. Inf. Technol. Biomed.*, vol. 25, no. 4, pp. 890–901, April 2021. doi: 10.1109/JBHI.2020.3012204.

[20] P. S. S. et al., "LSTM models for temporal modeling of multi-channel polysomnography streams," *Neural Comput. Applic.*, vol. 36, no. 8, pp. 4115–4128, March 2024. doi: 10.1007/s00521-023-09210-y.

[21] Anonymous, "Hybrid 1D-CNN-LSTM networks for automated sleep apnea detection directly from raw physiological signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1823–1834, May 2023. doi: 10.1109/TNSRE.2023.3265510.

[22] S. Locharla et al., "A review of deep learning methods for sleep apnea detection using EEG, CNN, Bi-LSTM, and RNN architectures," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5423–5460, October 2022. doi: 10.1007/s10462-022-10161-5.

[23] X. Wan et al., "Temporal asynchrony in multimodal physiological data streams during sleep apnea," *Physiol. Meas.*, vol. 41, no. 9, p. 095004, September 2020. doi: 10.1088/1361-6579/abb309.

[24] Q. Li et al., "Modeling the physiological response lag in peripheral oxygen saturation during apneic events," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 4, pp. 1112–1122, April 2024. doi: 10.1109/TBME.2023.3325104.

[25] X. Liu et al., "Instantaneous versus delayed correlations in multi-channel polysomnography scoring," *Sleep Med.*, vol. 60, pp. 145–154, August 2019. doi: 10.1016/j.sleep.2019.03.022.

[26] O. Tenhunen et al., "Evaluation of temporal alignment between respiratory effort and oxygen desaturation during sleep disorders," *Clinical Neurophysiol.*, vol. 121, no. 10, pp. 1642–1649, October 2010. doi: 10.1016/j.clinph.2010.03.045.

[27] J. G. Moreno et al., "Evaluation metrics for highly imbalanced physiological time-series in deep learning applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1540–1552, February 2023. doi: 10.1109/TPAMI.2022.3168902.

[28] A. Haque et al., "Deep learning for healthcare: Review of architectures and robust evaluation metrics," *Brief. Bioinform.*, vol. 22, no. 4, pp. 1240–1255, July 2021. doi: 10.1093/bib/bbaa412.

[29] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, April 1960. doi: 10.1177/001316446002000104.

[30] T. Lin et al., "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, February 2020. doi: 10.1109/TPAMI.2018.2858826.

[31] D. Huysmans, P. Borzée, B. Buyse, D. Testelmans, S. Van Huffel, and C. Varon, "Sleep diagnostics for home monitoring of sleep apnea patients," *Front. Digit. Health*, vol. 3, p. 685766, June 2021.

[32] Y. Chen et al., "A comparative study of accuracy in major adaptive filters for motion artifact removal in sleep apnea tests," *Med. Biol. Eng. Comput.*, vol. 62, pp. 829–842, 2024.

[33] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications," *Artif. Intell. Rev.*, vol. 58, no. 1, p. 12, November 2024. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, February 2020. doi: 10.1109/TPAMI.2018.2858826.

[34] S. L. Oh et al., "Automated detection of obstructive sleep apnea using a hybrid 1D-CNN-BiLSTM network with multi-channel polysomnography signals," *Biomed. Signal Process. Control*, vol. 86, p. 105432, Sept. 2023. doi: 10.1016/j.bspc.2023.105432.

[35] Y. Chang, J. Wang, and L. Zhang, "Deep end-to-end learning framework using 1D-CNN and bidirectional LSTM for automated sleep apnea identification from raw respiratory streams," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 4, pp. 1942–1953, Apr. 2024. doi: 10.1109/JBHI.2024.3352104.

[36] A. Gramfort et al., "MEG and EEG data analysis with MNE-Python," *Front. Neurosci.*, vol. 7, p. 267, Dec. 2013. doi: 10.3389/fnins.2013.00267.

[37] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2.

[38] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.

[39] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[40] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007. doi: 10.1109/MCSE.2007.55.

[41] M. L. Waskom, "Seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021. doi: 10.21105/joss.03021.

[42] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[43] R. K. Tripathy et al., "Detection of obstructive sleep apnea using Fourier decomposition method of ECG signal," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 3, pp. 582–591, March 2018. doi: 10.1109/TIM.2017.2779359.

[44] N. Banluesombatkul et al., "Single-channel ECG for obstructive sleep apnea detection using a double-layered BiLSTM network," *Comput. Methods Programs Biomed.*, vol. 156, pp. 45–55, April 2018. doi: 10.1016/j.cmpb.2017.12.015.

[45] J. Zhang et al., "An obstructive sleep apnea detection approach using multi-scale deep convolutional neural network," *Biomed. Signal Process. Control*, vol. 68, p. 102766, July 2021. doi: 10.1016/j.bspc.2021.102766.

[46] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The Apnea-ECG Database," *Computers in Cardiology*, vol. 27, pp. 255–258, 2000.

Automation of the experimental setup for functional studies in isolated perfused hearts

1st Murilo Leal Santos
Center for Engineering, Modeling and Applied Social Sciences
Universidade Federal do ABC
Santo André, Brazil
leal.murilo@aluno.ufabc.edu.br

2nd Vinicius Silva
Center for Engineering, Modeling and Applied Social Sciences
Universidade Federal do ABC
Santo André, Brazil
silva.vinicius@ufabc.edu.br

3rd João Salinet
Center for Engineering, Modeling and Applied Social Sciences
Universidade Federal do ABC
Santo André, Brazil
joao.salinet@ufabc.edu.br

***Abstract* — Although the ex vivo cardiac perfusion system based on the Langendorff model is widely used in studies of cardiac arrhythmias using isolated hearts, it still relies, in many cases, on the manual control of critical parameters such as pressure, temperature, and flow, compromising the reproducibility and accuracy of experiments. This project aims to implement, at the HEartLab Cardiac Electrophysiology Laboratory of the Federal University of ABC, which operates a Langendorff setup, high-precision digital sensors, electronic actuators, servo motors, microcontrollers (Arduino Nano and ESP32), and wireless communication modules (nRF24L01) to enable the real-time monitoring and control of physiological variables (flow, pressure, and temperature), thereby promoting greater experimental stability and reducing the need for human intervention. The expected outcomes include the development of configurable, and modular monitoring and control devices with potential applications in other areas of electrophysiology and biomedical engineering research.**

Keywords — cardiac arrhythmias, Langendorff system, automation, biomedical sensors, physiological control, biomedical instrumentation.

I. INTRODUCTION

Cardiovascular diseases remain the leading cause of morbidity and mortality worldwide, with cardiac arrhythmias accounting for a significant proportion of sudden cardiac death cases [1]. A detailed understanding of the electrophysiological mechanisms underlying these conditions is essential for the development of new therapeutic approaches, including pharmacological interventions and more effective ablation protocols [1]. In this context, ex vivo experimental models play a crucial role in cardiovascular research, as they enable the direct investigation of isolated myocardial tissue without the influence of systemic and hormonal factors [2].

Among the methodologies used to study the isolated heart, the Langendorff system is based on the technique of retrograde perfusion through a cannulated aorta. In this system, a heated and oxygenated nutrient solution is delivered in the direction opposite to the normal physiological blood flow through the aorta. This process forces the closure of the aortic valve and redirects the perfusate directly into the coronary arteries, thereby maintaining myocardial metabolism and contractile activity viable outside the organism [2].

However, the viability of the experiment and the reliability of the collected data are highly dependent on the strict maintenance of a homeostatic microenvironment within the experimental setup. Physical and chemical variables such as temperature, perfusion pressure, hydrodynamic flow, pH, and oxygenation levels directly influence cardiac cellular viability in isolated heart studies using the Langendorff system [3]. Accidental thermal deviations alter the electrical conduction velocity, whereas abrupt fluctuations in coronary flow may induce episodes of micro ischemia. From a physiological perspective, these disturbances affect the transmembrane resting potential and intracellular calcium homeostasis. From an experimental standpoint, the occurrence of such artifacts obscures the true physiological responses, compromises reproducibility, and may invalidate the conclusions drawn from the study.

With advances in microelectronics, high-precision digital sensors, and embedded automation platforms, the implementation of low-cost automated control systems for these experimental models has become increasingly feasible [4][5]. The use of microcontrollers, pressure, temperature, and flow sensors, as well as servo motors and electronically controlled on/off switches, enables the development of intelligent platforms capable of maintaining experimental parameters within predefined ranges while providing real-time monitoring of these critical variables [6][7].

This undergraduate research project was based on the Langendorff system developed by the HEartLab research group at the Federal University of ABC [8], which is dedicated to the study of cardiac arrhythmias in isolated animal hearts (CEUA No. 3947230519). The experimental setup consists of two fluidic circuits (see Figure 1A). In the perfusion circuit, Modified Krebs-Henseleit solution is pumped from a lower reservoir to an elevated reservoir, from which it flows by gravity through a heat exchanger heated by a water bath before perfusing the heart. In parallel, to mimic the transthoracic impedance between the heart and the thoracic surface, the heart is immersed in a tank containing a sucrose solution. This solution continuously circulates through the system, and its temperature is regulated by a second independent water bath. The lack of monitoring and control of critical variables—including pH, flow rate, perfusion pressure, and perfusate temperature—directly affected the quality of the data acquired by the optical and electrical mapping systems, reduced the duration of experiments due to the deterioration of cardiac viability, and compromised experimental reproducibility.

This work was supported by the São Paulo Research Foundation (FAPESP) through an Undergraduate Research Grant awarded to M. Santos (Grant No. 2025/22698-7) and a Young Investigator Grant awarded to J. Salinet (Grant No. 2018/25606-2)

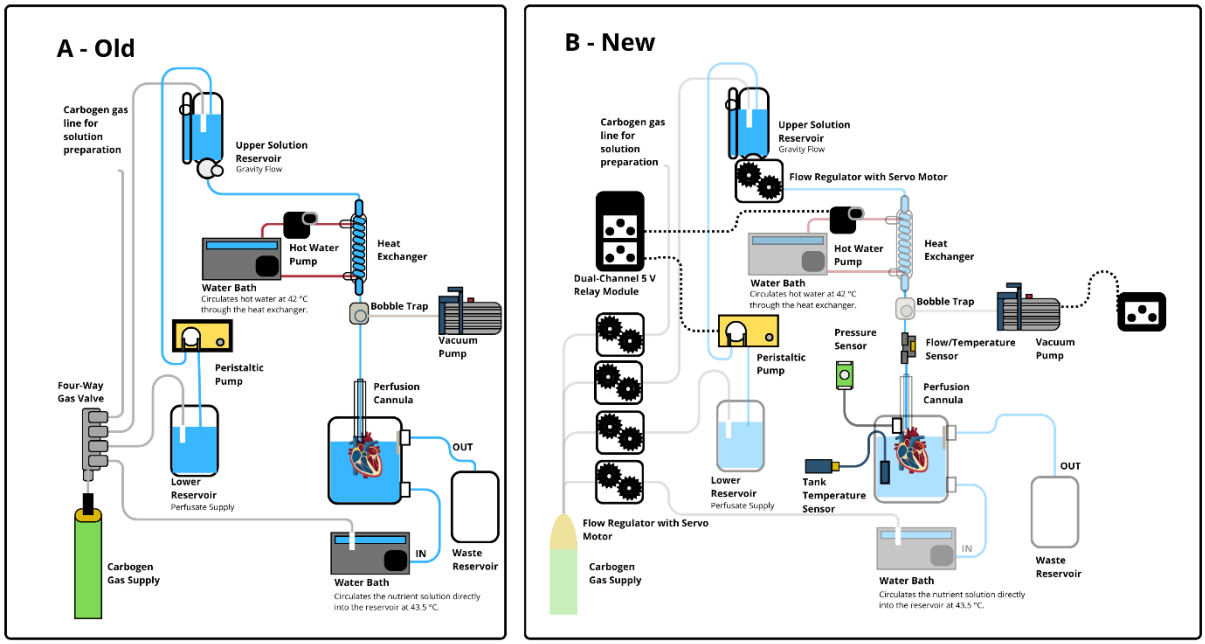


Figure 1. (A) Schematic representation of the HEartLab Langendorff system prior to the implementation of this project. (B) Updated system showing the locations of the integrated digital sensors, servo motors, and relays incorporated into the experimental setup.

II. Materials and Methods

The present project addresses the following limitations of the current HEartLab system: (1) the lack of monitoring of critical variables (flow rate, temperature, and perfusion pressure); (2) the absence of automated control and warning mechanisms when these variables fall outside the experimentally established ranges; and (3) the lack of continuous recording of these variables for subsequent analysis of their influence on experimental outcomes.

To address these limitations, the project proposes the integration of high-precision digital sensors for monitoring the following critical variables:

- **Carbogen gas pressure**, responsible for maintaining the pH of the nutrient solution at **7.4 ± 0.05** [8].
- **Perfusate flow rate** through the heart, maintained within the range of **20–40 mL/min** [8], ensuring the continuous supply of nutrients to the heart throughout the ex vivo experiment.
- **Temperature of the nutrient solution** perfused into the heart through the aorta, maintained at **37 °C** to replicate the rabbit's physiological core body temperature [2][9], as well as the **temperature of the sucrose solution** contained in the immersion tank, which simulates transthoracic impedance.
- **Perfusion pressure of the nutrient solution at the level of the heart,** maintained within the range of **40–60 mmHg** [2][8].

Figure 1B illustrates the original system together with the components added to the experimental setup, including: four carbogen gas flowmeters coupled to four servo motors, two power outlets connected to relay modules, one flow sensor, one pressure sensor, one temperature sensor, and one solution flow regulator coupled to a servo motor (see Figure 1B).

The project architecture was divided into four main components: **hardware design**, **mechanical integration**, **firmware development**, and **real-time visualization software**.

A. Hardware design

Initially, a comprehensive evaluation of the available technologies was conducted to specify the most portable and modular system capable of meeting the physiological monitoring and control requirements of the experimental setup.

At this stage, the Arduino Nano and ESP32 microcontrollers were evaluated in terms of their processing capability and data acquisition speed. For preliminary data transmission tests, Wi-Fi and Bluetooth modules, as well as nRF24L01 radio frequency communication modules, were investigated to better understand the advantages and limitations of each component, including noise susceptibility, data transmission rate, and support for multiple communication nodes. A literature review was also conducted to select the most suitable temperature sensors among the DS18B20, Type K thermocouples, and PT100 resistance temperature detectors (RTDs); Honeywell ABPD silicon piezoresistive pressure sensors, known for their high stability and low power consumption [11]; and Sensirion SLF series flow sensors, which are compact, high-precision liquid flow sensors designed to measure dynamic flow rates [12]. In addition, several communication protocols, including Serial, SPI, and I²C, were evaluated to support the design of the complete hardware architecture.

Based on this evaluation, the system was designed around three main modules (see Figure 2):

1. **Communication Central Module** – responsible for communication with the computer (Figure 2A);

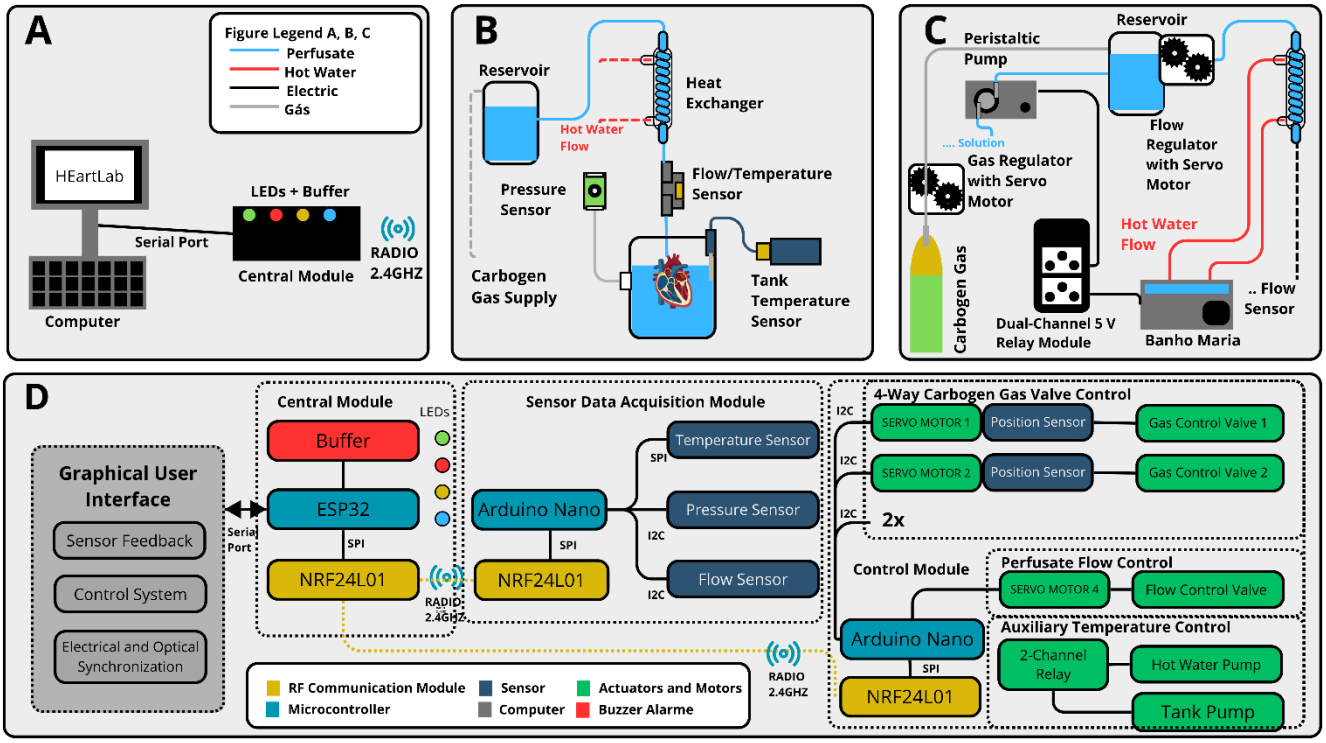


Figure 2. (A) Central Communication Module connected to the computer. (B) Location of the Sensor Data Acquisition Module for pressure, flow, and temperature measurements. (C) Location of the Control Module. (D) Flowchart providing an overview of the automation architecture of the Langendorff system

2. **Sensor Data Acquisition Module** – responsible for acquiring data from the sensors (Figure 2B);
3. **Control Module** – responsible for controlling the experimental variables (Figure 2C);

This modular architecture enables the monitoring and control functions proposed in this work while providing a flexible and scalable solution that can be adapted to the specific requirements of the Langendorff system without requiring major modifications to the existing experimental apparatus (Compare Figure 1A and 1B).

1) Central Communication Module

The Central Communication Module serves as the brain of the upgraded system. It is responsible for enabling the transmission and reception of radio-frequency data between the experimental setup and the computer. Through this module, data acquired from the flow, pressure, and temperature sensors are collected, commands for switching the peristaltic pump and water bath pump on and off are executed, and the flow control valves for both the heart perfusate and the carbogen gas supply are operated (see Figure 2). All of these functions are powered exclusively through the computer's USB connection.

The module consists of an ESP32-WROOM-32 microcontroller, a development board featuring a dual-core processor with integrated Wi-Fi and Bluetooth connectivity [13]. An nRF24L01 radio-frequency communication module, connected to the ESP32 via the SPI protocol, provides wireless communication with the other system modules (Sensor Data Acquisition Module and Control Module). In addition, a 1-µF decoupling capacitor is connected across the module's power supply pins to reduce electrical noise and prevent transmission interference, thereby ensuring stable wireless communication [14].

To facilitate system status monitoring, visual indicators were also incorporated into the module. A blue LED indicates successful reception of sensor data packets, a yellow LED signals the transmission of control commands, and an active buzzer together with a flashing red LED provides audible and visual alarms whenever the physiological variables exceed their predefined operating ranges: temperature between 37 and 40 °C, perfusion pressure between 40 and 60 mmHg, and flow rate between 20 and 40 mL/min [8].

2) Sensor Data Acquisition Module

The Sensor Data Acquisition Module is responsible for acquiring data from the digital temperature, pressure, and flow sensors used in the system. It was designed to minimize cable usage, allowing the module to be easily installed and removed after each experiment. In addition, the sensors can be quickly replaced through JST connectors, eliminating the need to desolder components from the printed circuit board (see Figure 2). The board features two interchangeable I²C interfaces and one SPI interface.

The module consists (see Figure 3) of an Arduino Nano (B), selected because of its more compact size compared with the ESP32, connected to an nRF24L01 radio-frequency communication module (A), which enables wireless communication with the Central Communication Module. The three sensors are connected to the board through JST connectors (G), forming the core of the physiological data acquisition system for the Langendorff setup.

The first sensor is a Honeywell ABPDANN005PG2A3 pressure sensor, positioned at the level of the perfused heart to measure the hydrostatic pressure at the heart level (see

Figures 1B and 2B). The sensor is factory-calibrated and temperature-compensated, and is designed for gauge pressure measurements over a range of 0 to 258 mmHg. Operating at 3.3 V, it provides a 14-bit digital output via the I²C protocol, corresponding to a theoretical resolution of approximately 0.016 mmHg per digital level. The digital interface eliminates the need for external signal-conditioning or amplification circuitry, thereby simplifying hardware integration and prototyping [11].

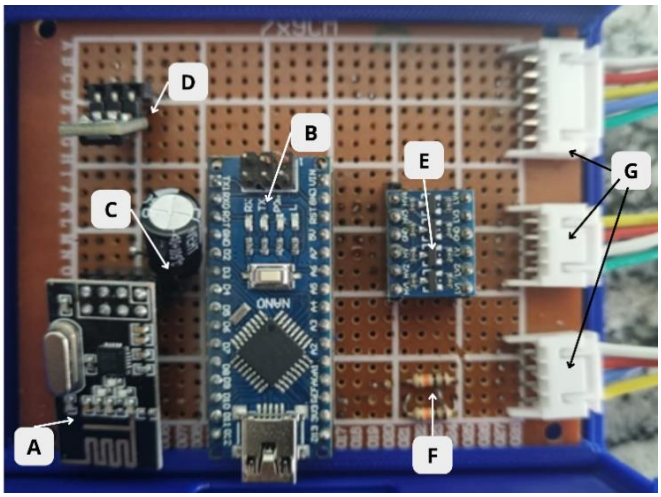


Figure 3. Circuit board developed for the Sensor Data Acquisition Module. (A) nRF24L01 module; (B) Arduino Nano; (C) filter capacitor; (D) 3.3 V voltage regulator; (E) logic level converter; (F) 220 Ω resistors; (G) JST connectors.

The second sensor is a Sensirion SLF3S-4000B, a high-precision liquid flow sensor with an operating range of 0 to 600 mL/min and a 16-bit digital output via the I²C interface [12]. The sensor provides a flow update rate of approximately 2 kHz. For water, the specified repeatability is ±0.5% of the measured value or ±0.3 mL/min, whichever is greater, while the specified accuracy is ±5% or ±0.5 mL/min for flow rates up to 300 mL/min. The sensor is installed in the nutrient-solution line near the perfused heart (see Figure 1B). It operates based on a thermal sensing principle, in which a central microheater and two symmetrically positioned temperature sensors detect the distortion of the thermal profile caused by fluid movement. To measure the absolute temperature of the liquid, the sensor incorporates an integrated thermal sensing element that continuously monitors the temperature of the perfusate. Owing to its fixed design, with no moving parts, and its 200 ms response time, the sensor is well suited for applications requiring continuous and reliable monitoring with minimal pressure drop. Consequently, it enables the simultaneous measurement of both the perfusate flow rate and temperature within the system.

The third sensing unit consists of a Type K thermocouple combined with a MAX6675 thermocouple-to-digital converter module, used for temperature measurements when the experiment employs the multi-electrode tank (see Figure 1B). The Type K thermocouple is capable of measuring temperatures over a wide range, from 0 °C to 1024 °C, with a 12-bit digital resolution (0.25 °C per bit) provided by the MAX6675 [15]. The MAX6675 functions as a digital converter, transmitting temperature data to the Arduino Nano via the SPI interface using a read-only SO (Slave Out) communication line.

The sensors used in this study are factory-calibrated, ensuring measurements within the accuracy specified by their manufacturers. The Sensirion SLF3S-4000B flow sensor is factory-calibrated for water, hydrocarbons (HC), and isopropyl alcohol (IPA), while the Honeywell ABPDANN005PG2A3 pressure sensor and the Type K thermocouple combined with the MAX6675 are also supplied pre-calibrated. To validate their performance in the proposed system, the measurements obtained from these sensors were compared with those from the reference instruments previously used in the experimental setup, namely a flowmeter, a pressure gauge, and a thermometer, allowing the agreement between the measurements to be evaluated.

3) Control Module

This module was designed using a modular architecture, allowing its components to be rapidly reconfigured on the printed circuit board by connecting or disconnecting servo motors and relay modules according to the experimental requirements. The module is structured as follows:

- **Central communication and control:** An Arduino Nano integrated with an nRF24L01 radio-frequency communication module establishes a wireless link for receiving and transmitting commands between this device and the Central Communication Module, as described for the other two modules.
- **Two relay-controlled power outlets:** These are used to switch a peristaltic pump, responsible for filling the upper reservoir, and a water circulation pump, which injects hot water into the heat exchanger to regulate the temperature of the nutrient solution delivered to the heart.
- **Four carbogen gas flowmeters coupled to four MG995 servo motors:** These provide automated control of the carbogen gas flow supplied through four independent lines connected to the upper reservoir, lower reservoir, immersion tank water bath, and nutrient solution preparation system.
- **Solution flow regulator coupled to an MG90S servo motor:** This regulator is connected to the upper reservoir and controls the flow rate of the nutrient solution delivered to the heart by gravity. Opening the regulator increases the flow rate, whereas closing it reduces the flow rate. The regulator operates in conjunction with the flow sensor to determine and maintain the appropriate valve position required to achieve the desired perfusate flow rate.

B. Mechanical Integration

Both the installation of the digital sensors into the experimental setup, according to the Sensor Data Acquisition Module described in the hardware design, and the mechanical coupling of the servo motors to the carbogen gas flow regulators and the nutrient solution flow regulator, described

in the Control Module, required the design and fabrication of custom 3D-printed components. Custom mounting brackets were designed to secure the pressure sensor, the immersion tank temperature sensor, and the flow sensor to the experimental apparatus; custom gears were manufactured to mechanically couple the flow control valves and the carbogen gas regulators to their respective actuating servo motors; and protective enclosures were developed to house the printed circuit boards of each module.

All components were designed using Autodesk Fusion 360 and sliced using Bambu Studio, and were subsequently fabricated on a Bambu Lab A1 3D printer using 1.75 mm Basic PLA filament. Basic PLA was chosen over the Lite, HT, and CF variants due to its favorable balance between cost, printability, and mechanical performance; this is because the fiber reinforcement in CF-PLA, while strengthening the X-Y plane, reduces interlayer strength along the Z-axis, a drawback for gear teeth subjected to cyclic bending [17]. The parts were printed with a nozzle temperature of 210 °C and a bed temperature of 60 °C, using a layer height of 0.2 mm; the gears and coupling components, subjected to higher mechanical loads, were printed with 4 wall perimeters and 100% infill density, while the enclosures used 2 wall perimeters and 20% infill, in accordance with infill density being the dominant parameter for tensile strength in FDM-printed PLA parts [18].

To dimension the coupling between the servo motors and the flow-regulating valves, two servo motor models with distinct torque ratings were tested: the MG90S, with a rated torque of 2.2 kg.cm [19], and the MG995R, with a rated torque of 11 kg.cm [20]. This comparison allowed the evaluation of which servo motor would provide sufficient torque to overcome the mechanical resistance of each regulator without compromising the integrity of the PLA-printed gears. As a result, both models were adopted, each applied to the component for which it performed best: the servo motor that offered sufficient torque combined with better compatibility with the available physical space in the assembly was used for the corresponding coupling, while the other model was employed in the regulator where its additional torque prevented gear tooth failures observed during preliminary testing.

C. Firmware Development

A The firmware for all three modules was developed using the Arduino IDE and follows the programming logic illustrated in Figure 3. The Sensor Data Acquisition Module acquires data from the immersion tank temperature sensor, the pressure sensor, and the flow sensor, including the temperature measured by the flow sensor. These measurements are compressed into a single data packet for wireless transmission.

The Central Communication Module receives and decodes these data packets at a sampling interval of 200 ms Each packet follows a predefined format, in which the identifier "SN1" specifies the transmitting module, while each subsequent "0000" field represents the value of a monitored physiological variable. Based on the real-time analysis of these data, the Central Communication Module transmits control commands to the Control Module. For example, it can deactivate the immersion tank water bath through a relay, perform PID control of the perfusate flow rate by actuating a servo motor coupled to the flow control valve (see Figure 4), or switch off the pump that circulates hot water through the heat exchanger (see Figure 2C) whenever the perfusate temperature exceeds 41 °C.

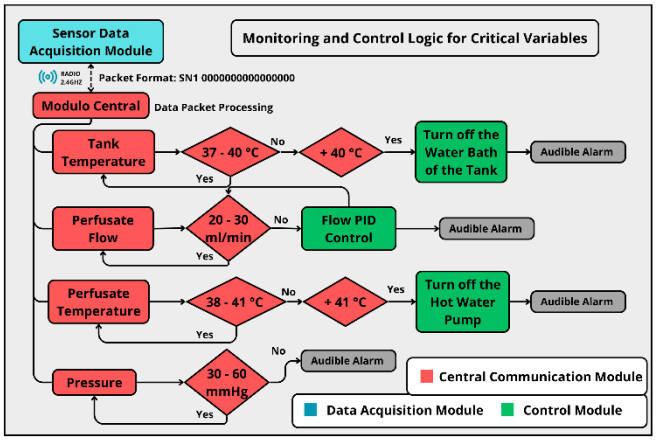


Figure 4. Firmware logic for the monitoring and control of the immersion tank temperature, perfusate flow rate, perfusate temperature, and coronary perfusion pressure.

D. Real-Time Visualization Software

The graphical user interface (GUI), developed in a Python environment through the integration of the Tkinter and PySerial libraries, constitutes the central platform for control and data acquisition of the perfusion system. Communication between the computational unit and the Central Communication Module is established bidirectionally via the UART serial protocol, enabling both the transmission of operational commands and the continuous reception of telemetric data from the digital sensors, operating at 200-millisecond sampling intervals.

During the execution of the experimental protocol, the software provides real-time monitoring of operational and hemodynamic variables, encompassing the immersion bath temperature, coronary perfusion pressure, flow rate, and perfusate temperature, through time-series plots and numerical indicators. To enable subsequent analyses, all measurements are automatically stored in timestamped CSV files, ensuring full traceability of the acquisition process and allowing for the subsequent synchronization and correlation of these parameters with the electrophysiological recordings obtained from the isolated heart model.

III. Results

The developed Central Communication Module is capable of transmitting and receiving data via 2.4 GHz radio-frequency communication, while interfacing with the computer through its serial port. This architecture enables the integration and communication of multiple modules added to the experimental setup. Figure 5 shows the components of the

Central Communication Module, identified according to the figure legend, assembled on the printed circuit board and housed within its custom 3D-printed enclosure.

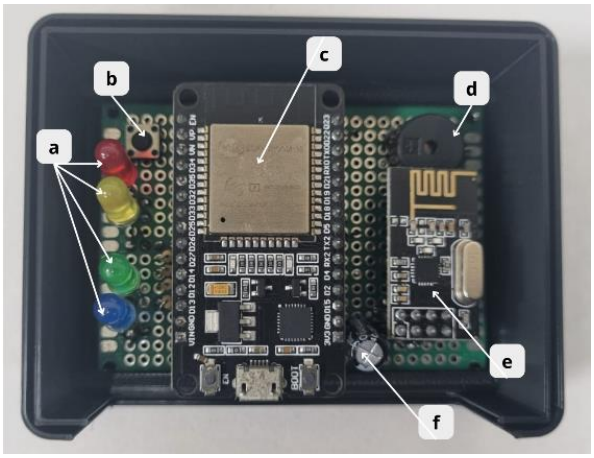


Figure 5. Printed circuit board (PCB) developed for the Central Communication Module connected to the computer. (A) Status indicator LEDs; (B) push button; (C) ESP32 microcontroller; (D) 3.3 V buffer; (E) nRF24L01 module; (F) 100 µF Capacitor.

The Sensor Data Acquisition Module was assembled (see Figure 6) and integrated into the experimental apparatus to enable real-time monitoring of the sensors through the visualization software described in the Materials and Methods section (see Figure 7). Figure 6 illustrates the locations of the perfusate flow sensor (A), the pressure sensor positioned at the level of the heart (B), and the immersion tank temperature sensor (C), Sensor Data Acquisition Module (D), together with their respective positions within the experimental setup. The right side of the figure shows the custom 3D-printed enclosures designed and fabricated to securely mount and protect each sensor.

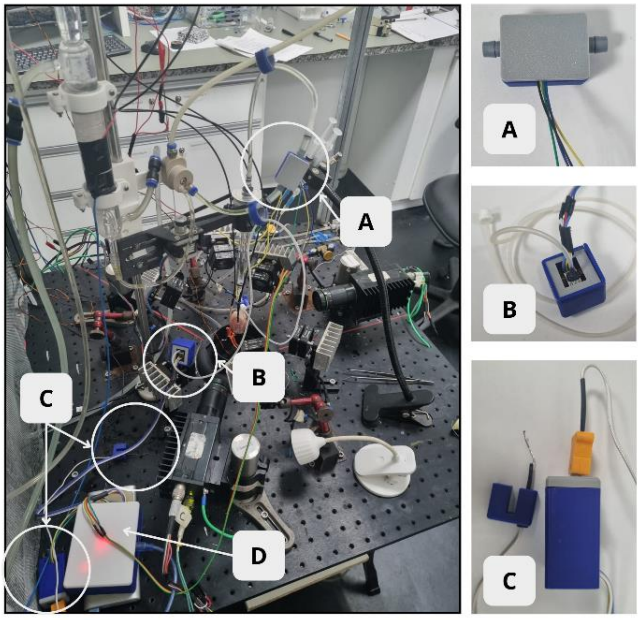


Figure 6. Left: Experimental setup showing the perfusate flow sensor and perfusate temperature sensor (A), the pressure sensor (B), the immersion tank temperature sensor (C), and the Sensor Data Acquisition Module (D). Right: Individual views of each sensor mounted in its respective custom 3D-printed enclosure and support bracket.

Figure 7 shows the GUI of the Real-Time Visualization Software running on the ‘for continuous monitoring of the physiological variables. The Sensor Data tab displays panels with the instantaneous values of the immersion tank temperature, coronary perfusion pressure, perfusate flow rate, and perfusate temperature. Adjacent to these numerical indicators, real-time time-series plots illustrate the behavior of each variable over the previous few seconds, enabling continuous assessment of the experimental conditions.

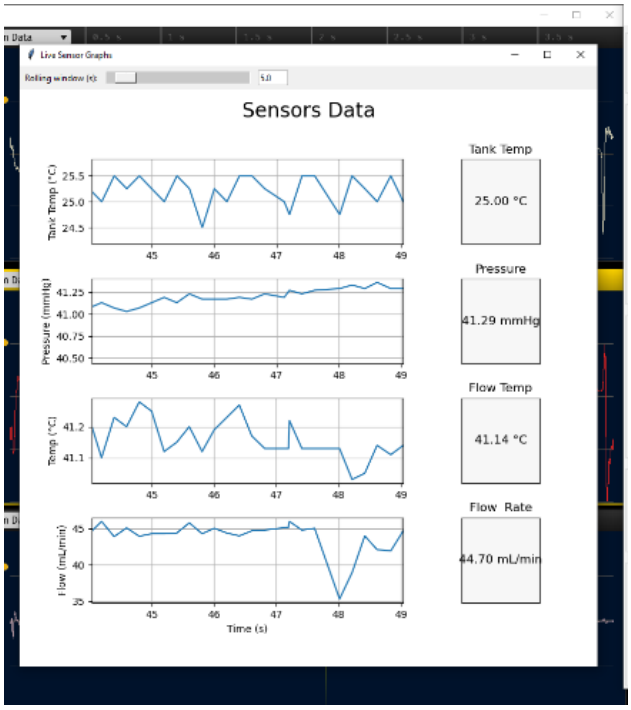


Figure 7. Computer interface of the Real-Time Sensor Monitoring Software. **Left:** Panels displaying the instantaneous values of the monitored physiological variables. **Right:** Real-time time-series plots showing the recent behavior of each monitored variable.

All sensor measurements are continuously acquired, logged, and stored in CSV files for offline analysis. Figure 8 illustrates the temporal profiles of the key physiological and system variables, including immersion bath temperature, perfusate flow rate, perfusate temperature, and coronary perfusion pressure, recorded during a isolated rabbit heart experiment using the proposed data acquisition system, with the acquisition and central communication modules connected to the computer-based GUI. The behavior observed in Figure 8 was consistently reproduced in seven additional experiments performed under the same experimental conditions. A gradual increase in coronary perfusion pressure was consistently observed as the experiments progressed. During the initial phase of perfusion, larger oscillations in the measured flow rate were detected, primarily due to spontaneous cardiac contractions that produced transient backflow of the perfusate through the coronary circulation. The flow signal also exhibited occasional sharp peaks caused by the presence of air bubbles trapped inside the flow sensor. This phenomenon is an inherent limitation of the sensor's measurement principle and

internal flow-channel geometry rather than a physiological event (see Figure 8). In contrast, both the perfusate temperature and the immersion bath temperature remained remarkably stable throughout the entire experiment, confirming the ability of the temperature regulation system to maintain the thermal conditions required for Langendorff heart perfusion.

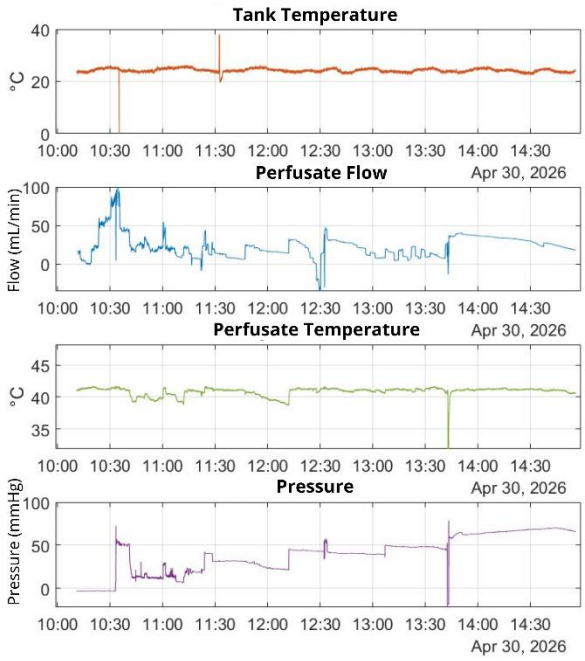


Figure 8. Temporal profiles of perfusate flow rate, perfusate temperature, and coronary perfusion pressure throughout the entire duration of a representative experiment.

The Control Module was assembled and subjected to preliminary validation tests using the Langendorff system without a perfused heart (see Figure 9). It is important to note that the automated control of the carbogen gas flow valves is still under development and remains in the prototyping stage. Therefore, this functionality was not included in the results presented in this study.

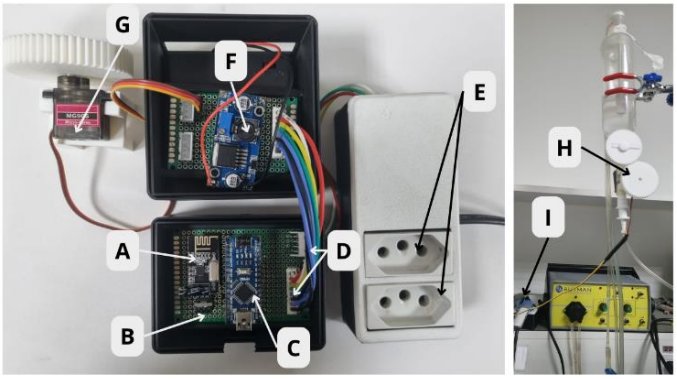


Figure 9. (A) nRF24L01 module; (B) 3.3 V voltage regulator; (C) Arduino Nano; (D) JST connectors; (E) dual-channel relay module connected to two power outlets; (F) servo motor power supply; (G) MG90S servo motor; (H) mechanical coupling between the servo motor and the flow control valve; (I) peristaltic pump and heat exchanger circulation pump connected to the Control Module.

The graph shown in Figure 10 demonstrates the effectiveness of the proportional–integral–derivative (PID) controller in regulating the perfusate flow rate through a servo motor coupled to the control valve, maintaining the variable at the predefined setpoint. In a control experiment performed without a perfused heart, a command to maintain the flow rate at 30 mL/min was issued at 16:36, and the system reached a stable steady state after approximately 2 minutes of minor oscillations. Subsequently, at 16:41, the flow setpoint was changed to 10 mL/min, again reaching stability after approximately two minutes. At 16:46, the flow rate was returned to 30 mL/min, and at 16:51, it was increased to 50 mL/min; in both cases, the controller stabilized the flow within approximately two minutes.

At 17:00, 17:01, and 17:02, successive forced closures of the flow path were performed to evaluate the system's response to potential obstructions in the perfusion line. Throughout the experiment, the fluid temperature remained between 37 °C and 41 °C, regulated by a relay connected to the hot-water circulation pump of the heat exchanger (Figure 5C). This relay automatically switched the pump on or off based on continuous temperature monitoring (Figure 3). Furthermore, when the monitored variables were intentionally driven beyond their predefined operating limits, the audible alarm was successfully activated.

A complementary analysis using the Pearson correlation coefficient ($r = 0.7251$) confirmed a positive relationship between the perfusate flow rate and fluid temperature, as illustrated in Figure 10. Under low-flow conditions (approximately 10 mL/min), a corresponding decrease in fluid temperature was observed. In contrast, at higher flow rates, the temperature stabilized at 41 °C, demonstrating the effectiveness of the thermal control system in compensating for the increased thermal demand imposed by higher perfusion flow rates.

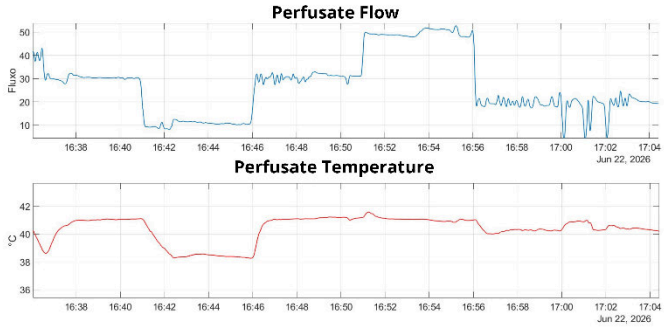


Figure 10. Temporal analysis of the relationship between perfusate flow rate and perfusate temperature during a control experiment performed without the perfused heart.

The results presented in this section demonstrate the development of a modular and configurable system capable of real-time monitoring of critical physiological variables, including coronary perfusion pressure, perfusate flow rate, and perfusate temperature. The monitoring system and its graphical user interface (GUI) were evaluated through a series of seven experiments using the Langendorff rabbit heart model. The control system is currently in the pre-experimental testing phase, regulating perfusate flow rate and

temperature without a perfused rabbit heart. Once fully integrated into the experimental setup, it will enable a more comprehensive analysis of the behavior and influence of these critical variables throughout each experiment.

IV. Conclusion

The results obtained demonstrate the feasibility of a scalable, and scientifically sound solution with direct application to an experimental setup for the study of cardiac arrhythmias in animal models. The proposed system enables the real-time monitoring and control of critical experimental variables, thereby reducing manual errors, improving the accuracy of experimental parameters, and increasing the efficiency of physiological control throughout the experiments.

Acknowledgments

This work was partially supported by the São Paulo Research Foundation (FAPESP) through grants awarded to M. Santos (Grant No. 2025/22698-7) and J. Salinet (Grant No. 2018/25606-2), as well as by CNPq INCT INTERAS (Call No. 58/2022). Additional support was provided by the Brazilian National Council for Scientific and Technological Development (CNPq) during the initial phase of the project.

The authors would like to thank all members of the HEartLab at the Federal University of ABC (UFABC) for developing the original experimental Langendorff setup that served as the foundation for this undergraduate research project.

References

[1] M. V. N. Alves *et al.*, "Arritmias cardíacas e prevenção de morte súbita," in *Cardiologia Contemporânea: Avanços, Desafios e Perspectivas na Prevenção, Diagnóstico e Tratamento das Doenças Cardiovasculares*, 1st ed. Brasil: Editora Humanize, 2025, pp. 107-120. doi: 10.5281/zenodo.17527565.

[2] M. Skrzypiec-Spring, B. Grotthus, A. Szelag, and R. Schulz, "Isolated heart perfusion according to Langendorff—still viable in the new millennium," *Journal of Pharmacological and Toxicological Methods*, vol. 55, pp. 113-126, 2007. doi: 10.1016/j.vascn.2006.05.004.

[3] H.-G. Zimmer, "The isolated perfused heart and its pioneers," *Physiology*, vol. 13, no. 4, pp. 203-210, Aug. 1998. doi: 10.1152/physiologyonline.1998.13.4.203

[4] S. T. Puente, A. Úbeda, and F. Torres, "e-Health: Biomedical instrumentation with Arduino," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 1-6, Jul. 2017. doi: 10.1016/j.ifacol.2017.08.1724 [5] N. T. Tsebesebe, K. Mpofu, S. Sivarasu, and P. Mthunzi-Kufa, "Arduino-based devices in healthcare and environmental monitoring," *Discover Internet of Things*, vol. 5, no. 46, pp. 1-15, Apr. 2025. doi: 10.1007/s43926-025-00139-z

[6] A. Nduka *et al.*, "Internet of things based remote health monitoring system using Arduino," in *Anais INTERNATIONAL CONFERENCE ON I-SMAC (IOT IN SOCIAL, MOBILE, ANALYTICS AND CLOUD)*, Palladam: IEEE, 2019. doi: 10.1109/I-SMAC47947.2019.9032438.

[7] P. Debnath, A. Mahmud, A. K. Hossain, and S. M. I. Rahman, "Design and application of IOT-based real-time patient telemonitoring system using biomedical sensor network," *SN Computer Science*, vol. 4, pp. 1-23, 2022. doi: 10.1007/s42979-022-01516-z

[8] J. Siles, V. Silva, T. Neves, I. Uzelac, I. Sandoval, G. Weber, and J. Salinet, "An integrated platform for 2-D and 3-D optical and electrical mapping of arrhythmias in Langendorff-perfused rabbit hearts," *The Journal of Physiology*, vol. 598, no. 15, pp. 3201-3215, Aug. 2020. doi: 10.1113/JP287815.

[9] D. M. Yellon, E. Pasini, A. Cargnoni, M. S. Marber, D. S. Latchman, and R. Ferrari, "The protective role of heat stress in the ischaemic and reperfused rabbit myocardium," *Journal of Molecular and Cellular Cardiology*, vol. 24, no. 8, pp. 895-907, Aug. 1992. doi: 10.1016/0022-2828(92)91102-b

[10] H. H. G. Hiwatashi *et al.*, "Fibrilação atrial: fisiopatologia, manifestações clínicas, estratégias diagnósticas e abordagem terapêutica atual," *Brazilian Journal of Implantology and Health Sciences*, vol. 8, no. 1, pp. 759-772, Jan. 2026. doi: 10.36557/2674-8169.2026v8n1p759-772..

[11] Honeywell, "Basic Board Mount Pressure Sensors, ABP Series - High Accuracy, Compensated/Calibrated," Datasheet, Charlotte, NC: Honeywell International Inc., 2021. [Online]. Available: https://automation.honeywell.com/content/dam/honeywell-edam/sps/common/en-us/industries/healthcare-and-life-sciences/medical-equipment/documents/sps-medical-abp-series.pdf

[12] Sensirion, "Datasheet SLF3S-4000B: Liquid Flow Sensor for High Flow Rates," Stäfa: Sensirion AG, 2023. [Online]. Available: https://sensirion.com/resource/datasheet/slf3s-4000b

[13] Espressif Systems (Shanghai), "ESP32-WROOM-32 Series Datasheet," v3.4, Shanghai: Espressif Systems, 2023. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32-wroom-32_datasheet_en.pdf

[14] Nordic Semiconductor, "nRF24L01+ Single Chip 2.4GHz Transceiver Product Specification," v1.0, Trondheim: Nordic Semiconductor, 2008. [Online]. Available: https://infocenter.nordicsemi.com/pdf/nRF24L01P_PS_v1.0.pdf

[15] TMRh20, "RF24: Arduino Driver for nRF24L01," v1.4, GitHub repository, 2023. [Online]. Available: https://github.com/nRF24/RF24

[16] Arduino, "Arduino Nano: Product Reference Manual," Ivrea: Arduino SA, 2023. [Online]. Available: https://docs.arduino.cc/resources/datasheets/A000005-datasheet.pdf

[17] O. Araque, L. A. Sánchez-Echeverri, and I. X. Cerón, "Impact of printing angle and layer height on the mechanical strength of PLA reinforced with chopped carbon fibres using FDM 3D printing," Polymers, vol. 17, no. 22, Art. no. 3069, Nov. 2025, doi: 10.3390/polym17223069.

[18] E. Hozdić and R. Hasanagić, "Analysis of the impact of cooling lubricants on the tensile properties of FDM 3D printed PLA and PLA+CF materials," Polymers, vol. 16, no. 15, Art. no. 2228, Aug. 2024, doi: 10.3390/polym16152228.

[19] Components101, "MG90S—Metal Gear Micro Servo Motor," Components101, Mar. 30, 2019. [Online]. Available: MG90S Servo Motor Datasheet. [Accessed: Jul. 22, 2026].

[20] Components101, "MG996R Servo Motor," Components101, Apr. 3, 2019. [Online]. Available: MG996R Servo Motor Datasheet. [Accessed: Jul. 22, 2026]

Exploratory Qualitative Screening of Benzoylecgonine and THC-COOH in Raw Sewage from Socioeconomically Distinct Areas of São Luís-Ma.

1st Cristovam Dervalmar Rodrigues Teixeira Filho
Biomedical Engineering
Universidade Brasil
São Luíz-Ma, Brazil
cristovamfilho17013@gmail.com

2nd Mardoqueu Martins da Costa
Universidade Brasil
São Paulo-SP, Brazil
mardoqueu.costa@universidadebrasil.edu.br

***Abstract*—Wastewater-Based Epidemiology (WBE) is a complementary approach for population-level surveillance of biomarkers associated with illicit drug use. This exploratory cross-sectional study qualitatively screened raw sewage from three wastewater treatment plants (WWTPs) serving socioeconomically distinct areas of São Luís-Ma, for cocaine- and cannabis-related immunoassay signals associated with benzoylecgonine (BE) and THC-COOH. Grab samples collected over the course of one week at each wastewater treatment plant were used and analyzed using rapid lateral-flow immunoassays originally developed for human urine. The manufacturer decision thresholds were 150 ng/mL for the cocaine-related assay and 50 ng/mL for the cannabis-related assay. Because the assays were not analytically validated for raw wastewater, these thresholds were used only for qualitative classification and were not interpreted as measured wastewater concentrations. BE-associated positivity was observed in 71% of Vinhais samples and 100% of Anil and Bacanga samples, whereas THC-COOH-associated positivity was 14%, 43%, and 43%, respectively. Pairwise Fisher's Exact Tests showed no statistically significant differences among WWTPs ($p > 0.05$). Therefore, the data do not support quantitative estimates of community drug consumption or an association between socioeconomic conditions and consumption. The findings were interpreted as preliminary qualitative screening results and require confirmation using matrix-validated quantitative methods and representative 24-h composite sampling.**

***Index Terms*—Wastewater-Based Epidemiology; qualitative screening; benzoylecgonine; THC-COOH; raw sewage; immunoassay; Fisher's Exact Test.**

I. Introduction

Illicit drug use is an important public health issue in Brazil. National survey data indicate substantial lifetime exposure to illicit psychoactive substances, including cannabis [5], [17], while official seizure data in Maranhão reinforce the regional relevance of cocaine- and cannabis-related markets [12] . Conventional surveillance methods, including surveys, clinical records, and seizure statistics, are useful but may be affected by underreporting, response bias, delays, and limited temporal or spatial resolution [8], [15].

Wastewater-Based Epidemiology (WBE) has emerged as a non-invasive complementary strategy for monitoring population-level biomarkers in sewage [6], [8], [15]. Quantitative WBE typically relies on validated measurements of biomarker concentration, representative wastewater sampling, flow data, contributing population, biomarker stability, and appropriate excretion/correction factors [6], [15].

Benzoylecgonine (BE) is commonly used as a cocaine-related biomarker, whereas THC-COOH is used in cannabis-related wastewater studies [9], [15].

The present study examined raw sewage from the Vinhais, Anil, and Bacanga WWTPs, which serve urban areas with different socioeconomic characteristics. These contextual differences motivated an exploratory comparison of qualitative detection frequencies. However, socioeconomic characteristics were not treated as evidence of individual exposure or as a causal determinant of drug use.

The objective was therefore to evaluate, as an exploratory qualitative screening study, the occurrence of cocaine- and cannabis-related immunoassay signals associated with BE and THC-COOH in raw sewage from the three WWTPs and to compare their detection frequencies using Fisher's Exact Test.

II. Background

A. Wastewater-Based Epidemiology and Analytical Context

Domestic sewage contains urine, feces, suspended solids, organic matter, detergents, and numerous chemical constituents [7]. Biomarkers excreted after drug use may enter sewer networks and reach WWTP influents, where they can be investigated for public health surveillance. Nevertheless, raw wastewater is a complex analytical matrix, and biomarker measurements may be affected by dilution, degradation, adsorption to solids, sewer travel time, and matrix interferences [15].

Validated quantitative WBE studies generally use representative composite sampling and instrumental analytical methods. The literature emphasizes the importance of uncertainty control in sampling, analytical recovery, biomarker

stability, flow measurement, population estimation, and back-calculation [6], [15]. These requirements differ fundamentally from qualitative screening based on a decision threshold.

In Brazil, WBE investigations have been reported in Brasília and other regions [1], [2], [9], [13], [14]. The present work differs by evaluating the feasibility and limitations of an initial qualitative immunoassay screening approach in São Luís. Because the assay used here was originally designed for urine and was not validated for raw wastewater, its findings are treated as preliminary screening observations rather than validated quantitative WBE estimates.

B. Study Areas and Socioeconomic Context

The Vinhais, Anil, and Bacanga WWTPs serve areas with contrasting socioeconomic characteristics according to the contextual information compiled in the original study from IBGE, SNIS, CAEMA projects, and author synthesis. These characteristics were retained only to describe the study setting and were not used to infer individual drug use or causal socioeconomic effects.

TABLE I
CONTEXTUAL SOCIOECONOMIC CHARACTERISTICS OF THE AREAS SERVED BY THE WWTPS.

Variable	***Vinhaes***	***Anil***	***Bacanga***
Approx. household income	3–6 minimum wages	1–3 minimum wages	0.5–3 minimum wages
Predominant profile	Middle/upper	Middle/lower	Low
Estimated sewer coverage	$> 60\%$	$< 50\%$	100%
Occupation type	Planned/regular	Mixed	Planned/socialhousing

Autor.

III. MATERIALS AND METHODS

A. Study Design and Sampling

The study had an exploratory, cross-sectional design. Raw sewage was collected at the inlets of the Vinhais, Anil, and Bacanga wastewater treatment plants over the course of a week during distinct periods. Seven grab samples were evaluated per WWTP. The sampling design was intended for initial qualitative pre-screening and did not constitute 24-h composite sampling or long-term temporal monitoring.

The possible influence of recreational, cultural, regional, or sporting events was not systematically controlled in the initial sampling design. Consequently, the data cannot be used to infer stable temporal patterns of drug use. A subsequent phase of the broader research project uses 24-h composite samples and instrumental quantitative analysis; because those analyses are still ongoing, those data are not included in this manuscript.

B. Qualitative Immunoassay Screening

Samples were collected in 250 mL PVC bottles, filtered, and applied to rapid visual lateral-flow immunoassays supplied by SABRE SAFETY LTDA. A critical methodological procedure was pH monitoring using colorimetric strips. It was observed that the pH of the raw sewage ranged from 6.0 to 8.0. Maintaining this near-neutral range is essential to ensure the integrity of the antibodies immobilized on the nitrocellulose strip, as extreme pH values can cause protein denaturation and compromise antigen-antibody affinity, leading to false-negative results [18]. According to the product labeling used in the study, the assays were intended for qualitative screening of cocaine-related (COC) and cannabis-related (THC) targets in human urine, with manufacturer decision thresholds of 150 ng/mL and 50 ng/mL, respectively.

The assays were not analytically validated in this study for raw wastewater. Therefore, the manufacturer cut-offs were not treated as measured concentrations of BE or THC-COOH and were not used to calculate mass loads or community consumption. Positive/negative results are interpreted as exploratory cocaine-related and cannabis-related immunoassay responses associated with the target biomarkers discussed in the study.

C. Statistical Analysis

Qualitative outcomes were coded as 0 (absence/negative) and 1 (presence/positive). Detection frequencies were summarized for each WWTP. Because the sample size was small and both variables were categorical, two-sided Fisher's Exact Tests were used for pairwise comparisons between Vinhais and Anil, Vinhais and Bacanga, and Anil and Bacanga [16]. The null hypothesis was that the detection frequency did not differ between the two WWTPs being compared. Statistical significance was set at 5% ($p < 0.05$).

IV. RESULTS AND DISCUSSION

The cocaine-related/BE-associated immunoassay was positive in 5 of 7 samples from Vinhais (71.4%) and in all 7 samples from both Anil and Bacanga (100%). The cannabis-related/THC-COOH-associated immunoassay was positive in 1 of 7 Vinhais samples (14.3%) and in 3 of 7 samples from both Anil and Bacanga (42.9%). Thus, Anil and Bacanga showed higher observed detection frequencies than Vinhais, particularly for the cannabis-related assay.

TABLE II
QUALITATIVE DETECTION FREQUENCIES BY WWTP.

WWTP	***BE-associated Positive***	***BE-associated Negative***	***THC-COOH-associated Positive***	***THC-COOH-associated Positive***
Vinhais	$5/7(71.4\%)$	$2/7(28.6\%)$	$1/7(14.3\%)$	$6/7(85.7\%)$
Anil	$7/7(100\%)$	$0/7(0\%)$	$3/7(42.9\%)$	$4/7(57.1\%)$
Bacanga	$7/7(100\%)$	$0/7(0\%)$	$3/7(42.9\%)$	$4/7(57.1\%)$

Autor.

However, all pairwise Fisher's Exact Test p-values were greater than 0.05. For the cocaine-related/BE-associated results, p = 0.462 for Vinhais versus Anil, p = 0.462 for Vinhais versus Bacanga, and p = 1.000 for Anil versus Bacanga. For the cannabis-related/THC-COOH-associated results, p = 0.559, p = 0.559, and p = 1.000, respectively. The null hypothesis was therefore not rejected for any comparison.

TABLE III
PAIRWISE FISHER'S EXACT TEST RESULTS

Target	***WWTP Comparison***	***p-value***	***Interpretation***
BE-associated	Vinhais × Anil	0.462	Not significant
BE-associated	Vinhais × Bacanga	0.462	Not significant
BE-associated	Anil × Bacanga	1.000	Not significant
THC-COOH-associated	Vinhais × Anil	0.559	Not significant
THC-COOH-associated	Vinhais × Bacanga	0.559	Not significant
THC-COOH-associated	Anil × Bacanga	1.000	Not significant

Autor.

The higher descriptive detection frequencies observed at Anil and Bacanga cannot be interpreted as evidence of higher drug consumption or as an effect of socioeconomic vulnerability. The qualitative design, small sample size, grab sampling, and lack of wastewater-specific analytical validation preclude such inference. The socioeconomic profile is therefore interpreted only as contextual information describing the areas served by each WWTP.

The higher detection frequency observed for the cocaine-related/BE-associated signal compared with the cannabis-related/THC-COOH-associated signal may reflect differences in biomarker excretion and physicochemical behavior. BE is widely used as a biomarker for cocaine monitoring in wastewater, whereas THC-COOH presents greater analytical challenges in wastewater-based surveillance. Its relatively low excretion, together with partitioning to suspended particulate matter and other matrix-related effects, may reduce its availability in the aqueous fraction and consequently affect detection. Therefore, the lower frequency of cannabis-related/THC-COOH-associated positive results observed in this study should not be interpreted as evidence of lower cannabis use in the populations investigated. Rather, it may partly reflect differences in excretion, wastewater partitioning, dilution, and the analytical characteristics of the qualitative immunoassay.

V. CONCLUSION

This exploratory study detected cocaine-related and cannabis-related immunoassay signals in raw sewage from the three investigated WWTPs in São Luís. The observed positivity frequencies differed descriptively among WWTPs, but Fisher's Exact Tests did not identify statistically significant pairwise differences.

Because the immunoassays were originally designed for urine, were not validated for raw wastewater, and were applied to grab samples collected week periods, the findings must be interpreted strictly as preliminary qualitative screening results.

The study nevertheless provides an initial feasibility assessment for qualitative wastewater screening in a setting with limited analytical infrastructure. More representative sampling and validated quantitative methods, preferably incorporating SPE and LC-MS/MS, are required to confirm biomarker identity, quantify concentrations, and support robust WBE estimates.

REFERENCES

[1] A. S. C. Guedes, "Drogas de abuso e seus metabólitos em águas residuárias coletadas durante o Réveillon do Distrito Federal," Master's thesis, Inst. de Química, Univ. de Brasília, Brasília, 2024.

[2] A. Pacheco-Ferreira, E. D. Wermelinger, and M. J. Cruz-Hernández, "Avaliação da presença e variabilidade de drogas ilícitas e metabólitos em estações de tratamento de esgoto do Sudeste do Brasil: um estudo abrangente de WBE," J. Hum. Environ. Health Promot., vol. 9, no. 2, pp. 63–71, 2023.

[3] CAEMA-ESSE, "Projetos executivos do sistema de esgotamento sanitário do município de São Luís: etapa 1 – sistemas São Francisco, Anil, Vinhais e Bacanga," vol. I, rev. A, São Luís: CAEMA, Jan. 2009.

[4] CAEMA-ESSE, "Estudos de concepção para o sistema de distribuição de água da Grande São Luís – MA: conclusão do estudo de concepção," vol. I, part A, São Luís: CAEMA, 2013.

[5] "Consumo de drogas ilícitas cresceu cerca de 80% no Brasil de 2012 a 2023," Revista Pesquisa FAPESP, 2025.

[6] E. Zuccato et al., "Estimating community drug abuse by wastewater analysis," Environ. Health Perspect., vol. 116, no. 8, pp. 1027–1032, 2008.

[7] E. P. Jordão and C. A. Pessoa, Tratamento de Esgotos Domésticos, 8th ed. Rio de Janeiro: ABES, 2017.

[8] European Union Drugs Agency, Wastewater analysis and normalization in WBE studies. Lisbon: EUDA Publications Office, 2025.

[9] F. F. Sodré et al., "Consumo de cocaína e maconha durante e após o Carnaval de 2019 em Brasília por meio da epidemiologia baseada em esgoto," Front. Anal. Sci., vol. 2, pp. 1–12, 2022.

[10] K. M. Paiva, "Epidemiologia de esgoto como ferramenta para estimativa do consumo de maconha," Undergrad. thesis, Univ. Fed. de Mato Grosso, Cuiaba, 2022.

[11] L. P. Favero and P. Belfiore, Manual de Análise de Dados: Estatística e Modelagem Multivariada em Excel, SPSS e Stata, 1st ed. Rio de Janeiro:Elsevier, 2017.

[12] Maranhão, "Forças de segurança do Maranhão retiram quatro toneladas de drogas de circulação e enfraquecem o narcotráfico com prejuízo," Secretaria de Estado da Segurança Pública, 2025.

[13] R. F. Silva, "Determinação de cocaína e benzoilecgonina em amostras de esgoto bruto do Distrito Federal como suporte à realização de estimativas de consumo," Master's thesis, Univ. de Brasília, Brasília, 2012.

[14] R. Z. Hahn et al., "Monitoramento de longo prazo de padrões de consumo de drogas durante a pandemia de COVID-19 em uma comunidade de pequeno porte no Brasil por meio da WBE," Chemosphere, vol. 309, Art. no. 136611, 2022.

[15] S. Castiglioni et al., "Evaluation of uncertainties associated with determination of community drug use through measurement of sewage drug biomarkers," Environ. Sci. Technol., vol. 47, no. 3, pp. 1452–1460, 2013.

[16] S. Vieira, Bioestatística: Tópicos Avançados, 2nd ed. Rio de Janeiro: Elsevier/Campus.

[17] UNIFESP, "Caderno temático LENAD III - Cannabis," 2025.

[18] R. Wong and H. Tse, Lateral Flow Immunoassay. New York: Humana Press, 2009.

CARACTERIZAÇÃO FRACTAL DA LAMININA POLIMERIZADA: IMPLICAÇÕES PARA SINALIZAÇÃO CELULAR, REGENERAÇÃO NEURAL E NEUROENGENHARIA

Istephany Vitória Batista Gomes Freitas
Laboratório de Estudos, Simulações e Sínteses de Materiais (LESSM.)
Engenharia Biomédica (EB.)
Universidade Federal do Norte do Tocantins (UFNT.)
Araguaína, Brasil
istephany.freitas@ufnt.edu.br

Matheus Pereira Lobo
Laboratório de Estudos, Simulações e Sínteses de Materiais (LESSM.)
Física (FIS.)
Universidade Federal do Norte do Tocantins (UFNT.)
Araguaína, Brasil
matheus.lobo@ufnt.edu.br

***Resumo*—**
A laminina polimerizada (*polyLM*) apresenta uma organização supramolecular complexa capaz de influenciar processos de sinalização celular e regeneração neural. Neste trabalho, realizou-se uma revisão bibliográfica com o objetivo de investigar como métricas fractais e modelos de crescimento não linear podem caracterizar a organização estrutural da polyLM e auxiliar na compreensão de suas propriedades funcionais. Foram analisados estudos relacionados à geometria fractal, à estimativa da dimensão fractal por meio do método de *box-counting* e às aplicações da polyLM em contextos de neuroengenharia e regeneração nervosa. Os resultados indicam que a polyLM apresenta comportamento fractal, com dimensão variando aproximadamente entre 1,55 e 1,70 durante sua formação, exibindo padrões compatíveis com modelos de Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA). Além disso, a literatura sugere que sua organização geométrica favorece a exposição de sítios de ligação celular, contribuindo para processos de adesão, diferenciação neuronal e crescimento de neuritos. Conclui-se que a geometria fractal da polyLM constitui um importante elo entre estrutura molecular e função neurobiológica, destacando o potencial de abordagens geométricas para o desenvolvimento de biomateriais aplicados à regeneração neural e à neuroengenharia.

***Palavras-chave*—Fractais, Matriz extracelular, Polilaminina, Neuroengenharia, Dimensão Fractal, Neurociência**

I. Introdução

A organização estrutural da matriz extracelular exerce papel fundamental na regulação de processos celulares, influenciando mecanismos de adesão, diferenciação, migração e regeneração tecidual. Nesse contexto, a laminina (*LM*) destaca-se como uma das principais glicoproteínas da matriz extracelular, participando da organização de membranas basais e da sinalização celular em diferentes tecidos [1].

Estudos recentes demonstram que a laminina polimerizada (*polyLM*) apresenta uma organização supramolecular complexa caracterizada por propriedades fractais, incluindo autosimilaridade e dimensão não inteira [2]. Segundo Hochman-Mendez, M. et al. [2], a morfologia da polyLM mantém padrões semelhantes em diferentes escalas de observação, indicando um comportamento compatível com sistemas fractais encontrados em diversos fenômenos naturais e biológicos.

Na Engenharia Biomédica e na Neuroengenharia, a compreensão dessas propriedades geométricas torna-se relevante devido à influência da organização espacial da matriz extracelular sobre a sinalização celular e o crescimento neural. Trabalhos anteriores demonstram que a estrutura fractal da polyLM favorece a exposição de sítios de ligação celular e contribui para a formação de nichos neurogênicos especializados, como os *fractones*, associados à diferenciação neuronal e à regeneração do tecido nervoso [1], [3].

Além disso, métricas fractais e modelos de crescimento não linear têm sido utilizados para descrever a formação e a evolução temporal de estruturas biológicas complexas, permitindo relacionar propriedades geométricas a funções biológicas específicas [4], [5]. Dessa forma, sua geometria fractal surge como um possível elo entre organização molecular e funcionalidade neurobiológica.

Diante desse contexto, o presente trabalho busca responder à seguinte questão de pesquisa: **Como métricas fractais e modelos de crescimento não linear podem contribuir para a compreensão da organização estrutural e funcional da laminina polimerizada (*polyLM*)?**

Para responder a essa questão, foi realizada uma revisão bibliográfica de estudos relacionados à caracterização fractal da polyLM, aos métodos de estimativa de dimensão fractal e às suas aplicações em neuroengenharia e regeneração neural.

II. Objetivos

Investigar como métricas fractais e modelos de crescimento não linear podem contribuir para a compreensão da

Trabalho desenvolvido com apoio da Universidade Federal do Norte do Tocantins por meio do programa PIBIC.

organização estrutural e funcional da polyLM, correlacionando propriedades geométricas com processos de sinalização celular, regeneração neural e aplicações em neuroengenharia.

A. Objetivos Específicos

- Descrever as principais características fractais da polyLM reportadas na literatura, incluindo propriedades de auto-similaridade e dimensão fractal.
- Analisar os métodos utilizados para caracterizar a complexidade geométrica da polyLM, com ênfase na estimativa da dimensão fractal por meio do método de *box-counting*.
- Comparar os valores de dimensão fractal da polyLM com padrões matemáticos e modelos físicos de crescimento, como a Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA).
- Investigar a relação entre a organização geométrica da polyLM e seus efeitos sobre processos de sinalização celular, diferenciação neuronal e regeneração neural descritos na literatura.
- Discutir limitações metodológicas das técnicas de caracterização fractal e apresentar perspectivas futuras para a aplicação de métricas geométricas no desenvolvimento de biomateriais voltados à neuroengenharia.

III. Fundamentação Teórica

A. Geometria Fractal

De acordo com Feldman [5], os fractais são estruturas geométricas complexas caracterizadas por propriedades como auto-similaridade, invariância de escala e dimensão não inteira. Diferentemente das formas descritas pela geometria euclidiana clássica, os fractais apresentam padrões que se repetem em diferentes escalas de observação, mantendo semelhanças estatísticas entre suas partes e o todo. Essa propriedade permite descrever sistemas naturais marcados por irregularidade, ramificação e elevada complexidade estrutural.

Segundo Barreiras [6], uma das principais características dos fractais é a presença de uma dimensão de Hausdorff não inteira, utilizada como métrica quantitativa da complexidade geométrica. Estruturas fractais podem apresentar dimensões intermediárias entre linhas e superfícies planas, permitindo descrever objetos biológicos que não podem ser representados adequadamente pela geometria convencional.

Conforme discutido por Oliveira [7], muitos fractais são gerados por processos iterativos capazes de produzir estruturas complexas a partir de regras simples de crescimento. Em sistemas biológicos, esses modelos auxiliam na compreensão de padrões de organização espacial e evolução morfológica observados em diferentes escalas, permitindo relacionar propriedades geométricas com características funcionais dos tecidos.

Entre os exemplos clássicos de fractais descritos na literatura, encontram-se a Curva de Koch [4], o Conjunto de Cantor [6] e o Triângulo de Sierpiński [7], frequentemente utilizados como modelos matemáticos para validação de métodos de cálculo de dimensão fractal, como o método de *box-counting*.

B. Dimensão de Hausdorff, Box-Counting e Agregação Limitada por Difusão

A dimensão fractal é uma métrica utilizada para caracterizar a complexidade geométrica e a capacidade de preenchimento espacial de estruturas irregulares. Entre as diferentes definições existentes, a dimensão de Hausdorff-Besicovitch (D_H) apresenta uma formulação matemática rigorosa para descrever conjuntos que possuem propriedades de auto-semelhança e dimensões não inteiras, permitindo a caracterização de estruturas cuja complexidade não pode ser representada adequadamente pelas dimensões euclidianas tradicionais [5], [6].

A dimensão de Hausdorff é definida a partir da medida de Hausdorff s-dimensional, baseada na construção de coberturas sucessivas de um conjunto por elementos de diâmetro menor ou igual a uma escala ϵ. O valor crítico de s, no qual a medida de Hausdorff sofre uma transição de infinito para zero quando $\epsilon \to 0$, corresponde à dimensão do conjunto analisado [5], [6]. Apesar de sua fundamentação matemática, o cálculo direto de D_H apresenta limitações computacionais para dados experimentais, pois envolve a avaliação de infinitas possibilidades de cobertura. Dessa forma, métodos aproximativos, como a dimensão de *box-counting*, são amplamente utilizados para estimar a dimensão fractal de imagens e estruturas biológicas discretizadas [2], [8].

A dimensão de *box-counting* (D_B), também denominada dimensão de capacidade ou dimensão de Minkowski-Bouligand, baseia-se na análise da ocupação espacial de um objeto em diferentes escalas. O método consiste na sobreposição de uma grade composta por caixas de tamanho r sobre a estrutura analisada, sendo $N(r)$ o número de caixas necessárias para cobrir o conjunto. A dimensão fractal é obtida pelo limite:

$$D_B(F) = \lim_{r \to 0^+} \frac{\ln N(r)}{\ln(1/r)} \tag{1}$$

Para aplicações em imagens digitais, a dimensão é estimada pela inclinação da reta obtida no gráfico log-log entre $\ln(N(r))$ e $\ln(1/r)$, utilizando regressão linear para determinar o coeficiente angular. O processamento das imagens normalmente envolve etapas de pré-tratamento, como equalização de histograma, segmentação binária e aplicação de diferentes escalas de grade para determinar a ocupação espacial da estrutura analisada [2], [8].

Embora a dimensão de Hausdorff seja a definição matemática mais geral, a dimensão de *box-counting* apresenta maior aplicabilidade computacional em sistemas reais. Para conjuntos matemáticos ideais, ambas podem apresentar valores equivalentes; entretanto, para estruturas gerais, observa-se a relação $D_H \leq D_B$, indicando que a dimensão de *box-counting* pode atuar como uma estimativa superior da dimensão fractal verdadeira [5]. Além disso, estruturas biológicas apresentam limitações associadas à resolução experimental e à ausência de detalhamento infinito, fazendo com que a análise fractal seja válida apenas dentro de determinados intervalos de escala [4].

Na caracterização da laminina polimerizada (*polyLM*), Hochman-Mendez *et al.* [2] utilizaram o método de *box-counting* para quantificar a organização fractal da matriz

supramolecular. O procedimento experimental envolveu a equalização das imagens microscópicas, binarização pelo método de Otsu e análise de múltiplos recortes para aumentar a estabilidade estatística da estimativa. Os resultados demonstraram que a *polyLM* apresenta dimensões fractais entre aproximadamente $1,55$ e $1,70$, enquanto a laminina não polimerizada apresentou comportamento compatível com estruturas de preenchimento bidimensional, com dimensão próxima a $2,0$ [2].

A organização fractal observada na *polyLM* pode ser interpretada por meio do modelo de Agregação Limitada por Difusão (*Diffusion-Limited Aggregation* – DLA). Esse modelo descreve processos nos quais partículas submetidas a movimentos aleatórios por difusão se agregam progressivamente a uma estrutura inicial, produzindo padrões ramificados e com propriedades fractais [4]. Devido à sua capacidade de gerar arquiteturas complexas a partir de regras simples, o DLA tem sido empregado para modelar processos de crescimento biológico, incluindo a formação de redes neurais e padrões de ramificação celular [4].

A proximidade entre a dimensão fractal característica do modelo DLA e os valores obtidos para a *polyLM* sugere que sua formação apresenta mecanismos de auto-organização compatíveis com esse modelo [2]. Essa arquitetura ramificada pode favorecer uma elevada relação superfície-volume e maior disponibilidade de regiões funcionais da matriz extracelular, contribuindo para processos de interação célula-matriz, crescimento de neuritos e regeneração neural [1], [3], [9].

C. Laminina Polimerizada: Estrutura Fractal e Função Biológica

A laminina polimerizada (*polylaminin* ou *polyLM*) constitui um modelo biológico de interesse para o estudo da organização da matriz extracelular e de sua influência sobre a regeneração neural. Segundo Hochman-Mendez et al. [2], sua formação ocorre em meio ácido, favorecendo a interação dos três braços curtos da molécula heterotrimérica da laminina e originando uma rede supramolecular altamente ramificada.

Diferentemente da laminina convencional, que tende a formar agregados mais compactos e contínuos, a polyLM apresenta uma arquitetura compatível com estruturas fractais, caracterizada por ramificações distribuídas em múltiplas escalas e elevada conectividade. Essas características aproximam sua organização dos modelos de Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA), discutidos na subseção anterior [2], [9].

Além dos aspectos estruturais, a organização da polyLM exerce influência sobre processos celulares. Mesquita et al. [7] demonstram que essa arquitetura favorece a exposição dos domínios de ligação da laminina, intensificando a interação com receptores celulares, como integrinas e distroglicanas. Como consequência, observa-se maior eficiência em processos de adesão celular, crescimento de neuritos e regeneração neural.

Assim, a polyLM representa um modelo biológico que evidencia a relação entre organização estrutural e função celular. Sua arquitetura ramificada fornece um contexto adequado para investigar como propriedades geométricas da matriz extracelular podem influenciar mecanismos de sinalização e regeneração neural, servindo de base para a análise quantitativa apresentada nas seções seguintes.

D. Da Estrutura Fractal à Função Neurobiológica

Segundo Bianchi [4], a caracterização matemática de sistemas biológicos complexos permite descrever padrões de crescimento, organização espacial e conectividade por meio de modelos fundamentados na geometria fractal e na teoria dos sistemas complexos. Essas abordagens são particularmente relevantes em processos neurobiológicos, nos quais estrutura e função estão intimamente relacionadas em diferentes escalas espaciais.

Conforme discutido por López et al. [10], estruturas com organização fractal apresentam elevada eficiência na distribuição espacial de recursos, na conectividade e na propagação de sinais. A auto-semelhança observada nesses sistemas possibilita a formação de redes altamente conectadas com menor custo estrutural, característica frequentemente encontrada em diferentes sistemas biológicos.

Nesse contexto, a caracterização matemática da polyLM por meio da dimensão fractal fornece uma ferramenta para quantificar sua organização espacial e relacioná-la ao comportamento biológico observado experimentalmente. A proximidade entre a dimensão fractal da polyLM e aquela prevista por modelos de Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA) sugere que mecanismos físicos de auto-organização podem contribuir para a formação dessa arquitetura supramolecular [2], [9]. Essa interpretação estabelece uma ponte entre os modelos matemáticos apresentados anteriormente e a organização estrutural da matriz extracelular.

Sob essa perspectiva, a geometria fractal deixa de representar apenas uma característica morfológica e passa a constituir um parâmetro quantitativo para investigar a relação entre estrutura e função em sistemas biológicos. A integração entre modelos matemáticos e evidências experimentais amplia a compreensão de como propriedades geométricas da matriz extracelular podem influenciar processos de adesão celular, organização espacial de sinais bioquímicos e regeneração neural, reforçando o potencial da análise fractal como ferramenta de investigação em neuroengenharia.

IV. Metodologia

A presente pesquisa caracteriza-se como um estudo qualitativo, de natureza exploratória e fundamentação teórica, desenvolvido por meio de revisão bibliográfica e análise comparativa de modelos matemáticos e biológicos relacionados à geometria fractal, à laminina polimerizada (*polyLM*) e às suas implicações na sinalização celular, regeneração neural e neuroengenharia.

A metodologia adotada foi estruturada a partir da leitura, seleção e análise de artigos científicos, dissertações, teses e livros especializados relacionados à dinâmica de crescimento da polyLM, aos métodos de quantificação fractal e

às aplicações neurobiológicas associadas à organização da matriz extracelular. Entre os principais referenciais utilizados destacam-se os trabalhos de Hochman-Mendez et al. [2], [9], Siqueira-Santos et al. [3], Bianchi [4] e Feldman [5].

Segundo Hochman-Mendez, M. et al. [2], a caracterização da natureza fractal da polyLM foi realizada por meio da estimativa da dimensão de Hausdorff utilizando o método de *box-counting*. Conforme descrito pelos autores, o procedimento inicia-se com o pré-processamento digital das imagens de microscopia, incluindo equalização de histograma e binarização pelo método de Otsu, com o objetivo de isolar a estrutura proteica do fundo da imagem.

Após essa etapa, uma grade de caixas quadradas de tamanho variável é sobreposta à imagem binarizada, contabilizando-se o número de caixas que interceptam qualquer parte da estrutura analisada. O cálculo da dimensão fractal é então obtido pela relação logarítmica entre o tamanho da grade e o número de caixas ocupadas, conforme descrito na literatura pelo método de *box-counting*. De acordo com os trabalhos analisados, esse procedimento permite quantificar matematicamente a complexidade estrutural da laminina polimerizada e acompanhar sua evolução temporal.

Os estudos revisados também descrevem que a evolução da dimensão fractal da polyLM ocorre progressivamente ao longo do tempo de adsorção, apresentando valores aproximados de 1,55 após 1 hora, 1,62 após 8 horas e 1,70 após 12 horas [2]. Esses resultados foram utilizados neste trabalho como referência comparativa para a análise dos padrões de crescimento fractal descritos na literatura.

Além da caracterização geométrica da polyLM, foram analisados estudos que investigam as relações entre a organização fractal da matriz extracelular e processos neurobiológicos, incluindo adesão celular, crescimento de neuritos, formação de nichos neurogênicos e regeneração neural. Também foram revisados trabalhos relacionados à aplicação de substratos biomiméticos funcionalizados com laminina em processos regenerativos do nervo ciático, especialmente aqueles que avaliam a influência da topologia de filamentos de policaprolactona (PCL) funcionalizados na orientação axonal e recuperação funcional *in vivo* [3].

Apesar de sua ampla utilização na literatura para estimativa da dimensão fractal, o método de *box-counting* apresenta limitações que devem ser consideradas na interpretação dos resultados. A estimativa da dimensão fractal depende diretamente das etapas de segmentação e binarização da imagem, sendo sensível à escolha de limiares, à qualidade do contraste e aos procedimentos de pré-processamento empregados [4], [8], [11]. Técnicas como equalização de histograma, suavização de ruído e correção de artefatos podem modificar a distribuição espacial dos pixels e influenciar os valores obtidos para a dimensão fractal.

Além disso, a precisão do método está relacionada à resolução da imagem e à faixa de escalas analisada. Diferentemente dos fractais matemáticos ideais, estruturas biológicas apresentam tamanho finito e apenas regiões limitadas de comportamento fractal, o que pode restringir a validade das estimativas obtidas [4], [5]. Outro aspecto relevante refere-se à sensibilidade do ajuste linear utilizado no gráfico log-log do algoritmo, uma vez que pequenas variações na distribuição dos dados podem produzir diferenças significativas na inclinação estimada e, consequentemente, na dimensão fractal calculada.

Dessa forma, os valores de dimensão fractal reportados para a polyLM devem ser interpretados como estimativas dependentes das condições experimentais, da qualidade das imagens analisadas e dos procedimentos computacionais empregados, não constituindo uma medida absoluta da complexidade geométrica da estrutura.

Também foram revisadas abordagens complementares utilizadas na caracterização de estruturas biológicas complexas, incluindo análise multifractal, lacunaridade, análise de textura, esqueletização (*skeletonization*), análise de redes e grafos e métodos de estimação fractal multiescala. Embora essas técnicas não tenham sido aplicadas diretamente neste estudo, sua discussão permitiu identificar potenciais descritores capazes de fornecer informações adicionais sobre a organização espacial da matriz extracelular, complementando as limitações associadas ao uso exclusivo da dimensão fractal.

A análise dos materiais selecionados foi conduzida de forma interpretativa e comparativa, buscando correlacionar propriedades geométricas fractais da polyLM com aspectos funcionais relacionados à sinalização celular, à organização de nichos neurogênicos e à regeneração neural. Dessa forma, a metodologia adotada permitiu integrar evidências provenientes da matemática, da biologia e da engenharia biomédica para compreender como a geometria da matriz extracelular pode influenciar processos neurobiológicos.

V. Resultados e Discussões

Os resultados apresentados nesta seção foram organizados a partir das análises e interpretações descritas nos trabalhos de Menezes et al. [1], Hochman-Mendez et al. [2], [9], Siqueira-Santos et al. [3] e demais autores utilizados como fundamentação teórica desta pesquisa.

A. Padrões Fractais e Comparações Matemáticas

Os estudos analisados demonstram que a laminina polimerizada (polyLM) apresenta uma organização geométrica compatível com sistemas fractais físicos. Segundo Hochman-Mendez, M. et al. [2], ela mantém padrões morfológicos semelhantes mesmo sob ampliações de diferentes ordens de magnitude, evidenciando propriedades de auto-similaridade estrutural.

Para validar a natureza fractal da polyLM e comparar o método de cálculo da dimensão fractal por estimativa de Hausdorff, os autores utilizaram padrões matemáticos clássicos reconhecidos na literatura. Entre eles, destacam-se o Conjunto de Cantor, com dimensão aproximada de 0,6309, a Curva de Koch, com dimensão aproximada de 1,2619, e o Triângulo de Sierpiński, cuja dimensão fractal é próxima de 1,585 [5], [6].

Esses modelos servem como referências matemáticas para validação do algoritmo de *box-counting*, permitindo comparar a complexidade estrutural da polyLM com sistemas fractais

teóricos. Além disso, o fractal quádrico, com valores próximos de 1,50 e 1,67 dependendo da iteração, apresenta comportamento geométrico semelhante ao observado nas fases iniciais da polimerização da polyLM.

Os trabalhos de Hochman-Mendez, M. et al. [2] demonstram que a polyLM apresenta dimensão fractal variando entre 1,55 e 1,70, enquanto a laminina comum (LM), preparada em pH neutro, apresenta dimensão próxima de 2,0. Esses resultados evidenciam diferenças quantitativas entre as duas organizações estruturais e reforçam o potencial da dimensão fractal como parâmetro de caracterização da matriz extracelular.

A Fig. 1 evidencia diferenças marcantes entre a organização da laminina convencional e da polyLM. Enquanto a LM apresenta deposição mais compacta e contínua, a polyLM forma redes ramificadas e lacunadas distribuídas em múltiplas escalas, em concordância com os valores de dimensão fractal observados.

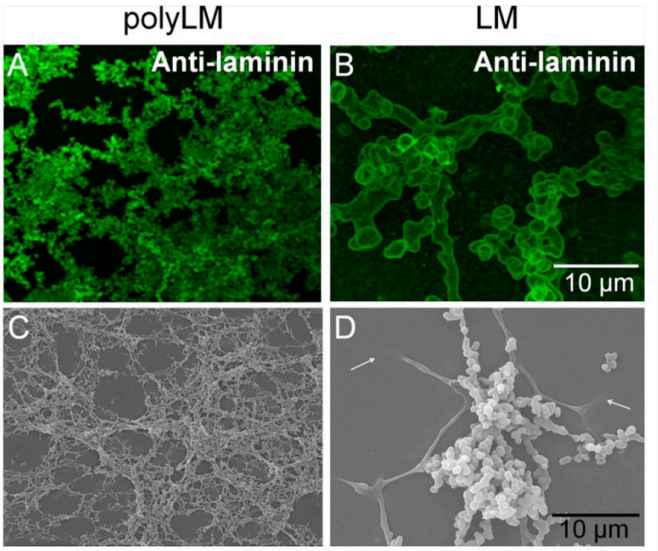


Figura 1. Estrutura tridimensional dos polímeros de laminina observada por microscopia confocal de fluorescência e microscopia eletrônica de varredura. A laminina foi incubada sobre lamínulas de vidro por 12 horas em meio ácido (polyLM) ou em tampão neutro (LM). (A, B) A imunofluorescência indireta foi realizada utilizando anticorpo policlonal anti-laminina. As imagens representam empilhamentos-z obtidos pela superposição de 74 cortes confocais renderizados pelo software 7.2.3. (C, D) Micrografias obtidas por microscopia eletrônica de varredura (MEV) dos polímeros em ampliação semelhante. As setas em D indicam depósitos lamelares de laminina. As barras de escala aplicam-se aos painéis A–D e representam 10 μm. Adaptado de Hochman-Mendez, M. et al. [2], sob licença Creative Commons CC BY.

Os resultados também indicam que a dimensão fractal da polyLM se aproxima daquela prevista pelos modelos de Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA), cujo valor típico é próximo de 1,7 [5], [9]. Essa proximidade reforça a hipótese de que mecanismos de auto-organização descritos por esse modelo físico contribuem para a formação da arquitetura supramolecular observada experimentalmente.

Além dos padrões matemáticos, os autores relacionam a organização da polyLM com estruturas biológicas naturais conhecidas como “fractones”, presentes em nichos neurogênicos do sistema nervoso central [1]. Essas estruturas apresentam organização espacial complexa responsável pela distribuição de fatores de crescimento e pela regulação da diferenciação celular.

Em conjunto, os resultados revisados demonstram que a geometria da polyLM é compatível com modelos matemáticos clássicos de crescimento fractal. A concordância entre as estimativas obtidas por *box-counting*, as referências fractais teóricas e o modelo DLA reforça a utilização da dimensão fractal como ferramenta para caracterização quantitativa de sistemas biológicos complexos.

B. Cinética de Crescimento e Evolução da Complexidade Fractal

Os trabalhos analisados demonstram que a cinética de crescimento da polyLM está diretamente relacionada à evolução de sua complexidade geométrica ao longo do processo de adsorção [2], [9].

A polyLM apresenta cinética de adsorção linear e organizada, atingindo aproximadamente 45% de adsorção após 20 minutos e cerca de 90% após 12 horas. Em contraste, a laminina convencional apresenta deposição rápida e heterogênea, ultrapassando 60% de adsorção nos primeiros 20 minutos e estabilizando em aproximadamente 80% após 12 horas [2].

A evolução da organização estrutural é acompanhada pelo aumento da dimensão fractal da polyLM. Após 1 hora de adsorção, a estrutura apresenta dimensão aproximada de 1,55; após 8 horas, esse valor aumenta para 1,62; e, após 12 horas, alcança aproximadamente 1,70 [2]. Esses resultados indicam que o crescimento da rede ocorre de forma progressiva, com aumento da complexidade espacial ao longo do tempo.

Os autores também descrevem que a formação inicial da polyLM ocorre a partir de estruturas poligonais planas, frequentemente hexagonais, com dimensões próximas de 30 nm. Essas unidades básicas propagam-se em escalas maiores preservando a conectividade estrutural, comportamento compatível com mecanismos de auto-organização descritos pelos modelos de Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA) [9].

A Fig. 2 apresenta o procedimento experimental empregado para estimar a dimensão fractal da polyLM por meio do método de *box-counting*. O processamento digital das imagens permitiu acompanhar a evolução da organização da rede proteica e quantificar o aumento da complexidade estrutural durante o processo de adsorção.

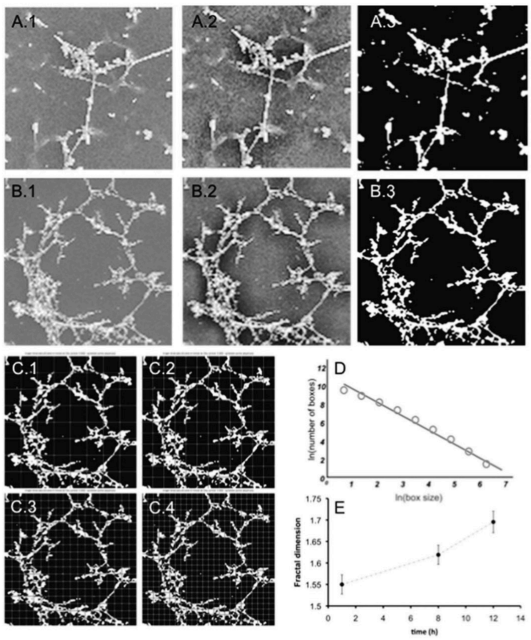


Figura 2. Cálculo da dimensão fractal (estimativa de Hausdorff). (A, B) O processamento das imagens foi realizado para preparar as imagens para o algoritmo de *box-counting* aplicado à LM (A) e à polyLM (B). 1, imagem original; 2, imagem original com histograma equalizado; 3, imagem binarizada utilizando o método de Otsu. (C) A partir das imagens A.3 e B.3, as estimativas da dimensão de Hausdorff podem ser calculadas por meio da sobreposição de uma grade de tamanho variável. (D) Repetindo o processo para diferentes tamanhos de grade e calculando o número de células que contêm qualquer parte da estrutura analisada, obtém-se a dimensão fractal. (E) Evolução temporal da dimensão fractal da polyLM ao longo do processo de adsorção. Adaptado de Hochman-Mendez, M. et al. [2], sob licença Creative Commons CC BY.

C. Proposta de Integração Conceitual

A análise conjunta dos trabalhos revisados permite propor um modelo conceitual que relaciona a geometria fractal da polyLM às suas propriedades estruturais e funcionais. Os resultados indicam que o aumento da dimensão fractal está associado à formação de redes progressivamente mais complexas e conectadas, refletindo um processo de auto-organização da matriz extracelular.

Nesse contexto, a dimensão fractal pode ser interpretada como um descritor quantitativo da organização espacial da polyLM, permitindo correlacionar propriedades geométricas com características estruturais observadas experimentalmente. A concordância entre os valores obtidos por *box-counting*, os modelos fractais clássicos e a aproximação com o modelo de Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA) reforça a consistência dessa abordagem para a caracterização da matriz extracelular.

A integração entre geometria fractal, crescimento não linear e organização estrutural fornece uma perspectiva capaz de conectar modelos matemáticos às evidências experimentais apresentadas nos estudos revisados. Sob essa perspectiva, a análise fractal constitui uma ferramenta quantitativa para investigar a arquitetura da matriz extracelular e sua evolução durante o processo de polimerização.

Embora os estudos revisados indiquem uma associação entre o aumento da complexidade geométrica e propriedades funcionais da matriz extracelular, as evidências disponíveis ainda não permitem estabelecer uma relação causal direta entre dimensão fractal e eficiência da sinalização celular. Assim, o modelo conceitual proposto deve ser interpretado como uma hipótese integradora, cuja validação depende de investigações experimentais e computacionais futuras.

D. Implicações Funcionais da Geometria Fractal da polyLM

Do ponto de vista funcional, a organização fractal da polyLM influencia diretamente os processos de sinalização celular e regeneração neural.

Os estudos analisados indicam que a geometria em rede favorece a exposição do braço longo da LM, região onde se localizam domínios responsáveis pela interação com integrinas e receptores celulares, como a distroglicana [9]. Essa disposição espacial facilita processos de adesão celular, diferenciação neuronal e crescimento de neuritos.

Segundo Menezes et al. [1], estruturas semelhantes à polyLM participam da organização de nichos neurogênicos no sistema nervoso central, contribuindo para a distribuição de fatores de crescimento e para a regulação do microambiente celular.

Além disso, os resultados apresentados por Siqueira-Santos et al. [3] demonstram que filamentos de policaprolactona funcionalizados com LM favorecem a regeneração do nervo ciático, influenciando diretamente parâmetros de densidade axonal, reinervação muscular e recuperação motora *in vivo*.

Os autores observaram que filamentos revestidos com LM comum apresentaram maior densidade de ramificações axonais e maior número de placas motoras marcadas com α-bungarotoxina, enquanto a polyLM promoveu crescimento mais linear dos neuritos. Esses resultados sugerem que diferentes geometrias da matriz extracelular podem ativar vias distintas de sinalização neurobiológica.

Nesse contexto, a geometria fractal da polyLM pode ser interpretada não apenas como uma propriedade estrutural, mas também como um possível elemento funcional associado à sinalização celular e à regeneração neural. Os estudos analisados sugerem que a organização fractal da matriz extracelular influencia a exposição de sítios de ligação celular, a distribuição espacial de sinais biológicos e a interação com receptores de membrana. Sob essa perspectiva, descritores geométricos como dimensão fractal, lacunaridade e conectividade podem contribuir para o desenvolvimento de biomateriais biomiméticos capazes de reproduzir características estruturais observadas em ambientes neurobiológicos naturais.

E. Limitações do Método de Box-Counting e Perspectivas Analíticas

Embora o método de *box-counting* seja amplamente utilizado para estimar a dimensão fractal de estruturas biológicas, sua aplicação apresenta limitações importantes que devem ser consideradas na interpretação dos resultados. Conforme discutido por Bianchi [4] e Silva [11], a estimativa da dimensão

fractal depende fortemente das etapas de pré-processamento da imagem, incluindo equalização de histograma, remoção de ruídos, correções de contraste e procedimentos de binarização.

A escolha do limiar de segmentação constitui uma das principais fontes de variabilidade do método. Mesmo quando técnicas automáticas, como o método de Otsu, são empregadas, pequenas alterações na separação entre objeto e fundo podem modificar significativamente o valor estimado da dimensão fractal. Em sistemas biológicos complexos, nos quais frequentemente existem regiões com baixo contraste ou fronteiras difusas, essa dependência torna-se particularmente relevante [4].

Outra limitação importante está relacionada à resolução espacial das imagens analisadas. Diferentemente dos fractais matemáticos ideais, que apresentam auto-similaridade em infinitas escalas, estruturas biológicas possuem limites físicos de observação. Dessa forma, a validade da dimensão fractal depende da existência de uma faixa adequada de escalas na qual a relação logarítmica utilizada pelo método permaneça aproximadamente linear [4], [7].

Além disso, procedimentos de pré-processamento utilizados em técnicas de microscopia, como suavizações gaussianas, correções geométricas e filtros de realce, podem alterar a distribuição espacial dos pixels e influenciar diretamente os valores obtidos para a dimensão fractal [11]. Consequentemente, a dimensão calculada deve ser interpretada como uma estimativa dependente das condições experimentais e computacionais empregadas.

Diante dessas limitações, abordagens complementares têm sido propostas para fornecer descritores mais abrangentes da complexidade geométrica de sistemas biológicos. Entre elas destacam-se a análise multifractal, a lacunaridade, a análise de textura, a esqueletização multiescala (*skeletonization*), a modelagem por redes e grafos e os métodos de estimação fractal multiescala [4].

A análise multifractal permite caracterizar heterogeneidades locais da estrutura por meio de múltiplos expoentes de escala, fornecendo informações que não podem ser capturadas por uma única dimensão fractal. Já a lacunaridade quantifica a distribuição dos espaços vazios da rede, possibilitando avaliar aspectos relacionados à conectividade e permeabilidade da matriz extracelular.

Por sua vez, técnicas de esqueletização permitem representar a estrutura como uma rede topológica simplificada, preservando sua conectividade e possibilitando a extração de parâmetros como comprimento de ramos, ângulos de bifurcação e organização hierárquica. A partir dessas representações, métodos baseados em grafos podem ser utilizados para investigar propriedades associadas à conectividade estrutural e à propagação de sinais biológicos [4].

Outra possibilidade consiste na utilização de métodos de dimensão fractal multiescala, nos quais a complexidade é analisada em diferentes intervalos de escala. Esses métodos permitem identificar regiões de maior fractalidade e avaliar a estabilidade do comportamento fractal ao longo da estrutura, reduzindo parte da sensibilidade observada nos algoritmos convencionais de *box-counting* [4].

No contexto da polyLM, essas abordagens podem fornecer descritores mais completos da organização da matriz extracelular, permitindo correlacionar propriedades geométricas com processos funcionais como adesão celular, orientação axonal, distribuição de fatores de crescimento e eficiência de sinalização neurobiológica. Dessa forma, a integração entre métricas fractais avançadas e modelos de sistemas dinâmicos constitui uma perspectiva promissora para investigações futuras em neuroengenharia e medicina regenerativa.

F. Lições Aprendidas, Lacunas de Pesquisa e Perspectivas Futuras

A análise da literatura permitiu identificar importantes contribuições para a compreensão da geometria fractal da polyLM e de suas implicações na organização de sistemas neurobiológicos. Os trabalhos revisados demonstram de forma consistente que ela apresenta propriedades de auto-similaridade e dimensões fractais compatíveis com modelos de crescimento auto-organizado, particularmente aqueles associados à Agregação Limitada por Difusão (*Diffusion Limited Aggregation* – DLA). Além disso, observou-se que o aumento progressivo da dimensão fractal ao longo do tempo de adsorção está associado à formação de redes cada vez mais complexas, sugerindo uma relação entre organização geométrica e potencial funcional da matriz extracelular [2], [9].

Outra lição importante refere-se ao papel biológico da sua geometria. Os estudos analisados indicam que a arquitetura fractal favorece a exposição de domínios de interação celular, contribuindo para processos de adesão, diferenciação neuronal, orientação axonal e regeneração tecidual. Dessa forma, a dimensão fractal não deve ser interpretada apenas como um parâmetro geométrico, mas como um possível indicador estrutural associado à eficiência de sinalização celular e à organização funcional de nichos neurogênicos [1], [3], [12].

Apesar dos avanços observados, diversas limitações metodológicas e conceituais permanecem em aberto. Uma das principais refere-se à dependência da análise fractal em relação à qualidade das imagens utilizadas. A resolução espacial limitada dos sistemas de aquisição de imagem restringe a faixa de escalas na qual o comportamento fractal pode ser investigado. Os próprios trabalhos pioneiros sobre a polyLM relatam dificuldades históricas relacionadas à ausência de imagens em magnificações intermediárias e às limitações técnicas de microscopia para observação de estruturas em escalas nanométricas [2].

Outra limitação importante está associada ao método de *box-counting*. Embora amplamente utilizado devido à sua simplicidade computacional, esse método fornece apenas uma estimativa da dimensão de Hausdorff e apresenta elevada sensibilidade às etapas de pré-processamento, segmentação e escolha de limiares. Além disso, a determinação da dimensão fractal depende da seleção da faixa de escalas utilizada na re-

gressão log-log, o que pode introduzir variabilidade estatística nos resultados obtidos [4], [7], [11].

Também foi identificada uma lacuna significativa relacionada à caracterização geométrica da matriz extracelular. A maior parte dos estudos concentra-se na utilização de uma única dimensão fractal como descritor da estrutura. Entretanto, propriedades relevantes da organização espacial, como distribuição de lacunas, heterogeneidade local, conectividade e hierarquia estrutural, permanecem pouco exploradas no contexto da polyLM.

Nesse sentido, abordagens modernas de análise geométrica representam oportunidades promissoras para pesquisas futuras. Técnicas como análise multifractal, lacunaridade, análise de textura, esqueletização multiescala (*skeletonization*), modelagem por grafos e métodos de dimensão fractal multiescala podem fornecer descritores mais completos da matriz extracelular. Essas ferramentas permitem investigar não apenas o grau de complexidade geométrica, mas também aspectos relacionados à conectividade estrutural, distribuição espacial de vazios, organização hierárquica e robustez da rede biológica [4], [11].

No contexto da neurociência, essas abordagens apresentam potencial para estabelecer relações quantitativas entre geometria e função biológica. A integração entre métricas fractais avançadas e modelos de sistemas dinâmicos pode contribuir para compreender como propriedades estruturais da matriz extracelular influenciam processos de sinalização celular, crescimento neural, plasticidade e regeneração do tecido nervoso.

Como perspectiva futura, propõe-se a ampliação desta investigação por meio da implementação computacional de métodos avançados de caracterização fractal e da construção de modelos matemáticos capazes de correlacionar parâmetros geométricos da polyLM com indicadores funcionais de atividade celular. Essa abordagem poderá contribuir para o desenvolvimento de novos biomateriais inspirados em arquiteturas fractais e para o avanço das aplicações da neuroengenharia e da medicina regenerativa.

VI. Considerações Finais

O presente estudo evidenciou que a laminina polimerizada (*polyLM*) apresenta uma organização estrutural complexa e não aleatória, compatível com propriedades fractais, incluindo auto-semelhança e dimensão não inteira. Durante seu processo de formação, a evolução da arquitetura da rede, marcada pelo aumento da complexidade ao longo do tempo de adsorção, apresentou parâmetros compatíveis com o modelo clássico de Agregação Limitada por Difusão (DLA), sugerindo a existência de mecanismos de auto-organização física associados à formação da estrutura supramolecular.

Essa caracterização geométrica apresenta implicações relevantes para a neurobiologia e para a engenharia de tecidos. A arquitetura ramificada e lacunar da *polyLM* é consistente com uma organização capaz de favorecer a disponibilidade de domínios bioativos, influenciando processos como adesão celular, neuritogênese e regeneração tecidual. Além disso, os resultados indicam que diferentes organizações supramoleculares da laminina podem desempenhar papéis distintos durante a regeneração neural, reforçando que a geometria da matriz extracelular atua não apenas como uma característica morfológica, mas como um elemento regulador da interação célula-matriz.

A consolidação dessa relação entre estrutura e função depende, entretanto, da superação das limitações técnicas associadas às metodologias convencionais de análise fractal, como a dependência de contraste, sensibilidade à escala e restrições do método de *box-counting*. Como perspectivas futuras, destaca-se a integração de simulações computacionais e métricas geométricas complementares, incluindo multifractalidade, lacunaridade, esqueletização multiescala e análise de redes, visando estabelecer correlações quantitativas entre os descritores matemáticos da matriz extracelular e o desempenho funcional de novos biomateriais aplicados à neuroengenharia.

Referências

[1] K. Menezes, J. R. L. de Menezes, M. A. Nascimento, R. S. Santos, and T. Coelho-Sampaio, "Polylaminin, a polymeric form of laminin, promotes regeneration after spinal cord injury," *The FASEB Journal*, vol. 24, no. 11, pp. 4513–4522, 2010.

[2] C. Hochman-Mendez, M. Cantini, D. Moratal, M. Salmeron-Sanchez, and T. Coelho-Sampaio, "A fractal nature for polymerized laminin," *PLoS ONE*, vol. 9, no. 10, p. e109388, 2014.

[3] R. Siqueira-Santos, G. Sardella-Silva, M. A. Nascimento, L. T. Oliveira, T. Coelho-Sampaio, and V. T. Ribeiro-Resende, "Biological activity of laminin/polylaminin-coated poly-epsilon-caprolactone filaments on the regeneration and tissue replacement of the rat sciatic nerve," *Materials Today Bio*, vol. 3, p. 100026, 2019.

[4] A. G. C. Bianchi, "Caracterização, modelagem e simulação matemático-computacional da dinâmica do crescimento e conexões de células neurais," Ph.D. dissertation, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, Brazil, 2003.

[5] D. P. Feldman, *Chaos and Fractals: An Elementary Introduction*, 1st ed. Oxford, U.K.: Oxford University Press, 2012.

[6] C. P. M. Barreiras, "O conjunto de cantor," Master's thesis, Universidade de Évora, Évora, Portugal, 2011.

[7] R. F. Oliveira, "Fractais e o crescimento microbiano: Um novo horizonte entre a matemática e a microbiologia," Paraná, Brazil, 2019.

[8] N. I. Maryenko and O. Y. Stepanenko, "Comparative analysis of fractal dimensions of human cerebellum: impact of image preprocessing," *Wiadomości Lekarskie*, vol. 74, no. 5, 2021.

[9] C. Hochman-Mendez *et al.*, "Generating a fractal microstructure of laminin-111 to signal to cells," *Journal of Visualized Experiments (JoVE)*, no. 186, p. e61134, 2022.

[10] J. López, M. Coccolo, R. Capeáns, and M. A. F. Sanjuán, "Controlling the bursting size in the two-dimensional rulkov model," *Communications in Nonlinear Science and Numerical Simulation*, vol. 120, p. 107184, 2023.

[11] G. R. Silva, "Microscopia de força atômica na análise de rugosidade e da dimensão fractal de membranas celulares," Master's thesis, Universidade Federal de Alagoas, Alagoas, Brazil, 2016.

[12] F. C. P. Mesquita *et al.*, "Polymerized laminin-521: A feasible substrate for expanding induced pluripotent stem cells at a low protein concentration," *Cells*, vol. 11, no. 24, p. 3955, 2022.

Characterization of 3D Printer Filaments for CT Anthropomorphic Pediatric Phantom Development

1st Lorena de Camargo Muchak
Postgraduate Program in Biomedical Engineering- Federal University of Technology- UTFPR
Curitiba, Brazil
ORCID 0009-0007-4888-1408

2nd Cristhiane Gonçalves
Postgraduate Program in Biomedical Engineering- Federal University of Technology- UTFPR
Curitiba, Brazil
ORCID 0000-0002-3894-2971

3rd Erik Mathêus Kachel
Postgraduate Program in Biomedical Engineering- Federal University of Technology- UTFPR
Curitiba, Brazil
ORCID 0009-0005-6082-8939

4th Jessica Villa Real
Imaging Diagnostic Unit- Hospital de Clínicas Complex of UFPR
Curitiba, Brazil
ORCID 0000-0002-4934-8361

5th Anna Luiza Metedieri Malthez
Postgraduate Program in Biomedical Engineering- Federal University of Technology- UTFPR
Curitiba, Brazil
ORCID 0000-0003-3109-1155

***Abstract*—The development of pediatric anthropomorphic phantoms faces limitations regarding high costs and the scarcity of customized devices on the market. This study characterizes three 3D printing filaments (Hyper PLA, PLA Stone, and TPU Silk) as affordable alternatives for simulating biological tissues for computed tomography (CT). Cubic samples with infill densities ranging up to 100% were printed for each material and scanned across multiple tube voltages (90, 120, and 140 kV). Linear regression showed a relationship between infill density and Hounsfield Units (HU), yielding(R^2) greater than 0.93. The PLA Stone filament demonstrated higher radiological density, proving suitable for simulating bone structures, while Hyper PLA and TPU Silk showed equivalence to soft tissues. The stability of HU values against voltage variation suggests that the radiological attenuation of these materials is predominantly dependent on physical density. The results validate the efficacy of these materials in the manufacture of customized phantoms, offering a solution for quality control and radiological protection in pediatrics.**

***Index Terms*—3D printing, computed tomography, 3D filaments, pediatric phantoms**

I. Introduction

Recent advancements in software design and 3D printing techniques have propelled the manufacturing of anthropomorphic phantoms, expanding their utility across the medical imaging field. Compared to high-cost and rigid commercial alternatives, 3D-printed phantoms manufactured via fused deposition modeling (FDM) stand out for their affordability, accessibility, and rapid prototyping capabilities. This technology enables the fabrication of customized medical simulators with complex geometries at significantly reduced costs. Within this approach, printed thermoplastic materials are combined to achieve tissue-mimicking structures with equivalent radiological properties [1]–[3].

Phantoms are used in computed tomography (CT) to evaluate dose and image quality, PMMA phantoms analyze dosimetric parameters such as CTDIvol and DLP, while quality assurance phantoms verify uniformity, noise, and CT number accuracy [4]. However, these devices present limitations regarding anatomical representation and adequate differentiation between tissues, particularly in simulating bone structures at different ages [3].

The evolution of diagnostic imaging technologies, particularly in CT, demands continuous advancements in quality assurance and radiological protection. In the pediatric context, this need is amplified by growth and changes in bone density. From a dosimetric perspective, these tissue-equivalent structures enable the precise and safe simulation of radiation dose deposition and spatial distribution within the human anatomy, allowing for realistic analyses without exposing patients to ionizing radiation [5]–[7]. This customization is particularly vital for pediatric applications. Standard adult phantoms fail to accurately represent the unique density, size, and elemental composition of children's tissues as recommended by the ICRU Report 44 [8], [9]. Practical studies reinforce the technical feasibility of constructing complex and detailed physical phantoms using CT images from real patients, thereby overcoming cost limitations and the simplified geometries of commercial models. Furthermore, The age ranges found in the literature cover phantoms for ages 1, 5, 10, and 15 years to represent the diversity of the pediatric population [9], [10].

Given these limitations, it is necessary to develop alternative materials capable of accurately simulating the Hounsfield

Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná financial support/fellowship (Protocols No. SUS2025831000030)

Unit (HU) range of biological tissues, thereby enabling the fabrication of more realistic 3D-printed phantoms.

Polylactic acid (PLA) is widely utilized in medical simulators due to its printing versatility, cost-effectiveness, low toxicity, and biodegradable nature [11]. Furthermore, PLA exhibits attenuation properties similar to those of soft tissues. Previous studies indicate that employing a rectilinear infill pattern with a density between 80% and 100% yields high radiological equivalence to adult and pediatric thoracic and abdominal tissues. Among the variations of this polymer, Hyper PLA stands out as a variant that maintains the chemical composition of classic PLA but is optimized to withstand high extrusion flow rates while exhibiting higher tensile strength than acrylonitrile butadiene styrene (ABS). This characteristic allows the material to reach printing speeds of up to 600 mm/s without compromising the structural integrity or dimensional accuracy of the parts, making it ideal for producing large-scale phantoms that require high mechanical strength [10].

The technique of incorporating additives into PLA is widely recognized in the literature as an effective method for developing composite filaments. A notable example is StoneFil, manufactured by FormFutura, which is used in 3D-printed phantoms to accurately simulate the radiological properties of human bone structures [1], [11]. In the Brazilian context, the addition of mineral filler powder from other brands is commercially available, featuring a composition starting from 35% stone powder.

Thermoplastic polyurethane (TPU) stands out in 3D printing for providing physical properties such as high elasticity, strength, and durability. However, its malleable nature demands greater technical complexity in equipment setup, requiring meticulous adjustments to printing parameters to ensure the final quality of the part [11], [12].

Although previous studies have characterized conventional polymers for use in anthropomorphic phantoms, the specific and comparative characterization of high-performance filaments such as Hyper PLA, PLA Stone, and TPU Silk for simulating pediatric tissues is currently lacking [12]. In this research, the experimental and quantitative characterization of 3D printing materials was performed to evaluate their radiological properties under clinical CT scanning conditions, aiming at the development of realistic human mimicking tissues for quality assurance anthropomorphic phantoms in diagnostic imaging examinations and the radiological protection of pediatric patients.

II. Materials and Methods

A. Materials

To characterize which biological tissues can be simulated through 3D printing, three materials were tested: Hyper PLA (Creality), PLA Stone (Voolt3D), a composite filled with stone powder, and thermoplastic polyurethane (TPU Silk). These filaments were selected to encompass a broad range of radiological attenuation, spanning from soft tissues to high-density bone structures. PLA and TPU are commercially prevalent materials across multiple brands. However, PLA Stone requires targeted selection to verify the presence of stone powder in the filament blend.

B. 3D printing

The samples were produced with infill densities between 10% and 100% using a Creality K2 PLUS printer, configured with a 0.4 mm nozzle, 0.35 mm line width, and 0.08 mm layer height (Fig. 1). The objective of this controlled variation was to enable the rigorous adjustment of Hounsfield Units (HU) for mimicking various biological tissues.

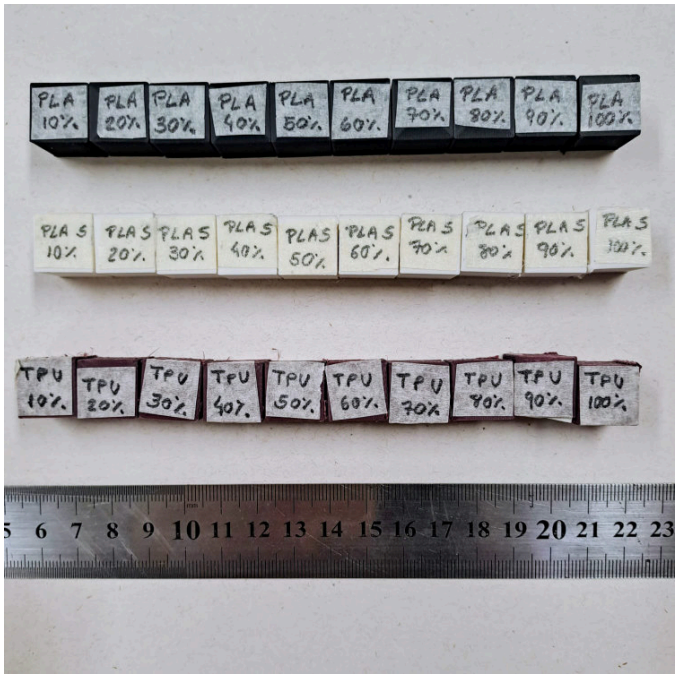


Fig. 1. Fabricated samples with Hyper PLA, PLA Stone, and TPU Silk.

The choice of the rectilinear pattern is based on published studies [13], [14], which claim that this pattern ensures better image quality and greater fidelity to real tissues compared to other infill patterns. For manufacturing solid structures, this rectilinear infill provides a more homogeneous laydown between layers, preventing the formation of interlayer voids.

The physical density of each sample was determined from the average of two weighings and the measurement of the faces of the cubes using a digital caliper, aiming for future comparison with ICRU 44 data at the density level of each material to develop tissue-equivalent density properties.

The printing parameters, including temperature and speed, were set according to the recommendations of each filament manufacturer, using Creality parameters for Hyper PLA and manufacturer specifications for TPU, for PLA Stone, nozzle temperature was adjusted based on prior temperature tower calibrations, and outer wall speed was set to 60 mm/s before infill deposition, as listed in Table I. To accommodate the stone powder composite in the PLA matrix, which demands lower processing speeds than standard PLA, the printing speed was manually reduced by 50% on the printer itself to ensure optimal infill.

TABLE I
SUMMARY OF PRINTING PARAMETERS FOR TESTED FILAMENTS

Parameter	Hyper PLA	PLA Stone	TPU Silk
Nozzle Temp. (°C)	220	215	230
Bed Temp. (°C)	50	65	60
First Layer Speed (mm/s)	60	50	60
Inner Wall Speed (mm/s)	120	70	100

C. Radiological Characterization

The cubes were subjected to computed tomography scans Philips MX 16-Slice for radiological characterization, using tube voltage variations of 90 kV, 120 kV, and 140 kV, combined with tube currents of 18 mAs, 60 mAs, 100 mAs, and 150 mAs, covering the typical clinical range for pediatric and adult CT protocols found in the literature. The selection of 120 kV and 140 kV is widely used in the literature for phantom testing and characterizing simulator materials, while 90 kV is suitable for pediatric protocols, where values between 70 kV and 90 kV are recommended. [15], [16].

Images were acquired with a slice thickness of 0.75 mm, applying mediastinum and bone filters for the subsequent analysis of the CT number (Hounsfield Units). Hounsfield Units (HU) value is calculated using [17]:

$$HU = 1000 \times \frac{\mu_{\text{material}} - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}} \tag{1}$$

where μ represents the linear attenuation coefficient of the tissue, μ_{water} is the coefficient of water, and μ_{air} is the coefficient of air. This scale allows a direct correlation between pixel values in the image and the physical density of the simulated tissues. Following the acquisition of the DICOM tomographic images, processing and volumetric analysis were performed using the open-source software 3D Slicer. For each printed sample, a centralized volumetric region of interest (ROI) of approximately 0.180 cm³ with a voxel count of 1009 was established using the Sphere Brush tool, avoiding the edges and sides in the axial, sagittal, and coronal planes to encompass a homogeneous volume of the cubes (Fig 2).

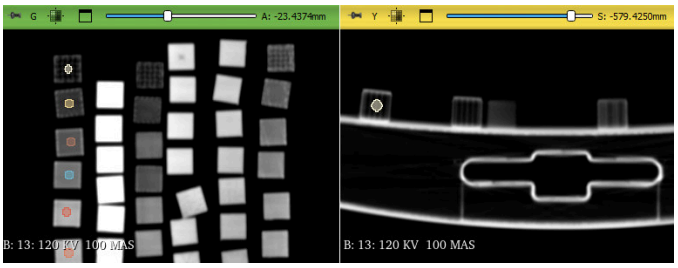


Fig. 2. 3D Slicer interface showing volumetric sample ROIs defined via the Sphere Brush tool in the coronal (left) and axial (right) views.

When defining a region of interest using tools such as the sphere, brush, or ROI, 3D Slicer calculates the simple arithmetic mean of the HU values across all voxels within that segment via modules like Segment Statistics, rather than performing voxel-by-voxel smoothing. For visualization and resampling along oblique planes where sampling coordinates do not align with a voxel's geometric center, Slicer applies default trilinear interpolation, computing a weighted average of neighboring voxel intensities to represent that specific point. Furthermore, individual voxel HU values are only replaced by the average of their immediate neighborhood if a spatial smoothing filter, such as the Mean Filter from the SimpleITK module, is explicitly applied manually.

Each acquisition was performed in triplicate, reporting the maximum, minimum, and mean HU values provided by the 3D Slicer software. This procedure aimed to extract the mean HU values and their respective standard deviations, ensuring that the edges of the cubes did not interfere with the measurement due to potential beam hardening effects or interface artifacts. The obtained data were organized and plotted in scatter plots to evaluate the dependence of HU values on the infill density percentage for each of the three tested materials: Hyper PLA, PLA Stone, and TPU Silk. The characterization of the radiological response was consolidated through a regression analysis, allowing the establishment of the mathematical relationship between the internal printing density and the observed attenuation, according to the linear fit equation:

$$HU = a \times \text{Infill} + b \tag{2}$$

In this expression, the angular coefficient 'a' represents the sensitivity of HU variation per infill (HU/%) increment, while the linear coefficient 'b' correlates to the attenuation of the base material with minimum porosity. In addition to the equation, fit quality indicators, such as the Coefficient of Determination (R^2), will be reported to demonstrate the correlation between the infill density (%). and the observed attenuation.

III. Results and Discussion

When comparing the physical densities obtained in this study with the reference values reported in ICRU Report 44, the Hyper PLA and TPU Silk filaments, with a density of 1.17 g/cm^3 at 100% infill, presented values close to soft and structural tissues of lower attenuation as shown in Table II. These results indicate that Hyper PLA and TPU Silk are more suitable for mimicking soft tissues, lung, and adipose tissue at infill densities between 80% and 100%.

Based on ICRU Report 44, the mean mass densities of the selected tissues are 916 kg/m^3 for adipose tissue, 1053 kg/m^3 for liver, 1050 kg/m^3 for deflated lung, 260 kg/m^3 for inflated lung, and 1100 kg/m^3 for skin. However, biological changes resulting from growth and maturation cause variations in tissue density, which prevents the direct application of adult model values to pediatric models. Therefore, the fabrication of pediatric phantoms requires specific materials calibrated for each age group and for the physical properties of pediatric tissues.

For the development of a dosimetric phantom, in addition to the physical density required to simulate tissue-equivalent

TABLE II
PHYSICAL PROPERTIES OF THE TESTED MATERIALS AT DIFFERENT INFILL PERCENTAGES.

Material	Infill (%)	Mass (g)	Volume (cm^3)	Density (g/cm^3)
PLA Stone	100%	4.39	3.398	1.29
	90%	3.94	3.398	1.16
	80%	3.53	3.443	1.03
	70%	3.24	3.420	0.95
Hyper PLA	100%	3.98	3.398	1.17
	90%	3.64	3.398	1.07
	80%	3.25	3.398	0.96
	70%	2.99	3.420	0.87
TPU Silk	100%	4.00	3.409	1.17
	90%	3.36	3.420	0.98
	80%	3.22	3.420	0.94
	70%	3.11	3.420	0.91

materials, it is essential to analyze the CT number for each utilized filament. Following the CT scan, the DICOM images enabled the extraction of HU values for each combination of voltage and current across the different tested filaments.

Based on the generated scatter plots for the radiological attenuation analysis, the parameters of the linear regression equations for the tested materials were extracted. The statistical fit, validated by the coefficient of determination (R^2), yielded values greater than 0.93 across all conditions, as detailed in Table III. When analyzing the influence of the tube voltage (kV) on the radiological response of the materials, a negligible variation in the HU values was observed for all tested samples (90, 120, and 140 kV). The coefficients of the linear fit equations showed notable stability across the different energy conditions, displaying minimal variations in both the slope (a) and y-intercept (b) parameters.

TABLE III
LINEAR FIT PARAMETERS FOR 18 MAS, ORGANIZED BY VOLTAGE AND MATERIAL

Voltage	Material	a	b	R^2
90 kV	PLA Stone	14.12	-981.01	0.984
	Hyper PLA	11.14	-986.44	0.997
	TPU Silk	9.46	-886.30	0.962
120 kV	PLA Stone	13.57	-992.07	0.995
	Hyper PLA	11.00	-982.41	0.997
	TPU Silk	9.42	-878.08	0.969
140 kV	PLA Stone	14.05	-973.08	0.983
	Hyper PLA	11.04	-978.93	0.996
	TPU Silk	9.49	-880.89	0.970

This behavior reveals that the radiological attenuation of the analyzed filaments, within the applied energy range, is essentially due to physical density and internal architecture, as already presented in literature [6], [17], [18], showing a negligible dependence on the X-ray beam energy. Since the attenuation of Hyper PLA, PLA Stone, and TPU Silk proved to be independent of the tube voltage within the tested range. We can inferred that the same 3D-printed object can be utilized in the CT scanner tests across multiple clinical scenarios. For the diagnostic energy range (90 to 140 kV) and for low atomic number (Z) materials, the dominant physical phenomenon in X-ray attenuation is Compton Scattering (μ_{Compton}). This process depends almost exclusively on the electron density (ρ_e), and consequently on the physical density (ρ_{mass}) of the material, as described [18]:

$$\mu \approx \mu_{\text{Compton}} \propto \rho_e \propto \rho_{\text{mass}} \tag{3}$$

Conversely, the Photoelectric Effect (μ_{photo}), which is highly dependent on both the atomic number and the beam energy (E), remains negligible in this regime due to the absence (or low concentration) of high-Z elements in the polymer matrices [18]:

$$\mu_{\text{photo}} \propto \frac{Z^3}{E^3} \approx 0 \tag{4}$$

Therefore, the independence of the linear attenuation coefficient relative to the beam energy validates these 3D-printed filaments as tissue-equivalent materials for multi-protocol phantom development. This stability represents a significant advantage for phantom construction, as the HU defined for a specific CT protocol can be faithfully preserved under various operating voltages.

In Figures 3 and 4, the results showed at 100% infill and 120 kV a considerable disparity between conventional Hyper PLA (99 to 113 HU) and PLA with mineral additive (372 to 379 HU). Due to its higher HU values, PLA Stone proves to be suitable for simulating denser biological tissues, specifically bone structures, with the exception of cortical bone. Regarding the simulation of cortical bone, the results obtained for PLA Stone, which exhibited values of 372 to379 HU at 100% infill, are significantly lower than the range typical of human cortical bone, which generally presents between 1000 and 1500 HU. This discrepancy indicates that current commercial PLA Stone filaments have limitations in fully replicating the radiodensity of dense cortical bone. To address this, future strategies may involve the development of composite filaments with higher concentrations of radiopaque mineral additives, such as hydroxyapatite or calcium carbonate, or the use of multi-material 3D printing techniques to increase the effective structural density.

To validate the radiodensity properties of the evaluated filaments, the CT number (HU) values obtained in this work were compared with data reported in previous studies for similar 3D printing materials.

Table IV summarizes these attenuation values alongside findings from Savi et al. [14]. Discrepancies observed across the literature highlights the influence of distinct manufacturing formulations, additive concentrations, and scanner reconstruction parameters on the final CT numbers.

Savi presents a comparison with Okkalidis, Shin, and Sole et al. for PLA and copper mixtures, as well as other filaments and variations such as ABS, ABS+, PETG, TPE, HIPS, PVA, and TPU. The authors present the CT number results for 120 kV and 100% infill using a rectilinear pattern. PLA+Cu is widely

TABLE IV
COMPARISON OF CT NUMBERS (HU) ACROSS LITERATURE

Study / Reference	Hyper PLA	PLA+ additives	TPU
This Study	104 ± 9	379 ± 5 (PLA Stone)	90 ± 9
Savi et al.	30 ± 13 (PLA)	330 ± 19 (PLA+Cu)	-37 ± 5 (TPU)
Okkalidis et al.	136	130 (PLA+Cu)	136 ± 27 (TPU)
Solc et al.	-59 ± 19	–	–
Shin et al.	168 ± 16	–	-53 ± 46

used and prominent in the literature, being one of the most commonly used PLA mixtures for testing filaments intended for phantoms. Copper additives typically account for around 25% to 30% of the composition found in these filaments [13], [14], [19], [20].

It is important to highlight that variations in mineral concentration among different manufacturers can influence the obtained HU values, a factor that must be considered according to the printing objective and the phantom's age. In contrast, the results revealed that TPU Silk maintained a equivalence with Hyper PLA, showing similarities in both physical density and radiological behavior. It is worth noting that, although both exhibit close performance, Hyper PLA demonstrated higher statistical precision, with R^2 values closer to unity.Additionally, as illustrated in Figures 3 and 4, TPU Silk exhibits stability in HU values between the infill densities of 50% and 60%, a pattern repeated in the 70% and 80% intervals, suggesting lower sensitivity to infill variation within these specific ranges. TPU Silk is a highly flexible material, variations in its radiological properties may be attributed to printing challenges, such as the requirement for low printing speeds and the material's tendency to absorb moisture if not stored properly, which can affect the consistency of the deposited layers.

When using the software for the mediastinum and bone filters, a contrast adjustment (window/level threshold) is necessary. Comparing the HU values and the standard deviation, we observe that the values for the bone filter applied to the cubes show a greater variation in CT numbers, in addition to having a significantly higher standard deviation. For values of 100 mAs and 90 kV with 100% infill in Hyper PLA, the mediastinum filter yields an average value of 100.3 HU ± 5.3, whereas for the same element using the bone filter, this value changes to 89.3 HU ± 28.6. For 10% infill values, this deviation reaches up to 150. For this study, the values acquired from the mediastinum filter were used.

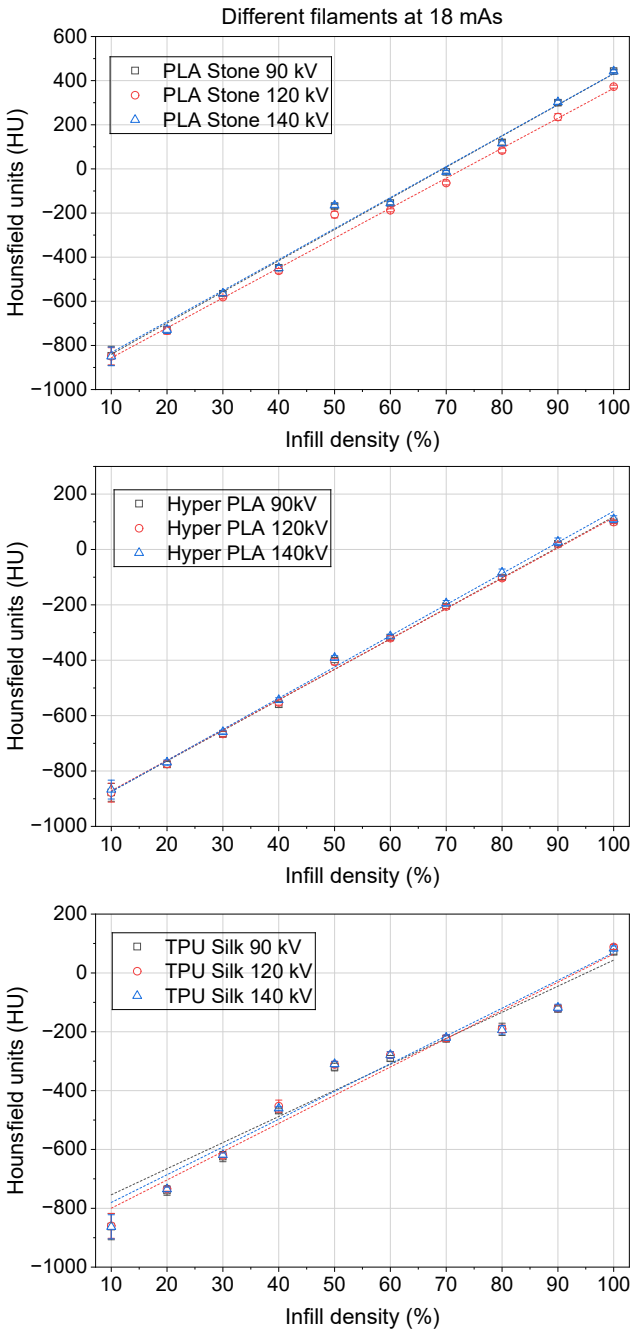


Fig. 3. CT Number in different filaments at each different infill at 18 mAs

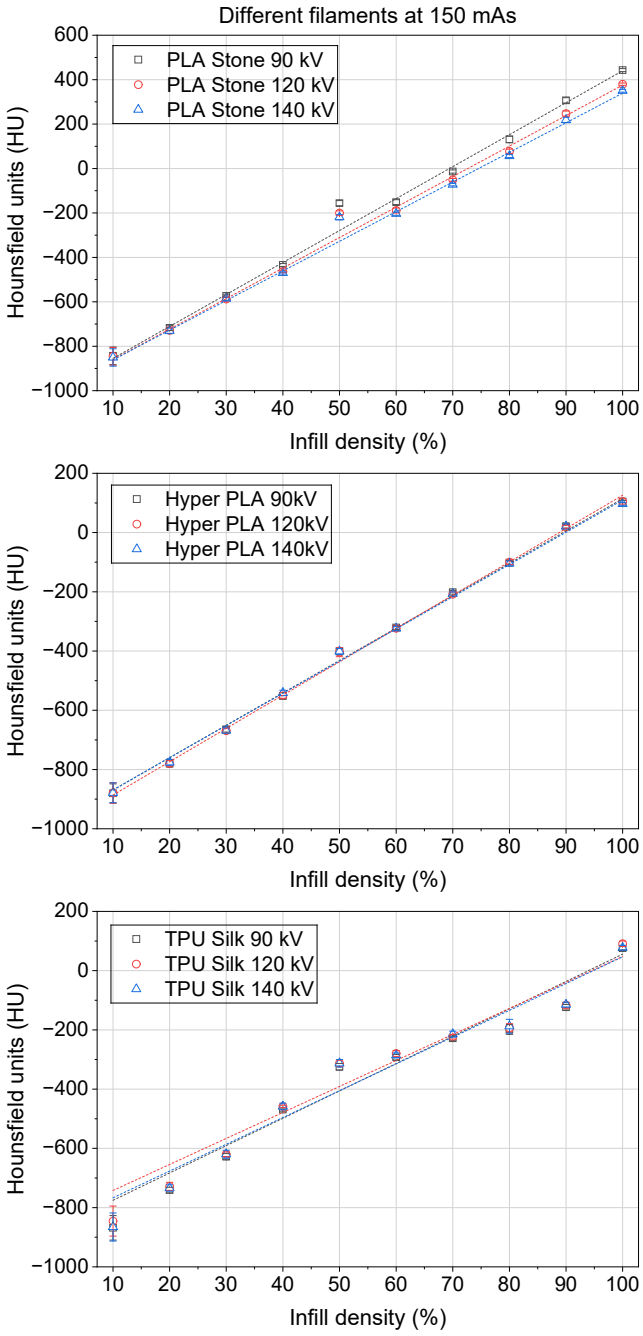


Fig. 4. CT Number in different filaments at each different infill at 150 mAs

Conclusion

When discussing quality assurance,3D-printed phantoms serve as standardized devices for calibrating and verifying parameters in CT systems, including geometric accuracy and HU stability. Therefore, using low-cost and customizable phantoms may facilitate their application in CT quality control. Additionally, it is worth emphasizing the importance of evaluating Hyper PLA and TPU Silk, which, despite presenting similar radiological attenuation, exhibit significant differences in workability and print quality.

The characterization of Hyper PLA, PLA Stone, and TPU Silk filaments validates the use of these materials as cost-effective alternatives for producing pediatric phantoms. The observed relationship between infill density and HU indicates that the radiological response of the printed samples can be adjusted by changing their physical density, highlighting density control through the correlation of attenuation coefficients and photon scattering in matrices that mimic biological tissues. The relatively stable HU values observed at 90, 120, and 140 kV also indicate that, within the conditions evaluated in this study, changes in tube voltage had a limited effect on the measured CT numbers.

Regarding the limitations of this study, we highlight the use of a limited number of filament brands, validation conducted on a single CT scanner model, and sampling restrictions, which may influence the generalizability of the results to other imaging systems or materials.

For future work, we recommend expanding this study to include mechanical and durability tests of the printed materials, as well as printing a phantom for validation with actual dosimetry. Furthermore, PLA Stone formulations and other mineral-filled filaments should be investigated to determine how composition and mineral concentration affect CT attenuation. These additional evaluations will help determine the suitability of these materials for specific anatomical structures and support their application in the development and validation of pediatric CT phantoms.

Acknowledgment

The authors would like to thank the Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná for the financial support/fellowship (Protocols No. SUS2025831000030 and No.SUS2025831000042), and the financial support from the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grants 420798/2023-4 and 312572/2025-5).

References

[1] K. Mei, S. Jiang, F. Kopp, M. Pinto, B. J. Schwaiger, A. S. Gersing, J. S. Kirschke, D. Pfeiffer, E. J. Rummeny, J. Herzen, F. Pfeiffer, and P. B. Noel, "Design and fabrication of 3D-printed patient-specific soft tissue and bone phantoms for CT imaging," *Scientific Reports*, vol. 13, no. 1, p. 17495, Oct 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-44602-9

[2] S. Rodgers, J. Atkinson, D. Cryer, C. Storm, R. Nezich, M. A. Ebert, and P. Rowshanfarzad, "Construction and validation of an infant chest phantom for paediatric computed tomography," *Physical and Engineering Sciences in Medicine*, vol. 47, no. 1, pp. 263–274, 2024.

[3] T. Fujibuchi, "Investigation of a method for creating neonatal chest phantom using 3d printer," *Journal of Physics: Conference Series*, vol. 1943, no. 1, p. 012056, jul 2021. [Online]. Available: https://doi.org/10.1088/1742-6596/1943/1/012056

[4] D. T. T. Mekonin, Tadelech Sisay, "Image quality and radiation dose assessment for the clinically applied 16-slice ct scanner using pmma phantom and quality assurance phantom," *Egyptian Journal of Radiology and Nuclear Medicine*, 2023.

[5] E. L. M. Guerra, "Construção de um fantoma físico do tórax para avaliações dosimétricas das radiações em medicina," *Anais do 20º Congresso Nacional de Iniciação Científica (CONIC SEMESP)*, 2020.

[6] H. A. Mahmoud and M. H. Ali, "Experimental determination of gamma ray attenuation and scattering coefficient in tissue-equivalent materials," *Journal of Physics: Conference Series*, vol. 3028, p. 012045, 2025.

[7] E. Schültke, "Biomimetic phantoms in X-ray-based radiotherapy research: A narrative review," *Biomimetics*, vol. 10, no. 1, p. 794, 2025.

[8] International Commission on Radiation Units and Measurements, "Tissue substitutes, phantoms and computational modelling in medical ultrasound," *ICRU Report 44*, 1989.

[9] D. E. Carver, S. D. Kost, M. Scripter, J. C. Yanch, M. G. Stavin, D. R. Pickens, and R. R. Price, "Realistic phantoms to characterize dosimetry in pediatric CT," *Pediatric Radiology*, vol. 47, no. 6, pp. 691–700, jun 2017.

[10] J. R. Stone, P. Rowshanfarzad, A. Polpo, C. Williams, and R. Nezich, "Development and validation of a 3d-printed dosimetry phantom for paediatric radiology," *Physical and Engineering Sciences in Medicine*, 2026.

[11] D. A. Santos, "Análise da estrutura e propriedades de compósitos de PLA com pós de madeira para impressão 3D por extrusão," *Trabalho de Conclusão de Curso, Dept. Engenharia Mecânica, UFSC*, 2025.

[12] D. Rahmatabadi, I. Ghasemi, M. Baniassadi, K. Abrinia, and M. Baghani, "3D printing of PLA-TPU with different component ratios: Fracture toughness, mechanical properties, and morphology," *Journal of Materials Research and Technology*, vol. 21, pp. 3970–3981, 2022.

[13] N. Okkalidis, C. Chatzigeorgiou, and D. Okkalides, "Assessment of 11 available materials with custom three-dimensional-printing patterns for the simulation of muscle, fat, and lung Hounsfield units in patient-specific phantoms," *Journal of Engineering and Science in Medical Diagnostics and Therapy*, vol. 1, no. 1, p. 011003, feb 2018.

[14] M. Savi, M. A. B. Andrade, and M. P. A. Potiens, "Commercial filament testing for use in 3D printed phantoms," *Radiation Physics and Chemistry*, vol. 174, p. 108906, apr 2020.

[15] R. A. J. Nievelstein, I. M. van Dam, and A. J. van der Molen, "Multidetector CT in children: current concepts and dose reduction strategies," *Pediatric Radiology*, vol. 40, no. 8, pp. 1324–1344, aug 2010.

[16] G. Salvi de Souza, M. Lanfredi, and A. Marques Da Silva, "Parâmetros de aquisição em tomografia computadorizada para pacientes pediátricos: uma revisão bibliográfica," *Revista Brasileira de Física Médica*, vol. 12, p. 30, 12 2018.

[17] E. Okuno and E. M. Yoshimura, *Física das Radiações*. São Paulo, Brasil: Oficina de Textos, 2010.

[18] D. R. Dance, S. Christofides, A. D. A. Maidment, I. D. McLean, and K. H. Ng, Eds., *Diagnostic Radiology Physics: A Handbook for Teachers and Students*. Vienna, Austria: International Atomic Energy Agency (IAEA), 2014.

[19] J. Solc, T. Vrba, and L. Burianova, "Tissue-equivalence of 3d-printed plastics for medical phantoms in radiology," *Journal of Instrumentation*, vol. 13, no. 09, p. P09018, sep 2018. [Online]. Available: https://doi.org/10.1088/1748-0221/13/09/P09018

[20] S. R. . S. G. Shin, J., "Imaging properties of 3d printed materials: Multienergy ct of filament polymers," *Journal of Digital Imaging*, 2017.

[21] P. Akhlaghi, H. M. Hakimabad, and L. R. Motavalli, "Determination of tissue equivalent materials of a physical 8-year-old phantom for use in computed tomography," *Radiation Physics and Chemistry*, vol. 112, pp. 169–176, 2015.

Deep Feature-Based Classification of Colorectal Histopathology: Effects of Class Balancing and Particle Swarm Feature Selection

Vitória Rocha de Oliveira Silva
Universidade Federal de Pernambuco
Recife, Brazil
vros@cin.ufpe.br

Eduardo Kesller Almeida Santiago
Universidade Federal de Pernambuco
Recife, Brazil
eduardo.kesller@ufpe.br

Sandrelly Mirely da S. Santos
Universidade Federal de Pernambuco
Recife, Brazil
sandrelly.santos@ufpe.br

Clarisse Lins de Lima
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
clarisse.lins@ufpe.br

Giselle M. M. Moreno
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
giselle.moreno@ufpe.br

Wellington P. dos Santos
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
wellington.santos@ufpe.br

Maíra Araújo de Santana
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
maira.araujosantana@ufpe.br

Juliana C. Gomes Cassemiro
Universidade Federal de Pernambuco
Recife, Brazil
juliana.cgomes@ufpe.br

Abstract*—Colorectal cancer diagnosis based on histopathological assessment remains challenging due to subtle morphological differences among tissue subtypes and interobserver variability. This study evaluates a computer-aided classification pipeline for hematoxylin-and-eosin (H&E)-stained biopsy patches from the EBHI-SEG dataset. Deep features were extracted using an ImageNet-pretrained EfficientNet-B0 network to generate a complete feature representation (Base 1), while Particle Swarm Optimization (PSO) was employed to produce a reduced feature subset (Base 2). Nine conventional classifiers implemented in Weka were evaluated using 30 repetitions of stratified 10-fold cross-validation under two experimental settings, with and without the Synthetic Minority Over-sampling Technique (SMOTE). The dataset comprised 2,228 highly imbalanced image patches distributed across six histopathological classes, with an approximately 13.7:1 ratio between the largest and smallest categories. The second-degree polynomial SVM achieved the highest composite score during cross-validation, reaching 0.8762 without SMOTE and 0.9737 with SMOTE. Although SMOTE substantially improved the training performance of most classifiers, these gains were not fully reflected on the independent test set. The complete feature representation achieved the best held-out performance, with 85.49% accuracy, a Cohen's kappa of 0.801, and a weighted AUC of 0.949. Class-wise analysis further revealed that high-grade intraepithelial neoplasia and serrated adenoma remained considerably more difficult to recognize than adenocarcinoma (recall > 0.93), demonstrating that weighted aggregate metrics may conceal clinically relevant deficiencies in minority classes. Overall, the proposed pipeline provides a reproducible baseline for AI-assisted colorectal histopathology and highlights the importance of class-aware evaluation when developing machine learning models for severely imbalanced multiclass datasets.

***Index Terms*—colorectal cancer, histopathology, EfficientNet, SMOTE, particle swarm optimization, EBHI-SEG**

I. Introduction

Colorectal cancer (CRC) is among the most common malignant diseases worldwide, with approximately 1.9 million new cases and more than 900,000 deaths estimated for 2022 [1]. Histopathological examination of hematoxylin-and-eosin (H&E) stained tissue remains central to confirming malignancy and characterizing precursor lesions. The diagnostic spectrum includes normal mucosa, polyps, serrated lesions, low- and high-grade intraepithelial neoplasia, and invasive adenocarcinoma. These categories may exhibit subtle glandular and cytological differences, while the growing volume of digital pathology creates a strong motivation for reproducible decision-support systems.

The clinical difficulty is not limited to separating normal tissue from invasive carcinoma. Colorectal biopsy interpretation involves a continuum of architectural and cytological alterations, including gland crowding, loss of polarity, nuclear stratification, and stromal response. In this setting, minority patterns are especially important because high aggregate performance can coexist with weak recognition of less frequent categories.

Artificial intelligence (AI) in computational pathology has progressed from small patch benchmarks to weakly supervised analysis of whole-slide images (WSIs). Deep models have classified colorectal polyps with high accuracy [2], learned cancer representations with limited dense annotation [3], and

supported prescreening or interpretable diagnosis across large WSI cohorts [4], [5]. Nevertheless, the literature consistently shows that high internal performance does not imply clinical readiness. Scanner variation, staining differences, dataset artifacts, patient leakage, and prevalence shifts can substantially alter performance [6], [7].

Although these advances clearly illustrate the capabilities of AI, most studies typically operate at the whole-slide image (WSI) level or rely on large cohorts with dense annotations [2], [4], [5]. For academic and clinical laboratories with limited computational infrastructure and access to modestly sized datasets, leveraging feature extraction from pre-trained backbones coupled with lightweight classifiers presents a feasible and inexpensive alternative. However, this strategy has received limited attention in settings characterized by severe class imbalance and morphologically subtle distinctions among categories-such as the EBHI-SEG dataset used as the benchmark for our experiments.

While full fine-tuning of neural networks remains a widely employed approach, it inherently couples the representation learning stage with the downstream decision model, thereby complicating the isolated assessment of post-hoc techniques such as class balancing and feature selection. Conversely, adopting a pre-trained convolutional network as a frozen feature extractor decouples these stages, yielding a stable and reproducible representation on which various conventional classifiers can be trained and systematically compared under controlled conditions [7], [8].

This work therefore addresses a narrower experimental question: how effectively can conventional classifiers distinguish six EBHI-SEG histological categories when trained on transferred EfficientNet features, and how do class balancing and particle-swarm-based feature selection affect performance? Tree research questions (RQs) guide the analysis:

- **RQ1:** Which conventional classifier provides the best overall discrimination using the complete EfficientNet representation?
- **RQ2:** How does SMOTE affect accuracy, agreement, sensitivity, specificity, AUC, and the composite ranking?
- **RQ3:** Does PSO-based feature selection improve predictive performance consistently across classifier families?

The study contributes a reproducible characterization of the local EBHI-SEG subset, a comparison of nine classifiers across complete and PSO-selected features with and without SMOTE, a paired analysis of PSO selection, and a critical assessment of the internal evaluation setting.

II. Background and Related Work

A. Clinical and Computational Context

Digital pathology enables computational processing of glass slides. Because a WSI can contain billions of pixels, typical systems require tissue detection, tiling, feature extraction, regional prediction, and slide-level aggregation [4], [7]. Patch datasets simplify these steps and support controlled experiments, but patch performance represents local morphological discrimination rather than complete case-level diagnosis [4], [5].

Current evidence supports decision assistance rather than autonomous replacement of the pathologist. High-sensitivity systems may prioritize suspicious cases or regions; consequently, sensitivity and negative predictive value may be more relevant than raw accuracy [4], [7].

B. AI-Assisted Colorectal Histopathology

Early deep-learning studies demonstrated that colorectal gland morphology and polyp type can be learned from histological images. Kainz *et al.* combined convolutional networks and regularization for gland segmentation and tissue classification [9]. Korbar *et al.* evaluated five polyp classes on independent WSIs and reported 93.0% accuracy [2]. More recent work has focused on reducing annotation requirements and moving closer to clinical workflows. Yu *et al.* reported semi-supervised CRC recognition approaching fully supervised and pathologist-level performance while using fewer labeled samples [3].

At WSI scale, weak supervision and multiple-instance learning are attractive because slide-level diagnoses are easier to obtain than exhaustive pixel annotations. Campanella *et al.* established a large-scale weakly supervised framework for clinical-grade pathology tasks [7], while attention-based methods such as CLAM improved data-efficient slide aggregation [10]. For CRC specifically, Bilal *et al.* developed an AI prescreening system for large-bowel biopsies from the United Kingdom and Portugal [4]. Neto *et al.* proposed an interpretable mixed-supervision system trained on approximately 10,500 WSIs, reporting 93.44% internal accuracy and high sensitivity, with lower performance on external cohorts [5]. The difference between internal and external results reinforces the importance of domain-aware validation.

Table I summarizes key AI studies in colorectal histopathology, highlighting that most operate at the whole-slide image (WSI) level using weak supervision or dense annotations. In contrast, our study adopts a controlled patch-level design, which allows us to isolate the effects of class balancing and feature selection across nine conventional classifiers.

III. Materials and Methods

A. Proposed Method

Figure 1 illustrates the methodological workflow adopted in this study. The proposed pipeline begins with the acquisition of histopathological images from the EBHI-SEG dataset, followed by deep feature extraction using a pre-trained EfficientNet-B0 network. The extracted feature vectors are then divided into training and independent test sets, after which class balancing using the Synthetic Minority Oversampling Technique (SMOTE) is applied to the training data. Two feature representations are subsequently evaluated: the complete feature set (Base 1) and a reduced feature subset obtained through Particle Swarm Optimization (PSO) (Base 2). Finally, conventional machine learning classifiers are trained and validated using repeated cross-validation, and the best-performing models are evaluated on the independent test set.

TABLE I
REPRESENTATIVE LITERATURE AND ITS RELATIONSHIP TO THE PRESENT STUDY

Study	Data and task	Method and contribution	Relationship to this work
Kainz *et al.* [9]	Colon glands; segmentation and benign/malignant classification	CNNs with total-variation regularization showed that gland morphology is learnable	Supports learned morphology but addresses segmentation rather than six-class classification
Korbar *et al.* [2]	239 independent WSIs; five polyp types	ResNet classification reached 93.0% accuracy and 88.8% F1	Provides a related benchmark with independent slide testing
Yu *et al.* [3]	Labeled and unlabeled CRC images	Semi-supervised learning reduced annotation requirements while retaining high AUC	Motivates data-efficient transfer learning
Bilal *et al.* [4]	UK and Portuguese large-bowel biopsies	Weakly supervised, externally validated prescreening	Illustrates the validation level required for clinical triage
Neto *et al.* [5]	Approximately 10,500 WSIs with external cohorts	Mixed supervision with interpretable tile severity	Represents a translational WSI target beyond this patch study

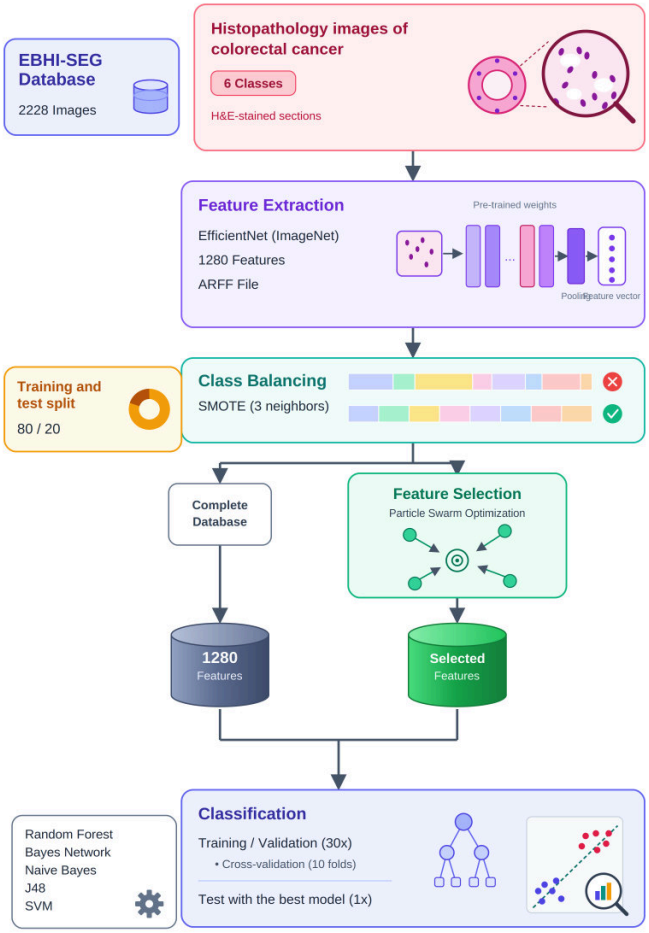


Fig. 1. Workflow of the proposed methodology, including the following stages: dataset, feature extraction, class balancing, feature selection, and classification.

Each stage of the proposed methodology is described in detail in the following sections.

B. Dataset

EBHI-SEG was introduced as a public enteroscope biopsy H&E dataset for segmentation and contains six tumor-differentiation categories with corresponding masks [11]. The complete release is described as 5,170 files comprising images and ground truth. The local copy audited for this study contained 2,228 RGB image patches of 224×224 pixels and 2,227 masks. Two low-grade intraepithelial-neoplasia masks were absent, but masks were not used in classification.

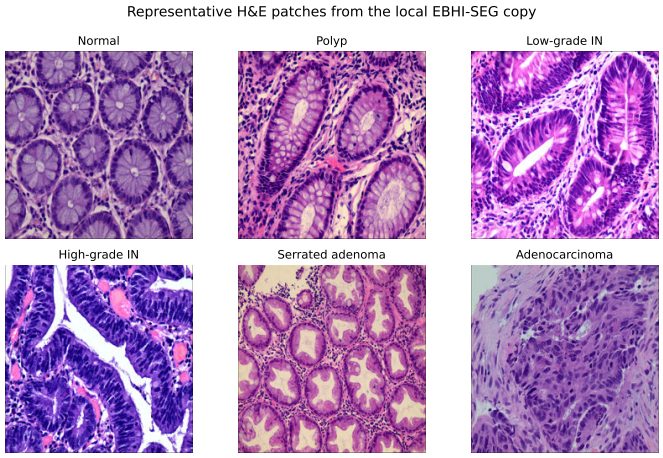


Fig. 2. Representative H&E patches from the six classes in the local EBHI-SEG copy.

TABLE II
CLASS DISTRIBUTION IN THE LOCAL DATASET

Class	Images	Share (%)
Normal	76	3.41
Polyp	474	21.27
Low-grade IN	639	28.68
High-grade IN	186	8.35
Serrated adenoma	58	2.60
Adenocarcinoma	795	35.68
Total	**2,228**	**100.00**

Fig. 2 illustrates the morphological diversity of the classes, and Table II quantifies the imbalance. EBHI-SEG does not provide sociodemographic variables in the distributed directory structure; therefore, age, sex, ethnicity, and other patient-level covariates were not analyzed. According to the dataset

documentation, the images originate from enteroscope biopsy histopathology and are distributed as PNG patches with segmentation masks.

The imbalance ratio between adenocarcinoma and serrated adenoma was approximately 13.7:1, while normal tissue was also scarce. Thus, high multiclass accuracy could coexist with poor minority-class recognition, motivating weighted sensitivity, specificity, kappa, and AUC.

C. Experimental Pipeline

The experiments were conducted in the Weka data mining platform [8]. Initially, an ARFF file was created to reference the image directories, and deep features were extracted from the histopathological images using a pre-trained EfficientNet-B0 network with ImageNet weights. EfficientNet employs a compound scaling strategy that jointly balances network depth, width, and input resolution, producing highly transferable image representations with relatively low computational cost [12]. This characteristic makes transfer learning particularly suitable for medical imaging applications, where labeled datasets are often insufficient to train deep architectures from scratch. In the proposed pipeline, EfficientNet-B0 was used as a fixed feature extractor, generating a 1,280-dimensional feature vector for each image, while the final classification was performed using conventional machine learning algorithms.

Two feature representations were subsequently investigated. **Base 1** consisted of the complete set of features extracted by EfficientNet-B0. **Base 2** was generated by applying Particle Swarm Optimization (PSO), a population-based optimization algorithm inspired by the collective behavior of biological swarms [13]. Feature selection aims to reduce redundancy and eliminate less informative attributes, potentially simplifying the classification models while preserving their discriminative capability.

The dataset was then divided into an 80% development set and an independent 20% test set. Since the EBHI-SEG dataset is highly imbalanced—with 795 adenocarcinoma patches but only 58 serrated adenoma and 76 normal patches—models trained directly on the original distribution may become biased toward the majority classes. To investigate this effect, both feature representations were evaluated under two experimental settings: with and without the Synthetic Minority Over-sampling Technique (SMOTE) [14]. SMOTE generates synthetic minority samples by interpolating neighboring instances in the feature space, potentially improving the recognition of underrepresented classes. To avoid information leakage, oversampling was applied exclusively to the training partitions within the development set, while the validation and independent test sets preserved the original class distribution.

Although transfer learning and class balancing have been extensively investigated in the literature, relatively few studies have systematically analyzed their combined effects in severely imbalanced histopathological datasets containing morphologically similar classes. Moreover, most previous studies emphasize global performance metrics, often overlooking detailed class-wise analyses of clinically relevant minority categories, such as serrated adenoma.

For each experimental scenario, nine classifier configurations were evaluated: Naive Bayes, BayesNet (K2 search with SimpleEstimator), Random Tree, Random Forest with 100, 300, and 500 trees, and Support Vector Machines (Weka SMO) with first-degree polynomial, second-degree polynomial, and radial basis function (RBF) kernels. These classifiers represent distinct learning paradigms, including probabilistic models, decision-tree ensembles, and margin-based classifiers, allowing a comprehensive evaluation of which family best exploits the transferred deep representations. Model performance was assessed through 30 repetitions of stratified 10-fold cross-validation, resulting in 300 performance measurements for each classifier and experimental scenario.

Finally, the best-performing classifier from each feature representation was selected and evaluated on the independent test set, comprising 448 histopathological image patches that were not used during model development.

D. Evaluation

The reported metrics were accuracy, Cohen's kappa, weighted sensitivity, weighted specificity, and weighted area under the receiver-operating-characteristic curve (AUC). For C classes with class weights $w_c = n_c/N$, weighted sensitivity and specificity were interpreted as

$$\mathrm{Sens}_w = \sum_{c=1}^{C} w_c \frac{TP_c}{TP_c + FN_c}, \tag{1}$$

$$\mathrm{Spec}_w = \sum_{c=1}^{C} w_c \frac{TN_c}{TN_c + FP_c}. \tag{2}$$

Cohen's kappa measures observed agreement beyond agreement expected by chance,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3}$$

where p_o is observed agreement and p_e is expected chance agreement. AUC summarizes ranking discrimination across decision thresholds. In the Weka multiclass output, class-wise AUC values were combined through prevalence weighting.

Accuracy was converted from percentage to the $[0, 1]$ interval. To summarize performance using the same rule applied in the project analysis, a composite score was defined as

$$S = \frac{\mathrm{Acc} + \kappa + \mathrm{Sens} + \mathrm{Spec} + \mathrm{AUC}}{5}. \tag{4}$$

The score is not proposed as a clinical utility function. It is an equal-weight descriptive index used to avoid selecting a model from accuracy alone. Means and sample standard deviations were calculated across the 300 exported measurements. Differences between Base 1 and Base 2 were explored by pairing equal run/fold identifiers. Wilcoxon signed-rank tests were used descriptively because the score distributions were not assumed to be Gaussian. Since repeated cross-validation folds are not statistically independent, these p-values are interpreted as exploratory.

IV. Results

A. Training Phase

Table III summarizes the best-performing classifier for each feature representation and balancing scenario during the training phase. Without SMOTE, the highest composite score was achieved by the second-degree polynomial SVM (SVM Poly-2) in Base 1 ($S = 0.8762$), whereas Base 2 slightly favored the first-degree polynomial SVM ($S = 0.8705$). After applying SMOTE, SVM Poly-2 consistently became the best-performing classifier for both feature representations, achieving composite scores of 0.9737 and 0.9704 for Base 1 and Base 2, respectively.

Random Forest with 500 trees produced the highest AUC in both feature bases, reaching 0.9618 and 0.9946 for Base 1, and 0.9613 and 0.9940 for Base 2, without and with SMOTE, respectively. This result indicates that the classifier providing the highest AUC was not necessarily the one with the highest overall composite score, highlighting that model ranking depends on the performance metric considered.

The influence of class balancing is illustrated in Fig. 3. SMOTE improved the composite score for all nine classifier configurations in both feature representations, with average gains of +0.1025 for Base 1 and +0.0956 for Base 2. The largest improvements were observed for Random Forest, reaching +0.1365 (500 trees) in Base 1 and +0.1332 (300 trees) in Base 2.

The performance gains were particularly evident for the SVM models. In Base 1, the best-performing classifier increased its mean accuracy from 84.66% to 96.42%, while weighted sensitivity improved from 0.8466 to 0.9642. Similar improvements were observed for Base 2. In addition to increasing the average performance, SMOTE also reduced variability across cross-validation runs. For SVM Poly-2 in Base 1, the standard deviation of accuracy decreased from 2.4475 to 0.9928 percentage points, whereas the standard deviation of Cohen's kappa decreased from 0.0334 to 0.0119, indicating more stable model behavior.

In addition, Fig. 3 also shows three consistent performance tiers: polynomial SVMs led, Random Forest and RBF SVM formed an intermediate group, and Bayesian or single-tree methods generally ranked lower. Increasing the forest size from 100 to 500 trees changed the score by only about 0.002, suggesting diminishing returns in the fixed feature space.

The effect of PSO-based feature selection was classifier dependent. While Base 2 substantially improved the performance of Naive Bayes (+0.2317 without SMOTE and +0.2053 with SMOTE), only marginal improvements were observed for BayesNet and Random Tree (up to 0.0140). In contrast, the SVM classifiers exhibited slight reductions in composite score after feature selection, with the largest decrease occurring for the RBF kernel under SMOTE ($\Delta S = -0.0380$). For the best-performing SVM Poly-2 model with SMOTE, the reduction from 0.9737 to 0.9704 was significant, although the magnitude of the difference remained very small ($\Delta S = -0.0034$). Overall, these results indicate that PSO can benefit classifiers that are more sensitive to redundant features, whereas the complete EfficientNet-B0 representation already provides an effective feature space for margin-based classifiers.

Based on these results, the best-performing classifiers from each feature representation were selected for evaluation on the independent held-out test set.

B. Held-Out Test Performance

The best-performing classifier from each feature representation was selected based on the training phase and evaluated on the independent test set, comprising 448 histopathological image patches not used during model development. Although the selected models were identified from the SMOTE-based experiments, the independent test set preserved the original class distribution, without any synthetic oversampling. Interestingly, the test results were more consistent with the ranking observed during the non-SMOTE cross-validation, suggesting that the substantial performance gains obtained with SMOTE during training did not fully translate into improved generalization.

Table IV summarizes the overall performance of the two feature representations on the held-out test set. Base 1 correctly classified 383 of the 448 patches, achieving an accuracy of 85.49%, whereas Base 2 correctly classified 378 patches (84.38%). The difference between the two approaches was modest, corresponding to 1.12 percentage points in accuracy and 0.014 in Cohen's kappa. Both models achieved the same weighted AUC (0.949), indicating that PSO-based feature selection largely preserved the discriminative capability of the EfficientNet representation, despite the slight reduction observed in the remaining performance metrics.

The class-wise performance presented in Table V reveals important differences that are not evident from the weighted metrics alone. Base 1 achieved higher recall for Polyp, Low-grade IN, Serrated Adenoma, and Adenocarcinoma, whereas Base 2 showed better performance only for the Normal and High-grade IN classes. Adenocarcinoma remained the easiest category to recognize, with recall exceeding 0.93 for both feature representations. In contrast, High-grade IN and Serrated Adenoma were consistently the most challenging classes, reaching maximum recall values of only 0.500 and 0.583, respectively.

The confusion matrices in Table VI provide further insight into the classification errors. The predominant confusion occurred between High-grade IN and Adenocarcinoma. Fourteen High-grade IN patches were classified as Adenocarcinoma by Base 1, compared with eleven in Base 2. Conversely, Base 2 misclassified eight Adenocarcinoma patches as High-grade IN, whereas Base 1 produced five such errors. These misclassifications are morphologically plausible because both categories share several architectural and cytological characteristics; however, they remain clinically relevant since they involve the distinction between preinvasive and invasive disease.

V. Discussion

Interestingly, although the best-performing models were selected from the SMOTE-based training experiments, their performance on the independent test set was much closer

TABLE III
BEST CONFIGURATION IN EACH FEATURE BASE AND BALANCING SCENARIO (MEAN ± STANDARD DEVIATION)

Base	Scenario	Model	Accuracy (%)	Kappa	Sensitivity	Specificity	AUC	Score
Base 1	Without SMOTE	SVM Poly-2	84.6629 ± 2.4475	0.7898 ± 0.0334	0.8466 ± 0.0245	0.9508 ± 0.0084	0.9473 ± 0.0092	0.8762
Base 1	With SMOTE	SVM Poly-2	96.4154 ± 0.9928	0.9570 ± 0.0119	0.9642 ± 0.0099	0.9928 ± 0.0020	0.9905 ± 0.0030	0.9737
Base 2	Without SMOTE	SVM Poly-1	83.9438 ± 2.4176	0.7800 ± 0.0330	0.8394 ± 0.0242	0.9493 ± 0.0082	0.9442 ± 0.0095	0.8705
Base 2	With SMOTE	SVM Poly-2	95.9721 ± 0.9802	0.9517 ± 0.0118	0.9597 ± 0.0098	0.9920 ± 0.0020	0.9887 ± 0.0032	0.9704



Fig. 3. Composite scores across classifiers. Base 1 contains complete EfficientNet features; Base 2 uses PSO-selected features.

TABLE IV
HELD-OUT TEST RESULTS ($N = 448$)

Metric	Base 1	Base 2
Correct / incorrect	383 / 65	378 / 70
Accuracy (%)	**85.491**	84.375
Kappa	**0.801**	0.787
Weighted precision	**0.848**	0.841
Weighted recall	**0.855**	0.844
Weighted F1	**0.849**	0.840
Weighted MCC	**0.810**	0.797
Weighted AUC	0.949	0.949
Weighted PRC area	**0.796**	0.789
MAE / RMSE	0.2266 / 0.3168	0.2268 / 0.3171
RAE / RRSE (%)	81.569 / 84.993	81.658 / 85.083

to that observed in the non-SMOTE cross-validation. This finding suggests that the substantial improvements produced by synthetic oversampling primarily benefited the internal training evaluation rather than the model's ability to generalize to unseen data. Moreover, SMOTE generates synthetic feature vectors from existing observations and therefore does not increase the diversity of patients, laboratories, scanners, or staining conditions represented in the dataset.

The class-wise results provide additional insight beyond the overall performance metrics. Although both feature representations achieved a weighted AUC of 0.949, High-grade IN and Serrated Adenoma remained the most challenging classes, with recall values ranging from 0.447 to 0.500 and from 0.417 to 0.583, respectively. The confusion matrices further showed that the predominant error involved misclassification between High-grade IN and Adenocarcinoma, a morphologically plausible but clinically important distinction. These findings demonstrate that high weighted performance metrics may conceal substantial deficiencies in clinically relevant minority classes.

The proposed six-class framework should therefore be interpreted as a morphology recognition system operating at the image-patch level rather than as a complete diagnostic solution. While the proposed pipeline achieved robust overall discrimination of histopathological patterns, reliable recognition of minority categories remains an important challenge for practical deployment.

Several limitations should also be considered. First, the dataset was partitioned at the image level, allowing patches originating from the same specimen to appear in different subsets, which may have introduced optimistic performance estimates. In addition, the composite score assigns equal importance to partially correlated performance metrics and has not been clinically validated. Finally, the EBHI-SEG

TABLE V
CLASS-WISE HELD-OUT TEST PERFORMANCE

	Base 1							Base 2						
Class	**TPR**	**FPR**	**Prec.**	**F1**	**MCC**	**ROC**	**PRC**	**TPR**	**FPR**	**Prec.**	**F1**	**MCC**	**ROC**	**PRC**
Polyp	.905	.042	.851	.878	.844	.969	.833	.884	.042	.848	.866	.829	.965	.820
Normal	.563	.012	.643	.600	.588	.941	.431	.750	.012	.706	.727	.717	.960	.602
Low-grade IN	.875	.047	.882	.878	.830	.947	.847	.852	.059	.852	.852	.792	.941	.810
High-grade IN	.447	.027	.607	.515	.484	.852	.362	.500	.037	.559	.528	.488	.841	.340
Serrated adenoma	.583	.002	.875	.700	.708	.952	.605	.417	.002	.833	.556	.582	.958	.470
Adenocarcinoma	.956	.062	.894	.924	.881	.963	.887	.937	.052	.909	.923	.879	.969	.904

TABLE VI
HELD-OUT TEST CONFUSION MATRICES. ROWS REPRESENT THE TRUE CLASSES AND COLUMNS REPRESENT THE PREDICTED CLASSES.

	Base 1							Base 2						
	P	N	L	H	S	A		P	N	L	H	S	A	
P	86	4	5	0	0	0	P	84	3	7	0	0	1	
N	6	9	0	0	0	1	N	3	12	0	0	0	1	
L	9	1	112	5	1	0	L	11	1	109	6	1	0	
H	0	0	7	17	0	14	H	0	0	8	19	0	11	
S	0	0	1	1	7	3	S	1	1	2	1	5	2	
A	0	0	2	5	0	152	A	0	0	2	8	0	149	

P: polyp; N: normal; L/H: low-/high-grade IN; S: serrated adenoma; A: adenocarcinoma.

dataset was obtained from a single public source and does not provide demographic metadata. Consequently, variability related to staining protocols, scanners, geographic origin, and patient populations cannot be assessed, limiting the generalizability of the reported results. Future studies should therefore investigate patient-level data partitioning, external validation on independent cohorts, and more diverse datasets to better assess the robustness and clinical applicability of the proposed methodology.

VI. CONCLUSION

This study evaluated conventional classifiers trained on EfficientNet-B0 representations for six-class colorectal histopathology classification. The polynomial SVM (Poly-2) with SMOTE achieved the highest cross-validation composite score (0.9737). However, when evaluated on the 448 held-out patches using non-SMOTE models, Base 1 achieved 85.49% accuracy ($\kappa = 0.801$, AUC $= 0.949$), marginally outperforming Base 2 (84.38%, $\kappa = 0.787$, AUC $= 0.949$). Although PSO substantially reduced the feature dimensionality, it did not improve the generalization performance of the best-performing classifiers, indicating that the original EfficientNet-B0 representation already provided highly discriminative features for this task.

SMOTE consistently improved the composite score, AUC, and sensitivity during cross-validation, although its application prior to data splitting may have introduced optimistic estimates. Moreover, the benefits of PSO-based feature selection were classifier-dependent: while Naive Bayes exhibited substantial improvements, the top-performing SVM models showed no practical gains and, in some cases, experienced slight performance degradation.

The most clinically relevant finding lies not in the aggregate metrics but in the class-wise results. Although adenocarcinoma achieved recall above 0.93, high-grade intraepithelial neoplasia and serrated adenoma remained considerably more difficult to recognize (recall ≤ 0.583). This discrepancy highlights that weighted performance measures may conceal important deficiencies in minority classes, particularly for lesions with malignant potential.

Overall, the proposed pipeline provides a reproducible baseline for AI-assisted colorectal histopathology rather than a clinically deployable system. The results highlight the importance of carefully evaluating the effects of class balancing and feature selection in severely imbalanced datasets, while reinforcing the need for patient-level data partitioning, stain normalization, and external validation in future translational studies.

ACKNOWLEDGMENT

We are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES), Fundação de Amparo a Ciencia e Tecnologia do Estado de Pernambuco - Brazil (FACEPE), and the National Council for Scientific and Technological Development - Brazil (CNPq), for the partial financial support.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] B. Korbar, A. M. Olofson, A. P. Miraflor, K. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep learning for classification of colorectal polyps on whole-slide images," *Journal of Pathology Informatics*, vol. 8, p. 30, 2017.

[3] G. Yu *et al.*, "Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images," *Nature Communications*, vol. 12, p. 6311, 2021.
[4] M. Bilal *et al.*, "Development and validation of artificial intelligence-based prescreening of large-bowel biopsies taken in the UK and portugal: A retrospective cohort study," *The Lancet Digital Health*, vol. 5, no. 11, pp. e786–e797, 2023.
[5] P. C. Neto, D. Montezuma, S. P. Oliveira, D. Oliveira, J. Fraga, A. Monteiro, J. Monteiro, L. Ribeiro, S. Goncalves, S. Reinhard *et al.*, "An interpretable machine learning system for colorectal cancer diagnosis from pathology slides," *arXiv preprint arXiv:2301.02608*, 2023.
[6] D. Tellez, G. Litjens, P. Bandi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, p. 101544, 2019.
[7] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, pp. 1301–1309, 2019.
[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
[9] P. Kainz, M. Pfeiffer, and M. Urschler, "Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization," *PeerJ*, vol. 5, p. e3874, 2017.
[10] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, pp. 555–570, 2021.
[11] L. Shi, X. Li, W. Hu, H. Chen, J. Chen, Z. Fan, M. Gao, Y. Jing, G. Lu, D. Ma *et al.*, "EBHI-Seg: A novel enteroscope biopsy histopathological haematoxylin and eosin image dataset for image segmentation tasks," *arXiv preprint arXiv:2212.00532*, 2022.
[12] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114.
[13] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 – International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.
[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

Development of a Sensorized Device for Respiratory Rehabilitation Assessment

Letícia Vitória Dulce Teixeira
PET Conexões Mecatrônica/BSI scholarship student
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
leticiavdtx@gmail.com

Lydia Massardi Ciotti
Mechatronics engineering - undergraduate student
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
lydiaciotti14@gmail.com

Sabrina Rodrigues de Araújo
PET Conexões Mecatrônica/BSI scholarship student
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
sabrina.engmect@gmail.com

Tales Pulinho Ramos
Electrical Department
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
tales.ramos@ifsudestemg.edu.br

Thiago Rodrigues Oliveira
Electronics Department, Tutor of PET Conexões Mecatrônica/BSI
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
thiago.oliveira@ifsudestemg.edu.br

Silvana Terezinha Faceroli
Computer Department
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
silvana.faceroli@ifsudestemg.edu.br

Ana Carolina Borges Valente
ConectaLab - Postgraduate student
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
fisiocarolborges@gmail.com

Sara Del Vecchio
Mechanical Department
IF Sudeste MG - Campus Juiz de Fora
Juiz de Fora, Brasil
sara.vecchio@ifsudestemg.edu.br

***Abstract*—Respiratory rehabilitation requires precise and continuous monitoring of lung parameters to ensure the efficacy of treatment and the participation of the patient. This work presents the mechatronic development of an affordable sensorized respiratory assessment system designed to capture airflow dynamics. The system integrates a 3D-printed adaptation tube designed to fit conventional mechanical incentive spirometers, transforming them into digital biofeedback-driven rehabilitation systems. Using an airtight pneumatic structural design, the adapter incorporates a differential pressure sensor (MPX5700DP) and an ESP32-C3-Mini microcontroller, enabling high-resolution data acquisition and transmission through Python-based processing interfaces. Data are processed to control a serious game interface, offering immersive patient motivation while generating objective evaluation reports for clinicians. Set in a dynamic deep-sea environment spanning the world's four oceans, the game utilizes real-time inspiratory and expiratory pressures to control avatar navigation and adapt difficulty levels.The proposed device addresses the limitations of high cost and mechanical fragility commonly associated with commercial spirometers, offering a robust alternative for remote, engaging, and decentralized clinical settings.**

***Index Terms*—pulmonary function, assistance, pressure measurement, physiotherapy, bioengineering**

I. Introduction

Pulmonary rehabilitation with mechanical training and incentive spirometers is critical to the management of chronic airway disorders and the acceleration of post-surgical trauma recovery [1], [2]. Traditional incentive spirometers are widely accessible and effective in clinical settings; however, they rely heavily on manual verification and lack objective digital logging, often leading to low patient compliance and decentralized monitoring deficits [3]. To overcome these barriers, contemporary research has focused on diverse state-of-the-art approaches to extract breathing parameters. Some advanced trends suggest that one completely abandons mechanical spirometry through alternative sensing topologies, such as machine learning-driven pharyngeal phonetic analysis [4], flexible near-field coupling printed radiofrequency resonators placed on the diaphragm [5], or complex multimodal wearable chest patches combining bioimpedance and mechanomyography [6]. While highly innovative, these methods lack direct physical resistance or active volumetric training loops, limiting their deployment in targeted physical therapeutic training.

Conversely, specialized hardware modifications have emerged to introduce targeted monitoring. These include

customized non-invasive extratracheal respiratory assessment ports tailored strictly for laryngectomized patients [7], individualized localized training units built from three-dimensional-printed prototype based on two spring valves to limit the air flux during inspiration and expiration [8], high-accuracy calibration-free architectures utilizing Fiber Bragg Grating (FBG) optical arrays [9], and dedicated portable closed-source wireless spirometers based on thin-sheet flexible flow sensor [10]. Furthermore, frugal design frameworks highlight the cost efficiency of 3D-printed pressure-differential Venturi implementations over standard delicate turbine units [11].

Nevertheless, a critical technological gap remains: the literature rarely explores low cost, open-source hybridization capable of retrofitting existing conventional clinical respiratory hardware. Most current platforms require purchasing entirely proprietary digital devices [1], [10]. There is an evident need for seamless physical adapters that maintain pneumatic integrity while introducing precise real-time biofeedback and automated clinical reporting through gamified structures.

Respiratory dysfunctions, whether resulting from chronic conditions such as asthma and chronic obstructive pulmonary disease (COPD) or from acute events such as stroke and rib fractures, impose a substantial burden on global healthcare systems [1]. Pulmonary rehabilitation programs incorporating individualized respiratory muscle training have demonstrated significant effectiveness in improving lung function, enhancing ventilatory capacity, and reducing the risk of severe complications [12], [13], [2]. Nevertheless, the provision of continuous, standardized, and accessible physiotherapy during and after hospitalization remains challenged by infrastructure constraints, limited availability of specialized professionals, and interruptions in follow-up care. These limitations highlight the need for affordable and reliable technological solutions capable of supporting objective respiratory assessment and facilitating the expansion of rehabilitation services beyond conventional clinical settings. To address these challenges, the objective of this work is to integrate a low-cost mechatronic device based on a Venturi tube geometry, utilizing differential pressure readings to estimate key respiratory parameters in real-time. Furthermore, aiming to mitigate patient drop-out rates and enhance engagement, this hardware system is embedded into an underwater-themed serious game. Through a bi-directional breathing control mechanism, patients embark on an ocean de-pollution mission, transforming repetitive clinical exercises into an immersive, adaptive, and educational therapeutic experience that fosters long-term rehabilitation adherence.

II. Methodology

The proposed hardware architecture is divided into three fundamental layers: the physical pneumatic adaptation, the electronic sensing framework, and the embedded processing unit.

A. 3D-Printed Adaptation Tube and Pneumatic Sealing

The proposed system is based on the mechatronic design of a sensorized adapter tube developed to interface a conventional mechanical incentive spirometer with a digital monitoring platform. Unlike mechanical turbine spirometers, the fluid-dynamic sensing structure employs a custom-designed Venturi-style flow-restricting tube modeled via computer-aided design (CAD) and fabricated through 3D printing. Grounded in Bernoulli's principle, the tube features restricted internal cross-sectional areas that generate a localized fluid velocity variation during patient expiration, creating a proportional differential pressure between two measurement ports [11]. Its internal geometry was designed to function as an in-line flow-conditioning structure, enabling the integration of pressure sensing elements without interfering with the operation of the original rehabilitation device.

A critical design requirement was the preservation of pneumatic sealing throughout all mechanical interfaces. Air leakage during inspiratory or expiratory maneuvers can significantly affect flow estimation accuracy and reduce the pressure load necessary for the proper operation of the mechanical spirometer. To address this issue, the adapter incorporates high-tolerance slide-fit connections and strategically positioned pressure ports for differential pressure measurement. This configuration allows the acquisition of the true pressure differential (ΔP) generated by the airflow while minimizing additional flow resistance and preserving the fluid dynamic characteristics of the conventional rehabilitation process. Applying Bernoulli's principle and the continuity equation, the theoretical volumetric flow rate (Q) is derived from the differential pressure ($\Delta P = p_1 - p_2$) across the Venturi constriction, 'Eq. 1'.

$$Q = A_1 \sqrt{\frac{2(p_1 - p_2)}{\rho \left(1 - \left(\frac{A_1}{A_2}\right)^2\right)}} \tag{1}$$

Where:

- Q is the volumetric flow rate (m^3/s),
- A_1 and A_2 represent the internal cross-sectional areas at the primary and constricted ports (m^2),
- p_1 and p_2 are the corresponding static pressures (Pa), and
- ρ is the fluid density of air.

The resulting hybrid architecture enables the simultaneous execution of standard respiratory therapy exercises and real-time digital interaction through a serious game environment. In this way, physiological data can be continuously monitored without altering the therapeutic protocol or the patient's experience with the conventional spirometer.

B. Electronic Sensing and Embedded Hardware

The conversion from pneumatic pressure to electrical signal is performed by the MPX5700DP (NXP Semiconductors), a piezoresistive, monolithic silicon differential pressure sensor. This sensor was selected for its robust signal conditioning and wide operating range (up to 700 kPa),

enabling the accommodation of forced expiratory maneuvers without saturation. Pressure sensing is accomplished via an MPX5700DP piezoresistive differential transducer. Although its nominal operating range extends up to 700 kPa—providing high structural headroom against overpressure during forceful exhalation—respiratory maneuvers typically induce pressures below 20 kPa. To compensate for reduced dynamic range resolution over lower pressure regimes, signal conditioning relies on full 12-bit ADC quantization on the ESP32-C3 sampled at 100 Hz. Future iterations will evaluate lower-range differential sensors (e.g., MPX5010DP, 0–10 kPa) to maximize signal-to-noise ratio during subtle inspiratory flows. The core processing and connectivity consist of an ESP32-C3-Mini microcontroller module, with 32-bit RISC-V single core and a 12-bit analog-to-digital converter (ADC) that samples the analog output of the MPX5700DP at a calibrated frequency of 100 Hz to prevent aliasing during sudden respiratory bursts. The processing flow involves signal filtration, analog-to-digital conversion, data buffering, and real-time packet transmission using USB cable in order to reduce latency.

C. Software and Processing Interface

The data stream is captured by a nearby workstation or smartphone running a Python processing pipeline. The game architecture and user loop are structured as follows:

Initialization and Customization: The protocol begins with a digital onboarding phase where patients select and customize their marine avatar, enhancing psychological ownership and initial compliance.

Bi-directional Motion Mapping: Once gameplay begins, the real-time differential pressure generated during the patient's expiration (blowing) and inspiration (inhaling) is mapped to multi-axial navigation. These pneumatic inputs control the avatar's upward, downward, and lateral movements, enabling the character to dodge underwater obstacles and defeat oceanic enemies. Concurrently, the magnitude of the measured pressure dynamically modulates the avatar's swimming speed, directly tying physical exertion to in-game velocity.

Phase-Dependent Sensor Calibration: To enforce progressive clinical training, the game is structured around the world's four major oceans, with each phase representing a different ocean. Programmatically, the software alters the input sensitivity of the sensor across different stages. Lowering or increasing the sensitivity scale changes the physical effort required to move the avatar a given distance, establishing an automated, firmware-driven rehabilitation progression.

Educational and Difficulty Scaling: As the player advances through the levels of each ocean, the difficulty increases, and the main characters, background landscapes, and enemies dynamically change to mirror that specific geographical region. Indigenous marine animals are displayed in the background, and each new level unlocks environmental or biological trivia (curiosities regarding the ecosystem or its fauna), combining clinical rehabilitation with active cognitive learning.

Win Condition: Patients collect rewards and extra lives while fulfilling the core mission of purging ocean pollution.

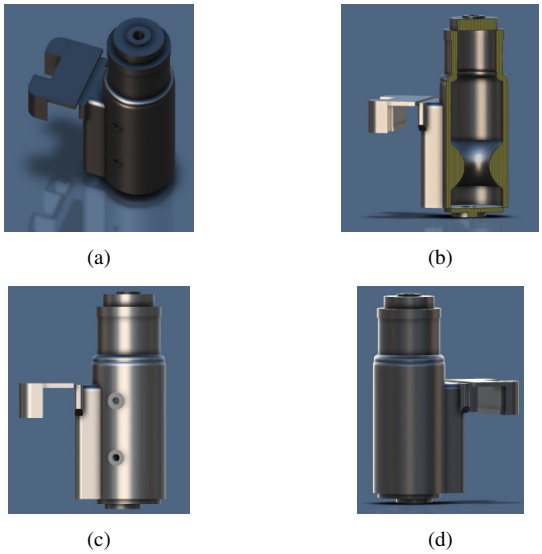


Fig. 1: Results obtained for the proposed method: (a) isometric view, (b) section view: Venturi detail, (c) front view , and (d) back view.

The game is won when the patient completes the entire aquatic journey and successfully unlocks a mystical sunken chest containing the ultimate water-cleansing antidote.

III. Results and Discussion

A. 3D-Printed Adaptation Tube and Pneumatic Sealing

The modeled device to be connected directly to a conventional mechanical incentive spirometer is presented in "Fig. 1" which details the CAD model of the adapter, highlighting the Venturi geometry and pressure ports (b). Grounded in Bernoulli's principle, the tube features restricted internal cross-sectional areas that generate a localized fluid velocity variation during patient breath, creating a proportional differential pressure between the two measurement ports. Although the airflow from breathing is pulsatile, simplifications in the modeling assumed a uniform and continuous flow, remaining steady during the interval of measurements and evaluations along the same streamline. In addition, the internal walls of the device were printed with minimal roughness to reduce the discharge coefficient of the duct, minimizing losses and the user's effort during breathing.

B. Electronic Sensing and Embedded Hardware

Initial bench tests were carried out to correlate the differential pressure read by the sensor with the volumetric flow of the air. Preliminary bench tests for the estimation of forced expiratory volume in one second (FEV_1) were performed with the proposed device, by a simulated flow, "Fig.2".

However, additional experimental cycles are required to optimize the device geometry and evaluate alternative filament materials. These improvements may reduce pressure losses within the duct and minimize the respiratory effort required from patients during training.

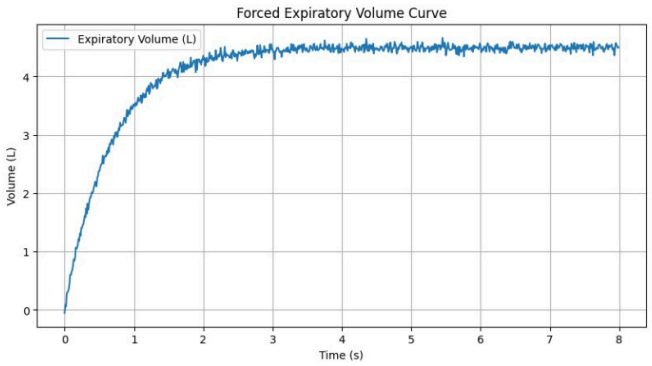


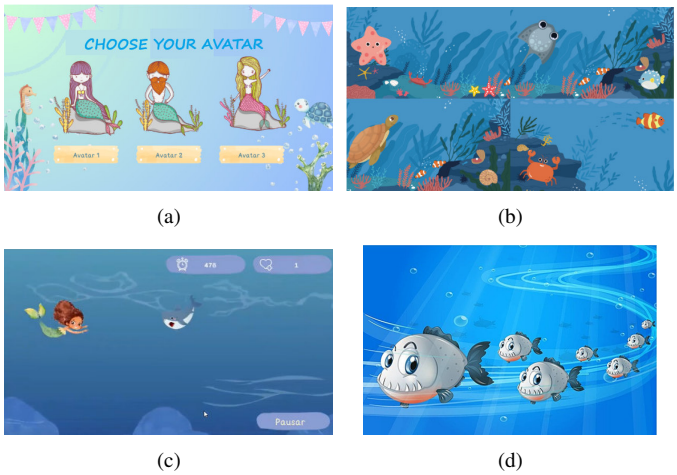
Fig. 2: Curve of Forced expiratory volume.

Furthermore, the MPX5700DP sensor has a measurement range of up to 700 kPa, which substantially exceeds the pressures typically observed during respiratory maneuvers. It was a choice for initial prototyping/bench availability. Therefore, sensors with narrower measurement ranges should be evaluated to improve resolution under the intended operating conditions. Nevertheless, the MPX5700DP provides a sensitivity of 6.4 mV/kPa and demonstrated an adequate response during forced expiratory volume in one second (FEV_1) measurements.

Upon completion of this stage, the physiological parameters associated with respiratory training can be estimated, enabling the generation of reports for subsequent analysis and comparison. Preliminary assessments demonstrated signal stability, a short response time, and the feasibility of obtaining measurements during both inhalation and exhalation.

C. Software, Processing and Serious Game Interface

The digital stream is captured by a localized workstation or smartphone running a Python processing pipeline. In the current stage of the research, differential pressure was used to control the serious game. At this stage of the work, the focus was on hardware development and bench validation of the signal processing algorithms, without conducting tests with humans. The graphical interface and game mechanics were designed to be intuitive and motivating, aiming for future clinical applicability. Fig. 3 shows the game interface developed based on deep sea. The introduction screen is shown in "Fig. 3a", the four game oceans in "Fig. 3b", a villain animal example in "Fig. 3c" and one of the obstacles, ocean current, in "Fig. 3d". Within the application, the mathematical conversion from raw differential pressure (ΔP) to volumetric flow rate (Q) is executed in real-time, mapping values to deliver parameters like Forced Vital Capacity (FVC) and Peak Expiratory Flow (PEF).These estimated physiological parameters serve as direct telemetry inputs for a serious game, where the patient's respiratory flow controls in-game mechanics (like overcoming obstacles or achieving targets), transforming repetitive clinical exercises into an immersive therapeutic experience. The effectiveness of the system with real users and the clinical validation of the estimated spirometry parameters will be the subject of future studies, subject to approval by the Research Ethics Committee (CEP).



(a) (b)

(c) (d)

Fig. 3: Game interface: (a) introduction screen, (b) four game oceans, (c) villain example and (d) ocean current.

IV. Comparative Analysis with State-of-the-Art Systems

To evaluate the mechatronic design and clinical feasibility of the developed system, its characteristics were cross-examined with other works of the literature considered as summarized in "Tab. 1".

V. Design Insights and Advantages

In recent years, the integration of biomedical technology with mobile health (mHealth) and telemedicine has emerged as a strategy to decentralize care, allowing objective home evaluation [1] and [12]. Traditional methods often rely on subjective self-reporting or fragile and expensive mechanical devices that are impractical for remote follow-ups [1]. Although portable spirometers exist, there is a clear technological gap for solutions that combine clinical accuracy, structural durability, and low cost. Frugal engineering guidelines address this discrepancy by eliminating delicate moving components—such as turbines or rolling seals—which are highly susceptible to dust, moisture, and mechanical wear, replacing them with solid-state pressure-differential techniques such as the one developed by [11]. In addition, [11] utilized an MPX5010DP sensor, which operates within a 10 kPa range. While highly accurate for quiet breathing, it risks saturation during forced expiratory maneuvers executed by robust individuals. The selection of the MPX5700DP provides an expanded dynamic headroom, assuring sensor integrity and linear output across diverse patient profiles. The platform implemented by [1] relies on a commercial mechanical-turbine device (ezOxygen). Turbine-based mechanisms introduce moving components that deteriorate due to saliva condensation, moisture, and dust contamination. By adopting a Venturi model, the developed system significantly drops maintenance costs and ensures device longevity in home-care scenarios. However, experimental

TABLE I: Comparison of Recent Respiratory Monitoring and Spirometry Systems

System	Sensing	Hardware	Biofeedback/UI	Report	Hybridization	Vulnerability
Proposed Adapter	Differential pressure (MPX5700DP)	3D-printed tube + ESP32-C3	Serious game interface	Yes	Retrofits standard incentive spirometers	Very low
Wallace et al. (2026)	Differential pressure (MPX5010DP)	3D-printed tube + Arduino	Character LCD	No	No	Low
Liao et al. (2025)	Commercial turbine	Proprietary ezOxygen device	Mobile app	Partial	No	High
Ren et al. (2025)	Airflow deformation	Wireless spirometer	Mobile graphs	No	No	Low
Zhang et al. (2025)	FBG optical sensor	Optical system + AI	Laboratory software	Yes	No	Very low
Abounasr et al. (2026)	RF resonator	Flexible wearable patch	Data logger	No	No	Low
Berkebile et al. (2026)	Bioimpedance + MMG	Wearable chest strap	Workstation monitor	No	No	Low
Nardi et al. (2025)	Spirometer interface	Silicone tracheal adapter	Commercial UI	Dependent	Yes	Low

evaluations of the behavior of airflow in the tube, taking energy losses and pressure drops into account, are needed in order to establish the uncertainties in pressure assessment, which will directly impact the evaluation of respiratory flow. Computational Footprint and Connectivity: Although the implementation by [11] relies on an external display (LCD/LEDs) on an Arduino platform, the proposed system channels the efficiency of the ESP32-C3-Mini. This enables a downscaled physical footprint, a lower production cost, and direct Native IoT cloud integration out-of-the-box, resembling the wireless versatility used in clinical trials, which can be evaluated in the future.

Unlike standalone digital devices such as those by [1] and [10], which oblige clinics to purchase redundant electronic spirometers, the proposed device utilizes an adapter configuration. It retains the validated physical mechanisms of conventional devices [3], augmenting them with precise sensing at a fraction of the cost. While flexible antennas [5] and multi-sensory chest patches [6] offer excellent passive vital sign logging, they cannot guide or enforce active respiratory volumes. The proposed 3D-printed structural enclosure ensures absolute sealing, keeping user breathing efforts directly linked to active exercise execution. Furthermore, eliminating turbine wheels eliminates mechanical failure modes caused by saliva moisture buildup, outperforming commercial clinical equivalents.

Standalone frugal systems [11] lack interactive software layer tracking, rendering home rehabilitation repetitive. The proposed integrated architecture bridges accurate solid-state transducer mapping with serious gaming, filling a prominent clinical gap by generating structural analytics for remote patient reporting.

Future work will focus on implementing the real-time mathematical conversion from raw differential pressure (ΔP) to volumetric flow rate (Q), mapping these values to deliver clinical parameters such as Forced Vital Capacity (FVC) and Peak Expiratory Flow (PEF).

ACKNOWLEDGMENT

The authors thank the Federal Institute of Education, Science, and Technology of Southeast Minas Gerais — Juiz de Fora Campus, and the Tutorial Education Program (PET), through the Pro-Rectorate of Teaching and the Ministry of Education (MEC), whose scholarships and resources were essential to support the research and extension activities presented here.

VI. CONCLUSION

The mechatronic development of the proposed device achieves a good balance between frugal engineering paradigms and the high-performance requirements of modern biomedical instruments. By replacing expensive closed commercial hardware with a low-cost and robust differential sensing array based on the ESP32-C3 and MPX5700DP, this solution proves to be capable of providing reliable metrics for decentralized respiratory rehabilitation, bridging the gap between clinical accessibility and engineering durability. Furthermore, the device was developed to ensure the airtightness of the pneumatic flow, allowing the patient to use conventional physiotherapy equipment while simultaneously interacting with the serious game. This hybrid integration between hardware and software makes it possible to provide real-time biofeedback, filling the literature gaps on the connection between precise measurement and gamification-based rehabilitation systems.

REFERENCES

[1] C.-A. Liao, T.-H. Young, L.-W. Kuo, C.-Y. Fu, S.-A. Chen, Y.-S. Tee, S.-C. Kang, C.-T. Cheng, and C.-H. Liao, "Empowering recovery: a remote spirometry system and mobile app for monitoring and promoting pulmonary rehabilitation in patients with rib fracture," *Trauma Surgery & Acute Care Open*, vol. 10, no. 1, p. e001309, Jan. 2025.

[2] B. Zhu, F. Xue, M. Li, and Y. Chen, "Long-term benefits of pulmonary rehabilitation with respiratory muscle training in people with COPD: A retrospective cohort study," *Respiratory Medicine*, vol. 245, p. 108214, Jan. 2025.

[3] N. Mera, E. C. W. Luna, and V. Benavides-Cordoba, "Comparative analysis of pulmonary ventilation distribution between low-cost and branded incentive spirometers using electrical impedance tomography in healthy adults: Study protocol," *PLoS ONE*, vol. 19, no. 5, p. e0302476, May 2024.

[4] A. J. Moshayedi, A. M. Aghda, S. A. Eftekhari, M. E. Andani, and S. Salahshour, "Respiratory parameter estimation using pharyngeal phonetics and machine learning: Breaking free from spirometry," *Engineering Applications of Artificial Intelligence*, vol. 160, p. 111989, Mar. 2025.

[5] J. Abounasr, M. E. Gharbi, R. Fernández-García, and I. Gil, "Flexible body-integrated breathing monitoring system based on near-field coupling printed sensor," *Measurement*, vol. 258, p. 119371, Jan. 2026.

[6] J. A. Berkebile, H. T. Crane, J. A. Sánchez-Pérez, K. Swamy, F. N. Rahman, and O. T. Inan, "reSPIRE: A wearable multimodal sensing system for monitoring ventilation, cardiovascular dynamics, and respiratory muscle activity," *Biosensors and Bioelectronics*, vol. 301, p. 118460, Mar. 2026.

[7] C. E. M. Nardi, M. A. F. de Castro, L. P. Kowalski, R. A. Dedivitis, and D. A. Ribeiro, "Development of an extratracheal device for pulmonary function assessment in laryngectomized patients," *European Archives of Oto-Rhino-Laryngology*, vol. 282, pp. 3771-3776, Apr. 2025.

[8] Ç. Emirza Cilbir, Ö. Ertan Harputlu, H. Uçgun, B. Ülker Ekşi, S. Kaya, E. Kiyan, G. Kuran Aslan, and B. Akinci, "Individualized respiratory exercise device originated in Türkiye: design, prototype, and first results," *Turkish Journal of Medical Sciences*, vol. 56, no. 1, pp. 99-106, Feb. 2026.

[9] R. Zhang et al., "A high-accuracy calibration-free spirometer based on Fiber Bragg Grating technique with neural network optimization," *Sensors and Actuators: A. Physical*, vol. 389, p. 116529, May 2025.

[10] H. Ren, J. Wang, L. Yang, T. Zhang, G. Li, Y. Tang, T. N. D. Yu, and Y. Shen, "A portable wireless spirometer device for long-term pulmonary function monitoring and training," *Biosensors and Bioelectronics*, vol. 270, p. 116937, Apr. 2025.

[11] J. Wallace, P. Checa Ríf á, T. Kannathasan, C. F. Hayfron-Benjamin, P. Anyanwu, and D. Piaggio, "A frugal arduino-based spirometer for low-resource settings: design, development and validation of a preliminary prototype," *Frontiers in Bioengineering and Biotechnology*, vol. 13, art. 1664127, Jan. 2026.

[12] Z.-K. Ma, H.-H. Jiang, Y.-H. Tang, J. Yang, Q.-Ch. Wei, and Q. Gao, "Effect of a comprehensive mobile-based respiratory training program on respiratory function in survivors of acute stroke: Randomized controlled trial," *JMIR mHealth and uHealth*, vol. 14, p. e78637, Apr. 2026.

[13] Z. Sánchez-Milá, V. Abuín-Porras, C. Romero-Morales, J. Almazán-Polo, and J. V. Saornil, "Effectiveness of a respiratory rehabilitation program including an inspiration training device versus traditional respiratory rehabilitation: A randomized controlled trial," *PeerJ*, vol. 11, p. e16360, Dec. 2023.

Integrating High-Resolution Electrophysiological Mapping with Histological Remodeling in a Rabbit Model of Ventricular Tachycardia

Mouhamed Zakiou K.A. Raimi
Department of Biomedical Engineering (HeartLab/UFABC)
University of Abomey Calavi (Biosciences Department/UAC)
Sao Paulo, Brazil
kolaraimi2016@gmail.com

Mohammad Moualla
Department of Biomedical Engineering (HeartLab/UFABC)
Sao Paulo, Brazil
moualla88@gmail.com

José Gomes
Department of Biomedical Engineering (HeartLab/UFABC)
Sao Paulo, Brazil
junior.gomes@ufabc.edu.br

João Salinet
Department of Biomedical Engineering (HeartLab/UFABC)
Sao Paulo, Brazil
joao.salinet@ufabc.edu.br

Abstract— Cardiac arrhythmias arise from a complex interplay between the heart's electrical activation and its structural architecture. Ventricular tachycardia (VT), a major cause of sudden cardiac death, typically requires detailed substrate mapping and signal analysis to delineate arrhythmogenic channels and scar regions for targeted ablation. Nevertheless, the quantitative relationship between local electrical and optical signals—acquired simultaneously via high-resolution optical and electrical mapping—and the underlying myocardial microstructure has rarely been characterized. This study aims to evaluate the spatial correspondence between electrophysiological mapping data and myocardial microstructure in the heart undergoing VT. Using a Langendorff perfusion system, hearts from two New Zealand White rabbits underwent VT induction with 1 µM carbachol and an S1–S2 stimulation protocol (100 bursts, 2 ms pulse duration, 1.3 s inter-burst interval); a third heart served as a non-arrhythmic control. High-resolution epicardial electrical and optical mapping were performed. Myocardial samples corresponding to each electrode site (n = 12 tissue samples per heart) were carefully excised and processed for histological analysis. Preliminary observations indicated that structural remodeling varied spatially across electrode-mapped regions. Some regions exhibited tissue organization comparable to that of the control heart, whereas others demonstrated mild myocyte disarray, increased intercellular spacing, and mild to moderate interstitial cellular infiltration. Notably, areas with near-normal histological appearance could still display abnormal electrical signals. These findings point to a complex, heterogeneous, and non-linear relationship between myocardial structure and local electrical activity. While cellular disorganization appears to contribute to arrhythmogenesis, it does not invariably manifest as abnormal electrical activity at the electrode site. Voltage mapping findings were concordant with these observations and suggest that voltage mapping alone may not fully capture the extent and pattern of underlying myocardial remodeling. This work underlines the importance of integrating detailed electrophysiological mapping with spatially comprehensive histological assessment in experimental models. Such an integrative approach may enhance mechanistic understanding of VT, refine identification of arrhythmogenic substrates, and ultimately improve the precision and efficacy of catheter ablation strategies.

Keywords— *Ventricular tachycardia, Electrophysiological mapping, Myocardial remodeling, Histological correlation, Arrhythmogenic substrate.*

I. INTRODUCTION

Cardiac arrhythmias arise from complex interactions between cardiac electrophysiology and myocardial microarchitecture, including fibrosis and myofiber orientation [1,2]. Ventricular tachyarrhythmias are associated with a substantial risk of sudden cardiac death; catheter ablation is an effective therapeutic modality but is limited by arrhythmia recurrence and procedural complications [3]. Ventricular tachycardia (VT) therefore remains a major contributor to sudden cardiac mortality, and detailed substrate characterization—encompassing high-resolution mapping and advanced electrophysiological signal analysis—is required to delineate arrhythmogenic channels and scar heterogeneities that can guide targeted ablation and improve clinical outcomes. Multimodal imaging approaches (optical and epicardial) in Langendorff-perfused rabbit hearts demonstrate good concordance for organized rhythms but reveal significant discrepancies during complex arrhythmias [4–6]. Structural remodeling, including myocyte disarray and alterations in wall thickness, generates heterogeneous patterns of activation and repolarization that facilitate reentrant activity. Fibrosis itself is highly heterogeneous in space: some fibrotic regions maintain conduction and exhibit pro-arrhythmic behavior, others are electrically quiescent,

Funding agencies: FAPESP-TWAS-UFABC.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

and, conversely, certain areas with complex electrical activity may appear structurally normal on standard histological assessment [7–9]. The impact of fibrosis on conduction properties and rotor dynamics remains incompletely understood, largely due to its heterogeneous spatial distribution [2,10–12]. Integrative methodologies that combine electrophysiological recordings with three-dimensional microstructural information indicate the presence of frequency-dependent conduction block within fibrotic regions, effectively acting as a “low-pass filter” that can facilitate reentry [10,11]. Multi-electrode arrays (MEAs) alone may fail to detect intramural fibrotic substrates [12], whereas optical mapping offers high-resolution, contact-free recordings but is only rarely co-registered with histology in the same preparation. MEAs enable high-density epicardial mapping across sinus rhythm, paced activation, and reentrant circuits [11,13]. Three-dimensional correlative studies have demonstrated that myocardial architecture, non-myocyte content, and gap junction remodeling collectively modulate arrhythmia dynamics [10–12].

Rabbits constitute a well-established experimental model for reentrant arrhythmias and pharmacological testing, with repolarization characteristics that closely resemble those of humans. The Langendorff-perfused rabbit heart model provides a robust platform to relate structural remodeling to arrhythmogenic conduction patterns. However, despite the availability of high-resolution mapping of activation sequences, rotor behavior, and fibrillatory dynamics, direct histological validation at individual recording sites remains relatively uncommon [4,6,8,11–13]. Consequently, the integration of MEA-derived electrophysiological data with the corresponding underlying microstructure remains a critical and largely unmet need.

This study proposes a spatially defined histology method at the MEA segment level by combining optical-electrical rabbit mapping [4,5] with correlative electrophysiological pipelines [10,11,12]. To determine whether electrophysiological patterns correlate with healthy versus remodeled arrhythmogenic tissue, this study performed histological analysis on distinct ventricular regions dissected from beneath the MEA during VT, while electrical activity was also mapped with optical mapping.

II. METHODOLOGY

The study was authorized by the Federal University of ABC's Animal Ethics Committee (protocol 3947230519).

A. Experimental design

Three New Zealand rabbits (3-4 kg) had their hearts isolated and perfused using the Langendorff system. One heart perfused for 4-5 hours without arrhythmia induction served as a control. Sustained VT was induced in the remaining two hearts using programmed stimulation and 1 μM carbachol and pinacidil. To induce cardiac arrhythmias, a custom-made bipolar pacing catheter with two electrodes on its tip and 4 mm intra electrode spacing was used. The system included a custom-built, battery-powered, programmable current-source stimulator (Aleksa Tech) designed to deliver timed and controlled cardiac stimulation. In the S1-S2 protocol, a train of ten S1 (pulse width of 2 ms 1.3 s intervals) pulses with a 300 ms period was followed by an S2 pulse, with the interval decreasing from 150 ms to 20 ms until the refractory period was reached. Optical mapping was conducted by utilising the voltage-sensitive dye Di-4-ANBDQPQ. Norpix-StreamPix 9 software was used to capture simultaneous recordings at 500 frames per second, with a resolution of 1600 × 1000 pixels, 12-bit depth, and a total duration of 10 seconds (5000 frames total) [4]. To evaluate the Langendorff system's effects, heart weight was compared before and after perfusion.

B. Electrophysiological mapping

Electrical and optical mapping, along with signal plotting, were used to confirm the mechanism of generated arrhythmias, as described before [4]. Voltage maps give a static assessment of cardiac tissue, with local electrogram (EGM) amplitude inversely related to the degree of remodelling or scar tissue replacement [14,15]. Voltage maps were used to characterize the epicardial substrates of the produced arrhythmias and to compare them with histological characteristics. The voltage cutoff for healthy epicardial tissue in unipolar configuration was defined as 17.4 mV for the ventricle (V) [14,15] . MEAs were produced from platinum electrodes removed from a reprocessed mapping catheter (St. Jude Medical Livewire, Medline ReNewal), insulated, and embedded in a PET substrate to improve flexibility, epicardial conformity, and signal acquisition. Voltage maps of the ventricular region were generated utilizing MEAs (with square shape 4×4 electrodes with a side length of 1 mm), specifically the 16 electrodes arranged from 81 to 97 with 4 mm distance between them, at 4 kHz, that make up each array [14]. Each voltage map was created using a single beat and 16 peak-to-peak amplitudes taken simultaneously from 16 sites. Peak-to-peak amplitudes were manually derived from unipolar signals.

Purple represents amplitudes above the relevant threshold (≥17.4 mV for the normal healthy ventricle). Values between 17.4 to 0.5 mV were classified as transition regions with reduced voltage but not compatible with dense scar, while values below 0.5 mV were defined as low voltage areas (LVAs), representing completely nonfunctional tissue (displayed in red), following the colour scale convention used in clinical electroanatomical mapping [14,16]. No spatial interpolation was used. Maps were displayed, with the 16 points represented as 16 squares (each holding the MEA electrode number), and the colour of each square denoting the amplitude range. Values over the healthy epicardial voltage cutoff are referred to as "Healthy Values" (HV), values below 0.5 mV are LVAs, and values within the transition region (17.4 to 0.5 mV) are "non healthy tissue".

C. Histology and tissue sampling

After labelling the MEA areas on the heart and photographing them, the hearts (control and arrhythmogenic) were perfused with 4% formaldehyde at the same flow rate as throughout the experiment and placed in a small beaker containing 4% formaldehyde for 30 minutes. Heart weight was measured again to compare pre- and post-Langendorff values. Keeping the same orientation as during Langendorff perfusion, the mapped tissue sections were dissected and sliced into four spatial pieces that were exactly aligned with electrode placements.

For each ventricular MEA, three-quarters of the mapped tissue area (12 mapped electrodes tissue area per heart = 36 tissue samples) was immersed in 4% formaldehyde overnight at room temperature, then refrigerated until histological processing. Each tissue segment (representing one electrode) was stained with hematoxylin and eosin (H&E) to evaluate

cellular morphology, myocyte alignment, and inflammatory alterations, as previously described [17]. The features examined were cardiac hypertrophy, intercellular spacing, nuclear clustering, and mononuclear or interstitial cell infiltration, normal myocytes are parallel orientated, but damaged tissue has scattered and disorganized myocytes with oblique fibers. Each feature was semi-quantitatively evaluated by an experienced investigator blinded to the electrophysiological mapping data, as follows: 0 = minimal, 1 = mild, 2 = moderate, and 3 = severe [18,19]. Based on these variables, a semi-quantitative histological analysis of mapped tissue areas at the electrode level was performed on the control and arrhythmogenic hearts to determine concordance or discordance with abnormal signals recorded at each electrode.

D. Statistical analysis

Statistical analyses were performed in MATLAB (version R2024a; MathWorks, Natick, MA, USA) using the Statistics and Machine Learning Toolbox. Spearman's rank correlation coefficient and paired t-test were computed, and figures were generated using native MATLAB plotting functions.

III. RESULTS

Heart weight before and after the experiment, voltage mapping and histological analysis are depicted in the following tables and figures.

As shown in Table I, on the average, the hearts showed a very slight decrease of 1.4% from the extraction to the end of the experiment. The average difference was -0.21 g, with a 95% confidence interval ranging from -0.87 to +0.45 g, which indicates that the true population effect could be a decrease, no change, or even a slight increase. Consequently, the paired t-test was not statistically significant ($t(2) = -1.34$, $p = 0.311$) difference in the heart weights of the animals at the time of immediate extraction following the initial surgical procedure or at the end of the experimental protocol either in control or non-control condition. This data show that the experimental interventions used throughout the study period did not cause any obvious or only minor detectable changes in heart mass, that would distinguish the non-control heart from the control heart. Furthermore, the lack of a significant difference at both time points suggests that any potential procedural or physiological variability impacting heart weight was probably low and uniformly distributed across the two conditions.

TABLE I: Animals' hearts weigh right after the extraction and at the end of experiment in control heart vs hearts under ventricular tachycardia (VT).

Heart weight / Experiments	Right after heart extraction (g)	End of the experiment (g)
Control	14.4	13.98
VT: Case 1	12	11.7
VT: Case 2	9.1	9.2

The figs 1 and 2 below depicted the simultaneous acquisition of optical (in blue) and electrical (in red) signals from the mapped region of the heart, as defined by the MEA recording area, during a VT episode.

Notably, the observed rhythms follow a fast and consistent pattern across both modalities, implying a stable and repeating electrophysiological activation sequence. The simultaneous nature of these recordings enables a direct comparison of electrical and optical activities. The detected rhythms' regularity and frequency indicate a sustained tachyarrhythmia.

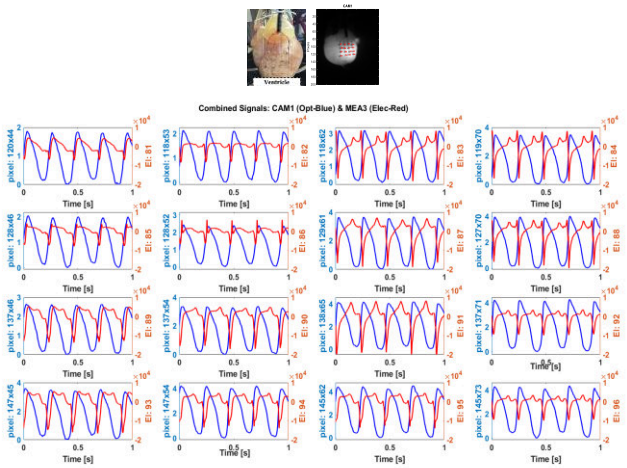


Fig. 1. Electrical and optical signals at the electrode scale of the heart under ventricular tachycardias: case 1 (Blue: optical, Red: electrical).

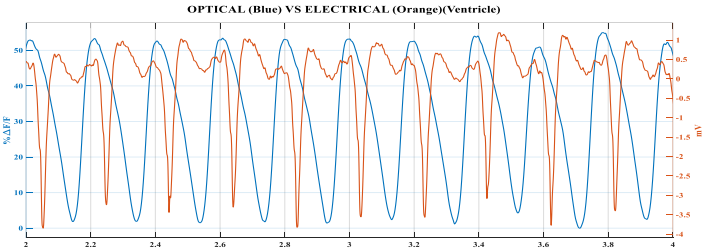
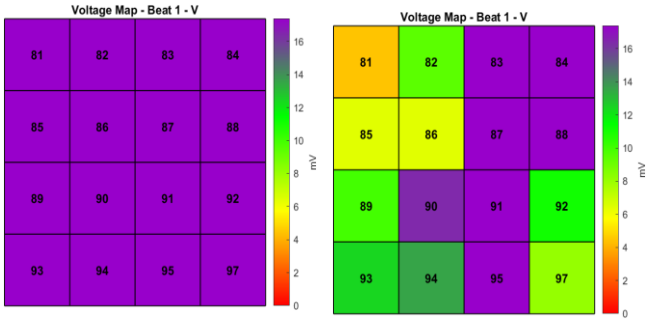


Fig. 2. Electrical and optical signals at one electrode of the heart under ventricular tachycardias: case (Blue: optical, Orange: electrical)

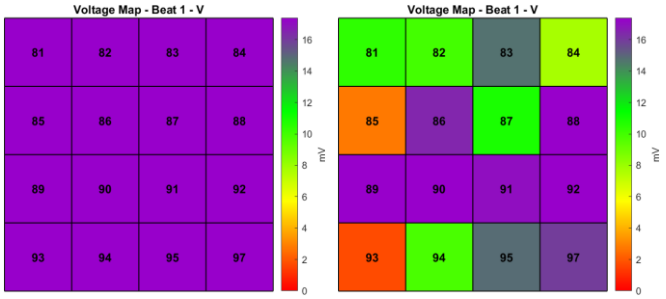
The peak-to-peak voltage mapping of the control heart's ventricular region (see figs. 3,4) showed that all 16 electrodes within the multi-electrode array recorded voltage values within the normal physiological range. The mapping area was consistently purple. This uniform distribution of normal voltages indicates that the control heart, which was perfused in the Langerdoff system without inducing cardiac arrhythmia, has intact myocardial electrical integrity. The absence of low-voltage zones indicates that the control condition maintained healthy despite perfused until heart is arrested, uniform tissue excitability, providing a reliable baseline.

In notable contrast, voltage maps obtained from hearts subjected to experimentally generated VT revealed a very distinct pattern. In these VT conditions, several of the electrodes reported peak-to-peak voltage values below the defined threshold. Low voltage areas often indicate re-entrant arrhythmias, such as VT. The increasing size of these abnormal voltage areas in the VT group emphasizes the electrophysiological alterations associated with sustained tachyarrhythmias.



Sinus Rhythm: Control heart heart VT: Case 1

Fig. 3. Peak to peak voltage mapping of control heart vs VT case 1 (Colours: Purple-normal voltage areas while Red-low voltage areas).



Sinus Rhythm: Control heart VT: Case 2

Fig. 4. Peak to peak voltage mapping of control heart vs VT case 2 (Colours: Purple-normal voltage areas while Red-low voltage areas).

The histological study (see fig.5), of the control heart showed well-preserved and organized tissue architecture across all myocardial areas. Each of the analyzed histological parameters—such as myocyte alignment, interstitial space width, and absence of inflammatory infiltrates —was assigned a score of zero, indicating a minimal degree of abnormalities. This uniform scoring suggests that the control heart, which received only standard perfusion without arrhythmia induction, was structurally intact throughout the experimental procedure. The lack of identifiable histological changes confirms the electrophysiological findings indicating the control preparation was healthy and free of pathological remodelling. In contrast, hearts from induced VT cases showed changes in tissue organization. These modifications included altered myofiber hypertrophy (black bar), increased intercellular spacing (yellow bar), isolated areas of myocyte disarray, and, in some cases, mild to moderate interstitial or inflammatory cell infiltration (blue circle). Importantly, the severity of these changes did not vary uniformly across the tissue; rather, histological scores varied inconsistently among the various tissue segments examined. Some segments exhibited very minor modifications (scores 0 or 1), whilst others showed more substantial anomalies (scoring 2 or 3), indicating a diverse pattern of structural remodelling. This inconsistency is consistent with the known pathophysiology of ventricular tachycardia.

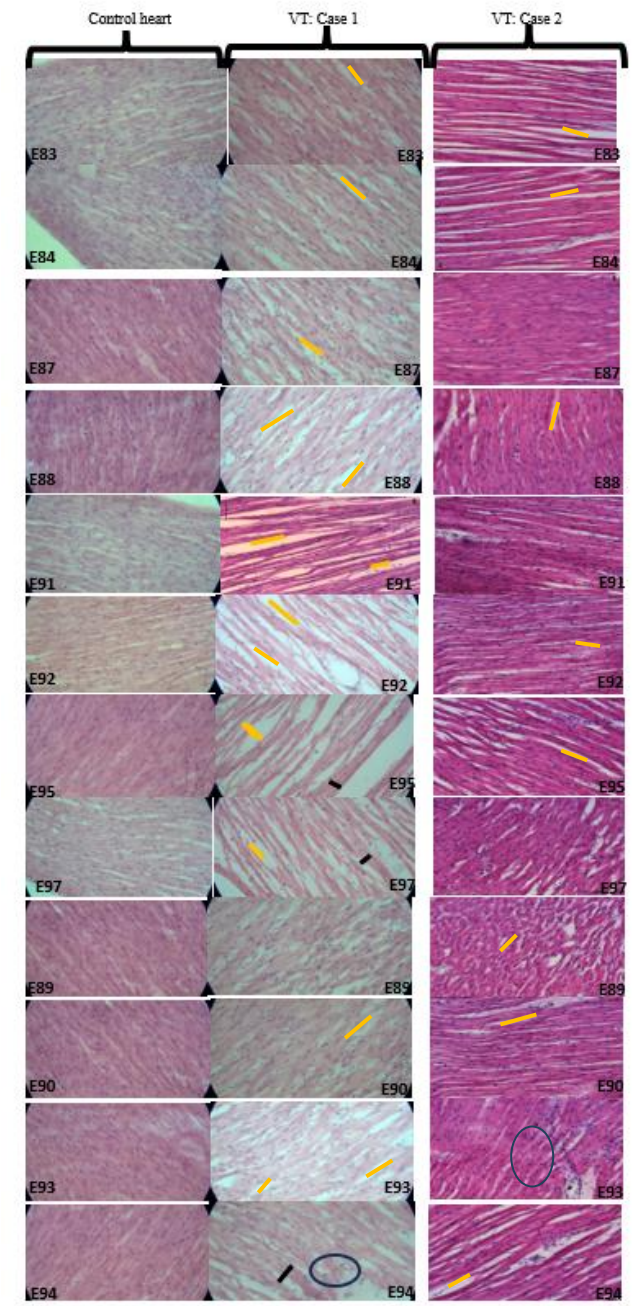
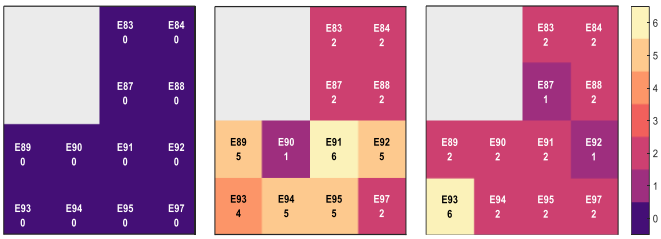


Fig. 5. Histological imaging (Hematoxylin Eosin 20x; scale bar 500 µm) of the mapped tissue area of each electrode in control heart and in the hearts under ventricular tachycardias (VT) cases. Hypertrophy fiber (black bar), intercellular space (yellow bar), Clusters of nuclei (blue circle).

After scoring each histological parameter in each electrode mapped tissue area (12 mapped electrodes area per heart = 36 tissue samples) in control heart and VT cases hearts, the sum of the score per electrode lead to generate the map depicted on fig.6. The total histology score in the control heart remained zero, whereas in the induced VT hearts it ranged from one to six. This range implies that the structural changes detected during VT were not equally severe across different tissue segments. A score of one represents minor alterations, whereas a score of six indicates more disruption of tissue architecture. This demonstrated the degree and distribution of non-fibrotic remodelling in the ventricular segments investigated during VT.



SR: Control heart heart VT: Case 1 VT: Case 2

Fig. 6. Ventricle MEA-Total histological score per electrodes (12 per condition) of the studied histological biomarkers in control and VT cases. Each colour indicated different score.

The pooled dataset (Control, VT1, and VT2 hearts; n = 12 electrode sites per heart) revealed a weak-to-moderate negative correlation between histological score and voltage amplitude (Spearman's ρ = -0.39), although this relationship did not reach statistical significance (p = 0.87).

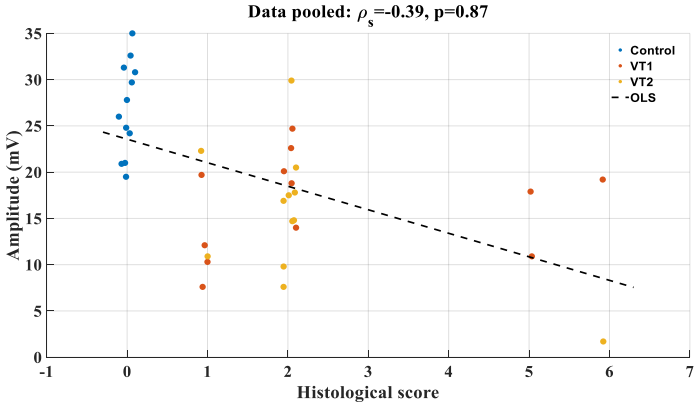


Fig. 7. Scatter plot graph of the histological score vs voltage amplitude in different condition.

The data in the table below show varied patterns among the electrodes examined in hearts subjected to induced VT. Specifically, some electrodes in the VT condition showed normal voltage values while having a total histology score that exceeded the normal reference range. Other electrodes showed abnormally low voltage values while also having total histology scores that are above the normal range.

TABLE II: Summary of histological scores for normal versus low peak-to peak voltage of the electrodes across VT cases.

VT Case	Voltage Status	Electrodes (n)	Total histological Score per electrode
1	Normal	7 (E83,E84,E87,E88,E90,E91,E95)	2, 2, 2, 2, 1, 6, 5
1	Low	5 (E89,E92,E93,E94,E97)	5, 1, 2, 1, 1
2	Normal	6 (E88,E89,E90,E91,E92,E97)	2, 2, 2, 2, 1, 2
2	Low	6 (E83,E84,E87,E93,E94,E95)	2, 2, 1, 6, 2, 2

IV. DISCUSSSION

The successful induction of sustained ventricular tachycardia (VT) in Langendorff-perfused rabbit hearts, combined with reproducible, simultaneous electrical and optical recordings obtained using the MEA system (Figs. 1 and 2), aligns with established rabbit and rodent models of experimental arrhythmogenesis [4]. The lack of a significant difference in heart weight before versus after the experiments may indicate that Langendorff perfusion did not promote an obvious myocardial water absorption or induce appreciable non-fibrotic structural remodeling (Table I). This interpretation is further supported by the preserved peak-to-peak voltage and the minimal histological score (0) observed in the control heart (Figs. 3–6; Table II).

In contrast, multiple electrodes in the VT group exhibited reduced peak-to-peak voltage accompanied by higher histological scores relative to the control, thereby demonstrating that voltage mapping can discriminate between electrophysiologically and structurally normal myocardium and arrhythmic substrate (Figs. 3, 4, 6, and 7). Notably, none of the mapped electrodes in the VT hearts showed histological scores comparable to those in the control heart, indicating persistent structural alterations following arrhythmia induction. These data further suggest that the pacing and VT induction protocols contributed to non-fibrotic structural remodeling. Although remodeling was evident in all VT cases, its distribution was regionally heterogeneous across the mapped tissue, consistent with prior multimodality investigations that have demonstrated spatial heterogeneity of structural and functional remodeling within the VT substrate [20,21].

Joint analysis of the site-matched histological and voltage data from the mapped electrodes (n = 12 per condition) revealed heterogeneous associations between tissue architecture and local electrical activity (Table II). Several electrodes displayed preserved peak-to-peak voltage despite elevated histological scores, indicating that structural remodeling does not uniformly manifest as a reduction in measured voltage. Comparable findings have been described in prior studies examining VT substrates and arrhythmogenic remodeling [21–23]. Conversely, tissue regions exhibiting both low voltage and high histological scores likely correspond to severely compromised myocardium with advanced remodeling. However, some recording sites displayed reduced voltage in the presence of only mild histological abnormalities, suggesting that electrophysiological impairment may precede overt structural remodeling or may reflect microstructural changes not captured by conventional histological evaluation.

Collectively, these observations corroborate earlier reports on the limited specificity of fixed voltage thresholds for identifying critical arrhythmogenic substrate [21–23]. Overall, this preliminary study supports the concept that structural remodeling and electrophysiological abnormalities are closely interrelated yet not uniformly coupled within the VT substrate, particularly in the context of post–myocardial infarction ventricular tachycardia [3,20,21]. The integration of simultaneous electrical and optical mapping with sites-matched histological analysis establishes a robust experimental framework for elucidating the mechanisms underlying arrhythmia initiation and maintenance, and provides mechanistic insight that is not attainable from clinical mapping studies alone.

V. CONCLUSION

This pilot study suggests that arrhythmogenesis and myocardial remodeling are closely related, but not always in a direct correlation. Using voltage mapping, as well as electrode-matched histology, we discovered that structural alterations can be arrhythmogenic or electrically quiet.

Furthermore, certain parts of tissue with aberrant voltage appeared structurally close to normal. On the other hand, not all low-voltage sites cause or maintain arrhythmia. The proposed spatial correlation pipeline could aid in understanding disease mechanisms and revising the criteria for ablation therapy targets.

ACKNOWLEDGMENT

We thank TWAS-FAPESP (grant no. 2024/08071-9) for financial support and UFABC Brazil for the research environment. This study is supported by grant no. 2018/25606-2, São Paulo Research Foundation (FAPESP) and CNPq: National Institute of Photonics (INTERAS, Grant: 409174/2024-6.

REFERENCES

[1] C. D. Cantwell, Y. Mohamied, K. N. Tzortzis, S. Garasto, C. Houston, R. A. Chowdhury, F. S. Ng, A. A. Bharath, and N. S. Peters, "Rethinking multiscale cardiac electrophysiology with machine learning and predictive modelling," *Computers in Biology and Medicine*, vol. 104, pp. 339–351, 2019. doi: 10.1016/j.compbiomed.2018.10.015.

[2] J. X. Quah, D. Dharmaprani, K. Tiver, A. Lahiri, T. Hecker, R. Perry, J. B. Selvanayagam, M. X. Joseph, A. McGavigan, and A. Ganesan, "Atrial fibrosis and substrate based characterization in atrial fibrillation: Time to move forwards," *Journal of Cardiovascular Electrophysiology*, vol. 32, no. 4, pp. 1147–1160, 2021. doi: 10.1111/jce.14987.

[3] J. Stoks, B. J. Hermans, B. J. Boukens, R. J. Holtackers, S. Gommers, Y. S. Kaya, K. Vernooy, M. J. Cluitmans, P. G. Volders, and R. M. Ter Bekke, "High-resolution structural-functional substrate-trigger characterization: Future roadmap for catheter ablation of ventricular tachycardia," *Frontiers in Cardiovascular Medicine*, vol. 10, p. 1112980, 2023. doi: 10.3389/fcvm.2023.1112980.

[4] J. Siles, V. Silva, T. Neves, I. Sandoval, A. Quadros, G. Weber, Ó. Barquero, I. Uzelac, and J. Salinet, "An integrated platform for 2-D and 3-D optical and electrical mapping of arrhythmias in Langendorff-perfused rabbit hearts," *The Journal of Physiology*, 2025. doi: 10.1113/jp287815.

[5] S. Ullah, G. Critelli, A. Quadros, J. Paredes, J. Gomes, and J. Salinet, "Mechanistic characterization of cardiac arrhythmia from electrical and optical mapping in Langendorff-perfused rabbit hearts," in *Computing in Cardiology Conference (CinC)*, 2025. doi: 10.22489/cinc.2025.095.

[6] G. S. Ramlugun, K. Kulkarni, N. Pallares-Lupon, B. J. Boukens, I. R. Efimov, E. J. Vigmond, O. Bernus, and R. D. Walton, "A comprehensive framework for evaluation of high pacing frequency and arrhythmic optical mapping signals," *Frontiers in Physiology*, vol. 14, p. 734356, 2023. doi: 10.3389/fphys.2023.734356.

[7] G. Gilbert, M. Youness, B. Thienpont, R. Willems, H. L. Roderick, P. Claus, and K. R. Sipido, "Heterogeneity of repolarization and cell-cell variability of cardiomyocyte remodeling within the myocardial infarction border zone contribute to arrhythmia susceptibility," *Circulation: Arrhythmia and Electrophysiology*, 2023. doi: 10.1161/circep.122.011677.

[8] C. Zhu, P. S. Rajendran, P. Hanna, I. R. Efimov, G. Salama, C. C. Fowlkes, and K. Shivkumar, "High-resolution structure-function mapping of intact hearts reveals altered sympathetic control of infarct border zones," *JCI Insight*, vol. 7, no. 3, p. e153913, 2022. doi: 10.1172/jci.insight.153913.

[9] C. Kuppe, R. Flores, Z. Li, S. Hayat, R. Levinson, X. Liao, M. Hannani, J. Tanevski, F. Wünnemann, J. Nagai, M. Halder, D. Schumacher, S. Menzel, G. Schäfer, K. Hoeft, M. Cheng, S. Ziegler, X. Zhang, F. Peisker, N. Kaesler, T. Saritas, Y. Xu, A. Kassner, J. Gummert, M. Morshuis, J. Amrute, R. Veltrop, P. Boor, K. Klingel, L. Van Laake, A. Vink, R. Hoogenboezem, E. Bindels, L. Schurgers, S. Sattler, D. Schapiro, R. Schneider, K. Lavine, H. Milting, I. Costa, J. Sáez-Rodríguez, and R. Kramann, "Spatial multi-omic map of human myocardial infarction," *Nature*, vol. 608, pp. 766–777, 2022. doi: 10.1038/s41586-022-05060-x.

[10] N. Trayanova, A. Lyon, J. Shade, and J. Heijman, "Computational modeling of cardiac electrophysiology and arrhythmogenesis: toward clinical translation," *Physiological Reviews*, vol. 104, pp. 1265–1333, 2023. doi: 10.1152/physrev.00017.2023.

[11] F. Giardini, C. Olianti, G. Marchal, F. Campos, V. Romanelli, J. Steyer, J. Madl, R. Piersanti, G. Arecchi, P. Vanaja, V. Biasci, E. Rog-Zielinska, G. Nesi, L. Loew, E. Cerbai, S. Chelko, F. Regazzoni, A. Loewe, M. Bishop, M. Mongillo, P. Kohl, T. Zaglia, C. Zgierski-Johnston, and L. Sacconi, "Correlative imaging integrates electrophysiology with three-dimensional murine heart reconstruction to reveal electrical coupling between cell types," *Nature Cardiovascular Research*, vol. 4, pp. 1466–1486, 2025. doi: 10.1038/s44161-025-00728-9.

[12] O. Baines, R. Sha, M. Kalla, A. Holmes, I. Efimov, D. Pavlovic, and C. O'Shea, "Optical mapping and optogenetics in cardiac electrophysiology research and therapy: a state-of-the-art review," *Europace*, vol. 26, 2024. doi: 10.1093/europace/euae017.

[13] C. O'Shea, J. Winter, S. Kabir, M. O'Reilly, S. Wells, O. Baines, L. Sommerfeld, J. Correia, M. Lei, P. Kirchhof, A. Holmes, L. Fabritz, K. Rajpoot, and D. Pavlovic, "High resolution optical mapping of cardiac electrophysiology in pre-clinical models," *Scientific Data*, vol. 9, 2022. doi: 10.1038/s41597-022-01253-1.

[14] Y. H. Kim, S. A. Chen, S. Ernst, C. E. Guzman, S. Han, Z. Kalarus, and M. K. Stiles, "2019 APHRS expert consensus statement on three-dimensional mapping systems for tachycardia developed in collaboration with HRS, EHRA, and LAHRS," *Journal of Arrhythmia*, vol. 36, no. 2, pp. 215–270, 2020. doi: 10.1002/joa3.12326.

[15] J. C. G. Junior, J. G. S. Paredes, T. C. Neves, A. D. S. de Quadros, S. Ullah., V. de Paula Silva, and J. L. S. Júnior, "Study of Peak-to-Peak Amplitude Distribution in Cardiac Rhythms: Electrophysiological Mapping in Langendorff-Perfused Rabbit Hearts," in *Computing in Cardiology Conference (CinC)*, 2025. doi:10.22489/CinC.2025.273.

[16] F. Zoppo, G. Gagno, L. Perazza, A. Cocciolo, G. Mugnai, D. Vaccari, and V. Calzolari, "Electroanatomic voltage mapping for tissue characterization beyond arrhythmia definition: a systematic review," *Pacing and Clinical Electrophysiology*, vol. 44, no. 8, pp. 1432–1448, 2021. doi: 10.1111/pace.14312.

[17] M. Z. K. A. Raimi, G. Hussain, N. Zafeer, F. Ahmad, M. Irfan, A. Ullah, and A. Shabbir, "Honey and black seed synergistically promote regeneration of oligodendrocytes in cuprizone intoxicated quail brain," *Pakistan Journal of Zoology*, vol. 56, no. 3, pp. 1163, 2024. doi: 10.17582/journal.pjz/20221107121103.

[18] M. Z. K. A. Raimi, S. Ullah, P. P. Martins, A. J. J. Gerber, F. F. Ruiz, M. S. C. Ramos, and J. S. Takada, "Structural remodeling analysis of induced arrhythmias in rabbit heart," in *2025 Computing in Cardiology Conference (CinC)*, 2025. doi: 10.22489/CinC.2025.092.

[19] Y. Takahashi, T. Yamaguchi, T. Otsubo, K. Nakashima, K. Shinzato, R. Osako, S. Shichida, Y. Kawano, A. Fukui, A. Kawaguchi, S. Aishima, T. Saito, N. Takahashi, and K. Node, "Histological validation of atrial structural remodelling in patients with atrial fibrillation," *European Heart Journal*, vol. 44, pp. 3339–3353, 2023. doi: 10.1093/eurheartj/ehad396.

[20] H. Imanli, K. Ume, J. Jeudy, T. Bob-Manuel, M. Smith, W. Chen, M. Abdulghani, Y. Ghzally, J. Mahat, R. Itah, A. Restrepo, V. See, S. Shorofsky, V. Dilsizian, and T. Dickfeld, "Ventricular tachycardia (VT) substrate characteristics: insights from multimodality structural and functional imaging of the VT substrate using cardiac MRI scar, 123I-metaiodobenzylguanidine SPECT innervation, and bipolar voltage," *Journal of Nuclear Medicine*, vol. 60, no. 1, pp. 79–85, 2018. doi: 10.2967/jnumed.118.211698.

[21] M. Josephson and E. Anter, "Substrate mapping for ventricular tachycardia: assumptions and misconceptions," *JACC: Clinical Electrophysiology*, vol. 1, no. 5, pp. 341–352, 2015. doi: 10.1016/j.jacep.2015.09.001.

[22] J. Donahue, J. Chrispin, and O. Ajijola, "Mechanism of ventricular tachycardia occurring in chronic myocardial infarction scar," *Circulation Research*, vol. 134, no. 3, pp. 328–342, 2024. doi: 10.1161/circresaha.123.321553.

[23] C. Glashan, A. Androulakis, Q. Tao, R. Glashan, L. Wisse, M. Ebert, M. De Ruiter, B. Van Meer, C. Brouwer, O. Dekkers, D. Pijnappels, J. De Bakker, M. De Riva, S. Piers, and K. Zeppenfeld, "Whole human heart histology to validate electroanatomical voltage mapping in patients with non-ischaemic cardiomyopathy and ventricular tachycardia," *European Heart Journal*, vol. 39, no. 29, pp. 2867–2875, 2018. doi: 10.1093/eurheartj/ehy168.

Estimation of membrane ion currents using an extended GHK equation which includes the ion-pumps activity

Marcio Nassif Maluf
Electrical and Computer Engineering (CPGEI)
Federal University of Technology – (UTFPR)
Curitiba, Paraná, Brasil
marcionmaluf@gmail.com

Paulo José Abatti
Electrical and Computer Engineering (CPGEI)
Federal University of Technology – (UTFPR)
Curitiba, Paraná, Brasil
pjabatti@gmail.com

***Abstract*—The electric potential across the cell membrane, known as the resting potential under quiescent conditions, is fundamental to the transport of nutrients, water, and salts to and from the cell. The GHK equation describes membrane electric potential as a function of the involved ions concentrations and respective permeabilities. However, GHK equation does not account for the active ion transport (ion pump). In other words, GHK equation does not explain the stability of intracellular ion concentrations over time. Recently, an extended GHK equation, including terms describing the ion pumps, has been proposed. The aim of this communication is to discuss the ionic currents through the cell membrane, including those due to the ion pumps, using this novel mathematical model. Because the developed model is based on the ion currents, a simplified version of it, valid for potassium and sodium ions, is presented and evaluated using experimental data available in the literature. A brief comment on energy consumption by the ion pumps is also provided.**

***Index Terms*—membrane potential, resting potential, ion pumps, GHK equation**

I. Introduction

The electrical potential difference across the cell membrane, known as the resting potential under quiescent conditions, is fundamental to solute and water transport, embryonic development, and cancer, among other processes [1]. Membrane potential (V) is classically attributed to the concentration difference of some ionic species across the cell membrane [2]-[4]. By applying Fick's first law and Ohm's law, and assuming, a constant electric field across the cell membrane, a relationship between V and the concentrations of the ion involved, each weighted by its respective membrane permeability, can be derived [5],[6]. This relationship is known as the Goldman-Hodgkin-Katz (GHK) equation and, for a cell membrane permeable to potassium, sodium, and chloride ions, it can be written as

$$V = \frac{kT}{q} \ln \frac{P_K[K]_o + P_{Na}[Na]_o + P_{Cl}[Cl]_i}{P_K[K]_i + P_{Na}[Na]_i + P_{Cl}[Cl]_o}, \qquad (1)$$

where k is the Boltzmann constant, T the absolute temperature, q the elementary electrical charge, [K] the potassium ion concentration, [Na] the sodium ion concentration, and [Cl] the chloride ion concentration, with the subscripts "o" and "i" indicating that the concentrations are in the external and internal media of the cell, respectively, and P_K the membrane permeability to the potassium ion, P_{Na} the membrane permeability to the sodium ion, and P_{Cl} the membrane permeability to the chloride ion. It should be noted that if an ion species is in equilibrium across the membrane, V can be calculated using the Nernst equation applied to this ion [2],[3]. For example, for chloride ion, the Nernst equation can be written as

$$V_{Cl_{NERNST}} = \frac{kT}{q} ln \frac{[Cl]_i}{[Cl]_o}, \qquad (2)$$

Whenever the membrane potential calculated using GHK coincides with the Nernst potential for one of the ions involved, in this example $V_{Cl_{NERNST}} = V$, it is easy to demonstrate that GHK can be simplified to

$$V = \frac{kT}{q} \ln \frac{P_K[K]_o + P_{Na}[Na]_o}{P_K[K]_i + P_{Na}[Na]_i}, \qquad (3)$$

The Nernst equation describes a situation of passive thermodynamic equilibrium [7]. Thus, it can be concluded that in this example, chloride ion does not require active ion transport across the membrane. On the other hand, potassium and sodium ions, which are not in passive thermodynamic equilibrium, require specific "ion pumps."

From the perspective of the mathematical model, it remains a matter of debate whether the relationship between V and ion concentrations is influenced indirectly by the action of ion pumps maintaining the stability of ionic concentrations over time or directly [8]-[13]. If the influence is direct, it can be stated that equation (3) is incomplete, as it lacks a specific term related to the action of ion pumps.

One experimental approach to evaluating (3) is to modify gradually one of the concentrations while observing the resulting effect on V. The literature contains studies that follow this methodology. For instance, the relationship between V and $[K]_o$ with the other concentrations held constant was studied in detail, showing that in certain situations there are discrepancies between (3) and the experimental values [14]. Thus, the direct influence of the ion pump on V was clearly

demonstrated; specially when temperature was reduced or ouabain was applied, both known methods of blocking ion-pump activity, because (3) can, this way, accurately account for the potential V across a wide range of $[K]_o$ values.

The derivation of an equation relating the concentrations of ions permeable to the membrane and V and explicitly incorporating terms describing the action of ion pumps was recently presented (work submitted to publication) [15]. This equation has a generic format as it can be applied to numerous ions permeable to the membrane simultaneously, including those with valence with modulus greater than one. In addition, several properties of the developed model had been discussed, such as its correlation with Donnan equilibrium and Nernst equation. However, the implications of this new mathematical model for interpreting ionic currents have not yet been discussed.

The aim of this work is to discuss the ionic currents through the cell membrane, including those due to ion pumps, using a novel mathematical model which relates the concentrations of ions permeable to the membrane and V, and explicitly incorporates terms describing the action of ion pumps [15]. Because the developed model is based on, both passive and active, ion currents, a simplified form of it (valid for potassium and sodium ions) is presented and compared with experimental data available in the literature [14]. As active transport of ion requires energy, a brief comment on energy consumption by the ion pumps is also provided. To facilitate the communication reading, a list of mathematical symbols used, and respective dimensions, is appended.

II. Development

Figure 1 schematically illustrates a typical cell with a membrane permeable to potassium and sodium ions represented by their concentrations $[K]$ and $[Na]$, where the subscripts "*o*" and "*i*" indicate the concentrations in the external and internal cellular environments, respectively. The x-axis indicates directions perpendicular to the cell membrane surface. It is known that these ion species cross the membrane only through specific channels [17]-[20]. Thus, it can be stated that the number of open channels determines the surface area through which each ion species crosses the membrane. Consequently, it is not possible to compare the ion fluxes (ionic current per unit area) of potassium and sodium; instead, it is necessary to analyze the absolute ion currents.

In Figure 1, I_C, I_D, I_E, and I_P represent the currents, associated with membrane equivalent capacitance (*C*), the ion concentration gradient, the electric potential gradient, and ion pumps, respectively, with the subscripts K, Na, and Cl indicating that the currents are those of potassium, sodium, and chloride ions, respectively.

From a conceptual point of view, I_C, I_D, and I_E can be considered passive currents, as they do not directly consume energy, whereas I_P can be classified as an active current because it directly consumes energy (ATP) directly during ion transport [13].

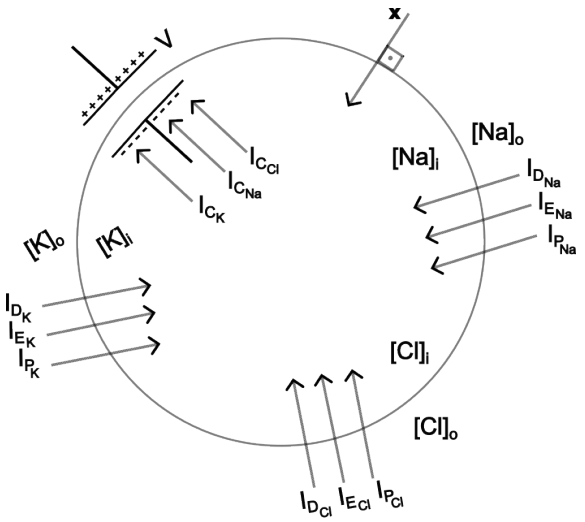


Fig. 1. Schematic representation of a typical cell indicating ion currents due to membrane capacitance (I_C), the ion concentration gradient (I_D), electric potential (I_E), and the ion pump (I_P). The x-axis indicates directions perpendicular to the cell membrane.

A. Potassium currents

The intracellular compartment is generally much smaller than the extracellular one, therefore, $[K]_o$ and $[Na]_o$ can be considered constant [13]. Consequently, for $[K]_i$ and $[Na]_i$ to remain constant over time, the total membrane currents (I_{T_K} and $I_{T_{Na}}$) must be zero.

Thus, for the potassium ion, it is possible to write

$$I_{T_K} = I_{C_K} + I_{D_K} + I_{E_K} + I_{P_K} = 0 \tag{4}$$

The capacitive current is given by

$$I_{C_K} = C\frac{dV}{dt}, \tag{5}$$

where *t* is the time.

The ion current due to the ion concentration gradient can be calculated using Fick's first law, and is given by

$$I_{D_K} = -A_K D\left(\frac{d[K]}{dx}\right), \tag{6}$$

where *D* is the diffusion coefficient, and A_K is the total area, through which K ions circulate passively.

The ion current due to the electric field (ε) can be determined using Ohm's Law, and can be represented as

$$I_{E_K} = A_K \mu_K [K]\varepsilon, \tag{7}$$

where μ_K is the ion mobility, while (ε) is given by

$$\varepsilon = -\frac{dV}{dx}. \tag{8}$$

Ion pumps can have a structure similar to that of selective ion channels, that is, they can be activated or deactivated by external factors such as temperature and the action of ion-pump blockers [21],[22]. According to [13],[23],[24], the

current density resulting from the ion pump is a function of *V*. However, it is known that ε directly influences I_{E_K} and the membrane characteristics (ε can influence the number of simultaneously opened potassium and sodium channels) that determine both I_{E_K} and I_{D_K} [25],[26]. Thus, assuming that ion concentrations remain constant for any value of *V*, it can be concluded that I_{P_K} must be a function of ε. In this work, it will be assumed that I_{P_K} is a linear function of ε. Therefore, using (8), one can write

$$I_{P_K} = -\alpha_k \frac{dV}{dx}, \tag{9}$$

where, α_k is a function of the other factors influencing I_{P_k}, such as $[K]_o$, $[K]_i$, $[Na]_o$, $[Na]_i$, blockers and *T*.

B. Potassium and sodium equilibrium electric potential

Substituting (6), (7), and (9) in (4) and considering that the electric potential is constant that is, $I_{C_K} = 0$ one obtain

$$-A_k D \frac{dK}{dx} - A_k \mu_k [K] \frac{dV}{dx} - \alpha_k \frac{dV}{dx} = 0. \tag{10}$$

Knowing that Einstein's relationship is given by $\frac{D}{\mu} = \frac{kT}{q}$ and that α_K is not a function of x, integrating from the outer side to the inner side of the cell, after manipulation (10) can be written as

$$V_K = \frac{kT}{q} ln \left(\frac{\alpha_K + A_K \mu_K [K]_o}{\alpha_K + A_K \mu_K [K]_i} \right), \tag{11}$$

where, the logarithmic term cannot be negative, that is

$$\left(\frac{\alpha_K + A_K \mu_K [K]_o}{\alpha_K + A_K \mu_K [K]_i} \right) > 0. \tag{12}$$

It is observed that, respecting (12), α_K can assume any value, including the null value, i.e., $I_{P_K} = 0$, for which from (11) it follows that

$$V_{K_{NERNST}} = \frac{kT}{q} ln \frac{[K]_o}{[K]_i}, \tag{13}$$

which is the Nernst equation applied to the potassium [2],[3].

Sodium, like potassium, is a monovalent positive ion. Therefore, for sodium ion the development is analogous to that for potassium ion. Thus, (11) and (12) can be rewritten replacing the potassium ion for the sodium ion providing

$$V_{Na} = \frac{kT}{q} ln \left(\frac{\alpha_{Na+} A_{Na} \mu_{Na} [Na]_o}{\alpha_{Na} + A_{Na} \mu_{Na} [Na]_i} \right), \tag{14}$$

and

$$\left(\frac{\alpha_{Na} + A_{Na} \mu_{Na} [Na]_o}{\alpha_{Na} + A_{Na} \mu_{Na} [Na]_i} \right) > 0. \tag{15}$$

Similarly to potassium, provided that (15) is respected, α_{Na} can to assume any value, including the null value, i.e., $I_{P_{Na}} = 0$, for which from (14) it follows that

$$V_{Na_{NERNST}} = -\frac{kT}{q} ln \frac{[Na]_o}{[Na]_i}, \tag{16}$$

which is the Nernst equation applied to the sodium [2],[3].

C. Calculation of the membrane electric potential

For the cell with membrane potential *V* necessarily we have

$$V_K = V_{Na} = V. \tag{17}$$

Finally , using (11), (14) and (17) together with the identity

$$y = K\ ln \frac{a}{b} = K\ ln \frac{c}{d} = K\ ln \frac{a+c}{b+d} \tag{18}$$

after manipulation, one obtains

$$V = \frac{kT}{q} \ln \frac{A_K \mu_K [K]_o + A_{Na} \mu_{Na} [Na]_o + \alpha_K + \alpha_{Na}}{A_K \mu_K [K]_i + A_{Na} \mu_{Na} [Na]_i + \alpha_K + \alpha_{Na}}. \tag{19}$$

It is observed that in this case doing $P = A \frac{kT}{q} \frac{\mu}{w}$, where *w* is the thickness of the membrane (*P* was originally defined as $P = \frac{kT}{q} \frac{\mu}{w}$ [3]), (19) becomes (3), or in other words, GHK applied to the sodium and to potassium ions is recovered for $\alpha_K = \alpha_{Na} = 0$. In addition, knowing that, in quiescent conditions, the membrane permeability to potassium ion is much greater than that to the sodium ion ($P_K \gg P_{Na}$) and that in intracellular media the sodium ion concentration is much lower than potassium ion concentration ($[K]_i \gg [Na]_i$) [17], (19) can be rewritten in a simplified form

$$V = \frac{kT}{q} ln \left(\frac{[K]_o + \frac{A_{Na} \mu_{Na}}{A_K \mu_K} [Na]_o + \frac{\alpha_{Na} + \alpha_K}{A_K \mu_K}}{[K]_i + \frac{\alpha_{Na} + \alpha_K}{A_K \mu_K}} \right). \tag{20}$$

III. EXPERIMENTAL MODEL AND RESULTS

To test the validity of (20) one can follow the classic methodology which consists, as already mentioned, in to obtain the relationship between V and $[K]_o$. However, first it is necessary to determine $\frac{\alpha_{Na} + \alpha_K}{A_K \mu_K}$. For this purpose, it was used the experimental data from the work of [14] obtained using the cell of the gastroesophageal ganglion of the marine mollusk *Anisodoris Nobilis*. Among the reasons for this choice, one can mention the differences between the theoretical, using GHK, and the practical results, experimental data obtained under various temperatures and under conditions in which the ion pumps cease to operate, such as low temperatures and the action of blockers [14]. Unfortunately, the experimental data are not presented in tabular form [14]. Thus, it was necessary import the figures into the AutoCAD application, allowing, using ID POINT function, to construct a table $[K]_o$ versus V. Note that it is not available a table $\frac{\alpha_{Na} + \alpha_K}{A_K \mu_K}$ versus $[K]_o$, which would allow to use conventional methods of curve fitting. However, using equation (20), it is known that $\frac{\alpha_{Na} + \alpha_K}{A_K \mu_K}$ is a function of $[K]_0$, it is not a function of V, and must have the dimensions of concentration. Moreover, figure (2) shows a point of minimum, and to increase in modulus a logarithmic function (20), the first choice is an exponential function.

Thus, using the data from [14] $\frac{A_{Na} \mu_{Na}}{A_K \mu_K} = \frac{P_{Na}}{P_K} = 0.033$, $[Na]_o = 480.5 mM/l, [K]_i = 235 mM/l$ and $T \approx 290K (\approx 17°C)$, after trial and error, it was observed that the experimental data were better adjusted using following expression

$$\frac{\alpha_K + \alpha_{Na}}{A_K \mu_K} = [K]_R \left[-p e^{-d \left(\sqrt{\frac{[K]_o}{[K]_R}} - 1 \right)^4} + s \right], \tag{21}$$

with the parameters $[K]_R$, p, d and s, adjusted for 7.5 mM /l ; 1.49 ; 14 and 0.16 , respectively . Combining (21) with (20) gives

$$V = \frac{kT}{q} ln \left[\frac{[K]_o + \frac{A_{Na}\mu_{Na}}{A_K\mu_K}\ [Na]_o + [K]_R \left(-pe^{-d\left(\sqrt{\frac{[K]_o}{[K]_R}}-1\right)^4} + s \right)}{[K]_i + [K]_R \left(-pe^{-d\left(\sqrt{\frac{[K]_o}{[K]_R}}-1\right)^4} + s \right)} \right] \quad (22)$$

Then, (22) was evaluated using the remained experimental data from [14]. Here also were used $\frac{A_{Na}\mu_{Na}}{A_K\mu_K} = \frac{P_{Na}}{P_K} = 0.033$, $[Na]_o$=480.5 mM/l, $[K]_i$=235 mM/l for temperatures 277K($\approx 4°C$), 284K($\approx 11°C$) and 290K($\approx 17°C$), with and without the application of a ion-pump blocker, ouabain. The parameters $[K]_R$, *p*, *d* and *s*, were adjusted as shown in Table I.

Figure 2 shows experimental data from [14] together with theoretical results obtained using (22) highlighting the curve (continuous black line) used to obtain empirically (21).

TABLE I
PARAMETERS *p*, *d*, $[K]_R$ AND *s* USED TO ADJUST THE GRAPHS GENERATED USING (21).

Adjusted Parameters	**Experimental data**				
	277K	284K	290K	284K + Ouabain	284K $[Na]_o$=0
p	0.16	1.14	1.49	0	0
d	14	14	14	14	14
$[K]_R$(mM/l)	7.5	7.5	7.5	7.5	7.5
s	0.16	0.16	0.16	0.16	- 0.67

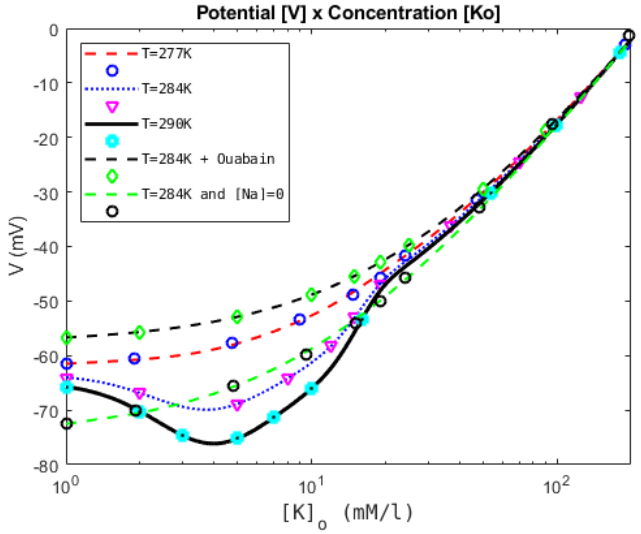


Fig. 2. Graph of *V* versus $[K]_o$ using (22) for the experimental data of [14]. The experimental conditions includes *T*=277K, *T*=284K,*T*=290K, *T*=284K with ouabain, *T*=284K with *Na*=0 and $\frac{A_{Na}\mu_{Na}}{A_K\mu_K} = \frac{P_{Na}}{P_K} = 0.033$, $[Na]_o$=480.5 mM/l, $[K]_i$=235 mM/l with $[K]_R$=7.5 mM/l. Parameters *p*,*d* and *s* were adjusted as show in Table I.

The correlation coefficient between the theoretical values using (22) and the available experimental values are greater than 0.95.

IV. Discussion and Conclusion

Based on the hypothesis that the electric current due to the ion pump is proportional to the electric field through membrane (see (9)), a mathematical model describing the relationship between V and the ion concentrations weighted by the respective permeabilities and the action of ionic pumps was presented. The correlation coefficient between the experimental data available in literature and the developed model was higher than 0.95. It is important to emphasize that the ion-pump characteristics are defined by α, which may not necessarily to be a linear function. In fact, (21) shows that $\frac{A_{Na}\mu_{Na}}{A_K\mu_K}$ is not a linear function of $[K]_o$.

Among the characteristics of the developed model, it can be observed that all the descriptive equations of the equilibrium state of involved ions, equations (11) and (14), the general equation (20), the simplified one (22) and the Nernst potential for chloride ion result in the membrane electric potential *V*. The equation to be used to determine *V* will depend on knowledge of the involved parameters. In this way, the empirical determination of the relationship $\frac{\alpha_{Na}+\alpha_K}{A_K\mu_K}$(see (21)) in the present work justifies the use of (20) or, more specifically for its simplicity, of (22). From a conceptual point of view, it is interesting to note that (11) and (14) explicitly show the difference between the equilibrium electric potential for the sodium and potassium ions (V_K=V_{Na}=V) and the respective thermodynamic equilibrium electric potential (Nernst potential) for the mentioned ions (chloride ion is in thermodynamic balance, $V_{Cl_{NERNST}}$=V). Figure 3 shows the comparison between electric potential of the membrane, which represents the equilibrium potential for all ions, and the Nernst electric potentials for sodium and potassium ions with $V_{Cl_{NERNST}}$=V, where here it was considered, as for most cells,$[K]_o < [K]_i$, $[Na]_o > [Na]_i$ and $[Cl]_o > [Cl]_i$, with the membrane permeability to sodium ion being much lower than that to potassium ion.

Note that, as demonstrated for sodium and potassium ions, in case of α for a given ion to be equal to zero, necessarily the membrane and Nernst electric potentials for this ion will equal to each other, see (13) and (16). In other words, for this ion I_P=0. Therefore, it is possible to conclude that for a membrane permeable to the sodium, potassium, and chloride ions, with $V_{Cl_{NERNST}}$=V, $I_{P_{Cl}}$ is equal to zero, while I_{P_K} and $I_{P_{Na}}$ are different from zero. In fact, several studies indicate that $V_{Cl_{NERNST}}$ is equal or nearly equal to *V* [7].

Unfortunately, in the present study, it is not possible to determine the absolute value of the electric current due to the ion pump. However, the current due to the ion concentration gradient I_D, for a given membrane electric potential *V*, can be considered constant. In other words, I_D can be taken as a reference.

The directions of I_E and I_D can be easily determined by observing the differences in ion concentrations between the

internal and external cell's media and the polarity of *V*. Thus, remembering that according to (4) the sum of the currents must be zero, and knowing that I_C=0, the value of $I_E + I_P$ can be determined, allowing also to determine the direction of I_P, as the direction of I_E is known. Figure 4 illustrates the currents due to the ionic concentration gradient I_D , to electric potential gradient I_E and to ion pumps I_P, for sodium, potassium and chloride ions.

The fact that, for a given ion, the sum $I_E + I_P$ being directly related to I_D may be important for understanding membrane energy consumption due to ion pumps. According to (22), the membrane potential has a value between the Nernst potential for sodium and for potassium (see figure 3). As $P_K \gg P_{Na}$, *V* tends to have a value closer to the Nernst potential for potassium, therefore, I_{E_K} has opposite direction and a lower value than that of I_{D_K}. Thus, based on the presented arguments, the current due to the potassium ion pump I_{P_K} has a relatively small value (see Figure 4). For sodium ion, $I_{E_{Na}}$ and $I_{D_{Na}}$ have the same direction, therefore the current due to the sodium ion pump $I_{P_{Na}}$ has a relatively high value. However, remembering that $P_K \gg P_{Na}$, the value of $I_{D_{Na}}$ must certainly be much smaller than I_{D_K}. This way, the values of I_{P_K} and $I_{P_{Na}}$ should not be very different from each other. In fact, it can be found in the literature works demonstrating that the ratio between the sodium and potassium ion pumps is 3 to 2 ions [27]. In conclusion, the energy consumption to maintain I_{P_K} and $I_{P_{Na}}$ may be minimized provided that I_{D_K}, and $I_{D_{Na}}$ are kept at minimum values.

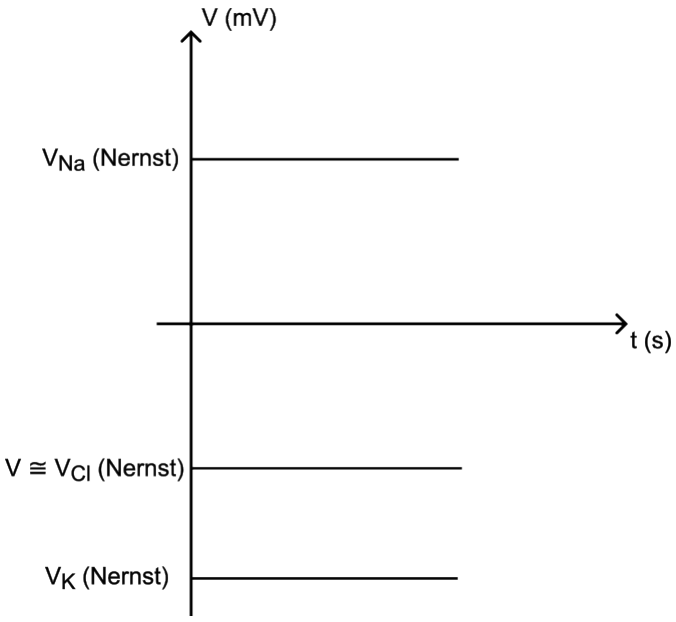


Fig. 3. Comparison between membrane electric potential (V), and K, Na and Cl Nernst's electric potentials.

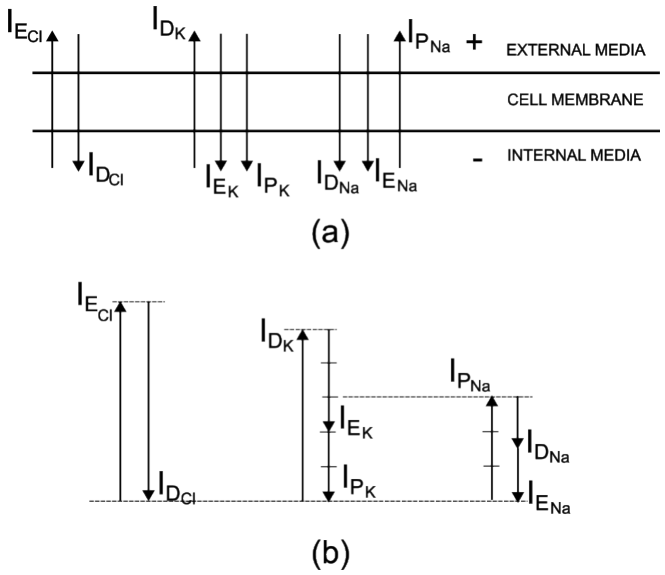


Fig. 4. Figure 4a) Ionic currents across the membrane due to the concentration gradient I_D, the electrical potential I_E, and the ion pump I_P, for chloride, potassium, and sodium ions across the membrane. b) The ratio between the sodium and potassium ion pumps is 3 to 2 ions.

New experimental data can confirm the validity of the developed mathematical model or suggest adjustments to improve it. In any case, the authors believe that developed model can be useful to understand the factors that affect the physiological/pathological conditions influenced by cell membrane electric potential, perhaps encouraging the development of automatic methods to measure *V*.

APPENDIX A
LIST OF USED MATHEMATICAL SYMBOLS AND RESPECTIVE DIMENSIONS.

Symbol	Definition	Dimension
A_K	Area of membrane through which potassium ion can cross	m^2
A_{Na}	Area of membrane through which sodium ion can cross	m^2
α_K	Potassium ion pump factor	Am/V
α_{Na}	Sodium ion pump factor	Am/V
ATP	Adenosine triphosphate	J/M
C	Membrane equivalent capacitance	F
$[Cl]$	Chloride ion concentration	M/l
D_K	Potassium ion diffusion coefficient	m^2/s
D_{Na}	Sodium ion diffusion coefficient	m^2/s
d	Fitting parameter	dimensionless
ε	Electric field	V/m
I_{TK}	Potassium ion total current	A
I_{CK}	Potassium ion capacitive current	A
I_{EK}	Potassium ion electric current	A
I_{DK}	Potassium ion diffusion current	A
I_{PK}	Potassium ion pump current	A
I_{TNa}	Sodium ion total current	A
I_{CNa}	Sodium ion capacitive current	A
I_{ENa}	Sodium ion electric current	A
I_{DNa}	Sodium ion diffusion current	A
I_{PNa}	Sodium ion pump current	A
k	Boltzmann's constant	J/K
$[K]$	Potassium ion concentration	M/l
$[K]_R$	Potassium ion reference concentration	M/l
μ_K	Potassium ion mobility	m^2/Vs
μ_{Na}	Sodium ion mobility	m^2/Vs
$[Na]$	Sodium ion concentration	M/l
p	Fitting parameter	dimensionless
P_{Cl}	Membrane permeability to chloride ion	m/s
P_K	Membrane permeability to potassium ion	m/s
P_{Na}	Membrane permeability to sodium ion	m/s
q	Elementary charge	C
s	Fitting parameter	dimensionless
t	Time	s
T	Absolute temperature	K
V	Membrane electric potential	V
x	Direction perpendicular to cell membrane	m
w	Membrane thickness	m

REFERENCES

[1] Ruhl, P., Langner, J. M., Reidel, J., Schonherr, R., Hoshi, T., Heinemann, S. T. 2021. Monitoring compound resting membrane potentials of cell cultures with radiometric genetically encoded voltage indicators. Comm. Biol. 1 – 11. https://doi.org/10.1038/s42003-021-02675-0

[2] Mountcastle, V. B., 1978. Fisiologia Médica. Guanabara Koogan. Rio de Janeiro (in portuguese). https://doi.org/10.1016/j.cub.2015.02.039

[3] Terry Bahill, A., 1981. Bioengineering: Biomedical, Medical and Clinical Engineering, Prentice-Hall, New Jersey.

[4] Hall, J.E., 2015. Guyton and Hall Textbook of Medical Physiology. E-Book, Elsevier, Amsterdam. doi Editora Elsevier Health Sciences, 2020 ISBN 0323640036, 9780323640039

[5] Goldman, D.E., 1943. Potential, impedance, and rectification in membranes. J. Gen. Physiol. 27,37 -– 60. https://doi.org/10.1085/jgp.27.1.37

[6] Hodgkin, A.L., Katz, B., 1949. The effect of sodium ions on the electrical activity of the giant axon of the squid. J. Physio. 108, 37-77.doi: 10.1113/jphysiol.1949.sp004310

[7] Russell, John M, 1984. Current Topics in Membranes and Transport. Elsevier; 22; 177–193 https://doi.org/10.1016/S0070-2161(08)60471-2

[8] Frumento, A. S., 1964, Sodium pump: its electrical effects in skeletal muscle. Science, 147, 1442–1443. DOI: 10.1126/science.147.3664.1442

[9] Moreton, R., 1969. An investigation of the electrogenic sodium pump in snail neurons using the constant-field theory. J. Experimental Biol., 51, 181—201. https://doi.org/10.1242/jeb.51.1.181

[10] Cohen, C. J. 1977. Characterization of the resting potential in chironomus salivary gland cells evidence for an electrogenic sodium pump. Exp. Cell Res., 106, 15–30.https://doi.org/10.1016/0014-4827(77)90236-1

[11] Lauger, P., 1991. Electrogenic ion pumps. Sinauer, Sunderland. https://cir.nii.ac.jp/crid/1971712334707304630

[12] Armstrong, C.M., 2003. The Na/K pump, Cl ion, and osmotic stabilization of cells. Proc. Natl. Acad. Sci., 100, 6257–6262. https://doi.org/10.1073/pnas.0931278100

[13] Jackle, J., 2007. The causal theory of the resting potential of cells. J. Theoretical Biol.,249, 445–463. https://doi.org/10.1016/j.jtbi.2007.07.027

[14] Gorman, A.L.F., Marmor, M.F., 1970. Contributions of the sodium pump and ionic gradients to the membrane potential of a molluscan neurone. J. Physiol. 210, 897-–917. 10.1113/jphysiol.1970.sp009248

[15] Maluf, M.N., Protachevicz1 P.R., Abatti1 P.J., 2026. Extending the Goldman-Hodgkin-Katz Framework to Include ion pump Mechanism. IEEE Engineering in Medicine and Biology. (submitted for publication).

[16] Marmor, M., Gorman, A., 1970, Membrane potential as the sum of ionic and metabolic Components. Science, 167, 65-–67. doi: 10.1126/science.167.3914.65

[17] Hamill, O.P., Marty, A., Neher, E., Sakmann, B., Sigworth, F.J., 1981. Improved patch-clamp technique for high-resolution current recording from cells and cell-free membrane patches. Pflugers Arch. European J. Physiol., 391, 85–100. https://doi.org/10.1007/BF00656997

[18] Mackinnon, R., Cohen, S. L., Kuo, A., Lee, A., Chait, B. T., 1998. Structural conservation in prokaryotic and eukaryotic potassium channels. Science, 280, 106–112. doi: 10.1126/science.280.5360.106

[19] Wright, S. H., 2004. Generation of resting membrane potential. Adv. Physiol. Educ., 28, 139 — 142. https://doi.org/10.1152/advan.00029.2004.

[20] Cardozo, D., 2016. An intuitive approach to understanding the resting membrane potential. Adv. Physiol. Educ., 40, 543–547. https://doi.org/10.1152/advan.00049.2016

[21] Alvarez, O., Latorre, R., 2017. The enduring legacy of the 'constant-field equation' in membrane ion transport. J. Gen. Physiol., 149, 10, 911—920 https://doi.org/10.1085/jgp.201711839

[22] Herrera-Valdez, M. A., A thermodynamic description for physiological transmembrane transport. F1000 Research. 1–30. doi:10.12688/f1000research.16169.3

[23] Rakowski, R.F., Gadsby, D.C., de Weer, P., 1989. Stoichiometry and voltage dependence of the sodium pump in voltage-clamped, internally dialyzed squid giant axon. J. Gen. Physiol. 93, 903-–941. https://doi.org/10.1085/jgp.93.5.903

[24] Lew, V.L., Bookchin, R.M., 1986. Volume, pH, and ion-content regulation in human red cells: Analysis of transient behavior with an integrated model. J. Membrane Biol. 92,57-–74. https://doi.org/10.1007/BF01869016

[25] Hodgkin, A.L., Huxley, A.F.,1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. 117,500—544.doi:10.1113/jphysiol.1952.sp004764

[26] Faber, D. S., Pereda, A. E. Two Forms of Electrical Transmission Between Neurons. Frontier in Molecular neuroscience. https://doi.org/10.3389/fnmol.2018.00427

[27] Jens Chr. Skou; The Identification of the Sodium Pump. Biosci Rep 10 August 2004; 24 (4-5): 436—451. doi: https://doi.org/10.1007/s10540-005-2740-9

Design and preliminary evaluation of a gear-based ratchet mechanism for controlled quarter-turn adjustment in Ilizarov external fixators

Macleane Ferreira Leite Monteiro
Centro de Engenharias - CE
Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró/RN,Brasil
macleane.34@gmail.com

Celso Gaona Lima Duarte
Centro de Engenharias - CE
Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró/RN,Brasil
celsogaonalimaduarte@gmail.com

Zoroastro Torres Vilar
Centro de Engenharias - CE
Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró/RN,Brasil
zoroastro@ufersa.edu.br

Diego Ariel de Lima
Departamento de Ciências da Saúde
Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró/RN,Brasil
diego.lima@ufersa.edu.br

Lana Lacerda de Lima
Departamento de Ciências da Saúde
Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró/RN,Brasil
lana.lima@ufersa.edu.br

Ramon Rudá Brito Medeiros
Departamento de Engenharias
Universidade Federal Rural do Semi-Árido (UFERSA)
Caraúbas/RN,Brasil
ramon.ruda@ufersa.edu.br

***Abstract* — Distraction osteogenesis using Ilizarov external fixators requires periodic quarter-turn adjustments to achieve controlled bone lengthening. These adjustments are typically performed manually by patients or caregivers, making the process susceptible to operational errors and variability. This study presents the design and preliminary evaluation of a mechanical ratchet mechanism capable of producing a nominal quarter-turn rotation of the adjustment nut with each actuation of the lever. The device employs a compact gear-based transmission and a unidirectional ratchet system that allows the lever to be manually returned to its initial position without reversing the adjustment. A preliminary laboratory test was conducted over 120 consecutive actuation cycles to assess the mechanical operation and cumulative linear displacement of the prototype. The mechanism completed the single 120-cycle test without critical transmission failures or operational interruptions. The theoretical cumulative displacement was 30.00 mm, while the measured displacement was 30.28 mm, corresponding to a deviation of 0.28 mm (0.93%). The findings support further development and mechanical validation of the proposed indexing concept before any assessment of clinical applicability.**

Keywords — Ilizarov Fixator, Distraction Osteogenesis, Mechanical Design, Biomedical Engineering, Assistive Device.

I. INTRODUCTION

Distraction osteogenesis is a well-established orthopedic technique for treating limb-length discrepancies, congenital deformities, bone defects, nonunions, and complex fractures. The method relies on the biological principle that the gradual, controlled mechanical distraction of two bone segments stimulates the formation of new bone tissue within the distraction gap, while simultaneously promoting the adaptation of adjacent soft tissues—including muscles, blood vessels, nerves, and skin [1], [8]. Since its introduction by Gavriil A. Ilizarov in the second half of the 20th century, the technique has become one of the most important procedures in reconstructive orthopedic surgery.

The Ilizarov external fixator remains one of the most widely adopted systems for distraction osteogenesis due to its versatility, mechanical stability, and ability to correct deformities in multiple planes [1]. During treatment, threaded rods connected by rings and fixation wires gradually separate the bone segments according to a prescribed distraction protocol. A commonly adopted regimen involves approximately 1 mm of distraction per day—typically divided into four equal increments of 0.25 mm—aimed at optimizing bone regeneration and minimizing complications [2], [3]. Consequently, patients or caregivers must perform multiple manual adjustments daily throughout the treatment, which can last several months depending on the clinical indication.

Although the mechanical principles of the Ilizarov system are well established, the adjustment procedure itself remains almost entirely dependent on manual operation. In conventional practice, patients or caregivers must manually turn each adjustment nut by a prescribed quarter-turn during every distraction cycle. This repetitive procedure requires strict adherence to the prescribed protocol and must be performed consistently throughout the entire treatment period [6], [7]. Since multiple threaded rods require independent adjustment at regular intervals, maintaining consistent angular displacement can be challenging, particularly outside the clinical setting [8], [9].

Various technological approaches have been proposed to enhance distraction osteogenesis, including computer-assisted external fixators, motorized distraction systems, and hexapod-based orthopedic frames [4], [5]. Although these systems offer advanced control capabilities, they are often associated with greater mechanical complexity, higher manufacturing costs, and specialized clinical requirements, which limit their widespread adoption in many healthcare settings. Consequently, conventional, manually adjusted Ilizarov fixators remain the standard treatment option at a large number of orthopedic centers worldwide [7].

Despite the extensive literature addressing biological responses, fixation biomechanics, and treatment outcomes, relatively few studies have focused on improving the manual adjustment mechanism itself [9], [10]. In particular, little attention has been paid to simple mechanical devices capable of assisting patients in performing standardized quarter-turn adjustments without the need for motors, electronics, or modifications to the original treatment protocol. Such solutions may provide a simple mechanical means of indexing

the adjustment movement while preserving the conventional configuration of the external fixation system [6].

The central hypothesis of this study is that a mechanically indexed ratchet device can produce nominal quarter-turn adjustment increments through a manual lever actuation while maintaining the adjustment direction during lever return. The present work therefore focuses on the preliminary mechanical evaluation of this operating concept under laboratory conditions. By introducing a mechanically simple, purely mechanical assistance mechanism, this work aims to contribute to the development of accessible technologies for mechanically indexed adjustment of distraction osteogenesis systems without increasing treatment complexity.

II. Materials And Methods

A. Device Design

The proposed device was developed to assist in the manual adjustment of threaded rods used in Ilizarov external fixation systems during distraction osteogenesis. The primary design requirement was to produce a nominal quarter-turn (90°) rotation of the adjustment nut through a single lever actuation, while maintaining the simplicity of conventional treatment protocols.

The gear transmission was designed using involute spur gears with a 20° pressure angle and a module of 1 mm. The driving gear, represented by the yellow component in Fig. 1, has 12 teeth, while the driven gear, represented by the red component, has 24 teeth. Both gears have a 7 mm face width. The resulting gear ratio is 2:1, such that the driven gear rotates at half the angular displacement of the driving gear. Therefore, a 180° rotation of the driving gear results in a nominal 90° rotation of the driven gear, corresponding to the required quarter-turn adjustment of the threaded mechanism.

The mechanical concept relied on a compact gear transmission integrated with a unidirectional ratchet mechanism. The transmission converts the rotational movement of the operating lever into rotational movement of the adjustment nut. The ratchet mechanism allows the lever to return to its initial position without reversing the rotation of the threaded rod, enabling consecutive actuations while preserving the accumulated distraction displacement. In other words, it enables the user to perform only the tightening movement in the correct direction for treatment.

The device was designed to operate independently on each threaded rod of the external fixator. Therefore, four identical units can be installed on a conventional Ilizarov frame when simultaneous adjustment of all rods is required. The design process was carried out using computer-aided design (CAD) software. During development, special attention was paid to geometric compatibility with commercially available Ilizarov fixators, ease of installation, compact dimensions, manufacturability, and user operation.

B. Mechanical Operating Principle

The operating principle of the proposed mechanism is illustrated in Figure 1.

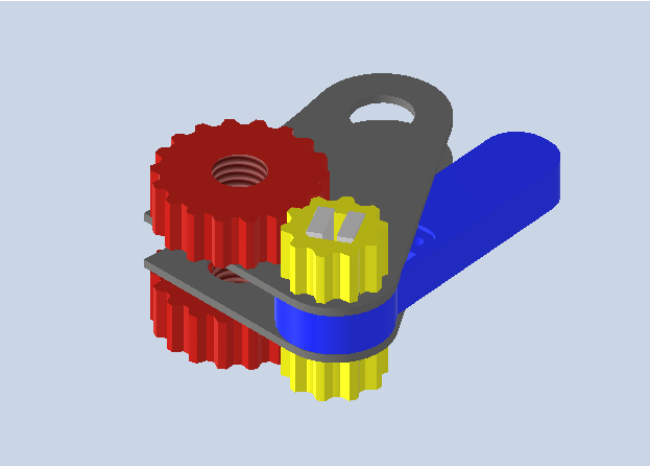


Fig. 1. System designed in CAD (original design).

When the user applies force to the operating lever, the lever rotates the yellow driving gear through a nominal 180°. The 12-tooth driving gear engages the 24-tooth red driven gear, producing a 2:1 reduction in angular displacement. Consequently, the red gear rotates by a nominal 90° for each complete lever stroke, corresponding to one quarter-turn of the threaded adjustment mechanism.

After the adjustment stroke, the lever can be manually returned to its initial position. During this return movement, the ratchet mechanism disengages the transmission, preventing reverse rotation of the adjustment mechanism while maintaining the previously achieved displacement. This allows successive quarter-turn increments to be applied through repeated lever actuation.

C. Prototype Development

A functional prototype was manufactured to evaluate the feasibility of the proposed concept. Additive manufacturing was used to produce a functional prototype for preliminary mechanical evaluation of the proposed concept. With the exception of the commercially available Ilizarov nuts and spring, the mechanism components were manufactured from ASA (Acrylonitrile Styrene Acrylate) using additive manufacturing. ASA was selected as the prototype material because of its suitability for functional polymeric components and its mechanical resistance compared with other commonly used prototyping polymers.

The complete prototype, including the polymeric components and metallic fastening elements, had an approximate mass of 20 g and overall dimensions of 5 × 3 × 2 cm. The system was designed to fit onto 6 mm threaded rods with a 1 mm pitch. Several prototype iterations were produced during the development process to optimize gear engagement, lever ergonomics, structural rigidity, and operational reliability.

D. Experimental Evaluation

Preliminary laboratory tests were conducted to evaluate the operational performance of the proposed mechanism.

1) Repeated Performance Test

A repeated actuation test was conducted to characterize the mechanical behavior of the prototype during consecutive operation. A single series of 120 consecutive actuation cycles was performed on the prototype. The test was considered a preliminary functional assessment rather than a statistical repeatability or fatigue test. Each operational cycle consisted

of a full forward actuation of the lever, followed by the lever's return. During the tests, the following parameters were observed:

a) Occurrence of mechanical failure;

b) Gear disengagement;

c) Ratchet malfunction;

d) Structural deformation;

e) Operational interruptions.

At the end of the test, the mechanism was visually inspected for wear or permanent deformation.

2) Linear Displacement Verification

The cumulative linear displacement produced by the mechanism was compared with the theoretical displacement expected based on the threaded rod's pitch. Given that each actuation of the lever corresponds to a quarter-turn of the adjustment nut, the theoretical displacement after 120 operating cycles was calculated based on the thread pitch. Cumulative linear displacement was measured at 12-cycle intervals using a caliper throughout the 120-cycle test. The measured values were compared with the theoretical displacement calculated from the 1 mm thread pitch and the nominal quarter-turn per actuation. Subsequently, the percentage error between the theoretical and measured displacements was calculated.

E. Experimental Test Bench

To enable the preliminary evaluation of the proposed mechanism under controlled laboratory conditions, a dedicated experimental test bench was designed and manufactured. The apparatus was developed to reproduce the basic threaded adjustment configuration of an Ilizarov external fixator in a simplified laboratory setup. The test bench was not intended to reproduce physiological loading conditions or the complete biomechanical behavior of a clinical Ilizarov frame.

The test bench consists of two structural frames connected by three threaded rods arranged in parallel, reproducing the threaded adjustment system commonly employed in circular external fixation devices. The proposed ratcheting mechanism was mounted on the central threaded rod, allowing repeated quarter-turn adjustments to be performed while the resulting linear displacement of the movable frame was observed.

The structural components of the test bench were fabricated by additive manufacturing using ASA thermoplastic, whereas the threaded rods and fastening elements were manufactured from steel. This configuration provided sufficient rigidity for repeated laboratory testing while enabling rapid assembly, component replacement, and geometric modifications during prototype development.

The primary purpose of the test bench was to evaluate the operational feasibility of the proposed mechanism, including its ability to produce successive indexed adjustments, maintain transmission integrity during repeated actuation cycles, and verify the cumulative linear displacement generated throughout the experimental procedure.

Figure 2 illustrates the experimental test bench developed for the preliminary validation of the proposed mechanism.

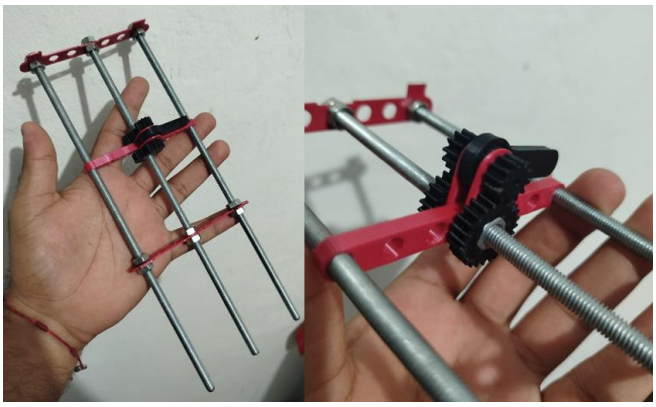


Fig. 2. Equipment similar to the Ilizarov apparatus for testing (original design).

III. Initial Prototype Results

Based on the model developed in the CAD software, it was possible to manufacture initial prototypes using 3D printing; these were the closest in dimensions and functionality to the final model produced in any other material. Following the tests and analyses outlined in the previous section, the following results were obtained.

A. Operational Performance

Figure 3 shows the initial prototype in yellow—a color that does not affect the material's properties—while Figure 4 displays the individual black components prior to their assembly into the testing equipment. In addition to the elements shown in the images, the system incorporates nuts from the commercial Ilizarov system and a spring (20 mm long, 3 mm in diameter) that actuates the ratchet mechanism. The experimental evaluation focused on the mechanical operation of the prototype during repeated actuation.



Fig. 3. Mechanism assembled without being inserted into the Ilizarov apparatus (original design).



Fig. 4. Disassembled mechanism (original design).

No critical failure of the gear train or ratchet mechanism was observed during this single test series. During the test, the gear train remained properly engaged, and no loosening of the adjustment gear mounted on the threaded shaft was observed. The ratchet mechanism also maintained its unidirectional operation, allowing the operating lever to return to its initial position without causing reverse rotation of the adjustment nut.

Given that each operational cycle produces a quarter-turn of the threaded nut, the expected theoretical linear displacement after 120 cycles was 30 mm, based on the thread pitch used in the experimental setup. The measured cumulative displacement was approximately 30.28 mm, representing a deviation of about 0.28 mm (0.93%) from the theoretical value; Table 1 summarizes the obtained data.

TABLE I. THEORETICAL AND EXPERIMENTAL CUMULATIVE DISPLACEMENT AFTER 120 ACTUATION CYCLES

Parameter	Theoretical	Experimental	Deviation
Linear advance	30mm	30.28mm	0.28mm (0.93%)

(Original design)

Figure 5 shows the theoretical and experimental cumulative displacement throughout the 120-cycle test, with measurements obtained at 12-cycle intervals. The experimental curve followed the theoretical trend over the evaluated range, reaching 30.28 mm at the end of the test compared with 30.00 mm theoretically expected. The accumulated deviation was therefore 0.28 mm (0.93%). This result indicates that the prototype maintained the expected direction and approximate magnitude of linear displacement during the evaluated test. However, because only one test series was performed, the result should not be interpreted as a statistical assessment of repeatability.

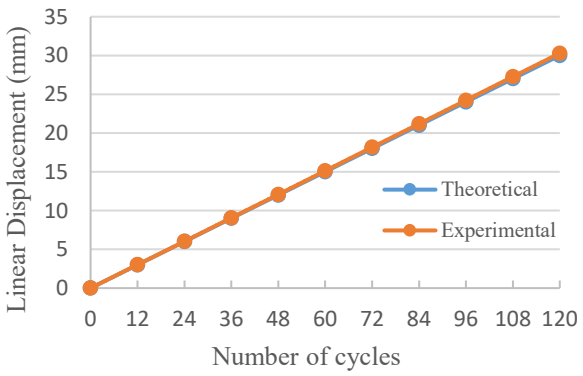


Fig. 5. Comparison between theoretical and experimental cumulative linear displacement during the 120-cycle actuation test. (original design).

Although the cumulative displacement differed slightly from the theoretical prediction, the mechanism consistently generated incremental movement throughout the test, demonstrating the feasibility of the proposed transmission concept. The observed deviation may be related to manufacturing tolerances inherent to the additive manufacturing process, dimensional variations in the printed gears, assembly clearances, and elastic deformation of the polymeric components. However, the contribution of each factor was not independently quantified in the present study and requires further investigation.

B. Structural Assessment

Visual inspection following the repeated actuation test revealed no evidence of failure in the gear transmission or the ratchet system. Likewise, the threaded adjustment gear remained securely attached to the shaft, indicating that the gear remained mechanically engaged throughout the evaluated test.

However, whitish cracks were identified on the wall surrounding the ratchet shaft mount—which is integrated into the operating lever—in the region shown in Figure 6. These cracks appeared after repeated use and suggest that this area represents the primary structural weak point of the current prototype.



Fig. 6. Area of possible fracture (original design).

The cracks did not result in loss of function during the 120-cycle test; however, their occurrence indicates localized structural damage and limits the conclusions that can be drawn regarding durability. Nevertheless, the results indicate that improvements to the structural design and material selection are necessary before clinical application can be considered. Future versions of the device could benefit from the use of engineering-grade polymers or metal components in high-stress regions, while retaining additive manufacturing as the rapid prototyping technique.

C. Limitations and Future Validation

The present study represents a preliminary mechanical proof of concept and has several limitations. The experimental evaluation consisted of a single 120-cycle test performed on a simplified laboratory test bench without physiological loading conditions. Therefore, the results do not establish the long-term durability, fatigue resistance, angular accuracy, or clinical safety of the mechanism. In addition, actuation force,

transmitted torque, mechanical backlash, gear wear, and cycle-to-cycle angular error were not quantitatively measured.

Further development should include repeated independent test series, angular displacement measurements, characterization of actuation force and transmitted torque, evaluation of backlash and gear wear, and extended fatigue testing. Tests under representative axial and mechanical loading conditions should also be conducted. A comparison with conventional manual adjustment would be necessary to determine whether the proposed mechanism provides any measurable improvement in adjustment consistency or usability. Material compatibility with sterilization or disinfection procedures was not evaluated in the present study and should be addressed in future development if clinical use is pursued.

IV. Conclusions And Future Work

This study presented the design and preliminary laboratory evaluation of a gear-based ratcheting mechanism intended to assist the manual adjustment of Ilizarov external fixators. The proposed mechanism was designed to produce a nominal quarter-turn adjustment of the nut for each lever actuation while allowing the lever to return without reversing the adjustment direction.

In the single 120-cycle laboratory test, the mechanism remained operational without critical failure of the gear transmission or ratchet system. The theoretical cumulative displacement was 30.00 mm, while 30.28 mm was measured experimentally, corresponding to a deviation of 0.28 mm (0.93%). A localized crack was also observed in the operating lever around the ratchet shaft mounting region, indicating a structural limitation of the current prototype.

These results provide preliminary evidence of the feasibility of the mechanical indexing concept under simplified laboratory conditions. However, the study does not establish angular accuracy, fatigue resistance, long-term durability, safety, or clinical applicability. Further work should include repeated independent tests, measurements of angular error, actuation force and transmitted torque, assessment of backlash and gear wear, fatigue testing under representative mechanical loads, and comparison with conventional manual adjustment before any consideration of clinical use.

Acknowledgment

The authors would like to thank the Federal Rural University of the Semi-Arid (UFERSA) for the institutional support provided during the development of this work, as well as the laboratories of the Center for Engineering and their respective faculty members.

References

[1] G. A. Ilizarov, "The tension-stress effect on the genesis and growth of tissues: Part I. The influence of stability of fixation and soft-tissue preservation," Clinical Orthopaedics and Related Research, no. 238, pp. 249–281, Jan. 1989.

[2] G. A. Ilizarov, "The tension-stress effect on the genesis and growth of tissues: Part II. The influence of the rate and frequency of distraction," Clinical Orthopaedics and Related Research, no. 239, pp. 263–285, Feb. 1989.

[3] D. Paley, Principles of Deformity Correction. Berlin, Germany: Springer-Verlag, 2002.

[4] S. R. Rozbruch, S. Ilizarov, W. A. Blyakher, and J. G. Shearer, Limb Lengthening and Reconstruction Surgery. Boca Raton, FL, USA: CRC Press, 2006.

[5] D. Paley, "Current techniques of limb lengthening," Journal of Pediatric Orthopaedics, vol. 8, no. 1, pp. 73–92, 1988.

[6] A. S. Green, "The evolution of the Ilizarov technique: Part 1," Strategies in Trauma and Limb Reconstruction, vol. 1, no. 1, pp. 3–8, 2006.

[7] A. S. Green, "The evolution of the Ilizarov technique: Part 2," Strategies in Trauma and Limb Reconstruction, vol. 1, no. 2, pp. 97–105, 2006.

[8] L. Kong, H. A. Li, Q. Kang, and G. Li, "An update to the advances in understanding distraction histogenesis: From biological mechanisms to novel clinical applications," Journal of Orthopaedic Translation, vol. 25, pp. 3–10, Nov. 2020.

[9] X. Zhang *et al*., "Evolution of the Ilizarov Technique: A Scoping Review of Improvement," *Langenbeck's Archives of Surgery*, 2025.

[10] K. N. D. Widanage, M. J. De Silva, T. D. Lalitharatne, A. M. J. Bull, and R. A. R. C. Gopura, "Developments in Circular External Fixators: A Review," *Injury*, vol. 55, no. 3, Art. 111365, Mar. 2024.

[11] R. G. Budynas and J. K. Nisbett, Shigley's Mechanical Engineering Design, 9th ed. New York, NY, USA: McGraw-Hill, 2011

H-Wave Detection in His Bundle Electrograms Using an LSTM Recurrent Neural Network

Naelso Alves Cunha
Department of Electronics and Systems
Universidade Federal de Pernambuco
Recife, Brasil
naelso.cunha@ufpe.br

Ana Vitória Morais Inocêncio
Department of Electronics and Systems
Universidade Federal de Pernambuco
Recife, Brasil
vitora.morais@ufpe.br

Emmanuel Andrade de Barros Santos
Department of Control and Automation
Universidade de Pernambuco
Recife, Brasil
emmanuel.andrade@upe.br

Gilson José Alves
Academic Department of Technical Courses
Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco
Recife, Brasil
gilsonj.alves@recife.ufpe.edu.br

Afonso Luiz Tavarez de Albuquerque
Pernambuco University Cardiology Emergency Hospital
Universidade de Pernambuco
Recife, Brasil
afonsoalbuquerque@terra.com

Marco Aurélio Benedetti Rodrigues
Department of Electronics and Systems
Universidade Federal de Pernambuco
Recife, Brasil
marco.rodrigues@ufpe.br

Abstract — Accurate identification of the components of the His Bundle Electrogram (HBE) is fundamental for mapping intracardiac bioelectrical activity. However, the automatic detection of the H-wave is challenged by pronounced interpatient morphological variability. This paper presents a methodology based on an LSTM Recurrent Neural Network for the automatic detection of the H-wave from invasive electrophysiological signals obtained exclusively through the Distal Ablation (ABLD) intracardiac channel. The proposed method utilizes sliding window segmentation and a post-processing algorithm based on a temporal persistence criterion, requiring continuous model activation for at least three consecutive windows to validate the event. The methodology was evaluated at the event level against an expert-defined gold standard. Quantitative results demonstrated the robustness of the classifier, which achieved maximum sensitivity (100%) by correctly identifying all 104 His potentials, with a precision of 86.67% and an F1-score of 92.86%. Visual analysis confirmed the consistency of the detections, showing that false activations were restricted to the signal's transition edges, without compromising the discrimination of the H-wave from adjacent atrial and ventricular activities. The system presents itself as a promising tool for future integration into automatic electrophysiological interval estimation systems and for supporting the diagnosis of conduction blocks.

Keywords — His Bundle Electrogram, LSTM, Biomedical Signal Processing, Pattern Recognition, Temporal Persistence.

I. INTRODUCTION

Cardiovascular diseases are the leading cause of mortality worldwide, representing an ongoing challenge for healthcare systems [1]. In addition to coronary artery disease, which stands out as the most lethal [2], disorders of the cardiac conduction system, such as atrioventricular blocks, are clinically critical as they present an asymptomatic evolution and a risk of severe hemodynamic compromise [3]. Various risk factors significantly increase the probability of acute cardiovascular events [4]. In this context, the early detection of alterations in electrical conduction is fundamental, with the conventional Electrocardiogram (ECG) and Holter monitoring being the primary non-invasive tools for evaluating myocardial bioelectrical activity [5], [6].

High-Resolution ECG (HRECG) systems employ enhanced electronics to detect low-amplitude signals on the surface [7], [8]. However, specific intracardiac potentials, such as the depolarization of the bundle of His, remain difficult to record superficially, requiring invasive procedures for their acquisition [9], [10]. The His Bundle Electrogram (HBE), acquired via electrophysiological studies, provides the H-wave, which is essential for the evaluation of ventricular conduction delays [3], [11]. However, the automatic determination of this parameter in a clinical environment remains a challenge due to the pronounced interpatient morphological variability of the H-wave [12]–[15].

The automatic detection of the H-wave is characterized as a pattern recognition problem in biomedical time series [16]. Although feedforward networks can incorporate temporal context [17], Long Short-Term Memory (LSTM) architectures prove to be highly suitable for modeling the complex temporal dependencies of these records, presenting promising results in the identification of short-duration physiological events [18]–[22].

Classical approaches based on amplitude thresholding or adaptive filtering exhibit severe limitations in H-wave detection due to the pronounced interpatient morphological and dimensional variability of the deflection [Shah et al., 2020; Boonstra et al., 2020]. Rigid techniques cannot adapt to changes in the shape of the His potential, frequently failing when attempting to parameterize dimensions that fluctuate from case to case. The combined use of an LSTM architecture with temporal persistence evaluation bridges this gap in the literature, since the network maps contextual dependencies regardless of the exact morphological dimensions, whereas the temporal criterion requires activation to be physiologically sustained, thereby overcoming the geometric inflexibility of traditional methods.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Therefore, this work proposes a method based on an LSTM network, trained with the ADAM (Adaptive Moment Estimation) optimizer [23], combined with a post-processing algorithm based on temporal persistence for the automatic detection of the H-wave. The system's performance is validated against expert delineations using exclusively the Distal Ablation (ABL D) intracardiac channel. The method aims to provide a robust tool to assist, in the future, in the automatic estimation of electrophysiological intervals and the identification of conduction blocks in clinical routine.

II. Materials and Methods

A. Signal Database

The data employed in this work originate from invasive electrophysiological studies performed at the Pronto-Socorro Cardiológico de Pernambuco (PROCAPE), conducted by cardiologists specializing in cardiac electrophysiology. All signals were acquired under the approval of the ethics committee (CAAE 53171921.6.0000.5208, opinion number 5.135.351). For the development of this work, the clinical records were partitioned into disjoint sets to ensure a robust hold-out validation. The independent test set, utilized to evaluate the model comprised continuous signals from 4 distinct patients. Each test record had a duration of 60 seconds with a sampling rate of 1000 Hz. The network was previously optimized on a separate, non-overlapping training set. The clinical electrophysiological procedure generates a set of multichannel records via catheters with intracardiac electrodes, which reflect different perspectives of the myocardial electrical activity. However, the proposed Recurrent Neural Network (RNN) architecture utilizes the ABL D channel as the input signal, since this is the only record in which the deflection associated with the His potential manifests itself viably for detection, due to the direct anatomical proximity of its electrode to the conduction bundle.

B. Use of ANN for H-Wave Identification

The HV (His-Ventricular) interval is a clinical metric widely used in intracardiac electrophysiology, defined as the interval between the H-wave deflection and the onset of the ventricular activation signal [3],[11]. As illustrated in Fig. 1, the system synchronously displays the surface records (leads I, II, III, aVF, V1, and V6, in gray), referring to the non-invasive signals of the cardiac cycle, and the invasive records obtained by intracardiac electrodes (ABL P and ABL D channels, in green). In this mapping of the atrioventricular conduction system, the His potential (H-wave) is observed, which manifests in the ABLD channel, being highlighted in orange in specific cycles. The precise acquisition of these intervals in a clinical environment therefore requires the exact identification of the temporal positions of the H-wave in the records of the His bundle electrogram.

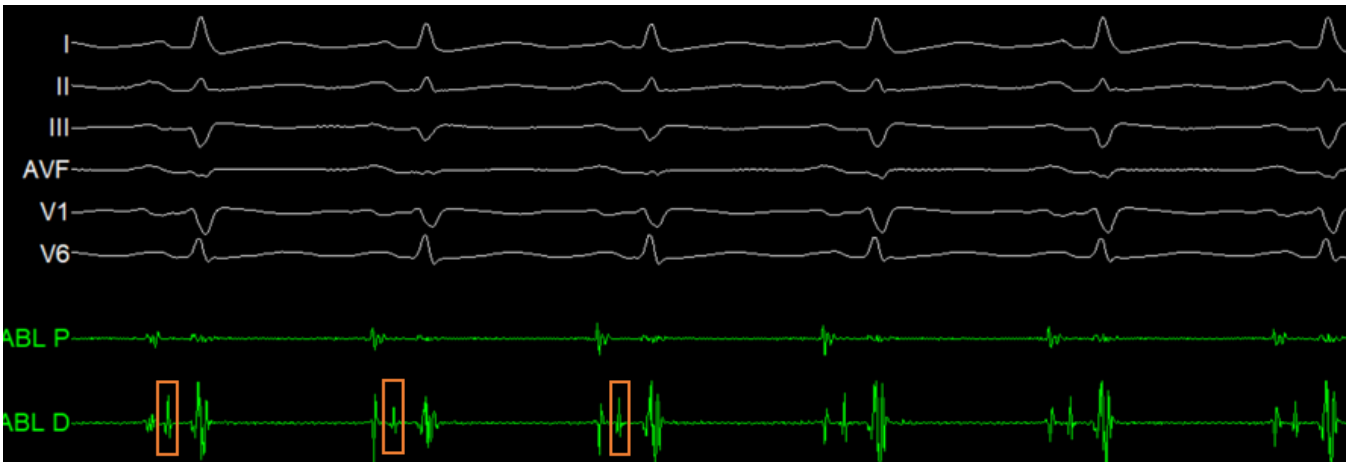


Fig. 1. Surface records (in gray) and invasive records obtained by intracardiac electrodes (in green). The His bundle electrogram (HBE) corresponds to the signal recorded in the ABL D channel, and the H-wave is highlighted in some cycles (in orange).

A widely employed strategy to incorporate temporal information in feedforward artificial neural networks is the use of the variation of activity over time, rather than the specific shape of the signal, such as the Artificial Neural Network with Input Dynamics (ANN-ID). However, its capacity to model longer or variable temporal dependencies is strongly conditioned by the choice of the window size and the number of delays considered [17]. In scenarios where the events of interest are transient, exhibit low amplitude, and are immersed in extensive and noisy sequences (as occurs in the detection of the His bundle electrogram), this dependency can limit the model's generalization and require specific adjustments for each dataset [12], [13].

Long Short-Term Memory (LSTM) architectures were designed to directly address these limitations by incorporating explicit memory mechanisms and controlling the flow of information over time. Unlike ANN-IDs, LSTMs allow temporal dependencies to be learned adaptively, without the need to predefine the extent of the relevant temporal dynamics. This characteristic is particularly advantageous in biomedical signals characterized by interpatient morphological variability, uncertainty regarding the exact temporal position of the event of interest, and the strong influence of the surrounding temporal context.

Reviews on the application of recurrent neural networks in electrocardiographic signals show that the LSTM presents consistent performance in detecting short-duration physiological events, reinforcing its suitability for the processing of intracardiac signals [18], [19].

In this work, a Long Short-Term Memory (LSTM) recurrent neural network was utilized, configured to operate directly on one-dimensional temporal sequences. This choice is based on the LSTM's ability to adaptively model temporal dependencies without the need for prior definition of the relevant temporal horizon, a characteristic especially suited

for the detection of the H-wave in noisy and low-amplitude bioelectrical signals.

C. *Architecture, Dynamic Nature, and Training Parameters of the ANN*

The proposed architecture presents a dynamic nature at both the input and the output: the input is provided in the form of time series (windows of consecutive samples), and the output consists of a continuous numerical value, between 0 and 1, representing the degree of activation or the probability of the H-wave occurrence for each evaluated window.

The HBE signal, previously pre-processed, is segmented into sliding windows of 40 samples (1000 Hz - 40 ms), with a step size of one sample. This window length was defined to fully encompass the typical duration of the deflection associated with the His bundle potential, as well as part of its immediate temporal surroundings, allowing the network to access the local context necessary to distinguish the H-wave from other components of the PR segment, without introducing excessively long windows that dilute the information of interest.

The processing flow of the proposed model was structured for the analysis of the signal's temporal dynamics, comprising three main stages (Fig. 2). Initially, a sequential input layer receives the one-dimensional time series in 40-sample windows. Next, an LSTM layer with 50 hidden units processes the sequence and extracts a compact vector that represents the entire temporal and morphological dynamics of the window. This high-level representation feeds a fully connected layer with a single neuron, which generates a continuous scalar prediction indicating the degree of H-wave presence. Finally, the output layer operates with a regression approach, using the Mean Squared Error (MSE) loss function to evaluate the prediction error and guide the adjustment of the network parameters via backpropagation during training.

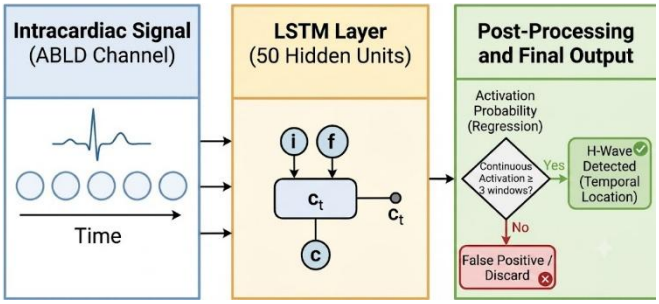


Fig. 2. Structure of LSTM Network

Subsequently, in the algorithm's detection stage, this continuous output is converted into a binary decision by means of a decision threshold: windows with an output greater than or equal to this value are considered as containing the H-wave, while the others are classified as negative.

The training of the network was performed using the ADAM algorithm [23] with an initial learning rate of 1×10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, executed over 150 epochs, utilizing the Mean Squared Error (MSE) as the loss function.

Once the training phase is concluded, the consolidated model is applied to perform a continuous scan across the unseen examination records (60 seconds per patient), generating a time series of predictions. The final H-wave annotations are then extracted by applying an amplitude threshold of 0.89 to these predictions, followed by the validation of the temporal persistence criterion across three successive windows, as structured in the post-processing stage in Fig. 2.

D. *HBE Identification Procedure*

The sequential integration of the described steps, starting from the raw data master file to the final mapping of the H-wave annotations, is structured chronologically in the flowchart of Fig. 3

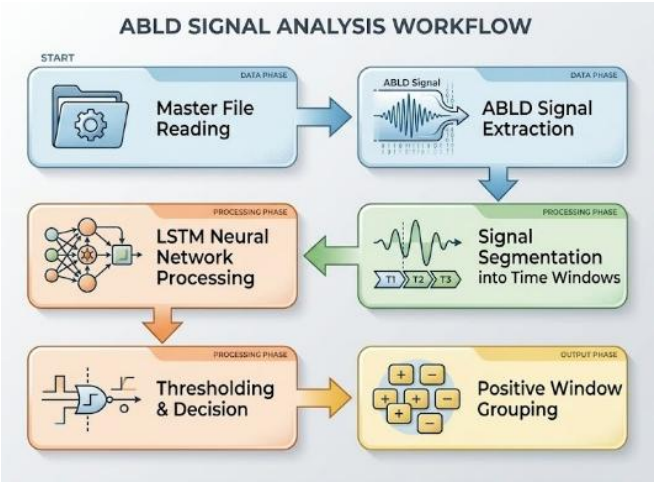


Fig. 3. Flowchart of the HBE identification procedure

The procedure is divided into sequential stages of pre-processing, inference, and post-processing:

- Signal Reading and Extraction: Initially, the system reads the master file containing the multichannel electrophysiological records. From this set, the ABL D intracardiac channel is specifically extracted, since it is the record in which the visualization of the His potential is viable, due to the direct anatomical proximity of its electrode to the bundle of His.
- Window Segmentation: The extracted signal is segmented using a sliding window technique. Each window has a fixed length of 40 samples (40 ms), with a unit step (1-sample stride), ensuring that the temporal dynamics of the H deflection are fully captured across multiple adjacent windows.
- LSTM ANN Processing: The segmented windows feed the LSTM neural network, which processes the morphology of each block and generates a continuous scalar output. This output represents the degree of activation or the probability of the His potential's presence for that specific instance in time.
- Thresholding and Decision: The continuous activation curve is subjected to a fixed decision threshold. Samples resulting in values above this threshold are candidates for positive classification, while the rest are discarded as noise or atrial/ventricular components.
- Grouping and Persistence: In the final post-processing stage, the mechanism for grouping positive windows is applied. To mitigate spurious activations and ensure physiological consistency, a detection is only validated as an effective His potential if there is temporal persistence of the activation for at least three consecutive windows above the threshold. This criterion ensures that only segments with a temporal

sustainment compatible with the physiology of the conduction system are identified by the system.

Based on the final temporal positions consolidated by the window grouping block (post-processing), the results generated by the model were subjected to performance evaluation. The metrics used to compare the model's results against the expert gold standard and to measure the system's accuracy are detailed below.

E. Data Evaluation Metrics

The quantitative evaluation of the classifier is conducted from the perspective of detecting discrete events localized at specific instants within a continuous signal. Each event of interest corresponds to the occurrence of the deflection associated with the His bundle potential (H-wave).

An actual His potential is considered correctly detected when a valid model activation occurs within the tolerance window associated with the gold standard.

This temporal tolerance window was established at ± 20 ms relative to the peak demarcated by the expert, corresponding to the typical physiological duration of the H-wave deflection. Furthermore, the decision threshold of 0.89 was optimized through Precision-Recall (PR) curve analysis during the validation phase to maximize the F1-score. To increase the model's robustness against false alarms, the training phase explicitly included 'hard negatives', defined as signal segments that did not contain the H-wave but exhibited high-amplitude artifacts or atrial/ventricular deflections with confounding morphologies.

For an activation to be considered valid, a temporal persistence criterion was established in which the network output must exceed the decision threshold for at least three consecutive windows, mitigating isolated spurious triggers caused by background noise. The classification process is based on three main components of the confusion matrix adapted for events:

- True Positives (TP): Number of His potentials correctly identified by the classifier within the tolerance window established by the expert.
- False Negatives (FN): His potentials present in the gold standard that were not accounted for or detected by the model.
- False Positives (FP): Classifier activations that occur outside the gold standard tolerance regions, characterizing false alarms or incorrect detections.

Based on these components, the overall performance of the system is measured using the metrics of Precision (P), Sensitivity (S, or Recall), and F1-score, mathematically defined by:

$$P = \frac{TP}{TP + FP},$$

$$S = \frac{TP}{TP + FN},$$

$$F1 = 2 \cdot \frac{P \cdot S}{P + S}.$$

The F1-score metric provides the harmonic mean between precision and sensitivity, synthesizing the dynamic balance between the exactness of the detections and the robustness of the classifier against the incidence of false alarms.

III. RESULTS AND DISCUSSION

The detection of the events precisely demarcated in the gold standard served as the basis for the quantitative evaluation of the classifier's performance. An actual event was considered correctly detected when at least one valid network activation occurred within the time interval corresponding to the His potential. In this mechanism, a detection is considered valid only when the classifier output exceeds the decision threshold across consecutive windows over time, which reduces isolated activations and ensures that only patterns with a temporal sustainment compatible with the His potential are considered as detected events.

The quantitative results are summarized in Table I, which gathers the values of TP, FN, and FP, along with the derived metrics after refining the training with hard negative windows and applying the consecutive window voting mechanism.

TABLE I. LSTM NETWORK METRICS AFTER TRAINING

Metric	Value
True Positives (TP)	104
False Negatives (FN)	0
False Positives (FP)	16
Precision (P)	86.67%
Sensitivity (Recall)	100.00%
F1-score	92.86%

It is observed that all His potentials present in the evaluated set were correctly identified (104), resulting in maximum sensitivity (100%). The occurrence of false positives remained limited to 16, indicating that the majority of the classifier's activations effectively corresponded to real events, as reflected by the high precision value of 86.67%. The resulting F1-score of 92.86% confirms the balance between detection capability and the control of false alarms.

The results from the table demonstrate that the combination of the LSTM-based recurrent architecture, the use of appropriate methodological choices in the decision stage, and the temporal aggregation mechanism of the decisions resulted in a robust and consistent classifier for the automatic detection of the His potential.

The individualized evaluation proves the classifier's consistency against morphological variability. When analyzing the detections separately for the 4 patients in the test set (Table II), sensitivity remained at 100% for all individuals. The occurrence of false positives ranged from 3 to 5 events per patient, generating stable F1-scores between 91.80% and 94.12%. The mean F1-score across patients was 92.90% with a standard deviation of only 0.97%. This narrow variance statistically attests that the method generalizes H-wave recognition, without overfitting bias to a specific individual.

TABLE II. LSTM NETWORK METRICS PER PATIENT

Patient	TP	FN	FP	P (%)	R (%)	F1 (%)
Patient 1	28	0	5	84.85	100.00	91.80
Patient 2	24	0	3	88.89	100.00	94.12
Patient 3	25	0	4	86.21	100.00	92.59
Patient 4	27	0	4	87.10	100.00	93.10
Total Test	**104**	**0**	**16**	86.67	100.00	92.86

To ensure clinical feasibility in the future estimation of the HV interval, the temporal localization error was evaluated. The Mean Absolute Error (MAE) between the activation instant detected by the model and the expert-demarcated H-wave peak was 2.3 ± 0.8 ms, demonstrating high temporal adherence and accuracy suitable for precise electrophysiological interval measurements. This minimal temporal offset confirms that the combination of LSTM-based modeling and temporal persistence filtering preserves the precise alignment required for clinical deployment [Chen & Upadhyay, 2021].

Fig. 4 presents the relationship between the original signal and the classifier's response within a recording interval. The original signal is located in the upper part of Fig. 4, in which the pattern of interest (gold standard) was identified by an expert, indicated by orange arrows. The lower part of the same figure presents the network output for the same sample interval, allowing a direct comparison of the model's behavior with the gold standard. The green curve represents the binarized reference signal previously marked by the expert. Overlaid on this trace are the components corresponding to true positives (TP), false positives (FP), and false negatives (FN). This trace allows for the verification, throughout the signal, of whether the network activation coincides with the presence of the pattern of interest, whether an undue activation occurs, or whether there is an absence of detection. Furthermore, from the same figure, it can be noted that in the regions indicated by the orange arrows, the network output exhibits concentrated activations that are temporally aligned with the intervals defined in the gold standard, characterizing correct detections of the His potential and demonstrating the model's capability to identify the occurrence of the physiological event.

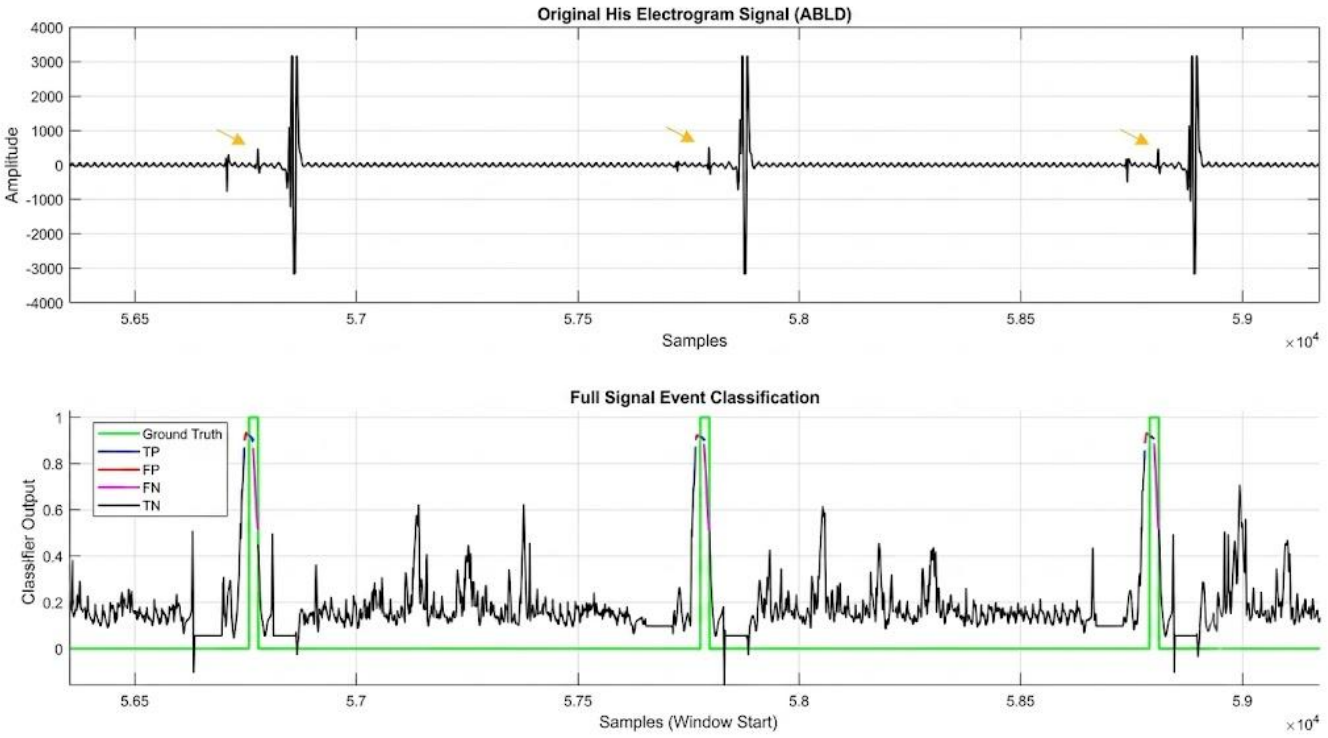


Fig. 4. Network Output

The occurrence of 16 false positives (FP), as indicated in Table I, is predominantly concentrated in small residual oscillations adjacent to the onset and offset limits of the H-wave (represented in red in Fig. 4). From the perspective of bioelectrical signal processing, these small false alarms at the margins of the deflection do not represent discrimination failures against atrial or ventricular activities, but rather a characteristic of the conservative nature of the unidirectional LSTM architecture, which tends to subtly extend the activation horizon before returning to the resting state. The aggregation mechanism using three consecutive windows proved robust by confining these undue alarms to the immediate surroundings of the His bundle itself, preventing erratic triggers along the PR segment and preserving the necessary rigor for the subsequent automatic estimation of the HV interval.

It is important to highlight the methodological distinction between the event-level evaluation, summarized in Table I, and the local dynamics observed in the signal trace in Fig. 3. Although Fig. 3 shows small regions of false negatives (FN, in magenta) at the transition edges of the H-wave, the final count of missed events in Table I is null (FN = 0). This behavior stems from the adopted validation criterion: the local discontinuities and the slight phase delay

in the network's initial activation merely reflect the model's sample sensitivity to the rapid morphological variation of the deflection. However, as the main body of the His potential is captured in a sustained manner, the post-processing algorithm consolidates the activation, ensuring that 100% of the reference physiological events are computed as true positives (TP).

The temporal synchronization of the two signals shows that the classifier maintains temporal coherence in identifying patterns of interest. The network activations occur temporally aligned with the intervals annotated in the gold standard, without undue prolongations or relevant shifts over time. This comparison reinforces the presented quantitative results, confirming that the network achieves good overall results and produces temporally consistent detections when analyzed at a detailed level on the original trace.

Although Fig. 4 has shown false negatives (segments in pink), there is no compromise in the final identification of the physiological event. The FN segments reflect the response time/phase delay of the unidirectional temporal LSTM modeling before the stabilization of the activation curve.

The model consistently detects the His Bundle Electrogram (HBE) after the temporal aggregation of the windows, and the punctual reductions in activation merely reflect the local sensitivity of the method, without impairing the localization, extent, or interpretation of the HBE. These FNs are physiologically irrelevant and methodologically expected, not interfering with the overall accuracy of the detections nor with the clinical visualization of the identified pattern.

During the model's development, temporal aggregation mechanisms were implemented to eliminate dispersed activations and reflect continuous segments, coherent with the morphology of the HBE.

From the perspective of clinical feasibility, the model demonstrated high computational efficiency. The mean inference time for processing one second of signal (considering the sliding window stride) was approximately 18.5 ms on a standard CPU (Intel Core i7), ensuring full capability for real-time execution in hospital routine.

The network maintained its detection capability in segments where the His potential occurs, without detecting deflections associated with atrial and ventricular activities. The algorithm is able to discriminate the morphology and the temporal position of the H-wave within the atrioventricular segment, resulting in a precise temporal delimitation of the event, with small local variations associated with the granularity of the temporal segmentation and the relative positioning of the analysis windows, without compromising the physiologically correct localization of the H-wave.

IV. Conclusions

This work presented a methodology based on an LSTM Recurrent Neural Network for the automatic detection of the H-wave in the His Bundle Electrogram (HBE). The proposed approach addressed the limitations arising from morphological variability and the intrinsic noise of the biosignal through a post-processing algorithm based on the temporal persistence criterion in sliding windows.

The quantitative results validated the robustness of the model, which achieved maximum sensitivity (100%) by correctly identifying all 104 gold-standard events, in addition to an F1-score of 92.86%. Visual analysis confirmed the temporal consistency of the detections and evidenced that the residual false positives were confined to the signal's transition edges, without compromising the clinical integrity of the diagnosis. The system demonstrated the ability to precisely discriminate the H deflection from adjacent atrial and ventricular components using exclusively the ABL D intracardiac channel.

As future work, the integration of this classifier into systems for the automatic estimation of electrophysiological intervals (such as AH and HV) is proposed, aiming to consolidate the tool as a robust support for the identification of conduction blocks in clinical routine.

Acknowledgment

The authors would like to thank the UFPE - Federal University of Pernambuco, the Brazilian Agencie Coordination for the Improvement of Higher Education Personnel (CAPES).

References

[1] M. Lindstrom et al., "Global Burden of Cardiovascular Diseases and Risks, 1990–2021: A Report from the GBD 2021 Diseases and Risk Factors Collaborators," Journal of the American College of Cardiology, vol. 80, no. 25, pp. 2372–2413, Dec. 2022.

[2] M. H. Moghadam and B. M. Asl, "Automatic detection of coronary artery disease from ECG signals using deep learning architectures," Biomedical Signal Processing and Control, vol. 85, p. 104899, Aug. 2023.

[3] D. P. Zipes, J. Jalife, and W. G. Stevenson, Cardiac Electrophysiology: From Cell to Bedside, 7th ed. Philadelphia, PA: Elsevier, 2017.

[4] P. Sarma and S. Biswas, "Cardiovascular disease risk factors in type 2 diabetes mellitus: A comprehensive review," Journal of Diabetes & Metabolic Disorders, vol. 22, no. 1, pp. 45–58, Jun. 2023.

[5] R. Neri et al., "The role of the electrocardiogram in modern clinical practice," International Journal of Cardiology, vol. 370, pp. 120–129, Jan. 2023.

[6] R. E. Klabunde, Cardiovascular Physiology Concepts, 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2012.

[7] T. Takuma et al., "High-resolution electrocardiography for the detection of low-amplitude cardiac potentials," IEEE Transactions on Biomedical Engineering, vol. 64, no. 4, pp. 812–820, Apr. 2017.

[8] N. A. Cunha et al., "High-Resolution ECG systems and advanced analog conditioning for low-amplitude potential amplification," Journal of Electrocardiology, vol. 82, pp. 45–53, Jan. 2024.

[9] C. Muresan et al., "Invasive vs. non-invasive assessment of the cardiac conduction system: A comparative review," Europace, vol. 21, no. 8, pp. 1140–1152, Aug. 2019.

[10] H. D. Huang and J. M. Miller, "Catheter ablation of cardiac arrhythmias: Technique and indications," Mayo Clinic Proceedings, vol. 94, no. 8, pp. 1655–1669, Aug. 2019.

[11] S. Chen and G. A. Upadhyay, "His bundle electrogram: Technical aspects and clinical applications," Heart Rhythm, vol. 18, no. 5, pp. 822–831, May 2021.

[12] S. Shah et al., "Morphological variability of the His deflection and challenges in automatic HV interval estimation," Pacing and Clinical Electrophysiology, vol. 43, no. 9, pp. 988–997, Sep. 2020.

[13] R. Boonstra et al., "Automated quantification of intracardiac electrograms: An open problem in electrophysiology," IEEE Reviews in Biomedical Engineering, vol. 13, pp. 204–215, Jan. 2020.

[14] L. Roudijk et al., "Signal-to-noise ratio challenges in local activation time estimation during electroanatomical mapping," Biomedical Engineering Online, vol. 20, no. 1, p. 42, Dec. 2021.

[15] J. Duchateau et al., "Mathematical and computational methods for intracardiac signal processing," Frontiers in Physiology, vol. 10, p. 1104, Sep. 2019.

[16] S. Kiranyaz et al., "1D convolutional neural networks and applications: A survey," Mechanical Systems and Signal Processing, vol. 151, p. 107398, Apr. 2021.

[17] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[18] S. Ansari et al., "Deep learning for ECG signal processing and pattern recognition: A comprehensive review," Information Fusion, vol. 91, pp. 312–333, Mar. 2023.

[19] R. Xiao et al., "Recurrent neural networks for cardiac arrhythmia detection from long-term ECG recordings," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 744–755, Feb. 2023.

[20] N. Yunita et al., "LSTM-based architectures for short-duration physiological event detection in intracardiac signals," Computer Methods and Programs in Biomedicine, vol. 258, p. 108502, Jan. 2025.

[21] M. Nemer et al., "Adaptive temporal modeling of noisy biomedical time series using recurrent networks," Artificial Intelligence in Medicine, vol. 159, p. 102741, Jan. 2025.

[22] W. K. Karri and S. R. Annavarapu, "Deep learning models for intracardiac electrogram analysis: A review," Cognitive Computation, vol. 15, no. 4, pp. 1201–1220, Jul. 2023.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, 2015, pp. 1–15.

Hand Gesture Classification from HD-sEMG Images Using Deep Learning and Decision Trees

Sandrelly Mirely da Silva Santos
Dept. de Ciências Farmacêuticas
Universidade Federal de Pernambuco
Recife, Brasil
sandrelly.santos@ufpe.br

Ingrid Bruno Nunes
Dept. de Ciências Farmacêuticas
Universidade Federal de Pernambuco
Recife, Brasil
ingrid.nunes@ufpe.br

Giselle Machado Magalhães Moreno
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
giselle.moreno@ufpe.br

Maíra Araújo de Santana
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
maira.araujosantana@ufpe.br

Clarisse Lins de Lima
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
clarisse.lins@ufpe.br

Wellington Pinheiro dos Santos
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
wellington.santos@ufpe.br

Juliana Carneiro Gomes Cassemiro
Dept. de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
juliana.cgomes@ufpe.br

Abstract - Motor impairments affecting the upper limbs constitute a significant challenge in rehabilitation. High-Density Surface Electromyography (HD-sEMG) provides detailed information on muscle activity; however, its high dimensionality poses significant processing challenges. This study proposes a hybrid approach that converts HD-sEMG signals into heat maps, extracts features using a pre-trained LeNet network, and classifies hand gestures through decision tree-based algorithms. The publicly available Hyser dataset, containing 10 functional gestures, was used. The data were divided into training (80%) and testing (20%) sets and evaluated using the J48, Random Tree, and Random Forest classifiers (100 to 500 trees). The Random Forest models achieved superior performance during cross-validation, with mean accuracy > 91.7% and AUC > 0.994. Considering the balance between performance and computational cost, the Random Forest model with 200 trees was selected for final validation. On the test set, it achieved an accuracy of 78.87%, a Kappa coefficient of 0.765, and a mean AUC of 0.962. The greatest misclassifications occurred between antagonistic gestures with similar muscle activation patterns, such as pronation and supination with a closed hand. The results demonstrate the feasibility of the proposed approach for technology-assisted motor rehabilitation systems by integrating the spatial information provided by HD-sEMG, automatic feature extraction via deep learning, and the rule-based nature of decision tree models, which-although still requiring complementary explainability analyses-provide a structurally interpretable decision framework.

Keywords - HD-sEMG; hand gestures; machine learning; decision trees.

I. Introduction

Motor impairments affecting the upper limbs constitute a significant challenge in rehabilitation, as they directly compromise an individual's ability to perform activities of daily living and maintain functional independence [1]. Such impairments may arise from a wide range of neurological and neuromuscular conditions, including cerebral palsy, traumatic brain injury, spinal cord injury, multiple sclerosis, and neurodegenerative diseases [1], [2]. Additionally, inherited ataxias, motor neuron diseases, polyneuropathies, and myopathies can result in impairments in hand strength, muscle tone, movement, dexterity, joint range of motion, and sensation. Regardless of etiology, the resulting loss of hand and wrist function imposes severe limitations on daily activities and diminishes quality of life. Consequently, the development of effective rehabilitation strategies targeting distal motor function is of paramount importance for promoting functional recovery and enhancing independence [1], [3].

In this context, some movement rehabilitation therapies require gesture recognition methods that can be employed in interfaces designed to support therapy for the affected limb [3]. Such recognition can be achieved through the analysis of High-Density Surface Electromyography (HD-sEMG) neuromuscular signals, providing a non-invasive means of decoding the user's motor intention, thereby enabling the control of external devices or providing feedback on muscle contraction [3], [4], [5].

HD-sEMG represents a significant advancement over conventional sEMG. By capturing detailed spatial information regarding muscle activity over a larger skin surface area, it enables the analysis of muscle activation patterns with a high level of detail, opening new possibilities for the recognition of more complex gestures and motor nuances. However, the high dimensionality inherent to HD-sEMG signals, which may contain hundreds of channels generating multivariate time series, imposes a substantial computational challenge for efficient processing and extraction of discriminative features [6], [7].

Traditionally, Machine Learning (ML) approaches applied to gesture classification from sEMG signals rely on feature extraction in the time and frequency domains. Although effective, these approaches may not fully capture the large amount of spatial information contained in HD-sEMG data. Strategies that convert multichannel signals into visual representations, such as heat maps, have emerged as a promising alternative [8], [6]. This conversion enables the problem to be transferred into the computer vision domain,

where Deep Learning architectures, particularly Convolutional Neural Networks (CNNs), have demonstrated strong performance in extracting hierarchical and spatially relevant features [8], [9].

Accordingly, there has been a growing number of studies and proposals focused on gesture classification using HD-sEMG signals. Among these, studies employing recurrent neural networks directly applied to instantaneous signal amplitude images, as presented in [10], achieved promising frame-by-frame classification performance and even higher accuracy when majority voting mechanisms over consecutive windows were employed. Another study implemented instrumented electrode grids combined with fully connected neural networks, achieving high accuracy rates for sequential movements, although relying on manual feature extraction [11]. More recently, [12] applied Vision Transformer (ViT)-based architectures to multi-day classification with ultra-short windows, demonstrating robustness even under reduced calibration conditions. Despite their high performance, these approaches depend on complex deep architectures with high computational cost and, in many cases, limited interpretability, which is a critical aspect for clinical applications.

In contrast, other authors emphasize the continued relevance of traditional machine learning methods, such as [13], who proposed hybrid feature selection algorithms combined with Support Vector Machine (SVM) classifiers, achieving competitive performance with significantly lower training time than deep networks. Nevertheless, these approaches still depend on manual feature engineering and do not exploit the automatic representation learning capabilities of CNNs. Therefore, a gap remains in the literature regarding the integration of automatic feature extractors based on deep learning (applied directly to visual representations of HD-sEMG signals) with structurally simpler, interpretable, and computationally efficient classifiers, such as decision trees.

Given this scenario, the present study aimed to evaluate the performance of Machine Learning algorithms implemented in the WEKA environment for hand gesture classification based on image representations of HD-sEMG signals. The proposed methodology was based on converting the signals into heat maps, extracting features using a pre-trained Convolutional Neural Network (LeNet), and classifying gestures using decision tree algorithms (J48, Random Tree, and Random Forest with 100 to 500 trees). These algorithms were selected due to their interpretability, robustness to noise, and suitability for high-dimensional data, which are essential attributes for clinical applications requiring transparency in decision-making processes.

II. Background

A. High-Density Surface Electromyography

Surface electromyography (EMG) records the electrical activity generated by muscle fibers during contraction through electrodes positioned on the skin [14], [15]. Unlike conventional EMG, which employs bipolar configurations with a limited number of channels and provides a global measure of muscle activation, High-Density Surface Electromyography (HD-sEMG) uses two-dimensional arrays containing dozens to hundreds of electrodes (typically 64-256), arranged in regular grids with reduced interelectrode spacing [14]. This configuration enables the acquisition of detailed spatial gradients of muscle activity, generating topographic maps that reveal intramuscular and intermuscular recruitment patterns [15], [16].

The main advantages of HD-sEMG include: (1) the generation of spatial activation maps, enabling visualization of the topographic distribution of muscle activity; (2) greater robustness due to spatial redundancy, reducing sensitivity to small electrode displacements; and (3) the potential ability to discriminate the activity of adjacent muscles [14], [16]. However, the high dimensionality inherent to these data (hundreds of multivariate time series) imposes significant computational challenges for the extraction of discriminative features, motivating the development of alternative processing strategies [16].

B. Transformation of HD-sEMG Signals into Visual Representations

A promising strategy for addressing the high dimensionality of HD-sEMG involves converting multichannel signals into two-dimensional visual representations, thereby transferring the problem to the computer vision domain [16], [8]. Among the available techniques, heat maps stand out due to their simplicity and their ability to preserve the anatomical topology of the electrode array: each position in the image corresponds to a specific electrode, while the color intensity represents the signal amplitude at that region [8].

Transforming the signal into the visual domain provides significant conceptual advantages. It enables structured dimensionality reduction by converting hundreds of time series into a single object (image) that preserves the essential spatial relationships [8], [17]. More importantly, this representation allows the automatic exploration of spatial patterns (such as contours, gradients, symmetries, and activation regions) through CNN architectures, which are specialized in detecting hierarchical features in grid-structured data [17]. Thus, the image-based representation functions as a methodological bridge connecting the spatial information of HD-sEMG data with advances in artificial intelligence and computer vision, expanding the possibilities for robust classification and analysis of complex motor patterns [9], [16].

C. Decision Trees

Decision trees are classifiers that recursively partition the feature space through tests applied to attributes, forming a hierarchical structure in which each internal node represents a decision and each leaf node corresponds to a predicted class. Their main advantage lies in their rule-based structure, which can be expressed as a set of IF-THEN rules. This structural property facilitates model inspection and constitutes a relevant starting point for clinical contexts that value decision transparency [18], [19], [20].

J48 is the implementation of the C4.5 algorithm, which employs the gain ratio criterion for attribute selection and incorporates post-pruning to avoid overfitting [20]. Random

Tree introduces controlled randomness into tree construction by considering a random subset of attributes at each node for splitting, thereby promoting greater structural diversity [19].

Random Forest is an ensemble method that combines bagging (bootstrap sampling of the training set) with the random attribute selection characteristic of Random Tree [19]. Multiple trees are trained in parallel, and the final prediction is obtained through majority voting. This dual randomness provides the model with high robustness, low sensitivity to overfitting, and excellent performance in high-dimensional scenarios while maintaining a reasonable degree of interpretability through feature importance analysis [19], [21].

D. Deep Learning and Convolutional Neural Networks

Deep learning is a subfield of machine learning that employs neural networks with multiple layers to learn hierarchical representations directly from raw data, eliminating the need for manual feature extraction [23]. Among its architectures are Convolutional Neural Networks (CNNs), which are designed to process grid-structured data, such as images, by exploiting properties of local correlation and translation invariance [24]. A typical CNN alternates convolutional layers, which apply learnable filters to detect features ranging from low-level to high-level patterns (edges, textures, semantic patterns), and pooling layers, which reduce spatial dimensionality [23], [24], [25].

Among the various applications of CNNs, their ability to function as automatic feature extractors is particularly noteworthy, transforming an input image into a compact vector of relevant attributes that can subsequently be used by other classifiers. The LeNet architecture, proposed by Yann LeCun in 1998, is a classical example of this paradigm, originally developed for handwritten digit recognition [26]. Transfer learning further enhances this capability by allowing a CNN pre-trained on a large dataset (such as MNIST) to provide generic visual representations (edges, corners, textures) transferable to other domains, including biomedical images [24], [26]. Pooling layers provide a relevant methodological trade-off: max pooling emphasizes salient features but is more sensitive to noise, whereas average pooling preserves the overall signal energy and produces representations that are more robust to local variations, making it more suitable when the spatial distribution of activity is informative [27].

E. Evaluation Metrics and Statistical Analysis

Several metrics can be employed to evaluate classifiers, each being sensitive to specific characteristics of the problem. Among the most commonly used metrics in healthcare applications, and those adopted in this study, are accuracy, the Kappa index, the Area Under the ROC Curve (AUC), sensitivity, and specificity [28], [29].

Accuracy represents the overall proportion of correctly classified instances. The Kappa index measures the agreement between predictions and true labels while correcting for agreement expected by chance, making it particularly useful in multiclass problems. AUC quantifies the discriminative capability of the model. Sensitivity (true positive rate) and specificity (true negative rate) enable a more granular analysis of class-wise performance, complementing the overall evaluation. In addition, the confusion matrix provides a detailed visualization of the classifier's correct and incorrect predictions for each gesture [28], [30].

Of particular relevance to this study is the Friedman test, a non-parametric test suitable for multiple comparisons involving related samples [22]. When the null hypothesis is rejected ($p < 0.001$), the Durbin–Conover post-hoc test may be applied. This test adjusts p-values to control the Type I error rate (false positives) in multiple comparisons, providing an appropriate balance between statistical power and error control [31], [32].

III. Methodology

The methodology adopted in this study is illustrated in Figure I, which describes the main stages addressed throughout the research and will be detailed in the following sections.

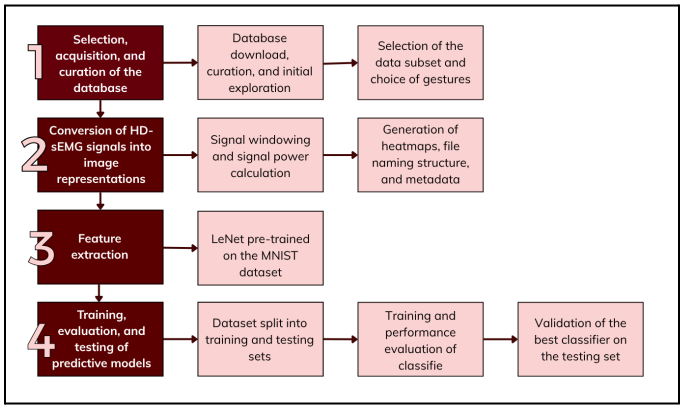


Figure I. Flowchart of the methodological stages of the study.

A. Dataset

For this study, the publicly available Hyser (High-density sEMG Recordings) dataset, available in the PhysioNet repository, was used. The dataset consists of HD-sEMG recordings from 20 healthy right-handed subjects (12 men and 8 women; age range: 21-34 years). Each participant performed two experimental sessions on different days, with an average interval of 8.5 ± 6.7 days between sessions [33].

The acquisition system employed four electrode arrays positioned on the forearm - two over the flexor muscles and two over the extensor muscles (distal and proximal) - totaling 256 channels arranged in a 16 × 16 grid. Signals were acquired using the Quattrocento system (OT Bioelettronica, Italy) with a sampling rate of 2048 Hz, a gain of 150, and 16-bit resolution, using high-pass (10 Hz) and low-pass (500 Hz) filters.

The Hyser dataset contains five experimental subsets. For this study, the Pattern Recognition (PR Dataset) subset was selected, comprising 34 hand gestures performed in two repeated trials. Only the maintenance segments (4 - second duration), during which participants sustained the gesture in a stable manner, were used in order to ensure greater signal stability.

Ten functional gestures (Figure II) were selected due to their relevance to activities of daily living and motor

rehabilitation: thumb extension, wrist flexion, wrist extension, wrist supination, radial deviation with a closed hand, pronation with a closed hand, supination with a closed hand, hand closing, hand opening, and three-finger pinch.

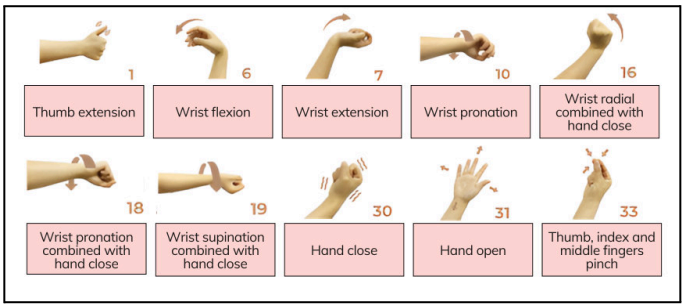


Figure II. Hand gestures selected for the study. Adapted from [34].

B. Preprocessing and Conversion to Images

The raw 4-second signals were segmented into 250 ms windows with 125 ms overlap. This configuration was adopted with future real-time applications in mind, such as human–machine interfaces or gamified rehabilitation systems, where short response times are essential.

For each temporal window and each of the 256 channels, the RMS value was calculated and reorganized into a 16 × 16 matrix reflecting the physical arrangement of the electrodes. To ensure that the visual representation corresponded to the actual anatomical orientation, geometric correction transformations were applied through horizontal mirroring followed by vertical mirroring using the NumPy library.

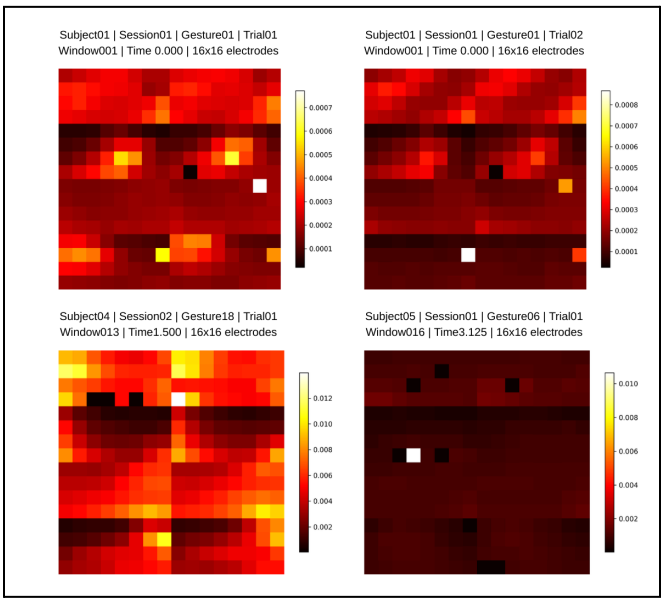


Figure III. Examples of heat maps generated from HD-sEMG signals. The images used in the machine learning experiments did not contain textual annotations in order to avoid bias during classifier training.

The corrected matrix was then converted into heat maps using the Matplotlib library in Python. The "hot" colormap was selected due to its effectiveness in representing intensity gradients: dark tones indicate low activation, orange/yellow tones represent moderate activation, and light/white tones correspond to high muscle activation. The images were generated at a resolution of 300 DPI, with all axis elements removed so that only the relevant visual information was retained.

At the end of the process, 24,707 images were generated - 2,400 per gesture, except for the three-finger pinch gesture (gesture 33), which totaled 2,387 images due to missing acquisitions in the original dataset. Each image was named according to a systematic protocol containing metadata (subject, session, gesture, trial, window, and time), enabling complete traceability (Figure III).

C. Feature Extraction with Deep Learning

To extract discriminative features from the heat maps, a transfer learning approach was employed using the LeNet architecture, a classical CNN widely validated for image recognition tasks. The network was initialized with weights pre-trained on the MNIST dataset, providing filters capable of detecting fundamental visual features (edges, corners, and textures) transferable to the analysis of HD-sEMG heat maps.

A fundamental adaptation consisted of replacing the original Max Pooling layer with Average Pooling, as the latter tends to better preserve the global spatial information of muscle activation by smoothing local peaks and producing a more robust and generalizable representation.

The extracted feature vectors (500 attributes) from the penultimate layer of the network were consolidated and structured into an ARFF (Attribute-Relation File Format) file, the native format of WEKA, containing the image paths and their corresponding labels (gestures). This structure enabled efficient integration between feature extraction in Python and the classification stage in WEKA.

D. Classification and Experimental Configuration

The dataset was divided into training (80%) and testing (20%) sets using the WEKA Resample filter with stratification to preserve the proportion of each gesture in both sets and without replacement, ensuring that training instances were not repeated in the test set. Due to the balanced distribution of the 10 classes in the original dataset, no additional artificial balancing techniques were required.

Eight decision tree-based classifiers were evaluated: J48, Random Tree, and Random Forest with 100, 200, 300, 400, and 500 trees.

The experiments were conducted in the Experimenter module of WEKA using the training set (80% of the data), employing 10-fold cross-validation with 10 repetitions. This protocol (100 executions in total) enables robust statistical estimation of model performance while compensating for the variability inherent to data partitioning.

The evaluated metrics were accuracy, Kappa index, Area Under the ROC Curve (AUC), sensitivity, and specificity. To verify statistically significant differences among classifiers, the Friedman test ($\alpha = 0.05$) was applied. Once the null hypothesis was rejected, the Durbin-Conover post-hoc test was employed to identify which pairs of algorithms differed significantly, with adjustments applied to control the Type I error rate.

E. Best Model Selection and Validation

The classifier that achieved the best balance among predictive performance (with emphasis on the Kappa index), low variability across repetitions, and lower computational cost (training and testing time) was selected for the final validation stage.

The selected model was then evaluated on the independent test set (20% of the data, not used in any training or cross-validation stage). From this validation, the confusion matrix and the final class-wise performance metrics were generated, providing a realistic estimate of the model's behavior when faced with unseen data.

IV. Results and Discussion

A. Classifier Performance

After training the classifiers using the 80% training set, the performance metrics summarized in Table I were obtained, presenting the mean values and standard deviations for the eight evaluated classifiers.

TABLE I - Performance metrics (mean and standard deviation) of the classifiers on the training set

Classifier	Accuracy	Kappa coefficient	Sensitivity	Specificity	ROC area
J48	79.80 ± 0.914	0.775 ± 0.010	0.842 ± 0.024	0.981 ± 0.003	0.930 ± 0.011
RF 100	91.70 ± 0.620	0.907 ± 0.007	0.947 ± 0.015	0.992 ± 0.002	0.994 ± 0.003
RF 200	91.90 ± 0.601	0.910 ± 0.006	0.948 ± 0.015	0.993 ± 0.002	0.995 ± 0.003
RF 300	92.00 ± 0.594	0.910 ± 0.006	0.948 ± 0.014	0.993 ± 0.002	0.995 ± 0.003
RF 400	92.00 ± 0.594	0.911 ± 0.006	0.948 ± 0.014	0.993 ± 0.002	0.995 ± 0.002
RF 500	79.70 ± 0.837	0.911 ± 0.006	0.948 ± 0.014	0.993 ± 0.002	0.995 ± 0.002
RT	79.40 ± 1.040	0.775 ± 0.009	0.830 ± 0.026	0.981 ± 0.003	0.910 ± 0.013

RF: Random Forest; RT: Random Tree.

The quantitative analysis of the data presented in Table I demonstrates distinct performance levels among the algorithms. The Random Forest classifiers consistently outperformed the single-tree algorithms (J48 and Random Tree). While J48 and Random Tree achieved mean accuracies of 79.8% and 79.7%, respectively, all Random Forest models achieved accuracies above 91.7%. An increase in mean performance was observed as the number of trees increased from 100 to 200, with the results stabilizing (saturating) from 200 trees onward, reaching a maximum accuracy of 92.0%.

A similar pattern was observed for the Kappa Index, with Random Forest values ($k \approx 0.907$-0.911) being classified as exhibiting "almost perfect" agreement according to the scale proposed by Jacob Cohen and substantially outperforming the single-tree models ($k \approx 0.775$). The Area Under the ROC Curve (AUC) was also notably superior for the Random Forest models (AUC > 0.994), indicating an almost ideal distinction among gesture classes.

The mean Sensitivity and Specificity metrics were consistently higher for the multi-tree models, with values close to 0.948 and 0.993, respectively. This result is clinically relevant because high sensitivity ensures that a user's or patient's motor commands are correctly recognized (minimizing false negatives), whereas high specificity prevents unintended activations in assistive systems (minimizing false positives).

The robustness of the Random Forest models is further evidenced by their low standard deviations across all metrics, especially when compared with the J48 and Random Tree algorithms. This indicates lower variability and greater stability in model performance across different validation data subsets.

B. Statistical Analysis of Performance Differences

To investigate whether the observed differences in classifier performance were statistically significant, the consolidated Kappa Index and ROC Area results were submitted to the Friedman test, a non-parametric test suitable for multiple comparisons involving related samples. Since the null hypothesis was rejected ($p < 0.001$ for both metrics), the Durbin-Conover post-hoc test was subsequently applied to identify specific classifier pairs with significant differences. The detailed results are presented in Tables II and III.

TABLE II - Results of the Durbin–Conover post-hoc test for the Kappa Index

Classifier 1	Classifier 2	Statistic	p
Random Tree	Random Forest 100	13.602	<.001
Random Tree	Random Forest 200	19.325	<.001
Random Tree	Random Forest 300	24.109	<.001
Random Tree	Random Forest 400	25.079	<.001
Random Tree	Random Forest 500	26.705	<.001
Random Tree	J48	00.250	0.803
Random Forest 100	Random Forest 200	05.722	<.001
Random Forest 100	Random Forest 300	10.507	<.001
Random Forest 100	Random Forest 400	11.476	<.001
Random Forest 100	Random Forest 500	13.102	<.001
Random Forest 100	J48	13.853	<.001
Random Forest 200	Random Forest 300	04.784	<.001
Random Forest 200	Random Forest 400	05.754	<.001
Random Forest 200	Random Forest 500	07.380	<.001
Random Forest 200	J48	19.575	<.001
Random Forest 300	Random Forest 400	00.969	0.333
Random Forest 300	Random Forest 500	02.595	0.010
Random Forest 300	J48	24.359	<.001
Random Forest 400	Random Forest 500	01.626	0.104
Random Forest 400	J48	25.329	<.001
Random Forest 500	J48	26.955	<.001

TABLE III - Results of the Durbin–Conover post-hoc test for the AUC

Classifier 1	Classifier 2	Statistic	p
Random Tree	Random Forest 100	17.927	<.001
Random Tree	Random Forest 200	21.981	<.001
Random Tree	Random Forest 300	23.608	<.001
Random Tree	Random Forest 400	25.036	<.001
Random Tree	Random Forest 500	24.208	<.001
Random Tree	J48	05.138	<.001
Random Forest 100	Random Forest 200	04.054	<.001
Random Forest 100	Random Forest 300	05.681	<.001
Random Forest 100	Random Forest 400	07.108	<.001
Random Forest 100	Random Forest 500	06.280	<.001
Random Forest 100	J48	12.789	<.001
Random Forest 200	Random Forest 300	01.627	0.104
Random Forest 200	Random Forest 400	03.055	0.002
Random Forest 200	Random Forest 500	02.227	0.026
Random Forest 200	J48	16.843	<.001
Random Forest 300	Random Forest 400	01.427	0.154
Random Forest 300	Random Forest 500	00.599	0.549
Random Forest 300	J48	18.470	<.001
Random Forest 400	Random Forest 500	00.828	0.408
Random Forest 400	J48	19.897	<.001
Random Forest 500	J48	19.069	<.001

The statistical analysis corroborates the initial observations. For the Kappa Index, the vast majority of comparisons between a Random Forest classifier and a single-tree classifier (J48 or Random Tree) showed statistical significance ($p < 0.001$). However, one particular finding deserves attention: although the mean Kappa values for the Random Forest models with 200 ($k = 0.910$) and 300 ($k = 0.910$) trees were virtually identical, the post-hoc test indicated a statistically significant difference between them ($p < 0.010$).

This apparent discrepancy between identical means and statistical significance can be explained by considering the nature of the analyses performed. Table I reports the central means of the 100 executions. In contrast, the Friedman test and its post-hoc procedure do not simply compare means, but rather the complete distribution of ranks assigned to each of the 100 executions of each classifier. Thus, it is plausible that throughout the 100 iterations, the RF300 classifier consistently achieved slightly higher ranks than RF200, even though the final mean values were identical. This small yet consistent difference in relative performance across all cross-validation repetitions is captured by the statistical test, resulting in a significant p-value.

In contrast to the statistical significance observed for the Kappa Index, the AUC analysis showed that differences among Random Forest models with adjacent numbers of trees (e.g., RF200 vs. RF300, p = 0.104; RF300 vs. RF400, p = 0.154) were not statistically significant, reinforcing the observation that performance gains saturate beyond 200 trees.

C. Processing Time Analysis of the Classifiers

To support the selection of the most suitable model, computational cost was considered in addition to predictive performance. Training times increased proportionally with the number of trees: the Random Forest model with 200 trees required 31,275 ms, whereas the model with 300 trees required 47,723 ms, representing an increase of approximately 52%. Testing times followed the same trend, as shown in Table IV.

TABLE IV - PROCESSING TIME OF THE CLASSIFIERS

Classifiers	Training time (ms)	Testing time (ms)
J48	22476	5.1
Random Forest 100	15722	109
Random Forest 200	31275	247
Random Forest 300	47723	414
Random Forest 400	62900	603
Random Forest 500	79105	790
Random Tree	00248	3.5

D. Model Selection and Validation on the Test Set

Based on the aggregated analysis of performance, stability, statistical significance, and computational cost (lower processing time), the Random Forest classifier with 200 trees was selected for the final validation stage. The choice of RF200 over configurations with a larger number of trees was based on the principle of parsimony: this model reached a performance plateau, presenting mean metrics equivalent to those of configurations with a greater number of trees (as shown in Table I), but with lower computational cost and inference time. As advocated by the principle, when equally valid explanations are available, less complex models are preferred. Furthermore, reduced computational cost is a critical factor for the development of real-time applications, such as brain-machine interfaces or biofeedback systems for motor rehabilitation.

The RF200 model was then subjected to a final critical evaluation on the independent test set (20% of the data). The results, presented in Table V, reveal solid performance, although lower than that observed during cross-validation, indicating a typical generalization scenario with controlled performance degradation.

The analysis of performance on the test set reveals that the model maintained robust discriminative capability, although with reduced classification precision. The accuracy of 78.87%, Kappa Index of 0.765, and sensitivity of 0.789 represent a significant decrease compared with the cross-validation values (91.9%, 0.910, and 0.948, respectively), yet still constitute clinically relevant performance for the complex task of classifying 10 hand gestures.

TABLE V - COMPARISON OF THE RF200 CLASSIFIER PERFORMANCE METRICS BETWEEN TRAINING AND TESTING

Metric	Training set	Test set	Variation
Accuracy	91.90%	78.87%	-13.03%
Kappa coefficient	0.910	0.765	-0.145
ROC area	0.995	0.962	-0.033
Sensitivity	0.948	0.789	-0.159
Specificity	0.993	0.977	-0.016

TABLE VI – DETAILED CLASS-WISE ACCURACY OF THE RF200 CLASSIFIER ON THE TEST SET

Class	TPR	FPR	Precision	F1 score	MCC	ROC area	PRC area
Gesture 1	0.883	0.025	0.759	0.839	0.821	0.974	0.915
Gesture 6	0.837	0.027	0.779	0.807	0.785	0.972	0.896
Gesture 7	0.873	0.027	0.783	0.826	0.806	0.979	0.904
Gesture 10	0.796	0.024	0.787	0.792	0.768	0.972	0.866
Gesture 16	0.802	0.018	0.831	0.816	0.796	0.956	0.866
Gesture 18	0.694	0.028	0.737	0.714	0.684	0.946	0.788
Gesture 19	0.698	0.026	0.752	0.724	0.695	0.946	0.772
Gesture 30	0.748	0.020	0.807	0.776	0.753	0.956	0.855
Gesture 32	0.750	0.021	0.797	0.773	0.748	0.955	0.842
Gesture 33	0.807	0.019	0.816	0.811	0.791	0.961	0.873

TPR: True Positive Rate/Sensitivity; FPR: False Positive Rate; MCC: Matthews Correlation Coefficient; ROC area: Area under the Receiver Operating Characteristic curve for class C; PRC area: Área under the Precision–Recall curve.Área sob a curva Precision–Recall.

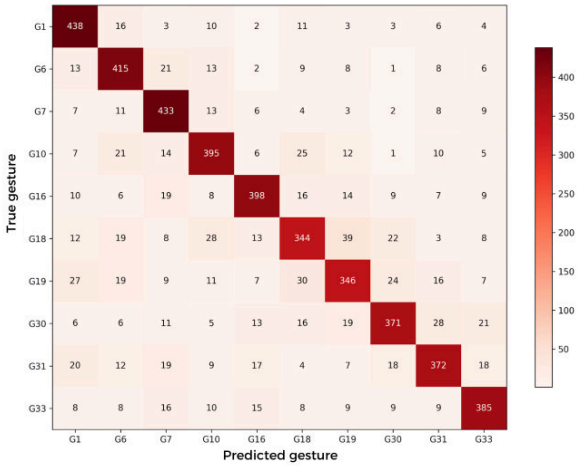


Figure IV. Confusion matrix of the RF200 classifier performance on the test set

The analysis of the Confusion Matrix (Figure IV) and the detailed class-wise accuracy table (Table VI) reveals important patterns regarding model behavior. Gestures 1, 6, and 7 exhibited the best results, with sensitivity values above 83%, indicating that the model can consistently identify these muscle activation patterns. In contrast, gestures 18 and 19 showed the lowest sensitivity values (69.4% and 69.8%, respectively), suggesting that these movements share similar electromyographic characteristics with other gestures, leading to greater classification confusion.

Based on the confusion matrix analysis, the most pronounced classification errors occurred between antagonistic gestures sharing similar muscle groups, particularly pronation and supination with a closed hand (gestures 18 and 19), which

accounted for 69 reciprocal misclassifications, representing the greatest challenge for the classifier. This pattern was also observed, although to a lesser extent, in the hand opening and hand closing pair (gestures 30 and 31), with 46 misclassifications, and in wrist flexion and wrist extension (gestures 6 and 7), with 32 misclassifications. In addition to these antagonistic pairs, other significant confusions following a functional logic were identified, such as wrist supination (gesture 10) being confused with wrist flexion (gesture 6) in 21 instances, and pronation with a closed hand (gesture 18) presenting considerable confusion with radial deviation (gesture 16) in 28 instances.

These patterns suggest that although the model demonstrates good discriminative capability across broad gesture categories, it faces difficulties in distinguishing subtle variations within similar functional groups, where muscle activation patterns overlap substantially. This finding is particularly relevant for rehabilitation applications, as it indicates that certain antagonistic gestures may require specific model training strategies or even the consolidation of functional classes to ensure greater robustness in human-machine interfaces.

Furthermore, the AUC metric remained high (mean value of 0.962), indicating that the model preserved strong discriminative capability among classes despite the reduction in pointwise accuracy. This is a positive indicator for clinical applications, where confidence in gesture discrimination is crucial.

In summary, the obtained results validate the feasibility of the proposed approach, with the 200-tree Random Forest classifier demonstrating robust performance in discriminating hand gestures from HD-sEMG heat maps, achieving an accuracy of 78.87% on validation using an independent test set. The analysis of confusion patterns provides valuable insights for future refinements, particularly regarding the discrimination of antagonistic gestures with overlapping muscular synergies.

V. Conclusion

The present study introduced a hybrid approach for hand gesture classification based on High-Density Surface Electromyography (HD-sEMG) signals, combining signal conversion into image representations (heat maps), deep feature extraction using a pre-trained Convolutional Neural Network (LeNet), and classification through decision tree-based algorithms in the WEKA environment. The methodology was evaluated using the publicly available Hyser dataset, composed of healthy participants. Ten functionally relevant gestures were selected for their potential in motor rehabilitation applications. Accordingly, the results provide a methodological foundation for future clinical validation.

The results demonstrated that ensemble-based models, particularly Random Forest, consistently outperformed single decision trees (J48 and Random Tree), achieving mean accuracy above 91.7%, Kappa Index values higher than 0.907, and Area Under the ROC Curve values greater than 0.994 during cross-validation. Statistical analysis using the Friedman test and the Durbin-Conover post-hoc test confirmed significant differences among classifiers, with the Random Forest models exhibiting statistically superior performance.

Considering the balance between predictive performance and computational cost, the Random Forest model with 200 trees (RF200) was selected for final validation. When applied to an independent test set, the RF200 model maintained robust discriminative capability, although with reduced performance metrics, achieving an accuracy of 78.87%, a Kappa Index of 0.765, and a mean AUC of 0.962. Analysis of the confusion matrix revealed that the greatest misclassifications occurred between antagonistic gestures with similar muscle activation patterns, such as pronation and supination with a closed hand, highlighting the model's difficulty in discriminating subtle variations within overlapping functional groups.

The results reinforce the potential of the proposed approach for technology-assisted motor rehabilitation applications, where the combination of HD-sEMG visual representations with classifiers that offer a structurally interpretable decision framework provides a feasible path toward the development of accessible and clinically relevant human–machine interfaces. The presented methodology constitutes a methodological contribution that integrates the informational richness provided by HD-sEMG, the representational capability of deep learning, and the rule-based structure of tree-based models, providing a starting point for explainability analysis. This aligns with the growing demand for rehabilitation support systems that are both effective and amenable to explainability analysis.

The data split was performed at the instance level, following standard practice in exploratory machine learning studies. The reported metrics therefore reflect the model's discriminative capability under controlled conditions, rather than its generalization to unseen participants. Given the well-documented inter-subject variability in HD-sEMG signals, subject-wise validation constitutes a necessary next step for assessing the robustness of the proposed approach in realistic deployment scenarios.

Future work will focus on expanding the hybrid approach to include a larger number of classifiers, enabling a broader comparison of their performance with that of decision tree-based methods. In addition, other deep learning architectures will be explored for feature extraction to evaluate whether alternative representations can improve the discrimination of antagonistic gestures. Finally, the model will be validated on independent datasets using a subject-wise split, ensuring that no data from test participants are used during training.

DECLARATION ON THE USE OF ARTIFICIAL INTELLIGENCE

The DeepSeek large language model was employed exclusively for orthographic and grammatical editing of the manuscript. No artificial intelligence was involved in the conception of the research, methodological design, data collection or analysis, interpretation of results, or generation of intellectual content. All AI-assisted processes and outputs were rigorously validated and audited by the human authors, who assume full responsibility for the accuracy, originality, and

scientific integrity of this work.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE).

REFERENCES

[1] C. O'Connell, M. Guo, B. Soucy, M. Calder, J. Sparks, and S. Plamondon, "All hands on deck: The multidisciplinary rehabilitation assessment and management of hand function in persons with neuromuscular disorders," *Muscle & Nerve*, vol. 71, no. 5, pp. 869–888, May 2025, epub 2024 Jun 6. PMID: 38845187; PMCID: PMC11998968

[2] K. Kiyono, S. Tanabe, S. Hirano, T. Ii, Y. Nakagawa, K. Tan, E. Saitoh, and Y. Otaka, "Effectiveness of robotic devices for medical rehabilitation: An umbrella review," *Journal of Clinical Medicine*, vol. 13, no. 21, 2024. [Online]. Available: https: //www.mdpi.com/2077-0383/13/21/6616

[3] L. F. Aguiar, "Reconhecimento de gestos de membros superiores usando eletromiografia de superfície para uso em reabilitação," Dissertação de Mestrado, Universidade de Brasília, 2017.

[4] K. Ito et al., "Electromyography‐controlled gamified exercise system for the distal upper extremity: a usability assessment in subacute post‐stroke patients," *Disability and Rehabilitation: Assistive Technology*, vol. 18, no. 6, pp. 883–888, Jun. 2021, doi: 10.1080/17483107.2021.1936663.

[5] S. M. Hatem et al., "Rehabilitation of motor function after stroke: a multiple systematic review focused on techniques to stimulate upper extremity recovery," *Frontiers in Human Neuroscience*, vol. 10, p. 442, Sep. 2016, doi: 10.3389/fnhum.2016.00442.

[6] G. Corvini, M. Arvanitidis, D. Falla, and S. Conforto, "Novel metrics for high‐density sEMG analysis in the time–space domain during sustained isometric contractions," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 5, pp. 760–768, 2024, doi: 10.1109/OJEMB.2024.3449548.

[7] E. Shandiz et al., "Applying high‐density surface EMG to the study of neuromuscular disorders: a systematic review," *Clinical Neurophysiology*, vol. 179, p. 2110983, Nov. 2025, doi: 10.1016/j.clinph.2025.2110983.

[8] J. Fan, J. Wen, and Z. Lai, "Myoelectric pattern recognition using Gramian angular field and convolutional neural networks for muscle–computer interface," *Sensors*, vol. 23, no. 5, p. 2715, Mar. 2023, doi: 10.3390/s23052715.

[9] J. Chen, S. Bi, G. Zhang, and G. Cao, "High-density surface EMG-based gesture recognition using a 3D convolutional neural network," *Sensors*, vol. 20, no. 4, p. 1201, Feb. 2020, doi: 10.3390/s20041201.

[10] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Hand gesture recognition using instant high-density EMG graph via deep learning method," in *2020 Chinese Automation Congress* (CAC), Shanghai, China, 2020, pp. 5143–5148, doi: 10.1109/CAC51589.2020.9326536.

[11] N. Tacca, C. Dunlap, S. P. Donegan, et al., "Wearable high‐density EMG sleeve for complex hand gesture classification and continuous joint angle estimation," *Scientific Reports,* vol. 14, p. 18564, 2024, doi: 10.1038/s41598-024-64458-x.

[12] Q. Hu, G. A. Azar, A. Fletcher, S. Rangan, and S. F. Atashzar, "ViT-MDHGR: Cross-day reliability and agility in dynamic hand gesture prediction via HD-sEMG signal decoding," *IEEE J. Sel. Top. Signal Process*., vol. 18, no. 3, pp. 419–430, Apr. 2024, doi: 10.1109/JSTSP.2024.3402340.

[13] X. Chen, H. Yang, D. Zhang, X. Hu, and P. Xie, "Hand Gesture Recognition Based on High-Density Myoelectricity in Forearm Flexors in Humans," *Sensors*, vol. 24, p. 3970, 2024, doi: 10.3390/s24123970.

[14] M. Rojas-Martínez, M. A. Mañanas, and J. F. Alonso, "High-density surface EMG maps from upper-arm and forearm muscles," *Journal of Neuroengineering and Rehabilitation*, vol. 9, no. 1, pp. 1–17, Dec. 2012, doi: 10.1186/1743-0003-9-85.

[15] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface EMG: an update," *Journal of Applied Physiology,* vol. 117, no. 11, pp. 1215–1230, Dec. 2014, doi: 10.1152/japplphysiol.00162.2014.

[16] R. Merletti and D. Farina, Surface Electromyography: *Physiology, Engineering, and Applications*. New York: The Institute of Electrical and Electronics Engineers, 2016.

[17] K. E. Ch Vidyasagar, K. Revanth Kumar, G. N. K. Anantha Sai, M. Ruchita, and M. J. Saikia, "Signal to image conversion and convolutional neural networks for physiological signal processing: A review," *IEEE Access*, vol. 12, pp. 66726–66764, 2024, doi: 10.1109/ACCESS.2024.3399114.

[18] E. B. Hunt, J. Marin, and P. J. Stone, *Experiments in Induction*. New York: Academic Press, 1966.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/a:1010933404324.

[20] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, pp. 235–240, 1994, doi: 10.1007/BF00993309

[21] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare," *in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2015, pp. 1721–1730, doi: 10.1145/2783258.2788613.

[22] P. A. Genuino, "Sistema inteligente para apoio ao diagnóstico de doenças de pele usando imagens e redes neurais artificiais profundas," M.S. thesis, Univ. Federal de Pernambuco, Recife, Brasil, 2024. [Online]. Available: https://repositorio.ufpe.br/handle/123456789/64345. [Accessed: may. 10, 2026].

[23] M. Santana, J. Seixas, A. Melo, C. Xavier de Medeiros, W. Machado, and W. Dos Santos, "Enhancing EEG-based ASD detection using wavelet transforms and hybrid deep learning models," in [Autism Diagnosis], 2026, pp. 110–124, doi: 10.1201/9781003531418-6.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun*. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," *in Proc. 27th Int. Conf. Mach. Learn.* (ICML), 2010.

[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management,* vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[29] T. Fawcett, "Introduction to ROC analysis," *Pattern Recognit.* Lett., vol. 27, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[30] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[31] R. L. Iman and W. J. Conover, "The use of the rank transform in regression," *Technometrics*, vol. 21, no. 4, pp. 499–509, Nov. 1979, doi: 10.1080/00401706.1979.10489820.

[32] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.,* vol. 7, pp. 1–30, Dec. 2006.

[33] X. Jiang, C. Dai, X. Liu, and J. Fan, "Open access dataset and toolbox of high-density surface electromyogram recordings," *PhysioNet*, Dec. 2023, doi: 10.13026/hxan-pe94.

High-speed Data Transfer Architecture in Portable Handheld Ultrasound Device using PCIe

Victoria Leite
Medical Device Department - DDM
Eldorado Research Institute
Campinas, Brazil
victoria.leite@eldorado.org.br

Tiago Dezotti
Medical Device Department - DDM
Eldorado Research Institute
Campinas, Brazil
tiago.dezotti@eldorado.org.br

Matheus Nogueira
Medical Device Department - DDM
Eldorado Research Institute
Campinas, Brazil
matheus.nogueira@eldorado.org.br

Bruno Dutra
Medical Device Department - DDM
Eldorado Research Institute
Campinas, Brazil
bruno.dutra@eldorado.org.br

Gustavo Morais
Medical Device Department - DDM
Eldorado Research Institute
Campinas, Brazil
gustavo.morais@eldorado.org.br

***Abstract*—Nowadays, portable handheld ultrasound systems continuously demand higher acquisition and processing capabilities, while maintaining the power and resource constraints imposed by embedded platforms. Heterogeneous ARM–FPGA architectures have emerged as an effective solution by combining the flexibility of embedded processors with the parallel processing capabilities of FPGAs. Consequently, a high-speed communication interface between these subsystems becomes essential. This work presents a PCIe-based Scatter-Gather DMA data transfer architecture for a portable embedded ultrasound platform composed of an NXP i.MX8M Plus processor and a AMD Artix-7 XC7A200T FPGA. The proposed architecture employs the AMD XDMA IP core and Linux XDMA driver to implement high-speed DMA communication between the FPGA and the embedded processor. Experimental tests were performed using data generated by the FPGA to simulate the ultrasound acquisition. The PCIe communication achieved a sustained throughput of up to 122 MB/s, exceeding the 40.6 MB/s required by the target portable handheld ultrasound application.**

***Index Terms*—PCIe, FPGA, Ultrasound, Embedded System**

I. Introduction

Portable handheld ultrasound systems have increasingly adopted heterogeneous embedded architectures that combine ARM processors and FPGAs to achieve real-time imaging under strict power and resource constraints [1] [2]. In these systems, the FPGA is commonly responsible for data acquisition and low-latency processing tasks, while the embedded processor performs system control, image processing, and visualization or streaming.

In ultrasound devices, as the number of acquisition channels and supported imaging modalities increases, the communication bandwidth between acquisition and processing devices becomes a critical design requirement [3]. Under real-time operating conditions, the system must sustain dozens of MB/s of continuous data throughput. High-speed interfaces such as USB 3.0, Ethernet and PCI Express (PCIe) can support these data rates [4], but present different trade-offs in terms of latency, implementation complexity, and power consumption.

In the ARM-FPGA architecture considered in this work, based on the NXP i.MX8M Plus and AMD Artix-7, USB 3.0, Ethernet, and PCI Express were evaluated as possible communication interfaces. USB 3.0 provides sufficient bandwidth for the target application, but its implementation on the FPGA side would require an additional proprietary USB 3.0 IP or a custom implementation, increasing the cost or complexity of the implementation. Ethernet also provides sufficient bandwidth and offers advantages when network connectivity or longer communication distances are required. However, its use for direct FPGA-to-processor communication introduces additional interface logic and protocol processing on the FPGA side.

PCI Express in particular, provides a direct point-to-point connection between the FPGA and the embedded processor, provides a low-latency and high-bandwidth communication interface, in addition, AMD provides the XDMA IP core and the corresponding Linux driver [5], which simplify the implementation of DMA-based communication between FPGA devices and embedded ARM platforms. The IP XDMA enables the DMA-based transfers, improving the transmission efficiency, because in this method, the transfer is made mostly by a dedicated controller for this hardware interface [6] [7]. Considering these trade-offs and the characteristics of the target hardware, PCIe was selected as the communication interface for the proposed architecture.

This work presents a firmware architecture for high-speed PCI Express communication between an AMD Artix-7 FPGA and an NXP i.MX8M Plus embedded platform through Scatter-Gather DMA based on PCI Express focusing to support the real-time data transfer in the portable handheld ultrasound device.

II. System Architecture

The proposed high-speed communication architecture is designed to support the ultrasound system, which can perform four main tasks:

- Signal generation - Transmission (TX).
- Scanline acquisition - Reception (RX).
- Image processing.
- Streaming of ultrasound images.

The proposed ultrasound platform has 86 acquisition elements, with up to 32 channels. The analog signals are sampled at 20 MHz, and the system supports imaging depths of up to 20 cm, corresponding to a maximum of 5,184 samples per scanline. Four imaging modes are supported: B, Harmonic B, Color Doppler, and Power Doppler. Under real-time operating conditions, the system must sustain approximately 40.6 MB/s of continuous data throughput to achieve 30 FPS in B and Harmonic B, 10 FPS in Color and Power Doppler.

A. Hardware Architecture

The hardware architecture is based on a heterogeneous ARM–FPGA platform composed of an AMD Artix-7 XC7A200T-L2SBG484E FPGA and an NXP i.MX8M Plus embedded processor connected through a PCI Express interface.

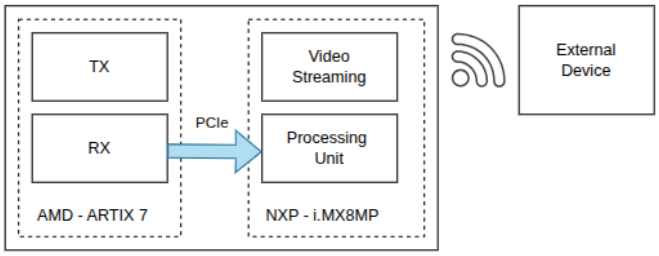


Fig. 1. System Architecture.

The FPGA is responsible for the time-critical ultrasound operations, including the generation of transmission pulses, the management of the reception chips, apply the beamforming technique on the receive data, also apply the apodization, and the management of data transfer. The acquired samples are stored in internal buffers and transferred to the embedded processor using the AMD XDMA IP core through a DMA-based PCI Express communication channel.

The i.MX8M Plus operates as the host processing platform and runs an embedded Linux operating system. On the processor side, the XDMA Linux driver is responsible for receiving the acquisition data transferred from the FPGA. The received data is then processed to generate ultrasound images, which are subsequently encoded and transmitted through a Wi-Fi connection for real-time visualization on external devices.

B. Communication Requirements

The ultrasound platform supports four acquisition modes: B, Harmonic B, Color Doppler and Power Doppler. In all cases, each acquired sample is represented using 16 bits (2 bytes).

For B and Harmonic B mode, each frame consists of 56 ultrasound scanlines. The number of samples per scanline depends on the selected imaging depth. For the maximum supported depth of 20 cm, each scanline contains 5,184 samples, resulting in 10,368 bytes per scanline and 580,608 bytes per frame. To sustain the target frame rate of 30 FPS, a throughput of approximately 17.4 MB/s is required.

The most demanding operating condition occurs during Color and Power Doppler modes. In this modes, each frame contains 56 B scanlines and 336 Doppler scanlines, resulting in a frame size of 4,064,256 bytes. At 10 FPS, the required throughput increases to approximately 40.6 MB/s.

TABLE I
COMMUNICATION BANDWIDTH REQUIREMENTS

Imaging Mode	Frame Size (Bytes)	FPS	Throughput (MB/s)
B	580,608	30	17.4
Harmonic B	580,608	30	17.4
Color Doppler	4,064,256	10	40.6
Power Doppler	4,064,256	10	40.6

As shown in Table I, the Color Doppler and Power Doppler represents the worst case communication scenario and was therefore used as the reference requirement for evaluating the PCI Express communication architecture.

III. PCI EXPRESS COMMUNICATION ARCHITECTURE

PCI Express (PCIe) is a high-speed serial communication standard widely used for data transmission. PCIe employs a point-to-point topology, providing dedicated communication resources between connected devices [7]. Each PCIe lane consists of two differential signal pairs, one for transmission and one for reception, enabling full-duplex communication. Multiple lanes can be aggregated to form wider links, such as x1, x2, x4, x8, and x16 configurations, allowing the available bandwidth to scale according to application requirements [8].

TABLE II
PCI EXPRESS THEORETICAL THROUGHPUT [8].

PCIe Version	Code	Transmission Speed	Throughput			
			x1	*x4*	*x8*	*x16*
1	8b/10b	2.5 GT/s	250 MB/s	1 GB/s	2 GB/s	4 GB/s
2	8b/10b	5 GT/s	500 MB/s	2 GB/s	4 GB/s	8 GB/s
3	128b/130b	8 GT/s	984 MB/s	3.9 GB/s	7.8 GB/s	15.7 GB/s
4	128b/130b	16 GT/s	1.969 GB/s	7.87 GB/s	15.754 GB/s	31.508 GB/s

The NXP i.MX8M Plus processor provides a PCIe Gen3 x1 controller capable of operating as either a Root Complex or an Endpoint. On the other side of communication, the AMD Artix-7 XC7A200T-L2SBG484E FPGA provides native hardware supporting a PCIe Gen2 x4 capable of operating as either a Root Complex or an Endpoint.

AMD provides three main IP cores for implementing PCIe communication in 7-Series FPGAs: the 7-Series Integrated Block for PCI Express, AXI Memory Mapped to PCI Express, and DMA/Bridge Subsystem for PCI Express (XDMA). In this work, the XDMA IP was selected because it provides a complete Scatter-Gather DMA engine, enabling efficient high-speed data transfers with minimal processor intervention [7]. Compared with a software-managed PCIe implementation, this approach reduces software complexity and frees processor

resources for ultrasound image processing and other real-time functions.

According to the minimum device requirements of the XDMA IP, the selected Artix-7 FPGA device supports only PCIe Gen1 operation and can operate only as a PCIe Endpoint [9]. Therefore, the FPGA is configured as a PCIe Gen1 Endpoint, while the i.MX8M Plus operates as the PCIe Root Complex. Since the processor provides a single PCIe lane and the negotiated link speed is limited by the lowest supported PCIe generation, the resulting communication link operates at PCIe Gen1 x1.

The PCIe configuration was limited by the choice of XDMA IP usage, which was adopted to reduce CPU intervention in data transfers and simplify the integration between the FPGA and the i.MX8M Plus. Despite this limitation, the resulting PCIe Gen1 x1 configuration provides sufficient communication bandwidth, since its theoretical maximum throughput is approximately 250 MB/s, which is above the 40.6 MB/s required by the proposed ultrasound application.

A. FPGA Endpoint Architecture

As ultrasound scanline samples are acquired, they are temporarily stored in a FIFO to synchronize the acquisition process with the data transfer subsystem. The samples are subsequently written to a dual-port RAM, in which one port is used by the acquisition of logic for writing incoming data while the second port is accessed by the XDMA engine for DMA transfers.

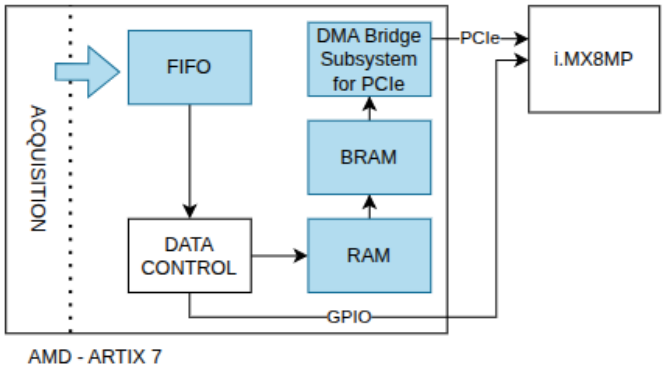


Fig. 2. FPGA Architecture.

Since PCIe transactions must be initiated by the Root Complex, the data transfer is controlled by the i.MX8M Plus processor. To synchronize both devices, a dedicated FPGA GPIO is connected to the processor and configured as a data-ready interrupt. Whenever a new set of scanline samples becomes available, the FPGA asserts the interrupt signal. Upon receiving the interrupt, the embedded processor initiates a DMA transaction through the XDMA driver to retrieve the acquired data from the FPGA.

The acquisition buffer was designed to store 28 ultrasound scanlines before initiating each DMA transfer. As a result, every PCIe transaction transfers to a fixed data block of approximately 300 kB. This organization reduces the number of PCIe transactions and decreases the frequency of GPIO interrupts generated by the FPGA. Although larger DMA transfers provide higher throughput, the buffer size represents a trade-off between communication efficiency, interrupt latency, and the available on-chip memory resources of the FPGA.

A custom data organization protocol is used to store the acquired scanlines in the FPGA RAM, as shown in Fig. 3. Each DMA buffer includes an 8-byte packet counter that allows the i.MX8M Plus to identify the transfer sequence. The buffer is further divided into scanline subpackets containing control information and padding fields to delimit each scanline. The control information identifies the ultrasound acquisition mode, scanline length, line number within the frame, and scanline number within the acquisition sequence. The scanline length depends on the configured imaging depth, while the line identification fields allow the processor to correctly reconstruct frames for B, Harmonic B, Color Doppler and Power Doppler types processing.

Each DMA transaction transfers an acquisition buffer containing up to 28 scanlines. The ultrasound frames are progressively assembled on the i.MX8M Plus from the received acquisition buffers. A B and Harmonic B frame consists of 56 scanlines and therefore requires two DMA transfers. Similarly, the Color and Power Doppler configuration contains 392 scanlines and requires 14 DMA transfers.

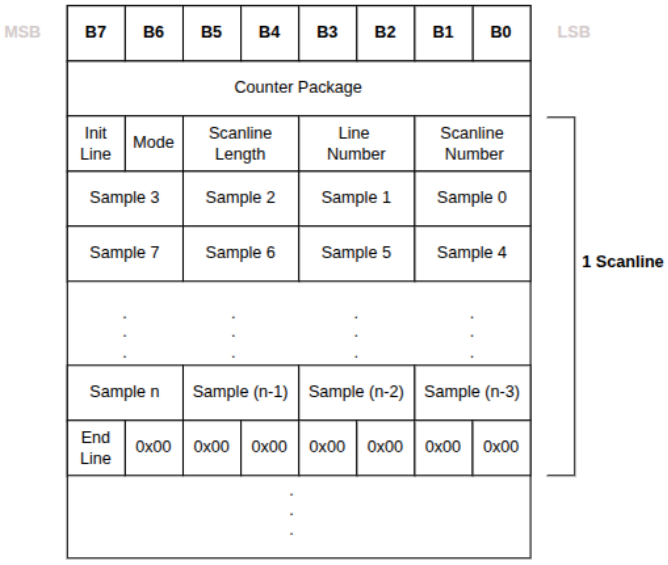


Fig. 3. RAM Memory Mapping.

B. i.MX8M Plus Root Complex Architecture

The embedded platform runs an Embedded Linux operating system generated using the Yocto Project. The AMD XDMA Linux driver XDMA was cross compiled for the ARM architecture and integrated into the system image.

The XDMA driver exposes character devices that provide user-space access to the DMA channels, including Host-to-Card (H2C) and Card-to-Host (C2H) interfaces. Since the

ultrasound data flow considered in this work is unidirectional, from the FPGA to the processor, only C2H channel is used.

A user-space application written in C was developed to manage ultrasound data after reception by the i.MX8M Plus. The FPGA delivers the ultrasound data after beamforming and apodization, organized as scanlines. In the embedded processor, the application is responsible for receiving the scanlines, performing the subsequent signal processing, and transmitting the resulting ultrasound images. Since these operations can be executed independently, the application was implemented using a multithreaded architecture divided into three main tasks: PCIe Data Manager, Signal Processing, and Video Transmission.

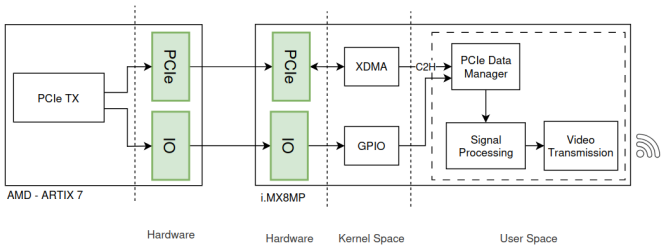


Fig. 4. i.MX8MP Architecture.

The PCIe Data Manager is responsible for communication with the FPGA and reassembly of the ultrasound frames from the received acquisition buffers. The thread waits for the data-ready GPIO interrupt and then initiates a C2H DMA transfer of 300 kB size. After each transfer, the packet counter field defined in the communication protocol as seen in Fig. 3, is checked to detect missing or out-of-sequence packets. The scanlines contained in the received buffer are then extracted and placed at their corresponding positions in the frame buffer. Once all scanlines belonging to the current frame have been received, the completed frame is forwarded to the Signal Processing thread. If a missing packet is detected, the incomplete frame is discarded and frame reassembly resumes when packets belonging to the next frame are received.

The Signal Processing thread performs the remaining processing required to generate a 640 × 480 ultrasound image. The resulting image is then forwarded to the Video Transmission thread. This thread divides each image into packets and transmits them through UDP at the target rate of 30 FPS.

IV. Results

The proposed architecture was evaluated using the hardware prototype composed of a custom board that integrates an AMD Artix-7 FPGA, connected to the variscite evaluation board "VAR-DT8MCustomBoard" through a Mini PCIe interface.

A. PCIe link

The first validation step consisted of verifying the PCIe communication between both devices. After system initialization, the PCIe link was successfully established and recognized by the embedded Linux operating system. As shown in Fig. 5, the PCIe link was negotiated as PCIe Gen1 x1.

```
root@imx8mp-us:~# lspci -v
00:00.0 PCI bridge: Synopsys, Inc. DWC_usb3 / PCIe bridge (rev 01) (prog-if 00 [Normal decode])
        Flags: bus master, fast devsel, latency 0, IRQ 227
        Memory at 18000000 (32-bit, non-prefetchable) [size=1M]
        Bus: primary=00, secondary=01, subordinate=ff, sec-latency=0
        I/O behind bridge: [disabled]
        Memory behind bridge: [disabled]
        Prefetchable memory behind bridge: 18100000-181fffff [size=1M]
        Expansion ROM at 18200000 [virtual] [disabled] [size=64K]
        Capabilities: [40] Power Management version 3
        Capabilities: [50] MSI: Enable+ Count=1/1 Maskable+ 64bit-
        Capabilities: [70] Express Root Port (Slot-), MSI 00
        Capabilities: [100] Advanced Error Reporting
        Capabilities: [148] Secondary PCI Express
        Capabilities: [158] L1 PM Substates
        Kernel driver in use: pcieport

01:00.0 Serial controller: Xilinx Corporation Device 7021 (prog-if 01 [16450])
        Subsystem: Xilinx Corporation Device 0007
        Flags: bus master, fast devsel, latency 0, IRQ 230
        Memory at 18100000 (64-bit, prefetchable) [size=64K]
        Capabilities: [40] Power Management version 3
        Capabilities: [48] MSI: Enable+ Count=1/1 Maskable- 64bit+
        Capabilities: [60] Express Endpoint, MSI 00
        Capabilities: [100] Device Serial Number 00-00-00-00-00-00-00-00
        Kernel driver in use: xdma
        Kernel modules: xdma

root@imx8mp-us:~#
CTRL-A Z for help | 115200 8N1 | NOR | Minicom 2.9 | VT102 | Offline | ttyUSB0
```

Fig. 5. PCIe link successfully established as Gen1 x1.

B. PCIe communication performance

The second step evaluated PCIe communication performance using the AMD XDMA *dma_from_device* application from the AMD XDMA Linux driver test suite, instrumented to measure each C2H DMA transaction using a monotonic clock. The measurements were used to calculate throughput, transfer latency and inter-transfer jitter. The tests were performed using transfer sizes ranging from 64 kB to 64 MB. The number of consecutive DMA transfers was adjusted according to the transfer size, resulting in a total of approximately 25 GB transferred for each configuration. In addition, CPU utilization was measured during the tests.

TABLE III
C2H XDMA PERFORMANCE RESULTS

Transfer Size	Transfer Count	Throughput		Latency		Mean Jitter (μs)	CPU Util. (%)
		Mean (MB/s)	Std. Dev. (MB/s)	Mean (ms)	Std. Dev. (μs)		
64 kB	390000	89.739	23.56	0.767	211.163	90.604	20.1
128 kB	195000	90.149	14.682	1.454	215.809	138.365	15.6
256 kB	98000	100.751	8.526	2.558	205.262	154.097	12.3
300 kB	83000	105.269	8.058	2.866	223.548	127.958	11.3
512 kB	49000	108.235	3.548	4.736	164.345	98.340	8.6
1 MB	25000	114.567	2.531	8.733	201.461	163.347	6.3
2 MB	12500	118.658	1.959	16.860	322.589	318.597	5.0
4 MB	6250	120.921	1.154	33.083	315.987	580.540	4.7
8 MB	3125	121.047	0.414	65.544	221.275	230.690	4.3
16 MB	1560	122.453	0.355	130.663	378.27	413.238	3.7
32 MB	780	122.436	0.327	261.362	699.348	770.519	3.6
64 MB	390	122.524	0.342	522.352	1463.595	1580.370	3.4

Table III summarizes the results for different DMA transfer sizes. The throughput increased with the transfer size and stabilized at approximately 122 MB/s for transfers above 8 MB. This represents 48.8% of the theoretical 250 MB/s bandwidth of a PCIe Gen1 x1 link. The difference from the theoretical limit is expected considering the hardware platform used in this work, which has more limited hardware resources than conventional PC-based architectures.

Despite the achieved throughput, the results also show that throughput variability and CPU utilization decrease with increasing transfer size, while transfer latency increases pro-

portionally to the amount of data transferred. Therefore, there is a trade-off between throughput, communication stability, CPU utilization, and latency.

To meet the system requirements, the communication must provide sufficient and stable throughput without affecting the video transmission rate, while maintaining low CPU utilization. The available processing time per frame is approximately 33 ms for B and Harmonic B mode, and 100 ms for Doppler mode. This time must be shared among the three threads implemented on the application, with signal processing being the most computationally demanding stage. Therefore, the communication stage should introduce only a small fraction of the available processing time.

To meet the minimum system latency of 33 ms in B mode, the transfer size is limited by the B mode frame size of 580,608 bytes, since a larger transfer would require data from more than one frame to complete a DMA transaction. On the other hand, increasing the transfer size also requires a larger FIFO in the FPGA to store the acquired scanlines before each DMA transfer. Considering these constraints, a transfer size of 300 kB was selected. With this configuration, each DMA transaction transfers up to 28 scanlines, corresponding to half of a B mode frame.

According to the acquisition buffer organization described previously, DMA transfers are triggered by a GPIO interrupt approximately every 16.5 ms in B mode and Harmonic B mode, and every 7.14 ms in Doppler modes. For the selected 300 kB configuration, the transfer latency remains below the shortest DMA trigger interval of 7.14 ms. Therefore, the selected transfer size provides a suitable trade-off between throughput, CPU utilization, FPGA memory usage, and transfer latency for the target application.

C. PCIe communication integration

After validating the PCIe communication subsystem, the proposed architecture was integrated into the portable handheld ultrasound application running on the i.MX8M Plus. Since the ultrasound acquisition hardware is still under development, a data generation module was implemented in the FPGA to emulate the output of the acquisition subsystem. The generated data reproduces the expected acquisition rate and buffer organization of the final ultrasound hardware, allowing the communication and processing pipeline to be evaluated under representative operating conditions.

The integrated system achieved frame rates of 29.3 FPS for B mode, 29.1 FPS for Harmonic B, 9.8 FPS for Color Doppler, and 9.2 FPS for Power Doppler. During these experiments, the CPU utilization of the i.MX8M Plus reached a maximum of 88% in the most demanding imaging mode.

The integrated system was also continuously tested for five hours to evaluate the communication reliability between the FPGA and the i.MX8M Plus. An average of 21 dropped acquisition packets per hour was observed. When a packet is lost, the corresponding incomplete frame is discarded, which may temporarily reduce the video frame rate. However, the observed packet loss rate was sufficiently low to maintain continuous video streaming during the experiments.

V. Conclusion

Experimental validation demonstrated that the proposed architecture provides sufficient communication throughput for the portable handheld ultrasound system. The PCIe communication reached a maximum throughput of approximately 122 MB/s. The communication subsystem was also successfully integrated into the ultrasound application running on the i.MX8M Plus, together with the signal processing and video streaming stages, while maintaining the target frame rates for the evaluated imaging modes.

Compared with previous PCIe-based FPGA communication architectures reported in the literature, the proposed solution achieves approximately 48.8% of the theoretical PCIe Gen1 x1 bandwidth while targeting an embedded platform with limited hardware resources. Du et al. [7] reported an efficiency of approximately 39% using a Xilinx Virtex-7 FPGA connected to a desktop PC, while Boyang [8] achieved approximately 64% with a Xilinx Kintex-7 FPGA and a desktop PC. Zhang and Li [1] also reported an efficiency of approximately 64% using an ARM-FPGA architecture based on an Artix-7 FPGA and an NVIDIA Jetson TX2. Although some of these implementations achieved higher PCIe efficiency, they were evaluated on platforms with greater processing and hardware capabilities, including Virtex-7 and Kintex-7 FPGAs, desktop PCs, and the NVIDIA Jetson TX2. In contrast, this work targets a mid-range embedded platform based on the Artix-7 and i.MX8M Plus. Despite these hardware limitations, the achieved communication bandwidth is sufficient to meet the requirements of the target ultrasound application.

The selection of the XDMA framework provides a DMA infrastructure and Linux driver support, reducing processor intervention and facilitating the integration between the FPGA and the embedded Linux platform. Although the selected XDMA limits the implemented interface to PCIe Gen1, this limitation does not affect the communication requirements of the target application. Alternative PCIe IP implementations could potentially provide higher bandwidth on compatible hardware, but would require additional implementation effort.

Future work includes integrating the proposed communication architecture with the complete ultrasound acquisition hardware, enabling validation using real ultrasound scanline data acquired from the final RX front-end.

Acknowledgment

The authors would like to thank the Eldorado Research Institute for its technical support and for providing the infrastructure that enabled the development and validation of this work.

References

[1] Z. Zhang and Z. Li, "ARM and FPGA Heterogeneous Accelerated Processing System Based on HLS and PCIe," 2021 4th International Conference on Information and Computer Technologies (ICICT), HI, USA, 2021, pp. 300-304, doi: 10.1109/ICICT52872.2021.00055.

[2] T. Matéo et al., ”A novel OEM Handheld Ultrasound Platform: from Research Innovation to Extended Portable Application Development,” 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS), Taipei, Taiwan, 2024, pp. 1-6, doi: 10.1109/UFFC-JS60046.2024.10793932.

[3] A. Vignoli et al., ”High-speed Data Transfer Architecture in Modular Ultrasound Systems for 3-D High Frame Rate Imaging,” 2025 IEEE International Ultrasonics Symposium (IUS), Utrecht, Netherlands, 2025, pp. 1-4, doi: 10.1109/IUS62464.2025.11201632.

[4] C. Muscari Tomajoli et al., ”USB 3.0 High-Transfer Rate Time-Tagging Module for High-Performance FPGA-based Time-to-Digital Converter,” 2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Italy, 2022, pp. 1-4, doi: 10.1109/NSS/MIC44845.2022.10399038.

[5] AMD, “XDMA for PCI Express Drivers and Software Guide,” AMD Adaptive Support. [Online]. Available: https://adaptivesupport.amd.com/s/article/65444. [Accessed: Jun. 22, 2026].

[6] L. Rota, M. Caselle, S. Chilingaryan, A. Kopmann and M. Weber, ”A PCIe DMA Architecture for Multi-Gigabyte Per Second Data Transmission,” in IEEE Transactions on Nuclear Science, vol. 62, no. 3, pp. 972-976, June 2015, doi: 10.1109/TNS.2015.2426877.

[7] L. Du, Z. Zhang, J. Tong, C. Wang and W. Wang, ”Research on XDMA High-Speed Data Transmission Architecture Based on PCIe,” 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 2020, pp. 1783-1787, doi: 10.1109/ICCC51575.2020.9345012.

[8] L. Boyang, ”Research and Implementation of XDMA High Speed Data Transmission IP Core Based on PCI Express and FPGA,” 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Kunming, China, 2019, pp. 408-411, doi: 10.1109/ICCASIT48058.2019.8973143.

[9] AMD, “DMA/Bridge Subsystem for PCI Express v4.1 Product Guide (PG195),” 2024. [Online]. Available: https://docs.amd.com/r/en-US/pg195-pcie-dma. [Accessed: Jun. 25, 2026].

Image-Based Evaluation System for Pilates: A Proof of Concept

Divonzir Douglas Demarchi
PPGEB/CEFID
Univ. Tecnológica Federal do Paraná
Polícia Militar do Paraná (PMPR)
Curitiba, Brazil
orcid.org/0009-0001-6411-5267

Lucas Daniel Ortiz Arias
DAINF/DAELN/CPGEI
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0009-0005-3393-4445

Mauro Aloisio Smaniotto
CPGEI
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0009-0003-9507-2022

Fabio Kurt Schneider
DAELN/CPGEI/PPGEB
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0000-0001-6916-1361

Julio Cesar Bassan
PPGEB/PPGEF
Univ. Tecnológica Federal do Paraná
Curitiba, Brazil
orcid.org/0000-0002-6910-9077

***Abstract*—This study presents a motion analysis tool applied to Pilates equipment exercises, designed to support the evaluation of movement quality through biomechanical parameters and graphical outputs. Based on a Machine Learning model (MoveNet) previously developed by the authors, the system identifies joint kinematics and generates biomechanical reports on range of motion, cadence, bilateral symmetry, and real-time feedback. Three exercises were analyzed: the lunge on the Reformer, lateral abduction and adduction, and hip flexion on the Pilates Cadillac Trapeze Table. The analysis focused on joint angles, lumbopelvic stability, and kinematic behavior across repetitions. The results revealed movement asymmetries, including reduced cadence and neuromuscular control between the lower limbs during the lunge, as identified through knee angle analysis. In lateral abduction and adduction, postural compensations were observed, reducing movement symmetry and indicating lateral dominance or deficits in hip mobility and strength. Additionally, deficits in lumbopelvic control were identified during the Stretch with Handles exercise, likely associated with insufficient activation of deep stabilizing musculature. Overall, the proposed tool provides an objective framework for biomechanical assessment in Pilates practice, offering potential applications in exercise monitoring, rehabilitation, and data-driven movement analysis aligned with current trends in biomedical engineering.**

***Index Terms*—Image analysis, MoveNet, Pilates exercises monitoring.**

I. INTRODUCTION

The Pilates method, developed by Joseph Hubertus Pilates (1883–1967), is widely recognized for its integration of body control, breath awareness, and precision in movement execution. Initially conceived as a rehabilitation technique, the method has evolved into a consolidated practice in training, rehabilitation, and health promotion settings, standing out for its focus on lumbar pelvic stability, postural alignment, and core strength development. Exercises can be performed both on the floor and on specific equipment—such as the Reformer, Cadillac, Chair, and Barrel—which offer graduated resistance and guided trajectories, allowing for greater control and safety in execution. The correct execution of the exercises is essential to achieve the benefits frequently attributed to the method, such as improved lumbar pelvic stability, increased body awareness, and optimized postural alignment. Biomechanics, in this context, becomes an indispensable tool for understanding joint and muscle behavior during movements, favoring more precise and safe interventions [1]. Deviations in joint range of motion, trunk compensations, or muscle imbalances can compromise the benefits of the method and increase the risk of overload. Biomechanical analysis in these contexts provides valuable insights into the interaction between joints, muscles, and motor patterns, allowing for individualized adjustments in exercise prescription. The analysis of human movement integrates knowledge of biomechanics, kinesiology, and motor control, allowing us to understand the behavior of the musculoskeletal system in dynamic tasks.

In Pilates, the importance of core stability and segmental alignment has been reinforced by recent studies [2], which highlight the need for precision in execution to avoid overload, especially in the lumbar region. The increasing incorporation of technologies for movement analysis, such as video capture systems, expands the potential for studying Pilates by allowing the identification of asymmetries, joint deviations, and compensatory patterns. Thus, investigating movement analysis in the Pilates method, especially with the support of technological resources, contributes to scientific, educational, and clinical advancements. Almeida Alves et al. [3] observed that biomechanics applied to Pilates makes it possible to identify inadequate joint patterns and functional limitations, contributing to personalized interventions. This understanding becomes even more robust with the use of motion analysis technologies. Lee et al. [4], for example, demonstrated that cameras and inertial sensors (IMUs) allow quantification of joint angles in different planes, with the potential for real-time correction. High-resolution cameras and body tracking soft-

ware have been explored for non-invasive analyses. Jongbae et al. [5] proposed, with high precision, recall, and f1-score, a real-time Pilates posture recognition system on a smartphone for exercise monitoring aimed to recognize 8 Pilates exercises (i.e., Bridge, Head roll-up, Hundred, Roll-up, Teaser, Plank, Thigh stretch, and Swan). Shakya et al. [6] used OpenCV and MediaPipe in order to evaluate motion analysis for Pilates ring-based exercises. They analyzed patient movements and evaluated posture accuracy by measuring the range of motion through qualitative assessments. Zhao et al. [7] employed MediaPipe and a convolutional neural network to extract features from human keypoints, and combined these with a Long Short-Term Memory (LSTM) algorithm to recognize ten Pilates poses (i.e., Criss Cross, Rolling Ball, Spine Twist, High Plank to Pike, One Hundred, Climber, Squat, Kneel Side Kick, Roll Up, and Bird Dog) in real-time video, achieving an accuracy above 94%. Although these studies have shown the capability of detecting the Pilates different actions, they are not prone for favoring longitudinal studies and objective assessments through movement details analyses. Therefore, this study aims to analyze, from a theoretical and conceptual perspective, the biomechanical fundamentals involved in the execution of Pilates movements with equipment, with emphasis on the evaluation of joint angles, postural stability, and movement precision. It seeks to discuss how movement analysis, supported by technological recording and processing resources, can contribute to technical improvement, postural correction, and injury prevention. Furthermore, the study presents foundations for the future validation of objective measurement systems in the context of Pilates.

II. MATERIALS AND METHODS

This study adopts a theoretical and exploratory approach, based on conceptual analysis of the biomechanical principles applied to the Pilates method. Essential aspects for the execution of movements on equipment were considered, including joint mobility, lumbar pelvic stability, postural alignment, and motor coordination, as indicated in research on motor control and segmental stability. The basis is the monitoring of joint angles achieved during the execution of exercises, taking as a reference standard described in biomechanical literature. The study incorporates a cohort evaluation of 18 healthy participants aged between 20 and 65 years to analyze inter-subject variability under identical experimental conditions. The sample comprised both regular practitioners and non-practitioners of physical activities, as well as individuals with and without prior experience in Pilates. The video protocol was captured using a standard smartphone camera positioned on a stable tripod, ensuring a perpendicular viewpoint of the sagittal plane recorded at 30 frames per second (FPS) with a resolution of 1080p. During the pose estimation pipeline, a confidence threshold of 0.5 was established for keypoint detection using MoveNet; missing keypoints or frames with high occlusion were handled using a linear interpolation algorithm based on the adjacent valid frames.

A. Proposed System Application

The proposed system was developed in Python, mainly using the TensorFlow and OpenCV libraries. TensorFlow is commonly used to train and apply machine-learning models for pattern recognition [8]. In this study, it was used simply to load the pre-trained model and run the inference needed to identify key body points in each video frame. The machine-learning model applied was MoveNet [8], developed by Google and available on Kaggle. MoveNet is designed to estimate human posture in images or videos by identifying key body landmarks—such as the head, shoulders, elbows, knees, and ankles—in real time. It uses compact deep-learning architectures that allow the model to operate efficiently even on devices with more limited processing capacity, including mobile phones and embedded systems. The analysis of movements in the Pilates method requires a detailed understanding of the biomechanical principles that guide the safe and efficient execution of exercises. During practice with equipment, variables such as joint range of motion, postural control, lumbar pelvic stability, and muscle activation play a determining role in the quality of movement and its impact in the health of the individual. Precise monitoring of joint angles has the potential for providing objective information to identify possible compensations and asymmetries, allowing for technical interventions and improvement of execution.

Studies reinforce that core stability and motor control are fundamental to the efficiency and safety of the method [9], [10]. Biomechanics plays a central role in the safe and efficient execution of Pilates exercises. Precise observation of joint angles allows for the identification of compensations, such as unwanted trunk rotations, excessive pelvic anteversion, or lumbar hyperextension—conditions frequently associated with pain and injuries [1]. Recent research reinforces the relevance of motor control in the practice of the method. Silva et al. [2] highlight that core stabilization is crucial not only for performance but also for the health of the lumbar spine. Similarly, Almeida Alves et al. [3] show that biomechanics helps differentiate between efficient and inefficient movement patterns, supporting therapeutic and preventive interventions.

The incorporation of motion analysis technologies significantly expands the scope of Pilates. Camera systems make it possible to identify small deviations that are not perceptible to the naked eye—such as subtle variations in pelvic tilt, internal rotation of the knees, or scapulohumeral misalignment. Studies such as those by Lee et al. [4] demonstrate that such tools offer promising resources for immediate feedback and evidence-based teaching. In this paper, the developed system intends to recognize joint points and segment angles to monitor: 1. Maximum range of motion achieved in abduction; 2. Symmetry between right and left sides; 3. Trunk and pelvic displacements; 4. Cervical and thoracic stability; 5. Execution time and cadence; 6. Initial and final isometric pauses; 7. Eccentric control in adduction. The system proposed in this study allows for measuring the rhythm of the movements, the execution time of each repetition, and the time spent in

the initial and final positions. These parameters are essential within the philosophy of the Pilates method, which emphasizes precision and control, including the ability to maintain a given position (example: for approximately 2 to 3 seconds), as traditionally guided in pedagogical practice [11], [12]. Thus, in addition to the angular aspects, the temporal component of the execution provides a more complete assessment of stability and fidelity to the proposed movement.

B. Angle Estimation and Validation

The angles between body-segment vectors were computed using the standard angle formula for two gradient vectors in 2D analytic geometry. Each vector was defined by the (x, y) joint coordinates obtained from the MoveNet model. For every pair of segments sharing a common point P_1, the corresponding vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ were defined as in Equations (1) and (2).

$$\mathbf{v}_1 = (x_2 - x_1, y_2 - y_1) \tag{1}$$

$$\mathbf{v}_2 = (x_3 - x_1, y_3 - y_1) \tag{2}$$

The absolute angle of each vector relative to the x-axis was computed using the `arctan2` function from the Python NumPy library, as shown in Equations (3) and (4). This function simultaneously evaluates both the horizontal and vertical components, ensuring correct quadrant identification and producing angles within the appropriate interval $(-\pi, \pi]$.

$$\alpha_1 = \arctan2(y_2 - y_1, x_2 - x_1) \tag{3}$$

$$\alpha_2 = \arctan2(y_3 - y_1, x_3 - x_1) \tag{4}$$

Equation (5), in radians, is used for determining the smallest angular difference θ.

$$\theta = \min(|\alpha_1 - \alpha_2|, 2\pi - |\alpha_1 - \alpha_2|) \tag{5}$$

This angle calculation method was validated using the Kinovea software [13], demonstrating strong agreement, with a mean bias of $-0.036°$, narrow limits of agreement ($-0.408°$ to $0.336°$) by Bland-Altman method [14], and minimal variability, indicating no evidence of systematic error [15]. To transition this framework into a quantitative evaluation tool, absolute angular errors across participants are evaluated relative to instructor-defined clinical targets [32]. For the Lunge on the Reformer, a target internal flexion angle of $90°$ was established for both knees, while for Lateral Abduction and Adduction, a target aperture angle of $75°$ was defined based on functional flexibility thresholds.

C. Biomechanical Report

As joint angles were calculated for each frame analyzed by the model, the results were stored in a `.csv` file containing the frame number, the corresponding video or recording time, and the angle of each joint. These data were later used to generate a biomechanical report summarizing key information for each joint, including range of motion (maximum and minimum values), indicators of movement symmetry when applicable, and graphs showing how joint angles changed over time [15].

D. Pilates Exercises

In the present work, three Pilates exercises were analysed to exemplify the usage of the tool: Lunge, Lateral Abduction/Adduction Movements on the Reformer, and Cat Stretch with Handles as presented in the following sections. All participants provided written informed consent for participation and publication of data.

1) Lunge Movement on the Reformer: The lunge movement on the Reformer is one of the examples in this study and involves placing one foot on the stable platform and the other on the moving carriage. It is generally the case the usage of springs for resistance to move the carriage in and out while maintaining a strong, upright torso and deep core engagement to work legs, gluts, and stability, with variations like low lunges (knee on carriage), reverse lunges, or explosive "flying lunges" for advanced control and strength. Therefore, the exercise begins with the practitioner in an upright position, with one foot on the fixed base and the other on the moving board. The controlled descent involves simultaneous flexion of the knees, aiming to reach approximately $90°$ in both joints.

2) Lateral Abduction and Adduction on the Reformer: Lateral abduction and adduction on the Reformer, performed standing, is an exercise that demands high motor coordination, lumbar pelvic stability, and overall postural control. Unlike other movements with a specific angular target, this exercise aims to achieve the greatest possible functional range of motion, limited by the practitioner's height, the length of the apparatus, and the spring adjustments. Thus, the quality of execution is more relevant than an absolute angular value in a movement. From a biomechanical point of view, the abduction phase requires controlled recruitment of the hip abductor muscles, with special attention to the gluteus medius, an essential muscle for pelvic stabilization and important in avoiding compensatory lateral tilts [16]. The adduction phase, in turn, represents an even greater challenge, as it requires active braking, segmental deceleration, and precise alignment of the lower limbs, preventing the return to the initial position from occurring abruptly. This eccentric control is fundamental for both strength development and joint stabilization [17]. The ability to adjust the spring resistance allows for individualizing the load, adapting the intensity to each student's specific capabilities. Stronger resistances tend to challenge eccentric control during adduction, while lighter resistances may require greater postural control during abduction. This variability is beneficial, but reinforces the importance of monitoring execution through objective parameters such as relative range of motion and trunk stability. In standing lateral abduction and adduction on the Reformer, the primary goal is to maximize functional range of motion while maintaining trunk stability and postural control. Although anthropometric factors and apparatus configuration may influence range, these are secondary to motor control and biomechanical integrity. It is important to emphasize that the proposed framework does not generate autonomous clinical text diagnoses; coordinate trajectories strictly output objective kinematic graphs while clinical interpretation is performed

by evaluators. Regarding model selection, MoveNet balances computational efficiency and joint angle accuracy for clinical biofeedback compared to OpenPose or heavy transformer architectures [7].

3) Stretch with Handles on Cadillac: Hip flexion on the Cadillac requires adequate dissociation between the hip and trunk, efficient activation of the trunk flexors, and maintenance of a neutral pelvis. The goal is to achieve approximately 90° of flexion without pelvic displacement or knee movement.

III. RESULTS AND DISCUSSION

A. Lunge Results and Discussion

Fig. 1(a) and (b) presents the Lunge exercise with the joint's angles measured after the processing by the tool and Fig. 1(c) presents the joint angle results relative the exercise session (b). Kinematic analysis from image capture enabled assessment of joint angles, movement cadence, and neuromuscular control. The mean repetition duration (∼10s), with balanced eccentric and concentric phases, is consistent with protocols targeting motor control, eccentric strength, joint stability, and increased time under tension [18], [19]. Despite appropriate cadence, qualitative deviations were observed, including prolonged end-range holds of 2–3 seconds, particularly in the posterior limb. This pattern suggests difficulty coordinating concurrent task demands and may represent a compensatory response to instability or insufficient force production, indicating deficits in intermuscular coordination, neuromuscular control, and proximal stability [20]. The quantitative analysis across participants demonstrated that the system effectively tracks multi-subject deviations from predefined instructor targets, verifying automated thresholding feasibility as supported by digital health literature [33].

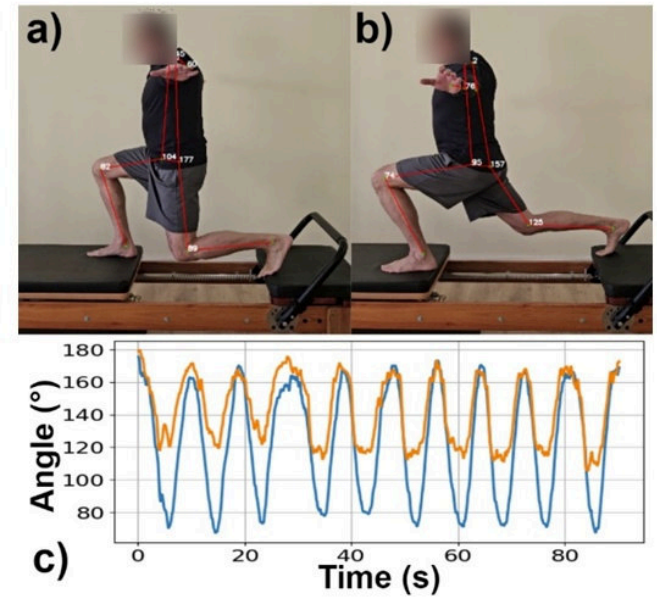


Fig. 1. (a) Lunge with the right leg at the front, (b) Lunge with the left leg at the front, (c) Knee internal angles for left (blue line) and right (orange line) legs relative the movement session (b). All images on the Reformer machine.

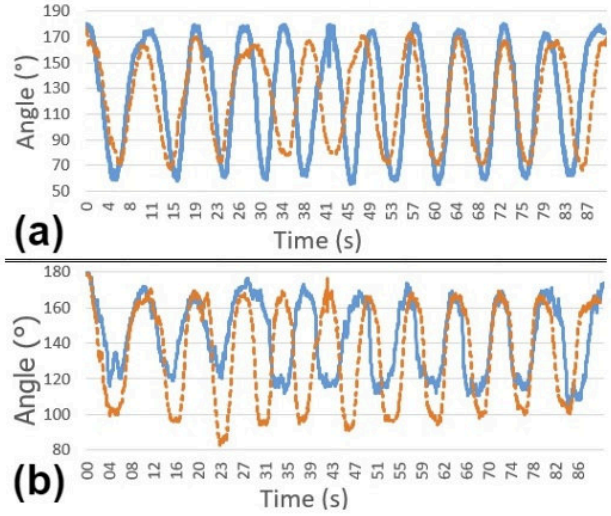


Fig. 2. (a) Right (blue line) and left (orange line) knees internal angles considering only the legs when they are in the frontal position, (b) Right (blue line) and left (orange line) knees internal angles considering only the legs when they are in the posterior position.

Additionally, the participant did not achieve and maintain the optimal knee flexion angle (∼90°) in both limbs during any repetition as shown in Fig. 1(c). This deviation suggests limitations in joint range of motion, strength, and/or neuromuscular control, potentially compromising load distribution, exercise effectiveness, and safety [21]. Stability analysis revealed increased signal oscillations, particularly in the limb on the fixed base, indicating difficulty in postural control and carriage displacement. These oscillations were most pronounced during movement initiation, suggesting impaired stabilization under unstable conditions, likely related to insufficient activation of core and hip stabilizing musculature [22], [23].

These findings indicate that, despite an appropriate cadence for motor control training, limitations persist in coordination, joint control, and postural stability. Accordingly, targeted training focusing on hip extensor and stabilizer strength and enhanced core activation and endurance is warranted to improve neuromuscular coordination, stability, and movement quality during Reformer-based exercises. Fig. 2(a) shows the left and right legs in front, there is a clear reduction in range of motion for the left foot in front compared to the right, as well as instabilities in maintaining the movement (at 28 s). Fig. 2(b) shows that for the knee angles, considering now the leg positioned behind, the left leg shows better results, possibly because with the right leg in front, the movement is more stable.

In general, the video analysis of the Reformer lunge shows transient increases in posterior-leg flexion during board displacement, likely reflecting compensatory stabilization demands. Across repetitions, reduced angular variability suggests improvements in motor control, neuromuscular balance, and movement precision, consistent with adaptive responses to functional exercise [24]. The lunge requires coordinated knee

flexion with trunk stabilization. Initial posterior-leg flexion exceeding 90° suggests compensatory frontal-plane stabilization, which decreases with repetition, indicating neuromuscular adaptation typical of closed kinetic chain tasks [17]. Trunk and pelvic oscillations and the absence of end-range isometric holds further suggest limited deep stabilizer activation and reduced integration of stability and proprioceptive control [11].

B. Lateral Abduction and Adduction Results and Discussion

Fig. 3 presents an example of lateral abduction and adduction performed with the left leg on the moving platform of the Reformer machine: (a) the initial position and (b) the final position of the leg-opening exercise. The (c) plot shows the hip angle trajectories for the left (blue) and right (orange) legs during the movement on the Reformer machine. Integrated analysis of image data and angular comparisons of the shoulders and hips enables assessment of postural control, movement symmetry, and biomechanical efficiency during lower-limb abduction and adduction on the Reformer. This exercise places high demands on dynamic stability, intermuscular coordination, and trunk control due to the mobile carriage and the need to maintain an upright posture throughout the movement.

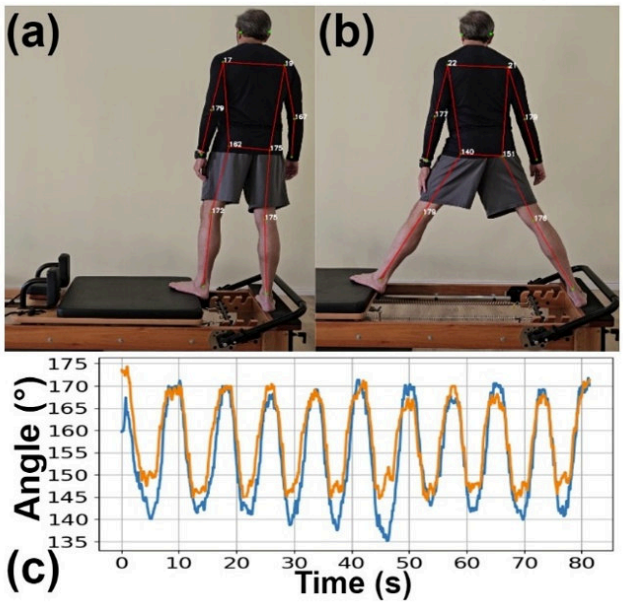


Fig. 3. Lateral Abduction and Adduction with left leg at the moving part of the Reformer showing the (a) initial and (b) final leg opening exercises. (c) Hip angles considering the left (blue) and right (orange) legs at the moving part of the Reformer Machine.

Hip angular analysis shows in Fig. 3(c) a cyclical pattern consistent with abduction and adduction phases, with a systematic side-to-side difference. The right hip exhibits higher angular values at specific time points, suggesting anterior projection and postural compensation that reduce movement symmetry. This asymmetry may reflect lateral dominance or deficits in hip mobility, strength, or motor control, particularly of the hip abductors and stabilizers, and has been associated with reduced mechanical efficiency and increased joint loading [25], [26]. Analysis of range of motion reveals variability in angular peak magnitude, indicating inconsistent movement across repetitions. The prescribed isometric hold at end-range abduction is not consistently maintained, suggesting limited capacity to sustain muscular activation under eccentric load. Irregular shoulder and hip angular patterns further indicate continuous postural adjustments in response to Reformer instability, reflecting compensatory balance strategies and highlighting the need for improved neuromuscular control and endurance of trunk and hip stabilizers [20], [27], [28]. In summary, integrated graph and image analysis indicates that, despite correct execution of the abduction-adduction pattern on the Reformer, deficits persist in movement symmetry, trunk stability, and hip position control. These findings support the need for targeted interventions to improve hip abductor strength, lumbopelvic and core stabilization, and movement consistency, with emphasis on upright posture and end-range isometric control.

C. Stretch with Handles Results and Discussion

Fig. 4 presents the (a) initial and (b) stretched positions, (c) hip angle and (d) shoulder angle for the Stretch with Handles exercises, respectively. Integrated analysis of image and angular data revealed biomechanical alterations during the Stretch with Handles exercise on the Cadillac, with failure to maintain optimal alignment of the pelvis, trunk, shoulders, and upper limbs, particularly under increased mechanical demand. This pattern reflects disruption of global kinetic chain integration, consistent with postural control principles of the Pilates method [11], [29]. Hip kinematic data show failure to maintain the reference 90° hip angle as shown in Fig. 4(c), associated with pelvic stabilization in quadruped. Continuous oscillations in left hip angulation, reflecting repeated anterior-posterior pelvic displacement, indicate deficits in lumbopelvic control, likely related to insufficient activation of deep stabilizing musculature, particularly the transversus abdominis and multifidus [31]. Pelvic oscillations impair spinal organization and postural stability, as inadequate pelvic control disrupts lumbar alignment and increases mechanical load on passive structures [16].

Shoulder angular data flexion peaks near 170–180° as shown in Fig. 4(d). The absence of stable plateaus and side-to-side differences could indicate scapular instability and asymmetric motor control, which might compromise efficient shoulder function during closed kinetic chain tasks. Angular variabilities would indicate difficulty sustaining shoulder flexion, likely due to insufficient scapular stabilizer activation and compensatory upper trapezius dominance [16]. Such instability might disrupt force transmission and global alignment, that would indicate reliance on compensatory pelvic and shoulder strategies and reflect into inadequate application of core Pilates principles, suggesting that task demands exceed the participant's neuromuscular control capacity [11], [29]. This analysis highlights the value of integrating biomechanical data with clinical observation to identify technical deficits and guide

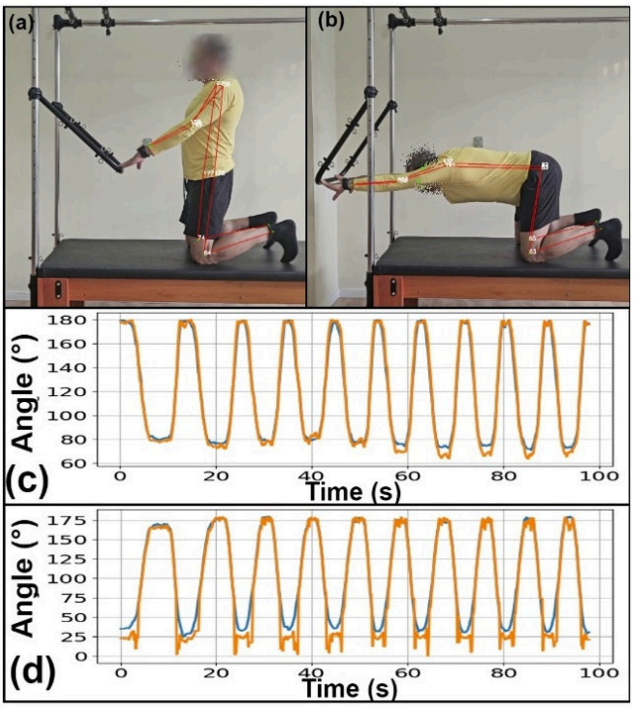


Fig. 4. (a) initial and (b) final stretched positions, (c) Hip left (blue) and right (orange) angle and (d) Shoulder left (blue) and right (orange) angle for the Stretch with Handles exercises.

safer, more effective interventions. Priority should be given to improving lumbopelvic and scapular stability before exercise progression. In general, the absence of sustained holds may reduce execution precision and reflect insufficient activation of deep stabilizers. Observed trunk compensations, including slight rightward cervical deviation, suggest asymmetries or deficits in core and shoulder-pelvic stabilizer balance, which are commonly reported during unilateral or lateral balance tasks [30].

IV. CONCLUSIONS

In general, low-cost computer vision tools can provide quantitative kinematic indicators that help monitor performance in Pilates equipment exercises, making movement analysis more accessible and objective. The proposed method allows the extracted joint angles and coordinate trajectories to be converted into clear visual reporting metrics, providing useful biofeedback to assist instructors in identifying execution patterns and guiding movement alignment.

Although this study provides a conceptual framework for motion-analysis systems and demonstrates how variations in motor control, lumbopelvic stability, and movement amplitude directly affect execution quality, subsequent investigations are planned to expand its translational value. Future works will focus on evaluating larger and more diverse clinical cohorts, specifically comparing performance variations between experienced practitioners, novices, and patient populations. Additionally, integrating multiple camera viewpoints to mitigate joint occlusion and validating these high-level indicators against optical 3D motion capture systems remain crucial steps to further consolidate the clinical utility of this reporting pipeline in real-world rehabilitation settings.

ACKNOWLEDGMENT

The authors express their gratitude for making the infrastructure of the sports practice facilities of the Paraná Military Police (PMPR) available for this research.

REFERENCES

[1] A. P. ROCHA, A. O. ANDRADE, J. L. A. CARVALHO, "Análise do torque de resistência e da força muscular resultante durante exercício de extensão de quadril no Pilates e suas implicações na prescrição e progressão," *Fisioterapia e Pesquisa*, vol. 29, no. 2, pp. 123-130, 2022.

[2] J. R. SILVA, M. L. PEREIRA, G. F. C. LIMA, "Controle motor e estabilidade do core no método Pilates: uma revisão integrativa," *Revista Brasileira de Fisiologia do Exercício*, vol. 20, no. 3, pp. 145-153, 2023.

[3] J. A. ALMEIDA ALVES, M. B. NUNES, G. A. PAZ, H. L. MIRANDA, "Nomenclatura dos exercícios estabilizadores e de fortalecimento do método Pilates: existe padronização?," *Revista Brasileira de Ciência e Movimento*, vol. 24, no. 1, pp. 145-152, 2016.

[4] M. LEE, S. KIM, D. PARK, "Prediction of Lower Limb Kinetics and Kinematics during Walking by a Single IMU on the Lower Back Using Machine Learning," *Sensors*, vol. 20, no. 1, p. 130, 2020.

[5] K. JONGBAE et al., "Real-time Pilates posture recognition system on smartphone," In: *ICOST 2023, Lecture Notes in Computer Science*, vol. 14237, pp. 3-15, 2023.

[6] A. SHAKYA, A. PAUL, R. KUMAR, N. SHARMA, A. K. SAHANI, "Automated Pilates Ring Tracking and Motion Analysis," *IEEE Access*, 2024.

[7] M. ZHAO, N. LU, Y. GUAN, "Classification of Pilates Using MediaPipe and Machine Learning," *IEEE Access*, vol. 12, pp. 77133-77140, 2024.

[8] R. VOTEL, N. LI, "Next-Generation Pose Detection with MoveNet and TensorFlow.js," *TensorFlow Blog*, 2021.

[9] N. T. B. OLIVEIRA, C. M. N. CABRAL, "Análise biomecânica do tronco e pelve em exercícios do método Pilates: revisão sistemática," *Fisioterapia e Pesquisa*, vol. 22, no. 4, pp. 443-455, 2015.

[10] P. A. O. MORAIS, "Efeitos de exercícios do método Pilates e do treinamento resistido sobre o equilíbrio de adultos jovens sedentários: um ensaio clínico randomizado controlado," Dissertação de Mestrado, Universidade Federal de Pernambuco, 2016.

[11] P. LATEY, "The Pilates method: history and philosophy," *Journal of Bodywork and Movement Therapies*, vol. 5, no. 4, pp. 275-282, 2001.

[12] A. CRUZ-FERREIRA et al., "Effects of Pilates-based exercise on life satisfaction, physical self-concept and health status in adult women," *Women Health*, vol. 51, no. 3, pp. 230-247, 2011.

[13] K. K. DALAL et al., "Effectiveness of prowling with proprioceptive training on knee hyperextension among stroke subjects using videographic observation," *Journal of Rehabilitation*, 2018.

[14] J. M. BLAND, D. G. ALTMAN, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307-310, 1986.

[15] T. M. A. SMANIOTTO, F. K. SCHNEIDER, H. S. LOPES, "A Machine Learning-Based System for Human Movement Analysis," In *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 490-498, 2026.

[16] F. P. KENDALL, E. K. MCCREARY, P. G. PROVANCE, *Muscles: Testing and Function with Posture and Pain*, 5th ed. Baltimore: Lippincott Williams Wilkins, 2015.

[17] D. A. NEUMANN, *Kinesiology of the Musculoskeletal System: Foundations for Rehabilitation*, 2nd ed. St. Louis: Elsevier, 2017.

[18] W. D. MCARDLE, F. I. KATCH, V. L. KATCH, *Fisiologia do exercício: nutrição, energia e desempenho humano*, Rio de Janeiro: Guanabara Koogan, 2016.

[19] C. KISNER, L. COLBY, *Exercícios terapêuticos: fundamentos e técnicas*, 6th ed. Barueri: Manole, 2017.

[20] D. G. BEHM, K. G. ANDERSON, "The role of instability with resistance training," *Journal of Strength and Conditioning Research*, vol. 20, no. 3, pp. 716-722, 2006.

[21] R. F. ESCAMILLA et al., "A three-dimensional biomechanical analysis of sumo and conventional style deadlifts," *Med Sci Sports Exerc.*, vol. 32, no. 7, pp. 1265-75, 2000.
[22] M. M. PANJABI, "Clinical spinal instability and low back pain," *Journal of Electromyography and Kinesiology*, vol. 13, no. 4, pp. 371-379, 2003.
[23] P. W. HODGES, C. A. RICHARDSON, "Contraction of the abdominal muscles associated with movement of the lower limb," *Physical Therapy*, vol. 77, no. 2, pp. 132–142, 1997.
[24] H. SUNG, J. KIM, J. KOO, "Neuromuscular adaptations during lower limb functional tasks with external support," *Gait Posture*, vol. 88, pp. 48-55, 2021.
[25] C. M. POWERS, "The influence of abnormal hip mechanics on knee injury," *Journal of Orthopaedic Sports Physical Therapy*, vol. 40, no. 2, pp. 42-51, 2010.
[26] M. P. REIMAN et al., "Hip function's influence on knee dysfunction," *Journal of Sport Rehabilitation*, vol. 21, no. 2, pp. 144-152, 2012.
[27] D. G. BEHM et al., "Effectiveness of traditional resistance training vs. instability resistance training," *Journal of Strength and Conditioning Research*, vol. 24, no. 12, pp. 3210-3221, 2010.
[28] V. AKUTHOTA, S. F. NADLER, "Core strengthening," *Archives of Physical Medicine and Rehabilitation*, vol. 85, no. 3, pp. 86-92, 2004.
[29] J. E. MUSCOLINO, "Pilates and the "powerhouse": II," *Journal of Bodywork and Movement Therapies*, vol. 14, no. 2, pp. 122-130, 2010.
[30] D. SANCHEZ-ZURIAGA et al., "Trunk muscle activation patterns and spine kinematics in individuals with and without chronic low back pain during functional tasks," *Clinical Biomechanics*, vol. 30, no. 7, pp. 698-703, 2015.
[31] S. MCGILL, *Low Back Disorders: Evidence-Based Prevention and Rehabilitation*, 3rd ed. Champaign: Human Kinetics, 2016.
[32] A. B. Khanghah, G. Fernie, and A. Roshan Fekr, "Design and Validation of Vision-Based Exercise Biofeedback for Tele-Rehabilitation," *Sensors*, vol. 23, no. 1, p. 452, 2023.
[33] A. Reinthal, D. Espy, and X. Luo, "Rule-Based Human Motion Tracking for Rehabilitation Exercises: Realtime Assessment, Feedback, and Guidance," *International Journal on Advanced Science Engineering and Information Technology*, vol. 16, no. 4, pp. 1203-1214, 2026.

Influência do Índice de Massa Corporal na Impedância da Pele Sob Corrente Alternada para Aplicações de Iontoforese

1st Ana Júlia de Paula Timóteo
Departamento de Engenharia Elétrica
Universidade Federal de Viçosa
Viçosa, MG, Brazil
ana.timoteo@ufv.br

2nd Anailson Raasch
Departamento de Engenharia Elétrica
Universidade Federal de Viçosa
Viçosa, MG, Brazil
anailson.raasch@ufv.br

3rd Fabiana N. Dornelles
TechPain Pesquisa & Desenvolvimento
Brazil
fabiana@techpain.com.br

4th Eduardo Souza-Silva
TechPain Pesquisa & Desenvolvimento
Brazil
edusosi@techpain.com.br

5th Leonardo B. Felix
Departamento de Engenharia Elétrica
Universidade Federal de Viçosa
Viçosa, MG, Brazil
leobonato@ufv.br

6th Rodolpho V. A. Neves
Departamento de Engenharia Elétrica
Universidade Federal de Viçosa
Viçosa, MG, Brazil
rodolpho.neves@ufv.br

***Abstract*—This study investigates the influence of body mass index (BMI) and other physiological variables on skin electrical impedance at the wrist joint in the context of iontophoresis, a promising technique for transdermal drug delivery in the treatment of chronic pain. For this purpose, data were collected from 30 volunteers (15 men and 15 women), classified into four BMI categories, who underwent impedance measurements by electrical impedance sprectrosocpy at 41 frequencies between 10 Hz em 500 kHz, using three distinct electrode positioning protocols. The data were modeled using a modified Cole equivalent circuit, incorporating a Constant Phase Element. Statistical analysis, performed using Linear Mixed Models, aimed to identify the influencing factors on impedance changes. The results reveal that the impact of BMI on skin impedance is not direct, but rather conditional, being distributed across multiple parameters of the bioimpedance model and modulated by factors such as volunteers' sex and electrode placement.**

***Index Terms*—Iontophoresis; Electrical bioimpedance; Skin impedance; Body mass index; Transdermal drug delivery; Alternating current.**

I. Introdução

O aprimoramento de técnicas terapêuticas seguras tem sido intensificado nos últimos anos, visando reduzir seus impactos físicos, mentais e funcionais [1], especialmente considerando que a prevalência dessa condição pós-pandemia nos Estados Unidos atingiu 24,3%, com destaque para dores articulares [2]. Nesse contexto, embora os anti-inflamatórios não esteroides (AINEs) sejam amplamente utilizados e eficazes no manejo dessas dores [3], seu uso prolongado está associado a graves efeitos adversos gástricos, renais e cardiovasculares [4]. Diante disso, torna-se urgente o desenvolvimento e o controle rigoroso de estratégias terapêuticas mais seguras para o tratamento.

Os autores agradecem à FAPEMIG pelo suporte financeiro.

Entre as alternativas terapêuticas em análise, a entrega transdérmica de fármacos tem se destacado devido ao seu elevado potencial terapêutico, que alia conforto, não invasividade, ação local e controle preciso da liberação do medicamento, reduzindo, assim, a exposição sistêmica e os efeitos adversos típicos das administrações orais [5]. Contudo, a principal limitação dessa via de administração é a resistência significativa oferecida pela camada mais externa da pele, o estrato córneo, que age como uma barreira seletiva à penetração de substâncias exógenas [6].

Nesse contexto, a iontoforese destaca-se como uma estratégia promissora para aumentar a permeação transdérmica por meio da aplicação de correntes elétricas com densidade de até 0,5 $\mathrm{mA/cm^2}$, favorecendo o transporte de fármacos por eletromigração e eletrosmose sem comprometer a integridade da pele e dos tecidos subjacentes [7]. Essa técnica pode ampliar a eficácia terapêutica e reduzir efeitos adversos sistêmicos no tratamento de condições dolorosas crônicas [8].

As propriedades elétricas da pele, especialmente sua impedância, influenciam diretamente a eficiência da iontoforese, pois determinam a potência necessária e a corrente efetivamente entregue ao tecido. Essa impedância varia em função de fatores como hidratação, temperatura, sexo, região anatômica e frequência da corrente aplicada [9; 10]. A bioimpedância elétrica permite caracterizar esse comportamento por meio da aplicação de correntes de baixa amplitude e da medição da oposição dos tecidos, enquanto modelos de circuitos equivalentes possibilitam estimar resistências intra e extracelulares e capacitâncias de membrana em diferentes frequências [11]. Na iontoforese, essa abordagem contribui para compreender a distribuição da corrente nas camadas cutâneas e dimensionar adequada-

mente os parâmetros do sistema de entrega transdérmica.

Diante desse contexto, este trabalho tem como objetivo investigar a influência do índice de massa corporal (IMC) e de variáveis fisiológicas sobre a impedância elétrica da pele em condições associadas à aplicação da iontoforese. Foram avaliados os efeitos da posição dos eletrodos (protocolos de coleta), da frequência de excitação utilizada na medição da bioimpedância e do sexo dos voluntários, bem como suas interações com o IMC, sobre os valores de impedância da articulação do punho. A partir da varredura de frequências, foram ajustados modelos de circuito equivalente para estimar e comparar os parâmetros elétricos da pele, contribuindo para a compreensão da distribuição da corrente e para o dimensionamento mais adequado de sistemas de entrega transdérmica.

II. Materiais e Métodos

A. Coleta do banco de dados

Os procedimentos de coleta da impedância elétrica foram realizados utilizando um sistema bipolar de dois eletrodos, constituído por um eletrodo positivo e um negativo. Para avaliar a influência do posicionamento dos eletrodos sobre os valores medidos, foram definidos três protocolos para as articulação do punho, conforme exibido na Figura 1. O eletrodo negativo foi mantido sobre o processo estiloide da ulna (marcador 1), enquanto o eletrodo positivo foi posicionado sobre o pronador quadrado proximal (marcador 2), o adutor do polegar distal (marcador 3) e o processo estiloide do rádio (marcador 4), constituindo assim, respectivamente, os Protocolos 1, 2 e 3. A coleta de dados foi aprovada pelo Comitê de Ética em Pesquisa da Universidade Federal de Viçosa, sob o CAAE 78939024.5.0000.5153.

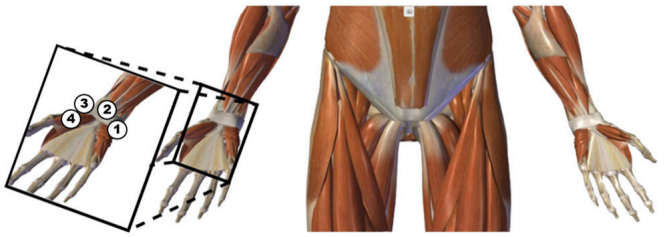
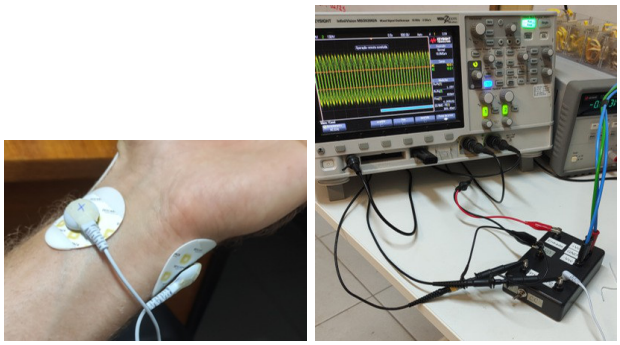


Fig. 1: Posicionamento dos eletrodos na articulação do punho: 1 - Processo estiloide da ulna; 2 - Pronador quadrado proximal; 3 - Processo estiloide do rádio e 4 - Adutor do polegar distal.

Os participantes foram recebidos na sala de coleta e, inicialmente, permaneceram em repouso por 10 minutos em ambiente climatizado (24°C), recebendo uma limpeza da articulação com álcool iodado.

Foram coletados dados de 30 voluntários, sendo 15 do sexo masculino e 15 do sexo feminino, com idade entre 19 e 40 anos e peso entre 42,7 e 94 kg. Para cada participante, cada coleta durou cerca de 340 segundos, sendo obtidos valores de impedância em 41 frequências diferentes para o punho por meio de espectroscopia de impedância elétrica (EIS), utilizando o osciloscópio MSOX2002A de 70 MHz.

A EIS foi realizada com eletrodos descartáveis de eletrocardiograma de Ag/AgCl, escolhidos por sua compatibilidade com a iontoforese e pela redução da impedância de contato com a pele. O sistema de aquisição, automatizado em *MATLAB*®, comunica-se via USB e protocolo VISA com o osciloscópio, cujo gerador de sinais excitava uma fonte de corrente de Howland melhorada. Uma foto do aparato e do posicionamento dos eletrodos do Protocolo 3 é apresentada na Figura 2. A fonte injetava corrente alternada de até $0,1\ mA_{RMS}$, enquanto o osciloscópio monitorava simultaneamente a tensão sobre a articulação e a queda de tensão em um resistor de $3,3\ k\Omega$, utilizada para estimar a corrente. A impedância foi calculada no



(a) (b)

Fig. 2: Coleta de dados nos voluntários: (a) Posicionamento dos eletrodos no Protocolo 3 e (b) equipamentos utilizados para estimulação de correntes em diferentes frequências.

domínio da frequência por meio da Transformada Rápida de Fourier, durante a varredura das 41 frequências distribuídas logaritmicamente entre 10 Hz e 500 kHz.

B. Cálculo de IMC

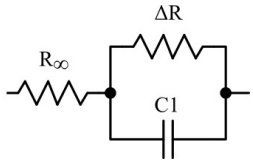
Para cada voluntário, foram registrados a altura e o peso, a partir dos quais foi calculado o IMC, definido como o peso dividido pelo quadrado da altura (kg/m^2). O cálculo do IMC e a classificação nutricional foram conferidos por meio da calculadora *on-line* disponibilizada pela Associação Brasileira para o Estudo da Obesidade e Síndrome Metabólica (ABESO).

Os participantes foram classificados em quatro categorias de acordo com o IMC: abaixo do peso (IMC $<$ $18,5\ kg/m^2$), peso normal ($18{,}5 - 24{,}9\ kg/m^2$), sobrepeso ($25{,}0 - 29{,}9\ kg/m^2$) e obesidade grau I (IMC $\geq$ $30{,}0\ kg/m^2$). Dois voluntários apresentaram baixo peso (1 do sexo masculino e 1 do sexo feminino); cinco apresentaram sobrepeso (3 do sexo feminino e 2 do sexo masculino), quatro foram classificados com obesidade grau I (3 do sexo feminino e 1 do sexo masculino) e os demais

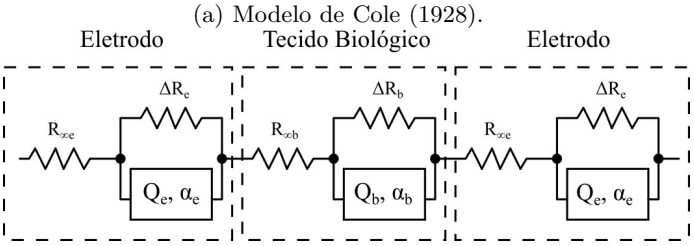
voluntários foram classificados como peso normal (8 do sexo feminino e 11 do sexo masculino).

C. Modelagem de Bioimpedância

Os modelos matemáticos de bioimpedância desempenham um papel essencial na compreensão do comportamento elétrico dos tecidos biológicos, ao modelar os parâmetros elétricos do comportamento tecidual. Entre as abordagens disponíveis na literatura, um dos primeiros foi o modelo proposto por [12], apresentado na Figura 3a, denominado de Modelo B, em que R_∞ representa a resistência dos meios intra e extracelular, ΔR denota a resistência das membranas celulares e do estrato córneo e C_1 corresponde à capacitância da membrana.



(a) Modelo de Cole (1928).



(b) Associação de modelos para representação do tecido de articulação.

Fig. 3: Circuitos equivalentes para a modelagem de impedâncias da pele: (a) Modelo de Cole (1928) e (b) representação do circuito elétrico equivalente completo, composto pela associação em série de dois modelos de eletrodo e o modelo tecidual. Adaptado de [13].

Para melhor representar a modelagem dos tecidos, foi utilizado um modelo de Cole que substitui o capacitor $C1$ por um componente denominado Elemento de Constante de Fase (Constant Phase Element – CPE) com impedância Z_{CPE} de acordo com (1), na qual Q é um coeficiente proporcional, j é a unidade imaginária, ω representa a frequência angular e α é um expoente que varia entre 0 e 1 determinando o grau de dispersão do sistema. O uso do CPE como substituto do capacitor ideal foi introduzido posteriormente ao modelo original de Cole e consolidou-se na literatura de impedância eletroquímica e bioimpedância, não sendo atribuído a um único autor [14].

$$Z_{\mathrm{CPE}}(\omega) = \frac{1}{Q(j\omega)^{\alpha}} \tag{1}$$

Os eletrodos também foram modelados e tratados em série com o tecido biológico. A Figura 3b mostra o circuito equivalente completo da associação. Os subscritos e e b foram utilizados para representar os parâmetros referentes aos eletrodos e tecido biológico (articulações), respectivamente. O comportamento dos tecidos biológicos frente à corrente elétrica varia de acordo com a frequência aplicada. Essa característica possibilita separar a impedância total medida em parâmetros específicos, que refletem os fluidos fisiológicos presentes, a capacidade dielétrica e a organização estrutural das membranas celulares.

1) Impedância em Alta Frequência (Z_{∞_b}): Corresponde à impedância quando a frequência tende ao infinito ($\omega \to \infty$).

$$Z_{\infty_b} \approx R_{\infty_b} \tag{2}$$

Fisiologicamente, em altas frequências, a reatância capacitiva das membranas celulares torna-se desprezível, permitindo que a corrente flua através dos meios intra e extracelulares. Portanto, R_{∞_b} é um indicador do volume condutivo total e da hidratação global.

2) Largura de Dispersão (ΔR_b): Representa a resistência na malha paralela ao CPE, quantificando a variação da impedância devido ao relaxamento dielétrico. Em frequências muito baixas ($\omega \to 0$), as membranas bloqueiam a corrente, fazendo com que a impedância total (Z_{0_b}) seja a soma dos componentes resistivos:

$$Z_{0_b} \approx R_{0_b} = R_{\infty_b} + \Delta R_b \tag{3}$$

3) Coeficiente de Dispersão (α_b): Este parâmetro adimensional ($0 \leq \alpha_b \leq 1$) modela o desvio do tecido em relação a um sistema homogêneo ideal ($\alpha_b = 1$). Fisiologicamente, reflete a heterogeneidade da arquitetura tecidual: como as células reais variam em tamanho, formato e orientação, elas geram uma distribuição estatística de tempos de relaxamento, em vez de uma constante única.

4) Pseudocapacitância (Q_b): Devido à rugosidade e à arquitetura fractal das membranas biológicas, o armazenamento de carga não é uniforme como num capacitor ideal. Assim, Q_b mede essa capacidade de armazenamento modulada pelo fator α_b. Sua unidade é $\mathrm{S} \cdot \mathrm{s}^{\alpha_b}$, tornando-a dimensionalmente dependente da heterogeneidade do tecido.

Para a obtenção dos parâmetros elétricos do modelo, adotou-se uma estratégia de otimização em duas etapas. Inicialmente, aplicou-se o Algoritmo Genético (GA), um método estocástico baseado em seleção natural, para realizar uma busca global ampla na superfície de erro. Através de operadores de seleção (roleta), cruzamento (*crossover*) e mutação, o GA avaliou a aptidão (*fitness*) de uma população de soluções candidatas.

Em seguida, a estimativa resultante do GA foi utilizada como ponto de partida para a Programação Quadrática Sequencial (SQP). Sendo um método iterativo local que utiliza informações de gradiente e aproximações da matriz Hessiana, a SQP refinou a solução, garantindo uma rápida taxa de convergência e alta precisão. Esta etapa final determinou os valores ótimos dos parâmetros elétricos.

D. Teste estatístico

A análise estatística foi conduzida por meio de Modelos Lineares Mistos (LMM), utilizando o *software* Jamovi (v. 2.7.11). Essa abordagem foi selecionada para acomodar a estrutura hierárquica e a dependência intrassujeito dos dados, em que as frequências das medidas estão aninhadas aos protocolos de posicionamento dos eletrodos e estes à articulação do punho. O modelo permitiu a inclusão simultânea de preditores categóricos e contínuos como efeitos fixos, controlando adequadamente as correlações residuais provenientes das medidas repetidas.

Devido à natureza não linear e à heterocedasticidade inerente aos dados de bioimpedância, os valores de impedância foram transformados em base logarítmica (log_{10}) antes da análise, assegurando a normalidade dos resíduos e a homogeneidade das variâncias, atendendo aos pressupostos dos Modelos Lineares Mistos [15]. Adicionalmente, a frequência foi submetida à transformação logarítmica, permitindo uma representação linear das variações de impedância ao longo do espectro de frequências [15].

A Figura 4 ilustra os valores de impedância obtidos para a coleta do punho utilizando o Protocolo 1 em todos os IMCs. Os valores representam a média dos valores de impedância de cada grupo de IMC juntamente ao seu respectivo desvio padrão. É possível observar que a impedância logarítmica apresenta valores de variância mais próximos do que teriam em escala linear, exemplificando a heterocedasticidade necessária para o LMM.

Na configuração final do modelo, o logaritmo da impedância foi definido como a variável dependente, enquanto o logaritmo da frequência foi inserido como covariável. Os parâmetros sexo, articulação, protocolo de coleta e IMC foram tratados como fatores de efeito fixo.

O teste LMM analisou os efeitos dos fatores sobre a impedância de forma isolada e em conjuntos, verificando a interdependência entre as variáveis. Para cada análise, reportam-se a estatística F, os graus de liberdade do efeito (df) e dos resíduos (df (res)), além do valor-p (p), adotando-se o nível de significância de 5%.

III. Resultados e discussões

A. Testes globais

Os resultados dos testes globais (*omnibus*) dos efeitos fixos e de suas interações encontram-se na Tabela I. Uma vez que a frequência de excitação foi incorporada ao modelo como preditor contínuo no eixo logarítmico (log_frequência) quando há interações envolvendo log_frequência, não são cabíveis comparações *post hoc* entre níveis (apropriadas a fatores categóricos).

1) Efeitos principais: Os testes globais indicaram efeitos principais significativos de Protocolo e Log_frequência ($p < 0{,}001$). Como esperado, mantidos os demais fatores constantes, a variação isolada de qualquer um desses preditores altera o módulo de impedância. Em contraste, IMC ($p = 0{,}997$) e Sexo ($p = 0{,}384$) não apresentaram efeito principal significativo.

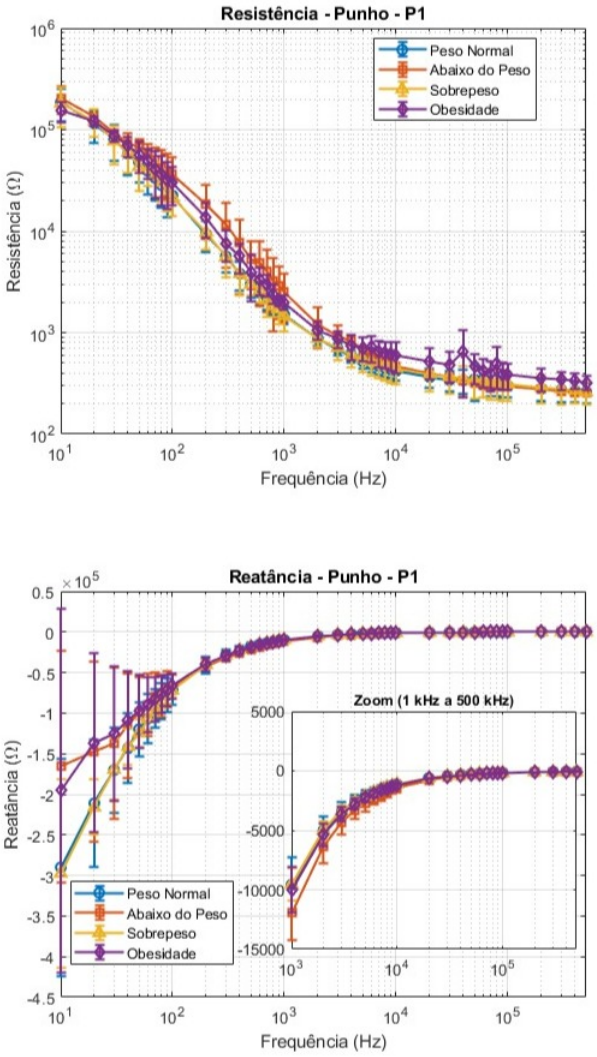


Fig. 4: Média e desvio padrão de impedância do punho (Protocolo 1) no domínio da frequência para todos os IMCs.

2) Teste 2 vias: Nos testes de duas vias que envolveram o IMC, apenas a interação com a Protocolo foi significativa ($p < 0{,}001$), indicando que a associação do IMC com a impedância varia conforme o protocolo de coleta utilizado. As interações com Sexo ($p = 0{,}417$) e Log_frequência ($p = 0{,}45$) não alcançaram significância, sugerindo que, isoladamente, o efeito do IMC não depende do sexo ou da frequência de excitação.

3) Teste 3 vias: Nos testes de três vias, observaram-se interações significativas do IMC com Sexo × Protocolo e Protocolo × Log_frequência (todas $p < 0.001$), indicando que a associação do IMC com a impedância depende simultaneamente desses fatores.

4) Teste 4 vias: No teste de quatro vias, o IMC revelou interagir com Sexo × Protocolo × Log_Frequência ($p = 0.004$), indicando que a associação entre IMC e a impedância é simultaneamente modulada pelo sexo,

pela posição do eletrodo e pela frequência da corrente de excitação.

TABLE I: Testes globais dos efeitos fixos

Efeito	F	df	df (res)	p
IMC	0,831	3	22	0,491
Sexo	0,113	1	22	0,740
Protocolo	23,635	2	3598	<0,001
Log_Frequência	7865,648	1	22	<0,001
Sexo × Protocolo	3,169	2	3598	0,042
IMC × Sexo	0,988	3	22	0,417
IMC × Protocolo	4,856	6	3598	<0,001
Sexo × Log_Frequência	0,771	1	22	0,389
Protocolo × Log_Frequência	6,558	2	3598	0,001
IMC × Log_Frequência	0,915	3	22	0,450
IMC × Protocolo × Sexo	5,956	6	3598	<0,001
Sexo × Protocolo × Log_Frequência	2,176	2	3598	0,114
IMC × Sexo × Log_Frequência	1,039	3	22	0,395
IMC × Protocolo × Log_Frequência	4,342	6	3598	<0,001
IMC × Protocolo × Sexo × Log_Frequência	3,223	6	3598	0,004

B. Análise Post Hoc

1) IMC × Protocolo: O IMC "Abaixo do peso" apresentou influenciar as impedâncias entre o protocolo 1 e 2, e entre 1 e 3. No IMC "Peso normal", a influência para o Protocolo 3 em relação aos demais.

2) IMC × Sexo × Protocolo: Para o sexo masculino, o IMC "Abaixo do peso" revelou uma influência na impedância para o Protocolo 1 em relação ao 2 e 3. Além disso, para esse mesmo sexo, o IMC "Peso normal" apresenta influência na comparação do Protocolo 3 em relação ao 1 e 2. O IMC "Peso normal" também demonstrou influenciar a comparação entre o Protocolo 1 e 3 no sexo feminino.

Esses resultados corroboram com os obtidos na interação de duas vias entre IMC × Protocolo. A adição do sexo como preditor agregou em uma análise mais precisa da influência dos fatores IMC e Protocolo.

Apesar de não apresentar efeito principal isolado, o IMC aparece em interações de ordem superior, como IMC × Protocolo × Sexo. Isso pode ser explicado pela quantidade de tecido adiposo presente na região analisada, variando entre os sexos, ou até mesmo o nível de hidratação entre eles. Embora este estudo traga contribuições consistentes, sua limitação está na baixa representatividade de voluntários em determinadas categorias de IMC compromete a precisão das médias obtidas individualmente.

C. Parâmetros de Bioimpedância

Para uma análise quantitativa, os parâmetros de bioimpedância do punho foram calculados para todos os 30 voluntários, independente de sexo ou IMC correspondentes, para os 3 protocolos de posicionamento do eletrodo. A Tabela II permite analisar os parâmetros de bioimpedância sob múltiplas perspectivas. É importante ressaltar que alguns agrupamentos possuem apenas 1 voluntário, o que influencia cálculos de média.

Dentre as alternativas possíveis, se o sexo feminino for selecionado como foco de análise no Protocolo 1, percebe-se que os valores de R_∞ apresentam médias próximas independente do IMC – média máxima de $263,68\ \Omega$ e mínima de $201,01\ \Omega$ – mas com grandes desvios ao redor desses valores – valor mínimo de $132,11\ \Omega$ e máximo de $349,58\ \Omega$.

Ainda nessa linha de análise, os valores médios de α_b do punho para o sexo feminino em Protocolo 1 de coleta apresentam valores mais próximos da homogeneidade ($\alpha_b = 1$), sendo o IMC de sobrepeso o que mais se distanciou do ideal, com uma diferença de $0,1047$. Essa observação pode ser explicada pela quantidade de voluntários que compõem esse grupo.

IV. Conclusão

A relevância do IMC em interações de ordem superior demonstra que parâmetros morfofisiológicos podem afetar os valores de impedância independente da frequência de excitação utilizada. As análises *post hoc* confirmaram a influência do IMC mesmo entre grupos com perfis de bioimpedância heterogêneos. Esse achado reforça que as variações de impedância decorrentes do IMC tendem a se distribuir por múltiplos parâmetros estimados, o que impossibilita isolar um único fator como o responsável exclusivo por essas alterações.

Entretanto, devido a limitação do estudo, para investigações futuras, sugere-se a adoção de uma amostra mais robusta e homogênea em cada faixa de IMC, fortalecendo a confiabilidade estatística e o poder de comparação dos resultados. A utilização de pelo menos 3 voluntários de cada sexo para cada IMC pode trazer valores mais confiáveis de média e desvio padrão.

Os resultados deste estudo podem nortear o dimensionamento de dispositivos direcionados a perfis específicos de IMC ou a protocolos particulares de coleta. Além disso, representam um avanço para o desenvolvimento de sistemas inteligentes de iontoforese, capazes de ajustar autonomamente a impedância de entrada em resposta às características anatômicas e fisiológicas do paciente.

References

[1] D. Riedl, J. Karnik, A. Lampe, C. Kirchhoff, K. Labek, and M. Schirmer, "Mentalizing, loneliness and pain-related depressive symptoms are associated with pain severity in patients with rheumatic diseases: Results of a cross-sectional secondary analysis," *Journal of Clinical Medicine*, vol. 14, no. 11, p. 3624, 2025.

[2] A. Zajacova, H. Grol-Prokopczyk, and R. L. Nahin, "Pain among us adults before, during, and after the covid-19 pandemic: a study using the 2019 to 2023 national health interview survey," *Pain*, pp. 10–1097, 2022.

TABLE II: Parâmetros de bioimpedância do punho.

		Protocolo 1			Protocolo 2			Protocolo 3		
Categoria	**Parâmetro**	**Média**	**Limite Máx.**	**Limite Mín.**	**Média**	**Limite Máx.**	**Limite Mín.**	**Média**	**Limite Máx.**	**Limite Mín.**
Masculino Peso normal	$R_\infty(\Omega)$	214,16	458,09	149,96	184,36	311,05	128,17	225,04	385,76	159,46
	$\Delta R(\Omega)$	$6,38 \cdot 10^5$	$9,89 \cdot 10^5$	$1,56 \cdot 10^5$	$2,46 \cdot 10^5$	$4,13 \cdot 10^5$	$8,11 \cdot 10^4$	$5,71 \cdot 10^5$	$9,99 \cdot 10^5$	$7,24 \cdot 10^4$
	$Q_c(S \cdot s^\alpha)$	$4,53 \cdot 10^{-8}$	$6,60 \cdot 10^{-8}$	$2,39 \cdot 10^{-8}$	$6,39 \cdot 10^{-8}$	$8,88 \cdot 10^{-8}$	$3,51 \cdot 10^{-8}$	$4,71 \cdot 10^{-8}$	$8,74 \cdot 10^{-8}$	$1,17 \cdot 10^{-8}$
	α_{bio}	0,9022	0,9390	0,8772	0,8503	0,8867	0,8109	0,8863	0,9157	0,8574
Masculino Abaixo do peso	$R_\infty(\Omega)$	167,45	167,45	167,45	228,11	228,11	228,11	351,78	351,78	351,78
	$\Delta R(\Omega)$	$1,83 \cdot 10^5$	$1,83 \cdot 10^5$	$1,83 \cdot 10^5$	$1,68 \cdot 10^5$	$1,68 \cdot 10^5$	$1,68 \cdot 10^5$	$3,12 \cdot 10^5$	$3,12 \cdot 10^5$	$3,12 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$4,00 \cdot 10^{-8}$	$4,00 \cdot 10^{-8}$	$4,00 \cdot 10^{-8}$	$3,55 \cdot 10^{-8}$	$3,55 \cdot 10^{-8}$	$3,55 \cdot 10^{-8}$	$2,67 \cdot 10^{-8}$	$2,67 \cdot 10^{-8}$	$2,67 \cdot 10^{-8}$
	α_{bio}	0,8630	0,8630	0,8630	0,8303	0,8303	0,8303	0,8472	0,8472	0,8472
Masculino Sobrepeso	$R_\infty(\Omega)$	171,69	203,90	139,48	246,67	302,61	190,72	180,61	188,10	173,11
	$\Delta R(\Omega)$	$8,01 \cdot 10^5$	$9,99 \cdot 10^5$	$6,02 \cdot 10^5$	$4,15 \cdot 10^5$	$4,94 \cdot 10^5$	$3,36 \cdot 10^5$	$6,78 \cdot 10^5$	$9,74 \cdot 10^5$	$3,81 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$4,15 \cdot 10^{-8}$	$4,45 \cdot 10^{-8}$	$3,85 \cdot 10^{-8}$	$4,65 \cdot 10^{-8}$	$4,98 \cdot 10^{-8}$	$4,32 \cdot 10^{-8}$	$3,57 \cdot 10^{-8}$	$4,18 \cdot 10^{-8}$	$2,97 \cdot 10^{-8}$
	α_{bio}	0,8966	0,9105	0,8828	0,8665	0,8677	0,8653	0,8957	0,9108	0,8805
Masculino Obesidade grau I	$R_\infty(\Omega)$	391,14	391,14	391,14	153,46	153,46	153,46	298,56	298,56	298,56
	$\Delta R(\Omega)$	$3,80 \cdot 10^5$	$3,80 \cdot 10^5$	$3,80 \cdot 10^5$	$3,10 \cdot 10^5$	$3,10 \cdot 10^5$	$3,10 \cdot 10^5$	$6,99 \cdot 10^5$	$6,99 \cdot 10^5$	$6,99 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$6,76 \cdot 10^{-8}$	$6,76 \cdot 10^{-8}$	$6,76 \cdot 10^{-8}$	$4,53 \cdot 10^{-8}$	$4,53 \cdot 10^{-8}$	$4,53 \cdot 10^{-8}$	$5,94 \cdot 10^{-8}$	$5,94 \cdot 10^{-8}$	$5,94 \cdot 10^{-8}$
	α_{bio}	0,8498	0,8498	0,8498	0,8759	0,8759	0,8759	0,8663	0,8663	0,8663
Feminino Peso normal	$R_\infty(\Omega)$	201,01	300,86	132,11	269,92	471,40	186,93	294,01	444,57	219,53
	$\Delta R(\Omega)$	$7,33 \cdot 10^5$	$9,96 \cdot 10^5$	$3,74 \cdot 10^5$	$4,16 \cdot 10^5$	$9,99 \cdot 10^5$	$1,66 \cdot 10^4$	$8,77 \cdot 10^5$	$1,00 \cdot 10^6$	$5,77 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$3,73 \cdot 10^{-8}$	$6,06 \cdot 10^{-8}$	$2,88 \cdot 10^{-8}$	$6,27 \cdot 10^{-8}$	$1,41 \cdot 10^{-7}$	$2,86 \cdot 10^{-8}$	$3,65 \cdot 10^{-8}$	$6,88 \cdot 10^{-8}$	$2,18 \cdot 10^{-8}$
	α_{bio}	0,9042	0,9212	0,8902	0,8481	0,9063	0,7913	0,8970	0,9210	0,8728
Feminino Abaixo do peso	$R_\infty(\Omega)$	209,88	209,88	209,88	243,91	243,91	243,91	215,53	215,53	215,53
	$\Delta R(\Omega)$	$6,91 \cdot 10^5$	$6,91 \cdot 10^5$	$6,91 \cdot 10^5$	$7,96 \cdot 10^5$	$7,96 \cdot 10^5$	$7,96 \cdot 10^5$	$8,30 \cdot 10^5$	$8,30 \cdot 10^5$	$8,30 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$3,52 \cdot 10^{-8}$	$3,52 \cdot 10^{-8}$	$3,52 \cdot 10^{-8}$	$2,96 \cdot 10^{-8}$	$2,96 \cdot 10^{-8}$	$2,96 \cdot 10^{-8}$	$4,10 \cdot 10^{-8}$	$4,10 \cdot 10^{-8}$	$4,10 \cdot 10^{-8}$
	α_{bio}	0,9016	0,9016	0,9016	0,9054	0,9054	0,9054	0,8821	0,8821	0,8821
Feminino Sobrepeso	$R_\infty(\Omega)$	263,68	349,58	182,14	233,67	312,33	192,14	249,47	291,39	221,08
	$\Delta R(\Omega)$	$7,04 \cdot 10^5$	$9,38 \cdot 10^5$	$3,56 \cdot 10^5$	$2,65 \cdot 10^5$	$4,35 \cdot 10^5$	$1,51 \cdot 10^5$	$7,61 \cdot 10^5$	$9,94 \cdot 10^5$	$4,23 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$4,45 \cdot 10^{-8}$	$6,11 \cdot 10^{-8}$	$3,53 \cdot 10^{-8}$	$5,13 \cdot 10^{-8}$	$6,39 \cdot 10^{-8}$	$4,07 \cdot 10^{-8}$	$3,87 \cdot 10^{-8}$	$5,36 \cdot 10^{-8}$	$2,74 \cdot 10^{-8}$
	α_{bio}	0,8953	0,9091	0,8801	0,8713	0,8763	0,8670	0,9123	0,9296	0,8948
Feminino Obesidade grau I	$R_\infty(\Omega)$	245,09	308,47	157,50	266,67	361,88	209,96	339,66	584,78	193,20
	$\Delta R(\Omega)$	$4,49 \cdot 10^5$	$9,85 \cdot 10^5$	$1,47 \cdot 10^5$	$3,38 \cdot 10^5$	$6,76 \cdot 10^5$	$8,99 \cdot 10^4$	$7,34 \cdot 10^5$	$9,86 \cdot 10^5$	$3,39 \cdot 10^5$
	$Q_c(S \cdot s^\alpha)$	$3,72 \cdot 10^{-8}$	$5,22 \cdot 10^{-8}$	$2,81 \cdot 10^{-8}$	$5,55 \cdot 10^{-8}$	$6,88 \cdot 10^{-8}$	$4,59 \cdot 10^{-8}$	$4,09 \cdot 10^{-8}$	$5,94 \cdot 10^{-8}$	$2,97 \cdot 10^{-8}$
	α_{bio}	0,9023	0,9162	0,8786	0,8535	0,8882	0,8218	0,8880	0,9111	0,8452

[3] M. Sisignano, A. S. Rice, and G. Geisslinger, "Topical analgesics: Pharmacology and clinical applications," *Anesthesiology*, vol. 143, no. 5, p. 1371, 2025.

[4] K. Wang and X. Li, "Comparison of cardiorenal safety of nonsteroidal anti-inflammatory drugs in the treatment of arthritis: a network meta-analysis," *Annals of Translational Medicine*, vol. 10, no. 24, p. 1388, 2022.

[5] Z. Guo, Y. Zhang, M. Zhao, W. Zhang, X. Li, F. Zhou, H. Peng, Q. Wang, and Z. Chen, "Intelligent transdermal nanoparticles as synergizing advanced delivery systems for precision therapeutics," *Materials Today Bio*, p. 102220, 2025.

[6] T. Murata, T. Honda, A. Mostafa, and K. Kabashima, "Stratum corneum as polymer sheet: Concept and cornification processes," *Trends in molecular medicine*, vol. 28, no. 5, pp. 350–359, 2022.

[7] V. Dhote, P. Bhatnagar, P. K. Mishra, S. C. Mahajan, and D. K. Mishra, "Iontophoresis: a potential emergence of a transdermal drug delivery system," *Scientia pharmaceutica*, vol. 80, no. 1, p. 1, 2011.

[8] J. F. Andrade, M. Cunha-Filho, G. M. Gelfuso, and T. Gratieri, "Iontophoresis for the cutaneous delivery of nanoentraped drugs," *Expert opinion on drug delivery*, vol. 20, no. 6, pp. 785–798, 2023.

[9] D. J. Bora and R. Dasgupta, "Estimation of skin impedance models with experimental data and a proposed model for human skin impedance," *IET Systems Biology*, vol. 14, no. 5, pp. 230–240, 2020.

[10] M. Machnia, W. Douzi, H. Jdidi, E. Miller, and B. Dugué, "Impact of regular cold exposure on electrical skin resistance in patients with joint degenerative diseases," *Clinical Physiology and Functional Imaging*, vol. 44, no. 2, pp. 164–170, 2024.

[11] S. Grimnes and O. G. Martinsen, "Cole electrical impedance model-a critique and an alternative," *IEEE transactions on biomedical engineering*, vol. 52, no. 1, pp. 132–135, 2005.

[12] K. S. Cole, "Electric impedance of suspensions of spheres," *The Journal of general physiology*, vol. 12, no. 1, p. 29, 1928.

[13] A. Raasch, "Estimação do perfil de impedâncias de articulações sinoviais para aplicação em equipamentos de iontoforese," Master's thesis, Universidade Federal de São João del Rei, 2025.

[14] D. D. Stupin, E. A. Kuzina, A. A. Abelit, A. K. Emelyanov, D. M. Nikolaev, M. N. Ryazantsev, S. V. Koniakhin, and M. V. Dubina, "Bioimpedance spectroscopy: basics and applications," *ACS Biomaterials Science & Engineering*, vol. 7, no. 6, pp. 1962–1986, 2021.

[15] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: a practical guide using statistical software*. Crc Press, 2022.

Kinematic and Multi-Feature Electromyographic Characterization of Functional Upper-Limb Tasks

Laura M. Montealegre
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
l260021@dac.unicamp.br

Ricardo G. Molinari
Center for Biomedical Engineering
University of Campinas
Campinas, Brazil
molinari@unicamp.br

Leonardo A. Elias
Dept. of Electronics and Biomedical Eng.
University of Campinas
Campinas, Brazil
leoelias@unicamp.br

***Abstract*—This study investigated the neuromechanical characteristics of two functional upper-limb tasks—the Box and Block Test (BBT) and the Southampton Hand Assessment Procedure (SHAP)—using kinematic and surface electromyographic (sEMG) analyses. Eight healthy participants performed both tasks with their dominant and non-dominant limbs. The kinematic analysis included marker traveled distance, average speed, and range of motion, whereas the sEMG analysis assessed mean absolute value (MAV), sample entropy (SampEn), the coefficient of variation of root-mean-square (RMS) values ($\mathrm{CoV_{RMS}}$), and Willison amplitude (WAMP). In both tasks, the wrist marker exhibited greater traveled distance and average speed than the shoulder and elbow markers ($p < 0.001$, $\eta_p^2 = 0.89$–0.996). The sEMG analysis revealed significant main effects of muscle during the BBT for all metrics ($p < 0.001$, $\eta_p^2 = 0.64$–0.94) and during the SHAP for MAV, SampEn, and $\mathrm{CoV_{RMS}}$ ($p < 0.001$, $\eta_p^2 = 0.72$–0.79). Furthermore, the SHAP revealed significant muscle-by-side interactions for SampEn, $\mathrm{CoV_{RMS}}$, and WAMP ($p = 0.004$–0.034). The trapezius muscle exhibited the most distinctive sEMG profile, characterized by greater temporal complexity and lower amplitude variability. These results show that the combined analysis of kinematic and sEMG variables enables the characterization of complementary aspects of neuromuscular coordination during functional upper-limb tasks. Given the small sample size (n = 8), these findings should be considered exploratory.**

***Index Terms*—Electromyography, functional assessment, kinematics, motor control, neuromechanics, upper limb.**

I. Introduction

The human upper limb is a versatile biomechanical system capable of performing a wide range of motor tasks, from gross movements to fine object manipulation [1]. Understanding the neuromechanical principles underlying object manipulation and goal-directed upper-limb movements is important not only for motor control research but also for clinical applications, including the assessment of neuromuscular disorders, the development of rehabilitation strategies, and the design of prostheses [2], [3].

This work was supported in part by CAPES through process nos. 88887.977117/2024-00 and 88887.145539/2025-00 (PROEX – Academic Excellence Program), granted to Laura Valentina Montealegre. Ricardo G. Molinari was supported by the Brazilian Public Ministry of Labor under Contract No. 002118.2019. Leonardo A. Elias is a CNPq Research Productivity Fellow, process no. 316320/2023-4. The kinematic data were collected using the Vicon Motion Capture system, a multi-user facility funded by FAPESP, process no. 2020/13293-0.

The Box and Block Test (BBT) and the Southampton Hand Assessment Procedure (SHAP) are widely used functional tests for assessing manual dexterity and motor performance [2], [4]. However, these tests rely primarily on global performance scores, such as the number of blocks transferred or the time required to complete the tasks. Although these scores quantify functional performance, they do not reveal the joint coordination and muscle activation strategies underlying movement execution [3], [5], [6].

Previous studies have shown that kinematic analyses of modified or instrumented versions of the BBT provide additional information on movement quality and compensatory mechanisms involving the trunk and upper limbs [5], [6]. Similarly, combining motion analysis with the SHAP has enabled the characterization of functional aspects that are not always reflected in the overall score [3]. Nevertheless, conventional functional assessments provide limited insight into the neuromechanical mechanisms underlying task performance. Similar global scores may result from different patterns of joint coordination or muscle activation. Integrating joint kinematics and muscle activity may therefore help characterize how movement is organized beyond the final task outcome.

Evidence combining kinematic analysis of the shoulder, elbow, and wrist with characterization of muscle activity using surface electromyography (sEMG) during the same functional tasks remains limited. Kinematic measurements describe the displacement and coordination of body segments, whereas sEMG provides noninvasive information on muscle activation. In addition to measures of overall amplitude, temporal features of the sEMG signal can provide complementary information on the variability and complexity of muscle activity [7]. Together, these measurements may help identify how specific functional demands are associated with different joint and muscle activation patterns.

A key question is whether the functional differentiation between proximal and distal segments, previously described in reaching and grasping, is also observed in more complex and functionally diverse tasks such as the BBT and SHAP. Proximal segments may contribute primarily to transporting and positioning the limb, whereas distal segments are more directly involved in object grasping and manipulation. This distinction is relevant because upper-limb control depends not

only on the final position of the hand but also on interjoint coordination [8], [9]. Moreover, temporal properties of muscle activation beyond average signal amplitude may provide additional information for differentiating muscle involvement across tasks with different motor demands [7].

Therefore, this study presents a neuromechanical characterization of the BBT and SHAP by combining kinematic analysis of the shoulder, elbow, and wrist with complementary sEMG features related to activation amplitude, variability, complexity, and transient changes: mean absolute value (MAV), sample entropy (SampEn), coefficient of variation of root-mean-square (RMS) values (CoV_{RMS}), and Willison amplitude (WAMP). By integrating these measurements within the same functional tasks, this study aimed to characterize the movement of proximal and distal upper-limb segments and identify differences in muscle activation patterns across muscles and limb sides. This approach may provide a basis for understanding the neuromechanical strategies underlying functional upper-limb performance. We hypothesized that these measurements would provide complementary information beyond the global performance indices of the BBT and SHAP.

II. Methods

A. Study Participants

Eight healthy volunteers participated in this study (7 males and 1 female; mean age ± standard deviation: 27 ± 7.8 years; body weight: 72.1 ± 15.1 kg; height: 1.73 ± 0.10 m). All participants provided written informed consent before the experimental procedures. The experimental protocol was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Research Ethics Committee of the State University of Campinas (UNICAMP), under CAAE 90477225.9.0000.5404, on August 10, 2025. Data collection took place after ethical approval, during the year 2025.

B. Experimental Protocol

Participants performed two standardized hand function tests: the BBT and the SHAP (Fig. 1). The BBT assesses manual dexterity by measuring the number of wooden blocks a participant can transfer from one side of the box to the other in 60 seconds, which was used as the functional index [4]. The SHAP assesses hand function using 26 timed tasks that involve manipulating abstract objects and performing activities of daily living [2]. The SHAP Index of Function (IOF) was calculated from the performance scores obtained across the 26 tasks, which are derived from task completion times and normalized into a global functional score.

Both tasks were performed while the participant was standing in front of a table whose height was individually adjusted to ensure the participant's comfort. Participants received standardized verbal instructions to complete the tasks as quickly and accurately as possible using only the limb being assessed. The tests were performed bilaterally in separate trials using the dominant and non-dominant limbs. The order of the tests (BBT or SHAP) and the limb tested (dominant or non-dominant) were randomized across participants to minimize potential learning or fatigue effects.

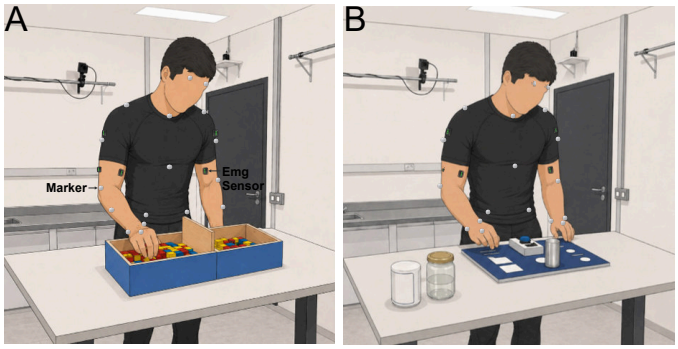


Fig. 1. Experimental setup for the simultaneous acquisition of kinematics and sEMG data during the standardized hand function tests. (A) Participant performing the Box and Block Test (BBT). (B) Participant performing a task from the Southampton Hand Assessment Procedure (SHAP).

C. Data Acquisition

1) Motion Capture: Kinematic data were acquired using a three-dimensional motion capture system (Vicon Motion Systems, Oxford, UK) consisting of eight infrared cameras operating at 100 Hz. A total of 39 retroreflective markers were placed on anatomical landmarks according to the full-body Plug-in Gait model [10], [11].

For this study, only markers corresponding to the head, trunk, and upper limbs were considered for the kinematic analysis (n = 23). This subset included four markers on the head, five on the trunk, and fourteen on the upper limbs, with seven markers on each side.

Marker trajectories were labeled in Vicon Nexus, and small trajectory gaps (up to 10 frames) were filled using a spline-based interpolation algorithm, ensuring signal continuity while minimizing distortion of the original trajectories.

2) Surface Electromyography: sEMG signals were recorded simultaneously with motion capture during task performance using a Delsys Trigno ™ system (Delsys Inc., Boston, MA, USA) at a sampling frequency of 2000 Hz and a resolution of 16 bits. sEMG activity was recorded from four upper-limb muscles: biceps brachii, the long head of the triceps brachii, deltoid, and trapezius. Before placing the electrodes, the skin was cleaned with alcohol to reduce skin-electrode impedance. The sensors were placed according to the recommendations of the Surface Electromyography for the Non-Invasive Assessment of Muscles (SENIAM) project [12], with the bipolar electrode pair positioned so that the line connecting both electrodes was aligned with the direction of the muscle fibers. Specifically, electrodes were positioned over the biceps brachii and triceps brachii muscle bellies, the anterior deltoid, and the upper trapezius, avoiding tendon regions and adjacent muscles.

D. Data Analysis

1) Kinematic Data Processing: Marker trajectories were smoothed using the Woltring spline filter, a generalized cross-

validation spline method commonly used in human motion analysis to reduce measurement noise while preserving the underlying signal dynamics. The level of smoothing was defined by the mean squared error (MSE) parameter, which specifies the tolerance to noise during spline fitting. An MSE value of 5 mm^2 was adopted, consistent with the expected noise level of the motion capture system [13]–[15]. The smoothed trajectories were subsequently used to calculate the kinematic variables analyzed in this study.

The coordinate systems of the segments and the definitions of the joints followed the recommendations of the International Society of Biomechanics for the upper limb [16]. Seven upper-limb joint angles were analyzed: three at the shoulder (flexion/extension, abduction/adduction, and internal/external rotation), one at the elbow (flexion/extension), and three at the wrist (flexion/extension, abduction/adduction, and rotation).

To reduce possible transient effects at the beginning and end of the recordings, the kinematic signals were trimmed before analysis. In the BBT task, the first and last 5 s of each trial were removed. In the SHAP task, the first and last second of each trial were removed.

2) Kinematic Variables: Trajectory and velocity metrics were obtained from the three-dimensional trajectories of upper-limb markers. In the dominant limb, the following markers were analyzed: RSHO (right shoulder acromion), RELB (right elbow lateral epicondyle), and RWRA (right wrist radial styloid). In the non-dominant limb, LSHO (left shoulder acromion), LELB (lateral epicondyle of the left elbow), and LWRA (radial styloid of the left wrist) were considered. These anatomical landmarks enabled characterization of arm, forearm, and hand movement during task performance.

The total distances traveled by the markers located on the shoulder, elbow, and wrist were calculated as the sum of the Euclidean distances between consecutive positions in three-dimensional space throughout each trial. The average speed was obtained by dividing the total distance traveled by the trial duration. In addition, the range of motion (RoM) for each degree of freedom was calculated as the difference between the maximum and minimum joint angles recorded during the trial. These variables were calculated for each trial and subsequently averaged across repetitions of the same experimental condition. In the BBT task, two repetitions per condition were averaged. In contrast, in the SHAP task, the variables were calculated across all subtasks and subsequently averaged per participant and joint to obtain a representative value for the task [3].

For the SHAP, the 26 subtasks were processed individually. Kinematic variables, including traveled distance, average velocity, and range of motion (RoM), as well as the sEMG features, were calculated separately for each subtask. Each subtask corresponded to a single execution, with no repeated trials. The entire available duration of each subtask after the predefined trimming procedure was analyzed. Because the subtasks had different durations, the signals were not temporally normalized or resampled to a common duration; each variable was therefore calculated using the original duration of the corresponding subtask.

3) sEMG Preprocessing: The sEMG signals were filtered using a zero-phase fourth-order Butterworth band-pass filter (10–500 Hz) to attenuate motion artifacts and high-frequency noise [12], [17]. The resulting filtered signals were then used for subsequent feature extraction. To avoid potential transients at the beginning and end of the recordings, the processed signals were trimmed before analysis. As with kinematic data, the first and last 5 s of each BBT recording were removed from the analysis, whereas the first and last 1 s of each SHAP trial were removed.

4) sEMG Feature Extraction: Features were extracted from the preprocessed sEMG signals to characterize muscle activity through complementary amplitude-, variability-, and complexity-related descriptors. To this end, four metrics were implemented: mean absolute value (MAV), sample entropy (SampEn), the coefficient of variation of the RMS signal ($\mathrm{CoV_{RMS}}$), and Willison amplitude (WAMP).

The overall amplitude of the signal was quantified using the mean absolute value (MAV) [7], defined in Equation 1, where x_i represents each sample of the sEMG signal and N is the total number of samples in the analyzed segment.

$$\mathrm{MAV} = \frac{1}{N}\sum_{i=1}^{N}|x_i| \tag{1}$$

As the sEMG amplitude was not normalised, the MAV was used solely as a descriptive measure. Direct comparisons of absolute MAV values between muscles should be interpreted with caution, as any differences may reflect anatomical factors or factors relating to the electrodes.

The temporal complexity of the sEMG signal was evaluated using sample entropy (SampEn, Equation 2). An embedding dimension $m = 2$ and a tolerance $r = 0.2 \cdot \sigma$ were used, where σ is the standard deviation of the signal [18]. In this equation, A represents the number of matches for patterns of length $m + 1$, and B represents the number of matches for patterns of length m. This measure reflects the signal's regularity, with lower values indicating more predictable patterns and higher values indicating greater temporal complexity.

$$\mathrm{SampEn}(m, r) = -\ln\left(\frac{A}{B}\right) \tag{2}$$

The amplitude variability was characterized using the coefficient of variation of the root-mean-square values of the sEMG signals ($\mathrm{CoV_{RMS}}$, Equation 3). The preprocessed sEMG signal was segmented into 200-ms windows with 50 % overlap over the entire trial duration. The RMS value was calculated for each window, yielding a sequence of RMS values throughout the trial. The 50 % overlap was used to maintain temporal continuity between consecutive windows. Here, σ_{RMS} and μ_{RMS} denote the standard deviation and mean of the RMS values across all windows in the experiment, respectively. This metric captures the variability of the signal amplitude over time.

$$\mathrm{CoV}_{\mathrm{RMS}} = \frac{\sigma_{\mathrm{RMS}}}{\mu_{\mathrm{RMS}}} \tag{3}$$

Finally, the Willison amplitude (WAMP) was calculated as the proportion of consecutive differences in the sEMG signal exceeding a predefined threshold T, defined by Equation 4, where N represents the number of samples and $f(\Delta x)$ is defined by Equation 5. The threshold T was set to 10 % of the standard deviation of each sEMG signal ($T = 0.1 \cdot \sigma$), within the low relative-threshold range recommended for amplitude-sensitive features [19]. The WAMP metric quantifies the frequency of significant amplitude changes between consecutive samples, reflecting the dynamic behavior of the sEMG signal. Higher WAMP values indicate a greater number of abrupt signal variations, which may be associated with changes in muscle activation dynamics and motor unit recruitment [20].

$$\mathrm{WAMP} = \frac{1}{N-1} \sum_{i=1}^{N-1} f(|x_{i+1} - x_i|) \tag{4}$$

$$f(\Delta x) = \begin{cases} 1, & \text{if } \Delta x > T \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

E. Statistical Analysis

Data normality was assessed using the Shapiro–Wilk test before hypothesis testing. All variables satisfied the normality assumption ($p > 0.05$), and parametric tests were therefore applied.

Statistical analyses were performed to evaluate differences in upper-limb kinematic variables across the experimental conditions. The variables analyzed included the total distance traveled and the average speed of anatomical markers associated with the shoulder, elbow, and wrist, as well as the RoM in degrees of freedom.

For total traveled distance and average speed, two-way repeated-measures ANOVAs (RM-ANOVAs) were performed considering side (dominant vs. non-dominant) and anatomical marker location (shoulder, elbow, and wrist) as within-subject factors. For RoM, separate two-way RM-ANOVAs were conducted for each movement type (e.g., flexion/extension, abduction/adduction, and rotation), considering side (dominant vs. non-dominant) and joint (shoulder, elbow, and wrist) as within-subject factors.

For each task (BBT or SHAP) and for each sEMG feature (MAV, SampEn, $\mathrm{CoV}_{\mathrm{RMS}}$, and WAMP), a two-way repeated-measures analysis of variance (RM-ANOVA) was employed, considering side (dominant vs. non-dominant) and muscle (biceps, triceps, deltoid, and trapezius) as within-subject factors. When a significant interaction between the factors was identified, post hoc comparisons with Bonferroni correction were performed. In the absence of a significant interaction, main effects were interpreted.

For the functional indices obtained in each task (BBT or SHAP), paired comparisons were performed between the dominant and non-dominant hands using paired Student's t-tests.

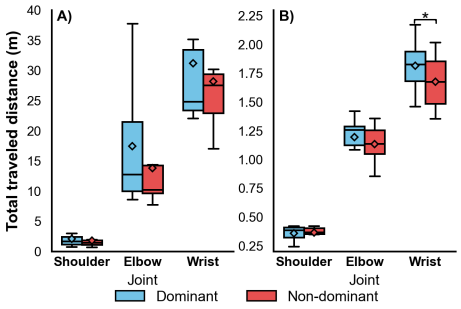


Fig. 2. Total traveled distance of the markers at the shoulder, elbow, and wrist during the Box and Block Test (BBT) (A) and the Southampton Hand Assessment Procedure (SHAP) (B). Boxplots show data from the dominant (blue) and non-dominant (red) limbs.

The significance level was set at $p < 0.05$. Effect sizes were calculated as partial eta squared (η_p^2) [21] and interpreted as small ($0.01 < \eta_p^2 < 0.06$), medium ($0.06 \leq \eta_p^2 < 0.14$), or large ($\eta_p^2 > 0.14$) [22]. All results are reported as mean ± standard deviation. Given the number of complementary kinematic and sEMG outcomes investigated, the statistical analyses were considered exploratory.

III. Results

A. Kinematic Characterization

Marker traveled distances for both the BBT and SHAP tasks are shown in Fig. 2. The total traveled distance significantly differed across marker locations in the BBT task ($p < 0.001$, $\eta_p^2 = 0.89$, Fig. 2A), with the lowest values at the shoulder, followed by the elbow, and the highest at the wrist. No significant effect of limb side was found ($p = 0.109$, $\eta_p^2 = 0.37$), and no significant side-by-marker-location interaction was observed ($p = 0.227$, $\eta_p^2 = 0.22$).

For the SHAP task (Fig. 2B), the analysis of total traveled distance revealed a significant main effect of marker location ($p < 0.001$, $\eta_p^2 = 0.98$). In addition, a significant effect of side was observed ($p = 0.043$, $\eta_p^2 = 0.59$), as well as a significant interaction between side and marker location ($p = 0.010$, $\eta_p^2 = 0.60$). Post hoc comparisons revealed a significant difference between limbs only at the wrist marker ($p = 0.035$).

The average speed of the markers at the shoulder, elbow, and wrist was assessed during the BBT and SHAP tasks, and the results are presented in Fig. 3. In the BBT task, the analysis revealed a significant main effect of marker location ($p < 0.001$, $\eta_p^2 = 0.89$), showing an increase in average speed from the shoulder to the wrist. Neither a significant effect of side ($p = 0.102$, $\eta_p^2 = 0.38$) nor a significant side-by-marker-location interaction ($p = 0.212$, $\eta_p^2 = 0.23$) was found.

For the SHAP task, the analysis of average speed revealed a significant main effect of marker location ($p < 0.001$, $\eta_p^2 = 0.996$). In addition, a significant effect of side was observed ($p = 0.005$, $\eta_p^2 = 0.82$), as well as a significant interaction between side and marker location ($p = 0.009$, $\eta_p^2 = 0.61$). Post hoc analysis of the interaction revealed significant differences between limbs only at the wrist marker ($p = 0.010$).

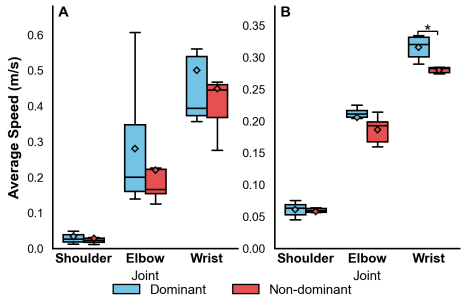


Fig. 3. Average speed of the markers at the shoulder, elbow, and wrist during the Box and Block Test (BBT) (A) and the Southampton Hand Assessment Procedure (SHAP) (B). Boxplots show data from the dominant (blue) and non-dominant (red) limbs.

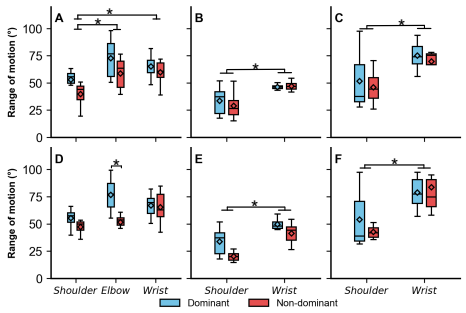


Fig. 4. Range of motion (RoM) during the Box and Block Test (BBT) and the Southampton Hand Assessment Procedure (SHAP). Panels A and D represent flexion/extension (Flex/Ext), panels B and E represent abduction/adduction (Abd/Add), and panels C and F represent rotation movements. Panels A–C correspond to the BBT task, whereas panels D–F correspond to the SHAP task. Boxplots show dominant (blue) and non-dominant (red) limbs. Diamonds indicate mean values.

Figure 4 shows the RoM associated with the degrees of freedom at each joint. In the BBT, the flexion/extension (Flex/Ext) movement (Fig. 4A) had a significant main effect of joint ($p = 0.002$, $\eta_p^2 = 0.66$). Post hoc comparisons revealed differences between the elbow and shoulder ($p = 0.004$), and between the shoulder and wrist ($p = 0.037$). For abduction/adduction (Abd/Add) movements (Fig. 4B), a significant main effect of joint was identified ($p = 0.026$, $\eta_p^2 = 0.59$), with post hoc comparisons showing differences between the shoulder and wrist ($p = 0.026$). For rotation (Fig. 4C), a significant main effect of joint was also found ($p = 0.009$, $\eta_p^2 = 0.71$), with post hoc comparisons indicating differences between the shoulder and wrist ($p = 0.009$). No significant main effects of side were found for RoM during the BBT (Flex/Ext: $p = 0.098$, $\eta_p^2 = 0.39$; Abd/Add: $p = 0.230$, $\eta_p^2 = 0.23$; Rotation: $p = 0.340$, $\eta_p^2 = 0.15$), and no significant side-by-joint interactions were observed (Flex/Ext: $p = 0.238$, $\eta_p^2 = 0.21$; Abd/Add: $p = 0.075$, $\eta_p^2 = 0.44$; Rotation: $p = 0.963$, $\eta_p^2 < 0.01$).

Conversely, for the SHAP task, the RoM associated with Flex/Ext (Fig. 4D) exhibited a significant main effect of side ($p = 0.015$, $\eta_p^2 = 0.65$) and joint ($p < 0.001$, $\eta_p^2 = 0.69$), along with a significant interaction between side and joint ($p = 0.012$). Post hoc analysis of the interaction showed that the difference between limbs was significant only at the elbow joint ($p = 0.021$). For Abd/Add (Fig. 4E), significant main effects of side ($p = 0.010$, $\eta_p^2 = 0.70$) and joint ($p = 0.002$, $\eta_p^2 = 0.81$) were found, with no significant interaction ($p = 0.561$). For rotation (Fig. 4F), a significant main effect of joint was identified ($p < 0.001$, $\eta_p^2 = 0.87$), with no significant effects of side ($p = 0.469$) or interaction ($p = 0.318$).

B. sEMG Characterization

For MAV, the BBT task (Fig. 5A) showed no significant muscle-by-side interaction ($p = 0.455$, $\eta_p^2 = 0.13$), whereas significant main effects were observed for both side ($p = 0.037$, $\eta_p^2 = 0.54$) and muscle ($p < 0.001$, $\eta_p^2 = 0.64$). However, post hoc comparisons for the main effect of muscle did not reveal significant pairwise differences after correction for multiple comparisons (all $p \geq 0.051$). In the SHAP task (Fig. 5B), MAV values showed no significant muscle-by-side interaction ($p = 0.920$, $\eta_p^2 = 0.03$) and no main effect of side ($p = 0.246$), but revealed a significant main effect of muscle ($p < 0.001$, $\eta_p^2 = 0.78$). Post hoc comparisons showed that, independently of limb side, the biceps brachii differed significantly from the triceps brachii ($p = 0.001$) and trapezius ($p = 0.003$).

In the BBT task, the SampEn values (Fig. 5C) showed no significant muscle-by-side interaction ($p = 0.258$, $\eta_p^2 = 0.20$) and no main effect of side ($p = 0.091$). However, a significant main effect of muscle was found ($p < 0.001$, $\eta_p^2 = 0.90$), indicating that SampEn values differed across muscles independently of side. Post hoc comparisons revealed higher SampEn values for the trapezius compared with the biceps brachii ($p < 0.001$), triceps brachii ($p = 0.008$), and deltoid ($p = 0.003$). In contrast, in the SHAP task (Fig. 5D), a significant muscle-by-side interaction was observed ($p = 0.034$, $\eta_p^2 = 0.37$), along with a significant main effect of muscle ($p < 0.001$, $\eta_p^2 = 0.79$), while the effect of side was not significant ($p = 0.328$). Post hoc comparisons showed that the trapezius had higher SampEn values than the biceps brachii in both limbs (dominant: $p = 0.004$; non-dominant: $p = 0.029$), as well as higher values than the deltoid ($p = 0.019$) and the triceps brachii ($p = 0.046$), only in the non-dominant limb.

The $\mathrm{CoV_{RMS}}$ analysis in the BBT task (Fig. 5E) revealed a significant main effect of muscle ($p < 0.001$, $\eta_p^2 = 0.89$), with no significant muscle-by-side interaction ($p = 0.404$, $\eta_p^2 = 0.15$) or main effect of side ($p = 0.646$). Post hoc comparisons showed that the trapezius muscle had significantly lower $\mathrm{CoV_{RMS}}$ values than the biceps brachii ($p < 0.001$) and the deltoid ($p < 0.001$), independently of limb side. In the SHAP task (Fig. 5F), however, a significant muscle-by-side interaction was found ($p = 0.004$, $\eta_p^2 = 0.51$), along with a significant main effect of muscle ($p < 0.001$, $\eta_p^2 = 0.72$) and no effect of side ($p = 0.777$). Post hoc comparisons indicated that, in the dominant limb, the trapezius muscle exhibited lower $\mathrm{CoV_{RMS}}$ values than the biceps brachii ($p = 0.032$) and the deltoid ($p = 0.004$). In contrast, in the non-dominant limb, $\mathrm{CoV_{RMS}}$ of the trapezius muscle was lower than that of

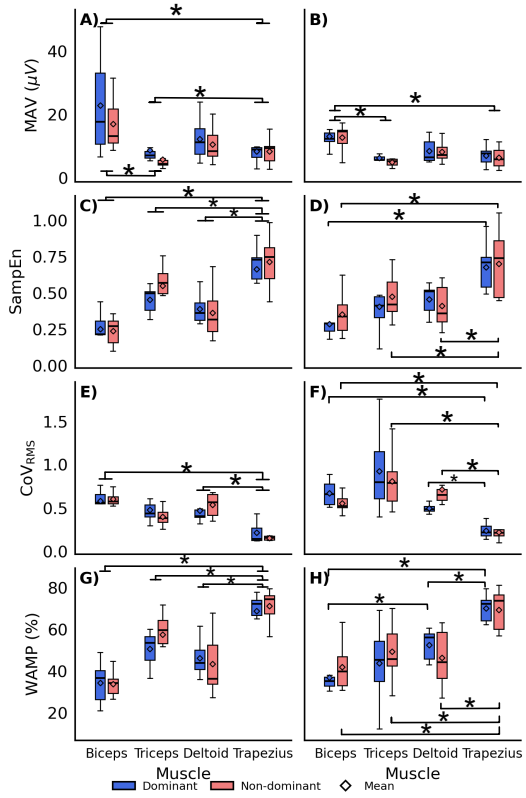


Fig. 5. Surface electromyographic (sEMG) characterization across muscles and tasks. Boxplots represent the distribution of MAV (A–B), SampEn (C–D), $\mathrm{CoV_{RMS}}$ (E–F), and WAMP (G–H). For each feature, the left panel shows the Box and Block Test (BBT) and the right panel shows the Southampton Hand Assessment Procedure (SHAP). Blue and red colors indicate dominant and non-dominant limbs, respectively. Asterisks indicate significant pairwise differences between muscles ($p < 0.05$).

the biceps brachii ($p < 0.001$), the deltoid ($p = 0.004$), and the triceps brachii ($p = 0.018$).

For WAMP, the BBT task (Fig. 5G) showed no significant muscle-by-side interaction ($p = 0.376$, $\eta_p^2 = 0.16$), nor a significant main effect of side ($p = 0.143$, $\eta_p^2 = 0.32$), whereas a significant main effect of muscle was observed ($p < 0.001$, $\eta_p^2 = 0.94$). Post hoc comparisons revealed that the trapezius muscle differed from the biceps brachii ($p < 0.001$), deltoid ($p < 0.001$), and triceps brachii ($p < 0.001$). For the SHAP (Fig. 5H), a significant muscle-by-side interaction was observed ($p = 0.005$, $\eta_p^2 = 0.50$). In addition to a significant main effect of muscle ($p < 0.001$, $\eta_p^2 = 0.80$). The main effect of side was not significant ($p = 0.637$, $\eta_p^2 = 0.04$). Post hoc comparisons revealed that the trapezius muscle differed from the biceps brachii (dominant: $p < 0.001$; non-dominant: $p = 0.002$) and the deltoid (dominant: $p = 0.007$; non-dominant: $p = 0.005$) in both limbs. A difference between the trapezius and triceps brachii was observed only in the non-dominant limb ($p = 0.030$). Additionally, a difference between the biceps brachii and deltoid was observed in the dominant limb ($p = 0.043$).

C. Functional assessment indices

Functional assessment indices obtained from the BBT and SHAP tasks are presented in Table I. The BBT scores were higher for the dominant limb compared with the non-dominant limb; however, no significant differences were observed between sides ($p = 0.126$, Cohen's $d_z = 0.67$). Similarly, SHAP scores were slightly higher for the dominant limb, with no significant differences between the dominant and non-dominant limbs ($p = 0.352$, Cohen's $d_z = 0.38$).

TABLE I
FUNCTIONAL TEST INDICES

Functional test	Dominant	Non-dominant
Box and Block Test	74 ± 6.69	71 ± 7.60
Southampton Hand Assessment Procedure	94 ± 2.08	92 ± 3.95

IV. DISCUSSION

This study investigated the neuromechanical properties of proximal and distal upper-limb segments during the BBT and SHAP, two functional motor tasks used to assess hand dexterity. To this end, both tests were characterized through kinematic analysis of the shoulder, elbow, and wrist joints, together with a multi-feature analysis of sEMG signals from upper-limb muscles.

Our results indicate that both tasks (BBT and SHAP) exhibited similar movement patterns, characterized by greater displacement and velocity of the wrist marker than of the elbow and shoulder markers. This similarity stems from both tasks involving reaching and grasping actions that require transporting and manipulating objects [1]. To perform these movements, the proximal joints primarily contribute to positioning and stabilizing the upper limb in space. In contrast, the distal joints are more directly involved in hand orientation, fine motor adjustments, and interaction with the object [1], [8]. Similar patterns have been described in the literature, namely, greater involvement of the distal joints during reaching and object-manipulation tasks [23].

In the BBT task, the contribution of each joint to the movement pattern varied by degree of freedom (flexion/extension, abduction/adduction, and rotation), and distal joints exhibited greater RoM than proximal joints, while distal markers showed higher speeds than proximal markers. However, we did not observe significant differences between the dominant and non-dominant limbs in total distance traveled, average speed, or RoM. This suggests that the repetitive and relatively stereotyped nature of the BBT may contribute to similar movement patterns on both sides [24]. Since the task consists of repeated reaching and grasping movements with limited variability, the observed movement patterns may be influenced primarily by the task's mechanical demands rather than by limb dominance [23].

In contrast, the SHAP showed different neuromechanical properties, with significant differences between the dominant and non-dominant limbs, particularly at the wrist joint. These findings suggest that differences between limbs may become

more evident as task complexity increases. Unlike the BBT, which involves repetitive movements, the SHAP comprises 26 subtasks that require different types of grasping, object manipulation, and movement transitions [2]. It is likely that the greater variability of motor actions in the SHAP, together with the complexity of some tasks, increases the demands on finer motor control and movement adaptation [25]. This is especially true in the distal segments that interact with objects. Under these conditions, lateralized motor strategies may become more pronounced, with limb dominance playing an important role in movement execution. Therefore, our results suggest that upper limb lateralization may be task-dependent and may become more evident during functionally complex activities [9], [26].

The sEMG features captured distinct properties of muscle activation during functional upper-limb tasks. The MAV showed significant muscle-dependent effects; however, post hoc comparisons were less consistent. Since this metric primarily reflects the overall intensity of muscle activation [17], [27], the results suggest that the functional tasks evaluated required a relatively homogeneous level of muscular effort across various muscles, possibly due to the coordinated and simultaneous nature of the reaching, grasping, and manipulation movements present in the BBT and SHAP tasks. Additionally, the lack of sEMG amplitude normalization may have contributed to the variability observed in the MAV values and, consequently, to the inconsistencies in the post hoc comparisons. Although normalization is recommended to improve the comparability of sEMG amplitudes across muscles and participants, selecting an appropriate normalization procedure for dynamic functional tasks remains challenging because no single method is universally applicable across movement conditions [28].

The temporal complexity of the sEMG signals, as measured by SampEn, was significantly higher in the trapezius than in the other muscles examined (biceps brachii, triceps brachii, and deltoid). This finding is physiologically consistent, as the trapezius, a muscle involved in stabilizing the shoulder and scapula, may require continuous fine-tuning to maintain postural control during movement, which could be reflected in greater signal irregularity [29]. Furthermore, we observed that the trapezius exhibited relatively low $\mathrm{CoV}_{\mathrm{RMS}}$ values, indicating lower variability in sEMG amplitude over time. In contrast, the biceps brachii and triceps brachii muscles showed greater amplitude fluctuations, despite exhibiting less complex activation patterns. This distinction suggests different functions during task performance. While the biceps brachii and triceps brachii are directly involved in the generation and modulation of movement-related forces, the trapezius appears to provide a more constant level of activation associated with shoulder position and upper limb support during task performance [30].

The WAMP analysis reinforced the functional distinction between stabilizing muscles and those more actively involved in movement execution. The trapezius muscle exhibited the most distinctive pattern in this measure, with more frequent changes in its activation amplitude. Together with the SampEn and $\mathrm{CoV}_{\mathrm{RMS}}$ findings, this result suggests that trapezius activation was characterized by a relatively constant overall amplitude accompanied by frequent local adjustments. These findings are consistent with a distinct pattern of trapezius muscle activity during functional tasks, characterized by continuous modulation of the recorded sEMG signal. In contrast, the biceps brachii and triceps brachii exhibited distinct activation profiles that were more closely related to movement execution during the task. Taken together, these findings suggest that combining multiple sEMG features enables a more comprehensive characterization of muscle activation during functional tasks.

An important point is that both tasks showed muscle-dependent differences in sEMG activity, but the structure of these responses differed across tasks. In the BBT, the absence of significant interactions across most metrics suggests that myoelectric activity was primarily driven by muscle-specific effects, with limited evidence of limb-dependent modulation. In contrast, the SHAP showed significant muscle-by-side interactions, indicating that muscle activation depended on the combination of muscle and limb. Therefore, although both tasks may share a general organization related to the functional role of each muscle, the SHAP appears to reveal task-specific and limb-dependent modulations in myoelectric activity.

Taken together, the kinematic and sEMG results enabled a more comprehensive characterization of upper-limb function during object manipulation tasks [31]. Kinematic analysis revealed greater displacement and average speed of the wrist marker than of the shoulder and elbow markers. At the same time, sEMG analysis showed functional differences among the evaluated muscles, suggesting distinct contributions to movement generation and postural maintenance. Although the BBT and SHAP functional indices showed higher values for the dominant side, no statistically significant differences were observed between sides. These results suggest that, while functional indices provide a global measure of performance during the evaluated tasks, differences associated with laterality may primarily emerge at the level of motor control strategies, which are not necessarily reflected in overall functional scores. In contrast, kinematic and sEMG variables allowed the identification of specific movement characteristics and muscle activation patterns, providing complementary information about the neuromotor control mechanisms involved during functional task execution. These findings reinforce the importance of combining kinematic and sEMG information to investigate motor control during functional tasks, as movement organization cannot be fully captured by joint variables or global measures of muscle activation alone [18].

Some limitations should be acknowledged: the sEMG analysis was performed on only four upper-limb muscles, and the sample comprised a small number of healthy participants. Despite these limitations, the approach enabled kinematic and sEMG characterization of functional upper-limb tasks. Future studies could expand on this analysis by including other muscle groups and different clinical populations.

V. Conclusion

This study characterized the kinematic and sEMG properties of functional upper-limb tasks by analyzing the BBT and SHAP in healthy individuals. The results revealed differences in movement features between the two tasks, particularly in the involvement of distal joints and the effects of handedness during more functionally complex and cognitively demanding tasks. Furthermore, sEMG characteristics revealed distinct patterns of muscle activation during task performance. Taken together, these findings contribute to a better understanding of motor behavior during object manipulation and demonstrate that kinematic and sEMG analyses extend the information provided by functional test indices, which primarily reflect overall task performance. By revealing task-specific movement patterns, joint-related characteristics, and muscle activity patterns, these neuromechanical measures provide a more comprehensive characterization of functional upper-limb performance. Given the small sample size (n = 8), the present findings should be considered exploratory.

Acknowledgment

We thank Guilherme A. G. De Villa for data collection.

References

[1] M. Jeannerod, "The timing of natural prehension movements," *Journal of Motor Behavior*, vol. 16, no. 3, pp. 235–254, 1984.

[2] C. M. Light, P. H. Chappell, and P. J. Kyberd, "Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: normative data, reliability, and validity," *Archives of Physical Medicine and Rehabilitation*, vol. 83, no. 6, pp. 776–783, Jun 2002.

[3] I. Vujaklija, M. K. Jung, T. Hasenoehrl, A. D. Roche, A. Sturma, S. Muceli, R. Crevenna, O. C. Aszmann, and D. Farina, "Biomechanical analysis of body movements of myoelectric prosthesis users during standardized clinical tests," *IEEE Transactions on Biomedical Engineering*, vol. 70, pp. 789–799, 3 2023.

[4] V. Mathiowetz, G. Volland, N. Kashman, and K. Weber, "Adult norms for the box and block test of manual dexterity," *The American Journal of Occupational Therapy*, vol. 39, no. 6, pp. 386–391, 1985.

[5] J. S. Hebert, J. Lewicke, T. R. Williams, and A. H. Vette, "Normative data for modified box and blocks test measuring upper-limb function via motion capture," *Journal of Rehabilitation Research and Development*, vol. 51, no. 6, pp. 918–932, 2014.

[6] E. S. Cocco, S. Pournajaf, C. M. Manzia, C. De Marchis, C. Massaroni, E. Schena, and M. Goffredo, "Comparative analysis of upper body kinematics in stroke, parkinson's disease, and healthy subjects: An observational study using imu-based targeted box and block test," *Gait & Posture*, vol. 114, pp. 69–77, 2024.

[7] D. Tkach, H. Huang, and T. Kuiken, "Study of stability of time-domain features for electromyographic pattern recognition," *Journal of NeuroEngineering and Rehabilitation*, vol. 7, p. 21, 2010.

[8] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *Journal of Neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[9] R. L. Sainburg, "Evidence for a dynamic-dominance hypothesis of handedness," *Experimental Brain Research*, vol. 142, pp. 241–258, 2002.

[10] M. P. Kadaba, H. K. Ramakrishnan, and M. E. Wootten, "Measurement of lower extremity kinematics during level walking," *Journal of Orthopaedic Research*, vol. 8, pp. 383–392, 1990.

[11] R. B. Davis, S. Õunpuu, D. Tyburski, and J. R. Gage, "A gait analysis data collection and reduction technique," *Human Movement Science*, vol. 10, pp. 575–587, 1991.

[12] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for semg sensors and sensor placement procedures," *Journal of Electromyography and Kinesiology*, vol. 10, no. 5, pp. 361–374, 2000.

[13] H. J. Woltring, "A fortran package for generalized, cross-validatory spline smoothing and differentiation," *Advances in Engineering Software*, vol. 8, no. 2, pp. 104–113, 1986. [Online]. Available: https://doi.org/10.1016/0141-1195(86)90098-7

[14] D. A. Winter, *Biomechanics and motor control of human movement*. Wiley, 2009.

[15] C. Seegelke, C. M. Hughes, and T. Schack, "An investigation into manual asymmetries in grasp behavior and kinematics during an object manipulation task," *Experimental Brain Research*, vol. 215, no. 1, pp. 65–75, 2011.

[16] G. Wu, F. C. T. van der Helm, H. E. J. Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz, "Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion — part ii: shoulder, elbow, wrist and hand," *Journal of Biomechanics*, vol. 38, no. 5, pp. 981–992, 2005.

[17] C. De Luca, "The use of surface electromyography in biomechanics," *Journal of Applied Biomechanics*, vol. 13, no. 2, pp. 135–163, 1997.

[18] A. Phinyomark, F. Quaine, S. Charbonnier, C. Serviere, F. Tarpin-Bernard, and Y. Laurillau, "Emg feature evaluation for improving myoelectric pattern recognition robustness," *Expert Systems with Applications*, vol. 40, no. 12, pp. 4832–4840, 2013.

[19] A. Waris and E. N. Kamavuako, "Effect of threshold values on the combination of emg time domain features: Surface versus intramuscular emg," *Biomedical Signal Processing and Control*, vol. 45, pp. 267–273, 2018.

[20] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "A novel feature extraction for robust emg pattern recognition," *Journal of Computing*, vol. 1, no. 1, pp. 71–80, 2009.

[21] J. Cohen, "Eta-squared and partial eta-squared in fixed factor anova designs," *Educational and Psychological Measurement*, vol. 33, no. 1, p. 107–112, Apr. 1973. [Online]. Available: http://dx.doi.org/10.1177/001316447303300111

[22] S. Mangiafico, *Summary and Analysis of Extension Program Evaluation in R*, 1st ed., Rutgers Cooperative Extension, New Brunswick, NJ, 2016. [Online]. Available: https://rcompanion.org/handbook/

[23] P. H. McCrea, J. J. Eng, and A. J. Hodgson, "Biomechanics of reaching: Clinical implications for individuals with acquired brain injury," *Disability and Rehabilitation*, vol. 24, no. 10, pp. 534–541, 2002.

[24] C. Schütz and T. Schack, "Hemispheric lateralization does not affect the cognitive and mechanical cost of a sequential motor task," *Experimental Brain Research*, vol. 237, no. 12, pp. 3133–3142, 2019. [Online]. Available: https://doi.org/10.1007/s00221-019-05652-6

[25] A. R. Sobinov and S. J. Bensmaia, "The neural mechanisms of manual dexterity," *Nature Reviews Neuroscience*, vol. 22, pp. 741–757, 2021.

[26] P. K. Mutha, K. Y. Haaland, and R. L. Sainburg, "Rethinking motor lateralization: specialized but complementary mechanisms for motor control of each arm," *PLoS ONE*, vol. 8, no. 3, p. e58582, 2013.

[27] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface emg," *Journal of Applied Physiology*, vol. 96, no. 4, pp. 1486–1495, 2004.

[28] A. M. Burden, M. Trew, and V. Baltzopoulos, "Normalisation of gait emgs: A re-examination," *Journal of Electromyography and Kinesiology*, vol. 13, no. 6, pp. 519–532, 2003.

[29] D. Falla, D. Farina, and T. Graven-Nielsen, "Experimental muscle pain results in reorganization of coordination among trapezius muscle subdivisions during repetitive shoulder flexion," *Experimental Brain Research*, vol. 178, no. 3, pp. 385–393, 2007.

[30] R. Osu, D. W. Franklin, H. Kato, H. Gomi, K. Domen, T. Yoshioka, and M. Kawato, "Short- and long-term changes in joint co-contraction associated with motor learning as revealed from surface emg," *Journal of Neurophysiology*, vol. 88, no. 2, pp. 991–1004, 2002.

[31] J. M. Lopez-Castellanos, J. L. Ramon, J. Pomares, G. J. Garcia, and A. Ubeda, "Multisensory evaluation of muscle activity and human manipulability during upper limb motor tasks," *Biosensors*, vol. 13, no. 7, p. 697, 2023. [Online]. Available: https://doi.org/10.3390/bios13070697

Device Architectures, High-Voltage Supplies, and Translational Challenges in Portable Electrospinning for Wound Care

Ianca Moraes Silva de Jesus
Electrical Engineering Dept.
Fed. Univ. of Mato Grosso do Sul
Campo Grande, Brazil
ianca.moraes@ieee.org

Ruben Barros Godoy
Electrical Engineering Dept.
Fed. Univ. of Mato Grosso do Sul
Campo Grande, Brazil
ruben.godoy@ufms.br

***Abstract*—Portable and handheld electrospinning devices enable the in situ deposition of polymeric nanofiber meshes for wound care, but require the integration of a compact, low-power high-voltage supply with a controlled infusion system. A structured narrative search of IEEE Xplore, Scopus, Web of Science, PubMed, and ScienceDirect through July 2026 retained 43 studies for qualitative comparative synthesis. The literature was analyzed across three complementary axes: biomedical, material, and process requirements; the architectures and performance of portable and handheld devices; and high-voltage topologies, with emphasis on Cockcroft–Walton multipliers and flyback-based supplies. Reported devices show increasing integration and biomedical validation; however, recurring limitations include incomplete electrical specifications, limited flow control and battery autonomy, and scarce electrical-safety data. The comparison indicates that Cockcroft–Walton multipliers remain attractive for extreme miniaturization, whereas flyback-based supplies are more frequently reported in electrospinning and provide transformer isolation together with high voltage gain. Repurposed cathode-ray-tube flyback transformers may support low-cost proof-of-concept prototyping, but their lack of traceable specifications and lot-to-lot reproducibility limits clinical translation. Galvanic isolation alone is insufficient to establish clinical suitability; output regulation, stored-energy and current characterization, component standardization, and regulatory validation remain priorities.**

***Index Terms*—electrospinning, portable devices, high-voltage power supply, flyback converter, Cockcroft-Walton multiplier, wound healing, nanofibers**

I. Introduction

Tissue injuries – damage to the skin from physical, thermal, chemical, or mechanical injury, or from underlying disease – are broadly classified as acute or chronic; chronic lesions, arising from conditions such as diabetes, impose a considerable burden on healthcare systems [1]. Whereas acute injuries often heal through intrinsic repair, extensive or full-thickness lesions require supporting materials that promote the migration and infiltration of reparative cells [2].

Wound healing is a complex process comprising several overlapping stages: hemostasis, inflammation, proliferation, and remodeling. A crucial step in tissue lesion treatment is dressing application, aimed at protecting the wound from external risk factors while accelerating the healing process. An ideal dressing should: (i) absorb excess exudate, (ii) prevent infection, (iii) maintain moisture at the lesion site, (iv) facilitate gas exchange, (v) be biocompatible and biodegradable, (vi) avoid adhering to the wound while remaining easy to remove, and (vii) induce angiogenesis and tissue remodeling [1].

Electrospinning is a versatile and low-cost technique for producing fibers with diameters ranging from nanometers to micrometers by stretching a charged polymer jet under a high-voltage electric field [3]. Electrospun nanofibers have emerged as a promising technology for meeting the wound-dressing requirements described above, particularly in the fabrication of 2D and 3D meshes. These structures offer the unique ability to be tuned in diameter, porosity, alignment, overlap, and surface functionalization, allowing precise control of cell migration – a control that is crucial for improving the repair and regeneration of tissues as diverse as nerve, skin, cardiac, vascular, and musculoskeletal tissue [2]. Electrospun nanofibers have additionally been widely studied as wound dressings owing to their high porosity and large surface area, properties that inhibit pathogen growth at the lesion site while facilitating the exchange of gaseous and liquid molecules [1].

Applying nanofibers directly onto a lesion in situ requires portable devices that overcome the bulk and mains dependence of benchtop equipment [22], [24]. Such devices need a compact, low-power high-voltage (HV) supply together with an infusion system delivering the solution at a controlled, reproducible flow rate – which directly affects Taylor-cone stability and fiber morphology [24], [26]. Compact, low-cost high-voltage sources are therefore a central design priority.

Research groups have proposed portable or handheld electrospinning platforms and, in parallel, a diverse array of high-voltage generation strategies to power them, ranging from Cockcroft–Walton (CW) voltage multipliers to flyback converters. These two lines of research – the biomedical characterization of the resulting nanofibers and the power electronics that render portable devices feasible – are, however, seldom reviewed jointly. This narrative review integrates the biomedical process requirements, the portable electrospinning

architectures, and the high-voltage power-supply design constraints, in order to identify the trade-offs that currently limit clinical translation. Section II describes the review methodology. Section III reviews the biological, physical, and chemical characterization of electrospun nanofibers for biomedical use, together with the physics governing jet formation and the analytical techniques used to characterize the resulting fibers. Section IV surveys the state of the art of portable and handheld electrospinning devices. Section V compares the high-voltage power-supply topologies reported for this class of device. Section VI discusses the feasibility of a flyback-based topology in terms of cost, safety, and component availability. Section VII summarizes the main open challenges, and Section VIII concludes the paper.

II. Review Methodology

This work is a narrative review with a structured and transparent search strategy. IEEE Xplore, Scopus, Web of Science, PubMed, and ScienceDirect were searched up to July 2026 using combinations of "portable/handheld electrospinning," "in situ wound dressing," "high-voltage power supply," "flyback converter," and "Cockcroft–Walton multiplier." Studies were included when they described a portable electrospinning device, its biomedical validation, or a high-voltage topology applicable to low-power electrospinning; foundational references on nanofiber characterization and jet physics were retained for background. After screening, 43 studies were retained for qualitative comparative synthesis. Given the heterogeneity of the devices and the frequent absence of standardized electrical data, the synthesis is qualitative and comparative rather than a quantitative meta-analysis.

III. Characterization of Electrospun Nanofibers

The performance of an electrospun mat as a wound dressing or drug-delivery vehicle depends on biological, physical, and chemical properties tuned jointly during electrospinning. This section summarizes recent findings and the underlying jet physics, kept concise so as to prioritize the portable-device comparison of Sections IV–VI.

A. Biological Characterization and Chemical Functionalization

An ideal wound dressing must absorb excess exudate, prevent infection, maintain a moist environment, and promote angiogenesis and tissue remodeling, while remaining biocompatible and biodegradable [6]; electrospun nanofibers can meet these requirements, in large part through improved fibroblast migration [2]. Therapeutic agents are incorporated either directly during electrospinning or by post-treatment, and this functionalization largely dictates biomedical performance [7]: antibiotics and silver or gold nanoparticles confer antimicrobial activity, including against multi-drug-resistant bacteria, while anti-inflammatory drugs such as dexamethasone or ibuprofen provide sustained, localized delivery. Representative primary studies illustrate the range of outcomes: doxycycline-loaded polybutylene adipate-co-terephthalate (PBAT)/gelatin meshes functionalized with an arginine-glycine-aspartate (RGD) peptide prevented *S. aureus* and *P. aeruginosa* infection [12]; pH-responsive coaxial fibers co-delivered lidocaine and curcumin for staged release [13]; hyaluronic acid/silk core–shell fibers with zinc oxide were active against *E. coli* and *S. aureus* [14]; chitosan/poly(3-hydroxybutyrate-co-3-hydroxyvalerate) (PHBV) fibers loaded with cephradine reached an encapsulation efficiency above 70% with sustained 48 h release in methicillin-resistant *S. aureus* (MRSA)-infected wounds [15]; a bilayered polycaprolactone/poly(vinyl alcohol) (PCL/PVA) system co-releasing LL-37 and vascular endothelial growth factor (VEGF) improved both bactericidal activity and epidermal regeneration in vivo [16]; and graphene-oxide composites enhanced bactericidal efficacy and reepithelialization, with optimal loadings yielding healing rates above 90% in murine models [11]. Loading strategy and fiber morphology are not independent, however: chrysin-loaded PCL–chitosan fibers, for example, increased in diameter from 191 to 313 nm as drug concentration rose [21].

B. Polymer and Solvent Selection

Polymers for biomedical use must meet criteria of biodegradability, safety, biocompatibility, moderate hydrophilicity, and adequate mechanical strength, and may be natural or synthetic [1]. Natural polymers such as chitosan, cellulose, collagen, and silk are valued for their biocompatibility and additional biological activity, but are comparatively costly and harder to process; synthetic polymers such as polylactic acid (PLA), poly(ethylene oxide) (PEO), and polyvinylpyrrolidone (PVP), together with PVA and PCL, offer tunable mechanical and degradation characteristics and are frequently blended with natural polymers to combine both advantages [8]. Solvent selection governs solution homogeneity and jet behavior: low-molecular-weight polymers tend to produce beads rather than fibers owing to limited chain entanglement, solvent volatility sets the jet solidification rate, and solvent dielectric constant modulates the electrostatic repulsion between surface charges [9], [19]. Commonly used solvents include alcohols, dichloromethane, chloroform, dimethylformamide (DMF), tetrahydrofuran (THF), acetone, dimethyl sulfoxide (DMSO), hexafluoroisopropanol (HFIP), and trifluoroethanol. Although aqueous electrospinning is well established for water-soluble polymers such as PVA and PEO, water's high surface tension and relatively slow evaporation can narrow the stable processing window; co-solvents or adjustments in temperature, distance, and flow rate may therefore be required depending on the polymer system [9].

C. Physical Characterization and Process Parameters

The electrospinning process is highly sensitive to a set of solution, process, and environmental parameters, whose optimization is crucial to achieving the desired fiber morphology, diameter, and mechanical properties [9]. Table I summarizes the effect of the main parameters reported in the literature.

TABLE I
EFFECT OF KEY ELECTROSPINNING PARAMETERS ON FIBER DIAMETER AND MORPHOLOGY.

Category	Parameter	Reported effect
Solution	Polymer concentration / M_w	Higher concentration or molecular weight increases chain entanglement, increasing fiber diameter; below a threshold, beads form instead of fibers [19].
	Surface tension	Higher surface tension raises the critical electric field and may promote bead formation or jet instability; its effect on fiber diameter is formulation-dependent [9].
	Conductivity	Higher conductivity increases surface charge density, accelerating stretching and reducing diameter, at the cost of increased Coulombic repulsion [9].
Process	Applied voltage	Higher voltage increases charge density and stretching, reducing diameter; excessively high voltage can raise flow rate and increase diameter again [17], [19].
	Flow rate	Higher flow rate increases fiber diameter and pore size; excessive flow leads to dripping and bead-on-string defects. Response-surface analysis identifies flow rate as the dominant parameter for mechanical properties [20].
	Tip-to-collector distance	Greater distance allows more time for jet stretching and solvent evaporation, reducing fiber diameter [17].
Environmental	Temperature	Higher temperature reduces solution viscosity and intermolecular Coulomb forces, reducing fiber diameter [9].
	Relative humidity	High humidity causes surface condensation and fiber wrinkling; very low humidity accelerates solvent evaporation at the needle tip, causing clogging [9].

Representative operating windows reported for portable or bench-characterized setups illustrate this parameter space: Haik et al. (2017) applied 25 kV at a 20 cm working distance, a 4.5 mL/h flow rate, and a 0.7 mm needle tip [25]; Mouthuy et al. (2015) operated at a fixed 20 cm distance with voltage adjustable between 0 and 14 kV and flow rate between 0 and 10 mL/h, tuned individually for each polymer solution [24]; and a parametric bench study using a polyacrylonitrile (PAN)/DMF solution at 10% (w/v) evaluated needle diameters of 0.45–0.80 mm and voltages of 15, 20, and 25 kV at a fixed 140 mm distance and 500 μL/h flow rate, obtaining fiber diameters as low as 364 $\pm$ 63 nm at 15 kV with a 0.55 mm needle [17]. A response-surface optimization of a PCL skin scaffold identified flow rate as the dominant parameter affecting both Young's modulus and tensile strength, ahead of polymer concentration and voltage [20], while a broader survey across cellulose- and chitosan-based biopolymers confirmed that fiber diameter increases with both solution concentration and applied voltage regardless of the specific polymer family [19]. Together, these results confirm the trends summarized in Table I and underscore that, for a portable device, the flow rate and the applied voltage must be co-regulated within a comparatively narrow, solution-specific window in order to sustain a stable Taylor cone.

D. Physics of Jet Formation and the Taylor Cone

The physical principle underlying nanofiber generation consists of electrically charging a polymer solution at the needle tip. The applied field must overcome the surface tension of the solution, deforming the pendant droplet into the characteristic conical shape known as the Taylor cone, from whose apex a charged jet is ejected toward the grounded collector [2], [4]. As the applied voltage increases, charge accumulates at the droplet surface until, at a critical voltage V_c, the electrostatic pressure $\rho_e = \epsilon E^2/2$ overcomes the capillary pressure $\rho_c = 2\gamma/r$ dictated by the Young–Laplace equation, where ϵ is the dielectric constant of the surrounding medium, E is the electric field intensity, γ is the surface tension, and r is the needle's internal radius. The critical voltage can be estimated by the semi-empirical relation

$$V_c^2 = \frac{4H^2}{h^2}\left(\ln\left(\frac{2h}{R}\right) - 1.5\right)(1.3\pi R\gamma)(0.09) \qquad (1)$$

where H is the needle-to-collector distance, h is the needle length, and R is the needle's external radius; the factor 1.3 derives from $2\cos(49.3^\circ)$, the semi-vertical half-angle at which the cone reaches equilibrium [2], [4]. Equation (1) is empirical and dimensionally inhomogeneous: its numerical constants (with $(1.3)(0.09) \approx 0.117$ matching the classical semi-empirical form) require the stated mixed units, namely H, h, and R in centimeters, γ in dyn/cm, and V_c in kV. It should therefore be treated as an order-of-magnitude design estimate rather than an exact physical law.

Once ejected, the jet initially follows a straight path – the near-field region – over a length sufficient to suppress the Rayleigh instability that would otherwise break it into droplets. Beyond this segment, the jet undergoes a whipping (bending) instability that stretches it by up to four to five orders of magnitude within tens of milliseconds, reaching drawing rates of up to 10^6 s^{-1} [2]. Under the whipping-instability model, the terminal jet diameter can be estimated as

$$d_t = \left(\gamma\varepsilon\frac{Q^2}{I^2}\frac{2}{\pi(2\ln\chi - 3)}\right)^{1/3} \qquad (2)$$

where γ is the surface tension, ε the permittivity of the surrounding medium, Q the volumetric flow rate, I the current carried by the jet, and χ the dimensionless wavelength of the whipping instability. Unlike Eq. (1), Eq. (2) is dimensionally consistent and is evaluated in SI units (γ in N/m, ε in F/m,

Q in m^3/s, I in A, and d_t in m); it is the analytical result derived by Fridrikh et al. for the terminal jet diameter and should likewise be treated as an idealized estimate [2], [5]. As the solvent evaporates, the jet solidifies; slower solidification allows longer stretching and thinner fibers, while residual surface charge tends to limit the achievable deposited mat thickness to roughly 0.5–1 mm [2]. For a portable device, this picture translates into a practical constraint: the needle-to-collector distance, the needle geometry, and the flow rate must be chosen jointly with the achievable output voltage of the supply in order to sustain a stable cone rather than an unstable, dripping, or multi-jet regime.

E. Post-Fabrication Characterization

Once fabricated, electrospun mats are characterized along morphological, mechanical, structural, chemical, and thermal dimensions – most commonly by scanning and transmission electron microscopy, tensile testing, X-ray diffraction, Fourier-transform infrared (FTIR) and Raman spectroscopy, and thermal analysis by differential scanning calorimetry and thermogravimetry [1], [10]. Porosity is particularly relevant: the high, interconnected porosity of these meshes governs exudate drainage and gas exchange, and is quantified by mercury or capillary-flow porosimetry per ASTM standards [6]. In vitro, fibroblasts and stem cells cultured on such meshes show no cytotoxicity and improved infiltration, migration, and proliferation [18].

IV. Portable and Handheld Electrospinning Devices: State of the Art

An electrospinning apparatus consists, essentially, of a high-voltage power supply, a mechanism for controlling the flow of the polymer solution, and a collector [10], as illustrated conceptually in Fig. 1. In benchtop equipment, all three subsystems are large, mains-powered units; portable devices must, instead, miniaturize each of them while preserving process stability. Table II summarizes representative portable and handheld electrospinning devices reported in the literature, together with their reported operating parameters and biomedical validation where available; it constitutes the core comparative contribution of this review.

Xu et al. (2015) reported one of the first battery-operated handheld apparatuses, replacing the syringe pump with a hand-actuated mechanism and a compact high-voltage module, enabling operation without a fixed electrical connection [22]. Mouthuy et al. (2015) combined an EMCO F101 high-voltage converter with a linear-actuator syringe pump under microcontroller control, powered by ten NiMH AA cells for about two hours [24]. Zhao et al. (2020) demonstrated a self-powered melt-electrospinning device reaching 20 kV from a hand-cranked or battery input, targeted at in situ wound dressing [23]. Haik et al. (2017) assessed the clinical feasibility of a handheld device – a precursor of the commercial SpinCare™ system – on a porcine wound model [25]. Owida et al. (2022) built an integrated, low-cost platform from off-the-shelf and open-hardware components at roughly US$202, against

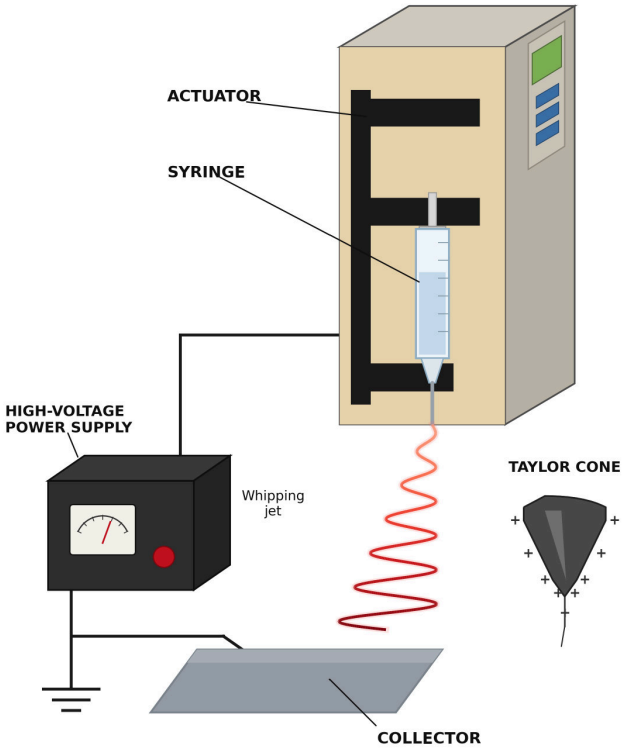


Fig. 1. Conceptual diagram of a portable electrospinning apparatus, common to the devices summarized in Table II: the battery-driven HV supply charges the syringe-dispensed polymer solution, which forms a Taylor cone at the needle tip; the resulting jet undergoes a whipping instability (coiled path) before depositing solid nanofibers on the grounded collector.

US$17,000–60,000 for commercial benchtop systems [26], underscoring cost as a first-order design constraint.

Cruz et al. (2025) introduced a glove-mounted, needleless device, stepping a 1 V input up to 50 kV via a miniaturized diode-capacitor ladder, producing cannabidiol-loaded microfibers validated on porcine skin [27]. Desai et al. (2026) presented "Fiber Gun," an integrated handheld system with high-voltage supply, syringe pump, stepper motor, and a 12 V battery, reporting over 10 h of operation and 100–151 nm fibers [28]. Cai et al. (2026) reported a hand-held multi-process printer combining photo-crosslinking extrusion, spraying, and electrospinning under wireless control, validated in rat and porcine wound models [29]. These devices show a trend toward tighter integration and lower input voltages, achieved by moving voltage multiplication or extra fabrication modes into the miniaturized platform.

Two patterns emerge from Table II. First, there is a persistent trade-off between output-voltage range, regulation quality, and cost: commercial HV modules offer stable, well-documented performance at higher cost, whereas self-built oscillator- or converter-based stages reduce cost but require in-house characterization of output voltage and ripple. Second, and more critical for translation, the reported limitations recur across designs: flow delivery is frequently hand-actuated or

TABLE II
REPRESENTATIVE PORTABLE/HANDHELD ELECTROSPINNING DEVICES: REPORTED OPERATING PARAMETERS, BIOMEDICAL VALIDATION, AND MAIN REPORTED LIMITATION. NR: NOT REPORTED.

Reference	Output voltage	Flow rate	Distance	HV topology	Power source	Biomedical validation	Main limitation
Xu *et al.*, 2015 [22]	NR	NR	NR	Compact HV module	Battery	NR	Hand-actuated; no regulated flow
Mouthuy *et al.*, 2015 [24]	0–14 kV	0–10 mL/h	20 cm	F101 converter (EMCO)	10× NiMH AA	Bench characterization	∼2 h runtime; no in vivo data
Zhao *et al.*, 2020 [23]	20 kV	NR	10 cm	Oscillator + voltage doubler	Self-powered/battery	In situ wound dressing (targeted)	Melt process; limited biological data
Haik *et al.*, 2017 [25]	25 kV	4.5 mL/h	20 cm	Commercial HV module	Battery	Porcine wound model	Feasibility only; proprietary module
Owida *et al.*, 2022 [26]	NR	NR	NR	Low-cost off-the-shelf HV module	Battery/mains	NR	No biomedical validation
Cruz *et al.*, 2025 [27]	50 kV	NR	NR	Miniaturized CW ladder	Wearable/glove battery	Porcine skin model	Microfiber scale; single drug
Desai *et al.*, 2026 [28]	NR	NR	NR	Integrated HV module	12 V battery	In vivo wound healing	Early access; few formulations
Cai *et al.*, 2026 [29]	NR	NR	NR	Multi-process (spin/spray/extrude)	Battery, wireless control	Rat and porcine wound models	Higher system complexity

only coarsely regulated; battery autonomy is short; and, above all, biomedical validation rarely advances beyond proof-of-concept or small-animal and porcine models, while quantitative electrical-safety data are essentially absent. The most recent platforms [27]–[29] improve integration and lower the input voltage, yet none reports the leakage-current or dielectric-strength characterization that clinical use would demand. These recurring gaps – regulated low-power delivery, longer autonomy, and documented electrical safety – frame the power-supply comparison developed in the following section.

V. HIGH-VOLTAGE POWER SUPPLY TOPOLOGIES FOR PORTABLE APPLICATIONS

Generating a stable output in the range of 10–30 kV from a low-voltage battery, while keeping the supply compact and light, constitutes the central power-electronics challenge of portable electrospinning. Two topologies dominate the literature on portable and low-cost high-voltage sources: the Cockcroft–Walton voltage multiplier and the flyback converter.

A. Cockcroft–Walton Voltage Multiplier

The CW multiplier, originally proposed for particle acceleration [30], cascades diodes and capacitors driven by an AC source to multiply the peak input voltage without a high-turns-ratio transformer (Fig. 2). It has been used as a compact high-voltage DC source for capacitor-bank charging in a tokamak [31], as a high-gain DC-DC converter for impulse loads [32], in a transformer-less 0.1/12 kV matrix converter with low-ripple input current and reduced device stress [33], and in a transformer-less high step-up converter that avoids the impractical duty ratios of a conventional boost stage [34].

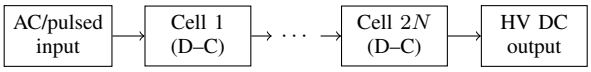


Fig. 2. Conceptual block diagram of a Cockcroft–Walton voltage multiplier. Each diode–capacitor cell transfers approximately one peak-input increment; an ideal N-stage (i.e., $2N$-cell) multiplier therefore approaches an unloaded output on the order of $2N\,V_{\mathrm{peak}}$. The practical output is lower, being reduced by load current, output ripple, and stray capacitance.

While the CW multiplier offers a modular, transformerless path to high voltage gain, its principal drawbacks for a handheld application are threefold: the number of discrete diode-capacitor cells required to reach tens of kilovolts; the increase in output impedance and voltage ripple as the cell count grows; and the absence of intrinsic galvanic isolation between the low-voltage input and the high-voltage output, unless an isolated driving stage is added beforehand. The device of Cruz et al. (2025) confirms that a CW-type ladder can nonetheless be miniaturized [27].

B. Flyback Converter

The flyback converter is a buck–boost-derived, transformer-isolated topology that stores energy in the transformer core during the switch-on interval and delivers it to the secondary during the switch-off interval. Because the transformer provides galvanic isolation by design, and a single transformer can be wound with a very high turns ratio, the flyback topology is the one most widely reported as the preferred choice for compact, low-power, high-voltage sources in electrospinning applications.

Aung et al. (2019) built a zero-voltage-switching (ZVS) flyback high-voltage supply for electrospinning, motivated by the high cost of commercial units [35]. Hapidin et al. (2017) used

a series-configuration Mazzilli ZVS flyback for needleless electrospinning [36], and Parkash et al. (2021) implemented a flyback high-voltage supply for nanofiber production [37]. Yusro and Kadarisman (2022) reported a low-cost setup with adjustable flow rate and voltage for biomedical fibers [38], while Muthi'ah et al. (2020) obtained 11–20 kV from a 12 V input with a Mazzilli ZVS driver, illustrating the high, low-cost gain of a repurposed flyback transformer [39].

Beyond electrospinning, Barsoum and Stanley (2015) found the flyback the most suitable approach for a small, simple, low-cost high-voltage supply [40]; Liu and Dong (2015) reported a 0–30 kV AC-input boost-doubler supply [41]; Manur et al. (2024) proposed a load-invariant, wireless-power-transfer high-voltage supply reaching 5 kV at 88% efficiency [42]; and Tang et al. (2017) presented a symmetric cascaded pulsed supply with reduced device voltage stress [43].

C. Comparative Discussion

Table III summarizes the trade-offs between the CW multiplier and the flyback converter for a portable, battery-powered electrospinning application.

The electrospinning-specific literature converges toward the flyback topology [35]–[39], whereas CW multipliers appear mostly in generic high-voltage or pulsed-power work [31]–[34], the wearable device of Cruz et al. showing that the CW family remains relevant when extreme miniaturization is the priority [27]. The ripple comparison in Table III is qualitative, since a rigorous one would require fixing switching frequency, load, output capacitance, control scheme, and equivalent impedance – rarely reported on a common basis. Subject to this caveat, the convergence toward the flyback, together with its inherent galvanic isolation, makes it the most frequently reported candidate for portable applications and motivates the feasibility discussion below.

VI. Feasibility of a Flyback-Based Stage: Cost, Safety, and Component Considerations

Building on the above, this section assesses – without proposing a specific design – the feasibility of a flyback stage repurposed from a cathode-ray-tube (CRT) television transformer, a strategy already reported for electrospinning [35]–[37], [39]. Such transformers are typically self-rectified: the high-voltage secondary embeds a rectifier diode and delivers pulsed DC directly, without an external multiplier. They can be energized by different switching arrangements, including single-switch converters and resonant ZVS drivers [39]. Figure 3 gives a simplified functional representation, emphasizing energy conversion and galvanic isolation rather than a particular switching topology.

From a cost and availability standpoint, CRT flybacks are discarded in considerable numbers as CRT displays are decommissioned, rendering them essentially free or very low-cost, in marked contrast with commercial high-voltage modules such as the EMCO F101 employed by Mouthuy et al. [24]. From an electrical standpoint, the flyback transformer isolates the high-voltage secondary from the low-voltage primary. Flyback

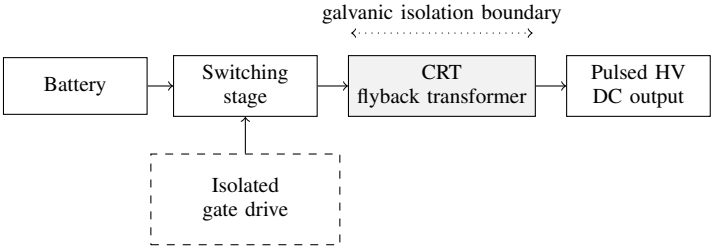


Fig. 3. Simplified functional diagram of a CRT flyback-based high-voltage stage. The switching stage is represented generically, while the transformer provides galvanic isolation between the low-voltage primary circuit and the high-voltage output.

operation can facilitate current limitation by controlling the energy transferred per switching cycle; however, the maximum output current and stored energy must be experimentally verified for each design.

Two important caveats must, however, be stated. First, galvanic isolation is a potentially safety-enhancing design feature, but it is insufficient by itself to establish suitability for clinical use. Establishing medical safety would additionally require leakage-current measurement, short-circuit-current characterization, dielectric-strength and insulation testing, and analysis of behavior under fault conditions – none of which follows automatically from the presence of an isolation transformer. Second, repurposed CRT flyback transformers may be suitable for low-cost proof-of-concept prototypes; however, their lack of traceable specifications and lot-to-lot reproducibility limits their suitability for clinical translation, which would instead call for a professionally specified and qualified high-voltage component. This absence of standardized documentation is one of the central challenges discussed in the following section.

VII. Challenges and Future Directions

Despite this convergence, several challenges remain. First, component standardization: designs relying on generic or salvaged flyback transformers lack public datasheets, so each group must bench-characterize its own units [35], [37]. Second, output regulation: low-cost flyback and CW sources typically run open-loop or with coarse duty-cycle control; closed-loop feedback on the output voltage itself – rather than the primary-side duty cycle – remains underexplored, and could exploit the critical-voltage relation of Eq. (1) to set an adaptive target per solution. Third, regulatory validation: clinical translation, as in the SpinCare™ case [25], demands leakage-current, short-circuit, and dielectric-strength testing well beyond a benchtop prototype, an area still sparsely reported. Finally, miniaturization versus output power: shrinking the HV stage and battery draw while sustaining a stable Taylor cone across solution viscosities remains an active problem for both topology families [27], [31], [42].

VIII. Conclusion

This review examined portable electrospinning devices for biomedical use along three axes: the characterization of the

TABLE III
COMPARISON OF HV TOPOLOGIES FOR PORTABLE ELECTROSPINNING.

Criterion	**Cockcroft–Walton multiplier**	**Flyback converter**
Galvanic isolation	Not inherent; requires an isolated driving stage	Inherent, provided by the transformer
Component count	Grows with the number of cells (diodes + capacitors)	Single magnetic component plus switch and driver
Cost	Low per cell, but rises with voltage target	Very low when repurposing salvaged CRT transformers
Output ripple	Application-dependent: set by switching frequency, load current, per-cell capacitance, and equivalent output impedance; tends to grow with cell count and load	Application-dependent: set by switching frequency, output capacitance, load, and control scheme; not intrinsically lower and must be evaluated per design
Reported use in electrospinning	Mostly generic HV applications [31]–[34]; miniaturized specifically in [27]	Directly reported for electrospinning [35]–[39]

resulting nanofibers and the physics of Taylor-cone jet formation; the state of the art of portable and handheld devices; and the high-voltage electronics that make them portable. While Cockcroft–Walton multipliers remain attractive for extreme miniaturization, flyback-based supplies are more frequently reported in electrospinning. Repurposed CRT flyback transformers represent a low-cost option for proof-of-concept prototyping, but their lack of traceability and reproducibility limits clinical translation. Integrating process requirements, device architectures, and power-supply constraints, we highlight the trade-offs that currently constrain clinical adoption. Galvanic isolation and low cost, though advantageous, do not by themselves establish clinical suitability; addressing component standardization, closed-loop regulation, and regulatory validation – including systematic electrical-safety testing – should guide the translation of these prototypes into clinically viable in situ applicators.

REFERENCES

[1] B. Abadi *et al.*, "Electrospun hybrid nanofibers: Fabrication, characterization, and biomedical applications," *Front. Bioeng. Biotechnol.*, vol. 10, 2022, doi: 10.3389/fbioe.2022.986975.

[2] J. Xue, T. Wu, Y. Dai, and Y. Xia, "Electrospinning and electrospun nanofibers: Methods, materials, and applications," *Chem. Rev.*, vol. 119, no. 8, pp. 5298–5415, 2019, doi: 10.1021/acs.chemrev.8b00593.

[3] J.-S. Park, "Electrospinning and its applications," *Adv. Nat. Sci.: Nanosci. Nanotechnol.*, vol. 1, p. 043002, 2010, doi: 10.1088/2043-6262/1/4/043002.

[4] G. I. Taylor, "Electrically driven jets," *Proc. R. Soc. Lond. A*, vol. 313, no. 1515, pp. 453–475, 1969, doi: 10.1098/rspa.1969.0205.

[5] S. V. Fridrikh, J. H. Yu, M. P. Brenner, and G. C. Rutledge, "Controlling the fiber diameter during electrospinning," *Phys. Rev. Lett.*, vol. 90, no. 14, art. 144502, 2003, doi: 10.1103/PhysRevLett.90.144502.

[6] X. Lu, L. Zhou, and W. Song, "Recent progress of electrospun nanofiber dressing in the promotion of wound healing," *Polymers*, vol. 16, no. 18, art. 2596, 2024, doi: 10.3390/polym16182596.

[7] M. M. Abdul Hameed, S. A. P. Mohamed Khan, B. M. Thamer, N. Rajkumar, H. El-Hamshary, and M. El-Newehy, "Electrospun nanofibers for drug delivery applications: Methods and mechanism," *Polym. Adv. Technol.*, vol. 34, no. 1, pp. 6–23, 2022, doi: 10.1002/pat.5884.

[8] N. Angel, S. Li, F. Yan, and L. Kong, "Recent advances in electrospinning of nanofibers from bio-based carbohydrate polymers and their applications," *Trends Food Sci. Technol.*, vol. 120, pp. 308–324, 2022, doi: 10.1016/j.tifs.2022.01.003.

[9] B. A. Chinnappan, M. Krishnaswamy, H. Xu, and M. E. Hoque, "Electrospinning of biomedical nanofibers/nanomembranes: Effects of process parameters," *Polymers*, vol. 14, no. 18, art. 3719, 2022, doi: 10.3390/polym14183719.

[10] M. Rahmati, D. K. Mills, A. M. Urbanska, M. R. Saeb, J. R. Venugopal, S. Ramakrishna, and M. Mozafari, "Electrospinning for tissue engineering applications," *Prog. Mater. Sci.*, vol. 117, art. 100721, 2021, doi: 10.1016/j.pmatsci.2020.100721.

[11] J. J. Grant, S. C. Pillai, S. Hehir, M. McAfee, and A. Breen, "Biomedical applications of electrospun graphene oxide," *ACS Biomater. Sci. Eng.*, vol. 7, no. 4, pp. 1278–1301, 2021, doi: 10.1021/acsbiomaterials.0c01663.

[12] J. Varshosaz *et al.*, "RGD peptide grafted polybutylene adipate-co-terephthalate/gelatin electrospun nanofibers loaded with a matrix metalloproteinase inhibitor drug for alleviating of wounds: an in vitro/in vivo study," *Drug Dev. Ind. Pharm.*, vol. 46, no. 3, pp. 484–497, 2020, doi: 10.1080/03639045.2020.1730397.

[13] H. Guo *et al.*, "Sequential release of drugs from a dual-delivery system based on pH-responsive nanofibrous mats towards wound care," *J. Mater. Chem. B*, vol. 8, pp. 1759–1770, 2020, doi: 10.1039/c9tb02522g.

[14] Z. Hadisi *et al.*, "Hyaluronic acid (HA)-based silk fibroin/zinc oxide core–shell electrospun dressing for burn wound management," *Macromol. Biosci.*, vol. 20, no. 4, p. 1900328, 2020, doi: 10.1002/mabi.201900328.

[15] J. Song, A. Razzaq, N. U. Khan, H. Iqbal, and J. Ni, "Chitosan/poly(3-hydroxy butyric acid-co-3-hydroxy valeric acid) electrospun nanofibers with cephradine for superficial incisional skin wound infection management," *Int. J. Biol. Macromol.*, vol. 250, p. 126229, 2023, doi: 10.1016/j.ijbiomac.2023.126229.

[16] S. Fahimirad, M. Khaki, E. Ghaznavi-Rad, and H. Abtahi, "Investigation of a novel bilayered PCL/PVA electrospun nanofiber incorporated chitosan-LL37 and chitosan-VEGF nanoparticles as an advanced antibacterial cell growth-promoting wound dressing," *Int. J. Pharm.*, vol. 661, p. 124341, 2024, doi: 10.1016/j.ijpharm.2024.124341.

[17] B. S. Vaz, J. A. V. Costa, and M. G. Morais, "Production of polymeric nanofibers with different conditions of the electrospinning process," *Matéria (Rio de Janeiro)*, vol. 22, no. 2, e11847, 2017, doi: 10.1590/s1517-707620170002.0180.

[18] M. Mejía Suaza, M. Moncada, and Y. Hurtado Henao, "Wet electrospinning and its applications: A review," *TecnoLógicas*, vol. 25, pp. 1–26, 2022, doi: 10.22430/22565337.2223.

[19] A. Refate *et al.*, "Influence of electrospinning parameters on biopolymers nanofibers, with emphasis on cellulose & chitosan," *Heliyon*, vol. 9, no. 6, e17051, 2023, doi: 10.1016/j.heliyon.2023.e17051.

[20] M. Hoseini, S. Hamidi, E. Salehi, A. Mohammadi, F. Mirhoseini, and M. Ravaghi, "Multi-variate multi-objective optimization of production conditions for electro-spun skin scaffold using RSM," *Heliyon*, vol. 10, no. 12, e32941, 2024, doi: 10.1016/j.heliyon.2024.e32941.

[21] M. Khasteband, Y. Sharifi, and A. Akbari, "Chrysin loaded polycaprolactone-chitosan electrospun nanofibers," *Int. J. Biol. Macromol.*, vol. 263, part 1, p. 130250, 2024, doi: 10.1016/j.ijbiomac.2024.130250.

[22] S.-C. Xu, C.-C. Qin, M. Yu, R.-H. Dong, X. Yan, H. Zhao, W.-P. Han, H.-D. Zhang, and Y.-Z. Long, "A battery-operated portable handheld electrospinning apparatus," *Nanoscale*, vol. 7, p. 12351, 2015, doi: 10.1039/c5nr02922h.

[23] Y.-T. Zhao *et al.*, "Self-powered portable melt electrospinning for in situ wound dressing," *J. Nanobiotechnology*, vol. 18, no. 1, p. 111, 2020, doi: 10.1186/s12951-020-00671-w.

[24] P.-A. Mouthuy, L. Groszkowski, and H. Ye, "Performances of a portable electrospinning apparatus," *Biotechnol. Lett.*, vol. 37, no. 5, pp. 1107–1116, 2015, doi: 10.1007/s10529-014-1760-6.

[25] J. Haik, R. Kornhaber, B. Blal, and M. Harats, "The feasibility of a handheld electrospinning device for the application of nanofibrous wound dressings," *Adv. Wound Care*, vol. 6, no. 5, pp. 166–174, 2017, doi: 10.1089/wound.2016.0722.

[26] H. A. Owida, B. Al-haj Moh'd, and M. Al Takrouri, "Designing an integrated low-cost electrospinning device for nanofibrous scaffold fabrication," *HardwareX*, vol. 11, e00250, 2022, doi: 10.1016/j.ohx.2021.e00250.

[27] O. B. Cruz, L. Lou, S. M. A. K. Mohammed, R. T. Murickan, L. Benedetti, Y.-M. Lin, T. Dolmetsch, and A. Agarwal, "Wearable, ultralow power, and needleless electrospinning equipment for cannabidiol-loaded patch fabrication," *ACS Appl. Mater. Interfaces*, vol. 17, no. 34, pp. 48145–48159, 2025, doi: 10.1021/acsami.5c14853.

[28] B. M. A. Desai, S. A. Nadaf, R. P. R. Shivakumar, and P. Biswas, "Fiber gun: A portable nanofiber delivery system for accelerated wound healing," *IEEE Trans. Biomed. Eng.*, early access, 2026, doi: 10.1109/TBME.2026.3679621.

[29] W. Cai, H. Liu, Z. Guo, S. Zhong, K. Luo, J. Zhang, Y. Zhang, H. Cai, W. Jin, X. Dong, and Y. Liu, "A hand-held digital printer for on-site fabrication of micro/nano composite dressings," *ACS Appl. Mater. Interfaces*, vol. 18, no. 25, pp. 34790–34805, 2026, doi: 10.1021/acsami.6c03701.

[30] J. D. Cockcroft and E. T. S. Walton, "Experiments with high velocity positive ions," *Proc. R. Soc. A*, vol. 136, no. 830, pp. 619–630, 1932, doi: 10.1098/rspa.1932.0107.

[31] N. Ahmed, N. Khan, and T. Abbas, "An inverter-fed Cockcroft-Walton multiplier based high voltage DC source for tokamak," in *Proc. 2022 Int. Conf. Emerging Trends in Electrical, Control, and Telecommunication Eng. (ETECTE)*, 2022, doi: 10.1109/ETECTE55893.2022.10007169.

[32] R. Saha, S. Bindu, and R. Harchandani, "Cockcroft-Walton voltage multiplier based high gain DC-DC converter as DC source for impulse load," in *Proc. 2023 World Conf. Commun. Comput. (WCONF)*, Raipur, India, 2023, doi: 10.1109/WCONF58270.2023.10235205.

[33] M. A. Razzak, S. Chakraborty, and Z. Tasneem, "Design of a multistage 0.1/12 kV DC-DC matrix converter using Cockcroft-Walton voltage multiplier topology," in *Proc. 2015 IEEE Int. WIE Conf. Electrical and Computer Eng. (WIECON-ECE)*, Dhaka, Bangladesh, 2015, doi: 10.1109/WIECON-ECE.2015.7443925.

[34] C.-M. Young, M.-H. Chen, T.-A. Chang, and C.-C. Ko, "Transformerless high step-up DC-DC converter with Cockcroft-Walton voltage multiplier," in *Proc. 2011 6th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, 2011, doi: 10.1109/ICIEA.2011.5975846.

[35] Z. P. P. Aung, S. S. Lwin, T. T. Win, and Y. M. Maung, "Design and construction of zero voltage switching flyback converter as a high-voltage power supply for electrospinning," in *Proc. Int. Symp. Environmental-Life Science and Nanoscales Technology (ISENT2019)*, Yangon, Myanmar, 2019.

[36] D. A. Hapidin, I. Saleh, M. M. Munir, and Khairurrijal, "Design and development of a series-configuration Mazzilli zero voltage switching flyback converter as a high-voltage power supply for needleless electrospinning," *Procedia Eng.*, vol. 170, pp. 509–515, 2017, doi: 10.1016/j.proeng.2017.03.081.

[37] V. Parkash, P. Kumar, P. Sharma, and G. Sapra, "Design and implementation of flyback converter as high voltage power supply for nanofibers production," *Mater. Today: Proc.*, 2021, doi: 10.1016/j.matpr.2021.01.857.

[38] M. Yusro and Kadarisman, "Development of low-cost electrospinning to fabricate structured nanofiber for biomedical designs with manageable flowrate and voltage," *Indonesian J. Electron. Electromed. Eng. Med. Informat.*, vol. 4, no. 3, pp. 123–130, 2022, doi: 10.35882/ijeeemi.v4i3.234.

[39] A. Muthi'ah, A. Rahardjo, F. Husnayain, and C. Hudaya, "Design of Mazzilli's zero voltage switching (ZVS) circuit as plasma glow discharge generator," *ELKHA*, vol. 12, no. 2, pp. 112–118, 2020, doi: 10.26418/elkha.v12i2.41769.

[40] N. Barsoum and G. I. Stanley, "Design of high voltage low power supply device," *Universal J. Electr. Electron. Eng.*, vol. 3, no. 1, pp. 6–12, 2015, doi: 10.13189/ujeee.2015.030102.

[41] J. Liu and Z. Dong, "A design for high voltage DC power supply," in *Preprints of the 5th Int. Conf. Electric Utility Deregulation and Restructuring and Power Technologies*, Changsha, China, 2015, doi: 10.1109/DRPT.2015.7432624.

[42] S. J. Manur, R. Mirzadarani, and M. Ghaffarian Niasar, "High voltage, wireless power transfer based DC power supply," in *Proc. 2024 IEEE 10th Int. Power Electronics and Motion Control Conf. (IPEMC-ECCE Asia)*, 2024, doi: 10.1109/IPEMC-ECCEAsia60879.2024.10567438.

[43] M. Tang, L. Shu, G. Qiao, G. Ning, W. Chen, X. Qu, and B. Ji, "A high voltage pulsed power supply with reduced device voltage stress for industrial electrostatic precipitators," in *Proc. 2017 IEEE Energy Conversion Congress and Exposition (ECCE)*, Cincinnati, OH, USA, 2017, pp. 2357–2361, doi: 10.1109/ECCE.2017.8096457.

Proposal of an Automated System for the Langendorff Isolated Heart Perfusion Method

André Rosa Ferreira
Department of Biomedical Engineering
Federal University of ABC-UFABC
São Bernardo do Campo, Brazil
andre.ferreira@ufabc.edu.br

João Salinet
Department of Biomedical Engineering
Federal University of ABC-UFABC
São Bernardo do Campo, Brazil
joao.salinet@ufabc.edu.br

Abstract* -The Langendorff method of isolated heart perfusion is a well-established ex vivo surgical technique in cardiovascular biological and pharmacological research. However, the accuracy of clinical trials is often limited by the oscillations and variability inherent in the manual control of homeostatic parameters. This research presents the development and benchtop validation of an automation platform applied to the Langendorff method., designed to ensure the rigorous stability of critical variables during cardiac arrhythmia trials, such as atrial fibrillation. The proposed modular architecture is based on the transposition of industrial automation technologies to the laboratory environment, using a Programmable Logic Controller (PLC WEG TPW03) vertically integrated with Elipse E3 SCADA data acquisition and supervision software. The PLC operates deterministically and with low latency to control temperature and actuate peristaltic pumps and solenoid valves, managed via Ladder logic. Preliminary fluid dynamic and thermal validation tests demonstrated that the system, using on-off thermal control on the test bench, stabilized the fluid temperature at a median of 37°C (IQR = 0.10°C; σ-sigma = 0.0699°C), exhibiting statistical superiority over the manual method ($p = 4.15 \times 10^{-7}$). Thermal control via a PID loop embedded in the PLC is proposed for future implementation directly on the aortic cannula. Coronary flow was precisely regulated and maintained within the physiological range with a median of 23.5 mL/min (IQR = 0.52 mL/min; σ-sigma = 0.39 mL/min), validating the alignment of the control with the predictive allometric model for leporine specimens. Benchtop results confirm that the platform's robustness and determinism eliminate human operational errors and ensure methodological reproducibility, validating the physical-operational control infrastructure as an indispensable preparatory step for future applications in actual biological assays.

Keywords* - Laboratory automation, Langendorff system, cardiac arrhythmia, temperature control, automated perfusion, biomedical engineering, supervisory system.

I. INTRODUCTION

The isolated heart perfusion method remains a widely used approach in cardiovascular and pharmacological research [1]. Its applicability can be continuously optimized through the development of new protocols for automated monitoring systems [2]. Historically, few organ-based experiments have been as extensively utilized as the Langendorff isolated heart preparation. Initially, this technique was applied to basic cardiac biology studies; subsequently, it was adopted to test the effects of various drugs on coronary vasculature, muscle contraction, and heart rate [1]. Currently, many researchers employ this method to investigate the heart, aiming to develop new treatments that can protect it from ischemia and other forms of injury [3].

Advances in biophysical and biochemical recording techniques over recent decades have transformed the landscape of experimental research. The integration of automated acquisition and control systems enables the simultaneous monitoring of multiple parameters with high precision and reproducibility [2]. This technological leap, combined with the improved cost-effectiveness of modern systems, offsets the limitations inherent in isolated models, allowing for an in-depth analysis of complex phenomena—such as atrial fibrillation (AF)—in controlled environments.

The development of cardiac arrhythmia leads to alterations in the heart's normal rhythm affecting the atria, causing the organ to beat rapidly and irregularly rather than contracting in a coordinated fashion, which impairs efficient blood pumping [3]. The overall prevalence of AF is known to reach approximately 2.9% of the global population. Therefore, the precise identification and quantification of the mechanisms underlying the generation and maintenance of AF are essential for better stratifying the pathology and enhancing the efficacy of treatment guidelines [4].

Automation is of great importance across various sectors of society, driving significant technological advancements and improving process efficiency and reproducibility [2]. In biomedical research, it reduces the occurrence of operational errors by minimizing human factors—such as fatigue, distractions, and performance fluctuations during the workday, thereby enabling continuous, precise control of the variables involved [2].

Automation also facilitates compliance with established standards or procedures by providing detailed process records [2]. This research addresses the development of an experimental testbed for the automation of cardiovascular assays, aiming for strict control of critical variables in the study of cardiac arrhythmia. The system is based on adapting industrial automation technologies to the laboratory environment, vertically integrating a PLC with SCADA software. Using Ladder logic, the PLC operates deterministically and with low latency to control temperature and actuate peristaltic pumps and solenoid valves. In the current bench tests, reservoir temperature was maintained using a dedicated On-Off controller, with the implementation of the PID control loop directly in the PLC reserved for a later stage of development.

The proposed automation for the Langendorff-style perfusion system relies on three critical aspects: control, monitoring, and data acquisition. The system will regulate nutrient supply, volume, flow rate, and temperature during the experiment, playing a significant role in research related to cardiac arrhythmia and atrial fibrillation. Consequently, this research proposes the development and validation of an automation infrastructure for an experimental perfusion module. The objective is to measure the technological precision of the hardware and software in a controlled benchtop environment, with the understanding that this study validates the operational performance of the automated

control system, serving as a physical foundation for future biological validation in atrial fibrillation assays.

II. MATERIALS AND METHOD

A. Evaluation of Current Technology

This section contextualizes available commercial technologies and outlines the biological scenario for the platform's future application. It highlights that the experiments described involving actual rabbit hearts and complex optical/electrical mapping systems represent the methodological goal for which this system is intended. However, the scope of this research is strictly limited to benchtop validation (without the use of animals), employing water as the test fluid to verify the feasibility and precision of the automation by integrating industrial automation technologies into the laboratory environment.

The technological benchmarking encompassed the main global solutions, ranging from conventional manual architectures to high-precision integrated systems, as described below:

• Radnoti (USA): Configured as a manual glass-built platform, based on classic physiological principles. Pressure control is dependent on the hydraulic load (height of the liquid column), and thermal stability is maintained by a water bath with manual thermostatic adjustment. The system does not have integrated feedback loops or automated digital recording mechanisms.

• emka Technologies (France): Presents an integrated solution focused on software processing (IOX2) for real-time acquisition and analysis. Although it allows for protocol programming and automated switching of solutions via solenoid valves, hardware control is dependent on the software layer, lacking industrial programmable logic controllers (PLCs) or SCADA supervisory systems.

• Harvard / Hugo Sachs Elektronik (USA/Germany): Utilizes the patented S2P2C (Solid State Physiological Perfusion Circuit) technology, which replaces conventional glassware with solid Perspex blocks with machined internal channels. This architecture favors laminar flow and offers superior thermal insulation. Operational control is semi-automatic, based on the SCP (Servo Pressure Controller) module, which employs electronic feedback to switch between pressure or constant flow modes. Data acquisition is managed by the ISOHEART software, which is responsible for decoding the analog signals from the transducers and converting physical variations (such as pressure, temperature, and flow) into hemodynamic and contractility parameters in an integrated manner. Using the system's transducers, the software continuously monitors aortic perfusion pressure (PAP) and coronary flow (CF), allowing for precise assessment of vascular resistance and coronary autoregulation under constant flow or pressure regimes.

• AD Instruments (Australia): Modular platform optimized for bio acquisition, structured on Power Lab hardware and LabChart software. The system prioritizes the fidelity of the physiological signal through the direct perfusion core, which positions pressure, flow, and temperature transducers adjacent to the aortic cannula, minimizing the dead volume that could dampen the captured waves. This proximity ensures the direct recording of aortic perfusion pressure (PAP), an analog electrical signal that translates the force exerted by the fluid at the entrance of the coronary system for the calculation of vascular resistance, and coronary flow (CF), which measures in real time the volume of fluid infused into the organ per minute. Flow control is performed by the STH module, which achieves a nominal thermal accuracy between ±0.02°C and 0.05°C. However, the absence of deterministic architecture (PLC) and industrial supervision systems limits the validation of real-time statistical capability.

• A viable alternative is the implementation of the Langendorff system, using independent modules in parts, adapting to local needs for specific lines of customized research, resulting in systems with lower operational costs when compared to ready-made commercial solutions. The Heart Lab at UFABC adopted this customized development strategy for its research line. For cardiac mapping, the experimental protocol uses a rabbit heart to capture electrical and optical signals under sinus rhythm and atrial fibrillation. Invasive electrical signals are obtained directly from the epicardium using a multi-electrode array, while non-invasive signals are recorded by electrodes arranged in a hexagonal tank that simulates the impedance of the human torso. After digital filtering of these electrical signals to attenuate noise, optical signals are acquired by panoramic fluorescence using the Di-4-ANBDQPQ dye and three high-speed cameras. The integration of these data allows the projection of isopotential maps and the calculation of activation fronts and conduction velocities on the 3D geometry of the organ. Another analysis, using fast Fourier transform spectral analysis, identifies the dominant frequency, differentiating the regularity of sinus rhythm from the high spatiotemporal complexity characteristic of atrial fibrillation. The scope of this research is limited exclusively to the physical and functional validation of the automated test bench using water, serving as an indispensable preparatory step for future biological assays.

B. Specification of the new system

Comparative analysis of commercial technologies focused on the Langendorff model prioritizes the acquisition of physiological signals, presenting limitations regarding robustness and industrial determinism.

This research project presents the development of an automated experimental bench module for cardiac perfusion, based on the transposition of industrial automation concepts to the laboratory environment.

The proposed architecture is based on the vertical integration between dedicated control hardware and supervisory software. This approach ensures that the processing of critical variables occurs independently of fluctuations in the host computer's operating system.

As a result, a robust and deterministic platform is obtained, capable of ensuring the precision and reproducibility requirements necessary for the feasibility and standardization of biological assays.

The feasibility of this operational stability and the fulfillment of high scientific rigor criteria are based on five structural technological pillars:

•Industrial SCADA Architecture (Elipse E3): Implements a centralized supervision layer unprecedented in Langendorff

systems. Through an integrated Human-Machine Interface (IHM), the system provides complete parameter control, full traceability, and automated data logging in a structured environment.

•Deterministic Control via Programmable Logic Controller (PLC) model (WEG TPW03): It has 24 digital inputs and 16 outputs, a total of 40 I/O points. The use of a PLC ensures hard real-time control. This architecture eliminates computational latencies inherent in conventional acquisition software, resulting in high experimental reliability and operational stability. The PLC was programmed using Ladder logic via dedicated TPW PCLINK software, which uses relay logic with contacts and coils following the IEC 61131-3 standard[2]. It will control the operation of 5 peristaltic pumps: 1 12VDC (Voltage Common Collector) pump, operating current ranging from 0.25A to 1A, with an average flow rate greater than or equal to 70mL/min, model Kamoer NKP-DC-B10B, applied for dosages that require refined resolution of smaller volumes, due to its reduced flow dynamics, and 4 24VDC peristaltic pumps, current ranging from 3A to 5A, with an average flow rate greater than or equal to 2000 mL/min, model Watson Marlow 313 D, intended for rapid pumping of larger volumes due to their high flow rate.

- The temperature control monitoring for the proposed research was planned in two complementary stages to ensure the thermal stability of the perfusion solution, accounting for the distinction between the benchtop validation performed and future developments. The first approach for the current validation involves basic control and maintenance of the reservoir temperature (water bath) using a commercial digital controller, model XH-W3001. This device operates in a closed-loop (on-off) mode, activating a 750W electric heating element to stabilize the water bath temperature within the desired physiological range. The second approach consists of a proposal for future work to introduce thermal monitoring directly at the critical point of the circuit. For this, the proposal to be tested is to fix a waterproof DS18B20 temperature sensor directly to the aortic inlet cannula of the ex vivo organ. The signals generated by this sensor are transmitted in real time to the PLC, which executes a Proportional, Integral and Derivative (PID) control algorithm. This approach allows the system to identify and compensate for the caloric losses that occur along the fluid path between the reservoir and the heart, ensuring that the nutrient solution reaches the myocardial tissue at the ideal physiological temperature and free from oscillations.

- Statistical Validation: The project introduces statistical analysis to verify the quality of the process. Measuring temperature and flow parameters. These indices attest that the process is statistically capable and controlled, increasing the scientific rigor and reproducibility of the tests. The Figure below shows the Langendorff proposed system.

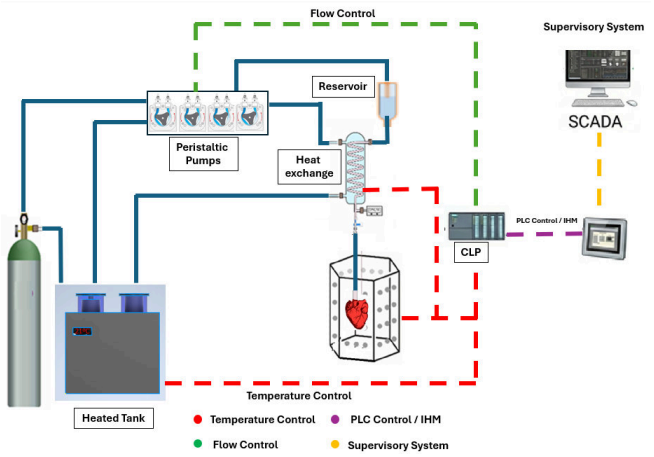


Fig. 1. Architecture of the automated experimental perfusion system

C. Development and validation

Platform validation was conducted on the benchtop by physically simulating the dynamic variables of the perfusion process—temperature, volume, and flow rate—using water as the test fluid. The testing focused strictly on the functional and thermo-fluid-dynamic verification of the automation system (PLC/SCADA), serving as a preparatory step prior to biological or hemodynamic validation using the isolated-organ Langendorff model.

The peristaltic pump drives, sized for the automated delivery of nutrients and medications, were validated using water as the test fluid. This enabled an assessment of the automated multivariable control of the fluidic circuit, ensuring the stability and precision required for the subsequent application of the protocol in biological assays.

The design of the benchtop tests was based on the hemodynamic and thermal parameters standardized by Dhein [5] for the ex vivo experimental model using rabbits, which typically has an average weight of 1.6kg and a cardiac mass between 8 and 9 g. Under these biological conditions, the reference coronary flow (CF) has an operating range of 22 to 25mL/min according to Dhein[5]. Perfusion pressure acts as an indirect indicator of coronary vascular resistance. To ensure experimental accuracy, coronary flow must be normalized as a function of cardiac mass, as demonstrated by the equation below[6]. This equation can be used to adjust the correct flow.

$$CF = 7.43 \times HW^{0.56} \quad (1)$$

Where CF represents the ideal coronary flow (mL/min) and HW corresponds to the heart mass (g). Applying the above equation and considering three different masses: 8.0g, 8.5g and 9.0g, we obtain the respectively CF as presented in Figure below. Allowing to obtain the linear allometric curve validation for CF.

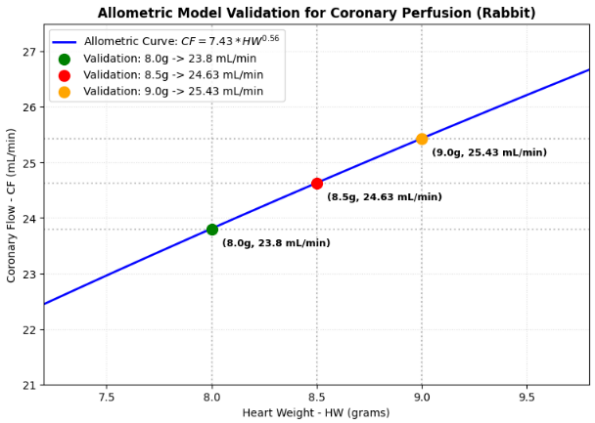


Fig. 2. Validation curve of the allometric model for determining ideal coronary flow as a function of weight

Under these biological conditions, the reference CF recommended in the literature presents a baseline operating range of 22 to 25mL/min[6], which validates the alignment of the mathematical model with the physiological behavior of the organ.

D. Methodology / test procedures

The validation of the automated dosing system was divided into three distinct, sequential and complementary phases, isolating the mechanical factors from the thermodynamic ones:

1) Volumetric and flow calibration tests of the peristaltic pumps: To conduct the tests, it is necessary to perform volumetric and flow rate calibration of the pumps. This test is carried out to evaluate mechanical precision. The repeatability of the obtained data formed the basis for the timing calculations and the parameterization of the counting blocks within the PLC logic.

Purging and Homogenization Procedure: Before starting any series of measurements, the purging cycle is manually activated using physical pushbuttons integrated into the control panel. This procedure ensured the complete elimination of residual air bubbles in the lines and the homogeneous filling of the flexible hoses, mitigating measurement errors. With the aid of graduated beaker to receive the liquid, the quantity in mL to be discarded was observed.

Low Flow Protocol for Kamoer NKP-DC-B10B 12VDC Microdosing Peristaltic Pump: The actuator was continuously activated to measure the time required to reach the fixed volume of 50mL, the graduated value found and noted on the beaker. The test was replicated 50 consecutive times to mitigate random errors. The time taken was recorded using a digital stopwatch, achieving an average of 31.25s. To calculate the flow rate (Q – mL/min), volume (V – mL) over time (t – s), the following flow rate formula was applied:

$$Q = V / t \quad (2)$$

Resulting in Q = 1.6mL/s or 96mL/min (average of readings)

Due to its high resolution and precision in small volumes, this actuator was indexed in the PLC logic for the controlled infusion of concentrated drugs and nutrients. Water was used in this test.

• High Flow Rate Protocol using a Watson Marlow 313 D 24VDC Main Circuit peristaltic pump: the same procedure as the first pump was performed, repeating the sampling for 50 consecutive repetitions with the same target volume (50mL). The average filling time is 1.88 s, and the resulting Q = 26.5 mL/s or 1590mL/min (average of readings).

2) Fluid Thermal Control and Stability Test (Water Bath): Once the mechanical flow rates were calibrated and validated, an independent protocol was implemented to evaluate the thermodynamic behavior of the system, focusing on heat exchange capacity and fluid stabilization at the target physiological temperature of 38°C (critical parameter established by Dhein[5][8] for the ex vivo rabbit model).

Procedure: The four circular glass reservoirs containing water with were immersed in a water bath inside a reservoir simulating different nutrients. The power electrical circuit consists of a 750W tubular electric resistance driven by a 127VAC industrial electromechanical contactor, its switching managed by a model XH-W3001 digital temperature controller. Continuous monitoring of the fluid temperature is performed by means of a sensor coupled to the controller as an RTD (resistance temperature detector) thermocouple. The controller compares the measured temperature with the desired temperature (setpoint) which is pre-defined at 38.5°C. With the temperature set in the controller, the contactor will be activated only when the controller determines that its output should be turned on, switching a supply voltage to activate the contactor, thus turning on the resistance. As soon as the temperature reaches 38.5°C, the controller determines that its output should be switched off, opening the contactor contacts and switching off the electrical resistance.

Monitoring: Temporal data were collected from the fluid at ambient temperature until it reached the setpoint of 37°C to 39°C, so that an average of 38°C was maintained, with the temperature adjusted for the process. The data was collected by activating the low-flow microdosing peristaltic pump, measuring the temperature with a Vonder brand portable digital probe thermometer with a temperature range of -50°C to 300°C, taking 50 samples directly in the beaker for reading. The temperature control system took 40 minutes to reach a temperature of 40°C, adjusting its temperature, and then it stabilized at 39°C at its setpoint.

3) Testing of systemic integration and reproducibility tests (PLC-IHM): The test can evaluate the reliability of the closed automation loop, the integration of the PLC with the supervisory system.

Procedure: Through the interface developed in the Easy Builder Pro software installed on the IHM, you can operate by inserting random and standardized volumetric values for each operational test of the peristaltic pumps, carrying out operational tests of the peristaltic pumps. The PLC processed the input volumes, internally converted the signals using the previously calculated flow constants 1.6 mL/s and 26.5 mL/s and activated the respective digital outputs to drive the pumps.

Measurement: The final volumes collected at the output of the automated system were checked in a graduated beaker

to perform the measurement in mL at a temperature established above.

In terms of thermal performance, the literature shows that the Krebs-Henseleit solution (perfusion) must reach the cranial portion of the aorta while maintaining a strict temperature of 37°C. Since thermal losses along the fluid circuit require the solution to be pre-heated indirectly in a water bath, adjusted between 41.5°C and 42°C, continuous monitoring at the aortic cannula inlet is mandatory. This approach is being developed with the aim of introducing thermal monitoring directly at the critical point of the fluid circuit to be tested using a waterproof DS18B20 temperature sensor directly on the aortic inlet cannula of the heart. The signals generated by this sensor will be transmitted in real time to the PLC, which executes a PID control algorithm. This approach will allow the system to identify and compensate for the caloric losses that occur along the fluid's path between the reservoir and the heart, ensuring that the nutrient solution reaches the myocardial tissue at the ideal physiological temperature and free from fluctuations.

The operating interface and experiment management were centralized in an IHM, developed specifically for monitoring process variables under the Elipse E3 SCADA platform. This system integrated the functions of supervision, control, and data acquisition, working together with the PLC to manage the dosing routines and thermal control of the assay in a coordinated manner.

E. Statistical Analysis and Experimental Validation

Validation of the automated system was carried out through benchtop tests, reproducing operational parameters in a controlled manner and without the presence of animals. The trials were divided into two constituent groups: the control group, operated via manual method, and the experimental group, conducted by the automated system. To ensure statistical robustness, a sample of 50 readings per group (n=50) was adopted, allowing a comparative analysis between the two groups.

The processing and statistical analysis of experimental data were carried out using a dedicated Python program, simulated on Google Collab.

Initially, data distribution was assessed using the Shapiro-Wilk normality test. To confirm whether thermal and hemodynamic variability follows the Gaussian distribution. The criterion was the value of $p>0.05$, which validated or applied parametric tests, thereby ensuring that the system operates in a predictable and controlled manner.

Once the normality of the data was confirmed ($p>0.05$), one-way analysis of variance (ANOVA) was applied to Compare the means and variability between the control and experimental groups.

To assess stability over a given period, repeated measures ANOVA were used, comparing readings at different time intervals (T1, T2, T3), where the null hypothesis (H0: $p>0.05$) indicates the maintenance of system stability[8].

The boxplot graph used the median as the central line; the value divides the data set in half. In the body of the boxplot, it is delimited by the 1st Quartile (Q1-25%) at the base and the 3rd Quartile (Q3-75%) at the top.

III. RESULTS AND DISCUSSION

A. Thermal Monitoring and Control

The thermal monitoring analysis was performed in real time and continuously, ensuring the variable was maintained within an operating range between 36.5°C and 37.5°C.

The statistical analyses, performed in a Python environment, consolidated the system's performance metrics. The boxplot diagram presented in figure below shows the comparative accuracy profile between the control and experimental groups, highlighting the marked reduction in data dispersion obtained through the automated system[8].

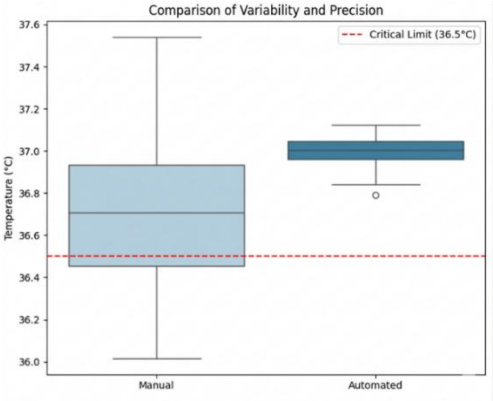


Fig. 3 - Statistical analysis of experimental temperature data

The experimental data measured for the automated method confirmed temperature stabilization at a median of 37°C, with an interquartile range (IQR) of 0.10°C. This narrow dispersion reflects the concentration of 50% of the central readings, with a standard deviation of 0.0699°C. The Gaussian distribution was validated by the normality test ($p = 0.2616$). The analysis of variance (ANOVA) confirmed the statistical distinction between the manual and automated methods ($p = 4.15 \times 10^{-7}$), while the temporal stability test ($p = 0.7172$) demonstrated the absence of drift in the system over time.

B. Coronary Flow Monitoring and Control

To ensure accuracy in perfusion control, the automated system operates under a control loop to maintain coronary flow between 22 and 25mL/min.

To validate the efficiency of the automated system, a lower specification limit (LSL) of 21mL/min was established, a level below which myocardial oxygenation becomes insufficient, and an upper specification limit (USL) of 26mL/min to prevent the risk of endothelial injury and edema formation. The system's operating point was set at 23.5mL/min, aiming to guarantee ideal stability for the experimental model.

Statistical analyses of the experimental data, processed in Python, consolidated the performance metrics presented in the box plot in the figure below, highlighting the precision comparison between the control and experimental groups and emphasizing the lower dispersion in the automated system [8].

Analysis of the measured experimental flow revealed a normal distribution of readings, validated by the normality test ($p = 0.742$). The system achieved stabilization within the desired central range, with a median of 23.5mL/min, an interquartile range (IQR) of 0.52mL/min, and a standard deviation of 0.39mL/min, demonstrating high precision in volumetric delivery. Analysis of variance (ANOVA) confirmed a statistically significant difference compared to the manual method ($p < 0.0001$).

Regarding temporal behavior, repeated-measures ANOVA confirmed overall baseline stability over extended periods ($p > 0.05$), demonstrating the absence of long-term operational drift. However, high-temporal-resolution analyses detected short-term transient fluctuations ($p < 0.05$) attributable to the intrinsic pulsatile dynamics of the peristaltic pumping mechanism.

Although these transient oscillations remain within acceptable physiological limits (lower and upper physiological limits), their identification provides a clear rationale for future system refinements, such as the implementation of pulsation dampeners and continuous closed-loop control algorithms.

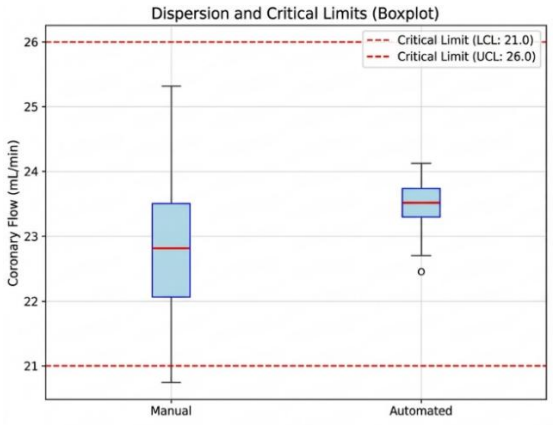


Fig. 4 - Statistical analysis of experimental flow data

The preliminary bench tests carried out can prove the feasibility of applying industrial automation concepts to the Langendorff system. The separation between the SCADA supervision layer (Elipse E3) and the real-time control executed by the PLC eliminated computational latencies typical of general-purpose operating systems. In fluid dynamics, the automated dosing routines and multivariable control ensured precision in the volumetric calibration of the peristaltic pumps.

In thermal control, the automated system stabilized the temperature of the solutions with Gaussian behavior and low dispersion (IQR = 0.10ºC and sigma(σ) = 0.0699ºC), showing statistical superiority over the manual method ($p = 4.15 \times 10^{-7}$). However, although coronary flow showed a significant reduction in variability compared to manual management, fluid dynamics revealed transient fluctuations and temporal instability ($p < 0.05$).

These results indicate the need for future refinement through dedicated continuous control algorithms and validate the proposal to integrate a thermal sensor directly into the aortic inlet cannula to mitigate caloric losses in the circuit. The developed platform establishes itself as a highly deterministic, reproducible, and low-cost alternative, offering cardiovascular research a robust tool for the standardization of ex vivo assays.

Therefore, benchtop tests were parameterized to reproduce this critical scenario, evaluating the ability of the PID algorithm embedded in the PLC to mitigate the ambient thermal gradient and maintain the target of 37°C at the aortic inlet under the stipulated flow rates (22 to 25mL/min).

The operation interface and experiment management were centralized in a Human-Machine Interface (HMI) developed specifically to monitor process variations within the Elipse E3 SCADA platform. This system integrates supervision, control, and data acquisition functions, working in conjunction with the PLC to coordinate the management of dosing routines and thermal control during the test.

IV. CONCLUSION

This study presents a review of major Langendorff systems, highlighting their key advantages and disadvantages. Based on this survey, a new system is proposed that utilizes a deterministic architecture based on a Programmable Logic Controller (WEG TPW03 PLC) integrated with Elipse E3 SCADA supervisory software.

This configuration enables greater precision and reproducibility in controlling critical ex vivo perfusion variables, eliminating the computational latencies inherent in commercial platforms that rely on conventional operating systems. As an initial validation, bench tests were conducted to confirm the efficiency of the industrial automation platform when applied to a laboratory setting; results demonstrated superior thermal and fluid-dynamic stability and a significant reduction in parameter variance compared to manual operation.

Volumetric calibration tests validated the precision of the peristaltic pumps against the allometric model for rabbit hearts. Thermodynamic control of the setup—operating in an on-off mode via secondary hardware—stabilized the temperature with low dispersion (IQR = 0.10°C and sigma-σ = 0.0699°C), demonstrating statistical superiority over the manual method ($p = 4.15 \times 10^{-7}$). Although coronary flow control reduced overall variability ($p < 0.0001$), the identification of transient fluctuations informs future developments in the Langendorff system's automation design; these plans include implementing a PID algorithm linked to a DS18B20 sensor placed directly on the aortic cannula to mitigate heat loss.

Thus, the platform establishes itself as a high-fidelity, low-cost automation solution for rigorous variable control, fully validated in both physical and technological terms to support—in future stages—complex biological protocols involving physiological signal acquisition and optical/electrical mapping.

CONFLICT OF INTEREST

There is no conflict of interest; this study simply presents a more reliable response to existing treatments, improving their accuracy.

ACKNOWLEDGMENT

This study is supported by grant no. 2018/25606-2, São Paulo Research Foundation (FAPESP) and CNPq: National Institute of Photonics (INTERAS, Grant: 409174/2024-6

REFERENCES

[1] K. E. Baker e M. J. Curtis, “Left regional cardiac perfusion in vitro with platelet-activating factor, norepinephrine and K+ reveals that ischaemic arrhythmias are caused by independent effects of endogenous ‘mediators’ facilitated by interactions, and moderated by paradoxical antagonism”, *British Journal of Pharmacology*, v. 142, n. 2, p. 352–366, 2004, doi: https://doi.org/10.1038/sj.bjp.0705767.

[2] C. C. de MORAES e P. de L. CASTRUCCI, Automation Engineering, 2nd ed.: LTC, 2006. Accessed on: January 8, 2026. [Online]. Available at: https://integrada.minhabiblioteca.com.br/reader/books/978-85-216-1976-5

[3] R. M. Bell, M. M. Mocanu, e D. M. Yellon, “Retrograde heart perfusion: The Langendorff technique of isolated heart perfusion”, *Journal of Molecular and Cellular Cardiology*, v. 50, n. 6, p. 940–950, jun. 2011, doi: 10.1016/j.yjmcc.2011.02.018.

[4] J. Salinet *et al.*, “Electrocardiographic Imaging for Atrial Fibrillation: A Perspective From Computer Models and Animal Experiments to Clinical Value”, *Front. Physiol.*, v. 12, p. 653013, abr. 2021, doi: 10.3389/fphys.2021.653013.

[5] S. Dhein, M. Delmar, e F. W. Mohr, Org., *Practical Methods in Cardiovascular Research*. em SpringerLink Bücher. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2005. doi: 10.1007/b137833.

[6] P. Kirchhof *et al.*, “Early Rhythm-Control Therapy in Patients with Atrial Fibrillation”, *N Engl J Med*, v. 383, n. 14, p. 1305–1316, out. 2020, doi: 10.1056/NEJMoa2019422.

[7] S. Dhein, “The Langendorff Heart”, em *Practical Methods in Cardiovascular Research*, S. Dhein, F. W. Mohr, e M. Delmar, Org., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, p. 155–172. doi: 10.1007/3-540-26574-0_9.

[8] V. R Richard D. SHARPE, Norean e VELLEMAN, Paul F., Applied statistics.

Corresponding author information

Author: André Rosa Ferreira
Institution: Federal University of ABC
City: São Bernardo do Campo
Country: Brazil
Email: andre.ferreira@ufabc.edu.br

CHARACTERIZATION OF THE SPEED OF SOUND IN PLASTISOL SPECIMENS FOR ULTRASOUND PHANTOM FABRICATION

I. C. A. de Andrade
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ignezaraujo@peb.ufrj.br

J. Oliveira-Junior
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
jessejunior@peb.ufrj.br

W. C. A. Pereira
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
wagner@peb.ufrj.br

ABSTRACT

Ultrasound phantoms require materials with acoustic properties close to those of the target tissues. This study characterized the speed of sound (SoS) in six plastisol specimens, three Soft and three Ultra. Twenty readings per specimen were obtained using through-transmission at 1 MHz. The analysis included descriptive statistics, Type A uncertainty evaluation, nested ANOVA, and Welch's t-test on specimen averages. The average SoS was 1602.34 m/s for Soft and 1473.05 m/s for Ultra. ANOVA detected no difference between formulations ($p = 0.142$), but provided evidence of variability among specimens within formulations ($p < 0.001$). The average SoS values for Soft and Ultra were numerically close to reference values for non-adipose and adipose tissues, respectively. Between-specimen variability and the small number of specimens limit generalizability and comprehensive uncertainty evaluation. The findings provide an exploratory characterization and inform the standardization of fabrication and measurement procedures for phantom applications.

Keywords - plastisol, ultrasound phantom, speed of sound, measurement uncertainty.

I. INTRODUCTION

Ultrasound phantoms are physical models used in the development, calibration, and evaluation of ultrasound systems, enabling controlled testing without exposing patients. For these purposes, a tissue-mimicking material must have acoustic properties consistent with those of the target tissue and sufficient stability to yield reproducible results [1]–[4].

The speed of sound affects time of flight (ToF) and the spatial localization of reconstructed interfaces. Its value depends on the properties of the medium and the measurement conditions; experimental characterization therefore requires careful comparison with biological tissue reference values [5], [6].

Ultrasound imaging systems calculate depth using an assumed speed of sound. A discrepancy between the actual and assumed speeds can cause axial scaling errors and affect focusing. Matching the speed of sound in a phantom to that of the represented tissue is therefore relevant to both geometric fidelity and equipment performance assessment [1], [3].

Plastisol, a dispersion of polyvinyl chloride particles in plasticizers, can be processed to achieve different degrees of flexibility and molded into various geometries. These features support its investigation as a material produced in Brazil for phantom fabrication [2], [4], [7]. Its use requires assessing both the dispersion of readings and the consistency of results across specimens.

This study aimed to characterize the SoS of Soft and Ultra formulations, estimate variability within and between specimens, and compare their average SoS values with biological tissue references.

II. THEORETICAL BACKGROUND

A. Speed of sound in tissues and tissue-mimicking materials

The speed of sound depends on the relationship between the elastic properties and density of the medium. In soft tissues, differences in water, lipid, protein, and collagen content give rise to characteristic ranges: lipid-rich tissues tend to have lower speeds, whereas tissues with higher protein or connective tissue content exhibit higher values [5], [6].

SoS depends jointly on density and the elastic properties of the medium. Flexibility alone does not determine longitudinal wave speed. Differences between formulations or specimens therefore cannot be attributed solely to density or flexibility without additional measurements [4]–[6].

Biological tissue references were used for descriptive comparison, accounting for differences in temperature, measurement method, and specimen preparation, without assuming that these references are free of uncertainty [5], [6].

B. Plastisol for phantom fabrication

Plasticized PVC materials combine flexibility with ease of molding. Formulation, homogenization, heating, bubbles, and cooling can affect their properties [4], [7]. Comparisons across specimens must also consider potential contributions from the measurement procedure.

Numerical proximity in SoS is an initial consideration when evaluating a tissue-mimicking material. Attenuation, acoustic impedance, backscattering, homogeneity, and temporal stability also influence phantom response; SoS alone does not establish acoustic equivalence with a tissue [3], [4].

III. MATERIALS AND METHODS

A. Study design and specimens

Three Soft specimens (S1–S3) and three Ultra specimens (U1–U3) were evaluated, with 20 thickness readings and 20 SoS readings per specimen. The specimen was the unit of comparison between formulations, and readings were treated as observations within each specimen. Inference between formulations assumes independent preparations. A fabrication batch effect was not estimated separately.

B. Fabrication

Approximately 100 g portions of plastisol were weighed using an Adventurer ARA520 balance (Ohaus, Brazil), homogenized for 30 min, and placed in a desiccator for 30 min. Microwave (Philco, model PM023E/127V) heating was performed at the 50% power setting, with an initial 90 s cycle followed by a 40 s cycle and, when necessary, a 15 s cycle. Heating was stopped near 150 °C to avoid signs of degradation; the material was then poured into molds and cooled.

Heating in successive cycles allowed the temperature to be monitored. For Soft, the additional 15 s cycle was used when necessary; for Ultra, the temperature limit was reached after the 40 s cycle. The formulations were processed separately, following the sequence of weighing, homogenization, bubble removal, heating, and cooling. Figure 1 identifies the six specimens evaluated.

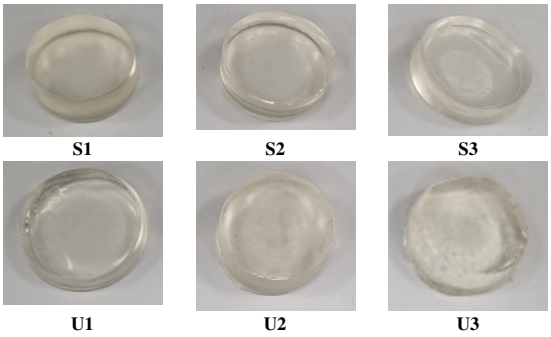


Figure 1 - Soft (S1–S3) and Ultra (U1–U3) plastisol specimens.

Source: authors.

C. Thickness and speed of sound measurements

Thickness was measured using a 500-172-30B digital caliper (Mitutoyo, Brazil). Acoustic measurements were performed at the Ultrasound Laboratory of COPPE/UFRJ in a glass tank containing distilled water, with temperature monitored near 24 °C. The setup comprised an AFG3021 signal generator (Tektronix), a 200 MHz WaveSurfer 3024 oscilloscope (Teledyne LeCroy), two 1 MHz Panametrics/Olympus transducers with a 0.70–1.30 MHz frequency band, and an MTH-1362W thermometer (Minipa).

Thickness readings were taken to represent dimensional variations within each specimen. During acoustic acquisitions, the transducers were aligned on opposite faces and kept in contact with the material. Bath temperature was monitored throughout the experimental sequence. The effect of contact pressure on thickness was not quantified.

Excitation consisted of a five-cycle sinusoidal burst at 1 MHz and 10 Vpp. The received signal was recorded on the oscilloscope and processed in LabVIEW with the standard substitution method [2]. The designation v_{PD} refers to SoS obtained along the direct path between the transducers; PD derives from the Portuguese term *percurso direto* (direct path). The internal equations and documentation of the ToF estimation algorithm, delay compensation, and use of thickness readings in the calculation are unavailable. No specific arrival-time detection method is therefore attributed to the routine. The analysis used the available SoS outputs, without reprocessing the signals. Figure 2 shows the through-transmission setup.

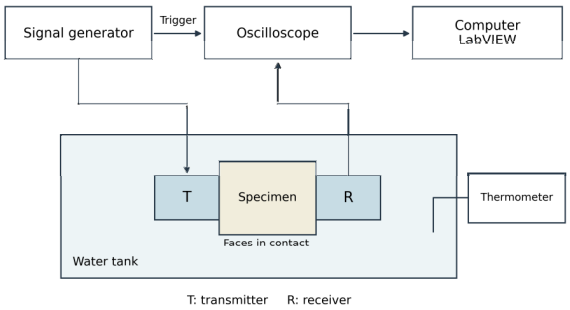


Figure 2 - Through-transmission setup. Adapted from Maggi [2].

D. Statistical analysis and uncertainty evaluation

For each specimen, the average, standard deviation (SD), coefficient of variation (CV), minimum, and maximum were calculated, and the Shapiro-Wilk test was applied to the readings. The Type A standard uncertainty of the average and the corresponding expanded uncertainty were estimated as [8]:

$$u_A = \frac{s}{\sqrt{n}};\ U_{95,A} = t_{0.975;\,n-1} \times u_A.$$

Here, s is the SD and n is the number of observations. For 20 readings, ν = 19 and k = 2.093. For the average across three specimens, the SD of specimen averages was divided by √3, with ν = 2 and k = 4.303. The 95% confidence intervals (95% CIs) rely on the t-distribution and independent observations at the relevant level; their coverage is approximate when these assumptions are not met.

The estimates u_A and $U_{95,A}$ describe the statistical contribution at the level analyzed. They do not constitute a complete uncertainty budget: calibration and resolution, thickness and its potentially shared contribution across readings, temperature, coupling, and internal processing were not fully evaluated. The lack of documentation for the routine prevents reconstruction of input uncertainty propagation. The uncertainty of the average across specimens also does not describe the dispersion expected for a future specimen.

Nested ANOVA was applied with material as a fixed effect and specimen nested within material as a random effect. Each variance term was calculated as the sum of squares divided by its degrees of freedom (SS/df). The material test used the specimen(material) term as its denominator; the between-specimen test used the within-specimen term. Welch's t-test on the three specimen averages per formulation was used as a sensitivity analysis for unequal variances. The results are exploratory because of the small number of specimens and limited assessment of assumptions; specimen-level Shapiro–Wilk tests do not assess normality of the between-specimen effects. Variance components were estimated from the ANOVA variance terms separately for each material. The intraclass correlation coefficient (ICC) was calculated as the ratio of between-specimen variance to total estimated variance. The significance level was α = 0.05.

Comparison with the values compiled by Mast [5] used difference magnitudes: $D = |\overline{c} - c_{ref}|$ and $D_{rel} = 100 \times \frac{D}{c_{ref}}$. No equivalence testing or statistical inference against tissue values was performed because the references are point values. Mass, volume, and density were not measured systematically before and after fabrication; their contributions to SoS could not be isolated.

IV. RESULTS

A. Descriptive statistics and within-specimen dispersion

Specimen averages ranged from 1410.49 to 1667.03 m/s, and CVs ranged from 1.28% to 3.99%. S2 had the greatest dispersion (SD = 63.19 m/s; CV = 3.99%), which is relevant to applications requiring precise SoS values. CVs were below 2.3% for the other specimens. The Shapiro–Wilk test indicated a departure from normality for S2 (W = 0.881; p = 0.018); failure to reject normality for the other specimens does not establish a normal distribution. Table I summarizes the 20 SoS readings for each specimen.

Table I - Descriptive statistics of SoS by specimen

ID	CV (%)	Average ± SD (m/s)	Min.–max. (m/s)
S1	1.28	1667.03 ± 21.28	1629.42–1698.49
S2	3.99	1582.55 ± 63.19	1467.19–1667.03
S3	1.33	1557.44 ± 20.69	1515.03–1593.57
U1	1.64	1598.00 ± 26.13	1551.49–1656.52
U2	1.91	1410.49 ± 26.87	1369.43–1471.69
U3	2.28	1410.64 ± 32.11	1349.26–1459.91

Note: 20 readings per specimen. S = Soft; U = Ultra; SD = standard deviation; CV = coefficient of variation.

Table II presents the Type A uncertainty evaluation for each specimen average and the corresponding 95% CI.

Table II - Type A uncertainty evaluation by specimen

ID	u_A	$U_{95,A}$	95% CI of average
S1	4.76	9.96	1657.07–1676.99
S2	14.13	29.57	1552.98–1612.12
S3	4.63	9.69	1547.75–1567.12
U1	5.84	12.23	1585.77–1610.23
U2	6.01	12.58	1397.92–1423.07
U3	7.18	15.03	1395.62–1425.67

Note: values in m/s; u_A = Type A standard uncertainty; $U_{95,A} = k \cdot u_A$, with k = 2.093 and v = 19. CIs are based on t; coverage is approximate for S2.

B. Comparison between materials and variance components

The average SoS of the Soft specimens was 1602.34 m/s, compared with 1473.05 m/s for the Ultra specimens. The Soft − Ultra difference was 129.30 m/s, equivalent to 8.78% of the average SoS of the Ultra specimens. Dispersion among specimen averages was greater for Ultra.

At the specimen level, $U_{95,A}$ ranged from 9.69 to 29.57 m/s for Soft and from 12.23 to 15.03 m/s for Ultra. For the average across three specimens, it was 142.63 m/s and 268.82 m/s, respectively. The increase reflects between-specimen variability and the larger t-factor associated with only two degrees of freedom. Table III summarizes the averages for each formulation, using specimens as the units of analysis.

Table III - Average SoS and dispersion of specimen averages

Material	Average ± SD	u_A	95% CI of average
Soft	1602.34 ± 57.42	33.15	1459.71–1744.97
Ultra	1473.05 ± 108.21	62.48	1204.23–1741.86

Note: three specimens per material. SD = standard deviation of specimen averages; u_A = SD/√3. Values in m/s; 95% CIs use k = 4.303 and v = 2.

ANOVA detected no material effect [F(1,4) = 3.342; p = 0.142], but provided evidence of variability among specimens within formulations [F(4,114) = 123.175; p < 0.001]. Welch's test yielded a consistent result [t(3.04) = 1.828; p = 0.164], with a 95% CI for the Soft − Ultra difference from −93.98 to 352.57 m/s. Inclusion of zero does not establish equivalence between formulations. Table IV presents the nested ANOVA results.

Table IV - Nested ANOVA of SoS

Source	df	SS	SS/df	F; p
Material	1	501522	501522	3.342; 0.142
Specimen (material)	4	600261	150065	123.175; < 0.001
Within specimen	114	138888	1218	—

Note: df = degrees of freedom; SS = sum of squares; SS and SS/df are in (m/s)². F and p are dimensionless.

The between-specimen component accounted for 66.4% of the estimated variance in Soft and 93.5% in Ultra. These percentages locate variability within the experimental design but do not distinguish fabrication effects from measurement contributions that persist within each specimen.

Table V presents the variance components for each material as SDs and their relative contributions. ICC is dimensionless and does not represent measurement accuracy.

Table V - Variability components and ICC

Component	Soft	Ultra
Between-specimen SD (m/s)	56.70	108.03
Within-specimen SD (m/s)	40.31	28.50
Between-specimen variance (%)	66.4	93.5
Within-specimen variance (%)	33.6	6.5
ICC (dimensionless)	0.664	0.935

Note: component SDs are in m/s; percentages refer to total estimated variance. The between-specimen SD is derived after subtracting the within-specimen contribution from the variance of specimen averages and therefore differs from the SD in Table III.

C. Comparison with biological tissues

For Soft, the smallest absolute relative differences were found for liver (0.46%), connective tissue (0.66%), skin (0.78%), and skeletal muscle (1.41%). For Ultra, they were found for fat (0.33%) and adipose tissue (1.59%).

Table VI compares formulation averages with human tissue references using absolute differences and absolute relative differences.

Table VI - Absolute differences from tissue reference values

Tissue	c_{ref}	Soft D; D_{rel}	Ultra D; D_{rel}
Adipose tissue	1450	152.34; 10.51	23.05; 1.59
Fat	1478	124.34; 8.41	4.95; 0.33
Breast	1510	92.34; 6.12	36.95; 2.45
Skeletal muscle	1580	22.34; 1.41	106.95; 6.77
Liver	1595	7.34; 0.46	121.95; 7.65
Connective tissue	1613	10.66; 0.66	139.95; 8.68
Skin	1615	12.66; 0.78	141.95; 8.79

Note: c_{ref} is in m/s; D = absolute difference in m/s; D_{rel} = 100D/c_{ref}, in %. Values from Mast [5]. "Fat" and "adipose tissue" retain the fat and adipose entries drawn from different sources in that compilation.

V. DISCUSSION

The average SoS values for Soft were numerically close to reference values for non-adipose tissues, whereas those for Ultra were close to adipose tissue references. This descriptive comparison guides further investigation of the materials but does not establish equivalence or demonstrate that flexibility determines SoS.

A formulation average does not guarantee the suitability of each specimen. For Ultra, U1 had an average SoS of 1598.00 m/s, whereas U2 and U3 had averages of 1410.49 and 1410.64 m/s. The proximity of the group average to the fat reference must therefore be considered alongside between-specimen dispersion. Furthermore, the references compiled by Mast [5] include tissue speeds measured at 37 °C or *in vivo*, whereas plastisol was measured near 24 °C; no temperature correction was applied.

The numerical difference between formulations was not statistically resolved, given the between-specimen variability and small number of specimens. The analysis accounted for readings nested within specimens, but generalizability depends on independent preparations and the model assumptions. A larger between-specimen variance component does not establish that fabrication caused the observed differences.

The greater dispersion in S2 is consistent with a possible contribution from limited repeatability of the procedure, including coupling, alignment, and ToF estimation. This explanation remains a hypothesis because the data do not distinguish these contributions from spatial variations in the material. Dispersion was therefore described quantitatively, without assigning a qualitative magnitude category. S2 was retained in the analysis, with its SD and Type A contribution explicitly reported.

Repeated readings quantify observed dispersion but do not reveal potential shared biases. The results characterize the outputs obtained under the study conditions; they do not constitute independent metrological validation of the routine or estimates supported by a complete uncertainty budget.

The absence of systematic mass, volume, and density measurements limits interpretation of effects. Future studies should include independently prepared batches, multiple specimens per batch, mass, volume, and density measurements at defined fabrication stages, local thickness measurements, and repeated acquisitions at defined positions. The ToF routine should also be documented and verified against an appropriate reference.

Such a design would allow the effects of formulation, batch, and measurement procedure to be distinguished more precisely. Characterization of other acoustic properties and temporal stability would complement the assessment of the materials for phantom applications.

VI. CONCLUSION

The Soft and Ultra specimens had average SoS values of 1602.34 and 1473.05 m/s, respectively, with no statistically detected difference between formulations. The between-specimen variance component predominated. Comparison with tissue references identified numerical proximities useful for initial material selection, without establishing equivalence or the suitability of each specimen. S2 dispersion was incorporated into the analysis; limited repeatability is a possible explanation, not a demonstrated cause. The findings inform the standardization of fabrication and measurement procedures and future validation of the materials and measurement procedure.

REFERENCES

[1] P. Fish, Physics and Instrumentation of Diagnostic Medical Ultrasound. Chichester, U.K.: Wiley, 1990.

[2] L. E. Maggi, "Characterization of acoustic properties of new materials Intended to manufacture ultrassonic phantoms" (in Portuguese), Ph.D. dissertation, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2011.

[3] J. E. Browne, K. V. Ramnarine, A. J. Watson and P. R. Hoskins, "Assessment of the acoustic properties of common tissue-mimicking test phantoms," Ultrasound Med. Biol., vol. 29, no. 7, pp. 1053–1060, 2003, doi: 10.1016/S0301-5629(03)00053-X.

[4] M. O. Culjat, D. Goldenberg, P. Tewari and R. S. Singh, "A review of tissue substitutes for ultrasound imaging," Ultrasound Med. Biol., vol. 36, no. 6, pp. 861–873, 2010, doi: 10.1016/j.ultrasmedbio.2010.02.012.

[5] T. D. Mast, "Empirical relationships between acoustic parameters in human soft tissues," Acoust. Res. Lett. Online, vol. 1, no. 2, pp. 37–42, 2000, doi: 10.1121/1.1336896.

[6] S. A. Goss, R. L. Johnston and F. Dunn, "Comprehensive compilation of empirical ultrasonic properties of mammalian tissues," J. Acoust. Soc. Am., vol. 64, no. 2, pp. 423–457, 1978, doi: 10.1121/1.382016.

[7] R. L. Q. Basto, "Construction of phantom material to mimic thermophysical and Acoustical properties of human biological tissues At 1.0 and 2.8 MHz," (in Portuguese), M.Sc. thesis, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2007.

[8] JCGM, Evaluation of measurement data — Guide to the expression of uncertainty in measurement, JCGM 100:2008, 2008.

A 3D-printed micropipette for open lab automation

Dariel Yared Lopez
Professional Interdisciplinary Unit of biotechnology.
National Polytechnic Institute
Mexico City, Mexico.
ORCID 0009-0004-3970-6892
dlopezm1808@alumno.ipn.mx

Alexis Chino Ulloa
Laboratory of Immunopharmacology
National Polytechnic Institute
Mexico City, Mexico.
ORCID 0000-0001-9541-5310

Itzel Pamela Tórres Ávila
Laboratory of Immunopharmacology
National Polytechnic Institute
Mexico City, Mexico.
ORCID 0009-0007-2125-9184

***Abstract*—**
Lab automation is increasing, improving efficiency, traceability and remote experimentation, however it is unevenly accessible. Small and medium labs support themselves on open-source 3D printed equipment for reliability and customization.
3D printing for liquid handling automation has a crucial challenge, removing leakage to get an electronically controlled micropipette to work.
This work proposes a model apt for the automated manipulation of cell cultures through 3D printing assisted by uncomplicated heat treatments.
To read volumes in microliters, digital imaging is used and the results show a reliable new open-source tool for micrometric volume control in labs.

Keywords*— *Automation, 3D printing, Biomaterials, Micropipette, Liquid Handler.

I. Introduction

Lab automation is the next step for improving efficiency and traceability, mainly used in North America, Europe and Asia while efforts in the rest of the world represent only 7% of the industry [1] Fig. 1.

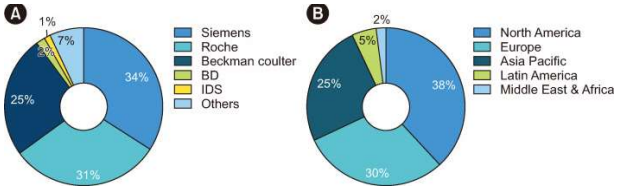


Fig. 1. Cake graphs for world distribution of the lab automation industry.

This disparity is more noticeable in public institutions and small enterprises, where technological adoption has the most resistance, even though these sectors are a potential customer base [1].

DIY (do it yourself) manufacturing allows smaller research teams to adopt and customize current technologies while still allowing replicability and automation due to the considerable number of labs adopting a similar methodology [2].

3D (three dimensional) printing is a promising technology that could allow us to create lab equipment to be shared globally [3].

During this work, 3D printers also served as a model of study for the adoption of an emerging technology fostering scientific work.

Liquid handling automation for lab use began developing at the start of the century and its usage in research doubled following the 2019 pandemic [4].

However, there is a gap in current research regarding standardization and flexible 3D printable designs, as the current devices use proprietary technology or require manual swapping of handheld universal micropipettes as will be shown in the current designs section.

This work intends to serve as the base for an automated 3D cell culture system; however, it is of widespread use for precise liquid management in any science project.

The result is the creation of an electronically controlled printable 3D micropipette using disposable syringes and stepper motors in contrast with current models to improve accessibility and maintain repeatability.

In this work we will:

1. Study current proposals for lab automation regarding pipetting (Sec. II).
2. Determine the optimal polymers and manufacturing processes (Sec. III).
3. Formulate a mathematical model for the micropipette (Sec. IV).
4. Create an electronically controlled micropipette prototype (Sec. V).
5. Design a replicable test to evaluate the micropipette's performance (Sec. VI).
6. Evaluate the pipette's performance for its adoption in lab environment and future uses (Sec. VII)

II. Current Designs

Every author mentioned in this section successfully solved a problem they encountered and designed ingenuous models for their specific use.

The goal of studying them in this work is to focus on their strengths to create a general micropipette that could spread lab automation adoption.

Current solutions can be divided into two groups, the ones that utilize manual micropipettes [5], [6] and the ones that implement proprietary volume control tools [7]-[11].

Using manual micropipettes presents a challenge to automation, because human intervention is needed to adjust precise volumes. Whilst proprietary volume control tools complicate widespread adoption by decreasing availability and repairability [12].

A. Manual pipette automation

The idea behind these models is to create a platform capable of moving a micropipette to pre-established coordinates triggering the pipette to suction or deposit a manually fixed volume, making repetitive tasks easier [5].

Another popular approach is adapting more familiar and available technologies, such as 3D printing and manual pipettes to automate liquid handling tasks [6]. Fig 2.

These modes enable replicability and take advantage of preexisting standardization, fostering open-source usage [12].

CBEB ©2026 IEEE

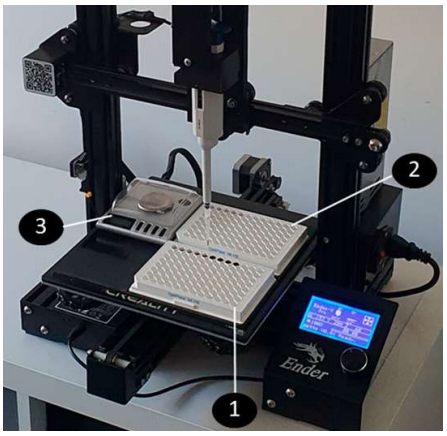


Fig. 2. Manual pipette mounted to a 3D printer [6]

B. *Proprietary volume control*

Privately developed technologies are more suitable for commercialization, but they lack standardization and flexibility due to corporate confidentiality, challenging repeatability for experimentation, in exchange, these models tend to offer tailored solutions for their consumers [7].

For example, the one-off automated cell culture system from Kyoto University [8] designed for automatic proliferation of stem cells, this system uses a robotic arm and different modules to replace human intervention.

Opentrons's liquid handling robot OT-2 is one of the most successful commercial models; it has an open software API and platform to share protocols and even semi-open hardware [7]. However, their pipetting technology is proprietary [9]. As the only manufacturers it is not viable to replicate their hardware and adopt their open software. Fig 3.



Fig. 3. Opentrons OT-2 liquid handler platform. [7]

Proprietary technologies can also be integrated into open-source 3D printable models, which improves repeatability if the proprietary hardware is sold in the region.

This compromise benefits researchers for highly specialized uses, the PHIL [10] or Sidekick [11] robots.

Phil was created to fit in a Nikon Ti-Eclipse microscope, and uses 3D printed peristaltic pumps, which are reliable at dispensing but not as much for suction [10].

Sidekick uses solenoid driven displacement pumps which are highly reliable although slightly less available worldwide [11].

C. *3D printing micropipettes*

Creating open tools for lab use goes beyond simply reducing costs, oftentimes it favors customization, streamlines the experimentation and provides the ability to share models globally for feedback [12].

Regarding 3D-printed micropipettes, the existing models deal with the mobile parts by using a syringe, this streamlines the printing process by preventing possible leaks while preserving repeatability and availability on a worldwide scale [13] Fig. 4.



Fig. 4. 3D printed micropipettes using disposable syringes [13]

After analyzing the current proposals, we can identify a clear need for an open-source standardized volume control system that can improve automation and worldwide replicability.

In addition, our proposal can be retrofitted to the systems that depend on manual micropipettes to simplify its adoption.

We can look at a brief comparison of the cases. Table 1.

TABLE I. CURRENT DESIGNS

Name	Automated volume control	Open-source design	Can use the automated pipette	Allows controlled suction
OTTO [5]	No	Yes	Yes	Yes
3D Printer system [6]	No	Yes	Yes	Yes
OT-2 [7]	Yes	Partial	No	Yes
Kyoto U [8]	Yes	No	Yes	Yes
PHIL [10]	Partial	Yes	No	No
Sidekick [11]	Partial	Yes	No	No

III. BIOMATERIALS AND 3D PRINTING

A. *Biocompatible polymers*

FDM (fused deposition modelling) 3D printers heat up a polymer filament and extrude it by moving on a flat plane to "draw" the layers that make up the volume of a digital object [14]

For life sciences, the use of polymers in general must consider biocompatibility, meaning the material must enable normal cell function, surface migration and proliferation, as well as be tolerated by the immunological cells [15].

One of the most complete studies on polymers used in life sciences was made before 3D printing popularized in the field but still includes melting temperatures due to the required sterilization processes [16].

A valid concern is that even if the polymers are biocompatible in the surface, the generation of microplastics can affect the cells, as it has been proven that their presence weakens the cellular membrane [17].

So, we will look for polymers suitable for FDM printing that do not degrade with disinfection. We will include UV (Ultra-Violet) light, chemical and heat and pressure conditions that ensure disinfection [16], [18], [19]. Table 2.

TABLE II. POSSIBLE POLYMERS

	PLA	ABS	TPU	PETG	Nylon UV/T	PC
Extrusion Temp. (°C)	130 – 180	200 – 230	~ 200	220 - 260	230 – 270	~300
Deforming Temp. (°C)	~ 60	~ 100	~ 70	80 – 90	80 – 120	>100
Ethanol (70%) resistance	No	Single event	Yes	Yes	Yes	Yes
Chlorine (0.5%) resistance	No	Single event	No	Yes	Yes	Single event
Formalde - hyde (10%) resistance	No	Single event	Single event	Yes	Yes	Yes
UV – B resistance	No	Single event	No	Single event	Yes	Yes
UV – C resistance	No	No	No	No	Single event	Yes

Where single event sterilization means it withstands the procedure once, making them single use.

The only materials that could resist standard sterilization due to temperature limits are PC (Polycarbonate) and certain Nylon UV tolerant varieties.

For disinfection, PETG (Polyethylene Terephthalate Glycol) resists chemical sterilization and can withstand cycles of UV-B radiation, however, it tends to deteriorate it overtime [17].

The problem with PC is it´s high fusion temperature, mid-range 3D printers are not designed to reach 300 °C.

The most promising filaments will be the PETG or UV tolerant Nylon, like PA6-CF [18].

We will use PETG-CF in a Creality Ender 3V3 printer due to availability, CF being carbon fiber, which increases slightly the melting temperature [20], [21].

B. 3D printing methodology

There are slight differences between modern 3D printers and their filaments, so a CAD model needs to be "sliced" (i.e., instructions to create the volume layer by layer [18]) selecting the correct model of 3D printer inside the software. Table 3.

Z-wobble (f) and layer shifting (c) appear to be hardware related with no direct solution, but the rest can be adjusted [18]. Fig. 5.

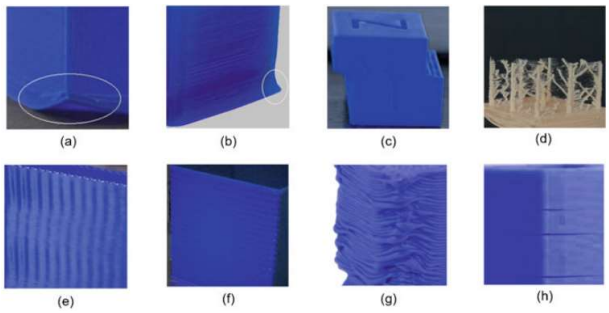


Fig. 5. Common 3D printing mistakes [18]

TABLE III. COMMON 3D PRINTING PROBLEMS [18]

Problem	Cause	Solution
Warping (a)	Thermic contraction detaches the piece from the base.	Heat up the base and avoid cold air around the piece.
Elephant's foot (b)	Melting in the base layers.	Cool down the base.
Stringing (d)	Plastic residue sticking to the tip.	Reduce tip temperature and extrusion rate.
Ringing (e)	Mechanical vibrations or air breezes affect the printing.	Fastening and isolating the 3D printer.
Curling (g)	Heat from the tip deforms lower layers.	Reduce tip temperature or cool down the piece.
Layer separation (h)	The extruded polymer isn't hot enough for another layer to stick.	Heat up the tip and avoid cold air around the piece.

Aside from this common 3D printing problems, there is also a prevalent challenge in avoiding leakage when working with fluids [11], [18], [22].

We analyzed three methodologies to avoid leaks to determine the optimal printing conditions for PETG-CF [20], [21], [23]:

- 0.16 mm layer height, as thin layers fuse easier.
- 90% Infill and 110% Extrusion rate, to overflow the borders and ensure minimal air bubbles.
- 4 Outer and Inner walls, pairing parallel walls reduces heat deformation and air bubbles.
- 50 mm/s printing speed, this is a moderate speed for most printers, which avoids unexpected vibrations.

C. After-printing treatment

Thermal treatments are common for manufacturing, however with FDM there is a possibility of deforming the polymer, so a new method has been developed, salt mold annealing [24], [25 (Unpublished)].

This methodology is used to improve the mechanical properties by reducing internal imperfections, but as a side effect it also allows us to fuse any possible leakage. Fig. 6

These essays use spectroscopy to show there is no chemical reaction happening, it is only PETG getting to its glass transition state and internally rearranging.

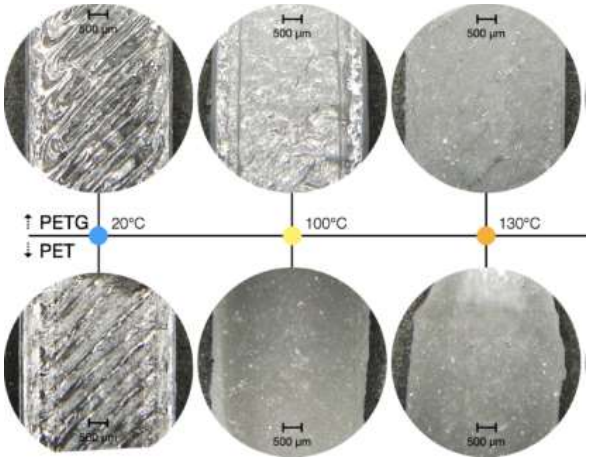


Fig. 6. Microscopic imaging of annealed polymers [25]

To avoid deformations from heat variation we will employ an 80 °C salt annealing process for 60 minutes, fount effective even if not ideal in both studies.

IV. Numerical Modeling

A. *Defining the function*

We will take Brennan et al.'s approach and use disposable syringes to control the volume [13], however, to avoid cross contamination and improve control we will use a flexible membrane [26]

This is the proposed model, it accepts both 200 µL and 10 µL pipette tips and connects to the syringe using medical grade disposable silicon tubes which have a fixed volume.

In this model the pipette has two chambers, the first one created by the mobile syringe, the silicon tube and the upper half of the pipette. Fig. 7.

The second one was created by the lower half of the pipette, the pipette tip holder and the universal pipette tips.

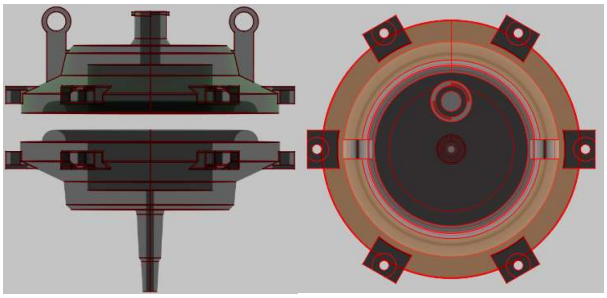


Fig. 7. Proposed pipette working for dispensing and suction.

Both halves of the pipette are separated by the flexible membrane, which is made from latex condoms after considering their global availability and standardization [27].

This membrane is characterized before [28] and after [29] vulcanization, as well as before packaging, in all cases the surface area presents homogenous elastic resistance up to 200% strain. This usage does not surpass this limit, so the surface behaves like a 2D Hooke's law spring.

Considering a 3mL syringe beginning at 1.5 mL, a 1 mm internal radius and our model's dimensions, we get a volume of 6.96 mL for the upper half (V1) and a volume of 5.35 mL for the lower half (V2). Fig. 8.

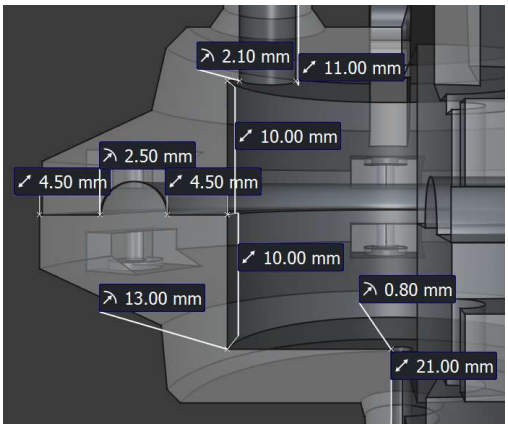


Fig. 8. Measurements from the proposed model

The small curve between the pipette parts ensures the gripping of the flexible membrane, it is held airtight by six M2 screws like a watertight watch.

Under the system's conditions we can consider the air to behave like an ideal gas [27] in two chambers Fig. 9.

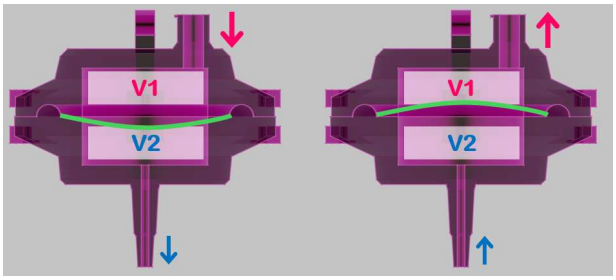


Fig. 9. 2D representation of the model's function, with the flexible membrane in green and air interacting trough it (V1 and V2)

$$PV=nRT \quad (1)$$

Considering a closed isothermal system with initial pressure, we can use Boyle's law (2 and 3).

$$P_o * V_0 = P_1 * V_1 \quad (2)$$

$$V_1 = (P_o * V_0) / \Delta P \quad (3)$$

The pressure varies as the volume compresses or expands because of the membrane's resistance (3).

Total pressure (P_T) will be the sum of external pressure (P_e) and the membrane's resistance pressure (P_m) leading to a change (4).

$$P_T = P_e + P_m \quad (4)$$

Unlike Beroz and Hart, our model's membrane has a lineal behavior for our range of displacements [29], [30].

$$P_m = K * \Delta V_{(P)} \quad (5)$$

We consider the displaced volume ($V_{(P)}$) directly as x because the scale in disposable syringes indicates volume from displacement (regular pistons use area * displacement).

$$\Delta V_{(P)} = x \quad (6)$$

Now that we have the main variable, we can substitute our defined values into Boyle's volume equation (3):

$$V(x) = (P_o * V_0) / (P_e + K* x) \quad (7)$$

Our initial (P_o) and external (P_e) pressures are both the atmospheric pressure at our altitude, the initial volume will be 6.96 mL based on our model, and K varies slightly with each disposable membrane used. Fig. 10.

$$V(x) = (P_{atm} * 6.96\ mL) / (P_{atm} + K* x) \quad (8)$$

Finally, to get the volume change (ΔV) we must consider the initial volume (7). Fig. 11.

$$\Delta V = V_0 - V(x) \quad (9)$$

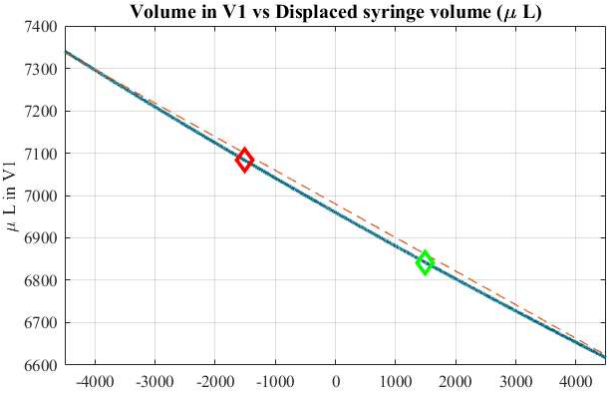


Fig. 10. Graph of V(x) according to equation 8 made, the red and green diamonds represent the physical minimum and maximum values for the syringe, and the orange dotted line exemplifies lineal behavior.

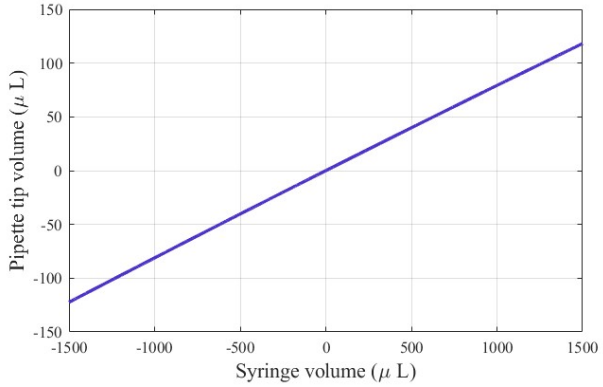


Fig. 11. This graph shows the final volume change according to equation 9, using the local barometric pressure (77 kPa) and a membrane stiffness ratio of 0.9 Pa / μL.

V. Printing the Pipette

The entire model uses a 3D printed platform to hold the syringe and the stepper motor in place.

During this stage of development, we use open loop control, the linear actuator uses a captive screw so the amount of steps can be accounted for reliably.

The 3D model for the pipette was designed independently from the motor platform to allow users to customize the control mechanism for their own needs, but our control model is also available if needed. Fig. 12.

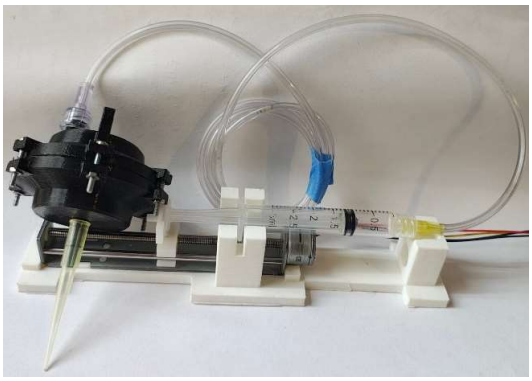


Fig. 12. 3D printable pipette and control system

The heat treatment done to the 3D pipette to avoid leakage differed from studied methodologies due to the uneven salt crystal sizes. Fig. 13.

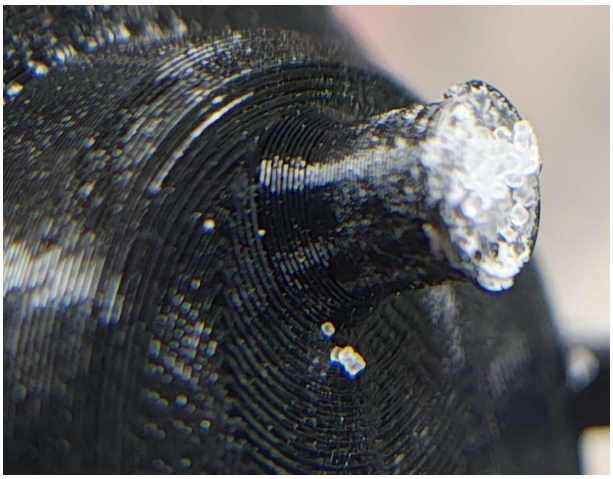


Fig. 13. Salt crystals and pulverized salt in the tip of the 3D printed pipette before salt annealing

To counteract this problem and ensure the least deformation in crucial spots pulverized salt was added where the pipette connects to the silicon tube.

Also, accounting for deformation the inside of the pipette was filled with salt and assembled with aluminum foil instead of the flexible membrane. Fig. 14.



Fig. 14. Aluminum foil placed into the pipette before salt annealing.

The resulting piece still shows the original filament positions, however, due to the heat treatment, layers joined together resulting in shallower, blurrier lines on the surface. Fig.15.

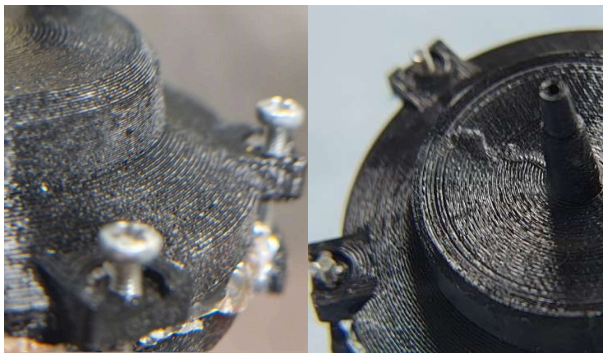


Fig. 15. 3D printable pipette post- salt annealing

VI. Testing the Pipette

To evaluate the pipette´s volume control accurately we resort to measuring the volume of each drop of water using digital imaging [30] using MATLAB. Fig. 16.

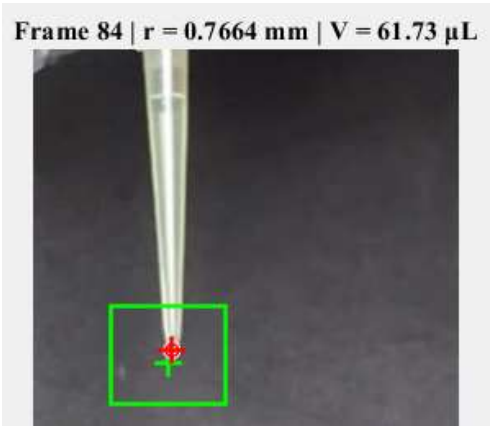


Fig. 16. MATLAB imaging program used to determine drop volume.

The program is simple, first it detects the tip and then it looks for a circle in that direction, we used the tip to determine the conversion from pixels to millimeters.

Footage will be captured from a smartphone camera, so, to avoid image deformation, we will keep a fixed 12 cm distance from the lens to the pipette's tip.

The phone's camera has a f/2.0 aperture and 3x optical zoom recording at 60 frames per second.

The volume will be controlled by the TB6600 stepper driver connected to an ESP32.

The controlled volume is determined by the amount of steps the stepper motor gives.

This number can vary depending on the dimensions of the disposable syringe, so it will be transformed into volume lineally by taking the maximum and minimum values for a 3mL disposable syringe Fig 17.

For reference, 50% of the syringe will be 0 so we have both positive and negative displacement like in our mathematical model.

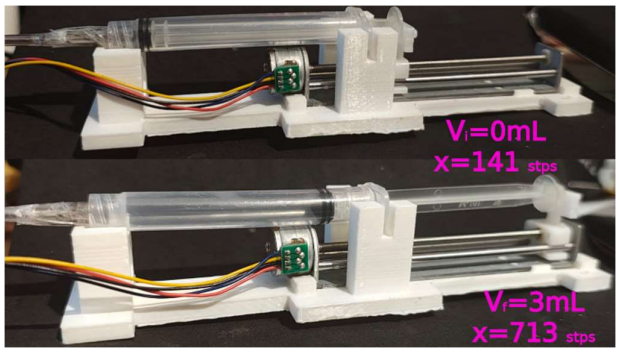


Fig. 17. Syringe max and min values.

To validate if the mathematical model is correct, we need to demonstrate that the membrane performs in a completely elastic deformation.

For this test the stepper motor will advance at a fixed speed by delaying each step 1 millisecond, thus we can analyze the syringes retraction (suction) and compression (expulsion).

After the first test (5 cycles of suction and expulsion). We will wait eight hours to repeat the same test to analyze the membrane's deformation and see if the elastic model still corresponds to the mathematical model.

VII. Results

From this experimentation model and advancing the linear actuator at a constant speed we can obtain the volume control graph in MATLAB for the droplets Fig. 18.

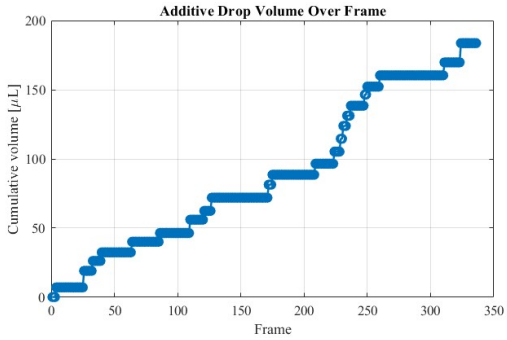


Fig. 18. Graph of total volume overtime as the syringe empties.

If we compare the theoretical graph [Fig. 15] to our obtained graph the only changing factor would be the membrane's stiffness, however behavior was not completely lineal.

As more tests were done it became clear that the membrane loosens overtime, needing compensation or diminishing the final range Fig. 19.

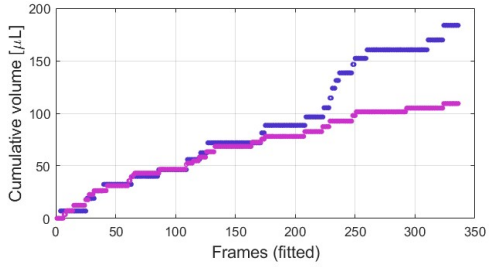


Fig. 19. Graph of total volume overtime as the syringe empties for test 1 and test 2 using the same membrane.

Finally, we can compare both tests to the mathematical model and analyze the R value to see if they correlate Fig. .

According to the R value from test 2, the membrane is no longer accurately represented by the mathematical model.

However, with a fresh membrane we get an R value above 0.95, making the mathematical model reliable for this case. Fig. 20.

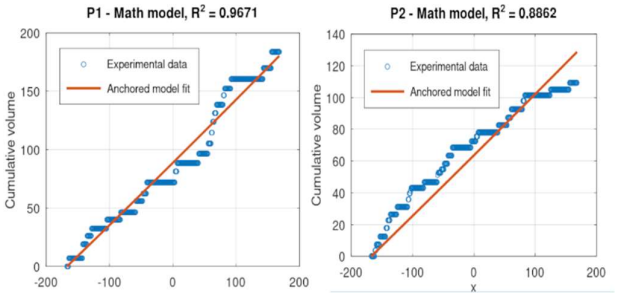


Fig. 20. Graphs of empirical values vs mathematical models with R squared values.

In conclusion, this 3D printable model can reliably control micrometric volumes only under ideal conditions (Thermically treated, with brand new membranes and with no degradation from repeated sterilization).

However, it does offer a base in which open-source development can be fostered by replicability and ease of access.

The design deals with leakage, one FDM's strongest challenge, by implementing manufacturing technologies from different fields like drone aeronautics and vastly available reliable and sterile materials like disposable syringes and condoms for its use in the biomedical sector.

To encourage replicability, we make available the 3D printable models and the MATLAB code.

Regarding further work we look forward to researchers wanting to experiment with this pipette into their own liquid handlers, as well as more sophisticated, closed loop control approaches using it to get more accurate volume conversions.

In our specific case, this pipette will be applied for the automation of hanging-drop 3D cell cultures, and we are looking for a programmer to create an open-source API to control it remotely.

Another approach to the API would be to integrate it into preexisting open software liquid handlers like OTTO [8] if we get to collaborate in their efforts.

For those who want to improve this proposal in material engineering aspects we encourage the development of flexible membranes designed specifically for micro pipetting or other creative approaches. We used condoms for their worldwide availability, but other solutions might make the system more reliable.

REFERENCES

[1] Nam, Y., & Park, H. D. (2025). Revolutionizing Laboratory Practices: Pioneering Trends in Total Laboratory Automation. Annals of laboratory medicine, 45(5), 472–483. https://doi.org/10.3343/alm.2024.0581

[2] Wenzel T. (2023) Open hardware: From DIY trend to global transformation in access to laboratory equipment. PLOS Biology, January 2023 https://doi.org/10.1371/journal.pbio.3001931

[3] Sahoo, S., & Lo, C. Y. (2022). Smart manufacturing powered by recent technological advancements: A review. Journal of Manufacturing Systems, 64, 236-250. Doi: https://doi.org/10.1016/j.jmsy.2022.06.008

[4] Torres-Acosta, M. A., Lye, G. J., & Dikicioglu, D. (2022). Automated liquid-handling operations for robust, resilient, and efficient bio-based laboratory practices. Biochemical Engineering Journal, 188, 108713. Doi: 10.1016/j.bej.2022.108713.

[5] Florian, D.C., Odziomek, M., Ock, C.L. et al. Principles of computer-controlled linear motion applied to an open-source affordable liquid handler for automated micropipetting. Sci Rep 10, 13663 (2020). https://doi.org/10.1038/s41598-020-70465-5

[6] Kopyl, A., Yew, Y., Ong, J. W., Hiscox, T., Young, C., Muradoglu, M., & Ng, T. W. (2024). Automated liquid handler from a 3D printer. Journal of Chemical Education, 101(2), 640–646. https://doi.org/10.1021/acs.jchemed.3c00855

[7] Opentrons Labworks, Inc., Opentrons OT-2 Instruction Manual. Opentrons Documentation. [Online]. Available: https://docs.opentrons.com/ot-2/ (accessed April. 24, 2026).

[8] A. Sigler, T. S. Goodman, N. P. Starno, and J. P. Badal, "Systems and methods for pipette robots," U.S. Patent 11 506 677, Nov. 22, 2022.

[9] Konagaya S; Ando T; Yamauchi T; Suemori H. & Iwata H. (2015) Long-term maintenance of human induced pluripotent stem cells by automated cell culture system Nature Scientific Reports Art. 16647(2015). https://doi.org/10.1038/srep16647

[10] Dettinger, P., Kull, T., Arekatla, G. et al. Open-source personal pipetting robots with live-cell incubation and microscopy compatibility. Nat Commun 13, 2999 (2022). https://doi.org/10.1038/s41467-022-30643-7

[11] Keesey R, LeSuer R, Schrier J (2022) Sidekick: A Low-Cost Open-Source 3D-printed liquid dispensing robot HardwareX 2468-0672, Vol. 12, October, 2022. https://doi.org/10.1016/j.ohx.2022.e00319

[12] Doloi S., Das M., Li Y., Cho Z., Xiao X., Hanna J., Osvaldo M. & Ng-Wei L. (2025) Democratizing self-driving labs: advances in low-cost 3D printing for laboratory automation. Digital Discovery Issue 7, 2025. https://doi.org/10.1039/D4DD00411F

[13] Brennan, M. D., Bokhari, F. F., & Eddington, D. T. (2018). Open Design 3D-Printable Adjustable Micropipette that Meets the ISO Standard for Accuracy. Micromachines, 9(4), 191. https://doi.org/10.3390/mi9040191

[14] Jandyal, A., Chaturvedi, I., Wazir, I., Raina, A., & Ul Haq, M. I. (2022). 3D printing – A review of processes, materials and applications in industry 4.0. Sustainable Operations and Computers, 3, 33–42. https://doi.org/10.1016/j.susoc.2021.09.004

[15] Mukherjee, S., Malik, P., Mukherjee, T. (2023). Mammalian Cell Culture: An Overview. In: Mukherjee, T.K., Malik, P., Mukherjee, S. (eds) Practical Approach to Mammalian Cell and Organ Culture. Springer, Singapore. https://doi-org.bibliotecaipn.idm.oclc.org/10.1007/978-981-19-1731-8_1-1

[16] L. McKeen, The Effect of Sterilization on Plastics and Elastomers, 3rd ed. Waltham, MA, USA: William Andrew, an imprint of Elsevier, 2012. ISBN: 978-1-4557-2598-4.

[17] Fleury, J. B., & Baulin, V. A. (2021). Microplastics destabilize lipid membranes by mechanical stretching. Proceedings of the National Academy of Sciences, 118(31), e2104610118. https://www.pnas.org/doi/epdf/10.1073/pnas.2104610118

[18] H. K. Dave and J. P. Davim, Eds., Fused Deposition Modeling Based 3D Printing. Cham, Switzerland: Springer Cham, 2021. doi: 10.1007/978-3-030-68024-4. ISBN: 978-3-030-68024-4.

[19] Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). 3D printing applications for healthcare research and development. Global Health Journal, 6, 217–226. https://doi.org/10.1016/j.glohj.2022.11.001

[20] S. Bhandari, R. A. Lopez-Anido, and D. J. Gardner, "Enhancing the interlayer tensile strength of 3D printed short carbon fiber reinforced

PETG and PLA composites via annealing," Addit. Manuf., vol. 30, p. 100922, 2019, doi: 10.1016/j.addma.2019.100922.

[21] K. S. Kumar, R. Soundararajan, G. Shanthosh, P. Saravanakumar, and M. Ratteesh, "Augmenting effect of infill density and annealing on mechanical properties of PETG and CFPETG composites fabricated by FDM," Mater. Today Proc., vol. 45, pp. 2186–2191, 2021, doi: 10.1016/j.matpr.2020.10.078.

[22] Leeder R., Rankin K. & Nightingale A. (2025) Microscale Computed Tomography (μCT) Imaging of Leak Pathways for Optimized Leak-Free 3D Printed Fluidics ACS Applied Polymer Materials Vol. 7 Issue 21. 14130–14137 https://doi.org/10.1021/acsapm.5c02274

[23] S. Valvez, A. P. Silva, P. N. B. Reis, and F. Berto, "Annealing effect on mechanical properties of 3D printed composites," Procedia Struct. Integr., vol. 37, pp. 738–745, 2022, doi: 10.1016/j.prostr.2022.02.004.

[24] M. Z. Baltić et al., "PETG as an alternative material for the production of drone spare parts," Polymers, vol. 16, no. 21, p. 2976, 2024, doi: 10.3390/polym16212976.

[25] R. Spina, "Annealing of PET/PETG specimens made with material extrusion," Sci. Rep., 2026, doi: 10.1038/s41598-026-50018-y. (Unpublished)

[26] Beroz J. & Hart A. (2016) Universal handheld micropipette, Review of Scientific Instruments, Noviembre 2016. USA. DOI: 10.1063/1.4966989

[27] International Organization for Standardization, “Natural rubber latex male condoms—Requirements and test methods,” ISO 4074:2015, 2015.

[28] Anand, K., Varghese, S., & Kurian, T. (2015). Effect of micro and nano zinc oxide on the properties of pre-vulcanized natural rubber latex films. Progress in Rubber Plastics and Recycling Technology, 31(3), 145-156. https://doi.org/10.1177/147776061503100301

[29] Keereerak, A., Lehman, N., Uthaipan, N., Nakaramontri, Y., Johns, J., Promsung, R., & Kalkornsurapranee, E. (2024). Exploring the influence of Hevea brasiliensis clones on the extraordinary properties of natural rubber vulcanizates. Polymer Bulletin, 81(12), 10991-11005. https://doi.org/10.1007/s00289-024-05218-9

[30] Zieger, V., Woehr, E., Zimmermann, S., Frejek, D., Koltay, P., Zengerle, R., & Kartmann, S. (2024). Automated Nanodroplet Dispensing for Large-Scale Spheroid Generation via Hanging Drop and Parallelized Lossless Spheroid Harvesting. Micromachines, 15(2), 231. https://doi.org/10.3390/mi15020231.

An Active Wrist Orthosis for Objective Rigidity Assessment in Parkinson's Disease: Biomechanical Correlates and Usability Evaluation

Daniel Hilário da Silva*
Dept. of Computer Science
IF Goiano (Catalão)
Catalão, GO, Brazil
daniel.hilario@ifgoiano.edu.br
0000-0002-0800-065X

Adriano Alves Pereira
Postgrad. Prog. Biomedical Eng.
Federal University of Uberlândia
Uberlândia, MG, Brazil
adriano.pereira@ufu.br
0000-0002-1522-9989

Caio Tonus Ribeiro
Postgrad. Prog. Biomedical Eng.
Federal University of Uberlândia
Uberlândia, MG, Brazil
caiot@ufu.br
0000-0003-0085-317X

Caroline Valentini de Queiroz
Postgrad. Prog. Biomedical Eng.
Federal University of Uberlândia
Uberlândia, MG, Brazil
caroline.valentini@ufu.br
0000-0002-1968-230X

Leandro Rodrigues da Silva Souza
Dept. of Computer Science
IF Goiano (Rio Verde)
Rio Verde, GO, Brazil
leandro.souza@ifgoiano.edu.br
0000-0003-2477-6893

Selma Terezinha Milagre
Postgrad. Prog. Biomedical Eng.
Federal University of Uberlândia
Uberlândia, MG, Brazil
selma.milagre@ufu.br
0000-0002-0807-9839

Eduardo de Moura Neto
Postgrad. Prog. Biomedical Eng.
Federal University of Uberlândia
Uberlândia, MG, Brazil
eduardonetofisio@gmail.com
0000-0003-0055-0822

Júlia Tannús de Souza
Postgrad. Prog. Biomedical Eng.
Federal University of Uberlândia
Uberlândia, MG, Brazil
julia.tannus@ufu.br
0000-0003-1521-5680

Lacordaire Kemel Pimenta Cury
Dept. of Computer Science
IF Goiano (Catalão)
Catalão, GO, Brazil
lacordaire.cury@ifgoiano.edu.br
0000-0003-4035-2126

***Abstract*—Parkinson's disease (PD) is the second most common neurodegenerative disorder worldwide, and rigidity affects most diagnosed individuals. Although a cardinal symptom, rigidity assessment remains largely subjective, relying on the ordinal MDS-UPDRS scale and susceptible to inter-rater variability. This study presents a biomechanical and usability assessment of an active wrist orthosis designed to objectively quantify wrist rigidity. The device performed standardized passive flexion-extension movements on 30 participants (15 PD, 15 healthy controls) while recording current, force, and angular displacement. Twelve biomechanical features were correlated with MDS-UPDRS Item 3.3 rigidity scores using Spearman correlation. As an exploratory, interpretable complement, a multiple linear regression model with second-order interaction terms was fitted to relate these features to the rigidity score. Usability was evaluated with the System Usability Scale (SUS) and the Short Flow State Scale (SFSS). Flexion-phase features showed moderate, statistically significant correlations with Item 3.3, while signal-variability features showed moderate negative correlations. In this exploratory analysis, a parsimonious three-interaction-term model reached an adjusted coefficient of determination of about 0.57 and an in-sample expanded accuracy near 73 percent, within one point of the clinical score. The overall mean SUS score of about 85, classified as excellent, confirmed strong device acceptance in both groups. These findings support the use of the orthosis as an objective and user-friendly tool for PD rigidity assessment.**

***Index Terms*—Parkinson's disease, wrist rigidity, active orthosis, MDS-UPDRS, usability, SUS**

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder and the second most common neurodegenerative condition worldwide, affecting an estimated 6.1 million people globally, with only Alzheimer's disease being more prevalent [1]. Its prevalence doubled between 1990 and 2015, driven by population aging [1], and rose by 274% from 1990 to 2021, with projections indicating continued acceleration [2]. In Brazil, PD represents a significant and growing public health burden, particularly given the country's demographic transition toward an older population.

The four cardinal motor signs of PD are frequently summarized by the acronym TRAP: Tremor, Rigidity, Akinesia (or bradykinesia), and Postural instability [3]. Despite advances in neuroimaging and biomarker research, PD diagnosis remains predominantly clinical, based on medical history and physical examination, introducing subjectivity and the potential for diagnostic inaccuracy. Diagnostic error rates ranging from 15% to 24% have been reported even in specialized centers [4].

Among the motor symptoms, rigidity is especially relevant in clinical practice. It is present in 83–89% of PD patients [5] and is defined as an increased resistance to passive joint movement that is independent of movement velocity and constant throughout the range of motion. Two subtypes are

recognized: “lead pipe” rigidity, characterized by a uniform, sustained resistance, and “cogwheel” rigidity, in which intermittent catches are superimposed on the underlying muscle tone [5]. Rigidity is influenced by both neural components (basal ganglia dysfunction, altered supraspinal drive) and non-neural components (viscoelastic properties of muscle and connective tissue), which contributes to the complexity of its quantification [5], [6].

Wrist rigidity, in particular, is highly disabling. The wrist joint is essential for daily activities such as eating, writing, and personal hygiene. Impaired wrist mobility reduces independence and quality of life [7], making its accurate measurement a priority for both therapy and monitoring. Furthermore, rigidity is a key outcome measure in clinical trials evaluating dopaminergic therapy and deep brain stimulation (DBS), where subtle changes in muscle tone must be reliably detected.

Currently, the gold standard for rigidity assessment is the Movement Disorder Society–Unified Parkinson’s Disease Rating Scale (MDS-UPDRS) [8], which revised the original UPDRS [9]. Its Part III, Item 3.3 evaluates rigidity at the neck, upper limbs, and lower limbs on a 0–4 ordinal scale. Although widely adopted, this instrument is inherently semi-quantitative and susceptible to inter- and intra-rater variability, with inter-rater disagreement of up to 20% reported [5]. The ordinal and subjective nature of the scale limits its sensitivity in detecting small but clinically meaningful changes in rigidity, particularly for monitoring disease progression or treatment response in longitudinal settings.

To overcome these limitations, several instrumented methods for objective rigidity quantification have been proposed. Devices such as the Bionics Institute Rigidity Device (BiRD) [6] and the PDMeter [10] perform automated passive joint movements and record biomechanical parameters, demonstrating the feasibility of objective rigidity measurement. However, few such devices have undergone simultaneous evaluation of clinical correlation with the MDS-UPDRS and usability in PD patients, both of which are needed before such tools can be adopted clinically.

This study reports a biomechanical and usability assessment of an active wrist orthosis developed at the Nucleus of Innovation and Technological Assessment in Health (NIATS/UFU). The orthosis performs standardized passive wrist flexion-extension movements and records biomechanical signals from which quantitative features are extracted. We quantified Spearman correlations between orthosis-derived features and MDS-UPDRS Item 3.3 rigidity scores, fitted a regression model relating these features to the clinical rigidity score, and assessed device usability via the SUS and SFSS in both PD and healthy control (HC) groups. A companion study used the same device for automatic PD-versus-control classification with machine learning [11]; here we instead examine how orthosis-derived measures agree with clinical rigidity scores and how usable the device is, treating the regression model as an exploratory, interpretable step and not a validated predictor.

II. Materials and Methods

A. Device Description

The active wrist orthosis used in this study (patent BR1020140232826) was developed in the NIATS laboratory as a custom instrumented tool for objective measurement of rigidity in PD. Weighing 850 g, the system is actuated by a Faulhaber DC linear servomotor (LM 2070-080-11, Faulhaber, Schönaich, Germany) driven by a high-precision controller (MCLM 3006 S RS, Faulhaber). The device performs passive flexion and extension of the participant’s wrist, and quantifies rigidity by measuring the motor’s current consumption, which is proportional to the resistive force exerted by the wrist joint during imposed movement. Fig. 1 illustrates the main structural components of the orthosis.

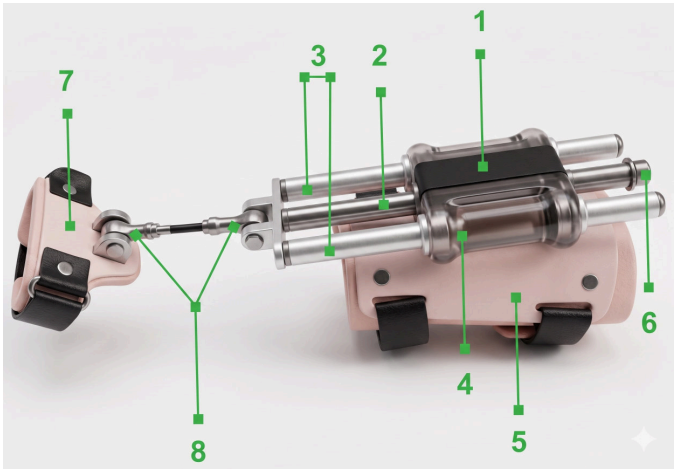
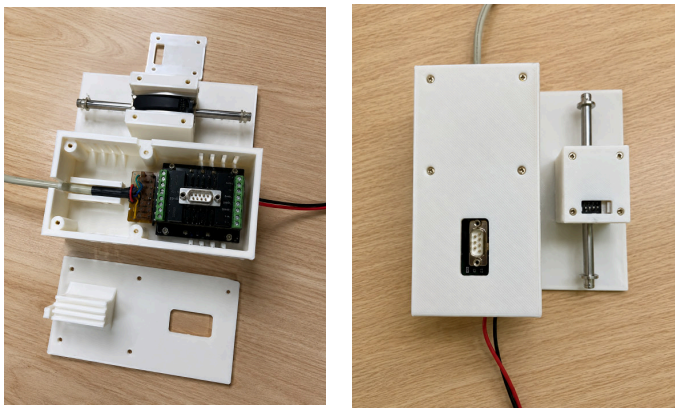


Fig. 1: Active wrist orthosis. (1) Actuator; (2) Actuator’s cylindrical shaft; (3) Auxiliary cylindrical shafts; (4) Actuator casing; (5) Forearm attachment structure; (6) Safety displacement limiter; (7) Hand coupling structure; (8) Spherical joints.

The electronic control unit consists of a Faulhaber MCLM 3006 S RS motion controller and the linear actuator, both housed in a custom 3D-printed PLA enclosure. The controller communicates via RS-232 and is programmed using Faulhaber Motion Manager 6 software. Signal acquisition was managed by the Faulhaber software running on a host PC, which communicated with the MCLM 3006 S RS controller over an RS-232/USB converter cable at 38,400 bps. ASCII commands defined the actuator trajectory as bidirectional displacement of $\pm$ 4,500 encoder increments from the neutral wrist position, at 100 increments/s with no commanded acceleration. Each acquisition produced one CSV file per set, containing time (ms), shaft position (increments), and motor current (mA). Motor current was used as the proxy for the resistive torque generated by the wrist, since it is proportional to the force exerted by the LM 2070-080-11 actuator through a fixed electromechanical constant.

Fig. 2 shows the opened housing revealing the control electronics.



(a) (b)

Fig. 2: (a) Aerial view of the MCLM 3006 S RS microcontroller and the LM 1247-060-01 actuator fixed in their respective housings. (b) Top view of the closed base. Note: the photographed unit corresponds to an earlier prototype of the housing assembly; the actuator used in the present study is the upgraded LM 2070-080-11 model.

Data collection took place in a standardized room with floor and chair markings to ensure reproducible positioning across sessions. The setup included a fixed chair, an adjustable forearm support for the assessed limb, a matched support for the contralateral arm, the orthosis with its dedicated power supply, disinfection materials, and a tripod-mounted smartphone (iPhone 14 Pro Max, Apple Inc., Cupertino, CA, USA) placed perpendicular to the sagittal plane of the assessed wrist to record the movement. Participants were positioned within the floor-marked area, with both forearms on their supports and the assessed wrist in a neutral pronation-supination position. Color-based markers were placed on the anatomical landmarks tracked by the lateral video. Fig. 3 shows an overview of the data collection environment.

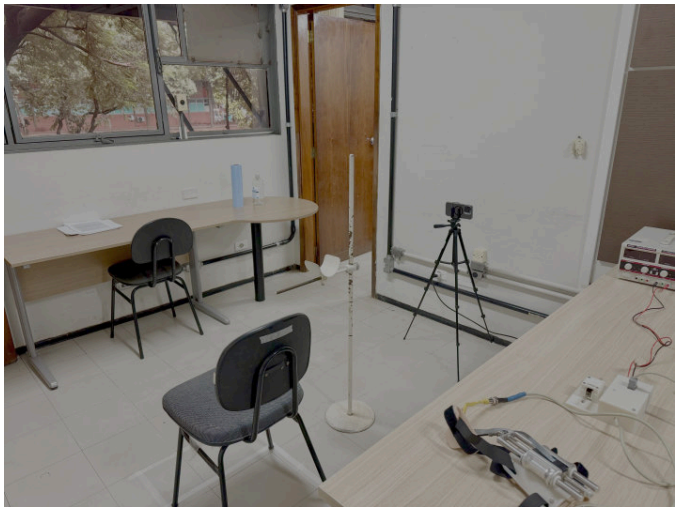
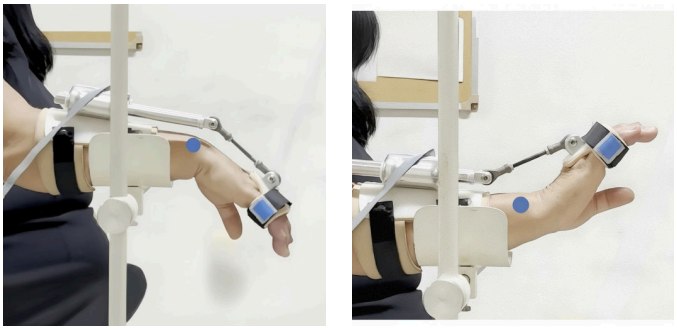


Fig. 3: Data collection setup including the chair, forearm supports, active wrist orthosis, and tripod-mounted smartphone.

Angular displacement of the wrist was estimated using a computer vision-based tracking system with color-based markers and the Kinovea® software (version 2025.1.1). This approach replaced the physical goniometer used in earlier validation studies [12], reducing equipment bulk and improving participant comfort. The orthosis was positioned on the participant's forearm 6 cm from the ulnar styloid process, following established biomechanical alignment recommendations [11]. Signals were acquired at a sampling rate of 166.67 Hz; video was recorded at 60 fps and interpolated to 166.67 Hz using cubic spline interpolation [11]. Fig. 4 shows the experimental setup with the orthosis fitted to a volunteer at maximum flexion (a) and maximum extension (b).



(a) (b)

Fig. 4: Experimental setup of the active wrist orthosis at maximum amplitude: (a) full wrist flexion; (b) full wrist extension. The angle θ indicates the wrist joint angle measured by color-based markers tracked with Kinovea®.

B. *Participants*

Following approval from the Research Ethics Committee of the Federal University of Uberlândia (CAAE: 84004924.1.0000.5152), 30 participants provided written informed consent and were enrolled: 15 with a confirmed PD diagnosis (PD group) and 15 neurologically healthy controls (HC group), matched by age and sex. Table I summarizes the demographic and clinical characteristics. Inclusion criteria for the PD group required a formal clinical diagnosis by a movement disorder specialist. Exclusion criteria for all participants included any other neurological condition, joint disease, or painful condition that could affect wrist movement. PD participants were evaluated in the OFF-medication state. In PD participants, the most affected limb (identified using MDS-UPDRS Item 3.3) was used for orthosis placement; in HC participants, the dominant hand was assessed.

Both groups were well-balanced for age, sex, and overall health profile, with no significant between-group differences in demographic variables. Within the PD group, the Motor Examination (Part III) score indicated mild-to-moderate motor impairment, with rigidity present in 14 of the 15 participants and broadly distributed across the 0–10 range of the Item 3.3 subscore.

TABLE I: Clinical and Demographic Characteristics

Item	HC	PD	Total
Gender – Male	8	8	16
Gender – Female	7	7	14
Age (years)	62.00 ± 5.21 (50–69)	62.07 ± 5.68 (53–72)	–
MDS-UPDRS Part III (Motor)[1]	–	20.93 ± 9.70	–
Rigidity Subscore (Item 3.3)[1]	–	4.53 ± 2.29	–
Total	15	15	30

[1] Part III: Motor Examination (3.1–3.18).
Gender values are counts; Age and MDS-UPDRS scores are mean ± SD.

C. Data Collection Protocol

Participants were seated with the forearm resting horizontally on a padded support, with the orthosis secured 6 cm proximal to the ulnar styloid process following established biomechanical alignment recommendations. The wrist was maintained in a neutral pronation-supination position throughout the protocol. Prior to data acquisition, a standardized 30-second passive stretching routine was applied in each direction (flexion and extension) to reduce viscoelastic stiffness and standardize initial joint conditions. Participants were verbally instructed to remain relaxed and not to resist or voluntarily react to the device's movement, allowing the actuator to act passively on the joint. Data collection commenced only after the researcher confirmed full participant familiarization. Each participant completed three consecutive sets of 10 passive wrist flexion-extension cycles, with 2-minute rest intervals between sets to prevent fatigue-related confounds. Electrical current (mA), time (s), and linear displacement were recorded continuously throughout each cycle at 166.67 Hz.

D. MDS-UPDRS Clinical Assessment

PD participants underwent evaluation of MDS-UPDRS Part III, Item 3.3 (neck and limbs rigidity, five sub-items a–e), conducted by two trained raters blinded to each other's scores. When the difference between raters' total motor scores exceeded five points, a consensus procedure was applied [13], [14]. The final score was computed as the mean of the two evaluations. Both the total motor score (Part III, maximum 132) and the rigidity subscore (Item 3.3, maximum 20) were used in subsequent analyses.

E. Feature Extraction

Twelve biomechanical features were extracted from the conditioned orthosis signals (1–16 Hz band-pass filtering and z-score standardization). As the features are used here only as relative descriptors for correlation and regression, they are treated as dimensionless quantities. The first pair, $C_{Extension}$ and $C_{Flexion}$, is the area under the current-time curve, a charge-like measure of the cumulative current during extension and flexion respectively. $W_{Extension}$ and $W_{Flexion}$ are the area under the force-angle curve, a work-like measure of the mechanical effort, while $FR_{Extension}$ and $FR_{Flexion}$ describe the force rate, obtained from the slope of the force-angle linear regression. The remaining features capture statistical properties of the signal: MAV denotes the mean absolute value, with MAVFD and MAVSD denoting the same measure applied to the first and second signal differences. The set is completed by PEAK, RANGE, and STD.

F. Correlation and Prediction Model

Correlation and regression analyses were restricted to the PD group, since MDS-UPDRS scores are defined only for these participants, and comprised 45 movement sets (three consecutive sets from each of the 15 PD participants). Spearman's ρ was computed between each feature and two clinical scores (MDS-UPDRS Part III and Item 3.3), with a significance level of 0.05. For the prediction model, the twelve features were combined with 66 second-order interaction terms, yielding 78 candidates that were evaluated via stepwise forward selection with an entry threshold of $p < 0.05$. Performance was assessed through R^2, adjusted R^2, RMSE, and Expanded Accuracy (tolerance ± 1 point), all computed on the fitting sample. The twelve-feature linear regression served as the internal baseline against which the optimized stepwise model was compared; formal external validation (e.g., k-fold cross-validation against a constant-mean predictor) is reserved for future work, consistent with the underlying thesis [15].

G. Usability Assessment

Device usability was assessed using two validated instruments administered to all participants immediately after the data collection protocol.

The System Usability Scale (SUS) [16] is a ten-item questionnaire rated on a five-point Likert scale ranging from 1 for strongly disagree to 5 for strongly agree, with items alternating between positively and negatively phrased statements. For scoring, odd-numbered items contribute the score minus one, while even-numbered items contribute five minus the score; all contributions are then summed and multiplied by 2.5 to yield a final value on a 0–100 scale. According to Bangor et al. [17], scores $\geq$68 are considered acceptable, $\geq$71 good, and $\geq$80.3 excellent. The validated Brazilian Portuguese version was used in this study.

The Short Flow State Scale (SFSS) [18] is a nine-item instrument that evaluates the psychological state of flow during task performance. Each item is rated on a five-point Likert scale, and the final score corresponds to the arithmetic mean of the responses, ranging from 1 to 5. The nine dimensions assessed are challenge-skill balance, action-awareness merging, clear goals, unambiguous feedback, concentration on task, sense of control, loss of self-consciousness, time transformation, and autotelic experience. Both instruments were administered in paper form immediately after device removal, with the researcher available for clarification.

III. Results

A. Sample Characterization

Groups were well-matched by age and sex (HC: 62.00 ± 5.21 years; PD: 62.07 ± 5.68 years; independent t-test, $p > 0.05$). Each group comprised eight men and seven women.

Within the PD group, the Mann-Whitney test revealed no significant difference in motor scores between sexes, and the Spearman analysis showed no significant correlation between age and motor severity, confirming a homogeneous and well-characterized sample.

B. MDS-UPDRS Scores

The PD group presented a mean total motor score of 20.93 with a standard deviation of 9.70, ranging from 11 to 42, which indicates mild-to-moderate motor impairment. The rigidity subscore corresponding to Item 3.3 reached a mean of 4.53 with a standard deviation of 2.29 and a median of 4.0, with scores ranging from 0 to 10. Rigidity was present in 14 of the 15 PD participants, one of whom scored 0 on Item 3.3 at the time of assessment, yielding a clinically meaningful spread of rigidity severity for the correlation analysis. No significant differences emerged between male and female subgroups for either score, and age showed no significant correlation with clinical severity.

C. Spearman Correlation Analysis

No feature showed a significant correlation with the total motor score of Part III, a result consistent with the broad and multi-symptom nature of this aggregate measure. In contrast, five features demonstrated significant moderate correlations with the rigidity-specific Item 3.3, as detailed in Table II. The strongest positive associations were observed for the flexion-phase variables $C_{Flexion}$ and $W_{Flexion}$, with Spearman coefficients of 0.408 and 0.342 and corresponding p-values of 0.005 and 0.022. The signal variability descriptors STD, MAVSD, and MAVFD, in turn, showed moderate negative correlations, reflecting the attenuated signal variability that characterizes rigid movement.

TABLE II: Spearman Correlations: Orthosis Features vs. MDS-UPDRS

Feature	ρ (Part III)	p	ρ (Item 3.3)	p
$C_{Flexion}$	0.238	0.115	0.408*	0.005
$W_{Flexion}$	0.210	0.165	0.342*	0.022
STD	−0.176	0.247	−0.326*	0.029
MAVSD	−0.281	0.062	−0.320*	0.032
MAVFD	−0.280	0.062	−0.320*	0.032
RANGE	0.081	0.597	0.270	0.073
PEAK	0.104	0.495	0.238	0.115
$C_{Extension}$	0.092	0.546	−0.144	0.346
$FR_{Extension}$	0.060	0.698	0.113	0.460
$W_{Extension}$	0.075	0.627	0.015	0.923
MAV	0.072	0.638	−0.055	0.721
$FR_{Flexion}$	0.037	0.810	−0.043	0.778

* $p < 0.05$. No correction for multiple comparisons applied (exploratory analysis).

Fig. 5 displays the full correlation profile across all features, with the magnitude of Spearman's ρ shown on the horizontal axis. Bars extending to the right of the origin indicate positive associations with rigidity, while bars to the left indicate negative associations; bars crossing the dashed thresholds at $\rho = \pm 0.3$ identify the five features that reached statistical significance and are highlighted in darker shading.

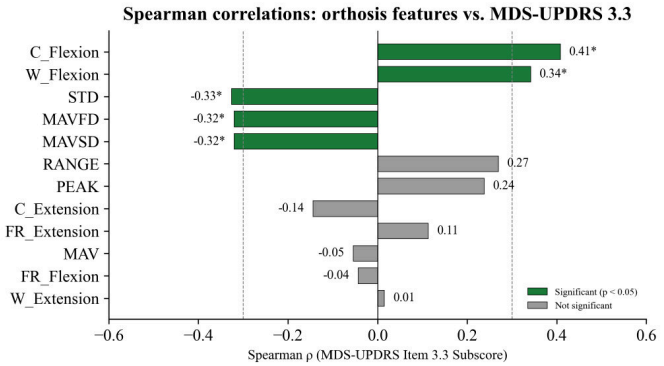


Fig. 5: Spearman correlation coefficients (ρ) between orthosis biomechanical features and MDS-UPDRS Item 3.3 (rigidity subscore). Dark bars indicate statistically significant correlations ($p < 0.05$). Dashed lines mark the $\rho = \pm 0.3$ threshold for moderate correlation.

D. Mathematical Prediction Model

Building on the correlations identified in the previous section, an exploratory stepwise selection refined the predictor set to three second-order interaction terms that proved most informative: MAVFD×STD, STD×$FR_{Flexion}$, and MAVSD×$C_{Flexion}$. To assess whether this parsimonious formulation preserved predictive performance, the resulting optimized model was compared against the internal twelve-feature baseline, with both configurations summarized in Table III. The reported metrics correspond to model fit on the complete sample and should be interpreted accordingly.

TABLE III: Regression Model Performance – Rigidity Score Prediction

Metric	Base Model (12 feat.)	Optimized (3 interact.)
R^2	0.682	0.603
R^2 adj	0.563	**0.574**
RMSE (points)	1.250	1.396
Expanded Accuracy (±1 pt)	75.56%	73.33%

The comparison shown in Table III reveals that the optimized model uses only three interaction terms instead of twelve features, yet its adjusted R^2 (0.574) slightly exceeds the baseline's (0.563), with comparable clinical accuracy. Its RMSE (about 1.40 points, near 7% of the 20-point maximum) is slightly higher than the baseline's 1.25 – a modest cost in raw error for a simpler, more interpretable model. STD appears in two of the three selected interactions, suggesting that signal variability adds information beyond the movement descriptors alone.

E. Usability Assessment

The overall SUS mean reached 84.83 with a standard deviation of 7.28, a value classified as "excellent" according to Bangor et al. [17]. Both groups individually surpassed the 80.3 excellence threshold: PD group 83.33 ± 7.66 (median 85;

range 72.5–97.5); HC group 86.33 $\pm$ 6.81 (median 90; range 70–95). The Mann-Whitney U test revealed no significant between-group difference ($p > 0.05$), indicating that PD-related motor impairments did not meaningfully affect technology interaction. Every individual SUS score in both groups exceeded the acceptable threshold of 68, and the majority surpassed the excellence threshold of 80.3, reflecting uniform device acceptance (Fig. 6).

SFSS scores confirmed a moderate-to-high flow state during device use in both groups (PD: 4.03 $\pm$ 0.26; HC: 4.38 $\pm$ 0.29; on a 1–5 scale; values are between-subject means $\pm$ SD). Both groups reached a high flow level, with the HC group showing a slightly higher mean (a 0.35-point advantage), suggesting that PD participants experienced somewhat lower, though still high, engagement during the protocol. Fig. 7 displays the SFSS score distribution across groups.

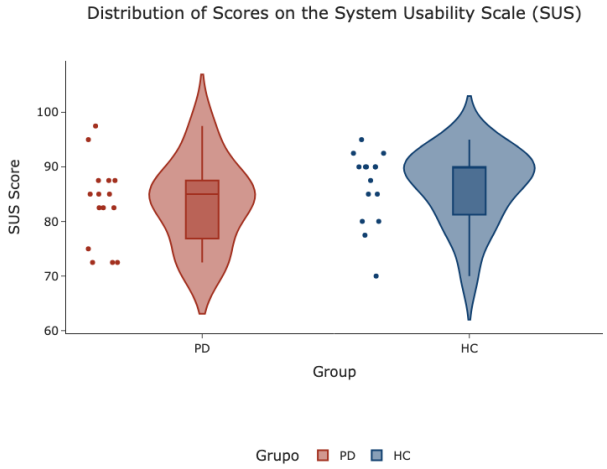


Fig. 6: Distribution of SUS scores by group. Boxes show the interquartile range; dots represent individual scores.

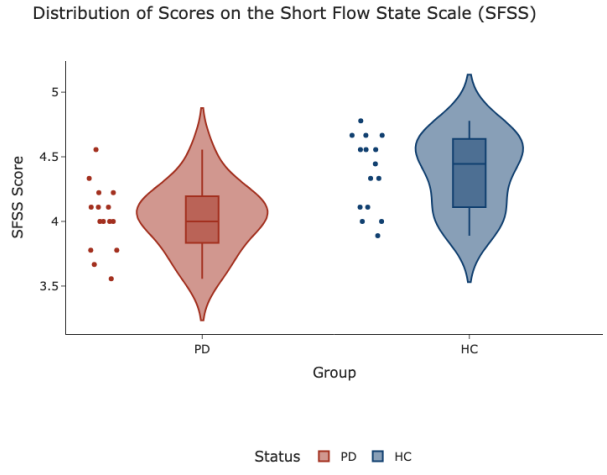


Fig. 7: Distribution of SFSS scores by group. Boxes show the interquartile range; dots represent individual scores.

IV. Discussion

The biomechanical measurements from the wrist orthosis correlated significantly with clinician-rated rigidity in Parkinson's disease, and the device was perceived as highly usable by both patients and healthy controls.

The correlations were selective: the orthosis features associated significantly with MDS-UPDRS Item 3.3, but not with the total motor score of Part III. Part III aggregates tremor, bradykinesia, postural instability, and other signs, which dilutes the rigidity-specific signal the device captures. Item 3.3 isolates the resistance to passive movement that the orthosis quantifies. The same selectivity was reported by Perera et al. [6] with the BiRD device, indicating that instrumented rigidity tools are more sensitive to rigidity-specific clinical items than to composite motor scores.

The strongest positive associations were observed for the charge-like flexion integral ($C_{Flexion}$, $\rho = 0.408$) and the work-like flexion integral ($W_{Flexion}$, $\rho = 0.342$). Both quantify, in relative terms, the cumulative current and mechanical effort required to overcome wrist resistance during flexion, mirroring the resistant force clinicians assess during passive movement. The negative correlations of STD, MAVFD, and MAVSD have a clear biomechanical basis: a rigid joint produces a more uniform force profile during passive movement. As rigidity increases, the wrist offers more constant resistance to the imposed motion, lowering the standard deviation and the derivative-based descriptors.

This flexion-phase specificity is reinforced by an independent group-level analysis reported in the underlying thesis [15], in which the only three features that differed significantly between PD and HC participants were all flexion-phase descriptors: $C_{Flexion}$ ($p = 0.003$), $W_{Flexion}$ ($p = 0.006$), and $FR_{Flexion}$ ($p = 0.015$). Two of them, $C_{Flexion}$ and $W_{Flexion}$, are precisely the features that track rigidity severity in the present analysis. The flexion phase therefore carries information both about group membership and about the degree of clinical impairment, which converges on the same conclusion from two independent directions.

The optimized prediction model suggests that wrist rigidity, as measured by Item 3.3, is better captured by feature interactions than by individual features. STD appears in two of the three selected interaction terms (MAVFD$\times$STD and STD$\times FR_{Flexion}$), indicating that signal variability modulates the dynamic movement descriptors, accounting for a sizeable share of the clinical rigidity variance in this sample, with an adjusted R^2 of 0.574. On the fitting sample, the model has a root mean squared error of about 1.40 points and an expanded accuracy of 73.3%, with its rounded estimate landing within one point of the clinician's score in about three of four cases. As these metrics are not cross-validated, they should be read as exploratory rather than as evidence of out-of-sample predictive performance.

The mean SUS score of 84.83 is encouraging, given the motor and cognitive difficulties that often complicate technology use in PD. An "excellent" rating in this population [17]

likely reflects the ergonomic refinements made across device iterations – above all, replacing the physical goniometer with computer-vision tracking. The 3.0-point PD–HC difference in mean SUS (86.33 versus 83.33) indicates that the device is accessible across functional levels. The SFSS scores point the same way: both groups reported a high flow state, with PD slightly below controls (4.03 versus 4.38 on a 1–5 scale). The small difference fits the higher motor and cognitive load in PD, but the still-high scores show that the device kept users engaged.

This study has limitations. The sample of 30 participants limits statistical power, and the correlation analysis was exploratory, without correction for multiple comparisons. The absence of test-retest reliability data and longitudinal assessments restricts conclusions about device sensitivity to disease progression or to therapy. Larger, stratified cohorts with longitudinal follow-up are needed to extend these findings. Each participant contributed three movement sets sharing the same clinical score, so the observations were not fully independent and the correlation and regression results should be read as exploratory. The reported R^2, adjusted R^2, and RMSE reflect model fit on the full sample rather than out-of-sample performance; future work will use per-subject aggregation or mixed-effects modeling with k-fold cross-validation against a constant-mean baseline, as discussed in the underlying thesis [15].

V. Conclusion

The active wrist orthosis serves as an objective complement to clinical rigidity assessment in PD. Flexion-phase features, namely the relative charge-like ($C_{Flexion}$) and work-like ($W_{Flexion}$) integrals, correlated moderately and significantly with MDS-UPDRS Item 3.3, while showing no significant association with the total motor score, confirming measurement specificity. An exploratory three-interaction-term regression model related these features to the clinical rigidity score ($R^2_{adj} = 0.574$; 73.3% expanded accuracy within one point), in line with values reported for the BiRD [6] and PDMeter [10] devices, and suggests that feature interactions deserve attention in future predictive models. PD patients and healthy controls rated the device as excellent in usability (SUS = 84.83) and reported a moderate-to-high flow state during use (SFSS = 4.20 on a 1–5 scale), supporting clinical deployment. The orthosis is therefore an objective, user-friendly tool for quantifying rigidity in PD, with potential use in diagnostic support and longitudinal monitoring.

Acknowledgment

The present work was carried out with the support of the Ministry of Science, Technology, and Innovation (MCTI); Federal Institute of Education, Science, and Technology Goiano (IF Goiano); National Council for Scientific and Technological Development (CNPq); Coordination for the Improvement of Higher Education Personnel (CAPES); and Foundation for Research Support of the State of Minas Gerais (FAPEMIG). A. A. Pereira is a CNPq research fellow (grant 309525/2021-7). The authors also thank the volunteers who participated in this study and acknowledge the infrastructure provided by NIATS/UFU.

References

[1] E. R. Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016," *Lancet Neurology*, vol. 17, no. 11, pp. 939–953, 2018.

[2] M. Li, X. Ye, Z. Huang, L. Ye, and C. Chen, "Global burden of Parkinson's disease from 1990 to 2021: A population-based study," *BMJ Open*, vol. 15, p. e095610, 2025.

[3] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, pp. 368–376, 2008.

[4] G. Rizzo *et al.*, "Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis," *Neurology*, vol. 86, no. 6, pp. 566–576, 2016.

[5] M. d. R. Ferreira-Sánchez, M. Moreno-Verdú, and R. Cano-de-la Cuerda, "Quantitative measurement of rigidity in Parkinson's disease: A systematic review," *Sensors*, vol. 20, no. 3, p. 880, 2020.

[6] T. Perera *et al.*, "A palm-worn device to quantify rigidity in Parkinson's disease," *Journal of Neuroscience Methods*, vol. 317, pp. 113–120, 2019.

[7] C. M. Alves *et al.*, "Wrist rigidity evaluation in Parkinson's disease: A scoping review," *Healthcare*, vol. 10, no. 11, p. 2178, 2022.

[8] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS)," *Movement Disorders*, vol. 23, pp. 2129–2170, 2008.

[9] MDS Task Force on Rating Scales for Parkinson's Disease, "The unified Parkinson's disease rating scale (UPDRS): status and recommendations," *Movement Disorders*, vol. 18, pp. 738–750, 2003.

[10] L. Raiano *et al.*, "PDMeter: A wrist wearable device for an at-home assessment of the Parkinson's disease rigidity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 6, pp. 1325–1333, 2020.

[11] A. A. Pereira *et al.*, "Parkinson's disease classification using machine learning and wrist rigidity measurements from an active orthosis," *Journal of Sensor and Actuator Networks*, vol. 15, no. 1, p. 1, 2026.

[12] C. T. Ribeiro, "Validação de uma órtese de punho instrumentada para avaliação da rigidez em pessoas com a doença de Parkinson," Master's thesis, Universidade Federal de Uberlândia, Uberlândia, Brazil, 2023.

[13] A. J. Noyce *et al.*, "Subtle motor disturbances in PREDICT-PD participants," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 88, no. 3, pp. 212–217, 2017.

[14] L. Marsili *et al.*, "Bradykinesia in neurodegenerative disorders: A blinded video analysis of pathology-proven cases," *Movement Disorders*, vol. 38, no. 3, pp. 496–501, 2023.

[15] D. H. d. Silva, "Avaliação quantitativa da rigidez de punho em pessoas com a doença de Parkinson por meio de uma órtese ativa," Ph.D. dissertation, Universidade Federal de Uberlândia, Uberlândia, Brazil, 2025.

[16] J. Brooke, "SUS: A retrospective," *Journal of Usability Studies*, vol. 8, pp. 29–40, 2013.

[17] A. Bangor, P. T. Kortum, and J. T. Miller, "Determining what individual SUS scores mean: adding an adjective rating scale," *Journal of Usability Studies*, vol. 4, pp. 114–123, 2009.

[18] S. A. Jackson, P. C. Martin, and R. C. Eklund, "Long and short measures of flow," *Journal of Sport & Exercise Psychology*, vol. 30, pp. 561–587, 2008.

*Corresponding author: daniel.hilario@ifgoiano.edu.br

Análise da Influência das Características do Sinal no Desempenho de Diferentes Algoritmos de Extração da Dimensão Fractal em Dados Biomecânicos da Marcha Humana

Nuno Brandão Di Barros Cachapuz Caiado
Programa de Mestrado em Engenharia Biomédica
Universidade Federal de Uberlândia
Uberlândia, Brasil
NunoBrandaoCaiado@gmail.com

Marcus Fraga Vieira
Programa de Pós-Graduação em Engenharia Biomédica
Universidade Federal de Uberlândia
Uberlândia, Brasil
Marcus.Fraga.Vieira@gmail.com

***Resumo*—A análise fractal tem sido amplamente empregada na biomecânica para investigar a complexidade presente em sinais associados à marcha humana. Entretanto, fatores relacionados à aquisição e ao processamento dos sinais, como a frequência de amostragem e a presença de ruído, podem influenciar diretamente a estabilidade das métricas obtidas. O objetivo deste estudo foi investigar a influência dessas características sobre o desempenho dos algoritmos Correlation Dimension (D_2) e Multifractal Detrended Fluctuation Analysis (MF-DFA). Foram analisados sinais biomecânicos de marcha provenientes de jovens, idosos não caidores e idosos caidores, submetidos a diferentes frequências de amostragem, velocidades e níveis de ruído branco gaussiano. Os resultados demonstraram que ambos os algoritmos foram influenciados pelas modificações impostas aos sinais, porém apresentaram comportamentos distintos quanto à robustez e sensibilidade. A Correlation Dimension mostrou maior sensibilidade à redução da frequência de amostragem, enquanto o MF-DFA apresentou maior influência das interações entre velocidade e ruído. Os achados fornecem subsídios para a utilização mais consistente de métricas fractais em estudos biomecânicos e clínicos.**

***Index Terms*—Marcha humana, Dimensão fractal, Correlation Dimension, MF-DFA, Complexidade, Biomecânica**

I. Introdução

A marcha humana representa um sistema dinâmico complexo cuja variabilidade emerge da interação entre mecanismos neuromotores, sensoriais e biomecânicos. A compreensão dessa variabilidade tem sido amplamente explorada por meio de métodos não lineares capazes de caracterizar a complexidade presente nos sinais biomecânicos, conforme discutido por Lopes e Betrouni [7].

Entre os métodos mais utilizados destacam-se a Correlation Dimension (CD), baseada na reconstrução do espaço de fase, e o Multifractal Detrended Fluctuation Analysis (MF-DFA), fundamentado na análise das propriedades multifractais das séries temporais. Esteller et al. [8] demonstram a relevância da análise por algoritmos de dimensão fractal e das diferenças associadas aos métodos utilizados. Embora ambos sejam amplamente empregados em estudos biomecânicos, ainda existe limitada compreensão sobre sua robustez frente às alterações frequentemente presentes nos sinais experimentais.

A frequência de amostragem e a presença de ruído representam dois fatores capazes de modificar significativamente a estrutura temporal dos sinais, podendo influenciar diretamente as estimativas fractais obtidas. Nesse contexto, Mehdizadeh [5] investigou a influência do ruído sobre métodos de análise não linear, enquanto Souza et al. [6] avaliaram a influência da adição de ruído em sinais biomecânicos da marcha. Dessa forma, torna-se importante compreender como diferentes algoritmos respondem a essas modificações.

O objetivo deste estudo foi investigar a influência da frequência de amostragem e do ruído branco gaussiano sobre o desempenho dos algoritmos Correlation Dimension e MF-DFA aplicados a sinais biomecânicos da marcha humana.

II. Materiais e métodos

A. Participantes

Foram analisados dados provenientes de um estudo experimental conduzido pelo Laboratório de Biomecânica da Universidade Federal de Goiás (UFG). A amostra foi composta por jovens saudáveis (J), idosos não caidores (INC) e idosos caidores (IC).

O estudo experimental que originou os dados foi previamente aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Goiás (UFG), sendo todos os procedimentos realizados em conformidade com as normas éticas para pesquisas envolvendo seres humanos. Todos os participantes foram incluídos no estudo mediante consentimento livre e esclarecido..

B. Aquisição dos Dados

Os sinais biomecânicos foram adquiridos por meio de um sistema tridimensional de captura de movimento com frequência original de amostragem de 120 Hz durante protocolos de marcha realizados em diferentes velocidades, 80PWS,

100PWS e 120PWS, (Preferred Walking Speed) correspondentes a 80%, 100% e 120% da Velocidade de Caminhada Preferida.

As séries temporais analisadas foram obtidas a partir do marcador posicionado sobre o processo espinhoso da décima vértebra torácica (T10). Esse ponto anatômico é amplamente utilizado em estudos biomecânicos como representação da dinâmica do tronco durante a locomoção, por estar localizado em uma região central do segmento torácico e apresentar estreita relação com os mecanismos de controle postural e estabilidade da marcha. Dessa forma, as métricas fractais avaliadas neste estudo refletem a complexidade associada ao comportamento dinâmico do tronco humano durante a marcha.

C. Processamento dos Sinais

Os sinais originais foram submetidos a duas formas de degradação controlada:

- redução da frequência de amostragem para 60 Hz e 30 Hz;
- adição de ruído branco gaussiano com relações sinal-ruído de 50, 40, 30, 20 e 10 dB.

D. Correlation Dimension

A Correlation Dimension foi calculada a partir da reconstrução do espaço de fase, conforme os fundamentos propostos por Grassberger e Procaccia [1] e Takens [2], utilizando atraso temporal obtido pelo primeiro mínimo da informação mútua e dimensão de imersão estimada pelo método dos falsos vizinhos mais próximos.

E. MF-DFA

O algoritmo Multifractal Detrended Fluctuation Analysis (MF-DFA), proposto por Kantelhardt et al. [3] e posteriormente detalhado por Ihlen [4], foi aplicado para caracterizar as propriedades multifractais das séries temporais biomecânicas. O método avalia o comportamento de escala das flutuações após a remoção de tendências em múltiplas escalas temporais, permitindo estimar correlações de longa duração e caracterizar a estrutura multifractal dos sinais analisados.

F. Análise Estatística

Os resultados foram analisados por meio de análise de variância (ANOVA) de medidas repetidas, utilizando o software JASP e adotando nível de significância de p¡0,05. Foram avaliados os efeitos principais e as interações entre os fatores velocidade da marcha, frequência de amostragem, nível de ruído e grupo, de acordo com as condições experimentais analisadas. As análises foram realizadas separadamente para os parâmetros obtidos pela Correlation Dimension (CD) e pelo Multifractal Detrended Fluctuation Analysis (MF-DFA), permitindo avaliar a influência das modificações impostas aos sinais sobre cada método.

Para a avaliação da influência da frequência de amostragem, foram consideradas as condições de 120, 60 e 30 Hz, juntamente com as diferentes velocidades de marcha e os grupos avaliados. Para a avaliação da influência do ruído, foram considerados os diferentes níveis de relação sinal-ruído aplicados aos sinais, permitindo analisar os efeitos do ruído e suas interações com a velocidade da marcha e o grupo.

Na análise da Correlation Dimension em função da frequência de amostragem, as diferenças identificadas entre as frequências foram posteriormente avaliadas por comparações pós-hoc. Adicionalmente, foram calculados coeficientes de correlação de Pearson (r), utilizando a condição de 100% PWS e 120 Hz como referência, com o objetivo de avaliar a associação entre os resultados obtidos nas diferentes condições experimentais. A incerteza das estimativas de correlação foi avaliada por meio de intervalos de confiança de 95% (IC95%), calculados utilizando a transformação z de Fisher.

III. Resultados

A. Influência da Frequência de Amostragem na Correlation Dimension

A análise da Correlation Dimension demonstrou efeito significativo da frequência de amostragem, indicando que alterações na resolução temporal modificam os valores estimados pela métrica. Em contraste, a velocidade e o grupo não apresentaram efeitos principais significativos. Entretanto, foram identificadas interações entre frequência e grupo e entre velocidade e frequência, demonstrando que o efeito da resolução temporal depende das condições analisadas. A interação entre velocidade, frequência e grupo não apresentou significância estatística. Os resultados completos são apresentados na Tabela I.

A Figura 1 permite visualizar esse comportamento na condição de 100% PWS. Os idosos caidores (IC) apresentaram maior variação dos valores de Correlation Dimension entre as frequências, com valores médios de aproximadamente 2,249 em 30 Hz, 2,167 em 60 Hz e 2,203 em 120 Hz. Em contraste, INC e J apresentaram trajetórias mais regulares, com redução gradual dos valores à medida que a frequência aumentou. Essa diferença entre as trajetórias é compatível com a interação significativa entre frequência e grupo observada na ANOVA, embora as diferenças apresentadas graficamente devam ser interpretadas como tendências descritivas.

A análise de correlação apresentada na Figura 2 complementa esses resultados ao avaliar a preservação do padrão da Correlation Dimension em relação à condição de referência de 100% PWS e 120 Hz. Mantendo-se 100% PWS, os grupos INC e J apresentaram coeficientes de Pearson próximos da unidade mesmo após a redução para 60 e 30 Hz, indicando elevada correspondência com os resultados obtidos em 120 Hz. Dessa forma, embora a frequência tenha influenciado significativamente a magnitude da Correlation Dimension, a relação entre os resultados permaneceu amplamente preservada nesses grupos.

O grupo IC apresentou maior variação, com correlações de aproximadamente 0,60 em 60 Hz e 0,79 em 30 Hz. Esse comportamento também foi observado na Figura 1 e pode estar relacionado, ao menos parcialmente, ao menor tamanho amostral desse grupo, que pode aumentar a variabilidade das

estimativas. Essa interpretação, entretanto, deve ser considerada com cautela, não sendo possível atribuir exclusivamente ao tamanho da amostra as diferenças observadas.

A Figura 2 também demonstra influência importante da velocidade da marcha sobre a preservação do padrão da métrica. Nos grupos INC e J, a alteração da velocidade produziu reduções acentuadas da correlação com a referência, mesmo quando a frequência permaneceu em 120 Hz. Esse comportamento foi particularmente evidente no grupo J, que apresentou correlações negativas nas condições de 80% PWS. Assim, embora a velocidade não tenha apresentado efeito principal significativo na ANOVA, a análise de Pearson indica que modificações na condição de marcha podem alterar a correspondência dos resultados com a referência.

Em conjunto, as Figuras 1 e 2 mostram aspectos complementares da resposta da Correlation Dimension. A Figura 1 evidencia alterações na magnitude da métrica associadas à frequência, enquanto a Figura 2 demonstra que essas alterações não necessariamente resultam em perda da correspondência com a condição de referência. Dessa forma, a Correlation Dimension apresenta sensibilidade quantitativa à frequência de amostragem, mas também relativa robustez quanto à preservação do padrão dos resultados, especialmente nos grupos INC e J.

Do ponto de vista metodológico, esses resultados reforçam a importância da resolução temporal na aplicação da Correlation Dimension. A redução da frequência diminui a densidade de pontos disponíveis para a reconstrução do espaço de fase e pode modificar a representação geométrica da dinâmica do sistema. Embora os elevados coeficientes de correlação indiquem que frequências inferiores podem preservar parte importante do padrão da métrica, a aquisição em 120 Hz permanece recomendável quando disponível, por preservar maior resolução temporal e reduzir a possibilidade de perda de informação.

Assim, os resultados indicam que frequência de amostragem e velocidade influenciam propriedades distintas da Correlation Dimension. A combinação da ANOVA, da análise dos valores médios e das correlações permite caracterizar de forma mais abrangente a resposta da métrica às modificações impostas aos sinais biomecânicos.

B. Influência do Ruído na Correlation Dimension

O fator ruído apresentou efeito significativo sobre a Correlation Dimension. Entretanto, não foram observadas diferenças significativas associadas aos grupos analisados, às velocidades de marcha ou às interações entre os fatores experimentais. Esses resultados indicam que a Correlation Dimension é influenciada pela degradação do sinal, porém essa influência ocorre de maneira relativamente homogênea entre as diferentes populações avaliadas. Os resultados completos da análise estatística são apresentados na Tabela II

C. Influência da Frequência de Amostragem no MF-DFA

Para o MF-DFA foram observados efeitos significativos da velocidade de marcha, da frequência de amostragem e

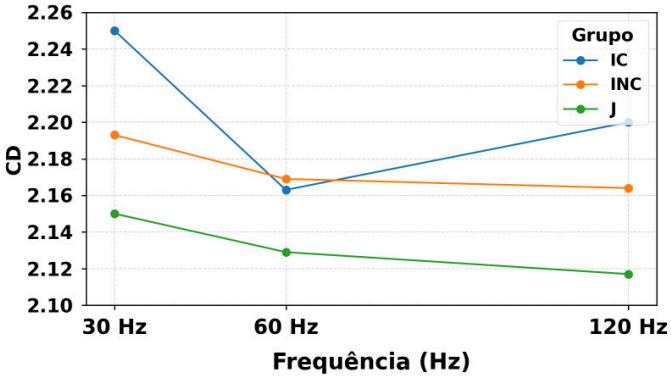


Figura 1. Valores médios da Dimensão de Correlação (CD) em função da frequência de amostragem na condição de 100% PWS para os grupos de Idosos Caidores (IC), Idosos Não Caidores (INC) e Jovens (J). Os valores representam a média com margem de erro considerada de $\pm 3\%$ para interpretação.

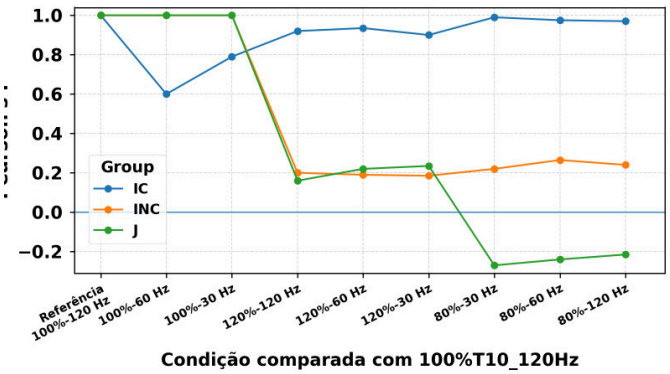


Figura 2. Coeficientes de correlação de Pearson (r) comparando as diferentes condições com a referência 100%T10_120Hz para os grupos IC, INC e J. As estimativas foram avaliadas considerando intervalo de confiança de 95% (IC95%), calculado pela transformação z de Fisher.

da interação entre velocidade e frequência. Não foram identificadas diferenças significativas envolvendo os grupos ou interações que incluíssem o fator grupo. Os resultados sugerem que o comportamento multifractal dos sinais é influenciado principalmente pelas condições experimentais relacionadas à execução da tarefa e pela resolução temporal dos dados. Os resultados completos da análise estatística são apresentados na Tabela III.

A representação dos valores médios do MF-DFA em função da frequência de amostragem, considerando a condição de 100% PWS, permite visualizar de forma mais detalhada a influência da resolução temporal sobre os valores obtidos pelo método (Figura 3). Observa-se uma tendência progressiva de aumento do MF-DFA com o aumento da frequência de amostragem em todos os grupos avaliados. Em 30 Hz, os valores permaneceram próximos de 0,97–0,98, aumentando para aproximadamente 1,26–1,28 em 60 Hz e atingindo valores próximos de 1,45–1,47 em 120 Hz. Dessa forma, a tendência observada graficamente é consistente com o efeito significativo da frequência de amostragem identificado pela análise

Tabela I
ANOVA PARA CORRELATION DIMENSION SOB DIFERENTES FREQUÊNCIAS DE AMOSTRAGEM

Parámetros	**Dados Estatísticos**	
	Correção de Esfericidade	***p***
Velocidade	Nenhum	.767
	Greenhouse-Geisser	.632
	Huynh-Feldt	.634
Velocidade × Grupo	Nenhum	.835
	Greenhouse-Geisser	.720
	Huynh-Feldt	.722
Frequência	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Frequencia × Grupo	Nenhum	.004
	Greenhouse-Geisser	.009
	Huynh-Feldt	.009
Velocidade × Frequencia	Nenhum	.002
	Greenhouse-Geisser	.003
	Huynh-Feldt	.002
Velocidade × Frequencia × Grupo	Nenhum	.066
	Greenhouse-Geisser	.079
	Huynh-Feldt	.071

Tabela II
ANOVA PARA CORRELATION DIMENSION SOB DIFERENTES NÍVEIS DE RUÍDO

Parámetros	**Dados Estatísticos**	
	Correção de Esfericidade	***p***
Velocidade	Nenhum	.764
	Greenhouse-Geisser	.653
	Huynh-Feldt	.656
Velocidade × Grupo	Nenhum	.610
	Greenhouse-Geisser	.541
	Huynh-Feldt	.543
Ruido	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Ruido × Grupo	Nenhum	.375
	Greenhouse-Geisser	.367
	Huynh-Feldt	.368
Velocidade × Ruido	Nenhum	.491
	Greenhouse-Geisser	.428
	Huynh-Feldt	.434
Velocidade × Ruido × Grupo	Nenhum	.186
	Greenhouse-Geisser	.255
	Huynh-Feldt	.251

estatística.

Além do aumento progressivo da métrica, a Figura 3 demonstra que a alteração observada entre 30 e 60 Hz é visualmente mais acentuada do que aquela observada entre 60 e 120 Hz. Esse comportamento indica que a relação entre a frequência de amostragem e os valores obtidos pelo MF-DFA não apresenta uma variação proporcional entre todos os níveis analisados. Entretanto, essa diferença na magnitude das variações deve ser interpretada de maneira descritiva, uma vez que a análise apresentada não permite estabelecer, apenas pela representação gráfica, que a alteração entre 30 e 60 Hz seja estatisticamente superior àquela observada entre 60 e 120 Hz.

Outro aspecto relevante apresentado na Figura 3 é a elevada proximidade entre as trajetórias dos grupos de idosos caidores (IC), idosos não caidores (INC) e jovens (J). Apesar de pequenas diferenças nos valores médios em cada frequência, os três grupos apresentam praticamente o mesmo padrão de resposta, caracterizado pelo aumento sistemático do MF-DFA entre 30, 60 e 120 Hz. A proximidade e o comportamento aproximadamente paralelo das trajetórias são consistentes com a ausência de interação significativa entre frequência e grupo observada na análise estatística, sugerindo que a influência da frequência de amostragem sobre o MF-DFA ocorreu de maneira relativamente uniforme entre as populações avaliadas.

Esse resultado também contribui para diferenciar o efeito associado à resolução temporal daquele relacionado às características dos grupos. Considerando que não foram identificados efeitos significativos envolvendo o fator grupo, enquanto a frequência apresentou efeito significativo, as alterações observadas na Figura 3 estão predominantemente associadas às modificações impostas à frequência de amostragem. Nesse sentido, dentro das condições avaliadas, a resposta do MF-DFA parece apresentar maior dependência das características temporais do sinal do que das diferenças entre as populações analisadas.

A dependência observada em relação à frequência de amostragem pode estar relacionada à própria natureza do MF-DFA, uma vez que o método caracteriza o comportamento das flutuações do sinal ao longo de múltiplas escalas temporais. A redução da frequência de amostragem diminui o número de pontos disponíveis na série temporal e modifica sua resolução temporal, podendo alterar a representação das flutuações utilizadas na estimativa dos parâmetros multifractais. Consequentemente, sinais analisados em frequências distintas podem produzir diferentes estimativas mesmo quando provenientes de uma mesma condição experimental. A tendência apresentada na Figura 3 reforça, portanto, a importância da padronização da frequência de amostragem em estudos que utilizem o MF-DFA para caracterização de sinais biomecânicos.

Embora a Figura 3 represente especificamente a condição de 100% PWS e demonstre elevada similaridade entre os grupos nessa velocidade, a interação significativa entre velocidade e frequência identificada pela ANOVA indica que a influência da resolução temporal sobre o MF-DFA também depende da velocidade de marcha considerada. Dessa forma, a tendência apresentada em 100% PWS não deve ser generalizada isoladamente para as demais velocidades avaliadas. Em conjunto, os resultados demonstram que a interpretação dos parâmetros obtidos pelo MF-DFA deve considerar simultaneamente a frequência de amostragem e as condições de execução da tarefa, uma vez que alterações nesses fatores podem modificar a caracterização multifractal dos sinais e, consequentemente, influenciar comparações entre diferentes protocolos experimentais.

D. Influência do Ruído no MF-DFA

O MF-DFA apresentou o comportamento mais complexo entre todos os cenários avaliados. Foram observados efeitos significativos para velocidade, ruído, interação entre veloci-

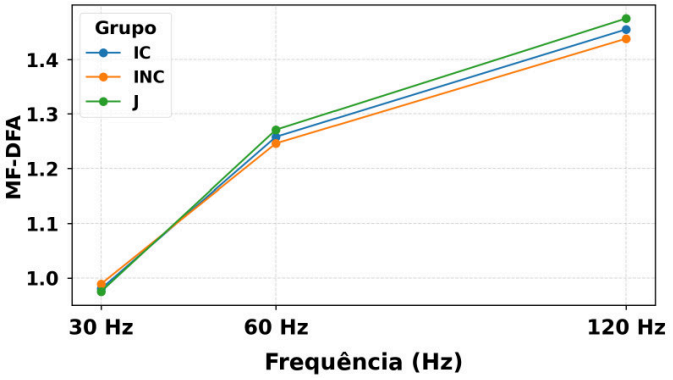


Figura 3. Valores médios do MF-DFA em função da frequência de amostragem na condição de 100% PWS para os grupos IC, INC e J. Os valores representam a média com margem de erro considerada de $\pm 3\%$ para interpretação.

Tabela III
ANOVA PARA MF-DFA SOB DIFERENTES FREQUÊNCIAS DE AMOSTRAGEM

Parâmetros	**Dados Estatísticos**	
	Correção de Esfericidade	***p***
Velocidade	Nenhum	.032
	Greenhouse-Geisser	.035
	Huynh-Feldt	.033
Velocidade × Grupo	Nenhum	.200
	Greenhouse-Geisser	.204
	Huynh-Feldt	.202
Frequência	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Frequência × Grupo	Nenhum	.367
	Greenhouse-Geisser	.348
	Huynh-Feldt	.348
Velocidade × Frequência	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Velocidade × Frequência × Grupo	Nenhum	.777
	Greenhouse-Geisser	.668
	Huynh-Feldt	.676

Tabela IV
ANOVA PARA MF-DFA SOB DIFERENTES NÍVEIS DE RUÍDO

Parâmetros	**Dados Estatísticos**	
	Correção de Esfericidade	***p***
Velocidade	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Velocidade × Grupo	Nenhum	.003
	Greenhouse-Geisser	.004
	Huynh-Feldt	.004
Ruido	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Ruido × Grupo	Nenhum	.680
	Greenhouse-Geisser	.518
	Huynh-Feldt	.518
Velocidade × Ruido	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001
Velocidade × Ruido × Grupo	Nenhum	<.001
	Greenhouse-Geisser	<.001
	Huynh-Feldt	<.001

dade e grupo, interação entre velocidade e ruído e interação tripla entre velocidade, ruído e grupo. Por outro lado, a interação entre ruído e grupo não apresentou significância estatística, indicando que a influência direta do ruído ocorreu de forma semelhante entre os grupos, embora suas interações com a velocidade tenham produzido respostas distintas. Esses resultados demonstram que o MF-DFA apresenta elevada sensibilidade às alterações impostas ao sinal e às condições experimentais da marcha. Os resultados completos da análise estatística são apresentados na Tabela IV

E. Discussão

O presente estudo investigou a influência da frequência de amostragem e da adição de ruído branco gaussiano sobre o desempenho da Correlation Dimension (D_2) e do Multifractal Detrended Fluctuation Analysis (MF-DFA) aplicados a sinais biomecânicos. De modo geral, os resultados demonstraram que ambos os métodos são influenciados pelas características dos sinais, porém apresentam comportamentos distintos frente às modificações introduzidas, evidenciando a importância das condições de aquisição e processamento na interpretação das métricas fractais.

Na Correlation Dimension, a frequência de amostragem apresentou efeito significativo. Embora não tenham sido observadas diferenças entre 120 e 60 Hz, a redução para 30 Hz produziu alterações significativas, indicando maior influência sobre a estimativa da métrica em reduções mais acentuadas da resolução temporal. Esse comportamento pode estar relacionado à reconstrução do espaço de fase, uma vez que a redução da frequência diminui a quantidade de informações temporais disponíveis e pode modificar a distribuição geométrica dos pontos que constituem o atrator reconstruído, considerando os princípios de reconstrução descritos por Takens [2].

A Figura 1 permite visualizar esse comportamento na condição de 100% PWS. Enquanto INC e J apresentaram trajetórias relativamente regulares entre 30, 60 e 120 Hz, o grupo IC apresentou maior variação entre as frequências, com comportamento não monotônico. A divergência entre essas trajetórias é compatível com a interação significativa entre frequência e grupo, indicando que a influência da resolução temporal sobre a Correlation Dimension não ocorreu de maneira completamente uniforme entre as populações analisadas.

A análise de correlação apresentada na Figura 2 complementa esses resultados ao comparar as diferentes condições experimentais com a referência de 100% PWS e 120 Hz. Quando a velocidade foi mantida em 100% PWS, INC e J apresentaram coeficientes próximos da unidade mesmo após a redução para 60 e 30 Hz. Assim, embora a frequência tenha produzido alterações significativas na magnitude da Correlation Dimension, a correspondência com a condição de referência permaneceu elevada nesses grupos.

O grupo IC apresentou maior variação, com correlações de

aproximadamente 0,60 em 60 Hz e 0,79 em 30 Hz. Esse comportamento, também observado na Figura 1, pode estar relacionado parcialmente ao menor tamanho amostral desse grupo, que aumenta a influência das características individuais sobre as estimativas. Entretanto, essa interpretação deve ser considerada com cautela, não sendo possível atribuir exclusivamente ao tamanho da amostra as diferenças observadas.

A Figura 2 também evidencia influência da velocidade sobre a preservação do padrão da Correlation Dimension. Nos grupos INC e J, a alteração para 120% ou 80% PWS reduziu consideravelmente as correlações com a referência, mesmo quando a frequência permaneceu em 120 Hz. Esse comportamento foi particularmente evidente no grupo J, que apresentou correlações negativas nas condições de 80% PWS. Embora a velocidade não tenha apresentado efeito principal significativo na ANOVA, esses resultados indicam que modificações na condição de marcha podem alterar a correspondência dos resultados individuais sem necessariamente produzir um deslocamento sistemático de sua magnitude média.

Em conjunto, as Figuras 1 e 2 demonstram que a Correlation Dimension apresenta sensibilidade quantitativa à frequência de amostragem, mas também relativa preservação do padrão em diferentes resoluções temporais, principalmente em INC e J. Considerando, entretanto, o efeito significativo da frequência e a dependência da reconstrução do espaço de fase em relação à resolução temporal, a aquisição em 120 Hz permanece metodologicamente recomendável quando disponível, por preservar maior quantidade de informação do sinal original.

Em relação ao ruído branco gaussiano, a Correlation Dimension apresentou alterações significativas associadas à degradação do sinal, porém não foram identificadas interações significativas envolvendo grupo ou velocidade. Esse resultado sugere uma influência relativamente homogênea do ruído nas condições analisadas. A introdução de componentes aleatórios pode modificar a distribuição dos pontos no espaço de fase e, consequentemente, interferir na estimativa da dimensionalidade, comportamento compatível com a influência do ruído sobre análises não lineares discutida por Mehdizadeh [5] e com os resultados apresentados por Souza et al. [6] em sinais biomecânicos da marcha.

O MF-DFA apresentou comportamento distinto. Na análise da frequência foram identificados efeitos significativos da velocidade, da frequência e da interação entre velocidade e frequência, sem interações significativas envolvendo o grupo. Esses resultados indicam que os parâmetros multifractais foram influenciados principalmente pela resolução temporal e pelas condições de execução da tarefa.

A Figura 3 evidencia esse comportamento na condição de 100% PWS. Os três grupos apresentaram aumento progressivo dos valores do MF-DFA com o aumento da frequência, partindo de aproximadamente 0,97–0,98 em 30 Hz, aumentando para 1,26–1,28 em 60 Hz e atingindo aproximadamente 1,45–1,47 em 120 Hz. As trajetórias permaneceram próximas e aproximadamente paralelas, comportamento compatível com o efeito significativo da frequência e com a ausência de interação significativa entre frequência e grupo.

A alteração observada entre 30 e 60 Hz apresentou magnitude visualmente maior do que entre 60 e 120 Hz, sugerindo maior influência de reduções mais acentuadas da resolução temporal. Essa tendência deve, entretanto, ser interpretada de forma descritiva. A dependência do MF-DFA em relação à frequência pode estar associada à própria natureza do método, uma vez que a caracterização das flutuações em múltiplas escalas temporais depende da quantidade e da distribuição temporal dos pontos disponíveis na série, conforme os fundamentos apresentados por Kantelhardt et al. [3] e Ihlen [4].

Embora a Figura 3 represente especificamente 100% PWS, a interação significativa entre velocidade e frequência demonstra que a influência da resolução temporal sobre o MF-DFA também depende da condição de marcha. Dessa forma, os resultados obtidos nessa velocidade não devem ser generalizados diretamente para as demais condições, reforçando a necessidade de considerar conjuntamente as características do sinal e da tarefa biomecânica.

Na análise do ruído, o MF-DFA apresentou efeitos significativos para velocidade e ruído, além de interações entre velocidade e grupo, velocidade e ruído e entre velocidade, ruído e grupo. Esses resultados demonstram maior dependência das combinações entre os fatores experimentais. Essa sensibilidade pode estar relacionada à natureza multifractal do método, que caracteriza a organização das flutuações em múltiplas escalas temporais e, consequentemente, pode responder de forma mais pronunciada às alterações simultâneas na execução da tarefa e na qualidade do sinal, considerando as características do método descritas por Kantelhardt et al. [3] e Ihlen [4].

A maior sensibilidade observada para o MF-DFA não deve ser interpretada necessariamente como menor desempenho. Essa característica pode representar uma limitação quando o objetivo é obter uma métrica robusta às alterações de aquisição e processamento, mas também pode favorecer a identificação de modificações na organização temporal dos sinais. Assim, sensibilidade e robustez devem ser consideradas de acordo com a finalidade da análise.

A comparação entre os métodos demonstra que eles caracterizam propriedades distintas da dinâmica dos sinais. A Correlation Dimension está associada predominantemente à geometria do atrator reconstruído, conforme os fundamentos de Grassberger e Procaccia [1], enquanto o MF-DFA caracteriza a organização das flutuações em múltiplas escalas temporais, conforme descrito por Kantelhardt et al. [3] e Ihlen [4]. Dessa forma, as diferenças observadas frente à frequência, velocidade e ruído são compatíveis com as distintas propriedades avaliadas pelos algoritmos.

Do ponto de vista metodológico, os resultados reforçam que métricas fractais obtidas sob diferentes protocolos de aquisição devem ser comparadas com cautela. Alterações na frequência de amostragem e na qualidade do sinal podem modificar os valores estimados e interagir com as condições experimentais, tornando importante a padronização dos protocolos de aquisição e processamento. A escolha entre Correlation Dimension e MF-DFA deve, portanto, considerar a propriedade dinâmica que se pretende investigar e as características dos

sinais disponíveis.

Dessa forma, os métodos podem ser considerados complementares na caracterização da complexidade dos sinais biomecânicos. A utilização conjunta dessas abordagens permite associar informações relacionadas à geometria da dinâmica reconstruída às características temporais e multiescalares das flutuações, contribuindo para uma análise mais abrangente da dinâmica da marcha humana.

F. Agradecimentos

Os autores agradecem ao Programa de Pós-Graduação em Engenharia Biomédica da Universidade Federal de Uberlândia, ao Laboratório de Bioengenharia e Biomecânica da Universidade Federal de Goiás e às instituições de fomento que contribuíram para a realização desta pesquisa.

REFERÊNCIAS

[1] Grassberger, P.; Procaccia, I. Measuring the strangeness of strange attractors. Physica D, 1983.

[2] Takens, F. Detecting strange attractors in turbulence. Lecture Notes in Mathematics, 1981.

[3] Kantelhardt, J. W.; Zschiegner, S. A.; Koscielny-Bunde, E.; Havlin, S.; Bunde, A.; Stanley, H. E. Multifractal detrended fluctuation analysis of nonstationary time series. Physica A, 2002.

[4] Ihlen, E. A. F. Introduction to multifractal detrended fluctuation analysis in Matlab. Frontiers in Physiology, 2012.

[5] Mehdizadeh, S. A robust method to estimate the largest Lyapunov exponent of noisy signals. Journal of Biomechanics, 2019.

[6] Souza, G. S. S.; Andrade, A. O.; Pereira, A. A.; Vieira, M. F. On the use of Hilbert transform for gait local dynamic stability analysis. Research on Biomedical Engineering, 2023.

[7] Lopes, R.; Betrouni, N. Fractal and multifractal analysis: A review. Medical Image Analysis, 2009.

[8] Esteller, R.; Vachtsevanos, G.; Echauz, J.; Litt, B. A comparison of waveform fractal dimension algorithms. IEEE Transactions on Circuits and Systems, 2001.

Aplicação de Metaestruturas Reentrantes no Desenvolvimento e Caracterização Mecânica de Prótese de Pé e Tornozelo

Rodrigo S. N. Silva
Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo/SP, Brazil

Humberto Rossini dos Santos
Engenharia Mecânica
Centro Universitário FEI
São Bernadro do Campo/SP, Brazil

Leonardo Maciel
Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo/SP, Brazil

Douglas Rocha Lima
Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo/SP, Brazil

Matheus Frazão Vieira Molina
Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo/SP, Brazil

Mohammad Hossein Shaterzadeh Yazdi
Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo/SP, Brazil

Wellington Cassio Pinheiro
Engenharia Mecânica
Centro Universitário FEI
São Bernardo do Campo/SP, Brazil
wellington@fei.edu.br

***Resumo*—A ausência de membros inferiores decorrente de amputações transfemorais é uma das principais causas de redução da mobilidade funcional, comprometendo o equilíbrio e o conforto biomecânico. As próteses convencionais de baixo custo distribuídas na rede pública apresentam limitações severas no amortecimento inicial da marcha devido a rigidez excessiva e ao comportamento mecânico, resultando em assimetrias significativas. O projeto propõe o desenvolvimento de uma prótese de pé e tornozelo fundamentada no uso de metamateriais mecânicos e manufatura aditiva para otimizar a absorção de impacto. Através da parametrização de metaestrutura auxeticas reentrantes, projetou-se um dispositivo com gradientes de rigidez adaptáveis. Os resultados obtidos via simulação por elementos finitos e ensaios experimentais de compressão demonstraram que a geometria de 0,4mm suportou cargas de até 5800N. Os resultados do modelo revelaram uma correlação de 88,9% entre a força estimada (4098N) e a força experimenta (3643 N) para uma deformação de 15,6%. Concluiu-se que a precisão obtida e a viabilidade econômica (R$ 186,54) têm uma tendencia possível da eficácia da proposta na atenuação de assimetrias de marcha.**

***Index Terms*—Prótese Modular de Tornozelo e Pé, Marcha Humana, Manufatura Aditiva, Absorção de impacto, Caracterização Mecânica.**

I. Introdução

A perda de membros inferiores representa um desafio crítico de saúde pública no Brasil. Dados da Sociedade Brasileira de Angiologia e de Cirurgia Vascular (SBACV), extraídos do Sistema Único de Saúde (SUS), indicam a realização de mais de 282 mil amputações entre janeiro de 2012 e maio de 2023, comum uma média alarmante de 85 procedimentos diários em 2022 [1]. Este cenário exige tecnologia assistivas que restaurem a independência funcional e mitiguem o comprometimento da fase de apoio terminal e o retorno de energia ineficiente

As limitações das próteses rígidas convencionais residem na incapacidade de reproduzir a resposta dinâmica não linear do complexo pé-tornozelo natural. Tais dispositivos falham na absorção de impacto gerando sobrecargas compensatórios que resultam em assimetrias de marcha crônicas. Dispositivos de alto desempenho, embora eficazes, possuem custo proibitivo para a maior parte da população, especialmente para adolescentes em fase de crescimento e idosos.

Apesar do avanço da engenharia, a realidade tecnológica das próteses de membros inferiores para grande parte da população ainda é marcada por limitações especialmente os dispositivos de baixo custo distribuídos em sistemas públicos, no qual baseiam-se em designs que utilizam materiais tradicionais e estruturas passivas de alta rigidez. Em modelos clássicos o uso de núcleos rígidos é incapaz de reproduzir as variações dinâmicas não lineares do complexo pé-tornozelo humano ao longo das diferentes fases de marcha [2].

Essa característica de "tamanho único" ou rigidez padrão gera um problema crítico na ausência de ajustes individualizados para diversas características biomecânicas únicas de cada indivíduo como a diferença entre massa corporal, idade, estatura e necessidades personalizadas que agravam a adapta-

O presente trabalho foi realizado com o apoio da Financiadora de Estudos e Projetos (Finep) – Contrato nº 01.25.0458.00 (0339/24). Os autores agradecem à Finep pelo suporte financeiro e institucional que possibilitou a realização desta pesquisa.

bilidade especialmente do público adolescente e idoso. A falta dessa personalização compromete a estabilidade e a eficiência funcional, uma vez que o dispositivo não consegue mimetizar a absorção de impacto necessária no contato inicial [3].

Como consequência dessa baixa adaptabilidade, o usuário desenvolve assimetrias de marcha significativas. O uso de sistemas rígidos sem amortecimento dinâmico exige do corpo a realizar compensação musculoesqueléticas que resultam em dores crônicas e doenças degenerativas ao longo prazo [4]. Portanto, existe uma lacuna tecnológica entre os dispositivos acessíveis e a necessidade biomecânica de uma marcha simétrica e confortável, justificando a busca por novos materiais e geometrias customizáveis.

O objetivo deste trabalho é projetar e caracterizar mecanicamente em bancada uma sola protética de pé e tornozelo baseada em metaestrutura auxética reentrante fabricada por manufatura aditiva, avaliando a rigidez ajustável do perfil celular, a concordância entre a previsão numérica e o ensaio de compressão, e o custo de material da solução proposta. A avaliação biomecânica da marcha e a validação clínica com usuários não fazem parte do escopo deste estudo.

II. Metodologia

A. Fundamentação Biomecânica e Materiais

A locomoção Humana exige que o pé e o tornozelo atuem como um sistema dinâmico de dissipação e restituição de energia. Durante a fase de Loading Response (LR), é imperativo que a estrutura promova a atenuação de impacto para proteger o membro residual. Na fase de propulsão, a rigidez deve ser suficiente para a transmissão eficiente de carga.

A escolha da metaestrutura auxética reentrante fundamenta-se em seu coeficiente de Poisson negativo. Sob compressão vertical, as paredes celulares contraem-se lateralmente, resultando em um aumento da densidade local e maior resistência ao impacto. Para a construção modular, o TPU 95A foi selecionado para a sola devido à sua resiliência e capacidade de amortecimento, enquanto o PETG foi adotado nas partes estruturais por sua estabilidade dimensional e resistência mecânica superior. A interface entre os materiais foi projetada com um encaixe do tipo “cauda de andorinha” (dovetail), garantindo a integridade mecânica para interferência e influenciado diretamente as condições de contorno das análises estruturais.

Os carregamentos adotados nas análises numéricas e nos ensaios preliminares foram definidos com base em referencias biomecânicas da macha humana e nos princípios gerais da ISSO 10328:2916, norma destinada à avaliação estrutural de próteses de membros inferiores.

B. Modelagem Geométrica e Simulação Numérica

A célula unitária auxética foi parametrizada no software Siemens NX, permitindo a construção de uma sola com gradiente de rigidez. A simulação numérica via Finite Element Analysis (FEA) foi conduzida no Ansys Workbench 2025R2, utilizando o modelo Bilinear isotropic Hardening para representar a não linearidade constitutiva do TPU 95A, os parâmetros inseridos incluíram uma tensão de escoamento de 42 MPa e um módulo de elasticidade tangente de 960 MPa, conforme caracterização de sólidos celulares e o TPU 95A foi modelado com densidade de 1,24 g/cm³ e coeficiente de Poisson de 0,35 [7].

As condições de contorno foram estabelecidas com foco na precisão do comportamento. Carregamento: Aplicou-se uma carga de 50 N sobre um módulo representativo da metaestrutura conforme ilustrado na Figura 1 lado A, cuja área é de 135,412 mm², correspondente a uma pressão de 369 kPa. Esta carga corresponde a uma célula unitária que ocupa 1,22% da área total da sola, escalonando matematicamente para o suporte das cargas observadas no ensaio experimental de compressão. Ressalta-se que a carga simulada e a carga do ensaio não são diretamente equivalentes: a comparação apresentada na Seção III foi obtida por escalonamento proporcional de área, admitindo-se distribuição uniforme dos esforços entre as células — hipótese simplificadora que, dado o caráter não linear da geometria, não é rigorosamente válida e não foi verificada experimentalmente. Malha e Configurações: Utilizou-se um refinamento de malha de 0,8 mm e a ativação da opção Large Deflection, essencial para capturar as grandes deformações e a reconfiguração geométrica das paredes auxéticas antes do platô de densificação gerando a malha da Figura 1 lado B. Foi utilizado auto time stepping com 25 substeps iniciais, mínimo de 10 e máximo de 100. Tentativas de refinamento adicional da malha não atingiram convergência, de modo que não foi realizado estudo formal de convergência. Restrições: A face superior foi modelada como suporte fixo, simulando o acoplamento por interferência à estrutura principal.

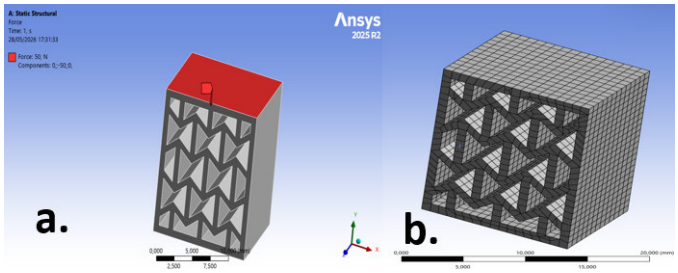


Figura 1. Configurações pré-simulação sendo o lado A - configuração das condições de contorno e o lado B - resultado da malha gerada pelo ANSYS.

C. Manufatura Aditiva e Parametros de Impressão

A fabricação foi realizada via FDM (Fused Deposition Modeling) utilizando uma impressora Bambu Lab A1 e P1S equipadas com bico de 0,4 mm. Foram selecionados dois materiais distintos, PETG para as partes estruturais (corpo principal, pé e tornozelo) pela tenacidade e estabilidade dimensional e a estrutura metamaterial da sola TPU 95A para a sola por sua elevada capacidade de recuperação elástica. Adotou-se altura de camada de 0,16 mm, com primeira camada de 0,20 mm e largura de linha entre 0,42 mm e 0,45 mm conforme a região da peça. A orientação das peças na mesa foi definida de modo a alinhar as camadas de deposição favorável

à direção predominante dos esforços, minimizando o risco de delaminação intercamadas.

- PETG: Bico extrusor a 223 °C e mesa a 80 °C, com 5 perímetros de parede para resistência mecânica e preenchimento tipo gyroid de 30% com ancoragem de 400% e comprimento máximo de âncora de 20 mm; seis camadas sólidas superiores em padrão monotônico; velocidades de 30 mm/s na parede externa, 80 mm/s na parede interna e 45 mm/s nas superfícies superiores.
- TPU: Bico extrusor a 250 °C e mesa a 60 °C com velocidade reduzida em 10 mm/s na primeira camada para assegurar a adesão interlaminar e precisão da célula auxética; paredes externa e interna a 70 mm/s e 80 mm/s; preenchimento a até 270 mm/s; limitação de vazão volumétrica em 3 mm³/s para reduzir instabilidades de extrusão típicas de materiais flexíveis. A impressão consumiu 156,9g de filamento em 12h 42min para o retropé.

D. Caracterização Mecânica Experimental

A caracterização mecânica foi conduzida por ensaios de compressão em máquina universal MTS 810. Foram produzidos três corpos de prova em TPU 95A, um para cada geometria avaliada — auxético de 0,4 mm, auxético de 0,6 mm e perfil combinado (Figura 2) —, todos fabricados com os parâmetros descritos na Seção II-C. As amostras, de 15 mm de altura, foram posicionadas entre as placas de compressão com alinhamento controlado, de modo a minimizar efeitos de carregamento excêntrico, e comprimidas até um deslocamento máximo de 8 mm com uma taxa de deslocamento de 8 mm/min. Adotou-se um corpo de prova por geometria, sem repetições; consequentemente, os valores reportados correspondem a observações individuais e não permitem estimativa de dispersão entre amostras. As amostras não foram submetidas a condicionamento controlado de umidade e temperatura previamente ao ensaio.

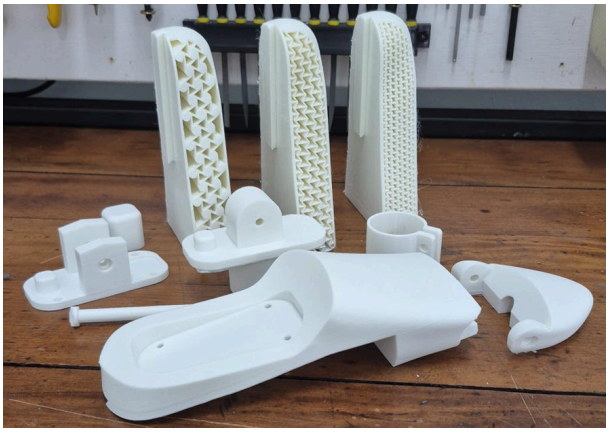


Figura 2. Prótotipo desmontado com todas as peças feitas em manufatura aditiva.

III. Resultados e discussão

Os ensaios experimentais em máquina MTS 810 caracterizam as respostas mecânicas dos três perfis distintos. Os dados obtidos estão sintetizados na Tabela I.

Tabela I
Resultado dos ensaios de compressão dos diferentes perfis.

Ensaio de compressão dos corpos de prova		
Perfil	Força Máxima (N)	Tensão Máxima (MPa)
Auxético 0,4 mm	∼5760	∼40
Auxético 0,6 mm	∼2450	∼17
Combinado	∼2800	∼19

O perfil auxético de 0,4 mm demonstrou o maior suporte de carga, suportando uma força máxima de aproximadamente 5760 N e com uma tensão máxima de 40 MPa. Em contrapartida, o perfil de 0,6 mm e combinado apresentaram um desempenho intermediário quando comparado ao perfil de 0,4 mm. Observou-se um comportamento não linear nas curvas de força-deslocamento, típico de sólidos celulares [8] com aumento progressivo da inclinação associado à densificação das células (Figura 3), sem colapso estrutural abrupto. Ressalta-se que os ensaios foram de compressão uniaxial sobre perfis isolados e não reproduzem as condições de carregamento combinado previstas nos ensaios estáticos da ISO 10328, referência que deve ser adotada em etapa posterior para qualificação do dispositivo montado.

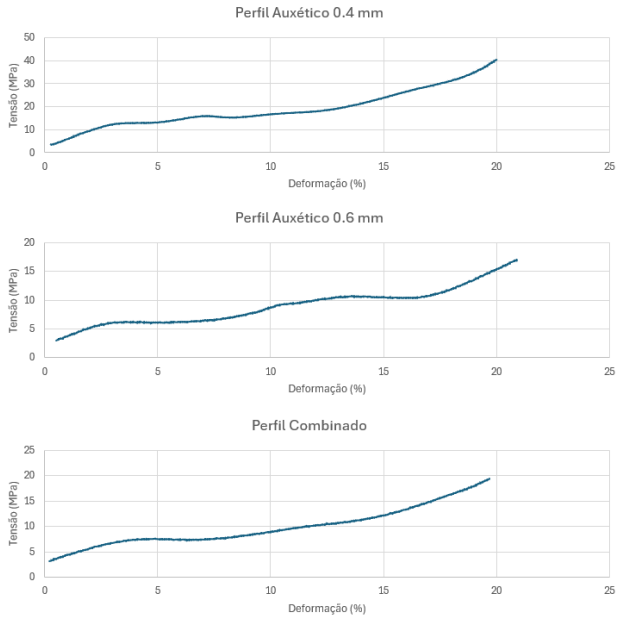


Figura 3. Curva tensão x deformação dos 3 peris auxeticos ensaiados e zonas de platô.

A comparação entre o modelo numérico e o ensaio experimental revelou que a simulação previu uma força de 4098 N para uma deformação de 15,6%, enquanto o ensaio físico

registrou 3643 N para o mesmo nível de deformação. Este resultado tem uma correlação de 88,9%, demonstrando o possível uso de elementos finitos para a predição do desempenho de metaestruturas impressas em 3D utilizando a tecnologia de impressão por Modelagem por Deposição Fundida (FDM) utilizando o modelo Bilienar Isotropic Hardening para prever o desempenho do protótipo.

A simulação ainda demonstrou que a estrutura auxética reentrante em TPU possui elevada capacidade de deformação controlada. Sob uma carga de 50N, o modelo apresentou um deslocamento máximo de 2,34 mm, o que apresenta 15,6% da altura total da amostra. A análise da energia de deformação atingiu 0,0419 mJ ao final da simulação, indicando a capacidade da geometria em armazenar e redistribuir energia.

A. Viabilidade Economica e Implicações Clinicas

O custo direto de fabricação do protótipo foi quantificado em R$ 187,66, decomposto em R$ 51,05 de materiais (0,157 kg de TPU a R$ 150,00/kg e 0,344 kg de PETG a R$ 80,00/kg), R$ 6,61 de energia elétrica (7,34 kWh a R$ 0,90/kWh) e R$ 130,00 de componentes comerciais (mola de inox do tornozelo e duas molas de torção dos dedos), a preços de mercado no período de execução do trabalho. Esse valor não inclui mão de obra, depreciação do equipamento, pós-processamento ou adaptação individualizada, itens que compõem o custo total de disponibilização do dispositivo ao usuário. Registra-se ainda que a mola inicialmente projetada em aço AISI 304 não atendeu ao critério de fadiga de Goodman (1,608 contra o limite de 0,77) para os dados de entrada adotados — força entre 500 N e 1000 N, deflexão de 15 mm e 3×10^6 ciclos conforme a ISO 10328; o projeto foi, portanto, alterado para aço SAE 1070, resultando em fator de segurança de 1,5 sob carregamento cíclico.

Este baixo custo de produção possui implicações clínicas diretas para o cenário brasileiro. A acessibilidade financeira é fundamental para adolescentes em fase de crescimento, que necessitam de substituições frequentes da prótese devido a alterações antropométricas. A modularidade e o baixo custo permitem que o sistema de saúde forneça dispositivos com amortecimento dinâmico a uma fração do preço das próteses de fibra de carbono importadas.

Ao comparar as outras opções no mercado, o conceito tecnológico apresentado evidencia a vantagem no custo-benefício conforme a Tabela II.

B. Comparativo com o Estado da Arte: Pé SACH

O pé SACH (Solid Ankle Cushion Heel) é utilizado como padrão de referência utilizado amplamente pelo SUS tendo sua construção com núcleo de madeira rígido e revestimento em poliuretano conforme ilustrado na Figura 4, sem qualquer articulação mecânica ou mecanismo de retorno de energia. Sua rigidez linear e a ausência de comportamento não-linear de amortecimento tornam-no incapaz de mimetizar a resposta dinâmica do complexo pé-tornozelo humano ao longo da fase da marcha [5].

Tabela II
COMPARATIVO ENTRE O CUSTO DE MATERIAL E COMPONENTES DO PROTÓTIPO E PREÇOS DE REFERÊNCIA DE DISPOSITIVOS COMERCIAIS (VALORES EM R$, BASE DE PREÇOS DO PERÍODO DE EXECUÇÃO DO TRABALHO; NÃO INCLUEM SERVIÇOS DE ADAPTAÇÃO).

Dispositivo	**Custo Estimado**	**Tipo de Amortecimento**	**Personalização**
Pé SACH convencional (SUS)	R$ 269,50 (reposição SUS) / R$ 200 ∼ R$380 (varejo)	Passivo, linear, núcleo rígido	Não
Pé SACH Ottobock 1S66 (importado)	R$ 1.299,09	Passivo, linear	Não
Próteses em fibra de carbono (ESR)	R$ 3.000 ∼ 15.000	Retorno de energia	Parcial
Protótipo (TPU/PETG)	R$ 187,66	Não linear, auxético, gradiente de rigidez	Sim

Do ponto de vista biomecânico, o pé SACH apresenta assimetria interlaterais significativas durante o ciclo de marcha, com redução da força de reação vertical no lado protetizado e incapacidade de simular a flexão plantar natural.

Contrapondo essa deficiência do pé de SACH, o protótipo apresentado, por meio do gradiente de rigidez das metaestruturas auxéticas reentrantes, oferece uma resposta mecânica não linear (Figura 3) que mimetiza a densificação progressiva do tecido adiposo plantar humano, potencialmente podendo contribuir na eventual redução de assimetrias de marcha. Ressalta-se que essa comparação é qualitativa e apoiada em dados de literatura, sendo necessário o estudo e ensaios clínicos para validar a hipótese acima posteriormente.

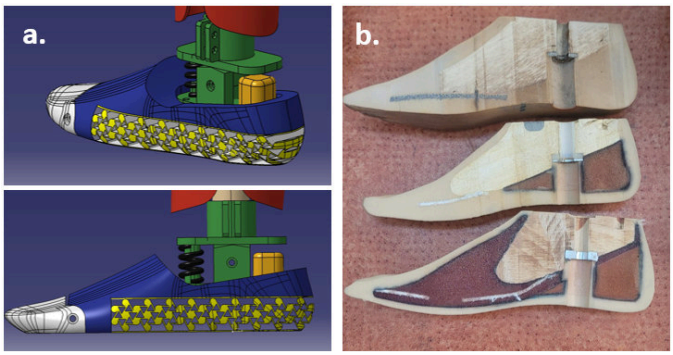


Figura 4. Comparação entre o prototipo (A) e o pé de SACH (B).

O pé SACH convencional não oferece mecanismo de ajuste de rigidez e por conta disso adolescentes em fase de crescimento tendem a mudar a sua massa corporal e biomecânica [6]. Sendo assim, a manufatura aditiva via FDM permite que o perfil auxético seja reimpresso e substituído com custo margi-

nal irrisório comparado aos custos atuais, aderindo diretamente as necessidades do sistema público brasileiro e ocupando essa lacuna tecnologia.

C. Limitações do Estudo

Este trabalho limita-se à caracterização mecânica em bancada de perfis celulares isolados, com um corpo de prova por geometria e sem repetições, o que impede estimativa de dispersão. As propriedades do TPU 95A empregadas no modelo numérico provêm de literatura e de ficha técnica, sem caracterização em tração do lote impresso. O refinamento adicional da malha não convergiu, de modo que não há estudo formal de convergência. O modelo de contato entre paredes celulares foi definido sem atrito e é válido apenas até o instante do primeiro toque entre elementos, não representando o deslizamento, as interferências adicionais nem a densificação subsequente. A comparação entre modelo e ensaio apoia-se em escalonamento proporcional de área, hipótese não válida em rigor para uma geometria de resposta não linear. Não foram realizados: condicionamento controlado das amostras; medição de armazenamento e retorno de energia; ensaios de fadiga e de carregamento cíclico; avaliação de envelhecimento ambiental e degradação polimérica; análise instrumentada de marcha, medição de pressão plantar ou avaliação funcional com usuários; nem ensaios comparativos diretos, sob condições equivalentes, com pés SACH ou com dispositivos de armazenamento e retorno de energia. Em decorrência dessas limitações as implicações clínicas discutidas devem ser entendidas como hipóteses de projeto.

IV. Conclusão

Os resultados obtidos indicam a viabilidade técnica de uma sola protética de pé e tornozelo baseada em metaestrutura auxética reentrante em TPU/PETG fabricada por manufatura aditiva, caracterizada em condições de bancada. Os perfis ensaiados apresentaram comportamento não linear com região de platô e densificação, característico de sólidos celulares, e rigidez ajustável pela espessura de parede da célula unitária, o que constitui a principal contribuição do trabalho: a caracterização experimental de uma geometria celular ajustável com potencial de aplicação em próteses de baixos custos.

Os principais achados técnicos destacam: Desempenho Estrutural: O perfil auxético de 0,4 mm demonstrou ser configuração mais eficiente, suportando cargas de até 5.760 N com uma tensão de 40 Mpa, superando as exigências típicas de carga para o ciclo de marcha humana sem apresentar deformação permanente. Concordância do Modelo: o desvio de 11,1% observado entre a simulação por elementos finitos e o ensaio na máquina MTS 810, em um único ponto de comparação, sugere que o modelo Bilinear Isotropic Hardening representa adequadamente a ordem de grandeza da resposta do TPU 95A na geometria estudada. A confirmação dessa capacidade preditiva depende da comparação de curvas completas para as diferentes geometrias, com repetições, e não é estabelecida por este resultado isolado.

Do ponto de vista socioeconômico, o protótipo apresentou um custo estimado de R$ 187,66 reais, valor drasticamente inferior às próteses comerciais de alto desempenho. Essa acessibilidade é fundamental para o cenário brasileiro e para o Sistema Único de Saúde (SUS), especialmente para o público adolescente, que demanda substituições frequentes devido ao crescimento antropométrico.

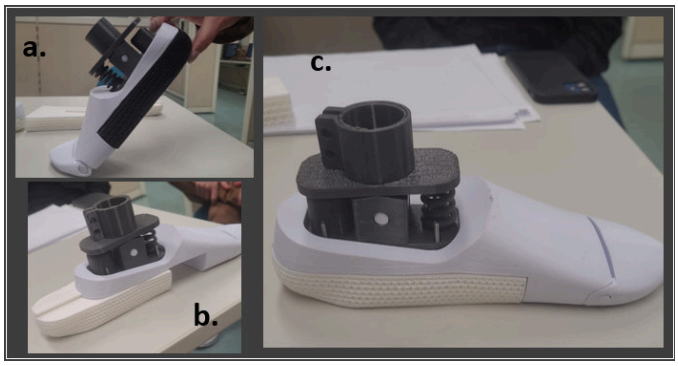


Figura 5. Prototipo em 3 vista sendo (A) demonstração do movimento da frente do pé, (B) subistituição do perfil auxetico, (C) prototipo completo.

Trabalhos futuros devem contemplar a caracterização em tração dos materiais efetivamente impressos, o estudo de convergência de malha e a comparação de curvas força-deslocamento pareadas para todas as geometrias, com repetições e reporte de erro e de modos de falha; a medição de armazenamento e retorno de energia; e a construção de bancada específica para os ensaios previstos na ISO 10328, incluindo fadiga dinâmica, visando à qualificação para uso clínico prolongado, além da análise da degradação polimérica sob exposição ambiental. Concluída a etapa de bancada e montado o dispositivo completo, a avaliação com usuários exigirá ensaios de segurança, aprovação por comitê de ética em pesquisa e adaptação individualizada por profissional habilitado, permitindo então a análise instrumentada da marcha, a medição de pressão plantar e a comparação funcional direta com pés protéticos comercialmente disponíveis, incluindo SACH e dispositivos ESR.

Uso de IAG

Em conformidade com a Portaria CNPq n° 2.664/2026, os autores declaram o uso da ferramentas de IAG nas seguintes etapas e finalidades: revisão de redação e padronização do estilo do texto (GPT-5.6, Notebook LM e Perplexity AI); busca de artigos para a revisão de literatura (Claude, Gemini e Sonar 2). A ferramenta não foi utilizada para gerar dados, resultados ou conclusões; não é listada como autora; e todo o conteúdo produzido com seu auxílio foi verificado e revisado pelos autores, que se responsabilizam integralmente pelo conteúdo final.

Referências

[1] Agência Brasil. “Amputações de pés e pernas em decorrência do diabetes batem recorde.” Agência Brasil, 19 set. 2023. Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2023-09/

amputacoes-de-pes-e-pernas-em-decorrencia-do-diabetes-batem-recorde. Acesso em: 10 set. 2025.

[2] B. J. Hafner, J. E. Sanders, J. M. Czerniecki, and J. Fergason, "Transtibial energy-storage-and-return prosthetic devices: A review of energy concepts and a proposed nomenclature," *J. Rehabil. Res. Dev.*, vol. 39, no. 1, pp. 1–11, 2002.

[3] WHITTLE, M. W. *Gait analysis: an introduction*. 4. ed. Edinburgh: Butterworth-Heinemann, 2007.

[4] PERRY, J.; BURNFIELD, J. M. *Gait analysis: normal and pathological function*. 2. ed. Thorofare: Slack Incorporated, 2010.

[5] BARR, A. E. et al. Biomechanical comparison of the energy-storing capabilities of SACH and Carbon Copy II prosthetic feet during the stance phase of gait in a person with below-knee amputation. *Physical Therapy*, v. 72, n. 5, p. 344–354, 1992.

[6] ABRÃO, J. S. *Design parametrizado em manufatura aditiva de pé protético modular com filamento de compósito polimérico reforçado com fibras descontínuas de carbono*. 2025. 199 f. Tese (Doutorado em Engenharia Mecânica) – Universidade Federal de Uberlândia, Uberlândia, 2025.

[7] SAMANTA, T.; QAIMI, M. G. S. Experimental and FEA analysis of 3D printed TPU material auxetic structure. *International Journal of Emerging Technologies and Innovative Research (JETIR)*, v. 7, n. 8, p. 279–287, ago. 2020. ISSN 2349-5162. Disponível em: https://www.jetir.org/view?paper=JETIR2008035.

[8] GIBSON, L. J.; ASHBY, M. F. *Cellular solids: structure and properties*. 2. ed. Cambridge: Cambridge University Press, 1997.

Comparison between Two PWM signal generation architectures for servomotors control in pediatric Socially Assistive Robots

Marcelo do Carmo Camargo Gaiotto
Graduate Program on Health Technology
Pontifícia Universidade Católica do Paraná
Curitiba, Paraná
0009-0006-9671-1350

Djovana Petrikic
Graduate Program on Health Technology
Pontifícia Universidade Católica do Paraná
Curitiba, Paraná
0009-0005-6905-4108

Percy Nohama
Graduate Program on Health Technology
Pontifícia Universidade Católica do Paraná
Curitiba, Paraná
0000-0002-8051-8453

***Abstract*— Simultaneous control of multiple servomotors is a challenge in the development of Socially Assistive Robots (SAR) with a high number of degrees of freedom. In this work, we compare two PWM architectures for servo control in pediatric SAR: (1) the software-based Time-Division Multiplexing (TDM) technique, implemented and validated in hospital field trials on the TIGRÃO robot since 2005, and (2) the dedicated PCA9685 integrated circuit, evaluated for the BORIS robot project. Timing analysis was conducted using an 8-channel logic analyzer at 24 MHz, comparing both approaches under equivalent operating frequency conditions (33 Hz). The results demonstrate that the actual TDM update frequency is 33 Hz, differing from the nominal value of 50 Hz, without compromising the robot's expressive performance. The effective resolution analysis, performed at the 33 Hz test frequency, reveals that TDM operates with 180 steps entirely within the servo's useful range (0.5–2.5 ms), while the PCA9685 uses approximately 270 of the 4096 available steps at 33 Hz, an effective resolution gain of 1.5 times; at the 50 Hz frequency planned for BORIS's servo operation this gain rises to 2.3 times (410 effective steps). The use of multiple PCA9685 units requires an external oscillator to ensure synchronization between ICs. Both solutions demonstrated feasibility for SAR applications, with PCA9685 being the chosen solution for BORIS as it fully frees the processor for other critical tasks.**

Keywords—socially assistive robotics; servomotors; PWM; time-division multiplexing; PCA9685; pediatric SAR; HRI

I. INTRODUCTION

The development of socially assistive robots has grown continuously in order to support different groups, including the elderly [1] hospitalized children [2] e and individuals with autism spectrum disorder (ASD) [3], presenting designs and human-robot interaction (HRI) [4] that range from anthropomorphic [5], to abstract and zoomorphic [6], under autonomous [7], pre-programmed [8] or telepresence [9] control strategies. Among these, the Wizard of Oz (WOZ) method stands out in clinical contexts by allowing the specialist to control the robot covertly and in real time [10]; this approach simulates autonomous social intelligence for users and ensures that responses are strictly empathetic, dynamic, and adapted to the current limitations of purely autonomous systems [6, 11].

In 1993 in Japan, the PARO robot was developed [12], inspired by a harp seal pup, with the purpose of providing therapeutic support to the elderly and individuals with cognitive impairment; it began to be commercialized in 2005, and in 2009 it was certified by the Food and Drug Administration (FDA) as a "biofeedback medical device." The robot was designed to be soft and cuddly, incorporating servomotors to control head, flipper, and eye movements, enabling the simulation of realistic emotional responses, favoring the establishment of affective bonds, and promoting user well-being [12].

In 2004, Aldebaran Robotics launched NAO, a medium-sized, autonomous, and programmable humanoid robot. Since then, NAO has been used in various research areas, including SAR with children with ASD. However, it is a platform that requires high investment, and its programming is not trivial [13].

Huggable was developed at the MIT Media Lab with the aim of socially interacting with children in hospital settings, and can be used for pain distraction, entertainment, and to inform them about upcoming medical procedures. The robot has the appearance of a cute and friendly teddy bear. Its structure comprises 12 degrees of freedom, enabling expressive and animated movements. It is also equipped with capacitive sensors, which allow it to respond to tactile stimuli, maintaining natural interaction [14].

TIGRÃO robot project [15] began development in 2005, becoming fully functional with its 37 servomotors (DoFs), executing movements from pre-recorded positions by the end of 2006. Its mission as a SAR began in 2010, when its WOZ capability was expanded with the integration of a joystick for head and mouth control, demonstrating its effectiveness as a SAR as a supportive playful therapeutic technology, aligning with global efforts to mitigate the stress and fear inherent to pediatric hospitalization [16], alidating the potential of SAR as a provider of well-being in highly sensitive clinical environments.

TIGRÃO robot was developed with the core concept that it should be simple to explain and reproduce. Thus, to explain its electronic structure, we use a description tool known as the Hardware Diagram, shown in Fig. 1, in which at the center, a PIC18F452I/P microcontroller operates at 20 MHz. It receives signals from five buttons/switches for quick access to poses stored in memory, three analog signals from the joystick, connection to external pose-storage memory via I^2C bus, USART serial communication signals for WOZ operator control, the internal PWM output for controlling the illumination of a frosted acrylic panel with red LEDs placed on the robot's cheek, and has twelve digital connections to the PWM signal multiplexing block [17], through which the software manages the PWM outputs, operating under Time-Division Multiplexing (TDM) [18].

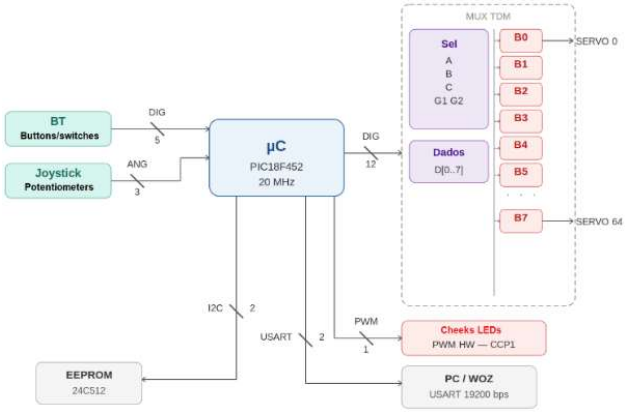


Fig. 1. Hardware Diagram of the TIGRÃO Robot

Since TIGRÃO robot requires at least 37 independent PWM signals (37 degrees of freedom - DoFs) to move its lower and upper limbs and all mobile parts of the head, it was necessary to expand the microcontroller's pin count. This task was accomplished through the TDM time-division I/O multiplexing circuit, increasing the physical capacity of digital signal outputs, mainly the PWM signals for the servos. The TIGRÃO mechanical structure was entirely built from flat, laser-cut aluminum sheets, with brass bushings and pins, resulting in an approximate weight of 13 kg, difficult to support even with the 20 kg HS805 motors used at that time, so it was decided to mount it onto an aluminum profile frame.

Inspired by the TIGRÃO robot, development of a new SAR began in 2026: BORIS (Buddy Operated Robot for Imagination and Stories), with at least 29 degrees of freedom (DoFs), in addition to sensors and haptic interfaces. It has the characteristics of a Siberian Husky dog and is being developed to be cute and huggable, intended primarily for use with hospitalized children during storytelling. It builds on almost 20 years of lessons learned from the TIGRÃO project, and its purpose is to go beyond a socio-emotional comfort tool and become an active physiological regulation agent.

BORIS robot is being fully designed to be built via 3D printing, which will facilitate its dissemination as a SAR development platform. Its electronic structure consists of five specialized microprocessor-based NODES, as shown in the robot's block diagram in Fig. 2. All of these modules are internal and communicate via a high-speed RS485 backbone (BORIS Bus daisy chain) and an RF network, both high-speed [19]. Some of the nodes have an SD card for synchronized data storage, allowing data retrieval and subsequent scientific evaluation. The microprocessors used in the NODES can be swapped for other models in the event of new research or the development of new models.

The microprocessors chosen for the BORIS robot project have two cores, and the primary strategy is to offload these cores by using as many hardware resources as possible. This created the need to replace software-based PWM signal generation and the previously developed I/O multiplexing structure with a dedicated hardware architecture. Accordingly, the first point considered was the possibility of using the microprocessor's internal hardware resources, which provide sixteen available PWM signals; however, this would limit the possibility of swapping microprocessors, would leave insufficient GPIO pins available on some modules, and would hinder future expansions. Therefore, the use of the PCA9685 integrated circuit was considered, which was originally designed to control sixteen PWM outputs for LEDs via I^2C, with a 12-bit resolution (4096 steps) and a guaranteed constant frequency across all sixteen outputs (24 Hz to 1526 Hz).

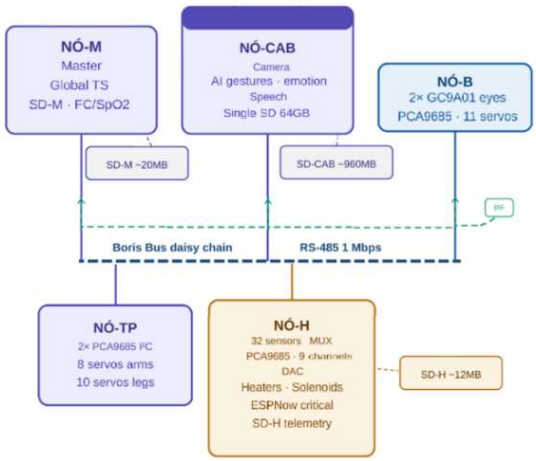


Fig. 2. Diagram of the BORIS Robot NODES.

The second point, in addition to I/O capacity and its respective load considerations, is that a SAR requires the latency between position changes of the facial servomotors to be less than 100 to 150 ms. This technical requirement is justified by the temporal dynamics with which the human brain decodes canine emotional expressions. Neurophysiological studies show that high-amplitude responses to dog emotions (such as aggressiveness or happiness) occur within a window of 100 ms to 150 ms [20]. Mechanical delays exceeding this neurological window would result in lethargic, unnatural behavior that breaks the consistency of realism and pushes the perception of the robot into the uncanny valley, an effect that is especially severe for moving entities [21].

Therefore, the efficient generation of PWM signals to control and synchronize multiple servomotors and haptic actuators remains a significant hardware challenge in the development of socially assistive robots (SAR). To address this challenge and ensure that the new platform meets real-time interaction requirements, the study described in this paper aims to compare two servomotor control techniques: software-based Time Division Multiplexing (TDM), currently used in the TIGRÃO robot, and the dedicated PCA9685 integrated circuit, proposed for the BORIS robot. This comparison is necessary to validate whether the dedicated integrated circuit can ensure the critical requirements of sub-150 ms latency and synchronization, without overloading the main microprocessors, and also to identify the most efficient approach that will determine BORIS's electronic architecture, ensuring the robot can deliver realistic, immediate emotional responses required for effective clinical intervention.

II. METHODOLOGY

To investigate the best PWM signal generation technique for the BORIS robot, we defined a strategy to temporally compare the two hardware solutions chosen for this purpose, namely the discrete I/O multiplexer and the dedicated PCA9685 integrated circuit. Since these are two very distinct

hardware solutions with very different techniques, we used a sequence of information common to both: the time between sending commands and the start of PWM signal generation at the outputs.

Another important analysis concerns the servomotor models used in the TIGRÃO robot project and those intended for use in the BORIS project, one being analog and the other digital, since each may have a different response in interpreting the PWM signals received.

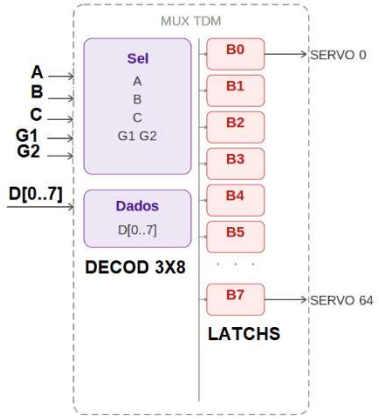


Fig. 3. PWM Output Multiplexing Board Circuit in block diagram.

To analyze the signals, an 8-channel logic analyzer with USB interface and a sampling frequency of 24 MHz was used, together with the Logic 2 software, which enables measurement of digital signals as well as monitoring and decoding of the I²C bus of the PCA9685. In addition to the logic analyzer, adapters were developed to connect the bank outputs of the TIGRÃO robot multiplexing board to the protoboard in order to facilitate connections, considering that the measurements do not require the servomotors to be in operation, nor any intervention with operator control or activation of a demonstration mode, since the outputs are constantly updated.

For the PCA9685 hardware test, four identical off-the-shelf Shield boards were used (Fig. 4.C), unmodified and mounted on the same protoboard as the TIGRÃO robot signal adapters, with I²C addresses configured from 0x40 to 0x43. The clock (SCL) and data (SDA) signals were connected in parallel and routed to an ARDUINO UNO. The input channels of the logic analyzer were connected to the selection signals and to the multiplexer bank outputs; it is not necessary to connect the SEL_G signal, since it is set to a high level as soon as the PIC18F452 microcontroller is configured, and the SEL_GA and SEL_GB signals remain continuously enabled in hardware, as shown in Fig. 3. TABLE 1 presents the configuration adopted for connecting the signals to the logic analyzer channels.

In the first test, the goal was to obtain the output signals from the TIGRÃO multiplexer, verifying, from the binary combinations of the selection signals, the moment at which the bank outputs begin their PWM cycles, connecting the logic analyzer channels according to TABLE 1. Once the signal acquisitions were performed, the interval between the configuration of the banks by the MUX selection signals and the start of the PWM cycle of the related output was measured.

TABLE 1. LOGIC ANALYZER CONNECTIONS.

LOGIC ANALYZER	MUX	PCA9685
SEL_0	SEL_A	SCL
SEL_1	SEL_B	SDA
SEL_2	SEL_C	0x40 – Servo 1
CHANNEL 3	Bank 0 – SR0	0x40 – Servo 9
CHANNEL 4	Bank 1 – SR8	0x41 – Servo 1
CHANNEL 5	Bank 2 – SR16	0x41 – Servo 9
CHANNEL 6	Bank 3 – SR14	0x42 – Servo 1
CHANNEL 7	Bank 4 – SR32	0x42 – Servo 9

After the first test was completed, the connector at the logic analyzer inputs was repositioned, now connecting the I²C bus and the outputs of the Shield boards to the input channels. Firmware versions for the ARDUINO UNO, which controls the PCA9685 units via the I²C bus, was recorded so as to simulate the same 8-output bank behavior as the TDM in each I²C transaction; that is, the PCAs was configured so that writing the position value of the 16 servos on each board is loaded in two transmissions to the same slave address on the I²C bus, first for outputs 1 to 8 and then 9 to 16, simulating the effect of loading one bank of 8 servomotors at a time, as in the multiplexer.

The test bench integrating the TIGRÃO robot with its PWM output multiplexing board, its connection cables to the protoboard, the PCA9685 Shield boards, and the logic analyzer are shown in Fig. 4.

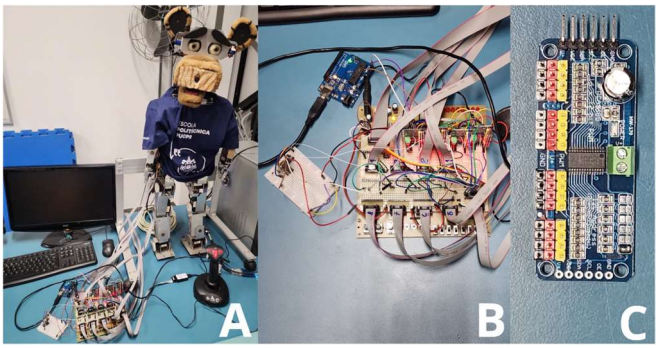


Fig. 4. Protoboard test setup for PWM signal evaluation. A. overall view; B. protoboard view; C. view of the PCA9685 board.

Initially, a write test was performed on all outputs of each PCA to confirm that everything is functioning correctly, and only afterward will signal acquisition begin, following the same timing measurement procedure used for TDM signal capture. In the case of the PCA9685, time measurement starts at the START of the I2C transmission frame and ends at the start of the PWM cycle of the related output. Since the PCA9685 allows all 16 PWM outputs to be sent in a single I2C transaction, the ARDUINO UNO firmware was modified for this test.

To perform the tests, a sequence was created that first resets all outputs of all PCA9685 units to zero, and after a delay, loads values for 8 and 16 outputs, depending on the test being performed. The signal capture procedure was carried out with a change in the connections between the logic analyzer and the PCA9685 Shield boards, following the configuration in TABLE 2.

TABLE 2. LOGIC ANALYZER AND PCA9685 – SINGLE 16-CHANNEL FRAME

LOGIC ANALYZER	PCA9685
SCL	SCL
SDA	SDA
CHANNEL 2	0x40 – Servo 1
CHANNEL 3	0x41 – Servo 1
CHANNEL 4	0x42 – Servo 1
CHANNEL 5	0x43 – Servo1
CHANNEL 6	Not used
CHANNEL 7	Not used

III. RESULTS

After collecting the signals from the TIGRÃO robot PWM multiplexing board, it was confirmed that the values implemented in its time-division multiplexing PWM generation technique are highly efficient. Although the multiplexer hardware has eight 8-bit banks (64 PWM outputs), the firmware of the PIC18F452 microcontroller limits activation to six banks, resulting in 48 updated PWM outputs, although only 37 are currently in use.

Another verification concerned the frequency of the generated PWM signal, which is 33 Hz, differing from the classic value typically used for servomotor control, normally 50 Hz. Since the TIGRÃO robot servomotors are analog, this frequency difference surprisingly causes no undesirable instability effect in their operation, allowing us to infer that they are more tolerant. Based on this observation regarding PWM frequency, and to maintain rigor in the analyses, the frequency of the PCA9685 units was changed to 33 Hz.

From the standpoint of the servomotors used in TIGRÃO robot (analog), each output bank's eight pins are set high as soon as the bank is selected as output within the ISR, followed by a fixed initial window (1000 to 1160 µs, depending on the bank) during which digital and, where applicable, analog inputs are read; this window is intentionally dimensioned in firmware so that its duration remains consistent across banks despite their different processing loads. Only after this window does the 180-step counting loop begin, with a fixed 4 µs resolution per step (720 µs total), during which each output individually falls to a low level once the step count exceeds its programmed position — this transition is what encodes the commanded pulse width. This 4 µs step resolution was dimensioned for the servos available in 2005 and safely exceeds the electromechanical deadband of every evaluated servo model (1 to 10 µs, depending on technology and era, TABLE 3), ensuring that the smallest programmable position change is always detected. An analog servo with an 8–10 µs deadband does not react to pulse variations smaller than this threshold, regardless of the controller's resolution, as shown in TABLE 3.

TABLE 3. SERVOMOTORS, DEADBAND, AND RESOLUTION

Model	YEAR	Technology	Deadband	Actual resolution
Hobbico CS-60	2005	Classic analog	8–10 µs	1,5°–2,0°/step
Hitec HS-805BB	2005	High-power analog	8 µs	0,8°–1,5°/step
SG90 TowerPro	2005	Micro analog	5 µs	0,8°–1,2°/step
MG90S	2026	Modern analog	5 µs	0,5°–1,0°/step
MG996R	2026	High-frequency digital	1–2 µs	0,44°/step (PCA bottleneck)

Timing measurements for the TDM technique were validated across n = 1435 complete bank cycles captured from a single continuous 44.25 s logic-analyzer session on the robot's original hardware and firmware, in use since the beginning of this study, summarized in TABLE 4 ("Boca ON/OFF" refers to whether the head-joystick analog reading was active). Banks without active analog input handling showed standard deviations (SD) on the order of tens of nanoseconds across all repetitions, indicating deterministic, hardware-level timing. Banks with conditional analog input reading showed two distinct, individually low-variance timing states corresponding to whether the joystick control was active, rather than measurement noise.

TABLE 4. BANK TIMING STATISTICS (SEL_G PULSE WIDTH, N = 1435 CYCLES)

Bank	State	n	Mean ms)	SD (ms)
0	Boca OFF	959	3.6231	0.01263
0	Boca ON	477	3.6199	0.00031
1	Boca OFF	959	3.6660	0.00002
1	Boca ON	477	3.6970	0.00821
2	dormant	1435	3.6662	0.00002
3	dormant	1435	3.6662	0.00002
4	no AD	1435	3.6666	0.00001
5	no AD	1435	3.6666	0.00001

Fig. 5 illustrates the actual behavior of the TDM technique operating on the TIGRÃO robot hardware, where it is possible to observe the logic selection signals Sel_0, Sel_1, and Sel_2 activated in sequence, and for each of their combinations, banks BANK0 through BANK4 being selected. The logic analyzer inputs were positioned on the first output bit of each bank (Q1 of the 74LS573), considering that the update of the 8 bits of the banks occurs simultaneously, since they are parallel registers; thus, it is possible to verify the moment at which the PWM signal begins its duty cycle at a high level in each bank. The markers from the start of a PWM pulse in one bank to the start of the subsequent bank indicate the delay time between the outputs of one bank and the next, which is 5 ms, the exact time of the timer interrupt call, and considering that there are six banks, the resulting period is 30 ms.

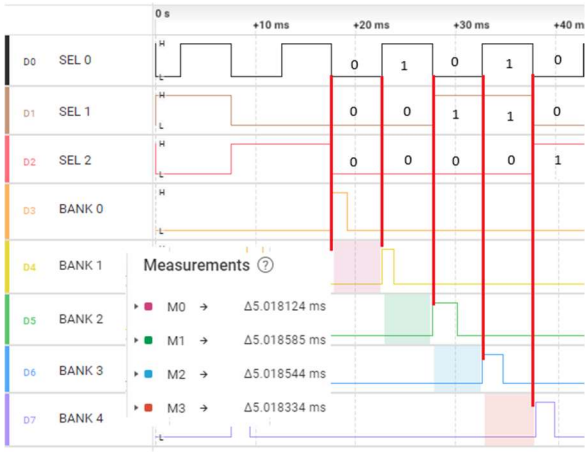


Fig. 5. Signal obtained from the logic analyzer for the TDM technique.

During the preliminary tests with the PCA9685 units, each with its own address and loading outputs in the classic manner, i.e., one output per I²C transaction, and using their internal clock generators, the synchronization problem between the outputs of one PCA relative to another became clearly evident, in addition to the variation in output frequency of each unit, as shown in Fig. 6. To mitigate the frequency variation problem, an external oscillator was developed.

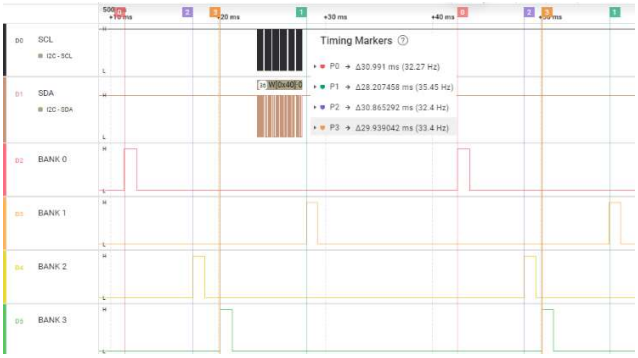


Fig. 6. Outputs of the Shield boards with internal oscillator and classic loading.

Since the Shield boards have the PCA9685 external clock input pin connected to GND through a via on the PCB, an intervention was performed using a drill bit to isolate the via from the GND plane on the bottom side of the board, and a wire was soldered to connect the external clock input pin to the developed clock generator. The external clock generator is a crystal oscillator with a frequency of 11.0592 MHz, based on inverting logic gates. This frequency value was chosen because it allows the best division ratio for the PCA outputs to operate at a precise PWM frequency of 50 Hz for the project; in this study, it was configured to operate at 33 Hz. The implemented oscillator circuit is shown on Fig. 7.

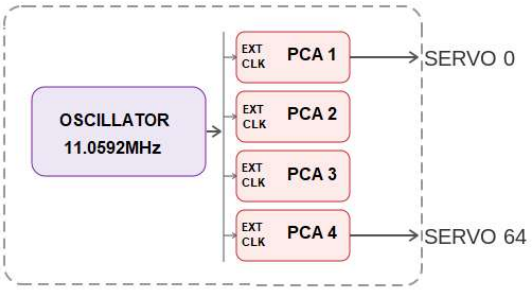


Fig. 7. External oscillator circuit and connection of the PCA9685 units in block diagram.

After implementing the clock generator circuit and confirming its operation together with the Shield board modifications and their connections to the external clock input pin, all PCA9685 units passed the individual basic tests and were connected to operate in a synchronized manner, allowing the acquisitions to be performed again.

A very important piece of information about PCA9685 concerns the update of the duty cycle of its outputs. Once a new value is loaded into the output registers, it only becomes available at the outputs after two conditions are met: the arrival of the STOP signal of the I²C transaction related to its address, and the next start of the PWM cycle. The times related to the update of the PCA9685 outputs after the arrival of the STOP signal of its I²C transaction were around 8 µs, and this procedure occurs regardless of the operating mode, whether loading a value for a single output, for blocks of 8 outputs (Fig. 8), which was one of our test objectives, or for 16 outputs (Fig. 9).

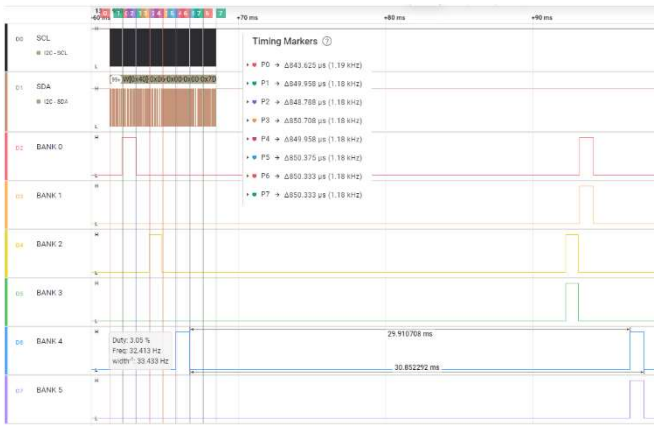


Fig. 8. Update times of the 8-output banks of the PCA9685 after receiving the channel stop signal.

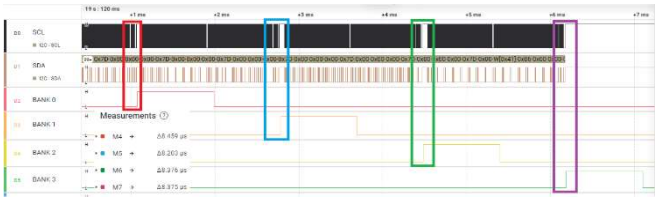


Fig. 9. Update times of the 16 outputs of the PCA9685 after receiving the channel stop signal.

One very important point that caught significant attention was the behavior observed when simulating operation in banks of 8 servos. As an example, Fig. 8 illustrates that when the I²C transaction for bank 0 ends (outputs 1 to 8 of the PCA at address 0x40), 8 µs after the STOP bit signal, the PWM

period of the entire PCA begins (the frequency is the same for all outputs), and loading of bank 1 (outputs 9 to 16 of the PCA at address 0x40) then begins; however, the PWM cycle of these outputs will only start at the beginning of the next PWM cycle of that PCA. This effect can be observed in all banks – PCAs in Fig. 8.

Since the ARDUINO UNO was being used for the tests, the source of the problem was identified as related to the size of its I²C buffer in the library, which allowed transmission of only 32 bytes per I²C transaction — coincidentally updating the positions of 8 outputs — with the second I²C transaction completing the remaining 16 outputs. The methodology originally specified the Arduino UNO for these acquisitions, however, upon identifying the limitation, the team switched to the ESP32devkit, redoing the data collection accordingly — all figures presented in this section (Fig. 8, Fig. 9, and Fig. 10) were captured using the ESP32devkit. Resuming the tests to evaluate the 16-output bank configuration, it was observed that, although the PCA9685 outputs have being updated and the I²C packets could be seen on the logic analyzer, the ATMEGA328 microcontroller kept writing two separate 32-byte packets for the outputs, instead of the expected 64 bytes that should appear in a single I²C transaction, since each position/output uses 4 bytes.

The solution would be to modify the standard library used or to change the microprocessor. We chose to migrate the test firmware to the ESP32devkit, which has up to 128 bytes of buffer, solving the buffer size problem, allowing the definition of 8- or 16-output updates per PCA9685, and bringing the hardware closer to the final version. Another change was made to increase the I²C bus frequency to 400 kHz, since with the ARDUINO UNO it operated at 100 kHz. With these changes, we solved the problem of writing 16 outputs (64 bytes) in a single I²C transaction, and the acquisitions for writing 16 outputs per transaction were repeated (Fig. 10).

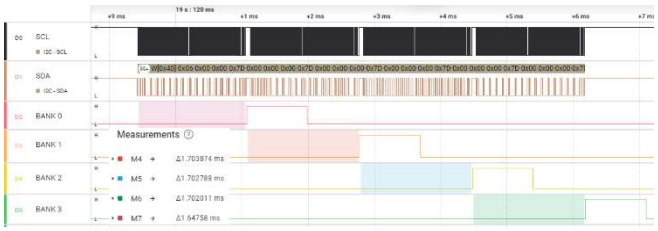


Fig. 10. Update times of the PCA9685 outputs measured from the start of the pulse of the previous PCA output – 16 outputs.

The procedure to measure this interval consisted of loading all 16 outputs of the PCAs with a value of 0, then starting the loading of duty cycle values for all 16 outputs of each PCA as well. Once the I²C transaction begins and the first STOP related to the PCA address occurs, the PWM cycle for that PCA begins for all 16 outputs. This procedure is repeated for each PCA slave on the bus.

IV. DISCUSSION AND CONCLUSION

Although the TIGRÃO robot is controlled by an 8-bit microcontroller, it still represents a major achievement of its development team, since its architecture is simple and discrete, based on performing multiple functions within each 5 ms window generated by a timer interrupt, optimizing the hardware to the maximum beyond what has already been described. Readings from 64 analog channels and 64 digital inputs were implemented using the same existing TDM selection signals, since these already had a planned destination: receiving signals from the WOZ operator suit, developed with potentiometers to capture the movements of arms, shoulders, legs, and other joints.

The scope of this study is limited to characterizing PWM signal generation. Treating the PWM signal generation electronics as a black box, its output does not interfere with the functioning of the servomotor and its attached mechanics since the same characteristics of the generated control signal are maintained, thereby ensuring an equivalent mechanical response regardless of the generation architecture used. Direct mechanical validation with load is beyond the scope of this comparison and is recommended as future work.

BORIS robot project inherits the concept of the Operator Suit, but now with current MPU technology, haptic feedback, and other features, aiming to offer greater interaction and realism for the operator, greater control of the robot, and the capture of scientific data for future evaluation. The behavioral implications of these technical improvements — including their effect on perceived expressiveness and child engagement — will be evaluated in the clinical trials of the BORIS project.

In TIGRÃO robot project, TDM operates with 180 steps, 100% within the useful range. At 33 Hz frequency used in this comparison, the PCA9685's 4096 total steps yield approximately 270 effective steps within the servo's useful operating range (0.5 ms to 2.5 ms), a 1.5× gain over TDM. At the 50 Hz frequency planned for BORIS's servo operation this effective count rises to approximately 410 steps, a 2.3× gain. In both cases, the PCA does not use its total counter range, but rather the effective steps that fall within the servo's commercial useful operating range.

Evaluating synchronization for the control of multiple servomotors, when using several PCA9685 units, the use of an external oscillator circuit must be considered to improve output synchronization and mitigate the drift effect caused by internal oscillators. This effect becomes concerning in a case similar to TIGRÃO, which requires 37 servomotors (DoFs). In the case of TDM, this problem does not occur, since the time base is the same for all banks, with only a 5 ms offset between banks in the TIGRÃO project, which could be reduced as the bank update rate is improved.

Traditional software-based PWM generation on a general-purpose microcontroller timer interrupt is inherently limited in channel count — the reference application note for this technique reports up to 24 PWM channels using its supplied code sample, with the demonstrated example implementing ten channels [18]. Microcontrollers such as the ESP32 offer up to 16 dedicated hardware PWM outputs [24], which can compromise the number of pins available for a project's other needs; the ATMEGA2560, in turn, has 54 digital pins, of which 15 can be configured as PWM outputs [25]. The microcontroller used in TIGRÃO project in 2005, however, was the PIC18F452, which has only two internal hardware modules for PWM generation [26]. TDM technique implemented in its architecture is therefore essentially discrete time-division multiplexing logic, using only 13 pins (2 control, 3 selection, 8 data) to drive external pin-expansion

hardware for up to 64 outputs. This output-expansion hardware could be migrated to a modern FPGA, drastically reducing the physical board space. Furthermore, the FPGA opens the possibility of implementing a dedicated multichannel PWM generation peripheral analogous to the PCA9685, with a configurable interface such as SPI, I²C, or even parallel as currently used, with position resolution defined by the design. However, the adoption of FPGAs in SAR projects must take into account the high cost of the devices, the complexity of the development tools (synthesis, simulation, programming), and the significant learning curve for the development team, factors that can be decisive in projects with limited resources, lower levels of education, or in a rapid prototyping phase.

In TIGRÃO robot, the facial expression servomotors are grouped in banks 0 and 1, ensuring that all are updated within a fixed 5 ms window, a hardware design parameter that remains well below the audiovisual asynchrony detection thresholds reported in the literature [22], and therefore imperceptible in HRI. This margin is preserved in BORIS robot, since transmitting the 16 outputs in a single I²C transaction allows practically all head servos to be updated within a comparably short interval, consistent with the same perceptual margin [23].

With its 37 degrees of freedom (DoFs) distributed among the head, upper limbs, and lower limbs, the TIGRÃO robot still represents a major achievement in the field of pediatric hospital-use SAR.

The comparative analysis between the TDM and PCA9685 techniques directly addresses the study's proposed objective, validating the hardware architecture for the BORIS robot. The results show that, although both solutions are effective for SAR development while keeping latency below 150ms, their application must be carefully analyzed, evaluating the context and functionality of each, and especially the technical knowledge required to implement them. TDM technique in the TIGRÃO robot was built entirely with discrete components, with the timer interrupt service routine (ISR) managing — not monopolizing — bank selection and PWM timing every 5 ms. Based on cycle-accurate analysis of the compiled instruction listing, the ISR remains active for approximately 3.4 to 3.8 ms per 5 ms bank period (69% to 75%), comprising a fixed initial window for digital and analog input handling followed by the 180-step comparison loop; the remaining 25% to 31% of each period is available to the main loop for joystick reading, serial menu commands, and EEPROM pose storage. In the case of the PCA9685, the load on the microprocessor is limited to I²C transactions, not to mention the reduction in the number of physical connections used.

Therefore, for the BORIS project, the intention is to use the PCA9685 based on the data collected during this study, since it proved to be the right path for innovation, in addition to meeting the goal of keeping the microprocessor nearly 100% free for other tasks — which does not invalidate the TDM technique, the alternative chosen in 2005, when a component with the flexibility and features of the PCA9685 was not yet available, and implementation with programmable logic was beyond the reach of the development team at the time.

ACKNOWLEDGMENT

The authors thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and NAPI TA - Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná (FA) for their important funding, support, and scholarships.

REFERENCES

[1] M. Heerink *et al.*, "Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model," *International Journal of Social Robotics 2010 2:4,* vol. 2, no. 4, 2010-09-04, doi: 10.1007/s12369-010-0068-5.

[2] A. L. Nunez, "The Future of Social Robots in Paediatric Care," *J Nurs Care,* vol. 13, no. 1, pp. 1-3, 2024.

[3] A. S. B. L. C. M, M. L. A. R. E. A. G. R. E, and F. C, "Socially Assistive Robotics for children with Autism Spectrum Disorder (ASD): a systematic review," in *2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2025, pp. 1-6.

[4] H. Ashraf, "Robôs Humanoides - Desenvolvimentos Recentes e Interação Humano-Robô: Uma revisão de artigos," 2023 2023

[5] A. Puglisi *et al.*, "Social Humanoid Robots for Children with Autism Spectrum Disorders: A Review of Modalities, Indications, and Pitfalls," *Children 2022, Vol. 9, Page 953,* vol. 9, no. 7, 2022-06-25, doi: 10.3390/children9070953.

[6] L. Lalla, *A design of actuation mechanisms for use in 'huggable' robotic teddy bear*. 2006, p. 24 leaves.

[7] A. A. Cantone, M. Esposito, F. P. Perillo, M. Romano, M. Sebillo, and G. Vitiello, "Enhancing elderly health monitoring: Achieving autonomous and secure living through the integration of artificial intelligence, autonomous robots, and sensors," *Electronics,* vol. 12, no. 18, p. 3918, 2023.

[8] Z. Paré, "Real Robots:: Estudo comparativo de robôs sociais," *Iluminuras,* vol. 25, no. 69, pp. 246-268, 2024.

[9] K. D. Bartl-Pokorny *et al.*, "Robot-based intervention for children with autism spectrum disorder: a systematic literature review," *Ieee Access,* vol. 9, pp. 165433-165450, 2021, doi: 10.1109/ACCESS.2021.3132785.

[10] C. Bettencourt *et al.*, "Investigating the Feasibility of a Wizard-of-Oz Robotic Interface (R2C3) in a Social Skills Group for Children with Autism Spectrum Disorder," *International Journal of Social Robotics 2025,* 2025-04-07, doi: 10.1007/s12369-025-01243-4.

[11] C. Clabaugh and M. Matarić, "Escaping oz: Autonomy in socially assistive robotics," *Annual Review of Control, Robotics, and Autonomous Systems,* vol. 2, no. 1, pp. 33-61, 2019.

[12] T. Shibata, "Therapeutic seal robot as biofeedback medical device: Qualitative and quantitative evaluations of robot therapy in dementia care," *Proceedings of the IEEE,* vol. 100, no. 8, pp. 2527-2538, 2012.

[13] S. Shamsuddin, H. Yussof, L. I. Ismail, S. Mohamed, F. A. Hanapiah, and N. I. Zahari, "Initial response in HRI-a case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot Nao," *Procedia Engineering,* vol. 41, pp. 1448-1455, 2012.

[14] S. Jeong *et al.*, "Designing a socially assistive robot for pediatric care," in *Proceedings of the 14th international conference on interaction design and children*, 2015, pp. 387-390.

[15] M. d. C. C. Gaiotto, "Projeto T.I.G.R.A.O – TIGRÃO," in *Anais da Mostra Nacional de Robótica (MNR)*, A. d. S. Simões, E. L. Colombini, and F. Tonidandel Eds. São João del Rei, MG, 2011, p. 111.

[16] M. Tripodi *et al.*, "Humanization interventions in general pediatric wards: a systematic review," *European journal of pediatrics,* vol. 178, no. 5, pp. 607-622, 2019.

[17] B. Croitoru, A. Tulbure, and M. Abrudean, "Microcontroller-based multiple-platform PWM signal generation procedures for industrial use," in *2014 IEEE International Conference on Automation, Quality and Testing, Robotics*, 22-24 May 2014 2014, pp. 1-6, doi: 10.1109/AQTR.2014.6857891.

[18] Microchip Technology Inc. (Atmel), "AVR136: Low-Jitter Multi-Channel Software PWM," *Application Note Atmel-8020C,* Sep. 2016.

[19] A. T. Spröwitz, A. Tuleu, M. Ajallooeian, M. Vespignani, R. Möckel, P. Eckert, M. D'Haene, J. Degrave, A. Nordmann, B. Schrauwen, J. Steil, and A. J. Ijspeert, "Oncilla Robot: A Versatile Open-Source Quadruped Research Robot With Compliant Pantograph Legs," *Frontiers in Robotics and AI,* vol. 5, art. 67, 2018, doi: 10.3389/frobt.2018.00067.

[20] M. V. Kujala, L. Parkkonen, and J. Kujala, "Empathy enhances decoding accuracy of human neurophysiological responses to emotional facial expressions of humans and dogs," *Social Cognitive and Affective Neuroscience,* vol. 19, no. 1, 2024, doi: 10.1093/scan/nsae082.

[21] D. Löffler, J. Dörrenbächer, and M. Hassenzahl, "The uncanny valley effect in zoomorphic robots: The U-shaped relation between animal likeness and likeability," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 261-270.

[22] R. Eg and D. M. Behne, "Perceived synchrony for realistic and dynamic audiovisual events," *Frontiers in psychology,* vol. 6, p. 736, 2015.

[23] A. Diel, W. Sato, C.-T. Hsu, and T. Minato, "Asynchrony enhances uncanniness in human, android, and virtual dynamic facial expressions," *BMC research notes,* vol. 16, no. 1, p. 368, 2023.

[24] Espressif Systems, "LED Control (LEDC)," *ESP-IDF Programming Guide.* [Online]. Available: https://docs.espressif.com/projects/esp-idf/en/stable/esp32/api-reference/peripherals/ledc.html

[25] Arduino, "Mega 2560 Rev3," *Arduino Documentation.* [Online]. Available: https://docs.arduino.cc/hardware/mega-2560

[26] Microchip Technology Inc., *PIC18F242/252/442/452 Data Sheet,* DS39564C, 2006.

Coordinative Variability Between Hip–Knee Joints during Gait using Generalized Wavelet Coherence Analysis

1st Guilherme A. Gomes De Villa
State University of Goiás (UEG)
Itaberaí-GO, Brazil
0000-0003-3331-9678

2nd Hygor Henrique da Silva Sales
University of Rio Verde (UniRV)
Aparecida de Goiânia-GO, Brasil
0009-0000-0210-2527

3rd Ramon Ribeiro de Lima
Federal University of Goiás (UFG)
Goiânia-GO, Brazil
0009-0005-5035-0142

4th Paulo H. Santana de Oliveira
State University of Goiás (UEG)
Itaberaí-GO, Brazil
0000-0002-7421-7881

5th Kevin Cavalcante Almeida
Goyazes University Center (UNIGOYAZES)
Trindade-GO, Brazil
0009-0007-3787-9098

6th Gustavo Antunes de Queiroz Gonçalves
University of Rio Verde (UniRV)
Aparecida de Goiânia-GO, Brasil
0009-0004-9358-892x

7th Marcus Fraga Vieira
Biomechanical and Biomedical L.
Federal University of Goiás
Goiânia-GO, Brazil
0000-0001-9096-1603

***Abstract*— This study applied the Generalized Wavelet Coherence Analysis (GWCA) technique, proposed by Espen Ihlen (2014), to investigate whether variability in hip–knee joint coupling during gait in elderly fallers and non-fallers is associated with falls, compared with young adults. Kinematic data were collected from 30 young adults, 17 older non-fallers, and 13 older fallers during treadmill walking at three speeds: preferred walking speed (PWS), 120% PWS, and 80% PWS. Participants performed three 1-minute trials at each speed in random order. Data from 16 reflective markers were used to compute hip and knee joint angles across four gait phases (first double support, single support, second double support, and swing), from which sagittal-plane coordinative variability was calculated. All participants exercise regularly at least three times a week. Kinematic data from 16 markers were used to calculate hip and knee joint angle values for four gait phases (first double stance, single stance, second double stance and swing), from which coordinative variability was calculated, in the sagittal plane. Results showed that GWCA distinguished gait patterns between elderly fallers and non-fallers. Exploratory ROC analysis further indicated that selected GWCA-derived parameters had potential to discriminate between elderly fallers and non-fallers, supporting their potential relevance for fall-related gait assessment. Significant speed effects were observed, with greater variability at 80% and 120% PWS, mainly indicating knee dominance over the hip.**

***Index Terms*— wavelet coherence analysis, older adults, aging, falls, gait**

I. Introduction

Population aging has become one of the main demographic transformations in Brazil[1]. According to recent estimates from the Brazilian Institute of Geography and Statistics (IBGE), the country has approximately 213.4 million inhabitants, with more than 34 million people aged 60 years or older, representing nearly 16% of the total population. Demographic projections indicate that this proportion will continue to increase over the coming decades, leading to a substantial rise in age-related health conditions[2]. Falls are among the most frequent adverse events affecting older adults, with approximately 28–30% of this population experiencing at least one fall annually[3]. Besides their impact on mobility, independence, and quality of life, falls account for a considerable number of hospitalizations and generate a substantial economic burden on the Brazilian public health system due to emergency care, surgical procedures, rehabilitation, and long-term treatment. Consequently, strategies aimed at identifying factors associated with fall risk and improving early detection have become increasingly important in the context of an aging population.

The gait can be described as a cyclical movement of body parts [4]. The coordination between segments provides a measure of the relative timing and magnitude of movement coordination variability measures the varied movement patterns that an individual uses during locomotion [5], [6]. Coordination variability represents the range of movement strategies that an individual can employ while walking [7], [8], [9].

Wavelet is a transient waveform of finite duration [10]. Generalized Wavelet Coherence Analysis (GWCA) combines wavelet coherence techniques with a matrix-based correlation method to estimate correlation coefficients among multiple joints simultaneously. The method divides the signal into smaller segments and analyzes them at different scales, allowing evaluation within short time windows [11]. The wavelet transform operates by decomposing a signal into multiple time-scale components, allowing each segment of the data to be analyzed at a resolution that is appropriate for its corresponding frequency content. In practice, the signal is examined within localized time windows using a reference function, commonly referred to as the

mother wavelet. High-frequency components are obtained by applying contracted versions of the mother wavelet, whereas low-frequency components are represented through its dilated forms, enabling the simultaneous investigation of both transient and long-duration signal characteristics [12]. The correlation between two signals is then computed across different time scales, producing time-frequency contour maps that provide a detailed representation of their dynamic relationship. From these analyses, an overall instantaneous correlation coefficient (Rt) is calculated, representing the global wavelet-based correlation between the paired signals throughout the analyzed period [11]. However, the discriminative potential of GWCA-derived coordination variability for distinguishing older adults with and without a history of falls remains insufficiently explored. Therefore, this approach may provide clinically relevant information for the assessment of fall-related gait alterations.

Based on this approach, the present study aimed to estimate coordination variability using the GWCA method between the hip and knee joints in three distinct groups—young adults, older adults without a history of falls(non-fallers), and older adults with a history of falls (fallers)—while they walked on a treadmill at different speeds. We hypothesized that (1) young people present greater coordination variability in relation to the other groups, that (2) the joint variability calculated by the GWCA technique for elderly non-fallers will be greater than that of fallers and (3) that the higher the speed greater the joint variability will be in all groups.

II. Material and Methods

A. Subjects

Thirty young adults (23.90 ± 5.05 years), 17 older non-fallers (64.68 ± 3.74 years), and 13 older fallers (68.35 ± 7.41 years) participated in this study. Older participants were classified as fallers if they reported three or more falls in the previous year. The protocol was approved by the local Research Ethics Committee (1.003.935–2018), and all participants provided written informed consent.

B. Protocol

Kinematic data were collected using sixteen retroreflective markers placed according to the Vicon Plug-in-Gait lower limb model and recorded with a 3D motion capture system (10 infrared cameras, 100 Hz). Signals were filtered using a fourth-order zero-lag Butterworth low-pass filter (8 Hz) and processed with custom MATLAB routines (R2025b). The preferred walking speed (PWS) was determined following a previous protocol [13], and participants performed three one-minute treadmill trials at 80%, 100%, and 120% of the PWS, in random order, with one-minute rest intervals.

The hip-knee joint pair was analyzed using 20 strides per trial, with time series normalized to 100 points. Analysis was performed in the sagittal plane, and the gait cycle was divided into first double support (0–10%), single support (11–50%), second double support (51–60%), and swing (61–100%) phases [14], [15].

Using Generalized Wavelet Coherence Analysis (GWCA), contour plots were obtained from instantaneous correlation coefficients (R_t) between joint angles at different time scales [11]. For each phase, the median R_t generated two maximum (Rmax) and two minimum (Rmin) values, representing respectively the double support phases (stronger coordination) and the single support and swing phases (weaker coordination). The Rmax and Rmin parameters were selected to characterize the highest and lowest coordination values within each gait phase, thereby capturing phase-specific variations in hip–knee coordinative variability.

C. Statistical analysis

The repeated measures analysis of variance (ANOVA)with mixed design was used to compare the three groups and the main effect of speed, followed by a post-hoc test with Bonferroni correction in the cases where the main or interaction effect was significant. Statistical analysis wasperformed using SPSS software, version 23 (SPSS Inc., Chicago, IL, USA), with a significance level set at $\alpha < 0.05$.

Additionally, an exploratory Receiver Operating Characteristic (ROC) analysis was performed to assess the discriminative capacity of GWCA-derived parameters between elderly fallers and non-fallers. The ROC analysis was restricted to gait phases and walking-speed conditions in which significant group effects were identified in the mixed ANOVA, and the area under the curve (AUC), 95% confidence interval, sensitivity, specificity, and optimal cutoff value based on the Youden index were calculated.

III. Results

Fig. 1 presents the GWCA contour maps for the hip-knee joint signals at scales from 1/20 to 1/5 of the average stride time for walking speeds of 80%, 100%, and 120% of the preferred walking speed (PWS). In these maps, red areas indicate stronger intra-articular coupling, according to the color scale. The figure suggests that young participants exhibited greater variability than the older groups.

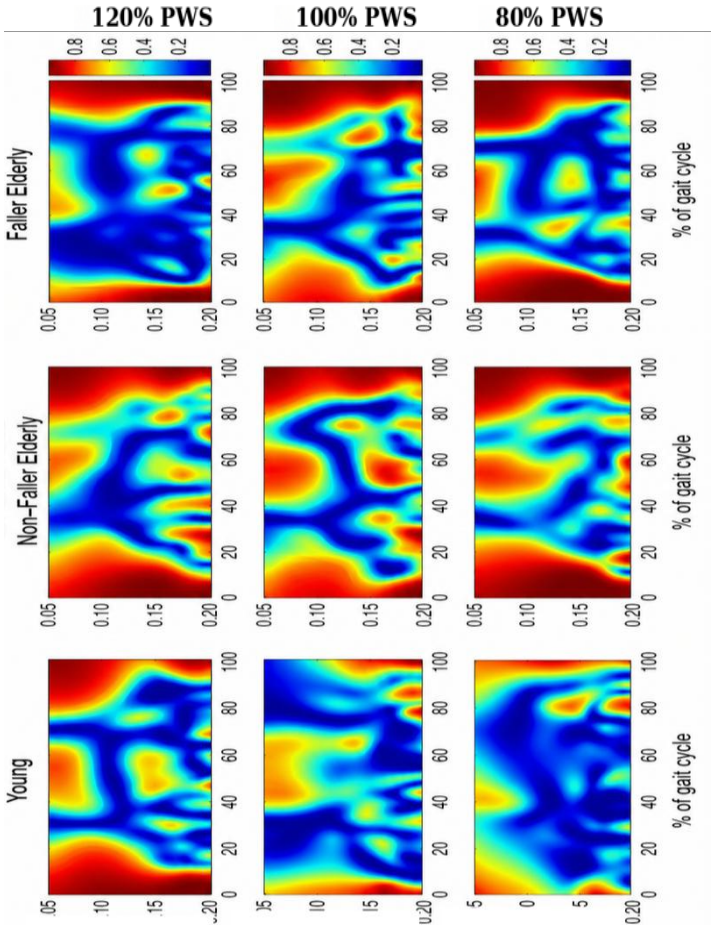


Fig. 1 Contour plotter for Hip-Knee joint signs, for scales from 1/20 to 1/5 of the average stride time.

To quantitatively characterize the wavelet coherence patterns, the median value of the instantaneous correlation coefficient (Rt) is computed for each phase of the gait cycle. Four representative values are extracted: two peak values (Rmax), corresponding to the first and second double-support phases, and two minimum values (Rmin), associated with the single-support and swing phases. The Rmax values indicate the gait events at which intersegmental coordination reaches its highest level, whereas the Rmin values identify the phases characterized by reduced coupling intensity between the analyzed signals.

The Fig. 2 show averages of coordinative variability of Rt, evaluated by GWCA, for the Hip-Knee joint pair at speeds 80% of PWS, PWS and 120% of PWS. From the Figure 2, it is possible to identify that in most situations' older non-fallers people (black curves) presented higher values of coordinative variability, followed by fallers people (red curves), and young people (dashed blue curve), this is not in line with hypothesis 1. Therefore, the graphical results suggest that gait speed may influence coordinative variability in a phase-dependent manner rather than producing a consistent monotonic increase in Rt. Accordingly, these findings provide partial graphical support for hypothesis 3, which should be interpreted in conjunction with the statistical analysis.

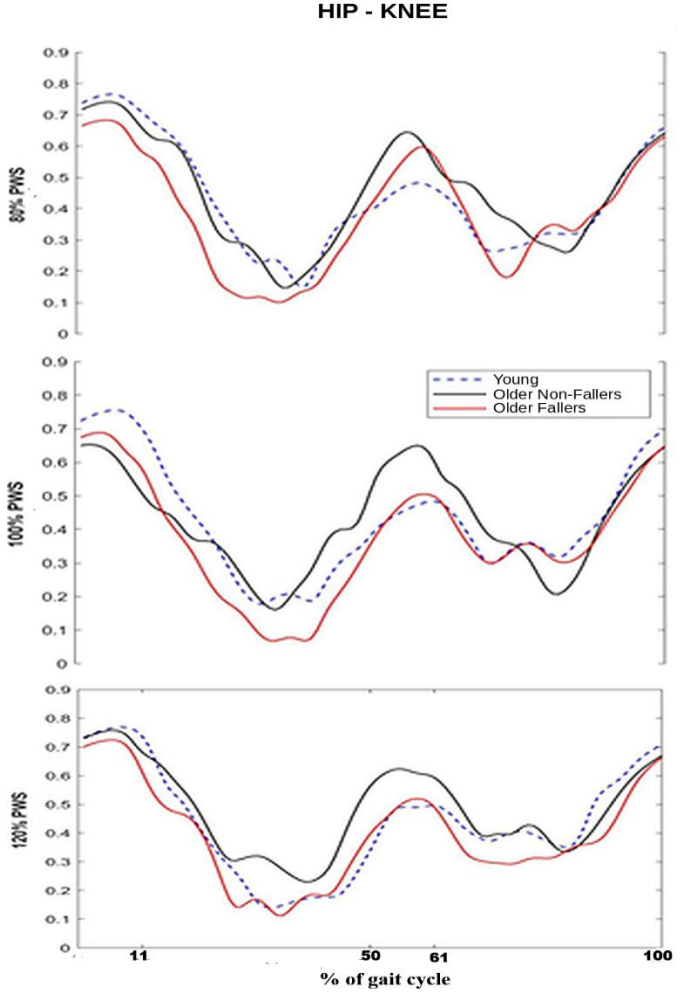


Fig. 2 Averages of coordinative variability, evaluated by GWCA, for the Hip-Knee joint pair at speeds 80% of PWS, PWS and 120% of PWS.

The statistical analysis (p-values) of the variability of $Rmax$ and $Rmin$ is presented in Table 1 below and was evaluated with a mixed repeated measures analysis (mixed ANOVA) and significance level $p < 0.05$.

TABLE I. VARIABILITY OF R_{max} and R_{min} FOR HIP-KNEE PAIR.

Effect	Phases of Gait	80%PWS	100%PWS	120%PWS
Group	***FDS***	NS	0.23	p< 0.001
	SS	NS	NS	p< 0.001
	SDS	0.039	NS	NS
	SG	NS	NS	**NS**
		Young	**Elderly Non-Fallers**	**Elderly Fallers**
Speed	***FDS***	p< 0.001	p< 0.001	p< 0.001
	SS	NS	p< 0.001	p< 0.001
	SDS	0.030	NS	NS
	SG	p< 0.001	NS	NS

Analysis of Repeated Measures (ANOVA). FDS = First Double Support; SS = Single Support; SDS = Second Double Support; SG = Swing.

Significant group-related differences were identified during the FDS phase, with higher Rt values observed in elderly non-fallers compared to elderly fallers. Significant group-related differences were identified in the FDS phase at 120% PWS ($p < 0.001$), with elderly non-fallers presenting higher Rt values than elderly fallers. A significant group effect was also observed in the SDS phase at 80% PWS ($p = 0.039$). Additionally, a significant group effect was found in the SS phase at 120% PWS ($p < 0.001$). Concerning walking speed, the young group showed significant differences between 120% and 80% of the preferred walking speed (PWS) in the FDS, SDS, and SG phases. In contrast, among elderly fallers, significant differences were detected only in the SG phase.

To further investigate the potential discriminative capacity of GWCA-derived coordinative variability between elderly fallers and non-fallers, an exploratory ROC analysis was performed. The analysis was restricted to gait phases and walking-speed conditions in which significant group effects were identified in the mixed ANOVA. Specifically, Rmax at the FDS phase at 120% PWS, Rmax at the SDS phase at 80% PWS, and Rmin at the SS phase at 120% PWS were evaluated for their ability to discriminate between elderly fallers and non-fallers. For each parameter, the area under the ROC curve (AUC), 95% confidence interval, sensitivity, specificity, and optimal cutoff value based on the Youden index were calculated. This exploratory analysis was intended to assess the potential discriminative value of GWCA-derived coordinative variability for identifying individuals with a history of falls, rather than to establish its predictive ability for future falls.

IV. Discussion

In terms of group, there were significant differences in the SDS phase with the Rt values of the non-falling elderly being higher than that of the falling elderly. The significant statistical differences between the Rt of the SDS phase for elderly non-fallers in relation to fallers may reflect a greater requirement for inter-articular coordination of the impulsion lower limb to accelerate the center of mass in gait progression compared to the lower limb of

impact on the ground that slows down the movement of the center of mass immediately after the heel touches the ground [16]. Therefore, age-related decline in inter-joint coordination only in these phases may indicate deficiencies related to push-off coordination [11]. These findings are consistent with previous evidence that older adults with a history of falls may exhibit altered inter-joint coordination and reduced ability to appropriately regulate lower-limb movements during critical phases of gait[17].

Regarding walking speed, significant differences were observed primarily between the slowest (80% PWS) and fastest (120% PWS) conditions. In the young group, these differences were evident during the first double-support (FDS), second double-support (SDS), and swing (SG) phases, whereas in older fallers significant changes were identified only during the FDS and SG phases. Similar findings have been reported in previous studies investigating gait modulation in young and older adults, indicating that locomotion at speeds below the preferred walking speed requires greater adjustments in lower-limb coordination than walking at faster speeds relative to PWS [7], [18]. These adaptations may reflect the increased neuromuscular demand required to preserve gait stability and maintain forward progression when the natural walking rhythm is altered[19].

Since the hip and knee joints play a fundamental role in controlling body support, propulsion, and limb advancement throughout the gait cycle, modifications in their coordination pattern are expected when walking speed changes[20]. In particular, the double-support phases require accurate coordination to ensure a smooth transfer of body weight between limbs, whereas the swing phase is characterized by precise control of limb advancement and preparation for the subsequent heel contact[21], [22]. The observed differences in *Rt* across walking speeds may represent adaptive strategies adopted to preserve dynamic stability while maintaining efficient locomotor performance.

The exploratory ROC analysis provides an additional perspective on the potential discriminative value of GWCA-derived coordinative variability between elderly fallers and non-fallers. The parameters selected for this analysis were those that showed significant group effects in the mixed ANOVA, namely Rmax during the FDS phase at 120% PWS, Rmax during the SDS phase at 80% PWS, and Rmin during the SS phase at 120% PWS. The observed ROC performance suggests that alterations in hip–knee coordinative variability may have potential to distinguish older adults with and without a history of falls. These findings are particularly relevant because the identified parameters were obtained under different walking-speed conditions and gait phases, suggesting that the discriminative behavior of inter-joint coordination may depend on the specific locomotor demands imposed by the task. However, given the exploratory nature of the analysis and the use of a cross-sectional classification based on fall history, these results should not be interpreted as evidence that GWCA-derived parameters can predict future falls.

This study focused exclusively on hip–knee coordination in the sagittal plane, future investigations should incorporate analyses of additional planes of motion and other lower-limb joint couplings to provide a more comprehensive understanding of the biomechanical adaptations associated with aging and fall risk.

V. Conclusion

The present study analyzed coordinative variability during gait between the hip–knee joints in young adults, older non-fallers, and older fallers, using the Generalized Wavelet Coherence Analysis (GWCA) technique, which demonstrated sensitivity to changes in intra-articular coordination of this joint pair. The exploratory ROC analysis further indicated the potential of selected GWCA-derived parameters to discriminate between older fallers and non-fallers, suggesting that coordinative variability may have potential as a complementary marker for fall-related gait assessment. However, further prospective studies with larger samples are needed to confirm these findings and investigate their applicability to future fall-risk assessment. Future studies should investigate the influence of physical activity, muscle strength, and a wider range of walking speeds on hip–knee coordinative variability.

Acknowledgment

The authors would like to thank the financial support of the National Council for Scientific and TechnologicalDevelopment (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Foundation for Research Support of State of Goiás (FAPEG). M. F. Vieira is CNPq fellow, Brazil (304533/2020-3).

REFERENCES

[1] S. C. Dumith and N. Feter, “Demographic shifts and health dynamics: Exploring the impact of aging rates on health outcomes in Brazilian capitals,” *Archives of Gerontology and Geriatrics Plus*, vol. 1, no. 3, p. 100044, Sep. 2024.

[2] “Brazilian Institute of Geography and Statistics (IBGE),” https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18318-piramide-etaria.html.

[3] L. Almidani *et al.*, “Gait Metrics in Elderly Fallers and Non-Fallers with Varying Levels of Glaucoma: A Longitudinal Prospective Cohort Study,” *Sensors*, vol. 25, no. 12, p. 3712, Jun. 2025.

[4] L. Tesio and V. Rota, “The Motion of Body Center of Mass During Walking: A Review Oriented to Clinical Applications,” *Front. Neurol.*, vol. 10, Sep. 2019.

[5] K. A. HAFER, J. F.;BOYER, “Variability of segment coordination using a vector coding technique: Reliability analysis for treadmill walking and running,” *Gait Posture*, vol. 51, pp. 222–227, 2017.

[6] T. Krauskopf *et al.*, “Altered gait stability and regularity in lower limb amputees observed across different body segments,” *J. Biomech.*, vol. 186, p. 112723, Jun. 2025.

[7] P. Floría, A. Sánchez-Sixto, A. J. Harrison, and R. Ferber, “The effect of running speed on joint coupling coordination and its variability in recreational runners,” *Hum. Mov. Sci.*, vol. 66, pp. 449–458, Aug. 2019.

[8] A. O. VIERA,M.F.; RODRIGUES, F.B.; SOUZA, G.S.S.; MAGNANI, R.M.; LEHNEN,G.C.; ANDRADE, “Linear and Nonlinear Gait Features in Older Adults Walking on Inclined Surfaces at Different Speeds,” *BIOMEDICAL ENGENEERING SOCIETY*, vol. 10, 2017.

[9] A. de Oliveira Assis *et al.*, “Influence of sampling frequency and number of strides on recurrence quantifiers extracted

from gait data," *Comput. Biol. Med.*, vol. 119, p. 103673, Apr. 2020.

[10] Derman Dondorur, *Acquisition and Processing of Marine Seismic Data*. Elsevier, 2018.

[11] E. A. F. Ihlen, "Age-related changes in inter-joint coordination during walking," *J. Appl. Physiol.*, vol. 117, no. 2, pp. 189–198, Jul. 2014.

[12] Amara Graps, "An Introduction to Wavelets," *IEEE Cpmputational Science and Engineering*, pp. 1–18, 1995.

[13] L. C. DINGWELL, J. B.; MARIN, "Kinematic variability and local dynamic stability of upper body motions when walking at different speeds.," *J. Biomech.*, vol. 39, pp. 444–452, 2006.

[14] N. C. Robert Needham, Roozbeh Naemi, "Quantifying lumbar-pelvis coordination during gait using a modified vector coding technique," *J. Biomech.*, vol. 47, pp. 1020–1026, 2014.

[15] K. A. Boyer *et al.*, "Age-related changes in gait biomechanics and their impact on the metabolic cost of walking: Report from a National Institute on Aging workshop," *Exp. Gerontol.*, vol. 173, p. 112102, Mar. 2023.

[16] H. Sadeghi *et al.*, "Lower-Extremity Intra-Joint Coordination and Its Variability between Fallers and Non-Fallers during Gait," *Applied Sciences*, vol. 11, no. 6, p. 2840, Mar. 2021.

[17] Z. Zeng, C. Ho, J. Zhou, J. Shen, and Y. Yang, "Effects of Dual-Tasking on Stepping Strategy and Inter-Joint Coordination During Walking in Older Fallers and Non-Fallers," *Innov. Aging*, vol. 9, no. 6, Jun. 2025.

[18] G. Spallone *et al.*, "Step-by-step insight into gait analysis: A narrative review unlocking knee biomechanics," *Knee Surgery, Sports Traumatology, Arthroscopy*, Sep. 2025.

[19] G. A. G. De Villa, L. A. Elias, and M. F. Vieira, "Quantifying Coordinative Variability Between Knee-Ankle Joints During Gait of Young, Elderly Fallers and Non-Fallers Using the Generalized Wavelet Coherence Analysis," 2025, pp. 340–346.

[20] X. Zhang *et al.*, "Biomechanical and neuromuscular adaptations in joint contributions during loaded countermovement jumps," *Sci. Rep.*, vol. 15, no. 1, p. 19167, May 2025.

[21] T. Ghanavati *et al.*, "Intra-limb coordination while walking is affected by cognitive load and walking speed," *J. Biomech.*, vol. 47, no. 10, pp. 2300–2305, Jul. 2014.

[22] J. P. Bailey, J. Freedman Silvernail, J. S. Dufek, J. Navalta, and J. A. Mercer, "Effects of treadmill running velocity on lower extremity coordination variability in healthy runners," *Hum. Mov. Sci.*, vol. 61, pp. 144–150, Oct. 2018.

Data Gaps in Trust Measurement for Physical Human-Robot Collaboration

Pedro M. Milhomem
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein, São Paulo, Brazil
0009-0009-6301-5262

Gustavo J. G. Lahr
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein, São Paulo, Brazil
0000-0002-0403-991X

***Abstract*—Operator trust governs whether part of a physical task can be handed to a robot: too little and the robot is underused, too much and the risk of collision rises. Estimating trust during the interaction depends on datasets that pair trust labels with recorded signals, and the field has not agreed on what to record or how. Through a documented search we retained eight collections that report a trust measurement in a physical task, comparing them on setting, mechanical coupling, instrument, time scale, recorded signals, and data availability. Each cell records whether its entry is documented, absent from the recorded signals, unreported, or without a referent in that setting. The first result concerns the settings rather than the measurements: only three of the eight place the operator in contact with the robot, and sustained contact occurs in one. Five gaps follow. No common protocol fixes the instrument or the time scale. A single trust score is often copied onto many observations and the model then tested on randomly drawn folds; the one study reporting both splits sees accuracy fall from 0.92 to 0.45 once participants are held out. No collection records force or torque, leaving its relation to trust untested rather than ruled out. Anonymization is never varied within a collection. The measured construct differs across collections, and only two of the eight are openly deposited. Four priorities follow: report the instrument, the sampling point, how the score was propagated, and how the evaluation folds were drawn; record force, torque, and trust on one clock under controlled loads; label one cohort at more than one time scale; and vary anonymization as a controlled factor.**

***Index Terms*—human-robot interaction, physical collaboration, trust, datasets, measurement.**

I. Introduction

The arrival of collaborative robots outside isolated industrial cells has moved the research focus from physical safety toward operator trust, now treated as a condition for their effective use and adoption [1], [2]. When trust is too low, the operator underuses the robot; when it is too high, the risk of collision and harm grows [3]. This matters most in physical human-robot interaction (pHRI), where the human and the robot share the workspace and often the same object, so a misjudged action carries an immediate physical cost. A robot that could track trust as the interaction unfolds could adjust to it, but it must first be measured, and that is where the difficulty begins. Trust cannot be read off directly, only inferred from what the operator reports and from the objective signals the interaction leaves behind. Estimating it therefore depends on data that pairs these signals with trust labels, and increasingly on learning models trained over such data. This makes curated pHRI datasets the foundation on which trust estimation rests. Yet the field has not agreed on what to record or how, and the datasets that exist reflect that disagreement.

Trust is not measured the same way across studies, and the differences start with the task. The collections reviewed here range from explicit handover [4] and shared-load transport [5] to co-located placement [3], object placement with a handover at failure recovery [6], shared-workspace assembly [7], [8], supervisory monitoring [9], and tasks reproduced in mixed or virtual reality [10], [11]. The physical coupling, the failure modes, and the signals available to an observer differ across these settings, so a trust measurement in one is not directly comparable to a trust measurement in another.

The clearest divergence is in the measurement instrument. The questionnaire is still the most common tool, but no single instrument is standard for pHRI, and the available scales were built for different purposes. Two scales come from the automation literature and rate reliance on a machine in general [12], [13]; two were developed for robots and rate the robot's perceived trustworthiness as a partner [14], [15]; another separates a cognitive judgment of the system from an affective attitude toward it [16]. Studies frequently assemble their own item sets from several of these scales; Alhaji et al. draw single items from five different questionnaires [10]. Two studies can therefore both report trust while operationalizing different constructs, and because the questionnaire records a single point in a process that changes over time [1], [8], results rarely transfer from one dataset to another.

The objective and physiological side is just as varied. Trust in pHRI is read from self-report, from observable behavior, and from physiological signals [17]. Physiological signals complement the questionnaire because they track the operator's internal state continuously and involuntarily, registering shifts in arousal, attention, and effort between the discrete moments a questionnaire samples and beyond what the participant reports. Across the collections reviewed here the recorded channels include raw video, eye tracking and gaze, body pose, facial expression and emotion, and central, muscular, cardiac, and autonomic physiology [4], [7], [9]. No two collections record the same combination, and there is no agreed answer to which signals a trust dataset should contain.

The contribution of this paper is a mapping of the trust data collected for physical collaboration onto a common set of mea-

surement parameters, read as a coverage matrix whose empty cells mark the gaps. We describe what has been collected rather than measure it against an ideal. A systematic review of trust assessment in HRI already covers how individual studies measure trust [1]. What is missing is a comparison at the level of the collections themselves, along the parameters that decide whether one dataset can be read against another, including whether its data can be obtained at all. Two further features separate the matrix from a narrative comparison. The retrieval that produced the set is stated, so the boundary of the set is reproducible, and each cell carries the status of its entry, so an absence attributable to a silent publication is not read as an absence of the signal itself. We split the parameters across two tables to keep the comparison readable. Table I covers the environment, the mechanical coupling, the task, robot, sample size, instrument, time scale, induced failures, and data availability, and Table II the objective signal modalities. Figure 1 shows three of the tasks. Section II sets the search, the scope, the terminology, and the cell coding, Section III describes the parameters, Section IV reads the matrix, its data volume, and its internal divergences, and Section V discusses the open problems. The reading begins with a result about the settings rather than the measurements: physical contact occurs in a minority of the collections gathered under the heading of physical collaboration.

II. Search, Scope, Terminology, and Coding

Search protocol. Candidate records were retrieved from IEEE Xplore and Scopus, searched over title and abstract. Two query branches were used: one combining trust with human-robot and a term for released data, run as two separate queries for *dataset* and for *database*, and one combining trust with a term for physical or contact-based interaction (physical human-robot, handover, or collaborative robot). On IEEE Xplore these three queries returned 22, 5, and 117 records; on Scopus, restricted to Computer Science and Engineering, the two branches returned 75 and 521 records. These are unfiltered return counts rather than counts after title/abstract screening: records were screened against the criteria below, and the reference lists of the retained records and of a systematic review of trust assessment in HRI [1] were followed for further candidates. Screening and coding were carried out by the first author and checked by the second. This is a targeted retrieval and not a systematic review: it is reported so that the boundary of the set is reproducible, not so that the set can be read as exhaustive.

Inclusion and exclusion. We restrict the scope to tasks in which the operator and the robot act on the same objects in one workspace, with or without contact: contact tasks, co-manipulation, shared load, object handover, and co-located work on shared parts. Contact is therefore a parameter we record, not a condition of entry, which is what makes the reading in Section IV possible. Robots acting as purely social or conversational agents fall outside this scope, since the nature of trust and the available signals differ from physical contact [1]. We include a record when its central task is of this kind and it reports some measurement of trust, whether by questionnaire, scale, or behavior-derived label. We exclude records that measure only adjacent constructs such as affect, comfort, stress, or cognitive load, even in physical tasks [18], [19], and conversational assistance scenarios [20].

The set is small by construction. We keep the works that set out to measure trust in a physical collaboration and that document the task, the instrument, and the recorded signals in enough detail to place them on the matrix, several of which were published specifically to release a trust dataset [4]–[6]. We leave aside studies where the trust data is incidental, undocumented, or not provided as a labeled target, since these cannot be read along the same parameters. The pattern therefore describes the collections that meet these criteria rather than the field in full.

Terminology. The sources label comparable arrangements differently, so we fix a vocabulary and apply it in place of the source wording. *Physical collaboration* covers any task in which the operator and the robot act on the same objects in one workspace. Within it, the *coupling* takes four values: *sustained*, when the two hold a common object or contact point for a whole repetition; *transfer only*, when contact exists at a handover and not otherwise; *none (co-located)*, when they act on the same parts without touching at the same time; and *none (no contact)*, when the robot is monitored or rendered rather than touched. A parenthetical qualifier after a coupling value names the phase of the task in which that contact occurs, and narrows the window rather than adding a fifth value. The *environment* is *real*, *mixed reality* when a virtual robot is rendered into a real workspace, or *virtual reality* when the whole scene is rendered. Where a source uses more than one of these terms for one setting, Table I records the value that follows from the definitions above. We use pHRI for the field, and *collection* for any of the eight records placed on the matrix, whether or not its data is documented as obtainable.

Cell coding. Every cell of Tables I and II carries the status of its entry. A cell that names a task, instrument, sensor, or number is documented in the cited publication, which is the level of evidence for the matrix as a whole: the entries state what the sources report, not what we verified in the data. Absences are separated into three cases. The mark *n.rec.* applies where the source enumerates its recording setup and the modality is not in that enumeration; reading an enumeration as complete is an assumption of ours rather than a statement by the authors. The mark *n.r.* applies where the source does not report the parameter at all, so its value is unknown, and in the availability column it marks the collections that state no route to the data. The mark *n/a* applies where the parameter has no referent, as with an interaction force in a task without contact. The distinction matters because an absence of the first kind is a property of the collection and one of the second kind is a property of its documentation.

III. The Coverage Parameters

Each parameter is a measurement axis established in the trust literature, listed with the reference that grounds it, and is

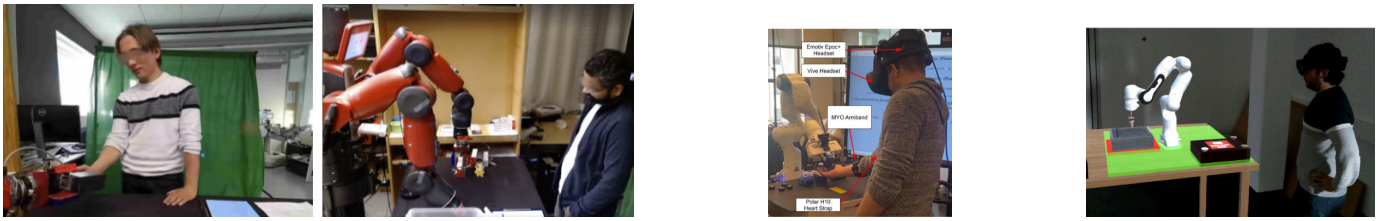


Figure 1. Physical tasks across three collections. **Left:** a human and a Baxter robot fill a shelf by exchanging objects (REFLEX [6]). **Center:** an instrumented handover with a real Franka Emika robot, the operator wearing EEG, EMG, ECG, and eye-tracking sensors (TrustBase [4]). **Right:** a battery disassembly task in *mixed reality*, where the Franka Panda and workpieces are virtual, rendered in the operator's head-mounted display rather than physically present (Alhaji et al. [10]). Images reproduced from the respective papers.

Table I

CONTEXT AND SUBJECTIVE MEASUREMENT PROTOCOL OF THE EIGHT PHRI COLLECTIONS. ENV.: ENVIRONMENT, REAL (R), MIXED REALITY (MR), OR VIRTUAL REALITY (VR). COUPLING: MECHANICAL COUPLING BETWEEN OPERATOR AND ROBOT, USING THE FOUR VALUES DEFINED IN SECTION II; A PARENTHETICAL QUALIFIER NAMES THE PHASE OF THE TASK IN WHICH THAT CONTACT OCCURS. AVAIL.: WHETHER THE SOURCE DOCUMENTS HOW THE DATA CAN BE OBTAINED. N.R.: NOT REPORTED BY THE SOURCE.

Collection	Env.	Coupling	Task (robot)	N	Instrument	Time scale	Failure induction	Avail.
TrustBase [4]	R	Transfer only	Handover (Franka Emika)	65	7-pt Likert (single item)	Per interaction	No (varies speed, height, distance)	n.r.
Benchmark Movement [5]	R	Sustained (transport)	Transport + draping (Kuka KR 300)	20	Trust Perception Scale-HRI	Per task (pre/post)	No failures (speed manipulated)	Open (Zenodo, CC BY 4.0)
Campagna et al. [3]	R	None (co-located)	Chemical mixing (Franka Emika)	14	Preference-based label	Per task iteration	No (varies trajectory)	n.r.
REFLEX [6]	R	Transfer only (failure resolution)	Object placement (Baxter)	55	Single item, 8-item scale (Hoffman)	Per round	Yes (failures + explanations)	Open (Zenodo)
Hopko / Zhang [7], [8]	R	None (co-located)	Assembly (UR10)	38	12-item trust survey; 1-item 7-pt	Per block and per trial	Yes (100% vs 76% reliability)	n.r.
Green & Iqbal [9]	R	None (no contact)	Supervisory (Franka Panda)	30	Muir survey, 5-pt Likert	Per task	Yes (success vs failure)	Announced, no deposit
Alhaji et al. [10]	MR	None (no contact)	Battery disassembly (simulated Panda)	32	Multi-factor (Muir), 7-pt	Post-hoc per phase	Yes (reliable / unreliable)	On request
Chauhan et al. [11]	VR	None (no contact)	Bricklaying (simulated UGV)	11	TPS-HRI + 5-pt dynamic item	Per task	No (varies workspace, interaction, speed)	On request

Table II

OBJECTIVE SIGNAL MODALITIES RECORDED BY EACH COLLECTION. A CELL NAMES THE SENSOR OR METHOD WHEN THE SOURCE DOCUMENTS IT. ABSENCES ARE SEPARATED BY STATUS: *n.rec.* MARKS A MODALITY THAT THE SOURCE DOES NOT LIST AMONG ITS RECORDED SIGNALS, *n.r.* A PARAMETER THE SOURCE DOES NOT REPORT AT ALL, AND *n/a* A MODALITY WITH NO REFERENT IN THAT SETTING. NO CELL IN THE F/T COLUMN NAMES A SENSOR.

Collection	F/T	Video	Eye / gaze	Body pose	Affect	Neural	Peripheral physiol.
TrustBase [4]	n.rec.	n.rec.	9 eye signals (Vive)	Forearm IMU (MYO)	n.rec.	EEG (14 ch)	EMG, ECG
Benchmark Movement [5]	n.rec.	n.rec.	n.rec.	Xsens (17 IMU, 60 Hz)	n.rec.	n.rec.	n.rec.
Campagna et al. [3]	n/a	n.rec.	Head-pose gaze proxy	Xsens (mocap)	n.rec.	n.rec.	n.rec.
REFLEX [6]	n.rec.	2 cameras	OpenFace	MediaPipe	Hume, FAUs	n.rec.	n.rec.
Hopko / Zhang [7], [8]	n/a	Egocentric (eye tracker)	Tobii Pro Glasses 2	n.rec.	n.rec.	fNIRS (46 ch)	ECG (HR, HRV)
Green & Iqbal [9]	n/a	Facial	Pupil Invisible	n.rec.	OpenFace AUs	n.rec.	BVP, EDA, temp.
Alhaji et al. [10]	n/a	n.rec.	n.rec.	n.rec.	n.rec.	n.rec.	n.rec.
Chauhan et al. [11]	n/a	n.rec.	n.rec.	n.rec.	Valence, arousal (EEG-derived)	EEG (Emotiv EPOC X)	EDA, temp., HRV

F/T: force and torque. AU: facial action unit. BVP: blood volume pulse. ch: channels. ECG: electrocardiography. EDA: electrodermal activity, also reported as GSR (galvanic skin response). EEG: electroencephalography. EMG: electromyography. fNIRS: functional near-infrared spectroscopy. HR: heart rate. HRV: heart rate variability. IMU: inertial measurement unit. temp.: skin temperature.

kept only when it varies across the collections. The parameters fall into six groups, split across Tables I and II.

Context. The environment, the task, the robot, and the number of participants (N) fix the setting in which trust is observed. They are not coverage claims, but N belongs to the minimum information expected of a measurement study, and the task and robot fix the physical regime.

Instrument and time scale. The instrument separates validated scales, ad hoc items, and behavior-derived labels; the time scale separates post-session, per-trial, and continuous labels, the last of which no collection in the set supplies. Validated trust scales for HRI exist [14], [15], and the same literature notes that a questionnaire misses variation within a task, which is what motivates real-time measurement [1], [3],

[8]. The instrument and the time scale are not independent, since a scale administered at one point per session cannot resolve a change inside the session.

Failure induction. This parameter records whether the collection deliberately manipulates trust, most often by provoking robot failures or violations, but also by varying a trust-relevant factor such as robot speed [5]. Trust is easier to observe around such manipulations; the effect of failures on real-time trust and the effect of a faulty robot on cooperation are established results [10], [21], [22].

Mechanical coupling. When the human and the robot are mechanically coupled, through a shared object or a contact point, forces and torques are exchanged at that point, and controllers already read them as the most direct available measure of the mechanical interaction [23], [24]. Whether they also carry information about an operator's trust is a separate question, taken up in Section V. Coupling of this kind is not present in every setting the literature places under physical collaboration, and it bounds the window in which an interaction force exists at all, so we record it as a parameter of its own.

Captured signal modality. This group separates force and torque from visual and physiological signals. Video, gaze, pose, and skin conductance are indirect indicators of the operator's state, and we keep the distinction from the mechanical channel as a parameter rather than folding the two together.

Availability. This parameter records whether the source documents a route to the data, since a collection that reports a measurement without releasing it constrains what a later study can reuse. It separates an open deposit, a statement that the data is available on request, an announcement with no deposit, and silence.

IV. The Coverage Matrix

Tables I and II bring the eight collections together. The two studies of the same cohort are kept as a single row because they share task and manipulation across two venues [7], [8].

Four patterns stand out. First, and in the coupling column rather than in any measurement: only three of the eight settings place the operator in contact with the robot, and only one holds the two on a shared load for a whole repetition. A set assembled under the heading of physical collaboration turns out to contain physical contact in a minority of its cases. No row names a force or torque sensor, which follows in part from this, since in five of the eight settings there is no common mechanical path on which such a sensor could read. Second, the instrument column is widely scattered: a validated scale [5], a single item [4], a multi-factor questionnaire [10], an item embedded in a satisfaction scale [6], a preference-based label [3], the Muir survey [9], and a validated scale adjusted by a single dynamic item [11]. Two instruments recur, the Trust Perception Scale-HRI [5], [11] and items from the Muir survey [9], [10], and in each case the two collections that share them differ in the sampling point or in how the items were assembled. The time scale is equally scattered, from a label taken per interaction or per trial to one taken per round, per phase, or per task iteration, with no dominant choice and no collection labeling trust continuously. Third, the objective modalities are scattered as well: visual and physiological signals are common, no two collections share a combination, and the most direct channel appears in none.

Fourth, the amount of data differs as sharply as its type, which Table II does not show. Cohort size runs from eleven to sixty-five, as the N column records, and cuts against the density of signal per participant. Small cohorts tend to carry dense signals and larger cohorts sparser ones, and none is both large in participants and rich in modalities. For training or evaluating a model, what matters is the number of trust labels rather than the raw signal volume, and this number is often small: a per-round or per-task label yields on the order of one to a few hundred labeled instances even when the underlying streams are large [6], [9]. The benchmark movement set is the sharpest case: roughly 950 task repetitions of 60 Hz kinematics across 17 IMUs are labeled by only two trust scores per participant, one per task, on the order of forty trust labels for the whole set [5].

Reading the matrix by cell status adds a fifth pattern: part of what looks like coverage is a property of the documentation rather than of the data. Two collections are deposited openly [5], [6], two state that the data is available from the authors on request [10], [11], and one announces a release without a deposit [9]; the remaining three report a measurement without documenting any route to it, so the matrix is in part a map of published studies rather than of retrievable material. The absences in the F/T column are of two kinds, five settings where the signal has no referent and three where it does but is not listed, and only the second is a design choice a later collection could reverse. Sample sizes are not uniformly defined either, since TrustBase describes its contents as 65 user studies in one place and as 65 participants in another [4]. The largest divergence sits inside one row: the two publications on a single cohort sample trust with different instruments at different rates, a 12-item survey per reliability block [7] and a single seven-point item per trial [8], so even one cohort does not fix an instrument or a time scale.

V. Gaps and Open Problems

The scarcity of contact just described, together with the empty cells of Tables I and II, points to five directions, each grounded in the collections reviewed here or in the adjacent literature. They are stated as gaps in this set, not as properties of the field as a whole.

No common measurement protocol. With the instrument and time-scale columns this scattered, two collections are rarely comparable: what one measures with a validated scale, another records as a single item or as a stated preference between trajectories. This matches the limitation that reviews attribute to the dominance of the questionnaire [1]. A practical step is to report which instrument was used, at what point in the task, and how the resulting score was propagated to the recorded signals, so that datasets can be read against one another, rather than to impose a single scale on the field.

One label, many observations. Trust is recorded once or twice per session. The signals meant to reveal it are recorded many times per second. To train a model, the single score has to be copied onto every observation it is supposed to cover, and that copying is not a preprocessing detail: it asserts that trust held constant across the whole interval, and it produces many near-identical examples that carry one label. Three of the collections do this, in different ways, counting the model trained on the benchmark movement set [25]. A five-point rating given after a four-minute task is copied onto every overlapping thirty-second window drawn from that task [9]. A per-task score is copied onto every one-second sample of physiology [11]. A questionnaire score is split at the sample median, and in the variant with the highest reported accuracy the training set keeps only the first five repetitions of the low-trust participants and the last five of the high-trust ones, on the assumption that trust rises within a session [25]. A fourth avoids the propagation, averaging each indicator over one task iteration, but its label is a stated preference between two trajectories rather than an absolute trust value, as the authors note, and its reported accuracy is computed after the 210 real comparisons are expanded to 840 by synthetic sampling [3].

The copying matters at evaluation. When folds are drawn at random, consecutive observations from the same participant fall on both sides of the split, and a model can score well by recognizing the participant rather than the trust level. One study reports both splits on the same data and the same architecture: accuracy is 0.92 and 0.90 with random folds, and 0.45 and 0.55 once the validation participants are held out [25]. Elsewhere in the set, a classification accuracy of 0.975 under ten-fold cross-validation [9] and a regression R^2 of 0.985 under five-fold cross-validation [11] are reported, neither described as participant-wise, and whether those folds separated participants cannot be told from the text. One of these targets a model personalized to an individual baseline, where validating within a participant is defensible; the qualification is that the figure then describes consistency within a session rather than detection of trust. Trust has no external ground truth against which the difference could be settled. Reporting how the folds were drawn, together with a majority-class baseline, would separate the two readings at no experimental cost. Until the measurement is standardized, such results are best read as evidence that a signal correlates with the label built from it, and the step from there to a robot adjusting its speed or proximity online is larger than the accuracies suggest.

No force and torque channel. None of the eight collections records force or torque, and the coupling column shows that the absence is not evenly distributed. In five of the eight settings there is no interaction force to record. The mixed- and virtual-reality tasks have no physical contact [10], [11], in the shared-workspace and supervisory settings the robot delivers parts to the operator's side of the workbench or is only monitored [7]–[9], and in the chemical-mixing task the operator holds one beaker while the robot pours into it from another, so the two never share a contact point [3]. Where contact does occur it is confined to the instant of transfer, at the handover itself [4] and at the failure-resolution step in which the operator places an object into the gripper [6]. Only the transport task of the benchmark movement set holds the human and the robot on the same load for the duration of a repetition, each carrying one end of a textile piece [5], and there only the human kinematics were logged. That coupling runs through a compliant textile, which transmits tension rather than the full wrench, so even the one available case bounds what the channel could carry.

For the settings with real contact the absence is not a technical limitation: handover failure-detection datasets already record force, torque, and gripper position [26], and trust has already been used to modulate impedance in co-manipulation [27], so the signal is both collectable with hardware these setups already run and usable by the controllers already in the loop. Trust also changes what the operator does physically [1], so a mechanical trace, if one exists, may surface in quantities these controllers already read [23], [24].

Any such trace would be superimposed on the mechanical response of the task itself, since the same forces are set by the payload, the commanded trajectory, the robot's impedance, external disturbances, and the operator's biomechanics, and any of these can change the measured wrench while trust does not [23], [24]. A study of this channel would therefore have to hold those conditions constant across trust conditions, or model them, before a difference in force could be attributed to trust, and a null result under such controls would be as informative as a positive one. The gap is therefore twofold. The signal is absent from these collections, and the setting in which it would carry the most information is itself nearly absent, so whether the mechanical channel carries trust information as a recorded signal paired with trust labels remains untested rather than ruled out.

Anonymization as an unused variable. How visible a participant is, from identifiable video to an anonymized silhouette, has been studied at length in computer vision, where it trades off against the utility of the data [28]–[30]. This trade-off falls directly on the trust-relevant channels: the facial expression, gaze, and body-pose signals that these collections record as indicators of the operator's state are the same signals that identify the participant, so anonymizing the data degrades precisely the cues a trust model reads. Blurring and similar traditional methods lower model performance, more so for the full body, while synthesis-based methods recover much of the loss [28]. The same trade-off shows up in keypoint detection, instance segmentation, and face detection, and re-identification often still succeeds when recognition rates are low [29]. No pHRI trust collection in this set varies this level of visibility. Only the collections that record raw video could be anonymized at all (Table II), and the few that do fix a single level, while the benchmark movement set, recording only IMU-derived skeleton kinematics, is anonymous by construction [5]. The set spans both endpoints of the axis but never varies it inside one collection, even though the level chosen decides whether the data can be shared at all. The trade-off above was measured for models trained on the anonymized data; for vision-language

models applied zero-shot, where no retraining absorbs the distortion, the effect of the anonymization level on a trust judgment is unstudied.

Fragmented construct and environment. The physical collaboration literature is split across constructs: several signal-rich datasets measure affect, comfort, stress, or cognitive load rather than trust [18], [19], which leaves trust labels in physical collaboration scarce. The environment varies too: some studies use a real robot, some mixed reality [10], some virtual reality [11], so the condition of physical collaboration is itself not held constant, as the environment and coupling columns of Table I record. The measured construct and the collection environment are part of the protocol that is still missing.

VI. Conclusion

This paper compared eight collections for physical human-robot collaboration by how each one measures trust. They were retrieved through a documented search, placed on a coverage matrix built from measurement parameters drawn from the trust literature, and read cell by cell with the status of each entry recorded. Physical contact occurs in three of the eight arrangements and sustained contact in one, so the condition the field is named after is largely absent from the data it has collected. The collections are individually rich, yet this set does not share a way to measure trust. Instruments range from validated scales to single items and preference-based labels, and time scales from post-session to per-trial, with none continuous, so two collections are rarely comparable, and in one case the two publications on a single cohort do not agree with each other. Visual and physiological signals are common but never co-occur across collections; no collection reports force or torque, only three of the eight settings put the operator in contact with the robot at all, and the sustained coupling under which that channel would be most informative is present in one setting only, so its relation to trust is untested and would in any case have to be separated from the payload, trajectory, and disturbance components of the same signal; and the level of participant anonymization is fixed rather than varied. The amount of data is uneven as well, and the count that matters for training a model, the number of trust labels, stays low even when the raw streams are large. That scarcity is then absorbed by copying one score across many observations, a step that also breaks the usual way of testing a model. Only two of the eight are openly deposited and three document no route to their data at all, so part of the matrix describes published studies rather than reusable material. The construct and the collaboration environment are fragmented too, since some collections substitute affect or comfort for trust and others replace real contact with mixed or virtual reality.

Laying these collections on one matrix gives a compact reading of what trust data for physical collaboration is documented and where it runs out, so that a new study can be positioned against the gaps rather than against a single prior work. Four experimental priorities follow from the empty cells, in the order in which each unblocks the next. First, a reporting protocol: state the instrument, the point in the task at which trust is sampled, how that score was propagated to the recorded signals, how the evaluation folds were drawn, and the route to the data, since without it the remaining three cannot be compared across studies. Second, a collection that records force, torque, and trust labels on a common clock under sustained contact, with the payload and the commanded trajectory held constant across trust conditions, so that a mechanical trace of trust can be looked for and, if absent, reported as absent. Third, labels at more than one time scale within one cohort, which is the only design that shows whether a per-session and a per-trial measurement of trust track the same quantity. Fourth, an anonymization level varied as a controlled factor within a single collection, which would measure its cost to a trust judgment, including for vision-language models applied zero-shot, where no retraining absorbs the distortion. The fifth gap, the fragmentation of construct and environment, is not a separate experiment but a consequence of the first: once the instrument, the sampling point, and the setting are reported in a common form, collections that measure different things can at least be told apart before they are compared. A journal extension will widen the scope beyond the parameters used here, toward other dimensions of trust and uncertainty in physical interaction and a larger, more systematic body of collections.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil, process Number 2025/09390-3 (Marcha+).

References

[1] G. Campagna and M. Rehm, "A systematic review of trust assessments in human-robot interaction," *ACM Trans. Hum.-Robot Interact.*, vol. 14, no. 2, art. 30, 2025.

[2] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[3] G. Campagna *et al.*, "Estimating trust in human-robot collaboration through behavioral indicators and explainability," *IEEE Robot. Autom. Lett.*, vol. 10, no. 10, pp. 10218–10225, 2025.

[4] J. Parron, T. T. Nguyen, and W. Wang, "Development of a multimodal trust database in human-robot collaborative contexts," in *Proc. IEEE 14th Annu. Ubiquitous Computing, Electronics & Mobile Communication Conf. (UEMCON)*, 2023, pp. 601–605.

[5] M. Rehm, K. Hald, and I. Pontikis, "Benchmark movement data set for trust assessment in human robot collaboration," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2024, pp. 934–938.

[6] P. Khanna *et al.*, "REFLEX dataset: A multimodal dataset of human reactions to robot failures and explanations," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2025, pp. 1032–1036.

[7] S. K. Hopko *et al.*, "Brain-behavior relationships of trust in shared space human-robot collaboration," *ACM Trans. Hum.-Robot Interact.*, vol. 13, no. 1, pp. 1–23, 2024.

[8] Y. Zhang *et al.*, "In gaze we trust: Comparing eye tracking, self-report, and physiological indicators of dynamic trust during HRI," in *Companion ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2024.

[9] H. N. Green and T. Iqbal, "Using physiological measures, gaze, and facial expressions to model human trust in a robot partner," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2025, pp. 11868–11875.

[10] B. Alhaji, M. Prilla, and A. Rausch, "Trust dynamics and verbal assurances in human robot physical collaboration," *Frontiers in Artificial Intelligence*, vol. 4, art. 703504, 2021.

[11] H. Chauhan, Y. Jang, and I. Jeong, "Predicting human trust in human-robot collaborations using machine learning and psychophysiological responses," *Advanced Engineering Informatics*, vol. 62, art. 102720, 2024.

[12] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[13] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[14] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the Trust Perception Scale-HRI," in *Robust Intelligence and Trust in Autonomous Systems*, Springer, 2016, pp. 191–218.

[15] R. E. Yagoda and D. J. Gillan, "You want me to trust a robot? The development of a human-robot interaction trust scale," *Int. J. Social Robotics*, vol. 4, no. 3, pp. 235–248, 2012.

[16] M. Madsen and S. Gregor, "Measuring human-computer trust," in *Proc. 11th Australasian Conf. Information Systems*, 2000, pp. 6–8.

[17] A. Alzahrani and M. I. Ahmad, "Real-time trust measurement in human-robot interaction: Insights from physiological behaviours," in *Companion Proc. ACM Int. Conf. Multimodal Interaction (ICMI)*, 2024.

[18] A. Bussolan *et al.*, "MultiPhysio-HRC: Multimodal physiological signals dataset for industrial human-robot collaboration," arXiv preprint arXiv:2510.00703, 2025.

[19] B. Gucsi, N. T. V. Tuyen, B. Chu, D. Tarapore, and L. Tran-Thanh, "HRI-SENSE: A multimodal dataset on social and emotional responses to robot behaviour," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2025.

[20] Y. Batra *et al.*, "RFM-HRI: A multimodal dataset of medical robot failure, user reaction and recovery preferences for item retrieval tasks," arXiv preprint arXiv:2603.05641, 2026.

[21] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2013, pp. 251–258.

[22] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, 2015, pp. 1–8.

[23] G. A. Cardona, S. S. Kumbhar, and P. Artemiadis, "Toward seamless physical human-humanoid interaction: Insights from control, intent, and modeling," arXiv preprint arXiv:2512.07765, 2025.

[24] A. Al-Yacoub *et al.*, "Improving human robot collaboration through force/torque based learning for object manipulation," *Robotics and Computer-Integrated Manufacturing*, vol. 69, 2021.

[25] M. Rehm, I. Pontikis, and K. Hald, "Automatic trust estimation from movement data in industrial human-robot collaboration based on deep learning," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 11245–11251.

[26] S. Thoduka, N. Hochgeschwender, J. Gall, and P. G. Plöger, "A multimodal handover failure detection dataset and baselines," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024.

[27] Z. Liao and Y. Wang, "Trust-based variable impedance control of human-robot cooperative manipulation," *Robotics and Computer-Integrated Manufacturing*, vol. 88, art. 102730, 2024.

[28] H. Hukkelås and F. Lindseth, "Does image anonymization impact computer vision training?," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 140–150.

[29] R. Stenger, S. Busse, J. Sander, T. Eisenbarth, and S. Fudickar, "Evaluating the impact of face anonymization methods on computer vision tasks: A trade-off between privacy and utility," *IEEE Access*, 2024.

[30] J. H. Lee and S. J. You, "Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision," *IEEE Access*, vol. 12, 2024.

Design and Implementation of PMSM Motor, Driver and Field Oriented Control Algorithm for Upper Limb Rehabilitation

Lucas Soares Leal
Faculty of Electric Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
lucassoaresleal@ufu.br, 0009-0008-1347-1588

Alcimar Barbosa Soares
Faculty of Electric Engineering (FEELT)
Federal University of Uberlândia (UFU)
Uberlândia, Brazil
alcimar@ufu.br, 0000-0003-1100-3533

***Abstract*—This work describes the use of a field oriented control (FOC) model to improve the performance of external-rotor Permanent Magnet Synchronous Motors towards the development of robotic rehabilitation devices. The proposed system encompasses the design of an external-rotor permanent magnet synchronous motor, a dedicated motor driver, and a FOC algorithm, constituting the actuator module upon which a complete robotic rehabilitation platform can be built. The simulation results demonstrate that FOC produces lower torque ripple than six-step commutation under equivalent operating conditions, supporting its adoption as the control strategy for applications requiring smooth and precise torque delivery.**

***Index Terms*—Upper-Limb Rehabilitation, FOC, Control Systems, Embedded Software.**

I. Introduction

The loss or impairment of upper limb motor function represents one of the main challenges in the field of physical rehabilitation, directly impacting individuals independence and quality of life. The World Health Organization (WHO) estimates that 2.4 billion people, with projections of 3.35 billion by 2050, require some form of rehabilitative intervention, with motor dysfunction resulting from conditions such as stroke being one of the leading causes of functional disability [1]. In this context, the restoration of mobility and functionality of the upper limbs becomes a central goal in therapeutic processes.

Conventional rehabilitation, based on repetitive and supervised exercises, faces limitations related to treatment variability, patient fatigue, and a shortage of specialized professionals. As an alternative, assistive robotics has emerged as a promising tool, enabling more consistent, intensive, and measurable therapies [2], [3]. These systems allow for the controlled application of forces and movements, as well as the acquisition of relevant data such as position, velocity, and torque, which are essential for assessing patients progress.

Among the most commonly used control strategies, Field-Oriented Control (FOC) is characterized by its ability to provide continuous, decoupled control of the motor currents, offering greater precision, superior dynamic response, and smoother torque generation—desirable characteristics in rehabilitation applications that require precise torque control and natural-like movement.

The generation of appropriate torque is a critical factor in robotic rehabilitation systems, as it directly influences physical interaction with the patient. More advanced control strategies can contribute to better adaptation to the required level of assistance, while adhering to fundamental principles of rehabilitation, such as individualization, intensity, and progression of therapeutic effort [2]. Actuators available in industry vary in configuration, some direct drive with Six-Steps as its commutation control strategy at a lower price (K-Tech MF6015) and FOC controlled actuators with reduction gear boxes (CubeMars AK10-9)

Given this context, it is important to investigate and compare the performance of different control strategies applied to robotic actuators, especially with regard to torque generation. Thus, this work aims to conduct an analysis of field-oriented control (FOC), evaluating its characteristics and implications for the performance of robotic systems, as well as the development and testing of a prototype designed for upper limb rehabilitation.

II. Literature review

A. Traditional Rehabilitation

Traditional Rehabilitation is primarily based on repetitive exercises supervised by physiotherapists and occupational therapists, aiming to restore movement, muscular strength, and functional independence. Among different strategies, task-specific training stands out as one of the most effective for upper limb rehabilitation. [5], [6]

Task-specific training is based on the repetitive execution of daily activities. It differs from motor deficit training by exploring functional abilities instead of focusing solely on muscle strengthening and increased range of motion. According to Hubbard et al., as cited by Rozevink et al., task-specific training should involve activities relevant to the patient, performed repeatedly, presented in different contexts, organized to reconstruct complete functional tasks, and accompanied by

performance feedback. These principles promote neuroplasticity and motor relearning. [5]

B. Robotic Rehabilitation

Advances in motor assistance technologies have enabled the development of robotic rehabilitation, in which electromechanical devices assist or partially perform upper limb movements during therapeutic sessions. These devices have emerged as an alternative for increasing the intensity, repeatability, and precision of rehabilitation exercises. [5]

Robotic devices for upper limb rehabilitation can be classified into two main categories: end-effector systems and exoskeletons. End-effector devices interact with the distal end of the limb, allowing for relatively free movement of the proximal joints, while exoskeletons have structures aligned with the arm joints, providing support and assistance directly to joint movements. [1]

In addition, these systems can operate in different training modes, including passive, active-assisted, assistive, and active modes. In assistive modes, the device provides supplemental torque to assist the patient's movement, while in active modes, the individual performs the movement independently, using the robot primarily as a monitoring or interaction tool. [5]

One of the main advantages of robot-assisted rehabilitation is its ability to provide a large number of repetitions with precise control over training conditions. In addition, these devices can automatically adjust the difficulty of exercises, record performance metrics, and provide real-time feedback to the user. These features encourage active patient participation and enable the implementation of individualized programs. [5]

The results presented by Rozevink et al. indicate that both robot-assisted rehabilitation and conventional task-based rehabilitation are effective in improving upper limb function after stroke. In the subacute phase, assistive devices yielded superior results in reducing motor deficits, whereas in the chronic phase, the benefits were similar to those obtained with conventional therapy. [5]

The Figure 1 represents the MIT-Manus, one of the designs for upper limb rehabilitation and the one that inspired the work. The MIT-Manus and H-Man are both end-effectors and can perform activities in all training modes previously mentioned.

C. Permanent Magnet Synchronous Motor (PMSM)

The operating principle of brushless DC motors is similar to that of brushed DC motors, in which rotational motion is generated through the attraction and repulsion of magnetic poles. However, these motors differ in the way a single DC voltage is converted into a three-phase voltage.

The DC current passing through the rotor coil windings generates an electromagnet with poles. These magnetic poles of the rotor interact with the poles of the permanent magnet fixed to the stator to cause the rotor to rotate. Commutation occurs mechanically every half-turn via the electrical contacts of the brushes. Due to this physical contact, the brushes wear out over time and cause fluctuations in torque.

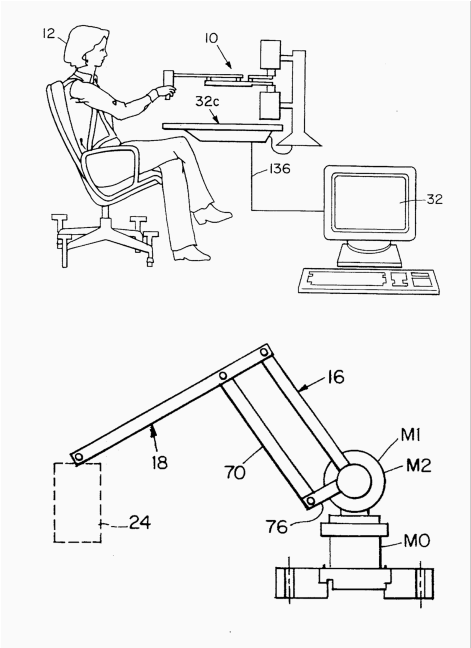


Fig. 1. Representation of the planar rehabilitation platform MIT-Manus, consisting of two actuators and a polar coordination configuration. Part of patent document US5466213A

The BLDC motor replaces mechanical commutation with electronic commutation. Compared to a brushed DC motor, permanent magnets can be installed on the rotor, while the coil windings form the stator.

BLDC and PMSM motors both operate synchronously; that is, the rotor rotates at the same speed as the stator's magnetic field. The main difference lies in the back-EMF, which determines the optimal control algorithm.

D. Field Oriented Control (FOC)

Field Oriented Control (FOC) is a control technique commonly used in induction machines and permanent magnet synchronous machines (PMSM) to achieve good control capability over the full torque and speed range. The block diagram in Figure 2 shows the architecture of the FOC control technique.

Clark and Park transforms a stationary frame abc into a rotating frame dq, the current controller consists of a PI controller for iq and id currents in the orthogonal rotating reference frame and a space vector modulator to convert a steady-state frame into a PWM signal to stator windings.

Hardware wise, the FOC uses a three-phase inverter circuit that consists of high and low power MOSFETs for each phase of the motor. The upper, lower, and no-switching states determine the phase state, keeping the motor at a constant speed. [7]–[10]

1) Clark and Park Transformation:

The Clark transformation (Equation 1) converts the time-domain components of a three-phase system into two components in a steady-state frame (Figure 3) [11].

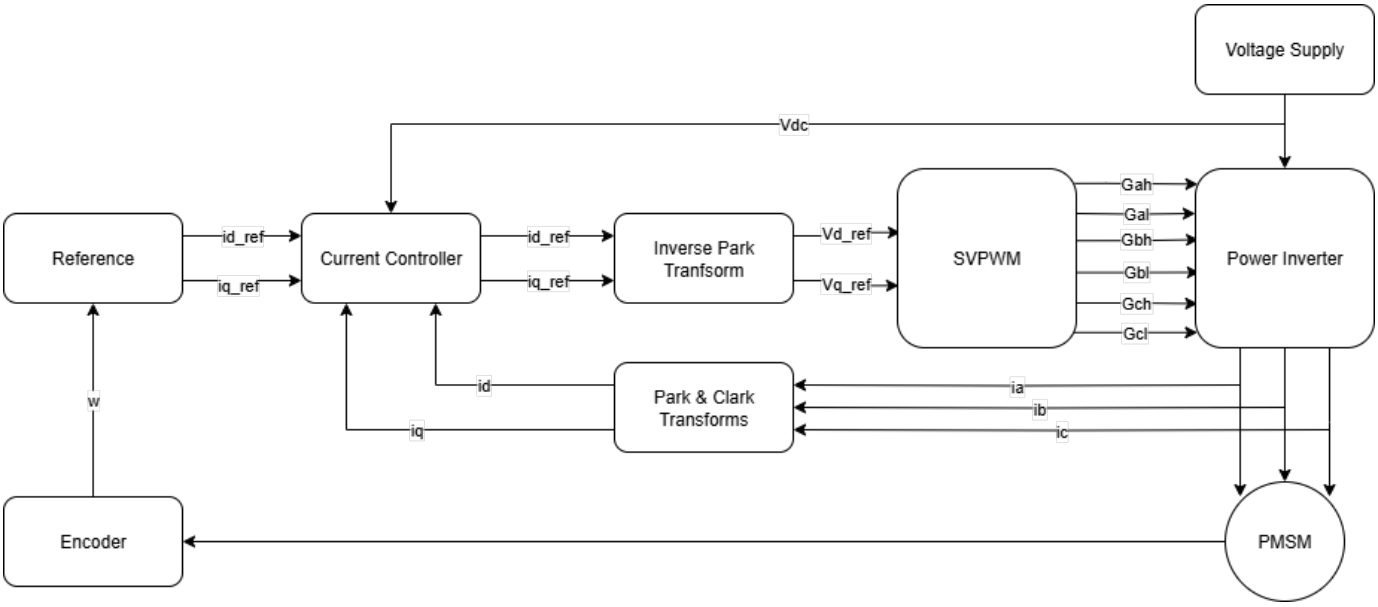


Fig. 2. Control block of Field Oriented Control architecture. Diagram made in draw.io

$$\begin{bmatrix} f_\alpha \\ f_\beta \\ f_0 \end{bmatrix} = \frac{2}{3} \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} f_a \\ f_b \\ f_c \end{bmatrix} \tag{1}$$

For a balanced system with motors, the zero-sequence component is always zero (Equation 2)

$$eu_a + eu_b + eu_c = 0 \;\Rightarrow\; e_{uc} = -\left(eu_a + eu_b\right) \tag{2}$$

This means that only two of the phases need to be monitored.

To put it simply (Equation 3):

$$\begin{bmatrix} f_\alpha \\ f_\beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} f_a \\ f_b \end{bmatrix} \tag{3}$$

Where:

- f_a, f_b, and f_c are three-phase components in the abc panel
- f_α and f_β are orthogonal two-phase components in the steady-state $\alpha\beta$ frame
- f_0 is the zero in the steady-state $\alpha\beta$

Park's transformation converts the two components of the $\alpha\beta$ reference frame into an orthogonal rotating dq reference frame (Figure 3)

Aligning axis f_α with axis d (Equation 4)

$$\begin{bmatrix} f_d \\ f_q \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} f_\alpha \\ f_\beta \end{bmatrix} \tag{4}$$

Where:

- f_α and f_β are components of the Clark transform
- f_d and f_q are orthogonal components in the dq reference frame

2) Space Vector Modulation (SVM):

SVM is a common tool when it comes to FOC in induction motors and PMSMs. This technique is responsible for generating PWM signals for inverter commutation, which in turn controls the motor [8], [12].

Using the circuit shown above (II-D) and the logic circuit presented, each configuration results in simple space vectors that describe the magnitude and direction of the space vector.

The switching states corresponding to the basic spatial vectors (direction) and the zero vectors (magnitude) are combined to approximate a voltage vector of any magnitude, at any position, within the spatial vector hexagon. For example, for each pulse width modulation (PWM) period, the reference vector 'Uref' is calculated using a switching sequence of two adjacent spatial vectors for a specified time period and a zero vector (U7 and U8) for the remainder of the period [12].

By controlling the switching sequence and, consequently, the duty cycle, any voltage vector with varying magnitude and direction can be achieved for each PWM period. The objective of the SVM technique is to generate switching sequences that correspond to the reference voltage vector in order to obtain a continuously rotating spatial vector [12].

For each PWM period, using the voltage vector as the input reference, the SVM algorithm:

- Calculates the turn-on and turn-off times based on the reference voltage vector
- Use transition times to generate double-hump modulation waveforms
- It uses switching times to generate gate pulses suitable for the inverter switches

The waveform of the modulation signal generated by a double-hump structure maximizes the utilization of the available bus voltage (Figure 4) [12].

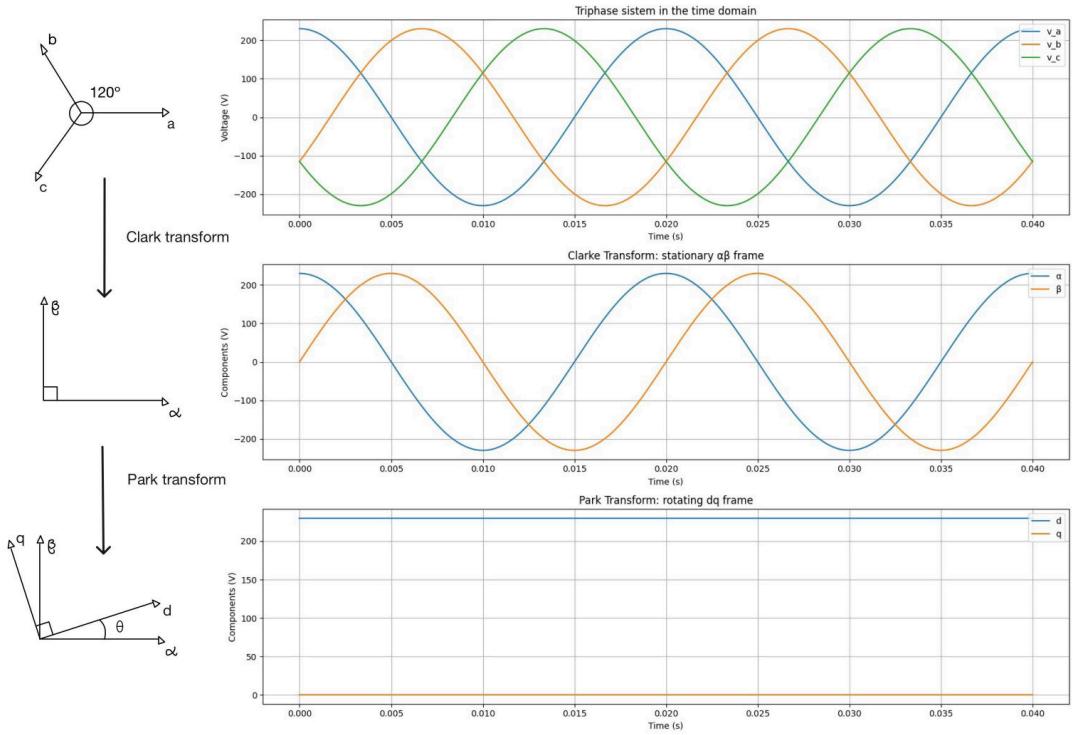


Fig. 3. Clark and Park Transforms representation in motor waveform. Made with python and GoodNotes

3) PI Current Control:

After the Clark and Park transformations, the three-phase stator currents are represented as two DC-like quantities, i_d and i_q, in the rotating reference frame. This characteristic allows for the use of conventional Proportional-Integral (PI) controllers to regulate the motor currents, since PI controllers are particularly effective at tracking constant reference signals with zero steady-state error, unlike the sinusoidal references required when controlling currents directly in the stationary abc frame.

As illustrated in the FOC block diagram (Figure 2), two independent PI loops are employed: one regulating the direct-axis current i_d and another regulating the quadrature-axis current i_q. For surface-mounted PMSMs, the reluctance torque component is negligible, so the electromagnetic torque is approximately proportional to i_q alone; the reference i_d^{ref} is therefore commonly set to zero, minimizing resistive losses whilei_q^{ref} is set according to the desired torque output. Each PI controller computes a voltage reference,v_d^{ref} and v_q^{ref}, based on the error between the reference and measured currents, which is then forwarded to the inverse Park transform and the space vector modulator (Figure 2).

It is worth noting that the d and q axes are not fully decoupled, since cross-coupling terms arising from the rotor's electrical angular speed appear in the motor's voltage equations. Although decoupling compensation terms can be added to improve dynamic performance at higher speeds, simplified implementations often neglect this coupling, relying instead on the bandwidth separation between the current loop and the mechanical dynamics of the system [8]–[10].

III. Methodology

The development of the proposed system followed a progressive approach, moving from control validation to a higher-fidelity simulation of the complete system. First, both six-step commutation and Field-Oriented Control strategies were modeled and compared in a MATLAB/Simulink environment to validate the proposed control approach and obtain reference parameters for the subsequent design stages, as detailed in Section III-A. Based on these results, a compact permanent magnet synchronous motor with an external-rotor topology was designed to provide the continuous torque required for upper limb rehabilitation, as described in Section III-B. In parallel, a dedicated motor driver was developed to execute the FOC algorithm with adequate computational and electrical performance, as presented in Section III-C. Given that physical validation of the assembled prototype was not feasible within the project schedule, a higher-fidelity simulation incorporating the identified motor and driver characteristics was employed as a substitute validation strategy, as discussed in Section III-D.

A. Simulation

The simulation phase aimed to validate the feasibility of both six-step commutation and Field-Oriented Control before proceeding to the design of the motor and driver electronics. At

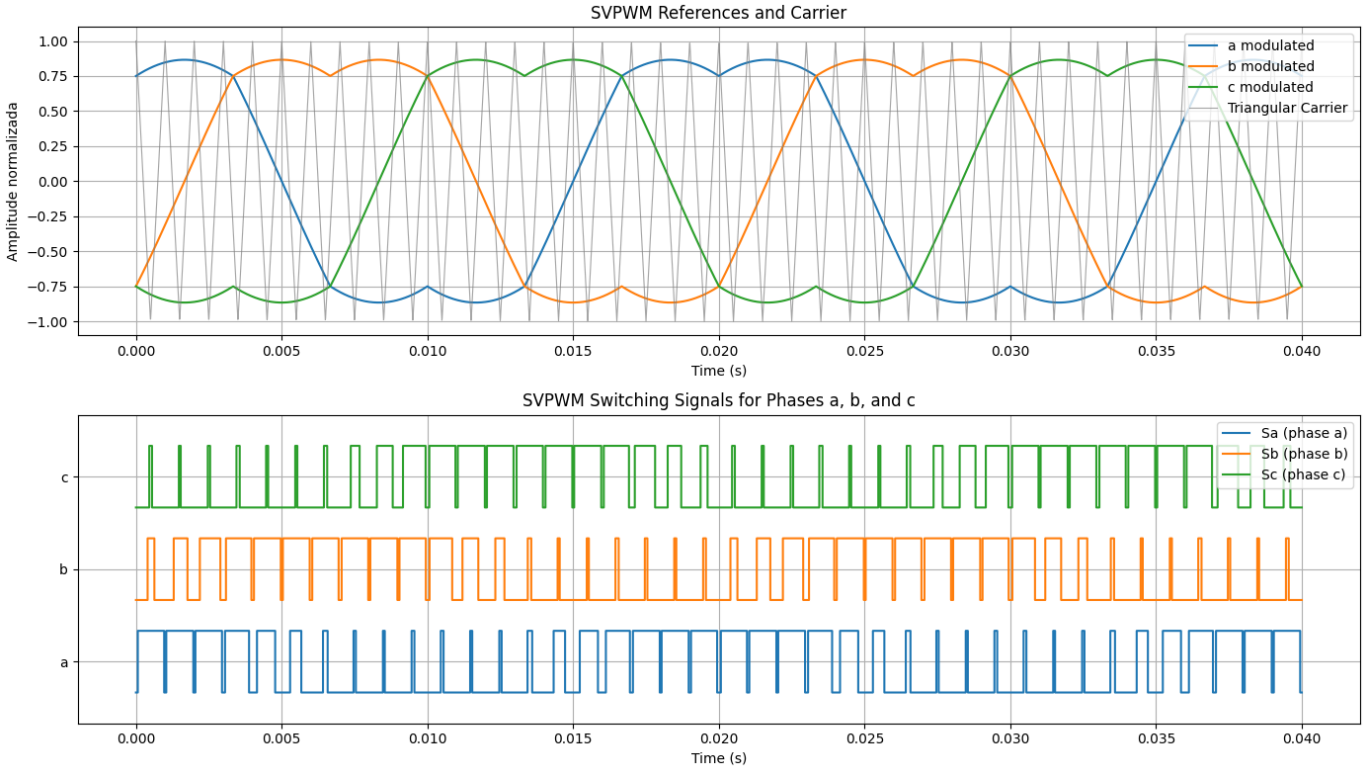


Fig. 4. Waveform and PWM output of Space Vector Modulation made in python.

this stage, since the custom motor had not yet been built, representative parameters for a generic surface-mounted PMSM (SM-PMSM) were used, consistent with the reference models adopted as a starting point for the control architecture [7], [9], [10]. These parameters were used to build a Simulink model combining blocks from the Simscape Electrical Specialized Power Systems library, which represent the power-electronics stage (three-phase inverter), and Simscape Electrical motor-control blocks, which represent the PMSM electromechanical behavior together with the Clark/Park transforms and current controllers shown in the FOC architecture (Figures 2 and 3). The resulting reference and switching signals generated by the space-vector modulator are shown in Figure 4. Both six-step commutation and FOC strategies were implemented in the same simulation environment and compared in terms of torque ripple, given its relevance for smooth, natural-like movement assistance in rehabilitation applications, as discussed in Section II-D.

B. Motor Design

Based on the parameters obtained from the simulation, a custom external-rotor (outrunner) PMSM with 14 pole pairs was designed. The out-runner topology, without the use of reduction gears, was selected to improve torque utilization at low speed without the addition of load, a desirable characteristic for rehabilitation actuators that must deliver continuous torque and backdrivability. By placing the permanent magnets on a rotating outer rotor cup around a fixed stator, the effective magnetic air-gap radius is increased, resulting in higher torque density when compared to an equivalent inrunner topology. The motor employs standard construction materials, namely copper wire for the stator windings, neodymium permanent magnets on the rotor, an aluminum structural housing, and soft magnetic steel for the stator core and rotor back-iron to efficiently channel magnetic flux.

The mechanical design was carried out in Autodesk Fusion 360, which was used for the geometric modeling and assembly verification of the rotor, stator, shaft, housing, and bearings (Figures 5 and 6). At this stage, design decisions were guided by the electrical parameters obtained from simulation and by the application requirements (continuous torque capability, compact footprint, and direct mechanical integration with the driver PCB); finite element analyzes (electromagnetic, thermal, or structural) were not performed due to hardware limitations.

C. Driver Architecture

The driver electronics were built around the STSPIN32G4, an STMicroelectronics system-in-package combining an STM32G4 microcontroller core with an integrated three-phase gate driver and motor pre-driver stage. This significantly reduced the component count of the board since most of the control and gate-driving functionality required by the FOC architecture (Figure 2) was already provided by the selected device. Power switching is performed by STL220N6F7, and rotor position feedback is provided by an AS5048A contactless

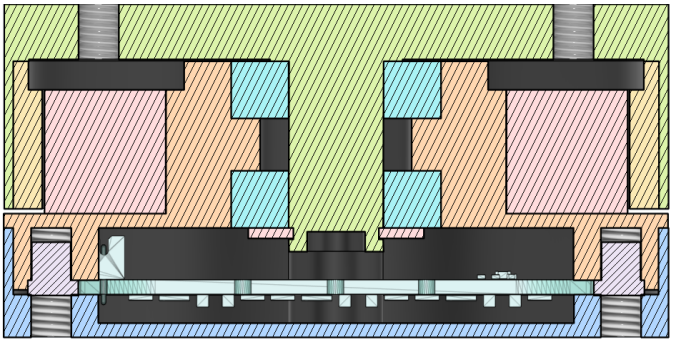


Fig. 5. CAD section analysis of the motor. Green: Rotor's case, Orange: Stator's case, Yellow: Rotor's magnetic ring, Pink: Stator, Light Blue: Bearings, Dark Blue: Driver's Case. Design and analysis made with Autodesk Fusion 360

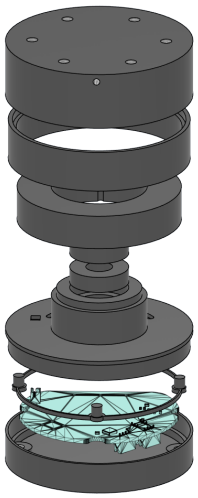


Fig. 6. CAD exploded view of the motor. Design and analysis made with Autodesk Fusion 360

14 bits magnetic rotary encoder, used as the position sensor in the current-control loop.

The printed circuit board was designed in KiCad 9 as a four-layer stack-up, with the outer layers dedicated to the ground reference and the inner layers routing the power-supply traces; digital/control circuitry and power circuitry were physically separated on the board layout to minimize switching noise coupling into the sensing and control signals. The board is supplied with a nominal input voltage of 24 V, with design operation up to a maximum of 48 V.

D. Validation

A high-fidelity simulation was conducted using the same Simulink model described in Section III-A, now updated with the parameters of the actual motor built during the design stage (Section III-B), in place of the generic representative parameters initially adopted. These parameters were identified directly from the physical motor using the ST Motor Pilot tool (v6.4.1) and the ST Motor Control Workbench, and are summarized as follows: stator resistance $Rs = 6.28\Omega$, stator inductance $L_s = 2.93$ mH, back-EMF constant $K_e = 41.89$ V/krpm, viscous friction coefficient of $16{,}06\,\mu Nms/rad$, rotor inertia of $4{,}66\,gcm^2$, 14 pole pairs, nominal current of $1.8A$, nominal DC-bus voltage of $23.35V$, and a maximum rated speed of $270rpm$. ST Motor Control Workbench provides a motor parameters extraction firmware for ST Nucleo Boards, in witch a set of values are tested inside a voltage and current range set by the user. As the motor spins the back-EMF is read and compared to the input. Motor Pilot is a graphical interface for a list of applications, such as torque, velocity and position control and open loop tests.

With the model updated to reflect the as-built system, the torque ripple comparison between six-step commutation and FOC was repeated under these conditions, allowing the conclusions drawn in the preliminary simulation to be compared against the characteristics of the real hardware developed in Sections III-B and III-C, in the absence of physical test results.

IV. Results

The torque ripple comparison between six-step commutation and Field-Oriented Control, conducted with the identified parameters of the custom motor, is shown in Fig. 7 and 8 and I. The results demonstrate that FOC produces lower torque ripple than six-step commutation under equivalent operating conditions, yielding a continuous torque output of 1 N·m. FOC also presents a faster response time when starting, causing it to overshoot, as shown in 8. This behavior is due to its characteristic alignment between the rotor end and the magnetic field. This difference is particularly relevant in the context of upper limb rehabilitation, where torque smoothness directly affects the quality of physical interaction with the patient, as discussed in Section II-D. The superior performance of FOC in this regard supports its adoption as the control strategy of choice for the proposed system, consistent with findings reported in the literature. Since the only hardware test made was in open-loop for the parameter extraction, results may vary in a closed loop test with both the motor and driver.

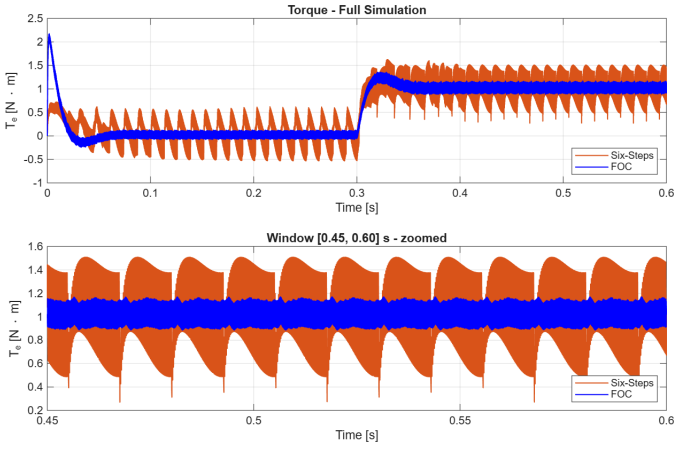


Fig. 7. Top figure shows the torque output of FOC in Blue and Six-Steps in Orange, bottom figure show a zoomed window of the same outputs

TABLE I
TORQUE RIPPLE ANALYSIS RESULTS FOR THE TWO TECHNIQUES.

Strategy	Mean[Nm]	Min[Nm]	Max[Nm]	P2P[Nm]	Ripple[%]
Six-Step	1,021	0,269	1,509	1,240	121.5
FOC	1,021	0,886	1,171	0,285	27.9

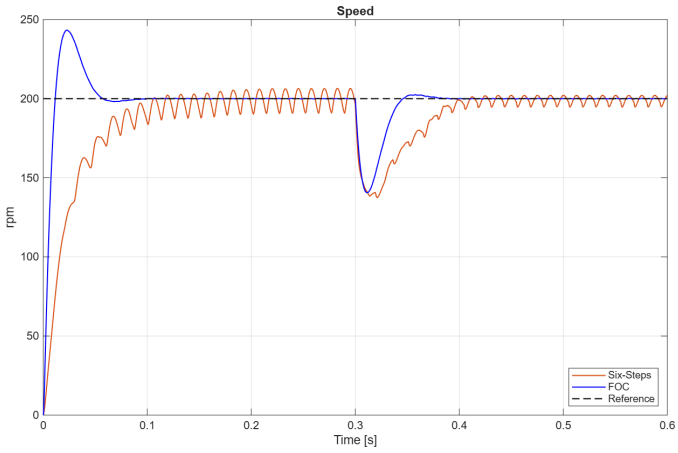


Fig. 8. Top figure shows the torque output of FOC in Blue and Six-Steps in Orange, bottom figure show a zoomed window of the same outputs

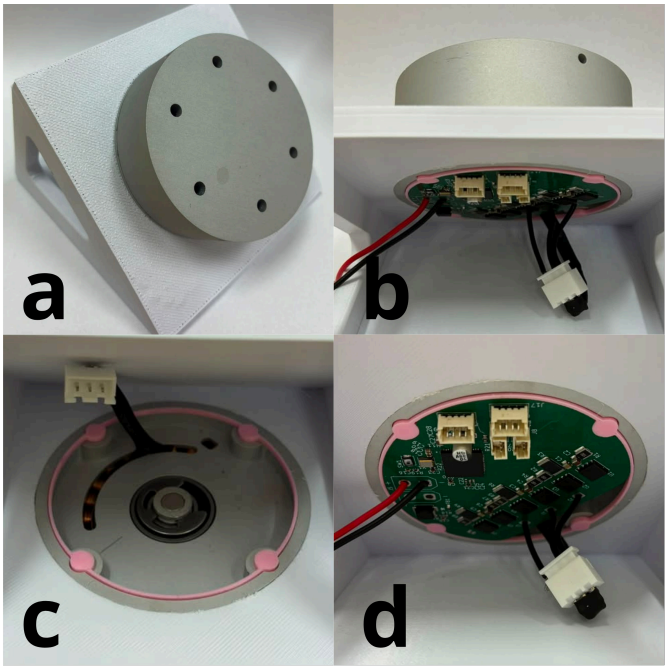


Fig. 9. a: Top view of the PMSM; b: Side view in test support; c: Bottom view; d: Bottom view with Driver attached.

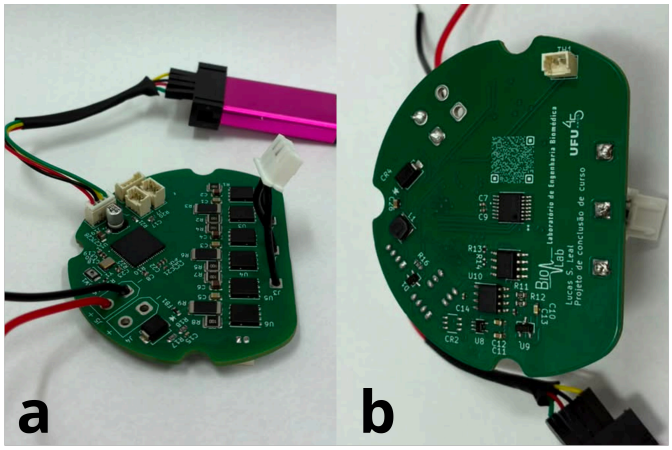


Fig. 10. a: Driver's top view with STlink; b: Driver's bottom view.

The custom external-rotor PMSM resulted in a compact actuator with a diameter of 70 mm and a height of 30.55 mm, with a total mass of 320g. These dimensions confirm the suitability of the outrunner topology for integration into a rehabilitation platform, where compactness and low weight are essential requirements. The assembled motor is shown in Fig. 9, together with the driver PCB mounted directly to its base, demonstrating the mechanical integration achieved by the design.

The manufactured four-layer driver board, shown in Fig. 10, integrates the STSPIN32G4 system-in-package, the STL220N6F7 power MOSFETs, and the AS5048A magnetic encoder interface into a compact footprint directly compatible with the motor housing. Other peripherals were also added, such as an EEPROM for storing configurations and other parameters, and a CAN communication interface. Some electrical protection was also added: SMBJ30CA for electrostatic discharge (ESD) protection in the power line; a Schottky barrier diode BAT54S for high speed switching applications, circuit protection, and voltage clamping; and ESDA5V3 for ESD protection in data lines. For temperature, both motor and driver present a slot for a thermistor, only for monitoring, no active cooling system was added to the project. The physical separation between digital and power domains on the board layout is expected to minimize electromagnetic interference on the current sensing and position feedback signals, which are critical for the stability of the FOC current loop. The board was designed to operate at a nominal supply voltage of 24 V, with the capability to support operation up to 48 V in case of usage with motors of up to 36 V, designed to present headroom for future iterations requiring higher torque output.

It is important to acknowledge that the simulation-based validation conducted in Section III-A, while sufficient to characterize the behavior of the control strategies under realistic motor parameters, does not substitute physical experimental validation. Furthermore, although FOC demonstrated a measurable reduction in torque ripple relative to six-step commutation, the perceptual relevance of this difference to the end user in a rehabilitation context remains an open question. Given the low operating speeds and the inherent compliance of the human musculoskeletal system, torque ripple at the levels produced by either strategy may fall below the threshold of clinical significance for upper limb rehabilitation quality, a hypothesis that can only be confirmed through patient-facing trials. These aspects constitute important directions for future

work, to be addressed once a fully functional prototype is available for physical testing.

V. Conclusion

This work presented the development of a FOC PMSM model to be used in the design of a robotic rehabilitation platform for upper limb motor recovery, inspired by end-effector systems such as the MIT-Manus (Figure 1). The proposed system encompasses the design of a custom external-rotor PMSM, a dedicated motor driver, and the implementation and evaluation of a Field-Oriented Control algorithm, constituting the actuator module upon which a complete rehabilitation platform can be built.

The simulation results demonstrated that FOC produces lower torque ripple than six-step commutation under equivalent operating conditions, supporting its adoption as the control strategy for applications requiring smooth and precise torque delivery. The custom motor achieved a compact form factor of 70 mm in diameter and 30.55 mm in height, with a continuous torque output of 1 N·m, and the driver integrated control, gate-driving, and position sensing functions into a single compact board.

Physical validation of the prototype remains the immediate next step, as the simulation-based approach adopted here does not fully capture real-world operating conditions. Future work includes experimental torque characterization, closed-loop performance evaluation, implementation of impedance and admittance controllers, and the integration of the actuator module into the mechanical structure of the rehabilitation platform, moving toward a complete system suitable for clinical assessment.

Acknowledgment

This work benefited from assistance with text review from other members of BioLab - UFU, as well as assistance with motor characterization, testing, and specific design-related questions. The authors would like to acknowledge the Brazilian funding agencies CAPES, CNPq and FINEP for the financial support.

References

[1] F. H. Diaz, C. Borrás Pinilla, and C. E. García Cena, "Exploring robotic technologies for upper limb rehabilitation: Current status and future directions," J. Sens. Actuator Netw., vol. 14, no. 3, art. no. 48, 2025.

[2] A. Guatibonza, L. Solaque, A. Velasco, and L. Peñuela, "Assistive robotics for upper limb physical rehabilitation: A systematic review and future prospects," Chin. J. Mech. Eng., vol. 37, no. 1, art. no. 69, 2024.

[3] P. P. Vu, C. A. Chestek, S. R. Nason, T. A. Kung, S. W. P. Kemp, and P. S. Cederna, "The future of upper extremity rehabilitation robotics: Research and practice," Muscle Nerve, vol. 61, no. 6, pp. 708–718, Jun. 2020.

[4] Y. Wang, X. Han, B. Xin, and P. Zhao, "Innovations in upper limb rehabilitation robots: A review of mechanisms, optimization, and clinical applications," Robotics, vol. 14, no. 6, art. no. 81, 2025.

[5] S. G. Rozevink, J. M. Hijmans, K. A. Horstink, and C. K. van der Sluis, "Effectiveness of task-specific training using assistive devices and task-specific usual care on upper limb performance after stroke: A systematic review and meta-analysis," Disabil. Rehabil. Assist. Technol., vol. 18, no. 7, pp. 1245–1258, Oct. 2023.

[6] B. French, L. Thomas, M. Leathley, C. Sutton, J. McAdam, A. Forster, P. Langhorne, C. Price, A. Walker, and C. Watkins, "Does repetitive task training improve functional activity after stroke? A Cochrane systematic review and meta-analysis," J. Rehabil. Med., vol. 42, no. 1, pp. 9–14, Jan. 2010.

[7] MathWorks, "Introduction to brushless DC motor control." [Online]. Available: https://www.mathworks.com/campaigns/offers/brushless-dc-motors-introduction.html

[8] M. Ulusoy, MathWorks, "Understanding field-oriented control – Motor control, Part 4," Apr. 2020. [Online]. Available: https://www.mathworks.com/videos/motor-control-part-4-understanding-field-oriented-control-1587967749983.html

[9] MathWorks, "Field-oriented control of PMSMs with Simulink." [Online]. Available: https://www.mathworks.com/videos/series/field-oriented-control-of-pmsms-with-simulink.html

[10] P. L. Chavan, D. Gowda, and S. K. Nayak, "Field oriented control technique for PMSM," in Proc. Int. Conf. Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022), 2023, pp. 474–484.

[11] MathWorks, "Clarke and Park transforms." [Online]. Available: https://www.mathworks.com/discovery/clarke-and-park-transforms.html

[12] MathWorks, "Space vector modulation (SVM) for motor control." [Online]. Available: https://www.mathworks.com/discovery/space-vector-modulation.html

Development of a Doppler Ultrasound Circulatory Phantom for Ventricular Assist Device Patients Based on a Co-Creation Framework

Jeferson Cerqueira Dias
Coordenação de Sistemas Biomédicos
Faculdade de Tecnologia Prefeito Hirant Sanazar
Osasco, Brasil
0000-0002-2999-1341

Jônatas Cerqueira Dias
Coordenação de Análise e Desenvolvimento de Sistemas
Faculdade de Tecnologia de Praia Grande
Praia Grande, Brasil
0000-0003-2515-8018

lvia L. A. Gonçalves Viana
ordenação de Sistemas Biomédic
culdade de Tecnologia Prefeito
rant Sanazar
sasco, Brasil
09-0006-3069-0286

illena Victória A. de Souza
epartamento de Engª Mecatrônica e
st. Mecânicos
niversidade de São Paulo - USP
'o Paulo, Brasil

Osasco, Brasil
0009-0008-1033-9127

Joilson Oliveira dos Santos
Coordenação de Sistemas Biomédic
Faculdade de Tecnologia Prefeito Hirant Sanazar

Diolino José dos Santos Filho
Departamento de Engª Mecatrônica e Sist. Mecânicos
Universidade de São Paulo - USP
São Paulo, Brasil
0000-0002-3578-8905

***Abstract*— Advanced Heart Failure (AHF) remains a major clinical challenge, often requiring Ventricular Assist Devices (VADs) as long-term mechanical circulatory support. Despite significant improvements in survival, VAD patients remain vulnerable to adverse events, including thrombosis, power supply failures, and electromechanical malfunctions. This study presents the development of an Internet of Medical Things (IoMT)-based architecture integrating telemonitoring and geolocation functionalities to improve emergency response and patient safety. The proposed system combines an ESP32 microcontroller, INA219 monitoring sensor, GPS module, MQTT communication protocol, cloud infrastructure, and a mobile application for real-time visualization and alarm management. The architecture adopts an Event-Driven Architecture (EDA) and Discrete Event Systems concepts supported by Interpreted Petri Nets. Experimental validation demonstrated stable communication, accurate geolocation, reliable telemetry acquisition, and effective transmission of critical alerts. The proposed solution establishes a scalable framework for remote supervision of VAD patients and contributes to reducing response time during critical adverse events.**

Keywords*— *Internet of Medical Things (IoMT), Ventricular Assist Devices, Telemonitoring, MQTT, Fault Diagnosis.

I. INTRODUCTION (*HEADING 1*)

Heart Advanced Heart Failure (AHF) remains one of the leading causes of morbidity and mortality worldwide, requiring therapeutic strategies capable of improving both patient survival and quality of life. In this context, Ventricular Assist Devices (VADs) have become an established therapeutic option for patients requiring bridge-to-transplant, bridge-to-recovery, or destination therapy by providing long-term mechanical circulatory support [1], [2]. Despite significant technological advances in pump miniaturization, durability, and hydraulic performance, VADs remain susceptible to adverse events such as pump thrombosis, ventricular suction, flow obstruction, electromechanical failures, and hemodynamic instability, all of which may compromise device performance and patient safety [2], [3].

Experimental cardiovascular platforms have become essential tools for investigating the behavior of VADs under controlled laboratory conditions. These systems enable the reproduction of physiological and pathological cardiovascular scenarios, allowing researchers to evaluate device performance, validate control algorithms, investigate adverse events, and assess new monitoring strategies without exposing patients to experimental risks [4], [5]. Similarly, vascular flow phantoms have been extensively employed for Doppler ultrasound calibration, operator training, and validation of blood flow measurements by reproducing different vascular geometries and flow conditions [5]–[7].

However, most vascular phantoms reported in the literature have been designed to simulate isolated blood vessels or specific cardiovascular diseases rather than the complex hemodynamic conditions observed in patients implanted with continuous-flow VADs. Furthermore, few experimental platforms integrate a realistic cardiovascular simulator, a ventricular assist device, Doppler ultrasound assessment, multiparameter monitoring, and data acquisition into a single experimental environment [4], [8]. This limitation restricts the development of standardized protocols for evaluating implanted VADs and validating Doppler ultrasound techniques under clinically relevant operating conditions.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

To accurately reproduce the physiological behavior of patients supported by VADs, it is necessary to define a minimum set of hemodynamic and operational parameters capable of representing the interaction between the native heart and the mechanical circulatory support system. Among the most relevant parameters are the Pulsatility Index (PI), blood flow rate, hydraulic pressure, pump rotational speed (RPM), electrical current, power consumption, and Doppler-derived blood velocity profiles. These variables provide valuable information regarding preload, afterload, ventricular contractility, pump operating conditions, and the occurrence of adverse events such as ventricular suction, thrombosis, partial obstruction, and abnormal flow patterns [2], [9], [10]. Continuous monitoring of these parameters is fundamental for understanding VAD behavior and improving both experimental investigations and future clinical monitoring systems.

The development of experimental platforms capable of reproducing these complex conditions requires multidisciplinary collaboration involving biomedical engineers, mechanical engineers, clinicians, cardiovascular surgeons, and ultrasound specialists. In recent years, **co-creation frameworks** have emerged as an effective methodology for integrating knowledge from different stakeholders throughout the design and validation process of healthcare technologies. By promoting collaborative definition of functional requirements, physiological scenarios, monitoring strategies, and validation procedures, co-creation enhances the clinical relevance, usability, and technological robustness of biomedical systems [11]–[13].

Therefore, this paper proposes the development of a Doppler ultrasound circulatory system Phantom capable of simulating patients implanted with Ventricular Assist Devices through the application of a co-creation framework. The proposed platform integrates a cardiovascular hydraulic model, a prototype blood pump representing the VAD, biomedical sensors, multiparameter monitoring, and Doppler ultrasound evaluation into a unified experimental environment. The Phantom is intended to reproduce clinically relevant physiological and pathological conditions, supporting the investigation of VAD-related adverse events, the validation of Doppler ultrasound protocols, and the development of intelligent monitoring technologies for mechanical circulatory support systems.

II. Theoretical Background

A. Advanced Heart Failure and Ventricular Assist Devices

Advanced heart failure (AHF) is a progressive clinical condition associated with high morbidity and mortality, often requiring mechanical circulatory support when pharmacological therapy is no longer sufficient [1], [2]. In this context, Ventricular Assist Devices (VADs) have become an important therapeutic option for patients undergoing bridge-to-transplant, bridge-to-recovery, or destination therapy, providing continuous support to systemic circulation and improving survival rates [1], [2]. Continuous-flow VADs are currently the most widely used devices; however, their operation is associated with complex hemodynamic interactions involving preload, afterload, ventricular contractility, and pump rotational speed [9], [10].

Despite significant technological advances, VADs remain susceptible to adverse events such as pump thrombosis, ventricular suction, flow obstruction, hemolysis, and electromechanical failures, which may compromise device performance and patient safety [3], [10]. These complications highlight the need for experimental platforms capable of reproducing realistic cardiovascular conditions and enabling the investigation of hemodynamic behavior under controlled laboratory environments [4], [5]. Furthermore, continuous monitoring of parameters such as blood flow, pressure, rotational speed, power consumption, and pulsatility index is essential for understanding device performance and supporting the development of safer and more reliable mechanical circulatory support systems [9], [10].

B. Cardiovascular Simulators and Circulatory Phantom Development

Experimental cardiovascular simulators have become essential tools for the development, validation, and optimization of mechanical circulatory support devices. These platforms reproduce physiological and pathological hemodynamic conditions under controlled laboratory environments, allowing researchers to evaluate the hydraulic performance of Ventricular Assist Devices (VADs), investigate cardiovascular responses, validate control strategies, and perform repeatable experiments without exposing patients to clinical risks [4], [5]. Mock circulatory loops have also been widely employed to support device certification, surgical training, and the development of innovative biomedical technologies by providing realistic representations of systemic circulation [4], [8].

In parallel, the development of vascular and circulatory phantoms has significantly expanded the application of Doppler ultrasound in biomedical research. Unlike conventional vascular phantoms designed primarily for equipment calibration or isolated flow measurements, next-generation circulatory phantoms seek to reproduce the complex interaction between the cardiovascular system and implanted VADs by integrating hydraulic components, prototype blood pumps, biomedical sensors, and data acquisition systems [5], [6]. Such experimental infrastructures enable the simulation of physiological and pathological flow conditions, including changes in pulsatility, pressure, flow distribution, and adverse events, thereby providing a robust platform for Doppler ultrasound validation and the evaluation of intelligent monitoring technologies for mechanical circulatory support systems [5], [7], [8].

C. Doppler Ultrasound Applied to Ventricular Assist Devices

Ultrasound imaging has become one of the most versatile and widely adopted diagnostic modalities in modern medicine due to its non-invasive nature, real-time imaging capability, absence of ionizing radiation, and relatively low operational cost. Recent technological advances have significantly expanded its clinical applicability through the development of portable and handheld ultrasound systems capable of providing point-of-care imaging in hospitals, ambulances, remote locations, and emergency scenarios. The incorporation of wireless communication, integrated beamforming electronics, and compact transducers has increased system flexibility while reducing equipment size and operational complexity, thereby facilitating the dissemination of ultrasound technologies across different healthcare environments [14]. These innovations have transformed Doppler ultrasound into an essential tool for cardiovascular assessment and continuous hemodynamic monitoring.

Ultrasound consists of mechanical waves with frequencies above the upper limit of human hearing, typically ranging from 20 kHz to several hundreds of megahertz. In diagnostic imaging, clinical systems generally operate between 2 and 15 MHz, depending on the target anatomical structure and the desired spatial resolution [15], [16]. The transducer contains piezoelectric elements that convert electrical energy into acoustic waves and subsequently transform returning echoes into electrical signals after their interaction with biological tissues. Differences in acoustic impedance among tissues determine the reflected echoes, enabling the reconstruction of B-mode images that represent the internal anatomical structures [17]. When combined with the Doppler effect, ultrasound systems become capable of estimating blood flow velocity, flow direction, spectral characteristics, and turbulence by measuring the frequency shift generated by moving red blood cells [6], [15].

Doppler ultrasound has become an important complementary technique for monitoring patients implanted with Ventricular Assist Devices (VADs), since changes in Doppler waveforms reflect the interaction between the native cardiovascular system and the mechanical circulatory support device. Parameters such as blood flow velocity, pulsatility, pressure gradients, and spectral flow profiles provide valuable information regarding preload, afterload, ventricular unloading, pump operating conditions, and the occurrence of adverse events including ventricular suction, thrombosis, obstruction, and abnormal flow patterns [8]–[10]. Nevertheless, the validation of Doppler measurements under clinically realistic VAD conditions remains limited because conventional vascular phantoms generally reproduce only simplified arterial geometries and physiological flow profiles [5], [7].

Therefore, the development of a circulatory Phantom specifically designed to simulate patients implanted with continuous-flow VADs represents an important contribution to biomedical engineering research. By integrating a cardiovascular simulator, prototype blood pumps, Doppler-compatible materials, biomedical sensors, and multiparameter data acquisition, the proposed platform enables controlled reproduction of physiological and pathological hemodynamic scenarios. This experimental infrastructure is expected to support Doppler ultrasound validation, investigation of VAD-related adverse events, evaluation of intelligent monitoring strategies, and future developments in mechanical circulatory support technologies [4], [5], [8].

D. Hemodynamic Parameters for VAD Phantom Development

The development of a circulatory Phantom capable of realistically simulating patients implanted with Ventricular Assist Devices (VADs) requires the definition of a representative set of hemodynamic and operational parameters that characterize the interaction between the native cardiovascular system and the mechanical circulatory support device. Unlike conventional vascular phantoms, which are primarily designed for equipment calibration or simplified flow measurements, a VAD-oriented Phantom must reproduce physiological conditions associated with continuous-flow support, enabling the evaluation of both normal operation and clinically relevant adverse events [4], [5]. Consequently, the experimental platform should incorporate variables that adequately represent cardiovascular function, pump performance, and Doppler ultrasound behavior under controlled laboratory conditions.

Among the most important parameters is the **Pulsatility Index (PI)**, which reflects the interaction between ventricular contractility and pump operation. The PI is influenced by preload, afterload, circulating blood volume, ventricular filling, and pump rotational speed, serving as an indirect indicator of changes in cardiovascular status [9], [10]. Reductions in pulsatility may occur during ventricular suction, hypovolemia, or reduced ventricular preload, whereas elevated pulsatility may indicate increased native ventricular contribution or changes in systemic vascular resistance [10], [18]. Continuous assessment of PI, combined with flow and power measurements, has been reported as an important strategy for monitoring VAD performance and supporting clinical decision-making [19].

Additional operational parameters include **blood flow rate**, **pump rotational speed (RPM)**, **electrical current**, **power consumption**, and **hydraulic pressure**, all of which provide complementary information regarding the mechanical behavior of the circulatory support system. Variations in these parameters may indicate changes in pump loading conditions, partial inflow or outflow obstruction, thrombus formation,

mechanical degradation, or abnormal operating conditions [2], [9]. Simultaneous acquisition of these variables enables comprehensive characterization of VAD performance and contributes to the interpretation of Doppler ultrasound findings during experimental investigations.

From the hemodynamic perspective, the Phantom should also reproduce physiological and pathological flow conditions, including laminar, pulsatile, turbulent, intermittent, and retrograde flow patterns. Furthermore, the experimental platform should be capable of simulating clinically relevant adverse events such as ventricular suction, pump thrombosis, aortic valve dysfunction, stenosis, and altered ventricular loading conditions, which produce characteristic Doppler signatures and distinct hemodynamic responses [4], [8], [20]. Computational cardiovascular models and experimental simulators have demonstrated that reproducing these scenarios is fundamental for understanding device behavior and validating monitoring methodologies before clinical application [4], [20].

Finally, integrating these hemodynamic and operational parameters into a Doppler-compatible circulatory Phantom provides a comprehensive experimental environment for evaluating VAD performance. The simultaneous acquisition of hydraulic, electrical, mechanical, and ultrasound-derived variables allows the development of reliable databases for future studies involving intelligent monitoring, predictive maintenance, adverse event detection, and artificial intelligence applied to mechanical circulatory support systems. Therefore, the proposed Phantom is expected to provide a realistic and reproducible platform for biomedical research, device validation, and the advancement of Doppler ultrasound applications in patients supported by Ventricular Assist Devices [5], [11], [13].

E. Co-Creation Framework for Biomedical Device Development

The increasing complexity of biomedical technologies has stimulated the adoption of collaborative development methodologies that integrate knowledge from multiple stakeholders throughout the innovation process. In this context, **co-creation** has emerged as an important approach for the conception, design, and validation of healthcare technologies by promoting the active participation of researchers, engineers, clinicians, end users, and other specialists during product development. According to Koning *et al.* [21], the literal meaning of co-creation derives from the combination of **"co"**, meaning "together", and **"creation"**, representing the joint production of something new. This collaborative philosophy encourages the integration of complementary expertise, resulting in technological solutions that are more robust, user-centered, and aligned with clinical needs.

Although co-creation has gained increasing attention in innovation management and healthcare engineering, its practical implementation remains heterogeneous. Roser *et al.* [12] analyzed more than fifty co-creation models described in the literature and concluded that there is no universally accepted definition or standardized methodology. Depending on the application, co-creation may be interpreted either as an open innovation strategy that promotes collaboration among organizations or as a participatory design methodology focused on involving stakeholders throughout the product development process. Despite these conceptual differences, most approaches share the common objective of integrating multidisciplinary knowledge to improve innovation outcomes and reduce development uncertainties [12].

Koning *et al.* [21] proposed a meta-model that organizes the co-creation process into six sequential phases: **Identify, Analyze, Define, Design, Realize, and Evaluate**. Within this framework, the **Design** phase represents the core of the collaborative process by incorporating activities such as inviting stakeholders, sharing knowledge, combining ideas, selecting alternative solutions, and continuously refining the proposed concept. Rather than representing a strictly linear methodology, these phases constitute an iterative innovation cycle in which technical requirements, clinical demands, experimental evidence, and user feedback are progressively incorporated into the development process, increasing the maturity and reliability of the resulting technology.

In the present study, the co-creation framework provides the methodological foundation for the development of a Doppler ultrasound circulatory Phantom intended to simulate patients implanted with Ventricular Assist Devices. The framework supports the collaborative definition of clinical requirements, hemodynamic parameters, Doppler ultrasound characteristics, hydraulic architecture, sensor integration, and experimental validation procedures. The participation of researchers from biomedical engineering, mechanical engineering, cardiovascular physiology, medical imaging, and clinical practice enables the systematic integration of multidisciplinary knowledge throughout all development stages. This collaborative approach contributes to the creation of a more realistic experimental platform capable of reproducing clinically relevant physiological and pathological conditions associated with VAD-supported circulation.

Furthermore, the proposed framework extends beyond the physical development of the Phantom by supporting the specification of monitoring variables, experimental protocols, validation criteria, and future technological evolution of the platform. By combining collaborative design principles with experimental cardiovascular simulation and Doppler ultrasound evaluation, the proposed methodology is expected to improve the reproducibility, usability, and scientific robustness of the experimental infrastructure. Consequently, the co-creation framework constitutes a strategic element for developing innovative biomedical devices capable of supporting research in mechanical circulatory support, intelligent monitoring, and validation of emerging healthcare technologies [5], [21]–[13].

III. Prepare Materials and Methods

This study adopted an applied and experimental research approach to develop a Doppler ultrasound circulatory phantom for patients implanted with Ventricular Assist Devices (VADs). The proposed methodology was structured around a co-

creation framework, integrating clinical requirements, engineering design, hemodynamic simulation, IoMT-based monitoring, and experimental validation. Initially, the functional and technical requirements were identified through a literature review and consultation of current VAD technologies. Subsequently, the circulatory phantom was designed and constructed, followed by the integration of the VAD, Doppler ultrasound acquisition system, embedded monitoring architecture, and IoMT communication infrastructure. Finally, experimental tests were conducted to evaluate the hydraulic behavior of the platform, the acquisition of hemodynamic parameters, and the suitability of the phantom for Doppler ultrasound studies. **Figure 1.** Methodological framework proposed for the development of a Doppler ultrasound circulatory phantom for patients implanted with Ventricular Assist Devices (VADs). The framework is organized into five sequential stages: co-creation, phantom development, hemodynamic simulation with Doppler setup, IoMT-based monitoring, and experimental validation with data analysis. This structured workflow integrates biomedical engineering, cardiovascular simulation, Doppler ultrasound, and digital health technologies to support the design and validation of an experimental platform for VAD research.

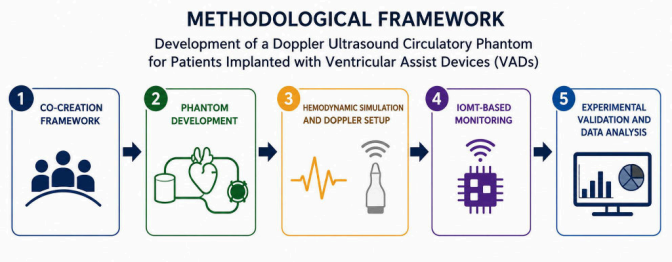


Fig. 1 Methodological framework proposed for the development of a Doppler ultrasound circulatory phantom for patients implanted with Ventricular Assist Devices (VADs).

A. System Architecture

The proposed supervisory platform was developed as a proof-of-concept Internet of Medical Things (IoMT) system for telemonitoring and geolocation of patients supported by Ventricular Assist Devices (VADs). The architecture adopts a modular layered organization that facilitates scalability, interoperability, and the future implementation of intelligent fault diagnosis algorithms.

The supervisory logic was conceived according to the principles of Discrete Event Systems (DES) and Interpreted Petri Nets (IPNs), allowing future representation of asynchronous events, alarm sequencing, and fault transitions in critical medical devices [13], [14].

The complete architecture is organized into four functional layers:

- Data Acquisition Layer
- Communication Layer
- Cloud Processing Layer
- Application Layer

The Data Acquisition Layer consists of an ESP32 microcontroller integrated with an INA219 current and voltage sensor and a NEO-6M GPS receiver. The ESP32 continuously acquires electrical variables generated by the VAD electronic emulator while simultaneously obtaining geographical coordinates from the GPS receiver. These data are processed locally and encapsulated into telemetry packets before transmission to the cloud infrastructure.

The proposed modular organization enables future integration of additional sensors and intelligent supervisory algorithms without requiring structural modifications to the overall system architecture.

B. Cardiovascular Simulators and Circulatory Phantom Development

Experimental cardiovascular simulators have become essential tools for the development, validation, and optimization of mechanical circulatory support devices. These platforms reproduce physiological and pathological hemodynamic conditions under controlled laboratory environments, allowing researchers to evaluate the hydraulic performance of Ventricular Assist Devices (VADs), investigate cardiovascular responses, validate control strategies, and perform repeatable experiments without exposing patients to clinical risks [4], [5]. Mock circulatory loops have also been widely employed to support device certification, surgical training, and the development of innovative biomedical technologies by providing realistic representations of systemic circulation [4], [8].

In parallel, the development of vascular and circulatory phantoms has significantly expanded the application of Doppler ultrasound in biomedical research. Unlike conventional vascular phantoms designed primarily for equipment calibration or isolated flow measurements, next-generation circulatory phantoms seek to reproduce the complex interaction between the cardiovascular system and implanted VADs by integrating hydraulic components, prototype blood pumps, biomedical sensors, and data acquisition systems [5], [6]. Such experimental infrastructures enable the simulation of physiological and pathological flow conditions, including changes in pulsatility, pressure, flow distribution, and adverse events, thereby providing a robust platform for Doppler ultrasound validation and the evaluation of intelligent monitoring technologies for mechanical circulatory support systems [5], [7], [8].

C. Doppler Ultrasound Applied to Ventricular Assist Devices

Doppler ultrasound is one of the most widely used non-invasive imaging techniques for evaluating blood flow dynamics and cardiovascular function. By measuring the frequency shift produced by moving blood cells, Doppler systems provide quantitative and qualitative information regarding flow velocity, flow direction, turbulence, and vascular resistance [6]. In patients implanted with Ventricular Assist Devices (VADs), Doppler ultrasound has become an important complementary tool for assessing the interaction between the native cardiovascular system and mechanical circulatory support. Variations in Doppler waveforms may indicate changes in ventricular unloading, pulsatility, flow

distribution, pump operating conditions, and the occurrence of adverse events, contributing to clinical decision-making and patient follow-up [8]–[10].

Despite its widespread clinical use, the validation of Doppler ultrasound techniques for VAD applications remains challenging because of the limited availability of experimental models capable of reproducing realistic hemodynamic conditions. Conventional vascular phantoms generally simulate simplified arterial geometries and steady or pulsatile flows but do not adequately represent the complex hemodynamic behavior observed in patients supported by continuous-flow VADs [5], [7]. Consequently, the development of a circulatory Phantom integrating a cardiovascular simulator, prototype blood pumps, Doppler-compatible materials, and multiparameter monitoring represents an important step toward improving ultrasound validation, investigating VAD-related adverse events, and supporting the development of intelligent monitoring systems. Such an experimental platform enables repeatable and controlled evaluation of Doppler signals under physiological and pathological conditions, strengthening research in biomedical engineering and mechanical circulatory support [4], [5].

D. *Validation Procedure*

The proposed IoMT platform was experimentally validated through integrated laboratory bench testing using a Ventricular Assist Device electronic emulator developed for supervisory system evaluation.

The validation procedure aimed to verify the correct operation of all functional layers of the proposed architecture, including embedded telemetry acquisition, wireless communication, cloud synchronization, geolocation services, and mobile application performance.

The experimental protocol evaluated the following performance indicators:

- Communication latency;
- MQTT connectivity stability;
- GPS positioning accuracy;
- Voltage measurement consistency;
- Current measurement consistency;
- Telemetry transmission reliability;
- Alarm notification reliability;
- Synchronization between embedded hardware and the supervisory application.

During the experimental tests, the ESP32 continuously acquired electrical measurements from the VAD emulator while simultaneously transmitting telemetry information and GPS coordinates to the cloud server. The received information was automatically processed and displayed by the supervisory mobile application, enabling real-time monitoring of the simulated device.

The adopted validation methodology follows a proof-of-concept approach commonly employed during the development of embedded IoMT systems for medical applications and was based on the experimental platform developed for VAD supervisory monitoring [16].

E. *Data Analysis*

The experimental results were analyzed qualitatively according to the operational behavior of the complete supervisory architecture. The stability of telemetry acquisition, wireless communication, cloud synchronization, GPS positioning, and graphical visualization were evaluated throughout continuous operation.

The validation focused on demonstrating the technical feasibility of integrating embedded hardware, IoMT communication, cloud services, and mobile supervision into a unified platform capable of supporting future intelligent fault diagnosis algorithms for Ventricular Assist Devices.

Although the present study was conducted using a laboratory emulator, the proposed architecture establishes a scalable technological framework for future implementation in commercial VAD systems and for the development of predictive maintenance and automated fault diagnosis strategies based on Discrete Event Systems and Interpreted Petri Nets [13], [14].

IV. RESULTS AND DISCUSSION

The proposed IoMT-based supervisory architecture was successfully implemented and validated through integrated bench testing using a Ventricular Assist Device (VAD) emulator. The experimental evaluation demonstrated the feasibility of combining telemetry acquisition, wireless communication, cloud processing, and real-time geolocation into a unified supervisory platform, consistent with the objectives established for the proof of concept. **Fig. 1** illustrates the complete architecture of the developed IoMT platform.

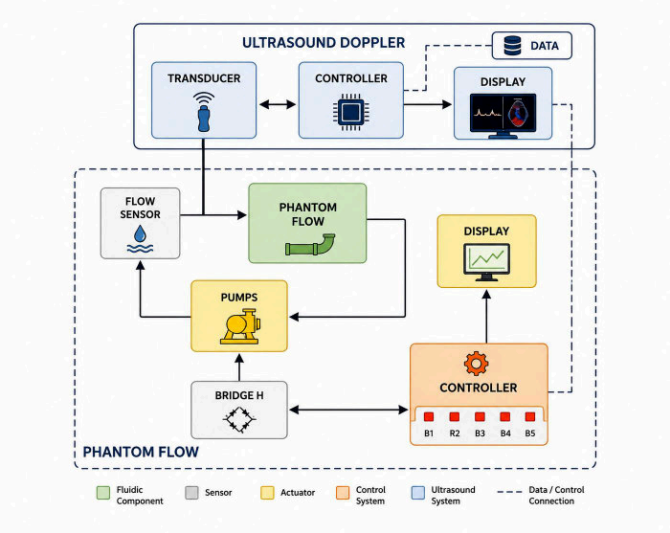


Fig. 1. The complete architecture of the developed IoMT platform

The system architecture integrates an ESP32 microcontroller, an INA219 current and voltage sensor, a

NEO-6M GPS receiver, an MQTT communication infrastructure, cloud services, and a mobile application operating under an Event-Driven Architecture (EDA). The developed platform continuously acquires electrical variables from the VAD emulator while simultaneously transmitting geographic coordinates, allowing remote supervision of both device status and patient location. These results validate the proposed architecture as a functional IoMT solution for VAD telemonitoring.

A. *Telemetry Acquisition*

Experimental tests demonstrated stable acquisition of the monitored electrical variables throughout the evaluation period. The INA219 sensor continuously measured voltage and current with consistent readings, while the ESP32 processed and transmitted the acquired information without communication interruptions.

As illustrated in **Fig. 2**, the developed prototype integrates all embedded hardware components responsible for data acquisition and communication. The experimental platform demonstrated stable operation during continuous measurements, confirming the feasibility of using low-cost embedded electronics for supervisory monitoring applications.

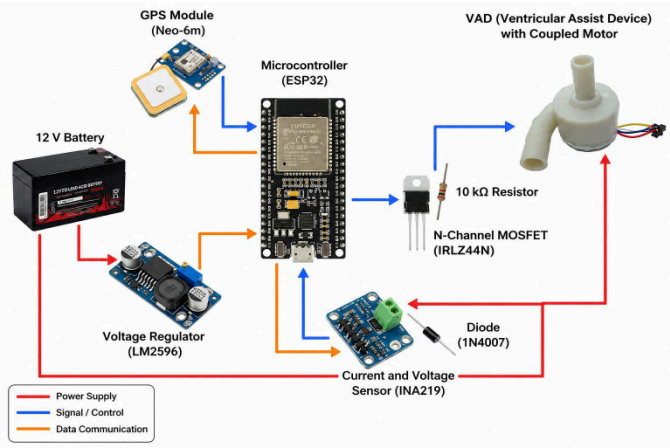


Fig. 2. The developed prototype integrates all embedded hardware components responsible for data acquisition and communication

The implemented acquisition module successfully generated periodic telemetry packets containing:

- Voltage (V);
- Current (mA);
- Device operating status;
- GPS coordinates;
- Alarm classification.

The obtained measurements remained consistent during prolonged operation, indicating that the embedded acquisition subsystem is suitable for continuous monitoring applications. Although the current implementation employs a VAD emulator rather than a clinical device, the adopted architecture can be directly extended to real LVAD controllers after appropriate hardware interface adaptation.

B. *MQTT Communication Prformance*

Communication between the embedded hardware and the cloud broker remained stable throughout the validation experiments. The publish-subscribe model based on MQTT enabled asynchronous data exchange with low communication overhead.

No packet losses capable of compromising system functionality were observed during the experimental evaluation. The communication architecture also demonstrated scalability, allowing additional sensing modules or monitoring devices to be incorporated without structural modifications.

These results confirm the suitability of MQTT for IoMT applications involving continuous telemetry of critical medical devices, where reliable low-latency communication is essential.

C. *Geolocation Validation*

The GPS subsystem successfully determined the geographic position during all outdoor validation tests. Coordinates acquired by the NEO-6M module were correctly transmitted to the cloud platform and visualized through the mobile application using Google Maps services.

The simultaneous availability of telemetry and geolocation information represents one of the principal contributions of the proposed architecture, since most currently available LVAD telemonitoring systems rely primarily on manual data entry and do not provide automatic patient localization.

The developed solution therefore enables emergency teams to identify not only abnormal operating conditions but also the patient's exact geographical location, supporting faster emergency response.

D. *Mobile Application Performance*

The supervisory mobile application successfully received, processed, and displayed all transmitted variables in real time.

As presented in **Fig. 3**, the graphical interface provides continuous visualization of the monitored parameters, including voltage, current, device status, GPS location, and alarm notifications. The Event-Driven Architecture ensures that the interface is automatically updated whenever new telemetry messages become available.

The graphical interface remained responsive throughout the validation tests, and all transmitted variables were correctly synchronized with the cloud database. The event-driven architecture allowed immediate updating of the graphical interface whenever new telemetry messages were received.

E. *Discussion*

The obtained results demonstrate that the proposed IoMT architecture is technically feasible for remote supervision of Ventricular Assist Devices.

Unlike conventional telemonitoring platforms, which generally depend on manual insertion of operational parameters, the proposed system automatically acquires electrical measurements directly from the monitored device while simultaneously providing continuous geolocation.

The combination of embedded sensing, MQTT communication, cloud services, and mobile visualization establishes the technological infrastructure required for future implementation of predictive diagnostic algorithms based on Discrete Event Systems and Interpreted Petri Nets.

Although the present study focused on proof-of-concept validation using a VAD emulator, the developed architecture constitutes an important technological step toward intelligent supervisory systems capable of supporting early fault detection, adverse event mitigation, and emergency logistics in patients supported by Ventricular Assist Devices.



Fig. 3. The graphical interface provides continuous visualization of the monitored parameters, including voltage, current, device status, GPS location, and alarm notifications Example of a figure caption.

The principal limitation of the present work is that validation was performed using an experimental emulator rather than commercially available LVAD hardware or a mock circulatory loop. Consequently, future investigations should integrate the proposed platform with physiological simulators and clinical VAD systems to evaluate its performance under realistic operating conditions.

F. Limitations of the Study

This study has several limitations that should be considered when interpreting the results. The proposed architecture was evaluated using a laboratory ventricular assist device emulator, and therefore the experiments demonstrate IoMT integration, telemetry, remote monitoring, control and geolocation functionality rather than clinical safety or patient-specific performance. Quantitative assessments of end-to-end communication latency, packet loss rate, long-term system availability, battery autonomy, alarm response time and GPS positioning error under different environmental conditions were not performed. In addition, fault-injection experiments, communication-failure tolerance, cybersecurity mechanisms, including authentication, encryption, access control and auditability, and medical interoperability were not experimentally validated. Therefore, the demonstrated results should be interpreted as proof of concept at the laboratory prototype level. Future work will include quantitative performance characterization, controlled fault-injection scenarios, cybersecurity assessment, network-failure testing and validation under increasingly realistic operating conditions prior to any clinical investigation.

V. CONCLUSION

This study presented the development and experimental validation of an Internet of Medical Things (IoMT)-based supervisory platform for telemonitoring and geolocation of patients supported by Ventricular Assist Devices (VADs). The proposed architecture integrates embedded hardware, wireless communication, cloud computing, and a mobile supervisory application into a unified platform capable of continuously monitoring electrical operating parameters while simultaneously providing real-time patient geolocation.

The proof-of-concept implementation demonstrated the technical feasibility of integrating an ESP32 microcontroller, INA219 current and voltage sensor, NEO-6M GPS receiver, MQTT communication protocol, and cloud-based services into a reliable supervisory system. Experimental results confirmed stable telemetry acquisition, consistent wireless communication, accurate geolocation, and successful synchronization between the embedded platform and the mobile application throughout the validation tests. These findings demonstrate that low-cost IoMT technologies can be effectively employed to support remote supervision of critical medical devices.

An important contribution of this work is the integration of automatic telemetry acquisition with real-time geolocation in a single supervisory architecture. Unlike conventional telemonitoring solutions that frequently depend on manual transcription of operational parameters, the proposed system automatically acquires electrical variables directly from the monitored device and transmits them to healthcare professionals through an event-driven communication infrastructure. This capability has the potential to reduce

response time during critical adverse events and improve emergency logistics for patients requiring immediate clinical intervention.

From the perspective of intelligent medical systems, the proposed architecture also establishes a technological foundation for future implementation of advanced fault diagnosis, predictive maintenance, and clinical decision-support algorithms. The modular organization adopted in this work facilitates the incorporation of Discrete Event Systems, Interpreted Petri Nets, and Artificial Intelligence techniques for automatic detection of abnormal operating conditions and prediction of adverse events.

The main limitation of the present study is that the experimental validation was performed using a laboratory VAD electronic emulator rather than a commercial ventricular assist device or a physiological mock circulatory loop. Consequently, additional investigations are required to evaluate the proposed architecture under realistic operating conditions and to validate its clinical applicability.

Future work will focus on integrating the supervisory platform with commercial VAD systems, implementing model-based fault diagnosis algorithms using Discrete Event Systems and Interpreted Petri Nets, incorporating machine learning techniques for predictive analysis of adverse events, and conducting experimental validation in physiological simulators and clinical environments. These developments are expected to contribute to safer, more reliable, and more intelligent remote monitoring systems for patients receiving long-term mechanical circulatory support.

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

VI. REFERENCES

[1] D. J. Slaughter *et al.*, "Advanced heart failure treated with continuous-flow left ventricular assist device," *N. Engl. J. Med.*, vol. 361, no. 23, pp. 2241–2251, 2009.

[2] M. C. Mancini and D. Colombo, "Left ventricular assist devices: a rapidly evolving alternative to transplant," *J. Am. Coll. Cardiol.*, vol. 65, no. 23, pp. 2542–2555, 2015.

[3] R. C. Starling *et al.*, "Unexpected abrupt increase in left ventricular assist device thrombosis," *N. Engl. J. Med.*, vol. 370, pp. 33–40, 2014.

[4] B. Fresiello *et al.*, "A cardiovascular simulator tailored for the evaluation of ventricular assist devices," *Artif. Organs*, vol. 39, no. 10, pp. E154–E163, 2015.

[5] J. C. Dias *et al.*, "Development and Validation of a Vascular Flow Phantom for Doppler Ultrasound Evaluation," *J. Biomed. Sci. Eng.*, vol. 19, pp. 189–209, 2026.

[6] P. Hoskins, K. Martin, and A. Thrush, *Diagnostic Ultrasound: Physics and Equipment*, 3rd ed. CRC Press, 2019.

[7] M. Tunkay *et al.*, "Patient-specific vascular phantom for Doppler ultrasound applications," *Med. Eng. Phys.*, vol. 78, pp. 59–67, 2020.

[8] M. Rapp *et al.*, "Simulation of ventricular assist device-related adverse events using cardiovascular simulators," *Artif. Organs*, vol. 46, no. 8, pp. 1503–1516, 2022.

[9] M. H. Chung, "Pulsatility Index in continuous-flow ventricular assist devices," *ASAIO J.*, vol. 64, no. 2, pp. 181–188, 2018.

[10] D. J. Slaughter *et al.*, "Clinical management of continuous-flow ventricular assist devices," *Ann. Thorac. Surg.*, vol. 89, pp. S48–S56, 2010.

[11] T. Roser, A. DeFillippi, and S. Samson, "Managing your co-creation mix," *J. Serv. Manag.*, vol. 20, no. 4, pp. 356–374, 2009.

[12] J. Koning, M. Crul, and R. Wever, "Models of co-creation," *Int. J. Des.*, vol. 10, no. 1, pp. 1–14, 2016.

[13] D. Fan and C. Luo, "Collaborative innovation through co-creation in healthcare technology development," *Technol. Forecast. Soc. Change*, vol. 161, Art. 120290, 2020.

[14] A. Iniewski, *Medical Imaging: Principles, Detectors, and Electronics*. Hoboken, NJ, USA: Wiley, 2012.

[15] R. K. Hobbie and B. J. Roth, *Intermediate Physics for Medicine and Biology*, 4th ed. New York, NY, USA: Springer, 2007.

[16] A. Carovac, F. Smajlovic, and D. Junuzovic, "Application of Ultrasound in Medicine," *Acta Informatica Medica*, vol. 19, no. 3, pp. 168–171, 2011.

[17] D. A. Christensen, *Ultrasonic Bioinstrumentation*. New York, NY, USA: Wiley, 1988.

[18] J. Thunberg, A. G. Garan, and M. C. Oz, "Hemodynamic determinants of pulsatility in continuous-flow ventricular assist devices," *Artificial Organs*, vol. 34, no. 10, pp. 852–859, 2010.

[19] H. Khemees and M. Shabsigh, "Clinical significance of pulsatility index monitoring in continuous-flow left ventricular assist devices," *ASAIO Journal*, vol. 58, no. 4, pp. 377–383, 2012.

[20] S. Lim, N. Howell, and A. Ranasinghe, "Continuous monitoring of pulsatility index and pump parameters in ventricular assist device patients," *Artificial Organs*, vol. 41, no. 9, pp. 859–867, 2017.

[21] J. I. J. Koning, M. Crul, and R. Wever, "Models of Co-Creation," *International Journal of Design*, vol. 10, no. 1, pp. 1–15, 2016.

Hands-Free Endoscopic Camera Control via Real-Time Head-pose Tracking for Robotic-Assisted Surgery

1st David D. Wainstein
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
0009-0003-3716-5538

2nd Rafael R. Baptista
Polytechnic School
University of São Paulo (USP)
São Paulo, Brazil
0009-0009-9437-2692

3rd Gustavo J. G. Lahr
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
0000-0002-0403-991X

***Abstract*—Camera repositioning in robot-assisted minimally invasive surgery is conventionally performed through foot-pedal clutching, forcing surgeons to interrupt instrument manipulation and increasing cognitive workload. Existing autonomous camera control approaches commonly rely on surgical instrument tracking, surgeon-specific learning, or additional wearable devices, which may limit exploratory camera motion, require personalization, or increase operational complexity. This paper presents a hands-free framework for intuitive endoscopic camera control based on real-time face-tracking. The proposed method estimates the surgeon's head orientation from facial landmarks, maps natural head movements into continuous camera motion commands, and enables independent control of the endoscopic view without interrupting the surgical workflow or requiring additional wearable hardware. Compared with the conventional pedal-based camera control strategy, the proposed approach achieved overall performance comparable to the conventional method while maintaining similar trajectory characteristics and providing statistically significant improvements in selected spatial accuracy metrics, particularly root-mean-square (RMS) deviation and angular deviation for specific participants. These results demonstrate that the proposed interface provides accurate and stable camera navigation while representing a feasible alternative to conventional endoscopic camera control.**

***Index Terms*—Robotic Surgery, Simulation and Modeling, Teleoperation, Human-Robot Interaction**

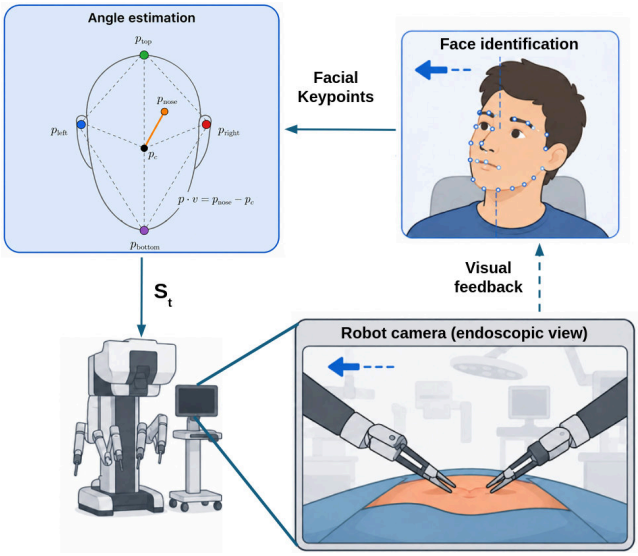


Fig. 1: Overview of the proposed framework. Head movements are tracked and translated into motion commands for the surgical robot's endoscopic camera, enabling hands-free camera control during robotic surgery.

I. INTRODUCTION

Robotic surgery has become a cornerstone approach in minimally invasive procedures [1], offering high precision while eliminating the need for conventional open interventions [2]. This technology has established applications in neurosurgery, orthopedics, urology, and many others [3]. In these systems, the surgeon typically controls the surgical manipulators and the endoscopic camera through hand-operated interfaces and foot pedals [4]. In this context, camera positioning control is a critical factor [3], [5], [6], given the inherent limitations of the surgical field of view [3]. However, current commercial platforms predominantly operate under a clutch-based control paradigm [4], which prevents the simultaneous manipulation of both the surgical instruments and the vision system.

The need to periodically interrupt the surgical procedure to reposition the camera directly disrupts the workflow. Given a restricted field of view and the dynamic nature of anatomical structures, the continuous effort to maintain an optimal frame heightens the operator's stress and cognitive load [7]–[9].To mitigate these limitations and preserve operational continuity, several autonomous camera manipulation strategies have been

proposed in the literature. These methods can be broadly categorized into: context-aware systems driven by anatomical structures, gesture and kinematic tracking approaches, and learning-based personalized frameworks [10].Within context-aware systems, Ji et al. [11] proposed the automatic detection of viewpoints to favor the visualization of critical anatomy. Regarding gesture and kinematic tracking, Pasini et al. [12] combined recurrent neural networks with forceps kinematics for automated positioning, while Elazzazi et al. [13] implemented vision control via voice commands. To capture surgical preferences, Iovene et al. [14] integrated camera movement with procedural phases and pre-mapped surgeon profiles, whereas Wagner et al. [15] developed a modular perception-interpretation-action framework that incrementally learns individual preferences into a knowledge base. Extending behavioral learning, Rivas-Blanco et al. [16] employed Learning from Demonstration (LfD) to guide an endoscopic assistant robot. Lastly, alternative hardware interfaces have been explored, such as the system by Dardona et al. [17], which leverages user head movements captured via a Head-Mounted Display (HMD).

Despite these advancements, some limitations remain unaddressed in the literature. The prevalent constraint of keeping surgical instruments strictly in focus [18], [19] restricts the surgeon's operational freedom, preventing the inspection of peripheral regions of interest without moving the active tools [3]. Learning-based and gesture-recognition systems inherently absorb the specific traits and habits of a single operator, requiring retraining to ensure interchangeability among different professionals [10]. Finally, alternative modalities that rely on additional hardware, such as HMDs or eye-trackers, could increase the user's cognitive burden [20], [21].

To overcome these limitations, this paper presents a camera control framework driven by the surgeon's head inclination and orientation via real-time face tracking, as illustrated in Fig.1. By decoupling camera motion from instrument tracking, the proposed solution restores exploratory autonomy to the surgeon, allowing independent field-of-view adjustments without cognitive drawbacks associated with wearable hardware.

II. METHODS

This section presents the methodology used to extract facial keypoints and estimate head inclination. It also describes the simulation environment used to emulate endoscopic camera movements and evaluate the proposed framework.

A. *Face-tracking*

The operator's facial pose is tracked using the MediaPipe framework [22]. After face detection, a set of facial landmarks is extracted to define a local reference plane. Since the proposed method relies solely on image information, the three-dimensional facial keypoints are represented as two-dimensional pixel coordinates. This plane is geometrically defined by four peripheral landmarks: the upper forehead region ($\mathbf{p}_{\text{top}}$) , the right ($\mathbf{p}_{\text{right}}$) and left ($\mathbf{p}_{\text{left}}$) facial extremities, and the lower chin region ($\mathbf{p}_{\text{bottom}}$), where each keypoint is

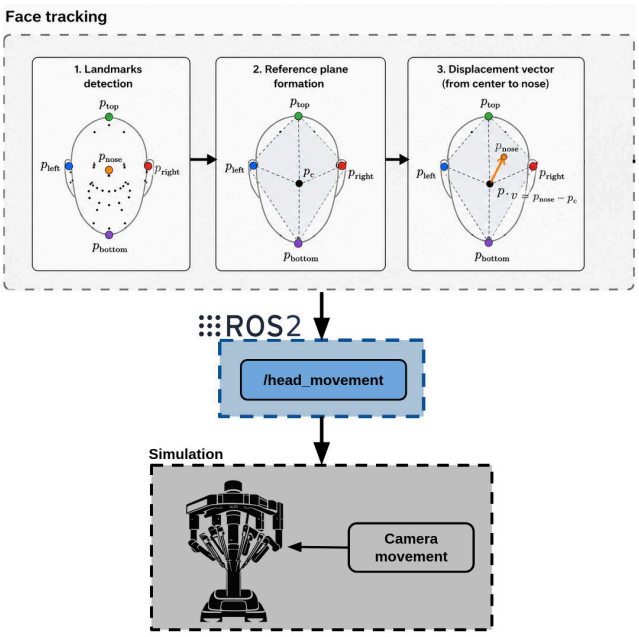


Fig. 2: Block diagram of the proposed framework. MediaPipe performs real-time face tracking to estimate the operator's head pose. Four peripheral facial landmarks are extracted to define a local reference plane, whose geometric center is computed and used as the origin of the local coordinate system. A displacement vector from the reference plane center to the nose landmark represents the head orientation through its direction and the movement amplitude through its magnitude, enabling continuous two-dimensional mapping of the operator's head motion. This information is published to a ROS2 topic, where the simulation environment subscribes and converts the received commands into endoscopic camera motion of the surgical robot.

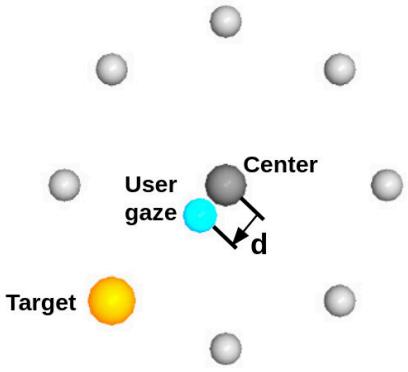


Fig. 3: Experimental environment for the target-tracking task. The sphere corresponding to the target sector is highlighted in yellow to indicate the gaze target, whereas the remaining spheres are displayed in gray. The user's estimated gaze position is represented by the blue sphere. The bidimensional length between the center and the user heead-gaze position is represented by d.

represented by a vector in $\mathbb{R}^{2\times1}$. The geometric center of this plane, denoted by $\mathbf{p}_c$, is computed according to (1) and used as the origin of the vector employed to estimate the head inclination:

$$\mathbf{p}_c = \frac{1}{4}\left(\mathbf{p}_{top} + \mathbf{p}_{bottom} + \mathbf{p}_{right} + \mathbf{p}_{left}\right) \quad (1)$$

Once the reference plane has been established, a projection vector $\mathbf{v}$ is defined by connecting the geometric center $\mathbf{p}_c$ to the landmark corresponding to the nose ($\mathbf{p}_{nose}$). This vector serves as an indicator of the face's relative displacement and is expressed by (2):

$$\mathbf{v} = \mathbf{p}_{nose} - \mathbf{p}_c \quad (2)$$

By taking $\mathbf{p}_c$ as the origin and baseline of the local coordinate system, the facial orientation and movement amplitude are directly represented by the direction and magnitude of the vector $\mathbf{v}$, respectively. Consequently, the operator's head movements are continuously mapped into a complete two-dimensional trigonometric domain. This representation provides an intuitive and omnidirectional control space, enabling unrestricted exploratory navigation of the endoscopic camera.

Finally, since dynamic head movements are inherently susceptible to involuntary tremors and high-frequency oscillations, a first-order exponential smoothing filter is applied to attenuate measurement noise and ensure the stability of the control signal. The filtering process is governed by (3):

$$\mathbf{s}_t = \alpha \cdot \mathbf{s}_{t-1} + (1-\alpha)\mathbf{v}_t \quad (3)$$

where $\mathbf{v}_t$ denotes the raw displacement vector measured at the current time step, $\mathbf{s}_t$ is the resulting filtered vector, and $\alpha \in [0, 1]$ is the exponential smoothing coefficient. The exponential smoothing coefficient α in Eq. 3 was empirically set to 0.6. A calibration step, performed at the beginning of each trial, recorded the operator's neutral head position as a baseline offset, which was subtracted from subsequent facial displacement measurements to account for individual variations in resting head pose and camera placement. To prevent unintended camera motion caused by small involuntary head movements, a deadzone threshold of 0.015 (in normalized image coordinates) was applied to the magnitude of the filtered displacement vector $\mathbf{s}_t$. The control loop operated at 100 Hz, facial landmarks were extracted from webcam frames captured at 30 fps (640×480 resolution), and the resulting camera pose commands were transmitted to the simulation environment at 30 Hz. The complete methodological workflow for facial pose estimation and kinematic mapping is illustrated in Fig. 2.

B. Simulation

The simulation of the da Vinci surgical robot, comprising its three instrument arms and the endoscopic camera, was implemented within the Asynchronous Multi-Body Framework (AMBF). AMBF was selected as it provides an open-source platform optimized for the real-time simulation of complex robotic systems, sensors, and environments, while offering baseline integration with existing da Vinci robot models [23]. Inter-process communication between the simulator and external command interfaces is handled via ROS2 topics, as illustrated in Fig. 2. Specifically, the simulator subscribes to the user's head movement orientation through ROS2 topics, translating these inputs to actuate the endoscopic camera in the corresponding direction.

III. EXPERIMENTAL EVALUATION

This section presents the experimental evaluation of the proposed system in comparison with the conventional foot-pedal-based camera control method.

The study was conducted with two volunteers, both non-surgeons, aged between 19 and 23 years. Regarding visual characteristics, one participant presented refractive errors: high myopia associated with astigmatism. Concerning familiarity with the task, only one participant had prior experience with both the simulator and camera manipulation.Both participants performed 15 trials with each of the two proposed methods.

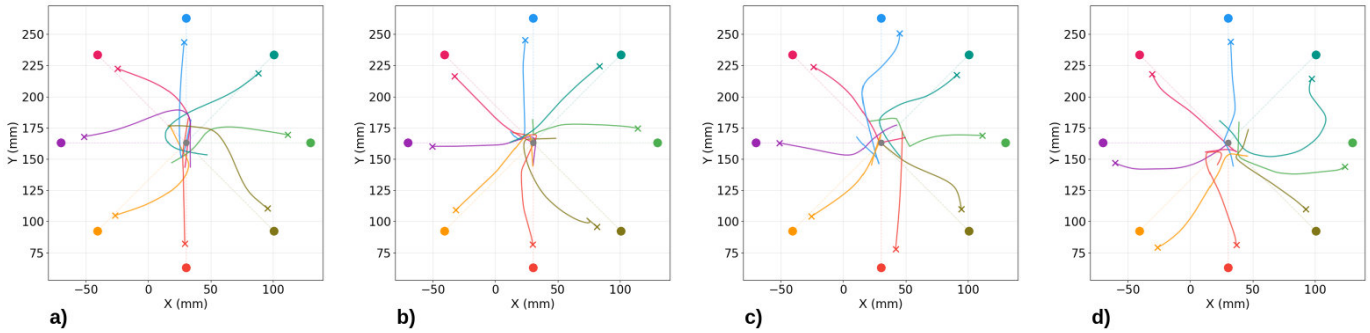
A. Experimental Setup

All experimental evaluations and the robotic system simulation were conducted on a ASUS ROG Strix G533ZS, with a 12th Gen Intel Core i9-12900H processor, 16 GB of RAM, NVIDIA GeForce RTX 3080 GPU, and Ubuntu 24.04.4. The simulation's graphical interface was rendered on the computer screen, with 15.6" and native resolution of 2560 x 1440 px. Operator facial tracking was performed using a Logi C270 HD WebCam camera.

B. Target Tracking

This experiment was designed based on established assessment tasks used in serious rehabilitation games [24], [25]. The visual stimulus consists of a circle divided into eight sectors, as depicted in Fig.3. At the beginning of each trial, the target sector is highlighted by changing its corresponding sphere's color from gray to yellow, while the remaining spheres remain gray. To ensure system robustness and mitigate potential failures in detecting the user's focus of attention, the pixel error between the target coordinates and the actual gaze fixation point is computed. After directing their gaze toward the target, participants are required to return their visual fixation to the center of the circle.

Based on this head-movement protocol, five performance metrics are quantified. The first metric is the total execution time required for the participant to complete the gaze shift from the central target to the highlighted peripheral target. The second metric is the trajectory length, defined as the total distance traveled by the gaze between the initial and final fixation points. The third metric is the trajectory efficiency, computed by comparing the performed gaze path with the shortest possible path connecting the two targets. The fourth metric is the perpendicular root-mean-square (RMS) deviation, computed as the RMS of the perpendicular distances between each sampled gaze point and the ideal trajectory, measuring the average lateral deviation from the intended path. The fifth



(a) Method 1 (b) Method 2 (c) Method 3 (d) Method 4

Fig. 4: Target-tracking trajectories obtained for each participant during the experimental evaluation. The solid lines represent the trajectories executed by the participants while moving from the central target to the highlighted peripheral target, whereas the dashed lines denote the ideal shortest path between the two targets. The ideal trajectory served as the reference for computing the trajectory efficiency, perpendicular RMS deviation, and angular deviation metrics. images (a) and (b) correspond to User 1, using the proposed and the traditional methods, respectively. Similarly, images (c) and (d) correspond to User 2, using the proposed and the traditional methods, respectively.

TABLE I: Performance metrics for each participant using the conventional and proposed methods.

User	Time (s)	Length (mm)	Eff. (%)	RMS (mm)	Angle (°)
Conventional Method Performance					
1	1.50 ± 0.13	104.81 ± 4.68	84.80	9.00 ± 1.43	9.79 ± 2.32
2	1.49 ± 0.11	112.98 ± 4.81	86.60	11.57 ± 1.84	10.17 ± 2.05
Proposed Method Performance					
1	1.63 ± 0.14	107.92 ± 10.59	84.80	9.01 ± 3.39	8.27 ± 2.19
2	1.55 ± 0.06	109.28 ± 5.20	86.60	9.34 ± 1.71	7.95 ± 1.94

metric is the angular deviation, defined as the angle between the ideal movement vector and the net displacement vector of the executed gaze trajectory, computed using the normalized dot product.

The results for the two participants are presented in Fig.4, where the solid lines represent the trajectories followed by the participants during the experiment, while the dashed line indicates the ideal trajectory. The quantitative results for each participant are summarized in Table.I, which reports the mean values of the performance metrics described before.

IV. DISCUSSION

The participants exhibited a mean completion time of $1.593, \mathrm{s}$ using the face tracking approach, which corresponds to the displacement from the central target to a peripheral target. Compared to the traditional method involving foot-pedal clutching and manual camera positioning, the conventional approach yielded a slightly lower average completion time of $1.500, \mathrm{s}$. Statistical analysis revealed that this difference was mainly influenced by User 1 ($p = 0.0128$), whereas no statistically significant difference was observed for User 2 ($p = 0.0906$). These findings indicate that no statistically significant difference in temporal performance was found between the proposed interface and the conventional control strategy, though this should be interpreted with caution given the small sample size, rather than indicating a consistent improvement in execution time. Nevertheless, the ability to maintain visual focus on the surgical field while simultaneously issuing camera commands remains an attractive characteristic of the proposed interface, as it minimizes interruptions associated with clutch activation and manual camera adjustments.

Regarding the trajectory, the average displacement between the central and peripheral targets was $108.60, \mathrm{mm}$ for the face tracking approach, while the traditional method resulted in an average trajectory length of $108.89, \mathrm{mm}$. Statistical analysis demonstrated that no significant differences were observed for either User 1 ($p = 0.3120$) or User 2 ($p = 0.0528$), although the latter exhibited a trend toward shorter trajectories with the proposed method. These results indicate that the face tracking interface is capable of generating camera movements with spatial efficiency comparable to the conventional control strategy. In this way, maintaining equivalent path lengths suggests that head-orientation commands can be translated into camera motions without introducing unnecessary endoscope displacement. Consequently, the proposed interface preserves the efficiency of camera repositioning while providing an alternative interaction modality that does not compromise trajectory quality.

The RMS deviation calculated between the ideal trajectory and the one executed via face tracking revealed an average deviation of $9.18, \mathrm{mm}$, whereas the traditional method presented an average deviation of $10.28, \mathrm{mm}$. Although no statistically significant difference was observed for User 1 ($p = 0.9890$), User 2 exhibited a significantly lower RMS deviation with the face tracking interface ($p = 0.0019$), indicating improved trajectory consistency for this participant. To complement this metric and ensure the robustness of the proposed method, the angular deviation of the participants was computed, yielding an average of $8.11°$ for the face tracking approach and $9.98°$ for the traditional method. This reduction was statistically

significant for User 1 ($p = 0.0049$), while no significant difference was found for User 2 ($p = 0.0770$). These findings suggest that the proposed interface is capable of producing camera trajectories with at least the same level of spatial precision as the conventional technique, while providing measurable improvements in specific aspects of trajectory quality depending on the individual user. Such consistency is particularly desirable in robot-assisted surgery, where predictable camera movements directly influence instrument positioning and the surgeon's situational awareness.

These results demonstrate that the proposed face tracking interface achieves performance comparable to the conventional pedal-based camera control while providing improvements in selected spatial metrics, including RMS and angular deviation for specific participants. Although a consistent reduction in task completion time was not observed across all users, the results indicate that head-orientation commands can serve as a viable alternative modality for robotic camera control without compromising navigation efficiency. Regarding the practical integration of this face tracking algorithm into the original architecture of the Da Vinci surgical system, an increase in face rotation detection sensitivity would be required. This calibration can be achieved by tuning the parameters of Equation 3, as surgeons exhibit subtle head rotations and translations when interacting with the master console viewer.

V. CONCLUSION

This paper presented a novel framework for controlling the endoscopic camera of the Da Vinci surgical robot through real-time face tracking, replacing the conventional foot-pedal clutching and manual camera manipulation paradigm. Unlike existing approaches that rely on surgical instrument tracking, surgeon-specific training, or additional wearable hardware, the proposed method enables intuitive hands-free camera control using only the surgeon's natural head movements. This framework preserves the surgeon's ability to freely explore the surgical field without requiring simultaneous movement of the surgical instruments, thereby improving workflow continuity and reducing interruptions caused by camera repositioning. The proposed approach was assessed through a preliminary experimental evaluation, in a comparative study against the conventional camera control strategy

The results demonstrated that the proposed interface achieved overall performance comparable to the conventional method while maintaining similar trajectory characteristics and providing statistically significant improvements in selected spatial metrics, particularly RMS and angular deviation for specific participants. Although statistically significant differences were not consistently observed across all metrics and users, the findings demonstrate that face-tracking interaction constitutes a feasible and promising human-machine interface for robotic surgical systems, providing intuitive operation while preserving the accuracy and consistency required for endoscopic camera navigation.

Despite these findings, some limitations should be acknowledged. The experimental evaluation was conducted exclusively with the authors of this paper, which limits the generalizability of the findings to a broader population of surgeons and users with different levels of experience. Second, the original Da Vinci surgical console differ from those implemented in the camera-and-computer setup used in this study. In particular, the interaction dynamics of the original console may impose constraints on lateral head movements, which could affect the direct transferability of the proposed interaction paradigm to the original surgical platform.

Future work will focus on integrating the proposed framework into the original Da Vinci hardware platform and evaluating its performance with a larger number of participants. In addition, future developments will investigate multimodal interaction by combining face tracking with complementary input modalities, motivated by the growing evidence that multimodal interfaces can provide greater robustness, flexibility, and usability than unimodal interaction paradigms.

REFERENCES

[1] J. R. Jeganathan, R. Jegasothy, and W. T. Sia, "Minimally invasive surgery: a historical and legal perspective on technological transformation," *Journal of Robotic Surgery*, vol. 19, no. 1, p. 408, 2025.

[2] G. Dagnino and D. Kundrat, "Robot-assistive minimally invasive surgery: trends and future directions," *International Journal of Intelligent Robotics and Applications*, vol. 8, no. 4, pp. 812–826, 2024.

[3] A. Pandya, L. A. Reisner, B. King, N. Lucas, A. Composto, M. Klein, and R. D. Ellis, "A review of camera viewpoint automation in robotic and laparoscopic surgery," pp. 310–329, 9 2014.

[4] F. Pugin, P. Bucher, and P. Morel, "History of robotic surgery: from aesop® and zeus® to da vinci®," *Journal of visceral surgery*, vol. 148, no. 5, pp. e3–e8, 2011.

[5] J. Conrad, A. Shah, C. Divino, S. Schluender, B. Gurland, E. Shlasko, and A. Szold, "The role of mental rotation and memory scanning on the performance of laparoscopic skills: a study on the effect of camera rotational angle," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 3, pp. 504–510, 2006.

[6] L. W. Way, L. Stewart, W. Gantert, K. Liu, C. M. Lee, K. Whang, and J. G. Hunter, "Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective," *Annals of surgery*, vol. 237, no. 4, pp. 460–469, 2003.

[7] R. Berguer, D. Forkey, and W. Smith, "The effect of laparoscopic instrument working angle on surgeons' upper extremity workload," *Surgical endoscopy*, vol. 15, no. 9, pp. 1027–1029, 2001.

[8] B. Zheng, M. A. Cassera, D. V. Martinec, G. O. Spaun, and L. L. Swanström, "Measuring mental workload during the performance of advanced laparoscopic tasks," *Surgical endoscopy*, vol. 24, no. 1, pp. 45–50, 2010.

[9] H. E. Douglas, M. Z. Raban, S. R. Walter, and J. I. Westbrook, "Improving our understanding of multi-tasking in healthcare: Drawing together the cognitive psychology and healthcare literature," *Applied ergonomics*, vol. 59, pp. 45–55, 2017.

[10] J. Colan, A. Davila, Y. Yamada, and Y. Hasegawa, "Human-robot collaboration in surgery: Advances and challenges towards autonomous surgical assistants," in *IEEE International Workshop on Robot and Human Communication, RO-MAN*. IEEE Computer Society, 2025, pp. 1566–1572.

[11] J. J. Ji, S. Krishnan, V. Patel, D. Fer, and K. Goldberg, "Learning 2d surgical camera motion from demonstrations," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2018, pp. 35–42.

[12] N. Pasini, A. Mariani, A. Deguet, P. Kazanzides, and E. D. Momi, "Grace: Online gesture recognition for autonomous camera-motion enhancement in robot-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 8, pp. 8263–8270, 12 2023.

[13] M. Elazzazi, L. Jawad, M. Hilfi, and A. Pandya, "A natural language interface for an autonomous camera control system on the da vinci surgical robot," *Robotics*, vol. 11, 4 2022.

[14] E. Iovene, H. K. Lev, Y. Sharon, U. Netz, A. Geftler, G. Ferrigno, E. De Momi, and I. Nisky, "A situation-aware autonomous camera alignment for enhanced suturing in robot-assisted minimally invasive surgery," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 2126–2133.

[15] M. Wagner, A. Bihlmaier, H. G. Kenngott, P. Mietkowski, P. M. Scheikl, S. Bodenstedt, A. Schiepe-Tiska, J. Vetter, F. Nickel, S. Speidel, H. Wörn, F. Mathis-Ullrich, and B. P. Müller-Stich, "A learning robot for cognitive camera control in minimally invasive surgery," *Surgical Endoscopy*, vol. 35, pp. 5365–5374, 9 2021.

[16] I. Rivas-Blanco, C. J. Perez-Del-Pulgar, C. López-Casado, E. Bauzano, and V. F. Muñoz, "Transferring know-how for an autonomous camera robotic assistant," *Electronics (Switzerland)*, vol. 8, 2 2019.

[17] T. Dardona, S. Eslamian, L. A. Reisner, and A. Pandya, "Remote presence: Development and usability evaluation of a head-mounted display for camera control on the da vinci surgical system," *Robotics*, vol. 8, 2019.

[18] T. Da Col, G. Caccianiga, M. Catellani, A. Mariani, M. Ferro, G. Cordima, E. De Momi, G. Ferrigno, and O. De Cobelli, "Automating endoscope motion in robotic surgery: A usability study on da vinci-assisted ex vivo neobladder reconstruction," *Frontiers in Robotics and AI*, vol. 8, p. 707704, 2021.

[19] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical image analysis*, vol. 35, pp. 633–654, 2017.

[20] J. Baumeister, S. Y. Ssin, N. A. ElSayed, J. Dorrian, D. P. Webb, J. A. Walsh, T. M. Simon, A. Irlitti, R. T. Smith, M. Kohler *et al.*, "Cognitive cost of using augmented reality displays," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 11, pp. 2378–2388, 2017.

[21] J. Buchner, K. Buntins, and M. Kerres, "The impact of augmented reality on cognitive load and performance: A systematic review," *Journal of Computer Assisted Learning*, vol. 38, no. 1, pp. 285–303, 2022.

[22] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[23] A. Munawar, Y. Wang, R. Gondokaryono, and G. S. Fischer, "A real-time dynamic simulator and an associated front-end representation format for simulating complex robots and environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 1875–1882.

[24] C. B. Moretti, T. Hamilton, D. J. Edwards, A. R. Peltz, J. L. Chang, M. Cortes, A. C. Delbe, B. T. Volpe, and H. I. Krebs, "Robotic kinematic measures of the arm in chronic stroke: part 2–strong correlation with clinical outcome measures," *Bioelectronic medicine*, vol. 7, no. 1, p. 21, 2021.

[25] H. I. Krebs, M. Ferraro, S. P. Buerger, M. J. Newbery, A. Makiyama, M. Sandmann, D. Lynch, B. T. Volpe, and N. Hogan, "Rehabilitation robotics: pilot trial of a spatial extension for mit-manus," *Journal of neuroengineering and rehabilitation*, vol. 1, no. 1, p. 5, 2004.

Kinematic Evaluation of a Series Elastic Actuator-Powered Prosthetic Knee Joint

Lucas P. Acosta
Instituto Superior de Investigaciones Biológicas (INSIBIO) CONICET
National University of Tucumán
Tucumán, Argentina
lacosta@herrera.unt.edu.ar

M. M. A. C. Moreira
Robotic Rehabilitation Laboratory,
University of São Paulo at São Carlos
Center for Robotics of São Carlos,
University of São Paulo
São Carlos, Brazil.
melkzedekue@usp.br

Y. M. Villamizar
Robotic Rehabilitation Laboratory,
University of São Paulo at São Carlos
Center for Robotics of São Carlos,
University of São Paulo
São Carlos, Brazil.
yecidmoreno@usp.br

D. G. Goroso
Center for Technological Research,
University of Mogi das Cruzes.
Mogi das Cruzes, São Paulo, Brazil.
danielg@umc.br

F. D. Farfan
Instituto Superior de Investigaciones Biológicas (INSIBIO) CONICET
National University of Tucumán
Tucumán, Argentina
ffarfan@herrera.unt.edu.ar

A. A. G. Siqueira
Robotic Rehabilitation Laboratory,
University of São Paulo at São Carlos
Center for Robotics of São Carlos,
University of São Paulo
São Carlos, Brazil.
siqueira@sc.usp.br

***Abstract*—For patients with lower limb disabilities, the primary goal of rehabilitation is to restore limb function, a process in which robotic knee prostheses act as a fundamental component. This paper presents the design of a mechanism adapted for the implementation of a Series Elastic Actuator (SEA) within a polycentric knee prosthesis. The proposed system utilizes a commercial polycentric knee joint, which comprises a four-bar mechanism and a pneumatic damping system. To evaluate the knee joint prosthesis, a flexion-extension cycle was executed to capture the sensor responses. During this dynamic test, distinct differences in displacement values between the two cycle phases were observed. Additionally, to test the SEA output, the knee joint was mechanically blocked with a rigid bar, enabling the measurement of signals generated exclusively by the elastic component's deformation. The testing of the elastic element demonstrated that its deformation values, relative to the drive pulses sent to the mechanism, scaled linearly. These results provide an essential foundation for the future implementation of an effective force control system on the prosthetic mechanism.**

Keywords—Knee joint prosthesis, Series elastic actuator.

I. INTRODUCTION

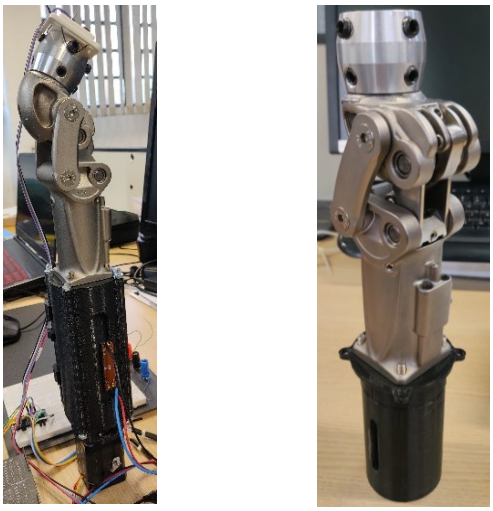
For individuals with lower limb disabilities, the restoration of limb function is the primary objective of physical rehabilitation, with robotic knee prostheses serving as the core functional component. These advanced prostheses enable amputees to walk by actively controlling shank posture, providing the necessary joint torque, supplying swing damping, and ensuring overall lower limb stability, thereby facilitating both physical recovery and psychological well-being. Due to its robust bionic and mechanical properties, the four-bar mechanism has been widely utilized in the design of robotic knee prostheses since the 1980s. Recently, research on four-bar robotic knees has increasingly focused on specialized locomotion functions, integrated actuation, and controlled elastic deformation. A polycentric knee mechanism is characterized by an instantaneous center of rotation that shifts its position dynamically as the knee flexion angle changes. While all polycentric devices control stability through this shifting center of rotation , there is an infinite variety of possible kinematic arrangements, each offering distinct functional characteristics.

The Series Elastic Actuator (SEA) is an alternative actuation configuration that emerged in the 1990s as a solution to robotic control challenges, specifically designed to maintain safe force levels during physical therapy and human-robot interaction. SEAs have since become a standard in force and impedance control applications. In an SEA, an elastic element is positioned in series between the mechanical transmission and the load, allowing the system to indirectly measure the applied force based on spring deformation, effectively transforming a complex force-control problem into a more manageable position-control problem. To further optimize SEA performance against disturbances caused by human biomechanical torques, advanced strategies such as optimal PID and feedforward controllers with disturbance observers have been implemented. The objective of this work is to design a mechanism specifically tailored for integrating an SEA into a polycentric knee prosthesis.

II. MATERIALS AND METHODS

A. Prosthesis Design

The experimental setup utilized a Wonderful 3F24P polycentric knee joint, which features a four-bar mechanism and a pneumatic damping system. The primary design goal was to replace the original passive system—composed of a spring and pneumatic valves—used to assist the gait cycle.



a b

Figure 1. (a) Complete assembly of the polycentric knee with the series elastic actuator. (b) Polycentric prosthetic knee.

The knee joint was modified by introducing a transmission mechanism driven by a NEMA stepper motor, which is linked to a lead screw via an elastic coupling. An extension is attached to the piston of the pneumatic damping system to transmit motion and force from the series spring assembly. This assembly is driven by a nut (8mm/rev pitch) connected to the lead screw. To optimize rapid prototyping costs during the testing phase, several custom components were manufactured using 3D printing. This compact design ensures full functionality and optimal space utilization within the knee, making it highly suitable for future developments of fully integrated lower limb prostheses.

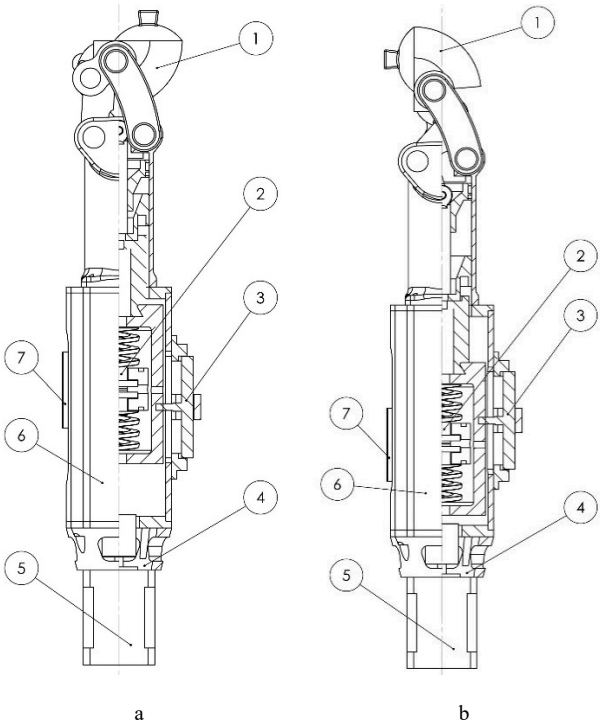


Figure 2. Half cut half view from the assembly during orthostatic stance phase (a) and sitting stance phase (b). 1: Polycentric Pneumatic Prosthetic Knee Joint. 2: Series Elastic Actuator (SEA) device. 3: Sliding potentiometer registering knee joint position. 4: Motor Case Joint. 5: Motor NEMA. 6: External Case from SEA. 7: Sliding potentiometer registering spring deflection.

The custom elastic device was constructed using two parallel springs with a combined elastic constant of 113 N/mm. The displacement of the internal sliding piston is recorded by a 20 kΩ sliding potentiometer. All these subcomponents are housed within a custom capsule attached to the sliding link of the polycentric knee prosthesis.

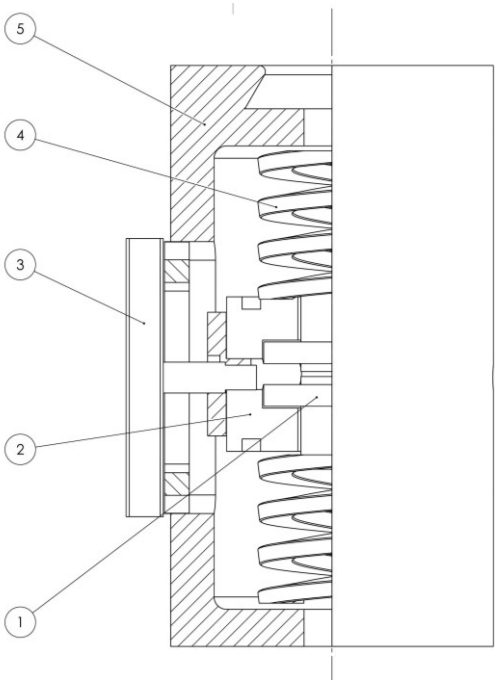


Figure 3. Half cut half view from the Spring Capsule. 1: Nut. 2: Piston. 3: Sliding Potentiometer. 4: Spring. 5: Case.

B. Specifications, Kinematics and Dynamics

For dynamic modeling, normative gait cycle torque data provided by Kirtley was applied to the kinematic model to determine the necessary force exerted by the motion system. Using a detailed CAD model, the structural efforts on the knee joint were simulated to calculate the resultant forces acting on the sliding link joint. These extracted values determined the required force along the displacement vector of the prosthetic link throughout a standard gait cycle. By correlating these force values with the displacement of the spring system, the varying elastic constants across different positions were calculated.

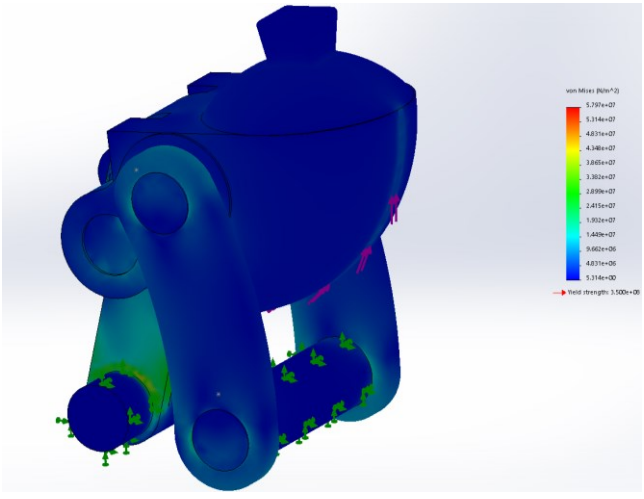


Figure 4. Forces on the prosthetic knee sliding link using FEM simulation.

The forces derived from the FEM simulation were plotted against Kirtley's normative torque data for corresponding knee joint angles. This comparison between the simulated sliding system forces (required for flexion and extension) and experimental torque values allowed for the precise determination of the electrical motor's necessary torque output.

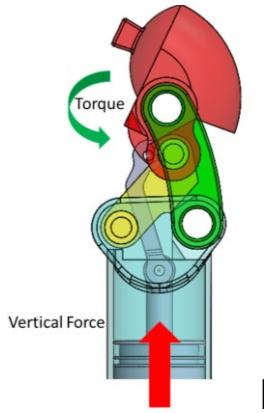


Figure 5. Vertical Force due an applied Torque schematic for static simulation.

The simulation allowed determine the Vertical Force values required to produce the torque value for the corresponding angle in [6] based in the geometry of the 3F24P prosthesis and executing a static simulation with the corresponding constraints.

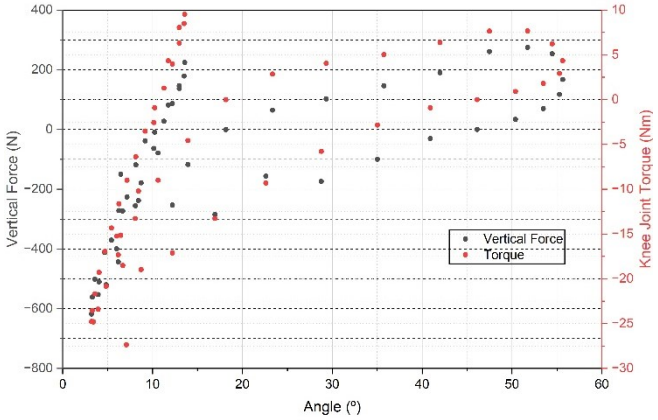
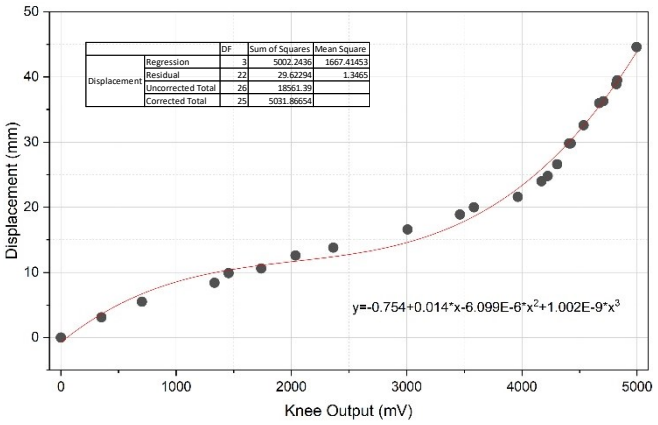


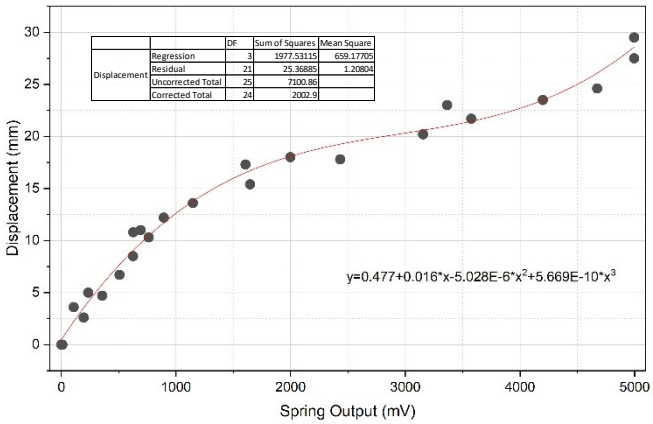
Figure 6. Vertical Force (black dots) and Torque on the Knee Joint (red dots) vs angular position of the knee joint.

To evaluate the knee joint's kinematic response, a sliding potentiometer was attached to the external case of the elastic system. A flexion-extension cycle was executed to map the relationship between the potentiometer's output and the actual knee angle relative to gravity, which was independently verified using an MPU9250/6500 IMU sensor communicating with an Arduino NANO (4.88 mV/Step) for data acquisition and command signaling.



		DF	Sum of Squares	Mean Square
Displacement	Regression	3	5002.2436	1667.41453
	Residual	22	29.62294	1.3465
	Uncorrected Total	26	18561.39	
	Corrected Total	25	5031.86654	

Figure 7. Nonlinear fit for the knee potentiometer response.



		DF	Sum of Squares	Mean Square
Displacement	Regression	3	1977.53115	659.17705
	Residual	21	25.36885	1.20804
	Uncorrected Total	25	7100.86	
	Corrected Total	24	2002.9	

Figure 8. Nonlinear fit for the spring potentiometer.

Because the components exhibited non-linear electrical responses, a polynomial curve fit was applied to translate the analog output values (in mV) into precise physical displacements. The R-squared value for the knee output fit was 0.994, while the spring output fit yielded an R-squared of 0.987. This non-linear behavior suggests that future iterations may benefit from using alternative position sensors. To specifically isolate and test the SEA output, the knee joint was mechanically blocked with a rigid bar. This constraint allowed the elastic component's deformation to be measured directly via the sliding potentiometer attached to the piston, while control pulses were sent to the stepper drive (configured at 800 steps/rev). The resulting spring deformation—caused by the piston's displacement driven by the nut and lead screw—was captured via the Arduino NANO, with analog voltage signals converted into discrete digital values for analysis.

III. PRELIMINARY RESULTS

During the dynamic flexion-extension cycle test, the output values from the knee joint sliding potentiometer were recorded by the microcontroller.

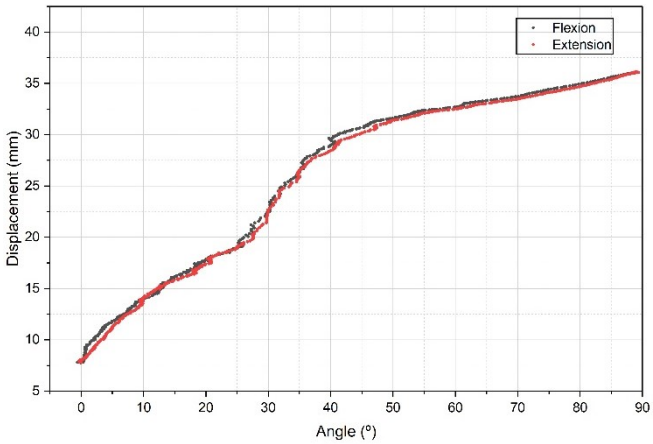


Figure 9. Potentiometer output during the flexion (black dots) and extension (red dots) cycle.

A distinct difference in the displacement trajectory was observed between the flexion and extension phases, strongly indicating the presence of mechanical backlash or other non-ideal mechanical behaviors within the physical assembly.

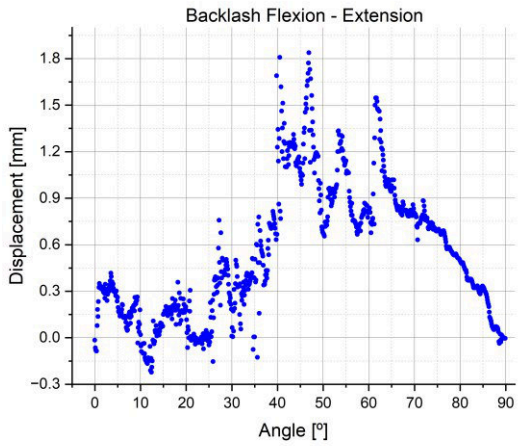


Figure 10. Backlash distribution for a cycle Flexion – Extension.

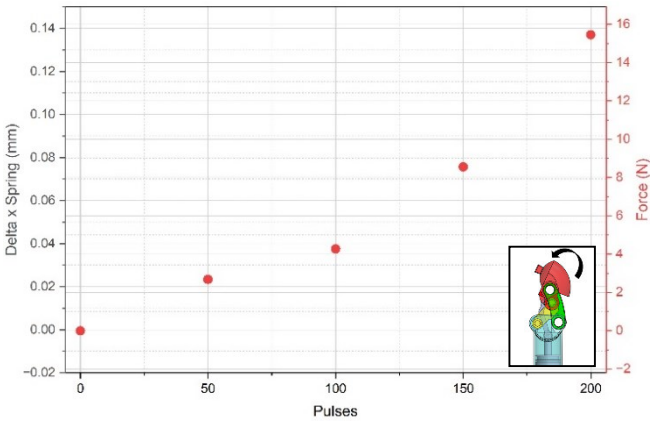


Figure 11. Output of the SEA´s spring test during the flexion phase.

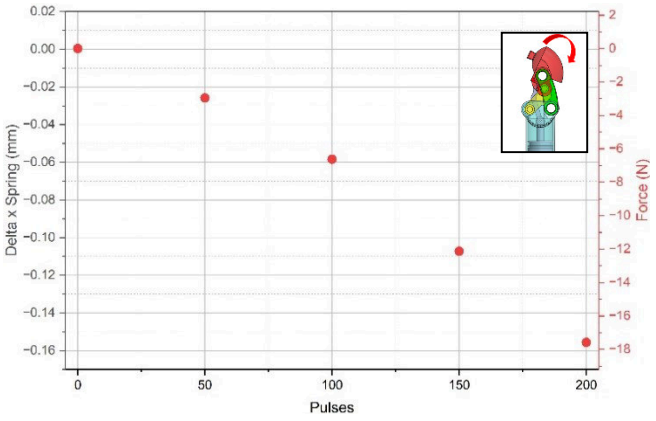
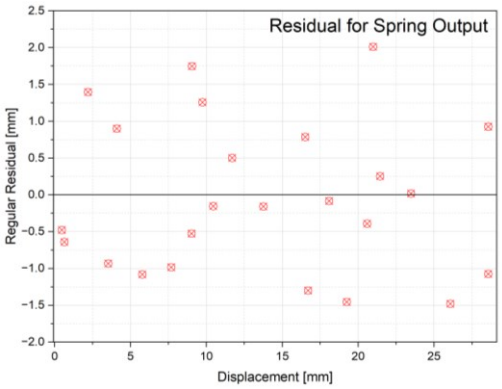
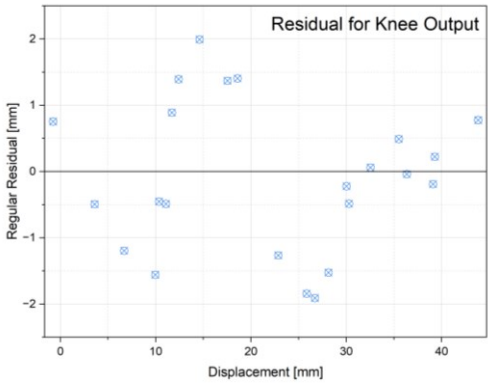


Figure 12. Output of the SEA´s spring test during the extension phase.

For the isolated SEA flexion test, the four-bar mechanism was blocked at a fixed angle of 30 degrees. Under these conditions, the registered spring deformation was 0.136 mm. Conversely, during the extension phase test—conducted under identical mechanical constraints with the knee fixed at 30 degrees—the elastic element exhibited a deformation of 0.155 mm in the opposite direction.



a



b

Figure 13. (a) Residual values in the nonlinear fit for Spring Output. (b) Residual in the nonlinear fit for Knee Output.

The residual analysis made for the Spring Output (red) nonlinear fit were +2.01 mm and -1.47 mm for maximum values, and the residual values for the Knee Output (blue) nonlinear fit were +1.99 mm and -1.91 mm in the extremes values.

IV. DISCUSSION

The continuous development of robotic prosthetics, particularly those integrating Series Elastic Actuators, represents a dynamic and critical field of research. This work is fundamentally driven by the objective of engineering solutions that not only restore mobility but tangibly improve the user's quality of life. Prior research by Herr and Grabowski on bionic ankle-foot prostheses has convincingly demonstrated how active devices can normalize the walking gait of amputees, underscoring the vital interplay between prosthetic mechanics and human biomechanics.

SEAs, pioneered by Pratt and Williamson, have proven to be exceptionally effective in applications requiring safe force thresholds and precise impedance control. By measuring force indirectly through spring deflection, SEAs successfully convert demanding force-control paradigms into highly manageable position-control algorithms. This foundational principle is critical for the deployment of sophisticated human-robot interaction controllers. Furthermore, comprehensive reviews, such as those by Sun, emphasize that optimizing structure, actuation, and control is essential for the next generation of responsive robotic knees. The integration of foundational biomechanical datasets, such as the CGA Normative Gait Database, remains an indispensable tool for

validating and calibrating dynamic models in these simulations. Combining robust theoretical frameworks with emerging rapid-prototyping technologies like 3D printing accelerates the advancement of customized robotic prosthetics. Moving forward, the integration of high-resolution linear encoders is identified as a critical next step to mitigate current non-linearities, improve measurement accuracy, and enhance the overall system response.

V. Conclusions

This research details significant advancements in the electromechanical design and kinematic evaluation of a robotic knee prosthesis powered by a series elastic actuator. The preliminary tests successfully demonstrate the feasibility and potential of integrating SEAs to enhance lower limb prosthetics. However, the experimental results also highlight the strict necessity of utilizing high-precision components to ensure absolute measurement accuracy and minimize mechanical backlash. Future work will focus on refining the mechanical design by replacing 3D-printed prototypes with machined metallic parts and integrating precise linear encoders. Ultimately, these iterative improvements aim to deliver a fully integrated lower limb prosthesis that seamlessly meets user demands, drives innovation in robotic rehabilitation, and profoundly enhances the quality of life for individuals with lower limb disabilities. Furthermore, integrating the SEA into the polycentric knee mechanism will provide proportional feedback signals to accurately monitor the joint state of the artificial limb. In this regard, future developments aim to implement user-interaction strategies that enhance the amputee's proprioceptive perception and functional embodiment of the prosthesis.

ACKNOWLEDGMENT

This research work was supported by:

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) Argentina. RESOL-2021-154-APN-DIR#CONICET.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 – Program MOVE LA AMÉRICA 88881.997609/2024-01.

São Paulo Research Foundation (FAPESP) under grants 2024//01120-4 and 2024/23815-4.

REFERENCES

[1] Yuanxi Sun, Review of Recent Progress in Robotic Knee Prosthesis Related Techniques: Structure, Actuation and Control - Journal of Bionic Engineering - Eng 18 (2021) 764–785

[2] CW Radclife “Above-Knee Prosothetics”; Prosthetics and Orthotics International 1977, 1, 146-160.

[3] G. Pratt, M. M. Williamson et al. “Series elastic actuators” in Intelligent Robots and Systems 95.’Human Robot Interaction and Cooperative Robots’, Proceedings. 1995 IEEE/RSJ International Conference on, vol. 1. IEEE, 1995, pp. 399–406.

[4] K. Kong, J. Bae, and M. Tomizuka, “Control of rotary series elastic actuator for ideal force-mode actuation in human robot interaction applications,” IEEE/ASME Transactions on Mechatronics, vol. 14, no. 1, pp. 105–118, Feb 2009.

[5] W. M. dos Santos, G.A.P. Caurin and A.A.G. Siqueira “Design and control of an active knee orthosis driven by a rotary Series Elastic Actuator”, Control Engineering Practice 58 (2017) 307–318.

[6] C. Kirtley, CGA Normative Gait Database, Hong Kong Polytechnic University. Available http://guardian.curtin.edu.au/cga/data/

[7] H. Herr e A. Grabowski, "Bionic ankle-foot prosthesis normalizes walking gait for persons with leg amputation," Science Translational Medicine, vol. 6, no. 255, pp. 255ra70, 2014.

[8] A. Dollar e H. Herr, "Active ankle-foot prosthesis," in Proceedings of the IEEE International Conference on Robotics and Automation, pp. 1395-1400, 2006.

Multimodal Sensory Feedback in a Game-Based Task for Robotic Rehabilitation: A Pilot Study in Healthy Adults

Jean Mendes Nascimento
Biomechatronics Laboratory, Escola Politecnica
University of Sao Paulo
Sao Paulo, Brazil
jeanmendes@usp.br

Guilherme Martins Mello
Biomechatronics Laboratory, Escola Politecnica
University of Sao Paulo
Sao Paulo, Brazil
guilhermemartinsmello@usp.br

Bruna M. Carlos
Biomechatronics Laboratory, Escola Politecnica
University of Sao Paulo
Sao Paulo, Brazil
bruna.carlos@usp.br

Arturo Forner-Cordero
Biomechatronics Laboratory, Escola Politecnica
University of Sao Paulo
Sao Paulo, Brazil
aforner@usp.br

***Abstract*—Sensory impairments after stroke can negatively affect upper limb function and daily life activities. However, sensory-oriented rehabilitation strategies are still less explored than interventions focused on motor training. This study presents a robotic handlebar integrated with a serious game capable of delivering multimodal sensory stimuli during gameplay, including visual effects, vibrotactile stimulation, and resistive haptic torque. Six healthy volunteers performed four game conditions: baseline, visual rain, visual rain combined with vibrotactile stimulation, and visual wind combined with resistive torque. Task performance was evaluated using the number of collected coins, while kinematic behavior was assessed through lateral velocity, acceleration, jerk and displacements. The coin collection rate remained higher than 92% accross all conditions, with a small tendency of increase after the first condition, suggesting a practice effect rather than a specific effect of the applied stimuli. On the other hand, kinematic outcomes indicated that the experimental conditions may modulate handlebar control behavior. The visual wind combined with resistive torque produced the clearest increase in movement intensity, suggesting a higher control demand. In contrast, vibrotactile stimulation did not show a significant additional effect beyond visual rain, although it presented a descriptive tendency toward lower trajectory deviation. These preliminary results support the feasibility of the proposed platform for investigating how multimodal sensory stimuli affect user interaction and handlebar control during game-based bimanual rehabilitation tasks.**

***Index Terms*—robotic rehabilitation, sensory feedback stimuli, serious game.**

I. Introduction

Stroke is recognized as the leading cause of adult disability and the second leading cause of death globally [1]–[3], with a global cost exceeding US$891 billion. Moreover, between 1990 and 2019, the global burden of incident strokes and stroke-related deaths increased substantially, by approximately 70% and 43%, respectively. This burden is disproportionately concentrated in low-middle income countries, which account for 86% of stroke-related deaths and 89% of disability-adjusted life years related to stroke [4].

Only a small proportion of individuals affected by stroke recover their ability to resume daily life without appropriate treatment. The extent of motor impairment depends on several factors, including stroke severity and location, and the effectiveness of early medical care [5]. In addition to motor deficits, up to 85% of stroke survivors experience upper limb sensory impairments, including reduced tactile, temperature, pain, and proprioceptive perception. These sensory deficits can further compromise motor function and negatively affect functional recovery [6], [7].

In recent decades, substantial efforts have been directed toward supporting the recovery of individuals with sensorimotor impairments. In this context, serious games have been increasingly applied in different fields with the purpose of improving users' abilities in real world contexts [8]. Their various applications have the ultimate goal of allowing users to achieve a specific objective in an entertaining and engaging manner through the experience of playing the game [9]. In rehabilitation, this approach has proven effective and has provided motivational benefits during therapy sessions, being widely combined with robotic devices to support goal-oriented tasks during limb movements [10], [11].

Rehabilitation technologies, particularly robotic devices, have emerged as promising tools to assist upper-limb therapy [5]. These systems can provide repetitive, intensive, and task-oriented training over prolonged periods, which may contribute to partial or complete functional recovery. Many robotic rehabilitation approaches have primarily focused on specific aspects of motor recovery, including shoulder rehabilitation [12], hand training and assessment to improve dexterity or grip force

[13]–[15], or therapeutic strategies such as Mirror Therapy (MT) and Bilateral Arm Training (BAT) [16], [17]. However, sensory rehabilitation is often insufficiently addressed. Current interventions for recovering upper limb sensory impairments are mainly based on electrical stimulation techniques, including noninvasive wearable devices with embedded electrodes [13], [18] and Functional Electrical Stimulation (FES) systems combined with task-specific training [19], [20].

This gap is clinically relevant, since upper limb sensory impairments have a substantial negative impact on activities of daily living in stroke survivors. Nevertheless, specific and effective interventions for upper limb sensory recovery remain limited, and this aspect of rehabilitation is still frequently overlooked in clinical practice [21]–[23].

Therefore, we propose a robotic device embedded with a system capable of providing different types of sensory feedback, including visual and haptic stimuli, integrated into a more immersive serious game. The haptic feedback can be delivered through force fields, which may be assistive or resistive, as well as through vibrotactile stimulation. These haptic stimuli can be applied either independently or in combination with visual feedback on the game screen, allowing the investigation of how different sensory conditions affect user interaction, movement control, and task performance. To evaluate this approach, we conducted an initial test with six healthy adult volunteers using a bimanual robotic handlebar [24] to control a coin-collection game, with multimodal stimuli directly synchronized with game events.

II. Materials and Methods

A. The Robotic Handlebar

The experimental procedure was conducted using an instrumented robotic handlebar that was originally designed to control an on-screen cursor toward predefined targets. In the present study, the system was expanded to control a computer game displayed on a screen positioned in front of the subject. The device consists of a one-degree-of-freedom system developed for bimanual rehabilitation training with a game-based interface and bidirectional communication with the computer. Its mechanical behavior is based on the dynamics of a bicycle handlebar, modeled as a rigid body rotating around a fixed steering axis. Six load cells are integrated into the rigid link and symmetrically distributed around the device's center, two of them are embedded inside the handles to measure grip forces, while the remaining four are used to estimate interaction torques. The rotational axis is coupled to a direct current (DC) motor equipped with an encoder, allowing continuous monitoring of angular position. System control is performed through an H-bridge and a microcontroller, which also manages the communication interface and data exchange between the handlebar and the game running on the PC. This complete experimental setup can be seen in Fig. 1.

In the current version of the device, not only the handlebar can control movements on the screen, but it can also receive commands from the computer so that the game can apply torques through the motor to provide haptic feedback in the form of either resistance or assistance to the user's movement. The motor uses impedance control to ensure a safe interaction between user and robot [25]. The impedance controller parameters are tuned to create a spring-like haptic effect.

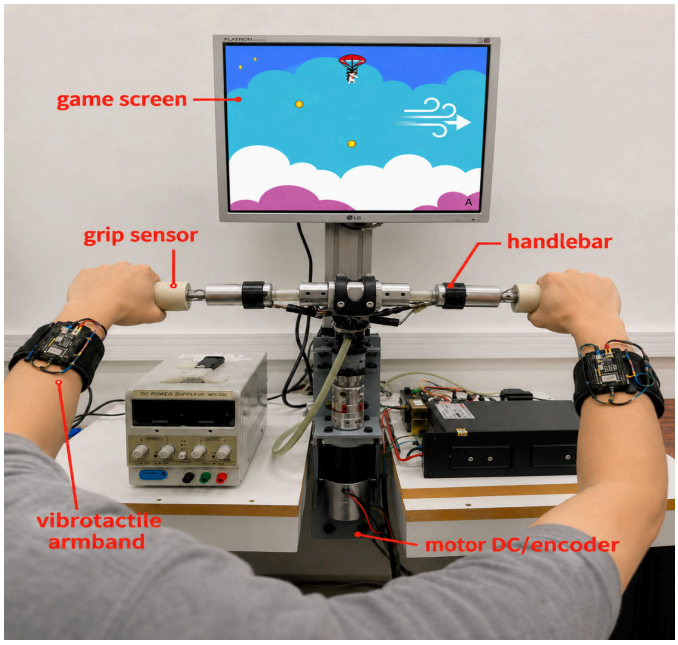


Fig. 1. Experimental setup showing the robotic handlebar, vibrotactile armbands, game screen, and main system components.

In impedance control, the robot is controlled so that the relationship between force and motion is governed by predefined mechanical parameters, such as inertia, damping, and stiffness. In the present system, this relationship can be expressed in terms of the handlebar rotational motion as

$$\tau_h = I_h(\ddot{\theta}_d - \ddot{\theta}_h) + B_h(\dot{\theta}_d - \dot{\theta}_h) + K_h(\theta_d - \theta_h), \quad (1)$$

where I_h, B_h, and K_h are the virtual inertia, damping, and stiffness terms, respectively, and θ_d is the desired angular position. In the present implementation, the desired angular position was kept as a constant during each disturbance, and no virtual inertial term was explicitly rendered [26]. Therefore, considering $\dot{\theta}_d = 0$, $\ddot{\theta}_d = 0$, the motor behavior was modeled as a rotational first order impedance controller (spring-damper behavior), in which the torque applied by the motor is given by

$$\tau_h = K_h(\theta_d - \theta_h) - B_h\dot{\theta}_h \quad (2)$$

The stiffness term K_h defines a virtual torsional spring, indicating that the restoring torque increases proportionally with the angular displacement from the desired position. The derivative term provides damping B_h, opposing motion and improving stability during interaction.

Fig. 2 shows a simulated response of the system when τ_h acts to maintain the handlebar at the setpoint defined by the game dynamics, which are described in the following section.

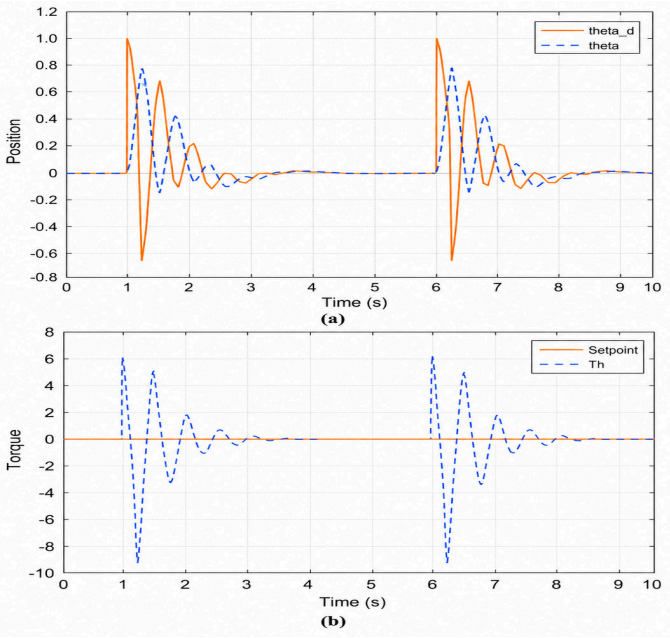


Fig. 2. Simulated response of the robotic handlebar under haptic torque control. The haptic torque τ_h drives the handlebar position toward the setpoint defined by the game dynamics. a) Position response and, b) Torque response.

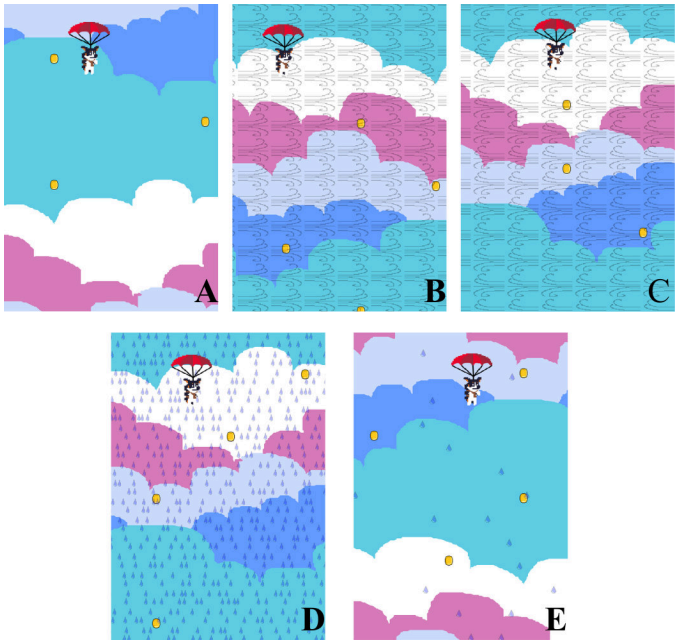


Fig. 3. The game task including the four game conditions. a) Without any visual feedback. b) With wind feedback to the left. c) With wind feedback to the right. d) With a rain intensity of 0.6 and e) With a rain intensity of 0.1.

B. The Wearable Vibrotactile Armband

To deliver the vibrotactile stimuli, two wearable armbands were developed using elastic bands coupled to embedded electronic circuits controlled by ESP32 microcontrollers, with each armband connected to two vibration motors. One armband was placed on each forearm. A third ESP32 board was connected to the computer running the game and acted as the host device, while the two armband circuits operated as feedback receiver boards. This configuration allowed the game to send commands wirelessly to both armbands and trigger the vibrotactile stimuli according to the visual events presented during gameplay.

The choice of ESP32 microcontrollers aimed to reduce the communication latency between the game and the vibrotactile actuation, thereby improving the responsiveness of the stimulation and the real time user experience [27]. Accordingly, the ESP-NOW communication protocol was selected for inter-board communication due to its low-latency performance [28].

Regarding the control system, the receiver boards were programmed to operate using predefined vibration pulses corresponding to an update rate of 5 Hz. This value was empirically defined to ensure user safety and to allow the stimulus to be interrupted in the event of communication loss or game system failure. The pulsed strategy was also adopted to reduce the perception of possible latency fluctuations during gameplay, since the pulse interval was longer than the frame-to-frame variations observed in preliminary tests.

Commands were sent from the computer to the host ESP32 board through serial communication. The host board then transmitted the vibration parameters to the receiver boards embedded in the armbands using a reduced data packet containing the vibration frequency and duty cycle. After receiving these parameters, each receiver board generated the corresponding pulse-width modulation (PWM) signal and triggered the vibration motors. This architecture allowed the game to control the vibrotactile stimulation in real time according to the rain events presented on the screen.

C. The Serious Game

The serious game adopted to guide the participants' movements was based on the platform previously presented in [29]. However, several adaptations were introduced to align the game mechanics with the objectives of the present experimental protocol. The game was configured to operate at 60 frames per second, providing smooth visual feedback during task execution, while experimental variables, including position, force data, and game events, were recorded at approximately 30 Hz.

During gameplay, the participant controls a character that moves downward on the screen while attempting to collect coins distributed along the path (Fig. 3). The horizontal position of the character is controlled through the encoder signal embedded with handlebar's DC motor, and the main task objective is to collect as many coins as possible during each trial. Each repetition of a given condition consisted of a 60-second trial in which 50 coins were available for collection. A set of five coin paths was generated using a randomization algorithm and then assigned across all game conditions. These paths were presented in the same order for each condition, ensuring that participants experienced equivalent trajectory sequences across the experimental phases.

The game presents different conditions based on visual or haptic feedback. These conditions can be implemented individually or together during gameplay. These feedback can be divided into visual stimuli, which appear as wind or rain on the game screen (Fig. 3, and physical stimuli, which are presented as haptic feedback. The haptic feedback can be applied in the form of torque, pushing the user's arms to each side, or in the form of vibrotactile cues. Usually, the haptic torques are applied together with the wind on the screen, in order to create a physical wind effect, while the haptic vibrations are applied together with the rain, simulating the effect caused by rain falling on an umbrella.

1) The Wind-Torque Effect: The wind-torque effect is applied in specific regions while the character descends and collects coins. The visual wind effect is introduced through an animation on the game screen (Fig. 3(b) and (c)), in which an unexpected gust appears and pushes the character laterally. Simultaneously, a corresponding physical torque can be applied by the robotic handlebar. Thus, when the wind appears in the game, it not only displaces the on-screen character but also sends a command to activate the motor, which pushes the user's arms in the same direction. This physical wind effect can act either as an assistive or resistive torque. When the wind direction is aligned with the target coin, the effect is considered assistive, otherwise, it is classified as resistive.

The horizontal position of the character is updated at each frame according to

$$x_{k+1} = x_k + \frac{E_k}{5} + v_w, \tag{3}$$

where x_k is the horizontal screen position at each frame k, E_k is the encoder value, and v_w is the wind-induced displacement in pixels per frame. When the character is inside the wind area $v_w = 5m_w$, being 5 the constant of wind magnitude, when bigger, stronger wind and m_w defines the direction of the wind ($m_w > 0$, push to the right and, $m_w < 0$, push to the left). When the character is outside the wind region $v_w = 0$.

For the haptic stimulus, the encoder position is converted to angular position as

$$\theta_h = \frac{E_k}{1000}\pi. \tag{4}$$

So, to define whether the haptic torque is assistive or resistive, let x_k be the horizontal position of the character on the screen and x_{c+1} the horizontal position of the next coin. The screen-based positional error is defined as

$$e_x = x_{c+1} - x_k, \tag{5}$$

where $e_x > 0$ indicates a required rightward movement and $e_x < 0$ indicates a required leftward movement. The assistive haptic torque is then applied in the same direction to the required movement, following the control law defined by (1), such that

$$\tau_h = \begin{cases} +|\tau_h|, & \text{if } e_x > 0 \\ -|\tau_h|, & \text{if } e_x < 0 \\ 0, & \text{if } e_x = 0 \end{cases}. \tag{6}$$

In the resistive condition, the haptic torque acts in the opposite direction to the required movement:

$$\text{sign}(\tau_h) = -\text{sign}(e_x). \tag{7}$$

In this context, rightward displacements on the game screen correspond to clockwise rotation of the handlebar described by positive τ_h in (6), whereas leftward displacements correspond to counterclockwise rotation and negative τ_h. Since the maximum actuator torque is 2.2 Nm, the resistive torque stimulus was limited to 70% of this value through the PWM command to avoid overloading the DC motor.

2) The Rain-Vibrotactile Effect: The rain stimulus was presented in two regions at different vertical positions along the character's descending trajectory. When the character interacted with a rain region, the corresponding vibrotactile stimulus was triggered. This interaction was quantified using a collision-based evaluation between the avatar and the raindrops, allowing the vibrotactile response to vary according to the avatar displacement during the game loop.

Two parameters defined each rain block: fill and sensitivity. The fill parameter was determined by rain intensity and speed. Rain intensity ranged from 0 to 1 and directly affected droplet distribution, whereas speed defined the number of rain layers generated at each game update. The combination of these parameters allowed different configurations of the same visual stimulus, producing variations in the number of droplets colliding with the avatar (Fig. 3(d) and (e)).

Because droplet collisions depend on both rain speed and intensity, modeling the relationship between raindrop interaction and vibration intensity would require a three-dimensional mapping. However, rain speed was kept constant at the default value of 5, since its manipulation was outside the scope of this study. This reduced the mapping to a two-dimensional problem, allowing a sigmoid curve to be fitted to the maximum number of droplets obtained across different rain intensities (Fig. 4). Thus, for each rain intensity value, the expected maximum number of droplet collisions could be estimated and used to define the threshold for the maximum vibration response.

To create different levels of association between the visual rain effect and the vibrotactile response, the sensitivity parameter was used to define the minimum level of interaction required to reach the maximum vibration intensity. In practice, sensitivity determined the minimum number of droplet collisions needed to trigger the highest vibrotactile response. Therefore, the sensitivity value (S), ranging from 0 to 1, was incorporated into the sigmoid approximation as

$$D_5 = \frac{268.43}{1 + e^{-5.21\cdot(S-0.81)}} - 3.98, \tag{8}$$

where D_5 represents the expected maximum number of droplets for a rain speed equal to 5. Both vibration intensity

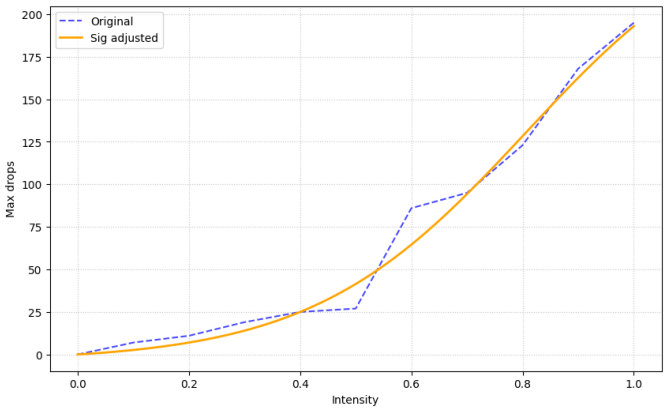


Fig. 4. Sigmoid fit on the curve of maximum droplets per intensity. The curve is rendered two-dimensional by holding the rain speed constant at 5, which establishes a fixed fall velocity of 5 pixels per update for the rain layers, in addition to a predefined minimum spacing between rain lines.

and frequency were determined from the vibration rate, which was calculated as the ratio between the number of droplets colliding with the avatar in the current frame and the previously defined value of D_5.

The vibration intensity was defined by multiplying the vibration rate by 255, corresponding to the maximum duty cycle accepted by the PWM function of the receiver board. For the vibration frequency, the dynamic response of the vibration motor was also considered, since low duty-cycle values combined with high frequencies may not provide sufficient activation to overcome motor inertia and generate a perceivable stimulus. Therefore, the frequency was adjusted according to the vibration rate using the following equation

$$V_f = 1 + (89 \cdot V_r) + (110 \cdot V_r^2), \tag{9}$$

where V_f is the vibration frequency and V_r is the vibration rate. This formulation makes the vibration frequency dependent on the vibration rate while limiting its maximum value to 200 Hz, ensuring an appropriate relationship between stimulus intensity and frequency.

D. Experimental Protocol

All experimental tests were performed at the Biomechatronics Laboratory of the Escola Politécnica, University of São Paulo. The study protocol was approved by the Ethics Committee of the University Hospital of the University of São Paulo (CAAE: 90632625.7.0000.0076), and written informed consent was obtained from all participants prior to experimental session.

The experimental protocol included six healthy adults volunteers (1 woman) with ages ranging from 20 to 41 years (mean: 27.00, SD: $\pm$ 7.13). All participants were naive to the experimental conditions, having no prior participation in any experimental procedure involving the handlebar or any recent musculoskeletal injuries to the upper limbs and trunk, as well as any cognitive-motor impairments.

Participants were equipped with vibrotactile armbands in both arms (Fig. 1) and performed the entire task while standing in front of the robotic handlebar, with their feet kept stationary. They were instructed to hold the handlebar with both hands and perform rotational movements of up to 45° to each side. Before the session, participants received a brief explanation about the game objectives and were informed that they might need to adapt to different game situations. However, they were not informed about the dynamics of the stimuli.

The first exposure to the game occurred during the first trial, as no prior familiarization or adaptation phase was applied. All volunteers performed the same four game conditions, which differed according to the type of stimulus provided. Each condition consisted of five trials and was performed in the same randomized coin path order by all participants. The first condition, considered the baseline, did not provide any visual or haptic sensory stimulus to the user, as shown in Fig. 3(a). The second condition introduced the visual rain stimulus (Fig. 3(d)), while the subsequent condition included vibrotactile stimulation combined with the appearance of rain on the game screen. The fourth and final one, implemented the visual wind on the game screen, combined with resistive haptic torque applied by the robotic handlebar actuator. The experimental protocol lasted approximately 20 minutes in total, consisting of 20 one-minute trials performed continuously, without rest intervals between trials or condition changes.

E. Data Analysis

The Friedman test was used to evaluate overall differences across the four experimental conditions. When a significant effect was observed, post-hoc planned paired Wilcoxon signed-rank tests with Holm correction were performed for specific condition contrasts. These choices are justified by the limited participant cohort and the non-parametric nature observed in the data.

Trajectory accuracy was quantified using the straight-line RMS error between consecutive coins. For each pair of consecutive coins, the ideal trajectory was defined as the straight-line segment connecting both positions. The actual player trajectory between these events was compared with this segment by computing, for each sample, the shortest Euclidean distance to the corresponding line segment. The RMS of these distances was then calculated for each trial and used as the trajectory error metric.

For all tests performed, statistical significance was set at 95% ($p = 0.05$). The analysis were performed using Python 3.13.0 and its specific statistical libraries, such as pandas, scipy.stats, statsmodel, and numpy.

III. Results

The coin collection rate differed across sequences ($\chi^2(3) = 14.60$, $p = 0.0022$). The percentage of available coins collected showed high values for all conditions, from $92.89 \pm 5.22\%$ in the baseline condition (Condition 1) to $97.93 \pm 1.91\%$ during visual rain (Condition 2), $98.88 \pm 1.30\%$ during visual rain combined with vibrotactile stimulation (Condition 3), and

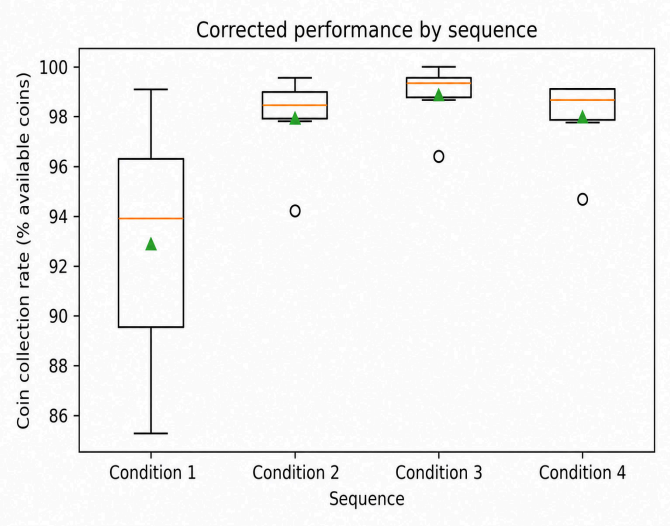


Fig. 5. Collected coin percentage across conditions

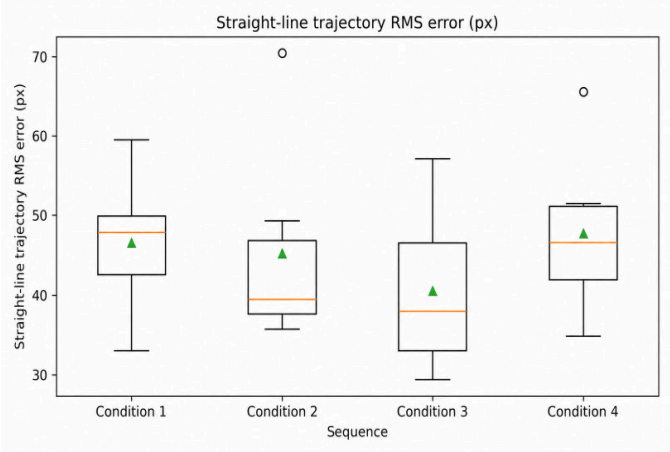


Fig. 6. Straight-line trajectory RMS error by condition.

$98.00 \pm 1.71\%$ during visual wind combined with resistive torque (Condition 4). Planned Wilcoxon comparisons suggested higher performance in the later conditions, Condition 3, ($\Delta = 0.95$, $p = 0.0313$, and Condition 4, ($\Delta = 5.10$, $p = 0.0313$. However, none of these planned comparisons remained significant after Holm correction, $p_{\text{Holm}} > 0.05$.

Significant Friedman effects were observed in device controlling behavior for mean absolute lateral velocity ($\chi^2(3) = 11.40$, $p = 0.0097$), acceleration RMS ($\chi^2(3) = 10.20$, $p = 0.0169$), jerk RMS ($\chi^2(3) = 11.00$, $p = 0.0117$), and displacement recorded by the encoder activity ($\chi^2(3) = 12.20$, $p = 0.0067$).

Relative to baseline, the visual wind combined with resistive torque, Condition 4, increased mean absolute lateral velocity ($\Delta = 29.60$ px/s, $p = 0.0313$), and its derivative terms, acceleration RMS ($\Delta = 227.36$ px/s^2, $p = 0.0313$), jerk RMS ($\Delta = 9457.99$ px/s^3, $p = 0.0313$), and encoder-derived lateral displacement ($\Delta = 0.042$ cm/frame, $p = 0.0313$).

In contrast, the additional effect of vibrotactile stimulation beyond visual rain was small for the full-trial outcomes. Compared with Condition 2, Condition 3 showed only minor changes in mean absolute lateral velocity ($\Delta = 2.22$ px/s, $p = 0.5625$), acceleration RMS ($\Delta = 13.73$ px/s^2, $p = 0.8438$), jerk RMS ($\Delta = 668.05$ px/s^3, $p = 0.8438$), and encoder-derived lateral displacement ($\Delta = -0.001$ cm/frame, $p = 0.8438$).

Since the stimuli were applied only at specific moments during each trial, it was performed an event-based analysis comparing periods in which the stimulus was active with periods in which it was inactive within the same trial. The difference between active and inactive changes in lateral velocity differed significantly across event conditions ($\chi^2(2) = 10.33$, $p = 0.0057$). The increase in lateral velocity during active events was 30.13 ± 11.62 px/s in Condition 2, 35.77 ± 18.11 px/s in Condition 3, and 69.12 ± 6.98 px/s in Condition 4. Paired comparisons showed that Condition 4 presented a higher active versus inactive velocity change than Condition 3 ($\Delta = 33.34$ px/s, $p = 0.0313$) and Condition 2 ($\Delta = 38.99$ px/s, $p = 0.0313$). However, these comparisons did not remain significant after Holm correction ($p_{\text{Holm}} = 0.0938$). Condition 1 was not included in the event-based analysis because it did not contain active visual or haptic events.

The event-related acceleration RMS, jerk RMS, and direction-change deltas did not show significant Friedman effects. No significant effect was found for acceleration RMS ($\chi^2(2) = 4.33$, $p = 0.1146$), jerk RMS ($\chi^2(2) = 4.33$, $p = 0.1146$), or direction changes ($\chi^2(2) = 3.00$, $p = 0.2231$). Nevertheless, the descriptive pattern suggested stronger event-related changes in Condition 4 for acceleration RMS and direction changes. The active and inactive relation in acceleration RMS was 203.15 ± 50.81 px/s^2 in Condition 2, 242.69 ± 66.71 px/s^2 in Condition 3, and 335.80 ± 70.56 px/s^2 in Condition 4. Planned Wilcoxon comparisons suggested higher acceleration changes in Condition 4 compared with Condition 3 ($\Delta = 93.11$ px/s^2, $p = 0.0625$, $p_{\text{Holm}} = 0.1875$) and Condition 2 ($\Delta = 132.65$ px/s^2, $p = 0.0625$, $p_{\text{Holm}} = 0.1875$), although these comparisons were not significant.

Similarly, active versus inactive direction changes were 0.25 ± 1.43 reversals/min in Condition 2, 0.77 ± 1.22 reversals/min in Condition 3, and 3.39 ± 2.26 reversals/min in Condition 4. Planned Wilcoxon comparisons indicated higher direction-change deltas in Condition 4 compared with Condition 3 ($\Delta = 2.62$ reversals/min, $p = 0.1563$, $p_{\text{Holm}} = 0.3125$) and Condition 2 ($\Delta = 3.14$ reversals/min, $p = 0.0938$, $p_{\text{Holm}} = 0.2812$). However, these effects did not remain statistically significant.

The straight-line trajectory RMS error variation is shown in Fig. 6 and did not differ statistically across sequences ($\chi^2(3) = 4.20$, $p = 0.2407$). The RMS error was 46.61 ± 8.96 px in the baseline, Condition 1, 45.25 ± 13.23 px during Condition 2, 40.54 ± 10.72 px in Condition 3, and 47.78 ± 10.62 px during Condition 4. The largest descriptive difference was observed between Condition 4 and Condition 3 ($\Delta = 7.24$ px, $p = 0.0313$, $p_{\text{Holm}} = 0.0938$).

IV. Discussion

The score analysis suggests ceiling effect across all experimental phases, as participants consistently collected over 92% of available coins regardless of condition (Fig. 5). Because small percentage shifts near ceiling level lack practical significance and failed to survive Holm correction, coin collection rates should not be interpreted as evidence of performance enhancement. Instead, the game task served primarily as an engaging tool to ensure active user participation, allowing us to evaluate the technical feasibility of synchronized multimodal feedback and analyze the underlying kinematic patterns during handlebar control.

Since this was a preliminary evaluation of the device and its game-based sensory stimuli, the interpretation of the results is limited by the small number of participants, the fixed order of conditions, and the inclusion of only healthy young adults. Future studies should use larger and counterbalanced samples and include participants with stroke or other motor impairments to better isolate the effects of each stimulus and determine whether the observed kinematic responses are relevant to rehabilitation.

Regarding the kinematic behavior during robotic handlebar control, the results indicate that the experimental conditions modulated movement intensity and device control patterns. The clearest effect was observed in the visual wind combined with resistive torque (Condition 4), which increased lateral velocity, acceleration, jerk, and encoder-derived displacement compared with the baseline condition. These findings suggest that resistive torque combined with the wind on the game screen imposed a higher control demand on the user, leading to more intense and variable handlebar movements during gameplay. This interpretation was also supported by the event-based analysis. When the stimulus events were analyzed within each trial, Condition 4 produced the largest increase in lateral movement velocity during active periods.

In contrast, vibrotactile stimulation did not show a significant additional effect when compared with the visual rain condition alone. This suggests that the visual stimulus may have been more relevant to task behavior, or that the effect of vibration was partially masked by the simultaneous visual rain stimulus. Therefore, future experiments should also include an isolated vibrotactile condition to better identify the specific contribution of vibrotactile stimulation to user performance and handlebar control.

The trajectory error analysis suggests that the sensory condition may modulate the straight line trajectory control, since the trajectory RMS error did not progressively decrease across trials, contrary to what would be expected if participants showed a clear trajectory-based adaptation effect. Although the differences across conditions were not statistically significant, a descriptive trend was observed in Condition 3, with vibrotactile stimulation, which presented the lowest trajectory deviation. This may indicate that this stimulation helped stabilize the movement trajectory, although this interpretation should be confirmed in future studies with a larger sample size.

Although most planned comparisons did not remain significant after Holm correction, the descriptive trends observed in this small cohort provide important preliminary information. These trends suggest the different effects that might be caused by sensory stimuli, helping to guide the use of the device and the design of future experimental protocols. Overall, the kinematic variations observed across our game conditions highlight the system's capacity to concurrently modulate motor demand and deliver sensory feedback, providing a more comprehensive sensorimotor challenge than standard setups that rely only on motor recovery or sensory stimulation.

While this pilot study demonstrates the feasibility of the multimodal feedback system, the fixed-order condition design followed by all participants limits our ability to draw causal conclusions. Because of this fixed order, the condition-associated differences observed may partially reflect practice effects, fatigue, or general adaptation to the game dynamics, rather than the isolated effects of the sensory stimuli alone. Furthermore, because later conditions added multiple sensory elements simultaneously, specific feedback mechanisms, such as isolating vibration from visual rain or torque from visual wind, cannot be independently evaluated in this work. Future research will employ a design with isolated stimulus conditions to analyze the impacts of each sensory modality.

From a clinical perspective, integrating multimodal sensory feedback into bimanual robotic therapy addresses a relevant gap in post-stroke rehabilitation, where sensory deficits are often under-treated [21]. To validate this system clinically, future studies will evaluate stroke patients across varying degrees of sensory and motor impairment. These upcoming trials will incorporate randomized and counterbalanced protocols with isolated stimulus conditions, adaptive force feedback tailored to individual patient capability, standard clinical outcome measures, such as the Fugl-Meyer Assessment (for both motor and sensory subscales), and kinematics-derived smoothness metrics to quantify therapeutic efficacy and functional recovery.

V. Conclusion

Even as an initial proof of concept, the present results demonstrated the technical feasibility of synchronizing bimanual robotic control of a serious game with multimodal visual, haptic, and vibrotactile feedback. The preliminary results support the feasibility of the proposed system for investigating how different sensory conditions affect task performance and handlebar control. The results also allowed us to identify kinematic patterns that may be influenced by the different types of sensory stimuli, although no causal effect of the different stimuli can be draw directly from the current protocol. Furthermore, the exposure to each stimulus appeared to modify the way participants controlled the handlebar, making the task more demanding or easier to execute depending on the condition. In particular, resistive torque combined with the visual wind effect tended to require more intense movements and greater control adjustments, whereas visual rain combined with

vibrotactile stimulation tended to produce straighter movement trajectories. The limited cohort of healthy volunteers restricts the generalizability of these findings, however future work will focus on expanding the sample size, counterbalancing conditions, and conducting clinical validation trials with post-stroke individuals to assess the therapeutic impact of multimodal sensory feedback on sensorimotor recovery.

ACKNOWLEDGMENT

AFC acknowledges support from FAPESP [Grant No 2024/01120-4], CNPq PQ [Grant No 311167/2022-5] and Pró-Reitoria de Pesquisa e Inovação da Universidade de São Paulo [ASSISTLAB].

REFERENCES

[1] M. A. Cervera, S. R. Soekadar, J. Ushiba, J. del R. Millán, M. Liu, N. Birbaumer, and G. Garipelli, "Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis," 2018.

[2] M. Coscia, M. J. Wessel, U. Chaudary, J. del R. Millán, S. Micera, A. Guggisberg, P. Vuadens, J. Donoghue, N. Birbaumer, and F. C. Hummel, "Neurotechnology-aided interventions for upper limb motor rehabilitation in severe chronic stroke," 8 2019.

[3] S. Micera, M. Caleo, C. Chisari, F. C. Hummel, and A. Pedrocchi, "Advanced neurotechnologies for the restoration of motor function," 2 2020.

[4] V. L. Feigin, M. Brainin, B. Norrving, S. Martins, R. L. Sacco, W. Hacke, M. Fisher, J. Pandian, and P. Lindsay, "World stroke organization (wso): Global stroke fact sheet 2022," 1 2022.

[5] C. Alia, C. Spalletti, S. Lai, A. Panarese, G. Lamola, F. Bertolucci, F. Vallone, A. D. Garbo, C. Chisari, S. Micera, and M. Caleo, "Neuroplastic changes following brain ischemia and their contribution to stroke recovery: Novel approaches in neurorehabilitation," 3 2017.

[6] H. Carlsson, G. Gard, and C. Brogårdh, "Upper-limb sensory impairments after stroke: Self-reported experiences of daily life and rehabilitation," *Journal of Rehabilitation Medicine*, vol. 50, 2018.

[7] S. D. Doyle, S. Bennett, and B. Dudgeon, "Upper limb post-stroke sensory impairments: the survivor's experience," *Disability and Rehabilitation*, vol. 36, pp. 993–1000, 6 2014.

[8] M. Graafland, M. Dankbaar, A. Mert, J. Lagro, L. De Wit-Zuurendonk, S. Schuit, A. Schaafstal, and M. Schijven, "How to systematically assess serious games applied to health care," 2014.

[9] E. Amengual Alcover, A. Jaume-I-Capó, and B. Moyà-Alcover, "PROGame: A process framework for serious game development for motor rehabilitation therapy," *PLoS ONE*, vol. 13, no. 5, 2018.

[10] Gutiérrez, D. Sepúlveda-Muñoz, Gil-Agudo, and A. de los Reyes Guzmán, "Serious game platform with haptic feedback and EMG monitoring for upper limb rehabilitation and smoothness quantification on spinal cord injury patients," *Applied Sciences (Switzerland)*, vol. 10, no. 3, 2020.

[11] M. Pirovano, E. Surer, R. Mainetti, P. L. Lanzi, and N. Alberto Borghese, "Exergaming and rehabilitation: A methodology for the design of effective and safe therapeutic exergames," *Entertainment Computing*, vol. 14, 2016.

[12] Y. Bouteraa, I. B. Abdallah, A. Ibrahim, and T. A. Ahanger, "Development of an iot-based solution incorporating biofeedback and fuzzy logic control for elbow rehabilitation," *Applied Sciences (Switzerland)*, vol. 10, 2020.

[13] A. Crema, M. Bassolino, E. Guanziroli, M. Colombo, O. Blanke, A. Serino, S. Micera, and F. Molteni, "Neuromuscular electrical stimulation restores upper limb sensory-motor functions and body representations in chronic stroke survivors," *Med*, vol. 3, pp. 58–74.e10, 1 2022.

[14] R. Ranzani, O. Lambercy, J. C. Metzger, A. Califfi, S. Regazzi, D. Dinacci, C. Petrillo, P. Rossi, F. M. Conti, and R. Gassert, "Neurocognitive robot-assisted rehabilitation of hand function: A randomized control trial on motor recovery in subacute stroke," *Journal of NeuroEngineering and Rehabilitation*, vol. 17, 2020.

[15] K. Wolf, A. Mayr, M. Nagiller, L. Saltuari, M. Harders, and Y. Kim, "Pori device: portable hand assessment and rehabilitation after stroke," *at - Automatisierungstechnik*, vol. 70, 2022.

[16] Y. chen Lee, Y. chun Li, K. chung Lin, G. Yao, Y. ju Chang, Y. yun Lee, C. ting Liu, W. ling Hsu, Y. hsuan Wu, H. ta Chu, T. xuan Liu, Y. ping Yeh, and C. Chang, "Effects of robotic priming of bilateral arm training, mirror therapy, and impairment-oriented training on sensorimotor and daily functions in patients with chronic stroke: study protocol of a single-blind, randomized controlled trial," *Trials*, vol. 23, 2022.

[17] Y. chun Li, K. chung Lin, C. ling Chen, G. Yao, Y. ju Chang, Y. yun Lee, and C. ting Liu, "A comparative efficacy study of robotic priming of bilateral approach in stroke rehabilitation," *Frontiers in Neurology*, vol. 12, 7 2021.

[18] M. Buist, S. Damercheli, M. T. N. Truong, A. Sanna, E. Mastinu, and M. Ortiz-Catalan, "Development and validation of a wearable device to provide rich somatosensory stimulation for rehabilitation after sensorimotor impairment," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 17, 2023.

[19] A. B. Conforto, A. G. Machado, I. Menezes, N. H. V. Ribeiro, R. Luccas, D. S. Pires, C. da Costa Leite, E. B. Plow, and L. G. Cohen, "Treatment of upper limb paresis with repetitive peripheral nerve sensory stimulation and motor training: Study protocol for a randomized controlled trial," *Frontiers in Neurology*, vol. 11, 3 2020.

[20] J. B. Kroth, B. Handfas, G. Rodrigues, F. Zepeda, M. A. Oliveira, D. J. Wang, R. M. de Azevedo Neto, G. S. Silva, E. Amaro, I. O. Sorinola, and A. B. Conforto, "Effects of repetitive peripheral sensory stimulation in the subacute and chronic phases after stroke: Study protocol for a pilot randomized trial," *Frontiers in Neurology*, vol. 13, 2 2022.

[21] K. N. Arya, S. Pandian, G. G. Agarwal, N. Chaudhary, and A. K. Joshi, "Effect of neuroplasticity-principles-based sensory-rehabilitation (nepser) on sensori-motor recovery in stroke: study protocol for a randomized controlled trial," *Neurological Research and Practice*, vol. 3, 2021.

[22] S. Lv, Y. Chen, M. Liu, L. Qin, Z. Liu, W. Liu, M. Cui, X. Zhang, H. Yan, F. Ning, H. Zhang, and Y. Xu, "Progress of proprioceptive training in the treatment of traumatic shoulder instability," 2022.

[23] A. Sidarta, Y. C. Lim, R. A. Wong, I. O. Tan, C. W. K. Kuah, and W. T. Ang, "Current clinical practice in managing somatosensory impairments and the use of technology in stroke rehabilitation," *PLOS ONE*, vol. 17, p. e0270693, 8 2022.

[24] L. R. Cardoso, L. M. Pedro, and A. Forner-Cordero, "Handlebar robotic system for bimanual motor control and learning research," *Sensors*, vol. 21, pp. 1–20, 2021.

[25] N. Hogan, "Impedance control: An approach to manipulation: Part i - theory," *Journal of Dynamic Systems, Measurement, and Control*, 1985.

[26] T. Hulin, C. Preusche, and G. Hirzinger, "Stability boundary and design criteria for haptic rendering of virtual walls," in *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 8, 2006.

[27] D. Sim, Y. Baek, M. Cho, S. Park, A. S. M. S. Sagar, and H. S. Kim, "Low-latency haptic open glove for immersive virtual reality interaction," *Sensors*, vol. 21, no. 11, 2021.

[28] D. Eridani, A. F. Rochim, and F. N. Cesara, "Comparative performance study of esp-now, wi-fi, bluetooth protocols based on range, transmission speed, latency, energy usage and barrier resistance," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 322–328, Sep. 2021.

[29] A. E. H. Dib, J. M. Nascimento, L. R. L. Cardoso, and A. Forner-cordero, "Analysis of motor control and learning in human-robot interaction during game guided movements," in *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2023*, vol. 6, pp. 2–5, IEEE EMBS, 2023.

Reliability and Usefulness of Kinovea Motion Analysis Software for Direct Measurement of Vertical Height During Counter Movement Jump in Down Syndrome

Gabriel de Oliveira Silva*, Thaynara Soares Amaral†, Nathalya Guedes Mota†, Clarissa Cardoso dos Santos Couto Paz†, and Adriano de Oliveira Andrade*,
*Centre for Innovation and Technology Assessment in Health (NIATS),
Faculty of Electrical Engineering, Federal University of Uberlândia (UFU), Uberlândia, Brazil
†Group in Physical Therapy for Neurological Disorders (GEFIN),
Faculty of Health Sciences and Technologies, University of Brasília, Brasília, Brazil

***Abstract*— Jumping performance is a critical indicator of motor development and neuromuscular function. Individuals with Down Syndrome (DS) often exhibit motor deficits, such as hypotonia and reduced muscle strength, which impact their ability to perform complex movements like the counter movement jump (CMJ). Although motion capture systems offer quantitative insights, their high cost and complexity often limit their clinical use. The objective of this study is to evaluate the intra-examiner and inter-examiner reliability of a low-cost, 2D video-based method using Kinovea® software to quantify CMJ height in individuals with DS. Thirty-two individuals with DS (mean age 21.71 ± 7.14 years) participated in this study. The protocol involved standardized instructions, a warm-up phase, and the performance of 10 consecutive CMJ cycles. An anatomical marker was placed on the greater trochanter, and movements were recorded and analyzed using Kinovea® (version 2025.2). Reliability was assessed using the intraclass correlation coefficient (ICC), standard error of measurement (SEM), minimum detectable change (MDC), and Bland-Altman plots, evaluating test-retest, intra-examiner, and inter-examiner agreements, Additionally, the Standard Error of Measurement (SEM), and the Minimum Detectable Change (MDC), obtained through, were calculated to verify the precision and sensitivity of the measurements. The analysis revealed high agreement, with mean differences between measurements approaching zero ($\boldsymbol{p > 0.05}$). Furthermore, the method proved robust, being applicable even among participants with varying cognitive levels, as identified by the Mini-Mental State Examination. This 2D video-based assessment protocol demonstrates excellent intra-examiner and inter-examiner reliability. Therefore, low-cost method can be a viable alternative for clinical and research settings to quantify jumping height in individuals with Down Syndrome, providing a methodology for data collection and analysis for this population.**

Keywords — Down Syndrome, Kinematics, Kinovea, Reliability, Vertical Jump.

I. INTRODUCTION

Jumping is one of the fundamental movements performed by humans after they learn to walk and run [1]. The vertical jump is often used as a performance test, as it is influenced by multiple factors, including maximal strength and neuromuscular coordination of the upper and lower body segments [2]. Throughout life, humans undergo progressive changes in motor development, characterized by continuous adaptations in motor behavior across the entire life cycle [3].

Among the stages of this process, the development fundamental motor skills is particularly significant as it plays a decisive role in because they bring about decisive changes in an individual's subsequent motor development. In early childhood, typically developing children motor skills. Children go through stages in which they initially attempt to perform basic movement patterns at around 3 years of age, acquire higher coordination and control between 4 and 5 years of age, and achieve efficiency by execution and movement coordination between the ages of 6 and 7 [4]. These skills — such as running, jumping, and catching — allow for extensive exploration of the environment [4]. This development is influenced by genetic, environmental, and learning factors, as well as the degree of motor experience exposure, which results in individual variations in the rate of motor skill acquisition among children [5]. However, achieving a mature level of motor skill proficiency requires appropriate stimulation and instruction [6], Furthermore, a lack of mastery of a locomotor skill acts as a barrier to the subsequent development of more advanced and complex skills [7].

From a genetic perspective, Down Syndrome (DS) is the most common chromosomal condition and is associated with motor development impairments [3]. The phenotype of DS includes manifestations that affect multiple systems, including the musculoskeletal, neurological, and cardiovascular systems [8].

The most frequent impairments include hypotonia, ligamentous laxity, decreased muscle strength, insufficient muscle co-contraction, postural control deficits, short stature, and cognitive delay [9]. In this context, jumping stands out as a versatile, functional, and essential locomotor skill and is considered one of the first motor skills that children are expected to master [10]. The vertical jump can be used to assess the stage of motor development, because it involves the vertical projection of the body into the air through propulsion from one or both feet, followed by a bipedal landing [3].

As such, it provides valuable information on human locomotion control [11]. Jump is a skill commonly present in games, play, and daily activities [12]. Moreover, jump height has been widely measured by professionals in sports performance and rehabilitation as a practical alternative to the

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

direct assessment of maximal lower-limb strength and power [13].

Although individuals with Down Syndrome exhibit poorer jumping skills than typically developing individuals, this deficit can be observed from childhood onward and persists into adulthood [9]. Various technologies can be used to quantitatively assess jumping performance, including inertial measurement units (IMUs) and three-dimensional motion capture systems [10]. However, the use of these devices presents barriers in clinical practice, mainly due to their medium-to-high cost, complexity of use, and limited usability [10]. In addition, their application often requires specialized knowledge to process and analyze raw data. In recent years, video analysis software has been developed to facilitate the use of low-cost quantitative assessment methods [11].

In this context, Kinovea® (version 2025.2) is a free motion analysis software designed for kinematic studies, with a primary focus is on capturing, observing, annotating, and measuring motion variables [12]. Therefore, the purpose of this study was to evaluate the intra-examiner and inter-examiner reliability of a low-cost method for quantifying counter-movement vertical jump (CMJ) height in individuals with Down Syndrome through the analysis of linear measurements using Kinovea® software.

II. METHOD

This study was approved by the Research Ethics Committee of the Faculty of Health Sciences and Technologies (FCTS) under approval number 6.046.308, on May 8, 2023. All procedures were conducted in accordance with the applicable ethical guidelines for research involving human participants and Resolution CNS 466/12.

For the kinematic analysis of spatiotemporal variables during the counter-movement jump (CMJ), the free motion analysis software Kinovea® (version 2025.2) was used [12]. Widely validated in the literature for clinical motor and biomechanical assessment, this two-dimensional (2D) video analysis tool enables the semi-automated tracking of anatomical landmarks and the measurement of kinematic parameters [12].

As mentioned throughout the text, in tables III, IV, and V present N as the number of evaluated jumps for jump height, and as the number of evaluated participants for maximum and mean height. Tables IV and V also report the Intraclass Correlation Coefficient (ICC), Standard Error of Measurement (SEM), and Minimum Detectable Change (MDC).

To classify cognitive status, the Mini-Mental State Examination (MMSE) instrument was used to screen for cognitive impairment, detect cognitive decline, and monitor the progression of diseases. It assesses eleven domains, such as memory (immediate and recall), temporal and spatial orientation, calculation, naming, repetition, following commands, reading, sentence construction, and drawing [13].

The data collection and analysis methodology was developed by the institutionalized extension project “Group in Physical Therapy for Neurological Disorders” (GEFIN) at the Faculty of Health Sciences and Technologies (FCTS) campus, on the University of Brasília (UnB), in partnership with Center for Innovation and Technological Assessment in Health (NIATS) at the Faculty of Electrical Engineering (FEE) of the Federal University of Uberlândia (UFU). Video recordings were imported into the software, in which spatial calibration and frame-by-frame analysis were performed to extract quantitative data.

A. Participant Selection Criteria

To ensure sample homogeneity and participant of the specific eligibility criteria were applied during the recruitment phase.

1) *Inclusion criteria*

- Individuals with a clinical diagnosis of Down Syndrome (DS);
- Aged between 10 and 36 years;
- Ability to understand and follow simple or adapted instructions.

2) *Exclusion criteria*

- Diagnosed atlantoaxial instability (confirmed by clinical examination or imaging);
- Presence of cardiovascular, respiratory, or metabolic conditions that contraindicate moderate-intensity physical activity (as determined by a medical evaluation);
- Presence of recent musculoskeletal injuries (within the previous 6 months) affecting the lower extremities;
- Use of devices that directly interfere with jump execution (e.g., rigid restraints).
- Low-resolution videos in which it was not possible to follow the protocol.

B. General Produces

The overall experimental workflow and chronological stages of the study are schematically represented in Fig. 1 to provide a clear, visual overview of the systems and procedures. Initially, participants and their legal guardians signed the Informed Consent Form (ICF) (Step 1). Subsequently, baseline anthropometric measurements and the Mini-Mental State Examination (MMSE) were completed to characterize the participants' profiles (Step 2). The experimental setup then began with orthostatic positioning and system calibration (Step 3), followed by the precise placement of reflective anatomical markers (Step 4).

To ensure a proper understanding of the motor task, the examiner performed a physical demonstrated jump (Step 5), after which participants performed three practice trials for familiarization (Step 6). During these trials, technical corrections and verbal feedback were provided during these trials to ensure correct execution (Step 7). Finally, participants performed the experimental protocol, consisting of 10 consecutive countermovement jumps (CMJ) (Step 8).

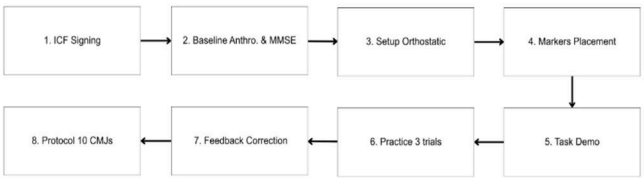


Fig. 1. Flowchart of the experimental procedures and chronological stages of the study, from initial screening to the countermovement jump (CMJ) protocol.

The marker should be positioned over the greater trochanter of the femur, on the side corresponding to the imaging plane, as shown in Fig. 2. To ensure proper placement, with the participant standing, the evaluator should palpate the lateral region of the hip while the participant is standing until the bony prominence of the greater trochanter is identified. Thereafter, the participant should be instructed to perform active internal and external hip rotation movements, following a demonstration by the evaluator, to confirm the correct identification of the structure.

The most prominent point of the greater trochanter identified during the rotational movements should be used as the anatomical reference. Once this point has been pinpointed, the hip should be returned to a neutral position, with no internal or external rotation, and the marker should then be attached to the verified anatomical landmark.

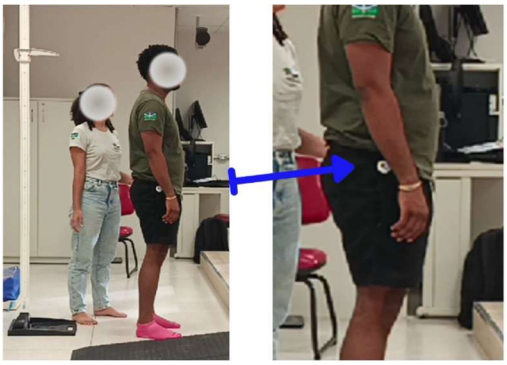


Fig. 2. Marker positioning

C. *Camera Positioning*

The camera or mobile device used for recording must be positioned horizontally, aligned with the participant's sagittal plane. The equipment must be mounted on a stable stand (e.g., a tripod), placed to the side of the participant, at approximately 3 to 5 meters. The camera height should be adjusted to approximately hip level, allowing a full-body view throughout the entire movement. The framing should ensure that the push-off, flight, and landing phases are fully captured, without the participant leaving the field of view at any point during the task.

D. *Participant orientation*

Table I below presents the stages and procedures established to ensure the proper execution of the vertical jump by the participants.

TABLE I. TRAINING AND FAMILIARIZATION PROCEDURES FOR PERFORMING THE COUNTER-MOVEMENT VERTICAL JUMP (CMJ).

Stage	Procedure
Verbal instruction	Participants received standardized instructions regarding the purpose of the test and were instructed to perform a maximum vertical jump using the countermovement jump (CMJ) technique.
Starting position	The participants stood upright, with their feet hip-width apart and their gaze directed straight ahead.
Positioning of the upper limbs	To prevent arm swing from affecting jump performance, participants were instructed to keep their arms extended at their sides or crossed over their chests, with their hands resting on their shoulders, throughout the entire exercise.
Visual demonstration	The evaluator performed a complete demonstration of the movement, including the starting position, the squat phase, and the execution of the jump.
CMJ Execution	Participants were instructed to perform a continuous movement consisting of a rapid flexion of the lower limbs followed immediately by a maximum vertical jump.
Security	Throughout the procedure, the evaluators remained close to the participants to provide support and ensure their safety while they performed the task.
Habituation	Before the official data collection, participants underwent practice sessions to ensure they understood the instructions and could perform the movement correctly.
Feedback	When necessary, guidance was provided regarding the starting position, the positioning of the upper limbs, and the continuous execution of the movement.
Start of collection	The official data collection began only after the participant demonstrated an understanding of the task and the ability to perform the movement correctly.

Measures

Video analysis and the determination of peak height during the countermovement jump (CMJ) were performed by a previously trained evaluator. To assess intra-examiner reliability, the same videos were analyzed twice by the same evaluator, with a 21-day interval between assessments. The same time points were used in both analyses to ensure consistency between measurements. Here, is reported that the sample size (N) was reduced from 42 to 32 analyzed videos in accordance with the exclusion criteria, especially because of the low-resolution videos in which it was not possible to follow the protocol.

Based on the variables presented in Table II, peak height, maximum jump height, and mean jump height were calculated as derived measures. The position of the greater trochanter over time was tracked using the anatomical marker placed directly over it.

TABLE II. Variables Analyzed

Variable	Description
Jump height (cm)	Vertical displacement of the trochanter, calculated as the difference between the maximum height and the initial position.
Maximum height of the greater trochanter (cm)	The highest vertical position reached by the trochanter during the jump.
Average jump height (cm)	The average of the heights recorded for the jumps performed.

III. Analysis Methodology

1) Calibration based on the Individual's height

To calibrate the software, the height (in centimeters) of the person being assessed is used. Following these procedures, as shown in Figure 3, the flowchart illustrates the process.

- Application of the Line Function: The program's line tool (step 3) should be positioned vertically, extending from the base of the feet to the top of the head (step 4);
- Postural Condition: The subject must be filmed in an orthostatic position (standing upright) for 3 seconds immediately before the start of the jump;
- Coordinate Reference: Ensure that the calibration line is perfectly aligned with the system's vertical axis (Y-axis).

2) Setting Up Tracking in Kinovea ("Track Path")

- This step involves using the "Track Path" tool on the marker to configure the software's tracking parameters (step 5);
- Anatomical Point Selection: The tool must be positioned exactly over the anatomical point of interest (marker);
- Tracking Window Dyna Tracking Window Dynamics: When tracking begins, Kinovea will display two bounding boxes around the point: The inner rectangle (Object Window) defines the texture and exact area of the marker that the software should recognize, and the outer rectangle (search window) defines the scan area where the software will search for the marker in the next frame. Adjusting its size is crucial for fast movements such as jumping, to prevent loss of tracking (step 6);

3) Frame-by-frame tracking of the anatomical landmark and corrections as needed until the end of the jump and return to the upright position (step 7);

4) Determination of the initial position and the maximum height of the trochanter in linear kinematics tools (step 8) to vertical positions for identifying peaks and basal values (step 9).

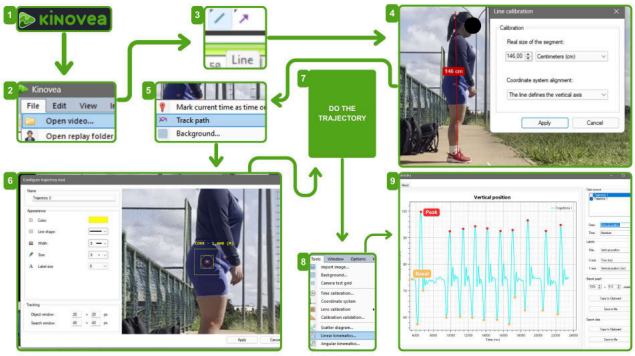


Fig. 3. Workflow for calibration, trajectory tracking, and kinematic analysis of the countermovement jump using Kinovea software.

IV. Statistical Analysis

First, a descriptive analysis of the participants was performed based on clinical and anthropometric variables. Furthermore, the distribution of the variables derived from Kinovea ® during the countermovement jump (CMJ) trials was assessed, and all variables were found to be normally distributed. Subsequently, test-retest and inter-rater reliability analyses were performed.

A descriptive analysis of the participants was initially performed based on clinical and anthropometric variables. Following this, the distribution of the variables derived from Kinovea® during countermovement jump (CMJ) tests was evaluated, and all variables were found to be normally distributed, as shown in Table III. Then, the inter-examiner and intra-examiner reliability for the variables height, maximum height, and average height of the 32 participants was assessed using the intraclass correlation coefficient (ICC) and Bland-Altman plots [14]. Additionally, the Standard Error of Measurement (SEM), and the Minimum Detectable Change (MDC), obtained through, were calculated to verify the precision and sensitivity of the measurements.

All statistical analyses adopted a significance level of $\alpha = 0.05$ and were performed using SPSS Statistics software (Statistical Package for the Social Sciences).

V. Results

This paper analyzed 314 jumps performed by 32 participants with Down syndrome. The participants had a mean age of 21.71 (±7.14) years, mean body mass of 54.88 (±13.04) kg, and a mean height of 148.52 (±10.93) cm. Regarding cognitive performance, the mean MMSE score of 28.72 (±7.16). Notably, five participants (15,63%) scored below 20 points, indicating cognitive impairment according to the cutoff adopted in this study. The Confidence Intervals (CI) and descriptive analysis of the height of the 10 consecutive jumps, including the mean and maximum height, is shown in Table III.

TABLE III. Descriptive analysis of countermovement jump heights obtained during the evaluation of 32 participants with Down syndrome

Variable	N	Mean	Standard deviation	CI (95%) Lower - Upper
Height (cm)	314	17.68	7.7	2.7 - 32.67
Maximum height (cm)	32	21.20	8.4	4.7 - 37.70
Mean height (cm)	32	17.66	7.6	2.7 - 32.60

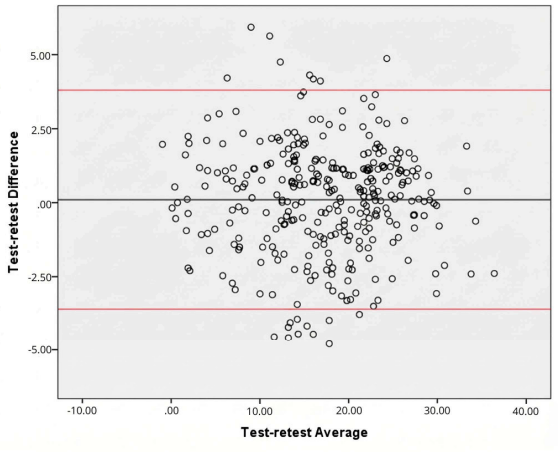
The reliability analysis demonstrated excellent consistency and reproducibility for all jump variables evaluated (height, maximum height, and mean height).

In the intra-examiner evaluation, as shown in Table IV, high ICC values were observed for jump height (0.977; SEM = 1.16; MDC = 3.22), maximum jump height (0.976; SEM = 1.30; MDC = 3.62), and mean jump height (0.982; SEM = 1.02; MDC = 2.83), with Bland-Altman plots confirming good agreement.

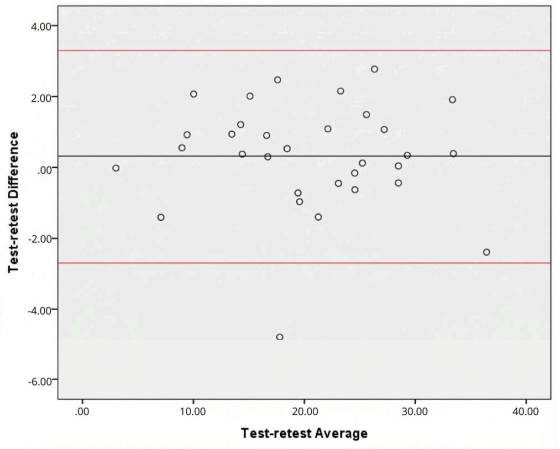
TABLE IV. Descriptive analysis of countermovement jump heights obtained during the evaluation of 32 participants with Down syndrome

Variable	N	ICC	SEM	MDC

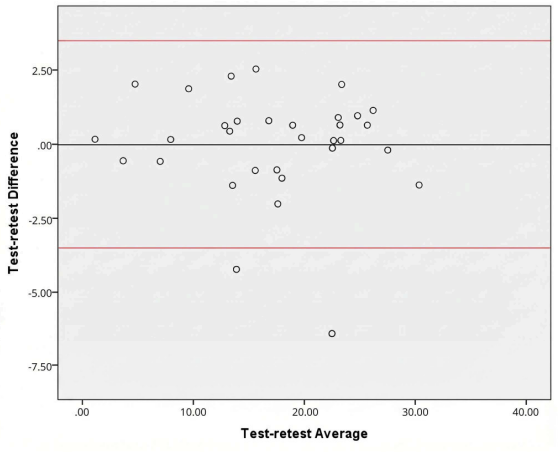
Height (cm)	314	0.977	1.16	3.22
Maximum height (cm)	32	0.976	1.30	3.62
Mean height (cm)	32	0.982	1.02	2.83



A



B



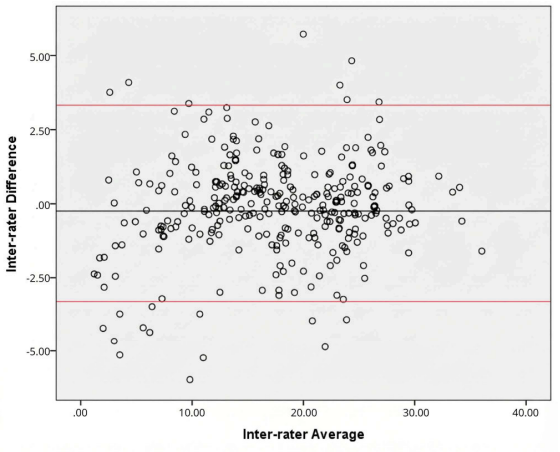
C

Fig. 4. Bland-Altman scatter plots regarding the intra-examiner reliability analysis of countermovement jump variables in participants with Down syndrome. Legend: (A) Graph regarding the dispersion of the height of the 314 jumps performed; (B) Graph regarding the maximum height of the jumps of the 32 participants; and (C) Graph regarding the mean height of the jumps of the 32 participants. The center lines indicate the mean difference, and the horizontal lines represent the 95% limits of agreement.

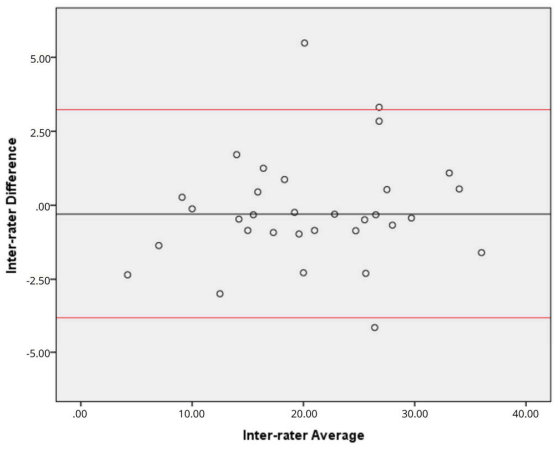
Likewise, the inter-examiner analysis, as shown in Table V, revealed high agreement between evaluators, with ICC coefficients of 0.971 for jump height (SEM = 1.30; MDC = 3.61), 0.976 for maximum jump height (SEM = 1.30; MDC = 3.61), and 0.992 for mean jump height (SEM = 0.68; MDC = 1.89), with the Bland-Altman plots in Figure 5 confirming agreement with mean differences close to zero. These findings indicate that the proposed method provides reliable and reproducible measurements for jump analysis in individuals with Down syndrome.

TABLE V. Inter-examiner reliability analysis of countermovement jump height data obtained during the evaluation of 32 participants with Down syndrome

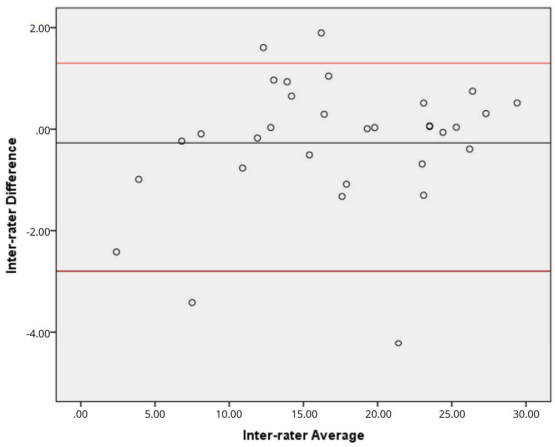
Variable	N	ICC	SEM	MDC
Height (cm)	314	0.971	1.3	3.61
Maximum height (cm)	32	0.976	1.3	3.61
Mean height (cm)	32	0.992	0.68	1.89



A



B



C

Fig. 5. Bland-Altman scatter plots regarding the inter-examiner reliability analysis of countermovement jump variables in participants with Down syndrome.Legend: (A) Graph regarding the dispersion of the height of the 314 jumps performed; (B) Graph regarding the maximum height of the jumps of the 32 participants; and (C) Graph regarding the mean height of the jumps of the 32 participants. The center lines indicate the mean difference and the horizontal lines represent the 95% limits of agreement.

VI. Discussion

The present study included individuals with DS with varying cognitive profiles. Despite the cognitive alterations observed in the sample, all participants successfully performed the jumps following prior instruction and training. These findings support the clinical applicability of the protocols. In addition to the countermovement jump, the participants' cognitive impairment was assessed using MMSE. In this case, of the total number of participants (n = 32), approximately (15.63% (n = 5) had cognitive impairment, based on the cutoff score from the article "Suggestions for the Use of the Mini-Mental State Examination in Brazil" [13].

The findings indicate that quantitative assessment of the CMJ with adequate inter-examiner and intra-examiner agreement is feasible even in individuals with cognitive impairment, provided that appropriate adaptations and visual and auditory cues are implemented, as described in Table I. It should be emphasized that the differences present in this population are significant and must be taken into account, given that children with Down Syndrome exhibit unique sensory characteristics when compared to their peers of the same mental age, whether those peers have autism spectrum disorder, typical development (TD), or a developmental disorder [15].

This characteristic may pose additional challenges in clinical and laboratory settings, particularly when using robust and high-cost motion capture systems. Such systems often require multiple reflective markers attached to the skin, specialized knowledge for system operation and data analysis, and are generally restricted to laboratory environments [10].

Another important aspect regarding clinical applicability is the type of jump analyzed. In the present study, the countermovement jump was selected. When performing this jump, not only is the flight phase considered, but there is also a preparation phase, during which the individual is instructed to perform a squat, followed immediately by the flight phase and takeoff [16]. The choice of the CMJ for clinical assessment should also be considered in relation to the musculoskeletal characteristics of this jump to the osteomyoarticular changes observed in people with Down Syndrome, including muscle hypotonia and joint hypermobility. Thus, the assessment of this type of jump goes beyond the analysis of an isolated movement. Rather, it requires prior mastery of the squat and motor coordination.

In this regard, Yorozu et al. [17] expand the classic concept of fundamental motor skills, which includes squatting and jumping — to foundational skills, which are goal-oriented movement patterns that directly and indirectly impact an individual's ability to be physically active. The development of these skills may enhance participation in different forms of physical activity, sports, and leisure throughout life, thereby contributing to health and social participation [9].

The experimental methodology adopted in this study focused on jump height, particularly maximum jump height, as these kinematic variables are commonly used to characterize CMJ performance [18], [19], [20], [21]. Given the demonstrated intra-examiner and inter-examiner agreement for maximum jump height, the measurements obtained using Kinovea, together with the anthropometric data obtained, allow for the calculation of other variables such as vertical velocity, acceleration, and time, making it mathematically possible to derive essential parameters using widely recognized equations [22]. Among these parameter, power and maximum concentric power stand out, as they provide a robust overview of the lower-limb explosive force-generation capacity in this population [22].

Overall, the findings of the present study indicate that a video-based assessment protocol using Kinovea® for CMJ analysis can provide objective and reliable measurements of jump height. These results are consistent with previous findings regarding the use of Kinovea software for motion analysis [22].

The main characteristics of the population with Down Syndrome who participated in the study include manifestations affecting the musculoskeletal, neurological, and cardiovascular systems [8], as previously reported. These manifestations include hypotonia, ligamentous laxity, decreased muscle strength, insufficient co-contraction of the [9] muscles, deficits in postural control, short stature, and cognitive impairment [23], all of which directly and indirectly affect jumping performance.

Therefore, the data from this study demonstrated that a video-based assessment protocol for countermovement jump analysis can provide objective and reliable measures of jump height. These results are consistent with previous findings regarding the use of Kinovea software for motion analysis. [22], [24], [25].

VII. Limitations

The main challenges of the study are related to video analysis. In scenarios with low video resolution and no fixed markers, tracking and tracing the movement trajectory of the anatomical point — although semi-automatic — requires a more thorough frame-by-frame correction process. Inter-individual variability was also observed in jumping performance, with differences in the number of subjects who

were able to complete the required 10 repetitions. While the proposed method showed high reliability, a limitation of this study was the absence of direct validation against the gold standard (i.e., force platform).

Consequently, there is a significant increase in the operational time required for data analysis and processing; given that there are 10 jump cycles, the analysis time ranges from 5 to 7 minutes under ideal conditions to up 1 hour until for rigorous analysis.

During the analysis, it was noted that the trochanter's location corresponds to roughly half of the subject's height, which helps locate the structure if the marker is missing. Moreover, despite careful evaluation, the final graph for a few videos was generated inverted, suggesting a possible program bug that requires correction. Therefore, if the graph is inverted, all lines must be cleared and the calibration repeated to ensure it is plotted correctly.

For future studies, the use of well-positioned anatomical markers with high visual contrast is strongly recommended. This approach may reduce the need for manual frame-by-frame corrections, optimize processing time, and improve the accuracy of the extracted variables.

VIII. Conclusion

Based on the findings of the present study, the assessment of countermovement jump height using two-dimensional (2D) video recording and Kinovea® demonstrated good intra- and inter-examiner reliability. Therefore, this low-cost assessment protocol may represent a feasible and accessible approach for evaluating countermovement jump performance in individuals with Down syndrome in both clinical and research settings.

Acknowledgment

The present work was carried out with the support of the National Council for Scientific and Technological Development (444437/2024-0, 442150/2023-7, and 405365/2023-3), Coordination for the Improvement of Higher Education Personnel (CAPES), Grant no. 88887.212098/2025-00. Foundation for Research Support of the State of Minas Gerais (FAPEMIG APQ-05015-24). AOA (302942/2022-0) is a CNPq fellow.

References

[1] M. Adrian and J. M. Cooper, The Biomechanics of Human Movement. Indianapolis: Benchmark, 1989. [Online]. Available: https://books.google.com/books/about/The_Biomechanics_of_Human_Movement.html?hl=pt-BR&id=F1qJQgAACAAJ

[2] M. Beerse and J. Wu, "Vertical stiffness and balance control of two-legged hopping in-place in children with and without Down syndrome," Gait Posture, vol. 63, pp. 39–45, Jun. 2018, doi: 10.1016/J.GAITPOST.2018.04.026.

[3] D. L. Gallahue, J. C. Ozmun, and J. D. Goodway, Compreendendo o Desenvolvimento Motor, 7th ed. Porto Alegre: AMGH, 2013. [Online]. Available: https://integrada.minhabiblioteca.com.br/reader/books/9788580551815/

[4] F. T. Melo, K. S. Oliveira, A. M. de Amorim, C. B. de Almeida, and R. F. de F. Mussi, "Padrões motores fundamentais de crianças de uma escola baiana," Diálogos Perspectivas Interventivas, vol. 1, p. e10164, Dec. 2020, doi: 10.52579/diapi.vol1.i.a10164.

[5] Y. Zi, C. E. M. van Beijsterveldt, M. Bartels, and E. J. C. de Geus, "Genetic and environmental effects on the early motor development as a function of parental educational attainment," Med. Sci. Sports Exerc., vol. 55, no. 10, pp. 1845–1856, Oct. 2023, doi: 10.1249/MSS.0000000000003209.

[6] J.-L. Wang, S.-H. Sun, and H.-C. Lin, "Relationship of quantitative measures of jumping performance with gross motor development in typically developed preschool children," Int. J. Environ. Res. Public Health, vol. 19, no. 3, p. 1661, Jan. 2022, doi: 10.3390/ijerph19031661.

[7] S. W. Logan, S. M. Ross, K. Chee, D. F. Stodden, and L. E. Robinson, "Fundamental motor skills: A systematic review of terminology," J. Sports Sci., vol. 36, no. 7, pp. 781–796, Apr. 2018, doi: 10.1080/02640414.2017.1340660.

[8] S. E. Antonarakis et al., "Down syndrome," Nat. Rev. Dis. Primers, vol. 6, no. 1, p. 9, Feb. 2020, doi: 10.1038/s41572-019-0143-7.

[9] F. Quinzi, V. Camomilla, C. Bratta, M. F. Piacentini, P. Sbriccoli, and G. Vannozzi, "Hopping skill in individuals with Down syndrome: A qualitative and quantitative assessment," Hum. Mov. Sci., vol. 78, p. 102821, Aug. 2021, doi: 10.1016/J.HUMOV.2021.102821.

[10] F. Bardid, G. Vannozzi, S. W. Logan, L. L. Hardy, and L. M. Barnett, "A hitchhiker's guide to assessing young people's motor competence: Deciding what method to use," J. Sci. Med. Sport, vol. 22, no. 3, pp. 311–318, Mar. 2019, doi: 10.1016/j.jsams.2018.08.007.

[11] K. Mills, "Motion analysis in the clinic: There's an app for that," J. Physiother., vol. 61, no. 1, pp. 49–50, Jan. 2015, doi: 10.1016/j.jphys.2014.11.014.

[12] J. Charmant, Kinovea, version 2025.2.0, 2026. [Online]. Available: https://www.kinovea.org/

[13] S. M. D. Brucki, R. Nitrini, P. Caramelli, P. H. F. Bertolucci, and I. H. Okamoto, "Sugestões para o uso do mini-exame do estado mental no Brasil," Arq. Neuro-Psiquiatr., vol. 61, no. 3 B, pp. 777–781, 2003, doi: 10.1590/S0004-282X2003000500014.

[14] J. A. Martelli Filho, L. Á. Maltagliati, F. Trevisan, and C. T. L. de A. Gil, "Novo método estatístico para análise da reprodutibilidade," Rev. Dental Press Ortod. Ortop. Facial, vol. 10, no. 5, pp. 122–129, Oct. 2005, doi: 10.1590/S1415-54192005000500012.

[15] E. B. Isralowitz, J. Sideris, L. I. Stein Duker, G. T. Baranek, and S. A. Cermak, "Comparing sensory processing in children with Down syndrome to a mental age matched sample of children with autism, other developmental disabilities, and typically developing children," Res. Dev. Disabil., vol. 134, p. 104421, Mar. 2023, doi: 10.1016/J.RIDD.2022.104421.

[16] L. Petrigna et al., "A review of countermovement and squat jump testing methods in the context of public health examination in adolescence: Reliability and feasibility of current testing procedures," Front. Physiol., vol. 10, Nov. 2019, doi: 10.3389/fphys.2019.01384.

[17] Z. Ji and Y. Zhou, "Fundamental motor skills as an important competence for child development: A bibliometric analysis (1995–2025)," Sport Sci. Health, vol. 22, no. 2, p. 125, Jun. 2026, doi: 10.1007/s11332-026-01710-z.

[18] T. B. Macedo Alfano Moura and V. H. Alves Okazaki, "Kinematic and kinetic variable determinants on vertical jump performance: A review," MOJ Sports Med., vol. 5, no. 1, pp. 25–33, Jun. 2022, doi: 10.15406/mojsm.2022.05.00113.

[19] L. K. Philpott, S. E. Forrester, K. A. van Lopik, S. Hayward, P. P. Conway, and A. A. West, "Countermovement jump performance in elite male and female sprinters and high jumpers," Proc. Inst. Mech. Eng. P J. Sport. Eng. Technol., vol. 235, no. 2, pp. 131–138, Jun. 2021, doi: 10.1177/1754337120971436.

[20] A. Struzik, "Biomechanical characteristics of the countermovement jump," in Measuring Leg Stiffness

During Vertical Jumps. Cham: Springer International Publishing, 2019, pp. 9–18, doi: 10.1007/978-3-030-31794-2_2.

[21] J. Xu et al., "A systematic review of the different calculation methods for measuring jump height during the countermovement and drop jump tests," Sports Med., vol. 53, no. 5, pp. 1055–1072, May 2023, doi: 10.1007/s40279-023-01828-x.

[22] L. Faj et al., "From laboratory to field: Concurrent validity of Kinovea's linear kinematics tracking tool for semi-automated countermovement jump analysis," Sensors, vol. 26, no. 1, p. 24, Dec. 2025, doi: 10.3390/s26010024.

[23] P. D. Jain, A. Nayak, S. D. Karnad, and K. N. Doctor, "Gross motor dysfunction and balance impairments in children and adolescents with Down syndrome: A systematic review," Clin. Exp. Pediatr., vol. 65, no. 3, pp. 142–149, Mar. 2022, doi: 10.3345/cep.2021.00479.

[24] P. Fernández-González, A. Koutsou, A. Cuesta-Gómez, M. Carratalá-Tejada, J. C. Miangolarra-Page, and F. Molina-Rueda, "Reliability of Kinovea® software and agreement with a three-dimensional motion system for gait analysis in healthy subjects," Sensors, vol. 20, no. 11, p. 3154, Jun. 2020, doi: 10.3390/s20113154.

[25] N. Shishov, K. Elabd, V. Komisar, H. Chong, and S. N. Robinovitch, "Accuracy of Kinovea software in estimating body segment movements during falls captured on standard video: Effects of fall direction, camera perspective and video calibration technique," PLoS One, vol. 16, no. 10, p. e0258923, Oct. 2021, doi: 10.1371/journal.pone.0258923.

Spectral Analysis of Ground Reaction Force on Standard Vertical Jump with Different Stretch-Shortening Cycle

Carlos Rodrigues
C-BER - Centre for Biomedical Engineering Research
INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal
https://orcid.org/0000-0001-6408-6032

Miguel Correia
C-BER - Centre for Biomedical Engineering Research
INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal
https://orcid.org/0000-0001-6065-9358

João Abrantes
MovLab - Laboratory of Technologies for Interactions and Interfaces
CICANT - Centre for Research in Applied Communication, Culture, and New Technologies
Lisbon, Portugal
https://orcid.org/0000-0002-7628-0462

Marco Rodrigues
Department of Electronic and Systems
Federal University of Pernambuco
Recife, Brazil
https://orcid.org/0000-0002-5143-3696

Jurandir Nadal
Biomedical Engineering Program
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
https://orcid.org/0000-0003-3921-7787

***Abstract*— This study presents and applies spectral analysis of vertical ground reaction force (GRF_z) during impulse and landing phase from long, short and no countermovement (CM) with lower limb muscle stretch-shortening cycle (SSC). Spectral analysis of GRF_z presents as a complementary tool to the time domain, with the ability to detect hidden predictors of force production at different CM and lower limb muscle SSC using the subject body weight and FFT digital signal processing thus expanding previous studies on standing, walking and running spectral analysis while providing key data to MVJ optimization.**

Keywords—ground reaction force, FFT, maximum vertical jump, lower limb, muscle SSC

I. Introduction

Muscle stretch-shortening cycle (SSC) is a mechanism of muscle action, with muscle stretching immediately preceding muscle contraction, for efficient submaximal and powerful maximal activities [1]. Despite muscle SSC is determinant at lower limb on gait and running, its higher expression and assessment is performed at maximum vertical jump (MVJ) on countermovement jump (CMJ) with long countermovement (CM) and SSC, drop jump (DJ) with short CM and SSC, and squat jump (SJ) with no CM and SSC [2]. According to well defined objective of MVJ whole-body (WB) center of mass (COM) height (ht) during flight phase and standardized nature of different MVJ protocol, higher attention has been focused on the relation of the impulse from vertical ground reaction force (GRF_z) with MVJ WB ht to assess different long, short and no CM contribution to MVJ performance. Thus, the area below the force-time profile of GRF_z during the impulse phase has been assessed as a determinant of each MVJ and CM performance [3]. Nevertheless, GRF_z time profiles present, during impulse phases at each MVJ type, different shapes determining the area below the force-time profile and thus influencing different MVJ type performance. Time domain analysis has been performed based on impulse, power and work predictors with differences between predicted and achieved performances and an open issue on explainable differences on each MVJ type [4]. Thus, based on the complementarity of the frequency analysis of the human body vibration [5] we applied fast Fourier transformations (FFT) to GRF_z time profiles for spectral analysis and GRF_z shape comparison on impulse and landing phases searching for hints and relationships with different MVJ and CM performances, by analogy to gait and running frequency analysis [6 - 8].

II. Materials and Methods

A. Trial tests

A total of fifty-four trials were performed by a sample of six sports students with no injury or specific training, ages 21.5 ± 1.4 years old, 76.7 ± 9.3 kg mass and 1.79 ± 0.06 m height executing three repetitions of CMJ, DJ and SJ maximum vertical jumps (MVJ). The study was approved CAE200903 by the ISMAI scientific committee 01/03/2009, the subjects signed written informed consent agreement, and the ethical standards were followed on human experimentation according to the Helsinki Declaration of 1975, as revised in 2000 and 2008. After warm-up, subjects were instructed on MVJ protocol, with CMJ starting from orthostatic standing position initiating downward movement with long countermovement (CM) to 90º knee flexion and reverse to upward movement, DJ depth jump from 40 cm height step downward and short CM reverse to upward movement, with SJ initiating from squat with 90º knee flexion without CM. The subjects were given initial period for protocol familiarization, the tests were performed by each subject starting with CMJ, SJ and DJ repetitions by this order, with 2 to 3 minutes rest between trials and failed trials were rejected.

B. Signal acquisition

During trial tests ground reaction forces (GRF) were acquired with force platform AMTI BP2416-6-4000CE (Advanced Mechanical Technology Inc., USA) at 1000 Hz with coupled amplifier Mini Amp MSA-6. The force platform was previously calibrated, and the amplifier gain was adjusted to prevent saturation of vertical force component during MVJ trials. Time profiles of GRF were segmented based on vertical component (GRF_z) at the start of the impulse phase on CMJ ($GRF_z < Fg$), DJ ($GRF_z > 0$), and SJ ($GRF_z > Fg$), flight phase ($GRF_z = 0$) and landing phase ($GRF_z > 0$) at CMJ, DJ and SJ. For each CMJ, DJ and SJ trial, whole-body (WB) center of mass (COM) MVJ height (ht) was computed from flight time (t).

C. Signal processing

Fast Fourier transform (FFT) was computed for the vertical ground reaction force component (GRF_z) during impulse and landing phases at CMJ, DJ and SJ, using MATLAB R2010b (The MathWorks Inc., Natick, MA, USA), according to (1) and (2). Fourier coefficients for decomposition of GRF_z were obtained using (1) and GRF_z reconstruction below a defined harmonic performed using (2),

$$\begin{aligned} A_0 &= \frac{1}{T}\int_0^T GRF_z(t)\,dt \\ A_k &= \frac{2}{T}\int_0^T GRF_z(t)\cos(k\omega t)\,dt \\ B_k &= \frac{2}{T}\int_0^T GRF_z(t)\sin(k\omega t)\,dt \end{aligned} \quad (1)$$

$$\begin{aligned} C_0 = A_0 \quad C_k = \sqrt{A_k^2 + B_k^2} \quad \varphi_k = \tan^{-1}\left[-\frac{B_k}{A_k}\right] \\ GRF_z(t) = A_0 + \sum_{k=1}^{n_k}[A_k\cos(k\omega t) + B_k\sin(k\omega t)] \\ GRF_z(t) = C_0 + \sum_{k=1}^{n_k} C_k\cos(k\omega t + \varphi_k) \end{aligned} \quad (2)$$

with $\omega = 2\pi/T$ the fundamental frequency, and $k = 1, \ldots, n_k$ the order of the harmonic. Fourier coefficients from GRF_z were analyzed according to the frequency of the harmonic with higher amplitude and 90th percentile of the Fourier series for each MVJ type. The convergence of the Fourier series was assessed according to the normalization of the partial sum expressed by (3)

$$SC(n_k) = \sum_{k=0}^{n_k} C_k / \sum_{k=0}^{n_k^{\max}} C_k \quad (3)$$

D. Statistical analysis

Selected parameters were independently and paired compared on CMJ, DJ and SJ for each subject, using parametric T-Test with Statistical Package for Social Sciences IBM SPSS 29.0.0.0 (IBM, 2022).

III. Results

Flight time t and MVJ height ht presented on independent comparisons, higher values at CMJ than SJ ($p \geq 0.05$), both higher than DJ ($p < 0.05$), Tab. 1.

TABLE I. Mean and Standard Deviation of the Flight Time and Whole-Body Center of Mass Maximum Vertical Jump Height.

Mean ± Std	Flight phase		
	CMJ	DJ	SJ
***t* (s)**	0.550 ± 0.036	0.479 ± 0.024	0.524 ± 0.042
***ht* (s)**	0.372 ± 0.049	0.282 ± 0.027	0.338 ± 0.051

Based on paired comparisons at Tab. 2 and 3, the flight time t and MVJ height ht presented similar higher means at CMJ in relation to SJ and both higher than DJ, all with $p < 0.05$ pointing to the paired comparison higher sensitivity for detection of statistically significant differences.

TABLE II. Mean, Standard Deviation and Significance of the Flight Time Paired Diferences of Maximum Vertical Jumps.

***t* (s)**	Flight time		
	***μ* (s)**	***σ* (s)**	***α***
CMJ - SJ	0.026	0.012	$< 10^{-3}$
CMJ - DJ	0.071	0.034	$< 10^{-3}$
SJ - DJ	0.045	0.041	0.006

TABLE III. Mean, Standard Deviation and Significance of the Flight Height Paired Diferences of Maximum Vertical Jumps.

***ht* (s)**	Flight height		
	***μ* (s)**	***σ* (s)**	***α***
CMJ - SJ	0.034	0.015	$< 10^{-3}$
CMJ - DJ	0.090	0.045	$< 10^{-3}$
SJ - DJ	0.057	0.052	0.006

Fig. 1 presents illustrative time profiles of the vertical ground reaction forces GRF_z and reconstructed FFT below $k = 10$ during impulse phase of CMJ, DJ, and SJ performed trials.

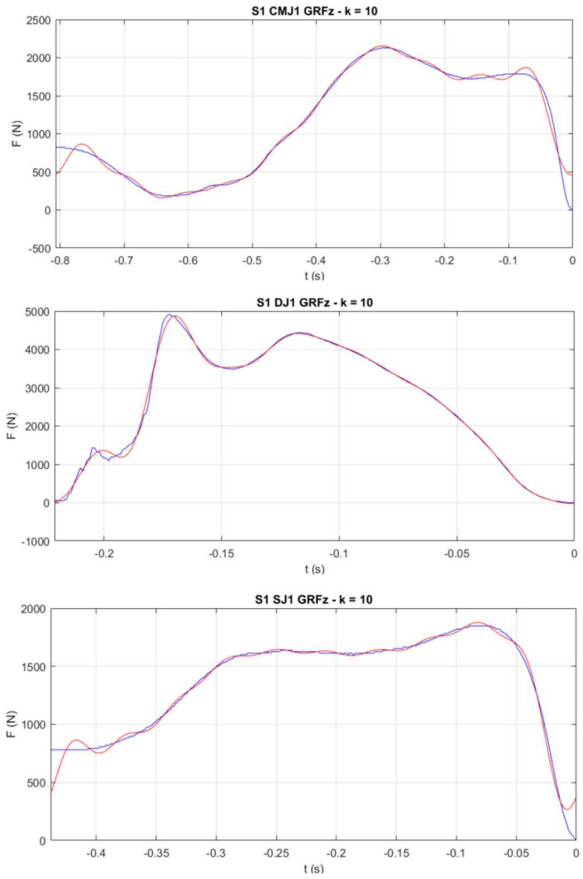


Fig. 1. Example of the vertical ground reaction force (GRF_z) (blue) and GRF_z reconstructed FFT (red) below $k = 10$ during impulse phase of CMJ, DJ, and SJ performed trials.

Fig. 2 presents corresponding time profile of the vertical ground reaction forces GRF_z and reconstructed FFT below $k = 20$ during landing phase at CMJ, DJ, and SJ performed trials.

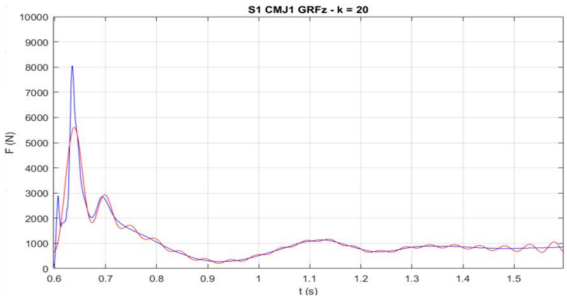


Fig. 2. Example of the vertical ground reaction force (GRF_z) (blue) and GRF_z reconstructed FFT (red) below $k = 20$ during landing phases of CMJ, DJ, and SJ performed trials.

A. *Impulse phase*

The maximum value $Max|FFT|$ resulting from Fourier decomposition (FFT) of GRF_z during impulse phase, Tab. 4, presents higher values in DJ relative to CMJ both higher than SJ ($p < 0.05$) with these maximum values being registered at higher frequencies $f(Max|FFT|)$ for DJ relative to SJ, both higher than CMJ ($p < 0.05$). Therefore, 90th percentile of FFT decomposition of GRF_z $f(p90|FFT|)$ occurs at lower frequencies for DJ relative to SJ and CMJ ($p < 0.05$) without statistically significant differences between SJ and CMJ ($p \geq 0.05$). The cumulative result of FFT decomposition of GRF_z up to the 25th harmonic $S|FFT|f25$ has higher values in DJ relative to CMJ, both higher than SJ ($p < 0.05$). The root mean squared deviation (RMSD) between the reconstructed signal up to the 25th harmonic and the GRF_z signal $RMSDf25$ presents lower values for DJ relative to SJ, both lower than CMJ ($p < 0.05$). The difference between the mean value, corresponding to the first harmonic of FFT decomposition and the original GRF_z signal $RMSDf0$ presents higher values in DJ relative to CMJ, both higher than SJ ($p < 0.05$). Therefore, when normalized to the mean value, the relative deviation between the 25th harmonic reconstruction and the GRF_z signal presents higher values in SJ relative to CMJ, both higher than DJ ($p < 0.05$).

Fig. 3 presents illustrative time profiles of the vertical ground reaction forces GRF_z and reconstructed FFT below the frequency $f(Max|FFT|)$ of the first harmonic $k = 1$ during the impulse phases of CMJ, DJ, and SJ performed trials.

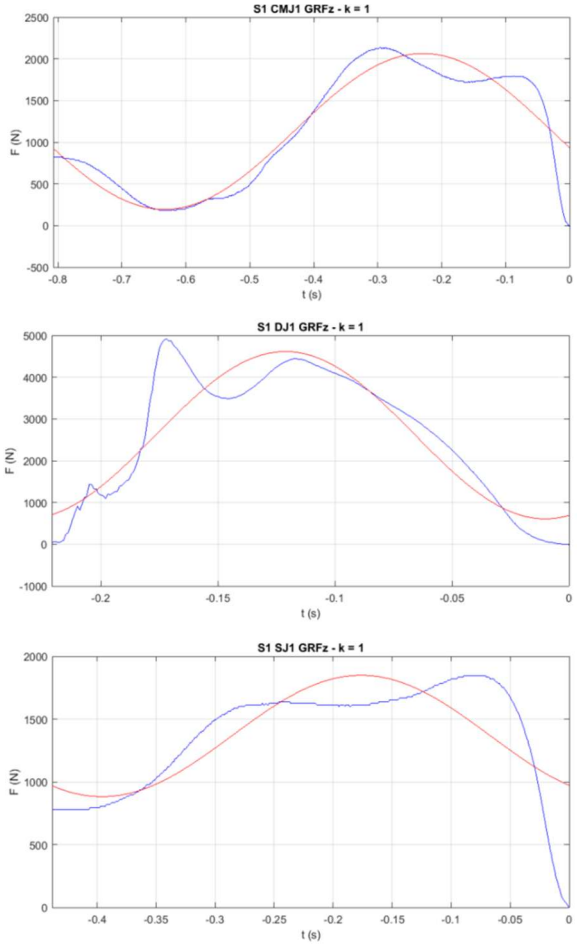


Fig. 3. Example of vertical ground reaction force (GRF_z) (blue) and GRF_z reconstructed FFT (red) below $k = 1$ during impulse phase of CMJ, DJ, and SJ performed trials.

TABLE IV. MEAN AND STANDARD DEVIATION OF THE MAXIMUM FFT, FREQUENCY OF THE MAXIMUM FFT, 90TH PERCENTILE, 25TH HARMONIC, RMSD OF 25TH HARMONIC, MEAN AND % AT CMJ, DJ, SJ IMPULSE PHASE.

Mean ± Std	Impulse phase		
	CMJ	DJ	SJ
$Max\|FFT\|$ (N)	825.22 ± 146.97	1927.88 ± 106.49	533.86 ± 74.06
$f(Max\|FFT\|)$ (Hz)	1.27 ± 0.18	4.93 ± 0.48	2.49 ± 0.40
$f(p90\|FFT\|)$ (Hz)	144.79 ± 13.9	35.05 ± 18.44	148.96 ± 7.74
$S\|FFT\|f25$	0.79 ± 0.02	0.97 ± 0.01	0.84 ± 0.01
$RMSDf25$ (N)	35.68 ± 3.68	23.00 ± 14.66	32.30 ± 3.10
$RMSDf0$ (N)	615.48 ± 103.83	1503.67 ± 109.82	448.94 ± 58.74
$RMSDf25$ %	5.94 ± 1.21	1.49 ± 0.90	7.29 ± 1.06

Paired comparison at Tab. 5 of the $Max|FFT|$ from FFT decomposition of GRF_z during impulse phase, with higher values at DJ than CMJ both higher than SJ, are in agreement with the corresponding metrics on time domain, and higher intensities of GRF_z values during DJ in relation to CMJ and SJ. Additionally, the results from the paired comparison of the harmonics frequency $f(Max|FFT|)$ on Tab. 6 for which the $Max|FFT|$ of GRF_z are registered, with higher values at DJ than SJ both higher than CMJ are in line with the short, long and no CM at DJ, CMJ and SJ, and the corresponding period of the impulse phases relation with the frequency of the fundamental harmonic at each MVJ type.

TABLE V. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE GRFz $Max|FFT|$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$Max\|FFT\|$ (N)	Maximum GRFz FFT		
	μ(s)	σ(s)	p
CMJ - SJ	292.1	151.6	$< 10^{-3}$
CMJ - DJ	-1102.7	193.2	$< 10^{-3}$
SJ - DJ	-1394.8	102.9	$< 10^{-3}$

TABLE VI. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $f(Max|FFT|)$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$f(Max\|FFT\|)$ (Hz)	Frequency of maximum GRFz FFT		
	μ(Hz)	σ(Hz)	p
CMJ - SJ	-1.26	0.25	$< 10^{-3}$
CMJ - DJ	-3.66	0.53	$< 10^{-3}$
SJ - DJ	-2.39	0.69	$< 10^{-3}$

The results from the paired comparisons for the 90th percentile of the FFT decomposition GRF_z $f(p90|FFT|)$ on Tab. 7 corresponds to the concentration at lower frequencies for the larger signal component at DJ in relation to CMJ an SJ.

TABLE VII. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $f(p90|FFT|)$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$f(p90\|FFT\|)$ (Hz)	Frequency of 9th percentile GRFz FFT		
	μ(Hz)	σ(Hz)	p
CMJ - SJ	-4.6	14.2	0.179
CMJ - DJ	107.7	27.3	$< 10^{-3}$
SJ - DJ	114.4	21.2	$< 10^{-3}$

Thus, DJ presents for the paired comparisons of the 25th harmonic $S|FFT|f25$ higher values in DJ than CMJ, both higher than SJ, supporting the larger concentration of GRF_z at lower frequencies on short CM than at long CM and no CM. Tab. 8. For this reason, the RMSD between the reconstructed signal up to the 25th harmonic and the GRF_z signal $RMSDf25$ presented for the paired comparison on Tab. 9 lower values on DJ in relation to SJ, and higher values on CMJ.

TABLE VIII. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $S|FFT|f25$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$S\|FFT\|f25$	Accumulated GRFz FFT up to the 25th harmonic		
	μ (Hz)	σ (Hz)	p
CMJ - SJ	-0.048	0.018	$< 10^{-3}$
CMJ - DJ	-0.176	0.029	$< 10^{-3}$
SJ - DJ	-0.128	0.017	$< 10^{-3}$

TABLE IX. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $RMSDf25$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$RMSDf25$ (N)	Root mean square error of GRFz FFT up to $k = 25$		
	μ	σ	p
CMJ - SJ	3.59	2.32	$< 10^{-3}$
CMJ - DJ	12.68	14.06	0.013
SJ - DJ	9.10	13.15	0.036

Regarding the paired comparisons of the differences between the mean values, corresponding to the first harmonic of FFT decomposition and the original GRF_z signal $RMSDf0$ on Tab 10, DJ presented higher values than CMJ and both higher than SJ, in line with the maximum values $Max|FFT|$ resulting from Fourier decomposition (FFT) of GRF_z. Thus, the relative deviation $RMSDf25\,\%$ between the 25th harmonic reconstruction and the GRF_z signal normalized to $RMSDf0$ presented on paired comparisons at Tab. 11, higher value on SJ than CMJ, both higher than DJ.

TABLE X. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $RMSDf0$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$RMSDf0$ (N)	Root mean square error of GRFz FFT for $k = 0$		
	μ	σ	p
CMJ - SJ	164.1	120.9	0.002
CMJ - DJ	-888.2	115.7	$< 10^{-3}$
SJ - DJ	-1052.3	80.7	$< 10^{-3}$

TABLE XI. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $RMSDf25\,\%$ PAIRED DIFERENCES AT CMJ, DJ, SJ IMPULSE PHASES.

$RMSDf25\,\%$	Root mean square error of GRFz FFT up to $k = 25$		
	μ	σ	p
CMJ - SJ	-0.013	0.017	0.030
CMJ - DJ	0.044	0.017	$< 10^{-3}$
SJ - DJ	0.057	0.011	$< 10^{-3}$

Fig. 4 present illustrative examples of the vertical ground reaction forces GRF_z FFT and relative accumulated FFT during the impulse phases of CMJ, DJ, and SJ performed trials with Fig. 5 corresponding landing phase on CMJ trial.

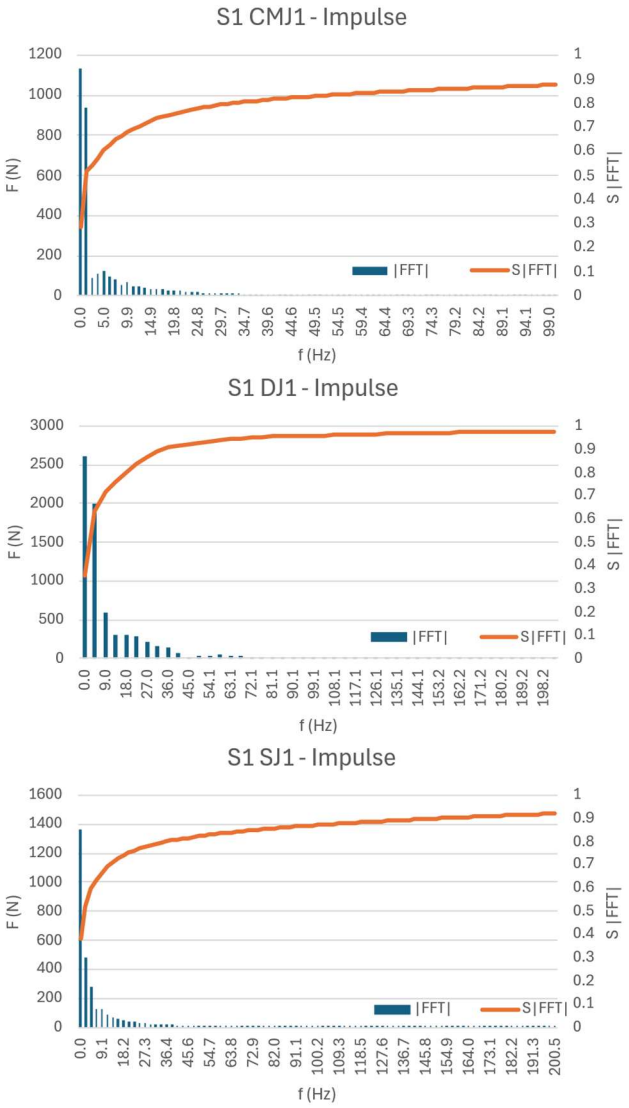


Fig. 4. Example of vertical ground reaction force (GRF_z) FFT and relative accumulated FFT during impulse phases of CMJ, DJ, SJ performed trials.

B. *Landing phase*

For the landing phases, the maximum value $Max|FFT|$ resulting from Fourier decomposition (FFT) of GRF_z presents higher values for CMJ relative to SJ, both higher than DJ ($p \geq 0.05$), Tab. 12, with the maximum values registered at similar frequencies $f(Max|FFT|)$, slightly higher for DJ relative to SJ, both higher than CMJ ($p \geq 0.05$). The 90th percentile of FFT decomposition of GRF_z $f(p90|FFT|)$ occurs at lower frequencies for DJ relative to CMJ, both lower than SJ ($p \geq 0.05$). The cumulative result of FFT decomposition of GRF_z up to the 25th harmonic $S|FFT|f25$ has higher values in DJ relative to CMJ and SJ ($p \geq 0.05$). The RMSD between the reconstructed signal up to the 25th harmonic and the GRF_z signal $RMSDf25$ presents lower values for DJ relative to SJ and CMJ ($p \geq 0.05$). The difference between the mean value corresponding to the first harmonic of FFT decomposition and the original GRF_z signal $RMSDf0$ presents higher values in CMJ relative to SJ, both higher than DJ ($p \geq 0.05$). Therefore, when normalized to the mean value, the relative deviation between the 25th harmonic reconstruction and the GRF_z signal $RMSDf25\,\%$ presents higher values in SJ and CMJ relative to DJ ($p \geq 0.05$).

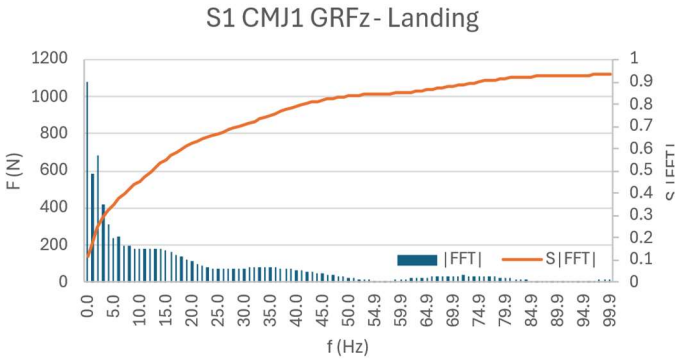


Fig. 5. Example of vertical ground reaction force (GRF_z) FFT and relative accumulated FFT during landing phase of CMJ performed trial.

TABLE XII. MEAN AND STANDARD DEVIATION OF THE MAXIMUM FFT, FREQUENCY OF THE MAXIMUM FFT, 90TH PERCENTILE, 25TH HARMONIC, RMSD OF 25TH HARMONIC, MEAN AND % AT CMJ, DJ, SJ LANDING PHASE.

Mean ± Std	**Landing phase**		
	CMJ	**DJ**	**SJ**
$Max\|FFT\|$ (N)	579.04 ± 171.74	487.36 ± 99.14	507.57 ± 69.97
$f(Max\|FFT\|)$ (Hz)	1.56 ± 0.53	1.67 ± 0.71	1.60 ± 0.52
$f(p90\|FFT\|)$ (Hz)	67.93 ± 10.35	60.16 ± 20.59	75.63 ± 15.37
$S\|FFT\|f25$	0.73 ± 0.07	0.77 ± 0.06	0.72 ± 0.07
$RMSDf25$ (N)	179.70 ± 93.62	125.70 ± 45.08	179.62 ± 108.44
$RMSDf0$ (N)	782.36 ± 164.98	738.83 ± 129.20	754.65 ± 182.44
$RMSDf25$ %	21.73 ± 9.08	17.30 ± 6.45	22.39 ± 10.16

The paired comparisons of the FFT parameters from GRF_z during the landing phases on Tab. 13 - 19 confirm the general trend detected with the independent comparisons on Tab. 12 for the absence of statistically significant differences among MVJ types. Nevertheless, the paired comparisons presented higher sensitivity for detection of statistically significant differences at $p < 0.05$ in association with maximum vertical jump height ht.

Thus, based on paired comparison of the FFT parameters from GRF_z during the landing phases, statistically significant differences were exclusively detected for the $Max|FFT|$ on Tab. 13 with higher value on CMJ than DJ in association with higher ht, the $f(p90|FFT|)$ on Tab. 15 with lower value at DJ than SJ and the $S|FFT|f25$ on Tab. 16 with higher value at DJ than CMJ due to higher FFT convergence associated with GRF_z time profile.

TABLE XIII. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE GRFZ $Max|FFT|$ PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$Max\|FFT\|$ **(N)**	**Maximum GRFz FFT**		
	μ**(s)**	σ**(s)**	p
CMJ - SJ	70.8	131.5	0.072
CMJ - DJ	91.7	83.0	0.005
SJ - DJ	20.9	78.3	0.223

TABLE XIV. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $f(Max|FFT|)$ PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$f(Max\|FFT\|)$ **(Hz)**	**Frequency of maximum GRFz FFT**		
	μ**(Hz)**	σ**(Hz)**	p
CMJ - SJ	0.00	0.50	0.500
CMJ - DJ	-0.11	0.93	0.364
SJ - DJ	-0.11	0.78	0.341

TABLE XV. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $f(p90|FFT|)$ PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$f(p90\|FFT\|)$ **(Hz)**	**Frequency of 9th percentile GRFz FFT**		
	μ**(Hz)**	σ**(Hz)**	p
CMJ - SJ	-9.11	15.1	0.054
CMJ - DJ	7.74	21.7	0.158
SJ - DJ	16.86	23.6	0.032

TABLE XVI. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $S|FFT|f25$ PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$S\|FFT\|f25$	**Accumulated GRFz FFT up to the 25th harmonic**		
	μ**(Hz)**	σ**(Hz)**	p
CMJ - SJ	-0.003	0.068	0.446
CMJ - DJ	-0.043	0.047	0.012
SJ - DJ	-0.040	0.069	0.061

TABLE XVII. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $RMSDf25$ PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$RMSDf25$ **(N)**	**Root mean square error of GRFz FFT up to $k = 25$**		
	μ	σ	p
CMJ - SJ	17.6	93.6	0.589
CMJ - DJ	54.0	73.8	0.059
SJ - DJ	36.4	92.1	0.269

TABLE XVIII. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $RMSDf0$ PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$RMSDf0$ **(N)**	**Root mean square error of GRFz FFT for $k = 0$**		
	μ	σ	p
CMJ - SJ	47.4	103.8	0.104
CMJ - DJ	43.5	161.6	0.221
SJ - DJ	-3.9	188.3	0.476

TABLE XIX. MEAN, STANDARD DEVIATION AND SIGNIFICANCE OF THE $RMSDf25$ % PAIRED DIFERENCES AT CMJ, DJ, SJ LANDING PHASES.

$RMSDf25\%$	**Root mean square error of GRFz FFT up to $k = 25$**		
	μ	σ	p
CMJ - SJ	0.009	0.105	0.814
CMJ - DJ	0.044	0.070	0.095
SJ - DJ	0.057	0.011	0.267

C. Correlations with MVJ height

Selected metrics from FFT decomposition of GRF_z presented during the impulse phases higher correlations, Tab. 20, between the vertical jump height ht with negative associations to $Max|FFT|$, $f(Max|FFT|)$, $S|FFT|f25$, $RMSDf0$ whereas positive associations were detected with $(p90|FFT|)$, $RMSDf25$, and $RMSDf25\,\%$ due to the clustering at DJ with lower ht, higher $Max|FFT|$, $f(Max|FFT|)$, $S|FFT|f25$ and lower $f(p90|FFT|)$, $RMSDf25$, $RMSDf0$, $RMSDf25\,\%$ and the opposite at CMJ and SJ. As regards to the landing phases, the dominant associations correspond to the positive correlations of ht with $Max|FFT|$, $f(p90|FFT|)$, $RMSDf25$, $RMSDf0$ and $RMSDf25\,\%$, negative correlation with $S|FFT|f25$, very low correlation with $f(Max|FFT|)$ and mixed variable association at different MVJ type.

TABLE XX. CORRELATRIONS OF THE MAXIMUM VERTICAL HEIGHT WITH MAXIMUM FFT, FREQUENCY OF THE MAXIMUM FFT, 90TH PERCENTILE, 25TH HARMONIC, RMSD OF 25TH HARMONIC, MEAN AND % AT CMJ, DJ, SJ IMPULSE AND LANDING PHASES

Variable	Correlations r with ht (m)			
	Impulse r, p		Landing r, p	
$Max\|FFT\|$ (N)	-0.528*	0.004	0.629*	$< 10^{-3}$
$f(Max\|FFT\|)$ (Hz)	-0.681*	$< 10^{-3}$	0.180	0.359
$f(p90\|FFT\|)$ (Hz)	0.581*	0.001	0.383*	0.044
$S\|FFT\|f25$	-0.654*	$< 10^{-3}$	-0.638*	$< 10^{-3}$
$RMSDf25$ (N)	0.548*	0.003	0.637*	$< 10^{-3}$
$RMSDf0$ (N)	-0.519*	0.005	0.554*	0.002
$RMSDf25\,\%$	0.556*	0.002	0.528*	0.004

IV. DISCUSSION

Whereas the selected metric of GRF_z FFT presented clear distinction at long, short and no CM on different MVJ impulse phases ($p < 0.05$), the corresponding metrics on landing did not present dominant statistically significant differences between CMJ, DJ and SJ ($p \geq 0.05$) pointing to clear differences of the GRF_z distribution in terms of FFT content during impulse phase at different CM, and similar GRF_z distributions assessed through FFT content during the landing phases, regardless of different CM and MVJ type.

A. Impulse phase

Thus, whereas the higher maximum value of $Max|FFT|$ at DJ in relation to CMJ and SJ is associated with the highest level of peak GRF_z during the impulse phase at DJ regarding CMJ and SJ, the fact that these maximum values are achieved at higher frequencies $f(Max|FFT|)$ for DJ relative to SJ, both higher than CMJ can be associated with the shortest duration of the impulse phase at DJ in relation to SJ, and both lower than CMJ.

The fact that the 90th percentile of FFT decomposition of GRF_z $f(p90|FFT|)$ occurs at lower frequencies for DJ in relation to SJ and CMJ points to the concentration at a lower number of harmonics for the same level of percentage, 90% of the signal content on DJ relatively to SJ and CMJ, and this being corroborated with the highest level of cumulative result from the FFT decomposition of GRF_z up to the 25th harmonic $S|FFT|f25$ at DJ in relation to CMJ, both higher than SJ.

This can be additionally confirmed by the lowest values at DJ in relation to SJ and CMJ of the root mean squared deviation (RMSD) between the reconstructed signal up to the 25th harmonic and the GRF_z sinal $RMSDf25$, pointing to the higher proximity of the reconstructed GRF_z with the initial 25 harmonics to the original GRF_z signal at DJ than SJ and CMJ, with higher concentration of the signal energy at lowest harmonics, intermediate dispersion at the signal content on SJ and the highest dispersion on CMJ.

Regarding the highest difference between the first harmonic $RMSDf0$ of FFT decomposition and the original GRF_z signal at DJ in relation to CMJ, both higher than SJ, this fact can be associated to a larger magnitude, with higher peak values of GRF_z signal in relation to its mean value at DJ, the next higher levels of GRF_z peak values at CMJ and the lowest peak values at SJ.

For this reason, despite the highest $RMSDf25$ at CMJ in relation to SJ and DJ, the normalization to $RMSDf0$ conduced to the relative deviation between the 25th harmonic reconstruction and the GRF_z signal $RMSDf25\%$ with higher value at SJ in relation to CMJ, both higher than DJ.

B. Landing phase

Despite the differences in the mean values and variabilities assessed through the standard deviations of the various selected metrics at FFT domain, the dominant absence of statistically significant differences ($p \geq 0.05$) of these metrics during the landing phases between CMJ, DJ and SJ clearly point that there were detected no shape distinction of GRF_z assessed through FFT domain during the landing, regardless of the CM type and MVJ flight height.

C. Correlations with MVJ height

Additionally, the groups of selected metrics for GRF_z FFT domain correlation to MVJ height presented clear distinction between the impulse and the landing phases. Thus, whereas the correlations at the impulse phases led to clear distinction among different MVJ with clear clustering for CMJ, DJ and SJ, the corresponding correlations at the landing phases conduced to indistinguishable clusters for CMJ, DJ, and SJ pointing to the selected FFT parameters as good predictors of ht both on impulse and the landing phases, with the ability for discrimination between MVJ type restricted to the impulse phase, preserving at the landing phases the association with ht performance independently of the MVJ and CM type.

V. CONCLUSION

According to the achieved results, spectral analysis of GRF_z presents as a complementary tool to the time domain analysis, focusing on the shape of GRF_z in terms of its FFT harmonics distribution and amplitudes. Thus, whereas FFT selected metrics led to clear distinction at the frequency domain of GRF_z during impulse phase of MVJ long, short and no CM, invariant remarkable behavior was detected for these FFT metrics during landing phase of different long, short and no CM from various MVJ, while preserving the ability for prediction of the explosive performance assessed through maximum vertical jump height.

The FFT allowed for the comparison of time-series spectra that would otherwise not be directly comparable due to differing curve shapes, such as those observed during the impulsion phases of the various maximum vertical jumps, and

especially during landing phases, where temporal profiles are incomparable on a point-by-point basis.

Based on this study involving the spectral analysis of the maximum vertical jumps, it is possible to optimize each subject's individual performance for each type of maximum vertical jump, using the detected frequencies in relation to the body segments' natural vibration frequencies estimated in previous studies.

Specifically, the performed spectral analysis also allows for the objective differentiation—based on detected frequencies associated with vertical force curve shapes during the impulsion phases—of tendencies toward a long countermovement in the CMJ, clearly distinguishing it from short countermovement executions and the non-countermovement actions seen in the DJ and SJ.

To the best of our knowledge, this is a pioneering study applying spectral analysis to various vertical counter-movements, expanding and deepening the knowledge gained from previous studies regarding body segment spectral behaviour during standing, walking, and running.

REFERENCES

[1] P. V. Komi, M. Ishikawa, and V. Linnamo, "Identification of stretch-shortening cycles in different sports," Portuguese Journal of Sport Sciences, vol. 11, suppl. 2, Vilas-Boas, Machado, Kim, Veloso, Eds., 2011, pp. 31–34.

[2] P. V. Komi, Strength and power in sport. Oxford, UK: Blackwell Science Ltd, 2003.

[3] C. Rodrigues, M. Correia, J. Abrantes, J. Nadal, and M. Benedetti, "Kinematic and dynamic predictors of maximum vertical jump performance on elite and non-elite performers," in *Book of Abstracts of the 25th Annual Congress of the European College of Sport Science*, Cologne, Germany (virtual), SporTools Gmbh, October 28-30 2020, pp. 143–144.

[4] C. Rodrigues, M. Correia, J. Abrantes, M. Rodrigues, and J. Nadal, "Impulse, Power and Work predictors comparison on standard maximum vertical jump of elite and non-elite performers," in *Book of Abstracts of the 43rd Annual Congress of the Iberian Society of Biomechanics and Biomaterials (SIBB)*, Porto, Portugal (virtual), September 5-9 2021, pp. 982–984.

[5] G. Ramussen, "Human Body Vibration Exposure and its Measurement," Technical Review, vol. 1, Marlborough, MA: Brüel & Kjaer Instruments Inc., 1982, pp. 3–31.

[6] R. M. Alexander, Optima for animals. New Jersey, NJ: Princeton University Press, 1982, pp. 50–53.

[7] C. Rodrigues, M. Correia, J. Abrantes, J. Nadal, and M. Benedetti, "Power spectrum analysis of contact forces and force moments during normal and modified gait," in *Book of Abstracts of the 5th Joint International Conference on Multibody System Dynamics*, Lisbon, Portugal, June 24-28 2018.

[8] S. D. Skejø, M. E. Lund, M. Stensvig, N. M. Kaae, and J. Rasmussen, "Running in circles: Describing running kinematics using Fourier series," Journal of Biomechanics, vol. 115, pp. 529–551, January 2022.

A 3D-Printing Workflow for Biomedical Educational Models: A DNA Double-Helix Proof of Concept

Vitor Hugo Felipe Noronha
Software Engineering
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
vitor.noronha@ges.inatel.br

Maysa Costa Alves
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

Elisa Rennó Carneiro Déster
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
elisa.renno@inatel.br

***Abstract*—Three-dimensional understanding of biological structures is one of the central challenges in biomedical education: students are routinely introduced to molecular, microbiological, and cellular concepts that cannot be directly observed, and conventional two-dimensional illustrations often struggle to convey their spatial organization.**

Additive manufacturing offers a promising way to address this gap, producing tangible models that students can manipulate and explore from multiple angles, with the added potential to make biological structures accessible through touch. This paper proposes a digital-to-physical workflow for producing this kind of biomedical educational model — from biological data acquisition through digital preparation, slicing, fabrication, and assembly — and illustrates it with a first proof of concept: an open-source DNA double-helix model, adapted and manufactured using fused deposition modeling (FDM). Masked stereolithography (MSLA), planned for future higher-resolution models, was not yet available during this stage, so the comparison between the two technologies discussed here draws on the literature rather than on our own measurements.

The resulting prototype shows that the workflow can be carried out end to end, from a digital reference file to an assembled, handheld object, and surfaces practical manufacturing considerations that will guide the next iterations. This is presented as a workflow proposal illustrated by a single case rather than a fully validated methodology: learning outcomes, accessibility benefits, durability, and a direct FDM-versus-MSLA comparison are left for future work.

***Index Terms*—additive manufacturing, biomedical education, fdm, msla, three-dimensional models, tactile learning, inclusive education**

I. Introduction

Biomedical engineering integrates concepts from biology, chemistry, physics, medicine, and engineering to develop technologies that improve healthcare. Because of its interdisciplinary nature, students must understand biological structures ranging from organs to molecular assemblies. Many of these structures exhibit complex three-dimensional geometries that are difficult to interpret when presented only through conventional two-dimensional educational resources.

Traditional teaching materials rely heavily on textbook illustrations, microscopy images, slides, and computer-generated renderings. Although these resources provide valuable information, they require students to mentally reconstruct three-dimensional geometries from flat representations, increasing the cognitive effort required to understand spatial organization. This limitation becomes particularly evident when studying structures such as DNA molecules, proteins, viruses, and cellular organelles, whose biological function is closely associated with their three-dimensional arrangement [1], [2].

Scientific education also presents accessibility challenges, as most biological concepts are communicated through visual resources. For students with visual impairments, tactile graphics and verbal explanations are often insufficient to represent complex spatial geometries. Three-dimensional printed models have been reported to overcome part of this limitation by allowing the exploration of shape, symmetry, and structural organization through touch [3]; the present work adopts this as a design motivation rather than as a claim evaluated in this study.

Recent advances in additive manufacturing have enabled educational institutions to produce customized teaching resources at relatively low cost. Compared with commercial educational kits, 3D printing provides greater flexibility by allowing educators to adapt dimensions, geometry, and modularity according to specific pedagogical objectives. Previous studies have reported improvements in spatial reasoning, classroom engagement, and active learning through the use of three-dimensional printed educational models [4].

Despite these advances, most reported applications focus on macroscopic structures such as bones, organs, and anatomical models. Comparatively fewer studies investigate molecular and microbiological educational models, whose development requires a structured workflow involving scientific validation, digital modeling, mesh preparation, and additive manufacturing before physical fabrication can be achieved [4].

This work proposes a reproducible workflow for developing biomedical educational models using additive manufacturing and illustrates it through a single case study. Rather than evaluating multiple printed artifacts, the paper documents one complete digital-to-physical process, including biological

data acquisition, digital preparation, slicing, fabrication, and technical assembly. As a proof of concept, an open-source DNA double-helix model was adapted and manufactured using fused deposition modeling (FDM). Since the laboratory resin-printing infrastructure was still being implemented during the experimental phase, FDM was used to exercise the workflow ahead of a future migration to masked stereolithography (MSLA).

The main contributions of this work are:

1) a proposed, reproducible workflow for transforming biological information into physical educational models through additive manufacturing;
2) a single proof-of-concept demonstration of this workflow, using a DNA double-helix prototype manufactured by FDM;
3) a manufacturing strategy that uses FDM to exercise the workflow end to end while identifying MSLA as the intended technology for future higher-resolution production;
4) a discussion of accessibility and tactile-learning considerations as design objectives to guide future iterations of biomedical educational models, rather than as outcomes evaluated in this work.

II. Related Work

A. Additive Manufacturing in Biomedical Education

The incorporation of additive manufacturing into education has expanded considerably over the last decade due to reductions in equipment cost, improvements in printing quality, and the growing availability of open-source digital models. Desktop 3D printers have become valuable educational tools by enabling students to transform theoretical concepts into tangible objects while integrating digital modeling, manufacturing, and engineering principles [4].

Within biomedical education, additive manufacturing has been widely adopted to produce anatomical replicas for teaching, surgical planning, and simulation. Physical models are reported to improve students' spatial understanding by allowing direct manipulation of structures that would otherwise be observed only through images or computer renderings. However, successful implementation depends on a well-defined workflow involving digital modeling, mesh preparation, slicing, and manufacturing, aspects that are often only briefly addressed in the literature [4].

B. Molecular Educational Models and Accessibility

Compared with anatomical applications, relatively few studies investigate additive manufacturing for molecular and microbiological education. Models representing DNA, proteins, viruses, and other microscopic biological structures require additional processing to convert scientific data into printable geometries while preserving their educational relevance.

Three-dimensional printed models have also been reported to contribute to inclusive education by transforming visual scientific information into tactile educational resources [3]. Recent studies have demonstrated the feasibility of integrating computational modeling and additive manufacturing into molecular biology education, reinforcing the potential of customizable three-dimensional educational resources in higher education [5].

Motivated by these advances, the present work proposes a standardized workflow that integrates scientific validation, digital modeling, additive manufacturing, and educational evaluation into a single reproducible methodology intended to support the future development of biomedical educational models.

III. Methodology

This research proposes a reproducible workflow for developing biomedical educational models through additive manufacturing. Rather than focusing on a single printed object as an end in itself, the methodology standardizes the digital-to-physical process so it can, in principle, be applied to different biological structures with minimal adaptation; this generality is a design intention and has so far been exercised on a single structure (Section V).

The workflow consists of six sequential stages:

1) Selection of the biological structure;
2) Scientific validation and data acquisition;
3) Digital modeling and mesh preparation;
4) Additive manufacturing;
5) Post-processing;
6) Technical and educational evaluation.

Table I summarizes the primary objective of each stage.

A. Biological Structure Selection

Candidate structures are selected according to three criteria:

- relevance to biomedical engineering education;
- three-dimensional complexity;
- feasibility of additive manufacturing.

This methodology is intended for structures such as DNA, proteins, viruses, bacteriophages, cellular organelles, and molecular assemblies.

B. Scientific Validation and Digital Preparation

Biological information is obtained from peer-reviewed literature and specialized databases. For molecular structures, three-dimensional data may be acquired from the Protein Data Bank (PDB) [5] and inspected using molecular visualization software such as UCSF ChimeraX [6]. For the DNA model reported in this paper, scientific accuracy was reviewed by Prof. Raphael Cardoso Mota Pereira, who also advised on the fabrication adjustments described in Section V-A that ultimately resolved the model's printing issues.

The models are then prepared for manufacturing through mesh repair, scaling, wall-thickness adjustment, removal of geometric artifacts, reinforcement of fragile regions, and export to STL format for slicing and printing.

C. Manufacturing Strategy

The proposed methodology combines two complementary additive manufacturing technologies.

The workflow was exercised using fused deposition modeling (FDM), since the laboratory MSLA infrastructure was still under acquisition during the experimental phase. FDM allowed the digital preparation, slicing configuration, fabrication, assembly, and handling stages to be carried out end to end without delaying the project.

Future production is planned to employ masked stereolithography (MSLA), whose higher resolution is better suited to educational models with fine biological detail, such as viruses, proteins, and molecular complexes. FDM was used to exercise the workflow in the meantime; no direct experimental comparison between the two technologies was performed in this work.

D. Post-processing and Evaluation

After fabrication, prototypes undergo support removal, cleaning, assembly when necessary, and visual inspection. The validation process considers:

- dimensional consistency;
- structural stability;
- manufacturing quality;
- ease of handling;
- educational interpretation;
- potential for tactile accessibility.

In the present work, this evaluation was carried out qualitatively (visual and manual inspection); it did not include quantitative metrology, mechanical testing, or assessment by students, instructors, or visually impaired users. The evaluation results are intended to be incorporated into subsequent design iterations.

TABLE I
WORKFLOW ADOPTED FOR BIOMEDICAL EDUCATIONAL MODEL DEVELOPMENT

Stage	**Objective**
Selection	Choice of a biologically relevant structure.
Scientific validation	Collection and verification of biological data.
Digital preparation	Mesh optimization and STL generation.
Manufacturing	Prototype fabrication using additive manufacturing.
Post-processing	Cleaning, assembly and inspection.
Evaluation	Technical and educational validation.

IV. MATERIALS, INFRASTRUCTURE, AND SCALABILITY

The proposed workflow was developed and exercised at FabLab Inatel, which provides the computational and manufacturing resources required for rapid prototyping. The laboratory supports digital modeling, mesh preparation, slicing, fabrication, prototype assembly, and iterative evaluation.

Fig. 1 presents the laboratory used during the experimental activities.



Fig. 1. FabLab Inatel, where the digital modeling, fabrication, and evaluation stages of the proposed workflow were carried out.

A. Manufacturing Technologies

The methodology intentionally combines FDM and MSLA technologies. FDM provides a fast, economical, and accessible platform for iterative development, while the literature reports MSLA as offering higher dimensional accuracy and finer surface quality, better suited to structures with intricate biological detail [4], [7]. Fig. 2 illustrates the general differences between both technologies, as reported in the cited source. This project uses FDM to exercise and document the workflow first, with MSLA reserved for future high-resolution production; the comparison itself has not been tested experimentally by the authors.

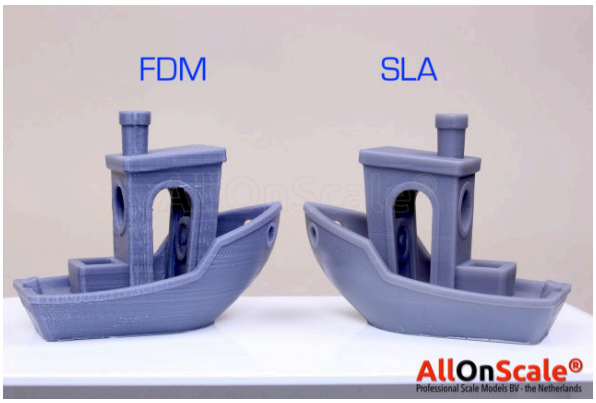


Fig. 2. General comparison between FDM and stereolithography-based resin printing technologies, as reported in the literature. Adapted from AllOnScale [7].

B. Available and Selected Equipment

FabLab Inatel provides three FDM printers, used across the different components of this project: a Creality K1 Max (build volume 300 × 300 × 300 mm), a Bambu Lab H2C (single-nozzle build volume ≈325 × 320 × 320 mm), and an UltiMaker 2+ Connect (build volume 223 × 220 × 205 mm). All three ship with a 0.4 mm nozzle as standard and support PLA, among other materials.

Future high-resolution production is planned to employ an Anycubic Photon Mono M7 Pro MSLA printer, already acquired by the research group. The equipment was not yet available during fabrication of the proof-of-concept prototype

(Section V-B); its incorporation represents the next stage of the proposed workflow. Fig. 3 shows the selected MSLA printer.



Fig. 3. Anycubic Photon Mono M7 Pro selected for future MSLA production of biomedical educational models.

C. Scalability

Once a digital model is prepared, it can in principle be reproduced and adapted without repeating the modeling process, supporting the gradual creation of an institutional repository of biomedical educational models. This is a design goal the current single-model proof of concept does not yet demonstrate, but is built to support.

V. Proof of Concept: DNA Double-Helix Model

To exercise the proposed workflow, a DNA double-helix model was selected as the first educational prototype. DNA is one of the central topics in biomedical engineering curricula, being addressed in molecular biology, genetics, biochemistry, biotechnology, and bioinformatics courses. Despite its importance, the molecule is commonly presented through two-dimensional textbook illustrations that often fail to convey its spatial organization and the relationships among its structural components.

The DNA model was chosen not because it represents the most complex biological structure, but because it provides an appropriate balance between educational relevance and manufacturing feasibility. The characteristic double helix incorporates repeated geometries, symmetry, continuous curvature, and modular organization, allowing the manufacturing workflow to be exercised before extending the methodology to more intricate biological structures such as proteins, viruses, and molecular complexes.

This proof of concept was intended to exercise the proposed methodology on a single case, rather than to demonstrate the maximum manufacturing capabilities of the selected technology or to statistically validate the workflow. The objective was to carry out every stage of the workflow at least once, from digital preparation to physical inspection, while establishing a reference process that can be reproduced and measured more rigorously in future prototypes.

A. Digital Model Preparation

The prototype was based on an open-source DNA double-helix model made available by the MakerWorld user WMZ [8]. The model's MakerWorld page confirms a Standard Digital File License: sharing, sub-licensing, selling, renting, hosting, transferring, or otherwise distributing the digital or 3D-printed object or any derivative work is prohibited, as is any use involving fees or payment; the specific file version was not separately recorded. Rather than being directly printed, the original model was adopted as a geometric reference and underwent a preparation stage consistent with the workflow proposed in this research.

Before fabrication, the geometry was inspected to verify mesh integrity, identify possible inconsistencies, and define the most suitable printing orientation. Digital preparation produced two related files: the assembled double-helix backbone (`Base.stl`), with a final size of 140.0 $\times$ 140.0 $\times$ 246.1 mm (17,158 triangles), and a combined file of modular components — nitrogenous bases and phosphate groups, printed separately for future assembly — scaled to 170% of the original open-source geometry (`Todos_DNA_170%.stl`), with a final size of 130.9 $\times$ 93.4 $\times$ 43.0 mm (9,628 triangles). Slicer inspection initially flagged 8 non-manifold edges in the combined modular-components file; redoing the scaling operation resolved these edges, and the file sliced without further errors afterward.

Because the model exists as a digital asset, modifications such as this scaling can be introduced before fabrication to improve manufacturability or structural resistance without requiring reconstruction of the entire geometry.

Fig. 4 illustrates the prepared digital model used throughout this stage.

B. Prototype Fabrication

During the experimental phase of this research, the laboratory MSLA infrastructure had not yet been fully implemented. The proof-of-concept prototype was therefore fabricated using fused deposition modeling (FDM), distributed across the FDM printers available at FabLab Inatel (Section IV-B): the double-helix backbone (`Base.stl`) was printed on the Creality K1 Max, whose 300 mm build height accommodated the piece's 246 mm height (exceeding the UltiMaker's 205 mm Z limit); the modular components (`Todos_DNA_170%.stl`) — printed individually in different colors, as visible in Fig. 6 — were fabricated using the lab's other available FDM printers. The parameters used for the backbone are summarized in Table II; Fig. 5 shows the sliced model in its printing orientation.

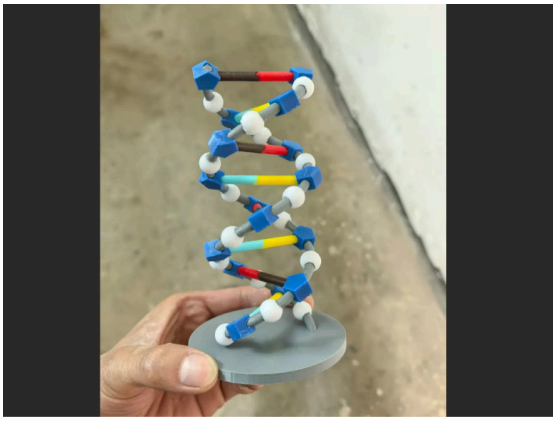


Fig. 4. Digital representation of the DNA double-helix model after preparation for additive manufacturing. Adapted from the open-source model published by WMZ [8].

TABLE II
FABRICATION PARAMETERS FOR THE DOUBLE-HELIX BACKBONE (BASE.STL)

Parameter	Value
Printer / model	Creality K1 Max (build volume 300 × 300 × 300 mm)
Nozzle diameter	0.4 mm
Filament	PLA
Layer height (initial layer)	0.2 mm (0.2 mm)
Default line width	0.42 mm
Wall (perimeter) loops	2
Top shell layers (thickness)	5 (1 mm)
Top surface density	100%
Infill density	15% (grid pattern, 45° direction)
Support strategy	None; the part was oriented specifically to be self-supporting
Print orientation	Vertical, standing on the circular base
Total print time	6h07min (6h01min model printing time)
Material mass used	90.50 g
Failed prints	0 for this piece (≈8 modular components failed separately, see below)
Material cost (FabLab service, invoiced)	R$ 184.00

For reference, the combined modular-components file (`Todos_DNA_170%.stl`) took approximately 3h50min to print and consumed about 70.44 g of PLA, at a cost of approximately R$ 80, fabricated using the lab's other available FDM printers. Of the kit's approximately 45 components, an early test batch printed at the original scale yielded 8 undersized, unusable pieces, which were discarded and reprinted at the corrected 170% scale (Section V-A) — a success rate of about 85% (45/53) for that batch; discarded material is not included in the totals above.

The printing process confirmed that a digital biological representation could be transformed into a physical educational artifact without major manufacturing difficulties, and it provided useful, qualitative information on orientation, mesh preparation, and handling that will carry over to future MSLA production. This evaluation remained qualitative (visual and manual inspection); dimensional metrology, mechanical testing, and user evaluation are left for future work.

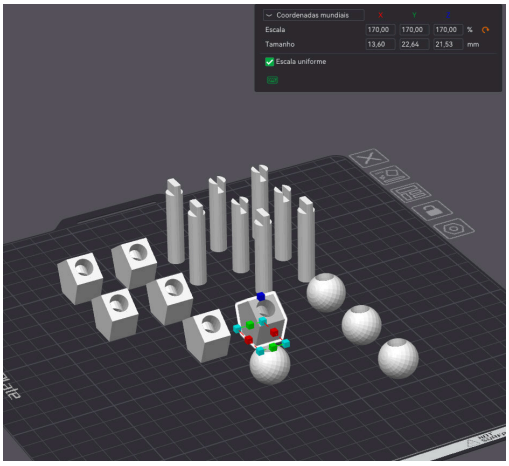


Fig. 5. Slicer preview of the double-helix backbone in its printing orientation, standing vertically on the circular base within the printer's build volume.

Fig. 6 presents the manufactured prototype obtained during this stage.



Fig. 6. DNA educational prototype manufactured using FDM. The assembled double helix demonstrates the manufacturing process end to end; individual components for a future modular educational assembly are under development. Adapted from WMZ [8].

C. Workflow Outcome

Fabrication of the prototype showed that the proposed workflow can be carried out end to end, from digital preparation to a physically assembled artifact — a demonstration on a single

case, with the workflow's applicability to other technologies or structures still to be tested.

The assembled prototype reproduces the macroscopic organization of the DNA molecule on visual inspection, and the project is evolving toward a fully modular educational kit: individual elements representing phosphate groups, deoxyribose sugars, nitrogenous bases, and carbon atoms are being designed and evaluated independently before integration into the complete model. Rather than observing a static representation, students would eventually be able to assemble the molecule themselves, exploring nucleotide composition, complementary base pairing, and the structural organization of the double helix — a design objective inspired by similar modular approaches reported for active learning in science education [9], still to be tested with real students.

Fabrication also surfaced practical, qualitative observations that will guide future prototypes: mesh optimization before slicing reduced printing failures (most visibly in the modular-components batch above), printing orientation affected dimensional quality and handling, and small geometric adjustments — including the rescaling that resolved the non-manifold edges described in Section V-A — improved handling without changing the model's scientific interpretation. These are working notes rather than measured results, but they already point to a clear next step: quantifying them in the SLA/MSLA prototypes planned for future work.

VI. Discussion

The outcome of this proof-of-concept stage indicates that additive manufacturing can be incorporated into biomedical education through a structured, documented workflow. Rather than reporting the fabrication and validation of multiple models, this work emphasizes documenting the entire digital-to-physical process on a single case, so that it can be replicated and evaluated more rigorously going forward.

FDM proved to be an adequate choice for exercising this workflow. Although the literature reports its geometric resolution as lower than that of resin-based systems [4], [7], filament deposition still allowed every stage — model preparation, slicing configuration, fabrication, assembly, and visual inspection — to be carried out at least once, reducing technical uncertainty ahead of adopting higher-resolution processes.

An important distinction underlying this work is between exercising a workflow and validating final manufacturing quality: the goal of the DNA prototype was not the highest possible geometric fidelity, but showing that the proposed steps can be carried out in sequence on a real case. Educational effectiveness, in this view, depends on digital preparation, scientific accuracy, manufacturability, and pedagogical adaptation, not on printing resolution alone.

The proposed methodology complements previous studies on additive manufacturing in education [3], [4]. While existing literature frequently focuses on anatomical replicas and medical simulation, this work extends the discussion toward molecular and microbiological educational models, whose development calls for additional attention to mesh preparation, structural simplification, modularity, and manufacturability.

Accessibility and tactile learning were treated as design objectives throughout: decisions about geometry, dimensions, component separation, and structural reinforcement were made with tactile handling in mind, in the spirit of prior studies on three-dimensional printed educational resources for visually impaired and sighted learners alike [1], [2]. No accessibility or usability evaluation with real users was carried out in this work, so these benefits remain a design objective rather than a demonstrated outcome — a natural next step once the modular kit described in Section V-C is complete.

Finally, the workflow is meant to be scalable: once exercised, it is expected to apply to different classes of biological structures with only minor adaptations during digital preparation, supporting the gradual development of an institutional repository of biomedical educational models.

VII. Limitations

The primary limitation of this study is that the complete proof of concept was carried out on a single structure (a DNA double helix). The stereolithography (SLA/MSLA) printer planned for this project was still under acquisition during the experimental phase and therefore could not be used to manufacture the model; the comparison between FDM and SLA/MSLA presented throughout this work is accordingly based on findings reported in the literature rather than on experimental measurements performed by the authors, and should be read as such [4], [7].

The validation procedure was qualitative: the manufactured prototype was evaluated through visual inspection and verification of assembly feasibility, without dimensional metrology, mechanical characterization, long-term durability testing, or educational assessment involving students, instructors, or visually impaired users. Claims regarding educational usability and accessibility in this paper should accordingly be read as design objectives motivating the work, not as demonstrated outcomes.

Table II reports a reasonably complete fabrication record for the backbone, reconstructed after the fact from the project's build history rather than logged systematically during printing. Notably, it does not include a measured dimensional error between the printed object and the digital file, which would require a follow-up caliper or scanner comparison.

Finally, only one biological structure was investigated. Although the proposed methodology is intended to extend to different educational structures, additional studies are necessary to evaluate its applicability to more complex geometries, other manufacturing technologies, and other areas of health education.

VIII. Future Work

Future work includes manufacturing the same DNA model using MSLA technology in order to experimentally compare dimensional accuracy, surface quality, feature reproduction,

production time, material consumption, and manufacturing cost under comparable conditions.

Additional investigations should include quantitative metrological analyses using precision measuring instruments, mechanical resistance tests, dimensional deviation mapping, and print success/failure rate tracking, together with a documented review of scientific accuracy against a defined checklist. Such analyses would strengthen the validation of the proposed workflow and provide objective criteria for selecting the most appropriate manufacturing technology for a given educational application.

A further direction, likely in the form of an anonymous questionnaire, is a structured educational evaluation involving students, instructors, and visually impaired users, to assess the learning gains, usability, and accessibility benefits that are, at this stage, design objectives rather than measured outcomes. The methodology may also be expanded to different biological structures and, eventually, to anatomical models such as organs and skeletal structures, enabling the creation of larger educational collections.

IX. Conclusion

This work presented a digital-to-physical workflow for producing biomedical educational models through additive manufacturing, illustrated by a first proof of concept: a DNA double-helix model, designed, fabricated, and assembled using the FDM printers available at FabLab Inatel. The process succeeded end to end and surfaced practical manufacturing considerations — around orientation, mesh preparation, and handling — that will directly inform the next iterations of the workflow, including the modular educational kit already under development (Section V-C).

The literature suggests FDM and SLA/MSLA are complementary: FDM for lower cost and simpler operation, resin-based processes for higher dimensional precision and surface quality [4], [7]. This work did not test that comparison experimentally, and reports it on the strength of the cited literature alone.

Learning gains, tactile-accessibility benefits, long-term durability, and a direct FDM-versus-MSLA comparison are left for future work, once the modular kit, the planned resin prototypes, and a structured evaluation with students, instructors, and visually impaired users are in place. What this paper does establish is a documented, reproducible starting point — a workflow, and a first working proof of it — for the biomedical educational models the group plans to build next.

References

[1] C. L. Price, J. M. Houghton, A. P. Preston, R. A. Brown, and S. K. McMahon, "Visible through touch: Open-source 3D-printed tools for inclusive microbiology education," *Access Microbiology*, vol. 8, 2026.

[2] S. Ford and T. Minshall, "Invited review article: Where and how 3D printing is used in teaching and education," *Additive Manufacturing*, vol. 25, pp. 131–150, 2019.

[3] L. Holloway, K. Marriott, and M. Butler, "Accessible maps for the blind: Comparing 3D printed models with tactile graphics," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2018.

[4] A. Aslan and Y. Celik, "A literature review on 3D printing technologies in education," *International Journal of 3D Printing Technologies and Digital Industry*, vol. 6, no. 3, pp. 592–613, 2022.

[5] S. K. Burley *et al.*, "Protein Data Bank: A comprehensive review of 3D structure holdings," *Biomolecules*, vol. 12, no. 10, 2022.

[6] E. F. Pettersen *et al.*, "UCSF ChimeraX: Structure visualization for researchers, educators, and developers," *Protein Science*, vol. 30, no. 1, pp. 70–82, 2021.

[7] AllOnScale, "FDM vs. SLA: Comparison of 3D printing technologies," 2020. [Online]. Available: https://allonscale.com

[8] WMZ, "DNA double helix model," MakerWorld, 2025. [Online]. Available: https://makerworld.com/pt/@WMZ_573004198

[9] W. C. Oliveira *et al.*, "Recurso pedagógico formativo: Aplicação de ferramentas computacionais e impressão 3D no ensino de biologia molecular," *REnCiMa*, vol. 15, no. 4, 2024.

Advertised Competencies for Biomedical Engineers in Brazil: A Descriptive Study Based on Job Announcements in 2026

Carolina Benetti
Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, Brazil
c.benetti@ufabc.edu.br

Anderson G. Santiago
Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, Brazil
gabriel.santiago@ufabc.edu.br

Andrea C. D. Rodas
Biomedical Engineering
Federal University of ABC
São Bernardo do Campo, Brazil
andrea.rodas@ufabc.edu.br

***Abstract*—Despite the growth of the healthcare technology sector in Brazil, there is still a lack of systematic information regarding the professional profile and competencies required for biomedical engineers. Building upon studies conducted in 2022 and 2024, this work aimed to analyze the profile of employment opportunities and the most frequently required skills for biomedical engineers in Brazil based on job advertisements. A total of 115 job postings were collected from the Indeed platform in April 2026 using a web crawler. After screening by three independent reviewers, 61 positions were classified according to professional specialization based on the 2017 Undergraduate Biomedical Engineering Program at the Federal University of ABC (UFABC), and their descriptions were analyzed. The results showed that most opportunities were concentrated in the General (39.3%), Instrumentation (26.2%), and Clinical Engineering (19.7%) areas, predominantly located in the Southeast region of Brazil (64.4%) and requiring on-site work (73.8%). The most frequently hard skills mentioned in job advertisements were quality control (33 job postings) and knowledge of regulatory standards (27 job postings). Communication and teamwork were the most frequently mentioned soft skills. The findings provide a descriptive overview of job advertisements for biomedical engineers, indicating a consolidation of the profession compared with analyses performed in previous years (2022 and 2024).**

***Index Terms*—Biomedical Engineering; Job Market; Hard Skills; Soft Skills.**

I. Introduction

One of the primary objectives of engineering education is to prepare professionals capable of succeeding in the labor market or pursuing entrepreneurial careers independently. Achieving this objective requires continuous monitoring of the competencies demanded by industry, enabling educational institutions to align pedagogical activities with curricular content and better prepare graduates to meet social and professional needs.

Recent studies have consistently demonstrated the increasing importance of soft skills in the engineering labor market across North America, regardless of engineering discipline or research methodology. Hirudayaraj et al. [3], in a meta-analysis, identified teamwork, proactivity, and communication across hierarchical levels and multidisciplinary teams as increasingly important competencies during recruitment processes. Similar conclusions were reported by Beagon and Bowe [4] for engineering education in general and by Akinci-Ceylan and Ahn [5] specifically for the aerospace engineering industry. Likewise, Fleming et al. [1], through an analysis of engineering job advertisements, reached the same conclusions, reinforcing the consensus that collaboration and interdisciplinary communication are essential competencies, regardless of whether evidence comes from employer interviews, graduate follow-up studies, or job advertisement analyses.

The Brazilian National Curriculum Guidelines for Engineering Undergraduate Programs [6] state that the profile of an engineer extends beyond technical expertise. The guidelines emphasize the integration of technical and social competencies throughout the educational process, supported by active learning methodologies, practical activities, interdisciplinary projects, and supervised internships. These experiences expose students to real-world situations that require both technical knowledge (hard skills) and interpersonal and managerial abilities (soft skills).

Specifically for Biomedical Engineering, despite the strategic importance of healthcare technologies, there remains a shortage of systematic information regarding the professional profile, scope of practice, and labor market demand in Brazil. This knowledge gap makes it difficult to align undergraduate curricula with the actual needs of the healthcare technology sector.

Therefore, this study extends previous investigations on Biomedical Engineering employment opportunities in Brazil [7], [8]. The objectives were to classify biomedical engineering job opportunities into different professional areas according to the 2017 Undergraduate Biomedical Engineering Program at the Federal University of ABC (UFABC) [9] and to identify the technical and interpersonal skills most frequently requested by employers based on job descriptions.

II. Materials and Methods

Job postings were collected using a web crawler developed in Python 3.12 with the Selenium library. The crawler extracted job advertisements using the search terms “Biomedical

Engineering" and "Biomedical Engineer" (in Portuguese) from the Indeed platform, which was selected because it provides a publicly accessible and free database of employment opportunities. Data collection was performed in a single batch during April 2026 and restricted to positions located in Brazil. The crawler automatically navigated through all available result pages for each search term and extracted the job title, full job description, and location. No additional filters were applied, and no specific sorting criteria were selected. Duplicate postings were automatically removed by comparing their unique identification numbers.

Three independent reviewers screened the collected advertisements to determine their relevance to Biomedical Engineering. The level of agreement among the reviewers was approximately 96.5%, and disagreements were resolved by consensus. Technical-level positions and advertisements unrelated to Biomedical Engineering, such as positions specifically intended for Nursing or Biomedical Sciences professionals, were excluded.

To evaluate overlap with other engineering and computing professions, additional job postings were collected using the keywords Electrical Engineering, Mechanical Engineering, Chemical Engineering, Mechatronics Engineering, Materials Engineering, Production Engineering, Engineering Management, Data Science, and Computer Science.

The reviewers subsequently classified all relevant Biomedical Engineering positions into six professional areas: Clinical Engineering, Instrumentation, Biomaterials, Computing, Biomechanics, and General. These categories were based on the 2017 Undergraduate Biomedical Engineering Program [9] at UFABC. The additional General category was introduced to include positions suitable for biomedical engineering graduates but involving professional activities shared with other engineering disciplines.

The analysis of job descriptions was organized into four categories:

- General characteristics (salary, geographic region, and work modality);
- Hard skills common to engineering professions (English proficiency, programming, CAD/CAE, risk management, validation, and quality control);
- Biomedical Engineering-specific hard skills (regulatory affairs, maintenance, and management of medical equipment);
- Soft skills (communication, time management, proactivity, and teamwork).

All evaluations were independently performed by the three reviewers based exclusively on the information provided in the job advertisements.

III. Results

The initial search on the Indeed platform returned 115 job postings using the keywords Biomedical Engineering and Biomedical Engineer. After screening by the reviewers, 61 positions were classified as relevant to undergraduate Biomedical Engineering education. The primary reason for exclusion of 54 was that the position required only a technical qualification rather than a bachelor's degree (84.5%). The remaining 17.2% correspond to positions in other fields (Biomedicine, Nutrition, and Veterinary Medicine) or to roles with requirements incompatible with biomedical engineering, such as software support functions or specialization in NR-13 (pressure vessels and boilers). The distribution of the selected positions according to professional area is presented in Figure 1. Most positions were classified in the General category (39.3%), whereas the Biomedical Engineering-specific areas with the largest number of opportunities were Instrumentation (26.2%), Clinical Engineering (19.7%), and Biomaterials (8.2%). No job posting was classified into two categories.

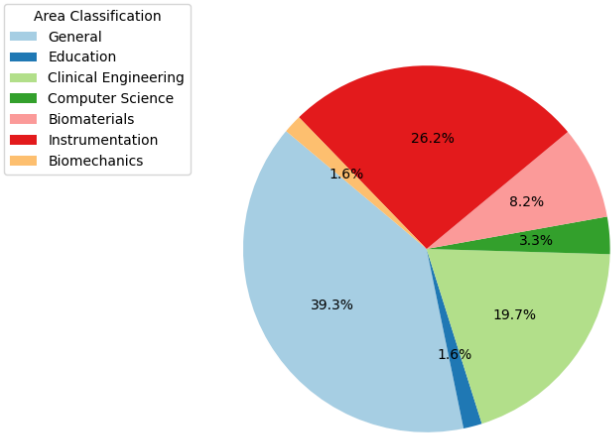


Fig. 1. Distribution of the selected job postings according to professional area.

The analysis of overlap between Biomedical Engineering positions and other professional fields identified five overlapping positions with Electrical Engineering, two with Chemical Engineering, five with Mechatronics Engineering, three with Materials Engineering, two with Production Engineering, one with Engineering Management, and one with Mechanical Engineering. No overlap was observed with Data Science or Computer Science positions.

The analysis of the general characteristics of the job postings showed that 86.9% required a completed undergraduate degree, whereas the remaining 13.1% corresponded to internship positions or required applicants to be currently enrolled in an undergraduate program.

In terms of educational requirements, of the 54 positions analyzed (internships excluded), 83.3% required a bachelor's degree with no mention of postgraduate or specialization qualifications, 13% listed postgraduate or specialization qualifications as desirable, 1.9% required a postgraduate degree, and 1.9% were open to candidates currently enrolled in an undergraduate program. Regarding seniority, the analysis was restricted to permanent positions and to explicit mentions of work experience in job descriptions. Among the same 54 positions, 11.1% required up to 2 years of experience, 11.1%

required 3 to 5 years, 5.6% required 5 to 7 years, and 1.9% required more than 7 years. The remaining 70.4% did not specify any experience requirement.

Most job postings (80.3%) did not disclose salary information. Among the positions requiring a completed undergraduate degree that reported salary, compensation ranged from R$3,000 to R$ 10,000, with an average monthly salary of R$6,400. Regarding geographic distribution, companies located in the Southeast region accounted for 67.2% of the job postings, followed by the Central-West (14.8%), South (9.8%), Northeast (3.3%), and North (1.6%) regions. Two postings did not provide location information (3.3%).

Concerning work modality, 73.8% of the positions (45 postings) required on-site work. Only three positions (4.9%) offered hybrid work arrangements, one position (1.6%) was fully remote, and in twelve postings (19.7%) the work modality was not specified. Furthermore, 21.3% of the advertisements explicitly mentioned business travel or field visits, all of which were associated with the Instrumentation or General professional areas.

Figure 2 summarizes the occurrence of hard skills commonly required across different engineering disciplines. Nearly half of the job postings (28) required English proficiency, and four of these also required Spanish proficiency. Only three advertisements explicitly mentioned programming skills. More than half of the positions (33 postings) required knowledge related to quality control or equivalent terms.

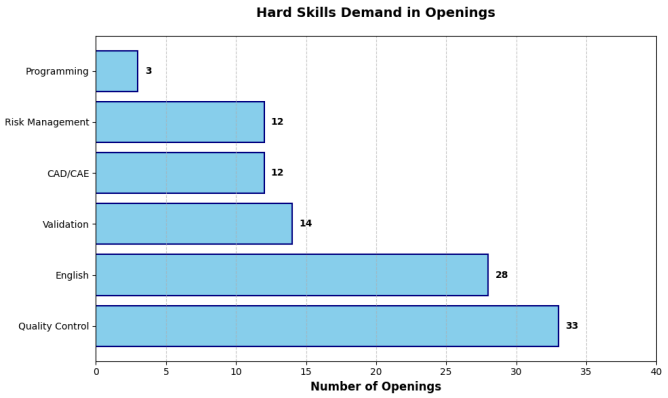


Fig. 2. Frequency of hard skills common to engineering disciplines mentioned in the job descriptions.

The Biomedical Engineering-specific hard skills are presented in Figure 3. Nearly half of the job postings (27) explicitly required knowledge of healthcare regulations and standards. Among these, the Brazilian Health Regulatory Agency (ANVISA) was the most frequently cited regulation (15 mentions), followed by ISO 13485 (12 mentions). Skills related to medical equipment were also commonly requested: 22 postings mentioned equipment maintenance, 12 referred to equipment management, and 10 required knowledge of medical imaging techniques.

Figure 4 presents the five most frequently requested soft skills. Communication, teamwork, proactivity and autonomy, organization/time management, and problem-solving/analytical thinking were the competencies most commonly mentioned by employers.



Fig. 3. Frequency of Biomedical Engineering-specific hard skills mentioned in the job descriptions.

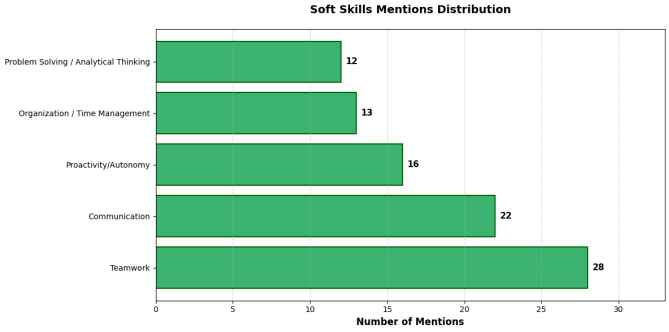


Fig. 4. Five most frequently requested soft skills in the analyzed job postings.

IV. Discussion

A total of 61 job postings identified during the first semester of 2026 were considered relevant to a Bachelor's degree in Biomedical Engineering. This result is consistent with the study conducted using the same methodology in 2024 (67 job postings) [7] and represents a substantial increase compared with the results obtained in 2022 (14 job postings) [8]. Although the linear trend test was not significant ($p > 0.05$), likely due to the small number of time points (n=3), the chi-square test indicated significant year-to-year differences ($p < 0.05$), mainly due to the post-pandemic outlier in 2022 [8]. The similarity between the 2024 and 2026 results may point to a continued consolidation of the Biomedical Engineering profession in the Brazilian labor market. However, future studies incorporating data from additional years are needed to support more robust trend estimations.

A. Professional Areas

The classification of job postings into the areas of Biomaterials, Biomechanics, Computing, Clinical Engineering, Instrumentation, Education, and General (Figure 1) identified employment opportunities in all professional categories. Overall,

39.3% of the positions were classified as General, consisting primarily of activities related to quality assurance, regulatory compliance, certification, sales, and commercialization, which require competencies developed during the common core of engineering programs. Instrumentation represented the second largest professional area (26.2%), mainly involving maintenance, repair, and calibration of medical equipment. Clinical Engineering ranked third (19.7%), including positions focused on equipment management, maintenance supervision, and hospital engineering services. Biomaterials accounted for 8.2% of the positions and primarily involved research, development, and validation of materials for medical devices. The limited number of positions classified in Computing (3.3%), Biomechanics (1.6%), and Education (1.6%) prevented the identification of representative employment profiles for these areas. Considering the significant presence of medical device manufacturers and healthcare technology companies in Brazil, the predominance of employment opportunities in General, Instrumentation, and Clinical Engineering suggests that the labor market recognizes biomedical engineers as professionals qualified to ensure regulatory compliance and support the maintenance and management of healthcare technologies [10].

The analysis of overlap with other engineering professions showed only limited overlap between Biomedical Engineering positions and those identified for other engineering disciplines, suggesting that employers recognize the unique competencies associated with Biomedical Engineering.

Compared with the 2024 results [7], the three most prominent professional areas remained unchanged—Instrumentation, Clinical Engineering, and General—although the proportion of General positions increased, moving from third place in the previous study to first place in the present analysis. The persistence of these three leading professional areas reinforces the perception that Biomedical Engineers are recognized as qualified professionals capable of contributing to multiple stages of the medical device commercialization process. It should be noted that the present study represents a temporal snapshot of the labor market. Consequently, the results may be influenced by temporary hiring demands, such as the establishment or expansion of departments within companies operating in Biomedical Engineering-related sectors or by seasonal bias. Furthermore, the analysis was restricted to a single recruitment platform, which may have influenced the observed employment profile.

B. General Characteristics

The results concerning educational qualifications and seniority indicate that the market remains receptive to candidates holding only a bachelor's degree, given that fewer than 15% of the positions listed postgraduate or specialization qualifications as mandatory or desirable. Additionally, the fact that 70.4% of the job postings did not specify experience requirements suggests that this information is not a priority in the advertisements, which may reflect positive prospects for newly graduated professionals.

The analysis of the job postings revealed an average reported salary of R$ 6,400, which is below the legally established minimum salary for engineers in Brazil, currently equivalent to six minimum wages for a six-hour workday, as established by Federal Law No. 4,950-A/1966 [11]. However, because only approximately 20% of the job postings disclosed salary information and most of them did not specify the corresponding workload, additional data are required to draw reliable conclusions regarding the average salary of Biomedical Engineers in Brazil. Furthermore, some job advertisements simultaneously targeted professionals with technical and undergraduate qualifications, broadening the salary range to accommodate both profiles and consequently affecting the estimated average.

The geographic distribution of the analyzed job postings showed a strong concentration in the Southeast region, which accounted for 67.2% of the 61 postings. This concentration is consistent with the fact that approximately 48% of the healthcare equipment currently in operation in Brazil is located in the Southeast region [12]. However, given the sample limitations (single platform, single month), these findings should not be interpreted as a complete representation of regional job availability for Biomedical Engineers in Brazil.

Most positions (73.8%) required on-site work, reflecting the professional activities typically performed by Biomedical Engineers in hospitals, healthcare institutions, manufacturing companies, and maintenance services. Additionally, 21.3% of the job advertisements explicitly mentioned business travel or field visits, all of which were associated with the Instrumentation or General professional areas, where professionals commonly provide technical support, installation, maintenance, and commercial assistance for medical devices.

C. Skills

The analysis of hard skills common to engineering disciplines (Figure 2) showed that more than half of the job postings (33 positions) explicitly mentioned quality control or equivalent terms. This observation aligns with the Brazilian National Curriculum Guidelines for Engineering Undergraduate Programs [6], which prescribe implementation, supervision, and control of engineering solutions as core competencies expected from engineering graduates. However, this alignment does not necessarily indicate that the guidelines are being effectively implemented in curricula or that the frequency of mentions reflects actual employer priorities, it only shows consistency between advertised requirements and policy expectations. Among the analyzed positions, references to quality control were particularly frequent in the General (14 postings) and Biomaterials (5 postings) categories, reflecting the regulatory and product development activities that characterize these areas.

The analysis of Biomedical Engineering-specific hard skills (Figure 3) suggests an important role for quality assurance and regulatory compliance within the profession. Nearly half of the analyzed job postings (27 positions) mentioned knowledge of healthcare regulations and standards. Skills related to medical

equipment also appeared frequently, with 22 postings mentioning equipment maintenance and 12 referring to equipment management, which may be associated with the Biomedical Engineer's responsibility for ensuring the safe operation, maintenance, and lifecycle management of healthcare technologies.

The analysis of soft skills (Figure 4) corroborates findings reported in previous studies [3]–[5]. Communication and teamwork were the two most frequently mentioned competencies in the analyzed job postings, followed by proactivity and autonomy, organization and time management, and problem-solving and analytical thinking. Given the inherently multidisciplinary nature of Biomedical Engineering, professionals routinely interact with physicians, nurses, technicians, engineers from other disciplines, regulatory specialists, and commercial teams. Therefore, it is expected that effective communication and collaborative teamwork are desirable for professional practice. These competencies are also explicitly recognized by the Brazilian National Curriculum Guidelines for Engineering Undergraduate Programs, which identify effective communication, teamwork, and autonomous learning as fundamental outcomes of engineering education.

However, it is important to note that the frequency of skill mentions in job advertisements does not necessarily equate to their importance in actual hiring decisions or professional practice. Advertised requirements may reflect ideal candidate profiles, marketing strategies, or template descriptions rather than the competencies most critical for job performance.

V. Conclusion

The findings of this study, together with comparisons with similar investigations conducted in 2022 and 2024, provide an updated perspective on the content of job advertisements for Biomedical Engineering on the Indeed platform. The results indicate that the analyzed postings primarily describe opportunities in positions related to the commercialization, maintenance, management, and regulatory compliance of medical devices. However, these findings should not be generalized to the entire Brazilian labor market without further validation using multiple data sources.

Most of the job opportunities analyzed are in positions related to the commercialization, maintenance, management, and regulatory compliance of medical devices. Among the technical competencies, quality control, knowledge of healthcare regulations, and medical equipment maintenance were the most frequently mentioned hard skills in the analyzed job postings. Regarding interpersonal competencies, communication and teamwork were also the most frequently mentioned soft skills. However, these findings reflect what employers choose to include in job advertisements, not necessarily a direct measure of the actual importance or prioritization of these skills in hiring decisions.

The lack of standardization among job advertisements, together with the frequent omission of information regarding working hours and salary, represented the main limitations encountered during the characterization of the analyzed job market.

Overall, the present results indicate a continued consolidation of Biomedical Engineering as suggested by the analyzed job postings. From a policy perspective, these findings may reinforce the importance of aligning undergraduate curricula with the competencies frequently mentioned in job advertisements. However, further researchers are required to correlate curricular content with graduate employment outcomes for curriculum development and educational planning.

Acknowledgment

The authors thank the Universidade Federal do ABC (UFABC) for the institutional support provided during the development of this research.

References

[1] G. C. Fleming, M. Klopfer, A. Katz, and D. Knight, "What engineering employers want: An analysis of technical and professional skills in engineering job advertisements," *Journal of Engineering Education*, vol. 113, no. 2, pp. 251–279, 2024.

[2] A. Ortiz-Rosario, N. H. Choe, A. Shermadou, D. A. Delaine, and T. M. Nocera, "Comparison of job market and employer interest in undergraduate engineering students: An exploratory analysis," in *2020 ASEE Virtual Annual Conference Content Access*, 2020.

[3] M. Hirudayaraj, R. Baker, F. Baker, and M. Eastman, "Soft skills for entry-level engineers: What employers want," *Education Sciences*, vol. 11, no. 10, Art. no. 641, 2021.

[4] U. Beagon and B. Bowe, "Understanding professional skills in engineering education: A phenomenographic study of faculty conceptions," *Journal of Engineering Education*, vol. 112, no. 4, pp. 1109–1144, 2023.

[5] S. Akinci-Ceylan and B. Ahn, "I think it's probably one of the most important skills we could ever know: Insights from early-career engineers on the preparedness of undergraduate degrees for the aerospace industry," *European Journal of Engineering Education*, vol. 50, no. 2, pp. 320–341, 2025.

[6] Ministry of Education (Brazil), "Diretrizes Curriculares Nacionais do Curso de Graduação em Engenharia." [Online]. Available: https://www.gov.br/mec/pt-br/cne/normas-classificadas-por-assunto/diretrizes-curriculares-cursos-de-graduacao. Accessed: Jul. 13, 2026.

[7] C. Benetti, A. G. Santiago, and A. C. D. Rodas, "Biomedical Engineering Career in the Brazilian Job Market: A Comparison Between 2021 and 2024," in *Brazilian Congress on Biomedical Engineering*, pp. 844–851, Springer, 2024.

[8] C. Benetti, A. D. Rodas, and A. G. Santiago, "Analysis of the Brazilian job market for biomedical engineering: A pilot study", in *Latin American Conference on Biomedical Engineering*, pp. 511–519, Springer, 2022.

[9] Federal University of ABC, "Projeto Pedagógico das Engenharias 2017". [Online]. Available: https://www.ufabc.edu.br/images/consepe/resolucoes/2017_-_10_-_projeto_pedagogico_engenharias_-_versao_2018.pdf. Accessed: Jul. 13, 2026.

[10] Aliança Brasileira da Indústria Inovadora em Saúde (ABIIS), "Mapa do Setor de Tecnologia Médica – Brasil e Américas." [Online]. Available: https://abiis.org.br/abiis-lanca-estudo-inedito-que-mapeia-o-setor-de-tecnologia-medica-no-brasil-e-nas-americas/. Accessed: Jul. 13, 2026.

[11] Brazil, "Federal Law No. 4.950-A/1966." [Online]. Available: https://www.planalto.gov.br/ccivil_03/leis/l4950a.htm. Accessed: Jul. 13, 2026.

[12] ABIMO, "Relatório Setorial da ABIMO: Dados consolidados de 2025 e perspectivas para 2026." [Online]. Available: https://abimo.org.br/dados-do-setor/relatorio-setorial/. Accessed: Jul. 13, 2026.

An analysis of the persistent availability of unmarked retracted articles in academic social networks

Floriano Dutra Neto Salvaterra
Programa de Engenharia Biomédica COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
floriano@peb.ufrj.br

Jessica de Andrade Gomes
Programa de Engenharia Biomédica COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jessicaandradeg@peb.ufrj.br

Aldo Fontes-Pereira
Programa de Engenharia Biomédica COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
aldo.fontes@gmail.com

Jurandir Nadal
Programa de Engenharia Biomédica COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
jn@peb.ufrj.br

Renan Moritz VR Almeida
Programa de Engenharia Biomédica COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
renan@peb.ufrj.br

***Abstract*— Objective: This study aimed to evaluate the prevalence of unmarked retracted publications across two major research-sharing platforms. Design: In this cross-sectional study, the Retraction Watch database on retractions was searched for papers fulfilling the following criteria: From Brazil, in all areas of health science, published between 2010 and 2020, and retracted for unreliability-related reasons (data fabrication, fraud, and data/methods/analysis problems). Expressions of Concern (EoCs) were also included. Once selected, these articles were searched on two research-sharing platforms: Research Gate and Academia.edu. Searches were conducted between Nov/24 and Jan/25, and papers were considered available when the manuscript could be found with no indication that a retraction/EoC had occurred and no retracted/EoC version could be found in the respective platform. Identified "abstracts only" within the conditions were also recorded, and, in addition, a standard Google Search was conducted, looking for papers/abstracts available in other platforms. All searches were performed by two researchers and independently double-checked by a third one. In cases of doubt, researchers convened to meet a consensus. The following information was collected: date of publication, date of retraction, title and (Google Scholar) citations received. Results: A total of 66 retractions could be selected. Out of these, 22 (1/3) were available unmarked at the platforms (13 Academia, 5 Research Gate, 4 in both). Additionally, 07 other unmarked abstracts were found (04 Research Gate, 03 in other repositories). The total post-year of retraction citations for the full papers was 730 (average 33.2), and for the abstracts, 166 (average 23.7). Conclusions: This study provides empirical evidence for the high persistence of unmarked retracted publications in academic social networks, thus highlighting potential noteworthy gaps in retraction-signaling practices on academic social networks.**

***Keywords*— Retraction, scientific publications, citation, research ethics**

I. INTRODUCTION

The retraction of scientific articles is a process in which a previously published article is invalidated due to serious problems that compromise its integrity, namely fraud, plagiarism, or significant methodological errors [1, 2]. These retractions serve as a crucial mechanism for maintaining research integrity within the scholarly ecosystem [3]. However, the efficacy and quality of retraction notifications across academic platforms remain unsatisfactory [3,4], potentially compromising the self-correcting nature of science [5].

The use of such papers can have consequences such as loss of scientific credibility, decline of science respect for ethical precepts, and the negative impacts of projects based on erroneous information. Thus, the fact that articles remain in databases without a clear indication of their invalidation is a cause for concern in the scientific community [6-8], which is investing in measures to deal with this problem.

Despite the rising interest in the problem, there are still significant gaps in understanding its causes. So, the extent of the presence of articles without proper retraction signposting in Internet databases is not fully understood, and neither are the reasons for their continued citation [3,4].

Therefore, the aim of this study was to identify retracted articles available on internet databases and to verify their citation rates, with the aim of contributing to an understanding of the impact of this problem and the development of more effective policies for controlling and signaling retractions.

II. METHODS

A. Data

In this descriptive study, the Retraction Watch (RW) database on retractions was searched for papers fulfilling the following criteria: from Brazil, in all areas of health science, published between 2010 and 2020, and retracted for unreliability-related reasons. Nationality was defined according to the first author's affiliation. Unreliability was defined by alleged retraction reasons related to data fabrication, fraud, and problems with data/methods/analysis.

Expressions of Concern (EoCs) were also included in the search. These are warnings, usually implemented when a paper is under investigation for retraction but no decision

has yet been found. They indicate serious concerns on the validity of a paper and thus should also impede paper citation. When a paper received a retraction after an EoC only the retraction was considered for analysis.

B. Search criteria

Once selected, the RW-identified retractions were searched on two research-sharing platforms. The first, Research Gate, has about 20 million members from 190 countries and presents itself as "the Facebook of scientists". The second, Academia.edu, claims to have more than 280 million users [9.10]. These are the two most prominent research-sharing platforms used by researchers.

Searches were conducted in Nov/24 and repeated in Jan/25, and papers were considered available when the manuscript could be found with no indication that a retraction/EoC had occurred and no retracted/EoC version could be found in the respective platform. Searching was perfomed with the standard search engines of the platforms, by looking for the article title, and, also, for the title in conjunction with the keywords "retraction", "retracted", "withdrawn" and "expression of concern" (in order to make sure that a retracted version was not available). In addition, once identified, a double-check was perfomed, so as to guarantee that the available version was the same one registered as "retracted" by the RW database (same venue and publication dates).

Identified "abstracts only" within the conditions above were also recorded, and, in addition, a standard Google Search was conducted, looking for these papers/abstracts in other platforms.

C. Information collection

All searches were performed by two researchers and independently double-checked by a third one. In cases of doubt or disagreement, researchers convened to meet consensus. The following variables were collected for the study: date of article publication, date of retraction, title, and (Google Scholar) citations received. Citations were divided as *post* and *pre-retraction*, where post-retraction citations were defined as those received from the calendar year following retraction onward. From these data, the percentage of post-retraction citations was calculated as: post-retraction citations / (pre-retraction + post-retraction citations).

III. RESULTS

A total of sixty-six (66) retractions could be selected from the Retraction Watch database accordingto the described criteria. Out of these, 22 (1/3) were available unmarked at the platforms (13 Academia, 5 Research Gate, 4 in both). Additionally to these 22 papers, 07 other unmarkedabstracts were found (04 Research Gate, 03 in other repositories).

Figure 1 shows the pre-post retraction rates of the analyzed papers. Total post-year of retractioncitations for the full papers were 730 (average 33.2), and for the abstracts were 166 (average 23.7). Table 1 provides a quantitative summary of the availability and citation data for the 22 full-text articles found unmarked on the platforms, complementing the individual-level visualization presented in Figure 1. Across these articles, citations accumulated both before and after the year of retraction, with the post-retraction period accounting for a substantial proportion of total citations in several cases. The seven abstracts identified without retraction labeling are reported separately, given that their availability was partial and their citation counts represent a distinct subset of the overall findings

Table 1. Summary of availability and citation data for retracted articles found unmarked on academic sharing platforms (2010–2020).

Characteristic	Full papers (n = 22)	Abstracts only (n = 7)
Platform distribution		
Academia.edu only	13	—
ResearchGate only	5	4
Both platforms	4	—
Other repositories	—	3
Total citations (pre + post retraction year)		
Total	1.332	279
Median per article	60,5	39,9
Median per article (IQR)	46,5 (24-75)	23,0 (12-76)
Range	0-320	0-130
Pre-retraction year citations		
Total	602	113
Median per article	27.3	16.1
Median per article (IQR)	16.0 (6–31)	2.0 (0–36)
Range	0–134	0–39
Post-retraction year citations		
Total	730	166
Mean per article	33.2	23.7
Median per article (IQR)	17.0 (6–36)	13.0 (11–35)
Range	0–186	0–61
Proportion of citations post-retraction (%)		
Mean (%)	53.4	74.4
Median % (IQR)	51.0 (36.1–66.0)	78.8 (59.3 –93.3)
Range	11.7 –100.0	36.8 –100.0

IQR: interquartile range. Citation counts retrieved from Google Scholar.

A Spearman correlation between time from publication to retraction (in days) and number of post-retraction citations indicated no statistically significant association ($\rho = -0.108$, $p = 0.631$).

IV. DISCUSSION

A retraction represents the nullification of a scientific paper, implying in its withdrawal from the valid literature. In this study, retractions were restricted to those pertaining to Brazilian first-author affiliations, a pragmatic delimitation adopted to maintain dataset homogeneity and manageability within the scope of the study. The focus on health sciences reflects the direct implications of retracted biomedical literature for clinical practice and public health policies. These choices limit the direct generalizability of findings to other national contexts or scientific disciplines (see "limitations" below).

High citation rates after the retraction of a paper are a well-known phenomenon inthe scientific literature [6,7,11-13]. Indeed, Gabehart [14] studied the persistence of citationsin retracted articles and concluded that around 30% of their citations occurred after the retraction date, demonstrating that, frequently, researchers continue to reference themwithout knowledge of their status.

An explanation for this phenomenon is that the indexing of scientific articles is many times deficient, introducing problems to the clear identification of their scientific status. Such flaws have been found even in renowned indexing databases and repositories, such as PubMed and Web of Science [3,4,6,15].

Another possibility, explored here, is that they can still be found over the Internet, without proper signposting, that is, with no indication that a retraction had happened. This was the possibility studied here, and indeed it could be observed that a large proportion of those retracted articles available in the analyzed databases had no indication whatsoever that they were not valid scientific literature anymore.

Worse yet, these retracted papers had a post-retraction citation rate of about 33 citations/paper. For comparison, Pfeifer and Mouton [17] and Wang et al. [18] reported post-retraction citation rates in of 733/82 and 954/128, respectively, values largely exceeded by the rates observed here. Although direct comparisons are limited by differences in study design, sample selection criteria, contexts and time frames, it is therefore possible that the lack of identification contributed for the dissemination and use of retracted material in other scientific reports.

The problem is even more serious if one considers that the analyzed retractions concerned health sciences, and, if left unchecked, may contribute to the incorporation of invalidated findings into literature reviews. This presents potential downstream implications for clinical guidance and evidence-based policy, a scenario that, while plausible, was not directly assessed in this study and warrants dedicated empirical investigation. Besides, time and resources will be diverted to replicate flawed research, adding to the already existing waste of resources introduced by shoddy scientific literature [16].

One could argue about who could be responsible for the persistent presence of unidentified retracted material on publicly available databases: authors or databases. Indeed, authors have a responsibility to, after a retraction, add clear indications that it has happened to impede the unwarned access to that material. However, databases also have their share of responsibility.Although a formal comparative analysis of their retraction-signaling policies falls outside of the scope of this study, ResearchGate and Academia.edu do not differ much in their governance models and content policies. Both operate as researcher-driven platforms, with some automated content controls and limited editorial vetting; and both would do well striving to develop guidelines and procedures to stop inappropriate dissemination of information.

Limitations: First, as mentioned, the sample is restricted to articles with Brazilian first-author affiliations in health sciences, limiting generalizability to other national or disciplinary contexts. Second, only two platforms were examined; the extent of the problem across other academic repositories, preprint servers, or institutional repositories remains unknown. Third, whether specific retraction reasons (such as data fabrication, fraud, or methodological problems) are associated with greater persistence on sharing platforms or with higher post-retraction citation rates remains an open question. The present dataset, while not powered for such subgroup analyses, may serve as a basis for future investigations examining retraction reason as a potential moderating variable in the persistence of invalidated content. Finally, citation counts were retrieved from Google Scholar, which may differ from other citation databases in coverage and update frequency.

V. CONCLUSION

In summary, this study provides empirical evidence for the high persistence of unmarked retracted publications in academic social networks, highlighting potential noteworthy gaps in retraction-signaling practices on academic social networks. The discussion of this topic addresses basic issues related to academic integrity and the reliability of scientific information. By understanding the factors that contribute to the dissemination of retracted articles and identifying solutions to reduce this problem, it could be possible to promote advances in the way scientific production is managed and disseminated.

Therefore, the availability of these unmarked versions may represent an additional route for the persistence and circulation of retreated content. These findings suggest that current practices for signaling retraction events on academic social networks may be insufficient to prevent continued access to invalidated literature. Potential responses could

include: (a) development of automatic notification mechanisms linking retraction databases (e.g., Retraction Watch, CrossRef) to content-sharing platforms; (b) establishment of clear editorial responsibilities for platform operators regarding content uploaded by registered users; and (c) enhanced guidance for authors regarding the obligation to update or remove uploaded manuscripts following retraction. These observations, while based on a bounded and exploratory sample, may inform future policy discussions on interoperability standards for retraction data.

ACKNOWLEDGMENTS

Part of this study was supported by a CAPES grant (Proex program, code 001), to whom we thank. Author JAG received a scholarship from Faperj (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro), and author RMVRA was also funded by the CNPq-PQ program, to whom we also thank.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest relative to the present work.

REFERENCES

[1]. Barbour V, Kleinert S, Wager E, Yentis S. Guidelines for retracting articles. Committee on Publication Ethics; 2009. https://doi.org/10.24318/cope.2019.1.4.

[2]. Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications. Proc Natl Acad Sci U S A 2012;109:17028–33.

[3]. Bakker CJ, Reardon EE, Brown SJ, et al. Identification of retracted publications and completeness of retraction notices in public health. J Clin Epidemiol. 2024; Sep; 173:111427. doi: 10.1016/j.jclinepi.2024.111427.

[4]. Salami MO, McCumber C, Schneider J. Analyzing the Consistency of Retraction Indexing. MetaArXiv Preprints 2024; Available from: doi.org Access Jan 2025. doi: 10.31222/osf.io/gvfk5.

[5]. Silva, JAT. Risks associated with the citation of retracted medical literature. Journal of Health and Social Sciences 2023; 7(4):363-367. doi: 10.19204/2022/RSKS2.

[6]. Bolland MJ, Grey A, Avenell A. Citation of retracted publications: A challenging problem. Accountability in Research 2022;29:18–25. https://doi.org/10.1080/08989621.2021.1886933.

[7]. Schneider J, Ye D, Hill AM, Whitehorn AS. Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. Scientometrics 2020; 125:2877–913. https://doi. org/10.1007/s11192-020-03631-1.

[8]. Almeida, Renan Moritz VR et al. Plagiarism allegations account for most retractions in major Latin American/Caribbean databases. Science and engineering ethics, v. 22, p. 1447-1456,2016. https://link.springer.com/article/10.1007/s11948-015-9714-5

[9]. Kintisch E. Is ResearchGate Facebook for science? 2021. https://doi.org/10.1126/science.caredit.a1400214.

[10]. Academia. Academia.edu | About. Academia.Edu 2023. https://www.academia.edu/about

[11]. Piller C. Disgraced COVID-19 studies are still routinely cited. Science 2021;371:331

[12]. Schmidt, M. Why do some retracted articles continue to get cited?.Scientometrics 129, 7535–7563 (2024).

[13]. Retraction Watch. Top 10 most highly cited retracted papers. Retraction Watch. https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-citedretracted-papers/ (accessed August 28, 2026).

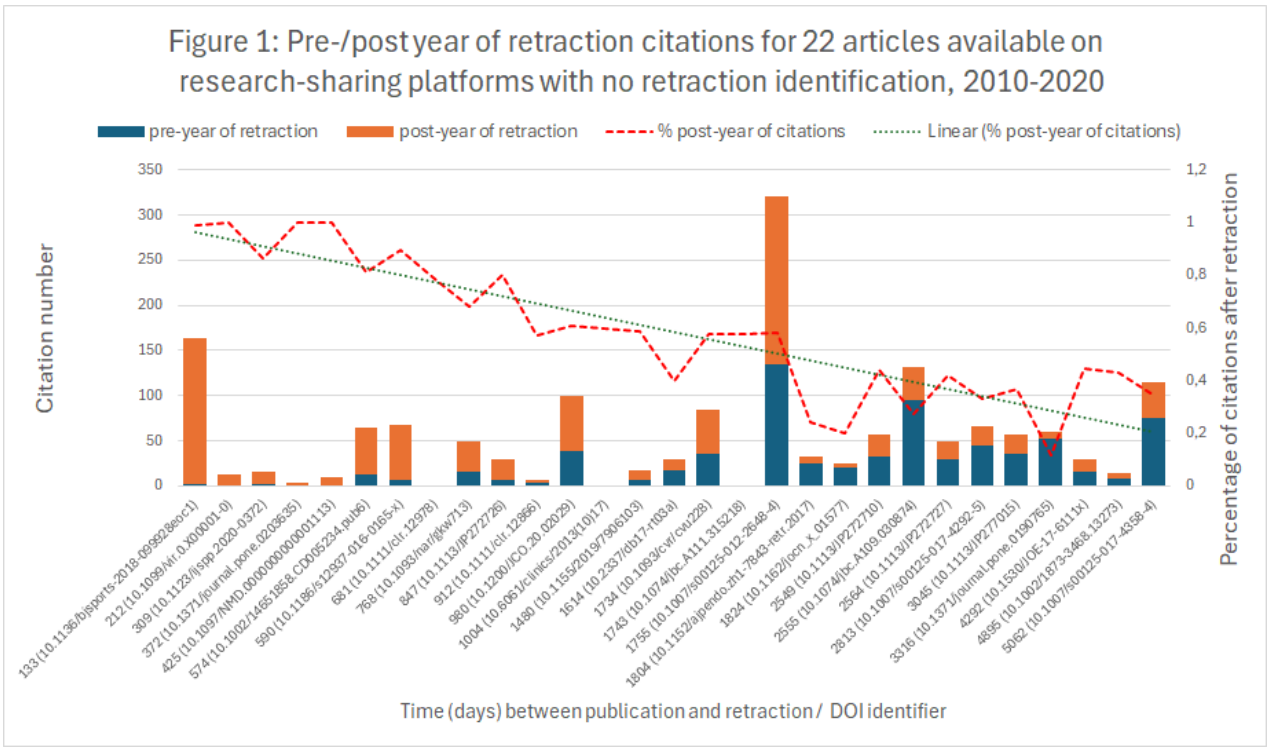
[14]. Gabehart, Mary E. An analysis of citations to retracted articles in the scientific literature. 2005. Available https://ils.unc.edu/MSpapers/3050.pdf

[15]. Ortega, José Luis; Delgado-Quiros, Lorena. The indexation of retracted literature in seven principal scholarly databases: a coverage comparison of dimensions, OpenAlex, PubMed, Scilit, Scopus, The Lens and Web of Science. Scientometrics, v. 129, n. 7, p. 3769-3785, 2024.

[16]. Moher, D., Shamseer, L., Cobey, K. et al. Stop this waste of people, animals and money. Nature 549, 23–25 2017.

[17]. Pfeifer MP. The Continued Use of Retracted, Invalid Scientific Literature. JAMA 1990;263:1420.

[18]. Wang Z, Shi Q, Zhou Q, Zhao S, Hou R, Lu S, et al. Retracted systematic reviews continued to be frequently cited: a citation analysis. Journal of Clinical Epidemiology 2022;149:137–45.


Figure 1: Pre-/post year of retraction citations for 22 articles available on research-sharing platforms with no retraction identification, 2010-2020
pre-year of retraction
post-year of retraction
% post-year of citations
Linear (% post-year of citations)
Citation number
Percentage of citations after retraction
350
300
250
200
150
100
50
0
1,2
1
0,8
0,6
0,4
0,2
0
133 (10.1136/bjsports-2018-099928eoc1)
212 (10.1099/mr.0.000001-0)
309 (10.1123/ijspp.2020-0372)
372 (10.1371/journal.pone.0203635)
425 (10.1097/NMD.0000000000001113)
574 (10.1002/1465-1858.CD005234.pub6)
590 (10.1186/s12937-016-0165-x)
681 (10.1111/cir.12978)
768 (10.1093/har/ykw713)
847 (10.1113/JP272726)
912 (10.1111/cir.12866)
980 (10.1200/JCO.20.02029)
1004 (10.5051/clinics/2013(10)17)
1480 (10.1155/2019/7906103)
1614 (10.2337/db17-1033a)
1734 (10.1093/cvr/cvu228)
1743 (10.1074/jbc.A111.315218)
1755 (10.1007/s00125-012-2648-4)
1804 (10.1152/ajpendo.zh1-7843-retr.2017)
1824 (10.1162/jocn_a_01157)
2549 (10.1113/JP272710)
2555 (10.1074/jbc.A109.030874)
2564 (10.1113/JP272727)
2813 (10.1007/s00125-017-4292-5)
3045 (10.1113/JP277015)
3316 (10.1371/journal.pone.0190765)
4292 (10.1530/JOE-17-611tx)
4695 (10.1002/1873-3468.13273)
5062 (10.1007/s00125-017-4358-4)
Time (days) between publication and retraction / DOI identifier

Aplicação de Metodologias Ativas na Formação em Engenharia Biomédica: Relato de Experiência na UFRN

Gideão Hercules da Silva*, Susana Garcia Sabino*, Pedro Victor da Costa Santos*,
Leticia Thaynna Xavier Araujo Oliveira de Azevedo*, Francisca Thais Queiroz Rego*,
Pablo Henrico Farias da Costa*, Rafael Medeiros e Albuquerque*, Thaise Silveira dos Santos Apolinario*,
Maria Fernanda de Franca Antunes*, Hadson Matheus de Lima Medeiros*, Danilo Alves Pinto Nagem*
*Departamento de Engenharia Biomédica, Centro de Tecnologia,
Universidade Federal do Rio Grande do Norte, Natal-RN, Brasil
{gideao.hercules.116, susana.sabino.093, pedro.costa.709, thaynna.araujo.016, thais.rego.092, pablo.costa.700,
rafael.medeiros.albuquerque.114, thaise.apolinario.049, fernanda.franca.078, hadson.medeiros.700}@ufrn.edu.br
danilo.nagem@ufrn.br

***Resumo*—O presente trabalho relata a experiência de aplicação de metodologias ativas na disciplina de Biomecânica para Engenharia Biomédica da Universidade Federal do Rio Grande do Norte (UFRN), com o objetivo de promover o protagonismo estudantil, a interdisciplinaridade e a integração entre teoria e prática. Foram desenvolvidas duas atividades principais: um projeto contínuo de análise biomecânica do movimento humano, estruturado em três etapas ao longo do semestre, e uma aula invertida baseada na produção de um vídeo educativo em formato de cordel. A experiência possibilitou aos estudantes aplicar conhecimentos de diferentes áreas, desenvolver competências técnicas e transversais, como trabalho em equipe, organização, comunicação, resolução de problemas e tomada de decisão, além de favorecer a aprendizagem significativa por meio de feedback contínuo. Os resultados sugerem uma associação entre a implementação das metodologias ativas e a melhoria dos indicadores de percepção discente.**

***Resumo*—This paper reports on the experience of applying active learning methodologies in the Biomechanics for Biomedical Engineering course at the Federal University of Rio Grande do Norte (UFRN), aiming to foster student agency, interdisciplinarity, and the integration of theory and practice. Two main activities were conducted: an ongoing project involving the biomechanical analysis of human movement—structured in three stages throughout the semester—and a flipped classroom activity based on the production of an educational video in the style of cordel (traditional Northeastern Brazilian folk poetry). The experience enabled students to apply knowledge from various fields and develop both technical and transversal skills—such as teamwork, organization, communication, problem-solving, and decision-making—while also promoting meaningful learning through continuous feedback. The results suggest a link between the implementation of active learning methodologies and improvements in indicators of student perception.**

***Index Terms*—Educação, Engenharia Biomédica, Aprendizagem, Biomecânica**

I. Introdução

Os modelos tradicionais de ensino em engenharia, durante muitos anos, consolidaram práticas pedagógicas centradas na transmissão unilateral do conhecimento, seja por meio de aulas expositivas, seja por atividades experimentais predominantemente demonstrativas. Nesses modelos, o professor costuma ocupar a posição central do processo educativo, enquanto o estudante permanece em uma postura mais passiva. De acordo com Geurts et al. [1], ambientes educacionais que não valorizam a participação discente tendem a limitar o potencial transformador da educação e o desenvolvimento da autonomia, da autoconfiança e do pensamento crítico. Metodologias excessivamente centradas no professor também podem contribuir para a desmotivação discente, resultando em menor participação em sala, baixo rendimento acadêmico e menor retenção do conhecimento [2]. Tais habilidades e competências são, atualmente, objetivos centrais dos novos currículos de engenharia [3].

Nesse cenário, promover o protagonismo estudantil ocupa posição central nos debates pedagógicos contemporâneos, sendo compreendido como a capacidade do estudante de tomar decisões, participar ativamente da construção do conhecimento e assumir a responsabilidade pelo próprio percurso formativo [4]. Como consequência, a ausência de protagonismo compromete tanto o desempenho acadêmico quanto a formação profissional e cidadã, reforçada muitas vezes pela insegurança e pelo medo de errar Gonçalves et al. [5].

Para superar essa realidade, torna-se necessária a implementação de estratégias pedagógicas capazes de transformar o aluno em agente central da construção do conhecimento, como o uso de espaços não formais de ensino, metodologias ativas e aulas expositivas dialogadas. Entre as metodologias ativas, pode-se destacar estratégias como a Aprendizagem Baseada em Projetos (PBL ou ABP), a resolução de problemas, as atividades colaborativas, e aulas invertidas que favorecem o desenvolvimento da responsabilidade, da autorregulação, do pensamento crítico e da autonomia intelectual. Essas abordagens apresentam grande potencial na área de exatas, uma vez

que exigem interação, tomada de decisão, gestão organizacional, autogestão e resolução conjunta de problemas [6].

Contudo, a implementação dessas metodologias exige mudanças estruturais, demandando mediação docente e estratégias organizacionais que orientem o planejamento e o trabalho colaborativo [6]. Soma-se a isso o desafio da fragmentação curricular, na qual a organização dos conteúdos em disciplinas isoladas dificulta a percepção das conexões entre diferentes saberes e prejudica a construção de uma visão sistêmica da futura atuação profissional [7]. Nesse contexto, práticas interdisciplinares tornam-se essenciais para integrar conceitos, aproximar a teoria de situações reais e conferir maior significado à aprendizagem [8], [9]. Nos cursos de engenharia, essa articulação é indispensável, sobretudo na Engenharia Biomédica, cuja atuação profissional exige a integração constante entre ciências exatas, biológicas e da saúde [10].

Diante da dificuldade histórica em conectar saberes com estruturas de ensino tão distintas, práticas integradoras apoiadas em metodologias ativas tornam-se essenciais. Nesse contexto, relatos de experiência tornam-se relevantes por documentarem estratégias pedagógicas aplicadas em contextos reais de ensino, permitindo compartilhar desafios, adaptações e resultados que podem subsidiar novas práticas no ensino de Engenharia Biomédica. Dessa forma, este texto apresenta um relato de experiência sobre a aplicação de duas metodologias ativas na disciplina de Biomecânica para a Engenharia Biomédica, na UFRN, realizadas no período letivo de 2026.1. A proposta concentra-se no desenvolvimento de competências e habilidades específicas, bem como na aplicação de conceitos de dinâmica.

Adicionalmente, com o intuito de complementar o relato de experiência, o trabalho apresenta um estudo comparativo descritivo fundamentado no histórico de avaliação da disciplina pelos próprios estudantes, compreendendo o período de 2018.1 a 2025.2, período esse em que ocorreu a inserção da disciplina na grade. A análise busca confrontar as avaliações docentes de semestres anteriores com os resultados coletados após a inserção estrutural das metodologias ativas, ocorrida a partir de 2023.1. Essa abordagem visa demonstrar o impacto do modelo ativo no engajamento, na motivação e na satisfação geral da turma ao longo dos anos.

II. Metodologia

No início da disciplina de Biomecânica para Engenharia Biomédica na UFRN em 2026.1 o docente apresentou o plano de curso contendo duas atividades práticas de caráter não tradicional a serem desenvolvidas ao longo do período letivo, com apresentações presenciais na sala de aula. Essa apresentação inicial permite que os estudantes saibam, previamente, quais conhecimentos serão necessários para sua apresentação final. Esse conhecimento também é reforçado durante as aulas, nas quais o professor associa o aprendizado às entregas das atividades. Há também aulas somente para discutir o desenvolvimento das atividades e esclarecer dúvidas sobre elas.

A turma apresentada neste relato é composta por 10 estudantes, que foram organizados em dois grupos, os quais atuaram de forma distinta desenvolvendo as duas atividades independentemente. Será apresentado aqui o resultado de somente um dos grupos em cada uma das atividades. Para garantir a participação de todos do grupo, além da avaliação pelo professor no conteúdo e na apresentação, cada aluno entregou um relatório de autoavaliação e de avaliação dos colegas, onde estimaram o desempenho, comprometimento, entendimento, entregas e trabalho em equipe, em cada uma das atividades.

A. Atividade 01 – Análise do movimento

A atividade 01 consistiu na análise biomecânica do movimento humano em um projeto desenvolvido em três entregas parciais ao longo do semestre. Ela iniciou como o registro em vídeo de movimentos do membro superior humano, seguido do rastreamento de pontos anatômicos via software Kinovea® e do desenvolvimento de modelagem cinemática e cinética bidimensional em linguagem Python, Matlab ou qualquer outra que o aluno tenha conhecimento para cálculos e construção de gráficos e análises biomecânicas.

1) A condução do projeto: As atividades foram desenvolvidas em três etapas ao longo do semestre, conforme apresentado na Figura 1: análise cinemática do ponto material, cinemática de corpo rígido e cinética de corpo rígido bidimensional. Todo o material de apoio foi disponibilizado aos estudantes por meio de sistema acadêmico online, contendo a descrição de cada etapa do projeto.

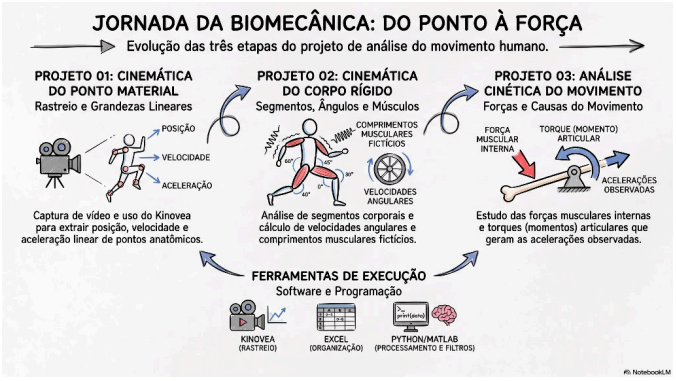


Figura 1: Divisão das atividades de acordo com as entregas. Fonte: autoria própria – Desenvolvido por IA, NotebookLM.

Após a primeira entrega do material solicitado na etapa 01, o professor realizou uma análise detalhada do relatório, destacando os pontos positivos e negativos do grupo. Também apresentou orientações para a condução das duas etapas seguintes, adequou o texto ao formato de relatório e buscou relacionar os conhecimentos trabalhados com conteúdo de outras disciplinas. Em seguida, devolveu o texto aos estudantes com correções, observações e análises de cada etapa e de cada fórmula apresentada.

A análise do segundo relatório ocorreu de forma mais pontual, durante encontros em sala de aula com cada grupo,

nos quais foram indicadas as falhas que ainda persistiam nos relatórios. Nessa etapa, não houve devolutiva escrita do documento, como ocorreu na primeira entrega.

A terceira etapa consistiu em uma apresentação expositiva realizada pelos estudantes, contemplando todo o material desenvolvido, os gráficos, as análises e a relação entre os valores determinados nos cálculos e os movimentos analisados, como flexão e extensão. Ao final, o professor avaliou o conteúdo apresentado, e questionou os alunos sobre as decisões durante o projeto.

2) O desenvolvimento do projeto: Os estudantes realizaram a filmagem do movimento das duas articulações em um único plano, com marcações anatômicas aplicadas ao participante. O vídeo foi processado no software Kinovea®, no qual foram rastreados os pontos anatômicos e exportados dados em formato CSV (comma separated values), contendo as informações de posição de cada ponto em cada instante de tempo. Com os dados obtidos, foi realizada uma análise cinemática e cinética bidimensional do membro por meio de programação em Python. O fluxo de execução estabelecido pelos discentes é apresentado na Figura 2.

Na primeira etapa, foram calculados os ângulos articulares, as velocidades e as acelerações lineares dos pontos anatômicos, a partir dos quais foram gerados gráficos animados e estáticos. Também foi realizada a comparação entre os valores da análise cinemática calculada pelos estudantes e os valores determinados pelo software Kinovea®, utilizado como referência para validação.

Na segunda etapa do projeto, foi analisada a dinâmica muscular do par agonista-antagonista bíceps e tríceps braquial e sua relação com a articulação do cotovelo. A escolha do par muscular a ser estudado foi realizada pelos alunos. Nessa etapa, foram calculadas as velocidades e acelerações angulares, além das velocidades e acelerações associadas à contração muscular. Novamente, foram gerados gráficos animados e estáticos; e relações biomecânicas do movimento foram descritas e estudadas, como velocidade de contração, movimento de flexão ou extensão entre outros, relacionando-as aos valores encontrados.

Na terceira e última etapa, foi realizada a análise das forças envolvidas no modelo criado de cálculo cinético. Foram determinados os ângulos entre cada músculo e o ponto de aplicação da força, o torque necessário para o movimento, e a força exercida por cada músculo do modelo. Mais uma vez, foram elaborados gráficos e discutiu-se a relação entre esses resultados e o movimento realizado.

3) Critérios avaliativos da atividade: A avaliação de cada relatório parcial e da apresentação final baseou-se em quatro eixos:

- Apresentação e qualidade do relatório: Clareza, organização e tipo de representação gráfica dos dados, avaliando negativamente a presença de gráficos confusos ou sem legendas, e a falta de objetividade no relatório.
- Qualidade do memorial de cálculo descritivo: Abrangência conceitual, organização lógica e modelagem físico-matemática clara das equações de movimento, garantindo que o documento seja autoexplicativo para a replicação por terceiros.
- Recursos: Adequação das ferramentas didáticas utilizadas (como animações) e profundidade das referências bibliográficas adotadas.
- Criatividade e trabalho em equipe: Complexidade do movimento escolhido, inovação na apresentação e nos conceitos apresentados.

B. Atividade 02 – Aula invertida

Os alunos deveriam apresentar uma aula criativa sobre um tema de Biomecânica, integrando uma apresentação presencial a um vídeo educativo de formato livre com até 15 minutos. Além da exibição, exigiu-se a entrega de um roteiro, de material de apoio e de questões relacionadas, que foram posteriormente compartilhados com toda a turma para estudo.

1) A condução da atividade: A proposta da atividade pode ser observada no infográfico da Figura 3, que apresenta o roteiro do desenvolvimento.

A proposta do vídeo e o material de apoio foram apresentados e discutidos com os estudantes, que selecionaram um dos temas propostos. A elaboração do vídeo teve formato livre, e pôde ser desenvolvida em forma de jornal, teatro, apresentação gravada ou qualquer outra estratégia adotada pelos alunos, que foram estimulados a serem criativos. Os alunos deveriam entregar para o professor o roteiro, o relatório, as fontes consultadas e exercícios sobre o tema. Ao longo do semestre, foram realizadas aulas destinadas ao esclarecimento de dúvidas e à orientação sobre o desenvolvimento da atividade. Ao final da apresentação, o grupo aplicou o questionário sobre o tema para o restante da turma.

2) O desenvolvimento da aula invertida: O grupo, buscando maior engajamento dos discentes e aproveitando o contexto festivo do mês de junho para valorizar a cultura nordestina, optou por estruturar o trabalho no formato de literatura de Cordel. Inspirado pela possibilidade de comunicar o conteúdo de forma acessível e popular, o grupo uniu a elaboração do cordel ao conteúdo relacionado com o crescimento do tecido ósseo. A primeira etapa do projeto consistiu na elaboração do relatório técnico, que serviu de base teórica para a adaptação dos conceitos biológicos e mecânicos à linguagem literária.

A segunda etapa consistiu no desenvolvimento da construção das rimas e da métrica tradicional do cordel. No entanto, o grupo não possuía experiência prévia com esse gênero literário nem ferramentas adequadas para a criação manual de xilogravuras. Por esse motivo, optou pela utilização de ferramentas de Inteligência Artificial generativa como suporte inicial para a estruturação dos versos e para a geração das ilustrações. Foi utilizado Google Gemini, com modelo Nano Banana, para gerar imagens e ChatGPT em conjunto com Google Gemini para composição dos versos.

A Inteligência Artificial ofereceu um ponto de partida ágil para a construção dos versos; porém, os resultados iniciais não alcançaram o padrão desejado pelo grupo, pois não apresentavam o rigor estético, cultural e literário exigido pela

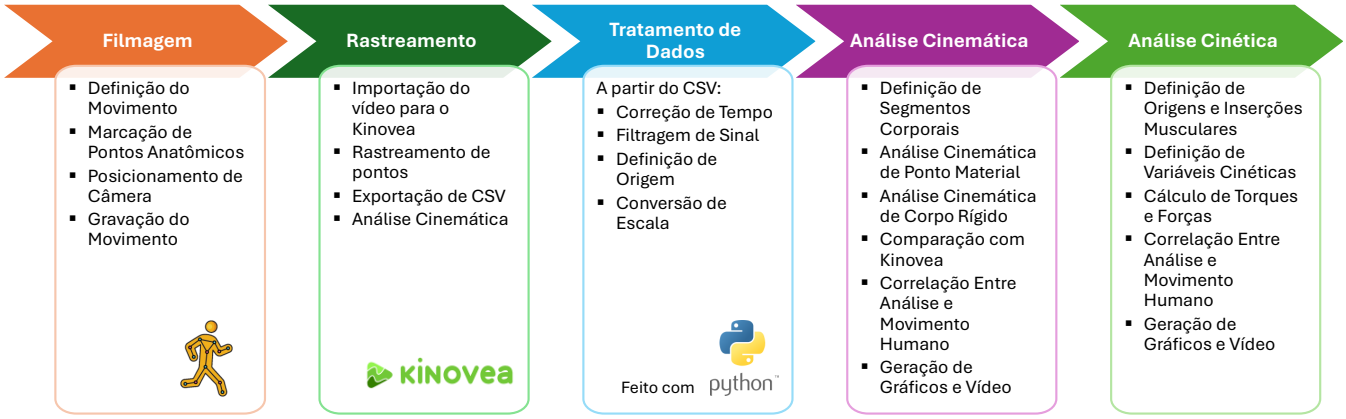


Figura 2: Fases da elaboração do projeto contínuo. Fonte: autoria própria.

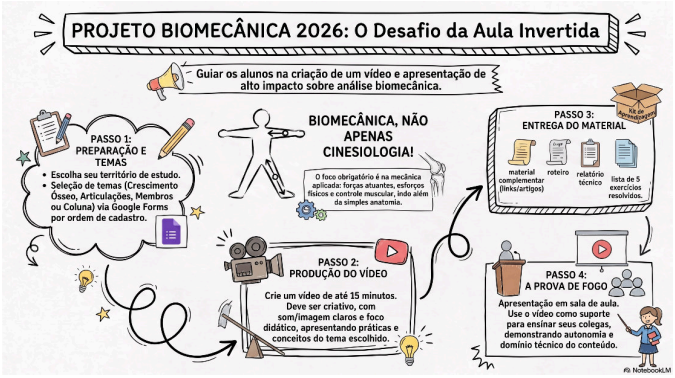


Figura 3: Roteiro da elaboração da atividade de aula invertida com vídeo. Fonte: autoria própria - Desenvolvido por IA, NotebookLM.

tradição poética. Foi necessária, portanto, uma intervenção criteriosa dos estudantes. Os versos passaram por um processo minucioso de revisão manual, no qual a equipe realizou correções de métrica, ajustes de rimas, substituições lexicais e inclusão intencional de expressões típicas nordestinas. Esse processo iterativo permitiu que a linguagem técnica sobre o comportamento mecânico dos ossos incorporasse ritmo, identidade e características próprias do cordel.

As imagens geradas inicialmente também não apresentavam a textura rústica, a identidade cultural e a expressividade próprias da xilogravura. Para compensar essa lacuna, o grupo valorizou a oralidade, elemento central dessa literatura historicamente concebida para ser cantada e declamada em voz alta diante do público. O vídeo foi narrado com sotaque típico por um discente natural do Sertão do Apodi (RN), promovendo uma imersão sonora que preservou características da apresentação dos poetas populares.

Por fim, durante a apresentação em sala, além da exibição do vídeo contendo as etapas do cordel, um dos estudantes, em referência ao legado de Luiz Gonzaga, o Rei do Baião, utilizou sua sanfona para reinterpretar o tema. Sustentando a apresentação final por voz e instrumento, o grupo articulou conceitos biomecânicos, linguagem regional e elementos da cultura popular nordestina.

3) A avaliação da atividade: Nessa atividade, realizada no meio do semestre, a avaliação foi calculada seguindo critérios específicos, descritos no documento entregue aos alunos e discutido em sala; são estes:

- Apresentação e qualidade do vídeo: Aspectos técnicos de gravação, edição, dinamismo visual e clareza de áudio.
- Qualidade do conteúdo: Profundidade conceitual com foco estrito em biomecânica aplicada, forças atuantes e torques no corpo humano, em detrimento de uma abordagem meramente anatômica ou cinesiológica tradicional.
- Recursos didáticos: Utilização de materiais de apoio inovadores para engajamento da turma, como slides, folhetos ilustrados, música, formato da apresentação entre outros.
- Relatório de produção e perguntas: Análise do material entregue, o roteiro, o relatório com o estudo realizado, a lista com os exercícios resolvidos entre outros.

Durante e após a apresentação os alunos foram questionados pelo professor da disciplina sobre os conceitos apresentados e sobre a participação de cada um.

C. Avaliação do desempenho dos estudantes

A UFRN possui sistema institucional de avaliação estudantil quanto à satisfação com as disciplinas cursadas a cada semestre. O sistema inclui avaliação da atividade docente, assim como autoavaliação do engajamento discente. Para avaliar o impacto pedagógico e a percepção discente em relação à disciplina de Biomecânica de forma longitudinal, realizou-se um estudo comparativo retrospectivo fundamentado nos relatórios do sistema de avaliação. O estudo abrangeu as quinze turmas entre 2018.1 e 2025.2, período no qual a grade curricular atual foi implementada no curso de Engenharia Biomédica. O número de estudantes por turma variou expressivamente, entre 6 e 35 alunos.

Em razão dessa variação e do baixo número de turmas analisadas, a taxa de aprovação pode não representar uma métrica robusta. Por esse motivo, o delineamento metodológico

concentrou-se na evolução dos indicadores quantitativos de autoavaliação dos estudantes e de avaliação docente, combinados à análise qualitativa do conteúdo de comentários e críticas registradas nos relatórios institucionais. Dessa forma realizou-se uma análise do grau de satisfação e do engajamento dos alunos com a disciplina. Essa abordagem permitiu contrastar o período conduzido sob o modelo expositivo tradicional com a fase de implementação contínua das metodologias ativas iniciada em 2023.1.

III. Resultados e Discussão

Desde a apresentação das atividades e da divisão da turma em grupos os discentes precisaram definir objetivos, levantar requisitos, pesquisar, organizar cronogramas e tomar decisões técnicas, permitindo a aplicação integrada de conceitos abordados em diferentes componentes curriculares anteriormente cursados. Inicialmente, os estudantes passaram por um processo gradual de reconhecimento das habilidades e afinidades de cada membro do grupo, o que permitiu a divisão em sub equipes focadas em programação e elaboração de relatório. Todos os grupos evoluiram durante o decorrer do semestre e as interações entre os membros foram cruciais para um desenvolvimento ótimo das atividades.

A. Atividade 01 – Análise do movimento

Projetos acadêmicos em grupo são normalmente desenvolvidos no final do semestre, a experiência acadêmica de projeto continuado, sendo elaborado durante o semestre com conhecimentos cumulativos foi de maior complexidade organizacional para os alunos.

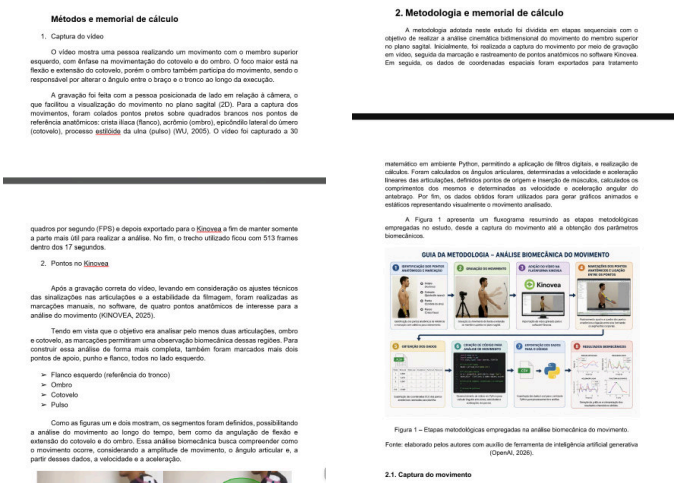
Durante a primeira etapa surgiram dificuldades relacionadas à falta de entendimento do problema, à falta de compreensão na comunicação entre os membros e à fragmentação das atividades, especialmente no desenvolvimento dos códigos e na escrita do relatório. A partir do primeiro *feedback* recebido do professor, de forma oral e escrita, os alunos reorganizaram o desenvolvimento e rearranjaram a estratégia de gerenciamento do projeto. Os estudantes adotaram, por conta própria, uma estratégia de gestão mais organizada e baseada em Kanban. Passaram então a realizar a definição de tarefas em tópicos, fazer o acompanhamento de pendências e realização de reuniões periódicas para alinhamento. Dessa forma, a avaliação contínua permitiu constante reiteração e adaptação do desenvolvimento do projeto.

Os resultados da experiência de aprender com base em projetos incluíram melhorias significativas na comunicação, na organização e na integração do trabalho em equipe. Os alunos relataram que a adoção de uma estrutura mais clara de divisão de tarefas possibilitou maior autonomia e melhor aproveitamento do tempo disponível. Como consequência, houve maior colaboração entre os responsáveis pela programação e pela escrita do relatório. Isso resultou em um relatório que a cada etapa ficou mais coeso, conciso e mais consistente quando comparado ao relatório inicial, além de códigos mais bem integrados com maior clareza sobre sua estruturação e capacidade de gerar os resultados apresentados no relatório. Toda a experiência proporcionou aos discentes a oportunidade de assumir o papel de executor e por consequência, sair do papel de ouvinte na turma.

Além disso, a experiência favoreceu não apenas a consolidação dos conhecimentos teóricos da disciplina, mas também contribuiu para o desenvolvimento de competências necessárias para a atuação de um engenheiro biomédico no mercado de trabalho, tais como: gerenciamento de projeto, organização de cronogramas, levantamento de requisitos, tomada de decisão, colaboração e resolução de problemas em grupo.

É interessante destacar que em sessões de *feedback* com os alunos foram levantadas discussões sobre que tipo de filtro se deve utilizar no sistema, butterworth ou o de Savitzky-Golay, pois os alunos durante a leitura da documentação do Kinovea® perceberam que o programa utiliza o segundo. Foi também observado pelos alunos a possibilidade de no cálculo da derivada numérica utilizar o método de Diferenças Finitas Centrais (DFC), ao contrário do método tradicional. Discussões essas que não seriam trazidas pelo professor em um modelo tradicional de ensino em uma disciplina de biomecânica.

Outro resultado importante é em relação a qualidade do material produzido. Do primeiro relatório para o último, é notória a diferença de qualidade, desde a capa até a qualidade de figuras e do texto como pode ser observado na Figura 4, que apresenta um recorte do primeiro relatório (4a) e do último relatório (4b) do grupo. Os relatórios são incrementais, dessa forma se utiliza do primeiro relatório para continuar a escrita para o segundo e para o terceiro. Porém o grupo entendeu a necessidade de mudanças em todo o texto durante o desenvolvimento.



(a) (b)

Figura 4: (a) recorte do início do capítulo de metodologia do primeiro relatório. (b) Recorte do início do capítulo de metodologia do último relatório. Fonte: autoria própria.

Fica claro que o desenvolvimento do projeto contínuo exigiu o resgate e progressão de conhecimentos adquiridos anteriormente em áreas diversas, como programação em linguagem

Python, análise de dados, anatomia, cinesiologia, e escrita científica. Somente a partir da integração destes, em conjunto com o conhecimento construído na disciplina de biomecânica, pôde-se desenvolver o projeto.

B. Atividade 02 – Aula Invertida

Essa experiência, apesar de não ser contínua durante todo o semestre, também contribuiu significativamente para o desenvolvimento de competências transversais exigidas na prática da Engenharia Biomédica. Competências essas que no ensino tradicional seria de difícil introdução em uma matéria como Biomecânica para Eng. Biomédica. O esforço de traduzir conceitos complexos de mecânica óssea para uma linguagem acessível e ritmada exigiu que os estudantes superassem a memorização passiva. Ao estruturarem as estrofes, os discentes precisaram dominar os fenômenos biomecânicos a ponto de conseguir explicá-los de forma acessível e popular. Essa transposição didática confirmou a eficácia das metodologias ativas em integrar o rigor científico à prática ativa.

Exemplos desse desafio podem ser observados na Figura 5 (5a) e (5b), as quais representam conceitos como o de remodelação óssea, uma sobre osteoblastos e osteoclastos (5a) e outra sobre a lei de Wolf (5b) traduzidos graficamente e na leitura do cordel.



(a) (b)

Figura 5: (a) Remodelagem óssea por meio de osteoblastos e osteoclastos, (b) Remodelagem óssea devido a tensões (lei de Wolf). Fonte: autoria própria, gerado com IA (Google Gemini).

Além do impacto pedagógico, uma reorganização do grupo também foi necessária. O uso da IA para realizar grande parte do trabalho se provou ineficiente da forma como pensado, o que levou a um maior estudo e dedicação por parte do grupo.

Diante dessas limitações, foi necessária uma intensa participação e intervenção dos alunos por meio do entendimento dos conceitos envolvidos na aula. A revisão manual dos versos pelos estudantes foi essencial para garantir a harmonia sonora exigida pelo gênero. Esse processo demonstrou na prática que a confecção autêntica do cordel exige um aprendizado que não se extrai apenas de regras gramaticais programáveis, mas de uma formação que provém do convívio, da tradição coletiva e de um saber transmitido ao longo de gerações.

No aspecto visual, a tentativa de simular ilustrações por IA revelou uma acentuada lacuna estética e identitária, sendo necessários refinos e mais refinos, nas descrições das imagens para uma textura mais rústica e com a força expressiva das metáforas visuais concebidas de forma orgânica pelos xilógrafos tradicionais. A Figura 06. apresenta imagens desenvolvidas para um mesmo tópico, mas gerados de forma diferentes e que foram adaptadas para a versão final. A Figura 6 apresenta imagens desenvolvidas para um mesmo tópico, mas gerados de forma diferentes e que foram adaptadas para a versão final.



(a) (b)

Figura 6: (a) e (b) Tentativas de representar a relação de carga e mudança óssea no desenho do cordel. Fonte: autoria própria, gerado com IA (Google Gemini).

Para suprir essa incapacidade tecnológica em reproduzir a "alma"do cordel, o grupo ancorou seus resultados práticos na força da oralidade. Dessa forma o grupo acabou desenvolvendo duas apresentações complementares com oralidades distintas, com conteúdo semelhante, mas com descrições diferentes. No cordel apresentado no vídeo foi adicionado um sotaque genuíno do sertão que narrou todo o desenvolvimento. Na apresentação em sala uma nova apresentação com nova oralidade, um cordel com performance movida a sanfona ao vivo, resgatou a atmosfera dos mercados públicos e praças onde os "poetas de bancada"interagem com a sociedade (Figura 7). O produto do trabalho tornou-se, assim, um híbrido que uniu a técnica das ciências exatas ao legado de uma cultura ancestral e conceitos técnicos e multidisciplinaridade em uma aula invertida com música e diversão.

C. Avaliação discente e Eficácia Pedagógica

A análise dos relatórios institucionais de avaliação da disciplina de 2018.1 a 2025.2 indica uma melhora positiva a partir de 2023.1, período em que as aulas puramente expositivas foram substituídas pelo projeto prático contínuo e pela sala de aula invertida. Conforme pode ser observado na Figura 8, os índices médios de autoavaliação e percepção docente (medidos de 0 a 10) eram de 8,6 e 7,8 respectivamente, e mudaram para 9,1 e 9,1 após a inserção da metodologia. Pode-se observar que os discentes chegaram a relatar um índice de engajamento estudantil superior a 9,5 em mais de uma ocasião.

Essa análise quantitativa deve ser seguida em trabalhos futuros de uma análise qualitativa do engajamento, da motivação e da satisfação geral dos alunos, visto que há uma grande variedade no número de alunos regulares e poucos semestres especificados, o que impossibilita uma análise estatística completa.

Ao se analisar o engajamento, por meio da participação em sala, e a motivação e a satisfação por meio dos comentários registrados nos relatórios institucionais de avaliação, essa



Figura 7: Apresentação em sala do cordel. Fonte: autoria própria.

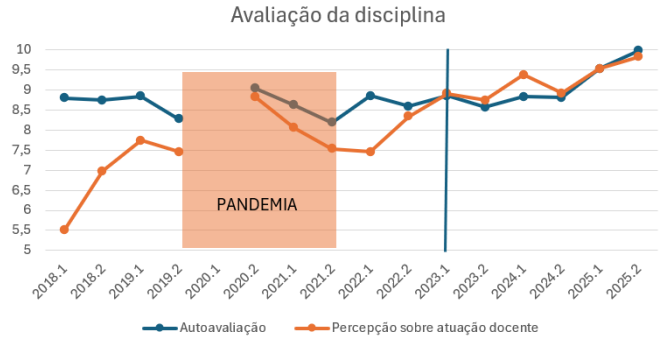


Figura 8: Gráfico de avaliação da disciplina ao longo dos anos. Fonte: autoria própria.

percepção e melhoria se amplia. Nos registros do período anterior ao modelo adotado em 2023.1, predominavam, na análise dos relatórios relatos de sobrecarga e descontentamento didático, evidenciados em apontamentos de que as atividades *"superlotam a agenda do aluno, mas pouco contribuem para a efetividade do ensino"* (2021.1) e críticas de que *"apesar de ter total conhecimento do assunto, o professor demonstrou uma didática pouco efetiva"* (2022.1).

Com as metodologias ativas, as críticas de natureza pedagógica foram superadas. Ocorreram críticas iniciais em relação a ajustes operacionais, para adaptações ao desenvolvimento do projeto. Em 2024 e 2025 as críticas foram substituídas por avaliações espontâneas de reconhecimento, culminando no registro de que a disciplina *"foi a melhor matéria do meu semestre"* (2025.1). Essa mudança no teor dos depoimentos demonstra que, apesar da impossibilidade de comprovação estatística na melhoria do ensino, a mudança de metodologia promoveu um ambiente de aprendizagem mais engajado, participativo e colaborativo.

IV. Conclusões

A experiência apresentada sugere que a utilização desses dois modelos de metodologias ativas (PBL e Aula invertida), contribuíram para aumentar o engajamento, a motivação e satisfação dos alunos durante o curso. As atividades possibilitaram a integração teoria-prática, fazendo interface com a interdisciplinaridade e aumentando o interesse e participação dos alunos.

O desenvolvimento dessas atividades exigiu a aplicação integrada de conhecimentos de diferentes disciplinas, isso gerou novos aprendizados que foram além do escopo inicial da disciplina.

A experiência contribuiu para o desenvolvimento de habilidades e competências importantes para a atuação do engenheiro biomédico, como planejamento, organização, trabalho em equipe, comunicação, pesquisa, resolução de problemas e tomada de decisão. As entregas parciais e os *feedbacks* contínuos também possibilitaram ajustes durante a execução do projeto, tornando o processo de aprendizagem mais reflexivo e colaborativo.

Este relato, em conjunto com a análise qualitativa do engajamento da motivação e satisfação dos alunos, indica que está ocorrendo uma melhora no andamento da disciplina após sua reformulação. Espera-se que esta experiência possa servir como referência para a implementação de estratégias semelhantes em disciplinas da Engenharia e contribuir para futuras discussões sobre inovação no ensino superior principalmente em disciplinas com uma alta carga técnica de física e matemática.

Nota sobre o uso de Inteligência Artificial

Durante a elaboração deste texto, ferramentas de Inteligência Artificial foram utilizadas como apoio à revisão linguística, à organização das ideias, à reestruturação de trechos e ao aprimoramento da clareza textual. O uso dessas ferramentas ocorreu de forma auxiliar, sem substituir a análise crítica, a autoria intelectual e a responsabilidade dos autores pelo conteúdo apresentado. Todas as informações, interpretações e decisões finais foram revisadas, ajustadas e validadas pelos autores, garantindo a coerência acadêmica, a adequação ao objetivo do trabalho e a preservação da originalidade do relato.

Referências

[1] E. M. A. Geurts, R. P. Reijs, H. H. M. Leenders, M. W. J. Jansen e C. J. P. A. Hoebe, "Co-creation and decision-making with students about teaching and learning: A systematic literature review," *Journal of Educational Change*, v. 25, n. 1, pp. 103–125, mar. de 2024, ISSN: 1389-2843, 1573-1812. DOI: 10.1007/s10833-023-09481-x. acesso em 13 de jul. de 2026. endereço: https://link.springer.com/10.1007/s10833-023-09481-x.

[2] R. Z Hauzel, T. Pattnaik, R. Vara e S. Prabha Mandela, "Investigating factors contributing to student disengagement and ownership in learning: A case study of undergraduate engineering students," *Journal of Information Technology Education: Innovations in Practice*, v. 23, p. 008, 2024, ISSN: 2165-3151, 2165-316X. DOI: 10.28945/5336. acesso em 13 de jul. de 2026. endereço: https://www.informingscience.org/Publications/5336.

[3] BRASIL. Ministério da Educação. Conselho Nacional de Educação. Câmara de Educação Superior, *Resolução CNE/CES nº 2, de 24 de abril de 2019*, Institui as Diretrizes Curriculares Nacionais do Curso de Graduação em Engenharia, Brasília, DF, 26 de abr. de 2019. acesso em 18 de jul. de 2026. endereço: https://www.gov.br/mec/pt-br/cne/pdf/normas-classificadas-por-assunto/diretrizes-curriculares-cursos-de-graduacao/rces002_19.pdf.

[4] C. M. Ribeiro, L. S. T. Alarcão, K. M. Ribeiro, H. S. Tavares e Z. P. d. Anjos, "REINVENTANDO O APRENDER: QUANDO O ALUNO VIRA PROTAGONISTA," *Missioneira*, v. 27, n. 9, pp. 111–123, 6 de set. de 2025, ISSN: 1518-0263. DOI: 10.46550/zgv26457. acesso em 13 de jul. de 2026. endereço: https://cemipa.com.br/revistas/index.php/missioneira/article/view/297.

[5] B. Gonçalves, A. Tomás e J. C. Jota, "Os desafios da educação em Moçambique na era digital: do modelo da memorização ao desenvolvimento da autonomia do aluno," *Revista EDaPECI*, v. 24, n. 1, pp. 106–119, 15 de abr. de 2024, ISSN: 2176-171X. DOI: 10.29276/redapeci.2024.24.119939.106-119. acesso em 13 de jul. de 2026. endereço: https://periodicos.ufs.br/edapeci/article/view/19939.

[6] T. d. A. Rodrigues, L. R. R. Lima, A. V. Araújo e A. d. M. Santos, "Análise da utilização de metodologia ativa para o ensino universitário em gestão de projetos," *Caderno Pedagógico*, v. 21, n. 13, e11779–e11779, 9 de dez. de 2024, ISSN: 1983-0882. DOI: 10.54033/cadpedv21n13-152. acesso em 13 de jul. de 2026. endereço: https://ojs.studiespublicacoes.com.br/ojs/index.php/cadped/article/view/11779.

[7] A. C. Gerhard e J. F. B. d. Rocha, "A FRAGMENTAÇÃO DOS SABERES NA EDUCAÇÃO CIENTÍFICA ESCOLAR NA PERCEPÇÃO DE PROFESSORES DE UMA ESCOLA DE ENSINO MÉDIO," *Investigações em Ensino de Ciências*, v. 17, n. 1, pp. 125–145, 2012, ISSN: 1518-8795. acesso em 13 de jul. de 2026. endereço: https://ienci.if.ufrgs.br/index.php/ienci/article/view/210.

[8] K. F. Antunes, S. G. d. Silva e G. J. Mello, "PROJETO INTEGRADOR INTERAÇÃO: propostas e desafios de interdisciplinaridade no ensino médio integrado do IFMT Câmpus Várzea Grande," *Revista Educação, Cultura e Sociedade*, v. 14, n. 2, pp. 7–18, 27 de dez. de 2024, ISSN: 2237-1648. DOI: 10.30681/ecs.v14i2.12928. acesso em 13 de jul. de 2026. endereço: https://periodicos.unemat.br/recs/article/view/12928.

[9] E. A. d. Jesus, A. d. L. e. R. Guerra e A. R. G. Pereira, "A INTERDISCIPLINARIDADE COMO ESTRATÉGIA PARA o DESENVOLVIMENTO DE UMA APRENDIZAGEM SIGNIFICATIVA," *International Contemporary Management Review*, v. 5, n. 2, e87–e87, 9 de mai. de 2024, ISSN: 2595-0428. DOI: 10.54033/icmrv5n2-003. acesso em 13 de jul. de 2026. endereço: https://icmreview.com/icmr/article/view/87.

[10] A. Kolmos, J. E. Holgaard, H. W. Routhe, M. Winther e L. Bertel, "Interdisciplinary project types in engineering education," *European Journal of Engineering Education*, v. 49, n. 2, pp. 257–282, 3 de mar. de 2024, ISSN: 0304-3797, 1469-5898. DOI: 10.1080/03043797.2023.2267476. acesso em 13 de jul. de 2026. endereço: https://www.tandfonline.com/doi/full/10.1080/03043797.2023.2267476.

Cognitive Load Assessment in Second Language Academic Instruction Using a Low-Cost BCI

Kevin Mateo Gutiérrez-Ortiz
Robotics and Assistive Technology Laboratory
Federal University of Espírito Santo (UFES)
Vitória, Brazil
kevin.ortiz@edu.ufes.br

Luis Carlos Rodríguez-Timaná
Faculty of Engineering / Rob. and Assist. Tech. Lab
Santiago de Cali Univ. (USC); Federal University of Espírito Santo
Cali, Colombia; Vitória, Brazil
luis.rodriguez11@usc.edu.co

Eliete Maria de Oliveira Caldeira
Robotics and Assistive Technology Laboratory
Federal University of Espírito Santo (UFES)
Vitória, Brazil
eliete.caldeira@ufes.br

Teodiano Freire Bastos-Filho
Robotics and Assistive Technology Laboratory
Federal University of Espírito Santo (UFES)
Vitória, Brazil
teodiano.bastos@ufes.br

***Abstract*—The objective assessment of cognitive load in educational environments remains a methodological challenge due to high inter-individual variability. This work presents a pilot study functioning as a proof of concept to evaluate the preliminary technical feasibility of using a Brain-Computer Interface (BCI) for real-time monitoring during audiovisual instruction in a second language. An Emotiv EPOC X device and a Python-based architecture were used to record the electroencephalographic (EEG) signals of nine participants in an ecologically realistic laboratory environment. To mitigate ocular artifacts, the neurophysiological Frontal Theta / Parieto-Occipital Alpha index was implemented. The analysis using robust statistical estimators revealed significant inter-subject variability and a significant correlation (Spearman rank correlation coefficient = 0.85, p-value = 0.006) between the baseline and the task Theta/Alpha index, suggesting possible limitations of relying on static group metrics. The system integrated personalized dynamic thresholds, identifying candidate episodes of increased cognitive demand in an average of 17.75% of the instruction time. These threshold exceedances should be interpreted as preliminary evidence of individualized monitoring feasibility, not as definitive cognitive overload. The results suggest that individualized calibrations provide a methodological advantage in educational BCIs and expose the potential susceptibility of short baselines to anticipatory anxiety, emphasizing the need for behavioral ground-truth validations in future educational neurotechnology applications.**

***Index Terms*—Cognitive Load, EEG, BCI, Emotiv EPOC X, Second Language, Neuroeducation, Dynamic Threshold.**

I. Introduction

Learning novel concepts in a foreign language imposes a high extrinsic and intrinsic load on a person's working memory [1]. The difficulty in real-time monitoring of when a student reaches their cognitive limit often results in a decrease in academic performance. Traditionally, the evaluation of this effort has been performed using deferred questionnaires, such as the NASA-TLX [2]. However, advances in neurotechnology make the exploration of continuous mental state measurement viable [3].

Electroencephalographic (EEG) signals offer direct access to attentional states [12], [13]. Several studies [14], [15] have used the frontal Beta/Alpha ratio in EEG signals; however, such an approach is susceptible to contamination by muscular (electromyography-EMG) and eye (electrooculography-EOG) artifacts, especially eye blinks [4]. Current literature suggests that verbal working memory demand is more robustly reflected in the ratio between Theta (4–8 Hz) in the frontal, and Alpha (8–12 Hz) in parieto-occipital areas [5]–[7].

The main methodological contribution of this work is to present a proof of concept for the implementation and critical analysis of a neurofeedback platform developed using the Emotiv EPOC X device [8]. Unlike approaches based on static universal metrics, this study explicitly evaluates the technical advantage of individualized dynamic thresholds to monitor transient episodes of increased cognitive demand in an ecologically valid environment [9]. By proposing robust statistical estimators over traditional mean-based thresholds, it aims to better account for inter-individual variability and identify methodological boundaries regarding physiological interpretations when behavioral ground truth is absent.

II. Methodology and System Design

The system is based on a modular Front-end and Back-end architecture for the acquisition and digital processing of EEG (Fig. 1).

A. Theta/Alpha Index Processing

EEG acquisition was performed using the Emotiv EPOC X with a sampling frequency of 128 Hz. The device features 14 EEG channels positioned according to the international 10-20 system. The Band Power stream is calculated internally through Fast Fourier Transform (FFT) applying a 2 s Hanning window. To isolate cognitive effort from muscular noise, the pipeline extracts spectral power as follows:

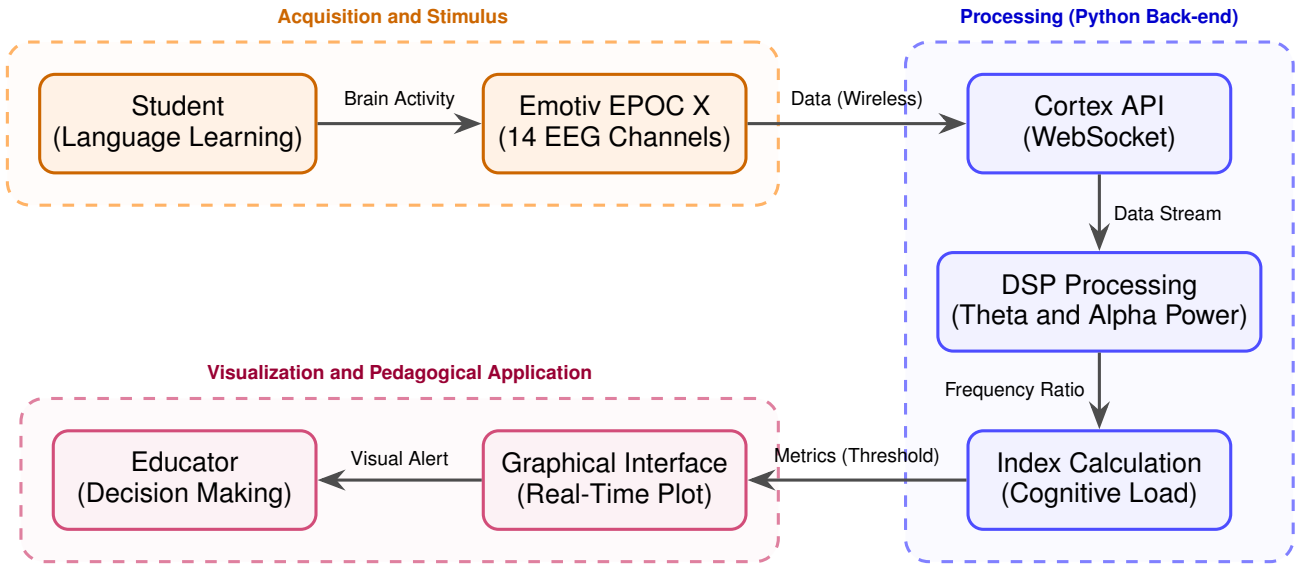


Fig. 1. Block diagram of the cognitive load assessment system architecture.

1) **Frontal Load (Theta):** The spectral band in frontal channels (AF3, F3, F4, AF4) is averaged, theoretically associated with mental effort.
2) **Posterior Processing (Alpha):** Activity in parieto-occipital channels (P7, O1, O2, P8) is averaged, linked to working memory.
3) **Threshold Calculation:** Given the non-normal distribution of the biometric data, the system established an adaptive threshold based on robust estimators (Median$+$ $1.5 \times$ MAD), where MAD represents the Median Absolute Deviation, calculated during the first 60 s of the user's baseline. The factor 1.5 was selected as a heuristic to define individualized exceedances; it is not intended as a formal statistical cutoff.

For each subject, the instantaneous Theta/Alpha ratio was smoothed with a 16-sample moving average. Values exceeding 1.5 interquartile ranges (IQR) beyond that subject's own first or third quartile were then clipped to the corresponding outlier boundary (MATLAB `filloutliers`, quartile method), prior to computing that subject's robust baseline and task statistics.

B. Participants and Ethical Considerations

This pilot study included 9 independent university students (19 to 25 years old) without diagnosed neurological disorders. The sample integrated native Spanish and Portuguese speakers with an English intermediate-advanced level (B1 or higher) as a second language, ensuring comprehension of the stimulus. Participants were informed about the nature of the test and signed an informed consent form, guaranteeing the confidentiality of their biometric data.

C. Stimulus and Experimental Environment

To induce a double cognitive load (linguistic processing and assimilation of technical concepts), the English video *"ALL OF PHYSICS explained in 14 Minutes"* was used. Prioritizing ecological validity—defined here as an experimental design aiming to better approximate the conditions of a natural learning environment outside of a sterile, soundproof booth [9]—the experiment took place in a shared classroom condition. Each session lasted 16 min: 1 min of basal calibration followed by 15 min of stimulus. During the 60-second baseline, participants remained seated with eyes open and were instructed to relax and prepare to watch the video. Fig. 2 presents the average temporal evolution of the ratio.

D. Statistical Analysis

To address the non-normal distribution typical of EEG signals, non-parametric statistical measures were implemented [11]. The Wilcoxon signed-rank test was used to compare baseline and task medians. Spearman correlation was used to assess the association between baseline and task Theta/Alpha index. Effect sizes were estimated per subject as a robust Cohen's d, defined as the standardized median difference between task and baseline Theta/Alpha values, using a pooled estimate of the median absolute deviation (MAD) from both conditions.

III. Results

The descriptive analysis of the Theta/Alpha index was performed using robust statistical estimators (Fig. 3). The Jarque-Bera test applied to the baseline medians did not follow a normal distribution ($p = 0.018$), justifying the use of the median and MAD. Extreme inter-individual variability was observed, with a basal median interquartile range (IQR) of 6.29.

The comparison of the basal median against the instruction task using the Wilcoxon signed-rank test did not reveal a significant global increase ($W = 17.0$, $p = 0.570$). An approximate retrospective power analysis based on a paired

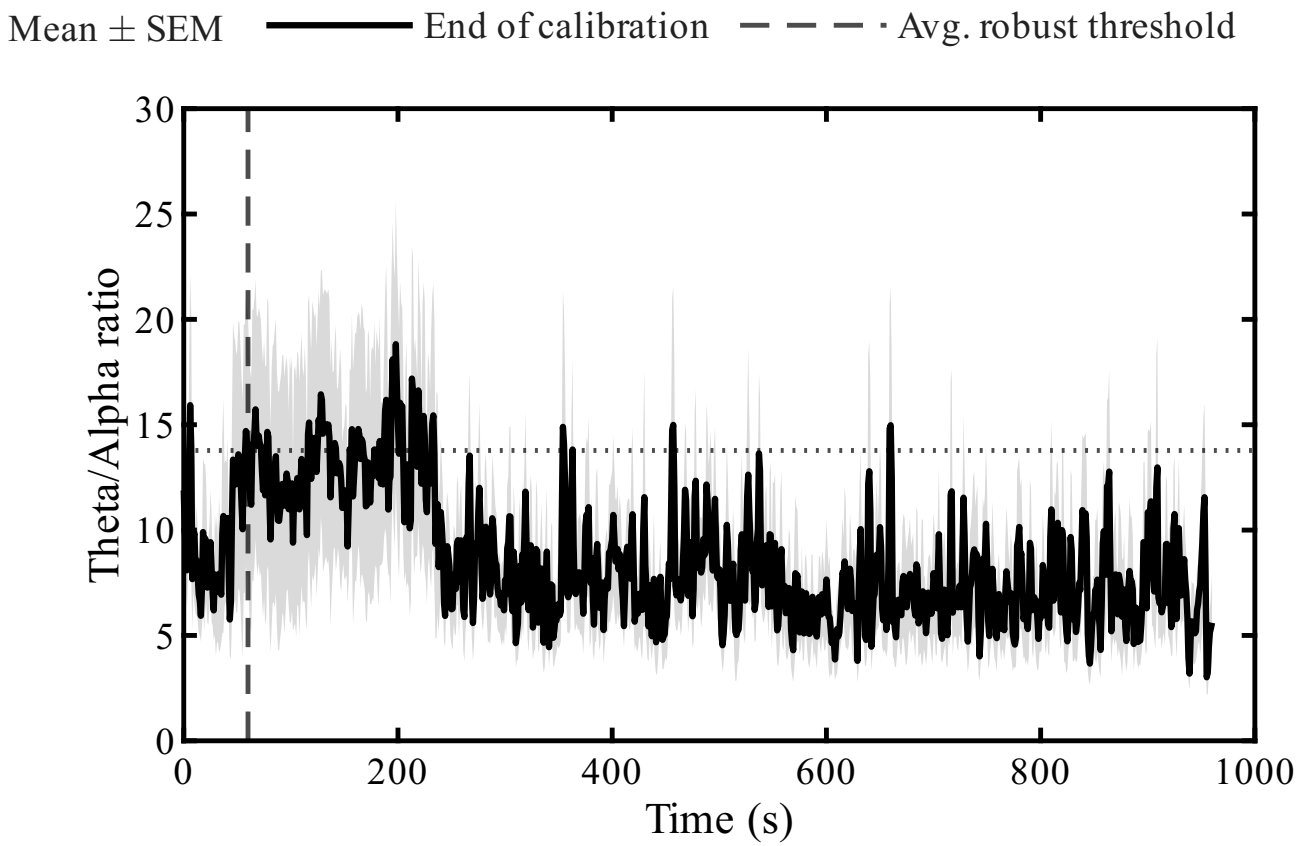


Fig. 2. Average temporal evolution of the Theta/Alpha ratio ($n = 9$). The shaded band represents ± 1 Standard Error of the Mean (SEM). The solid black vertical line indicates the end of calibration; the dashed line shows the average robust threshold.

Table I
INDIVIDUAL METRICS USING ROBUST THRESHOLDS BASED ON THE MEDIAN AND MAD.

Subject	Threshold	Median (Task)	Time > Thresh. (%)	Cohen's d
S01	2.73	1.16	9.6%	-0.95
S02	4.00	1.72	10.6%	-0.74
S03	17.54	9.52	22.1%	0.18
S04	40.77	13.06	18.8%	-1.65
S05	17.92	12.43	21.1%	0.13
S06	13.30	3.18	0.0%	-1.73
S07	8.17	2.65	0.0%	-2.07
S08	6.96	3.70	22.7%	0.11
S09	12.62	13.45	54.8%	1.33

t-test model indicates that, with $n = 9$, this study had approximately 26% power to detect a medium effect ($d = 0.5$) and 56% power to detect a large effect ($d = 0.8$) at $\alpha = 0.05$; a sample of approximately 33 participants would be required to reach 80% power for a medium effect. Although this approximation assumes normality, it provides a useful reference for interpreting the null result. The non-significant Wilcoxon result should therefore be interpreted as inconclusive rather than as evidence of no change between baseline and task. Despite the overall stability of the median, the system detected that users crossed their personalized dynamic threshold in an average of 17.75% of the instruction time (Table I).

A key statistical finding was the Spearman correlation test, demonstrating a strong positive association between the basal median and the task median ($\rho = 0.85$, $p = 0.006$), as shown in Fig. 4. This correlation supports the individual stability of the index across the transition from rest to cognitive effort, although it may alternatively indicate that the index is influenced by trait-like individual factors (e.g., baseline cortical activity, electrode impedance) rather than purely reflecting task-induced cognitive state changes.

IV. DISCUSSION

This study demonstrated the preliminary technical feasibility of using dynamic thresholds in environments with ecological validity, providing important methodological insights for computational neuroeducation.

It is important to distinguish the constructs potentially reflected in the Theta/Alpha index. Frontal theta has been associated with working memory engagement and cognitive effort, while parieto-occipital alpha is more closely linked to visual attentional allocation [5], [6]. Neither directly indexes learning outcomes, and both are also sensitive to arousal and affective state, including anxiety. Consequently, an increase in the Theta/Alpha ratio observed in this study is best interpreted as a marker of increased neural engagement or spectral vari-

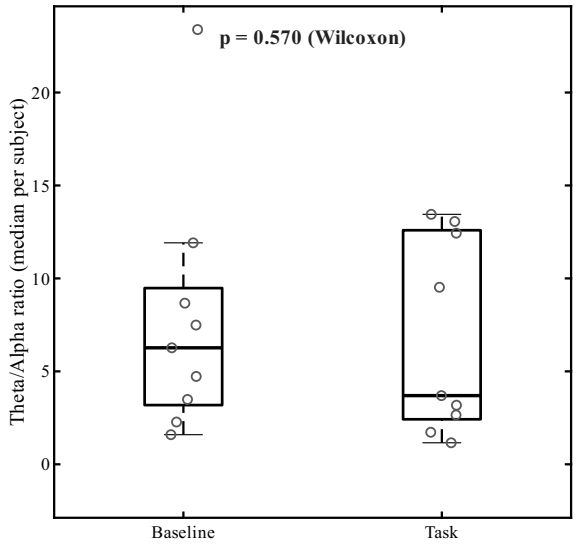


Fig. 3. Comparison of the Theta/Alpha ratio median distribution per subject in the baseline and during the task ($p = 0.570$, Wilcoxon).

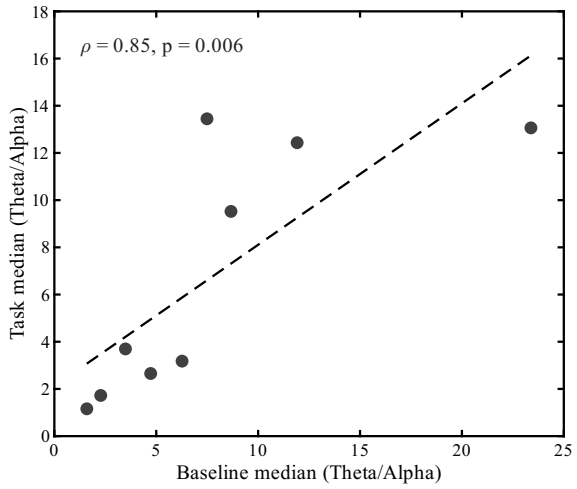


Fig. 4. Scatter plot of baseline versus task Theta/Alpha medians. The dashed line represents a linear least-squares fit for visual reference; the reported correlation is Spearman's $\rho = 0.85$, $p = 0.006$.

ation, which may reflect cognitive effort, attentional shifts, or anticipatory stress — constructs that are not distinguishable using this single index alone.

The absence of a statistically significant increase in the group's cognitive load index ($p = 0.570$, though likely underpowered as noted above) and the descriptively observed reduction of task medians compared to baselines raise the plausible hypothesis of an anticipatory anxiety effect. Anticipatory anxiety, linked to stress regulation when an individual expects an impending evaluation [10], might artificially elevate baseline activity. Upon starting the video, students may have transitioned into a more stable attentional state. While this aligns with literature suggesting pre-task EEG states can be contaminated by expectation stress, it remains a hypothesis requiring validation through subjective scales or heart rate monitoring. To minimize this effect in future protocols, strategies such as a 3-to-5 minute acclimatization period prior to baseline recording should be implemented.

The strong basal-task correlation ($\rho = 0.85$) supports the core premise of this work: marked inter-individual variability suggests caution when imposing universal "group thresholds", although this study does not directly compare individualized versus group-threshold performance. Because an individual's resting state heavily influences their operational baseline, the dynamic parameterization (Median $+ 1.5 \times$ MAD) emerges as a plausible engineering approach to capture transient variations (average 17.75% time above threshold) during complex academic stimuli. Exceeding this statistical threshold indicates spectral variation, though it does not automatically equate to cognitive overload.

It is imperative to acknowledge the limitations of this pilot study. First, the small sample size ($n = 9$) restricts statistical generalization. Second, the use of a single experimental activity limits the extrapolation of the results to other pedagogical formats. Third, reliance on Band Power values provided by the commercial API acts as a "black box", severely limiting control over fundamental preprocessing steps such as frequency filtering and advanced artifact rejection techniques like Independent Component Analysis (ICA). Finally, the absence of behavioral measures of learning or subjective questionnaires (e.g., NASA-TLX) prevents unequivocally linking EEG peaks with actual mental saturation.

Sensitivity analyses across baseline durations and test-retest reliability could not be conducted in this pilot, as each participant completed a single session with one fixed 60-second baseline; these analyses require a dedicated multi-session protocol and are identified as priority next steps (Section V).

Lessons Learned and Research Gaps

This pilot study highlights that ecological BCI protocols must account for pre-task psychological stressors, as short immediate baselines are easily corrupted by anticipatory anxiety. A critical lesson learned is that static, universal thresholds may be poorly suited to educational settings due to extreme inter-subject variability. Consequently, a significant research gap exists in standardizing dynamic, individualized baselines. Future research must bridge this gap by combining these dynamic EEG thresholds with behavioral learning metrics and raw signal processing to ensure accurate pedagogical monitoring.

V. Conclusion and Future Works

This study demonstrated the preliminary feasibility of a proof-of-concept system for individualized EEG monitoring in an ecologically valid educational environment. The robust non-parametric statistical threshold (Median$+1.5\times$MAD) provided a methodological alternative for individualized monitoring under high population variability. This preliminary evidence suggests potential advantages over static group metrics, but direct comparative validation is still required. The results suggest that ecological protocols utilizing short 60-second baselines

are susceptible to anticipatory anxiety biases, indicating that the true pedagogical utility of this marker requires deeper, multi-modal investigation.

As future work, the statistical sample must be expanded to include behavioral validations of acquired knowledge and subjective cognitive load surveys to establish a reliable ground truth. Additionally, sensitivity analyses across different baseline durations and threshold definitions will be conducted to optimize the calibration period. Regarding signal processing, a critical transition will be made to extracting raw EEG data to apply ICA algorithms, optimizing the BCI's diagnostic precision and artifact rejection capabilities for its eventual integration into intelligent tutoring systems.

REFERENCES

[1] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[2] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in Psychology*, vol. 52, pp. 139–183, 1988.

[3] H. Zeng, X. Huang, Y. Liu, and X. Gu, "Exploring neural evidence of attention in classroom environments: A scoping review," *Brain Sciences*, vol. 15, no. 8, 2025, Art. no. 860. doi: 10.3390/brainsci15080860.

[4] M. Fatourechi *et al.*, "EMG and EOG artifacts in brain computer interface systems: A survey," *Clinical Neurophysiology*, vol. 118, no. 3, pp. 480–494, 2007.

[5] B. Raufi and L. Longo, "An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload," *Frontiers in Neuroinformatics*, vol. 16, 2022, Art. no. 861967. doi: 10.3389/fninf.2022.861967.

[6] W. Zhou and X. Wu, "The impact of internal-generated contextual clues on EFL vocabulary learning: insights from EEG," *Frontiers in Psychology*, vol. 15, 2024.

[7] F. Bueno-Palomeque, E. Lema-Condo, S. Castro-Villalobos, L. Serpa-Andrade, and J. Ordoñez-Morales, "Variation of the EEG-energy in a second language class," in *Proc. NEUROTECHNIX 2018*, 2018, pp. 110–116.

[8] A. Nirabi, N. Abd Rahman, M. H. Habaebi, K. A. Sidek, and S. Yusoff, "Cognitive load assessment through EEG: A dataset from arithmetic and Stroop tasks," *Data in Brief*, vol. 13, 2025, Art. no. 111477. doi: 10.1016/j.dib.2025.111477.

[9] M. L. R. Menezes and R. M. Vicari, "EduEmotion: An EEG Dataset of Emotional Responses in Simulated Learning Tasks," in *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, 2025.

[10] Z. Qiao, E. S. Poppelaars, and X. Li, "In the anticipation of threat: Neural regulatory activity indicated by delta-beta correlation and its relation to anxiety," *Biological Psychology*, vol. 187, 2024, Art. no. 108769.

[11] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177–190, 2007.

[12] C. Berka *et al.*, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, Space, and Environmental Medicine*, vol. 78, no. 5, pp. B231–B244, 2007.

[13] P. Aricò *et al.*, "Passive BCI beyond the lab: current trends and future directions," *Physiological Measurement*, vol. 39, no. 8, 2018, Art. no. 08TR02.

[14] A. T. Pope, E. H. Bogart, and D. S. Bartolome, "Biocybernetic system evaluates indices of operator engagement in automated task," *Biological Psychology*, vol. 40, no. 1-2, pp. 187–195, 1995.

[15] A. T. Kamzanova, A. M. Kustubayeva, and G. Matthews, "Use of EEG workload indices for diagnostic monitoring of vigilance decrement," *Human Factors*, vol. 56, no. 6, pp. 1136–1149, 2014.

Design, Construction and Experimental Testing of a Low-Cost Horizontal Tensile Testing Machine for Mechanical Characterization of Materials

1st Guilherme Souza Emergente
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
guilherme.emergente@ges.inatel.br

2nd Laís Carlos Rodrigues
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
lais.rodrigues@ges.inatel.br

3rd Alexandre Baratella Lugli
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
baratella@inatel.br

4th Raphael Cardoso Mota Pereira
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
raphaelpereira@inatel.br

5th Maysa Costa Alves
National Institute of Telecommunications (Inatel)
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

***Abstract*—Tensile testing machines are used to characterize polymers, membranes and scaffolds relevant to Biomedical Engineering, but commercial equipment is expensive and often unavailable in teaching laboratories. This paper reports the design, construction and experimental testing of a low-cost horizontal tensile testing machine assembled from 3D-printed parts, an aluminum structural profile and off-the-shelf electronics. The machine uses a NEMA 23 stepper motor coupled to a TR12 trapezoidal lead screw of 8 mm lead, an S-type load cell with an HX711 24-bit amplifier, and an Arduino Mega 2560 controlled from a PyQt6 interface that records force and crosshead displacement and computes stress, strain and the elastic modulus from the specimen geometry. The mechanical and electronic architecture, the open bill of materials, the published CAD models and source code, and a costed build of approximately R$ 3,702 (2026 prices) are presented. Three specimens whose geometry was defined experimentally — a rigid PVC strip, an HDPE strip and an elastomeric ring — were tested to failure with a 40 mm free length between grips, and the recorded curves reproduce the behaviour expected for a brittle polymer, for a semi-crystalline polymer with yielding and cold drawing, and for a hyperelastic material. The computed moduli span roughly four orders of magnitude across the three materials and place them in the correct relative order, showing that stiffness, strength, deformability and toughness are distinct properties and that the same acquisition chain resolves material classes that differ widely in stiffness.**

***Index Terms*—Tensile testing, low-cost instrumentation, open hardware, stepper motor, load cell, arduino, stress-strain curve, 3D printing, engineering education.**

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) through a research scholarship, and by the National Institute of Telecommunications (Inatel).

I. Introduction

Mechanical characterization requires instrumentation capable of relating an applied force to the resulting deformation of a specimen. Tensile testing machines produce the stress–strain curve, from which the elastic modulus, the yield strength and the tensile strength are obtained [1]. In Biomedical Engineering this is relevant for polymeric materials used in devices, implants, membranes and scaffolds for tissue engineering, whose mechanical behaviour affects clinical performance.

A. Problem and motivation

Professional tensile testing machines are expensive for two main reasons. Most commercial models are universal machines able to perform tensile, compression, bending and fatigue tests, which increases complexity; and industrial-grade frames must be stiff enough that their own deformation does not corrupt the measurement, which requires costly components. This equipment is therefore frequently out of reach for teaching institutions and small laboratories, where contact with mechanical testing is limited by availability. A low-cost machine lets students observe stress, strain, the elastic regime, yielding and rupture directly; the present work targets low-strength materials, mainly polymers.

B. Related low-cost initiatives

Several groups have published simplified prototypes. An Arduino-based uniaxial tensile tester built for under one hundred dollars has been reported for teaching mechanical properties and characterizing soft materials, including biomedical applications [2]. Educational tensile testing machine projects are also documented in Brazilian teaching material [3]–[5].



Fig. 1. Instron 3400 Series dual-column universal testing system, a currently marketed commercial machine of the type discussed in this section, with capacities from 500 N to 300 kN.
Source: Instron [10].

As the main constructive reference, this project adopted an extensively documented open-source horizontal model [6].

C. Contributions

The contributions of this work are: (1) a reproducible horizontal tensile testing machine architecture, with an open bill of materials, published CAD models [7] and published firmware and interface source code [8]; (2) a costed build under 2026 Brazilian market conditions, with the reference year and exchange rate stated explicitly; (3) tensile tests on three polymer specimens, recorded on video [9], with the resulting curves and computed quantities; and (4) an analysis of those results relating the observed behaviour of each material to the mechanical properties the stress–strain curve is meant to expose.

II. Related Work

A. Commercial testing machines

Commercial tensile testing machines from manufacturers such as Instron (Fig. 1) [10], Shimadzu and EMIC are used in industrial and research laboratories because they provide verified force accuracy, repeatability and compliance with testing standards, such as ASTM D638 for plastics [11] and ASTM E8/E8M for metals [12]. Their force-measuring systems are periodically verified according to ISO 7500-1 [13], and strain is normally measured with a class-rated extensometer [14] rather than inferred from crosshead motion. In the Brazilian market, prices range from about R$ 4,300 for manual benchtop models to more than R$ 40,000 for automated universal machines [15].

B. Low-cost models studied

Two low-cost models were analysed during development: a vertical prototype previously available in the institution's laboratory (Fig. 2), and the open-source horizontal model of [6] (Fig. 3).



Fig. 2. Vertical testing machine model previously available in the laboratory, used as a case study.
Source: The authors (2025).

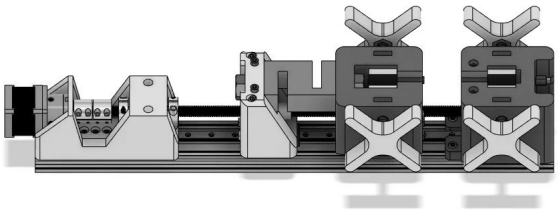
1) Previous vertical model — case study: The first model was used only as a source of design problems to avoid. Two vertical wooden pillars carried a movable upper grip connecting the specimen to the load cell, actuated by a conventional (non-stepper) motor with belt transmission, driven from an Arduino Uno through a relay. A conventional motor rotates continuously, so position and speed cannot be commanded and displacement control would depend entirely on external sensing; the wooden frame has low axial stiffness compared to metal profiles; and the vertical orientation makes alignment between the load axis and the cell harder to guarantee. These limitations motivated the move to a stepper-driven horizontal architecture.

2) Reference horizontal model — constructive basis: The second model [6] was adopted as the direct constructive basis. It shares the mechanical architecture used here: a carriage moving along a linear guide, driven by a lead screw and nut rotated by a stepper motor. In its original form it uses a NEMA 17 stepper with a planetary gearbox and a 500 kg–1 t S-type load cell, and two microcontrollers — one for machine control and an ATtiny85 dedicated to load cell acquisition. The adaptations made here are summarized in Table I: a direct-drive NEMA 23 stepper (3 N m), a TR12 trapezoidal lead screw of 8 mm lead, control centralized in a single Arduino Mega 2560 with a TB6600 driver, and a computer graphical interface written from scratch, replacing the stand-alone button-and-display operation of the original.

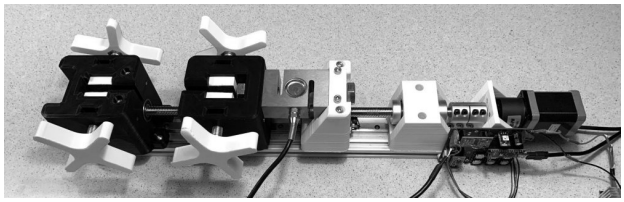
III. System Design

A. Mechanical design

The machine is built on a structural aluminum profile ($20 \times 60\,\mathrm{mm}$, channel 6) that acts as base and structural reference. Linear displacement is guided by a 15 mm linear guide with two carriages. One end holds the stationary grip; the opposite end holds the moving grip. Both grips are 3D-printed jaws tightened by wing screws, which allows the clamping force to be adjusted and specimens of different thicknesses to be accommodated. An S-type load cell is mounted between the moving grip and the carriage, through the moving load cell mount, so that the cell travels with the moving grip and the



(a) CAD model.



(b) assembled machine.

Fig. 3. Open-source horizontal model adopted as the constructive basis. Source: Adapted from [6].

TABLE I
COMPARISON BETWEEN THE STUDIED MODELS AND THE ADOPTED DESIGN

Aspect	**Vertical (previous)**	**Horizontal (ref.)**	**This work**
Orientation	Vertical	Horizontal	Horizontal
Structure	Wood	3D printing	Aluminum profile + 3D printing
Actuator	Conventional motor	Stepper (NEMA 17 + gearbox)	Stepper (NEMA 23, direct drive)
Lead screw	Toothed belt	Lead screw	TR12, 8 mm lead
Position command	Open loop	Step-metered	Step-metered
Force acquisition	Load cell + Arduino	Load cell + ATtiny85	S-type cell + HX711
Operation	Buttons	Buttons + display	Computer interface

Source: The authors (2026).

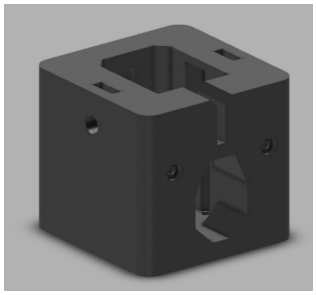
whole load path (stationary grip, specimen, moving grip, load cell, carriage) passes through it.

Actuation is provided by a stepper motor at the end of the rail, whose shaft is connected through a rigid coupling to a TR12 trapezoidal lead screw with a *lead* of 8 mm per revolution. The lead, not the thread pitch, sets the displacement per revolution in (2). The screw spans the full length of the frame and passes through the moving assembly, so that the load path stays coaxial with the rail over the whole stroke. A nut mounted on the carriage converts screw rotation into linear travel of the moving grip.

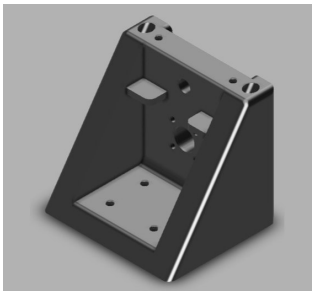
The maximum axial force the actuator can produce follows from the screw relation

$$F_{\max} \approx \frac{2\pi\,\eta\,T}{L_s}, \tag{1}$$

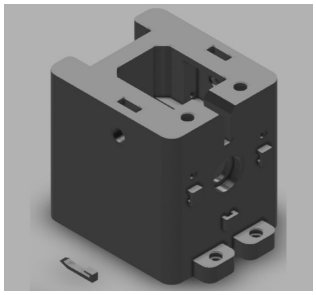
where T is the motor torque, L_s the screw lead and η the screw



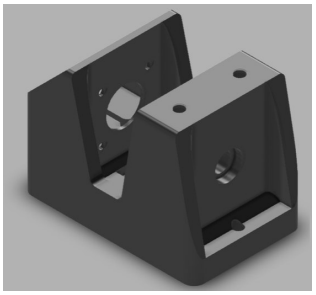
(a) moving grip.



(b) moving load cell mount.



(c) stationary grip.



(d) stepper motor mount.

Fig. 4. Main parts designed in CAD and 3D printed. Source: The authors (2026).

efficiency. With $T = 3\,\mathrm{N\,m}$ and $L_s = 8\,\mathrm{mm}$, this gives about 707 N for a typical unlubricated trapezoidal screw ($\eta \approx 0.3$). Reducing the lead relative to the reference design increases the force available at a given torque and halves the displacement per step, at the cost of a lower maximum crosshead speed. Three protections are implemented: a software force limit, a maximum-displacement stop, and a physical emergency button.

1) 3D-printed components: A large part of the functional structure consists of parts designed in Autodesk Fusion 360 and manufactured by fused-filament 3D printing, which lowers cost and makes reproduction easier. Fig. 4 shows the main parts: the moving grip, which holds the moving end of the specimen; the moving load cell mount, which connects the cell to the carriage; the stationary grip, which holds the fixed end; and the stepper motor mount, which fixes the motor and aligns its shaft with the screw. A grip block completes the jaw assembly. All CAD models are publicly available [7].

B. Control and acquisition system

The control system has three loosely coupled layers: the graphical interface (frontend) and the orchestration logic (backend), both on the computer, and the firmware in the microcontroller, linked by a USB serial connection at 9600 bit/s. Widgets expose Qt signals and contain no hardware logic; the main window connects those signals to the controller, which is the single coupling point between frontend and backend; the controller translates them into commands for the serial bridge, which converts command enumerations into bytes. Fig. 5 shows the layers and the data flow. The repository also includes a mock firmware that emulates the hardware

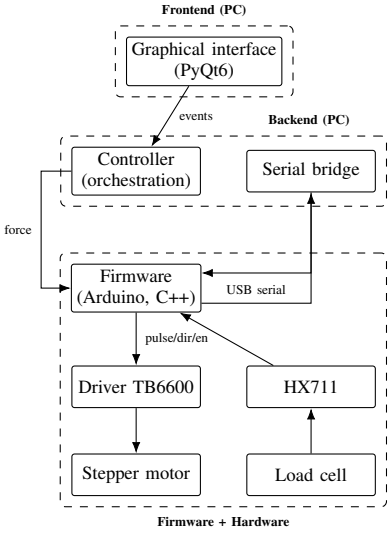


Fig. 5. Block diagram of the control and acquisition system, in three layers, with the interface decoupled from the hardware.
Source: The authors (2026).



Fig. 6. Settings tab of the control interface, showing the manual grip control, the serial connection and the entry of the specimen geometry.
Source: The authors (2026).

responses, allowing the interface to be exercised without the machine, and continuous-integration workflows [8].

1) Electrical system: An Arduino Mega 2560 generates the logic control signals, and a TB6600 driver converts them into coil current, isolating the control circuit from the power circuit. Force is measured by an S-type load cell connected to an HX711 module, a 24-bit ADC dedicated to bridge sensors. The main components are: Arduino Mega 2560; TB6600 stepper driver; NEMA 23 stepper motor with step angle of 1.8° (200 full steps per revolution), 3 N m torque, 3 A per phase, 3.9 V rated and 8 mm shaft; HX711 amplifier; S-type load cell; and a 12 V, 5 A supply.

2) Control interface: The interface was written in Python with PyQt6 [8]. It has two tabs. In the settings tab the operator jogs the grips manually, enters the specimen geometry (cross-section type and dimensions, and free length) and configures the stop conditions: maximum load, maximum displacement, or manual stop (Fig. 6).

In the test tab, force versus displacement and stress versus

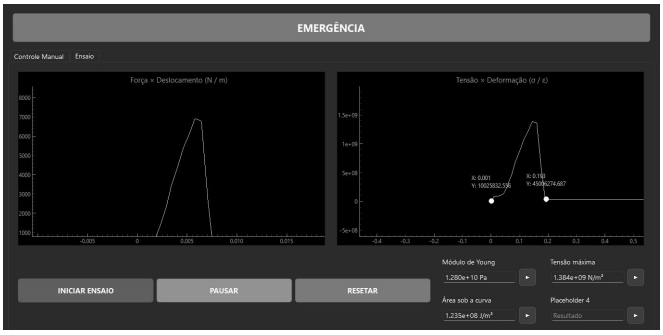


Fig. 7. Test tab of the control interface during an acquisition, with the force–displacement and stress–strain plots and the numerical indicators.
Source: The authors (2026).

strain are plotted in real time, together with numerical indicators for the elastic modulus, the maximum stress and the area under the curve, and the record can be exported. An always-visible emergency control stops the drive (Fig. 7).

3) Firmware: The firmware runs a non-blocking state machine based on `millis()` rather than `delay()`, so that command parsing, step pulse generation and load cell reading proceed concurrently. Its states are idle, up and down (manual grip adjustment), and test, in which the motor advances while the load cell is read every 50 ms and the value is streamed to the computer. Each command travels as a single byte over the serial port. The emergency button is served by a hardware interrupt on a dedicated pin, so the drive stops regardless of the current state and without depending on the main loop.

IV. Cost of the Build

Table II lists the items purchased and printed for this prototype, separated into electromechanical components and manufacturing of the 3D-printed parts. All prices are the amounts paid in the Brazilian domestic market during the first half of 2026; the reference exchange rate was R$ 5.10 per US dollar [16]. The component subtotal is R$ 1,558.32 and the total build cost is R$ 3,701.53.

Two aspects deserve comment. First, the manufacturing line exceeds the whole component subtotal. This is not a filament-material cost: the parts were produced through an outsourced printing service, so the figure includes machine time and labour. A laboratory with its own printer would replace this line by the filament mass alone. Second, the reference project reports about US$ 200 [6], and converting that figure does not reproduce the cost of building the same machine in Brazil, since import duties, freight and retail margins raise component prices and the NEMA 23 motor adopted here is more expensive than the geared NEMA 17 of the original. The build cost is below the entry price of commercial benchtop machines in the same market [15], but that is a comparison of purchase price, not of capability.

TABLE II
PROTOTYPE ITEMS AND COSTS (BRL, 2026 PRICES)

Item	Qty.	Total (R$)
Electromechanical components		
NEMA 23 stepper motor	1	229.00
15 mm linear guide kit + 2 carriages	1	261.90
Aluminum profile 20 × 60 (600 mm)	2	166.04
Arduino Mega 2560 board + cable	1	159.90
TR12 trapezoidal lead screw + nut	1	101.46
T-slot nut kit M3/M4/M5	1	65.25
Stainless Allen screw kit M3/M4/M5	1	72.48
Breadboard, 1660 points	1	48.90
Rigid coupling 8 × 12 mm	1	48.00
Ball bearings	3	43.77
Emergency button	1	34.99
Jumper cable kit	1	31.47
Step-down voltage regulator	2	33.40
Structural M5 screws (100 and 120 mm)	2	103.81
M12 nuts and bolts	2	46.57
M8 wing screw kit	1	23.99
12 V power supply	1	23.99
Kit of 10 adapters for 12 V supply	1	20.79
HX711 module	1	19.61
Stepper driver	1	23.00
Component subtotal		**1,558.32**
Manufacturing		
3D-printed parts (outsourced printing service)	–	2,143.21
Total		**3,701.53**

Source: The authors (2026).

V. DATA REDUCTION

A. Commanded displacement resolution

The displacement commanded per microstep follows from the screw lead, the motor step count and the driver microstepping factor:

$$\Delta x = \frac{L_s}{N_{\text{steps}} \times \mu}, \tag{2}$$

where $L_s = 8\,\text{mm}$ is the lead, $N_{\text{steps}} = 200$ the full steps per revolution and μ the microstepping factor. With $\mu = 16$ the commanded increment is 2.5 µm. This is a commanded increment rather than a measured positioning accuracy, since backlash in the screw–nut pair and any lost steps are not captured by (2).

B. Stress, strain and modulus

Engineering stress is computed from the measured force F and the initial cross-sectional area A_0 [1], and engineering strain from the crosshead travel Δx and the free length L_0:

$$\sigma = \frac{F}{A_0}, \qquad \varepsilon = \frac{\Delta x}{L_0}. \tag{3}$$

Because Δx is taken at the crosshead, it also contains the elastic deflection of the frame, of the grips and of the printed parts, and any slip of the specimen in the jaws. Writing C_m for the machine compliance and δ_s for the slip, the specimen strain is

$$\varepsilon_{\text{sp}} = \frac{\Delta x - C_m F - \delta_s}{L_0} \;\leq\; \varepsilon. \tag{4}$$

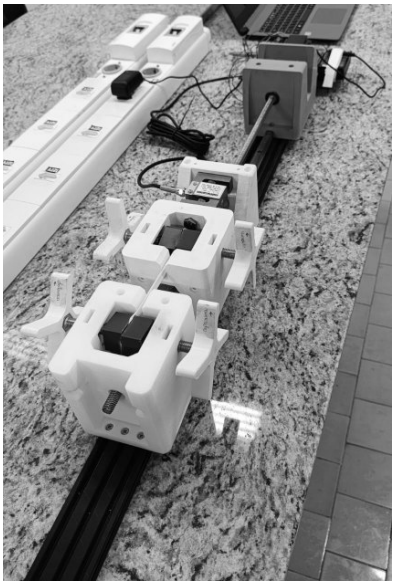


Fig. 8. Assembled prototype during a test, with an elastomeric ring mounted between the 3D-printed grips. From front to back: stationary grip, moving grip, S-type load cell, trapezoidal lead screw and NEMA 23 stepper motor on the aluminum profile.
Source: The authors (2026).

The elastic modulus is the slope of the initial linear portion of the stress–strain curve, obtained by a least-squares fit over a declared strain interval $[\varepsilon_1, \varepsilon_2]$ rather than from a single-point ratio:

$$E = \frac{\sum_i (\varepsilon_i - \bar{\varepsilon})(\sigma_i - \bar{\sigma})}{\sum_i (\varepsilon_i - \bar{\varepsilon})^2}, \qquad \varepsilon_i \in [\varepsilon_1, \varepsilon_2]. \tag{5}$$

The indicator labelled area under the curve is the trapezoidal integral $u = \int_0^{\varepsilon_f} \sigma \, d\varepsilon$, in energy per unit volume.

VI. EXPERIMENTAL TESTS

A. Test setup

All tests reported below were carried out on the assembled prototype shown in Fig. 8, in which the parts described in Sections III-A and III-B are integrated into a working machine. The specimen is clamped between the stationary grip, at the front, which is fixed to the aluminum profile, and the moving grip, which carries the S-type load cell and is driven along the profile by the trapezoidal lead screw and the NEMA 23 stepper motor visible at the back. The load cell therefore travels with the moving grip and remains in series in the load path throughout the test, and the whole assembly is operated from the computer through the interface of Figs. 6 and 7. Placing the specimen horizontally at bench height keeps the grips accessible during mounting and makes the deformation of the specimen directly observable throughout the test, which is a practical advantage for teaching use.

B. Specimens

Three materials were tested: a rigid PVC strip, a high-density polyethylene (HDPE) strip and a commercial elastomeric ring (Fig. 9). The geometry and the dimensions of



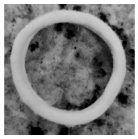
(a) rigid PVC strip.



(b) HDPE strip, before testing.



(c) HDPE strip, after testing.



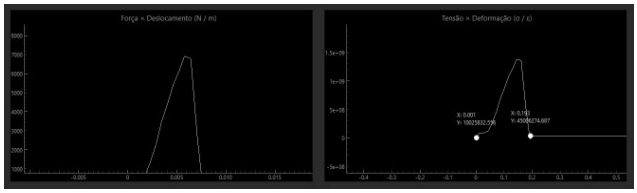
(d) elastomeric ring.

Fig. 9. Specimens used in the tests. The drawn and necked region produced by cold drawing is visible along the whole gauge length in (c), which should be compared with the undeformed strip in (b).
Source: The authors (2026).

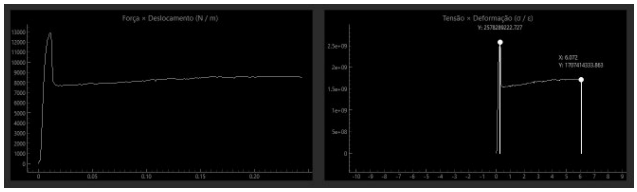
the specimens were defined experimentally, from a series of preliminary tests carried out with the developed machine. Different configurations were evaluated with respect to test stability, specimen clamping, load application capability and the quality of the resulting curves. Based on these tests, dimensions of $170 \times 5 \times 1\,\mathrm{mm}$ were selected for the rectangular specimens, and 40 mm outer diameter with 5 mm width for the circular specimen, since these presented the behaviour best suited to the operating conditions of the equipment. In every test the free length between the grips was set to 40 mm, so that L_0 in (3) is the same for all specimens and the strain axes are directly comparable across materials.

C. Procedure

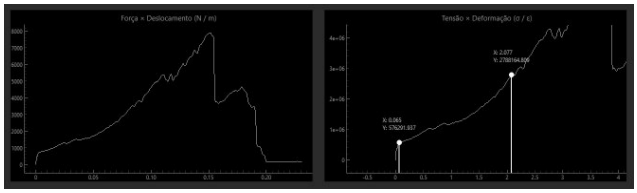
The specimen tests and load cell calibration were performed in triplicate. The specimens were tested under displacement control, with a constant speed applied to the crosshead, until failure or until the previously established displacement limit was reached. The load cell was calibrated using a reference mass, with the system zero set with the grips unloaded before each test. Force was sampled every 50 ms and streamed to the interface, which computed stress, strain, the elastic modulus and the area under the curve in real time from the entered geometry. The complete sequence of the three tests, including specimen mounting, grip adjustment and failure, was recorded and is available as supplementary material [9].



(a) rigid PVC.



(b) HDPE.



(c) elastomer.

Fig. 10. Force versus crosshead displacement (left) and the derived stress–strain curve (right) for the three tests.
Source: The authors (2026).

TABLE III
QUANTITIES COMPUTED BY THE INTERFACE FOR EACH TEST

Quantity	**PVC**	**HDPE**	**Elastomer**
Elastic modulus E (Pa)	1.280×10^{10}	1.482×10^{10}	9.468×10^{5}
Maximum stress (N/m^2)	1.384×10^{9}	2.584×10^{9}	6.289×10^{6}
Area under curve (J/m^3)	1.235×10^{8}	1.004×10^{10}	1.398×10^{7}
Strain at failure (–)	0.19	6.07	≈ 4.2
Behaviour observed	brittle	yield + drawing	hyperelastic

Source: The authors (2026).

D. Results

Fig. 10 shows the interface output for the three tests, and Table III collects the quantities computed by the software at the end of each test.

E. Analysis

1) Rigid PVC: The PVC strip exhibited a steep and approximately linear increase in the curve, indicating predominantly elastic behavior up to failure, which occurred at a strain of 0.19. No clearly defined yield point or formation of a neck or necking was observed during the test. After reaching the maximum load, an abrupt decrease in load occurred, and the two recovered fragments exhibited a clean transverse fracture. This behavior is consistent with that of a predominantly amorphous and rigid polymer, which exhibits limited plastic deformation before failure. The absence of an extensive plastic deformation region is also consistent with the small area under the curve presented in Table III. The PVC absorbed approximately two orders of magnitude less energy per unit volume than HDPE before failure, despite exhibiting a similar calculated modulus.

This result demonstrates that the elastic modulus, associated with material stiffness, is not sufficient to characterize its ability to absorb energy up to failure. Therefore, the test clearly distinguishes stiffness from toughness. While PVC exhibited predominantly elastic behavior, with no clearly defined yield point or neck formation, and relatively abrupt failure, HDPE exhibited yielding, neck formation and propagation, and large deformation before failure. This difference demonstrates one of the main characteristics that the analysis of the stress–strain curve is intended to reveal.

2) HDPE: The HDPE strip exhibited the most characteristic curve among the three materials. Initially, there was a sharp increase in load corresponding to the elastic deformation region until the yield point was reached, identified by the peak of the curve. From this point onward, plastic deformation of the material began, accompanied by the formation of a neck or localized narrowing, characterized by a localized reduction in the strip width. After neck formation, a drop in load occurred, followed by a long, nearly constant portion of the curve that extended to a strain of 6.07, corresponding to more than six times the initial 40 mm free length. During this stage, the neck propagated along the free length while the material continued to elongate. This process, known as cold drawing, occurs with the load remaining approximately constant. The formation and propagation of the neck were also visually observed in the specimen. After the test, the drawn region extended over practically the entire free length (Fig. 9c), leaving the strip visibly thinner and longer than before the test (Fig. 9b). The long deformation region under approximately constant load explains why HDPE exhibited the largest area under the curve among the three materials, approximately two orders of magnitude greater than those of the other materials. This indicates that HDPE was able to absorb a much greater amount of energy before failure. This test was also the most demanding for the machine, as it was necessary to maintain movement over a long distance while the specimen remained under load, without loss of grip at the clamps. Thus, the test demonstrated both the characteristic behavior of a semi-crystalline polymer and the machine's ability to perform tests involving large deformations.

3) Elastomer: The elastomeric ring exhibited low stiffness and a high capacity for elongation. The curve showed a progressive increase in stiffness as the material was elongated, a behavior characteristic of elastomeric materials with a hyperelastic response. Near failure, an irregular decrease in load was observed. The calculated elastic modulus was 9.468×10^5 Pa, representing the lowest value among the three materials evaluated, approximately four orders of magnitude lower than the values obtained for HDPE and PVC. This result highlights the difference in stiffness among the materials: while HDPE and PVC exhibited behavior characteristic of rigid polymers, the elastomer showed greater ease of deformation under an applied load. This result is also important for evaluating the performance of the machine because, using the same acquisition system and changing only the specimen, it was possible to characterize materials with very large differences in stiffness. The difference between the calculated moduli was approximately 13,500-fold, and the machine correctly identified the relative order of stiffness of the materials. Thus, the test clearly demonstrates the difference between a rigid polymer and an elastomer, which is particularly relevant for the educational purpose of the equipment.

Therefore, the results presented in Table III show clear differences among the three materials. PVC and HDPE exhibited similar elastic moduli, indicating comparable stiffness; however, they showed distinct behaviors after the onset of deformation. PVC exhibited low deformation at break (0.19) and abrupt fracture, whereas HDPE exhibited large deformation (6.07), with yielding, neck formation, and stretching, resulting in the highest energy absorption. The elastomer exhibited the lowest elastic modulus, approximately four orders of magnitude lower than those of the rigid polymers, demonstrating its low stiffness and high deformation capacity. Therefore, the results demonstrate that stiffness, strength, deformability, and toughness are distinct properties, and that the elastic modulus alone is not sufficient to describe the mechanical behavior of the materials.

VII. Future Work

Although the machine is operational, the project has considerable room for improvement in its structure, hardware and software.

Gearbox coupled to the motor. Adding a gearbox to the stepper motor would convert part of the motor speed into torque, increasing the available force and reducing the rotational speed, which allows finer control of the crosshead movement.

Replacing the microcontroller. The project currently uses an Arduino Mega, which lacks some resources that would benefit the system. A more suitable microcontroller would be an ESP32, which supports multithreading and can therefore execute two software processes in parallel. This makes it possible to separate motor control from the process responsible for reading the load cell, preventing a malfunction in one from directly affecting the other.

Soldered circuit. A soldered circuit allows better organization of the hardware and removes the risk, inherent to the breadboard, of wires disconnecting accidentally.

Maximum and minimum points within the defined interval. At present the minimum point of the stress–strain plot is not computed, and the maximum point does not respect the interval defined by the user, since the maximum of the complete plot is displayed instead. Making both points respect the user-defined interval would improve the accuracy of the calculations.

Data persistence. A feature to save the data obtained during a test can be implemented, building over time a broad catalogue of records from different tests.

In addition to these improvements, the planned continuation includes calibrating the force chain over 10–100 % of the working range in ascending and descending sequences [13], reporting linearity, hysteresis, zero return and repeatability with

an uncertainty statement [17]; determining the machine compliance and the screw backlash; and comparing ASTM D638 specimens [11] from a single batch against a calibrated universal testing machine, with at least five replicates per condition.

VIII. Conclusion

This paper presented the design, construction and experimental testing of a low-cost horizontal tensile testing machine for teaching laboratories, built from 3D-printed parts, an aluminum profile and off-the-shelf electronics, with a documented architecture, published CAD models and source code, and a costed build of approximately R$ 3,700 at 2026 Brazilian prices. Three specimens, whose geometry was defined experimentally from preliminary tests with the machine itself, were tested to failure with a common 40 mm free length, and the machine reproduced brittle fracture in rigid PVC, yielding followed by neck formation and cold drawing in HDPE, and hyperelastic stiffening in an elastomeric ring, with the drawing mechanism confirmed on the specimen itself. The computed moduli span about four orders of magnitude across the three materials and place them in the correct relative order, which shows that the acquisition and data reduction chain resolves material classes as intended and that stiffness, strength, deformability and toughness are distinct properties that the stress–strain curve makes visible to the student.

Acknowledgments

The authors thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the research scholarship, and the National Institute of Telecommunications (Inatel) for supporting the development of this work. They also thank Felipe Emanoel Chaves and Wanderson Eleuterio Saldanha for their support and contributions to the project.

References

[1] W. D. Callister and D. G. Rethwisch, *Ciência e Engenharia de Materiais: Uma Introdução*, 9th ed. Rio de Janeiro, Brazil: LTC, 2016.

[2] J. H. Arrizabalaga, A. D. Simmons, and M. U. Nollert, "Fabrication of an economical Arduino-based uniaxial tensile tester," *J. Chem. Educ.*, vol. 94, no. 4, pp. 530–533, 2017.

[3] Instituto Federal de Minas Gerais (IFMG), Campus Arcos, "Máquina de ensaio de tração (Trabalho de Aplicação Interdisciplinar 4)," 2018. [Online]. Available: https://www.ifmg.edu.br/arcos/ensino-1/tai/20182_TAI4_Mquinadetrao.pdf [Accessed: Mar. 6, 2026].

[4] Instituto Federal de Minas Gerais (IFMG), Campus Arcos, "Máquina de ensaio de tração (Trabalho de Aplicação Interdisciplinar 3)," 2018. [Online]. Available: https://www.ifmg.edu.br/arcos/ensino-1/tai/20181_TAI3_Mquinadeensaiotrao.pdf [Accessed: Feb. 9, 2026].

[5] Universidade Estadual Paulista (UNESP), "Projeto de máquina de ensaio de tração," Repositório Institucional UNESP. [Online]. Available: https://repositorio.unesp.br/server/api/core/bitstreams/500f28b3-7690-4dba-b182-4ec470342bed/content [Accessed: May 14, 2026].

[6] X. Zhang, "Universal tensile testing machine, version two," Instructables, 2021. [Online]. Available: https://www.instructables.com/Universal-Tensile-Testing-Machine-VERSION-TWO/ [Accessed: Mar. 2, 2026].

[7] G. S. Emergente and L. C. Rodrigues, "Tensile testing machine — 3D printed parts, CAD models," Autodesk Fusion 360, 2026. [Online]. Available: moving grip, https://a360.co/443s2RN; moving load cell mount, https://a360.co/4wgyLUo; stationary grip, https://a360.co/3QBAzs5; grip block, https://a360.co/4gi13ZX; stepper motor mount, https://a360.co/4v6Lisq [Accessed: Aug. 27, 2026].

[8] G. S. Emergente, "UniversalTestingMachine: firmware and control interface source code," GitHub repository, 2026. [Online]. Available: https://github.com/Guilherme-souzae/UniversalTestingMachine [Accessed: Aug. 27, 2026].

[9] G. S. Emergente and L. C. Rodrigues, "Tensile tests on PVC, HDPE and elastomer specimens using the developed machine," video recording, 2026. [Online]. Available: https://youtu.be/TkaLVOvfQI4 [Accessed: Aug. 27, 2026].

[10] Instron, "3400 Series universal testing systems," product page, 2026. [Online]. Available: https://www.instron.com/wp-content/uploads/2024/07/3400-series-universal-testing-systems-10.pdf#page=18 [Accessed: Aug. 28, 2026].

[11] *Standard Test Method for Tensile Properties of Plastics*, ASTM D638-22, ASTM International, West Conshohocken, PA, USA, 2022.

[12] *Standard Test Methods for Tension Testing of Metallic Materials*, ASTM E8/E8M, ASTM International, West Conshohocken, PA, USA.

[13] *Metallic Materials — Calibration and Verification of Static Uniaxial Testing Machines — Part 1: Tension/Compression Testing Machines*, ISO 7500-1:2018, International Organization for Standardization, Geneva, Switzerland, 2018.

[14] *Standard Practice for Verification and Classification of Extensometer Systems*, ASTM E83, ASTM International, West Conshohocken, PA, USA.

[15] Impac Instrumentos de Medição, "Máquina universal de ensaios mecânicos ALX-B-500" and "Máquina universal de ensaios 1000 kgf 2 colunas IP-AEV-10000N," online catalogues, 2026. [Online]. Available: https://loja.impac.com.br/ [Accessed: Jul. 3, 2026].

[16] Banco Central do Brasil, "Cotação do dólar norte-americano (USD/BRL)," time series, 2026. [Accessed: Jul. 5, 2026].

[17] *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement*, JCGM 100:2008, Joint Committee for Guides in Metrology, 2008.

Development of a Low-Cost Pulse-Echo Teaching Apparatus for Brazilian Middle School Science Classes

Bruno Murakami
Universidade Federal do ABC
São Bernardo do Campo, Brazil
bruno.murakami@aluno.ufabc.edu.br

Caio de Castro e Freire
Independent Researcher
Santo André, Brazil
cdcfreire@gmail.com

John Andrew Sims
Universidade Federal do ABC
São Bernardo do Campo, Brazil
john.sims@ufabc.edu.br

***Abstract*—The Brazilian National Common Curricular Base includes the investigation of mechanisms involved in the transmission and reception of sound and images, the classification of electromagnetic radiation, and the discussion of technologies used in medical diagnosis and treatment. In this context, ultrasound provides an opportunity to connect wave physics, sensing technologies, and biomedical applications. This paper presents the technical development of a portable and low-cost teaching apparatus based on an Arduino Uno and an HC-SR04 ultrasonic sensor, together with a pedagogical sequence for introducing pulse-echo principles to middle school students. The apparatus estimates the distance to an object from the ultrasonic time-of-flight and converts the measurement into a visual light-emitting diode (LED) display, providing a tangible representation of an otherwise invisible process. The estimated cost of the prototype was approximately 83.50 Brazillian real. The proposed sequence connects everyday echoes, animal echolocation, audible and inaudible frequencies, wave reflection, and time-of-flight measurements. A preassembled version of the apparatus was operationally implemented during a 50-minute teaching-practice session at Universidade Federal do ABC. The implementation indicated that the core demonstrations and guided discussions can be conducted within one class period, whereas student-led assembly and a quantitative distance investigation would require additional time. Rather than evaluating student learning outcomes, this work focuses on the technical development, reproducibility, operational feasibility, pedagogical organization, and limitations of the proposed teaching resource. The resulting apparatus offers an accessible means of connecting wave physics, educational robotics, and introductory biomedical engineering concepts in basic science education.**

***Index Terms*—Ultrasound education, HC-SR04, low-cost apparatus, BNCC, physics teaching, educational robotics**

I. Introduction

The Brazilian National Common Curricular Base (BNCC) includes items related to waves, communication technologies, and applications of radiation in health [1]. Brazilian initiatives have also encouraged the inclusion of educational robotics, physical computing, and computational thinking in basic education [2]. Research involving educational robotics in Brazilian public-school contexts suggests that such activities may create opportunities for interdisciplinary work involving science, mathematics, electronics, and programming, provided that the technological resources are associated with explicit educational objectives [3].

The use of experimental resources in science education nevertheless requires careful planning. Limited preparation time, lack of ready-to-use activities, maintenance difficulties, and insufficient access to equipment may restrict the use of practical activities in schools [4]. Moreover, the presence of an apparatus alone does not demonstrate educational effectiveness; the physical resource must be connected to well-defined questions, representations, and opportunities for conceptual discussion [4], [5].

Skill EF09CI05 proposes the investigation of mechanisms involved in the transmission and reception of images and sounds and skill EF09CI07 extends this discussion to technologies used in medical diagnosis and treatment, including X-ray imaging, ultrasound, nuclear magnetic resonance, radiotherapy, and optical laser surgery [1]. Although ultrasound is a mechanical wave rather than electromagnetic radiation, the progression among these competencies provides a useful curricular context (while sound propagates as a mechanical wave, distinct from electromagnetic phenomena, there are physical similarities between the two which can be made use of in teaching). Concepts previously in skill EF09CI06 such as frequency, propagation, reflection, and signal reception can be connected to the classification of infrasound, audible sound, and ultrasound. This connection can then support an introductory discussion of how reflected acoustic signals are used in medical imaging.

The literature on the use of ultrasonic sensors in educational contexts and on the teaching of ultrasound-related concepts has shown a wide range of approaches. In some studies, the focus was on scientific literacy and the contextualization of medical technologies through didactic sequences, discussions, and other educational activities [6]–[8]. Other studies employed playful approaches and gamification-based strategies to address concepts related to waves and ultrasound [9], [10]. Another line of research focused on experimentation as a teaching methodology, using ultrasonic sensors and platforms such as Arduino to investigate physical phenomena or perform

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, through a master's scholarship.

data acquisition [11]–[13]. Despite the diversity of the studies identified, a significant proportion of these proposals are aimed at upper secondary or higher education students and, in some cases, involve extensive didactic sequences or greater technical complexity. This scenario suggests an opportunity for the development of simplified experimental approaches aimed at lower secondary education and compatible with the time and resources available in the classroom.

In this context, the objective of this work is to describe the development of a low-cost ultrasonic teaching apparatus and to organize a pedagogical sequence for introducing ultrasound principles in middle school science classes. The article evaluates the construction, cost, reproducibility, classroom timing, and technical limitations of the resource.

II. Theoretical Background

A. Ultrasound, Pulse-Echo and Time-of-Flight Measurement

Medical ultrasound is a diagnostic imaging technique that exploits the acoustic properties of human tissues to provide anatomical or functional information. Medical ultrasound includes multiple techniques - such as B-mode pulse-echo imaging, Doppler imaging, elastography, and tomographic methods - beyond the scope required for an introductory middle school activity. The present proposal therefore focuses on the fundamental sequence of pulse emission, propagation, reflection, echo reception, and time-of-flight estimation.

In a pulse-echo mode, a short acoustic pulse is emitted into a medium, such as the human body, and part of its energy is reflected when it encounters an object or an interface. The distance to a reflector can be estimated from the time interval between pulse transmission and echo reception. For a uniform propagation speed, the distance is calculated as shown in (1) [14]:

$$d = \frac{vt}{2}, \tag{1}$$

where d is the distance between the transmitter and the reflector, v is the speed of sound in the propagation medium, and t is the measured round-trip time. The factor of two accounts for the outward and return paths of the pulse.

Modern diagnostic ultrasound systems use transmitted pulses and returning echoes. The arrival time of an echo contributes to the estimation of depth, while echo amplitude and spatial scanning contribute to image formation. For basic education, the complete imaging chain is not required. The distance relation in (1) is sufficient to introduce the principle that information about an object can be obtained without direct visual contact by measuring the return time of a reflected wave.

B. Conceptual Scope of the HC-SR04 Analogy

The HC-SR04 is an ultrasonic ranging module containing a transmitter, a receiver, and control electronics. According to its technical documentation, fig. 1 shows that the module emits an eight-cycle burst at approximately 40 kHz and returns an electrical pulse whose duration is proportional to the acoustic

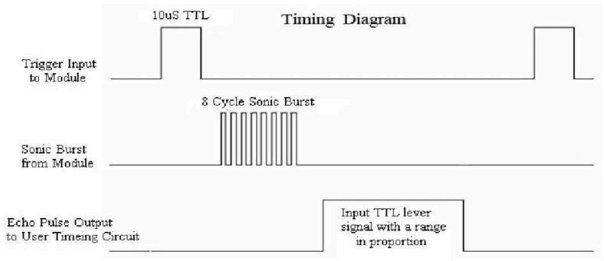


Fig. 1. Timing diagram of the HC-SR04 ultrasonic sensor, showing the 10 μs Transistor-Transistor Logic (TTL) trigger pulse, the 8-cycle ultrasonic burst, and the Echo output pulse whose width is proportional to the echo time-of-flight. Source: [15]

round-trip time. The manufacturer recommends a measurement cycle longer than 60 ms to reduce interference between successive measurements. The documented operating range is approximately 2 cm to 400 cm under suitable conditions [15].

The apparatus proposed in this work is a conceptual model of pulse-echo ranging and does not take into account many design aspects of medical ultrasound. Diagnostic ultrasound scanners use frequencies in range of 2 to 10 MHz, with specialized ultrasound applications up to 50 MHz. [14] In medical applications, ultrasound travels through the human body, which is composed of liquids, soft tissues and some solids, such as bone. The HC-SR04 operates in air at approximately 40 kHz. Medical image formation also requires beam steering or scanning, signal processing, amplitude encoding, and assumptions regarding tissue propagation speed [14]. These aspects are not reproduced by the HC-SR04 apparatus. Therefore, the analogy is restricted to the basic pulse-echo concepts of acoustic emission, propagation, reflection, reception, and round-trip time, without suggesting that the low-cost sensor reproduces a clinical ultrasound imaging system. These differences must be stated explicitly so that the analogy supports the intended physical concepts without suggesting that the low-cost sensor reproduces a clinical imaging system.

III. Methodology

A. Development of the Teaching Apparatus

The prototype was designed to provide an observable representation of an otherwise inaudible pulse-echo process. It was assembled using an Arduino Uno microcontroller, an HC-SR04 ultrasonic sensor, a breadboard, four light-emitting diodes (LEDs), resistors and jumper wires. Table I presents the estimated component cost at the time of construction.

The Arduino generates the trigger signal required by the sensor and measures the duration of the echo pulse. The round-trip time is converted into distance using (1), with the speed of sound in air used as the propagation speed (approximately 340 m/s). The measured value is then converted into a visible output through the LEDs, from 0 to 100 cm, with the four LEDs corresponding to 0 to 25 cm, 25 to 50 cm, 50 to 75 cm, and 75 to 100 cm, respectively. These LEDs provide an

TABLE I
APPROXIMATE COST OF THE COMPONENTS USED IN THE TEACHING APPARATUS

Component	Approximate Cost (R$)
Arduino Uno	45.00
HC-SR04 ultrasonic sensor	10.00
Breadboard	15.00
Light-emitting diodes (4 units)	2.00
Jumper wires	7.50
Resistors	4.00
Total	**83.50**

output reference for the distance being measured. The distance intervals associated with the LEDs can be modified in the source code according to the objectives of the activity. Fig. 2 summarizes the functionality of the code. In addition, an Arduino Serial Monitor output is displayed for the distance calculated and the time-of-flight measurement of the ultrasonic sensor.

Fig. 3 presents the circuit arrangement of Arduino with the ultrasonic sensor. Further details about the Arduino code and I/O connections were made available in the references [16].

B. Development of the Pedagogical Sequence

The pedagogical sequence was designed around two guiding questions:

1) How can a sensor operate using a sound that humans cannot hear?
2) How can the sensor estimate the distance to an object?

The conceptual progression of the class was organized in the following sequence:

everyday echo → animal echolocation → sound waves and frequency → pulse-echo ranging → medical ultrasound context

Table II summarizes a suggested lesson plan for a 50-minute class using a preassembled apparatus. The times are approximate and may be adapted to the school context.

The activity begins with the operation of the apparatus before a formal explanation is provided. Students may move an object in front of the sensor, observe changes in the visible output, and propose hypotheses regarding the type of signal used by the device. The teacher can then connect the observations to parking sensors, sonar systems, echoes in enclosed spaces, and animal echolocation.

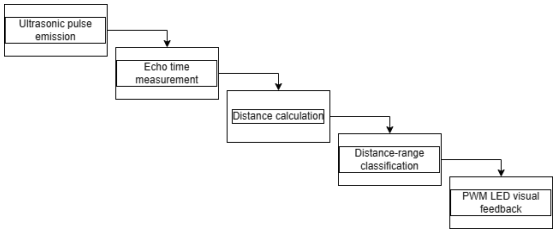


Fig. 2. Functional diagram of the ultrasonic distance-measuring apparatus.

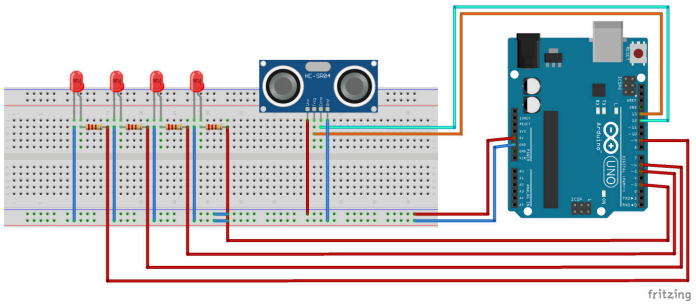


Fig. 3. Circuit arrangement of the proposed ultrasonic teaching apparatus Source: Authors.

Students may produce an initial representation through a drawing, a short written explanation, or an oral model of how the sensor operates. This representation is included as a formative pedagogical resource that helps the teacher identify ideas requiring discussion. It is not treated as research data in the present work.

The teacher can introduce the concept of wave reflection through the analogy of a ball traveling toward a wall and returning after impact. The analogy makes the outward and return paths visible, but the teacher should clarify that a sound wave is not a material object and that the analogy represents only propagation direction and reflection. Students may measure the time between throw and catch.

The teacher can introduce a discussion that addresses frequency and the limitations of human hearing. Audio signals of progressively higher frequency may be used to introduce the idea that sensory perception is limited and that inaudible sound can still be detected by an electronic device. Examples of animal communication and echolocation provide a transition to the HC-SR04 operating frequency: some whales communicate using low-frequency sounds (infrasound), while bats use high-frequency sounds (ultrasound) for echolocation.

The distance demonstration should connect the measured echo time to (1). A simple extension consists of positioning a flat object at different known distances and comparing the sensor reading with a ruler or measuring tape. This extension shifts the activity from demonstration towards a quantitative investigation, but requires additional time for data recording, calculation, comparison, and discussion.

TABLE II
SUGGESTED ORGANIZATION OF THE PEDAGOGICAL SEQUENCE.

Stage	Time
Exploration of the apparatus and guiding questions	5 min
Initial representation of the sensor operation	5 min
Everyday examples and animal echolocation	6 min
Ball-reflection analogy and discussion of echoes	7 min
Audible range, infrasound, and ultrasound	7 min
Pulse-echo distance demonstration and (1)	12 min
Revised representation and final discussion	8 min
Total	**50 min**

At the end of the sequence, students may revise their initial representation of how the equipment works. In the proposed pedagogical use, the revision is intended to promote self-explanation and teacher-mediated discussion.

C. Operational Implementation

A preassembled version of the apparatus and the proposed sequence were implemented during the teaching-practice component of the undergraduate course *NHT5013-15 (4-0-4) – Práticas de Ensino de Ciências e Matemática no Ensino Fundamental* at Universidade Federal do ABC (UFABC), taught by Prof. Caio de Castro e Freire in 2025.

A group of 14 seventh-grade students from a municipal public school located in the eastern region of the city of São Paulo visited UFABC to participate in the activity. Seventh-grade students in the Brazilian educational system are typically approximately 12 to 13 years old. The activity was designed and conducted by the first author under course supervision, with approximately 50 minutes available for its implementation.

The activity was conducted in a computer laboratory equipped with round tables. The students were divided into three groups. Only one preassembled apparatus was available during the session; therefore, each group approached the apparatus in turn to observe its operation and interact with it during the proposed activities. Despite the limited class time, all stages of the proposed pedagogical sequence were completed.

All lesson-plan materials and circuit arrangements were made available \cite{github} and can be adapted to different instructional contexts.

IV. Results

A. Prototype Operation and Cost

The apparatus was pre-assembled and used to demonstrate object detection based on ultrasonic pulse emission and echo reception. During operation, changes in the position of an object produced corresponding changes in the distance-dependent output of the system. The LEDs provided a visible response associated with the sensor measurement, allowing the teacher to discuss an acoustic process that is not directly audible to students.

The estimated component cost for the equipment was approximately R$83.50. This value should be interpreted as a local estimate rather than a fixed cost because component prices vary by supplier and date. The circuit uses commonly available components and can be assembled on a solderless breadboard, facilitating maintenance and modification.

B. Feasibility of the Pedagogical Sequence

With the apparatus assembled in advance, the central demonstrations and conceptual discussions could be accommodated within the available 50-minute session. The sequence allowed the teacher to relate the behavior of the US sensor with wave reflection, inaudible frequencies, animal echolocation, and time-of-flight distance estimation.

The limited class duration did not allow the implementation to include both a complete student-led circuit assembly and systematic quantitative investigation with repeated measurements and data analysis. Therefore, the single-session format is more appropriate for a guided demonstration and conceptual exploration. An additional class or extended workshop could include the construction of the complete ultrasound apparatus from its components. From this perspective, previous student contact with robotics activities in the school context facilitated the introduction of the basic elements: ultrasonic sensor, LEDs, and breadboard connections.

The classroom session also indicated the value of teacher mediation. The apparatus produces a distance measurement, but it does not independently explain why the echo time is related to distance, why the path is divided by two, or how the classroom sensor differs from a medical ultrasound scanner. These connections must be made explicit through questioning, diagrams, analogies, and mathematical discussion.

C. Technical and Pedagogical Limitations

The implementation of the circuit on a breadboard limits both student and instructor interaction with the apparatus. Although breadboards are well suited for static measurements, circuit prototyping, and direct demonstrations of sensors or microcontroller operation, they provide limited mechanical stability for activities that require physical manipulation of the device.

The apparatus should also be tested before each class to verify proper operation and should be used within a distance range that provides stable and reliable measurements under the environmental conditions of the classroom.

Another limitation regarding middle school classes is the limited class time. Considering that, several minutes are typically spent organizing students and taking attendance. Instructional time becomes a scarce resource. Although it would be ideal to have unlimited time for engaging discussions with active participation from all students, this is often not feasible within the constraints of a regular class period.

The comparison with medical ultrasound is limited to the general pulse-echo logic. The prototype does not produce anatomical images, does not operate in biological tissue, and does not reproduce the frequencies, transducers, beamforming methods, or signal-processing stages of diagnostic equipment. During the presentation, these topics were addressed only briefly. As the target audience consisted of middle school students, they were introduced primarily to clarify the limitations of the apparatus and to avoid an overly literal interpretation of the analogy.

V. Discussion

The main contribution of the proposed apparatus is to provide a physical bridge between an everyday distance sensor and the abstract pulse-echo principle. The HC-SR04 makes it possible to discuss emission, propagation, reflection, reception, and round-trip time using low-cost electronics. The apparatus also offers a point of connection among physics, basic

programming, sensor technology, and biomedical engineering without requiring the introduction of advanced ultrasound modalities.

The pedagogical proposal follows the BNCC progression by connecting sound transmission and reception with frequency classification and health-related technologies [1]. Nevertheless, its central content is wave physics rather than a comprehensive treatment of medical imaging. Medical ultrasound is used as a contextual application after the pulse-echo principle has been established.

From the teacher's perspective, one useful discussion point is the possibility that students initially interpret the sensor as a light-based or infrared device because many familiar object-detection systems are associated with visible or invisible light. This possible interpretation can be used pedagogically to contrast acoustic ranging with optical sensing and to prepare later comparisons between ultrasound and X-ray imaging. In this sense, the apparatus does not replace conceptual mediation; it creates a concrete situation in which the teacher can ask why an inaudible sound can still be emitted, reflected, received, and converted into a visible signal.

The 50-minute format imposes an important design choice. A preassembled apparatus permits more time for conceptual discussion, whereas student assembly can support electronics and programming objectives but reduces the time available for investigating the physical relationship between echo time and distance. Others experimental activities may also strengthen the investigative character of the sequence, as proposed by Martins and Lisbôa [12]. However, considering the target audience, the introduction of more complex activities should take into account students' prior knowledge and their familiarity with electronics, programming, and the physical concepts involved.

Although the proposed apparatus is based on low-cost and widely available components, its implementation in different school contexts may depend on factors beyond the cost of the materials. The availability of computers and electronic components, the school infrastructure, the time available for preparing the activity, and teachers' familiarity with electronics and programming may influence how the proposal is implemented. Therefore, adaptations may be necessary according to the resources and characteristics of each educational context. The work of Nunes and Rocha [6], for example, provides an alternative approach to the broader topic of "Radiations and their applications in health" through debates and a simulated jury in a school context characterized by limited resources and infrastructure. This example illustrates how different methodological approaches may be adopted according to the conditions of each educational context.

The inclusion of an evaluation with an initial and a revised representation of the apparatus functionality represents a pedagogically useful resource as it provides a structured opportunity for students to articulate and reconsider an explanatory model. However, in the present paper, these representations are components of the proposed lesson plan rather than data supporting educational claims. Therefore, although assessment activities may be incorporated into classroom practice, the present study does not provide evidence that the experimental resource improves students' conceptual understanding.

A study employing a more rigorous methodology could be implemented to systematically assess student learning through, for example, pre- and post-activity tests, questionnaires containing conceptual questions, or qualitative analysis of the explanations produced by students. Likert-scale instruments could also be used as a complementary approach to investigate aspects such as interest, engagement, or students' perceptions of the topic. An example of obtaining evidence of learning was presented by Nunes and Rocha [6], through the classification and analysis of transcripts of student interactions during the activity. The use of more systematic assessment procedures in future studies could provide more robust evidence regarding the contribution of the proposed experimental resource to the learning of the concepts addressed in the classroom.

Taken together, the implementation of the proposed activity highlights several practical lessons. First, the use of a preassembled apparatus makes it possible to prioritize conceptual discussion within a regular 50-minute class, whereas student-led construction requires additional instructional time. Second, the low cost of the apparatus alone does not ensure its applicability, since teacher preparation, infrastructure, and students' previous familiarity with electronics may influence its implementation. Finally, the apparatus is most appropriately used as a teacher-mediated conceptual model of pulse-echo rather than as a direct representation of diagnostic ultrasound. At the same time, important research gaps remain. The present study did not systematically evaluate student learning, conceptual change, or engagement, nor did it examine how students interpret the analogy between the HC-SR04 and medical ultrasound. Future studies should therefore investigate these aspects and replicate the activity across different school contexts.

VI. Conclusion

This work presented the development of a portable, low-cost ultrasonic teaching apparatus and a pedagogical sequence for introducing pulse-echo principles in middle school science classes. The prototype was assembled from commonly available components at an estimated cost of R$83.50 and was operationally suitable for demonstrating ultrasonic object detection and a simplified time-of-flight distance estimation.

A preassembled version of the apparatus can support the core sequence within a 50-minute class. However, student-led assembly and systematic distance measurements require additional time and are better suited to a second class or an extended workshop. The apparatus should be presented as a conceptual model of pulse-echo ranging, not as a reproduction of a diagnostic ultrasound imaging system.

These results support the feasibility of the proposed resource as a low-cost approach for introducing pulse-echo and time-of-flight concepts. However, educational effectiveness was not formally assessed in this study. Future work should include

technical calibration, classroom replication in different contexts, and formal pedagogical studies with learning assessment to investigate how students understand and use the proposed representations and measurements.

ACKNOWLEDGMENT

The authors acknowledge the financial support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, through a master's scholarship. The authors also thank Universidade Federal do ABC (UFABC) and the participants of the course *NHT5013-15 – Práticas de Ensino de Ciências e Matemática no Ensino Fundamental* for supporting the development of the teaching activity.

REFERENCES

[1] Brasil, Ministério da Educação, "Base Nacional Comum Curricular (BNCC)," Brasília, DF, Brazil, 2018. [Online]. Available: https://www.gov.br/mec/pt-br/escola-em-tempo-integral/BNCC_EI_EF_110518_versaofinal.pdf [Accessed: May 20, 2026].

[2] Brasil, Fundo Nacional de Desenvolvimento da Educação, "Audiência Pública nº 04/2017 – Robótica Educacional," 2017. [Online]. Available: https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/acoes/compras-governamentais/compras-nacionais/audiencias-publicas/anos-anteriores/audiencia-publica-no-4-2017-2013-robotica-educacional [Accessed: May 20, 2026].

[3] M. R. dos Santos and R. C. G. Meneghetti, "Aspectos da interdisciplinaridade em dissertações e teses que versam sobre a Robótica Educacional com alunos de escolas públicas de Educação Básica," *Ciência & Educação*, vol. 30, pp. 1–20, 2024, doi: 10.1590/1516-731320240010.

[4] A. T. Borges, "Novos rumos para o laboratório escolar de ciências," *Caderno Brasileiro de Ensino de Física*, vol. 19, no. 3, pp. 291–313, 2002.

[5] M. A. Moreira, "Desafios no ensino da física," *Revista Brasileira de Ensino de Física*, vol. 43, suppl. 1, Art. no. e20200451, 2021, doi: 10.1590/1806-9126-RBEF-2020-0451.

[6] T. M. Nunes and D. M. Rocha, "Fake news e o ensino de ciências: explorando indicadores de alfabetização científica," *Revista de Ensino de Ciências e Matemática*, vol. 14, no. 3, pp. 1–23, 2023, doi: 10.26843/rencima.v14n3a11.

[7] N. C. S. Garajau, T. Santos, L. R. Oliveira, and J. R. Neves, "Radiação e suas aplicações na saúde: contribuições para o ensino de Ciências," in *Anais do VII Congresso de Inovação e Tecnologia em Diagnóstico do Agreste Alagoano (CINTEC)*, Arapiraca, AL, Brazil, 2025, pp. 24–29.

[8] S. Sudirman, D. Kennedy, and S. Soeharto, "The teaching of physics at upper secondary school level: A comparative study between Indonesia and Ireland," *Frontiers in Education*, vol. 8, Art. no. 1118873, 2023, doi: 10.3389/feduc.2023.1118873.

[9] W. Teixeira Júnior, C. A. P. Monteiro, and V. H. L. de Barros, "Objeto de aprendizagem Morcego Tech: pensamento computacional, ensino-aprendizagem de Ciências, Matemática e robótica educacional," in *Anais do XVIII Seminário Sul-Mato-Grossense de Pesquisa em Educação Matemática*, vol. 18, no. 1, pp. 191–203, 2024.

[10] T. F. Couto and J. V. Zampieron, "Gamificação e robótica educacionais no ensino de ondulatória no Ensino Médio: desenvolvimento e avaliação de uma intervenção pedagógica," *Experiências em Ensino de Ciências*, vol. 21, no. 1, pp. 257–274, 2026.

[11] H. G. Riter, E. A. Soares, L. Mota, M. P. P. de Castro, and M. G. da Silva, "Experimentos de baixo custo para o ensino de física," in *Anais do VIII Congresso de Ensino, Pesquisa e Extensão (CONEPE)*, 2021.

[12] J. F. T. Martins and R. S. M. Lisbôa, "Aprendizagem de reflexão sonora com Arduino por meio da construção e investigação de um protótipo ultrassônico em ambiente maker," *Experiências em Ensino de Ciências*, vol. 21, no. 1, pp. 380–395, 2026.

[13] M. B. Valentín, C. R. de Bom, M. P. de Albuquerque, M. P. de Albuquerque, E. L. Faria, and M. D. Correia, "Aquisição de imagens ultrassônicas como experimento didático," *Revista Brasileira de Ensino de Física*, vol. 40, no. 2, Art. no. e2503, 2018, doi: 10.1590/1806-9126-RBEF-2017-0224.

[14] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt Jr., and J. M. Boone, "Ultrasound," in *The Essential Physics of Medical Imaging*, 3rd ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2012, ch. 14, pp. 500–576.

[15] ElecFreaks, "Ultrasonic Ranging Module HC-SR04: Product Manual," [Online]. Available: https://cdn.sparkfun.com/datasheets/Sensors/Proximity/HCSR04.pdf [Accessed: Jun. 19, 2026].

[16] B. Murakami, "Low-Cost Ultrasonic Distance Sensor for Science Education," GitHub repository, 2026. [Online]. Available: https://github.com/murakami-bruno/Repository-title-Low-Cost-Ultrasonic-Distance-Sensor-for-Science-Education [Accessed: Jun. 26, 2026].

Inatel Museum of Biomedical Engineering: A Proposal for an Interactive Collection of Resin-Printed Biological Educational Models

Vitor Hugo Felipe Noronha
Software Engineering
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
vitor.noronha@ges.inatel.br

Elisa Rennó Carneiro Déster
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
elisa.renno@inatel.br

Maysa Costa Alves
Department of Biomedical Engineering
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, Brazil
maysa.alves@inatel.br

***Abstract*—Teaching biological, microbiological, and chemical concepts in Biomedical Engineering programs has traditionally relied on two-dimensional representations, which limit the understanding of complex morphologies, microarchitectures, and molecular structures. This paper presents the conceptual design of the *Inatel Museum of Biomedical Engineering*, a proposal for a permanent and traveling collection of three-dimensional educational models based on an additive manufacturing workflow previously validated by the same research group through the development of a DNA double-helix model fabricated using fused deposition modeling (FDM) [1]. The planned transition to photopolymer resin printing (MSLA) is expected to provide greater geometric accuracy, improved surface quality, transparency, color customization, and detachable components, enabling the production of educational models representing DNA, atomic orbitals, bacterial cell walls, viruses (HIV and SARS-CoV-2), and a macrophage performing phagocytosis. The collection also includes a flagship technological exhibit consisting of an articulated bacteriophage model equipped with LED lighting and a microcontroller to dynamically demonstrate the viral infection process. The museum is conceived as an educational walkthrough featuring exhibits protected by acrylic domes or displayed on interactive tactile stations accompanied by accessible explanatory texts. The proposal is discussed against a comparable Brazilian museum initiative and a critical review of its own technological choices. This paper presents the conceptual design and future implementation plan of the project rather than completed experimental results.**

***Index Terms*—Science Communication, Additive Manufacturing, Resin 3D Printing, Accessibility, Science Museum, Biomedical Engineering, Tactile Learning**

I. Introduction

Biomedical Engineering integrates concepts from biology, chemistry, physics, and medicine to develop technologies for healthcare applications. Many of these concepts—from the organization of bacterial cell walls to the three-dimensional conformation of viral proteins—are inherently spatial and therefore difficult to convey through static illustrations found in textbooks and lecture slides. This limitation has been discussed in previous work by the same research group, which identified additive manufacturing as an effective approach for producing manipulable educational models capable of enhancing spatial understanding [1].

As a proof of concept, a complete workflow for manufacturing three-dimensional educational models was previously developed and validated, encompassing biological data acquisition, digital modeling, slicing, fabrication, post-processing, assembly, and evaluation through the production of a DNA double-helix model using fused deposition modeling (FDM) [1]. This validated workflow established the methodological foundation upon which the present proposal is built.

Building upon that experience, and motivated by the superior resolution and geometric fidelity offered by masked stereolithography (MSLA) resin printing compared with FDM, the authors, together with the project supervisor, conceived a natural extension of the previous research: the creation of the **Inatel Museum of Biomedical Engineering**. The proposed museum consists of a physical, traveling, or hybrid collection of tactile three-dimensional educational models representing biological structures relevant to Biomedical Engineering curricula, pursuing three primary objectives:

1) promote science communication among students, children, and the general public;
2) enhance accessibility for visually impaired individuals through tactile educational exhibits;
3) strengthen the relationship between the institution and society while increasing academic and institutional visibility.

It is important to clarify the current stage of this project. While the DNA model produced using FDM has already been fabricated and experimentally evaluated [1], the museum proposed in this work—including the illuminated bacteriophage exhibit—remains a **conceptual proposal and future development plan** that has not yet been fully implemented. Therefore, this paper presents the conceptual design of the museum rather than reporting experimental results regarding the museum itself.

The main contributions of this work are summarized as follows:

1) the proposal of a conceptual layout for the museum based on an educational walkthrough;

2) the definition of an initial catalog of resin-printed educational models, together with their biological and pedagogical motivations;
3) the proposal of integrating additive manufacturing with embedded electronics to create interactive educational exhibits;
4) a discussion of the expected social, educational, and institutional impacts of the initiative, situated against a comparable Brazilian museum initiative and a critical review of the technological choices made.

II. Conception of the Museum

A. Format and Architecture

A hybrid museum format is proposed, combining a permanent installation—located in hallways and adjacent institutional spaces—with traveling modules that can be taken to schools, scientific events, and educational exhibitions. The central architectural concept is an **educational labyrinth**, in which visitors follow a sequential pathway beginning at the microscopic scale (the origin and diversity of bacteria), progressing through historical pandemics and the search for treatments, and culminating in a structural biology section where molecular and atomic structures are presented at larger scales and with greater detail. Fig. 1 illustrates the conceptual layout of the proposed educational pathway.



Fig. 1. Conceptual sketch of the proposed educational labyrinth for the Inatel Museum of Biomedical Engineering, guiding visitors from introductory microbiology to structural biology. Illustrative representation of the proposal, not yet physically implemented. Conceptual art generated by Artificial Intelligence.

B. Exhibition Modes

Two exhibition modalities are proposed, selected according to the fragility and educational purpose of each model:

- **Protected exhibits**, displayed under acrylic domes, intended for delicate models containing small detachable components or detailed painting;
- **Interactive exhibition benches**, featuring durable models designed for direct handling, particularly those intended for tactile learning and visitors with visual impairments.

Each exhibit is accompanied by accessible explanatory text covering three complementary dimensions: the biological nature of the represented structure, its functions or applications (for example, in industrial, biotechnological, or physiological processes), and, whenever applicable, the diseases or harmful effects associated with the represented organism or molecule.

C. Initial Collection of Models

The proposed initial collection consists of six educational models described below. With the exception of the DNA model—which has already been validated using FDM technology [1]—all remaining models are currently at the conceptual design stage.

1) Realistic DNA Model: A detailed double-helix model representing nucleotides, phosphate groups, and deoxyribose molecules. This model is an evolution of the FDM prototype previously fabricated and evaluated by the research group [1] and is intended to be reproduced using photopolymer resin to achieve greater geometric fidelity and detachable nucleotide components.

2) Schrödinger Atomic Model: A representation of electron probability clouds (s, p, and d orbitals) manufactured with translucent resin, allowing simultaneous visualization of the atomic nucleus and the spatial distribution of electron density—a concept that is traditionally difficult to understand in chemistry and physical chemistry courses.

3) Bacterial Cell Wall: Comparative tactile models of Gram-positive and Gram-negative bacteria, highlighting differences in peptidoglycan thickness, the presence of an outer membrane, and the location of lipopolysaccharides, thereby supporting the understanding of antibiotic mechanisms of action.

4) HIV and SARS-CoV-2 Models: Comparative virus models displaying the viral capsid and major surface proteins (gp120 glycoprotein for HIV and the *Spike* protein for SARS-CoV-2), illustrating their role in host-cell receptor recognition.

5) Macrophage Engulfing a Bacterium: A model representing the phagocytosis process, illustrating the innate immune response to invading microorganisms through a cross-sectional view that reveals phagosome formation.

6) Bacteriophage Model (Flagship Technological Exhibit): The centerpiece integrating Biomedical Engineering and Electronics, described in detail in Section III, consists of an articulated bacteriophage infecting a bacterium. The model is designed to incorporate LED lighting and a microcontroller capable of simulating, through sequential light patterns, the injection of viral genetic material into the bacterial host.

III. Technological Integration

A. Photopolymer Resin Manufacturing

The transition from FDM to masked stereolithography (MSLA) printing, already planned and partially supported by the acquisition of an Anycubic Photon Mono M7 Pro printer by the research group [1], provides the technical capability required to manufacture the high-resolution models proposed for the collection, particularly the translucent atomic model and the viral capsids. Standard, translucent, and flexible photopolymer resins are proposed for different models according to the transparency, flexibility, and mechanical strength required for each educational application.

The printer was acquired in August, and initial resin test prints have since begun; an early run was interrupted when the vat's build-plate film (FEP) was punctured, and a replacement is being sourced before fabrication of the collection's models proceeds. No museum exhibit has yet been produced on this equipment.

B. Embedded Electronics

For the bacteriophage model (Fig. 2), the integration of an addressable LED strip controlled by a low-cost microcontroller (Arduino- or ESP32-based) is proposed. The controller will execute a programmed lighting sequence designed to simulate, in an educational rather than literal manner, the injection of viral genetic material into the bacterial host. The model is conceived with articulated legs, allowing manual positioning over the bacterial surface. This is the only electromechanical and dynamic exhibit planned for the initial collection; its development, electrical safety testing, and durability validation remain future tasks that must be completed before any public exhibition.



Fig. 2. Artistic concept of the articulated bacteriophage model with LED illumination, designed to visually simulate the injection of viral genetic material into a bacterial host. The exhibit is currently at the conceptual design stage and is based on the 3D model designed by Leon Morlion [2].

C. Scalability and Modularity

As with the workflow previously validated for the DNA model [1], every new museum exhibit will follow the same sequence of scientific validation, digital modeling, manufacturing, post-processing, and educational evaluation. This standardized workflow enables the collection to expand incrementally without requiring a redesign of the manufacturing process for each new model, while also facilitating the future incorporation of educational models proposed by faculty members from other disciplines.

IV. Social Impact and Institutional Visibility

A. Accessibility and Inclusion

The availability of tactile educational models aims to provide visitors with visual impairments an alternative means of understanding the shape, spatial organization, and structural relationships of biological objects, consistent with previous evidence demonstrating the educational benefits of 3D-printed tactile resources for this audience [3], [4].

B. Science Communication and University Outreach

By welcoming elementary and high school students, families, and visitors during public events and weekend activities, the proposed museum is intended to serve as a continuous university outreach initiative, making advanced topics in structural biology and microbiology accessible to audiences that would otherwise have limited exposure to such scientific content.

C. Institutional Visibility

Initiatives of this nature have the potential to generate local and regional media coverage, strengthen partnerships with schools, and attract prospective undergraduate students to the institution. Although these outcomes are not the primary educational objective of the project, they contribute to its long-term institutional sustainability and visibility.

V. Implementation Plan

A. Proposed Timeline

Table I presents a preliminary four-phase implementation schedule, starting from the completion of the validated FDM manufacturing workflow.

TABLE I
Preliminary implementation schedule for the proposed museum

Phase	Planned Activities
1	Expansion of the digital catalog and resin fabrication of the first models (DNA, atomic orbitals, bacterial cell wall).
2	Acquisition and assembly of the exhibition infrastructure (acrylic display domes, interactive benches, and accessible signage).
3	Design, prototyping, and testing of the electromechanical bacteriophage model (LEDs and microcontroller).
4	Pilot opening of the collection and structured collection of feedback from visitors and faculty members.

B. Preliminary Resource Estimation

Table II presents a preliminary estimate of the additional resources required for the pilot phase, considering the materials already budgeted for the ongoing resin-printing project and the electronic components required for the bacteriophage model. These values are approximate and intended solely for planning purposes, and they remain subject to formal quotations before acquisition.

TABLE II
PRELIMINARY RESOURCE ESTIMATE FOR THE MUSEUM PILOT PHASE

Item	Estimated Cost (R$)
Additional resins (standard and transparent)	900.00
Acrylic display domes and exhibition supports	700.00
Microcontroller, LED strips, and wiring	350.00
Accessible signage (including Braille)	400.00
Assembly and finishing materials	250.00
Estimated Total (Pilot Phase)	**2,600.00**

VI. Discussion and Limitations

A. Comparison with Related Initiatives

Brazil already has a well-established example of a science museum built around microbiology and biomedical content: the Museu de Microbiologia Prof. Isaias Raw, part of Instituto Butantan's Parque da Ciência in São Paulo [8]. Open since 2002, it uses three-dimensional models, microscopes, and interactive media to guide visitors through the history and scale of microorganisms, and is explicitly accessible to visitors with disabilities, offering tactile models, enlarged specimens, audio description, and Braille materials. Its scale — a permanent, professionally staffed exhibition within a major public research institute — is far beyond what the present proposal can achieve as an undergraduate project, but it demonstrates that the core premise of this work, a microbiology-themed, tactile, accessible museum, is a proven and valued format in the Brazilian context rather than a speculative one. The present proposal differs mainly in scope and origin: a smaller, curriculum-specific collection built from synthetic, designed models (DNA, atomic structures, engineered viral models) rather than biological specimens, developed as part of a Biomedical Engineering undergraduate research project rather than a dedicated museum institution.

B. Critical Analysis of Technological Choices

The decision to move from FDM to resin (MSLA) printing, and to add embedded electronics to the bacteriophage exhibit, involves trade-offs worth stating explicitly. Resin offers finer detail than FDM — a property reported as the single strongest driver of museum visitors' preference for 3D-printed replicas [7] — which favors the translucent atomic model and viral capsids proposed in Section II-C. The same source, however, found robustness a meaningful secondary concern for handled objects, and resin parts are typically more brittle than FDM parts under the repeated handling expected on the interactive benches described in Section II-B; print technology may therefore need to be chosen per exhibit rather than applied uniformly. Likewise, the LED and microcontroller system is the collection's most technically ambitious element, and also its highest-risk one in cost, electrical-safety review, and maintenance under repeated public use, none of which apply to the collection's five static exhibits. A simpler, passive alternative, such as an audio- or QR-code-triggered explanation, would communicate the same infection sequence without embedded electronics. The more complex illuminated version was nonetheless chosen because peer review specifically identified it as the collection's clearest technical contribution to Biomedical Engineering, distinct from an education or outreach initiative.

C. Current Stage and Limitations

The main limitation of this work is its inherently propositional nature. With the exception of the DNA model, which has already been manufactured and evaluated using FDM [1], the remaining models in the proposed collection—particularly the LED-illuminated bacteriophage—have not yet been prototyped or experimentally tested, although initial resin test printing has begun (Section III-A). The technical feasibility of the electronic integration, electrical safety assessments, durability testing, and pedagogical validation with students and museum visitors remain subjects for future work.

Likewise, the values presented in Tables I and II represent planning estimates rather than measured data and should be revised according to actual quotations and any scope adjustments approved during project development.

Nevertheless, because the proposal is grounded on an additive manufacturing workflow that has already been experimentally validated [1], together with well-established literature on accessibility and three-dimensional educational resources [3]–[6], it goes beyond a purely speculative concept by providing a concrete and technically supported roadmap for expanding the undergraduate research project.

VII. Future Work

Because this paper reports a conceptual design together with the very early stages of physical implementation (Section III-A), most of the museum's technical and pedagogical validation remains ahead of it. A more concrete pilot plan, with systematic collection of pedagogical data, is described below to guide that work and to be reported in a follow-up publication.

A. Pedagogical Evaluation Plan

Once the bacteriophage exhibit and an initial set of resin models are fabricated, a preliminary pedagogical evaluation is planned using an anonymous questionnaire administered to a convenience sample of volunteer students, instructors, and, where possible, visually impaired users. The instrument will be applied before and after interaction with the exhibits and will cover three dimensions: comprehension (a small number of questions on the biological content represented, to estimate

learning gain), usability (Likert-scale items on the clarity of the models and, for the bacteriophage, its lighting sequence), and accessibility and educational perception (Likert-scale items on engagement, tactile clarity, and perceived educational value, with wording adapted for visually impaired respondents where needed). Anonymity is intended to encourage candid feedback, particularly on usability problems. As with the group's earlier DNA prototype [1], this evaluation is explicitly exploratory: with a convenience sample, it is meant to surface usability issues and a first, non-generalizable signal of learning gain and accessibility benefit, not to statistically validate the museum's pedagogical impact.

B. Fabrication and Technical Validation

Future work also includes completing resin fabrication of the five remaining exhibits, resolving the FEP-related interruption described in Section III-A and reporting fabrication time, material cost, and any print failures for each model at the same level of detail already reported for the group's FDM DNA prototype [1]; constructing and electrically safety-testing the bacteriophage's LED and microcontroller system before any public exhibition; and durability testing of both the resin models and the exhibition infrastructure (acrylic domes, interactive benches) under repeated handling.

C. Broader Directions

Beyond the initial collection, future work may extend the catalog with models proposed by faculty from other disciplines, formalize partnerships with local schools for recurring visits, and, if the pilot evaluation above indicates a learning or accessibility benefit, pursue a larger, more rigorously sampled study of the museum's educational impact.

VIII. Conclusion

This paper presented the conception of the Inatel Museum of Biomedical Engineering, a proposal for a tactile and interactive collection of three-dimensional educational models intended to support the teaching of biology, microbiology, and chemistry within Biomedical Engineering education. Building upon the additive manufacturing workflow previously validated by the research group through the FDM DNA model, the proposal introduces an initial catalog of six educational models, an architectural concept based on an "educational labyrinth," a flagship technological exhibit combining resin 3D printing with embedded electronics, and an implementation plan discussed against a comparable Brazilian museum initiative and a critical review of its own technological choices.

This paper accordingly presents a conceptual design and implementation plan rather than completed experimental results. Concrete progress toward execution has begun: the MSLA printer required for the collection was acquired in August, and initial resin test prints followed, though an early run was interrupted by a punctured build-plate film whose replacement is being sourced (Section III-A). No museum exhibit has yet been fabricated. Completing resin production of the collection, constructing and safety-testing the bacteriophage exhibit, and carrying out the pedagogical evaluation detailed in Section VII remain the next steps, to be reported in a follow-up publication.

References

[1] V. H. F. Noronha, M. C. Alves, and E. R. C. Déster, "Development of three-dimensional educational models for biomedical engineering using 3D printing," unpublished manuscript.

[2] L. Morlion, "Bacteriophage vs. bacteria diorama," MakerWorld. [Online]. Available: https://makerworld.com/pt/models/464688

[3] C. L. Price, J. M. Houghton, A. P. Preston, R. A. Brown, and S. K. McMahon, "Visible through touch: Open-source 3D-printed tools for inclusive microbiology education," *Access Microbiology*, vol. 8, 2026.

[4] L. Holloway, K. Marriott, and M. Butler, "Accessible maps for the blind: Comparing 3D-printed models with tactile graphics," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2018.

[5] A. Aslan and Y. Celik, "A literature review on 3D printing technologies in education," *International Journal of 3D Printing Technologies and Digital Industry*, vol. 6, no. 3, pp. 592–613, 2022.

[6] S. Ford and T. Minshall, "Invited review article: Where and how 3D printing is used in teaching and education," *Additive Manufacturing*, vol. 25, pp. 131–150, 2019.

[7] P. F. Wilson, J. Stott, J. M. Warnett, A. Attridge, M. P. Smith, and M. A. Williams, "Museum visitor preference for the physical properties of 3D printed replicas," *Journal of Cultural Heritage*, vol. 32, pp. 176–185, 2018.

[8] Instituto Butantan, "Museu de Microbiologia Prof. Isaias Raw," Parque da Ciência Butantan. [Online]. Available: https://parquedaciencia.butantan.gov.br/programacao/atracoes/museu-de-microbiologia

Learning Through Innovation: Student-Led Scientific Events as Transformative Learning Ecosystems in Biomedical Engineering Education: An Integrative Review

Paula Celine Barbosa da Silva
PPG em Engenharia Biome´dica Universidade Federal de Pernambuco Recife, Brazil
paula.celine@ufpe.br
ORCID: 0009-0002-6483-9908

Lisiane Lima Felix
PPG em Engenharia Biome´dica Universidade Federal de Pernambuco
Recife, Brazil
lisiane.felix@ufpe.br ORCID:
0000-0002-3044-3465

Melissa Soares de Souza
PPG em Engenharia Biomédica Universidade Federal de Pernambuco Recife, Brazil
melissa.soaress@ufpe.br
ORCID: 0009-0001-8098-5860

Isla Pereira da Silva
PPG em Engenharia Biome´dica Universidade Federal de Pernambuco
Recife, Brazil
isla.silva@ufpe.br ORCID:
0009-0003-8093-5745

Neferson Barbosa da Silva Ramos
PPG em Ed. Matema´tica e Tecnolo´gica Universidade Federal de Pernambuco
Recife, Brazil neferson.barbosa@ufpe.br
ORCID: 0000-0001-9962-120X

Cristine Martins Gomes de Gusma˜o *PPG em Engenharia Biome´dica Universidade Federal de Pernambuco* Recife, Brazil
cristine.gusmao@ufpe.br
ORCID:
0000-0001-8831-217X

***Abstract*—This study presents an integrative literature review on the impact of student-led scientific events as transformative learning ecosystems in Biomedical Engineering education. The analysis of 27 studies published between 2023 and 2026 demon-strates that these initiatives promote student leadership and the development of essential technical and socio-emotional skills. The findings indicate that such events articulate teaching, research, and outreach, aligning with Sustainable Development Goals (SDG 3, 4, and 9). A conceptual model is proposed, recognizing these events as structured pedagogical devices with the potential to transform academic training.**

***Index Terms*—Biomedical Engineering; Student-led Scientific Events; Learning Ecosystems; Pedagogical Innovation; Active Learning.**

I. Introduction

In recent decades, the healthcare sector has been deeply impacted by intensifying digitalization and the expanding use of Artificial Intelligence (AI)—a phenomenon directly linked to the exponential growth of healthcare data volume. This scenario, frequently termed the ”data era,” has enabled the development of systems capable of analyzing large clinical datasets, contributing to more accurate diagnoses, early risk identification, and therapeutic decision support. Such advance-ments have driven significant improvements in the quality, safety, and agility of healthcare [1]–[5]. In the Brazilian context, initiatives like the Digital Health Strategy for Brazil 2020–2028 reinforce this movement by proposing the inte-gration of information systems and the strengthening of the Unified Health System (SUS), expanding access and service efficiency. This research is intrinsically aligned with SDG 3 (Good Health and Well-being) by training professionals for the sector’s new technological demands, and SDG 4 (Quality Education) by proposing ecosystems that modernize higher education through student leadership [6]. Furthermore, the advancement of AI in areas such as telemedicine, hospital management, and predictive analysis highlights that the mere incorporation of technologies is insufficient to guarantee real impact on clinical practice. One of the main contemporary challenges lies in training professionals capable of under-standing, interpreting, and applying these tools in a critical, ethical, and efficient manner. In this sense, issues such as the explainability of AI systems and the reliability of algorithms become central, requiring a multidisciplinary approach that articulates technical, clinical, and social knowledge. Thus, higher education, especially in Biomedical Engineering, de-mands innovative pedagogical strategies that prepare students to handle complex and constantly transforming scenarios. This need for pedagogical adaptation reinforces Target 4.7 of SDG 4, which aims to ensure that all learners acquire the theoretical and practical knowledge needed to promote sustainable devel-opment and healthy lifestyles. In this scenario, active learning strategies gain prominence as effective alternatives to foster the development of essential 21st-century skills, such as critical thinking, problem-solving, creativity, and teamwork. By deal-ing with real challenges from industry and healthcare services, students are encouraged to integrate theoretical and practical knowledge, promoting a more meaningful and contextualized education. Approaches such as challenge-based learning and transdisciplinary experiential learning have demonstrated pos-

itive results in Biomedical Engineering education by bringing students closer to the real demands of healthcare systems and encouraging innovative solutions [14], [24]. The promotion of this educational technological innovation also dialogues with SDG 9 (Industry, Innovation, and Infrastructure) by strengthening scientific research and the technological capacity of the academic sector.

In parallel, the construction of collaborative educational ecosystems has become a fundamental element in the training of future professionals. These ecosystems involve the inte-gration of teaching, research, outreach, and extracurricular activities, fostering a dynamic environment for knowledge ex-change and continuous development. In this context, initiatives that strengthen academic mentoring, diversity, and inclusion contribute to expanding learning opportunities and stimulating student leadership—essential elements for consolidating more robust and innovative academic trajectories [18], [23]. By focusing on inclusion and equal educational opportunities, these initiatives support Target 4.5 of SDG 4, eliminating gender disparities and ensuring equal access to high-level vocational and professional training.

Therefore, it is within this framework that student-led scien-tific events emerge, configuring themselves as strategic spaces for active learning, interdisciplinary collaboration, and knowl-edge dissemination. These events enable practical experience in scientific endeavors, stimulate academic communication, and promote integration among different fields of knowledge. The Academic Symposium on Biomedical Engineering and In-novation (SABIO), for instance, represents a relevant initiative by articulating teaching, research, and innovation, allowing students to assume a central role in knowledge construction and in developing solutions for real healthcare challenges. This student-centered approach to solving complex public health problems directly contributes to Target 3.c of SDG 3 by helping to develop and train a skilled and technologically capable health workforce [3], [17].

However, despite the evident relevance of these forma-tive environments, studies that scientifically systematize and analyze the impact of student-led events on academic and professional training remain scarce. This gap highlights the need for investigations exploring how these experiences con-tribute to skill development, educational innovation, and the transformation of teaching-learning processes in Biomedical Engineering. Given this, it becomes essential to investigate this theme in a structured manner, recognizing that such events can serve as pillars for achieving Target 4.4 of SDG 4, which aims to substantially increase the number of youth with rele-vant skills for employment. Conducting an integrative review allows for the aggregation of available scientific evidence, the identification of gaps in knowledge, and the proposal of guidelines to strengthen the integration of innovative practices in higher education. Furthermore, these collaborative spaces contribute to forming more critical, reflective professionals prepared to act in a scenario of constant technological and social transformation, aligning with contemporary demands in healthcare and engineering.

Despite the growing valuation of active methodologies and learning ecosystems, there is a significant gap in the literature regarding the understanding of student-led scientific events as structured pedagogical devices. In particular, studies analyzing them as complex learning ecosystems that integrate pedagog-ical, organizational, and innovation dimensions are lacking.

II. RESEARCH PROBLEM

Despite advancements in teaching methodologies within Biomedical Engineering, a significant gap remains in the literature regarding the analysis of the impact and effectiveness of student-led scientific events as structured formative spaces. Although these events are increasingly frequent, their contri-bution to academic and professional development is seldom investigated systematically, particularly concerning the lack of clear guidelines for their implementation, monitoring, and evaluation.

Within this context, there is a clear need to understand how these events contribute to the development of technical and socio-emotional skills—such as leadership, communication, and teamwork—as well as to identify the primary challenges faced during their organization and management, including financial, structural, and institutional limitations. Furthermore, it is relevant to discuss which practices have proven most effective and how these spaces can be improved to maximize their integration with teaching, research, and the job market, thereby amplifying their impact on student training.

Consequently, the relevance of this study is linked to the necessity of generating knowledge that supports educational institutions, faculty, and students in understanding the trans-formative potential of these events. By advancing the sys-tematization of evidence and proposing parameters for their organization and evaluation, this work aims to contribute to consolidating these spaces as more structured, replicable learning ecosystems aligned with the contemporary demands of Biomedical Engineering education.

III. METHODOLOGY

A. Type of Research

This study was conducted as an integrative literature review based on the methodological recommendations of Whittemore and Knafl [27], Mendes, Silveira and Galvão [13], and Sousa et al. [25]. This approach was selected because it allows the integration of findings from qualitative, quantitative, mixed-methods, and descriptive studies. The review focused on student-led scientific events as learning ecosystems in Biomedical Engineering education, particularly their relationship with active learning, experiential learning, innovation, student leadership, scientific communication, competency development, and institutional support.

B. Research Question

The review was guided by the following question: How do student-led scientific events contribute to learning, competency development, and innovation in Biomedical Engineering education? The analysis considered the

organization and format of events, students' active participation, educational strategies, mentoring processes, technical and socioemotional competencies, and the institutional conditions that support or limit these initiatives.

C.Inclusion and Exclusion Criteria

Studies were included when they were full-text, peer-reviewed articles published between January 2023 and April 2026 in English, Portuguese, or Spanish. Eligible studies addressed student-led scientific or academic events, student research activities, active or experiential learning, scientific communication, leadership, mentoring, pedagogical innovation, or competency development in higher education, particularly in Biomedical Engineering, Bioengineering, Engineering, or related STEM fields. Editorials, opinion papers, conference abstracts, books, theses, dissertations, and studies that did not directly address the research question were excluded. Theses and dissertations were consulted only as complementary theoretical sources when relevant, without being included in the final sample.

D. Justification of the Period

The period from 2023 to April 2026 was selected to identify recent educational practices and emerging approaches associated with student participation, scientific events, digital transformation, interdisciplinary learning, and innovation in Biomedical Engineering education. Nevertheless, foundational publications published before 2023 were considered to support the theoretical discussion of active learning, experiential learning, engineering education, student leadership, and competency-based education. These earlier studies were used as conceptual references and were not necessarily included in the final sample.

E. Databases and Search Strategy

The searches were conducted in PubMed, SciELO, Scopus, Web of Science, IEEE Xplore, Engineering Village, and Google Scholar. The search strategy combined descriptors and related terms such as "student-led scientific events," "student-led academic events," "student-led conferences," "student research conferences," "biomedical engineering education," "bioengineering education," "engineering education," "active learning," "experiential learning," "transformative learning," "pedagogical innovation," "scientific communication," "student leadership," "mentoring," and "competency development." The Boolean operators AND and OR were used to combine the terms. The search procedures, databases, descriptors, and date of access were recorded to improve transparency and reproducibility.

F. Identification and Selection of Studies

The identification and selection process occurred in four stages. First, the studies were retrieved from the selected databases. Second, duplicate records were removed. Third, titles and abstracts were screened according to the previously established eligibility criteria. Finally, potentially relevant studies were assessed through full-text reading. Studies were excluded at the full-text stage when they did not address student-led scientific or educational practices, did not focus on higher education, lacked sufficient methodological information, or did not contribute to the research question. The selection process was organized according to the principles of PRISMA, with records of the number of identified, screened, excluded, and included studies.

G. Data Extraction

Data were extracted using a structured instrument developed for this review. The extracted information included authors, publication year, country, study design, educational context, participant characteristics, type and format of the scientific event, level of student leadership, learning strategies, mentoring activities, interdisciplinary collaboration, competencies developed, reported educational outcomes, assessment instruments, barriers, facilitating factors, methodological limitations, and DOI or other identification information. When available, the review also recorded whether the outcomes were assessed through qualitative perceptions, self-reported questionnaires, academic performance indicators, validated instruments, or objective measures of competency development.

H. Critical Appraisal

The methodological quality of the included studies was assessed according to their respective designs. Qualitative studies were examined regarding the clarity of their objectives, theoretical framework, participant selection, data collection, analytical procedures, and consistency between findings and interpretations. Quantitative studies were assessed according to sample description, measurement procedures, validity of instruments, statistical analysis, and clarity of results. Mixed-methods studies were evaluated based on the integration between qualitative and quantitative components. Descriptive studies and experience reports were analyzed for the clarity of the educational intervention, contextual description, evidence supporting the reported outcomes, and limitations. This assessment was used to interpret the strength and reliability of the available evidence.

I.Risk of Bias

Potential risks of bias were considered throughout the analysis, including selection bias, self-selection of participants, response bias, recall bias, publication bias, institutional bias, and limitations related to self-reported outcomes. Particular attention was given to studies conducted in a single institution, studies with small or non-probabilistic samples, and reports based mainly on students' perceptions. The review also distinguished between short-term outcomes, such as satisfaction, engagement, and perceived learning, and more objective or long-term outcomes, such as academic performance, sustained competency development, professional identity, and employability. These aspects were considered when comparing findings and discussing their generalizability.

J.Synthesis of Results

The findings were organized through thematic synthesis

into four categories: scientific event formats and student leadership; active, experiential, and collaborative learning processes; technical and socioemotional competency development; and institutional conditions, barriers, facilitating factors, and evaluation strategies. The results were compared according to similarities, differences, methodological characteristics, and the strength of the evidence. The synthesis also identified limitations in the existing literature, including the predominance of qualitative and self-reported data, the lack of standardized assessment instruments, the limited evaluation of long-term outcomes, and the restricted generalizability of single-institution studies. These gaps were used to indicate directions for future research on student-led scientific events as transformative learning ecosystems in Biomedical Engineering education.

IV. RESULTS

The initial database search identified 31 records. After title and abstract screening, four studies were excluded because they did not meet the eligibility criteria. Following full-text assessment, 27 studies were included in the final integrative review. The selection process should be presented in a PRISMA flowchart, detailing the number of records identified, duplicates removed, studies screened, full texts assessed, exclusions, and studies included in the final synthesis.The study selection process is illustrated in Fig. 1.

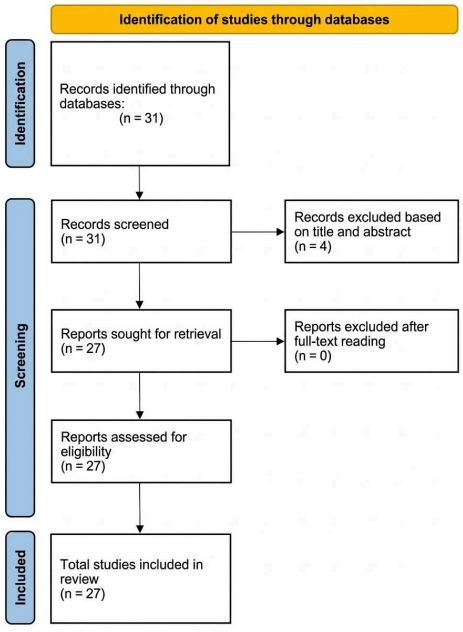


Fig. 1. PRISMA flow diagram detailing the identification, screening, and inclusion process of the reviewed studies.

An overall synthesis of the 27 selected studies was conducted to analyze their primary scope and contributions to the field. A detailed overview of each included study, including references, titles, and main focus, is presented in Table I.

TABLE I. Summary of the 27 Included Studies in the Integrative Review.

Ref.	Authors & Year	Study Title	Document Type / Primary Focus Scope
[1]	J. Amann *et al.* (2020)	*Explainability for artificial intelligence in health-care: A multidisciplinary perspective*	Multidisciplinary perspective on AI explainability in healthcare systems.
[2]	C. S. Ang (2025)	*Developing AI literacy in healthcare education*	Educational strategies for artificial intelligence literacy in health training.
[3]	K. Billiar *et al.* (2022)	*Learning environments and evidence-based practices in bioengineering and biomedical engineering*	Evidence-based pedagogical practices and learning environments in BME education.
[4]	Brazilian Ministry of Health (2020)	*Digital Health Strategy for Brazil 2020–2028*	Strategic institutional framework for national digital health implementation.
[5]	M. M. Da Silva (2024)	*Innovation in organizational and educational strategies in nursing: An integrative review*	Integrative literature review on educational and organizational innovations.
[6]	A. B. Dogan	*Feasibility of integrating*	Feasibility study on medical student

	(2026)	*medical student mentors into undergraduate biomedical engineering research*	mentorship within BME research.
[7]	S. Freeman *et al.* (2014)	*Active learning increases student performance in science, engineering, and mathematics*	Meta-analysis evaluating active learning efficacy in STEM education.
[8]	L. H. Jamieson & J. R. Lohmann (2012)	*Innovation with Impact: Creating a Culture for Scholarly and Systematic Innovation in Engineering Education*	Institutional report on systematic innovation culture in engineering education.
[9]	F. Jiang *et al.* (2017)	*Artificial intelligence in healthcare: Past, present and future*	Comprehensive literature review on AI development and applications in healthcare.
[10]	C. J. Kelly *et al.* (2019)	*Key challenges for delivering clinical impact with artificial intelligence*	Critical analysis of barriers to clinical translation of AI solutions.
[11]	L. R. Lattuca *et al.* (2017)	*Engineering curricula and student learning: Development of competencies*	Curriculum design impact on student competency development in engineering.
[12]	J. L. Leight *et al.* (2025)	*Creating inclusive environments to support learning and professional development in biomedical engineering education*	Inclusivity and professional development strategies in BME programs.
[13]	K. D. S. Mendes *et al.* (2008)	*Integrative review: Research method for incorporating evidence in health and nursing*	Methodological guidelines for integrative literature reviews.
[14]	L. Montesinos *et al.* (2023)	*Transdisciplinary experiential learning in biomedical engineering education for healthcare systems improvement*	Experiential and transdisciplinary learning model in healthcare-focused BME.
[15]	National Academy of Engineering (2010)	*The Engineer of 2020: Visions of Engineering in the New Century*	Strategic vision report on future competencies and engineering roles.
[16]	OECD (2026)	*OECD Digital Education Outlook 2026*	Global policy benchmark and trends in digital educational practices.
[17]	United Nations (2015)	*Transforming Our World: The 2030 Agenda for Sustainable Development*	Global development framework establishing the Sustainable Development Goals.

[18]	B. W. L. Packard *et al.* (2025)	*Synergy as a strategy to strengthen biomedical mentoring ecosystems*	Framework for building synergistic mentoring ecosystems in biomedical fields.
[19]	J. M. Parker & D. Amorelli (2024)	*Oral presentation abstracts, 27th annual conference of the IAMSE*	Compiled conference proceedings on medical science education practices.
[20]	C. Peng (2025)	*Interdisciplinary integration in undergraduate education: The practice of biomedical engineering program in a Chinese STEM university*	Practical implementation of interdisciplinary integration in BME curricula.
[21]	M. Prince & R. Felder (2006)	*Inductive teaching and learning methods: Definitions, comparisons, and research bases*	Review of definitions, comparisons, and evidence bases for inductive teaching.
[22]	M. Richards *et al.* (2026)	*Implementation of a biomedical research training program for early-stage undergraduate students to inspire development of a STEM professional identity*	Program implementation evaluation for early STEM identity development.
[23]	C. L. A. Santana *et al.* (2025)	*Advancing equity and diversity in engineering education through community-based participatory research*	Community-based participatory research applications for equity in engineering.
[24]	A. Santos-Díaz *et al.* (2024)	*Implementing a challenge-based learning experience in a blended bioinstrumentation course*	Course implementation study on challenge-based learning in bioinstrumentation.
[25]	L. M. M. Sousa *et al.* (2017)	*The methodology of integrative literature review in nursing*	Methodological framework for executing integrative literature reviews.
[26]	E. J. Topol (2019)	*High-performance medicine: The convergence of human and artificial intelligence*	Conceptual review on human and artificial intelligence convergence in clinical care.
[27]	R. Whittemore & K. Knafl (2005)	*The integrative review: Updated methodology*	Foundational methodological standards for integrative literature review synthesis.

The included studies were predominantly conducted in international settings, especially in the United States, the United Kingdom, and China, with additional contributions from Latin American contexts. The methodological approaches were mainly qualitative, mixed-methods, descriptive, and exploratory. This pattern indicates that the field is still largely focused on understanding students' experiences, perceptions, engagement, and educational

processes rather than measuring long-term outcomes through standardized or objective instruments. Recent Biomedical Engineering literature emphasizes that student engagement involves behavioral, cognitive, and emotional dimensions and is influenced by active learning, mentoring, belonging, institutional support, and the organization of educational experiences.

The thematic analysis resulted in four categories: (i) characteristics and formats of student-led scientific events; (ii) learning, engagement, and competency development; (iii) mentoring, inclusion, and collaborative learning ecosystems; and (iv) challenges, enabling factors, and sustainability conditions. This organization expands the previous three-category structure by distinguishing the educational effects of the events from the institutional and relational conditions required for their implementation.

A. Characteristics and Formats of Student-Led Scientific Events

The studies identified different formats of student-led academic and scientific activities, including conferences, symposia, scientific workshops, hackathons, research presentations, interdisciplinary challenges, and project-based initiatives. These activities generally involve students in the planning, organization, communication, scientific production, and presentation of results. Their educational value is associated not only with participation in the event itself, but also with the process of preparing abstracts, developing projects, discussing research findings, presenting oral or poster communications, and interacting with peers, faculty members, and external professionals.

The reviewed literature indicates that these formats promote the integration of theoretical knowledge with practical and real-world problems. In Biomedical Engineering, this integration is particularly relevant because students must articulate knowledge from biology, medicine, engineering, mathematics, technology, and clinical practice. Student-led workshops and research activities may also reduce barriers to participation by offering structured opportunities for students to access scientific training, develop research skills, and establish relationships with mentors. For example, the Starting Undergraduate Research Experience initiative used interactive workshops, peer mentoring, and student leadership to support early undergraduate students and facilitate access to future research opportunities.

B.Learning, Engagement, and Competency Development

The findings suggest that student-led scientific events contribute to both technical and socioemotional development. Technical competencies include scientific writing, literature searching, research planning, data interpretation, scientific presentation, communication of results, and the application of specialized Biomedical Engineering knowledge. These activities also encourage students to understand the relationship between scientific evidence, technological innovation, and healthcare needs.

The development of socioemotional competencies was associated with leadership, teamwork, interpersonal communication, organization, autonomy, creativity, critical thinking, and problem-solving. By assuming responsibilities in the organization and execution of scientific activities, students move from passive recipients of information to active participants in the production and dissemination of knowledge. This finding is consistent with evidence supporting student-centered and active learning strategies as mechanisms for increasing participation, engagement, and learning opportunities in engineering education.

The results also indicate that engagement should not be understood only as attendance or participation in an event. It includes behavioral involvement, cognitive investment, reflection, self-regulation, interest, sense of belonging, and emotional connection with the educational environment. Therefore, student-led scientific events may function as learning ecosystems when they combine scientific production, collaborative interaction, mentoring, reflection, and opportunities for students to exercise autonomy and responsibility.

C.Mentoring, Inclusion, and Collaborative Learning Ecosystems

Mentoring and peer collaboration emerged as important elements for the success of student-led initiatives. The analyzed studies indicate that students benefit when they receive guidance from faculty members, more experienced peers, researchers, and institutional partners. Mentoring may support scientific learning, improve confidence, strengthen professional identity, and facilitate access to research networks and future academic or professional opportunities.

The reviewed evidence also highlights the importance of inclusive environments. Structured research activities, peer mentoring, and welcoming learning spaces can reduce inequalities in access to scientific experiences, particularly among students who have limited previous contact with laboratories, research groups, or professional networks. The literature on Biomedical Engineering education emphasizes that belonging, support, and participation are associated with students' engagement and academic development.

In addition, collaborative learning ecosystems extend beyond the event itself. They may involve partnerships among students, faculty members, research groups, professional organizations, innovation centers, and external institutions. Such networks create opportunities for knowledge exchange, interdisciplinary collaboration, scientific dissemination, and the development of projects connected to social and healthcare demands.

D.Challenges, Enabling Factors, and Sustainability

The main challenges identified were limited funding, insufficient infrastructure, lack of dedicated spaces, restricted access to technological resources, workload, dependence on voluntary student participation, and difficulties in maintaining institutional continuity. The sustainability of student-led initiatives may also be affected when activities depend on a small group of students or on individual faculty members without formal institutional mechanisms.

Conversely, the main enabling factors included faculty support, institutional recognition, mentoring, partnerships, access to laboratories and technological resources, collaborative networks, and the integration of teaching, research, and outreach. Evidence from studies on student-led initiatives indicates that mentoring and networking are perceived as high-impact mechanisms, whereas formal institutional structures and dedicated facilities may vary in availability and effectiveness.

The results further suggest that institutional support must include more than authorization to conduct an event. It should involve financial resources, access to infrastructure, training, dissemination support, recognition of student and faculty participation, and mechanisms for continuity between academic cycles. In Biomedical Engineering education, these conditions are especially important because scientific activities frequently require specialized equipment, interdisciplinary supervision, ethical considerations, and interaction with healthcare-related problems.

E. Overall Interpretation of the Findings

Overall, the evidence indicates that student-led scientific events can operate as transformative learning ecosystems by combining active participation, scientific communication, collaborative problem-solving, mentoring, and professional development. Their contribution is not limited to the acquisition of technical knowledge; they also support autonomy, leadership, teamwork, confidence, scientific identity, and the connection between academic learning and professional practice.

However, the strength of this conclusion must be interpreted cautiously. Much of the available evidence is based on qualitative reports, exploratory studies, institutional experiences, and self-reported perceptions. There remains a limited number of studies using validated instruments, control or comparison groups, longitudinal designs, and objective indicators of academic performance, competency development, employability, or professional insertion. The literature also presents limitations related to small samples, single-institution contexts, self-selection of participants, and differences in how engagement and learning outcomes are defined.

Therefore, although student-led scientific events show strong potential for promoting innovation and student development in Biomedical Engineering, further research is required to establish standardized evaluation models and to measure their short-, medium-, and long-term effects.

V. DISCUSSION

This integrative review demonstrates that student-led sci-entific events are consolidating as relevant spaces for skill development in higher education, especially in interdisci-plinary fields such as Biomedical Engineering. In light of active learning theories, these environments favor student participation as the central agent of the educational process, promoting greater engagement, autonomy, and meaningful knowledge construction [7], [21]. The literature points out that collaborative and interactive strategies directly contribute to better academic outcomes and to training more critical professionals prepared for complex challenges. In this regard, the discussion highlights that such ecosystems are fundamental to operationalizing the goals of SDG 3, 4, and 9 within daily academic practices.

When analyzing the historical evolution of engineering education, it is evident that initial innovation proposals were centered on curricular reformulation, seeking to align aca-demic training with market demands. Classical studies had already indicated the limitations of the traditional lecture-based model, highlighting the need to incorporate practical experiences and extracurricular activities. Currently, an advancement toward the conception of integrated educational ecosystems is observed, in which scientific events, hackathons, and technological challenges expand formative possibilities and strengthen interdisciplinarity [6], [8], [11], [15]. The anal-ysis of the studies demonstrates convergence regarding the role of these events as practical learning platforms for developing skills such as critical thinking and problem-solving.

However, there are divergences regarding how to measure these impacts, especially between international and national studies. While international literature tends to present more structured and quantitative evaluations, Brazilian studies still concentrate mostly on descriptive reports, which limits the comparability of results and the consolidation of robust evi-dence. This methodological disparity suggests a challenge for Target 9.5 of SDG 9, which encourages scientific research and an increase in the number of researchers, demanding greater analytical rigor in national productions by 2026 [2], [5], [19].

A critical appraisal of the 27 reviewed studies reveals significant methodological limitations and divergence in findings. The literature is predominantly reliant on cross-sectional, self-reported perception surveys, which introduces response, recall, and social desirability biases. Furthermore, potential publication bias is evident, as existing studies overwhelmingly report successful implementations and positive student outcomes, rarely documenting failed event models, logistical bottlenecks, or low engagement rates. Generalizability across diverse institutional and cultural contexts also remains constrained: educational frameworks implemented in highly funded North American or European institutions cannot be directly extrapolated to developing academic settings, such as Brazilian public universities, where infrastructure, funding availability, and integration with health systems (e.g., SUS) present unique operational constraints.

Regarding the implications for the Biomedical Engineering curriculum, the results indicate the need for greater integration between formal and non-formal learning activities. Incorporat-ing scientific events as complementary pedagogical strategies can contribute to developing skills that are not always fully addressed in traditional courses. Furthermore, the importance of faculty training focused on active methodologies and the creation of institutional policies that encourage student lead-ership becomes evident. The sustainability of these initiatives depends directly on institutional support, such as infrastructure and continuous funding—elements that still present challenges in the Brazilian context compared to the international scenario. Another relevant aspect concerns the dynamic nature of these learning ecosystems, which promote interaction among students, faculty, researchers, and the productive sector. This articulation strengthens interdisciplinary training and brings students closer to real-world demands in the

field. Moreover, these environments contribute to developing socio-emotional skills, such as leadership and adaptability, which are essential for contemporary professional practice. Despite the advance-ments, the review highlighted significant limitations, such as the scarcity of longitudinal studies and the lack of standardized instruments for impact evaluation.

Assessing the objective impact of student-led events on technical skill acquisition, academic performance, and long-term career insertion poses substantial evaluative challenges. Most current studies fail to isolate the specific causal effect of extracurricular event participation from standard curricular progression. To overcome these limitations, future research must transition from non-standardized feedback forms to validated measurement instruments. Competency growth can be quantified using established frameworks such as the AAC&U VALUE Rubrics (for critical thinking, teamwork, and problem-solving), the Motivated Strategies for Learning Questionnaire (MSLQ), or validated STEM Professional Identity Scales. Implementing pre- and post-test designs paired with control groups would provide the empirical rigor necessary to validate these informal learning spaces.

Based on the findings of this review, a conceptual model is proposed that conceptualizes student-led scientific events as transformative learning ecosystems structured across four interdependent dimensions: (i) inputs, which include institu-tional support, financial resources, and academic mentoring (ii) processes, related to the adoption of active methodolo-gies, interdisciplinary collaboration, and real-world problem-solving; (iii) products, represented by the development of tech-nical and socio-emotional skills; and (iv) outcomes, associated with qualified professional training, innovation, and impact on healthcare systems. This model contributes to understanding these events not merely as extracurricular activities, but as structured pedagogical devices with the potential to transform Biomedical Engineering education.

VI. CONCLUSION

This integrative review synthesized evidence reinforcing the role of student-led scientific events as transformative ecosystems in Biomedical Engineering education. From this synthesis, key lessons learned emerge regarding their pedagogical value: first, student leadership in event management shifts learners from passive knowledge consumers to active educational architects; second, robust mentoring ecosystems involving faculty, senior peers, and industry partners are essential to sustain event quality and student engagement; and third, these events serve as vital practical devices bridging technical bioengineering concepts with real-world healthcare demands, thereby directly advancing SDG target alignments.

Despite these contributions, critical research gaps remain in the current literature. Most available evidence relies on short-term, single-institution qualitative reports, leaving a major gap regarding the long-term impact of these events on academic retention, graduation rates, and professional career insertion. Furthermore, there is a notable absence of multicenter comparative studies and quantitative evaluation models capable of measuring the specific educational return on investment or skill acquisition isolated from standard curricular progression. To address these gaps, future research priorities must focus on implementing longitudinal cohort designs, multicenter evaluations across diverse institutional and socioeconomic contexts, and validated quantitative measurement models to rigorously substantiate the transformative potential of student-led learning ecosystem.

REFERENCES

[1] J. Amann et al., "Explainability for artificial intelligence in health-care: A multidisciplinary perspective," *BMC Medical Informat-ics and Decision Making*, vol. 20, no. 310, 2020. Available: https://doi.org/10.1186/s12911-020-01332-6

[2] C. S. Ang, "Developing AI literacy in healthcare education," *BMC Medical Education*, vol. 25, no. 1, pp. 1–10, 2025. Available: https://doi.org/10.1186/s12909-025-04212-z

[3] K. Billiar et al., "Learning environments and evidence-based prac-tices in bioengineering and biomedical engineering," *Biomedi-cal Engineering Education*, vol. 2, pp. 1–16, 2022. Available: https://doi.org/10.1007/s43683-021-00062-z

[4] Brazilian Ministry of Health, *Digital Health Strategy for Brazil 2020–2028*, Ministry of Health, 2020. Available: https://www.gov.br/saude/pt-br/assuntos/saude-digital

[5] M. M. Da Silva, "Innovation in organizational and educational strategies in nursing: An integrative review," *Revista Brasileira de Enfermagem*, vol. 78, p. e20240001, 2024. Available: https://doi.org/10.1590/0034-7167-2024-0001

[6] A. B. Dogan, "Feasibility of integrating medical student mentors into undergraduate biomedical engineering research," *Biomedical En-gineering Education*, vol. 6, no. 1, pp. 1–10, 2026. Available: https://doi.org/10.1007/s43683-025-00203-8

[7] S. Freeman et al., "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 111, no. 23, pp. 8410–8415, 2014. Available: https://doi.org/10.1073/pnas.1319030111

[8] L. H. Jamieson and J. R. Lohmann, *Innovation with Impact: Creating a Culture for Scholarly and Systematic Innovation in Engineering Education*, American Society for Engineering Education, 2012.

[9] F. Jiang et al., "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.
Available: https://doi.org/10.1136/svn-2017-000101

[10] C. J. Kelly et al., "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 195, 2019. Available: https://doi.org/10.1186/s12916-019-1426-2

[11] L. R. Lattuca et al., "Engineering curricula and student learning: Development of competencies," *Higher Education*, vol. 74, pp. 1–15, 2017. Available: https://doi.org/10.1007/s10734-017-0129-3

[12] J. L. Leight et al., "Creating inclusive environments to support learning and professional development in biomedical engineering education," *Biomedical Engineering Education*, vol. 5, pp. 121–134, 2025. Avail-able: https://doi.org/10.1007/s43683-025-00189-3

[13] K. D. S. Mendes, R. C. C. P. Silveira, and C. M. Galva˜o, "Integrative review: Research method for incorporating evidence in health and nursing," *Texto & Contexto - Enfermagem*, vol. 17, no. 4, pp. 758–764, 2008. Available: https://doi.org/10.1590/S0104-07072008000400018

[14] L. Montesinos, D. E. Salinas-Navarro, and A. Santos-Diaz, "Transdis-ciplinary experiential learning in biomedical engineering education for healthcare systems improvement," *BMC Medical Education*, vol. 23, no. 207, 2023. Available: https://doi.org/10.1186/s12909-023-04171-x

[15] National Academy of Engineering, *The Engineer of 2020: Visions of Engineering in the New Century*, The National Academies Press, 2010. Available: https://doi.org/10.17226/12635

[16] OECD, *OECD Digital Education Outlook 2026*, OECD Publish-ing, 2026. Available: https://www.oecd.org/en/publications/oecd-digital-education-outlook-2026_062a7394-en.html

[17] United Nations (UN), *Transforming Our World: The 2030 Agenda for Sustainable Development*, UN, 2015. Available: https://sdgs.un.org

[18] B. W. L. Packard, B. L. Montgomery, and J. L. Mondisa, "Syn-ergy as a strategy to strengthen biomedical mentoring ecosystems," *Frontiers in Education*, vol. 10, no. 1474063, 2025. Available: https://doi.org/10.3389/feduc.2025.1474063

[19] J. M. Parker and D. Amorelli, "Oral presentation abstracts, 27th annual conference of the International Association of Medical Science Educa-tors," Medical Science Educator, vol. 33, pp. 1–100, 2024. Available: https://doi.org/10.1007/s40670-024-01981-6

[20] C. Peng, “Interdisciplinary integration in undergraduate education: The practice of biomedical engineering program in a Chinese STEM univer-sity,” Engineering Education Review, vol. 1, no. 1, p. 878, 2025.

[21] M. Prince and R. Felder, “Inductive teaching and learning meth-ods: Definitions, comparisons, and research bases,” Journal of En-gineering Education, vol. 95, no. 2, pp. 123–138, 2006. Available: https://doi.org/10.1002/j.2168-9830.2006.tb00809.x

[22] M. Richards et al., “Implementation of a biomedical research training program for early-stage undergraduate students to inspire development of a STEM professional identity,” Biomedical Engineering Education, vol. 6, no. 1, pp. 1–10, 2026. Available: https://doi.org/10.1007/s43683-026-00219-8

[23] C. L. A. Santana et al., “Advancing equity and diversity in engi-neering education through community-based participatory research,” Frontiers in Education, vol. 10, no. 1696510, 2025. Available: https://doi.org/10.3389/feduc.2025.1696510

[24] A. Santos-D´ıaz et al., “Implementing a challenge-based learning experi-ence in a blended bioinstrumentation course,” BMC Medical Education, vol. 24, no. 510, 2024. Available: https://doi.org/10.1186/s12909-024-05462-7

[25] L. M. M. Sousa et al., “The methodology of integrative literature review in nursing,” Revista Investigac¸a˜o Enfermagem, vol. 2, no. 21, pp. 17–26, 2017.

[26] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019. Available: https://doi.org/10.1038/s41591-018-0300-7

[27] R. Whittemore and K. Knafl, “The integrative review: Updated method-ology,” Journal of Advanced Nursing, vol. 52, no. 5, pp. 546–553, 2005. Available: https://doi.org/10.1111/j.1365-2648.2005.03621.

The Paradox of AI-Assisted Academic Writing in Biomedical Engineering and Health Sciences Research: A Critical, Historical, and Predictive Approach

Rafael Fabre Rodrigues e Souza
Faculty of Technology (FT)
State University of Campinas (UNICAMP)
Limeira, São Paulo, Brazil
E-mail: rafaelfabre.rs@gmail.com

Cíntia Cristina Soares
School of Technology (FT)
State University of Campinas (UNICAMP)
Limeira, SP, Brazil
E-mail: soares.cristina.cintia@gmail.com

***Abstract*—The growing adoption of generative Artificial Intelligence (AI), such as Large Language Models (LLMs), has introduced a disruptive paradigm in scientific communication, with sensitive implications for Biomedical Engineering and the Health Sciences, fields in which published knowledge serves as direct input for medical devices, diagnostic algorithms, and clinical protocols. This scenario exposes the paradox of assisted writing: productivity gains contrast with risks of authorial homogenization, factual hallucinations, and the erosion of academic integrity. This work investigates this paradox in the submission processes of conferences and journals in Biomedical Engineering and the various Health Sciences fields, from a historical and predictive perspective, through a literature review and critical analysis of empirical studies on AI detection systems, focusing on false positives, systemic failures, and resilience mechanisms used to evade plagiarism detection. The results show that verification technologies are statistically unstable: both algorithms and human reviewers fail to reliably distinguish organic from synthetic text, unjustly accusing authentic researchers. A survey of biomedical literature indexed in PubMed identified a tenfold increase in the rate of fabricated references, surging to one in fewer than 280 papers by early 2026, evidencing measurable contamination of the evidence base in health sciences. Unregulated AI adoption also exacerbates structural disparities in science, and absolute prohibition policies prove ineffective. Projecting the future of publishing through 2035, this study concludes that the "Academic Ecosystem" must transition toward guidelines of radical transparency and ethical-use regulation. This shift is essential to preserve the reliability of the knowledge base that underpins the transfer of technology from academia to the market. To this end, evaluation criteria must safeguard researchers' critical reasoning and methodological accountability.**

***Index Terms*—Generative Artificial Intelligence, Academic Integrity, LLM, Biomedical Engineering, Synthetic Content Detection, Scientific Writing.**

I. Introduction

The evolution of scientific communication and research support tools has reached a disruptive milestone with the adoption of generative Artificial Intelligence and Large Language Models (LLMs), altering the foundations and validation processes of academic knowledge synthesis [1]. These architectures use natural language processing to generate text with fluency and coherence, mimicking human writing to the point of challenging even experienced evaluators to distinguish it [2].

The substantial productivity gains in scientific writing have, however, introduced the paradox of AI-assisted writing, in which technological efficiency directly contrasts with the risks of homogenized thinking, loss of authorial integrity, and the emergence of synthetic plagiarism [3]. The rationale for investigating this phenomenon lies in the tension experienced by journals and scientific conference committees, particularly within Biomedical Engineering and the Health Sciences, which face a growing volume of submissions containing factual inaccuracies or algorithmic hallucinations, often masked by scientifically plausible rhetoric [4].

In response, automated tools for detecting AI-generated content have been developed, but the stability and effectiveness of these platforms remain statistically questionable. This fragility translates into frequent false positives, capable of unjustly incriminating the original work of human researchers and deepening the disparities already inherent to the publishing ecosystem [5]. Strict prohibition of these tools disregards their potential for co-creation, requiring the community to urgently transition toward policies grounded in transparency and responsible, ethical use [6].

This work aims to investigate the paradox of generative AI-assisted academic writing within the scientific community, focusing on submission processes in Biomedical Engineering conferences and journals. To this end, the study adopts a historical and predictive approach. The research analyzes technological evolution, evaluates the methodological flaws of current detection systems, and discusses the ethical implications for the future of scientific publishing. Finally, it proposes that evaluation guidelines be adapted to preserve critical reasoning in the face of inevitable automation.

II. Methodology

To comprehensively investigate the impact of LLMs on academic integrity, this study adopted a narrative evidence-synthesis methodology rather than a systematic review protocol. Sources were identified through exploratory searches in the Google Scholar, PubMed, and Web of Science databases, conducted between February and July 2026. The search expressions combined key terms, including "Generative AI," "ChatGPT," and "Large Language Models," crossed with "academic integrity," "peer review," "authorship," and "scientific publishing." Sources were selected based on their relevance to the paradox investigated, prioritizing peer-reviewed articles, institutional guidelines, and official regulatory documents that directly addressed the ethical implications, structural impacts, or detection of synthetic content in academic writing. A total of 145 sources were retrieved through the initial database searches, of which twelve core references were retained for qualitative analysis [1]–[12]. This narrative approach, rather than a systematic protocol with fixed inclusion and exclusion screening at each stage, is acknowledged as a methodological limitation in Section V.

The review was organized into three main analytical axes: an analytical and historical review of documentary evolution, a critical synthesis of empirical studies on detection tools, and an analysis of the incidence of false positives in AI detection systems [4].

Important methodological clarification: this work did not conduct its own empirical tests with AI detectors or synthetic text generation. Instead, it adopted an evidence-synthesis approach, compiling and critically analyzing the results of empirical studies already published in high-impact scientific literature [2], [3]. This methodological choice enabled a comprehensive and comparative view of the state of the art without the need to replicate already robustly documented experiments.

Initially, a methodological mapping was carried out of the transition from traditional documentary support technologies to current natural language processing architectures, contextualizing the ethical implications, copyright restrictions, and emerging editorial policies in light of these innovations [5]. This theoretical foundation allowed for an in-depth structuring of the debate on the paradox of human-machine co-creation in the scientific writing process [1].

In the second axis, the research focused on the critical synthesis of empirical studies on originality-verification technologies designed to identify synthetic content. The study was based on an analysis of the resilience of traditional and AI-focused detectors against evasion strategies, that is, techniques employed by authors to bypass detection, such as successive text rewriting by another AI (recursive paraphrasing) and linguistic prompt optimization [3]. Based on data compiled from the literature, the performance of algorithmic classifiers was examined when confronted with mixed datasets composed of rigorously authentic research abstracts and artificially generated abstracts [2].

The third axis consisted of investigating the intrinsic limitations of algorithmic behavior and human evaluation, focusing on classification failure rates. Trials were synthesized in which blinded human reviewers, together with plagiarism detection platforms, evaluated academic material to identify the machine's signature [2]. Attention was primarily directed toward quantifying the impact of the false-positive phenomenon, an alarming situation in which entirely human-authored texts are unjustly flagged as synthetic plagiarism [2].

This methodological triangulation, crossing the evolutionary history of artificial intelligence with the synthesis of practical tests of its effectiveness and detection, provided the necessary substrate to understand the instability of current control mechanisms and to ground a responsible approach to the use of AI within the publishing ecosystem [6].

III. Results

A. Historical Evolution of Academic Writing Tools

To grasp the magnitude of the current paradox, it is essential to trace the historical evolution of writing technologies in academia, which can be periodized into three distinct eras:

Era 1: Typographic and Mechanical Assistance (1960-1995). In this phase, tools such as the typewriter and, later, the first word processors (WordStar, Microsoft Word), operated exclusively in formatting and editing. The researcher was the sole generator of ideas and text. The technology reduced physical effort and facilitated revisions but did not alter the genesis of scientific thought. Authorship was unquestionably human.

Era 2: Information Retrieval and Originality Policing (1995-2022). With the internet and the digitization of journals, the challenge shifted from production to information retrieval and verification. Unlimited access to articles gave rise to the problem of digital plagiarism and "salami science." Academia's historical response was the development of textual similarity detection software (Turnitin, iThenticate, Copyscape). The focus of the technology was to control originality, requiring the human to prove that the text was theirs through the absence of overlap with existing databases. Authorship remained exclusively human but now had to be "certified" by algorithms.

Era 3: Probabilistic Generation and Algorithmic Co-authorship (2022-Present). Generative AI broke the previous paradigm. The tool no longer merely retrieves or formats text; it generates text based on statistical probabilities learned from vast text corpora. The paradox sets in: the very tool used to "police" originality can now perfectly simulate it. Academia has moved from a posture of assistance to a crisis of authorship, in which the boundary between "productivity tool" and "ghost co-author" becomes systemically blurred.

B. Instability of Detection Systems and False Positives

The results extracted from the synthesis of empirical studies reveal that the ability to discriminate between human writing and synthetic content is highly flawed, both for human evaluators and algorithmic systems [3]. In a blinded evaluation involving 50 original and 50 ChatGPT-generated abstracts from five high-impact medical journals (*Nature Medicine*,

JAMA, *NEJM*, *BMJ*, and *The Lancet*), four human reviewers correctly identified only 68% of artificially generated texts and, alarmingly, incorrectly classified 14% of strictly authentic abstracts as AI-produced [2]. In the same study, an automated AI output detector achieved an area under the ROC curve of 0.94, indicating that algorithmic tools, while imperfect, outperformed unaided human judgment under these specific experimental conditions [2]. This documented incidence of false positives in the literature evidences a harmful distrust instilled in the peer-review process, in which human originality is frequently placed under unwarranted suspicion [2].

Additionally, contemporary synthetic-content detection tools were found to exhibit acute vulnerabilities and drastic drops in effectiveness when confronted with structured evasion techniques. In a controlled experiment, an 11-billion-parameter paraphrasing model (DIPPER) was used to rewrite text generated by large language models including GPT3.5; this evasion strategy collapsed the detection accuracy of the DetectGPT classifier from 70.3% to 4.6% at a fixed false-positive rate of 1%, without appreciably altering the semantic content of the text [13]. This vulnerability to paraphrase-based evasion is consistent with the broader survey findings on detector fragility [3]. Another severe impact observed in academic production generated by these architectures was the occurrence of hallucinations, resulting in nonexistent bibliographic citations and entirely fabricated numerical data that nonetheless perfectly mimic the structure and plausibility of real studies [4].

C. Specific Impact on Scientific Conferences

Scientific conferences, as rituals of social validation and rapid knowledge dissemination, have felt the impact of this paradox in specific and concerning ways:

Flood of Synthetic Abstracts. Between 2021 and 2023, quantitative analyses of submissions to major international conferences and scientific journals identified a measurable increase in AI-generated content, both in oncology conference abstracts [14] and in artificial intelligence conferences such as NeurIPS and ICLR [15].

Crisis of Peer Review at Events. Reviewers began using LLMs to summarize and evaluate full conference papers, often without organizers' permission, leaking confidential data to third-party servers and compromising the confidentiality of the evaluation process.

Devaluation of the Poster Format. With the ease of generating visually flawless posters and dense text via AI, the poster format has lost part of its value as an indicator of the researcher's synthesis effort and content mastery.

Asymmetry in Oral Presentations. Researchers who used AI to prepare presentation scripts demonstrated artificially elevated fluency, while researchers native to non-hegemonic languages, even when experts in the content, continued to face linguistic barriers unresolved by simple automated translation.

D. Structural and Ethical Impacts

The findings also reveal impacts of a structural and ethical nature on the dissemination of science. It was found that unregulated adoption of automated writing carries the intrinsic risk of propagating biases consolidated in the vast training databases, threatening the diversity of thought [5]. Furthermore, it was observed that AI's predictive behavior tends to exacerbate phenomena of academic inequality, such as the Matthew Effect, by systematically favoring the recommendation and citation of already widely referenced works to the detriment of more recent output or research from peripheral researchers [5]. The convergence of this data strongly indicates that the current infrastructure for submission and verification at conferences and in journals is technologically outdated for ensuring content integrity. This fragility not only allows sophisticated fraud to circulate but also penalizes legitimate scientific labor [1].

E. The Specific Case of Biomedical Engineering and Health Publications

Biomedical Engineering has an essentially applied character, in which published knowledge frequently serves as direct input for the development of medical devices, diagnostic algorithms, and clinical protocols. For this reason, the paradox of AI-assisted writing takes on particularly high-risk contours in this field. A survey published in The Lancet, which audited approximately 2.5 million biomedical articles indexed in PubMed Central between January 2023 and February 2026, identified the proportion of articles containing at least one fabricated bibliographic reference. The data are striking: such articles grew more than tenfold during this period. While in 2023 a researcher had to scan nearly 3,000 publications to find a single paper with a fabricated reference, by early 2026 this density had surged, requiring the review of fewer than 280 papers to encounter one, with the sharpest inflection coinciding with the popularization of generative AI writing tools from 2024 onward [7]. The same study found that literature review articles, a category widely used to inform clinical guidelines and care protocols, presented a reference-fabrication rate 57% higher than that of original research articles, amplifying the risk of systemic contamination of the evidence bases used in health decision-making [7].

Vulnerability is not restricted to the writing of primary articles: the integrity of the peer-review process itself in Biomedical Engineering journals has also been directly compromised. In a letter to the editor published in the journal Annals of Biomedical Engineering, researchers reported identifying hidden textual instructions deliberately inserted into preprint manuscripts, designed to manipulate reviewers using AI assistants during review by instructing the system to issue a favorable opinion regardless of the work's scientific merit [8].

In response to analogous risks, funding agencies such as the National Institutes of Health (NIH) and the National Science Foundation (NSF) have formally prohibited the use of generative AI in the peer-review process for research proposals, recognizing the confidentiality and integrity of evaluation as non-negotiable assets [9], [10]. On the national scale, this rigor was consolidated with the publication of Ordinance No.

2,664/2026 by the Brazilian National Council for Scientific and Technological Development (CNPq) [11]. According to an interpretation raised in recent academic debate by researcher Fernanda Scussel [12], the explicit prohibition on inputting third-party projects into AI tools for drafting reviews implies, in practice, a restriction on the use of automated detectors as well. This is because detection platforms require the upload of others' manuscripts for analysis on third-party servers, violating the confidentiality guaranteed by the regulation. This legal perspective directly dialogues with the empirical evidence that such tools are statistically unstable. Given that detectors generate false positives [2], [3] and that legislation itself restricts the transit of third-party data on these platforms [11], it becomes untenable to keep scientific evaluation focused solely on textual policing. In light of this, the urgency is reinforced for a transition in how reviews are conducted, shifting them toward a context focused more on data auditing and methodological transparency.

These findings have direct implications for the theme of this conference edition, "Building Bridges Between Academia and the Market": the transfer of scientific knowledge into the development of health technologies presupposes that this knowledge base be traceable and verifiable at its source. Contamination of the literature by fabricated citations or undetected synthetic data compromises, above all, the very raw material of technological innovation in health, constituting a risk that transcends the strictly academic sphere and directly affects the reliability of the products, devices, and processes derived from it.

IV. Discussion and Predictive Implications (2026-2035)

A. The Organic versus Synthetic Paradox

The scenario outlined by the results evidences the consolidation of the "organic versus synthetic" paradox, in which the boundary between a legitimate co-creation tool and authorial fraud becomes increasingly blurred in academia [1]. The key insight for understanding this phenomenon lies in the observation that the current punitive stance of most journals and conference committees rests on a systemically fragile premise; after all, if a discovery or textual formulation possesses rigor, originality, and scientific value, the purely artificial nature of its genesis should not be the sole criterion for its invalidation [6]. The introduction of these tools does not create ethical problems but rather acts as a magnifying glass that exposes the structural fractures of an academic system historically anchored in the "publish or perish" culture [1], [5].

B. The Unsustainability of the Technological Arms Race

The predictive implications for the future of publishing indicate that the technological arms race between text-generation platforms and detection software is unsustainable and unproductive. Since detectors generate false positives and are easily fooled by paraphrasing, trust in the peer-review process will suffer severe strain if it relies exclusively on algorithmic audits [3].

C. Projections for the Academic Ecosystem (2026-2035)

The four transformations projected below were derived through analytical extrapolation by the authors, based on the structural fractures identified in the empirical evidence discussed in Section III (namely, the instability of detection systems [2], [3], [13], the fabrication crisis documented in biomedical literature [7], and the regulatory responses already adopted by funding agencies [9]–[11]). Unlike the findings reported in Section III, these projections do not derive directly from empirical studies; rather, they represent the authors' reasoned interpretation of where current trends are likely to lead. No formal forecasting method (such as a Delphi panel or structured scenario analysis) was employed, and this speculative nature is further addressed as a limitation in Section V. Given the above, it is projected that journals and scientific events will need to urgently migrate from blind-prohibition policies toward requiring radical methodological transparency. Specifically, four major transformations are anticipated through 2035:

1) Redefinition of Authorship and "Epistemic Origin Seals" [authors' projection]. The requirement for an "AI use declaration" will cease to be a footnote and become a mandatory, auditable metadata field. The creation of "Epistemic Origin Seals" is anticipated, whereby researchers must map exactly what percentage of ideation, writing, and data analysis was delegated to AI. AI "co-authorship" will continue to be denied by major journals, but conferences will create hybrid presentation categories (Human + AI Agent), with specific transparency requirements.
2) Migration from Text to Data Auditing [extrapolated from the detector-instability evidence in Section III-B [2], [3], [13]]. As written text loses its value as proof of competence (since anyone can generate a flawless article with AI), academic evaluation will shift focus to what AI cannot easily falsify: raw data, programming code, laboratory protocols, and the methodological decision tree. In the specific context of Biomedical Engineering, this auditing will translate into the traceability of biosignal databases (such as ECG and EEG), the verification of medical image repositories, and the availability of open-source code for the diagnostic algorithms employed. The article itself will come to function strictly as an "executive summary" of an underlying database, one that is traceable and subject to audit.
3) Rebirth of the Oral Defense at Conferences [authors' projection]. The paradox of writing will trigger a pendulum reaction: the revaluation of orality. Scientific conferences will cease to be showcases for read-aloud texts and will become arenas for the defense of thought. Q&A sessions will be rigorously expanded. A researcher's ability to debate, improvise, and defend their premises against challenges from peers (and auditing AIs) will become the true proof of authorship and understanding.
4) Peer Review Hybridized with AI "Devil's Advocates"

[extrapolated from the regulatory precedents in Section III-E [9]–[11]]. Peer review at conferences will become fully hybridized. AIs will not replace reviewers but will act as pre-programmed "Devil's Advocates," tasked with finding logical flaws, statistical biases, and literature gaps before the article even reaches the human reviewer. In this new arrangement, the expert is repositioned: rather than acting as the manuscript's "primary reader," the expert assumes the role of "auditor of audits." Far from constituting a demotion, this transition elevates the expert to the indispensable status of guardian of scientific ethics and common sense, focusing on questions of relevance and impact that require contextual judgment. This human role will be crucial precisely to catch possible algorithmic hallucinations or biases generated by the evaluating AI itself, since the same architecture capable of fabricating references when writing can also "hallucinate" a nonexistent methodological flaw when evaluating, making it the human reviewer's responsibility to ensure that the AI's critique actually makes sense.

D. Radical Transparency and Open Science

The evaluation ecosystem will need to focus on the open declaration of AI use, requiring that final responsibility, data curation, and the validation of critical reasoning remain inalienable duties of the human researcher, while simultaneously strengthening open science practices [4], [6]. Transparency will no longer be an ethical option but an epistemic requirement for the validity of the knowledge produced.

V. Study Limitations

This study presents certain limitations that should be weighed when interpreting its results and projections. First, from a methodological standpoint, the work is grounded in a narrative evidence synthesis rather than a systematic review protocol; database search results were not exhaustively screened with fixed inclusion and exclusion counts at each stage, and it did not conduct primary empirical tests (such as generating original synthetic texts or novel testing of detection algorithms). Second, given the exponential pace of updates to Large Language Models (LLMs) and verification tools, the empirical data on detector effectiveness represent a temporal snapshot that may become rapidly obsolete. Finally, the quantitative analysis of bibliographic contamination and fabricated references focused primarily on the Health Sciences and Biomedical Engineering databases (such as PubMed, IEEE Xplore, Scopus, and Web of Science), meaning the exact magnitude of the phenomenon may vary if extrapolated to other fields of knowledge, such as the Humanities or Exact Sciences.

Regarding the predictive implications outlined for the academic ecosystem (2026-2035), it is necessary to acknowledge their inherently speculative nature. As explicitly noted in Section IV-C, these four transformations were derived through the authors' analytical extrapolation from documented structural fractures, without recourse to a formal forecasting method such as a Delphi panel or structured scenario analysis. The projected transformations, such as the implementation of origin seals and the hybridization of peer review, constitute logical extrapolations based on current trends and structural fractures within the publishing system. The definitive realization of these scenarios is not deterministic and remains subject to unpredictable variables, such as the pace of updates to current legislation, the commercial policies of major publishing conglomerates, and the degree of cultural acceptance of these new norms within the scientific community itself.

VI. Conclusion

It is concluded that the symbiosis between researchers and generative artificial intelligence models is an irreversible and profoundly transformative technological reality. The demonstrated instability and high failure rate of synthetic content detectors show that the preservation of scientific integrity will not be achieved through banning or imprecise automated tracking. The historical approach demonstrated that academia has always absorbed and domesticated new communication technologies, but generative AI represents a qualitative rupture, as it is not merely an assistance tool but a probabilistic generation agent that challenges the very foundations of authorship. The predictive projections indicate that the future of scientific communication will depend on restructuring evaluation methods and formulating ethical guidelines that prioritize radical transparency, originality of methodological reasoning, and the non-transferable human responsibility over generated knowledge. Scientific conferences, in particular, will have the opportunity to reclaim their lost essence: live debate, the confrontation of ideas, and the oral defense of knowledge. For publications and conferences to maintain their relevance and reliability, systemic adaptation that transcends mere policing of tool use is imperative. The future of academia does not belong to those who write best with the help of machines, but to those who think most clearly to guide them. In the specific field of Biomedical Engineering, this systemic adaptation is a necessary condition for the bridge between Academia and the Market, proposed as the theme of this conference edition, to be built on a verifiable evidence base, rather than on a silent accumulation of fractures in scientific integrity.

Acknowledgments

The authors thank the Faculdade de Tecnologia of the Universidade Estadual de Campinas (UNICAMP) for institutional support and the faculty for their critical contributions that enabled the maturation of this ethical debate. Rafael Fabre Rodrigues e Souza also thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for granting an institutional scholarship quota (CAPES/DS Scholarship), fundamental for the subsistence and exclusive dedication to graduate academic research.

AI Use Declaration

The authors acknowledge the use of generative artificial intelligence tools in this work. Google's Gemini 3.1 Pro was

used as an auxiliary resource for exploratory research, initial idea structuring, spelling review, textual cohesion enhancement, and translation. Anthropic's Claude Sonnet 5 was used solely to perform a critical review of the manuscript after its finalization. All ideas, analyses, interpretations, results, and final content were reviewed, validated, and remain the sole and exclusive responsibility of the researchers.

REFERENCES

[1] A. Bozkurt, "Genai et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative ai," *Open Praxis*, vol. 16, no. 1, pp. 1–10, 2024.

[2] C. A. Gao *et al.*, "Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–7, 2023.

[3] S. Pudasaini *et al.*, "Survey on ai-generated plagiarism detection: The impact of large language models on academic integrity," *Journal of Academic Ethics*, vol. 23, pp. 1137–1170, 2025.

[4] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, pp. 1–20, 2023.

[5] B. D. Lund *et al.*, "Chatgpt and a new academic reality: Ai-written research papers and the ethics of the large language models in scholarly publishing," *Journal of the Association for Information Science and Technology*, vol. 74, no. 5, pp. 570–581, 2023.

[6] Z. Lin, "Why and how to embrace ai such as chatgpt in your academic life," *Royal Society Open Science*, vol. 10, no. 8, p. 230658, 2023.

[7] M. Topaz *et al.*, "Fabricated citations: an audit across 2.5 million biomedical papers," *The Lancet*, vol. 407, no. 10541, 2026.

[8] L. Giray, "Hidden prompts in manuscripts threaten the integrity of peer review and research: Recommendations for journals and institutions," *Annals of Biomedical Engineering*, vol. 53, no. 10, pp. 2385–2388, 2025.

[9] National Institutes of Health, "The use of generative artificial intelligence technologies is prohibited for the nih peer review process," https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html, 2023, accessed on: June 27, 2026.

[10] National Science Foundation, "Notice to research community: Use of generative artificial intelligence technology in the nsf merit review process," https://www.nsf.gov/news/notice-to-the-research-community-on-ai, 2023, accessed on: June 27, 2026.

[11] Brazil. Ministry of Science, Technology and Innovation. National Council for Scientific and Technological Development, CNPq, "Ordinance no. 2,664, of march 11, 2026," Official Gazette of the Union, Brasília, DF, Mar. 11, 2026, 2026, accessed on: July 14, 2026.

[12] F. Scussel, "Interpreting cnpq resolution 2.664/2026 in practice," https://www.instagram.com/reel/DatFZNEFLfo, 2026, accessed on: July 20, 2026.

[13] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[14] F. M. Howard, A. Li, M. F. Riffon, E. Garrett-Mayer, and A. T. Pearson, "Characterizing the increase in artificial intelligence content detection in oncology scientific abstracts from 2021 to 2023," *JCO Clinical Cancer Informatics*, vol. 8, p. e2400077, 2024.

[15] Pangram Labs, "Ai conference papers are increasingly being written by ai: up 370% since 2023," https://www.pangram.com/blog/academic-papers, 2025, accessed on: July 18, 2026.

Science Dissemination and Female Participation in Engineering Applied to Health Technologies

Kaylane de Sousa Silva
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
kaylane.silva@itec.ufpa.br

Yasmim Lisboa
Faculty of Electrical and Biomedical Engineering
Federal University of Pará
Belém, Brazil
yasmim.lisboa@itec.ufpa.br

Iris Rodrigues
Electrical Engineering Graduate Programation
Federal University of Pará
Belém, Brazil
iris.rodrigues@itec.ufpa.br

Janise Maria Monteiro Viana
Nucleus of Higher Amazonian Studies
Federal University of Pará
Belém, Brazil
janise@ufpa.br

Wellington da Silva Fonseca
Electrical Engineering Graduate Programation
Federal University of Pará
Belém, Brazil
fonseca@ufpa.br

***Abstract*— The low representation of women in Science, Technology, Engineering, and Mathematics (STEM) fields highlights the need for initiatives that increase girls' and women's contact with these fields and raise their visibility in academic and professional careers. This study analyzes three live streams organized by the Girls in Technology Project – North Connection, covering topics related to Assistive Technology, Biomedical Engineering, and Biomedicine. The evaluation took into account the reach of promotional posts on Instagram, the number of views of the YouTube streams, the age distribution of the audience, and the questions asked by viewers. The posts received a total of 10,581 views on Instagram, while the live streams reached 333 views on YouTube. The discussions covered the training of educators in the use of assistive technology, career prospects in biomedical engineering, and the challenges faced by an Indigenous student in science. The results provide a preliminary mapping of the digital reach of these initiatives and identify specific instances of audience engagement based on recorded interactions related to training, career opportunities, and women's academic and professional trajectories in technology and health.**

Keywords*— Science promotion; women in science; health technologies; STEM.

I. INTRODUCTION

Women's presence in science, technology, engineering, and mathematics (STEM) is still marked by inequalities, with barriers related to entering and remaining in higher education, as well as to entering the workforce [1]

In engineering degree programs in Brazil, for example, there is still a significant male predominance [2]. However, this situation cannot be attributed solely to female students' individual choices regarding these programs, since their paths are also influenced by gender stereotypes, prejudice, isolation, insufficient institutional support, and a lack of visible female role models [1], [2]

Biomedical Engineering is a relatively recent interdisciplinary field in Brazil, with its first undergraduate program established in 2001 by the University of Vale do Paraíba (UNIVAP) [3]. By integrating knowledge from engineering, the exact sciences, the biological sciences, and the health sciences, the field is interdisciplinary in nature and focuses on the development of techniques, equipment, and technologies for diagnosis, treatment, rehabilitation, and other health-related applications [4].

In this context, scientific outreach and public engagement initiatives can influence perceptions of STEM careers and increase the visibility of female scientists and engineers as professional role models [5]. Thus, presenting the fields of study, career opportunities, and the professional trajectories of women in these areas can also broaden female students' knowledge and bring them closer to the applications of engineering in healthcare [6], [7].

As ways to broaden the reach of science dissemination, social media and digital platforms make it possible to share scientific content with geographically dispersed audiences [8]. In this regard, Phillips et al. [9] found that an initiative carried out on Instagram brought visibility to female researchers of different racial identities and nationalities, highlighting the potential of these channels to showcase female role models to international audiences.

Habibi and Salim also observed that scientific content presented in dynamic formats garnered greater engagement than static posts [8]. In this regard, live streams constitute a format for scientific outreach capable of combining content presentation, sharing academic and professional experiences, and interaction with audiences from different institutions and locations.

In this context, the Girls in Technology Project—North, South, and Northeast Connection was launched, a collaborative initiative among universities in these three Brazilian regions, developed with the goal of promoting participation and increasing the visibility and representation of women in the fields of exact sciences, engineering, and technology.

As part of North Connection, the project carries out initiatives to disseminate scientific knowledge and engage with students, promoting the sharing of academic and professional experiences among women working in STEM fields. This inter-institutional collaboration fosters the exchange of practices and experiences focused on gender equity in science and technology, while also increasing the visibility of women and encouraging the participation of girls and women in these fields.

Among the activities carried out by the project, three live streams were selected for analysis based on their thematic relevance to the intersections of technology, engineering, and health. The study aims to analyze the reach and thematic

contributions of these activities, taking into account the number of views, the content presented, and the questions raised by the audience.

In addition, the study seeks to discuss the potential of these activities to increase female representation and bring girls and women closer to STEM fields, particularly at the intersections of engineering, technology, and health.

II. METHODOLOGY

The methodology was structured into four stages: 1) defining topics related to STEM fields; 2) selecting participants and organizing the team; 3) planning and conducting live streams; and 4) monitoring indicators and audience feedback. These stages made it possible to assess the potential of the activities and their contributions to encouraging greater female participation in engineering.

The project team, shown in Fig. 1, consists of professors and undergraduate and postgraduate students from various fields of study, including Electrical Engineering, Biomedical Engineering, and Chemical Engineering, which contributes to the multidisciplinary nature of the initiatives carried out.



Fig. 1. Girls in Technology Project Team - North Connection

A. *Themes definition*

The events were streamed live on YouTube and lasted an average of 60 minutes. Launched in May 2025, the initiative covered various themes related to inclusion, leadership, accessibility, science, and technology.

The selection of topics for the livestreams was largely guided by suggestions and comments from the participating audience, who identified topics of interest for future editions, thereby fostering a program aligned with the community's needs. In addition, the project team prioritized topics recognized for their scientific and educational relevance, especially those focused on promoting health, inclusion, and technological development.

To this point, the project has held 14 live streams, three of which are included in the scope of this study because they address topics related to biomedical engineering, titled: "Accessibility and Assistive Technology in Professional and Technological Education"; "Women in Health Innovation: Challenges and Opportunities in the Field of Biomedical Engineering"; and "Between Traditional Knowledge and Biomedicine: The Journey of an Indigenous Student in Science."

B. *Selection of Participants*

After the topics were chosen, the selection of participants prioritized professionals whose academic, scientific, and social activities were aligned with the themes addressed in each broadcast. Accordingly, preference was given to female researchers and professionals from the North Region with proven experience in teaching, research, extension, or the development of projects related to biomedical engineering, accessibility, assistive technology, innovation in health, and social inclusion.

In addition to academic background and technical-scientific output, consideration was given to involvement in initiatives with social impact, participation in research and outreach projects, and the representativeness of their professional trajectories, in order to promote discussions grounded in scientific evidence and practical experience.

C. *Live Stream Planning*

The planning and execution of the live streams followed the methodological workflow presented in Fig. 2, structured in three stages. The pre-broadcast stage included defining the theme, selecting the speakers, organizing the team, and preparing support and promotional materials.

During the transmission, the moderator led the session and the technical script was followed. The post-transmission phase involved issuing certificates and collecting feedback from participants, which informed the evaluation of the project's activities.

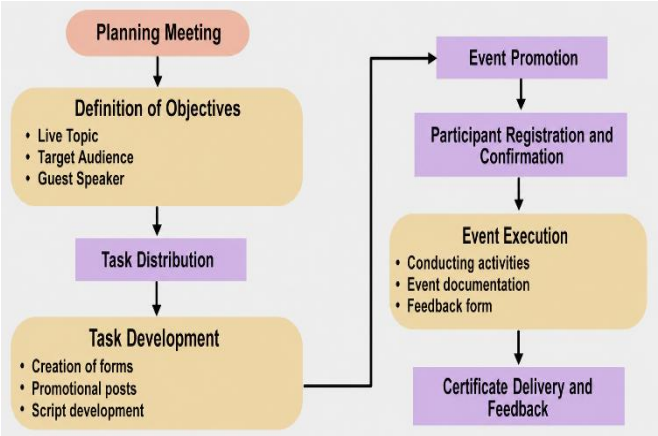


Fig. 2. Planning Flowchart

To expand the reach of the activities, the events were promoted on Instagram—the project's primary communication channel—using promotional graphics, as shown in Fig. 3. The live streams were hosted on YouTube, with technical support and management provided by StreamYard software, which facilitated interaction among participants and the simultaneous broadcast of the event.



Fig. 3. Promotional material for the live streams.

D. *Indicators and Feedback*

In this study, the analysis of the live streams was based on metrics generated by Instagram and YouTube, which allow for an assessment of the reach of the content among followers and non-followers, the demographic profile of the audience (age group and gender), the number of views, and engagement metrics such as likes, shares, comments, and other recorded interactions.

In the case of live streams, comments and interactions that occurred during the broadcast were also considered, making it possible to analyze the level of viewer participation and engagement with the topics discussed. In addition, the feedback forms completed at the end of each live stream were quantitatively analyzed.

The analysis considered the respondents' demographic and educational profiles, the communication channel through which they learned about the live streams, and their evaluations of the organization, clarity of language, fulfillment of expectations, and contribution of the activities to their interest in STEM fields. The rating-scale responses were analyzed using descriptive statistics, including frequencies and percentages, while open-ended responses were considered as complementary qualitative information.

III. RESULTS

The indicators obtained from the three live streams are presented in Table I. On Instagram, the reach of the posts used to publicize the initiatives was analyzed, while the YouTube data corresponds to the views of the live streams. Thus, Instagram metrics were considered indicators of reach and dissemination, whereas YouTube views were used to characterize audience participation in the live broadcasts.

TABLE I. PROJECT LIVESTREAMS

Theme	*Instagram Post Views*	*YouTube Live Views*
Accessibility and Assistive Technology in Vocational and Technological Education	3.888	77
Women in Health Innovation: Challenges and Opportunities in the Field of Biomedical Engineering	3.578	81
Between Traditional Knowledge and Biomedicine: The Journey of an Indigenous Student in Science.	3.115	175

The three promotional posts analyzed reached a total of 10,581 views on Instagram, with an average of 3,527 views per post. The corresponding live streams recorded 333 views on YouTube, comprising 77 views for the Accessibility and Assistive Technology initiative, 81 for Biomedical Engineering, and 175 for Biomedicine. These results show that the publications achieved broad dissemination on Instagram, while the live broadcasts reached a more specific audience on YouTube. Therefore, the two indicators provide complementary information: Instagram reflects the visibility achieved through the dissemination strategy, whereas YouTube views provide an indication of participation in live activities.

Among the Instagram publications, the initiative on Accessibility and Assistive Technology accounted for 36.7% of the views, followed by Biomedical Engineering, with 33.8%, and Biomedicine, with 29.4%. The relatively close distribution of views among the three themes indicates consistent interest in different relationships between technology and health, including accessibility, innovation, professional education, and diversity in science.

The live broadcast entitled "Accessibility and Assistive Technology in Vocational and Technological Education" addressed the relationship between technology, inclusion, and professional education. Among the questions raised by viewers during the broadcast, two stood out:

> "Many educators lack the training and knowledge required to use Assistive Technology in education, which limits the effectiveness of including students in educational practices. How can this situation be changed?"

> "Do you believe that accessibility is considered from the beginning of research projects, or does it still tend to appear as a subsequent adaptation?"

These questions directed the discussion toward two aspects highlighted in the literature: the need for teacher training for the appropriate selection and use of Assistive Technology, and the adoption of universal design approaches, in which accessibility is considered in advance during the planning of educational practices and resources, thereby reducing the need for subsequent adaptations [10], [11]. In this context, the initiative presented Assistive Technology as a field of research and professional practice committed to developing more inclusive solutions.

During the live broadcast entitled "Women in Health Innovation: Challenges and Opportunities in the Field of Biomedical Engineering," which highlighted female

participation at the intersection of engineering, innovation, and health, questions such as the following were recorded:

> "What distinguishes Biomedical Engineering from Engineering Physics?"

> "What is the job market like for biomedical engineers?"

These questions indicate an interest in understanding the identity of the field and its professional opportunities. The responses made it possible to present professional activities related to research, development, management, and healthcare technologies. These activities are consistent with the professional competencies of Biomedical Engineering established in Article 8 of CONFEA Resolution No. 1,156/2025, in conjunction with the professional activities defined in Article 5 of Resolution No. 1,073/2016 [12], [13].

By sharing the guest speaker's academic and professional journey, the live broadcast also increased the visibility of women working in engineering and research. This aspect is consistent with studies that emphasize the importance of visible female role models in encouraging girls to identify with STEM fields [5].

The live broadcast entitled "Between Traditional Knowledge and Biomedicine: The Journey of an Indigenous Student in Science" recorded the highest number of views on YouTube and Instagram. Most of the audience consisted of non-followers, accounting for 83.7% of the total.

During the broadcast, participants asked:

> "What were the greatest challenges you faced when entering and remaining in the scientific field as an Indigenous student?"

> "What were the greatest cultural or scientific shocks you experienced when entering university?"

These questions broadened the discussion to include the challenges faced by Indigenous women in accessing and remaining in higher education, as well as the relationships among academic education, cultural identity, and a sense of belonging. These aspects are discussed in the literature as important elements of the university experiences of Indigenous women [14]. By presenting the guest speaker's journey, the live broadcast also increased the visibility of women and Indigenous and Amazonian individuals within scientific and academic environments. This is consistent with studies that associate such representation with empowerment, the recognition of traditional knowledge, and the strengthening of ethnic identity [14].

Furthermore, the age distribution indicated that the Instagram content reached audiences from different age groups, as illustrated in Fig. 4.

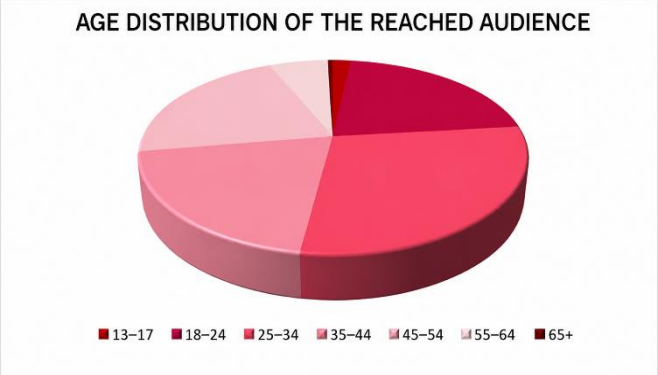


Fig. 4. Age distribution associated with the promotional posts for the three Instagram live broadcasts.

In Assistive Technology, the 45–54 age group stood out, accounting for 46.7% of the audience. In Biomedical Engineering, 67.0% of the audience was between 25 and 44 years old, whereas in Biomedicine, 73.0% was between 18 and 34 years old. This variation suggests that the topics attracted individuals at different stages of their academic education and professional careers.

The combined analysis allows the identification of three main contributions: the discussion on the inclusive application of technologies, the dissemination of professional opportunities at the intersection of engineering and health, and the recognition of diverse female trajectories in science.

The quantitative analysis of the feedback forms, completed by a total of 28 participants, complemented the indicators obtained from the digital platforms, allowing the characterization of the participants' profiles and their evaluation of the activities. Among the respondents, 57.1% were women and 42.9% were men. Regarding educational level, undergraduate students predominated (82.1%), followed by secondary school students (14.3%), while postgraduate participation was lower (3.6%). Most respondents reported that they had learned about the live streams through Instagram, reinforcing the relevance of this platform as a dissemination channel for the activities.

Regarding the evaluation of the live streams, 92% of respondents positively assessed the organization of the activities. In the overall evaluation, 95% assigned the highest score, while the remaining 5% assigned a score of 4, corresponding to a "good" rating. The feedback forms also recorded predominantly positive evaluations regarding the clarity of the language, fulfillment of expectations, and the contribution of the activities to interest in STEM fields. These quantitative results indicate a high level of acceptance among the participants who completed the feedback forms and complement the reach and participation metrics obtained from Instagram and YouTube.

In addition, records of audience participation during the live broadcasts, particularly through comments and questions, complement the digital metrics by identifying specific topics that generated audience interaction. Within the scope of the data collected, these results indicate point

instances of engagement with topics related to accessibility, professional opportunities, and women's trajectories in science, engineering, and health.

IV. CONCLUSION

The study provided a preliminary mapping of the digital reach of scientific outreach initiatives related to female participation in technology and health. The analysis of Instagram and YouTube metrics, feedback forms, and audience interactions allowed the identification of specific instances of engagement with topics related to accessibility, professional opportunities, and diverse female trajectories in science. It is important to highlight that this information is limited to the interactions recorded during the analyzed activities, which restricts broader interpretations regarding female participation in STEM fields.

V. FUTURE WORKS

As a next step toward improving the project, it is proposed to expand the analysis by including a larger number of participants and live streams, as well as longitudinally monitoring their academic and professional trajectories. It is also intended to investigate whether exposure to the live streams and female role models in STEM fields is associated with students' interest in and persistence in Engineering programs over the years. For future evaluations, quantitative metrics and statistical validation methods are recommended to enable a more robust assessment of the impacts of these initiatives. These improvements may contribute to a broader understanding of their effectiveness in promoting women's participation and supporting academic and professional pathways in STEM fields.

ACKNOWLEDGMENTS

The authors would like to thank the Federal University of Pará (UFPA), the National Council for Scientific and Technological Development (CNPq), and the Girls in Technology Project – North, South, and Northeast Connection, the Department of Undergraduate Education (PROEG), and the Center of Excellence in Energy Efficiency in the Amazon (CEAMAZON) for their support and encouragement in carrying out this work.

REFERENCES

[1] L. F. C. Costa et al., "Women's Journey in STEM Education in Brazil: A Rapid Review on Engineering and Computer Science," IEEE Access, vol. 12, pp. 112576–112593, 2024, doi: 10.1109/ACCESS.2024.3442879

[2] B. Camila Francisco de Castro, T. M. David, A. Maria Tonini and P. Maria Silva Rocha Rizol, "Gender Disparity in Engineering Courses in Brazil," in IEEE Access, vol. 12, pp. 159784-159793, 2024, doi: 10.1109/ACCESS.2024.3483660.

[3] Sociedade Brasileira de Engenharia Biomédica, "Graduação em Engenharia Biomédica no Brasil.". Disponível: https://sbeb.org.br/graduacao/. Acesso em: 30 jul. 2026.

[4] M. A. B. de Paula, A. de C. Grilo e M. A. Ávila, "Engenharia biomédica: escolhas profissionais nas trajetórias docentes," Revista Internacional de Educação Superior, vol. 9, Art. nº e023013, 2023, doi: 10.20396/riesup.v9i00.8660655.

[5] L. Fogg-Rogers and L. Hobbs, "Catch 22—Improving visibility of women in science and engineering for both recruitment and retention," Journal of Science Communication, vol. 18, no. 4, Art. no. C05, 2019, doi: 10.22323/2.18040305.

[6] F. M. Martinez-Licona and J. R. Jimenez-Alanis, "Gender Parity in Biomedical Engineering: Analysis of Admission to the Bachelor's Degree at UAM," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 6036-6039, doi: 10.1109/EMBC44109.2020.9175308.

[7] L. A. Thompson and N. Zhang, "Young Women Exposed Actively to the Value of Biomedical Engineering," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 6013-6017, doi: 10.1109/EMBC44109.2020.9176290.

[8] J. Steinke, K. Baumel, and M. M. Turner, "Cultivating authenticity as role models: Women in STEM influencers on Instagram," Science Communication, vol. 46, no. 5, pp. 543–583, 2024, doi: 10.1177/10755470241239942.

[9] S. A. Habibi and L. Salim, "Static vs. dynamic methods of delivery for science communication: A critical analysis of user engagement with science on social media," PLoS ONE, vol. 16, no. 3, Art. no. e0248507, 2021, doi: 10.1371/journal.pone.0248507.

[10] J. M. Fernández-Batanero, M. Montenegro-Rueda, J. Fernández-Cerero, and I. García-Martínez, "Assistive technology for the inclusion of students with disabilities: A systematic review," Educational Technology Research and Development, vol. 70, pp. 1911–1930, 2022, doi: 10.1007/s11423-022-10127-7.

[11] A. Sewell, A. Kennett, and V. Pugh, "Universal Design for Learning as a theory of inclusive practice for use by educational psychologists," Educational Psychology in Practice, vol. 38, no. 4, pp. 364–378, 2022, doi: 10.1080/02667363.2022.2111677.

[12] Conselho Federal de Engenharia e Agronomia, "Resolução nº 1.156, de 24 de outubro de 2025: Consolida normativos acerca das atividades e competências profissionais dos engenheiros da modalidade eletricista," Brasília, DF, Brasil, 2025. Disponível: https://normativos.confea.org.br/Ementas/Visualizar?id=82360&utm. Acesso em: 30 jul. 2026.

[13] Conselho Federal de Engenharia e Agronomia, "Resolução nº 1.073, de 19 de abril de 2016: Regulamenta a atribuição de títulos, atividades, competências e campos de atuação profissionais," Brasília, DF, Brasil, 2016. Disponível: https://normativos.confea.org.br/Ementas/Visualizar?id=59111&utm. Acesso em: 30 jul. 2026.

[14] T. J. da S. Gomes and S. de F. M. Nina, "A presença de mulheres indígenas nas universidades brasileiras: uma revisão de escopo," Revista de Psicología, vol. 33, no. 2, pp. 1–12, 2024, doi: 10.5354/0719-0581.2024.71620.

Automatic Classification of Short Videos on Anorexia and Bulimia with Validation Based on Expert Agreement

1st Maria Fernanda R. B. P. Costa
Computing Institute
Universidade Federal de Alagoas
Maceió, Brazil
mfrbpc@ic.ufal.br

2nd Gabriel L. B. Germano
Computing Institute
Universidade Federal de Alagoas
Maceió, Brazil
glbg@ic.ufal.br

3rd Thiago D. Cordeiro
Computing Institute
Universidade Federal de Alagoas
Maceió, Brazil
thiago@ic.ufal.br

***Abstract*—Mental health content about disorders like anorexia and bulimia has grown rapidly on platforms such as TikTok and Instagram, especially among adolescents, but the sheer volume makes it hard to distinguish reliable information from misleading content. This study presents a pipeline to automatically classify short videos into three categories (useful, misleading, and personal experience), combining mobile-based collection, transcription, expert annotation, and supervised classification. Inter-annotator agreement (Fleiss' Kappa) was incorporated into model training via sample weighting. Since the annotated corpus was small and imbalanced (33 videos, no misleading examples), a three-stage synthetic augmentation pipeline was built using LLM-generated seed texts, BERTimbau embeddings, class-conditional GANs, and LLM paraphrasing. Four classifiers (Logistic Regression, Random Forest, SVM, Naive Bayes) were trained on TF-IDF features with Kappa-oriented Grid Search, achieving 0.885–0.893 accuracy and 0.817–0.840 Kappa, surpassing per-disorder baselines and comparable literature. Logistic Regression was selected for deployment for its low inference cost and calibrated probabilities, and integrated into a real-time web interface that classifies videos by URL, reporting a confidence score adjusted by inter-annotator reliability and a semantic-affinity indicator per disorder.**

***Index Terms*—Mental Health; Natural Language Processing; Machine Learning; Eating Disorders; Interrater Agreement; Data Augmentation; Social Media.**

I. Introduction

Mental disorders are recognized as a major public health concern due to their substantial impact on quality of life, functioning, and mortality [1]. Among the conditions that demand particular attention are anorexia and bulimia, eating disorders associated with severe physical, emotional, and behavioral harm that affect millions of people worldwide [2]. Early identification of symptoms, access to reliable information, and adequate follow-up are essential to prevent worsening and to promote mental health.

In parallel, social media platforms have become central to the production, sharing, and consumption of health-related information [3]. Platforms such as TikTok and Instagram have consolidated themselves as major digital communication channels, allowing users to publish personal experiences, opinions, guidance, and educational content in accessible, rapidly disseminated formats. Short-form videos have become one of the primary means of information consumption, reaching millions of views daily and significantly influencing how people seek knowledge on a wide range of topics [4].

Although these platforms contribute to democratizing access to information and foster discussion of previously underaddressed topics, they also raise challenges related to the quality and reliability of published content. The absence of effective validation mechanisms allows inaccurate, misleading, or decontextualized information to spread at scale, often reaching visibility comparable to that of content produced by qualified professionals [5], [6]. For sensitive topics such as eating disorders, the circulation of inadequate information can reinforce stigma, normalize harmful behaviors, and contribute to misconceptions about diagnosis, treatment, and prevention; recent experimental evidence shows that even brief exposure to pro-anorexia short-form video content produces measurable declines in body image satisfaction among young women [7].

The sheer volume of published content makes manual analysis infeasible, motivating automated approaches to identify, organize, and classify this information. Natural Language Processing (NLP) and Machine Learning (ML) techniques have shown potential to support the analysis of large volumes of data in a fast, consistent, and scalable manner [8].

This work aims to develop machine-learning-based computational tools to collect, transcribe, and classify short videos published on Instagram and TikTok related to anorexia and bulimia, and to assess the reliability of the information shared on these platforms by measuring agreement among human evaluators. Specific objectives include: collecting short videos related to the disorders under study on both platforms; building a dataset through human classification of the collected videos; designing a text preprocessing pipeline for the collected content; training machine learning models for automatic validation and classification informed by inter-annotator agreement; analyzing model performance and the reliability of the generated classifications; and developing a web interface to make the classification system available.

II. Methodology

A. Data Collection

Short videos were collected from TikTok and Instagram Reels using Appium-based mobile automation [9]. The collection targeted hashtags related to anorexia and bulimia in both female- and male-oriented contexts, evenly distributed across the two platforms. To ensure sufficient public reach, only videos meeting a minimum engagement threshold of 10k views on Instagram or 10k likes on TikTok were considered.

Audio was extracted using the Python `youtube-dl` library. Videos containing excessive background noise, dominant music, or other audio characteristics unsuitable for textual analysis were filtered using TensorFlow's YAMNet model [10] in combination with Fourier-domain signal analysis, retaining only those in which the audio was predominantly classified as speech. The remaining audio was converted to WAV format and transcribed using the Python `SpeechRecognition` library. The resulting transcripts were stored together with each video's original metadata.

Initially, 400 videos were collected for each disorder, comprising 200 videos associated with female-oriented hashtags and 200 with male-oriented hashtags. After automated audio filtering and manual verification of topical relevance and language (Portuguese only), only 19 anorexia-related videos and 14 bulimia-related videos satisfied the final inclusion criteria. This substantial reduction highlights the challenge of identifying directly relevant content on social media, as many videos employed the target hashtags only incidentally or in contexts unrelated to eating disorders.

Videos were labeled by expert judges from the Alagoas Psychometrics and Psychological Assessment Laboratory (LAPAP) into three categories: useful, misleading, and personal experience (adapted from [11]). Two independent panels of judges (n = 10 for anorexia, n = 11 for bulimia) rated all videos, and the inter-annotator agreement was quantified with Fleiss' Kappa, both overall and per category. The study was approved by the UFAL Research Ethics Committee (protocol No. 7.587.554).

B. Class imbalance and synthetic data augmentation

As shown in Table I, no videos were labeled as misleading for either disorder, resulting in a severe class imbalance that precluded reliable supervised learning. Expanding the dataset through additional rounds of manual annotation was deemed impractical due to the time required and the limited availability of domain experts. To address this limitation, a three-stage synthetic data augmentation pipeline was developed. It must be stated explicitly that, because not a single real, human-annotated example of the *misleading* category was available, the classifier's performance on this class cannot be validated against any held-out real-world sample. Consequently, the reported metrics for this category should be interpreted strictly as a proof-of-concept (PoC) demonstration of the proposed pipeline's feasibility, rather than as evidence of generalization to real.

TABLE I
DISTRIBUTION OF CLASSIFICATION CATEGORIES

Category	**Disorder**	
	Anorexia	***Bulimia***
Useful	5	6
Misleading	0	0
Personal Experience	14	8

The first stage consisted of seed generation. Ten synthetic misleading texts were created for each disorder using the Gemini large language model (LLM). A structured prompt instructed the model to assume the role of a digital content creator and produce pseudoscientific narratives that promoted or normalized behaviors associated with anorexia or bulimia. Students with relevant domain knowledge manually reviewed all generated texts before being incorporated into the dataset.

In the second stage, all available texts were encoded into 768-dimensional embeddings using BERTimbau (`neuralmind/bert-base-portuguese-cased`), with the representation extracted from the final hidden layer. These embeddings were used to train class-conditional Generative Adversarial Networks (GANs), which generated synthetic embeddings for the underrepresented class.

In the final stage, each synthetic embedding was mapped back to text. When the cosine similarity between a synthetic embedding and its nearest real embedding was greater than or equal to 0.70, the corresponding original text was used directly. Otherwise, the closest synthetic seed text was rewritten by the Gemini LLM using the prompt: "Rewrite the following sentence using different words while preserving its original meaning, without translating it into English." This procedure increased lexical diversity while maintaining semantic consistency.

In this initial study, the data augmentation pipeline expanded the anorexia dataset from 29 to 329 samples (+1034.5%) and the bulimia dataset from 24 to 324 samples (+1250%), resulting in datasets that were approximately balanced across the three classification categories. This amount was empirically determined and used to increase other mental disease databases used by our research group.

Text was tokenized and stop words were removed before feature extraction. Data were split 80/20 into development and held-out test sets before synthetic augmentation, ensuring the test set reflects real-world, independent data (though it does include the manually added misleading seed texts). Each classifier (Logistic Regression, Random Forest, SVM, Naive Bayes) was wrapped in a scikit-learn Pipeline that combined TF-IDF vectorization with the classifier, so the vectorizer was fit only on training folds, preventing data leakage.

Each training sample was weighted according to

$$w_i = 0,8 + \kappa_c \tag{1}$$

where κ_c is Fleiss' Kappa of the category to which the sample belongs. The 0.8 floor was implemented to ensure that categories with near-zero agreement still contribute to training rather than being effectively excluded.

Hyperparameters were tuned via Grid Search with 3-fold stratified cross-validation (chosen given the small dataset size to avoid folds with fewer than four examples per class), using Cohen's Kappa as the optimization metric—more appropriate than raw accuracy under class imbalance. Table II lists the hyperparameter grids evaluated.

TABLE II
HYPERPARAMETER GRIDS USED IN GRID SEARCH

Model	Evaluated Hyperparameters
Logistic Regression	**C**: 0.01, 0.1, 0.5, 1.0, 10.0 **penalty**: 11,12 **solver**: liblinear
Random Forest	**n_estimators**: 10, 50, 100, 200 **max_depth**: 1, 10, 20, 50 **min_samples_split**: 2, 5, 10, 20 **min_samples_leaf**: 1, 5, 10, 20 **max_features**: sqrt, log2
SVM	**C**: 0.01, 0.1, 1.0, 10.0 **gamma**: scale, auto **kernel**: linear, rbf
Naive Bayes	**alpha**: 0.01, 0.1, 0.5, 1.0, 10.0 **fit_prior**: True, False

We performed a complementary analysis on the unified dataset using Rocchio-style TF-IDF centroids per disorder [12], computed exclusively on the training subset. Cosine similarity between each text and both centroids was normalized to a percentage semantic-affinity score, which was later exposed to end users.

C. Deployment interface

The selected model (Logistic Regression, trained on the unified augmented dataset with isotonic probability calibration) was integrated into a Python/Gradio web application. Given a video URL, the pipeline downloads audio, transcribes it with OpenAI Whisper (chosen over `SpeechRecognition` for its greater robustness to variable real-world audio quality and noise), applies the same TF-IDF/preprocessing steps, classifies the content, and computes an adjusted confidence score:

$$C_{aj} = \alpha \cdot p_m + \beta \cdot \hat{\kappa}_c \tag{2}$$

where p_m is the calibrated model probability for the predicted class, κ_c is Fleiss' Kappa of that class normalized by the maximum observed category Kappa, and α = 0.7, β = 0.3 were set empirically to prioritize model sensitivity while still incorporating annotation reliability. Scores at or below 0.40 (the "fair" threshold on Altman's scale) trigger an automatic low-confidence visual warning recommending expert review. The interface also reports the semantic-affinity indicators using a separate TF-IDF vectorizer fitted to the combined vocabulary of both disorders.

It should be acknowledged that the weighting coefficients were set empirically rather than derived from a formal statistical calibration procedure. As a direct consequence, given that the *misleading* category consistently exhibits the lowest Fleiss' Kappa across both disorders ($\kappa \approx 0.09$–0.15), the resulting adjusted confidence score for this class is systematically pulled below the 0.40 alert threshold by construction, regardless of the model's own predicted probability. This behavior should therefore be understood as a hardcoded business rule designed to flag inherently low-reliability categories for human review, rather than as a probabilistically calibrated estimate of the true likelihood that a given prediction is correct.

III. RESULTS AND DISCUSSIONS

A. Anorexia dataset

Overall Fleiss' Kappa reached 0.659 (SE = 0.0033, 95% CI [0.595, 0.724]), indicating *good* inter-rater agreement according to Altman's scale. Category-level agreement was also classified as *good* for the personal experience ($\kappa = 0.719$) and useful ($\kappa = 0.717$) classes. In contrast, the misleading category achieved only *very weak* agreement ($\kappa = 0.150$), reflecting both its limited representation in the original dataset and the greater ambiguity associated with identifying misleading content.

Following the three-stage LLM-based synthetic data augmentation process described in Section II, the dataset increased from 29 to 329 samples. This increase in both class balance and textual diversity created more favorable conditions for learning discriminative patterns, particularly for the originally underrepresented classes.

The expanded dataset substantially improved the performance of most classifiers. As shown in Table III, Random Forest exhibited the largest relative gain, with Kappa increasing from 0.368 to 0.795 and accuracy improving from 0.667 to 0.864. Logistic Regression and Naive Bayes also benefited from the augmented dataset, both achieving Kappa values above 0.79 and accuracies exceeding 86%, demonstrating that the synthetic samples effectively enriched the training set without compromising overall predictive consistency.

TABLE III
CLASSIFICATION PERFORMANCE BEFORE AND AFTER LLM-BASED SYNTHETIC DATA AUGMENTATION FOR THE ANOREXIA DATASET.

Model	Metric	Original Data	Augmented Data
Logistic Regression	Accuracy	0.8333	0.8788
	F1-score	0.6190	0.8808
	Precision	0.5833	0.8899
	Recall	0.6667	0.8800
	Kappa	0.7000	0.8181
Random Forest	Accuracy	0.6667	0.8636
	F1-score	0.4722	0.8588
	Precision	0.5333	0.8668
	Recall	0.5000	0.8613
	Kappa	0.3684	0.7952
SVM	Accuracy	1.0000	0.8939
	F1-score	1.0000	0.8958
	Precision	1.0000	0.9056
	Recall	1.0000	0.8945
	Kappa	1.0000	0.8407
Naive Bayes	Accuracy	0.8333	0.8636
	F1-score	0.8222	0.8647
	Precision	0.8333	0.8833
	Recall	0.8889	0.8649
	Kappa	0.7500	0.7958

Although SVM presented a reduction in Kappa from 1.000 to 0.841 after augmentation, this decrease should not be interpreted as a deterioration in predictive performance. Instead,

it reflects the elimination of the optimistic bias introduced by overfitting on the original test set, which contained only four samples. Consequently, the post-augmentation results provide a more realistic estimate of the classifier's performance on unseen data.

Overall, the results indicate that LLM-based synthetic data augmentation improved not only predictive performance but also the stability of the learning process. The most expressive gains were observed for Random Forest, suggesting that this algorithm was particularly sensitive to the increased volume and diversity of training examples. In contrast, the more modest improvements achieved by Logistic Regression and Naive Bayes indicate that these models were already capable of capturing part of the underlying textual patterns in the original dataset. However, they still benefited from the improved class balance.

B. Bulimia dataset

Inter-rater reliability was assessed using Fleiss' Kappa, yielding an overall coefficient of 0.655 (SE = 0.0033, 95% CI [0.590, 0.719]), which corresponds to *good* agreement according to Altman's scale. Agreement was strongest for the personal experience ($\kappa = 0.708$) and useful ($\kappa = 0.673$) categories, whereas the misleading category achieved only *very weak* agreement ($\kappa = 0.090$). As observed in the anorexia dataset, the lower agreement on misleading content is likely due to its limited representation in the original corpus and the greater subjectivity involved in identifying misleading health-related information.

Following the three-stage LLM-based synthetic data augmentation process described in Section II, the dataset expanded from 24 to 324 samples, providing a substantially larger and more balanced corpus for supervised learning.

Table IV summarizes the classification results before and after augmentation. In contrast to the anorexia dataset, all four classifiers produced the same performance on the original dataset, achieving an accuracy of 0.800 and a Kappa coefficient of 0.667. This uniform behavior is unlikely to reflect equivalent learning capabilities; instead, it is a consequence of the extremely small test set, where the limited number of evaluation samples constrained the classifiers to identical prediction outcomes.

The expanded dataset produced consistent improvements across all evaluated algorithms. Logistic Regression and Random Forest achieved the best overall performance, with accuracies of 0.985 and Kappa coefficients of 0.977. At the same time, SVM and Naive Bayes obtained accuracies of 0.954 with Kappa values close to 0.931.

Taken together, these findings indicate that the proposed LLM-based augmentation strategy effectively mitigated the limitations imposed by the small and initially imbalanced dataset. Besides improving predictive performance across all classifiers, the expanded corpus enabled a more reliable assessment of model generalization.

TABLE IV
CLASSIFICATION PERFORMANCE BEFORE AND AFTER LLM-BASED SYNTHETIC DATA AUGMENTATION FOR THE BULIMIA DATASET.

Model	Metric	Original Data	Augmented Data
Logistic Regression	Accuracy	0,8000	0,9846
	F1-Score	0,6000	0,9848
	Precision	0,5556	0,9855
	Recall	0,6667	0,9848
	Kappa	0,6667	0,9769
Random Forest	Accuracy	0,8000	0,9846
	F1-Score	0,6000	0,9845
	Precision	0,5556	0,9855
	Recall	0,6667	0,9841
	Kappa	0,6667	0,9769
SVM	Accuracy	0,8000	0,9538
	F1-Score	0,6000	0,9545
	Precision	0,5556	0,9559
	Recall	0,6667	0,9538
	Kappa	0,6667	0,9307
Naive Bayes	Accuracy	0,8000	0,9538
	F1-Score	0,6000	0,9542
	Precision	0,5556	0,9551
	Recall	0,6667	0,9538
	Kappa	0,6667	0,9308

C. Unified dataset

The anorexia and bulimia datasets were merged to evaluate cross-disorder generalization, using averaged per-category Fleiss' Kappa (Useful: 0.695; Misleading: 0.120; Personal experience: 0.713). The unified corpus initially comprised 53 annotated samples and, after the three-stage LLM-based synthetic augmentation process, expanded to 653 samples with an almost perfectly balanced class distribution.

Unlike the individual datasets, the original merged corpus already produced competitive performance for three of the four evaluated classifiers, with accuracies ranging from 0.818 to 0.909, as shown in Table V. Only Random Forest exhibited noticeably lower performance (0.636), suggesting that the linguistic characteristics shared among disorder-related content already yield relatively consistent discriminative patterns, even before synthetic augmentation.

After augmentation, all classifiers converged to a narrow performance range, with accuracies between 0.885 and 0.893. Compared with the original dataset, the variation in accuracy across models decreased from 0.273 to only 0.008, indicating that synthetic augmentation not only improved predictive performance but also reduced variability among different learning algorithms. Logistic Regression, Random Forest, and SVM all benefited from the expanded dataset, whereas Naive Bayes showed a slight decrease in some evaluation metrics. One possible explanation is that the richer semantic relationships introduced by the synthetic texts partially violate the conditional independence assumption on which Naive Bayes relies.

To place these findings into context, they were compared with the results reported by Benítez-Andrades *et al.* cite benitezandrades. In their study, Random Forest achieved an accuracy of 73.7% with an F1-score of 49.2%. At the same time, a RoBERTa-based model achieved 84.4% accuracy in classifying English-language eating-disorder posts as informative or non-informative. The proposed approach achieved

TABLE V
CLASSIFICATION PERFORMANCE BEFORE AND AFTER LLM-BASED SYNTHETIC DATA AUGMENTATION FOR THE UNIFIED DATASET.

Model	Metric	Original Data	Augmented Data
Logistic Regression	Accuracy	0,8182	0,8855
	F1-Score	0,7852	0,8861
	Precision	0,8667	0,8872
	Recall	0,7667	0,8861
	Kappa	0,7027	0,8282
Random Forest	Accuracy	0,6364	0,8779
	F1-Score	0,4667	0,8780
	Precision	0,4222	0,8852
	Recall	0,5333	0,8790
	Kappa	0,3889	0,8168
SVM	Accuracy	0,8182	0,8931
	F1-Score	0,7852	0,8934
	Precision	0,8667	0,8954
	Recall	0,7667	0,8931
	Kappa	0,7027	0,8396
Naive Bayes	Accuracy	0,9091	0,8931
	F1-Score	0,8586	0,8931
	Precision	0,9444	0,8943
	Recall	0,8333	0,8946
	Kappa	0,8493	0,8397

competitive performance despite relying on conventional machine learning classifiers and a substantially smaller annotated corpus, with classification results exceeding those reported for both the Random Forest and RoBERTa-based models in terms of accuracy. Nevertheless, these comparisons should be interpreted with caution because of differences in language, social media platforms, annotation protocols, and classification objectives.

The learning curve obtained for Logistic Regression (Fig. 1) on the augmented unified dataset provides additional evidence of the benefits of synthetic augmentation. A persistent gap between the training and validation curves was observed until approximately 400 training samples, with the validation F1-macro remaining close to 0.59 while the training score exceeded 0.80. Beyond approximately 420 samples, however, both curves began to converge while continuing to improve, suggesting that the availability of additional synthetic samples progressively mitigated intermediate-stage overfitting. This trend also indicates that incorporating more real annotated data could further improve the model's generalization capability.

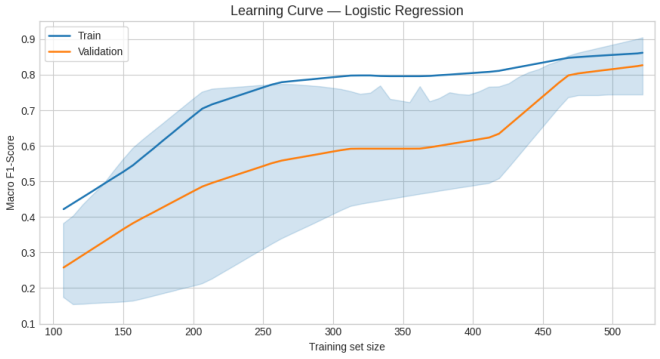


Fig. 1. Learning curve for the Logistic Regression classifier.

Prediction confidence exhibited a behavior consistent with the annotation reliability analysis. After applying the confidence adjustment defined in (2), predictions belonging to the useful and personal experience classes remained strongly concentrated near the maximum confidence level, both before and after augmentation. In contrast, confidence scores for the misleading class were considerably more dispersed, with most predictions concentrated between 0.70 and 0.75 after augmentation. This behavior is consistent with the low inter-rater agreement observed for the misleading category (Fleiss' Kappa = 0.120), suggesting that annotation uncertainty propagates into the classifier's confidence estimates.

Finally, semantic-affinity analysis based on Rocchio centroids revealed high within-disorder relevance, reaching 80.5% for anorexia and 76.6% for bulimia, whereas cross-disorder relevance remained substantially lower (19.2–23.4%). These findings indicate that, although both disorders share eating disorder-related vocabulary, each exhibits a distinct semantic profile. This semantic separation helps explain the strong classification performance achieved on the merged dataset and supports the feasibility of both disorder-specific and cross-disorder text classification.

D. Deployed Interface

Initially developed to analyze Portuguese-language videos, the application was designed with a focus on validating the user experience. The workflow begins with the initial input screen (Figure 2), where users can submit the URL of an Instagram or TikTok video for classification. Upon submission, the system processes the video and presents the comprehensive analysis on the output screen (Figure 2).



Fig. 2. Results screen displaying the predicted class.

This interface details the outcome of the Natural Language Processing (NLP) pipeline, reporting the predicted class, the base model probability, the Fleiss' Kappa of the predicted class, and an adjusted confidence score with its corresponding Altman interpretation. Additionally, it displays the full video transcript alongside the semantic-affinity percentages toward specific disorders (Figures 3 and 4).

To demonstrate this functionality, the figures include an illustrative example using an Instagram Reels URL. In this case, the system predicted the "personal experience" class with a 71.2% model probability and a 79.9% adjusted confidence score (Fleiss' Kappa = 0.713). Furthermore, the semantic



Fig. 3. Results screen displaying the model probability, adjusted confidence, and the video transcript.

analysis revealed an affinity of 55% toward bulimia and 45% toward anorexia.



Fig. 4. Detailed metrics screen illustrating the semantic-affinity percentages for the analyzed video.

IV. Conclusion

This work presented a complete pipeline for the automatic classification of eating disorder-related short-form video content, encompassing automated data collection, expert annotation, LLM-based synthetic data augmentation, supervised learning, and a deployed reliability-aware classification interface. Unlike conventional approaches that provide only a predicted label, the proposed system explicitly incorporates inter-rater agreement through Fleiss' Kappa.

Experimental results demonstrated that synthetic data augmentation substantially improved classifier performance and produced more stable estimates of model generalization. Across the anorexia, bulimia, and unified datasets, the augmented corpora consistently outperformed their original counterparts, with the unified dataset achieving up to 0.893 accuracy and a Kappa coefficient of 0.840 using classical machine learning classifiers.

Despite these promising results, several limitations should be acknowledged. The manually annotated corpus remained relatively small after language and relevance filtering, making the reported performance dependent on synthetic data whose ability to generalize to broader real-world scenarios still requires further validation. In addition, the consistently low inter-rater agreement observed for the misleading category suggests that this class requires a more precise operational definition to improve annotation consistency. Finally, the study was restricted to Brazilian Portuguese, and the automatic transcription process remains sensitive to factors such as audio quality, regional accents, and code-switching.

A further methodological concern is a mismatch between the transcription method used to build the training corpus and the one used at deployment. The training dataset was transcribed with the `SpeechRecognition` library, whereas the deployed interface uses OpenAI Whisper; the latter was used only for testing and will be replaced by an API in future work, with `SpeechRecognition` as the main library.

Another weak point of this study is that none of the 33 collected videos was classified as "Misleading". Future work will remake the video collection process looking for this class, thereby increasing both the volume and diversity of the corpus. Additional annotation rounds will refine the definition of misleading content and investigate alternative annotation protocols. From an application perspective, future versions of the proposed system will include mechanisms for collecting feedback on model predictions and support for direct text and audio submissions without requiring URLs from external social media platforms. Overall, the proposed framework provides a practical foundation for developing reliable decision-support tools to help identify and analyze eating disorder-related content on social media while maintaining transparency about prediction reliability.

References

[1] GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," The Lancet Psychiatry, vol. 9, no. 2, pp. 137–150, 2022.

[2] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR, 5th ed. Washington, DC: American Psychiatric Association Publishing, 2022.

[3] A. M. Khalaf, A. A. Alubied, A. M. Khalaf, A. A. Rifaey, "The impact of social media on the mental health of adolescents and young adults: a systematic review," Cureus, vol. 15, p. e42990, 2023.

[4] A. Z. Priyambodo, N. A. Zhafirah, R. Shabur, P. W. Handayani, H. Fitriani, "Health information-seeking using short video platforms," DIGITAL HEALTH, vol. 11, 2025.

[5] A. Hudon et al., "Navigating the maze of social media disinformation on psychiatric illness and charting paths to reliable information for mental health professionals: Observational study of TikTok videos," J. Med. Internet Res., vol. 27, p. e64225, 2025.

[6] C. E. Kirkpatrick, L. L. Lawrie, "TikTok as a source of health information and misinformation for young women in the United States: survey study," JMIR Infodemiology, vol. 4, p. e54663, 2024.

[7] M. R. Blackburn, R. C. Hogg, "#ForYou? The impact of pro-ana TikTok content on body image dissatisfaction and internalisation of societal beauty standards," PLOS ONE, vol. 19, no. 8, p. e0307597, 2024.

[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, "Deep learning–based text classification: A comprehensive review," ACM Comput. Surv. , vol. 54, no. 3, pp. 1–40, 2021.

[9] N. Verma, Mobile Test Automation with Appium . Packt Publishing, 2017.

[10] S. A. Kumar, S. Sasikala et al., "Parkinson's speech detection using YAMNet," in 2023 2nd Int. Conf. Adv. Electr. Electron. Commun. Comput. Autom. (ICAECA) , IEEE, 2023, pp. 1–5.

[11] SAMUEL, L.; KUIJPERS, K.; BLEAKLEY, A. Therapytok for depression and anxiety: A quantitative content analysis of high engagement TikTok videos. Journal of Adolescent Health, v. 74, n. 6, p. 1184–1190, 2024.

[12] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval . Cambridge University Press, 2008.

[13] BENíTEZ-ANDRADES, J. A. et al. Traditional machine learning models and bidirectional encoder representations from transformer (BERT)–based automatic classification of tweets about eating disorders: Algorithm development and validation study. JMIR Medical Informatics, v. 10, n. 2, p. e34492, Feb 2022.

University hospital as drivers of healthcare innovation: analyzing HCPATec's experience in fostering startups and intrapreneurship

Paulo Roberto Stefani Sanches
Dept of Biomedical Engineering
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
psanches@hcpa.edu.br

Bruno Rodriguez Tondin
Dept of Biomedical Engineering
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
btondin@hcpa.edu.br

Hugo Goulart de Oliveira
Innovation and Technology Transfer Office
Hospital de Clínicas de Porto Alegre / UFRGS
Porto Alegre, Brazil
holiveira@hcpa.edu.br

Alessandro Nakoneczny Schildt
Innovation and Technology Transfer Office
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
aschildt@hcpa.edu.br

Valter Ferreira Da Silva
Innovation and Technology Transfer Office
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
vfsilva@hcpa.edu.br

Rafael Leal Zimmer
Innovation and Technology Transfer Office
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
rzimmer@hcpa.edu.br

Larissa Pierozan
Innovation and Technology Transfer Office
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
lpierosan@hcpa.edu.br

Fernanda dos Santos De Oliveira
Innovation and Technology Transfer Office
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
fesoliveira@hcpa.edu.br

Wendy Beatriz Haddad Carraro
PPG em Controladoria e Contabilidade
UFRGS
Porto Alegre, Brazil
wcarraro@hcpa.edu.br

***Abstract*— University hospitals play a strategic role in fostering innovation by integrating healthcare delivery, education, research, and technological development. This study describes the experience of the Hospital de Clínicas de Porto Alegre (HCPA) in establishing a healthcare innovation ecosystem through the Health Technology and Innovation Park (HCPATec), including the pre-incubation, incubation, and intrapreneurship programs promoted by the Starts-HCPA Technology-Based Business Incubator. A cross-sectional observational study was performed based on institutional records from the Starts-HCPA innovation programs conducted between 2023 and 2025. During this period, nearly 30 entrepreneurial initiatives were supported, covering areas such as artificial intelligence, digital health, medical devices, wearables, mental health, pharmaceuticals, rehabilitation, and assistive technologies. The programs engaged more than 150 participants and contributed to the creation and consolidation of technology-based healthcare startups. The results demonstrate the potential of university hospitals as hubs for open innovation, technology transfer, and the generation of socioeconomic impact.**

Keywords—healthcare innovation; university hospital; entrepreneurship; startups; business incubation; technology transfer.

I. INTRODUCTION

Innovation has been widely recognized as a key driver for addressing the contemporary challenges faced by healthcare systems. Population aging, increasing demand for healthcare services, and the rapid incorporation of emerging technologies require organizational models capable of translating scientific and technological knowledge into practical healthcare solutions with meaningful clinical and societal impact [1].

In this context, university hospitals occupy a strategic position by integrating patient care, education, and research, thereby providing an ideal environment for identifying real-world healthcare challenges and developing technologies tailored to users' needs [2]. In addition to concentrating clinical, academic, and scientific activities, university hospitals possess characteristics that make them particularly conducive to intrapreneurship. Although hierarchical structures, regulatory requirements, and patient safety concerns may constrain experimentation, the inherent complexity of healthcare delivery creates opportunities for innovation in processes, services, and organizational models. International evidence demonstrates that structured intrapreneurship programs foster clinical leadership, interdisciplinary collaboration, and the development of solutions aligned with institutional needs [3][4][5].

Within this context, the Hospital de Clínicas de Porto Alegre (HCPA), an academic hospital affiliated with the Federal University of Rio Grande do Sul (UFRGS), established its healthcare innovation ecosystem through the creation of the HCPATec Health Technology and Innovation Park. Located within the hospital premises, HCPATec promotes synergies between long-established scientific research activities and a technology-driven innovation approach, fostering collaboration with industry while supporting the creation of new healthcare-focused companies. To further stimulate innovation-driven

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

entrepreneurship, HCPA established the Starts-HCPA Technology-Based Business Incubator, which operates pre-incubation, incubation, and intrapreneurship programs.

Against this background, this study describes HCPA's experience in developing a healthcare innovation ecosystem through HCPATec, including the pre-incubation, incubation, and intrapreneurship programs promoted by the Starts-HCPA Incubator. To achieve this objective, a cross-sectional observational study based on institutional records from the Starts-HCPA innovation programs conducted between 2023 and 2025 was performed, as detailed in the methodology section. The results are presented according to the selected evaluation indicators, followed by a discussion highlighting the role of university hospitals as key actors within innovation ecosystems and as drivers of technology transfer. Finally, the study concludes by discussing the main contributions of this experience and its implications for strengthening healthcare innovation ecosystems.

A. HCPA's Innovation Ecosystem

HCPA's innovation model was developed based on the principles of open innovation, scientific entrepreneurship, and healthcare needs-driven technological development. To support this strategy, the HCPATec Health Technology and Innovation Park was established in 2022 to serve as a hub connecting the multiple stakeholders of the Quadruple Helix of Innovation [6] [7] within the healthcare sector, including healthcare professionals, researchers, students, startups, companies, investors, government agencies, and other relevant partners.

The international literature has emphasized that healthcare innovation ecosystems rely not only on scientific production but also on structured mechanisms capable of transforming knowledge into practical solutions. Innovation, incubation, and technology venture development programs foster capacity building, collaborative networks, and stronger connections among academia, industry, and society, thereby enhancing the ability of institutions to generate economic and social impact [8][9]. University hospitals are uniquely positioned within this context because they combine research infrastructure, highly qualified professionals, and direct access to real healthcare challenges, creating favorable conditions for innovation generation and technology transfer [3][5].

HCPA promotes innovation through a structured innovation pathway composed of five sequential stages:

1. Problem Identification: This stage focuses on recognizing concrete challenges within clinical care, hospital operations, academic activities, or healthcare management. The objective is to identify inefficiencies, risks, costs, limitations in patient care, and opportunities for process improvement. Innovation opportunities may arise from healthcare professionals, researchers, students, managers, patients, or entrepreneurs, particularly when they originate from unmet needs identified in clinical practice or institutional routines.

2. Solution Development: During this stage, research groups or multidisciplinary teams transform validated problems into potential healthcare products, services, processes, software applications, medical devices, digital platforms, or other technological solutions by developing a Minimum Viable Product (MVP). Activities include defining the value proposition, technical requirements, solution design, prototype development, functional specifications, feasibility assessment, and adaptation to the intended use environment. Healthcare innovation projects must also address ethical and regulatory requirements, patient safety, data protection, interoperability, and compatibility with clinical workflows.

3. Pre-incubation: During pre-incubation, teams receive guidance to further refine both the identified problem and the proposed MVP. This stage includes business model development, market assessment, identification of potential users and customers, cost estimation, risk analysis, and strategic planning for subsequent development phases. Pre-incubation serves as a preparatory stage in which early-stage projects are structured to evolve into more robust and sustainable innovations. It also enables the identification of initiatives with the greatest technical, institutional, commercial, or social potential to advance through the innovation pipeline. At this stage, participants are not required to have an officially registered company.

4. Incubation: The incubation program is designed for projects and startups that have reached a more advanced level of technological and business maturity. Participants receive structured support for technology consolidation, business development, legal incorporation, mentoring, networking with strategic partners, fundraising, user validation, and market entry. The primary objective is to reduce technical and commercial uncertainties, strengthen execution capabilities, and increase the long-term sustainability of the venture through the value-added services provided by the incubator. Within the HCPATec and Starts-HCPA ecosystem, this stage plays a critical role in supporting the transition from pre-incubation—whether conducted within or outside HCPA—to technology-based healthcare startups and companies.

5. Scaling: Once a solution has demonstrated initial technical and commercial feasibility, its potential for expansion to other departments, healthcare organizations, institutions, or markets is evaluated. Scaling involves process standardization, technological robustness, operational capacity, commercialization strategies, revenue models, financial sustainability, regulatory compliance, and adaptation to different implementation contexts. Within the HCPATec ecosystem, this stage may occur during the incubation period or subsequently if the startup becomes a resident company within the Technology Park.

In 2025, HCPA also launched its first Intrapreneurship Program, specifically designed for hospital employees. This initiative expanded internal innovation activities while strengthening the institutional culture of developing solutions for organizational challenges. The program complements the activities of HCPATec and the Starts-HCPA Incubator by creating structured mechanisms to identify and address improvement opportunities emerging directly from the hospital environment. Consequently, HCPA's innovation ecosystem now encompasses both the development of technology-based startups and the promotion of internal innovation initiatives

aimed at improving healthcare processes, services, and organizational practices, further reinforcing the integration of patient care, research, education, and innovation.

II. METHODOLOGY

The study employed an observational, cross-sectional design based on documentary analysis. It was considered cross-sectional because the available data were analyzed as a snapshot of the innovation programs conducted during the study period, with the consolidation of the variables available for each year.

The unit of analysis comprised each team, startup, incubated company, or intrapreneurship project registered within the innovation programs. The variables analyzed included the year in which each program cycle began, type of initiative, technological area, number of participants, completion or continuation status, formal business registration, and evidence of progression through the innovation pipeline, particularly the transition from pre-incubation to incubation.

III. RESULTS

A. *Evolution of the Starts-HCPA Innovation and Intrapreneurship Programs*

Between 2023 and 2025, the Starts-HCPA programs included pre-incubation, incubation, and intrapreneurship initiatives, covering projects in digital health, artificial intelligence, medical devices, wearable technologies, mental health, rehabilitation, biotechnology, and pharmaceuticals.

During the study period, the following were identified:

- 38 teams applied to the Pre-incubation Program, of which 20 were selected through the competitive selection process, while 3 (three) teams decided to forsake..
- 11 startups applied to the Incubation Program, of which 8 were selected during the study period.
- 11 teams applied to the 2025 Intrapreneurship Program, with 3 teams selected to participate in that year's activities.

Considering all the teams and startups that have applied, there were more than 150 participants directly involved in the innovation programs, supported by 25 institutional mentors and more than five partner companies, which contributed expertise, mentoring and value-added services throughout the programs.

TABLE I. EVALUATION OF PARTICIPATION IN THE STARTS-HCPA INNOVATION AND INTRAPRENEURSHIP PROGRAMS

Year	*Pre-incubation*	*Incubation*	*Intrapreneurship*
2023	11 teams	–	–
2024	3 teams	5 startups	–
2025	3 teams	3 startups	3 teams
Total	**17 teams**	**8 startups**	**3 teams**

The incubation program attracted fewer applicants than the pre-incubation program. This difference reflects the operational model adopted by the Starts-HCPA Incubator, which emphasizes intensive support for incubated ventures through structured mentoring, regular follow-up meetings, and continuous monitoring of startup development indicators.

In 2024, five startups were admitted to the incubation program, which has an expected duration of 18 months. In 2025, three additional startups joined the program, bringing the total number of supported companies to eight during the study period. Unlike the pre-incubation program, which has an educational focus and accommodates a larger number of teams, the incubation program concentrates resources on a smaller portfolio of ventures, enabling individualized support and closer monitoring of business development.

The results demonstrate the effectiveness of this strategy. By the end of 2025, three startups had successfully completed their first incubation cycle and had their residency renewed for 2026 to continue executing their business development plans. In addition, one startup successfully completed the graduation process after achieving its planned milestones, illustrating the progressive maturation of companies supported by the incubator. These findings highlight the incubator's role not only in attracting innovative initiatives but also in fostering the consolidation and growth of technology-based healthcare businesses.

B. *Distribution of Initiatives by Technological Area*

Approximately 60% of the initiatives focused on software, digital health, and artificial intelligence, highlighting the alignment of HCPA's innovation ecosystem with the ongoing digital transformation of the healthcare sector.

TABLE II. DISTRIBUTION OF INITIATIVES BY TECHNOLOGICAL AREA (2023-2025)

Technological Area	Number of Initiatives
Software / Digital Health	18
Artificial Intelligence	5
Digital Mental Health	4
Medical Devices and Wearables	3
Pharmaceuticals and Biopharmaceuticals	4
Therapeutic Virtual Reality	1
Healthcare Marketplaces and Digital Platforms	1
Rehabilitation	1
Social Technologies	2

C. *Participation of Professionals and Entrepreneurs*

The innovation programs engaged multidisciplinary teams composed of healthcare professionals, researchers, students, managers, and entrepreneurs.

TABLE III. NUMBER OF PARTICIPANTS BY YEAR

Year	*Participants*
2023	87
2024	21
2025	47
Total	**155**

The results demonstrate HCPA's growing integration into the regional innovation ecosystem through the activities of HCPATec and the Starts-HCPA Incubator. This integration is evidenced by the active participation of professionals from medicine, nursing, pharmacy, physiotherapy, psychology,

nutrition, engineering, information technology, and several other disciplines, reinforcing the multidisciplinary nature of healthcare innovation.

D. Startup Generation and Business Formation

The structured training pathway and continuous mentoring provided by the innovation programs enabled several pre-incubation teams to naturally progress to the incubation stage.

Among the incubated startups are:

- Florence Artificial Intelligence: home-based self-monitoring platform for multiple health parameters.
- Artpsy Software: digital platform providing group therapy for mental health treatment.
- Formula3D: personalized medicine pills manufactured using 3D printing technology.
- Lumen Corporate Mental Health: digital mental health application for organizations.
- Warux Health: software for post-discharge patient monitoring and follow-up.
- Otus Solutions: software for the management and monitoring of healthcare research projects.
- Alora: cosmetic products for skin and oral ailments from oncology treatment.
- Move Sapiens: exergame solutions for complementary treatment for ADHD.

Several initiatives evolved into formally established companies during or after participation in the innovation programs.

The progression from pre-incubation to incubation and subsequently to company formation demonstrates the effectiveness of HCPA's innovation model in fostering technology-based entrepreneurship within the healthcare sector.

E. Intrapreneurship Program

In 2025, HCPA launched its first dedicated Intrapreneurship Program. There were 15 participants organized into three project teams. The initiatives developed included:

- Precision Care: a system for monitoring the volume of contrast agents administered during endovascular procedures, contributing to improved patient safety.
- Biobankers: a digital platform designed to optimize the operational processes of the HCPA Biobank.
- Supplier Management: a system for automating supplier registration, qualification, and performance evaluation processes.

The successful implementation of the program demonstrates its ability to generate innovation directly from institutional needs, reinforcing a culture of continuous improvement within the hospital.

IV. Discussion

The findings demonstrate the progressive maturation of HCPA's healthcare innovation ecosystem.

The diversity of supported initiatives reflects strong alignment with global trends in digital health, artificial intelligence, medical devices, personalized medicine, and the digital transformation of healthcare services.

The results also reveal the existence of a structured innovation pipeline, in which ideas originating from the hospital environment evolve into research projects, startups, and incubated companies. Examples such as Florence AI, Formula3D, Artpsy, and Lumen illustrate the program's ability to foster technology-based ventures addressing real healthcare challenges.

These findings highlight the importance of institutional structures capable of supporting the entire innovation journey—from problem identification to market-ready healthcare solutions. HCPA's experience demonstrates that integrating a university hospital, a technology park, a business incubator, and structured entrepreneurship programs creates a sustainable healthcare innovation pipeline. Beyond supporting startup creation and business development, this model strengthens the connection between research, clinical practice, and technological development, expands opportunities for knowledge transfer to society, and reinforces the role of university hospitals as drivers of economic and social transformation.

It also reinforces the perspective presented by ANPROTEC [10] which recognizes innovation environments as strategic platforms that integrate research, entrepreneurship, technology transfer, and institutional collaboration, enabling innovative ideas to progress from early-stage development to market-ready solutions with societal impact.

Overall, HCPA's experience confirms that university hospitals are not only centers for healthcare delivery, education, and research but also strategic actors within innovation ecosystems and important facilitators of technology transfer.

V. Conclusion

HCPA's experience demonstrates that university hospitals can serve as structured environments for innovation, entrepreneurship, and technological development.

Between 2023 and 2025, HCPATec's innovation programs supported dozens of innovative initiatives spanning multiple healthcare domains, including artificial intelligence, medical devices, digital health, and personalized medicine.

The findings indicate that combining a health technology park with structured pre-incubation, incubation, and intrapreneurship programs represents an effective strategy for transforming unmet clinical and organizational needs into technological solutions with the potential to generate significant clinical, social, and economic impact.

The innovation ecosystem developed at HCPA may serve as a model for other healthcare institutions seeking to strengthen innovation capacity, promote technology transfer, and foster entrepreneurship.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the Brazilian Funding Authority for Studies and Projects (FINEP), Grant number 0384/22, and the Research Support Foundation of the State of Rio Grande do Sul (FAPERGS) – Grant number 25/2551-0002188-0. Their support was essential for the development of this work and the advancement of the innovation initiatives described in this study.

REFERENCES

[1] WORLD HEALTH ORGANIZATION. Global Strategy on Digital Health 2020–2025. Geneva: WHO, 2021.

[2] Thune, T.; Mina, A. Hospitals as innovators in the health-care system: A literature review and research agenda. Research Policy, v. 45, n. 8, p. 1545–1557, 2016.

[3] B. Urban and E. Wood, "The innovating intrapreneur: Recognizing the role of intrapreneurship in innovation," International Journal of Innovation Management, vol. 21, no. 7, p. 1750060, 2017. doi: 10.1142/S1363919617500603.

[4] Kuratko, D. F., & Audretsch, D. B. (2013). Clarifying the domains of corporate entrepreneurship. International Entrepreneurship and Management Journal, 9(3), 323–335.

[5] Fayad, M. A. S. (2024). Fostering leadership and innovation in UK healthcare: Insights from the NHS Clinical Entrepreneur Programme and application of clinical leadership models. JMIR Preprints.

[6] Carayannis, E. G.; Campbell, D. F. J. 'Mode 3' and 'Quadruple Helix': toward a 21st century fractal innovation ecosystem. International Journal of Technology Management, Geneva, v. 46, n. 3/4, p. 201-234, 2009.

[7] Etzkowitz, H. The Triple Helix: University-Industry-Government Innovation in Action. New York: Routledge, 2008.

[8] W. R. Garney, K. L. Wilson, K. Garcia et al., "Supporting and enabling the process of innovation in public health: The Framework for Public Health Innovation," International Journal of Environmental Research and Public Health, vol. 19, no. 16, p. 10099, 2022. doi: 10.3390/ijerph191610099.

[9] M. Bonnett, M. M. Kennedy, S. Becker and A. Panzera, "A cross-sectoral approach to innovation in public health," Journal of Public Health Innovation, vol. 12, no. 2, pp. 45–62, 2024. doi: 10.61152/hbtw2644.

[10] ANPROTEC. Estudo sobre Ambientes de Inovação e Empreendedorismo no Brasil. Brasília: ANPROTEC, 2023.

Vancolearn: Web Platform for Teaching Vancomycin Therapeutic Drug Monitoring

1st Roger Rodrigues da Silva
Programa de Pós-Graduação em Engenharia Biomédica
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
ORCID: 0009-0008-3126-7409

2nd Gustavo Martins Pires
Programa de Pós-Graduação em Engenharia Biomédica
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
ORCID: 0009-0005-4809-4506

3rd Marcia Aparecida Silva Bissaco
Programa de Pós-Graduação em Engenharia Biomédica
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
ORCID: 0000-0002-3219-2567

4th Robson Rodrigues da Silva
Programa de Pós-Graduação em Engenharia Biomédica
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
ORCID: 0000-0003-1082-0777

5th Fabiano Bezerra Menegidio
Programa de Pós-Graduação em Engenharia Biomédica
Universidade de Mogi das Cruzes
Mogi das Cruzes, Brasil
ORCID: 0000-0002-4705-8352

***Abstract*—Vancomycin therapeutic monitoring based on the AUC/MIC ratio is recommended to guide antimicrobial exposure, but interpreting the pharmacokinetic concepts involved may pose an educational challenge. This study aimed to develop the VancoLearn web platform as a support resource for teaching clinical pharmacokinetics and AUC/MIC-based vancomycin therapeutic monitoring. A responsive application was implemented featuring four sequential educational modules, deterministic pharmacokinetic simulation, automated calculation, concentration-time curve visualization, guided activities, and conversational support. The implementation enabled the demonstration, in hypothetical scenarios, of the influence of clinical and anthropometric variables on estimated clearance, 24-hour AUC, trough concentration, and the AUC/MIC ratio. The presented results are functional and demonstrative, and do not constitute evidence of educational effectiveness or clinical validation. In conclusion, VancoLearn presents initial functional viability as an educational environment. The evaluation of usability, acceptance, and learning with students and professionals constitutes the next phase of the research, which has already been approved by the Research Ethics Committee.**

***Index Terms*—Vancomycin, Therapeutic Drug Monitoring, Pharmacokinetics, AUC/MIC Ratio, Medical Education, Digital Learning.**

I. Introduction

Vancomycin is a glycopeptide antimicrobial agent used in the treatment of severe infections caused by Gram-positive bacteria, including methicillin-resistant Staphylococcus aureus (MRSA) [1]–[3]. Its clinical management requires careful attention to systemic exposure, as insufficient concentrations may compromise the therapeutic response, whereas elevated exposures are associated with an increased risk of nephrotoxicity [4]–[6]. In this context, therapeutic drug monitoring (TDM) facilitates the integrated interpretation of pharmacokinetic and clinical parameters.

Historically, vancomycin monitoring relied on serum trough concentrations, typically obtained prior to the subsequent dose [6], [7]. The consensus guidelines published in 2020 shifted toward recommending the estimation of exposure using the ratio of the 24-hour area under the concentration-time curve to the minimum inhibitory concentration (AUC/MIC), particularly for severe MRSA infections [6], [8]. For the scenarios addressed by these guidelines, an AUC range of 400 to 600 mg·h/L is utilized as a target to balance antimicrobial exposure and the risk of toxicity [6].

The application of AUC/MIC-based monitoring involves concepts such as clearance, volume of distribution, half-life, exponential decay, and concentration-time curve integration [4], [5]. The literature reports difficulties among students and healthcare professionals in learning and applying quantitative content related to drug administration [7], [9], [10]. Digital resources and simulations can support the representation of these concepts, allowing for the observation of how alterations in physiological and dosing variables modify calculated outcomes [11], [12]. However, the educational effects of these strategies depend on empirical evaluation with the target audience.

Given this context, this study aimed to develop and describe the initial functional viability of VancoLearn, an educational web platform designed for teaching clinical pharmacokinetics and AUC/MIC-based vancomycin therapeutic monitoring. The proposed system integrates theoretical content, data visualization, simulation of hypothetical scenarios, guided activities, and conversational support within a single environment. The platform was designed exclusively for educational purposes and does not constitute a validated tool for prescription or

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

dose adjustment in patients.

II. Materials and Methods

A. Study Design

This work is characterized as a methodological study of technological development with a quantitative and deterministic approach. The study encompassed architecture specification, interface implementation, pharmacokinetic modeling, and the organization of a simulated educational environment [5]. At this stage, the evaluation was restricted to the implementation and functional demonstration of the components; usability, acceptance, cognitive load, and learning were not measured.

B. Platform Development

The platform was developed as a responsive web application, with a separation between the interface, the processing of simulated parameters, and the visualization of results. The technological choices aimed to enable the immediate updating of values and curves during the exploration of hypothetical scenarios, as well as to facilitate access across different devices:

1) User Interface (Front-end): The React library with TypeScript was utilized, a combination that enables reusable components, reactive interface updating, and static data type checking. These features were selected to mitigate implementation inconsistencies and allow changes in scenario parameters to be reflected without a page reload [12].
2) Instructional Design Support and Doctor AI: Google AI Studio (https://aistudio.google.com/) was employed as an auxiliary resource for the ideation and the textual and visual refinement of the educational pathways; its suggestions were subjected to review by the authors and did not constitute an autonomous source of scientific content. The Doctor AI chatbot was included as conversational support for inquiries regarding the content presented on the platform. Its responses are generated by a language model and, therefore, may contain inaccuracies, omissions, or outdated information. The resource is identified as educational, does not replace bibliographic sources or professional judgment, and displays a disclaimer advising that clinical decisions should not be made based on its responses. The definition of a restricted knowledge base, periodic content curation, and systematic evaluation of the responses constitute measures planned for subsequent stages.
3) Graphical visualization: The Recharts library (https://recharts.org/) was utilized to represent the concentration-time curves and update the graphs following parameter modifications. This choice allows for visually correlating the scenario inputs with the calculated outcomes, without assuming, at this stage, any learning gains [12].
4) Deployment and availability: The application was hosted on the Vercel platform. The hosting enables public access (https://vancolearn.mathematicalmodels.app/) to the demonstrative version; performance, scalability, and compatibility with different access conditions remain to be systematically evaluated.

C. Educational Modules

The content was organized into four sequential modules to progress from introductory concepts to application in simulated scenarios: pharmacokinetic fundamentals, principles of TDM, interpretation of the AUC/MIC ratio, and interactive simulation. This sequence was adopted to present definitions prior to their mathematical representation and, subsequently, to allow for the manipulation of variables. The guided activities and visual feedback were designed as strategies for guided exploration [12]. At this stage of development, no formal evaluation of the pedagogical architecture by external experts was conducted; this aspect will be incorporated into the subsequent validation phase.

Module 1 – Fundamentals of Vancomycin Pharmacokinetics

Presents the first-order linear one-compartment model and the definitions of volume of distribution (V_d), clearance (Cl), elimination half-life ($t_{1/2}$), and steady state [4], [5]. The module provides the conceptual foundation necessary to interpret the subsequent simulations.

Module 2 – Therapeutic Drug Monitoring (TDM)

Presents the principles of TDM and discusses the relationship between exposure, therapeutic response, and the risk of nephrotoxicity. The content contrasts traditional monitoring based on trough concentration with the AUC/MIC-based approach and contextualizes the AUC range of 400 to 600 mg·h/L within the situations covered by the guidelines [6], [13].

Module 3 – AUC/MIC Ratio

Explores the definition and interpretation of the ratio between the 24-hour AUC and the MIC. Relates this indicator to current monitoring recommendations, highlighting that the values presented on the platform are educational estimates and not individualized treatment guidance [4], [6].

Module 4 – Interactive Simulation

Allows the input of real or hypothetical values for sex, age, weight, serum creatinine, dose, dosing interval, and MIC. These variables were selected because they participate in the implemented estimations and allow exploring, in a controlled manner, relationships between renal function, dosing regimen, and exposure. The system generates concentration-time curves and calculated indicators for comparison between scenarios [5], [6].

D. Mathematical Modeling

For instructional purposes, the computational engine implements first-order deterministic relationships and a simplified one-compartment representation [1], [5]. The calculation assumes constant parameters over the simulated interval, linear kinetics, instantaneous distribution, and the absence of

interindividual variability not represented by the inputs. Vancomycin clearance estimation utilizes a linear relationship with creatinine clearance ($CrCl$), with conversion from mL/min to L/h using the factor 0.06, according to Equation (1) [11]. The model does not incorporate Bayesian inference, population pharmacokinetic models, additional covariates, observed serum concentrations, or adjustments for special clinical situations. For these reasons, it is intended for the demonstration of mathematical relationships and not for individualized clinical calculation.

$$CL = 0.8 \times CrCl \times 0.06 \tag{1}$$

The 24-hour AUC is estimated by the ratio between the total dose administered in the period and clearance, according to Equation (2) [5]. This approximation assumes linear kinetics and conditions compatible with steady state. The implementation was verified for dimensional consistency and the direct reproduction of the presented equations; however, external validation against reference clinical software or an independent set of cases was not performed at this stage.

$$AUC = \frac{Dose}{CL} \tag{2}$$

The AUC/MIC ratio is obtained by dividing the estimated exposure by the provided MIC, according to Equation (3) [6]. The result serves to compare educational scenarios and illustrate the interpretation of the indicator. It should not be employed as a dosing recommendation or a substitute for TDM and clinical evaluation.

$$TR = \frac{AUC}{MIC} \tag{3}$$

III. Results

A. Platform Architecture

The implementation resulted in a functional web application that receives user-defined parameters, processes the described mathematical relationships, and updates the indicators and visualizations without a page reload. Fig. 1 presents the flow between the interface, the deterministic processing, and the visualization layer. The results in this section describe the observed operation of the developed version, without inferring scalability, clinical performance, or educational effectiveness.

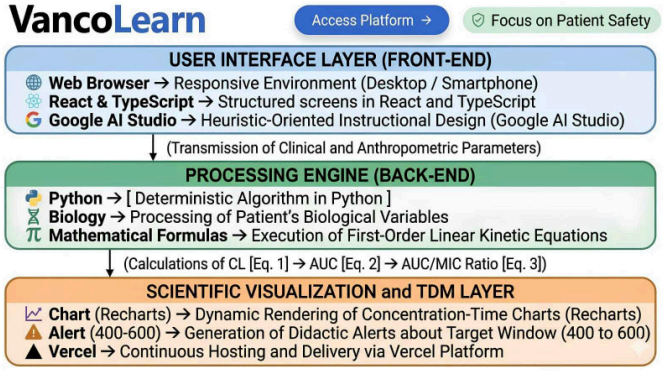


Fig. 1: General architecture and data flow of the VancoLearn platform.

Alterations in dose or physiological parameters of the virtual patient update the estimated values and displayed curves. The platform presents disclaimers regarding its exclusively educational purpose and has neither been designed nor validated to support prescription, dosage adjustment, or clinical decision-making. The perceived responsiveness of the interface updates was not subjected to formal performance testing.

B. User Interface

The interface was organized to consolidate related inputs, explanations, and results within a single visual sequence. Sliders, informational cards, and graphs were employed to explicitly demonstrate the relationship between the manipulated variables and the calculated outcomes. The suitability of this organization, its accessibility, and its effect on cognitive load have yet to be evaluated with users.

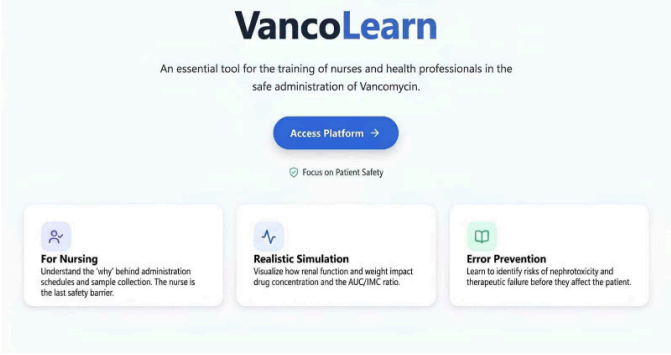


Fig. 2: Initial welcome and navigation screen of the platform.

The simulator (Fig. 3) allows for the configuration of biological sex, age, total body weight, serum creatinine, dose, dosing interval, and MIC within hypothetical scenarios. Sliders and selectors constrain input values and update the calculated outcomes. Based on sex, age, weight, and creatinine, the system estimates creatinine clearance using the Cockcroft-Gault method and presents an educational categorization of renal function. These results do not substitute for laboratory or clinical evaluation.

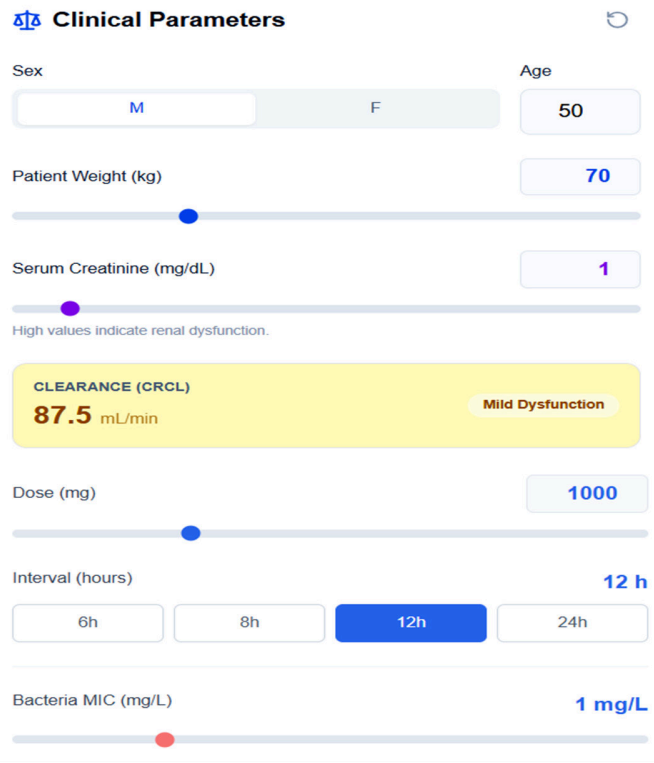


Fig. 3: Interface of the practical pharmacokinetic simulator.

The Fig. 4 demonstrates the visualization of the concentration-time curve and the calculated indicators for the configured scenario. The panel consolidates three informational cards updated following modifications to the inputs:

AUC/MIC Card: displays the calculated ratio and indicates its position relative to the reference range of 400 to 600, utilized in the module for educational interpretation purposes [6].

Trough Concentration Card: presents the estimated concentration prior to the subsequent dose and allows for its comparison, within the simulated scenario, with values traditionally utilized in monitoring [6], [7].

AUC (24h) Card: presents the estimated 24-hour exposure, in mg·h/L, utilized as the numerator for the AUC/MIC ratio and as an element for quantitative exploration [5].

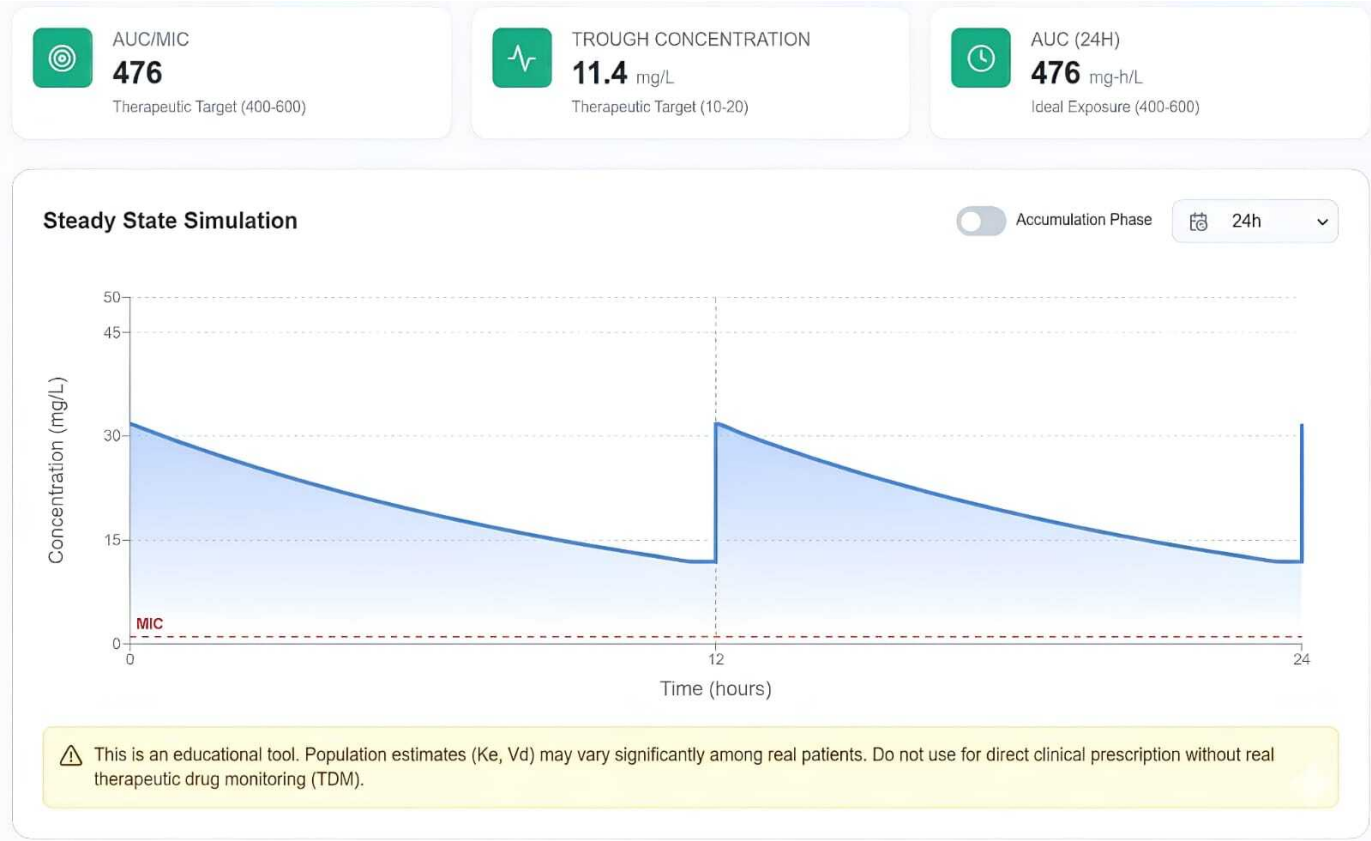


Fig. 4: Dynamic graph of the plasma concentration-time curve and PK/PD indicator panel of the VancoLearn platform.

C. Implemented Features

The platform consolidates calculation, visualization, and guided exploration routines. Table I summarizes the implemented features and their educational purpose. The presence of these components demonstrates the functional integration of the prototype, but it does not allow for the conclusion that it produces knowledge gain, reduces math anxiety, or promotes active learning.

TABLE I: Analysis and Description of the Features Implemented in the Platform

Feature	Description
AUC Calculation	Automated estimation of the 24-hour Area Under the Curve [5].
Clearance Estimation	Deterministic pharmacokinetic calculation based on renal function [11].
Concentration Curves	Interactive and dynamic plots of temporal decay [4].
Clinical Scenarios	Educational simulations based on the patient's biological and anthropometric variation.
Educational Activities	Three practical guides integrated into the "Activities" page, intended for the guided exploration of concepts [10].
AI Support (Chatbot)	Doctor AI module for conversational support, subject to the limitations of AI-generated responses and restricted to educational use.
Scenario Interpretation	Visual indication of the AUC/MIC ratio relative to the reference range, without clinical recommendation [6].

IV. Discussion

The proposed differential for VancoLearn does not reside in the creation of new pharmacokinetic equations, but in the integration, within a single educational environment, of sequential content, variable manipulation, deterministic calculation, graphical visualization, guided activities, and conversational support. This organization allows exploring the relationships between renal function, dosing regimen, and exposure indicators. The pedagogical utility of this integration remains a hypothesis to be tested with the target audience [4], [5], [12].

Platforms such as DoseMeRx and InsightRX are directed toward clinical support and utilize population models and Bayesian methods, whereas SimPHARM adopts simulations in pharmacology education [3], [7], [9], [15]. VancoLearn distinguishes itself through its specific focus on teaching AUC/MIC-based vancomycin monitoring and through the articulation between conceptual explanation, manipulable scenarios, and immediate visualization. This comparison pertains to scope and educational strategy; it does not constitute a demonstration of superiority. Doctor AI broadens the forms of interaction, but generative responses may be incorrect or outdated, requiring usage disclaimers, source review, and systematic reliability evaluation prior to any expansion of its educational role.

Limitations include the absence of empirical validation with students, faculty, or professionals, of pedagogical comparison, of usability and accessibility metrics, of objective learning assessment, and of external validation of the mathematical model against a reference clinical tool. The simplified model does not represent all pharmacokinetic variability nor special clinical situations. The experience across different student profiles, devices, and connectivity conditions, the hosting scalability, and the update process in light of new recommendations also remain to be evaluated. The next phase, already approved by the Research Ethics Committee, entails a study with students and healthcare professionals to evaluate usability, acceptance, user experience, and learning indicators. Its results will guide pedagogical, technical, and accessibility revisions.

V. Conclusion

This study described the development and demonstrated the initial functional feasibility of VancoLearn as an educational web environment for teaching vancomycin pharmacokinetics and AUC/MIC-based monitoring. The platform integrated conceptual modules, deterministic simulation, automated calculation, graphical visualization, guided activities, and conversational support, preserving the distinction between educational exploration and clinical decision-making. The results do not allow for the assertion of educational effectiveness, cognitive load reduction, or a proven contribution to professional training. The next phase of the research, already approved by the Research Ethics Committee (CAAE: 97742926.7.0000.5497), will evaluate the platform with the target audience through usability, acceptance, experience, and learning metrics.

VI. Acknowledgment

The authors express their sincere gratitude to the Universidade de Mogi das Cruzes (UMC) and the Graduate Program in Biomedical Engineering for the institutional support, infrastructure, and technical assistance that made the development of this research possible, as well as to the faculty, students, and collaborators who contributed to the conception and initial refinement of the VancoLearn platform. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 (Process No. 88887.254418/2026-00), and the Teaching, Research and Extension Support Fund (FAEP).

References

[1] Editors of Encyclopaedia Britannica, "Vancomycin," *Encyclopedia Britannica*, 2025. [Online]. Available: https://www.britannica.com. [Accessed: Feb. 12, 2026].

[2] German Center for Infection Research (DZIF), "Antibiotic therapy for MRSA," *DZIF Press*, 2025. [Online]. Available: https://www.dzif.de. [Accessed: Feb. 15, 2026].

[3] R. Goud *et al.*, "Community prevalence of methicillin and vancomycin resistant Staphylococcus aureus in Bangalore," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 44, no. 3, pp. 330–333, 2011.

[4] L. Shargel, S. Wu-Pong, and A. B. C. Yu, *Applied Biopharmaceutics & Pharmacokinetics*, 6th ed. New York: McGraw-Hill Medical, 2012.

[5] S. Storpirtis *et al.*, *Farmacocinética Básica e Aplicada*. Rio de Janeiro: Guanabara Koogan, 2011. [Online]. Available: https://integrada.minhabiblioteca.com.br. [Accessed: Feb. 10, 2026].

[6] M. J. Rybak *et al.*, "Therapeutic monitoring of vancomycin for serious methicillin-resistant Staphylococcus aureus infections: A revised consensus guideline and review by the American Society of Health-System Pharmacists, the Infectious Diseases Society of America, the Pediatric Infectious Diseases Society, and the Society of Infectious Diseases Pharmacists," *American Journal of Health-System Pharmacy*, vol. 77, no. 11, pp. 835–864, 2020.

[7] R. Morales Júnior *et al.*, "Monitoramento terapêutico de vancomicina: barreiras na implementação do modelo AUC/MIC na rotina hospitalar," *Revista Saúde em Foco*, vol. 15, no. 2, pp. 78–89, 2023.
[8] J. Liu *et al.*, "Is it time to recommend AUC-based vancomycin dosing? A survey of hospitals in China," *Frontiers in Pharmacology*, vol. 15, p. 1370040, 2024.
[9] E. Ampe *et al.*, "Implementation of a clinical decision support system for vancomycin dosing: a-before-and-after study," *International Journal of Antimicrobial Agents*, vol. 55, no. 3, p. 105871, 2020.
[10] A. G. Melo, M. Struchiner, and J. B. Frant, "A matemática da administração de medicamentos," *Revista de Estudos e Pesquisas sobre Ensino Tecnológico*, vol. 8, p. e184222, 2022.
[11] N. Holford *et al.*, "A physiological approach to renal clearance: From premature neonates to adults," *British Journal of Clinical Pharmacology*, vol. 90, no. 4, pp. 1066–1080, 2024.
[12] R. R. Silva *et al.*, "Tecnologias educacionais e o ensino de farmacocinética: uma revisão sistemática do cenário brasileiro," *Revista Brasileira de Informática na Educação*, vol. 31, pp. 112–135, 2023.
[13] S. S. Alghanem *et al.*, "Knowledge, Attitude, and Practice of Clinicians towards Vancomycin Therapeutic Drug Monitoring," *Antibiotics*, vol. 12, no. 3, p. 556, 2023.
[14] L. M. Fleuren *et al.*, "Clinically relevant pharmacokinetic knowledge on antibiotic dosing among intensive care professionals is insufficient: a cross-sectional study," *Critical Care*, vol. 23, no. 1, p. 185, 2019.
[15] A. Albassam *et al.*, "Vancomycin therapeutic drug monitoring: knowledge and practice among pharmacists," *Journal of Pharmaceutical Policy and Practice*, vol. 15, no. 1, p. 43, 2022.

Characterization of 3D-Printed PLA Composites for Pediatric CT Phantom

Azir R. B. Pires
Hospital de Clínicas Complex of the Federal University of Paraná (CHC-UFPR/Brazil)
Federal University of Technology–Parana (UTFPR/Brazil)
Curitiba, Brazil ORCID: 0009-0003-0186-4783

Erik Matheus Kachel
Department of Electronics Engineering Federal University of Technology–Parana (UTFPR/Brazil)
Curitiba, Brazil
ORCID: 0009-0005-6082-8939

John H. Hamann
Hospital de Clínicas Complex of the Federal University of Paraná (CHC-UFPR/Brazil)
Federal University of Technology–Parana (UTFPR/Brazil)
Curitiba, Brazil
ORCID: 0009-0004-0879-421X

Frieda Saicla Barros
Postgraduate Program in Biomedical Engineering Federal University of Technology–Parana (UTFPR/Brazil)
Curitiba, Brazil
ORCID: 0000-0002-3962-1192

Cristhiane Gonçalves
Postgraduate Program in Biomedical Engineering Federal University of Technology–Parana (UTFPR/Brazil)
Curitiba, Brazil
ORCID: 0000-0002-3894-2971

***Abstract*—Introduction: Dosimetry and quality control in pediatric computed tomography (CT) require physically realistic phantoms that replicate the radiological properties of the developing cranial bone. Objective: To validate PLA and PLA Stonefill materials, processed by FDM 3D printing at 220 °C, for simulation of the cranial bone of a 2-year-old child, characterizing the Hounsfield Units (HU) as a function of the filling percentage from 10% to 100% and digitized on a Canon Aquilion Prime SP CT scanner (80, 100 and 120 kVp; 200 mA; 100 mAs) with FC21 and FC30 kernels. The analysis included volumetric ROI, linear modeling HU = a·(%filling)+b, and spatial line profiles. Results: Both materials showed a significant linear correlation with the infill (R^2 = 0.956 for PLA; R^2 = 0.966 for PLA Stonefill), with slope coefficients of ≈10 HU/% and ≈14 HU/%, respectively. The inter-kVp variation at 100% filling was 29.5% (Δ = 130.17 HU) for the PLA Stonefill. Inter-kernel variation was less than 15 HU (PLA) and less than 39 HU (Stonefill). Line profiles confirmed the absence of structural voids at 100% filling, with a plateau difference of 325 HU between the two materials. Conclusion: 100% filling of PLA Stonefill reproduces the cortical bone of the pediatric skull (300–600 HU; 80 kVp: +440.7 HU); 80% and 90% fillings simulate muscle tissue and trabecular bone, respectively. Conventional PLA covers lung tissue (10–30%), adipose tissue (≈90%), and dense soft tissue (100%). The linear model allows for fine-tuning of attenuation with each 1% increase in filling.**

***Index Terms*—3D printing, PLA, PLA Stonefill, pediatric phantom, computed tomography, Hounsfield Units, cranial bone, FDM, tissue equivalence.**

I. Introduction

Computed tomography (CT) plays a relevant role in pediatric diagnostic imaging, particularly in cranial evaluation of children under 3 years of age [1]. Optimizing pediatric computed tomography (CT) protocols requires anthropomorphic phantoms that mimic the attenuation properties of developing tissues, thus enabling better quality control and reducing patient exposure to ionizing radiation [2]. In pediatric patients, especially around two years of age, the calvaria presents very different radiological characteristics from those of an adult. In this age group, the structure is predominantly monocortical, transitioning to the adult trilaminar configuration (outer table, diploë, and inner table). Diploë identifiable by CT scan is present in more than 90% of cases and only from 3 to 5 years of age in the frontal and posterior parietal regions [3]. The average calvaria thickness in this age group is between 3.7 mm (anterolateral) and 3.9 mm (posterolateral) [4], with a thickness of less than 3 mm in 85–92% of patients under 4 years of age [5]. Liu et al. [6] developed a normative model based on 2,068 individuals (0–10 years) with a predictive uncertainty on the Hounsfield scale of 76.7 ± 91.5 HU, reflecting the high heterogeneity of this segment. Gajawelli et al. [7] confirmed, using structural MRI in a three-dimensional model of 56 children (0–3 years), that cranial thickness and geometry vary with age and anatomical location. Delye et al. [8] demonstrated that cranial bone density increases at an accelerated rate in the first 3 years of life without reaching a plateau until the third decade of life.

From a densitometric point of view, the calvaria at 2 years of age presents a mosaic of Hounsfield Units (HU): Yamada et al. [9] documented an upward trajectory of density from birth, with a plateau around 20 years of age with a HU value around 1,164 ± 143 HU; in the parietal and frontal regions of children aged 1–3 years, the values are predominantly in the range of 300–600 HU [6], [9]. Takahashi et al. [10] identified that only specific regions such as the mastoid reach 1000 HU around 1.7 years. Rowbotham et al. [11] documented adult neurocranial variation from 78 to 1323 HU, a range that the pediatric cranial bone still traverses throughout development. The consolidated reference ranges for adjacent tissues are: inflated lung tissue between −500 and −1000 HU; adipose tissue between −50 and −100 HU; muscle tissue between +10 and +40 HU; spongy bone between +100 and +300 HU; dense cortical bone between +300 and +1000 HU [12]–[14].

Fused deposition modeling (FDM) additive manufacturing allows for fine-tuning of the effective density of phantoms through control of the percentage of internal filling [15], [16]. Hatamikia et al. [16] obtained a range of −482 to 968 HU with a single filament; Arnold et al. [17] achieved an increase of 196 HU by weight percent with PLA/bismuth oxide composites; Okkalidis et al. [18] confirmed the feasibility of bone tissue replication by FDM; and Mei et al. [19] reproduced bone marrow with differences smaller than 15 HU using calcium-doped filament. Conventional PLA ($\rho \approx 1.24$ g/cm3) is the most widely used

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

thermoplastic in medical physics applications, being characterized as equivalent to soft tissue in certain configurations [20], but with limitations for bone simulation even at 100% filling [20]. Mineral-filled composite filaments expand this range: Ahmed, Buschmann, et al. [15] reported values of 57 to 460 HU with mineral PLA, and Lai et al. [22] obtained 547 HU with PLA supplemented with calcium carbonate for adult skull simulation.

Despite these advances, a gap persists in the specific validation of FDM materials for cranial bone simulation, particularly in the 2-year-old age range [6]. Given this scenario, the present study aims to validate two FDM 3D printing materials — conventional PLA and PLA Stonefill — for simulating pediatric skull bone at 2 years of age, through their characterization in Hounsfield Units under the kVp variation of the pediatric skull protocol in CT and reconstruction algorithms.

II. Methodology

A. Printout of Test Specimens

Test specimens were manufactured by FDM additive manufacturing (Fused Deposition Modeling) using two commercial filaments with a diameter of 1.75 mm, both produced by Earth Renewable Technologies — ERT (Curitiba, PR, Brazil). Each piece was printed in the form of a cylinder 20 mm in diameter by 20 mm high, for a total of 20 cylinders (10 of conventional PLA and 10 of PLA Stonefill), with infill densities ranging from 10% to 100%, in 10% increments.

1) Conventional PLA: bio-based resin FC 10030 [23], with a nominal density of 1.24 g/cm³, tensile modulus of 3500 MPa and tensile strength of 50 MPa (ISO 75-1/2). The manufacturer's datasheet establishes a maximum processing temperature of 230 °C; the extrusion temperature adopted in this study (220 °C — Table I) respects this limit with a safety margin of 10 °C, avoiding thermal degradation of the polymer chain and ensuring rheological stability during deposition.

TABLE I. FDM 3D Printing Parameters

Parameter	Specification
Slicing Software	Creality Print V2.10.4414
Nozzle Diameter	0.6 mm
Filament Diameter	1.75 mm
Filament Density (PLA)	1.24 g/cm3
Filament Density (PLA Stonefill)	1.35 g/cm3
Extruder Temperature	220 °C
Table Temperature	50 °C
Chamber Temperature	35 °C
Number of Layers	83
Maximum Height Z	20 mm
Contraction / Shrinkage Comp.	1.5 mm / F2400
Fill Pattern	Gyroid

2) PLA Stonefill: FC 50325 composite [24], formulated from a physical blend of PLA with an antiblocking additive based on calcium carbonate ($CaCO_3$), giving it a marble-like texture and a nominal density of 1.35 g/cm³. The material has a tensile strength of 16 MPa and an elongation at break of 400% (ASTM D638-22). The maximum processing temperature limit is also 230 °C, respected by the adopted extrusion temperature of 220 °C. The presence of $CaCO_3$ in the polymer matrix is primarily responsible for shifting the attenuation values to higher ranges of Hounsfield Units compared to pure PLA, due to the effective atomic number of the composite.

The printing parameters were standardized as per Table I. The relationship between filling density and HU values was modeled by simple linear regression, described by the model:

$$HU = a \cdot (\%filling) + b \qquad (1)$$

where a represents the slope coefficient (HU/%), corresponding to the change in HU per percent unit of filling, and b is the linear intercept corresponding to the HU value extrapolated to 0% infill (i.e., contribution of the polymer matrix and residual air).

B. CT Acquisition Protocols

The acquisitions were made using the computed tomography equipment model Canon Medical Systems Aquilion Prime SP (EBSERH-UFPR). For acquisition, the CRANIO 0–2 YEARS protocol was used with tube voltage variation as per Table II. The images were reconstructed using two algorithms: FC21 (soft) and FC30 (bone).

TABLE II. CT Acquisition Parameters by Protocol

Parameter	Prot. 1	Prot. 2	Prot. 3
Voltage (kVp)	80	100	120
Current (mA)	200	200	200
Exposure (ms)	500	500	500
Charge (mAs)	100	100	100
Rotation (s)	0.5	0.5	0.5
Collimation (mm)	20	20	20
Thickness (mm)	1.0	1.0	1.0
Kernel	FC21/FC30	FC21 (soft)	FC21 (soft)
Filter	Al-Cu	Al-Cu	Al-Cu

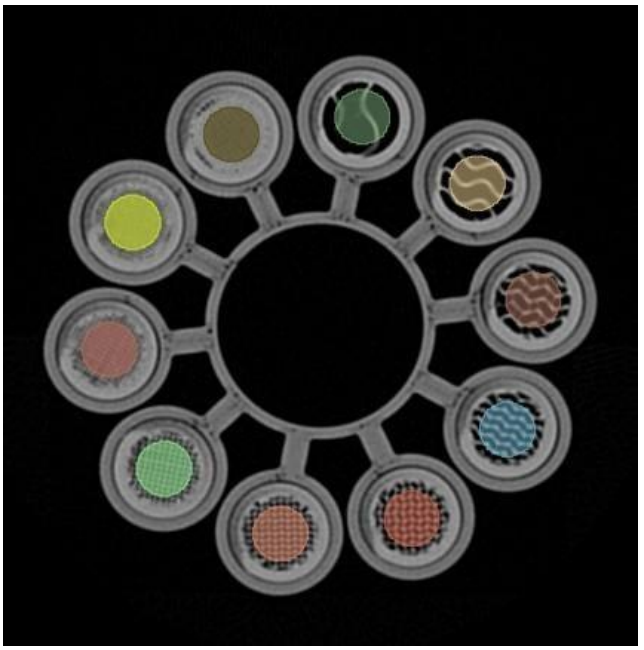


Fig. 1. Axial CT scan of ten PLA and PLA Stonefill specimens showing the ROI regions for evaluation.

C. Data Analysis

HU values were extracted from regions of interest (ROI), as shown in Fig. 1, segmented in each specimen using the free software 3D Slicer, version 5.12.0. The mean and standard deviation were determined for each combination of material, filler, protocol, and reconstruction kernel. The measurements obtained with the soft kernel (FC21) were compared with the bone kernel (FC30) to evaluate the influence of the reconstruction algorithm on the quantitative values of HU. The results were correlated with reference values of biological tissues [12], [13] and with published data on pediatric cranial bone density [6], [8], [10].

III. Results

A. Gradual Progression of Radiopacity in the Test Specimens

Fig. 2 shows the ten PLA test specimens distributed across their respective filling levels, visually illustrating the density spectrum produced by the parametric variation.

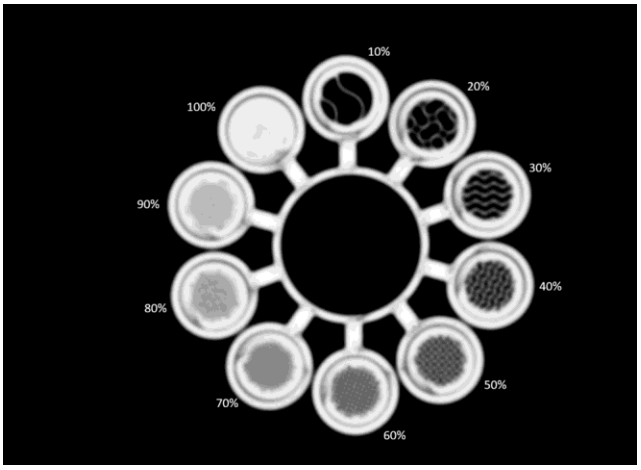


Fig. 2. Axial CT slice of the ten PLA specimens. The gradual progression of radiopacity — translucent pattern (10%) to homogeneous and dense appearance (100%) — demonstrates the sensitivity of the FDM method in modulating radiological density, as in the model of Eq. (1).

B. HU Values — PLA and PLA Stonefill (Soft Kernel, 80 kVp)

Table III presents the mean values and standard deviations of HU at all ten fill levels acquired at 80 kVp, with FC21 soft-kernel reconstruction for PLA and PLA Stonefill materials.

A monotonic relationship was observed between the percentage of filling and the HU value for both materials. PLA Stonefill showed systematically higher values than conventional PLA, with an increasing difference: from 51.40 HU at 10% to 332.51 HU at 100%. The linear fit by Eq. (1) produced coefficients of determination of $R^2 = 0.956$ for PLA and $R^2 = 0.966$ for PLA Stonefill, confirming that the linear model describes with high fidelity the densitometric response of the materials in the evaluated range. Fig. 3 presents the experimental data with uncertainty bars (±SD) and their respective regression lines.

TABLE III. Tissue Equivalence of Filling Materials (80 kVp)

Infill (%)	PLA (HU)	Stonefill (HU)
10	−904.98 ± 155.84	−853.58 ± 205.06
20	−801.31 ± 161.17	−719.09 ± 215.94
30	−698.04 ± 125.97	−603.71 ± 164.89
40	−592.88 ± 79.84	−473.33 ± 108.07
50	−493.36 ± 42.17	−349.52 ± 62.71
60	−388.81 ± 19.92	−223.50 ± 40.12
70	−280.85 ± 12.85	−90.40 ± 38.62
80	−177.70 ± 12.04	+36.15 ± 31.49
90	−87.89 ± 11.46	+148.99 ± 32.79
100	+108.22 ± 14.25	+440.73 ± 43.52

C. Tissue Equivalence by Filling Range

The HU values obtained were compared with the reference ranges for biological tissues (Table IV).

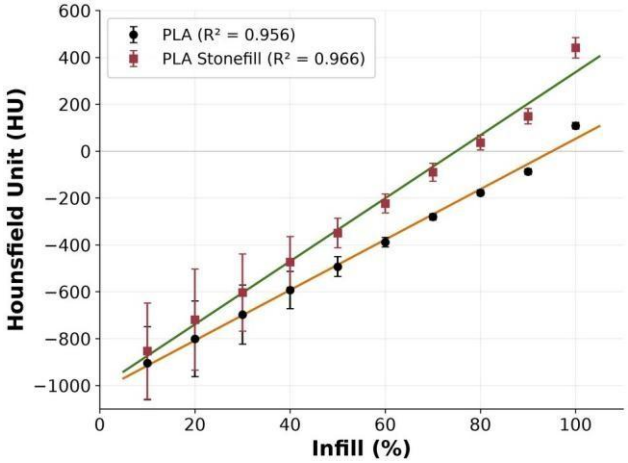


Fig. 3. Linear regression between percentage of filling and mean HU values (80 kVp, kernel FC21) for PLA (black circles, orange line, $R^2 = 0.956$) and PLA Stonefill (red squares, green line, $R^2 = 0.966$). Error bars represent ±1 SD. The greater slope of the Stonefill line (a ≈ 14 HU/%) reflects the contribution of $CaCO_3$ to the attenuation coefficient.

TABLE IV. Tissue Equivalence of Materials by Filling Range (80 kVp)

Infill (%)	PLA (HU)	Stonefill (HU)
10–30	Lung tissue (−698 to −905 HU)	Lung tissue (−604 to −854 HU)
40–60	Lung/adipose transition (−389 to −593 HU)	Pulmonary/adipose transition (−224 to −473 HU)
70	Adipose transition (−281 HU)	Adipose tissue (−90 HU)
80	Adipose (−178 HU)	Soft tissue/muscle (+36 HU)
90	Adipose (−88 HU)	Trabecular bone (+149 HU)
100	Dense soft tissue (+108 HU)	Pediatric cortical bone (+441 HU)

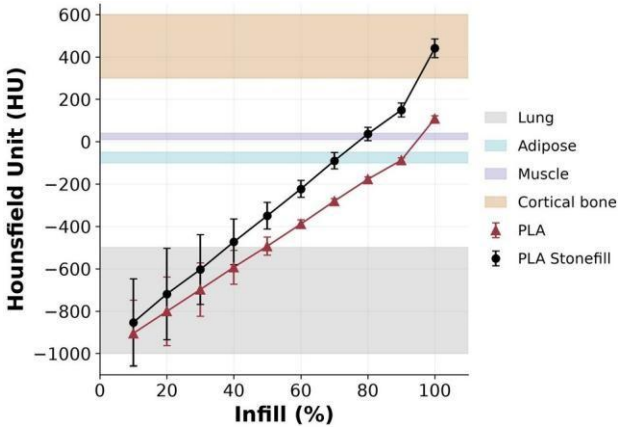


Fig. 4. Tissue equivalence as a function of filling percentage (80 kVp, kernel FC21). Red triangles: conventional PLA; black circles: PLA Stonefill. Colored bands delimit reference ranges in HU: inflated lung (gray, −500 to −1000 HU), adipose tissue (green, −50 to −100 HU), muscle/soft tissue (blue, +10 to +40 HU) and pediatric cortical bone tissue (beige, +300 to +600 HU) [6], [9]. Stonefill reaches the bone range from 90% filling, while PLA remains below dense soft tissue throughout the evaluated range.

It is noteworthy that PLA Stonefill at 80% filling (+36.15 HU) is located exactly in the muscle/soft-tissue range (+10 to +40 HU) [5]. Conventional PLA at 90% (−87.89 HU) corresponds to the adipose tissue range (−50 to −100 HU) [13]. Fig. 4 graphically summarizes this correspondence, superimposing the experimental values onto the reference ranges of the main biological tissue groups.

D. Qualitative and Spatial Analysis of Homogeneity

Fig. 5 shows an axial CT image of ten specimens with different percentages of filling identified directly in the samples, acquired with a bone kernel (FC30). The numerical

indication allows for the correlation between the printing configuration and the radiological density, highlighting an increase in radiopacity. Fig. 6 shows an axial CT scan of a phantom containing a sample of PLA Stonefill at ten filling levels, highlighting the internal geometric macrostructure of the specimens and allowing a qualitative assessment of the spatial distribution of the material in each infill configuration.

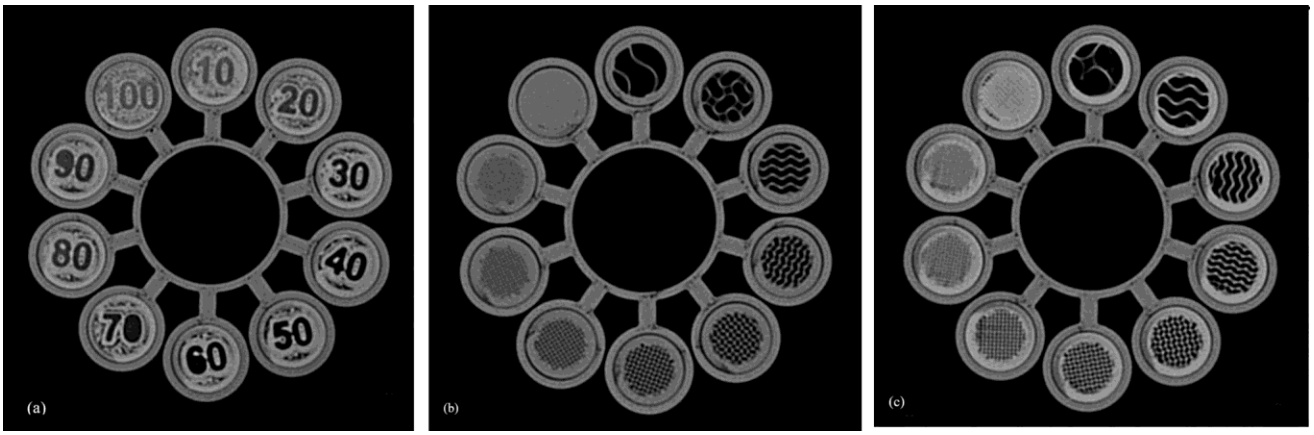


Fig. 5. Axial CT image (bone kernel — FC30) of the ten specimens. (a) Percentage of filling recorded on each disc (10% to 100%); (b) PLA specimens at the ten filling levels — the bone kernel shows the microstructure of the infill pattern at intermediate percentages (40–70%), where the filament network is clearly distinguishable from the interstitial air space, while in high-density samples (90–100%) the appearance is uniform with no discernible internal structure; (c) PLA Stonefill specimens at the ten filling levels — the greater radiopacity of the calcareous compound ($CaCO_3$) is evidenced by the more intense brightness in the filaments even at low percentages, contrasting strongly with the interstitial air space. In fillings above 80%, the appearance approaches the homogeneity of pediatric cortical bone, consistent with the HU values in Table III.

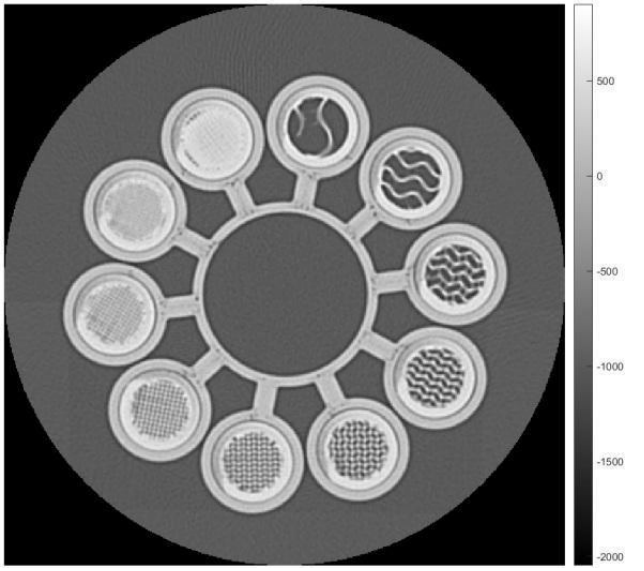


Fig. 6. Axial CT scan of the ten specimens showing the geometric macrostructure of the filling patterns at the ten levels of infill (grayscale sidebar).

A continuous progression of apparent radiological density (grayscale) is observed from the lowest fillings (10% and 20%) — a clearly defined reticular pattern with ample interstitial air space — up to 100%, where the appearance is close to a solid filling. The visual distinction of the internal layers in the 40–70% fillings confirms that the CT scanner individually differentiates the material layers and air spaces, so that the HU values measured by ROI express the weighted average between the two phases, as described by Bustillo et al. [25].

Figs. 7 and 8 show the spatial line attenuation profiles obtained in MATLAB. For the analysis, the total diameter of the 100%-filling specimens was evaluated for PLA and PLA Stonefill. These attenuation profiles demonstrate that there are no abrupt depressions (negative values in the plateau region), confirming the absence of structural voids in the test specimens. Direct comparison reveals an important point: the perimeter peaks of PLA (~450–490 HU) numerically coincide with the internal plateau of Stonefill (~433 HU), demonstrating that the overlapping perimeter walls of PLA reach a radiopacity equivalent to Stonefill in solid configuration — a value more than four times higher than the internal plateau of PLA itself. This intrasectional heterogeneity, inherent to FDM phantoms, should be considered when defining ROIs for quantitative measurements and when designing thin-walled geometries [25]. At the same time, Stonefill 100% confirms its suitability to simulate the monocortical architecture of the 2-year-old pediatric calvaria — with perimeter peaks in the cortical range and an internal plateau in the trabecular/lower-cortical bone range [9], [10] — while PLA 100% remains in the dense soft-tissue range (+108 HU), suitable for simulating non-bone tissues [12], [13].

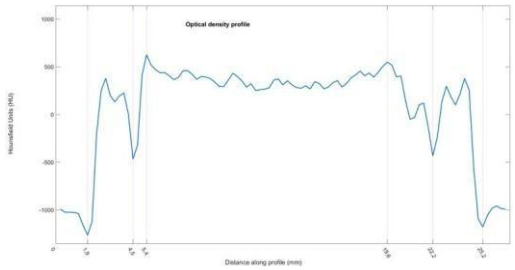


Fig. 7. Spatial line profile of the 100% PLA Stonefill sample at 80 kVp. Perimeter peaks (627 and 550 HU) correspond to the walls (shells) and a homogeneous central plateau with a median of 433 HU. Horizontal axis: distance along the profile (mm).

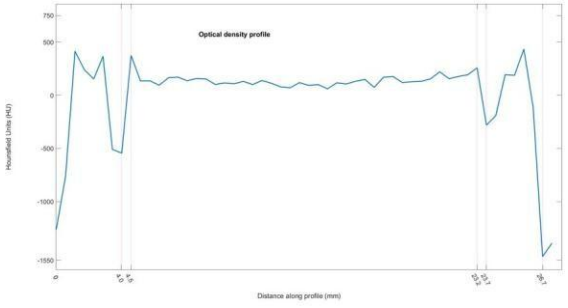


Fig. 8. Line profile of the 100% PLA sample at 80 kVp. Perimeter peaks (450 and 490 HU) correspond to the walls (shells) and a homogeneous central plateau with a median of ≈108 HU. Horizontal axis: distance along the profile (mm).

E. Voltage Dependence — PLA

Table V presents the average HU values of PLA at the three kVp values. PLA showed low energy dependence: maximum variation of 8.91 HU between 80 and 120 kVp (for 100% filling). At other filling levels, the variation was less than 7.5 HU. Thus, we observed that PLA has an effective atomic number close to that of water [13], [17].

TABLE V. AVERAGE HU VALUES OF PLA AS A FUNCTION OF KVP (KERNEL FC21)

Infill (%)	80 kVp	100 kVp	120 kVp
10	−904.98	−898.93	−900.89
20	−801.31	−798.73	−796.90
30	−698.04	−694.97	−692.32
40	−592.88	−589.30	−586.46
50	−493.36	−488.19	−486.02
60	−388.81	−385.12	−383.18
70	−280.85	−279.11	−277.63
80	−177.70	−176.62	−175.76
90	−87.89	−87.62	−87.41
100	+108.22	+115.03	+117.13

F. Voltage Dependence — PLA Stonefill

Table VI presents the average HU values of PLA Stonefill at the three kVp values. PLA Stonefill exhibited pronounced energy dependence, with a variation of up to 130.17 HU between 80 and 120 kVp (at 100% infill), corresponding to a 29.5% reduction. This effect became progressively more intense with increasing infill, indicating a growing contribution from the photoelectric effect of the filament additive [13], [17].

TABLE VI. AVERAGE HU VALUES OF PLA STONEFILL AS A FUNCTION OF KVP (KERNEL FC21)

Infill (%)	80 kVp	100 kVp	120 kVp
10	−853.58	−858.18	−873.62
20	−719.09	−733.47	−742.29
30	−603.71	−616.50	−625.61
40	−473.33	−494.68	−508.92
50	−349.52	−375.06	−391.29
60	−223.50	−254.72	−273.12
70	−90.40	−125.48	−147.85
80	+36.15	−6.18	−28.46
90	+148.99	+99.43	+73.19
100	+440.73	+359.86	+310.57

G. Comparison between Reconstruction Kernels — PLA

Table VII compares the HU values of PLA obtained with a soft kernel (FC21) and a bone kernel (FC30) at 80 kVp.

The difference between kernels for PLA ranged from 0.25 HU (100% fill) to 14.19 HU (10% fill), demonstrating high agreement between the reconstruction algorithms, especially at high fill values.

TABLE VII. PLA COMPARISON: SOFT KERNEL (FC21) VS. BONE KERNEL (FC30) AT 80 KVP

Infill (%)	FC21 (HU)	FC30/bone (HU)	|Δ| (HU)
10	−904.98	−890.79	14.19
20	−801.31	−791.69	9.63
30	−698.04	−692.58	5.46
40	−592.88	−584.03	8.84
50	−493.36	−484.63	8.73
60	−388.81	−382.74	6.07
70	−280.85	−276.21	4.64
80	−177.70	−174.03	3.67
90	−87.89	−87.45	0.44
100	+108.22	+107.96	0.25

H. Comparison between Reconstruction Kernels — PLA Stonefill

Table VIII compares the HU values of PLA Stonefill with a soft kernel vs. bone kernel at 80 kVp. For PLA Stonefill, the difference between kernels ranged from 0.02 HU (80% fill) to 38.45 HU (100% fill). Agreement was excellent at intermediate fills (60–80%), with the most pronounced divergence at 100% fill.

TABLE VIII. COMPARISON OF PLA STONEFILL: SOFT KERNEL (FC21) VS. BONE KERNEL (FC30) AT 80 KVP

Infill (%)	FC21/soft (HU)	FC30/bone (HU)	|Δ| (HU)
10	−853.58	−845.64	7.94
20	−719.09	−707.64	11.90
30	−603.71	−591.59	12.12
40	−473.33	−468.70	4.63
50	−349.52	−344.03	5.49
60	−223.50	−221.01	2.49
70	−90.40	−88.65	1.75
80	+36.15	+36.17	0.02
90	+148.99	+142.90	6.10
100	+440.73	+402.29	38.45

IV. DISCUSSION

A. Suitability of Materials for the Simulation of Biological Tissues

The results demonstrate that the two materials evaluated allow coverage of a wide range of the HU spectrum of biological tissues. Considering the reference values consolidated in the literature — inflated lung tissue (−500 to −1000 HU), adipose tissue (−50 to −100 HU) and muscle tissue (+10 to +40 HU) [12], [13] — a quantitative correspondence with a good degree of agreement is identified.

Conventional PLA in 10–30% fillings produced averages of −905.0 ± 155.8 to −698.0 ± 126.0 HU (coefficient of variation, CV = 17.2% to 20.1%), placing it within the range of inflated lung tissue. The absolute difference between the average obtained at 10% (−905.0 HU) and the upper limit of the lung range (−500 HU) is 405 HU, demonstrating that fillings of 30–40% can be adjusted to simulate partially inflated or emphysematous lungs. At 90% filling (−87.89 ± 11.5 HU, CV = 13.1%), PLA coincides with the central range of adipose tissue (−50 to −100 HU), with an absolute mean deviation of only 12.1 HU from the midpoint of the range (−75 HU). At 100% (+108.22 ± 14.2 HU, CV = 13.2%), it is positioned in the transition between dense soft tissue and trabecular bone. These values are consistent with Walt et al. [20], who characterize PLA as equivalent to soft tissue, and with Kamomae et al. [26], who reported 126 ± 18 HU for a PLA phantom wall — a difference of only 17.8 HU (< 14%) compared to the 100% PLA in this study.

PLA Stonefill exhibits a shifted equivalence range toward higher densities with systematically more variable behavior (CV = 9.9% to 24.1%). At 70% (−90.40 ± 38.6 HU), it is located in the adipose-tissue range with 79.3% overlap with the reference interval. At 80% (+36.15 ± 31.5 HU), it reaches agreement with muscle tissue (+10 to +40 HU): the average of +36.15 HU is located only 11.2% from the upper limit of the muscle range, with an overlap of ±1 SD fully covering the target range. At 90% (+149.0 ± 32.8 HU), it corresponds to the cancellous-bone range. At 100% (+440.7 ± 43.5 HU at 80 kVp), it falls within the pediatric cortical-bone range.

B. Statistical Analysis with Pediatric Cranial Bone

Validating PLA Stonefill as a pediatric cranial-bone simulator for the 2-year-old age group requires quantitative comparison with densitometric data available in the literature for this age.

Pediatric reference range. Yamada et al. [9], in a study with 495 individuals (ages ranging from months to 80 years), documented that the peak cranial-density histogram gradually increases from birth, reaching an adult plateau of

1164 ± 143 HU (women, 20–30 years). The developmental curve indicates that children aged 1 to 3 years are in the range of 300–600 HU for parietal and frontal regions [6], [9]. Delye et al. [8], in a cohort of 187 individuals (0–20 years), indicate that the rate of mineralization is highest in the first 3 years of life, corresponding to approximately 25–50% of the adult plateau in that time window.

C. Quantitative Agreement

PLA Stonefill at 100% filling at 80 kVp (440.73 ± 43.5 HU) is located at the 47th percentile of the pediatric reference range (300–600 HU), with an overlap of ±1 SD (397–484 HU) covering 62% of the target range. At 100 kVp (359.86 ± 38.4 HU), the material moves to the 20th percentile of the same range, and at 120 kVp (310.57 ± 33.8 HU) it approaches the lower limit. Taking the pediatric midpoint of 450 HU (estimated from the curves of Yamada et al. [9] for the age of 2 years), the absolute deviation of PLA Stonefill at 80 kVp is only −9.27 HU (2.1%), indicating excellent agreement.

Regional heterogeneity. Takahashi et al. [10] demonstrated that only the lateral surface of the mastoid reaches ∼1000 HU by around 1.7 years of age, while the parietal and frontal regions remain significantly below this value. Rowbotham et al. [11] documented adult neurocranial variation from 78 HU (mastoid) to 1323 HU (temporal squama). The 310–441 HU range of PLA Stonefill (depending on kVp) is therefore specifically suited for simulating the parietal and frontal regions — anatomically more relevant for pediatric cranial phantom structures in radiological-protection and neurosurgical-planning studies [14], [26].

Comparison with phantoms for adult skulls. Lai et al. [22] obtained 547 HU with PLA/calcium carbonate for adult skull simulation — a value 24.1% higher than PLA Stonefill at 80 kVp and 76.1% higher at 120 kVp, a difference consistent with lower pediatric mineralization. Kim et al. [27] reported 680 HU for a PLA-plaster phantom (adult skull), representing a difference of 54.3% compared to PLA Stonefill at 80 kVp. Ahmed, Buschmann, et al. [15] reported 57 and 460 HU with PLA of variable mineral load. The value of 440.73 HU found in the present study is close to the upper limit of that research (460 HU), showing a difference of only 4.4%. These data confirm that PLA Stonefill is positioned in the pediatric densitometric window, distinct from the values reported for mature-bone simulation.

Variability and precision. The coefficient of variation of PLA Stonefill 100% (CV = 9.9% at 80 kVp) is lower than the predictive error of the normative model of Liu et al. [6] (76.7 ± 91.5 HU, CV > 100% in children under 3 years old), indicating that the reproducibility of the printed material is superior to interindividual biological variability. The standard deviation of 43.5 HU for Stonefill 100% at 80 kVp corresponds to only 30% of the target-range amplitude (300 HU), suggesting adequate stability for use as a reference in calibration protocols.

D. Energy Dependence and Spectral Behavior

The variation of 130.17 HU (29.5%) observed for PLA Stonefill 100% between 80 and 120 kVp (Δ = 440.73 − 310.57 HU) is clinically relevant, since pediatric protocols frequently use 80 to 100 kVp for dose reduction [26]. This spectral behavior is attributed to calcium carbonate ($CaCO_3$) in the Stonefill filament: the high effective atomic number of calcium (Zeff ≈ 14.0) increases the photoelectric contribution ($\propto Z^4/E3$) at low energies, mimicking biological bone (hydroxyapatite, Zeff ≈ 13.8) [13], [17]. For comparison, pediatric cranial bone also exhibits energy dependence: Yamada et al. [9] used 120 kVp in their reference measurements, while values obtained at 80 kVp in pediatric clinical studies are typically 20–30% higher [15]. The 29.5% variation of PLA Stonefill between 80 and 120 kVp is consistent with this magnitude, suggesting that the material replicates not only the absolute density but also the spectral behavior of mineralized pediatric cranial bone.

Conventional PLA showed a variation of only 8.91 HU (8.2%) between 80 and 120 kVp at 100% filling, behaving as a near water-equivalent material (Zeff ≈ 7.0). The inter-kVp coefficient of variation of 4.1% is lower than that observed for biological soft tissue (∼5–8%) [13], confirming the spectral stability of pure PLA. This spectral difference between the two materials (8.2% vs. 29.5%) is analogous to the difference between soft tissue and biological bone [17].

E. Influence of the Reconstruction Kernel

The comparison between soft (FC21) and bone (FC30) kernels at 80 kVp revealed distinct behaviors depending on the material and the filling level.

For PLA, the correlation between the means of the two kernels was excellent for all levels ($r > 0.999$), and the maximum difference of 14.19 HU represents only 1.4% of the total dynamic range of the material (−905 to +108 HU). In clinical terms, this agreement indicates that PLA measurements in phantoms are virtually independent of the selected reconstruction kernel.

For PLA Stonefill, inter-kernel agreement was equally high at intermediate fillings (40–90%, $|\Delta| < 7$ HU), but divergent at the extremes: 7.94 HU at 10% and 38.45 HU at 100%. The divergence of 38.45 HU (8.7% of the average value) for Stonefill 100% indicates that the bone kernel (FC30) systematically underestimates the density of this material in solid configuration. Concomitantly, the standard deviation obtained with FC30 (74.6 HU) was 71.4% higher than that obtained with FC21 (43.5 HU), suggesting that the bone kernel — optimized for edge enhancement in biological bone with hydroxyapatite [20] — amplifies the textural heterogeneity of $CaCO_3$ in the polymer matrix.

The interpretation of these differences is relevant for clinical protocols: at low fillers, the geometric heterogeneity of the filler pattern generates air/polymer interfaces with a high contrast gradient, highlighted differently by the algorithms [16]. It is therefore recommended to use the soft kernel (FC21) for quantitative HU measurements in phantoms with PLA Stonefill 100%, avoiding negative bias and artificially high dispersion.

F. Implications for the Construction of Pediatric Cranial Phantoms

The combination of the two materials in selected fillings allows for the replication of pediatric cranial microarchitecture at 2 years of age with radiometric fidelity.

- a) Pediatric cortical bone (300–600 HU in the frontal-parietal region [6], [9]): PLA Stonefill 100% (310–441 HU as per kVp) — 73–100% agreement with target range.
- b) Trabecular bone/diploë (100–300 HU [13]): PLA 100% (+108 HU) or PLA Stonefill 90% (+149 HU) — both within the reference range.
- c) Periosteal/pericranial soft tissue (+10 to +40 HU [5]): PLA Stonefill 80% (+36.15 HU) — deviation < 4 HU from the midpoint.

- d) Subdural space/CSF (~0 to +15 HU [13]): PLA Stonefill 80% can be calibrated via fine fill adjustment (78–82%).

According to the linear model described by Eq. (1), the slope coefficient quantifies the intrinsic calibration of each material: a ≈ 10 HU/% for PLA and a ≈ 14 HU/% for Stonefill. This coefficient allows for the accurate prediction of the HU value for any intermediate infill percentage, giving the system fine-tuning capability in 1% increments. This predictive resolution constitutes a significant advantage over commercial phantoms with fixed densities [15], [16]. Huang et al. [3] demonstrated that diploë is inconsistent in children under 3 years of age; therefore, for phantoms representing the 2-year-old calvaria, a 100% Stonefill monolayer configuration with a thickness of 3.7–3.9 mm (according to Garfin et al. [4]) is more reliable than the adult trilaminar configuration.

G. Limitations and Perspectives

The cylindrical geometry of the specimens does not replicate the pediatric cranial curvature, which may influence the beam-material interaction in clinical protocols with angulation. Comparison with reference data from the literature is limited by the interindividual variability of pediatric cranial mineralization (CV > 100% according to Liu et al. [6]), by the absence of age-stratified HU tables specifically for 2 years [9], and by the dependence on a single CT scanner, which prevents assessment of inter-scanner variability — known to be relevant in cranial densitometry [21]. Future studies should: (1) manufacture phantoms with anatomical cranial geometry derived from pediatric statistical models [14]; (2) validate on multiple low-dose clinical instruments and protocols; (3) evaluate mechanical properties (Young's modulus, Vickers hardness) to verify biomechanical equivalence in addition to radiological equivalence; and (4) investigate the temporal stability of the radiological properties of PLA Stonefill under prolonged storage conditions.

V. Conclusion

This study aimed to demonstrate the feasibility of FDM 3D printing with PLA and PLA Stonefill filament for fabricating phantoms simulating biological tissues across a wide range of Hounsfield Units (HU), with specific application to simulating pediatric cranial bone at 2 years of age. The main findings were:

- 1) PLA Stonefill at 100% infill (310–441 HU, depending on kVp) simulates pediatric cranial cortical bone, falling within the 300–600 HU range reported for 2-year-old children.
- 2) PLA Stonefill at 80% fill (+36.15 HU) shows precise equivalence with muscle/soft tissue (+10 to +40 HU).
- 3) Conventional PLA at 90% infill (−87.89 HU) simulates adipose tissue (−50 to −100 HU), and at 100% (+108.22 HU) it positions itself as dense soft tissue/trabecular bone.
- 4) Energy dependence: PLA Stonefill (130 HU between 80–120 kVp) mimics the spectral behavior of biological bone. Conventional PLA is almost energetically independent (< 9 HU).
- 5) The agreement between reconstruction kernels (soft vs. bone) was high for both materials, with differences of less than 15 HU in most fillings.

The combination of the two materials in distinct fillings provides a quantitative basis for the construction of low-cost, multicompartmental pediatric phantoms, applicable to dosimetry and quality assurance in pediatric cranial CT.

Acknowledgment

The authors thank Fundação Araucária for the support.

References

[1] S. Idriz et al., "CT of normal developmental and variant anatomy of the pediatric skull: distinguishing trauma from normality," Radiographics, vol. 35, no. 5, pp. 1585–1601, 2015. [doi: 10.1148/rg.2015140177]

[2] D. E. Carver, S. D. Kost, N. D. Fraser, W. P. Segars, D. R. Pickens, R. R. Price, and M. G. Stabin, "Realistic phantoms to characterize dosimetry in pediatric CT," Pediatr. Radiol., vol. 47, no. 6, pp. 691–700, 2017. [doi: 10.1007/s00247-017-3805-1]

[3] A. H. Huang, H. H. Sun, G. B. Skolnick, and A. S. Woo, "Thickness of calvarium and diploic space in children ages 0 to 17 as assessed by computed tomography," Eur. J. Plast. Surg., vol. 38, no. 3, pp. 193–198, 2015. [doi: 10.1007/s00238-014-1060-3]

[4] S. R. Garfin, R. Roux, M. J. Botte, R. Centeno, and S. L.-Y. Woo, "Skull osteology as it affects halo pin placement in children," J. Pediatr. Orthop., vol. 6, no. 4, pp. 434–436, 1986. [doi: 10.1097/01241398-198607000-00008]

[5] P. Domenech-Fernandez, J. Yamane, J. Domenech, C. Barrios, F. Soldado-Carrera, J. Knorr, and F. Canavese, "Analysis of skull bone thickness during growth: an anatomical guide for safe pin placement in halo fixation," Eur. Spine J., vol. 30, no. 2, pp. 410–415, 2021. [doi: 10.1007/s00586-020-06367-x]

[6] J. Liu, C. Elkhill, S. LeBeau, B. French, N. Lepore, M. G. Linguraru, and A. R. Porras, "Data-driven normative reference of pediatric cranial bone development," Plast. Reconstr. Surg. Glob. Open, vol. 10, no. 8, p. e4457, 2022. [doi: 10.1097/GOX.0000000000004457]

[7] N. Gajawelli, S. Deoni, J. Shi, M. G. Linguraru, A. R. Porras, M. D. Nelson, B. Tamrazi, V. Rajagopalan, Y. Wang, and N. Lepore, "Neurocranium thickness mapping in early childhood," Sci. Rep., vol. 10, art. 16651, 2020. [doi: 10.1038/s41598-020-73589-w]

[8] H. Delye, T. Clijmans, M. Y. Mommaerts, J. Vander Sloten, and J. Goffin, "Creating a normative database of age-specific 3D geometrical data, bone density, and bone thickness of the developing skull: a pilot study," J. Neurosurg. Pediatr., vol. 16, no. 6, pp. 687–702, 2015. [doi: 10.3171/2015.4.PEDS1493]

[9] K. Yamada, S. Endo, S. Yoshioka, J. Hatazawa, H. Yamaura, and T. Matsuzawa, "Age-related changes of the cranial bone mineral: a quantitative study with computed tomography," J. Am. Geriatr. Soc., vol. 30, no. 12, pp. 756–763, 1982. [doi: 10.1111/j.1532-5415.1982.tb03366.x]

[10] K. Takahashi, Y. Morita, S. Ohshima, S. Izumi, Y. Kubota, and A. Horii, "Bone density development of the temporal bone assessed by computed tomography," Otol. Neurotol., vol. 38, no. 10, pp. 1445–1449, 2017. [doi: 10.1097/MAO.0000000000001566]

[11] S. K. Rowbotham, C. G. Mole, D. Tieppo, M. Blaszkowska, S. Cordner, and S. Blau, "Reference measurements for average human neurocranial bone density to inform head trauma interpretations," Aust. J. Forensic Sci., vol. 56, no. 2, pp. 124–143, 2024. [doi: 10.1080/00450618.2022.2141320]

[12] X. Ma, M. Figl, E. Unger, M. Buschmann, and P. Homolka, "X-ray attenuation of bone, soft and adipose tissue in CT from 70 to 140 kV and comparison with 3D printable additive manufacturing materials," Sci. Rep., vol. 12, art. 14580, 2022. [doi: 10.1038/s41598-022-18741-4]

[13] U. Schneider, E. Pedroni, and A. Lomax, "The calibration of CT Hounsfield units for radiotherapy treatment planning," Phys. Med. Biol., vol. 41, no. 1, pp. 111–124, 1996. [doi: 10.1088/0031-9155/41/1/009]

[14] J. Liu, J. Chaij, M. G. Linguraru, B. French, R. Keating, A. L. Alexander, and A. R. Porras, "Cranial bone thickness and density anomalies quantified from CT images can identify chronic increased intracranial pressure," Neuroradiology, vol. 66, no. 10, pp. 1817–1828, 2024. [doi: 10.1007/s00234-024-03393-0]

[15] A. M. M. Ahmed, M. Buschmann, et al., "Tailoring the mass density of 3D printing materials for accurate X-ray imaging simulation by controlled underfilling for radiographic phantoms," Polymers, vol. 16, no. 8, p. 1116, 2024. [doi: 10.3390/polym16081116]

[16] S. Hatamikia et al., "3D printed patient-specific thorax phantom with realistic heterogeneous bone radiopacity using filament printer

technology," Z. Med. Phys., vol. 32, no. 4, pp. 438–452, 2022. [doi: 10.1016/j.zemedi.2022.02.001]

[17] J. Arnold et al., "3D printed bismuth oxide-polylactic acid composites for radio-mimetic computed tomography spine phantoms," J. Biomed. Mater. Res. B Appl. Biomater., vol. 109, no. 6, pp. 789–796, 2021. [doi: 10.1002/jbm.b.34744]

[18] N. Okkalidis, K. Bliznakova, and N. Kolev, "A filament 3D printing approach for CT-compatible bone tissues replication," Phys. Med., vol. 102, pp. 96–102, 2022. [doi: 10.1016/j.ejmp.2022.09.009]

[19] K. Mei et al., "Design and fabrication of 3D-printed patient-specific soft tissue and bone phantoms for CT imaging," Sci. Rep., vol. 13, art. 17495, 2023. [doi: 10.1038/s41598-023-44602-9]

[20] M. Van der Walt, T. Crabtree, and C. Albantow, "PLA as a suitable 3D printing thermoplastic for use in external beam radiotherapy," Australas. Phys. Eng. Sci. Med., vol. 42, no. 4, pp. 1165–1176, 2019. [doi: 10.1007/s13246-019-00818-6]

[21] C. Cavaliere et al., "A customized anthropomorphic 3D-printed phantom to reproducibility assessment in computed tomography: an oncological case study," Front. Oncol., vol. 13, p. 1123796, 2023. [doi: 10.3389/fonc.2023.1123796]

[22] M. Lai et al., "Development of a CT-compatible, anthropomorphic skull and brain phantom for neurosurgical planning, training, and simulation," Bioengineering, vol. 9, no. 10, p. 537, 2022. [doi: 10.3390/bioengineering9100537]

[23] Earth Renewable Technologies (ERT), "FC 10030 Technical Data Sheet," PRO_TDS_15, Rev. 0, Jun. 2022, Curitiba, PR, Brazil.

[24] Earth Renewable Technologies (ERT), "FC 50325 Product Data Sheet," PRO_FDS_62, Rev. 04, Feb. 2026, Curitiba, PR, Brazil. https://ertbio.com/wp-content/uploads/2026/03/PRO_TDS_62-FC-50325_v4-PT.pdf

[25] J. P. O. Bustillo et al., "3D printed heterogeneous paediatric head and adult thorax phantoms for linear accelerator radiotherapy quality assurance: from fabrication to treatment delivery," Biomed. Phys. Eng. Express, vol. 10, no. 5, p. 055037, 2024. [doi: 10.1088/2057-1976/ad6f13]

[26] T. Kamomae et al., "Three-dimensional printer-generated patient-specific phantom for artificial in vivo dosimetry in radiotherapy quality assurance," Phys. Med., vol. 44, pp. 205–211, 2017. [doi: 10.1016/j.ejmp.2017.10.005]

[27] S. Y. Kim et al., "Fabrication of 3D printed head phantom using plaster mixed with polylactic acid powder for patient-specific QA in intensity-modulated radiotherapy," Sci. Rep., vol. 12, art. 17500, 2022. [doi: 10.1038/s41598-022-22520-6]

Adaptation of a Public Dental Radiology Service to Anvisa RDC No. 611/2022: Experience Report at LABIM/UFPR

1st Cristiano Miguel
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
cristiano.miguel@ufpr.br

2nd Stella Siqueira Campos
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
stella.campos@ufpr.br

3rd Katia Regina Heiden
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
katiaheiden@ufpr.br

4th José Ângelo Guerra
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
angeloguerra@ufpr.br

5th Lúcio Flavo Martins Tosta
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
luciotosta@ufpr.br

6th Gracielly Soares Nunes
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
gracielly.nunes@ufpr.br

7th Valdeci Tomas
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
valdeci.tomas@ufpr.br

8th Eduardo da Silva Garcia
Clinical Hospital Complex
Federal University of Paraná
Curitiba, Brazil
edusg36@hotmail.com

9th Mauren Abreu de Souza
Graduate Program in Health Technology
Pontifical Catholic University of Paraná
Curitiba, Brazil
mauren.souza@pucpr.br

10th José Vinícius Bolognesi Maciel
Department of Stomatology
Federal University of Paraná
Curitiba, Brazil
viniciusmaciel@ufpr.br

***Abstract*—The publication of Anvisa Collegiate Board Resolution No. 611/2022 (RDC No. 611/2022) introduced new regulatory requirements for the arrangement, operation, and quality control of diagnostic and interventional radiology services. This study aimed to describe the process of adapting a public dental radiology service to RDC No. 611/2022 and to critically discuss regulatory issues related to dental practice. This descriptive case study used a mixed qualitative-quantitative approach and involved document analysis of technical reports, dosimetry records, quality control tests, radiometric surveys, and the institution's radiation protection plan (RPP). Nine imaging units were assessed: six intraoral units, two extraoral units, and one cone-beam computed tomography scanner, along with ten lead protective garments. Radiometric surveys, protective garments, and eight of the nine units complied with the applicable requirements. One intraoral unit failed the tube voltage accuracy and exposure time accuracy tests and was referred for corrective maintenance. Occupational dosimetry readings remained below the service's monthly recording level. This finding was consistent with the radiometric surveys and quality control tests, although dosimetry alone cannot establish a causal relationship with the effectiveness of the protection measures in place. The service team prepared the RPP and descriptive report at no additional cost to the institution. Overall, the service had largely met the requirements of RDC No. 611/2022; however, the periodic training program and the installation of red warning lights above the examination room doors were still pending at the time of assessment. Document organization, technical follow-up, and staff engagement facilitated progress toward compliance with RDC No. 611/2022. The analysis also highlighted potential regulatory tensions between Brazilian requirements and recent international recommendations on the use of lead protective garments in dental imaging, particularly in examinations where shielding may interfere with the acquisition field or automatic exposure control (AEC).**

***Index Terms*—Quality control, public health management, radiation protection, dental radiology, RDC No. 611/2022.**

I. Introduction

According to the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), medical exposures are the main artificial source of human exposure to ionizing radiation. Approximately 1.1 billion dental radiology examinations are performed worldwide each year, representing a 130% increase compared with the previous report, published in 2010. Dental radiology accounts for 26.3% of human exposures to ionizing radiation but only 0.2% of the total collective dose [1].

In Brazil, SVS/MS Ordinance No. 453/1998 was replaced by RDC No. 611/2022 in 2022. This regulatory update was needed to modernize radiation protection requirements and keep pace with technological advances, particularly the expansion of digital image acquisition and processing systems. Ordinance No. 453/1998 had limitations regarding newer technologies and included requirements and equipment assessment parameters that could overlap with other standards, creating uncertainty about which regulation applied.

In addition to updating regulatory requirements for radiology services, RDC No. 611/2022 aligned its provisions with specific technical and regulatory standards, including guide-

lines issued by the Brazilian Nuclear Energy Commission (CNEN) [3], [5].

The dental radiology service evaluated in this study comprises two settings: the Dental Radiology Service and the Imaging Teaching and Research Laboratory (LABIM). The Dental Radiology Service has one room equipped with an ORTHOCEPH extraoral X-ray unit and one cubicle containing an XDENT X70 intraoral unit. LABIM has five cubicles with five intraoral/periapical units—three DABI ATLANTE SPECTRO 70 X units and two XDENT X70 units—one room with an ORTHOCEPH extraoral unit, and one examination room with an ICAT cone-beam computed tomography scanner. Extraoral radiographs are performed in the Dental Radiology Service. At LABIM, cone-beam computed tomography examinations and extraoral radiographs are performed, and practical intraoral radiology classes are provided to undergraduate dentistry students. In 2025, the service performed 5,708 extraoral radiographs and 1,345 cone-beam computed tomography examinations.

This study aimed to describe the process of adapting a public dental radiology service to RDC No. 611/2022 and to critically evaluate selected regulatory requirements.

II. Methodology

A descriptive case study with a mixed qualitative-quantitative approach was conducted through document analysis of technical reports, institutional records, occupational dosimetry reports, quality control tests, radiometric surveys, and the radiation protection plan of a public dental radiology service. The analysis compared the service records with the requirements established by RDC No. 611/2022 and by the Normative Instructions applicable to extraoral and intraoral dental radiology. The analysis included the assessment results for six intraoral X-ray units, two extraoral X-ray units, and one cone-beam computed tomography scanner. Ten lead aprons were also evaluated according to the performance and integrity criteria established by RDC No. 611/2022 and the applicable Normative Instructions. For the intraoral units, the following parameters were assessed: tube voltage indicator accuracy; exposure time indicator accuracy; half-value layer; air kerma rate linearity; air kerma rate reproducibility; field size; focus-to-skin distance; representative dose value; image artifacts; image quality; darkroom sealing; effectiveness of the erasure cycle in digital processing; room radiometric survey; and leakage radiation. For the extraoral X-ray units and the cone-beam computed tomography scanner, the following parameters were assessed: accuracy and reproducibility of the tube voltage indicator; accuracy and reproducibility of the exposure time indicator; half-value layer; air kerma rate reproducibility; field size; representative dose value; dose indicator accuracy; presence of image artifacts; indicative computed tomography (CT) number value; image uniformity in cone-beam computed tomography; noise and image uniformity in direct or computed radiography; effectiveness of the erasure cycle; and radiometric surveys of the rooms.

III. Results

The radiation protection plan (RPP) includes the professional team, work instructions, team training, signage system, warning signs and controls, area monitoring program, description of protective garments, description of the record system, standard operating procedures, quality assurance program, and accident procedures. The document contains seven appendices: the architectural design of the facility; quality control test and radiometric survey reports; the list of staff members and personal dosimetry reports; work procedures and examination techniques; archived information on periodic training; investigations of events related to occupational risk; and records of changes to the physical structure and equipment.

The examination room doors are fitted with lead-glass viewing windows for visual monitoring of the patient during examinations. They are identified with the international radiation symbol and the warning signs recommended by RDC No. 611/2022 and RDC No. 50/2002 [3], [4].

Analysis of the monthly occupational dosimetry reports showed that readings remained below 0.2 mSv per month, corresponding to the recording level adopted in the reports evaluated, with no readings above this level during the period analyzed [3], [5].

The radiometric surveys of all examination rooms were considered compliant with the regulation.

The assessment results for the extraoral X-ray units and the cone-beam computed tomography scanner were considered compliant with the applicable regulatory requirements.

Among the six intraoral units, only one XDENT X70 unit was found to be noncompliant in the tube voltage accuracy and exposure time accuracy tests. The unit underwent corrective maintenance and will be reevaluated.

All ten lead aprons were considered compliant with the regulation in terms of integrity and the quantity available in each examination room.

The periodic training program was still being implemented, with activities scheduled for the following two months.

Table I summarizes the degree of compliance with the requirements evaluated.

IV. Discussion

Although dental radiology examinations involve low individual doses compared with other radiological modalities, the high frequency of these procedures supports the continued use of quality assurance programs, dose control, protocol optimization, and periodic staff training. In this context, the principles of justification and optimization remain essential to reduce unnecessary exposures and avoid repeat examinations [6]–[8].

Ordinance No. 453/1998 needed to be replaced because it had become outdated in relation to newer technologies and overlapped with other standards, including Regulatory Standard No. 32 of the Brazilian Ministry of Labour and Employment (NR 32), CNEN Standard NN 3.01, which serves as the national reference for occupational dose limits, and some state and municipal laws. In the state of Paraná, State Law No.

TABLE I
SUMMARY OF COMPLIANCE WITH THE REQUIREMENTS EVALUATED

Requirement	Observed status	Status
RPP and documentation	Elements and records described in the RPP available	Met
Radiometric surveys	All rooms evaluated were compliant	Met
Extraoral equipment and CBCT	Test results compliant	Met
Intraoral equipment	Five of six compliant; one under maintenance and pending reevaluation	Partially met
Lead protective garments	Integrity and quantity considered compliant	Met
Occupational monitoring	No readings above the recording level in the reports evaluated	Met
Periodic training	Program under implementation	Pending
Red warning light	Not installed above the examination room doors	Pending

19,125/2017 also requires the availability and use of thyroid collars and lead shielding barriers to protect radiosensitive organs such as the gonads, thyroid, and lens of the eye, without explicitly accounting for examination-specific technical factors [2], [5], [9], [10].

Under RDC No. 611/2022, companions must use personal protective equipment when their presence in the examination room is indispensable. During radiological examinations, the gonads, lens, and thyroid must be protected when these organs are directly exposed to the primary radiation beam or located within 5 cm of it, provided that such protection does not obscure or degrade relevant diagnostic information or increase patient dose [3].

In contrast, recent international recommendations have questioned the routine use of patient shielding in dental imaging. The 2024 ADA recommendations no longer advocate the routine use of thyroid collars or lead aprons during dental radiographic examinations because shielding may interfere with the primary beam, obscure relevant anatomy, generate artifacts, or lead to repeat exposures [13].

Similar concerns have been raised for imaging systems using automatic exposure control (AEC), because highly attenuating material within the acquisition field may interfere with exposure regulation. However, the present study did not include experimental measurements of dose or image quality; therefore, this issue should be interpreted as a regulatory and technical concern rather than as an effect demonstrated in the service evaluated. [11]–[13].

The issue is presented here as a regulatory comparison between national requirements and international recommendations, not as an effect verified through dose or image-quality measurements obtained in the service evaluated in this study.

The situation is further complicated by the requirement in RDC No. 611/2022 to post a notice instructing patients to request protective garments. This requirement may call for additional explanations from professionals and complicate communication with patients, since technical guidance may appear to conflict with the institutional notice required by regulation. Standardized notices that do not distinguish among imaging modalities and technologies may therefore make patient communication more difficult and increase operational uncertainty for the team. The examination room doors display the warnings required by RDC No. 611/2022 and have lead-glass viewing windows that allow observation of the patient during examinations. At the time of the assessment, however, the red warning lights above the examination room doors had not yet been installed. They remain a pending item in the adaptation process [3], [4].

Another gap identified in RDC No. 611/2022 is the absence of a definition of the professional categories qualified to assume technical responsibility and radiation protection supervision, unlike Ordinance No. 453/1998. By using the term "legally qualified professionals," RDC No. 611/2022 leaves this definition to professional councils and local health surveillance authorities. This may create administrative and legal uncertainty, as well as professional disputes over which specialties or professionals have the legal authority to sign reports and assume responsibility for the service [3].

RDC No. 611/2022 is also complemented by Normative Instructions that specify technical requirements and the quality control tests applicable to each imaging technology. In dental radiology, operating routines must follow Normative Instruction No. 94/2021 (IN 94) for extraoral equipment parameters and Normative Instruction No. 95/2021 (IN 95), which establishes the minimum constancy tests for intraoral dental radiography equipment [3], [14], [15].

Regarding quality assurance, the institutional assessment showed compliance after the institution hired an external specialized company in 2025. This enabled the issuance of radiometric survey reports for all physical barriers, quality control test reports, and leakage radiation measurements for all X-ray units. Although technical outsourcing is permitted under the current regulation, it introduces costs that radiology services may find difficult to absorb. Dosimeter readings remained below 0.2 mSv per month throughout the period analyzed, corresponding to the recording level adopted in the evaluated reports. Radiometric surveys demonstrated compliance with the shielding barriers, while periodic quality control tests assessed equipment performance [3], [5].

The dental radiology staff prepared the descriptive report and RPP at no cost to the institution. Because these documents were developed specifically for the service rather than using generic commercially prepared templates, they provided detailed information on the infrastructure, waste management, and dosimetry control.

V. CONCLUSION

The service's experience showed that systematic document organization, staff involvement, technical monitoring, and quality control contributed to progress toward compliance with

RDC No. 611/2022. Full adaptation had not yet been achieved during the study period: one intraoral unit was noncompliant and required corrective action, while the periodic training program and red warning lights remained pending. Because this was a case study of a single service with its own institutional characteristics, the findings should not be generalized to other institutions. The analysis also highlighted regulatory issues related to the use of lead protective garments in AEC equipment; their effects on dose, image quality, and repeat examinations require additional data and specific investigation. Regulatory guidance tailored more closely to dental radiology may help reduce operational and legal uncertainty while preserving quality of care and continuity of services.

REFERENCES

[1] United Nations Scientific Committee on the Effects of Atomic Radiation, *Sources and Effects of Ionizing Radiation: UNSCEAR 2021 Report to the General Assembly, with Scientific Annexes*. New York, NY, USA: United Nations, 2021.

[2] Brazil, Ministry of Health, Health Surveillance Secretariat, "Ordinance No. 453, of June 1, 1998: Radiation protection guidelines in medical and dental radiodiagnosis," *Diario Oficial da Uniao*, Brasilia, DF, Brazil, 1998.

[3] Brazilian Health Regulatory Agency, "Collegiate Board Resolution–RDC No. 611, of March 9, 2022: Provides for the sanitary requirements for the organization and operation of diagnostic or interventional radiology services," *Diario Oficial da Uniao*, Brasilia, DF, Brazil, 2022.

[4] Brazilian Health Regulatory Agency, "Collegiate Board Resolution–RDC No. 50, of February 21, 2002: Technical regulation for the planning, programming, design, and evaluation of physical projects for healthcare facilities," Brasilia, DF, Brazil: ANVISA, 2002.

[5] Brazilian Nuclear Energy Commission, "CNEN Resolution No. 306, of February 22, 2024: Approves standard CNEN NN 3.01–Basic radiation protection guidelines," Rio de Janeiro, RJ, Brazil: CNEN, 2024.

[6] International Commission on Radiological Protection, *The 2007 Recommendations of the International Commission on Radiological Protection*, ICRP Publication 103, *Annals of the ICRP*, vol. 37, nos. 2–4. Oxford, U.K.: Elsevier, 2007.

[7] P. Abbott, "Risk of radiation-induced cancer from dental radiography," *Australian Endodontic Journal*, vol. 26, no. 1, pp. 33–35, 2000.

[8] T. C. Oliveira and R. A. Medeiros, "Proteção radiológica na prática odontológica atual: Uma revisão de literatura sobre riscos estocásticos e aplicação do princípio ALARA," *Revista Brasileira de Odontologia*, vol. 78, Art. no. e1921, 2021.

[9] Paraná, Legislative Assembly of the State of Paraná, "State Law No. 19,125, of September 18, 2017: Provides for the mandatory posting of signs informing about the use of neck protectors in places where radiological, mammography, or tomography examinations are performed for thyroid cancer prevention," Curitiba, PR, Brazil: ALEP, 2017.

[10] Brazil, Ministry of Labor and Employment, "Regulatory Standard No. 32–Occupational safety and health in healthcare services," Brasilia, DF, Brazil: MTE, 2005.

[11] International Commission on Radiological Protection, *Radiological Protection in Fluoroscopically Guided Procedures Performed Outside the Imaging Department*, ICRP Publication 117, *Annals of the ICRP*, vol. 40, no. 6, 2011.

[12] International Atomic Energy Agency, *Radiation Protection and Safety in Medical Uses of Ionizing Radiation*, IAEA Safety Standards Series No. SSG-46. Vienna, Austria: IAEA, 2018.

[13] E. Benavides et al., "Optimizing radiation safety in dentistry: Clinical recommendations and regulatory considerations," J. Am. Dent. Assoc., vol. 155, no. 4, pp. 280–293.e4, Apr. 2024, doi: 10.1016/j.adaj.2023.12.002.

[14] Brazilian Health Regulatory Agency, "Normative Instruction–IN No. 94, of May 27, 2021: Establishes sanitary requirements for quality assurance and safety in extraoral dental radiology systems," *Diario Oficial da Uniao*, Brasilia, DF, Brazil, 2021.

[15] Brazilian Health Regulatory Agency, "Normative Instruction–IN No. 95, of May 27, 2021: Establishes sanitary requirements for quality assurance and safety in intraoral dental radiology systems," *Diario Oficial da União*, Brasilia, DF, Brazil, 2021.

Analysis of Sunglass Frame Light Blocking Assessment Methods

1st Pedro da Silva
São Carlos School of Engineering
University of São Paulo
São Carlos, Brasil
pedrotxs@usp.br

2nd Augusto de Andrade
São Carlos School of Engineering
University of São Paulo
São Carlos, Brasil
augusto.367@usp.br

3rd Pedro Guedes
São Carlos School of Engineering
University of São Paulo
São Carlos, Brasil
pmlguedes@usp.br

4th Liliane Ventura
São Carlos School of Engineering
University of São Paulo
São Carlos, Brasil
lilianeventura@usp.br

***Abstract*—Ultraviolet radiation (UVR) is a significant factor correlated with various ocular pathologies, making the geometry of sunglass frames a essential element for peripheral blocking. This study presents a comparative analysis between two techniques designed to assess the efficacy of sunglass frames in UVR: an experimental measurement system assembled from a polystyrene sphere internally irradiated by a UVA source from 385 to 395 nm, with a tolerance of plus or minus 5 nm, over a mannequin headform, and a computational image processing system based on mathematical modeling of the visual field of view. The methodologies were applied to five distinct sunglass samples, each characterized by a different frame geometry. The results indicate a consistent qualitative agreement between the two systems in ranking and identifying the relative blocking capability across the samples. While differences in absolute values were observed, they can be attributed to the physical sensor's 80-degree acceptance angle compared to the 100-degree temporal angle considered in the computational model. Furthermore, both methodologies provide a preliminary comparative approach within controlled environments to support future measurement procedures.**

***Index Terms*—eye protection, scattered light, sunglasses, sunglass frames.**

I. Introduction

Studies indicate that ultraviolet radiation (UVR) is a significant factor correlated with ocular pathologies affecting the cornea, lens, and retina [1]–[5]. While the use of sunglasses is fundamental, their protective efficacy is not solely defined by the UV transmittance of the lenses.

The vast diversity of protective eyewear models on the market highlights the necessity of considering both frame geometry and material to characterize the contribution to ocular radiation blocking. This consideration may be essential because peripheral radiation incidence becomes highly significant in environments with high levels of diffuse and reflected UVR, as demonstrated by Parisi *et al.* [6], Sliney [7], and Rosenthal *et al.* [8].

Supported by FAPESP and CAPES

Based on extensive literature in this field, the guidelines published by the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [9] establish safe UVR limits for an 8-hour exposure period. These guidelines specify that the maximum radiant exposure—encompassing the UV-C (100 nm–280 nm), UV-B (280 nm–315 nm), and UV-A (315 nm–400 nm) bands weighted against the spectral efficacy curve—must not exceed 30 J/m^2.

The work published by Backes *et al.* [10] simulated solar exposure environments to isolate the fractions of direct, diffuse, and reflected radiation incident on the eyes of a computer-modeled human head. Their study utilized three sunglass models with distinct geometries to quantify the UVR incident on the ocular and periorbital regions. Furthermore, these evaluations were categorized according to the season, time of day, and head tilt of the user.

The primary objective of this study is to provide a comparative analysis of two techniques currently under development at the Ophthalmic Instrumentation Laboratory (LIO-EESC/USP). Although independently introduced in preliminary works [11], [12], both methodologies are applied in this work to a new set of sunglass samples to assess their efficacy in blocking UVR.

The first approach is experimental [11], utilizing a physical setup composed of UV-A LEDs as the radiation source, paired with a photodiode sensor sensitive to the UV-A spectrum. In contrast, the second strategy is computational [12], employing digital image processing techniques alongside mathematical modeling to map the optical paths of radiation reaching the human eye. Although currently under development, both protocols aim to establish a standardized measurement procedure within a controlled environment to ensure consistent testing conditions in future work.

II. Methodology

To compare the methods, both the experimental and computational techniques were applied to evaluate the same group of sunglass frames. This section outlines the implementation

of both protocols [11], [12]. The following subsections present the physical development of the experimental setup, followed by the algorithmic steps of the computational image-based model.

A. Measurement System

To achieve the objective of measuring scattered UVR within a controlled diffuse UV-A exposure environment, a set of optical instruments was selected to develop an electronic apparatus capable of conducting measurement tests on sunglass frames of various geometries.

A polystyrene sphere (Fig.1) was employed to establish a controlled UVR exposure environment, containing a headform with its eye positions centered within the sphere. An array of four high-power LEDs, emitting at 385 ± 5 nm and 395 ± 5 nm, was positioned behind the headform at a zenith angle of $60°$. Photodiode sensors were integrated at the eye locations of the headform to measure the incident irradiance, featuring an $80°$ acceptance half-angle, covering from $-80°$ to $+80°$ (Fig. 2). The laboratory-developed apparatus is illustrated in Fig.3.

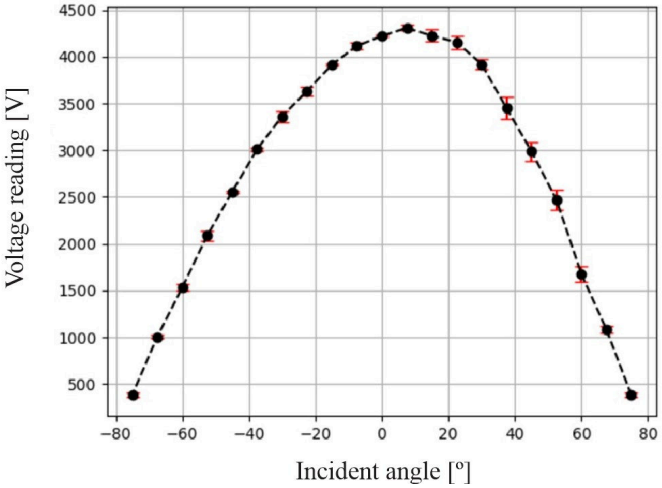


Fig. 2. Sensor incidence angle response.

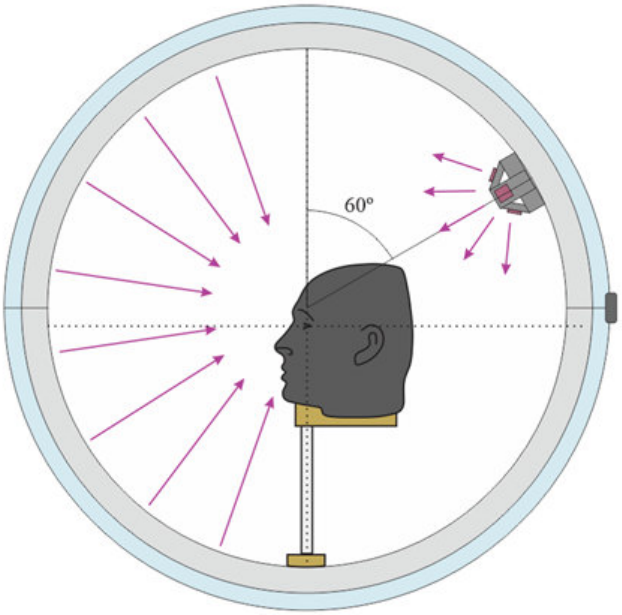


Fig. 1. Prototype of the spherical environment used to create a controlled diffuse UV-A source.



Fig. 3. Developed physical prototype of the measurement system.

Furthermore, an embedded system was developed based on an ESP8266 microcontroller to integrate the instrumentation and communicate with a custom graphical user interface (GUI). This software application ensures ease of use and efficiency during testing, displaying the processed results from the irradiance readings.

The UVR blocking measurement (B_{m}) is automatically calculated through auxiliary calibration steps that quantify the irradiance detected by the sensor under minimum and maximum exposure baselines, denoted as E_{off} (sources turned off) and E_{on} (sources turned on without the sunglasses), respectively. The irradiance incident through the gaps of the frame (E_{frame}) is measured by positioning the sunglasses on the headform. The blocking percentage is then computed according to (1). It is important to highlight that an opaque material was applied to the inner surface of the lenses to eliminate any transmitted or reflected optical contributions from the lenses, thereby isolating the geometric blocking capacity of the frame.

$$B_{\text{m}} = 100 \cdot \left(1 - \frac{E_{\text{frame}} - E_{\text{off}}}{E_{\text{on}} - E_{\text{off}}}\right) [\%] \tag{1}$$

B. Computational Image-Based Method

This computational method for estimating the percentage of light blocking caused by sunglass frames integrates three primary components: a trigonometric projection model, an image acquisition system, and a digital processing pipeline. The objective of this procedure is to propose a standardized metric capable of quantifying the light blocking capacity provided by the frame's geometry across the user's visual field of view (FOV).

At the core of this methodology is mathematical modeling, which is essential for accurately delimiting the regions of interest (ROIs) on the image plane corresponding to the physiological limits of the human FOV. This critical step ensures standardization when evaluating luminous obstruction from both frontal and lateral perspectives. The specific ROIs are derived by performing trigonometric projections of the physiological FOV boundary angles, specifically $60°$ superior, $75°$ inferior, and $100°$ temporal [13], onto the two-dimensional (2D) view plane, accounting for the physical distances between the sunglasses and the ocular surface of the headform.

The image acquisition setup was mounted on an optical bench. Images were captured using a Bright 2k WC577 camera with a resolution of 2560×1440 pixels, placed at a fixed distance of 40 cm from a yellow headform (Fig.4) to maximize contrast against the sunglass frames. Furthermore, a white LED ring light was employed to ensure uniform illumination throughout the experiments.

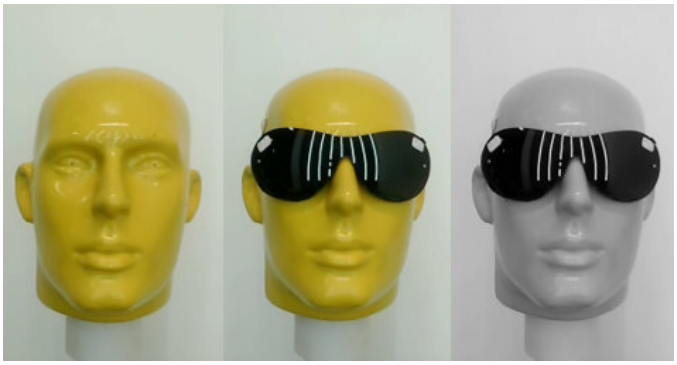


Fig. 4. Headform images from frontal perspective showing baseline setup and sample placement.

The captured images are processed through a custom Python-based software pipeline. The workflow begins with system calibration, where a feature of known physical dimensions on the headform is measured in pixels to establish a spatial scale (px/mm). Using this scale, the software applies the mathematical projection model to automatically segment the ROI. The segmented ROI is then converted into a binary image using color-based thresholding to isolate the frame (represented by white pixels) from the background (black pixels). Subsequently, morphological operations (Fig.5) are applied to reduce noise, such as shadows and unwanted reflections.

Finally, the percentage of light blockage obtained via image processing (B_{i}) is quantified by calculating the ratio of the blocked pixel area (comprising the frame and lens region, considering 100% light blockage, equivalent to the opacified lenses in the physical method) to the total pixel area of the ROI. Additionally, the results analyzed from the frontal and lateral viewpoints are assigned distinct weighting factors. This proportional allocation is determined by the percentage of the total visual field that each perspective encompasses for each eye, establishing the frontal contribution at 81% and the lateral contribution at 19%.

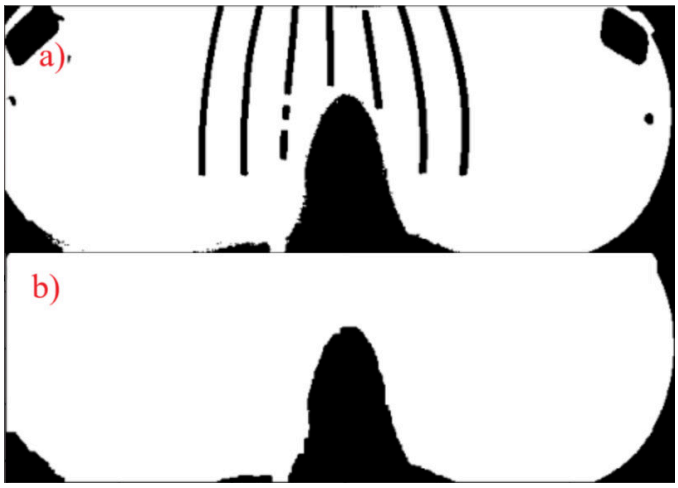


Fig. 5. Morphological filter application: (a) unfiltered image and (b) filtered image.

III. Results

The methodologies were applied to five distinct sunglass samples, each characterized by different frame geometries, as depicted in Fig. 6. These samples represent a new dataset evaluated specifically for this comparative study, distinct from the preliminary sets reported in [11] and [12].

The experimental results obtained from the measuring system are presented in Table I. The data includes the blockage percentages for the right eye, left eye, and the mean value for each sample.

TABLE I
EYE BLOCKAGE RESULTS OBTAINED BY THE MEASURING SYSTEM.

Sample	$B_m[\%]$		
	Right Eye	Left Eye	Mean
I	91.89	92.91	92.40
II	92.69	94.80	93.74
III	89.25	89.75	89.50
IV	93.82	95.02	94.42
V	91.24	91.59	91.41

Table II presents the results obtained through the computational image-based method. This technique calculates the radiation blockage within the projected physiological FOV for the right eye, the left eye, and the mean value for each sample.

The mean blockage percentages obtained from both the physical measurement system and the computational image-based method are summarized in Table III. The data include the absolute difference between the two approaches.



Fig. 6. The five distinct sunglass samples evaluated in the study, showcasing diverse frame geometries.

TABLE II
EYE BLOCKAGE RESULTS OBTAINED BY THE COMPUTATIONAL IMAGE-BASED METHOD

Sample	B_i (%)		
	Right Eye	Left Eye	Mean
I	86.74	87.81	87.27
II	87.26	89.87	88.56
III	82.76	83.08	82.92
IV	91.78	93.15	92.46
V	85.41	86.02	85.72

Both methodologies consistently identified Sample IV as the most effective and Sample III as the least effective in blocking radiation. Under the physical measurement system, Sample IV achieved a mean blockage of 94.42%, compared to 89.50% for Sample III. A similar trend was observed in the computational method, with Sample IV reaching 92.46% and Sample III obtaining 82.92%. Overall, the experimental measurement system consistently yielded higher blockage values than the computational image-based method across all evaluated samples.

TABLE III
COMPARISON OF THE MEAN BLOCKAGE RESULTS BETWEEN BOTH METHODS

Sample	B_m (%)	B_i (%)	Difference (%)
I	92.40	87.27	5.13
II	93.74	88.56	5.18
III	89.50	82.92	6.58
IV	94.42	92.46	1.96
V	91.41	85.72	5.69

IV. DISCUSSION

The comparative analysis between the two systems reveals a consistent agreement in ranking the efficacy of the samples, demonstrating that both methodologies successfully distinguish blocking differences among distinct frame geometries. Both methodologies identified Sample IV as providing the highest coverage and Sample III as the least effective. This qualitative agreement demonstrates the initial potential of the computational image-based method for evaluating frame geometry.

While the ranking trends are consistent, the quantitative results show that the physical measurement system consistently recorded higher blockage percentages than the computational method. This offset may be partly explained by the difference between the physical sensor's acceptance angle and the FOV limits modeled in the software. As detailed in Section II-A, the photodiode sensors possess an acceptance angle of $80°$. In contrast, the computational method is modeled to evaluate blockage up to a temporal limit of $100°$. Although both methodologies aim to estimate the total radiation percentage blocked by the frames, they approach it through different mechanisms. The physical system measures the percentage blockage of UV-A irradiance, while the computational method models the projected geometric occlusion of the visual field to approximate the actual measured blocking.

An additional point of agreement is observed in the lateral asymmetry of the results. In both Table I and Table II, the blockage values differ between the right and left eyes across all samples. This systematic difference suggests a slight asymmetry in the positioning of the sunglasses on the headform during the tests. The fact that both systems detected the same variation with a similar magnitude between the eyes indicates congruent results between the physical measurements and the computational model.

V. CONCLUSION

This study presented a comparative analysis between a physical measurement system and a computational image-based method, both developed to assess the efficacy of sunglass frames in blocking peripheral ultraviolet radiation.

While the computational method requires further parameter refinement to enhance absolute quantitative accuracy, it offers a non-destructive testing methodology that eliminates the need for physical sample preparation, thereby accelerating the evaluation throughput. In contrast, although the physical measurement system provides reliable direct irradiance readings, it requires temporary lens opacification to isolate the geometric blocking contribution of the frame and eliminate any transmitted or reflected optical contributions from the lenses.

Ultimately, although based on preliminary results, this study demonstrates that both methodologies under development show potential to establish standardized measurement procedures within controlled environments. By minimizing external variables, both protocols aim to provide consistent testing conditions across the evaluated samples, representing

complementary methodologies under development for sunglasses frame analysis.

ACKNOWLEDGMENT

The authors acknowledge the financial support from FAPESP (grant number 2024/22508-0) and the Coordination for the Improvement of Higher Education Personnel (CAPES) – Finance Code 001.

REFERENCES

[1] D. G. Pitts, "The ocular ultraviolet action spectrum and protection criteria," *Health Phys.*, vol. 25, no. 6, pp. 559–566, Dec. 1973, doi: 10.1097/00004032-197312000-00002.

[2] D. G. Pitts, "The ocular effects of ultraviolet radiation," *Am. J. Optom. Physiol. Opt.*, vol. 55, no. 1, pp. 19–35, Jan. 1978.

[3] J. P. G. Bergmanson and P. G. Söderberg, "The significance of ultraviolet radiation for eye diseases: A review with comments on the efficacy of UV-blocking contact lenses," *Ophthalmic Physiol. Opt.*, vol. 15, no. 2, pp. 83–91, Mar. 1995, doi: 10.1046/j.1475-1313.1995.9598237h.x.

[4] D. H. Sliney, "Ultraviolet radiation and the eye," in *Environmental UV Radiation: Impact on Ecosystems and Human Health and Predictive Models*, F. Ghetti, G. Checcucci, and J. F. Bornman, Eds. Berlin, Germany: Springer, 2006, pp. 259–278.

[5] M. Coroneo, "Ultraviolet radiation and the anterior eye," *Eye Contact Lens*, vol. 37, no. 4, pp. 214–224, Jul. 2011, doi: 10.1097/ICL.0b013e318223394e.

[6] A. V. Parisi, A. Green, and M. G. Kimlin, "Diffuse solar UV radiation and implications for preventing human eye damage," *Photochem. Photobiol.*, vol. 73, no. 2, pp. 135–139, Feb. 2001, doi: 10.1562/0031-8655.

[7] D. H. Sliney, "Photoprotection of the eye - UV radiation and sunglasses," *J. Photochem. Photobiol. B, Biol.*, vol. 64, no. 2–3, pp. 166–175, Nov. 2001, doi: 10.1016/S1011-1344(01)00229-9.

[8] F. S. Rosenthal, A. E. Bakalian, C. Q. Lou, and H. R. Taylor, "The effect of sunglasses on ocular exposure to ultraviolet radiation," *Am. J. Public Health*, vol. 78, no. 1, pp. 72–74, Jan. 1988, doi: 10.2105/AJPH.78.1.72.

[9] International Commission on Non-Ionizing Radiation Protection, "Guidelines on limits of exposure to ultraviolet radiation of wavelengths between 180 nm and 400 nm (incoherent optical radiation)," *Health Phys.*, vol. 87, no. 2, pp. 171–186, Aug. 2004, doi: 10.1097/00004032-200408000-00006.

[10] C. Backes *et al.*, "Sun exposure to the eyes: Predicted UV protection effectiveness of various sunglasses," *J. Expo. Sci. Environ. Epidemiol.*, vol. 29, no. 6, pp. 753–764, Nov. 2019, doi: 10.1038/s41370-018-0087-0.

[11] A. Perez de Andrade and L. Ventura, "Method for obtaining the relative proportions of ultraviolet radiation incident on the eyes as a function of different sunglasses frame geometries using electronic instruments," in *Proc. 19th Int. Joint Conf. Biomed. Eng. Syst. Technol. (BIODEVICES)*, SciTePress, 2026, pp. 147–151, doi: 10.5220/0014339600004070.

[12] P. T. X. da Silva and L. Ventura, "Visual field light blocking of sunglass frames: A quantitative imaging method," in *Proc. 19th Int. Joint Conf. Biomed. Eng. Syst. Technol. (BIODEVICES)*, SciTePress, 2026, pp. 123–129, doi: 10.5220/0014330600004070.

[13] R. H. Spector, "Visual fields," in *Clinical Methods: The History, Physical, and Laboratory Examinations*, 3rd ed., H. K. Walker, W. D. Hall, and J. W. Hurst, Eds. Boston, MA: Butterworths, 1990, ch. 116.

Assessment of the Acoustic Power Emitted by a Newly Manufactured Dermatological Ultrasound Device

Kacie L. R. M. Leme
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0009-0005-3447-1649

Thaís Rodrigues
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0009-0009-5598-934X

Sarah R. De Annunzio
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0000-0001-5477-2135

José F. S. Costa-Júnior
Postgraduate Program in Biomedical Engineering
Brazil University
São Paulo, Brazil
https://orcid.org/0000-0002-6526-2760

Abstract* — This study evaluated the acoustic power output of a dermatological ultrasound device operating at frequencies of 5.5 and 10.10 MHz. Three applicators with different effective radiating areas (ERAs) were analyzed using a calibrated radiation force balance with up-to-date preventive maintenance. Ten consecutive measurements were performed at each nominal power level to assess measurement accuracy and repeatability. Accuracy was evaluated using the mean relative error between the measured and reference values, whereas repeatability was quantified using the standard deviation and coefficient of variation. Overall, the acoustic power values measured using a radiation force balance showed good agreement with the corresponding reference values, with most results falling within the ±20% interval adopted as a comparative criterion with reference to IEC 61689:2022, although the evaluated frequencies are outside the scope of the standard. In general, measurement accuracy was lower at the one or two lowest nominal power settings**. *The applicators* **with a nominal ERA of 3 cm² showed high repeatability (CV < 2%), whereas the applicators with a nominal ERA of 1 cm² exhibited greater dispersion at lower power settings. At the lowest nominal power setting (0.3 W), the relative error was negative and exceeded 20% in magnitude, indicating that the measured acoustic power was higher than the nominal value and could therefore result in greater acoustic energy delivery for a given exposure time. These findings underscore the importance of periodic acoustic power assessment and performance monitoring to ensure the safety and reliability of dermatological ultrasound applications.

Keywords*—dermatological ultrasound, acoustic power, accuracy, precision, device performance

I. Introduction

The increasing incorporation of electromedical technologies into healthcare services has heightened the need for robust quality control, calibration, and performance assessment programs for medical devices [1]. The reliability of the physical parameters generated by these devices is fundamental to ensuring patient safety, treatment effectiveness, and the reproducibility of clinical and scientific results [2–4]. In this context, metrology applied to biomedical engineering plays an essential role by providing methods and criteria for verifying whether devices operate in accordance with the specifications established by manufacturers and international technical standards [5].

Among the devices used in clinical practice, therapeutic ultrasound (TUS) is an electrophysical modality frequently employed in physical therapy, including both rehabilitation and dermatofunctional applications [6]. Its physiological effects on biological tissues result from the propagation of acoustic waves and their interaction with the medium, producing thermal and non-thermal responses. However, the effects of TUS also depend on tissue properties, such as density, perfusion, and absorption coefficient, as well as treatment parameters, including frequency, operating mode, pulse duration, exposure time, acoustic intensity, and acoustic power [7,8].

When there is a discrepancy between the dose indicated by the equipment and the dose actually delivered, both the treatment outcome and patient safety may be compromised [9,10].

To promote measurement reliability and support the safe and effective use of therapeutic ultrasound, international bodies have established standards for the metrological evaluation of TUS equipment [11]. The IEC 61689 standard [12] stipulates that the power emitted by TUS equipment may vary by up to approximately ±20% relative to the values indicated by the device. Despite the existence of international standards defining calibration and verification procedures, studies have reported discrepancies between the programmed nominal power and the power actually emitted by equipment in clinical use [13,14].

One of the most commonly used methods used for measuring acoustic power is the radiation force balance, in which a conical target positioned in the path of the ultrasound beam. The radiation force exerted on the target is employed to determine the acoustic power output of the equipment. Measurement repeatability is also essential for obtaining consistent results and ensuring their reliability [15,16]. Thus, the accuracy of the acoustic power output of the equipment is critical to ensuring that the planned therapeutic dose is effectively delivered to the patient [14].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

In this context, the aim of this study was to evaluate the output power of a therapeutic ultrasound device operating at two nominal frequencies (5.5 and 10.10 MHz) by comparing the measured values with the corresponding nominal settings. The findings were also used to emphasize the importance of periodic performance assessment and proper maintenance as part of quality assurance for dermatological ultrasound (DUS) equipment.

II. Methods

A. Acoustic power measurement

A new ILIFT therapeutic ultrasound device (Ibramed, Amparo, Brazil) was evaluated in its first use. The device included three applicators with distinct electroacoustic characteristics. Output power was measured using a UPM-DT-1000PA acoustic radiation force balance (OHMIC Instruments Inc., St. Charles, USA), which had a valid calibration and up-to-date preventive maintenance at the time of the measurements.

The tests were conducted over two consecutive days, with five hours of operation per day. During the measurements, the applicators was secured in a mounting fixture, centered and partially submerged in the water reservoir of the acoustic radiation force balance, with its emitting surface perpendicular to the metal cone. A fixed distance of 3 cm was maintained between the emitting surface and the cone throughout the tests. The water was replaced at the beginning of both test days and allowed to stabilize for 1.5 hours before testing.

To monitor the thermal conditions during the experiments, water and ambient temperatures were recorded at the beginning and end of each test period. Water temperature was measured using a TM-947SD thermometer (Lutron Electronic Enterprise Co., New Taipei, Taiwan), whereas a digital thermometer was used for ambient temperature. Other potential sources of interference were also controlled: the air conditioning remained off during testing, and the ultrasound device was placed on a separate table from the acoustic radiation force balance to minimize the transmission of mechanical vibrations.

The ILIFT device comprises three applicators operating at 5.5 or 10.10 MHz, with different ERAs and maximum output powers, as summarized in Table 1. Output power was measured individually for each applicator at all available nominal power settings.

TABLE I. Operating characteristics and output power settings evaluated for the three ILIFT dermatological ultrasound applicators

	Applicator 1	**Applicator 2**	**Applicator 3**
Frequency (MHz)	5.5	10.10	10.10
Nominal ERA (cm^2)	3	3	1
Maximum output power (W)	9	9	3
Power range evaluated (W)	0.3–9.0	0.3–9.0	0.1–3.0
Increment (W)	0.3	0.3	0.1

The obtained data were entered and organized into Microsoft Excel® spreadsheets, where an automated procedure was also developed to calculate the acoustic power accuracy test, as expressed in equation (1). This test is well-established in the literature for evaluating therapeutic ultrasound equipment and consists of comparing the measured acoustic power with the nominal power specified by the manufacturer [6]. The relative error was calculated as the percentage difference between the measured and nominal output power values. A ±20% interval, based on IEC 61689:2022, was adopted as a comparative criterion; however, the operating frequencies evaluated in this study fall outside the scope of the standard.

$$E = \left(\frac{P_{Medida} - P_{Nominal}}{P_{Nominal}}\right) \times 100 \quad (1)$$

B. Analysis of the repeatability of sound power measurements

For each nominal power setting of each DUS applicators, 10 consecutive measurements were performed under the same experimental conditions. The mean relative error (MRE) and standard deviation of the relative error (SDE) were calculated to characterize the average deviation from the nominal power and the dispersion of the relative errors, respectively, following the approach described by Costa Júnior et al. [17]. The coefficient of variation (CV) was also calculated to evaluate the repeatability of the output power measurements. These parameters are defined in Eqs. (2)(4). Data analysis was performed using MATLAB® software (Version R2019a; Mathworks, Natick, USA) and Microsoft Excel. MRE and SDE were calculated and the figures were generated in MATLAB, whereas the remaining parameters were obtained using Microsoft Excel.

$$MRE = \frac{1}{n}\sum_{i=1}^{n}(E_i) \quad (2)$$

$$SDE = \sqrt{\frac{\sum_{i=1}^{n}(E_i - MRE)^2}{n - 1}} \quad (3)$$

$$CV = \frac{s}{\bar{x}} \times 100 \quad (4)$$

where E_i is the relative error associated with the ith measurement for a given applicator and nominal power setting, $\bar{x}$ is the mean output power obtained from the repeated measurements, and s is the standard deviation of these measurements.

III. Results

All measurements were performed at night. Throughout the experiments, the water temperature ranged from 26.0 to 27.9 °C, whereas the ambient temperature decreased from 25.9 to 25.1 °C.

Figure 1 presents the results obtained for the applicators with a nominal frequency of 5.5 MHz and a nominal ERA of 3 cm². The top panel shows the mean relative error in acoustic power as a function of the selected nominal power, along with the ±20% tolerance limits specified by the IEC standard for therapeutic ultrasound equipment. The bottom panel shows the standard deviation of the relative error in acoustic power, thereby allowing measurement repeatability to be assessed.

Figure 2 presents the same analysis for an applicator with a nominal effective radiating area of 3 cm² operating at a

nominal frequency of 10.10 MHz. Figure 3 presents the corresponding results for an applicator with a nominal frequency of 10.10 MHz and a nominal effective radiating area of 1 cm². The top panel shows the mean relative error in acoustic power, while the bottom panel presents the corresponding standard deviation.

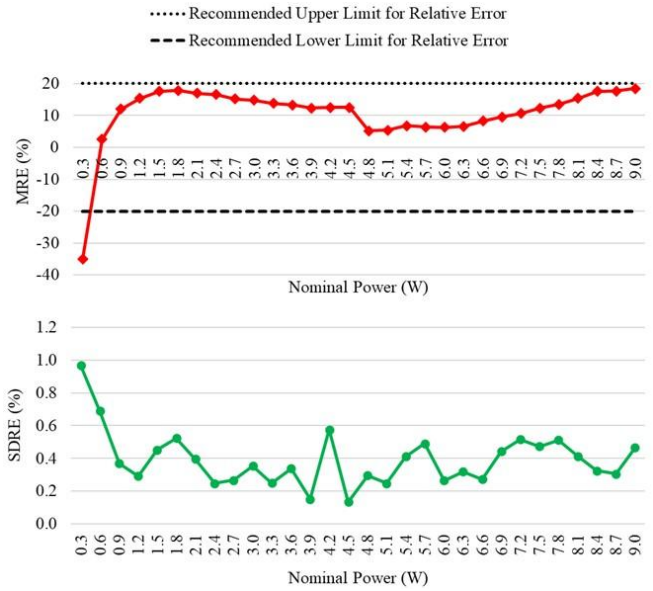


Fig. 1. Mean relative error in acoustic power (top panel) and corresponding standard deviation (bottom panel) as a function of nominal power for an applicator with a nominal frequency of 5.5 MHz and a nominal ERA of 3 cm².

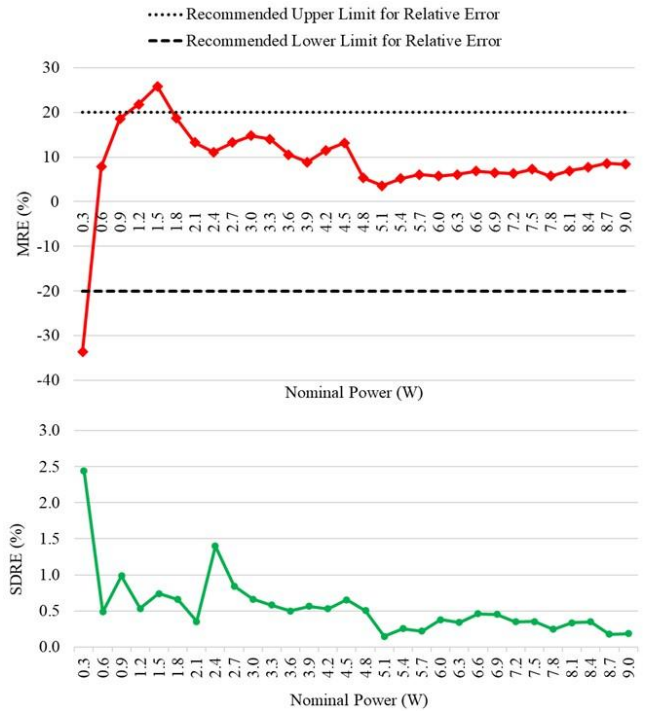


Fig. 2. Mean relative error in acoustic power (top panel) and corresponding standard deviation (bottom panel) as a function of nominal power for an applicator with a nominal frequency of 10.10 MHz and a nominal ERA of 3 cm²

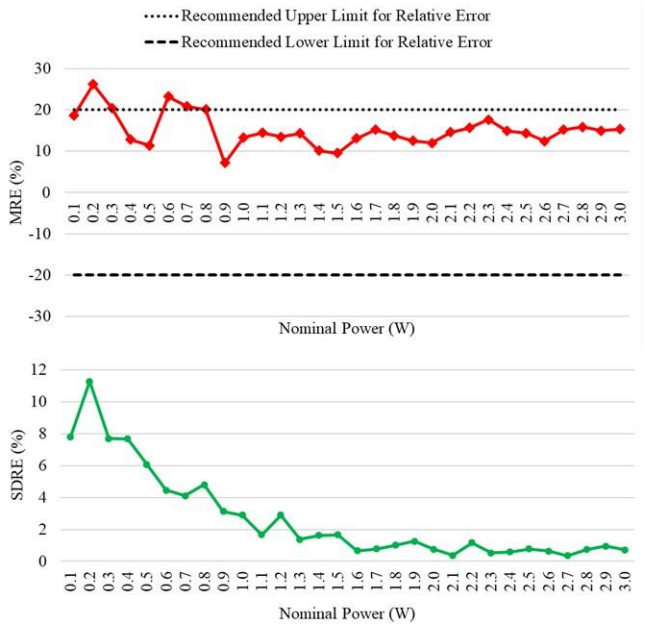


Fig. 3. Mean relative error in acoustic power (top panel) and corresponding standard deviation (bottom panel) as a function of nominal power for an applicator with a nominal frequency of 10.10 MHz and a nominal ERA of 1 cm²

Figure 4 presents the coefficient of variation of the acoustic power measurements as a function of nominal power for the three applicators evaluated. Overall, the CV remained below 2% for the applicators with a nominal acoustic response area (ARA) of 3 cm², regardless of the operating frequency, indicating high measurement repeatability across most of the investigated power range.

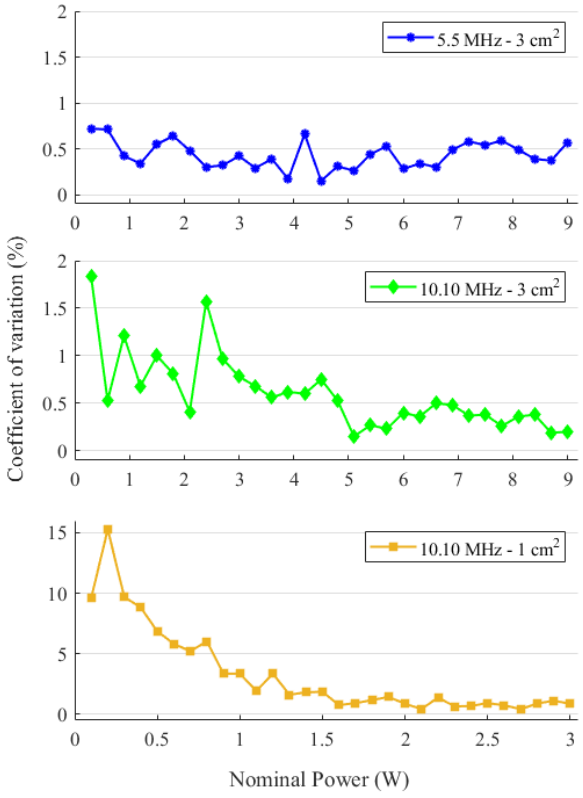


Fig. 4. Coefficient of variation of acoustic power measurements as a function of rated power for the applicators evaluated.

In contrast, the applicators with a nominal frequency of 10.10 MHz and a nominal acoustic area (ERA) of 1 cm² exhibited substantially higher coefficient of variation (CV) values at low power levels, reaching approximately 15% at the lowest power settings evaluated. However, variability decreased markedly as nominal power increased, approaching 2% at power levels above approximately 1.1 W. This behavior suggests a reduction in measurement uncertainty as the emitted acoustic power increases, which is consistent with the improvement in signal-to-noise ratio typically observed at higher power levels.

Overall, the results demonstrate satisfactory repeatability of the radiation force balance measurements, particularly for applicators with a nominal ERA of 3 cm². The greater variability observed for the 1 cm² applicators at low power levels is likely due to its lower acoustic output power and the consequent reduction in measurement system sensitivity under these operating conditions.

IV. Discussion

This study evaluated the metrological performance of dermatological ultrasound equipment operating at 5.5 and 10.10 MHz based on the agreement between measured and nominal output power values and the repeatability of measurements obtained using a radiation force balance. Overall, the three applicators demonstrated measured acoustic output power within the ±20% reference criterion derived from IEC 61689 across most of the operating range, although deviations were observed at low power levels. Under these conditions, the combination of high repeatability and reduced accuracy indicates the predominance of systematic errors, a pattern typically observed in radiation force measurements when the forces exerted on the target are small [16,18].

Because the operating frequencies of the dermatological ultrasound equipment (5.5 and 10.10 MHz) are outside the 0.5–5 MHz frequency range specified in IEC 61689:2022 for physiotherapy ultrasound systems, this standard was not considered an applicable conformity standard. In addition, no specific standard applicable to the type of non-focused dermatological ultrasound equipment evaluated in the present study was identified. Nevertheless, the ±20% criterion specified in IEC 61689:2022 was adopted as a reference for comparison with the measured acoustic output power.

The results corroborate previous findings showing that the acoustic power output of therapeutic ultrasound equipment may not remain uniform throughout the entire operating range. Several studies have reported discrepancies between programmed and measured values, which have been attributed to the electroacoustic characteristics of the applicators, operating conditions, and wear resulting from clinical use [14,15,19,20].

However, unlike the equipment evaluated in previous studies, the device tested here was new and had not been used previously, which may have contributed to the close agreement between the measured and nominal output power values. This result reinforces the fact that periodic calibration, preventive maintenance, and quality control programs play a decisive role in the metrological reliability of the equipment [6,21].

The literature also highlights the importance of assessment at multiple points across the operating range. While Artho et al. [19] found that approximately 39% of the evaluated devices exhibited acoustic power levels outside acceptable limits at least one intensity setting, Nascimento et al. [6] found deviations between measured and nominal acoustic power even in devices undergoing preventive maintenance, with some measurements exceeding the limits established by IEC 61689. These results indicate that periodic maintenance alone does not ensure adequate metrological performance, underscoring the need for regular performance assessments and calibration.

A distinguishing feature of this study is the evaluation of applicators operating at frequencies above those commonly used in therapeutic applications. Most previous studies have focused on devices operating at 1 and 3 MHz, whereas data on therapeutic ultrasound equipment operating between 5 and 10 MHz remain limited [19, 15, 22]. This lack of data warrants further investigation of the metrological performance of these applicators, particularly at low power levels.

Differences in accuracy and precision were observed among applicators with different effective radiating areas. The applicator with an ERA of 3 cm² generally exhibited higher accuracy and precision across the operating range, whereas lower accuracy and greater variability at low power levels were observed for the applicator with an ERA of 1 cm². Although these findings do not establish a causal relationship between ERA and measurement performance, the differences may be partly associated with the lower acoustic output of applicators with a smaller radiating area. At a given intensity, a smaller ERA results in lower total acoustic power and, therefore, a lower radiation force. At low output levels, instrumental effects account for a larger proportion of the measured signal and may contribute to the lower accuracy and precision observed. [23].

The observed deviations were predominantly concentrated at the lower nominal power levels. For the 5.5 MHz applicators, only the 0.3 W power setting exceeded the ±20% limit established by IEC 61689, although repeatability remained high. A similar trend was observed for the 10.10 MHz applicators with a 3 cm² ERA, where variability progressively decreased as power increased. In contrast, the 10.10 MHz applicators with a 1 cm² ERA exhibited greater dispersion at lower power levels but also showed progressive improvement in repeatability as power increased. These results reinforce that the observed limitations are primarily associated with the low magnitude of radiation force at low power levels, a condition that reduces the signal-to-noise ratio of the measurement system [24].

A notable finding was observed at the lowest nominal power setting (0.3 W), for which the mean relative error was approximately −35%. The negative error indicates that the measured acoustic power exceeded the nominal value, falling outside the recommended lower limit of −20%. This discrepancy is clinically relevant because an acoustic output greater than intended may increase tissue exposure and potentially compromise treatment safety and dosimetric accuracy.

From a biomedical engineering perspective, the results reinforce that the metrological assessment of ultrasound equipment should not rely on accuracy testing at a single power setting, as performance may vary across the operating

range. Quality control procedures conducted during maintenance and calibration should therefore include acoustic power measurements at multiple output levels to detect deviations that might otherwise go unnoticed. Regular performance testing, combined with appropriate maintenance and calibration, is important for the safety and reliability of therapeutic ultrasound equipment [15, 23].

This study has limitations, as only one unit of a single commercial model from one manufacturer was evaluated, precluding an assessment of unit-to-unit variability and limiting the generalizability of the results. Moreover, the ultrasound device was new, and the measurements reported here were obtained during its first use. The results may therefore differ from those obtained with equipment subjected to routine clinical use, maintenance, or calibration over time. Future studies should include multiple units of the same model and manufacturer to assess inter-unit variability, as well as models from other manufacturers. Investigations covering a broader range of operating frequencies, applicators, and power settings are also needed for a more comprehensive characterization of the metrological performance of high-frequency therapeutic ultrasound equipment.

In summary, the results demonstrate that dermatological ultrasound equipment operating at frequencies above 5 MHz can exhibit satisfactory metrological performance when subjected to appropriate calibration, although evaluation at low power levels requires greater attention due to the increased influence of uncertainties inherent to the measurement method. These findings contribute to expanding knowledge regarding the metrological characterization of high-frequency therapeutic ultrasound devices and provide a basis for improving quality control programs and future normative recommendations applicable to this class of devices.

A. Limitations

The present study evaluated a single unit of one model of dermatological ultrasound equipment. Therefore, the results cannot be used to distinguish device-specific characteristics from variations associated with the technology as a whole. Despite this limitation, previous studies have evaluated the acoustic output power and/or acoustic output intensity of multiple physiotherapy ultrasound devices, with each study including a single model of equipment [10, 25]. In particular, Guirro and Santos [25] evaluated multiple physiotherapy ultrasound devices that were new and had not been previously used, providing a precedent for assessing the acoustic output of new equipment of a specific model. The equipment evaluated in the present study is also a recently introduced system, which began commercial distribution in April 2026, limiting the availability of units with different periods of use at the time of this study. Further investigations involving multiple units of the same model and, when available, units with different periods of use are necessary to assess inter-unit variability and long-term stability.

V. CONCLUSION

This study demonstrated that the dermatological ultrasound equipment evaluated complied with the accuracy limits established by IEC 61689 across most of its operating range, showing satisfactory metrological performance following calibration. The observed deviations were limited to the lower nominal power settings, underscoring the need for assessments across the entire operating range. Furthermore, this work makes an original contribution to the literature by evaluating the acoustic power of therapeutic applicators operating at frequencies of 5.5 MHz and 10.10 MHz, given that existing studies focus predominantly on 1 MHz and 3 MHz equipment. Thus, the results highlight the importance of periodic calibration and proper maintenance as essential strategies for ensuring equipment reliability, patient safety, and the quality of therapeutic ultrasound treatments

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

[1] E. C. Monteiro and M. L. Lessa, "A metrologia na área de saúde: garantia da segurança e da qualidade dos equipamentos eletromédicos," *Engevista*, vol. 7, no. 2, pp. 51–60, 2005.

[2] P. Squara et al., "Metrology part 1: definition of quality criteria," *Journal of Clinical Monitoring and Computing*, vol. 35, no. 1, pp. 17–25, 2021.

[3] P. Squara et al., "Metrology part 2: procedures for the validation of major measurement quality criteria and measuring instrument properties," *Journal of Clinical Monitoring and Computing*, vol. 35, no. 1, pp. 27–37, 2021.

[4] A. H. Zamzam et al., "A systematic review of medical equipment reliability assessment in improving the quality of healthcare services," *Frontiers in Public Health*, vol. 9, 2021.

[5] A. Badnjević and L. G. Pokvić, "Legal metrology framework for medical devices," in *Clinical Engineering Handbook*. Academic Press, 2020, pp. 774–779.

[6] A. R. M. Nascimento, M. G. M. Pinto, T. P. Omena, W. C. A. Pereira, and J. F. S. Costa-Júnior, "quão confiáveis são os valores de potência acústica emitidos por equipamentos de ultrassom em fisioterapia?," *Revista Brasileira de Física Médica*, vol. 17, e709, 2023.

[7] S. M. Z. Uddin, D. E. Komatsu, T. Motyka, and S. Petterson, "low-intensity continuous ultrasound therapies—A systematic review of current state-of-the-art and future perspectives," *Journal of Clinical Medicine*, vol. 10, no. 12, p. 2698, 2021.

[8] M. Lafond, A. Payne, and C. Lafon, "Therapeutic ultrasound transducer technology and monitoring techniques: a review with clinical examples," *International Journal of Hyperthermia*, vol. 41, no. 1, p. 2389288, 2024.

[9] D. Dalecki, "Mechanical bioeffects of ultrasound," *Annual Review of Biomedical Engineering*, vol. 6, pp. 229–248, 2004.

[10] A. R. A. Martins, A. A. S. Junior, T. P. Omena, W. C. A. Pereira, and R. P. B. Costa-Félix, "Aferição da intensidade de dois aparelhos de ultrassom fisioterapêutico pós-manutenção," in *Anais do XXV Congresso Brasileiro de Engenharia Biomédica (CBEB)*, Foz do Iguaçu, Brazil, 2016, pp. 1985–1988.

[11] L. S. Lima et al., "Exatidão da potência acústica de equipamentos comerciais de ultrassom fisioterapêutico," in *Anais do XXIII Congresso Brasileiro de Engenharia Biomédica (CBEB)*, Porto de Galinhas, Brazil, 2012.

[12] *IEC 61689: Ultrasonics—Physiotherapy systems—Field specifications and methods of measurement in the frequency range 0.5 MHz to 5 MHz*. Geneva, Switzerland: International Electrotechnical Commission, 2013.

[13] R. Guirro and S. C. B. dos Santos, "A realidade da potência acústica emitida pelos equipamentos de ultra-som terapêutico: uma revisão," *Fisioterapia e Pesquisa*, vol. 4, no. 2, pp. 76–82, 1997.

[14] S. M. Schabrun, H. L. Walker, and L. S. Chipchase, "The accuracy of therapeutic ultrasound equipment: a systematic review," *Physical Therapy Reviews*, vol. 13, no. 6, pp. 443–449, 2008.

[15] R. R. J. Guirro, F. Serrão, D. Elias, and A. J. Bucalon, “Calibration of therapeutic ultrasound equipment,” *Physiotherapy*, vol. 83, no. 8, pp. 419–422, 1997.

[16] S. E. Fick and F. R. Breckenridge, “Ultrasonic power output measurement by pulsed radiation pressure,” *Journal of Research of the National Institute of Standards and Technology*, vol. 101, no. 5, pp. 659–669, 1996.

[17] J. F. S. Costa Júnior, G. C. Parcero, and J. C. Machado, “Comparison analysis of four processing methods employed in dynamic elastography to estimate viscoelastic parameters of a medium: Tests using computational simulation and experiment,” *Biomedical Physics & Engineering Express*, vol. 3, Art. no. 025010, 2017.

[18] S. E. Fick, “Ultrasound power measurement by pulsed radiation pressure,” *Metrologia*, vol. 36, no. 4, pp. 351–356, 1999.

[19] P. A. Artho et al., “A calibration study of therapeutic ultrasound units,” *Physical Therapy*, vol. 82, no. 3, pp. 257–263, 2002.

[20] S. Schabrun, H. Walker, and L. Chipchase, “How accurate are therapeutic ultrasound machines?,” *Hong Kong Physiotherapy Journal*, vol. 26, no. 1, pp. 39–44, 2008.

[21] G. Durando and C. Guglielmone, “Ultrasound physiotherapy devices: how to measure them,” *IEEE Instrumentation & Measurement Magazine*, vol. 19, no. 5, pp. 15–48, 2016.

[22] G. ter Haar, “Therapeutic applications of ultrasound,” *Progress in Biophysics and Molecular Biology*, vol. 93, nos. 1–3, pp. 111–129, 2007.

[23] A. V. Alvarenga and R. P. B. Costa-Félix, “Metrological aspects on therapeutic ultrasound parameters: effective radiating area and non-uniformity ration,” *Physics Procedia*, vol. 3, no. 1, pp. 643–649, 2010.

[24] P. A. Oliveira, A. V. Alvarenga, and R. P. B. Costa-Félix, “Ultrasonic output power measurement according to IEC 61161:2013,” in *Proc. XXVI Brazilian Congress on Biomedical Engineering (CBEB 2018)*, IFMBE Proc., vol. 70/1, Springer, Singapore, 2019, pp. 851–854.

[25] R. Guirro, S.C.B. Santos. “Evaluation of the acoustic intensity of new ultrasound therapy equipment”. *Ultrasonics*. vol. 39, no. 8, pp. 553-7, 2002.

Calibração de Medidores de Campo Elétrico de Baixa Frequência

1st Regis Renato Dias
Laboratório de Metrologia Elétrica
Instituto de Pesquisas Tecnológicas
São Paulo, Brasil
rrdias@ipt.br

2nd Carla Costa Guimarães
Engenharia Biomédica
Pontifícia Universidade Católica
São Paulo, Brasil
cguimaraes@pucsp.br

***Abstract*—Este trabalho tem a intenção de fortalecer e complementar a cadeia de rastreabilidade brasileira para a grandeza campo elétrico de baixa frequência, detalhando o desenvolvimento de um sistema e a comparação experimental de um padrão para calibração de medidores de campo elétrico de baixa frequência, baseado no sistema de placas paralelas conforme a norma IEC 61786-1. Durante o processo, testamos o procedimento recomendado pela norma, até os limites de exposição estabelecidos por organismos internacionais como a International Commission on Non-Ionizing Radiation Protection (ICNIRP). O sistema montado no Instituto de Pesquisas Tecnológicas do Estado de São Paulo, com placas de alumínio de 1,5 m x 1,5 m e distância de separação de 0,6 m, foi avaliado quanto às principais componentes de incerteza: tensão aplicada (divisor de tensão e multímetro digital), distância entre placas, não uniformidade do campo, efeitos de borda, deslocamentos do sensor e alinhamento. A incerteza expandida de calibração obtida foi de 1,2% (k=2), resultando em "Test Uncertainty Ratio" (TUR) de aproximadamente 4 vezes em relação à incerteza típica dos medidores comerciais (5%), valor dentro da faixa recomendada para atender aos requisitos da maioria dos usuários que necessitam de rastreabilidade no Brasil. Além disso, este trabalho visa futuramente preencher a lacuna existente na Rede Brasileira de Calibração, que atualmente não possui laboratório acreditado para essa grandeza.**

***Index Terms*—Metrologia, Medidores, Campo Elétrico, Campo Magnético, limite de exposição, calibração, placas paralelas, incerteza de medição**

I. Introdução

A expansão da infraestrutura elétrica no Brasil, com o aumento do número de linhas de transmissão e subestações [3], tem intensificado a preocupação com a exposição da população a campos elétricos e magnéticos de baixa frequência. Como resultado, normas nacionais [1] e regulamentações [3] estabeleceram limites de exposição e requisitos para instrumentos de medição. Entretanto, até o momento, não há laboratório acreditado pela Rede Brasileira de Calibração (RBC) para calibração de medidores de campo elétrico em baixa frequência [7], nem estrutura de rastreabilidade em campos elétricos, levando parte dos medidores usados no Brasil a ser enviada para calibração fora do país, com mais custos e prazos maiores, o que motivou o desenvolvimento deste trabalho. Este trabalho apresenta o desenvolvimento de um padrão baseado em placas paralelas no Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT), bem como a validação do procedimento de calibração, visando suprir essa lacuna e a conformidade com os requisitos normativos.

II. Efeitos biológicos e Limites de exposição

Os limites de exposição a campos elétricos e magnéticos de baixa frequência, estabelecidos por normas como a ABNT NBR 25415 e agências reguladoras, como a Agência Nacional de Energia Elétrica (ANEEL) e Agência Nacional de Telecomunicações (ANATEL), são fundamentados em diretrizes internacionais de proteção radiológica não ionizante [8]. Estas diretrizes visam evitar efeitos adversos comprovados, como a estimulação de nervos e músculos e o surgimento de correntes induzidas no corpo humano (em baixas frequências) e aquecimento excessivo de tecidos (em altas frequências). Embora estudos subsequentes [9], [10], [11] continuem a investigar possíveis efeitos de longo prazo e lacunas de conhecimento, o consenso científico atual, corroborado por pareceres em comitês científicos de risco ambiental e à saúde [10], é de que não há evidências convincentes de riscos à saúde para exposições dentro dos limites estabelecidos. A Tabela I apresenta os níveis de referência para 50Hz e 60Hz, adotados pela regulamentação brasileira [3].

Os limites de exposição são definidos por comitês internacionais de saúde "International Commission on Non-Ionizing Radiation Protection" (ICNIRP) e Organização Mundial da Saúde (OMS) e adotados pela norma nacional de referência ABNT NBR 25415 e normas (portarias) publicadas pelas agências reguladoras nacionais (ANEEL e ANATEL).

Tais limites visam evitar prioritariamente os efeitos adversos já comprovados, além de evitar os efeitos de exposição a longo prazo que ainda estão sendo estudados.

TABLE I
Níveis de referência de exposição a campos elétricos e magnéticos

Frequência	**50 Hz**		**60 Hz**	
Campo	**Elétrico (kV/m)**	**Magnético (μT)**	**Elétrico (kV/m)**	**Magnético (μT)**
Público em Geral	5,00	200,0	4,17	200,0
População Ocupacional	10,0	1000,0	8,33	1000,0

III. Requisitos da Norma Brasileira para Medição e Calibração

A norma ABNT NBR 25415 — versão brasileira da IEC 61786-1 [2] — estabelece métodos de medição e níveis de referência para exposição a campos elétricos e magnéticos nas frequências de 50Hz e 60Hz. A norma também define os requisitos para instrumentação de medição e os procedimentos recomendados para calibração.

O foco deste trabalho foi o desenvolvimento de um procedimento de calibração de medidores de campo elétrico de baixa frequência (50Hz e 60Hz) até 40 kV/m, grandeza para a qual não há laboratório acreditado pela RBC. A calibração desses instrumentos, utilizados em medições em ambientes industriais, subestações e linhas de transmissão, é essencial para assegurar a rastreabilidade da medidas, o cumprimento dos limites de exposição estabelecidos e a segurança da população exposta.

Vale destacar que, para campos magnéticos, o IPT já oferece serviço de calibração de medidores de campo magnético AC e DC, com escopo acreditado pela RBC nas faixas de amplitude e incertezas exigidas pela norma.

A. Medidores de Campo Elétrico

Segundo a norma ABNT NBR 25415, os medidores de campo elétrico utilizados nas medições de exposição humana a campos devem ter incertezas de ±(10% + 2V/m) ou menores. Entretanto, os medidores mais novos comumente encontrados no mercado, têm incerteza de 5% ou melhor.

B. Padrão de Campo Elétrico

Para a calibração dos medidores de campo elétrico, a norma ABNT NBR 25415 referencia diretamente a norma IEC 61786-1, para que seja seguida.

Dentre as possibilidades de padrões para calibração de campo elétrico recomendadas na norma IEC 61786-1 [2], escolhemos uma das mais utilizadas no mundo, o Sistema de placas paralelas. Esse sistema tem a vantagem de permitir calibrar todos os tipos de medidores, além da possibilidade de calibrar sensores com dimensões maiores, devido ao tamanho do dispositivo construído no IPT.

A norma contém descrições de dimensões e proporções mínimas do sistema de placas paralelas, para que possa ser utilizada a incerteza recomendada.

Para esse tipo de sistema, o cálculo do campo elétrico teórico vem da equação básica do capacitor de placas planas e paralelas, $E = Vaplicado/D$, onde E é o campo elétrico do sistema, $Vaplicado$ é a tensão aplicada nas placas e D é a distância de separação entre elas.

IV. Incerteza da Medição de Campo Elétrico

A incerteza da medição de campo elétrico com o sistema de placas paralelas é calculada a partir da combinação de incertezas dos principais padrões envolvidos.

A seguir, discutiremos as principais componentes de incertezas (efeitos) que foram consideradas no cálculo final do sistema construído, Fig.1.

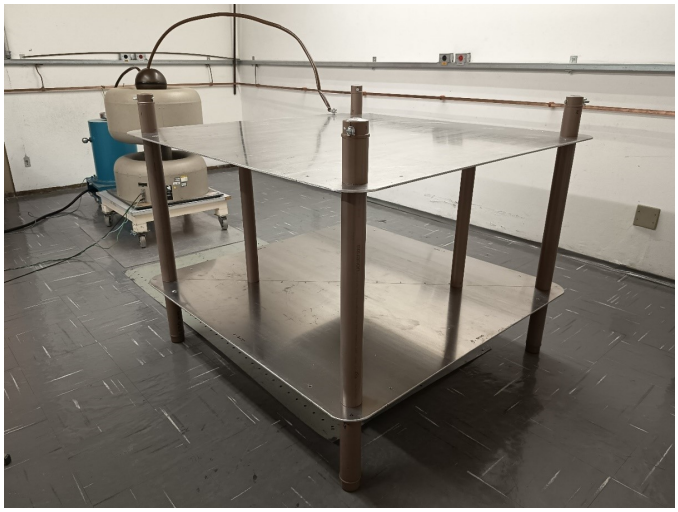


Fig. 1. Arranjo experimental para calibração de medidores de campo elétrico: sistema de placas paralelas com placas de alumínio de 1,5 m × 1,5 m com separação de 0,6 m (relação D/L = 0,4), instalado no IPT. Ao fundo, observa-se o aplicador de tensão (fonte de tensão) e divisor de tensão padrão.

A. Incerteza da tensão aplicada ($Vaplicado$)

Para o sistema montado, as incertezas relacionadas à tensão aplicada ao mesmo vêm de 2 fontes principais: um divisor resistivo de tensão de relação de 100000:1 e um multímetro digital 6 ½ dígitos.

A incerteza recomendada pelo fabricante do divisor de tensão usado com tensão AC é de 0,05% da relação, enquanto a incerteza recomendada pelo fabricante do multímetro digital para a medição de 60 mV, equivalente à leitura de 6 kV com o divisor, é igual a 0,13%. Essa é a amplitude de tensão necessária para o ponto de 10 kV/m, com distância de 0,6m.

B. Incerteza da distância entre as placas (D)

A incerteza relacionada à distância entre as placas paralelas é devido à régua graduada utilizada para a medição da distância. A régua graduada utilizada tem incerteza de 0,5 mm em toda a extensão de medição, o que representa 0,083% no ponto de medição de 0,60 m de separação entre as placas.

C. Incerteza do dispositivo de placas paralelas (k)

Considerando as proporções recomendadas na norma e o tamanho típico dos medidores comerciais (normalmente menores que um cubo de (0,1 m x 0,1 m x 0,1 m) e maior diagonal menor que 0,23 m), foi montado um capacitor com 2 placas paralelas, de alumínio com dimensões (L) de 1,5m x 1,5m, com distância de separação entre elas (D) de 0,6 m, que segundo a norma, seria suficiente para ter uma incerteza de 1%, somente devido ao dispositivo gerador de campo.

Segundo a norma, a incerteza do dispositivo será da ordem de 1%, desde que a relação entre a distância das placas (D) e o seu lado (L) deve ser $D/L \leq 0,5$. O sistema construído possui L= 1,5 m e D= 0,6 m, resultando em uma relação $D/L = 0,4$, que é ainda mais conservadora que a recomendada.

A não planicidade e não paralelismo das placas, devido às imperfeições geométricas na construção das placas, causam variações no campo elétrico gerado. Embora essas componentes sejam geralmente pequenas quando as placas estejam bem construídas, elas podem ser avaliadas durante a medição do espaçamento das placas e devem ser consideradas na análise de incerteza.

Uma correção k, referente à variável do dispositivo, pode ser levada em consideração, com os efeitos empíricos do dispositivo montado.

D. Validação da incerteza do dispositivo de placas paralelas

Para a validação da incerteza do dispositivo, foi utilizado um medidor de campo elétrico com o princípio de placas paralelas com dimensões de aproximadamente 17 cm x 7 cm e espaçamento de 4 cm entre as placas. Durante a validação, esse medidor teve resolução de 0,1% ou melhor.

Por causa de suas dimensões, esse "probe" mede um campo elétrico médio, mas que já está inserido no volume central dentro do dispositivo com campo uniforme.

A uniformidade do campo, devido à presença do medidor, com suas dimensões controladas (relação entre a dimensão do medidor, as dimensões da placa e o espaçamento entre elas, fornece uma incerteza de 0,5% [4].

Deslocando o "probe" do centro, em relação ao eixo X ou Y (paralelamente às placas) em distâncias de até 100 mm, o campo indicado pelo medidor variou até 0,1%.

Deslocando o "probe" do centro, em relação ao eixo Z (aproximando de alguma das placas energizadas) em +20 mm, o campo indicado pelo medidor variou -0,6% e em -20 mm, o campo indicado pelo medidor variou +0,6% (comportamento simétrico).

Também relevante, é a incerteza devido ao efeito das bordas finitas das placas, que tem um efeito de aproximadamente +0,3% e que esse efeito seria "compensado" pelo efeito de proximidade do sensor com as placas -0,4% [5].

Por fim, o ângulo de inclinação do sensor em relação às placas paralelas é importante e deve ser considerado um sistema de alinhamento do mesmo. Para uma incerteza no alinhamento de até 5°, isso resulta em uma incerteza de aproximadamente 0,31% [6].

Na tabela II, apresentamos as principais componentes relacionadas ao dispositivo de placas paralelas. Nela, também apresentamos a incerteza combinada calculada do dispositivo para comparação com o valor proposto na norma.

TABLE II
ORÇAMENTO DE INCERTEZAS PARA O DISPOSITIVO DE PLACAS PARALELAS

Componente de incerteza	**Incerteza (%)**	**Distribuição**
Não uniformidade do campo	0,50	Retangular
Variação no eixo X ou Y	0,10	Retangular
Variação no eixo Z	0,60	Retangular
Dimensão finita das placas	0,40	Retangular
Alinhamento do sensor e das placas	0,31	Retangular
Incerteza combinada	0,54	-
Incerteza expandida (k=1,732)	**0,94**	**Retangular**

Considerando os efeitos avaliados experimentalmente, juntamente com outros efeitos relevantes da literatura, obtivemos uma incerteza de medição de 0,94%, valor levemente inferior à incerteza recomendada pela norma IEC 61786-1 [2].

Embora as medidas tenham sido executados com somente um tipo de medidor (em 2 medidores de fabricantes distintos), esse resultado reforça a adoção de 1% da proposto na norma.

E. Cálculo da Incerteza final

A incerteza final de calibração de um medidor de campo elétrico com o sistema de placas paralelas é calculada a partir da combinação de incertezas dos principais padrões envolvidos; além da variabilidade e resolução dos padrões e do próprio instrumento sob calibração.

Como vimos, a disposição e a geometria do sistema influencia também no valor do campo elétrico de referência e, desta forma a equação de medição do campo elétrico deriva para $E = k * (Vaplicado/D)$, onde E é o campo elétrico da medição, k é a correção aplicada de acordo com o dispositivo, o $Vaplicado$ é a tensão aplicada nas placas e D é a distância de separação entre elas.

Para simplificação da tabela de componentes de incertezas, omitimos as componentes relativas à variabilidade (0,05%) e resolução dos padrões (0,01%) e do instrumento sob calibração (0,1%), por serem consideradas de menor magnitude e não impactarem a incerteza final. Dessa forma, apresentamos apenas as principais componentes relacionadas ao sistema, conforme visto na tabela III.

A incerteza expandida de medição foi calculada como a incerteza padrão combinada, multiplicada pelo fator de abrangência k=2, supondo variação normal, correspondente a uma probabilidade de abrangência de aproximadamente 95% [12]. Resultando numa incerteza de calibração de medidores de campo elétrico de 1,2 % para o sistema montado, Fig.2.

Ressaltando que, a garantia da rastreabilidade do campo elétrico no sistema montado no IPT se dá por duas frentes; a primeira relativa à metodologia utilizada no laboratório ser conforme o descrito na norma e a segunda, devido à rastreabilidade de todos os instrumentos envolvidos: o divisor de tensão padrão, o multímetro 6 ½ dígitos e a régua graduada (nas grandezas e escalas utilizadas durante o procedimento de calibração) terem sido calibrados em laboratórios acreditados pela Coordenação Geral de Acreditação (Cgcre - RBC).

TABLE III
ORÇAMENTO DE INCERTEZAS PARA CALIBRAÇÃO DE CAMPO ELÉTRICO

Componentes de incerteza	**Incerteza (%)**	**Distribuição**	**Incerteza padrão (%)**
Divisor de tensão (medição da tensão)	0,05	Retangular	0,029
Multímetro digital na saída do divisor (medição da tensão)	0,13	Retangular	0,075
Distância entre placas - D (régua graduada)	0,08	Retangular	0,048
Dispositivo de placas paralelas	1,00	Retangular	0,577
		Incerteza combinada	0,58
		Incerteza expandida (95%, k=2)	**1,2**

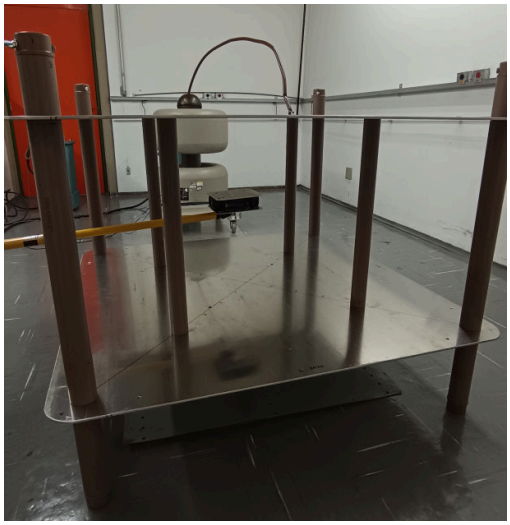


Fig. 2. Sistema de placas paralelas montado do IPT, com um sensor posicionado no centro, pronto para calibração

Um ponto importante, verificado durante as medições executadas, é que a proximidade de pessoas e objetos ao dispositivo, interfere nos valores medidos (provavelmente pela distorção das linhas de campo elétrico).

Outro ponto chave é o suporte do medidor de campo elétrico, que também fica sob efeito do campo de referência, podendo alterar o valor indicado pelo medidor. Por isso, é recomendado o uso do suporte do próprio medidor sob calibração, que também será utilizado nas medidas em externas. Caso contrário, será uma nova componente de incerteza.

V. Conclusão

Recomenda-se que a relação de incertezas entre um instrumento sob calibração e um padrão utilizado nessa calibração "Test Uncertainty Ratio" (TUR) seja entre 3 a 10 vezes.

Como os medidores de campo elétrico utilizados nas medições de exposição humana a campos têm normalmente incertezas de cerca 5%, e que a incerteza de medição calculada do sistema foi de 1,2%, chegamos na relação TUR próxima de 4 vezes, valor dentro da recomendação para oferecer o serviço de calibração para a maioria dos clientes que necessitem de rastreabilidade para as medições de campo elétrico no Brasil.

Conforme relatado, temos a linha de rastreabilidade ininterrupta e a metodologia aprovada pelas normas (nacional e internacional), por isso já estamos oferecendo para qualquer cliente do IPT, o serviço de calibração de Campos Elétricos de baixa frequência rastreado ao Sistema Internacional (SI).

VI. Próximos Passos

Como próximos passos, pretendemos solicitar a acreditação do serviço de calibração de medidores de campo elétrico em baixa frequência para a Cgcre.

Para tanto, uma comparação interlaboratorial com algum laboratório já acreditado pela norma ABNT NBR ISO/IEC 17025, uma garantia da confiabilidade dos resultados obtidos [13], se faz necessária.

Adicionalmente, buscaremos aprimorar o sistema desenvolvido, testando também outros tipos de medidores de campo elétrico, com o objetivo de entender cada componente de incerteza envolvida na instrumentação, a fim de poder reduzir a incerteza de medição, aplicando fatores de correção onde for possível.

Agradecimentos

Gostaria de agradecer principalmente ao engenheiro mecânico Ramon Valls Martin do IPT, pela construção do dispositivo de placas paralelas, que através de sua habilidades, conseguiu montar o dispositivo de acordo com as proporções recomendadas nas normas, mantendo assim as características elétricas esperadas.

Por fim, gostaria de agradecer à chefia do Laboratório de Metrologia Elétrica do IPT e à gerencia do mesmo, por acreditar e investir em uma capacitação tão importante para a segurança das pessoas no entorno das instalações elétricas de alta tensão.

References

[1] Associação Brasileira de Normas Técnicas (ABNT). Métodos de medição e níveis de referência para exposição a campos elétricos e magnéticos na frequência de 50 Hz e 60 Hz. NBR 25415. Rio de Janeiro, 2016.

[2] International Electrotechnical Commission (IEC), IEC 61786-1:2013: Measurement of DC magnetic, AC magnetic and AC electric fields from 1 Hz to 100 kHz with regard to exposure of human beings - Part 1: Requirements for measuring instruments. Geneva, 2013.

[3] Agência Nacional de Energia Elétrica (ANEEL), Resolução Normativa ANEEL Nº 915, Fevereiro 2021.

[4] MISAKIAN, M., Calibration of flat 60-Hz electric field probes, Bioelectromagnetics 5, pp. 447-450, 1984.

[5] Takuma, T., et al.. Analysis of calibration arrangements for AC field strength meters. IEEE Transactions on Power Apparatus and Systems, PAS-104(2), 489–496, 1985.

[6] LIU Zhipeng, LIN Haoyu, WU Yanli, LI Enguang, ZHANG Bolin. Low-frequency electric field standard device based on parallel plates, 2024.

[7] Instituto Nacional de Metrologia Qualidade e Tecnologia (INMETRO), Rede Brasileira de Calibração – RBC, Rio de Janeiro. Disponível em: https://www.gov.br/inmetro/pt-br/centrais-de-conteudo/sistemas/rbc. Acesso em: junho 2026.

[8] International Commission on Non-Ionizing Radiation Protection (ICNIRP). "Guidelines for limiting exposure to time-varying electric and magnetic fields (1Hz - 100kHz)". Health Physics, vol. 99, no. 6, 2010, pp. 818-836.

[9] International Commission on Non-Ionizing Radiation Protection (ICNIRP). "Gaps in knowledge relevant to the 'Guidelines for limiting exposure to time-varying electric and magnetic fields (1Hz – 100kHz)". Health Physics, vol. 118, no. 5, 2020, pp. 533-542. DOI: 10.1097/HP.0000000000001261.

[10] Scientific Committee on Health, Environmental and Emerging Risks (SCHEER). Potential health effects of exposure to electromagnetic fields (EMF): Update with regard to frequencies between 1Hz and 100kHz – final opinion. European Commission, 2024. Disponível em https://health.ec.europa.eu/document/download/85ef39d5-49dc-4b5a-b875-54e578d1d2bc_en.

[11] Swedish Radiation Safety Authority (SSM). Recent Research on Electromagnetic Fields and Health Risk, Nineteenth Report from SSM's Scientific Council on Electromagnetic Fields, 2025.

[12] European Co-operation for Accreditation (EA). EA-4/02 M: Evaluation of the uncertainty of measurement in calibration. 2022. Disponível em https://european-accreditation.org/wp-content/uploads/2025/10/EA-4-02.pdf.

[13] Instituto Nacional de Metrologia Qualidade e Tecnologia (INMETRO), DOQ-CGCRE-001: Orientação para o processo de acreditação de laboratórios, produtores de materiais de referência e provedores de ensaio de proficiência. Maio 2023, Rio de Janeiro.

Development and Validation of an Experimental Protocol for In Vitro Radiobiological Studies Using Low-Energy Proton Beams

1st Lara Beatriz B. da Silva
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
lara.ds@einstein.edu.br

2nd Manuela S. de Souza
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
manuela.dsouza@einstein.edu.br

3rd Cecília L. Prioste
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
cecilia.lprioste@einstein.edu.br

4th Maira S. Cucielo
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
Maira.Cucielo@einstein.br

5th Juliana Magdalon
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
juliana.magdalon@einstein.br

6th Roger Borges
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
roger.b@einstein.br

7th Tiago F. da Silva
Institute of Physics
University of São Paulo (USP)
São Paulo, Brazil
tfsilva@if.usp.br

8th Manfredo H. Tabacniks
Institute of Physics
University of São Paulo (USP)
São Paulo, Brazil
tabacniks@if.usp.br

9th Paulo de Tarso D. Siqueira
Nuclear Engineering Center (CEN)
Nuclear and Energy Research NInstitute (IPEN/CNEN)
São Paulo, Brazil
ptsiqueira@ipen.br

10th Julian M. B. Shorto
Nuclear Engineering Center (CEN)
Nuclear and Energy Research NInstitute (IPEN/CNEN)
São Paulo, Brazil
jmshorto@ipen.br

11th Hélio Yoriyaz
Nuclear Engineering Center (CEN)
Nuclear and Energy Research NInstitute (IPEN/CNEN)
São Paulo, Brazil
yoriyaz@usp.br

12th Paula C. G. Antunes
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein
Hospital Israelita Albert Einstein
São Paulo, Brazil
paula.antunes@einstein.br

***Abstract*—Proton therapy offers a dosimetric advantage through the concentrated energy deposition at the Bragg peak, but its biological response remains insufficiently characterized and is still managed clinically using a fixed Relative Biological Effectiveness (RBE = 1.1) that does not capture its true variability. Developing more accurate RBE models requires experimental data collected under well-controlled physical conditions. This preliminary study aimed to develop and initially validate an experimental protocol for in vitro radiobiological studies using a low-energy (2.6 MeV) proton beam. MG-63 cells, derived from a human osteosarcoma, were irradiated, the absorbed dose was estimated through Monte Carlo simulations using the PHITS code, and cell viability was assessed by MTT assay at different incubation periods. The 48-hour protocol showed a strong, significant negative correlation between absorbed dose and cell viability, whereas the 6-day protocol showed no consistent relationship, supporting the selection of the 48-hour incubation period for the proposed experimental protocol. These results support the initial validation of the proposed methodology and demonstrate its feasibility for obtaining dose-dependent biological responses using a low-energy research accelerator, providing an experimental foundation for subsequent radiobiological studies and future RBE estimation.**

***Index Terms*—radiobiology, cell culture, low-energy proton beam, cell viability**

I. Introduction

Clinical access to proton therapy has expanded considerably over the past decades, establishing it as a relevant treatment

modality for aggressive tumors, particularly for complex tumor locations or pediatric patients, where preserving healthy tissue is essential [1]. Its main physical advantage stems from the Bragg peak, where protons deposit most of their energy near the end of their range, followed by a sharp drop in dose beyond that point [2]. Because the depth of this peak depends on the initial beam energy and the properties of the traversed medium, dose can be concentrated within the tumor volume while minimizing exposure of surrounding structures, reducing the risk of late toxicity and radiation-induced secondary malignancies [3].

Despite these clinical advances, the biological response induced by protons is still not fully understood [4]. Current radiotherapy planning relies on a fixed Relative Biological Effectiveness (RBE = 1.1), defined as the ratio between the dose of a reference radiation (typically photons) and the proton dose required to produce the same biological effect. Although widely used in clinical practice, this value was derived from experimental studies conducted under specific conditions [5]. Recent studies, however, indicate that RBE varies with parameters such as LET, dose per fraction, cell type, and tissue radiosensitivity [6], [7]. This variability can lead to tumor underdosing or healthy-tissue overdosing, compromising treatment safety and efficacy, particularly in the distal beam region where LET is higher.

Several studies point to the need for mechanistic models capable of predicting RBE by integrating physical parameters, such as cellular dose and LET, with biological characteristics specific to each cell type [8]. Advancing these models, however, depends on experimental data obtained under rigorously controlled physical conditions [9], [10]. Accordingly, there is a growing effort within the scientific community to clarify the relationship between LET and RBE through in vitro experiments, computational simulations, and clinical analyses, aiming to develop more accurate radiobiological models that can be incorporated into proton therapy planning [11].

Obtaining this experimental data, however, depends on the availability of experimental platforms capable of irradiating cell cultures under well-defined physical conditions [6]. Although radiobiology studies are often conducted at clinical proton therapy centers or dedicated biological research facilities, access to this equipment remains limited [12], [13]. In this context, research accelerators represent an important alternative for expanding the production of experimental radiobiology data. Adapting in vitro experiments to these platforms, however, involves significant technical challenges due to the horizontal beam geometry and the short range of low-energy protons. These factors impose constraints on the experimental geometry, cell culture conditions, and control of the physical irradiation parameters, requiring the development of specific protocols capable of ensuring rigorously controlled physical conditions and reliable, reproducible quantitative data that can serve as a reference for radiobiology studies and for validating predictive models.

Within this context, this study evaluated the effect of low-energy proton irradiation, produced by the accelerator at the Laboratory of Materials Analysis by Ion Beams of the University of São Paulo (LAMFI-USP), on the viability of tumor cell lines, investigating the consistency and reproducibility of the biological responses, as well as the adequacy of the experimental adaptations required to use this platform for in vitro radiobiology studies. These results aim to establish criteria and standardize the experimental protocol for obtaining consistent quantitative experimental data, capable of supporting the development of mechanistic models for RBE estimation.

II. Materials and Methods

This section describes the experimental protocol developed for the in vitro irradiation of tumor cells with low-energy protons and the evaluation of their biological effects through viability assays. The protocol was established and optimized throughout the study, including the definition of culture and plating conditions and the standardization of the irradiation and MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] viability assay steps.

Fig. 1 illustrates the experimental workflow adopted, from cell culture and seeding through proton irradiation and subsequent viability assessment. Each step is described in detail in the following subsections.

A. Pre-irradiation

1) Cell culture: The cell line used as the primary model for the irradiations was MG-63, derived from a human osteosarcoma, widely used as in vitro models for tumors that are difficult to control therapeutically, characterized by high growth and invasive capacity [14], [15].

MG-63 cells were cultured in low-glucose (1 g/L) Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (pen-strep). The cells were kept in T-75 culture flasks in an incubator at 37°C with a 5% CO_2 atmosphere, with medium changes three times per week and trypsinization once confluence (the percentage of the culture surface occupied by adherent cells) reached 80%.

2) Sample preparation: Cell cultures were prepared in 33 mm Petri dishes (KASVI, model K13-0035, growth area = 8.5 cm^2), allowing the culture medium to be removed immediately before irradiation while keeping the cells adhered to the dish surface as a monolayer, oriented toward the beam, as shown in Fig. 2. The cell suspension volume was adjusted to yield 9×10^5 cells per dish, a concentration previously defined to reach approximately 80–90% confluence at the time of irradiation. The cell suspension was prepared in a sufficient volume of culture medium to dispense 1.5 mL per dish, homogenized, and then distributed into the Petri dishes. Finally, to ensure a homogeneous cell distribution, the dishes were vortexed at 250 rpm for 2 min.

Twenty-four hours after plating, the dishes showing the highest confluence and most homogeneous cell distribution were selected for irradiation. Samples were then sealed with parafilm and transported from Faculdade Israelita de Ciências

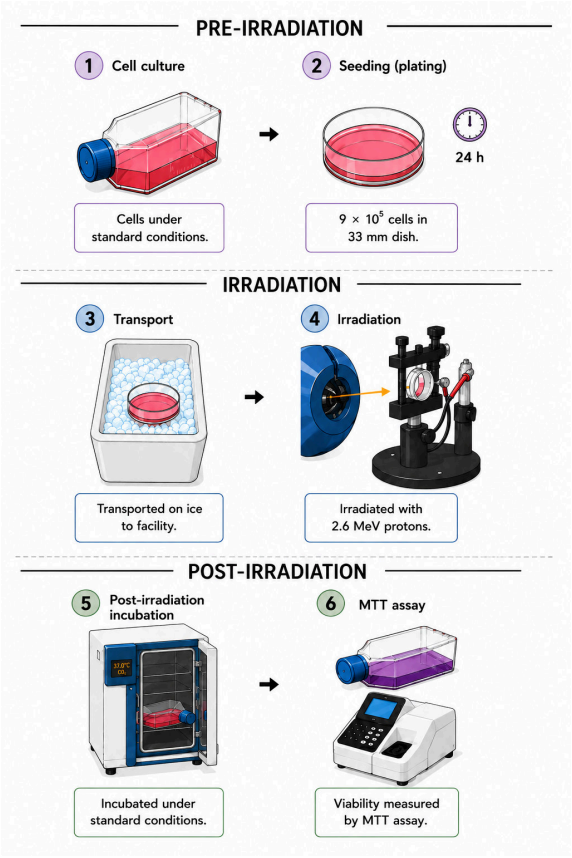


Fig. 1. Schematic flowchart of the experimental protocol steps. The pre-irradiation stage comprises (1) maintaining cells under standard culture conditions and (2) plating 9×10^5 cells in a 33 mm dish 24 h before irradiation. The irradiation stage includes (3) transporting the dishes on ice to LAMFI-USP and (4) irradiation with a 2.6 MeV proton beam. Post-irradiation: (5) incubating samples under standard culture conditions and (6) assessing cell viability using the MTT assay.

da Saúde Albert Einstein (FICSAE) to LAMFI-USP in an insulated container with crushed ice, to preserve cell viability during transport.

B. Proton irradiation

1) Experimental irradiation platform: Irradiations were performed using the proton beam available at the Laboratory of Materials Analysis by Ion Beams of the University of São Paulo (LAMFI-USP). Experiments were conducted on a 5SDH Pelletron-tandem electrostatic accelerator, with a beam energy of 2.6 MeV selected for corresponding to the equipment's most operationally stable energy range. The charge delivered in each irradiation was monitored using a charge integrator coupled to the sample holder.

The irradiation line has a horizontal geometry, requiring the culture dishes to be positioned vertically, with the cell monolayer perpendicular to the beam direction. Since the beam has a 1 mm diameter, a distance of 97.44 mm was adopted between the beam exit and the cell monolayer to increase the irradiated area of the Petri dish and, consequently, the number of cells exposed to the beam, as shown in Fig. 2.

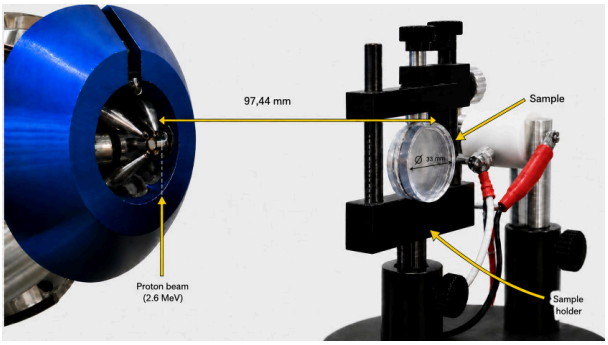


Fig. 2. Schematic representation of the irradiation experimental setup.

Given the limited range of 2.6 MeV protons in the culture medium, approximately 117.6 μm according to the PSTAR database (National Institute of Standards and Technology - NIST) [16], the culture medium was completely removed and the dishes were left uncovered during irradiation to avoid attenuating the beam before it reached the cell monolayer. After irradiation, 1 mL of culture medium was added, and the dishes were sealed with parafilm for transport.

2) Experimental design: Samples were divided into two experimental groups, control and irradiated, with all conditions evaluated in technical duplicate. The control group was subdivided into C0 and C1, while the irradiated group comprised different irradiation conditions, defined by the delivered proton beam charge.

The C0 control group remained in incubation throughout the entire experimental procedure, without undergoing transport or the irradiation steps, and was used as the biological reference control for comparing cell viability results.

The C1 control group underwent the same experimental conditions as the irradiated groups, including transport to LAMFI-USP, removal of the culture medium, and time spent in the experimental environment corresponding to the longest irradiation performed (4 minutes), but without exposure to the proton beam. This group was used exclusively to assess possible systematic effects arising from the experimental protocol itself, such as sample transport, handling during irradiation preparation, and the temporary absence of culture medium.

The irradiated group was subdivided into different exposure conditions, with charges ranging from 0.12 to 4.8 μC, obtained by combining beam current (2 to 20 nA) and irradiation time. In this study, a replicate was defined as a full irradiation session, using technical duplicate samples for each dose condition within that session. However, due to experimental uncertainties in the proton beam during irradiation, the effectively delivered integrated charge showed small variations between samples, including between technical duplicates. The integrated charge was therefore recorded individually for each sample.

C. Post-irradiation

1) Post-exposure cell handling and maintenance: After irradiation, samples were transported on ice over a 20-minute trip from LAMFI to FICSAE. They were then trypsinized

and transferred to a T25 flask with 5 mL of medium, and incubated under standard culture conditions for subsequent viability assessment. Two protocols were tested, one with a 6-day incubation period and another with 2 days, to account for late effects of radiation-induced damage [17], [18]. During incubation, cells were kept under standard culture conditions for the period corresponding to each protocol.

In the 6-day incubation protocol, samples were checked after 72 hours, if confluence exceeded 80%, they were transferred to T75 culture flasks or simply had their medium changed, to prevent excessive confluence from restricting cell proliferation and altering cell metabolism, which could interfere with the viability assessment. In the 2-day incubation protocol, due to the shorter culture period, cells remained in T25 flasks throughout.

2) Cell viability assessment: Cell viability was assessed using the colorimetric MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assay, which estimates the metabolic activity of viable cells through the formation of formazan crystals. For this, the culture medium was discarded and 2 mL of a solution containing 10% MTT solution – prepared by dissolving the salt in PBS at 5 mg/mL – and 90% culture medium was added. Samples were incubated for 3 hours to allow formazan crystal formation.

Colorimetric quantification was performed using a Kasvi K37-VIS spectrophotometer at a wavelength of 570 nm. To keep sample readings within the equipment's detection limit, the formazan solution was diluted in isopropyl alcohol, maintaining a final volume of 2 mL for all readings. The aliquot volume used for dilution was defined according to the intensity of sample coloration, so that more concentrated samples were subjected to higher dilution factors, aiming for absorbance readings of up to 1.5.

To correct for the dilution effect during data analysis, the obtained absorbance values were multiplied by the ratio between the final solution volume (2 mL) and the aliquot volume used for dilution. The corrected values were then normalized relative to the C0 control group, considered as 100% cell viability.

Given the variations in absorbed dose observed between technical duplicates, each sample was analyzed individually by evaluating the correlation between dose and the corresponding cell viability response for each irradiation experiment. Because of the small sample size and the lack of normality in the data, Spearman's correlation test was used, with the correlation coefficient (r) calculated independently for each irradiation (GraphPad Prism, version 11.0.2). The r coefficient was used to assess the strength of the association between variables, while the p-value was used to determine the statistical significance of that association, adopting a 5% significance level ($p < 0.05$).

Samples grouped by absorbed dose were then compared against C0 using the non-parametric Kruskal-Wallis test, followed by Dunn's post hoc teste for multiple comparisons. The analysis included four independent irradiation experiments ($N = 4$). The Kruskal-Wallis test was used to determine whether significant differences existed among the groups, while Dunn's test was subsequently applied to identify specific pairwise differences. A significance level of 5% ($p < 0.05$) was adopted. Given the limited number of independent experiments, the statistical analysis was interpreted within the context of an exploratory pilot study aimed at the initial validation of the proposed experimental protocol.

D. Absorbed dose estimation

The absorbed dose corresponding to each experimental condition was estimated through Monte Carlo simulation using the PHITS (Particle and Heavy Ion Transport code System) code, version 3.35. The irradiation geometry was faithfully reproduced, using the same experimental conditions described in Section II-B, including the irradiation line configuration, the composition and dimensions of the materials, the position of the cell monolayer relative to the beam, and the source proton energy.

The absorbed dose was estimated for a representative region of the cell monolayer using the T-Deposit estimator, which calculates the energy deposited per unit mass in the region of interest, in units of Gy per source particle. Since the integrated beam charge was monitored during each irradiation by the LAMFI-USP acquisition system, the total number of source particles for each condition was obtained by dividing the integrated charge by the elementary proton charge (1.602×10^{-19} C). The total absorbed dose in the cell monolayer, in Gy, was then calculated by multiplying the dose per source particle estimated by PHITS by the number of source particles derived from the integrated charge.

This approach made it possible to convert the experimentally measured electric charge into absorbed dose within the cell monolayer, allowing the biological assay results to be quantitatively correlated with the estimated absorbed dose.

III. Results and Discussion

A. Morphological assessment

Fig. 3 shows MG-63 cell images obtained with an EVOS M5000 microscope (Thermo Fisher Scientific), taken immediately after irradiation (40× magnification) and 48 hours after irradiation (10× magnification), for the C0 control group and the C1 control group.

The 40× images, taken after the experimental procedure, show that the C1 control group preserved the same spindle-shaped morphology and adhesion pattern observed in the C0 control group, with no evident structural alterations resulting from transport, handling, or the temporary absence of culture medium. Similarly, the 10× images, taken 48 hours later, show comparable confluence and cell distribution between the two groups.

Fig. 4 shows MG-63 cell images obtained with an EVOS M5000 microscope (Thermo Fisher Scientific), taken immediately after irradiation (40× magnification) and 48 hours after irradiation (10× magnification), for the C0 control group and two irradiated groups at different absorbed doses.

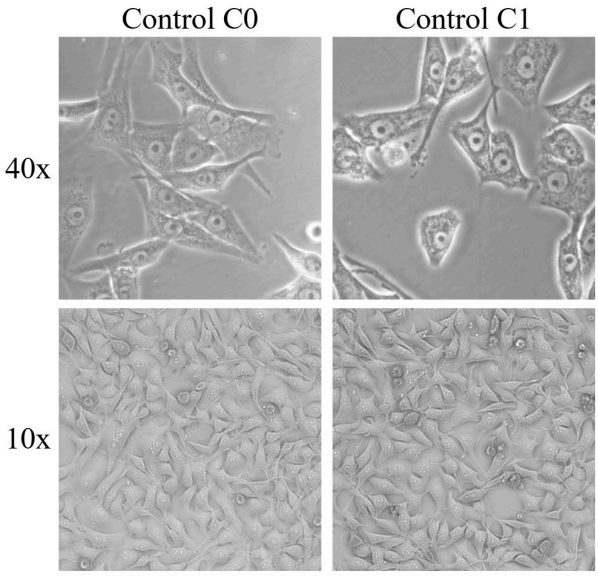


Fig. 3. Microscope images comparing the control groups, as indicated in the image. The top row shows samples immediately after irradiation at 40× magnification.The bottom row shows samples 48 hours after irradiation at 10× magnification; Columns represent the different groups: control C0 and control C1.

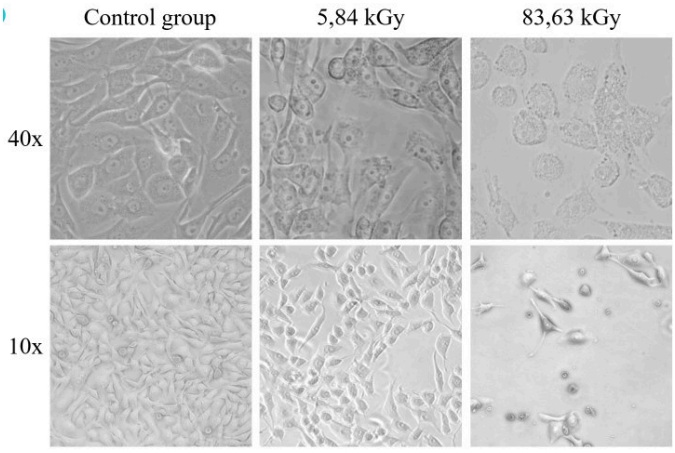


Fig. 4. Microscope images comparing the control group with the irradiated groups, as indicated in the image. The top row shows samples immediately after irradiation at 40× magnification.The bottom row shows samples 48 hours after irradiation at 10× magnification; Columns represent the different groups: control and different absorbed doses.

The 40× images, taken immediately after irradiation, reveal the initial morphological changes induced by proton beam exposure. The lower-dose group preserved the spindle-shaped morphology and adhesion pattern characteristic of the MG-63 line, with no evident structural changes compared to the control group, consistent with the absence of visible morphological alterations at this early stage. In contrast, the higher-dose group showed more retracted and rounded cells compared to the control, suggesting an acute cellular response to irradiation. These morphological changes precede the reduction in confluence observed after 48 hours, indicating that structural modifications may represent one of the earliest manifestations of the biological response to proton beam exposure.

The 10× images, taken 48 hours after irradiation, allow the effect of dose on cell confluence to be assessed. The control group shows high confluence and the characteristic spindle-shaped morphology of the MG-63 line, with cells well adhered and organized into a monolayer. This behavior confirms the stability of the culture conditions under incubation. In the lower-dose group (5.84 kGy), morphology similar to the control was observed, with no evident alterations, although confluence was slightly reduced. This pattern suggests that this dose induces damage insufficient to substantially compromise cell viability, allowing part of the cell population to survive and maintain partial confluence. In contrast, the higher-dose group (83.63 kGy) shows a more pronounced reduction in confluence progression, consistent with the increased cell death observed in this group.

The absorbed dose values obtained, on the order of kGy, are related to the extremely small mass of the irradiated cell monolayer. Since absorbed dose corresponds to energy deposited per unit mass, depositing energy in a region of extremely small mass results in numerically high dose values. This order of magnitude therefore does not, by itself, indicate excessive energy deposition in absolute terms, but rather reflects the normalization of that energy by the small cellular mass. For this reason, the dose values estimated in this experiment should not be directly compared to doses used in clinical practice, which are determined for macroscopic tissue volumes and under different irradiation conditions.

B. Biological response as a function of absorbed dose

The biological response was assessed using the MTT assay at two distinct time points after irradiation. Results obtained after six days of incubation are presented first, following the approach commonly adopted in the literature, followed by the results obtained two days after irradiation.

1) Cell viability 6 days after irradiation: The C1 control group showed a mean relative deviation in cell viability of $(26 \pm 21\%)$ compared to C0, which may be associated with the experimental conditions preceding irradiation, including transport and the period without culture medium. The relatively high standard deviation indicates considerable variability in the relative difference between C1 and C0 across the four independent experimental sessions.

Fig. 5 shows a scatter plot of MG-63 cell viability as a function of absorbed dose. Each point represents the mean of an irradiated experimental group, and the horizontal and vertical error bars correspond to the standard deviation of the absorbed dose and cell viability, respectively.

Fig. 5 shows that cell viability varied among the experimental groups across the different absorbed doses. Overall, higher doses were associated with lower viability compared to groups irradiated with lower doses. However, this reduction did not occur uniformly, and no clear dose-dependent trend was observed.

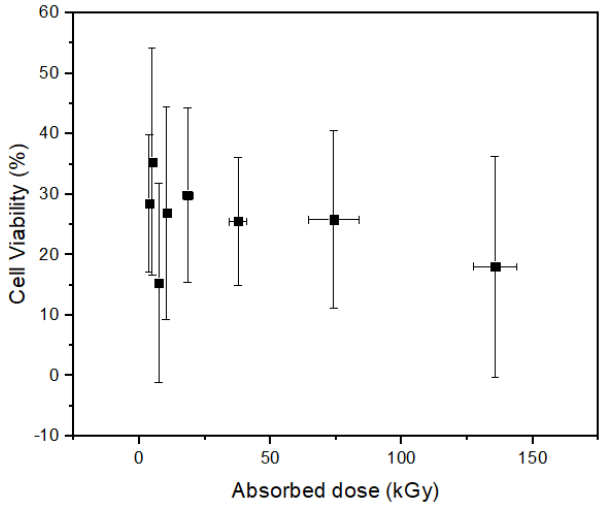


Fig. 5. MG-63 cell viability as a function of absorbed dose, determined by the MTT viability assay 6 days after irradiation. Data represent the mean viability of the experimental groups, normalized to the corresponding non-irradiated control group, set as 100% viability. Horizontal error bars indicate the standard deviation of the absorbed dose, and vertical error bars represent the standard deviation of cell viability.

The reproducibility of the absorbed dose was assessed by comparing technical duplicates irradiated during the same experimental session. Duplicates showed a mean relative difference in absorbed dose of $(7.0 \pm 0.6)\%$, representing the uncertainty associated with the irradiation procedure. This variability contributes to the dispersion observed among the experimental groups and should be taken into account when interpreting the biological results.

Spearman correlation analysis yielded the following values: irradiation 1 ($r = 0.16$; $p = 0.66$), irradiation 2 ($r = 0.22$; $p = 0.50$), irradiation 3 ($r = -0.42$; $p = 0.12$), and irradiation 4 ($r = 0.05$; $p = 0.89$). These results indicate no relationship between increasing absorbed dose and decreasing cell viability after 6 days of incubation.

The Kruskal-Wallis test indicated a significant difference among groups ($H = 31.55$; $p = 0.0001$). Dunn's test identified significant differences between the control and several groups: the intermediate-dose groups (63.38 kGy, $p = 0.0001$; 31.69 kGy, $p = 0.0006$; and 15.85 kGy, $p = 0.0177$), the higher-dose groups (253.53 kGy, $p = 0.0074$, and 126.77 kGy, $p = 0.0032$), and the lower-dose groups (12.67 kGy, $p = 0.0006$; 7.92 kGy, $p = 0.9601$; and 6.34 kGy, $p = 0.0609$). Despite these significant differences, no ordered relationship between increasing dose and decreasing viability was found. These results indicate that, with a 6-day incubation period, no direct relationship between beam exposure and biological effect could be observed.

The absence of correlation between absorbed dose and cell viability is related to the prolonged incubation period, which favors biological processes that modify the radiation response. Although irradiation initially induces cell cycle arrest, part of the surviving cells resume proliferation while another fraction enters radiation-induced senescence, remaining metabolically active despite losing proliferative capacity [19], [20]. Since the MTT assay reflects cellular metabolic activity, these processes can attenuate the relationship between absorbed dose and the biological response observed after a longer incubation period.

The 6-day interval between irradiation and the MTT assay was initially adopted based on studies described in the literature for evaluating radiobiological responses in cell cultures [21]. However, the results obtained did not show a consistent relationship between absorbed dose and cell viability. In addition, daily monitoring of the cell cultures revealed that morphological changes were visually more evident in the first days after irradiation. Accordingly, the incubation period before the viability assay was reduced to 2 days.

2) Cell viability 2 days after irradiation: The C1 control group showed a mean relative deviation in cell viability of $(22 \pm 2\%)$ compared to C0, which may be associated with transport and experimental handling. The lower standard deviation compared with the 6-day protocol indicates greater consistency in the relative difference between C1 and C0 across the four independent irradiation sessions using the 2-day incubation protocol.

Fig. 6 shows a scatter plot of MG-63 cell viability as a function of absorbed dose for a 2-day incubation period. Each point represents the mean of an irradiated experimental group, and the horizontal and vertical error bars correspond to the standard deviation of the absorbed dose and cell viability, respectively.

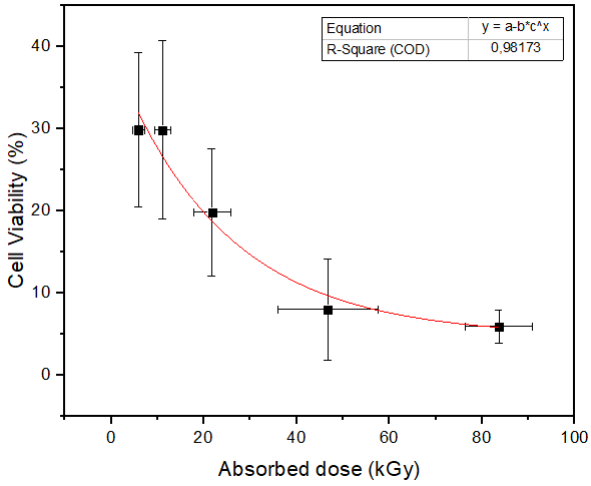


Fig. 6. MG-63 cell viability as a function of absorbed dose, determined by the MTT viability assay 2 days after irradiation. Data represent the mean viability of the experimental groups, normalized to the corresponding non-irradiated control group, set as 100% viability. Horizontal error bars indicate the standard deviation of the absorbed dose, and vertical error bars represent the standard deviation of cell viability.

Fig. 6 shows a progressive reduction in cell viability with increasing absorbed dose. Groups exposed to lower doses showed higher viability compared to those irradiated at higher doses. Unlike the 6-day results in Fig. 5, this dataset shows

a clearer, approximately exponential decline in viability as absorbed dose increases.

Spearman correlation analysis showed a strong, statistically significant negative correlation: irradiation 1 ($r = -0.87$; $p = 0.0022$), irradiation 2 ($r = -0.77$; $p = 0.0126$), irradiation 3 ($r = -0.89$; $p = 0.0011$), and irradiation 4 ($r = -0.89$; $p = 0.0011$), indicating a progressive and consistent reduction in cell viability with increasing absorbed dose.

The Kruskal-Wallis test indicated a significant difference among groups ($H = 38.66$; $p < 0.0001$). Dunn's post-test revealed a significant reduction in viability in the higher-dose groups relative to the C0 control: 126.77 kGy ($p < 0.0001$), 63.38 kGy ($p < 0.0001$), and 31.69 kGy ($p = 0.0064$). In contrast, the lower-dose groups, 15.85 kGy ($p = 0.27$) and 7.92 kGy ($p = 0.23$), showed a trend toward reduced viability without reaching statistical significance.

A clear trend of decreasing cell viability with increasing absorbed dose was observed. In addition, the mean relative difference between duplicates was $(4.2 \pm 1.8)\%$, indicating good experimental reproducibility. This behavior suggests that, in the first days after irradiation, the cellular response is dose-dependent, since the immediate effects of radiation on viability have not yet been influenced by cellular adaptation mechanisms, such as the resumption of proliferation by surviving cells. These results validate the choice of the 48-hour protocol as the most suitable for assessing the relationship between absorbed dose and biological response under the adopted experimental protocol.

In addition to the MTT assay, direct counting of viable cells was performed using trypan blue exclusion, to verify whether the trend observed by MTT was consistent with an independent viability assessment method. Fig. 7 shows the mean number of viable cells as a function of absorbed dose for each experimental group.

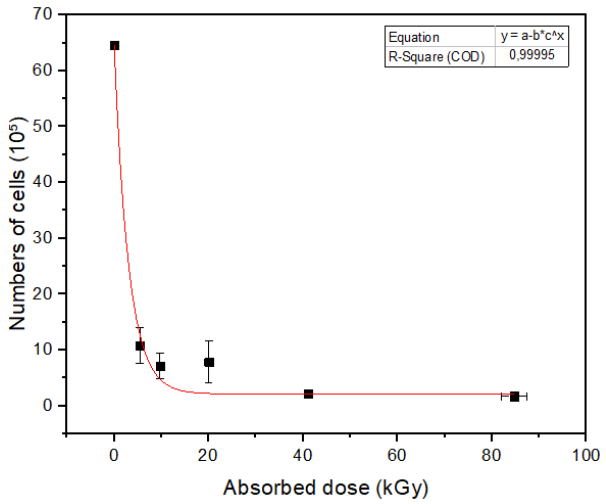


Fig. 7. Number of viable MG-63 cells as a function of absorbed dose, determined by trypan blue counting, 2 days after irradiation. Horizontal error bars indicate the standard deviation of the absorbed dose, and vertical error bars represent the standard deviation of the cell number.

Fig. 7 shows a reduction in the number of viable cells with increasing absorbed dose. This behavior is consistent with the results obtained by MTT 2 days after incubation, confirming agreement between the two methods used to assess the cellular response to irradiation.

Overall, these results show that the 2-day incubation protocol allowed for a consistent analysis of the relationship between dose and cell viability, yielding an appropriate dose-dependent biological response. These findings demonstrate that the proposed experimental protocol is capable of generating reproducible biological data under well-defined physical irradiation parameters. This established protocol therefore enables progression to the next stage of the pilot study, consisting of clonogenic assays and subsequent RBE determination.

IV. Conclusion

The results obtained support the initial validation of the experimental protocol established for irradiating cells at LAMFI-USP and its ability to produce consistent biological data for cell cultures irradiated with low-energy protons. Assessment of MG-63 cell viability showed that the 2-day MTT viability protocol is the most suitable for establishing the relationship between absorbed dose and biological response, showing a strong and consistent negative correlation across the independent experiments. In contrast, the 6-day protocol proved less suitable for this purpose, as no consistent relationship between absorbed dose and cell viability was observed under the evaluated conditions. Overall, these results show that, despite the inherent limitations of using low-energy proton beams at research accelerators, it is possible to establish an experimental platform capable of generating quantitative radiobiological data under well-characterized physical conditions. Considering the exploratory nature and the limited number of independent experiments, these findings represent an initial validation of the developed protocol. As a next step , additional independent experiments wiil be performed to strengthen its validation, and the characterization of the cellular response will be complemented by the clonogenic assay, a method that evaluates reproductive survival after radiation exposure. Accordingly, the methodology developed in this pilot study can be applied in future investigations integrating biological measurements, Monte Carlo simulations, and radiobiological models, enabling the application of mechanistic models for RBE estimation.

Acknowledgment

Lara Beatriz B. da Silva acknowledges the National Council for Scientific and Technological Development (CNPq) for the Scientific Initiation scholarship support. Process Number #168975/2025-4

The authors would like to thank the Hospital Israelita Albert Einstein for the financial support provided through the Seed Money Program.

This study was financed, in part by the São Paulo Research Foundation (FAPESP), Brazil. Process Number #2025/17052-0

REFERENCES

[1] S. Frisch and B. Timmermann, "The Evolving Role of Proton Beam Therapy for Sarcomas," *Clinical Oncology*, vol. 29, no. 8, pp. 500–506, Aug. 2017.

[2] M. Musielak, W. M. Suchorska, M. Fundowicz, P. Milecki, and J. Malicki, "Future Perspectives of Proton Therapy in Minimizing the Toxicity of Breast Cancer Radiotherapy," *Journal of Personalized Medicine*, vol. 11, no. 5, p. 410, May 2021.

[3] H. Paganetti, *Proton Therapy Physics*, 2nd ed. Boca Raton, FL: CRC Press, 2018.

[4] L. Vanderwaeren, R. Dok, K. Verstrepen, and S. Nuyts, "Clinical Progress in Proton Radiotherapy: Biological Unknowns," *Cancers*, vol. 13, no. 4, p. 604, Feb. 2021.

[5] R. Mohan, "A review of proton therapy – Current status and future directions," *Precision Radiation Oncology*, vol. 6, no. 2, Apr. 2022.

[6] C. B. Chen, C. Yang, T. Y. Chua, Z. Mi, M. Wee, M. Chen, *et al.*, "Facilitating focussed proton radiation of single cells by automated cell targeting and novel design of live cell chamber," *Nuclear Instruments and Methods in Physics Research Section B*, vol. 545, no. 165140, Dec. 2023.

[7] M. Pustovalova, R. Mohammad, Y. Wang, W. Xue, P. Malakhov, V. Nekrasov, *et al.*, "High-LET Proton Irradiation Significantly Alters the Clonogenic and Tumorigenic Potential of Human Breast Cancer Cell Lines In Vitro and In Vivo," *Frontiers in Bioscience-Landmark*, vol. 30, no. 4, Apr. 2025.

[8] A. McNamara, H. Willers, and H. Paganetti, "Modelling variable proton relative biological effectiveness for treatment planning," *The British Journal of Radiology*, vol. 93, no. 1107, p. 20190334, Mar. 2020.

[9] H. Paganetti, E. Blakely, A. Carabe-Fernandez, D. J. Carlson, I. J. Das, L. Dong, *et al.*, "Report of the AAPM TG-256 on the relative biological effectiveness of proton beams in radiation therapy," *Medical Physics*, vol. 46, no. 3, pp. e53–e78, Feb. 2019.

[10] S. J. McMahon, A. L. McNamara, J. Schuemann, H. Paganetti, and K. M. Prise, "A general mechanistic model enables predictions of the biological effectiveness of different qualities of radiation," *Scientific Reports*, vol. 7, no. 1, Sep. 2017.

[11] A. L. McNamara, J. Schuemann, and H. Paganetti, "A phenomenological relative biological effectiveness (RBE) model for proton therapy based on all published in vitro cell survival data," *Physics in Medicine and Biology*, vol. 60, no. 21, pp. 8399–8416, Oct. 2015.

[12] Y. Tao, H. Q. Tan, Z. Mi, C. B. Chen, V. Y. M. Lam, T. Osipowicz, *et al.*, "The radiobiology beam line facility at the Centre for Ion Beam Applications, National University of Singapore," *Nuclear Instruments and Methods in Physics Research Section B*, vol. 456, pp. 26–31, Oct. 2019.

[13] M. Straticiuc, M. Bacalum, C. Mircea Rusu, R. Andrei, I. Burducea, I. Cenusa, *et al.*, "New setup for basic radiobiology studies using a 3 MV Tandetron: Design and developments," *Nuclear Instruments and Methods in Physics Research Section B*, vol. 528, pp. 45–53, Oct. 2022.

[14] V. Bravatà, F. P. Cammarata, L. Minafra, P. Pisciotta, C. Scazzone, L. Manti, *et al.*, "Proton-irradiated breast cells: molecular points of view," *Journal of Radiation Research*, vol. 60, no. 4, pp. 451–465, May 2019.

[15] M. Yagi, Y. Takahashi, K. Minami, T. Matsuura, J. M. Nam, Y. Onodera, *et al.*, "A Consistent Protocol Reveals a Large Heterogeneity in the Biological Effectiveness of Proton and Carbon-Ion Beams for Various Sarcoma and Normal-Tissue-Derived Cell Lines," *Cancers*, vol. 14, no. 8, p. 2009, Apr. 2022.

[16] National Institute of Standards and Technology (NIST), "PSTAR: Stopping Power and Range Tables for Protons," Physical Measurement Laboratory. [Online]. Available: https://physics.nist.gov/PhysRefData/Star/Text/PSTAR.html

[17] E. Firat, S. Gaedicke, C. Tsurumi, N. Esser, A. Weyerbrock, and G. Niedermann, "Delayed cell death associated with mitotic catastrophe in γ-irradiated stem-like glioma cells," *Radiation Oncology*, vol. 6, no. 1, Jun. 2011.

[18] T. Bernardo, L. Heuchel, F. Heinzelmann, J. Esser, L. Lüdemann, B. Timmermann, *et al.*, "Linear energy transfer dependent variation in viability and proliferation along the Bragg peak curve in sarcoma and normal tissue cells," *Physics in Medicine & Biology*, vol. 69, no. 19, p. 195005, Sep. 2024.

[19] B. Baselet, N. Belmans, E. Coninx, D. Lowe, A. Janssen, A. Michaux, *et al.*, "Functional Gene Analysis Reveals Cell Cycle Changes and Inflammation in Endothelial Cells Irradiated with a Single X-ray Dose," *Frontiers in Pharmacology*, vol. 8, Apr. 2017.

[20] R. Mirzayans, B. Andrais, and D. Murray, "Impact of Premature Senescence on Radiosensitivity Measured by High Throughput Cell-Based Assays," *International Journal of Molecular Sciences*, vol. 18, no. 7, p. 1460, Jul. 2017.

[21] O. Keta *et al.*, "Radiosensitivity of human ovarian carcinoma and melanoma cells to γ-rays and protons," *Archives of Medical Science*, vol. 3, pp. 578–586, 2014.

Development of a *in vitro* Protocol for Tumor Cell irradiation with Low-Energy Protons

Manuela S. de Souza
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein (FICSAE)
Hospital Israelita Albert Einstein
São Paulo, Brazil
manuela.dsouza@einstein.edu.br

Cecília L. Prioste
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein (FICSAE)
Hospital Israelita Albert Einstein
São Paulo, Brazil
cecilia.lprioste@einstein.edu.br

Lara Beatriz B. da Silva
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein (FICSAE)
Hospital Israelita Albert Einstein
São Paulo, Brazil
lara.ds@einstein.edu.br

Maira S. Cucielo
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein (FICSAE)
Hospital Israelita Albert Einstein
São Paulo, Brazil
maira.cucielo@einstein.br

Welbert O. Pereira
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein (FICSAE)
Hospital Israelita Albert Einstein
São Paulo, Brazil
welbert.pereira@einstein.br

Paula Cristina G. Antunes
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein (FICSAE)
Hospital Israelita Albert Einstein
São Paulo, Brazil
paula.antunes@einstein.br

***Abstract*—This preliminary study aims to assess the feasibility of an experimental procedure that enables the adaptation of cell culture techniques, using different tumor cell lines, for irradiation with a low-energy proton beam originally intended for nuclear physics research. Considering the physical constraints imposed by the beam and the experimental setup, different conditions were investigated to determine whether cell cultures could be maintained under the procedure required for irradiation and to assess their influence on cell viability independently of radiation exposure. The results provide an initial methodological basis for establishing a reproducible pre-irradiation protocol, enabling subsequent radiobiological experiments with the proton beam and supporting future microdosimetric investigations**

***Index Terms*—proton therapy, cell culture, low-energy proton beam, radiobiology**

I. Introduction

Proton therapy has become clinically consolidated as a high-precision alternative to conventional radiotherapy, owing to its ability to concentrate the therapeutic dose within the target volume while sparing surrounding healthy structures [1]. Its rise is sustained by an expanding global infrastructure, which currently exceeds the milestone of 100 active clinical centers worldwide, according to data from the Particle Therapy Co-Operative Group (PTCOG) [2].

The main advantage of this technique lies in the proton dose deposition profile, characterized by the Bragg peak, a region in which protons deposit most of their energy abruptly, sparing adjacent tissues. However, from a biological standpoint, important gaps persist, particularly regarding parameters such as Linear Energy Transfer (LET), a physical quantity that indicates the average rate and spatial density at which the particle dissipates its kinetic energy along its microscopic path, and Relative Biological Effectiveness (RBE), which expresses the dose ratio required for protons and a reference radiation to produce the same level of cellular damage [3] [4].

Clinical practice adopts a constant RBE of 1.1, comparing the efficacy of proton therapy relative to the use of photons. However, current studies show that these factors vary depending on tissue type, depth, and dose, and can influence both tumor response and adverse effects [5] [6]. Thus, variations in the response of different cell lines to the treatment, such as the mechanisms of response and DNA damage induced by protons, are not considered [7] [8]. However, the actual biological variations resulting from the increase in LET, especially in the distal region of the Bragg peak, present critical gaps that demand investigation [9]. In this scenario, *in vitro* assays constitute the primary tool for generating controlled experimental data, providing the necessary basis to validate biological models and refine microdosimetric approaches [9] [10].

One of the main obstacle to expanding these assays is the limited access to suitable irradiation infrastructure. The majority of *in vitro* studies are conducted in clinical proton therapy centers or in facilities dedicated exclusively to radiobiology [11]. Under these conditions, the experimental geometry is already compatible with cell culture parameters, and the beam possesses high energy, sufficient to penetrate isolation windows, physical supports, and culture medium layers. However, access to this type of facility remains restricted and limited worldwide [2], particularly in Brazil, compromising the advancement of exploratory and pre-clinical studies.

In order to mitigate this restriction, linear accelerators designed for basic research in nuclear physics emerge as a

promising alternative to expand the capacity for investigating fundamental interactions between protons and biological systems [1]. On the other hand, because they were not designed for biomedical applications, the use of these facilities imposes methodological challenges, such as the low energy of the beam. The reduced range of these particles requires the adaptation of biological cell culture conditions to ensure that the beam reaches the cellular target [12]. However, these adaptations may themselves introduce cellular stress and affect cell viability independently of radiation exposure. Therefore, before investigating proton-induced biological effects, it is necessary to determine whether cells remain viable under the experimental conditions required for irradiation.

In this context, the proton accelerator at the Laboratory of Materials Analysis with Ion Beams (LAMFI) of the University of São Paulo (USP) represents a potential platform for this type of investigation [13]. Although the system exhibits excellent beam stability and precision, its maximum energy of approximately 5.1 MeV imposes significant constraints on the biological preparation of samples; under these conditions, conventional cell culture protocols must be adjusted to enable irradiation.

Against this backdrop, this preliminary study aimed to evaluate the feasibility of adapting adherent cell cultures for irradiation with a low-energy proton beam from an accelerator originally designed for nuclear physics research. To this end, the main experimental factors capable of influencing cell viability independently of radiation were systematically investigated, including exposure time without culture medium, cell concentration, homogeneous cell seeding, and the conditions required to allow direct beam incidence on the cells. The results are intended to provide an initial methodological basis for establishing a reproducible pre-irradiation procedure, minimizing biological effects associated with the adaptations required for subsequent proton irradiation experiments.

II. Materials and Methods

A. Experimental platform

The development of the experimental methodology was based on the characteristics of the proton beam available at USP's LAMFI. This facility features a Pelletron-tandem electrostatic accelerator, model 5SDH, originally designed for research in Nuclear Physics and Materials Science, which is capable of accelerating different ionic species with energies up to 5.1 MeV [13]. The system delivers beams with high operational stability, excellent reproducibility, and rigorous control of irradiation parameters, characteristics that make it suitable for experiments demanding high precision in dose delivery. Beam energy control is performed by a closed-loop electronic feedback system that is periodically calibrated, while fluence is monitored by a charge integrator connected to the sample holder, allowing control of the delivered charge in each irradiation. For this study, an energy of 2.6 MeV was used, corresponding to the operating range that exhibits the greatest stability of the accelerator.

LAMFI's irradiation beamline has a horizontal configuration, such that the beam impinges laterally on the experimental region. This characteristic imposes constraints on the positioning of cell cultures and constitutes one of the challenges in developing the experimental platform employed in this study. To ensure that the cell growth surface remained parallel to the beam incidence direction, the culture plates were positioned vertically in the sample holder, with their bottom perpendicular to the ground, as shown in Fig. 1. Due to these conditions, the cultures were necessarily performed using adherent cells in monolayers.

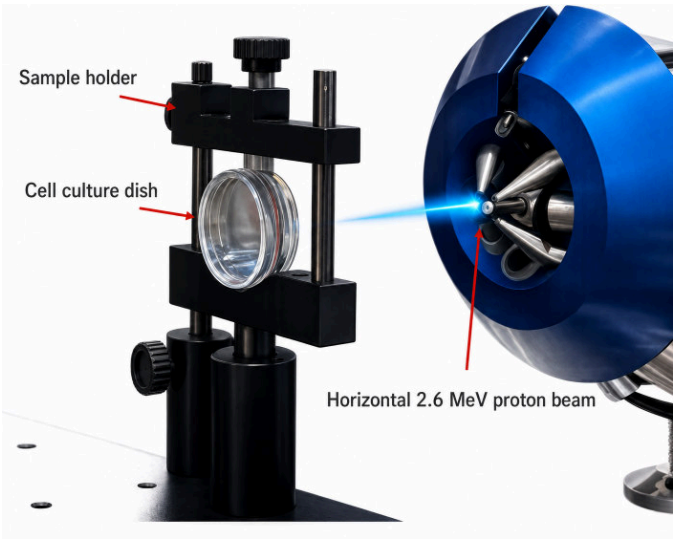


Fig. 1. Experimental geometry adopted in the LAMFI irradiation beamline. The horizontal incidence of the proton beam requires the vertical positioning of the culture plate in the sample holder during irradiation.

The limited penetration depth of low-energy protons represented a second challenge for the development of the experimental methodology. At 2.6 MeV, the range of protons in water, and consequently in culture medium,is approximately 117.6 μm, according to data from the PSTAR database (National Institute of Standards and Technology – NIST) [14].

The depth-dose curve presented in Fig. 2A demonstrates that the Bragg peak occurs very close to the surface of the irradiated medium. Under conventional culture conditions, cells remain adhered to the bottom of the culture plate, covered by a layer of culture medium with a thickness on the order of millimeters (Fig. 2B). However, at the available energy, the proton range is insufficient to penetrate the medium and reach the cells adhered to the plate, as illustrated in Fig. 2C. In order for the Bragg peak to reach the cell surface, complete removal of the culture medium was required during the irradiation period.

B. Cell culture

All biological experiments were conducted at the Cell and Tissue Culture Laboratory of the Albert Einstein Israelite College of Health Sciences (LCCT-FICSAE) under standardized culture conditions. Based on the imposed limitations, the

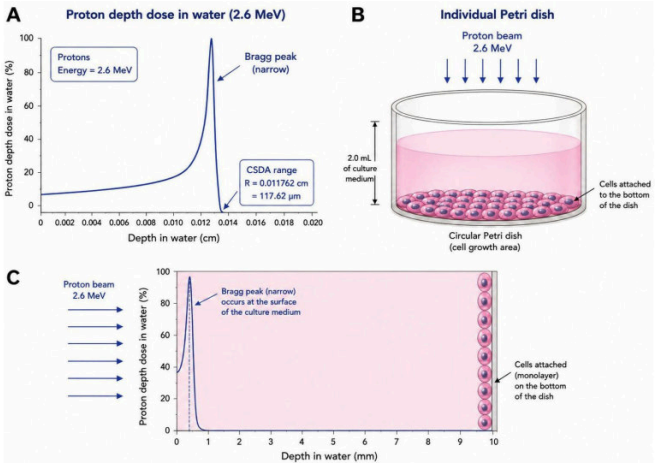


Fig. 2. Physical justification for the development of the experimental methodology. (A) Depth-dose curve (DDC) in water for 2.6 MeV protons, highlighting the Bragg peak. (B) Conventional configuration of a cell culture plate, in which cells remain adhered to the bottom of the plate and covered by a layer of medium. (C) Illustration of the culture plate in the irradiation position, with the bottom positioned vertically, parallel to the beam incidence direction. Inset: the position of the Bragg peak relative to the cells, demonstrating that the proton range is insufficient to penetrate the culture medium and reach the cells.

chosen experimental methodology was the culture of adherent monolayers, using the human osteosarcoma cell line MG-63, and subsequently validated with the MDA-MB-231 line (human breast adenocarcinoma) to evaluate its reproducibility across different cellular models.

Cells were cultured in T75 flasks, maintained at 37 °C in a humidified atmosphere containing 5% CO_2, with culture medium replenishment every two days. Upon reaching high confluence, the cultures were washed with phosphate-buffered saline (PBS), trypsinized, and prepared for the experiments according to the protocols recommended by the American Type Culture Collection [15] [16]. The specific culture conditions for each cell line are presented in Table I.

For all assays described herein, a standard protocol was established to plate cells between 20 and 24 hours before irradiation, an interval defined based on the time required for cell adhesion and the operational logistics of the laboratories. All described samples were seeded in culture petri dishes with a diameter of approximately 33 mm, using models K13-0035 and K12-006 from KASVI.

TABLE I
CULTURE CONDITIONS USED IN THIS STUDY

Cell line	Origin	Culture medium	Supplements
MG-63	Human osteosarcoma	DMEM Low Glucose	10% FBS + 1% Penicillin-Streptomycin
MDA-MB-231	Human breast adenocarcinoma	DMEM Low Glucose	10% FBS + 1% Penicillin-Streptomycin + 0.1 mM NEAA

C. Employed methodology

The physical limitations described in Section II-A required the development of a specific experimental protocol for culturing adherent cells. To this end, the maximum exposure time of cells without culture medium, the cell concentration required to obtain a uniform monolayer, and the homogeneity of cell distribution in the irradiation area were systematically evaluated. The procedures adopted for each stage are described below.

1) Evaluation of exposure time without culture medium: To investigate the feasibility of exposing cells without culture medium to a non-sterile environment and the impact of exposure time on their viability, an experimental protocol was developed and divided into three sequential stages, as illustrated in Fig. 3.

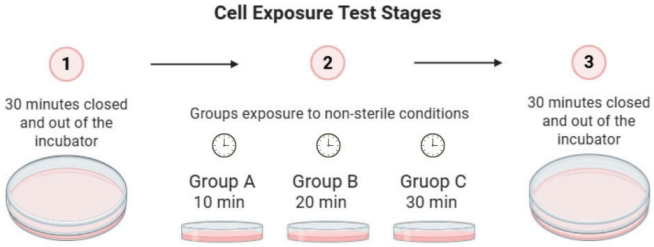


Fig. 3. Workflow and timeline of the experimental stages evaluated during the cell exposure test.Group A was exposed for 10 minutes, group B, for 20 minutes and group C, for 30 minutes.

Cells previously plated with 1.5 mL of culture medium were sealed and kept outside the incubator for 30 minutes, simulating the transport time between laboratories. Subsequently, the culture medium was completely removed, and the adhered cells remained exposed to the environment for different exposure times: Group A (10 minutes), Group B (20 minutes) and Group C (30 minutes). The selected times were longer than those typically required to perform the irradiations at LAMFI, thus representing a conservative condition for cell viability assessment.

At the end of the exposure period, 1 mL of culture medium was added to each well. The plates were sealed again and remained outside the incubator for an additional 30 minutes, reproducing the return transit to the LCCT laboratory after irradiation at LAMFI. In parallel, a control group was maintained, in which cells remained with 1.5 mL of culture medium in the incubator under optimal conditions of temperature, humidified atmosphere, and CO_2 supply.

Following 48 hours of incubation, the MTT assay was performed, which is based on the metabolic capacity of viable cells to reduce the tetrazolium salt (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) by the action of mitochondrial enzymes, forming insoluble formazan crystals [17]. The amount of formazan formed is proportional to the number

of metabolically active cells, allowing the estimation of cell viability compared to the control group and the quantification of the impact of the stress conditions imposed by the experimental protocol. Following the incubation period with the reagent, the formed crystals were solubilized in isopropyl alcohol, and the absorbance of the solutions was determined using a Kasvi K37-VIS spectrophotometer.

2) Optimization of cell concentration: To investigate the cell concentration required to achieve approximately 80% confluence, that is, 80% of the plate area occupied by cells after the incubation period preceding irradiation, concentrations ranging from 1.4×10^5 cells/well to 1.0×10^6 cells/well were evaluated.

Sample preparation consisted of trypsinizing T75 flasks, followed by cell counting. The volume of cell suspension to be added to each well (in μL) was calculated using Equation (1), utilizing the mean concentration obtained from the cell count, followed by the addition of culture medium to a final volume of 1.5 mL.

$$V_{\text{cells}} = \left(\frac{\text{Desired concentration}}{\text{Mean count}} \right) \times 1000 \qquad (1)$$

Following approximately 24 hours of incubation, the cultures were observed under an optical microscope to verify the level of confluence of the monolayer across the culture surface. Optimal confluence should represent a continuous monolayer, maximizing the utilization of the irradiated area without cellular overlap. Analysis allowed the determination of the optimal seeding concentration and the evaluation of the efficiency of the employed homogenization strategies.

3) Monolayer homogeneity evaluation: To evaluate cell homogeneity after seeding, two distinct homogenization techniques were tested. In the conventional protocol, the volume of cell suspension was dispensed directly into the well of the culture plate. Subsequently, culture medium was added to reach a final volume of 1.5 mL, without performing any prior homogenization step. After seeding, gentle manual shaking of the plate was performed with the objective of promoting cell spreading over the entire culture surface, as illustrated in Fig. 4A.

In the adapted protocol, a pre-homogenization step for the cell suspension was incorporated. In this method, the calculated volume of the cell suspension and the volume of culture medium required to reach 1.5 mL were combined in a conical tube. The mixture was homogenized by mechanical resuspension (pipetting up and down) and by mechanical vortex agitation for 2 minutes at a constant speed of 250 rpm, as shown in Fig. 4B.

III. Results and Discussion

A. Maximum exposure time without culture medium

Cell viability was initially evaluated to verify whether the temporary removal of the culture medium could be incorporated into the experimental protocol without compromising cell integrity. Fig. 5 presents the means and standard deviations

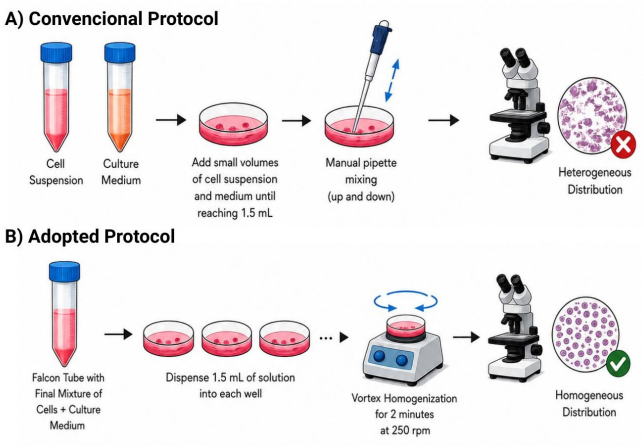


Fig. 4. Comparison between the conventional (A) and adapted (B) cell homogenization methods.

of the absorbance values for each group, which were normalized by the control group. Technical replicates (duplicates) were condensed into their respective means prior to the inter-assay analysis. The final mean and variability of each group were calculated strictly based on the number of independent experiments (N=4).

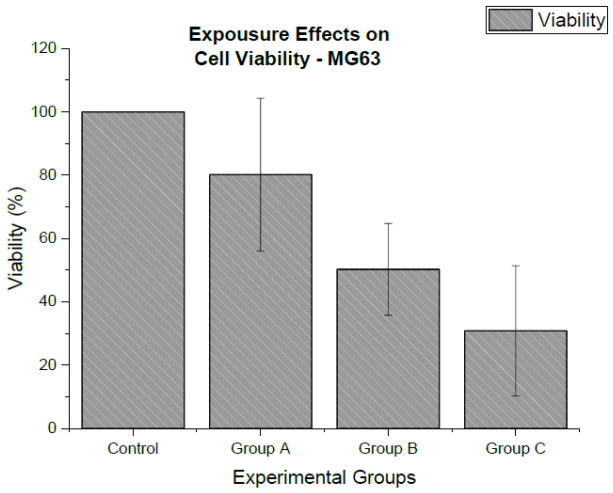


Fig. 5. MTT assay results for MG-63 cell groups A, B, and C previously defined. Data are presented as mean and standard deviation (SD) from two independent experiments, each performed in technical duplicate (n = 4 independent experiments × 2 technical replicates).

Fig. 5 presents the results of the MTT assay performed on the MG-63 cell line. As expected, cell viability decreased with increasing exposure time. For the statistical analysis, two non-parametric tests were performed. First, the Kruskal–Wallis test was used to evaluate the null hypothesis that no differences existed in cellular response among the different exposure times and the control group. Subsequently, Dunn's multiple comparisons test was applied as a post hoc analysis to assess the hypothesis that each exposure time (10, 20, and 30

minutes) would individually exhibit a statistically significant difference in response compared with the control group.

The Kruskal–Wallis test yielded a p-value of 0.0013, indicating a statistically significant difference among the groups. Graphical analysis and Dunn's post hoc test indicated that that cell viability was maintained at approximately control levels for exposure times of up to 10 minutes, with a viability of $(80.18 \pm 24.12)\%$ and the highest p-value in Dunn's multiple comparisons test ($p > 0.999$). In contrast, longer exposure times had a greater impact on cell viability. As shown in the graph, Group B (20-minute exposure) exhibited approximately a 50% reduction in viability , reaching $(50.25 \pm 14.49)\%$, whereas Group C (30-minute exposure) showed a viability of only $(30.18 \pm 20.58)\%$. When compared with the control group, Dunn's test yielded p-values of 0.0744 and 0.0083 for Groups B and C, respectively.

Given the promising results obtained with the MG-63 cell line, the same experimental protocol was replicated using MDA-MB-231 cells, as shown in Fig. 6.

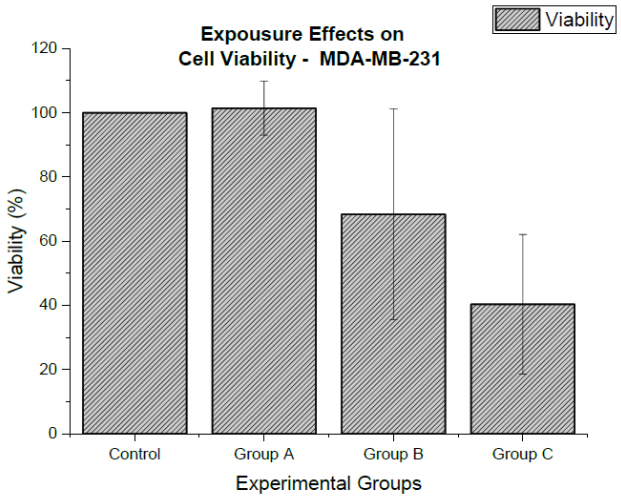


Fig. 6. MTT assay results for Groups A, B, and C of the MDA-MB-231 cell line. Data are presented as mean and standard deviation (SD) from two independent experiments, each performed in technical duplicate (n = 4 independent experiments × 2 technical replicates).

Fig. 6 presents the results obtained for the MDA-MB-231 cell line. The Kruskal–Wallis test indicated a statistically significant differences among the groups ($p = 0.0014$). To further investigate these findings, Dunn's multiple comparisons test was also performed. Group A showed the greatest similarity to the control group, with a cell viability of $(101.40 \pm 8.38)\%$ and a p-value greater than 0.9999 in Dunn's test. Graphical analysis again revealed a decline in cell viability for Groups B and C, reaching $(68.34 \pm 32.86)\%$ and $(40.32 \pm 21.74)\%$, respectively. However, the post hoc analysis yielded p-values of 0.3485 for Group B and 0.0407 for Group C, indicating that neither group differed significantly from the control.

Statistical analysis showed no significant difference between Group A and the control group for either cell line in Dunn's post hoc test ($p > 0.9999$ for both the MG-63 and MDA-MB-231 cell lines). These findings suggests that exposure for up to 10 minutes under the evaluated experimental conditions did not significantly affect cell viability. To provide an additional margin for subsequent irradiation experiments, a safety factor of 2 (Fs = 2) was applied to this exposure time, establishing 5 minutes as the maximum exposure time adopted during proton irradiation.

B. Cell seeding concentration

The different cell concentrations evaluated for the MG-63 cell line resulted in distinct patterns of culture surface coverage. Following a preliminary screening, lower cell concentrations were found to produce areas with incomplete cellular coverage, whereas higher concentrations promoted cell overlap.

Fig. 7 presents representative microscopic images of cultures seeded at concentrations of 8.0×10^5, 9.0×10^5, and 1.0×10^6 cells/well after 24 hours of incubation. The images were acquired using the Cell AutoCount tool of the EVOS M5000 microscope (Thermo Fisher Scientific) at 4× magnification.

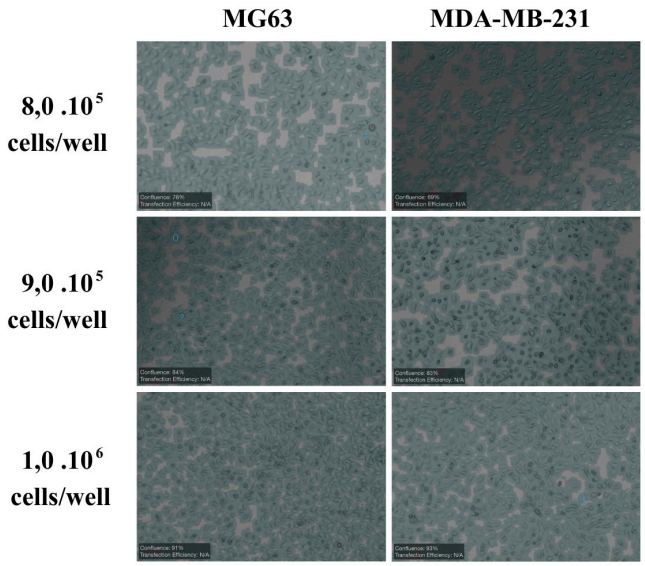


Fig. 7. Relationship between cell seeding concentration and culture surface coverage for the different cell lines, by the AutoCout confluence result. Representative images were acquired 24 hours after cell seeding.

In Fig. 7, the first column displays the different cell seeding concentrations for the MG-63 cell line, while the second column presents the corresponding concentrations for the MDA-MB-231. For both cell lines, the concentration of 8.0×10^5 cells/well did not achieve the target confluence, (MG63=78%, MDA-MB-231=69%) as evidenced by the presence of large cell-free areas. In contrast, the concentration of 1.0×10^6 cells/well exceeded the desired confluence (MG63=91%, MDA-MB-231=93%). This effect was particularly evident in the MG-63, where cell overlap and limited intercellular space indicated excessive cell density.

Finally, a cell seeding concentration of 9.0×10^5 cells/well produced a uniform cell monolayer with minimal cell overlap while maximizing the use of the available culture surface, achieving 84% for MG63 cells and 83% for MDA-MB-231. Therefore, this concentration was considered optimal and was adopted for all subsequent experiments.

C. Cell monolayer homogeneity

Fig. 8 illustrates the difference in monolayer homogeneity between the conventional cell homogenization protocol and the optimized protocol developed in this study, 24 hours after MG-63 cells seeding.

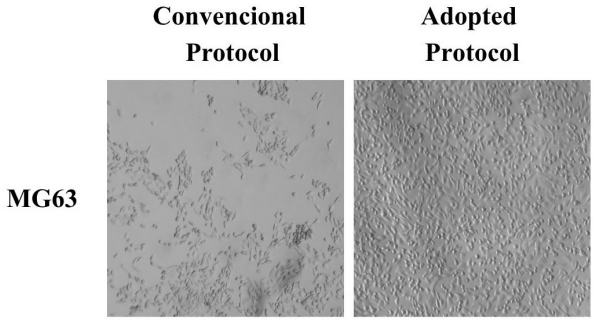


Fig. 8. Microscopic analysis of the impact of cell homogenization techniques on the distribution of MG-63 cells.

The optimized protocol produced a homogeneous cell monolayer throughout the irradiation area, minimizing local variations in cell density that could influence the biological response to the proton beam. In contrast, the conventional protocol resulted in cell accumulation at the center of the well, followed by a gradual decrease in cell density toward the periphery. These findings demonstrate that the proposed modifications were effective in achieving a homogeneous monolayer. Moreover, the optimized protocol improved the reproducibility of the experimental procedure while reducing the potential for human error.

The establishment of a homogeneous cell monolayer is a fundamental requirement for irradiation experiments using low-energy proton beams. A homogeneous monolayer ensures that all cells are located at approximately the same distance from the incident beam and receive a comparable radiation dose, thereby minimizing geometric variations in energy deposition.

D. Final experimental protocol

Table. II presents the final flowchart of the protocol developed throughout this preliminary study. The experimental design was conceived to establish compatibility between cell culture and the available proton beam by selecting adherent cell lines and using culture wells with a diameter of 33 mm. To adapt the experimental conditions to the irradiation geometry, the proposed protocol defines the cell seeding density (9.0×10^5 cells/well), the volume of culture medium (1.5 mL/well), the cell homogenization procedure during seeding, and the incubation time and conditions. Finally, the results showed that, under the evaluated conditions, the adaptations required by the physical constraints of the proton beam did not significantly compromise cell viability, supporting the feasibility of the proposed pre-irradiation procedure.

TABLE II
FINAL IRRADIATION CONDITIONS

Culture Conditions	Adherent Cells in 33 mm diameter plate
Seeding Density	9.0×10^5 cells/well
Monolayer Homogeneity	Use the adapted protocol, available in Section II-C-2
Expousure Time	5 minutes Without Culture Medium

IV. CONCLUSION

This study presented a preliminary evaluation of an experimental methodology as the first stage of adapting cell cultures to irradiation with a low-energy proton beam generated by an accelerator primarily intended for nuclear physics research.The proposed protocol was designed to address the physical limitations of the beam, including its low energy and spatial configuration.

Adherent cell culture was selected as the most suitable approach to satisfy the physical constraints imposed by the irradiation system. Based on this strategy, experimental procedures were evaluated to define an appropriate cell seeding density, obtain homogeneous cell monolayers within the region reached by the proton beam, and assess the effects of the experimental conditions required for irradiation on cell viability, resulting in an initial protocol for sample preparation.

The results indicate that cell culture, transport and exposure procedures can be performed without significantly compromising cell viability under the evaluated conditions. These findings support the feasibility of the proposed pre-irradiation procedure and provide an initial methodological basis for subsequent radiobiological studies using low-energy protons.

This preliminary assessment represents the first stage of a broader research project aimed at determining the absorbed dose at the cellular level, performing the microdosimetric characterization of the proton beam, and ultimately estimating the RBE in different tumor cell lines. As the next steps, a parametric analysis of the microscopic results presented in Sections III-B and III-C will be performed, followed by pilot irradiations using the LAMFI proton beam according to the proposed protocol. These experiments will evaluate the biological response under the established irradiation conditions and provide the experimental basis for subsequent quantitative radiobiological investigations.

ACKNOWLEDGMENTS

Manuela S.de Souza acknowledges the Brazilian Association of Projects Management (ABPG) fot the Scientific Initiation scholarship support.

The authors would like to thank the Hospital Israelita Albert Einstein for the financial support provided through the Seed Money Program.

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Process Number #2025/17052-0.

REFERENCES

[1] T. Bortfeld and S. Yan, "Our journeys through the fascinating world of proton radiation therapy," Medical Physics, vol. 50, no. S1, pp. 27–34, Jun. 2023, doi: 10.1002/mp.16118. Available: https://pubmed.ncbi.nlm.nih.gov/36502491/

[2] J. S. Kim, "Global development of particle therapy centers," Particle Therapy Co-Operative Group (PTCOG) Newsletter, Mar. 6, 2026. [Online]. Available: https://ptcog.online/global-development-of-particle-therapy-centers/

[3] C. Hahn et al., "Towards harmonizing clinical linear energy transfer (LET) reporting in proton radiotherapy: a European multi-centric study," Acta Oncologica, vol. 61, no. 2, pp. 206–214, Feb. 2022, doi: 10.1080/0284186X.2021.1992007.

[4] M. Joiner and A. van der Kogel, Eds., Basic Clinical Radiobiology, 4th ed. London, UK: Hodder Arnold, 2009. [Online]. Available: https://phyusdb.wordpress.com/wp-content/uploads/2013/03/basic-clinical-radiobiology.pdf

[5] H. Thomas and B. Timmermann, "Pediatric proton therapy," The British Journal of Radiology, vol. 93, no. 1107, p. 20180601, Mar. 2020, doi: 10.1259/bjr.20180601. Available: https://pubmed.ncbi.nlm.nih.gov/31529979/

[6] M. Mizumoto et al., "Proton beam therapy for children, adolescents and young adults (AYAs): JASTRO and JSPHO guidelines," Cancer Treatment Reviews, vol. 98, p. 102209, Aug. 2021, doi: 10.1016/j.ctrv.2021.102209. Available: https://pubmed.ncbi.nlm.nih.gov/33984606/

[7] H. Paganetti et al., "Relative biological effectiveness (RBE) values for proton beam therapy," International Journal of Radiation Oncology Biology Physics, vol. 53, no. 2, pp. 407–421, Jun. 2002, doi: 10.1016/s0360-3016(02)02754-2. Available: https://pubmed.ncbi.nlm.nih.gov/12023146/

[8] A. Lühr et al., "Recommendations for reporting and evaluating proton therapy beyond dose and constant relative biological effectiveness," Technical Reports & Consensus Guidelines, 2024, PMCID: PMC11750264. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11750264/

[9] H. Paganetti, "Significance and application of RBE variations in proton beam therapy," Technology in Cancer Research & Treatment, vol. 2, no. 5, pp. 413–426, Oct. 2003, doi: 10.1177/153303460300200506. Available: https://pubmed.ncbi.nlm.nih.gov/14529306/

[10] H. H. Rossi and M. Zaider, "Elements of microdosimetry," Medical Physics, vol. 18, no. 6, pp. 1085–1092, Nov. 1991, doi: 10.1118/1.596616. Available: https://pubmed.ncbi.nlm.nih.gov/1753880/

[11] Y. Tao et al., "The radiobiology beam line at the Centre for Ion Beam Applications of the National University of Singapore," Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, vol. 461, pp. 241–245, Dec. 2019, doi: 10.1016/j.nimb.2019.09.043. Available: https://www.sciencedirect.com/science/article/abs/pii/S0168583X19304707

[12] H. Yoriyaz, I. S. Branco, I. P. Almeida, and G. P. Fonseca, "Fundamentos de Transporte e Cálculo de Dose em Tratamentos com Feixes de Prótons," Revista Brasileira de Física Médica, vol. 13, no. 1, pp. 109–115, Set. 2019, doi: 10.29384/rbfm.2019.v13.n1.p109-115. Available: https://www.rbfm.org.br/rbfm/article/view/510

[13] Laboratório de Análises de Materiais por Feixes Iônicos (LAMFI), "Acelerador Pelletron-tandem modelo 5SDH," Instituto de Física, Universidade de São Paulo (USP). [Online]. Available: https://portal.if.usp.br/lamfi/pt-br/node/335. [Accessed: Jul. 14, 2026].

[14] National Institute of Standards and Technology (NIST), "PSTAR: Stopping Power and Range Tables for Protons," Physical Measurement Laboratory. [Online]. Available: https://physics.nist.gov/PhysRefData/Star/Text/PSTAR.html. [Accessed: Jul. 14, 2026].

[15] American Type Culture Collection (ATCC), "MG-63 Cell Line," ATCC Product Sheet CRL-1427. [Online]. Available: https://www.atcc.org/api/pdf/product-sheet?id=CRL-1427. [Accessed: Jul. 14, 2026].

[16] American Type Culture Collection (ATCC), "MDA-MB-231 Cell Line," ATCC Product Sheet CRM-HTB-26. [Online]. Available: https://www.atcc.org/api/pdf/product-sheet?id=CRM-HTB-26. [Accessed: Jul. 14, 2026].

[17] "Analysis of cell viability by the MTT assay," Cold Spring Harbor Protocols, vol. 2018, no. 6, Jun. 2018, doi: 10.1101/pdb.prot095505. Available: https://cshprotocols.cshlp.org/content/2018/6/pdb.prot095505.short

Development of low-cost bone phantoms for application in Ho-166 brachytherapy

Ana Luisa Carradore Sorato*
Biomedical Engineering
Faculdade Israelida de Ciências da Saúde Albert Einstein, Hospital Israelita Albert Einstein
São Paulo, Brazil
ana.csorato@einstein.edu.br

Gabriel Azevedo Poli Toffoli*
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein, Hospital Israelita Albert Einstein
São Paulo, Brazil
gabrieltoffoli2007@gmail.com

Bianca Peixoto
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein, Hospital Israelita Albert Einstein
São Paulo, Brazil
peixas7@gmail.com

Paula Cristina Guimarães Antunes
Biomedical Engineering
Faculdade Israelita de Ciência da Saúde, Hospital Israelita Albert Einstein
São Paulo, Brazil
paula.antunes@einstein.br

Roger Borges
Biomedical Engineering
Faculdade Israelita de Ciência da Saúde, Hospital Israelita Albert Einstein
São Paulo, Brazil
roger.b@einstein.br

Abstract*— This project aims to develop and characterize a low-cost bone-equivalent phantom for brachytherapy applications using commercially available materials. Because brachytherapy employs low-energy radioactive sources that generate steep dose gradients, accurate experimental dosimetry remains challenging, making tissue-equivalent phantoms essential for quality assurance and dosimetric validation. To address this need, several formulations based on plaster, epoxy resin with hardener, and PLA powder were produced and experimentally characterized by gravimetric analysis, density measurements, energy-dispersive X-ray spectroscopy (EDS), and scanning electron microscopy (SEM). Their dosimetric performance was further investigated through Monte Carlo simulations. The combined experimental and computational approach enabled the evaluation of the physical, morphological, elemental, and dosimetric properties of the proposed materials, demonstrating their potential for use in brachytherapy research and quality assurance.

***Index Terms*— brachytherapy, low-cost phantoms, dosimetry, bone tissue-equivalent materials, bone phantoms**

I. Introduction

Brachytherapy (BT) is a radiotherapy modality based on the use of sealed radioactive sources positioned inside or near the treatment region, allowing the administration of high radiation doses in a highly localized manner [1]. The growing precision of calculation algorithms and planning systems has increased the need for experimental validation of dose distributions, as recommended by the TG-186 report of the American Association of Physicists in Medicine (AAPM) [2].

In this context, physical phantoms play a fundamental role in dosimetric studies, quality control, and the validation of computational models. For experimental measurements to be representative, these devices must reproduce not only the geometry of the irradiated region but also the radiological properties of the biological tissues of interest [3].

This characterization becomes particularly important in brachytherapy applications using low-energy brachytherapy sources, where radiation interactions strongly depend on the elemental composition of the medium due to the predominance of the photoelectric effect. [4]. Consequently, small differences in density, effective atomic number, and energy absorption coefficients can produce significant changes in the dose distribution. This behavior is especially relevant for bone tissue, whose high concentration of calcium and phosphorus, associated with the variability between different bone types, complicates the development of equivalent materials capable of adequately reproducing its radiological properties.

Although several bone phantoms have been reported in the literature using polymeric and ceramic materials, as well as additive manufacturing techniques, many exhibit high manufacturing costs, complex fabrication processes, or limited characterization for brachytherapy applications involving low-energy sources. Consequently, there remains a need for

* These authors contributed equally to this work

alternative materials that combine low cost, ease of fabrication, and radiological equivalence to bone tissue [5].

In this context, the present study aims to develop and characterize low-cost bone phantom formulations based on plaster, high-viscosity epoxy resin, and polylactic acid (PLA) [6]. Their physical, morphological, and elemental properties were evaluated to assess their suitability for dosimetric studies involving low-energy brachytherapy sources.

II. Materials and Methods

A. Preparation of Bone Phantom Formulations

Four bone phantom formulations were developed using three commercially available materials: plaster (Fortaleza), high-viscosity epoxy resin (Real Epoxi) with its corresponding hardener, and microparticulated polylactic acid (PLA). Plaster and epoxy resin were selected because of their ability to reproduce the density and radiation attenuation properties of bone tissue, while PLA powder was incorporated into the formulations as a porosity-controlling and density-modifying agent.

The formulations were prepared by manually homogenizing the components according to the mass proportions presented in Table I.

The mass of each component was measured using an analytical balance with a resolution of 0.0001 g, maintaining a total mass of 10 g for all specimens. To facilitate the discussion of the results, the formulations were designated according to the relative proportions of their main constituents: G (plaster), R (epoxy resin), and P (PLA). Accordingly, four formulations were defined and organized into two groups, as summarized in Table I.

TABLE I
IDENTIFICATION AND COMPOSITION OF THE SAMPLES

Sample	Composition (g)			
	Plaster	Resin	Hardener	PLA
7G3R	7.000	2.000	1.000	-
4G4R1P	4.125	3.000	1.500	1.375
3G4R3P	3.000	2.670	1.330	3.000
7G2R	7.500	1.670	0.830	-

After homogenization, the samples were placed in silicone molds (Figure 1) and kept at room temperature (~20 °C) for 24 hours for curing.

Subsequently, to ensure complete curing in the internal part and the removal of residual water, the samples were subjected to drying in a muffle furnace. The thermal cycle was programmed starting from room temperature, with a controlled heating rate of 1 °C/min, until reaching the maximum temperature of 100 °C for 12 h, followed by natural cooling to room temperature. After cooling, the specimens were removed and weighed to obtain the total mass after complete drying.

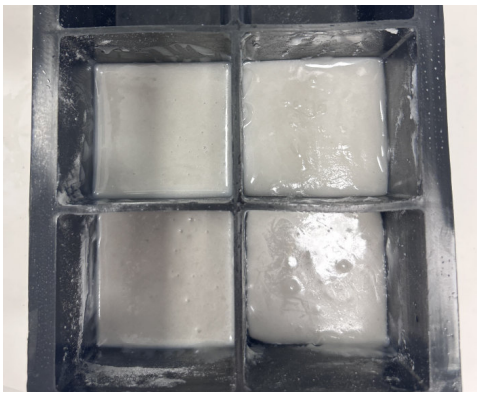


Fig. 1. Homogenized samples placed in silicone molds for room temperature curing.

B. Gravimetric Analysis

With the purpose of quantifying the mass variations resulting from the heat treatment and curing stages, gravimetric analysis was performed on both the isolated commercial plaster and the manufactured complete formulations.

Initially, the finished structures were weighed before and after thermal curing, which allowed for the determination of the total mass post-drying. Measurements were performed using an analytical balance, with an error of 0.0001 g, and the percentage mass loss was calculated using (1):

$$\Delta m\% = \frac{mi - mf}{mi} \times 100 \quad (1)$$

Subsequently, the plaster matrix was subjected to the same muffle furnace thermal cycle used in the production of the samples in order to determine the degree of hydration of the raw material.

Additionally, after the complete curing and drying process, the experimental density of each manufactured formulation was determined based on Archimedes' principle using the liquid displacement method. The volume of each sample was obtained by measuring the displaced volume of water, and the experimental density was subsequently calculated by its final mass by the measured volume.

C. Scanning Electron Microscopy (SEM)

The microstructural characterization of the specimens was performed using a Scanning Electron Microscopy (SEM) system, model JSM-6010LA (JEOL), applying a voltage of 8 kV, which allowed the evaluation of the surface morphology and homogeneity of the matrix formed by the constituent materials.

The samples were initially cut using a diamond disc cutter into dimensions of 10 mm x 10 mm, obtaining a uniform surface suitable for microscopic analysis.

D. Energy-Dispersive X-ray Spectroscopy (EDS)

Chemical composition analyses were performed using energy-dispersive X-ray spectroscopy (EDS) coupled with SEM. Three complementary approaches were conducted:

EDS spectra for the quantitative determination of the elemental composition of each sample, in order to distinguish the main elements present in each formulation; elemental mapping (EDS Mapping) to examine the spatial distribution of chemical elements along the sample surface; and morphological-compositional correlation, correlating the SEM images with the results obtained by EDS, aiming to relate morphological characteristics to the observed elemental distributions. The quantitative data obtained were compared with X-ray attenuation data from NIST XCOM.

E. Computational Simulation Using the Monte Carlo Method

To complement the experimental characterization and evaluate the dosimetric behavior of the developed formulations, Monte Carlo simulations were performed using the MCNP6.3 (Monte Carlo N-Particle Transport) code [7].

The computational model consisted of a homogeneous cylindrical phantom with a diameter of 4.0 cm, representing the material under investigation (Figure 2). The elemental composition of each formulation was defined according to the experimental results obtained by energy-dispersive X-ray spectroscopy (EDS). Furthermore, the experimental density values obtained via Archimedes' principle for each sample were directly applied to the material definitions in the MCNP6.3 input. For comparison, commercially available bone-equivalent materials from CIRS (Tissue Simulation & Phantom Technology) were also simulated using their reference compositions, together with the ICRP cortical bone standard, whose elemental composition was validated against the NIST database.

A Ho-166 brachytherapy source was modeled as a circular disk with a diameter of 0.2 cm positioned in direct contact with the phantom surface, as illustrated in Figure 2. The radionuclide was represented by the emission of β^- particles with a mean energy of 0.665 MeV, while secondary photon transport was also taken into account. Coupled electron-photon transport was performed using the physics models available in MCNP6.3, employing the MCPLIB08 library for photon transport and the e303 library for electrons.

The energy deposited within the phantom was calculated using the *F8 tally, which estimates the energy deposited in each cell as a result of particle interactions. To obtain the percentage depth dose (PDD) distribution, the phantom was subdivided into fifteen adjacent parallelepiped regions aligned with the direction of radiation incidence, as shown in Figure 2. The energy deposited in each region was then used to determine the depth-dependent energy deposition profile.

To quantitatively evaluate this dosimetric equivalence, the mean relative error between the doses deposited in the developed phantoms and the reference materials was calculated for the depth range of 0.01 - 0.11 cm.

All simulations were carried out on a workstation equipped with an Intel® Core™ i9-14900KF processor. The number of primary particle histories was adjusted to ensure statistical uncertainties below 1.0% (Type A, $\pm 1\sigma$) for all calculated quantities.

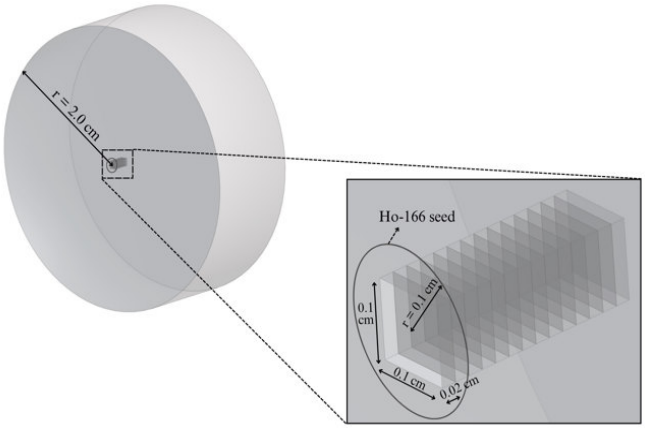


Fig. 2. Three-dimensional representation developed for simulations. On the left, the isometric view of the cylinder with a 2.0 cm radius. On the right, magnified detail of the central region, demonstrating the Ho-166 source and the arrangement of the 14 blocks used.

III. Results and Discussion

A. Gravimetric Analysis and Mass Loss Evaluation of the Specimens

The gravimetric characterization of the sample formulations was performed after the thermal cycle in the muffle furnace. The mass quantification of the specimens pre- and post-treatment provided the measurement of the percentage magnitude of the loss (m%), primarily associated with dehydration and the release of remaining volatile fractions from the ceramic matrix. The obtained experimental data are summarized in Table II.

TABLE II
GRAVIMETRIC PROPERTIES OF THE FORMULATIONS PRE- AND POST-MUFFLE FURNACE HEAT TREATMENT

Sample	Initial mass (g)	Final mass (g)	Mass loss (%)
7G3R	10.00	7.21	27.93
4G4R1P	10.00	5.47	45.33
3G4R3P	10.00	6.10	38.96
7G2R	10.00	8.27	17.27

The gravimetric profiles obtained in Table II demonstrate a general decrease in the mass of all specimens, corroborating the occurrence of kinetic processes of water elimination and volatile byproducts during the proposed thermal cycle. It is also observed that the magnitude of the percentage mass loss was directly influenced by the composition of each system, especially in systems with a higher fraction of resin, indicating that volatile organic compounds evaporated during the drying processes. The final experimental densities determined for the developed formulations are summarized in Table III.

TABLE III
EXPERIMENTAL DENSITY OF THE MANUFACTURED FORMULATIONS

Sample	Experimental Density (g/cm³)
7G3R	2.04
4G4R1P	1.93
3G4R3P	1.94
7G2R	2.67

B. Morphological Characterization via SEM

Figure 3 presents the scanning electron micrographs of the four developed formulations, allowing the influence of composition on the surface morphology and internal organization of the composite matrix to be qualitatively evaluated. In general, formulations composed only of plaster and epoxy resin exhibited a more compact and homogeneous microstructure, whereas the incorporation of PLA produced noticeable changes in surface morphology and in the spatial distribution of the constituent phases.

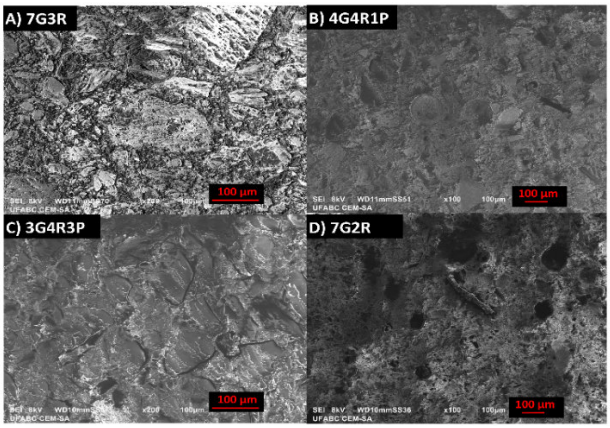


Fig. 3. SEM showing the surface morphology and topography of fragments from the developed formulations. The interpenetration domains of the organic and ceramic phases are observed, as well as the presence of porosities with predominantly spheroidal geometry and micrometric voids. Images obtained under an acceleration voltage of 8 kV.

The 7G3R (Figure 3A) and 7G2R (Figure 3D) formulations, composed exclusively of plaster, epoxy resin, and hardener, exhibited a continuous and relatively homogeneous matrix. In sample 7G2R, predominantly spheroidal pores were observed, heterogeneously distributed across the analyzed surface. These voids are likely associated with air entrapment during the mixing process and/or the evaporation of residual moisture during the curing and drying stages, phenomena commonly observed in plaster-based composite materials.

In contrast, the formulations containing PLA (4G4R1P and 3G4R3P) exhibited rougher surfaces and a higher density of micrometric interfaces than the formulations without this additive. These observations indicate that the incorporation of PLA modifies the organization of the composite matrix during curing, promoting a more heterogeneous distribution of the ceramic and polymeric phases and reducing the matrix continuity observed in the other formulations.

The microstructural differences revealed by SEM demonstrate that the formulation composition directly influences the internal architecture of the developed composites. Considering their intended application as bone-equivalent phantoms for brachytherapy, these morphological characteristics may affect relevant physical properties, such as density, elemental homogeneity, and radiation transport. Therefore, the SEM observations complement the gravimetric, compositional, and Monte Carlo analyses, providing a more comprehensive understanding of how each formulation influences the overall performance of the proposed phantom materials.

C. Chemical Characterization by EDS

The elemental composition of the developed formulations was investigated by energy-dispersive X-ray spectroscopy (EDS) in order to evaluate the influence of each constituent material on the chemical composition of the composites. Figure 4 presents the EDS spectra obtained for the four formulations, while Figure 5 shows the corresponding elemental distribution maps.

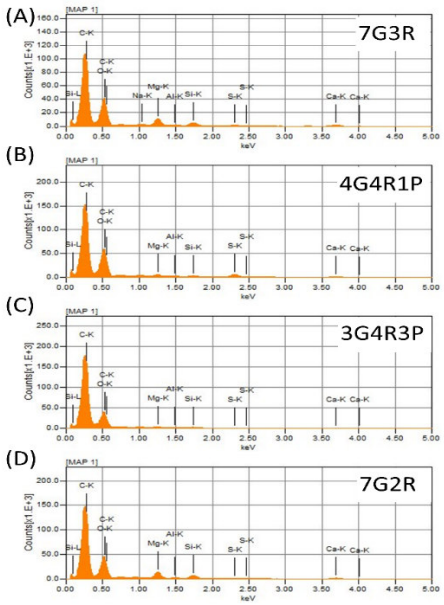


Fig. 4. EDS spectra of the developed formulations, showing the peaks of each identified element in the different compositions.

The generated spectra confirmed the presence of characteristic energy peaks of the elements theoretically expected from the raw materials used, including carbon (C), oxygen (O), magnesium (Mg), aluminum (Al), silicon (Si), sulfur (S), and calcium (Ca). Carbon and oxygen concentrations were identified in the lower energy regions of the spectrum, likely corresponding to the organic fraction of the polymeric base derived from the resin and PLA, as well as emission lines related to sulfur and calcium, presumably derived from the mineral phases of the commercial plaster. Furthermore, a residual presence of magnesium, aluminum, and silicon was detected

distributed across the samples, likely due to impurities in the raw materials, which are for industrial or commercial use and therefore do not possess high purity. The average elemental composition obtained for each of the formulations, along with their respective standard deviations, is presented in Table IV.

TABLE IV
MEAN AND STANDARD DEVIATION OF ELEMENTS PER SAMPLE

Sample	C	O	Mg	Al	Si	S	Ca
7G3R	57.4 ± 6.8	29.0 ± 1.0	4.1 ± 1.3	0.5 ± 0.1	2.7 ± 1.0	1.5 ± 1.3	4.8 ± 3.5
4G4R1P	63.1 ± 3.2	28.6 ± 1.3	2.2 ± 0.5	0.4 ± 0.2	1.5 ± 0.4	1.8 ± 1.6	2.5 ± 1.7
3G4R3P	70.1 ± 0.4	26.3 ± 0.1	1.4 ± 0.2	0.5 ± 0.1	1.1 ± 0.2	0.2 ± 0.02	0.4 ± 0.3
7G2R	57.0 ± 1.9	29.2 ± 1.9	4.8 ± 0.5	0.6 ± 0.1	2.9 ± 0.6	0.8 ± 0.3	4.8 ± 1.1

It can be observed that the chemical composition varied among the formulations due to the distinct proportions of each matrix used. Regarding the identification of the expected elements, the analyses pointed out that the calcium concentration was lower than the predicted nominal content, due to the volume of the ceramic matrix added during the manufacturing process. The elemental mapping by EDS presented in Figure 5 allows for the evaluation of the spatial distribution of the elements in the samples.

The EDS spectra confirmed the presence of the major elements expected from the raw materials used in the formulations, namely carbon (C), oxygen (O), magnesium (Mg), aluminum (Al), silicon (Si), sulfur (S), and calcium (Ca). Carbon and oxygen were predominantly associated with the polymeric constituents (epoxy resin and PLA), whereas calcium and sulfur originated mainly from the calcium sulfate phase of the commercial plaster. Minor amounts of magnesium, aluminum, and silicon were also detected in all formulations, most likely due to naturally occurring impurities present in the commercial-grade raw materials.

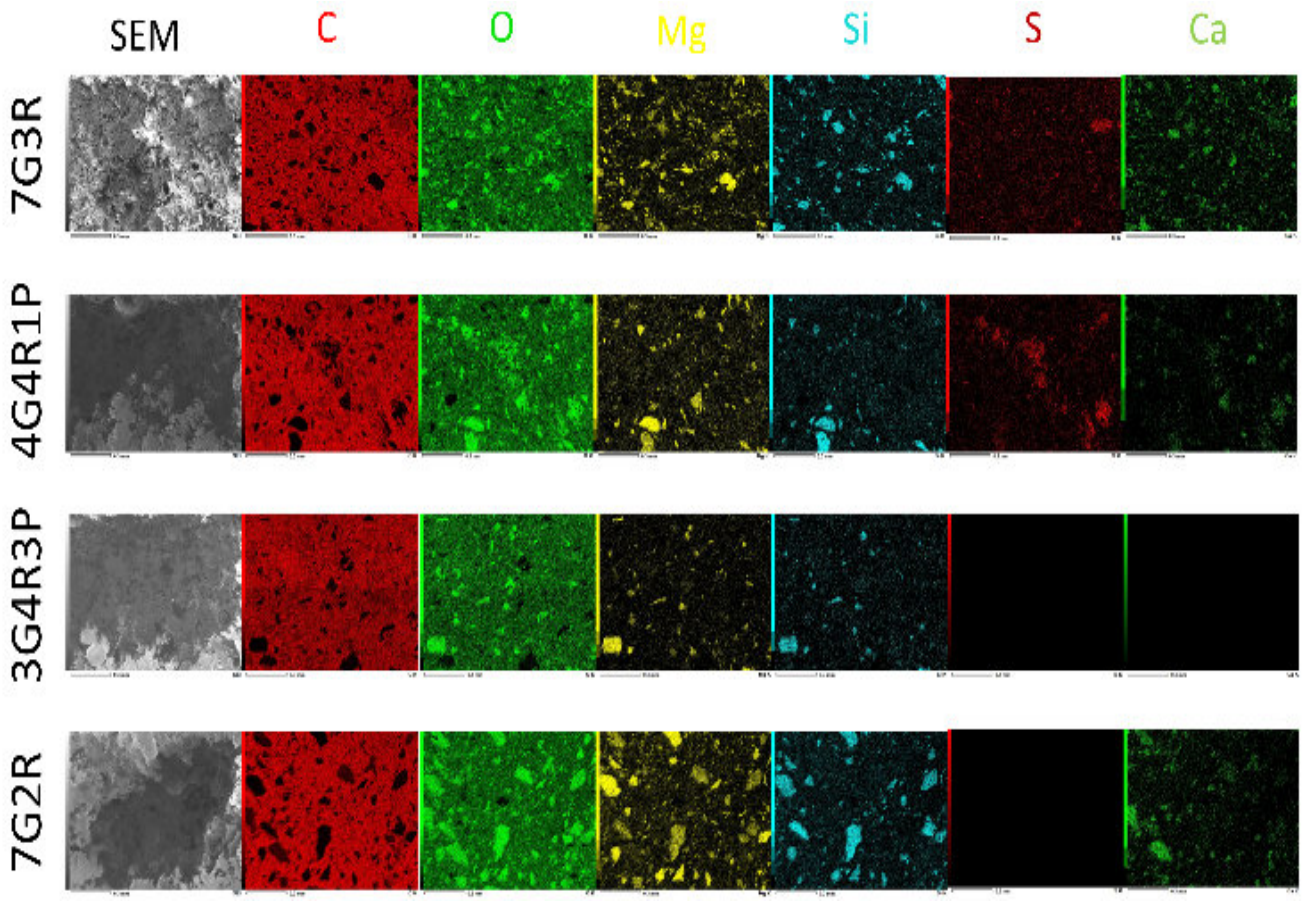


Fig. 5. Elemental mapping by energy-dispersive X-ray spectroscopy (EDS Mapping) of the surfaces of the developed formulations: (a) 7G3R, (b) 4G4R1P, (c) 3G4R3P, and (d) 7G2R." The spatial distributions correspond to the characteristic signals of Carbon (C), Oxygen (O), Magnesium (Mg), Silicon (Si), Sulfur (S), and Calcium (Ca).

The average elemental composition of each formulation, together with the corresponding standard deviations, is presented in Table IV. As expected, differences in elemental composition were observed among the formulations, reflecting the varying proportions of plaster, epoxy resin, and PLA used during fabrication. In particular, formulations containing higher amounts of polymeric material exhibited an increase in the relative carbon content, accompanied by a decrease in the relative concentrations of calcium and sulfur, consistent with the lower proportion of plaster incorporated into these composites.

The elemental maps shown in Figure 5 revealed a homogeneous spatial distribution of the detected elements throughout the analyzed regions, with no evidence of significant elemental segregation or phase clustering. This result indicates that the manufacturing process promoted adequate dispersion of the constituent materials within the composite matrix, contributing to the chemical homogeneity of the developed formulations.

From the perspective of phantom development for brachytherapy applications, this chemical homogeneity is particularly important because the interaction of ionizing radiation with matter depends directly on its elemental composition. Therefore, the EDS results demonstrate that the proposed manufacturing process produced chemically homogeneous phantoms, supporting the reliability of the subsequent Monte Carlo simulations and dosimetric evaluations.

D. Monte Carlo Dosimetric Analysis

With the purpose of evaluating the dosimetric equivalence of the manufactured formulations in relation to the reference bone materials, the ratio between the dose deposited in each proposed sample and the dose obtained in the simulated reference materials was calculated. The dose ratio was determined point-by-point along the radiation penetration depth.

Comparisons were made between each of the four manufactured specimens and the bone materials from the literature adopted as references, with densities ranging between 1.161 and 2.150 g/cm3. The analysis was conducted limited to depths between 0.01 and 0.11 cm, since, for greater depths, the dose deposited by the Ho-166 source decreases significantly due to the limited range of the particles, resulting in values with a negligible contribution to the evaluation.

Figure 6 presents the Percentage Depth Dose (PDD) curves for the different simulated materials, comparing the proposed formulations to the bone materials cited in the literature with distinct densities. The reference materials from CIRS are represented in green, while the developed formulations are shown in their respective colors.

Overall, the developed formulations exhibited depth-dose profiles comparable to those of the reference bone-equivalent materials. Nevertheless, differences in the position of the dose fall-off were observed, reflecting the influence of material density and composition on beta-particle transport and energy deposition within the materials. Among the investigated formulations, 7G2R exhibited the earliest dose fall-off, indicating a more rapid reduction in the deposited dose with depth. In contrast, formulations 7G3R, 4G4R1P, and 3G4R3P displayed very similar depth-dose profiles throughout the region of interest, suggesting comparable radiological behavior despite differences in their compositions.

To further quantify these differences, Figures 7A–D present the ratio between the dose deposited in each developed formulation and that calculated for the CIRS reference materials with known densities. A dose ratio close to unity indicates greater agreement between the developed formulation and the corresponding reference material.

At shallow depths (up to approximately 0.05 cm), all formulations exhibited dose ratios close to unity, indicating excellent agreement with the reference bone materials in this region. As the depth increased, larger deviations from unity became evident, particularly for the higher-density reference materials. This behavior is associated with the limited penetration range of the beta particles emitted by the Ho-166 source. As the deposited dose decreases rapidly with depth, small differences in material composition and density become more pronounced in the deeper regions.

Within the analyzed depth range (0.01-0.11 cm), formulation 7G3R (Figure 7A) exhibited the closest agreement with the reference materials of densities 1.820 and 1.850 g/cm³, presenting mean relative errors of 4.56% and 5.07%, respectively. Additionally, formulations 4G4R1P (Figure 7B) and 3G4R3P (Figure 7C) demonstrated the highest dosimetric equivalence with the 1.660 g/cm³ reference material, yielding mean relative errors of 3.68% and 4.20%, respectively. The similar behavior observed for these formulations suggests that the incorporation of PLA did not significantly modify the depth-dose characteristics of the developed composites, preserving their radiological behavior relative to the reference materials.

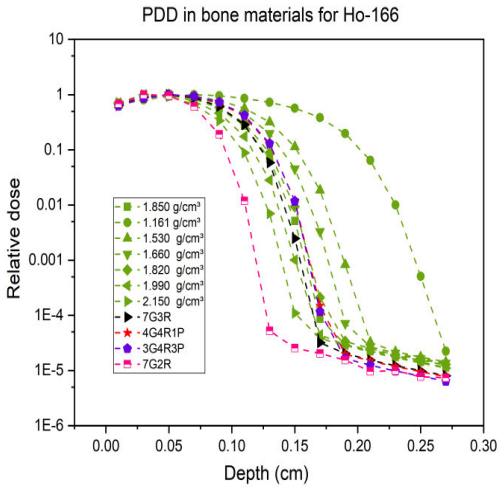


Fig. 6. Percentage Depth Dose (PDD) curves obtained via Monte Carlo simulation for the developed samples and reference materials with different densities.

In contrast, formulation 7G2R (Figure 7D) exhibited a progressive decrease in the dose ratio beyond approximately 0.06 cm for most of the reference materials, indicating a greater deviation from the expected dose deposition profiles. This behavior suggests that differences in its composition and density resulted in a distinct beta-particle transport pattern compared with the reference materials.These observations indicate that formulations 7G3R, 4G4R1P, and 3G4R3P more closely reproduce the radiological behavior of the selected CIRS bone-equivalent reference materials over the investigated depth range, whereas formulation 7G2R exhibits a distinct dose deposition profile.

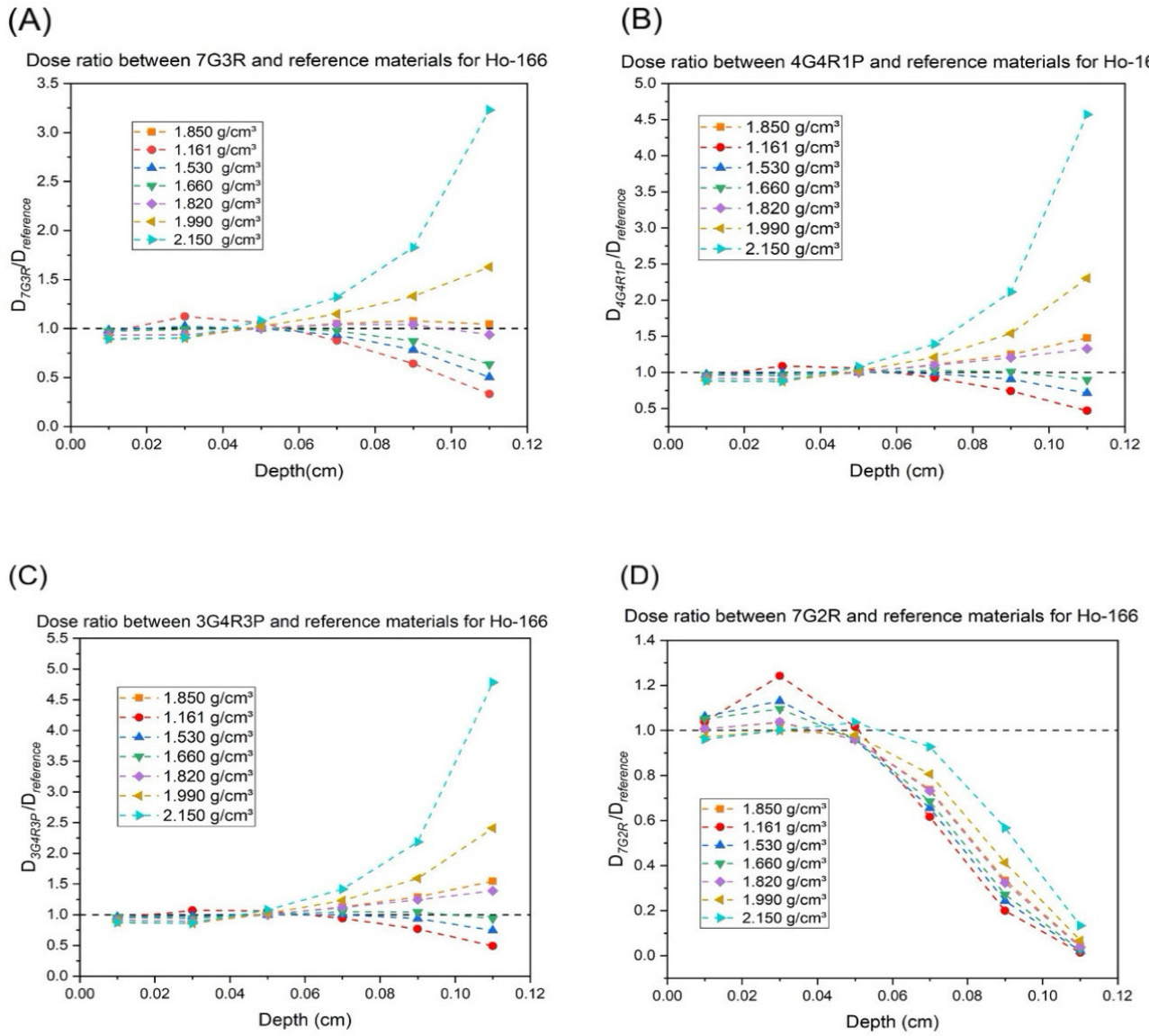


Fig. 7. Dose ratio between the developed formulations and reference materials with different densities for Ho-166.

IV. Conclusion

The present study demonstrated the feasibility of developing low-cost bone-equivalent phantoms using commercially available materials for potential applications in brachytherapy with Ho-166. Gravimetric, morphological, and chemical characterizations showed that variations in the proportions of plaster, epoxy resin, and PLA directly influence the physical properties of the developed formulations. In addition, Monte Carlo simulations demonstrated that these compositional differences affect the radiological behavior and depth-dose distributions of the materials under the simulated irradiation conditions.

The criteria for determining the best overall equivalence were based on two main factors: dosimetric agreement, defined by a dose ratio as close to unity as possible, and the lowest mean relative error compared to the reference materials across the evaluated depth range (0.01 - 0.11 cm). Among the investigated formulations, 7G3R, 4G4R1P, and 3G4R3P exhibited the highest agreement with the dose deposition profiles of the selected CIRS bone-equivalent reference materials. In particular, formulation 7G3R most closely reproduced the radiological behavior of the reference materials with densities of 1.820 and 1.850 g/cm³, while formulations 4G4R1P and 3G4R3P exhibited optimal equivalence with the 1.660 g/cm³ reference. The comparable behavior of the PLA-containing samples suggests that the presence of this polymer does not substantially impact the depth-dose properties of the composites. However, the 7G2R formulation displayed a faster dose fall-off as depth increased, pointing to a reduced radiological agreement under the simulated conditions.

Overall, the proposed methodology provides a simple and low-cost approach for developing bone-equivalent materials, demonstrating primarily manufacturing feasibility and preliminary computational equivalence to reference bone densities for

Ho-166 irradiation. To confirm their potential for dosimetry, calibration, and validation procedures, subsequent stages of this research will incorporate experimental measurements using physical detectors to validate the simulated data. Furthermore, future studies will evaluate the manufacturing reproducibility of their experimental densities across multiple batches and expand the clinical applicability by utilizing other brachytherapy sources.

REFERENCES

[1] Institute for Quality and Efficiency in Health Care (IQWiG), "In Brief: Brachytherapy," InformedHealth.org, Cologne, Germany, Sep. 15, 2022.

[2] AAPM Task Group No. 138, "A dosimetric uncertainty analysis for photon-emitting brachytherapy sources: Report of AAPM Task Group No. 138 and GEC-ESTRO," American Association of Physicists in Medicine, College Park, MD, USA, AAPM Rep. 138, 2011.

[3] P. C. G. Antunes, P. T. D. Siqueira, J. M. B. Shorto, and H. Yoriyaz, "A versatile physical phantom design and construction for I-125 dose measurements and dose-to-medium determination," *Brachytherapy*, vol. 22, pp. 80–92, 2023.

[4] P. C. G. Antunes, P. T. D. Siqueira, J. M. B. Shorto, and H. Yoriyaz, "Heterogeneous physical phantom for I-125 dose measurements and dose-to-medium determination," *Brachytherapy*, vol. 23, pp. 73–84, 2024.

[5] G. Bielęda, G. Zwierzchowski, K. Rosłan, A. Adamus, and J. Malicki, "Dosimetric assessment of the impact of low-cost materials used in stereolithography in high-dose-rate brachytherapy," *Journal of Contemporary Brachytherapy*, vol. 13, no. 2, pp. 188–194, Apr. 2021.

[6] L. Wu, G. Yan, F. Wang, J. Xie, S. He, and Y. Wang, "Clinical observation of customized 3D-printed templates assisted 125I radioactive seed implantation for pelvic malignant tumors," *Frontiers in Oncology*, vol. 14, Art. no. 1421045, Jun. 2024.

[7] S. Sahoo, T. P. Selvam, R. S. Vishwakarma, and G. Chourasiya, "Monte Carlo modeling of 60Co HDR brachytherapy source in water and in different solid water phantom materials," *Journal of Medical Physics*, vol. 35, no. 1, pp. 15–22, Jan.–Mar. 2010.

[8] R. S. Sloboda, H. Morrison, B. Cawston-Grant, and G. V. Menon, "A brief look at model-based dose calculation principles, practicalities, and promise," *Journal of Contemporary Brachytherapy,* vol. 9, no. 1, pp. 79–88, Feb. 2017.

[9] G. P. Delpino, R. Borges, T. Zambanini, J. F. S. Joca, I. Gaubeur, A. C. S. Souza, and J. Marchi, "Sol-gel-derived 58S bioactive glass containing holmium aiming brachytherapy applications: A dissolution, bioactivity, and cytotoxicity study," *Materials Science and Engineering: C*, vol. 119, Art. no. 111595, Feb. 2021.

[10] T. Zambanini, R. Borges, A. C. S. Souza, G. Z. Justo, J. Machado, D. R. Araujo, and J. Marchi, "Holmium-containing bioactive glasses dispersed in poloxamer 407 hydrogel as a theragenerative composite for bone cancer treatment," *Materials*, vol. 14, no. 6, Art. no. 1459, Mar. 2021.

[11] R. Borges, T. Zambanini, A. M. Pelosine, G. Z. Justo, A. C. S. Souza, J. Machado, J. F. Schneider, D. R. Araujo, and J. Marchi, "A colloidal hydrogel-based drug delivery system overcomes the limitation of combining bisphosphonates with bioactive glasses: In vitro evidence of a potential selective bone cancer treatment allied with bone regeneration," *Biomaterials Advances,* vol. 151, Art. no. 213441, Aug. 2023.

Dimensional Validation Methodology of Low-Cost Phantoms for Magnetic Resonance Imaging Manufactured by Advanced Manufacturing

Julia Abbondanza Schahin
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein,
Hospital Israelita Albert Einstein
São Paulo, Brazil
julia.aschahin@einstein.edu.br

Rafael Ribeiro Henriques Netto
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein,
Hospital Israelita Albert Einstein
São Paulo, Brazil
rafael.hnetto@einstein.edu.br

Paula Cristina Guimarães Antunes
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein,
Hospital Israelita Albert Einstein
São Paulo, Brazil
paula.antunes@einstein.br

Roger Borges
Biomedical Engineering
Faculdade Israelita de Ciências da Saúde Albert Einstein,
Hospital Israelita Albert Einstein
São Paulo, Brazil
roger.b@einstein.br

***Abstract*—Magnetic resonance imaging (MRI) presents geometric distortions intrinsic to the acquisition process, arising mainly from the nonlinearity of spatial gradients and the inhomogeneity of the static magnetic field. These distortions have a relevant clinical impact on applications that require millimetric spatial precision, such as stereotactic radiosurgery and imageguided radiotherapy. Phantoms constitute the main methodological tool for quantifying these deviations, but their reliability depends directly on the manufacturing process employed and on the dimensional validation method adopted. This study aimed to develop and dimensionally compare two phantom models, one geared toward the evaluation of repetitive linear features of decreasing dimensions (Model 1, Comb) and another toward the evaluation of concentric hollow geometries at multiple scales (Model 2, Squares). Both were manufactured by additive manufacturing (3D printing by PETG filament extrusion, polyethylene terephthalate glycol) and by subtractive manufacturing (laser cutting in PMMA, polymethyl methacrylate). Dimensional validation was carried out by two independent methods, a physical caliper and image processing via ImageJ, with overall accuracy quantified by the Relative Root Mean Square Error (rRMSE) and statistically tested by Student's t-test ($\alpha = 0.05$), preceded by a normality check via Shapiro-Wilk. The findings indicate that PETG is preferable for reproducing fine linear patterns, while PMMA remains a viable alternative for robust volumetric structures, reinforcing the physical caliper as the gold standard for high-sensitivity dimensional validation.**

***Index Terms*—magnetic resonance imaging, geometric distortion, dimensional validation, low-cost phantoms**

I. Introduction

Magnetic resonance imaging (MRI) is one of the leading medical imaging modalities for clinical diagnosis. Unlike other techniques, it does not use ionizing radiation, but rather intense magnetic fields associated with radiofrequency (RF) pulses to obtain high-resolution cross-sectional images of soft tissues and anatomical structures, both in 2D and 3D [1].

Physically, the method exploits the magnetic moment of hydrogen protons: under a static field (B_0), the spins align and precess at the Larmor frequency; an RF pulse at the same frequency shifts the magnetization to the transverse plane, and the return to equilibrium (T1 and T2 relaxation) generates the signal detected by the receiver coils. Since different tissues have distinct relaxation times, due to the interactions of protons with the adjacent chemical components of each tissue, the resulting contrast enables anatomical differentiation [2].

Since the first clinical equipment, in the early 1980s, it was observed that the dimensions reconstructed in the image did not correspond exactly to the real dimensions of the objects, with larger divergences at positions farther from the center of the magnet. This is the result of a systematic error caused mainly by the nonlinearity of the gradient field and the inhomogeneity of B_0, added to patient-dependent effects such as magnetic susceptibility and water-fat chemical shift [3]. The problem was already known to the scientific community in the 1980s [4], but its clinical impact only became critical when MRI began to support millimetric-precision applications, such as stereotactic radiosurgery and image-guided radiotherapy, in which distortions of a few millimeters compromise the coverage of the therapeutic target [3]. The choice of acquisition sequence also introduces relevant variability; in patients with hepatocellular carcinoma, the tumor measurement error by MRI varied between 0.75 and 0.87 cm depending on the sequence, resulting

in incorrect classifications according to the criteria in up to 14.3% of cases [5].

Phantoms emerged as the main tool for quantifying these errors, as they possess precisely known geometry and composition. The first three-dimensional phantoms made it possible to map geometric distortion across different equipment manufacturers [3], although more recent studies show that the deviations measured in phantoms do not always reflect the behavior observed in real anatomical images [6]. Today, commercial phantoms are used together with normative criteria, such as the AAPM's TG-284, to validate MRI simulation protocols in radiosurgery [7]. Their use, moreover, is not limited to MRI, as they are employed in practically all medical imaging modalities (CT, PET, SPECT, ultrasound) as a control standard for validation and quality control, with the advantage of dispensing with patient exposure to risk and allowing longitudinal comparisons between equipment [8].

The manufacturing of phantoms has been driven by advanced manufacturing. FDM (fused deposition modeling) 3D printing in polyethylene terephthalate glycol (PETG) offers low cost and good dimensional stability, and relaxometry characterization studies (T1, T2, T2*) have already evaluated this and other thermoplastics as possible tissue mimics in MRI [9, 10]. Polymethyl methacrylate (PMMA), historically machined by laser cutting or CNC (computer numerical control), remains widely used in commercial and research phantoms owing to its good magnetic susceptibility compatibility with water and mineral oils, even though laser cutting introduces thermal edge effects that can compromise dimensional fidelity in small features [11].

The objective of this work was to determine whether additive or subtractive manufacturing techniques present greater dimensional accuracy in MRI calibration. This study reports the methodological development of two phantoms for MRI calibration, comparing additive manufacturing (FDM 3D printing in PETG) and subtractive manufacturing (laser cutting in PMMA), focusing on the comparison between these manufacturing methods and on the proposal of a dimensional validation methodology capable of indicating which process produces the geometrically more faithful phantom.

II. Materials and Methods

A. Geometric Design of the Phantoms

To comprehensively map geometric errors across different size scales and volumetric configurations, two computational phantom models were developed using the Onshape CAD platform:

1) Model 1 (Comb): Designed to evaluate spatial resolution and the fidelity of linear features and repetitive slits, with decreasing nominal dimensions (total length of 64.0 mm).

2) Model 2 (Square Sequence): Composed of a matrix of concentric square cavities, designed to evaluate the stability of robust, orthogonal hollow geometries, in dimensions ranging from 2.0 mm to 20.0 mm.

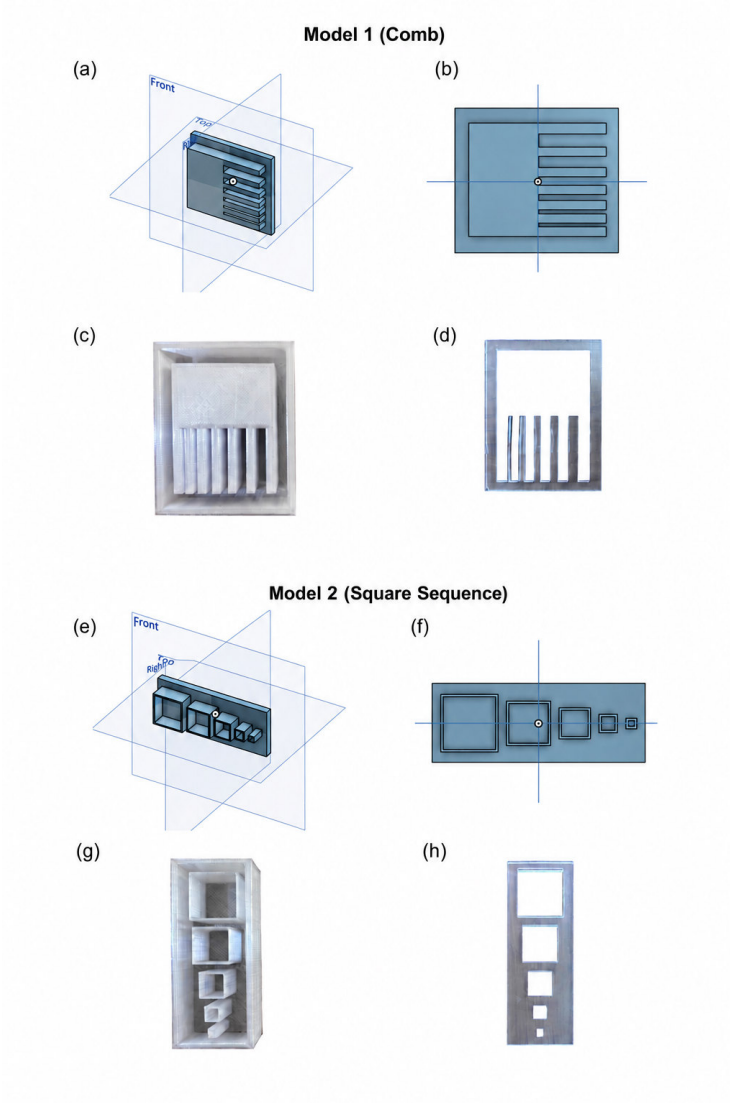


Fig. 1. Physical phantom models developed for metrological evaluation. For Model 1 (Comb), panels (a) and (b) show the three-dimensional and front-view CAD models created in Onshape, respectively; panel (c) shows the FDM 3D-printed PETG phantom, and panel (d) the corresponding laser-cut PMMA phantom. For Model 2 (Square Sequence), panels (e) and (f) present the three-dimensional and front-view CAD models created in Onshape, respectively; panel (g) shows the FDM 3D-printed PETG phantom, and panel (h) the corresponding laser-cut PMMA phantom.

B. Advanced Manufacturing Parameters

1) Subtractive Manufacturing (Laser Cutting – PMMA): The PMMA sheets were processed on an industrial router laser cutter, Prime Cutmaker – CMK 6040 model. The cutting profile was configured with a constant linear speed of 5 mm/s, minimum CO_2 tube power set at 75%, and maximum power at 80%. For the structural bonding of the multilayer acrylic plates, the commercial structural adhesive S-330 Sinteglas was used, prepared strictly at a volumetric ratio of 25 mL of resin base to 1 mL of liquid chemical catalyst.

2) Additive Manufacturing (3D Printing – PETG): The phantoms were fabricated on a high-precision kinematic 3D

printer, Core A1V2 model. Commercial GTmax PETG filament was used, with a nominal diameter of 1.75 mm. The selected extruder nozzle had an outlet diameter of 0.4 mm. The kinematic and thermal slicing parameters were extrusion temperature set at 235 °C, heated bed temperature at 85 °C, standardized layer height of 0.2 mm, print speed of 80 mm/s, and internal infill density set at 50% using a crosshatch rectilinear geometric pattern [10].

C. Metrological Validation Protocol

To mitigate measurement bias and establish statistical error control, dimensional validation was conducted in a binary fashion. For the physical approach, the six measurements collected for each nominal dimension were performed by six different evaluators, all following the same standardized measurement instructions, in order to account for inter-operator variability in the manual measurement process:

1) Physical Approach (Physical Gold Standard): Direct measurements performed with a calibrated caliper (0.01 mm resolution). Six independent measurements were collected (n = 6) for each nominal dimension of interest, minimizing parallax error and operator grip-pressure variability. Despite being adopted as the gold standard, the caliper technique presents inherent limitations that must be acknowledged: its resolution (0.01 mm) approaches the same order of magnitude as the deviations measured in the smallest features (1.5–2.0 mm), its accuracy remains partially dependent on operator technique even under standardized instructions, and it is restricted to two-point linear contact measurements, being unable to capture curved or internal geometric profiles as comprehensively as image-based methods.

2) Digital Approach (Image Processing): Standardized high-resolution photographic capture, under controlled, orthogonal lighting. The images were imported into the open-source software ImageJ [12]. The protocol consisted of spatial calibration in pixels per millimeter (px/mm), based on a known scale, followed by semi-automated digital measurement of the geometric profiles.

D. Statistical Analysis and Error Protocol

The overall cumulative precision of each technical scenario was quantified through the Relative Root Mean Square Error (rRMSE), calculated according to Equation 1:

$$rRMSE = \sqrt{\frac{1}{M} \times \sum \left(\frac{(X_{exp} - X_{theor})}{X_{theor}} \right)^2} \times 100 \qquad (1)$$

Where X_exp represents the mean of the experimental measurements, X_theor represents the theoretical nominal dimension extracted from the CAD, and M indicates the total number of distinct features evaluated in the geometric model. The normality of the residuals and relative errors was tested using the Shapiro-Wilk criterion. Once the parametric distribution was confirmed, Student's t-test for independent samples was applied to assess the existence of statistically significant differences between the materials (PETG vs. PMMA) and between the metrological methods (Caliper vs. ImageJ), adopting a significance level of $\alpha = 0.05$.

III. Results

A. Direct Comparative Analysis

After a series of measurements using both protocols mentioned above, we obtained the difference between the expected values (exact CAD model values) and the mean of the experimentally obtained values. These data are compiled in Tables I and II.

TABLE I

Difference Between Expected and Experimental Value in Model 1 (Comb)

Feature (Model 1)	Measurements relative to CAD (mm)				
	CAD (mm)	*PETG (Cal.)*	*PETG (ImgJ)*	*PMMA (Cal.)*	*PMMA (ImgJ)*
Total Length	64.00	-0.38	+2.51	+0.23	+0.22
Structural Height	51.00	-0.34	+1.88	+0.39	+0.49
Comb Thickness	4.00	+0.01	-0.18	+0.35	+0.06
Gap 1 (Largest)	6.00	-0.06	+0.04	+0.16	-0.04
Gap 2	5.00	-0.05	+0.05	+0.22	-0.33
Gap 3	4.00	+0.07	+0.02	+0.11	-0.09
Gap 4	3.00	+0.01	+0.11	+0.19	+0.23
Gap 5	2.00	+0.02	+0.02	+0.23	+0.14
Gap 6 (Smallest)	1.50	+0.04	+0.08	+0.27	+0.19

Note: Cal. = Physical measurement via Caliper; ImgJ = Digital measurement via ImageJ software.

TABLE II

Difference Between Expected and Experimental Value in Model 2 (Squares)

Feature (Model 2)	Measurements relative to CAD (mm)				
	CAD (mm)	*PETG (Cal.)*	*PETG (ImgJ)*	*PMMA (Cal.)*	*PMMA (ImgJ)*
Square 1 (Largest)	20.00	-0.30	+0.34	+0.31	+0.40
Square 2	15.00	-0.38	-0.12	+0.35	+0.47
Square 3	10.00	-0.29	+0.24	+0.28	+0.36
Square 4	5.00	-0.17	+0.22	+0.22	+0.40
Square 5 (Smallest)	2.00	-0.13	-0.01	+0.15	+0.33

Note: Cal. = Physical measurement via Caliper; ImgJ = Digital measurement via ImageJ software.

B. Overall rRMSE Analysis

In addition to the direct comparison of expected values (exact CAD model values), we compared the rRMSE, aiming to compensate for these differences as a function of the size of each part of the phantoms. The consolidated values of the overall rRMSE (%) indicator, broken down by phantom model, polymeric material, and applied metrological methodology, are summarized in Table III.

TABLE III
OVERALL RRMSE VALUES (%) CALCULATED FOR THE DIFFERENT PHANTOMS, MATERIALS, AND MEASUREMENT METHODS

Validation method	Models rRMSE Values (%)			
	Comb PETG	*Comb PMMA*	*Squares PETG*	*Squares PMMA*
Caliper	1.29	8.11	3.81	6.51
ImageJ	3.22	5.96	2.35	8.49

Note: rRMSE represents the square root of the overall cumulative relative mean square error for each geometric set. Lower values indicate greater overall accuracy and manufacturing process replication stability.

C. Overall Statistical Analysis and Dimensional Fidelity

The Shapiro-Wilk test confirmed the adequacy of a parametric model for all groups of relative percentage errors calculated ($p > 0.05$), statistically validating the subsequent use of Student's t-test. The internal variability and the behavior of the deviations with respect to the measurement methods (Caliper and ImageJ) and to the phantom geometries are illustrated in the Box Plot of Figure 2.

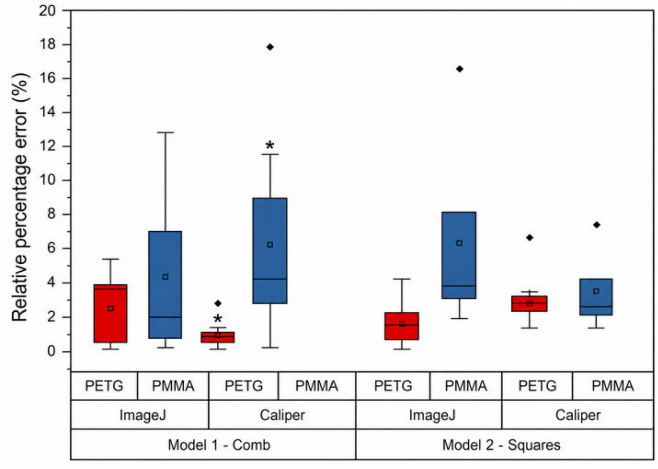


Fig. 2. Distribution of the relative percentage error (%) for PETG and PMMA phantoms in Model 1 (Comb) and Model 2 (Square Sequence), evaluated using digital image analysis (ImageJ) and caliper measurements. Boxplots show the median (horizontal line), interquartile range (box), whiskers extending to 1.5 × IQR, and outliers (individual points). The asterisk (*) denotes a statistically significant difference between PETG and PMMA for the corresponding comparison ($p < 0.05$).

For Model 1 (Comb), the physical approach via caliper demonstrated a robust and statistically significant disparity between materials ($p = 0.0176$), confirming the superiority of the geometric stability of the injected PETG filament over the thermal ablation cutting of PMMA. However, when the same model was evaluated by the digital method in ImageJ, no statistical significance was detected in the difference between materials ($p = 0.285$). For Model 2 (Square Sequence), an absence of statistical significance was observed on both metrological fronts: caliper-based validation resulted in $p = 0.829$, while ImageJ analysis returned $p = 0.120$.

D. Scale Effect on Geometric Deviations

The influence of the nominal feature size on the relative percentage error was further evaluated to investigate the scale effect on geometric deviations. The corresponding results are presented in Figure 3 and 4.

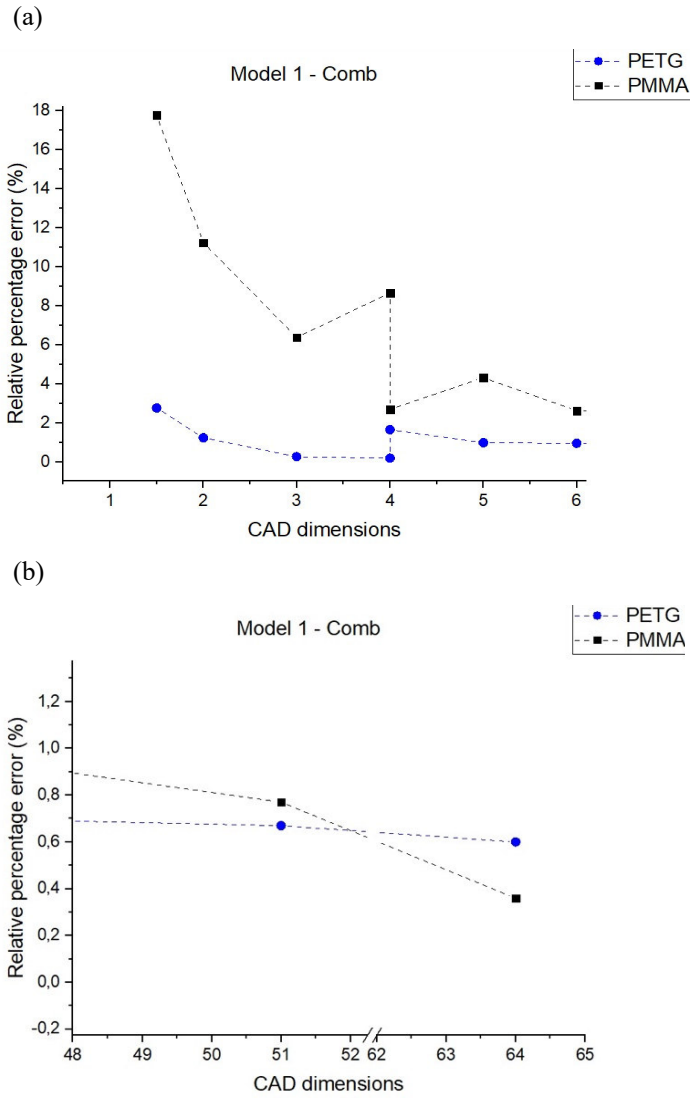


Fig. 3. Distribution curve of relative percentage errors measured using a caliper as a function of the nominal dimensional scale of the designed features: (a) Model 1 (Comb) at low nominal dimensions; (b) Model 1 (Comb) at higher nominal dimensions.

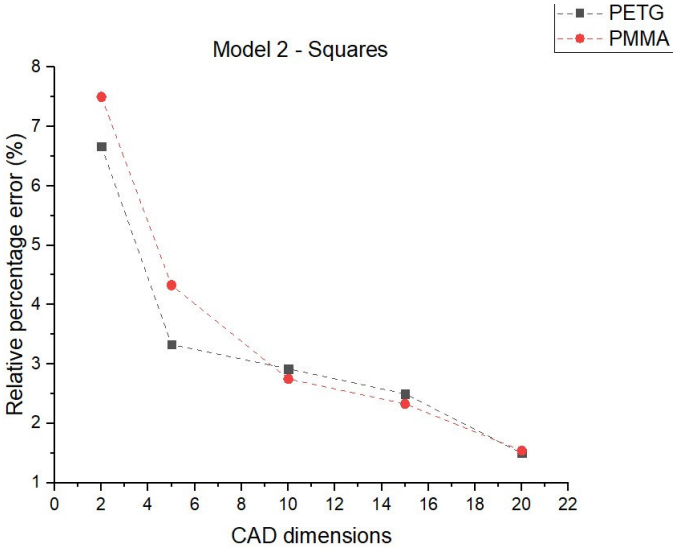


Fig. 4. Distribution curve of relative percentage errors measured using a caliper as a function of the nominal dimensional scale of Model 2 (Squares) across all its dimensions.

The distribution of errors as a function of the designed characteristic size reveals a behavior inversely proportional to the nominal size of the feature. The maximum error percentage peak was invariably concentrated at the smallest nominal dimensions of both phantoms (1.5 mm slits in Model 1, and 2.0 mm edges in Model 2). As the nominal dimension increases toward larger features (such as the 20.0 mm structural base or the 64.0 mm total length), a stable exponential attenuation of the relative percentage error is observed, converging to levels below 1% for both advanced manufacturing methods. Additionally, a singular structural discontinuity behavior (a "step" effect) is noted, located exactly at the 4.0 mm geometric transition point of Model 1.

IV. Discussion

In this work, MRI phantoms were produced using additive (FDM 3D printing) and subtractive (laser cutting) manufacturing techniques. Subsequently, geometric fidelity to the CAD model was analyzed by means of dimensional measurements with a caliper or optical microscopy images in the ImageJ software. The statistical divergence observed between the physical (caliper) and digital (ImageJ) validation approaches in Model 1 (Comb) evidences a metrological attenuation phenomenon arising from digital optical signal processing. Two-dimensional image capture adds an intrinsic background uncertainty, characterized by pixel noise at shadowed edges and by the operator's manual positioning of measurement vectors in ImageJ. This background noise acted as an attenuation filter: it artificially raised the baseline error of the precise PETG parts (from 1.29% to 3.22%), while at the same time masking the severe micro-cutting deviations and physical burrs suffered by the acrylic (PMMA) at the laser cut edges (reducing the apparent error from 8.11% to 5.96%). This finding methodologically validates the use of the physical caliper as the most appropriate technique for high-sensitivity dimensional validation in this study.

On the other hand, the absence of statistical significance observed in Model 2 (Squares), both in the physical scenario ($p = 0.829$) and the digital one ($p = 0.120$), demonstrates that robust, symmetrical hollow volumetric geometries with regular spacing are substantially less susceptible to severe manufacturing distortions. In these structural scenarios, the intrinsic deficiencies of each advanced manufacturing method end up equalizing in terms of overall mean error.

The superior geometric fidelity observed in the fine linear features of Model 1 (Comb) manufactured in PETG by 3D printing (rRMSE of 1.29%), compared with the pronounced deviation of PMMA processed by laser cutting (rRMSE of 8.11%), corroborates the trend described in the literature in other biomedical fields, such as dentistry, in which products can be produced by additive or subtractive methods in digital restorations. Similar dimensional accuracy studies have shown that additive manufacturing methods, such as stereolithography (SLA), frequently achieve significantly smaller mean deviations (RMS) and greater fidelity in complex and marginal details than subtractive processes based on mechanical wear (wax and zirconia milling) or laser ablation. Whereas conventional subtractive manufacturing tends to introduce errors arising from thermal edge effects in micro-cutting features or imperfections induced by drill and mill passes (milling bur passes), additive techniques provide a more controlled and faithful reproduction of geometric profiles relative to the original computational model (CAD). On the other hand, the statistical equivalence observed in the robust, concentric geometries of Model 2 (Squares) between both manufacturing methods in the present study ($p > 0.05$) is also consistent with previous findings in the fixed prosthesis literature, in which no statistically significant differences were detected in the marginal adaptation of robust structural copings when comparing laser sintering (additive) and CAD-CAM milling (subtractive) methods [13, 14, 15].

The macro-mapping of the scale effect validates a classic tendency of the physics of digital polymer manufacturing. The sharp error peak at the smallest nominal dimensions (1.5 mm and 2.0 mm) occurs because small absolute variations in fusion or cutting (on the order of hundredths of a millimeter, inherent to the mechanical resolution of the equipment) carry a greater mathematical weight when normalized by very small design dimensions. When these same absolute variations are diluted across large nominal dimensions (such as 20.0 mm or 64.0 mm), the relative error becomes negligible. This experimentally proves that the final geometric deviation of a phantom does not depend solely and exclusively on the absolute nominal size of the feature but is strongly determined by the volumetric configuration and the geometric neighborhood of the part in the cutting or deposition plane.

V. Conclusion

This work presented the development and dimensional validation of two phantom models intended for future

applications in MRI calibration, manufactured by additive (PETG) and subtractive (PMMA) manufacturing, further comparing two measurement methods, physical caliper and image processing via ImageJ. The results indicate that, for fine linear features, PETG offers greater geometric stability than PMMA, a statistically robust disparity when measured by the physical gold standard but attenuated to the point of not being detected by digital analysis. This reinforces that the chosen validation method is not neutral and can, depending on the processing involved, either artificially raise the error of precise parts or mask real manufacturing defects. For hollow, symmetrical volumetric geometries, on the other hand, dimensional fidelity proved statistically equivalent between the two materials, evidencing that the intrinsic deficiencies of each advanced manufacturing process tend to equalize when the part does not depend on isolated fine features. In both models, the relative error was concentrated at the smallest nominal dimensions and converged to values below 1% for the largest-scale features, indicating that the final geometric deviation of a phantom is determined not only by the absolute size of the feature but also by the volumetric configuration and the geometric neighborhood of the part in the cutting or deposition plane.

Therefore, these results provide an initial answer to the central research question regarding which manufacturing process produces the more dimensionally accurate phantom. Under the conditions evaluated in this study, PETG showed better dimensional accuracy for reproducing fine linear patterns, while PMMA remains a viable alternative for robust volumetric structures, with the additional advantage of lower equipment cost in scenarios where 3D printing is not available. It should be emphasized, however, that this study represents an initial, methodological-development stage, restricted to the dimensional characterization of the phantoms themselves. Thus, the present results allow comparison of the dimensional accuracy of the manufacturing processes but do not yet demonstrate the performance of the phantoms for calibration or geometric distortion control in an actual MRI environment.

Three complementary steps are therefore proposed to consolidate these findings. First, functional validation must be performed by scanning the phantoms on real MRI equipment, across different field-of-view positions and acquisition sequences, in order to quantify the geometric distortion intrinsic to the imaging system itself and not only to the manufacturing process. Second, future work should include multiple phantoms of each model manufactured independently, enabling a proper assessment of inter-piece manufacturing repeatability, in addition to inter-operator agreement in the caliper-based measurements. Third, the inherent resolution limits of the caliper as a gold standard, discussed in Section II.C, should be addressed through complementary high-resolution metrological techniques (e.g. optical or laser scanning) for the smallest evaluated features. Together, these steps would extend the present dimensional validation methodology toward a fully validated protocol for quality control in MRI-guided procedures requiring submillimetric spatial precision.

REFERENCES

[1] T. Yousaf, G. Dervenoulas, and M. Politis, “Advances in MRI methodology,” in *International Review of Neurobiology*, vol. 141. Amsterdam, The Netherlands: Elsevier, 2018, pp. 31–76.

[2] A. S. Minhas and R. Oliver, “Magnetic resonance imaging basics,” in *Electrical Properties of Tissues*, R. Sadleir and A. S. Minhas, Eds. Cham, Switzerland: Springer, 2022, pp. 47–82.

[3] D. Wang, W. Strugnell, G. Cowin, D. M. Doddrell, and R. Slaughter, “Geometric distortion in clinical MRI systems Part I: evaluation using a 3D phantom,” *Magn. Reson. Imaging*, vol. 22, no. 9, pp. 1211–1221, Nov. 2004.

[4] M. O’Donnell and W. A. Edelstein, “NMR imaging in the presence of magnetic field inhomogeneities and gradient field nonlinearities,” *Med. Phys.*, vol. 12, no. 1, pp. 20–26, Jan./Feb. 1985.

[5] M. Armbruster, M. Guba, J. Andrassy, M. Rentsch, V. Schwarze, J. Rübenthaler, T. Knösel, J. Ricke, and H. Kramer, “Measuring HCC tumor size in MRI—the sequence matters!,” *Diagnostics*, vol. 11, no. 11, Art. no. 2002, Oct. 2021.

[6] K. Nousiainen, T. Mäkelä, and J. I. Peltonen, “Characterizing geometric distortions of 3D sequences in clinical head MRI,” *Magn. Reson. Mater. Phys., Biol. Med.*, vol. 35, no. 6, pp. 983–995, Dec. 2022.

[7] M. Barzegar, A. Rostami, A. Y. Mkanna, T. Torfeh, S. P. Paloor, A. Basith, S. Aouadi, R. Hammoud, and N. Al Hammadi, “Characterizing geometrical distortions in MRI for radiotherapy: evaluating MR-SIM protocols and dose accuracy in stereotactic radiosurgery using AAPM TG-284 criteria,” *J. Appl. Clin. Med. Phys.*, vol. 26, no. 10, Art. no. e70199, Oct. 2025.

[8] T. Yousaf, G. Dervenoulas, and M. Politis, “Advances in MRI methodology,” in *International Review of Neurobiology*, vol. 141. Amsterdam, The Netherlands: Elsevier, 2018, pp. 31–76.

[9] B. E. Yunker, K. F. Stupic, J. L. Wagner, S. Huddle, R. Shandas, R. F. Weir, S. E. Russek, and K. E. Keenan, “Characterization of 3-dimensional printing and casting materials for use in magnetic resonance imaging phantoms at 3 T,” *J. Res. Natl. Inst. Stand. Technol.*, vol. 125, Art. no. 125028, 2020.

[10] Y. Habeeb, P. E. Zorn, F. Blindauer, G. Bierry, S. Kremer, and J. P. Dillenseger, “High-fidelity anatomical phantoms for MRI practical training,” *Physica Medica*, vol. 127, Art. no. 104832, Nov. 2024.

[11] H. Li, A. C. Özen, A. Juerchott, M. Breckwoldt, J. Jesser, D. F. Vollherbst, D. Schwarz, M. Bendszus, S. Heiland, and T. Hilgenfeld, “3D quantification of metal-induced geometric distortions in MRI,” *Sci. Rep.*, vol. 15, Art. no. 7195, Feb. 2025.

[12] *ImageJ*, National Institutes of Health, Bethesda, MD, USA. [Online]. Available: https://imagej.net/ij/

[13] P. Kharade, K. Dholam, and A. Gorakh, “A technique for fabrication of cranial prostheses using high-temperature vulcanizing silicone material,” *J. Prosthet. Dent.*, vol. 118, no. 1, pp. 113–115, Jul. 2017.

[14] F. G. Mangano, D. Cianci, N. Pranno, H. Lerner, F. Zarone, and O. Admakin, “Trueness, precision, time-efficiency and cost analysis of chairside additive and subtractive versus lab-based workflows for manufacturing single crowns: an in vitro study,” *J. Dent.*, vol. 141, Art. no. 104792, Feb. 2024.

[15] K. Sathish, S. S. Kumar, R. T. Magal, V. Selvaraj, V. Narasimharaj, and R. Karthikeyan, “A comparative study on subtractive manufacturing and additive manufacturing,” *Adv. Mater. Sci. Eng.*, vol. 2022, Art. no. 6892641, 2022.

Heart Rate Calibration in Wearables via Biometric Signal Mimicry

1st Fabrício Gonçalves Torres
Laboratório de Metrologia Elétrica
Instituto de Pesquisas Tecnológicas
São Paulo, Brazil
fabrigt@ipt.br

2nd Ryan Wictky Sallatti
Laboratório de Metrologia Elétrica
Instituto de Pesquisas Tecnológicas
São Paulo, Brazil
ryansallatti@ipt.br

3rd Ernesto Colombelli Agostini
Laboratório de Metrologia Elétrica
Instituto de Pesquisas Tecnológicas
São Paulo, Brazil
colombelliagostini.ernesto@gmail.com

Abstract*—Wearables are devices such as smartwatches that integrate sensors for the continuous monitoring of physiological parameters, playing an auxiliary role in the early detection of health deviations. This study explores the effectiveness of photoplethysmography (PPG) sensors in heart rate measurement under the IEC 63203-402-3 standard, while addressing the gap in technical information provided by manufacturers. Through a proof of concept of optical mimicry using infrared LEDs, the response of different smartwatch models to simulated signals was tested. The results demonstrate the feasibility of the proposed approach for the evaluated conventional devices and indicate its potential for identifying hardware anomalies, while also revealing limitations in devices incorporating skin-impedance detection. It is concluded that the lack of technical transparency from manufacturers poses significant challenges to the standardization and calibration of these wearables.

***Index Terms*—wearables, smartwatch, photoplethysmography, heart rate, calibration**

I. INTRODUCTION

A smartwatch can measure heart rate through transmission or reflection photoplethysmography (PPG). Transmission PPG is used in finger pulse oximeters to measure oxygen saturation and heart rate through direct transmission of red and infrared light that passes through the finger and reaches a photodiode on the other side. This method, commonly used in hospitals and health centers, has the advantage of reading both beats per minute (BPM) and oxygen saturation.

Reflection PPG uses a photodiode on the same side as the LED transmitter, reaching the blood flow on the surface of the skin and reflecting the signal back to the sensor. This configuration enables PPG measurements in compact wearable devices and is therefore commonly used in smartwatches, including during physical activity. However, reflected PPG signals can be significantly affected by motion artifacts. Heartbeat measurements are obtained through the variation in light attenuation in the receiver. This variation stems from the dilation and contraction of blood vessels, which occur in synchronization with each heartbeat.

Green light is strongly modulated by pulsatile changes in blood volume because hemoglobin exhibits significant absorption in this spectral region. Its relatively shallow penetration depth in biological tissue makes it particularly sensitive to superficial vascular changes, which are relevant for heart-rate monitoring at the wrist (Fig. 1).

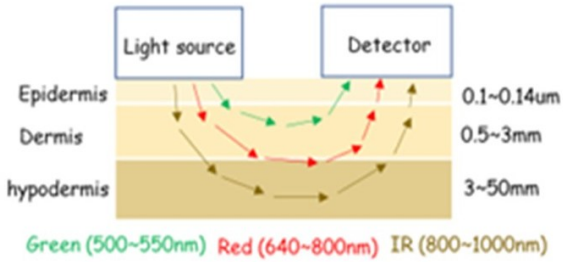


Fig. 1. Light transmission based on wavelength, Source: [3]

Several physiological and environmental factors can affect wrist-worn PPG measurements, including motion artifacts, ambient light, peripheral perfusion, sensor-to-skin contact, and skin optical properties. The influence of skin tone on heart-rate measurement accuracy remains under investigation, with the literature reporting device- and condition-dependent effects [2]. These factors are particularly relevant in clinical or real-world validation, but they are outside the primary scope of the present study. Here, the objective is to isolate the wearable's response to a controlled optical stimulus, reducing physiological and environmental variability in order to investigate the feasibility of a traceable calibration approach at the sensor level.

The IEC 63203-402-3 standard [1] seeks to standardize how heart rate tests are performed and reported, requiring manufacturers to provide detailed information about the

hardware and software of the devices. The goal is for results to be fairly and reproducibly comparable. However, in practice, a lack of transparency is observed, where manufacturers omit critical parameters to maintain confidentiality and prevent competitors from attempting to reproduce their technologies. This information gap reinforces the importance of exploring tests and methodologies to validate the integrity and reliability of wearables.

Recent studies have proposed optical PPG simulators capable of providing repeatable optical signals with controlled characteristics for the characterization and comparison of photoplethysmographic devices [9]. Such approaches demonstrate the potential of artificial optical stimulation for controlled PPG testing. The method investigated in this work follows a related principle but focuses on the practical calibration of commercial wearables and on the synchronization and optical-coupling requirements necessary for the simulated pulsation to be interpreted by proprietary smartwatch algorithms. A direct performance comparison with commercial PPG simulators was not performed in this proof-of-concept study and remains an important topic for future investigation.

Charlton et al. [10] demonstrated that wrist PPG signal quality depends not only on sensor characteristics but also on physiological and measurement conditions, including posture, sensor position, peripheral perfusion, skin optical properties, and LED intensity. These factors reinforce the importance of distinguishing controlled sensor-level calibration from clinical or real-world validation.

II. METHODOLOGY

The experiment was designed as a proof of concept (PoC) to verify the feasibility of mimicking biometric signals via determining the technical requirements necessary for an artificial signal to be processed by the smartwatch's algorithms as a real physiological pulsation.

The light source was controlled via a function generator calibrated in an accredited laboratory, configured with the following parameters:

- Signal Amplitude: 5 V peak-to-peak.
- Sine wave form.
- Pulsation Frequency: 0.5 Hz to 3.5 Hz as established by the IEC 63203-402-3 standard.
- The function generator was externally referenced to a Rubidium Frequency Standard (Model 910R), whose measurement uncertainty is 4×10^{-11}.

Although green light is commonly used in wrist-worn PPG because of its interaction with superficial blood-volume changes, silicon photodiodes are also responsive to near-infrared radiation. The purpose of using an IR LED in the present setup was not to reproduce the wavelength normally employed by the smartwatch, but to determine whether a controlled periodic optical stimulus within the photodetector's spectral response could be interpreted by the device as a pulsatile signal. The successful detection of the IR stimulus by Watches A1–A3 supports this feasibility for the evaluated devices; however, it does not imply wavelength-independent photodiode responsivity.

The test was performed on 5 watches, 3 of the same model (A1, A2, and A3), and two of different models (B and C). Watches A1, A2, and A3 are less advanced and do not have protective barriers against external interference. Watch B, although it has no protection systems, is more modern than the previous ones. Watch C, on the other hand, has an impedance sensor that detects if the watch is in contact with human skin. For this watch, it would be required to develop an electronic system that could simulate human skin impedance and, additionally, improve the synchronization of the pulses generated by the standard function generator with the pulses of the watch under calibration.

A. Experimental Procedure and Repeatability

The five devices were selected based on their availability for this proof-of-concept study, as the acquisition of a broader range of commercial smartwatch models was beyond the resources available for the experimental campaign. Although this convenience-based selection limits the representativeness of the sample, the availability of three units of the same model (A1, A2, and A3) allowed the consistency of the proposed method to be evaluated across nominally equivalent devices. Watches B and C, in turn, provided additional observations involving different hardware characteristics, including the skin-contact detection mechanism identified in Watch C.

At each calibration point, four repeated measurements (n = 4) were performed under the same experimental conditions. Each measurement was recorded only after the smartwatch displayed a stable heart-rate indication. When no stable indication could be obtained at a given calibration point, the result was classified as unavailable (N/A). The mean of the valid measurements was used as the reported heart-rate value, while their dispersion was used to estimate the Type A uncertainty. The optical coupling arrangement was maintained unchanged throughout the repetitions, with the infrared LED positioned inside the optical coupler at a fixed position relative to the smartwatch sensor,

minimizing variations in alignment and exposure to ambient light.

The combined uncertainty was estimated by evaluating Type A and Type B uncertainties. The Type A uncertainty (uA), representing statistical variations, was calculated using (1), where s is the standard deviation and n = 4 is the number of repeated measurements performed at each calibration point.

$$u_A = \frac{s}{\sqrt{n}} \quad (1)$$

The Type B uncertainty (u_B) associated with the smartwatch resolution was estimated assuming a rectangular distribution, as defined in (2). The error of the Rubidium Frequency Standard (Model 910R) is 4×10^{-11}, which is negligible.

$$u_B = \frac{Resolution}{\sqrt{3}} = 0.29\ \text{BPM} \quad (2)$$

Finally, both components were aggregated to determine the combined uncertainty (uc) using the root sum of squares method, as shown in (3).

$$u_C = \sqrt{(uA^2 + uB^2)} \quad (3)$$



Fig. 2. Test setup for 60BPM. Source: Elaborated by the authors.

The experimental setup followed the principle of the PPG simulator test proposed in the IEC 63203-402-3 standard [1], using an infrared (IR) emitting LED as the simulated signal source. To ensure beam integrity and minimize light dispersion, the LED was inserted into an optical coupler designed to accurately direct the photon flux (Fig. 2). This allowed the sensor to be isolated from environmental interference and ensured that the captured intensity variation came exclusively from the controlled source.

III. RESULTS

To ensure methodology compatibility with the synchronization triggers of modern high-sensitivity PPG sensors, such as the MAX30101/MAX30102 [6], [7] or AFE4950 [8], an alignment circuit utilizing an MCP6002 operational amplifier was implemented (Fig. 3).

While the amplifier's finite slew rate causes the output signal to assume a triangular waveform at extreme frequencies above 100 kHz (Fig. 4), the circuit operates without significant waveform distortion within the low-frequency range required for heart-rate calibration.

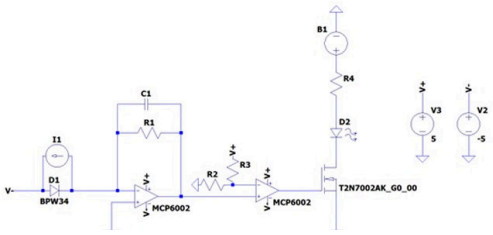


Fig. 3. Simulation of the heartbeat synchronization and mimicry circuit. Source: Elaborated by the authors.

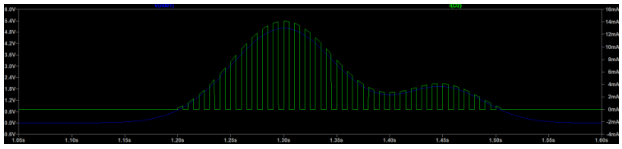


Fig. 4. System response at 100 kHz, showing the distortion of the current signal. Source: Elaborated by the authors.

Table I presents the results obtained in the calibration of the heartbeats measured by watches A1, A2, A3, B, and C. The reported value for each calibration point represents the mean of four repeated measurements, while the associated combined uncertainty was calculated according to Equation (3). According to IEC 63203-402-3 [1], heart-rate accuracy is evaluated using the Mean Absolute Percentage Error or MAPE (4), where HRi is the heart-rate indication provided by the wearable for the i-th measurement, HRref,i is the corresponding reference heart rate derived from the frequency applied by the simulator, and N is the total number of measurements considered in the calculation. Considering the four repeated measurements performed at each of the seven calibration points, a total of 28 measurements were evaluated for each device. The resulting MAPE values were 1.3%, 1.3%, and 1.7% for Watches A1, A2, and A3, respectively. All values were below the 10% acceptance criterion considered from IEC 63203-402-3,

indicating satisfactory performance of the three devices under the controlled conditions evaluated in this proof-of-concept study.

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{HR_i - HR_{Ref,i}}{HR_{Ref,i}}\right| \quad (4)$$

TABLE I
RESPONSE OF DEVICES TO PULSATION MIMICRY VIA INFRARED EMISSION (RESULTS ARE EXPRESSED AS MEAN ± COMBINED STANDARD UNCERTAINTY)

Expected (bpm)	Dev A1 (bpm)	Dev A2 (bpm)	Dev A3 (bpm)	Dev B (bpm)
30	30,3 ± 0.9	29.8 ± 0.7	31.0 ± 0.8	N/A
60	60,3 ± 0.6	59.8 ± 0.8	59.3 ± 0.7	59.8 ± 0.9
90	89,8 ± 0.7	88.5 ± 0.6	90 ± 1	N/A
120	119,8 ± 0.9	119.8 ± 0.9	118.5 ± 0.4	121.3 ± 0.6
150	151,3 ± 0.6	151.3 ± 0.6	149.8 ± 0.7	N/A
180	180.8 ± 0.6	180.8 ± 0.6	180.5 ± 0.6	N/A
210	209.3 ± 0.7	209.3 ± 0.7	210 ± 1	N/A

Watches A1, A2, and A3 exhibited stable responses across all evaluated calibration points. The measured values remained close to the expected heart-rate values, indicating consistent response to the controlled optical stimulus among the three nominally equivalent devices.

Watch B: No signal was detected at several calibration points. Subsequent technical analysis revealed probable damage to the receiving photodiode, making it impossible to convert the photon flux into electrical voltage, regardless of the external source's intensity. Despite the hardware failure, stable readings were obtained at specific frequency points such as 60BPM and 120BPM.

Watch C: The device prevented data collection due to the presence of a skin impedance sensor. This mechanism acts as a safety lock that requires electrical contact with human skin to activate the optical reading stages, demonstrating that mimicry in devices with this security feature requires the simultaneous simulation of other electrical biometric properties.

IV. DISCUSSION

The experimental results indicate that optical mimicry is feasible for the evaluated devices without impedance-based presence detection. Watches A1, A2, and A3 demonstrated an almost immediate response to the infrared optical pulses, converting the frequencies applied by the standard function generator within the maximum permissible error established by the IEC 63203-402-3 standard. This result supports the feasibility of infrared optical mimicry for the evaluated models, although it should not be interpreted as evidence of wavelength-independent photodetector responsivity.

The calibration process also indicated a possible hardware anomaly in Watch B. The absence of stable readings at several calibration points was consistent with a probable failure in the optical receiving stage, possibly associated with the receiving photodiode. This observation suggests that the proposed setup may also be useful for identifying abnormal sensor behavior during calibration.

Finally, the barrier encountered in Watch C highlights the evolution of security mechanisms. The requirement for galvanic skin impedance detection to activate the PPG stage acts as a logic lock. It is concluded that, for the calibration of devices in this segment, it became necessary to develop a hybrid simulator capable of simultaneously providing the optical stimulus and emulating the electrical impedance of the skin-device interface.

A. Limitations and Metrological Implications

The results should be interpreted within the scope of a proof-of-concept study. The experiments were performed under controlled optical coupling conditions and were intended to evaluate the response of the wearable to a controlled artificial optical stimulus with traceable pulsation frequency rather than to reproduce the full optical and physiological behavior of human tissue. Consequently, factors that affect PPG performance in real-world use, including motion artifacts, skin optical properties, low peripheral perfusion, sensor pressure, and variations in ambient light, were not reproduced in the present setup.

From a metrological perspective, this distinction is important. Calibration using a controlled optical stimulus can evaluate the relationship between an applied pulsation frequency and the heart-rate indication generated by the wearable, while clinical validation evaluates device performance under physiological and environmental variability. Therefore, satisfactory performance under the proposed calibration conditions should not be interpreted as evidence of equivalent accuracy during unrestricted use or under all clinical conditions.

The present study was also limited to five devices and three different models. The results therefore cannot be generalized to all wearable PPG architectures. Further validation should include a larger number of devices and manufacturers, as well as controlled variations in LED wavelength, optical intensity, coupling geometry, waveform shape, ambient-light conditions, and synchronization strategy. Comparison with commercial

PPG or physiological simulators would also provide an important benchmark for assessing the advantages and limitations of the proposed approach.

Furthermore, recent calibration approaches have demonstrated that adaptive sensor parameters, such as optical gain, may be used to mitigate PPG response variability associated with skin pigmentation [11]. This reinforces that the controlled optical calibration proposed here represents only one component of the broader validation required for wearable PPG systems.

V. CONCLUSION

The experiments demonstrated the feasibility of optical signal mimicry as a proof-of-concept approach for heart-rate calibration and integrity verification in the evaluated wearable devices. Watches A1–A3 produced heart-rate indications within the acceptance criterion considered from IEC 63203-402-3 under the controlled experimental conditions. Watch B exhibited anomalous behavior associated with hardware damage, whereas Watch C could not be evaluated using optical stimulation alone because its operation requires skin-contact validation. These results indicate the potential of the proposed method for controlled wearable calibration, but further experimental validation involving a broader range of devices and sensing architectures is required.

The lack of transparency by manufacturers proved to be a significant obstacle, as hardware configurations remain undisclosed. Access to specific sensor information and internal barrier mechanisms is essential to optimize calibration and facilitate objective device benchmarking. Therefore, practical experimentation remains vital to bridge these informational gaps and identify the operational variables of these devices.

VI. FUTURE WORKS

Future work will focus on the calibration of devices such as Watch C, which require skin impedance detection for operation. To address this, we propose an impedance simulation circuit based on wrist tissue parameters described by Bora and Dasgupta [4]. Additionally, we are developing a coupling mechanism to integrate these circuits with the device under test while minimizing external interference. This mechanism will incorporate a translucent silicone interface to function as an attenuating filter. By preventing photodiode saturation and enhancing light scattering, this layer will more accurately simulate the optical response of human tissue.

REFERENCES

[1] International Electrotechnical Commission (IEC), IEC 63203-402-3:2024: Wearable electronic devices and technologies – Part 402-3: Performance measurement of fitness wearables – Test methods for the determination of the accuracy of heart rate, Edition 1.0. Geneva: IEC, 2024.

[2] D. Koerber, S. Khan, T. Shamsheri, A. Kirubarajan, and S. Mehta, “Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: a systematic review,” J. Racial Ethn. Health Disparities, 2022. doi: 10.1007/s40615-022-01446-9.

[3] K. B. Kim and H. J. Baek, “Photoplethysmography in wearable devices: A comprehensive review of technological advances, current challenges, and future directions,” Electronics, vol. 12, no. 13, p. 2923, Jul. 2023.

[4] N. Kelly, B. Santoyo, and C. Lin, “Pulse oximeter calibration study,” Phys 371: Design Like a Physicist, 2022.

[5] D. J. Bora and R. Dasgupta, “Estimation of skin impedance models with experimental data and a proposed model for human skin impedance,” IET Syst. Biol., vol. 14, no. 5, pp. 230–240, Oct. 2020.

[6] Analog Devices, “MAX30101: High-sensitivity pulse oximeter and heart-rate sensor for wearable health,” Maxim Integrated. [On-line]. Available: https://www.analog.com/en/products/max30101.html. Accessed: Mar. 17, 2026.

[7] Analog Devices, “MAX30102: High-sensitivity pulse oximeter and heart-rate sensor for fitness & healthcare,” Maxim Integrated. [On-line]. Available: https://www.analog.com/en/products/max30102.html. Accessed: Mar. 17, 2026.

[8] Texas Instruments, “AFE4950: Ultra-small integrated AFE for wearable optical heart-rate monitoring and SpO2.” [Online]. Available: https://www.ti.com/product/AFE4950. Accessed: Mar. 17, 2026.

[9] M. Caruso et al., “An Optical Signal Simulator for the Characterization of Photoplethysmographic Devices,” Sensors, vol. 24, no. 3, 1008, 2024. DOI: 10.3390/s24031008.

[10] P. H. Charlton, V. Marozas, E. Mejía-Mejía, P. A. Kyriacou, and J. Mant, “Determinants of photoplethysmography signal quality at the wrist,” PLOS Digital Health, vol. 4, no. 6, e0000585, 2025. DOI: 10.1371/journal.pdig.0000585.

[11] C. MacIsaac, M. Nguyen, A. Uy, T. Kong, and A. Hedayatipour, “A Programmable Gain Calibration Method to Mitigate Skin Tone Bias in PPG Sensors,” Biosensors, vol. 15, no. 7, Art. no. 423, 2025, doi: 10.3390/bios15070423.

Impact of cold head replacement on the constancy of quality control parameters in magnetic resonance imaging: A case study

Thiago Michel de Brito Farias
Dept. of diagnostic Imaging
Escola Paulista de Medicina
Federal University of São Paulo
UNIFESP
São Paulo, SP, Brazil
thiago.michel@unifesp.br

Paula Regina Rodriguez Coelho
Dept. of diagnostic Imaging
Escola Paulista de Medicina
Federal University of São Paulo
UNIFESP
São Paulo,SP, Brazil
paularrodriguez@gmail.com

Camila Hitomi Murata
Dept. of diagnostic Imaging
Escola Paulista de Medicina
Federal University of São Paulo
UNIFESP
São Paulo, SP, Brazil
camila.murata@unifesp.br

Benedito Herbert de Souza
Dept. of diagnostic Imaging
Escola Paulista de Medicina
Federal University of São Paulo
UNIFESP
São Paulo, SP, Brazil
Prof.herbertsouza@gmail.com

Kellen Adriana Curci Daros
Dept. of diagnostic Imaging
Escola Paulista de Medicina
Federal University of São Paulo
UNIFESP
São Paulo, SP, Brazil
daros.kellen@unifesp.br

***Abstract*—Cold head devices are essential for maintaining superconductivity and field homogeneity in modern MRI equipment. This study analyzed the impact of coldhead replacement on the quality control (QC) parameters of a 3.0T MRI, evaluating system stability before and after the component exchange. The results showed that wear of the cold head led to instability in helium pressure, a drop in the signal-to-noise ratio, slice thickness deviations, and an increase in ghosting artifacts. Following the intervention, the magnetic field was stabilized and image quality parameters returned to ideal standards with the recalibration of the center frequency. Thus, it can be concluded that periodic monitoring of QC data validates the effectiveness of predictive maintenance in preventing critical failures, allowing potential failures to be observed through fluctuations in periodic data.**

Keywords—Magnetic resonance, cold head, quality control.

I. INTRODUCTION

In recent decades, the evolution of magnetic resonance imaging scanner has enabled the development of new technologies, highlighted by scanners with more intense magnetic fields based on superconductivity principles. On the other hand, these systems have become more sensitive to small fluctuations in environmental parameters and physical hardware components. Scanners with fields of 3.0 T or higher, for example, utilize ultra-fast gradients (slew rates) that exhibit high susceptibility to external disturbances [1, 2].

The operation of Magnetic Resonance (MR) systems with 3.0 Tesla (3.0 T) magnets imposes rigorous thermodynamic challenges due to the amount of energy stored in the superconducting solenoid. Maintaining the superconductivity of Niobium-Titanium alloys (NbTi) requires the filament temperature to remain strictly below its critical temperature ($T_c = 9.2$ K), being immersed in liquid helium at approximately 4.2 K under normal atmospheric pressure. [3], [4]

In this scenario, the natural evaporation rate of helium is intensified by the heat generated from the activation of magnetic field gradients and radiofrequency (RF) pulses. To prevent this loss, Zero Boil-Off (ZBO) technology is employed, in which the gaseous helium evaporating inside the cryostat is captured and re-condensed into a liquid state via a mechanical cryocooler, returning to the reservoir in a closed loop [3],[4].

The cold head is the component of the cryocooler responsible for providing the thermal exchange required for this liquefaction. Thus, ZBO technology plays a crucial role by eliminating continuous cryogen loss, preventing substantial financial losses and the imminent thermal instability of the magnet [4], [6].

The performance and diagnostic accuracy of high-resolution MRI systems directly depend on the mechanical and thermal stability of the cold head and the cryostat. While cycles, such as the pulse tube, minimize vibrations and microphonics by having no moving parts in the cold head, systems based on mechanical pistons — such as the Gifford-McMahon (GM) — require constant monitoring against micro-frictions that generate thermal instabilities and radiofrequency noise (spike noise). [6],[8]

Furthermore, strict thermal control on the millikelvin scale is essential to prevent micro-drifts in the static magnetic field (B_0) and oscillations in the resonance frequency, factors that degrade field homogeneity and the image signal-to-noise ratio (SNR).[6],[8],[9]

Although modern scanners perform automated calibrations before and during each pulse sequence (such as real-time active shimming and dynamic center frequency adjustment), severe deviations in B_0 homogeneity still generate

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

macroscopic geometric distortion artifacts and chronic failures in the chemical shift fat saturation (fat-sat) technique [8]. Echo-Planar Imaging (EPI) based sequences, such as Diffusion-Weighted Imaging (DWI), suffer severe phase distortions if the internal thermal gradient fluctuates, causing displacement and c artifacts in the phase-encoding direction. [8], [9], [10].

In clinical engineering, the integrity of magnetic resonance imaging (MRI) systems depends on the integrated operation between the cold head, the compressor, and the chiller. Currently, preventive and predictive monitoring of these components relies on telemetry sensors, helium pressure analysis, and vibrational signatures to anticipate mechanical failures [8], [13]. The failure of these cooling systems raises the thermal energy of the cryostat, which can induce a quench, causing the sudden loss of superconductivity along with the violent expansion of liquid helium [13], [14], [15], [16], [17], [18].

Given this scenario and the need for metrological validation, this work aims to experimentally analyze the impact of cold head (helium level and pressure levels) replacement on fundamental image quality parameters, specifically central frequency, slice thickness, SNR, and ghosting factor.

II. Materials and Methods

This study consists of a longitudinal analysis conducted before and after the replacement of the cold head (a system based on Zero Boil-Off technology). Image data were acquired on a 3.0 T magnetic resonance imaging (MRI) scanner (Achieva, Philips Healthcare) using an 8-channel head coil (Model SENSE head coil 8ch) and the Magphan® RT phantom (Model SMR170). The quality control (QC) protocol utilized spin-echo sequences with the following parameters: TR = 500 ms, TE = 20 ms, a slice thickness of 5.0 mm, a 256 × 256 pixel matrix, a FOV of 256 x 256 mm² , and a bandwidth of 364 Hz/pixel. Additionally, the physical operating parameters of the equipment—specifically internal pressure and liquid helium level—were monitored weekly.

Signal-to-noise ratio (SNR), slice thickness, and ghosting artifact metrics were calculated from the phantom images by positioning regions of interest (ROIs), following the guidelines of the established QC protocol. For SNR quantification, a circular ROI was placed at the center of the phantom's homogeneous region to obtain the mean signal intensity ($\bar{S}$). Simultaneously, elliptical ROIs were distributed in the background region (air), outside the object's perimeter, to determine the background noise (σ_{bkg}) (equation 1).

$$SNR = \frac{0.655 \times \bar{S}}{\sigma_{bkg}} \qquad (1)$$

The evaluation of the ghosting parameter (PSG) was performed using the same background elliptical ROIs, positioned in the peripheral directions (top, bottom, right, and left) relative to the phantom. The mean intensities of these regions were compared against the mean intensity of the phantom's central ROI, allowing the calculation of the ghosting factor (Figure 1b, equation 2).

$$PSG = 100 \times \frac{(Left+right)-(Top+Bottom)}{2 \times \bar{S}} \qquad (2)$$

Slice thickness was determined by analyzing the specific ramp profiles of the phantom for this test by full width at half maximum (FWHM), as specified by the phantom manufacturer. Initially, the maximum and minimum pixel intensity values in the x and y directions were mapped to calculate and adjust the optimal window level (WL). Following segmentation by the WL, the boundaries of the ramp structure were defined for the execution of geometric measurements. Finally, the obtained linear dimensions were applied to the mathematical formulations of the QC protocol to determine the effective slice thickness (Figure 1c).

The central frequency was obtained directly from the DICOM image metadata, using the imaging frequency information found in the file header.

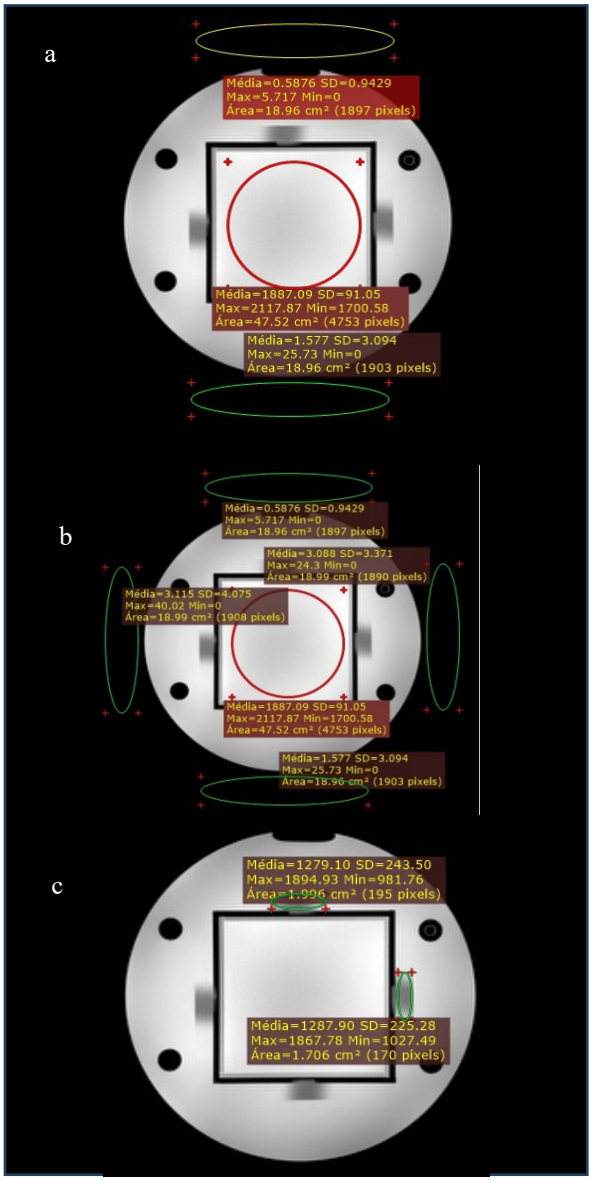


Figure 1(a) signal-to-noise measurements, (b) ghosting measurements, and (c) slice thickness.

The quantitative analysis was performed using RadiAnt software, with subsequent tabulation and data constancy analysis conducted in Microsoft Excel [19].

The measurements were compared before and after the replacement carried out on February 4, 2026.

III. Results and Discussion

To evaluate system constancy in an integrated manner, the image quality control test results were correlated with the historical physical parameters of the equipment. For this purpose, time-trend plots of internal pressure and liquid helium level were generated and detailed, making it possible to verify whether any variations in image quality settings showed a direct correlation with the behavior of the cryogenic system pre-and post-intervention (Figure 2).

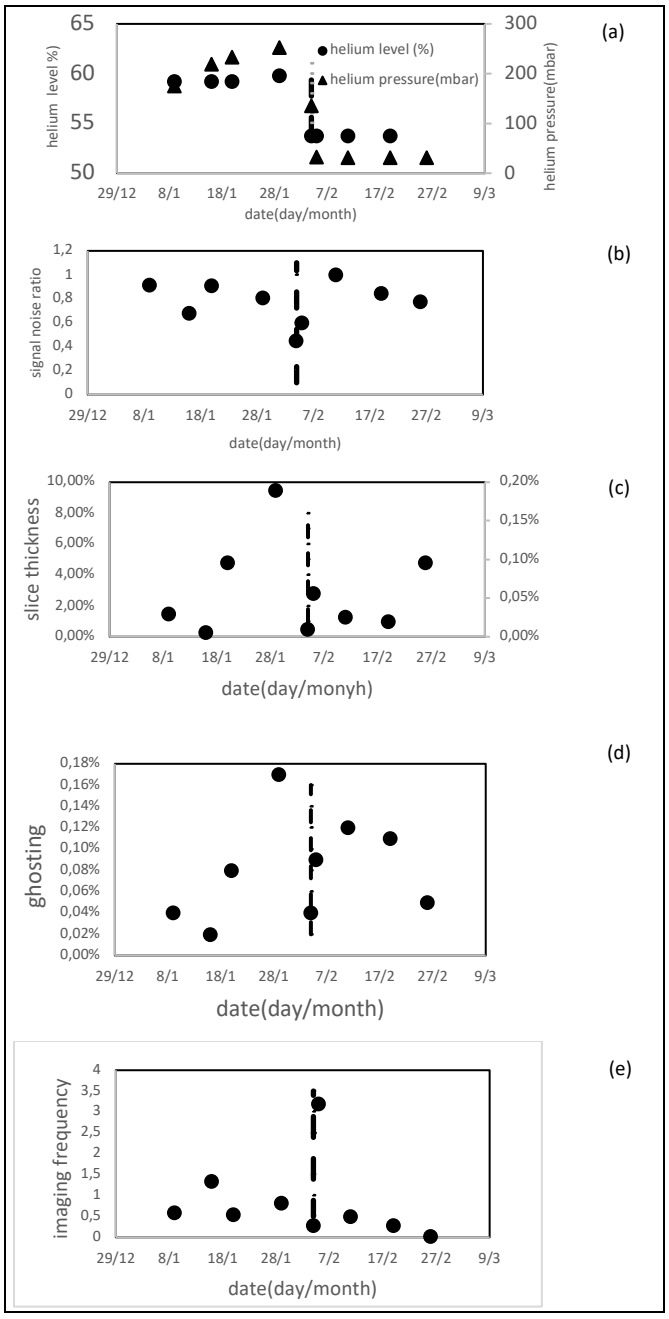


Figure 2:Timeline of 9 measurements pre- and post- intervention (dashed vertical line);(a)Helium level and pressure stability over time;);(b)signal noise ratio;(c)slice thickness;(d) ghosting; imaging frequency

Figure (2a) highlights an instability in the helium pressure, indicating that the old cold head was no longer recondensing the gas adequately. After the component was replaced, the pressure stabilized at lower levels. This thermal stabilization halts fluctuations in the helium evaporation rate and, consequently, stabilizes the magnetic field—a process that, however, requires at least a week to fully stabilize, as observed in the data obtained.[13]. Furthermore, the data obtained are consistent with the manufacturer's recommended limits (30% helium level and 30 mBar helium pressure). [4] [20]

For the parameters shown in Figures 2b to 2d, the tolerance levels recommended by the Normative Instruction IN 97 of the Brazilian Health Regulatory Agency (*Agência Nacional de Vigilância Sanitária - ANVISA*) were applied. [21]

The analysis of the SNR data shown in Figure (2b) can be directly correlated to how thermal and pressure instability affects the homogeneity of the B0 field. It is noted that the SNR had a downward trend and reached a peak just before the replacement, which suggests that a failing coldhead generates mechanical vibrations and thermal fluctuations that will interfere with signal quality.

PSG in MRI is also related to possible mechanical vibrations caused by wear of the cold head and signal phase instabilities. Figure (2d) shows that before the replacement, the value was high, reaching a peak in January. Although there was a short delay before full stabilization was achieved, the general trend after the replacement is a decrease and control of the artifact, returning to low levels by the end of the analyzed time series [13],[18],[19].

Figures (2c) and (2e) show the slice thickness and imaging frequency parameters, respectively. These parameters depend on the stability and calibration of the gradients, as well as the resonance frequency. They are affected by the temperature and pressure of the magnet. It is noted that, most of the time, the slice thickness varied within a low range (usually below 5%), but an isolated peak occurs where the deviation reaches 9.5%, and after the replacement, the slice precision returns to the expected values. It is noted that exactly after the cold head replacement, there are peaks in the Imaging Frequency graph, which is acceptable and described in the literature. Following the replacement, the pressure system stabilizes, and the equipment needs to undergo a center frequency recalibration to adapt to the new stable condition of the magnetic field.

The data showed an end-of-life scenario for a cold head, which presented helium pressure instability that, consequently, likely generated noise, SNR decrease, and artifacts. The replacement intervention allowed the image quality (SNR and slice thickness) to be restored to ideal operating standards

The analysis of the graphs prior to the maintenance and cold head replacement date offers a case study for the development of predictive models in MRI systems. Based on the evaluation of the obtained measurements, it can be verified that despite having high values immediately before the cold head replacement, only the center frequency presented values above the reference parameters of the current national regulations.

Visual inspection of the phantom images on a medical diagnostic display revealed no perceptible differences pre- and post- intervention during quality control tests.

Differences between pre- and post-intervention phases were evaluated using the two-tailed Wilcoxon signed-rank exact test.

Due to the small sample size, the statistical power was limited, preventing the detection of subtle differences between phases (*p-value*>0.75). Nevertheless, a clear trend toward stabilization and a return to normative tolerance parameters was observed, highlighting a meaningful operational impact. Future studies with larger cohorts are recommended to confirm these findings

The cold head replacement was effective in reducing noise in the MRI images of the analyzed scanner, restoring image quality parameters (SNR and slice thickness) to ideal operating limits. Furthermore, the observed trends during the parameter fluctuation period demonstrate the potential of using these historical data to develop predictive maintenance models, preventing unplanned system downtime through periodic quality control (QC) analysis. Although high fluctuations were recorded immediately prior to the component replacement, only the center frequency exhibited values exceeding the reference parameters of the current national regulation.

ACKNOWLEDGMENT

We would like to thank the Financiadora de Estudos e Projetos (FINEP) of Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTI) and Fundação Instituto de Pesquisa e Estudo de Diagnóstico por Imagem (FIDI) for providing the resonance Imaging scanner. We would also like to thank the technicians who helped us to create protocols for this study.

REFERENCES

[1] Westbrook, Catherine; Talbot, John. Ressonância Magnética: Aplicações Práticas. 5. ed. Rio de Janeiro: Guanabara Koogan, 2019

[2] Gifford, W.E.- The Gifford-McMahon Cycle. Advances in Cryogenic Engineering ,V.11. Springer, Boston,152-159, 1966.

[3] Mahesh M. The Essential Physics of Medical Imaging. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2013.

[4] Philips Healthcare. Achieva 3.0T System Technical Description a nd Pre-installation Manual. Best, The Netherlands. Philips Medical Systems , 2015.

[5] Hashemi, R. H., Hashman R, Bradley Jr, The Basics. 3rd ed. New York: Lippincott Williams & Wilkins, 2010.

[6] Radebaugh, R. Refrigeration for superconductors. IEEE Transactions on Applied Superconductivity, v. 92, n. 10, p. 1719-1734. 2004.

[7] Sumitomo Heavy Industries. Operating Manual for SHI Cryocooler and Compressor Series. Tokyo: SHI Cryogenics Group, 2020.

[8] Brown R.W., Cheng Y.C,N, Haacke, E. M., Thompson M.R, Venkatesan R. Magnetic Resonance Imaging: Physical Principles and Sequence Design. New York: John Wiley & Sons, 2014.

[9] Mcrobbie, D. W., Moore E.A., Graves M.J., Prince M.R- MRI from Picture to Proton. 2nd ed. Cambridge: Cambridge University Press, 2007.

[10] Smith, M. A. The techonology of magnetic resonance imaging. Clinical Radiology, 36(6),553-559. 1985.

[11] Melz, T., Melcher J., Matthias M. Active Reduction of Cryocooler Induced Vibrations Using Smart Materials and Adaptive Control. Int.Cong on Sound and vibration. 4-7.07.2000

[12] Philips Healthcare- Cryogenics Management and Maintenance Guide for Achieva Systems. Philips, 2018.

[13] Philips Healthcare- Achieva 3.0T Service and Field Engineering Manual. Philips, 2017.

[14] Leybold GmbH. Fundamentals of Vaccum Technology. Cologne: Leybold GmbH, 2019.

[15] Safety Committee.Physics Division Cryogenic Safety manual. Argonne National Laboratory, 2020.

[16] Steckner, M. Hazards associated with superconducting magnet quench events: System requirements, aplication and consequences.Magma, 2026.

[17] Montoya, R.A., Delgado, S.,Castilla, J., Navarrete J., Contreras N.D., Marijuan, J.R., Barrena V., Guillamón I., Suderow H., Methods to simplify cooling of liquid Hellium cryostats. HardwareX V.5, 2019

[18] International Electrotechnical commission IEC 60601-2-33: Medical electrical equipment - Part 2-33: Particular requirements for the basic safety and essential performance of magnetic resonance equipment for medical diagnosis. Geneva: IEC, 2022.

[19] American Association of Physicists in Medicine. Acceptance Testing and Quality Assurance Procedures for Magnetic Resonance Imaging Facilities: Report of MR Subcommittee Task Group I. College Park, MD: AAPM; 2010. (AAPM Report No. 100).

[20] Philips Healthcare. Achieva 3.0T Operator Manual. Best, The Netherlands.Philips Medical Systems; 2014.

[21] Brazil. National Health Surveillance Agency (Anvisa). 2021. Instrução Normativa - IN nº 97, de 27 de maio de 2021[Normative Instruction No. 97, of May 27, 2021]. Diário Oficial da União, May 31, 2021. Accessed August 18, 2026. https://www.gov.br/anvisa

Safety and Conformity Evaluation of Medical Beds in the Post-Market Context

1st Miguel R. Nunes
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0009-1757-5568

2nd Diego A. O. Rosa
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0004-2422-8453

3rd Edison Silva Filho
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
009-0002-5397-0881X

4th Cristina Q. A. Fais
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0007-3632-7223

5th Gilberto S. Zunta
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0009-0002-6006-2142

6th Henrique T. Moriya
Biomedical Engineering Lab. (LEB)
University of São Paulo (USP)
São Paulo, Brazil
0000-0002-1595-1779

***Abstract*—Medical beds (MBs) are essential devices for patient care and must comply with regulatory requirements to ensure safety and performance throughout their life cycle. This study evaluated the conformity of MBs submitted to a public procurement process conducted by the São Paulo State Department of Health under Reference Term 330/2025. The evaluation was performed at the Normative Testing Division of the Biomedical Engineering Laboratory of the University of São Paulo and included tests based on ABNT NBR IEC 60601-2-52: 2013 and ABNT NBR IEC 62353: 2019 standards. The evaluation covered documentation, labeling, electrical safety, mechanical construction, mobility, patient entrapment protection, stability, and functional performance requirements. MBs from different manufacturers were initially received for evaluation. One manufacturer was disqualified during the inspection phase due to noncompliance with mattress specifications, and the remaining products proceeded to testing. The results demonstrated that none of the evaluated beds fully complied with the applicable requirements. Nonconformities were identified in many tests. The findings revealed discrepancies between expected regulatory compliance and the actual characteristics of products available on the market, indicating potential risks to patients and healthcare professionals. The study also highlights the importance of integrating safety tests, clinical engineering practices, and post-market surveillance to support continuous product improvement and patient safety.**

***Index Terms*—hospital bed, functionality, security, entrapment, adverse events**

I. Introduction

A medical bed (MB) is a medical device intended for patient rest or sleep, designed to assist in the diagnosis, monitoring, and treatment of diseases, or to compensate for injuries and disabilities [1]. To fulfill these clinical functions, MBs must incorporate a reinforced metal frame and mechanisms that allow the position of the bed to be adjusted, arfticulated side rails as a means of protection against falls and features to facilitate the entry and exit of the patient from the bed. These adjustments may include head, legs and foot elevation, as well as height range or inclination, facilitating both therapy and patient rest and the activities of healthcare professionals.

According to the Brazilian Health Regulatory Agency (ANVISA), MBs are commonly classified as low-risk medical equipment [2]. They are widely used in healthcare facilities such as hospitals, clinics, nursing homes, or in homes when patient requires continuous care or has reduced mobility.

Due to their relevance to life care and support, all health marketed in Brazil must comply with the Brazilian Health Regulatory System (SNVS) by being registered by ANVISA [3]. In addition, electrically operated medical beds and be certified under the Brazilian Conformity Assessment System (SBAC), ensuring that these devices meet both regulatory and technical standards for safety and essential performance.

Additionally, the establishment of the Brazilian Sentinel Hospital Network by ANVISA and the mandatory implementation of Patient Safety Centers (NSP) in healthcare institutions, enabled the systematic monitoring of adverse events occurring during the post-market scenario. These events are reported and monitored by the SNVS through the Surveillance Notification System's module (NOTIVISA). Notifications can be made via the web platform by health professionals and citizens [4].

An analysis of incident reports available in the ANVISA public dashboard revealed 28 incident reports involving hospital beds between 2007 and 2026. Among these notifications, 24 lacked a specified root cause, three were directly associated with patient falls, and one was officially classified as a "drowning" event [5], a critical clinical occurrence that warrants further technical examination of this study.

To ensure product quality and promote good safety practices in hospital environments, the São Paulo State Department of Health (SES-SP) established, through Reference Term (RT) 330/2025, prepared by the General Administrative Coordination (CGA), a public procurement process for the acquisition of MBs. The procurement was conducted through an

electronic auction, adopting evaluation criteria based on the lowest price combined with cost-benefit considerations and conformity evaluation. Compliance was demonstrated through the execution of tests performed according to methodologies aligned with the requirements of the ABNT NBR IEC 60601-2-52: 2013 - Particular requirements for the basic safety and essential performance of medical beds and ABNT NBR IEC 62353: 2019 - Recurrent test and test after repair of medical electrical equipment [1], [6], [7].

In cooperation with SES-SP, this study evaluated the quality and conformity of the MBs defined under RT 330/2025. The evaluation included tests to verify electrical safety, electromechanical construction, essential performance, accessories, labeling, and user instruction manuals. By presenting these findings, this study aims to inform manufacturers about potential gaps between regulatory compliance and real-world clinical use while contributing to improving medical device quality, patient safety, and supporting the objectives of the National Patient Safety Program (PNSP) [4].

II. Methodology and materials

According to the RT, sixteen manufacturers of MBs with weighing scales and twenty manufacturers of MBs without weighing scales competed in the bidding process. The evaluation followed a sequential line: one unit with a weighing scale and one without were tested concurrently. If both MBs fulfilled all requirements, they were accepted, the procurement item was awarded, and the evaluation of subsequent candidates in line was terminated. If the equipment under test failed any item, the testing process was immediately halted, the company was officially notified and the process moved to the next ranked manufacturer. Based on this pricing registry workflow, the laboratory received and evaluated a total of four MBs with weighing scales and four without weighing scales between November 2025 and April 2026. Although additional MBs remained on the waiting list, they were not subjected to experimental testing as the subsequent manufacturers did not meet the mandatory minimum price criteria. It is important to mention that manufacturer's representatives were present throughout the entire evaluation period, and no objections regarding the test procedures, execution, or results were reported.

The tested MBs had the following technical specifications: safe working load of 250 kg; internal bed dimensions of 900x205 mm (width x length); external bed dimensions of 1000x2150 mm; adjustable height without mattress from 420 to 750 mm; viscoelastic mattress with a thickness of 14 cm, steel frame with epoxy paint, articulated side rails with descent damping system, articulated bed in 4 sections (backrest, seat and leg rest with two sections), electric fowler-type movements (backrest, legs, Trendelenburg, reverse and elevation) activated by 4 motors, cardiopulmonary arrest device (electric and mechanical), remote control, minimum IPX4 protection rating against water ingress, antistatic double casters (with hubcap and 150 mm swivel), movement/ brake system activated by a single pedal, steering system or fifth wheel, automatic bivolt power supply with rechargeable battery option, wired remote control and with membrane keyboard control on the side rails for easy cleaning and disinfection (with the possibility of locking the rail control) and equipped with an aluminum or stainless steel IV pole.

To perform the experimental procedures, the infrastructure of the Normative Testing Division (DEN) of the Biomedical Engineering Laboratory (LEB) at the Escola Politécnica of the University of São Paulo (EPUSP) was utilized. The laboratory is equipped with metric scale (1000 mm, DEN ID: ESC-053), caliper (330-114, Mitutoyo, Japan), Weighing Indicator (I35, Precia Molen, France), stopwatch (Superact, DEN ID: CRO-149), balance (ID-U300/4, Filizola, Brazil), inclinator (CR20232, Digimees, Brazil), floor caliper (100460, Digimees, Brazil), electrical safety analyzer (LGI 3000, Entran, Brazil), and an environmental conditions meter (176H , Testo, Germany). All measurement instruments used complied with the specifications established by the ABNT NBR IEC 60601-2-52: 2013 and ABNT NBR IEC 62353: 2019 standards. Furthermore, the measurement instruments were duly calibrated and traceable to recognized metrological standards, ensuring the reliability and accuracy of the tests performed.

To meet the test requirement involving the application of dynamic forces, a mechanical cycling apparatus was developed using a pneumatic cylinder actuator (C95SDB250-XC6, SMC, Brazil) coupled to an air compressor (MSV5.2VL/130, Schulz, Brazil). The system was designed to simulate a patient sitting on the bed and performed 5,000 loading cycles. Before testing, the apparatus was validated using a load cell and weighing indicator to ensure accuracy and repeatability of the applied forces.

Additional materials, as specified by the ABNT NBR IEC 60601-2-52: 2013 standard, included: cone tool to simulate the patient's head, a cylinder tool to simulate the width of the patient's trunk and arm, two inclined planes at 6 degrees, one with vinyl flooring and the other with concrete, a load cushion, two obstacles to simulate a frontal collision with a door frame and another as a floor threshold. After being manufactured, the cone and cylinder tools were calibrated dimensions, mass and roughness. The two inclined planes were validated using an inclination gauge. The two obstacles were validated using a caliper.

A systematic inspection protocol was established for the MBs under test. Upon receipt, an initial checklist was utilized to register the equipment's status. This inspection verified full electromechanical functionality, the presence of standard accessories, and the completeness of the technical documentation (instruction manuals, service manuals, and risk management files). To minimize external variables, all testing was performed under stable and controlled climatic conditions.

MB functionality was evaluated by verifying whether the final test results met the compliance conditions described in the regulatory requirements.

A. Tests for MBs According to ABNT NBR IEC 60601-2-52: 2013 (Including Amendment 1) and ABNT NBR IEC 62353: 2019

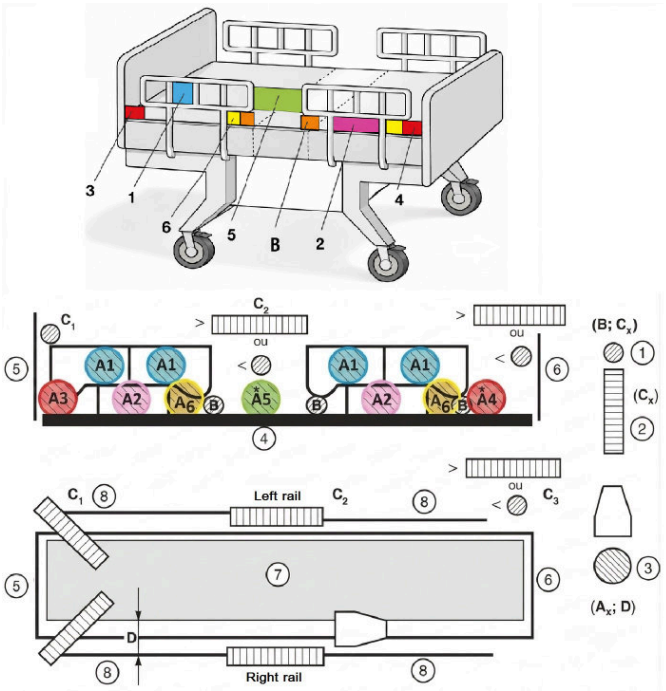
Table I describes the complete relationship between the specific normative clauses and the respective tools employed.

TABLE I
SUMMARY OF MBS TEST PROTOCOLS

Test	Normative Technical Specifications		
	Standard Clause	***Normative Requirement***	***Method***
T1	**201.7.2.2**	Manufacturer identification and labeling	Visual inspection
T2	**201.7.4.2**	Control devices	Functional verification
T3	**201.7.6.3**	Symbols for control and performance	Visual inspection
T4	**201.7.9.2**	Instructions for use	Document inspection
T5	**201.8.11.3.2**	Power cord: length $\geq$ 2.5 m and protection	Metric scale and Inspection
T6	**201.9.1.101**	Entrapment protection (rails and boards)	Cone and cylinder; Load cell (250 N)
T7	**201.9.2.2.5**	Continuous activation of controls	Functional verification
T8	**201.9.2.3.1**	Prevention of unintentional movement	Functional verification
T9	**201.9.101**	Protection against inadvertent patient falling	Side rail height protection Metric scale ($\geq$ 220 mm)
T10	**201.15.4.101**	Head/foot board removal ($<$ 15 s)	Chronometer
T11	**201.15.3.5**	Rough handling: 10 impacts at 0.4 m/s	Door frame and floor markings
T12	**201.9.4.2.3**	Instability from horizontal and vertical forces (1100 N)	Calibrated masses and balance
T13	**201.9.4.2.2**	Instability excluding transport (2200 N)	Calibrated masses and balance
T14	**201.9.4.3.1**	Transport stability on 6° inclined plane	Inclined plane, Inclinator balance (250 kg load)
T15	**201.9.4.2.4.3**	Caster resistance: 10 threshold collisions	Threshold, Caliper balance (250 kg load)
T16	**201.9.8.3.3**	Dynamic loading: 1350 N for 5000 cycles	Mechanical cycling apparatus
T17	**201.9.8.3.2**	Static forces due loading from persons 5000 N for 1 h	Floor caliper balance (250 kg load)
T18	**201.11.8**	Power Supply Interruption	Chronometer
T19	**5.3.2**	Grounding resistance measurement	Electrical Safety Analyzer (ESA)
T20	**5.3.4**	Leakage current limits verification via direct methods	Electrical Safety Analyzer (ESA)

Tests T1 through T4 focus primarily on labeling, control identification, and user instruction compliance. Mechanical integrity, patient entrapment mitigation, and structural stability under extreme vertical, horizontal, and transport loads were addressed from tests T5 to T17, using specialized apparatus such as load cells, calibrated masses, metric scales and standard threshold obstacles. Finally, power supply interruptions (Test T18), grounding resistance measurement (Test T19), and leakage current (test T20) were evaluated to ensure electrical safety.

To evaluate the risk of patient entrapment in non-moving parts (Test T6), the dimensions and critical boundary openings of the beds were checked using the cone and cylinder recommended by the ABNT NBR IEC 60601-2-52: 2013 standard. As illustrated in Figure 1, these evaluations focused on specific areas, such as the spaces within the side rails, between the rails and the head/footboards, and the clearances under the rail profiles.



- Ax represents the different A1, A2, A3, A4, A5 and A6.
1 Area of tool representing neck diameter (60 mm).
2 Area of tool representing chest breadth (318 mm).
3 Area of tool representing head breadth (120 mm).
4 Mattress Support Platform.
5 Head Board.
6 Foot Board.
7 Mattress.
8 Side rail.
* Only applies when the area C above is $<$ 60 mm.

Fig. 1. Patient entrapment zone for an MB with a segmented or split side rail, adapted from ABNT NBR IEC 60601-2-52: 2013 and Guidance for Industry and FDA Staff. [1] [8].

To evaluate the leakage current (Test T20), an Electrical Safety Analyzer was utilized to perform measurements through both the alternative and direct methods. As illustrated in Figure 2, both methodologies rely on a standardized measuring device (MD) prescribed by the ABNT NBR IEC 62353: 2019.

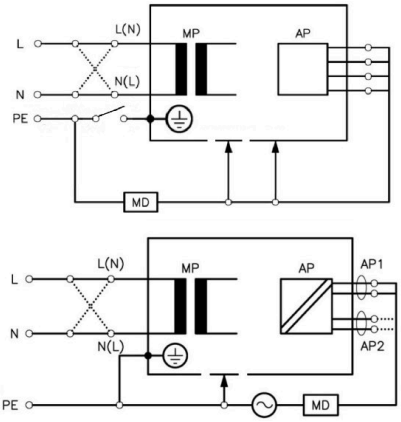


Fig. 2. Electrical diagram of the circuit for measuring leakage current to ground and from the patient, adapted from ABNT NBR IEC 62353: 2019 [8].

III. Results

The inspection and functional verification performed upon delivery of the MBs, as documented in the checklist forms, revealed the following findings:

- All beds demonstrated apparent structural integrity of the mattress support platform, articulated movement sections, side rails, headboard and footboard assemblies, castors, brake and locking systems, with no visible signs of damage or defects.
- All beds demonstrated proper operation under both mains and battery power. Functional verification confirmed correct performance of the height adjust-ment, backrest, leg rest, Trendelenburg, reverse Trendelenburg, and CPR functions using both the remote control and patient controls.
- The other parts of the beds, such as the power supply cord or electrical cables, connectors, and IV pole, were in good condition to use.
- A few MBs were delivered with instructions for use and risk management file. In those cases, they were requested to be sent by e-mail to realize the test T4 - 201.7.9.2.

Of the four manufacturers initially selected in the bidding process, one was disqualified during the initial inspection phase because the delivered mattress did not meet the technical specifications defined in the RT, namely the required thickness of 140 mm. Considering this fact, only three manufacturers proceeded to the evaluation tests.

During the initial evaluation, finger trapping zones were identified in two MBs in the area beneath the mattress support platform, associated with the side-rail articulation mechanism (Figure 3). Although this condition is addressed in item 201.9.2.2 of ABNT NBR IEC 60601-2-52: 2013, the requirement was not included in the RT. Therefore, the finding was communicated to the manufacturers as a safety-related observation and was not considered a criterion for disqualification.

According to the RT, it consisted of a total of 20 sets test for each MB, as described in Table I. Table II shows a summary of the conformities and non-conformities regarding the results obtained in the tests of the MBs that performed. However, only 15 tests were performed on one of the MB equipped with a weighing scale (Second-place finishers) failed to meet the requirements of clause 201.9.4.2.4.3 due to a structural failure of its castors while attempting to overcome the threshold and this rendered it impossible to carry out the remaining tests, as required by the bidding process.

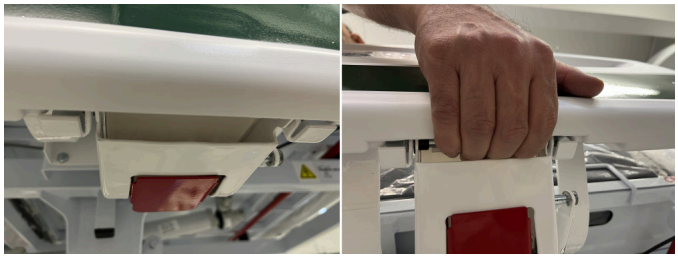


Fig. 3. Finger trapping zones

TABLE II
Results Medical Beds.

Test	MBs without weight scale				MBs with weight scale			
	Compliant			Non Compliant			Compliant			Non Compliant		
201.7.2.2	□	□	■	■	■	□	■	□	■	□	■	□
201.7.4.2	■	■	■	□	□	□	■	■	■	□	□	□
201.7.6.3	■	■	■	□	□	□	■	■	■	□	□	□
201.7.9.2	□	□	□	■	■	■	□	□	□	■	■	■
201.8.11.3.2	□	■	■	■	□	□	□	□	■	■	■	□
201.9.1.101	□	□	■	■	■	□	□	□	□	■	■	■
201.9.2.2.5	■	■	■	□	□	□	■	■	■	□	□	□
201.9.2.3.1	■	■	■	□	□	□	□	□	■	■	■	□
201.9.101	■	■	□	□	□	■	■	■	□	□	□	■
201.15.4.101	■	■	■	□	□	□	■	■	■	□	□	□
201.15.3.5	■	■	■	□	□	□	■	■	■	□	□	□
201.9.4.2.3	■	■	■	□	□	□	■	■	■	□	□	□
201.9.4.2.2	□	■	■	■	□	□	■	■	■	□	□	□
201.9.4.3.1	■	■	■	□	□	□	■	■	■	□	□	□
201.9.4.2.4.3	■	■	■	□	□	□	■	□	■	□	■	□
201.9.8.3.3	■	■	■	□	□	□	■	—	■	□	—	□
201.9.8.3.2	■	■	■	□	□	□	■	—	■	□	—	□
201.11.8	■	■	□	□	□	■	■	—	□	□	—	■
5.3.2	■	■	□	□	□	■	■	—	■	□	—	□
5.3.4	■	■	■	□	□	□	■	—	■	□	—	□

(−) Test not performed due to technical failure.

■ Stage 1: First-place finishers.

■ Stage 2: Second-place finishers.

■ Stage3: Third-place finishers.

IV. Discussion

Patient safety is a globally recognized healthcare priority and a fundamental component of quality of care. Adverse events associated with healthcare delivery remain a major challenge worldwide and have received considerable attention from organizations such as the World Health Organization (WHO).

The tests conducted in this study revealed that none of the evaluated MBs fully complied with the RT´s 20 sets test and applicable regulatory requirements, including documentation, mechanical safety, and electrical safety criteria. These findings demonstrate a gap between expected regulatory compliance and the actual characteristics of products specifically tested in this study, despite their disposable on the market, potentially affecting their safe use in healthcare environments. Given that all measuring instruments were calibrated, ensuring metrological reliability and that the manufacturers observed the tests without raising objections, it was made not to report the values of the results obtained, due to the large volume. Instead, the following comments were made.

Regarding documentation (T4), deficiencies were identified in the instructions for use that could compromise safe operation. Missing information included guidance on bed–mattress compatibility, inspection of critical components, and warnings related to mechanical hazards. In addition, several MBs lacked adequate identification markings (T1) on major components and accessories, omitting essential information such as manufacturer identification, model designation, and part numbers. Such deficiencies hinder traceability, maintenance activities, replacement part procurement, and post-market surveillance.

Some mechanical non-conformities were also observed. Openings within the side-rail structures and gaps between the side rails and the headboard or footboard presented potential entrapment hazards (T6), because they failed to withstand the forces exerted by the conical (250 N) and cylindrical (weight force) test tools when these were inserted into or applied against the relevant areas. Furthermore, the height between the mattress surface and the upper edge of the side rails (T9) was not greater than minimum limit of 220 mm established by the standard. Two of the tested MBs, from different manufacturers, exhibited stability deficiencies by failing to withstand a lateral force of 2200 N in the instability excluding transport test (T13), as well as mobility issues resulting from caster systems unable to traverse floor irregularities (T15). These issues could compromise bed stability and increase the risk of patient falls. Another concern regarding the two tested MBs involved finger-crushing risks associated with the side rail's articulation mechanism.

Electrical safety concerns were also identified. First of all, it is worth noting that the power cords for two of the MBs did not meet the minimum length of 2.5 meters established by the standard (T5). In some MBs, the impedance of the protective earth conductor exceeded the limits of 300 miliohm established by the standard. Additionally, certain models were classified as Class II equipment and employed detachable power supply cords. Although Class II protection is based on double insulation, its application to electrically operated hospital beds deserves careful consideration due to the presence of extensive accessible metallic structures and electromechanical systems. In the event of insulation failure, accessible conductive parts could become energized, potentially exposing patients and healthcare professionals to electric shock hazards. In this context, Class I protection systems may provide an additional safety layer through the use of a protective earth conductor capable of safely conducting fault currents.

The relevance of these findings is reinforced by reports from international surveillance systems. Functional failures, patient entrapment, falls, and operational limitations associated with hospital beds have been recognized worldwide as important patient safety concerns. According to the Food and Drug Administration (FDA), between 1985 and 2009, 803 patient entrapment incidents involving hospital bed rails were reported in the United States, resulting in 480 deaths and 138 nonfatal injuries [9]. Similarly, the Medicines and Healthcare products Regulatory Agency (MHRA) reported 18 deaths and 54 serious injuries associated with hospital beds and bed rails in the United Kingdom between 2018 and 2022 [10]. In Canada, surveillance data identified 125 patient entrapment incidents between 1980 and 2017, including 47 deaths [11]. These reports demonstrate that the risks identified in the present study are not merely theoretical but are associated with documented adverse events in clinical practice.

Despite the relevance of these risks, Brazil still faces challenges in characterizing and analyzing medical device-related adverse events. Although the NOTIVISA system provides a structured framework for adverse event reporting and supports preventive actions at the local and national levels, important limitations remain, including underreporting. Moreover, the system does not provide specific categorization for incidents involving patient falls or entrapment associated with MBs, limiting the ability to identify trends and establish targeted risk-reduction strategies.

A notable example is the single notification classified as a "drowning" event associated with MBs. While this report suggests the potential involvement of an MB in a severe adverse event, the lack of clinical and technical details prevents a proper understanding of the circumstances and identification of root causes. Consequently, opportunities to learn and implement corrective or preventive measures remain limited. Such information gaps may contribute to the under-recognition of risks associated with MBs in the Brazilian post-market surveillance system.

Finally, it should be noted that the present study evaluated only twenty sets of test requirements established in the RT, however the ABNT NBR IEC 60601-2-52: 2013 includes additional tests that were outside the scope of this investigation but remain highly relevant to patient safety and product performance, including evaluation of audible acoustic energy, side-rail locking strength and reliability, protection against excessive temperatures and other hazards, resistance to water ingress, and requirements applicable to machine-

washable medical beds.

V. Conclusions

The tests conducted in this study revealed that none of the particularly evaluated MBs fully met the requirements of the twenty sets test established by the RT in accordance with the ABNT NBR IEC 60601-2-52: 2013 standard, as well as the ABNT NBR IEC 62353: 2019 standard. These findings demonstrate that regulatory certification alone may not be sufficient to ensure that products available on the market consistently meet the expected standards of safety and performance, potentially indicating noncompliance with Good Manufacturing Practices (GMP) despite pre-market registration and certification.

The identified nonconformities indicate the existence of risks that may affect both patients and healthcare professionals during routine clinical use on MBs evaluated.

This study also demonstrates the need for a more integrated approach involving manufacturers, healthcare institutions, clinical engineering services, and regulatory authorities throughout the entire product lifecycle. Effective interaction among these stakeholders is essential to ensure that information obtained during real-world use contributes to continuous product improvement and risk reduction.

Finally, the results reinforce the importance of strengthening the connection between conformity evaluation, post-market surveillance, and clinical practice. Such integration is fundamental for improving the safety, performance, and reliability of medical devices.

Acknowledgment

This study was carried out within the framework of a partnership between DEN-LEB-EPUSP and SES-SP. This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (309464/2022-6 to HTM).

References

[1] Associação Brasileira de Normas Técnicas (ABNT), ABNT NBR IEC 60601-2-52: Equipamento eletromédico – Parte 2-52: Requisitos particulares para a segurança básica e o desempenho essencial de camas médicas. Rio de Janeiro, Brazil: ABNT, 2013.

[2] Agência Brasileira de Desenvolvimento Industrial (ABDI), Manual de Registro e Cadastramento de Materiais de Uso em Saúde. Brasília, DF, Brazil: ABDI, 2022.

[3] Agência Nacional de Vigilância Sanitária (ANVISA), Resolução RDC nº 751, de 15 de setembro de 2022. Brasília, DF, Brazil, 2022.

[4] Agência Nacional de Vigilância Sanitária (ANVISA), Documento de Referência para o Programa Nacional de Segurança do Paciente. Brasília, DF, Brazil: Ministério da Saúde, 2014.

[5] Agência Nacional de Vigilância Sanitária, "Painel Power BI – Notificações em Tecnovigilância", Brasília, Brazil. [Online]. Available: https://www.gov.br/anvisa/pt-br/acessoainformacao/dadosabertos/informacoes-analiticas/tecnovigilancia/notificacoes-tecnovigilancia, [Accessed: Jun. 22, 2026].

[6] Coordenadoria Geral Administrativa (CGA), Termo de Referência nº 330/2025. São Paulo, SP, Brazil, 2025.

[7] Associação Brasileira de Normas Técnicas (ABNT), ABNT NBR IEC 62353: Equipamento Eletromédico – Ensaio Recorrente e Ensaio Após Reparo de Equipamento Eletromédico. Rio de Janeiro, Brazil: ABNT, 2019.

[8] Food and Drug Administration (FDA), Guidance for Industry and FDA Staff: Hospital Bed System Dimensional and Assessment Guidance to Reduce Entrapment, Silver Spring, MD, USA, Mar. 2006.

[9] Food and Drug Administration (FDA), A Guide to Bed Safety Bed Rails in Hospitals, Nursing Homes and Home Health Care: The Facts, USA govermment - revsised April 2010.

[10] Medicines and Healthcare products Regulatory Agency (MHRA), Guidance Bed rails: management and safe use Guidance on managing and using bed rails safely, London, U.K., published August 2023.

[11] Winnipeg Regional Health Authority (WRHA), Bed System Safety: Long-Term Care Resource Guide. Winnipeg, MB, Canada, revised October 2025.

Multimodal Characterization of a Single Magnetic Nanoparticle Formulation

1° Maria Eduarda Domingues Alves
Department of Physics
University of São Paulo
Ribeirão Preto, Brazil
maria.e.domingues@usp.br

2° Thiago Tibúrcio Vicente
Department of Physics
University of São Paulo
Ribeirão Preto, Brazil
thiagotiburcio37@usp.br

3° Raffael Willmann
Institute of Biosciences
São Paulo State University
Botucatu, Brazil
r.willmann@unesp.br

4° Guilherme Piloto Fernandes
Department of Physics
University of São Paulo
Ribeirão Preto, Brazil
guilherme.santos.fernandes@usp.br

5° Ricardo Ricci Bordonal
Department of Physics
University of São Paulo
Ribeirão Preto, Brazil
ricardo.bordonal@usp.br

6° Theo Zeferino Pavan
Department of Physics
University of São Paulo
Ribeirão Preto, Brazil
theozp@usp.br

7° José Ricardo Miranda
Institute of Biosciences
São Paulo State University
Botucatu, Brazil
jmiranda@ibb.unesp.br

8° Antonio Adilton Oliveira Carneiro
Department of Physics
University of São Paulo
Ribeirão Preto, Brazil
adilton@usp.br

I. Abstract

Magnetic nanoparticles (MNPs) were evaluated through complementary magnetic, mechanical, optical, and thermal techniques, including Alternating Current Biosusceptometry (ACB), Magnetomotive Ultrasound (MMUS), Photoacoustic Imaging (PAI), and Magnetic Hyperthermia (MH). The nanoparticles were synthesized by chemical coprecipitation and characterized in terms of their magnetic and morphological properties. A series of nanoparticle concentrations was investigated using tissue-mimicking gelatin/agar phantoms for ACB with two-dimensional spatial scanning, MMUS and PAI and aqueous suspensions for ACB and MH. All investigated modalities detected the MNPs, with the measured responses varying according to nanoparticle concentration and the physical mechanism probed by each technique. MMUS displacement and PAI contrast increased with increasing concentration, demonstrating their ability to provide mechanical and photoacoustic contrast, respectively. In MH, higher nanoparticle concentrations resulted in greater temperature increases, while the specific loss power (SLP) decreased with increasing concentration. ACB measurements enabled the detection and concentration-dependent characterization of the magnetic response before and after the investigated hyperthermia cycles. Overall, the results demonstrate the feasibility of characterizing the same MNP formulation using complementary experimental modalities and highlight its potential for multimodal phantom characterization and magnetic hyperthermia studies.

Keywords: Magnetic Nanoparticles, Alternating Current Biosusceptometry (ACB), Magnetomotive Ultrasound (MMUS), Photoacoustic Imaging (PAI), Magnetic Hyperthermia (MH).

II. Introduction

Magnetic nanoparticles (MNPs) have been widely investigated due to their ability to respond to external magnetic fields and to exhibit properties that can be explored through different physical mechanisms. In particular, magnetite (Fe_3O_4) nanoparticles have attracted attention because their magnetic behavior can be investigated using sensing and imaging techniques, while their interaction with alternating magnetic fields can also result in heat generation. These characteristics make MNPs suitable for comparative studies in which different measurement modalities are used to probe distinct aspects of the same material.

The response of MNPs to external stimuli depends on factors such as composition, particle size, morphology, magnetic properties, and concentration. Consequently, measurements based on magnetic, mechanical, optical, and thermal effects may provide complementary information about the same nanoparticle formulation [2], [5]. Evaluating these responses under controlled experimental conditions is therefore relevant for establishing how nanoparticle concentration influences the measured signals and for identifying the potential and limitations of each characterization approach.

In ACB, MNPs are detected through changes in the magnetic flux produced by their response to an alternating magnetic field. The sensing system employs a gradiometric coil configuration, in which magnetic contributions common to the detection regions tend to be compensated. In the absence of

magnetic material near the sensitive region, this arrangement keeps the system approximately balanced. However, when MNPs are exposed to the excitation field, their magnetization introduces an additional magnetic contribution, disturbing the balance between the coils and resulting in an electrically induced detectable signal. The amplitude and phase of this response are related to the magnetic behavior of the nanoparticles and may vary with their concentration and position relative to the sensing system [1]. In addition, when performed over the phantom, displacement of the sensor enabled two-dimensional spatial scanning, allowing the distribution of MNP-containing inclusions to be mapped and the signal variation across the investigated region to be evaluated.

In addition, Magnetomotive Ultrasound (MMUS) combines magnetic excitation with ultrasound detection to characterize magnetic materials through mechanically induced motion. When MNPs are exposed to an external magnetic field, magnetic forces acting on the particles can produce mechanical displacements in the surrounding medium. These displacements can be detected by ultrasound and used to generate mechanical contrast between regions containing and not containing MNPs. The magnitude of the displacement depends on factors such as the amount of magnetic material and the mechanical properties of the surrounding medium [2].

Photoacoustic Imaging (PAI) provides another complementary contrast mechanism by combining optical excitation with ultrasonic detection. Following the absorption of a short laser pulse, the absorbed optical energy is converted into localized thermal expansion, generating acoustic waves that can be detected by an ultrasound transducer. The resulting photoacoustic signal depends on the optical absorption properties of the investigated material and can therefore provide contrast between regions with different compositions [3]. For MNP-containing samples, PAI can be used to detect the inclusions and evaluate changes in photoacoustic contrast as a function of nanoparticle concentration.

Beyond imaging applications, Magnetic Hyperthermia (MH) explores the thermal response of MNPs under an alternating magnetic field. Energy dissipation by the nanoparticles can result in heat generation through magnetic relaxation mechanisms, including Néel and Brownian relaxation, whose relative contributions depend on nanoparticle and medium properties [4]. The heating efficiency can be quantified by the Specific Loss Power (SLP), which represents the power dissipated per unit mass of magnetic material. In this study, MH measurements were performed using MNP suspensions at different concentrations, allowing the temperature increase and SLP to be evaluated under controlled experimental conditions.

Tissue-mimicking phantoms are experimental models designed to reproduce selected physical properties of biological tissues under controlled laboratory conditions. They provide a standardized environment for evaluating imaging responses while controlling parameters such as geometry, concentration, and spatial distribution of the investigated material. In this study, gelatin/agar phantoms containing MNP inclusions at different concentrations were used for MMUS and PAI measurements, as well as for two-dimensional spatial scanning with ACB, while MNP suspensions in Eppendorf tubes provided a controlled configuration for complementary measurements, particularly magnetic and thermal characterization.

Taken together, ACB, MMUS, PAI, and MH enable the investigation of complementary magnetic, mechanical, photoacoustic, and thermal responses of the same MNP formulation. Therefore, this study aimed to evaluate a magnetite-based MNP formulation at different concentrations using these four techniques, investigating how nanoparticle concentration affects the measured responses and comparing their performance under the respective experimental conditions.

III. Results and Discussion

The synthesized MNPs exhibited well-defined nanoscale structural and magnetic characteristics. In Figure 1(a) the TEM analysis revealed approximately spherical MNPs with a relatively homogeneous size distribution. The particle-size analysis yielded a mean diameter of 6.8 ± 0.25 nm, with most particles distributed approximately between 5 and 8 nm, as shown in the corresponding size-distribution histogram in Figure 1(d). In Figure 1(b) the room-temperature magnetization curve obtained by vibrating sample magnetometry (VSM) exhibited low coercivity and reached magnetization values of approximately 60–65 emu/g. These results are consistent with the magnetic response expected for nanoscale magnetite particles. The MNPs were stabilized using sodium citrate during synthesis, and all concentrations reported in this study correspond to the total mass of MNPs per unit volume of suspension or phantom inclusion.

Figure 1 presents the characterization of the nanoparticles together with the gelatin/agar phantom developed for the experiments, containing $200\,\mu$L inclusions with MNP concentrations of 10, 20, 30, 40, and 50 mg/mL. The use of the same phantom for MMUS and PAI enabled the response of the two imaging modalities to be evaluated using the same concentration series and phantom geometry as shown in Figure 1(c).

For the ACB measurements, the excitation frequency was set to 10 kHz, with a lock-in amplifier gain of 0 dB and a sensitivity of 10 mV. The excitation coil was driven with 13.2 V RMS (approximately 37.6 Vpp) and a current of 56.4 mA. Measurements were performed for MNP concentrations of 10, 20, 30, 40, and 50 mg/mL. For the suspension measurements, the response was corrected by subtracting the signal obtained from a reference sample containing 200 μL of Milli-Q water. The ACB response was characterized in terms of the magnitude (R) and phase (θ) of the detected signal.

Figure 2(a) presents the magnitude (R) and phase (θ) obtained by ACB for MNP suspensions at different concentrations. A progressive decrease in the values of R and θ was observed with increasing MNP concentration. However, this variation represents the displacement of the measured signal relative to the reference condition containing only Milli-Q water. Considering the difference between the reference signal and the signal obtained for each concentration, a progressive

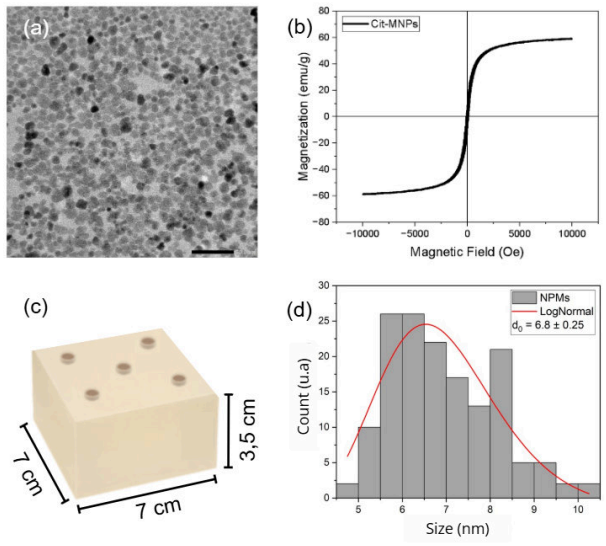


Fig. 1. (a) transmission electron microscopy (TEM) image; (b) room-temperature magnetization curve (emu/g) as a function of the applied magnetic field (Oe); (c) illustration of the gelatin/agar phantom containing inclusions with different MNP concentrations ; (d) histogram of the MNP size distribution.

increase in the response associated with the MNPs was observed. The mean magnitude variation (ΔR) increased from 1.19 mV at 10 mg/mL to 6.23 mV at 50 mg/mL, while the mean phase variation ($\Delta\theta$) increased from 2.38° to 18.77° over the same concentration range.

To quantitatively assess the dependence of the ACB response on MNP concentration, linear regressions were performed between the concentration (C) and both ΔR and $\Delta\theta$. The magnitude variation was described by $\Delta R = (0.1252 \pm 0.0403)C + 0.286$, with $R^2 = 0.9702$, where the uncertainty corresponds to the 95% confidence interval of the slope. Similarly, the phase variation was described by $\Delta\theta = (0.4108 \pm 0.0512)C - 1.948$, with $R^2 = 0.9954$. The high coefficients of determination indicate a strong positive linear relationship between the ACB response and MNP concentration, demonstrating the sensitivity of the technique to the amount of magnetic material present in the samples.

The detection capability of the ACB system was further evaluated from the calibration curve. The limit of detection (LOD) was estimated according to

$$\mathrm{LOD} = \frac{3.3\sigma_{\mathrm{blank}}}{S}, \tag{1}$$

Where σ_{blank} is the standard deviation of the blank response and S is the slope of the calibration curve. The blank response was obtained from $n = 3$ repeated measurements of an Eppendorf tube containing 200 μL of Milli-Q water, measured under the same experimental conditions as the MNP suspensions. For the magnitude response, the standard deviation obtained from the repeated blank measurements was approximately 0.021 mV, and the slope of the ΔR versus concentration calibration curve was 0.1252 mV/(mg/mL). Thus, an estimated LOD of approximately 0.55 mg/mL was obtained.

Figure 2(b) presents the two-dimensional map obtained from the spatial scanning of the biosusceptometer over the phantom, followed by interpolation of the acquired data. The map enabled the spatial visualization of the regions corresponding to the inclusions containing MNPs, with different response levels according to the concentration of magnetic material, thereby facilitating the identification and localization of the inclusions. In the representation shown, the magnitude scale was inverted during data processing, such that lower displayed values correspond to regions with higher experimentally measured magnetic responses. Thus, the dark-blue region, associated with the lowest value on the displayed scale, corresponds to the inclusion with the highest MNP concentration, whereas the other regions exhibit progressively lower responses according to the MNP concentration. This inversion was applied exclusively for graphical representation and does not alter the experimental values obtained during the spatial scan.

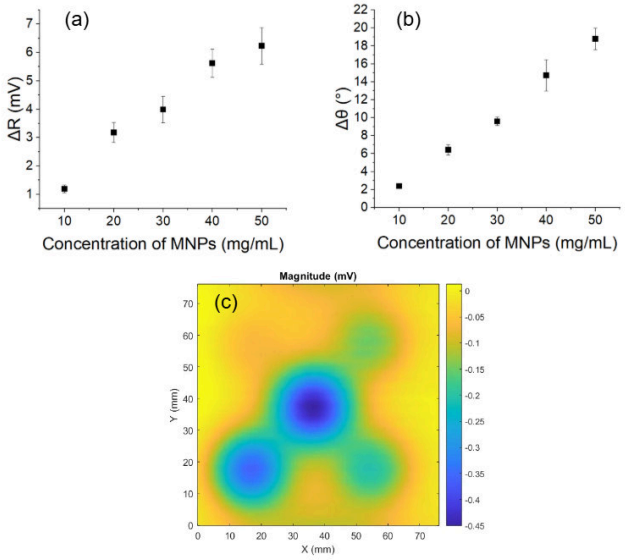


Fig. 2. ACB results: (a) magnitude (R) and (b) phase (θ) variations relative to the reference for different MNP concentrations; (c) interpolated two-dimensional map obtained from the spatial scan performed over the phantom.

In the MMUS experiments, the mechanical displacement of the surrounding medium induced by the magnetic excitation was detected in the regions containing MNPs. Each inclusion was subjected to magnetic field pulses of approximately 0.1 T with a duration of 6 ms, generated by a 130-turn coil with a ferromagnetic core. Ultrasound signal acquisition was performed using a programmable Verasonics ultrasound platform coupled to a linear L14-5/38 transducer, with a center frequency of 9.5 MHz and an acquisition rate of 4 kHz. The transducer was positioned to monitor the displacements produced in the MNP-containing regions during the application of the magnetic pulses. The MMUS response was quantified as the magnetically induced displacement, expressed in μm.

Larger displacements were observed at higher nanoparticle concentrations, consistent with the greater amount of magnetic material within the inclusions. Despite the inherent variability of the phantom, the results demonstrated the ability of MMUS to generate mechanical contrast associated with the presence and concentration of MNPs.

Three repeated acquisitions were performed for each MNP concentration. An example of the acquired B-mode image for the 30 mg/mL inclusion is shown in Figure 3(a), together with the corresponding MMUS displacement map in Figure 3(b).The mean displacement increased from 9.33 ± 1.15 μm at 10 mg/mL to 31.67 ± 3.51 μm at 50 mg/mL. The corresponding mean displacements for 20, 30, and 40 mg/mL were 12.00 ± 2.00, 20.00 ± 0.00, and 23.00 ± 2.65 μm, respectively, as shown in Figure 3(c). The zero standard deviation at 30 mg/mL reflects the displacement quantization of the estimation method rather than perfect repeatability. A linear regression between MNP concentration and mean displacement yielded $D = (0.57 \pm 0.06)C + 2.5$, with $R^2 = 0.969$, indicating a strong positive association between MNP concentration and the magnetically induced displacement under the investigated conditions.

For PAI, the inclusions containing MNPs exhibited enhanced photoacoustic contrast relative to the surrounding gelatin matrix. This behavior is attributed to the increased optical absorption promoted by the MNPs, which enhances the generation of photoacoustic waves detected by the ultrasound transducer [3].

An increase in the maximum temperature was observed with increasing MNP concentration, rising from 26.4 °C to 34.9 °C between the concentrations of 10 and 50 mg/mL. In contrast, the mean Specific Loss Power (SLP) decreased from 5.87 W/g to 4.56 W/g, indicating a reduction in the heating efficiency per unit mass of nanoparticles. This behavior is consistent with previous reports and may be attributed to enhanced dipolar interactions and nanoparticle aggregation in more concentrated suspensions, which restrict the magnetic relaxation mechanisms responsible for heat generation [7].

The same phantoms used in the MMUS experiments were employed for photoacoustic imaging (PAI). Optical excitation was performed using a Nd:YAG laser coupled to an OPO system (AcousticX), operating at 750 nm and 10 Hz. The photoacoustic signals were acquired using an L14-5/38 linear transducer with a sampling frequency of 40 MHz. The images were normalized by the energy of each laser pulse, and the photoacoustic response was quantified using the signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR).

For the PAI measurements, a single acquisition was performed for each MNP concentration; therefore, no repeated measurements were available to estimate experimental variability. The results are consequently presented as descriptive values and interpreted primarily as qualitative trends rather than as statistically validated differences between concentrations. For quantitative analysis, a square region of interest (ROI) of 5×5 mm was positioned within each MNP-containing inclusion, and an additional ROI of the same size

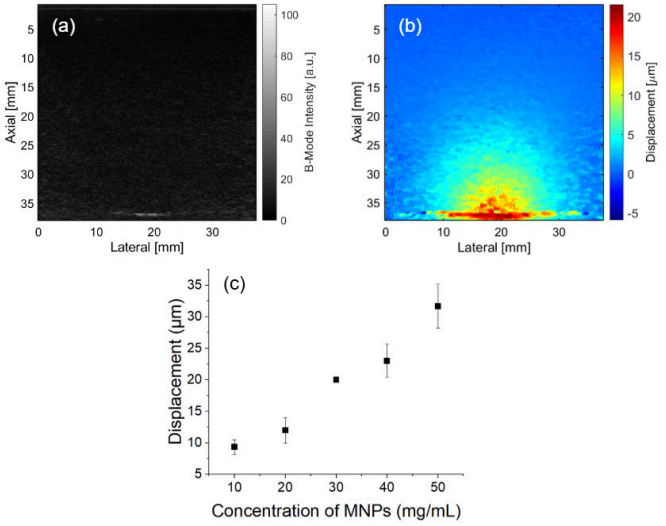


Fig. 3. MMUS results obtained from the phantom containing different MNP concentrations: (a) B-mode image of the 30 mg/mL inclusion; (b) MMUS displacement map for the 30 mg/mL inclusion; (c) displacement amplitude as a function of MNP concentration.

was selected in the surrounding background region. The B-mode image of the phantom is shown in Figure 4(a), while the corresponding photoacoustic image acquired at 750 nm is presented in Figure 4(b). The PAI response was quantified using the signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) calculated from the ROI. The measured SNR values were 22.10, 23.31, 24.80, 26.76, and 27.25 dB, while the corresponding CNR values were 11.20, 13.07, 15.90, 19.81, and 21.39 dB for MNP concentrations of 10, 20, 30, 40, and 50 mg/mL, respectively, as shown in Figure 4(c) and Figure 4(d). The increase in these parameters indicates enhanced photoacoustic signal and contrast with increasing MNP concentration, which can be associated with the greater optical absorption provided by the nanoparticles [3]. Although the results indicate a concentration-dependent trend, statistical analysis with a larger number of repetitions would be required to establish the significance and reproducibility of the observed differences.

Overall, the PAI results demonstrate a concentration-dependent increase in both SNR and CNR under the investigated experimental conditions. However, because the present analysis does not include statistical tests between concentrations, the observed differences should be interpreted as a trend rather than as statistically confirmed differences. Additional repeated measurements and statistical analysis would be necessary to assess the significance and reproducibility of the concentration dependence.

MH experiments were performed using 200 μL of MNP suspension exposed to an alternating magnetic field at 222 kHz and 300 A for 10 min using an EasyHeat system (Ambrell). Temperature was monitored using a fiber-optic temperature sensor positioned directly in the suspension, with an initial temperature of approximately 22.6 °C. The maximum

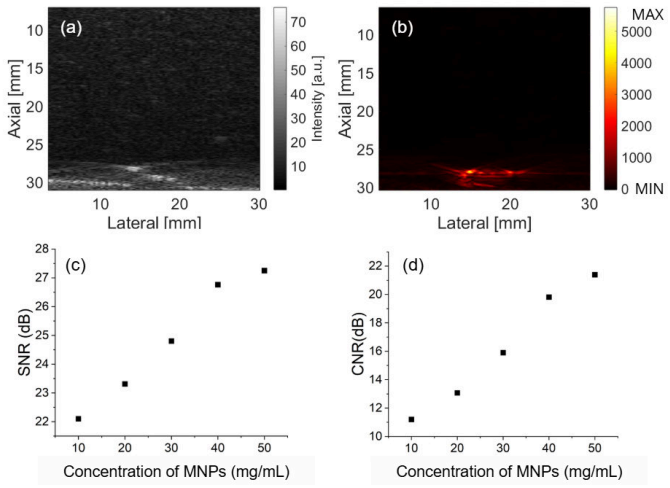


Fig. 4. PAI results: (a) B-mode image of a phantom containing inclusion with MNPs; (b) PA images, acquired at 750 nm; (c) SNR and (d) CNR as a function of MNPs concentration.

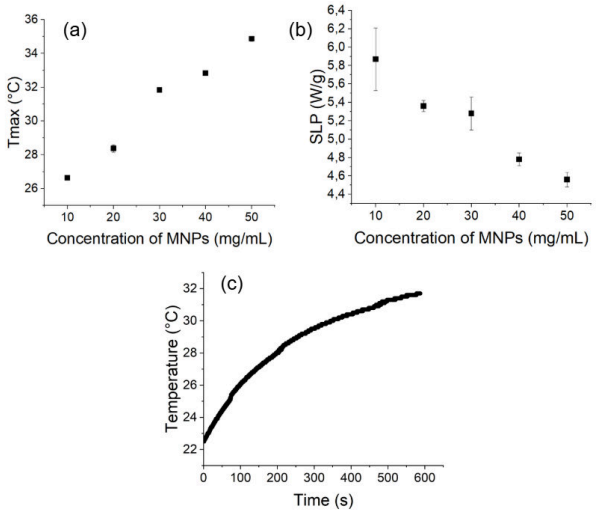


Fig. 5. MH results: (a) maximum temperature variation; (b) specific loss power (SLP); (c) temperature evolution.

temperature increased from 26.4 °C to 34.9 °C for MNP concentrations from 10 to 50 mg/mL, respectively, as shown in Figure 5(a), indicating greater heating with increasing nanoparticle concentration.

The specific loss power (SLP) was calculated according to

$$\mathrm{SLP} = \frac{C_{\mathrm{eff}}}{m_{\mathrm{MNP}}} \left(\frac{dT}{dt} \right) t \to 0, \tag{2}$$

where C_{eff} is the effective heat capacity of the sample (J/K), m_{MNP} is the total mass of Fe_3O_4 nanoparticles in the sample (g), and dT/dt is the initial heating rate (K/s), resulting in SLP values expressed in W/g. No correction for heat losses to the surroundings was applied. The obtained SLP values were 5.87 $\pm$ 0.34, 5.36 $\pm$ 0.06, 5.28 $\pm$ 0.18, 4.78 $\pm$ 0.07, and 4.56 $\pm$ 0.08 W/g for 10, 20, 30, 40, and 50 mg/mL, respectively, as shown in Figure 5(b). Despite the increase in the total amount of magnetic material, SLP decreased by approximately 22% over the investigated concentration range. This behavior may be related to concentration-dependent magnetic interactions and changes in the relaxation mechanisms; however, possible aggregation or structural changes after heating were not directly characterized in this study and therefore cannot be established as the cause.

The temperature profiles showed a progressive increase in the final temperature with MNP concentration, as illustrated in Figure 5(c). After MH, the ACB measurements exhibited changes in both amplitude (ΔR) and phase ($\Delta\theta$), with ΔR values of 0.028, -0.058, 0.154, 0.175, and 0.225 mV and $\Delta\theta$ values of 0.023, -0.367, 0.183, 0.213, and 0.267° for 10–50 mg/mL, respectively. The corresponding changes in ACB amplitude (ΔR) are shown in Figure 6(a), while the changes in phase ($\Delta\theta$) are shown in Figure 6(b). The increase in ΔR at higher concentrations suggests a concentration-dependent change in the magnetic response after heating. However, the non-monotonic behavior at lower concentrations and the absence of post-heating structural characterization prevent a definitive attribution of these changes to nanoparticle aggregation or dipolar interactions. Complementary characterization before and after MH would be required to investigate these effects.

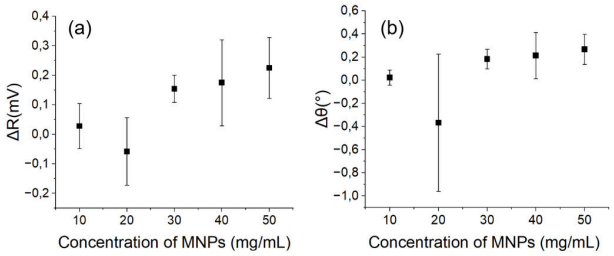


Fig. 6. ACB results after MH: (a) changes in ACB amplitude (ΔR) as a function of MNP concentration; (b) changes in ACB phase ($\Delta\theta$) as a function of MNP concentration.

Although ACB, MMUS, PAI, and MH rely on different physical mechanisms, all techniques exhibited consistent responses to the presence of the MNPs. These results demonstrate that a single nanoparticle formulation can provide complementary magnetic, mechanical, optical, and thermal information. Further studies involving biocompatibility assessment, stability in biological media, post-hyperthermia structural characterization, and integrated multimodal experiments are required before biomedical applications can be established.

IV. Conclusion

The present study investigated the performance of MNPs synthesized by the coprecipitation method in four distinct modalities: ACB, MMUS, PAI, and MH. The results demonstrated that all techniques were capable of detecting the presence of the MNPs, although with different sensitivities and contrast mechanisms.

In ACB, ΔR and $\Delta\theta$ varied linearly with MNP concentration ($R^2 = 0.9702$ and 0.9954; estimated LOD ≈ 0.55 mg/mL), and two-dimensional scanning enabled the localization of the inclusions. Concentration-dependent changes in amplitude and phase were observed after the hyperthermia cycles.

In MMUS, an increasing trend in the induced mechanical displacement was observed with increasing nanoparticle concentration, rising from approximately $9.33\,\mu$m at the lowest concentration to nearly $31.67\,\mu$m at 50 mg/mL. A linear relationship was observed between MNP concentration and mean displacement over the investigated concentration range. In PAI, photoacoustic contrast increased with nanoparticle concentration, highlighting the potential of the MNPs as optical contrast agents in tissue-mimicking environments. In MH, the maximum temperature increased with MNP concentration, exhibiting an approximately 31% increase between 10 and 50 mg/mL. Conversely, SLP decreased by approximately 22%, a behavior attributed to enhanced interparticle interactions and the consequent reduction in magnetic relaxation efficiency at higher concentrations.

Overall, the results demonstrate the feasibility of characterizing a single magnetite-based MNP formulation through complementary magnetic, mechanical, optical, and thermal modalities. Each modality probes distinct physical properties of the same nanoparticle formulation, highlighting their complementary responses across the investigated concentration range.

REFERENCES

[1] BRUNO, A. C., *Magnetoacustic transducer application in gastric studies: with phanthoms and in vivo.* Ph.D. Thesis, Faculty of Philosophy, Sciences and Letters of Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil, 2015. doi: 10.11606/T.59.2015.tde-22042015-153203.

[2] VICENTE, T. T.; ARSALANI, S.; QUIEL, M. S. et al., *Improving the Theranostic Potential of Magnetic Nanoparticles by Coating with Natural Rubber Latex for Ultrasound, Photoacoustic Imaging, and Magnetic Hyperthermia: An In Vitro Study.* Pharmaceutics, vol. 16, no. 11, Art. no. 1474, 2024. doi: 10.3390/pharmaceutics16111474.

[3] SAMPAIO, D. R. T., *Laser- and X-Ray-Induced Photoacoustic Imaging*, Ph.D. Thesis, Faculty of Philosophy, Sciences and Letters of Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil, 2020. doi: 10.11606/T.59.2020.tde-21052020-163021.

[4] ARSALANI, S. et al., *Hybrid Nanoparticles of Citrate-Coated Manganese Ferrite and Gold Nanorods in Magneto-Optical Imaging and Thermal Therapy.* Nanomaterials, 2023. doi: 10.3390/nano13030434.

[5] E. E. MAZON, S. ARSALANI, J. H. ULIANA, A. A. O. CARNEIRO, A. J. GUALDI, and T. Z. PAVAN, *A pulsed magnetomotive ultrasound imaging system for magnetic nanoparticle detection.* 2021 IEEE UFFC Latin America Ultrasonics Symposium (LAUS), Gainesville, FL, USA, 2021, pp. 1-4. doi: 10.1109/LAUS53676.2021.9639204.

[6] J. H. ULIANA, D. R. T. SAMPAIO, A. F. SANCHES, T. Z. PAVAN, and A. A. d. O. CARNEIRO, *MUSA-WAVE: A Software Platform for Magnetomotive Ultrasound and Shear Wave Analysis 2024 IEEE UFFC Latin America Ultrasonics Symposium (LAUS)*, Montevideo, Uruguay, 2024, pp. 1–4. doi: 10.1109/LAUS60931.2024.10553040.

[7] HADADIAN, Y.; RAMOS, A. P.; PAVAN, T. Z., *Role of zinc substitution in magnetic hyperthermia properties of magnetite nanoparticles: interplay between intrinsic properties and dipolar interactions* Scientific Reports, vol. 9, 18048, 2019. doi: 10.1038/s41598-019-54250-7.

V. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by the São Paulo Research Foundation (FAPESP, Grant Nos. 2018/16939-8, 2021/06728-2, 2022/11418-5 and 2022/07463-5), the Brazilian National Council for Scientific and Technological Development (CNPq, Grant Nos. 311224/2021-0 and 421531/2025-8), and the Coordination for the Improvement of Higher Education Personnel (CAPES). The authors also thank the BIOMAG Research Group (UNESP, Botucatu) and the GIIMUS Research Group (USP, Ribeirão Preto) for their scientific and technical support, as well as for providing the infrastructure necessary to carry out this work.

Synthesis and Characterization of Cobalt Ferrite Magnetic Nanoparticles for Biomedical Applications

Leonardo Sonego
Dept. of Biophysics and Pharmacology
São Paulo State University (UNESP)
Botucatu, Brazil
l.sonego@unesp.br

Roberto Morato Fernandez
Dept. of Biophysics and Pharmacology
São Paulo State University (UNESP)
Botucatu, Brazil
morato.fernandez@unesp.br

Pedro Teixeira Ferraz
Dept. of Biophysics and Pharmacology
São Paulo State University (UNESP)
Botucatu, Brazil
p.ferraz@unesp.br

Guilherme Augusto Soares
Dept. of Biophysics and Pharmacology
São Paulo State University (UNESP)
Botucatu, Brazil
guilherme.soares@unesp.br

Michael Almeida
Dept. of Biophysics and Pharmacology
São Paulo State University (UNESP)
Botucatu, Brazil
mj.almeida@unesp.br

José Ricardo de Arruda Miranda
Dept. of Biophysics and Pharmacology
São Paulo State University (UNESP)
Botucatu, Brazil
jose.r.miranda@unesp.br

***Abstract*—Cobalt ferrite nanoparticles have high magnetocrystalline anisotropy, making them attractive candidates for magnetic hyperthermia and theranostics. However, their coercivity and magnetic dipole attraction favor macroscopic aggregation and rapid sedimentation. This work reports the synthesis of citrate-functionalized cobalt ferrite nanoparticles via co-precipitation, evaluating the effect of washing procedures and pH control on colloidal stability. Characterization was conducted using Alternating Current Biosusceptometry (ACB), Dynamic Light Scattering (DLS), Small-Angle X-ray Scattering (SAXS), and Electrophoretic Light Scattering (ELS). The non-optimized protocol showed macroscopic aggregation (hydrodynamic radius $R \approx 1380.5$ nm) and a negligible ACB signal (0.027 mV) due to particle settling. Exhaustive washing to lower ionic strength combined with surface charge adjustment to pH 8.85 yielded an ACB signal of 56.6 mV, stability across 30 days, and a mean zeta potential of -22 ± 1 mV. DLS and SAXS analyses confirmed a bimodal size distribution dominated by primary nanoparticles ($R = 6.9$ nm, 80.6% mass fraction) along with a secondary nanocluster fraction ($R = 52.8$ nm, 19.4% mass fraction) bounded by a maximum inorganic dimension of $D_{\max} \approx 75$ nm. Electrolyte removal and surface charge optimization were essential to obtain stable cobalt ferrite suspensions in aqueous media.**

***Index Terms*—Cobalt Ferrite, Alternating Current Biosusceptometry (ACB), Dynamic Light Scattering (DLS), Small-Angle X-ray Scattering (SAXS), Zeta Potential, Co-precipitation, Colloidal Stability.**

I. Introduction

Magnetic nanoparticles are widely investigated as platforms for combined diagnostic imaging and localized therapy in nanomedicine. Cobalt ferrite ($CoFe_2O_4$), an inverse spinel oxide, exhibits higher magnetocrystalline anisotropy and Specific Loss Power (SLP) than classical magnetite (Fe_3O_4), which favors magnetic fluid hyperthermia [1]–[3]. Previous reports indicate that small cobalt ferrite nanoparticles ($\approx$ 12 nm) produce high heating rates under alternating magnetic fields [4].

In $CoFe_2O_4$, Fe^{3+} cations occupy tetrahedral and octahedral interstitial sites, whereas Co^{2+} ions reside predominantly in octahedral sites, yielding a strong uncompensated magnetic moment and orbital angular momentum coupling [5]. In bulk form, cobalt ferrite exhibits high coercivity and remanence. When nanoparticles exceed the superparamagnetic limit or lack adequate surface charge, magnetic dipole-dipole attraction causes particle aggregation [6].

In physiological settings, macroscopic aggregates lead to sedimentation, vessel occlusion risks, and rapid clearance by the reticuloendothelial system (RES) in the liver and spleen [7], [8]. Uncontrolled clusters can also modify cellular toxicity profiles through reactive oxygen species generation [9].

Electrostatic stabilization using citrate ions is commonly employed to control aggregation. Citrate binds to surface metal sites via carboxyl groups, leaving uncoordinated carboxylates (COO^-) to generate repulsive forces. Still, electrostatic stabilization remains vulnerable to ionic screening. Residual synthesis salts, such as NaCl, compress the electric double layer (EDL), screening surface charges and enabling dipolar agglomeration [10].

Here, we evaluate the synthesis, purification, and colloidal properties of citrate-functionalized $CoFe_2O_4$ nanoparticles prepared via co-precipitation. We compare a preliminary purification route with an optimized washing protocol, examining the relationship between electrolyte removal, pH, hydrodynamic size, inorganic core structure, zeta potential, and magnetic response in vitro. Characterization was carried out using Alternating Current Biosusceptometry (ACB), Dynamic Light Scattering (DLS), Small-Angle X-ray Scattering (SAXS), and Electrophoretic Light Scattering (ELS).

II. Materials and Methods

A. Chemical Synthesis and Stoichiometry

Cobalt ferrite nanoparticles were synthesized by chemical co-precipitation. Divalent cobalt and trivalent iron hydroxides

were precipitated in an alkaline medium, followed by ferritization at $80\,^{\circ}$C:

$$\mathrm{Co}^{2+}_{(\mathrm{aq})} + 2\mathrm{OH}^{-}_{(\mathrm{aq})} \rightarrow \mathrm{Co(OH)}_{2(\mathrm{s})}, \tag{1}$$

$$2\mathrm{Fe}^{3+}_{(\mathrm{aq})} + 6\mathrm{OH}^{-}_{(\mathrm{aq})} \rightarrow 2\mathrm{Fe(OH)}_{3(\mathrm{s})}. \tag{2}$$

At pH $\approx$ 12 under continuous stirring, the hydroxides condense to form the spinel phase:

$$\mathrm{Co(OH)}_{2(\mathrm{s})} + 2\mathrm{Fe(OH)}_{3(\mathrm{s})} \rightarrow \mathrm{CoFe_2O_{4(s)}} + 4\mathrm{H_2O_{(l)}}. \tag{3}$$

The overall reaction using chloride precursors and sodium hydroxide is:

$$\begin{aligned}\mathrm{CoCl_{2(aq)}} + 2\mathrm{FeCl_{3(aq)}} + 8\mathrm{NaOH_{(aq)}} \rightarrow \\ \mathrm{CoFe_2O_{4(s)}} + 8\mathrm{NaCl_{(aq)}} + 4\mathrm{H_2O_{(l)}}.\end{aligned} \tag{4}$$

A stoichiometric $1:2$ molar ratio ($\mathrm{Co}^{2+}:\mathrm{Fe}^{3+}$) was set to synthesize 1.0 g of $\mathrm{CoFe_2O_4}$ (231.83 g/mol). Precursor salts (0.558 g $\mathrm{CoCl_2}$, 129.84 g/mol; 1.395 g $\mathrm{FeCl_3}$, 162.20 g/mol) were dissolved in 10 mL of deionized water acidified with concentrated HCl to prevent premature Fe^{3+} hydrolysis. This solution was added to 100 mL of 2 mol/L NaOH under magnetic stirring (1000 rpm).

B. Purification, Surface Treatment, and pH Control

Two purification methods were compared:

- **Preliminary Protocol:** Magnetic decantation with a standard ferrite magnet, followed by 3 washing cycles with deionized water without monitoring electrical conductivity or pH.
- **Optimized Protocol:** Magnetic decantation using high-flux NdFeB magnets, followed by surface treatment with dilute nitric acid ($\mathrm{HNO_3}$, 0.01 mol/L) to remove uncoordinated surface metal species. The precipitate was coated with trisodium citrate (0.1 mol/L) at $80\,^{\circ}$C and repeatedly washed with deionized water. Washing continued until supernatant electrical conductivity dropped below $10\,\mu$S/cm, confirming the removal of residual spectator ions (Na^{+}, Cl^{-}). The final suspension was adjusted to pH 8.85.

The target pH of 8.85 was chosen because it lies above the isoelectric point of ferrite (IEP $\approx$ 3.5–4.0) and exceeds the pK_a values of citric acid (p$K_{a1} = 3.13$, p$K_{a2} = 4.76$, p$K_{a3} = 6.40$). At this pH, carboxyl groups are deprotonated into COO^{-}, generating negative surface charge to reduce sedimentation.

C. Alternating Current Biosusceptometry (ACB)

AC magnetic susceptibility was evaluated using an ACB system consisting of a first-order gradiometric pickup coil array coupled to a lock-in amplifier. In the absence of a sample, differential flux is zero.

When a magnetic sample is introduced, the induced magnetic flux variation (Φ_B) produces an electromotive force (emf) according to Faraday's law:

$$\mathrm{emf} = -d\Phi_B/dt. \tag{5}$$

The induced voltage (mV) is proportional to the AC susceptibility and nanoparticle mass within the active sensing volume. ACB measurements were recorded in triplicate at 10-day intervals over 30 days under constant excitation (10 kHz, 2.0 mT).

D. Dynamic Light Scattering and Zeta Potential

Hydrodynamic radius (R_h) and polydispersity (%Pd) were measured by DLS (DYNAMICS, Wyatt Technology). Samples were diluted $1:1000$ in deionized water. The translational diffusion coefficient (D) was converted to R_h using the Stokes-Einstein equation:

$$R_h = \frac{k_B T}{6\pi\eta D}, \tag{6}$$

where k_B is the Boltzmann constant (1.38×10^{-23} J/K), $T = 298.15$ K, and $\eta = 0.89$ mPa $\cdot$ s. Regularization algorithms were applied for multimodal distribution analysis [11].

Surface charge and zeta potential (ζ) of the optimized sample were determined in triplicate by Electrophoretic Light Scattering (ELS) under an applied field of 18.51 V/cm, using the Smoluchowski relation.

E. Small-Angle X-ray Scattering (SAXS)

Small-Angle X-ray Scattering (SAXS) was used to assess inorganic core dimensions and size distribution in liquid dispersion. Scattering profiles $I(q)$ were acquired as a function of the momentum transfer $q = (4\pi/\lambda)\sin\theta$. Indirect Fourier transformation was used to calculate the volume-weighted size distribution $D_v(r)$ as a function of the characteristic real-space dimension r and determine the maximum cluster dimension ($D_{\max}$).

III. Results and Discussion

A. ACB Magnetic Signal and Temporal Stability

ACB measurements showed clear differences between the preparation routes (Table I). The preliminary protocol gave a low induced voltage (0.027 ± 0.005 mV), remaining near the electronic noise floor of the system.

TABLE I
Temporal ACB Induced Voltage Signal Comparison

	Induced Voltage Signal (mV)	
Measurement Interval	**Preliminary Protocol**	**Optimized Protocol**
Day 10	0.030 ± 0.005	55.4 ± 0.8
Day 20	0.030 ± 0.004	56.6 ± 0.8
Day 30	0.020 ± 0.003	56.6 ± 0.7
Overall Mean	$\mathbf{0.027 \pm 0.005}$	$\mathbf{56.2 \pm 0.8}^{*}$

*Significant difference compared to preliminary protocol ($p < 0.001$).

The optimized protocol produced a stable average signal of 56.2 ± 0.8 mV. An unpaired two-tailed Student's t-test indicated a statistically significant difference between protocols

($p < 0.001$). One-way ANOVA across the 30-day window showed no significant signal decay ($p > 0.05$), confirming that the particles remained suspended over time in deionized water.

B. Hydrodynamic Size, Zeta Potential, and DLVO Analysis

DLS data accounted for the differences observed in the ACB response. In the preliminary protocol (Fig. 1), DLS revealed substantial aggregation. Although primary particles ($R = 4.6$ nm) represented 74.5% of the estimated mass, light scattering intensity was dominated (95.3%) by a macro-aggregate population ($R \approx 1380.5$ nm, $\%\text{Pd} = 161.1\%$). Because Rayleigh scattering scales with $I \propto R^6$, large aggregates accounted for nearly all scattered light and settled rapidly, removing magnetic material from the sensing volume of the ACB coils.

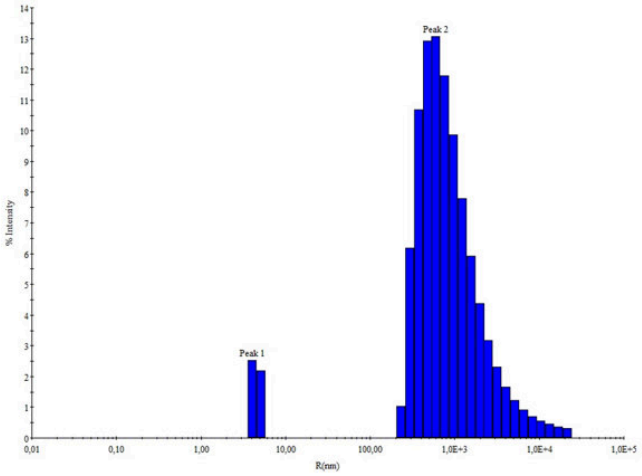


Fig. 1. Hydrodynamic radius intensity distribution for the preliminary non-optimized protocol, showing macroscopic aggregation ($R \approx 1380.5$ nm).

According to DLVO theory, unwashed Na^+ and Cl^- ions increase solution ionic strength and compress the double layer. This reduces electrostatic repulsion, allowing attractive van der Waals and magnetic dipole interactions to cause coagulation [10].

In the optimized protocol, electrolyte removal and pH adjustment suppressed the > 1000 nm aggregate population (Fig. 2). The ACB signal increased by roughly 2080 times (from 0.027 mV to 56.60 mV), with 80.6% of the mass concentrated in primary nanoparticles ($R = 6.9$ nm, hydrodynamic diameter $D_h \approx 13.8$ nm), as summarized in Table II.

TABLE II
COMPARATIVE DLS MULTIMODAL SIZE DISTRIBUTION PARAMETERS

	Preliminary Protocol		Optimized Protocol	
Parameter	**Peak 1**	**Peak 2**	**Peak 1**	**Peak 2**
Radius (R)	4.6 nm	1380.5 nm	6.9 nm	52.8 nm
Polydispersity (%Pd)	11.9%	161.1%	13.8%	43.6%
Intensity (%Int)	4.7%	95.3%	2.9%	97.1%
Mass (%Mass)	74.5%	23.5%	80.6%	19.4%

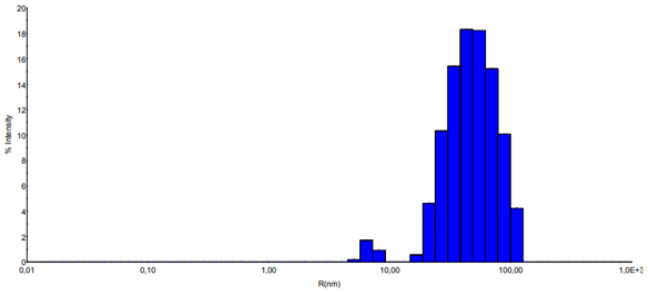


Fig. 2. Hydrodynamic radius intensity distribution for the optimized protocol, showing a bimodal distribution ($R_1 = 6.9$ nm, $R_2 = 52.8$ nm).

Electrophoretic light scattering indicated a mean zeta potential of -22 ± 1 mV (Fig. 3, Table III), confirming negative surface charge from deprotonated carboxylates at pH 8.85.

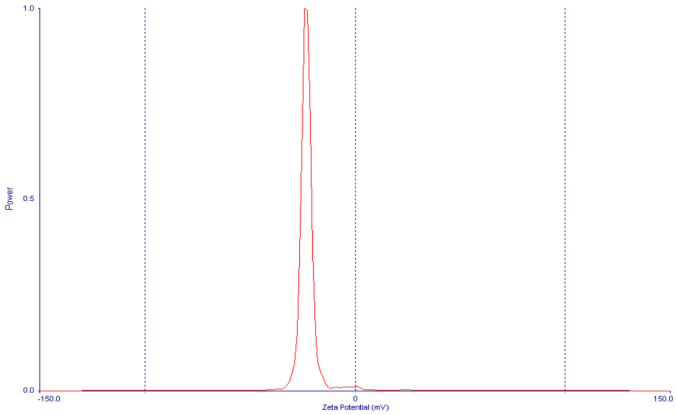


Fig. 3. Zeta potential distribution of the optimized citrate-functionalized $CoFe_2O_4$ suspension.

TABLE III
ELECTROPHORETIC LIGHT SCATTERING ZETA POTENTIAL RUNS

Run	**Zeta Potential (mV)**	**Half Width (mV)**	**Data Retention (%)**
1	-19.76	5.91	44%
2	-22.46	2.80	44%
3	-23.87	2.60	67%
Mean $\pm$ SD	$\mathbf{-22 \pm 1}$	$\mathbf{4 \pm 1}$	**52%**

From a DLVO perspective, a zeta potential of -22 mV provides moderate stability, whereas absolute values above 30 mV are normally required for full electrostatic repulsion. Because cobalt ferrite has strong magnetic anisotropy and permanent dipoles, this moderate charge does not completely eliminate dipole-dipole attractions [6]. As a result, a secondary population of small clusters ($R = 52.8$ nm, 19.4% mass fraction) remained in suspension.

C. Nanostructural Validation via SAXS

SAXS measurements were used to probe inorganic core dimensions in liquid dispersion (Fig. 4). The scattering curve

$I(q)$ decayed smoothly and matched the model fit across the entire q-range, showing no scattering features associated with micron-sized aggregates (Fig. 4a).

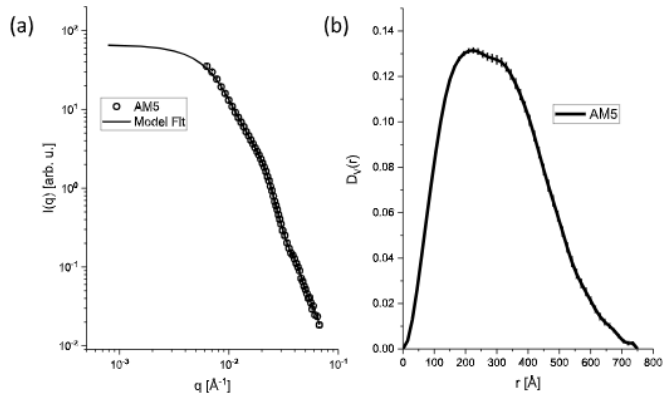


Fig. 4. SAXS data for the optimized $CoFe_2O_4$ suspension: (a) scattering intensity $I(q)$ and model fit, and (b) volume distribution $D_v(r)$ as a function of characteristic real-space dimension r.

The volume-weighted size distribution $D_v(r)$ extracted by indirect Fourier transformation shows a peak centered at $r \approx 230$ Å (23 nm) and reaches a maximum dimension of $D_{\max} \approx 750$ Å (75 nm) (Fig. 4b).

When comparing SAXS and DLS, the physical differences between both techniques must be considered. DLS measures the hydrodynamic radius (R_h), which includes the inorganic core, surface citrate molecules, and the surrounding hydration layer. In contrast, SAXS detects the electron-dense inorganic core. Because scattering scales with particle volume ($I \propto V^2$), the SAXS $D_v(r)$ curve is weighted toward the small cluster population rather than isolated cores ($R_h = 6.9$ nm, $D_h \approx 13.8$ nm).

The maximum inorganic dimension ($D_{\max} \approx 75$ nm) from SAXS is consistent with the cluster population identified by DLS ($R_h = 52.8$ nm, or $D_h \approx 105.6$ nm). The ≈ 30 nm difference between techniques stems from the distinct weighting in SAXS (volume-weighted) versus DLS ($I \propto R^6$), the contribution of the citrate/hydration layer (1–3 nm), and the increased hydrodynamic drag of non-spherical clusters. SAXS and DLS therefore point to the same conclusion: the secondary fraction consists of sub-100 nm clusters rather than macro-aggregates. This prevented gravitational settling and preserved the ACB signal over time.

D. Perspectives and Biological Considerations

The synthesis yielded primary cores ($R = 6.9$ nm) and clusters limited to < 75 nm, but stability in deionized water does not ensure the same behavior in physiological environments. At physiological salt concentrations (≈ 150 mmol/L NaCl), electrolyte screening compresses the electrical double layer, reducing the -22 mV electrostatic barrier [8]. Serum protein adsorption also forms a protein corona, which can trigger agglomeration and accelerate macrophage clearance in the liver and spleen [7], [8].

Future studies should assess colloidal stability under physiological conditions. In particular, DLS, zeta potential, and ACB measurements in PBS and protein-supplemented media (such as FBS or plasma) should be monitored immediately after dispersion and after 24–48 hours. To suppress the 19.4% cluster fraction and prevent salt-induced aggregation, steric coatings like PEGylation or amphiphilic polymers will be required prior to in vitro cytotoxicity, hemocompatibility, and in vivo biodistribution experiments [12].

IV. Conclusion

Controlling electrolyte content and pH improved the colloidal stability of cobalt ferrite nanoparticles synthesized by co-precipitation. Washing the suspension to an electrical conductivity below $10\,\mu$S/cm and setting the pH to 8.85 produced a zeta potential of -22 ± 1 mV and increased the ACB signal to 56.6 mV. DLS and SAXS confirmed a bimodal size distribution dominated by primary nanoparticles ($R = 6.9$ nm, 80.6% mass), alongside a cluster fraction bounded below 75 nm (19.4% mass), demonstrating that citrate electrostatic repulsion alone cannot fully overcome magnetic dipole attractions. Future work will focus on stability tests in PBS and protein media across 24–48 hours, combined with steric polymer coatings (such as PEG) before proceeding to biological evaluation.

References

[1] R. Hergt et al., "Magnetic particle hyperthermia: nanoparticle magnetism and materials development for cancer therapy," *J. Phys. Condens. Matter*, vol. 18, no. 38, pp. S2919–S2934, 2006.

[2] K. Maaz et al., "Synthesis and magnetic properties of cobalt ferrite (CoFe2O4) nanoparticles prepared by wet chemical route," *J. Magn. Magn. Mater.*, vol. 308, no. 2, pp. 289–295, 2007.

[3] S. Khizar et al., "Aminodextran Coated CoFe2O4 Nanoparticles for Combined Magnetic Resonance Imaging and Hyperthermia," *Nanomaterials*, vol. 10, no. 11, p. 2182, 2020.

[4] A. S. Garanina et al., "Cobalt ferrite nanoparticles for magnetic hyperthermia," *Nanomedicine*, vol. 17, no. 4, pp. 211–225, 2022.

[5] D. S. Mathew and R. S. Juang, "An overview of the structure and magnetism of spinel ferrite nanoparticles and their synthesis in microemulsions," *Chem. Eng. J.*, vol. 129, no. 1–3, pp. 51–65, 2007.

[6] V. Socoliuc et al., "Magnetic Nanoparticle Systems for Nanomedicine—A Materials Science Perspective," *Magnetochemistry*, vol. 6, no. 1, pp. 1–36, 2020.

[7] S. Laurent et al., "Magnetic Iron Oxide Nanoparticles: Synthesis, Stabilization, Vectorization, Physicochemical Characterizations, and Biological Applications," *Chem. Rev.*, vol. 108, no. 6, pp. 2064–2110, 2008.

[8] Z. Belhadj et al., "A combined 'eat me/don't eat me' strategy based on extracellular vesicles for anticancer nanomedicine," *J. Extracell. Vesicles*, vol. 9, no. 1, p. 1806444, 2020.

[9] M. J. Akhtar et al., "Size-dependent cytotoxicity and toxicity mechanisms of cobalt ferrite nanoparticles," *Int. J. Nanomedicine*, vol. 14, pp. 2457–2471, 2019.

[10] L. Jiang et al., "pH-dependent aggregation of citrate-capped Au nanoparticles induced by Cu2+ ions," *Colloids Surf. A Physicochem. Eng. Asp.*, vol. 346, pp. 216–220, 2009.

[11] S. W. Provencher, "A constrained regularization method for inverting data represented by linear algebraic or integral equations," *Comput. Phys. Commun.*, vol. 27, no. 3, pp. 213–227, 1982.

[12] K. Huang et al., "Size-dependent localization and penetration of ultrasmall gold nanoparticles in cancer cells, multicellular spheroids, and tumors in vivo," *ACS Nano*, vol. 6, no. 5, pp. 4483–4493, 2012.

Electrospun PVA-Based Nonwoven Mats with Silver Nanoparticles and Neomycin Sulfate: Development and Antimicrobial Evaluation

Thainara Gouvêa
Insitituto de Pesquisa e Desenvolvimento – Lab.Bionanotecnology
Universidade do Vale do Paraíba
São José dos Campos, Brazil
thaigouvea20@gmail.com

Lara Rebouças
Insitituto de Pesquisa e Desenvolvimento – Lab.Bionanotecnology
Universidade do Vale do Paraíba
São José dos Campos, Brazil
larasilvasouzareboucas@hotmail.com

Lúcia Vieira
Insitituto de Pesquisa e Desenvolvimento – Lab.Nanotecnology and Plasma Process
Universidade do Vale do Paraíba
São José dos Campos, Brazil
lucia.vieira@univap.br

Maiara Castilho
Insitituto de Pesquisa e Desenvolvimento – Lab.Bionanotecnology
Universidade do Vale do Paraíba
São José dos Campos, Brazil
mcastilho@univap.br

***Abstract*— Bacterial resistance and chronic wound infections motivate the development of antimicrobial dressings able to combine structural support and local active delivery. This study produced electrospun polyvinyl alcohol nonwoven mats functionalized with silver nanoparticles and neomycin sulfate, isolated or associated, and evaluated their chemical, morphological and antimicrobial performance against *Staphylococcus aureus* and *Escherichia coli*. A 10 percent PVA solution with 0.5 percent citric acid was electrospun after incorporation of the active agents. The systems were characterized by scanning electron microscopy, energy dispersive spectroscopy, ultraviolet visible spectroscopy and Fourier transform infrared spectroscopy, and antimicrobial activity was tested by agar diffusion. Scanning electron microscopy showed spherical silver nanoparticles with mean diameter of 77.7 nm and continuous fibers with mean diameter close to 314 nm after functionalization. Energy dispersive spectroscopy confirmed silver in the composite, and infrared spectroscopy suggested preservation of the PVA structure with interactions between the polymeric matrix and the incorporated agents. Pure PVA did not inhibit bacterial growth. The combined silver nanoparticle and neomycin system produced the largest inhibition halos for both bacterial strains, increasing inhibition from 16.07 mm to 20.53 mm for *S. aureus* and from 8.97 mm to 15.64 mm for *E. coli* compared with neomycin alone. The results show larger inhibition halos for the combined formulation under the tested conditions and support further investigation of electrospun PVA containing silver nanoparticles and neomycin for antimicrobial wound dressing applications.**

***Keywords*—electrospinning; polyvinyl alcohol; silver nanoparticles; neomycin sulfate; wound dressing; antimicrobial activity**

I. INTRODUCTION

Chronic and infected wounds remain an important public health challenge because delayed tissue repair increases the risk of local infection, systemic complications, prolonged treatment, and higher healthcare costs [1]. The skin acts as the main physical barrier of the organism, and its disruption becomes especially critical in patients with impaired healing, such as individuals with diabetes or other comorbidities, in whom inflammatory response, collagen deposition, angiogenesis, and tissue regeneration may be compromised [1], [2]. Therefore, wound dressings are no longer considered only passive protective barriers; advanced dressings are expected to maintain a moist environment, manage exudate, protect the lesion, and, when possible, deliver therapeutic agents directly to the wound site [2], [3].

This need becomes even more relevant in the context of antimicrobial resistance. The excessive and inappropriate use of antibiotics has accelerated the emergence of resistant bacterial strains, making infections more difficult to treat and increasing the clinical relevance of localized antimicrobial systems [4]. In wound infections, *Staphylococcus aureus* and *Escherichia coli* are frequently used as representative experimental models because they differ in cell-wall organization and antimicrobial susceptibility. Gram-positive bacteria, such as *S. aureus*, have a thick peptidoglycan layer, whereas Gram-negative bacteria, such as *E. coli*, possess an additional outer membrane that may act as a permeability barrier and reduce the effectiveness of antimicrobial compounds [5].

Thus, the development of wound dressings capable of combining structural support with local antimicrobial action represents a promising strategy to address infected wounds while reducing dependence on systemic antibiotic exposure. For this purpose, the dressing matrix must provide adequate porosity, high surface area, close contact with the wound bed, and the possibility of incorporating active agents in a controlled and homogeneous manner. In this scenario, electrospinning has

emerged as a simple and versatile technique for producing polymeric nanofibers from solutions subjected to a high-voltage electric field [6], [7], [8]. During the process, the polymeric solution is stretched into fine fibers that are deposited as a nonwoven mat, producing materials with high surface area, interconnected porosity, and structural similarity to the extracellular matrix [6], [7]. These characteristics are especially attractive for wound dressing applications because they may support fluid exchange, improve contact with the wound bed, and allow the incorporation of bioactive agents, including nanoparticles and antibiotics [3], [9].

Among the polymers used for electrospun wound dressings, polyvinyl alcohol (PVA) is widely investigated due to its hydrophilicity, biocompatibility, low toxicity, low cost and ability to form continuous fibers [9], [10]. Its hydroxyl-rich structure favors water retention and interaction with incorporated compounds, making it suitable for the development of functional wound dressing platforms [9], [10]. However, although PVA-based materials present favorable physicochemical properties for biomedical use, PVA itself does not exhibit intrinsic antibacterial activity. Therefore, the incorporation of antimicrobial agents into the PVA matrix is an important strategy to improve its biological performance [10], [11].

Neomycin sulfate is an aminoglycoside antibiotic commonly associated with topical antimicrobial strategies due to its activity against clinically relevant bacteria and its applicability in localized treatments. In wound-related applications, the local delivery of antibiotics may increase the antimicrobial concentration at the infection site while minimizing systemic exposure and associated adverse effects. However, the widespread and prolonged use of antibiotics has contributed to the emergence of antimicrobial resistance, reducing the effectiveness of conventional therapies and highlighting the need for alternative approaches to control wound infections [4]. In this study, neomycin sulfate was selected as the antibiotic model because it is compatible with topical antimicrobial approaches and allows the evaluation of antibacterial performance when incorporated into an electrospun PVA matrix. Its association with the polymeric dressing also enables comparison between the isolated antibiotic effect and the combined effect with silver nanoparticles.

In this context, silver nanoparticles (AgNPs) have attracted considerable attention because of their broad-spectrum antimicrobial activity and their multiple mechanisms of action. AgNPs may interact with bacterial membranes, proteins, and nucleic acids, promoting membrane disruption, oxidative stress, enzyme inhibition, and impairment of microbial replication [5], [12]. Furthermore, their incorporation into electrospun polymeric fibers can provide localized antimicrobial activity and improve the functionality of wound dressing systems [3], [10]. Previous studies have reported enhanced antibacterial responses when AgNPs are combined with conventional antibiotics, with proposed mechanisms including increased bacterial membrane permeability and simultaneous action on multiple cellular targets [13]. These literature reports provide a mechanistic rationale for evaluating the combined formulation, but they do not establish synergy in the present material.

Based on this rationale, the present study aimed to develop electrospun PVA nonwoven mats containing AgNPs and neomycin sulfate, either alone or in association, and to evaluate their antimicrobial activity against *S. aureus* and *E. coli*. This approach provides a materials-engineering basis for further evaluation of multifunctional nanofibrous wound dressing systems.

II. Materials and Methods

This section describes the experimental strategy used to produce and evaluate electrospun PVA-based nonwoven mats associated with silver nanoparticles (AgNPs) and neomycin sulfate (Neo). The methodology comprised the preparation of the biological model and active compounds, fabrication of the electrospun mats, and evaluation of their physicochemical, morphological and antimicrobial properties.

A. Electrospinning of PVA-based nonwoven mats

The electrospinning solution consisted of 10% (w/v) poly(vinyl alcohol) (PVA) and 0.5% (w/v) citric acid dissolved in distilled water under continuous magnetic stirring at approximately 45 °C until complete dissolution. The solution was then cooled to room temperature before use.

Four electrospinning formulations were prepared: PVA, PVA/AgNPs, PVA/Neo, and PVA/AgNPs/Neo. The control formulation (PVA) consisted only of the polymeric solution. For the PVA/AgNPs formulation, 2.13 mg of silver nanoparticles purchased from Nanostructured & Amorphous Materials, Inc. were added to 1.5 mL of the polymeric solution and stirred at approximately 45 °C until a homogeneous dispersion was obtained. The PVA/Neo formulation was prepared by incorporating 3.75 mg of neomycin sulfate into 1.5 mL of the polymeric solution under the same conditions. For the dual-functionalized formulation (PVA/AgNPs/Neo), 2.13 mg of AgNPs and 3.75 mg of neomycin sulfate were simultaneously added to 1.5 mL of the polymeric solution and mixed until complete homogenization. These values represent the nominal amounts added during formulation; retained loading in the final mats and the release of silver and neomycin were not quantified in this study.

Electrospinning was performed using a syringe fitted with a stainless-steel needle connected to a syringe pump operating at a flow rate of 1 mL h^{-1}. A grounded copper collector covered with aluminum foil was positioned 10 cm from the needle tip, and a voltage of 15 kV was applied between the needle and the collector. All samples were produced under low relative humidity conditions and electrospun for 3 h. Identical processing parameters were employed for all formulations to ensure reproducibility and enable direct comparison of their physicochemical, morphological, and antimicrobial properties.

B. Morphological characterization

Scanning electron microscopy (SEM) was employed to evaluate the morphology of the silver nanoparticles (AgNPs) and the electrospun PVA-based nonwoven mats. Samples were cut into small fragments, mounted on aluminum stubs using conductive carbon tape, and sputter-coated with a thin layer of gold to improve surface conductivity prior to imaging. SEM analyses were performed under high-resolution conditions using an accelerating voltage ranging from 9 to 47.5 kV. Micrographs were acquired at different magnifications to assess fiber morphology, diameter uniformity, and nanoparticle distribution within the polymeric matrix. Fiber and AgNP diameters were measured from the SEM images using ImageJ software and expressed as diameter distributions. The measurements were used descriptively; no independent batch-to-batch inferential analysis was performed for the diameter distributions. In addition, energy-dispersive X-ray spectroscopy (EDS) was performed to confirm the elemental composition and verify the presence and distribution of silver in the PVA/AgNPs and PVA/AgNPs/Neo nonwoven mats.

C. Antimicrobial assay and data analysis

Reference strains of *Escherichia coli* ATCC 25922 (Gram-negative) and *Staphylococcus aureus* ATCC 25923 (Gram-positive) were provided by the Oswaldo Cruz Laboratory (São José dos Campos, Brazil). Before the antimicrobial assays, both strains were cultured in Brain Heart Infusion (BHI) broth and incubated at 37 °C for 24 h. The cultures were subsequently streaked onto BHI agar plates to obtain isolated colonies, and bacterial purity was confirmed by Gram staining. Fresh bacterial suspensions were prepared from isolated colonies and standardized to 0.5×10^6 CFU/mL for the antimicrobial assays.

The antimicrobial activity of the electrospun nonwoven mats (PVA, PVA/AgNPs, PVA/Neo, and PVA/AgNPs/Neo) was evaluated using the agar diffusion method. Circular samples (10 mm in diameter) were placed onto BHI agar plates previously inoculated with the standardized bacterial suspensions. The plates were incubated at 37 °C for 24 h, after which the inhibition zones were photographed and measured using ImageJ software. The agar-diffusion measurements were performed in triplicate, and inhibition zone diameters were expressed as mean ± standard deviation. The triplicates were treated as replicate measurements of this screening assay. Results were analyzed descriptively, and no inferential statistical test or formal interaction/synergy analysis was performed. Therefore, comparisons among formulations are reported as differences in inhibition-zone diameter under the tested conditions.

III. Results and Discussion

The results were organized to connect the structural characterization of the materials with their antimicrobial performance. First, SEM analysis was used to evaluate the morphology and size distribution of the AgNPs and the electrospun fibers. EDS was performed to confirm the elemental composition and the presence of silver in the electrospun nonwoven mats. Finally, the disk diffusion assay was used to compare the antimicrobial activity of the formulations containing individual and combined active agents.

SEM analysis of the commercial AgNPs (Fig. 1) revealed predominantly spherical nanoparticles with a mean diameter of 77.7 nm. Although some degree of aggregation was observed, the particles remained within the nanometer size range, which is desirable for antimicrobial applications. The morphology shown in Fig. 1A is relevant because nanoparticle size and shape may influence dispersion in the polymeric solution, interaction with the electrospun fibers, and antimicrobial performance. This finding agrees with studies reporting that AgNPs present antimicrobial potential associated with their nanoscale dimensions and interaction with bacterial structures [5], [12]. Furthermore, the particle size distribution histogram (Fig. 1B) confirmed the predominance of nanoparticles within the nanometric range, supporting the dimensional analysis obtained from the SEM micrographs.

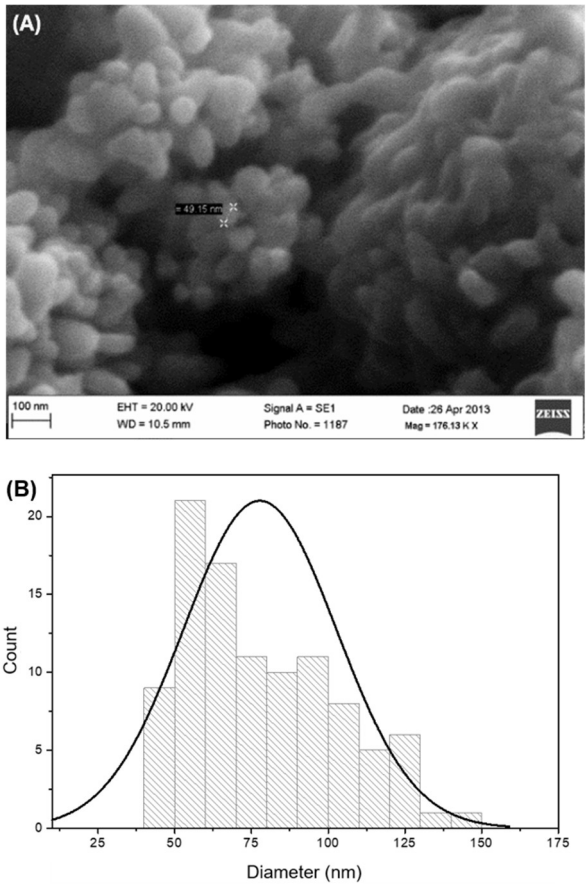


Fig. 1. Morphological characterization of the commercial silver nanoparticles (AgNPs) by scanning electron microscopy (SEM). (A) Representative SEM micrograph showing predominantly spherical nanoparticles. (B) Particle size distribution histogram based on diameter measurements obtained from the SEM images.

After confirming the nanometric dimensions of the AgNPs, SEM was used to evaluate the morphology of the electrospun nonwoven mats before and after incorporation of the

antimicrobial agents. As shown in Fig. 2B, the neat PVA mat exhibited a continuous, homogeneous, and defect-free fibrous network, indicating stable electrospinning under the selected processing conditions. Similarly, the functionalized PVA/AgNPs/Neo mat (Fig. 2C) preserved a homogeneous fibrous architecture, demonstrating that the incorporation of AgNPs and neomycin did not adversely affect fiber formation or the structural integrity of the electrospun matrix. The fibers exhibited a mean diameter of approximately 314 nm (Fig. 2A). In addition, discrete bright regions were observed along the fibers, suggesting the presence of electron-dense particles associated with the polymeric matrix.The preservation of the fibrous morphology is consistent with previous reports showing that electrospinning can produce polymeric nanofiber mats with high surface area, porosity, and structural similarity to the extracellular matrix, which are important characteristics for wound dressing applications [6], [7], [9]. In addition, PVA-based electrospun mats containing AgNPs have been described as suitable platforms for antimicrobial dressings, since the polymeric matrix can support nanoparticle incorporation while maintaining a nanofibrous structure [10]. Therefore, the morphology observed in Fig. 2 suggests that the selected processing conditions were adequate for producing PVA-based nonwoven mats containing antimicrobial agents.

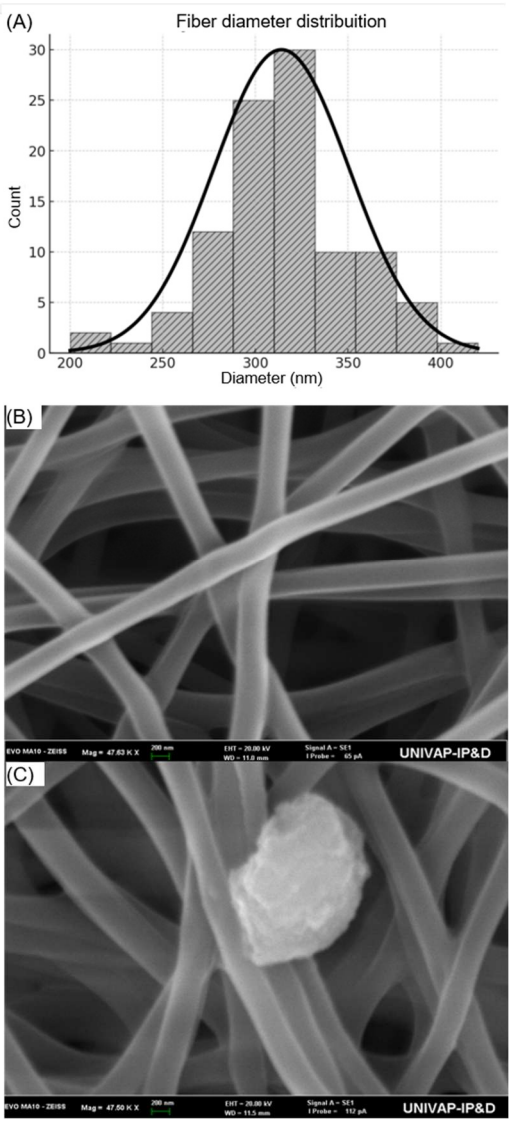


Fig. 2. SEM characterization of the electrospun PVA-based nonwoven mats. (A) Fiber diameter distribution histogram obtained from SEM image analysis. (B) Representative SEM micrograph of the neat PVA nonwoven mat showing a continuous and homogeneous fibrous network. (C) Representative SEM micrograph of the PVA/AgNPs/Neo nonwoven mat showing the incorporation of antimicrobial agents while maintaining the fibrous morphology.

To verify the elemental composition of the functionalized fibers, energy-dispersive X-ray spectroscopy (EDS) was performed. As shown in Fig. 3, the characteristic silver peak confirmed the incorporation of AgNPs into the PVA/AgNPs/Neo nonwoven mat. Gold peaks were also detected due to the gold sputter-coating applied prior to SEM analysis. In contrast, neomycin could not be directly identified because it is predominantly composed of low atomic number elements, such as C, H, O, and N, whose signals overlap with those of the PVA matrix. This limitation is consistent with the analytical nature of SEM/EDS, which is more suitable for identifying elemental composition, especially heavier elements, than for distinguishing organic molecules within polymeric matrices [14]. Therefore, EDS primarily served to confirm the

presence of the metallic component within the electrospun composite.

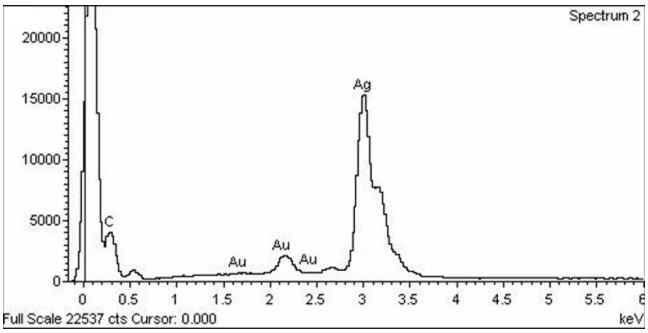


Fig. 3. Energy-dispersive X-ray spectroscopy (EDS) analysis of the PVA nonwoven material incorporated with silver nanoparticles (AgNPs) and antibiotic.

The antimicrobial performance of the electrospun nonwoven mats was evaluated by disk diffusion assay against *Staphylococcus aureus* ATCC 25923 and *Escherichia coli* ATCC 25922. The representative inhibition zones are shown in Fig. 4, while the quantitative results are summarized in Tables I and II. The neat PVA mat exhibited no inhibition halo against either bacterial strain, confirming that the polymeric matrix did not present intrinsic antibacterial activity under the tested conditions. This result is in agreement with previous reports indicating that PVA-based materials, despite their biomedical potential, generally require association with antimicrobial agents to provide antibacterial activity [10], [11]. Because diffusion-zone diameter depends on both antimicrobial action and the mobility of the active compounds through agar, halo size was interpreted as a screening outcome rather than as a direct measure of antimicrobial potency.

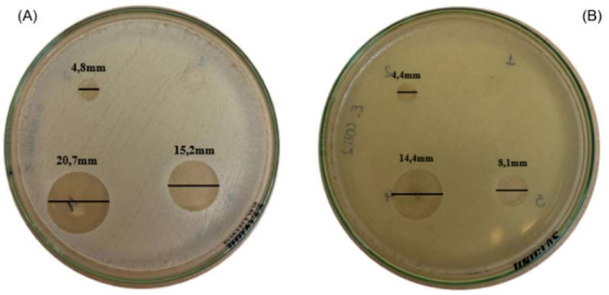


Fig. 4. Antibacterial activity of the electrospun nonwoven mats evaluated by the disk diffusion assay. (A) Staphylococcus aureus ATCC 25923. (B) Escherichia coli ATCC 25922.

For *Staphylococcus aureus*, the results presented in Table I demonstrate that the incorporation of AgNPs into the PVA matrix resulted in a moderate antibacterial response, with an inhibition halo of 5.954 ± 0.99 mm. This effect can be attributed to the antimicrobial properties of AgNPs, which have been associated with bacterial membrane damage, interaction with proteins and nucleic acids, and inhibition of microbial growth [5], [12]. However, the relatively smaller halo observed for PVA/AgNPs may be related to the immobilization of nanoparticles within the polymeric network, which can limit their diffusion through the agar medium compared with freely dispersed antimicrobial agents.

TABLE I
ANTIBACTERIAL ACTIVITY OF ELECTROSPUN NONWOVEN MATS AGAINST *Staphylococcus aureus* ATCC 25923.

Sample	Mean (mm)	SD (mm)
PVA	0.000	0.00
PVA with AgNPs	5.954	0.99
PVA with neomycin	16.068	0.70
PVA with AgNPs and neomycin	20.528	0.24

Source: Authors.

The PVA/Neo formulation exhibited a greater inhibitory effect against *S. aureus*, producing a mean inhibition halo of 16.068 ± 0.70 mm (Table I). This result indicates antibacterial activity of neomycin against the Gram-positive strain under the tested conditions. The largest inhibition halo was observed for the PVA/AgNPs/Neo formulation, which reached 20.528 ± 0.24 mm (Table I and Fig. 4A). Compared with the individual formulations, the combined system increased the inhibition halo by approximately 27.8% compared with PVA/Neo and by more than three-fold compared with PVA/AgNPs. Thus, the combined formulation produced an enhanced halo response under the tested conditions.

For *Escherichia coli*, the results summarized in Table II showed a similar trend, although the inhibition halos were lower than those observed for *S. aureus*. The PVA mat exhibited no antibacterial activity, while the PVA/AgNPs formulation produced a mean inhibition halo of 5.473 ± 1.19 mm. The PVA/Neo formulation showed an increased inhibitory effect, reaching 8.967 ± 0.72 mm (Table II, Fig. 4B). The reduced susceptibility of *E. coli* compared with *S. aureus* may be associated with the presence of an additional outer membrane characteristic of Gram-negative bacteria, which can limit the penetration and action of antimicrobial compounds.

TABLE II
ANTIBACTERIAL ACTIVITY OF ELECTROSPUN NONWOVEN MATS AGAINST *Escherichia coli* ATCC 25922.

Sample	Mean (mm)	SD (mm)
PVA	0.000	0.00
PVA with AgNPs	5.473	1.19
PVA with neomycin	8.967	0.72
PVA with AgNPs and neomycin	15.640	1.03

Source: Authors.

The PVA/AgNPs/Neo formulation exhibited the largest inhibition halo against *E. coli*, with a mean value of 15.640 ± 1.03 mm (Table II and Fig. 4B). Compared with PVA/Neo, the combined formulation resulted in an increase of approximately 74.4% in halo diameter. This enhanced halo response was more

pronounced for *E. coli* than for *S. aureus*. However, the mechanism responsible for the larger halo cannot be distinguished by agar diffusion alone; proposed effects such as bacterial-envelope disruption and facilitated antibiotic access remain literature-based mechanistic interpretations rather than experimentally demonstrated interactions in the present system [13].

Overall, the results from Tables I and II demonstrate that the association of AgNPs and neomycin within the PVA electrospun matrix improved antibacterial activity compared with the individual components. The enhanced response observed for the combined formulation may be associated with complementary antimicrobial mechanisms: AgNPs may affect membrane integrity and cellular functions, while neomycin contributes antibiotic activity against susceptible bacteria [5], [12], [13]. These findings support the potential of PVA/AgNPs/Neo nonwoven mats as antimicrobial materials for wound dressing applications, especially considering that electrospun nanofibrous systems can combine structural support with local delivery of active agents [3], [9], [10].

IV. CONCLUSION

Electrospun PVA-based nonwoven mats associated with AgNPs and neomycin sulfate were successfully produced under the selected processing conditions. SEM analysis showed that the AgNPs presented predominantly spherical morphology with nanometric dimensions, while the electrospun mats maintained a continuous fibrous structure after incorporation of the antimicrobial agents. EDS analysis confirmed the presence of silver in the PVA/AgNPs/Neo formulation, supporting the incorporation of the metallic component into the polymeric matrix.

The antimicrobial assay demonstrated that pure PVA did not inhibit bacterial growth, confirming the absence of intrinsic antibacterial activity of the polymeric matrix. In contrast, the incorporation of AgNPs or neomycin individually resulted in inhibition halos against both *S. aureus* and *E. coli*. The combined PVA/AgNPs/Neo formulation produced the largest inhibition halos for both bacterial strains, reaching 20.528 ± 0.24 mm for *S. aureus* and 15.640 ± 1.03 mm for *E. coli*. These results show that the combined formulation produced larger inhibition halos than either isolated agent under the tested conditions.

Therefore, the developed PVA/AgNPs/Neo nonwoven mat provides a basis for further investigation as an antimicrobial wound dressing material. Further studies involving release kinetics, MIC or viable-count reduction, cytocompatibility, mechanical behavior, swelling, degradation, sterilization, resistant clinical isolates, and wound-simulating conditions are necessary to confirm its safety and applicability as an advanced wound dressing.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – 88887.270867/2026-00.

REFERENCES

[1] G. M. Gomes et al., "Complications faced by diabetic patients in the healing process: an integrative literature review," Research, Society and Development, vol. 10, no. 14, e349101419993, 2021, doi: 10.33448/rsd-v10i14.19993.

[2] R. Dong and B. Guo, "Smart wound dressings for wound healing," Nano Today, vol. 41, p. 101290, 2021, doi: 10.1016/j.nantod.2021.101290.

[3] Y. Dong et al., "Electrospun nanofibrous materials for wound healing applications," Progress in Polymer Science, vol. 61, pp. 67-105, 2017, doi: 10.1016/j.progpolymsci.2016.11.001.

[4] World Health Organization, "Antibiotics most responsible for drug resistance are overused - WHO report," Apr. 2025.

[5] G. Franci et al., "Silver nanoparticles as potential antibacterial agents," Molecules, vol. 20, no. 5, pp. 8856-8874, 2015, doi: 10.3390/molecules20058856.

[6] D. Li and Y. Xia, "Electrospinning of nanofibers: reinventing the wheel?" Advanced Materials, vol. 16, no. 14, pp. 1151-1170, 2004, doi: 10.1002/adma.200400719.

[7] Z.-M. Huang, Y.-Z. Zhang, M. Kotaki, and S. Ramakrishna, "A review on polymer nanofibers by electrospinning and their applications in nanocomposites," Composites Science and Technology, vol. 63, no. 15, pp. 2223-2253, 2003, doi: 10.1016/S0266-3538(03)00178-7.

[8] G. Taylor, "Disintegration of water drops in an electric field," Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, vol. 280, no. 1382, pp. 383-397, 1964, doi: 10.1098/rspa.1964.0151.

[9] N. Charernsriwilaiwat, T. Rojanarata, T. Ngawhirunpat, and P. Opanasopit, "Electrospun chitosan/polyvinyl alcohol nanofibre mats for wound healing," International Wound Journal, vol. 11, pp. 215-222, 2014, doi: 10.1111/j.1742-481X.2012.01077.x.

[10] S. Lin et al., "Facile and green fabrication of electrospun poly(vinyl alcohol) nanofibrous mats doped with narrowly dispersed silver nanoparticles," International Journal of Nanomedicine, vol. 9, pp. 3937-3947, 2014, doi: 10.2147/IJN.S64985.

[11] S. Saito et al., "Modification of antibacterial copolymers on the surface of PVA-based microfibers via thermal cross-linking and their antibacterial properties," ACS Omega, vol. 9, no. 46, pp. 45961-45969, 2024, doi: 10.1021/acsomega.4c05637.

[12] X.-F. Zhang, Z.-G. Liu, W. Shen, and S. Gurunathan, "Silver nanoparticles: synthesis, characterization, properties, applications, and therapeutic approaches," International Journal of Molecular Sciences, vol. 14, no. 10, pp. 20715-20757, 2013, doi: 10.3390/ijms141020715.

[13] D. S. Ipe, P. T. S. Kumar, R. M. Love, and S. M. Hamlet, "Silver nanoparticles at biocompatible dosage synergistically increases bacterial susceptibility to antibiotics," Frontiers in Microbiology, vol. 11, p. 1074, 2020, doi: 10.3389/fmicb.2020.01074.

[14] D. E. Newbury and N. W. M. Ritchie, "Is scanning electron microscopy/energy dispersive X-ray spectrometry (SEM/EDS) quantitative?" Scanning, vol. 35, no. 3, pp. 141-168, 2013, doi: 10.1002/sca.21041.

Graphene Oxide-Silver Nanocomposite (GO-PEI-Ag): In Vitro Cytotoxicity Associated with Radiofrequency Ablation

1st Gabriella Gomes dos Santos
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
0009-0002-0169-6874

2nd Ronni Geraldo Gomes de Amorim
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
0000-0001-6532-3087

3rd Matheus Sevilha da Silva
Department of Cell Biology
Institute of Biological Sciences, University of Brasília
Brasília, Brazil
ORCID: 0009-0001-7419-5985

4th Melissa Silva Monteiro de Mendonça
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
ORCID: 0000-0001-6146-1482

5th Sonia Nair Bao
Department of Cell Biology
Institute of Biological Sciences, University of Brasília
Brasília, Brazil
0000-0002-9873-3098

6th Leonardo G. Paterno
Polymer and Nanomaterials Research Laboratory
University of Brasília
Brasília, Brazil
0000-0001-7716-0318

7th Saulo L. P. da Silva
Polymer and Nanomaterials Research Laboratory
Institute of Chemistry, University of Brasília
Brasília, Brazil
ORCID: 0000-0001-7280-1286

8th Janara Hellen Marques da Silva
Graduate Program in Biomedical Engineering
University of Brasília
Brasília, Brazil
0009-0003-3862-0184

***Abstract*—Radiofrequency ablation (RFA) is a minimally invasive technique widely used for the treatment of solid tumors. However, its efficacy is limited in large tumors and in lesions located near blood vessels because of heat dissipation and increased tissue impedance. Graphene oxide (GO)-based nanocomposites functionalized with polyethyleneimine (PEI) and silver nanoparticles (AgNPs) have emerged as promising multifunctional platforms for cancer therapy. In this study, the GO-PEI-Ag nanocomposite was synthesized and evaluated for its cytotoxic effects, alone and in combination with RFA, using Walker 256 tumor cells in vitro. The nanocomposite exhibited a dose-dependent cytotoxicity. However, the combined treatment with RFA did not produce a synergistic reduction in cell viability compared with RFA alone. These findings differ from previous reports involving gold nanoparticle-based systems, in which enhanced RFA efficacy has been associated with the higher electrical and thermal conductivity of gold. Overall, GO-PEI-Ag demonstrated intrinsic antitumor activity and may represent a promising platform for future investigations combining nanotechnology with non-ablative therapeutic approaches.**

***Index Terms*—Cytotoxicity, nanocomposite, cancer, ablation.**

This study was financed in part by the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil – Finance Code 001. The first author was supported by a Master's scholarship from the National Council for Scientific and Technological Development (CNPq), Brazil (Grant No. 133832/2025-2)

I. Introduction

According to GLOBOCAN, about 20 million new cancer cases and 9.7 million cancer-related deaths were reported worldwide in 2022 [1]. This annual incidence is expected to surpass 35 million new cases by 2050, a growth driven by population aging, demographic expansion, and continued exposure to factors associated with carcinogenesis [2].

Among minimally invasive therapeutic modalities, radiofrequency ablation (RFA) has established itself as an effective alternative for treating various solid tumors, particularly in patients ineligible for surgical resection. The technique relies on the application of high-frequency alternating current through an electrode inserted into the tumor lesion, generating tissue heating and coagulative necrosis at temperatures above approximately 60 °C [3].

Nevertheless, its efficacy remains limited in larger tumors ($\geq$ 3 cm) or those located near large blood vessels, owing to

the so-called *heat sink effect*, which dissipates the heat generated during the procedure. Compounding this limitation, the rise in electrical impedance caused by tissue carbonization and necrosis reduces current conduction and restricts the expansion of the ablation zone [4]–[8]. As a result, viable tumor cells may persist after treatment, favoring local recurrence.

Given this scenario, nanomedicines have gained prominence for enabling greater biocompatibility, selectivity, and therapeutic efficacy, even in the face of tumor microenvironment heterogeneity. This potential has driven the development of strategies based on nanoparticle surface engineering aimed at optimizing their interaction with tumor cells [9], [10]. Within this context, graphene oxide (GO) nanoplatforms functionalized with polyethyleneimine (PEI) stand out: GO presents a high surface area and broad potential for chemical functionalization, features that make it a versatile platform for the transport of bioactive molecules and metallic nanoparticles [11]. Functionalization with PEI, in turn, contributes to the colloidal stability of the nanocomposite, promotes electrostatic interaction with the plasma membrane, and enhances its internalization by tumor cells [12]–[14].

The incorporation of silver nanoparticles (AgNPs) into the GO-PEI platform results in a multifunctional nanocomposite with potential for oncological applications. Beyond their well-established antimicrobial activity, associated with silver ion release, generation of reactive oxygen species (ROS), destabilization of the cell membrane, and damage to microbial genetic material [**?**], [15], AgNPs have also been shown to induce oxidative stress, mitochondrial dysfunction, DNA damage, cell cycle arrest, and apoptosis in tumor cells, in addition to enhancing the efficacy of antineoplastic agents [17], [18].

These properties have also motivated their investigation as sensitizing agents in energy-based therapies, including modalities such as photothermal therapy and photodynamic therapy, given their capacity to intensify intracellular oxidative stress and increase the susceptibility of tumor cells to treatment-induced damage [17]–[19].

In light of the above, the combination of the GO-PEI-Ag nanocomposite with radiofrequency represents a strategy with the potential to enhance the cytotoxic effects of thermal therapy, promoting greater cell destruction and, consequently, increasing antitumor efficacy. Thus, the objective of this study was to investigate the potential synergistic effect between GO-PEI-functionalized silver nanoparticles and radiofrequency in a tumor cell model, evaluating their influence on cell viability and the mechanisms associated with the selective antitumor response.

II. Materials and Methods

A. Synthesis of the GO-PEI-Ag Nanocomposite

Silver nanoparticles (AgNPs) were synthesized via chemical reduction in the presence of GO-PEI, yielding a hybrid material with controlled properties and marked biological potential. GO acts as a scaffold, preventing AgNP aggregation, while PEI increases the solubility and functionality of the system, contributing to improved cellular interaction [20].

B. B.In Vitro Experimental Procedures

1) Cell line used:: For the biological evaluation, the Walker 256 tumor cell line (LLC-WRC-256) was used. Cells were cultured in Roswell Park Memorial Institute (RPMI-1640) medium, supplemented with 10% inactivated fetal bovine serum (FBS) and 1% antibiotic solution containing penicillin and streptomycin at a concentration of 100 μg/mL each. Cultures were maintained in an incubator at 37 °C, under a humidified atmosphere with 90% humidity and 5% CO_2. For the biological assays, cells were seeded in culture plates.

2) Cell Viability Assays (Without Radiofrequency): Cell viability was assessed using the Alamar Blue colorimetric assay, with spectrophotometric readings taken in the 560–590 nm range. Cells were treated with ten different concentrations of the nanocomposite: 15.0; 7.5; 5.0; 3.75; 1.87; 0.94; 0.47; 0.23; 0.12; 0.06; and 0.03 μg/mL. Treatments were evaluated after 24 and 48-hour period. All experimental conditions were tested in triplicate, across three independent experiments.

3) Combined Therapy Assays with Radiofrequency (RF): To evaluate the response to combination therapy, cells previously treated with the nanocomposite were subjected to radiofrequency (RF) using the SOFIA® generator, protected under Brazilian patent BR 10 2017 002919 0 and affiliated with the University of Brasília (UnB). Sterile syringe needles served as electrodes, and RF was applied for 3 minutes. The experimental groups consisted of: (1) control; (2) combined treatment with GO-PEI-Ag (1.87 μg/mL) + RF; (3) combined treatment with GO-PEI-Ag (3.75 μg/mL) + RF; and (4) RF alone.

Data distribution was assessed using the Shapiro–Wilk test. For normally distributed data, differences among groups were evaluated by one-way analysis of variance (ANOVA), followed by Tukey's multiple comparisons test. When the normality assumption was not met, the Kruskal–Wallis test was applied, followed by Dunn's post hoc test with Bonferroni adjustment for multiple comparisons. Effect sizes for pairwise comparisons were estimated using Cliff's delta (δ) and reported with 95% confidence intervals (95% CIs). Statistical significance was set at $p < 0.05$ in all cases. Statistical analyses and data visualization were performed using RStudio, version 2025.05.1+513.

III. Results and Discussion

1) Cell Viability Assays (Without Radiofrequency): Cell viability assays revealed reduced viability in the untreated control wells located near those containing the highest concentrations of the GO-PEI-Ag nanocomposite. This spatial pattern suggests interference between adjacent wells, possibly resulting from the formation of microaerosols during pipetting, droplet transport via condensation on the plate lid, and factors inherent to 96-well plate culture, such as the *edge effect*, which promotes evaporation of the medium and alters culture conditions in peripheral wells [21], [22].

To investigate this possible interference, three different experimental plate designs were evaluated (Fig. 2, 3, 4, in which the spatial arrangement of nanocomposite concentrations was

modified. Fig. 1 shows the legend for the nanocomposite concentrations.

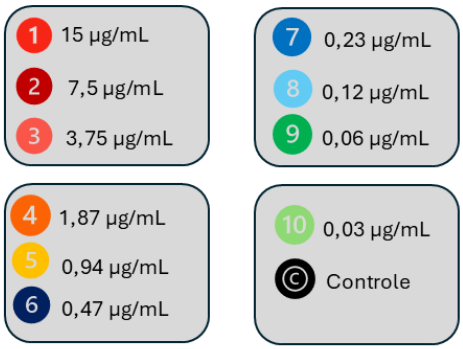


Fig. 1. Concentration legend for the GO-PEI-Ag nanocomposite used in the cell viability assays. Concentrations are represented by a color scale, ranging from warmer, more vibrant tones for higher concentrations to cooler, pastel tones for lower concentrations. Based on the cytotoxicity observed across the experimental designs, concentrations were classified as high (15.0, 7.5, and 3.75 μg/mL), intermediate (1.87 and 0.94 μg/mL), and low (0.47, 0.23, 0.12, 0.06, and 0.03 μg/mL), corresponding respectively to marked, partial, and negligible reductions in cell viability. Numerical labels correspond to: 1 (15.0 μg/mL), 2 (7.5 μg/mL), 3 (3.75 μg/mL), 4 (1.87 μg/mL), 5 (0.94 μg/mL), 6 (0.47 μg/mL), 7 (0.23 μg/mL), 8 (0.12 μg/mL), 9 (0.06 μg/mL), and 10 (0.03 μg/mL); C denotes the untreated control group.



Fig. 2. Plate design No. 1 with a legend of the concentrations used. The nanocomposite concentrations were distributed in alternating rows, with the highest concentrations (15,0–3,75 μg/mL) positioned near the control wells, followed by the intermediate concentrations (1,87–0,47 μg/mL) and the lowest concentrations (0,23–0,06 μg/mL).

Across all experimental designs, the concentrations of 15.0, 7.5, and 3.75 μg/mL resulted in a marked reduction in cell viability, reaching values equal to zero (Fig. 5, Fig. 6, and Fig. 7). However, at concentrations of 1.87 and 0.94 μg/mL, low cell viability was observed in Designs 1 and 2 (Fig. 5 and Fig. 6), whereas in Design 3 (Fig. 7), where these groups were positioned farther from the highest concentrations, high cell viability was observed. Notably, this same pattern was observed in the untreated control wells located near the highest-concentration wells, reinforcing that the reduced viability is not attributable to the nanocomposite's intrinsic cytotoxicity but rather to the spatial arrangement of the wells on the plate.

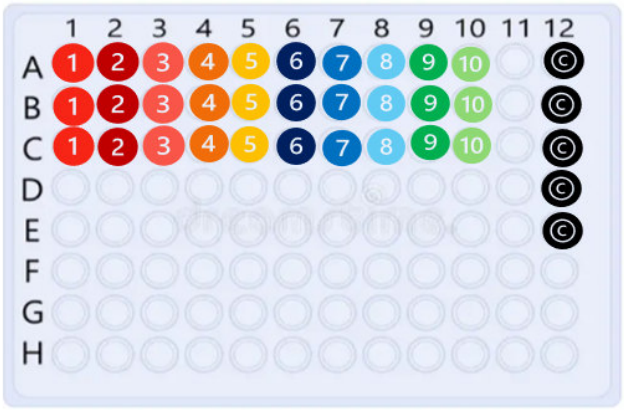


Fig. 3. Plate layout No. 2 with legend for the concentrations used: The concentrations were arranged in sequential rows, starting from the highest concentration to the lowest, with the control placed away from the highest concentrations to reduce their interference with the untreated group.

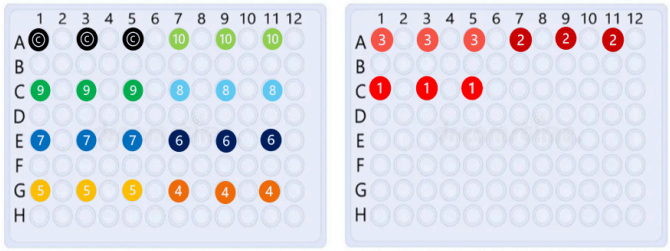


Fig. 4. Plate layout No. 3 with a legend of the concentrations used. The same experimental design was carried out on two culture plates, since the highest concentrations (15.0, 7.5, and 3.75 μg/mL), relative to the other concentrations (1.87–0.06 μg/mL), made it impossible to distribute all experimental concentrations on a single plate. A uniform spacing was maintained between adjacent wells, both laterally and between rows

This behavior is consistent with interference between adjacent wells, likely arising from the dispersion of microaerosols and vapors during pipetting and from droplet transport via condensation on the plate lid. Aerosols originating from wells with higher nanocomposite concentrations carry a proportionally greater cytotoxic load than those from lower-concentration wells; as they disperse across the plate, they preferentially affect the immediately neighboring wells, exposing them to an additional, non-intrinsic source of toxicity beyond their nominal concentration [23]–[25]. This effect is compounded by factors inherent to 96-well plate culture, such as the *edge effect*, which promotes evaporation of the medium and alters culture conditions in peripheral wells [21], [22].

2) Combined Therapy Assays with Radiofrequency (RF): In the experiments combining the GO-PEI-Ag nanocomposite

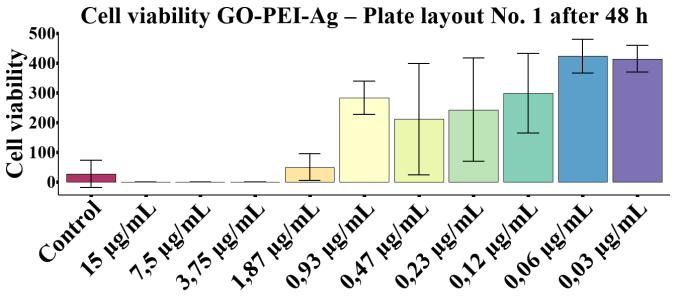


Fig. 5. The cell viability results for plate design No. 1 showed a concentration-dependent cytotoxic effect. The highest concentrations (15, 7.5, and 3.75 μg/mL) resulted in cell viability close to zero. At intermediate concentrations (1.87 to 0.47 μg/mL), a gradual increase in cell viability was observed, whereas the lowest concentrations (0.23 to 0.03 μg/mL) exhibited high cell viability. It is important to note that the untreated control group (C) showed low viability due to interference associated with the plate design, as discussed above. Therefore, given the low viability observed in the control group, the data were not normalized and are presented as absolute fluorescence values, representing cellular activity.

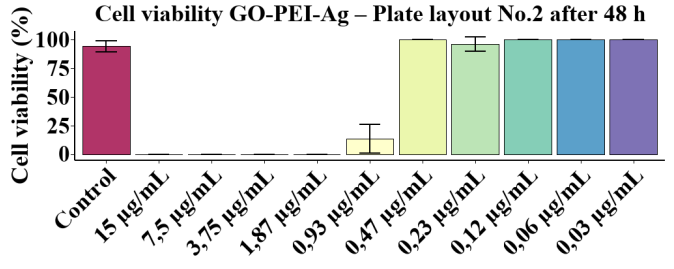


Fig. 6. The results from plate design No. 2 demonstrate that higher concentrations of the nanocomposite maintained high cytotoxicity, with cell viability close to zero. Starting at a concentration of 1.87 μg/mL, a significant increase in cell viability was observed, which remained high at subsequent concentrations, approaching the values observed in the control group, which exhibited maximum viability.Cell viability is expressed relative to the untreated control, which was set as 100%.

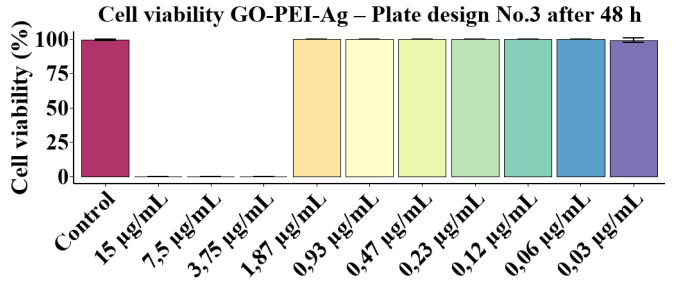


Fig. 7. Viability data based on plate design No. 3 reveal that the highest concentrations of the nanocomposite exhibited high cytotoxicity, with cell viability close to zero, while intermediate and low concentrations maintained high viability, with values similar to those observed in the control group, which exhibited maximum viability.Cell viability is expressed relative to the untreated control, which was set as 100%.

with radiofrequency (RF) ablation, two concentrations were selected: 3.75 and 1.87 μg/mL. The first was chosen for comparison with the study by [25], in which the same assay was performed using a gold-based nanomaterial (GO-PEI-Au) instead of silver. The second, 1.87 μg/mL, was selected as an intermediate dose expected to be sublethal, given the high cell viability observed at this concentration in Design 3 (Fig. 7). This choice was deliberate: the aim of the combined therapy was to potentiate the effect of radiofrequency ablation, rather than to induce tumor cell death through the cytotoxic just the action of the nanomaterial.

The Shapiro-Wilk test indicated that the assumption of normality was not met across the experimental groups at 24 and 48 h. Therefore, differences in cell viability among the experimental groups were assessed using the Kruskal-Wallis test, followed by Dunn's post hoc test with Bonferroni adjustment.

At 24 h, Dunn's post hoc test showed significantly lower viability in the 1.87 μg/mL + RF, 3.75 μg/mL + RF, and RF-only groups compared with the untreated control (adjusted $p < 0.001$ for all comparisons), with very large effect sizes (Cliff's $\delta = 1.000$; 95% CIs: 0.994–1.000). No significant differences, however, were observed among the treatment groups themselves (adjusted $p = 1.000$), with Cliff's δ ranging from -0.247 to 0.235 and all corresponding 95% CIs including zero (Tab. I).

At 48 h, a similar pattern was observed: viability was significantly lower in the 1.87 μg/mL + RF and 3.75 μg/mL + RF groups compared with the untreated control (adjusted $p < 0.001$; Cliff's $\delta = 1.000$; 95% CI: 0.986–1.000), as well as in the RF-only group (adjusted $p = 0.003$; Cliff's $\delta = 0.944$; 95% CI: 0.688–0.991). Again, no significant differences were detected between the treatment groups (adjusted $p = 1.000$), with effect sizes ranging from -0.206 to -0.095 and 95% CIs that included zero (Tab. II).

Taken together, these findings indicate that the addition of the GO-PEI-Ag nanocomposite to RF did not result in a statistically significant increase or decrease in cell viability compared with RF alone at either 24 or 48 h (Fig. 8). No statistically significant differences in cell viability were observed at either time point when combined treatment was compared with RF alone, suggesting that the intrinsic characteristics of the silver nanoparticles may not favor synergistic energy transfer with GO, in contrast to the synergistic effect observed by the research group with the GO-PEI-Au nanocomposite [25].

TABLE I
STATISTICAL ANALYSIS OF CELL VIABILITY TREATED WITH COMBINATION THERAPY AFTER 24 H

Comparison	**Dunn adjusted p**	**Cliff's δ (95% CI)**
Untreated (24 h) vs. 1.87 μg/mL + RF	< 0.001	1.000 (0.994–1.000)
Untreated (24 h) vs. 3.75 μg/mL + RF	< 0.001	1.000 (0.994–1.000)
Untreated (24 h) vs. RF	< 0.001	1.000 (0.994–1.000)
1.87 μg/mL + RF vs. 3.75 μg/mL + RF	1.000	-0.247 (-0.575–0.149)
1.87 μg/mL + RF vs. RF	1.000	-0.037 (-0.382–0.317)
3.75 μg/mL + RF vs. RF	1.000	0.235 (-0.136–0.547)

TABLE II
STATISTICAL ANALYSIS OF CELL VIABILITY TREATED WITH COMBINATION THERAPY AFTER 48 H

Comparison	**Dunn adjusted p**	**Cliff's δ (95% CI)**
Untreated (48 h) vs. 1.87 μg/mL + RF	< 0.001	1.000 (0.986–1.000)
Untreated (48 h) vs. 3.75 μg/mL + RF	< 0.001	1.000 (0.986–1.000)
Untreated (48 h) vs. RF	0.003	0.944 (0.688–0.991)
1.87 μg/mL + RF vs. 3.75 μg/mL + RF	1.000	-0.128 (-0.520–0.309)
1.87 μg/mL + RF vs. RF	1.000	-0.206 (-0.602–0.270)
3.75 μg/mL + RF vs. RF	1.000	-0.095 (-0.532–0.381)

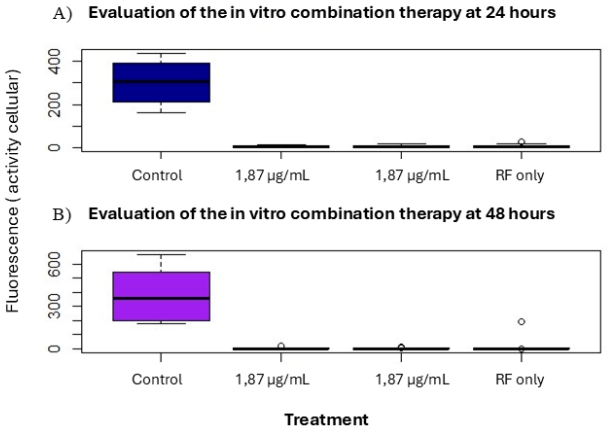


Fig. 8. Cell viability after combination therapy with the GO-PEI-Ag nanocomposite and radiofrequency (RF) ablation at 24 h (A) and 48 h (B). Box plots represent cellular activity measured by fluorescence in the untreated control, GO-PEI-Ag + RF (1.87 and 3.75 μg/mL), and RF-only groups. At both time points, a reduction in cell viability was observed in the treatment groups compared with the untreated control, while no significant differences were detected among the treatment groups.

This lack of an additional effect may suggest that the incorporation of silver nanoparticles into the nanocomposite did not enhance the interaction with RF energy under the conditions evaluated. This finding differs from previous observations by our research group using the GO-PEI-Au nanocomposite, in which the combination with RF produced an enhanced antitumor effect [25].

IV. CONCLUSION

The results demonstrated that GO-PEI-Ag possesses significant intrinsic antitumor activity, inducing a dose-dependent reduction in cell viability, particularly at concentrations of 15, 7.5, and 3.75 μg/mL. However, when evaluated as a potential sensitizing agent for radiofrequency ablation (RFA), the combined treatment with GO-PEI-Ag did not result in a synergistic reduction in cell viability compared to RFA alone, as also reported by Monteiro et al. (2024) [25].

This lower efficacy, when compared to systems based on gold nanoparticles (AuNPs), is likely related to differences in the physicochemical properties of the two nanomaterials. AuNPs can act as conductive dopants, reducing tissue impedance and favoring current conduction and the deposition of radiofrequency energy, which in turn may expand the ablation zone [25]–[27]. The heating observed under RF fields has also been linked to the electrophoretic oscillation of surface charges and the Joule effect, both of which favor greater energy conversion [28]. AgNPs, on the other hand, despite their well-recognized cytotoxic activity, seem to have less impact on local impedance modulation. The combination of this weaker interaction with the RF field and the already high intrinsic cytotoxicity of AgNPs may explain the limited additional benefit brought by radiofrequency, and consequently the lower synergy observed in this study.

Taken together, these findings point to the silver nanocomposite as a candidate for future use either as a standalone antitumor agent or in combination with non-ablative therapeutic strategies that can take advantage of its cytotoxic profile.

V. DECLARATION ON THE USE OF ARTIFICIAL INTELLIGENCE

During the preparation of this manuscript, the generative artificial intelligence tools ChatGPT (OpenAI) and Gemini (Google) were used to support text revision, improve writing quality, organize and connect ideas, and standardize bibliographic references. DeepL was used to assist with the translation of the manuscript and the adaptation of the text to scientific English. None of these tools were used to generate, analyze, interpret, or manipulate scientific data, or to formulate the study's conclusions. All content produced with the assistance of these tools was critically reviewed by the authors, who verified its accuracy, made all necessary revisions, and assume full responsibility for the content and the final version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the advisor, Dr. Ronni Geraldo Amorim, for his guidance and support throughout the development of this work. The authors also thank Dr. Melissa Silva Monteiro for project coordination and leadership; Prof. Dr. Leonardo Giordano Paterno for his support in the production and physicochemical characterization of the nanocomposites; Dr. Sônia Nair Bao for conducting the biological assays at the Microscopy and Microanalysis Laboratory of the University of Brasília (UnB); as well as all participants who contributed to the development of this work.

The authors also acknowledge the funding agencies—the National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Federal District Research Support Foundation (FAPDF), through the StartBSB Call for Proposals, and the Agency for the Financing of Studies and Projects (FINEP)—for their financial and institutional support, which was essential for the completion of this research.

REFERENCES

[1] F. Bray et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 74, no. 3, pp. 229–263, 2024.

[2] International Agency for Research on Cancer, "Global cancer burden growing, amidst mounting need for services," Lyon, France: IARC, 2024. [Online]. Available: https://www.iarc.who.int/news-events/global-cancer-burden-growing-amidst-mounting-need-for-services/

[3] J. Wu, Z. Zhou, and Y. Huang et al., "Radiofrequency ablation: mechanisms and clinical applications," MedComm, vol. 5, Art. no. e746, 2024, doi: 10.1002/mco2.746.

[4] B. M. Künzli, P. Abitabile, and C. A. Maurer, "Radiofrequency ablation of liver tumors: actual limitations and potential solutions in the future," World J. Hepatol., vol. 3, no. 1, pp. 8–14, 2011, doi: 10.4254/wjh.v3.i1.8.

[5] M. Ahmed et al., "Image-guided tumor ablation: standardization of terminology and reporting criteria—a 10-year update," Radiology, vol. 273, no. 1, pp. 241–260, 2014, doi: 10.1148/radiol.14132958.

[6] R. Levigard, J. Mendes Luz, and M. Nita, "Ablação percutânea de metástases: revisão da literatura e análise de custo no cenário brasileiro," Pulmão RJ, vol. 30, no. 1, pp. 53–61, 2021.

[7] H. Zheng et al., "Microwave ablation versus radiofrequency ablation for subcapsular hepatocellular carcinoma: a propensity score-matched study," Eur. Radiol., vol. 32, no. 7, pp. 4657–4666, 2022.

[8] H. Wang et al., "Radiofrequency ablation of hepatocellular carcinoma: current status, challenges, and prospects," Liver Res., vol. 7, no. 2, pp. 108–115, 2023.

[9] M. Ikeda-Imafuku, L. L. Wang, D. Rodrigues, S. Shaha, Z. Zhao, and S. Mitragotri, "Strategies to improve the EPR effect: a mechanistic perspective and clinical translation," J. Controlled Release, vol. 345, pp. 512–536, 2022, doi: 10.1016/j.jconrel.2022.03.043.

[10] B. Lahooti et al., "Targeting endothelial permeability in the EPR effect," J. Controlled Release, vol. 361, pp. 212–235, 2023, doi: 10.1016/j.jconrel.2023.07.039.

[11] H. Sharma and S. Mondal, "Functionalized graphene oxide for chemotherapeutic drug delivery and cancer treatment: a promising material in nanomedicine," Int. J. Mol. Sci., vol. 21, no. 17, Art. no. 6280, 2020, doi: 10.3390/ijms21176280.

[12] P. S. Teo, H. N. Lim, N. M. Huang, C. H. Chia, and I. Harrison, "Room temperature in situ chemical synthesis of Fe_3O_4/graphene," Ceram. Int., vol. 38, no. 8, pp. 6411–6416, 2012, doi: 10.1016/j.ceramint.2012.05.014.

[13] A. P. Pandey and K. K. Sawant, "Polyethylenimine: a versatile, multifunctional non-viral vector for nucleic acid delivery," Mater. Sci. Eng. C, vol. 68, pp. 904–918, 2016, doi: 10.1016/j.msec.2016.07.066.

[14] H. L. Jiang, M. A. Islam, L. Xing, J. Firdous, W. Cao, Y. J. He, Y. Zhu, K. H. Cho, H. S. Li, and C. S. Cho, "Degradable polyethylenimine-based gene carriers for cancer therapy," Topics Curr. Chem., vol. 375, no. 2, Art. no. 34, 2017, doi: 10.1007/s41061-017-0124-9.

[15] A. S. Rodrigues et al., "Advances in silver nanoparticles: a comprehensive review on their potential as antimicrobial agents and their mechanisms of action elucidated by proteomics," Front. Microbiol., vol. 15, Art. no. 1440065, 2024, doi: 10.3389/fmicb.2024.1440065.

[16] S. K. Mondal, S. Chakraborty, S. Manna, and S. M. Mandal, "Antimicrobial nanoparticles: current landscape and future challenges," RSC Pharm., vol. 1, no. 3, pp. 388–402, 2024, doi: 10.1039/D4PM00032C.

[17] P. Takáč et al., "The role of silver nanoparticles in the diagnosis and treatment of cancer: are there any perspectives for the future?," Life, vol. 13, no. 2, Art. no. 466, 2023, doi: 10.3390/life13020466.

[18] H. Jangid, S. Singh, P. Kashyap, A. Singh, and G. Kumar, "Advancing biomedical applications: an in-depth analysis of silver nanoparticles in antimicrobial, anticancer, and wound healing roles," Front. Pharmacol., vol. 15, Art. no. 1438227, 2024, doi: 10.3389/fphar.2024.1438227.

[19] C. H. Rodrigues et al., "Silver nanoprisms as plasmonic enhancers applied in the photodynamic inactivation of *Staphylococcus aureus* isolated from bubaline mastitis," Photodiagnosis Photodyn. Ther., vol. 34, Art. no. 102315, 2021.

[20] S. L. P. Silva, "Síntese fotoquímica, caracterização e aplicação em sensores de nanocompósitos de óxido de grafeno e nanopartículas de prata," M.S. thesis, Inst. Química, Univ. Brasília, Brasília, Brazil, 2022.

[21] B. K. Lundholt, K. M. Scudder, and L. Pagliaro, "A simple technique for reducing edge effect in cell-based assays," J. Biomol. Screen., vol. 8, no. 5, pp. 566–570, Oct. 2003, doi: 10.1177/1087057103256465.

[22] M. Mansoury, M. Hamed, R. Karmustaji, F. Al Hannan, and S. T. Safrany, "The edge effect: a global problem. The trouble with culturing cells in 96-well plates," Biochem. Biophys. Rep., vol. 26, Art. no. 100987, 2021, doi: 10.1016/j.bbrep.2021.100987.

[23] M. Houdkova, G. Albarico, I. Doskocil, J. Tauchen, K. Urbanova, E. E. Tulin, and L. Kokoska, "Vapors of volatile plant-derived products significantly affect the results of antimicrobial, antioxidative and cytotoxicity microplate-based assays," Molecules, vol. 25, no. 24, p. 6004, 2020, doi: 10.3390/molecules25246004.

[24] S. L. Avlasevich, E. Briggs, K. Tichenor, A. Conrad, N. Hall, S. M. Bryce, J. C. Bemis, and S. D. Dertinger, "Mitigation of volatile chemicals' effect on adjacent microtiter plate wells," Environ. Mol. Mutagen., vol. 66, no. 3, pp. 134–141, Mar. 2025, doi: 10.1002/em.70005.

[25] M. S. Monteiro et al., "Radiofrequency driving antitumor effect of graphene oxide-based nanocomposites: a Hill model analysis," Nanomedicine, vol. 19, no. 5, pp. 397–412, 2024, doi: 10.2217/nnm-2023-0312.

[26] D. L. Castro-López, E. Berjano, and R. Romero-Mendez, "Radiofrequency ablation combined with conductive fluid-based dopants (saline normal and colloidal gold): computer modeling and ex vivo experiments," Biomed. Eng. Online, vol. 20, no. 1, Art. no. 4, 2021, doi: 10.1186/s12938-020-00842-8.

[27] M. O. Estevari, M. Zabihzadeh, N. Riahi-Alam, and M. Momen-Gharibvand, "Measurement of electrical properties of tissues with and without gold nanoparticles (GNPs) at different temperatures," J. Nanostructures, vol. 13, no. 2, pp. 442–452, 2023, doi: 10.22052/JNS.2023.02.014.

[28] S. J. Corr, M. Raoof, Y. Mackeyev, S. Phounsavath, M. A. Cheney, B. T. Cisneros, M. Shur, M. Gozin, and P. J. McNally, "Citrate-capped gold nanoparticle electrophoretic heat production in response to a time-varying radio-frequency electric field," J. Phys. Chem. C, vol. 116, no. 45, pp. 24380–24389, 2012, doi: 10.1021/jp309053z.

Green Synthesis of Cerium Oxide Nanoparticles for Crystal Violet Removal with Potential Application in Healthcare Wastewater Treatment

Lourdes Maria Muraro Favarin*, Daniel Moro Druzian*, Leandro Rodrigues Oviedo*, Vinícius Rodrigues Oviedo*, Marcos Vinicius Pasqualoto Prior*, Luiz Batista Cardoso*, Yolice Patricia Moreno Ruiz†, Giovani Pavoski‡, Jorge Alberto Soares Tenório‡, Denise Crocce Romano Espinosa‡, William Leonardo da Silva*, Maria Amélia Zazycki*

* *Applied Nanomaterials Research Group (GPNAp)*
Franciscan University (UFN)
Santa Maria, RS, Brazil
lourdes.favarin@ufn.edu.br; leandro.oviedo@ufn.edu.br; luiz.bcardoso@ufn.edu.br; vinicius.oviedo@ufn.edu.br; marcos.prior@ufn.edu.br; daniel.moro.druzian@gmail.com; w.silva@ufn.edu.br; zazycki@ufn.edu.br

† *Strategic Technologies Center of Northeast (CETENE)*
Federal University of Pernambuco (UFPE)
Recife, PE, Brazil
ypmoreno@gmail.com

‡ *Polytechnical School of Chemical Engineering (EPEQ)*
University of São Paulo (USP)
São Paulo, SP, Brazil
giovanipavoski@gmail.com; jtenorio@usp.br; espinosa@usp.br

***Abstract*—The contamination of aquatic environments by healthcare-related pollutants represents a challenge. This study investigated the green synthesis of cerium oxide nanoparticles (CeO_2-NPs) using *Ilex paraguariensis* leaf extract and their performance to adsorb Crystal Violet (CV), a model cationic pollutant. The nanoparticles were characterized by X-ray diffraction, FEG-SEM/EDS, N_2 physisorption, and zeta potential, confirming crystalline mesoporous CeO_2-NPs with a surface area of 85 $\pm$ 1.6 m^2 g^{-1}, pore diameter of 7.8 $\pm$ 1.5 nm, and zeta potential of -10.9 $\pm$ 2.4 mV. Batch adsorption assays at 0.5 g L^{-1} and pH 9 showed concentration-dependent CV removal, reaching 95.6% at 2 mg L^{-1} and decreasing to 51.4% and 36.4% at 5 and 15 mg L^{-1}, respectively, after 180 min. The Sips model provided the best fit (R^2 = 0.999), indicating heterogeneous adsorption and supporting the material's potential as a nanoadsorbent for cationic pollutants, pending validation with healthcare effluents.**

***Index Terms*—adsorption, biomedical engineering, hospital effluents, nanotechnology**

I. Introduction

The management and treatment of wastewater generated by hospitals and clinical laboratories have become an emerging challenge at the interface between biomedical engineering, public health, and environmental sustainability [1]. Unlike municipal wastewater, healthcare effluents contain a complex mixture of persistent contaminants, including pharmaceuticals, antiseptics, and histological dyes (e.g., Crystal Violet, CV), which are frequently resistant to conventional treatment processes due to their complex molecular structures, high chemical stability, low biodegradability, and resistance to microbial degradation [2]. Their continuous release into aquatic environments has been associated with mutagenicity, carcinogenicity, endocrine disruption, and the dissemination of antibiotic-resistant microorganisms, representing a growing risk to both environmental and human health [3].

Among the available treatment technologies, adsorption has attracted considerable attention as an effective approach for removing persistent organic pollutants from aqueous systems due to its operational simplicity, versatility, and potential for pollutant recovery [4]. Cerium oxide nanoparticles (CeO_2-NPs) have been investigated for this application due to their large specific surface area, chemical stability, and favorable surface properties, making them promising candidates for the treatment of wastewater [5]. Furthermore, their surface charge and physicochemical properties make CeO_2-NPs particularly suitable for the adsorptive removal of cationic pollutants, such as CV [6].

However, conventional synthesis routes for CeO_2-NPs commonly employ toxic reducing agents and stabilizers that compromise the environmental sustainability of the process [7].

Fundação de Amparo à pesquisa do Estado do RS (FAPERGS)

Green synthesis has emerged as an alternative approach by replacing hazardous chemicals with renewable plant extracts [8]. Among these, *Ilex paraguariensis* has attracted attention because of its high content of polyphenols and flavonoids, which act as natural reducing and stabilizing agents during nanoparticle formation and enable the production of functional nanomaterials through a simple, low-cost, and environmentally friendly route [9].

Despite growing interest in green-synthesized CeO_2-NPs for aqueous pollutant removal, previous studies have generally investigated adsorption under specific experimental conditions, focusing primarily on individual pollutants and selected kinetic and equilibrium models [10], [11]. Therefore, the present study systematically investigates the adsorption of CV over a broad range of initial concentrations (1–15 mg L^{-1}), combining kinetic modeling based on the pseudo-first-order (PFO), pseudo-second-order (PSO), and intraparticle diffusion (IPD) models with equilibrium analysis using the Langmuir, Freundlich, and Sips isotherms. This comprehensive approach enables a more detailed assessment of concentration-dependent adsorption behavior and provides further insights into the potential of plant-mediated CeO_2-NPs as sustainable nanoadsorbents for removing cationic pollutants [12].

Moreover, developing sustainable technologies for healthcare wastewater treatment directly contributes to Sustainable Development Goals (SDG) 3 (Good Health and Well-being) and SDG 6 (Clean Water and Sanitation) by reducing environmental risks associated with emerging healthcare contaminants [13]. In this sense, the present work aimed to develop and characterize green-synthesized CeO_2-NPs using *Ilex paraguariensis* leaf extract and evaluate their adsorption performance toward CV dye as a model cationic dye, with potential future applications in healthcare wastewater treatment.

II. Materials and Methods

A. Preparation of the Ilex paraguariensis extract

The aqueous extract of *Ilex paraguariensis* leaves was prepared by the infusion method [14]. Fresh leaves of *Ilex paraguariensis* were collected in Rio Grande do Sul, Brazil, thoroughly washed with distilled water to remove surface impurities, and dried at 25 $\pm$ 2 °C. The dried leaves were manually sliced and ground in a Wiley knife mill (TECNAL, model TE-650/1) for three consecutive cycles and subsequently sieved (#250) to standardize the particle size. The aqueous extract was prepared by adding 5 g of the powdered leaves to 50 mL of distilled water maintained at 60 $\pm$ 2 °C under magnetic stirring for 10 min. Subsequently, the suspension was vacuum-filtered to remove the plant residues, and the resulting extract was stored in an amber bottle under refrigeration at 4 $\pm$ 2 °C. The obtained extract, naturally rich in polyphenols and flavonoids, was employed as a renewable reducing and stabilizing agent during the green synthesis of CeO_2-NPs.

B. Biosynthesis of the CeO_2-NPs

CeO_2-NPs were synthesized through a green biosynthesis route employing the aqueous extract of *Ilex paraguariensis* as a natural reducing and stabilizing agent [15]. Initially, a precursor solution was prepared by dissolving 8.628 g of cerium(III) nitrate hexahydrate ($Ce(NO_3)_3{\cdot}6H_2O$, 99%, Sigma–Aldrich) in 100 mL of distilled water. Subsequently, a 50 mL aliquot of the precursor solution was slowly added to the previously prepared 50 mL of *Ilex paraguariensis* extract under magnetic stirring (300 rpm, 25 $\pm$ 2 °C) for 40 min, allowing the phytochemicals present in the extract to promote the reduction of cerium ions, nanoparticle nucleation, and stabilization of the growing particles.

Following the reaction, the suspension was allowed to decant, and the resulting precipitate was collected and dried (60 $\pm$ 2 °C for 12 hours) to remove residual moisture. Finally, the dried material was calcined at 450 $\pm$ 2 °C for 4 hours using a heating rate of (10 °C min^{-1}) to eliminate residual organic compounds from the extract and promote the crystallization of CeO_2-NPs.

C. Characterization of the CeO_2-NPs

The crystalline phases of the green-synthesized CeO_2-NPs were identified by X-ray Diffraction (XRD) using a Bruker D2 Advance diffractometer with Cu Kα radiation (λ = 0.15418 nm), operating over the 2θ range of 20–80° [16]. The morphology and elemental composition were analyzed by Field Emission Gun Scanning Electron Microscopy coupled with Energy-Dispersive X-ray Spectroscopy (FEG-SEM/EDS) using a TESCAN MIRA3 microscope equipped with an Oxford EDX system [17]. N_2 physisorption measurements were performed using an ASAP 2020 (Micromeritics) surface area analyzer at 77 K to determine the specific surface area (S_{BET}), pore volume (V_p), and average pore diameter (D_p), according to the BET method and IUPAC recommendations [18]. Surface charge was evaluated by zeta potential (ZP) measurements using a Zetasizer Nano ZS (Malvern Instruments) at room temperature (25 $\pm$ 2 °C) [19].

D. Adsorption experiments

Batch adsorption experiments were conducted to evaluate CV dye removal using green-synthesized CeO_2-NPs obtained from *Ilex paraguariensis* leaf extract. Aqueous CV solutions were prepared at four nominal initial concentrations (C_0 = 1, 2, 5, and 15 mg L^{-1}) at pH 9, while CeO_2-NPs were added at a dosage of 0.5 g L^{-1}. Under these conditions, the negatively charged surface of the nanoparticles, as characterized in Section III.A, was considered favorable for the electrostatic attraction of the cationic CV molecules [20].

Each suspension was magnetically stirred in the dark at 25 $\pm$ 2 °C, and aliquots were collected at 0, 15, 30, 45, 60, 75, 90, 105, 120, 150, and 180 min. The samples were subsequently centrifuged, and the residual CV concentration (C_t) was determined by UV–Vis spectrophotometry at 590 nm using a calibration curve established for each concentration range

[21]. The removal efficiency (%R) and adsorption capacity at time t (q_t, mg g^{-1}) were calculated according to (1) and (2):

$$\%R = \left(\frac{C_0 - C_t}{C_0}\right) \times 100 \quad (1)$$

$$q_t = \frac{(C_0 - C_t) \times V}{m} \quad (2)$$

where C_0 (mg L^{-1}) is the initial CV concentration, C_t (mg L^{-1}) is the residual concentration at time t, V (L) is the solution volume, and m (g) is the mass of adsorbent used in each assay.

Since all assays were conducted in the dark and in the absence of any catalytic or oxidizing agent, CV was expected to remain chemically stable in aqueous solution under these conditions [22], so a dye-only (adsorbent-free) blank was not deemed necessary to isolate the contribution of adsorption: any decrease in C_t measured throughout the assays could thus be attributed to adsorption onto the CeO_2-NPs rather than to spontaneous degradation of the dye.

The adsorption kinetics of CV onto CeO_2-NPs were evaluated using the nonlinear pseudo-first-order (PFO), pseudo-second-order (PSO), and intraparticle diffusion (IPD) models, expressed by (3), (4), and (5), respectively, for each of the four CV concentrations tested. The equilibrium behavior was described by fitting the Langmuir, Freundlich, and Sips isotherm models, expressed by (6), (7), and (8), respectively, to the q_e versus C_e equilibrium adsorption data, obtained at the end of the 180 min contact period for each of the four initial CV concentrations tested and combined into a single dataset [23].

$$q_t = q_1 \times [1 - \exp(-k_1 \times t)] \quad (3)$$

$$q_t = \frac{k_2 q_2^2 t}{1 + k_2 q_2 t} \quad (4)$$

$$q_t = k_d \times t^{0.5} + C \quad (5)$$

where k_1 (min $^{-1}$) and q_1 (mg g $^{-1}$) are the PFO rate constant and theoretical adsorption capacity; k_2 (g mg $^{-1}$ min $^{-1}$) and q_2 (mg g $^{-1}$) are the PSO rate constant and theoretical capacity; and k_d (mg g $^{-1}$ min $^{-0.5}$) and C (mg g $^{-1}$) are the intraparticle diffusion rate and boundary-layer constants, respectively.

$$q_e = \frac{q_{max} K_L C_e}{1 + K_L C_e} \quad (6)$$

$$q_e = K_F C_e^{1/n} \quad (7)$$

$$q_e = \frac{q_{max} (K_S C_e)^m}{1 + (K_S C_e)^m} \quad (8)$$

where q_{max} (mg g^{-1}) is the maximum adsorption capacity; K_L (L mg^{-1}) is the Langmuir constant; K_F (mg g^{-1})(L mg^{-1})$^{1/n}$ and $1/n$ are the Freundlich capacity constant and heterogeneity factor; and K_S (L mg^{-1}) and m are the Sips constant and heterogeneity exponent, respectively. Model quality was assessed by the coefficient of determination (R^2), following the same statistical treatment described in Section E.

E. Statistical Analysis

Experimental data were organized and processed using OriginPro 2025 (OriginLab Corporation, Northampton, MA, USA) [24]. The software was employed for data visualization, nonlinear curve fitting, and regression analysis of the adsorption kinetic and equilibrium isotherm models. To verify the prediction performance of the isotherm and kinetic models tested, the following error metric functions were used in Statistic® 10.0, expressed by (9), (10), and (11), respectively [25].

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - p)}{(n - p - 1)}\right] \quad (9)$$

$$SSE = \frac{1}{n}\sum_{i=1}^{n} (q_{exp} - \bar{q}_{pred})^2 \quad (10)$$

$$ARE = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{q_{exp} - \bar{q}pred}{qexp}\right] \quad (11)$$

where q_{exp} and q_{pred} are the experimental and predicted adsorption capacities (mg g^{-1}), respectively; $\bar{q}pred$ is the mean of the predicted values (mg g^{-1}); n is the number of data points; p is the number of model parameters; R^2 is the coefficient of determination (%); $Radj^2$ is the adjusted coefficient of determination (%); SSE is the sum of squared errors; and ARE is the mean relative error (%).

III. Results and Discussion

A. Characterization of the CeO_2-NPs

The XRD diffractogram of the green-synthesized CeO_2-NPs is presented in Fig. 1. The diffractogram exhibited characteristic diffraction peaks at approximately 28.6°, 33.1°, 47.5°, and 56.3°, indexed to the (020), (200), (220), and (311) crystallographic planes, respectively, corresponding to the cerianite phase of CeO_2-NPs (JCPDS/PDF No. 34-0394). These diffraction peaks confirm the formation of crystalline

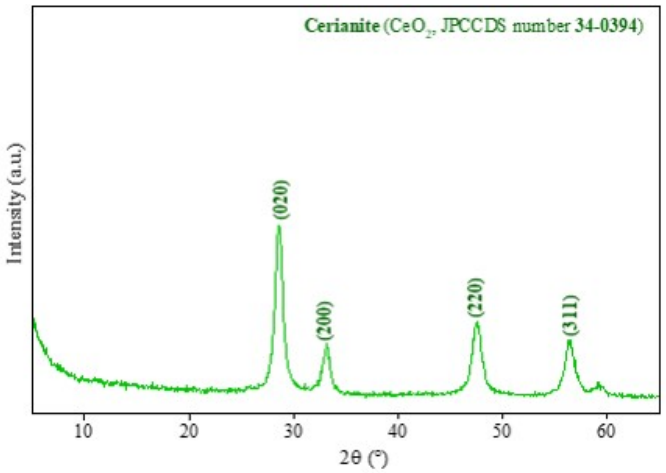


Fig. 1. XRD diffractograms of the CeO_2-NPs from *Ilex paraguariensis* extract.

CeO_2-NPs, in agreement with previous reports on green-synthesized CeO_2-NPs materials [26].

Furthermore, no additional diffraction peaks associated with secondary crystalline phases or precursor-derived impurities were detected, indicating the high phase purity of the synthesized material. The diffraction pattern agrees with previous studies on plant-mediated synthesis of CeO_2-NPs, demonstrating that the biosynthetic route produced crystalline cerium oxide without detectable secondary phases [27].

FEG-SEM micrographs of the green-synthesized CeO_2-NPs are presented in Fig. 2.

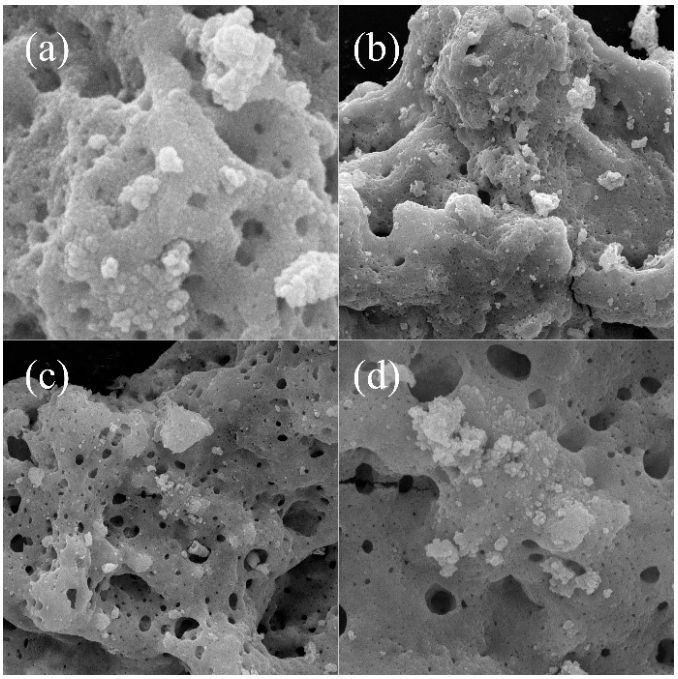


Fig. 2. FEG-SEM micrographs of the green-synthesized CeO_2-NPs obtained using *Ilex paraguariensis* leaf extract at different magnifications: (a) 160 kx, (b) 20 kx, (c) 20 kx, and (d) 80 kx.

Fig. 2a, acquired at 160 kx, reveals agglomerated, quasi-spherical nanoparticles distributed across the material surface. This agglomeration is common in metal oxide nanoparticles and is mainly attributed to their high surface energy and attractive interparticle interactions during synthesis and calcination [28]. Fig.s 2b and 2c, obtained at 20 kx, provide a broader view of the material, revealing a porous architecture composed of interconnected pores distributed throughout the sample. This morphology indicates the formation of an accessible porous network that may facilitate the adsorption of pollutant molecules and improve the contact between the adsorbent surface and the aqueous solution during the adsorption process [29].

At an intermediate magnification (80 kx), Fig. 2d shows the agglomerated CeO_2-NPs deposited on the porous matrix, providing a clearer visualization of the relationship between the nanostructured particles and the porous framework. Similar morphological features have been reported for green-synthesized CeO_2-NPs obtained using plant extracts, in which naturally occurring phytochemicals act as reducing and stabilizing agents during nanoparticle formation, influencing particle nucleation and aggregation [30].

Overall, the FEG-SEM analysis showed that the green synthesis route produced agglomerated CeO_2-NPs with a porous architecture. These morphological characteristics are consistent with previous reports of biosynthesized CeO_2-NPs and are expected to favor interaction between the adsorbent surface and organic pollutants, contributing to the adsorption of dye molecules from aqueous solution [31].

The elemental composition determined by EDS (Fig. 3) confirmed cerium and oxygen as the predominant elements in the synthesized material, corroborating the formation of CeO_2-NPs [32]. The Au signal originated from the conductive coating applied during SEM sample preparation and therefore does not represent the composition of the synthesized nanomaterial. The EDS spectrum revealed cerium as the predominant element (91.9 wt.%), whereas oxygen accounted for 4.5 wt.% of the detected composition. The remaining balance is tentatively attributed to residual carbon from the *Ilex paraguariensis* extract, which was not fully eliminated during calcination at 450 °C, although EDS quantification of light elements should be interpreted with caution.

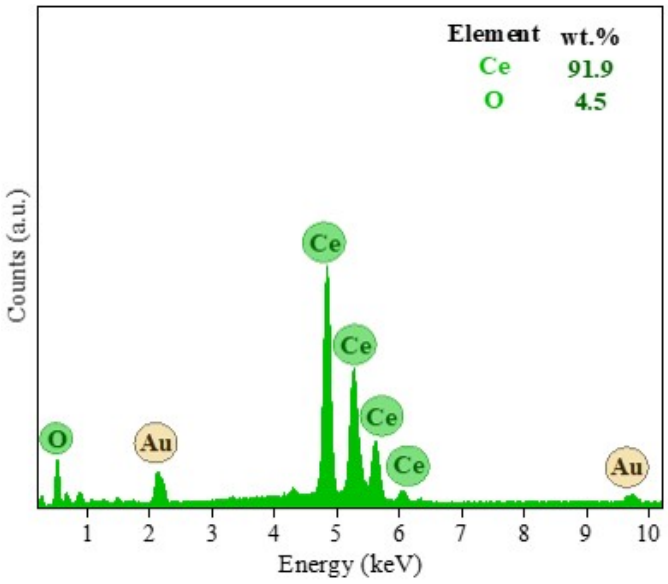


Fig. 3. Elemental analysis of the CeO_2-NPs from *Ilex paraguariensis* extract.

The structural, textural properties, and surface charge of the nanoparticles, quantified via N_2 physisorption and ZP, are consolidated in Table I.

TABLE I
TEXTURAL PROPERTIES AND ZETA POTENTIAL OF THE GREEN CeO_2-NPS

SBET ($m^2\ g^{-1}$)	V_p ($cm^3\ g^{-1}$)	D_p (nm)	ZP (mV)
85 ± 1.6	0.1 ± 0.01	7.8 ± 1.5	-10.9 ± 2.4

According to Table I, the green-synthesized CeO_2-NPs

exhibited a specific surface area of 85 ± 1.6 m^2 g^{-1}, a pore volume of 0.10 ± 0.01 cm^3 g^{-1}, a pore diameter of 7.8 ± 1.5 nm, and a zeta potential of -10.9 ± 2.4 mV. Based on the IUPAC classification, the obtained pore diameter indicates a mesoporous material (2–50 nm) [33]. The combination of a relatively high specific surface area, mesoporosity, and negatively charged surface contributes to the adsorption of the cationic CV dye molecules and supports their interaction with the adsorbent surface during the adsorption process [34]. Although the zeta potential of -10.9 ± 2.4 mV indicates moderate surface charge density, consistent with the agglomeration observed in FEG-SEM micrographs, the electrostatic attraction toward the cationic CV molecules remains operative [35].

B. Adsorption of Crystal Violet

The adsorption profiles of CV onto the green-synthesized CeO_2-NPs (0.5 g L^{-1}, pH 9) showed concentration-dependent removal behavior over the 180 min contact period, as summarized in Fig. 4.

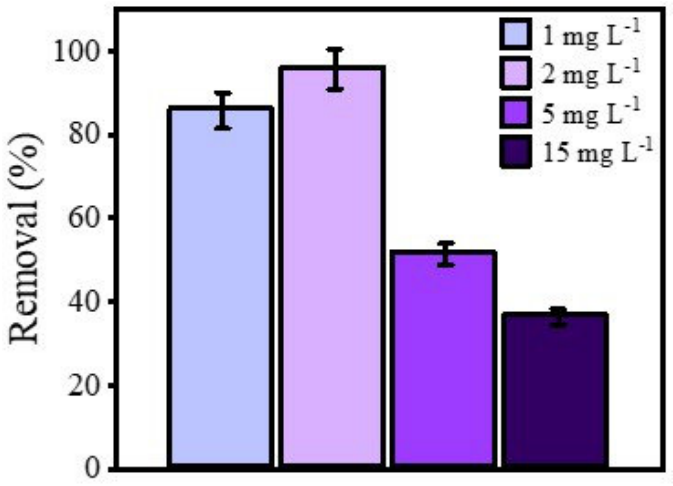


Fig. 4. Adsorption tests of CV under the conditions of CeO_2-NPs = 0.5 g L^{-1}, pH = 9, and T = 25 ± 2 °C, varying the CV concentration.

At the two lowest initial concentrations (1 and 2 mg L^{-1}), removal exceeded 85%, reaching a maximum of 95.6% at C_0 = 2 mg L^{-1}. Increasing the initial CV concentration to 5 and 15 mg L^{-1} decreased the removal efficiency to 51.4% and 36.4%, respectively. This behavior can be attributed to the fixed adsorbent dosage (0.5 g L^{-1}), which provides a limited number of active sites for dye uptake. As the initial CV concentration increases, the ratio of available adsorption sites to dye molecules decreases, resulting in progressive saturation of the adsorbent surface. Similar concentration-dependent decreases in dye removal have been reported for CeO_2-based adsorbents, where increasing pollutant concentration led to lower removal percentages as available surface sites became increasingly occupied [36], [37].

Table II summarizes the kinetic parameters obtained from the PFO, PSO, and IPD models. For the two lowest concentrations (1 and 2 mg L^{-1}), both PFO and PSO described the experimental data equally well ($R^2 \geq 0.99$), with fast rate constants (k_1 = 0.051 to 0.077 min^{-1}), indicating that equilibrium was essentially reached within the experimental time frame. At 5 and 15 mg L^{-1}, the fitted rate constants were markedly lower ($k_1 < 0.006$ min^{-1}), and the extrapolated theoretical capacities (q_1, q_2) considerably exceeded the q_t values actually attained at 180 min, suggesting that equilibrium had not yet been reached under these conditions and that the adsorption process is slower at higher CV loads. The IPD model gave lower R^2 for the low-concentration runs (0.629 to 0.738) with non-zero intercepts ($C \neq 0$), indicating that intraparticle diffusion is not the sole rate-limiting step and that boundary-layer (film) diffusion also contributes to the overall adsorption rate [38].

The equilibrium behavior, described in Table III, showed that the two-parameter Langmuir and Freundlich models provided a comparatively poor description of the data (R^2 = 0.998 and 0.999, respectively), while the three-parameter Sips model, which combines Langmuir- and Freundlich-type behavior, consistently gave the best statistical fit across all CV concentrations (R^2 = 0.999). This result indicates a heterogeneous adsorption process that approaches a finite maximum capacity, in agreement with the mesoporous, negatively charged surface (S_{BET} = 85 ± 1.6 m^2 g^{-1}; D_p = 7.8 ± 1.5 nm; ZP = -10.9 ± 2.4 mV) previously characterized for these nanoparticles, which favors electrostatic interaction with the cationic CV molecule but with site-to-site variability typical of green-synthesized, agglomerated metal-oxide nanoadsorbents [39].

The results in Table III indicate that the Sips model best described the adsorption of CV onto CeO_2-NPs. In this context, it exhibited the highest coefficients of determination (R^2 = 0.999 / R^2_{adj} = 0.999) and the lowest error values, outperforming the Langmuir and Freundlich models. The maximum theoretical adsorption capacity obtained using the Sips model was q_{max} = 0.289 mg g^{-1}.

Regarding the adsorption mechanism, the strong fit to the Sips model reflects the heterogeneous nature of the CeO_2-NP surface. The Sips exponent (m = -3.975) indicates that the process occurs at active sites with different binding energies, with heterogeneous adsorption at low CV concentrations and a tendency toward monolayer saturation at the available sites as the equilibrium concentration increases [40]. Under the experimental conditions employed, the predominant mechanism involves electrostatic attraction between CV molecules, a cationic dye with a positive charge, and the CeO_2-NP surface, which is deprotonated and negatively charged under alkaline conditions, along with possible hydrogen-bonding and Van der Waals interactions [41].

Overall, the results demonstrate that the green-synthesized CeO_2-NPs exhibit promising adsorptive performance toward CV under dark conditions, achieving a maximum removal efficiency of 95.6% at an initial concentration of 2 mg L^{-1}. Adsorption strongly depended on the initial CV concentration, with rapid uptake at lower concentrations and slower adsorption at higher pollutant loads, where the fixed adsorbent dosage limited active-site availability and delayed the

TABLE II
KINETIC ADSORPTION RESULTS OF CV USING THE KINETIC MODELS UNDER CONDITIONS: CeO_2-NPS = 0.5 G L^{-1}, PH = 9, AND T = 25 ± 2 °C, VARYING THE CV CONCENTRATION.

PPO						
[CV] (mg L^{-1})	**q_1 (mg g^{-1})**	**k_1 (min^{-1})**	**R^2**	**R_{adj}^2**	**ARE (%)**	**SSE**
1	0.00158	0.0509	0.991	0.988	2.76	3.57×10^{-9}
2	0.00761	0.0765	0.998	0.997	1.25	1.31×10^{-8}
5	0.3562	0.0004	0.960	0.995	86.02	6.04×10^{-6}
15	0.2595	0.0056	0.981	0.976	11.27	9.67×10^{-5}
PSO						
[CV] (mg L^{-1})	**q_2 (mg g^{-1})**	**k_2 (g mg^{-1} min^{-1})**	**R^2**	**R_{adj}^2**	**ARE (%)**	**SSE**
1	0.00179	40.12	0.992	0.990	3.11	3.21×10^{-9}
2	0.0081	17.38	0.997	0.996	1.58	1.31×10^{-8}
5	0.744	0.0002	0.960	0.950	85.90	6.04×10^{-6}
15	0.434	0.0079	0.980	0.975	11.26	9.90×10^{-5}
IDM						
[CV] (mg L^{-1})	**k_{dif} (mg g^{-1} $min^{-0.5}$)**	**C (mg g^{-1})**	**R^2**	**R_{adj}^2**	**ARE (%)**	**SSE**
1	0.000013	0.000809	0.738	0.672	12.00	4.09×10^{-8}
2	0.000051	0.00457	0.629	0.536	11.49	9.93×10^{-7}
5	0.00031	0.0012	0.963	0.953	46.39	5.39×10^{-6}
15	0.00183	0.0135	0.969	0.961	9.13	1.37×10^{-4}

Abbreviations: PPO = pseudo-first-order model; PSO = pseudo-second-order model; IDM = intraparticle diffusion model; k_1 = kinetic constant for PPO; q_1 = theoretical adsorption capacity for PPO; k_2 = kinetic constant for PSO; q_2 = theoretical adsorption capacity for PSO; k_{dif} = kinetic constant for the intraparticle diffusion model; C = constant related to diffusion resistance; R^2 = coefficient of determination; R_{adj}^2 = adjusted coefficient of determination; ARE = mean relative error; and SSE = sum of squared errors.

approach to equilibrium. The superior performance of the Sips model indicates a heterogeneous adsorption process, consistent with the mesoporous structure and negatively charged surface of the nanoparticles, which favor interactions with the cationic CV molecules. These findings support the potential of green-synthesized CeO_2-NPs as sustainable nanoadsorbents for cationic dye removal, while further studies involving real healthcare wastewater, competing contaminants, regeneration, reuse, and long-term stability are required to assess their practical applicability in healthcare wastewater treatment.

TABLE III
ISOTHERM ADSORPTION RESULTS OF CV USING THE MODEL ISOTHERMS UNDER CONDITIONS: CeO_2-NPS = 0.5 G L^{-1}, PH = 9, AND T = 25 ± 2 °C, USING THE EQUILIBRIUM (C_e, q_e) DATA COMBINED FROM ALL CV CONCENTRATIONS TESTED (1–15 MG L^{-1}).

Langmuir					
q_{max} (mg g^{-1})	**K_L (L mg^{-1})**	**R^2**	**R_{adj}^2**	**ARE (%)**	**SSE**
0.0701	-0.182	0.998	0.997	0.0763	1.15×10^{-7}
Freundlich					
K_F (mg g^{-1})(L mg^{-1})$^{1/n}$	**1/n**	**R^2**	**R_{adj}^2**	**ARE (%)**	**SSE**
5.192	-0.652	0.999	0.998	0.0356	2.50×10^{-8}
Sips					
q_{max} (mg g^{-1})	**K_S (L mg^{-1})**	**m**	**R^2**	**R_{adj}^2**	**ARE (%)**	**SSE**
0.289	0.098	-3.975	0.999	0.999	0.0081	1.40×10^{-9}

Abbreviations: q_{max} = maximum adsorption capacity of the material; K_L = Langmuir constant; K_F = Freundlich adsorption capacity constant; $1/n$ = constant related to surface heterogeneity; K_S = Sips constant; m = fractional exponent related to the Sips adsorption mechanism; R^2 = coefficient of determination; R_{adj}^2 = adjusted coefficient of determination; ARE = mean relative error; and SSE = the sum of squared errors.

C. Lessons Learned, Research Gaps and Perspectives

Despite these promising results, several research gaps remain before practical application can be considered. The present study evaluated a single model cationic dye in aqueous solution and did not investigate real healthcare wastewater, competing contaminants, different ionic strengths, adsorbent regeneration and reuse, long-term stability, potential cerium leaching, or process-level sustainability aspects such as synthesis yield, energy consumption during drying and calcination, and reagent/water use. Moreover, although a single additional temperature condition was evaluated, a systematic thermal study was not performed, preventing the determination of thermodynamic parameters such as the Gibbs free energy change (ΔG°), enthalpy change (ΔH°), and entropy change (ΔS°), which could provide further insight into the spontaneity, energetic nature, and temperature dependence of the adsorption process. The absence of direct comparisons with conventional or commercial CeO_2 and other reference adsorbents also limits the quantitative assessment of the material's relative performance.

Future studies should therefore investigate the adsorption process under more realistic conditions, including real or synthetic healthcare wastewater, different pH and ionic strength conditions, competing solutes, and contaminant mixtures. Temperature-dependent adsorption experiments should also be conducted to determine the thermodynamic parameters

and clarify the energetic characteristics and possible changes in the adsorption mechanism with temperature. Furthermore, regeneration cycles, nanoparticle recovery, reuse, long-term stability, and potential cerium release should be evaluated to assess the material's environmental safety and practical feasibility. Comparative studies with commercial CeO_2 and established adsorbents, along with a process-level or life-cycle assessment of the green synthesis route, would also be valuable for benchmarking adsorption performance and sustainability. From a Biomedical Engineering perspective, these investigations may contribute to the future development of sustainable nanoadsorbent-based strategies for treating healthcare-related wastewater.

IV. Conclusion

The present study demonstrated that the green synthesis of CeO_2-NPs using *Ilex paraguariensis* leaf extract is an effective plant-mediated route for obtaining crystalline mesoporous nanomaterials with physicochemical characteristics favorable for adsorption. Structural, morphological, elemental, and surface characterizations confirmed the formation of CeO_2-NPs, with a specific surface area of 85 $\pm$ 1.6 m^2 g^{-1}, a mean pore diameter of 7.8 $\pm$ 1.5 nm, and a negative zeta potential of -10.9 ± 2.4 mV.

Batch dark-adsorption assays demonstrated that CV removal by the green-synthesized CeO_2-NPs (0.5 g L^{-1}, pH 9) was strongly dependent on the initial pollutant concentration, reaching up to 95.6% removal at low CV concentrations (1–2 mg L^{-1}) and decreasing at higher concentrations (5–15 mg L^{-1}), consistent with the increasing demand for available adsorption sites at a fixed adsorbent dose. The kinetic and equilibrium analyses indicated concentration-dependent adsorption behavior, while the Langmuir, Freundlich, and Sips models provided high-quality fits to the equilibrium data, with the Sips model yielding R^2 = 0.999. These findings are consistent with heterogeneous adsorption behavior and highlight the role of the physicochemical characteristics of the CeO_2-NPs in CV uptake.

From a Biomedical Engineering perspective, the results demonstrate the potential of green-synthesized CeO_2-NPs as an adsorbent for the removal of CV, used here as a model cationic pollutant, from aqueous systems relevant to environmental and healthcare-related applications. Importantly, the present results establish adsorption performance under controlled laboratory conditions rather than treatment of real healthcare wastewater. Further studies involving real and complex wastewater matrices, competing contaminants, pH and ionic-strength effects, regeneration and reuse, long-term stability, and possible Ce leaching are necessary to determine the practical applicability of these nanomaterials in healthcare-related wastewater treatment systems.

Acknowledgment

The authors gratefully acknowledge the Franciscan University (UFN) for institutional support and for providing the infrastructure necessary for the development of this study. This research was financially supported by the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) through the Auxílio Recém-Doutor (ARD – Call 10/2024), Process No. 25/2551-0000968-5. The authors are also grateful to the collaborating universities for providing access to the facilities and instrumentation used for the physicochemical characterization of the materials. Finally, they acknowledge the continued support of public funding programs that foster scientific research and innovation in Brazil.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Boutros, R. Puig, E. Bartoli, and M. El Bachawati, "Sustainability assessment of hospital wastewater treatment techniques: a comprehensive review," *Sustainability*, vol. 17, no. 11, pp. 1–25, 2025.

[2] B. Lellis, C. Z. Fávaro-Polonio, J. A. Pamphile, and J. C. Polonio, "Effects of textile dyes on health and the environment and bioremediation potential of living organisms," *Biotechnology Research and Innovation*, vol. 3, no. 2, pp. 275–290, 2019.

[3] G. Canan-Rochenbach, M. A. Barreiros, A. O. Lima, A. D. Conti-Lampert, R. Ariente-Neto, W. Pimentel-Almeida, R. Laçoli, R. Corrêa, C. M. Radetski, and S. Cotelle, "Are hospital wastewater treatment plants a source of new resistant bacterial strains?" *Environmental Science and Pollution Research*, vol. 30, no. 50, pp. 108 635–108 648, 2023.

[4] J. O. Ighalo, P.-S. Yap, K. O. Iwuozor, C. O. Aniagor, T. Liu, K. Dulta, F. U. Iwuchukwu, and S. Rangabhashiyam, "Adsorption of persistent organic pollutants (POPs) from the aqueous environment by nano-adsorbents: a review," *Environmental Research*, vol. 212, no. 1, pp. 1–21, 2022.

[5] D. Joshy, S. Chakko, Y. A. Ismail, and P. Periyat, "Surface basicity mediated rapid and selective adsorptive removal of congo red over nanocrystalline mesoporous CeO_2," *Nanoscale Advances*, vol. 3, no. 23, pp. 6704–6718, 2021.

[6] Q. U. Din, M. C. Iannaco, I. De Marco, V. Vaiano, and G. Iervolino, "Non-thermal plasma-driven degradation of organic dyes using CeO_2 prepared by supercritical antisolvent precipitation," *Nanomaterials*, vol. 15, no. 23, pp. 1–26, 2025.

[7] S. Drummer, T. Madzimbamuto, and M. Chowdhury, "Green synthesis of transition-metal nanoparticles and their oxides: a review," *Materials*, vol. 14, no. 11, pp. 1–30, 2021.

[8] S. Jadoun, R. Arif, N. K. Jangid, and R. K. Meena, "Green synthesis of nanoparticles using plant extracts: a review," *Environmental Chemistry Letters*, vol. 19, no. 1, pp. 355–374, 2020.

[9] S. M. S. Palavicini, B. M. S. Puton, R. A. Jacques, E. Valduga, G. T. Backes, N. Paroul, C. Steffens, and R. L. Cansian, "Bioactive compounds of *Ilex paraguariensis*: a critical update on extraction, gastrointestinal stability, and technological applications," *Journal of Food Science*, vol. 90, no. 11, pp. 1–14, 2025.

[10] C. Liu, X. Liu, Y. Wu, Z. Chen, Z. Wu, S. Wang, H. Han, Z. Xie, Y. Wang, and T.-H. Ko, "Green synthesis of nanostructure CeO_2 using tea extract: characterization and adsorption of dye from aqueous phase," *Bioinorganic Chemistry and Applications*, vol. 2021, no. 1, pp. 1–15, 2021.

[11] N. Masood, M. A. Irshad, R. Nawaz, T. Abbas, M. A. Abdel-Maksoud, W. H. AlQahtani, H. AbdElgawad, M. Rizwan, and A. H. A. Abeed, "Green synthesis, characterization and adsorption of chromium and cadmium from wastewater using cerium oxide nanoparticles; reaction kinetics study," *Journal of Molecular Structure*, vol. 1294, no. 1, pp. 1–10, 2023.

[12] X. Borgohain, E. Das, and M. H. Rashid, "Facile synthesis of CeO_2 nanoparticles for enhanced removal of malachite green dye from an aqueous environment," *Materials Advances*, vol. 4, no. 2, pp. 683–693, 2023.

[13] United Nations. (2015) Transforming our world: the 2030 Agenda for Sustainable Development. Department of Economic and Social Affairs. Accessed: July 6, 2026. [Online]. Available: https://sdgs.un.org/2030agenda

[14] D. F. dos Santos, V. Alves, E. Costa, A. Martins, A. F. F. Vieira, G. H. F. dos Santos, C. T. d. P. Francisco, and V. Z. Pinto, "Yerba mate (*Ilex paraguariensis*) processing and extraction: retention of bioactive compounds," *Plant Foods for Human Nutrition*, vol. 78, no. 3, pp. 526–532, 2023.

[15] D. Moro Druzian, L. Rodrigues Oviedo, S. Nunes Loureiro, R. Dias Wouters, B. Stefanello Vizzotto, E. de Oliveira Pinto, N. Julia Schűssler de Vanconcellos, Y. Patricia Moreno Ruiz, A. Galembeck, G. Pavoski, D. Crocce Romano Espinosa, C. dos Santos, and W. Leonardo da Silva, "Cerium oxide nanoparticles: biosynthesis, characterization, antimicrobial, ecotoxicity and photocatalytic activity," *Journal of Photochemistry and Photobiology A: Chemistry*, vol. 442, no. 1, pp. 1–12, 2023.

[16] Fauzia, M. A. Khan, M. Chaman, and A. Azam, "Antibacterial and sunlight-driven photocatalytic activity of graphene oxide conjugated CeO_2 nanoparticles," *Scientific Reports*, vol. 14, no. 1, pp. 1–17, 2024.

[17] C. Hano and B. H. Abbasi, "Plant-based green synthesis of nanoparticles: production, characterization and applications," *Biomolecules*, vol. 12, no. 1, pp. 1–9, 2021.

[18] M. Thommes, K. Kaneko, A. V. Neimark, J. P. Olivier, F. Rodriguez-Reinoso, J. Rouquerol, and K. S. Sing, "Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (IUPAC technical report)," *Pure and Applied Chemistry*, vol. 87, no. 9-10, pp. 1051–1069, 2015.

[19] A. Haleem, A. Shafiq, S.-Q. Chen, and M. Nazar, "A comprehensive review on adsorption, photocatalytic and chemical degradation of dyes and nitro-compounds over different kinds of porous and composite materials," *Molecules*, vol. 28, no. 3, pp. 1–38, 2023.

[20] N. Fifere, M. M. Lazar, I. E. Raschip, A. Airinei, C.-D. Varganici, and M. V. Dinu, "Preparation and characterization of chemically crosslinked Xanthan/Poly(Vinylalcohol) hydrogel films containing cerium oxide nanoparticles for potential application in removal of methylene blue and crystal violet dyes," *Gels*, vol. 11, no. 10, pp. 1–22, 2025.

[21] Q. Maqbool, I. Cavallini, N. Lasemi, S. Sabbatini, F. Tittarelli, and G. Rupprechter, "Waste-valorized nanowebs for crystal violet removal from water," *Small Science*, vol. 4, no. 4, pp. 1–12, 2024.

[22] P. Sharma, S. Kar, M. Sahu, and M. Ganguly, "Copper-based nanoparticles for the removal of the crystal violet dye *via* degradation and adsorption: a comparative account," *RSC Advances*, vol. 15, no. 34, pp. 27 995–28 020, 2025.

[23] A. S. Alawam, S. M. Mahgoub, A. A. Allam, A. M. Radalla, M. H. Shemy, and R. Mahmoud, "Sustainable zeolite/activated carbon nanocomposite for dual-dye adsorption, kinetic and thermodynamic studies, and urea electro-oxidation applications," *RSC Advances*, vol. 15, no. 51, pp. 43 983–44 006, 2025.

[24] F. H. Emamy, A. Bumajdad, and J. P. Lukaszewicz, "Adsorption of hexavalent chromium and divalent lead ions on the nitrogen-enriched chitosan-based activated carbon," *Nanomaterials*, vol. 11, no. 8, pp. 1–24, 2021.

[25] F. F. Severo, P. C. L. Muraro, J. S. De Oliveira, L. F. Rodrigues Junior, W. L. Da Silva, and M. A. Zazycki, "Adsorção do corante rodamina b usando cinzas de casca de arroz," *Disciplinarum Scientia - Ciências Naturais e Tecnológicas*, vol. 26, no. 2, pp. 51–64, 2025.

[26] R. A. El-Salamony and W. A. Aboutaleb, "Preparation, characterization of green synthesis CeO_2 nanoparticles and their photocatalytic activity towards malachite green dye," *ChemistrySelect*, vol. 7, no. 9, pp. 1–10, 2022.

[27] Y. Wu and H. T. Ta, "Different approaches to synthesising cerium oxide nanoparticles and their corresponding physical characteristics, and ROS scavenging and anti-inflammatory capabilities," *Journal of Materials Chemistry B*, vol. 9, no. 36, pp. 7291–7301, 2021.

[28] F. Sargin and F. Ak Azem, "Hydrothermal synthesis of CeO_2: Structure–adsorption performance relationship in methyl orange dye removal," *Nanomaterials*, vol. 16, no. 5, pp. 1–28, 2026.

[29] Y. Xu, L. Gao, J. Yang, Q. Yang, W. Peng, and Z. Ding, "Effective and efficient porous CeO_2 adsorbent for acid orange 7 adsorption," *Materials*, vol. 16, no. 7, pp. 1–16, 2023.

[30] F. T. Geldasa, M. A. Kebede, M. W. Shura, and F. G. Hone, "Experimental and computational study of metal oxide nanoparticles for the photocatalytic degradation of organic pollutants: a review," *RSC Advances*, vol. 13, no. 27, pp. 18 404–18 442, 2023.

[31] A. A. Belew and M. A. Assege, "Solvothermal synthesis of metal oxide nanoparticles: a review of applications, challenges, and future perspectives," *Results in Chemistry*, vol. 16, no. 1, pp. 1–20, 2025.

[32] A. Dhouib, B. Mezghrani, G. Finocchiaro, R. Le Borgne, M. Berthet, B. Daydé-Cazals, A. Graillot, X. Ju, and J.-F. Berret, "Synthesis of stable cerium oxide nanoparticles coated with phosphonic acid-based functional polymers," *Langmuir*, vol. 39, no. 23, pp. 8141–8152, 2023.

[33] I. Matsukevich, N. Kulinich, L. Kulbitskaya, T. Kuznetsova, V. Popkov, M. Chebanenko, D. Moskovskikh, K. Kuskov, and V. Romanovski, "Mesoporous nanocomposites based on CeO_2 and MgO: preparation, structure and photocatalytic activity," *Journal of Chemical Technology & Biotechnology*, vol. 98, no. 10, pp. 2497–2505, 2023.

[34] G. Xiang, S. Long, and A. Dang, "Fabrication of the ordered mesoporous nZVI/Zr-Ce-SBA-15 composites used for crystal violet removal and their optimization using RSM and ANN–PSO," *Sustainability*, vol. 14, no. 11, pp. 1–26, 2022.

[35] M. Loganathan, A. S. Raj, A. Murugesan, and P. S. Kumar, "Effective adsorption of crystal violet onto aromatic polyimides: kinetics and isotherm studies," *Chemosphere*, vol. 304, no. 1, pp. 1–14, 2022.

[36] E. Rápó and S. Tonk, "Factors affecting synthetic dye adsorption; desorption studies: a review of results from the last five years (2017–2021)," *Molecules*, vol. 26, no. 17, pp. 1–31, 2021.

[37] N. Semwal, D. Mahar, M. Chatti, A. Dandapat, and M. C. Arya, "Adsorptive removal of congo red dye from its aqueous solution by Ag–Cu–CeO_2 nanocomposites: adsorption kinetics, isotherms, and thermodynamics," *Heliyon*, vol. 9, no. 11, pp. 1–18, 2023.

[38] M. Ashour, A. E. Alprol, M. Khedawy, K. M. Abualnaja, and A. T. Mansour, "Equilibrium and kinetic modeling of crystal violet dye adsorption by a marine diatom, *Skeletonema costatum*," *Materials*, vol. 15, no. 18, pp. 1–28, 2022.

[39] C. Hu, Y. Huang, and R. Rao, "Tuning surface termination of CeO_2 nanoparticles for adsorption of anionic dyes via urea content in a simple pyrolysis technique," *Applied Surface Science*, vol. 635, no. 1, pp. 1–11, 2023.

[40] G. de Vargas Brião, M. A. Hashim, and K. H. Chu, "The Sips isotherm equation: often used and sometimes misused," *Separation Science and Technology*, vol. 58, no. 5, pp. 884–892, 2023.

[41] S. M. Mahgoub, A. S. Alawam, A. A. Allam, M. R. Mahmoud, A. Elrafey, A. H. Abdelmohsen, and R. Mahmoud, "Crystal violet removal from aqueous solution: adsorption studies, box–behnken optimization, safety, and reuse in methanol oxidation," *RSC Advances*, vol. 15, no. 52, pp. 44 373–44 391, 2025.

Antimicrobial Activity of Plasma-Activated Water and Silver Nanoparticles Against Staphylococcus aureus

Sophia Franzoni Azevedo Souza
Laboratory of Bionanotechnology Institute of Research and Development Universidade do Vale do Paraíba
São José dos Campos, Brazil
Sophiafranzoni2001@gmail.com

Maiara Lima Castilho
Laboratory of Bionanotechnology Institute of Research and Development Universidade do Vale do Paraíba
São José dos Campos, Brazil
mcastilho@univap.br

***Abstract*— The progressive increase in bacterial resistance to conventional antimicrobial agents demands new therapeutic strategies. This study evaluated the antimicrobial activity and associative potential of silver nanoparticles (AgNPs) and Plasma-Activated Water (PAW) against *Staphylococcus aureus* (ATCC 25923). PAW was produced using a cold atmospheric plasma discharge in a gliding arc system. Microbial growth assays by optical density (OD600) demonstrated that PAW alone (activated for 5, 15, and 30 minutes) did not significantly reduce cell growth. The isolated AgNP treatment showed a massive reduction in bacterial proliferation only at concentrations of 12.5% and 25%. However, combining PAW (5 minutes of activation) with a lower, sub-inhibitory AgNP concentration (6.25%) resulted in an equivalent severe reduction in bacterial growth. The results, validated by one-way ANOVA followed by Dunnett's and Tukey's post hoc tests, indicate that this association is highly effective, enabling a 50% reduction in the required nanomaterial dosage for massive bacterial inactivation. Ultimately, this approach effectively reduces the required nanomaterial dosage, representing a promising *in vitro* antibacterial strategy that warrants further investigation regarding cytotoxicity and clinical applicability.**

Keywords—Silver nanoparticles, Plasma-activated water, Antibacterial association, Staphylococcus aureus

I. INTRODUCTION

The progressive increase in bacterial resistance to conventional antimicrobial agents represents one of the most critical challenges for global public health in the 21st century [1, 2]. Opportunistic microorganisms, such as *Staphylococcus aureus* (Gram-positive), stand out due to their high clinical relevance and their capacity to develop resistance to multiple drugs, which complicates the treatment of hospital and chronic infections [1, 3]. In clinical practice, infected cutaneous wounds are particularly concerning, as the presence of resistant pathogens can severely impede the natural healing process and lead to serious complications [4, 5].

In this context, nanotechnology has emerged as a strategic frontier, with silver nanoparticles (AgNPs) being widely investigated due to their broad spectrum of biocidal action [2, 3]. AgNPs synthesized by the bottom-up method allow for rigorous control of physicochemical properties that determine their interaction with the cell wall, destabilization of cytoplasmic membranes, and induction of intracellular oxidative stress [6, 7]. However, the search for strategies that enhance their efficacy at reduced concentrations is fundamental to mitigate potential systemic toxicity and adverse side effects, ensuring safer applications in human healthcare and wound management [8].

Concurrently, atmospheric pressure cold plasma technology has consolidated itself as an innovative tool with promising biomedical and therapeutic applications [9, 10]. The interaction of plasma with water generates Plasma-Activated Water (PAW), a solution rich in reactive oxygen and nitrogen species (RONS), such as hydrogen peroxide (H_2O_2), nitrites (NO_2^-), and nitrates (NO_3^-) [11, 12, 13]. These species promote an oxidative and acidified environment that causes irreversible structural and metabolic damage to bacterial cells, while also regulating inflammatory responses to aid in tissue repair [5, 9, 10].

The rationale of this work lies in investigating the association potential between AgNPs and PAW produced by a gliding arc discharge. The central hypothesis is that the use of PAW as a washing and dispersion medium for nanoparticles can potentiate the bactericidal activity of the system, allowing for a significant reduction in the metabolic viability of clinical pathogens like *S. aureus* even at low concentrations of silver [2, 13]. Therefore, the aim of this study was to evaluate the antimicrobial activity of silver nanoparticles dispersed in Plasma-Activated Water (PAW) produced by a gliding arc discharge against Staphylococcus aureus.

II. METHODOLOGY

A. Synthesis of Silver Nanoparticles (AgNPs)

The silver nanoparticles were synthesized via the bottom-up method [14, 15]. Prior to the procedure, all glassware and magnetic stir bars were washed with aqua regia (a 1:3 ratio of nitric acid to hydrochloric acid) to remove contaminating metals and subsequently rinsed with ultrapure water. In this method, silver nitrate (AgNO3) is reduced by sodium borohydride (NaBH4) to form a colloidal solution. Initially, 30 mL of a 1 mM NaBH4 solution was cooled to approximately 0 °C in an ultrathermostatic bath and maintained under constant magnetic stirring at 1100 rpm for 1 hour. Subsequently, 10 mL of a 2 mM AgNO3 solution was added to the mixture, allowing the reaction to proceed for 7 minutes. The colloidal solution was then removed from the cooling system, stirred at 300 rpm for an additional hour, and left to rest under light for 24 hours. The sample was characterized by UV-Vis spectroscopy and stored in the dark.

B. Production of Plasma-Activated Water (PAW)

The Plasma-Activated Water was produced by exposing 100 mL of distilled water (under magnetic stirring at 400 rpm) to a gliding arc plasma discharge. The system operated at a power of 18.4 W and a frequency of 60 Hz, utilizing a gas mixture flow of 6 L/min of Argon and 4 L/min of compressed

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

air. The distance between the plasma outlet and the sample surface was maintained at 2 mm. The water samples were activated for intervals of 5, 15, and 30 minutes.

C. Bacterial Strains and Antimicrobial Assays

The antimicrobial efficacy against *Staphylococcus aureus* (ATCC 25923) strains was evaluated using microbial growth assays by measuring the optical density at 600 nm (OD600). The bacterial inoculum was standardized at 10^5 CFU/mL. All experiments were conducted in 96-well microplates and incubated for 24 hours at 37 °C. A positive control (C+) consisting exclusively of BHI broth and the bacterial inoculum was included in all assays to establish the baseline for uninhibited bacterial growth.

The experimental design was divided into three stages. For the isolated PAW assessment, the bacterial inoculum was treated with water activated for 5, 15, or 30 minutes. In the isolated AgNPs assessment, the inoculum was exposed to an initial nanoparticle concentration of approximately 1.21×10^{10} AgNPs/mL, which was serially diluted (two-fold) to evaluate the dose-dependent inhibitory effect. For the associative assessment, the lower sub-inhibitory concentrations of AgNPs (3×10^9 and 1.5×10^9 AgNPs/mL) were prepared using PAW activated for 5 minutes as the diluent medium instead of standard water, prior to incubation with the bacterial inoculum.

D. Statistical Analysis

All quantitative data obtained from the OD600 readings were statistically analyzed using GraphPad Prism 8.0.1 software. Data normality was verified prior to analysis using the Shapiro-Wilk test. Subsequently, a one-way Analysis of Variance (ANOVA) was performed, followed by Dunnett's post hoc test (to compare isolated treatments against the control) and Tukey's post hoc test (to compare differences among multiple specific groups in the associative assays), considering $P < 0.05$ as statistically significant.

III. Results and Discussion

The synthesis of silver nanoparticles was confirmed and characterized by UV-Vis spectroscopy. As shown in Fig. 1, the colloidal AgNP solution exhibited a distinct surface plasmon resonance (SPR) absorption band centered at approximately 400 nm, a hallmark optical feature widely reported for structurally stable silver nanoparticles [7, 16]. Ultimately, this spectral behavior is characteristic of the well-controlled bottom-up chemical reduction method using sodium borohydride, corroborating the structural parameters described in previous literature [15].

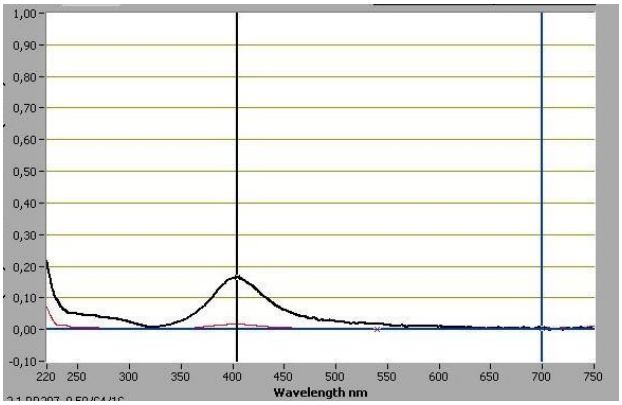


Fig. 1. UV-Vis absorption spectrum of the synthesized AgNPs, displaying a characteristic surface plasmon resonance peak at ~400 nm (n=3).

Furthermore, the physical characterization of the synthesized nanomaterial was complemented by Dynamic Light Scattering (DLS) analysis. As depicted in Fig. 2, the colloidal solution exhibited a Z-average hydrodynamic diameter of 10.99 nm with a Polydispersity Index (PdI) of 0.866. The symmetry and narrow profile of the previously observed SPR peak, combined with this DLS profile, strongly indicate the successful formation of nanoparticles with a homogeneous size distribution and low polydispersity [14, 16]. Achieving a low degree of polydispersity is a crucial parameter in nanomaterial synthesis, as it reflects a uniform colloidal suspension with minimized particle aggregation. This structural homogeneity ensures a high and consistent surface-to-volume ratio, which is vital for maximizing the contact area and optimizing the subsequent antimicrobial interaction with bacterial membranes [14, 16].

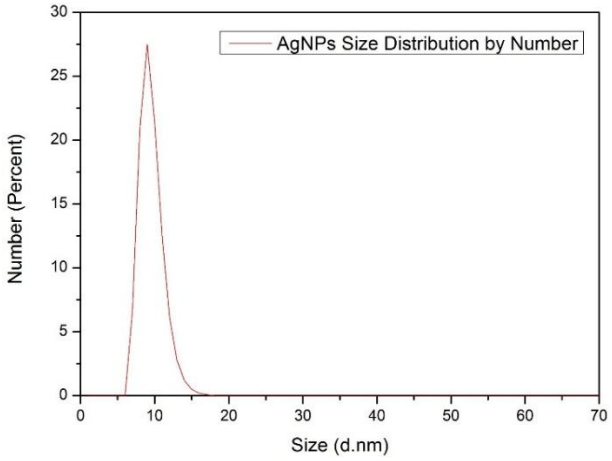


Fig. 2. Dynamic Light Scattering (DLS) analysis of the synthesized AgNPs, showing the hydrodynamic size distribution (n=3).

Regarding the antimicrobial efficacy, the isolated treatment with silver nanoparticles demonstrated a clear dose-dependent inhibitory effect on bacterial growth (Fig. 3). Statistical analysis showed that lower concentrations (3.125% and 6.25%) significantly differed from the positive control (*$p \leq 0.05$), promoting a partial growth reduction of approximately 30%. However, these sub-inhibitory doses did not completely arrest cellular proliferation. Complete inactivation was observed at a concentration of 12.5%, achieving an approximately 96.5% reduction in bacterial growth compared to the control (****$p \leq 0.0001$), a profile that was maintained at the 25% concentration.

This dose-dependent behavior is consistent with the mechanisms described by P. Parvekar et al. (2020), who observed that while AgNPs efficiently disrupt cell membranes and induce intracellular oxidative stress, their isolated biocidal potential relies heavily on the applied concentration. The necessity of a relatively high dose (12.5%) to achieve massive bacterial death corroborates the discussions raised by K. Markowska et al. (2025) and L. O. Camargo et al. (2022). These authors noted that treating resilient pathogens with isolated nanoparticles often requires elevated concentrations, which can be a critical limiting factor in clinical and environmental applications due to potential residual cytotoxicity.

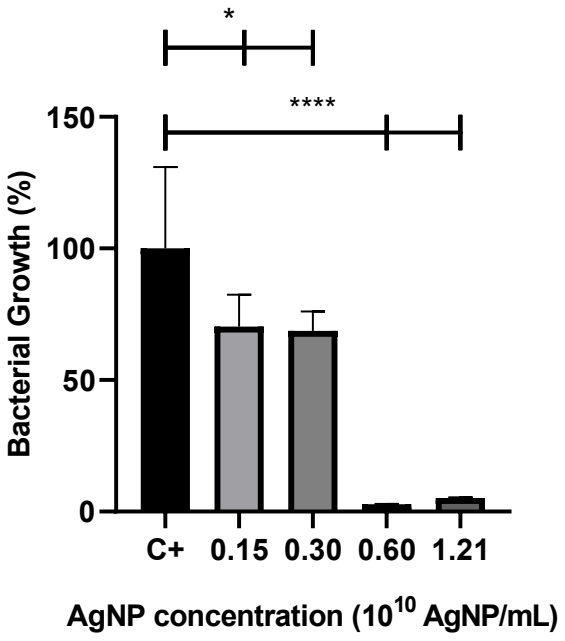


Fig. 3. Analysis of *S. aureus* ATCC 25923 growth following isolated treatment with different concentrations of AgNPs. Data are expressed as a percentage of growth relative to the positive control. *p ≤ 0.05; ****p ≤ 0.0001 (n=3).

The evaluation of the bacterial growth of *S. aureus* ATCC 25923 exposed exclusively to Plasma-Activated Water (PAW) is presented in Fig. 4. The optical density readings indicated that treatments with PAW activated for 5, 15, and 30 minutes did not induce a massive reduction in bacterial growth compared to the positive control (C+). The one-way ANOVA revealed no statistically significant differences among the distinct activation times, nor between them and the control. Consequently, the 5-minute activation time was selected for the subsequent associative assays, as extending the plasma discharge time provided no additional inhibitory benefit.

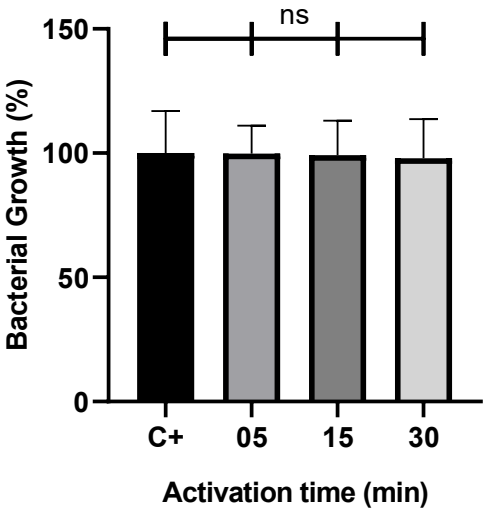


Fig. 4. Analysis of *S. aureus* ATCC 25923 growth following isolated treatment with PAW activated at intervals of 5, 15, and 30 minutes. Data are expressed as a percentage of growth relative to the positive control. ns = not significant, p > 0.05 (n=3).

These results differ from studies such as those by Y. M. Zhao et al. (2022) and A. Mai-Prochnow et al. (2021), which report high isolated biocidal activity of PAW, including the effective inactivation of planktonic cells and the disruption of mature biofilms. However, they strongly align with findings by Z. Xu et al. (2020), that highlighted Gram-positive bacteria, such as *S. aureus*, exhibit higher resistance to reactive oxygen and nitrogen species (RONS) due to their remarkably thick peptidoglycan layer, which acts as a robust physical barrier. Furthermore, it´s been demonstrated that PAW efficacy is highly dependent on the plasma source parameters. In our study, the activation of a relatively large liquid volume (100 mL) at a power of 18.4W likely resulted in a diluted concentration of RONS. Physicochemical measurements of the PAW supported this hypothesis. Immediately after activation, PAW treated for 5 minutes exhibited an acidic pH and a conductivity of approximately 33 µS/cm. Interestingly, monitoring this solution post-activation revealed aprogressive increase throughout the 2-hour evaluation period. In contrast, extending the plasma discharge to 15 and 30 minutes further acidified the initial medium and increased the starting conductivity to 107 µS/cm and 122 µS/cm, respectively. However, the conductivity for these longer activation times remained relatively stable post-activation, returning near their initial values after 2 hours. Despite these distinct chemical dynamics, the RONS concentration generated in this specific volume remained insufficient to independently breach the cellular defense and cause lethal metabolic damage across all evaluated times, validating the choice of the 5-minute activation for the associative assays.

The association potential was evaluated utilizing PAW (activated for 5 minutes) as the dispersion medium for sub-inhibitory concentrations of AgNPs (3.125% and 6.25%) (Fig. 5). According to Dunnett's multiple comparisons test, the combined treatment with the lowest concentration (3.125% AgNP + PAW) did not exhibit a statistically significant difference when compared to the positive control (ns; p = 0.0799).

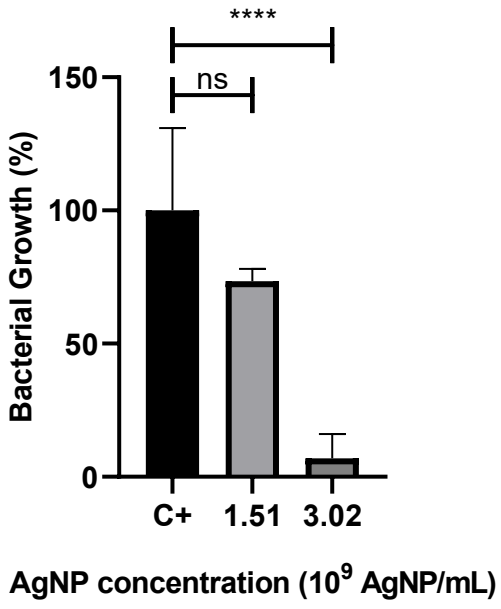


Fig. 5. Analysis of *S. aureus* ATCC 25923 growth following associated treatments (AgNP + PAW) using sub-inhibitory nanoparticle concentrations. Data are expressed as a percentage of growth relative to the positive control. ns = not significant (p > 0.05); ****p ≤ 0.0001. (n = 3).

However, when the 6.25% AgNP concentration (3.02 × 10^9 AgNP/mL) was associated with PAW, a severe drop in bacterial growth was observed, reaching basal viability levels (approximately 93% reduction). According to Tukey's multiple comparisons test, this specific combination presented a highly significant difference not only against the positive control (****; $p < 0.0001$) and the lower combined dose (****; $p < 0.0001$), but also when compared directly to the isolated 6.25% AgNP group (***; $p = 0.0002$) (Fig. 6). This confirms that the associative efficacy is strictly dependent on reaching this specific threshold, effectively reducing the required silver dosage by 50% to achieve massive inactivation.

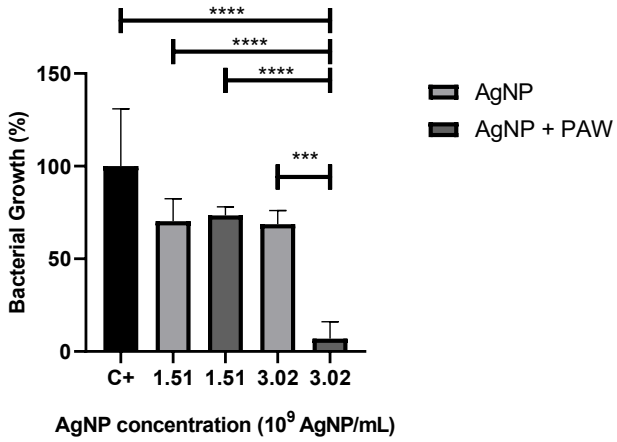


Fig. 6. Statistical comparison of *S. aureus* ATCC 25923 growth inhibition highlighting the associative efficacy between isolated AgNPs and the combined treatment (AgNP + PAW). Data are expressed as a percentage of growth relative to the positive control. ns = not significant ($p > 0.05$); ***$p = 0.0002$; ****$p < 0.0001$. ($n = 3$).

These results strongly support the synergistic mechanisms described in recent literature by B. Nowruzi et al. (2024), S. Wang et al. (2021) and H. R. Lee et al (2021). As proposed by these authors the initial oxidative stress provoked by the PAW, is insufficient to cause cellular death on its own, but compromises the bacterial envelope and begins to deplete endogenous antioxidant defenses, such as staphyloxanthin. This preliminary destabilization explains our findings, as it accelerates the internalization of silver ions. As demonstrated by B. Nowruzi et al. (2024) in combined therapies, this simultaneous attack from RONS and AgNPs overwhelms the bacterial repair pathways, causing irreversible metabolic collapse at significantly lower nanomaterial concentrations. This associative potential is driven by the structural destabilization of the thick Gram-positive cell wall, facilitating a severe and continuous transmembrane influx of silver ions.

Although these in vitro results against a reference strain are highly promising, the exact associative mechanism requires further formal synergy modeling and direct biological validation. Ongoing future studies are focused on the direct chemical quantification of specific RONS, such as hydrogen peroxide, nitrite, and nitrate using targeted colorimetric and reduction assays, to fully elucidate the lethal oxidative pathways. Furthermore, comprehensive nanomaterial characterization and cytotoxicity assessments, will be essential to application biomedical safety and efficacy of this combined approach.

IV. Conclusion

The present study demonstrated the highly effective associative antimicrobial action between silver nanoparticles (AgNPs) synthesized via the bottom-up method and Plasma-Activated Water (PAW) against *Staphylococcus aureus* ATCC 25923. While PAW activated for 5 minutes exhibited no isolated inhibitory effect, and isolated AgNPs required a concentration of 12.5% to achieve a massive reduction in cellular proliferation, their combination allowed for a 50% reduction in the required nanomaterial dosage. Specifically, utilizing PAW as a dispersion medium for a 6.25% AgNP concentration achieved an equivalent basal bacterial viability, an associative efficacy that was statistically validated by Tukey's multiple comparisons test. Ultimately, this *in vitro* combined approach overcomes the limitations of isolated treatments by minimizing the necessary nanomaterial concentration. Future validations involving cytotoxicity assessments and resistant clinical strains are required to fully establish its translational biomedical applicability.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 88887.134975/2025-00.

References

[1] GBD 2021 Antimicrobial Resistance Collaborators,"Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," *Lancet*, vol. 399, no. 10325, pp. 629–655, 2024, doi: 10.1016/S0140-6736(21)02724-0.

[2] K. Markowska, A. M. Grudniak, K. Krawczyk, I. Wróbel and K. I. Wolska, "Modulation of antibiotic resistance and induction of a stress response in *Pseudomonas aeruginosa* by silver nanoparticles," *J. Med. Microbiol.*, vol. 63, no. 6, pp. 849–854, 2025, doi: 10.1099/jmm.0.068833-0.

[3] H. Al-Momani *et al.*, "The efficacy of biosynthesized silver nanoparticles against *Pseudomonas aeruginosa* isolates from cystic fibrosis patients," *Sci. Rep.*, vol. 13, no. 8876, 2023, doi: 10.1038/s41598-023-35919-6.

[4] A. Mai-Prochnow, R. Zhou, T. Zhang, K. Ostrikov, S. Mugunthan, S. A. Rice, and P. J. Cullen, "Interactions of plasma-activated water with biofilms: inactivation, dispersal effects and mechanisms of action," *npj Biofilms Microbiomes*, vol. 7, no. 11, 2021, doi: 10.1038/s41522-020-00180-6.

[5] Z. Xu *et al.*, "*In vitro* antimicrobial effects and mechanism of air plasma-activated water on *Staphylococcus aureus* biofilm," *Plasma Process. Polym.*, vol. 17, no. 1900270, 2020, doi: 10.1002/ppap.201900270.

[6] S. Khana, M. Zahoorb, R. S. Khana, M. Ikramc and N. U. Islam, "The impact of silver nanoparticles on the growth of plants: The agriculture applications," *Heliyon*, vol. 9, no. 6, 2023, doi: 10.1016/j.heliyon.2023.e16928.

[7] [7] H. K. Tawfeeq, N. K. Abdalameer, R. H. Jassim, and M. M. Shehab, "Silver nanoparticles synthesized by cold plasma as an antibiofilm agent against *Staphylococcus epidermidis* isolated from acne," *Nano Biomed. Eng.*, vol. 16, no. 1, pp. 101–109, 2024, doi: 10.26599/NBE.2023.9290042.

[8] A. H. Aldujaily, K. F. Hassoon, A. A. Hatem, and M. Abbas, "The effect of biosynthesized silver nanoparticles on *Pseudomonas aeruginosa*-infected dogs wounds," *PLOS ONE*, vol. 20, no. 0313166, 2025, doi: 10.1371/journal.pone.0313166.

[9] Y. M. Zhao, S. Ojha, C. M. Burgess, D. W. Sun, and B. K. Tiwari, "Inactivation efficacy of plasma-activated water: influence of plasma treatment time, exposure time and bacterial species," *Int. J. Food Sci. Technol.*, vol. 56, pp. 721–732, 2021, doi: 10.1111/ijfs.14708.

[10] S. Wang, D. Xu, M. Qi, B. Li, S. Peng, Q. Li, H. Zhang, and D. Liu, "Plasma-activated water promotes wound healing by regulating inflammatory responses," *Biophysica*, vol. 1, no. 3, pp. 297–310, 2021, doi: 10.3390/biophysica1030022.

[11] R. Zhou *et al.*, “Plasma activated water (PAW): generation, origin of reactive species and biological applications,” *J. Phys. D: Appl. Phys.*, vol. 53, no. 30, 2020, doi: 10.1088/1361-6463/ab81cf.

[12] V. S. Urquia, “Água ativada por plasma: estado da arte e estudos preliminares do efeito da temperatura sobre os mecanismos de geração de espécies reativas de oxigênio e nitrogênio,” Bachelor's in Physics, UNESP, Guaratinguetá, 2025.

[13] R. Thirumdas, A. Kothakota, U. Annapure, K. Siliveru, R. Blundell, R. Gatt, and V. P. Valdramidis, “Plasma activated water (PAW): Chemistry, physico-chemical properties, applications in food and agriculture,” *Trends Food Sci. Technol.*, vol. 77, pp. 21–31, 2020, doi: 10.1016/j.tifs.2018.05.007.

[14] I. Fontoura, T. S. Veriato, L. J. Raniero, and M. L. Castilho, “Analysis of capped silver nanoparticles combined with imipenem against different susceptibility profiles of *Klebsiella pneumoniae*,” *Antibiotics*, vol. 12, no. 535, 2023, doi: 10.3390/antibiotics12030535.

[15] L. C. S. Belusso *et al.*, “Synthesis of silver nanoparticles from bottom up approach on borophosphate glass and their applications as SERS, antibacterial and glass-based catalyst,” *Appl. Surf. Sci.*, vol. 473, pp. 303–312, 2019, doi: 10.1016/j.apsusc.2018.12.155.

[16] L. O. Camargo, I. Fontoura, T. S. Veriato, L. Raniero, and M. L. Castilho, “Antibacterial Activity of silver nanoparticles functionalized with amikacin applied against multidrug-resistant Acinetobacter baumannii,” *Am. J. Infect. Control*, vol. 000, pp. 1–8, 2022, doi: 10.1016/j.ajic.2022.12.009.

[17] P. Parvekar, J. Palaskar, S. Metgud, R. Maria, and S. Dutta, “The minimum inhibitory concentration (MIC) and minimum bactericidal concentration (MBC) of silver nanoparticles against *Staphylococcus aureus*,” *Biomater. Investig. Dent.*, vol. 7, no. 1, pp. 105–109, 2020, doi: 10.1080/26415275.2020.1796674.

[18] B. Nowruzi, H. Beiranvand, F. M. Aghdam, and R. Barandak, “The effect of plasma activated water on antimicrobial activity of silver nanoparticles biosynthesized by cyanobacterium *Alborzia kermanshahica*,” *BMC Biotechnol.*, vol. 24, no. 75, 2024, doi: 10.1186/s12896-024-00905-x.

[19] H. R. Lee *et al.*, “Liquid plasma as a treatment for cutaneous wound healing through regulation of redox metabolism,” *Cell Death Dis.*, vol. 14, no. 119, 2021, doi: 10.1038/s41419-023-05610-9.

[20] T. S. Veriato, I. Fontoura, L. D. Oliveira, L. J. Raniero, and M. L. Castilho, “Nano-antibiotic based on silver nanoparticles functionalized to the vancomycin–cysteamine complex for treating *Staphylococcus aureus* and *Enterococcus faecalis*,” *Pharmacol. Rep.*, vol. 75, pp. 951–961, 2023, doi: 10.1007/s43440-023-00491-3.